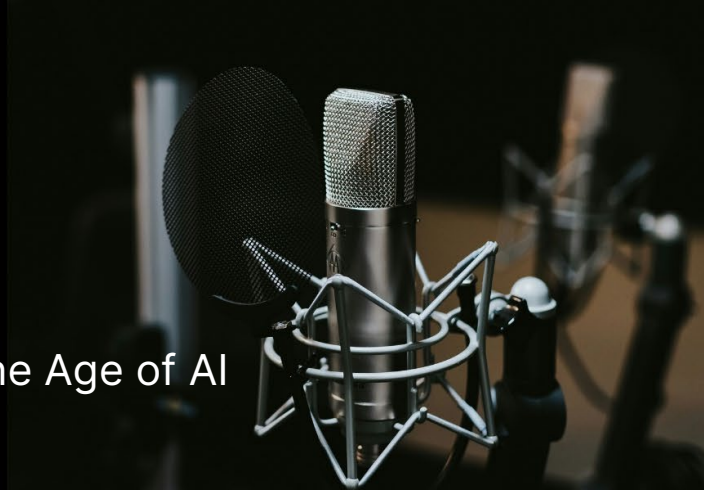
the art of boring™

EP 207 | Data Moats in the Age of AI



[00:00] Rob Campbell: Today on the Art of Boring, Josh Samuel, on competitive advantages tied to data. We talk about the four types of data that matter, why not all data is created equal, and how the rules are shifting in the age of large language models. Josh has a really useful framework for thinking through the strengths and weaknesses of these moats, plus the shockingly low number of likes that it takes for your social media profile to know you even better than your spouse.

[00:29] Disclaimer: This podcast is for informational purposes only information relating to investment approaches or individual investments should not be construed as advice or endorsement. Any views expressed in this podcast are based upon the information available at the time and are subject to change.

[00:46] Rob Campbell: Josh, welcome back to the podcast
And I'm glad you're back so soon, because there's a topic, or at least a research note, that you put into our system, M42, a number of months ago that I thought was particularly insightful and I really wanted to have a conversation with you about it.

It has to do with the notion of just very basically, what are some of the sources of competitive advantages, and then specifically, data as a competitive advantage and how that might change. Can you set the stage for us? Why is this something that you're grappling with today?

[01:37] Joshua Samuel: Yeah, sure. To take a step back, this notion of data being a competitive advantage, I think it's been around since the time of Google. If you go back to how Google search engine works, I think we've been aware of this kind of data flywheel, where more data on consumers equates to better search results and better targeting.

This concept of data being an advantage has been here for at least over a decade, at the very least. And what's happened in the past couple of years is just a continuation of the trend or reinforcement of the trend, if you want to call it that. And the big change that has occurred is model architecture, neural networks, and machine learning.

These have been around for decades, and I've listened to podcasts on this before. I think all these concepts have been designed at least three or four decades ago, or even more.

So don't shoot me if I'm wrong, but the general idea is that for every generation there are new model architectures being introduced, and this continually enhances the importance of data. And perhaps why we're having this conversation is the buzzword of the LLMs - Large Language Models.

This is just another kind of architecture that has been discovered in recent years. What is has done is provide an improvement on the neural networks that we've already been using for some time. With the transformer architecture, it further allows people to throw in more data and achieve even better output.

It's clear to the consumer now when you talk to chat GPT, if you've used it, maybe two years ago—and the research team has been playing around with this for at least two to three years now—we've seen step changes in the quality of output. It all boils down to parameter count, because there's more parameter count. And of course, the models themselves, there's been some architecture changes and some innovations in reasoning.

But in general, you're training the model with more and more parameters, and the output quality is continuing to scale. So, that's the underlying thoughts behind why we're thinking about data and what it means for moats.

[03:34] Rob Campbell: Well, you mentioned Google, and I think it's a great example as one that's very familiar to all of us. But even thinking back to Google's founding, it's not just these tech companies that have really built serious competitive advantages and real business models around this notion of data. You mentioned the quality and the availability of data increasing, really with the dawn of the internet age. It's far beyond the tech sector where we've seen these types of advantages accrue.

