

(no title)

34 MIN · YOUTUBE · [HTTPS://WWW.YOUTUBE.COM/WATCH?V=0DJ7NDSKNTS](https://www.youtube.com/watch?v=0DJ7NDSKNTS)

<https://www.youtube.com/watch?v=0DJ7NdsKnts>

SUMMARY

Apoorv, the instructor for a 9-week course on AI, shares his background and outlines the course logistics, emphasizing the importance of understanding the economic landscape of AI. He discusses the current state of AI investments, the profitability challenges in the ecosystem, and the anticipated evolution of AI applications and infrastructure.

- Apoorv has a diverse background in AI and investment, having worked at Palantir and now leading Altimeter.*
- The course will include guest speakers from various AI sectors and will focus on the economic implications of AI technologies.*
- Key topics include the profitability of AI applications versus infrastructure, with Nvidia dominating the semiconductor market.*
- The course aims to equip students with the knowledge to evaluate AI business models and investment opportunities.*
- Apoorv encourages active participation and questions to enhance learning.*
- The current AI ecosystem is compared to previous tech revolutions, highlighting its early stage and potential for growth.*
- The discussion includes the challenges of monetizing AI applications and the potential shift towards ad-based revenue models.*
- The course will explore the dynamics of AI market players and their strategies for profitability and growth.*

All right, folks. we're going to have some fun in this session. my name is Apoorv. I'm going to be your instructor for the next 9 weeks or so. here's what we're going to go through today. I'm going to talk a little bit about myself. Why do I do this? Some logistics on what to expect. Quiz, yep. I'm that guy. We're going to have quiz on day one. And the biggest question that we've all been wrestling with, where's the money? Where's the money in AI?

Some of you know me, but you know, my journey started in India. I moved to Singapore. I met a couple of Singaporean folks here earlier. I started my career at Palantir with Sunil and a couple of other folks about 13 years ago, 14 years ago. Led a variety of engineering teams. all that to say, we wrote a lot of Spark in government buildings. And I came back to Stanford for grad school, which is when I got tired of writing Spark in government buildings.

And now I lead Altimeter. I don't know how many of you have heard of Altimeter, but Altimeter is an investment firm. We focus on fairly concentrated form of investing. We've got two businesses. We got a public business and a private business. And then I got the biggest promotion of my life 6 months ago. I'm now the proud dad. Woohoo! this is the as as people have told me, the biggest investment I will make. Some have called it

the one with the most guaranteed negative IRR.

I think it's the most guaranteed positive IRR, not financial. but that's me. I live across the street. reach out with questions. I want to be here as a resource to you guys and make the most of it. we've got a great session lined up for you guys. The course is designed to be no more than 3 hours a week. and that includes class, readings, all the time you spend arguing with chat GPT, with Claude, about whether it can do your assignment for you. So, it's about an hour of class.

it's about an hour or two of readings. And basically the format is we're we're going to do guest speakers every single class next class onwards. Adam House rules, a lot of guest speakers will will share, maybe overshare. So, please don't record, what they're what they're saying. we'll have an optional dinner with some of them after right afterwards. You guys are welcome to join. arrange the the logistics. grading is easy, 50/50. Show up to class. If Ali Gots he can show up to class, you can show up to class.

And the other half is an assignment that we will release at the, end of the course. yeah, it's conversational. Ask questions, be involved. The more you're involved, the more you're going to get out of it. And honestly, you know, what's in it for me is I'm going to learn the most from you guys. This is the course schedule over the next 9 weeks. Lots of great speakers, one more impressive than the other. As for me, you know, honestly putting this calendar together has been my my job for the last couple of couple of weeks and months.

but be be present. These are all incredible leaders running incredible businesses across the stack from semis to to to infra to energy on on on the infrastructure side to models. You know, we're going to have folks from OpenAI and Anthropic and and a bunch of applications and agents. so, ask all your hardest questions. Save them for the speakers. they're going to love it. and we're going to assign some readings. So, why should you take this course? What should you achieve in this course?

what is a good thing to get out of this? You know, honestly, I I thought about this and you know, I was just telling one of the students here of how it all began. It's you know, I come back to campus once a year and I talk about all that's happening and you know, typically it's in the context of finding great people. It's I realized that this is such a big super cycle. We know it. We We believe it at Ultimeter, you know, we have positioned our entire focus around it.

