Data Quality by Design



Ensuring Reliable Inputs for AI Systems

peter.aiken@anythingawesome.com +1.804.382.5957

© Copyright 2026 Peter Aiken, PhD

RELTIO
An SAP company

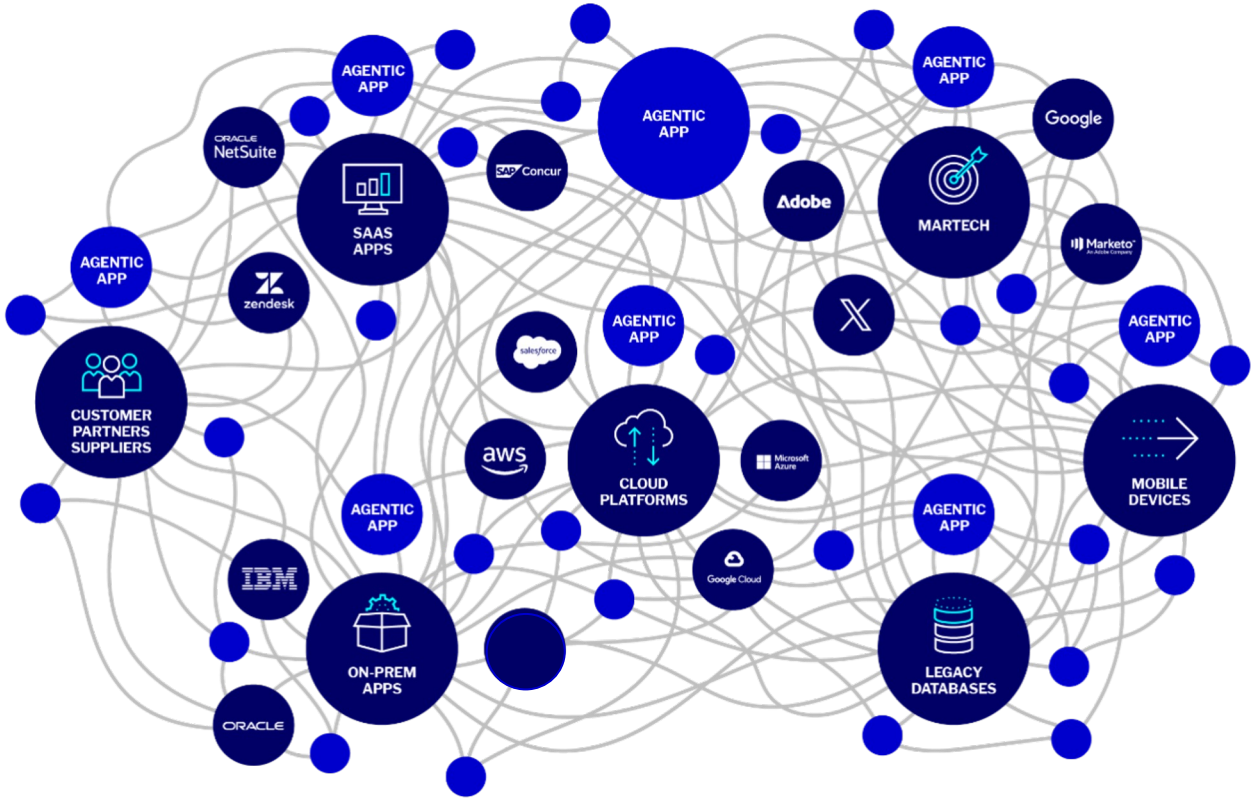
Data Quality by Design with Reltio Context Intelligence Platform

Phil D'Amico
Senior Solutions Consultant, Reltio



©2026 Reltio. All Rights Reserved.

Siloed data creates a quality and context gap for AI



Data silos fragment context

AI agents lack a true 360 understanding

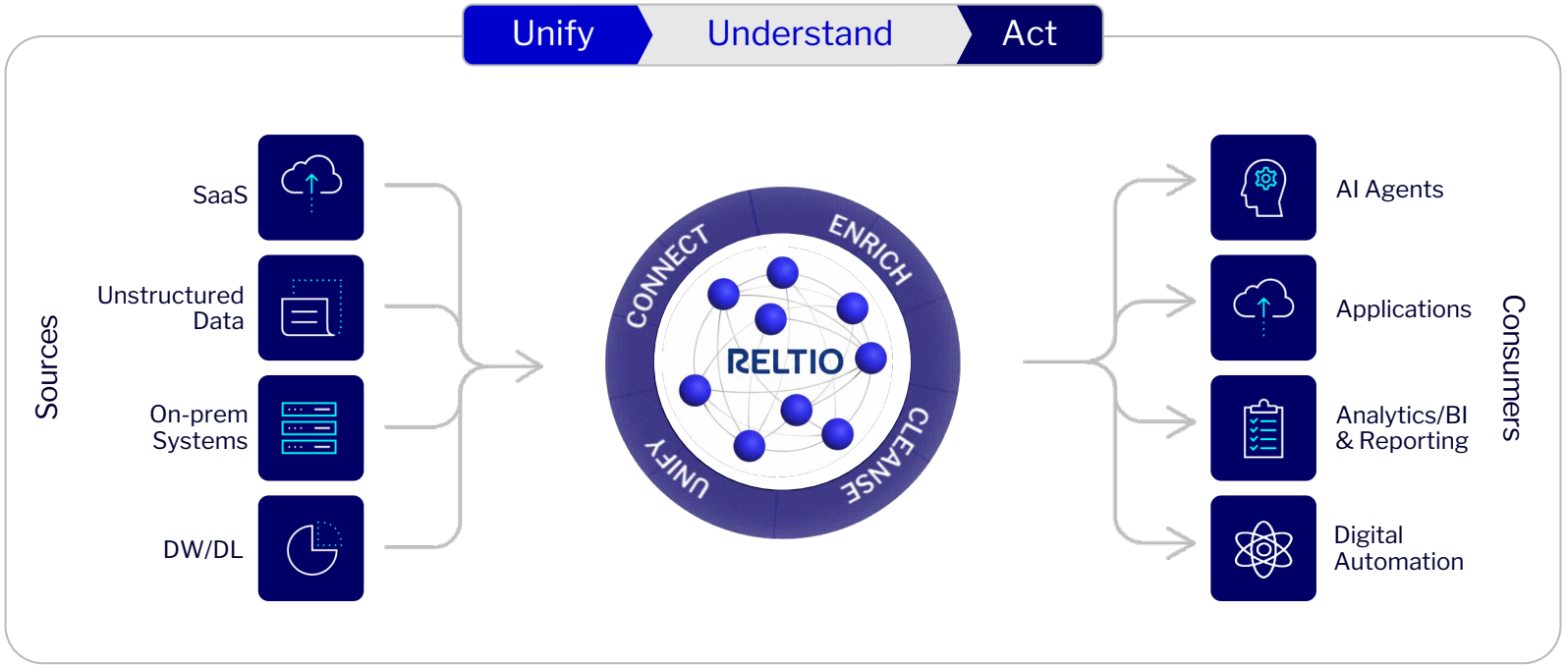
AI initiatives break down without a human in the loop

RELTIO

©2026 Reltio. All Rights Reserved.

The content provided is for informational purposes only. ©2026 Reltio. "Reltio" is a registered trademark of Reltio, Inc. All Rights Reserved.

Reltio powers context intelligence



RELTIO
©2026 Reltio. All Rights Reserved.

The content provided is for informational purposes only. ©2026 Reltio. "Reltio" is a registered trademark of Reltio, Inc. All Rights Reserved.

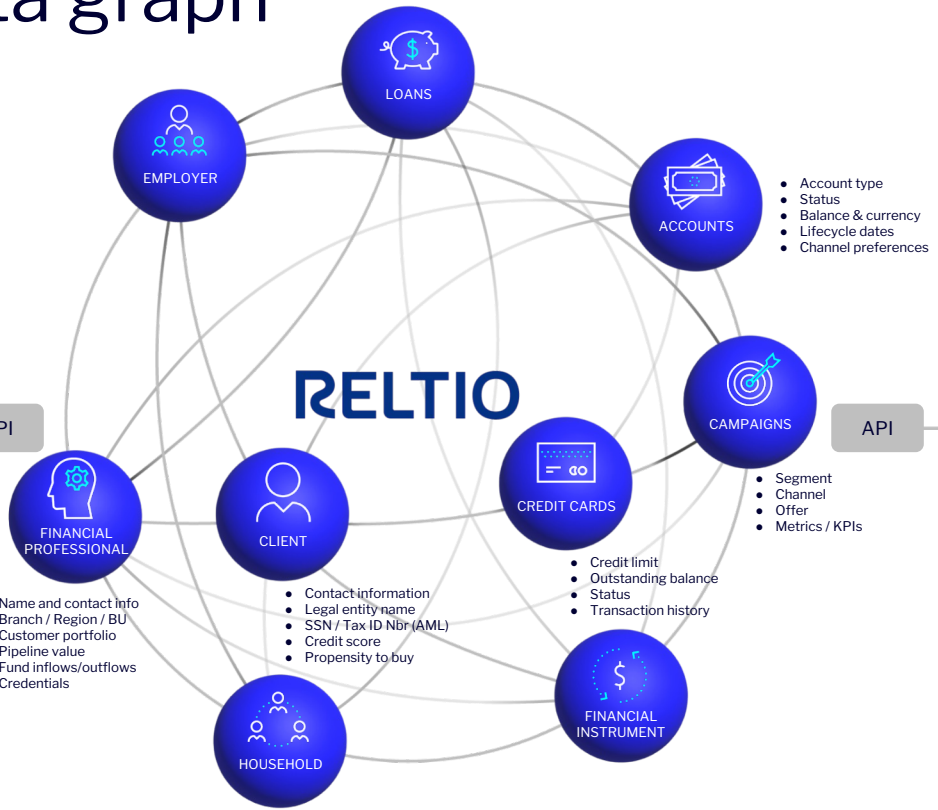
Your system of context runs on our intelligent data graph

Financial Services

Sources

- APPLICATIONS
- DATA LAKES/DW
- CATALOGS
- DATA ENRICHMENT
- UNSTRUCTURED SOURCES

API



- Account type
- Status
- Balance & currency
- Lifecycle dates
- Channel preferences

- Segment
- Channel
- Offer
- Metrics / KPIs

- Name and contact info
- Branch / Region / BU
- Customer portfolio
- Pipeline value
- Fund inflows/outflows
- Credentials

- Contact information
- Legal entity name
- SSN / Tax ID Nbr (AML)
- Credit score
- Propensity to buy

- Credit limit
- Outstanding balance
- Status
- Transaction history

API

Consumers

AUTONOMOUS AGENTS

Data and business ops. agents
Resolver
Address Enricher
Work Assigner
Product Recommender

- DATA LAKES
- APPLICATIONS

RELTIO

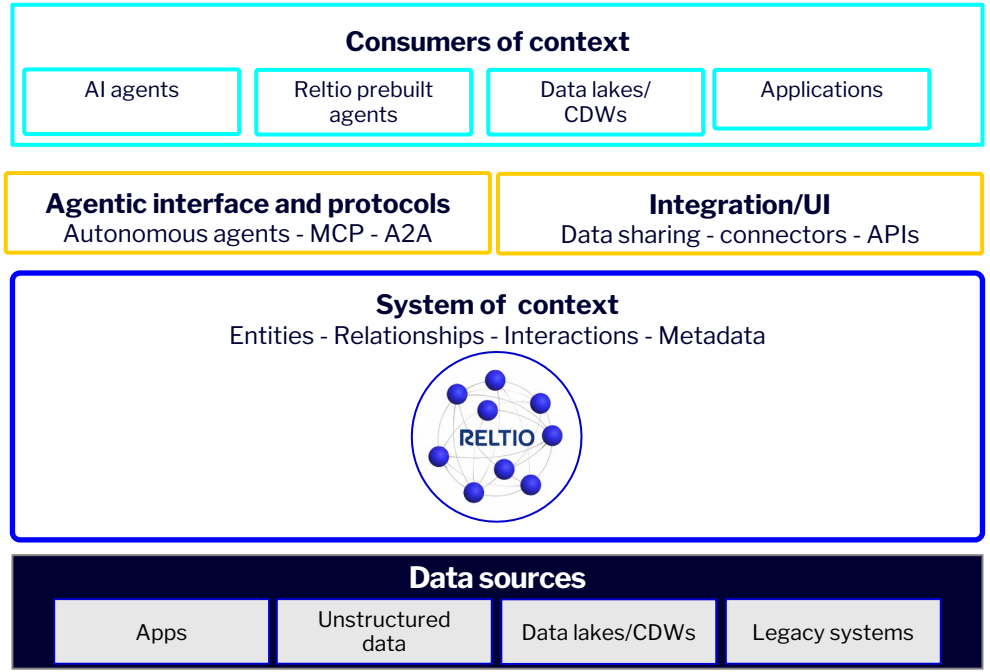
©2026 Reltio. All Rights Reserved.

The content provided is for informational purposes only. ©2026 Reltio. "Reltio" is a registered trademark of Reltio, Inc. All Rights Reserved.

System of context is the bridge between enterprise AI and data

Without system of context

Fragmented and siloed
Agentic AI as a science project



With system of context

Clean and connected
Agentic AI as business transformation

RELTIO

©2026 Reltio. All Rights Reserved.

The content provided is for informational purposes only. ©2026 Reltio. "Reltio" is a registered trademark of Reltio, Inc. All Rights Reserved.


Thank you
Reltio.com

RELTIO

CDATA

Data Quality by Design: Ensuring Reliable Inputs for AI Systems

November 2026

© 2026 CData Software. All rights reserved.

CData Software

Based in Chapel Hill, NC, with offices in Europe, Japan, and India, CData is the leader in real-time data connectivity software for enterprises, data engineers, developers, and data scientists.

10,000+

Active customers

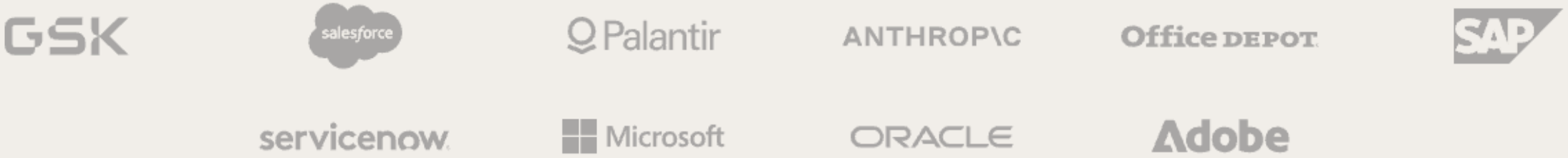
140+

Embedded partners

400+

Employees

Over 10,000 customers leverage CData to simplify data connectivity, eliminate data silos, and break down barriers to better integration and insights.



© 2026 CData Software. All rights reserved.

Data and AI leaders agree - data quality and connectivity is holding AI back

71%

of AI teams spend more than a quarter of implementation time on **data connectivity** alone

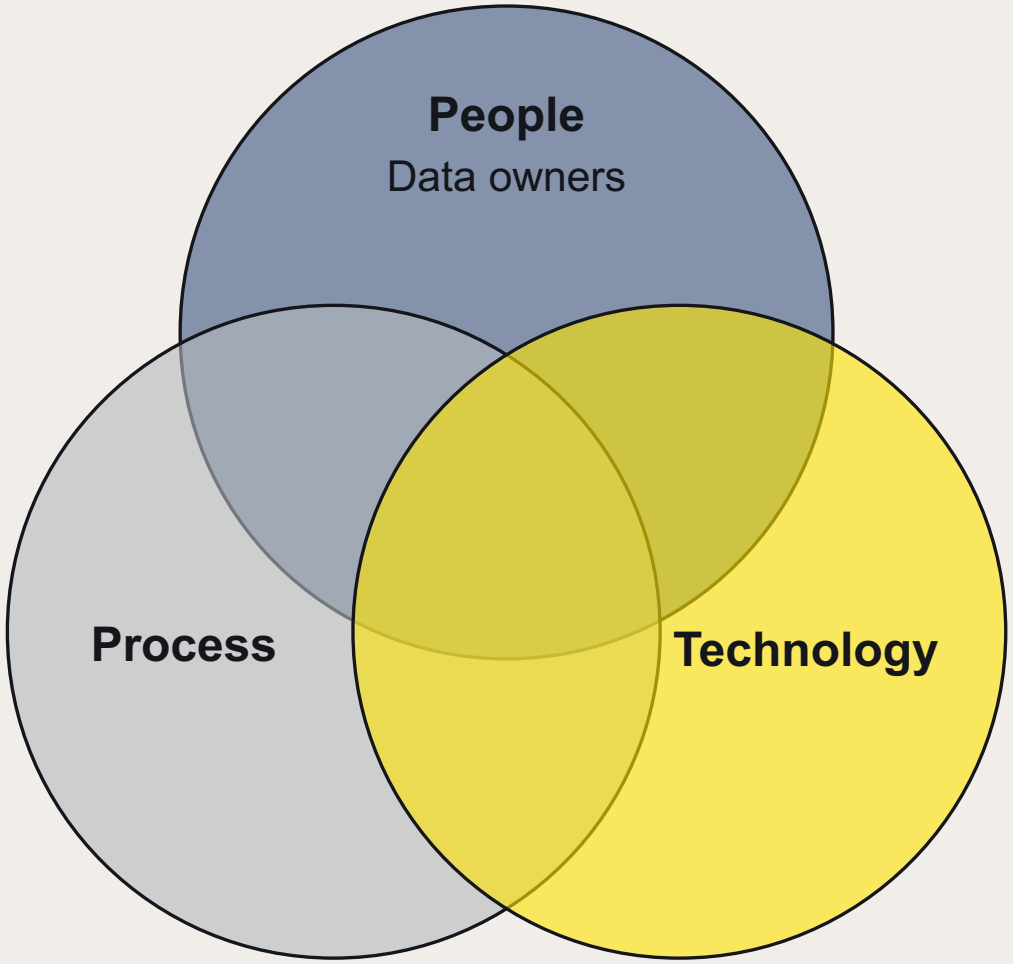
73%

of organizations cite **data quality and integration** as a top barrier to AI adoption

Source: CData State of AI Data Connectivity Report: 2026 Outlook—survey of 200+ enterprise data and AI leaders

© 2026 CData Software. All rights reserved.

