PROFESSOR: All right, everyone, so we are very happy to have Andy Beck as our invited speaker today. Andy has a very unique background. He's trained both as a computer scientist and as a clinician. His specialty is in pathology. When he was a student at Stanford, his thesis was on how one could use machine learning algorithms to really understand a pathology data set, at the time, using more traditional regression-style approaches to understanding what the field is now called computational pathology. But his work was really at the forefront of his field.

Since then, he's come to Boston, where he was an attending and faculty at Beth Israel Deaconess Medical Center. In the recent couple of years, he's been running a company called PathAI, which is, in my opinion, one of the most exciting companies of AI in medicine. And he is my favorite invited speaker--

ANDY BECK: He says that to everyone.

PROFESSOR: --every time I get an opportunity to invite someone to speak. And I think you'll be really interested in what he has to say.

ANDY BECK: Great. Well, thank you so much. Thanks for having me. Yeah, I'm really excited to talk in this course. It is a super exciting time for machine learning in pathology And if you have any questions throughout, please feel free to ask.

And so for some background on what pathology is-- it's so like, if you're a patient. You go to the doctor, and AI could apply in any aspect of this whole trajectory, and I'll kind of talk about specifically in pathology.

So you go to the doctor. They take a bunch of data from you. You talk to them. They get signs and symptoms. Typically, if they're at all concerned, and it could be something that's a structural alteration that's not accessible just through taking blood work, say, like a cancer, which is one of the biggest things, they'll send you to radiology where they want to-- the radiology is the best way for acquiring data to look for big structural changes.

So you can't see single cells in radiology. But you can see inside the body and see some large things that are changing to make evaluations for, like, you have a cough, like are you looking at lung cancer, or are you looking at pneumonia?

And radiology only takes you so far. And people are super excited about applying AI to

radiology, but I think one thing they often forget is these images are not very data-rich compared to the core data types. I mean, this is my bias from pathology, but radiology gets you some part of the way, where you can sort of triage normal stuff.

And the radiologist will have some impression of what they're looking at. And often, that's the bottom line in the radiology report is impression-- concerning for cancer, or impression-- likely benign but not sure, or impression-- totally benign. And that will also guide subsequent decisions. But if there's some concern that something serious is going on, the patient undergoes a pretty serious procedure, which is a tissue biopsy.

So pathology requires tissue to do what I'm going to talk about, which is surgical pathology that requires tissue specimen. There's also blood-based things. But then this is the diagnosis where you're trying to say is this cancer? Is this not cancer? And that report by itself can really guide subsequent decisions, which could be no further treatment or a big surgery or a big decision about chemotherapy and radiotherapy.

So this is one area where you really want to incorporate data in the most effective way to reduce errors, to increase standardization, and to really inform the best treatment decision for each patient based on the characteristics of their disease.

And the one thing about pathology that's pretty interesting is it's super visual. And this is just a kind of random sampling of some of the types of different imagery that pathologists are looking at every day. I think this is one thing that draws people to this specialty is a saying in radiology, you're sort of looking at an impression of what might be happening based on sending different types of images and acquiring the data and sort of trying to estimate what's going on.

Whereas here, you're actually staining pieces of tissue and looking by eye at actual individual cells. You can look within cells. You can look at how populations of cells are being organized. And for many diseases, this still represents sort of the core data type that defines what's going on, and is this something with a serious prognosis that requires, say, surgery? Or is this something that's totally benign? All of these are different aspects of benign processes.

And so just the normal human body creates all these different patterns. And then there's a lot of patterns of disease. And these are all different subtypes of disease that are all different morphologies. So there's sort of an incredible wealth of different visual imagery that the pathologist has to incorporate into their diagnosis.

And then there's, on top of that, things like special stains that can stain for specific organisms, for infectious disease, or specific patterns of protein expression, for subtyping disease based on expression of drug targets. And this even more sort of increases the complexity of the work.

So for many years, there's really nothing new about trying to apply AI or machine learning or computation to this field. It's actually a very natural field, because it's sort of laboratory-based. It's all about data processing. You take this input, things like images, and produces output, what a diagnosis is.

So people have really been trying this for 40 years or so now. This is one of the very first studies that sort of just tried to see, could we train a computer to identify the size of cancer cells through a process they called morphometry, here on the bottom? And then could we just use sort of measurements about the size of cancer cells in a very simple model to predict outcome?

And in this study, they have a learning set that they're learning from and then a test set. And they show that their system, as every paper that ever gets published shows, does better than the two competing approaches. Although even in this best case scenario, there's significant degradation from learning to test.

So one, it's super simple. It's using very simple methods, and the data sets are tiny, 38 learning cases, 40 test cases. And this is published in *The Lancet,* which is the leading biomedical journal even today.

And then people got excited about AI sort of building off of simple approaches. And back in 1990, it was thought artificial neural nets would be super useful for quantitative pathology for sort of obvious reasons. But at that time, there was really no way of digitizing stuff at any sort of scale, and that problem's only recently been solved.

But sort of in 2000, people were first thinking about once the slides are digital, then you could apply computational methods effectively. But kind of nothing really changed, and still, to a large degree, hasn't changed for the predominance of pathology, which I'll talk about.

But as was mentioned earlier, I was part of one of the first studies to really take a more machine learning approach to this. And what we mean by machine learning versus prior approaches is the idea of using data-driven analysis to figure out the best features.

