Andrew Saintsing: Hi, you're tuned into 90.7 FM KALX Berkeley. I'm Andrew Saintsing, and this is The Graduates, the interview talk show where we speak to UC Berkeley graduate students about their work here on campus and around the world. Today I'm joined by Lucy Li from the School of Information. Welcome to the show, Lucy.

Lucy Li: Thanks for having me.

Saintsing: It's so great to have you here. And we're also joined by Katie Keith, who is not a student at Berkeley. She actually just graduated from University of Massachusetts Amherst in a very similar department to Lucy, but they do a podcast together. We're going to talk about their podcast in a little bit. Welcome to the show, Katie.

Katie Keith: Thanks for having me.

Saintsing: So, great to have you both here. So, I just wanted to start: you're both in information programs. What is an information program in particular?

Li: So, I actually think of myself as a computer scientist. I have a master's in computer science, and when I was looking for grad schools, I was kind of looking based on like what I wanted to do, who did similar work as me, and it turns out that some of the people who did similar work that I wanted to do were at information science departments or information schools. And when I was looking around, I was like, "What is information?" And in fact, in the School of Information we have times where they sit us all down and we like spend an hour talking about "what is information?" And then people give all these ideas of what information is, and I guess for me the information that I look at is language, so it makes a little bit of sense as to why I'm at an information school.

Keith: I think another thing that kind of connects digital information to like traditional information schools is that information schools have a long history of being library schools. So, library schools have books that have like text and information in them. And so now a lot of that is digital, and when things go digital, they scale up. Data is bigger. And so, sometimes we'll want to use computational methods on them. At UMass, right before I got there, they split themselves off to the College of Information and Computer Sciences, where it used to be just the Department of Computer Science, and my rough understanding was, it's kind of a rebranding in that data science is really hot right now. Informatics is really hot. Also digital humanities and computational social science, which overlap with what Lucy and I do, is also gaining interest in momentum. So, the idea I think with branding yourself as an information school, it covers a lot more than just pure computer science. When you think about computer science, I think you typically think about algorithms, AI systems, and information. I think in my mind kind of lends itself more to the data science and statistics side of things. So, it's kind of a larger umbrella I would say than just purely computer science.

Saintsing: Okay, I see. And so Lucy was kind of mentioning language also, and so I guess the idea I'm taking away from that is that information is kind of things that are being communicated. Is that kind of a correct assessment?

Li: Kind of. I don't quite remember, but I remember we had this whole really long conversation in one of my classes once where we were like, “What is the difference between knowledge and information?” And I don't remember the answer to that question, but even the question itself is interesting to think about. And one thing that I think sets apart a lot of information programs from computer science programs. So, some schools don't have an information school or a separate program, it's kind of just like sprinkled throughout everything else. But some schools do have a separate program. So, like Berkeley is an example, or Cornell is another example. And I feel like students in information science departments often care more about people than people in computer science departments, which is like maybe like a gross generalization, but like it's even just like on the school information page, it's just like “looking at people and technology” or something like that. I'm probably misquoting that, but “people” is there, and so I think a big component to us looking at information is the people being involved in information.

Keith: Yeah, and I would just jump in and say there's this term digital trace data, and I don't know who coined that term, but I know Matt Salganik, who is a really big name in computational social science from Princeton. He wrote a book *Bit by Bit* that talks about computational social science or social science in the digital age. And I love this idea of digital trace data because that's the sort of found observational data that you can say is really in my mind propelling a lot of this information science and information schools. And so, that's think about all of social media think, about every time you click on a website. Think about all this data from your mobile phones, anything that's passively being collected all the time. It's massive scale, so you need to have understanding of large-scale databases and data science and computer science in order to automate analysis of really, really big data. But it's also that social human connection, like Lucy was saying. Because it typically involves human behavior.

Saintsing: So, you mentioned people, social. Would you say that whereas knowledge is something that someone could have to themselves, kind of information is something that is necessarily being exchanged over whatever platform that you're looking at.

