Tyler: This is the healthcare.ai live broadcast with the Health Catalyst Data Science team, where we discuss the latest machine learning topics with hands-on examples. And here's your host, Levi Thatcher.

Levi (L): Hi everyone, I'm Levi Thatcher. I'd like to introduce you to Adam Frisbee [and] Mike Mastanduno from Health Catalyst Data Science. If this is your first time joining us, we are gonna reach out each week through these broadcasts to help interact with the healthcare community and the machine learning community, to help people learn how to do hands-on machine learning and get involved, yourself, with your data.

L: To start off todays broadcast, we're gonna start with the mailbag, see what people have been chatting about and what questions you guys might have that we can help with. Mike?

Mike (M): Great. Thanks, Levi. So, just to give an overview of the session, a couple of house keeping things. We want to make sure you guys are interacting with us through the chat, because that's the way we're gonna get your questions addressed and be able to have a conversation, make this interactive. To do that, make sure you're logged in to YouTube. You can chat right on the "healthcare.ai/broadcasts" page or you can do it on the YouTube page itself. Make sure to change your resolution. We may be looking at text on the screen, so if you can support a high definition resolution, that'll definitely help. Please remember to show us some love and subscribe by clicking the bell for notifications on this channel as well.

M: With that, we will move onto the mailbag. We've had some questions from users over the past week and we're excited to share our opinions on them. Then, we'll talk about some really interesting news articles that came out this week in deep learning. And finally, move onto the real beef and potatoes—

L: Yeah. The bread of the lesson.

M: of the lesson: [we'll] be talking about MOOCs, which are "Massively Online Courses." There's an extra "o" in there.

Adam (A): Open.

M: ["Massively Open Online Courses."] Thank you, Adam. And finally, we'll cover some of the questions that the chat has. Let's move on to the mailbag. Levi, can you talk to us about the first question? We had some, a lot of, people kind of asking just for a more broad overview of machine learning. So like, *is machine learning for predictive only?*

L: Great question whoever that was. Do we have a username?

M: We don't have a username with that one.

L: Well, whoever you were, thank you. So, the idea of machine learning is that it's not just predictive. Predictive is a major part of it, but the idea is that you can make predictions or you can analyze data. Let's say that you have high cost patients coming into your hospital. They're there for a long time and they have expensive treatments, surgeries, what have you. One thing that you can do with machine learning, that's not predictive, is do what's called k-means

clustering. It's an unsupervised type method, and we won't get too technical, but the idea is that if you want to identify the different kinds of patients that are coming in that are high cost, maybe there's various attributes about these folks that are putting them into three or four groups that you want to learn about. That would be one application of machine learning that's not predictive in nature, but more for analysis. Great question.

M: So does it require big data?

L: Great question as well! No it doesn't. That's one of the awesome things about healthcare.ai and machine learning in general is that you can do it on your own computer. With healthcare.ai and these packages from R and python, you can put them on your laptop, play with a few thousand rows or a couple million rows, depending on how much memory you have. But, it doesn't require big data. And big data, it seems like the buzz around it has kind of died down a little bit because of that. Machine learning's kinda taken over and we're excited about that fact that you can do it right at home.

A: I think people should remember that our personal computers are so powerful now. The phone in my pocket is 10x more powerful than the most powerful thing in the 1970s. So, it's very interesting to realize the power that we can actually do with our own individual systems. We don't need huge clusters of machines, like in the past.

L: Yeah in the future, we'll be doing it on our smartphones.

A: I'm gonna look for the contact. [*motions to his eyes*]

L: There as well.

A: Wearable tech.

