1:40

Alexey

**This week, we'll talk about practical data privacy. We have a special guest today, Katharine. Katharine is a privacy activist, machine learning engineer, and a principal data scientist at ThoughtWorks, Germany. I was really afraid that I would mispronounce this name. It's a mouthful. [chuckles] Previously, she held numerous roles at large companies and startups in the US and Germany, where she was in charge of implementing data processing and machine learning systems with a focus on reliability, testability, privacy, and security. She's also a book author. She recently published a book about data privacy, and we'll talk about this book today and about data privacy in general. Welcome to our event.**

2:22

Katharine

Thank you, Alexey. I'm really happy to be here.

2:25

Alexey

**As always, the questions for today's interview were prepared by your Johanna Bayer. Thanks, Johanna, for your help.**

# Katharine's background

2:32

Alexey

**Before we go into our main topic of data privacy, let's start with your background. Can you tell us about your career journey so far?**

2:41

Katharine

Yeah. It was a bit odd. I'm not sure if you and I are contemporaries in age, but I just reached about my 40th year in life – differential privacy, plus minus. [chuckles] And this means that when we were first in school and in uni, we didn't have a term called “data science”. I've kind of always been interested in doing data science, but I studied math and informatics (computer science). Then I switched my major when I ran out of math classes, because I really didn't like Java and I just liked math. [laughs] So I did political science and economics to work on statistical reasoning and so forth.

Then I went and I did public school teaching for a while, as a school teacher. Then I left that to get a degree in journalism. When I got my degree in journalism, I got hired as a data journalist (or what we would today call it data journalist, but back then, we just called it a “news application developer”). I worked at The Washington Post, which is a large newspaper in the US.

3:53

Alexey

**They have very nice visualizations, right?**

3:57

Katharine

Yeah! Yeah, at that point in time, they were the largest Django... remember Django, the Python-based web framework? They were the largest Django installation in the world at that time. So I learned Python. I had never seen Python in school. They just had us work on Java applets, which I did not like.

4:16

Alexey

**Applets! [chuckles]**

4:18

Katharine

Applets! Java applets, yeah. [chuckles] I was like, “Can I just write it in PHP? Why am I using an applet?” Anyways, that kind of got me back into data. Then I went and I led another team doing data journalism at a different paper at USA Today, which is a large national paper chain. Then I got recruited to go do natural language processing back in LA, where I'm from – Los Angeles. I jumped at the opportunity and we were working on trend analysis and trendiness using natural language processing.

From there on, I kind of went... that would have been 2010, and then, of course, word vectors took off a few years later. NLP changed greatly, which is really fun and exciting to be a part of. Around that time I moved to Berlin and I first started doing independent consulting when I moved here. I don't know how long you've been in Berlin. Were you in Berlin in 2014?

5:17

Alexey

**Since 2014, yeah.**

5:19

Katharine

Yeah. Okay. I don't know how your experience was, but I would go talk to companies and I would try to be like, “Oh, I can help you with machine learning and data science.” And they were like, “What are those things?” [laughs] I was like, “Oh, no! Did I make the wrong choice moving here?” [chuckles] But it was just starting to take off here, I think.

5:38

Alexey

**I was at university at that time and then I graduated in 2015. There were already jobs then. I would go to LinkedIn, type “data scientist” and I would find some stuff. So it took a year for Berlin companies to actually realize what this role was.**

6:01

Katharine

Yeah. I was mainly working with startups and I think it wasn't on the agenda so much. When I first came, I consulted some Rocket internet companies and so forth, starting out... [cross-talk]

6:13

Alexey

**Those were the only companies back then, I guess. [chuckles] Am I right?**

6:17

Katharine

Yeah. [laughs] Exactly. And now it's much nicer. Now everybody knows what you mean when you say data science and machine learning. Yeah. So that's what I did.

6:27

Alexey

**And then, at some point, I guess you joined ThoughtWorks?**

6:31

Katharine

Oh, yeah. After coming to Berlin, I was doing some independent consulting for a while, and I founded a startup here focused on privacy for machine learning, where we worked on, basically, stream encryption mechanisms for pseudonymisation, for link pseudonymisation, and so forth. With my co-founder here, Andreas Dewes, the company was called KI Protect. Then, after I left that, I worked at an encrypted machine learning company, based mainly in the US, but we were kind of all remote. I worked with some amazing cryptographers there, and we built an encrypted machine learning platform that was used by some folks in finance to do machine learning (deep learning, primarily) on encrypted data, but also other types of data science processing on encrypted data. We can talk about that more.

It was very exciting. And then the company chose to go in a different direction, so I left that there, and that's when I started at ThoughtWorks. That would have been January of 2022. ThoughtWorks was starting to see an uptick in data privacy interests and advanced privacy. And so now, I focus a lot on public sector and finance sector work and even advising globally on new privacy initiatives that people are rolling out across the world and then also leading teams here in the public sector of Germany to build things with privacy built in.

8:08

Alexey

**Yeah. Wow, there are so many things that you have done that it's probably easier to list what you have not done – political science, economics, school teaching, journalism, NLP, crypto – that's like everything. [chuckles] must be a fun journey, right?**

8:26

Katharine

It's been fun, yeah. [chuckles] Keep learning, you know? Right? You feel the same way. Have you kept learning, you think?

