1:47

Alexey

**This week, we'll talk about being a professor and leading data research. We have a special guest today, David. David is the director of the Institute for Data Science at the New Jersey Institute of Technology. He is also a distinguished professor at the Department of Data Science in Ying Wu College of Computing. His interests are at the intersection of data science, Big Data, high performance computing and real-world applications, including cybersecurity, massive scale analytics, and computational genomics.**

**That's quite a lot of interests. So what is interesting when I read David's bio was that he co-authored over 300 articles. This is actually one of the topics – one of the things we'll cover today – how is it humanly possible to actually do this? But yeah, that's a bit of a teaser. Welcome, David.**

2:44

David

Thank you. And thank you for that kind introduction. That's right. I'm David Bader. And now I've been, for just over three years, a Distinguished Professor at the New Jersey Institute of Technology, where I launched a brand new institute for data science back in summer of 2019. That institute includes a number of centers, ones related to big data, to cybersecurity, to medical informatics, AI and machine learning and also FinTech, which is a strength of the New Jersey / New York Region.

# David’s background

3:20

Alexey

**I kind of cut your biography a little bit because it was too long. [chuckles] And I apologize, maybe I left out the important part. But your biography is amazing and that's one of the reasons we want to have a chat today. Maybe, before we start, before we go into today’s main topic of leading data research and being a professor, can you tell us about your career journey so far?**

3:46

David

Sure. I'll tell you about my career in brief. I think my full CV is over 100 pages, so I'll try to limit it to just a word or two. But I grew up in Pennsylvania in the United States, and I did my undergraduate and Master's degree in Electrical and Computer Engineering at Lehigh University. Then I did a PhD in electrical and computer engineering at the University of Maryland.

I've held faculty positions at the University of New Mexico, where I was a Regents lecturer and also a professor there. I joined Georgia Tech in 2005 to launch the School of Computational Science and Engineering. I was at Georgia Tech for about 14 years before moving to the New Jersey Institute of Technology, where I am today. So I've spent time at quite a number of universities and have a research career spanning almost 30 years.

# A day in the life of a professor

4:53

Alexey

**That's impressive. What do you do as a professor? You already mentioned things such as launching a school, or some other things. Is this something professors do? What do you usually do?**

5:08

David

Great question. I, of course, do research, teaching in service to the Institute and also to the national and international community. But a typical day is meeting with my students on research and also working with my institute – my faculty and staff – to make sure that we have projects running, new proposals that we're submitting to sponsors.

I also teach, which means preparing for lectures, giving lectures, and interacting with students. The core of what I do is really interacting with students. In my research group today, I have high school students, so they're pre-college. I have undergraduate students, Masters and PhD students, as well as research scientists who have completed their PhD in previous years. So it's quite a diverse group of students, both men and women, and at all levels.

# David’s current projects

6:11

Alexey

**So as a professor, you do meetings with students, you keep projects running, you make sure proposals have been submitted to sponsors, and you interact with students. And some of the things, such as making sure projects are running, that proposals are being submitted, are some sort of project management, right? Your research is a project, and you need to manage that it's actually executed – that people actually work on things. Is this a correct interpretation?**

6:41

David

Currently, I have three projects from the National Science Foundation in the US. One is looking to build out massive scale graph analytics using an open source framework called Arkouda. Another is developing a streaming analytics platform called StreamWare. These types of projects often require coordination among a number of personnel. We are doing research and then we're writing papers about the research that we do. Though, at the same time, we're writing proposals for new projects that will launch after this. I also work with industry quite a bit. I have active engagements today with Accenture labs – we're looking at a cybersecurity problem involving the use of graphs, where we're trying to find vulnerabilities across open source software packages. We also work with other companies like NVIDIA.

I have an NVIDIA AI lab (or NVAIL) award at NJIT and I've worked with a lot of companies in the past as well, such as Intel, Exxon Mobil. I've worked with Yahoo, with Microsoft Research – quite a number of companies. And that's really exciting, to be able to be at the forefront of developments and looking at data science and also at the intersection with high performance computing, and to have ideas that we developed with our students that can then be transitioned into practice, whether it's through industry or startup companies or other types of organizations. I find it quite exciting.

# Starting a school

8:30

Alexey

**This is not something I actually prepared for – it wasn’t in the list of questions – but I'm really curious, what does it take to actually start a school? Like, to launch a school? Maybe it's a very simplistic picture, but I guess you need to come up with a bunch of projects, a bunch of ideas, and then you also need to have connections with industry, because you need money for running and establishing the school. What else is required to start a school?**

9:01

David

That's a great question. Most faculty will join a department at a university and the department's been around for anywhere from a couple of years to decades or even centuries (in Europe) and that's fine. But twice in my career, I've been able to essentially do a startup within academia. At Georgia Tech, as I mentioned, I founded the School of Computational Science and Engineering and here at the New Jersey Institute of Technology, last year, I founded the Department of Data Science. So I've now done startups twice within academia. And what it requires – the first thing that you need – are people.