And now you can do that in an even more explicit way with machine learning, but there's sort of a progression from measuring one or two things in a very tedious way on very small data sets to, I'd say, this way, where we're using some traditional regression-based machine learning to measure larger numbers of features. And then using things like those associations, those features with patient outcome to focus your analyses on the most important ones.

And the challenging machine learning task here and really one of the core tasks in pathology is image processing. So how do we train computers to sort of have the knowledge of what is being looked at that any pathologist would want to have? And there's a few basic things you'd want to train the computer to do, which is, for example, identify where's the cancer? Where's the stroma? Where are the cancer cells? Where are the fibroblasts, et cetera?

And then once you train a machine learning based system to identify those things, you can then extract lots of quantitative phenotypes out of the images. And this is all using human-engineered features to measure all the different characteristics of what's going on in an image. And machine learning is being used here to create those features. And then we use other regression-based methods to associate these features with things like clinical outcome.

And in this work, we show that by taking a data-driven approach, sort of, you begin to focus on things like what's happening in the tumor microenvironment, not just in the tumor itself? And it sort of turned out, over the past decade, that understanding the way the tumor interacts with the tumor microenvironment is sort of one of the most important things to do in cancer with things like fields like immunooncology being one of the biggest advances in the therapy of cancer, where you're essentially just regulating how tumor cells interact with the cells around them.

And that sort of data is entirely inaccessible using traditional pathology approaches and really required a machine learning approach to extract a bunch of features and sort of let the data speak for itself in terms of which of those features is most important for survival.

And in this study, we showed that these things are associated with survival. I don't know if you guys do a lot of Kaplan-Meier plots in here.

PROFESSOR: They saw it once, but taking us through it slowly is never a bad idea.

ANDY BECK: Yeah, so these are-- I feel there's one type of plot to know for most of biomedical research, and it's probably this one. And it's extremely simple. So it's really just an empirical distribution

of how patients are doing over time.

So the x-axis is time. And here, the goal is to build a prognostic model. I wish I had a predictive one in here, but we can talk about what that would look like. But a prognostic model, any sort of prognostic test in any disease in medicine is to try to create subgroups that show different survival outcomes.

And then by implication, they may benefit from different therapies. They may not. That doesn't answer that question, but it just tells you if you want to make an estimate for how a patient's going to be doing in five years, and you can sub-classify them into two groups, this is a way to visualize it. You don't need two groups. You could do this with even one group, but it's frequently used to show differences between two groups.

So you'll see here, there's a black line and a red line. And these are groups of patients where a model trained not on these cases was trained to separate high-risk patients from low-risk patients.

And the way we did that was we did logistic regression on a different data set, sort of trying to classify patients alive at five years following diagnosis versus patients deceased, five years diagnosis.

We build a model. We fix the model. Then we apply it to this data set of about 250 cases. And then we just ask, did we actually effectively create two different groups of patients whose survival distribution is significantly different?

So what this p-value is telling you is the probability that these two curves come from the same underlying distribution or that there's no difference between these two curves across all of the time points. And what we see here is there seems to be a difference between the black line versus the red line, where, say, 10 years, the probability of survival is about 80% in the low-risk group and more like 60% in the high-risk group.

And overall, the p-value's very small for there being a difference between those two curves. So that's sort of like what a successful type Kaplan-Meier plot would look like if you're trying to create a model that separates patients into groups with different survival distributions

And then it's always important for these types of things to try them on multiple data sets. And here we show the same model applied to a different data set showed pretty similar overall

effectiveness at stratifying patients into two groups.

So why do you think doing this might be useful? I guess, yeah, anyone? Because there's actually, I think this type of curve is often confused with one that actually is extremely useful, which I would say-- yeah?

PROFESSOR: Why don't you wait?

ANDY BECK: Sure.

PROFESSOR: Don't be shy. You can call them.

ANDY BECK: All right.

AUDIENCE: Probably you can you use this to start off when the patient's of high-risk and probably at five years, if the patient has high-risk, probably do a follow-up.

ANDY BECK: Right, exactly. Yeah, yeah. So that would be a great use.

PROFESSOR: Can you repeat the question for the recording?

ANDY BECK: So it was saying like if you know someone's at a high risk of having an event prior to five years, an event is when the curve goes down. So definitely, the red group is at 40, almost double or something the risk of the black group.

So if you have certain interventions you can do to help prevent these things, such as giving an additional treatment or giving more frequent monitoring for recurrence. Like if you can do a follow-up scan in a month versus six months, you could make that decision in a data-driven way by knowing whether the patient's on the red curve or the black curve.

So yeah, exactly right. It helps you to make therapeutic decisions when there's a bunch of things you can do, either give more aggressive treatment or do more aggressive monitoring of disease, depending on is it aggressive disease or a non-aggressive disease.

The other type of curve that I think often gets confused with these that's quite useful is one that directly tests that intervention. So essentially, you could do a trial of the usefulness, the clinical utility of this algorithm, where on the one hand, you make the prediction on everyone and don't do anything differently. And then the other one is you make a prediction on the patients, and you actually use it to make a decision, like more frequent treatment or more

frequent intervention.

And then you could do a curve, saying among the high-risk patients, where we actually acted on it, that's black. And if we didn't act on it, it's red. And then, if you do the experiment in the right way, you can make the inference that you're actually preventing death by 50% if the intervention is causing black versus red.

Here, we're not doing anything with causality. We're just sort of observing how patients do differently over time. But frequently, you see these as the figure, the key figure for a randomized control trial, where the only thing different between the groups of patients is the intervention. And that really lets you make a powerful inference that changes what care should be.