8:34

Alexey

**Yeah, but maybe in a less entertaining way? I was just a boring Java developer and then I saw this thing called “machine learning” and then I decided to do it. So I did, and then I worked since then. Now I'm focusing on community. So it's just a lot fewer things than you got to try in your career. But yeah, who knows what happens for me next, right?**

8:58

Katharine

I know, yeah. I think I have about 10 years of career on you, so you've got some time. [chuckles]

9:04

Alexey

**Like seven, maybe. [chuckles]**

9:05

Katharine

Okay. [chuckles]

# Katharine's ML privacy startup

9:08

Alexey

**So this startup that you mentioned, KI Protect, that you and your co-founder started – this startup was about privacy in machine learning, right? So what kind of problem did you see back then that you decided, “Okay, I need to have a startup.” And why this particular domain? Why privacy?**

9:32

Katharine

Yeah. So there were numerous things that I think were happening at the time. First off, GDPR was about to be rolled out around that time. I think that we definitely noticed some need in the industry to start to address privacy by design in machine learning, which I think a lot of people didn't know how can we even do machine learning with privacy and how that would look. We had some customers that we were working with, that wanted to train on datasets, where they were going to share datasets either across companies, or even across pretty highly secure datasets. This is also later what I focused on at Cape Privacy and Dropout Labs, the company encrypted learning.

But the problem in the industry, and I still see it all the time, is that you have data partners or you have numerous companies or even within a company, you have a highly secure data zone, and then more widely available data. And yet the machine learning problem you're actually trying to solve needs that secure data. But if you could, you'd rather not send the data over or centralize the data. So I think this is a huge problem. It's a problem that I've now worked on since 2017. There's numerous ways and that's some of what's covered in the book – there are numerous ways.

One of it is thinking through link pseudonymisation, that's probably the least amount of protection that you can offer. Another is moving towards encrypted machine learning or federated machine learning, or both. I think it's a real need that will only get stronger over time, for better or worse. Now we're starting to see those conversations also pop up in the US, which is not really known for its “privacy first” approach.

# GDPR, CCPA, and the “opt-in as the default” approach

11:33

Alexey

**In the States there is a thing that is similar to GDPR, right?**

11:38

Katharine

So there are a few things. The States – each of the states themselves have started passing newer privacy regulation in the last few years. Some of the newest stuff I haven't even read through. There are so many that are proposed everyday and only a few pass. But California's new ones – first, they pass CCPA in California, and that's to cover how data is collected and how it's consented to and how it can be shared.

Then they actually passed, essentially, I would say an enhancement of it last year, which is called CPRA (California Privacy Rights Act). This one essentially makes it a lot stronger. For the CCPA, it was only if you had a data breach could you be fined or sued, but the new one is actually going to set up (exciting!) one of the first regulatory enforcement authorities in the US that is just around privacy. They're hiring now. So if you're in California and you want to work on this, they're hiring this person right now and you get to lead the first Protection Authority (only on privacy) in the US, in the state of California. So, yeah. That's kinda cool.

12:59

Alexey

**Now people in California will also need to consent to cookies, right?**

13:05

Katharine

Yeah, they're gonna have to consent to all that collection. There's also some certain restrictions on how the data can be shared, so third-party data sharing becomes a lot more restrictive. You can delete your data – so you can request deletion. It doesn't really have a lot on portability, which we have here in Europe – of taking your data with you. But some measures around that I think are coming. We'll see what happens.

13:33

Alexey

**Before, you didn't really have these pesky banners that you have over your entire screen saying, “Hey, consent to cookies.” Right?**

13:45

Katharine

Which now... Okay, by the way, since like six to nine months ago, if you don't see a one click of opting out of “everything but necessary,” it's technically not GDPR-compliant, because the French authorities came down and said, “Absolutely not. No more with these 10 clicks to turn off cookies.”

14:11

Alexey

**I think people just accept them, though. [chuckles]**

14:15

Katharine

[chuckles] Well, there should be a one-click reject now and the default should be to reject. But yeah, it's a long road, as you know.

14:24

Alexey

**Do you accept cookies or reject them?**

14:25

Katharine

I reject cookies. I accept real cookies. [laughs]

14:32

Alexey

**So why do you reject cookies?**

14:35

Katharine

Because I think I've seen how the collected data gets shared and used in advertising optimization and I'm not a big fan of personalized advertising myself. I find it to be annoying. So I'd rather not participate if I can opt out. But I think everybody has to make their personal choice. But yeah, I think the default should be “opt-in only” so if I don't want personalized ads, why can't I just tell you, “Cool, I'm on this website that sells data engineering books to show me ads about data engineering. Why are you trying to show me ads about chews on a data engineering website?” It makes zero sense to me.

15:24

Alexey

**Well, now when you say this, it does make sense. You just know what the site is about, so a person who landed on that website is probably interested in the topic of the website, right? [Katharine agrees] I did not think about that. I also happened to work at some point in an ad tech company and we were those people who were doing this nasty stuff of collecting data.**

15:48

Katharine

And you still opt in for cookies?! [chuckles]

15:51

Alexey

**I do. Yeah.**

15:54

Katharine

You want to feed the algorithm! [chuckles]

15:57

Alexey

**Yeah. I mean, after working at this startup, I removed the ad blocker from my computer.**

16:03

Katharine

Oh, yeah?

16:04

Alexey

**Yeah. Because I realized how these companies make money. [chuckles] It's interesting.**

16:12

Katharine

Okay. I'm curious as to what your decision matrix is like there. But yeah.

