ISAAC SAMUEL KOHANE: So let me give you a case history. This is work that I've done with one of my former interns, actually, who's a superb scientist at the Dana-Farber Cancer Institute, David Rowitch, who also is a part time neonatologist. I call this finding the needle in the haystack.

And it's about studying the cerebellum. So I impressed, I hope, upon you on the last lecture, just how bad microarrays are. How noisy they are, and how difficult it is to interpret them. So now, I'll show you how-- darn it-- how, if you actually do the computation right, you can actually extract useful biomedical information out of it.

So the cerebellum is a nice system to study because it has this nicely layered structure. The cerebellum is this part of the brain that sits in the posterior fossa, at the back. And has this very nicely layered structure which is detailed in this cross section here so that each layer has this distinct group of cells. And often, these group of cells are distinguished by different genes expressed at different times, specifically, for that cell type or cell layer.

And the cerebellum has a pivotal role in locomotion and posture, but the fact that I can stand approximately straight and move around is almost solely a function of the activity of the cerebellum. And because of its nicely layered structure, the cerebellum has been the subject of a lot of neuroscience because it's much less complex than the other central neural tissues, in that regard. So there was a particular system that David was interested in.

Sonic Hedgehog, it's one of these supposedly amusing names developed by fly biologists because fly biologists may frequently have a very bad sense of humor and they apply that to all the genes that they discover. Sonic Hedgehog is one such gene, turns out to be a very important signaling molecule that's very important in proliferation and differentiation of cells.

And David wants to know, what was the role of Sonic Hedgehog in this outermost of the layer of the cerebellum, the external granule layer A, EGLA. It was known that cyclin D1 was expressed in this layer. And different other genes were expressed in other layers. It was known that Sonic Hedgehog was important in a variety of different ways.

For instance, genes that had previously been implicated in a Sonic Hedgehog signaling pathway when a mutant cause in the mouse model here-- a posterior cerebellar tumor-- which is very much akin to the medulloblastomas that are a big cause of juvenile brain cancer, and a lethal one at that. And it's certainly been shown that medulloblastomas is are involved in the Sonic Hedgehog pathways, a large percentage of them. So the question was, could we find other culprits, other genes that are involved in the Sonic Hedgehog pathway?

And it's a big haystack that we're looking into because we're looking in two large haystacks, actually. One is a large number of probes on the microarray. Two, we're looking at an entire cerebellum. A complex tissue-- even though it's one of the simpler neural tissues-- when in fact, we're trying to extract a signal out of one part of this complex tissue, there's a very superficial layer of cells called the EGLA.

So can we find the signal in time and space? So we're not only looking for the right number of probes in the right location, but the right point in development. And this was quite a challenge for us because David is a personal friend and former intern. And he heard me pontificate about the importance of bioinformatics. And here was this closed friend who is now engaging in this type of collaboration.

And if I gave him the wrong candidates, that means I'd be wasting the time of his post-doc, of which he had only one at that time, and delaying him by six months if I gave him some wrong things to look at. So it was a relatively high stakes game. So the experiment was very simple.

To harvest, initially, just day five-- postnatal day five cerebellum, expose it to Sonic Hedgehog vehicle and growth arrest media. Forget about the growth arrest media for a second. But basically, exposing it to Sonic Hedgehog or vehicle, hybridize it to the now obsolete affymetrix mouse array. The mouse arrays are obsolete in this generation, it's no longer available.

And we did what we refer to in the grant as noise models in our bioinformatic analysis, which I'll detail shortly. And then we gave a list of nine significantly up-regulated genes and 10 genes not significantly up-regulated and these will be subsequently screened with insight through hybridization to see whether, in fact, these predictions were correct. And so what we did was the very simple thing.

If you recall, I showed you in the last lecture that I gave you how noise grew dramatically at lower levels of expression because you could flip the sign of the up or down regulation just by virtue of having low levels of expression affected by low levels of noise. So clearly, you need to have some of statistical measure of the separation of the two conditions.

So in this case, the signal was the expression of a gene in response to the signal, namely, Sonic Hedgehog. And the other is the vehicle treated ones. And the question is, each such gene, could we find those genes that were clearly separated across those two conditions? And there's a variety of ways that you could have done it.

And in fact, as you promise it, I'll have you explore some of the standard methods that, for instance, are things like a modified t-test, like the significance analysis of microarrays. But I will make the following point, which is, try to understand what you're doing by looking at ratios. For instance, just looking at a t-test, where you're looking at the difference in means, is fundamentally misunderstanding the computational challenge in a very simple way.

What average do we want to calculate? And I realize not all of you are computationally involved, so I'll try to keep it simple. There's two types of averages that we might consider among many calculating. One is the simple arithmetic average, where you take where you sum the signal and divide it by the number of samples. And you sum the control divide it by the number of samples and you take that ratio. Or you take the geometric mean of them.

And why would you do one or the other? Well, let me ask you the following thing. When I first thought about this, I was having a mortgage in my apartment, so it was incredibly important to me. If you have four different interest rates over 4 years, such that, the amount you have to pay out is 1 plus R1 times 1 plus R2, to 1 plus R3, 1 plus R4 and you're going to have to pay different amounts on that principle.

