Week 1 Transcript (Updated June 2021)
If you require closed captioning for accommodation reasons, please contact the instructor and we will make the necessary accommodations.

1.1. What Is Data Science?
We'll get quite varying definitions of data science depending on the domain that you're in. But we really see it as a discipline that occupies a space between computer programming and Master's of Business Administration, or MBA. And the broad goal is to produce insights to improve decision making. And so let's talk about the broad role of data scientists, and also some of the skills that you'll need.

So broadly speaking, on the business side, we communicate with decision makers in our organization to help frame the question. And it's our responsibility to understand the ask. We often play the role of consultant here within our organization. So that's on the business side.

On the computer side, we often work with data that's developed as part of an existing business practice, or we work with the computer folks, the software and other engineers, to design, develop, and execute a data pipeline. In short, we have to work with both the business and computer people. And I think that we're uniquely situated in this middle ground. And this enables us to make great impact.

We wear many hats on the team. We might be the cognitive scientist or we might be the artist. So let's think pretty concretely about some of the skills we'll need to do our job well. We'll work with the business people to help them reframe their vague, poorly posed questions. We'll need to develop a system to collect and store data.

We'll use statistics to answer specific tactical and strategic decisions. For example, how can we put exciting content in front of people so that they'll remain engaged with our product? And we'll also build and deploy models that may answer a repeated question a million times-- or maybe we answer that question in a slightly different way each time, such as, what video or what song to recommend to Mike at a given time.

We'll communicate in a way that resonates with our nontechnically minded decision makers. They get paid for their leadership, it's our responsibility to communicate to them in a way that they understand, in a way that speaks to their level of technical expertise. We'll confront ethical questions, such as, I know quite a lot about our users and customers, is that OK? We'll deal with both structured data and we'll need to restructure it. And we'll deal with data that needs to be structured for the first time.

And so we'll talk more in other videos, but I think what differentiates a good from a great data scientist is their ability to ask the well-formulated actionable questions, and communicate that insight to key decision makers who are often nontechnical experts. In this course, we're going to

make an argument to do things in the quote, unquote, "right order." And you'll find out why I put that in quotes a little later.

That is, start with a question, be willing to spend some time on ideation, and then allow the question to guide you toward an appropriate analytical approach. It's so easy to jump to the data or the analytics. But if we fail to spend time thinking about the question and the appropriate design to answer that question, whatever answers or insight we gather won't be very useful.

1.1.2 What Does Data Science Mean for This Class?

We've spent a lot of time thinking about the rationale for including a course like this in a data science degree program, in fact, including it right at the beginning. And so we wanted to share some of that thinking with you so that you understand where we're going, why we're going there, and how we're going to integrate what we're going to learn in this course into all the other courses that are such a core part of this curriculum.

And so the first thing that everybody acknowledges and that we share is the undoubted situation about technical skills. Technical skills in data science today are really quite demanding, and they're going to become more demanding. And excellence in those technical skills has to be the foundation of any data science program, and it's the foundation of ours.

We've also designed those kinds of inputs and those kinds of courses so that the knowledge that you gain will be ready for constant updating because this is a field in constant motion. We all know that, and we're really aware of it. But our view and one of the core rationales for this course is the notion that all these technical skills are necessary but actually are not going to be sufficient and, in fact, probably aren't really even sufficient today for an effective data scientist.

What we're thinking is a great data scientist today, much of that really focuses on technical skills. So when you read about someone in the press or in the media, they've perhaps found some new algorithm or created something new that found some sort of insight. But sometimes, you often don't hear much more about it after that, and that's what we want to change.

We want to look at what your role is as a data scientist in an organization that is a much higher level. So you both know the technical skills, but you also operate on a strategic level-- so working with an organization making really big high-stakes decisions. And we'll talk a lot about that in this class. What are the decisions that organizations face, different types of ones?

And we won't be speaking abstractly. We'll bring in real companies, real experts, and so forth and go, what sort of questions did you face? How did you do it? And one of the things we'll look at, as well, at the strategic level is, how do you take the information that you find when you run a query or run an analysis and how do you convey that to others because that, we think, is one of the key strategic values we offer, as well.

It's not just finding that finding but actually, how do you convey that finding to others and persuade them to maybe take an action, maybe something that's risky? And that's a skill that we think you're going to need going forward.

So we want to share with you our ambition for this class and how it relates to the really high ambitions for the program as a whole, because our vision is that data science is not simply acting in the service of requests from others or directives from others. That may be the case in many companies today, but we see data science as moving right into the center of the company, right into the center of the organization, first as strategist, then into the C-suite, and ultimately and probably not very far in the future to the position of the chief executive officer.

And when we say that, we mean not just the chief executive officer of a data science company or of the kinds of companies that we think of as being data-native, digitally born, comfortable with the idea that data ought to inform all decisions. Well, I think a lot of those companies could be the ones that you know, where it's Facebook, Google, Yahoo, or anybody born in the internet realm, but also companies before that.

Think of insurance companies, where the pure existence of those companies and why they can stay in business is through people analyzing data. And Stephen and I will call those the "data-native firms." So whenever you hear that in the course, that's who we're thinking about, groups who create products and services where everything is really about analyzing and using data in different sorts of ways. And that's a big opportunity right now.