This one, you're just like, OK, maybe we should do something differently, but not really sure, but it makes intuitive sense. But if you actually have something from a randomized clinical trial or something else that allows you to infer causality, this is the most important figure.

And you can actually infer how many lives are being saved or things by doing something. But this one's not about intervention. It's just about sort of observing how patients do over time.

So that was some of the work from eight years ago, and none of this has really changed in practice. Everyone is still using glass slides and microscopes in the clinic. Research is a totally different story. But still, 99% of clinic is using these old-fashioned technologies-- microscopes from technology breakthroughs in the mid-1800s, staining breakthroughs in the late 1800s. The H and E stain is the key stain.

So aspects of pathology haven't moved forward at all, and this has pretty significant consequences. And here's just a couple of types of figures that really allow you to see the primary data for what a problem interobserver variability really is in clinical practice. And this is just another, I think, really nice, empirical way of viewing raw data, where there is a ground truth consensus of experts, who sort of decided what all these 70 or so cases were, through experts always knowing the right answer.

And for all of these 70, called them all the category of atypia, which here is indicated in yellow. And then they took all of these 70 cases that the experts that are atypia and sent them to hundreds of pathologists across the country and for each one, just plotted the distribution of different diagnoses they were receiving.

And quite strikingly-- and this was published in *JAMA,* a great journal, about four years ago now-- they show this incredible distribution of different diagnoses among each case. So this is really why you might want a computational approach is there should be the same color. This should just be one big color or maybe a few outliers, but for almost any case, there's a significant proportion of people calling it normal, which is yellow-- or sorry, tan, then atypical, which is yellow, and then actually cancer, which is orange or red.

PROFESSOR: What does atypical mean?

ANDY BECK: Yeah, so atypical is this border area between totally normal and cancer, where the pathologist is saying it's-- which is actually the most important diagnosis because totally normal you do nothing. Cancer-- there's well-described protocols for what to do.

Atypia, they often overtreat. And that's sort of the bias in medicine is always assume the worst when you get a certain diagnosis back. So atypia has nuclear features of cancer but doesn't fully. You know, maybe you get 7 of the 10 criteria or three of the five criteria. And it has to do with sort of nuclei looking a little bigger and a little weirder than expected but not enough where the pathologist feels comfortable calling it cancer.

And that's part of the reason that that shows almost a coin flip. Of the ones the experts called atypia, only 48% was agreed with in the community.

The other interesting thing the study showed was intraobserver variability is just as big of an issue as interobserver. So a person disagrees with themselves after an eight month washout period pretty much as often as they disagree with others. So another reason why computational approaches would be valuable and why this really is a problem. And this is in breast biopsies.

The same research group showed quite similar results. This was in *British Medical Journal* in skin biopsies, which is another super important area, where, again they have the same type of visualization of data. They have five different classes of severity of skin lesions, ranging from a totally normal benign nevus, like I'm sure many of us have on our skin to a melanoma, which is a serious, malignant cancer that needs to be treated as soon as possible.

And here, the white color is totally benign. The darker blue color is melanoma. And again, they show lots of discordance, pretty much as bad as in the breast biopsies. And here again, the intraobserver variability with an eight-month washout period was about 33%. So people

disagree with themselves one out of three times.

And then these aren't totally outlier cases or one research group. The College of American Pathologists did a big summary of 116 studies and showed overall, an 18.3% median discrepancy rate across all the studies and a 6% major discrepancy rate, which would be a major clinical decision is the wrong one, like surgery, no surgery, et cetera. And those sort of in the ballpark agree with the previously published findings.

So a lot of reasons to be pessimistic but one reason to be very optimistic is the one area where AI is not-- not the one area, but maybe one of two or three areas where AI is not total hype is vision.

Vision really started working well as, I don't if you've covered in this class but with deep convolutional neural nets in 2012. And then all the groups sort of just kept getting incrementally better year over year. And now this is an old graph from 2015, but there's been a huge development of methods even since 2015, where now I think we really understand the strengths and the weaknesses of these approaches.

And pathology sort of has a lot of the strengths, which is super well-defined, very focused questions. And I think there's lots of failures whenever you try to do anything more general. But for the types of tasks where you know exactly what you're looking for and you can generate the training data, these systems can work really well.

So that's a lot of what we're focused on at PathAI is how do we extract the most information out of pathology images really doing two things. One is understanding what's inside the images and the second is using deep learning to sort of directly try to infer patient level phenotypes and outcomes directly from the images.

And we use both traditional machine learning models for certain things, like particularly making inference at the patient level, where n is often very small. But anything that's directly operating on the image is almost some variant always of deep convolutional neural nets, which really are the state of the art for image processing.

And we sort of, a lot of what we think about at PathAI, and I think what's really important in this area of ML for medicine is generating the right data set and then using things like deep learning to optimize all of the features in a data-driven away, and then really thinking about how to use the outputs of these models intelligently and really validate them in a robust way,

because there's many ways to be fooled by artefacts and other things.

So just some of the-- not to belabor the points, but why these approaches are really valuable in this application is it allows you to exhaustively analyze slides. So a pathologist, the reason they're making so many errors is they're just kind of overwhelmed. I mean, there's two reasons.

