ALVIN THONG-JUAK KHO: Transforming data, as well as modeling and genomics. It's a very broad category. And I'm sure that in the other, lectures the people who are giving them have touched on several aspects of these same topics, as well.

The lecture outline is as follows. The first thing we'll do is to go through two very prototypical study designs, one of which is a two-way comparison which occurs all the time in studies. The second one is the time series, or a dosage study. The idea being that there is a one parameter in the study that moves on, progresses, and you are making measurements alongside that one parameter.

And finally, we get to the topic of data representation itself, properly. We'll talking about what it is, what it means to go from a measurement device to the spreadsheet, because in order to do analysis, we must somehow end up in the world of numbers. So how does that mapping actually happen? We'll touch on it very little.

Then we talk about the numbers themselves. Do they have a dimensionality attributed to them? Are there scales attributed to them? These things actually matter.

Because depending on whether or not it's dimensioned, certain, fundamental theorems of mathematics actually do apply in guiding you on how to formulate, say, a null hypothesis for your experiments. What would the outcome be had it been random? Things like that.

And then we talk about how, once you have the data down, you transform it so as to uncover internal or relationa structures within the data. And if we get through to this and we have time left, we will go through background concepts which crops up very often when we talk about genomics or microarray analysis, namely, things or words like noise, what is a replicate? What does it mean to be reproducible?

Do you have to normalize your data, and what is the notion of a fold in the context of, say, microarray data? I mean, a fold is a very natural measure of change when it comes to PCR. But is it that natural, after all, when you are dealing with microarray data? And finally, because by going through some miscellaneous rules of thumb.

So two prototypical study designs-- the first is a two-way comparison. You see this all the time, in fact. For instance, you want to find molecular gene protein differences in the white blood cells of type 2 diabetes human patients versus normal human patients.

For instance, you have 27 diabetes patients and 11 normal ones, and you microarray them. Actually, there's a little typo here. It should be the word "micro."

So these studies are often carried out. And there are very obvious things to keep in mind when you do these studies. Namely, before you do them, you should check for stratification, say, gender, age, things like that-- other clinical parameters that could enter later on and bedevil your subsequence analysis.

Now, a partial mathematical formulation of this problem would be, say, if you were going to microarray every person, every patient goes onto a chip, then the j-th patients, j-th diabetic patient, you could represent by this symbol, d sub j. And j normal patient, you could reproduce by this symbol.

And this quantity here is actually a vector or matrix. It's multi-factorial. We are measuring, say, 10,000 genes, or proteins, or RNA of one person. So it's a multi-dimensional quantity. And you go on to do the subsequence analysis.

So this is the first prototypical study design. The second one is equally common. It has to do with a time series or dosage study. For instance, you are observing a developing organ, and you are assaying that organ for RNA or protein at different stages of its development as it unfolds.

Again, there are reality checks that you should do before you begin, such as that system which you are assaying, how heterogeneous is it? Is it one cell type, several cell type? How is that going to affect your interpretation of the signals later on?

And again, a partial math formulation of this design would be let T sub j represent the chip data of the j-th developmental stage. Again, it is a vector. And you can get very fancy pictures like this of the expression profile as a function of time-- of one gene and another gene or RNA or protein.

Now, what actually is data representation. It's a very commonly-used term, and I don't think anyone can really agree on what it means, actually. But it could go anywhere from a mathematical formulation of a scientific problem. Or it could be mapping observations and measurements into the set of symbols.

Well, it's not just any old set of symbols. Typically, they are symbols which you can do an algebra on, say, numbers, integers, things like that. It's pointless to map it into a set of symbols where you can't actually do higher-order operations on, that being the point.

If you go on a website and type in data representation, the governmental agency for atmospheric science gives you this definition here, which you can read for yourself in this slide print out later. Another notion of data representation is something to do with database annotation and standards. You've heard of Miami, presumably, in one of the lectures before or to come. I'm not going to be talking about that at all.