16:19

Alexey

**It's not like I had a decision matrix. [chuckles]**

16:22

Katharine

Okay. [chuckles]

# What is data privacy?

16:24

Alexey

**So what is data privacy? I think we kind of talked about this, but we didn't explicitly define it. So what is it?**

16:32

Katharine

Data privacy, I mean, it's a huge topic. I have this one diagram that I use in the book, which, by the way, you can preview for free on Amazon and eBookNow and you could see this diagram. I like to talk about the fact that there are legal definitions of data privacy, that's a large aspect. Obviously, depending on where you live, there might be a different legal definition of what data privacy means and there might even be a local level, or a contractual level, and so forth. So there are all those legal definitions.

And then there's societal definitions and cultural definitions. When I talk to you and I say, “Hey, I'd like to make this conversation private,” what do you understand from that? What do you understand about the ways that we communicate privately with our families and friends and colleagues and so forth? So there's a whole area of data privacy, that's just kind of how we understand it as humans and humanity and then there's technical data privacy, and that's the field that I work in. It's kind of about, “How do we rigorously define privacy? How do we measure privacy? How do we reason about privacy in statistics, mathematics, and in areas like machine learning.”

I think it's cool because all these can intersect, and you can take a legal interpretation and map it to social interpretation and map that to a technical interpretation. I think maybe your individual definition of privacy is very much influenced by what sphere you work in. If you talk to a lawyer who works on privacy rights, they have a very different interpretation of their own personal privacy, compared to if you talk to somebody that's never worked in law. Just like I've been working for a while now in privacy and I think I probably have a different definition than a lot of people.

It's kind of cool that there are all these different factors, and people probably move around over time and in their life. I think we can see that today with how younger people operate on social media and so forth and creating fake profiles, or this or that and whatever. And there are other people who are really open and they share everything – they stream their entire life. So you have these different choices that people make, which makes it kind of interesting to think about.

19:01

Alexey

**I have a fake profile on Facebook.**

19:04

Katharine

Ah, okay. What's your name? [chuckles]

19:08

Alexey

**Well, my first name was my first name, but my last name is different. And the reason I did this was because I was afraid of FSB – this Russian security service tracking me. [chuckles]**

19:17

Katharine

Yes. Yep. Which is a real thing, obviously.

19:24

Alexey

**So yeah. [chuckles] I don't know if they actually do this. Maybe they do.**

19:29

Katharine

I have seen it with my own two eyes. So indeed, I can confirm that.

19:36

Alexey

**The interesting thing, since we're talking about privacy, is that people still find me, surprisingly. There are services where you can upload a picture of somebody – for example, I speak at a conference and somebody takes a picture of me, they upload it to the service, and this service finds my profile on Facebook. Can you imagine? And then they contact me through Facebook. [Katharine agrees]**

**For me, this was like, “Oh, wow. How is it even possible?” [chuckles nervously] Technically it's easy – you just scrape all these forbidden profiles or whatever and index them... Yeah, that was a surprise for me. I guess that's another definition of privacy, right? I don't want to be discovered, but people still do this even though I have a fake profile. Interesting.**

20:27

Katharine

Yeah, there's this concept that I talk about in the book too and a lot of what I talk about when I discuss these topics is – there's this translation from the real world into the digital world or technology. And I think we haven't been exactly very good at figuring out how to translate privacy from one to another. I should be able to infer, based on you changing your name, that maybe you don't want to actually be contacted randomly on Facebook. [chuckles] But there's also no way for you to tell Facebook “Hey, I'd rather people not find me using these types of services.” And the fact that the services even exist, there's no choice to opt out of them saying, “Please don't do this for me at all.”

21:13

Alexey

**It's not like I want to opt out. I want to not opt in, right? I want to be opted out by default – I don't want my profile to be scraped. If I want other people to contact me, why did it not occur to them that they could use LinkedIn or something, right? [chuckles]**

# Finding Katharine's book – Practical Data Privacy

21:35

Alexey

**Anyway, there is a question that I see someone asked, “How can we find this book?” The book is called Practical Data Privacy. If you just Google it, I think that first thing will be this book. There's only one book with this name, right?**

21:48

Katharine

Yes. I hope it's the first link. If not, then find it and click on it, so it'll be the first link one day. [chuckles]

21:55

Alexey

**A more certain way to find it is to also add your first and last name after, right?**

22:00

Katharine

Yeah, yeah, yeah. Then you'll for sure find it. The e-book went out yesterday, so now it's available. The print copy should be available in the next week or two.

22:13

Alexey

**I think we should reach out to O'Reilly and ask them to give us a few copies. Right now we will record the podcast episode (the interview) and then we will publish it in a couple of weeks. By the time we publish, the book will be out and we will probably have a few copies. So everyone who is listening, keep an eye out – when we release this episode, there will probably be some sort of giveaway where you can get a free copy.**

# The various definitions of data privacy and “user profiles”

22:38

Alexey

**Anyway, coming back to the definition of data privacy. You said there are a bunch of different definitions. Maybe you can give us an example of how a lawyer would define data privacy versus how a data science manager would define privacy?**

22:53

Katharine

Yeah. Really, these two need to eventually agree, right? At the end of the day, when we're working at a company and we're building out data products or services and so forth, and they're operating in some sort of regulatory context, which is all the time in Europe and most of the time in other places, we want the legal team to agree with the technical team about not only the data we're collecting and processing, but how we intend to use that data and the risks that we hold in using that data and the mitigations that we have for that risk. So it's a whole beautiful picture there of all the different things that you need to think about at once.

