THE GREAT COURSES®

Topic
Science & Mathematics

Subtopic
Mathematics

Big Data: How Data Analytics Is Transforming the World

Course Guidebook

Professor Tim Chartier
Davidson College

PUBLISHED BY:

THE GREAT COURSES
Corporate Headquarters
4840 Westfields Boulevard, Suite 500
Chantilly, Virginia 20151-2299
Phone: 1-800-832-2412
Fax: 703-378-3819
www.thegreatcourses.com

Tim Chartier, Ph.D.

Associate Professor of Mathematics
and Computer Science
Davidson College

Professor Tim Chartier is an Associate Professor of Mathematics and Computer Science at Davidson College. He holds a B.S. in Applied Mathematics and an M.S. in Computational Mathematics, both from Western Michigan University. Professor Chartier received his Ph.D. in Applied Mathematics from the University of Colorado Boulder. From 2001 to 2003, at the University of Washington, he held a postdoctoral position supported by VIGRE, a prestigious program of the National Science Foundation that focuses on innovation in mathematics education.

Professor Chartier is a recipient of a national teaching award from the Mathematical Association of America (MAA). He is the author of *Math Bytes: Google Bombs, Chocolate-Covered Pi, and Other Cool Bits in Computing* and coauthor (with Anne Greenbaum) of *Numerical Methods: Design, Analysis, and Computer Implementation of Algorithms*. As a researcher, he has worked with both Lawrence Livermore National Laboratory and Los Alamos National Laboratory, and his research was recognized with an Alfred P. Sloan Research Fellowship.

Professor Chartier serves on the editorial board for *Math Horizons*, a magazine published by the MAA. He chairs the Advisory Council for the National Museum of Mathematics, which opened in 2012 and is the first mathematics museum in the United States. In 2014, he was named the inaugural Math Ambassador for the MAA.

Professor Chartier writes for *The Huffington Post*'s "Science" blog and fields mathematical questions for *Sport Science* program. He also has been a resource for a variety of news outlets, including Bloomberg TV, the *CBS Evening News*, National Public Radio, the *New York Post*, *USA TODAY*, and *The New York Times*. ■

Table of Contents

Table of Contents

Table of Contents

Big Data:
How Data Analytics Is Transforming the World

Scope:

Thanks to data analytics, enormous and increasing amounts of data are transforming our world. Within the bits and bytes lies great potential to understand our past and predict future events. And this potential is being realized. Organizations of all kinds are devoting their energies to combing the ever-growing stores of high-quality data.

This course demonstrates how Google, the United States Postal Service, and Visa, among many others, are using new kinds of data, and new tools, to improve their operations. Google analyzes connections between web pages, a new idea that propelled them ahead of their search engine competitors. The U.S. Postal Service uses regression to read handwritten zip codes from envelopes, saving millions of dollars in costs. Visa employs techniques in anomaly detection to identify fraud—and today can look at all credit card data rather than a sampling—and with such advances comes more accurate methods.

This course will help you understand the range of important tools in data analytics, as well as how to learn from data sets that interest you. The different tools of data analysis serve different purposes. We discuss important issues that guide all analysis. We see how dangerously prone humans are to finding patterns. We see how the efficiency of algorithms can differ dramatically, making some impractical for large data sets. We also discuss the emerging and important field that surrounds how to store such large data sets.

An ever-present issue is how to look at data. Important questions include what type of data you have and whether your data is robust enough to potentially answer meaningful questions. We discuss how to manage data, and then how to graph it.

Graphing the data, or some portion of it, is a key exploratory step. This, if nothing else, familiarizes you with the data. This can help focus your

questions, because aimless analysis can be like searching haystacks with no idea of what counts as a needle. Good graphics can also figure centrally in the final presentation of stories found in the data. In between, graphic analysis can also produce meaningful results throughout a data analysis.

A key issue early in the process of data analysis is preparing the data, and we see the important step of splitting data. This important but overlooked step makes it possible to develop ("train") a meaningful algorithm that produces interesting analysis on some of the data, while holding in reserve another part of the data to "test" whether your analysis can be predictive on other data.

This course shares a large variety of success stories in data analysis. While interesting in their own right, such examples can serve as models of how to work with data. Once you know your data, you must choose how to analyze your data. Knowing examples of analysis can guide such decisions.

Some data allows you to use relatively simple mathematics, such as the expected value, which in sports analytics can become the expected number of wins in a season based on current team statistics. Such formulas led to the success of the Oakland A's in 2002, as detailed in the book and movie *Moneyball*.

Is the recency of the data important, with older data being less predictive? We see how techniques for weighting and aggregating data from polls allowed Nate Silver and others to transform the use of polling data in politics.

Data analytics draws on tools from statistical analysis, too. Regression, for example, can be used to improve handwriting recognition and make predictions about the future.

If you know, in a general way, which variables are important and don't need to assess their relative importance, then artificial intelligence could be a good next step. Here, a computer learns how to analyze the data—from the data itself.

Anomaly detection enables credit card companies to detect fraud and reduce the risk of fraud. It also enables online gaming companies to detect anomalous patterns in play that can indicate fraudulent behavior.

When data involves vast numbers of possibilities, analysis can turn to simulating a phenomenon on a computer. Such techniques enable the aerodynamics of cars to be tested before a prototype is constructed and lead to the special effects we see in movies and scientific visualizations.

The ability to determine which variables are influential is quite important. In fact, including too many variables can lead to the pitfall known as overfitting, where methods may perform stunningly well on past data but are terrible at predicting future data.

Data mining, which involves looking for meaning within larger data sets, often makes use of linear algebra. This mathematical tool starts like high school algebra, except we put our equations into a matrix form. From there, performing even a complex matrix analysis can be as simple as pushing a button on a computer. So, the key becomes understanding what we are doing. Linear algebra lies at the core of Google's ability to rank web pages, the determination of schedule strength for a sports team to better predict future, and the entire field of data compression.

Another approach early in an analysis, if the data is looking at a single "root" variable, is decision trees, which split data in order to predict disease, for example. Sometimes, decision trees suffice as a stand-alone analytical tool. Other times, they can be used like a sieve, to prepare the data for other methods, thereby jump-starting the analysis. And when no single master variable is targeted, many other methods for clustering are used—for example, Netflix and many other companies profile their customers.

We can also study data about relationships, allowing one to determine who is at the center of Hollywood or professional baseball, along with the validity of the claim that everyone on our planet is connected by six people, or by six degrees of separation.

A key insight we keep in mind amid all the hype about "big data" is that small data sets continue to offer meaningful insights. Beware of thinking that you need more data to get results; we see how more data can make the analysis more difficult and unwieldy. Returning to the haystack analogy, we want to avoid making a bigger haystack without including any more needles.

Thinking like a data analyst also involves realizing that previous ideas can be extended to other applications. Conversely, no single tool answers all questions equally: A different tool may tell a different story.

Our modern data deluge offers a treasure trove of exciting opportunities to unveil insight into our world. We can understand how data analytics has already transformed many current practices, as well as how we can better navigate further changes into the future. ■

Data Analytics—What's the "Big" Idea?

Lecture 1

The field of data analysis relates to, and impacts, our world in unprecedented ways. Right now, millions, even billions, of computers are collecting data. From smartphones and tablets to laptops and even supercomputers, data is an ever-present and growing part of our lives. What makes data analytics so powerful are the fundamental techniques you will learn for analyzing data sets. Data analysis is a set of existing and ever-developing tools, but it is also a mindset. It's a way of improving our ability to ask questions, and it's an expectation that data can make possible new answers.

Big Data

- A 2012 Digital Universe study estimated that the global volume of digital data stored and managed in 2010 was over a trillion gigabytes—which is equivalent to a billion terabytes (so, less than one terabyte per person at that point), a million petabytes, a thousand exabytes, or a zettabyte. That was in 2010, and the number was predicted to double every year, reaching 40 trillion gigabytes by the year 2020.

- Those numbers are for all the data—no one person or computer has all the data that's distributed over all the computing devices everywhere. Still, even individual data sets are huge. In fact, so many applications are creating data sets that are so big that the ways we traditionally have analyzed data sometimes do not work.

- Indeed, the ideas we have today might not solve the questions we have for the data tomorrow. As more and more data is collected, and as the technology we use to collect that data changes, new questions will arise, which may mean we need new ways to analyze the data to gain insight.

- Data analysis is a fairly new combination of applied mathematics and computer science, available in ways that would have been difficult to imagine a few decades ago and inconceivable 100 years ago. Why? A lot of it has to do with data. And this flood of new data is being organized, analyzed, and put to use.

- For example, online companies like Amazon, Netflix, and Pandora are gathering rating after rating from millions of people and then putting all of that data to use. Similar transformations are taking place in politics, sports, health care, finance, entertainment, science, industry, and many more realms.

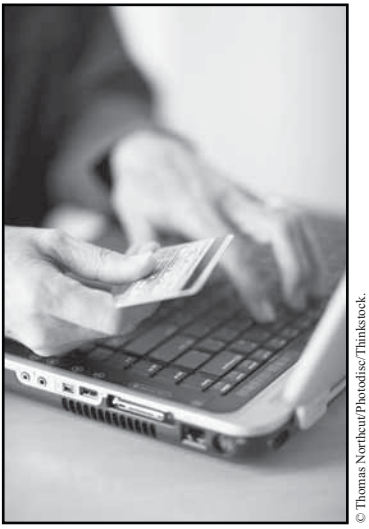
- We are collecting data as never before, and that creates new kinds of opportunities and challenges. In fact, so many applications are creating data sets that are so big that the ways we traditionally have analyzed data don't work. Indeed, the ideas we have today might not solve the questions we have for the data of tomorrow. This is the idea behind the term "big data"—where the sheer size of large data sets can force us to come up with new methods we didn't need for smaller data sets.

The Three Vs of Big Data

- Big data is often defined as having three Vs: volume, velocity, and variety. First, in terms of volume, which would you say is bigger: the complete works of Shakespeare or an ordinary DVD? The complete works of Shakespeare fit in a big book, or roughly 10 million bytes. But any DVD—or any digital camera, for that matter—will hold upward of 4 gigabytes, which is 4 billion bytes. A DVD is 400 times bigger.

- And data is not merely stored: We access a lot of data over and over. Google alone returns to the web every day to process another 20 petabytes—which is equal to 20,000 terabytes, 20 million gigabytes, or 20 quadrillion bytes. Google's daily processing gets us to 1 exabyte every 50 days, and 250 days of Google processing may be equivalent to all the words ever spoken by mankind to date, which

have been estimated at 5 exabytes. And nearly 1,000 times bigger is the entire content of the World Wide Web, estimated at upward of 1 zettabyte, which is 1 trillion gigabytes. That's 100 million times larger than the Library of Congress. And of course, there is a great deal more that is not on the web.



The Internet is revolutionizing the ways we send and receive data.

- Second is the velocity of data. Not only is there a lot of data, but it is also coming at very high rates. High-speed Internet connections offer speeds 1,000 times faster than dial-up modems connected by ordinary phone lines. Every minute of the day, YouTube users upload 72 hours of new video content. Every minute, in the United States alone, there are 100,000 credit card transactions, Google receives over two million search queries, and 200 million e-mail messages are sent.

- Third, there is variety. One reason for this can stem from the need to look at historical data. But data today may be more complete than data of yesterday. We stand in a data deluge that is showering large volumes of data at high velocities with a lot of variety. With all this data comes information, and with that information comes the potential for innovation.

- We all have immense amounts of data available to us every day. Search engines almost instantly return information on what can seem like a boundless array of topics. For millennia, humans have relied an each other to recall information. The Internet is changing that and how we perceive and recall details in the world.

- Human beings tend to distribute information through what is called a transactive memory system, and we used to do this by asking each other. Now, we also have lots of transactions with smartphones and other computers. They can even talk to us.

- In a study covered in *Scientific American*, Daniel Wegner and Adrian Ward discuss how the Internet can deliver information quicker than our own memories can. Have you tried to remember something and meanwhile a friend uses a smartphone to get the answer? In a sense, the Internet is an external hard drive for our memories.

Smartphones are becoming increasingly able to complete transactions that would be difficult for our brains.

- So, we have a lot of data, with more coming. What works today may not work tomorrow, and the questions of today may be answered only to springboard tomorrow's ponderings. But most of all, within the data can exist insight. We aren't just interested in the data; we are looking at data *analysis*, and we want to learn something valuable that we didn't already know.

- You don't need large data sets to pose computationally intensive problems. And even on a small scale, such problems can be too difficult to allow for optimal solutions.

- Data analysis doesn't always involve exploring a data set that is given. Sometimes, questions arise and data hasn't yet been gathered. Then, the key is knowing what question to ask and what data to collect.

- How big and what's big enough depends, in part, on what you are asking and how much data you can handle. Then, you must consider how you can approach the question.

Misconceptions about Data Analysis

- There are several misconceptions about data analysis. First, data analysis gives you *an* answer, not *the* answer. In general, data analysis cannot make perfect predictions; instead, it might predict better than we usually could without it. There is more than one answer. Much of life is too random and chaotic to capture everything—but it's more than that. Unlike math, data analytics does not get rid of all the messiness. So, you create an answer anyway and try to glean what truths and insights it offers. But it's not the *only* answer.

- Second, data analysis *does* involve your intuition as a data analyst. You are not simply number crunching. If you build a model and create results that go against anything anyone previously has found, it is likely that your model has an error.

- Third, there is no single best tool or method. In fact, many times, part of the art and science of data analysis is figuring out which method to use. And sometimes, you don't know. But there are some methods that are important to try before others. They may or may not work, and sometimes you simply won't know, but you can learn things about your data and viable paths to a solution by trying those methods.

- Fourth, you do not always have the data you need in the way you need it. Just having the data is not enough. Sometimes, you have the data, but it may not be in the form you need to process it. It may have errors, may be incomplete, or may be composed of different data sets that have to be merged. And sometimes just getting data into the right format is a big deal.

- Fifth, not all data is equally available. It is true that some data sets are easy to find. They already exist on the Internet. You can download them and immediately begin analyzing the data. But other pieces of data may not be as easily available. It doesn't mean that you can't get it. It's out there, but you need to figure out how to grab it.

- Sixth, while an insight or approach may add value, it may not add enough value. Not every new and interesting insight is worth the time or effort needed to integrate it into existing work. And no insight is totally new: If everything is new, then something is probably wrong.

Suggested Reading

Davenport, *Big Data at Work*.

Paulos, *Innumeracy*.

Activities

1. One way to open your mind to the prevalence of data is to simply stay attuned to your use of it. As you rate films or songs, use a credit card, make a phone call, or update your status on Facebook, think about the data being created. It is also interesting to look for news stories on data and to take note when new sources of data are available.

2. If you pay any bills online, look for the availability to download your own data. Whether you can or do, what might you analyze? What might you be able to find or see? How much data do you expect to be in the file?

3. An important part of this course is learning tools of data analysis and applying them to areas of your personal interest. What interests you? Do you want to improve your exercise? Do you want to have a better sense of how you use your time? Furthermore, think about areas of our world where data is still unruly. What ideas do you have that might tame the data and make it more manageable? You may or may not be able to implement such ideas, but beginning to look at the world in this way will prepare you to see the tools we learn as methods to answering those questions in the data deluge.

Got Data? What Are You Wondering About?

Lecture 2

In very broad strokes, there are three stages of data analysis: collecting the data, analyzing the data (which, if possible, includes visualizing the data), and becoming a data collector—not of everything, but in a purposeful way. After all, no one has time to gather and analyze all data of potential interest, just as no one has time to read every worthwhile book. In fact, you have a better chance of being able to analyze every book that interests you than of being able to analyze the quantity and variety of all types of data available today.

Collecting Data

- Data analysis is not just about large organizations and large data sets. Data analytics is also about individual people. Your financial details can be monitored and analyzed as never before. Your medical data and history can be organized to give you, as well as professional caregivers, unprecedented insight.

- It is now much easier to track, adjust, and understand any aspect of everyday life: including eating, sleep, activity levels, moods, movements, habits, communications, and so on. These are exciting changes. They make possible new fields such as personalized medicine, lifelogging, and personal analytics of all kinds. And they are coming together for some of the same reasons that data analytics as a whole is taking off.

- First, the wide range of technologies and methods available to large organizations are also increasingly available to individuals. Virtually all the tools you will learn in this course—such as grouping data into clusters, or finding correlations with regression, or displaying data with infographics—can also be used on your own data. You can use the ideas and techniques of data analytics to live a healthier life, save money, be a better coach to others in many areas of life, and so on.

- Second, large organizations are already accumulating more and more data about you and your world. So, they are, in effect, doing a lot of heavy lifting on your behalf, if you choose to access and make use of the data thereby accumulated. Even so, there is plenty of personal data analysis you can do, even without needing access to large data sets or sophisticated tools.

- There is a lot of data that could be kept on whatever you want to analyze. But the possibilities, and the realities, of having a lot of data can also be overwhelming. Keep in mind that even simple analysis, without extensive data, can give insight.

- A data analysis cycle involves collecting, analyzing, questioning, making a change, and then reanalyzing the data to understand what is happening.

- Today, there are many digital devices that aid with collecting exercise data, for example. The devices may keep track of how far you ran, biked, or swam. They can break down your exercise by the mile or minute, and some also give you some analysis. A device may log your steps, or how many calories you've burned, or the amount of time you've been idle during the day, and sometimes your location, whether you are climbing, your heart rate, and so on. Then, they can connect to applications on your computer or smartphone and give you feedback.

Smartphones are now able to log, store, and analyze lots of data, including running data.

- Why bother with all that data? It can offer insight. If you like to walk, does it make a difference where you walk? If you walk on a trail versus pavement, or if you walk one scenic route versus another, do these make differences? It isn't always that you need to change something; sometimes, it is simply a matter of having the knowledge to be informed about your choice.

- But while gathering data is worthwhile, be careful not to assume that data automatically means insight. In particular, just buying a device and collecting data does not mean that you'll gain insight. Many companies have made that mistake—collecting evermore data to try to take advantage of data analytics. They have more data. But, again, that doesn't necessarily mean that they'll gain insight, even if they attempt an analysis.

- In 2013, *The Wall Street Journal* reported that 44 percent of information technology professionals said they had worked on big-data initiatives that got scrapped. A major reason for so many false starts is that data is being collected merely in the hope that it turns out to be useful once analyzed. In the same *Wall Street Journal* article, Darian Shirazi, founder of Radius Intelligence Inc., describes the problem as "haystacks without needles." He notes that companies "don't know what they're looking for, because they think big data will solve the problem."

- On the other hand, once you have a goal, or a question you want to answer, you'll have much more success, both immediately and over the longer term. In fact, once you know what you are trying to learn, you can often think quite creatively about how to collect the data.

- So, having a clear goal makes a huge difference. Instead of just piling up data in the hopes that insight will pop out, having a clear goal guides you into gathering data that can produce insight. And with a bit of creativity, gathering the data may be much less onerous than you think. In fact, sometimes you already have data, but you may not realize that you do.

Analyzing Data

- When comparing data for two things in an attempt to analyze the data, you could compare two projects at work, two schools, two recipes, two vehicles, or two vacations—really anything that interests you.

- Or you can compare just one case to typical values for that one case. This is what a student learning analytics did at Mercer College. He was taking a class from Dr. Julie Beier, who asked her students to track personal data. The student was concerned about his aunt, who used the free clinic in town and had diabetes. The student felt that her medication was incorrectly calibrated. So, he kept track of her glucose levels, which she was already measuring.

- The student gathered this data and compared it with acceptable values, and it looked high. He could have easily stopped there. In fact, in data analytics, that's often where we do stop. But he even used a statistical test to see how likely it was that the readings would be that high, just by chance. That's called hypothesis testing, and it's sort of statistical inference traditionally called in when your sample is only a tiny part of a large population. But in data analytics, we are often studying a whole population or zooming in on a specific case.

- The main point, from the perspective of data analytics, is that he collected the data and compared to see what it meant. With his newfound information, the student and his aunt walked into the doctor's office with the data and conclusions. The result was a change in the aunt's medication. What is needed is data aimed at answering a question.

Becoming a Data Collector

- The nature of data analysis is that we don't have everything, but we can work with the data we do have to learn and gain insight. And the process of questioning is important. Data analytics offers new insight, but not all at once. With insight comes knowledge but also the potential to learn more. So, be prepared. Once data helps you

answer one question, you are likely to have another, and you will have to go back for more data. But you can keep digging, learning, and improving your decisions along the way.

- So, this is what it's like to begin as a data analyst. Collect data associated with a question that interests you. Keep your own interests in mind. These days, there are various ways to share your data, giving you and others more opportunities to learn from whatever you gather. Today, many devices can directly display and share the data, not only with your own computer but even with social media.

- What do you care about? If you see a connection to your life, jot it down so you can look at it. And then think about how to gather the necessary data. If you have it, great; if not, think about gathering it. Remember that it doesn't have to be a lot of data. Start somewhere. Also, look for opportunities to share, which is fun and helps you learn more from your data.

- Next, as you learn the tools of data analytics, think about which tools might apply to the data and address questions you have. Keep in mind that you may want to try a few methods so that you get different insights on the data. And remember that visualizing data often helps a lot. Whether the data is big or small, visualization can help you see when your sleeping patterns changed, for example, or what is happening during a sport or other physical activity.

Suggested Reading

Gray and Bounegru, *The Data Journalism Handbook*.

Russell, *Mining the Social Web*.

Activities

1. A key to data analytics is data. Papers, such as the *Guardian*, have data sites with downloadable data related to their articles. Look for such sites and see what data sets interest you.

2. The following are a few other data repositories for you to explore and begin thinking of questions that interest you. Having your questions can help you frame what you might do with the tools we will learn.

https://www.data.gov/
http://r-dir.com/reference/datasets.html
http://www.pewresearch.org/data/download-datasets/

A Mindset for Mastering the Data Deluge

Lecture 3

We stand within a data explosion of sorts. Organizations talk about trying to drink from a "fire hose" of information. Commentators refer to a "data deluge." But there is no need to drown in data. In this lecture, you will learn how data analysts of many kinds think about their data—the amount of data, the types of data, what constraints there may be on an analysis, and what data is not needed. In this way, you will learn how the deluge can be put to work, answering questions you have by developing the mindset of a data analyst.

The Size of Data

- There is a lot of data, and it can be difficult to wrap one's mind around the huge numbers. But it is possible. As data analysts, this is what we do. With data analysis, data can be managed in a way that's both timely and useful.

- Keep in mind that advances in storage play into this. Consider 50 GB, which is about the amount of storage on a Blu-ray disc. This would hold the textual content of just about a quarter of a million books. That's simply a disc you might have laying around. Storage capacity of a high-end drive from companies like Seagate or Western Digital can hold 5 terabytes or more. A terabyte is 1,000 gigabytes.

- With all this data, we begin to see why there began to be a lot of talk about big data. But without analysis, the data is essentially a lot of 1s and 0s. If you can't analyze it, it may not be helpful. Proper analysis can enable one to gain insight even with big data. But "big" is a relative term. We may call something "big" and only mean it in the context of data that in some other arena might seem small.

- Our sense of size changes with time, too. The Apollo 11 computers were fast and stored a lot of data for the time. Later, during the time of floppy discs, holding 1.44 MB, a gigabyte seemed about as remote as the petabyte or exabyte are for many of us today.

- To learn about size, we need to learn about how we measure and have a sense of what each measurement means.
    - A bit is a single binary digit. It equals 0 or 1—"on" or "off" in the hardware. A bit is a single character of text.

    - Eight bits make up a byte. Ten bytes is a written word.

    - One kilobyte is equal to 1,000 bytes and equals a short paragraph. Two kilobytes is equal to a typewritten page.

    - A megabyte is equal to 1,000 kilobytes and equals a short novel. Ten megabytes is enough for the complete works of Shakespeare.

    - Seven minutes of high-definition television video is 1 gigabyte, or 1,000 megabytes. A DVD can hold from 1 to 15 gigabytes; Blu-ray disks can hold 50 to 100 gigabytes.

    - One thousand gigabytes equates to 1 terabyte. Ten terabytes equals all the text information in books held by the U.S. Library of Congress, and 400 terabytes might be sufficient to hold all the books ever written.

    - A thousand terabytes is equal to 1 petabyte, or 10 million four-drawer filing cabinets filled with text.

    - One thousand petabytes is 1 exabyte. All the words ever spoken by mankind one decade into the 21st century may have equaled about 5 exabytes.

- One thousand exabytes equals 1 zettabyte. This is roughly the scale of the entire World Wide Web, which may be doubling in size every 18 months or so, with 1 zettabyte reached perhaps in the year 2011.

- One thousand zettabytes equals 1 yottabyte, which is 1 quadrillion gigabytes. Using a standard broadband connection, it would take you 11 trillion years to download a yottabyte. For storage, 1 million large data centers would be roughly 1 yottabyte.

Analyzing Big Data

- Many people are working with big data and trying to analyze it. For example, NASA has big data on a scale that can challenge current and future data management practice. NASA has over 100 missions concurrently happening. Data is continually streaming from spacecraft on Earth and in space, faster than they can store, manage, and interpret it.

- One thing about some of the largest data sets is that they are often being analyzed to find one specific thing. But many data sets are much more complex. In fact, when things get too big, you sometimes peel off part of your data to make it more manageable.

- But excluding some of the data is important on much smaller scales, too. How much data can you omit from a large data set and still be okay for the question you are investigating? Moreover, do you lose insight if you omit? Could excluding such data be relevant to issues that interest you, but you simply don't know it yet? Such challenges are inherent in data analysis, making it both more difficult and more interesting.

- In fact, a concern with the term "big data" is that although you can do amazing things with big-data sets, the field is not merely about a few big businesses. The same lessons can apply to all of us. Sometimes, the data is already available. The trick is recognizing how much to use and how to use it.

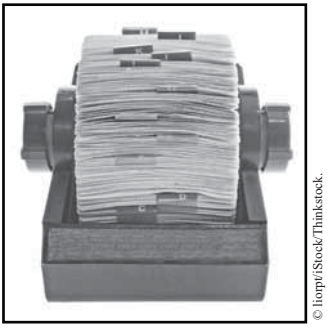
- Sometimes, relevant information comes from returning to the same places many times. In other situations, we might not even know what we don't know. Gus Hunt of the Central Intelligence Agency stated this really well in a 2013 talk. He noted, "The value of any information is only known when you can connect it with something else which arrives at a future point in time." We want to connect the dots, but we may not yet have the data that contains the dot to connect. So, this leads to efforts to collect and hang on to everything.

- Furthermore, data from the past may not have been stored in a usable way. So, part of the data explosion is having the data today and for tomorrow. The cost of a gigabyte in the 1980s was about a million dollars. So, a smartphone with 16 gigabytes of memory would be a 16-million-dollar device. Today, someone might comment that 16 gigabytes really isn't that much memory. This is why yesterday's data may not have been stored, or may not have been stored in a suitable format, compared to what can be stored today.

Structured versus Unstructured Data

- A common way to categorize data is into two types: structured data and unstructured data. Just this level of categorization can help you learn more about your data—and can even help you think about how you might approach the data.

- First, structured data is the type of data that many people are most accustomed to dealing with, or thinking of, as data. Your list of contacts (with addresses, phone numbers, and e-mail addresses) and recipes are examples of structured data. It can be a bit surprising that most experts agree that structured data accounts for only about 20 percent of the data out there.



Lists of addresses, phone numbers, and e-mail addresses are examples of structured data.

