MARCO RAMONI: What I'm going to talk about after this little introduction about microarrays is how to analyze this BLAST data. And the principle that I try to present to you is that there is no such a thing as putting your data into a freaking machine and expecting to get an answer. The type of analysis you make is always related to the question you're asking.

This has to be a completely stupid point. But the tragedy of a lot of this field is that it's not. And a lot of people usually try to answer the same question using different methods and different questions using the same methods, which is even more disturbing. What I'm trying to tell you today is what kind of problems you can tackle with this kind of data and what kind of analysis you need to answer every different question.

It's going to be very basic. I will introduce some kind of advanced notions at the end. But most of the rest is very basic and is what is routinely done by people in papers, in genome centers, and things like this. And this is important for you. Because at the very end, I will tell you the bad news. You have an assignment. And you have to use a couple of programs that are describing this thing.

So what I'm going to do is start from the microarrays, tell you what you do with supervised classification and differentiable analysis, argue the prediction and validate your results. How do you do unsupervised analysis using basically clustering for different type of methods and different types of experiments? And then, at the end, I'm going to talk to you about base networks, which are those things that a few of you know are my passion.

This is exactly the same slide I had last week, two weeks ago. This is central to molecular biology. I will not recapitulate it here, except to say that DNA gets copied in RNA. RNA gets copied in proteins. Proteins do all the job. So last week, we talked about DNA. Today, we are talking about RNA, OK?

So all your cells in your body come from the same cell. And they all have the same DNA code. What makes a neuron different from a fingernail is that the proteins that are made of this are different. So the same code will express different type of RNA, which in turn will be turned into different proteins. And these will give different cells their different in nature.

The idea of studying this is called functional genomics. What we talked about last week was structural genomics. We were looking at the structure of the DNA. Now we're looking at the action of this DNA. And we're looking at the function that each different cell performs while it exists. The aim of the game is elucidate functions and interactions among genes.

Now functional genomics is a very old thing. I mean, you don't need computer scientists to do functional genomics. Functional genomics means try to understand what is expressed in a particular cell. And you can do it by hand. You can do it one gene in the time. And people have been doing one gene at a time forever.

What is changed in modern functional genomics is the introduction of microarrays, which are these platforms that allows us to look at the whole transcripts, all the RNA, every gene in a cell and see what is expressed and what is not. This is what is changed. And this is why you need, at this point, computer scientists. But there is a little different change in this. There is an intellectual change in this, a very dramatic intellectual change in this.

If I have to pick up one gene to see if this gene is expressed in a cell, I have to go in and read papers and, at some point, decide how to allocate my next two weeks or two years to see the expression of this particular gene. Sometimes I need money. So I need to put in writing-- to somebody else that hopefully will give me money-- why this darn gene is important, right?

Now with microarrays, they don't need to do that. With microarrays, I use one microarray. And for a tissue, what they observe are 40,000 genes. I don't need to justify which gene I'm interested in. I look at all of them. And this is a very interesting consequence. One-- one is cool. One, I can look everything in action. I can try to be surprised by my results. But the other thing is that I get very different type of information.

That is, suppose I have been spending two years seeing if a particular gene is expressed in a particular cell. And what you get at the end are two pictures-- one of this gene in a normal cell and the other in the cell you're interested in-- and see if there is some kind of change. And on one side you see a ball this size. And on the other side you see a ball this size. You take pictures and you send in your paper, right? This is what people do.

And probably, you are going to make in your-- after you spend two years-- maybe even if the ball is not really that big, you are going to make a very long argument that this ball is really, really big. And there is some particular reason why you shouldn't take this ball as small as it is, right? Now when you have microarrays, you're measuring all the genes. So even if the ball is not that small, maybe there are lot of other balls that are much bigger.

So at that point, it's kind of difficult to say, you know, my gene is kind of interesting. Because it ends up that your gene is expressed as another 20,000 genes, right? So you get another piece of information by looking at the entire genome. You also get what are the most dramatic changes. What are the most dramatic things that happen in that particular cell?

And this is kind of interesting because you produce a new intellectual style. The new intellectual style is not hypothesis-driven and has been a disaster for the biomedical research culture-- for us, basically-- and still is a big problem when you submit grants. When you submit grants, the way you write the grants is to say, this is my hypothesis. This is why I think this is true. This is why I think this is important and interesting. This is what other people have seen-- and myself, I have seen-- to support this hypothesis. And this is what I plan to do.

Now here, the hypothesis-- what the hell is the hypothesis? I don't know. I have a very vague scheme. And I can say, well, you know, my hypothesis is that the genes expressed in prostate cancer are different from the genes not expressed in the normal tissue. Gee, what a hypothesis. I mean, you don't need to go to grad school to come up with an hypothesis like this, right?

But then the intellectual styles is completely different. So we are heading into something that has some simple statistics and some simple technology but has a much broader impact on the way people think about biology. One of my dearest quotations is from a physicists of the 19th century that used to say that, "there are two types of science, physics and STEM collection."

What this mean is that physics provides mathematical, quantitative models of phenomena. STEM collections is going around and measure animals, and put them in your collection, your album, and show them to friends, and maybe arrange them in some way. What this thing is doing to the intellectual landscape of modern biology is, hopefully, to turn STEM collection into highly quantitative science.

The characters that are behind all these things are microarray technology-- microarrays. They are able to measure the expression of thousands of genes at the same time. And now we have microarrays that are able-- on a little piece of plastic this size-- they are able to measure the expression of 55,000 transcripts, which includes all the estimated 35,000, 40,000 genes in the human genome.

Technically, a microarray is-- although it is called an array, it's actually a vector. So for each cell in my array, I have a label that tells me the name of the gene there. And then I have the value of expression for that particular cell. So at the end, when they put them together, they become a vector that associates to each gene-- its particular value of expression in a particular cell or a tissue.

Therefore, the arrays-- because, of course, putting down a vector of 25,000 genes, of 25,000 little cells, is less convenient from a geometrical point of view than putting down a square thing. It takes less space. They are called arrays but that can be kind of misleading. And there are two types of arrays that are currently used more frequently. One are called cDNA and the other are called oligonucleotide microarrays. I'm going to tell you in a second what they are.

