Hi, I'm Craig Smith and this is Eye on AI. This week I talked to Ben Sorscher, a PhD student at Stanford, about his work in reducing the size of data sets used to train models, particularly large language models, which are pushing the limits of scaling because of the enormous cost of training and the environmental impact of generating the electricity they consume. Ben and his colleagues, won an outstanding paper award at the NeurIPS conference in December for their paper: Beyond Neural Scaling Laws, beating power law scaling via data pruning.

Before we begin, I want to mention our sponsor, ClearML, an open source ML ops solution for data scientists who need an end-to-end platform for building machine learning models.

Try it for free at clear. Ml.

That's C L E A R ml. Now, I hope you find my conversation with Ben as fascinating as I.

**CRAIG:** The audience is pretty sophisticated, even if I'm not. But I think it's good because I ask questions that, maybe people would be embarrassed to ask if they're in the game. So why don't you by introducing yourself, and then I'll go from there.

**BEN:** Sure. Thanks a lot. It's really good to meet you. My name's Ben. I'm a PhD student at Stanford, and before that I studied physics and math at Harvard, and toward the end of my time there, I discovered that there were physicists using some of the mathematical tools that had proved useful in physics to study neural networks in the brain and in deep learning models. And I became really interested in this idea. And in particular, I was really captivated by this beautiful, or what I thought was beautiful idea that you can take all these really simple units, neurons and somehow when you wire them up together, they conspire to do these incredibly complex, intricate tasks like navigation and visual object recognition and motor planning.

**BEN:** And turns out that physicists over the past 60 years or so, have developed some really powerful techniques for studying these kinds of systems where you have lots of interacting, simple degrees of freedom that conspire to produce emergent really interesting behaviors. And so I've been working on that ever since, trying to use these physical techniques to study networks of neurons in the brain and in AI models.

**CRAIG:** A lot of the AI researchers I've spoken to came out of psychology or what in my day they call neuropsychology. I don't know what it's called now, but through study of the brain or cognitive models and that sort of thing. I mean, Geoff Hinton started wanting to understand the brain.

**CRAIG:** And there's some fascinating stuff at Harvard. I've been meaning to go up there. There's a guy with a machine that slices the brain really thin. And he does these beautiful visualizations

**BEN:** Oh yeah. Yeah. Jeff Lipman

**BEN:** I've seen that machine.

**CRAIG:** Wow. It's incredible. Yeah. So we're gonna talk about your paper on data pruning, is that specifically for language models or just for scaling anything? That's one question. But one of the many goals of AI is to create a cognitive model that at least mimics the brain, so are you interested in deep learning as a model of how the brain might be functioning?

**BEN:** Ultimately my interest is in understanding the brain rather than deep learning models.

**BEN:** And I don't know at the moment how good of a model deep learning systems are for the brain, but I think of deep learning as almost a warmup problem for understanding the brain, and unfortunately, it's an extremely difficult warmup problem, but in some sense, it's a lot easier than understanding the brain because we have access to the entire system.

**BEN:** We specify the architecture, the learning rule, the data, and an analogy to doing experiments in the brain. You can go in and record from any neuron in your machine learning model you like. You can measure any of the weights you like. You can do any experiment you like to these systems and nevertheless we have almost no idea what's going on in these systems.

**BEN:** They're extremely complicated. And despite having such incredible access to them, we still have a long ways to go in terms of understanding them. And so for that reason, I think of it as a first step for trying to understand the brain. If we can understand these models as physical systems, then hopefully we'll have some direction for trying to understand our own brains.

**CRAIG:** And how then did that lead you to writing a paper on scaling and data pruning? Was it through the lens of large language models?

**BEN:** There have been all these algorithmic and architectural advances In machine learning over the last 10 years.

**BEN:** , but the dominant trend has been simply scaling things up. Training larger and larger models on larger and larger data sets. And in some sense that's the simplest thing you can do. But nevertheless, it's yielded consistent improvements.

**BEN:** And that's part of the reason , you'll hear people evangelizing about how scale is all you need to reach a human level intelligence. Right? And I don't know whether that's true, but more recently, a growing body of work has tried to study the scaling empirically, and they found that performance of models tends to improve like a power law with the number of training examples you give or the number of parameters in your model.

