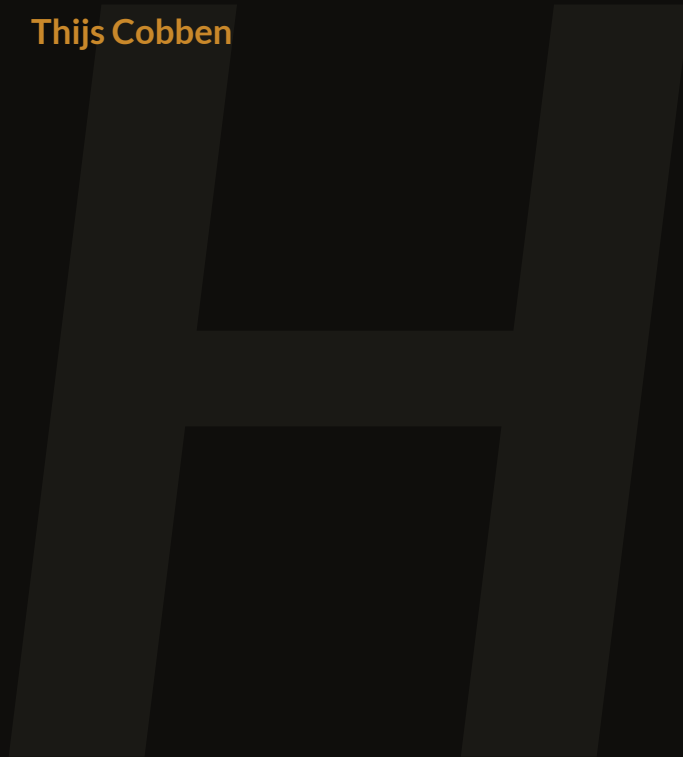

HISTORY OF DATA

From Cave Paintings to Lakehouses —
Forty Years of Patterns

Thijs Cobben



Contents

IN THIS SAMPLE

Foreword · 3

1. From the Origin of Language to the Birth of Data · 9

FULL BOOK — COMING SOON

2. The First Tablets — Sumer, the Edubba, and the Birth of the Scribe

3. Rome — Industrialising the Census

4. Meanwhile, in China and the Andes

5. The House of Wisdom and the Curated Word

5.5 The Bible as Data Product

6. Print, Ships, and the First Mass Data

7. The Wires Before the Web

8. Punched Cards to Mainframes — When the State Got a Memory

9. The Relational Decade — Codd, SQL, and the Reporting Hangover

10. Enter the Warehouse — OLAP, BI, and the Department That Owns the Truth

11. The Spam Years and the Birth of Modern Governance

12. The Lakehouse and the Age of AI

This sample contains the Foreword and Chapter 1. The full book traces 47,000 years of data history across twelve chapters.

FOREWORD

Foreword

Reading time: ~8 min.

Hagia Sophia. Holy Wisdom. The cathedral named after it, in Nicaea — today's *znik* in Turkey — was at the time of the council held there from the twentieth of May to the twenty-fifth of July, 325 AD, a freshly consecrated royal chapel, and Emperor Constantine presided over the gathering of some three hundred and eighteen bishops, priests, mystics and theologians from a modest golden throne.

"Division within the Church — the guardian of a Christianity that had just been declared the state religion — would be worse for the unity of his recently united Empire than the external wars of which he had won so many," or so he reasoned; and therefore, as he intended, it was time to arrive at a broadly supported, undisputed definition of what that "religion" actually meant. The council produced what would later be called the Nicene Creed. It declared the equal substance of God, his Son, and the Holy Spirit — the Holy Trinity. It fixed the date of Easter. It declared the teaching of the priest Arius, who held that the Son was not co-eternal with the Father, a heresy. The biblical canon itself was not settled here — that took two further councils, at Hippo (393) and Carthage (397) — but the doctrine of the Trinity was Nicaea's, and it stuck.

The nine-week conference ended with a thunderous feast, and the established dogmas would go on to shape European history for the next thirteen hundred years: from the alleged slap of Saint Nicholas against an Arian who refused to accept the Trinity, to the persecution of the Cathars and the Spanish Inquisition, but perhaps most of all — on a positive note — in the unity of Church and Faith that proved a powerful binding force in medieval Europe and thereby had enormous influence. Impact. Church and Pope were "infallible," so revision was, for most of that long period, hardly even thinkable. But the credibility of the authority began to crumble as conflicting information became easier to spread and more broadly accessible.