How does these things work? It works by reversing the natural phenomenon of transcription, right? So the idea is that I have special glue in each cell that is shaped exactly as the transcript that they want to measure. Then, I will have my tissues that go on this microarray. And by some diverse method, they will hybridize. They will attach to the cells that are specific to them. And then, I wash them away.

And what they have to do at the end is simply to measure how many of this RNA or how much of this RNA is left of a particular cell. How do I do it practically? It works like this. I get a bunch of tissues. Let's say I have only one tissue, for the time being. I'm producing one, single microarray, right? I get my tissues. I extract the RNA. And I tag this RNA with pieces of transcribed RNA with some fluorescent dye.

Then, I put them in the dishwasher, [INAUDIBLE] station, which is microarray washer actually. And I hybridize it. I put it there. And then I scan it. So what happened is that-- you remember, this RNA is left attached to this particular cells. And because I have tagged it-- it's tagged with a fluorescent dye-- I would have more intensity in those places in which a lot of things are attached.

So once I wash it, I can actually use a scanner, like the one for your picture at home, exactly a scanner. And that scanner will come up with a picture that looks like this one, in which each ball represents how much RNA is left in that particular spot. Because I have created or somebody else I pay has created that particular spot, knows exactly what is the transcript that is there. And they can tell you that the third spot from the left is gene [INAUDIBLE] [? alfalfa. ?]

I can go measure the intensity of this gene and then turn everything into a database. I know for each probe what it represents. And for each microarray, I can measure how much of that is estimated in my sample. Question? This is how it looks like. This is cDNA microarrays. cDNA microarray-- I lied a little. You can't use scanners for this. You use some kind of laser stuff to read these points.

But the idea is that in this microarrays, you copy the entire transcript of a gene. So I know how a particular gene-- what is the sequence that is transcribed for a particular gene. I make 1,000 clones or a million clones of this. And I put them in one spot. Now this microarray has two channels that I can read using laser scanning, OK?

So I will have two samples. And I dye one in red and one in green. And I put them on this microarray. And they will compete competitively-- hybridize to this one. So if both things are highly expressed, what I will see is something that is yellow. If none of them is expressed, it would be some kind of black-grayish thing. And if the green is more expressed than the red-- will be greenish. And if the red is more expressed than the green, it would be reddish. And you can see here-- you see a lot of yellow balls, few green balls, and few red, and a lot of black. What's that?

AUDIENCE: Just a basic question. What kind of different information will you get from the [? array ?] as opposed to [INAUDIBLE] intensity. Do you get any different--

MARCO RAMONI: No, this was just because it was the original-- the original one were made this way. To use a scanner-- which I guess what you're going to-- to use a scanner, you need to use silicon technology. And these are glass slides. So to build this kind of things is much easier. You can build these things at home. You can buy a robot that will spot the stains for you.

The oligonucleotide microarrays-- I'll show you in second-- use another type of technology that requires really a production line. So it's not something you can do. So these things actually offer you flexibility. If you're interested in 1,000 genes rather than 40,000, you can do it. The other thing is that it costs much less.

The problem is that because you copy the entire transcript, when you take your RNA out of a cell, you crash it. Then you dye it. Now what happened is that there may be a lot of crap floating around that is really not related to your gene. There may be very, very small fragments that are going hybridized to some random sequence in your clones, just because they are very small to hybridized there.

But still, they will bring fluorescent dye to that particular spot. So the precision of this measurement is not really as great as it could be. You get a lot of random hybridization. There are tricks you can play for this. So if you want to make a comparative experiment, you can put one condition in one and one condition in the other channel. But you can use some kind of RNA soup that is not supposed hybridized to anything and put it on one channel.

So if you get a lot of random hybridization on one side, this will pick it up and makes your point, your spot, yellow. In this case, you will treat yellow and black exactly in the same way. It's not like, no, I'm undecided. Both are up. You will say, well, this is up, but it's up because of some possible random hybridization. The resolution is to use a computational methods.

And these microarrays produced by Affymetrix are like Microsoft Office. It's something that everybody uses. They are far more expensive than they should be. And everybody hates Affymetrix. But still, you can't live without Microsoft. You can't live without Affymetrix.

AUDIENCE: Was the last slide a cDNA array?

MARCO RAMONI: Yes, it was. So cDNA means that you put the entire transcript. And usually you have this two-channel dye. [INAUDIBLE] is a great expert in cDNA microarrays, which actually comes from one of the very first departments that made cDNA microarrays. And is [? Formica, right? ?] So oligonucleotide microarrays follow this idea. OK, I have my entire transcript. And my problem is this random hybridization. What can I do?

Well, because I have the human genome-- the entire draft of the human genome-- I can take this gene and find out if there is a sequence, a small sequence in this gene, that is unique to this particular gene, right? So in this case, even if the sequence is small, if the broken part of RNA in my sample is small, it will not hybridize there because this sequence is too small and too specific.

So the idea here for oligonucleotide microarrays is to say, I'm going to take my transcript-- the entire transcript, the subsequence of the transcript-- and I'm going to sample it 20 times, between 16 and 20 times, and find out these 25 small sequences that are specific to that thing. And, for good measure, I will create another sequence, a sequence that is exactly to this very specific sequence, except that I have the base in the middle that is flipped.

And then we check that the sequence with that base in the middle is not specific to any other gene. So I had a positive control and a negative control. And this way, what I'm going to do is to have a very specific measurement and a very specific measurements of random hybridization. Then, once I have these 20 measurements, I will find some statistics to put them together. It's not easy because these measurements are not independent measurements. But it doesn't matter.

I will put these things together somehow. And the measure I get at the end is going to be pretty accurate, right? This is why they cost a lot of money. This is why a lot of computational work goes into it, because you have to search for all these sequences. And this is why sometimes these microarrays get completely screwed up.

There was a famous case a couple of years back in which they created a new mouse microarray. And somebody-- I don't remember where-- reanalyzed the sequences of their microarray-- that new edition of the mouse microarray-- and found out about 25% of those were screwed up. They were not specific. They were not following the standard design. Hello? OK, so this is how they look like.