- There are two sources of structured data: computer-generated data and human-generated data. The boundary between computer-generated and human-generated data is not fixed. For example, a doctor may personally input medical information into a case file, but that might appear in combination with data read automatically from routine scans or computer-based lab work.

- Second, unstructured data doesn't follow a prespecified format. While perhaps 80 percent of identified data comes in this form, until recently we didn't have mechanisms for analyzing it. In fact, there were even problems just storing it, or storing it in a way that could be readily accessed.

- Scientific data often is unstructured and can be anything from seismic imagery to atmospheric data. Everyday life also produces a lot of unstructured data. There are e-mails, text documents, text messages, and updates to sites like Facebook, Twitter, or LinkedIn. There is also web site content that's added to video and photography sites like YouTube or Instagram.

- If data is structured, it is more likely that a method, possibly around for some time, has been developed to analyze it. If data is unstructured, this is much less likely. A million records in a structured database are much easier to analyze than a million videos on YouTube. Unstructured data can still have some structure, but overall, the data is much more unstructured.

- Part of what it means to think like a data analyst is deciding what to analyze and how. This is always important. In addition to getting the data and having a sense of what form it comes in, you need to consider how quickly it comes in. Is the data going to come in real time? If so, how quickly will you need to analyze it?

- Today, knowing where data is coming from, how much of it is coming, and how quickly you are going to need to analyze it are all very real and very important questions. The amount of data needed for a problem depends in part on what you are asking and how much data you can handle. Then, you must consider how you can approach the question.

Suggested Reading

Brenkus, *The Perfection Point*.

Mayer-Schönberger and Cukier, *Big Data*.

Activities

1. An interesting exercise is to simply look for data sets. What is available and what is not, at least easily? Then, several months later, you may want to look again and see what may have changed. The landscape of available data is always changing, and keeping this in mind is very important as a data analyst.

2. What data do you have? What data does someone else have who might be willing to share it? Students can e-mail campus groups to ask questions about their data they are interested in.

3. When you hear about data or people using data to come to conclusions, think about what you might do. Even if you don't have the data available, simply thinking about what you would do will improve your ability to work with the data you do and will have. You'll be honing your data analyst mindset.

Looking for Patterns—and Causes

Lecture 4

It's in our nature to find connections—real or not—and this ability is what lets us take surprising correlations from data analysis and find impressive connections. Beware of just rushing in where angels fear to tread. Some of those connections are real. And because of that, we will continue to see them. Top athletes, investors, and researchers will continue to look for patterns to improve their performance. We all love an interesting pattern, especially if it comes with a plausible story. The difference in good data analysis is that we don't stop there. Finding a pattern is a great start, but it's also just a beginning.

Pareidolia and Seeing Patterns

- We organize information into patterns all the time; our mind has a way of organizing the data that we see. This type of thinking is a part of how we think. In fact, we can also make up patterns, seeing things that are not there—for example, when we look at clouds, or inkblots, or other random shapes. Psychologists call this pareidolia, our ability to turn a vague visual into an image that we find meaningful.

- We do this with what we see, and we also do this in how we think about cause and effect. Professional athletes, for example, often look for patterns of behavior that lead to success. They are under constant pressure to perform at a high level, so if a player finds something that meets success, he or she repeats it. Maybe it helps; maybe it doesn't. But this can be taken to an extreme.

- In basketball, Michael Jordan, who led the Chicago Bulls to six NBA championships, had his rituals. The five-time MVP wore his University of North Carolina (UNC) shorts under his uniform in every game. Jordan led UNC to the NCAA Championships in 1982—which was a really good outcome—so he kept wearing

that lucky pair. Players sometimes see a correlation between their success and some activity, so they repeat it.

- We all look for correlations. When is a pattern real, and when is it merely spurious or imagined? For example, increased ice cream sales correspond to increased shark attacks. Correlation picks up that two things have a certain pattern of happening together: more ice cream sales and more shark attacks. However, there is a well-known aphorism in statistics: Correlation doesn't mean causation. It could be that the connection is simply a random association in your data.

- But if there is a connection, many other things can cause the connection. The two factors may themselves not be particularly connected but, instead, be connected to another factor. For example, maybe weather is warmer in a particular area at the time when sharks tend to migrate in that area. Maybe the warmer weather causes an increase in the presence of sharks *and* an increase in people eating ice cream. Ice cream consumption and shark attacks just happen to be correlated, but one does not cause the other.

- A published medical study reported that women who received hormone replacement therapy were less likely to have coronary heart disease. It turns out that more affluent women had access to the hormones, and that same female population had better health habits and better access to all kinds of health care, which was probably a much better indicator of less heart disease.

- Such research results can have worldwide effects—seemingly positive effects at first, but actually quite harmful ones. News of a big result can spread quickly, but if found wrong, the impact of such news can be difficult to reverse.

- *The Wall Street Journal* reported in 2011 that retractions of scientific studies were surging. This can put patients at risk, and millions of dollars' in private and government funding can go to waste. Some research is retracted due to researchers unethically fabricating

results or for plagiarism, but in other cases, people find connections that do not offer the level of insight touted. And the increasingly powerful tools for data analysis and data visualization now make it easier then ever to "over-present" the results of a study. It is very important to keep in mind that our tendency is to essentially overexplain and overpredict what we find.

- Indeed, scientists are finding this to be hardwired into the human brain. Psychologists have long known that if rats or pigeons knew what the NASDAQ is, they might be better investors than most humans are. In many ways, animals are better predictors than people when random events are involved. People keep looking for higher-order patterns and thinking they see one. Attempts to use our higher intelligence leads people to score lower than rats and pigeons on certain types of tasks.


© Steve Mason/Photodisc/Thinkstock.

The gambler's fallacy is what keeps people pulling slot machine levers over and over again.

- We can also look too hard for patterns in the randomness of financial markets. A few accurate predictions on the market, and an analyst can seem like an expert. But how will the person do over the long run?

- We have a tendency to overlook randomness. If you just flipped a fair coin and got heads 13 times in a row, what do you think you would flip next? Does some part of you think that it is more likely to be tails? This type of thinking is common enough that it has a name: the gambler's fallacy. It's what can keep us at the tables in Las Vegas or pulling the slot machine levers.

- However, there can be a lot at stake in such thinking. Interestingly, people don't always pick the option with the highest probability of success over time. But humans will move into that type of thinking when the outcomes really matter and the stakes are high.

- This can play into financial decisions. If we are determined that there is a pattern where there isn't one, we could be making the wrong decision. In such a way, we could actually make our worst financial decisions on small amounts of money, but before long, that can equate to a larger decision. So, one way to look at that type of thing is to convince yourself that there is no small or casual investment when it comes to finance.

Apophenia and Randomania

- Be careful thinking that there aren't patterns. There are. And if we find them, the consequences can be very powerful. But the fact that there is a correlation doesn't mean that the correlation will predict, or will continue to predict, as well as before. Again, correlation doesn't necessarily mean causation. We simply have a tendency to think in that way. It's like it is hardwired into us.

- It was to our ancestors' advantage to see patterns. If you saw a bush shake and a tiger jump out, then it might behoove you to keep that in mind: Even if for the next 100 times that shaking is the wind and not a tiger, on the 100th time, if a tiger jumps out, that's an important connection to notice! There is a correlation. A tiger could rustle a bush before pouncing, but remember, that doesn't mean that a bush's rustle *is* a tiger.

- There is a name for this. Apophenia is the experience of seeing patterns or connections in random or meaningless data. The name is attributed to Klaus Conrad and has come to represent our tendency to see patterns in random information. But Conrad was actually studying schizophrenia in the late 1950s. He used the term to characterize the onset of delusional thinking in psychosis. In 2008, Michael Shermer coined the word "patternicity," defining it as "the tendency to find meaningful patterns in meaningless noise."

- On the other end of the spectrum is randomania, which is where events with patterned data are attributed to nothing more than chance probability. This happens when we overlook patterns, instead saying, "It was just totally random." But the most common reason we overlook patterned data is that we already have some *other* pattern in mind, whether it is a real connection or not.

- In his book *On the Origin of Stories*, Brian Boyd explains why we tell stories and how our minds are shaped to understand them. He argues that art is a specifically human adaptation. Boyd further connects art and storytelling to the evolutionary understanding of human nature.

- For Boyd, art offers tangible advantages for human survival. Making pictures and telling stories has sharpened our social cognition, encouraged cooperation, and fostered creativity. How can this help us from an evolutionary point of view? Humans depend not just on physical skills but even more on mental power. We dominate that cognitive niche, and as such, skills that enhance it can aid us. Looking for patterns from that point of view aids us, and when we see patterns, we create meaning and may even tell a story.

- Whatever the case, we do have a tendency to look for patterns. And that can be a real problem and a real strength in data analysis. The important part is to realize and recognize that we may unearth a pattern in data analysis. It may even be surprising. But even if we can offer a possible explanation, that still doesn't mean that we have found something meaningful.

- In data analysis, we look for patterns in the data. We as people are good at it, but sometimes, we are good at finding something that isn't there. It's an ever-present balance. As you look for and find correlated data, be careful.
    - First, look in *both* directions, and see if you can think of why one might cause the other. Maybe causality is there, but maybe it goes in the opposite direction from what you expected.

- Second, like warm weather explaining shark attacks and ice cream consumption, check to see whether something else offers a better explanation.

- Lastly, always keep in mind that it might just be your hardwired ability that's leading you to expect something that isn't there.

Suggested Reading

Boyd, *On the Origin of Stories*.

Devlin, *The Math Instinct*.

Activities

1. Consider the following series of heads and tails.

TTHHHHHTTH

THTHHTHTHT

Which is random? The first series is a series of actual flips of a quarter. The second is one that was made up in trying to keep the number of heads and tails equal. Often, people will think that a random set of flips isn't random, because often a number of heads or tails will consecutively fall.

2. Visit Google flu trends at https://www.google.org/flutrends/ and see what it is predicting for your area or for an area of interest. Have you recently conducted a search on the flu? When you do, what do you search on?

3. To see more examples of domino mosaics, visit Robert Bosch's web site at http://www.dominoartwork.com/.

Algorithms—Managing Complexity

Lecture 5

Data sets are getting bigger, so a fundamental aspect of working with data sets today is ensuring that you use methods that can sift through them quickly and efficiently. In this lecture, you will learn about a core issue of computing: complexity. Managing complexity is an important part of computer science and plays an important role in data analytics. You will discover that algorithms are the key to managing complexity. It is algorithms that can make one person or company's intractable problem become another's wave of innovation.

Algorithms and Complexity Theory

- Billions of dollars are spent with credit card numbers flying through the World Wide Web to make online purchases. When you make such a purchase, you want to be on a secure site. What makes it secure? The data is encrypted. And then it's decrypted by the receiver, so clearly it *can* be decrypted. Broadly speaking, encryption techniques are based on factoring really huge numbers—such as 10^{75}.

© Fuse/Thinkstock.

When you make online purchases on a secure site, the data is encrypted, and then it is decrypted by the receiver.

- How do we know that someone isn't going to be able to factor such a number on a computer or some large network of computers? Computers keep increasing in speed. Maybe tomorrow there will be a computer that's fast enough, and suddenly Internet sales are insecure. But a computer simply cannot move at these speeds; having a computer even 1,000 times

faster isn't going to be enough to crack the code. The ideas that ensure the safety of such methods also determine how large a data set someone can look at, or if someone's data analysis technique can be done in real time.

- Computer scientists, long before the massive data sets of today, have studied how fast algorithms work. This is called complexity theory. It can help you compare algorithms and know if one will work as problems grow. For example, it can tell you if what you are doing will work if your company suddenly spikes in users. If you go from 150 users to half a million, will things still be done efficiently?

- Complexity theory can help us know how big of problems we can analyze. If we double the problem, is it going to take just a tad longer, or twice as long—or more?

- Can we write the sum of 1 through 100 in a clean way? There is a famous story that this problem was given to a primary school class in the late 1700s as punishment. The young mathematician Carl Friedrich Gauss was a child in the class. He later became a prominent mathematician with work that is used today in many fields of mathematics.

- Gauss saw a pattern. Take the numbers 1 through 100, and under them write the numbers 100 through 1.

$$\begin{array}{rcrcrcrcr} 100 & + & 99 & + & 98 & + & \dots & + & 1 \\ 1 & + & 2 & + & 3 & + & \dots & + & 100 \\ \hline 101 & + & 101 & + & 101 & + & \dots & + & 101 \end{array}$$

- Next, add each column of numbers. So, you are adding 100 and 1, which is 101. Then, you add 99 and 2, which is 101. That's the key—you will get 101 in every case. And there are 100 of them. So, the sum of 1 through 100 and 100 through 1 is 100 times 101. This is twice what we need, so summing 1 through 100 is 50×101.

- What if we added the integers between 1 and 500? This would equal 250 × 501. The formula in general for adding the integers between 1 and n is

$$\frac{n(n+1)}{2}.$$

Exponential Growth

- Suppose that a highly contagious virus hits. It starts with one person. The next day, that person and someone else are sick. The next day, four people are ill. Suppose that the illness stays in Manhattan, which has about 1.6 million people. Also assume that it would take 10 days to create a vaccine.

- On the first day, one in 1.6 million people is ill. Even after a week, only 64 people are ill. After two weeks, about 8,000 people have been ill, or half a percent of the population. After three weeks, or 21 days, a million people are ill, and the next day, everyone is infected.

- Note that on day 20, you only have 30 percent ill. This means that if you wanted the vaccine available on day 20, when 32 percent of the population is infected, then you needed to start it on day 10, when 512 people, only 0.032 percent of the population, are infected. That's difficult to see. Exponential growth can, to a certain extent, appear to be doing almost nothing and then spike. That's why paying attention to small changes can be quite important.

- Having a bigger, faster computer isn't enough. If you have an inefficient algorithm, it may take thousands of years to solve certain styles of problems. They are simply that difficult. No quick algorithm is known. In fact, some problems can be shown to be computationally intensive. The encryption problem for credit cards is this type of problem.

- Searching is one big thing to do with data. Another is sorting. One way to sort is to put everything in a pile and take the items one by one and insert them into their proper place. It works. However, for large data sets, that can be slow.

- There is another technique that is faster, which usually means it's a bit more complicated. This method uses the idea of dividing and conquering. The idea is to split the array into two lists. One list contains items less than some value, and the other list contains items greater than or equal to some value. Sort both lists and recombine, which is easy. This is called Quicksort, because it is quick.

- Suppose that our list is 5, 3, 7, 4, 6. So, we pick the value that will determine what goes into the two piles. We call that the pivot. Let's pick 5, because it is the first element in our list. Now, 3 goes in the first pile, because it is less than 5, as does 4. So, 5, 7, and 6 go in the other pile.

- So, we sort the first pile into 3 and 4. We sort the other pile into 5, 6, and 7. Note that when you combine, the left pile is sorted, and all elements are less than those in the right pile. Combining is trivial.

- If your list is bigger, then you simply apply the same idea to the two piles you create at the first step. You again divide them into two pieces. This is called recursion. Keep using the idea to produce smaller versions of the same problem until the problem is small enough that it is easy to do. When you get down to 10 items, for example, it is quick and easy to sort. Then, backtrack up the ladder of recursion and combine those lists

- Try this the next time you need to sort through a pile. Computers make this much faster, of course, but it's the algorithm that's the key. The algorithm is what makes it easier to manage complexity, whether it's slips of paper or big data in a computer.

- Keep in mind that sometimes we can work with small data sets and gain great insight. However, we also want to keep in mind what can change when we work on large data sets, especially the even larger-sized data that might follow—or else what we do today may become obsolete tomorrow. And beware of thinking that bigger computers alone make things quicker.

- Imagine if Google's search algorithm didn't scale. The number of web pages has grown at an alarming rate—a genuinely exponential rate. If they could not find or develop scalable algorithms, then Google may have been great with millions of web pages but failed to keep up with billions of web pages—or failed to keep up today, when the number of web pages is estimated in the trillions. By scaling up by 100 or a million in size, the time needed could have suddenly become problematic.

- Imagine if, a year from now, given the growth in the size of the web, it suddenly didn't take a second or less to get a search result from Google. Now, it took an hour, or overnight. Google as we know it wouldn't be Google. Part of Google's success was its ability to adapt its algorithms to the growing size and complexity of the web.

- Today's developers of applications and services must think ahead. There may be 10,000 users now, but if things go well in today's age, there could be several million users. Can the current methods deal with huge, even nonlinear, increases in scale? Sometimes, faster algorithms have a more complex design, with some details that enable increased speed. More complexity in the algorithm can mean simplifying what your computer needs to do to finish the job.

Suggested Reading

Levitin and Levitin, *Algorithmic Puzzles*.

Johnson, *Simply Complexity*.

Activities

1. Search Google news for "exponentially," and look for examples where the growth is most certainly not exponential but instead just fast.

2. You may wish to watch the movie or read the book *Pay It Forward.* The thesis of the book is how quickly an idea can spread. By expecting a doubling in an event, we see an example of the impact of exponential growth.

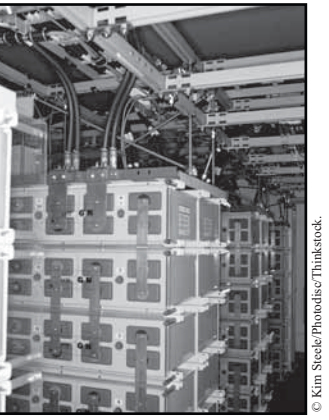
The Cycle of Data Management

Lecture 6

Storing information is something that data analysts have been doing electronically for decades. But today, we are also dealing with huge data sets, and in order to analyze them, we have to have a better understanding of how they can be stored, and then how data can be retrieved. There are a variety of approaches, but some of the questions that need to be addressed are as follows: What data do you need to collect? How much of it do you need? What happens if you run out of storage? These issues are ever-present with today's large and growing data sets.

Data Storage

- In data analysis, data is often stored in digital rather than physical form. Digital storage and reference isn't new. Computers can search large data sets, like the 155.3 million items in the Library of Congress, very rapidly. And we are now accustomed to such speed. These days, a second may be too long to wait for results. Google engineers have found that if search results are just a tiny bit slower, even by the blink of an eye, people search less.



Finding enough storage space for large amounts of data can be a challenge.

- There are multiple search engines. And in the cyberland of searching, it is definitely survival of the fittest. The fastest to retrieve what you queried and return results is, indeed, the fittest. Research showed that we will visit a web site less often if it is slower than a close competitor by more than a quarter of a second (250 milliseconds)—it takes 4/10 of a second (400 milliseconds) just to blink your eye!

- The Library of Congress is archiving all of America's tweets. This isn't all the content of Twitter, but it's still a lot of information, and depending on the moment, it can be a whole lot. The volume of tweets isn't uniform from moment to moment, but Twitter typically gets 500 million tweets per day, for an average of about 5,700 tweets per second.

- When you hear that an organization like the Library of Congress is archiving something like tweets, think about the amount of data they are storing. Also think about how fast the act of storage has to be during a spike. Then, think about the further challenges of making the archive accessible. Users will want it to be stable, fast, and convenient. Then, multiply the hundreds or thousands of terabytes just from Twitter across a lot of other organizations: There are many, many data sets today that are very, very large.

The Cycle of Data Management

- For a data set to be meaningful and useful, it must first be stored and second be accessible. With such amazing volumes of data, how can this be done? In a nutshell, not with yesterday's methods. Some of the storage techniques that worked with data sets that were considered large a decade ago are now dated.

- There can be different approaches used to handle large data sets, especially very large data sets. However, size alone isn't the only determining factor in how to approach managing data. Another issue is whether the data is in motion or at rest.

- Using the example of Twitter, data at rest would consist of analyzing past tweets about a product to get a sense of customer satisfaction. Data in motion could have the same goal, but the difference is immediacy. In this case, the company might keep track of tweets as they appear. When a product is rolled out, what are people saying? This could help, for example, with quickly changing some aspect of the promotion, or customer service, or even the product, if that's possible.

- There is also a cycle to data management. Of course, you must collect data. But what data you collect depends heavily on what problem or question you are considering. The cycle of data management involves the following questions: First, does the information even exist? Is there data to reach any conclusions? And then, if it is there, how do we capture it and know what it is stating?

- Once we have data collected, we must decide how to store it and how to organize it. We may also need to bring together data from multiple sources—this is called data integration. And if your data is in motion, it may be better to have data integration happening over and over. You may even want data integration in real time, which is the opposite of dumping everything just once into a static and fixed database.

- An important aspect of data integration, even if you do not have data coming from disparate sources, is preparing your data for analysis. After the data is integrated, we can finally analyze it. For example, Amazon will look at past customer actions from all the data that's been integrated in its system. Then, they can make an action—for Amazon, that may be recommending a book.

- But the cycle of data management goes back up to capture. Once we act, we need to capture more data and validate the action. For Amazon, if they keep recommending books that aren't of interest, then the feature could even become a nuisance.

Data Warehouses

- When we begin a project in data analysis, we are fundamentally looking at how data will be managed. First, how much data will you have? A second important issue is the security of the data. How secure must it be? Third, how precise should the data be? Can we aggregate information?

- Let's assume that you are, in fact, using quite a bit of data—enough that you can't easily store it on your own computer or even with one external hard drive. At one time, you'd need a supercomputer to aid in this. For most of us, that means that any such questions simply couldn't be considered. But today, that's changed. The change came when companies like Yahoo, Google, and Facebook turned their attention to helping offer storage mechanisms for the data that their services were helping produce.

- An important part of storing and accessing data is data warehousing, which serves as a central repository of data collected and then integrated from one or more disparate sources. Data warehouses often contain both current and historical data, enabling one to look at current patterns and compare them to past behavior. This is an important feature; data warehouses generally are used to guide management decisions.

- A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. As mentioned, a data warehouse often has multiple sources feeding into it. For example, there may be multiple branches of a bank in several countries with millions of customers and various lines of business from savings to loans. Each bank's database may have been developed or tweaked internally, with each application designer making individual decisions as to how an application and its associated database should be built. As mentioned, these sources are combined in a data warehouse. A smaller subset of the data warehouse, aimed at end users, might be called a "data mart."

- The key, though, is that the data sources are also accessible. That's the "relational database" part. A relational database is a type of database that organizes data into tables and links them based on defined relationships. These relationships enable you to retrieve and combine data from one or more tables with a single query.

- Parallel processing is one of the cornerstones behind modern-day large-scale data analysis. In particular, Apache Hadoop is a free programming framework that supports the processing of large data sets over multiple computers. Interestingly, the word "Hadoop" is named after the creator Doug Cutting's child's toy elephant. But the technology was initially sponsored by Google so that they could usefully index all the rich textural and structural information they were collecting.

- However, the key is that they didn't want to just store it—they wanted to present meaningful and actionable results to users. Nothing like that existed, so they created it themselves. This isn't a project of Google alone. Yahoo has played a key role in developing Hadoop for enterprise applications.

- What is characteristic of modern big data storage issues is that new techniques need to be created, so they are. Ironically, we aren't always aware they happened, because sometimes they are created so that we never see their impact. The data world doesn't feel any bigger to the user. We can keep posting updates on Facebook, for example, and Facebook is in a new big data center, ready to capture, analyze, and present it until they inevitably need to move again.

- The issue of data storage is one that comes up more and more in data analysis. If your business is especially dependent on analyzing huge amounts of web data, then Hadoop can be part of the solution. But you don't need any of that to tap the power of relational databases, which are your first step up when a spreadsheet is no longer big enough to hold your data. And if you have only megabytes or even gigabytes of data, you are unlikely to gain either speed or flexibility from putting your data into a Hadoop rack of computers.

Suggested Reading

deRoos, *Hadoop for Dummies*.

Hurwitz, Halper, and Kaufman, *Big Data for Dummies*.

Activities

1. Find a data set related to your interests and create a relational database. Note that we worked with a rather small data set in the lecture. You may wish to try a small and a large set of data and see how it organizes the data.

2. To gain an appreciation for data warehouses and Hadoop, conduct an Internet search and see the array of topics, news articles, and companies on this topic.

Getting Graphic and Seeing the Data

Lecture 7

Graphics are a terrific first step to really getting a grip on your data. But it's only recently that graphical tools have become easy to create. In the past, graphics were used only at the end of an analysis, to summarize what's been done—and sometimes as a tool of persuasion. When done well, graphics tell a story. They tell a limited story, but they can tell it clearly. And seeing the story can offer new insights.

Graphics as Storytellers

- In modern times, we have many tools that make it very easy to create graphics. This has led to a realization that graphics can be used throughout an analysis, not just to present final results. In 1977, John Tukey published a highly influential book entitled *Exploratory Data Analysis*. He recommended graphing your data when you start. He especially recommended graphing to see five things: the two extreme values, the median value (the one in the middle), and the two quartile values (cutoff points where 25 percent of the values are above or below).

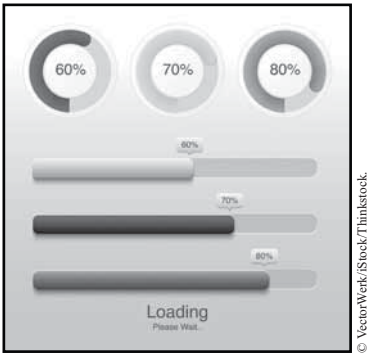
- The general process of data visualization starts with your data. Then, you often need to transform the data—for example, by looking at it in a few different ways. Then, you visualize the data. With that, you can analyze and interpret what you see. Keep in mind that sometimes this leads to wanting to look at the data in a new way. Then, you visualize again and analyze your new results. In this way, you create a cycle that allows you to gain more insight.

- In his 1993 book *Visualizing Data*, William S. Cleveland of Bell Labs declared that "visualization … provides a front line of attack, revealing intricate structure in data that cannot be absorbed in any other way." Data visualization can help explore data, but it only begins the process. A picture is worth a thousand words, but no one

picture conveys all the data equally. Graphics, like good writing, tell a story. And, like any storyteller, graphics tell the tale from their point of view.

- As an example, a pie chart yields less information than some other graphs. We can use a pie chart to highlight the top few values, but it tells us less about the rest of the data. So, an important part of any graphical process is deciding what design to use. It isn't always obvious what you need to see. You may not even know that some insight resides in the data. But graphical displays can help you efficiently and quickly see information.

- Look carefully at today's information graphics, and you'll often see common graphs like bar charts or pie charts. They are artistically done. But it's the same basic graphical techniques. In fact, we've probably all been stuck looking at a very common bar chart used to give the status of a download or update. It doesn't always work accurately, but when it does, it gives some potentially useful information.



Data visualization is an important part of data analysis.

- Why do graphics help convey information? Why can they, when done right, tell a story so quickly? Part of this comes from the way our brains think. Pictures, unlike text, are processed all at once. Approximately 30 percent of our total gray matter in the brain is responsible for visual activity. That's why we can communicate with pictures so effectively. We need only to look at stories from 35,000 years ago as told with drawings on rocks and walls to see that pictures have been working well for a very long time.

- On the other hand, the brain processes text linearly as it moves through letters or words. That takes longer, and most readers skip a lot. A web usability expert named Jakob Nielsen found in a small study that the average person will read only about 20 percent to 28 percent of the words on a web page. So, in this era, visual content that communicates quickly and effectively can tell a story that words cannot—at least if someone never reads those words. This helps explain why many of today's magazines and online resources use graphics and infographics to tell their stories.