Li: Yeah, I think so, and we actually… the cool thing about sitting in the school information is that like, even though I'm trained as a computer scientist, there's a lot of other students who are trained as sociologists and they come from such different like disciplinary backgrounds, and you're just like in the same room as them. And so, you get all these connections to other types of work and even other types of methods. So, like Katie said that there's a lot of computational methods out there, and that's mostly what I work in, but we also have people who are doing ethnographic work. So, they do things like digital ethnographies, which I don't know too much about, but I think it means that they're like basically doing ethnography but in like online spaces when traditionally these have been

happening in like non-online spaces. And so, it's really, it's just like such a sometimes when I explain to like people who are part of academia, I'm just like, “We're in this like really special corner of Berkeley!” And like I totally sell it as like extra, extra special because of how like uniquely interdisciplinary we are compared to everywhere else I feel like. But I'm biased. I'm in the school, so of course I'm going to hype it up.

Saintsing: Is that… do you have a similar relationship with your, the program you just graduated from, Katie?

Keith: I think it was slightly different than lucy's in that it was sort of self-contained. I think the college that I was in came from more of like a very traditional computer science department or college, and there was a big focus on artificial intelligence and natural language processing, which I focus on and lucy focuses on. But they're trying to build up the data science and expand into HCI, more pure data science like informatics. And I think there's a huge demand from the students for that, too, because that's where a lot of jobs in the workforce are. So, I think Berkeley is a little bit unique in that they've already sectioned themselves off and are growing this program kind of separately from their computer science departments.

Saintsing: Are you seeing people wanting to move more towards what Berkeley already has? Or is Berkeley just unique, and it’s going to stay unique?

Keith: My understanding is that they're kind of growing across the country, especially data science, informatics, information schools. Is that your sense, Lucy?

Li: Yeah, I think so. One thing is like a lot of people now are a big fan of like rebranding, like HCI, which is human computer interaction, by giving it all these fancy new names like human-centered AI or human something. Something attached, some like algorithmic thing, computer thing next to it. And so, I think there's a growing interest in humans in computer science, and so a lot of schools are trying to find ways to integrate that into their curriculum, and some different schools have different strategies of doing that, like some schools maybe they'll like create a whole new institute that's siloed away. Maybe they'll try to like actually integrate into every single department. I don't really know what people's strategies are for that, but like when I was looking for PhD programs, I guess maybe I didn't even know what I was signing myself up for because I came from a traditional computer science department. And so when I arrived in the School of Information, someone asked me, “Are you qualitative or quantitative?” And I was like, “I'm quantitative.” And then I realized that quantitative includes surveys, and like other types of like things that involve numbers. And I was like, “I don't run surveys that often. I mostly just like run like machine learning classifiers.” And so it was just like situating myself in this new area where, before I came from a place where everybody just like slapped an algorithm on something and didn't think twice about it, but I feel like a school of information would think more carefully about that and also has more of a social justice leaning because of how they care about people more.

Saintsing: Okay, now I'm interested to know actually what you're both doing. So, you've mentioned natural language processing. So, I'll start with Katie for this one. Could you give me a definition of natural language processing and then maybe also go into a little bit about what it means for the work that you do?

Keith: I think about natural language processing as this broad umbrella of, “How do we apply computational and statistical and linguistic techniques to text data?” And what distinguishes NLP to me is kind of like the automated nature of it and scaling it up. But I think it comes out of a tradition of computational linguistics, so it was you know originally had a deep linguistic focus. But there were several paradigm shifts. One being that a lot of machine learning and statistical techniques came into the field and ended up doing really, really well. So, if you go to our NLP conferences today, it's overwhelmingly about machine learning and most recently deep learning neural network stuff. So, I would say the modern sense of NLP is really about applying machine learning techniques to language to text data.