This is the scan microarray. This is Affymetrix microarray. This is how it scanned. The little spots here on the longer transcripts but are these probes that are sampling for one particular, specific sequence. And they are scatter in the microarray so that if something bad happened to a corner, it would not affect something else. Otherwise, you could have biases in the entire microarray. There were one next to each other. Your whole microarray will be screwed up.

And that's how, in theory, one probe should look like when it's hybridized. Up there, all these transcripts are more or less hybridized. And down there, the random hybridization is not really hybridized. But the resulting hybridization will be the difference between the real hybridization and the random hybridization for each probe, for each probe pair, the negative and the positive, and then a global measure to put them together, which I will not bother you with.

So what's the problem? The problem is that this stuff costs $1,000 a pop. $1,000 a pop is a lot of money. And you remember what I was talking about-- the hypothesis-driven thing. When people draw balls for a single microarray, usually they do it twice, at most, three times. But if you're measuring 40,000 genes at the same time-- well, measuring twice is going to be a little problem, also, because you don't have any hypothesis to prove, right?

So this is where the major cultural clash comes in. That when people analyze data, even in medical domains, database looks like this. There are some variables and a lot of cases. The microarrays data set usually look like this. You have a lot of variables, thousands of variables, and very little measurements. And it's kind of funny to see these people that work on the genetics and define from a genomic side.

So when they design a genetic experiment for SNPs, they collect 5,000 patients, 2000 patients, 3,000 patients because that sample size is required to analyze a couple of SNPs. But then when they do microarrays, they expect to find the [INAUDIBLE] 5 microarrays out of 45,000 probes. It's exactly the same people. So what you can do with this?

Well, let me introduce you to a notion that will remain very precious to you. When people will be confused-- as some people are-- you will have a very easy and fast answer. What is the difference between supervised and supervised? Is exactly the difference between a normal movie and a PG movie.

Supervised means that there is something or somebody supervising. They're telling you things. This is what a supervisor does, tells you things. So a supervised thing means that they have either a human or some kind of signal that will tell me what a particular sample means, right? A typical supervised problem is-- let's try to decide what characterized the people in a particular room.

I get measurements from this room and that room. And in this room, this is a graded course in functional genomics. In the other room is a class from the dental school. And let's make differences. We have properties of these people. Let's see what is different. But I will tell you that people in this room are different from the people in that room. And this is what is called your training signal, the difference between two clusters.

In an unsupervised thing I know I have no supervision. I'm old enough to go to a R-rate movie. So the question in this case is-- I get a bunch of people-- are there groups among them? There are people that look more like others. There are people that control other people. It looks more like, I can say, gossip, finding stories in these things. But they answer two very different questions.

One is, what is different between these two groups of people? And the other is what is similar, or what is related, or what are the stratification, or what are the things that we have in common among these different people? It becomes clear. So what can we do with the microarray? With the microwave-- well, the first thing we can say is, OK, I have two experimental conditions. And my aim is to see which genes are expressed more and which genes are expressed less in this condition, right?

So typical example is cancer. I get a bunch of people with cancer. I get a bunch of people without cancer. I run microarrays and then see what is different. What does it mean? Well, it means that they have tissues from healthy cells and from tumor cells. And for each sample, I will create a microarray. And this is how my database, at the end, will look like.

So the first column represents the name of the gene, name of the transcript directly looking at. And the second column represents the value for that transcript for sample one, sample two, sample three, sample four, and sample five. And then, I'm going to tell you, well, sample one, sample two, sample three belongs to one category, are in this room. And sample four and five are in that room. Go and find what's different between these two things. Now, what does it really mean-- what is different?

Remember, so if we do it by hand, we can take pictures of balls and say this ball is bigger than this one. But if I have 50,000, 40,000 balls, what I'm going to do about this? Well, what they want to do, in this case, is to find what is more expressed in one side rather than another. And the currency provision-- that I will tell you in a second-- the currency of these measurements are called folds. Fold is how many times one condition is more expressed than the other condition.

Now the problem is that this is good when you have one single ball. If I have 50 patients, what the hell I do? Should I take the mean? Sure. I can take the mean, but then I don't have any measure of the variance in my data. Maybe I have two things that the mean are very far apart. But because the variance is very big, they will overlap. So there is not much evidence that they can collect.

So other measures are things like standardized differences, and make the difference, and standardize them by the variance, which will somehow take into account the variance, that is under the assumption that these things are normally distributed, so that kind of variance has any statistical meaning. Then what do you do? Well, then I decide the threshold. I get the landscape of this thing. And I'm going to say the top 50 genes are what I actually like and the bottom 20 genes.

The top 50 are the ones that are more change in one condition and the bottom are the ones that are more change in the other condition. I'm going to pick up this stuff and see if there is anything interesting. What people do typically is to make up stories or to pull up a protein-- like people are doing with their project-- pull out the protein and see if I can actually-- a gene, find out the protein, and actually find out if this protein does something to my particular phenotype.

This project, which is a project about preeclampsia the investigator there-- you get only two microarrays from a preeclamptic-- preeclampsia is a disease that women get during pregnancy. It's a very bad disease-- and the normal placenta, compare them, and pull out the protein. Put the protein into mice and found out the mice were getting preeclampsia. There are this kind of an exploratory thing.

In this case, what I'm interested in is find out new hypothesis. Then they can test in some kind of laboratory setting. As I said-- because, in this case, I have only two samples. But suppose I have several patients, what I can do? One other problem we have here is that we are not really sure what kind of distributions are running this microarray.

So what people say is, well, because we didn't know the distribution, let's use some distribution-free method, which is a good idea. But it's an idea that rests on the hope that there is some free lunch of life. And there is no free lunch in life. No parametric method, distribution-free method requires a lot of data because you have to do two things. First, you have to decide what kind of distribution you have, implicitly. And then, you have to run your test.

People use parametric method typically because they have an idea of the distribution. And so they need less data to fit this test. If you have few data and no idea of the distribution, you are screwed. And running this kind of test tends to be kind of a dangerous. One, because usually your sample size is too small to run a proper parametric test. Two, because frequentist people have these things called p-values. P-values are very interesting animals. What is the p-value? Who gives me a definition of p-value?

AUDIENCE: [INAUDIBLE]

MARCO RAMONI: Speak up.

AUDIENCE: The probability that--

MARCO RAMONI: Two things are different?