We think the bigger one, though, is really in those non-data-native firms, which is everybody else out there in the world--

The people who were sort of products of the previous Industrial Revolution, the companies that think of themselves as dropping heavy things on your foot or building big--

Ford Motor Company--

General Electric--

General Electric.

Imagine you're General Electric, and that's a non-data-native firm in many ways. General Electric's huge, but you think about they make airline engines. And how would you take the engines and gather data off of that and analyze it? And they do that with locomotives, as well. Those are the types of opportunities that we'll have.

In our view today, mostly, data science is being used by other decision makers in that setting. But it won't be long before the data scientists own the farm.

1.1.3. Effective Research Design Will Help You Make Meaningful Impact

The goal of this kind of discussion is, again, upfront, to make sure that you become the brilliant scientist who changes the world for the better, not the brilliant data scientist who nobody listens to. And let's acknowledge that sometimes in some of our work, the brilliant data scientist who changes the world for better can feel like you're involved in what might seem a slightly pedestrian project.

So I don't know, consider, for example, it's your job to reduce the cost of password resets in a hospital by 20%. And so you collect some interesting data. You look to see who is resetting their password, what time of day are they doing it, what devices are they doing it on, against like, for example, where they are in their shift. And you discover that if the system were to prompt people earlier in the shift for their password, you could get a 22% or 23% reduction in password resets.

Now that may not seem like a terrific scientific breakthrough, but actually, it could make a real difference in health care economics. And, in fact, that's the kind of work that many of us are going to find ourselves doing much of the time. So it's important to understand that you got to be the brilliant data scientist who makes a difference in the world. Sometimes it's not going to be the scientific breakthrough that we would hope for in other circumstances where it might be more exciting.

And so that kind of project, or I should say, project-based mind set for us imbues our teaching philosophy how we want to think about this class. We've given a lot of thought to that. We asked the question when we started, does this program, do the students in this program, need to come in and have a traditional 15-week class on standard research design and methodology, the way you would teach it in a PhD program, for example, in the social sciences. And we decided that actually that's probably not what we needed to do.

We need to do some elements of that. And, of course, a lot of research design questions get picked up in statistics courses, the way we think about selecting on the dependent variable or explaining why we do that, selection bias, too few equations, too many variables, don't over-fit. Many of those things are going to be covered in other classes, statistics, experiments, machine learning.

What we wanted to do instead was to think about the bigger frame for context. And so our question, really, for this class is how to design a research project that matters in an applied sense and in an efficient sense. And so I think one way to think about that, maybe in a more formal sense, is to imagine us trying to construct together almost an optimization problem, bringing together a point of view on the resources we can bring to the table, the capabilities that we're going to have in place, our ability to leverage action or decisions in the world against the importance of a question, and creating an optimized mix of all those ingredients so that we can ask better questions and find ways to answer them in a useful way that will change what

somebody else does in the world in a way that makes things better. Kind of that simple, and that's what our teaching philosophy is going to be for the semester.

We've thought about creating a sort of rough template for each week. We're going to try to give a map of what the week's going to look like. We're going to offer some examples. Sometimes that will be in video. Sometimes it will be in written story. Sometimes it might be in a short interview. We're going to try to extract some concepts, use some more examples, talk about some cases. We're going to engage in some discussion, maybe again, mix in some interviews with people who are actually trying to engage in those parts of problem solving and offer some closing thoughts.

I think it's really important for us, and actually a big part of the fun of the course is for us to be able to move across different domains. So some of them you'll be interested in, some of them you might not. We're going to talk about medicine, we're going to talk about sports, we're going to talk about manufacturing, we're going to talk about government, retail, security. And I hope that you'll bring in more of those kinds of domain examples during synchronous sessions, because, after all, comparing applications across domains can sometimes yield some really interesting findings.

Couple more points I want to make just to kind of frame up the way we're going to teach the course. We're going to be talking a lot about models. And when we talk about models in the social science world as compared to the technical world, sometimes people get uncomfortable with what they see as bizarre or obnoxious oversimplifications. Well, I'm going to defend that now and we'll come back to it later by simply saying we treat models almost exactly the same way scientists and technical people treat models.

We're looking for the variables that we think really matter the most, and we try to simplify them. Then we try to exaggerate them in the model so that you can see how they impact upon dependent variables or outcomes of concern. And then once we kind of get that picture in our heads of what's happening, we slowly add the complexity back in an organized way so that we can improve the richness of the model.

In other words, we're going to simplify, exaggerate, and then try to add the complexity back. And on occasion we're going to have to ask for your patience through that period of what may sometimes feel like oversimplification in order to get to that place where we feel like we've organized the complexity in a way we can do something about it.

So that's the first thing we wanted to talk about. There's a second which I think of as a trade off between stretch and relevance. When we're focused on decision making, and as many of us experience in our day-to-day jobs, people are asking us what's the takeaway. How do we make this immediately relevant to the problem that's on the table today?

And we often find that in research positions that are kind of a bit of a trade-off between the immediate relevance and stretch that means sometimes asking people to ask a slightly different question or to look beyond the problem of today to upframe it into a larger problem, which actually might be more significant but which we might not be able to solve right away. And so we're going to try to manage that trade-off between stretch and relevance and just at least be aware and mindful of where we are on that dimension at any given time.

