Episode 35 – Andrew Su: How artificial and community intelligence are shaping medicine

Drew (00:04):

A sincere welcome back listeners. Thank you for being here. This is another serving of Science Changing Life and I'm Drew Duglan. Today we expand our knowledge networks, exploring how artificial intelligence continues to shape medicine and help us join the dots between increasingly complex fields. Helping me to join the dots today is computational wizard Andrew Su, a leader in developing digital tools to accelerate biomedical discoveries. Before he tells us how we can all be citizen scientists, let's find out how he bridged the areas of science and technology.

Andrew (00:38):

I went to Northwestern for undergrad, where they had this program called the integrated science program. And it was a fantastic fit for me because it was, it was like we, we did sort of half of a major in all of the natural sciences and computer science at the time was specifically in a sort of computational chemistry. So thinking about molecular docking and how we physically simulate how molecules interact. And then when I started grad school, I was right when genomics was getting more and more popular with micro arrays to do gene expression profiling. I was a first sort of real molecular profiling of what cells we're doing at a really broad scale. And so that's when I sort of discovered bioinformatics and all the data challenges that we're now really familiar with in terms of biology, data and biology research.

Drew (01:34):

So even in undergrad, did you have this foresight that science, that natural sciences let's say in biology was gonna rely so heavily on computers and sort of big data?

Andrew (01:47):

I would love to say yes that I was ahead of the curve, but the truth is no. I mean, you know, I, I would say chemistry and computer science, even at the time to me, I, I was like, oh, what am I gonna do with this? But even that is sort of a different than what I'm doing now, right? That is I would put, you know, sort of squarely in sort of computational chemistry, computational biology, where we're doing a lot of physical modeling of the world. And, you know, I think the data driven revolution that's come around in the last 20 years or so, right, is less about the physical modeling of the world and more about the statistical analysis and the data management challenges. Right? So here we're looking for patterns in data. So what genes are differentially expressed in a disease system? What genes tend to be differentially expressed together across many disease systems. So it's a different set of challenges that yes, I'll use computational methods, but bias more towards sort of statistical analysis and pattern recognition.

Drew (02:55):

Right. So you just mentioned bioinformatics then. So can you kind of break that down for people of what that is and what is it that your lab does now broadly?

Andrew (03:03):

So I think of bioinformatics as this intersection between biology and chemistry, computer science and statistics, and there's some really interesting scientific questions right at that interface. And again, as we've alluded to a lot of those questions are driven by the fact that it's very easy to generate data in biology beginning the last 20, maybe 30 years ago, first with gene expression technologies, then really massive genome sequencing technologies. This has sort of been a revolution within biology, where we go from a very sequential type of science where we come up with a hypothesis, we design an experiment, we do the experiment, we interpret the results. And then we design the next hypothesis, right? That is 200 years of, of biology sort of basic framework. And again, in the last 20 years where we have the rise of biotechnology, lots of robotics, lots of miniaturization. We can do thousands or tens of thousands of experiments in parallel, and that's opened up some great challenges in bioinformatics.

Andrew (04:16):

So my lab right, really focuses on bioinformatics. And so we think about two main areas. The first is to create tools in bioinformatics. And we think about this as bioinformatics infrastructure. So that means we are creating tools to organize information, to analyze the data and to visualize it. And so these are tools where we actually want to provide access to the broader scientific community. And in some cases, the broader public access to scientific information so that they can use that in their research or in their daily lives. So one half of the lab really focuses on bioinformatics infrastructure. And then the other half of the lab focuses on trying to use tools to make discoveries, right? So this is really data mining. So it's taking the tools that we develop, combining it with all the tools that exist in the field that other people are developing and really applying those tools to large data sets. So for example, trying to understand what pathways are dysregulated in cancer versus not, or what variants are most important for coronavirus, for example. So doing large scale data mining within that to come up with testable hypotheses that may lead to, to new discoveries.