One is humans aren't good at interpreting visual patterns. Actually, I think that's not the real reason, because humans are pretty darn good at that. And there are difficult things where we can disagree, but when people focus on small images, frequently they agree. But these images are enormous, and humans just don't have enough time to study carefully every cell on every slide. Whereas, the computer, in a real way, can be forced to exhaustively analyze every cell on every slide, and that's just a huge difference.

It's quantitative. I mean, this is one thing the computer is definitely better at. It can compute huge numerators, huge denominators, and exactly compute proportions. Whereas, when a person is looking at a slide, they're really just eyeballing some percentage based on a very small amount of data.

It's super efficient. So you can analyze-- this whole process is massively paralyzable, so you can almost do a slide as fast as you want based on how much you're willing to spend on it.

And it allows you not only do all of of these, sort of, automation tasks exhaustively, quantitatively, and efficiently but also discover a lot of new insights from the data, which I think we did in a very early way, back eight years ago, when we sort of had human-extracted features correlate those with outcome. But now you can really supervise the whole process with machine learning of how you go from the components of an image to patient outcomes and learn new biology that you didn't know going in.

And everyone's always like, well, are you just going to replace pathologists? And I really don't think this is, in any way, the future. In almost every field that's sort of like where automation is becoming very common, the demand for people who are experts in that area is increasing.

And like airplane pilots is one I was just learning about today. They just do a completely different thing than they did 20 years ago, and now it's all about mission control of this big system and understanding all the flight management systems and understanding all the data they're getting. And I think the job has not gotten necessarily simpler, but they're much more

effective, and they're doing much different types of work.

And I do think the pathologist is going to move from sort of staring into a microscope with a literally very myopic focus on very small things to being more of a consultant with physicians, integrating lots of different types of data, things that AI is really bad at, a lot of reasoning about specific instances, and then providing that guidance to physicians. So I think the job will look a lot different, but we never really needed more diagnosticians in the future than in the past.

So one example, I think we sent out a reading about this was this concept of breast cancer metastasis is a good use case of machine learning. And this is just a patient example. So a primary mass is discovered.

So one of the big determinants of the prognosis from a primary tumor is has it spread to the lymph nodes? Because that's one of the first areas that tumors metastasize to. And the way to diagnose whether tumors have metastasized to lymph nodes is to take a biopsy and then evaluate those for the presence of cancer where it shouldn't be.

And this is a task that's very quantitative and very tedious. So the International Symposium on Biomedical Imaging organized this challenge called the Chameleon 16 Challenge, where they put together almost 300 training slides and about 130 test slides. And they asked a bunch of teams to build machine learning based systems to automate the evaluation of the test slides, both to diagnose whether the slide contained cancer or not, as well as to actually identify where in the slides the cancer was located.

And kind of the big machine learning challenge here, why you can't just throw it into a off-the-shelf or on the web image classification tool is the images are so large that it's just not feasible to throw the whole image into any kind of neural net. Because they can be between 20,000 and 200,000 pixels on a side. So they have millions of pixels.

And for that, we do this process where we start with a labeled data set, where there are these very large regions labeled either as normal or tumor. And then we build procedures, which is actually a key component of getting machine learning to work well, of sampling patches of images and putting those patches into the model.

And this sampling procedure is actually incredibly important for controlling the behavior of the system, because you could sample in all different ways. You're never going to sample exhaustively just because there's far too many possible patches. So thinking about the right

examples to show the system has an enormous effect on both the performance and the generalizability of the systems you're building. And some of the, sort of, insights we learned was how best to do the, sort of, sampling.

But once you have these samples, it's all data driven-- sure.

AUDIENCE: Can you talk more about the sampling strategy schemes?

ANDY BECK: Yeah, so from a high level, you want to go from random sampling, which is a reasonable thing to do, to more intelligent sampling, based on knowing what the computer needs to learn more about. And one thing we've done and-- so it's sort of like figuring-- so the first step is sort of simple. You can randomly sample.

But then the second part is a little harder to figure out what examples do you want to enrich your training set for to make the system perform even better? And there's different things you can optimize for, for that. So it's sort of like this whole sampling actually being part of the machine learning procedure is quite useful. And you're not just going to be sampling once. You could iterate on this and keep providing different types of samples.

So for example, if you learn that it's missing certain types of errors, or it hasn't seen enough of certain-- there's many ways of getting at it. But if you know it hasn't seen enough types of examples in your training set, you can over-sample for that.

Or if you see you have a confusion matrix and you see it's failing on certain types, you can try to figure out why is it failing on those and alter the sampling procedure to enrich for that. You could even provide outputs to humans, who can point you to the areas where it's making mistakes. Because often you don't have exhaustively labeled. In this case, we actually did have exhaustively labeled slides. So it was somewhat easier.

But you can see there's even a lot of heterogeneity within the different classes. So you might do some clever tricks to figure out what are the types of the red class that it's getting wrong, and how am I going to fix that by providing it more examples?

So I think, sort of, that's one of the easier things to control. Rather than trying to tune other parameters within these super complicated networks, in our experience, just playing with the training, the sampling piece of the training, it should almost just be thought of as another parameter to optimize for when you're dealing with a problem where you have humongous slides and you can't use all the training data.

AUDIENCE: So decades ago, I met some pathologists who were looking at cervical cancer screening. And they thought that you could detect a gradient in the degree of atypia. And so not at training time but at testing time, what they were trying to do was to follow that gradient in order to find the most atypical part of of the image. Is that still believed to be true?

ANDY BECK: Yeah. That it's a continuum? Yeah, definitely.

PROFESSOR: You mean within a sample and in the slides.