That's the second point. The third thing we're going to do is we're going to ask you, and we're going to engage in a number of different thought experiments. And by thought experiments, I mean often asking ourselves what if this precious assumption that we've held on to is wrong, just completely wrong? That's the most common kind of thought experiment.

There's a second and sometimes even more interesting kind of thought experiment which is to ask ourselves, for example, what if this precious assumption is even more right than anyone thinks it is? For example, there are a lot of assumptions about how consumers behave in relationship to data and privacy. Some of it may turn out to be completely wrong, and some of which, interestingly, may turn out to be even more right than we think. We want to play with both ends of that continuum.

And then, finally, we want to make sure we're having fun. Because I really believe the work we're doing and the overall educational experience of a program like this, some revolutionary stuff, and I'm of the conviction that if we're not having fun along the way, we're probably not doing it right. So again, kind of a template for the course. We're going to work with models, we're going to experiment with the stretch relevance kind of continuum, we're going to do some thought experiments, and along the way, we better make sure we're having some fun.

1.1.4. **An Information School Perspective on Data Science**

Welcome. We are privileged to have with us, now, the Dean of the Information School, who was, really, the motivating force behind the development of this program. Her name is AnnaLee Saxenian, but those of us who know her, and now you know her, too, refer to her as Anno, or Dean Anno, as the case may be. Anno, thank you so much for joining us.

Yeah, it's a pleasure to be here.

So, Anno, thinking back over the trajectory of your career, you've spent a lot of time around the technology sector, and Silicon Valley, in particular. Cam you tell us a little bit about how your interests have evolved to now be sort of focused like a laser beam on the next phase of data science?

Well, my academic career started with an effort to understand the dynamism of the Silicon Valley, the Bay Area. This was in the late 1970s, early 80s. It was before anybody really had even thought about data science. It was on the heels of the microprocessor invention, it was a little bit before the PC off. And I, essentially, watched the region evolve over several decades.

And in comparison with the other regions in the country that I knew, I started to understand that the really remarkable pace of innovation and change in the Valley didn't have to do with the technical skills of the people there. There are people all over the world that had technical skills, and have technical skills. They have capital. They have universities, great universities, they have science parts. They of all the things that we think goes into innovation.

But with what happened in Silicon Valley was a set of social structures that allowed information to move very quickly between firms, and people to learn very quickly about organizational change, about technology, about business models, and whatnot. And I realized that the social and organizational aspects of the region were as important as the technical skills of the individuals, or the capital markets, or the universities. So, essentially, that informed my desire to work in the School of Information.

Most academic disciplines are narrowly focused. I have a degree in political science. There are computer science programs. There are math programs. There are economic programs. And what I have seen in Silicon Valley led me to think that we needed to educate people who could work across those domains. People understood the social world, understood managements, and understood the technology and how technology was evolving. And that's really what the School of Information has tried to do over the past couple of decades. And I think data science is simply the latest and newest, and I think one of the biggest, things that we are going to be able to engage as a school of information.

Now, If any of our students wanted to just go back and understand your big picture view of the region and how it works, the book they would read is Regional Advantage. I forgot what year it was published. It was--

1994. It mentioned the internet only one time in the conclusion.

But I suspect that people would still find it interesting, and this fundamental argument within it, quite relevant to how data science might unroll in the Valley.

It's pretty relevant to today, interestingly, although you have to substitute different company names and technologies. But I think the evolution of the PC industry, and then the mobile industry-- mobile devices-- and now, big data, data science fits right into that account.

I remember in Regional Advantage, you talk about the role of universities and so forth as being part of the ecosystem of Silicon Valley in this area, and Stanford, and Cal, and the other programs, and so forth, and the exchange and flow of information. By having this course and this program be online, how does that change the spread of what Silicon Valley can offer?

Well, it's wonderful. I think the thing that I love about it is that those-- in the previous years, universities have been very localized. I mean, to some extent, they were ivory towers. But in the

best cases in Silicon Valley, they opened up their boundaries a bit. This opens up our boundaries much more widely. Now, we can reach students anywhere in the world. And we educate students to work in Africa or in Asia, anywhere in the US. So this is a huge opportunity. And I think we're going to see the spread of data science accelerating because of the online technology.

So you have a history of coming to technological environments, or highly technical environments, with a kind of social science microscope, and using that, in a way, to kind of parse out what are the most important and most long lasting technical developments. And so you're placing a bet that data science is one of those now. Tell us a little bit more about that bet and how you see that evolving over the course of this program.

Great. Well, when I started studying the Bay Area, Silicon Valley, it was in the late 70s. The microprocessor had been invented about 10 years previously, less than 10 maybe. The PC was just on the verge of emerging. And so I've seen successive waves of technology. And one of the indicators of something that is going to be big is that it diffuses very quickly out of one particular sector across domains. And we've seen that happen very quickly in data science.

It started in social media companies, in so technology companies, especially technology companies. But very quickly, we're seeing it being picked up. First of all, finance has been doing it. But then it's been picked up in advertising and marketing. And you're seeing, now, that it's being picked up in health care and just a wide range of domains.