And I did not find a course that goes deep in a way that I would have liked to be when I was an undergrad here or or a grad student here. And you know, I thought about in 5 years, everybody's going to ask you, "Hey, did you see it coming? You were at the start of it. You were around when ChatGPT was launched. You were around when, you know, the the the tectonic plates were forming and clay was forming and I think you want to be able to say yes."

So, half of you are going to start an AI company. The other half are going to fund it. So, at least you should know where to spend the series A money that you're going to raise. And at the very minimum, you know, you'll have a sense of what not to go or at least have mental models for, "Hey, this business that I'm looking at or

considering starting or considering funding or considering joining, what are the right questions to be asking? What are the laws of physics that govern this this business at this at this part of the cycle?"

I think it's going to be the biggest one yet. And I'm excited that you guys are here to study alongside us. I'm going to spend some time on the slide because this is the this is sort of the punchline. how many of you have seen a version of this before? Well, those of you did the readings, thank you. I appreciate it. We did include a notebook LM for those who were more auditory inclined. But let's talk about this for a second.

you know, what is going on here is is actually the probably the biggest the biggest question in generative AI right now, which is the if you listen to any of the earnings calls from the hyperscalers or or or even Nvidia and and and and others is we are investing so much into the CapEx. We're We're investing so much into building these data centers. We're you know, it's a five-layer cake as Jensen calls it.

Energy, chips, power, interconnects, memory, all that to give you a data center that you can either rent by the hour or by the token that you can go train models on and serve those models. And then the question is, hey, these models that you built, are they creating economic value? That is basically the right-hand side of this chart. And to to to make an analog of the biggest technology revolutions that I have seen, you know, internet 25 years ago, mobile 20 years ago, cloud, probably the most recent one, 10 years ago.

I put up one of those charts on cloud, but you know, under readings you'll see the same for internet and mobile and and cloud. And that's the shape of the cloud ecosystem. The cloud ecosystem looks dramatically different than the AI ecosystem. Anybody have guesses as to why that's the case or your theory on why it's so different or reasons why this might look like We're not going to call it a pyramid. We're going to call it a triangle. In what a triangle?

Go ahead. Is that because it's still early in the cycle for AI? Yeah, definitely. Definitely early. That's a good guess, yeah. Any others? Any other thoughts? Well, maybe cuz Nvidia has like a monopoly, so they can charge whatever. Can you ask that question again next week when we have the when we have the folks from Nvidia? But no good. It's a It's a great point. They do have a stranglehold, right? one of the one of the charts we had in the readings was the market share that Nvidia has on all of the compute right now and it's it's up there.

Any other thoughts or hypotheses on why this is so different? Yeah, I don't know. Is it the the cloud sector seem to be able to leverage that hardware to generate value for the apps and then the AI in Nvidia being You know, we know how software ate the world, as Marc Andreessen said. Software ate the world because, you know, I could build software, could build software, and I could distribute it to millions of people, and the marginal cost of running that software was close to zero.

These software businesses ran at 80, some even at 90% gross margins. That is not the case with this new economic model of AI because if we have a set of users using Cursor or using you know, you you hear all these stories about large bit scale businesses that are still not profitable at billions of dollars of revenue scale is because of that. Is because the incremental user of an AI application is not free. It's not marginally free. It's actually quite

a bit more expensive to have AI users because turns out you've got to burn those GPUs.

And I would say everything you guys said from it being early to Nvidia being dominant, we'll we'll call it, to you know, the the physics of the problem are very different of how inference is run. It's certainly where we are right now. So, I think that's the case right now. I might add another dimension to it, which I spoke about in the readings, was you know, we analyzed what happened in internet. We analyzed what happened in mobile and cloud and how many years did it take to for these triangles to flip?

And you know, one of the examples we take is AWS. AWS started in the year 2004. AWS has its first customer in Netflix in 2010. And ultimately Amazon shifted fully to AWS in 2012. Eight years from breaking ground, eight years from first CAPEX investment cycle. I don't know if any of you were around reading Ernest earnings reports 20 years ago, but the big debate was, "Hey, is Amazon going to go bankrupt?"

