Week 13 transcript--The Future of Data Science (updated June 2021)

13.1 Data Science Project Lifecycle

13.1.1 Question-Driven Research

This element addresses the following learning objective of this course:

- LO2: Design and apply research questions.

We've made a case in this course for a relatively specific approach to research. We admit that there's not a one-size-fits-all design for every question. The design will vary depending on if we want to describe, predict, or explain an outcome. But the one thing we are pretty set on is that the question you ask should drive the design.

I know it's tempting to jump into the method, the tools, or the data. But remember, the goal is to design research in a very instrumental fashion. If we want to impact the decisions in our organizations, we need to craft careful questions, and then each design decision we make should be guided by the question. Any design decision that does not speak to the question is wasted effort.

You could think of your research question like the argument of a paper. Any paragraph, any sentence that does not help support your argument is wasted effort. And in fact, it may weaken your argument. Focus on the question. Be open to some iteration. We know that you will not get the ideal question the first time around. But keep an eye on that question, and use it to guide you through your decisions.

Keep focused. It's so easy to get sidetracked and focus on data or methods that do not serve you. The punchline is to spend time on the front end of the project. Ideation matters. Allow the question to drive your research. And this will give you the best chance to generate the intended insight.

13.1.2 Beware Of Shiny Tools and Methods

This element addresses the following learning objectives of this course:

- LO4: Justify an analytic approach that informs decision making.

- LO7: Imagine, plan, and design a data science project.

By this point in the course, you're probably pretty familiar with the idea that the question should dictate the design, the method, and the tools you use. Really, the question should guide everything. Continue to think of the question as the core principle that guides your decision. In particular, beware of shiny tools and methods.

Technology has become cheaper and more available. And there might be some perceived need to keep up with competitors, even though the technology may not be ideal to answer the questions we care about. Questions should come first.

Now, people in your organization or even you might think we need to invest in artificial intelligence. Everyone else is doing that, and it's the thing to do right now. But instead, ask yourself, do the questions we ask require the use of advanced methodologies? Can we do it a different way? When possible, keep it simple. Think back to earlier in the course. If explainability is important to your clients, then maybe certain methods are more attractive than others.

Now, if you are set on using something new, if you are set on using some kind of new cutting-edge method, go ahead and try it. But also use the tried and proven method that you're comfortable with. If they both point in the same direction, great. If they don't, then you need to think about how much you understand about this newer method.

13.1.3 Knowledge of Project Lifecycle Is Key

This element addresses the following learning objectives of this course:

- LO7: Imagine, plan, and design a data science project.

The hope is that this class has prepared you to say yes if someone asks you to design and lead a project. The goal is for you to have a holistic view of what it takes to move from an idea or question to an appropriate design that will produce actionable insight.

You don't need to be an expert in every aspect of the project. But if you understand the fundamental elements of what makes a quality project, you'll be prepared to take a lead role. Now, it's not enough to know the steps and the relative order.

You also need to know how the different steps are related. How does each component of the design fit together? For example, how does the way we measure concepts influence our ability to make an inference?

If we decide to operationalize a concept in a different way, how much does it change the interpretation of our findings? If you adjust one component of the design, what happens? How are the pieces interrelated? This is the knowledge that will make you very effective.

If a manager asks you to come up with a short memo to discuss how you might tackle a given problem, the hope is that you could summarize the process and could justify your approach.

13.2 Can You Look Into This for Me?

13.2.1 Discussions Prompt
Spend five minutes on the following prompt:

Imagine the following scenario:

Your boss asks you to figure out how you can use machine learning to reduce the number of customers who stop using your service (i.e., churn).

You are curious about the motivation behind the ask, and rather than say, "Ok, I'll look into it," you want to ask for additional information in order to best serve the needs of your colleague.

In 2–3 sentences, describe how you might ask clarifying questions to better understand the context around the ask.