Li: Yeah, that was a pretty good definition. I feel like I have this like unusual perspective of NLP, too, because I think Katie and I both think of ourselves as the intersection between computational social science and NLP. So, like I personally am not that interested in things like building the next best dependency parser (So, dependency parser is something that will like link together different parts of the text and say like what the syntax tree might kind of look like. This is a very like simplification of what it actually is.) So, for me, I think, how I think of NLP is that there are people who like make a lot of models, and these models can do certain tasks, but then there's also people who are like using those models and trying to answer like social scientific questions, which is kind of like the corner that Katie and I are more in. And one word for that is like “text as data.” So, it's like some sort of like content analysis, and maybe Katie could talk a little bit about like text as data and like that sort of community.

Keith: Yeah, so text as data is both, I would say, like a community of practice, like people all trying to have similar goals of “let's use NLP and computational techniques to analyze text data and come out of it with some sort of quantitative insight into some sort of social science phenomena.” But it also is an annual conference, and so Lucy and I have both been there together, and it's an amazing conference of a mix of social scientists, some historians, some humanists, and then a lot of computer science, some stats folks. So, it really sits at this intersection, and Lucy and I are actually academic cousins, meaning our advisors had the same advisor, so we come out of this tradition of both of our advisors, and that community is really central I think to both of our research agendas, and the research agendas of the labs we're being trained in, or we're trained in.

Saintsing: So, basically in NLP, you're building a program that can comb through a bunch of text, like maybe tweets or maybe even like essays, articles, some huge compilation of writing, and you are trying to put in a program that will regularly pull out certain things that you're interested in?

Li: In NLP, there's this paradigm of thinking about tasks. So, like I have a computer, I want it to do something with a bunch of text, and that thing that they could be doing could be things like dependency parsing, which is kind of like “What is the subject of the sentence? What is the verb associated with the sentence, of the sentence?” It could also be something like any question answering task where it's like you give the computer a question, and then it reads through a bunch of things, and it tries to retrieve the answer. And so, it's all formalized in terms of tasks I guess. For like people who are interested in the intersection of computational social science and NLP, the tasks might instead be something like, “Given this user's tweets in the past few years, are they more likely to leave this community or stay in it? Or given this book, what is the narrative arc that's happening?” Or something like that, and then like these (some of the questions I'm putting out are less well-formed than others in terms of like what are the inputs, what are the outputs), but the general idea is that you have a computer, you get a bunch of tasks, and you ask it a question. You're like, “Can you please do, answer this question for me?” But the question might not be like a question-answer-type question, but it might be like a research question of like, “I'm curious of whether this word has gotten more popular on twitter in the past few years.” And then the computer will like do something, and then it'll hopefully tell you whether something's popular or not.

Keith: I would even add you know an even simpler task that sometimes is something that you can conceptualize is your input is a sentence, your output is “I want a part of speech tag on every single word.” So, I want like a noun or a verb, and something a lot of these words are a one-to-one mapping. But something like fish depends on the context that it's in. So, if you want to automatically have as input a sentence, an output, “is fish in this sentence a verb or a noun?” People have built algorithms to do that and try to infer that from a lot of data. And so, I think dependency parsing is even sort of a higher level computational task because the input is still a sentence, the output is now this structure that linguists have come up with that relates all of these words based on their syntactic, so like the structure of the sentence dependencies to one another. So, I think Lucy is totally right in characterizing all of these sort of subtests are what a lot of mainstream NLP works on, and what both Lucy and I do is “how do we take various tasks in mainstream NLP, apply it to broader social science questions, and then reformulate potentially brand new tasks that could really help social scientists with what they're trying to do with text data?”

Saintsing: Can you run the same NLP algorithm for English text and other languages?

Li: Yeah, this is a good question. I feel like this is like an open question in NLP is like “how do we build models for a larger variety of languages?” And we also have this issue right now where there's this thing called “low resource languages,” which is like languages that don't have a lot of training data out there, and the models for those languages are not quite as like, they don't perform quite as well.

