



BEYOND THE HIRE

Data Scientists Without Data Infrastructure: Why Analytics Ambition Outruns Reality

On the widening gap between analytics ambition and data reality in the enterprise, 2014

Giovanni Leonardi

March 2014

Contents

1. Executive Summary	3
2. The team that arrived before the plumbing	3
3. The glamour and the plumbing	4
4. Reading the wrong scarcity	5
5. The proof-of-concept that hides the problem	6
6. The case for hiring first	6
7. Why the forcing function misfires	7
8. What the pattern reveals	8

Data Scientists Without Data Infrastructure: Why Analytics Ambition Outruns Reality

Transformation · No. 228

Originally written March 2014 | Re-edited and re-rendered for the WhereBeyond Edition,
2025

*“An organisation can hire brilliance in a quarter; it
cannot hire, in a quarter, the years of patient
plumbing that brilliance needs in order to be of any
use.”*

1. Executive Summary

In the space of three or four years, the data scientist has travelled from a job title almost nobody recognised to the most coveted hire in the enterprise. Boards approve the headcount without the usual haggling; recruiters chase the same few hundred people across the same handful of cities; the trade press has taken to calling it the defining profession of the decade. The ambition behind all this is genuine, and so, very often, is the talent. Yet walk into many of the teams that have been assembled at such expense and you find people of real ability spending most of their week not doing the work they were hired to do. They are hunting for data across systems that were never meant to talk to one another, reconciling it by hand, exporting it to their own laptops, and apologising in meetings for numbers that refuse to agree.

This essay is an attempt to understand why so many organisations keep making the same mistake — buying the science before laying the foundation the science depends on — and why the mistake is so stubborn that pointing it out rarely seems to cure it. My argument is that the pattern does not persist through ignorance. It persists because the visible half of analytics is easy to buy and the invisible half is hard to fund; because leaders consistently misread which resource is actually scarce; and because the one artefact that ought to expose the problem — the dazzling proof-of-concept — instead conceals it, and defers the reckoning to a later quarter and, often, a later hire. Along the way I want to take seriously the strongest case for hiring the scientists first, because it is not a foolish case, before explaining why in practice the forcing function it relies on so often misfires. The pattern, read closely, is a small and honest mirror of a much larger gap: the distance between what we say transformation is and what we are actually willing to build.

2. The team that arrived before the plumbing

Picture a mid-sized consumer bank early in the big-data years. The strategy is sound and unremarkable: the organisation believes, correctly, that it is sitting on a fortune in customer data and using almost none of it. A new head of analytics is hired at some expense, and given a mandate and a modest team — three analysts, two of them with doctorates, all fluent in the tools that the moment has made fashionable. There is a launch. There is a slide with the phrase actionable insight on it. Expectations, quietly, become enormous.

Within a quarter the picture on the ground looks nothing like the slide. There is no single place where the bank's data lives. The current account system knows one version of the customer; the mortgage platform knows another; the cards business, acquired years earlier and never fully integrated, knows a third. None of them share a common customer identifier. The warehouse that was meant to unify them was built for regulatory reporting, refreshes overnight in a batch that frequently fails, and contains perhaps a third of the fields the analysts actually need. So the team improvises. Someone sets up a read replica of a production database as a favour. The analysts write long, defensive SQL against it, export the results to comma-separated files, and clean them in R on their laptops. The most senior person on the team, hired to build predictive models, spends her days as what people have taken to calling, only half in jest, a data janitor.

Put a number on it and the waste becomes vivid. Three scientists on premium salaries — comfortably north of a quarter of a million pounds a year, all in — spending, by their own rueful estimate, four days in every five locating, moving, and repairing data, and one day doing the analysis for which they were recruited. The organisation is paying the rarest talent it has ever hired to do the least skilled work in the building. The first genuine model the team ships — a decent, defensible churn predictor — takes nine months, and eight of those months have nothing to do with modelling. They are spent discovering, painfully, that the same customer appears under four different keys in four different systems, and negotiating, meeting by meeting, a way to stitch them together that everyone can live with.

None of this is a story about weak people. The team is good. It is a story about an organisation that bought the top of the pyramid and assumed the rest of the pyramid would somehow already be there.

3. The glamour and the plumbing

Why does an intelligent organisation do this? The beginning of an answer is that the two halves of an analytics capability are wildly unequal in how they present themselves to the people who hold the budget.