ANDY BECK: Yeah, I mean, you mean just like a continuum of aggressiveness. Yeah, I think it is a continuum. I mean, this is more of a binary task, but there's going to be continuums of grade within the cancer. I mean, that's another level of adding on.

If we wanted to correlate this with outcome, it would definitely be valuable to do that. To not just say quantitate the bulk of tumor but to estimate the malignancy of every individual nucleus, which we can do also. So you can actually classify, not just tumor region but you can classify individual cells. And you can classify them based on malignancy. And then you can get the, sort of, gradient within a population.

In this study, it was just a region-based, not a cell-based, but you can definitely do that, and definitely, it's a spectrum. I mean, it's kind of like the atypia idea. Everything in biology is pretty much on a spectrum, like from normal to atypical to low-grade cancer, medium-grade cancer, high-grade cancer, and these sorts of methods do allow you to really more precisely estimate where you are on that continuum.

And that's the basic approach. We get the big whole site images. We figure out how to sample patches from the different regions to optimize performance of the model during training time. And then during testing time, just we take a whole big whole site image. We break it into millions of little patches. Send each patch individually.

We don't actually-- you could potentially use spatial information about how close they are to each other, which would make the process less efficient. We don't do that. We just send them in individually and then visualize the output as a heat map.

And this, I think, isn't in the reference I sent so the one I sent showed how you were able to combine the estimates of the deep learning system with the human pathologist's estimate to make the human pathologist's error rate go down by 85% and get to less than 1%.

And the interesting thing about how these systems keep getting better over time and potentially they over-fit to the competition data set-- because I think we submitted, maybe, three times, which isn't that many. But over the course of six months after the first closing of the competition, people kept competing and making systems better. And actually, the fully automated system on this data set achieved an error rate of less than 1% by the final submission date, which was significantly better than both the pathologists in the competition, which is the error rate, I believe, cited in the initial archive paper.

And also, they took the same set of slides and sent them out to pathologists operating in clinical practice, where they had really significantly higher error rates, mainly due to the fact, they were more constrained by time limitations in clinical practice than in the competition. And most of the errors they are making are false negatives. Simply, they don't have the time to focus on small regions of metastasis amid these humongous giga pixel-size slides.

AUDIENCE: In the paper, you say you combined the machine learning options with the pathologists, but you don't really say how. Is that it that they look at the heat maps, or is it just sort of combined?

ANDY BECK: Yeah, no, it's a great question. So today, we do it that way. And that's the way in clinical practice we're building it, that the pathologists will look at both and then make a diagnosis based on incorporating both.

For the competition, it was very simple, and the organizers actually did it. They interpreted them independently. So the pathologists just looked at all the slides. Our system made a prediction. It was literally the average of the probability that that slide contained cancer. That became the final score, and then the AUC went to 99% from whatever it was, 92% by combining these two scores.

AUDIENCE: I guess they make uncorrelated errors.

ANDY BECK: Exactly. They're pretty much uncorrelated, particularly because the pathologists tend to have almost all false negatives, and the deep learning system tends to be fooled by a few things, like artefact. And they do make uncorrelated errors, and that's why there's a huge bump in performance.

So I kind of made a reference to this, but any of these competition data sets are relatively easy

to get really good at. People have shown that you can actually build models that just predict a data set using deep learning. Like, deep learning is almost too good at finding certain patterns and can find artefact. So it's just a caveat to keep in mind.

We're doing experiments on lots of real-world testing of methods like this across many labs with many different standing procedures and tissue preparation procedures, et cetera, to evaluate the robustness. But that's why competition results, even ImageNet always need to be taken with a grain of salt.

And then but we sort of think the value add of this is going to be huge. I mean, it's hard to tell because it's such a big image, but this is what a pathologist today is looking at under a microscope, and it's very hard to see anything. And with a very simple visualization, just of the output of the AI system as red where cancer looks like it is. It's clearly a sort of great map of the areas they need to be sure to focus on.

And this is real data from this example, where this bright red area, in fact, contains this tiny little rim of metastatic breast cancer cells that would be very easy to miss without that assistant sort of just pointing you in the right place to look at, because it's a tiny set of 20 cells amid a big sea of all these normal lymphocytes.

And here's another one that, again, now you can see from low power. It's like a satellite image or something, where you can focus immediately on this little red area, that, again, is a tiny pocket of 10 cancer cells amid hundreds of thousands of normal cells that are now visible from low power.

So this is one application we're working on, where the clinical use case will be today, people are just sort of looking at images without the assistance of any machine learning. And they just have to kind of pick a number of patches to focus on with no guidance. So sometimes they focus on the right patches, sometimes they don't, but clearly they don't have time to look at all of this at high magnification, because that would take an entire day if you were trying to look at 40X magnification at the whole image.

So they sort of use their intuition to focus. And for that reason, they end up, as we've seen, making significant number of mistakes. It's not reproducible, because people focus on different aspects of the image, and it's pretty slow.

And they're faced with this empty report. So they have to actually summarize everything

they've looked at in a report. Like, what's the diagnosis? What's the size?

So let's say there's cancer here and cancer here, they have to manually add the distances of the cancer in those two regions. And then they have to put this into a staging system that incorporates how many areas of metastasis there are and how big are they? And all of these things are pretty much automatable.

And this is the kind of thing we're building, where the system will highlight where it sees cancer, tell the pathologist to focus there. And then based on the input of the AI system and the input of the pathologist can summarize all of that data, quantitative as well as diagnostic as well as summary staging.

