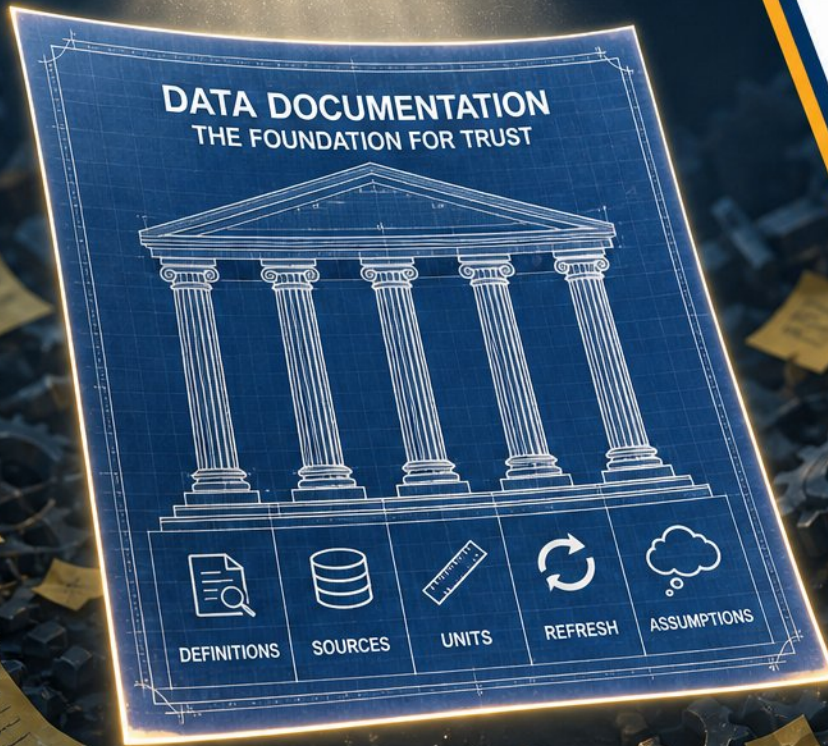


Data Documentation: Definitions, Sources, and Assumptions



- Undocumented data becomes unusable, costing time and money



- 71% of organizational know-how lives only in people's heads



- Knowledge workers lose 2.8+ hours/week searching for info



- Five pillars: definitions, sources, units, refresh, assumptions



- Goal: clear, reusable docs that prevent costly misunderstandings

The Real Cost of Undocumented Data



THE INVISIBLE



CONTRACT

- A renamed column broke a production AI agent, sending 12 overpriced deals in 6 hours — \$180K cost.
- Mars Climate Orbiter: thruster data in pound-force-seconds, software expected newtons — \$327M loss.
- Knowledge workers lose ~2.8 hours/week searching; 71% of know-how lives only in people's heads.
- Poor documentation costs ~\$84,000 per developer per year in lost productivity.
- Documentation is a trust mechanism that protects revenue, reputation, and team sanity.

Who Relies on Your Documentation?



- Four audience roles: analysts reproducing metrics, engineers debugging pipelines, decision-makers acting on dashboards, and your future self



- Documentation is a contract between producers and consumers; without it, every handoff risks silent failure



- A single renamed column (`net_revenue` → `revenue_net`) broke an AI agent, sending 12 overpriced deals worth \$180K in 6 hours



- Knowledge workers lose 2.8 hours/week searching for information; 71% of organizational know-how lives only in people's heads



- Test question: Can a stranger use your data correctly from documentation alone?





Pillar 1 — Writing Field Definitions That Last

FIELD SPECIMEN



BUSINESS MEANING

The real-world concept this field represents.



DATA TYPE

The kind of data and its structure.



VALID VALUES

The allowable values this field can contain.



EXCLUSIONS

What this field does NOT include.



- A usable definition: business meaning, data type, valid values, and explicit exclusions



- Technical column names vs. human-readable definitions — both are essential



- Avoid circular logic, incomplete enumerations, internal slang, and missing context



- Test: Can a new analyst use this field correctly if you leave tomorrow?