**BEN:** And so what I mean by that is if you look at the models error on some task, that error tends to fall off. Like the number of training examples, you give it to some exponent. That's the power law. And so this is really good news in the sense that performance just improves consistently. And we can expect to keep improving our models by doing the simplest thing of just training them on more and more data.

**BEN:** But it's bad news in the sense that power laws are really slow. And to give an example, if you have a model trained on a million training points, and it achieves a 2% test error on some task you're evaluating it on. If you wanna shave that 2% down to 1% test error, you often have to double your dataset size to 2 million or more realistically, maybe 10 million training examples or something like that.

**BEN:** So you get diminishing returns with these power law scalings, and if you believe that the future of machine learning will look as it looks now, like simply scaling things. Then you should believe that the most important thing to do is to understand how we can improve that scaling, improve, or even beat those power laws .

**BEN:** these models take tremendous compute to train, millions of dollars and tons and tons of CO2 emitted into the atmosphere. And so if you wanna be able to save those kinds of things, I think it's important to think about how you can try to scale things up more efficiently.

**CRAIG:** That's important because we're pretty much reaching the limits of scaling in terms of electrical consumption at this point, there's a backlash among the public and the media toward these massive language models that take tremendous amount of electricity to train.

**CRAIG:** So is that the primary motivation

**BEN:** Yeah, and also just a general interest in understanding how these models work as physical systems . physicists get very excited when they see things like power laws, which are indicative of some underlying simplicity or some kind of critical behavior or something like that.

**CRAIG:** That's an interesting thought, an underlying simplicity. The original large language models were dense models, and then people began coming out with much larger models, but they were sparse models.

**BEN:** There's this whole complimentary field of pruning model parameters and trying to produce sparse models by getting rid of weights rather than data pruning, which is getting rid of training examples.

**BEN:** So you take a model with a fixed number of parameters. There have been a number of recent works, including some nice works by some of my colleagues where they've shown that if you take a standard data set like an image recognition data set, a lot of the training examples in that data set are redundant or uninformative, and you can throw them out or prune them away from your data set without sacrificing any performance in the trained model.

**BEN:** They've come up with very clever ways for picking which examples to train on and which ones to prune away. And if you choose them cleverly, then you can construct much, much smaller data sets. The standard data set is Cifar 10, and you can prune it down 50% without sacrificing any performance.

**CRAIG:** Andrew Ng who I've had on the podcast is now focusing on the quality and quantity of data, and he says that, you can do computer vision system on as few as 50 training examples if you have a very specific task and

**CRAIG:** so his argument is we should be looking at the quality of the data, not just scaling up randomly. Is that the direction that you're going ?

**BEN:** That's exactly, the motivation for this work, and we wanted to understand when and why this kind of data pruning works in practice, when you should expect to be able to train on only a small number of examples from your training set.

**BEN:** And to understand that we set out to write down a theory for data pruning. And to do that, we have to restrict ourselves to the very simplest possible models. So we study these extremely simplified models that don't look anything like deep neural networks. They're linear models or sometimes called perceptrons and

**BEN:** there we can use some of these techniques from physics, from statistical physics, like replica theory, and using those techniques we can write down analytical expressions for the performance of these models as a function of how many examples we prune away, for instance. And so this is one major part of the work, and spare you the details of the theory, but there's one or two

**BEN:** interesting predictions, I think from this extremely simplified model that we can go and then test on real deep neural networks trained in realistic settings.

**CRAIG:** Are you looking at the features in the data set, in the single data examples?

**CRAIG:** I saw on Your paper you talk about ranking them, and then , depending on what direction you're ranking, you take the top however many and train the model on that and throw away the rest. But in order to do that, you need some way of measuring the efficacy of what you're looking for in a

**CRAIG:** piece of data and you need to do that statistically, right? Or I suppose if computer vision , you could use ai, a computer vision algorithm to go through and figure out which images have the features that you need to train a particular model.

**BEN:** Right exactly. That's what we do and what others have done before.

**BEN:** If you take, for instance, a data set of images and you wanna learn to classify different types of animals or something like that, maybe you've already seen a lot of white swans before. And so having another image of a white swan in your data set doesn't help you much. You can throw that one out.