What I want you to notice — because you will see it again and again in this book — is the mechanism by which Nicaea installed a proposition that is, on its face, semantically paradoxical (three persons, one substance, fully equal, eternally co-existent) and made it durable for civilisations. The same mechanism, give or take an institution, kept Galenic medicine and the four humours running through every European medical school for nearly two thousand years; it kept Ptolemy's geocentric astronomy at the centre of the Western cosmos for some fourteen hundred years, with epicycle piled upon epicycle to make the observed data fit; it killed millions in Stalin's Soviet Union when Trofim Lysenko's anti-Mendelian agricultural genetics was elevated to state doctrine and dissenting biologists were sent to the camps. Authority, plus institutional curation,

plus the absence of permitted counter-evidence, manufactures a "truth" that drives behaviour for centuries — regardless of whether the underlying proposition can survive a careful reading.

A little less than seventeen hundred years later, a Lead Architect — the author of this book — at a large Dutch online auction company watched, with some bemusement, the influence of symbol and belief on leadership, group dynamics, politics, and policy. Here is how it went. Out of the twenty creaking and screeching systems on which the site ran, every kind of data stream converged in the company's Business Intelligence department, which then — with an Excel model that had more tabs than the number of data sources, and formulas full of cross-references that nobody around the table could fully read — sent out a daily "State of the Company" email. Based on FTE, working days, and gross revenue of the same or comparable day in the previous fiscal year, was the target — defined by thirty interlocking formulas — met? Then the thumb pointed up, green, for the day. If not, the red thumb pointed down. Depending on that indicator everyone went to work cheerfully and the CEO came round with a compliment about how things were going; or people avoided each other a little and there was indirect communication, unspoken reproaches, a generally irritable mood.

Belief. Belief in truth, acceptance as the authoritative direction-giver. That is what makes a statement successful in influencing the behaviour of its recipient. It is the *Impact* at the apex of Ackoff's well-known pyramid of Data, Information, Knowledge, and Wisdom — to which modern practice tends to add Impact as the modern apex — a model I will return to often in this book in assessing the impact that data and its processing methods have themselves had on everyday life today, and on the history of humanity.

These are large topics for a data engineer to broach. But this is a data engineer with forty years of experience: from my first job at fifteen, when, with my typing diploma and a knack for computers, I suddenly found myself being paid to program a customer database. From the time before that, when I watched the first digital word processors, library terminals, home computers and then personal computers take over society. And from the forty years after — in which, by way of a degree in Artificial Intelligence, I worked first as a web developer, then as a CTO, and later as a data engineer and data architect at well-known organisations at the forefront of the technological developments in data, but also as interim director of the Netherlands Fortress Museum, and with a lifelong fascination for human history (*according to my children, dad knows everything*) — perhaps I have a sufficiently unusual perspective to have developed a view on this material that is worth listening to.

I should be honest that this is not the first book in the neighbourhood. James Gleick's *The Information* (2011) approached the long history of information from the angle of information theory, with the eye of a journalist. Ann Blair's *Too Much to Know* (2010) is a serious academic history of pre-modern information overload, covering many of the same centuries. Cukier and

Mayer-Schönberger's *Big Data* (2013) is mainly a 2010s data-policy book but it does nod at the historical antecedents. I will not pretend to be in those authors' league as a historian, and I will not pretend to summarise arguments I have not lived inside long enough to do justice to. What I think *this* book adds — and the reason it felt worth writing rather than recommending one of theirs — is the *practitioner's reading*: the perspective of someone who has been building these systems for forty years and is willing to say, plainly, which patterns recur, which are accidents of vendor marketing, and which ones the senior data leader of 2026 is paying for the second time without realising it.

This book is *not* a tools manual, not a vendor comparison, not a tour of architectural fashion, not an exhaustive academic history, and certainly not a consulting brochure. It is what I have just described, and not more. The promise, then, is a modest but real one: by the end of the book, when somebody hands you a "data strategy" document, you will know what to ask of it.

The book is divided in 12 chapters, each covering a certain period, or breakthrough in data procession, in human history. Each chapter contains sections with "Fun Facts" - historical anecdotes worth telling, "War Stories" - relevant comparisons to my own personal life and career and a section where I translated the relevance of historical questions to the practise of the data professional today.