Data Quality takes people, process, and technology



People

Own the standards

Process

Runs the process/continuous improvement

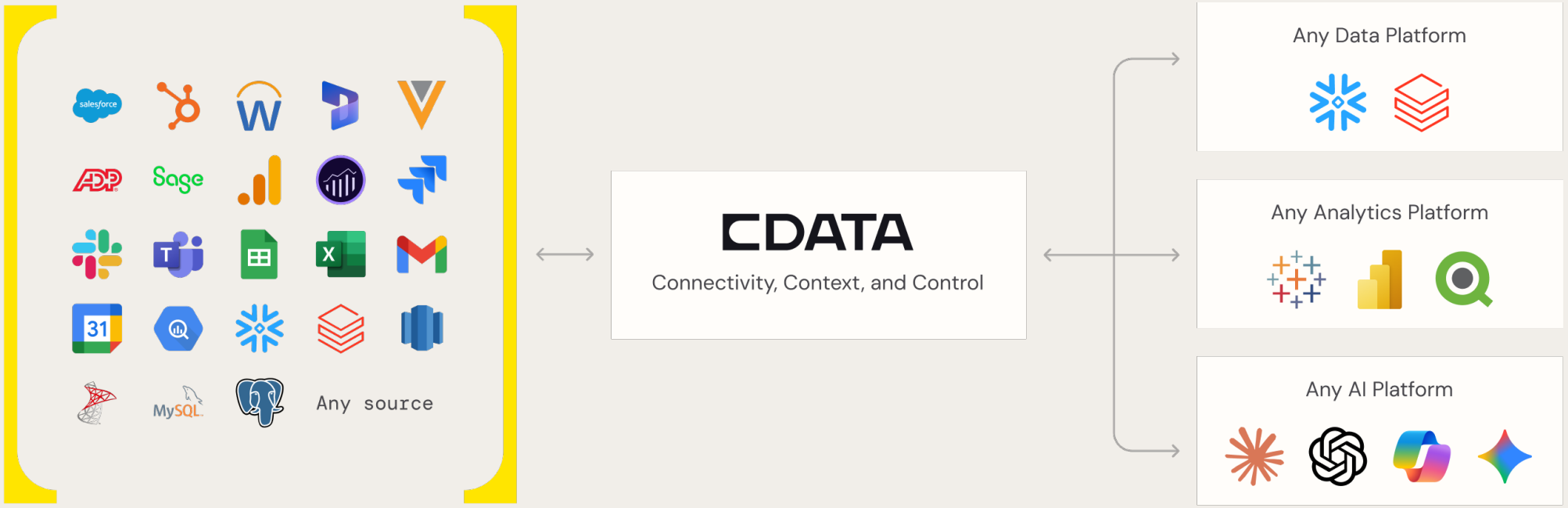
Technology

Applies the standards, transforms and automates, governs

CDATA

© 2026 CData Software. All rights reserved.

One data layer that connects, understands, and governs



Connectivity • Context • Control

CDATA

© 2026 CData Software. All rights reserved.

Learn More About CData

Learn more about CData

www.cdata.com

Contact:

sales@cdata.com

© 2026 CData Software. All rights reserved.

Peter Aiken, Ph.D.

- I've been doing this a long time
- My work is recognized as useful
- Associate Professor of IS (vcu.edu)
- DAMA International (dama.org)
  - IDEAL Affinity Group
  - GenAI in Data Affinity Group
  - Data Educators Affinity Group (forming)
  - Data Project Managers Affinity Group (forming)
- Anything Awesome (anythingawesome.com)
- Experienced w/ 1,000+ data management practices worldwide
- 13 books and dozens of articles
- Multi-year immersions
  - US DoD (DISA/Army/Marines/DLA)
  - Nokia
  - Deutsche Bank
  - Wells Fargo
  - Walmart
  - HUD
  - NGA …



$1,500,000,000.00 USD



Program Overview

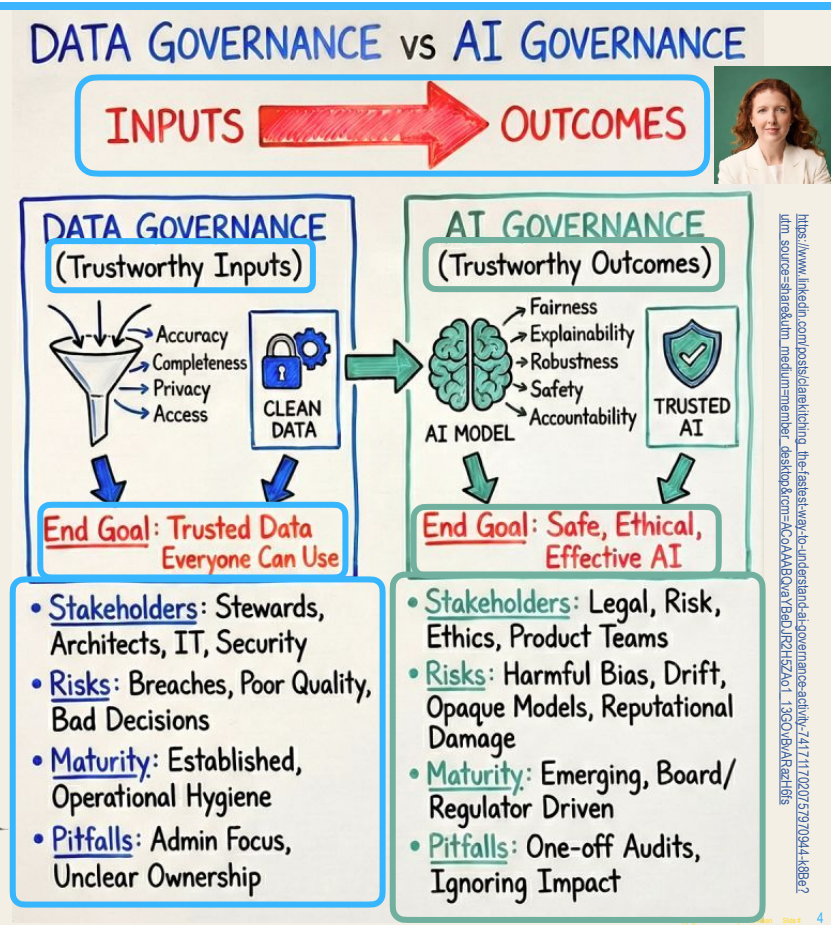
Data Quality by Design

Ensuring Reliable Inputs for AI Systems

- Definitions
  - Data governance and AI governance
  - Defensive driving is a mindset
  - Trying to clean too much data is unproductive
  - Data quality definitions
  - Practice vs. structural challenges
- Data quality challenges manifest in many ways
  - Hidden data factories
  - Obscure data quality challenges
  - Obvious data quality challenges
- Approaching Data Quality
  - Causes can be difficult to discern
  - The role of data debt
  - Solutions must be built on leverage
- Approach Basics
  - Programmatic approach is most effective
  - Early projects have a dual purposes
  - High quality data requires architecture and engineering
  - Systems thinking
  - Misunderstanding the role of culture
  - Developing repeatable capabilities in core data quality expertise
- How do we get better?
  - Refocus the discussion around business outcomes
  - Leadership
  - Math (cost or investment?)
  - Program focus
  - Storytelling/Practice
- Takeaways and Q&A

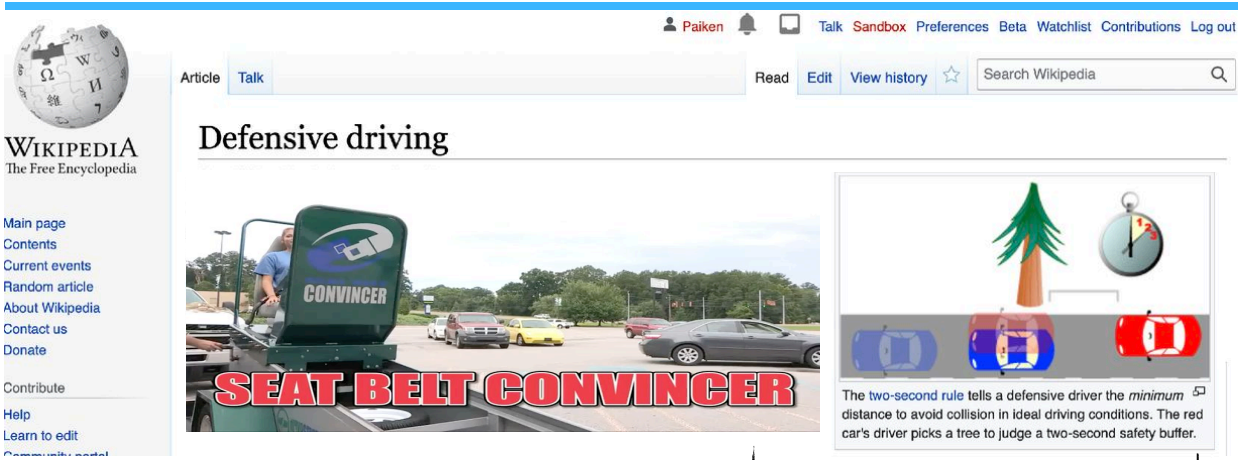


- This diagram has produced lots of value by getting everyone on the same sheet of music
- Clearly data is of concern to both data governance & AI governance
- Anything done to improve data quality as part of either of these activities, helps other parts of the organization as well



https://www.linkedin.com/posts/clarekitching_the-fastest-way-to-understand-ai-governance-activity-7417117020757970944-k8Be?utm_source=share&utm_medium=member_desktop&rcm=ACoAAABQvaYBeDJR2H5ZAo1_13GOvBvARazH6fs

https://www.linkedin.com/in/clarekitching



General principles:
- Controlling your speed.
- Looking ahead and being prepared for unexpected events.
- Being alert and distraction free.

Regarding other participants in traffic:
- Preparedness for all sorts of actions and reactions of other drivers and pedestrians.
- Not expecting the other drivers to do what you would ordinarily do.
- Watching and respecting other drivers.

Regarding your own vehicle:
- Maintaining a safe following distance.
- Driving safely considering (adjusting for) weather and/or road conditions.
- Adjusting your speed before entering a bend, in order to avoid applying the brakes in the middle of a bend.

- When taught as part of basic driving, behavior becomes part of a lifelong practice
- Difficult to practice when learned after driving has been taught
- The Convincer creates a change in mindset

https://anythingawesome.com
© Copyright 2026 by Peter Aiken Slide #

Data in the Cloud Should Have Three Attributes That Data Outside the Cloud/Warehouse Should Not Have. It Should Be:

Cleaner





Smaller

Sharable-er

SHAREABLE

Making —Cloud— Successful



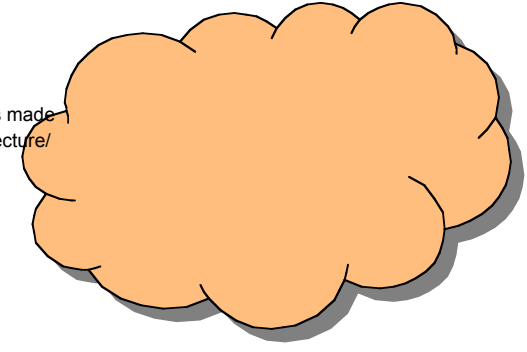


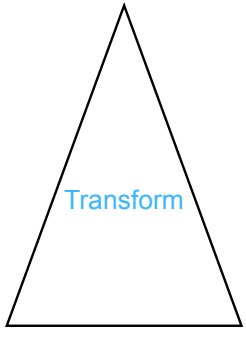




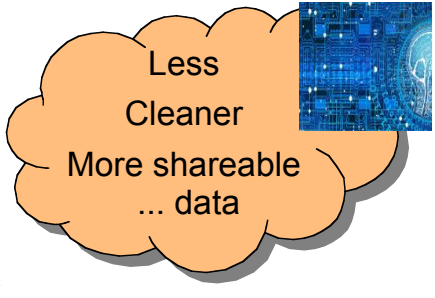
Problems with forklifting

1. no basis for decisions made
2. no inclusion of architecture/ engineering concepts
3. no idea that these concepts are missing from the process
4. 80% of organizational data is ROT





Data Branding



Data Branding



AI Ready?



https://mindsparklemag.com/design/arcadia-data/

© Copyright 2026 by Peter Aiken

Augment

© Copyright 2026 by Peter Aiken



https://anythingawesome.com

The Professional Version of AI is a great search tool

- What has been said about and by others who have been involved in similar situations?
- Where are the best sources to learn more about data quality in my specific industry?



Direct Information Synthesis: Converts fragmented web pages and blue links into single, unified explanations, extracting key findings directly without requiring manual site-by-site browsing.
Context and Intent Recognition: Employs natural language processing to decode underlying search intent, technical terminology, and contextual relationships rather than relying on exact keyword matches.
Multi-Step Reasoning: Deconstructs complex, multi-layered queries into coordinated sub-questions, retrieving disparate information across sources to cross-examine and answer edge-case problems.
Inline Attribution and Traceability: Embeds source citations, metadata, and direct verification links next to specific claims to support factual auditing.

Iterative Conversational Refinement: Retains dialogue context across multi-turn exchanges, allowing users to drill down, reframe scope, or change formatting without restarting the search.
Enterprise Document and Data Grounding: Integrates external web indexing with private internal documents, data stores, and extended context windows to ground answers in proprietary research.
Multimodal Search Capabilities: Interprets diagrams, charts, user-uploaded data tables, and images alongside standard text queries to identify patterns and troubleshoot technical schemas.
High Signal-to-Noise Ratio: Bypasses commercial SEO-optimized filler, click-heavy landing pages, and web-page clutter by querying, extracting, and summarizing raw textual data directly.

© Copyright 2026 by Peter Aiken

Everyone Wants To Do Better Data Analysis …

- Some data preparation is inevitable
- What would a 'good' ratio be?
- "Everyone knows"



© Copyright 2026 by Peter Aiken

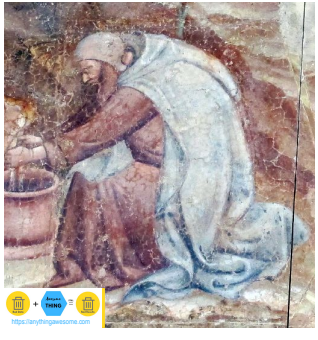
Munge

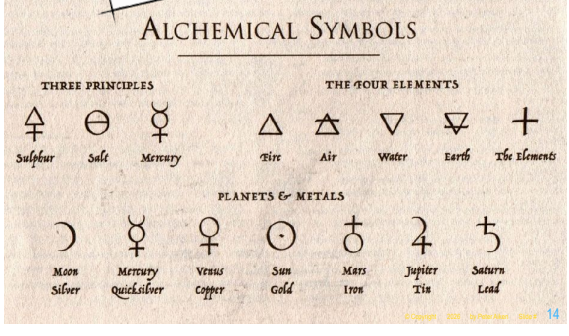
© Copyright 2026 by Peter Aiken

Munge Https://En.Wikipedia.Org/Wiki/Mung_(Computer_Term)

- Computer jargon
- For a series of potentially destructive or irrevocable
- Changes to a piece of data or a file
- Vague data transformation steps that are not yet clearly defined
- Common munging operations include:
  - Removing punctuation
  - Removing html tags
  - Data parsing
  - Filtering
  - Transformation

Data Cleansing!