And thirdly, it could be multimedia. How you present your data-- graphically, and charts, plots, et cetera, things like that. What I shall be actually concentrating on is actually the first two points. It is arguable that these two points actually have to do with data representation. If anything, the second point certainly does-- mapping observations into a set of symbols.

Now, so you make an observation, and a device takes a picture of it. But eventually, it has to get translated into the set of symbols. So it's a very obvious thing that one has to actually understand at least the basic principles of how the measurement device works if you are going to be using it in your undertaking of any biological endeavor.

You should know the relevance of the scanner setting, things like that. I'm not going to go through all of this. It's quite obvious, a lot of these things.

And you also have to be sure that the journey from an image to the actual number itself-- does it go through analytical or statistical processing software? How far removed are you from the reality of the observation? One should check for these things, in fact.

Because subsequent stuff that you're going to be doing on this numbers will matter depending on the condition upon these factors. So you get, in the case of microarrays, basically a bright spot. And somehow, you have to translate this brightness.

And the grid here is imaginary. It doesn't just-- you won't see just very discrete grids lighting up. Typically, there is a diffusion of light from one feature to another. So you have to basically translate this whole mess here down to a simplistic-looking table of numbers. So you have genes measured across different samples and their, quote, unquote, "intensity."

And next thing we'll do is to understand if these intensities themselves have a dimension associated with them, say, centimeters, Fahrenheit, things like that, or they have no dimensions at all. So dimensionality and scales. And I guess we can concentrate for a while on microarray data.

In the case of two-channel or competitive hybridization microarrays where you have two dyes, a size 3 and size 5, competing-- being hybridized onto different tissue types and put onto one array, the readout is arguably dimensionless because it has to do with a ratio. And arithmetic tells you that the ratio operation nullifies or cancels out any dimension there is in the data.

Second type species of microarrays are, say, oligonucleotide microarrays where there is no competition. You have, essentially, one tissue type. You tag it with some fluorescence, and then you hybridize it. There is no competition whatsoever.

And the registered intensities have some kind of units associated with them. If anything, you don't have to know the actual units. It may be sufficient just to if units exist or not-- dimensions or dimensionless. Why? Because different math techniques actually will apply depending on whether the quantity is a dimension quantity or dimensionless one.

For example, if you are having intense-- you are dealing with a set of numbers which have to do with radiation measurements-- clearly phosphor images or radiation-- then physics tells you that the underlying, or the most sensible distribution to this background to be studying this system would be the gamma distribution.

Secondly, there are certain, what you've probably heard as power or scaling laws, that could be useful in detecting errors in this set of numbers if they don't have a dimension or they have a dimension-- things, laws such as Zipf's law or Benford's law. For instance, the third digit of every telephone number in this country, things like that. There is a certain underlying distribution that could be very useful in informing you what actually is random and what actually isn't.

Because if you get a result after some analysis that you do, whatever it is, you would like, at least, to be sure that that result is not arising completely out of chance. And that is the whole point of this, that knowing the dimension-ness of the numbers could help you or guide you in formulating a null hypothesis for this data set. So that's dimensionality.

Why do you-- so let's say you already have the numbers in front of you. Why would you want to transform it or rewrite it in a different format? There are actually several very good reasons for doing it, and I'm sure it comes very obvious to a lot of people's heads.

Number one-- it could simplify mathematical manipulation. And secondly, rewriting it in a certain way could uncover certain structures in the data. We will see examples of these very shortly.

So number one-- simplifying mathematical manipulation. It can be argued that any spreadsheet is, essentially, a matrix. Of course, the entries of the matrices themselves are not homogeneous. They can be very different, and they could affect what kinds of operations you can sensibly perform on them. But essentially, it is a matrix if there are numbers in these entries.

And as such, if there are matrices and the entries are homogeneous, they are subject to formal and very basic linear algebraic roles. You could do a matrix addition, subtraction, et cetera. And you could investigate the eigenvalues that are eigenvectors-- basically, invariant structures within the data.

