Getting Data Quality Right

Engineering Business Success Stories

peter.aiken@anythingawesome.com +1.804.382.5957

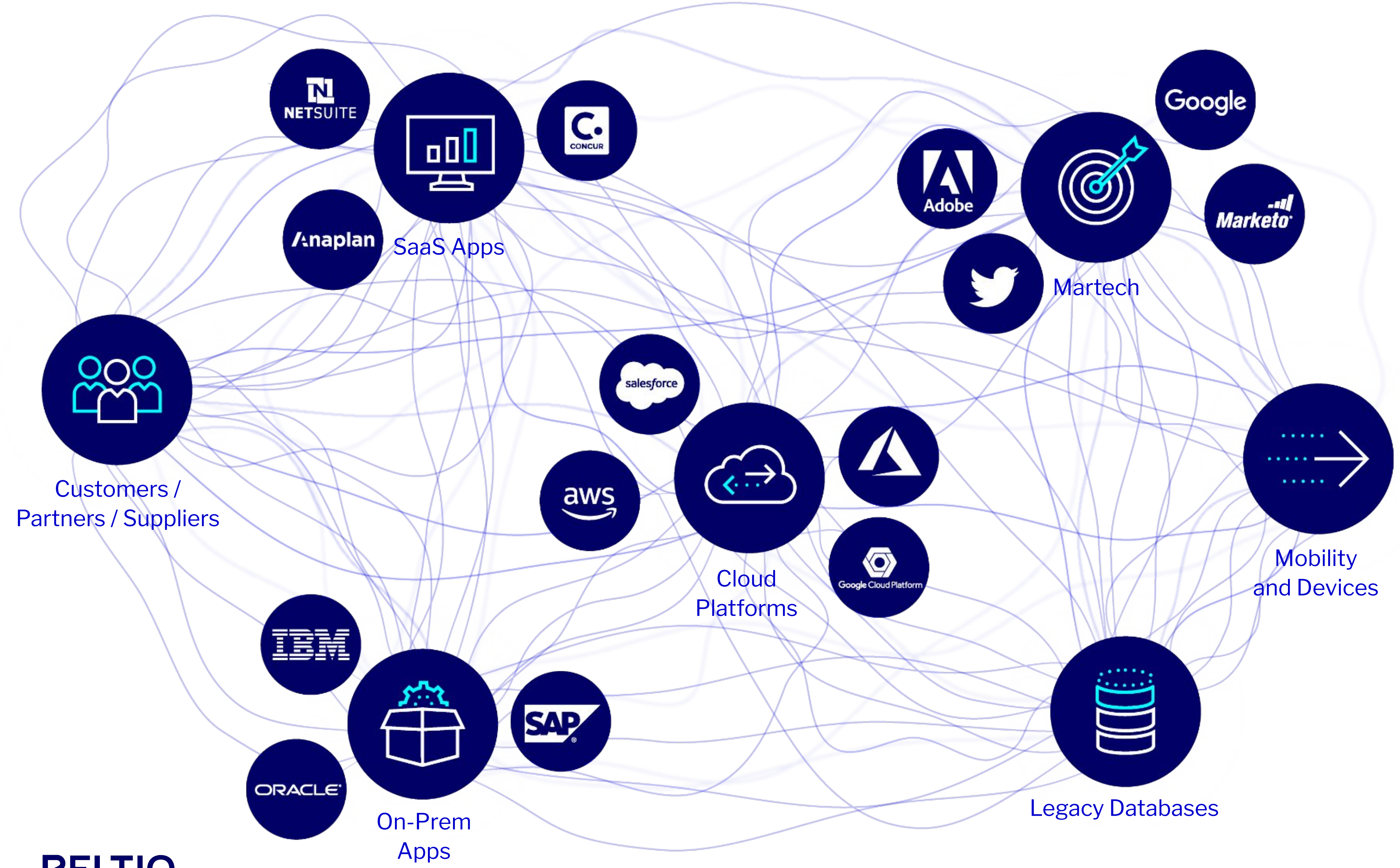
Peter Aiken, Ph.D.

