Week 7 transcript--Data, Method, and Analysis
(updated Feb 2022)

7.1 Data Collection

7.1.1 Data Collection Philosophy

This element addresses the following learning objective of this course:

- LO3: Assess and select specific data and the data collection methods that best fit a specific outcome or need.

An earlier data collection philosophy was to collect as much data as possible and then worry about what you'll do with it later. Now in part because of various international and regional policies on data protection and privacy, this philosophy perhaps has shifted to collect as much data as you need and then perhaps not any more. Let's think more about this new data collection philosophy.

First, if you collect more data than you need, then you have to worry about how that data is subject to different regulations. The strategic collection of data can help minimize any unnecessary risk. Second, if you collect a ton of data, then you have to store it. And when you want to analyze it, you'll have to sift through larger quantities of data. It's best to collect data with a purpose. Otherwise, you're just paying for server space.

7.1.2 Existing or New Data

This element addresses the following learning objective of this course:

- LO3: Assess and select specific data and the data collection methods that best fit a specific outcome or need.

Oftentimes, we have to decide between the use of existing or new data. What are some of the trade-offs between using off-the-shelf data versus collecting new data? The overly simplistic but perhaps helpful framework we could use is, is the pain worth the gain to collect additional information? Now, as you try to make this determination, ask yourself what was the purpose of the initial data collection?

Now, just because the purpose of the initial data collection was similar to the purpose of the new project, that doesn't mean that you could automatically use it. On the other side of things, just because there was a different goal behind the initial data collection, that doesn't mean you automatically can't use it. So we'll go through an example of where existing data may be sufficient and then one example of where existing data may not be

sufficient.

So let's think about-- let's imagine you work for an auto manufacturer. And you collect data on different configurations and on production. This is a case where existing data might be sufficient. So you collect data on these kind of-- on production and on output. And a year later, a new piece of equipment is made available. And you think it could improve efficiency.

Now, you'd like to do another experiment to test the effectiveness of a new piece of equipment. But leadership just wants you to deploy it and deploy it in the lines with the most potential for impact. Now, you are aware of regression to the mean. And so you're cautious about interpreting data after you implement the new equipment in lower yield production lines, but nevertheless, leadership decides to use the equipment in those lower yield production lines.

Now, here's an example of where existing data may not be sufficient. You work for a major retailer. And you recently started curbside pickup. You have data on the popularity of curbside for your company. And you want to consider the viability of delivery to users' homes. Now, you're tempted to extrapolate from the curbside data you have to home delivery. But there's reasons to believe that users may view the two services quite differently.

You have a few options. You can collect new data about users' willingness to embrace home delivery with your company, or maybe you could use industry data that's not specific to your company to make a decision. The broad reminder is to be aware of the initial intent of why the data was collected. You need to articulate how this existing data will help you answer your question.

7.1.3 Acquire and Clean Data

This element addresses the following learning objective of this course:

- LO3: Assess and select specific data and the data collection methods that best fit a specific outcome or need.

You may spend a lot of time in your job acquiring and cleaning data. If you have a role in the acquisition of the data, or the creation of the data, perhaps you could help ensure that you're getting it in the right format. Even if you have a role in that process, you still might struggle to get the data in the desired format.

You'll spend a lot of time cleaning the data to transform in a way that's usable for your analysis. For example, let's say you get data from equipment in some kind of manufacturing process. The data comes in one-minute intervals with a long numerical

code. The code is the four-digit time code, the day, the month, the year, and the machine ID, and the assembly line number.

Your manager asked you to provide daily logs grouped by fiscal quarter. And so you're probably used to-- depending on what region of the world you live in-- the month, date, year format. But you don't realize that the equipment provides information about the date in the day, month, year format until you look at the data, and you ask yourself, what the heck is this 13th month?

7.2 The Intent of Data Collection

Spend five minutes on the following prompt:

Think about a time you had to decide whether or not to use off-the-shelf data vs. collecting new data. Describe the scenario and what the trade-offs were between using new vs. existing data. In particular, what was the initial intent behind the data collection effort?

If you have not confronted this type of scenario, describe what you think are the top two or three trade-offs between using existing vs. new data.

7.3 Method Selection

7.3.1 Overview of Methods

This element addresses the following learning objectives of this course:

- LO2: Design and apply research questions.
- LO4: Justify an analytic approach that informs decision making.

Remember that your research approach is your broad plan. It includes your philosophical worldview about research, your design, and your methods. Your design-- in your design, you articulate the type of inquiry-- qualitative, quantitative, mixed-- and the specific procedures of the study. You could think of your design as a kind of recipe.

Your method is your plan, data collection, analysis, and interpretation. Now, I'm not going to mention specific tools or programs because those vary quite a bit based on the industry, based on the task, or even based on the year. These tools progress pretty quickly.

But let me outline a few different methods. Surveys of individuals, in-depth interviews, semi-structured interviews, or structured interviews, ethnographies, experiments to test an intervention on equipment or people, observational studies, some use a

cross-section, a single slice in time, or a time series.

7.3.2 Tips on How to Choose a Method

This element addresses the following learning objectives of this course:

- LO2: Design and apply research questions.
- LO4: Justify an analytic approach that informs decision making.