[04:10] Joshua Samuel: It's not just the tech sector. In the finance sector, on the FinTech side of things, we're also observing how you have new business models that are more tech-native, and they're approaching the lending sector with a blank sheet of paper. And, if this is the 21st century, how would we do lending, right?

So, we have FinTech companies like Kaspi, like Bajaj, and even some of the banks that we own. They're also migrating from old-school credit bureaus that were just based on a couple of data points about a person, like salary, and where they work, etc. And they're now embracing the era of data because there's so much data that's available.

How do we use all these data points to better inform our decisions? It's not just tech companies. It's gone into your FinTech and finance companies, and it's also going into e-commerce. How it's targeting people is also data driven. And that's on the B2C side.

On the B2B side, there's how corporates make decisions, ERP companies, and a whole bunch of SaaS companies. It's all built around data and how we use data for an advantage. And then, of course, don't get me talking about defense. I can do a whole podcast on data and defense!

I won't do that today, but even in the defense space, I was just at Rheinmetall's investor day, and they were showing how they're incorporating data and analytics into helping the militaries better target and automate the warfighting. Ultimately, it improves the decision and speed of decision.

I think that's the key with data, in any field. And that results in, most of the time, better outcomes for businesses.

[05:46] Rob Campbell: I'm just thinking, in the sports world, you've seen how data and data analytics have really reshaped the way that decision making takes place. Both in terms of how games are played, and in-game decision making. It's everywhere. But you mentioned earlier, the notion of LLMs and the step change in terms of how we need to think about data advantages.

Can you give us a framework for how you think of what a robust data advantage or a data moat looks like today? Are there characteristics of a particular data advantage that you're really looking for as you scan the world for investment opportunities?

[06:17] Joshua Samuel: Yes, good question, Rob. I think there are four key points here when we think about data moats.

Number one, is it unique, exclusive, proprietary data?
Number two, is the data continuously refreshed or dynamic?
Number three, is it high-dimensional, interactive data?
And number four, is it closed-loop data?

And it can be any of these four, and just to clarify, they're not mutually exclusive. Having one of the four is very good. You may not have all four together, but those are the four key points that I have in mind when thinking about data and data moats.

[06:50] Rob Campbell: Okay, and I presume that underlying all this is the data actually has to be useful to somebody for it to have value. I might have a whole bunch of chicken scratch on my notepad here, but it's largely useless information.

[07:01] Joshua Samuel: Precisely. I mean, there's no point having unique data that's useless.

[07:05] Rob Campbell: Yes. Can we now unpack each of those four categories, starting with that unique piece?

[07:09] Joshua Samuel: Yeah. So proprietary data—the easiest example that everyone would be familiar with—is the social graph. The information that Meta, Instagram, Facebook, WeChat has on you and your social network, your likes and dislikes, that is very, very valuable data.

I think the easiest analogy here is that there was a study done 10 years ago, in 2015, that said with about 300 likes or interactions on Facebook, the company is able to know you better than your spouse. I find it very funny because it's probably true and you're getting shown ads or stuff that is really in line with what you're thinking. And sometimes, it's even able to figure out what you like before you even figure out you like it. Spooky.

[07:51] Rob Campbell: Yeah.

[07:52] Joshua Samuel: I think that is the social graph at play. There's also interaction logs, which is a bit of an overlap with social graphs. What you're searching for on Google or Amazon, and even what you hover on when you're scrolling on a feed—I think all that gets locked somewhere. There are data tags associated with it. For the consumer side, that's proprietary data that a few companies have which is difficult to replicate but very useful.

Outside of consumer, let's talk about pharmaceuticals. We have real-world test data, scientific data that is when you're folding molecules or when you're trying out different combinations of molecules. Do they work? Do they not work? That sits somewhere within our pharmaceutical companies. That's proprietary and unique. And you would have to replicate a ton of experiments to get that same data set.