Field Definitions — Anatomy and Examples



- **Strong definition:** full business meaning, units, and explicit exclusions



- **Example:** ``customer_lifetime_value`` — total net revenue, USD, first purchase through churn, excludes trials and chargebacks



- **Weak definition:** ``CLV`` — just repeats the acronym, no units, no exclusions



- **Practice:** pick one field and write a definition a stranger could rely on

Strong Definition



Weak Definition



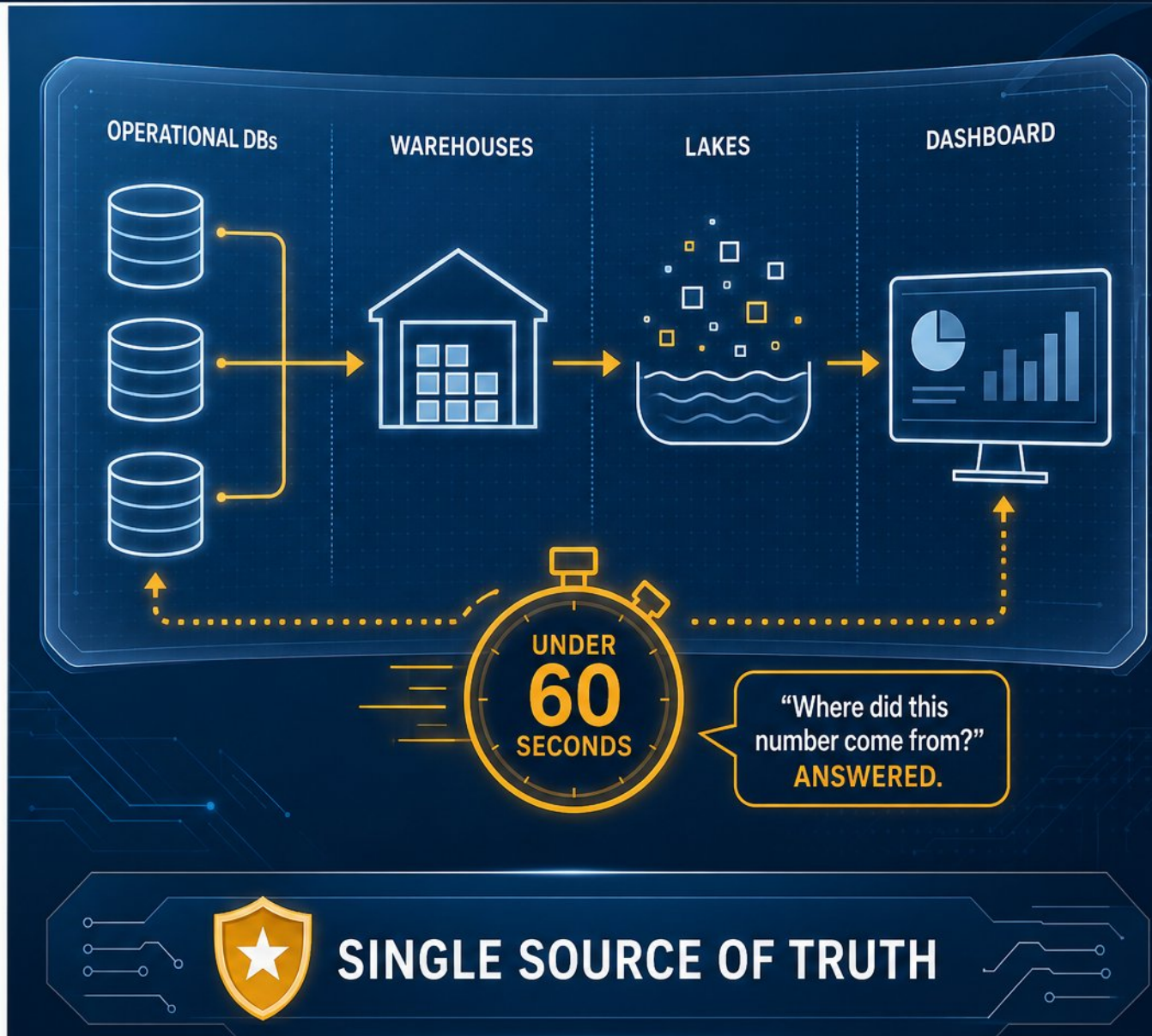
Pillar 2 — Documenting Data Sources

- Capture: system of origin, extraction method, data owner, access path, and refresh schedule.
- Distinguish raw from derived sources; document joins and blending transparently.
- Record reliability: known quality issues, latency expectations, and failure modes.
- Use a minimal source doc template you can adapt today.