- For example, *USA TODAY* has their daily snapshot. In 1982, the paper departed from text-centric, black-and-white-newspaper format and moved to color and graphics to tell part of the story. The British *Sunday Times* and *Time* magazine have used graphics to simplify and enhance the understanding of their stories. *The New York Times* has invested more effort in producing sophisticated graphics, many of which are interactive.

Making a Good Graphic

- A potential pitfall of infographics is that it's possible for a graphic to omit the important conclusion you are trying to find. Data can contain information that can be perceived quickly with a graph, but we must choose the right graph in the right way, or it's possible that the important information might be concealed.

- Ideally, graphics portray data in a quickly consumable and easily understood fashion. This can be, and is often, done. You can see this in Leonardo da Vinci's Vitruvian Man from the late 1400s. This classic image graphically shows how to understand the proportions of the human body as written by the Roman writer Vitruvius 15 centuries earlier.

- For example, the head as measured from the forehead to the chin is one-tenth of the total height. One's outstretched arms are always as wide as the body is tall. Now, this may not be true; Vitruvius might

be wrong. But until Leonardo da Vinci, no one cared much about the data analysis presented by Vitruvius. It's the graphic that tells the story in a compelling way.

- What goes into making a good graphic? Two questions lie at the heart of a graphic. Who is the graphic for? What does it need to communicate? Again, think of a graphic as a story. For example, you'll tell a story of a company differently to stakeholders than to customers—they have different needs. What story are you telling?

- Edward Tufte's 1983 book entitled *The Visual Display of Quantitative Information* is a definitive resource on visually displaying data for many statisticians. In his book, he notes, "Graphics reveal data. Indeed graphics can be more precise and revealing than conventional statistical computations." But we also want to be aware that even a great graphic does not include everything.

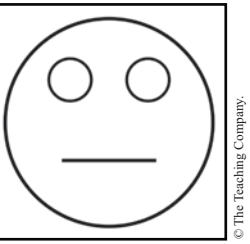
- So, there is another important key in data visualization: What data do you visualize? This can be both an art and a science. It's not enough to know the general area for an archeologist to dig to find ruins. While there may be hundreds of miles to consider, being half a mile off can still lead to empty results. Visualizing the right data can led to discoveries and new insights.

Graphing in Four Dimensions

- Let's step beyond the third dimension and graph data in more than 3-D. First, let's see how we can graph three dimensions in 2-D. We see this all the time on weather maps. Think about how weather maps start with 2-D spatial data and temperature data on top, where the temperatures are represented by color. Even 4-D is possible on a flat map, and that fourth dimension doesn't have to be spatial. Think of those weather maps. The third dimension is color. A fourth dimension could be humidity, or precipitation, or storms.

- Circular histograms, or rose diagrams, are another way to put several dimensions on a single graph, with each wedge representing one dimension. Star plots are a cruder version of the same idea, where the result looks like a star instead of a flower.

- One of the more interesting ideas for graphing more than three dimensions is to plot data using Chernoff faces. Herman Chernoff invented this graphical technique to display data in the shape of a human face. A face has eyes, ears, a mouth, and a nose—that's four dimensions. It's almost like a Mr. Potato Head for data visualization, except the data itself determines what we see. And it's not just the size of the features; it can also be the shape, placement, and orientation of the facial features.



Chernoff faces allow people to display data in the shape of a human face.

- Humans easily recognize faces and notice even small changes without difficulty. Chernoff faces allow us to pack a lot of dimensions on a small graphic that's easy to compare. This type of graphic encompasses a lot of information, and it means more if you already know and care about the underlying data. But even with this type of graphic, as always, the graphic cannot portray everything.

- Another issue is how the graphic itself can push your story in a particular direction. For example, with Chernoff faces, we perceive some features more than others. We notice eye size and eyebrow slant a lot, so they tend to carry more weight.

- You need to be careful: How you map the variables to the graphic can affect what people who view the graphic notice. In general, you should think carefully about which features of a graphic will attract the most attention. Such limitations are inherent in any graphic and simply should be kept in mind when presenting or interpreting graphic results.

Suggested Reading

Smiciklas, *The Power of Infographics*.

Tufte, *The Visual Display of Quantitative Information*.

Activities

1. *The New York Times* is a leader in interactive graphics. Look at their site and find your favorites. You may also want to look at the D3.org site for examples of such graphics and, if you like programming, download templates for your own use and exploration.

2. Create your own infographic. You can create a single infographic of a piece of data or create a poster, which often works best in three sections with some overarching theme. You may also want to look online for tools that help. There are a variety of free tools to aid.

3. Search Google for "infographic" and enjoy the image gallery of options and creative ideas. Be critical in your viewing. Does the graphic tell its story effectively?

Preparing Data Is Training for Success

Lecture 8

All data is not created equal. There can be errors and ambiguities in data, and we must determine what to include or exclude. We need what is called data preparation. In addition, all data does not have—or should not have—the same purpose. That's because we often need to peel off part of the data to help us figure out how predictive or effective our analysis is. We divide our information into training data, which we use first to build our analysis, and test data, which we use afterward to check.

Training Data and Test Data

- The set of data that is the data we develop our idea on is called the training set. We also need to test our idea, and it often doesn't work to use the set we already have. So, we turn to another set of data. This is called the test set. The data we are using is separated in order to be used this way. It is very possible to design a method that does great on the training set but fails on the test set.

- On one level, it sounds simple enough to have the two sets of data. But it's not always that simple. The issue of really testing an idea can be quite subtle, and it can become troublesome when not done correctly.

- In fact, a poor training set can doom data analytics. There are a number of issues to consider. First, the training set must cover the full range of values that the problem might present. Suppose that you are creating a method to predict housing prices. You'd want expensive and inexpensive houses, big and small houses, one- and two-story houses, and houses with and without garages. The more features that exist, the larger your training set should be. There isn't, though, an easy way to know how large the set should be. You do want dozens, if not hundreds or thousands, of examples of each feature.

- An issue with training and test sets is that they often come from the same set of data. You have access to only one set of data. From it, you want to create a training *and* a test set. Ideally, the training set would be representative of the whole set. In the same way, without being identical, the test set would also be representative. However, you probably don't know in advance exactly what representative means—especially when dealing with really large data sets.

- So, how do we do this? Let's assume that we have our data in a table. Each row represents a person's ratings of movies. We can create the training set by selecting random rows—about 80 percent—of the data set. The remaining rows are the test set. We probably don't know if this particular test set works, but the rows of data were created at random, so we can simply do it again. In this way, we can create a few training and test sets and see how we do. This tests the robustness of our method. As long as we keep any given training set entirely distinct from its corresponding test set, we're okay.

- This assumes that your data is ordered randomly, or at least placed randomly into training and test sets. If that's not the case—if the training data differs from test data in some way—then this approach might not work. Clearly, there is subtlety here, and what to do in those cases is still an area of continued research.

- In many ways, this shouldn't be all that surprising. For example, we have been trying to create representative samples in political polling for years. We have, at various times, created successful ways to do this. However, when the technological landscape changes—or the issues change, or the nature of the population changes—we need new ways to specify our training and test data, too. Note that putting data into training and test sets is done prior to creating methods that analyze them. If the goal is to predict some phenomenon, then we process the data in this way prior to creating our method.

Dirty Data

- Data can be what is called dirty. If data is dirty, how can you trust your results? Data can be incomplete, noisy, and inconsistent. This can be due to human error, limitations of measuring devices, or flaws in how the data was collected. It can be that nothing really went wrong but the data isn't in necessarily the assumed form. For example, one person could have two entries in the data that the U.S. Postal Service is using due to changing addresses. It could be important not to double-count the individual.

- However, if we end up with data that isn't descriptive, it can be essentially garbage. If this is the case, how can you expect to make meaningful insights? This leads to the saying "garbage in, garbage out." This is a common expression in computing.

- So, what does garbage look like? How dirty is it? Well, it may not look all that dirty to us. Consider this simple question. Are Bobby Miller from Charlotte, North Carolina, and Bob Miller also from Charlotte, North Carolina, the same person? We might guess that it's the same person. But a computer would treat this as two distinct people. These issues can be very real.

- Conversely, there can be two addresses for one person. Mail can come to a house not addressed to the very small town that it is actually located in, but to a neighboring larger town. As such, the house might be listed in some databases as being located in the neighboring larger town. At some point, it would be easy for such a data set to list the house as being located in two, or even three, different towns.

- A computer, without help from specialized software, could easily see such entries as being two, or more, separate people. This may not be a big deal by itself, but when merging corrupt or erroneous data into multiple databases, the problem may be multiplied by millions. What's the point in having a comprehensive database if that database is filled with errors and disputed information?

- One way to deal with this problem is to buy software to clean up the data. This isn't all that easy, as can be reflected in the price of such software. A general tool powerful enough to address your particular issues can cost between $20,000 and $300,000. What is it fixing? What could be going wrong? One issue is duplicate data, such as the house being listed twice. A variety of issues can occur, and how to deal with them isn't always all that obvious. To ensure that this doesn't happen, data analysts generally must look at the data to ensure that it makes sense.

- To see how easily variation from differing data sources can introduce small, but serious, inconsistencies, consider a few cases. Were postal codes recorded using a uniform format? In the United States, that translates to the following: Do the zip codes all use the same five-digit format? If addresses are encoded with something like bar codes, does the code have enough flexibility to describe every type of address?

- Are dates all the same? If you have data coming from the United States and from Europe, they likely are not. This is because in the United States, we write dates in day-month-year format. However, in Europe, dates are often written in year-month-day format. A date like 10/11/12 can cause a lot of confusion.

- Is data in the same units? If you are looking at currency, is the data global—and if so, are the countries reporting in their native currency or some common currency? Is everyone reporting values with the same accuracy?

- Is some data missing? This is not unusual. The data might not have been collected. Maybe some people declined to report their age or weight. It may be that some data isn't applicable. For example, not everyone has a middle name.

- Regardless of why the data is missing, what will you do when it isn't present? There are several options. First, you can eliminate data objects that have missing elements. But then you are throwing

out data, and if the data is missing from an important subset, you might be removing most or all of that set. You could also delete an attribute that has missing values. For example, maybe you no longer include middle names or ages. Again, this removes data and should be done with caution.

- Sometimes, you can estimate a value; sometimes, you can simply ignore that a value is missing. This depends on the analysis and the type of method being derived.

- Another common attribute is having inconsistent values. Sometimes, this is easy to check. For example, heights can't be negative, and adult ages shouldn't be single digits. Other times, possible mistakes are not easy to check but can still affect the results.

Making Data Clean

- To aid in dealing with these issues, data is often preprocessed, or made clean—or at least clean enough. Data preprocessing comes in two main forms: 1) selecting data objects and attributes for analysis and 2) creating or combining attributes to create new attributes more suitable for analysis. In selecting data objects, you might sample the population or look at only a subset of the available features of your data. Other times, you might decide that you need to add an attribute—some data you hadn't collected before. Other times, you might combine data.

- In his 2012 book entitled *Best Practices in Data Cleaning*, Jason W. Osborne stated that recent surveys of top research journals in the social sciences reveal that many academic authors are not suitably concerned about dirty data.

- The main problem is that most statistical tests may not be robust or reliable with dirty data—at least to the degree that researchers might hope. But even if you've done everything right initially, similar problems can pop up later. Things can quickly and

almost unexpectedly go astray. Many issues can occur that cause inconsistencies, changes, missing values, or other problems in data, creating results that are truly garbage.

- The other problem is that we have to create results that work on good, clean data. But our goal is generally not to simply do well on existing data—that can lead to a problem called overfitting. We want to create a method that can use that data to perform well, but also offer insight into data that is yet to come.

Suggested Reading

Osborne, *Best Practices in Data Cleaning*.

Tan, Steinbach, and Kumar, *Introduction to Data Mining*.

Activities

1. Sometimes, rather than splitting your data, you can test your data on future events. This is also a way to train and test your analysis. Think of questions that have interested you in this series. Would you need to split the data, or could you test your analysis on future events?

2. Keep in mind that just because you test your data, it may not always work. First, random events can happen, and even more behavior can change. So, even a well-designed model may, in time, no longer work. As such, you sometimes will want to test again—even on analysis that is working well.

How New Statistics Transform Sports

Lecture 9

Data plays a sometimes fondly remembered role in sports. Numbers shape how we watch sporting events. Numbers affect how players approach the competition. And numbers are important to how athletic performance is understood and assessed by everyone. In modern times, data analytics is playing a huge—and growing—role in sports. In this lecture, you will gain insight into why it's growing, and you will learn how to do some sports analytics.

The Pythagorean Expectation and Runs Created

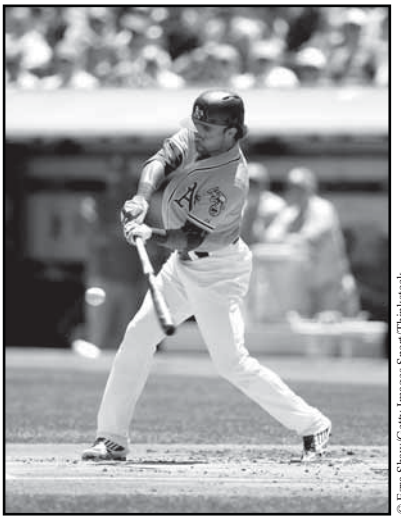
- Keeping track of data has long been part of many sporting events. Sports data can offer insight that is otherwise difficult, if not impossible, to perceive. Baseball box scores, for example, not only record which team is winning and losing as a game progresses, but they also include summary information about the performance of every player in the game. Beyond individual games, there are statistics summarizing the data for entire seasons—and for entire careers.

- Statistics and box scores can unveil various aspects of the game. But they simplify at that very same time. In some cases, a particular statistic isn't very instructive. In other cases, it is a matter of having that statistic plus several other statistics.

- But under every summary statistic is the need for data, and it is the data that we are analyzing. Today, data collection and analysis is an art and science of its own. In modern times, we can look at all of the data—not just statistics that summarize the data.

- Let's consider the math behind the 2011 film *Moneyball* and the 2003 book by the same name. First, let's set the stage for the role of analytics in the story. The drama comes from the tension between two factors: winning and money.

- Billy Beane became the general manager of the Oakland Athletics baseball team in 1997. But he had one of the smallest budgets in Major League Baseball. Other teams were still building their rosters with the conventional wisdom of the time. They used their allotted budgets to sign big-name hitters and rocket-armed pitchers. Even the A's had followed that strategy in the 1980s, winning three consecutive World Series appearances in 1988 through 1990. But in the 1990s, new ownership wanted to spend less on the team, so the cash-strapped A's needed a new strategy.



© Ezra Shaw/Getty Images Sport/Thinkstock.

Billy Beane's method of data analysis led the Oakland A's to victory while cutting costs.

- Beane and his staff decided to try something different. They built their team by being really good at data analysis. The most expensive players were overvalued at the time, and even the wealthiest teams didn't have unlimited budgets to spend. So, Beane's strategy was to buy players that had less value as measured by the traditional techniques, which often involved intuition, foresight, and experience.

- Beane's method worked. In 2002, his team became the first in over 100 years of American League baseball to win 20 consecutive games. They also made the playoffs, reserved for the top eight teams in baseball. This all came while having the smallest player payroll of any Major League Baseball team. So, how did they cobble together such a team for a fraction of the budget available to many other teams?

- In the film, Billy Beane's assistant, the character of Peter Brand, states that the team needs to win at least 99 games to guarantee a playoff spot. What type of team can win 99 games? The statistic we'll learn now was developed by Bill James, a baseball writer and statistician who is connected to many of the techniques detailed in *Moneyball*. What James discovered was a statistic combining the total number of runs a team scores and the total number it allows.

- Here's how it works. A fraction will estimate the percentage of games a team will win. The numerator equals the square of the total number of runs scored by the team that season. So, if a team scored 100 runs, for example, the numerator would be 100 squared. To get the denominator, we just add the numerator (the square of the number of runs scored by the team) to the square of the number of runs allowed by the team. This formula is known as the Pythagorean expectation.

- As an example, let's look at the 2002 Oakland A's. The team scored 800 runs and allowed 654 runs during the regular season.

  So, the Pythagorean expectation equals $\frac{800^2}{800^2 + 654^2}$.

  Put this into a calculator and you'll find that the A's were expected to win 59.94 percent of their games. The team played 162 games in the regular season, and 59.94 percent of 162 is 97.1.

- There are some interesting aspects of data analysis here. First, the Pythagorean expectation didn't quite reach the 99-win threshold. They won 103 games—not 97.1—and they made the playoffs. The number we calculate is a calculated guess; it's not a crystal ball.

- It can be easy, as a data analyst, to get attached to your computations. But remember that they give insight—an opinion. In the case of the A's, they won slightly more games than this Pythagorean expectation calculates. So, the formula wasn't

perfect. Maybe it could be refined. The important point is that just by using the Pythagorean expectation, the A's got much closer to identifying their own winning formula.

- Ok, so now we know that the Pythagorean expectation uncovers a distinctive relationship between wins and runs scored and allowed during a season. This is something a coach and the players can concentrate on. But how? We can use more data on the players themselves to think about how to reach our target of 99 wins.

- In the film, Brand indicates that the team can allow no more than 645 runs (which is 9 fewer than were allowed in the 2002 season). Taking this number as fixed, how many runs must be scored to reach an expected number of wins greater than or equal to 99?

- Let's put these pieces into the Pythagorean expectation formula.

$$\frac{99}{162} = \frac{x^2}{x^2 + 645^2}.$$

- So, we get

$$x^2 + 645^2 = \frac{162x^2}{99}.$$

- Moving the x^2 to the same side, we get

$$\frac{63}{99}x^2 = 645^2.$$

- So,

$$x = \sqrt{\frac{99}{63}}645 \approx 808.55.$$

- This means that the team must score more than 808 runs.

- So, we know how many runs we need to score, and we know the maximum number of runs we can allow. But how do we know whom to sign? For this, we turn to another Bill James statistic called runs created, which quantifies approximately how many runs a player contributes to the team.

- Like the Pythagorean expectation, runs created is a fraction. The numerator equals (hits + walks) times total bases, where a single is worth 1 base, a double is worth 2, and so forth. The denominator equals the number of plate appearances by that player.

- This is a great example of how we can use data analysis to come up with innovative solutions. In particular, note how we combined ideas to create workable insight. This is an important element of data analysis. You create tools, but you use many of them to create something.

- This approach to baseball is now well known and used by many teams. But that doesn't mean that all opportunities are gone. In later seasons, for example, the A's began selecting players more for undervalued skills in defense. And they also began using data on players who had not even begun to play professionally.

Advances in Sports Analytics

- From building a team to recognizing improbable performances, math can give us insight when it comes to sports. What else can we analyze? That's a question that many professional athletes and teams are answering. With modern technology, there are new answers to that question as new discoveries are made. Some are made public; others are kept very private, because they give a competitive advantage. So, look for new statistics being kept in a sport or new graphics appearing in the news to detail a performance in a sport. This often reflects advances in sports analytics.

- A key of sports analytics is that it can help us know when something special is happening. Keeping box scores enables us to keep track of battering averages. It can help us recognize who to recruit for

the A's and make a play for the World Series—without spending the most money. It can also help ensure that a team on a losing streak simply may be on the unfortunate side of randomness. This can improve coaching decisions about a team or player. This can enable a recruiter to recognize greatness on the field. If we turn to predictive analytics, it might even indicate an attribute that correlates well with future play.

- The key is knowing which analytics are useful and insightful. No statistic ever tells the whole story, but a good statistic can tell us which part of the story to look at more carefully. And for coaches and players, better use of data makes it possible to change the story—better data analytics is game-changing.

Suggested Reading

Baumer and Zimbalist, *The Sabermetric Revolution.*

Lewis, *Moneyball*.

Winston, *Mathletics*.

Activities

1. Are you interested in sports analytics? Which sport? Look online for downloadable data sets. For example, the Olympics offers downloadable data at http://odf.olympictech.org/. This isn't always the easiest format. You may want to search for data sites. For example, the *Guardian* has data for the London 2012 games. You can view it at http://www.theguardian.com/sport/series/london-2012-olympics-data.

2. Pay attention to the statistics and predictions made in sports programs. Many use new ideas to predict future events. You are hearing the new waves in sports analytics.

Political Polls—How Weighted Averaging Wins

Lecture 10

Statisticians have been conducting election polls for many decades. But such polls aren't always accurate, especially for close elections. But recent data analytics does much better than ordinary polling. The secret is combining multiple polls. It turns out that clever aggregation of multiple data sources produces much more accurate predictions. And that also transforms political campaigns. But data aggregation is not just for politics. Weighting and aggregating data can work well for any messy, complex field where no single variable explains everything.

Predicting Elections Results

- In the 1920s, political predictions were notoriously unreliable, and people were looking for new ideas on how to forecast the upcoming presidential election between Herbert Hoover and Alfred Smith.

- A popular magazine of the time was *Literary Digest*, which employed a powerful new tool, called a poll, to accurately predict Hoover's landslide victory in 1928. Their poll also accurately predicted Hoover's demise in 1932.

- Then came the next presidential election. The republican nominee in 1936 was Kansas governor Alf Landon, who was running against first-term president Franklin D. Roosevelt. The *Literary Digest* predicted that Landon would receive 57 percent of the popular vote to Roosevelt's 43 percent. They came to this conclusion after sending out 10 million surveys and getting 2.4 million back. FDR, our future four-term president, won with 62 percent of the popular vote.

- Why was the *Digest* so wrong in predicting this election, especially when it was so accurate in previous ones? This election occurred in 1936, which put it deep in the shadow of the Depression. As a result, many people were rapidly eliminating luxuries from their lives.

- The *Literary Digest* got their mailing list for their surveys from telephone directories, automobile registration records, and their list of subscribers. The key is that each one of these activities—whether we're talking about using the telephone, driving a car, or subscribing to magazines—was considered an unnecessary luxury during the Depression.

- This means that the *Digest*'s method analyzed how the *wealthy* population of the United States would vote. But this isn't what they thought they were predicting. They thought they had accurately predicted the election.

- There was a young pollster of the time who recognized the errors in the *Digest*'s data and made his own poll. He surveyed only 50,000 voters, compared to the 10 million voters surveyed by the *Digest*, but he made his set of voters more representative of the voting population. And as a result, this new poll predicted FDR's win with only half a percent of the data used by the *Digest*.

- Maybe the most impressive part is that this new pollster saw the *Digest*'s error and was able to compute what their prediction would be within one percent. The new pollster was named George Gallup, and this is the beginning of the Gallup poll we see used everywhere today. George Gallup's work in this election launched his career and gave the Gallup poll its initial credibility.

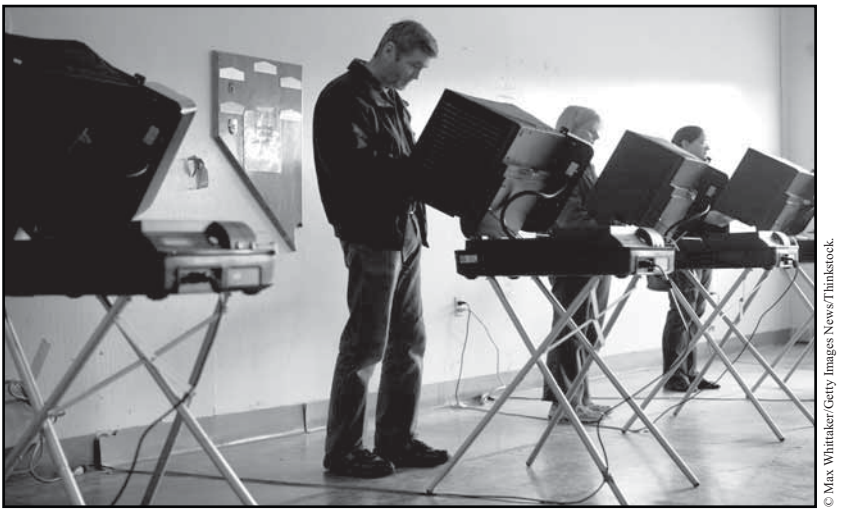
- Almost 100 years later, in the 2008 and 2012 elections, Nate Silver emerged, in a way similar to George Gallup, as a new force in polling. Silver is a statistician and writer who initially made his name analyzing in-game baseball activity in the realm of sabermetrics. Then, he turned that statistical toolbox toward elections.

- In the 2008 presidential election, Silver established his web site, FiveThirtyEight.com. At first, he didn't reveal his identity. Soon after he did, he began to appear in various media outlets as an electoral and political analyst. In 2008, Silver correctly predicted

the winner of 49 of the 50 states in the presidential election. The only state he missed was Indiana, which went for Barack Obama by one percentage point. In that same election, he correctly predicted the winner of all 35 U.S. Senate races. It wasn't difficult to predict some of the results; what was impressive was coming so close to giving correct predictions for *all* of the results.

- On the morning of the 2012 presidential election, Silver posted the final update of his model, giving President Barack Obama a 90.9 percent chance of winning a majority of the 538 electoral votes. By the end of the day, Mitt Romney conceded to Barack Obama. Silver also correctly predicted the winner of all 50 states and the District of Columbia.

- Silver created a poll by combining the results from multiple pollsters. He was not alone in this approach. But it is important to note that individual pollsters were less successful. Rasmussen



Predicting how people will vote in elections is a classic challenge that stumps even prediction specialists.

Reports missed on six of its nine swing state polls. Gallup had among the worst results. In late October, their results consistently showed Mr. Romney ahead by about six percentage points among likely voters. This differed significantly from the average of other surveys.

- How can so many specialists dedicated to the science of predicting elections vary and even struggle? First, just like Roosevelt's second election, a major issue is polling a representative group. Gallup polls typically sample the opinions of 1,000 national adults with a margin of error of plus or minus four percentage points. If Gallup is creating a poll to gauge public opinion about a national issue, how can the opinion of 1,000 people represent the opinion of millions? That's a key issue in the science of polling.

- In polling, just like with the *Literary Digest*, you must poll a representative group. We must have a well-mixed sample of people. Not only should they be mixed, but in a way, the mix of the group should look the same as the large group.

- An issue with this is that you must contact them. In the 2012 election, some polls were done with live interviewers, other with automated telephone interviewers, and even others via the Internet. Of these three modes, automated polls had the largest average error of five points, having a Republican bias for that selection.

- Another issue is whether you call cell phones. There are legal restrictions regarding automated calls to cell phones. As seen in the time of the 1936 *Literary Digest* poll, this can be tricky. How you contact poll respondents affects whom you reach.

- For someone who wants to get started with making predictions, keep it simple: Just start with a single state, assign weights to each poll from that one state, come up with a prediction, and update your prediction as more polls come in.

- In an address at the Joint Statistical Meetings, which is the largest gathering of statisticians held in North America, Silver noted that the average is still the most useful statistical tool. It's a cornerstone of all the methods that are making election predictions so much more reliable.

Running Political Campaigns

- Using data analysis, campaigns can better understand the electorate. In particular, they are able to better identify who they want to communicate with. In the 2012 election, the Obama campaign used data analytics to help know what demographic to reach for votes. Then, they could study what media markets tend to reach that group. They promoted their candidate there rather than buying time on media outlets that reached a large demographic. This approach was much cheaper and, in many ways, more valuable.

- Political campaigns have always tried to identify and mobilize "their" voters. And campaigns have been amassing more sophisticated files on potential voters since the 1990s. Some experts credit this approach to the 1996 Bill Clinton campaign, which focused on winning swing votes rather than the entire electorate. George W. Bush narrowed the focus further by concentrating resources on swing voter Republicans. This makes sense, but to do this right, they needed to figure out who these people were, where they lived, and what they cared about.