- I've been doing this a long time
- My work is recognized as useful
- Associate Professor of IS (vcu.edu)
- Institute for Defense Analyses (ida.org)
- DAMA International (dama.org)
- MIT CDO Society (iscdo.org)
- Anything Awesome (anythingawesome.com)
- Experienced w/ 500+ data management practices worldwide
- Multi-year immersions
  - US DoD (DISA/Army/Marines/DLA)
  - Nokia
  - Deutsche Bank
  - Wells Fargo
  - Walmart …
- 12 books and dozens of articles

Reltio - Data Quality

Mike Frasca
Field CTO

Data silos and complexity accelerate data quality challenges

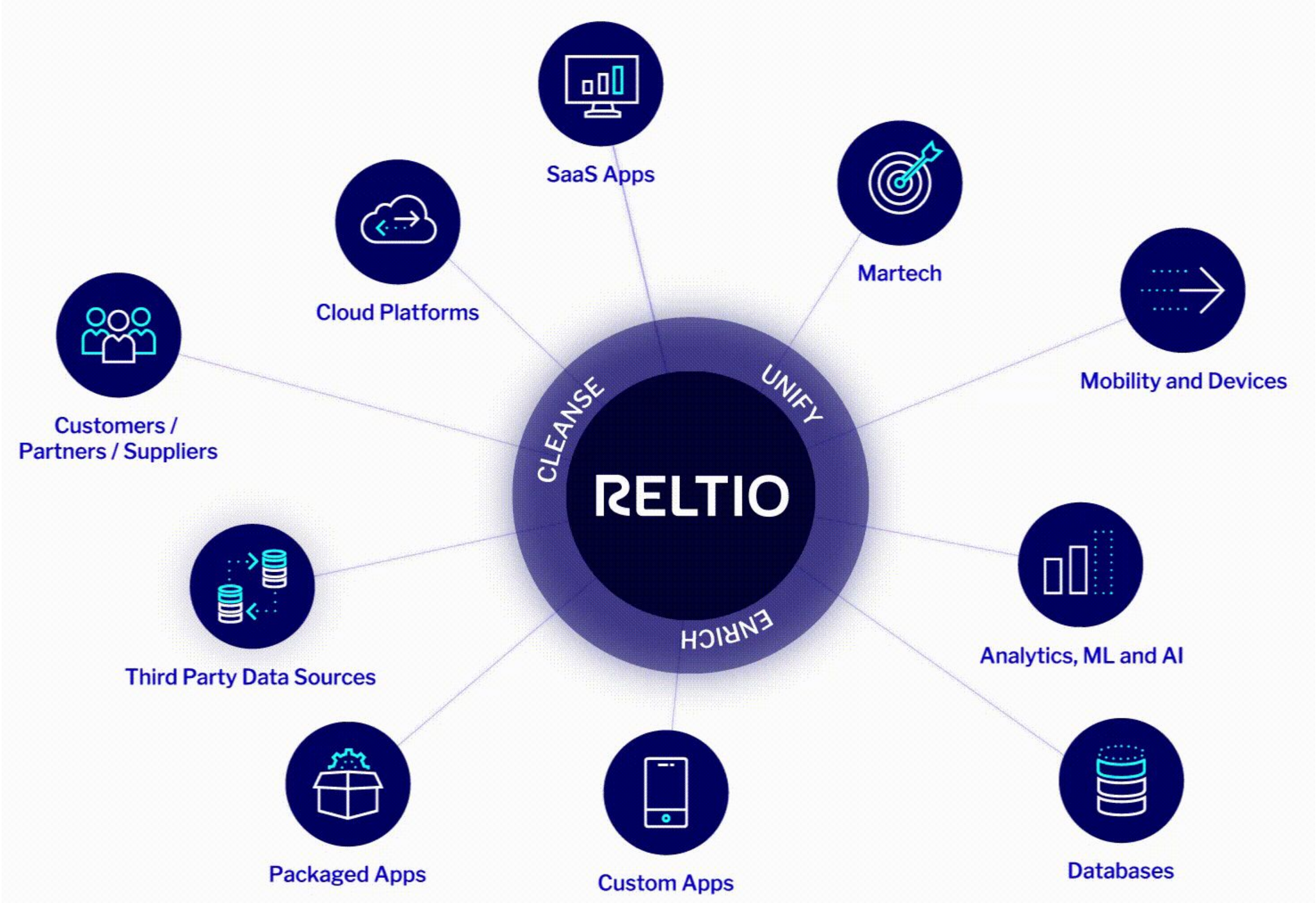


Highly fragmented

Poor quality, duplicate data

Delays in activating data

Reltio delivers unified, trusted core data in real-time



Single source of reliable core data

Trusted data with milliseconds latency

Time to value <90 days

Simple, flexible MDM saves costs

Reltio Connected Data Platform

Velocity Packs for Market Segments

Life Sciences
- HCO/HCP
- Study, site
- Product/IDMP
- …

Healthcare
- Patient
- Provider
- Plan, payer
- …

Financial Services
- Client
- Account
- Household
- …

Insurance
- Customer
- Policy
- Broker
- …

Horizontal (B2B/B2C)
- Customer
- Consumer
- Location
- …

Core Capabilities

- Entity Resolution
- Relationship Management
- Reference Data Management
- Entity 360
- Data Quality
- Data Governance
- Data Integration

Cloud Foundation

Hyperscale Clouds
aws, Azure, Google Cloud Platform

Certified Secure
HIPAA, AICPA SOC, AICPA SOC 2, HITRUST CSF Certified, GDPR, MLPS 2.0 Compliant

Data Sovereignty Regions

Real-time for Low Latency at Scale

High Availability and DR

AI/ML Ready

Our vision: the end to end Data Quality Lifecycle

Harness real-time data quality insights to drive upstream and downstream business impact

Assess

Prior to onboarding data into Reltio platform

- Understand the DQ gaps in your source data