AUDIENCE: --that the means [INAUDIBLE] very different.

MARCO RAMONI: OK, so this is what patients believe. But to do that-- you work with patients too much. And that's a very reasonable measure. I'm interested in finding what is the probability that these two things are different, right? This is not the p-value. The p-value is the probability that you will make a mistake if you repeat the experiment N times and compute it as the number of times you will be mistaken by repeating this study, which is an extremely [? masturbational ?] measure.

There is no relationship with the probability of your hypothesis. And it's very difficult to put into practice. First of all, people should explain me why I should repeat my experiment 100 times when I already repeat 20. And this is what I know, right? The rest is educated or uneducated guess. But the p-value, in this case, has this other little problem. That because I repeat the experiment a lot of times, sometimes things may come up just at random.

So if I say, OK, I'm going to accept something if my probability of is 5%-- so the p-value is 0.01-- if I test two hypothesis, to maintain the same level of error, because I have the probability that something will come out at random-- assuming that these two tests are independent-- I have to multiply the probability of this p-value, right? So my real threshold to get a 5% evidence of a p-value would be the product of these two 5%s to maintain the same level of strength of evidence.

Now imagine if you have to multiply 0.05 40,000 times. What kind of threshold you get? Nothing. Nothing will pass that particular test. This is called Bonferroni correction. Nothing will pass the test. I have a very dear friend of mine who is very frustrated by this and decided to be a biologist after trying to use p-test on this kind of experiments. Because the threshold, the accepted evidence is 0.05%, will turn against you when you are testing your hypothesis 40,000 times.

Besides, this is under the most lenient condition. Because you'll assume that all your hypotheses are independent. But we know that this is not true. We know that these genes regulate each other. So the probability something is up is not independent of the probability something else is up, OK? So even under the most simple condition, we have a little problem with this.

So what can we do? Well, I will tell you in a second what can we do. But what we can do further-- not for the experiment but in general-- so once I have the differences-- OK, I can go back to my lab and put the protein into a couple of mice and see what happened. But isn't there anything better that I can do using these differences? Well, maybe I can make predictive models.

Predictive models, rather than using proteins-- one protein at a time, what are called markers-- are able to put together a batch of proteins and provide a profile, a prediction for a particular outcome. In this case, maybe I can predict if a particular tissue is a tumor or it's not. I can predict if a particular tissue is a type of tumor or is not a type of tumor. Maybe this type of tumor require different therapies. Maybe I can predict how long it will take for a particular tissue to come back as a cancer, because they find a particular signature.

Now how do I find the signature? I have to run a game called feature selection. Feature selection is-- I have a class. I have all these predictors down together. And I'm going to select some of them as good predictors for [? C. ?] I cannot use all of that, right? Why cannot use all of them? Because good prediction comes from specificity, right? I'm glad you agree with this, because it's not really such a normal statement to accept.

People believe that you use 40,000 variables, you're going to give better prediction if you use five variables independently of the quality of this variable. I mean, as long as these five variables are a subset of the 40,000 variables. But we all agree that this doesn't happen, right? Right? If somebody has a doubt, I have a joke. OK, no joke. So what do I do? Well, I want to identify those genes that predict my class, the set of genes that predict the class.

So if I do feature selection, I typically increase the predictive accuracy. I get a more competitive presentation. I can get some insight in the process that may happen. Although, remember, this is not just differentiable analysis. It's something that I want to use as a prognostic or a diagnostic set of markers when combined. And why differences are important? Well, because we start from the assumption is that if two things are exactly the same across two samples, it's very difficult that they will be able to discriminate between them.

So classification, which is this task in feature selection, looks sometimes very much like differentiable analysis, but it's not. I have a twist at the end. And the aim of my game is not really to find out what is different. It's finding what is predictive. And the example is-- supposed I give you two groups of people. And you don't know it, but one group are men and the other group are women. And then I give you a list of properties of these people. And there would be a lot of differences.

Women tend to be slightly shorter than men. Women tend to have more hair at certain age. Women tend to make less money. But there are a couple of anatomical differences that are really good predictors of these differences. Doesn't mean that there is no other difference. But it means that particular anatomical feature is a perfect predictor between male and female.

So if you're doing differentiable analysis, you may be also interested in the fact that these people have differences in income. But if you include these factors into your predictive model, maybe because I'm short and don't make a lot of money, you end up classifying me as a woman, OK? May confuse your ideas.

So we were saying, non-parametric method has kind of little problems with this because we don't have enough samples. But we have classifiers that are parametric classifiers. In this case, we make an assumption about the distribution of our data. And then we try to fit our data into this distribution, thus saving us a lot of effort in collecting more data.

Because data are very complicated and hypotheses are cheap. We can actually go and validate our hypotheses afterward. So this is, as many of you know, is called a Naive Bayes Classifier, in which I assume that each gene down there is conditionally independent given the class. It doesn't mean that it is independent, right? Like we were doing before-- we were doing independent tests, you remember?

The independent tests assume that they are marginally independent. In this case, they are conditionally independent. Conditional independent is that once I know the class, I don't give a dime about the dependency between these two genes. Maybe there is a very complicated relationship between gene one and gene two. But because my interest, in this case, is to find a classification, I don't care.

Because as far as the classification of the class is concerned, these things are not related, right? So it's like a weak independence assumption where weak, in this case, means good. Because we are not forcing an assumption that is too strong within your data analysis. Once I have that, I run-- let me go back a second to this-- the other one. I want the other picture. Come on.

See this? In this case, my genes are marginally independent. The arrow is going the other direction, right? So all these genes [? cause ?] my class, but they are independent. And this is the structure of a standard classifier. In the other case, they are conditionally independent given the class. So once I have this particular model, I have selected which are the genes that I like. I have estimated the parametric model. Then, I can make predictions.

So if I had used some kind of differentiable analysis using a non-parametric test, by definition, I don't have parameters. It's non-parametric. So I cannot really make a prediction with the parameter set [? lower. ?] What people use are things called mixture of expert, in which they assign some kind of arbitrary weight to different genes. And each gene will be like an expert, judging if this particular tissue is a cancer or is not a cancer.

