

BEYOND THE MODEL

The Foundation Nobody Funds: Why AI Stalls on the Data, Not the Model

On the persistent gap between the appetite for machine learning and the willingness to pay for the data it must learn from

Giovanni Leonardi

July 2022



The Foundation Nobody Funds: Why AI Stalls on the Data, Not the Model

Transformation · No. 280

Originally written July 2022 | Re-edited and re-rendered for the WhereBeyond Edition, 2025

"The intelligence in an intelligent system turns out to be mostly plumbing."

1. When the Demo Outruns the Data

The demonstration went well. It usually does. On a screen in front of the steering group, a demand-forecasting model trained on three years of trading history called the next quarter to within a few percent, and for twenty minutes the room believed the hard part was over. Someone asked when it could go live. The analyst who had built it said something careful about "getting it production-ready," and the question moved on. The minutes recorded a success. What they did not record was that the model lived inside a single notebook on a single laptop, fed by a table the analyst had assembled by hand across several late evenings, stitched together from three systems that could not agree on what a customer was.

Thirteen months later, that model was still not live. Not because the mathematics was wrong — the mathematics was never the problem — but because no one could reliably reproduce the data it had learned from. The definitions existed only in the analyst's head. The joins existed only in her notebook. And the moment anyone tried to run the same logic on this week's data rather than last year's extract, the numbers drifted, because the source systems had shifted underneath it in ways nobody had catalogued.

This is not an unusual story. It is, in my experience, the usual one. We have spent the better part of a decade being told that machine learning will reorder the enterprise, and we have responded by hiring data scientists, buying platforms, and standing up centres of excellence. What we have been far more reluctant to do is fund the unglamorous work on which all of it depends: making the data trustworthy, reproducible, and owned. That work has no demo. It photographs badly. And so it goes unfunded, year after year, while the same organisations wonder why their models keep dying on the road between the notebook and production.

2. The Invisible Prerequisite

Say what the industry now repeats about the "AI project failure rate" out loud and something odd surfaces. The figures vary — the ones in circulation put the share of models that never reach production somewhere north of eight in ten — but the reason is remarkably consistent. When you trace a stalled initiative back to its root, you rarely find

an exotic modelling failure. You find missing data, or dirty data, or data that exists but cannot be got at, or data whose meaning nobody can pin down. The intelligence in an intelligent system turns out to be mostly plumbing.

The model is the part everyone can see, and therefore the part everyone wants to pay for. The data foundation is the part nobody can see, and therefore the part everyone assumes someone else has already built.

That asymmetry of visibility is the heart of the problem. A model produces a number, and a number can go on a slide. A cleaned-up customer master, a documented lineage, a pipeline that reproduces a feature the same way every time — these produce nothing you can show a board. They are the structural engineering of the building, and structural engineering only becomes visible when it is absent.

I want to be precise about what data readiness actually means, because the phrase has been worn smooth by overuse. It is not a data lake. It is not a platform procurement. It is a small number of unglamorous properties holding true at once:

- The data you need exists, and you can find it without asking three people where it lives.
- Its meaning is agreed and written down, so that "active customer" denotes one thing across finance, marketing, and operations rather than three quietly different things.
- It can be reproduced — the path from raw source to the feature a model consumes is a pipeline, not a heroic act of memory.
- Someone owns it, and that ownership survives the departure of the individual who happened to understand it.

None of these is technically hard. All of them are organisationally hard, which is a different and more stubborn thing.

3. Why the Foundation Goes Unfunded

If the prerequisite is this well understood, why does the pattern persist? Not through ignorance. It persists because the incentives arranged around it are almost perfectly designed to starve it.

Consider how the money moves. Investment cases are written around outcomes a sponsor can claim — a forecasting capability, a propensity model, a recommendation engine. The data work is folded in as an assumption, a line that reads "subject to data availability," and that line is where the risk quietly accumulates. When the initiative runs late, the overspend is attributed to the model, because the model is the named deliverable. The foundation, having never had a name of its own, cannot even be credited with the delay it caused.

Consider, too, how governance has been allowed to define itself. In a great many organisations, data governance means a committee, a policy document, and a compliance obligation — a mechanism for saying no and for satisfying the regulator. It very rarely means a capability for making data usable. The result is the peculiar situation in which an

organisation is simultaneously over-governed and unready: heavy with policy, light on pipelines; able to tell you who is accountable for the customer record, unable to hand you a clean one.