- Compare to industry benchmarks
- Receive recommendations on how to improve DQ

Manage

A unified platform for measuring and managing real-time DQ

- Identify bad data & anomalies
- Reduce manual effort
- Increase effectiveness

Enrich

As Data is flowing through Reltio

- Recommendations for enrichment
- Pre-integrated providers for enrichment
- Measurement of ROI post enrichment

Our vision: Reltio Real-Time Data Quality

Harness real-time data quality insights to drive upstream and downstream business impact

Real-Time, Integrated

Around the clock monitoring with a fully integrated platform

- Real time performance
- Rapid issue detection
- One-click remediation

Industry Benchmarks

Unparalleled industry specific insights

- Monitor global trends
- Define better key performance indicators
- Increase effectiveness

Recommendations

Make informed decisions effortlessly with AI

- Identify bad data
- Detect anomalies
- Reduce manual effort

SUCCESS STORY

Schneider Electric uncovered several million dollars in new sales opportunities

Realized cost savings, sales opportunities, and more effective service operations

BUSINESS CHALLENGES

- Wanted to improve customer experiences and speed resolution of customer issues
- Need to streamline sales and service processes
- Maintaining home-grown MDM solution was unsustainable for one person

SOLUTION

- Unified data across 20 systems comprised of 5M organizations and 13M individuals
- Enriched customer data with integration to Dun & Bradstreet
- Reltio Connected Data Platform as the authoritative source

OUTCOMES

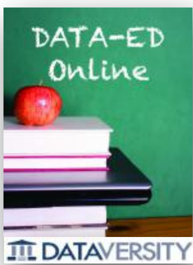
- Several millions of new potential sales opportunities uncovered
- Saves hundreds of thousands in shipping costs per year
- 50% less time to create a new account in operational systems
- Increased efficiency for service and support teams with significantly reduced manual data entry