How am I going to estimate the average interest rate I'm paying per year? Is it the arithmetic average of these Rs? No because it's compounding. I know that's a simple point, but few people realize it's actually the geometric mean. And the same error is being made if you look at the arithmetic mean versus the geometric mean to look at ratios because, fundamentally, we're looking at a ratio.

And we did the analysis both ways. And I can tell you, we knew in this case, the answer is because we did the insight to hybridization. And first of all, the good news. The good news that we published in PNAS a couple of years ago was that, essentially, 80% of the time that we said a gene was expressed in the EGLA based on this simple geometric mean, sure enough, by insight of staining, it was.

And when we said it wasn't, it wasn't. But the point is, if we use that arithmetic mean, we had double the false positive rate. So just understanding something as simple as the difference in the arithmetic mean and geometric mean, which is a non-obvious notion to anybody who's just been used to doing t-tests, screws up the analysis. Let's make this more relevant to human disease.

How can we leverage this developmental view of the mouse? This we're talking about mice. How do we actually try to understand how this is reflected in its relevance to these tumors that are talking about, medulloblastoma? So we had, actually, some medulloblastoma. And I want to say about microarrays is that this community, along with the SNP community has been unique in making freely available the data to its community.

For those of us who have MDs-- or will have MDs-- let us be aware that still today, most clinical research papers do not publish the primary data from which the conclusions are made. Still do not. And therefore, you are unable to independently verify these results or come out with better results or with improved results. And that's why it was really incredibly important.

I'm sorry I didn't emphasize this before, that the early practitioners of microarrays, such as Todd Golub. And Pat Brown published all the data right away after publication. In fact, I looked a half year ago at Todd's original paper in 1999, on AML versus ALL. And I think there were 700 citations of just that one paper. And it would have not been the same had they had not published data because there was a lot of secondary analyzes and a lot of problems found with the paper and the size of the paper. It sets up a whole scientific process of motion that you don't have otherwise.

So we have this data on a variety of different human medulloblastomas that have been published a couple of years ago. And then, we could go to a database called Homology-- and by the way, your next problem set, not the one I'm describing to you, but your next problem set will be a treasure hunt where we'll make you hop across all the different databases that you should know about as basic-- not basic biologists, but basic consumers of the very simple kinds of information around about SNPs and the microarrays.

But we'll get to treasure hunt later. Suffice it to say that there is a database called humology, which allows you to map the genes in humans to those that are present in mice. And it turns out that there were present on those two microarrays, there on the order of about 3,000 genes that are homolog from mice to men. And so what we did then was to do a principal component analysis of these data sets.

Now hands up those of you who really understand principal component analysis? Two, two of you. So those who don't, let me try to give you the simple version of it. Principal component analysis, essentially, tries to refactor the data so that you're looking at these new variables that are linear sums of the original variables. In this case, the genes. These principal components are linear versions of these original measurements that capture, in successive components, the major amount of variation of the data set.

So the first principle components is this linear combination that gets the largest amount of the variation in the data set. The second principle components are orthogonal. Essentially, at right angles to that first principal component and get the second most amount of variants. And so on. And in fact, have as many principal components as you have original variables.

But the nice thing is that, heuristically speaking, by the time you've gone through about two or three or four principal components, you've typically captured on the order of 70% to 80% to 90% of all the variants in the data set. In other words, by doing these linear combinations of orthogonal vectors, you're able to capture most of the variance.

And the nice thing about that is for data sets, such as gene expression, where you have a lot of highly correlated behavior. In other words, there's large numbers of genes that are really not that independent of one another, this allows you to reduce the number of variables that you have to look at an analysis from a very large number-- tens of thousands-- to adjust a handful-- the first principle components that allow you to capture the majority of variation.

And that's, by the way, the intuition that some of these face recognition software that you may have heard about from Homeland Security has been inspired. By essentially, taking the eigenface, where they look at the principal components that make up the features of the face and reduce it-- reduce the complexity to these principal components of face so that not only can you generate an eigenface, this generic face, but you can also map people to which faces they're most like.

So it allows you a little bit of play and slot between different angles and different ways that you might look in a different day because it's still capturing the major variants of your face. So what we did is the following. Not only was there day five of the mouse, but we actually captured multiple time points on the order. I can't remember now, exactly. Maybe 15 time points.

And it's a publicly available data set, if you're interested. 15 time points, going from embryonic days to postnatal days. Day 60 I believe, of the mouse. And then when we did the principal components of the gene's expression pattern across this time series, we saw the following when we just plotted the genes by that position in this new coordinate system, the first and second principal components. And I've colored them by whether they fell into this section or that half section of the egg, so to speak.

And why did I color it so? Because if you actually look at when these genes had their maximal period, their maximal time of expression. The ones in this component, in this area of the two components, had an early peak and the genes in this component had a late peak. So the late-- sorry, the early principal component and the late set of principal components. In plane English, by just doing this simple decomposition-- the simple factoring into two principal components, we've taken that time series and re-plotted the genes.