13.3 Think Big, Part 1

Much earlier in the semester, one of our interviewees told us that CEOs and the kind of people who aspire to be CEOs tend to be really grounded people and they like to have their feet on the ground very much bound by reality. But the best CEOs also have among their key advisors almost always someone whose job it is to be a little less grounded in fact and a little bit more visionary about what's happening in the future. Now admittedly, these people tend to be people with a little less power, they have a little less control over budgets, they tend to have a little more time. It's almost an inverse relationship between those things. But that's kind of intentional because their job is to help the company see a little bit around the bend and do a little bit about foresight about

what's coming next. The goal here is really about surprise reduction. Surprise production and upside identification. And most of the companies that do this well don't try to predict the future. Instead, they engage in imagining several, multiple futures. This is sometimes called the scenario method. It was actually identified and developed at Shell Oil in the 1970s, which was grounded in essentially an intractable Type 2 problem. The people at Shell realized it was hopeless to try to predict the price of oil. Just couldn't do it. So instead what they tried to do was imagine a couple of worlds in which the price of oil might be really different than it is today and try to understand how we might get to those worlds. And most importantly, how you would know a little bit in advance of others if that world was coming about. And that developed into a discipline, really, that's now known as scenario thinking. And it really has a modest goal in a very serious way. The idea is if you can imagine alternative futures that stretch the plausible out just a little bit, and then tie those scenarios to leading indicators about how you would know they are coming, like in a world of this sort, we'd soon start to see this kind of thing or that kind of thing. Companies that can do this successfully will make themselves just a little less vulnerable to big surprises and a little more likely to see big discontinuities in their business environment. Here's a classic example. IBM is dealing with the transition from the mainframe to the PC. The critical scenario might have been, what if somebody links those PCs back together? What would that look like? Well, that would be called the internet, and interestingly it wasn't science fiction. At the time that IBM was thinking about this, TCP/IP was already in use in the ARPANET so they could have imagined it. So we're going to get a perspective on this issue. Now keep this in mind as you listen. This kind of thinking might seem, in a really serious way, countercultural for data scientists. Like it's not grounded in fact. But you know, it shouldn't be counter-cultural. It's never going to be the mainstay of the data scientist's work. In fact, it's going to be quite the opposite. But it's going to be part of your work because you better than anyone are going to understand what are the really big uncertainties that need to be in those scenarios. You're going to understand that better than anyone else inside or outside the organization.

13.4 Think Big, Part 2

13.4.1 Ask Questions, Challenge Assumptions, Ask Why?

This element addresses the following learning objectives of this course:

- LO6: Navigate organizational, personal, legal, and ethical constraints to facilitate better decision making and improve communication.
- LO7: Imagine, plan, and design a data science project.

Some of the most influential innovations started with questions that challenge assumptions. They asked why. Part of the reason this is the case is because if we don't challenge the status quo, nothing happens. We will only see change if we ask why do we do it this way or is there a different way to do it?

Now, I'm not giving you license to kind of just go around and disrupt things and turn over tables just for the sake of creating disruption. But I encourage you not to be complacent. We are uniquely situated to create change. We could support change, but we could also plant the seed that would become the next big idea.

13.4.2 Ask Pointed Questions

This element addresses the following learning objectives of this course:

- LO6: Navigate organizational, personal, legal, and ethical constraints to facilitate better decision making and improve communication.
- LO7: Imagine, plan, and design a data science project.

Some of the most influential innovations started with questions that challenge assumptions. They asked why. Part of the reason this is the case is because if we don't challenge the status quo, nothing happens. We will only see change if we ask why do we do it this way or is there a different way to do it?

Now, I'm not giving you license to kind of just go around and disrupt things and turn over tables just for the sake of creating disruption. But I encourage you not to be complacent. We are uniquely situated to create change. We could support change, but we could also plant the seed that would become the next big idea.