You need people who are really thinking about new directions. I'd like to think that innovation within academia is really finding interesting work at the interface between traditional departments, and especially in computer and data science. We find so many new areas that are just outside of a single discipline. For instance, my own research in massive scale analytics requires expertise in data science, in high performance computing, in systems, in algorithms, and also in application areas. So we have to weave together many areas of knowledge to be able to produce students who are able to really be impactful as they graduate, and they go on in their careers. So very briefly – what does it take to create these schools and departments?

It takes people and also new academic programs. We spend quite some time thinking about “What does a new program look like, for instance, in data science?” This past fall, we launched one of the earliest Bachelor's degree programs in data science at the New Jersey Institute of Technology. We've had a Master's program since 2017 and we're at the cusp of launching a new PhD program in data science. So I think it's quite fascinating to be able to think about “What does it take to train students to have degree programs in these emerging areas, like data science?” And I hope other universities will also repeat the model that we've created for this training and preparation of students for data science.

11:22

Alexey

**Yeah, now I realize that when I was asking this question, I was more thinking about research labs, rather than schools. But then, actually with schools – the main reason for a school to exist is to teach people. These people then graduate and are then qualified for the job. To start the school, you need to see “Okay, there is *this* university, and then there is *this* area. And for this area, there is no school for this university.”**

**What happens then? You identified this gap, and then do you just approach the university and say, “Hey, how about we just start a data science department?” Or do you first start working there and then you say, “Okay, but these students are really great. Let's start the department there.”? How does it work?**

12:13

David

What I've done is, really, looked at “Where's the need?” And we see there's such a demand right now for students educated in data science. That need is really the main driver, because we don't want to just create programs where students won't be able to find jobs afterward. We want students to be productive as they graduate and data science is a growing area. So first is identifying the need. We'd also look at the regions. It doesn't make sense, for instance, for every university to launch every degree program and have it the same as every other university out there.

We really have to look at “What are the needs of the region?” Here at New Jersey Institute of Technology, over a third of our students are the first time anyone in their family is going to college. We're taking students from a very diverse background, and some who are really new at going into higher education within their families and we're making them well prepared to enter the job market – the workforce – in this region, in New Jersey, New York, and the tri-state area, as well as to be national players as well. Also potentially international, because we have students who will either return to countries that they came from in Europe and Asia, or they may find jobs that are international in nature. So we want to have very well rounded students.

# The different types of professors

13:55

Alexey

**Also, one of the things in your biography is that you're a distinguished professor. I was wondering – what is the difference between just a professor and a distinguished professor? Which one is better? Or can you even say that one is better than the other? [chuckles]**

14:12

David

That's a designation that's unique to every university – different universities have different ranks. Generally, in the US, we have assistant professors – usually they’re before a tenure decision, which is a guarantee of employment. There are associate professors who typically gain that promotion once they're tenured. And then there are full professors who are what you would think of as an “internationally known professor” – one whose research has really resonated internationally.

At NJIT, we have those three ranks as well, but we've added a fourth rank “distinguished professor,” where the bar is significantly higher for that promotion to distinguished professor. Each year, just a handful of senior faculty at the “professor” rank are then promoted to “distinguished professor,” once they are at the top echelon of their fields.

15:17

Alexey

**If I'm interpreting this correctly, a professor without any associate assistants and so on – it describes the type of work you do. I think this is what you told us – you work with students, you work with research projects, you work with sponsors – this type of work is what we can call “professor,” right? And then there are different grades, so to say. For example, maybe an assistant professor would have a small scope, and the next level professor would have a wider scope and so on. Is that right?**

15:59

David

In the US, which is slightly different from the European system, all professors – assistant, associate, and fuller – are typically doing research, teaching, and service. The designation of assistant, associate, full and even distinguished, really is a statement of the impact that that professor has had so far. Normally, an assistant professor may be just out of graduate school, or has done a postdoc and joined a faculty, but they're still very early in their career. Usually, after about six to seven years, a promotion and tenure decision occurs for promotion to associate professor. Then, some faculty are associate professors their whole year – sort of their whole careers.

Others, once they achieve national or international recognition, then may try to become a full professor at their university, which normally happens maybe another six to ten years at a minimum beyond their promotion to associate professor. So it takes a long time and really, it's measured by their impact and the candidate is evaluated by peers from other universities who then write letters as to whether or not that professor has achieved the rank of a full professor. It’s the same with distinguished professors, but the bar for distinguished is even higher than just a full professor at NJIT.

17:41

Alexey

**What does the process look like before the professorship? I assume that it all starts with being a PhD student, right? This is like the “entry-level” role in academia. Maybe a graduate student, and then a PhD student. Then after a student graduates, they can become a postdoc. Right?**

18:08

David

That's right. Some PhD graduates will enter a postdoc. Normally, a postdoc is a limited position from one to three years in the US, where they may join another research group, either at a university or a national laboratory. After the postdoc, they either enter full time technical staff positions, or assistant professor positions within universities. But it's not required to do a postdoc.