And that was the biggest that was the biggest question everybody had about the buildout of AWS and you know, thankfully nobody at least yet is on the verge of bankruptcy, but these are large numbers. So, we'll we'll go back to this slide, but I would say it this is the central theme of the course that we're going to explore. We're going to have speakers from some of the company companies that are listed here to others and the central theme that we're going to pick around is like, "Hey, in your with the Nvidia speakers, are you a dominant force? How long are you going to stay to be the dominant force? What are the forces that you're most worried about?

Who are the ASICs that you're most worried about? What are the pricing compression vectors for your business? To you know, the folks at Anthropic and OpenAI, we're going to talk a lot about profitability. Is your serving a billion user franchise at at OpenAI? With the with the Anthropic folks, honestly, 100% of this class is on on Claude, so we'll ask them about, "Is this group of users profitable? How do you think about profitability? is ads going to be a bigger source of revenue than subscriptions?" And then for the folks in the middle, which is the inference layer, you know, this is the most competitive part of the whole ecosystem. There's a lot of startups that are doing really well. They're winning so far.

But you've also got the hyperscalers who want to have a dominant say in that layer. So, honestly, the jury's still out. And the biggest question there is are you a feature or a platform? A lot of new in you know, businesses that we're seeing on the infrastructure side that they feel like very good ideas, but you if you ask yourself the question, "Hey, why is this not a part of AWS?" you're thinking about maybe it should be a part of AWS.

So, for the speakers, they will we're going to talk a lot about that. Any questions before we jump into the quiz? Go ahead. I'm curious how you think about so on the right-hand side like the triangle being like the application layer being small. Like how do you think about including like incumbent platforms into that? Like maybe like Salesforce, maybe they're high revenue. Like would you include them as part of like analyzing that pyramid and how it shifts over time?

It's a great question. And I might add, you know, Salesforce, Palantir. There's a series of let's call them old economy businesses that are reinventing themselves to have SKUs of products that are, you know, in the case of

Salesforce, Einstein. In the case of Palantir, AIP. In the case of you know, there's a series of these. And the answer is yes, they should be. The answer is yes, that they should be. The way I solve for that in this calculation is I get the model revenue.

And so, if you were running Salesforce, you're probably running either one of the big models or running inference. So, their spend is captured in the app layer by way of the substrate. it's very hard to extract that out from public disclosures, but yeah, we should. Yeah. Is a launch part of the bottom part of the dioxide pyramid which basically is buying capacity for future revenue which will be at the top which is what we're not seeing in the brochure. It went well. It's a great question.

and maybe just to rephrase the question. The question is hey is there a timing mismatch in the build out of the semis layer because typically you build semis for a 5-year period or a 6-year period. but the application revenue is for right now. it's a great question and that's what makes the lower half of this call it triangle somewhat cyclical and you go through phases of capex cycles. think of it as like laying down the rail railroads. that is very much the case and so there's a chart in the readings for what happened in the mobile super cycle. Something very similar happened.

The first inning had inflated market caps for a lot of the capex heavy businesses and so if you think about a basket of capex names to to call it steady state names. You you should expect that and so I suspect we're so early that that's happening as well. I'm curious to see how you think about Google because I see that you labeled Google as Google Cloud there on the digital layer. But Google also have their own Gemini models and they have their own their own GPUs instead of the fully integrated. How do you view where they are positioned in this diagram?

Any large conglomerate like Google deserves to be you know, we have to call business units. So I would put the TPU business unit in semis. we include that here as we counted the revenue. their GCP unit is in the infra infrastructure layer and then the Gemini unit unit is at the apps layer. the you know, we have a chart later that we'll talk a little bit about Gemini is actually one of the most used consumer applications. It's the second most used consumer application right now. And the biggest question there is how much of that is coming from the distribution advantage that Google has to it meritocratically being such a good application. jury's still out, but we'll get into it.

Yeah. Well, it's go ahead. I have a question about prediction. Right now, it looks like this triangle shape. And if it were to be successful, perhaps it should become inverted. but what does a an unsuccessful new new technology look like? Does it stay a triangle? And how would you predict whether this would invert or not? Yeah. I'm not sure the Maybe rephrasing your question, I I might say it like what is the stable equilibrium of this industry?