Thank you

Program Overview

Getting Data Quality Right
Engineering Success Stories

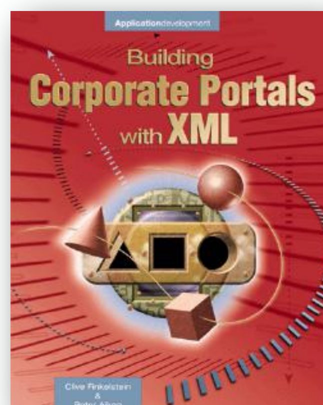
- Approaching Data Quality
  - Cloud considerations
  - Data quality attributes
  - Structural versus practice-related challenges
  - Digitization depends on quality data
  - Definitions
  - Must be built on leverage
  - Data quality examples
  - Causes can be difficult to discern
  - High quality data requires architecture/ engineering

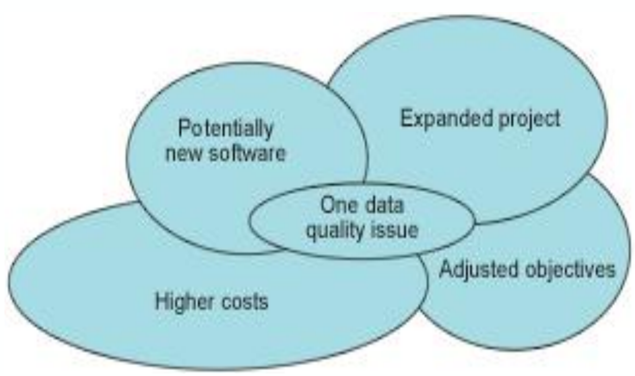


- What do we need to get better at?
  - Systems thinking
  - Not looking at data quality in isolation
  - Developing repeatable capabilities/core data quality expertise
  - PDCA
- How do we get better?
  - Refocus the request around business outcomes
  - Get good at munging
  - Strategy
  - Investment characteristics
  - Conversations
  - Leadership
  - Programatic focus
  - Team development
  - Tangible ROI
- Takeaways and Q&A

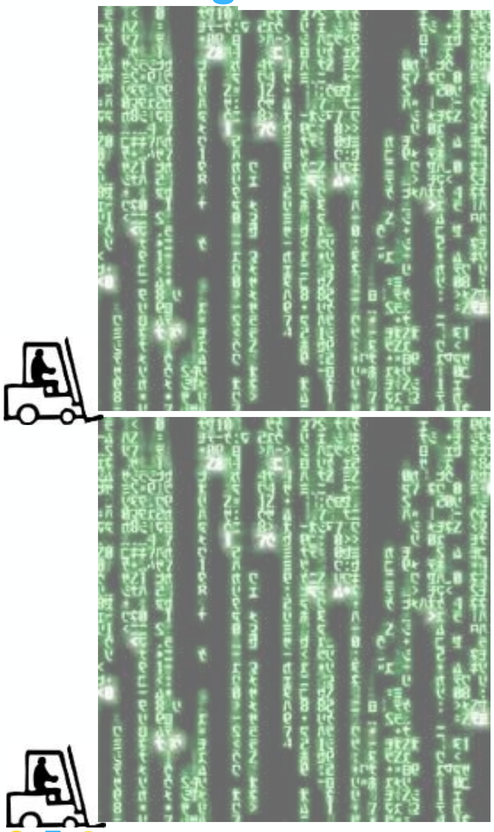
Famous 1990's Words?

- Question:
  - Why haven't organizations taken a more proactive approach to data quality?
- Answer:
  - Fixing data quality problems is not easy
  - It is dangerous -- they'll come after you
  - Your efforts are likely to be misunderstood
  - You could make things worse
  - Now you get to fix it
- A single data quality issue can grow into a significant, unexpected investment