There are faculty who joined as an associate professor immediately after completing their dissertation and getting awarded their PhD. The PhD is really an entry degree to do research. Some PhD graduates will join universities, others will join research labs at companies and do quite well in industry.

Industry CVs versus academia CVs

19:03

Alexey

**Just curious – I took a look at your CV (I think we talked a bit about that) and your CV is 106 pages long. In industry, if you listen to any podcast about career and CV recommendations, they will tell you that you should keep your CV at a minimum of one page and two pages max. I think there is even the rule of thumb that you should have one page for every 10 years of experience. It looks like in academia, it's the complete opposite. Is it typical that professors have CVs that are that long?**

19:55

David

I think 106 pages is probably excessive and lynched for academia. But we're expected to list everything that we've done in terms of students mentored, classes we've taught, papers we've published, research projects that we have been the lead investigator on, and so on. And so my CV is just naturally long, because I've done a lot of things. You might have noticed the first page is really a one-page extended biography and that's really my one-pager of highlights if you don't want to read the next 105 pages – that one page is great. But I've had a great career.

I graduated with my PhD in 1996, which was just over 25 years ago, and have been very active. I've led over 90 projects from the National Science Foundation, Department of Defense, Department of Energy, NASA, as well as working with leading companies. And I've graduated quite a number of students, and as you mentioned earlier, published and co-authored over 300 papers. It's been a very productive career, hence the extra pages in my CV. This is a typical format for academia versus industry – you’re right, CVs are typically one to two pages of highlights. But in academia, we're expected to list everything.

21:28

Alexey

**The reason that this usually happens in industry is that the hiring managers – people who decide to hire for a role – receive quite a lot of applications and they simply don't have time to go through every CV. That's why there is this suggestion that if you want a hiring manager to actually look at your CV and read it, then you should keep it at a minimum. But my understanding is that in academia, that's different. People will actually go through and check. The professor is quite a big position, so if you want to get hired as a professor, people take time to evaluate all the work. Right?**

22:08

David

That's right. And it's a privilege to be a professor. The work that I've done – I'm also very proud of the service that I've accomplished. For instance, I been chairing a committee for the National Science Foundation – a committee of visitors for the office of advanced cyberinfrastructure, and evaluating the NSF office that looks at advanced cyberinfrastructure, which includes networks and workforce development, and some of the most capable systems in the world for computing and data science. There's a lot of service that I do as well, that's quite well-documented in my CV. I'm very proud of that work and I apologize if you had to make it through those 106 pages.

23:03

Alexey

**[chuckles] Just curious, out of these 106 pages, how many pages are about your papers? You said you have 300 papers, right? And then you have 106 pages in your CV, so is a third of the CV papers or less?**

23:17

David

Maybe. I didn't look at the actual length, but maybe approximately that. Every time we publish a paper in a journal, or present at a conference, there's another line that gets added to the CV. It's been quite a lot of work, but I've had some great co-authors and students. We normally publish a few papers a year. As you can see, “a few” adds up over a period. That's almost 30 years in length.

# David’s recent papers

23:52

Alexey

**Can you tell us about some of your recent papers? I think you mentioned a few projects that you do. There was a project about graph analytics, right? I assume that this is one of the active projects that you're working on right now. Maybe you can tell us about some papers you published recently?**

24:10

David

Sure. We're just finishing up a paper right now and it's on a framework that we're calling ARACHNE, the Greek word for “spider” and looking at interactive graph analytics at scale. There's an open source framework called Arkouda and Arkouda is spelled A R K O U D A – it's the Greek word for “bear”. You can find this on GitHub. This project started just about three years ago as an open source framework for doing massive scale data science. Often, you may find that you have datasets that are terabytes in size, maybe tens or dozens of terabytes, and no existing enterprise framework is able to interact with datasets that large.

We have analysts who want to be able to run queries. They're trained in Python, they like using NumPy, and Pandas. And what we've tried to do with Arkouda is develop a framework that is able to take an analyst who knows Python, and substitute out NumPy for Arkouda, to be able to look at running where the dataset may sit in a supercomputer on the back end because of its size, and then operate in near real-time, just like you're in your Jupyter notebook and you're running Python and you issue a command, you want it to return fairly quickly. It’s the same way here, we don't want to wait hours and hours – we want a near real-time response.

We want the productivity of Python, with the performance of a supercomputer and that's what Arkouda is providing. Now, as I mentioned earlier, what we're building out in Arkouda is a sub-piece of the framework called ARACHNE for graph analytics. Often our datasets represent graphs, where we have relationships between entities, and these graphs can come from system logs – maybe we're doing some cybersecurity analysis of our syslog. It could come from information about our customers, it could come from social media.

Many, many sources generate large volumes of data and we want to be able to manipulate these datasets running graph analytics, such as connected components between a centrality breadth-first search and others. We want to count triangles, compute clustering coefficients, find K-trusses, run new centrality measures, like triangle centrality, and we're building out the analytics to be able to do that. This has been joint work with my students. We're publishing a paper coming up in September at the IEEE HPEC conference (high-performance extreme computing) that will be held in Massachusetts in September. We're really excited about this work.

