Brian Gracely (00:01.742)
Three, two, one. Good morning, good evening, wherever you are. And welcome back to the Cloudcast. We are coming to you live from the massive Cloudcast studios here in Raleigh, North Carolina. March is upon us. Hard to believe we're already into the third month of the year. Hope everybody's doing well. Spring is right around the corner. I always have a little bit of a beef with spring because for every other month, it starts on the 21st. For some reason, we decided that spring starts on March 20th, which was always kind of a weird thing. know, seemed a little bit odd that we were.

one day off, but I guess we have an odd number of days in the year. So we have to live with that. For those of you that are new to the podcast, if you've been listening for any period of time, any short period of time, you're going to probably see over the next couple of days, couple of weeks, couple of months, number of shows, at least for a couple of maybe six, eight weeks or so, you might pick up a little bit of a different kind of tone or twang to our voice, especially for both Aaron and I, since we're here in North Carolina.

pollen season is starting to come upon us. is, it comes on like a, like a velvet blanket, like a velvet glove. kind of snuffs out all sorts of things here in the deep south, the wonderful yellow blanket of pollen. But what that means is it makes everything more nasally, everything more scratchy from a voice perspective. And we always apologize for that, but there's not much we can do to it. So we are just the beginning of that. And, but anyways, hope everybody's doing well. Hope everybody's starting to get into spring. Hopefully you're enjoying the daylight being out there a little bit more, at least for those of you in

parts of the world where daylight savings is coming to an end very soon. It feels like the days are getting a little bit longer, both in the morning and in the evening. So anyways, want to dive into something. Aaron did a talk this last week, he had a conversation this week with some folks about developer co-pilots, AI developer co-pilots. And just coincidentally, I had a long conversation with a number of people who were...

Aaron's talk was very much about the state of what's going on, some positive things that are happening, some progress that's happening. And I had a conversation with a number of folks in development organizations this week and just listened to their progress, their journey, their frustrations about developer co-pilots, but more so being pushed to use the tools.

Brian Gracely (02:22.058)
What I want to kind dive into today is in typical kind of weekend perspective is not to dive into that as a topic again, not to kind of reiterate it too too much, but to kind of get into this beginning to hear a little bit about kind of confusion around GNI, frustration in some cases around it. And I think what we're seeing is just sort of the natural evolution. Maybe it's what Gartner would call the trough of disillusionment.

around this technology that has been given a lot of hype, obviously been given a ton of funding. And a number of people are both bullish and bearish sometimes at the same time. Sometimes they're not exactly sure what's going on. think there's a certain amount of, are very good at evolving, but we're not necessarily good at learning really fast and then having to put that into practice. And I think we're kind of at a stage where right now there is so much going on

with AI and not only trying to keep up with it, but then trying to figure out like, how should I apply it? If I do apply it, does it work? Does this next announcement mean anything? Is it an incremental thing? Is it a good thing? Does it take us off in a Northeast direction, a Southwest direction? So it feels like we're beginning to get into, I don't want to call it a trough of disillusionment, but the natural cycle that every one of our technologies go through when

you know, we oftentimes look for the next silver bullet and then we have to work through the difference between expectations and execution. The difference between hype and hope and the realities of like, you know, can I make that work for myself? And so I made a list of some stuff that I've seen over last couple of weeks. And again, you know, it may just be purely coincidental. think sometimes if we simply just follow sort of new cycles, new cycles these days especially tend to be very

you know, either boom or bust, right? There's oftentimes very little in between. And the reality of where we live from a technology perspective is we oftentimes have to figure out the in-betweens because the in-betweens are the realities that the vast majority of us live in, right? The, you know, the 1 % on one end or the other of the spectrum, you know, that live in, you know, kind of crazy land in terms of number amount of funding or something like that, or crazy land in terms of being sad and depressed about something.

Brian Gracely (04:47.106)
I don't want to say we ignore those things, but in the middle is kind of where the reality is for the vast majority of our markets. And that's sort of the area that we want to kind of dig into today. So there was a number of things, and I put a bunch of links in the show notes. They're going to feel a little bit like a mixed bag of some negativity things, but also some future things and so forth. So some of the stuff that I'm seeing, and again, these are all just kind of cherry picking data points, but kind of going, OK.