Source Lineage — From Operational DB to Dashboard

- Operational DBs, warehouses, and lakes each leave different lineage breadcrumbs — document the full chain.
- “Where did this number come from?” must be answerable in under 60 seconds.
- Modern catalogs automate technical lineage at column-level; human context fills the business gaps.
- Combine automated ingestion with written business context for traceable, trustworthy data.





Pillar 3 — Units and Measurement Context



- State units for all numeric fields: currency, time, rates, normalized values
- Mars Climate Orbiter: a 4.45x trajectory error from lb-s vs N-s silence
- Document precision, rounding rules, and effective decimal places
- When to convert at source vs. store multiple representations



Unit Disasters Are Preventable — Case Studies

THE INVISIBLE CONTRACT

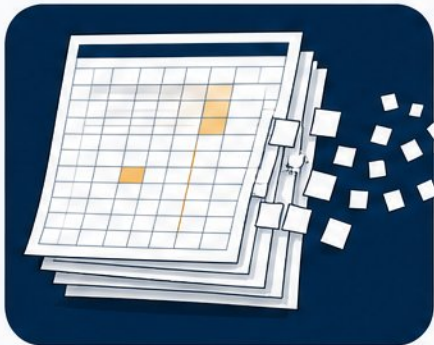
1999

UNWRITTEN ASSUMPTION

2025



- **Mars Climate Orbiter (1999):** thruster data in lb-s, expected N-s; \$327M lost over an unwritten unit contract.



- **Procurement comparison failures:** vendor A IOPS at 4K block, vendor B at 8K — matrix looks comparable, assumptions are incompatible (\$130–390B/yr).



SAME ROOT CAUSE:
AN UNWRITTEN ASSUMPTION



- **The fix is process:** explicit unit documentation, automated cross-checks, and peer review across team boundaries.



EXPLICIT UNIT
DOCUMENTATION



AUTOMATED
CROSS-CHECKS



PEER REVIEW ACROSS
TEAM BOUNDARIES

Pillar 4 — Update Frequency and Freshness



- Define refresh cadence and set escalation contacts when it fails



- Separate extraction timestamps from actual event timestamps



- Set staleness thresholds so consumers know when to stop trusting data



- Modern catalogs auto-surface freshness, but humans must add business context

FIVE PILLARS ARCHITECTURE



Pillar 5 – Assumptions and Known Limitations



- Document all formulas, business rules, and filters applied before data reaches consumers



- List known gaps: missing date ranges, excluded categories, and sampling biases



- Hidden filters break dashboards: a retention model silently excluded zero-order customers



- Unwritten assumptions caused a \$180K pricing error when a renamed column defaulted to list price