Hiring a data scientist is a legible act. It produces a name, a face, a start date, a line in the board pack. It photographs well. It signals, to the market and to the organisation itself, that the firm is serious about the future. And it is fast: a good recruiter can turn the hire around in weeks. Building the infrastructure beneath that hire is the opposite in every respect. It is slow, and it is invisible until the moment it is missing. Nobody is promoted for the pipeline that quietly did not fail last night. There is no launch event for a reconciled customer key, no board slide that reads we spent six months agreeing

what a customer is. The work is plumbing, and plumbing is admired only in its absence.

The visible half of analytics can be bought in a quarter. The invisible half has to be built over years, by people the organisation has not thought to value, against a return that will not photograph well. Almost every incentive in the building pushes towards buying the first and hoping about the second.

There is a deeper asymmetry beneath the visible one, and it is a matter of temperament as much as of budgets. We are, as a profession, far more comfortable acquiring capability than constructing it. Acquisition feels like progress; construction feels like cost. The vendors and the commentary of the moment reinforce this at every turn, because they sell the destination — the insight, the edge, the model that sees what competitors cannot — and are quiet about the journey, which is mostly integration, definition, and the grinding improvement of data quality. When the whole surrounding narrative celebrates the arrival and ignores the road, it is no surprise that organisations budget for arrivals.

4. Reading the wrong scarcity

Underneath the incentives sits a simple analytical error, and it is worth stating plainly because so much follows from it. Leaders embarking on this journey almost universally believe that the scarce resource is talent. The people are hard to find, the salaries prove it, the recruiters confirm it, and so the mental model forms: get the scarce people and the value will follow. It is an intuitive model and, in this setting, it is wrong.

For most organisations at this stage, analytical talent is not the binding constraint. The binding constraint is the availability of data the talent can actually trust and reach. A brilliant scientist with no clean, accessible, well-defined data is idle, or worse than idle — she is expensively idle, and she knows it, and she updates her CV accordingly. A competent analyst with a good pipeline and a trustworthy definition of a customer will out-deliver her every week of the year. The bottleneck is not the mind. It is what we feed the mind.

“The scarce resource was never the scientist. It was data the scientist could trust.”

Misreading the scarcity has a compounding cost, because it sends the organisation back to the same well for the wrong bucket. When the celebrated hires fail to produce the promised transformation, the instinct is rarely we did not build the foundation. It is more often we need better people, or more people, and so a second round of hiring is approved while the actual constraint stands untouched. An organisation can hire brilliance in a quarter; it cannot hire, in a quarter, the years of patient plumbing that brilliance needs in order to be of any use. Confusing the two is the central error, and it is remarkably durable.

5. The proof-of-concept that hides the problem

If the foundation is so plainly missing, why does the reckoning not arrive quickly and force a correction? Because of a particular and seductive trap, and it deserves its own name.

A capable data scientist, handed a single well-defined question and permission to work around the organisation rather than through it, can produce something genuinely impressive in a few weeks. She will pull an extract, clean it heroically by hand, build a model on that curated slice, and present a result that is real, defensible, and exciting. The room is delighted. The hire is vindicated. The pilot becomes the centrepiece of the next steering committee, and the number in it — the uplift, the saving, the accuracy — is repeated until it hardens into an expectation.

Everything about that success is true and almost everything about it is misleading. It worked because it bypassed the very problem the organisation most needs to solve. The heroic hand-cleaning does not scale; the curated slice cannot be refreshed; the extract cannot become a production feed without the pipeline that does not exist. The proof-of-concept proves the talent and disproves nothing about the foundation — yet it is read, precisely backwards, as evidence that the foundation is adequate. So the moment of reckoning, instead of arriving, is deferred. The organisation commits to scaling something that cannot be scaled, and the gap between the pilot and the production system — which is almost entirely a gap in infrastructure — becomes someone's problem next year.

- The pilot succeeds by working around the missing foundation, not on top of it.
- Its success is read as evidence that the foundation is sound.
- The organisation commits to scaling on that false reading.
- The infrastructure gap resurfaces at production scale, larger and more expensive, and is inherited by whoever is unlucky enough to be there when it does.

6. The case for hiring first

It would be too easy to end there, with a tidy moral about plumbing before people. The honest difficulty is that the opposite sequence has a serious argument behind it, and anyone who has watched infrastructure programmes fail will feel its force.