**BEN:** And the ones you wanna keep are the sort of rare or informative ones like the black swan,

**CRAIG:** and so how do you do that? Is there some algorithm that measures the standard deviation or something within a data set, and outside of that,, anyway, you explain it.

**BEN:** Very much as you guessed. I think, and this isn't our work, this is work that predates us, has come up with algorithms for selecting which examples to keep and which ones to throw away. But they mostly involved training other maybe smaller or less powerful models which learn something about the data. And those smaller models can decide which ones are the easy examples and which ones are the hard examples.

**BEN:** You can rank each example in your data set by how difficult it is for this smaller model. And then if you have enough of them, the best thing to do is to throw away the easy examples and give only the hardest examples to the model you'd ultimately like to train.

**BEN:** These neural networks pretty quickly pick up on the coarse features they need to classify lots of examples, but to learn the really fine grained features that you only get from rare examples in the long tail of your data distribution requires a lot more data if you just naively, randomly sample new data points.

**BEN:** But these methods are trying to seek out those rare examples, seek out the hardest ones to feed into your model.

**CRAIG:** And so how

**CRAIG:** do you do that? In your paper you say you use supervised learning. Can you explain that?

**BEN:** Sure. Actually, so a lot of the previous works have used supervised learning algorithm.

**BEN:** Which require having labels for each image to decide which ones to keep. But one of the predictions from our theory is that the really big improvements we expect from these sorts of data pruning active learning strategies will come at at much larger scales. Data sets much larger than the standard data sets we look at today, like Cifar 10 and ImageNet and those data sets often don't come with labels. So another contribution of our work, and this is mostly due to my colleagues on the paper, is to come up with a new self supervised algorithm for selecting which examples are easiest and which ones are hardest. And this we hope might scale up to these really large data sets where we don't have access to the labels for each image.

**CRAIG:** And self supervise, what, is that through reinforcement

**CRAIG:** learning or?

**BEN:** it's through just simple sort of contrastive loss, where you feed your model two different views of the same image, maybe two different random crops, or two different distortions of the same image, and you ask the model to decide whether those two distortions come from the same image or from two different images.

**BEN:** Somehow if you train a model with this kind of objective, it learns useful features for solving a lot of downstream tasks. And also by training a model. In this way, you can learn whether an example is easy or an example is hard,

**CRAIG:** What size data sets are you applying this to, and can you give an example of how tightly you've pruned a data set and come up with equally strong results

**BEN:** So maybe just to briefly summarize the results of this theory, our prediction, is that sort of the best strategy for pruning your data set depends on how much data you have, and we find that if you don't have much data, you want to keep the easiest example. And if you have quite a lot of data, you wanna keep only the very hardest examples.

**BEN:** And so we can write down in this simple setup that I described earlier, an exact theory for what the optimal strategy for pruning your data looks like. And if you follow this optimal pruning strategy, what we find is that you scale much faster than those power laws I was describing before.

**BEN:** Those power laws have very slow scaling, but if you do this sort of optimal pruning, You can beat those power laws and actually achieve something like exponential scaling or even faster than exponential scaling, which is a tremendous speed up and part of the reason we're so excited about this line of work. This of course, though, is all in this highly simplified theory with linear models that don't look anything like deep neural networks in practice.

**BEN:** So we have this prediction from these simple linear models and then we go on to test it. Mostly this is done by my colleagues who have access to quite a bit more compute than I do, and they test this prediction in real deep neural networks trained on real object recognition tasks.

**CRAIG:** You're doing primarily computer vision right now?

**BEN:** Yeah.

**BEN:** Most of the experiments we've done so far are in computer vision, but the direction we'd like to go is into these large language models. In computer vision, the standard data sets are data sets like ImageNet, which has about a million training examples or something like that. And there even the best data pruning metrics can only prune down to not more than 50% before you start to sacrifice performance of ImageNet.

**BEN:** So while substantial and exciting these data pruning methods, haven't really taken off . One prediction from this theory is that a lot of these methods haven't really taken off because so far we've been studying these sort of smaller data sets with only say a million or so training examples, and there you can

**BEN:** cut down about 50% of them. But a prediction from this theory, because it's exponential rather than power loss scaling, is that the really big returns, the really big savings in compute or tons of CO2 will start to come once you look at really big data sets with billions of examples or trillions of tokens or something like that.