There's no strict formula that you could use to determine which method to choose. Some things you may think about are, what method can give us the type of answer we're looking for? Do we want a more qualitative or quantitative approach, or maybe a little bit of both?

What types of methods are familiar to our audience? Do they want a measure of statistical significance or some kind of metric of confidence? Do we need to be able to explain to others exactly how the method works?

What is your expected sample size? Some methods require many observations. How long will you have to complete the project? What existing information is out there? Can you build upon it or do you have to start from square one? Remember, if you remember nothing else, the questions should determine the method you choose.

7.3.3 Internal and External Validity

This element addresses the following learning objectives of this course:

- LO2: Design and apply research questions.
- LO4: Justify an analytic approach that informs decision making.

You could think of internal validity threats as aspects of our design that limit the validity of the conclusions we draw from our study. For example, let's imagine that you design an experiment among human subjects where you give students tablets and you want to see how this affects their attention span in class.

You might experience diffusion, which is an internal validity threat where those assigned to a control condition get some form of the treatment. In this case, the kids who are not assigned tablets might interact with students who were assigned tablets. Let's assume that you can't keep the kids away from each other. Perhaps, you need to assign treatment at the class level instead of at the individual student level.

External validity is when the particular sample, setting of the experiment, or timing of the project may limit the validity of the conclusions we draw from our study. For example, let's say you implement a project. And midway through, there's a major economic downturn.

It's unclear if those findings can be extrapolated to other time periods where there isn't major economic turmoil. This factor was external to the inner workings of the study. And oftentimes, these external validity threats are not within our control.

7.4 Interview or Survey

Spend five minutes on the following prompt:
Imagine you want to get a better sense of how customers use your product. You are considering the use of either interviews or surveys to gain that insight.

1. What are the high-level trade-offs of using interviews vs. surveys? 2. What questions should you ask to help determine whether interviews or surveys are the most appropriate for this task?

Example: What type of product is it? What type of organization? What type of insight do you want? What is the timeline?

7.5 The Intersection of Data, Method, and Analysis

7.5.1 Introduction to a Perspective from IBM's Journey

This element addresses the following learning objectives of this course:

● LO4: Justify an analytic approach that informs decision making. ● LO6: Navigate organizational, personal, legal, and ethical constraints to facilitate better decision making and improve communication.

So now, I'd like to introduce you to Irving Wladawsky-Berger, from IBM or was from IBM. Irving has a pretty distinctive intellectual and commercial history with that company. As he'll tell you, he's retired now, but he was really quite central to the development of data science and business intelligence, strategic decision making at IBM. And because he's been there, he lived through that history, I think he's got a really unique perspective on the development of business intelligence and its transformation to data science within IBM. And that's not to say that IBM's experience is the same as everybody else's, but I think it's a really powerful and interesting example of a really big player in this space.

So as you listen to Irving, I think, in particular, it's worth paying attention to a couple of

things. I would highlight, first, some of his really interesting insights on the sources and motivations for early manifestations of what would later be called data mining. It's not a phrase many of us use anymore, but it was an important piece of the development at IBM.

Second, he's got a really interesting point of view on the relationship between what he calls pattern matching and intelligence. And again, historically, how that relationship or thinking about that relationship really informed IBM's experiments with Deep Blue, the computer that eventually beat Kasparov in a chess game.

And then, finally, I think Irving has a really interesting perspective on models overall as a kind of philosophy of science and I think, particularly his perspective on the term elegance, which we'll come back to. He associates that principally with deductive modeling. As I said, we'll turn to that a bit later in the course. But for now, what I'd like to do is sort of tee up the question, does elegance, as we think about it in modeling terms, really matter so much anymore? Or is elegance a kind of obsolete way of thinking about the relationship between data and models that's best left behind? It's a pretty important question, and some of the reading from this week bears on it. I would listen to Irving closely and get his sense of how that notion has evolved historically at IBM.

So let's listen to Irving. I mean, apart from everything else, he's a humanist. He's always thought deeply and cared a lot about the impact of his work and IBM's work on people in society and so I think that comes through very strongly in his interview, and I hope you'll appreciate that. And we'll come back around to the other side and talk about some of the insights that come out of that.

7.5.2 Human Decision Making Is Complex

Interview with **Irving Wladavsky-Berger**

This element addresses the following learning objectives of this course:

> ● LO4: Justify an analytic approach that informs decision making. ● LO6: Navigate organizational, personal, legal, and ethical constraints to facilitate better decision making and improve communication.

So my name is Irving Wladavsky-Berger. I got a PhD in physics at the University of Chicago in the 1960s, which is relevant to what I'm going to discuss later, even though it's such a long time ago. I then switched over to computer sciences and joined IBM Research Labs in 1970 and worked in IBM in various capacities.

I spent the first 15 years of my career in the IBM Watson Research Lab outside of New York City, north of New York. I then switched over to IBM's business unit. And I led a number of initiatives to bring new technologies to market, including parallel supercomputing, the internet, Linux, and open source, and what we called on-demand computing and so on.

I retired from my full-time job at IBM in 2007. Since then, I continued consulting for IBM for a few years. And since March of 2008, I've been a strategic advisor at Citigroup, working on the evolution to digital money and payment. And I am affiliated as a visiting faculty with three universities-- MIT, Imperial College in London, and most recently NYU, where there is a new program at the new Center for Urban Sciences that the Mayor Bloomberg in New York City helped launch.