The case runs like this. Infrastructure built speculatively, ahead of any concrete demand, is one of the most reliable ways to waste money that large organisations possess. We have all seen the enterprise data warehouse that took three years and tens of millions to build, was architected to answer every conceivable future question, and answered almost none of the actual ones because no real user was in the room while it was designed. Build the plumbing first, the argument goes, and you will build the wrong plumbing — elegant, comprehensive, and unused. Far better to hire the scientists, let them collide with the organisation's data reality, and use their well-paid, highly visible frustration as the forcing function that finally justifies the investment. The scientists

generate the demand that makes the infrastructure fundable; their pilots prove the value that unlocks the capital; their pain, loudly expressed, tells you exactly which pipes to lay first, because it tells you which data people actually want. Demand should pull infrastructure into being, not the other way around.

This is not a weak argument, and I have some sympathy with it. Infrastructure without a user is a genuine failure mode, and it is at least as common as the one this essay is about. Demand-led building is usually wiser than speculative building. If the story ended with the organisation reading its scientists' frustration correctly and funding the foundation in response, the sequence would have worked exactly as designed.

7. Why the forcing function misfires

The trouble is not with the logic. The trouble is that the forcing function relies on an organisational learning loop that, in practice, is broken — and it is broken by the very same invisibility that caused the problem in the first place.

For frustration to function as a signal, someone with budget has to read it correctly: to hear a scientist say I spend four days in five moving data by hand and conclude we must build the pipeline, rather than this expensive person is underperforming or perhaps we hired the wrong sort. But infrastructure is invisible, and invisible causes are the hardest to credit. It is far more intuitive to blame the person you can see than the pipeline you cannot. So the frustration is frequently misread as an individual failing rather than a systemic one. The scientist, sensing that the diagnosis has gone wrong and that nothing structural will change, leaves — these are, after all, the most mobile employees in the organisation. A replacement is hired, at similar cost, into the identical conditions. And here is the quiet tragedy of the pattern: because each hire arrives fresh and the infrastructure lesson was never institutionally absorbed, the loop does not tighten with each turn. It simply repeats. The organisation experiences the same frustration several times over without ever converting it into the investment it was supposed to force.

Two readings of the same frustration	What follows
"Our expensive scientists are underperforming"	Churn the people; re-hire into identical conditions; repeat
"Our scientists have nothing they can trust to work with"	Fund the pipeline, the definitions, the access; the next hire is productive in weeks

So the pro-hiring argument is right about sequence and wrong about assumption. Yes, demand should pull infrastructure into being. But that only works where leadership can read the demand as a signal about the foundation, and the whole difficulty of this domain is that leadership systematically cannot, precisely because the foundation is the part nobody can see. Hiring first is defensible in an organisation with a functioning learning loop. In one without it — which is most — hiring first simply means paying, repeatedly, to relearn a lesson it never manages to keep.

8. What the pattern reveals

I have spent this essay on a narrow phenomenon — analytics teams starved of the data they were hired to use — because it is such a clean and legible instance of a much larger one. The gap between analytics ambition and analytics reality is a specimen of the gap between transformation intent and transformation reality, and it fails in the same way, for the same reasons.

We are fluent, as a profession, in the language of destinations. We can describe the data-driven organisation, the customer-centric bank, the insight-led business, with real conviction and in considerable detail. We are far less fluent, and far less willing, when it comes to the years of unglamorous foundational work that any of those destinations actually requires. We fund the arrival and starve the journey. We buy what is legible and neglect what is not. And when the destination fails to materialise, we reach again for the legible lever — another hire, another tool, another restructure — because it is the one we know how to pull, rather than the invisible foundation whose absence is the real cause.

The correction is not, I think, primarily technical, though it has technical components. It is a correction of attention. It means learning to fund the invisible, to value the people who build foundations as highly as the people who stand on them, and — hardest of all — to read a talented person's frustration as evidence about the system rather than about the person. Sequence the science and the plumbing however the situation demands; there is no single right order. But do not mistake the visible half of the work for the whole of it. The organisations that eventually pull ahead in this decade will not, I suspect, be the ones that hired the most data scientists. They will be the ones that finally understood what those scientists needed, and were willing to build it before anyone could take a photograph of the result.