Are they coming from places that are backed by a bunch of data or are they backed by things that could be relative to a lot of people? So the first thing I had on my list and I just sort of made a list of things where I said there's a ton of AI out there, and then I kind of want to dig into five or six topics that we're seeing and kind of try and put some reality to them, some rationalization to them. And again,

One of the things that we do a lot on this show, and some people enjoy it some people hate it, is I always try and provide a certain amount of historical context for when new things come along, do they look anything like previous things that we've done, previous technology trends, previous economic meets technology trends, all those sort of things. And I kind of feel like I'm doing a lot of that with AI, which is kind of a weird thing because on one hand, it's

pushing things in different ways than we've seen before, or at least that we've seen in many, many, many years, whether that's the cost of it or the potential for it to have impact or its involvement with both politics and economics and all that sort of stuff. And then at the same time, you do sort of go, well, technology has a history of sort of repeating itself. You just have to sort of know where it repeats, where it rhymes, and all that sort of stuff.

I'm kind of kind of going through that again just because it does feel like we are at least hitting a little bit of a bump in terms of people questioning What's going on or people really kind of questioning? You know, how do I make sense of this stuff and that that feels like more the sentiment than necessarily sort of a negative thing So the first thing I have my list is there's a ton of AI out there but it's not always simple and and what I mean by that is There was a very good conversation this week

Brian Gracely (07:11.278)
Ben Thompson, who somebody we follow quite a bit and we recommend quite a bit, had an interesting conversation with Benedict Evans. So both very well known on the internet in terms of being analysts and kind of big thinkers and very good at analyzing markets in sort of as things change at the macro level. And one of things that they sort of highlighted out was that there's a lot of stuff going on out there. But one of the struggles that people are having a little bit and this is again kind of

comparing technology trends as a whole to where we are today. Benedict Evans made a really interesting comment. said, you know, there's a lot of AI out there, but it's sort of based on this premise that people are gonna have to learn it. They're gonna have to figure it out. And the context he used was he said, you know, if you look at chat GPT, there are going to be people who have figured out all sorts of things. They've figured out prompt engineering and they've...

They've learned to do a bunch of things. But if you think about it in the context of sort of mass adoption of stuff, it really begins as nothing more than a blinking cursor. And for many, many people, when you start them with something that's essentially kind of a blank slate, right, a blinking cursor, as opposed to even just having a GUI on top of it, which has certain constraints about maybe what use cases you should do more frequently, sort of the defaults lifestyle, all that kind of stuff.

it's not always simple. And yes, there are people that are figuring it out. But in essence, it is a power user's game, if you will, at this point. there's some sentiment out there that, yes, there's a lot of AI out there, but it's not always simple. And that was one of the things I heard this week in talking to a lot of people, is they sort of said, look, I could tell you all of the things at my day job that

I don't like that I feel are repetitive, that I feel are time consuming. But it's not simple for me to figure out like, where is that sort of AI that's an assistant for me, as opposed to an AI that I have to adapt to the limitations of what it has, or I have to learn a whole bunch of things in order to be great at it. I just sort of want to, you know, kind of have the, the nirvana of what sort of been promised in some cases, which is like,

Brian Gracely (09:29.378)
take away the work that I don't want to do so that I can focus on higher level stuff. But that's not necessarily a simple thing to do. So I think there is a sentiment out there that while there's a lot of AI out there, it's not always simple. And I think when we think about the biggest trends over the last couple of decades, whether it was like, hey, if you want to search for stuff, there's just this box and you type stuff in and then stuff comes back.

anybody can kind of figure that out. Or, you know, here is a device that if you want to make a phone call, you push the phone button. And if you want it to, you know, play a game for you, you push the game button. And I know I'm oversimplifying some things, but there is a certain amount of, you know, kind of things that the computer industry has to do to make the amazing things that they do behind the scenes.

usable for mass numbers of people. Right. So anyways, I think that's the first sort of sentiment that I'm hearing from places, not necessarily everybody, but places that it's not always simple enough. And we haven't yet reached a point where, you know, the blinking cursor concept of, sort of endless possibilities, but not necessarily enough constraint to help figure out how to get started is, has, has progressed enough. Right. So that's the first one.

The second one is there's a ton of AI out there, but it's not always affordable. And we're starting to see a little bit of that, right? Where obviously it feels weird on this show to constantly be checking ourselves and be like, wait, that number says a hundred billion, not a hundred million. That number says that company lost $12 billion this year, or it's going to, there's a new project to take on.

