Research @ Citi Podcast, Episode 46: AI — The Apex Technology of the Information Era

Recorded: August 26, 2025

Published: August 28, 2025

Host: Lucy Baldwin, Global Head of Research, Citi

Guest: Heath Terry, Global Head of Tech and Communications, Citi

Transcript:

Lucy Baldwin (00:00)

Hello, and welcome to the Research @ Citi podcast. I'm your host, Lucy Baldwin, Global Head of Research at Citi.

Today, I'm delighted to be joined by Heath Terry, our Global Head of Tech and Communications Research here at Citi. Welcome to the podcast, Heath.

Heath Terry (00:16)

Thanks, Lucy. Really excited to be here.

Lucy Baldwin (00:18)

Fantastic. Now, I know this is pretty timely because we've got our big Tech conference upcoming. You've just launched with a new piece of work here at Citi, looking at the private-company ecosystem and in particular what's happening in the AI space.

So maybe I can delve straight in, given people have just gone through a summer with a lot of earnings reports, an awful lot of product announcements, things like ChatGPT 5, DeepSeek version 3.1. It's been a lot of conferences and there's a lot of research papers, too, that have come out across the industry debating how valuable some of this enterprise AI actually is — whether it actually works or not.

So where do you think the investors are now, Heath, given all of this? How are they digesting all of that new information?

Heath Terry (01:05)

Yeah, it's been fascinating to watch this summer because we've really had this pendulum swinging back and forth between real optimism, sort of glass-half-full view of AI and the world that we're going into, and more caution, the real concerns that companies are over-investing and over-spending.

You know, we're just a little over a year past when we had the CEOs of Microsoft and Meta and Alphabet all get on their earnings calls and say some version of “the danger of under-investing is much greater than the danger of over-investing.” And they all acknowledge that maybe they're over-investing. And what's really interesting to look at a year later is all of those same companies then got on their earnings calls and made note that they're still running well below what they need from a capacity standpoint. So even though a year ago, they were afraid that they might be over-investing, a year later, it turns out they were actually under-investing relative to what they need and the kind of enterprise adoption that we've seen.

So I think where investors' heads are at right now is kind of in the middle, maybe a little bit more on the cautious side at the moment when you look at the way certain areas are performing. And — I know we'll talk about this more as we get into it — we are seeing a little bit of dispersion around how individual parts of the AI trade are acting. It's clear this is not a monolithic trade anymore. And it's really about how the individual companies are leveraged to it.

Lucy Baldwin (02:34)

Heath, on this concept that the AI trade isn't sort of some big aggregated thing, it's how individual companies are playing through it: Do we know whether the estimated $400 billion of investment in AI infrastructure is playing out well in people's minds? Is the return on that looking like it's going to be sufficiently attractive?

And to link it back to what you were saying, how does this then actually play out at the enterprise level? Because obviously we've seen some huge capex numbers from certain companies being announced.

Heath Terry (03:10)

Yeah, it's interesting. Over the last few weeks, we've gotten sort of dueling reports on this. MIT put out a report last week that got a lot of headlines suggesting that 95% of the AI proof-of-concept initiatives don't actually return any value and that all the value out of AI is coming from that remaining 5%.

Now, to be fair, that was the headline. You know how headlines are. If you read the actual report, it's much more nuanced than that in terms of what the data shows in terms of the value that companies are getting out of this.

But look, on the other side of the spectrum, you had an IDC report that came out a few months earlier that showed that companies were getting a 3.7x return on every dollar and spend on AI. So, the idea that AI is this printing press where you put a dollar in on one side and get $3.70 out on the other is incredibly encouraging.

And like most things, the answer is probably somewhere in the middle. We surveyed 120 companies, startups, and existing enterprises about their use of AI. And generally, the response was that companies are getting positive returns off their investments on this. They feel good about the investments that they're making in it.

And I think, as I referenced before, that's really showing up in the company data. You get a company like Microsoft that is exposed to the broadest of broad sets of clients that comes on and says, “We're still running behind capacity, we've got more enterprise demand than we can fulfill.” That wouldn't exist. They wouldn't be able to make that kind of statement unless customers were getting value out of what they're putting into it. And I think that's kind of what the right answer is going to be when it comes to how investors should think about this.

Really different answers if you're talking about specific software companies vs. specific hardware companies vs. enablement companies. But as a whole, this technology does seem to be working.

Lucy Baldwin (05:06)

That makes sense. And I think there's obviously been a huge amount of concern about this potential for over-investment in the infrastructure. And I think parallels have been drawn to what we saw with fiber when we went through the early internet era and people ended up laying down all this fiber. And, you know, some numbers suggest that 85% of it was never used.