And of course, the hope in any scientific endeavor is that these invariant structures somehow are related to a physical phenomenon from which these numbers come from. There is a hope of all scientists who are trying to study a system. And this warning is obvious. If your matrix-- if you are measuring, say, temperature, and height, and weight, clearly, you can't just perform blindly linear algebraic manipulations in this data and hope to get something sensible out of it. But that's obvious.

OK, why transform data number two-- second reason. Revealing intrinsic geometries. What is meant by intrinsic geometry? It could be a very unstructured data. You may not notice upon looking at the numbers themselves in a table that there is a group of variables, which are acting a certain way. But it could turn out that if you write it or transform it, this feature would stand out. And we actually will see graphical examples of when that happens.

Let's see, so there are actually-- I'm not sure if you need to know if there are clues to the existence of these structures. But oftentimes, it's not obvious, and you could use some help by transforming the data.

Now, these internal structures may be explicit or implicit. In what sense? The explicit structures are clearly the notation you give, like gene 1, gene 2, gene 3, or condition 1, 2, and 3. These are very obvious things of clinical phenotypes.

The implicit ones are the relation between gene-- how they interact with one another, at least as captured by these numbers. Now, that, you don't know for sure. You actually have to go in and look for them. But it's just good to keep in mind that this aspect-- explicit and implicitness of these structures.

Well, the whole idea of-- it's a bit of a jump from data transformation to modeling use. This is the real-world system, and here is what exists in our heads, or at least in our computers. You're trying to understand a system-- a physical phenomenon, say. And it's going to be-- there clearly has to be some physical quantity one is measuring. And this physical quantity is subject to some kind of perturbation.

You make your measurement here. This is where the data representation occurs you map it into a set of numbers or symbols, which there's an algebra on it. And then any perturbation of this physical system will manifest itself as numeric fluctuations.

And oftentimes, when you get a lot of data, it's not clear if you're actually seeing a fluctuation or you're just seeing, quote, unquote, "noise." And then you form a model of the system. And hopefully, your model somehow will corroborate with the physical phenomenon. I mean, these are very obvious ideas of modeling.

Now, I give an example of uncovering internal structures. Now, let's say that you have two different patient populations, X and O. Say the X's are cancer patients, and the O are normal patients. And they're being controlled for age, gender, et cetera, things like that.

And for each patient, you make two gene measurements-- RNA measurements, protein, doesn't really matter. This is just an illustration. And let's call the measurements G1 and G2. OK, so each patient, you measure two of these quantities.

Now, what let's graphically represent this data. After all, we're talking about data representation, right? And let's suppose that the measurements came out this way.

Now, what is the point here that I'm trying to make? Well, if you were simply using measurement 1 and trying to use measurement 1, somehow, to discriminate between the crosses and the zeros, it would not work right. Do you see why it does not work?

Basically, you simply project the axis and the zeros onto the G1 axis, here. If you project it, you notice that there is just a alternation of crosses and X's, right? So G1 by itself certainly does not discriminate the cancer from the normal.

The same goes for quantity G2. If you project this onto the G2 axis, you don't see that G2 actually segregates the two different population samples. However, now, when you perform something called principal component analysis, which we will briefly touch on, but it is very standard in any textbook, basically, on multivariate analysis, the data is simply rotated. It's an Fn transformation. It is rotation and translation.

And now, in this new coordinates, principal component 1 and 2-- this is how it looks. It's the same picture. And what do you see? It's the same thing. You're just rotating.

But the simple act of rotation itself is highly useful. Why? Well, you notice, then, that while PC1 does not distinguish the two populations, PC2 certainly does. And the discriminating quantity is G1 minus G2. If G1 minus G2 is positive or negative, you are either cancer or you're normal.