Drew (05:41):

And this issue of just the scale of data, I mean, people don't realize it, but if you take any given topic in, you know, biology, chemistry, whatever it is, and then the data that comes from different labs all around the world, the discoveries have just been exponential, right? There's just papers and papers and papers every single day. It's like, how do you keep on top of all that information?

Andrew (06:01):

Yeah. That's, that's spot on. The NIH - the National Institutes of Health invests something like 30 billion a year on biomedical research on top of whatever's investments are done in industry and in other countries. And so the biomedical research enterprise is hugely productive, right? We're uncovering new insights every single day. One metric of those outputs is just the publications that we put out. And so there, you know, in PubMed the primary resource for the biomedical literature, right, there are over a million new articles every single year averaged out over the year. That's like one new article, every 25 seconds or so <laugh> so it's incredible, right? And there's no way that any single person is reading any substantial part of that literature just too massive. And so a lot of what my lab thinks about is, well, how do we take all that knowledge, all the knowledge that's being generated through all those publications, how do we get that into a form where we can computationally provide access to it, computationally, analyze it and visualize it because again, no person is reading by hand more than a high number, probably hundreds, you know, maybe maybe low thousands numbers of articles per year.

Drew (07:24):

Right, this almost seems like a database approach to sort of systems biology because medicine, for example, is just so highly specialized now. And like you said, so few doctors are able to, or even have the time to sort of bring these different threads together and different systems in the body. So it's kind of nice to have these tools that could potentially, uh, solve some of these problems.

Andrew (07:45):

Yeah. In an ideal world, every time we make a new hypothesis, it is based on the entirety of what is known up to that date, right. We want to be make the most informed hypothesis that we can at any given moment as the field gets more and more massive. The percentage of it that we are personally expert in just shrinks, shrinks, and shrinks. So we, we get more and more specialized. And one byproduct of being more and more specialized is that we miss the opportunities to connect the dots between seemingly disparate fields. And those are the places where we have huge scientific advances. So my hope is that computational tools, in addition to having, you know, interdisciplinary science, my hope is that computational sciences can also help build those bridges between, uh, areas where there's good synergy to be discovered.

Drew (08:45):

And so when you mentioned the two halves of your lab there, I think, uh, one of the elements was sort of this public interface. And I think some of the projects you've been involved with have sort of versed on this idea of community intelligence and citizen science. So could you kinda speak about that and some examples of it out there and, and then maybe some specific ones that you've been working on,

Andrew (09:05):

I'm gonna start with seemingly tangential story, but I think it's the best sort of origin of online citizen science. So I'm forgetting some of the details, but a couple decades ago there was an astronomy graduate student who was faced with classifying images of galaxies. This was his PhD project, right. He had a million images from the Hubble space telescope and other space telescopes, and he needed to classify these galaxies based on essentially their shape, uh, whether they were elliptical or circular, how many arms they had and sort of details and so on and so forth. And it was going to be a pretty daunting and mind numbing challenge for him to, to slog through this. So what he did instead was I think in retrospect, I mean completely brilliant. He put it online and he said, Hey, here are some cool looking images of the universe.

Andrew (10:00):

Come help me do this. And he just asked people, it's like, if you want to be involved in scientific research, here's a great opportunity to do that. And so he called this project, the galaxy zoo, and essentially within weeks, he had millions of classifications of galaxies, right? Tens of thousands of people sort of volunteered their time just cause they thought it was cool, because people, if you give them sort of an opportunity to be engaged in scientific research, a lot of people like that. And so this was sort of the spawn of the field of online citizen science, where we have platforms online to allow people to engage productively with professional scientific researchers.

Drew (10:45):

Wow. That's amazing. I love that he crowdsourced his PhD.

Andrew (10:48):

<laugh> exactly, exactly. Um, right. That, that whole effort, right. It started out as a galaxy zoo and it's expanded into this project called Zooniverse. So if anybody's interested in citizen science, you go check out Zooniverse I think it's zooniverse.org. And so there's a variety of different citizen science projects. Uh, many of which are focused on astronomy, but they also have other projects focused on biology and

ecology and history. And so it's a really great way for people to find citizen science projects that align with their interests.

