PETER SZOLOVITS: So today I'm going to talk about precision medicine. And we don't really have a very precise idea of what precision medicine is. And so I'm going to start by talking about that a little bit. David talked about disease subtyping. And if you think about how do you figure out what are the subtypes of a disease, you do it by some kind of clustering on a bunch of different sorts of data.

And so we have data like demographics, comorbidities, vital signs, medications, procedures, disease trajectories, whatever those mean, image similarities. And today, mostly I'm going to focus on genetics. Because this was the great hope of the Human Genome Project, that as we understood more about the genetic influences on disease, it would help us create precise ways of dealing with various diseases and figuring out the right therapies for them and so on. So I want to start by reviewing a little bit a study that was done by the National Research Council, so the National Academies, and it's called "Toward Precision Medicine." This was fairly recent, 2017.

And they have some interesting observations. So they start off and they say, well, why is this relevant now, when it may not have been relevant before? And of course, the biggie is new capabilities to compile molecular data on patients on a scale that was unimaginable 20 years ago. So people estimated that getting the first human genome cost about $3 billion. Today, getting a human genome costs less than $1,000. I have some figures later in the talk showing some of the ads that people are running.

Increasing success in utilizing molecular information to improve diagnosis and treatment, we'll talk about some of those. Advances in IT so that we have bigger capabilities of dealing with so-called big data-- a perfect storm among stakeholders that has made them much more receptive to this kind of information. So the fact that costs in the health-care system in the US keep rising and quality doesn't keep rising proportionately makes everybody desperate to come up with new ways of dealing with this problem. And so this looks like the next great hope for how to do it.

And shifting public attitudes toward molecular data-- so how many of you have seen the movie *Gattaca?* A few. So that's a dystopian view of what happens when people are genotyped and can therefore be tracked by their genetics. And it is true that there are horror stories that can happen. But nevertheless, people seem to be more relaxed today about allowing that kind of

data to be collected and used. Because they see the potential benefits outweighing the costs. Not everybody-- but that continues to be a serious issue.

So this report goes on and says, you know, let's think about how to integrate all kinds of different data about individuals. And they start off and they say, you know, one good example of this has been Google Maps. So Google Maps has a coordinate system, which is basically longitude and latitude, for every point on the Earth. And they can use that coordinate system in order to layer on top of each other information about postal codes, built structures, census tracts, land use, transportation, everything. And they said, wow, this is really cool, if only we could do this in health care.

And so their vision is to try to do that in health care by saying, well, what corresponds to latitude and longitude is individual patients. And these individual patients have various kinds of data about them, including their microbiome, their epigenome, their genome, clinical signs and symptoms, the exposome, what are they exposed to. And so there's been a real attempt to go out and create large collections of data that bring together all of this kind of information.

One of those that is notable is the Department of Health-- well, NIH basically started a project about a year and a half ago called All of Us, sounds sort of menacing. But it's really a million of us. And they have asked institutions around the United States to get volunteers to volunteer to provide their genetic information, their clinical data, where they live, where they commute, things like that, so that they can get environmental data about them. And then it's meant to be an ongoing collection of data about a million people who are supposed to be a representative sample of the United States.

So you'll see in some of the projects I talk about later today that many of the studies have been done in populations of European ancestry. And so they may not apply to people of other ethnicities. This is attempting to sample accurately so that the fraction of African Americans and Asians and Hispanics and so on corresponds to the sample in the United States population.

There's a long history. How many of you have heard of the Framingham Heart Study? So a lot of people.

So Framingham, in the 1940s, agreed to become the subject of a long-term experiment. I think it's now run by Boston University, where every year or two they go out and they survey-- I can't remember the number of people. It started as something like 50,000 people-- about their

habits and whether they smoke, and what their weight and height is, and any clinical problems they've had, surgical procedures, et cetera.

And they've been collecting that database now over several generations of people that descend from those. And they've started collecting genetic data as well. So All of Us is really doing this on a very large scale.

Now, the vision of these of this study was to say that we're going to build this information commons, which collects all this kind of information, and then we're going to develop knowledge from that information or from that data. And that knowledge will become the substrate on which biomedical research can rest. So if we find significant associations, then that suggests that one should do studies, which will not necessarily be answered by the data that they've collected. You may have to grow knock-out mice or something in order to test whether an idea really works. But this is a way of integrating all of that type of information. And of course, it can affect diagnosis, treatment, and health outcomes, which are the holy grail for what you'd like to do in medicine.

Now, here's an interesting problem. So the focus, notice, was on taxonomies. So Sam Johnson was a very famous 17th century British writer. And he built encyclopedias and dictionaries, and was a poet and a reviewer and a commentator, and did all kinds of fancy things. And one of his quotes is, "My diseases are an asthma and a dropsy and, what is less curable, 75," years old. So he was funny, too.

Now, if you look up dropsy in a dictionary-- how many of you have heard of dropsy? A couple. So how did you hear of it?

AUDIENCE: From Jane Austen novels. [LAUGHS]

PETER SZOLOVITS: Sorry? From a novel?

AUDIENCE: Novels.

PETER SZOLOVITS: Yeah.

AUDIENCE: I've heard of dropsy [INAUDIBLE].

PETER SZOLOVITS: Yeah. I mean, I learned about it by watching Masterpiece Theatre with 19th century people, where the grandmother would take to her bed with the dropsy. And it didn't turn out well, typically. But it took a long time for those people to die.

So dropsy is water sickness, swelling, edema, et cetera. It's actually not a disease. It's a symptom of a whole bunch of diseases. So it could be pulmonary disease, heart failure, kidney disease, et cetera.

And it's interesting. I look back on this. I couldn't find it for putting together this lecture. But at one point, I did discover that the last time dropsy was listed as the cause of death of a patient in the United States was in 1949. So since then, it's disappeared as a disease from the taxonomy.

