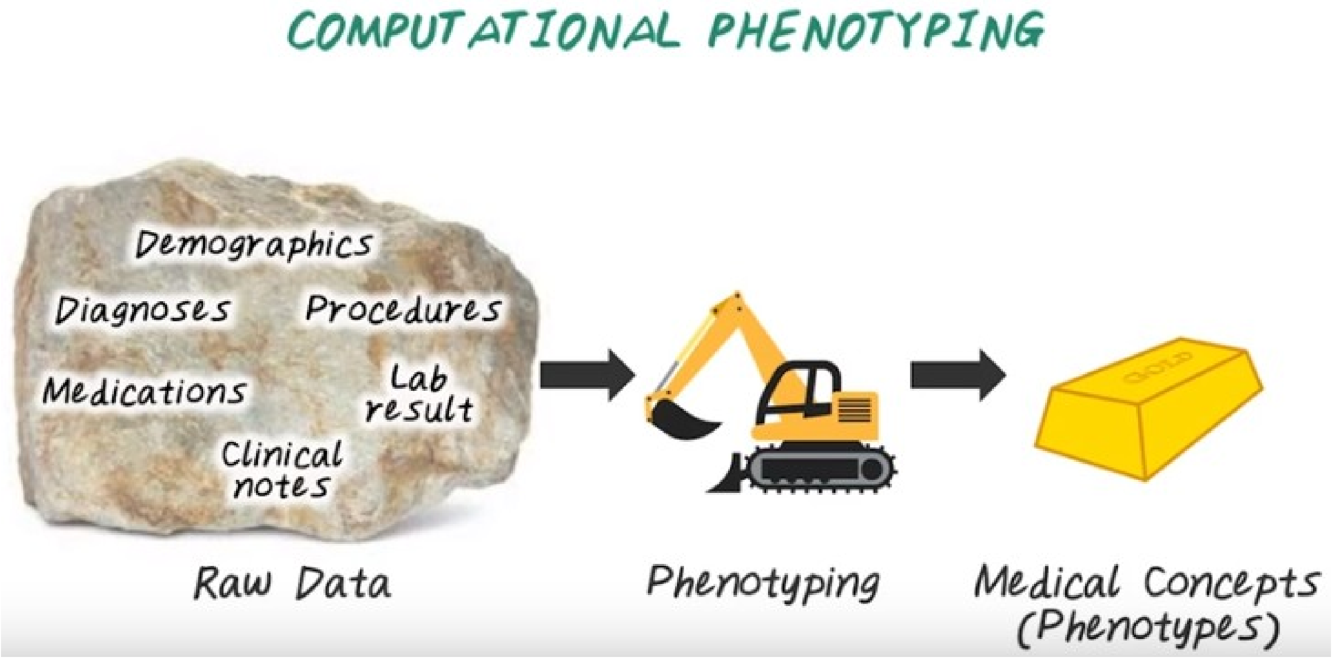
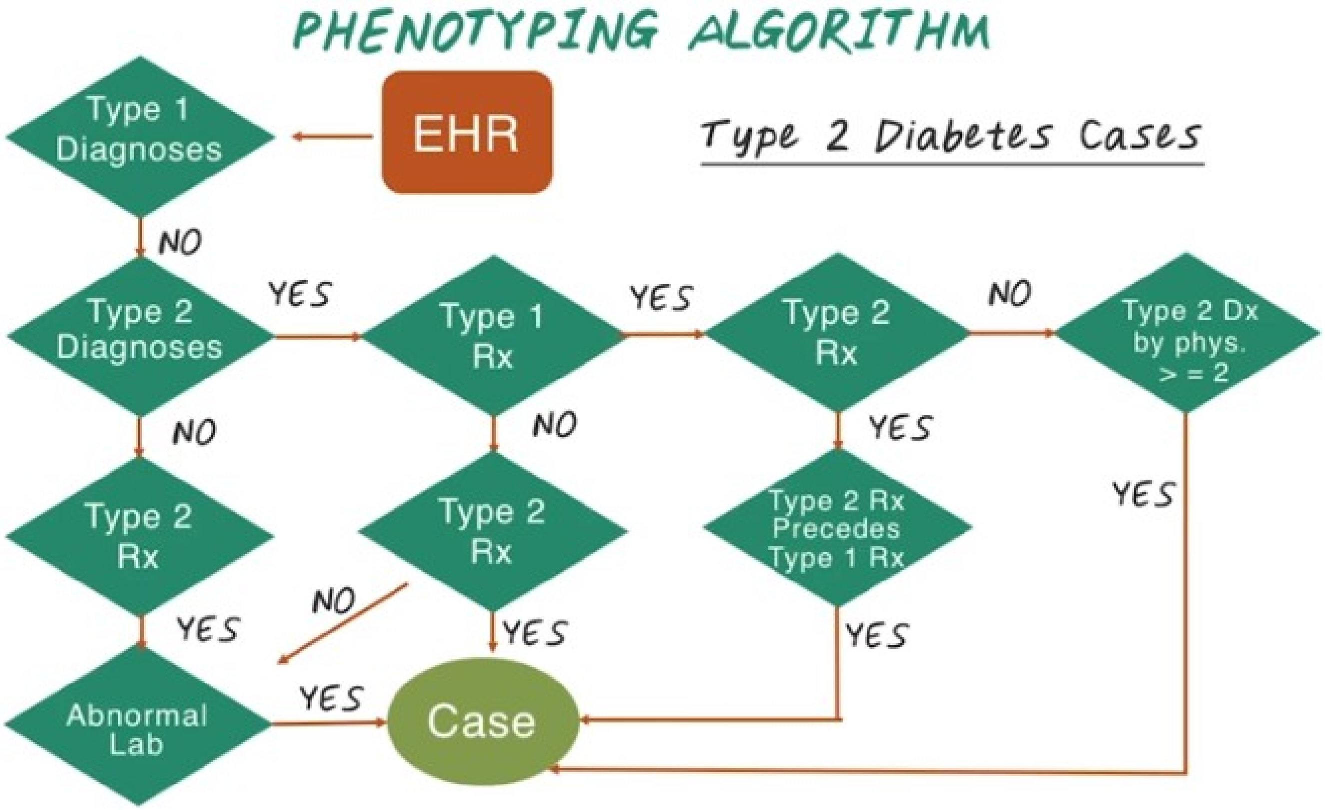
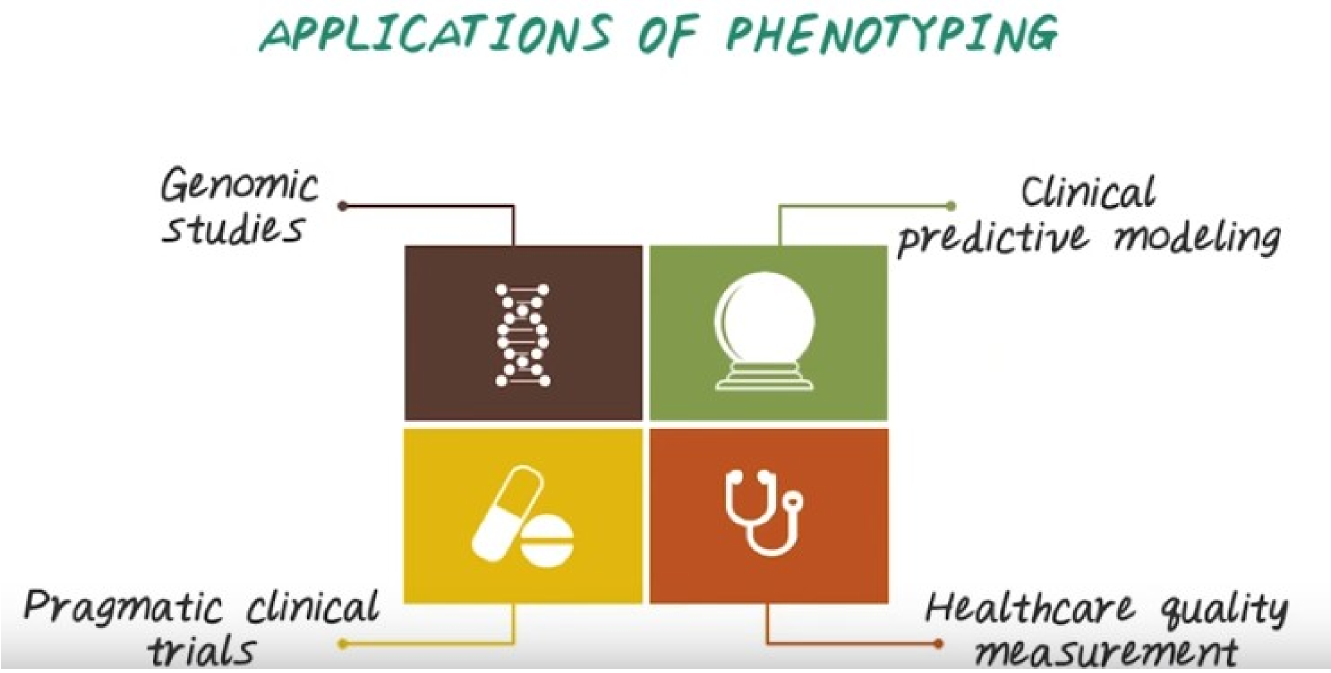
1. In this lesson, we're going to discuss a healthcare application of clustering called phenotyping. Phenotypes(表现型) are medical concepts such as diseases or conditions. We know many phenotypes of patients based on existing medical knowledge such as major diseases. But there are many more phenotypes and their subtypes out there that we haven't discovered. Computational phenotyping is a way to use data available to us to discover those novel phenotypes. Phenotypes aren't just for disease diagnosis, though. We can also use those phenotypes for predicting healthcare cost, readmission risk, and supporting genomic studies.