And the last category here would be like sensor data. Think about Tesla and self-driving cars. There's a lot of data that went into training these vehicles. And even SpaceX, when they recovered their rockets, there's trial and error, but there's unique proprietary data from all the sensors on the rockets for it to be able to land so precisely. That's number one: the proprietary data set that is a valuable moat.

[09:02] Rob Campbell: That makes a lot of sense. That's long been the case. Your second category was data sets that get refreshed frequently. It seems to me like that might be an area in which AI has changed the notion of what that might look like. Am I right?

[09:14] Joshua Samuel: Yes. To some extent, I think there is debate going on here. And let me inform the user what continuously refreshed data looks like.

I think the easiest example to give would be on the FinTech or lending side of things. You need data that is up to date. You do not want my credit score from three years ago or even two years ago. You want it right now and can ask “What's Josh’s most up-to-date credit scoring?” Even with credit scoring, the data being captured has evolved over time. It used to be very simplistic, you know, not one or two data points, but less than 10.

And now with more sensors out there, and I'm using the term very loosely, stuff like: where you live, your cell phone logs, where you spend most of your time, etc. Whatever can be measured and taken in can be fed into a model with more variables. It's probably able to give you a better understanding of the credit profile of the person, the risk level of the person.

And it doesn't end there. It also goes on into classified. If you think about online classifieds, like you want to buy a car, or buy a house, you need the most current data. You don't want last year's catalog. These are examples where you need up-to-date, continuously refreshed data.

[10:22] Rob Campbell: And just to stay on the second one, am I right in that the frequency of those updates is that much more important today for that data moat to be as strong as it was maybe 10 or 20 years ago?

[10:34] Joshua Samuel: Yes. I think it's always been important. What's changed was the introduction of LLMs. I don't think that changes the frequency of update because buying a house pre LLMs and after LLMs, it's probably the same. Where LLMs would come in is maybe the user interface or the way that you buy the house could evolve. So, you can imagine now going onto one of these online classified websites and specifying, "I want a house that has five windows, two doors, nice mountain view, far away from a main road." It is a different overlay that goes on top of that. But I think the frequency, in my opinion, should be the same.

[11:12] Rob Campbell: Okay. Got it. The next two categories, or the last two categories should I say, those are terms I was less familiar with. Can we start with the first one?

[11:21] Joshua Samuel: Sure. This is very interesting because I got these concepts from two gentlemen who are building an agentic AI software company. It's called Axium.

It's very interesting because one of them is formerly from Databricks. And Databricks, for your information, is one of the data warehouses behind, I can't give you the name, but it's behind one of the best LLMs. It is the most frontier technology and it was explained to me by them.

The first category is high-dimensional interaction data. In simple terms, high-dimensional means large variable. It's interactive data. You can think about this as, it's not just transaction logs, but it is also communication between buyers and suppliers. It can be between customer and the company. It can be also even how human agents resolve problems that have come up.

And this is very valuable data because it's multi-variable. It's not fixed and defined. Solving a problem, let's say a customer service log, it's the same, but it's also slightly different every time.

So, it's multiple variables and it's interactive. With this data, it overlaps with the unique proprietary data, but the additional caveat is the interactive portion and how it has so many variables. This enables you, especially in the age of LLMs, to then have much better outcomes because you're dealing with real world situations.

And I think we're just at the starting point here because this is more B2B. Everything that you and I use on ChatGPT, that's very B2C, it's like a one-way kind of interaction. But here on a B2B level, if you think five years from now and how a company will employ agentic AI, you need more of these kinds of datasets for it to be useful.

The second category, and there's a bit of overlap here, is called a closed-loop. Action and outcome combined together in a dataset that encompasses both. This is valuable because you have the feedback loop in the data. You know that taking a certain action results in a certain outcome.

For example, in a procurement process, if you're going to get an agentic AI to eventually help your procurement team, you need to know historically if these are the items we have procured.

And then, you need to check five years from now, what were the outcomes? Such as which equipment broke down early and which did not etc, stuff like that. Once again, this gives you a very rich dataset that you will be able to train your AI with.