And if you talk to pulmonary people, they suspect that asthma, which is still a disease in our current lexicon, may be very much like dropsy. It's not a disease. It's a symptom of a whole bunch of underlying causes. And the idea is that we need to get good enough and precise enough at being able to figure out what these are.

So I talked to my friend Zack Kohane at Harvard a few weeks ago when I started preparing this lecture. And he has the following idea. And the example I'm going to show you is from him. So he says, well, look, we should have this precision medicine modality space, which is this high-dimensional space that contains all of that information that is in the NRC report. And then what we do is, in this high-dimensional space, if we're lucky, we're going to find clusters of data.

So this always happens. If you ever take a very high-dimensional data set and put it into its very high-dimensional representation space, it's almost never the case that the data is scattered uniformly through the space. If that were true, it wouldn't help us very much. But generally, it's not true. And what you find is that the data tends to be on lower-dimensional manifolds. So it's in subsets of the space.

And so a lot of the trick in trying to analyze this kind of data is figuring out what those lower-dimensional manifolds look like. And often you will find among a very large data set a cluster of patients like this. And then Zack's approach is to say, well, if you're patient-- it's hard to represent three dimensions in two. But if you're patient that falls somewhere in the middle of such a cluster, then that probably means that they're kind of normal for that cluster, whereas if they fall somewhere at the edge of such a cluster, that probably means that there's something

odd going on that is worth investigating, because they're unusual.

So then he gave me an example of a patient of his. And let me give you a minute to read this. Yeah?

AUDIENCE: What's an armamentarium?

PETER SZOLOVITS: Where does it say armamentarium?

AUDIENCE: [INAUDIBLE]

PETER SZOLOVITS: Oh, yeah. So an armamentarium, historically, is the set of arms that are available to an army. So this is the set of treatments that are available to a doctor.

AUDIENCE: Is that the only word you don't know?

[LAUGHTER]

It's the only word--

AUDIENCE: If I start asking--

AUDIENCE: Based on [INAUDIBLE].

AUDIENCE: Oh, OK.

AUDIENCE: In the world. Some of it, I thought I could understand.

PETER SZOLOVITS: Well, you probably know what antibiotics are. And immunosuppressants, you've probably heard of. Anyway, it's a bunch of different therapies.

So this is what's called a sick puppy. It's a kid who is not doing well. They started life, at age three, with ulcerative colitis, which was well-controlled by the kinds of medications that they normally give people with that disease. And then all of a sudden, 10 years later, he breaks out with this horrible abdominal pain and diarrhea and blood in his stool. And they try a bunch of stuff that they think ought to work, and it doesn't work.

So the kid was facing some fairly drastic options, like cutting out the part of his colon that was inflamed. So your colon is an important part of your digestive tract. And so losing it is not fun

and would have bad consequences for the rest of his life.

But what they did is they said, well, why is he not responding to any of these therapies? And the difficulty, you can imagine, in that cloud-of-points picture, is, how do you figure out whether the person is an outlier or is in the middle of one of these clusters, when it depends on a lot of things? In this kid's case, what it depended on most significantly was the last six months of his experience, where, before, he was doing OK with the standard treatment. So that cloud might have represented people with ulcerative colitis who were well-controlled by the standard treatment. And now, all of a sudden, he becomes an outlier.

So what happened in this case is they said, well, maybe there are different groups of ulcerative colitis patients. So maybe there are ones who have a lifelong remission after treatment with a commonly used monoclonal antibody. So that's the center of the cluster.

Maybe there are people who have multi-year remission but become refractory to these drugs. And after other treatments, they have to undergo a colectomy. So that's the removal of the colon. And then there are people who have, initially, a remission, but then those standard therapy works. So that's what this kid is in, this cluster.

So how do you treat this as a machine learning problem from the point of view of having lots of data about lots of different patients? And the challenges, of course, include things like, what's your distance function in doing the kind of clustering that people typically do? How do you define what an outlier is? Because there's always a continuum where it just gets more and more diffuse.

What's the best representation for time-varying data, which is critical in this case? What's the optimal weighting or normalization of dimensions? So does every dimension in this very high-dimensional space count the same? Or are differences along certain dimensions more important than those among others? And does that, in fact, vary from problem to problem? The answer is probably yes.

So how do we find the neighborhood for the patient? Well, I'm going to give you some clues by starting with a shallow dive into genetics. So if you've taken a molecular cell biology class, this should not be news to you. And I'm going to run through it pretty quickly. If you haven't, then I hope at least you'll pick up some of the vocabulary.

So a wise biologist said, "Biology is the science of exceptions." There are almost no rules.

About 25 years ago, the biology department here taught a special class for engineering faculty to try to explain to us what they were teaching in their introductory biology, molecular biology classes.

And I remember, I was sitting next to Jerry Sussman, one of my colleagues. And after we heard some lecture about the 47 ways that some theory doesn't apply in many, many cases, Jerry turns to me and he says, you know, the problem with this field is there are just too many damned exceptions. There are no theories. It's all exceptions. And so even biologists recognize this.

Now, people have observed, ever since human beings walked the earth, that children tend to be similar to their parents in many ways. And until Gregor Mendel, this was a great mystery. Why is it that you are like your parents? I mean, you must have gotten something from them that sort of carries through and makes you similar to them.

So Mendel had this notion of having discrete factors of inheritance, which he called genes. He had no idea what these were. But conceptually, he knew that they must exist. And then he did a bunch of experiments on pea plants, showing that peas that are wrinkled tend to have offspring peas that are also wrinkled. And he worked out the genetics of what we now call Mendelian inheritance, namely dominant versus recessive inheritance patterns.

Then Johann Miescher came along some years later, and he discovered a weird thing in cells called nuclein, which is now known as DNA. But it wasn't until 1952 that Hershey and Chase said, hey, it's DNA that is carrying this genetic information from generation to generation. And then, of course, Watson, Crick, and Franklin, the following year, deciphered the structure of DNA, that it's this double helix, and then figured out what the mechanism must be that allows DNA to transmit this information.

