Camille Nebeker: So, I'm talking about this term called ethical sourcing. Just so you have some background, it's coming from the NIH initiative called Bridge2AI, which began about four years ago, and UC San Diego is one of the partners. One of the things that occurred early on when I first started working and I'm a principal investigator for the Ethics Corp. The bridge2AI is $130 million. It could be more initiative to create four datasets that I'll talk about in just a second. The idea behind Bridge2AI is that they wanted to generate new data and best practices to propel modern machine learning toward improved health care. They funded and initiated the development of four data repositories. Each of these are very unique data repositories. Part of the goal is to advance a culture of ethical reflection around the lifecycle of a data repository creation and to create a trained workforce skilled in new and responsible methods of data creation, management, and use. The pillars for Bridge2AI are data, ethics, and people. When we first started working together, it was really an interesting way of bringing people together. It wasn't just a group of people who had proposed to do something. The NIH liked certain parts of what different groups proposed, and so they mixed and matched them. From the very early beginning, we were partnering with people that we didn't really know, with disciplines that we didn't really understand. It took some time to start learning about what it was we were trying to do, how we were going to create these repositories. These four grand challenges, what we call data generation projects also one of them is a data repository that's based on detecting disease and health through voice recordings. How many syllables would it take for me to identify that you are predisposed to Parkinson's? Really sensitive data that needs to be stored in very secure ways. Who can have access to those data? How are those data collected so that they can be used in a way that can inform health? Another one was called AI-ready, and these are still active, so I won't say was, it is. Its goal is to recruit 4,000 people to learn more about how AI can be used in clinical care, focusing on Type 2 diabetes. The course grand challenge is looking at over 100,000 patients from 14 intensive care units. Those are existing datasets, not prospective. The first two that I mentioned are prospective recruitment of people to contribute data, and then CM4AI, which is led by Trey Ideker here at UC San Diego, is looking to map cell structure function and disease context using cell lines, looking at 200 genes and proteins for coordinated experiments in these three modalities. Very different kinds of data and across these four data generation programs. We're looking at electronic health records, digital and wearable data, waveform data, survey data, clinical data, and imaging, as well as genomic data. One of the consortium goals is how do we empower researchers, clinicians, and communities to collaboratively generate high-quality, ethically sourced datasets that reflect the diversity of human experience. About a year to 15 months into the program, I was working with a tribe from South Dakota trying to help the National Institutes of Health understand what data they may be willing to share, where that data may be able to go, who might be able to touch it. Working with the program officers, there was a real disconnect about what was possible. I asked the question because of the branding of Bridge2AI, focused on creating four flagship databases that were trustworthy, AI-ready, and ethically sourced. I asked the question, what does it mean to be ethically sourced? There was a pause, and the response was, " We're hoping that you would define that. This is a year-plus into the program where the data are already being generated. They're already being put into a repository. The whole goal is to make this open-access public data that can be used by all kinds of people who want to learn more about AI and machine learning, so it's a training database. It really caused me pause. Obviously, I'm part of the ethics cores, and I thought, well, gosh, should we have known this before now, and I think we probably should have. But it really set me on a path of trying to figure out what does that mean. The reason it matters is because ethically sourced is the foundation of what makes something trustworthy AI-ready, and a flagship dataset. It's foundational. It's not supposed to be, let's get that ethics in here at the very end and make it look like we did something here, which oftentimes happens, so I'm not being facetious, but Bridge2AI very intentionally put ethics at the forefront. Creating an ethically sourced database has implications of public trust about who wants to use these data and downstream performance. Thinking about this, I ended up using my friend GPT to ask about the origin of ethically sourced, where did that come from? It turned out, and I'm not a business person, but it came from supply chain management. How many of you remember back in the day

when Nike did some bad things, polluted rivers, hired children, really poor labor practices? I was thinking, Wow, this is super interesting. I started reading the literature on supply chain management. The whole purpose of that was to get these companies to not just create a product, but to tell people that they were partnering with good partners, that they were committed to not polluting, that they were going to have good labor practices. When you bought the product, you knew that it was created by a company that had committed to social values and norms. Then they started creating branding around these terms. I learned a lot, and I started thinking about how do we map what we're doing with the data repository development and map that into the supply chain management framework. What we did is we took what's called value sensitive design, which helps us understand what the various stakeholder values are and where those ethical tensions lie, because there's always going to be, well, if we do this, then we're going to give up that. Having those reflections so that you could think about what those tensions are and making those visible so that decision-making would be informed, knowing