I am writing this Foreword from a sailboat in a small harbour in Normandy. That is not colour: spending sustained time away from the screens has done at least as much of the shaping work on this book as the four decades of practice that precede it. Some patterns become visible only when you stop standing on top of them.

It remains for me to close this Foreword with a few words of gratitude. To my own teachers — from Hanno, the manager of the video store who at fifteen gave me the chance to learn programming with dBase III, to Professor Van Dalen, who encouraged me to immerse myself in the *Intuitionism* of L.E.J. Brouwer and its mystical world at the foundations of mathematics and the philosophy of language. To the colleagues, clients, and students of the last twenty-five years of my working life — and to all that I have learned from them — particularly the systems administrator at the university who taught me Linux, the Ukrainian business partner who in 1998 told me to switch from Perl to Python (*you still use Perl?*), and the trust placed in me by clients such as MediQuest, where I built data platforms for the storage and handling of medically sensitive information on behalf of clinicians.

Most of all to my parents — for kindling my curiosity and my inner drive, and for setting me down on this beautiful world precisely in the era in which the technological revolutions in data have followed one another at high speed. At least in part because of their parenting and their guidance through my study choices, I have ended up in a position from which I have been able to watch

those revolutions from the front row.

And finally to you, for being willing to take in this collection of first-hand experience, historical material, and theoretical reflection on data, language, interpretation, cost, value, and trust. You are the reason I think I had to write this book. I wish you a great deal of pleasure and wisdom in the reading.

If any of this sparked a question about your own platform — that is what a Data Health Call is for.

Book a 30-minute call



430,000 BC
TRINIL, JAVA — THE FIRST MARK



CHAPTER 1

From the Origin of Language to the Birth of Data

Reading time: ~13 min.

Before we begin a 47,000-year journey through the history of data, it is worth pausing on a question we almost never ask: what is data, exactly?

Almost all humans develop a remarkably capable brain during nine months in the mother's womb. The first action many take upon entering the world is an alarm test: a loud cry for food and warmth. Soon thereafter we learn to smile — a biological trait shared with many other mammals. But around our first year, we do something no other species manages in quite the same way: we name things. We utter our first deliberate word — "Mommy" — and somehow, Mama teaches us to associate that sound with her identity. Our first verbal act is not a cry or a smile. It is a label.

To answer the question of what data is, we need to go back. Much further back than computers, further back than writing, further back even than agriculture. We need to find the moment a brain first produced a symbol that meant something to another brain — and ask how long that actually took.

The long silence

About 3.3 million years ago — roughly 700,000 years before the first members of the genus *Homo* appeared — someone picked up a stone and deliberately shaped it. We know this because the site is still there: Lomekwi 3, on the shores of Lake Turkana in Kenya, where a research team stumbled on 150 stone flakes and cores in 2011 after taking a wrong turn. The maker was probably an australopithecine, with a brain barely larger than a chimpanzee's. The act does not imply language. But it implies intention: a mental model of what the stone should become, held in mind before the first strike.

By around 2.6 million years ago, *Homo habilis* was producing the Oldowan tools most of us learned about in school — simple core choppers, consistent across generations, clearly transmitted socially. Then, around 1.7 million years ago, *Homo erectus* produced something genuinely new: the Acheulean hand-axe. Teardrop-shaped, bilaterally symmetrical, requiring forethought about the finished form before the knapping began. A real cognitive step.

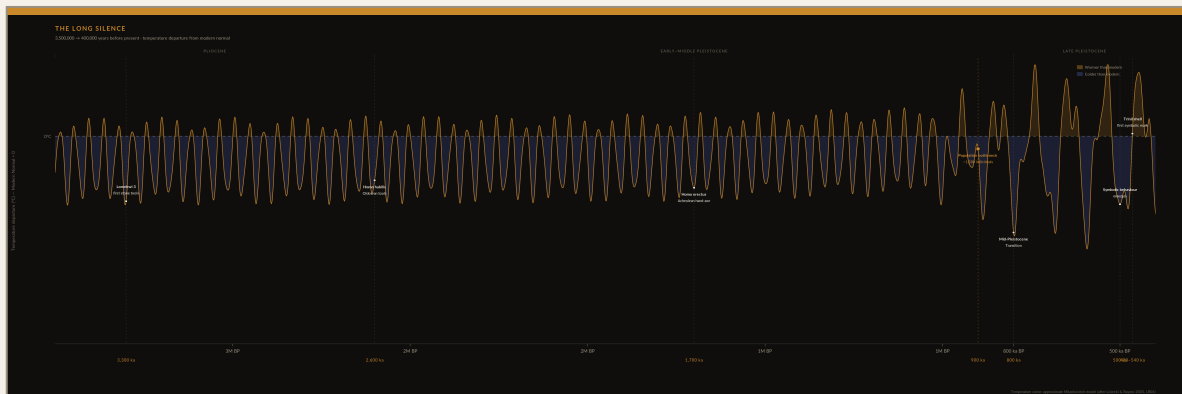
And then almost nothing changed for 1.4 million years.