I think what you've just said probably suggests, given where demand is now vs. supply, that that isn't happening now. But how do people get comfort that that is the case and that this is fundamentally different than that experience was?

Heath Terry (05:41)

I think you get comfort by digging into those use cases. Is this working for companies that are putting it into development? Because if it is, then we're about to see an explosion in the amount of data-center capacity that's needed. Because when you're running a proof of concept in a Fortune 500 company, you're running that against maybe 5,000 users. But when it goes into production, that can go out to 200,000, 300,000, 400,000, a million users at companies that have those kind of employees or those kind of customer bases.

And so, the amount of inference and the amount of compute that's required as you go from proof of concept to production is exponentially larger. If you get comfortable with the idea that the technology works and that there's value in putting this technology into play, then you can get pretty comfortable that the $400 billion that's being spent in data-center buildout is going to be worthwhile.

And for what it's worth, that's $400 billion this year. When we look out over the next five years, just the hyperscalers alone are spending $2.3 trillion on data-center buildout. You've got another trillion that's already been committed from sovereign governments announcing their own data-center initiatives. So, we are going to see a lot of capacity here. And there's a real debate over how much of that we need at the data-center level vs. what we're going to be able to utilize at the edge level. But at least at the moment, I think investors can feel a pretty high degree of comfort that this is not another dark fiber situation.

Lucy Baldwin (07:11)

Got it. And Heath, you just said something there that I think will resonate with a lot of people. This debate around how much of the investment we've seen to date, and in the near-future continuing, is really in the picks and shovels, if you like. It’s really at the data-center server level, rather than actually starting to see the value creation move down the stack.

What's your view on where that ends up? And in particular, you talked about the edge. That's obviously super topical, obviously there's a lot of growth there, but it's perhaps from a smaller base. Can you talk us through how you see that evolution coming through?

Heath Terry (07:49)

As a firm, we've done a lot of really interesting work specifically around the edge. Peter Lee out of Korea did a phenomenal piece back in May going deep into opportunity around edge computing, what Jensen Wong has referred to as an AI server in your pocket. We all have these massively powerful computers that are incredibly underutilized at this point. And that's the compute power that AI can consume very, very quickly once given the opportunity.

It doesn't really work for that right now because you need these very powerful data centers. But as we start to use more small language models, as we start to see more efficient models being built specifically for that and more of this compute push to the edge, it does create incremental capacity that ultimately should bring down pricing for these workloads in a way that increases the ROI that enterprises are seeing on their usage and enables more usage.

Right now, one of the things that we interestingly have not seen in the development of this AI cycle is any consumer applications beyond things like ChatGPT. Typically, at the beginning of a technology cycle, like we saw with internet or mobile, you get travel companies and social-media companies and messaging companies and calendar companies and this explosion of companies meant for consumers. And we really haven't gotten that because the economics don't work.

But when you bring the cost down — and we've seen 99% declines in cost over the last three years already — DARTs do enable a lot more applications. And I think edge computing is going to be an incredibly important part of that.

Lucy Baldwin (09:28)

That makes a lot of sense. And just to stick with this infrastructure layer for a moment, Heath, there's been huge press around what's been going on in the land of semis, in terms of capacity there in particular, and obviously the constraints in GPUs.

Who are you actually seeing have success now in terms of creating supply within the chip space? And is there really a lot of demand or significant demand beyond just GPUs in other chips?

Heath Terry (09:56)

You know, it's interesting. We have seen just such incredible demand for what NVIDIA has created. And their CUDA ecosystem has for so long kept anyone out from really being successful in creating meaningful competitors, able to take real market share at the GPU level. But the level of investment that we're seeing from companies like Amazon with their Trainium and Inferentia chips, Google with their tensor processing units that are what they use internally to train all of their models, as well as early initiatives from companies like Microsoft and Meta to build out their own chip ecosystem around the GPUs that are available from NVIDIA just gives you a sense of how important this is to other companies.

We're still in a chip-constrained environment. And so that increase, it's largely being driven through Broadcom, is really I think probably where you're seeing the most impact from these custom ASICs that are being built. Laura Chen out of our Taiwan team has done some really interesting work around this balance between custom ASICs and GPUs. This is not winner take all, and there is real value from having a more diverse ecosystem of infrastructure at the chip level.

Lucy Baldwin (11:17)

Got it. And with this kind of concept of a diverse ecosystem in mind, Heath, what do you think everything we've discussed so far today actually means for the existing enterprise tech stack? Is it at risk of obsolescence? Where do you think it's at?