Keith: We have a huge bias towards English because that's where a lot of the money is unfortunately. A lot of the money is in processing English text for English consumers were driven really heavily by industry especially nowadays, which I think, in my opinion, is somewhat unfortunate, but is the reality.

Saintsing: A lot of the money in the research that you do is coming from like big tech companies?

Keith: Yeah, maybe not in like my research specifically. I think my advisor got a lot of like NSF funding, but I was on a fellowship from Bloomberg, which was a private company, and they're a great company, but still it's a private industry. And I'd say our conferences are funded by big tech a lot. Lucy, I don't know what you're funded on.

Li: Yeah, I'm funded on NSF as well, but it's not unusual for a PhD student in NLP to be working part-time maybe at a big tech company or interning at big tech companies. I interned at Microsoft this summer, so I have to admit I was also connected to big tech and fellowships coming from big tech. It's an interesting question of the relationship between industry and academia especially in computer science given the fact that like these companies hold so much wealth and so much power.

Saintsing: But I guess it's nice, at least Lucy's been talking a lot about the social justice bent of the Berkeley program, so I guess do you at least feel that that's nice that you kind of can, if you're going to these internships and other students like you're going to these internships, do you feel like that's having kind of a bringing a social justice perspective that might not otherwise be going into these companies?

Li: Yeah, I really hope so. I think a big challenge… Well, there's a lot of challenges with this sort of thing. One is that you don't really know if leadership is going to agree with the people who are doing the research as we've seen with some of the issues coming out of Facebook. We know that at Google, Timnit Gebru and Margaret Mitchell got fired. And so it's like, it just seems a little murky right now in terms of big tech. And another issue might be ethics washing. Like if there's a lot of like really good seeming people going to these tech companies, they might seem like they're doing better than they actually are in terms of on the social justice front. Yeah, balancing these things as hard as a PhD student because you're just like, "I want to have a career. I want to like have money more than like the stipend that I'm getting right now." But then like you see the options of where you can do like research, and sometimes you're just a little, you're just a little worried about how your research might be seen, especially since a lot of it is critical of the technology that are coming out of these companies.

Keith: There's a lot that's been written on this, and one of my favorite papers is titled "The Gray Hoodie Project," which references an analogy to the White Lab Coat Project of big tobacco. The focus is you know, there's a lot of parallels between big tobacco and the money they had and the way that they were buying academics and what it looks like from the outside with big tech right now. So, I think we're at a really pivotal time, and

like lucy said, there was a lot of stir in our field just this summer I think when Google fired two of their most prominent researchers who worked on fairness in artificial intelligence. I think it made a lot of people question whether these big tech companies are actually following through with what they say they're doing in terms of fairness and ethics, so I think we're at a very, very pivotal moment in our field when it comes to the roles of these big tech companies who have enormous power and enormous amounts of money.

Saintsing: We're talking about jobs now, and Katie, you just graduated, and so I was wondering if you could kind of tell us generally what type of job you've taken on and kind of how these… Did these considerations we've just been talking about factor into that decision at all?

Keith: Oh, yeah. I don't want to be disparaging of any company, but there were certain companies in particular that I just didn't even apply to because I am very concerned with the ethics of those particular companies. So, I knew I wanted to stay in academia, and I think it's a really important time that academics think about how we're broadening the field and trying to be more inclusive and bring people in. And for me that meant trying to, being at an undergrad institution where a lot of the so-called leaky pipeline, especially for women and people of color (students might be interested in computer science and then leave because there are a lot of toxic people. There's a lot of great people too, but there's a lot of toxic people that turn people away from computer science and tech). So, I ended up going on the undergraduate job market, and I'll be at Williams College next fall as a tenure track professor, which I'm really excited about. And then they were very generous and let me defer for a year, and this year I'm at the Allen Institute for Artificial Intelligence. So, Paul Allen donated a pretty large sum of money to create this institute that's about AI for the social good. They're a non-profit, and they have a lot of really great researchers you know trying to work on impactful work. It's a wonderful place. It kind of feels like a tech company, but it doesn't have the the profit driven motivation of a lot of these other places. So, I got very lucky I think to end up at institutions that I think align with my values.