- Jump to the 2008 Obama campaign and its well-executed web campaign. They raised about half a billion dollars online. At the same time, they gathered a lot of data—around 13 million e-mail addresses and 5 million friends across social media platforms. E-mail addresses, combined with information from voter registration records, helped the campaign uncover which potential voters they should reach with rides to polling places, or which phone calls to make addressing specific points raised online.

- Once campaigns have a digital profile of voters they'd like to target, they can also develop very customized political advertising. Campaigns can use online advertising techniques to be sure they present their message to their desired audience.

- Online ads also offer quick feedback about how well they are working. Did you click the ad? How long did someone engage in pre-roll video before skipping it and getting to the featured video? In the last weeks of an election, this kind of rapid feedback can be especially valuable, but it also helps a campaign stay more on track throughout an election.

- What makes the approach we've been discussing especially new and powerful is the effort to grab and use *all* of the relevant data about voters. There had been an entirely different approach to political prediction over the years that just ignored polling data. A model from that approach might have tried to predict voting by looking at something else: How is the economy doing? Is there a popular or unpopular war? The idea was that if you had a model explaining why voters will vote one way or the other, then maybe you don't need polling data tracking how they say they will vote.

- Data analytics has transformed national politics, in part, by taking the opposite approach. Instead of hoping for master variables to explain overall elections, get lots of data about voters. See how they vote, and see how they say they'll vote. As long as there is plenty of voter data available to aggregate, this will very easily beat more speculative kinds of prediction.

- This approach works well when there is a lot of data. The less data you have, the less confidence you can have in this kind of analysis. But the basic approach remains valid even for smaller elections. There's just more room for error.

- If you want to know how people will vote, then focus on data about the voters. You may need to focus on other types of voter data, such as who they voted for previously, changing demographics of the district, and political contributions. Especially for small races, it may also become important to have very granular metrics about candidates themselves, especially how they interact with voters.

Suggested Reading

Bradburn, Sudman, and Wansink, *Asking Questions*.

Silver, *The Signal and the Noise*.

Activities

1. Search the Internet for political polls that might interest you. Can you find several? What elements can you derive that could help weight their importance, or do you believe that recency is enough?

2. Are you interested in presidential polls? Either looking back at the past election, or if one is coming, some claim that state polls are enough for aggregated polls, which may not be as complicated as Nate Silver's work but can give insightful results. What data can you find?

When Life Is (Almost) Linear—Regression

Lecture 11

A big part of data analysis is predicting future outcomes by studying the past. And we predict, where possible, by describing a perceived trend with a formula. The formula might be linear, exponential, parabolic, and so on, but whatever the formula, it's rare to exactly match the data. Even so, a formula can be close to the data and capture enough of its essence to give insight on what might happen in the future. This lecture will teach you about the power of predicting data just by using equations of lines. In this sense, we are modeling life as if it were linear.

Least Squares Regression

- In 2012, Usain Bolt electrified the Olympic track and field stadium in London as he won a second consecutive gold medal in the 100-meter dash. This was Bolt's second time to win this title. In 2008, Bolt ran the fastest 100-meter dash in the Olympics. No other Olympic gold medalist could have beaten him—until he ran even faster in 2012.

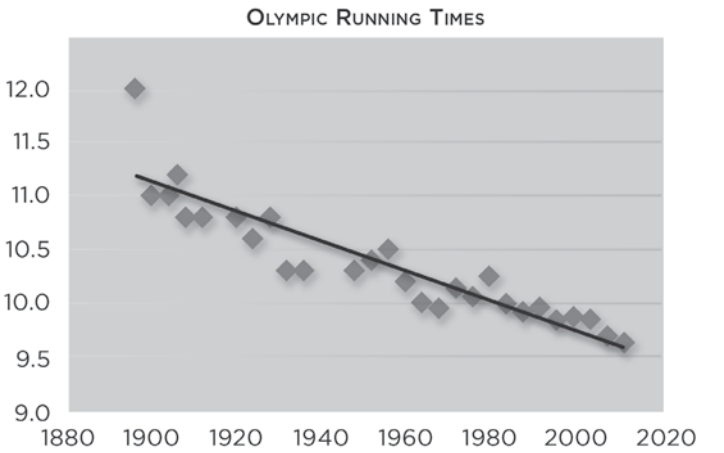
- There have been 28 gold medalists in the 100-meter dash between 1896 and 2012. The slowest time was Tom Burke's 12-second sprint to gold in 1896. The fastest was Usain Bolt in 2012. In fact, in most of the Olympic Games, the gold medal time is faster than the previous Olympic winning time.

- Bolt ran 100 meters in 9.69 in 2008 and 9.63 in 2012. It takes 400 milliseconds, or 0.4 seconds, to blink an eye. So, Bolt beat his own time by less than the blink of an eye in 2012 versus 2008. Carl Lewis's 1984 run took only 9.99 seconds; this was about a blink of an eye slower than Bolt's 2012 race. Jesse Owens ran in 10.3 seconds in 1936, which is almost 2 blinks of an eye slower than Bolt.

- This gives a bit more insight on the numbers, but how far ahead would one runner be than another? To help with this, we'll compute average speed. Bolt ran the 100-meter race in 9.63 seconds, which averages to a speed of 23.23 miles per hour. Carl Lewis averaged 22.39 miles per hour. So, at the end of the race, Carl would have been 3.6 meters, or just under 12 feet, behind Bolt. Jesse Owens would have been 6.5 meters, or about 21 feet, behind.

- Having a sense of the speed that these numbers represent helps us see how a slight variation can change the time, but very small changes are important in races at these speeds. Next, let's get a sense of the numbers as a whole. The numbers are generally decreasing. One quick way that data analysts get a sense of data is to graph or visualize it.

- There is some variability, but the data can be approximated by a line. While the line won't pass through every point, it can get close. If you wanted to predict future times in this race, the insight you'd gain simply from graphing would be huge. You can model the data with a line, and if the trend in the data continues in the future, you can make predictions at how fast people might be running in 2040, for example.



- When trying to fit a line to the data, your tinkering could bring your biases into the computations. So, it is better to use all the data and let the data, without any fiddling on your part, create the line. If the points don't lie on a line, we use the tool of regression, where we "regress" the data back toward whatever model we create.

- Before producing the line, let's think about what line to choose. We want to find a line that approximates the data. We want the line to be as close to all the points as possible, but the line may not pass through any of the points. We'll find how far off the points are by their vertical distance to the line.

- We could just measure the distances from the line, add those up, and be done. But that ignores the fact that some values are going to be close to any line we choose, while others will be farther away. We'd like to choose a line that minimizes those maximum distance points—keeps those far distances as small as possible. So, instead of taking the least distance for each point, we take the least squares distance. This is clever because by squaring, our biggest values are now huge, and minimizing those huge values gets much more attention. So, the least squares method takes the square of all those distances and adds those up. Whatever line makes that sum smallest is our best fit.

- Using a least squares regression line for our Olympic times, we find the line $y = -0.0133x + 36.31$. There are a variety of tools—from Excel, to JMP, to SASS, to R, and many others—that can help you find the equation of the line. Using Excel, you simply have a table of data: one column containing the years and another containing the times. Then, you essentially hit a button and the formula comes out. Regression is a powerful technique that is quickly calculated on a computer.

- The slope of the least squares line is -0.0133 (x is the year, and y is the time). The slope predicts that for every year, we expect the Olympic gold medal time to drop by just over a hundredths of a second. So, over 4 years, we expect it to drop by just over 5 hundredths of a second.

- However, beware of assuming that every regression line continues indefinitely. Clearly, there is some limit at which a human can not run any faster. For example, the 100-meter dash will never be completed in 2 seconds.

Applications of Regression

- So far, we have used only two variables, and we have used one to predict the other. But regression can also be a much more powerful tool. Regression is often used to determine how strong a correlation is. For the Olympic winners, the correlation coefficient (r) was −0.91. This is an extremely high correlation value. Depending on your field, a much lower value may still give insight.

- The square of the correlation coefficient (r^2) is another commonly used statistic. What counts as a "good" r^2 value varies enormously from field to field. But the correlation coefficient is where the r^2 actually comes from. Your regression may also have far more than two variables; even 100 variables or more could be involved. Once you have the data in place, a regression can be done very quickly, even when you have data in more than a few dimensions.

- Regression is used in various fields, with economics being one of the leading areas. One reason for this is because regression enables us to artificially change one variable while holding all the others constant. For the gold medal times, each year reduces the time by about a hundredth of a second.

- This can be quite helpful in business. Suppose that you have data on sales, prices, and promotional activities. Regression can give you insight as to what would happen to sales if prices were to increase by 5 percent. What if promotional activities were increased by 10 percent? This helps marketing.

Logistic Regression

- In their very popular book *Freakonomics*, Stephen J. Dubner and Steven D. Levitt look to regression as their tool to bust some of the myths about parenting, for example. They use data to see what factors correlate to test scores. They're calculating correlation coefficients and seeing which factors have the highest r^2 value.

- What is correlated? Keep in mind that correlation can be positive or negative. Test scores and the fact that the child has highly educated parents are positively correlated. Test scores are also positively correlated with the fact that the child's parents have high socioeconomic status. But these aren't all that surprising.

- We shouldn't always be seeing what we don't perceive or expect in our world, but we can also get new insight that isn't as expected. This can be the case for some of the factors Dubner and Levitt found that don't correlate—not positive correlation, but not negative correlation either. They found that moving to a better neighborhood doesn't mean better test scores, but the possible disruption from moving doesn't hurt test scores. So, there's no correlation.

- In addition, the fact that the child's family is intact doesn't help test scores. On the other hand, the child's family not being intact doesn't hurt test scores, so there's no correlation. We often have variables like intact versus not intact, and they are just as easy to use. In fact, there is another form of regression we can use in such cases—it's called logistic regression.

- With the Olympic Games, we had two variables: x was the year of the Olympic Games, and y was the gold medal winning time from that year in the men's 100-meter dash. But we can also have more inputs. Instead of just one input variable x, the studies would have used 10, 20, 100, or even hundreds of variables. But all the x variables, however many, combine to produce just one y.

- In logistic regression, we still produce a y and have various input variables. But with logistic regression, the x values take on only 0s and 1s. They are "on" and "off": "intact family" or "not," or "male" or "female," for example. There are just two values.

Suggested Reading

Chartier, *Math Bytes*.

Levitt and Dubner, *Freakonomics*.

Activities

1. Download a data set of interest, plot it, and see if the data or data cloud, depending on the size, follows some curve. See how you do at predicting future events or past events that you exclude from the data.

2. Pay close attention to the way we visually identify things. For example, we identify handwriting by looking at it and what it looks like. This is a key that a computer might be able to do the same thing, as we learned in this lecture. Many cameras put boxes around faces or identify when people are smiling. This uses similar ideas. What other types of visual identification do you see?

Training Computers to Think like Humans

Lecture 12

Usually, data analysts look at data, analyze it, and learn from it. But in this lecture, you will learn how computers can be programmed to look at data and learn by themselves. The computer—all on its own—looks at the data and figures out how to predict what is happening. It's programmed to learn, like our brains. That can sound like science fiction, but artificial intelligence is not just a technique used in novels and movies. It is used to explore and use large data sets that we may not otherwise understand.

Artificial Intelligence

- In the game 20 questions, you think of anything—at all—and your partner gets to ask 20 questions about whatever you are thinking about that you must answer truthfully. If your partner guesses your object in fewer than 20 questions, he or she wins. If he or she doesn't, then you win.

- This could seem impossible; there are trillions of things to choose from. The key is to ask questions that essentially divide the options you currently have in half with each guess.

- This game has been played for over 100 years. It grew in popularity during the late 1940s, when it became the format for a weekly radio program. But it was transformed by data analytics when the computerized game 20Q was invented by Robin Burgener in 1988. Today, you can play 20Q online or with the electronic handheld version.

- Initially, 20Q was an Internet game, and it became a sensation. The link was e-mailed from person to person. With no promotion, the link to 20Q was sent around the world. In fact, this viral success played an important role in why 20Q did so well at guessing.

- How can a computer possibly guess what you're thinking of? How did anyone figure out how to get a computer to make those decisions? The computer, in a sense, figured it out. How can a computer possibly figure something out? The quick answer is artificial intelligence. How does that work? In this case, lots and lots of practice.

- It's possible that inventor Robin Burgener could have created a database containing attributes of common objects. Instead, he taught it only one object: a cat. In fact, according to 20Q's published history, the system knew only one question. But he also built into the program the ability to learn. Play a game, and it learned from the game. So, he began playing, and to help the program learn quicker, he put it on a floppy disk and shared it with friends. As each of them played, the game learned. By playing again and again and again, the program learned more objects, more questions, and what series of questions tended to correlate with an object.

- Then, Burgener shared his program with even more people by putting it on the Internet, which helped it get better even faster. And as it got better, even more people wanted to try it. That viral success meant that it was playing many, many times—and learning a lot. In the end, 20Q had built a database of 15,000 objects.

- According to the 20Q web site, the game guesses correctly within 20 questions 76 percent of the time and 98 percent of the time if you let it ask 25 questions. It can eventually guess correctly, even if one or more of your answers does not agree with answers from most other people. Even with a vast database, it continues to learn.

- It's also interesting that the game could be converted into a handheld device that is not connected to the Internet. There would be a difference. Rather than that database of 15,000 objects, it would be brought down to 2,000. The good news is that people think of those 2,000 objects 98 percent of the time.

Neural Networks

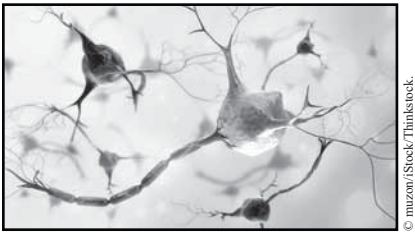
- The 20Q game uses an artificial neural network, which is a computational model inspired by the brain. Just like 20Q, these so-called neural networks are capable of learning and pattern recognition. They've been used not only for language, but also for computer vision and speech recognition.

- Amazingly, the work originated in 1943, even before the digital computer. Warren McCulloch, a neurophysiologist at the University of Illinois, and Walter Pitts, a logician, postulated a simple model to explain how biological neurons work. They were working to understand the human brain. In the end, it provided a foundation for computerized learning.

- When the digital computer became available in the 1950s, the ideas of McCulloch and Pitts were implemented as what were called perceptrons. They could balance a broom standing upright on a moving cart by moving the cart back and forth. The computer learned to do this in the same way that a human would—with lots and lots of practice, by noting what did and didn't work as it learned.

- How do we get a computer to learn? A neural network, natural or artificial, creates complex behavior from simple units; get enough of them acting together, and you can create that behavior. A neuron is a single cell that communicates with others.

- A typical neuron in the human brain receives between 1,000 and 10,000 inputs from other neurons. These signals are relayed to the cell body, where they combine. If the stimulation is high enough, the cell "fires," sending an electric signal to its downstream



Neurons are cells that send and receive information between the brain and other parts of the body.

neighbors. The input layer of biological neurons receive their inputs from the environment. The axons of neurons relay their signals to other neurons, and others are connected to other cells. This is how we learn.

- For a computer, the first layer of the neural network is the input layer. In the human body, this interacts with the environment. The second layer is known as the hidden layer. It contains the artificial neurons. They receive multiple inputs from the input layer. Sometimes there is more than one hidden layer. Then, the artificial neurons summarize their inputs and pass the results to the output layer.

- So, you give the network training inputs. For 20Q, this is many, many people playing the game. You want it to generalize the results. The computer is learning, so it can be difficult, if a neural network doesn't work, to fiddle with it. Neural networks are good for prediction and estimation when the following are true.
  - The inputs are well understood. (You have a pretty good idea of what is important but not how to combine them.)
  - The output is well understood. (You know what you are trying to model.)
  - Experience is available. (You have plenty of examples to train the data.)

- A black box model is okay. You don't need to interpret or explain the model. Why did it get what it got? You may not know, but it will predict from what it learned.

- After the online version of 20Q had played one million games, the neural network had built up 10 million synaptic connections. How do they fire and communicate when asking questions? That's what you won't know. But, clearly, if you've played the game, it does pretty well.

Applications of Neural Networks

- Neural networks and artificial intelligence underscore how computers can learn from data. Indeed, they can learn to guess what we are thinking, to navigate a room, or to predict various phenomena. Neural networks have also been used since the early 1990s to predict what stocks might rise or fall. A key to effective neural networks is training the data—or having it learn on a rich set of data.

- Since the early 1990s, neural networks have been used extensively for a wide variety of applications in stocks, from selecting or diversifying a portfolio to rating the risk of fixed-income investments. Early uses included using neural networks to time when to buy and sell stocks. Banks developed neural networks to predict interest rates; companies with international operations developed neural networks to predict exchange rates. These early predictions worked well to attract interest from other organizations. As these innovations spread, the expression "neural network" somewhat fell out of use. For many users, machine learning was becoming just another routine aspect of what computers do.

Neural networks are used to predict various elements of the stock market.

- Clearly, it is wonderful when neural networks work. Here also lies an important issue in machine learning: We can't always "open the hood" and see the wiring—meaning that you may not be able to know exactly what the computer learned, why, and how. If you need to know that, a variety of machine learning techniques may not suffice.

- However, if you simply need a predictive method, and you have the data and the output you want to predict, then using these techniques can be great. All you need to get started with neural networks is software that performs neural networks, such as Excel, JMP, SAS, or R. Then, you take your data and think about your inputs and the output, and then you will have to think about the layers in a neural network and the number of neurons. Depending on the software, you'll break the data into three groups: training, validation, and test data sets. The validation enables the software to know when it can stop learning and consider itself done. It is then ready to be tested on the data it doesn't see.

- If neural networks don't work well, there are still a variety of ways to improve them. For example, ensemble methods use multiple models to obtain better predictive performance than could be obtained from any single one. You may not know exactly how to build your machine learning program, but you can use several versions to build a better one. It is important to recognize that machine learning algorithms can take tuning. And, sometimes, you simply have to try another approach—which is true in many parts of data analysis.

- Finally, remember that these techniques are focused on a defined task. Our minds learn all sorts of things. The 20Q algorithm trained from many, many people playing the game and now is able to work through various possibilities. This helps underscore that data alone won't allow a computer to learn. The machine must be taught effectively. But with the ever-present and growing amount of data, there is a lot of interest in learning from the data and automating how we can learn and predict from our past.

Suggested Reading

Hsu, *Behind Deep Blue*.

Warwick, *Artificial Intelligence*.

Activities

1. Visit the 20Q web page. How does the algorithm do at guessing your word? Given your knowledge of how the algorithm works, can you fool it? See how someone else does, and then describe how 20Q learned to guess, and see if that person can pick a word that falls outside its range of experience.

2. A fun exercise in artificial intelligence is to look at the world and think of tasks that you think could be automated if a computer could learn them. This is how many innovations in the field occur. Pay particular attention to tasks in which you look at numbers or lists of numbers for specific outputs or patterns—even if you aren't entirely sure what pattern you should be seeing.

Anomalies and Breaking Trends

Lecture 13

In this lecture, you will learn about a process called anomaly detection, which involves trying to spot differences in data. The goal is to find objects or behavior that are different from most other objects or behavior described by the data. Anomalies, by definition, aren't easy to find. They are the surprises, the exceptions, the countertrends, and the peculiarities. As time goes on, more data unfolds, leaving a longer trail for some types of anomalies. And research continues to create a more robust toolbox of techniques to find these special cases. The good news is that we find such anomalies every day, improving our lives and saving money.

Anomalies

- You might have run into this before: You make a big purchase with your credit card, and suddenly you get a call from your credit card company checking in. Among the many, many purchases happening, how did they see yours, and why did they call? Criminal behavior can be detected with anomaly detection. And so can health risks—in fact, the techniques for identifying risks to your health are quite similar.

- The goal, in either case, is to find objects or behavior that are different from most other objects or behavior described by the data—but different only in a relevant way. In data analysis, we often won't know in advance if and when the differences occur. So, we don't know if anomalies are there, but if so, we want to find them, or at least flag that something might be happening.

- At first, this may not seem that difficult to do. Essentially, you are looking for an outlier, which might be easy to spot on a graph or in a list. But it can be much more difficult to spot all of the anomalies.

- We need to be careful in thinking that anomalies rarely occur. In percentage terms, that will be true. But let's say that an anomaly happens only once in 1,000 times. It will still be pretty easy to spot, if you only have 1,000 events. But if what you're tracking happens billions of times, then an anomaly that's one time in 1,000 may end up happening millions of times. We have to remember this when working with large data sets. Statistically, an anomaly might be unlikely, but it still might occur a million times.

- Anomalies, by definition, are not common. In the natural world, many events and objects we notice and care about are common. Yet anomalies are of considerable interest, too—maybe even more so.

- One thing that is very helpful is to see a variety of forms anomalies come in. There are several areas where anomalies are important. The first is fraud detection. You may make a big purchase that garners a call from the credit card agency—or maybe you've made a purchase far from home. They call you because that unusual purchase is uncharacteristic of you, and they want to ensure that the purchase was, in fact, made by you. Noticing a change in behavior, in this case in spending, can aid in detecting fraud more quickly.

- Next is intrusion detection on the Internet. In 2011, a security company named Imperva monitored 10 million attacks targeting 30 different enterprise and government web applications. On average, there were 27 attacks per hour, or roughly one attack every two minutes. The attacks appeared to mostly be probing for vulnerabilities on various sites. If a vulnerability was exposed, Imperva found that the automated attacks could grow to upward of 25,000 per hour—or seven attacks per second. From overwhelming or crashing a system to intruding and secretly gathering information, attacks on computers happen or are being attempted every second of every day.

- The third area where anomalies are important is with ecosystem disturbances. There can be atypical events that have significant impacts on humans, including earthquakes, hurricanes, floods, droughts, and heat waves. In these cases, the goal is to predict such anomalies.

- Fourth is public health. These techniques can help with outbreaks. The Carnegie Mellon system ran an algorithm to determine how likely some anomaly happened by chance. Suppose that normally 8 percent of cases with patients over 50 involve respiratory problems but that today this number is 15 percent. This system could figure out that the probability that this happened by chance is 20 percent—so, not as unlikely as you might have expected. Therefore, maybe this incidence of higher activity can be taken less seriously.

- Finally, let's take an example from individual medicine. When you go to a doctor, you may have tests done. Generally, you don't want unusual test results. Keep in mind, though, that what constitutes an anomaly may depend in part on the age and sex of the person. Correctly identifying an anomaly relates to costs, too. Unneeded tests can cost money. If a condition isn't noticed, it can be harmful.

- If you are working with data, finding anomalies can be helpful. But if you are looking for meaningful averages and statistics, such things can be an issue. So, sometimes, anomaly detection is part of data preprocessing.

Outliers

- The causes of anomalies aren't all the same, and this is important to keep in mind. First, data may come from a different class or source. If someone has stolen a credit card, he or she is of a different class than credit card users. In mathematical terms, a purchase made with a stolen credit card is an extreme value. It's what in regression we call an outlier. Statistician Douglas Hawkins's definition of an outlier is as follows: an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism.

- Second, there is natural variation in data. A bell curve is a normal distribution, and a large amount of data can follow this type of curve. Height is often an example. A height of 6 feet 11 inches is not very common among men, so this height is at an extreme value of the normal distribution for heights.

- Another source of anomalies is errors in data collection or measurement. If the data you collect seems flawed, spend some time looking at your data to ensure that it is correct. You may be more likely to detect a problem if you have a reasonable idea about what distribution to expect in your data—and what would be highly unlikely.

- Sometimes, models are difficult to build. For example, if you don't have data in advance—if you don't know in advance what things are going to look like—then other techniques may be needed. Other methods look at other factors, such as how close an object is to the others or the density of objects in a region. For example, is a credit card purchase wildly different from the value or frequency or location of typical purchases? Do you have one huge purchase? Do you have a flurry of smaller purchases? Sometimes, these types of anomalies can be seen on a graph.

- One way to detect such an outlier is to use the method of clustering. A cluster is a dense glob of dots. A cluster of one element, called a singleton, would be quickly identified. If we wanted to measure the distance of points, we could measure the distance from points to the distance of the center of the cluster, called the centroid. This would help us see that the one main is much farther than anything in the dense group of points. Clustering finds a group of points in two, three, or more dimensions. You don't need to graph it.

- There are some very specialized statistical tests to calculate what counts as an outlier, at least for specific types of data. Clustering is an example from the area of data mining. We might use machine learning techniques, from the area of artificial intelligence. Other times, you might use statistical techniques, which calculate, among

many things, the probability of something occurring. There are also other techniques from areas such as information theory and spectral theory. A lot of methods depend on assumptions about your data that may not be valid. So, you may need to try several methods, especially if you don't know what to expect.

Advances in Data Analytics

- Advances in data analytics have a direct impact on how and when we can find fraud. There are data sources that were previously ignored because they change too quickly. Traditional techniques simply couldn't handle them. The insurance industry is an example. They can refresh fraud scoring in real time, but then that entire data set changes. Aberrant behavior can be analyzed in employees, too, with huge log files from claims or bill processing systems.

- At one time, analysis of such data would take hours or days to run. Now, billions of rows of data can be analyzed in seconds. This speed has profound impacts on data analysis. Applications that once demanded a sample, or subset, of the data used can now run on the entire data set. There isn't a need to find a representative sample. You simply run it on the entire data set to learn and explore it.

- Furthermore, given the speed of modern algorithms and technologies, models can be retuned and tested quickly. If a model seems to be failing and reducing in its ability to detect today's fraud, then a refined model can be tested quickly and analyzed. If it is effective, it can be deployed quickly. At one time, refined models might have been deployed only once or twice a year.

- The larger lesson is that whatever you might consider today that's beyond your computing resources should be logged for tomorrow. Sometimes, in a year or two, something that was impossible becomes possible. It may be better to postpone ideas rather than write them off entirely. They may, in time, be viable—if you don't lose track of them. Returning to an old hunch with fresh tools can be exciting and challenging.

Suggested Reading

Bari, Chaouchi, and Jung, *Predictive Analytics For Dummies*.

Gladwell, *Outliers*.

Activities

1. If something is 99 percent likely not to happen, it still happens about three times per year on average. Sometimes, we expect unlikely things to happen much less than they will. We expect less of a pattern, in a way, than will happen.

2. When something seems odd, it may just be unexpected to you, or it may be an anomaly. If you sense an experience or observation that is an outlier, how did you recognize that it is? Are you sure it is? If you learn to narrow down how you know, you are thinking like a data analyst and honing in on how to do this with data.

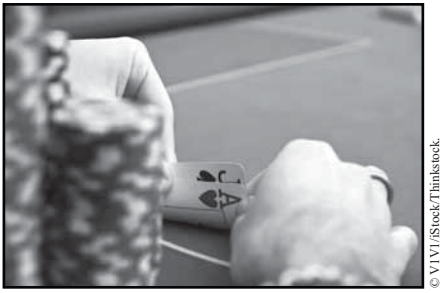
Simulation—Beyond Data, Beyond Equations

Lecture 14

Sometimes, we have too many possibilities to consider, and sometimes, phenomena are simply too difficult to capture in an equation. Simulation is a powerful tool in our world. In many cases, rather than analyzing lots of data, you can produce a simulation and analyze what that says about a physical phenomenon. This tool of data analytics allows us to make better medicines and faster cars and to explore new realms of scientific study—all with the speed and safety of a computer.