# Similarities and differences between research labs and startups

27:10

Alexey

**What you described sounds like a typical startup. What you said “Productivity of Python with...” What was it?**

27:22

David

With supercomputer performance.

27:24

Alexey

**Exactly. It's like a really good elevator pitch, right? Does a research group have to be like a startup in academia?**

27:35

David

Right, the research groups are like a startup. That's a great analogy. What we want to do is really have an impact. Instead of just publishing papers, we also produce code and it's open source on GitHub as well. I should mention that the productivity front end for our work is Python. Our users really use Python within Jupyter Notebooks – very similar to any Python developer. We're doing all of the hard work to be able to bring in a supercomputer in the back end to make it seamless, so that you don't need a heroic programmer. You don't need to even know that there's a supercomputer back there.

We're trying to democratize supercomputing and make it easy. We are leveraging an open source compiler framework called “chapel,” spelled C H A P E L, that Cray developed under a DARPA program about 20 years ago, and is now supported by HPE that acquired Cray recently. So the HPE/Cray Chapel compiler is the framework that we're using in the backend to be able to run, whether we're on a laptop, on a cluster or a supercomputer – we're able to leverage this compiler framework to get truly high performance for the backend, where we do all of that hard work so that our user can just call a Python function and get the result and not even know all of the complexity running with a supercomputer in the back end.

29:09

Alexey

**I guess the difference between a typical startup and a research lab is that you actually have a research lab, right? You need to publish papers and you keep your research open. This is really good, because not every startup company would just open source their know-how.**

29:32

David

That's right. We're very much like a startup in that we have to acquire funds for supporting our research lab and students. And we are also pushing out code. But our real deliverable is producing students who are educated and able to contribute in the workforce and also the papers that disseminate our ideas. Those are the prime deliverables that we have.

I have been involved with startup companies, so I've seen also from the other side, creating some new value and entrepreneurship. I’m quite excited by the work that I've done either advising or launching startup companies as well. So I love both sides. I love the academia side where everything's open, and I also love the motive of startup companies, taking some new idea and getting it to the market and really impacting people's lives.

# Finding (or creating) good datasets

30:30

Alexey

**You mentioned that you need to work with datasets that terabytes in size. In our community, we have a course about data engineering and sometimes it is a problem for us to find a good dataset for a project. Can you recommend some of these datasets? You mentioned that there are system logs data sets? Are these datasets even open? Are there good open datasets?**

30:58

David

Most companies and organizations have massive datasets. Often, what we do in the research lab is either use synthetic data sets that we create or use some of the repositories online, where we find datasets that model social networks or other types. For instance, in our work, we look at graphs – and Stanford has a very good set of datasets called SNAP that has graph datasets. We use many of those.

But if you work with a company, or any organization, they'll have terabyte-sized datasets and we're trying to train people to use those. I don't imagine that they're going to be opened and given to researchers. I think we have to go to them. In the past, what I used to do is create our code on synthetic datasets, and then work with companies by taking my code to the company and then running internally on their datasets.

32:00

Alexey

**It's a challenge to find a dataset that looks like datasets from industry – not something pre-cleaned and already a CSV file, like with all the data in one CSV file, compressed, and it's like two megabytes. But rather something that looks like real-world messy data. That's a difficult thing to find.**

32:21

David

Correct. That's hard to find within academic research, but very easy to find within industry, in large organizations. We often work with many organizations that have these types of massive, massive datasets.

# David’s lab

32:38

Alexey

**How large is your lab?**

32:41

David

My lab right now – I have a Principal Research Scientist (a senior PhD) – I have about three PhD students, about two or three Master's students, two or three undergrads, and then about two or three high school students as well. [cross-talk] The boundary between our summer semester and our fall starts and so I have some students graduating and then the numbers are plus or minus one as students graduate and new students join.

33:18

Alexey

**That's why “two or three Master’s students”.**

33:21

David

Correct.

33:22

Alexey

**It's back to school now. Right? It starts soon.**

33:25

David

Our first day of classes is in about a week. I'm quite excited. I'll be teaching an introductory class to Big Data for graduate students. I love teaching that class and really training students in data science.

33:41

Alexey

**That's the New Jersey University, right?**

33:45

David

Right, at New Jersey Institute of Technology.

33:47

Alexey

**Institute, yeah. I'm sorry. Is this publicly available? Or is it just for the students of the Institute?**

33:54

David

It's just for students. Students have to enroll to take the class.

33:58

Alexey

**Right. Yeah. So these are the people – the lab – that is doing all this research in graph analytics. Right?**

34:08

David

That's right. I should mention that that's my current lab, but I have quite a large number of alumni who've gone on to the bigger and better things. Some are faculty at other universities, some are working in major companies like Google and Microsoft and Facebook, and some are doing startups. I have students who are now across many different sectors and geographically all around the world.