Saintsing: Well, congrats! That's great.

Keith: Thanks, thanks.

Saintsing: So, I'm really interested to know kind of what sort of research questions you both address. So, Lucy, why don't we start with you? What are you actually using NLP to ask?

Li: I'm interested in a lot of different things, and it's getting to the point where that means I have a lot of projects, and I'm very busy, but like some of the main themes, so one of the main things that I'm really interested in is the language of online communities, and this could just be an excuse for me to go on the internet and procrastinate, but I have published a paper in a journal about the language of online communities and just looking at what kind of communities tend to have very community specific language.

And it turns out it's the ones that are not too big, the ones that have a lot of user activity and stuff like that. It's a fun paper, and now I'm like taking that work, and I'm going further with it where I'm just like thinking about like language changing communities over time. And also extremist communities and the language of those communities and stuff like that. And so, that's like ongoing work. Another line of ideas that I'm pursuing is "how are people described in text, and how are people discussed in text?" And so I have a paper on Texas textbooks, Texas US history textbooks. I'm looking at the people described in there, and we're planning on extending that work to looking at maybe how people are talked about in fiction books and looking at maybe discussions of race in fiction books and stuff like that. Those are kind of like the type of work that I'm into these days.

Saintsing: You mentioned extremist communities and the way people are discussed, and I guess it's nice you kind of work with algorithms so maybe more of your work is about kind of working on code, but is that… do you have to deal with upsetting… like how do you deal with what you're actually looking at in terms of text if you have to look at text that's maybe unpleasant?

Li: Yeah, that's a really good question. I think when I first started I was a little naive in the sense where I felt a little bit more removed from the work because in my head I was like, "These are online extremists. They're extremists for a reason. They're on the extreme end, and therefore, they have, they're further away from me." But then I started like reading into them a bit more and then noticing some of that rhetoric come up in more mainstream areas or just like my daily life and stuff like that, and it kind of started hitting a little bit too close to home. And I was like, "Oh, the things that are extreme, are considered extreme, are actually not that extreme. The things that are close to me in the world are also not happy." And like, so like things like racism exist everywhere, and you know misogyny exists everywhere. All of these things that you think of as extreme on the internet. So, like I've been looking at like incel communities and stuff like that, so those types of communities that you think are extreme, little bits and pieces of them show up in surprising places. And so, it can get a little difficult to cope, which I think the way I deal with it is that I don't work on just one project. So, like I also have other projects. I think another way that I try to cope is I don't know. I think like I keep telling my advisor that maybe in the future we should work on something happy. Maybe we should do something like well-being on the internet or like happy gardening people on the internet or something you know.

Saintsing: Yeah, I hope that you get to do that.

Li: Yeah.

Saintsing: OK, thanks, Lucy. Okay, and Katie, you just graduated. So, could you tell us a little bit about what your dissertation was on?

Keith: Yeah, so the first half of my PhD focused on specific methods for social measurement with NLP and text data, I had one paper that we were trying to measure and create like an automated method for identifying civilians who have been killed by police from corpus, corpora of news reports. And so that involves a lot of event extraction, it involves a lot of thinking about "how do we aggregate outputs from our models that are potentially statistically uncertain?" And we need to somehow aggregate these in a statistically rigorous manner. So, I thought a lot about measurement, and then the second half of my PhD was focused on causal inference with text, so with these large digital trace data sets and observational data, oftentimes people want to ask causal questions. But a big part of that is having confounding bias, so there's a lot of confounders when you're not actually running a randomized control trial. And so, a lot of people have tried to use, say, all of Twitter to get at confounders between individuals they're trying to study or particular confounders coming from text. So, some of my work looked at "how do we outline methods by which we can do this? And what are some of the open problems in thinking about text as causal confounders?" And then one of the last projects I worked on, I recently had a really great collaboration with a legal study scholar who really wanted to look at supreme court oral arguments and thinking about "what's the causal relationship between the gender of lawyers on the supreme court and whether they're interrupted by justices?" And so their language here is really a causal mediator that we want to say "well, what's the direct effect between gender and interruptions? What's the indirect effects that goes through the particular ways and the kind of arguments that people are making in these oral arguments?"

