Interviews with Experts
Public

Join

Discussion Media Files Members About

Back

GP Giovanni Paternostro
17 days ago

Interview with Trey Ideker

Trey Ideker has been appointed as the new Director of the University of Oxford's Big Data Institute (BDI). He will take up the role in June. He is currently a Professor in the Department of Medicine at the University of California, San Diego (UCSD), where he also leads and co-leads several major data-driven research initiatives, including an ADAPT Center for Precision Oncology, the Cancer Cell Map Initiative, and the Bridge2AI Functional Genomics Data Generation Program.

Dear Trey,
What are your plans for the Oxford's Big Data Institute? What is your vision of data science and AI, and, more broadly, of science in this age of advancing AI?

Trey:

Coming to the BDI at this time presents a significant opportunity since, as you know, AI is at such an inflection point. As a direct result, biomedicine and biomedical data science are also at an inflection point. What the Big Data Institute has done extraordinarily well so far, and arguably better than anyone else in the world, is biobanks. An example is the UK Biobank, but there are many other Oxford biobanks, based in either the Big Data Institute or the Nuffield Department of Population Health. China Kadoorie Biobank is another example, yet another is the Mexican Biobank. Some of those efforts have inspired other UK national initiatives, like Genomics England and Our Future Health.

One of the immediate values of these biobanks is to make lots of associations, for example between the incidence of two phenotypes, such as a behavior and an ensuing disease, but also associations of genetics with disease in so-called genetic risk associations. Many of the genome-wide association studies that we've seen published in the past decade are aided directly from UK Biobank samples or from other Oxford biobanks. Now, what do all those associations do? They produce many potential links between genes and disease, which if you understood the mechanistic basis of those, would provide a key, not just to predicting risk, but to understanding how to treat those diseases. They help find the drug targets within the causal pathways connecting genotype to phenotype. And that's where the Big Data Institute, and someone like me coming in as director, really stands to make an impact.

Members

GP Giovanni Paternostro Follow

See All Members (1)

In particular, my lab has spent the past several decades advancing the fields of functional genomics and proteomics, more broadly known as "systems biology". On the one hand, lots of progress has already been made at the BDI and elsewhere in connecting structural or functional omics to genotype-phenotype associations. On the other, there's clearly a need to step up those efforts and I certainly imagine this was recognized by the Oxford leadership in considering candidates like me.

Accordingly, my clear mandate here is to roll out infrastructure and concepts that will let us understand the mechanistic, molecular pathways underlying the major diseases and conditions represented in the biobanks. These biobanks have thousands of phenotypic traits, not just "disease/non disease", but things like full body CAT scans, blood work, and now also in-depth proteomics.

When cell and systems biologists think "big data", we typically mean something different than population resources. We mean big omics datasets, like a massive proteomics repository or a single cell study. Of course, clearly you need both types of big data — big omics alongside big population health data. We will include data at both the molecular and the population level.

As a side note, the new position at Oxford BDI greatly magnifies a role I have thus far been playing mainly within NIH centers, which can function like virtual institutes. NIH-funded centers typically knit together 8 or 10 faculty, and over the past decade, my lab has been involved in several of these. One example is the Cancer Cell Map initiative, which is now just finishing its second five-year term. The Big Data institute at Oxford will take these efforts to a level where you can really bring a vision that impacts and coordinates the research of many people. The Big Data institute is certainly an order of magnitude larger than any of the centers that I've been involved with so far. And it's a physical institute as opposed to a virtual institute, with a dedicated building at the heart of the Oxford health sciences campus.

Giovanni:
Since you are talking about big data, someone with whom we have been talking often is Soren Brunak, which is involved in similar initiatives in Denmark. Denmark is another country where they have a large amount of long-term health data, collected over several decades from the entire population.

Trey:

Soren Brunak is someone I've known for a very long time. My first postdoc, Chris Workman, came from Soren's lab. And even starting back in those days twenty years ago, Soren has been an enormous supporter of my research and mentor to me personally. I look forward to connecting with Soren again. There are many potential synergies between his work and what we have planned at the Big Data Institute.

Giovanni:
In another interview you mentioned that both cells and brains exhibit intelligent behaviors, and that they have a complexity of the same order. There might also be analogies with AI and collective intelligence. Are there similarities among these network-based types of intelligence, even if they have different hardware? Can knowledge of one type of intelligence inspire questions to be asked in another type? Neuroscience researchers have been talking about distributed intelligence in the brain, a concept that is difficult to

grasp completely, but now, with AI, we built one. So now we know that it is possible.