M: *What is the machine learning process like?* We've also had a lot of questions about that. *Is it just model building or is there more to it?*

L: Yeah so it starts with getting to know your data, *what are the distributions of particular variables? Do they make sense? Do you have a lot of missing values*, like *do you need to cut out certain rows because most of those rows have missing data?* So you'll want to consider some of those type of questions. Maybe you'll want to do some plots to look at distributions. Maybe you'll want to transform certain columns in a certain way. Let's say that you have a latitude and longitude for a particular set of patients and you would need to turn that into maybe a zip code sort of thing. So, that's sort of the first step. And then after that, you try a couple different algorithms and compare your column set with those different algorithms to say, *okay well algorithm A gives us this accuracy for our patients. Algorithm B gives a certain other accuracy.* And then from there you go and deploy the algorithm that worked best, such that each night you're getting those refresh predictions for things like CLABSI predication, or readmissions prediction. So that's kinda the general work flow, and we can go into a little more detail in perhaps another episode. That's a great idea.

M: Yeah, in fact next week's episode, we're gonna be focusing on the machine learning platform, just so that we can make sure that we know everybody's on the same page and talk about the nuances of how you actually build a machine learning model from start to finish.

L: Yeah I'm excited for that.

M: Really looking forward to that episode. We had one more question that's a bit nitty gritty, so I'll try and frame it for you guys. We had someone doing some machine learning, and they were playing with different algorithms. When they got their results from each algorithm back, they were seeing [that the] feature importance for the two algorithms was vastly different. Meaning that the two algorithms were using different information to come to similar conclusions about their predictions. So, basically, the question was *why is that?*

L: Yeah it's a great question. That's one thing, in healthcare.ai, we've tried to offer up this guidance to whoever the end user is. So when you run an algorithm on your data, not only do you get an accuracy score or a performance metric that's spit out, but you also get guidance as to which features perhaps you can leave out of the data set. Like Mike was saying, that differs. If I'm doing the Lasso algorithm, the order list would be different than if I'm doing the RandomForest algorithm. And that's just due to the different ways that algorithms work, right? Lasso is a linear model. RandomForest is an ensemble method that combines output from a lot of different trees. There's actually a blog post up on healthcare.ai that touches on that. It's a few weeks back, but the idea is that different algorithms calculate things in different ways. Keep that in mind when you're looking at the output for healthcare.ai.

M: Great. Yeah. That's awesome. Thanks for all those great questions from the users and we'll be sure to keep monitoring the chat, and answering your questions as they come up. We'll also be collecting them throughout the week so that we can answer them at the beginning of every show.

M: Let's move on to machine learning in the news. This week there was a ton of really interesting news surrounding deep learning, especially. One study in particular, a group in France was using a neural network, which is a deep learning technique, to synthetically age faces or make them look younger.

L: This is crazy.

M: That's really interesting because, right now, the only tools to do that sort of thing are photoshop. [And deep] learning researchers have had the capability to do artificial aging for a while. But usually what happens is that you kind of lose the identity of the face at the same time. So this study is really exciting because they found a way to use two different deep learning methods towards the same output to not only preserve the identity of the face, but also to help age or make them younger.

L: That's amazing.

M: And so we thought that was really, really interesting.

L: So you can….Is that Richard Gere? What celebrity is the middle one there? I think it might be him.

M: It does look like him.

L: Yeah so you're able to pick which decade you prefer for Richard Gere. 50's Richard Gere? 30's Richard Gere? Adam?

A: 50's Adam. Wait. What?

L: 50's Richard Gere?

A: Oh I'm all about 30's Richard Gere.

[*All in agreement*]

L: Yeah that's a pretty good decade for him. Well that's interesting, Mike.

M: So deep learning's being used on the cutting edge of research. But then, also, it's being used in everyday products. There's a long read here, [*pulls up the article*] but the "Inside Facebook's AI Machine" kinda goes through how deep learning and artificial intelligence are really just Facebook's platform. So, whether or not you know it, some of those recommendations they're serving up, the way they're interacting with all the data they collect is with artificial intelligence. So, it's kind of an interesting article about how the director of engineering transformed them into an artificial intelligence first company in the last few years. I think that's definitely gonna be a common theme in companies these days.