2. Now let's talk about computational phenotyping. So recall, computational phenotyping is about converting raw electronic health record through phenotyping algorithms into a set of meaningful medical concepts, or phenotypes. For example, a specific disease can be a phenotype. Such as type 2 diabetes. And the raw data, in this case, consists of many different sources. Such as demographics about patients, diagnosis code, medication information, clinical procedure, lab result, and clinical notes. There are many reasons why phenotypes are not represented consistently or reliably in the raw data. First, the data are noisy, there are missing datas and raw information in the raw data. And second, the main usage of this data is to support clinical operations, such as billing. And it's not designed directly for supporting research. Third, there are many overlapping and redundant information. For example, diagnosis information can be found in the structure field corresponding to diagnosis code. But the same information can also be present as end structure information in the clinical notes. This information is overlapping and redundant in the raw data. And phenotyping is this process of deriving research grade phenotypes from clinical data, using computer algorithms.

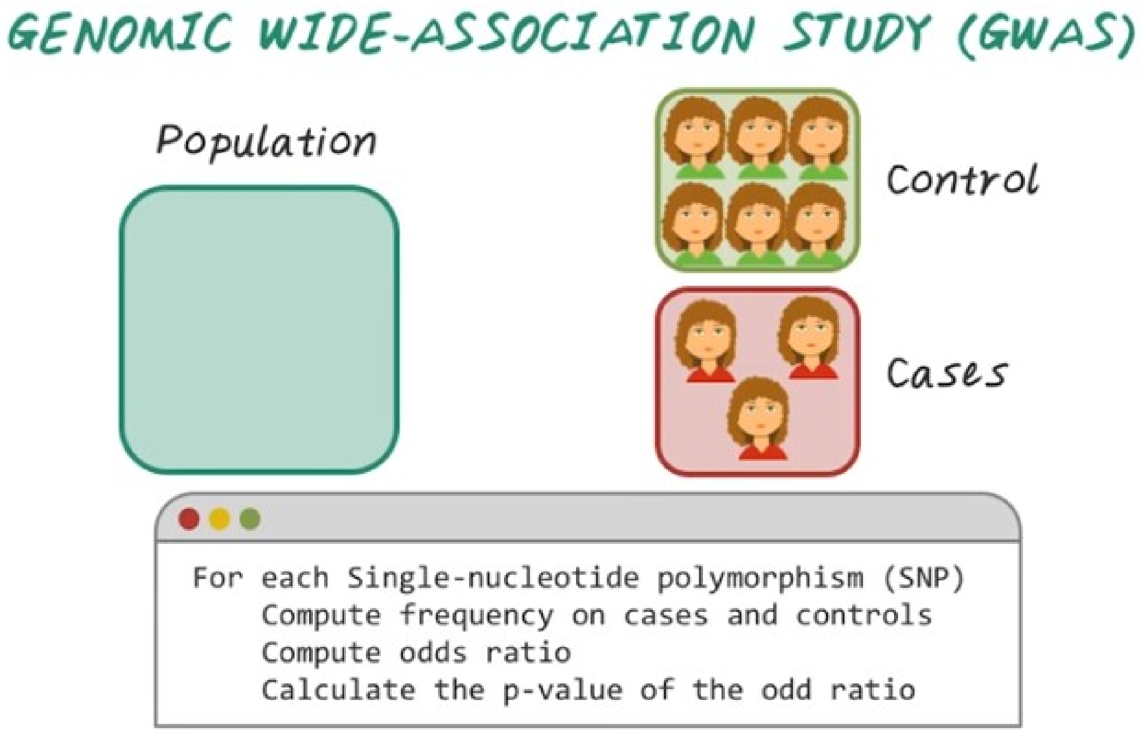


3. Here's a phenotyping algorithm for Type 2 diabetes. And the goal here is, we want to determine whether patient has Type 2 diabetes based on her electronic health records. For example, we can first check whether the patient has Type 1 diabetes code in her record. If no, then we check whether Type 2 diabetes diagnosis are present in her record. If still no, then we check medication for Type 2 diabetes. Then, check abnormal lab result related to diabetes. If these two steps are confirmed then we confirm she has Type 2 diabetes. And there are many different paths that can lead to the confirmation of Type 2 diabetes, and this decision flow is a phenotyping algorithm. And this was developed manually by clinical experts. More details about this algorithm can be found in the instructor note.