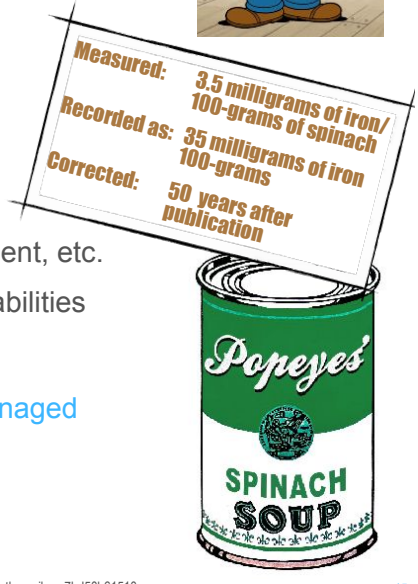




Definitions

- Quality Data
  - Fit for purpose meets the requirements of its authors, users, and administrators (from Martin Eppler)
  - Synonymous with information quality, since poor data quality results in inaccurate information and poor performance
- Data Quality Management
  - "Planning, implementation and control activities that apply quality management techniques to measure, assess, improve, and ensure data quality"
  - Encompasses life cycle activities
  - Include supporting processes from change management, etc.
  - Continuous improvement process requiring core capabilities
- Data Quality Engineering
  - Recognition that data quality solutions cannot not managed but, instead, must be engineered
  - Data quality engineering concepts are generally not known and understood within IT or business!





Spinach/Popeye story from: https://medium.com/hurb-labs/data-quality-a-lesson-from-the-myth-behind-popeye-the-sailor-a7bd50b61510



Data Quality Attributes

Tend to be
- Practice-related data quality challenges
- Numerous
- Closer to the user

Tend to be
- Structure-related data quality challenges
- Fewer
- Remote to uses/management

Data Representation Quality	Data Value Quality	Data Model Quality	Data Architecture Quality
as presented to the user	as maintained in the system	as understood by developers	as an organizational asset
Completeness	Completeness	Completeness	Completeness
Correctness	Correctness	Correctness	Correctness
Timeliness	Currency	Conceptual Correctness	Enterprise Model Utility
Conciseness	Frequency	Conceptual Completeness	Data Management Quality
Clarity	Time Period	Syntactic Correctness	Data Sharing Ability
Detail	Precision	Syntactic Completeness	Data Engineering Quality
Order	Reliability		Data Operation Quality
Presentation	Relevance		Data Evolvability
Media	Scope		Data Self Awareness
Unambiguous	Granularity		

Quality Dimensions

Data Value Quality

Data Representation Quality

practice-related

- Failure in rigor when capturing/ manipulating data
- Allowing incorrect data to be collected when requirements specify otherwise
- Presenting data out of sequence

Data Quality

Data Quali



Data Quali

"Fit for purpose"

Data Model Quality

structure-related

- Data arranged imperfectly
- Street address → GPS coordinates
- Data is captured but inaccessible
- When incorrect data is provided as the correct response

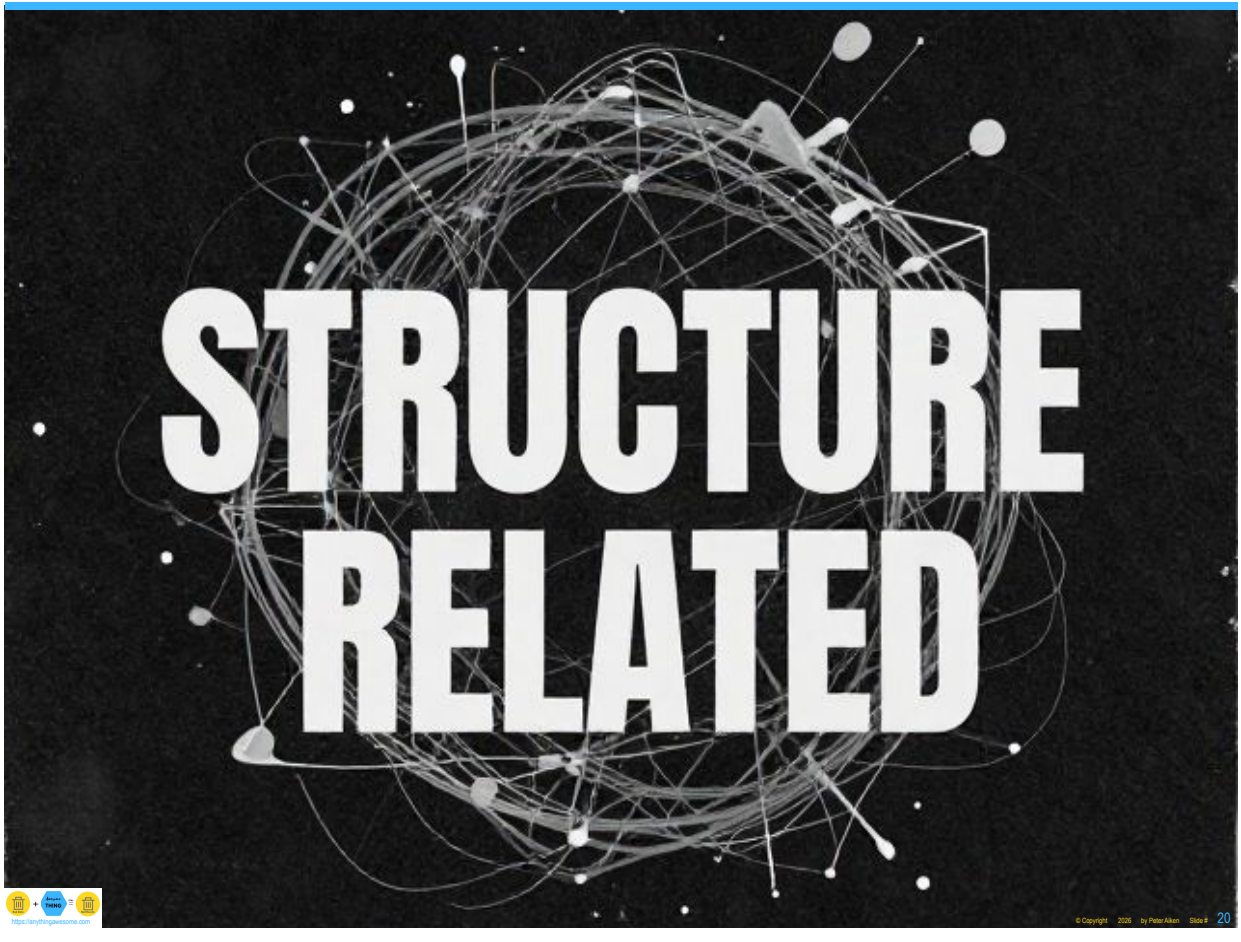
Data Architecture Quality

Practice Related





© Copyright 2026 by Peter Aiken



https://anythingawesome.com

NYC Tree Maintenance Analysis

- NYC
  - 2,500,000 trees
  - Annual deaths by falling tree limbs
  - 4 people were killed in Central Park 2009-10
- Belief
  - Arborists believe that pruning and otherwise maintaining trees can keep them healthier and make them more likely to withstand a storm, decreasing the likelihood of property damage, injuries and deaths
- Until recently
  - No research or data to back it up
- Question
  - Does pruning trees in one year reduce the number of hazardous tree conditions in the following year?





- Lots of data but challenges
  - Pruning data recorded block by block
  - Cleanup data recorded at the street address level
  - Trees have no unique identifiers
- After downloading, cleaning, merging, analyzing and intensive modeling
  - Pruning trees for certain types of hazards caused a 22 percent reduction in the number of times the department had to send a crew for emergency cleanups
- The best data analysis
  - Generates further questions
- NYC cannot prune each block every year
  - Building block risk profiles: number of trees, types of trees, whether the block is in a flood zone or storm zone

© Copyright 2026 by Peter Aiken

Program Overview

Data Quality by Design
Ensuring Reliable Inputs for AI Systems

- Definitions
  - Data governance and AI governance
  - Defensive driving is a mindset
  - Trying to clean too much data is unproductive
  - Data quality definitions
  - Practice vs. structural challenges
- Data quality challenges manifest in many ways
  - Hidden data factories
  - Obscure data quality challenges
  - Obvious data quality challenges
- Approaching Data Quality
  - Causes can be difficult to discer
  - The role of data debt
  - Solutions must be built on lever
- Approach Basics
  - Programmatic approach is most effective
  - Early projects have a dual purposes
  - High quality data requires architecture and engineering
  - Systems thinking
  - Misunderstanding the role of culture
  - Developing repeatable capabilities in core data quality expertise
- How do we get better?
  - Refocus the discussion around business outcomes
  - Leadership
  - Math (cost or investment?)
  - Program focus
  - Storytelling/Practice
- Takeaways and Q&A



Hidden Data Factories

Department A

Work products are delivered to

Department B
1. Check A's work
2. Make any corrections
3. Complete B's work
4. Deliver to Department C



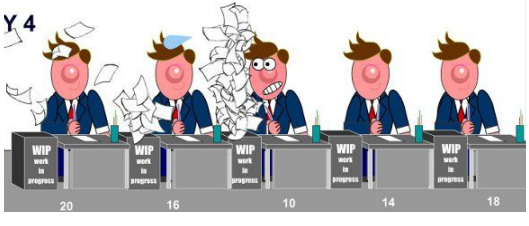
Department C
1. Check B's work
2. Make any corrections
3. Complete C's work
4. Deliver to Customer
5. Deal with consequences





Customers

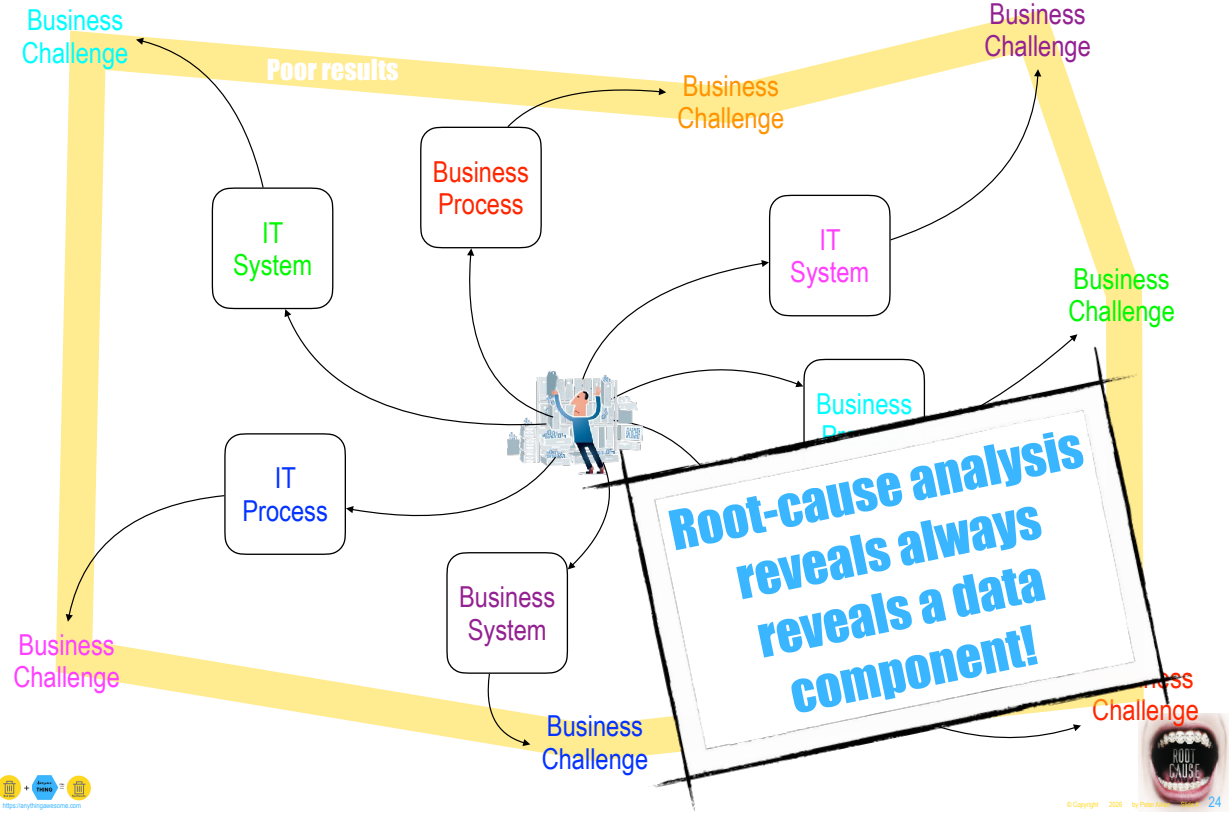
Any weak link in the value chain causes everything to slow



https://en.wikipedia.org/wiki/Theory_of_constraints

https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year

Poor Data Quality Manifests as Multifaceted Organizational Challenges



HOW TO FIND HIDDEN DATA FACTORIES

Uncovering shadow pipelines, local transformation layers, and rogue spreadsheets

Every hidden factory is proof of an **unmet business requirement** and a blueprint for what to automate.



1. IDENTIFY "DATA FRICTION POINTS"

- Search for **manual pre-meeting preparation,** especially recurring calendar blocks like "Monthly Prep."
- **Audit reconciliation scripts** bridging gaps between disparate systems.
- Look for **persistent departmental translation layers** and **"mapping_FINAL" files.**



3. TRACK API KEYS & EXTRACTION PATTERNS

- **Audit read-heavy service accounts** with repetitive, scheduled extract behavior.
- **Identify mass "Export to CSV/Excel" behavior,** especially regular high-volume reports.
- **Scan database query logs** for large-scale data extractions originating from user IPs.



5. FOLLOW "SINGLE POINT OF FAILURE" PEOPLE

- Look for the **"shadow data engineers"**—unofficial gurus who "know how to get the actual numbers."
- **Check onboarding checklists** to see if new hires are instructed to run undocumented desktop macros.
- Identify teams that **panic** when a specific analyst is on leave because "reports won't run."



2. INSPECT SAAS & CITIZEN-INTEGRATOR TOOLS

- **Audit low-code/no-code workflows (Power Automate, Zapier, Make)** transferring data out of core systems.
- **Check BI workspace gateways** using personal connections or local file sources (SharePoint, Google Sheets).
- **Scan cloud expense line items** for unmanaged database or compute instances.



4. SCAN SHARED STORAGE & LOCAL AUTOMATION

- Search for **script extensions** like **.py, .ipynb, .r, .xlsm** (macro-enabled), **.pbix** (BI files) in shared directories.
- Check **local task schedulers and cron jobs** running on department virtual machines.
- **Map shared and personal drives** for local storage arrays of raw datasets.

QUICK DETECTION CHEAT SHEET

	Check	Tell
Identity & Access	API & Export Logs	Repetitive scheduled user extracts
Compute & Network	Egress & DB Logs	Heavy SQL reads from user/non-server IPs
Workflow Automation	Low-code Tenants	Flows outputting to rogue sheets/local drives
File Artifacts	Corporate Drive Crawl	Unmanaged scripts and macros (.py, .xlsm, .pbix)

Obscure Data Challenge



doslie thalonge

https://anythingawesome.com

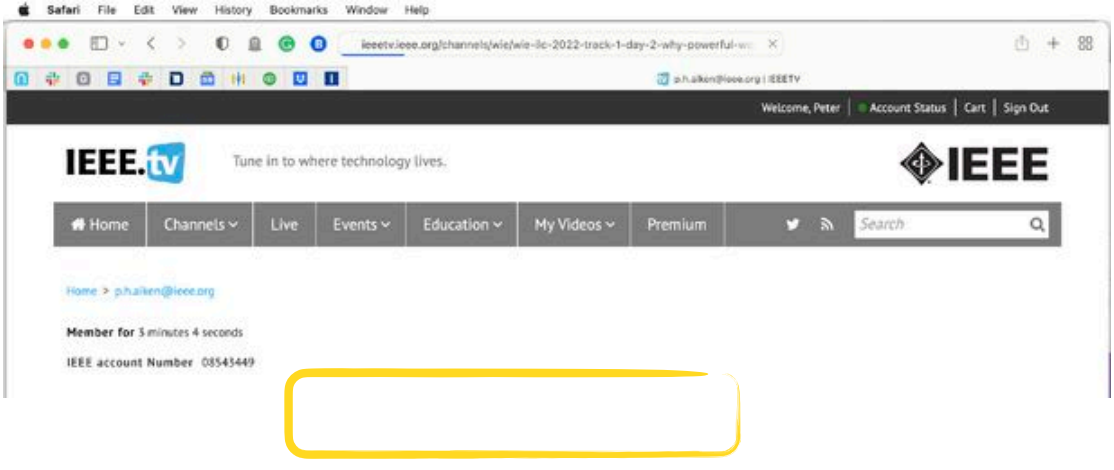


A congratulations letter from another bank

Problems

- Bank did not know it made an error
- Tools alone could not have prevented this error
- Lost confidence in the ability of the bank to manage customer funds

IEEE Senior Member Status (30+ Years Membership)



© Copyright 2020 by Peter Aiken

2 Obvious Data Challenges



Mizuho Securities

- Wanted to sell 1 share for 600,000 yen
- Sold 600,000 shares for 1 yen
- $347 million loss
- In-house system did not have limit checking
- Tokyo stock exchange system did not have limit checking ...
- … and doesn't allow order cancellations

TIMESONLINE



December 9, 2005

Fat fingered typing costs a trader's bosses £128m

CLUMSY typing cost a Japanese bank at least £128 million and staff their Christmas bonuses yesterday, after a trader mistakenly sold 600,000 more shares than he should have. The trader at Mizuho Securities, who has not been named, fell foul of what is known in financial circles as "fat finger syndrome" where a dealer types incorrect details into his computer. He wanted to sell one share in a new telecoms company called J Com, for 600,000 yen (about £3,000).