And I really believe that the biggest impact of data science will be in the people-oriented disciplines.

Yes.

And obviously that includes the social sciences, health care, business, management, government. So you actually are absolutely in the right place to drive this. Quite a few people are trying to drive it from computer sciences.

Right, right. So you and I first met, I think, in the mid 2000s, when our interests converged around the open source software movement and how that was affecting large scale businesses in the mainstream. But I'd like to take it back even a little further and probe your reflections on the history and evolution of the business intelligence movement, as it was seen from IBM. Could you tell us a little bit about what you learned from being right in the middle of that for so long?

Well, from my point of view, in anything complicated, different people will have differing story lines. But I attribute data mining to the high energy physics community. I think they were the first who came up with the notion of gathering gigantic amounts of data and then applying very sophisticated algorithms and powerful supercomputing to look for patterns in the data.

And the reason the high energy physics community did that is because that's the way you found new particles. You did experiments. They run accelerators whether it is in CERN, which, as you know, just found the Higgs boson last year and just got the Nobel Prize for the find, or Fermilab outside Chicago, or SLAC (SLAC National Accelerator Laboratory [formerly Stanford Linear Accelerator Center]) in Stanford. And so they

developed the techniques for data mining.

Now, because I've been involved with supercomputing through my career, first getting my PhD at Chicago, I worked with some of the pioneers in computational science, and later on when I did parallel supercomputing at IBM, I was pretty close to these various labs we're talking about, with CERN and Fermilab. And in fact, for about three years, I was on the board of governors of Fermilab. And I saw the work they were doing.

What I was interested in is that when parallel computing came along-- I think it really started taking off in the world, I would say, in the early '90s. There were experimental machines before that. But the technology wasn't powerful enough. You needed microprocessors that were inexpensive and you could aggregate in large numbers. That really started to happen in the late '80s, early '90s, when the PC-based microprocessors were becoming powerful enough. And then you had the Unix-based microprocessors, the RISC microprocessors, that were even more powerful.

And what we saw is that a number of companies started applying data mining to business problems, what they called business intelligence. They started looking for patterns in marketing-- how are their customers buying things, what are they buying? So they were doing the equivalent of what the high energy physicists were doing. But they were applying it to information about sales and markets and so on.

The financial community was applying it for fraud detection, that is, start looking for patterns that indicate that something was awry in the transaction. So things like that. And so they continued. And that work has continued.

If you put a lot of information into computers, you ask the computer to look for certain patterns that you were asking it about, something akin to intelligence started to emerge from just extracting patterns from a lot of data. And I think what really, really, really convinced people that this was going to be a very fruitful approach to AI is when, in 1997, IBM's Deep Blue supercomputer beat the then world champion in chess, Garry Kasparov, in a very renowned chess match.

I remember that. It garnered a lot of public attention. But I just want to follow the line. So the Deep Blue was principally a pattern matching induction machine, is that correct?

Totally, totally. Deep Blue just had lots and lots of information about different chess moves and what had worked and what didn't work. And what it did better than anybody could do before is surge all that information so that at any point in time, it could evaluate the probabilities that different moves were good moves. This is totally different from the way chess champions play chess-- totally, totally different.

Now, I remember in the '80s, when I was in grad school, people were struggling to extract those patterns from experts' minds to build expert systems and found that, again, they didn't know the right questions to ask because the knowledge was there, but it was so tacit, that it really couldn't be extracted into algorithms. Is that--

Yeah, I agree. And I don't think, even today, we know how to do this. We don't know how to do this, even today. For example, you know, the way humans communicate their knowledge is to write. I mean, they also do mathematical things and so on. But you know, when you have a great idea in your field, you write a paper.

Today, we have computers that can sort through the natural language and figure out what's in the paper. But there is a huge difference between getting lots of facts and lots of connections about your paper which you can then search and transforming that into the equivalent of the Schrodinger equation or Maxwell's equations, that is, into to an incredibly elegant model that lets you see things in the world around you that you couldn't do that.

So going back to, say, '97, '98 and what you've said about the elegance of, let's call them, human algorithms. It sounds like it would be reasonable to say-- and I remember people at IBM articulating this-- that there was so much inefficiency in so many business processes, that actually finding those inductively derived patterns, if you could do it, could actually generate an enormous amount of value without actually having the truly elegant formulation, mainly because they were essentially so much low hanging fruit to be taken out. And IBM built a very significant business doing exactly that. And in no ways is that a pejorative. In fact, it's incredibly powerful. Is that a fair characterization?

Yeah. I don't know truly whether, in those days, IBM was able to apply what we call big data now to business processes. I think that the sort of business process, or the re-engineering of processes, was probably more model driven than big data driven. That's my feeling.

Now, the reason for that is we didn't have enough data for that. But now we are moving into the whole data science world, the hope is that we will be able to do that more now and into the future. But I don't think-- this is my personal recollection-- that in the late '90s we knew enough to be able to apply this kind of sophisticated business analysis to the whole way companies operated.

So let me pose a more challenging, or potentially more challenging, scenario for the present. Lots of folks here in the Valley would like to move the data science world out of the kind of, let's call it the nuts and bolts of inventory management and linking logistics systems and other such issues, which are incredibly important, but into a different realm

of what they would call strategic decision-making of the kind that I think you've characterized in some of your writing as complex, rather than merely complicated.