APPLICATIONS OF PHENOTYPING
Genomic studies
Clinical predictive modeling
Pragmatic clinical trials
Healthcare quality measurement

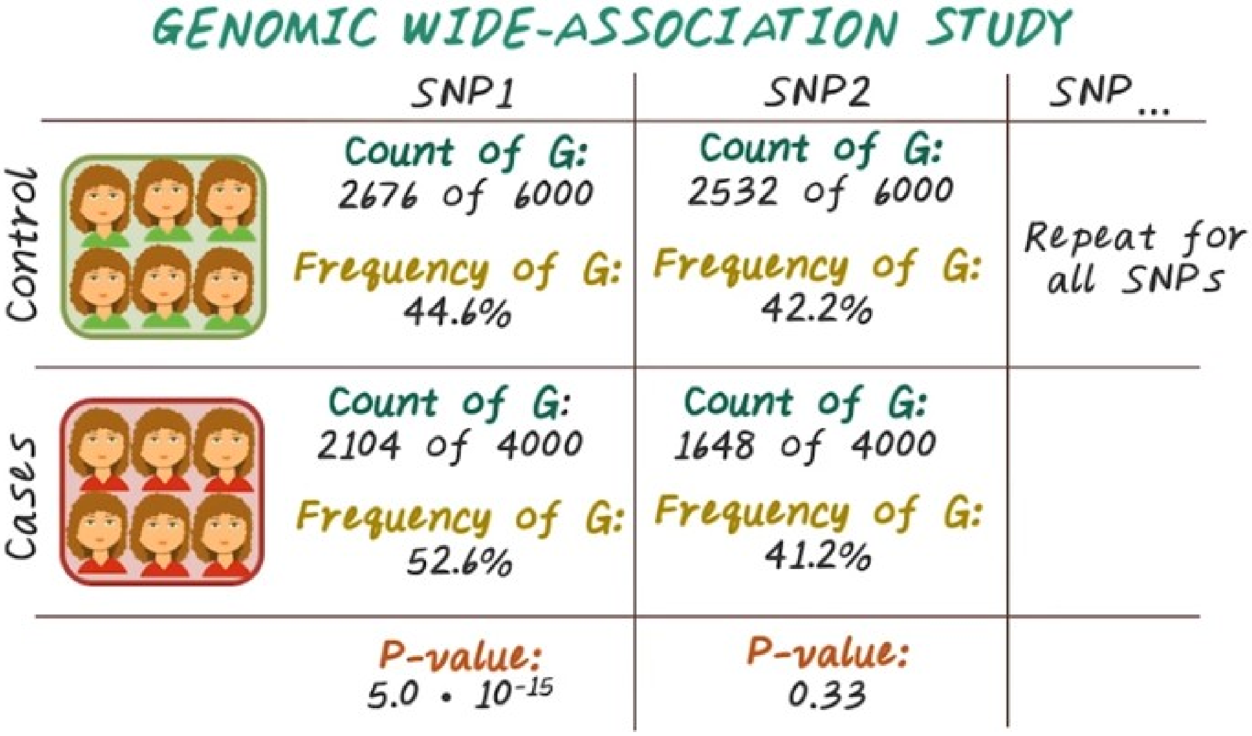
後面便是分別介紹這幾種應用

4. There are many different applications that require phenotyping. For example, genomic study, which is about finding relationship between genomic data and phenotypic data. Clinical predictive modeling, which is about building an accurate, robust, and interpretable prediction model about disease onset and other related targets, such as hospitalization. And pragmatic(注重实效的) clinical trials, which is about comparing treatment effectiveness in the real world clinical environment using observational data, like electronic health records. And healthcare quality measurement, which is about measuring efficiency and quality of care across different hospitals. All those applications depends on phenotyping algorithms. We'll show them in more details next.



```
For each Single-nucleotide polymorphism (SNP)
    Compute frequency on cases and controls
    Compute odds ratio
    Calculate the p-value of the odd ratio
```

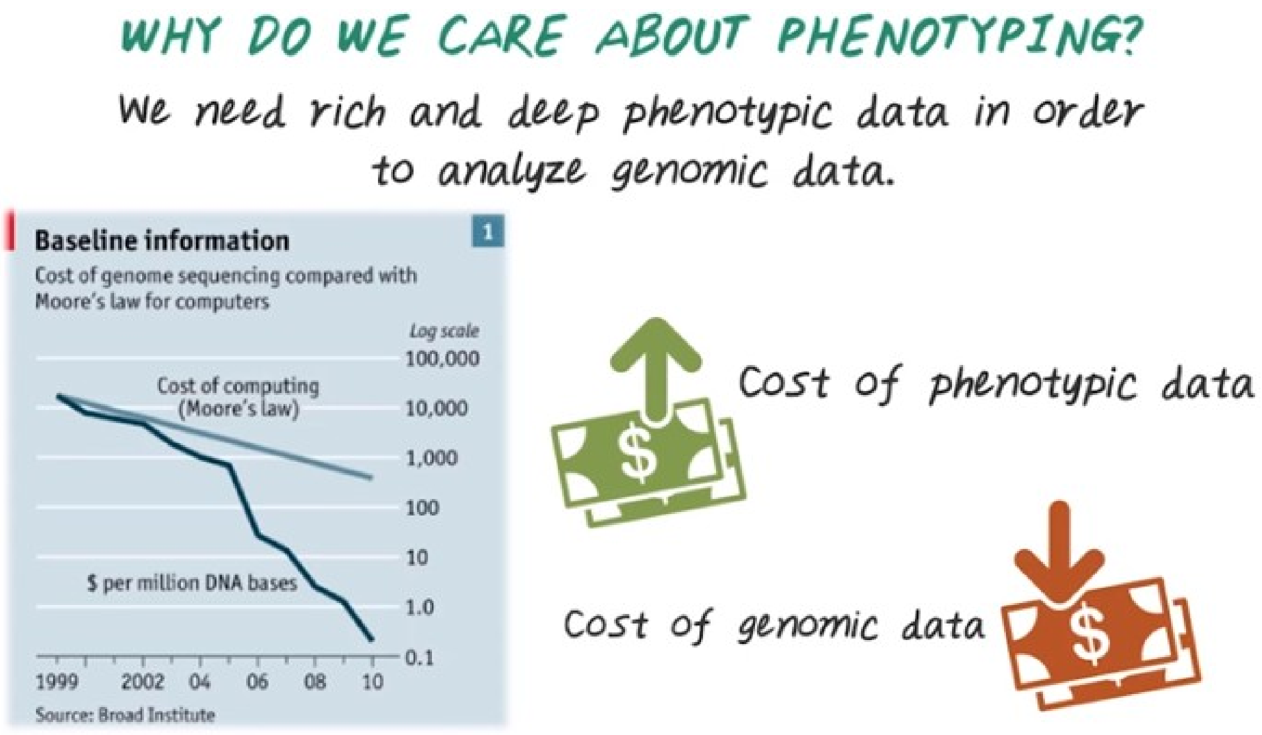
5. Phenotyping algorithm is very important in supporting genome-wide association study. What is a genome-wide association study? It's an approach that involves scanning biomarkers such as single nucleotide polymorphism (单核酸多态性), or SNP's from DNA of many people, in order to find genetic variation, associated with a particular disease field phenotypes. Once new genetic associations are identified, researchers can use that information to develop better strategies. To detect and treat and prevent diseases. So, how are genomic wide-association studies conducted? To run a genomic wide-association study, or GWAS, we first identify the disease phenotypes. Then group the participants into these two groups, cases and these are people with disease phenotypes. And controls, those are similar patient without the disease phenotype. Then we need to obtain DNA samples from all these participants. >From DNA samples, we can use lab machines to quickly survey each participant's genomes for genetic variation. Which are called single-nucleotide polymorphism or SNP. If certain genetic variation have found to be significantly more frequent in the people with the disease phenotypes compared to people without the disease phenotype. These variations are said to be associated with the disease. We do this by computing the frequencies of SNPs on the cases and on the controls. Then based on the frequency, we calculate the odds ratio. >From there, we can calculate the corresponding p-value for the odds ratio. If the p-value is small, then we conclude this variation is significant. The associated genetic variations can serve as powerful pointers to the region of the human genome that may cause the disease.