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

Transform

Data Branding

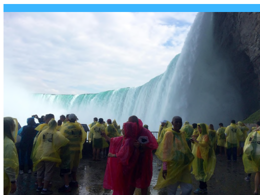
Less

Cleaner

More shareable

... data

Fixing Data in the Cloud Is Like Using a Glovebox

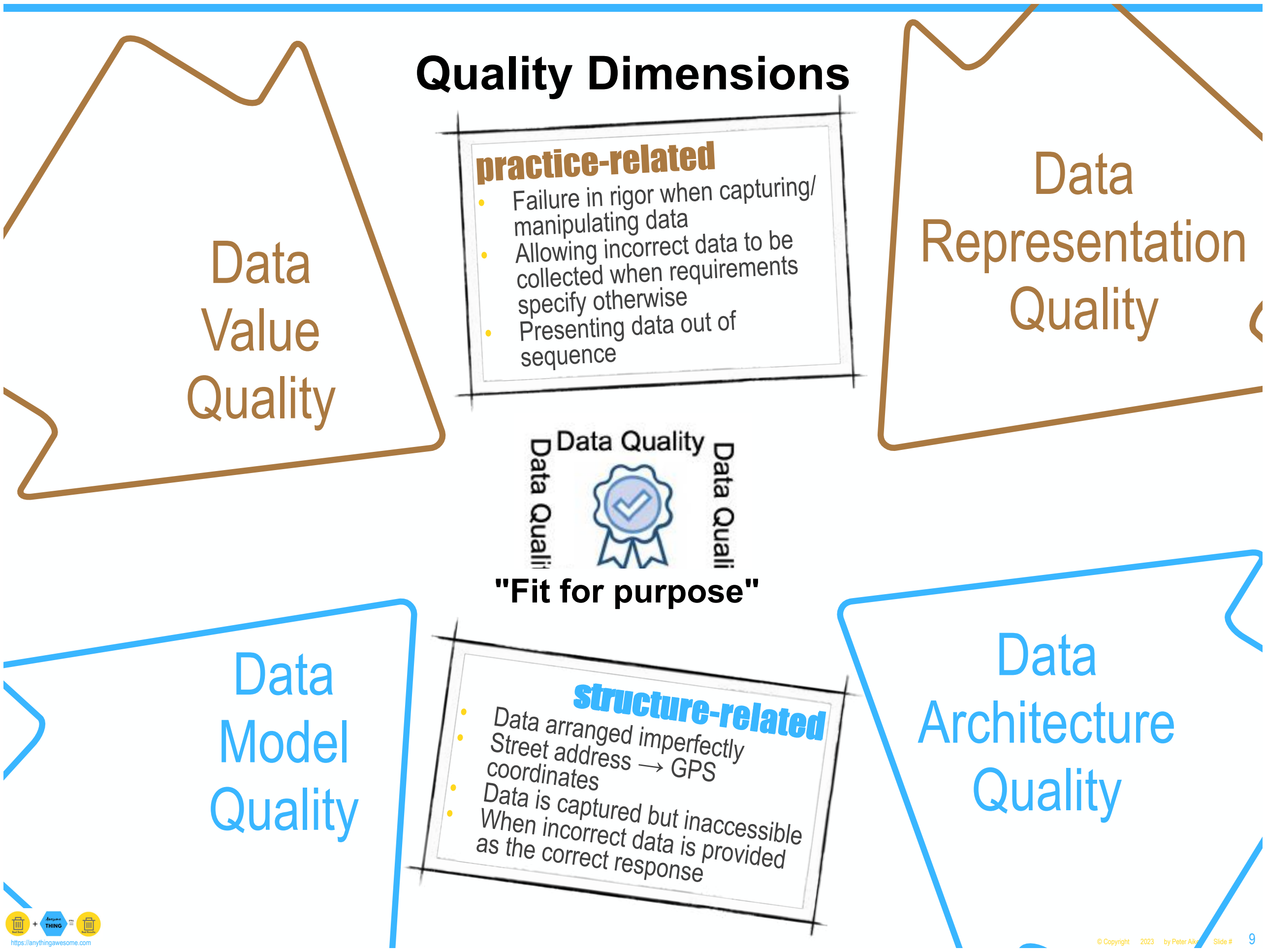


Data Quality Attributes

Tend to be
- **practice-related data quality challenges**
- **many**
- **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 Correctness Timeliness Conciseness Clarity Detail Order Presentation Media Unambiguous	Completeness Correctness Currency Frequency Time Period Precision Reliability Relevance Scope Granularity	Completeness Correctness Conceptual Correctness Conceptual Completeness Syntactic Correctness Syntactic Completeness	Completeness Correctness Enterprise Model Utility Data Management Quality Data Sharing Ability Data Engineering Quality Data Operation Quality Data Evolvability Data Self Awareness