This is the beauty of this technique. And why is this? So this is a very kindergarten example, but there is something to be learnt from here. Because as human beings, we can't visualize beyond three or four dimensions. And when you're making multiple measurements and-- so imagine that you are a person who lives in one dimension, and you are making these two measurements.

Your vision of this whole process is only somehow captured in these projections under G1 and G2. So in your one-dimensional world, or in mine, I would not have realized that G1 and G2 actually does anything. However, this transformation helps me. Because then, using this one dimension alone-- remember, I am a one-dimensional beast-- I can tell that this quantity here is a linear combination actually distinguishes one from the other.

And imagine, now, you have 10,000 variables, not just three. And the power of this method immediately comes to the fore. Another example-- say you have two different-- so each dot is a gene. And you have-- let's just suppose that you have two patients, or two persons, or two animals-- red and blue animals.

And for each animal, you measure 5,000 genes. And these 5,000 genes are measured under three conditions. So you could get a brown mouse and a white mouse-- microarray 5,000 protein levels or something-- under three different conditions, say, heat shock, starvation, exercise. And then you plot it.

I hope you're not colorblind. But let's suppose that the red and blue are not there. All you see is one color. And you did this measurement. And let's say you have a three-dimensional vision of this whole process. Look at the projections here, here, and here-- two-dimensional projections.

You don't actually see that blue and red are just together, right? Now, if you do the same principle component analysis, simply rotating the data, remember, is simply an Fn transformation. The first projection, you see nothing. But this is where the power comes.

Again, you have to somehow remove the green-- the blue and the red, and you see that clearly. The two populations reveal themselves. In a lower dimension, that's the whole point. So we are cutting down from three to two-- basically just one dimension. Clearly, one dimension actually distinguishes the two different populations.

The power is you could have 10,000 dimensions. It cuts it down all the way to the first five or 10. It depends on the system. It's just a demonstration. But this is simulated data, by the way. It's not real.

Data transformation example two-- I'm not quite sure now why I give you all this principal component analysis examples. But this is actually real data-- pancreatic development time series. And at-- I think it was 11 time points versus something like 10,000 genes, I believe. I'm not quite sure. So it's a matrix.

Now, there are two ways to look at the system. It's basically, say, 10,000 genes by 11 conditions. You can look at it as time points in a gene space-- 10,000 dimensions. That's one vision of this experiment.

The other vision is the transpose vision, which is that the objects, the graphical objects number are genes, and they live in 11-dimensional sample space or temporal space. I hope you see that. So there are two ways to look at this study. And the two ways are, when you do further analysis on them, will actually bring out different aspects of the experiment.

So you can look at the system sample-wise, meaning that dots, genes in time-space, or you can look at it gene-wise-- actually, their names could have been switched-- which is each point is a time, is a whole pancreas, and the space it's sitting on is actually genes. So it's like 10,000 dimensions.

And CLT is, basically-- Central Limit Theorem-- scaling. It's a fancy term for saying, you normalize the data to mean 0 variance 1. There are reasons that actually do it, but we come to that later.

So the vision of each dot being a gene sitting in time-space-- I'm only showing you 3 times 11, OK? Well, when you do that-- when you don't do anything at all, you just plot it, you get this cloud. It's not very informative, really. Maybe there is some information, but I didn't go deeper into see what these things did.

Now, when you do a principal component analysis of the time axis and re-plot this, what you see is a circular object. And the reason it's actually circular or looks like an egg is because of the scaling. This scaling, if you know linear algebra, actually maps everything to the unit hypersphere, which is why it's not a surprise you get an egg-shape like this.

But why is this more informative than the previous one, the previous slide? There is a reason. So I will claim, actually, that the density of the-- oh, and the first principal component captures 45% of the variance. The second principal component captures 15%.

So the idea of principal component analysis is that as you-- the first principal component captures the greatest amount of variance. The second captures the second greatest, third, et cetera. They are all orthogonal, one to the other.

So the utility of doing this is now you can, I claim, that you can actually represent 10,000 different profiles in a convenient egg-shape like this. How so? Well, the first component captures a lot of variance in the system. And so what is the profile of a gene?