Results: It Is Not Always About Money



- Increased safe matches from 3 out of 10 to 6 out of 10

Program Overview

Data Quality by Design
Ensuring Reliable Inputs for AI Systems

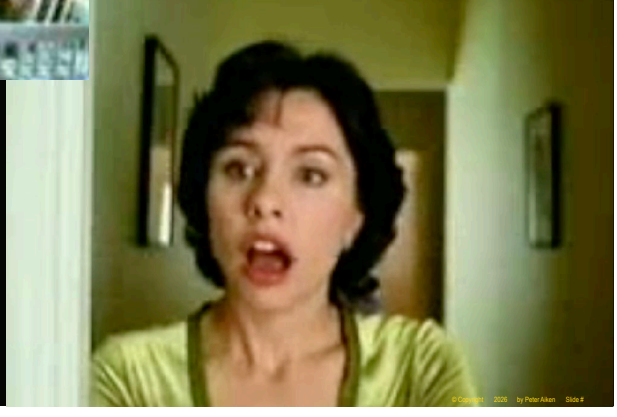
- Definitions
  - Data governance and AI governance
  - Defensive driving is a mindset
  - Trying to clean too much data is unproductive
  - Data quality definitions
  - Practice vs. structural challenges
- Data quality challenges manifest in many ways
  - Hidden data factories
  - Obscure data quality challenges
  - Obvious data quality challenges
- Approaching Data Quality
  - Causes can be difficult to discern
  - The role of data debt
  - Solutions must be built on leverage
- Approach Basics
  - Programmatic approach is most effective
  - Early projects have a dual purposes
  - High quality data requires architecture and engineering
  - Systems thinking
  - Misunderstanding the role of culture
  - Developing repeatable capabilities in core data quality expertise
- How do we get better?
  - Refocus the discussion around business outcomes
  - Leadership
  - Math (cost or investment?)
  - Program focus
  - Storytelling/Practice
- Takeaways and Q&A



Recall the Short Film on the Very First Slide

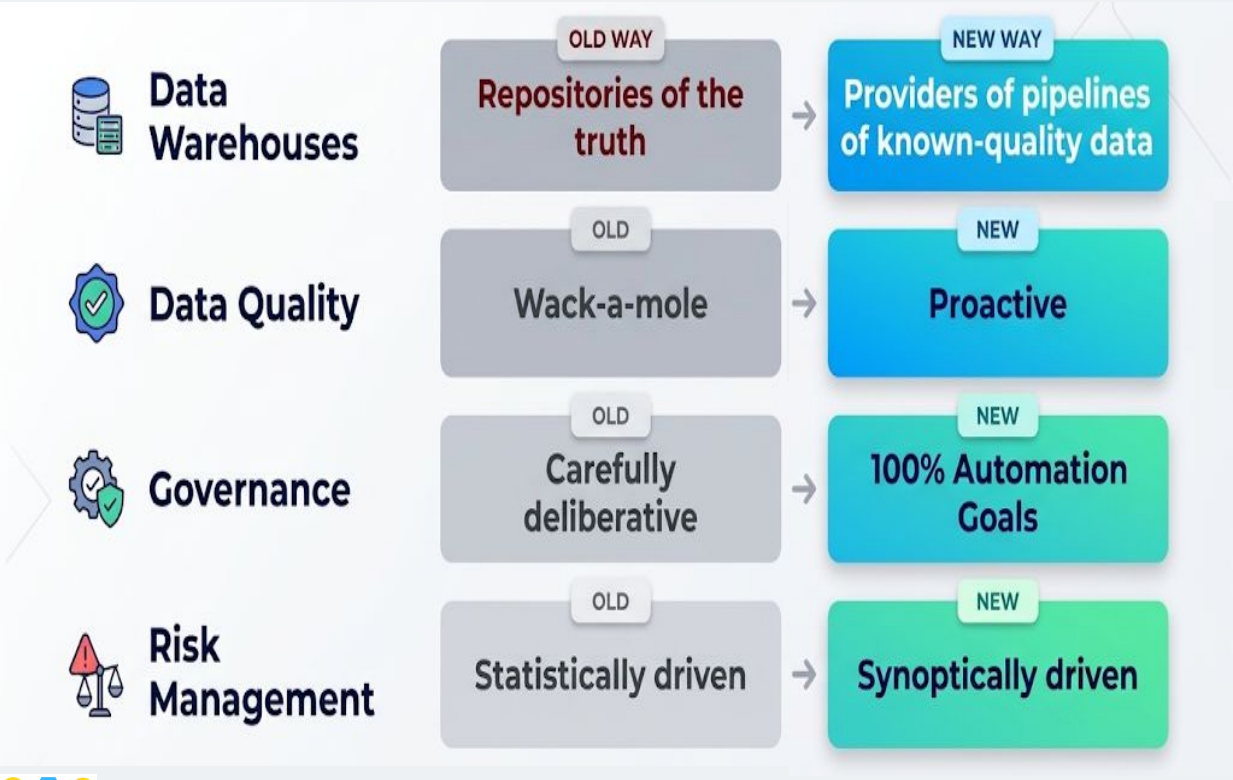


Daughter takes on the task of cleaning Mom's data



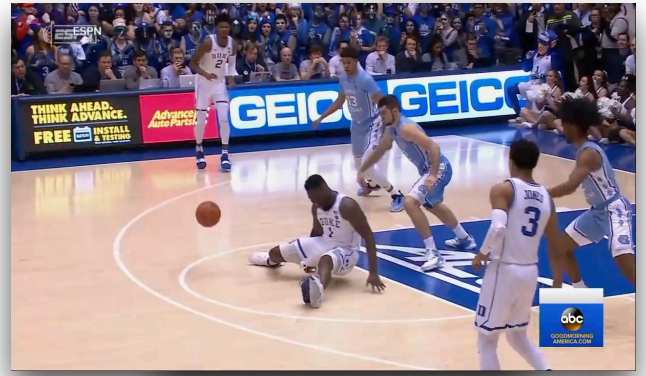
© Copyright 2026 by Peter Aiken Slide #

Rethink …


Data Warehouses
OLD WAY
Repositories of the truth
NEW WAY
Providers of pipelines of known-quality data
Data Quality
OLD
Wack-a-mole
NEW
Proactive
Governance
OLD
Carefully deliberative
NEW
100% Automation Goals
Risk Management
OLD
Statistically driven
NEW
Synoptically driven

Burning Bridge (Zion Williamson's Footwear Debut)

- Something bad happened
  - Imperfect data was to blame
- Someone needs to fix
  - Poor quality data
- You currently have management's attention
  - It is wise to ensure you also have their understanding
- "Do something" often leads to "Buy something"
  - Mostly technology-based
- *Get data quality-ing!*
  - A fool with a tool is still a fool
- Something is accomplished
  - Most often all the project funding is used up



- Early cases have a dual purpose
  - Make the case that this will fix the immediate challenge
  - Illustrate why a programmatic approach is preferable

© Copyright 2026 by Peter Aiken



All organizational challenges have data roots


80% of organizational data is ROT
- Redundant
- Obsolete
- Trivial
It is not easy to visualize the cost of Data Debt or that it depletes organizational resources:
- Slowing progress
- Decreasing quality
- Increasing costs
- Presenting greater risks

HOW DATA PROFESSIONALS USE AI TO ERADICATE DATA DEBT

Tackling legacy mess, shadow pipelines, and decaying pipelines with intelligent automation

AI transforms data debt from a manual burden into an automated, ongoing cleanup process.



1. MAP & DISCOVER THE MESS

- AI-driven discovery crawls unstructured drives and databases to inventory dark data.
- Automatically maps data lineage and identifies abandoned shadow pipelines.



2. AUDIT & DIAGNOSE

- ML algorithms detect data quality anomalies, duplicates, and drift in real-time.
- LLMs analyze and evaluate legacy code, schemas, and stored procedures for redundancy.



3. AUTOMATE DOCUMENTATION

- AI generates instant, accurate metadata, data dictionaries, and catalog descriptions.
- Automatically documents complex SQL queries, ETL jobs, and legacy transformation logic.



4. REFACTOR & CLEANSE

- AI suggests and drafts refactored code to consolidate messy, redundant pipelines.
- Automated workflows auto-correct schema drift and standardizes naming conventions.



5. MONITOR & ALERTS

- Predictive models identify potential system and pipeline failures before they happen.
- Sets intelligent, dynamic thresholds to trigger alerts for unexpected data behavior.



6. PREVENT FUTURE DEBT

- Implements automated guardrails to ensure new ingestion matches governance standards.
- Continuous training and feedback loops keep documentation and code optimized.

THE RETURN ON ERADICATING DATA DEBT

COST SAVINGS
Reduced storage, compute, and licensing fees for redundant

SPEED & AGILITY
Faster query times, rapid pipeline deployment, and agile analytics.

TRUST & COMPLIANCE
Accurate data, clear audit trails, and lower regulatory risk.

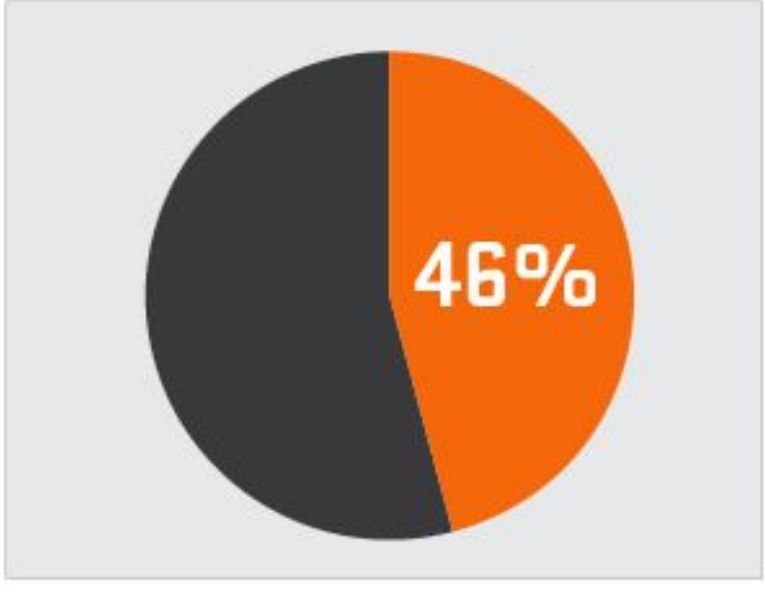
INNOVATION
Freedom from maintenance to focus on strategic AI and analytics.

Bad Data Decisions Spiral

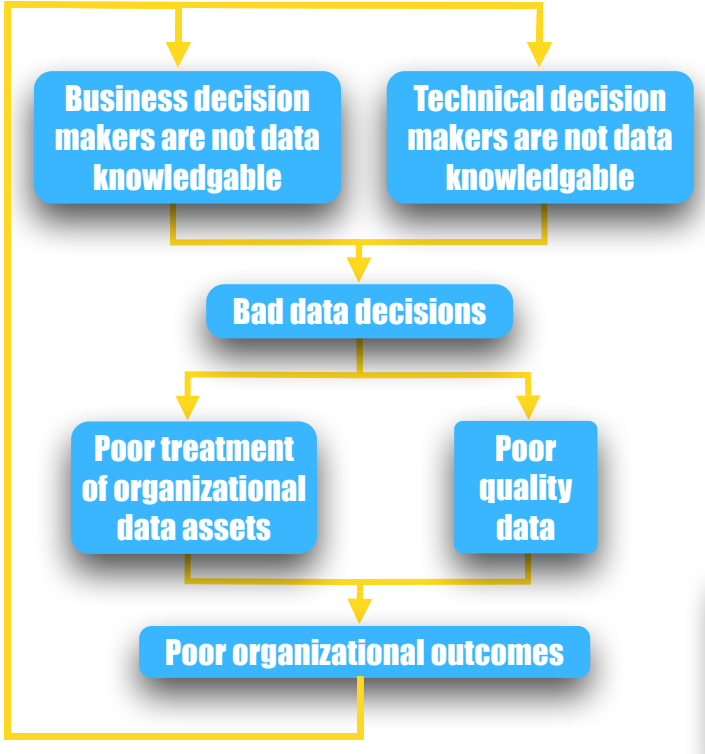
NEWS FLASH!

46% of companies report they made an inaccurate business decision based on bad or outdated data. Bad data leads to bad business decisions. Companies need to be careful that their data is sound – especially when dealing with investors.

Like Comment Share



Bad Data Decisions Spiral







Who Is Joan Smith?



© Copyright 2026 by Peter Aiken

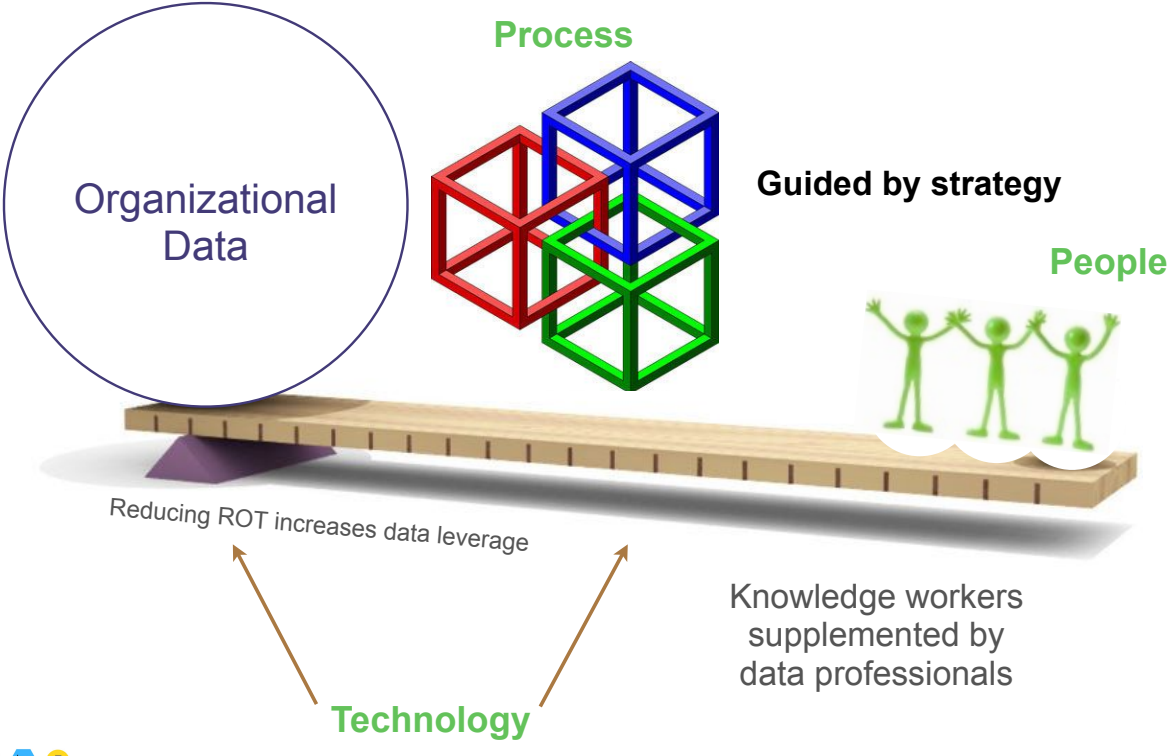
Leverage Is an Engineering Concept

- Using proper engineering techniques, a human can lift a bulk that is weighs much more than the human



Data Quality Is a Wholistic Approach to Obtaining Data Leverage





https://www.computerhope.com/jargon/f/framework.htm

Concrete Example of Data Quality Leverage

- Reference
  - Controls accessible data values
- Master (Main … Golden … ?)
  - Controls access to system capabilities
- Transaction
  - Instances of values