But these weights are actually embedded in any parametric model you derive, which is the probability of observing that particular gene expressed, given the fact that you have a change in your class, that the class is tumor or not tumor. So you can apply Bayes' theorem, and reverse those errors, and obtain the posterior probability that your particular sample is a tumor or is not a tumor. This is how it works.

Well, this is what I just said. [INAUDIBLE] had this one. So I have a class. And I'm interested in the probability of the class, given the sample molecular profile, which is my new patient coming in. And by applying Bayes' rule, I can actually compute. Because the probabilities I have are the probability of each feature given the class, which is the direction of the arrow.

Bayes' rule will allow me to flip this rule backward, apply this as a product, and put all these things together into a single posterior probability. It's just the sum of this probability. I have another interesting thing with this-- another goodie with this thing that I can actually validate my stuff. Well, validate this would mean to go back in my lab and look at a couple of things. Validate means to see how my model is good to fit the entire 40,000 genes.

And the best way to validate something is to have an independent test set. I collect patients here at Harvard. I build my model. And then, I call up my friends in San Antonio and say, listen, I have this model. Do you have 50 patients. for me that I can classify, and you know diagnosis already? And if he has them, then cool. I can really say, this is the accuracy of my model from here to there. But sometimes, we don't have these things. Well, quite often we don't have these things.

So how can we do this cheaply? Cheaply-- we can use cross-validation. Cross-validation means that I take my data set. I split it in five parts. And I use four parts to learn my model. And then, I predict the fifth part. And then, I take other four of these five parts. I build another model. And I predict the remaining fifth part. This decreases the sample size I originally I had already.

So what happened is that people use a thing that is called leave one out cross-validation, where the number of sets is equal to the number of samples. So what means that they pull out one sample. I build a model on the other one. I try to predict the sample that's taken out of which I know the classification. This is an example. One of the first predictive models that came out in 1999.

We have two types of leukemia-- ALL and AML, acute lymphoblastic leukemia, acute myeloid leukemia. And as you can see under the microscope, they are very difficult to diagnose. So what these people at [? Wycliffe ?] did was to say, well, let's collect, I think, 27 and 11 patients, right? And what they did was to create a dummy vector of zeros and ones and then correlate the gene expression-- sorry. The columns are patients. The rows are genes, right?

And now I don't remember if the blue is underexpressed or overexpressed. But what does it mean is that they take some kind of average to represent this picture. And the positive distance of the point from this average is the intensity of the red. And the negative distance is the intensity of the blue. So the more intense is the color, the farther would be your point compared to the mean of these values.

And at the same time, the direction of this distance would be given by the color. So if it's dark blue, it would be very negative. If it's dark red, it would be very positive. So what they did was to correlate these genes and pull out the top 50. So the 50 that correlated more with the gene, with this dummy vector with the positive correlation and with the negative correlation, 50 and 50.

And what they did then was to make a mixture of expert prediction and see what was the accuracy they could get of their own patients. And since then, there have been gazillions of paper written like this. I want to stress the fact that, in this case, we are not interested, again, in what is really different. We are interested in finding a molecular classification for these things.

The hope here is that one day you can build a little check-- and they are really doing this for literature-- on which you can put some specific genes and have a classification that will tell you this patient has this particular type of leukemia. This patient has this other type of leukemia. OK, so I have talked about something that is [INAUDIBLE] question, is controversial thing. But I thought last night about including this thing.

But then I said to myself, yeah, as long as I tell you that what I'm going to tell you may be kind of controversial. It's OK if I tell you that, right? And this is why you want to go to school to be a professor. Because then you can say controversial things. They cannot fire you, hopefully. One of the things people do to identify differences easier, even with pass it down, cut the threshold, is to deflate the variance, OK?

If I have two samples that are very far-- but if I find a way to squeeze the variance of these cases, then I will have a much smaller variance. And I have more chances that my changes will be significant, right? Because the variance would be smaller. Now this is something that for any other type of data analysis, will send you at least in disrepute, sometimes in jail. If you do this on a company budget or if you do this on a clinical trial, you go to jail.

In microarrays, people don't go to jail. Because it is an original thing that made originally sense. Remember cDNA microarrays? cDNA microarrays have two channels. Now we know that, by design, there is an imbalance between these two channels. One channel is more intense than another. So if I'm comparing to samples, what I may come up with is something that looks like this.

So I have the two microarrays that are lying on two parallel things, right? And you see that there is a bias. All the red on one side and all the blue are on the other side. So what people used to actually do for this kind of platform, because you have two channels, is to try to reconcile these two channels by studying the distribution of these two things and try to put them one over the other. So as a form of correction, you do.

Because, by design, you know that your platform will introduce some bias. And this is fair. This is good. The problem is that when oligonucleotide microarrays were introduced, people just blindly took these things and try to apply to microarrays. And you start coming up with a couple of problems. First of all, oligonucleotide microarrays are not two channels. They're one channel.

So suppose there have 50 patients. What do I do there? Which patient do I take to be my baseline? I'm going to reconcile all the patients with the first patient at the beginning? And what happens if I change this patient? Are my genes going to change? Yeah, you bet so. So now if you really want to have a great success talk with biologists, go and tell them that they shouldn't normalize.

Because there are about 100 different normalization methods of this type. And people are confused. But people are confused because there is really no need. People are not confused on normalization and cDNAs. People are confused for normalization to squeeze your variance and get better results. Because in reality, even when you have design with two channels-- so I have a pair case and control with microarrays. You get actually results that look like this one.

Now these are microarrays that come from an institution from this [? street ?] to which I am not affiliated and nobody here is affiliated to. So I can actually speak on them. And this is a good example of why not to do normalization. So these are people before and after treatment, OK? These are paired experiments because it's the same person that is sample before treatment and after treatment.

So you remember those lines that were like going one after the other? Means that we were plotting the intensity of one channel against the intensity of the other channel. So look. We plot this microarray against this microarray, which is the microarray before and after, right? So in this case, yeah, more or less, it looks like the other one. You remember? Now look at this one. Can you imagine any transformation that will put those things along the same line?

Yeah. Look at this one. So in this case, what happened is that there is something that is highly screwed up. And, again, these are following exactly the same design that cDNA followed, the experimental design, although the end practice is absolutely different. So my advice as far as normalization is concerned is don't change your data that may be useful. But try to look at your data because they may contain some important information. This microarray is completely screwed up and should be either removed, redone, or done something about it.