that these tensions exist. Then the supply chain management operationalizes the processes and accountability. Then by combining value-sensitive design with supply chain management features, we could start to define what ethically sourced looks like from the very initial stages of planning through implementation and management. It was so fascinating to me early on, I'd ask in my various meetings, What does that mean to you? Many times I'd hear, hey, ours is exempt. IRB is not involved. We did our regulatory stuff. We're good. Oh, we got consent. We're good, and so everyone that I spoke to thought, This is all we got to do. I kept thinking to myself, no, and I think part of this is the term ethics. I think we have a gap or a barrier in how people interpret that word. Maybe it's responsible, maybe it's accountable, maybe it's transparent. Maybe it's a lot of these different kinds of values. One of the things that I needed to set up is success and not penalization. Because every single one of these data repository generation projects had already started. They had already started collecting data. They had begun their process of even releasing it. It wasn't that they wouldn't have been able to do it. It's just that we're working in parallel. Here comes the data repository, and now we're defining ethical sourcing. My goal was to say that it doesn't have to be all upfront that you can actually strengthen what we think of as ethical sourcing retroactively, and that when we identify the gaps, we can minimize those gaps through documentation, governance updates, community engagement, and transparency. The values, as I mentioned, transparency is talking about where did the data come from. What was the origin of those data? Were there permissions? Were permissions even needed? But knowing where these data originated, knowing and documenting the metadata that is needed to make sure we understand the source of data. Accountability, who's involved in governance? Where is it happening? How is it audited and whose role, who's doing what? I think in any of these big data generation projects, and the reason we're doing this is because we really need to know what the data are, how they were collected, even the vendor name, the model number, so we can make sense of what the data are. You think about using electronic health records as a source of training data. It's very messy. You may correct me on that, Dr Singh, but I think it's pretty messy. We have integrity issues. We think about data quality and validity. Thinking about sustainability. What is the long-term stewardship? What happens when the NIH chooses not to continue to support these data repositories? What happens at that point? There's a lot of questions there. Then respect for human rights. We have to think about consent, dignity, and community engagement. It's a very important part of this. With this supply chain management lifecycle, we're thinking about these steps in supply chain management. The first thing you do is demand planning. What does that translate to when you're thinking about building a biodata repository? What is the demand? How do we move and define the scientific need, and what the ethical risks are? What is the design consideration, governance, and architecture? Sourcing, inclusive, representative data acquisition, processing, quality control, distribution, and control access. When you build something that is going to be used for machine learning AI. When you're building something like this, what do you need to think about? She said, "Well, it's almost like you have to think about building a building. You have to think about HVAC. You have to think about soil density. You have to think about all of these things that are going to hold up the structure." When the NIH provides funding to say, do this fast, a lot of that structure doesn't

happen in ways that I think we would all want to see it happen. Clearly, we at a crossroads, but it's coming together. What I did here in this table in this paper, I have it at the end of the slide. It's open access. Any of you can download it. What we did is we looked at the supply chain management step. We talked about how it looks in supply chain management, like forecast demand, understanding the objectives. Then we said, "What would it look like in a health data repository counterpart? Then what are the ethical sourcing values? Who are the actors?" This is really important. You think here, we have clinicians, we have researchers, we have data scientists, we have sustainability experts, we have business analysts, we have the public, we have architects for data, IT experts. We have people that are sourcing managers, we have procurement, we have legal people that are looking at the agreements between the company that wants to do retinal imaging, and getting that retinal image to the data repository, and do we trade the use of equipment in exchange for giving that company back free data? There's a lot of different places, things can be a little questionable that we have to start unpacking and looking at. I have no idea about technical privacy, how to create blockchain, but together, we can learn to talk to each other and start solving some of these really difficult, challenging parts of this infrastructure. When I think about the actors and Bridge2AI, and this is true for anybody that's building a health data repository. This just happens to be a use case. We need to know that the sponsors are aware of what risks that they're promoting when they say, do this very fast in four years, and we can't be confident we can fund you after that. Then the data contributors and the communities that they represent. What time are we spending not only in thinking that we can do consent, but what's the implication for the communities that these people represent? What forms of governance and stewardship are there? We don't think about there's an IRB somewhere that somebody saw. Governance has to happen and stewardship has to happen at every single touchpoint across the life cycle. We think about the clinical and research teams that are going to be involved, the data creators, the designers, the users, and then again, regulatory and ethical perspectives happen throughout. This table talks about who the actors