If you pick any gene from here, say, and you plot its profile, how does it look like? It turns out that the profile looks like that. So in English, essentially, it is a gene of a protein which is highly-expressed early on and it goes down later on.

What about if you pick something here from the complete opposite end of PC1? That's the shape that you get. It goes the opposite direction.

In fact, if you picked samples going all the way from here to there, you will notice that there is a gradual shifting of this shape to that. It morphs one into the other. And so this is an example of something picked out of the 90 degrees, counting from 12:00. That's how it looks like.

So it's a very convenient way to display everything. And another utility is a density of dots. You notice that there is a huge absence of anything here, right? So basically, you could state-- you could claim that there is a family of profiles that is actually missing from all these genes.

No genes at all express a theoretical profile here. I think it is something which looks like that but with some variation. But there is a density here. I'm not sure how that looks like.

So it's a convenient way to display everything as opposed to what? As opposed to seeing 10,000 of these things. So it's like a dictionary.

So that was looking at genes sitting in time-space. The other transpose way of looking at the same system, the same data is to view it as samples, time points. I've labeled them 1 through-- oh, actually, it's 13-- 1 through 13 sitting in a genomic space of 10,000 dimensions.

If you just picked any three genes at random and plotted the samples-- the numbers 1 through 13 are consecutive. I didn't just randomly assign them. They're consecutive with time. So time 13 is going to be greater than time 12, greater than time 11, et cetera, et cetera.

So you've picked any two genes, and what do you notice? Well, you notice nothing. It's just a mess. You pick any three, and I'm just picking three, for example.

I mean, you'd be very lucky if we pick a set of three that actually reveals some beautiful structure in here. But then you have to wonder, what is our noise here? That goes back to understanding, what is the underlying null hypothesis of the system?

So now you perform principal component analysis. What happens? The first principal component, second, and third. The most salient thing that jumps out at you is that PC1 looks-- looks-- to be correlated with time. You get 1, 2, 3, 4, 5, 6, 7-- well, 12 is an anomaly. I don't know why it landed there-- and 13.

So I'm not sure what the other principal components mean. Maybe there is some biologic import to them. It's unclear to me. But certainly, one captures the progression of time, it looks to me.

And you are sort of immune from the possibility that this is actually due to noise. Why? Because a principal component, now, it's not just a single gene or two. It's actually a linear combination of something like 10,000.

So that's the power of the method. Had you just picked any tree randomly and found a configuration like this, you have to wonder at the randomness of this thing happening. But this is a linear combination of all of them. This is the power of this methodology.

I give you another example, but this is very kindergarten. So Fourier decomposition-- Fourier transforms-- that is another way of transforming data to reveal structures within the data. And the point of doing Fourier analysis is, you want to reveal-- the objects that you're looking for are basically frequencies.

So let's say that you have a-- is not real-world data, of course. It was completely cooked out of a machine. If you have this red sinusoid here of period 1, clearly, the frequency is just one frequency. I don't know if sequence is something 1 over 2 pi or whatnot.

When you apply Fourier transforms on it-- well, discrete, fast Fourier transform-- you will find you enter the realm of complex numbers, actually. But suffice to know that you get a point in frequency space. That's the point.

And to give you some bearing, let's take another waveform, which is twice the frequency. The frequency shouldn't be surprising that it's actually twice. It's 5, 2.5. And now you take yet another waveform which is even faster.

So the mapping-- so this is the entire waveform maps to just one point, one point, one point. So the object of interest here are frequencies. It's not localized in time. For localization, the are transformation techniques such as wavelets.

Now, of course, the world does not give you data so nicely in this uniform, band-limited, three signals like this, right? So let's suppose that the world makes it more complicated, adds up these three sinusoids. When you add up these three sinusoids, what happens? This is what you would see.