So you have a double helix. You match the four letters A, C, T, G opposite each other, and you can replicate this DNA by splitting it apart and growing another strand that is the complement of the first one. Now you have two. And you can have children, pass on this information to them. So that was a big deal.

So a gene is defined by the National Center for Biotechnology Information as a fundamental physical and functional unit of heredity that's a DNA sequence located on a specific site on a chromosome which encodes a specific functional product, namely RNA or a protein. I'll come back to that in a minute. The remaining mystery is it's still very hard to figure out what parts of

the DNA code genes.

So you would think we might have solved this, but we haven't quite. And what does the rest, which is the vast majority of the DNA, do if it's not encoding genes? And then, how does the folding and the geometry, the topology of these structures, influence their function?

So I went back and I read some of Francis Crick's work from the 1950s. And it's very interesting. This hypothesis was considered controversial and tentative at the time. So he said, "The specificity of a piece of nucleic acid is expressed solely by the sequence of its bases, and this sequence is a simple code for the amino acid sequence of a particular protein." And there were people arguing that he was just flat wrong, that this was not true. Of course, it turned out he was right.

And then the central dogma is the transfer of information from nucleic acid to nucleic acid or from nucleic acid to protein may be possible. But transfer from protein to protein or from protein to nucleic acid is impossible. And that's not quite true. But it's a good first approximation. So this is where things stood back about 60 years ago.

And then a few Nobel prizes later, we began to understand some of the mechanism of how this works. And of course, how it works is that you have DNA, which is these four bases, double stranded. RNA gets produced in the process of transcription. So this thing unfolds. An RNA strand is built along the DNA and separates from the DNA, creating a single-stranded RNA. And then it goes and hooks up with a ribosome. And the ribosome takes that RNA and takes the codes in triplets, and each triplet stands for a particular amino acid, which it then assembles in sequence and creates proteins, which are sequences of amino acids.

Now, it's very complicated. Because there's three-dimensionality and there's time involved. And the rate constants-- this is chemistry, after all. So again, a few more Nobel prizes later, we have that transcription, that process of turning DNA into RNA, is regulated by promoter, repressor, and enhancer regions on the genome. And the proteins mediate this process by binding to the DNA and causing the beginning of transcription, or causing it to run faster or causing it to run slower, or they interfere with it, et cetera.

There are also these enhancers, some of which are very far away from the coding region, that make huge differences in how much of the RNA, and therefore how much of the protein, is made. And the current understanding of that is that, if here is the gene, it may be that the strand of DNA loops around. And the enhancer, even though it's distant in genetic units, is

actually in close physical proximity, and therefore can encourage more of this transcription to take place.

By the way, if you're interested in this stuff, of course MIT teaches a lot of courses in how to do this. Dave Gifford and Manolis Kellis both teach computational courses in how to apply computational methods to try to decipher this kind of activity.

So repressors prevent activator from binding or alters the activator in order to change the rate constants. And so this is another mechanism. Now, one of the problems is that if you look at the total amount of DNA in your genes, in your cells, only about 1 and 1/2% are exons, which are the parts that code for mRNA, and eventually protein. So the question is what does the other 98 and 1/2% do?

There was this unfortunate tendency in the biology community to call that junk DNA, which of course is a terrible notion. Because evolution would certainly have gotten rid of it if it was truly junk. Because our cells spend a lot of energy building this stuff. And every time a cell divides, it rebuilds all that so-called junk DNA. So it can't possibly be junk.

But the question is, what does it do? And we don't really know for a lot of it. So there are introns-- I'll show you a picture. There are segments of the coding region that don't wind up as part of the RNA. They're spliced out. And we don't quite know why.

There are these regulatory sequences, which is only about 5%, that are those promoters and repressors and enhancers that I talked about. And then there's a whole bunch of repetitive DNA that includes transposable elements, related sequences. And mostly, we don't understand what it all does.

Hypotheses are things like, well, maybe it's a storehouse of potentially useful DNA so that if environmental conditions change a lot, then the cell doesn't have to reinvent the stuff from scratch. It saved it from previous times in evolution when that may have been useful. But that's pretty much pure speculation at this point.

So just recently, the Killian Lecture was given by Gerald Fink, who's a geneticist here. And his claim is that a gene is not any segment of DNA that produces RNA or protein. But it's any segment of DNA that is transcribed into RNA that has some function, whatever it is, not necessarily building proteins, but just anything. And I think that view is becoming accepted.

So I promised you a little bit of more complexity. So when you look at your DNA in eukaryotes, like us, here's the promoter. And then here is the sequence of the genome. When this gets transcribed, it gets transcribed into something called pre-mRNA, messenger RNA.

And then there's this process of alternative splicing that splices out the introns and leaves only the exons. But sometimes it doesn't leave all the exons. It only leaves some of them. And so the same gene can, under various circumstances, produce different mRNA, which then produces different proteins.

And again, there's a lot of mysteries about exactly how all this works. Nevertheless, that's the basic mechanism. And then here, I've just listed a few of the complexity problems.

So there are things like, RNA can turn into DNA. This is a trick that viruses use a lot. They incorporate themselves into your cell, create a DNA complement to the RNA, and then use that to generate more viruses. So this is very typical of a viral infection.

Prions, we also don't understand very well. This is like mad cow disease, where these proteins are able to cause changes in other proteins without going through the RNA/DNA-mediated mechanisms.

There are DNA-modifying proteins, the most important of which is the stuff involved in CRISPR-CAS9, which is this relatively new discovery about how bacteria are able to use a mechanism that they stole from viruses to edit the genetic complement of themselves, and more importantly, of other viruses that attack them. So it's an antiviral defense mechanism. And we're now figuring out how to use it to do gene editing.

