



BEYOND THE CELEBRATED HIRE

Hiring the Scientist, Forgetting the Laboratory

An essay on the gap between the promise of big data and the quiet plumbing it depends upon

Giovanni Leonardi

March 2014

Contents

1. Executive Summary	3
2. The hire who arrives to a cold laboratory	3
3. Why we buy the scientist first	4
4. The invisible layer, and why it has no champion	4
5. The forces that keep the pattern alive	5
6. But 'infrastructure first' is a trap of its own	6
7. A worked figure	6
8. What the pattern tells us about transformation	7

Hiring the Scientist, Forgetting the Laboratory

Transformation · No. 229

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

"The models work in the laboratory and die on the way to the branch."

1. Executive Summary

There is a familiar sequence in organisations that have resolved to become data-driven. First comes conviction — a board persuaded that data is the decisive competitive terrain of the decade. Then comes recruitment: the search for the scarce, celebrated talent, the data scientist lately anointed as the holder of the sexiest job of the twenty-first century, who will turn raw information into advantage. And then comes the quiet disappointment. Eighteen months on, the expensively assembled team has produced a handful of dashboards and a great deal of frustration, because most of its time has gone not into analysis but into archaeology: finding data, extracting it, cleaning it, and reconciling the three systems that each insist they hold the definitive record of the same customer.

This essay examines that pattern — analytics ambition running years ahead of the data foundations it depends upon — and asks why it persists among organisations that are neither naïve nor short of money. The argument is that the gap is not born of ignorance. It is sustained by a structural asymmetry between the visible and the invisible. A brilliant hire is visible; it signals ambition to a board, to a market, to a workforce. The pipelines, agreed definitions, access rights and quality controls that would make that hire productive are invisible; they signal nothing and photograph badly. Transformation, being as much an exercise in signalling as in delivery, reliably funds the symbol and starves the substance. Yet the essay resists the tidy conclusion that seems to follow — build the infrastructure first — because infrastructure built in the abstract, ahead of any real question, carries a failure mode of its own. What the pattern finally asks of us is harder than sequencing.

2. The hire who arrives to a cold laboratory

Consider the first month of a newly recruited data scientist: a doctorate in some numerate discipline, a portfolio of models, hired after a long search and against genuine competition. The expectation, held sincerely on both sides, is that within weeks there will be insight — a pattern in customer behaviour, a churn model, something the business has never seen before. What actually fills that first month is a hunt for the data. Much of it lives in a warehouse built for regulatory reporting, which was designed

to answer an entirely different set of questions. Some of it sits in a system the marketing function runs, to which access requires a request that joins a queue. The definitions refuse to agree: an active customer means one thing to finance and another to operations. Dates are stored three ways. By the close of the quarter the scientist has become, in practice, a very expensive data-cleaning resource — doing by hand, and in isolation, the work that ought to have been done once, centrally, and for everyone.

This is not the story of one unlucky appointment. It is close to the modal experience of the first wave of data science hiring, and its sameness across sectors and organisations is precisely what makes it worth examining. When the same disappointment recurs in banks and retailers and insurers that share little else, the cause is unlikely to be local incompetence. It is structural. Something about the way organisations approach data reliably produces the scientist before the conditions in which a scientist can work.

3. Why we buy the scientist first

The first reason is narrative. The story of the age is that data is valuable and that insight is where the value is realised, and the human embodiment of that story is the analyst-genius who sees what others miss. That story has a hero, and the hero is a person you can hire. It does not have a memorable role for the pipeline. When a board asks what the organisation is doing about data, we have recruited a team of data scientists is an answer that lands; we are re-platforming our customer data and reconciling our master definitions is an answer that produces polite nods and a change of subject.

The second reason is the shape of the decision. A hire is discrete, approvable, and owned: a headcount request, a salary, a start date, a name. Infrastructure is diffuse, long, and owned by no one in particular. It crosses every functional boundary, benefits everyone and therefore belongs to nobody, and its business case is the least exciting sentence in the language of change — that things will work better and that problems you cannot currently see will not occur. Faced with a decision that is easy to make and one that is hard even to frame, organisations make the easy one and postpone the hard one.

The visible investment advertises the ambition; the invisible investment delivers it. When the two compete for the same budget in the same quarter, the one that can be photographed tends to win.

The third reason is that the market sells it this way. The scarce talent and the striking tool are what the ecosystem is organised to supply and to celebrate. The pipeline is nobody's flagship product and nobody's conference keynote. An organisation forming its sense of what doing data looks like from the outside will absorb, without ever quite deciding to, the belief that the work is modelling and the plumbing is a detail.

4. The invisible layer, and why it has no champion

It is worth saying plainly what the neglected layer actually is, because part of its problem is that it resists a glamorous description. Data infrastructure is reliable access to data; agreement on what the data means; assurance that it is correct and current; knowledge of where it came from and how it has been transformed; and the machinery that moves it from the places it is created to the places it is used. It is the difference between a scientist who can pull a clean, trustworthy, documented customer table in an afternoon and one who must first spend six weeks establishing which of four tables can be believed.