Now, let's say the world presents you with this. There are many things can do. You can actually do principal component analysis, but the question is, what are you trying to look for? If you are looking for the predominant frequencies embedded in this waveform, the most natural thing to do is Fourier analysis.

And when you do Fourier analysis, it should not shock you that the answer, when you map it to frequency space, is three dots. Happens to be the same three dots up there. So that's the beauty of it.

And there are applications, very real applications, actually, when you enter into sequential genomics. Because the alphabets, A, T, C, G can easily be mapped into 0, 1, 2, 3. Of course, the ordering-- I'm not quite sure if the ordering actually matters or the ordinality, but I don't think it does.

And if you are interested in repeating structures in the genome, it could be helpful. I'm sure people have done this, in fact. So the summary of data transformation is, basically, somebody gives you a vector x, data x.

And you simply rewrite it into a different form based on a set of basis elements that are different from the original ones. How do I say it? Typically, when somebody gives you data, there's going to be real numbers, let's say a matrix.

And the standard basis is the basis they're thinking of, that 10000100, et cetera. What you can do is actually transform it by principal component analysis, or Fourier transforms, or even wavelets.

All of these techniques are simply names to describe these basis elements. They are actual new representation forms are come out. But in these A subjects-- you can actually read about this in several textbooks that say, do the heart and pattern recognition, which I shall provide in leaflet handouts.

And clearly, not all these transformation techniques are equal. They are going to reveal for you very different things, very different internal structures in the data. That should be obvious.

And I claim that there is almost always a geometric interpretation of any given data set. Secondary users would be denoising. And feature reduction we have seen, actually, in the case of PCA. Denoising is, say, Fourier transforms, you could-- if you believe that noise are higher frequencies, you could band limit or you could model all the dots which appear much higher in that frequency space.

No, I don't think I'll talk about this. Do we actually have time for the next part? We do?

Now that we are done with data representation, I'm going to try to cover some common terms that occurs over and over again in the area of microarray analysis or genomics. You hear it all the time, but you wonder what they mean. I sometimes wonder what they mean, but this is my understanding, at least, of what may mean.

A very important thing I believe, as do a lot of people, I'm sure, is that nature makes no leaps, that physical phenomenon, at least at the microscopic level, what you observe-- microscopic atoms bumping around-- at least at a microscopic level, it cannot occur abruptly. There has to be a continuity to these processes.

And this is a very important guiding principle, at least, in definition of noise. So I'll give an example. I don't know why I called it example four. I make 100 separate measurements of the room temperature in this room in a 1-minute interval at different locations.

Depending on the accuracy of the device I'm using, it is not very likely that all these measurements are going to be the same. So the question is-- and it is an ill-posed question-- what is the temperature in this room? On average, or what is the temperature distribution in this room? Questions like this.

Now, this is my working definition of noise. And I'm sure that there are going to be-- it can be argued. In a narrow sense, noise is any measurable divergence from axiom 1-- this idea-- or more generally, any applicable axiom in a studied system.

So if you believe that-- if you are very, very sure that the room temperature-- example, if you believe that the room temperature in this room cannot be so different from me from here to there, then you make 100 measurements. And you believe that there is an idealized temperature or some belief that there should be a static quantity.

Any fluctuation, any fluctuating observations you made away from this idealized temperature-- which you don't know, anyways. The best you can do is estimate with the average-- any fluctuation is noise. So it's practical. I'm not sure if it is useful at all.

And in ideal situations, math theorems will apply-- things like the central limit theorem and law of large numbers. They are very, very robust if you have a lot of-- the question is, or the problem is that you need to have a lot of observations in order for these to kick in and help you.

Now, what is a replicate and repeated measurement? And what I'm going to say here is not going to be new to you-- how people define replicate. It, in a way, also depends on-- in fact, replicates and reproducibility go hand in hand.

