1:23

Alexey

**This week we'll talk about industrial data challenges. We have a special guest today, Rosona. Rosona is a trained mathematician who works in the data space for the last six years, with the last three, working in industrial data. She is currently a machine learning engineer in a technical leadership role around synthetic tabular data in an AI innovation team. She's particularly intrigued by industrial R&D problems. Welcome to our event today, Rosona.**

1:57

Rosona

Thanks for the invitation. [chuckles] I'm glad that it finally worked out. We've been trying to meet up… [cross-talk]

2:02

Alexey

**It was a year or maybe more. It was always fun with you. Anyways, every time we spoke previously, you always had an opinion about many interesting topics. Today, we finally have a chance to put this on the record and have a podcast interview about that. The questions for today's interview are prepared by Johanna Bayer. Thanks a lot, Johanna, for your help.**

# Rosona’s background

2:28

Alexey

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

2:38

Rosona

I mean, you sort of gave the broad strokes, right? I was originally a PhD mathematician. This is actually what brought me to Germany – an academic career. I did several postdocs first in Hamburg. And you reach a point where either it’s gonna work or it's not [chuckles] relative to your constraints in your life. So I decided, “Okay, industry it is.” Then that's been about six years. I've tried many roles. I sometimes tell people that looking at my career path looks kind of like I'm a data Roomba, “Let's try this. I go in this direction, I hit a block, spin around.”

3:15

Alexey

**Sorry, a data what?**

3:16

Rosona

A Roomba. You know those vacuum machines that you…? You don't. A robot – a robot that cleans your house. Except that makes me sound like I'm a data cleaner by trade, which I'm not. But I mean, just like… there was intention, which is more than I could say for a robot, I guess, or this particular robot, but I explored a lot of things and each time, I pivoted. as someone on LinkedIn was saying, “Have a goal and then you have to keep pivoting.” Yeah, sure. But you should also adjust the goal as you go. Before you get into industry, you don't really know what it means to be in industry. Anyway, a mixture of things – financial data, CRM data, and then industrial data about three years ago. A contract position for about nine months to a year and then where I am now permanently.

I think industrial data is something that really frustrated me and why I wanted to do this topic is – I feel like we're ignored by a lot of tools and a lot of demos. You get kind of tired of being that person in the audience being like, “But what about our use case?” So I'm here to say, “We have all of these data. It's cool! We have great challenges. If you're looking for a new area, or in the future, what might you do in industry, this is a great place that I think is under-represented, maybe in the space of events.” Also, a caveat I have to say. I'm here as a private person. None of my opinions are those of my employer. I will do my very best not to use profanity, but I make no promises. [chuckles]

4:56

Alexey

**[chuckles] What did you actually research as a mathematician?**

5:00

Rosona

Oh, I meant to have a prop. Can I get off camera and come back? Is that okay? Alright, I'll be right back.

5:14

Alexey

**I'm really curious as to what happens next.**

5:22

Rosona

Originally it was going to be at my desk and I would have had this there. So I didn't plan for that part. This is the shortest answer [holds up 3D-printed plastic item].

5:32

Alexey

**For those who do not see the video because we will also release it as audio-only, what is that?**

5:37

Rosona

I actually don't know. I was thinking more… [cross-talk]

5:39

Alexey

**Those who are seeing the video are also wondering. [laughs]**

5:45

Rosona

These are 3D printed topological spaces. This is an inverted sphere. I think this might be a Mobius band, it might not.

5:51

Alexey

**It’s a Mobius strip, right? [cross-talk]**

5:52

Rosona

Yeah, I think so, but actually, I think it's like an equivalent, but it's not actually the same. Geometrically it's different, but topologically it is. So that's the long answer to the short answer of “I was an algebraic topologist”. I was way off in the far… have you seen the XKCD where he's got applied fields and he's got “sociologist, scientist, blah, blah, blah, blah, blah, engineers,” and then way over here is mathematicians? If you did the same graph, but inside of mathematics, I feel like topology, especially algebraic topology, especially my sub- sub- sub-discipline, is similarly way off the deep end. Like, “Hey, guys, we're over here!” So it’s the study of spaces. Some people call it “rubber sheet geometry”. It was fun. I liked it, I think.

6:41

Alexey

**You were inverting spheres?**

6:42

Rosona

Inverting spheres. There you go. There is actually applied algebraic topology, but that's not what I did. But if you're excited about things like this like, “Hey, this sounds cool. What can I do?” Maybe go that way. I can name my first advisor, who was Rob Ghrist, who then changed to more pure discipline. But he has a lot of cool presentations.

# How mathematics knowledge helps in industry

7:02

Alexey

**So one of the skills you gained is that you know a lot of mathematics. What parts of these mathematics were actually useful for your industrial career?**

7:14

Rosona

[chuckles] Sorry. [laughs] This is not fair. Because it's totally not a question. It's also a little mean, I feel like. [chuckles]

7:22

Alexey

**[laughs] I'm sorry. We can skip that one.**

7:24

Rosona

No, no, no. It's fine. It's fine. It just reminds me. I had this interview with a group of (I'm not gonna name the employer, obviously) but a group of applied mathematicians. I came into this interview and they’re like, “How did you end up in IT?” Like, “Okay, hostile environment…” [chuckles]

7:39

Alexey

**[chuckles] Because I'm curious, it was probably very useful when you were picking up some basics of machine learning. But maybe not all of that, right?**

7:48

Rosona

Yeah. But I think this is also a deeper question of “How is mathematics useful?” And “What is mathematics?” And I think most… [Alexey chuckles] No, I'm dead serious. Give me a second to address this. I think most of what we learned in school of mathematics isn't mathematics. And what really is mathematics to me is thinking logically. What I bring… Really, another question that I like better is “What do you bring, as a pure mathematician, into this space?” And I think, in math, especially in pure math – in a really niche discipline, where people are not willing to repeat themselves, you have one chance to learn something from someone. You're at a conference. There's that fancy guy over there and you're like, “Oh, my God! He wrote this paper (or she wrote this paper). I need to talk to them.” And you have maybe 40 minutes of their time, if you're lucky. Probably 20. Maybe 10. And you have to optimize, the thing I want to know and what questions to ask and like, “Have I answered these questions?”

