Welcome to the Big Ideas Monday mini-series, brought to you by the For Your Innovation podcast.
Big Ideas is meant to enlighten investors on the long-term impact of innovation.
This annual research report seeks to highlight the technological breakthroughs evolving today and creating the potential for super-exponential growth tomorrow.
We believe that innovation is taking off now, corroborating our original research and boosting our confidence that our strategies are on the right side of change.
To learn more, visit arc-invest.com.
Arc Invest is a registered investment advisor focused on investing in disruptive innovation.
This podcast is for informational purposes only and should not be relied upon as a basis for investment decisions.
It does not constitute either explicitly or implicitly any provision of services or products by ARC.
All statements may regarding companies or securities are strictly beliefs and points of view held by ARC or podcast guests and are not endorsements or recommendations by ARC to buy, sell, or hold any security.
Clients of ARC investment management may maintain positions in the securities discussed in this podcast.
Welcome to the Artificial Intelligence Overview of Big Ideas.
I'm Will Subberlin.
I co-lead venture investing and also cover AI on the research team at ARC.
And my name is Frank Downing.
I'm the director of research for ARC's next generation internet theme.
So this was clearly an exciting year in AI.
And this presentation will cover three topics primarily.
The first is generative AI.
The second are cost declines and are forward expectations for cost declines through 2030.
And third is the impact in AI on the productivity of knowledge workers.
So starting with generative AI.
Generative AI has come a really long ways in the last few years.
If you go back to just a few years ago and asked a neural net to produce an image of say a monkey on the moon, the image that was produced would be rather terrible.
It would be grainy and it would probably not look like a monkey on the moon.
But in the last few years, we've seen a lot of architectural advancements in AI.
Transformers being one great example.
And importantly, the cost to train large AI models has continued to decline at a very rapid rate.
And that's allowed organizations like OpenAI to train much larger models that are much more performant and capable.
We've seen this in the image space.
Look at Dolly 2, for example.
In this presentation, we give an example of a prompt asking to create a picture of an astronaut riding a horse.
And you can see here the image that was generated by the neural network.

You can actually use these models yourself.
You can go to Dolly 2 or many platforms that allow you to train or use models like stable diffusion.
But we've also seen generative models beyond images.
For example, META released a make a video model where you can provide a text prompt and from that text prompt, the neural net will actually generate a video.
So really incredible capabilities.
We're also seeing innovations on the 3D front and the audio front.
And we expect this pace to continue for the next few years.
While these models are really interesting and cool and fun to show your friends, they're actually impacting the productivity of knowledge workers already.
In the case of coding, we've seen AI coding assistants like GitHub co-pilot more than doubling the productivity of software engineers when it comes to coding tasks.
And so these models are abstracting away some of the more redundant tasks that would be done by software engineers, making them more productive.
We've also seen examples on the design front.
We did a survey and found that if we wanted to have a graphic designer create a relatively complex image, it would cost about $150 in human labor.
We can ask a neural net to create a similar image.
And after it generates a few images, getting a satisfactory one costs all in about $0.08.
We expect that cost to continue to decline.
And so in the case of image models, we've seen a cost decline from $150 in unassisted human labor down to $0.08 with generative AI.
Now to talk about cost declines, I'll hand it over to Frank.
Thanks, Will.
A great example of cost declines as applied to image generation.
But if we step back and look at large language models as a whole, we're always looking at cost declines to try and understand how markets are going to involve around innovative technologies.
And AI is a great example of a technology that is very compute intensive and thus very expensive to train and run inference on these large models right now.
And one of the things we've been really shocked by is the cost declines that we're seeing already with AI just in a short term time horizon.
So if you look at GPT-3, which is OpenAI's model that powers chatbots like ChatGPT, that original base model cost $4.6 million to train in 2020.
And we've already seen a company called MosaicML demonstrate that they can train a model to similar levels of performance for just $450,000 in 2022, just two years later.
And so that comes to about a 70% annualized cost decline.
And if you compare this to what we put in our Big Ideas presentation last year, where we thought costs were declining at about 60% per year, that would have been a $740,000 training cost.
So this is one case where the actual numbers that we're seeing are outpacing our original cost decline forecast from last year.
To double click a little bit further here, what we like to do to understand where these