Quality Dimensions
Data Value 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 Representation Quality
Data Quality
"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
https://anythingawesome.com
© Copyright 2023 by Peter Aiken Slide # 9


Right Knee
Day of
Surgery
8/12/14
← Data Quality Remediation
https://anythingawesome.com
© Copyright 2023 by Peter Aiken Slide # 10

New York Turns to Data To Solve Big Tree Problem

- NYC
  - 2,500,000 trees
- 11-months from 2009 to 2010
  - 4 people were killed or seriously injured by falling tree limbs in Central Park alone
- 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

https://anythingawesome.com

http://www.computerworld.com/s/article/9239793/New_York_Turns_to_Big_Data_to_Solve_Big_Tree_Problem?source=CTWNLE_nlt_datamgmt_2013-06-05

NYC's Big Tree Problem

- Question
  - Does pruning trees in one year reduce the number of hazardous tree conditions in the following year?
- Lots of data but granularity challenges
  - Pruning data recorded block by block
  - Cleanup data recorded at the 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

https://anythingawesome.com

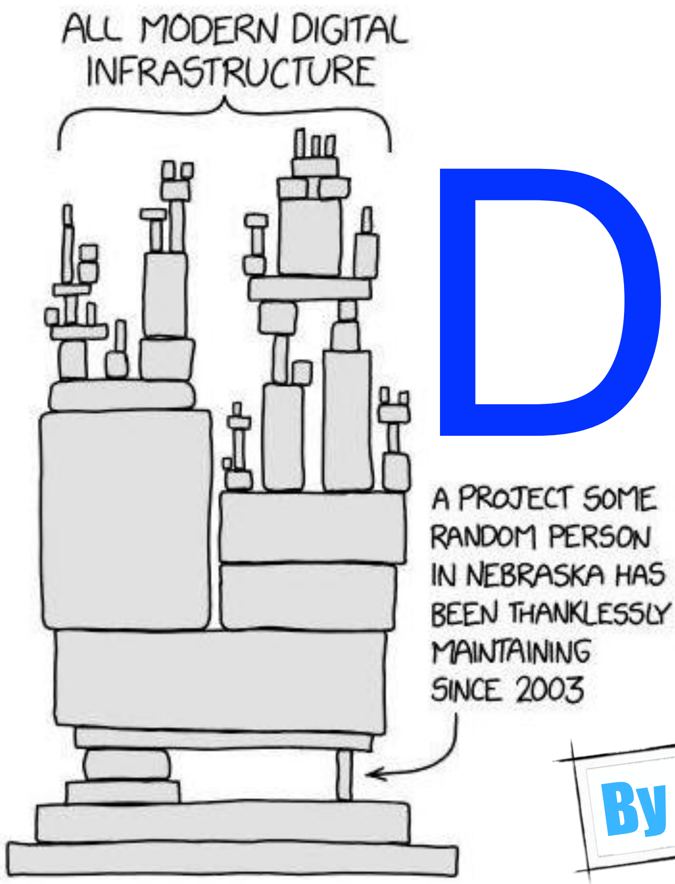
http://www.computerworld.com/s/article/9239793/New_York_Turns_to_Big_Data_to_Solve_Big_Tree_Problem?source=CTWNLE_nlt_datamgmt_2013-06-05

Who Is Joan Smith?