34:36

Alexey

**So I guess there is this natural cycle. A Master’s student comes and then they spend two years working at the lab, some of them stay as PhD students, but I guess most of them go and apply for PhD in other universities and then you have some PhD students coming in from other universities. So you have this natural cycle of people coming and working for a couple of years and then leaving, and then new people come in, right?**

35:07

David

That's right. As a research lab in a university, I'm used to a very dynamic and changing workforce. I maintain a very diverse set of students and I have them (as I train them) for usually just a couple of years. Maybe a Master's student up to two years, a PhD student maybe three to four years, and then they go up and take the next stage of their career, whether it's in school or doing research at a company or university.

# Balancing research and teaching as a professor

35:39

Alexey

**You said that you also teach an introductory course to Big Data, right? I'm just wondering – how much time do you spend on teaching versus doing research?**

35:49

David

That's a great question. Because I'm research active, I usually teach just a couple of classes per year. Normally… maybe I would estimate about a quarter of my time is spent on teaching. As mentioned earlier, I direct an institute for Data Science at NJIT, and I have four centers and one research thrust that report to the institute. There's about 40 faculty around NJIT that are part of those centers. We have activities related to the Institute.

For instance, a weekly Virtual Data Science Seminar that gets broadcast through YouTube. We have other activities to bring students and faculty together. So that takes quite a lot of time. And then I have my own research as well, in my research group. I do a lot of service. For instance, right now I'm the editor in chief of the ACM Transactions on Parallel Computing, and some other service roles that take my time as well, whether it's inside NJIT, or for the betterment of the broad computing and data science communities.

# David’s most rewarding research project

37:06

Alexey

**We have a question from the listeners. The question is, “What is the most rewarding research project for you that you have done?”**

37:14

David

What is the most important research project? [Alexey corrects] Oh, most *rewarding*, sorry. Most rewarding. That's a great, great question. Huh. I’ve done so many research projects. Maybe I'll mention my highest-cited paper. It is one on finding an algorithm for the linear time distance between sign permutations. Let me just describe this a little bit more in laypersons terms. There's something called the “pancake flipping problem,” where you have a stack of pancakes of different sizes, and you want to count the minimum number of times you can put a spatula into the stack and flip them over to sort them from biggest on the bottom to smallest at the top. This problem was one that Bill Gates actually worked on when he was an undergraduate student at Harvard with Christos Papadimitriou. He opened the door for solutions for looking at this problem.

Many years ago in my career, I looked at a very similar problem related to this, where instead of a single stack of pancakes, you actually have a circular stack that you're putting two spatulas in and then flipping sections. That's called an inversion and it's a very useful mechanism in biology, looking at evolutionary histories. So I worked on this problem and the algorithms from the best theoreticians were extremely complex, some head running times, like order n4, order n3. And as I worked on this problem, we discovered that you could solve this problem in just a couple lines of code in linear time – true linear time – and it was very easy to implement. The only data structure it used was a stack.

This problem was something opened by Bill Gates and we closed it – and we did it in such an elegant and simple way. So that's one of the most rewarding examples that I have of being able to work on a problem that many others have worked on before, but getting that spark – getting that innovation – that takes a problem that used to be really, really complex, and making it almost trivial. That paper has been cited hundreds of times now, and it’s just a delight for me to see that we can still improve things, even if they seem like others have worked on it and got so far, there may be a different way of solving things, or there may be a new thought and we can make these *big* discoveries. To me that work is something that I really found rewarding to do.

# David’s most underrated research project

40:29

Alexey

**I guess when people appreciate and cite your paper, that's good. But sometimes it happens that you put a lot of effort into something, you really liked the outcome, but people just don’t notice it. Or maybe like one or two researchers cite it and then that's it. Are there papers like that, which you wish more people knew about?**

40:53

David

Another great question. One of the works that we've done, it's been cited a number of times – it received the Best Paper award at the IEEE HPEC conference (high-performance extreme computing) – and it was work that we did on a framework called STINGER. STINGER stands for Spatial Temporal Interaction Networks and Graphs. Essentially, this was a foundational paper looking at analytics when your data is in motion and you can form that data into a graph. We described one of the earliest processing frameworks for streaming graphs.

I believe the paper is now about 10 years old and we have had a lot of developments since, as we ported that work to GPUs and accelerated it, it became a package called a Hornet, and then cuGraph that you'll find in, for instance, NVIDIA’s RAPIDS AI framework for data science. The graph analytics are actually based on some of our work with STINGER. I've been quite excited by that. It's been somewhat of a niche, looking at streaming graphs, I think they were a little bit ahead of their time. Now, it's become mainstream to look at graphs and especially streaming datasets. But I really enjoyed that work and working with the students, as well, that helped make it possible.