Why I brought this up is because I think a lot of focus right now is on B2C, as it's the most clear and evident. However, with B2B, there is so much data that is still unlocked. We're just scratching the surface here and the reality is, it's going to take time. Corporates need to get the data in place first. You can't apply agentic AI if you don't have the datasets ready to use.

So, these are the four categories of data, which when we think about moats long term, having this data and having it ready to use, that's going to give company an advantage.

[14:12] Rob Campbell: Okay. And as you said at the outset, having one of them is great, but having multiple of these things and if they can interact is even better with respect to the strength of that data moat.

[14:20] Joshua Samuel: Yes.

[14:21] Rob Campbell: Can I ask you just the flip side? As we've seen advancements with LLMs, where are we seeing erosion in what might have been considered a pretty strong data moat historically and where those competitive advantages are perhaps no longer as strong as they once were?

[14:37] Joshua Samuel: It is once again linked to the first part. We talked about dynamic data. So, on the flip side, static data or slow-moving data or knowledge, that's probably easier to disrupt. And what we've seen, even in our portfolio, is databases like scientific journals, legal data, even medical data.

And I mean the caveat that, because there is a difference between what I mentioned earlier: when you're doing scientific data for drug discovery, that's proprietary, but then there's the medical data and when you have certain indications, such as what that disease is likely to be, then that's more open source.

If the data is publicly available, I think there are risks that your LLMs, which are trained basically on all of humanity's knowledge, have it in there somewhere. Right now, I think we're at the point that we've exhausted all the written words that humanity has had collectively.

Somewhere in the LLMs parameter count, when it was training these models, is data on legal issues. There's also data on medical issues. This data that is slow moving in nature, it doesn't change from time to time, such as classifying if someone has the flu, the indication is not going to change. It's not going to evolve that rapidly. That is just one example.

And there has been debate around Wolters Kluwer and RELX, and are these businesses going to be affected? We can always dive into that if you want, but these are examples where there's some question marks around data and the moat around the data. How useful is it really?

[16:02] Rob Campbell: And I imagine another category might be, the data is not out there, but it can be almost synthesized based on data that's out there.

[16:10] Joshua Samuel: Yes. I attended a conference earlier this year, and there was a startup company, I think it's called Hippocratic AI. And the founder was saying that ChatGPT can actually diagnose most of your diseases.

Recently, one of our research team members shared on our channel that there was a study done that basically verified this. Your general purpose LLM can outperform even your specialized LLMs in diagnosing diseases. You can think of it as the mosaic theory. Even though the LLM doesn't have access to some of these proprietary

datasets, because it's trained on so much, it can piece together enough information to give you outcomes that are pretty decent.

And it's not just in medical; it is even in stuff that is paywall. If you try hard enough, and you just put the content there, you can reverse engineer probably what's behind the paywall if you wanted to. It is a mosaic theory.

And the fact that the newest models are so good at reasoning, they can figure stuff out without having access to the full information. That's the other risk. And I think this is more like a generalist approach where the general models are getting very good at this, especially with a reasoning function.

[17:18] Rob Campbell: How do you think about the criticality of data to a customer? Because I can imagine if the LLM can synthesize and get 98% of the way there, in most cases, as the stakes go up to your customer, that 2% can be really costly. Does the criticality of the data to a customer's business processes matter too in this context?

[17:40] Joshua Samuel: It's great that you brought that up because if we talk about Wolters Kluwer and RELX, you're talking about legal and medical data. There's little room for error and there's an element of trust. So, it's critical. At a B2B level, there's a level of trust. Are you going to get fired for using ChatGPT to diagnose a patient, or are you better off using the previous vendor you've been using for the past 20 years?

In theory, the data is widely available and it's easy to synthesize. But the moat of these companies doesn't just lie in data. There is a trust element that is also very important. And especially in B2B settings, that's worth something. Data is just one possible moat.

There could be many other moats that a company has. For some of our companies, the data side is being attacked, but they're not a one-trick pony. There's other stuff. Having the customer relationships on a B2B level, that is quite difficult to do.