**BEN:** And so in ongoing work, we're trying to really test this prediction at scale. So we're training some of these very, very large language models, trillion parameter language models on very, very large data sets with billions of, or trillions of examples to see how long this exponential scaling might hold up.

**CRAIG:** That's fascinating. And you're doing that training now?

**BEN:** Yeah. Well, I'm not doing it myself. I don't have access to enough GPUs, nor do I have probably the coding skills to do it. My colleagues are.

**CRAIG:** And are they building a large language model to run these tests on? Or are they using an existing large language model?

**CRAIG:** And if so, who's?

**BEN:** They're taking a fixed model architecture and training it on ever larger dataset size. And then using these pruning strategies to cut down on the dataset size to see if we can achieve faster than power loss scaling in the test loss with the dataset size.

**CRAIG:** When do you think you'll see whether or not that's achievable?

**BEN:** I was speaking with some of my colleagues today who've been training these models and they're pretty excited about the results so far. I don't know when we'll have anything conclusive to show, but hopefully in one of these upcoming conferences, we'll have something to report on.

**CRAIG:** And if that's the case, then that sort of gives large language models a new lease on their lives. Right? I mean, you suddenly can train them much faster , there'll be an incentive to build more of them. I mean, one of the disincentives is the cost of training.

**BEN:** Right. Yeah. I think the future of training these models will be quality of the training data over quantity of the training data.

**BEN:** I think that's really where you'll start to see these speed ups and also I think will be very interesting from interpretability standpoint to understand which kinds of training examples these models care about, which sentences in the text the models pick up on used to train. Which images in your image recognition data set are the crucial ones.

**CRAIG:** Can you give some metrics of how far you expect, the data sets to shrink and maintain performance?

**BEN:** Yeah, so from this theory, there are sort of two bottlenecks to this data pruning strategy. There's two places it can go wrong.

**BEN:** One of them. Is the quality of your metric for deciding which examples are easy and which examples are hard. And what we find is that, at least in our theory, if you have a perfect metric, an Oracle or something like that, that can decide with complete confidence, which examples are the easy ones and which ones are the hard ones, you can achieve this kind of exponential scaling forever.

**BEN:** But in real life, we don't have that kind of metric. We only have a poor one. And when you have a poorer metric, you'll only achieve exponential scaling for some time. And eventually you'll settle back onto that same asymptotic, power law scaling. And so I think another very promising direction for future investigation will be finding better metrics for deciding which examples to train on which ones are the important ones and which ones to throw away.

**BEN:** I should say that the metrics we come up with in our work are really very simple, the very first things we thought of and

**CRAIG:** such as

**BEN:** such as passing examples into your model, collecting their representations in the hidden layers of your model and clustering those and keeping the outliers and throwing away but that the sort of more typical prototypical one.

**BEN:** That's the kind of strategy we use, and it's very simple and there are reasons, I think, to believe that we can do a whole lot better than that. And the theory tells us that if we can do better than that, then the gains that we'll achieve are really quite significant. So I think very exciting direction for future work will be coming up with better metrics for how to decide which examples to keep and which examples to throw away.

**BEN:** So this is operating on unlabeled data. What's the largest computer vision data set available?

**BEN:** I think the largest image recognition data sets was actually just released at NeurIPS .

**BEN:** It's called LAION-5 billion. I think they have something like 5 billion images or something like that.

**CRAIG:** Okay. So if you took LAION-5B and you were training a large language model, if your theory bears out. How much do you think you could prune that, to what size do you think you could reduce that and still have performance?

**BEN:** There's an important question of how long this exponential scaling will hold up, and I think we'll only know the answer once we go and actually do these experiments. But the most exciting prediction of this theory is not a quantitative one. Not that you can prune X percent of your data, but a qualitative one, which is that the scaling, the power law scaling we're used to can be qualitatively beaten and achieve exponential scaling, which you can compare the performance of

**BEN:** different strategies on any fixed dataset size, but I think ultimately the thing we care about is how these models will perform in the far future as we continue to train on larger and larger and larger data sets. And that's where this qualitative prediction, I think really makes a difference.