$500 billion of data center stuff, right? Like we're talking about numbers and the B's as if they are, you know, the cash that you have in your pocket, right? So there's a certain amount of that that I think people, they say the numbers, but kind of wrapping what that means and then how does that become profitable, right? Because one of the things that we have to be very conscious of in this industry is like, there are no lack of ideas of things that are possible, whether it's like,

Brian Gracely (11:50.232)
putting people on other planets or whatever it might be, like desalinizing the ocean or any other gigantic idea. But at some point, they have to be paid for and then there needs to be a return on that payment so that it can continue to be a viable thing. so I think there is a sentiment out there that maybe the latest and greatest of all computing, and I point to a couple of articles. There's an article from Vinod Kasla, who is one of the

super well known VCs is kind of an OG internet VC. And he's got an article out there that says, most AI investments will lose money as the market enters a greed cycle. And I think what he means by that is we still are at a stage where we are adapting to this idea that for the last 20 or so years, all compute has essentially been marginally zero. And that's how the internet's been built and it's how we've had scale out services.

But it's marginally zero because we figured out a revenue stream for it. And in many cases on the consumer side, it's because of advertising or selling data or whatever. On the enterprise side, you had to build a service that would be viable for your business. And so he's just sort of raising the flag that like right now, valuations were probably very high.

The amount of money that went into things was probably more than it should have been. I don't think he's necessarily making a statement about like interest rates and so forth because they're pretty reasonable at this point.

But again, he's just sort of highlighting that people who couldn't figure out a way to monetize things, especially when you either didn't figure out how to monetize it at any price, but especially when the cost of compute is very high, even though it does come down some, but still GPU shortages are out there and all that kind of stuff. And then there was a second article that was written, I believe, by the Goldman Sachs CIO, or maybe CTO, sort of an internal officer of Goldman Sachs, not necessarily an investment officer.

And he said, there's just too much spend and too little benefit to generative AI. And again, what he was really highlighting was there's quite a bit of investment you have to make in order to see success. And again, he was comparing it back to the days when you could just spin up a website or try and spin up an e-commerce type of thing. You could see immediate returns. And so I think these things highlight the idea that if we think about computing

Brian Gracely (14:15.096)
you know, the history of computing in the last multiple decades, computing takes off when we have breakthroughs that continue to drive the cost curve down, to drive the utility of it up, you know, the whole Jevons paradox idea. But right now we haven't yet figured that out for AI. And, you know, again, what that means is when you have technology that's potentially very, powerful, but you haven't figured out the economics of it, you start to get into things where

people get creative, right? We saw this with DeepSeek to a certain extent, but I think we're probably going to start to enter an age in which people have to really think about both the power of the technology, but also the economic viability of the technology, right? So whether that is smaller models, whether that is technology that's being used to sort of decouple certain choke points. So maybe that's Nvidia GPUs or something, projects like VLM and other stuff like that.

You know, I think we're going to start to get into a stage where people are as conscious about the power of the technology, the what could you do with the, I afford to do anything with, the economics that are, that are sort of prevalent and then what are things that could happen in order to make that more, more viable? Right. So, again, not a necessarily a negative thing, but a, you know, push for new innovation.

to really kind of to address some of these things. The third thing is on a different sort of swing or sort of, you know, going in a different direction. There's a ton of AI out there. And sometimes it blows our mind, right? I, I've had a number of friends who have just been working on side projects lately, and they're like, you can't believe, you know, what what it was able to do, right? You know, what it was able to do, whether it was a reasoning model, or just

its ability to build charts and graphs and images and take on things that would have taken them days and days and days to do, it was done in 30 minutes. And so I don't think we've lost the sort of shock and awe that we see sometimes when some things happen from a generative AI perspective and we go, it's amazing what it was able to do. And it may not be 100 % accurate. I think we're going to figure out how we live in a world in which

Brian Gracely (16:41.582)
We're not always sure that things are 100 % accurate in a technology sense, right? Like we already live in that world from a politics and a media perspective, but we're to have to learn how to figure that out. You know, and we're either going to get smart about that or tools are going to help us do those sort of things. But I definitely don't feel like we are in the same sort of trough of disillusionment in which when the technology does work or it is applied in ways that we, you know, we hope it will work, that it doesn't do some amazing things for people.

or it doesn't produce some amazing results. And so, you know, to me, that is the really interesting part of this and the part that continues to make me think that while there may be some concerns about profitability and affordability, and is it simple enough, the bones of what are there, you know, unlike some of the other kind of recent kind of buzzy things in technology, you know, a web three or metaverse or something along those lines, certain crypto things.