And now, we have a set of genes that are clearly separating themselves out from another set of genes. And they're characterized by being early versus late peaking. Now, so what? We then took the human homologues of those genes and we asked in the medulloblastomas, which genes compared to non-medulloblastomas were upregulated and downregulated in the human tumors? Please?

AUDIENCE: Describe the x-axis and y-axis, what's the--

ISAAC SAMUEL KOHANE: This is the first and second principal component.

AUDIENCE: Yeah, but I mean, the number 0,1, 2, 3 and--

ISAAC SAMUEL KOHANE: It's, I can't-- I don't think it's zero based. I think zero is somewhere around here.

AUDIENCE: So it's just-- I mean, I know, so what does that mean?

ISAAC SAMUEL KOHANE: Its according to the coefficient. If you position that gene according to its position along those two principal components. So each one of these is a single gene. So then what we do is we plotted those up and down regulated genes by whether they fell into-- we just took those up and down regulated genes without knowing anything more. And we say, where do they fall into this late and early decomposition based on the principal component analysis?

And the simple answer was essentially, 90% of the genes that were upregulated fell into the early period of expression. And similarly, around 90% of the genes that were downregulated-- in humans, mind you-- fell into the later phase of development of the mouse. Go ahead.

AUDIENCE: I have a question the way you're presenting the data doesn't make apparent why you need the second component.

ISAAC SAMUEL KOHANE: That's a very insightful question. And the answer is that for what I'm describing it, does not because the separation that you're seeing is just in one dimension and you're absolutely right. It doesn't add anything. Turns out, there were actually different processes that I'm not describing here that actually were separated out by going into the second principal component.

But I think you're showing that you're understanding what the hell I'm talking about by saying, for this distinction early and late, all you need is the first principal component. You're absolutely right. So I just want to-- those of you who've taken basic pathology in medical school-- will hear the following. I don't know if I have a pointer to it or a slide, but it's been known by Cohnheim and others from the turn of the century have speculated that there's something about the embryonic program of development that is recapitulated by tumors. That tumors are similar-- not the same thing-- are similar to essentially a poorly controlled embryonic development because when you look under the microscope, a lot of the tissues in tumors are not as well differentiated as the tissue from which they came from.

Liver cancer, if the liver cancer looked exactly like liver tissue, it wouldn't be liver cancer, it would be liver. And so they're actually more primitive versions. And this actually adds support to this. In fact, just to check on this, we did the following experiment, which I don't show here. So what we've done here, just to place this in an appropriate perspective, is take neural tissue from humans-- human tumors-- and projected it against the components of mouse development.

And what we C is this nice separation between upregulated genes and the early phase. Now maybe all we're showing are general markers of hyper-proliferation and not something specific to the system. So what we did is we took a similar pair, namely, lung cancer and lung development. And lung cancer and lung development showed, essentially, the same relationship.

But then when you took medulloblastoma against the lung development background, the separation was no longer as good. And when you took the lung cancer against the neural development, it was also not as good. So what that tells us, it's something much more nuanced than simply embryogenesis recapitulates generically angiogenesis or the other way around I should say.

It says that there is part of the differentiation program of that particular tissue, which is recapitulated by the tumor. And that's already interesting because it gives us some insights and development, but it gets more interesting when you start thinking, well, I know a lot about different stages of development. Can I actually start pulling out my understanding of the mechanisms operating at different stages of the development to understand what's different about these different cancers?

So this is just showing you in the time series, what the principal components were showing you before. And in red and green, we're distinguishing the genes that are up and down regulated. And you see they have, in the human tumors, they have different time courses in the mouse data versus in the-- different courses of the mouse data, depending on whether they were upregulated or downregulated in the human tumors.

And as I said here, this is the slide I wanted to Lobstein and Cohnheim were among the first who theorized similarities among human embryogenesis and biology of cancer cells. And it's actually not until now that we've gotten really much more objective evidence that this is the case. And the brain tumor classification system that is used and that was devised by Daly and Cushing in 1926-- from which our modern taxonomy is derived-- emphasized this. But they are very crude taxonomies and descriptive taxonomies.

And here we're providing perhaps-- and I'm saying perhaps-- a much finer classification. Is that true? Well, let's turn around the principal components. Let's look at each tumor as a function of the genes that are upregulated. Sorry, let's look at the position of the tumors, now-- not the genes-- by the position in the principal component analysis. And to make a long story short, we were able to dissect out different days of development. So these are different mouse samples, and projecting onto different human cancers onto the same principal component analysis of these mice.

And this is very interesting in a number of different ways. One is, we're separating out different days of development into its principal component space. But also, separating in that same space, the tumors. Now that's fundamentally interesting because it's telling us and the reason why I'm I bring this up in the genomic medicine class because if one of the fundamentals of modern medicine is taxonomization of disease. And here, I'm providing a very, very quantitative measure of differentiation of these different tumors based on their position in the space of development from a data set that you can never get from humans, but from mouse.