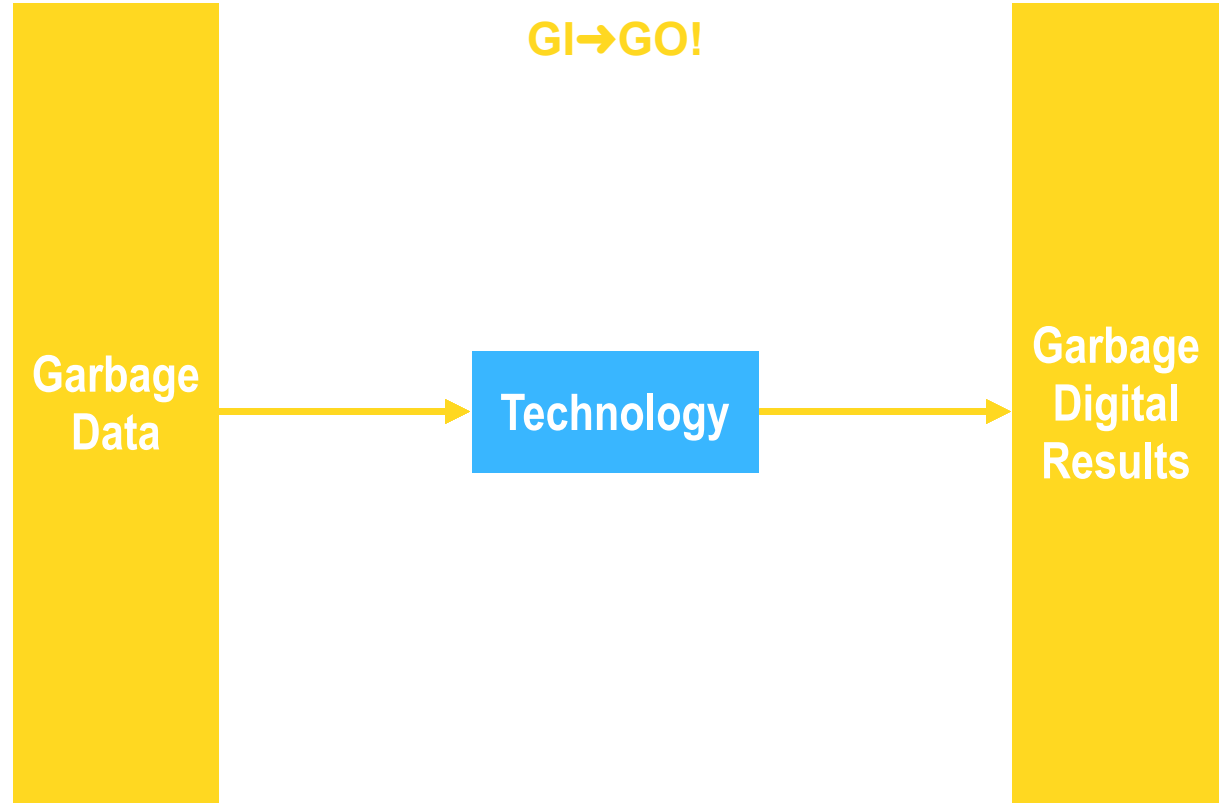
Countries where we do business?
Types of accounts available?
Controlled vocabulary items

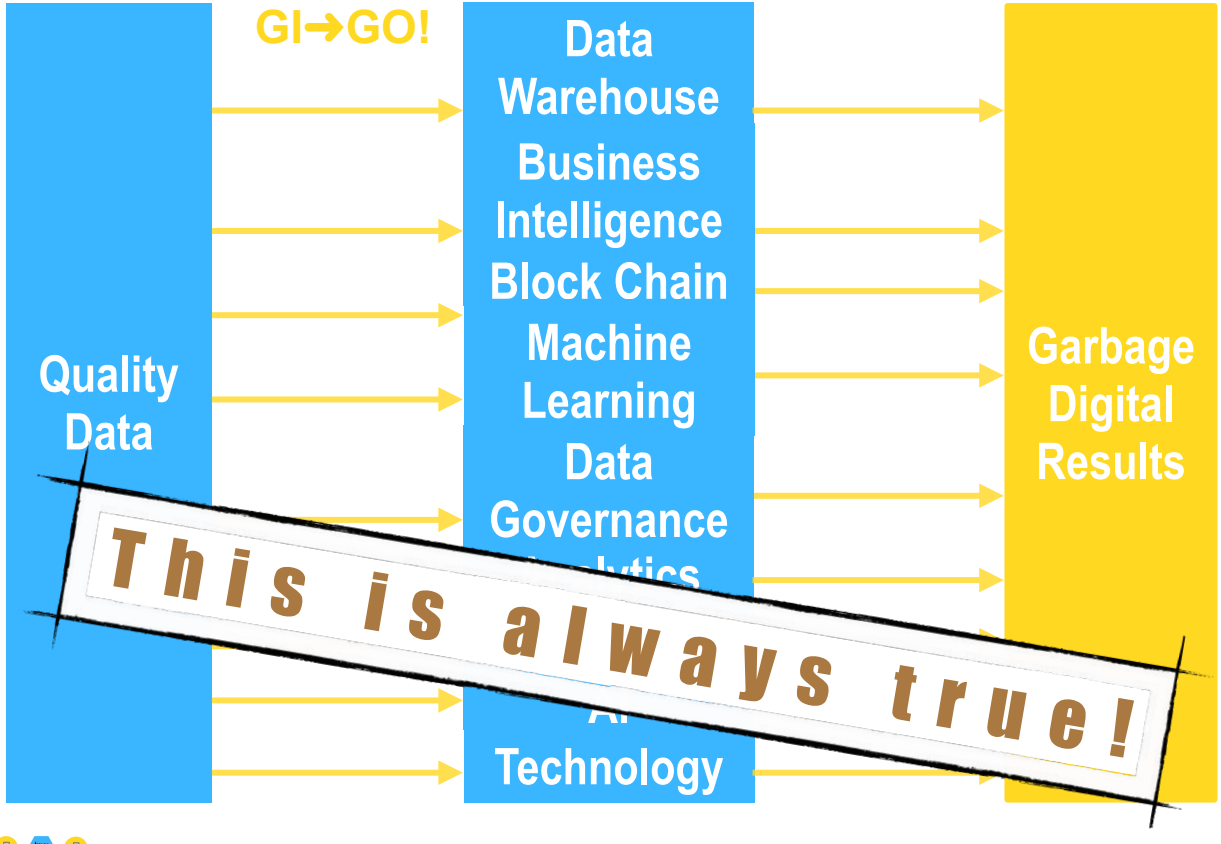
Are you a member of our premium club?
Authorizing uses/users?
Common/standard data structures

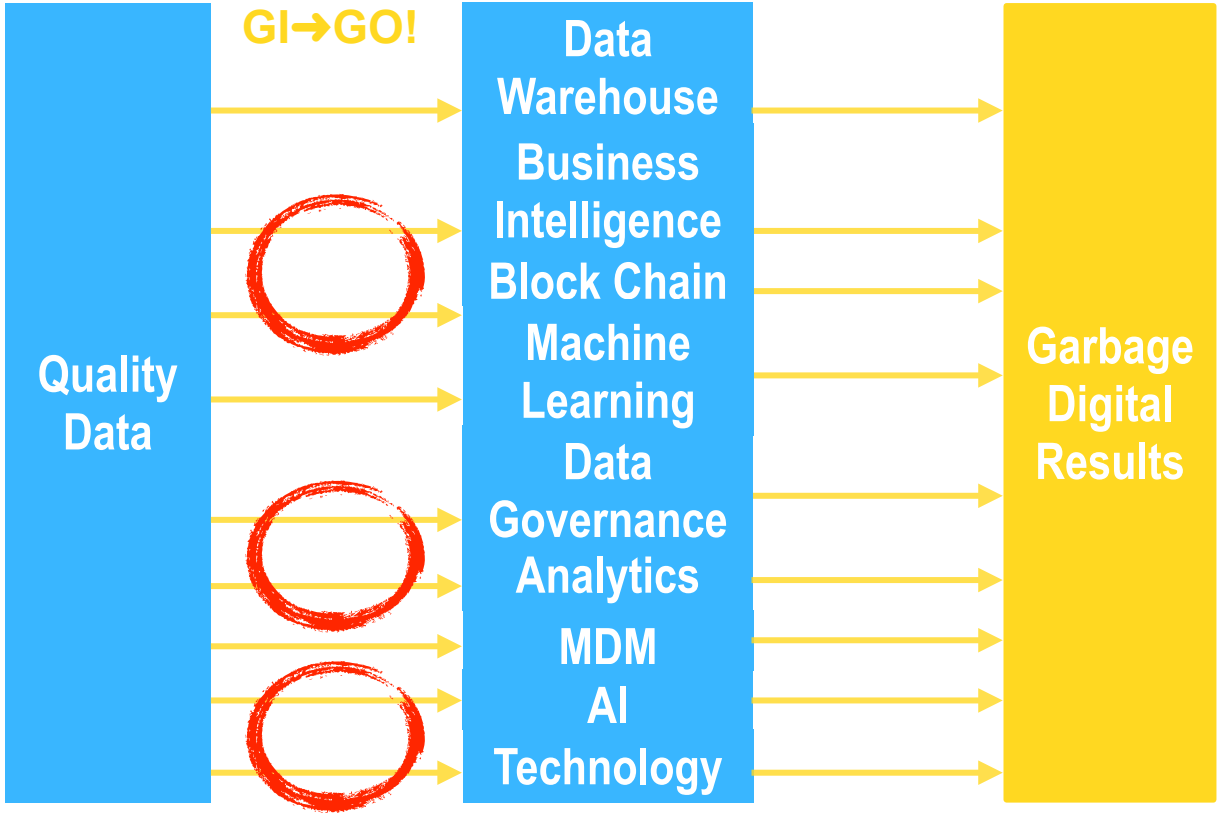
$5
Authorized
Like 👍

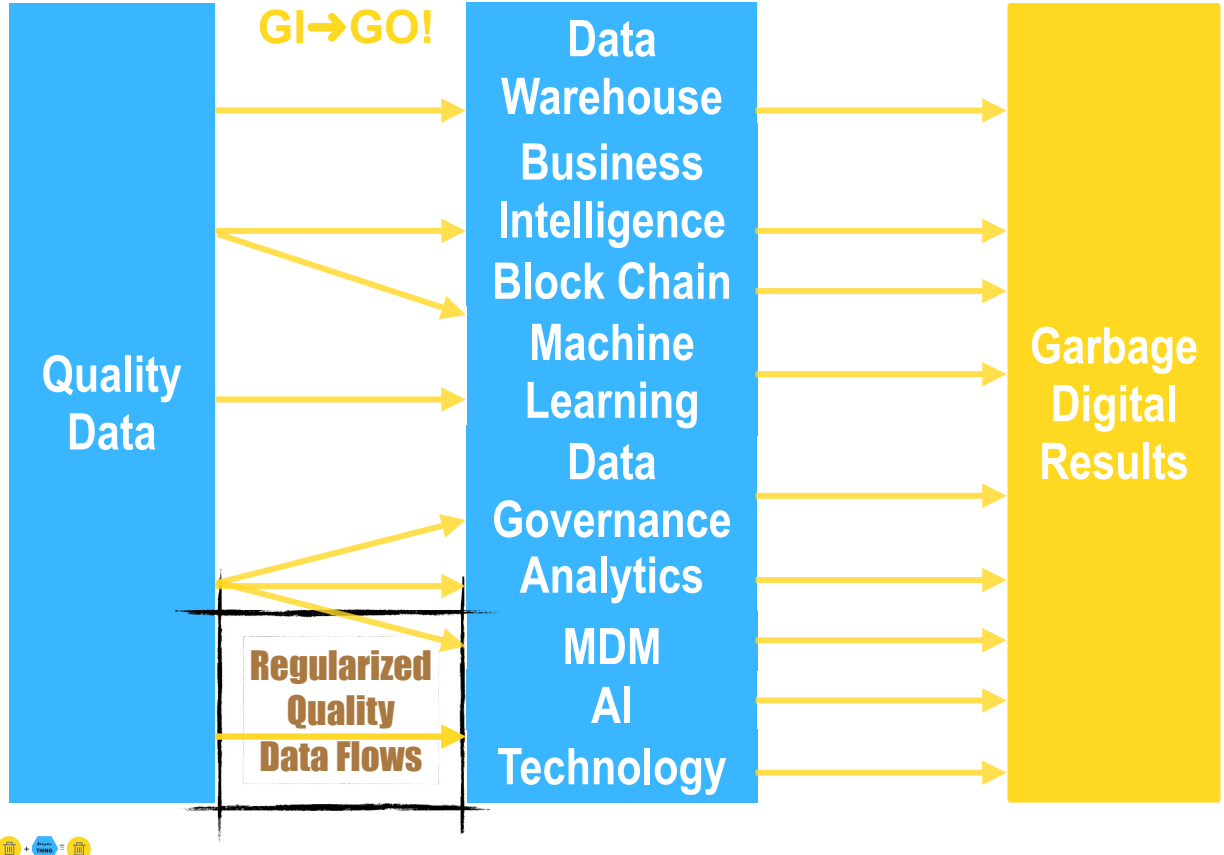
Cannot do business overseas?
Cannot determine product origin?
Cannot add a foreign language to the website?
Cannot select a valid menu item?

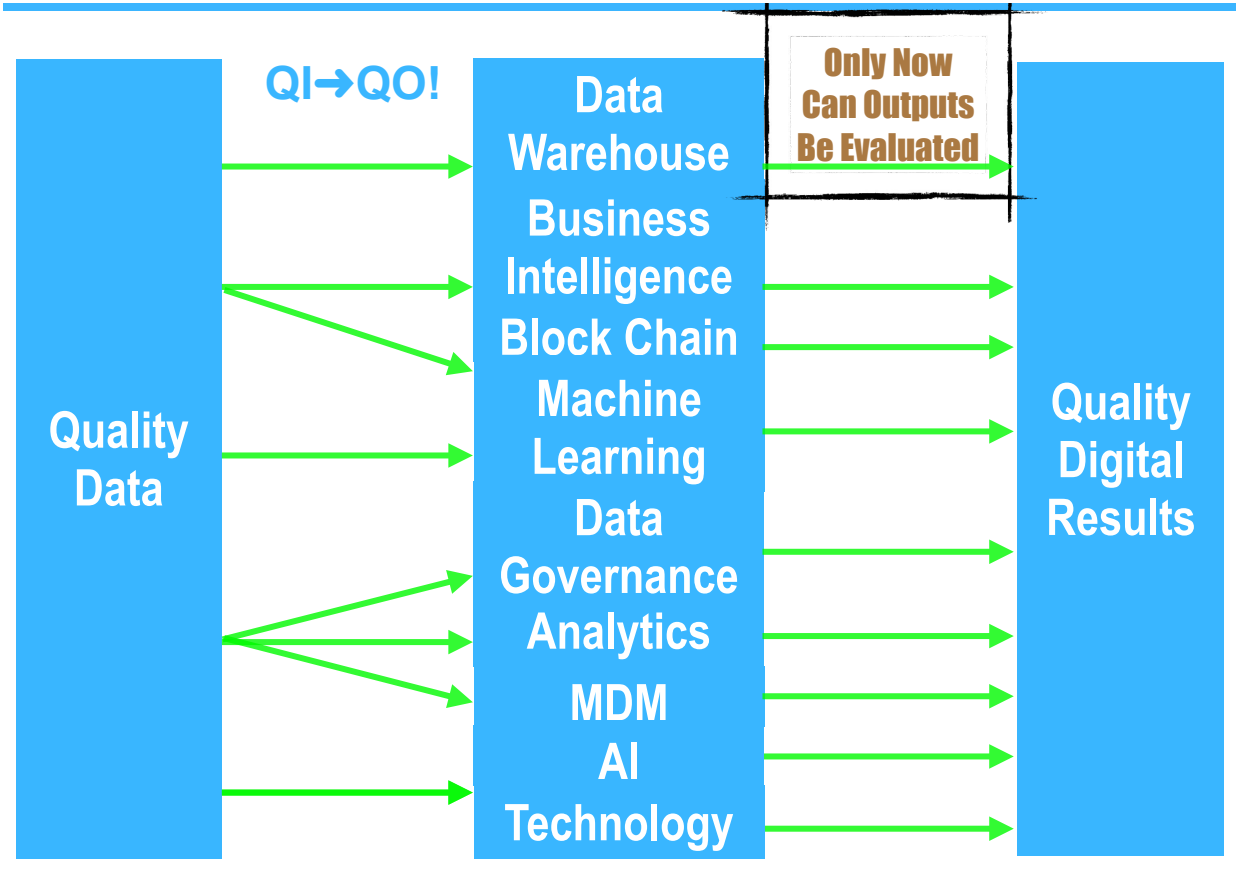
Example based on: Dr. Christopher Bradley of DMAdvisors–he has more, ping him at chris.bradley@dmadvisors.co.uk


GI→GO!
Garbage Data
Technology
Garbage Digital Results


GI→GO!
Quality Data
Data Warehouse
Business Intelligence
Block Chain
Machine Learning
Data Governance
Technology
Garbage Digital Results
This is always true!