Saintsing: And this is actually a great segue into your podcast, right? Because your podcast is kind of tackling these more interdisciplinary papers. So, could you tell us you know what is your podcast? And also I'm really interested to know how you got started on it? Like what was the process that led to you all, you two doing a podcast together?

Keith: Our podcast is called *Diaries of Social Data Research*, and we sort of, we started early 2021. I had the idea for a while. New York Times does this *Diary of a Song* that they interview artists, and they go back to old versions of the song, and then they walk through what's the historical progression of the idea of the song to what we actually hear as a final finished project. And sometimes when I'm really burnt out of code or writing papers, I like to you know dive into pop culture a little bit, so I found this really entertaining and interesting. And I wanted the same thing for this interdisciplinary computational social science research because all you see is the final, finished product, and that was really frustrating for me. I think it took me many, many years in my PhD to get to the point where I understood the process of doing research, and I felt like that was a big barrier and one that I had to really struggle through myself. So, I had this idea that I wanted to do. Behind the scenes, take us through the idea, conception all the way to the final finished product and all the challenges and roadblocks in between. And I knew I work best when I have really good collaborators, and so I had met Lucy because she had interviewed with my PhD advisor. She had gotten into UMass and was really considering us, and we do a candidate Friday. We host PhD candidates. So, Lucy actually stayed with me, and my impression of Lucy was, "Wow, this is someone who is super

brilliant, super hard working, so enthusiastic." And so, as I was sort of thinking about who I wanted to collaborate with, she like came to the top of my list. So, I just reached out and pitched her the idea, and in classic Lucy fashion, she was super enthusiastic and had all these ideas of what we should do and how. And I was like, "This is gonna be great." And I kind of need that like… I like having an idea, but then you need someone who can be there when it's kind of hard to push through. And so, now we've done, yeah, I don't know, how many episodes, but a lot, and we're going to continue to do so, and it's one of the things I really love.

Li: This podcast was entirely Katie's idea. I just got really lucky that I made a good impression, and she decided to pick me as her collaborator. I have no podcast experience before this, but I also had this problem of like being an early career PhD student, I kept seeing the shiny outcomes of everyone's work, but I felt like I was just struggling all the time behind the scenes. And so, I wanted to kind of like normalize the fact that like other people who have these really shiny outcomes might also have stories behind the scene. Because I was just, I was just entirely behind the scenes for quite a while before I came out with anything. So, that was like a really interesting aspect of the podcast for me, and I still think it's like maybe one of the most important parts that we look at processes behind papers, not just the papers themselves.

Saintsing: Well, unfortunately, it looks like we are running out of time on the interview. Is there anything you'd like to leave us with before we go?

Li: Listen to our podcast. Listen to the advice that these very, very accomplished and smart people are saying about how to do interdisciplinary collaborations really well. And then take some of that advice and carry it out in your own work.

Keith: Yeah, I would just say that if anybody's listening and thinking about a liberal arts college job in computer science and wants to understand how to navigate that, please do reach out because I think it's a little bit of an unconventional path for people in tech and CS, and I'm happy to talk to whoever is interested.

Saintsing: Great, thanks. Today I've been speaking with Lucy Li and Katie Keith, hosts of the *Diaries of Social Data Research* podcasts, and you can Google them to find more information about them. Google their podcast, look it up, and if you want to get in touch with Katie Keith, she just graduated from the University of Massachusetts Amherst. Lucy Li is here at Berkeley. Thanks again so much for being on the show, Katie and Lucy.

Li and Keith: Thank you. Yeah, thank you for having us.

Saintsing: Tune in in two weeks for the next episode of The Graduates.