The same basic hand-axe shape, made by multiple hominin species, on three continents — Africa, Europe, Asia — for over a million years. This is the Acheulean stasis, one of the strangest facts in human prehistory. The current best explanation: without language capable of abstraction, you can copy a tool by watching someone make it, but you cannot improve on it by discussing it. The design was transmitted through imitation, not instruction. Low-fidelity copying, generation after generation, for fourteen hundred thousand years.

Then a finding from 2023 adds a darker dimension to this silence.

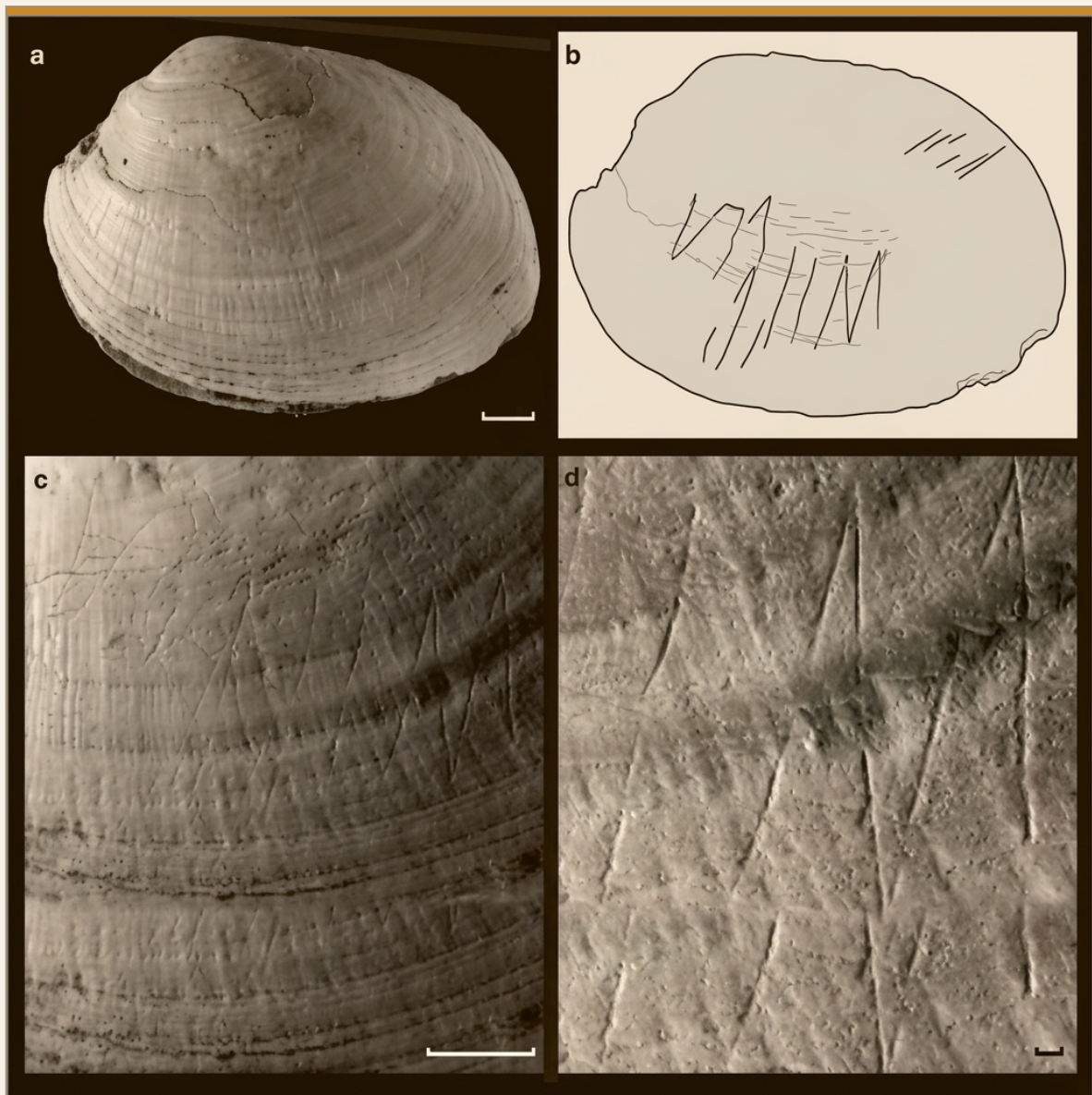
Fun fact — The population of a market town