You may have read about this Chinese guy who actually went out and edited the genome of a couple of girls who were born in China, incorporating some, I think, resistance against HIV infections in their genome. And of course, this is probably way too early to do experiments on human beings, because they haven't demonstrated that this is safe. But maybe that'll become accepted.

George Church at Harvard has been going around-- he likes to rattle people's chains. And he's been going around saying, well, the guy, he was unethical and was a slob, but what he's doing is a really great idea. So we'll see where that goes.

And then there are these retrotransposons, where pieces of DNA in eukarya just pop out of wherever they are and insert themselves in some other place in the genome. And in plants,

this happens a lot. So for example, wheat seems to have a huge number of copies of DNA segments that maybe it had only one of, but it's replicated through this mechanism.

Last bit of complexity-- so we have various kinds of RNA. There's long non-coding RNA, which seems to participate in gene regulation. There is RNA interference, that there are these small RNA pieces that will actually latch onto the RNA produced by the standard genetic mechanism and prevent it from being translated into protein. This was another Nobel Prize a few years ago. Almost everything in this field, if you're first, you get a Nobel Prize for it.

Once the proteins are made, they're degraded differentially. So there are different mechanisms in the cell that destroy certain kinds of proteins much faster than others. And so the production rate doesn't tell you how much is going to be there at any particular time.

And then there's this secondary and tertiary structure, where there's actually-- what is it? It's a mile of DNA in each of your cells. So it wouldn't fit. And so it gets wrapped up on these acetylated histones to produce something called chromatin.

And again, we don't quite understand how this all works. Because you'd think that if you wrap stuff up like this, it would become inaccessible to transcription. And therefore, it's not clear how it gets expressed. But somehow or other, the cell is able to do that. So there's a lot yet to learn in this area.

Now, the reason we're interested in all this is because, if you plot Moore's law for how quickly computers are becoming cheaper per performance, and you plot the cost of gene sequencing, it keeps going down. And it goes down much faster even than Moore's law. So this is pretty remarkable. And it means that, as I said, that $3 dollar first genome now costs just a few hundred dollars.

In fact, if you're just interested in the whole exome, so only the 2%, roughly, of the DNA that produces genetic coding, you can now go to this company, which I have nothing to do with. I just pulled this off the web. But for $299, they will give you 50-times coverage on about six gigabases. And if you pay them an extra $100, they'll do it at 100x coverage. So these techniques are very noisy. And so it's important to get lots of replicates in order to reassemble what you think is going on.

A slightly more recent phenomenon is people say, well, not only can we sequence your DNA but we can sequence the RNA that got transcribed from the DNA. And in fact, you can buy a

kit for $360 that will take the RNA from individual cells-- so these are like picoliter amounts of stuff. And it will give you the RNA sequence for $360 for up to 100 cells, so $3, $3.50 per cell.

So people are very excited. And there are now also companies that will sell you advanced analysis. So they will correlate the data that you are getting with different databases and figure out whether this represents a dominant or a recessive or an x-linked model, if you have family familial data and functional annotation of candidate genes, et cetera.

And so, for example, starting about three years ago, if you walk into the Dana-Farber with a newly diagnosed cancer, a solid-tumor cancer, they will take a sample of that cancer, send it off to companies like this, or their own labs, and do sequencing and do analysis and try to figure out exactly which damaged genes that you have may be causing the cancer, and maybe more importantly, since it's still a pretty empirical field, which unusual variants of your genes suggest that certain drugs are likely to be more effective in treating your cancer than other drugs. So this has become completely routine in cancer care and in a few other domains.

So now I'm going to switch to a more technical set of material. So if you want to characterize disease subtypes using gene expression arrays, microarrays, here's one way to do it. And this is a famous paper by Alizadeh. It was essentially the first of this class of papers back in 2001, I think. Yeah, 2001. And since then, there have been probably tens or hundreds of thousands of other papers published doing similar kinds of analyses on other data sets.

So what they did is they said, OK, we're going to extract the coding RNA. We're going to create complementary DNA from it. We're going to use a technique to amplify that, because we're starting with teeny-tiny quantities. And then we're going to take a microarray, which is either a glass slide with tens or hundreds of thousands of spotted bits of DNA on it or it's a silicon chip with wells that, again, have tens or hundreds of thousands of bits of DNA in it.

Now, where does that DNA come from? Initially, it was just a random collection of pieces of genes from the genome. Since then, they've gotten somewhat more sophisticated.

But the idea is that I'm going to take the amplified cDNA, I'm going to mark with one of these jellyfish proteins that glows under light, and then I'm going to flow it over this slide or over this set of wells. And the complementary parts of the complementary DNA will stick to the samples of DNA that are in this well. OK-- stands to reason.

An alternative is that you take normal tissue as well as, say, the cancerous tissue, you mark the normal tissue with green fluorescent jellyfish stuff and you mark the cancer with red, and then you flow both of them in equal amounts over the array. That lets you measure a ratio. And you don't have as much of a calibration problem about trying to figure out the exact value.

And then you cluster these samples by nearness in the expression space. And you cluster the genes by expression similarity across samples. So it used to be called bi-clustering. And I'll talk in a few minutes about a particular technique for doing this.

So this is a typical microarray experiment. The RNA is turned into its complementary DNA, flowed over the microarray chip. And you get out a bunch of spots that are to various degrees of green and red. And then you calculate their ratio. And then you do this bi-clustering. And what you get is a hierarchical clustering of genes and a hierarchical clustering, in their case, of breast cancer biopsy specimens that express these genes in different ways.

So this was pretty revolutionary, because the answers made sense. So when they did this on 19 cell lines in 65 breast tumor samples and a whole bunch of genes, they came up with a clustering that said, hmm, it looks like there are some samples that have this endothelial cell cluster. So it's a particular kind of problem. And you could correlate it with pathology from the tumor slides and different subclasses. And then this is a very typical kind of heat map that you see in this type of study.