L: Yeah you'll see that from every tech company in Silicon Valley will pretty much have a story like this on the web soon. We'll actually drop these links in our show notes as well afterwards, so—

A: You know what's weird about that, if I can interrupt you…

L: Yeah please.

A: When I buy something at Walmart or some place that has a really sophisticated point of sale system, and I come home or I log into Facebook in my phone, I'll often see an advertisement related to the physical thing that I purchased. And that's when things start to get a bit scary for me.

L: A little creepy.

A: Little bit, yes.

L: But maybe you made that purchase that was suggested?

A: Yeah well, I wasn't going to announce that publicly, but yes I did.

L: Yes, okay. So, effective and creepy.

A: They got me.

L: It happens.

M: Deep learning, we've seen it all across the news. It's a big topic. In a very data science and machine learning-esque way, someone compiled a list of all the most cited deep-learning papers into a GitHub repository. So, bibliographies are over. We're all on GitHub now. So we might as well just post our reference list on GitHub too.

L: But check this out. I'll scroll down a little bit because they break it down by category. They have an "Awesome Badge," which is awesome. I guess you can put an awesome badge on there.

A: That's so subjective.

M: Could we get one of those badges?

L: We should get an awesome badge on our [refos?]. haha Oh they have an "awesome list criteria." So, I guess if they define what it takes.

M: That's interesting because [in] the data science process, you always have to define.

L: As long as you're defining it, you're good to go. So, definitely some awesome articles this week. Check them out again in the show notes on the YouTube live page. We'll drop those in here shortly after the broadcast. Again, with the mailbag, reach out, let us know what you're wanting to talk about, what questions you have. This show can be whatever you want it to be, so please drive it. We're happy to answer anything in machine learning and healthcare, and the intersection of those two. Anything else in the news this week?

M: There's a lot of great stuff, but I think we're out of time for the news, so let's move on to our main program. We'll make sure we get to talk about some of those new articles next week.

L: Sounds great. The main topic for today....We'll always have a main lesson or tutorial or something like that. So, today we want to focus on MOOCs.

A: [*like a cow*] Moo-ooCs.

L: Yeah, so MOOCs are "Massively Open Online Courses." And we brought Adam Frisbee on as a guest today as he is an instructor at the University of Utah. Do you want to give a little bit about how long you've been up there and what kind of courses you teach?

A: Yeah, actually, my full-time job is here at Health Catalyst. I am lucky enough and privileged enough to teach in the Information Systems and Masters of Science and Information Systems programs at the University of Utah. I've been doing that since 2014.

L: Three years, coming up on.

A: Yeah, coming up on three years, and it's just been really incredibly enjoyable, and rewarding for me. And I've got to teach some data related classes recently, like data warehousing, for example.

L: Very topical.

A: Yeah but the other side of that is being a teacher for me means being a perpetual lifelong learner. I have never learned so much as when I've decided to teach. I think that's sort of true for [all teachers]. If you were to take a poll of college teachers, that probably would be a sentiment that a lot of us share. I love to do online courses and it helps if they're free. MOOCs are "Massively [Open Online] Courses." Many of them are free. Some of them are free or you pay a little bit of money to get a verified certificate or something.

L: Okay so that's an interesting point. So, they've gotten huge in data science and Adam's taken a few of these MOOCs himself, always diversifying, always learning. Some people out there might ask, okay well—we'll get to the certificate in a second, but like *what's the benefit of a MOOC over more formalized education, like at your University?*

A: Great question, so I'm gonna be a little bit biased on the answer to that, and I'm going to say that, in my opinion, nothing can take the place of a University education, where you physically go somewhere with a bunch of learners who are interested in the same topic and you enroll in an official degree program or do some sort of professional education. I don't think anything can take the place of that because it's just been true that debating with other students and with professors and teachers in real-time, is a really, really effective way to learn. It's how we've always done it, since medieval universities.