A lot of times, the conversation between the legal people and the technical people can get really difficult because both fields use quite a bit of special words or jargon and that means that being able to translate between those spaces is kind of like a superpower that you can build over time. Something that we have to really work on from the technical side of the house as data people, because a software engineer does not have the same depth of knowledge about data and statistical reasoning, and I think to fully understand the data privacy problem, you have to have some exposure to statistical reasoning and thinking about data in real way.

So I think, if you're a data professional, you can really actually start leading this conversation. Because when we think about things like – let's think about privacy risk – if you're in the complete middle of a distribution, the chunky part, and the chance of you being singled out in any way, shape or form is almost zero, then your privacy risk is much, much smaller than somebody that's at the very thin tail somewhere in the distribution.

25:06

Alexey

**Can you give an example? Because for me, it's too abstract to understand.**

25:11

Katharine

Okay. Yeah.

25:12

Alexey

**We talked about this website – about data engineering, right? [Katharine agrees] So maybe you can use the online advertisements as an example. I visit a website and there is an option to have personalized advertisements or generic ones. In this example, what are the privacy risks?**

25:33

Katharine

Yeah, okay. Let's go with that. So, Firefox actually studied this and web histories – the series of things that you visit in your browser. Even over the course of a fairly short time span (let's say one week) are uniquely identifiable most of the time. This is what we can think about. We all use some major services – you log into your email, you maybe log in to social media and other websites, and you do that every day. Even news websites are quite popular. So that would be a large percentage of us who just have those. Those, essentially, don't leak very much information because so many people use them every day.

But then, let's say you go and you click on this book link, and you go look at that book. And then, let's say later today, you look at two or three different YouTube videos. And then, let's say later tonight, you log into a fitness website that you use, or you go to a gaming portal that you use, and then maybe you go to a subreddit that you like to post in. When you start to trail these, with all these data points together, maybe not any one of them, but that timeline of history together, starts to become uniquely identifying. That's been studied and proven and I think, as statistical people, or people that understand that, you can start to see that to remain non-identifiable, you'd have to really restrict and think about how you use the internet. And so it's just not...

27:15

Alexey

**We had a name for this thing when I worked in an ad tech company. We couldn't monitor the entire internet, of course, and we could only see the apps that users use. But that's enough. It's pretty much the same. For example, you use Bubble Witch and then you use some other app and then... It's the same story as you just said. We call this thing a “user profile”. And then for each user, we had this profile, and we basically knew everything about this person, including the geographical coordinates. So we knew where they live, which is mind blowing, right?**

27:59

Katharine

Yeah. Just to go back to the user experience, I think that the average user, even those of us that have worked in this space, we don't actually think a lot about the privacy risks that that poses to people and the fact that some people would probably rather not have that amount of private information exposed or that fingerprinting via various mechanisms – you could use browser signatures to fingerprint, you can use operating system information to fingerprint. There's even been people that fingerprint mouse activity and so forth, or typing flow. I mean, there's a lot of ways to violate someone's privacy using technology.

28:45

Alexey

**Why do companies do that? Don't we like to have other ways to identify that you are you?**

28:55

Katharine

I mean, yes, presuming that you log in and that you participate. For sites that don't have login, or for sites that want to offer personalized tracking for people that are not logged in and later login – so a lot of shopping cart optimization and so forth – also uses these types of things to link profiles of logged in and logged out users. What I guess the bigger question here is – what value does it actually bring for us to know exactly who somebody is?

29:35

Alexey

**Zero, most of the time? It's also just creepy.**

29:39

Katharine

[laughs] It's like, “What are we actually trying to measure here?” That's a lot of what I try to ask.

29:47

Alexey

**The goal would be for an advertisement company to serve the ad that a user is more likely to click rather than ignore, and the rest is not important. We don't really need... Well, I also remember that there are use cases for geographical coordinates. For example, if there is a store nearby that has some campaigns or some discounts, then we can promote that.**

30:15

Katharine

Yeah. One of the things that I'm excited to see is that I think there's a lot more thinking through intent-based recommendation, so “What is the intent of the user? And how do we improve?” The first time I noticed it was, I think, with Amazon and Netflix, who are both starting to talk about this, where they wanted to do session-driven recommendations. It's also kind of a bit of a “cold start” problem. The user comes and they might not have previous information about the user. Given the way they're interacting with the site right now, at this moment – what they're searching for, what they're scrolling to, what they're clicking on – how do we give them something that matches their intent?

And I think there's ways to design these systems to be extremely privacy preserving, particularly if you use a training group that has fully opted in to being tracked. Then you use almost an ephemeral inference that you can just throw away the data when this user leaves the session. You could really design this in an extremely privacy-friendly way, if you wanted to. I think that they were starting to experiment and showing, “Oh, hey. It's kind of almost better than personalization if what the person is trying to do, we immediately give them what they're looking for.”

31:50

Alexey

**In this case, as you said in “privacy preserving methods,” we do not necessarily need to look at this user profile that we mentioned. We don't necessarily need your five years of browsing history. We just look at your session – the last 10 minutes, for example.**

32:08

Katharine

Yeah. What are you trying to do?