training efficiencies are coming from is split out hardware cost declines and software cost declines.
On the hardware side, we see cost declining at 57% per year.
So this is through a variety of innovations, not just Moore's law, but building on top of adding more transistors on a chip by adding more high speed memory and increasing the number of what are called tensor cores on a chip.
These are compute cores designed specifically for training AI models that allow, over time, these chip architects to build more and more efficient chips, specifically tailored for the task of AI training.
On the software side, we're seeing similar advancements as AI is becoming more prolific and there's more and more research being poured into it.
There's various algorithmic tricks and efficiencies that AI researchers are coming up with that have shown to decline costs at 47% independent of the compute.
And when you combine these two forces that are moving in parallel, you get this 70% cost decline that we think, based on our research, could sustain for the next 10 years through 2030 and beyond.
So these cost declines, which we think will open up a broad market for AI, they beg the question, what will we do with this cheaper compute?
We won't just be able to run more inferences, but we'll be able to train larger models.
If you look at GPT-3, which is already a large model at 170 billion parameters, there's a question of how big models will get in the future and what data will be needed to train those models.
DeepMind, an AI inside of Google, published a paper regarding a model they created called Chinchilla in the beginning of last year that seeks to come up with a scaling function for how large models can get based on the amount of training data that we can feed them.
And based on our estimates, the largest model they put forth in that paper is a hypothetical 10 trillion parameter model.
That's 57 times more parameters than the GPT-3 model and 720 times more tokens as this largest kind of theoretical model that they put forth.
If you overlay our cost declines on that, we think that model, which is orders of magnitude larger than GPT-3, could be trained for just $600,000 by 2030.
So that's just 13% of the cost to train GPT-3.
Now the trick there is 216 trillion tokens of training data.
So that's larger than the total size of Wikipedia, for example.
There's going to need to be a lot of language data generated to feed models once we get training capacity this large.
I think that brings about an interesting question in terms of what is most important with getting an edge in AI training that's particularly related to proprietary data advantages.
And I'll turn it back to Will to talk a little bit more about that.
As Frank mentioned, we believe proprietary data is really important, both in terms of training a model, but also in terms of fine-tuning and continually improving a model.
When we think about a use case like self-driving, data is really important.
A company like Tesla, in our opinion, is in an advantageous position, in part because of the data that they've acquired.

They have millions of cars on the road driving billions of miles, and they're using those cars to continually collect more data that improves their full self-driving capability.
For example, if a driver presses a brake while autopilot is engaged, Tesla marks that as an error on the part of the AI.
They can then use that example to continually train the model and further improve it.
We believe data advantages in this feedback loop exist in other verticals as well.
For example, in software development, AI coding assistants are continually improving based on the feedback they're getting from usage.
So a company like Replet that has millions of software engineers on their platform and is continually improving their AI models, we think will be in an advantage position moving forward.
When we think about the impact of cost declines on model performance, we look at a variety of applications of AI, one of which is AI coding assistants.
As we mentioned earlier, AI coding assistants are already doubling the productivity of many software engineers, and when we take that 70% cost decline and model forward performance expectations, we think that the models will be performing enough to more than 10x the productivity of software engineers by 2030.
When we talk to researchers who are building these models, or Amjad, who is the founder and CEO of Replet, they actually think that a 10x uplift in productivity may be too conservative, but based on our cost declines, we are confident in our base case of 10x.
To give another example of how we think cost declines will unlock new markets, I think it's important to touch on chat GPT and similar sophisticated AI chatbots.
So chat GPT has exploded in popularity, reaching over a million users in just its first five days.
And part of the reason why is because you're able to ask in very natural dialogue a wide range of questions that the AI can answer based on the corpus of public internet data that it's trained on.
To give a sense of current performance, and this is changing rapidly as kind of new iterations of these chatbots are released, chat GPT when launched could already answer SAT questions at around the national average, which is pretty incredible, and that's only getting better as more training data is fed into these models and it gets feedback from real use.
And this is already being productized and launched by Microsoft and Google as competing product coming to market as well.
The question here, the interesting dynamic is doing this inference on these large language models is relatively expensive.
So we've estimated the cost of one query on chat GPT at about one penny.
It may not sound like much, but when you compare it to the 8.5 billion searches per day on Google costs that up very quickly.
So if you were looking at doing about a billion inferences per day, right now we think that would cost about $10 million per day, and that's not even the total scale of Google.
But by 2030 using our cost declines, we think that will come down to about $650.
So a much more reasonable number that will allow these chatbots to be deployed at scale that we think is going to happen quite quickly.
When we think about the impact of AI on knowledge worker productivity, in our base case we

believe
AI will increase the productivity of knowledge workers by more than fourfold by 2030.
In our base case, that would result in companies spending about $14 trillion a year on AI software, and we believe that could create more than $90 trillion in enterprise value.
In our bull case, at 100% adoption rate, we think companies could spend $41 trillion on AI software to increase labor productivity by more than $200 trillion.
That's compared to the $32 trillion they currently spend on knowledge worker salaries today.
So we believe that AI could actually break consensus GDP estimates given the potential for knowledge worker productivity.
That's all.
Thank you for listening to the AI section of Big Ideas.
ARC believes that the information presented is accurate and was obtained from sources that ARC believes to be reliable.
However, ARC does not guarantee the accuracy or completeness of any information and such information may be subject to change without notice from ARC.
Historical results are not indications of future results.
Certain of the statements contained in this podcast may be statements of future expectations and other forward-looking statements that are based on ARC's current views and assumptions and involve known and unknown risks and uncertainties that could cause actual results, performance
or events that differ materially from those expressed or implied in such statements.