And this is truly exciting because I don't have time, well actually, I don't have a slide to show you because we just submitted this-- sorry, it just got accepted to the genes and development, which is an important journal in this space. We can actually separate the metastatic from the non-metastatic medulloblastoma in the same way. They separate in the same individual space. So what does this tell us more generically?

Principal component analysis is not a commonly used tool for clinicians, obviously. And yet, here we have a tool that is creating a much better taxonomy of disease with predictive power previously not available to us. And allows us to understand-- because we're saying what are the genes? Because remember I said the principal components are a linear component or linear combinations of genes? We can say why is this gene in this position? Which genes have the major weights which make it responsible for that tumor being in that position.

Might give us some insights into the underlying tumor mechanism. And so here, we have a fine grain [INAUDIBLE] which was not previously available to us. And I think it's, therefore, not a leap to think that-- in similar kinds of processes-- that computational approach to these data sets will allow us to recast our very, very squishy current taxonomies. I think I told you in the first lecture, with a few exceptions, the most notable exceptions being microbiology, most of our disease classifications are not mechanism-based.

In microbiology, this classification is mechanism-based. It's the organism that's infecting you that's the name of the disease. But most of the diseases like lupus, probably a mixed bag of inflammation inflammatory diseases that we don't know, polycystic ovary diseases, we're describing symptoms, not pathophysiology. And I think this is allowing us to get much, much closer and therefore, a much more robust and fine-grained understanding of that.

So my interim summary here is that this computing allows us to identify some pathways and provides more importantly, a natural classification of disease with further insight. Apparently, I don't know how to open my book. Go ahead.

AUDIENCE: Did you try to redefine bacterias that were [INAUDIBLE] against the optimized use of combination of genes that you expect to be a developmental program, or something to that level?

ISAAC SAMUEL KOHANE: So Jose is asking a very interesting question which the short answer is no, but I think it's a very good question, which is-- but I wouldn't know exactly where to start. But surely there is a thoughtful way to respond to it. Could you take linear combinations of genes-- and I wouldn't know exactly which linear combination of genes-- but maybe just using those as a starting point, perhaps, genes that have been implicated at various stages of development and using those.

And I think that's a good way to go. My understanding of developmental Biology is unfortunately, it's very sparse knowledge at different stages. So it'd be hard to know which genes are generally informative. Nonetheless, I'm a big fan of the approach that you're suggesting, which is a knowledge based approach. What we're doing here was clearly knowledge free. We were just taking the data and just saying, what explains the maximum variance and going from there.

But if we could conceive that process or the set of genes that we knew were implicated, I bet we'd have much better resolution than we have currently. So I think it's a very good thought, but we have not done that. Which brings me now to another aspect of computing. I've shown you guys this slide. Does it look familiar? Good. So here's the fantasy that most deans of medical schools and heads of hospitals have. And, again, this is not a very technical point. It's more of a sociological point, but it will end up being technical because it can tell you about a whole other area of biocomputing that I think is ripe for the picking, but it's also a rate limiting step.

So they all have the fantasy. Now that they've seen things like, oh, yes I can predict different outcomes of different tumors, let's say, I want to be able to develop a set of targets that a drug company is going to pay me a lot for. So the fantasy is the nice doctor talks to the nice patient. He gets consent in the nice hospital. They have a family history.

They bank the tissue obtained. They do some genomic analyzes. They do a clinical annotation. They do some fancy mumbo jumbo bio-informatics. And lo and behold, they're going to have the target that a drug company is going to pay a lot of money for. That indeed is-- for those of you who are more entrepreneurial oriented, that's the fantasy that's launched. Many, many Venture Capital funded ships, like the decode project and others. And the problem is the following.

It's that this central piece, the phenotypic annotation, ends up being the hardest part. Understanding what really happened to the patient. Did the patient really have a tumor? How old were when they had the tumor? What drugs were they responsive to?

That is the hard part. And any of you have ever looked at a clinical chart can understand why that is. There's just very little useful machine-able data in the clinical chart. Because in fact, the real fantasy is the following. And what I call [? Ratwitchz ?] at the Medical College of Wisconsin-- what they do is they take rats, consomic rats, where they have systematically substituted one chromosome for another.

So which would give them a very big efficiencies of identifying linkage with a different trait. What they do is they take these different chromosomic rats-- different strains with well-understood genotypes-- and they exposed them to very well characterized environmental exposures, such as hypoxia. They put them in the oxygen equivalent of top of the mountain for several weeks.

Or they give them a high salt diet or high fat diet. Or they volume deplete them. And after that wonderful experience, they then physically reconstruct them and rip out their heart and put their hearts on these prep machines where they can look at the contractility curves of the heart in the perfused prep. And that's great because now you have a well-defined phenotypic environmental characterization. You know the genotype and you have the expression.

And by the way, let me say and point you to the following site, PGA.mcu.edu. They have all that data online. The phenotypes, the genotypes and expression. And they do it for cardiac phenotypes, kidney phenotypes, lots of different phenotypes and different parts of medicine. And so this is the fantasy. But of course, patients would probably objectively subjected them to this kind of treatment.