Drew (11:22):

It turns out that there are citizen science initiatives open to the public in all kinds of disciplines. Some of these initiatives have you identifying clogged blood vessels in the brain to advance Alzheimer's research while others let you design new potential drugs to help treat misfolded proteins similar in concept, but different in approach, Andrew's lab developed a web-based app called Mark2Cure, aimed at advancing our knowledge of NGly1 deficiency, a rare genetic disease, which can lead to developmental delays, movement disorders and liver disease.

Andrew (11:53):

So I'll say that the majority of citizen science projects to date rely on people's visual abilities. So you're looking at images and you're trying to classify them. You're trying to look at color shape, you're counting things. And that makes a lot of sense because humans, you know, visual ability is nearly universal, right? So our effort in citizen science actually didn't use the visual capabilities. We used a different capability that I think satisfy the same criteria of being nearly universal and humans are better than computers. And that is in the area of language comprehension. So our ability for humans to read and, and comprehend and infer context, uh, again, these are all things that are very well developed. And so we tried to harness citizen scientists to try to read biomedical literature, to try to tackle the big challenge I mentioned before the fact that there are, you know, a new paper every 25 seconds.

Andrew (12:56):

And how do we extract sort of the key findings, the key nuggets of information. So we developed an application, we called Mark2Cure and it was focused on rare disease, a particular rare disease actually focused on NGly1 deficiency. And our goal was to say, okay, let's get all the papers we can that are even remotely related to NGly1 and try to extract what we call a knowledge network out of those papers. And the knowledge network essentially is a graph where we have nodes and edges. The nodes are things in biology, they're genes, their diseases, their phenotypes, their chemical compounds, and the edges are relationships between those things. So how different chemical compounds interact with genes and gene products and things like that. And we wanted to know, can we train essentially citizen scientists to help us identify these entities and extract relations between those objects out of some pretty dense, scientific texts? Uh, we had some great contributors actually from a, a high school in Alabama where their teacher sort of made it part of their curriculum or maybe it was an extracurricular activity, but that class in total, you know, contributed quite a bit of content and quite a bit of effort to our initiative.

Drew (14:18):

And that kinda speaks to incentives. So you just brought up that great example of a teacher getting their high school students to engage in this as part of class, but then sort of on the broader scale, how do we get the public to engage in this type of community science? I mean obviously if they find the topic interesting, like astronomy, then that's great, but otherwise, you know, should there be financial incentives there? Could there be different grants, you know, make it a game like, and how would that maybe skew the science either positively or negatively?

Andrew (14:46):

Yeah. Well you hit on a whole number of, of great incentives. And I think there are people who are trying all of those and more and combinations of those. I'll say that gamification one that you mentioned is absolutely one approach that has been used successfully in a, a lot of context because people like having fun. And if you can couch your problem in a way that it's fun, then even if you can just harness a small slice of the overall gaming community's time, there's a huge amount of human resource there. We looked into this also quite a bit, and I remember one statistic said something on the effect of we as a society spends something like 9 billion hours per year on gaming. And it was something like, it took 3 billion human hours to create the Panama canal. So for all the human effort that goes to gaming, right, we can make three Panama canals a year. So again, if you, if you can harness even a small sliver of that, by making a game, that's engaging to some small population of the gaming and community, even that could be a huge amount of human, uh, resource that could be tapped into,

Drew (15:54):

I think you just summarized our culture, right there. <laugh>

Andrew (15:57):

There's value in games. Right?

Drew (15:59):

I know. I know. I like it too.

Andrew (16:01):

Yeah. But aside from gaming and, and enjoyment, right, there's a lot of citizen scientists that are motivated by other things, right? So we surveyed our Mark2Cure participants. A lot of people were retired and wanted to keep intellectual stimulation right. And learn things. So education is obviously one important part of citizen science. So giving people access to experts and sort of how science actually works, I think is incredibly motivating. A lot of people not surprisingly are just do-gooders, right? They wanna make a positive impact in the world. And so that altruistic spirit is definitely something we saw. We saw many contributors who weren't affected by NGly1 deficiency, which is this ultra-rare disease, but were affected by other diseases. And so it wasn't directly benefiting them, but they felt very a personal connection to the idea of biomedical research in particular. So there's a lot of really great and interesting motivations that bring people to the citizen science.