32:09

Alexey

**Just activity on Amazon or whatever? Right?**

32:12

Katharine

Yeah, exactly. I think people consider, especially since it's been shown with research over decades now, that if sharing my location right now is going to help improve my experience, I am willing to do that. But I don't want to turn on location sharing forever. I just want to say, “Hey, look. I'm looking for restaurants near me. Here's my location right now. Tell me what choices I have. And then let me opt out immediately after I'm done with that search.” So I think there are some pretty cool, if you wanted to, to build privacy into these things and to also meet users where they're at – to let them opt in and out on a case-by-case basis. It could even be something as simple as, “Hey, we'd like to use your location to improve the search results *for this search*,” or something like this. So I think there are ways to build it in quite easily.

# Privacy engineering and privacy-enhancing technologies

33:08

Alexey

**Thinking about these things and incorporating them in our systems. Would you call this privacy engineering as a term? Privacy engineering is exactly about these things? Right?**

33:21

Katharine

Yeah. I think that could be some of the basics. You look at the data side, or you look at the application side of what you're trying to do, and then you look at the privacy risks and constraints – you try to build out requirements that can meet both. This also means learning to talk with legal staff and so forth. But then also, it can go even further. Some of what I talked about in the book, and what I'm really excited about, is privacy-enhancing technologies, or what I often just call privacy technologies. They are particular ways that we can architect, or particular types of algorithms or cryptography that we can use, that takes these problems even further, and it takes them to the next level, to allow us to do more advanced data science and machine learning, but also have an option for rigorous definitions of anonymity, for keeping data localized (for not centralizing so much data) and for doing real data science on encrypted data, where the data is never decrypted.

All of those, I think, are the really cool stuff. But as you can imagine, it takes a bit of time to integrate them into a data architecture and a data science workflow. I would almost always advise people, “Start with the simple stuff, get that conversation going at the organization, and then they can move into more advanced privacy engineering.” But I'm a big fan of broadening the definition. I think the definition of “privacy engineer” is all of that. I would say to use it liberally. We need more people talking about privacy engineering. So if you're listening to this and you want to be a privacy engineer – do it! [chuckles]

# Why data privacy is important

35:09

Alexey

**I'm still somewhat skeptical about this. Let's say I run a company. Why would I care about these privacy-preserving techniques? For me, from what it feels like, if I use all this information I have about the user, then I will be able to serve them with better ads and have a better user experience. For example, if we take Google – Google knows that I am me. When I type something, it shows me a Stack Overflow article, and not something else. Because it knows me so well, it knows what kind of stuff I'm into and then the experience I have as a result, is much better for me. Presumably, right? [cross-talk]**

35:58

Katharine

How much of that is personalization and how much is just a very nice search algorithm or series?

36:05

Alexey

**Well, there is a lot of ambiguity in many things. For example, for me, it's a Stack Overflow article, but for you, it could be something like... I don't know, just an article about something else. Because I clicked on Stack Overflow links so many times, it knows that I like this site more than other websites.**

36:31

Katharine

I mean, I think individualized search ranking is probably something that Google moved away from a long time ago, but I'm not 100% certain. But I'm pretty sure about that one. I think there's some of that. There's also, of course, some more real-time generation of results, which can change results over time. But going back to the example of you running a company – why integrate privacy?

37:01

Alexey

**It feels to me that it's better without privacy than with privacy. [chuckles] How do you prove me wrong?**

37:07

Katharine

Yeah, yeah. There's a few things. First off, there's a ground statement here of, “What does it mean about your product and your company if you can't build privacy in?” I think there's a moral and ethical thing that you need to think about where it's like, “If the entire basis of your company or your product is not going to work if you allow people to opt out or to share less data, then what exactly are you building?” I think that's a big question to ask. But let's say you say, “Katharine, whatever. We tested it and it works better with private information.” What you need to think about is the managed risk that you hold at that point in time.

I've worked in finance and the public sector now for a while, and, therefore, the risk appetite of the people that I work with is much smaller than a lot of industries. But I think what I'm starting to also see, even just ThoughtWorks Global, some of the things that are coming across my desk, as one of the privacy SMEs of the company, is that there's a growing concern for a lot of people who hold data. And I think there's a lot of companies that you wouldn't expect are thinking and prioritizing this problem that actually are. I think the reason is, there's regulatory pushback and there's customer and social blowback.

I think both of those are real and if you can build the product and deliver what you're trying to do by collecting less data and holding less risk, it's just a smart financial decision. Because if something happens and you get sued, or if something happens and your insurer drops you, that's gonna cost you a lot more than the half a percentage bump that adding location to the machine learning regressor will give you. So I think you have to actually think that the trade-off you're making there is not 0% risk/100% reward – it's a balance between those two.

39:20

Alexey

**I guess for me as, let's say the CEO or whoever of a company, I need to rely on the data protection officer, or these privacy engineers that we talked about, who can tell me, “Look, keeping this data might be dangerous in terms of maybe...”**

39:40

Katharine

Or illegal. [chuckles]

39:43

Alexey

**Illegal, yeah.**

39:44

Katharine

It could just be illegal. [laughs]

39:47

Alexey

**But in the case of using location data, that's not illegal. It's legal. As long as you have consent.**

39:52

Katharine

As long as it's consented, yes. Yeah.