OK, so what have we learned? We have learned that we can actually find differences among samples in different conditions. We can make predictors. Have we learn anything interesting about the genome cells? Not really. We have learned nothing about the relationship among genes. Although we are measuring all of them at the same time, we have completely disregarded-- actually we have fight against the idea that these things could be related.

We are simply interested in finding something that was different in two conditions or simply interested in finding something that, put together, could predict this condition. That's it. This is where supervised classification brings us. If we want to take advantage of the fact that we measured all these genes and we observe the genome in action to try to decode something about the genome, then we need some different method. And we don't need supervision all the time.

It's like when you're a kid. If there is supervision, there is very little fun. So the easiest thing we can do is to say, well, OK, forget about supervision. I got this bunch of genes in different conditions. Forget about these conditions. I don't care about these conditions. What I want to see is which are the genes that behave more similarly across all these different conditions? It's like having a car, right? Try to understand how it works by kicking it, and kicking in different points of the car, and then see how the things go together under different stresses.

So if I kick the wheel, if I keep the trunk, if I kick the door, what happens? How these things move together? What is the relationship among these things? Was a nice analogy some time ago. But you study these things. And the way in which you study these things-- this was for sequencing the genome. Well, it's like you have, in the future, somebody comes up with the Volkswagen. And they discover a Volkswagen [INAUDIBLE] somewhere. And they have no idea what it is.

So to understand how this works, they take the Volkswagen and they throw it off the cliff. And then, when it's down, they try to put the pieces together again, right? This is what somehow we're trying to do. We are breaking this down with some kind of solicitations and trying to see which parts behave together. So, in this case, like we had 1,000 Volkswagens-- well, 100 Volkswagens. And we keep throwing them down.

And at the end, when they are down-- because we don't know how to open the engine-- when they are down, we will see there are some pieces that are closer together. And they remain closer together. And this is independent of the fact that these two things fall to the left or to the right of the main body of the Volkswagen. So a simple thing is to say, well, let's measure correlation among these things.

Genes [INAUDIBLE] supervision. They have a lot of solicitations. They could be different compounds that are treating a particular disease. These are maybe different type of cancer. I don't care. I don't want to find classification. I just want to find out which are the genes that go together around these conditions. If I use correlation, the only thing I can do, though, is to look at pairwise comparisons, right? I can only say that one gene go to another gene.

A correlation is a distance between two points. I cannot have groups of three, or five, or 15. How can I put these things together? Well, I can use another type of clustering called hierarchical clustering. Hierarchical clustering start putting things together. But when it puts two things together, it creates some kind of a dummy gene, which make us feel like the average of these two genes or something like this. And then try to correlate this average profile, this average gene, with other genes.

So, at the end, the result would be something like this. Again, it's like the blue and the red. In this case, is green and red. These are the Stanford color. Wycliffe uses the blue and pink. Duke, I think, use yellow. John Hopkins use green and blue-- well, a few combinations of this. But you can actually recognize at least the platform they're using by the color of their pictures.

So in this case, this is a Stanford picture. Again, the green is down and the red is up or vice versa. And you can actually, by visual inspection, see that there are some points that are very highly expressed away from the mean, are very down expresses, very [? low ?] expressed from the mean. So this is the zoom of that picture. And you can see that these things are creating a tree or Venn diagram. And this tree will put together groups of genes, not only two genes.

And the problem here is that you don't really have a good measure to decide when you've made a group. Because, again, you have one single tree that will combine all of them in different order. So, technically, this is not-- although it's called clustering-- clustering means to put things together and divide them. Technically, this is a sorting algorithm by which I put a particular order-- in this case, a partial order-- over these things.

And then some knowledgeable biologists will come and say, oh, among these people here in this group, I see that there are among this group-- I see that these genes are all related to a particular process. So maybe also these genes that is right embedded between them is related to the same process. And maybe it's apoptosis. And these are five apoptotic genes. And then, they find something else. And we create another group. But these groups, these different coloring-- the pink, the purple, and the red there-- are handmade by somebody with a lot of patience that put them together.

AUDIENCE: [INAUDIBLE].

MARCO RAMONI: Say again. What?

AUDIENCE: The trees are made by hand?

MARCO RAMONI: No, no. The trees-- sorry. The tree itself is built through some kind of metric. I don't know why it's not coming-- OK. So I compute the correlation between these two points, these two vectors of values. Then, I create, let's say, an average value here. And then, I draw these two points. And they consider this new value that I have created as a new member of my data set. I didn't see what this correlates to.

In this case, this correlates to this one. So the highest correlated thing is a gene. And this creates a new hypothetical thing, which is the average of these two and this one. So what happens is that, at the end, I create a structure like this. But the problem is that because they are all measures, at the end, they will have one single tree. So how do I create blocks? The way in which blocks are created-- and I say, color this in purple and this in pink. These were handmade.

I will tell you in a second how you can avoid to do this handmade. I can do something more interesting also. That was a temporary experiment, [INAUDIBLE] second temporary experiment. So I knew the order of these microarrays. But sometimes, I'm not really interested only in the way in which genes go together. I'm also interested in finding some new class among patients, right? This is a very interesting paper from 2000 in which what these people did was to try to cluster simultaneously genes and patients.

And what they came up with were groups. You see those groups up there. The groups up there is not a Venn diagram. Those are group of patients based on some selection of genes that are more expressed across the two conditions. And then what they did was to find out that-- if you look at the survival time-- how many of what is a Kaplan-Meier curve? Everybody. OK, so if you look at the Kaplan-Meier curves of those groups, you see that there are very significant difference in survival, OK?

So in this way, I can discover not something really that is about genes but something that is about the overall classified disease. I find out new classification for diseases with interesting clinical consequences. Again, problem is I have to do this darn coloring by hand. Is there a way by which we can actually avoid coloring this stuff? Yeah. There is a way. And this is the idea.

If you want to cluster, it means that you have to make differences among things. So you can decide arbitrarily the number of clusters. And say, OK, I have 50 clusters. And you divide everything in 50 parts. But why not 49, or 38, or 15, or two. So central notion of the clustering is similarity. If we have a definition of similarity that is specific enough, then this similarity will allow us to say when we can actually cluster without creating a threshold, just a conceptual definition of similarity.