13.5 Technology and Decision Making,

Part 1

Interview with **Phil Nolan, IBM**

13.5.1 Expert Decisions, Technology, and Society

This element addresses the following learning objectives of this course:

- LO1: Describe the role of data science in organizations.
- LO4: Justify an analytic approach that informs decision making. ● LO5: Identify the audience and the most effective method to communicate a persuasive argument.

Lots of stories have been told in the past about expert systems making people who define themselves as experts feeling very unhappy, threatened, and displaced. I know that none of those personality characteristics or emotions could ever be felt by anyone inside the federal government. But let's imagine that it was possible that someone inside the government could feel that way. Have you run into any of those kinds of situations?

And without naming names or naming organizations, could you tell us-- oh, look at that-- anything about how they've evolved and what that feels like and what works and what doesn't work?

Sure. It's way more threatening in theory than in practice. What I discovered with Watson, a year and a half ago I hired three subject matter experts. These were analysts who had been working on a variety of topics for-- they had been recently retired, so like 25, 30 years. And they were the brightest and the very best I could come across.

And I had them working with a fairly raw Watson system. The performance was low. The UI was miserable. And they were initially antagonistic, aggressive. How come you got it wrong, Watson?

Then what we discovered is each one of them went through a curve, a slight conversion, a road to Damascus-- no, it was actually nothing as dramatic as that-- to realize that Watson was a slave, not a master. It was a tool. And at no point was it ever going to give you an answer that you could act on by itself.

Nobody plays Jeopardy. They were asking questions about what is the right treatment

for a healthy I'm going to go for a 47-year-old professor at Cal who does blah, blah, blah, blah? You know what, there's no single answer. It's really messy. And if you're a doctor, you're happy to come back with some of these intermediate ideas from Watson and then make your own judgment.

If you're trying to answer a national security question about is it a good idea to accept Putin's offer vis-a-vis serious chemical weapons, well, you know what, Watson's never going to give you an answer. It could bring back some really useful information that could allow you, the analyst, to become a rock star.

So what I've discovered is it starts off with kind of trepidation and resistance. But in fact, there's a quick realization that if I ask Watson the range of a missile and I say, what was the range of the missile that was launched over the mountains, that'll really confuse Watson, because now you're using the word range, which has to do with missile, and mountains, and they're both in the same sentence. And you're going to get something that's absolute garbage.

On the other hand, if you ask a question, which is, how was Martin Luther King-- what did Martin Luther King and Nelson Mandela have in common-- and I asked Watson this question a year and a half ago, when it was still raw, and it came back with a bunch of information about Gandhi and his writings, not that they were two men who won I think a Nobel Prize, they're were two African-American-- one was African, one was African-American, et cetera. But it was way smarter than I thought and at the same time way dumber if I asked about what happens when you shoot a missile over the mountains.

Interesting. Well, Phil, let's move on to just a couple last

questions. Sure.

When you think about the trajectory of this technology, could you talk about the stuff that worries you or excites you or both, first maybe from a business strategy perspective, and then if you think that there are any really significant sort of ethical/legal concerns that are starting to appear that people are starting to think about in a serious way as compared to a science-fiction-y way? What would a business strategist want to have in his sights? And what would a legal/ethical person want to have in her sights looking forward over the next couple of years?

That's a great question, Steve. I think the business strategist-- my analogy is when 10 or 15 years ago American companies realized that they had vast amounts of intellectual property that they could license and many of them discovered that a set of patents sitting on the shelf in fact could turn into a cash flow. I believe there are lots of

companies that are sitting on data that they have collected for one purpose or another, that they had no idea can be used to answer questions that they care about or quite possibly questions other people care about.

So I think there is a potential intellectual property equivalent explosion to be had. And this can be done in a way-- this can be done in a way that protects individuals' privacy. I don't see that it will be done in that way.

Let me flip over to the other side of big data. It's Watson and it's also many of the structured data sets. There's been a presumption of obscurity, in America at least, around individual behavior. And it's a presumption which is socially rooted but not technologically rooted anymore.