L: Since medieval times! But you're doing MOOCs for data science because maybe the program's not there at the U, or you want to do it in your spare time?

A: Well the program is there at the U. However, I think MOOCs are an excellent—I'd even go so far as to say perfect—addendum to one's formal education because MOOCs tend to be more focused, in my opinion. If you want a MOOC specifically on R, there are those available. If you want a MOOC specifically on machine learning or a sub topic of machine learning, they are available, for sure.

L: So you cut out the fluff [and] focus on the topic at hand?

A: Right. And usually for free.

L: Awesome. Usually for free, that's a great question.

A: Or at least there's a free option out there.

L: So what is the difference? Sometimes courses will let you do either. So why would one get the certificate—and have you gotten the certificates?

A: I usually do the certificates. But I do them for me because, for example, one of the main MOOC leaders is Coursera, and they offer a certificate program where you pay like fifty dollars a month and then you can get a verified certificate. Then what you can do is log onto Coursera at any time and you can look at your transcript and look at all the certificates you've earned. Whereas if you don't do that, you can take the classes and do everything, but they don't keep a history and they don't keep a transcript for you. So that's an advantage for me.

L: A little motivation.

A: Yeah and you're paying money, so that's a little motivation to actually complete a MOOC that you start. And the other thing is, I really don't think employers necessarily care if you get a verified certificate or not, so you have to do it for yourself.

L: Yeah I guess it's nice to see on the resume. If two people are equal, that maybe would be beneficial, but it'll always come out whether you know the stuff or not, so learn the stuff primarily.

A: Yeah and if nothing else, [put it] on your LinkedIn because you can put little badges and stuff.

L: Oh the badge, yeah. Like the little awesome badge?

A: Yeah I have tons of those awesome badges. Haha

L: Oh that's nice. Totally. Should we go into *what are some of our recommended MOOCs?* We've been looking around, Mike's been really helpful, do you want to just give a little bit of that?

M: Before we get into our recommended classes, I'm kind of curious [because there are] so many of them, what would you look for in a MOOC?

A: Ratings. Just like anything else on the interwebs, ratings are king. You don't buy an Amazon product that has one star. One of the big paid MOOC providers is udemy.com. And they always have sales. Their classes are usually like two hundred bucks, but they always—every time I've ever logged in there, it's always like ~~$200~~, and then it's twenty dollars to take the class. So every time. Don't worry about the cost on those. It's twenty dollars a class. That's totally worth it. This one instructor that I really like, that I kind of follow on Udemy, is Kirill. And he teaches a lot of the data science stuff, but he's always got five stars and like two thousand reviews. I trusted that and so I bought his course and it turned out that it was a really cool course. So, I do recommend him as an instructor.

L: So like good material or good style? What was it about him?

A: Engaging style, not boring, like me. No haha. Hopefully I'm not. Yeah, really engaging style. But, another thing about him is everything in the course is live.

L: So he's typing the code and like actually working?

A: Yes he's typing the code, he's showing you exactly what he's doing, and he actually writes a little R comment right before what he's about to type, saying "this is about to do this." So, it's really easy to follow along with, and he has data sets for you already prepared for his exercises. That's really good too.

L: That's really nice.

A: Other ones I've taken, like Coursera are really good, don't get me wrong, but they tend to be like PowerPoint or slideshow focus to where it's not really live coding. There's a preference thing there too because some people might do fine with that.

L: It seems like hands on is good. That's kind of our theme as well.

A: For me, yeah.

M: Yeah so a couple questions have come up kind of on these lines. We had one person ask where we can find good publicly available data sets.

A: Oh. Those are everywhere.

M: You know sometimes they come with the MOOCs, but not always. Where are some good places to look?

A: Can I answer that?

L: Yeah please do.