I think it is pretty clear that AI is unlikely to be a fad. It is unlikely to be an unsuccessful endeavor. And I think about the stable equilibrium of this chart quite a bit. In fact, I got into a little bit of a debate on Twitter with with with somebody quite smart and who thought a lot about this about this exact question of like what is the stable equilibrium? And you know, my guess is that it might stay this way for longer than I anticipated.

In the cloud, I think that range is about a decade. I have a feeling this might stay longer this way longer because of just how hard it is to get the substrate right. But there will be one or two unlocks. I don't I couldn't tell you what they are, but for example, if one of the ASIC programs at one of the hyperscalers be it Google's TPU or Meta's MTIA or you know, the folks at Amazon and OpenAI and Microsoft and all the labs that we don't even know about exist has a break out has break out success. I suspect that'll be the biggest repricing off that layer.

The other catalyst could be you know, I think about the hyperscaler capex guidance and earnings calls which by the way I recommend everybody here to listen to four times a year. You'll have public company CEOs tell you their biggest questions, their biggest things that they're thinking about and you know, I recommend listening to those. if they got stop guiding to big numbers on capex because that would imply that the current equilibrium does not work. so that that's why the second thing that could happen.

And so you see there's a lot of news about the guidance that all the hyperscalers give about their capex for that reason. Go ahead. Is there also an element of training versus inference? because my sense is if the only way this flips is if inference only needed to be larger than training. And and I'm curious to hear your thoughts on when you think that will happen because that you're saying that I mean these all the hyperscalers will stop spending on training because they're not seeing what is the performance that they're getting fully and hence what they're not getting on inference.

It's a great question and it is probably the one of one of the nuggets of information that Nvidia's earnings calls have the like the most sought-after nugget of what is Nvidia's share of inference in their fleet. last I checked it was about 40% or or they quoted it to be at 40% meaning that if they were selling a million GPUs assuming full utilization about 40% of them were used for inference and the other 60% for training. I suspect that number will increase over time in favor of inference.

But I I I wouldn't I couldn't tell you when and how it'll happen because there's a lot of training still going on in the world. And the shape of the training workload as you know looks very different from the shape of the inference workflow workload. A training workload is very predictable, high utilization for for a short period of time. The inference workload is very burst usage, typically when humans are awake, until the agents take over. Maybe then it'll be 24/7.

And harder to predict. It goes down around Christmas for some reason. It goes down around Thanksgiving for some reason. But I think that might be the case that we, you know, at least in this calculation we try to capture it because it it's it's a mix and but it's a good it's a good hypothesis. What what what what would What what would probability be? It's on slide 16. We'll come to it. We'll come to it. I'll give you the answer. the most profitable part of the stack is the semis layer by a long shot. And video's data center revenues earn the most margin of about 75%. Don't quote me on it. It's like plus minus a couple percentage points from there.

whereas, you know, I estimate some of the application layer revenues to be somewhere between, depending on who you ask, like between 0 and 30%. And so the gap is quite wide. And I think the reason to that, I mean, it's it's it's a it's a it's a theory that a gentleman here had is there's one player who kind of runs the tables on the

semis layer. And so it's very much the case. And in fact, if you looked at this from a profitability perspective, it's even more concentrated.

The the the triangle is even more concentrated. I'll I'll I'll I'll I'll flash that in a second when we when we get to it, but it's a great question. Go ahead. Do you think that the machine learning is probably like 10 fish years on machine learning AI? What do you think about that in the first 10 fish years? Yeah, that is definitely a big part of it is that we've gone through the investment cycle in cloud.

It's definitely an element to it. Go ahead. If all these infra companies like Google, AWS, Nvidia, and all their own TPUs, and Nvidia is also doing their own TPUs. Open AI is also searching for some ASICs. who like where do does all these ASIC infra startups going to sell to? If everyone is building their own chip. There's a \$300 billion of revenue to fight about.