Quality Data
GI→GO!
Data Warehouse
Business Intelligence
Block Chain
Machine Learning
Data Governance
Analytics
MDM
AI
Technology
Garbage Digital Results


GI→GO!
Quality Data
Data Warehouse
Business Intelligence
Block Chain
Machine Learning
Data Governance
Analytics
MDM
AI
Technology
Regularized Quality Data Flows
Garbage Digital Results


QI→QO!
Quality Data
Data Warehouse
Business Intelligence
Block Chain
Machine Learning
Data Governance
Analytics
MDM
AI
Technology
Only Now Can Outputs Be Evaluated
Quality Digital Results

General Limitation



[Bad Data] + Anything Awesome [will always yield] Bad Results



Goodness In → Goodness Out

© Copyright 2026 by Peter Aiken

Program Overview

Data Quality by Design

Ensuring Reliable Inputs for AI Systems

- Definitions
  - Data governance and AI governance
  - Defensive driving is a mindset
  - Trying to clean too much data is unproductive
  - Data quality definitions
  - Practice vs. structural challenges
- Data quality challenges manifest in many ways
  - Hidden data factories
  - Obscure data quality challenges
  - Obvious data quality challenges
- Approaching Data Quality
  - Causes can be difficult to discer
  - The role of data debt
  - Solutions must be built on lever

- Approach Basics
  - Programmatic approach is most effective
  - Early projects have a dual purposes
  - High quality data requires architecture and engineering
  - Systems thinking
  - Misunderstanding the role of culture
  - Developing repeatable capabilities in core data quality expertise
- How do we get better?
  - Refocus the discussion around business outcomes
  - Leadership
  - Math (cost or investment?)
  - Program focus
  - Storytelling/Practice
- Takeaways and Q&A



Data Quality Engineering

- Allow the form of the problem to guide the form of the solution
- Provide a means of decomposing the problem
- Feature a variety of tools simplifying system understanding
- Offer a set of strategies for evolving the design of a programmatic solution
- Provide criteria for evaluating the quality of the various solutions
- Facilitate development of a framework for developing organizational knowledge



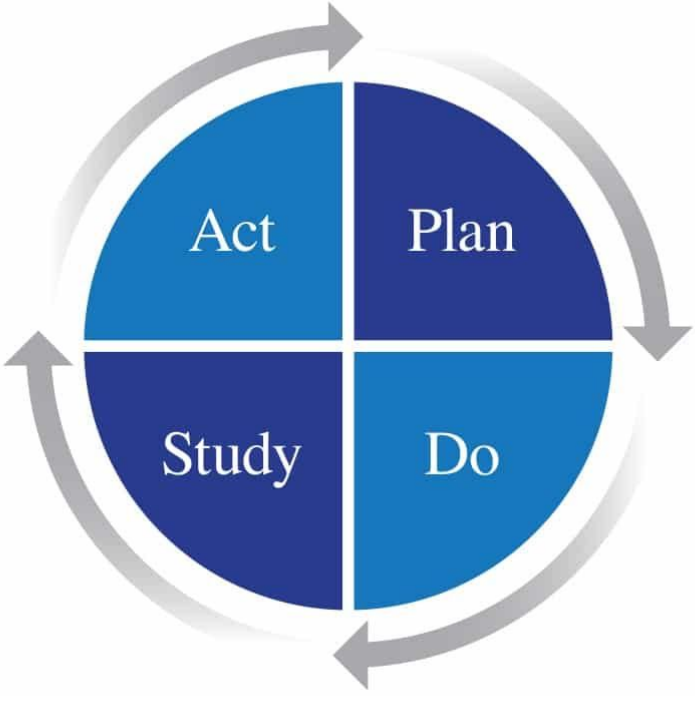
Edward Yourdon and Larry L. Constantine, 1979, *Structured Design: Fundamentals of a Discipline of Computer Program and Systems Design*

The Deming Quality Cycle (aka. Deming Wheel or Deming Cycle)

Deming cycle

"Plan-do-study-act" or "plan-do-check-act"

- Identifying data issues that are critical to the achievement of business objectives
- Defining business requirements for data quality
- Identifying key data quality dimensions
- Defining business rules critical to ensuring high quality data



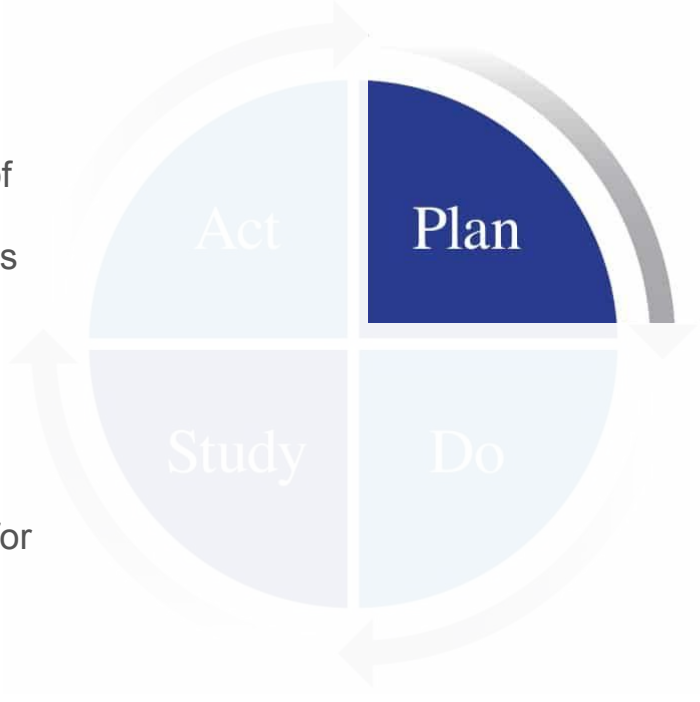
https://deming.org/explore/pdsa/

(1) Plan

Plan for the assessment of the current state and identification of key metrics for measuring quality

The team assesses the scope of known issues

- Determining cost and impact
- Evaluating alternatives for addressing them



https://anythingawesome.com

(2) Do (aka. Deploy)

Deploy processes for measuring and improving the quality of data:

Data profiling

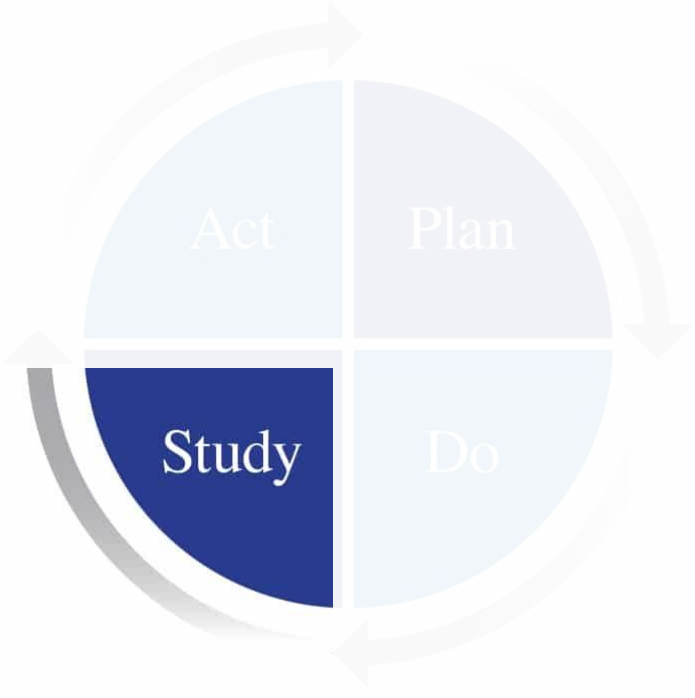
- Institute inspections and monitors to identify data issues when they occur
- Fix flawed processes that are the root cause of data errors or correct errors downstream
- When it is not possible to correct errors at their source, correct them at their earliest point in the data flow



https://anythingawesome.com

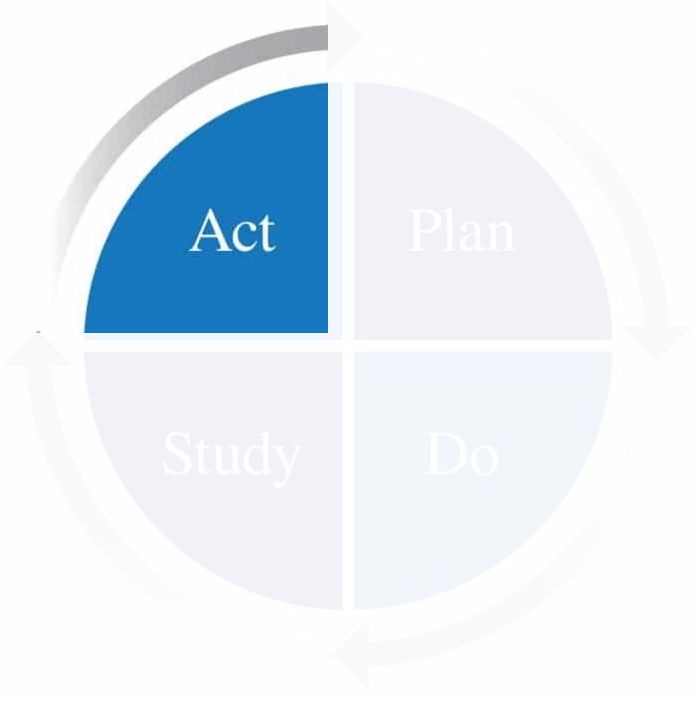
The DQE Cycle: (3) Study (Aka. Check)

- Monitor the quality of data as measured against the defined business rules
- If data quality meets defined thresholds for acceptability, the processes are in control and the level of data quality meets the business requirements
- If data quality falls below acceptability thresholds, notify data stewards so they can take action during the next stage



(4) Act

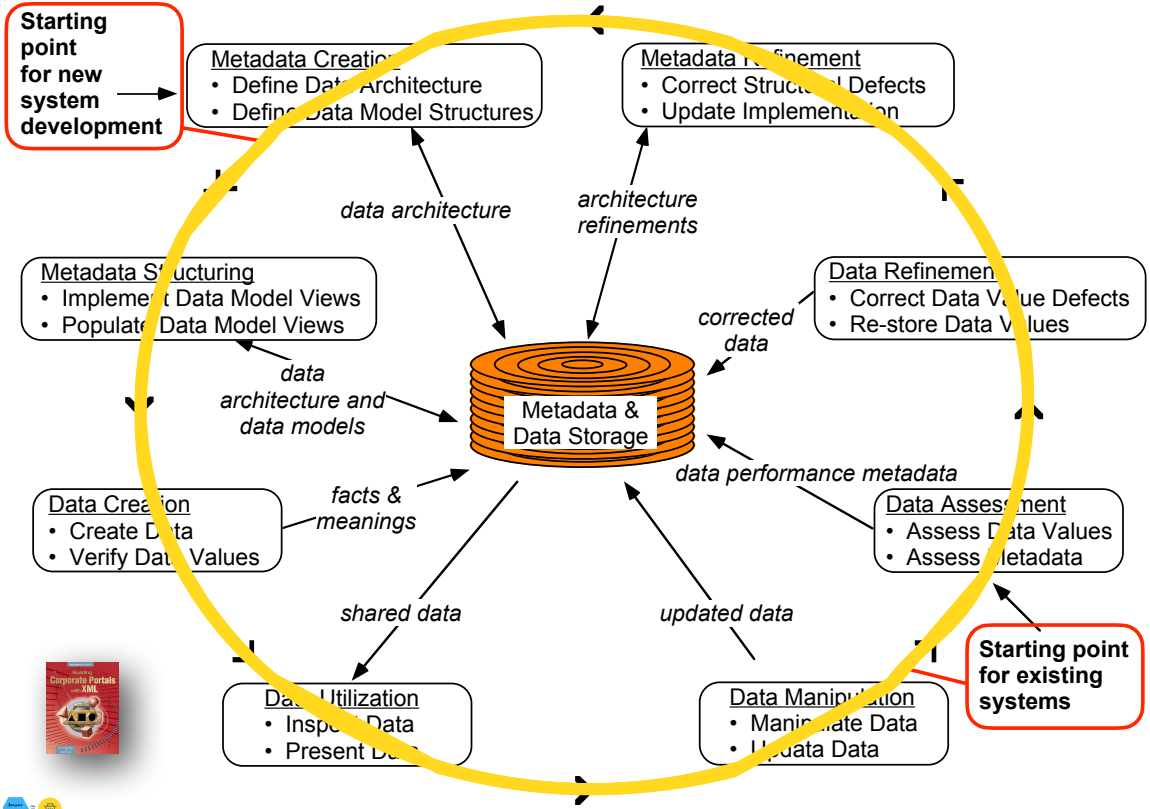
- Act to resolve any identified issues to improve data quality and better meet business expectations
- New cycles begin as new data sets come under investigation or as new data quality requirements are identified for existing data sets



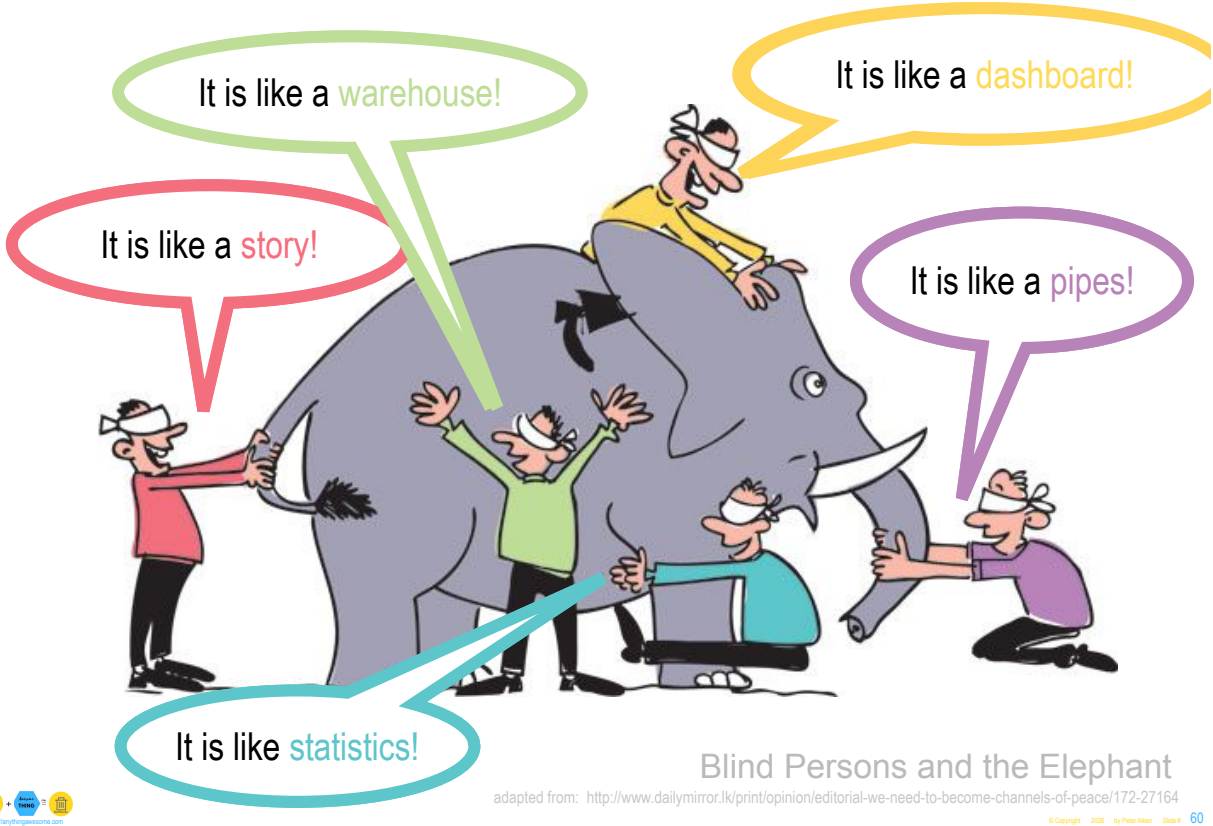


https://anythingawesome.com

Extended Data Life Cycle Model With Metadata Sources and Uses



Data Is Not Broadly or Widely Understood



Blind Persons and the Elephant

adapted from: http://www.dailymirror.lk/print/opinion/editorial-we-need-to-become-channels-of-peace/172-27164

Understanding Systems Thinking

- A framework that is based on the belief that the component parts of a system can best be understood in the context of relationships with other systems, rather than in isolation.

- The only way to fully understand why a problem or element occurs and persists is to understand the part in relation to the whole.