Another study from 65 breast carcinoma samples, using the gene list that they curated before, looks like it clusters the expression levels into these five clusters. It's a little hard to look at. I mean, when I stare at these, it's not obvious to me why the mathematics came up with exactly those clusters rather than some others. But you can see that there is some sense to it. So here you see a lot of greens at this end of it and not very much at this end, and vise versa. So there is some difference between these clusters. Yeah?

AUDIENCE: How did they come up with the gene list? And does anyone ever do this kind of cluster analysis without coming up with a gene list first?

PETER SZOLOVITS: Yes. So I'm going to talk in a minute about modern gene-wide association studies, where basically you say, I'm going to look at every gene known to man. So they still have a list, but the list is 20,000 or 25,000. It's whatever we know about. And that's another way of doing it.

So what was compelling about this work, this group's work, is a later analysis showed that

these five subtypes actually had different survival rates, and at p-equal 0.01 level of statistical significance. You've seen these survival curves, of course, before from David's lecture. But this is pretty impressive that doing something that had nothing to do with the clinical condition of the patient-- this is purely based on their gene expression levels-- you were able to find clusters that actually behave differently, clinically. So some of them do better than others.

So this paper and this approach to work set off a huge set of additional work. This was, again, back in the Alizadeh paper. They did a similar relationship between 96 samples of normal and malignant lymphocytes. And they get a similar result, where the clusters that they identify here correspond to sort of well-understood existing types of lymphoma.

So this, again, gives you some confidence that what you're extracting from these genetic correlations is meaningful in the terms that people who deal with lymphomas think about, about the topic. But of course, it can give you much more detail. Because people's intuitions may not be as effective as these large-scale data analyses.

So to get to your question about generalizing this, I mean, here's one way that I look at this. If I list all the genes and I list all the phenotypes-- now, we're a little more sure of what the genes are than of what the phenotypes are. So that's an interesting problem. Then I can do a bunch of analyses.

So what is a phenotype? Well, it can be a diagnosed disease, like breast cancer. Or it can be the type of lymphoma from the two examples I've just shown you.

It can also be a qualitative or a quantitative trait. It could be your weight. It could be your eye color. It could be almost anything that is clinically known about you.

And it could even be behavior. It could be things like, what is your daily output of Twitter posts? That's a perfectly reasonable trait. I don't know if it's genetically predictable. But you'll see some surprising things that are.

So then, how do you analyze this? Well, if you start by looking at a particular phenotype and say, what genes are associated with this, then you're doing what's called a GWAS, or a Gene-Wide Association Study. So you look for genetic differences that correspond to specific phenotypic differences. And usually, you're looking at things like single nucleotide polymorphisms. So this is places where your genome differs from the reference genome, the most common genome in the human population, at one particular locus. So you have a C

instead of a G or something one place in your genes.

Copy number variations, there are stretches of DNA that have repeats in them. And the number of repeats is variable. So one of the most famous ones of these is the one associated with Huntington's disease. It turns out that if you have up to 20-something repeats of a certain section of DNA, you're perfectly healthy. But if you're above 30 something, then you're going to die of Huntington's disease.

And again, we don't quite understand these mechanisms. But these are empirically known. So copy number variations are important, gene expression levels, which I've talked about a minute ago. But the trick here in a GWAS is to look at a very wide set of genes rather than just a limited set of samples that you know you're interested in.

Now, the other approach is the opposite, which is to say, let's look at a particular gene and figure out what's it correlated with. And so that's called a PheWAS, a Phenome-Wide Association Study. And now what you do is you list all the different phenotypes. And you say, well, we can do the same kind of analysis to say which of them are disproportionately present in people who have that genetic variant.

So here's what a typical GWAS looks like. This is called a Manhattan plot, which I think is pretty funny. But it does kind of look like the skyline of Manhattan.

So this is all of your genes laid out in sequence along your chromosomes. And you take a particular phenotype and you say, what is the difference in the ratio of expression levels between people who have this disease and people who don't have this disease? And something like this gene, whatever it is, clearly there is an enormous difference in its expression level. And so you would be surprised if this gene didn't have something to do with the disease.

And similarly, you can calculate different significance levels. You have to do something like a Bonferroni correction, because you are testing so many hypotheses simultaneously. And so typically, the top of these lines is the Bonferroni-corrected threshold. And then you say, OK, this guy, this guy, this guy, this guy, and this guy come above that threshold. So these are good candidate genes to think that may be associated with this disease.

Now, can you go out and start treating people based on that? Well, it's probably not a good idea. Because there are many reasons why this analysis might have failed. All the lessons that

you've heard about confounders come in very strongly here.

And so typically, what biologists do is they do this kind of analysis. They then create a strain of knock-out mice who have some analog of whatever disease it is that you're studying. And they see whether, in fact, knocking out a certain gene, like this guy, cures or creates the disease that you're interested in in this mouse model. And then you have a more mechanistic explanation for what the relationship might be.

So basically, you're looking at the ratio of the odds of having the disease if you have a SNP, or if you have a genetic variant, to having the disease if you don't have the genetic variant. Yeah?

AUDIENCE: I'm just curious on the class size. It seems like the Bonferroni correction is being very limiting here, potentially conservative. And I'm curious if there are specific computational techniques adapted to this scenario that allow you to sort of mine a bit more effectively than those.

PETER SZOLOVITS: Yeah. So if you talk to the statisticians, who are more expert at this than the computer scientists typically, they will tell you that Bonferroni is a very conservative kind of correction. And if you can impose some sort of structure on the set of genes that you're testing, then you can cheat. And you can say, well, you know, these 75 genes actually are all part of the same mechanism. And we're really testing the mechanism and not the individual gene. And therefore, instead of making a Bonferroni correction for 75 of these guys, we only have to do it for one.