[18:31] Rob Campbell: Well, that stickiness, that trust and reputation takes time to build. Yeah. Josh, in your view, as you look out at the world, what company or set of companies are best positioned with respect to a data advantage? Or to put it differently, where are you seeing the most formidable data advantages out there in the world today?

[18:50] Joshua Samuel: Let's go back to the list of the four key points that I mentioned on proprietary data, dynamic data, high-dimension interaction data, and closed-loop data. I'm on the international equity team and my expertise is more ex-North America, but the easiest one for me is probably Tencent.

If you think about it, they have data on 1.3 billion consumers in China. That's a huge amount. And if you think of all the four categories, they have all four. The data is proprietary. They have a social graph, for example, that's continuously refreshed. They also have transaction data. Tencent is not just a WhatsApp of China. They have payment rails, they're like the visa of China. So, they have all the transaction data that's also continuously refreshed. They have high-dimension interaction data. Tencent has mini programs. I think they have like 10 million corporates or mini businesses. There's a huge number of corporates that are sitting on their platform and there's interaction data between the mini programs and the customers, including the chat logs. It's all getting captured within the Tencent ecosystem. They have that high-dimension interaction data. And is it closed-loop? Yes, all of it is closed-loop. The entire point of what Tencent is trying to do in e-commerce is to own the top of the funnel. The top of the funnel is when you first start interacting or thinking about a purchase, all the way down to purchasing something.

Tencent owns the customer journey right from start to finish. And this is different from say, e-commerce, where e-commerce is most of the time at the bottom of the funnel. So, they must acquire you by placing ads on TikTok,

and Instagram etc. Tencent owns the whole stack making it closed-loop from start to finish. When you think of all those four categories, they have it all.

It is a very good starting point, but it doesn't end there. To the point that businesses can have many moats, if you think about where LLMs or this whole agentic AI development is going to go, one thing I'm very excited about is AI agents that can help you do things.

Think about booking a holiday, or about shopping for yourself. For example, you need to top-up your fridge. So, you take a photo of your fridge, and the AI agent knows what's missing or not and does your shopping for you.

For this to happen, you need to have permissions because in most ecosystems, WhatsApp for example that Meta owns, they own you as the consumer. You have their apps, such as Instagram, and WhatsApp, but they do not own the payment rails. They do not own the e-commerce. They need permission for their AI agent to then interact with Visa, Mastercard, or maybe Amazon, to go and buy your groceries.

In Tencent's case, because they own the full stack, they have 10 million mini program businesses on it. In theory, they could facilitate that whole transaction seamlessly from start to finish. I think that's very valuable because they don't have to ask for permission from a third-party app. It's all on their platform and the permissioning is done since they have payment rails to facilitate transactions.

When you think of the next step in AI, and this is not immediate, I think there are many problems that need to be solved. But AI agents that can take economic action, that is where we're going. I think Tencent's very well positioned for that, and it's not something that's been focused on by the street right now. In the near term, what I've talked about is vision, that maybe five years from now, we'll get there.

Even in the near term, AI and LLMs have already had a big impact on them because Tencent has multiple apps within this app. They're most well-known for WeChat, but they have Tencent Video, the games business, and data on different users that is being collected. Historically, it's been very difficult to unify all this data.

With LLMs, I'm being told that it's ushering in a new era because you can basically collect all of your interactions across different platforms. Instead of having to do patchwork of algorithms to try to unify it all together, you just throw all your data into a LLM model. This is going to figure out all the stuff you like, and they can then target you better.

In the next couple of quarters, you'll probably see a continuation of this trend where Tencent is able to grow ad revenue double digits, despite China having horrible macroeconomic conditions because they're getting better and better at targeting. Short term or long term, I think it's a very nice story that's playing out there for Tencent, and it all boils down to data in most cases.

[23:06] Rob Campbell: Just the sheer amount of data that Tencent collects on its customers, given the products that they've rolled out over the years and the capture that they've had, it all sounds extremely compelling.