So this high throughput phenotype fantasy just is unlikely to happen anytime soon. So we took a very different approach when we were dealt with this problem. And that was how to get sufficient numbers of samples in the right amount. And you should know that there are companies like Ardeas, which have been given millions of dollars of venture capital unsuccessfully to try to solve this problem. To get enough tissues in the right amount.

So the National Cancer Institute put out a request for applications saying, listen, there's thousands of tissues available throughout the pathology repositories of our country that we want to be able to do genomic studies. How to go about that? So inspired by Napster and Gnutella-- and for those of you who don't know what the-- those of you who are not aficionados of file sharing, Napster is a file sharing service where there actually is a centralized. But the data themselves are decentralized in multiple directories. Guntella, in order to avoid getting anybody sued, it's fully decentralized. The Directories are decentralized and the data is decentralized.

And what we built, the shared Pathology Informatics network, SPIN, that's funded by about $7 million from the NCI, is exactly the same thing. And what we did is to take advantage of the twin obsession compulsions of pathologists. Pathologists are probably the best first taxonomists of the medical profession.

They actually categorize organs and specimens better than anybody else. Second, they're obsessive compulsive collectors of tissue. And they keep tissue around for years and years. And so we thought that if we could take advantage of those twin compulsions, we could actually be successful. So at a bird's eye view what we did was the following, which was to create a network whereby a user, with a web browser, could send a query to his Query Composer, which would then send out a query to this nebulous thing called SPIN network. Which would, using this peer to peer technology, respond which samples were out there.

More specifically, each tissue bank or institution would have its own node on the network, on the internet. And we provide them these open source three tools which allow them to extract from these pathology databases because there's a textual report with each pathology sample. Both the anonymized-- because we have this anonymization program-- the anonymized textual report, and code a few data elements to enable search. So that when you put in the query, the query percolates in a fully distributed way.

There's no center to this network. This query percolates throughout the system. All the nodes respond eventually, if they can, and you get back the sum of all specimens involved. And those of you even in the medical profession probably don't appreciate the following point. The reason we created a peer to peer system is that pathology tissue banks are jealously guarded by the pathologists and by the surgeons who built them. They just don't want to share data.

And it's essentially impossible today to know-- even across the street and the Brigham-- what samples they have. And by allowing them to actually fully control their own node and control what they expose to the outside, we're able to overcome the sociological obstacles for data sharing so that we can actually do this. So for instance, we've done this now, this is actually an out-of-date slide, not only across Harvard and UCLA, but now we've gone live with Pittsburgh and a few other sites.

So that today, if you're a registered user, you can actually send out a query to all those institutions that you see on the right. And the query then percolates across the country. And this query takes on the order of 10 seconds on a bad day to execute.

So for instance, we want to know how many distinct specimens can we have of renal transplant patients? And that's typically hard to come by. And here is a response. So this is out of seven nodes, this is a query I did back in October of last year. Out of the seven nodes that we had up at that time, with only two nodes responding, we already had 20,000 specimens identified with an age distribution shown as here.

By the way, I'm aghast at the fact that apparently 34 specimens of patients ages 90 to 99. I wish it was true that was an error, but I looked into it and it's not. So some weirdos are actually transplanting things in very old patients. Would not happening in England. In any case, here's the age distribution. And if you click on the full text, and you have the appropriate privileges, you can actually see the full anonymized text that show a report of that patient.

And so I make this point for a number of reasons. First, we're successfully mining the obsessive compulsive feature of pathologists. Second, I hinted to you that some of these outcomes analysis using functional genomic measures, such as the lymphoma study, are essentially not well reproduced in different studies. Whatever the reason, they are many.

A lot of the ills in the analysis are excused by large numbers. So if we went from having 100 patients to having 20,000 patients, I assure you that the differences in different machine learning techniques would matter very little. And so this is how we're beginning to approach, how do we harness the phenotypes that are out there that are available, that are not available otherwise through a completely different kind of bioinformatics? This is more infrastructure bioinformatics.

But it's the rate-limiting step. We can work on a dozen samples until we're blue in the face. We really won't know how much we're overfitting our data until we have large samples.

AUDIENCE: But the very first thing that you can request is that for [INAUDIBLE].

ISAAC SAMUEL KOHANE: Yeah, I wish it was that simple. So that shows you're awake. The next question is, how do I get the samples? So I wish it was true that once you identified a sample, you just click the box, put your credit card and you get FedExed the tissue sample. So what happens at that point is that you start a dialogue with investigator and the Institutional Review Board of that tissue bank.

So that's the bad news. The good news is, we've solved one part of the problem is figuring out where the specimens are to find out if your study is feasible. And people are already being used as tools right now to write grants because they can say, I know that there's [INAUDIBLE] specimens out and they don't necessarily get access to the tissue for that. But they get in-principle agreement from the collaborators that if they get funded, they'll do this as a collaboration.