I'll give you three examples of a replicate measurement. When you talk about replicate, it always involves two things. You have something, and you need something to compare it with. So let's say that you want to do replicate assays of mice pancreas-- normal mice pancreas, whole pancreas, RNA analysis. And they had been controlled for weight, for gender, et cetera.

Now, there are three different situations, or three different ways of defining replicates, right? Number one-- you take the pancreas from each of these mice-- they are, say, the same litter-- and hybridize it, et cetera. This could be a definition one of a replicate.

The other is you take the pancreas from one mouse and you split it and hybridize the other two. That's the other one. Notice, in this case, the biological variation. There is none. The call comes to one. Here, clearly, a biological variation is going to be very important.

The third one is to somehow homogenize the biological variation at the very top level. You just mix them all up and split them. And there are arguments for using this way, this way, this way-- what you want to control. It can also be argued that you can remove biological variation but only at a later stage when you have the numbers, whereas, you are pulling them here, you can pull them later on.

There is no better or worse definition of a replicate. But one should be aware, when you are reading papers, you should be aware, at least, what do they mean by replicate? Because in a way, you will also notice that replicate-- the notion of a replicate will actually guide or affect how you define noise.

Because if you believe that these two mice should give you identical reading, then any deviation of this mouse from that one is going to be noise. And that deviations-- it seems very obvious things.

Now, yeah, actually, here it is. The definition of replicate will have implication on how you define noise. And there is-- we talked about this-- biological versus measurement variation. This being biological variation, this being measurable variation. And arguably, this could be measurement variation, too.

However, if you're too restrictive in your definition of replicate, it could actually hinder the generalizability of your study. Your study is only applicable to mice with a certain genetic background, very, very restricted, et cetera. One has to consider these things.

But despite taking all precautions, it is not very likely that your, quote, unquote, "replicate assay" will, down the line, give you numerically identical results. And there is this old saying, I think from the Greek alchemist, that you never step into the same river twice. It's very true here.

And as a result, people often will try to massage the data later on after the test tube comes the numbers-- well, it has to go through certain scanners and machines. But when the numbers come out, then they massage it-- they normalize, in other words-- to somehow account for biological variation or measurement variation.

There are arguments for and against normalization. And there is no blanket principle whether to do this-- to normalize or not. It depends on your experiment design

OK, I give an example of that pancreatic development data-- embryonic day 12, 14, 18, postnatal day 2, and the adult pancreas. And I'm simply plotting that 10,000 genes against itself, here-- embryonic day 12 against itself. And E12 against its aliquot. What is aliquot?

I think, actually, aliquot, in this case, was this. It's a measurement variation. Though, in a way, it's not very wise in the case of-- clearly, you cannot remove the whole pancreas from the same mouse twice. So it may have been wiser to have done this, come to think of it, now, to account for biological variation.

Because in this case, then, it was measurement variation. Then, clearly, there are two confounding factors here. Number one is measurement variation, number two, biological variation. So it would have been nice, in a way, if we had biological variation captured here.

So E12, you see it start to spread like a comet. And as you progress with development, you can see that it spreads even worse. So it looks less and less alike. The most alike is here.

They should align up right-- you're supposed to be taking the same reading of the same thing. Of course, by the time it reaches adulthood, anything can happen.

We don't know-- what are your priority assumptions about the system guides you on whether or not to normalize. If you believe-- if you have done 10 whole microarrays for some system and you believe that your system should be remaining consistent despite the biological variation, et cetera, then you could claim that the average of these array readings have to be the same, or the variance should be a certain quantity.

But that's actually putting your own assumptions of how the system behaves into this actual phenomenon that's unfolding. So one has to be careful.

The common normalization technique that people use are the one which we have actually talked about, central limit theorem scaling, meaning that this is a vector, actually-- vector x against the reference r. So you simply subtract the average from each component of the vector x and you divide it by the standard deviation. The end result, actually, is a quantity which has mean 0 and variance 1.

Some people have reason to do this. Some people don't have a reason to do it, but they do it anyways. You just have to be very careful. So what happens when-- well we just saw what happens.