And I think there are many more that it will be picked up in that are slower for organizational or institutional reasons. You can think about education, or about government, which badly need, I think, analysis of their data to help improve their processes. But we'll take a while for people to start to really move into those sectors and overcome some of the barriers that exist.

So one last question on this. It still seems that you roughly, casually compare the speed of diffusion of the data science community as compared, to say, the speed of diffusion of the world wide web community in the early 1990s. It really feels like it's moving roughly twice or three times as fast to me. Does it seem like?

It's very fast. But I think each wave builds on the previous wave. So the speed of the web allows the next wave to be faster. The PC wasn't as fast as the web. The web now enables the data size to be even faster. And in between that, you've got the mobile device.

Right.

So these things, the sensor networks, all of these technologies are building on one another. So I think each one is faster. It's accelerating.

When looking at the data science program here at Berkeley, in constructing and crafting it, what was some of your perspectives on, what did we want to offer? What is Berkeley's unique perspective in this space? You mentioned from joining the iSchool and having a unique perspective. What is that perspective now for data science?

I think one of the things that I really tried to do-- and I built this program in conjunction with all of my faculty colleagues-- but I think we really wanted to make sure that our students had an understanding of the relevant technical skills. They need to understand algorithms and machine learning. They need to understand statistics and and they need to understand databases. But they also needed to understand the context in which those would be applied.

So they needed to understand the domain, but also how to ask the good questions of data. How to think about how decisions will be made with data. What are the broader policy contexts that will shape how data is being used. I think we are increasingly aware of the issues about privacy. And those issues are front and center, the issues of context, of decision making, of communication, and of policy are front and center of this program, alongside the harder skills of programming and of statistics. And I think what I learned from Silicon Valley is that that mix is very powerful in organizations.

It seems like the natural history of technologies in the Valley has often been that the technology leads and then we go back later and backfill all these other issues when they become urgent and necessary to actually moving forward and would it be appropriate to say we're hoping to short circuit that a little bit?

Exactly. That's exactly what we're trying to do. We'll see if it works, but you're right. Technology has always outpaced our social arrangements, our institutions, our law. And we hope that we can start to build those things in from the beginning.

And where do you see graduates working following their degree?

My hope is that graduates will be found in any organization from your nonprofit to the public sector to international organizations to banking, health care. I think any organization that has data will want to be working within and trying to understand, see what it can help them learn about their own organization and how they might provide services better, or define find new products, or make decisions better based on their data.

We're going to introduce, later in the course, a distinction between what we call roughly date native companies-- those that are sort of born digital-- and non-data native companies, let's say sort of main line, 18th century industrial firms into which the data world is incorporated almost like a new limb or something like that. Do you see a distinction between that in the way our graduates are going to move, or are we trying to bring those things together? How are you viewing that?

That's a good question. I think the data native companies are at the leading edge of all of these techniques of developing the new algorithms. And I think that some of our students may well move into those places. But I think the places where there's the most growth, the most potential growth in the future, is all of the other companies in the economy and organizations that don't natively know how to integrate it, and are not really even organized to do it. Their data may be in one part of the organization, the engineers may be somewhere else, the decision maker somewhere else. They may not know how to make use of their data. So I'm thinking that our students will actually help transform these old line companies to move them into the world of data-intensive strategy.

Yeah, presumably that requires all the data technical skills, but also a strategic mindset that's able to use those skills, not even in the service of solving particular data problems, but in the service of actually transforming the organization to be data-friendly.

Absolutely. I mean, I think it is a bigger challenge in a way. It's a more multifaceted challenge about thinking about how organizations are going to evolve in this new world.

I guess the one thing we can say for sure is that if the data scientists don't take on that task, someone else will, and they probably won't do it very well.

Absolutely. Good point.

Some of the things that we've talked about in this class, and you've shared with us, is looking at the responsibilities for data scientists. And there's your responsibilities within your organization, maybe within your industry of using data, and so forth, in different ways. For this first group of people, it's almost they're like pioneers of data science. What kind of opportunities are them for maybe being like an ambassador of data science outside of their industry or their domain, and things like that. What are other ways for people to get involved in being an ambassador for data?

Well I think the first way, actually, is to be a role model for the broader, more well-rounded data scientists who understands context and decisions and whatnot. I think getting involved in the broader conversation about the use of data, about how data can contribute to good decision making, but also to understanding the potential pitfalls. I think there's a lot of potential for a backlash against data and data scientists if it's misused. And so I think at the same time that we want our graduates to be role models, we want them to be understanding those ethical dilemmas and models of how to address the broader societal and organizational issues, as well as their own domain and organizational issues.

Yeah. One of the things we've mentioned a little bit in the class is our responsibilities and how do we engage with regulatory bodies, things like that. Because there is, understandably, justified fears about data, how it can be misused and things like that. Can you talk me about how an individual can maybe get involved in say the state, or local, or government level-- federal level--

for being able to ensure that these sorts of methods in this field can continue to grow going forward? I'm thinking of the Federal Trade Commission, things like that. What are those various roles you see out there for people to be advocates?

Well, I think I think almost at any level, starting, really, from the firm and the community level, informing their colleagues, starting the conversation in a direction that understands the potential and also understands how to think about the pitfalls and avoid them. I think it could be writing op-ed pieces. It could be joining local organizations. It could go all the way up to the FTC. But I think right now, mobilizing at a very grassroots level is actually quite important.