These are really valuable skills. I think just going in and understanding what questions to ask, how you learn enough about something to sort of keep a model of it in your head, and then (here's the pure math part) say, “How is this going to break?” Because this is the difference between a proof and a simulation. When you try to prove things, you're actually not trying to prove something. I'm not out there being, “Here's my argument for why you should vote this way.” I'm out there trying to first see, “Why will this fail?” and then fill those holes. And that's incredibly valuable and I think it’s the biggest thing I bring as a pure mathematician. It’s this “proof” viewpoint.

9:25

Alexey

**How long did you actually do this for?**

9:29

Rosona

Academics?

9:30

Alexey

**Mathematics, yeah.**

9:31

Rosona

Oh. [chuckles thoughtfully] I mean… You mean from the point I decided to study pure mathematics? That was like two years of undergrad plus seven years of Master’s plus PhD plus six years of postdoc.

9:47

Alexey

**Okay. That's a lot of mathematics.**

9:48

Rosona

It's a lot. This is a full career. I think this is also something that makes it really difficult – jumping to the end of the talk to “What is this transition like?” I think it's very… people don't know where to put you because you're lopsided. You have this whole career in a completely separate space. So you're senior. You're obviously senior. I mean, look at your age. Look at how long you've been working. But you aren't senior in a business sense. There are a lot of things you don't know, right? You come in and maybe you don't know SQL or whatever. You can learn it, obviously. But you're this weird, lopsided human that people have to figure out, “How do I deal with the development of this person?”

10:29

Alexey

**Yeah. If you can invert a sphere, the SQL is like ten minutes, right? [Rosona laughs] It’s very easy to pick up.**

# What is industrial data?

10:37

Alexey

**Okay. So we actually wanted to talk about industrial data. So what is industrial data?**

10:45

Rosona

Let me first make a caveat that I really like this topic. I'm working on this topic. But I think there are obviously people who are super experts on this. So before I offend anybody, industrial data is not a monolith. I think it's a great question, “What is it?” It's very broad. Let me say that the simplest answer is “A productive industry generates the data.”

11:11

Alexey

**That’s pretty much everything, right?**

11:12

Rosona

Not really. [cross-talk] I find that CRM data does not fall into this. HR data does not fall into this. Obviously, we have this data.

11:22

Alexey

**Why doesn’t it fall into this? There is a process that generates data – somebody is hired and there's a new record.**

11:31

Rosona

Sure, but it's not like I have a thing that I'm trying to make. I mean, in industries like the chemical industry, or the semiconductor industry, or our friends in the automobile industry – people who have a thing that they sell at the end of the day.

11:52

Alexey

**Like a physical thing or process or something like that?**

11:54

Rosona

A physical thing that they produce.

11:55

Alexey

**Not necessarily a program.**

11:57

Rosona

Not a program, not a person, not recommendations on a website. An actual physical thing.

12:06

Alexey

**So say I want to create blue paint, there is a process for creating the paint, and there is some data produced by this process. [Rosona agrees] And when you work with this data, this is industrial data. Right?**

12:23

Rosona

Right. And then within industrial data, I just want to… [audio cuts out] R&D data, which is small. This is why there's this like, “Oh, small R&D data,” because that's where I have historically been sitting. But I've been sort of half. I've also done some of what I would call “the productive”. So there's industrial R&D, where experiments can be very expensive, which is why it's small data – because it's expensive. And then you develop a process for a new product. And then there's something like a pseudo-plant where you do this – you run the process with a lot of Q&C, a lot of stopping it, adjusting things.

13:03

Alexey

**Q&C?**

13:06

Rosona

QC. Quality control, where you stop it. Good call, I'm usually better about abbreviations. Right. So you stop it, and you adjust it, and you're developing the process of making the thing. Step one, “Here's the formulation for the thing.” Step two, “How do we actually do it? And what information do we need to record to make sure that it's produced at a good quality?” And then step three is “We've developed that process. We've got it live. It's running.” And then there are live quality controls that you have to do like, “This machine has a part that welds and it needs to be replaced regularly because it wears out. How often does it need to be replaced?” Sort of productive maintenance or planned maintenance kinds of tasks happen there.

I would say there are broadly three large areas. Then within those, there are all sorts of directions people can go. There's agricultural stuff, where data is expensive because it just takes so long. You have to wait until corn grows. [chuckles] Or toxicology – you have to kill a rat (it's kind of expensive to kill animals) to measure the toxicology. I mean, there's a lot of stuff, not even just R&D, but all over. There's a lot of directions.

14:18

Alexey

**Just to summarize, there are three categories of industrial data. First are R&D experiments. These experiments are expensive, so we have very little data there. Then there is production data. That's the other side of the spectrum. And then the thing in between, I kind of missed. What was that?**

14:39

Rosona

I mean, I haven't worked with this myself. I just know it exists. This is how people have explained it to me. I've worked on the two extremes. In the middle is sort of a mock-up of a running process. In step one of your R&D, you've started developing this process. Step two is maybe where you've built the sketch of a plant. Your plant’s there, but you need to… [cross-talk]

15:01

Alexey

**Like a proof of concept?**

15:02

Rosona

Yeah, like a POC. Instead of a whole giant warehouse, you build a room running this process and you're trying to fine-tune how this should work.

# Setting up an industrial process using blue paint

15:10

Alexey

**So if we take this example of blue paint, then the expensive experiment part would be combining different chemicals to see that the shade of blue is the right one. Then the mock-up could be like, “How can we combine it in a more automatic way?” [cross-talk]**

15:30

Rosona

Exactly. Right. “How do we automate this process? How do we get the right volume? How do we make sure that it's always this color blue?”