http://www.dataflux.com

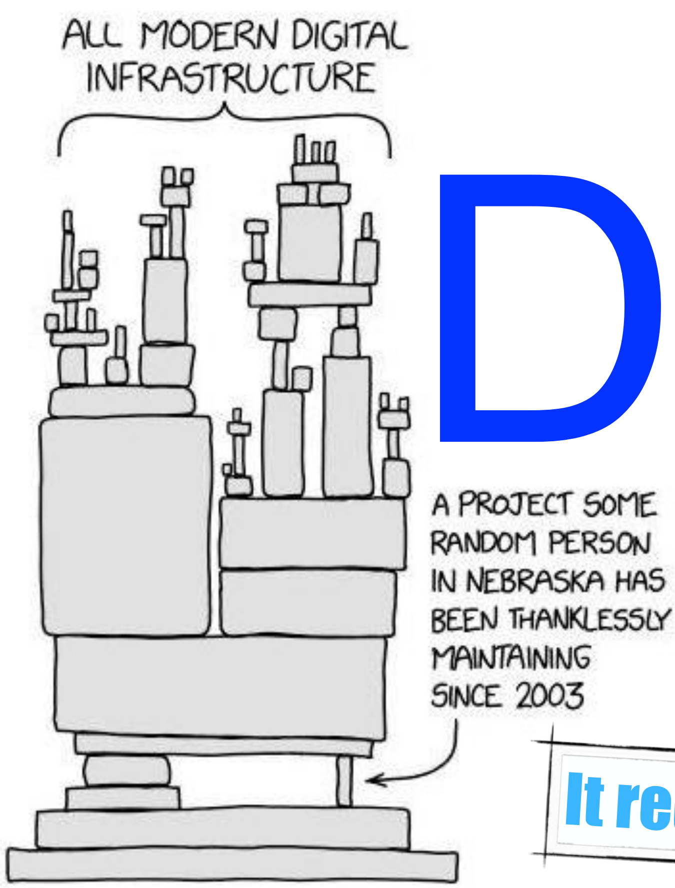
Digital

It isn't possible to go digital



Dat

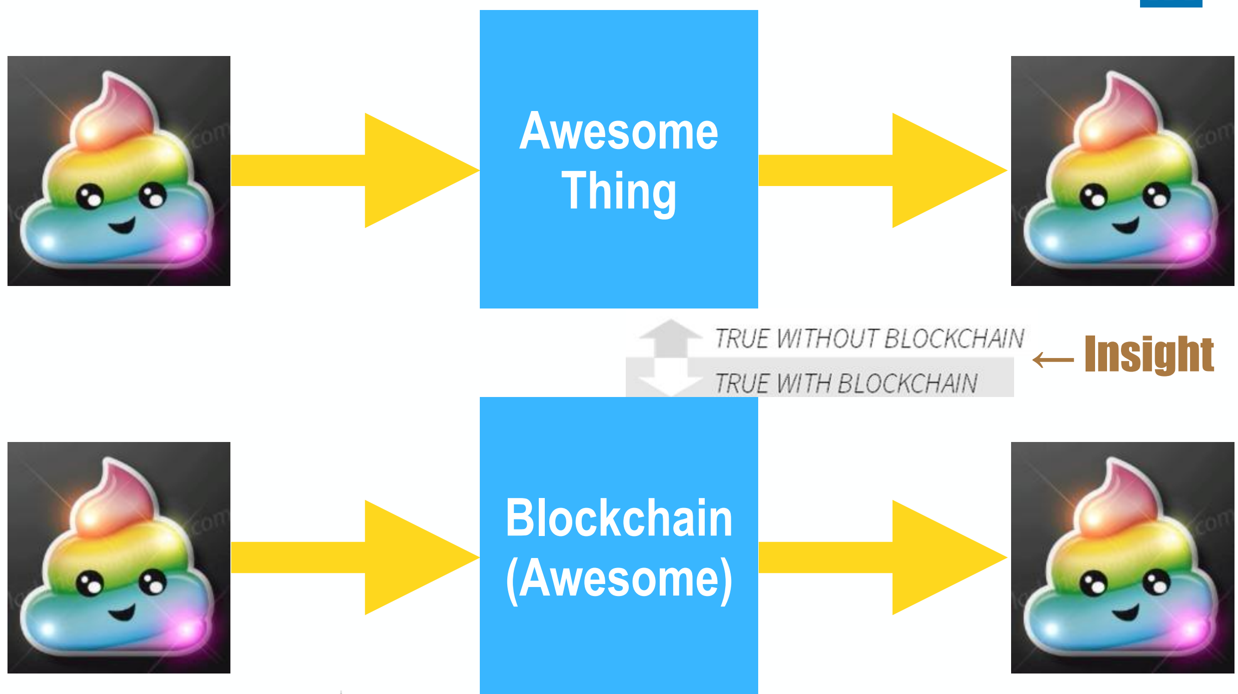
By just spelling 'data'



Data

It requires more work!