Capra, F. (1996) The web of life: a new scientific understanding of living systems (1st Anchor Books ed). New York: Anchor Books. p. 30





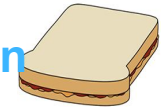


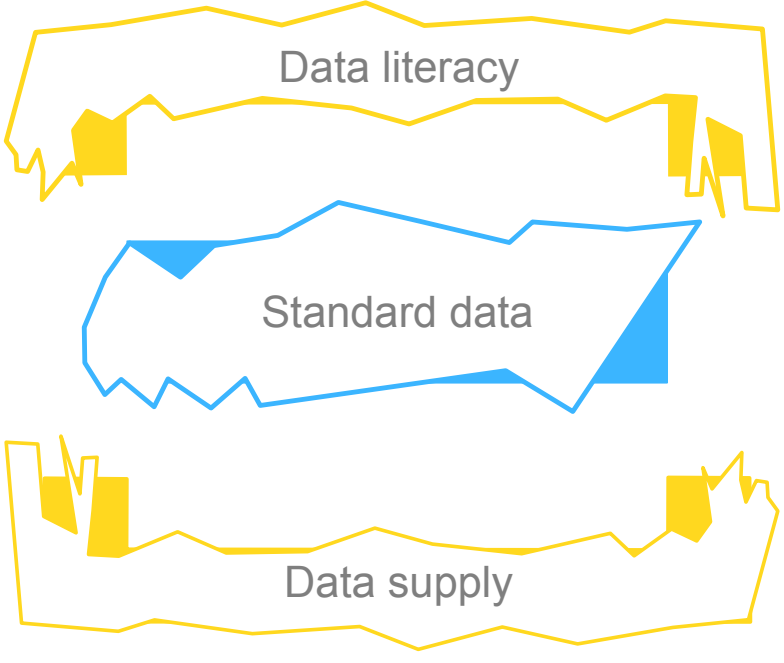
Many DQ Challenges Are Unique and/or Context Specific!

© Copyright 2026 by Peter Aiken


Data Quality
Sandwich
Ingredients

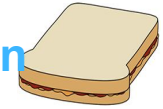
Leverage Point - High Performance Automation





© Copyright 2026 by Peter Aiken

Leverage Point - High Performance Automation

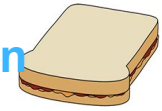


Data literacy

Standard data

Data supply

Leverage Point - High Performance Automation



This cannot happen without investments in engineering and architecture!

Data literacy

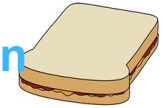
Standard data

Data supply



Quality engineering/ architecture work products do not happen accidentally!

Leverage Point - High Performance Automation



This cannot happen without investments in data engineering and architecture!

Data literacy

Standard data

Data supply



These are necessary but insufficient prerequisites to successful AI augmentation

Logistics Company

- Fortune 450
- Room of 100 associates
- Manually correcting every item on every customer invoice
- Upon noting this to the responsible manager - replied:
  - *This is the best quarter*
  - *Of the best year*
  - *I've ever had*
  - *Perhaps I need to double the number in that room?*



CFO: If you can fix this in 1 month for less than $750,000,000 then its a positive ROI in 30 days!



Mobile Phone Policy



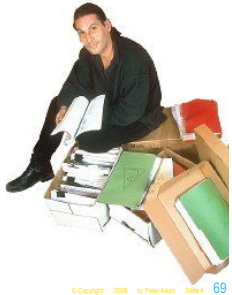


- Must turn in if not used for 90 days
- Many providers of mobile phones
- Warehouse integration strategy
  - Months and millions to develop technologies
  - Dozens of ETL streams to feed the warehouse



- Does not address some fundamental questions
  - It has been 88 days, it that close enough?
  - It has been 92 days, should I wait a bit more?



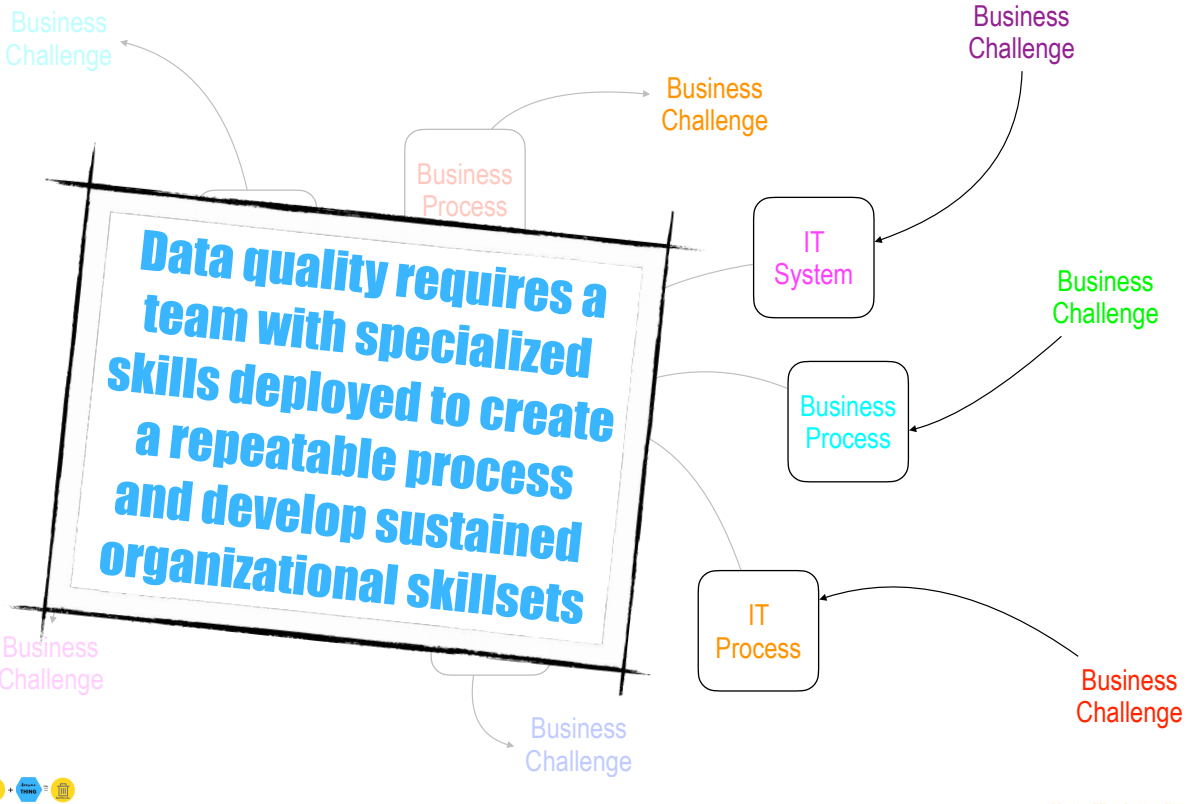


- Reset - 3 billing cycles
  - Eliminates technology need - work level is clerical not technical

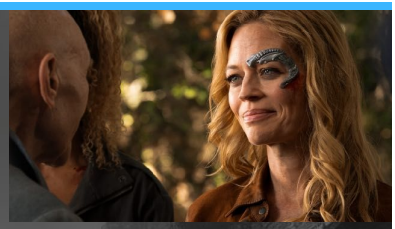
https://anythingawesome.com

© Copyright 2026 by Peter Aiken

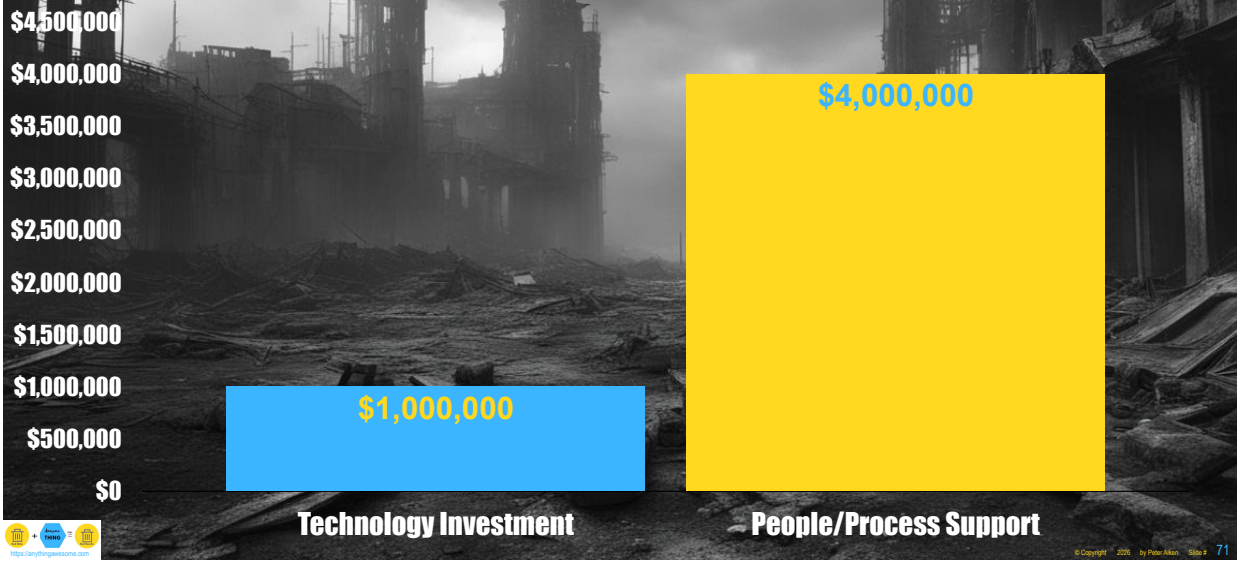
Consistency Encourages Quality Analysis



Data Investment Evaluation



- Remember **7-of-9** from Star Trek?
- Peter's version is **1-to-4**
  - If you invest $1 million in any tool/technology?
  - It requires $4 million in people and process support



© Copyright 2026 by Peter Aiken

WINNING CARDS FOR DATA QUALITY SUCCESS

5 Prerequisite Cards for Data Quality Program Success & The Reality Check

• CARD 1



THE PROJECT NEEDS TO BE SMALL

PREREQUISITE

Small, focused scope to ensure delivery.

REALITY CHECK

Projects shouldn't begin until the data requirements for the entire project are **verified.**

• CARD 2



SKILLED OWNERS & SPONSORS

PREREQUISITE

The product owner, sponsor, or executor executor must be highly skilled in data.

REALITY CHECK

Few in IT currently have the requisite data skills and knowledge.

• CARD 3



AGILE-READY PROCESSES

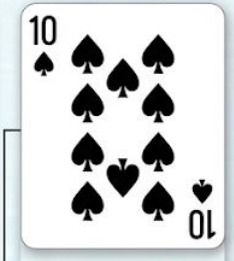
PREREQUISITE

The process must be agile-ready and adaptive.

REALITY CHECK

Agile is a construction technique—data requires more planning before construction.

• CARD 4



HIGHLY SKILLED TEAMS

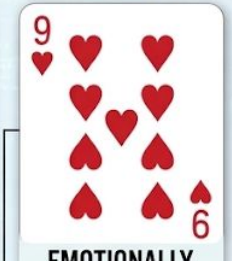
PREREQUISITE

The team must be **highly skilled** in both data quality processes & technology.

REALITY CHECK

Very few teams possess the requisite levels of specialized data skills.

• CARD 5



EMOTIONALLY MATURE ORGANIZATION

PREREQUISITE

The organization must be highly skilled at emotional maturity and change.

REALITY CHECK

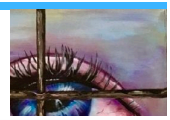
Few organizations understand data 'stuff' and the cultural change it requires.

SUCCESS DEMANDS BALANCING THE IDEAL STATE WITH REALITY

https://anythingawesome.com

© Copyright 2020 by Peter Aiken Slide #

Sample from: https://artist.com/kathy-linden/on-outside-looking-in/?artid=4385



External Comprehension

Everything Else Data

Data Management

Data Governance Program

Data (blah blah blah)

Most do not appreciate the difference between Data Governance and the other data stuff that needs to be done

Data Program

Program Overview

Data Quality by Design
Ensuring Reliable Inputs for AI Systems

- Definitions
  - Data governance and AI governance
  - Defensive driving is a mindset
  - Trying to clean too much data is unproductive
  - Data quality definitions
  - Practice vs. structural challenges
- Data quality challenges manifest in many ways
  - Hidden data factories
  - Obscure data quality challenges
  - Obvious data quality challenges
- Approaching Data Quality
  - Causes can be difficult to discer
  - The role of data debt
  - Solutions must be built on lever

- Approach Basics
  - Programmatic approach is most effective
  - Early projects have a dual purposes
  - High quality data requires architecture and engineering
  - Systems thinking
  - Misunderstanding the role of culture
  - Developing repeatable capabilities in core data quality expertise
- How do we get better?
  - Refocus the discussion around business outcomes
  - Leadership
  - Math (cost or investment?)
  - Program focus
  - Storytelling/Practice
- Takeaways and Q&A



© Copyright 2026 by Peter Aiken

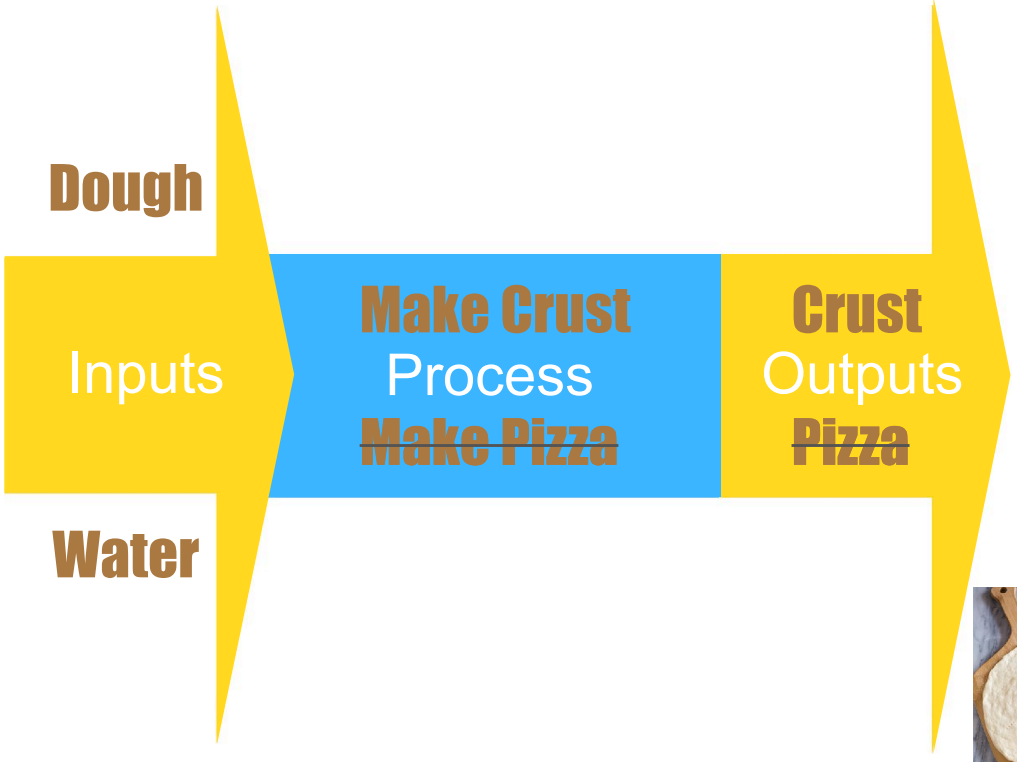


Compare Story Types

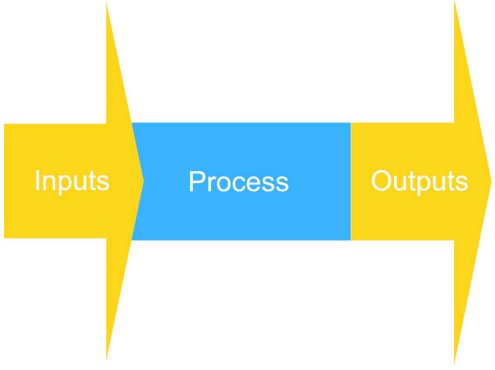


Technical	Business
Clean some data	Decrease the number of undeliverable targeted marketing ads
Reorganize the database	Increase the ability of the salesforce to perform their own analyses
Develop a taxonomy	Create a common vocabulary for the organization
Optimize a query	Shaved 1 second off a task that runs a billion times a day
Reverse engineer the legacy system	Understand what was good about the old system so it can be formally preserved and, what was bad so it can be improved

Input ➜ Process ➜ Output Diagram







Data Steward Quality Responsibilities

- Inputs
  - From where, do each of these my responsible data items come?
  - Why are they produced?
  - *What level of quality is required by 'my processes?'*
- Process
  - What business processes use the data within my fiduciary responsibility?
  - For what business purpose do they use each data item?
  - *What role does quality play for my processes to contribute?*
- Output
  - What downstream business processes consume data that was under my fiduciary care?
  - For what purpose are each data items consumed?
  - *What quality attribute are required by each downstream consumer?*

Simple Math

If X is invested in Y then outcome Z will result (Z > X)

- At the beginning of a project,
- Where the parties know the least about each other
- All are expected to agree on the meaning of price, timing, and functionalities
- Define X (some resources)
- Define Y (cleaning 1 set of data)
- Define Z (that data will be clean)



Simple Math

If $100 is invested in cleaning 1 set of data then outcome $1000 will result

- Define X ($100)
- Define Y (cleaning 1 set of data)
- Define Z ($1000)



THE ANATOMY OF A DATA QUALITY PROGRAM

Key pillars of an ongoing commitment to data excellence

A successful program is not a one-time project—it is a continuous journey that requires governance, leadership, and resources.

THE FOUNDATION: AN ONGOING COMMITMENT

- **WHAT IS IT?**
  A perpetual "Data Quality Program" is an ongoing commitment, not a one-time initiative.
- **EVOLUTIONARY**
  Permits continuous, evolutionary improvement of the data quality approach over time.