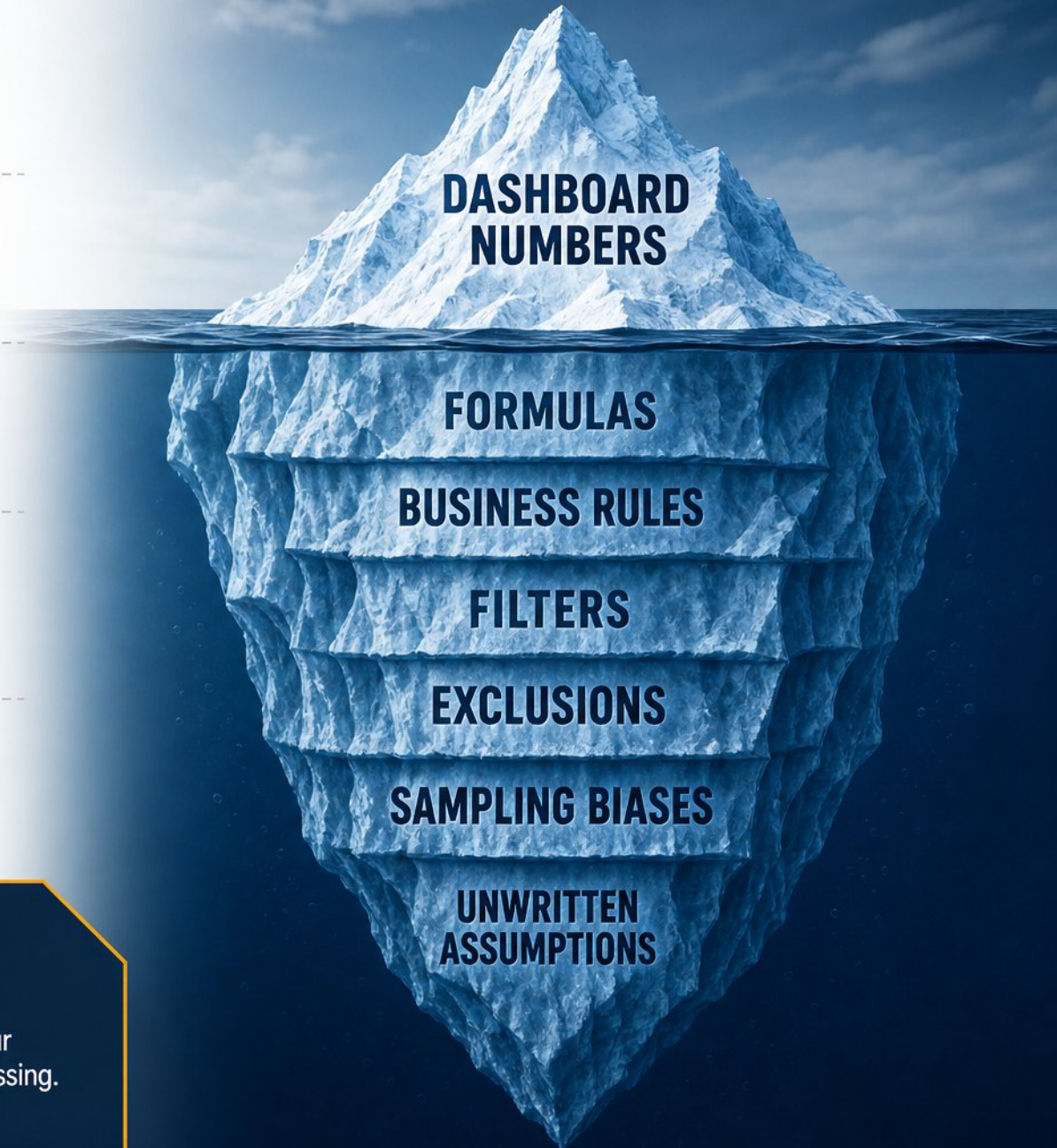


- Version your documentation and maintain a change log when assumptions change



THE FORGOTTEN FUTURE SELF

Document today so your future self isn't left guessing.



Where Documentation Lives



DATABASE SCHEMA

Documentation close to data

```
CREATE TABLE orders (  
  order_id    BIGINT    NOT NULL, -- Unique order identifier  
  customer_id BIGINT    NOT NULL, -- References customers  
  order_date  TIMESTAMP NOT NULL, -- When the order was placed  
  status      VARCHAR(20) NOT NULL, -- Current order status  
  total_amount DECIMAL  NOT NULL, -- Order total in USD  
);  
  
-- Table: orders  
-- Purpose: Stores customer orders  
-- Source: ERP system (orders_service)
```



**SINGLE
SOURCE OF
TRUTH**



DATA CATALOG

Documentation for human search

Search documentation...



order_id
Unique order identifier.



customer_id
References customers.



order_date
When the order was placed.



status
Current order status.



total_amount
Order total in USD.