Monte Carlo Simulation

- The World Series of Poker in Las Vegas is a tournament that begins with thousands of competitors and narrows down to the final two sitting at a table with hundreds of thousands of viewers tuned in to see the outcome. The winner is considered the World Champion of Poker and receives a multimillion-dollar cash prize. When the game is broadcast on ESPN, a card is dealt, and quickly the probability of each player winning given the current hand is updated.



© V1V1/iStock/Thinkstock.

Simulation is a great tool that can help you play the game of poker.

- How is this done? Could there be a big database of all possible combinations? In that case, a card is dealt, and someone looks up the probabilities for that given state of the game. Consider how big such a database would need to be—and consider how fast we want our answer. We possibly could do it that way, but there is another simpler way: use a computer to simulate the game.

- To see how, we turn to another card game and travel to Los Alamos, New Mexico, in the 1940s. Stanislaw Ulam, while working on the Manhattan Project that developed the nuclear bomb during World War II, pondered the probabilities of winning a card game of solitaire. Because the computations of the probabilities are inherently complex, Ulam explored another route. On an early mainframe computer that he programmed to simulate solitaire, he would play the game a large number of times and computed the proportion of times that he won.

- Such an approach became known as Monte Carlo simulation, because the methods often depend on an element of chance, such as what cards will be dealt. Today, an ordinary spreadsheet can generate and insert random numbers for you.

- Such methods can be used to simulate more important real-world phenomena, too. At Los Alamos National Laboratory, Ulam and John von Neumann also used the methods to simulate nuclear reactions. Today, Monte Carlo simulation is used to study applications in such areas as physics, mechanics, and economics.

- Let's return to poker, and specifically the game Texas Hold'em, to see how simulation could save us from developing a database of trillions and trillions of probabilities. The rules of the game are as follows. Two cards are dealt face down to each player. Then, five community cards are revealed, face up. Each player takes his or her best five-card poker hand from his or her two down cards and the five community cards, and the player with the best hand wins. During the process of dealing, there are several rounds of betting, and much of the strategy in Texas Hold'em comes from betting.

- Here's where a simulation can help you play the game. You won't know, unlike the TV broadcasters in the World Series of Poker, what hand everyone holds. We want to find the odds of winning from a given two-card starting hand, assuming that no players fold. These probabilities can be computed quickly on a modern computer with simulation. Today, a spreadsheet can do this kind of

simulation. Like Ulam, we put the current state of a game into the computer. Then, we let the computer play thousands or millions of random games and count the fraction of wins, losses, and draws for each player.

- Monte Carlo simulations must be run many, many times. We need a lot of numbers—a lot of data—to find what we're looking for. From the law of large numbers in mathematics, as we run more and more tests, we will tend toward the expected value. The issue, which isn't a huge one for computers, is that we need to run hundreds of thousands of experiments. Then, we begin to see the values that we want and expect to see.

Simulations in Our World

- Simulation can help us understand our world. It can help answer questions in probability that can be difficult to answer. This is what Ulam was doing when he simulated solitaire.

- Another example is the Monty Hall problem. It's based on the game show hosted by Monty Hall, in which you are told that there is a 100-dollar bill behind one of three doors and that there is nothing behind the other two. You choose one of the doors. Then, you are told one of the other doors that does not contain the money. At that point, you may change your guess to the remaining door—the one that you did not choose the first time and that you were not told did not contain the 100 dollars.

- Is it a better strategy to stick with your first choice or switch? This question appeared in the "Ask Marilyn" column of *Parade* magazine in 1990. It caught wide attention. The problem was stated as having goats and a car behind the doors. In her column, Marilyn vos Savant asserted that switching is the best strategy. She got thousands of letters, and 92 percent of them insisted that she was wrong. She settled the argument with a simulation. She called upon "math classes all across the country" to simulate the probabilities using pennies and paper cups. She was right, and of course, the simulation backed it up.

- What else can simulation do? Have you been in a fast-food drive-through and noticed that they time how long it takes to fill your order? Such information can be quite important and helpful. This branch of simulation is called queuing theory. If we know the rate at which customers arrive and the length of time it takes to fill the order, we can simulate queuing up, or lining up, under different scenarios.

- When should you have the cashier take orders only and leave the filling of orders to someone else? Simulation can help you determine the impact of such choices. You can see what happens on average. You can also see the extreme cases, or outliers, and determine their frequency and if they are acceptable.

- Similar concepts allow one to model emergency room intake to reduce waiting times. In traffic studies, you can model the difference between a roundabout and an intersection with a stoplight. Simulation is a great tool when you might change parameters in the problem. Sometimes, you need an analytical solution computed directly from the data, but often, a computed number will do just as well. If so, simulation can save a lot of time—and allow you to quickly test many more ideas.

A simulation can help planners decide between a roundabout and a stoplight for a particular intersection.

- While it's only a model, a simulation can, if carefully constructed, have enough realistic behavior that it will uncover enough characteristic behavior to offer insight. With that, decisions can be made. Simulation in general requires some computer programming—but not too much.

Simulation in Hollywood

- Simulation can model phenomena in our world. Blockbuster movies often contain stunning special effects—particularly computer-generated images (CGI). Such images often rely heavily on simulation.

- In the 1980 *Star Wars* film *The Empire Strikes Back*, Yoda was a puppet controlled by Muppeteer Frank Oz, the one behind Fozzie the Bear, Miss Piggy, and Grover. In *Episode II*, *Attack of the Clones*, from 2002, Yoda was created using CGI. Frank Oz was still the voice, but he no longer controlled the movement as he did when Yoda was a puppet.

- In order to operate Yoda in a computer as opposed to the hand of a puppeteer, animators create a digital wire frame of the character. Such a model can contains over 50,000 vertices connected by lines. That number of vertices is needed to capture the detail of Yoda.

- To move and animate Yoda, animators sometimes simply decide on specific places for Yoda's arm, for example, to be in space and time. Then, it is the computers' job to figure out where that limb will be in intervening frames.

- Animating Yoda's hair is even more complicated. Unlike the movement of his body, the movement of his hair may not be specified, except in the first frame of the scene. Generally, it is up to the computer to determine how his hair would move given the movement of his body. Often, the computer is also figuring out the movement of his body.

- Simulation is used to determine how his hair will move. A model is built. In particular, animators model hair as springs. You can determine how springy hair is in the model, too. Think of a bed, where some springs are bouncier than others. Then, you let the computer determine this given the force acting on the hair. This allows animators to put digital doubles into scenes. By simulating

the movement of hair, it may not be exact and perfect, but it is close enough that the audience buys into it.

Simulation in Science

- Of course, Hollywood is different from science. In science, a simulation is used to predict or explain behavior. In the movies, a simulation needs only to produce images that give the appearance of reality. But simulations in entertainment and science are becoming closer, too. CGI simulations for scientific purposes can make it easier to visualize large data sets in motion, further blurring the line between scientific simulation and entertainment-quality CGI.

- Simulation not only visualizes imaginary worlds for Hollywood, but it can also help us understand our universe scientifically. The Bolshoi simulation is a massive, incredibly detailed model of the universe's 14-billion-year history. The images it is producing are amazing and being closely studied.

- The simulations create frame after frame of video. They simulate the evolution of the universe. They do this by first examining the data from NASA's WMAP explorer, which maps out the cosmic microwave background radiation. That radiation is the light that was left over from the big bang. That data can be used as the starting conditions of the universe, and then the supercomputer can simulate how the universe evolved.

- While there are reaches of the universe we have yet to explore, there are many regions that we do know. So, the supercomputer's results are compared to parts of the universe that we do know. And they match up really, really well.

Suggested Reading

Gladwell, *The Tipping Point*.

Neuwirth and Arganbright, *The Active Modeler*.

Shapiro, Campbell, and Wright, *The Book of Odds*.

Activities

1. You can think about how to simulate many aspects of life. If you can simulate it, what questions might you explore with it? How heavy must traffic be for a roundabout to be less effective than a four-way stop? Which board games could be simulated, and could you compare strategies that people play? This is the first step in building a simulation and knowing why are you creating it.

2. Many video games are a form of simulation. Video games must have a quick response to your decisions. Can you determine what type of underlying model the program is using? Sometimes, possibly even often, you may not know. But create your own models and play the game as a data analyst.

3. When watching movies, pay attention to the presence of special effects. What looks real, and where is the special effect less than real? It can be difficult to concentrate on this. You might have to watch the movie more than once.

Overfitting—Too Good to Be Truly Useful

Lecture 15

Sometimes data analysis is too good to be true—or, it is too good to be truly useful. Data analysis builds models that take data to predict future outcomes or explain past events. The goal is to extend a model into something we've not yet observed and make predictions. To do this, you usually have to be less predictive in the past. If we can find that sweet spot between predicting the past and predicting the future, then data analysis is at its best. It can improve our forecast and give us more time to respond to the forecast.

Overfitting

- In 2014, Warren Buffett offered a billion dollars to anyone who could perfectly predict winners of just over 60 games in the NCAA's March Madness tournament. What if someone offered to sell you a method that is a system of linear equations, one equation for each game teams have already played? It uses all the data available from the last 10 years, and it can create a perfect set of predictions for all of those years. How much would you be willing to pay for this method?

- Hopefully, this promise sounds too good to be true. Here's why the method described won't work: If this method includes the tournament that you hope to predict in the data, then you already have information that you're trying to predict. If the method knows that its job is to predict the tournament and that that data is included, it is actually possible for some methods to simply look at those games and predict the winner that way, because it already knows the outcome.

- In other words, if you know the outcome of a game you need to predict, then you simply won't pay attention to the outcome of any other game. You need to separate training data from data you use to test and predict.

- This is an example of overfitting data. In this example, it is pretty obvious that something went wrong; it is too good to be true. You haven't kept your training data separate from the data you use to predict. The model will not generalize at all to new data, because the model is overly fitted to past times that will never occur again.

- Overincluding the past isn't the only problem. Overfitting can also happen if we include a variable that really has no insight for the analysis at all. With so much attention and interest, especially in presidential elections, there is a lot of curiosity about ways to explain the election through something totally unrelated. For example, did you ever notice that if the Redskins won their last home game before the election, the incumbent party would hold the White House? This has been true in 16 of the past 18 elections. 2012 was an exception. The Redskins lost to the Carolina Panthers on November 4th. But Barack Obama won the election for his second term.

- As we have seen before, it is easy for us to see correlation and think causation. This variable, while correlated with past data, doesn't mean it will have predictive value in the future. However, if you allow this to enter your data analysis, it could end up looking like a highly predictive variable. We often find patterns where there may not really be any, and if we throw such patterns into our model without thinking, the model may mistakenly tell us that it's actually helping.

- Too many variables can lead to great results on predicting past data but poor performance for future data. Again, we are overfitting the past. Said another way, we are trying too hard to predict the past when our real goal is to predict the future. In many cases, it is much better to have fewer variables than more.

Ockham's Razor

- The idea of striving for fewer variables and less theory in our model connects to the principle called Ockham's razor, which is attributed to the 14th-century logician and Franciscan friar William of Ockham. The principle states that entities should not be multiplied unnecessarily. Many scientists have adopted or reinvented Ockham's razor. A more current way of saying this is that if you have two competing models, each making the same predictions, go with the simpler model.

- Ockham's razor is especially useful as a heuristic in the development of theoretical models. When it arbitrates between published models, that may be a sign that the theory being rejected wasn't ready for publication after all.

Underfitting

- The opposite problem of overfitting is called underfitting. As the name suggests, underfitting is where things become too simple—so simple that they don't adequately describe the phenomenon. And here lies the difficulty: One must make a model complex enough that it can predict both the data you have and future data you have yet to see, but it must not become so complex that is performs really well on current data to the degree that it does not perform well on future data. This inherent tension is ever-present.

- This can sound hopeless. But we do have predictive models. And they can save lives. For example, modern data and computing have dramatically improved modern hurricane prediction. Underfitting in hurricane predictions is no longer the problem it once was.

- Hurricane forecasting overcomes the problem of underfitting with lots and lots of data. Satellites collect information about the hurricane's position, wind movement, and the atmosphere's temperature and moisture levels. This information is improving all the time.

Scientists learn how and why hurricanes form and strengthen by using hurricane forecast systems.

- This data is used to calculate temperature, pressure, and humidity changes usually in 30-second intervals at points on a grid of about 100 trillion points. That's a lot of computing. And that's what makes it possible to predict a path two or more days down the line.

- But there is still room for improvement: Underfitting is still a large issue for forecasts about intensity, which are not significantly better than a few decades ago. In fact, you may notice that fields where there is still plenty of room for improvement will leave more room for error by referring to "forecasts" rather than predictions.

Overfitting and Underfitting

- In a sense, the worst thing about underfitting is how it leads to overfitting. When we don't have enough data, we overfill the gaps. We may have some shiny gems of data in the mix, but that dazzles us into thinking we know more than we do. Overall, it's hard to be fooled by underfitting. Your training data and your test data give poor results. Your model doesn't work, and you immediately know it doesn't work.

- But one must be more careful to avoid overfitting, which can occur from more than one direction. Sometimes an entire variable or model can cause overfitting. Another culprit is not appreciating the place of noise, or error, in real data; you must assume the presence of measurement error.

- What if we wanted 100 percent accuracy for the data that we measured? One bad possibility is just to change the data. For example, Gregor Mendel, the famous 19th-century heredity researcher, published data about heredity in peas that was much later found to be too perfect. Decades later, a statistical analysis of Mendel's data by R. A. Fisher found there to be too little noise in Mendel's result. We always want to avoid overfitting data, even if it's for a cherished model.

- Another way to strive for 100 percent accuracy is to overfit the model itself: We could leave the data alone and make too many adjustments to the model. But, remember, data is generally not exact and can contain spurious components. It may help to think of static in a phone call. When someone calls and has static in the connection, you try to hear the person talking. So, you work to listen to the voice within the static of the call.

- In his book *The Signal and the Noise*, Nate Silver lays out a case for the Fukushima nuclear disaster as a dangerous outcome of overfitting. Earthquakes continue to be unpredictable. Even with today's supercomputers, modern geophysicists are only marginally better at predicting than simple historical means were.

- With hurricanes, we can collect data—lots of it. With earthquakes, relevant data is much harder to collect. We cannot directly measure stresses 20 kilometers or more underground. So, the data is scarce, and underfitting is an issue. In addition, the current data we have is approximate—that is, it is also noisy. So, we're at risk for overfitting what we do have.

- If we look over long geological timescales, it turns out that earthquakes are far from random. You can even fit the data well with an underlying function. Then, you can pick any point on the Earth and compute a future earthquake probability.

- So, it depends on what you are studying. A hurricane may be predicted within hundreds of miles based on short-term factors that have become easy to measure. Earthquakes, by contrast, depend on data occurring not only deep in the Earth, but also spread over decades, if not centuries. The kind of data needed is different, and adjustments needed for lack of data are also different.

- Noise isn't the only issue in overfitting. It is possible not to have enough data. And it is possible to have too little noise. In such a case, you actually may not have enough noise to accurately predict what is happening. If you have only two points, you will fit the data exactly with a line. Or, if you have only one point, then you can't draw any line at all. But if you can get 10 to 20 points, then you are not overly weighting just one or two initial measurements. And if you can get a few hundred points, or a thousand, then you are not overly weighting a couple dozen points.

Suggested Reading

Berry and Linoff, *Data Mining Techniques*.

Silver, *The Signal and the Noise*.

Activities

1. Remember that overfitting can be caused by including too many variables. When you think of phenomena of interest that you'd like to predict, think not only about the variables that you think will be predictive, but also what you don't expect to be predictive. What do you think you could drop? When building models, it is often good to start as simple as possible and see what meaningful information comes out, and then build up from the basic levels.

2. If you hear a result that seems too good to be true, check if it is describing past events or happened only once. Methods can be overfit or can simply get lucky due to randomness. Remember, if you have 70 people, someone is likely to flip five heads in a row. They aren't a better heads flipper; you just have enough people to make that probable.

Bracketology—The Math of March Madness

Lecture 16

Every March, the United States goes entirely mad for March Madness, an NCAA Division I basketball tournament. In 2012, it was estimated that 86 percent of employees devote part of a workday in March to the tournament. And this was before Warren Buffett offered a billion-dollar prize to whomever could correctly predict the entire tournament. Once you know how to create brackets driven by data and appropriate weighting, you can apply this method to any subject—and come up with a winner that may surprise you.

March MATHness

- For March Madness, our tool of choice is bracketology, where we use math to fill in a diagram of brackets. For NCAA basketball, things begin on a Sunday in March with the announcement of which teams are in the tournament, along with the first-round matchups. There are men's and women's tournaments.

- Then, it's your turn. Who do you predict will win the first-round games? From there, you choose who will win in your predicted second-round matchups. You keep selecting winners for each round until you have selected a national champion. This creates your bracket for the tournament.

- If you are in a pool, how do you know who wins? It really depends on how you keep score. Some pools award each correct prediction the same number of points; others give more points for correct predictions in later rounds of the tournament. This is what the sports network ESPN does.

- There are more than 9×10^{18} possible brackets, or more than 9 quintillion brackets. So, we can't simply pick a random bracket and hope for the best—the chances are against us. If we wanted to win just by guessing, our chances would be better in a lottery, with odds of just a billion to one.

- How can we do better than mere chance when we predict our winners? The math we'll learn will rate every team in the tournament from the best to the worst. We'll follow the math and predict that the team with the better rating always wins. In fact, we will use methods previously used to rank teams in NCAA college football, which historically has had holiday bowl games instead of a full tournament.

- Let's use some real data from the 2012–2013 season. For now, let's work with winning percentage. Simply take a team's total number of wins and divide by the total number of games. If a team won 2 and lost 3 games, then that team has a rating of 2 divided by 5, which equals 0.4. The higher the rating, the better the team is predicted to perform.

- We'll use the result of every game between Division I men's basketball teams leading up to March Madness. Even though we will only rank the 64 teams in the first round, we will see how they

played against everyone to find our ranking. This amounts to 5,000 games. Because we look at every game, we actually rank every Division I team, which is about 350 teams.

- This data is gathered and made available on web sites, such as masseyratings.com. This is game-by-game data. Massey's site gives you the date of the game, who played, if either team played at home (because sometimes games are at a neutral location), and the score of the game. To rank the teams, you have to convert the raw data into a linear system, which you probably won't do by hand.

- Winning percentage doesn't take into account the strength or quality of the two teams playing, so the results you get using this method are not much better than flipping a coin. It's not just whether you win or lose, but whom you play against. So, we can design mathematical methods that integrate strength of schedule by adding weights to rank each team.

The Massey Method

- It will help to learn a method used by Professor Kenneth Massey of Carson Newman College to rank NCAA college football teams. Early in the 2012 season, team A beats B by 14 points, and team B beats C by 3 points. Can we now predict that team A will beat team C by 14 + 3, or 17 points? Of course not. If we could, scores would be what is called transitive.

- So, if transitivity doesn't always hold, we can only assume that it holds approximately. If a team wins one game by 61 and another by 59, then we can approximate with 60. In data analytics, assumptions are often made. You assume a property, and as long as it is close to holding, insightful information emerges.

- Let's see how this assumption leads to a linear system that produces rankings of teams. Let's look at 3 teams: x, y, and z. We need a rating for each. Assume that team x beats team y by 10 points. The Massey method, as it has come to be known, writes this as $x - y = 10$. This means that x wins, y loses, and 10 is how much x

won by. And we can add rating 5 for each team. So, if team x has a rating of 13, then $x = 13$. If team y has a rating of 3, then $y = 3$. So, team x would be predicted to beat team y by $13 - 3 = 10$ points, which we can write as $x - y = 13 - 3 = 10$.

- We don't know the ratings of the teams. That's what we are finding. But we do know the difference in points in the games. So, if we have a game where team x beats team y by 10, then we form the system $x - y = 10$. Assume, additionally, that team y beats team z by 5 points, which would give us $y - z = 5$. Finally, suppose that $z - x = 1$. Team z won that game.

- We can quickly see that transitivity didn't hold. Team x beat team y by 10, team y beat team z by 5, but team x didn't even beat team z, but instead lost by 1. So, transitivity doesn't hold. That means that we won't be able to find values for x, y, and z that exactly solve this linear system.

- In fact, we are usually going to have quite a few more games than the number of teams. Again, for March Madness, we have around 5,000 games and about 350 teams. This means that we have more equations than variables.

- For our type of problem, we can approximate the equations using what is called the least squares method. Instead of averages, we square each of our numbers, add them up, and take the square root. It's like a giant Pythagorean problem, but computers quickly and easily solve such problems. So, it is a matter of entering numbers into a computer and pressing enter.

- This is going to be a system—in the case of March Madness, with 5,000 equations, one for each game. We can, though, form a smaller system with only 350 rows, one for each team. That's still a lot, but it's quicker, and it's also how Massey has done it. Many programming languages have a function that performs a linear solve directly, but even in a spreadsheet like Excel, this can by done.

- We are using the version of the Massey method that considers scores, but you don't have to include scores. In fact, when NCAA football used Massey's method, they didn't do it with scores, due to their concern that weights based on scores could reward blowouts. A big win in one game against a weak team could increase a team's overall rating. This method could also be adapted to look only at wins or to dampen the reward for large wins.

Using Mathematical Software

- For March Madness, the linear system we create will have about 350 equations and 350 unknowns. A computer can solve this system very quickly. We might use a command like LinearSolve on the web site Wolfram|Alpha or "linalg.solve" in Python, which is a free programming language.

- You can use mathematical software called Matlab, or you can use Excel, but you might not want to use Excel if you're working with a rather large system. Sage is a free mathematical software that it is often described as a combination of Matlab and Mathematica. You can also code it in Java, but Java doesn't solve linear systems as easily as other languages. An advantage with Java is that it allows you to post codes on web pages.

- So, we can take all the data from a complete season of Division I men's (or women's) basketball season and create the ratings. Once we have the ratings, it's easy: The higher the rating, the better we considered the team. Then, we create a bracket by choosing the higher-rated team in every matchup.

- If we submit this score-weighted bracket to the ESPN Tournament Challenge, we find that it beats over 73 percent of the over 8 million brackets submitted. That's a very, very stark increase over our winning percentage bracket, which only beat 1.8 percent of the brackets and was competing with coin flipping.

- We've added weights based on scores, but we haven't yet decided how much to weight each game. The key is determining the importance of a game. There are many, many ways to weight the games in a season. You can weight games higher if the team won the previous game. This weights a team's ability to win consistently—to sustain a winning streak. You can weight games that are won on the road, because essentially every game in March Madness is on the road.

- Another option is breaking a season into four parts. Then, we decide how to weight each part. We could count the games in the first quarter of the season as half a win and loss for the respective teams. In the second quarter of the season, the games could count as 0.75 a game. How do we weight games in the third and fourth pieces of the season? Is the last part of the season leading into the tournament the most predictive of a team's success? If so, maybe you could weight it as 1 game or even 2.

- How do we use this when we form the linear system? It's pretty straightforward. You simply count the number of weighted games. So, once we determine our weighting, we can form a revised linear system. Once that's formed, a computer can produce the rating, and we are ready to form a bracket.

Suggested Reading

Langville and Meyer, *Who's #1?*

Oliver, *Basketball on Paper*.

Activities

1. Several professors have used sports-ranking methods to rank teams only by their play against each other to predict end-of-season conference play. This makes for much smaller systems. How predictive can you be

without using a system of 350 unknowns? You may want to create a pool with friends who don't use math to see how you compare. Remember, it can depend on the variability, or "madness," of the tournament.

2. You might want to create three brackets and see how they compare. The first bracket is created before running any of the math methods. The second is created using *only* the math methods and letting it choose the winners. The final method takes the math and overrides some of its decisions. Which does better? Try this over several years and see if any particular one tends to do better. Does one do better in early rounds than another? What about later rounds of the tournament?

Quantifying Quality on the World Wide Web

Lecture 17

Search engines like Google play a huge role in our world. The importance of search engines also makes them big business. In this lecture, you will learn the data analytics of search engines. Google has stayed relevant, from the time when the Internet had millions of web pages and into the era of trillions of web pages. The reason that the Google algorithm has continued to work for more and more users is because the algorithm itself adjusted to downgrade attempts to hijack search results, while also finding evermore ways to deliver meaningful results.

Google's Algorithm

- There are hundreds of billions of web pages. But someone can still manipulate a system of that size by knowing how Google analyzes that system to form its search results. The PageRank algorithm quantifies the quality of a web page. The quality of a web page helps determine which results are listed earlier in the results from a search engine. If two web pages are equally relevant to a query that you input, then the page with higher quality is listed first.

- The ability to quantify quality of web pages is part of what allowed a once-struggling Internet company, Google, to overtake their competitors and become the most visited web site in the world. The founders of Google were Sergey Brin and Larry Page, graduate students in computer science at the time they started creating what became Google. When Google's algorithm was unveiled to the world, there was Google, with its new algorithm, and then everyone else.

- The Internet before Google was a rather tangled web. Millions of web pages existed, and the other search engines weren't as helpful. You could go to a search engine, input your query, and, just like today, get a list of web pages. But, at that time, the web pages at the top of the list often weren't very useful. It wasn't uncommon for a

much better page to appear farther down the list. This almost forced you to look at several pages of search results. The way Google analyzed the World Wide Web changed everything.

- What did Google do that was different? They looked at the connectedness of the web. This was part of their billion-dollar idea. They didn't just look at the content of web pages; they also looked at the structure of the web.

- A fundamental idea in Google's search engine is the concept of endorsement. Higher-quality web pages tend to be linked by other high-quality web pages. How did Brin and Page determine this? Their model can be seen as recommendations, but also as a model of surfing the Internet. It's a model, so it won't exactly replicate how people surf the web. In fact, it's entirely possible that no one person surfs by the rules of the model. But the model captures enough of the characteristic behavior to return meaningful and reflective results.

- Because all people are different, the model assumes that surfing is random. So, a random surfer moves through the web following very simple rules. You can think of it almost like a game. At each web page, the surfer rolls a die to decide which web page to go to next. Which link should the surfer follow? Two people might have different preferences on which link to click, so Google assumes that you pick a random link on a web page.

- But how does this produce a measure of the quality of a web page? Brin and Page calculate the probability of the random surfer being at any given web page if this model is followed. That probability is the associated measure of quality. You can try surfing the network you see on the screen to determine which is the highest-quality web page under this model.

- Once this model is set up and there is a network of web pages, you simply randomly surf the network following these rules. But you must do it long enough that the results settle down. You need to

surf for a very, very long time. How does Google know that eventually the numbers will settle? A nice feature of Brin and Page's algorithm is that it will converge for any web network—anything ever created.

Google's search engine uses the idea of endorsement to analyze the World Wide Web.

- A web page with no links on it is called a dangling node. PDF files, movie files, and music files are often dangling nodes. If you end up at a web page with no links on it, what do you do? You either input a new web address or hit the back button. Currently, our model assumes that you're stuck. It's almost as if you'd give up and simply close the browser, because there is nowhere to go. But we don't do that.

- Google, of course, assumes that you go somewhere else. Where? Again, that is assumed to be a random choice. You go anywhere on the Internet with equal probability. This relates to the final aspect of the algorithm. We don't always follow links on web pages; sometimes, we simply decide to go somewhere else. Brin and Page called this teleporting. You teleport either when you are at a web page, possibly with outlinks, or you teleport because you are at a web page with no links from it.