A 2023 study in *Science*, analysing the genomes of 3,154 modern humans, found that roughly 900,000 years ago our entire ancestral breeding population collapsed to approximately 1,280 individuals — and stayed there for 117,000 years. That is 98.7% of the human population gone. The whole of humanity, for over a hundred thousand years, would have fitted into a mid-sized market town. The timing coincides almost exactly with the deepest trough of the Acheulean stasis and a well-documented gap in the African and Eurasian fossil record. The cause is unknown — a prolonged climate crisis is the leading theory. Information cannot compound without people to carry it.



3,500,000 400,000 years before present. Temperature departure from modern normal as a visual substrate — amber above, indigo below the 0 °C baseline.

What broke the silence? The record becomes gradually richer from around 500,000 years ago. At Trinil in Java — the same site where *Homo erectus* was first identified in 1891 — a museum collection yielded a mussel shell bearing a geometric zigzag engraving. Dating placed it at 430,000 to 540,000 years old. *Homo erectus* made it. This is the oldest known non-utilitarian mark in the human record, predating the previously assumed dawn of symbolic behaviour by 300,000 years.



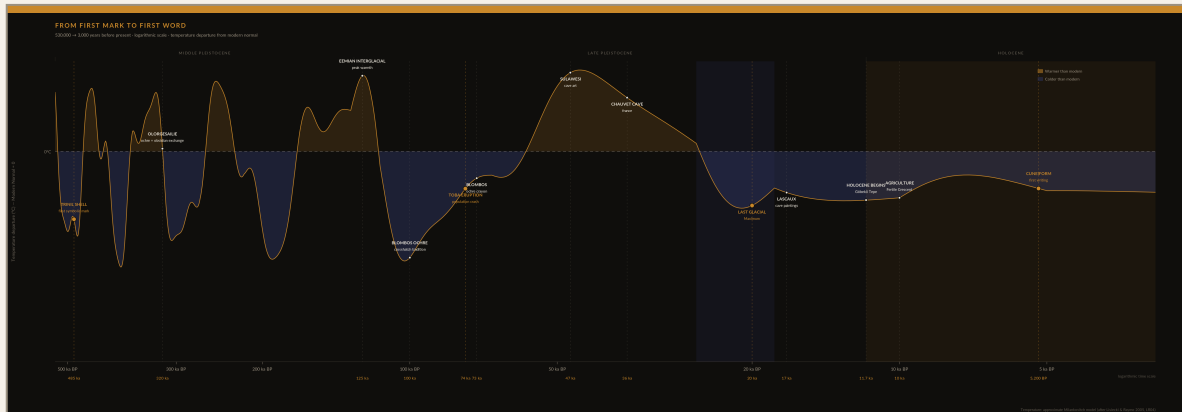
Pseudodon (DUB1006-fL) shell with geometric incisions attributed to *Homo erectus*, showing deliberate zigzag engraving on the exterior surface. c. 430,000–540,000 BP. Photographer unknown. Naturalis Biodiversity Center, Leiden. [VERIFY LICENCE – see Wikimedia Commons file page before publishing].

By around 320,000 years ago, at Olorgesailie in Kenya, something further: shaped pieces of red ochre — pigment, intentionally prepared — found alongside obsidian imported from 60 kilometres away. Ochre with no obvious functional use, plus long-distance exchange, requires shared symbolic value. That is information encoded in an object and carried across space. A data transaction.

Between 142,000 and 73,000 years ago, the record becomes unmistakable. Shell beads from Bizmoune Cave in Morocco — 33 of them, deliberately perforated — are the oldest known personal ornaments. At Blombos Cave in South Africa, ochre blocks engraved with crosshatch

patterns maintained the same geometric tradition for 25,000 years: not one person's invention but a shared, transmitted, cultural convention. By 73,000 years ago, someone was drawing with an ochre crayon on a silcrete flake.