**CRAIG:** Okay, let me get my head around that.

**BEN:** I'm sure I could explain that better.

**BEN:** I mean by quality, your model's learning curve, how well its performance improves with the size of the data set you train it on. Different strategies, will have different learning curves, and so far, most models seem to have a power law type learning curve, which is a very slow learning curve. And so when I say qualitatively, I mean that with this type of strategy, we believe we might be able to achieve an exponential learning curve.

**BEN:** Which is a tremendous speed up over a power law learning curve, and that's the qualitative difference I'm talking about. You can look at the improvement of an exponential over a power law for any fixed dataset size and get some number, but really the important thing is that as you continue to increase that dataset size, that number will grow and grow and grow.

**BEN:** The speed up will grow and grow and grow. The amount of compute you can save by using one of these other strategies only increases as the size of your data set increases.

**CRAIG:** How does this relate to your original interest of understanding the

**CRAIG:** brain?

**BEN:** Well, understanding the brain, I feel, unfortunately, we're still a long way off. But what does excite me about this kind of work is that we can study very simple models like this linear model I described using some powerful techniques from physics and understanding and studying these simple models can actually lead to predictions about how big, deep neural networks will behave in realistic settings that we can then go and measure.

**BEN:** And my impression of the theory of deep learning has mostly gone in the opposite direction. Where we observe some interesting phenomenon in a deep neural network, say your listeners might be familiar with the phenomenon of double descent, for instance. This was a very intriguing and surprising feature of deep learning, where if you train larger and larger models with more and more parameters, classical, statistical wisdom would suggest that you'd start to overfit your data.

**BEN:** Once you have more parameters, more knobs to tune, then you have training examples. But in reality, what happens is that you don't overfit surprisingly, somehow, even if you continue increasing the size of your model, you can learn generalizable features. And so this was shocking to a a lot of people in the field, I think.

**BEN:** And then we went back and looked at the simple models and discovered that even linear regression, even kernel regression exhibit, the same sorts of behavior. So you can find signatures of that, even the very simple model.

**BEN:** So all it is to say, I think a lot of the theory of deep learning has been post talk. We observe an interesting thing and then we go back and find a model to explain it.

**BEN:** And it's kind of exciting in projects like these to study the simple linear models and notice intriguing phenomena there, and then make predictions about how those phenomena will hold up in real deep neural networks trained in realistic settings.

**CRAIG:** Just thinking about the brain, this begins to address one of the paradoxes of the brain versus deep learning neural nets, and that's the brain generalizes on very few

**CRAIG:** training examples, right? So do you think that you're moving in a direction that would explain how the brain can generalize on so few examples?

**BEN:** That'd be very exciting. If we were, and I think it's at least a little bit inspired by the brain, or at least by human psychology, where we seem to be able to pick up new tasks with very few examples.

**BEN:** And I think part of the reason is that we seek out the hard or interesting or surprising examples. there's all this work in the psychology literature about humans, even infants being driven by curiosity or surprise. They're drawn toward rare or surprising or intriguing behaviors in the physical world.

**BEN:** And don't just engage with the world passively taking in random inputs. And so I think. You know, an example where we might get some inspiration from psychology or for neuroscience for training deep neural networks by, rather than just randomly presenting inputs, presenting the most uncertain or difficult or intriguing inputs to try to get the models to learn faster in the same way that humans do.

**CRAIG:** Although humans, You mentioned toddlers. Toddlers. Learn what an elephant is with maybe three examples, right? I don't know. I'm sure it's been studied how many examples, but you see a couple in a book and then you go to the zoo and your mom says that's an elephant, and from then on you recognize elephants,

**BEN:** right

**CRAIG:** So you're not seeking out difficult examples from then on, you kind of refine your understanding, self supervised way I guess. The first three are supervised. I mean, does that relate at all to what

**CRAIG:** you're

**BEN:** so. I think there's a lot of hope when people who are scaling these models up that.

**BEN:** Having seen enough examples before these deep models will start to learn relevant features. And so when presented with an example of something they've never seen before, they'll be able to very quickly incorporate that among the other kinds of objects they've seen before like giraffes and zebras.

**BEN:** And this explains the surprising one and few shot abilities of these really deep, large language models. One in few shot, meaning just a single presentation of an elephant or a few presentations of an elephant. But I think also, in order to be able to generalize from few examples the way humans do, I think we probably will have to do more than simply training these models on more and more data learning richer and features.