There's a lot of anxiety. I think about my parents and what they read about data, and it sounds scary. They're collecting all this data about us. So I think we need to help them understand that there's real value that can be gained from the data, but also to understand that the people that are working with this data are considered, and have thought, and will continue to think about the issues that they worry about.

Now, I don't want to lose the thread that in the midst of this, of course, it's necessary for all sorts of different reasons, almost like as a baseline capability, for our graduates to have the best set of technical skills that they can possibly gain at this moment, along with the recognition, as you've said, that these skills are going to be evolving very, very quickly. So how do you think about that sort of upgrading, that ability to upgrade over time?

I think that's a really important question because we've seen already, I mean in the short life of this field, a lot of evolution. A lot of new tools are coming out-- new databases, new storage mechanisms, whatnot. So I think our students are going to have to assume that they will be continuing to learn as they go along. And I think that means participating in the broader community, professional communities, where education happens. Hopefully they'll stay in touch with one another and they will have an alumni community that can help. But I think at every level, this is moving very quickly. And anybody who doesn't stay on top of those skills will probably fall behind.

It's been an interesting challenge, I think, for our colleagues in developing classes to, again, transfer state of the art knowledge for today, but really keeping a core part of their agenda. How do we make sure that people who have taken these classes are prepared to upgrade and adapt as quickly as they possibly can as the technology moves forward? And we're not teaching a static subject here. It's really quite--

No. I mean, I'll give you an example. When I became Dean of the School of Information in 2000, we really were not using mobile devices at all, as mobile internet. And so we didn't have any curriculum on that. So we've had to continue to upgrade the curriculum, and to change it to bring in new faculty, to bring in new concentrations. And I think we're going to do the same with data science. And our alumni are going to need to find it. I certainly think this program will try to

provide some of the material to help, but we'll see institutions evolving to help people keep up-to-date with these things.

So our final question. We'd like you to share, if you're willing, a little bit about your personal passion in this subject. People don't come to this purely for intellectual reasons. Why do you really, truly care so much?

It's a great question. I think, as I've moved through my career, I see myself now more as an institution builder, and as an educator, more broadly, than I did previously. And I really want to see us educating the kind of students that can really make a difference and be responsible contributors to society more broadly.

And I'm already doing it. The school is already educating those people. But I think, with data science, we're going to make a global impact. We're going to make an impact beyond just the San Francisco Bay Area. And I really would like to see students who have this mix of skills that will allow them to be more effective ambassadors of data science and then of Berkeley, more broadly.

Great. Well, thank you so much for this introduction. And we'll look forward to seeing you later on in the program.

Great. Thank you.

1.4. Where do you start?

So this segment, we want to ask a big question and at least start to construct a few ideas about an answer. The question is what does data science mean to the world in the context of this class? Well, it depends upon the question you're asking. So again, let's turn to a concrete example around something like climate change.

And imagine for the sake of discussion that you've just been given a one month sabbatical to use your new skills in the data world to come up with a new proposal for mitigating climate change. And the question that I want you to think about for this segment is how do you start that month? What do you do on the first day, the very first day?

So maybe you want to start by focusing on the question that you actually want to ask. Because, look, you've only got a month to do the project, so you can't afford to go to down too many blind ends. So you could ask, for example, who are the worst polluters in the world? And you could bring data to bear on that question.

Of course, you could also ask who can reduce carbon dioxide emissions most cheaply and efficiently? That's what you would want to know if you were building a carbon exchange market or what sometimes people call cap and trade. You might want to ask who, right now, has the worst cost benefit ratio from CO2 emissions? You might want to ask who in the world is going to

suffer the most irrecoverable harm from climate change? There's probably 10 other questions you could ask, and you can't answer them all in the course of your month sabbatical.

So the point, really, here is to reflect on how would you start constructing your own data science enterprise from the very start to the very end in order to be most effective in the world? Because ultimately, that's what you care about here. And I think this kind of question overlaps with the big, broad, abstracted desire to try to have a point of view and develop a point of view over the course of this semester on what data science actually does and what it can mean to the world.

For me, the meta question that embeds that is how do you decide what question you want to ask and answer with these powerful tools? And in fact, a lot of this class, a lot of the semester is going to be spent thinking about what are the various ideas, arguments, and tools that when used together, can help people to frame and point to a better question? In other words, how do we choose better questions?

And in the context of that, we should recognize, look, we're just getting started. We're really so far away, all of us, right now from a truly data-enabled society that it may be that choosing the right questions to answer is as powerful, if not more powerful, than getting the answers precisely right. A lot of people have opinions on where we are and what that means, some informed, some uninformed. And as we've said before, those opinions are made up of expectations and fears, hopes, desires.

But I think most importantly, and interesting for now-- not a little desire. Important to recognize-- not a little desire to look for magic bullets for problems that have seemed unsolvable up till now. But that's actually not what most of us are going to be in a position to be able to create.

1.6. Behavioral Science and Critical Thinking

Let's outline some of the critical skills we'll cover in this course, theories of decision making, political and bureaucratic power structures in organizations, talk about behavior science and how the brain works, incentives, cognitive biases, logical fallacies, how to persuade and tell compelling narratives. These are going to be all skills that we're going to cover throughout the course. We'll talk about how to communicate with a variety of audiences in both written and verbal form.