This is the birth of data — not a moment, but a slope. Intentionality, then abstraction, then transmission, then tradition.

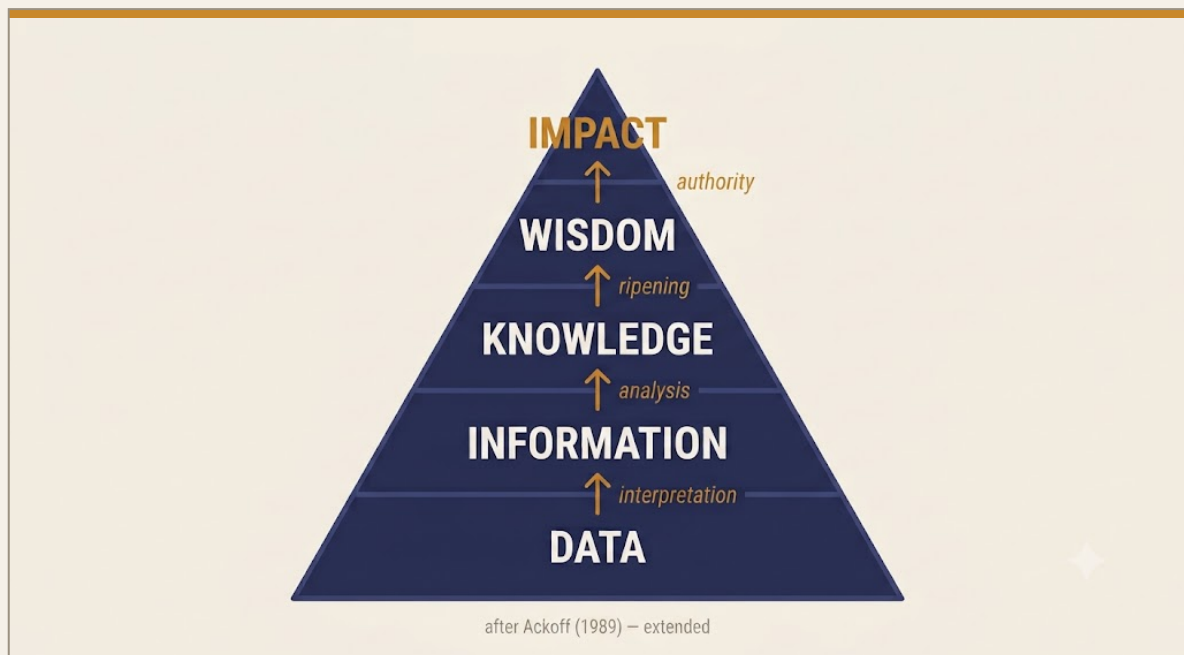


530,000 3,000 years before present. Logarithmic scale — the last 50,000 years is stretched to show the acceleration.
Temperature substrate continues: Eemian warmth, the Last Glacial Maximum, the Holocene opening.

Where the meaning lives

The word *data* comes from the Latin *datum* — "that which is given." Not "that which is true," nor "that which is observed." Data are not facts, and they are not raw observations. They come into existence the moment we accept a set of symbols as a starting point and agree to reason from them.

According to Russell Ackoff's DIKW hierarchy (Data Information Knowledge Wisdom), data are the uninterpreted physical form — the lowest rung. Meaning does not reside in the data; we add it. We "give" meaning to symbols through interpretation (the domain of semantics). Only then do they become information, which we can evaluate as true or false, and — through aggregation and abstraction — turn into justified knowledge.



The DIKW hierarchy extended: each upward step requires a human act — interpretation, analysis, ripening, authority. Data alone produces none of it.

Data in itself is neither true nor false. Truth and belief enter only after interpretation. This process is not automatic; even primitive language requires training and education, feasible only in a relatively narrow window in childhood when the brain is still plastic enough to wire the circuitry for speaking, listening, reading, and writing. How exactly fully symbolic language emerged in human evolution remains genuinely open — social gradualism à la Steven Pinker versus more sudden transitions with universal grammar à la Noam Chomsky, alongside biological changes like a descended larynx and larger brains, plus cultural factors like trust within groups.

The "gift" in data is therefore twofold: the transfer of a physical information bearer from sender to recipient, and the shared acceptance of interpretive rules that allow symbols to produce the same meaning in different minds.