- How do we know when to do what? You can think of it as a game. At each web page, you roll a die. If it comes up with 1 through 5, the surfer will click a link on the current web page and follow that link to a new web page—each link is equally likely to be chosen. If the die comes up 6, the surfer will go to any web page on the Internet; again, each web page is equally likely to be chosen. Finally, if you end up at a dangling node, you teleport to any web page with all choices equally likely.

- That's the model that Brin and Page created that started Google. There is only one difference: Their model assumed that you were 85 percent likely to follow a link on a web page. By rolling a die, we made that 86.6 percent likely.

Google's Beginnings

- Do you surf like this? Probably not. You're probably not equally likely to go to any web page. It's estimated that there are more than a trillion web pages. The subset that you would visit or want to visit is very, very small relative to this size. You probably don't even follow links on a web page 85 percent of the time. Then, how did Brin and Page know that this model would lead to good search results?

- This was probably the question that the existing search engines asked. In the 1990s, blue-chip venture capital firms, Yahoo!, Alta Vista, and many other major companies were approached by Stanford with Brin and Page's algorithm. They turned down the chance to buy Google's search system for 1 million dollars. This search system became the foundation of the Google company, and by the summer of the 2005, each of the founders had a net worth of more than 10 billion dollars.

- David Filo, a Stanford alum and founder of Yahoo!, encouraged the pair to start their own company and return when the product was fully developed. So, in 1998, the two left their doctoral programs and moved their computers into the garage of a friend.

- Before long, Google became not just a business, but a verb—with people saying, "I'll google that." By integrating the connectivity of web pages, Brin and Page folded in a very important piece of information that set their search engine apart.

Perron's Theorem

- How exactly does Google know that the algorithm will stabilize, or converge, given enough time? No matter how long we surf, could the number one site change if we double the number of steps we've surfed? A math theorem called Perron's theorem guarantees that Google will always find a unique answer for *any* configuration of the web. It turns out that teleportation enables this.

- If there is some nonzero probability of surfing from any web page to any other web page, then we are guaranteed to find an answer—*and* guaranteed for that answer to be unique. We must have those nonzero probabilities for Perron's theorem to hold. We do, we have our unique answer—guaranteed regardless of any configuration of the web. That's a huge result, and teleportation guarantees that this happens. The probability may be quite small, but it is nonzero, which is all we need. This tiny aspect of Google's model guarantees that the algorithm will always work.

Internet Politics

- Brin and Page developed a powerful model that created the company Google. But with such prominence comes much attention. In particular, companies want to be at the top of the web searches. PageRank can help raise a web page's spot in a list of search results.

- If you can get web pages with high PageRank to link to yours, then your PageRank rises. So, some companies go into business offering to help raise your PageRank. One way is by developing pages with high PageRank. Then, for a fee, they link to yours.

- Link farms are created by what are called spammers to essentially fool search engines like Google and raise the rank of a web site. Generally, a link farm has several interconnected web sites about a popular topic and with significant PageRanks. The interconnected nodes then link to a client's page.

- When Google figures out that a site sells links, then their PageRank gets hit. For example, in 2007, Google decided that some web pages were selling links and lowered their rankings. It isn't illegal to sell links, but if you get caught, Google can penalize you. This can have a major impact on your business.

- PageRank isn't the thing Google uses to create search results, so presumably one can exploit other aspects of their search, too. A key is knowing what is being done. In November and December 2003, if you put "miserable failure" into Google, the official White House biography of the President was returned as the highest-ranked query. Somehow, this was happening, even though the words "miserable failure" were nowhere on that web page. This is what became known as a Google bomb.

- The architect was George Johnston, who had starting building this one a month earlier. Putting together a Google bomb was relatively easy; it involved several web pages linking to the White House biography of the President with an agreed-upon anchor text, which is the hyperlinked text that you click to go to the linked web page.

- In the case of the "miserable failure" project, of the over 800 links around the web that pointed to the President's biography, only 32 were part of the Google bomb that used the phrase "miserable failure" in the anchor text. But from November through December of 2003, if you put the words "miserable failure" into Google, the top search result was the White House biography for President George W. Bush.

- What part of the Google algorithm did this exploit? Google wasn't just looking at the text on a web site to determine its content. It treated the hyperlinked anchor text that links to a site as a summary. For example, if you link to The Great Courses, you might link with the words The Great Courses or The Teaching Company. These words summarize the content. By using such words, one can connect the web page to words that possibly don't appear on the site.

- Google has filed patents and taken other steps aimed at reducing the impact of Google bombs. The history of changes to the Google search engine offers more general lessons for data analytics. The core approach was very good, but Google stayed on top by continuing to make the algorithm work better and better.

Suggested Reading

Langville and Meyer, *Google's PageRank and Beyond*.

Vise, *The Google Story*.

Activities

1. When you conduct searches on Google, it can be fun to think about whether you agree with PageRank's measure of quality. Would you rank one page over another? When you don't agree with PageRank, is that just your taste, or do you think it would be broadly preferable? It can be an interesting exercise to think about how to capture such a preference in modeling.

2. It is interesting to try different search engines and see the difference in the order of the web pages returned. It's also an interesting mental challenge to think about what algorithm or modeling decisions might account for these changes. You may not, and probably won't, know, but you can improve your ability to think like a data analyst even without the answer.

Watching Words—Sentiment and Text Analysis

Lecture 18

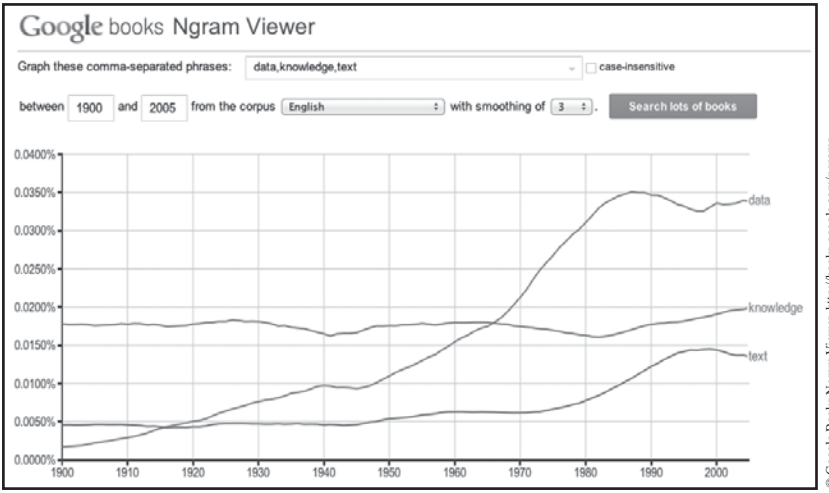
We are approaching a time when every book ever written is available in a digital form that is both readable and searchable. We are now digitally creating, storing, and analyzing many, many more words than appear in traditional publications. All of these words offer rich data sets to understand and predict phenomena in our world. The challenge is that all of this data is unstructured, so in this lecture, you will learn about how to watch the words and gain insight from unstructured data. These are the realms of text analysis and sentiment analysis.

Text Analysis

- We all have had some experience with text analysis. At the simple end, think of spelling and grammar checks. These already contain a level of text analysis. They tell you how many words you have, and many programs will also autocorrect.

- Authorship is a traditional area for text analytics. For example, you could decipher who wrote the works of Shakespeare without knowing that it was, indeed, Shakespeare. You can also look at the frequency distributions of individual letters, which are important in cryptography, where the distribution of letters can make some ciphers too easy to break.

- As for word frequencies, at the end of 2010, Google released an extraordinary tool that looks at frequency distributions across millions of published books—all at once. You put in a word or phrase, and it looks for the frequency across all of Google's scanned books. The tool is called Google Books Ngram Viewer, and it's an astonishingly powerful tool. You can drill down to search the underlying publications year by year. You can refine your search in all sorts of ways, to include what you want and exclude what you don't.



Google Books Ngram Viewer analyzes data across millions of published books.

- What about social media? On January 8, 2012, Denver Broncos quarterback Tim Tebow threw the longest overtime touchdown pass in NFL history, ending the game (and the Steelers' season) in incredibly dramatic fashion. The fervor in the Denver stadium created an Internet tsunami. *USA TODAY* reported that thumbs were typing at a rate of about 9,400 tweets per second on that night as football fans tweeted about the victory.

- We can look at the volume of posts, which can be huge. But what about the words themselves? Researchers at IBM and elsewhere have programmed computers to automatically generate journalistic summaries of soccer matches using only Twitter updates. Spikes in the volume of tweets on a topic help identify key moments in the match. No reporter was sent to the soccer match. In a sense, the people tweeting were the only reporters, and data analysis synthesized the information.

- One can write a program to grab the data straight from Twitter using what is called an application programming interface (API), which defines ways you can retrieve information—in this case, from Twitter. You define a function that calls for username, number of re-tweets, or whatever information you want and have appropriate access to. Some information may require a user password, but often, you can do a lot even without that.

- There are web sites that offer to download data for you. For example, DataSift.com offers the ability to enter search terms and download the data live.

Sentiment Analysis

- Social media can do more than summarize an event. In 2010, a Facebook page created in Egypt called "We Are All Khaled Said" gathered 250,000 followers within three months, and the swelling sentiment that began online culminated in street rallies and a change of regime for the entire country seven-and-a-half months later.

- But there are some issues with sentiment analysis. First, accuracy is never perfect. How well a computer can assess sentiment is generally judged by how well it agrees with human judgments. But humans don't always agree with each other. This is sometimes called inter-rater reliability, and you can never expect 100 percent accuracy.

- Second, data sets and documents vary in their amount of sentiment. A lot of content lacks sentiment. One web page may have a lot of comments, but it might have a less-charged discussion than another web page with fewer comments. So, the quantity and character of sentiment may be independent of one another. You may also get a different picture using blogs compared to Facebook, and both may be different from Twitter.

- Third, words don't always mean the same thing. The word "bad" famously can mean "good," as in "Man, that was a bad song!" This is important for companies, because they watch social networks for sentiment on their products. There may be different lingo, or different inflections, for those who market video games than for those who market men's suits. Furthermore, some words are simply difficult to track.

- The fourth issue for sentiment analysis is getting a baseline. What counts as negative? Some things simply get a higher level of negative responses all the time. For example, a president with an 80 percent approval rating would be doing very well. But what if a company is seeing 20 percent negative sentiment about its product—is that bad? It's not so bad if all of its competitors are creating 30 percent negative sentiment. On the other hand, if 20 percent negative sentiment translates into 20 percent returns of the product, that could be very costly.

Search Engines

- Let's turn to how a search engine determines which web pages are more and less related. This involves a vector space model and linear algebra. A vector space model relies on two things: a list of documents and a dictionary. The list of documents is a list of the documents on which you can search. The dictionary is a database of key words, which could be some or all of the words in the documents. (Sometimes, it's not computationally possible to include all of the words.) Words not in the dictionary return empty searches.

- We then translate all the words into a table of numbers. Each row in the table is a key word in the dictionary. Each column is a document. The entry in a particular row and column is a 1 if that key word is in the corresponding document. Otherwise, the element in the table is 0. This is often called the document matrix.

- Even with this matrix, there are computational issues. Should every document be searched, in its entirety, for each key word? Because of the sheer size of the Internet, some search engines read and perform analysis on only a portion of a document's text. They read only so many words and index from those.

- But this isn't the only issue. Do you require exact matching? Will you allow synonyms? If you don't, then you are requiring searches to have the exact words that appear in the document. Then, there is the issue of word order. For example, do we treat queries of "boat show" and "show boat" as the same or as different? All of these choices impact which documents are deemed most relevant.

- What can we do with an ability to find similarity between texts? Imagine being a legal firm tasked with looking at a large data set. How would you begin? Clustering analysis can't resolve everything, but it can find groups and speed up the process of grouping information that could unlock important aspects of the case, all just by clustering the words in e-mail messages.

Applications of Text and Sentiment Analysis

- Sentiment analysis is a big deal in data analysis. It allows companies to hear their customers in unprecedented ways. They can hear you when you aren't even talking to them. You aren't calling Customer Service, but Customer Service is still listening. If you post on Twitter, they hear it and might even respond. Getting this type of information can be worth big money.

- Today, we can find sentiment available on many topics. When you are thinking about buying a new product, you can look around the web for online reviews posted by people. You are, in your own way, sifting through the data and looking for sentiment. A computer that is programmed can zip through many, many web pages, and if trained in the right way, the idea and outcome is the same. This is a rich resource for businesses.

- Text analysis can also be used in science and applied fields of all kinds. Consider that more than 8,000 scientific papers are published every week on Google Scholar, and top-down classification systems increasingly do not capture every way of looking at published results that could bring insight.

- Also, ironically, research published in academic journals is among the most difficult for ordinary users, or even researchers, to access in bulk. To address this problem, in 2004, the University of Manchester began hosting Britain's National Centre for Text Mining, the world's first publicly supported center for text analysis in the world, with an initial focus on biomedical information.

- Biology is a huge area for textual data analytics. For example, a project to link locations on the human genome back to specific research articles about those locations is underway at a project called text2genome. Every scientific field could potentially have this sort of mapping between research conclusions and the research where those conclusions were proposed and confirmed.

- There is also fun and creative analysis going on with texts such as cookbooks—what you might call recipe analysis. In fact, even IBM's Watson team has gotten into recipes. Their work uses not only the text of recipes, but also other data contributed by expert chefs to create new recipes. In fact, ways of combining text analysis with other tools will probably become more prominent in the future.

Suggested Reading

Berry and Browne, *Understanding Search Engines*.

Montfort, Baudoin, Bell, Douglass, Marino, Mateas, Reas, Sample, and Vawter, *10 PRINT CHR$(205.5+RND(1)); : GOTO 10*.

Activities

1. Sentiment analysis is an active area of research. Look for it in the news, or look for news stories on it periodically to see what is happening lately. Keep up with the field, and when you learn of new research, see if you can identify new products or features in software that take advantage of it, or think about how you would use it in products.

2. When you look at tweets, text messages, or e-mails, how do you discern the sentiment? When you misread someone's sentiment, why does this happen?

3. Word clouds are a very simple version of text analysis. Online software will allow you to cut and paste words into a web page and create a word cloud. Try it. Do you agree with the sentiment that the word cloud is conveying? Can you create examples where this simple version of graphical data analysis would be misleading? Remember, data analysis won't tell the whole story and sometimes can easily convey a misleading one.

Data Compression and Recommendation Systems

Lecture 19

Data compression is powerful: It allows us to store, access, and use far more images. But the mathematics inside data compression is even more powerful and can be used in other ways, too. In particular, you will learn that the same methods for throwing out data can be adapted to improve an online recommendation system. The mathematics of data reduction makes it possible to decompose a recommendation based on thousands of people. Of course, the recommendations work best if you offer some ratings yourself. Data analytics can then predict not just whether you'd like a particular movie or not, but also what score you might give it.

The Netflix Prize

- Netflix put together a million-dollar competition to improve their recommendation system. To enter the millionaires' club, you needed to do much more than recommend a film to a user. In a sense, you had thousands and thousands of people telling you the films they like and how much they like them. It came via the Netflix ratings data. To win the cash, your computer algorithm had to do a better job of predicting than Netflix's existing recommendation system at the time, called Cinematch[SM].

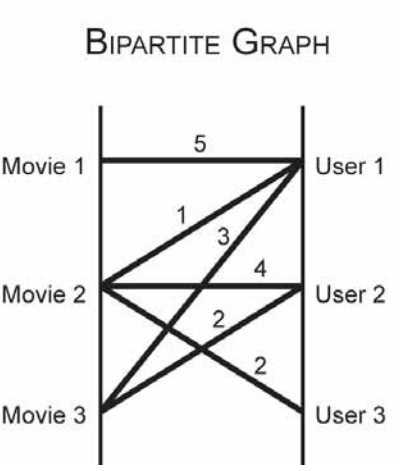
- Your goal was to take Netflix's data set of users ratings, which are integers from 1 to 5, for all movies. From that, predict how a user would rate other films. Note that this is more difficult than just recommending. You are actually trying to predict how much a user would like it.

- In the competition, Netflix supplied data on which you could test your ideas. When you thought you had something, you'd test your recommendation method by predicting ratings of movies and users in another data set. Your recommendation system had to predict future ratings. So, the second data set had actual ratings of movies, and you'd test your predicted ratings against the real ratings of

movies. If your predictions were at least 10 percent better than the recommendation system Netflix's used, Netflix would write a check for a million dollars.

- In data analysis, you need to know the nature of the data. Netflix gave you users and movies. One way to store this is in a table. A column is one user's ratings for all the movies, with a 0 being a movie that wasn't rated. So, a row contains all the ratings for one movie. We connect a user and a movie with a line if the user rated the movie. We associate a number with the line, where the number equals the rating of that user for the movie.

- This type of diagram is called a bipartite graph. The lines are called edges. For this application, the numbers or weights represent the number of stars a Netflix user gives it. For another application, the weights could be a 1 for a thumbs-up or a −1 for a thumbs-down. So, you have all of this training data in the form of preferences. Remember that the goal is to use that data to predict ratings.



- Suppose that a user enjoys *Braveheart*. What can you recommend? You have a slew of data on this user, and on other users and movies. One approach is to find the user whose opinions are closest to this particular user. What did a different person like? Recommend that movie to the *Braveheart* user. If you are trying to predict this user's rating, what rating did the other user give?

- But how do you figure out who in the data is most like this user? There are a few ways to do this. One way is to simply treat a row of data as a point. So, if you have two data entries per row, then you might have the point (1, 5) and (4, 5). You simply compute the distance between these points as if they were points in two dimensions. If you have five entries per row, then you find the distance between the points in five dimensions.

- Another way to measure distance is known as Jaccard similarity. This equals the number of preferences in common divided by the total number of things. This method captures the fact that two users have similar things that they like and dislike.

- These two distance measures are different. So, you'd want to think carefully about what they both mean in terms of the question you are asking. For example, Jaccard similarity only considers when users rate movies exactly the same. Any other ratings are ignored, whether they are close or not.

- Measuring distance like this can have double-counting problems that lead to overfitting and bad performance. To help with this, we need to reduce what is called the dimension of the problem. We need to get rid of such redundancies.

- When the Netflix competition was announced, initial work quickly led to improvement over the existing recommendation system. It didn't reach the magic 10 percent, but the strides were impressive nonetheless. The key was using linear algebra, specifically a technique called the singular value decomposition (SVD).

- If this algorithm is so straightforward and creates good results, why did the Netflix Prize take time? Had the competition been to improve the algorithm by 5 percent, it would have been over quickly. But it wasn't. Those last percentage points took more work, because not everything is easily predictable. For example, there's the issue of time: How you rate movies can depend on when you rate them. Furthermore, some movies are simply difficult to predict.

The Winner of the Netflix Prize

- The Netflix Prize took years of work. As such, it is rather amazing that the first-place winner crossed that digital finish line of 10 percent mere minutes ahead of the competitors. The team was BellKor's Pragmatic Chaos, a team of computer scientists, electrical engineers, and statisticians.

- Interestingly, the first-place winner and the second-place winner, the Ensemble, were amalgamations of teams that started off competing separately for the million-dollar prize. It's when separate teams joined forces with other teams that the final leap beyond the 10 percent was made. It was by combining teams and algorithms into more complex algorithms that those final advances were made.

- The ideas that were further away from the mainstream proved the most helpful at making final improvements that won the prize. For example, what about the number of movies rated on a given day? This information didn't predict much on its own. But movies rate differently on the day they are seen compared to movies reviewed long after viewing. And it turned out that how many movies were reviewed at once could be used as a proxy for how long it had been since a given viewer had seen a movie. In a sense, the prize-winning algorithm was a meta-algorithm that combined weights for a variety of simpler algorithms.

Other Recommendation Systems

- Recommendation systems appear all over the Internet. For example, Amazon recommends movies and books based on the one you are looking at. A different kind of example is Pandora, which has over 70 million active listeners each month. Pandora's success is rooted in an idea that was a commercial failure: the Music Genome Project.

- Pandora was launched in 2000 by Tim Westergren and a small team of entrepreneurs. They wanted to create a database of musical characteristics for a given song in order to identify other songs with similar qualities. Instead of starting with listener ratings, they built up from data about the music itself.

- How do you find songs that have similar attributes? Computer algorithms can do this, but the Music Genome Project team believed that such identification required a human touch. So, they hired musicians, who knew music theory, for example, to listen to each song. Then, they broke the song down by as many as 450 predetermined attributes, giving each a numerical value.

- They didn't get much response to license the music recommendation data. So, in 2005, Tim Westergren cofounded Pandora. It still uses Music Genome Project data to generate the custom playlists for its users. And it still uses people to listen to and evaluate music. This is very different from what happened to in-house reviewers at Amazon, who were dropped very early in Amazon's history.

- There is another human factor in the process: the user's interaction with the program. You can skip a song to hear a new one. Or, if you like a selection, click a thumbs-up. If you don't, click a thumbs-down.

- Pandora tracks users' interactions. Thumbs-up, thumbs-down, or skipping a song immediately affects what is played as the next song. But these things don't all affect what is played next to the same degree. Skipping a song carries less weight than a thumbs-down. Skipping a song doesn't have as clear of a meaning. You might skip a song because you've heard it too much or because you simply aren't in the mood at the moment. But it also depends on how often you use a feature. If you rarely skip and have been using Pandora for some time and then do skip, that says a lot about that interaction.

- Yet another mechanism is to use the wisdom of the crowd to filter and evaluate material. This is used by Reddit, which is a social news and entertainment web site. On this site, registered users submit content in the form of links or text posts. Users then vote submissions "up" or "down" to rank the post and determine its position on the site's pages. In 2013, there were over 100,000 subscribers.

- This type of site uses what is called collaborative filtering. It filters large amounts of information by spreading the process of filtering among a large group of people. Rather than having one main editor or group of editors, like a newspaper or magazine might have, the collaboratively filtered social web has its entire set of subscribers as editors, which also encourages participation.

- There are two basic principles involved in collaborative filtering. First, there is the wisdom of crowds and the law of large numbers. According to this, as communities grow, they make better decisions. In fact, this same idea is behind YouTube. People submit videos, and the wisdom of the crowd enables the best videos to bubble to the top.

- The second principle of collaborative filtering depends on the community being large enough, with enough data on individual participants and on how the individual participants collaborate or correlate with each other. The second principle suggests that we can make predictions about what these users will like in the future based on what their tastes have been in the past. That is, we can include collaboration when we create recommendations.

Suggested Reading

Eldén, *Matrix Methods in Data Mining and Pattern Recognition*.

Takahashi, Inoue, and Ltd. Trend-Pro Co, *The Manga Guide to Linear Algebra*.

Activities

1. From Amazon, to Pandora, to Netflix, to shopping sites, recommendations occur frequently. Think about what data might be leading to a recommendation that you see. When a recommendation is wrong, can you offer more data that might "recalibrate" the data? For example, you can make sure to rate songs or movies that you don't like, and then try to observe a change in subsequent recommendations.

2. When you make a recommendation, can you discern how you are making it? What factors do you consider most heavily? Does it make a difference what you are recommending or to whom? As a data analyst, could you automate such a process to possibly make broad recommendations like you see online?

Decision Trees—Jump-Start an Analysis

Lecture 20

Decision trees are one of the most transparent and powerful techniques in all of data analytics. In this lecture, you will learn how decision trees can help you analyze many kinds of variables. Sometimes, we create a decision tree, and we are done with our analysis; other times, decision trees are the first tool that paves the way for other methods. Either way, decision trees can carve quickly through your data, offering insight and possibly predictions about the future.

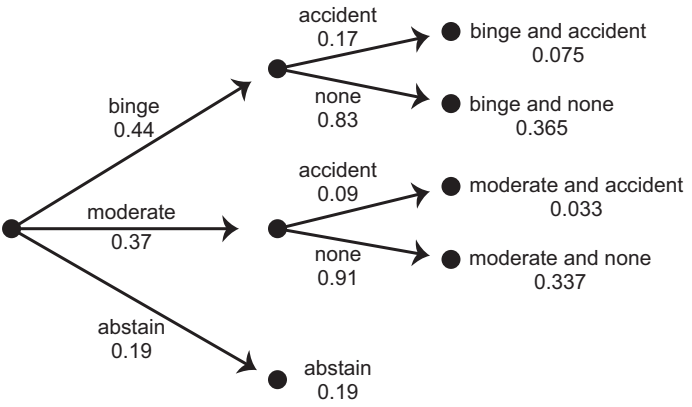
Decision Trees

- Let's start with decision trees involving probability, using a medical example from the text *Stats: Modeling the World* by David Bock, Paul Velleman, and Richard De Veaux. There are two studies. The first is a study by the Harvard School of Public Health called "Binge Drinking on America's College Campuses." They found that 44 percent of college students engage in binge drinking, 37 percent drink moderately, and 19 percent abstain entirely.

- A second study appeared in the *American Journal of Health Behavior*, and it reported that among binge drinkers between the ages of 21 and 35, 17 percent were involved in alcohol-related auto accidents, while 9 percent of non–binge drinkers in the same age group were involved in such accidents.

- Ignoring the fact that college students are often a bit younger than the 21 to 35 year olds of the second study, can we then combine the results of these two studies? That would let us determine the probability of a randomly selected college student being a binge drinker and in an alcohol-related car accident. A decision tree diagram makes it easy to combine two studies and answer such questions.

Binge Drinking and Alcohol-Related Car Accidents



- We're interested in college students, so the first step is to branch out according to college drinking habits. So, we have a branch representing the 44 percent likelihood of a college student being a binge drinker, 37 percent likelihood of being a moderate drinker, and 19 percent likelihood of abstaining.

- After we have visualized the results of the first study (about college students), we fold in the second study (about adult binge drinking). First, we have that 17 percent of the binge drinkers were involved in alcohol-related auto accidents. So, we add this to the branch related to binge drinking. Then, we also know that 9 percent of non–binge drinkers were involved in alcohol-related auto accidents.

- We are interested in the probability of a randomly selected student being a binge drinker who had been in an alcohol-related accident. We find this by first following the branch for binge drinking and then the branch for an accident. We can multiply the two probabilities along each path. So, we find 0.44 times 0.17, which equals 0.075. In this way, we can fill out the entire tree.

- We can also use the tree to find answers not already displayed on the tree. For example, if someone is a drinker involved in an alcohol-related accident, what's the probability that the person is a binge drinker?

- To find this probability, we are interested in a ratio involving the top branch (a binge drinker who was involved in an accident, which is 0.075), and we divide that by the sum of both the branches involving an accident, which is 0.075 + 0.033, or 0.108. So, we find this probability as 0.075/0.108 = 0.694, or about 69 percent.

- We started with two entirely separate studies, but we combined their data to find a clear result: The chance that a student who has an alcohol-related car accident is a binge drinker is more than 69 percent. Results like this might not be obvious from the two studies, but the results become very clear with tree diagrams.

- We combined the data of two studies to answer our own question. But you can, of course, also collect the data yourself. Decision trees can enable us to combine results to answer new questions or use probabilities from a large data set to study particular cases.

- Decisions trees have been a powerful research technique in medical research on heart disease, which is the leading cause of death in the world. A nice attribute of decision trees is that they produce a questionnaire that a doctor, or patient, can ask. They produce essentially an if-then-else type of structure: If this, then ask this, or else ask this other question. Furthermore, at each step of the process, we have a probability of someone having heart disease.

- There are nice but expensive programs that can help you create decision trees, but there are also inexpensive add-ins to spreadsheets like Excel. JMP software is not cheap, but you click a button and the data splits. You feed the program a table of data where one column is what you are trying to predict, and the other columns contain what might be predictive of that outcome. Then, you click to see what this might find.