15:36

Alexey

**Uh-huh. And then production would be… [cross-talk]**

15:37

Rosona

Because volume can affect the color of paint. Also, paint is… I mean, color is a hard problem, as we all know. I've painted a few apartments. [chuckles]

15:49

Alexey

**I don't know why I decided to use this example. I think maybe five or six years ago, I interviewed with a company. It was a chemical company in a very small German village. And that was the only company in that village. It was a huge company – a huge chemical company. And they were doing paints.**

16:08

Rosona

It's a good example. There's a lot of cool stuff with paints. Actually, another reason why it's expensive is that one of the standard tests you do with paints is called the “Florida test”. I think originally, it was like “We leave it out in Florida for 30 years.” But obviously, that's way too long for R&D length. So you paint some paint and you send it to Florida and you let it sit there for I don't know how long. But this is one of those… paint is a big application.

16:36

Alexey

**That's also an expensive experiment, right? If you live in Germany, sending something to Florida… [cross-talk]**

16:39

Rosona

Yeah. Imagine – you have to ship it to Florida.

16:43

Alexey

**[chuckles] So Spain will not work, right? It has to be Florida?**

16:46

Rosona

No. There's something special. You can look it up. I encourage people in the audience to look up the “Florida paint test” or whatever. You can see pictures of these huge fields. They explain it to you like “There’s a special tropical environment, blah, blah, blah.” I don't know the technical aspect. I just know that it exists and I was shocked.

# Internet companies’ data vs industrial data

17:03

Alexey

**Okay. So we already discussed that industrial data is different from the usual internet companies’ data. So the main difference is that there is a physical process that creates a physical thing, most of the time, in industrial data. There is an assembly line that creates a car, right? While in usual internet companies’ data, it’s like a person comes, clicks, and then there is a recommendation and that’s it. It’s all virtual, right?**

17:29

Rosona

Yeah, and I think it's also hard to adjust what data you get once it's productive. There's a lot of things that seem to surprise people and one of them is you can have tons of data on the productive end. Processes are not designed with data in mind, necessarily, which I think is a difference with CRM – you run experiments, you can say, “What information do I actually need about my customers?” You can adjust what information you're collecting very quickly. In industry, you can't necessarily…

If your plant’s in another country, you first need to explain what it is you want them to do. I mean, things like what sensors are available to you. There are also issues where you might have a work floor where you've got a consistent process, but you have machines from four different manufacturers and each machine has different sensors and different positions. How much do you trust your model that you're building with these? First, of course, you have to do cleaning, and you have to sort of pull this data together so that it makes sense as a whole.

# Explaining industrial processes using packing peanuts

18:29

Alexey

**Can you maybe give an example? I don't know to what extent you can. It can be a made-up example of the process of creating something physical. Then, what kinds of things can we observe and record and use for…?**

18:42

Rosona

Sure. I'm gonna make one up. This is a made-up example. I was thinking about this, like, “How do I make something up?” Because I obviously shouldn’t be talking about work. [chuckles] So what's something that we all know and understand? I went with packing peanuts, because this for me illustrates several problems all at once. I have no idea how you make packing peanuts, by the way. I don't even know. [cross-talk]

19:08

Alexey

**What is packing peanuts? You have peanuts and then you want to pack them?**

19:12

Rosona

Oh, no, no. I'm sorry. I thought this was a standard phrase. You know when you open a box that's been shipped to you, and there's various ways to pack things? One of them is with those big cushions of air and one of them is – there's these little Styrofoam things.

19:28

Alexey

**Oh. So it's not actual peanuts.**

19:30

Rosona

They look sort of vaguely like a peanut. That's what we call them – packaging peanuts.

19:33

Alexey

**Ah. I got overly excited. I thought we would be talking about peanuts. [chuckles]**

19:36

Rosona

Sorry. No. [laughs]

19:39

Alexey

**Well packing peanuts is also a process. You need to add salt or roast them or whatever.**

19:48

Rosona

Okay. So this is a production process. I have no idea how you make packing peanuts. I'm just going to make up a process, just because. Okay, packing peanuts. I'm going to assume you extrude them or something. Something is mixing some kind of polymer and then there's an extruder that spits it out.

20:08

Alexey

**What is an extruder?**

20:11

Rosona

Imagine your toothpaste.

20:18

Alexey

**Okay, instead of toothpaste, you have this material.**

20:20

Rosona

Right. And then maybe something that cuts it off. I don't know. Maybe there's mold. Maybe it's a molding process. Somebody is going to come see this with a transcript and tell us next week, “How you make packing peanuts.” But whatever. We're going to pretend this is how you make packing peanuts.

Alright. It gets extruded. Maybe it gets dried. You obviously might collect them together at some point right into a big batch of packing peanuts. Now you have them and you ship those packing peanuts off. But obviously, you need some kind of… you need to measure the quality of your packing peanuts. I don't know what's important. What's important about packing peanuts? That they actually take up space, I guess? That they're not flat?

21:01

Alexey

**Yeah, and they're not squishy?**

21:06

Rosona

Not squishy. That they crumble. That they're not wet, I guess. That would be good. But let's just say I want to make sure that they are not flattened. Maybe flattening is a big problem. You extrude them and they dry wrong. I don't know. How can I measure this? At this point, I've got some sensors I can put in. I feel like visual sensors are always expensive.

I don't know if it's true, but my impression from what kind of data is available is that other kinds of sensors are cheaper and easier to deal with. You might have its weight, I guess. It's rolling along the belt and there's a weight sensor underneath it and it says, suddenly, there was a drop in weight. Okay, what happened? Did the batch of peanuts actually get excluded?

21:51

Alexey

**Maybe this thing that extrudes them can measure the volume? How much comes out each time?**

21:58

Rosona

Sure, yeah. You can. There you go. You can measure with each one “Is the volume consistent? Am I always extruding 10 grams?” Or whatever. Liters? No, milliliters. [laughs] I don't want a 10-liter packing peanut. [chuckles] Right.

22:14

Alexey

**So you monitor every step of this process.**

22:17

Rosona

You might. But you have to decide. There's a person designing this process. This is also the thing – the person designing the process is not optimizing for the data collection necessarily. I feel like it's generally an afterthought. I have had this told to me by someone who runs a huge process that, “Data is an afterthought.” And they're working on it. This is an industry-wide, “Oh, yeah. We should really work on this problem.” It's changing. But in general, historically, when you design a workflow, what you're optimizing for is – you don't want the people doing the things to cross paths too many times. Otherwise, they're going to slow things down by running into each other.