39:53

Alexey

**Okay, but we have to provide an option for users to opt out of this, right? [Katharine agrees] Then our website still must still function without location data. [Katharine agrees] At least in Europe, right?**

40:11

Katharine

In Europe, and I think in Brazil now with LGPD. Chile has a new law coming out. China has some – the China law that came out last year. Let's see how it gets enforced. So it's an increasing number of places.

40:28

Alexey

**Actually, I recently had this experience where there was a website (that I will not name) and when you reject the cookies, it stops working.**

40:42

Katharine

[laughs] Obviously by design, everybody. A great user experience.

# What is differential privacy?

40:50

Alexey

**I heard this term called “differential privacy”. What is this? How is it different from usual data privacy?**

40:59

Katharine

Differential privacy is a concept that we use to reason about privacy loss. It's often used to think about how stringent we need to be about the amount of privacy or information that we give out about any one individual in the dataset, in order to provide them with the highest level of privacy we can offer. In Europe, and in a lot of places, it is also often considered by legal experts. If you talk to a privacy lawyer, they actually know what differential privacy is, which is really cool. It's kind of the highest standard for thinking through the problem of anonymization.

To go briefly through it – I'm not sure if you're familiar with Cynthia Dwork's work? Cynthia Dwork was a researcher at Microsoft for a long time and she developed some of the fundamental theorems in a few different hard crypto problems and then also, this thing of differential privacy. She was, I think, working in machine learning at the time. This was a quite a while ago. She and a few of her peers had debunked the idea that data could be released and be anonymous. Because the problem is, we work in information theory, and in terms of information theory, when you release information, the chance of the information being about individuals in the data is quite high. Then the chance that something can be learned about the individuals is quite high, too. This is kind of the baseline of why data science works. You take information from a population, you synthesize it, and you learn. Right?

So what Cynthia Dwork did is, after they proved that there was no such real thing as an anonymization from a scientific sense of the world, if you're going to release data, then they were like, “Oh, no. We broke it. Maybe we should come up with a new definition that we could use so people can still keep doing data.” So what they developed is this reasoning about the probability that somebody can learn something, in the classic definition, before and after a person was added to a dataset. So you're building a query mechanism, you want to give an analyst access to a query mechanism, the analyst queries before a person is added, and then after a person is added, and we hold the probability that they learned something about this person who is added (or even that the person was added) within tight bounds. That's the formal definition of differential privacy.

That original definition has been expanded to include a lot of cool concepts like computational complexity, thinking through the chance of particular types of error happening and responding, and so forth. There's a chapter on it in the book. I got to work with Damien Desfontaines, who's one of the global leading experts on differential privacy, and is based down in Zurich. He led the Google differential privacy implementations when they first did theirs and now is at a really cool startup called Tumult Labs. They have a Spark-based open source differential privacy library that you can use at this moment, that you can install. I was one of their beta testers. This is a really cool library, so I can recommend it.

44:36

Alexey

**What's the difference between anonymization and what you just described? I guess, we can take a dataset with some private information, like emails, phone numbers, geographical locations, etc. – this information that we can use to identify a specific person – and we just hash the data using some sort of MD5 hash or something. Is it not enough? Are we not within this differential privacy thing?**

45:08

Katharine

Those are kind of what I would call more “old school” methods of anonymization. People might have also heard about K-anonymity, which is another one of what I call an old school definition of anonymization. It's just like, “Oh, I dropped all the personally-identifiable information, so it's anonymized.” And I think what Dwork (and numerous folks after) were able to prove is that that doesn't actually cover anything about anonymization. Because if you don't release any data – you don't release any of the attached rows or other information from the dataset, then sure, then you can call it anonymized because no information is released.

But depending on how you release your groups and what other information is in there, again, like we were talking about, there are quite a few things that are personally identifiable, like what apps you use, what phone you use, your browser, this and that, whatever. If you don't remove all those as well, then that's just more information leakage. This is why people that are outliers have a special amount of privacy risk, because if let's say there's one row in the data set that you're releasing, where somebody is singled out, so to speak, these can lead to all different types of privacy attacks that people can perform. One of the great ones around de-anonymization was the Netflix prize, in which Narayanan and Shmatikov were able to prove that they could they de-anonymized (they re-identified) a group of reviews that Netflix released and they said they were anonymized. They had just done exactly what you said. It was like, “Yeah, we totally anonymized it.” [chuckles]

# The importance of keeping privacy in mind when designing systems

47:00

Alexey

**At DataTalks.Club, we run courses. As a part of that, when students submit their homework, we ask things like, “Hey, how much time did you spend on this particular homework? How much time did you spend watching lectures?” They answer all that. Then, at some point, we release this data so that others can explore it. What we did is take the email and just applied SHA1 hashing.**

47:28

Katharine

With salt.

47:33

Alexey

**Without... [chuckles] Is this bad-bad?**

47:41

Katharine

[laughs] We can do a privacy analysis later. You show me the script you use and I'll help you out. Okay? [chuckles] I mean... Sorry, did you do any other randomization of the data or anything? [Alexey says “No.”] Okay.

47:57

Alexey

**It was not so much data. We had this information – how much time they spent on different activities, plus an anonymized email, so we can identify that this is the same user across different modules of the course.**

48:13

Katharine

Yeah. Again, I think probably people have opted in for this by using the site and by submitting the form, so I think... I'm not here to try to give you legal advice. [chuckles] I'm also not a lawyer.