- Like any technique, this may not work, because it is splitting one variable at a time and thereby missing something when two variables must happen concurrently. Still, it can be interesting and is frankly a fun way to explore data.

Classification Methods

- Decision trees are part of the larger field of classification. We use data to classify. Classification methods help detect a spam e-mail message from its header and content. Galaxies can be classified based on their shape as spiral or elliptical, and then split further. Banks can take data to determine if they believe someone should be given a home loan. In this case, you are predicting if someone might default on a loan. The resulting probability indicates if someone could be seen as safe or risky for a loan.

- Who visits each web page offers another example. When you visit a web page, a digital trail of sorts is left. Not a lot of information is available, but some is. For example, a data analyst would be able to see that you use the Chrome browser to connect with your Mac from a particular IP address.

- You can also know if a person visited a page by clicking a link on another web page. If not, that individual directly inputted the web page for the visit. This can be analyzed to determine the total number of visitors for a web page, along with how many visitors came from .edu, .com, and .gov sites. You can also tell what time and which day of the week people visit.

- We can use such a log to determine when we have a person visiting versus when it is a web robot. Web robots, often called web crawlers, automatically move through the web, retrieving information about web pages. For example, search engines use crawlers to see which web pages link to each other and even some or all of the textual content on the page.

- Suppose that a business would like to know if people repeatedly visit a web site, and if so, which products they view. If someone does visit the same product more than once, do rebates or free shipping aid in the customer making a purchase? To analyze such things, we first must remove web robot movement through the web pages.

- Rather than analyze such data immediately, we can combine the data in order to better analyze it. That is, we use the original data to construct a number of attributes not directly in the original data. In short, we are taking the data and essentially making a new data set that we can then analyze to gain insight on the original data. The resulting decision tree allows us to clear a lot of the noise out of our data and do a more meaningful analysis.

Advantages and Limitations of Decision Trees

- Decision trees have some important advantages. First, they are simple to understand and interpret. You don't need to know about data mining, or even much about the data itself, to understand how to use the results.

- Second, decision trees require little data preparation, apart from maybe combining pieces of data. Other methods may perform better, but preparing the data can take a much longer time.

- Finally, decision trees are often called a white-box method. We are able to take the results and see from the data why each split is made. Other methods, by contrast, give a black-box result, meaning that the result is often harder to see or explain.

- Decision trees do have limitations. First, decision trees aren't necessarily performing the best split. For example, a method might only make one split at a time. If we want an analysis that could be varied by changing the splits, we could use other tools, such as support vector machines—a model for machine learning that is basically a more nuanced way to do splits.

- Regardless of the underlying method, we can't consider every possible split, especially for larger data sets with more attributes. So, we choose some way of splitting and do the best we can. Just keep in mind that the split may not be the best.

- On the other hand, if you split too many times, you may have great results for the data you are looking at, but it may not work well for future data. This is called overfitting. Another problem with splitting too much is that the rules become quite complicated. Statistically, you may have a good descriptor of who to give a loan to, but it may become more difficult for a loan officer to implement.

- There could be subtle interdependencies in the data that a decision tree will not capture. Even so, it can bring us down to a manageable number of variables. Then, we can turn to regression and neural networks to refine the analysis. If you include too many variables in regression or neural networks, the data gets memorized. As such, you perfectly describe the data you have—that is overfit—but then you can't predict future behavior, except in the rare case that it already matches (perfectly) with past events.

- Decision trees are a powerful technique not only for decisions we make, but also for understanding all kinds of factors and probabilities contributing to a set of outcomes. Always ask yourself whether you can use a decision tree whenever you look at data. With decision trees, you create a huge sieve in the data deluge, keeping a lot of the good stuff and getting rid of a lot of the noise.

Suggested Reading

Bock, Velleman, and De Veaux, *Stats*.

Conway and White, *Machine Learning for Hackers.*

Activities

1. If you have a data set that you want to use, think about whether you run a decision tree on it. Again, it is one of the best initial tools, along with graphing, to use on data.

2. A key to decision trees is the output variable. What are you predicting? Is it one thing? Even if you don't have data, simply looking at life for things that could be analyzed with tools you learn if you *did* have the data is increasing your ability to think like a data analyst.

Clustering—The Many Ways to Create Groups

Lecture 21

Clustering is a powerful family of analytics for sorting data into groups—what we call clusters. There is more than one way to sort data into clusters, and which clustering method you use depends in part on your data. In fact, the choice of method can also affect the results you get—all clustering is not the same. Clustering is a widespread technique in data analysis. From political science, to medicine, to sports, to economics, clustering can be a tool to find connections and similarities in large data sets that otherwise can go unnoticed.

Clustering

- We saw one type of clustering in the last lecture, where we used decision trees to carve the data into groups. Decision trees are great when you have a directed flow of all your data toward a single target variable. But often, we want to look for groups where there is no carving of the data based on a single master variable. In those cases, clustering techniques are more appropriate.

- Fighting crime through predictions might sound like science fiction, but in fact, the analytics leading to this sort of police work started in the mid-1990s. The idea began with researchers Lawrence Sherman and David Weisburd, who developed a concept of clustering known as hot spots.

- They defined hot spots as "small places in which the occurrence of crime is so frequent that it is highly predictable, at least over a 1-year period." According to their research, crime is approximately six times more concentrated among places than it is among individuals.

- For example, one can find hot spots for robberies in the Bronx, which can help direct police to that area, for that type of crime. Even better, doubling or tripling the frequency of police patrols in these crime hot spots was found to reduce street crime rates by two-thirds.

- This idea of grouping items has many applications. In education, clustering can help identify schools or students with similar properties. In geology, clustering can help evaluate reservoir properties for petroleum.

- Market researchers group data from surveys and test panels. They group consumers into market segments. This can help find previously unidentified customers, develop new products, and select test markets.

- Deciding how many groups to make is a common and very important question in clustering. Algorithms can be used to form groups—by math. So, we must first see if we can make sense of how the math grouped, and second, see if there is something unexpected that the math found. There's an inherent balance there. Clustering at its best can discover something surprising. However, if everything is surprising, it's entirely likely that the method isn't working well with that set of data. So, you look to another clustering method.

- One way clustering methods differ is over how to calculate distance. Euclidean distance is measuring distance in space—for example, in the *xy*-plane. Another measure is by angle and is often called cosine similarity because you can measure by the cosine of the angles.

Clustering Methods

- Hierarchical clustering is a clustering method in which you don't have to decide on the number of clusters until you're done. This is a big feature of this clustering algorithm, and it's not true of the others in this lecture.

- To begin, each object is assigned to its own cluster. Then, we find the distance between every pair of clusters. We merge the two clusters with the shortest distance. Then, with this new group of clusters, we find the distance between every pair of clusters. We again merge the two closest clusters. We repeat until everything is in one single cluster. Then, we look at the process visually and make decisions on the final clustering that we may use.

- One part of this process is how to measure the distance between clusters. One way is to measure it as the shortest distance between a pair of points in each cluster. This method is based on the core idea of objects being more related to nearby objects than to objects that are farther away. These algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances.

- Hierarchical clustering is more flexible than some methods, but the results are not always as clear. This method provides a multilevel interpretation. You can easily zoom in and out to find something like subgenres, and you don't have to the number of clusters in advance. But you do have to specify how to measure the distance between clusters. It could be the distance between centers of your clusters, but it also might be the two closest points between clusters or even the two points farthest from each other in the two groups.

- Hierarchical clustering methods can potentially help with diagnosis. When a new patient with an unknown classification arrives, their data can be compared with the data from the existing classified clusters, and a classification for this new patient can be determined.

- Another common clustering method is called k-means clustering. This method is useful for point-wise data with distances. It creates what can be thought of as globs of data, where you choose the number of globs in advance. The k-means algorithm finds *k* cluster centers and assigns the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.

- First, you must choose the number of clusters at the outset. This is generally associated with the variable *k*, which is why the method is called k-means. Choose *k* center points, or centroids, for the clusters you are about to form. You can pick *k* points from the set of points you are about to cluster, or even *k* random points to begin.

- Assign each element of the data set to the nearest centroid. These are the clusters. Next, calculate new centroids: These are the average, or mean, of the data points in each cluster. Again, assign each element of the data set to the nearest centroid. Once every point is assigned, you have *k* clusters. Continue calculating new centroids and then new corresponding clusters until the clusters don't change.

- One big decision that has to be made is the initial centroids. It turns out that different initial choices can lead to different clusters. That can be a downside to k-means, but it is easy to run and quick to check. That's a huge benefit. One use of k-means is with downsampling images, which is usually done to reduce the memory of an image.

- Another method is spectral clustering. If you have data that's a graph (with vertices and edges), then you should use spectral clustering, which also looks for globs, but now in graphs. Spectral clustering uses eigenvectors from linear algebra to get more of the connections into a submatrix and not as many outside of it. A nice attribute of spectral clustering is that you get a unique solution, and it even finds an optimal solution.

- With globs defined using k-means, you have to pick the number of means in advance. With spectral graphs, you automatically get powers of two, which may not be what you want. You can't tell it to find seven, for example.

- These are only some of the most common methods; many more exist. Regardless of which method you use, clustering is used as an exploratory method. Clustering is about grouping items, so they lose their individuality. A particular description might not apply entirely to any one person you know, but parts of it might describe people you know. We look for a method that groups each person more accurately than alternatives.

- It is important to keep in mind that data analysis or math does the grouping. With the exception of decision trees, we generally can't immediately see *why* the groups turn out the way they do. So, once we get the results, we might need to collaborate with an expert who knows the data or application.

Using Software to Cluster

- Many times, data analysts take tables of data and use software to cluster. Many packages come with k-means and hierarchical clustering methods. There are many more. However, inherent to this work is deciding what to cluster.

- A few things can go wrong: The distance measure makes no sense, or the clustering itself doesn't fit the application. The easiest way to see why distance measures change is to think of driving. Do you want to know the distance between locations "as the crow flies" or how far they are given the roads you would have to drive on? Those aren't necessarily the same, and which one you want probably depends on why are you asking.

- Clustering algorithms generally have a style of problem they do well on. You start with the data, think about how to measure distance or similarity, and then choose a clustering algorithm. Clustering often starts by telling you pretty much what you'd expect. You may not have needed math to tell you much of what you're seeing, but then comes something unexpected. And with that unexpected result can come the insight.

- You look at the surprising data carefully. If possible, you verify the results, and you look at supporting data to ground your insight. And those surprising results can be where you learn something about that item—but often about the overall data, too. In the end, clustering can be most useful when it produces something that you in no way expected.

Suggested Reading

Foreman, *Data Smart*.

Gan, Ma, and Wu, *Data Clustering*.

Activities

1. Suppose that you have a data set of ratings where the rows are users and the columns are movies. Then, taking each row as a data point gives you a vector of user ratings. So, clustering the rows gives you similar users and could help with recommendation. Clustering over the columns can identify mathematical genres of movies.

2. Just like with decision trees, it can be fun and hone your ability to think like a data analyst if you look for aspects of life you would cluster if you had the data, regardless of whether you do. What would you cluster? What algorithm would you choose? What might you find?

3. If you find yourself grouping people or objects (like movies) together, what attribute are you using to cluster? Noticing these things in your life can help you think about how to automate it in your data analysis when the data is available.

Degrees of Separation and Social Networks

Lecture 22

We definitely live in a connected world. E-mail, mobile phones, social media, and video chatting all enable unprecedented connections. But how do we analyze just how connected we are? This is the area of networks—what are often called social networks, even when there is nothing obviously social about the network. Networks, social or not, offer a richer, deeper dive into the relations among points in your data set. You can and should look to networks whenever you have a relationship between objects in a system, because any set of relationships can be modeled as a network.

Degrees of Separation

- One of the more famous ideas of social networks is degrees of separation. Part of the popularity of this idea comes from a 1993 movie, and 1990 play, called *Six Degrees of Separation*. There are over 7 billion people on this planet, and the concept of six degrees of separation is that you can pick anyone on the planet, and there exists a path of acquaintances from that person to you (or anyone else). In particular, that path only consists of six people.

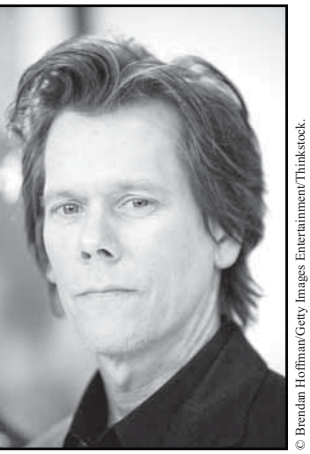
- In Europe, the concept of six degrees of separation started in 1929 with Hungarian author Frigyes Karinthy, who used it in a short story translated as "Chains." In the United States, this concept dates back to at least 1967, when social psychologist Stanley Milgram conducted what became known as his small-world experiment.

- The concept was popularized even more by Jon Stewart's *Daily Show* in the mid-1990s, when he referenced a game created by three Albright College students called Six Degrees of Kevin Bacon. The challenge there was to connect every film actor to Bacon in six cast lists or fewer.

- The Internet Movie Database has over 2.6 million movies and 5.3 million names. That's a huge data set. But even though all of these people make movies together, it's still surprising how few steps it takes to get from one actor or actress to another.

- In finding these connections, we are working through a mathematical structure called a graph. In such a graph, each vertex is an actor, and a link or edge is drawn between two vertices when both people appear in a film.

- Rather than degrees of separation, we talk about the distance between two vertices as the minimum number of edges that connect those vertices. For example, there are two edges between Kevin Bacon and Daniel Day-Lewis. Then, the eccentricity of a vertex is the maximum graph distance between that vertex and any other in the graph. The eccentricity of Kevin Bacon is claimed to be six or less. Finally, the center of a graph consists of all vertices that have the smallest eccentricity possible.



A game called the Six Degrees of Kevin Bacon involves connecting every film actor to Bacon in six cast lists or fewer.

- In 2008, Microsoft, after studying billions of electronic messages, computed that any two strangers have on average 6.6 degrees of separation. Researchers at Microsoft mined through 30 billion electronic conversations among 180 million people in various countries.

- The database covered the entire Microsoft Messenger instant-messaging network in June 2006. This was roughly half the world's instant-messaging traffic at that time. Two people were acquaintances if they had sent one another a message. The average distance between people was 6.6. Some were separated by as many as 29 steps.

Social Networks

- To investigate the connectivity of the Twitter network, social media analytics company Sysomos Inc. examined more than 5.2 billion Twitter friendships (the number of friend and follower relationships). So, a graph of that would have 5.2 billion edges. After an impressive amount of careful computing, they reported in April 2010 that there is an average of 4.67 steps between people.

- In November 2011, Facebook announced that there are, on average, just 3.74 intermediate friends separating one user from another. There were 721 million vertices in that graph. Interestingly, for a while, the average eccentricity got smaller as Facebook got bigger: In 2008, a much smaller Facebook had an average of 4.28 intermediate friends.

- These types of graphs, with edges representing connections between the vertices, are called social networks. They aren't necessarily social in context. They can be electrical power grids, telephone call graphs, or the spread of computer viruses.

- It is important to note that the type of graph needed for an application can differ. For example, Facebook friendships go both ways. If someone is your friend, you are also that person's friend. That's a graph with undirected edges. Twitter is different. You can follow someone, but you may or may not be followed by that person in return. This is a directed network, where the edges point from one vertex to another.

- While similar, these can be quite different in terms of the analysis tools available. You can easily cluster the Facebook graph with a very powerful technique. It's not that you can't for the Twitter graph, but the technique for undirected graphs doesn't immediately port over.

- Then, there are other layers of analysis—just in modeling a system as a graph. Is the existence of any connections between objects or vertices all you want? Is it enough to know that you are friends with

someone else? Sometimes. But maybe you'll want to integrate the number of interactions you have had with that person.

- For Facebook, this might be the number of times you have tagged each other in photos and left comments on each other's pages and such. For Twitter, interactions would probably be retweeting and mentioning. The number of interactions can be included as a weight on an edge.

- Is there any other information lost? Can you integrate that into a graph as well? Each time you switch to a slightly different graph structure, that generally means different algorithms for analysis. As such, you can't always answer the same questions on every graph. So, sometimes, the issue is not only what graph do you have, but which graphs might reveal the information you're wondering about.

- Is there a way to look at directed edges as undirected? Think about what happens if you simply remove the arrows. Twitter has directed edges, and Facebook doesn't. If you remove the directed edges from Twitter, you just lost the information that you might follow Bill Gates, but he doesn't follow you. So, does this make sense as a modeling decision? Maybe, but maybe not.

- Social networks, with the degrees of freedom, are called small world networks, which is a very active field of research. It's also a wonderful field for beginners, because it can be quite accessible.

Network Analysis

- In his book *Networks: An Introduction*, Mark Newman notes that networks really look for the pattern of connections between components. That pattern of interactions or structure of the network can have a big effect on the behavior of the system. It can affect how quickly news spreads and can influence how we form opinions or even how often we might see someone we know.

- Your Facebook friends can indicate where you are from. That's important to Facebook, because then they can provide ads and services in your area. But only about six percent of users enter their address.

- So, how does Facebook know your address? By pictures? Maybe. They're working hard on face recognition. By where you post? They probably do use that. But what's interesting is that they can already connect you to locations just by using those six percent of known addresses.

- Most people are geographically close to many of their active friends. So, Facebook can look at your connections to people with known addresses. Then, they can weight the importance of the edges by how recently and how much you and someone have been active.

- Address is not all that can be predicted. Students at MIT demonstrated that sexual orientation and religion could also be identified by Facebook, even when such preferences weren't mentioned. How? Again, by looking at the links to people for whom such details are known.

- One surprise from the study of networks is that, on average, your friends have more friends than you do. Said a bit more technically, the average number of friends of friends is always greater than the average number of friends of individuals. This comes from a 1991 paper by Scott Feld of the State University of New York at Stony Brook. It offers an interesting insight on friendship.

- Do you ever feel like your friends have more friends than you? They do. And the same is generally true for them, too. How can this be true? It seems wrong, like a paradox. In fact, it is called the friendship paradox.

- This has implications in social networks like Facebook and Twitter. On a directed network like Twitter, the people a person follows almost certainly have more followers than that particular follower has. But the same is true on a bidirectional network like Facebook. Either way, the reason for the apparent paradox is there: People are more likely to be friends with, or follow, those who are popular than those who are not.

- Today, connections can be found almost everywhere: people, information, events, and places. This connectivity is all the more evident with the advent of online social media. In this lecture, we see how we can gain from analyzing such connections. And there lies a key as you move ahead and consider network analysis. What you need is a connection. It doesn't need to be between two people. It can, but doesn't necessarily have to be.

- This is why network analysis has been applied to such fields as sociology, mathematics, computer science, economics, and physics. The World Wide Web is a vast network, and how you personally access anything on the Internet travels through a network of routers. Phone calls take place more efficiently thanks to advances in how networks of landlines and wireless transmitters are understood and managed.

- Network analysis is also shedding light on the connections between neurons in the brain. Better understanding of this network has led to advances in artificial intelligence. In addition, biology and ecology have long looked at networks of living species. Weights for the edges of such networks are moving them from the realm of cartoon drawings to powerful tools for analysis.

Suggested Reading

Easley and Kleinberg, *Networks, Crowds, and Markets*.

Watt, *Six Degrees*.

Activities

1. One way to play the degrees-of-separation game is on Wikipedia. Think of a target page, such as the page for The Great Courses, on Wikipedia. Then, go to a random page, which is an option on Wikipedia. Then, click a link on that first page that you think will get you closer to your target page. How many times do you have to click a link to get from the random page to your target page? This is sometimes called Wiki golf. If you play with others and start at the same beginning page, you can compare logic.

2. As you think about social networks, such as actors, athletes, friends, or colleagues, who do you think might be a center, or at least near the center, of that graph? Why? Play with friends. Everyone is a winner, because you learn to think more like a data analyst.

Challenges of Privacy and Security

Lecture 23

You have learned that there is a lot of data to analyze, and a lot of insight can be gained from analyzing it. In this lecture, you will learn about data privacy and security. What data is being analyzed? How difficult is it to keep data secure? And what can be done to improve security? With so much data, we often have less privacy than we assume we have, and quite a lot of information can be gleaned and known from our data.

Security Issues with Netflix

- In 2014, familiar smartphone apps, including Google Maps and Facebook, were shown to reveal personal information in unexpected ways—not just to the companies, but also to the U.S. National Security Agency and the Government Communications Headquarters in Britain.

- The Netflix Prize was the million-dollar contest that challenged experts in data analysis to use Netflix's data to produce better recommendations than Netflix did. But there was also a personal security angle. Netflix knew that they were supplying personal data, so they made an effort to remove identifying information.

- For the first Netflix Prize, they were successful. The data set covered about 480,000 customers. But when a second challenge was announced, Netflix got sued. The lawsuit claimed that Netflix indirectly exposed the movie preferences of its users by publishing user data, even though efforts had been made to remove identifying information and make the users anonymous.

- The initial data was out there, so it had already been analyzed by academics. In the end, 50,000 contestants participated; it was the fact that so many people were pouring over the data that made privacy advocates more worried.

- Plaintiff Paul Navarro and others sought an injunction to prevent Netflix from offering that follow-up challenge. Netflix wanted to take the recommendation challenge another step. They promised to include even more personal data, such as genders and zip codes, which could provide interesting answers to some fundamental questions. But the lawsuit was settled, and the sequel competition was cancelled. Netflix also settled a negotiation with the Federal Trade Commission.

- Netflix released data having already thought about security; they didn't propose to release additional data without any forethought about the issue. So, how could one possibly figure out someone's identity in the ratings data when Netflix believed that such info had been removed?

- One way is with other data. For example, two privacy researchers showed that comments on another site, such as the popular Internet Movie Database, could help triangulate the identity of an "anonymous" Netflix customer. The dates for posting on both sites were often virtually the same. This made it easy to match entries from one database with the other.

- There are ways people propose that the data could have been masked. The technique called data masking can randomize the data, making it even harder to trace an entry back to any specific person. Such additional precautions might have allowed the second Netflix challenge to go forward. But the fundamental issue remains: How secure is secure, and when do you know?

Security Issues with Facebook

- When does your privacy or security actually change without you knowing? Clearly, it is in Facebook's interest to keep information secure. But it is also in Facebook's interest for users to share as much information as possible

- In 2007, Facebook released a new feature called Beacon, which was created to enhance how people share information with their friends on the web about things they do. Facebook benefits when users share more information with potential monetary value, so Facebook was very excited with the new feature. However, problems were discovered.

- A security researcher indicated that the online advertising system went much further than anyone had imagined in tracking people's Internet activities outside the popular social networking site. Beacon reported back to Facebook on members' activities on third-party sites that participate in Beacon, even if the users were logged off from Facebook—and even if they declined having their activities broadcast to their Facebook friends. In fact, users wouldn't even know this was happening or be given the option to block it.

- Beacon tracked certain activities of Facebook users on more than 40 participating web sites, such as Blockbuster and Fandango. Then, that user's Facebook friends were notified. There was much negative buzz around this emerging news. In 2009, Beacon became defunct. In the end, there was a 9.5-million-dollar settlement in a class-action lawsuit against Facebook. The money set up a not-for-profit group that addresses online privacy rights.

- Both Netflix and Facebook would have been aware of privacy concerns about these projects. Yet their projects did not have privacy as a primary objective. And they did end up having trouble—in both cases, costing them millions. And in both the cases, the U.S. Federal Trade Commission also got involved, in an effort to establish clearer rules about privacy online.

- Why is Facebook so eager to know so much about you? The more Facebook knows, the more it can produce better, more relevant results for you. It can show you information about those that it deems are most important to you, simply by knowing who is connected to whom. Facebook and other companies are likely to know more about you than you intend.

The Security of the U.S. Government

- Incursions by large companies, actual or potential, can seem small or harmless when compared with the vast and unexpected reaches of the U.S. National Security Agency. Revelations during 2013, by a former technical contractor for the NSA named Edward Snowden, leaked sensitive documents. A vast system of government monitoring and archiving from phone lines and online services was unveiled.

- Snowden made it harder for the U.S. government to spy on U.S. citizens and other law-abiding people. This led some to say that he should be protected, like a corporate whistleblower. To others, he was a villain. He broke the law. He made it harder to block cyber attacks from other countries. He made it harder to catch terrorists.

- There are two security issues here. First, the security of U.S. government information was breached by the leaked documents. Second, the revelations showed how readily the security and privacy of information about individual U.S. citizens could be breached. Calls by U.S. citizens were revealed to have been recorded and stored in a vast database. The U.S. government defended the program as court-supervised and as a powerful tool that has thwarted terrorist attacks and protected citizens.

- Part of what has made the new surveillance techniques so powerful was the ability to analyze previously neglected data in new ways. For example, a 2012 study published in *Nature* reported that just four data points about the location and time of mobile phone calls were needed to identify a specific caller 95 percent of the time.

- These debates underscore two very different opinions people can have about security, how it is implemented, and what makes sense. On the one hand, the security of government information can be essential to the defense of the nation. On the other hand, the security of personal information against unreasonable search can be an essential value of the nation.

- We can expect to see at least three big practical consequences from these disclosures. First, organizations are rethinking how to effectively encrypt their most sensitive data. Second, international organizations consider doing less business with U.S. companies, because the NSA has methods and even agreements to see the data of U.S. companies. Third, many organizations are more hesitant to put their data on what was the fast-moving field of cloud computing.

- We can also expect further disclosures. NSA and other organizations can sometimes install and use covert radio-wave technology to spy on computers that are not on the Internet. Computers with wireless technology can be secretly accessed from increasingly remote distances.

Making Data More Secure

- Passwords are gateways to many forms of information about you. IT security consultant Mark Burnett collected and analyzed over 6 million passwords, and over 91 percent of these passwords were from a list of just 1,000.

- Changing passwords frequently doesn't help much, at least not if the password is short or common. It is better to have a long password—the longer the better. Experts like Burnett also advise using nonstandard spellings, including capital letters in nonstandard ways, and using non-letters.

- It's a really bad idea to use the same password across more than one account; not all organizations have equally good security. If your one-and-only password gets compromised anywhere, then you've just allowed it to be compromised everywhere.

- At many levels, we want data available. We want a rich set of ingredients available for our computational laboratory. But that data is often about us—our society, our companies, our families, even our most personal details. So, we want it to be secure.

Creating long, complicated, unique passwords is an important step you can take to keep your digital information secure.

- Let's look at how companies and the government make their data more secure. First, they keep outsiders out. Smartphones are less secure and more vulnerable than commonly believed, so they are not allowed to be used in the White House Situation Room, for example.

- Second, they keep inside things inside. For example, a virtual private network (VPN) extends a private network across a public network, such as the Internet. Using VPN, you can send and receive data across shared or public networks as if it were directly connected to the private network.

- There is a lot of interest in how to increase the privacy and security of specific data sets. One approach is more use of cryptography. What works for passwords can be extended to databases, even to

the creation of encrypted data. The technology that is used to make commercial web sites secure can be extended to many other kinds of web sites.

- Another approach is called differential privacy. This can be thought of as aiming to solve the Netflix Prize problem by adding increments of noise to a database. Work on this has been supported partly by Microsoft. The idea is not so much to hide the information entirely, but instead to mask just enough so that a data set can be used freely, without revealing sensitive data.