I imagine there are risks though as well, whether it's data privacy or just data concerns, perhaps privacy concerns. Does that not become that much more important, as you mentioned, as all these disparate data sets can interact? There's a part of it that is real scary.

[23:32] Joshua Samuel: Yes, I totally agree with you because we are reaching the point where these systems will probably know more about you than not just your spouse, but you, yourself. You can ask ChatGPT for your Myers-Briggs profile and it's able to guess based on your interactions. So you can imagine that if this thing is collecting data, and it's not just Tencent but for all tech companies and the amount of data they collect, it's

capturing stuff from your subconscious. It's down to mapping what's going on in your subconscious mind because it's manifesting in how you click and how you're scrolling.

Yes, I think data privacy is a very real issue. For Tencent specifically, I think they're actually quite ahead of the curve. Historically, WeChat has been run by a guy named Allen Zhang, and he's been very conservative with sharing the data with the rest of the group. So, if anything, I think they've been putting the brakes on that. In very conservative, I mean with how they're monetizing it. But in general, I think the direction is that they have to. If they don't do it, someone else will. They're probably aware of that.

This is almost an ethical question because how much is too much? The amount of data being collected allows people to really map your subconscious. I think that's really scary.

If you think of shorts videos, I've heard that people are being shown stuff for e-commerce that they didn't even know they wanted but somehow want them. I don't know where this goes, but that's just data science at work. It's taking all these data points and trying to make sense of it. And we humans may not. I think we are reaching the point where we do not understand how these models work.

[24:58] Rob Campbell: I want to ask you another question. I'm struck that you chose Tencent as a company that, I think the question was, has the most formidable moat and that is the one that came top of mind for you.

What strikes me about that is that is not new, as I think that's always been a feature of Tencent. Part of the discussion today has been: how does the nature of a data moat shift in a world of LLMs?

Should I interpret from your comments, that despite changes... and we started the discussion with Google and the so-called Magnificent Seven, which really capitalized on creating platforms that really harnessed a vast array of data...should I interpret from your comment, that despite these changes and real shifts occurring right now in terms of what a data advantage is, that it's those incumbents who are probably best positioned for the era going forward?

[25:48] Joshua Samuel: Yes, I think that's fair. I think Google, and internally we've talked and held this for a while, that some of my colleagues think Google is probably better than Tencent. I haven't looked at it in depth enough to give an informed opinion, but from what I've heard in the debates in our Teams chat, it seems reasonable. On top of all the stuff that Tencent has, they [Google] also own the TPUs down to the hardware, where they are designing the TPUs.

But yes, I think the Magnificent Seven, or at least some of them, are in a good position. Now, the caveat here is that, will there be a platform shift in the way people interact with these LLMs? I think that's the risk for all these companies, including Tencent.

So being well positioned doesn't guarantee you win. It means you have a head start. The odds are in their favour. So, to answer your question, yes, the odds are in their favour. It’s not a sure win, because you've seen ChatGPT come out. It's become a verb, “I'm going to ChatGPT this,” right?

[26:43] Rob Campbell: We've probably used it 20 times in this podcast.

[26:44] Joshua Samuel: Exactly. We should be saying, “I'm going to LLM this,” or something, but we're not. There is an example where, if a new platform comes out, such as AR or VR glasses, where the way in which people interact with the world changes, then these incumbents will need to be fast to adapt to that trend and make sure they're on it. For example, if AR glasses come out and WeChat is not part of that communication process for you

to speak to someone else, and that becomes the de facto kind of mode of communication, then Tencent is in trouble.

Similar for Google, if the way in which people interact with these apps changes over time, that's the big risk for them. But they have data. They have all the data and conditions are in their favour right now. They would have to screw up on execution in order to lose the race.

[27:32] Rob Campbell: Fascinating. Well, Josh, I think this is podcast number three for you and me this year, and I hope we have three more next year. It is always a pleasure speaking with you and hearing your thoughts.

[27:41] Joshua Samuel: Awesome. Thanks for having me on board.