42:13

Alexey

**I think this is actually an emerging topic. If you think about cybersecurity and like fraud detection, this is actually a graph, but then you get a stream of events and then you somehow need to build a graph from this stream of events. And you need to be able to do this fast – if there is a fraudster, you want to catch them as fast as possible. Is this how it’s used?**

42:40

David

Correct. I've been involved with parallelizing graph algorithms since the 1980s. I don't know if you even remember the 1980s or were born yet, but… [chuckles]

42:51

Alexey

**I was at school in second grade. So I don't remember that. [laughs]**

42:55

David

I've always been interested in graphs, but what has really taken off is the fact that there are tables and databases that you can't do a join of those tables because the space requirement really blows up. So we move to graphs because you could form a graph between relationships and your trading the table joins for traversing through vertices in that graph. There are many problems – as mentioned, from cybersecurity, from biology, from social network analysis – that are amenable to graph representations.

There, what I do is take all of the raw data, which is really relationships and attributes about objects, and all of the objects become vertices in the graph, and the relationships become edges, and those relationships could have attributes, they could have timestamps, there could be directions on those relationships, there could be an ontology (or not) that's associated with it. But these graphs really give us a raw and natural representation for many things that we see in the real world. And so I abstract away our problems to graphs and then I solve the algorithm that we're looking for within a graph framework, then map it back to the application domain.

But these graphs – we've been doing this, as I mentioned, for decades – it's now becoming mainstream as more data scientists realize the power of graphs. So I'm really excited by this shift and all of the frameworks out there that have given us great capabilities for processing graphs.

44:35

Alexey

**Yeah, that's interesting that you mentioned that. Maybe you know, with deep learning – there were some researchers in the 90s that did some of the work on deep learning and nobody really recognized their efforts until it was actually the right time. Now (or like 10 years ago) people realize that there are these GPUs that could be used and then all of a sudden, deep learning became popular. Now these researchers who started the research way, way back – now they are very well-known. Probably the same thing is happening here. [cross-talk]**

45:06

David

That's right. Everything comes full circle. I think it's cyclic and we see things rediscovered, and it's great. We have new capabilities now. I think what's also different from when I first saw neural nets in the 1990s, now that we see it, we have more computational capabilities and we have data sets that are available. Whereas before, we didn't have as much data. So I think we've had the perfect storm of data sets, computational capability, and then really bright students who are looking to do this type of research.

# David’s virtual data science seminars on YouTube

45:45

Alexey

**You mentioned that you're doing some seminars and you broadcast them to YouTube. How can people find these seminars? And what do you actually talk about there? Do you talk about things like we discussed now – like graph analytics and things like this?**

45:58

David

That's right. For the past two years, I've had a virtual data science seminar series during the academic semesters, Wednesdays at 4pm eastern time, if you want to join it live. You can find those – the previous seminars that we posted – on YouTube, if you look for “NJIT data science,” you'll find our channel that's got all of these rich (dozens of) talks that we've recorded. We're still planning our fall semester. We're gonna launch our seminar series soon. So stay tuned. But if you subscribe to our YouTube channel, you'll be able to get access, as those talks are live. Also, you can see the old talks as well. For those that join live, you can either watch the broadcast on YouTube, or join our Zoom by registering. And there, you can interact with the speakers as well.

# Teaching at universities without doing research

46:52

Alexey

**We will make sure to include links in the description. I remember that I prepared a lot of questions for you. One thing I really wanted to ask you about is – I like teaching, but if I want to go to university and I want to work at the university, it feels like it's kind of expected that I also do research. You mentioned that you actually devote 25% of your time to teaching and the rest, I think, research and all these coordination activities. My question is, “Is it possible to join a university just to teach students? Or is university not the right place for that?”**

47:44

David

I have to apologize. I'm in New York City and something loud just passed by the street outside. I caught most of the question, but can you repeat it just one more time?

47:55

Alexey

**Can I teach at university without working…? Do I always have to take part in research if I work at university? Or can I just work there and teach?**

48:05

David

Great question. There are many different types of universities. Some are research-oriented universities, where faculty are expected to do both research *and* teaching. But there are other universities that are focused solely on teaching. Many, many universities are like that as well. Some even focus just on undergraduate students.

There are some students who get a PhD and go to a teaching school, where they teach undergraduate students. That's fantastic as well. We typically have a classification for universities in the US called the “Carnegie Classification.” There, there are research-extensive and research-intensive schools. But also we have teaching schools as well. There are ample universities – hundreds upon hundreds of universities of all of these different types.

# Staying up-to-date in research

49:01

Alexey

**But I guess if you join a university just to teach, then what may happen is that in five or 10 years what you teach becomes obsolete, right? So you need to somehow know what's the cutting edge and that's why you need to do research and teach at the same time. Is that right?**

49:18

David

Correct. In computing and data science, of course, everything becomes old quite quickly. I've had to relearn and reinvent every few years to stay on top of what it means to do computing. For instance, as an undergrad, for me in the late 1980s, probably none of that technology… We'd see it in a museum today. But the concepts are still very similar. So I've had to stay on top of the technology, but the foundations usually still remain the same. Still, when you're teaching, there are great ways to stay on top.