So I have to postulate this description of similarity. And I need a piece of theology before this. But let me postulate this. In statistics, you don't believe that what you observed was directly created by God. What you believe is that there are some processes that you don't observe that generate the data that you observe with some randomness, some measure of uncertainty.

Now let's make an example. Let's suppose we take the electrocardiograms of each of us. And, hopefully, especially for me, all these electrocardiograms would be different. But, hopefully, they would be coming from the same process, which is the process of a healthy heart, mine will be slightly different because it's small but probably will not be different enough from yours to say that this is a completely different stuff.

Now suppose we go to Brigham, to cardiology at Brigham, and we take electrocardiograms of people there. There, I expect people to have differences between themselves that are great enough to be generated by different processes, different pathology of the heart. Now I will pose to you that two things are similar if they are generated by the same process. And two things are different if they are generated by two different processes.

And if you buy this story, then I can give you a method to compute when something is generated by the same process and when something is not. How? Well, we know that these processes that we do not observe but they underpin the data that we actually observe, generate our data with some kind of uncertainty, that is a random process that is generating data from this. An example is aging, right?

Aging has a particular effect on people, usually make you wealthier, usually, after at a certain point, make you stronger, after a certain point, make you weaker, has affect on your marital status. You tend to get married, and then divorced, or widowed, or whatever. When coupled with other variables like gender, can have other physical effects like you can lose your hair if you're male and so forth, right?

So if I find somebody that at 13 is on the verge of his third divorce, that's not impossible. But I would find it kind of unlikely. Why? Because there is a process called aging that dictates, more or less, that people to be on their third divorce usually have to be at least 35. So if this guy is 13, it's difficult. It's not impossible, but it's difficult.

So we have these general expectations that stem from the fact that there is these processes generating the observation that we have and is constrained by other things. Like we're saying losing her is constrained by gender, probabilities change by gender. But at the same time, once I observe the data, I can tell you that something is probable to be generated by a particular process and something less probable to be generated by a particular process, right?

And this is what we want to do. We want to compute the perceived probability that a set of processes, as responsible of my data-- so M given D will is the data-- for each class we model, for each way of combining my clusters. And then, I can combine the score and find out what is the most probable way of combining these clusters. And at the very end, what I will have is a bunch of clusters, not simply a tree, not something that I have to cut with the threshold.

But I would be able to tell you that if two things are put together, they are N times more probable to be generated by the same process than they are to be generated by two different processes. Interesting paper-- you're going to read it. This is how it works. The probability of the model given the data by Bayes' theorem is equal to the probability of the data given the model times [? provision ?] of model on the [? unprovision ?] of the data.

Now I will not get delve into details. But at the end of the day, under some assumptions-- like the assumption that before looking at any data, all models are equally probable and the assumptions that we are trying our models on the same data, which is usually what we do. We have the same set of expression data and want to find the best model. What we can compute is that probability, which is the probability of the data given the model, which is proportional to the probability of the model given the data. And, therefore, we can use that as a score.

These things is kind of compared with this to compute. It's called marginal likelihood. And so we can search all these combinations and find out which is the most likely combination, which is the most probable combination, of, in this case, genes, given the data that we observe. Now let me go-- so these are a couple advanced topics. From now on, this is not subject to examination, for the test.

Suppose I'm interested in something like control. Have we learned anything about control so far? Well, we have learned that things go together, things are similar. But we haven't really learned anything about how things control things. To see how genes control other genes, we need a very important experiment design which is a temporal experiment. We need to see what happened from one point to another.

And you say, well, it's kind of easy. I take this clustering method. I use this clustering method, and I put them together. And then, I will find some kind of similarities. Can I do that? No. Why? Because measures like correlation or distance measures assume that all the observations you have are marginally independent. What happened to patient one in gene one is completely unrelated to what happened to patient two on gene one, right?

But when it's time-- well, when it's time, it's really, really different. Time means that where I am now depends on where it was five minutes ago, 10 minutes ago, 30 minutes ago, 100 minutes ago. So if I keep measuring the same system a long time, my observations will not be independent. Let's put it this way. If I measure things a long time, I don't have ground to safely assume that my assumptions are independent. Because assuming that assumptions are independent is a simplification, right?

If I have a model that is able to account for dependency, I can always reduce it to a model of independence. But I cannot do the other way around. And let me give you a practical example. These are two pairs, two genes up and two genes down. So you are measuring the distance between these two genes. Now the correlation of the two genes up there is something like 0.6. And the correlation of the genes down there is about 0.8, right?

But now consider the memory of time. And look at the first picture. Except for the first point, when the first gene goes in one direction, the second gene goes exactly the same direction, right? They never intersect each other. The second point-- it goes from one point, goes down. The second gene goes down. The third point-- the first goes up and the second goes up. And then it goes down, and goes down, and goes down again. And the other goes down again. It goes down a little less. It goes-- look at it.

Now look at the other one, which has a higher correlation. These genes are always one against the other. Every time one gene goes up, the other gene goes down. So if I am actually interested in the dynamics of my system-- why correlation would put these things more similar to those ones-- my good measure-- by keeping in mind that I'm interested in the dynamics of change of this thing-- would actually require a different perspective, a different measure that takes into account what happened before and that we put those two together, those two closer than these pairs.

How can I model these things? I can use a thing called autoregressive models. Autoregressive models-- it is very simple. There are a lot of way of doing this. This is just an example of how to take into account your past. How can I do this? Well, I have a time series of dependent observations. And what I can say is I assume that my observed point, at in this moment, is independent of the remote past, given its recent past, right?

So to know that I'm here now, you don't really need to know where I was the day before yesterday. You need to know where I was 10 minutes ago, an hour ago, maybe two hours ago. But the predictive ability of two days ago, where I was five miles from here, is going to be very, very weak. So you can actually summarize your data, summarize your expectation on somebody being here by forgetting the remote past and considering only the recent past.