That's right.

And I wonder if you could talk a little bit about how you see that evolving. How close are we to being able to do that? What are the main obstacles? And where do you think that might be going?

Well, I think that decision-making, applying big data, data science principles to the way we make decisions of all sorts, is definitely one of the most exciting promises of data science. But it's one of those things that is going to take us longer to understand and get right. But when we do eventually, will turn out to be incredibly deeper than we realize today.

The reason I think it's going to take us longer is that, whereas, let's say, decisions like fraud detection or inventory management are, let me say, relatively straightforward. And I mean relatively that the more data you have and the better the algorithms, usually the better you do.

When it comes to making a complex decision-- whether it's in health care diagnosis or whether it's in a business strategy or a government decision-- you are evaluating multiple options. You have to figure out how to give probabilities to different options. And you can only do that if you have models of how to assign the probabilities to the different options. You need models of what does it mean to have a good option. And those models have to be built because, as it turns out, we don't understand how people make complex decisions well.

So to really dig deeper and apply data science, I think we will be developing a whole new set of disciplines around decision-making. How do people make decisions?

7.5.3 Intelligent Systems and Philosophy

This element addresses the following learning objectives of this course:

- LO4: Justify an analytic approach that informs decision making. ● LO6: Navigate organizational, personal, legal, and ethical constraints to facilitate better decision making and improve communication.

So I want to take the opportunity to highlight once again one really important point from the reading and Irving's interview around human decision making. And Irving said it really clearly. The impact of data on complex human-oriented decisions, in his view, will take longer than we think. But it will also be much deeper and more profound than people can imagine.

And part of the reason, of course as Irving put it, is that we probably will eventually start focusing very intently on models of what people actually value in these settings in order to train algorithms. And it's actually really a simple idea. After all, if we want artificial intelligence or machine learning systems that are goal-aligned-- in other words, align with human goals-- then again, we need to know what those goals really are. And of course, in most cases, it's not simply a matter of being more efficient or getting to a well-understood and consensual goal in a simpler way.

That problem becomes really clear as it gets stretched to what are logically consistent but sometimes slightly absurd conclusions. In the last couple of years, really, this has become tagged with the notion of Nick Bostrom's work on superintelligence. So for those of you familiar with that work, you know what I'm talking about. For those of you who aren't, let me just sum it up really quickly.

It's an evocative notion. Imagine a super intelligent, artificial intelligent system that has exceeded the capability of human beings to think. And let's imagine that we have now decided that we want that system to solve problems that human beings just haven't been smart enough to solve. So we tell that system, human beings want there to be no more wars. We want the system to figure out how to end wars on the planet.

So what does the artificial intelligence system decide? Well, maybe it decides that because humans by nature always seem to be fighting wars against each other, then it may choose to end the phenomenon of war by killing off all the humans on the planet. Now, again, this is kind of an absurd science fictiony example. But Nick Bostrom has made many interesting arguments about less absurd examples that make the same point.

And so what it's really kind of pointing to is that the human decision processes and the values that people are trying to maximize are really complex. And so it's really often hard to know what it is that human beings really are optimizing for as they go about their daily lives.

Just in the things we deal with every day, are we optimizing for justice? Are we optimizing for equality? Are we optimizing for aggregate welfare? Are we optimizing for fairness? All of those values might be present. But knowing what it is is actually on the table is going to affect what we want our systems to do.

And look. I get it. These are basic philosophical debates. And sometimes, people in a course like this might react by saying, look, that doesn't really have anything to do with data science. But I can't accept that because ultimately, despite the fact that they are basic philosophical debates, the act of choosing a training set for algorithmic decision

making is an engagement with exactly those questions. And when you choose that training set, you can't avoid engaging those discussions in some fashion.

So we've included in the suggested readings on the syllabus a paper that starts to dig into this in an interesting way. It's called accountable algorithms. And I would suggest you read it if you're interested in this stuff. In any event, we'll take up this issue in a little more detail in sync session. But the point for now is if you're going to be working in an area that involves important human goals, you actually can't avoid thinking about the relationship between the data set and the goals that you're trying to serve.

Now, I want to move on to a more contemporary product at IBM and introduce you to another IBMer. In this case, his name is Phil Nolan, old colleague of mine, good friend. Phil is not a data scientist by training. He's a guy that has specialized over time in decision making and has worked, actually, in many of the organizations that he's now trying to sell IBM's artificial intelligence products to.

So he has a really different and, I would say, a complementary perspective to what Irving offered. Sort of more of a like a demand function perspective than a supply function perspective. And he's going to talk about that with a particular twist on IBM's Watson system and how it's being used. Plus, Phil has a crazy sense of humor, and I hope you will let that loose at some point in the discussion. So I hope you'll enjoy listening to Phil, and we'll come back around and discuss some of those points afterwards.

7.5.4 Difficult Problems, Precise Definitions, and Repurposed Data

Interview with **Phil Nolan, IBM**

This element addresses the following learning objectives of this course:

- LO2: Design and apply research questions.
- LO4: Justify an analytic approach that informs decision making. ● LO6: Navigate organizational, personal, legal, and ethical constraints to facilitate better decision making and improve communication.