Sort of if the pathologist then takes this is their first version of the report, they can edit it, confirm it, sign it out. That data goes back into the system, which can be used for more training data in the future and the case is signed out. So it's much faster, much more accurate, and standardized once this thing is fully developed, which it isn't yet.

So this is a great application for AI, because you really do need-- you actually do have a ton of data, so you need to do an exhaustive analysis that has a lot of value. It's a task where the local image data in a patch, which is really what this current generation of deep CNN's are really good at, is enough.

So we're looking at things at the cellular level. Radiology actually could be harder, because you often want to summarize over larger areas. Here, you really often have the salient information in patches that really are scalable in current ML systems.

And then we can interpret the output to the model. So it really isn't-- even though the model itself is a black box, we can visualize the output on top of the image, which gives us incredible advantage in terms of interpretability of what the models are doing well, what they're doing poorly on. And it's a specialty, pathology, where sort of 80% is not good enough. We want to get as close to 100% as possible.

And that's one sort of diagnostic application. The last, or one of the last examples I'm going to give has to do with precision immunotherapy, where we're not only trying to identify what the diagnosis is but to actually subtype patients to predict the right treatment. And as I mentioned earlier, immunotherapy is a really important and exciting, relatively new area of cancer therapy, which was another one of the big advances in 2012.

Around the same time that deep learning came out, the first studies came out showing that targeting a protein mostly on tumor cells but also on immune cells, the PD-1 or the PD-L1 protein, which the protein's job when it's on is to inhibit immune response.

But in the setting of cancer, the inhibition of immune response is actually bad for the patient, because the immune system's job is to really try to fight off the cancer. So they realized a very simple therapeutic strategy just having an antibody that binds to this inhibitory signal can sort of unleash the patient's own immune system to really end up curing really serious advanced cancers.

And that image on the top right sort of speaks to that, where this patient had a very large melanoma. And then they just got this antibody to target, to sort of invigorate their immune system, and then the tumor really shrunk.

And one of the big biomarkers for assessing which patients will benefit from these therapies is the tumor cell or the immune cell expressing this drug target PD-1 or PD-L1. And the one they test for is PD-L1, which is the ligand for the PD-1 receptor.

So this is often the key piece of data used to decide who gets these therapies. And it turns out, pathologists are pretty bad at scoring this, not surprisingly, because it's very difficult, and there's millions of cells potentially per case. And they show an interobserver agreement of only 0.86 for scoring on tumor cells, which isn't bad, but 0.2 for scoring it on immune cells, which is super important.

So this is a drug target. We're trying to measure to see which patients might get this life-saving therapy, but the diagnostic we have is super hard to interpret. And some studies, for this reason, have shown sort of mixed results about how valuable it is. In some cases, it appears valuable. In other cases, it appears it's not.

So we want to see would this be a good example of where we can use machine learning? And for this type of application, this is really hard, and we want to be able to apply it across not just one cancer but 20 different cancers.

So we built a system at PathAI for generating lots of training data at scale. And that's something that a competition just won't get you. Like that competition example had 300 slides. Once a year, they do it. But we want to be able to build these models every week or something.

So now, we have something 500 pathologists signed into our system that we can use to label lots of pathology data for us and to really build these models quickly and really high quality. So now we have something like over 2 and 1/2 million annotations in the system. And that allows us to build tissue region models.

And this is immunohistochemistry in a cancer, where we've trained a model to identify all of the cancer epithelium in red, the cancer stroma in green. So now we know where the protein is being expressed, in the epithelium or in the stroma.

And then we've also trained cellular classification. So now, for every single cell, we classify it as a cell type. Is it a cancer cell or a fibroblast or a macrophage or a lymphocyte? And is it expressing the protein, based on how brown it is? So while pathologists will try to make some estimate across the whole slide, we can actually compute for every cell and then compute exact statistics about which cells are expressing this protein and which patients might be the best candidates for therapy.

And then the question is, can we identify additional things beyond just PD-L1 protein expression that's predictive of response to immunotherapy? And we've developed some machine learning approaches for doing that.

And part of it's doing things like quantitating different cells and regions on H and E images, which currently aren't used at all in patient subtyping. But we can do analyses to extract new features here and to ask, even though nothing's known about these images and immunotherapy response, can we discover new features here?

And this would be an example routinely of the types of features we can quantify now using deep learning to extract these features on any case. And this is sort of like every sort of pathologic characteristic you can sort of imagine. And then we correlate these with drug response and can use this as a discovery tool for identifying new aspects of pathology predictive of which patients will respond best.

And then we can combine these features into models. This is sort of a ridiculous example because they're so different. But this would be one example where the output of the model, and this is totally fake data but I think it's just to get to the point. Is here, the color indicates the treatment, where green would be the immunotherapy, red would be the traditional therapy, and the goal is to build a model to predict which patients actually benefit from the therapy.

So this may be an easy question, but what do you think, if the model's working, what would the title of the graph on the right be versus the graph on the left if these are the ways of classifying patients with our model, and the classifications are going to be responder class or non-responder class? And the color indicates the drug.

AUDIENCE: The drug works or it doesn't work.

ANDY BECK: That's right but what's the output of the model? But you're right. The interpretation of these graphs is drug works, drug doesn't work. It's kind of a tricky question, right? But what is our model trying to predict?

AUDIENCE: Whether the person is going to die or not? It looks like likelihood of death is just not as high on the right.