You know, this feels like something that has very, very broad human applicability and at the same time, you know, has at least shown itself to be such that you're like, wow, that is something that we really couldn't do. And it does create viable sort of end results. Right. So keep that on the positive side of things. Another thing is there's a ton of AI out there and open source is starting to be disruptive.

And I put in parentheses again. And what I meant by this is, you know, open source is, going through a moment in the AI world because the definitions that we used for open source software, the sort of framing of, know, what made something really open source, the licensing, the ability to create stuff, stuff from, from source and have all the parts such that you could create it is going through some things in the AI world because of, you know, how much

because the cost of some of these gigantic models is so high, you know, take something like a llama model, you know, hundreds of millions of dollars to build or something along those lines. You know, there is some hesitancy to necessarily release everything, all of the data, right? And we are seeing this in some things and in some instances, so this isn't an absolute. And that's causing some concern in the open source software community of like, are we going to be as

Brian Gracely (19:08.142)
you know, specific in our definition of what is open source versus, you know, what is useful. And again, I understand the pros and cons of both of that, but I think we are beginning to also see kind of a forming of a stack that is becoming more more prevalent in open source. And again, you know, I don't know if that's going to be purely at the, you know, the model level, right? We're starting to see some commoditization even of, I don't even know if it's commoditization, but just like so much choice.

whether it's a DeepSeek or a Llama or lots of other things that are out there, granite from IBM or a bunch of other things. But we are seeing lots of options and variety out there, so it's good and they seem very viable. We're seeing things like PyTorch become very sort of de facto. We're beginning to see things like VLLM, as I mentioned earlier, becoming a disruptor in the space that sort of CUDA, which is oftentimes thought of as the moat for NVIDIA,

you know, potentially could be for the AI hardware accelerator. And again, if we can, you know, attack that area in terms of trying to bring down costs, trying to bring down availability, that typically is viable and good for the broader sense of the marketplace. So, and then, you know, obviously we're seeing things like, like LamaStack from, from Meta beginning to gain a lot of traction and offer a lot of very, very interesting capabilities from, you know, how to build applications, how to coordinate applications and agents and all that kind of stuff. so there are

beginning to be a bunch of viable, and I'm leaving out tons of them, but they're beginning to be a bunch of viable open source communities and projects and things that offer a very good alternative to some of the very locked in proprietaryness, proprietary stack, proprietary location. And I think we've seen over and over again, open source has been a huge unlocker of value and creator of value.

when used in certain ways for the broader industry. So I'm encouraged by that sort of thing. Now, for stuff that is a little more complicated, the next one I have on my list is there's a ton of AI out there and people are trying to forecast a future that is really changing very, very quickly. I mean, two weeks ago, three weeks ago, we were talking about DeepSeq and like, was it going to be an industry changing type of thing?

Brian Gracely (21:28.28)
you know, we went back three or four months before that, we were talking about, you know, the amount of funding that, you know, people were pouring into data centers or pouring into models. You know, if we go back a year from before, you know, we were talking about, you know, Amazon's position and how they were very creative in what they were doing. And they had this great position with open AI, and then you move forward a couple of months and, know, their relationship with open AI seems chaotic and so forth. And so I think it's, it's very difficult for people to kind of forecast a future.

when so many variables are changing. We really haven't stabilized or standardized around any of the variables in the stack. You can argue, the chat GPT type of thing is sort of a standard, but not necessarily. It's not a profitable model. It's not widely used everywhere. mean, it's used in a lot of places. But again, it's got

or 400 million users versus billions for mobile phones or billions for web browsers and so forth. it just feels like we have everything as a variable. so it becomes very hard when everything's a variable to kind of then build models or build projections or build framing of what the industry might look like when there's so many things. And again, this is not unusual. mean, we've obviously gone through periods of time when

there was lots and lots of choice for companies and for consumers. There was lots of choices in terms of, you know, is there a standard for things yet? And we are in very, very early stages. And again, you know, we're, in a weird stage in which, you know, if we think about how long it took for the internet to sort of grow up, I mean, the, the, foundations of the internet took many decades. The sort of commercial building of the internet took probably 15 years to call it, you know, 2000, you know, 1995 to about 2000.