Trey:

I think the essence of your question relates to the main differences between engineering and science. As you mentioned, for AI we built it. For the brain we did not; for human cells we did not. AI, brains and cells, however, share something very important in common. They are all complex systems. When we build a thing, that activity falls into the realm of engineering, the use of knowledge to design and build solutions to practical problems. When we seek to understand a thing in nature, that activity is considered to be entirely different – it falls into the realm of science, the systematic study of the natural world.

What we've seen emerge over the last year is that AI is now so complex that people study it scientifically, much more like a product of nature than something that has been engineered and therefore is by definition completely understood. For example, many recent attempts to understand large language model and agentic systems resemble methodology that is quite like studying biology. The number of parameters, layers of artificial neurons, and so on, has grown so large that the way that people study these artificial brains is very similar to the way that we study real brains, or other complex biological systems like cells and tissues. First you perturb, then you see what happens in response to your perturbation in terms of causes and effects. That's the way those computer science papers are now being written, because we do not fully understand the behavior of these systems, even though we engineered them. AI developers now talk about the principle of "emergence", just as people talked about emergent properties in biology or physics many years ago.

It's worth mentioning how the Big Data institute can leverage and interact with AI and collective intelligence moving forward. Certainly, machine learning is not newly important to biology or to understanding the molecular basis for disease. However, we're now adding to the classical mode of machine learning, where you look for patterns in inputs that predict biological outputs with high accuracy. Agentic systems and large language models now provide another mode of biology that is very different, and I think both modes are going to be pursued moving forward. We will still, on the one hand, try to gather enormous data sets, aiming to learn how patterns in molecular and physiological inputs cause a behavior or a disease. But the other way to approach this problem is by literature and data synthesis followed by hypothesis generation using large language models.

The classic approach would be: here's a massive data set, please build a machine learning model that can accurately classify diabetes cases from controls, using the genome and the molecular profile of the individual. The new version is: Dear Chatbot, here are 1000 single nucleotide polymorphisms that have been associated with diabetes. Please formulate a variety of plausible, alternative hypotheses for their mechanisms of action. At the moment there's a lot of activity in this space, including from many startup companies. But where I'm most interested in pushing deeply is into large-scale experimental biology. That's where the infrastructure established at Oxford surrounding the Big Data institute, the Oxford Center for Human Genomics and the Oxford Centre for Medicines Discovery really has the potential to blow open the space.

How do we get there from here? I think the primary bottleneck currently falls on the scientists. I mean, ultimate success at implementing AI in

the life sciences is almost certainly going to be driven by the scientists themselves. I realize I have to be careful here, because the engineers of course are the ones responsible for nearly all recent developments in AI so far. But science and engineering mindsets are so very, very different.

To relate a personal example, my undergraduate degree at MIT was in computer science and electrical engineering. I switched over to biology only for my PhD in the late 90's, then spent the next 5 years deep in the molecular biology laboratory carrying out some of the first microarray gene expression screens with Leroy Hood. I was then in my late 20's, but I have to admit I was probably 40 before my lab was really productive at hypothesis-driven scientific research. In contrast to efforts to build tools (aka "if you build it they will come") or to generate large-scale experimental screens (aka "discovery science"), hypothesis-driven science is such a different mode of thinking.

To sum up, the introduction of AI in life sciences over the next few years really must be done carefully, deliberately and incrementally, much more so than many are seeming to claim at the moment, and it's got to be with life scientists in the loop. From here on out, experimental science will almost certainly be done only by AIs and scientists working together.

Giovanni:
My last question is about virtual cells and digital models of organisms. These are broad efforts for which many biomedical scientists can contribute knowledge and ideas. And then people like those working at your institute could integrate them.

Trey:

Yes, virtual cell (and tissue) models are quite central to the BDI's vision for how you understand, and ultimately alter, genotype-phenotype relationships. I know we are almost out of time, but you might check out digitaltumors.org, our virtual tumor cell effort funded by the ARPA-H ADAPT precision oncology program. Exactly as you're talking about.

Giovanni:
Many thanks Trey, I really enjoyed this conversation!