A: Okay well in my university class, currently data warehousing, where one of the projects that we're doing is: take a data set, put it in a data warehouse, [and] do some analytics on it. That's the group project. So I had to find these. There's a lot of lists of data, if you just Google "public data sets," they're all over the place. And I told students that I would give them extra credit if they did some of the weirder ones. So if you Google "weird public data sets," you'll find "Big Foot sightings," like GIS maps of Big Foot sightings; there's also UFO sightings. So there's a bunch of weird ones out there.

L: So your class did some of these?

A: Well the group project is not due until the end of the semester, but—

L: Hopefully they're working through them.

A: Yeah exactly. I want to know. I said, if you can accurately predict when I'm going to see Big Foot and where, then you will get an automatic A in the class.

L: Quite the prediction.

A: Yeah there are so many [public data sets]. Data.gov has all kinds of stuff about the demographics of the United States: health-related, or otherwise. Census data.

L: Yeah it's been kind a nice initiative. I think in the past administration, there was a lot of effort under way to actually put those data sets out there.

A: Get it before Trump takes it away is what I'm gonna say.

L: That's probably good advice. So I had one as well. There's a UCI, so the University of Californi**a:** Irvine [*searches for the Irvine UCI data sets*].

M: While Levi pulls that up, I'd also just like to make a plug for APIs, which are live streams of information coming from different companies and websites. Uber has one. Twitter has one. Reddit has one. Any website that's collecting data, you can kinda tap into that hose and pull data off their website to use for your own analysis.

L: That would make for really fun projects. So like, where are Ubers in my city? Where are the buses in my city? Are they late?

M: Exactly.

L: [*referring to the UCI data sets*] This one Irvine is amazing for machine learning, in particular. They have a nice break down in terms of whether it's a classification or regression-type data set, the number of rows, the number of columns, the year it came out. And we'll throw that one up and some of these other ones while in the show notes after we're done.

M: Yeah definitely don't worry about writing down everything we say. It's gonna go in the show notes. Every class we talk about, all the links we pull up, everything will be in the notes. So don't worry about that at all.

L: Yeah sounds good. So do you want to talk about some specific MOOCs that we'd recommend?

M: Yeah. I guess I'm still kind of concerned about how to choose one.

L: Yeah let's go over that.

M: We want a good teacher. We want good content, but what is good content? What are we looking for? How do we choose it?

L: Being a data science focused, and machine learning focused [broadcast] is making it appropriate that we talk about that topic, so things with R, python, data science, and the sort of subfields that lie under those. There were a few that we looked at that talked about data science, but talked about maybe older technologies or technologies that we really don't use, and our colleagues don't use, so we kind of ignored those ones. Do you want to kinda talk about this freeCodeCamp article we found a little bit?

M: Yeah so in a very data science-oriented fashion, someone wrote a blog post about MOOCs. We thought it was really engaging because it kind of covered all these points that we've talked about: How the class has to cover the content, it has to use the right tools, and it has to be highly rated. Someone kind of scraped the website "Class Central" for reviews of all these different online courses, and came up with a list of great ones. This person did it as the culmination of their self-guided MOOC-based data science degree, that they claim they paid a lot less for than a pompous academic degree.

L: Sometimes pompous.

A: These guys have PhD's, so…

M: Just as much success.

L: Rebels fleeing academia.

M: We really like that because…

A: I do want to keep my job.

M: Of course you do, Adam. We really liked the flow of that article and some of the recommendations based on the fact that they're grounded in data. They have reviews to back them up. They have what people say about them. It's not just your friend saying *Hey, you should try this one.*

L: Exactly. We thought we'd throw this out there since a lot of people ask us, both in the community and Health Catalyst—they say *How do I get more into machine learning and data science?* We liked a couple of these that he mentions, in particular, so do we want to go into the course [into the Udemy one?]?

A: Let me just interject one thing.

L: Please do.

A: For every MOOC that I've ever come across, you can always watch some sort of preview video or preview of it.