I'd like to welcome Phil Nolan, an old colleague of mine now working for IBM on Watson as well as another large set of big data applications. Phil, thanks for joining us. Nice to see you.

My pleasure.

Yeah. So, Phil--

My pleasure.

--tell us a little bit about your background, and how you came to be doing what you're doing today, and a little bit about what it is that you are doing for IBM.

Sure. I have been wrestling with issues at the intersection of technology and analytics for, god knows, 20-odd years. And until recently, when I was working with you, Steve, it was more from the analytics side-- how do you think about really difficult problems. And the last couple of years, I've been with IBM and I've been asking different questions, which is how do you use technology to solve some of those same analytic problems.

So from my point of view, nothing's changed. And yet, I think it's something that we're going to talk about over the course of this interview, which is it is an arbitrary separation to say there's analytic problems and big data technology problems, when in fact the oh aha moment is that they're really the same thing.

Very interesting. And, Phil, are you working mostly with the government clients, and/or, private sector clients, or a combination?

So it's a combination. Most of my customers right now are federal government. And I work with-- and use this as you will-- when I think of big data, it's a remarkably sloppy term. One of the things I like to talk about is structured data and unstructured data.

So-- stuff I'm sure that your students already know. Structured data, rows and columns like an Excel spreadsheet. Unstructured data is like New York Times, something where it is text, or video, or audio, which the information comes not just from the element but from the context around it. And that puts a big wrinkle on-- an additional complication on-- what we think of as just normal big data, rows and columns.

Great, Phil. Now just to clarify, I want to make sure that everyone knows you're speaking on behalf of yourself and not on behalf of IBM or any other organization. We're just asking for your opinion here and your thoughts based on your experience. But we would like to hear a little bit of your view on the trajectory and current status of this thing called Watson technology. It's one of these things about which there is both profound innovation but also a great deal of hype for people who are not using it and aren't inside of IBM. What can you tell us about Watson as a technology, where it's come from, where it is today, where it's going?

Sure. Watson has a remarkably interesting background. It was created as a big challenge-- a Grand Challenge at IBM research to find a way to demonstrate IBM's technical expertise around two areas, natural language processing and machine learning. And they were lucky, in my mind, to stumble upon a easy-to-understand problem-- how do you win at Jeopardy. In some ways, this is like JFK's "I want to go to the moon." Every moron, every person in the country can understand winning at Jeopardy.

Now, I will be the first to tell you I don't think there was-- well, not a single thing that was really innovative about Watson, but not that many. What was innovative and brilliant was bringing all the different technologies together to solve a problem which required Watson to be able to take natural language questions as an input and to deliver answers and evidence with confidence when drawing from a large corpus of unstructured text.

I'm going to just hit those again. So, inputs natural language from a user, you return answers and evidence with confidence from a large corpus of unstructured text. That's easy now-- everybody nods their head and says, yeah, yeah, we can do it-- and it was a remarkably hard thing to do.

I've also sat down with many non-IBMers, have walked through the technology, and they go, oh yeah, that's pretty straightforward, that was straightforward, that was straightforward. And it's true. There was like 150 straightforward things which were brought together at the same time in order to deliver this solution to a really difficult problem.

Interesting. And presumably, all the time ignoring the mustache of Alex Trebek.

Alex he was an interesting-- when I started working with Watson, they said, how many users do you want to have on him? For my customer I said, well, initially I want 150, and then they want to get up to god knows how many hundred users. How many do you do? And they go, we've only ever had one at a time. Because they built a system to play a game, and that's a long, long way from a system which can be useful for real-world problems. So the only user at the time was Alex.

So, Phil, we've spend some time in this unit talking about what are sometimes called type 1, type 2, and type 3 problems. Some people call them tame, complicated, and wicked. Is that distinction a distinction that's useful to you or to your clients when you're thinking about using Watson against big unstructured data sets for the federal government? Is that something that means anything for you? And if so, could you talk a little bit about the kinds of aspirations or applications that Watson is being used for in

that respect?

Well, let me start with the second and talk a bit about the way people are using Watson and then get back to your categorization. Watson is being used in three or four areas. One is medical, in health care-- a little bit of diagnostics and a lot of treatment decisions. Second area is in banking, but in fact what it is is financial research-- I want to invest in this company, this bond, this god knows what, and all of the information is dispersed in text all over the place.

The third area with the federal government was trying to dig through a vast amount of unstructured text they have to help with analytics. If you're trying to answer a national security question, I'm reading all the newspapers that come out of Syria, what in god's name is going on? Well, good luck with that. But it is a similarly messy problem.

The interesting thing is that the wickedness or the simplicity of the problem I believe is definable. None of these problems are intrinsically that hard if you define the use case tightly enough. But if you don't, they quickly explode.

I talked a few minutes earlier about JFK and the man to the moon idea. Well, there were some parts of the whole lunar mission that were really, really difficult, but a lot of the other ones were pieces that had been solved before and needed to be lumped together and strung together in the right way. That's a bit like the way Watson works.

And the flip side is, if you're trying to have a Watson system that's going to diagnose your health, Steve, well, it ain't going to do that. If on the other hand, you say, I'm going to input a bunch of unstructured information-- your doctor's notes, some outputs from various tests I've had on you, information I know about your lifestyle-- and try to help a doctor identify which one of these five or six treatments is best for your high blood pressure, well, now you're starting to narrow it down and Watson can be really useful.