Heath Terry (11:41)

You know, every cycle that we go through, whether it's mainframe to client server, client server to cloud SaaS, now cloud SaaS to AI-optimized, which is what we're kind of calling this at the moment, you're going to see these kinds of transitions and it'll be gradual.

One of the things that we spent a lot of time detailing in the 120-page report that we published earlier this week is what this AI-optimized tech stack will look like. And we spent a lot of time talking to technologists and founders and CIOs and companies that are designed to make this transition and to get a sense of what that tech stack starts to look like in an AI-optimized world.

The earliest stages are just what we were talking about in terms of GPUs, this transition from CPUs to GPUs at the foundational level. We're seeing it in data. As you see vector databases being more utilized here, we're starting to go further up the stack. And it's one of the reasons that we think there's a lot of opportunities and companies that are enabling that further up the stack, that are really helping enterprises like Citi and others in the Fortune 500 make this transition from cloud SaaS to an AI-optimized stack.

We're still super early-stage in this, where when these happen, they tend to be sort of 10-year processes and arguably we're maybe a year and a half, two years into this one. But that's what we're seeing so far.

Lucy Baldwin (13:15)

Yeah, that's really interesting. And building on that concept, when you think about some of the software companies, if you look at software as a space in aggregate, it's obviously been a pretty big underperformer since the AI boom kicked off, right? And that's despite a lot of those big companies in the software space launching their own premium AI tools, etc. Do you think the market's got this right? Like, is software going to be a perennial underperformer? Is the pricing side going to be unbelievably difficult for them in an AI-led world? What are your thoughts there?

Heath Terry (13:46)

Yeah, you touched on the two big issues. The one that's getting the most attention is because companies like OpenAI when they demo do such a good job of painting the picture of this future where you and I just sit down in front of a chat engine, tell it what we want, and it creates it for us out of whole cloth. It's fascinating when it works that way.

I struggle to believe, though, that we're going to want to live in a world where we do that every day because so much of our daily tasks revolve around the same thing, right? You and I are analysts; we're not going to want to walk in every day and the first thing that we do is sit down and have ChatGPT build a version of Excel so that we can build a model. We already have a pretty good version of Excel. We don't need somebody else to build it for us.

And I think that's going to be the case in most versions of verticalized software. You can create and refine and build, and especially enable with AI, a much better user experience with that kind of specialized environment, than a more general AI application is going to be able to create on the fly for you.

I think where there's a bigger impact, and you touched on it, is on the pricing side of things. The really basic way of thinking about this is if a Salesforce automation tool makes my salesforce 30% more efficient and I need 30% fewer salespeople, then that's 30% fewer seats.

And so just to stay even, they've got to raise prices 30%. We've seen a lot of this over the last few years as companies have gone from seat-based or license-based to consumption-based models. Those pricing transitions can be painful. And I think that's what's being discounted in the stocks more so than anything right now is that there is this pricing transition or just this unknown out there. Unknowns, as we know, are always the enemy of valuation, particularly premium valuations. And I think that's really what we're seeing more in the stocks right now more than any sort of real determination as to what kind of an impact AI is going to have on software.

Lucy Baldwin (15:51)

Yeah, fascinating. And linked to that, Heath, you mentioned earlier this relative lack of consumer applications that we've seen coming through so far. And I think a lot of people listening to this will say, “Well, you know, maybe I use AI at home, but I only use the free stuff. I'm not paying for anything.” Again, a lot of that links to what you've said.

I think when you look across the space, it's the internet ad models that perhaps seem to be furthest along in terms of actually being able to leverage AI in order to generate some real revenues.

How instrumental is that as a test case and example in terms of the future success for other companies in other industries, do you think?

Heath Terry (16:32)

Yeah, I think it's the best example of the way that this is going to be used across every application. So clearly it should come as a shock to no one that the technology companies are the ones leading the way on actually implementing this stuff. They have the talent to be able to do it. They've got the use cases to be able to do it. They've got the compute power to be able to do it. And so companies like Meta and Alphabet, who led here by using AI on their own products.

And so when people say, “I actually don't use AI” or maybe “I only use ChatGPT,” realize that if you touch Instagram, you touch Google, you touch any of these consumer applications that we all use daily, you're touching AI. And that feed that you're seeing on Instagram or on Facebook has been optimized by AI. The advertising that you're seeing is being targeted better because of AI. The creative that you're seeing is better because of AI, in most cases.

And clearly, you look at the acceleration in growth, the improvement in things like revenue per search or revenue per query at Alphabet or revenue per user at Meta, you're clearly seeing a really positive impact from their adaptions of AI.