And so you can reduce the Bonferroni correction that way. But people get nervous when you do that. Because your incentive as a researcher is to show statistically significant results. But that whole question of p-values keeps coming under discussion.

So the head of the American Statistical Association, about 15 years ago-- he's the Stanford professor. And he published what became a very notorious article saying, you know, we got it all wrong. Statistical significance is not significance in the standard English sense of the word. And so he called for various other ways and was more sympathetic to Bayesian kinds of reasoning and things like that. So there may be some gradual movement to that. But this is a huge can of worms to which we don't have a very good mechanistic answer.

All right. So if you do these GWASs-- and this is the real problem with them is that most of what you see is down here. So you have things with common variants. But they have very

small effect sizes when you look at what their effect is on a particular disease.

And so that same Zach Kohane that I mentioned earlier has always been challenging people doing this kind of work, saying, look-- for example, we did a GWAS with Kat Liao, who was a guest interviewee here when I was lecturing earlier in the semester. She's a rheumatologist. And we did a gene-wide association study. We found a bunch of genes that had odds ratios of like 1.1 to 1, 1.2 to 1.

And they're statistically significant. Because if you collect enough data, everything is statistically significant. But are they significant in the other sense of significance?

Well, so Zach's argument was that if you look at something like the odds ratio of lung cancer for people who do and don't smoke, the odds ratios is eight. So when you compare 1.1 to eight, you should be ashamed. You're not doing very well in terms of elucidating what the effects really are. And so Zack actually has argued very strongly that rather than focusing all our attention on these genetic factors that have very weak relationships, we should instead focus more on clinical things that often have stronger predictive relationships. And some combination, of course, is best.

Now, it is true that we know a whole bunch of highly penetrant Mendelian mutations. So these are ones where, one change in your genome, and all of a sudden you have some terrible disease. And I think when the Genome Project started in the 1990s, there was an expectation that we would find a whole bunch more things like that from knowing the genome. And that expectation was dashed.

Because what we discovered is that our predecessors were actually pretty good at recognizing those kinds of diseases, from Mendel on, with the wrinkled peas. If you see a family in which there's a segregation pattern where you can see who has the disease and who doesn't and what their relationships are, you can get a pretty good idea of what genes or what genetic variants are associated with that disease. And it turns out we had found almost all of them. And so there weren't a whole lot more that are highly penetrant Mendelian mutations.

And so what we had is mostly these common variants with small effects. What's really interesting and worth working on is these rare variants with small effects. So the mystery kid, like the kid whose case I showed you, probably has some interesting genetics that is quite uncommon, and obviously, for a long time, had a small effect. But then all of a sudden, something happened.

And there is this whole field called unknown disease diagnosis that says, what do you do when some weirdo walks in off the street and you have no idea what's going on? And there are now companies-- so I was a judge in a challenge about four or five years ago, where we took eight kids like this and we genotyped them, and we genotyped their parents and their grandparents and their siblings. And we took all their clinical data. This was with the consent of their parents, of course. And we made it available as a contest.

And we had 20-something participants from around the world who tried to figure out something useful to say about these kids. And you go through a pipeline. And we did this in two rounds. The first round, the pipelines all looked very different.

And the second round, a couple of years later, the pipelines had pretty much converged. And I see now that there is a company that did well in one of these challenges that now sells this as a service, like I showed you before, different company. And so you send them the genetic makeup of some kid with a weird condition and the genetic makeup of their family, and it tries to guess which genes might be involved in causing the problem that this child has.

That's not the answer, of course. Because that's just a sort of suspicion of a problem. And then you have to go out and do real biological work to try to reproduce that scenario and see what the effects really are. But at least in a couple of cases out of those eight, those hints have, in fact, led to a much better understanding of what caused the problems in these children.

That was fun, by the way. I got my name as an author on one of these things that looks like a high-energy physics experiment. The first two pages of the paper is just the list of authors. So it's kind of interesting.

Now, here's a more recent study, which is a gene-wide association of type 2 diabetes. It's not quite gene-wide, because they didn't study every locus. But they studied a hundred loci that have been associated with type 2 diabetes in previous studies. So of course, if you're not the first person doing this kind of work, you can rely on the literature, where other people have already come up with some interesting ideas.

So they wound up selecting 94 type 2 diabetes-associated variants. So these are the glycemic traits, fasting insulin, fasting glucose, et cetera; things about your body, your body mass index, height, weight, circumference, et cetera; lipid levels of various sorts, associations with different

diseases, coronary artery disease, renal function, et cetera. And let me come back to this.

So what they did is they said, OK, here's the way we're going to model this. We have an association matrix that has 47 traits by 94 genetic factors. So we make a matrix out of that.

And then they did something funny. So they doubled the traits. The technology for matrix factorization is called non-negative matrix factorization. And since many of those associations were negative, what they did is, for each trait that had both positive and negative values, they duplicated the column. They created one column that had positive associations and one column that had the negation of the negative associations with zeros everywhere else. So that's how they dealt with that problem.

And then they said, OK, we're going to apply matrix factorization to factor X into two matrices, W and H. And I drew those here on the board. So you have one matrix that-- well, this is your original 47 by 94 matrix. And the question is, can you find two smaller matrices that are 47 by K and K by 94, that when you multiply these together, you get back some close approximation to that matrix.

Now, if you've been looking at the literature, there are all kinds of ideas like auto-encoders. And these are all basically the same underlying idea. It's an unsupervised method that says, can we find interesting patterns in the data by doing some kind of dimension reduction? And this is one of those methods for doing dimension reduction.

So what's nice about this one is that when they get their W and H, they predict X from that. And then they know, of course, what the error is. And they say, well, minimizing that error is our objective. So that also lets them get at the question of, what's the right K?