But you're absolutely right to do it. So this is just another example how we're using computation to overcome the obstacles to doing genomic medicine. Let me give you one last case history and then we'll go to the problem set. So let's talk about how we can use computing again to have a different understanding, and how we use computing allows to have a different understanding of for instance, gene regulation.

So this is time on x-axis. And this is some arbitrary value of expression. Are these two time courses the same? Are those two genes developed by the same process? Yes or no? You could make an argument either way, right? Maybe they're the same.

How about this one? Maybe they're the same, maybe not. I tell you, the stuff was pretty noisy. Maybe that's just noise, a priori, who knows. Now does order matter? It does in our analysis. Does time matter? Yes, good. Because it should be true that the likelihood of me being here now is informed much more by where I was five minutes ago and where I was a day ago, right?

That's a fundamental Markovian property of Zack, and most physical processes. Now if there's a correlation coefficient at all, address that Markovian process. In other words, if you take two stocks, and shuffle the timing. Let's say two stocks, Apple and Microsoft, let's say for the sake of argument, that they are highly correlated. They the high correlation coefficient.

If you permute the days that they maintain the same pair of stock values for that day together, to shuffle the days, is the coefficient the same or different? Same. So correlation coefficient does not in fact capture the effects of time. Now, as I hope you're aware from our previous discussions, a lot of the clustering that's been done even on time series data, is done on the basis of correlation coefficients.

People in those dendrograms bring together genes that share the same expression pattern as measured by the correlation coefficient. So what consequences does that have? Let's go back to a very old study. This is [INAUDIBLE] study back in 1999. A classical Jewish study where they snipped off the-- no, they took a fibroblast from foreskins and they measured the gene expression pattern over time of the transcriptome of these foreskins after being exposed to serum.

So you see here, every column is zero hour, 50 minutes and so on till 24 hours of each gene. And then they did this act of creativity, which is they draw lines next to the dendrogram and then cut out the blocks each line and say, oh look, there is a bunch of these genes and they seem to -- for instance, coagulation hemostasis. There's a tissue factor, pathway inhibitor and so on.

These are coagulation involved. These are cell cycle proliferations. This is inflammation. Oh look, these are angiogenesis involved genes. These are cytoskeletal. And they all seem to be clustered together. And it's again, I use this metaphor all too often, but it's so true. This is very much like the dog at the opera. It's a miracle it can sing at all and you don't criticize how well it's singing.

Because what they've done is this act of creativity. They've actually looked at the data of genes and they draw these lines and say, these belong a chunk. Now there are other lines they could draw on. And in fact, if you look closely at their original paper, there's large numbers of genes that they do not draw those lines next to. They don't have anything to say. So what's really going on here?

So again, these clusters are defined by correlation coefficient. So what if we try to take advantage of the Markovian property of gene expression, like any physical process. Namely, that in a time series. The recent past informs you more than the distant past.

So you can approximate the conditional probability of this value upon the recent past rather than the whole time series. That is, basically the assumption is that T0 is independent of the remote past, given the recent past. And how many steps back and look is what you well know as the mark of order. So mark of order is two, that means you look back two time slices. Markov order is one, you look back one type slices.

If time ordering does not matter-- just like the correlation coefficient-- then your mark order is zero. The recent past actually doesn't inform you at all. Now let's look instead at clustering as this fiscal model selection problem. So rather than looking for highly correlated pairs, we'll consider each of the time series of his genes as being examples of a process. And the process is going to be represented therefore, by one or more genes, which are driving that process.

And the question is, for any given pair of genes, do they belong to the same process? And essentially, with a very simple Markovian analysis, which is just a pain in the butt to implement, you apply Bayes' theorem and you say, is it more likely that these two gene signatures were generated by the same process or by a different process? If they are generated by the same process, if the model that they were generated by the same process has a higher probability than the one then they have a different process then you cluster them together. Just as you would in a regular dendrogram.

Of course, the threshold you pick-- the probabilistic threshold that you pick-- is going to determine exactly which branches are brought together, which gene traces are brought together. But nonetheless, you now have a strong probabilistically based and Markovian based reason for bringing together two gene expression signatures, so then based on the marginal likelihood of these two gene expression patterns being generated by one process versus the other.

And you can thereby eliminate them appropriately. Since I'm running short of time, let me just tell you that the patients have a number of nice tricks in their armamentarium. Which essentially, depend on the following observation that if we use the same data for the models, if we assume all models are initially equally likely, then you can do that very nice equivalents shown at the bottom such that, you only have to calculate the marginal likelihood rather than absolute probability in calculating which models are most likely. And therefore, simplifies a lot miscalculation.

And incidentally, this program that we called CAGE is available publicly for-- if you look at time series. But what does it do? So if we took the same time series that you looked at with the expression data from the foreskin through this same Markovian clustering program, we only found four clusters. Two small clusters and two large clusters. So what's with that?

The two small clusters-- the two small clusters were one, a cytokine cluster, and then an apoptose cluster. The large two clusters had a bunch of different genes in them. And they contained essentially, all the other clusters that had been obtained in the original analysis. So you could say at this point, well, Zack, you guys didn't do a very good job because despite all this convincing talk about the Markovian nature of gene expression, they were able to really pull apart a lot of processes and you were not.

