# Video #0:

## Slide #1:

I'm Scott Alberts, and I'll be walking you through this module:

PDAT 625G Big Data Security and Ethics -   
 Module 7: Legal considerations for

Big Data Security and Ethics.

For the most part we're going to be learning from YouTube videos that cover these topics, I think, having lawyers and security experts talk about it is maybe better than having a statistician talk about it, although each of these are from for-profit sites, so they're trying to convince you to hire them. I will add some context here, and share some things I learned from some friends of mine who work in the industry.

## Slide #2:

The broad concept to consider here is that we tend to think about these four terms as somewhat independent goals, each worthy of time and attention. But, do they always point in the same direction? Do they sometimes pull in opposite ways that mean that we have to make hard choices? What's their Venn diagram look like?

It's somewhat comforting to know that most ethical things are also legal, but of course, we know that the converse is not always true. Similarly security and privacy often go hand in hand, but both of them push against our broader goal as data scientists, having access to data that allows us to do what we need to do, and ideas like reproducibility and replication.

With small data, we have some sense of literal cloud privacy, because we keep things off of the could completely. I can have an excel file on my flash drive or external hard drive, analyze it in R on my laptop with the WiFi turned off, and take it home with me when I'm done, or put it into a safe.

But, with Big data, we can't really do that, when we're talking about gigantic amounts of data on multiple servers with multiple processors, it's a bit harder to unplug it from the internet completely, even though there are times that might be exactly what you want to do.

Many companies to have an intranet that doesn't connect to the internet, but those are usually in specialty areas, like factories that use computers to control their production, or hospitals keeping their literal life support systems on a mini-grid not connected to the rest of their own network.

Prior to the mid-2010s, there really wasn't a common standard for privacy. Some laws from the days of telephones were all we had, or from the dawn of the internet. Specialty areas made their own, a few companies did, and individual privacy was not something that we collectively worried about. The wild west, indeed.

Starting in 2016, the GDPR was a European standard that really revolutionized the field, and nowadays, we talk about most every other standard in terms of how it compares to the GDPR.

Like we've done throughout this course, we end by swinging back to why this matters to you, a data scientist. You probably aren't managing your company's firewall or their VPN, but you very well might be at the table for a project large or small, so you need to know enough that you can be part of the conversation.

Again, our interest is sometimes pointing away from the data engineer - how can we get access to the data we need to make our model, to make the best decisions we can?

Still, we do want to protect our data, and data that's been entrusted to us - we're a kind of steward that has to take care of things in our care, not just because of the legal responsibilities, but because our ethical sense of trying to do the right thing as often as we can. "Stupid sinks more ships than evil."

### Slide #3:

We're going to start by talking about a standard called the General Data Protection Regulation, the GDPR, and this is the European Union standard, passed in 2016 and enacted in 2018. In general, people think it's the best one out there -- Again, not perfect, but that it really balances protections for users with what's reasonable and possible for companies to actually do, which clear expectations and consequences.

There's a YouTube video that we'll watch from a british company to explain it to us. In fact, speaking of the UK, even with Brexit, the UK maintained the GDPR standard almost totally intact.

Federally, there aren't any data privacy standards as broad as GDPR, but there are specific ones that you probably already know about: HIPAA for health, FERPA for educational records, GLBA for financial data, and more. Sometimes they are more strict than GDPR, but because GDPR was meant to be a global standard, and is decades newer, following GDPR probably means that you are following the more specific ones.

Some States set their own requirements. Our own state of Missouri state has no specific standards for data privacy, other than what you have to do if there's a data breach, so if someone breaks into your system and gets access to your data, there's a law that covers who you have to tell, how quickly, and how you tell the people who are affected by the data breach.

California, though, has the toughest state rules. It's loosely based on the GDPR, although a little bit looser. Some other states are making standards based on the California standard. Those of us have a certain age remember the auto emission standards -- that California had tougher emission standards than other states, and states could choose between the federal standard or the California standard.

In our materials for this module is a chart that highlights the differences between them

Your company, especially if it does business in multiple states might decide to just go ahead and follow the California standard, since you'll automatically meet all other state standards. Or, you might decide to follow the GDPR, since then you'll be fine anywhere in the world. Many big multinational companies just do that, although a few split their standards for work in different countries.

## Slide # 4

Two broad principles:

1) Data Protection by Design is the idea that you should think about data privacy standards at the beginning of your process and throughout your design. You're intentional about what you do, not reactive to try to minimally meet a bar.

You don't want to set up your system and then figure out how to make it secure afterwards, how to make it private afterwards -- you want to start with a secure system.

2) Data Protection by Default - that's the idea that unless there's a reason to give someone access to private data, you should have it secure. Individuals at your company should only have access to the minimum they need to do their job.

This bugs those of us who are statisticians because, in general, we think data wants to be free and that by sharing as much information as we can, we can make the best decisions and the best models. But, when it's private data, you should be intentional and purposeful with you has access to the data, and ideally, no one has access to everything. If you do it correctly, there's not even a way that hacker could get a lot of secure data.

3) A third principle is that, because your work is probably interconnected with vendors and clients, you want to think about how the data you steward is protected by them. You might be liable for their mistakes.

This privileges large companies.

If you sell through an online marketplace, you can sell things through Amazon. Well, you know Amazon might be using your data for nefarious purposes, but as a site, they're probably more secure and legal than anything you might build yourself, and if you aren't big enough to want to think about that, you probably aren't big enough to sell things on a locally created site.