You can read publications, for instance, from professional societies. There are typically journals and magazines that help you stay on top. You can also follow research, even if you're not doing it yourself. Usually, there are papers that are very accessible. I love to read and I love to stay on top of many different areas, because who knows what we'll be doing in five years from now? How the world could change – we'll be here and talking about quantum computing and new trends that are emerging as well.

50:37

Alexey

**But how do you even find time to stay on top? Not only have you published 300 papers, meaning that you are busy all the time writing these papers, co-authoring, managing students and so on – how do you even find time to read papers? I imagine that you need to read a lot of papers – far more papers than you write. Right? [chuckles]**

51:03

David

Right. So I'm sitting here, and I always have a journal next to me or some papers to read. You know, it's fun. I love reading and staying on top of it. To me, it doesn't feel like work. I can't believe that I get paid to do this. It's a lot of fun. It's exciting to look at what's happening out there. I often have my phone close by – I call up colleagues and see what they're working on. I will meet up with colleagues and ask about what their research is, even if it's in a completely different area, like architecture, or physics, or in the humanities. It's just great to be able to interact with others.

At the end of the day, what I want to do is make the world a better place. I want to solve global grand challenges and do real-world problem solving. And that requires not just knowing everything in my niche discipline in computing and data science, but really knowing how the world works and what can I do to solve problems that really matter to people and populations around the world?

52:07

Alexey

**Do you have a favorite mailing list?**

52:10

David

Do I have a favorite what?

52:13

Alexey

**A mailing list. How do you know which papers you want to read? Do you just, I don't know… there is a conference, you open the schedule for this conference and see what's there? Or do you have some sort of mailing list that you follow?**

52:28

David

I guess at this point in my life, email is overwhelming. So I don't have a favorite mailing list. Often, I'm trying to triage email to find the very important pieces to respond to versus the advertisements and this and that. But I typically read the general magazines. For instance, from the IEEE Computer Society, there's the IEEE Computer Magazine. And then from the ACM, there's Communications of the ACM. Those typically have some great summaries that will give me pointers to maybe papers in journals and conferences that are something to pay attention to.

Then I'll watch my favorite conferences in the areas of data science and parallel computing, high-performance computing, and I'll track what's happening. I'll attend some of those conferences. Now, it's easy being able to attend many things virtually. I'll scan the agendas and see what's interesting – what the trends are and what catches my eye. I just stay on top by trying to follow it. Again, it's a lot of fun. I have some books that I'd love to read. But I also like being able to read these professional articles as well.

# David’s favorite conferences

53:52

Alexey

**You said you prefer to go to your favorite conferences. So what are those? I know there is one called SIGGRAPH. Is there such a conference? SIGGRAPH?**

54:03

David

SIGGRAPH is for computer graphics. I…[cross-talk]

54:07

Alexey

**So it’s not related to graphs, right? [chuckles]**

54:10

David

Right. That's more visualization and graphics. SIGGRAPH is the top conference in that area. I'm often doing high performance data analytics and go to conferences like Supercomputing from the IEEE and ACM, or IEEE HPEC, which is high-performance extreme computing. And one of my favorite conferences is from the IEEE called IPDPS, the International Parallel and Distributed Processing Symposium. That's been one of the longest-running conferences in parallel computing, where I have my main community there. There's other conferences as well. There's quite a lot that are blossoming in the area of data science, and I'm excited to see where those go as well.

# Selecting topics for research

54:58

Alexey

**How do you select topics for research? You read these summary papers and think, “Okay, this is actually something I can contribute to.” Or how does it work?**

55:07

David

That's a great question. Many faculty look at an area and say, “Hey, what can I do in this area?” I'm probably somewhat of an outlier, where I first want to find a person in another discipline who may be struggling and they don't have the computational capability or the data science tools that they need to solve their problem. So I first go to find “What's the need?” And then I really look at their domain in detail, and “How can I help enable them to solve their problem?” I think that's the way to have more impactful research, rather than just creating something that I'd love to do but maybe nobody else will be interested in.

I always try to think about “Who needs this?” And “Can I help them?” Earlier in my career, I worked with geographers as well as with many computational biologists and those working on Genome Sciences, where they had datasets and problems that they'd like to solve, but didn't know how to do it. They didn't have the algorithms, they didn't have the right data structures – whereas I could help assist them. Through the course of my career, I've repeatedly been able to work with domain scientists and help them solve the problems that they have. By doing so, I get to publish some great computer and data science papers. But more importantly, I get to solve real problems, where it makes a difference to a scientific and technical committee out there.

56:40

Alexey

**I guess that's the recipe of how to get such a long list of credentials like you have, right? Just start with a need, find somebody who’s struggling, find out what they’re struggling with and offer them tools, and then work together.**

56:55

David

That's right. You know, at the start of my career, I thought it took longer to do my research. For instance, NASA, the space agency in the US – I did my PhD and I had a fellowship and worked on problems from NASA on satellite image processing. And I remember working on that problem, where there are a number of academic papers, but the academic papers cut corners and abstracted away from the real problem. I wanted to build a system that real NASA scientists would use.