We'll think quite systematically through the research process. This is one of the major takeaways of the course is to slow down and to be very systematic and deliberate with our thinking. We want to craft carefully-thought-of research questions and design projects that will give us the best shot at answering the question that we started with.

We will cover many of the ethics that guide high-level decision making but also that guide the decision process. For example, ethics come into play not only when we're thinking about the go or no-go decision. But throughout the process when we're thinking about different modeling and

sampling strategies and sampling decisions, there will be some ethical questions that will surely come up.

Now, some people may call these soft skills, but we really kind of think of this framework as somewhat pejorative. It doesn't really capture the value added of thoughtful question formation and careful research design. A systematic approach can make all the difference in a project. These are not optional soft skills. These are critical, even though you may not be able to express the ideas in a formal equation.

So let's explore the importance of careful thinking by exploring a cautionary tale. And so let's go through a quick real life example that I engaged with of what can happen when we don't spend time on a thoughtful design, and we ignore decisions-- or we ignore how decisions are made within a organization.

So let's imagine you-- or again, I'm describing a situation that I was intimately involved in. Let's imagine that you care about voter turnout. And you work for an organization that has tools to increase the vote and to engage its volunteers. They designed one intervention to try to do both, that is, to both increase the vote and engage its volunteers.

But instead, they didn't do either one very well. They developed an experimental intervention, a letter-writing campaign. And in an effort to engage volunteers, they had them write personalized letters to voters. Now, again, this was aimed to both increase the turnout.

But because they encouraged individual volunteers to write personal letters to each voter, this introduced variation in the treatment in a nonsystematic way. And as a result of this issue and other issues that we'll revisit throughout the course, they were not able to determine if the intervention worked.

So I came in a little later in the project. I'm not deflecting blame, but I came a little later in the project. And I played the role of cleanup crew. I could have said, well, you completely messed this up. We can't salvage this. But instead, I had to be mindful of the way that I communicated with the group.

Because even if I didn't work with this group in the future, the community is a lot smaller than you think. And it's important to keep that in mind as you work with different clients. Be deliberate with your thinking when you design a project. That's the punch line. Why? Because it'll save time. It'll save money. It'll save headache.

It'll also help with replicability if you or your client want to reproduce this effort a carefully-thought-out design will help you do that. And if necessary, you'll be able to justify your approach to your clients or your manager if they ask you about specific decisions as to why you did something a particular way versus another way.

1.8.1 Ensure Impact

So we often need to think ahead to make sure that we're going to ensure that our projects have impact. How might we change our behavior? Or how might we encourage our clients to change their behavior based on the anticipated effect?

We want to be careful that we don't want our priors to unintentionally color the decisions we make throughout the process. And we'll talk at some other point about theory and how that might guide our projects. But we want to think about best case scenario. If this worked exactly how we wanted it to, how might people change their behavior?

Also, what's the standard of proof that's going to get our clients to change their mind? And if we could get our intended audience to commit beforehand that if we show them certain evidence, then we could change their mind, then we'll be golden. So this is a really big ask. But if you can get your clients to agree to a standard of proof beforehand, you're going to be in really good shape.

You generally want to place emphasis on how a project can influence a decision given the motivations and desired insight of a client. In other words, how do we frame a decision in a way that our client will listen?

It's important also to remember that we want to design a project that has a good chance of getting this kind of desired insight in the timeline that is provided. We want to avoid coming up with a plan that's a little bit too ambitious that gets us closer to the, quote unquote, "truth" but that takes too long for it to be actionable by your decision maker.

As you design your project and as you get into the analytics, think about impact. Think about significance. But think about significance both in a statistical sense and in a substantive sense. Something can be statistically significant, but it may not have substantive meaning.

1.8.2 **Data-Native and Non-Data-Native Organizations**

Let's talk about decision making and how to deploy a data science approach in different types of organizations. Let's think about a big bohemoth company, it's been around for decades, that makes a physical product, let's say a widget. And then let's think about a company that was founded in the past 10 years and their product is software. Think of a data native music platform or video platform that collects data every time a customer clicks versus this widget company that does track its production and distribution but it doesn't have a good idea of how customers use the item once they take it home.

A data native organization built from the start with a data backbone and that makes every decision with fine grained data compared to an organization that doesn't have this kind of data driven approach is different in at least two ways. One, the orgs will have different abilities to generate, process, and store data. And two, they will have different cultures about how data drives decisions.

And why does this matter? The more we know about how decisions are made, the better chance we have to make an impact. If we want to innovate in a company that has less experience with data science approaches, we can't expect to just come in with a data drives everything mentality and expect to change the way that things are done.

1.10.1. How to Create the "Best" Research Design, in Theory

How do you design a project so that you have the best chance to produce meaningful, actionable insight? We'll spend quite a bit of time talking about this in class. But I think it's good to start early. Here's some of the questions you should ask yourself when you're at the beginning stages of planning. I'll cover the core components of how to create the best design in theory. And then in another video, we'll cover how to take things in a more practical approach. So what are some of the questions you want ask yourself?