I see this again and again not just with Watson but with other big data problems. If you're trying to use a technology to solve an almost intractable problem, the first step is to define it a bit better, because it can often be tractable in the smaller scale.

So it's about the unstructured data is workable when you have a reasonably well-structured or fairly defined problem, but having an unstructured problem and unstructured data at the same time-- like, what's wrong with me, why do I feel so lousy today-- is not going to be a very interesting, or, at the moment, not a very tractable application.

That's a nice way of putting it. Recently I've been doing a lot more work with structured data. And so if you're looking at some of those remarkably large data sets that telcos,

insurance companies, et cetera have, they're actually looking for patterns that matter to them, either patterns of fraud and abuse or patterns of I want to find this set of customers and sell them something else.

A huge volume of data, but at some level it's actually not intellectually that tough, because you've got some very clear structured data and you're digging through it looking for some things. You need some powerful tools to do it, you need data scientists to help you get there. But that's a different thing than that very messy health problem that you just threw out.

So, Phil, when you're doing work like that, normally in a smaller data set setting one would contrast a deductive and an inductive method. It's starting to sound like actually in your work that distinction is a little bit fuzzy, like what's induction or what's deduction?

Or are you actually just doing very large induction? How do you think about it, and does that matter at all?

I don't think it matters in the situations that I've been working on. So one way to think about the unique characteristics of big data is it's often you're dealing with a data set which has been collected for a purpose other than the question that you're trying to answer. And if you're faced with such a data set, you in fact need to have some-- I would say some brilliant insights about what's in that data before you can apply the right tools and perhaps solve the problem that you're getting to. So if I was to think about it the way you're describing it, there's probably a mixture of inductive and deductive.

Interesting. So talk.

7.6 Analysis

7.6.1 Design Research to Generate Applied Insight

This element addresses the following learning objectives of this course:

- LO2: Design and apply research questions.
- LO4: Justify an analytic approach that informs decision making.

Ensure that your design will allow you to say something meaningful. What's your goal? Do you want to provide descriptive insight? Do you want to predict? Or do you want to explain? Based on your answer, your design might look quite different.

Do you need to provide a recommendation? Your project will likely look different if you want to explore various decisions versus simply predict some kind of outcome. What

level of confidence does the audience want? Will you be able to provide the level of confidence with the current design?

If the current design will not allow you to provide the level of confidence desired, you should communicate to your audience what you need to provide that level of confidence. Do you need more data, more time, more money to get to that desired level of confidence?

Think about statistical significance. Remember that something can be statistically significant, but not substantively significant. You want to avoid comments from your audience like, oh, that's nice, but how should we take action on that information? Or great, you got statistical significance. But it's unclear if that very small effect size matters.

The headline takeaway is-- or the headline question you should ask yourself is what is the goal of this project? And how do I design a project to get me as close as possible to that goal?

7.6.2 Keep the Audience in Mind

This element addresses the following learning objectives of this course:

- LO2: Design and apply research questions.
- LO4: Justify an analytic approach that informs decision making. ● LO5: Identify the audience and the most effective method to communicate a persuasive argument.

When you develop your analytical strategy, make sure to keep the audience in mind. Keep in mind their priors and their incentive structure. And think about what you will have to show your audience to convince them. What language or terms will you use to persuade? The narrative you build and the vocabulary you use may be a consideration that occurs to you after you conduct the analysis. But really, your strategy to communicate should influence the analysis you choose.

What's your audience's level of technical expertise? One approach is to design a project that your audience will be comfortable with. But if you want to use a framework that's not entirely familiar to your audience, make sure you provide sufficient information so that they can understand what you did. Does your audience want you to inform them about a particular domain? Or do they want you to help them make a decision? Your approach to the project will differ in these two scenarios.

7.7 Deliverables

7.7.1 Communicate Insight, Hypotheses Testing, and the Word “Prove”

This element addresses the following learning objectives of this course:

- LO4: Justify an analytic approach that informs decision making. ● LO5: Identify the audience and the most effective method to communicate a persuasive argument.

We won't always operate in a world of hypothesis testing. I totally get that. If we care about prediction, we should make sure that our audience is comfortable with whatever explanation we plan to provide about how the prediction works. Some may care about how the sausage is made. Others may not. Let's talk about a term that is mainly used in the context of hypothesis testing, but it's broadly applicable to research. Let's talk about the word prove.

Why do we say that we want to reject the null versus prove something? I'm going to quote Richard Heuer. He wrote The Psychology of Intelligence of Analysis. He says, quote, "A hypothesis can never be proved by the enumeration of even a large body of evidence consistent with that hypothesis because the same body of evidence may also be consistent with other hypothesis," end quote.

Think about an example. Let's say our null hypothesis is that there's no relationship between one's age and salary. Our alternative hypothesis is that the older one is, the greater their annual salary will be. We collect data that shows that older people make more money than younger people. Because we find evidence that makes our null hypothesis look ridiculous, we reject the null.