48:26

Alexey

**[chuckles] Noted.**

48:28

Katharine

But there's ways that...

48:29

Alexey

**You should have said this from the very beginning of this interview. [chuckles]**

48:33

Katharine

There's ways that we can offer a bit more privacy for the users, or also there's consent – people that say, “I want my data to be used,” or “I want my data used with anonymization,” or “I only want this data used for these things.” But essentially, differential privacy as a mechanism – we use certain algorithms to implement differential privacy and then release data – it can help you actually reason, “What's the privacy loss of each individual?” And it can start to cap individuals at a certain point in time.

There's a cool graphic in the book and also Damien has a blog series Introduction to Differential Privacy. But there's a cool graphic where you can actually start to view the probability distribution of the different values of certain parameters we use to tune differential privacy. With this, you can actually then sit there and reason, and say, “How much privacy are we willing to have so that this data could still have high information, high utility, but tune it?” And maybe even you could tune it on an individual level, if you ever wanted to give that option to people. But I mean, if you have an iOS device, differential privacy is already running on your phone on whatever iOS device you have.

This is something that's been deployed in quite large scale ways in a lot of the larger tech companies and I think it's more usable now, for those of us that didn't get PhDs in differential privacy and how we can experiment to use them. I have a few notebooks for the book too. There's a repository. So if you want to play around, I think there are three or four notebooks on building your own differential privacy mechanism and then one notebook using the Tumult analytics library, which I found to be the most user friendly of the ones I reviewed for the book.

50:27

Alexey

**Now, I'm kind of freaking out and thinking whether I should delete all these things. [chuckles] [Katharine laughs] Like, am I breaking any laws?**

50:36

Katharine

No, no – sorry, I don't want to scare people with this information. I want people to feel empowered.

50:41

Alexey

**Too late. [chuckles]**

50:42

Katharine

I want people to feel empowered. So, look – some of it is about intention and awareness and I think folks who participate in DataTalks.Club understand. There's a FAQ chapter of the book, because I get a lot of questions on this stuff, and one of the FAQs is something like, “Oh, no. We released data and we didn't apply any of this stuff. What should I do?” And it's like, “Just think about it the next time. Nobody's gonna die (hopefully). Nobody's gonna get arrested (hopefully).” [laughs] I meant based on your data release, you know? “Hopefully, next time, just have a think through it.” I think, honestly, if people just started writing up some privacy requirements as part of the normal data science process that, in and of itself, would solve like 95% of the issues we have.

It's just actually sitting there thinking for a second, “Hmm. Is there more privacy we can offer? Or should we offer people the chance to opt out? Or can we try out a cool new library that we heard about?” Just start to build it in. It's the same thing as everybody talking about building CI/CD for data science processes, building testing into data science processing, building validation – I mean, these are all concepts that we can slowly work into our workflow, and nobody's gonna report you. I'm definitely not gonna report you. [chuckles]

52:24

Alexey

**Some of the students who are listening to this and freaking out, please just reach out to me and I will remove your data. Don't sue me, please. [chuckles]**

# Data privacy on the example of ChatGPT

52:35

Alexey

**What do you think about ChatGPT? How privacy-preserving is it?**

52:41

Katharine

It's been interesting. I have a newsletter on data privacy and machine learning and in the past two issues, I just only wrote about ChatGPT. You heard about it being banned in Italy for a while, right?

52:54

Alexey

**Yes. It's not anymore?**

52:55

Katharine

Not anymore. So, cool – here's an amazing privacy story, everybody. [chuckles] ChatGPT was banned in Italy. I don't know if you remember the big privacy leak that happened, but some users were logged in and they could see other people's chat histories. This happened a month and a half ago or something like that. You could see the chat histories of other people and people started sharing screenshots. Then they took down history for a while and then they put it back up after they got it fixed.

But in that time period, they actually had to notify people in Europe whose data was exposed, if they found that there was private information in the chat histories. Good job GDPR – that it exists. Right? So this is a win. So, some people got notified and there were things like email addresses, credit card numbers – there was all this private data that people were feeding into ChatGPT.

53:58

Alexey

**Why would you put credit card information into ChatGPT?**

54:02

Katharine

You're probably copying and pasting from internal emails or something like this, right? Or other stuff.

54:09

Alexey

**Because you don't care about this? What bad thing could happen, right?**

54:14

Katharine

Yeah, but this is a problem. People are just copying and pasting anything into ChatGPT, which is... [cross-talk]

54:20

Alexey

**Is it ChatGPT's fault, or ours?**

54:22

Katharine

Well, it is definitely not the user's fault when privacy comes into question. Because I think the thing that we have to think through is, if we don't build things where the easiest path is security and the easiest path is private, or at least gives people consent options – this is how we have to design stuff. Because people are going to use stuff however it is easiest for them to use it, and if there's a big text box and says, “Hey, I'll help you write email responses,” people are gonna put their email in it. [chuckles] That's exactly what they're gonna do. And there's even OpenAI people being like, “Please do that.” So they have to assume that people are gonna do this.

Anyways, they got notified and Italy was like, “Absolutely not! We're shutting this down.” So Italy shut it down and said, “You have until the end of the month to prove you complied.” It was literally like a day or two before the end of the month, which is when they launched the turn history on and off. So now you can turn history off, which is what they should have done in the first place, but they didn't. But at least you could turn history off.