ANDY BECK: I think the overall likelihood is the same on the two graphs, right versus left. You don't know how many patients are in each arm. But I think the one piece on it-- so green is experimental treatment. Red is conventional treatment. Maybe I already said that.

So here, and it's sort of like a read my mind type question, but here the output of the model would be responder to the drug would be the right class of patients. And the left class of patients would be non-responder to the drug. So you're not actually saying anything about prognosis, but you're saying that I'm predicting that if you're in the right population of patients, you will benefit from the blue drug.

And then you actually see that on this right population of patients, the blue drug does really well. And then the red drug are patients who we thought-- we predicted would benefit from the drug, but because it's an experiment, we didn't give them the right drug. And in fact, they did a whole lot worse. Whereas, the one on the left, we're saying you don't benefit from the drug, and they truly don't benefit from the drug.

So this is the way of using an output of a model to predict drug response and then visualizing whether it actually works. And it's kind of like the example I talked about before, but here's a real version of it. And you can learn this directly using machine learning to try to say, I want to find patients who actually benefit the most from a drug.

And then in terms of how do we validate our models are correct? I mean, we have two different ways. One is do stuff like that. So we build a model that says, respond to drug, don't

respond to a drug. And then we plot the Kaplan-Meier curves.

If it's image analysis stuff, we ask pathologists to hand label. Many cells, and we take the consensus of pathologists as our ground truth and go from there.

AUDIENCE: The way you're presenting it, it makes it sound like all the data comes from the pathology images. But in reality, people look at single nucleotide polymorphisms or gene sequences or all kinds of clinical data as well. So how do you get those?

ANDY BECK: Yeah, I mean, the beauty of the pathology data is it's always available. So that's why a lot of the stuff we do is focused on that, because every clinical trial patient has treatment data, outcome data, and pathology images. So it's like, we can really do this at scale pretty fast.

A lot of the other stuff is things like gene expression, many people are collecting them. And it's important to compare these to baselines or to integrate them. I mean, two things-- one is compare to it as a baseline. What can we predict in terms of responder, non-responder using just the pathology images versus using just gene expression data versus combining them?

And that would just be increasing the input feature space. Part of the input feature space comes from the images. Part of it comes from gene expression data. Then you use machine learning to focus on the most important characteristics and predict outcome.

And the other is if you want to sort of prioritize. Use pathology as a baseline because it's available on everyone. But then an adjuvant test that costs another $1,000 and might take another two weeks, how much does that add to the prediction? And that would be another way.

So I think it is important, but a lot of our technology to developing our platform is focused around how do we most effectively use pathology and can certainly add in gene expression date. I'm actually going to talk about that next-- one way of doing it. Because it's a very natural synergy, because they tell you very different things.

So here's one example of integrating, just kind of relative to that question, gene expression data with image data, where the cancer genome analysis, and this is all public. So they have pathology images, RNA data, clinical outcomes. They don't have the greatest treatment data, but it's a great place for method development for sort of ML in cancer, including pathology-type analyses.

So this is a case of melanoma. We've trained a model to identify cancer and stroma and all the different cells. And then we extract, as you saw, sort of hundreds of features. And then we can rank the features here by their correlation with survival.

So now we're mapping from pathology images to outcome data and we find just in a totally data-driven way that there's some small set of 15 features or so highly associated with survival. The rest aren't. And the top ranking one is an immune cell feature, increased area of stroma plasma cells that are associated with increased survival. And this was an analysis that was really just linking the images with outcome.

And then we can ask, well, what are the genes underlying this pathology? So pathology is telling you about cells and tissues. RNAs are telling you about the actual transcriptional landscape of what's going on underneath. And then we can rank all the genes in the genome just by their correlation with this quantitative phenotype we're measuring on the pathology images. And here are all the genes, ranked from 0 to 20,000. And again, we see a small set that we're thresholding at a correlation of 0.4, strongly associated with the pathologic phenotype we're measuring.

And then we sort of discover these sets of genes that are known to be highly enriched in immune cell genes. Sort of which is some form of validation that we're measuring what we think we're measuring, but also this sets of genes are potentially new drug targets, new diagnostics, et cetera, that was uncovered by going from clinical outcomes to pathology data to the underlying RNA signature.

And then kind of the beauty of the approach we're working on is it's super scalable, and in theory, you could apply it to all of TCGA or other data sets and apply it across cancer types and do things like find-- automatically find artefacts in all of the slides and kind of do this in a broad way. And then sort of the most interesting part, potentially, is analyzing the outputs of the models and how they correlate with things like drug response or underlying molecular profiles.

And this is really the process we're working on, is how do we go from images to new ways of measuring disease pathology? And kind of in summary, a lot of the technology development that I think is most important today for getting ML to work really well in the real world for applications in medicine is a lot about being super thoughtful about building the right training data set. And how do you do that in a scalable way and even in a way that incorporates

machine learning?

Which is kind of what I was talking about before-- intelligently picking patches. But that sort of concept applies everywhere. So I think there's almost more room for innovation on the defining the training data set side than on the predictive modeling side, and then putting the two together is incredibly important.

And for the kind of work we're doing, there's already such great advances in image processing. A lot of it's about engineering and scalability, as well as rigorous validation. And then how do we connect it with underlying molecular data as well as clinical outcome data? Versus trying to solve a lot of the core vision tasks, which there's already just been incredible progress over the past couple of years.