The But to answer your question, about half of that as as Jensen discloses on the earnings call is from the big hyperscalers. so those are those are probably going to be your primary customers. So if you were starting a chip company today, you would have a very the the shape of your customer base is a very small number of very large orders. it's a very different shape from building a consumer business or an enterprise software business.

and then you might have a long tail of other enterprises, though I wouldn't I wouldn't bank on it because I think they just go to the cloud providers. if you were if you were thinking about starting a chip company, it is a it is a it should be your number one consideration is like which of the five are you going to sell to first? Last question. And then I was able to get a small info earners in each of these layers. It takes multiple years for that to play out.

maybe for the very long way, I feel like in the past there haven't been a fully vertically integrated higher than just one. Mhm. I get that Google is fully vertically integrated on the hardware side, but I'm wondering how that shifts the balance of power in this space. What a great question. The biggest winner on the internet super cycle was probably Google. it's about 3 trillion in market cap, has near 99% market share in search. I would say that that's a pretty vertically integrated player, right?

They're on their own file server to search to ads on top to the user experience. Let's see the next one is mobile. the winner of that super cycle is Apple. What like two and a half trillion or so in market cap? You called it already. The next one is let's say social. Meta is probably the big winner in in in in social. They're not as fully integrated and what is their market cap like two trillion or something right now? Pretty dominant, but maybe they lost a trillion because they didn't fully go down to the servers.

And then the cloud is fairly heterogeneous. We don't have a single player that won You've got the three big oligopolies in AWS, GCP, and Azure. But they're not fully integrated. And you know, Nvidia's been trying a lot. Nvidia's been trying I don't know if you've heard of DGX Cloud, which is their cloud effort to to to build a cloud ecosystem. Obviously, they've they've got a series of vertical apps that they're trying. So, yeah, you might you might you might be onto something.

Yeah. Folks, I know it's a Thursday evening at 5:00. Probably the last thing standing between the weekend and your weekends. I don't want to be that person. So, I'm going to jump into the part that wakes you up. I do actually have This is This is a quiz that we're going to go go through. I'm going to give you a hint about the companies that we're going to go through. I do have a prize for the winner.

This is a prize. So, you're motivated. And it you win points on on on two grounds. One is by being right and the other is by being fast. The fastest way to be fast is to do fast inference. And drop that thing into cloud. Please, you're welcome to do that. Just give the human players 5 seconds. Let them go at it. And let let them win the analog way. And if you really want to use cloud, you're welcome to do it. Just give them 5 seconds.

All right. So, this is question number one. >> Ready for the next?

The software engineers in the room might have a have an unfair advantage. So, you know, I wanted to spend the next maybe 10 minutes or so going into some of the hypotheses that I have about what's going on and why the value is accruing in the manner that it is. I think there was a question, a very good question about profitability and how it gets magnified. So, we'll jump through that, but again, feel free to stop me if you have any questions.

I have a feeling we're going to have very little time left, and I do want to end on time. You guys remember this? and I I painted it slightly differently on the next chart, which is I did the same exercise that I did that I posted about 2 years ago. And what it looked like 2 years ago was this thing on the left, where the ecosystem was obviously a lot smaller. It was about five times smaller. Shockingly, the shape of it hasn't changed much. This is despite This is despite heroic growth.

And you know, for if if if you look at the revenue that was added about 350 billion or so of revenue added, a good like 75% of it just went straight to semis. in the last 2 years, despite apps having grown, you know, more than 10x, it still hasn't made that big of a dent. And so I was like, "Okay, well, let's dig deeper into this." if you started to open up each of these cells and you're like, "Hey, what what companies make up each of those parts?"

most of that 300 is Nvidia, as you guys know. The apps is actually two companies make up about 90% of it. Anybody want to guess which those two are? The infra segment is the one that has the most competitive intensity, as we discussed. It is probably the place where there's the biggest battle brewing both sideways, but also across the stack. it's also the place that has the highest metabolic rate in that there's a lot of companies being formed, there's a lot of companies that are getting bought out, and I would say it's the most competitive, but also the most unstable of the equilibriums that we have right now.

And the And you know, the question that we think about as we think about investing, as we you guys will think about investing your times, is how much time will this chart that has moved such little in the last 2 years, what what is the amount of time it'll take to get to cloud software shape like like shape? is it 5 years? Is it 10 years? Is it 15 years? Is it never? Maybe it's just stays that way.

We do think it'll happen. We think it'll happen at some point, but it's it's it's not happening nearly fast enough. The second thing that we've been thinking a lot about as we think about the future of AI is you know, this is this I don't know if you guys saw this chart in the readings, but consumer AI, which is the biggest, call it, market for AI right now outside of coding, has incredibly high usage on, you know, at ChatGPT, most of it is free you know about 95% of the users are free.