People who go on Facebook feel that they are only posting toward two or three of their friends. But in reality, they're posting everything to the world. And every single credit card transaction or phone call you've made or anything else like that you kind of assume is lost. But with big data, it's not.

And I believe that there is a vast tension that tends to appear when we have technical abilities that go beyond our social norms. So that is the big data problem. It will be interesting when it comes together if AT&T wants to sell your phone records to somebody who wants to market Viagra to you, wow, it may be a great business opportunity for AT&T, and in fact you will find it to be a huge violation of your personal privacy.

So what I have seen is that users quickly accept that-- they accept the technology a lot faster than you ever expect. And I think my belief is that socially, we will as individuals accept it faster than our legal structure will. So what I tend to believe is this leads not to a dystopian world, but one in which there are a number of different behaviors that we engage in that allow us to live with the kind of lack of obscurity-- maybe it's not lack of privacy, but lack of obscurity that we're used to.

So the lawyers win.

This is America. Lawyers always win.

13.6 Technology and Decision Making, Part 2

The power of the technology we've been talking about today leads me naturally to think

about a really important argument that was made by Bill Joy about a decade ago. And in fact, Bill Joy seems to turn up over and over again in stories about computer science, software engineering, and the technology industry generally. For those of you who don't know him or don't know who he was, Bill Joy actually started his career as a graduate student here at the Berkeley Computer Science Department. He was instrumental in the development of BSD Unix, and later, he went to Sun and was instrumental in the development of Solaris. In 2000-- he was actually at Sun at the time-- he wrote a really important and provocative article that showed up in Wired magazine. The title of the article was "Why the Future Doesn't Need Us." And it's worth a read, actually, if you haven't read it. But the core argument of the article is really very simple, and it basically says that there were a bunch of really dangerous technologies that human beings produced in the 20th century. But ultimately, they were limited in their scope. It takes a large government, it takes a lot of money, it takes a lot of planning-- it takes access to really expensive machines and some pretty hard to find raw materials to make a nuclear weapon. But most importantly, from Bill Joy's perspective, it takes a human being to make a decision every single time another weapon gets built. In other words, nuclear weapons don't build more nuclear weapons. So at least there's a point in the process where a human being is always intervening and making a decision about what to do next. What Bill Joy says and leads us to think about is that some of the technologies that we're talking about in the 21st century, well, we know they tend to be cheaper for people to be able to get. They're perfectly viable in the hands of smaller organizations or maybe even just individuals working in garages. This is one of the things we really like about data, about IT. The access to machinery is actually not a big deal for people who can get access to machine tools in a place like a tech shop. In fact, this is one of the things that most people like about the information technology revolution. We call it "democratization." And the presumption behind many of the people here in Silicon Valley is that this is absolutely essential and a critical part of what leads to real profound innovation. But here's the thing that Joy pointed to. Really what we ought to worry about, as well, is that some of these technologies can remake themselves. He called them "self-replicating technologies." The point was it doesn't have to be a human being deciding to make the next one. So an engineered virus is an obvious example. It'll just kind of spawn itself without any human intervention. People have talked about robots that can assemble other robots without a human being having to be in the middle of that-- software code that writes other software code. Now, today, a lot of these things seem like science fiction, at least if we're talking about any meaningful level of complexity, but possibly not for long. And that idea scares Bill Joy to pieces. And maybe it should be scary. Some people dismiss the argument in Wired magazine as a little bit sensationalistic or science fictiony or even downright crazy. And it probably didn't help that Bill Joy quoted in his article the Unabomber-- remember, Ted Kaczynski-- and his kind of anti-utopian technology fear. But others and I think we today need to take this argument as a really serious thing to think about. How much power do we really want to

endow technology with? In fact, many films, science fiction and others, are really good at experimenting with some of those ideas, and they have for a very long time. Let's consider two of those well-known films and take a look.

13.7 Is Technology Too Powerful?

13.7.1 Discussions Prompt

Spend five minutes on the following prompt:

Choose one position. In 2–3 sentences, defend the position you chose.