So you design your shop floor full of machines and processes to minimize people running into each other. But that can make data collection and identification and tracking difficult. Why would you track one peanut? You wouldn't track one peanut, right? I wouldn't care about one peanut. I care about maybe 10 grams of peanuts. But then that's sort of tough, because at each stage of the process, maybe they get mixed up. Maybe you have data at the beginning that says this peanut was made with 10 milliliters of gunk. And then later on the process, you have 10 kilos of peanuts. And you can say, I guess, what their weight is. And then “Okay. Well, how do I know what peanuts were in here? How do I pull this data together so that it makes sense all together?” There’s a sort of coarseness and fineness of data and you mix things with processes that mix them together. Like when you're drying them – you could imagine that you put them in some big thing and they dry and then you have some kind of cycle to squish them around, mush them around.

So I feel like coarseness and fineness of data is a problem. How you model the data when you finally want to work with it as a problem. Or challenge, sorry. A challenge. I mean, it's exciting. It's really cool. There's also processes where… I was told sewing (when you cut things) is really messy and it's a really hard problem to solve, “How do you model anything with a process that does cutting?” Because there's so much vibration. Whatever sensors you have are going to have a lot of noise.

# Why productive industry needs data

24:34

Alexey

**Okay. So we have this process. This process produces a lot of data and not necessarily all the data we need. But what I’m now wondering is, why do we actually care about this data? Do we want to make sure that the quality is good? Do we want to make sure that nothing breaks or what's the purpose?**

24:53

Rosona

On the productive end, definitely quality. For the research end, it's “How do I make the best product?” And then on the production, “It's gone live. What am I doing?” Maybe this thing that you're doing gets turned into something else later. This is often the case in the productive industry. I wouldn't buy a semiconductor (I suppose I could) but I would buy something that has this chip in it. What you worry about is, “What does my quality control tell you about the quality control of the downstream product?” That's a big challenge. So this is something “Why do I even care?”

Okay, maybe packing peanuts is a terrible example. But I don't want to ship a box full of flattened packing peanuts, because then they don't do their job. So quality control going out, quality control coming in of the product we're using to make our product. Also, stuff like, “Can I improve processes?”This example where I have a machine that's doing something and it needs to be changed regularly? Often you do this with just a rule of thumb. But maybe the costs are, especially now, so much of a pressure that you really want to optimize this process and change this part one less time a day, if possible. So you need to keep an eye on “Is this process producing anomalous pieces? How good is the quality? Is it within the range I need?”

26:22

Alexey

**Basically, we use this data for monitoring, I guess, at the beginning. [Rosona agrees] For example, we have one kilo or one liter of the material from which the peanuts are produced and we expect one kilo out. Right? [Rosona agrees] Or whatever volume. We can monitor it and we see “Okay. Something is happening. We put one kilo/one liter in but then something is happening, so we have less or more (or whatever) coming out.” Then there are some charts and we can observe this.**

**That's one thing – monitoring the process and making sure that things do not go wrong. And if something goes wrong, some of the metrics deviate from the usual things. You notice that, you detect an anomaly, and then you can do something with this – maybe send a technician to check what's happening.**

27:17

Rosona

I’m just here nodding. For the listeners.

# Measuring product qualities

27:21

Alexey

**The other thing you mentioned is maybe the qualities of this thing. For example, when you squash them, they need to recover their shape. Maybe there is a part of the process that does that and we also record it?**

27:37

Rosona

We should also talk about tiny data. [chuckles] We should also talk about R&D, but this one's much easier because I think it's understandable and faster. But my impression is that there are two kinds of quality controls: live, on the conveyor belt or whatever. During the process, you take a picture and keep going and you leave the thing in the line. There's a second kind, where you destroy the product.

28:07

Alexey

**I see. Okay.**

28:11

Rosona

So that answers your question, which is “How do you tell if this is viscous enough (or whatever)?” You take the thing out and you squash it. [laughs] You'd imagine this giant bin of packing peanuts, and you take your shovel, and you go in and you examine one shovel full of them and destroy that, so that you're only impacting a small percentage of what you're actually producing. Then, of course, there's a whole science around, “How much of a sample do I need? How often do I need the sample?” Statisticians have certainly been working on this for a while.

# How data specialists use industrial data

28:43

Alexey

**But what I was going to ask is – as a data person, what do you do with this data? How do you use it? Do you build all these anomaly detection models that we just talked about? Or is there more?**

28:54

Rosona

Yeah, you can certainly do that. There are people who are adjacent to data science, or maybe in the broader data science family. You can also talk about econometrics, people who are computational scientists – I don't know how they would like to describe themselves – where they actually have the scientific background and have models based on these measurements, “What do I expect about this particular quality of this product?” Certainly, when you have to physically have to reach in and look at things and touch them, it's expensive. Anything you can automate is great.

My experience talking to people has been primarily quality control. I'm sure there are other things right, like predictive maintenance and things like anomaly detection. Why do you care about anomaly detection? Or maybe the volume of anomalies? If suddenly you go from where your process has a 5% failure rate to a 20% failure rate, this is bad. Right? And then it's, but it's not something that you can… it's like a pre-warning. Your role is to give them a flag that says something's terribly wrong. Then someone has to make the decision about what you do about it. Do you stop the machine? Do you stop the process? That also costs money. It really depends on the situation, what people need and what you can do.

# Defining and measuring sustainability

30:24

Alexey

**I have no idea how it works in this industry, but in internet companies (where I work) it works like this. There is a problem, users complain, or we want to improve something. We think, “Okay, what kind of data do we have for that?” And if there is data – good. We just take this data and try to see if we can use this data to solve this problem.**

**If there is no data, we need to collect this data, and then eventually hope that it will be enough to solve the problem. Is it a similar situation with industrial data? Say there is a problem where we need to make sure that the percentage of defects is less than 1%.**

31:10

Rosona

In answer to what you were just saying, I was thinking of a new requirement. Right? Your downstream customer says, “Actually, we really want to track X. We have sustainability requirements.” And then suddenly, it's a new requirement. Then the first thing is, “Does our data tell us that? Can we actually answer your question with our data? And if we can't, what do we need to do? Do we need to add a sensor? Is it just one sensor? Can we get away with one? Can we add a camera? Can we bring in some computer vision guys and gals?”