And that's an important problem. Because normally clustering methods like hierarchical clustering, you have to specify what the number of clusters is that you're looking for. And that's hard to do a priori, whereas this technique can suggest at least which one fits the data best. And so the loss function is some regularized L2 distance between the reconstruction, W times H and X, and some penalty terms based on the size of W and H coupled by these relevance weights that-- you can look at the paper, which I think I referred to in here and I asked you to read. And then they do give sampling and a whole bunch of computational tricks to speed up the process.

So they got about 17,000 people from four different studies. They're all of European ancestry.

So there's the usual generalization problem of, how do you apply this to people from other parts of the world? And they did individual-level analysis of all the individuals with type 2 diabetes from these.

And the results were that they found five subtypes-- again, five-- which were present on 82.3% of iterations. By the way, total random aside, there's a wonderful video at Caltech of the woman who just made the picture of the black hole shadow. And she makes arguments very much like this. We tried a whole bunch of different ways of coming up with this picture. And what we decided was true is whatever showed up in almost all of the different methods of reconstructing it. So this is kind of a similar argument.

And their interpretations, medically, are that one of them is involved with variations in the beta cells. So these are the cells in your pancreas that make insulin. One of them is in variations in proinsulin, which is a predecessor of insulin that is under different controls. And then three others have to do with obesity, bad things about your lipid metabolism, and then your liver function.

And if you look at their results, the top spider diagrams, so the way to interpret these is that the middle circle, octagon, the one in the very middle, is the one with negative data. The one in between that and the outside is with zero correlation. And the outside one is with positive correlation.

And what you see is that different factors have different influences in these different clusters. So these are the factors that are most informative in figuring out which cluster somebody belongs to. And they indeed look considerably different. I'm not going to have you read this. But it'll be in the slides.

Now, one thing that's interesting-- and again, this won't be on the final exam. But look at these numbers. They're all tiny.

Some of them are hugely statistically significant. So DI, whatever that is, contributes 0.05 units to having beta-cell type of this disease at a p-value of 6.6 times 10 to the minus 37th. So it's definitely there. It's definitely an effect. But it's not a very big effect. And what strikes me every time I look at studies like this is just how small those effects are, whether you're predicting some output like the level of insulin in the patient, or whether you're predicting something like a category membership, as in this table.

So as I said, PheWAS is a reverse GWAS. And the first paper that introduced the terminology was by Josh Denny and colleagues at Vanderbilt in 2010. And so they did not quite a phenome-wide association. But they said, we're going to take 25,000 samples from the Vanderbilt biobank, and we're going to take the first 6,000 European Americans with samples, no other criteria for selection.

Why European Americans? Because all the GWAS data is about European Americans. So they wanted to be able to compare to that.

And then they said, let's pick not one SNP but five different SNPs that we're interested in. So they picked these, which are known to be associated with coronary artery disease and carotid artery stenosis, atrial fibrillation, multiple sclerosis and lupus, rheumatoid arthritis and Crohn's disease. So it's a nice grab-bag of interesting disease associations.

And then the hard work they did was they went through the tens of thousands of different billing codes that were available. And they, by hand, clustered them into 744 case groups and said, OK, these are the phenotypes that we're interested in. And that data set, by the way, is still available. And it's been used by a lot of other people, because nobody wants to repeat that analysis.

So now what you see is something very similar to what you saw in GWAS, except here, what we have is the ICD-9 code group. I guess by the time this got published, it was up to 1,000. And these are the same kinds of odds ratios for the genetic expression of those markers.

And what you find, again, is that this is the p-equal 0.05. That's the Bonferroni-corrected version. And only multiple sclerosis comes up for this particular SNP, which was one of the ones that they expected to come up. But they were interested to see what else lights up when you do this sort of analysis.

And what they discovered is that malignant neoplasm of the rectum, benign digestive tract neoplasms-- so there's something going on about cancer that is somehow related to this single-nucleotide polymorphism, not at a statistically high enough level, but it's still kind of intriguing that there may be some relationship there. Yeah?

AUDIENCE: So is this data at all public? Or is this at one particular hospital? Or who has this data? Would it be combined?

PETER Yeah. I don't believe that you can get their data unless-- I think, if-- I mean, they're pretty good

SZOLOVITS: about collaborating with people. So if you're willing to become a volunteer employee at Vanderbilt, they could probably take you. But I just made that up. But every hospital has very strong controls.

Now, what is available is the NCBI has GEO, the Gene Expression Omnibus, which has enormous amounts-- like, I think, hundreds of billions of sample data. But you don't often know exactly what the sample is from. So it comes with an accession number and an English description of what kind of data it is.

And there are actually lots of papers where people have done natural language processing on those English descriptions in order to try to figure out what kind of data this is. And then they can make use of it. So you can be clever. And there's a ton of data out there, but it's not well-curated data.

Now, what's interesting is you don't always get what you expect. So for example, that SNP was selected because it's thought to be associated with multiple sclerosis and lupus. But in reality, the association with lupus is not significant. Its p-value of 0.5, which is not very impressive. The association with multiple sclerosis is significant.

And so they found, in this particular study, a couple of things that had been expected but didn't work out. So for example, this SNP, which was associated with coronary artery disease and thought to be associated with this carotid plaque deposition in your carotid artery, just isn't. p-value of 0.82 is not impressive at all.

OK, onward. So that was done for SNPs. Now, a very popular idea today is to look at expression levels, partly because of those prices I showed you where you can very cheaply get expression levels from lots of samples. And so there's this whole notion of Expression Quantitative Trait Loci, or EQTL, that says, hey, instead of working as hard as the Vanderbilt guys did to figure out these hundreds of categories of disease, let's just take your gene expression levels and use those as defining the trait that we're interested in.

So now we're looking at the relationship between your genome and the expression levels. And so you might say, well, that ought to be easy. Because if the gene is there, it's going to get expressed. But of course, that's not telling you whether the gene is being activated or repressed or enhanced, or whether any of these other complications that I talked about earlier are present.