1. One should be concerned about how powerful technology is becoming. While technology has improved the lives of many, we should be aware of the potential for harm.

2. One should NOT be concerned about how powerful technology is becoming. Technology has improved the lives of many, and we should NOT be concerned about the potential for harm.

13.8 The Data Scientist's Role in Responsible Technology

13.8.1 Data Science: Technology and Our Responsibility

This element addresses the following learning objectives of this course:

- LO6: Navigate organizational, personal, legal, and ethical constraints to facilitate better decision making and improve communication.
- LO7: Imagine, plan, and design a data science project.

What role do data scientists have in the debate about the changing role of technology in society? I'm going to get a little bit on my soapbox here, so hold on. One could say that we're principally technical people, and it's really someone else's job to worry about this stuff.

But I don't think that's a responsible position to take. I think it's extremely important and maybe even necessary for data scientists to grapple with these issues. Because after

all, who else is qualified to understand the issues we're talking about?

People are going to spread information. People are also going to spread misinformation. And they're going to hype the subject. And there will be fear. Some of it will be unintentional by people who just don't understand the technology. And others will understand the technology but try to make people afraid for one reason or another. We miss an opportunity if we say that we'll let ethicists, lawyers, or politicians or other people deal with the social and political implications of technology.

But I argue that data scientists have to be spokespeople for this technology. Who else is going to represent this technology in an honest, grounded way? You don't get to stand on the sidelines of these debates.

And you for sure don't get to complain about others' ignorance or misrepresentation of the issues you care about unless you're out there trying to bring discipline and reality to the debate. This is true of data science. This is true of other technology. And it's true of most of what we care about in social life.

13.9 A Responsible and Ethical Data Scientist

13.9.1 Discussions Prompt

Spend five minutes on the following prompt.

Imagine you are in a job interview. One of your prospective colleagues asks:

What does it mean to be a responsible and ethical data scientist?

In a few sentences, share your response.

13.10 Remember Your Roots

13.10.1 Focus on The Core Skills

This element will address the following learning objectives of this course:

- LO1: Describe the role of data science in organizations.
- 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.
- LO5: Identify the audience and the most effective method to communicate a persuasive argument.
- LO6: Navigate organizational, personal, legal, and ethical constraints to facilitate better decision making and improve communication.
- LO 7: Imagine, plan, and design a data science project.

Let's revisit the example of the nonprofit group that wanted to increase turnout. They had the technical capacity to execute the project and send letters to hundreds of thousands of individuals, but they didn't have the correct design. As a result, I was not able to provide them with specific insight related to their intervention. I was able to articulate best practices for the next project, but that didn't help them with the current endeavor.

This example reminds us to focus on how to design a project that will give us the best chance of answering the question we care about. Yes, you'll need technical skills to execute the design, but the core skills are critical. A systematic approach can be the difference between a successful and unsuccessful project. These are not optional skills. They are necessary skills.

13.10.2 Technical Skills Necessary But Not Sufficient

This element will address the following learning objectives of this course:

- LO5: Identify the audience and the most effective method to communicate a persuasive argument.
- LO6: Navigate organizational, personal, legal, and ethical constraints to facilitate better decision making and improve communication.
- LO 7: Imagine, plan, and design a data science project.

There's no substitute for technical skills. However, you must also understand enough of the context around a problem to determine if you're asking the right question. The question motivates the design. Then we translate the data and analysis into narratives that people will remember. It's our responsibility to be effective communicators. The onus is on us to persuade, not on the audience to understand.

Remember to spend some time in the communication phases of your work. These are often the last steps in the research process. They are often rushed. But remember, this is the only part of the process that your audience sees. As you go out and change the

world, remember that with great power comes great responsibility. Use your data science superpowers for good. Do no harm.

13.11 Final Project Preparation

Use the remainder of your asynchronous time to finish up your project deliverables.