L: A little taster.

A: So that's one thing I did with this Kirill guy on Udemy is I watched his preview video and that helped sell his MOOC.

M: And the syllabus is online too, so if you just wanna read through it, you can read through it.

M: *Where is a good place to start?* We want to learn about the machine learning [pipeline]. We want to learn about what data science even is.

L: Yeah so there's the data science toolbox, getting at Mike's comments there, on Coursera. If you're looking to learn about the field on a broad sense, and kinda get a flavor for what tools you'd even be using, we'd recommend this one. Is this one by Roger Pang?

A: It is. Roger Pang and his Johns Hopkins cohort there, and that course is really awesome because they show you what tools data scientists use, how to get them, and the basics of how to use them.

L: So we're talking GitHub, R—

A: R and RStudio.

L: Awesome. So if those things are new to you, those concepts are not familiar, check this MOOC out first. And we'll put this in the show notes…

A: They also go into a little bit of theory of what you want to do as a data scientist, for example, ask the right questions.

L: Oh yeah, a little bit of process too. That's really helpful.

M: That's really fantastic because a lot of times you'll see a list of "So you want to become a data scientist? You must learn all of these things."

A: It's overwhelming.

M: Yeah its overwhelming. And the lists, they compete with each other, so you don't know [what you really need]. This is kinda like it's a trailer to the movie. It kinda fills in on getting that big picture oriented in your mind so that you know what you might want to go pursue a little bit more.

A: Speaking of movies, do you guys remember Star Trek Four?

M: Can't say I saw it. Sorry.

L: Not off the top of my head.

A: Okay so Spok's dad in Star Trek Four has an awesome quote. He says, "It is difficult to provide an answer when one does not know the question."

L: Ooo That's pretty deep. Show notes!

A: And remember they're talking about the whales are trying to talk to them in Star Trek Four.

L: The whales are? Are they underwater?

A: And there are no whales in the twenty third century because they become extinct by that time.

L: From pollution et cetera.

A: Right. Because we suck at maintaining our planet.

L: To a certain extent. Okay so you gotta know the question if you're looking for an answer, is what we're getting at here. So, if you're wondering if you should drop out of your PhD program and become a data scientist, you may want to do this MOOC first, and then decide to drop out or not.

M: So then programming is also a big part of data science. It's kind of the most basic tool that you have to know. So what languages?

L: Well what have you been delving into, Adam? You're going down this path.

A: The MOOCs that I have done have been focused on R because I think R and python are kind of the two big players right now, but I think I just chose one and went with it. R just kind of—

L: It was concise naming.

A: —spoke to me a little bit more than python, but I think there's a lot of personal preference that goes into play there.

L: Yeah both are awesome. We found this amazing MOOC here that you've actually done, and we found this in the list we referenced earlier. So this is by Jose Portilla. And you recommend Jose?

A: Yeah he's really good. Not my favorite—

L: But good.

A: —but second probably on Udemy.

L: And this is "Data Science and Machine Learing Bootcamp with R."

A: Now look at the price on that. See, they all say 90% off all the time. I've never paid full price for a Udemy course. I think that's how they get you, but it's worth it.

L: Oh yeah? They reel you in?

A: It's worth twenty bucks. Trust me.

L: Check it out. Okay. And then if you're looking for something a little bit more broad, [*pulled up another course titled "Machine Learning A-Z™: Hands-On Python & R in Data Science"*]

A: I'm enrolled in this one now.

L: How is it? Break it down. It's python and R, so is each segment in both languages?

A: Well he—I actually just started this one, so I can't answer that. But,

L: But it is by Kirill.

A: This is my favorite instructor on Udemy. Kirill Eremenko. And he also runs a data science podcast called "Super Data Science." So sign up for that too. I mean it's not as cool as this broadcast, but it's pretty cool. He interviews some really accomplished data scientists every week, so it's pretty [cool]. So, superdatascience.com if I can plug his.