And Gemini who's I don't know if you guys saw but Demis who leads DeepMind announced that they were not planning to do ads as a subscription as as a revenue model. We've been thinking a lot about hey how big do these businesses get? What is the monetization engine of these businesses? Do you think a subscription business will be larger or ads business will be larger? And so what I did was I looked at the largest consumer franchises outside of AI.

And so you will see that I mean you you all know these products. There's a class of products that have gotten to 3 billion users scale. These are almost near mandatory products to live your lives to to you know this is WhatsApp and Chrome which you could not live without. Then there's a class of products at the 2 billion 1 and 1/2 to 2 billion users scale which are social these are like social products like Instagram Tik Tok and Facebook.

They're not mandatory but they are exhibit very good network effects if my on one of these I'm more likely to be there. And then you've got the third category of you know mainstream consumer products that are neither mandatory that are neither extremely social but I would call them like niche products. If you're shopping you go on Amazon if you're looking for music you go on Spotify if you're looking for a good debate you're going on Twitter or cat videos.

Any guesses on where closer to which of these will chat GPT and Gemini are right now? And the answer's on the next slide so we'll get it quickly. Would you guess that chat GPT's or the leading AI applications terminal scale will be closer to a mandatory app like YouTube or WhatsApp? A social app like Instagram or Tik Tok? >> >> Or a niche app like Spotify or Twitter? Any guesses if not I'll rebuild the question go ahead.

I would say on the YouTube WhatsApp scale because it would be a daily utility. Yeah. People would just be using daily as part of their normal life. Yeah. You are Well, I'll let me show you the answer and I'll we'll come back to your advice. Do you have any other guesses? Any different guesses? Go ahead. It's closer to Facebook time frame only existed around nine months exactly. Yeah. Yeah, that's right. You're certainly right right now. can I show you guys the answer?

This is this is how they fare if you plot them all together. ChatGPT has just overtaken the niche app category. Gemini still has not. You were right that it's heading towards social. Personally, I would have loved as a as an investor in OpenAI, I would have loved for it to start heading towards the full utility, but you know, one of the biggest questions that we ask ourselves is is knowledge work work that everybody does. Is knowledge is the work of you know, ChatGPT is not a place where you're messaging other folks yet. It's not a place where you're getting your email inbox or your or your or your document fix. It's a place where you go and you have to do active work. You have to go ask a question.

And the number of people in the world who are asking active questions of technology is not the entirety of the population that's online. You know, there's about 8 billion people on the planet. 4 billion of them are online. The graph economics of consumer applications that you know, Alphabet has about 4 billion users. They monetize them at about \$100 a user a year. Meta's got about 3 and 1/2 billion users that monetize at about \$70 a user a year.

The leading AI provider ChatGPT has got about a billion users that are monetized at about \$10 a user a year. And so, the question is how do we get the billion up to 4 billion? I'm not sure knowledge work is the answer. I think we'd have to go beyond knowledge work. and then the second question is then how do we get the \$10 a user per year up from 10 to 100? And I'm not sure ads and I'm not sure subscription's the answer. I suspect we'll have to go into ads.

And I suspect the ads that ChatGPT will be able to serve or cloud will be able to serve will have a lot better pricing because they will understand your intent that you'll be logged in, very good attribution, a lot more trust. and I think that'll be the big other big headline this year. And you heard it here first. It'll be a big deal. There's a lot of alpha in understanding the ad model really well. And once again, 10 years ago at the Facebook IPO, there was a lot of short reports on Facebook because people said, "Hey, well, these ads worked on on a computer. They're not going to work on a phone. Why? Because there's no space on a phone."

Shocker, we found the space on a phone. The same thing's going on right now, which is while I'm having this conversation, it is a very personal conversation. I don't want to be shocked by advertisements. That's the big debate. I couldn't tell you what it's going to be like, but I am optimistic that we will find it. And I think that's going to be, you know, a big a big unlock a big unlock for this economic model.

And so we'll dig into that in in one of the speaker sessions later this year. I've got a lot of new slides. We're out of time. Thank you.