It took an extra effort to make sure that I was scientifically valid, that the results were quality checked and controlled, and that I was solving the problem – the *real* problem – not just a computer science abstraction that I could publish a paper on. It took a lot more to create the right interfaces and to really maintain all the science that was in that code. And I've had to do that multiple times in my career, when I build systems for biologists or others, where I really have to make sure that it's a system that has all of the corner cases, all of the complexity that's in the real data and the real problem, rather than just writing a paper that I can publish and put on my resume, but no one will ever use.

# Convincing students to stay in academia and competing with industry

58:11

Alexey

**And maybe the last question. I remember when I was a Master’s student, when I almost graduated, my professor called me to his room and said, “Okay, we're doing this cool research. How about you join us and work as a PhD student?” I thought, “Okay, the salary is not that great. How about I think for a couple of years and work in industry?” Then I did that and didn't come back.**

**As a professor, do you have this problem that you need to compete for students who are maybe motivated not by research, but by things like money? They don't stay in academia, they don't pursue a PhD – they just work on something like running SQL queries and calculating click-through rates.**

59:09

David

I love attracting PhD students to this research. That's why I'm here talking with you and I'd encourage anyone who's interested in this area to come seek me out at NJIT. You could do a Google search and find us. I'm looking for some great PhD students. There is always a competitive market for PhD students in research and there are many different areas that PhD students can work on. I've had PhD students that do their whole research with me. Others work with me for a few years and then they may find someone else where their research is more exciting – and that's great for them to find a research topic that they can do their dissertation on and then become the world expert in.

I’m always looking for fantastic PhD students. I have a great lab. It's very diverse. I've men and women and have been able to graduate quite a number of students over the years. I think, for students, there are a lot of choices to be made, especially for research and that there's probably a research lab for everybody. No matter what your interest is, you'll probably find a person that is working in that area. I encourage students to really look at the professors, rather than looking just at the university name – look at the professors and what research they're doing and decide, “Do you want to be an expert in that area?” And “Who can you apprentice with to do your research?” Again, there are many funded PhD positions that we have. Students are typically supported while they do their PhD, so they get a stipend and their tuition paid. So what better way could you do your graduate degree than a funded PhD position and come out and be the expert in your field?

I have had my students go on and, as I mentioned, some are faculty at Penn State, at University of Florida and other places. And others who are now leaders at research at major companies where they're really the thought leaders. So it's exciting. I should mention, I was an undergraduate student when I first did research. I got involved as an undergraduate with a faculty members research program during a research experience for undergraduates. And that's when the bug hit me. It was like, “My gosh! This is so fun!” I had no idea what research was until I spent a few summers working with that faculty member. I'd encourage *all* students to think about a research experience, find a faculty member that you think their work is interesting, and see if you can work in their lab. I'm sure that they would love to have you and I think once you touch research and you see research, maybe it will lead to a lifelong career.

62:02

Alexey

**The example I gave you at the beginning – of my professor talking to me and trying to sort of convince me to stay – do you need to do this? Do you need to compete with industry? Or you don't have this problem because there are enough motivated students – not necessarily from your group, but coming from elsewhere, who want to join your group?**

62:26

David

That's a great question. Often our Master's students are with us for about a year or two and many of them go to industry. A few will continue on for a PhD, either at their current institution or another. But the PhD students typically want to get a PhD because that's the entry level to be a researcher within some of the top industries and companies. We work with many companies that want to recruit our students, but they want them at particular levels.

They may not want to take the PhD students before they're completed to make sure that they have a PhD versus the Master’s students, who are more readily accessible to industry. I haven't really had to compete against industry. In fact, often we collaborate with industry, finding shared research topics that our PhD students can do. What better way to train them so that when they do get their PhD, they have a company that's ready to hire them and continue that research? It's really more collaboration at the PhD level than a competition with industry.

# Finding David online

63:31

Alexey

**Okay, makes sense. If anyone has questions and they want to reach out to you and ask them, what's the best way to do this?**

63:40

David

If you find my webpage – DavidBader.net – there's a contact form where anyone can put in a question, comment, or ask me and I’ll reply.

63:51

Alexey

**I guess this is the page where I found your CV, if I'm not mistaken.**

63:55

David

That's right. You'll find 106 pages if you'd like to read it, along with some advancements and copies of the papers and other fun stuff at the website.

64:06

Alexey

**Maybe by the time we release, it will be 107. No? [chuckles] We’ll see. Okay. Thanks for the chat. Thanks, everyone, for joining us today, for asking questions. I didn't cover two questions. My apologies for that. But I think we will stop now. I just want to thank you again, David, for joining us today, for sharing your experience. For me, as somebody who is working in industry, this is an entirely different world. Now I have some idea of what exactly you do there in academia. That was very interesting. Thank you.**

64:41

David

Thanks, Alexey. Great to talk with you and I hope your listeners really enjoyed the conversation.

64:45

Alexey

**I'm sure they did.**

64:49

David

Have a great day.

64:50

Alexey

**You too. Bye, everyone.**