So how do we know that we did a better job or not? Well it turns out that back when they did their original experiment, Unigene, which is a database that Alberta will tell you about was an earlier stage. And what do I mean by that? Every time GenBank grows by a certain amount, by which I mean, more sequences are deposited into the human genome database, there's a periodic reassembly of the putative genes based on the better and better populated jigsaw puzzle of GenBank sequences.

And sometimes it means that two things that were supposed to be different genes, when you have enough bridging sequences, ends up being the same gene. And a gene that was previously thought to be the same with enough distinguishing sequences, ends up being split into two. So these Unigene builds are different. And consequently, when we look back in 1999, 238 out of 517 genes were unknown. And we relabeled the genes according to the current state of the art at which point, only 20 were left unknown.

And there were 19 genes that were present twice in the data set. And the original clustering puts four of these in completely different clusters. Whereas, we only did this once. If we put our marker order back to zero, that means we ignore the past, we get exactly the same misclassification of these identical genes. So the conclusion here is that the temporal ordering does matter. And doing just a simplistic correlational analysis will not cut the mustard when you're really try to dissect some of the processes.

And using a statistical measure of clumpiness of these clusters, rather than a looks right test, which had been the standard for most of these papers from 1999 to 2003 probably is also not a safe thing to do. So it turns out, those two choices I showed you before were in fact, the same gene. Although, it was not known to be the case when it was first published. And these two traces were put in two different clusters.

And likewise, these two traces also ended up being the same gene. And they too were put into two different clusters. All right. Let me wrap up by talking about another aspect of genomic medicine. Those of you who've done clinical research know the following is true, which is most clinical trials are censored. By which I mean, either some patients drop out of the trial before it's completed. Or the patient has died by in the trial, if mortality is your endpoint.

So it's unclear whether, for instance, a patient who just died right after in the trial should have been included or not. And the big picture is that by having a particular cutoff point on the study, you're not allowing yourself to see the full evolution of the patient's history. And therefore, depending on what the nature of the study, you may actually have a very strong bias.

And it's been shown in clinical research again and again that if this censoring is not attended to, you'll have a misinterpretation of the clinical trials. So that's well known in the clinical research arena. And the question is, is it equally true in genomics as applied to clinical research? Well, let's understand what most of the studies published actually do. They often do the following.

They take a set of gene expression patterns, and do some clustering operation such as what [? Eliza ?] did originally in their paper. And they say, oh look, there's two or more in actual clusters just based on the gene expression profile. And then they ask themselves after they found these expression profiles, is there anything different about these patient populations? And they rummage around.

And in this case, they found a wonderfully different mortality between the two groups of patients as defined by the expression pattern. So let's review that. You first look for a difference in expression pattern. And then you say, what makes these patients different? And it makes for a very impressive publication because you're not looking at the patient, just saying, I'm looking at gene expression pattern. And then when I open my eyes and allow myself to see what's different in a patient in these two groups, I'm finding something that's really clinically different.

Now what's the problem with that? The problem with that is that you may be looking at the wrong thing for an example. With 10,000 genes to look at, there might be a cluster of genes that for instance-- let me just actually, before I come up with a fanciful example-- it might be that the reason that these patients have different expression profiles is because they weigh a different amount. One group is fatter than the other because one group is sicker than the other, except it was not picked up clinically.

But one is going down the tubes and one is not. And what we're really seeing here is not something intrinsic about the tumor, but something intrinsic about the weight. And so getting at is a very indirect measure of mortality. And it may be very much unrelated. Can we do better by going directly from the gene expression pattern to the clinical phenotype of interest rather than going through this two step process where we first cluster the genes without any knowledge of the phenotype and then see how the phenotype is different.

Specifically, can we directly find genes or linear combination genes that are highly correlated with survival times? For example, gene A plus 0.5 times gene B, plus 2 times gene C equals some probability of survival. Can we do that? And can we do that in the context of the kind of censoring that I described? So in summary, how can we use the survival times directly to find good predictors?

So here's the fundamental problem. We have gene expression data. We have phenotypic data. And we want to find out how one predicts the other. Now previously, if I asked you that question, you'd say oh Zack, let's just use linear regression, logistic regression, for instance. It turns out with 20,000 variables, it breaks down. It just does not work. And so a nice answer to something called partial least squares. And what is partial least squares?

Well, let me give you the intuitive feeling for it. I explain to you what principal components were. Principal components, being these components that capture a large component of the variance. And subsequent components are orthogonal to each other. This is like principal components, but as they relate to a specific outcome variable like survival.

So these are principal components that are correlated essentially, with an outcome variable. So that's interesting. But the problem with perfect least squares is that it does not actually allow you to use sensor data. it assumes that the data is complete throughout. And that's going to immediately bias your data.

So how do we take advantage of censoring? Now in classical clinical research in medicine, what you use is something called the Cox model, which essentially is a probabilistic model that says, what is the hazard of dying for this population of patients? And basically, imputes for the missing data fundamentally. And for a small number of variables or genes, you actually can use a Cox model to actually figure out, essentially, what would have happened if you'd had a complete data.