Language is the uniquely human capacity that makes this possible. For over three million years we were, in essence, clever apes with stone tools. What ultimately set us apart was descriptive language — not just alarm calls, but the ability to pass on recipes, methods, stories, and shared value systems across generations and time. The birth of words, nouns, and verbs was also the birth of the first data carriers: oral tradition, information that broke free from a single brain.

The earliest data in human history must have been oral stories (ephemeral, living only in voice and memory) and cave paintings (remarkably persistent: some are still legible after tens of thousands of years).

War story — The youngest of six

I grew up the youngest of six, separated from my nearest sibling by six years. At home, language was competitive sport. My brothers and sisters — children of academics, sparring partners to each other since birth — debated grammar at dinner and settled disputes with the encyclopedia. Films were always aimed at teenagers, at least a year or two above me. I learned to read before school, taught partly by siblings who did not adjust their vocabulary for my age.

In secondary school, I scored 98–100% on interpretive reading tests without studying for them — the kind that make up half the grade in Dutch pre-university foreign language exams. I'm not saying this to brag. I say it in awe: in awe of the human brain, of my brain, and of the culture that had been quietly training it for over a decade before I noticed anything was happening.

We were doing probability theory at the Monopoly board at age nine. Linear algebra and the meaning of π at ten, while my classmates were still practising multiplication tables. Not as schoolwork — as the price of keeping up with people six years my senior who were playing to win.

The Sumerians had a name for what I was accidentally getting at home: the *edubba*, the tablet house, where boys as young as eight were drilled in cuneiform for years before being trusted with a legal inscription. Every literate civilisation seems to have understood, instinctively, that sustained early challenge in language and symbolic reasoning produces a qualitatively different kind of mind. We are noticeably less confident about that today, and I wonder what it costs us.

Before writing, before Sumer, before the first school — when the first human somewhere managed to abstract a symbol that others could use, who recognised what had happened? What did the group do with what they had found?

Every data system in this book was built by a mind that had been trained to see patterns others missed. The question, across forty-seven thousand years, is always the same: who taught them, and how early?

Translation to today

As Ackoff shows, information is interpreted symbolic form. We see similarity in the surrounding landscape and lines on a cave wall, and we interpret it as a map: we give meaning to the form.

How does this work in a modern Medallion architecture, where raw data is captured, filtered, combined, and aggregated into a record in a fact table — registering the simple fact that churn in January was 4.02%?

What does "raw data" actually mean here?

It means that some source business system produced a long set of electrical signals that were stored on a landing zone as magnetic microstructures, later translated back into electrical signals

and rendered as letters in a table on a monitor. It is assumed that the business system created comparable microstructures on its own operational hard drive — for instance, registering each time a customer ends a subscription.

This "raw data" actually reflects an almost incomprehensible, and largely unknown, complex business logic: timed instances of reference tables, GUIDs pointing to missing tables (compensated with data from other systems), and more. Yet in our intermediate layer, we proudly declare we have a "cleansed" dataset with actual cancellations.

Then Johnny from Sales walks in and explains that a batch of these were inaccurate sales registrations cancelled for technical reasons and should not count towards churn. You define the indicator, apply the filter in the fact table.

Now what does it mean when the Marketing colleague says "churn was 4.02% in January"?

How does your organisation transform data into information? How do your users value that information? Is the derivation transparent and verifiable? Do people follow the data, or do they follow trust and authority?

For the purpose of this book, I want to make a clear distinction between form (data) and function (information — interpreted data). Based on modern information theory and language philosophy:

Data *(n.)*

Observations about the world that have been captured in a form that can be stored, transmitted, or processed.

Data is not raw reality itself, but reality made portable: a measurement recorded, a transaction written down, a signal digitised. What counts as data — and who gets to decide — has changed with every era of human history.

The history of data does not begin with computers. It begins with humans trying to make observations transferable across time and space. Today data is everywhere and almost free, but the core problems remain exactly the same as those faced by the first scribes: how do we turn raw symbols into shared meaning, trust, and impact?

A modern data platform solves scalability and performance. But a database without a reporting mechanism, without often implicit metadata required to understand meaning and mentally compile the data into information, is just a very fast bucket of, meaningless, symbols. The reason many "data-driven" organisations feel data-confused is not purely technical. It is the ancient question of who is allowed to declare meaning, and on what authority.

The chapters that follow trace 47,000 years of partial answers to that question — from ca