- Should the government have more data available to help fight crime, espionage, or terrorism? How much data should businesses share with each other or the government? How much should businesses or government share with the rest of us?

- Even if we make up our minds about this—even if we know what we want—we can be going along thinking the data is secure, but it may not be. We will continue to see disclosures in the news, stories revealing data that was suddenly found to be insecure. Do not think that you are doing things that won't be noticed. If you are sharing digital data, it might be visible. In fact, somewhere your digital data almost certainly is visible, at least potentially.

Suggested Reading

Angwin, *Dragnet Nation*.

Singer and Friedman, *Cybersecurity and Cyberwar*.

Activities

1. The balance between gaining insight with data and security and privacy is a tension we will have, likely for some time. As news stories emerge about data and security, think about that balance. What part of this tension is at the heart of the issue? Do you agree with the decision? What other issues do you think might emerge?

2. Security breaches of data inevitably emerge, from small to big companies. In what ways did we think we were safe? Watch the news and reflect on your own thinking and actions. How does this new information change your behavior or outlook?

Getting Analytical about the Future

Lecture 24

Throughout this course, you have discovered that data analysis allows us to predict the future. The fancier, more recent term for this is predictive analytics. In the future, data analytics will continue to become more powerful at cracking all sorts of problems, and you may need to keep trying to get a successful outcome. Hopefully, the tools of data analytics help you explore paths you would not otherwise take. Once you find something, like explorers of old, plant your stake and begin exploring even more.

Predicting the Future

- Sometimes, it's easy to predict the future, while other predictions aren't very clear. Sometimes, that's because we have lots of uncertainty, even if we have lots of data. Other times, predictions seem impossible because we don't see any data or patterns to analyze at all.

- The same range of issues comes up when we discuss trends in data analysis itself. Some things are clear, some have uncertainty that can be reduced, and others may yield to a new approach to gathering data.

- First, it is clear that analytics will change. If a company you know today is still relevant and strong a few years from now, it will have adapted and changed. The world of data analytics will, in many cases, pass you by if you stand still.

- Where will data analytics go? In part, it will come from what we are able to dream as today's innovations emerge around us. In fact, an important feature of data analytics is that the more things change, the more things *change*! They do not just stay the same.

- No matter where we go in this changing landscape, there are some fundamental principles that are unlikely to change. The following are four principles that we can keep in mind as we think about predicting the future.
    - Prediction comes from insight, not just from data. When new technology comes, we often think that it will make things better. That's not always the case. The same is true with more data: It's how we use the data, not simply its presence. More data, if not properly analyzed, can be a problem. More data can mean less insight. The goal is good predictions and more insight, not necessarily more data.

    - The value of prediction comes from context. Analysis is not just what you know. You need to situate what you know so that it becomes the right insight for the right problem. For example, the insight you glean from a data set may seem relatively unimportant to your context but turn out to be more important in a larger or different context.

    - Don't overestimate the value of any one prediction. Results in data analysis are better seen as an informed opinion. We have the ability to read a lot into results and assume causation where there is simply correlation. Analytics can help improve our predictions, but they are not necessarily stating truth or absolute outcomes.

    - Don't underestimate the ability of prediction to anticipate and transform the future. Don't miss the boat. Have the courage to follow sound predictions and believe in the results.

Predicting the Future of Predictive Analysis

- Keeping these fundamental principles in mind, let's predict the future of predictive analysis. First, tools evolve. Google adapted its algorithm to stay current. Sports teams look for the competitive edge with new algorithms that offer new insight—from biomechanics, to training, to coaching strategies that can range

from draft picks to player lineups. So, watch for tools we've seen in this course to show up in new combinations, such as clustering plus ranking, simulations plus differential equations, or decision trees plus regression or neural networks.

- Second, we can predict that there will be new data sets. Every year, new medical research emerges, and with it, often new data to analyze and study. News media sometimes makes their data available for analysis. Sports data is often available, and with the advent of new measuring devices, entirely new data is available—whether play-by-play data or new data about conditioning and training.

- Third, new technologies will continue to impact data analysis. This can be software like SportVU technology that led to a massive influx of new data into basketball and soccer. However, it can also mean new technologies that create new questions in data analysis.

- Fourth, data analysis itself can drive new technologies. Gordon Moore of Intel famously jump-started wider thinking about information technology with a very simple analysis. He didn't have a lot of data, but just by looking at a very small data set about improvements in Intel microprocessors over time, he was able to make a powerful prediction. He uncovered a rapid and regular doubling of processing power, and that prediction became the starting point for a technology road map that has shaped the entire information technology industry for decades.

- New materials may become another key source of technological change for newfangled devices. For example, lithium-ion batteries, which are part of a family of rechargeable batteries, have applications from cars to computers. However, a general road map for such innovations is not yet available. Creating a new, revolutionary material can be a very slow process, especially when compared to the rate at which other new products are conceived and introduced to the market.

- Projects such as the Materials Genome Initiative announced in 2013 are aimed at using data analytics to accelerate such innovations. By gathering and analyzing data about materials in new ways, the goal is to reduce costs and cut the development time for new materials in half.

- The name of this initiative resembles that of the Human Genome Project, which set out to map the underlying structure of human genes. The Materials Genome Initiative, in a similar way, is attempting to gain a deeper understanding of how elements interact to form more complex materials. Both of these are essentially vast data analysis projects. And as more becomes understood, scientists and engineers will be able to create new materials.

The analysis of genetic information is the way of the future.

- The Human Genome Project itself continues to stimulate new analytics, not only in biological research, but also in medicine. Sequencing a single reference genome of 3 billion base pairs for human DNA was a huge first step, but research is now uncovering a vast amount of variation between individuals—and even between different cells within a single person.

- In the most distant future, we have more non-human cells in our bodies than we have human cells, due to all the friendly bacteria. The data challenges here will be even more enormous, as researchers try to track all the members of the entire ecosystem that make up a single person. As human beings, we have so much genetic material in common, yet genetic data will also show that each of us is unique to a mind-boggling degree.

- Medicine will increasingly be a matter of managing and using large data files that track and accommodate more and more of the variation that exists among people. A genetic variant in one context might be detrimental, but neutral or beneficial in another context—and only detailed data, both in big studies and from individual patients, will be able to make sense of this. Traits may be expressed, or recessive, only in combination with other traits, making the combinatorics of variation even more complicated.

- The tools we've seen for visualization will be enhanced in order to visualize such complex information, which will be a big gain for basic research. And personal data will become the cornerstone of personalized medicine.

- Another spin-off from having individual genomes is identification and visualization. In fact, we had DNA testing in the late 1980s, long before the entire reference genome was sequenced. But with much more data, it's becoming increasingly possible to predict one's facial features just by looking at one's DNA.

- Paleontologists have already been using DNA visualizations to show us revised pictures of dinosaurs and our hominid relatives, such as the Neanderthals. And the work on current humans goes even further. Researchers began using data on gene markers for just a few prominent facial landmarks, such as the tip of the nose or the middle of each eyeball.

- But more recent work already uses highly detailed data grids, with thousands of reference points, and these data grids are superimposed onto scans of 3-D images taken with stereoscopic cameras. The same progression toward a data reference map is taking place with the Materials Genome Initiative, in which scientists and engineers are seeking a road map that will make it easier to tune a new material to the exact properties needed for a particular application—and to do it faster and cheaper than ever before.

- There are many, many combinations that can be tried and arranged at the atomic level. A huge number of them could have useful properties. However, most won't. And going one by one simply isn't possible. So, the Materials Genome Initiative is using computers to model known and unknown materials and to simulate their behavior. In the end, they have lots and lots of data to analyze in order to help find areas that deserve a more careful examination.

- What do you see as a need around you? Where does data seem unruly? Where do we need better decisions? Where could you get access to data that might contain insight? What insights would be genuinely valuable? Answers to these questions could predict the next area of innovation. These could be the next advances, large or small.

Suggested Reading

Brabandere and Iny, *Thinking in New Boxes*.

Siegel, *Predictive Analytics*.

Activities

1. As innovations in data analysis emerge, consider the four principles discussed in this lecture. Does the innovation fit into one of the four, or is there possibly another category? What tools it is using? Does the tool relate to one we've learned, or is it something else, or is it new?

2. As you read about data analytics or find data sets, what questions could you ask? What new questions might be possible? It's fun to learn what others have done, and it's even more fun to ask your own questions and make your own discoveries.

March Mathness Appendix

This supplement is intended to guide you in creating your own personal bracket using either the Massey or Colley methods. In Lecture 16, you learned the Massey method, and a very slight change in the linear system allows you to also rank with the Colley method. Both methods have been used to rank college football teams and help place them into bowl games. The Massey method integrates scores into its ranking, and the Colley method only uses win and loss information.

If you prefer not to create the matrix systems yourself but instead make the modeling decisions that give personalized brackets, using the weighting ideas outlined in the lecture, visit Professor Tim Chartier's web page at Davidson College. Each year, he posts links to online resources that will allow you to rank and create brackets.

The key to these methods is taking the season of data and creating the linear system. Then, you only need to solve the linear system using mathematical software or even Excel. This produces the ratings, and sorting the ratings in descending order creates a ranked list from first to last place.

Massey Method

Let's denote the linear system for the Massey method as $M\mathbf{r} = \mathbf{p}$, where M is the Massey matrix and $\mathbf{r}$ is the ratings as a vector. The diagonal entry in row i of matrix M equals the total number of games that team i has played. The off-diagonal entries convey information about the games played between two teams. The element in row i and column j of M equals the product of -1 and the number of times teams i and j have played each other. Finally, the i^{th} row of $\mathbf{p}$ is the sum of the point differentials of all the games played by team i, where wins equate to a positive point differential and losses to negative values. For example, if team i won a game by 10 and lost a game by 8, then the accumulated point differential would be $10 - 8 = 2$. Once the matrix is filled out, we replace the last row of M by a row of 1s. The last element of $\mathbf{p}$ is set to 0.

Colley Method

A small change to the linear system $M\mathbf{r} = \mathbf{p}$ results in the Colley method. In particular, you take the matrix M, prior to replacing the last row with 1s. You simply add 2 to every diagonal element. Next, you use a different right-hand side. We'll call this new system $C\mathbf{r} = \mathbf{b}$. To form the i^{th} row of $\mathbf{b}$, you compute $1 + (W - L)/2$, where the i^{th} team won W games and lost L games. Note that this method does not include scores but only win and loss information.

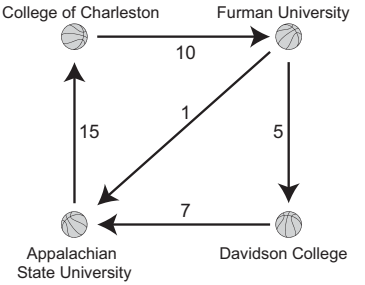


Figure 1. A fictional season played between NCAA basketball teams.

Rating Madness

To illustrate both methods, the following example uses a fictional series of games between NCAA Division I men's basketball teams. The records of the teams are represented by a graph in which an arrow points from the winning team to the losing team, and each edge is assigned a weight equaling the difference between the winning and losing scores. From the graph, you can see that each team plays every other team once, in a round-robin fashion.

For the Colley method, the linear system is as follows.

$$\begin{pmatrix} 4 & -1 & 0 & -1 \\ -1 & 5 & -1 & -1 \\ 0 & -1 & 4 & -1 \\ -1 & -1 & -1 & 5 \end{pmatrix} \begin{pmatrix} C \\ F \\ D \\ A \end{pmatrix} = \begin{pmatrix} 1.0 \\ 1.5 \\ 1.0 \\ 0.5 \end{pmatrix}, \text{ implying } \begin{pmatrix} C \\ F \\ D \\ A \end{pmatrix} = \begin{pmatrix} 0.5833 \\ 0.4167 \\ 0.4167 \\ 0.5833 \end{pmatrix},$$

where C, F, D, and A correspond to the ratings for Charleston, Furman, Davidson, and Appalachian State, respectively. So, the ranking (from best to worst) is Furman, a tie for second between Charleston and Davidson, and finally Appalachian State.

For **Figure 1**, the linear system for the Massey method is as follows.

$$\begin{pmatrix} 2 & -1 & 0 & -1 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 2 & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} C \\ F \\ D \\ A \end{pmatrix} = \begin{pmatrix} -5 \\ -4 \\ 2 \\ 0 \end{pmatrix}, \text{ implying } \begin{pmatrix} C \\ F \\ D \\ A \end{pmatrix} = \begin{pmatrix} -2.125 \\ -1.000 \\ 1.375 \\ 1.750 \end{pmatrix}.$$

So, the ranking (from best to worst) is Appalachian State, Davidson, Furman, and Charleston.

Personalized Brackets

A bracket is formed from such rankings by simply assuming that a higher-ranked team wins. A simple adjustment to the system requires mathematical modeling decisions and results in personalized brackets. You simply weight the games differently. For example, you may decide that the recency of a game is predictive of performance in a tournament like March Madness.

One way to measure this is to break the season into n equal parts and assign weights w_i for i from 1 to n. For example, assigning $n = 4$ breaks the season into quarters. Letting $w_1 = 0.25$, $w_2 = 0.6$, $w_3 = 1$, and $w_4 = 1.5$ assumes that a team's play increases in its predictability as the season progresses. For example, games in the first quarter of the season count as 0.25 of a game. So, it would be worth 0.25 of a win or loss for the teams involved. Similarly, in the last quarter, games count as 1.5 of a win or loss. As such, games differ in their contribution to the final ratings, increasing the potential of assigning a higher rating to teams that enter the tournament with stronger records in periods of the season that one deems predictive.

In addition, given the underlying derivation of both the Colley and Massey methods, teams that win in such predictive parts of the season *and* do so against strong teams receive a greater reward in their increased rating. Furthermore, the change to the linear systems is minor. Now, a game is simply counted as the weight of the game in its contribution to the linear systems. Before the matrices C and M were formed with each game counting as 1, now it is simply the associated weight.

So, returning to our example of breaking a season into quarters, a game would count as 0.6 games in the second quarter of the season. As such, the total number of games becomes the total number of weighted games. The only other difference is the right-hand side of Massey. Now, the point differential in a game is a weighted point differential, where the weighted differential for a game equals the product of the weight of the game and the point differential in that game. Again returning to our example, a game in the first quarter of the season that was won by 6 points would now be recorded as a 6(0.25), or 1.5, point win.

Got Data?

The last, and important, piece is downloading data that is processed to form the linear system. Professor Chartier uses http://masseyratings.com/. The page changes in format occasionally, so specific places to look for content can change. On the homepage, you want to look for a link to "Data," which may be under the section for "Information." Once there, you want to find your sport of choice, such as "Basketball," and then select "College" and "2014," if that's the year of interest. Note that this data is generally at the bottom of the web page below the series of links. Select your sport, and generally you need to click "All" to get to the schedules and scores information. You then want to request the data and select "Intra" games and download the "Matlab Games" and "Matlab Teams" data. These are simply text files that you save and can process without Matlab. With that, you are ready to process the data line by line, create your linear system, and rank the teams in your sport of interest. You may want to try this first for a professional sport, such as the National Football League or Major League Baseball. The process is the same, the linear systems are smaller, and the navigation of data is easier.

Bibliography

Angwin, Julia. *Dragnet Nation: A Quest for Privacy, Security, and Freedom in a World of Relentless Surveillance*. New York: Henry Holt & Co., 2014. An award-winning investigative journalist looks at who's watching you, what they know, and why it matters.

Bari, Anasse, Mohamed Chaouchi, and Tommy Jung. *Predictive Analytics For Dummies*. Hoboken: John Wiley & Sons Inc., 2014. This introductory-level book teaches you how to use big data in a way that combines business sense, statistics, and computers in a new and intuitive way.

Baumer, Benjamin, and Andrew Zimbalist. *The Sabermetric Revolution: Assessing the Growth of Analytics in Baseball*. Philadelphia: University of Pennsylvania Press, 2014. An all-star lineup recommends this book that discusses the past, current, and future of analytics in baseball. Read what's happening and think about what you might do in baseball or another sport of interest.

Berry, Michael J. A., and Gordon S. Linoff. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. Indianapolis: Wiley Publishing, Inc., 2011. This book contains a wealth of examples of data analysis in marketing, sales, and customer service. Their examples on the dangers of overfitting are helpful in framing one's thinking about this effect.

Berry, Michael W., and Murray Browne. *Understanding Search Engines: Mathematical Modeling and Text Retrieval*. Philadelphia: Society for Industrial and Applied Mathematics, 2005. This book gives many more details into using the singular value decomposition for textual analysis.

Bock, David E., Paul F. Velleman, and Richard D. De Veaux. *Stats: Modeling the World*. New York: Pearson, 2014. This is actually a text for high school students. It has great content on decision trees and the Titanic example.

Boyd, Brian. On the Origin of Stories. Cambridge: Belknap Press, 2010. This book explains why we tell stories and how our minds are shaped to understand them. This gives a sense of why humans are so prone to see patterns.

Brabandere, Luc De, and Alan Iny. Thinking in New Boxes: A New Paradigm for Business Creativity. New York: Random House, 2013. Now that you are equipped with a data analyst's toolbox, this book can help you begin to think broadly and innovatively. Innovation comes in various aspects of life, and as you've learned, data is a great place to gain new perspective to think far outside the box.

Bradburn, Norman M., Seymour Sudman, and Brian Wansink. *Asking Questions: The Definitive Guide to Questionnaire Design—For Market Research, Political Polls, and Social and Health Questionnaires*. San Francisco: John Wiley & Sons, 2004. This is considered the classic guide to designing questionnaires. It illuminates one of the many issues in collecting data on what you want to be asking.

Brenkus, John. *The Perfection Point: Sport Science Predicts the Fastest Man, the Highest Jump, and the Limits of Athletic Performance*. New York: Harper Collins Publishers, 2010. This is a book that uses data and math modeling to predict the limits in sports. Read the book and see if you agree with the analysis. If not, run your own tests and create your own predictions.

Chartier, Tim. *Math Bytes: Google Bombs, Chocolate-Covered Pi, and Other Cool Bits in Computing*. Princeton: Princeton University Press, 2014. In this book, I show how I found my celebrity look-alike among a library of 16 celebrities. These types of ideas model how facial recognition is done. In Lecture 11, we discussed these types of ideas as a means to identify handwriting.

Conway, Drew, and John Myles White. *Machine Learning for Hackers*. Sebastopol: O'Reilly Media, 2012. This book is intended for coders but helps the reader understand machine learning, a field in which decision trees lie, and their use in hands-on case studies.

Davenport, Thomas H. *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*. Boston: Harvard Business Review Press, 2014. This book looks at data analytics from the business perspective, using examples from companies that include UPS, GE, Amazon, United Healthcare, Citigroup, and many others. It really helps you get the big idea of big data in business.

deRoos, Dirk. *Hadoop for Dummies*. Hoboken: John Wiley & Sons, 2014. To really get an appreciation for Hadoop, you may want to dive into this book to see its flexibility and power.

Devlin, Keith. *The Math Instinct: Why You're a Mathematical Genius (Along with Lobsters, Birds, Cats, and Dogs)*. New York: Basic Books, 2006. Among its various topics, this book examines how we improve our math skills by learning from dogs, cats, and other creatures that "do math."

Easley, David, and Jon Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. New York: Cambridge University Press, 2010. If you want to learn and do research in social networks, this is the book for you.

Eldén, Lars. *Matrix Methods in Data Mining and Pattern Recognition (Fundamentals of Algorithms)*. Philadelphia: Society for Industrial and Applied Mathematics, 2007. This book gives an in-depth look at matrix methods in the field of data mining. If you want a deep dive, this is a great book to help you on your journey.

Foreman, John W. *Data Smart: Using Data Science to Transform Information into Insight*. Indianapolis: John Reilly & Sons, 2014. This book covers far more than clustering, with a chapter also on, among other things, simulation. It also discusses k-means and other clustering method, such as spherical k-means and graph modularity.

Gan, Guojun, Chaoqun Ma, and Jianhong Wu. *Data Clustering: Theory, Algorithms, and Application*. Philadelphia: ASA-SIAM, 2007. This is a more technical book, but it is one that student researchers can use when working in clustering.

Gladwell, Malcolm. *Outliers: The Story of Success*. New York: Little, Brown, and Company, 2008. This book is a good reminder that an outlier can lead to innovation and success. There are many reasons to identify an outlier. As you read this book, think about what indicators might have signaled someone's success as an outlier.

———. *The Tipping Point: How Little Things Can Make a Big Difference*. Boston: Little, Brown, and Company, 2000. A well-built simulation can help identify how small changes can lead to big changes in events. This book can help you gain a perspective on this effect from a large variety of well-written examples.

Gray, Chambers, and Liliana Bounegru. *The Data Journalism Handbook*. Sebastopol: O'Reilly Media, 2012. This book provides a discussion of getting data without having to be a computer programmer.

Hsu, Feng-Hsiung. *Behind Deep Blue: Building the Computer That Defeated the World Chess Champion*. Princeton: Princeton University Press, 2002. This book tells the compelling tale of the work of humans behind making a computer that shocked the chess world by defeating the defending world champion.

Hurwitz, Alan Nungent, Fern Halper, and Marcia Kaufman. *Big Data for Dummies*. Indianapolis: John Wiley & Sons, 2003. This book contains a really helpful discussion on data management and Hadoop.

Johnson, Neil. *Simply Complexity: A Clear Guide to Complexity Theory*. London: Oneworld Publications, 2010. This book discusses complexity theory and connects to traffic jams, stock market crashes, and predicting shopping habits.

Langville, Amy N., and Carl D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton: Princeton University Press, 2006. This book is great for the deeper mathematics behind search engines—not just Google, but a variety of approaches. It also tells the history with wonderful and valuable insights along the way.

———. *Who's #1?: The Science of Rating and Ranking*. Princeton: Princeton University Press, 2012. This is the book of ranking. It gives you the math and the details of numerous methods and the math behind them, which can help you adapt and extend them depending on your interests.

Levitin, Anany, and Maria Levitin. *Algorithmic Puzzles*. Oxford: Oxford University Press, 2011. Do you want to learn different types of computer algorithms that programmers use to tackle problems? This book teaches it without coding. You learn through solving puzzles, some of which are even offered on job interviews.

Levitt, Steven D., and Steven J. Dubner. *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*. New York: HarperCollins Publisher, 2005. This book, along with the companion web site http://freakonomics.com/, model how to look at the world as a data analyst. Given Lecture 11's topic, pay close attention to how many topics discuss regression.

Lewis, Michael. *Moneyball: The Art of Winning an Unfair Game*. New York: W. W. Norton & Company Inc., 2004. This is the book that detailed the surge in data analytics with the Oakland A's.

Mayer-Schönberger, Viktor, and Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. New York: Eamon Dolan/Houghton Mifflin Harcourt, 2013. This book can help you see the wide range of problems data can tackle, helping broaden your perspective as a data analyst. For example, which paint color is most likely to tell you that a used car is in good shape?

Montfort, Nick, Patsy Baudoin, John Bell, Jeremy Douglass, Mark C. Marino, Michael Mateas, Casey Reas, Mark Sample, and Noah Vawter. *10 PRINT CHR$(205.5+RND(1)); : GOTO 10*. Cambridge: Massachusetts Institute of Technology, 2013. Starting with a single line of code, this book dives into creative computing. This book is collaboratively written, taking text that appeared in many different printed sources. It's an example of textual analysis in the humanities with an ever-growing field.

Neuwirth, Erich, and Deane Arganbright. *The Active Modeler: Mathematical Modeling with Microsoft Excel*. Belmont: Thomson/Brooks/Cole, 2004. This book contains a wide range of applications in math modeling and details how to analyze them with a spreadsheet program—in this case, Excel.

Oliver, Dean. *Basketball on Paper: Rules and Tools for Performance Analysis*. Washington, DC: Brassey's Inc., 2004. This book is for every student interested in an exceptional study that explains the winning, or losing, ways of a basketball team with data.

Osborne, Jason W. *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data*. Los Angeles: Sage Publications Inc., 2013. Remember, if you are going to work with data, you may need to spend time preparing it. This book gives easy-to-implement strategies to aid you.

Paulos, John Allen. *Innumeracy: Mathematical Illiteracy and Its Consequences*. Boston: Holt McDougal, 2001. This classic book underscores the possible implications of not understanding numbers. If you read this book and think about it in our modern world of data, the points and stories become all the more compelling.

Russell, Matthew A. *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. Sebastopol: O'Reilly Media, 2014. Students use this book to learn to tap into social web data.

Shapiro, Amram, Louise Firth Campbell, and Rosalind Wright. *The Book of Odds: From Lightning Strikes to Love at First Sight, the Odds of Everyday Life*. New York: HarperCollins Publishers, 2014. Simulations are often built on probabilities and odds of events occurring. This book can supply a wealth of such information and, if nothing else, can be a very fun read about data.

Siegel, Eric. *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. Hoboken: John Wiley & Sons Inc., 2013. This book will review many concepts of this course, including the Netflix Prize and machine learning, and it also offers predictions for the future.

Silver, Nate. *The Signal and the Noise: Why So Many Predictions Fail—But Some Don't*. New York: Penguin Press, 2012. This book, while not giving a lot of details of Silver's work, really underscores and lays out his thinking. This can help you understand the mindset that led to Silver's innovations. This book can also help frame your thinking about data and looking for insight in the noise of everyday events.

Singer, P. W., and Allan Friedman. *Cybersecurity and Cyberwar: What Everyone Needs to Know*. New York: Oxford University Press, 2014. This book discusses critical issues in this field while keeping focused on the key questions in cyberspace and its security: how it all works, why it all matters, and what we can do.

Smiciklas, Mark. *The Power of Infographics: Using Pictures to Communicate and Connect with Your Audiences*. Upper Saddle River: Pearson Education Inc., 2012. This book not only gives a wonderful array of ideas for infographics but also discusses why we are so hardwired to digest visual information quickly.

Takahashi, Shin, Iroha Inoue, and Ltd. Trend-Pro Co. *The Manga Guide to Linear Algebra*. Translated by Fredrik Lindh. San Francisco: Oluumsha Ltd. and No Starch Press Inc., 2012. Linear algebra is a powerful tool of data mining. This can teach you many ideas in the field in an entertaining style.

Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Boston: Pearson Addison-Wesley, 2005. This is a text, but it is a very complete guide to data mining. It also has very instructive and complete sections on data preparation.

Tufte, Edward. *The Visual Display of Quantitative Information*. Cheshire: Graphics Press, 2001. This is a definitive resource on visually displaying data for many statisticians.

Vise, David A. *The Google Story: For Google's 10th Birthday*. New York: Random House Inc., 2005. This book tells you even more of the tale behind Google's journey, from struggling for funding in 1998 to a data analytics mega success story.

Warwick, Kevin. *Artificial Intelligence: The Basics*. New York: Routledge, 2012. What are the blended boundaries of robots? Can machines think? This book can help give you insight on artificial intelligence as a larger field and one that has important implications for data analysis.

Watt, Duncan J. *Six Degrees: The Science of a Connected Age*. New York: W. W. Norton & Company Inc., 2004. This author is an expert in network theory and discusses a wide range of ideas in networks, from computers, to economies, to terrorist organizations. Learn about this field and how it is emerging to uncover insight.

Winston, Wayne L. *Mathletics: How Gamblers, Managers, and Sports Enthusiasts Use Mathematics in Baseball, Basketball, and Football*. Princeton: Princeton University Press, 2009. This entertaining book looks at a wide range of sports and delves into answering questions in each sport.