GENOMIC WIDE-ASSOCIATION STUDY

	SNP1	SNP2	SNP...
Control	Count of G: 2676 of 6000 Frequency of G: 44.6%	Count of G: 2532 of 6000 Frequency of G: 42.2%	Repeat for all SNPs
Cases	Count of G: 2104 of 4000 Frequency of G: 52.6%	Count of G: 1648 of 4000 Frequency of G: 41.2%	
	P-value: $5.0 \cdot 10^{-15}$	P-value: 0.33	

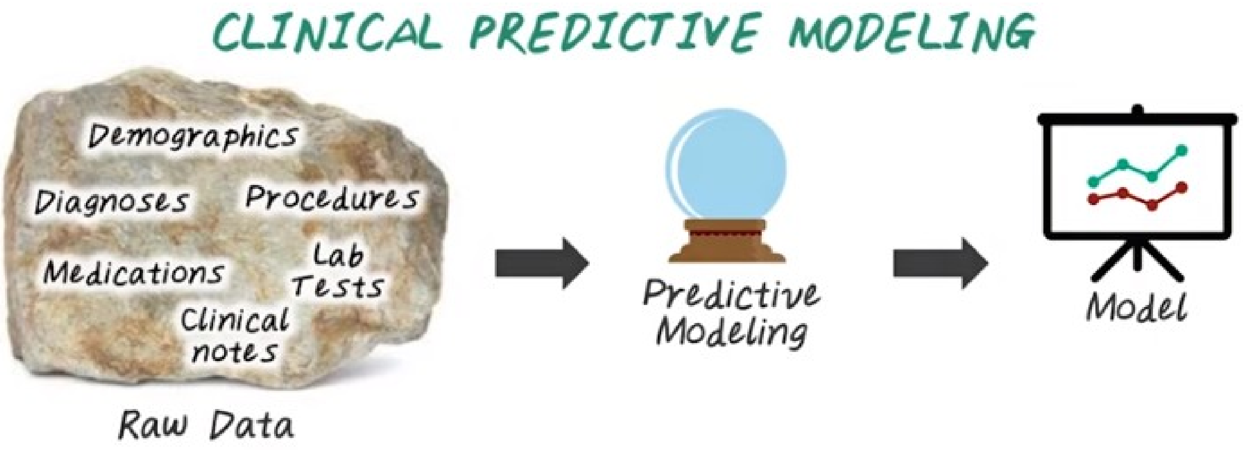
Here's an example showing more details on how the GWAS is computed. We first identified the cases and controls. That is, the people with the disease phenotype and the people without the disease phenotype. In this case, we have 4000 patients with the disease phenotype and 6000 patients without the disease phenotype. Then we iterate over all the SNPs to compare the relevant frequencies. For instance, for SNP1 for the control group, we have 2676 out of 6000 has the corresponding variation G, at this location. And the frequency of G in this case is 44.6%. in the case group, we have 2104 out of 4000 with the corresponding variation G at this location. So, the frequency is 52.6%. If we go through the calculation, we'll find now the P-value is 5 times 10 to the minus 15. Which means, this is extremely significant. We can conduct the same calculation on SNP2 and find out the P-value here is 0.33 which is not significant. And there's support GWAS study, we need to know high quality phenotypes on the cases and controls. In order to perform this calculation, that's why phenotyping algorithm, is very important.



WHY DO WE CARE ABOUT PHENOTYPING?

We need rich and deep phenotypic data in order to analyze genomic data.

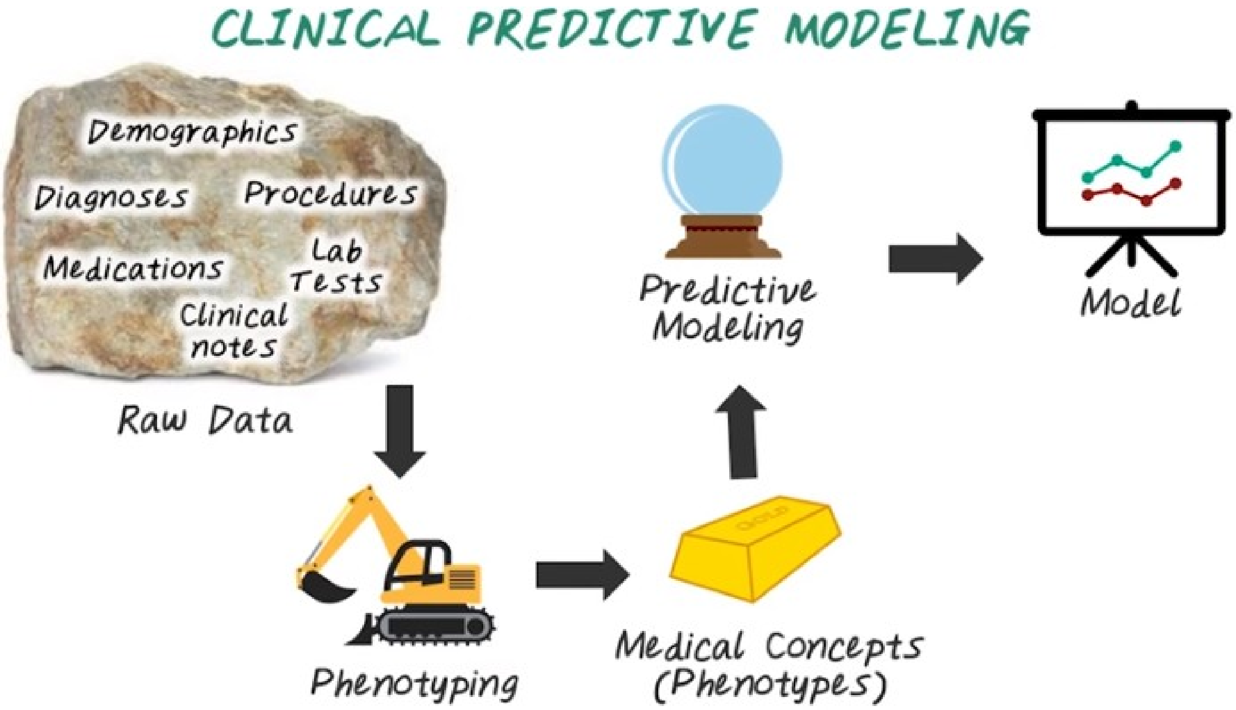
6. As we have shown in the genome-wide association studies, we need phenotypic data in order to analyze genomic data. But in general, why do we care about phenotyping algorithms in genomic study? In fact, many people argue that we need rich and deep phenotypic data in order to analyze genomic data. Especially as sequencing technology improves, the cost of generating genomic data is dropping so fast over time. While the cost of computing or Moore's Law cannot keep up with the improvement of sequencing technology, which means we'll have more and more genomic data in near future about many individuals. However, due to the complexity of the electronic health record, the cost for generating high quality phenotypic data is actually increasing, while the cost of genomic data is dropping drastically. That is why we really need to invent better phenotyping algorithms to reduce the cost of acquiring high quality phenotypic data and to support genomic studies.