are, what the roles are, what the challenges are, how you would mitigate issues that are challenging, what those tensions are that we were talking about. What are the values that you rely on to draw on when a conflict arise? Then what are the open questions that we still have? Now, again, this is very conceptual. It was something that It was a starting point to get us to what does ethically sourced mean? But I didn't want to stop there. I wanted to move from getting this paper out that starts conversations. But what would it look like if we could take this concept and get it into a tool that leadership could use to start thinking about, what do I need to think about? How do I evaluate myself? How do I identify? Can there be things that I can improve? Using it as a tool to do better work. As I said, Bridge2AI set a very ambitious goal to have ethically sourced AI-ready data. Ethically sourced meant different things to different teams. Our paper, we proposed a framework, and the checklist that I'm working on now puts that into a very practical and useful tool. The checklist covers this entire life cycle, which looks like I have 11 steps. What's the purpose of doing this? What's the provenance and traceability of the data? Did consent happen, or should it have happened? What about community involvement, equity and bias mitigation, data quality and integrity, documentation, data sheets, health sheets, all of the pieces that go into that, governance and accountability, access and use controls, privacy and security, stewardship, sustainability, and documentation and public communication? I really think this part is where you not only evaluate yourself, but you make public, where you stand. This is where I think we get closer to having a brand that can be meaningful to the public, and the public is, who's paying for this? Again, I took my paper because it was published, and I put it in one of our subscription large language models, and I said, "What would a checklist look like?" It came out with a really bad Excel thing. I'm not an Excel user, so I thought, this makes me uncomfortable. I said, can you make it more aesthetically pleasing? Without compromising the content that's in this paper. It gave me some pretty nice stuff, and I like it so far. It's accurate. It hasn't done anything misleading, it hasn't added things that I didn't have. I'm checking it twice. But now my next step is to start working with different groups like the standards working group, the groups that are creating the other parts of the infrastructure, the tools group, the data group, our ethics working group, to go through each one of these things to

try to find out is how we're seeing this meaningful to you, or does it not make sense, or would you say it differently? We have columns where somebody can say, yes, I did this. Nope, I didn't do it, or it wasn't applicable. Then you document your evidence, and I'll show you what that might look like. These are just snapshots of what the checklist is looking like so far. It has all of those little boxes that I showed you earlier. But this is what it would look like if it were to be filled out. On the left, you see consent language covers actual data uses, including the use of machine learning and artificial intelligence. Then under the documentation piece, you would say, well, consent forms currently explicitly reference data analysis and algorithmic development, but early versions did not explicitly name downstream AI/ML model training or data reuse across institutions. What the action is, is that we could revise consent language for future cohorts to explicitly reference the machine learning, secondary analysis, cross-site data sharing, and add transparency statements for legacy participants. The idea is that you would have actionable items. You would know between your team, you could identify what should we prioritize and start working on. This one's about data sharing and stewardship practices, so data use agreements reflect ethical constraints. Does the data use agreement prohibit re-identification, redistribution, non-approved uses, reinforcement mechanisms? What the action is is to periodically review DUAs as a new use case emerges. Again, It can be a lot, but it's a starting point. Right now, there's nothing. I think having a tool that people can actually use is going to be a good starting point. How they could use this is they could do a self-assessment. They can identify strengths and gaps, they can develop action plans, and they can publish transparent public statements about their strengths, their gaps, and improvements. Operational opportunities, I'm clearly working on the checklist, but this could also be something that we would use for onboarding to start socializing people about what they need to think about. There are so many different actors that come from different disciplines that have been socialized to ethical and responsible practices in very different ways. You have clinicians that are licensed. You have researchers that may have been exposed to scientific ethics. You have sometimes business people that have not had anything to do with that and may not really understand what the culture and the norms are. Bringing these people together during onboarding and learning what the expectations are, and having a checklist, and Karen DeBiasi was in this Google-Hastings Center bioethics AI thing about what, five years ago. I was in a breakout room with a Google engineer who said, "I want to do things that are responsible and ethical. I don't know what that means. I just need a checklist, and engineers really like checklists." That's where I'm seeing this going, is that we give people what they may not know that they don't know, so that they can start getting familiar, and it may not be their responsibility in one section, but it could be in another. This is a fault of mine. I like to get things into the workforce, practically applied, get things out there, and so that's what we're doing now. Again, this is a goal to identify gaps, figure out what you would operationalize to do this in your work, and what supports would help you to get there. This is the paper. You can type the PMC number right into your search bar, and it will come up. Thank you. [Applause] [Music]