Drew (17:08):

If these knowledge networks sound complex and foreign to you, the chances are that they're at work every time you go online or watch your favorite show on TV. In fact, much of the headway made in integrating vast streams of information has been driven by the commercial interests of the tech industry.

Andrew (17:24):

The baseline application of these massive knowledge networks is simply to find the information that we'd wanna get. Right? Show me all papers that describe genes involved in neurological cancers, right? Show me all papers that involve a certain class of chemical compounds. So it's, it's just a way to more easily find and retrieve articles relevant to what you're interested in. But that to me is just the start knowledge graphs are becoming a, a common way of representing information in large part driven by high tech industry. And as you alluded to, right, social networks are a great example of how they're being used in high tech. Okay. And so if you think about how social networks or knowledge networks more broadly in high tech are being used, you could think about it like this, right? Netflix wants to develop algorithms to predict movies that you may want to watch next.

Andrew ([18:22](#)):

And that is gonna be based on your movie, watching history. It's based on everybody else's movie watching history, it could be based on relationships, you know, family relationships or friend relationships between you and other people. And so this idea of predicting potential links in a large network that would join for example, you and this potential movie you might wanna watch. That's what we call knowledge, graph reasoning. So we're trying to predict edges in a knowledge graph that don't yet exist, but maybe should exist that idea of predicting what movie you want to view next may also be used to predict what drugs may be used to treat given diseases; can be used to predict what chemical compounds be used to modulate different biological pathways. It's the same idea. You have massive knowledge networks, and you're trying to predict edges that are supported by the evidence, but aren't yet explicitly enumerated there. So there's great hope that we can use the same algorithms or variance of them for knowledge graph, reasoning that are being developed in high tech, use those algorithms or adapt those algorithms to biological knowledge networks, to make predictions of new biological insights that then can be tested in the lab.

Drew ([19:44](#)):

Yeah. That <laugh> those Netflix algorithms definitely work. <laugh> Talk about 9 billion hours... <laugh>

Andrew ([19:51](#)):

Yes, yes.

Drew ([19:52](#)):

Yeah. Cool. And opening this up now into sort of the broader view on AI and medicine going forward, I'm curious where you think either where we have seen the biggest progress or where you think we're gonna see some major progress going forward, either in knowledge networks or just AI in general.

Andrew ([20:07](#)):

Well, look, I mean, we've already in the past maybe 10 years or so seen huge strides in image based AI. So we now have all sorts of algorithms for predicting, for example, melanomas, right? If you have a dark spot on your skin, right. Predicting whether that's a melanoma or not, we have algorithms for predicting breast cancer based off of mammography images, predicting diabetic retinopathy based off of, uh, images of your eyes. So image based analysis really is leading the charge. And I think that that will continue to be a fruitful area of innovation, both in medicine and in biomedical research. And yeah, I'm very hopeful that knowledge graphs and mining of these information networks will continue to mature. I think we're not quite there yet in terms of where image based AI is, but definitely there are many academic labs who are interested in that topic as well as companies and startups that are pursuing that area.

Drew (21:13):

Yeah. It's amazing progress, especially in those visual fields. And it just makes me wonder, you know, what's gonna happen to the field of say dermatology or X-rays just with doctors not being as good or not being needed for that field. I don't really know.

Andrew (21:27):

Look the way I view it is that I hope that again, we push experts further down the value chain that we better utilize their talents and their time and their energy towards hard earned harder cases. Uh, let's leave the easy stuff to machines and crowdsourcing systems and really free experts up to do things that really require their expertise.

Drew (21:51):

Got it. Yeah. Direct the productivity elsewhere.

Andrew (21:53):

Exactly.