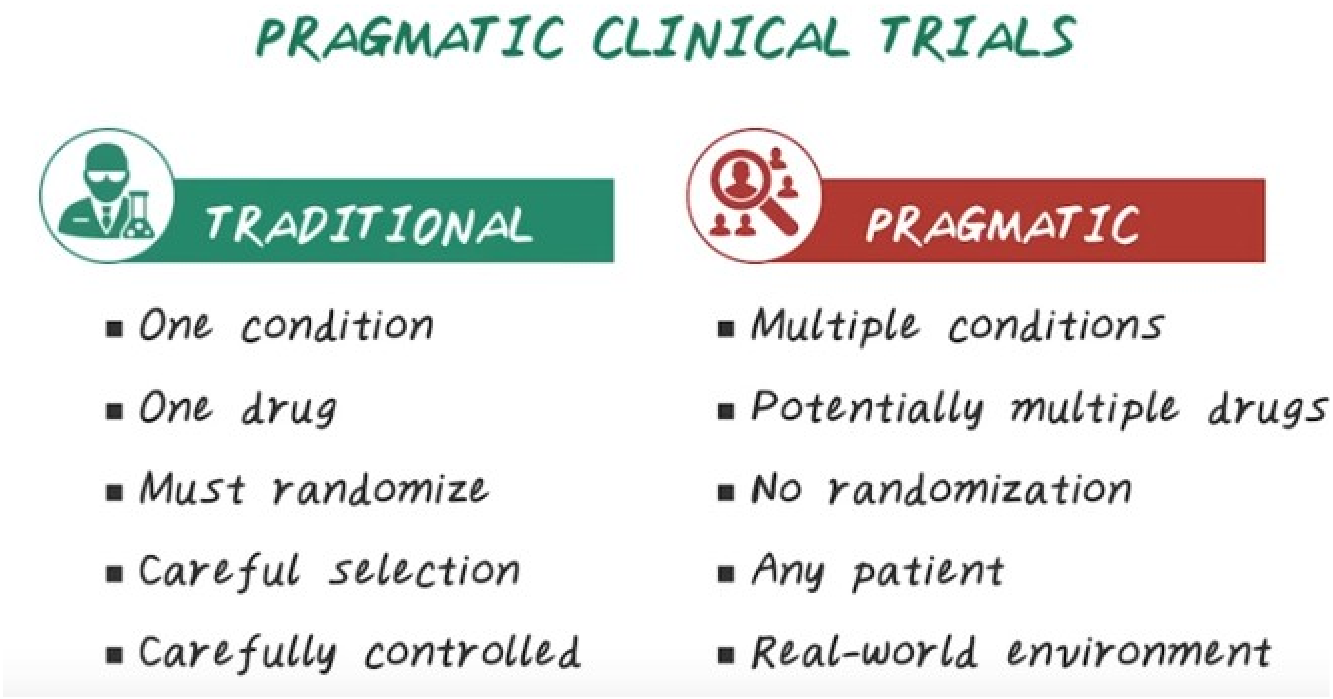
CLINICAL PREDICTIVE MODELING

上圖是 predictive modeling directly on the raw data, 是不好的.

7. Phenotyping algorithms can also help with clinical predictive modeling. We have talked about predictive modeling in other lectures. Clinical predictive modeling starts with the raw EHR data as the input, then goes through the predictive modeling phase. Then we come up with accurate predictive model, such as predicting whether a patient will develop heart failure or not in the next six months. There are many problems with predictive modeling directly on the raw data. First, as we all know, the raw data are very noisy. The resulting model may not perform as well because of the noise in the data. The raw data are also very complex and height dimensional. The resulting model may be difficult to interpret. Third, because the model is tied directly to the raw data, it can be difficult to adapt the model from one hospital to another, because their input data format can be different.



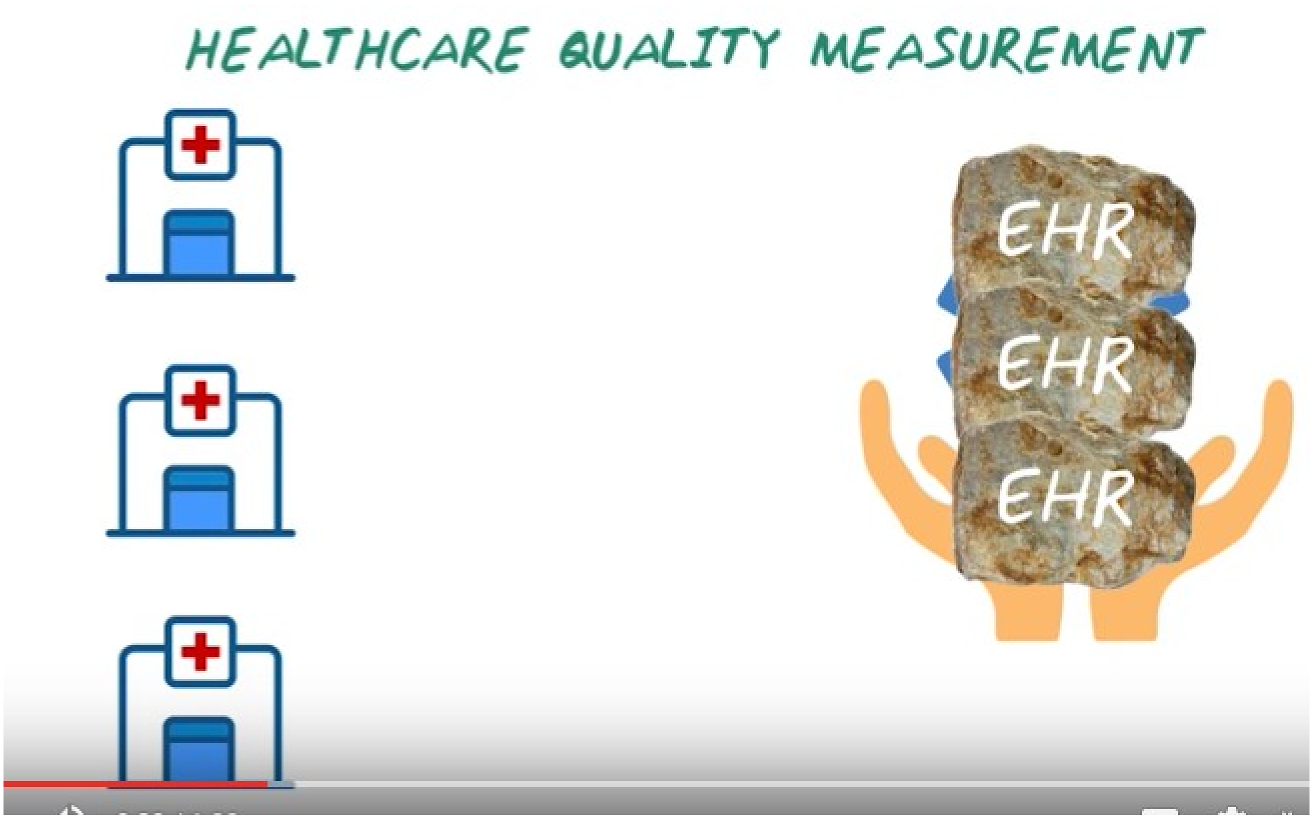
So, instead of directly modeling over the raw data, we can first convert the raw data through phenotyping to high quality, low dimensional medical concept, or phenotypes. Then use the phenotypes as input to the predictive modeling process to get the accurate model. So this way we can remove a lot of noise from the raw data, thanks to the phenotyping algorithm. We can also get better interpretival model, since the input to the model are meaningful phenotypes instead of complex raw data from EHR. The resulting model can be applied across hospitals, because the input are general phenotypes, as opposed to a specific data format from a hospital.