And in terms of in our world, things we think a lot about, not just the technology and putting together our data sets but also, how do we work with regulators? How do we make strong business cases for partners working with to actually change what they're doing to incorporate some of these new approaches that will really bring benefits to patients around quality and accuracy in their diagnosis?

So in summary-- I know you have to go in four minutes-- this has been a longstanding problem. There's nothing new about trying to apply AI to diagnostics or to vision tasks, but there are some really big differences in the past five years that, even in my short career, I've seen a sea change in this field.

One is availability of digital data-- it's now much cheaper to generate lots of images at scale. But even more important, I think, are the last two, which is access to large-scale computing resources is a game-changer for anyone with access to cloud computing or large computing resources. Just, we all have access to a sort of arbitrary compute today, and 10 years ago, that was a huge limitation in this field. As well as these really major algorithmic advances, particularly deep CNN's revision.

And, in general, AI works extremely well when problems can be defined to get the right type of training data, access, large-scale computing, as well as implement things like deep CNNs that work really well. And it sort of fails everywhere else, which is probably 98% of things. But if you can create a problem where the algorithms actually work, you can have lots of data to train on, they can succeed really well.

And this sort of vision-based AI-powered pathology is broadly applicable across, really, all image-based tasks and pathology. It does enable integration with things like omics data-- genomics, transcriptonics, SNP data, et cetera. And in the near future, we think this will be incorporated into clinical practice. And even today, it's really central to a lot of research efforts.

And I just want to end on a quote from 1987, where in the future, AI can be expected to become staples of pathology practice. And I think we're much, much closer than 30 years ago.

And I want to thank everyone at PathAI, as well as Hunter, who really helped put together a lot of these slides. And we do have lots of opportunities for machine learning engineers, software engineers, et cetera, at PathAI. So certainly reach out if you're interested in learning more. And I'm happy to take any questions, if we have time. So thank you.

[APPLAUSE]

AUDIENCE: Yes, I think generally very aggressive events. I was wondering how close is this to clinical practice? Is there FDA or--

ANDY BECK: Yeah, so I mean, actual clinical practice, probably 2020, like early, mid-2020. But I mean, today, it's very active in clinical research, so like clinical trials, et cetera, that do involve patients, but it's in a much more well-defined setting. But the first clinical use cases, at least of the types of stuff we're building, will be, I think, about a year from now.

And I think it will start small and then get progressively bigger. So I don't think it's going to be everything all at once transforms in the clinic, but I do think we'll start seeing the first applications out. And they will go-- some of them will go through the FDA, and there'll be some laboratory-developed tests. Ours will go through the FDA, but labs themselves can actually validate tools themselves. And that's another path.

AUDIENCE: Thanks.

ANDY BECK: Sure.

PROFESSOR: So have you been using observational data sets? You gave one example where you tried to use data from a randomized controlled trial, or both trials, you used different randomized control trials for different efficacies of each event.

The next major segment of this course, starting in about two weeks, will be about causal

inference from observational data. I'm wondering if that is something PathAI has gotten into yet? And if so, what has your finding been so far?

ANDY BECK: So we have focused a lot on randomized controlled trial data and have developed methods around that, which sort of simplifies the problem and allows us to do, I think, pretty clever things around how to generate those types of graphs I was showing, where you truly can infer the treatment is having an effect.

And we've done far less. I'm super interested in that. I'd say the advantages of RCTs are people are already investing hugely in building these very well-curated data sets that include images, molecular data, when available, treatment, and outcome. And it's just that's there, because they've invested in the clinical trial. They've invested in generating that data set.

To me, the big challenge in observational stuff, there's a few but I'd be interested in what you guys are doing and learn about it, is getting the data is not easy, right? The outcome data is not-- linking the pathology images with the outcome data even is, actually, in my opinion, harder in observational way than in RCT. Because they're actually doing it and paying for it and collecting it in RCTs.

No one's really done a very good job of-- TCGA would be a good place to play around with because that is observational data. And we want to also, we generally want to focus on actionable decisions. And RCT is sort of perfectly set up for that. Do I give drug X or not?

So I think if you put together the right data set and somehow make the results actionable, it could be really, really useful, because there is a lot of data. But I think just collecting the outcomes and linking them with images is actually quite hard. And ironically, I think it's harder for observational than for randomized control trials, where they're already collecting it.

I guess one example would be the Nurses' Health Study or these big epidemiology cohorts, potentially. They are collecting that data and organizing it. But what were you thinking about? Do you have anything with pathology in mind for causal inference from observational data?

PROFESSOR: Well, I think, the example you gave, like Nurses' Health Study or the Framingham study, where you're tracking patients across time. They're getting different interventions across time. And because of the way the study was designed, in fact, there are even good outcomes for patients across times. So that problem in the profession doesn't happen there.

But then suppose you were to take it from a biobank and do pathologies? You're now getting

the samples. Then, you can ask about, well, what is the effect of different interventions or treatment plans on outcomes? The challenge, of course, drawing inferences there is that there was bias in terms of who got what treatments. That's where the techniques that we talk about in class would become very important. I just say, I appreciate the challenges that you mentioned.

ANDY BECK: I think it's incredibly powerful. I think the other issue I just think about is that treatments change so quickly over time. So you don't want to be like overfitting to the past. But I think there's certain cases where the therapeutic decisions today are similar to what they were in the past. There are other areas, like immunooncology, where there's just no history to learn from. So I think it depends on the--

PROFESSOR: All right, then with that, let's thank Andy Beck.

[APPLAUSE]

ANDY BECK: Thank you.