31:47

Alexey

**What kind of requirements… you mentioned sustainability. This means that the process we have for producing packing peanuts is not too bad for the environment? There isn’t too much emission?**

32:00

Rosona

Let's pick another industry that I'm further away from. [chuckles] Like the airlines, and also the train companies have started saying “We're sustainable,” or whatever. How are they guaranteeing that? I bought something recently that said, “We're 100% sustainable.” Your T-shirt, right? There you go. Clothing is technically also a manufacturing industry. How do they certify that? The person making the shirt is taking fabric and thread and whatever from different customers and putting it all together.

So they have to go to each of them and say, “We want this to be sustainable.” And then they have to say what that means. I think this is a developing label, where people say, “Sustainable means X, Y, Z.” And I don't know what it means. I haven't looked it up. I just threw it out there as a word. Maybe it means it has to be organic. Maybe it means you have to be CO2-neutral. I don't know. So you take these requirements to your cotton manufacturer, and then they have to go and see, “Okay, who do we buy our cotton from? What are the pesticides? What are the blah, blah, blah, blah? Do we even know that?”

I feel like a lot of the problem is that no one thought about this problem at the time they developed the process. And so you have to figure out, “Can I approximate? Can I create an avatar that answers this question given what I have? Or do I need to figure out a way to get that information going forward?”

# Using data in reactionary measures to changing regulations

33:29

Alexey

**Do you have other examples (different examples) for these kinds of problems?**

33:36

Rosona

Right. Let’s do toxicology. A friend of mine did this. Well, not toxicology, but he's doing brain research. But this is all just anything that involves animal trials. Pharmacology. I have never worked in this field, so this is another one where I'm making things up speculatively. My apologies to people who work with drugs. [cross-talk]

33:52

Alexey

**Or paint, right? Paint is not supposed to be toxic.**

33:56

Rosona

No, no, but I mean… No. But paint can be toxic. Right? I mean, lots of things can be toxic.

34:01

Alexey

**But better if it’s not, right? [cross-talk]**

34:01

Rosona

But take drugs, right? You have to have drug trials. I know in the States, we have changing rules on what you're allowed to have in various things. At some point, it came out that BPA is terrible. So suddenly, all the plastics…

34:23

Alexey

**BPA? What is BPA?**

34:24

Rosona

It's a chemical. That’s all I know. All I know is that at some point, we hit this, “Oh, no! This is terrible!” And so, all the chemical companies that made plastics had to suddenly make BPA-free plastics for food-safe things. So there was this sort of sudden reactive change. What’s another example? Right – toxicology. I was just thinking, laws change, regulation changes, “the allowable amount of X in Y” changes regularly.

You have to keep on top of regulations – what's allowed. Then, “Okay, can we still keep selling this thing? Can we still keep producing this thing with the current regulation? Can we adjust our processes so that we have no BPA? Or only 5% BPA? Or we just don't sell water bottles anymore because we can't avoid this?”

35:17

Alexey

**Okay. So there is a new requirement about a certain percentage of certain elements – BPA, or whatever element. Then you think, “Okay. Now how do I measure this?” Then maybe there is a sensor that you can add to one of the pieces of equipment that you have. [cross-talk]**

35:35

Rosona

This actually would be a good transition into small data. Because at that point, if there's a law that says – well, I don't think there was a law about BPA, but whatever – if there's a customer demand or a law that says, “We cannot have this thing in our product,” then you have to go back to the drawing board. You have to go back into research and say, “Can we develop an equivalent product with similar enough properties without using this substance?”

36:09

Alexey

**So how do we do this?**

36:11

Rosona

[chuckles] Experiments!

36:14

Alexey

**[chuckles] Real experiments, right? Not A/B tests.**

36:15

Rosona

Real experiments! I think that's also what's exciting about industry, is that there's real science… [chuckles nervously] I shouldn’t say it like that, should I? [laughs]

36:26

Alexey

**Well I mean data science – is it a real science or not? Maybe not so much.**

36:28

Rosona

Sorry, natural sciences. People who have spent many years in a windowless office blowing things up and losing fingers and whatever. [chuckles] Maybe not losing fingers.

36:41

Alexey

**Hopefully not.**

36:43

Rosona

It's a really different kind of place. I think people really hyper-specialize into their area, which is great. It's really valuable. It's really cool that people can explain to you… they give you a dataset, and you're like, “Man, this is really hard to model this.” And they're like, “Yeah. We're not surprised because blah, blah, blah.” The way that you even define a measurement can be highly subjective. Right. Back to the point. The question? Sorry.

37:13

Alexey

**The question was, there is a new requirement and we need to change the process. I imagine that in an internet company it could be something like, for example, four or five years ago GDPR appeared. All of a sudden, all the processes of collecting data needed to be redesigned or changed. For us, internet companies, that meant going there, changing the code and thinking “Okay. Does the new code we have and the processes we have now satisfy the new requirements?” But it's code. It’s all virtual. So it's just changes in Git or whatever. But in this case, in industry, it's actual scientists going there and starting to experiment with different chemicals or other things.**

38:05

Rosona

Yeah. If you're in a situation – going back to your blue paint. Let's say we had our blue paint, and suddenly, it came out that there's something in this particular blue dye that's toxic when exposed to water.

38:19

Alexey

**Does not pass the Florida test. [chuckles]**

38:20

Rosona

I don't know, whatever. Something happens and we have to remove one of the components of this blue dye. All production on it gets stopped. We go back to R&D. So if it's something like that, where you're redeveloping something, that's an interesting place to start. Because you do also have the old data. You have the historic data, you have the experiments that got you there, and maybe you can reuse some of that data to help you plan your next experiments.

38:48

Alexey

**What kind of data is there? Because it's different from the sensor data we discussed.**

38:53

Rosona

Yeah, super different.