At this moment the work barely even has a name. The title data engineer is only beginning to appear; for the most part the people who do this work are the ETL team, or are borrowed from the warehouse, or are the scientists themselves, quietly doing the plumbing they were not hired for and cannot put on their next curriculum vitae. A discipline without a name and without a hero has no natural champion inside the organisation and no natural advocate in the talent market. It is judged, when it is judged at all, by the absence of problems — and the prevention of trouble that never arrives is the least rewarded form of success in any institution.

5. The forces that keep the pattern alive

If the pattern were merely a mistake, it would be corrected once and learned from. That it recurs suggests it is held in place by forces stronger than any individual decision. Four of them seem to me to do most of the work.

1. The economics of the visible. Budgets flow toward what can be shown. A named team is a line a leader can point to; a message queue and a set of reconciled definitions are not. So the visible is funded to completion and the invisible is funded to the minimum that keeps the lights on.
2. The ownership vacuum. Data crosses every silo and sits inside none. The customer record is touched by marketing, sales, operations, finance and risk, and is owned, in any accountable sense, by none of them. Infrastructure that would serve all of them is therefore everyone's benefit and no one's responsibility.
3. The framing as cost. Infrastructure enters the conversation dressed as IT capital expenditure — a cost to be minimised — rather than as the thing that determines whether the analytics investment ever pays back. Framed as a cost, it competes to be small. Framed as the multiplier on a much larger spend, it would compete instead to be sufficient.
4. The talent market's own incentives. The scientist is trained, credentialed and rewarded for modelling, not for pipelines. Asked to spend a year building foundations, a capable analyst experiences it as a demotion and, quite rationally, as a threat to the very skills that keep them employable. The people best placed to notice the missing layer are the people least incentivised to build it.

Held together, these forces do not describe a failure of intelligence. They describe a system reliably producing an outcome that no one within it would have chosen on purpose.

6. But 'infrastructure first' is a trap of its own

The temptation at this point is to reach for the obvious correction: build the foundation before you hire the scientist. I want to resist it, because the organisations that took that lesson literally have assembled a graveyard of their own. It is populated by the multi-year enterprise data warehouse and master-data programmes that consumed budget for three years, set out to model the whole of the business in the abstract, and delivered a platform beautifully engineered to answer questions that, by the time it arrived, no one was still asking.

Infrastructure built without the pull of a real question tends toward completeness rather than usefulness. It tries to be right about everything and so takes years, and years is exactly the currency these programmes do not have. The scientist's questions are what reveal which data actually matters, which quality problems are worth the cost of fixing, and which of the four irreconcilable definitions of customer must be reconciled first. Remove the question and you remove the discipline that keeps the foundation honest.

*"The scientist without a pipeline does archaeology;
the pipeline without a scientist builds a monument.
The value lies in the narrow channel dug between a
real question and the specific data that answers it."*

So the true answer is not a sequence but a co-evolution. A real, narrow question pulls a real, narrow pipeline into being; that pipeline, once built, lowers the cost of the next question; and the foundation grows in the shape of the problems it has actually been asked to solve rather than in the shape of an architect's model of a business. This is less satisfying than a clean rule about what to do first, and considerably harder to put on a plan. It is also, I think, the only version that works.

7. A worked figure

Let me make the cost concrete, using a composite that will be recognisable to anyone who has lived through this period. A mid-sized retail bank resolves to become data-driven and hires a team of six — a lead and five scientists — at a fully loaded cost in the region of £600,000 a year. Eighteen months later the team has produced two dashboards that reach production and several genuinely promising models — a churn predictor, a next-best-action engine — that never reach a customer, because the data needed to run them in production cannot be refreshed reliably or quickly enough to be of use. The models work in the laboratory and die on the way to the branch.

The investment that would have changed this was neither large nor exotic: two data engineers and a modest customer-data pipeline, at perhaps a fifth of the annual cost of the science team. It was deferred twice. Not because anyone judged it worthless, but

because it had no sponsor, appeared on no slide about the bank's data-driven future, and could always wait one more quarter. The arithmetic is unkind. Something on the order of £600,000 a year was spent to purchase insight in principle and archaeology in practice, for want of a £120,000 investment that no one owned. The expensive resource was rendered ineffective by the cheap one that was never made.

8. What the pattern tells us about transformation

The data scientist without infrastructure is a specific instance of a general habit, and it is the general habit that is worth carrying away. Transformation, again and again, buys the visible emblem of the future and underfunds the invisible substance that would make the emblem function — because the emblem is legible to boards, to markets and to staff, and the substance is not. The emblem changes with the decade. It has been the strategy team, the innovation unit, the flagship system; today it is the data scientist. The dynamic beneath it does not change at all.

The discipline this asks of us is unromantic. It is to treat the pipeline as part of the capability rather than as the plumbing beneath it; to fund the two in the same breath rather than the symbol now and the substance later; and to resist the particular satisfaction of the visible hire — the appointment that feels like progress because it can be announced. The organisations that will, in the end, extract real value from their data are unlikely to be the ones that hired the most scientists. They will be the ones that made the unglamorous investment nobody thought to put on a slide, and thereby gave their scientists a laboratory to walk into rather than a building site.