Now, this doesn't mean that we've proved the hypothesis that age has a direct impact on salary. As Heuer says, quote, "The same body of evidence may also be consistent with other hypotheses," end quote. In this case, an alternative explanation is that it isn't about age. It's about the number of years of work experience or education that lead to higher salaries. If we focus on proving a particular hypothesis, we can unintentionally ignore alternative explanations. We can get tunnel vision and focus exclusively on the one thing we want to prove.

The word prove also has a sense of certainty and conclusiveness that scientists are often uncomfortable with. That is, if we prove something, it's as if that's the last word, and we shouldn't study the topic anymore. Some domains are more or less comfortable with the
word proof. But most formally-trained scientists use that word infrequently. In fact, I'm pretty skeptical any time I hear an academic use the term.

So I want to close with a few comments from some industry folk I talked to about the word prove. One colleague works at an organization where there is quite a bit of comfort with the idea of evidence-based decision making.

But most business folks in the organization admit that they don't understand the statistics. So he often writes summaries with a short statement of statistical rigor and then a business recommendation. The word prove doesn't fit into this business context because they're often looking for recommendations.

Another colleague says they never use the word prove. They like the word suggest and find. They also use the term support because it's strong enough, but it still sets a tone that is open to more evidence and thoughts.

Finally, a colleague says that the main reason they usually avoid the word prove is because of how strong it sounds. Usually, in a meeting room of business people trying to make a decision, each of them has their own reason to support a choice. And some are very passionate about their position and firmly believe that their choice is the right one or the best one.

When one uses the word prove, they're communicating that some of the people in the room are right, and some of them are wrong. And by doing this, you'll almost certainly get a fierce response. And that could spark more extreme and unproductive conversations.

Furthermore, this particular colleague works in sales. And it's even more important not to make enemies as they are usually more vocal than those who are in favor of your solution. This colleague usually uses terms such as suggest, find, and show because they demonstrate some level of certainty but without an extreme sense of right and wrong without this extreme sense of right or wrong that the word prove brings into the conversation.

7.7.2 Different Deliverables

This element addresses the following learning objectives of this course:

- LO4: Justify an analytic approach that informs decision making. ● LO5: Identify the audience and the most effective method to communicate a persuasive argument.

Your final deliverable will depend based on the audience and the objective of your research. Some of you may deliver a slide deck and a presentation. For others, you may deliver a paper or a report. You may be required to show proof of concept or

minimum viable product. And some of you will have code-based deliverables. Maybe you're expected to produce a new model.

Whatever you produce, especially if you produce a code-based deliverable, make sure to future proof your work. Think about your future self. That's you five months from now or even two weeks from now. Or think about how someone else will interpret your code. I can't remember what I did last week let alone what I meant by some code a few months ago. Remember to annotate.

7.7.3 Different Audiences

This element addresses the following learning objectives of this course:

- LO4: Justify an analytic approach that informs decision making. ● LO5: Identify the audience and the most effective method to communicate a persuasive argument.

We expand on this elsewhere, but it's important to remember who your audience is. When we communicate, when we write, when we speak, we're talking to someone. Even if our audience is implied, we're communicating with somebody. It's important to make the audience explicit. And similar to how you use the research question as a foundation for your project, you should also use the intended audience as a guiding heuristic.

The way you talk to a technical group will be different than the way you talk to a team of C-suite executives on at least two levels. One, the level of technical expertise. Depending on the level of technical expertise, you will likely change the vocabulary you use. And you will emphasize some aspects of the project more than others.

Two, incentive structure. The C-suite audience cares more about recommendations and high-level conclusions. The C-suite might be more persuaded by the narrative you tell compared to a technical team that might be sufficiently convinced about your work if you only focus on the nuances.

7.8 Creswell and Creswell Textbook

7.8.1 Creswell and Creswell, Chapter 10: Mixed Methods Procedures

This will help you digest the required Creswell and Creswell reading. Please use these ideas as a starting point.

This element addresses the following learning objective of this course:

- LO2: Design and apply research questions.

- LO3: Assess and select data and the data collection methods that best fit a specific outcome or need.
- LO4: Justify an analytic approach that informs decision-making.

This chapter covers mixed methods. It's important that we're familiar with both qualitative and quantitative forms of research, because there's no one size fits all approach to a problem. And if you say, oh, well, I only do qualitative, or I only do quantitative research, you may be limiting yourself. Keep in mind that each approach has trade-offs.

Now one way to kind of respond to the fact that there's trade-offs to each method-- and there's no perfect way to tackle a problem-- is to take a mixed approach. Now remember, mixed method is not just a little bit of qualitative and a little bit of quantitative. It's not like you just sprinkle a little bit here and a little bit there and all of a sudden, you have mixed methods.

Instead, you need to make sure that the purpose of the study and the purpose of the mixed method approach is made clear and make sure that your use of both the quantitative and qualitative is intentional. So I turn your attention to the page 218 of the fifth edition of Creswell. And there's a really cool visualization of these three high-level approaches that I'm about to highlight.

So one is that you could do both quantitative and qualitative at the same time and see if the results converge. Now that's one way to kind of take a mixed approach. Another way is that the qualitative can inform the quantitative. Maybe your qualitative work helps identify what relationships you want to look at in a qualitative way-- excuse me, in a quantitative way. So again, in the second method, you do qualitative to inform the quantitative.