The same is true for Google, Indeed.com, monster.com and other larger companies are probably going to make you less liable than you building it yourself. That's one less thing for you to worry about. They follow the California or European standards, as well as the financial and other industry standards, so you don't have to.

## Slide #5

Contrary to popular belief, most companies don't actually want your private information, and not just because they could get into trouble with the EU or California. They want something much more granular - signal convergence. This is the idea that they can track an (anonymous) user across sites to see how you went from start to finish, from "here's an idea" to a click to a "click here to pay." This is what makes their A/B testing work to see which marketing ideas have legs, this is how they can see connections between different products, "people who donate to environmental charities also watch *Family Guy* overwhelmingly voted for Hilary Clinton in 2016"

Originally, this was done by putting a tiny transparent photo on a web page that required connection to a remote server, which is why it was called a Pixel. It's more commonly now called a "web beacon," and it's where the action is. We'll watch a short video with someone from Facebook explaining how this worked before GPDR and CCLA and how it's changing today.

## Slide #6

Scott Alberts (he/his): As we conclude this module we're going to get to this idea of what you should do and really, you as a data scientist, as a person with a master's degree from this program -- you're probably not going to be your corporate data security guy for your company. That's not what we're teaching here.

But you might want to ask: Do we have someone who does think about this? Who is responsible? If no one else is, then it might be you whether you want it to be you or not. One specific point in both GDPR and CCAP is that you have a point person for security, whether it's their whole job, an outside vendor/consultant, or, for a smaller company, it's part of their larger job.

Beyond that, assuming, you do have somebody that's either part of their job or their whole job or maybe there's an office of people, depending on how big your company is, how do you work with that person or that office. You need to make sure that you have access to the data that you need to do your job. Again, we tend to think about how open we can make data. What can we publish? What can we share?

Once we get the data, what is that security person going to expect us to do? How do we set up our own processes?

The whole idea of reproducibility is the idea that someone else could recreate our analysis and get the same results. Or, replicate our analysis with different data. How can we do that if we want to maintain privacy?

Again, you're not going to be the decider of that, most likely, but how are you part of the discussion about that, and how can you be inside the "need to know" circle when you "need to know."

At the very least, you probably need to know about firewalls, about VPN, about these common tools meant to keep things private. Other kinds of secure connections make it a lot easier to be secure, but we're not really the people who think about how hardware decisions are made or how.

We connect through these different secure servers. So how do we know enough that we can be in the conversation, and how we can be helpful and useful as our companies are trying to make those decisions?

The last idea is DPaaS, "Data Privacy as a Service" Maybe your company needs to hire someone else to do this for you. You can also host the servers yourselves and hire a company to manage them for you. A couple of those companies are who made the videos we're watching this week, so you might take note of when they switch from supplying information into offering a commercial.

If your data lives in the cloud, you already are, even if that's just Google Drive or Amazon Web Services. It's only one more step (and one more fee) for them to help you with your end of the interface. One example, Yelp, has no servers of it's own - everything lives on AWS and they pay "per-click." They have to think about security on their end, but once the data is in AWS, that's part of what they pay for.

Each of their nodes is kept very separately, so if they broke into yelp they probably aren't going to get access to your data, because each of their sub sites are kept secure with firewalls and virtual machines and such.

## Slide #7

Scott Alberts (he/his): This discussion, and I think this module will give you a good introduction to thinking about these legal issues. Again, we're not trying to solve all the problems, but we want you to be able to have a seat at the table.

As you know, your company or your board or your boss, or the "data guy" they've hired inside or outside of the company, is trying to make decisions about how to avoid liability, but also to keep data available enough that we have enough access to that we can use but it's still secure and private so that.

We meet the legal responsibilities as well as that ethical idea that we don't want to share private data when we don't have to. Enjoy.

## Other Module Materials

Course Learning Objectives: A successful student will:

* • identify vocabulary and issues important to Big data security,
* • describe algorithmic decision making and its benefits and challenges,
* • explore techniques and methods for keeping big data secure,
* • critique cases where big data security was an issue, and
* • discuss legal issues for big data.

Module 7: Important legal considerations for data ethics, privacy, and security.

Upon completion of this module, a successful student will be able to:

* Understand and explain data protection standards
* Outline the EU's (and UK's) General Data Protection Regulation (GDPR).
* Understand and distinguish how American standards (especially the California CCPA) is different from GDPR
* Explain considerations in selecting, adopting, and complying with a data protection standard.
* Outline, assess, and synthesize different actions that could/should be taken to ensure the data that you work with meets a data protection standard

Data protection law and the ethical use of analytics (2010) pp.18-26

<https://www.yumpu.com/en/document/read/46664015/data-protection-law-and-the-ethical-use-of-analytics-international->

Legal issues related to Big Data (2013)

<https://blog.ipleaders.in/legal-issues-related-to-big-data/>

Big data: Six critical areas of legal risk (2014)

<https://www.techrepublic.com/article/big-data-six-critical-areas-of-legal-risk/>

A contrary view (2020);

<https://www.forbes.com/sites/forbestechcouncil/2020/07/29/the-privacy-mindset-of-the-eu-vs-the-us/?sh=63b1489d7d01>

Indeed.com's privacy statement: <https://hrtechprivacy.com/brands/indeed>