INHERITED FRAMEWORK

- **THE APPROACH INHERITS**
  - Enterprise Budget & Funding
  - Strategic Organizational Priorities
- **DIRECT ALIGNMENT**
  Directly shapes the program based on the organization's overarching business goals.

GOVERNANCE

- **COORDINATION & CONTROL**
  Requires senior-level coordination, direction, and operational control.
- **MANAGEMENT**
  Integrates comprehensive change management and proactive risk management into the data lifecycle.
- **LEADERSHIP**
  Mandates strong, active executive leadership capabilities to champion and enforce data quality standards.

SENIOR-LEVEL ENABLERS

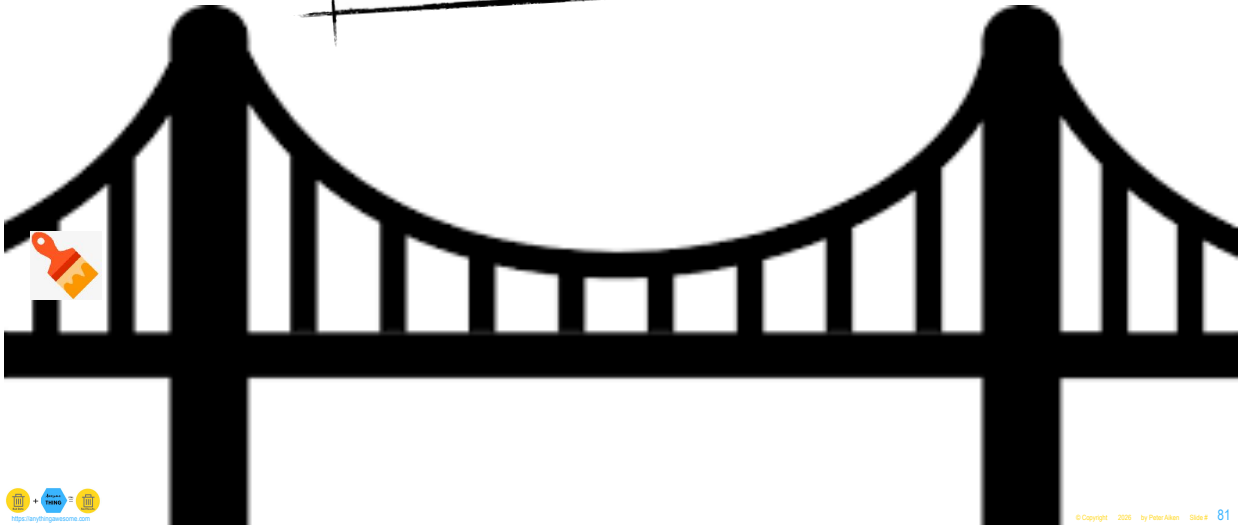
- **ATTENTION**
  Ongoing senior-level attention and improving topical facility (expertise and understanding).
- **REALISTIC OUTCOMES**
  Setting reasonable timelines and expectations, recognizing that data quality is a long-term journey.

GOVERNANCE • LEADERSHIP • EVOLUTION

https://anythingawesome.com
© Copyright 2026 by Peter Aiken Slide # 80

Bridge Maintenance

- Painting is continuous and optimized
- As is much of the remaining maintenance
- Ensures a knowledgeable workforce



https://anythingawesome.com

Program Overview

Data Quality by Design
Ensuring Reliable Inputs for AI Systems

- Definitions
  - Data governance and AI governance
  - Defensive driving is a mindset
  - Trying to clean too much data is unproductive
  - Data quality definitions
  - Practice vs. structural challenges
- Data quality challenges manifest in many ways
  - Hidden data factories
  - Obscure data quality challenges
  - Obvious data quality challenges
- Approaching Data Quality
  - Causes can be difficult to discern
  - The role of data debt
  - Solutions must be built on leverage
- Approach Basics
  - Programmatic approach is most effective
  - Early projects have a dual purposes
  - High quality data requires architecture and engineering
  - Systems thinking
  - Misunderstanding the role of culture
  - Developing repeatable capabilities in core data quality expertise
- How do we get better?
  - Refocus the discussion around business outcomes
  - Leadership
  - Math (cost or investment?)
  - Program focus
  - Storytelling/Practice
- Takeaways and Q&A



SLATE News & Politics Culture Technology Business Human Interest Search Sign In

https://slate.com/business/2020/06/irs-stimulus-check-dead-people-yawn.html

MONEYBOX

Who Cares if the IRS Sent Stimulus Checks to Dead People?

This isn't the scandal journalists are making it out to be.

By JORDAN WEISSMANN JUNE 25, 2020 • 3:17 PM

TWEET SHARE COMMENT



Leave the IRS alone. Zach Gibson/Getty Images

Payments sent	160,000,000
to dead taxpayers	1,100,000
error rate	0.4%
(substantive?)	**$1.4 billion**

Note: IRS lawyers determined "did not have legal authority to deny payments to those who filed a return for 2019, even if they were deceased at the time of payment" and checking with SSA might have slowed the distribution process down.

Economic Policy

Treasury sent more than 1 million coronavirus stimulus payments to dead people, congressional watchdog finds

The checks sent to dead people as of April 30 totaled nearly $1.4 billion, according to the Government Accountability Office

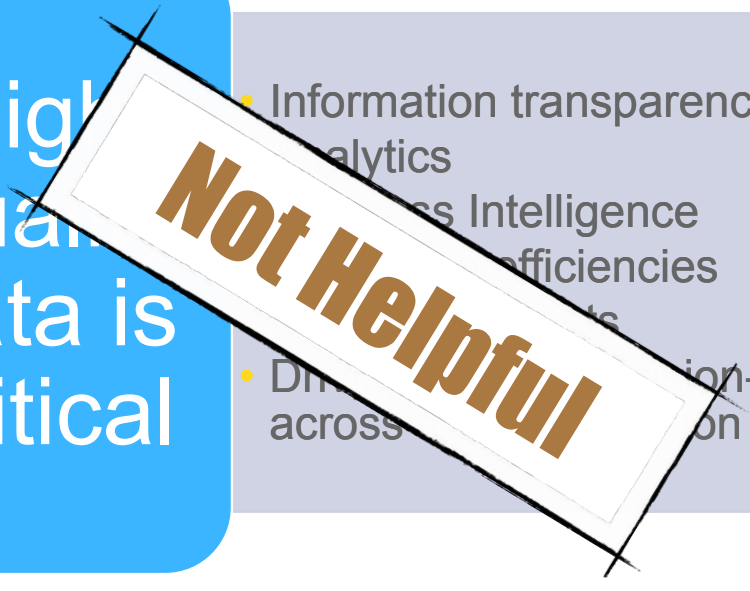
https://www.washingtonpost.com/us-policy/2020/06/25/irs-stimulus-checks-dead-people-gao/

From SLATE:

- The real headline here should really be that the government did its job pretty well!
- Speed was the priority
- Within two weeks, the IRS delivered 80 million payments electronically
- "Speed is part of the reason that poorer households were able to start spending normally again by May despite an unemployment rate rivaling the Great Depression's."

High Quality Data is Critical

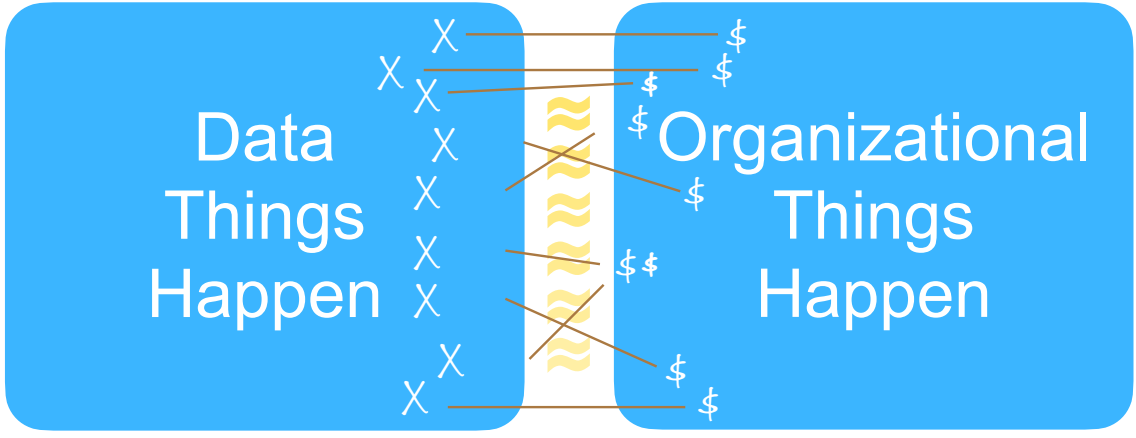
- Information transparency
- alytics
- ss Intelligence
- fficiencies
- ts
- Dri ion-making across on



© Copyright 2020 by Peter Aiken

This Approach Only Works if

- We know where the data that needs to be fixed–resides
- We can communicate precisely and correctly amongst team members, sponsors, collaborators
- We are adept with the correct technological support
- …



Data Value Quality

Data Value Quality: Attributes & Associated Definitions		
V1	Correctness/Accuracy	Data values maintained are free from fault, recording defects or damage
V2	Currency	Data values maintained are the most up-to-date and match user expectations
V3	Time Period	Data values maintained cover the time period required by the users
V4	Clarity	Data values maintained match the breadth and depth of the user request parameters
V5	Precision	Data values are maintained with the amount of precision or detail required by the user
V6	Reliability	Data values stored can be depended up on by the user understated conditions
V7	Consistency	Data values continue to be maintained in a steady, dependable manner
V8	Timeliness	Data values are updated as often as the user requires
V9	Relevance	Data values stored are directly responsible to the specific user needs
V10	Completeness	Attributes of entities requiring values are not null

© Copyright 2020 by Peter Aiken

Data Representation Quality

Data Representation Quality: Attributes and Associated Definitions		
R1	Timeliness	Data should be promptly presented to the users at the time when it is needed
R2	Conciseness	Data presented to the users match user breadth/depth requirements without data loss.
R3	Clarity	Data are presented in a form that is easiest for the user to understand given the request circumstances
R4	Consistency	Data presented to the users lacks nothing with respect to the user's information requirements
R5	Detail	Data are presented in the level of detail most appropriate for the user's need
R6	Accessibility	Data presented to the users is free from retrieval fault, data displayed unaltered from what was stored
R7	Order	Data are presented in a sequence fitting the user's need and their cognitive style
R8	Flexibility	Data are able to be easily transformed between systems, formats, media to best match user needs
R9	Portability	Data are able to be migrated from application to application without data loss
R10	Presentation Appropriateness	Data are presented in a format facilitating user comprehension
R11	Media	Data are presented using media most effective for user comprehension
R12	Unambiguousness/ Interpretability	Data presented to the users requires no interpretation to comprehend the correct value

Data Model Quality

Data Model Quality: Attributes & Associated Definitions		
M1	Completeness	The model is comprehensive enough to be used for a reference – containing complete enough subject areas to be of use
M2	Definition Clarity/Unambiguity	The model is developed and maintained according to generally accepted modeling principles indicating the modelers consistently and correctly applied the techniques
M3	Relevance	The model contents represent the facts of interest to the user
M4	Value Obtainability	The data model is structured so that users can obtain the facts they require
M5	Comprehensiveness	This quality attribute addresses the issue "Did the modelers include all of the information they desired to in the model? Is this model populated with sufficient data to be useful?"
M6	Essentialness	The model contains only those elements fundamentally required to describe the subject
M7	Attribute Granularity	The model is structured so that it manipulates the level of detail desired by the users
M8	Domain Precision	The model maintains the factual precision desired by users
M9	Naturalness	The model "fits" with the way users assimilate facts into their work processes
M10	Occurrence Identifiability	The model maintains sufficient access means to uniquely identify facts required by users
M11	Robustness	Both the model component definitions and the relationships between the entities are free from interpretation-based faults
M12	Flexibility	The model is maintained in a fashion where it is able to be useful in multiple applications

Data Architecture Quality

Data Architecture Quality: Attributes & Associated Definitions		
A1	Architectural Completeness	The architecture is comprehensive enough to be used by any functional area of the organization wishing to utilize it
A2	Architectural Correctness	The information describing the architecture is correctly represented with the appropriate methodology. That is, the organization can use the methodology to maintain uniform data definitions throughout the organization
A3	Management Utility	The data architecture is widely used by the organization in strategic planning and systems development as an indication of its utility. In practice, architectures too often wind up as shelf ware
A4	Data Management Quality	The organization as a whole is data-driven. Data models are developed and managed from an organization-wide perspective, guided by the organizational data structure. Data are managed with distributed control from a centralized unit
A5	Data Sharing Ability	The data architecture serves as the basis for negotiating and implementing
A6	Functional Data Quality	Data are engineered in support of business functional area requirements where data elements for individual systems are derived from organizational metadata requirements and implemented using organizational systems designed to support information representation
A7	Data Operation Quality	Data quality engineering is established as a functional area actively and consistently applying data quality engineering methods to data elements
A8	Evolvability	The organizational data architecture is maintained in a flexible, evolving fashion to enable the fulfillment of future user requirements
A9	Organizational Self-Awareness	Organization ability to investigate architecture use and determine the types of value that it provides to end-users. Feedback helps data architects refine the architecture to make it more useful organizationally.

Time: 19:00 UTC (2:00 PM NYC) | Presented by: Peter Aiken, PhD

Upcoming Events

Data Governance/Stewardship in Action: Governing AI's Data for Long-Term Value

8 December 2026

Data Governance for the AI Era: Building Trust, Accountability, and Value

13 *January 2027*

Modern Data Architecture: Enabling Value from Data Products

9 February 2027



Note: In this .pdf, clicking any webinar title opens the registration link

Brought to you by:

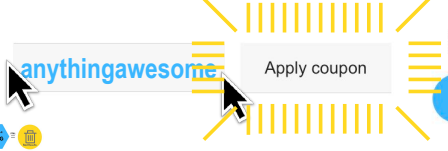
DATA-ED ONLINE

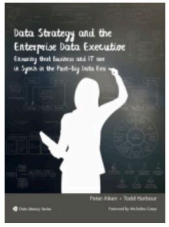
DATAVERSITY

©Copyright 2026 by Peter Aiken Slide #

Event Pricing on Peter's Books

- 20% off directly from the publisher on select titles
- My 'Book Store' @ https://anythingawesome.com/books-overview.html
- Enter the code "anythingawesome" at the Technics bookstore checkout where it says to "Apply Coupon"

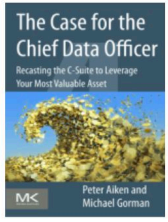




Data Strategy and the Enterprise Data Executive

Ensuring that Business and IT are in Synch in the Post-Big Data Era

Learn More of Data Strategy

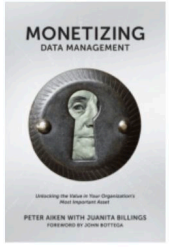


The Case for the Chief Data Officer

Recasting the C-Suite to Leverage Your Most Valuable Asset

(The Chinese Translation Title is: Chief Data Officer Combat)

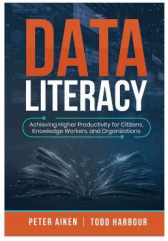
Learn More of the Case for Data Leadership



Monetizing Data Management

17 Case Studies Illustrating How Data Leveraging (Big and Small) Can Produce Quantifiable Results That Are of Keen Interest to C-Suite Occupants

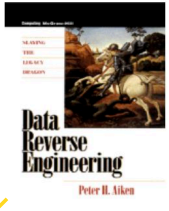
Learn More of Monetizing Data



Data Literacy: Achieving Higher Productivity for Citizens, Knowledge Workers, and Organizations

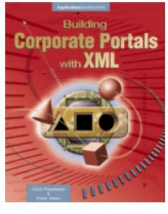
Citizens and organizations need to improve their data literacy to 'do more with data'

Learn More of Data Literacy



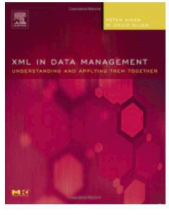
Data Reverse Engineering

Learn More of Data Reverse Engineering



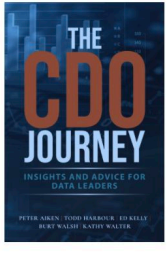
Building Corporate Portals with SML

Learn More Corporate Portals (& XML)



XML in Data Management

Learn More of XML and Data Management



The CDO Journey: Insights and Advice for Data Leaders

Learn More of the CDO Journey

Critical Design Review?

Mentoring?

Executive Data Literacy Training?

Collaboration?



Independent Verification & Validation



Peter.Aiken@AnythingAwesome.com +1.804.382.5957

Reverse Engineering Expertise?

Hiring Assistance?

Thank You!

Use your data more strategically?

Tool/automation evaluation?

Book a call with Peter to discuss anything - https://anythingawesome.com/OfficeHours.html