# Types of industrial data

38:56

Alexey

**What kind of data is there? For example, paint or packing peanuts or whatever.**

39:00

Rosona

I'll try to give you a spectrum. Oh! That's a great example, actually – spectra. It depends on what stage of the process we're at, honestly. A year ago, I was working with polymers so I know a little more about polymers. There's a paper, which I think I added to the document for linked recommendations on polymer informatics and its challenges. It’s a really cool field. So if you're doing something with chemicals, whether it's polymers or not, you have ingredients. It’s the same for pharma – any field that has chemicals. Also in drug development. You have the ingredients that you put together. You have the recipe of how you make it.

Like if you're baking a cake. If I tell you, “This cake is made out of… I don't know.” This works with some recipes, but not all of them. Cake is actually chemistry. If I tell you, “It's 5 eggs, 100 milliliters of milk, 100 milliliters of flour.” I made something terrible – some terrible slurry at this point. But if I just tell you the ingredients and you know nothing about baking a cake, you're not going to be able to bake a cake. [Alexey agrees] So there's effectively two sets of data and I think one of them tends to be held secret, which is one of the challenges of industrial data I wanted to come to. There's the data you are allowed to see and there's the data that you are not allowed to see.

40:45

Alexey

**You as a data person – you're not allowed to see some data.**

40:49

Rosona

Yeah. So, baking a cake, right? Sure, the ingredients tell you something. They tell you, for instance, if you are celiac (if you're allergic or to gluten) you know whether the cake contains gluten. So you can certainly do something with this information. But if you want to study the cake making process and figure out if this cake is gonna be tasty or not, I don't think I can tell you from the ingredients if the cake is gonna be tasty.

41:20

Alexey

**But what do you do with this information? I imagine that there is a table, there are different sorts of ingredients with different… [cross-talk]**

41:28

Rosona

Sure, let’s say eggs, flour, milk.

41:33

Alexey

**The amount of eggs… for this cake, you use 100 grams of sugar. For this, you use 200. And then you record everything you have in a database. So what do you do with this? In the end, maybe you predict whether the cake is tasty?**

41:48

Rosona

You also measure your stuff. There are several kinds of measurements you can make. This is a beautiful transition. Thank you for the question. There's material properties, like hardness. There's different kinds of hardness. Viscosity, which there are also different kinds of. There's surface tension. There's things where you can pull it. I imagine I wouldn’t do this with a cake. I wouldn't pull my cake apart. But certainly, if I were an industrial cake baker (they exist) I probably would pull a cake. I probably would measure how much force it takes to pull my cake apart. And then I would try to work out backwards, “Does that mean my cake is delicious?”

42:31

Alexey

**I really want to try it now. [chuckles]**

42:33

Rosona

Because you can't just have people sit down and eat cake all day. You need approximations to that, which tell you something. So there's material properties, like I said the viscosity, or like the color, moistness. I don't know. Material properties. Then there's also “application tests,” which are, “I've put this thing into another thing and I've done something to it.” I don't have a good example… What would a cake application test be? [chuckles]

43:04

Alexey

**Well, eating it?**

43:06

Rosona

Um. Honestly, it would be something like you put it in a box and you ship it. You drive it around Germany for a month with improper conditions and then you measure if it's moldy. I can imagine that'd be an application test. It's more extended. It's not a pure property of the cake. It's something you've done to it.

43:28

Alexey

**Okay, so you have… [cross-talk]**

43:29

Rosona

Everybody's going to be like, “What are you really talking about Rosona?” [chuckles] Yeah, I don't know if I can give you a real example. That’s the thing.

43:39

Alexey

**Let's stick to the cake example. So what we record in our dataset is the ingredients we use and how much of these ingredients we use. Then there are some properties of the process that produces the cakes. And then there are all these properties of the end result of the cake, like what happens if it's stretched. [chuckles]**

44:07

Rosona

I like this example where we drive it around Germany and then we measure if it's moldy.

44:10

Alexey

**All these things – you do all these things and maybe one of the tests is actually eating the cake or trying to cut it and see how much stuff there is on the knife?**

44:18

Rosona

Cutting it is a nice example. Yeah, there you go. That's how you actually test the cake, right? You put the knife in to see if it's done. [chuckles]

44:29

Alexey

**So then I can imagine something like 20-30 different tests and you put all this data in a database. What happens next? What do you do with this?**

44:40

Rosona

Well, first of all. It's there and they do whatever they do with it. I'm not a person who works in a lab. I’m a data scientist.

# Solving problems and optimizing with industrial data

44:44

Alexey

**But as a data scientist, what do you do with this?**

44:46

Rosona

As a data scientist, at that point, the reason that they're coming to you is that they want to optimize something. This is a slight divergence to – if you like mathematical optimization or Villard mathematical optimization, there are tons of those problems in industry. There's also a fight – okay a “gentle disagreement” – on whether mathematical optimization falls under the larger ML/data science/AI world. Putting those aside, it's really important. I had an interview somewhere else where it was like a traveling salesman problem.

We have logistics, we have 10 places where we store stuff – how many trucks do we need to get things from A to B in the correct time? We want to minimize that, obviously, for various reasons. And what route should they take? Planning, graph theory, mathematical optimization – these kinds of problems come in at this point. Coming back to your actual question, it's an optimization problem. They want the best cake. And they come to me and they say, “We want the best cake. Here's our data. We've done our experiments.”

45:59

Alexey

**Do they define what is “best”?**

46:03

Rosona

Yeah. But before that even happens – in an ideal world, let me tell you the ideal. The ideal workflow, which does not always happen, is they first come to you and say, “We want to run experiments such that we can then determine the best cake relative to these requirements.” And if they do that, then you can do a design of experiment.

46:22

Alexey

**I was thinking of a situation, for example, when you bake a cheesecake, sometimes there could be cracks in the cheesecake. After you bake it.**

46:32

Rosona

So “best” could be minimizing cracks, or there's no cracks, right? So you got things that you measured. What do they want to optimize? They want maybe the lowest cost possible relative to the best taste and, the least cracks, the best color. Maybe you want a shiny cake. If you got one of those (I don't know if you watch baking shows) but there are these baking shows where they make really shiny cakes. I don't know how they make them shiny, but they do look really tasty. So maybe you want a shiny cake. And then you can measure the shininess. There was a way to measure that using light and reflection and sensors.