But the Cox model, just like regression, does not work well for tens of thousands of variables. So what are we going to do? How do we get to use partial least squares? To make a long story short, I guess I'll post a paper to myCourses website. What you do is actually model the patient's data using a Poisson regression. So that you actually transform this data set from one with missing data, to one with complete data based on this Poisson regression.

But the problem is now we can create essentially many more data points, essentially. And just to give you an intuition behind it because I'm going to run out of time to go into a full description is, this will allow us to estimate the fine problems. So let's say that a patient lasted until this point and dropped out of the study, where might they have died? What can we assume, based on this prior data?

So by using a plus model, we can enter all these new data points. And so we can actually figure out precisely, or probabilistic-ally I should say, what the values were the various time points and completely in this probabilistic sense, so now we can apply partial least squares. So what does that mean, pragmatically?

So Bhattacharjee, how do you pronounce that?

AUDIENCE: Bhattacharjee, yeah.

ISAAC SAMUEL KOHANE: Bhattacharjee, and basically working with Todd Golub, did a study of lung cancer, which had in fact censored data. And each patient had a survival time and it was marked whether they were censored or not. And the question was, could they identify different outcomes? And the short answer is, they were able to identify, for instance, metastatic versus non metastatic, but there are a lot of other outcomes that they could not distinguish from one another using these 125 samples.

So we asked ourselves the question. If we took into account these partial least squares on top of this Poisson model, could we actually tease apart in a reliable fashion some new phenotypes. And the short answer is these are uncorrected P values, whoops. These are uncorrected P values, but tenth of -7, if you do correction for multiple hypothesis testing, can see that they're actually going to be highly significant.

We were able to obtain components that were highly predictive from the partial squares analysis. Highly predictive of clinical outcome. Where previously, you could not. And to make that very clear, these were two groups of patients that previously were not distinguishable. But by using this method, we were able to distinguish them with a P value of this amount. And provide a very sharp distinction between these two groups of patients that was actually not resolvable using the standard analysis that done previously.

And the reason is because I just want to bring this home to you, because we were able to go directly from gene expression to the outcome of interest. We didn't ask ourselves, let's cluster the patient data and then see how the patients are different after the fact. Who said, I know the different characters of the patients in this case mortality. What combination of genes best predicts that difference?

And by looking at that direct signal, rather than indirect signal with all the complications of having to deal with sensored data, we're able to separate out very, very well these subgroups of patients. So, again, showing you how by using a little bit more sophisticated computation, we're able to identify subgroups of such patients. So this brings me to the problem and I want to spend a little bit of time to make sure you all understand the problem set.

So I think it's very important for you to get your hands dirty with the data. So there is a data set that's available at the following URL. And if you just Google Kunkel, Kohane, Haslett, you'll probably find it. But that's the full URL. And I'll send out to the group the URL through our group email.

And that's a data set from a paper that we published, I think, a year or two ago about Duchenne muscular dystrophy. Duchenne muscular dystrophy is a degenerative disease of the muscles. It's the major cause of what Jerry Lewis raises money for. And we have data on the muscles of patients who have this disease and people who don't have that disease.

And what I want you to look at is that comparison. And I want you to identify those genes that are differentially expressed. Now, if you want to work in pairs, that's fine. But no more than pairs of you. And I'd like to use at least two different tests. One is SAM, Significance and Analysis of Microarrays. The other is the t-test, which I hope most of you know.

There's many, many toolkits that you could use to do this comparison. Let me recommend to you the MEV, the Multiple Experiment Viewer at Tiger. Tiger is the Institute for genomic research and at Tiger.org. So if you Google MEV at Tiger, it does many more things than just these two tests. But I recommend it to you for these doing these two tests.

So given these data, I want you to answer the following question. What is the difference between the top ranked 50 genes differentially expressed, both up and down by the two methods? And why are the lists of genes different? I'd like to use another program called Map Finder to classify these genes by their function.

And so the output I'm expecting from you is two sets of differentially expressed genes based on these two different methodologies, your explanation why these sets of genes are different and the classification of these two sets of genes by functional annotation. It's a very simple exercise, but I think it's important because if I've learned anything about this area, it's just doing it by yourself a few times opens your eyes up to how bad the data is and what the problems are in analyzing it.

And if you're not comfortable in downloading tools, then partner with someone in the group who is. All right.

AUDIENCE: Your course is on the website?

ISAAC SAMUEL KOHANE: [INAUDIBLE] will post this on the website and I will send an email out of this URL. Any other questions?

AUDIENCE: When's the due date?

ISAAC SAMUEL KOHANE: That's an excellent question. It's now February 24. How about-- when is March 15? Can you look up what day of the week that is?

AUDIENCE: It's a Monday.

ISAAC SAMUEL KOHANE: So March 16. And please let me if you're having problems. The guy who actually built MEV is a friend of mine at Tiger, so I can kick his butt if it doesn't work for you at a particular time. All right.