One, what's the business problem? Two, what's the research question? Three, what are your expectations about how things work i.e. theory? Four, based on your theoretical expectations about the question, what's the best data to use? Five, what's your analytical method? Will my intended audience be familiar with this form of analysis?

Six, do I want to engage in descriptive, predictive, or explanatory research? And finally, how will I communicate my findings? This is simple, right? But perhaps it's not that simple. And we'll unpack each of these steps in this class.

1.10.2 How to Create the "Best" Research Design, in Practice

How we develop the ideal design in theory is quite distinct from how we will do that in practice. When we articulate the business problem, we need to articulate why the status quo is unacceptable, why should we expend time and effort to change things, especially when we know that sometimes we get punished for sins of commission more than sins of omission. In other words, often easier to do nothing.

As we develop the research question, we should explain why the project matters. This is the, quote unquote, "so what" that we'll revisit throughout the course. Is the idea answerable in the available timeframe and budget? The best design, the best project might takes six months. But guess what? You only have three. Make sure you make it pretty dang concrete that you can do whatever you plan to do given the time constraints that you experience. Let's talk a little bit about theory.

Is our expectation about how things work in line with the dominant paradigm? We'll talk more about this in other parts of the class. We'll dive deep into paradigms. But in brief, do you anticipate that you'll get pushback based on the way you've set up the problem and based on the way that you've set up your expectations?

What about data? What's the best data to use? Will you analyze the entire universe of data? Or will you need to think about sampling? Will you use off-the-shelf data? Or will you collect new data? Is the pain worth the gain if you're going to collect new data? If you repurpose data, what was the original intent behind the collection?

A strategy that collects a lot of information may allow greater insight, but that may also raise privacy concern. And it might be more difficult to extract signal from the noise. Granular data might also be more precise, but again, we might have to deal with more noise.

What about method? What's the best way to approach the problem? Do we want to conduct a descriptive, predictive, or explanatory project? Will we use a method that your prospective audience will be familiar with? Do you think a black box approach is sufficient?

Do we only care about prediction? Or will your client or audience really want to understand the mechanism? Maybe you want to survey clients or customers. And maybe we think that more contact may help us better gauge their satisfaction. This approach could also backfire if you constantly ask for feedback.

Think about how you're going to communicate findings. Is your primary audience technical or nontechnical? Even if your audience is technical, will you be able to communicate findings to a nontechnical decision maker if they jump on the call unexpectedly? That's a really important skill to practice.

And so the punch line of all this is that we want to do research for a purpose. We want to make sure that our design is going to help us answer the question that we care about, want to make sure that our setup will give us actionable insight. We want to develop a design for a specific purpose.

Make sure the goal is clear. Otherwise, it's wasted effort. So this approach that I've laid out might help you design projects in your current organization, or maybe knowledge of this approach will help you get your next job.

Let me talk about an example that happened recently. A former student came back to me and said that they applied for a job. And they got past the initial interview process. And they were given a prompt to prepare for the second interview.

The prompt said, quote, "design an empirical study that outlines your hypothesis, desired data company, desired company data, analysis approach, expected findings, and potential limitations," unquote. Guess what? This class prepared them for that task. And the goal after this class is that you too will be able to design a similar study.

1.11 Define Your Terms

It's really important that we are explicit with our terminology. And it's important that we define our terms. Often, we do use terms that we may think our audience should know, and we skip definitions. Or maybe we use colloquial terms and technical ways.

In both these scenarios, it's important to make our definitions explicit. So let's imagine that you're presenting to leadership at a major institution, a major university. You were tasked to identify ways to increase student outcomes on campus. And your focus was on retention, graduation, and time to degree challenges.

Many of folks in the room likely understand the definitions of these terms. But there's a chance that someone does not. Or you got to realize that these folks have their attention occupied by 100 different things. And so it's important to get your audience up to speed as quickly as possible.

Let's imagine that someone in your audience doesn't one of these terms. Best case scenario, they raise their hand. They ask you to clarify, and we're good to go. But another possibility is that they don't want to reveal their confusion.

Or they may be uncomfortable raising their hand because they feel that they will reveal their lack of knowledge. And instead, they disengage because it's not clear on how using these terms. If your audience is confused and disengages, we've lost our opportunity to communicate and persuade.

So one solution is to define these terms at the beginning of the talk. You can say something like I'll use these three terms throughout the talk. And I'll start with a few definitions to make sure we're all on the same page.

You can also say something like I know we're all experts in this domain, but it's my job to make sure I'm being clear. And so I'll start with a few definitions. So I'm going to role play for just a second.

So I might say something like OK, the focus today is on freshmen entrance. And for some metrics, I will talk about transfer students. And I'll make that switch clear when I talk about transfer students.

So let's talk about some of the metrics I'll use for freshmen. When I refer to retention rate among freshmen, I mean how many freshmen who enroll in year 1 that return in year 2. When I referred to the graduation rate, I mean of new freshmen in a given cohort who graduate in six years. And finally, when I refer to median time to a degree, I mean, the median time it takes a freshman to earn their degree.

And a separate but related challenge is when you discuss interrelated terms. You've got to make sure that it's clear when you're shifting from one term to the other. In this example, it could be a challenge for your audience to track between these different metrics that you're focusing on.