L: Yeah! Superdatascience.com. It's a community effort.

A: That's where his data sets are too for his courses.

L: Oh that's handy. Speaking of open data sets.

A: Right.

L: Yeah, awesome. So again, we'll throw those all up afterwards. Don't feel like you need to write them down. And again, we love the mail. Keep that coming in. Next week we'll be going through how to deploy a model on your laptop, on your server; we'll give you the full run through as to the data prep, developing a model, deploying a model, and actually using the predictions for something cool. But anything else today? I think we—

M: Yeah I just had one other thing I wanted to say. So maybe you've gone through all these classes and you start to feel pretty good about the toolset. Where do you go next?

L: Are you still kind of in the process?

A: I am still. I am both in the process and outside of the process, very different. I'm kinda like Batman in that regard.

L: Teaching. Yeah. You're doing everything.

A: Kinda like Bruce Wayne and Batman. Although I wish I was rich. Anyway, what I would say is there are a lot of places online where there are…kind of like competitions, where it's like "take this data set and whoever can do the coolest thing with it or whatever, or whoever can make the best prediction or whatever." There are a lot of websites like that. So that is really good practice.

M: Kaggle.

L: Kaggle! I believe we're pronouncing it correctly.

[*discussion about how to pronounce Kaggle*]

A: So we have—slightly off-topic, but we have a BI expert, Curtis Harris, who does a lot of Tableau stuff, and every Monday there's like a Tableau Monday thing that they do.

L: Competition?

A: Yeah it's like a competition. And he does those *every* Monday, and so he is so good at Tableau just from doing those. So I would say, just like math or anything else, I mean practice, practice, practice, practice, until you're blue in the face really, is the way to do it.

L: Yeah yeah. That's fantastic. Did you have anything, Mike, in terms of where to go?

M: Yeah and I think Paul just suggested it as well. If you're looking for a data science job, using GitHub to showcase the work you've done is a really great way to get yourself a portfolio, almost. You know?

L: Yeah yeah yeah!

M: These degrees are so new. The best way to show that you're capable of doing the work is to kind of put it online and show it to people on GitHub.

L: Yeah.

A: But what's a huge advantage to these degrees being so new? Anybody can get into it.

M: That's true.

L: It's open. It's wild west out here.

A: If you have the desire, anybody can get [into it]. I think it's possible for anyone that has a desire to step into something like this.

L: That's one of the awesome things about this whole open source data science movement is that it broadens the scope and anybody can get involved. And speaking of portfolios, blogs are another fantastic way. We actually just talked about this. I gave a little bit of a talk at Adam's class the other night and went into how blogging—

A: You did a great job, by the way.

L: Oh thanks. That was nice of you to let me come up. But, bloggings an awesome way to work through different data sets: ask questions of the data, make some visualizations, maybe some predictions, and talk about what you're learning and your process.

A: And be vulnerable.

L: Yeah, people love vulnerability.

A: If you—I do some light blogging, but I like, I mean I'm fine pointing out my mistakes because I want people to see—If they ever come across my blog, which is unlikely—I want people to see that my learning process is no different than anybody else's.

L: It's endearing. And you learn as you type. That's one thing with writing in general, you have to really learn something to be able to write about it.

A: Yeah.

L: So just some random thoughts on data science there, kind of meandering, but I liked it. So anything else from the mailbag and chat?

A: We're not a scripted show.

L: Yeah, unscripted.

M: Yeah there's been quite a bunch of comments on the chat. Maybe we can just run through them real quick?

L: Yeah.

M: Paul is wondering if we have any thoughts on marrying clinical data from the electronic health records with socio-economic data.

L: Oh yeah!

M: Oh man. That is such a great idea. Usually when you're doing machine learning, the more data sets or sources you can pull together to get a complete picture of what's going on, the better your models are gonna be. So, definitely something we try to do.