Now you take those same fundamental tools, because at the end of the day, they're just optimization engines. And so, you point that same optimization engine at drug discovery, at cost controls within a hospital environment, at energy discovery, at preventative maintenance in a utility, and you start to get those same efficiencies, those same optimizations, and the impact turns into trillions of dollars. And we've done the math around what we're kind of calling the first three of the big enterprise-use applications, which are code development, customer service, and knowledge retrieval.

That's about $275 billion a year in cost savings that will be realized just from those three. As I said before, when you apply that across everything else that AI is going to touch, you get into the trillions of dollars really quickly.

Lucy Baldwin (18:39)

Yeah, that's pretty fascinating and pretty exciting, Heath. And linked to some of those big numbers, the last big question for you is if you think about VC investment that we've seen into AI, I think its run rating is currently at about 180% up year on year, Heath. But alongside that huge investment, we haven't really seen any meaningful exits from the last big cycle in terms of VC investing.

Given that, how sustainable do you think a run rate like that is going to be in terms of investment? And then what happens if you end up in a scenario where VC can't fund the kind of capex into AI that is really needed? Where does the world go to next for that funding?

Heath Terry (19:24)

The thing that I would say we're seeing in venture right now is there's this incredible appetite to put money into AI if you're already winning. And so, if you're OpenAI, if you're Anthropic, if you're any of these companies that have shown that they're at the front of the queue in terms of leading this AI environment, there's almost unlimited demand to invest in what you're trying to do.

It starts to get really narrow after that, and it starts to get really thin, even for profitable businesses that are out there. What will be challenging, I think, in this is AI is a much more capital-intensive business than really anything that we've seen in technology in a very long time.

You think about what it costs to start an internet company, to start a SaaS company, you were talking about a few hundred million dollars. Meanwhile, OpenAI has already raised close to $30 billion and deployed. That's not even including some of the commitments that they have for future raising. And the cost of building out these kinds of data centers or the cost of consuming and training these large language models is so high that it's going to be really tough to see a lot of smaller companies raise the kind of capital that they need to.

We're also going to need to see new structures, right? Venture and private in general is going through a real transition right now, because companies are staying private so much longer, they're getting so much bigger, they're getting so much more capital-intensive.

And that's why you've seen the growth in crossover investing that we have. It's why you're seeing venture tap into or try to tap into retail investors in a way that they've really never been able to before. And we think it's one of the reasons that focusing on the private-company space is so incredibly important.

It was one thing to stand on the sidelines around a sector while it developed when the companies involved were $1 billion, $5 billion, $10 billion companies. We've got $300 billion-plus companies in this space right now and multiple $100 billion companies in this space; we just can't wait for them to go public before we really start paying attention to the impact that they're having on enterprises, on consumer behavior.

And so venture is going to continue to fund this, as investors get more creative. I think we're going to start to see more novel investment structures around private companies, and particularly around the infrastructure that they need for this, really separating out, in some cases, the technology investors from the infrastructure investors or the more cash-flow-driven investors. And you need that to really tap into the amounts of capital that you need to fund the amounts of investment that we need here.

AI is a big-company game. And that's what we're seeing right now. And the amount of money involved is going to require a lot from investors, which is what makes this so incredibly interesting.

Lucy Baldwin (22:17)

That is fascinating and a great place to leave it. And we'll certainly welcome you back, Heath, for part two of this kind of conversation, because there's a lot to cover that we haven't had a chance to. But it's very much, as you describe it in your report, the apex technology of the information era.

And as you've just articulated, it's going to require some innovation in terms of how to fund and finance it, not just by the big companies involved, but of course by the governments as well. So thank you for all of your insights today.

Heath Terry (22:44)

Thanks, Lucy.

Lucy Baldwin (22:46)

This episode of Research @ Citi was recorded on Tuesday, August the 26th, 2025. I'm your host, Lucy Baldwin. Please do join us next time when we will be discussing innovation in biotechnology with Anne Malone and Geoff Meacham. Thank you.

Disclaimer (22:59)

This podcast contains thematic content and is not intended to be investment research, nor does it constitute financial, economic, legal, tax, or accounting advice. This podcast is provided for information purposes only and does not constitute an offer or solicitation to purchase or sell any financial instruments. The contents of this podcast are not based on your individual circumstances and should not be relied upon as an assessment of suitability for you of a particular product, security, or transaction. The information in this podcast is based on generally available information and although obtained from sources believed by Citi to be reliable, its accuracy and completeness are not guaranteed. Past performance is not a guarantee or indication of future results. This podcast may not be copied or distributed, in whole or in part, without the express written consent of Citi.