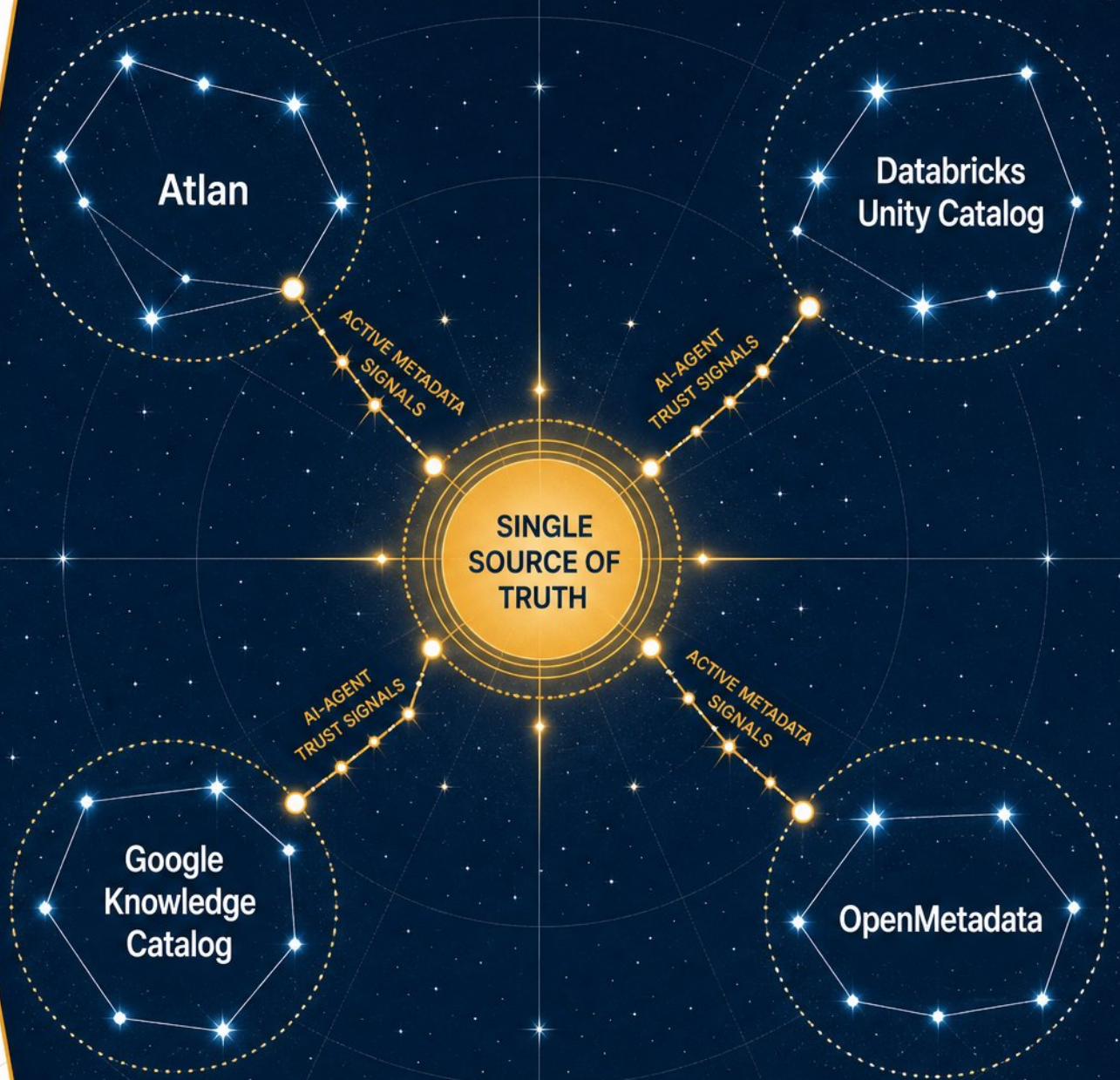
- ▶ - Options: data catalogs, wikis, repos, READMEs, and schema comments
- ▶ - Schema comments stay close to data; catalogs enable human search—use both
- ▶ - Treat docs as code: review in PRs, flag staleness with CI/CD
- ▶ - Start sustainable: field definitions and source lineage first



SINGLE SOURCE OF TRUTH

Tools and Platforms – A 2026 Landscape

- Catalog category converging into a validated context layer
- Continuously vouches for meaning, ownership, quality, and trust state
- Key players: Atlan, Unity Catalog, Knowledge Catalog, OpenMetadata
- Selection criteria: integration footprint, time-to-production, governance maturity, AI agent readiness



Building a Documentation Habit



- Embed documentation steps into code reviews, definition of done, and sprint checklists