**BEN:** Because I think beyond having a very rich set of features in our own brains, we have strategies for attending to the relevant features of the new object you're presented with, thinking critically about what distinguishes an elephant from a zebra or something like that. And so those sorts of strategies might be crucial for understanding few shot learning,

**BEN:** I should say also, I've done some previous work with Haim Sompolinsky and Surya Ganguli trying to come up with a theory for this problem of few shot learning, of learning a new concept from just a few examples. And we do it based on the geometry of representations in the deep layers of neural networks.

**BEN:** And then we take recordings from primate brains, and we look at the geometry of those representations and we find interesting connections between deep neural networks and brains.

**CRAIG:** By geometry you mean the configuration of the network or

**BEN:** Yeah, if you look at the activity you record from a hundred neurons in the brain as you present, you know, an image of a cat.

**BEN:** You'll get some point in a high dimensional space. And if you look at the different responses of those neurons to many different images of cats, you'll get many different points in this high dimensional space. And those points lie on some kind of manifold and you can study the geometry of that manfold.

**CRAIG:** Oh, that's fascinating.

**CRAIG:** Wow. So where do you think all of this is going? it sounds as though you're not gonna stop with data pruning. That's kind of a step toward a larger goal for you, but also for the field. Do you think that large language models are kind of the way toward a more general intelligence? and I don't want to get into the AGI discussion, but for example, already you have these models that are multi modal. And, are increasingly general. And so , are large language models the direction that you think the field will go in order to reach a generalized higher level intelligence? Whether or not it's human level intelligence?

**BEN:** Well, I don't know enough to speculate, but the field certainly seems to be moving in that direction.

**BEN:** These language models are just getting bigger and they seem to be acquiring all kinds of new abilities every day. And I think if you ask me what the most pressing question in my own research is, it's in trying to understand and interpret how these models work. At the moment, our understanding is extremely limited.

**BEN:** We don't know how they process input. How they make decisions. And I'm a little frightened by the fact that these models are being deployed in the world today and we have no idea how they work. And so the most pressing research question for me is trying to understand and interpret these models , using whatever lens I can, maybe studying them as physical systems because I am afraid about the fact that they're being deployed at scale and we don't really know how to peer inside them and understand what they're doing.

**CRAIG:** I'm a little puzzled by why people are afraid of them . is it that they're going to be used to inform decisions and the decisions are going to be based on faulty inferences? Is that the

**CRAIG:** fear?

**BEN:** I think so, and they already are today being used to make decisions.

**BEN:** And when we talk about the quality of data versus the quantity of data, the data we feed into these models itself is biased, independent of whether we bake any bias into the models ourselves, just by collecting examples at random from the web. We're presenting our own biases to these deep models, and so they can pick up on prejudices and things like that.

**BEN:** And when we use these models to make decisions, those prejudices which arise from bias in the data are baked into the outputs.

**CRAIG:** Although large language models aren't really being used to make decisions right now, are they?

**BEN:** But in deciding what kind of content we present to people on social media or on news platforms, what kind of ads, things like that.

**BEN:** I think those are all kinds of decisions that can affect the way people behave.

**BEN:** I do have a variety of projects.

**BEN:** Most of them are centered around trying to understand what these deep neural networks learn and how these deep neural networks learn. And the hope is if we can extract some principles about how they learn, maybe we can start to extract some principles about how the brain learn. And I'm also really lucky to collaborate closely with experimental neuroscientists who are doing recordings in mouse and primate brains every day.

**BEN:** And we're trying to take some of the lessons we learned from these deep neural networks and look for insights about how our own brains work.

**CRAIG:** If this line of inquiry is successful, won't it revolutionize the field of large language models? Cause suddenly there will be no constraints on costs and electricity use.

**BEN:** Well, I think, if not our theory or our algorithm, something like it probably will change the way we train these large language models in the future.

**CRAIG:** That's it for this episode. I want to thank Ben for his time. I also want to thank our sponsor, clear ml. If you're building machine learning models, check them out at clear.ml.

**CRAIG:** That's C L E a r.ml, and there's always, you can find a transcript of this episode on our website. That's E Y E hyphen O N A. I find that transcripts deepen my understanding because the eye catches information that the ear does not. And remember, the singularity may not be near, but AI is about to change your world, so pay attention.