1. The work is invisible, so it cannot be demonstrated.
2. Because it cannot be demonstrated, it cannot easily be funded.
3. Because it is not funded as a thing in itself, it is smuggled into modelling projects as an assumption.
4. When the assumption fails, the modelling project takes the blame, and the foundation escapes scrutiny once more.

Round the loop turns, and each lap deepens the conviction that "the AI didn't work," when what actually happened is that the AI was asked to stand on ground that was never laid.

There is a human dimension beneath the structural one. Nobody is promoted for fixing the data. The reputational rewards in this field accrue to the people who build the clever thing, not to the people who spend eighteen months reconciling reference data so that the clever thing has something honest to learn from. Until that is no longer true, the most capable people will keep routing around the foundation rather than toward it — and it is hard to blame them.

4. The Objection, Taken Seriously

The strongest argument against everything I have said is not foolish, and it deserves its full weight rather than a straw version. It runs like this: data readiness is a bottomless pit. If you wait until the data is clean, governed, and perfect, you will wait forever and ship nothing. The whole lesson of recent years is to stop boiling the ocean — pick a narrow, valuable use case and let that use case pull exactly the data it needs into shape. Readiness follows delivery; it does not precede it.

I agree with almost all of that. A programme that sets out to make all the data ready before it will attempt anything is a programme that will spend three years producing a data dictionary and no value. Use-case pull is the right tactic. The demand for perfection is the enemy of readiness, not its ally.

But notice what the objection quietly assumes: that readiness is a gate you pass through once, rather than a capability you hold continuously. That assumption is the actual mistake, and the ocean-boilers and the move-fast camp make it alike. The first group treats the foundation as a giant up-front project. The second treats it as something that will accrete for free as a by-product of shipping models. Both are wrong in the same way. The foundation is neither a project nor a by-product; it is a standing asset that decays when it is not maintained, exactly like any other piece of infrastructure.

"Data readiness is not a gate you pass through on the way to the model. It is a capability you either hold or lack when the next model arrives — and there is always

a next model."

The reconciliation, then, is not "foundation or delivery." It is this: let the first use case pull a thin, real slice of the foundation into existence — but fund that slice as a permanent, owned thing rather than as scaffolding to be discarded once the model ships. The pull tactic builds the right piece. The mistake is throwing the piece away.

5. What Changes When You Fund It

Return to the forecasting model that sat unlaunched for thirteen months. What eventually moved it was not a better algorithm. It was a decision that felt, at the time, like an admission of defeat: the sponsor stopped funding a second model and redirected the money to the foundation beneath the first. A single owner was made accountable for the customer definition across the three systems. The hand-built table became a documented pipeline that ran on a schedule and produced the same features the same way every time. The reconciliation logic came out of the analyst's head and into version-controlled code that outlived her eventual move to another team.

The numbers are worth stating plainly, because they are the mechanism and not merely the moral. The original model reached a convincing prototype in roughly six weeks, then consumed thirteen months without shipping. The second model the team attempted — once the foundation under it was real — reached production in under a month, because the expensive part had already been built and, crucially, had been built to be reused. The foundation did not slow the second model down. It was the reason the second model was fast. That is the whole argument in one before-and-after: the work that looked like a tax on the first initiative was in fact the enabling capital for every initiative after it.

Where the effort goes	The model-first way	The foundation-first way
First initiative	6 weeks to prototype, 13 months stalled	Slower start, foundation built once
Each later initiative	Rediscovered the same data problems from scratch	Reuses definitions, pipelines, ownership
Where blame lands when late	On "the AI"	On a named, fundable gap

6. The Argument, Plainly

None of this requires believing that models do not matter, or that data science is overrated. It requires only noticing where the difficulty actually lives, and being honest enough to fund it there. The organisations that get real value from machine learning in the years ahead will not, for the most part, be the ones with the cleverest models. They will be the ones that treated their data as an asset to be deliberately built and owned,

rather than a raw material assumed to be lying around.

We are, as a profession, fluent in the language of models and fluent in the language of governance. We are far less fluent in the plain, expensive, unglamorous language of making data trustworthy — and it is precisely that fluency the moment now demands. The next demo will go beautifully, as demos do. The only question worth asking in the room afterwards is whether anyone has funded the ground the thing is standing on.