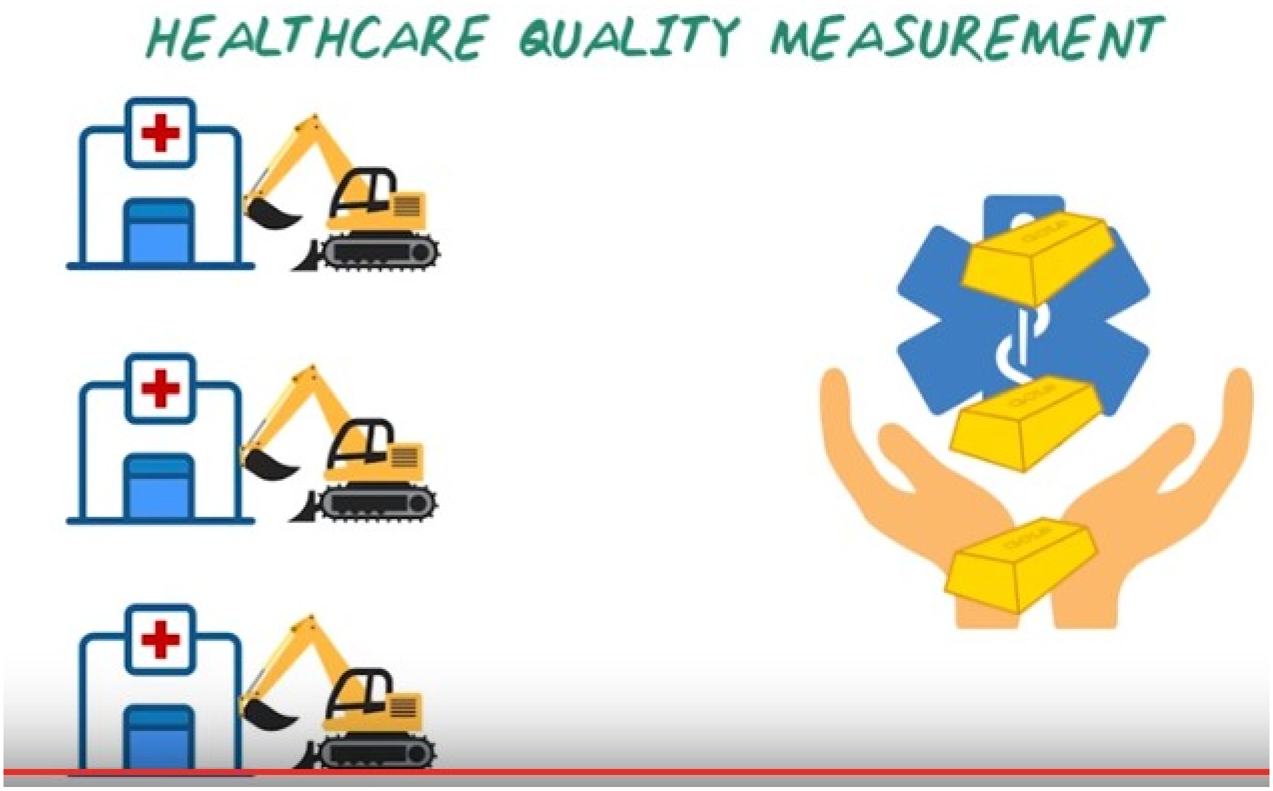
PRAGMATIC CLINICAL TRIALS

TRADITIONAL	PRAGMATIC
▪ One condition	▪ Multiple conditions
▪ One drug	▪ Potentially multiple drugs
▪ Must randomize	▪ No randomization
▪ Careful selection	▪ Any patient
▪ Carefully controlled	▪ Real-world environment

8. Another application of phenotyping algorithms is to support pragmatic clinical trials. So clinical trials can be described as either traditional trials or pragmatic trials. So traditional clinical trials generally measure efficacy, which means the benefit that treatment produces under ideal conditions. So, it deals with one condition. The pragmatic trial deals with real-world patients, often have multiple conditions simultaneously. In the traditional trials, we only test one drug at a time. In the pragmatic trials, patients potentially can take multiple drugs at the same time. In the traditional trial, randomization is required, which means some patient will be given the drug, some patient will be given a placebo. And this randomization is very important because it deal with the bias in the clinical research. However in the pragmatic trials, randomization is often not possible because it's real world environment. The other difference is traditional clinical trials recruits homogeneous population. In the traditional trials, the patients are carefully selected, often with very strict inclusion and exclusion criteria. So for pragmatic trials, there's no patient selection criterias introduced, any patient can be potentially included. So in summary, traditional trial really is designed with a very well controlled environment, while pragmatic trials deal with real world environment. High quality phenotyping algorithms are very important for pragmatic trials because we need to know what disease condition patient has and what medication they're on. Those are all can be derived as phenotypes.



9. Phenotyping algorithm is also very important for health care quality measures. It is important to compare house quality measure across hospitals. One way for doing that is to have all those hospital sending their raw EHR data to the central side. The central side can be an insurance company or public health agency such as Centers for Disease Control. Then the central side has to aggregate all those raw information in order to compute all those health care quality measure. And this become very difficult task because all those hospital can use very different format to represent their raw data and this centro side has to figure it out how to process them differently.