I think some of my writing on ChatGPT from the start was, “Hey, let's not even touch GPT 3, 3.5 and 4 and how much private information is in that. We're gonna leave that for a second. That's another ballpark. Let's just talk about the interface and how you roll it out.” I used it a little bit in the beginning to try to see how all of it worked, obviously, and see if I could get it to tell me interesting things. And one of the things I noticed was that they were like, “Please do not put personal information here,” in like a tiny little box. [laughs] I don't think anybody read that.

So it's just how you communicate to the user and I think most people didn't know that it's a real... obviously, people working on machine learning know, but they don't know it's a reinforcement learning system that's under active training, so any of your chats can be used at any point in time to optimize the reinforcement learning with human feedback system that's underlying it. The amount of probably proprietary, confidential, and personal data that's now in that reinforcement learning system (or several of them). Hooey – they have probably a big job to figure out and sort that out.

57:00

Alexey

**They will need to identify that in their historical data, in all these logs, “We have credit card information, addresses, email addresses, dates of birth, and all these kinds of identifiable information,” they will need to mask it (put placeholders there) and only then use it for reinforcement learning. Right?**

57:24

Katharine

Yeah, that's an ideal first step. I think the other thing to question is, “Under what consent was that collected? Is it allowed to be stored for a long time and used for training? Is there a retention period for that? Were people informed? How can people delete it?” So there have to be deletion abilities. And then one of the things that I don't know if you saw or not, but Amazon's corporate lawyers had to send out a big memo to Amazon and start doing mitigation with OpenAI because there were internal proprietary Amazon documents that they found in ChatGPT because people were using it. [chuckles]

Again, I don't blame any users, because users just want to do what users want to do. That's the whole point of offering a service. It's OpenAI's responsibility, and now Microsoft's responsibility in a lot of ways, to actually build a user interface and machine learning system where they think about this stuff and they make it a little bit easier to use. I say we should respect and protect even the most irresponsible user, because that's part of our job. That's a responsibility we take on when we give somebody a huge text box and we say “Go to town.”

58:45

Alexey

**Yeah, I just realized how irresponsible it was on me when I recently got a contract. So I had this contract. And I'm not a lawyer and I thought, “Okay, let's ask ChatGPT what it thinks about this contract. Are there any suspicious parts?” So I took the contract, Ctrl+A, Ctrl+C, Ctrl+V to ChatGPT, and asked it something like, “Is this a good contract?” And it said, “I'm not a lawyer, but this contract looks fine.” And now when you talk about this, people who put credit card information there, and I think, “I'm the same.” [laughs] I just took that contract and put it there without realizing what could actually happen with this information.**

59:29

Katharine

Yeah. But, again, I don't think the responsibility should ever lie on the user. You're doing exactly the way the product is designed to be used. It is not your fault that it's not thought through. How you could be like, “Oh, hey. For this interaction, could you just forget everything that I'm going to tell you?” Which should be available. One of the other cool things I learned with the Italian stuff and everything going on – obviously, I've been advising a few different groups of folks on how to use ChatGPT in a more compliant manner. And I think one of the coolest things is that Azure launched the ability to localize ChatGPT for you.

So not only is it localized in whatever your Azure cluster or cloud looks, so the data is not being sent to the US (which is a part of fulfilling some of the GDPR requirements) but then also, you can even do fine tuning and have an individual reinforcement learning system on top of the underlying language model. And all of that is yours – you actually own that. So I think that's pretty cool. Maybe it's something that is worthwhile for people that are using it – to put some money into the Azure credit machine so that you have your own ChatGPT. [chuckles] Of course, there's like a million open source repos where people try to recreate some of these things. I'm excited to see how that develops over time.

61:03

**Alexey  
Okay. I realized that we're slightly over time. Unfortunately, there's one question that we did not cover. Maybe next time.**

# Katharine's resource suggestions for learning about data privacy

61:15

Alexey

**But one thing I wanted to ask you before we finish was – if I want to learn more about this topic, of course there is book that I can go and read, but is there anything else you would recommend in addition to the book that would be useful for me to check out if I want to learn more about this topic?**

61:31

Katharine

Yeah, obviously I will mention that my newsletter is called Probably Private. [chuckles] ProbablyPrivate.com So you can check it out. I'll try to cover things there. I know I mentioned Damien's work earlier. You can find him online if you search “Ted on Privacy” or “Damien Desfontaines”. His work is really amazing around differential privacy. There's a bunch of cool references in the book, too. There's like a million other people whose work I really love and admire. If you go on Amazon right now, you could preview the book up to the end of the first chapter. You can also do that on ebooks.com and if you look through those you'll see a bunch of references to the beginning concepts. I'm thinking of promoting a series called “What does it mean to be a privacy engineer?” and interviewing some of my colleagues, peers, and also folks whose work I admire. So maybe stay tuned in and you can see some blurbs on what it means to work in privacy.

62:37

Alexey

**You'll probably announce it in your newsletter, right?**

62:40

Katharine

Yeah. There, and definitely on social media and so forth.

62:44

Alexey

**Okay. So, keep an eye on all social media accounts and the newsletter. And with that, that's all we have time for today. Thanks a lot for joining us today, for telling us about all this. I also realized how irresponsible I was in a few cases, [Katharine laughs] which I guess is a good outcome? [chuckles] Thanks, everyone for joining us today, too.**

63:08

Katharine

Thanks so much.