So one approach would be to report all on retention. So you're like here are all the retention metrics. Here are all the graduation metrics. And here are all the time to degree metrics.

But you might want to compare these three concepts at the same time. Maybe you want to look at a particular cohort and look at how these different metrics compare. Or maybe you want to look at how these metrics have changed over time.

And so if you are looking at these related metrics at the same time, be careful when you bounce back and forth between concepts. You might say something like OK, now let's shift and talk about graduation rate. And again, by graduation rate I'm referring to the percent of new freshmen in a given cohort who graduate in six years or fewer.

We'll spend more time elsewhere talking about communication. But one of the key takeaways is that the onus, the responsibility is on us to communicate effectively. It's our responsibility to be clear. If we assume the audience knows what we mean, they may get it wrong.

All right. Now, let's shift gears real quick and just pretend that instead of studying retention, graduation rate, and time to degree, let's pretend instead you're giving a talk on a university campus about bias. It's another example of how to define our terms. Bias is one of my favorite words that emphasizes the need to define our terms because bias, in this example that I'm going to go through, means at least six different things.

And so I think, especially when we're having conversations surrounding ethics or just kind of data science in general, we'll use the term bias. And if we're not explicit about exactly what we mean, our audience could have a bunch of different definitions. So let's go through some of the different examples of bias.

One could be stereotypes and differential treatment of people based on some kind of physical characteristic or based on belonging to a certain group. That's one. Two could be sampling or modeling or model bias. Maybe that are-- we don't have much faith in our, let's say, results because the way that subjects or observations were selected into our study makes our sample or our findings non-representative.

A third way is algorithmic bias. Maybe our models were trained on data that reinforces some of these systematic issues that I talked about when I mentioned stereotypes and differential treatment of folks. Four, sometimes people say things are biased when they simply don't like it. Generally, the word has a negative connotation. So to claim something is biased can kind of tarnish the idea of conversation.

Five, you might say that folks have a biased perspective. Let's say that you are looking at research from the Heritage Foundation, which is a conservative public policy research organization. Or you're looking at research from the Center for American Progress, which is a liberal public policy research organization.

As we'll talk about throughout this course, the perspective-- an individual perspective or a paradigm could influence the type of questions we ask and our approach to the research. And sixth and finally, cognitive biases. These are information processing errors or mental shortcuts.

So as you can see, in this word bias, we've defined it right now in six different ways. And so if you go in front of an audience and say something about bias or most terms, just be very precise and explicit about what you're talking about. Otherwise, it could lead to some confusion.

1.12. Creswell and Creswell Textbook

This will help you digest the required Creswell and Creswell reading. Please use these ideas as a starting point. Focus on the writing ideas in the chapter; we will directly engage the rest of the concepts in subsequent weeks.

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.

Here, I'll talk briefly about writing and about ethics. Let's first talk about writing strategies. First, I want to acknowledge that writing is difficult. It takes a lot of drafts, and I personally find it pretty challenging.

I think one of the most useful guiding principles is to know your audience. And you've probably heard this, or you probably will hear this throughout the semester. What does that mean? That means who are you writing for, what's their incentive structure, and what are their motivations? What type of language would resonate with them the most?

Really, your approach to writing depends on the task. You're going to use a different approach if you're writing an email or a Slack message or something similar compared to if you're writing a final report for a client. And again, I acknowledge that the advice I'd give one or give you depends on the task. But here are a few pointers if you're writing something of somewhat importance-- i.e., not an email.

So first, you want to build time to step away from your writing. Why do you need to step away? Sometimes you get tunnel vision, and we can't identify our own errors because we've been staring at the same thing for so long. And that might mean stepping away for a day or maybe just stepping away for 30 minutes. So that's the first piece of advice I'd give.

The second piece of advice I'd give is to get someone else to look at your work. We know when we reread our work-- we know what we mean to say, and we kind of struggle to separate what's on the page from what's in our head. And so get someone else to read it that doesn't have all that kind of additional context and who doesn't know exactly what's going on in your head.

Second big idea is ethics. So Creswell pushes us to think about the many ethical issues we may confront in our work. We might confront ethical issues prior to conducting the study, at the beginning of the study, when we collect data, when we analyze data, when we share, report, and store data. This is the kind of framework that Creswell uses.

And I think there's a nice chart in the text that highlights the different types of ethical issues we may experience. And as you read this part of the chapter, I encourage you to think and engage with the following question. How often do we disclose the intent of our research or our analysis to participants or subjects or customers?

This is a challenge that we'll continue to address, because as researchers or data science practitioners, we may want to be transparent and ethical. And at the same time, we may want to avoid revealing so much to our subject that we unintentionally impact the results of our study. For example, imagine I have to choose how I want to inform subjects about the survey in front of them. I can say something like, "Here, we invite you to participate in a study to better understand how customers use our product." Or I can also say, "We want you to participate in a study to determine how appeals to emotions impact our customers' engagement with our product."

The second might be more transparent and honest. But it also might change the behavior of our customers because they were primed to think about emotion. And their behavior may be different because of this and may not be as generalizable to users on your platform that are not primed to think about emotion. So overall, I'd encourage you to think about how you disclose the intent of your research, and just be mindful of that.