47:06

Alexey

**So you have all these requirements and then you have an optimization problem. [Rosona agrees] How do you combine your ingredients in such a way, or how do you optimize your process in such a way that you still satisfy all the requirements. You maybe maximize some of the things you minimize some of the other things. Is it called linear programming? Simplex method? Things like that?**

47:28

Rosona

Yeah, related.

47:30

Alexey

**That's what you use. Or what do you typically use?**

47:32

Rosona

I'm gonna say we're coming a little too close to things I probably shouldn’t talk about. [chuckles]

# Industrial solvers

47:39

Alexey

**Well, let's say for this hypothetical example of a cake. What would you use?**

47:44

Rosona

No, I’m not going to.

47:48

Alexey

**Simplex method? Would you use the simplex method?**

47:52

Rosona

I think honestly, and I think this is probably obvious that we have tools that help us. There are companies who specialize in doing mathematical optimization and making it super easy and we might make use of their offerings to make our lives easier. Then what are we actually doing? Well, we're setting up the problem. And if there's not enough data, there are ways to deal with that. That's where data science comes in, as well. And then you hand all that to a solver.

48:22

Alexey

**Okay. So there is some industrial solver. There is a company that maybe specializes in optimizing these kinds of processes. For you, what you need to do is define the problem and then put the data there. Then you click a button and watch the solver find the best solution?**

48:41

Rosona

If you only have like 10 data points, you can't feed that into a solver. There's modeling in the middle. And that's where data science comes in as well.

48:50

Alexey

**What kind of modeling is it? Creating new data?**

48:55

Rosona

Alexey…

48:56

Alexey

**That's the most interesting stuff. [laughs]**

48:57

Rosona

I know, I know. But I'm not officially here to officially talk about this. I haven't had a deep heart-to-heart discussion with communications about what I'm allowed to say. And we're recording, and this is live. I have no way to edit this out later if I say something I'm not supposed to.

49:16

Alexey

**Okay. Understood.**

49:21

Rosona

Now I have to figure out if I can answer your question at all? [chuckles] We do have hack… “We” now I'm talking from my company's perspective. There are hackathons that are faced to the outside world. And that's not a bad place to learn about what kind of modeling one does. And there certainly are, with tiny data, the answer is not… This is obvious. It is not a secret that the answer is not going to be a neural net, unless it's something that's been pre-trained and you can do transfer learning. Make good friends and statisticians, if you think tiny data is exciting.

And I think there's no one answer. That's what I also found exciting. We have a working group at work to talk about, “Hey, I have this problem. What do I do with it?” It's not just optimization. There's also “Can we do prediction?” And sometimes the smallness of the data is actually – we have an infrared spectrum, or some kind of measurement that's like a time series. So it looks huge, but it's actually small. And then, “How do I take this data and figure out things from it?” [chuckles] I keep wanting to say that there are all these cool challenges.

50:31

Alexey

**I imagine a binary classification model that says “With these ingredients and this process of baking the cake, will it crack at the end or not?”**

50:44

Rosona

Well, that's another problem. Yeah. If you don't want to do optimization, if you don't want to do “What's the best formulation for this cake,” which is what I was working on, then what you want is, “Given these things, what's going to happen?” “Will the cake crack?” Completely separate question? We have those questions as well.

I think something I was sort of struggling to explain to someone the other day is that with this kind of work, you don't necessarily have a high volume of in-production models. Back when I hung out in the MLOps community, I was trying to figure out, “Can I use this structure for my problems?” But if you don't have high volume things, because you only have tiny data or you have a consulting project to do this one thing, then you don't need a model that’s there and tells you the optimal solution. You have one.

# Tiny data vs Big data in productive industry

51:42

Alexey

**Well, since you mentioned MLOps, I'm wondering – in internet companies, I know that people use Kubernetes and all this stuff, Cloud for deploying the models. I imagine that this is a pretty different situation in the real industry.**

52:03

Rosona

If you go to the productive side, I would say that it should feel to you like everywhere else. So if you're coming from anywhere else in IT. Can we call IT? In the data space. If you're working in the data space, and you're used to how people deploy things, CI/CD, MLOps, whatever, and you like working in that space and you're looking over here saying, “Hmm. Manufacturing. What can I do?” I would definitely say the productive side, where you're generating, depending on the speed of this conveyor belt, if each of our packing peanuts is gonna have a time series associated with it. It's giant piles of data.

At one point, I was so excited because I was using Spark to pivot like two terabytes of data and that was like a day of data. So there is big data and there are these things where people have models and they're doing experiments, and they're constantly deploying and all that. It's there. But I would say with the tiny data, it should feel obvious that that's not necessarily the challenge that you're dealing with. I knew someone who was embedding their model into a software that was being used for visual inspections. There's also deployment in a kind of local sense as well. Not a big…

53:25

Alexey

**I understand. So for the production settings, you probably have all these Kafkas and streaming data.**

53:33

Rosona

Again, I can’t tell you what our stack is.

53:35

Alexey

**It doesn't matter. What I mean is, there are some systems that maybe you will find in other places, like internet companies. Because you want to react to this data reliably. Right? You want to observe this data, you want to make sure all their alerting are set and all the models work in real-time, as the process is actually happening. And then there is the other side of the spectrum, which is these experiments, where maybe you have like 10 data points and you put them in Excel and then you look at them, and then…**

# The advantages of coming from academia into productive industry

54:10

Rosona

You don't even just look at them. I appreciate your summary and I like what you're doing and you're trying to bring us to an end. But I think part of what's really interesting is you have to learn about the data. If you're a scientist now and by scientist, I mean a real… [chuckles] if you're a physical scientist, if you're a chemist, you're a physicist, you're a – I don't know what other sciences we have, sorry, guys [chuckles] – it's a really powerful place to be because then you can learn the machine learning methods and then you have the background. You go into it.