By so doing, you have mapped, actually, this entire vector into an element of the unit hypersphere, it turns out. So you do lose information. Maybe, you lose all notion of absoluteness, absolute intensities.

The second common method of normalization is you have a reference data set and you regress against that reference data set. What do I mean? Well, if you-- back to this thing.

If you take the E12 to be the reference data set and you're trying to regress everything to it or normalize everything to it, then, for example, the E12-- second aliquot against the first E12. In an ideal world, you have to get scatter that's clear. However, even through the scatter, some people believe that the linear regression line has to be of slope 1 and going through the origin.

If it isn't, make it so. So it's a linear transformation of the second aliquot by, basically, you subtract and you divide. It's just a translation of the data so that the regression of one against the other is now going through the origin and has slope 1. You do it for the entire time series, in this case, for example.

There are reasons to do it. There are reasons not to do it. What actually happens after you do it? Well, clearly then, all the samples are now going to have regression intercept 0, slope 1 against the reference.

A second thing that comes out of this is all the vectors, the newly-normalized vectors, will have the same average. You can show that for yourself. You can do the arithmetic. You'll see that come out for free.

Then, there is a notion of a fold that you hear all the time which is very natural when it comes to PCR and blots. But the question is, does a fold actually make sense in the context of Affymetrix chips? For example, you often encounter this.

You have a sample population, A, and you have three readings for that sample population A versus B, which has another three readings. And people will ask, what is the fold change from here to there? And notice that there is a negative here, too.

Yes, many ways have been proposed to solve this problem. You could take the arithmetic average here then divide it by the arithmetic average there. There's a negative there, though-- negative 1. So does the fault actually make sense in its setting?

Alternatively, some people have used the geometric average rather than the arithmetic average. Or they have logged it, somehow. But logs only-- well, posed on a set of positive numbers-- none, not even zero.

One argument against using fold actually, is that it is not stable. It is incredibly not stable and highly-sensitive to its denominator. For instance, 20 over 10, 50 over 25, and 200 over 100 are all equal to 2, right?

Let's actually perturb them by the same quantity, epsilon. This is a perturbation epsilon. Can be positive or negative.

Now, 20 over 10-- when you perturb it by epsilon, it swings all the way from 3 fold to something so small-- 1.8 folds there. But this quantity here is more stable.

So does it mean that folds actually only makes sense if the absolute numbers involved or the absolute averages involved are large, so it's much more flatter here? You can say it's more robust. So these are things that one actually should be aware of when doing analysis or when reading the papers and what they mean by folds.

Again, you recall that we have covered, actually, all of this. The study design-- prototypical study designs on data representation, background, and the last slide, essentially, is just miscellaneous. So our discussion so far, if you notice, does not require biology. It's just makes nominal reference to biology.

The approaches are very general, if you think about it. It would apply in any data set, whether or not it came from biology. But one should be aware that the math or the statistics will only provide a tool for biological discovery. And the key thing we are all here to do is to understand the biology. And because that, actually, will dictate for you the experiment design, the appropriate measure or similarity space to formulate your problem, and, also, in reading and making sense of what your model gives you after you have done this.

And a very, very important thing is, no study is ever hypothesis-free. You are going to read things into your study, no matter what. It is best to be very explicit in the beginning what your hypotheses are.

And a lot of studies, I noticed, have completely deluded themselves into thinking that they are unsupervised or hypothesis-free. There is some hypothesis, I'm sorry to say. Even principal component analysis has been claimed to be unsupervised. That's not true.

Well, it is unsupervised, in the sense that you just let things fall. But the underlying assumption is that the measure of similarity there is Euclidean distance. That matters, actually. And I end early with a quote from

An evolutionary by a biologist, a French one, from the 18th century that the discoveries that one can make with a microscope amount to very little. So you can see with your mind's eyes and without the microscope the real existence of these little beings. I think he was referring to microbes or very small features. Thank you.