And so this is an interesting empirical question. And so people say, well, maybe a small genetic variation will cause different expression levels of some RNA. And we can measure these, and then use those to do this kind of analysis.

So differential expression in different populations-- there is evidence that, for example, if you take 16 people of African descent, then 17% of the genes in a small sample of 16 people differ in their expression level among those individuals; and similarly, 26% in this Asian population and 17% to 29% in a HapMap sample. Of course, some of these differences may be because of confounders like environment, different tissues, limited correlation of these expression levels to disease phenotypes. Nevertheless, this type of analysis has uncovered relationships between these EQTLs and asthma and Crohn's disease. So I'll let you read the conclusion of one of these studies.

So this is saying what I said before, that we probably know all the Mendelian diseases. So the diseases that we're interested in understanding better today are the ones that are not Mendelian, but they're some complicated combination of effects from different genes. And that makes it, of course, a much harder problem.

There is an interesting recent paper-- well, not that recent-- 2005-- that uses Bayesian network technology to try to get at this. And so they say, well, if you have some quantitative trait locus and you treat the RNA expression level as this expression quantitative trait locus, and then you take C as some complex trait, which might be a disease or it might be a proclivity for something, or it might be one of Josh Denny's categories or whatever, then there are a number of different Bayesian network-style models that you can build.

So you can say, ah, the genetic variant causes a difference in gene expression, which in turn causes the disease. Or you could say, hmm, the genetic trait causes the disease, which in turn causes the observable difference in gene expression. Or you can say that the genetic variant causes both the expression level and the disease, but they're not necessarily coupled. So they may be conditionally independent given the genetic variant.

Or you can have more complex issues, like you could have the gene causing changes in expression level of a whole bunch of different RNA, which combined cause some disease. Or you can have different genetic changes all impacting the expression of some RNA, which causes the disease. Or-- just wait for it. Oops.

You can have models like this that say, we have some environmental contributions and a

bunch of different genes which affect the expression of a bunch of different EQTLs, which cause a bunch of clinical traits, which cause changes in a bunch of reactive RNA, which cause comorbidities. So the approach that they take is to say, well, we can generate a large set of hypotheses like this, and then just calculate the likelihood of the data given each of these hypotheses. And whichever one assigns the greatest likelihood to the data is most likely to be the one that's close to correct.

So let me just blast through the rest of this quickly. Scaling up genome-phenome association studies-- the UK Biobank is sort of like this All of Us project. But they do make their data available. All of Us will, also, but it hasn't been collected yet.

UK Biobank has about half a million de-identified individuals with full exome sequencing, although they only have about 10% of what they want now. And many of them will have worn 24-hour activity monitors so that we have behavioral data. Some of them have had repeat measurements. They do online questionnaires. About a fifth of them will have imaging.

And it's linked to their electronic health record. So we know if they died or if they had cancer or various hospital episodes, et cetera. And there's a website here which publishes the latest analyses. And so you see, on April 18, genetic variants that protect against obesity and type 2 diabetes discovered, moderate with meat-eaters are at risk of bowel cancer, and research identifies genetic causes of poor sleep. So this is all over the place. But these are all the studies that are being done by this.

I'll skip this. But there's a group here at MGH and the Broad that is using this data to do, large-scale, many, many gene-wide association studies. And one of the things that I promised you, which is interesting, is from these studies, they say, well, the heritability of height is pretty good. It's about 0.46 with a p-value of 10 to the minus 109th. So your height is definitely determined, in large part, by your parents' height.

But what's interesting is that whether you get a college degree or not is determined by whether your parents got a college degree or not. This is probably not genetic. Or it's only partly genetic. But it clearly has confounders us from money and social status and various things like that. And then what I found amusing is that even TV-watching is partly heritable from your genetics. Fortunately, my parents watch a lot of TV.

The last thing I wanted to mention, but I'm not going to have time to get into it, is this notion of

gene set enrichment analysis. It's what I was saying before, that genes typically don't act by themselves. And so if you think back on high school biology, you probably learned about the Krebs cycle that powers cellular mechanisms. So if you break any part of that cycle, your cells don't get enough energy. And so it stands to reason that if you want to understand that sort of metabolism, you shouldn't be looking at an individual gene. But you should be looking at all of the genes that are involved in that process.

And so there have been many attempts to try to do this. The Broad Institute here has a set of, originally, 1,300 biologically-defined gene sets. So these were ones that interacted with each other in controlling some important mechanism in the body. They're now up to 18,000. For example, genes involved in oxidative phosphorylation and muscle tissue show reduced expression in diabetics, although the average decrease per gene is only 20%.

So they have these sets. And from those, there is a very nice technique that is able to pull-- it's essentially a way of strengthening the gene-wide associations by allowing you to associate them with these sets of genes. And the approach that they take is quite clever.

They say, if we take all the genes in a gene set and we order them by their correlation with whatever trait we're interested in, then the genes that are closer to the beginning of that are more likely to be involved. Because they're the ones that are most strongly associated. And so they have this random walk process that find sort of the maximum place where you can say anything before that is likely to be associated with the disease that you're interested in. And they've had a number of successes of showing enrichment in various diseases and various biological factors.

The last thing I want to say is a little bit disappointing. I was just really looking for the killer paper to talk about that uses some really sophisticated deep learning, machine learning. And as far as I can tell, it doesn't exist yet. So most of these methods are based on clustering techniques on clever ideas, like the one for gene set enrichment analysis. But they're not neural network types of techniques. They're not immensely sophisticated.

So what you see coming up is things like Bayesian networks and clustering and matrix factorization and so on, which sort of sound like 10-, 15-, 20-year-old technologies. And I haven't seen examples yet of the hot off the presses, we built a 83-layer neural network that outperforms these other methods. I suspect that that's coming. It just hasn't hit yet, as far as I know. If you know of such papers, by all means, let me know.

All right. Thank you.