L: Definitely, yeah. That's the idea of the data warehouse, you might have heard of it, the more data sets the better. We always say. So any other chat questions or?

M: Yeah we also have someone asking if adversarial neural nets are good for oncology work?

A: Is that English?

L: Well those are all the rage these days, I hear. GANs, Generative Adversarial Nets. Yeah probably cool with oncology data.

M: Probably.

L: We haven't tried it yet, but give it a whirl, let us know.

M: It seems like cancer prediction and classification is really the cutting edge in machine learning right now—

A: Protein folding.

M: Just because [of] the application. There's so much low-hanging fruit in other areas, like hospital readmissions and hospital-based infections that it's hard to justify spending all that resource on cancer classification.

L: Yeah you gotta start with the most practical use case and move up from there, which is what we're trying to do.

M: But eventually that's gonna move on from the academic realm and into the [industry], we'll see that in hospitals.

L: Yeah exactly, and that's what healthcare.ai is all about: pacing that transition.

M: And then Adam, we had a question for you specifically. *Do you have any experiences with classes on Redshift?*

A: Actually yeah. That's what we're focusing on in my class, currently.

L: This wasn't scripted.

A: I see that one of my former students asked that question. Hi Amelia. Actually it's been really great. If you don't know, Redshift is a cloud data warehousing platform offered by Amazon. And it's a distributed data warehouse. It's really fast, honestly. The other classes are using an older technology, but this semester, I asked our department chair, I said "I really want to use this new technology because it's what people are actually using in the working world." And my department chair was extremely supportive of me, so I'm very excited about this. Students are seeming to like it.

L: That's really cool.

A: Yeah so Amazon Redshift, if you guys are interested in data warehousing. There's a lot of pre-built labs you can do.

L: Oh and the nice thing with Amazon or AWS is you can use it for free if you don't use that much compute power. They have a free tier right?

A: Yes.

M: Yeah and we've been told that Azure and Adobe also do.

L: Awesome. That's right.

M: So whatever you want to use.

A: Yeah for sure.

L: Everything is very open.

A: To a point.

L: Small data.

A: New term?

L: Not mine.

A: Okay.

M: So we have one more question concerning the data and Bob asked if Health Catalyst could make patient level data that's de-identified available to educators and researchers and analysts, people who want to tinker with it. What are the problems or how possible is that?

L: That's one of the things we're excited about with healthcare.ai and machine learning in general is not only offering up the algorithms, but these open data sets to accelerate research in healthcare and machine learning. We have this CAFÉ project, which is gonna aggregate data from multiple health systems. There's some way, and the details haven't been sorted out, but in the future we're eager to provide such data sets for exactly those reasons you mentioned.

A: And we're starting to do that internally, so it's not unreasonable to think that at some point, they can be public for sure.

M: Definitely a challenging subject with all of the—

A: HIPAA.

M: HIPAA compliance and the patient-specific information that has to get removed.

A: I mean think about what's on a medical record though. I mean, everything about you. And identity theft is such a huge thing anyway.

L: So we're skating that, but it might be a bit.

M: Great, unless you guys have other stuff to cover.

A: Let me just say something about this Redshift question. Sorry.

L: Back to Redshift.

A: I will say if you want to learn anything about RedShift or anything on Amazon, go to qwicklabs.com. They're an official Amazon partner. They're not exactly MOOCs because it's individual labs. The lab spins up an individual instance just for you, and it's like ten or fifteen dollars a lab, or something.

L: That's not bad.

A: Really really valuable. So, that's what we're doing in my class.

L: Very cool. That's awesome. We'll throw that in the show notes for afterwards. Other than that, thanks for joining us, Adam.

A: Oh yeah. Anytime. Sure.

L: Okay and yeah feel free to subscribe and you'll notice that on healthcare.ai and in the YouTube live chat, as well. Thanks for joining, we'll see you next week.