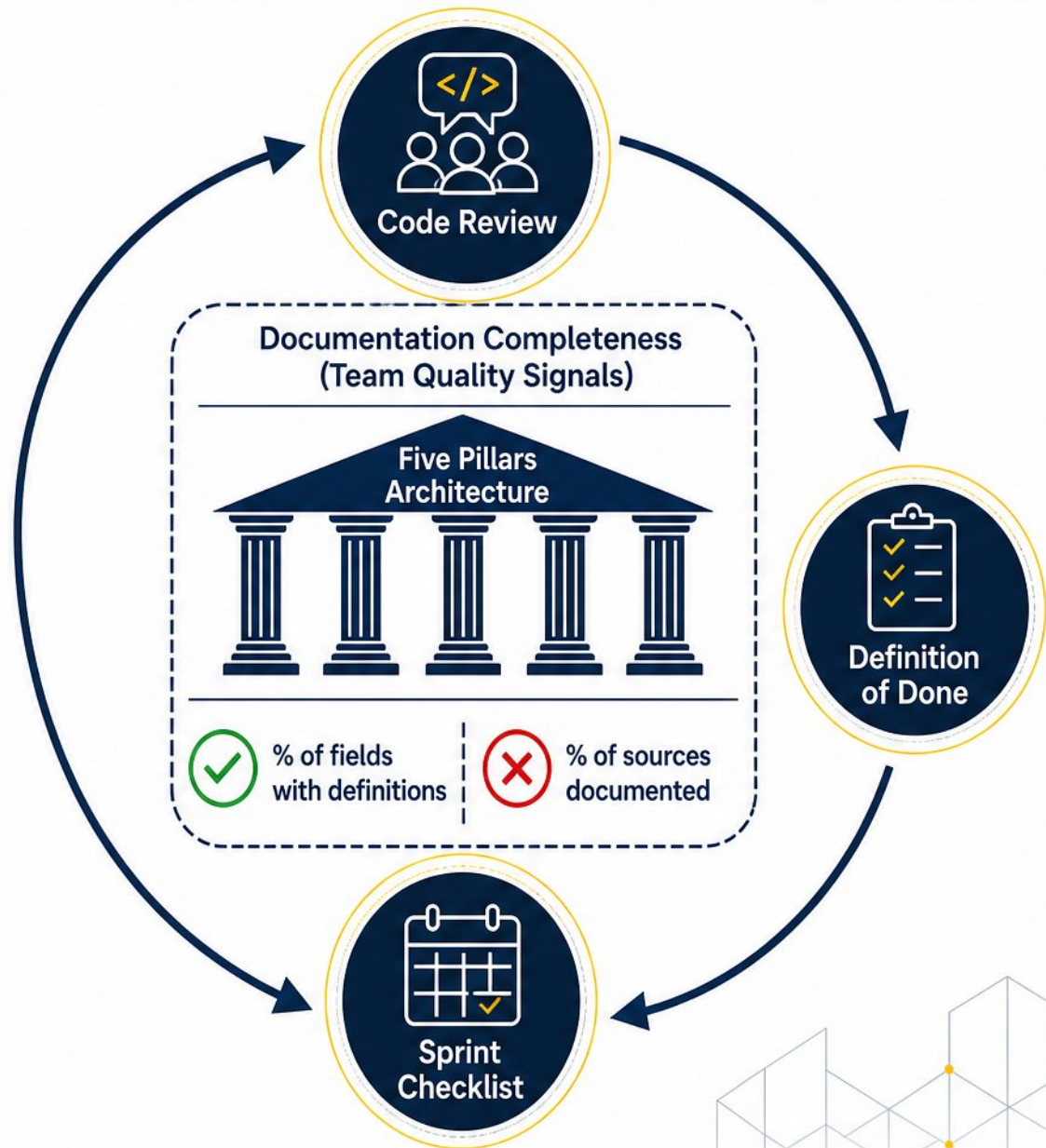
- Start with a one-page template covering all five pillars to reduce friction



- Define review triggers: schema changes, new sources, or updated assumptions



- Track % of fields with definitions and % of sources documented as team quality signals





Immediate Actions — Start With One Dataset



- Pick one dataset: define its top 5 columns and source system.



- Document refresh cadence; list 3 assumptions or known gaps.



- Share with a consumer and ask: "Can you use this correctly without me?"



- Turn their follow-up questions into next week's documentation priorities.



DATASET SPECIMEN CARD

TOP 5 COLUMNS

1. Customer ID	Unique customer identifier
2. Customer Name	Full legal name
3. Account Type	Type of account held
4. Open Date	Date account was opened
5. Status	Current account status

SOURCE SYSTEM

Customer Master System

REFRESH CADENCE

Daily refresh overnight

3 ASSUMPTIONS / KNOWN GAPS

ASSUMPTION:
Names are kept up to date

ASSUMPTION:
Status reflects current state in source

GAP:
No history of status changes available



Summary and Next Steps

THE BRIDGE OF DOCUMENTATION

PRODUCER

Creates and documents with clarity



YOUR FUTURE SELF

Thankful you documented it well



CONSUMER

Understands and makes confident decisions



- Five pillars: definitions, sources, units, refresh cadence, assumptions



- Documentation protects team time, decisions, and your future self



- Use the one-page checklist for each pillar



- This week: document one dataset, share with a consumer

PersonWise.

PersonWise • AI digital-human course platform



Scan the code or click anywhere to watch this course live, with voice Q&A

personwise.ai/t/a818f861/public