HEALTHCARE QUALITY MEASUREMENT

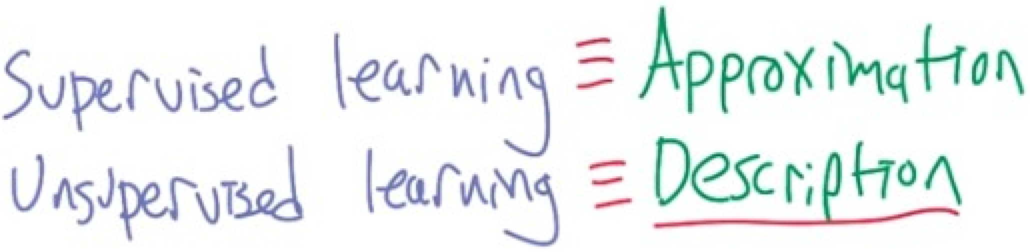
In more scalable way for dealing with this problem was to process all those raw EHR data through phenotyping first, then obtain the high quality phenotypic information then share that with the central site. With those consistent phenotypic information sending from different hospitals, now the central side can aggregate those information to compute the healthcare quality measures then compare them across hospitals. So in this case, high quality and consistent phenotypic data are crucial to enable this house care quality measure comparison across hospitals.

PHENOTYPING METHODS

SUPERVISED LEARNING

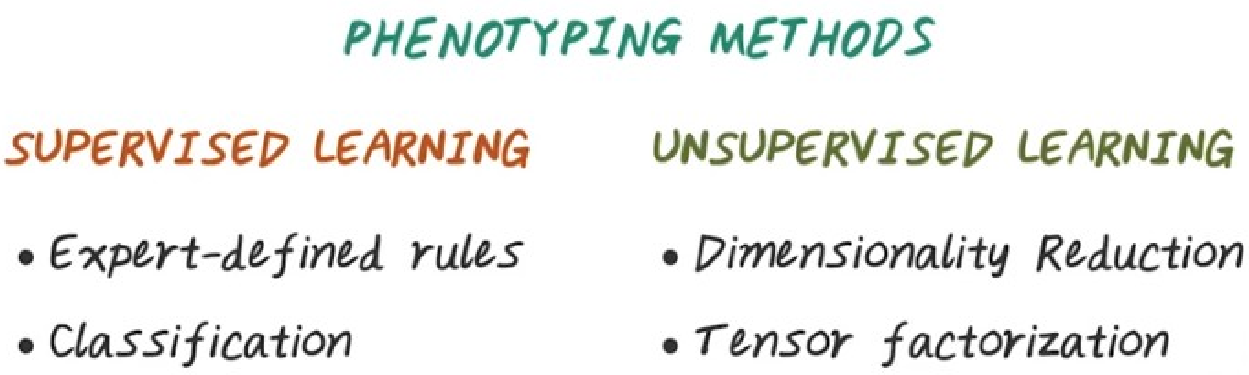
UNSUPERVISED LEARNING

10. Now understand phenotyping is a very important process, then what are the phenotyping method? There are two main categories of phenotyping method, supervised learning and unsupervised learning. They're actually corresponding to two important topics in Machine Learning. I'll have Charles and Michael to explain what they are from Machine Learning.



Supervised learning ≡ Approximation
Unsupervised learning ≡ Description

然後本段後面都是ML的視頻！>> So what do you think supervised learning is? >> I think of supervised learning as being the problem of taking labeled data sets, gleaning information from it, so that you can label new data sets. >> That's fair. I call that function approximation. So here's an example of supervised learning. I'm going to give you an input and an output. And I'm going to give them to you as pairs, and I want you to guess what the function is. >> Sure. >> Okay, okay. >> Wait, hang on, is one the input, and one the output? >> Yes. >> And 2 the input, 4 the output? >> Correct. >> Okay, I think I'm on to you. >> Nice, this is very hip data set. >> It is. What's the function? >> It's hip to be squared. >> Exactly, maybe. >> Now if you believe that's true, then tell me if the input is ten, what's the output? >> A hundred. >> And that's right, if it turns out in fact that the function is x squared. But the truth is we have no idea whether the function is x squared or not, not really. >> I have a pretty good idea. >> You do? Where does that idea come from? >> And it comes from having spoken with you over a long period of time and plus math. >> And plus math. Well, I'm going to- >> You can't say I'm wrong. >> You're wrong. >> You just said I was wrong. >> Yeah, I did. No, you've talked to me for a long time and plus math, I agree with that. >> Okay. >> But I'm going to claim that you're making a leap of faith, despite being a scientist, by deciding that the input is 10, and the output is 100. >> Sure, I would agree with that. >> What's that leap of faith? >> Well, I mean, from what you told me, it's still consistent with lots of other mappings from input to output. Like 10 gets mapped to 11. >> Right, or everything's x squared except ten. >> Sure. >> Where everything's x squared up to ten. >> Right, that would be mean. >> That would be mean. >> But it's not logically impossible. >> Or would it be the median? >> Ha. >> Thank you very much, man. I was saving that one up. What about unsupervised learning? >> Right, so unsupervised learning, we don't get those examples. We have just essentially something like input. And we have to derive some structure from them just by looking at the relationship between the inputs themselves. >> Right, so give me an example of that. >> So when you're studying different kinds of animals, say, even as a kid, you might start to say, there's these animals that all look kind of the same. They're all four-legged. I'm going to to call of them dogs, even if they happen to be horses or cows or whatever. But I have developed, without anyone telling me, this sort of notion that all these belong in the same class, and it's different from things like trees. >> Which don't have four legs. >> Well, some do, but I mean, they both bark, is all I'm saying. >> [LAUGH] Did I really set you up for that? >> Not on purpose. >> I'm sorry, I want to apologize to each and every one of you for that. But that was pretty good. >> Michael's very good at word play, which I guess is often unsupervised as well. >> No, I get a lot of supervision. [LAUGH] >> [LAUGH] You certainly get a lot of feedback. >> Yeah, that's right, please stop with that. >> So if supervised learning is about function approximation, then unsupervised learning is about description. It's about taking a set of data and figuring out how you might divide it up in one way or the other. >> Or maybe even summarization. It's not just a description, but it's a shorter description. >> Yeah, it's usually a concise, compact- >> Compression. >> Description. So I might take a bunch of pixels like I have here and might say male. >> [LAUGH] Wait, wait, wait, wait, I'm pixels now? >> As far as we can tell. >> That's fine. >> I however am not pixels. I know I'm not pixels. I'm pretty sure the rest of you are pixels. >> That's right. >> So I have a bunch of pixels and I might say male, or I might say female, or I might say dog, or I might say tree, but the point is I don't have a bunch of labels that say, dog, tree, male, or female, I just decide that pixels like this belong with pixels like this as opposed to pixels like something else that I'm pointing to behind me. >> Yeah, we're living in a world right now that is devoid of any other object. Chairs. >> Chairs, right. >> We got chairs. >> So these pixels are very different from those pixels, because of where they are relative to the other pixels. Exactly, right? So if you were- >> I'm not sure that's helping me understand unsupervised learning. >> Go outside and look at a crowd of people and try to decide how you might divide them up. Maybe you'll divide them up by ethnicity. Maybe you'll divide them up whether they have purposely shaven their hair in order to mock the bald, or whether they have curly hair. Maybe you'll divide them up by whether they have goatees- >> Facial hair. >> Or whether they have gray hair. There are lots of things that you might do in order- >> Did you just point at me and say gray hair? >> I was pointing, and your head happened to be there. >> Come on. Where's the gray hair? >> Right there. It's right where your split curl is. >> All right. >> Okay, so imagine you're dividing a world up that way. You can divide it up male and female. You can divide it up short and tall, wears hats, doesn't wear hats, all kinds of ways you can divide it up, and no one's telling you the right way to divide it up, at least not directly. That's unsupervised learning. That's description, because now, rather than having to send pixels of everyone or having to do a complete description of this crowd, you can say there were 57 males and 23 females exactly. Or there are mostly people with beards or whatever. >> I like summarization for that. >> I like summarization for that, it's a nice concise description. >> Good. >> That's unsupervised learning. >> Very good. And that's different from supervised learning. >> It's different from supervised learning, and it's different in a couple of ways. One way that it's different is all of those ways that we could have divided up the world? In some sense, they're all equally good. So I can divide it by sex, or I can divide it by height, or I can divide it by clothing or whatever, and they're all equally good absence some other signal later telling you how you should be dividing up the world. But supervised learning directly tells you, here's a signal, this is what it ought to be, and that's how you train. They're very different. >> But I can see ways that unsupervised learning could be helpful in the supervised setting. Right, so if I do get a nice description, and it's the right kind of description, it might help me do the function approximation better. >> Right, so instead of taking pixels as input and labels like male or female, I could just simply take summarization of you, like, how much hair, relative, the height to weight, and various things like that might help me do it, that's right. And by the way, in practice, this turns out to be things like tensity estimation. We do end up turning it into statics at the end of the day. Often. >> It was statics from the beginning, but when you say density estimation- >> Yes. >> Are you saying I'm stupid? >> No. >> All right, so what is density estimation? >> Well, they'll have to take the class to find out. >> I see. >> Okay.