The most recent could be one point. And, in this case, you can create a model like this in which you plot your present-- that is, my time now-- with your immediate past. And, in this case, you're assuming that everything-- my observation is independent of my past, given my most recent observation. This is the simplest autoregressive model. Now this kind of experiments, again, tell us something about the similarity of things. Actually this is a kind of analysis. The data are always the same.

Once we have this temporal data, if we do some clustering, we may see that things are working in the same way a long time but are hardly going to tell us that something controls something else. In this case-- as I was saying in the beginning-- it's not really the data of the design of the experiment. It's the type of analysis you make. So if your interest is to find out which are functional clusters of genes that work together, well, clustering is your solution.

But if you're interested in dissecting what is the regulation, the mechanism of regulation among genes, that will not tell you. I may have things that behave kind of similarly, but they not necessarily behave together equally. To be extreme, I would consider that something that controls something else will not really have exactly the same temporal behavior, right? So if I want to have you here today, I have to call you yesterday or I have to be here yesterday. I have to do something before you're here if I'm controlling you, right?

So a way to use these things, to try to dissect this kind of information, if this is the question you have, is to use a thing called Bayesian networks. Bayesian networks are regulating genes-- in this case, relating variables in general-- by looking at how probable it is that one particular set of variables will control another set of variables. Originally, these things were built for humans, humans you want to clone information from, knowledge from.

You are buying lunch or dinner to your physician friend, getting drunk, and distract the promise that will come to your lab the day after. And they will draw a network of this knowledge saying this gene versus other gene versus other gene and then add some probability that describe the function by which a particular gene controls another gene. This particular example-- I'm sorry. There are a couple of people who have seen this example at least 100 times. This is not about genes, just the intuition of what is there.

This network tells you that your age your education affects your income. So this is easy to draw. The problem is how age and education affect your income? This is specified by that particular set of distributions. And those distributions tell you that if you are young and if you have a low education, your probability of having a low income is 0.9. And as you grow older and you get more educated, your probability of having a higher income increases. It's not one because you can always choose to be an academic.

The problem is that we're not interested in doing these things by hand. We are interested in finding these things from data, right? And we can play exactly the same game we're playing with the clustering thing. We can find out what is the most probable set of nodes, which is the set of nodes that are most probable to control a particular gene. And we can do this for each gene.

So the final picture-- lost it. Oh, the final picture of this is this one. Come on. Give me a picture. Here it is. Each ball represents a gene, except these three blue balls. OK, so these are about 40 patients, say, 41 patients, pediatric patients with leukemia. And for these patients, we have measured some phenotypes.

But the most important thing we are interested in is the molecular classification, so the type of is called oncology status, oncogene status, which is the molecular classification of the tumor. And this is their survival. And this is [? finding, ?] it's how many days they've been in the hospital, OK? And what they're interested in is find out if there is a relationship between-- you remember when we were analyzing the other things into different conditions? We were doing one analysis for each different phenotype.

We couldn't put the phenotypes together in a single picture. In this case, we can put the two phenotypes in a single picture and see, for instance, if there is any link that will go from oncogene to survival and how this process is mediated by other genes. And what we can also find out are dependencies among genes and other genes. And you see there are directions in those arrows. And those directions mean actually that one gene controls the other gene.

Example I usually run is suppose we want to discover which of these flickers control these lights, right? So I can do it this way. I can change the flicker. I can change these things. And this will affect these lights to be on and off. But if I try to unscrew those lights, they will not change the state of this one, right? The metrics we use is very similar to this one.

So the metrics will actually take into account the fact that you are measuring the influence of a directed influence from one gene to another gene. It's not just a simple distance. It's not just a pairwise measure. And actually, it's not pairwise because, as you can see, you can have more than one parent. This node here-- just to make an example-- is three parents, this one, this one, and this one. No, sorry. This is a child. And this is a parent. And this is another child. And he also has a grandchild, here.

So you can actually use this kind of information to create a molecular landscape of the control mechanism of your things. And you remember what we were saying about how probable it is? I can actually measure how more probable is something to be affected by some variables than is to be affected by other variables through something that they I will not bother you with. But it's called basically base factor. Base factor is the ratio between the probability of two models, which tells you how more probable is one model compared to another.

And these are the numbers we get. So we say that oncogene status, which had these three parents, we choose these three parents-- these are all the other possible combination of parents we have explored. And this picture tells you that the runner-up-- which is this other thing down there, the second one-- is seven times less probable than the top one to be responsible for the oncogene status. And the third one is going to be 56 times, and [INAUDIBLE] times, and down, down, down, down, down.

And you see, basically, the runner-up, which tells you how more probable is the model you have compared to the best scenario of any other model. So it gives you some kind measure of confidence. OK, so this is an example how you can validate these things. You can actually do cross-validation. You remember? We were saying you pull out one case and make a prediction. And first validation is here. It was 100% and something like this.

But the interesting thing is-- the take-home message for today and the thing that is important is that because there are no hypotheses here, the way you collect the data is important. But the way in which you analyze the data is the thing that is going to give you the answer. So if you are interested in mechanism of control, comparative analysis will tell you squat. If you're interested in molecular classification, clustering will tell you nothing. If you're interested in discovering new types of disease, these metrics will tell you nothing.

Each type of analysis, as a particular type of answers-- is designed to answer them. And this is the most important thing you want to consider. There is a review up there. If you want to be bored to tears, then you can take it down from that website. But it was the state-of-the-art until six months ago, nothing has changed much. So the second is a [INAUDIBLE] book, which is part of your school equipment, right? Didn't you have to buy this book, yes, for the course?

AUDIENCE: No.

MARCO RAMONI: No? OK, go and--

AUDIENCE: Not that I know of.

MARCO RAMONI: --see it because he's the director of the course. You may want to kiss some ass. Gene cluster and SAM are the-- you member The two non-parametric statistics I was describing before? Age is the thing that implements the Bayesian metrics and the temporal analysis. And what I'm going to do is send around an assignment, which will probably be a data [? study. ?] And you will do two different analysis for it.

I don't remember if you have to do both of them or if you have to choose which one you want to do. And one is going to be a supervised analysis using either a gene cluster or SAM, two different statistics. And the other is going to be an unsupervised analysis using gene [? calssification. ?] Gene cluster is two components, one doing clustering and one doing a supervised differentiable analysis. OK, thank you.