Drew (21:55):

Alright. Well, I'm gonna put you on the spot and see if you have any concerns at all about AI. So people talk about the singularity, you know, this point in time, which technological growth kind of could become uncontrollable and irreversible and often people sort of conflate that with kind of machines taking over. So does that keep you up at night at all?

Andrew (22:14):

I do think a lot about making sure our advances are equitable because, you know, for example, right, we just think about the amount of genome sequencing that we have, right. And how much of it is based on country is in the global north and United States and, and Europe in, in particular and whether or not sort of the benefits of biotechnology and AI and all of these advances are equitably sort of distributed among people. And so, I think there's certainly awareness of this issue and, and many initiatives to, for example, if we're staying on genome sequencing to get more broader representation among genome sequencing and human variation, but I certainly also worry about some of those biases creeping into the knowledge networks that we're generating, right? Much of machine learning relies on training data, right. Relies on real examples. So we can learn the patterns in the knowledge networks that are indicative of the type of thing we're trying to find. For example, drugs, treating diseases. We need many examples of true drugs that treat true diseases. So we can learn the patterns in the knowledge network that join those two entities. But if the knowledge networks or our training data are biased in certain ways, then essentially the benefits are not equally shared. And that does, that does worry me.

Drew (23:45):

I see the machine learning could run away with a data set that is sort of not representative to begin with. And then all the outputs are then wrong.

Andrew (23:53):

Exactly.

Drew (23:54):

Yeah, yeah. Good perspective there. Okay. Well, when you're not messing around with knowledge networks, do you have any other hobbies or obsessions when you're not in the lab other than going to Disneyland? Cause I know you just went… <laugh>

Andrew (24:06):

<laugh> well, um, uh, Disneyland is an example of the broader activity, which is chasing around my two sons and, and trying to keep them entertained baseball for my older one or arts and crafts for the younger one. There's always something, uh, with those guys that keeps me busy.

Drew (24:23):

Yeah. It probably gives you a nice creative break away from the data too. It's probably helpful when you go back to it.

Andrew (24:29):

Yep. That is absolutely true. Yeah. I will say that there's one silver lining of pandemic is that, you know, my lab's been working remotely for oh yeah. We just passed two years. Um, working remotely and for computational work working remotely is actually not that big of a deal. And so what I've discovered is that it's actually great. I can take a lot of afternoons out to go to little league practice and then do works in the morning, work in the evenings and maintain, uh, productivity, but still interact a lot with my, my kids, uh, during the middle of the day.

Drew (24:59):

Yeah. Yeah. You chose the right field for a pandemic. Definitely.

Andrew (25:02):

<laugh> it's it's worked out. Okay.

Drew (25:04):

All right. Well we'll just wrap things up and I'll ask my final Roundup question, which I like to throw all of my guests, which is if you could give one piece of advice or your wisdom to anyone, and this could be in the realm of work or career progression, life health, self-improvement really anything. What do you think it would be and why?

Andrew (25:19):

I think my one tidbit that I think a lot about recently is in the area of being the head of a lab and especially in academic science, there's a lot of smart people who do science, but a lot of responsibility that comes with that. There's a lot of great people who do science, but distinguishing one's self for being a good mentor; I think is one of the most valuable things for me. We hear a lot about bad apples who run labs that are overly aggressive or overly high pressure. And to me, one of the most satisfying parts of this job has been helping trainees progress in whatever direction they want. And so I think encouraging, especially young investigators to distinguish yourself, uh, not just by your science, but essentially how you treat your trainees has served me well in my lab in any case.

Drew (26:09):

Solid final words there. I mean, who are we without those mentors inspiring us and guiding us along the way? I'm grateful that Andrew could join me today and show us the value of community intelligence. Perhaps this might inspire you to become a citizen scientist. If that does pique your interest, then check the show notes to learn more about current projects. We'll also have some more links in there to Andrew's work and our latest feature content from the Scripps Research magazine. Thank you as always for listening and remember to hit subscribe and leave us a review. As you now know, it really helps with those tech algorithms. And before we meet again for more breakthrough topics, stay curious, stay grounded and be well.