2001, 2002, when Google started to figure out how to monetize things. know, the cloud took a decade before Amazon was even making a billion dollars. You know, it took, you know, a long time for the mobile phones to sort of replace the old mobile phones and mobile applications. these things take decades plus to sort of shake themselves out. And right now, I think because there is so much money pouring into this, and it feels like money is trying to replace time in terms of

Brian Gracely (23:53.804)
let's build the AI internet, AI world in half the time, a third of the time, a quarter of the time. It's making it very, very difficult to have any sort of bearings about that. I think to a certain extent, we've often seen when so much money is thrown at a problem, it often doesn't create the right kind of behaviors in terms of building things that are sustainable. just simply building things for trying to win a three month race, be the next

Press release pressure announcement so on so forth. So I think people are struggling with that and the last one I had on my list was There's a lot of a there's a ton of AI out there and people are struggling with The fact that it feels like every time we talk about it. We are intersecting technology and politics and You know, it's not a statement on politics per se politics are part of our lives and whether you

choose to be actively involved with it or not, the ramifications of politics impact people's lives. But it does feel like more and more technology and politics have overlaps in which you can't necessarily pull them apart. And I talked about this a little bit earlier in the year. We're going to try and do our best to not necessarily take a political standpoint on things. But I think the reality is, again, whether we're talking about

Government funding or subsidizing of certain things whether we're talking about geopolitical Issues that could you know arise if you know one country invades another country? You know laws that prohibit whether they're tariffs or restrictions about where technology can flow around the world like we are at a stage You know whether you like it or not that that the two things are overlapping much more than they probably ever have at least in in the last

couple of decades and that both makes things more complicated for people because we're not just talking about the technology or how do we evaluate the technology but we have to think about it in terms of you know are there political decisions being made political regimes having opinions on things or you know geopolitical things that you know are bigger than the world that you typically live in that are impacting the the maybe narrow scope you tend to think about technology in and so I think that is

Brian Gracely (26:16.14)
It's confusing people. It's adding more burden on them evaluating things and dealing with things and so forth. And again, it's the reality of where we are today. We're talking about technologies, especially in the AI concept, that are

impacting people in a similar way that the Industrial Revolution did and the invention of the automobile did and the invention of the internet did and other things. It's just that right now,

Everything feels very compressed and compact and the worlds are colliding much more so than we've ever seen it. then obviously we've got the power of the internet to sort of spread communication, to spread information. And so it drives more immediate discussion than maybe we would have 20 years ago, 25 years ago when the internet wasn't as prevalent or obviously go back to the 1900s with the industrial revolution. Like you needed the Pony Express to show up or you were

500 miles away from it. So you weren't going to hear the news for six months or something. anyways, I think on the whole, we are in a stage of what feels like constant change, which is very hard for people to get their bearings. We are in a stage of constant innovation, some of which feels very tangible, some of which feels very hard to grasp because of the pace that it's changing. so

you know, I think we're having to figure out how do we adapt in that world? How do we adapt to finding a normal finding a north star finding what applies to what you do? You know, what doesn't apply to what you do? How do you filter those things out? So anyways, I just kind of want to throw some things out there. Because I, you know, again, this this, this AI thing, right is going to be a decades long kind of evolution, you know, for good for bad for whatever it might be.

but the little micro time segments we have within it sometimes feel more than they have in the past and so forth. anyways, try to find some balance, try to find some things that allow you, if AI is in some way part of your world to impact you in a way that benefits what you're doing as opposed to feels like it's disruptive or...

Brian Gracely (28:41.742)
disruptive to what you're doing, negative to what you're doing. But we will do our best here on Cloudcast and Cloud News of the Week and Weekend Perspectives and all those things to try and find some middle to put some reality and rationalization around it. So anyways, welcome to March. Hope everybody's doing well. Thank you all for listening. Thank you for telling a friend and giving us some feedback. And Aaron has been taking over.

a lot of the weekday shows, a lot of the interview shows. And sometimes when you give Aaron a project, Aaron goes crazy with it. So he has an amazing set of guests lined up for a lot of shows. I've been traveling a bunch. My work's been kind of crazy. He's been steering the ship and guiding a lot of those interviews. So I thank him for sort of taking over a bigger role in that. And this isn't a permanent thing. Him and I are both going to be on a lot of the different shows. But thank you all for listening. Thanks for.

putting up with some of the variability and again, thanks for putting up with some of the scratchiness and nasoliness as we get through what I call season season or pollen season here in the South. anyways, hope you're all doing well. Thanks and have a good day and we will talk to you next week.