A third approach is to take a quantitative approach first, and then that will inform your qualitative approach. So you do a quantitative study. You identify some interesting patterns. And then you follow up with a qualitative approach.

So before I close, I want to think a little bit about the worldview. It's something that we touch upon several times throughout the course. Our view about science, and more generally, our worldview, or your worldview about how the world works can influence problem definition, the questions you ask, and which findings you emphasize.

And I think it's best to acknowledge our worldview at the beginning of a project. And I encourage us to be mindful of how that influences our thinking throughout the project. For example, I was trained in a very quantitative, post-positivist way. I was trained to think about measurement. This is both a blessing and a curse.

On one hand, I'm careful about whether or not the metrics we gather capture the idea we care about. But on the other hand, I'm sometimes quick to dismiss things that can't be easily measured. And because I'm aware of this, I'm kind of able-- with some level of success-- I'm able to apply the brakes when I say something or think something. But we can't measure that.

And so, instead I say, OK. That sounds tough to measure, but let's think. So the goal here is to kind of slow down our thinking.

7.9.1 The Data Detective by Tim Harford

This will help you digest the required Harford reading. Please use these ideas as a starting point. *The Data Detective: Ten Easy Rules to Make Sense of Statistics*. In the video below, Mike places focus on the following three chapters.

- Introduction: How to Lie With Statistics
- Rule 6: Ask Who is Missing
- The Golden Rule: Be Curious.

This element addresses the following learning objective of this course:

- LO2: Design and apply research questions.
- LO3: Assess and select data and the data collection methods that best fit a specific outcome or need.
- LO4: Justify an analytic approach that informs decision-making.

So in this unit you will engage to Harford, the data detective. This is an excellent text. This I think does a great job of summarizing the kind of different ways that we view statistics and how data science can improve decision-making. I think this text does a great job kind of summarizing that. And so I encourage you to, or you'll be required to read the entire text. But I want to highlight a couple key points. So I want to highlight the kind of important takeaways from the introduction. Rule six, ask who is missing and the conclusion. Be curious. So an introduction, I think Tim Harford does a great job kind of laying out his argument. And I think in contrast to the way that statistics is approached in How to Lie with Statistics. The idea that that Tim hard for promote is to embody a healthy level of caution or a healthy level of skepticism. However, in contrast with how to Lie with Statistics and that text, we don't want to fear statistics, but rather just kind of embrace that healthy informed level of skepticism. The other kind of big kind of motivation, I think that, that's behind most of the texts is that statistics and analytics are key. And so our persuasion, storytelling and communication,and those also are key tenants of this course. And finally, I think the introduction, and we'll come back to this point when I talk about the conclusion, really pushes the reader to ask questions. That is, I think at the core of Tim Harford idea, and I think as at the core of the 20 one idea which is to ask kind of better or different question than to embody that healthy level of skepticism. Okay, Let's turn to rule 6 asked who is missing. So in this context, in this particular chapter, we are encouraged to think about who is missing from our data sets, who is missing from our studies? And as Tim will cover, in many academic and business cases, it's often women or gender minorities and racial and ethnic minorities that are often missing. And so I think it's important to be mindful of at least two different types of kind of error slash bias. So one is sample error. So sample error is that the sample by chance does not reflect the population. And so you'll cover sample arrogant in other courses, in particular in stats and

also sample bias. And these two things are distinct. Sample bias is, we could think of it also as the selection effect. Here. The idea that who enters or what enters our sample is a systematic process that excludes some units or for talking about people that excludes some folks. And so think about the data generating process. Ask questions. What was the initial intent for this data collection process? For this data collection effort? How was it collected? Who is in the data, and also importantly, who is missing? And is this via some systematic process? And remember, the set of users that you engage with or your sample is not equivalent to all people. Because we might be unintentionally generalizing from our sample or from our users in a way that's not appropriate because oftentimes our samples or users does not reflect the population that we care about. Finally, I want to really focus your attention on the concluding chapter, the golden rule. Be curious. So there's kind of brings it back to this idea that a healthy level of skepticism, a healthy level of curiosity, and healthy level of caution is really important. I think we learn more on a personal level if we embrace this curiosity, I think we will solve problems kind of differently in the workspace and potentially solve those problems in a better way. Similarly, encourage curiosity in your teams. Part of that is embracing the idea that there will be failure, that things won't always work out. But that's part of the process, that's part of this kind of scientific approach to decision-making. There will be failures. But if we intentionally design our projects, we can learn from those failures. To find out what's storytelling matters. The statistics are important. A robust plan. Research design is absolutely important, as is that kind of communication step. And make sure to spend enough time kind of building out that final step. Because as null Nasa mom or Knaflic from it, from a different reading, mentions this kind of final story time, excuse me, final storytelling presentation, part of the analytical or research process. That's the only step that your audience sees. And so I'll make sure to spend time and I kind of story time. Finally, I think I think we could summarize ten Harford text with three words. And that's tell me more. Tell me more, I think is one of those kind of awesome open-ended invitations to learn more from your audience. So when someone says, Hey, let's use this data, all right, Tell me more. Where did this come from blows the data-generating process. Tell me more, I think are those three words that we and our team can use to embrace this kind of healthy level of skepticism, to embrace his curiosity, and to potentially do things different and potentially do things better. Thanks.