People don't write everything down. Everybody knows when you start making whatever it is – when you start making blue paint – the first thing you do is you heat up the container to 100 Celsius. Or whatever it is. But everybody knows it so we don't write it down. So if you are a chemist who did this thing, then you know that thing. You know all the context, you know the baked-in knowledge. So it's more than just the Excel file. And I've talked to people about that and their challenges with past projects and they're like, “Yeah, we just threw a data scientist at it and got nowhere.” And it's because they didn't ask the right questions.

55:24

Alexey

**And the data scientist had no idea about all this hidden knowledge, right?**

55:28

Rosona

You have to work really closely with your partners. You can't just be a hermit data scientist. You have to be incredibly engaged with your partners and interested about their processes and ask them.

55:40

Alexey

**You cannot just open your CSV file with pandas and then throw it to XGBoost and then…**

55:44

Rosona

No. First you talk to them and you ask them what the problem is, and that you understand it. Then they explain the process to you and what everything means and then you go digest it. You do your EDA, and you come back and you're like, “These things are both called shininess. Shininess 1, Shininess 2. And their ranges are really different. Their distributions are really different. What is this? What is this shininess? Is it a data collection problem? Is it a definition? Do I just not understand the difference? Should I be taking their differences?”

Also, people measure stuff, after 6 days, after 12 days, after a month, and then they view these as really separate measurements, but to me, they're just the same measurement but over time. “Which way should you model it? How do I cut the data? How do I model it? How many models do I need to bring together to actually give what they need?” There's a lot of really cool challenges.

# Materials and resources for industrial data

56:49

Alexey

**There any open materials about this subject? Courses, books, whatever, where people can just go and learn about this?**

56:58

Rosona

There must be. [laughs]

57:02

Alexey

**Oh, okay. I should ask “Do you know any and can you recommend some?”**

57:06

Rosona

No, no. I mean, I'm happy to try to figure out the answer to that question. I wish I had gotten you an answer beforehand. I did put on this document – there's this machine learning repository, which has a very realistic dataset on sensors. It's from the semiconductor industry. You can go play with that, if you want to build yourself into an anomaly detection model for whatever the process is that they're modeling. [cross-talk]

57:35

Alexey

**The** [**UCI SECOM dataset**](https://www.kaggle.com/datasets/paresh2047/uci-semcom)**, right?**

57:38

Rosona

Exactly that, right. That's a short answer. There must be a book. I just don't know what it is.

57:53

Alexey

**Well, we will release this interview in a couple of weeks as an audio-only version with transcripts and footnotes. If by then you find anything interesting, then please let us know and we will add this to the show notes. By the way, I just shared the link in the live chat and I'm going to put this in the description too.**

58:16

Rosona

We’re at the end, but there was a Slido. Did you check the questions? Do you want to do a quick question?

# Women in industry

58:19

Alexey

**Yeah, there was a question. There's one – I don't know, it’s a bit personal if you want to answer that or not. Are you the only woman on the team?**

58:33

Rosona

In my current team? No. In my larger previous team, definitely not. I'm not going to share a picture, but we had a dinner on Friday and let's see… There were ten people and four women.

58:52

Alexey

**Okay. So gender balance is…**

58:54

Rosona

Gender balance is fine. I don't think it's always – it's not always half. I think some of this is, of course, who actually comes out to dinner. But it's not terrible.

# Why Rosona decided to shift to industrial data

59:05

Alexey

**Yeah, we have more questions, but I think we already answered some of them. For the rest it’s “Can you talk more about your decision to work in industry?” It will probably require another episode and another interview. So maybe we can just wrap up.**

59:22

Rosona

Yeah. Hmm. “Why Industry.” You mean productive industry? There's sort of two questions, depending on how you define “industry”. Do we mean just not academia? Or do we mean productive industry?

59:40

Alexey

**Well, I guess maybe both?**

59:43

Rosona

Okay. Not academia because I like it all right here. I reached the point where the next academic position I could take would have been six months with a possible renewal for an extra six months. Moving to another country, learning another language. And I thought, “You know, I'm not… No.” [laughs] At that point, obviously, I needed to do something not in academia. And then why I started where I started is – I applied for everything and I asked everybody, “Do you know people who are in industry?” [chuckles] “Oh, yeah! I know this guy. Oh, yeah.” I talked to him and it's over someone I knew.

60:32

Alexey

**Was it just something attractive to you? You thought, “Okay, this is something I want to work with because it sounds way cooler than your usual e-commerce store.”**

60:40

Rosona

Six years ago, it was “I need a job.” [laughs] And then three years ago it was the way the project was pitched to me, which was, “Hey, we're effectively this little research group in industry and we have all these POCs. We want someone to help us try to figure out how to take them and turn them into productive product projects.” And I thought that sounded really cool. That's what got me into it. And like the people. I have to say, I hung out with a lot of physicists in grad school. If you're listening – Hi. [laughs] There's a lot of physicists and chemists. It's just a different worldview – the people that you end up interfacing with. And in data science as well, the people touching the data and generating the data.

It's just a very different flavor of workplace. My previous employer, the people who worked on the shop floor were walking around in these ridiculous shiny jumpsuits, because of something in the fabric that reduces the static. So you know people just wear whatever, right? People show up in hiking clothes. It’s just very casual. It's much more about the work and the science. It's very… the people. The people – that’s the answer.

61:58

Alexey

**I also guess that it depends on where you are geographically. When it comes to Germany, for example, if you're in Berlin, there are not so many companies in this kind of industry. You'll probably end up working for an internet company.**

62:15

Rosona

There's like a little bit.

62:19

Alexey

**A little bit, but you need to be specifically looking…**

62:21

Rosona

A little further away. Like closer to Dresden or something. There are a few down there. You have to get outside of Berlin itself. Yeah.

62:30

Alexey

**Yeah. Well, I think that's all we have time for today. Thanks a lot. That was a really good discussion. I'm pretty hungry now after talking about peanuts and cakes. [laughs] I should probably go and eat something. Thanks again. And thanks, everyone, for joining us today. It was nice talking to you.**

62:51

Rosona

Yeah, likewise.