PHENOTYPING METHODS

SUPERVISED LEARNING

- Expert-defined rules
- Classification

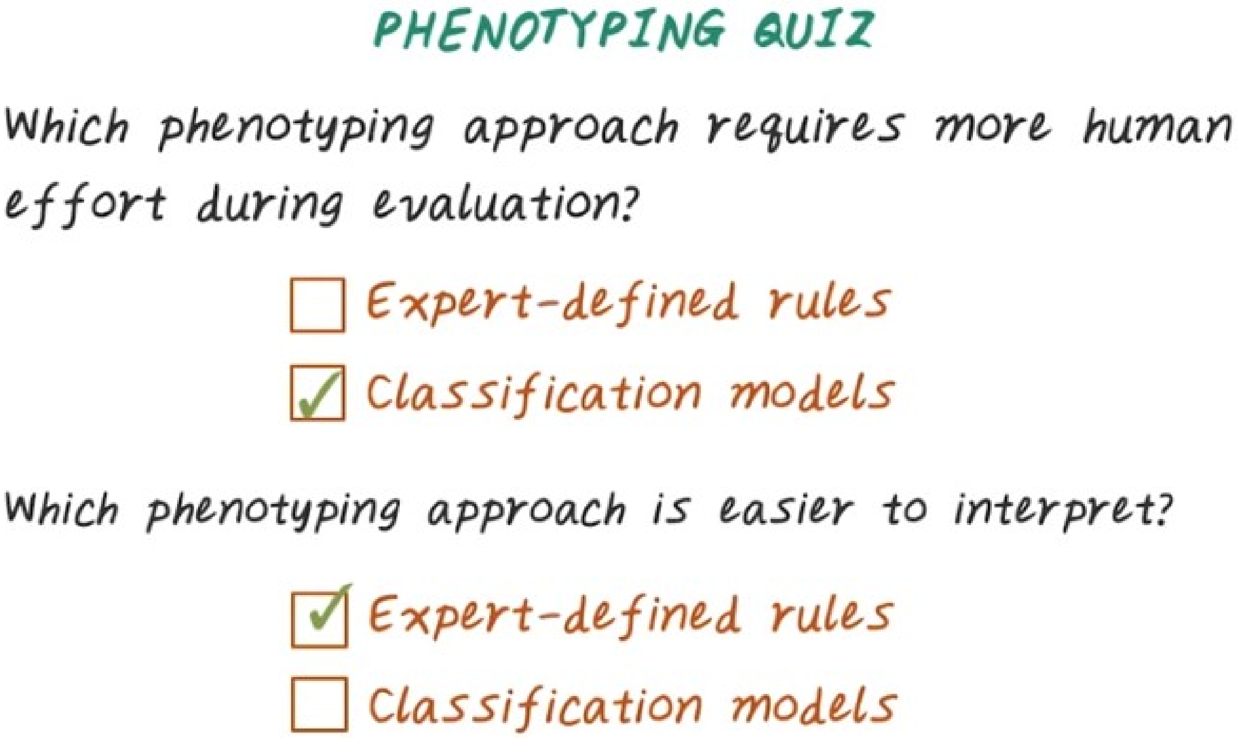
UNSUPERVISED LEARNING

- Dimensionality Reduction
- Tensor factorization

11. One way to defining Supervised Learning is to develop expert-defined rules like the ones we have seen in the early slide for type 2 diabetes. And this is probably the most widely adopted method for

phenotyping. And this approach begins with manually develop the algorithm, often use Boolean logic or scoring threshold or decision tree based on domain expertise. Then the logic is iteratively enhanced through validation and chart review on EHR data. So the advantage of this approach is, it provides a human interpretable algorithm. The number of chart review to validate this algorithm can be low because often times the expert can come up with a pretty good algorithm to start with. However, the effort and time for developing such an algorithm can be significant because it requires clinical and informatic knowledge. And this approach cannot be used to identify phenotypes that are not well understood by the clinical experts. We can also use supervised machine learning to train a classifier to differentiate the cases and the controls. Depending on the algorithm classification models sometimes can be difficult to interpret. And it requires significant amount of training data and it may not transfer well from one hospital to another. As the model may learn features that are unique to a specific hospital. Unsupervised learning provide approaches to cluster EHR data into patient groups corresponding to phenotypes or subtypes. Unsupervised learning does not require expert labels which tremendously reduce the time needed for manual chart review. However, the validation of the resulting phenotypes can be challenging because there's no ground truth on what those phenotypes are. While this method often require very large amount of training data, they do not carry the cost of manually labeling individuals as cases or controls, as what is required in the supervised method. So example of unsupervised learning for phenotyping include dimensionality reduction, such as modeling and tensor factorization. We'll talk more about them in other lectures.

12. So here a quiz of phenotyping methods. Which of the phenotyping approach require more human effort during evaluation? Is it expert-defined rules or classification models? And which phenotyping approach is easier to interpret? Is it expert-defined rules or classification models?



PHENOTYPING QUIZ

Which phenotyping approach requires more human effort during evaluation?

☐ Expert-defined rules

☑ Classification models

Which phenotyping approach is easier to interpret?

☑ Expert-defined rules

☐ Classification models

13. And the answer is classification models often require more human effort during evaluation, because they require a large amount of label training data in order to be a good model. However, the expert-defined rules, because the quality of those rules are often very good, so, during evaluation phase, it

doesn't require a lot of human effort. The expert-defined rules are easier to interpret because they're designed directly by clinical experts, follows clinical intuition and knowledge. While the classification models sometimes can be difficult to interpret, because they're derived directly from data, may or may not follow the clinical intuition.