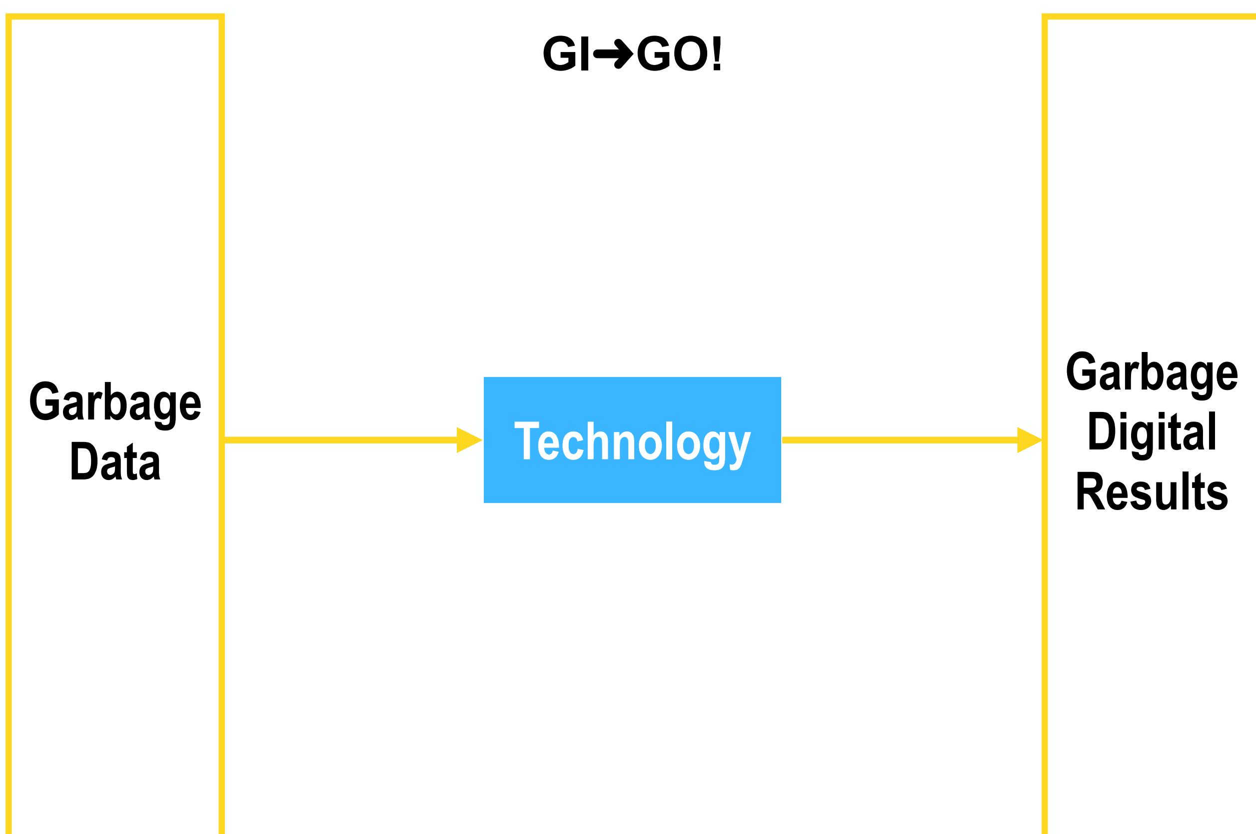
"I've Just Had a Recent Technology Realization"

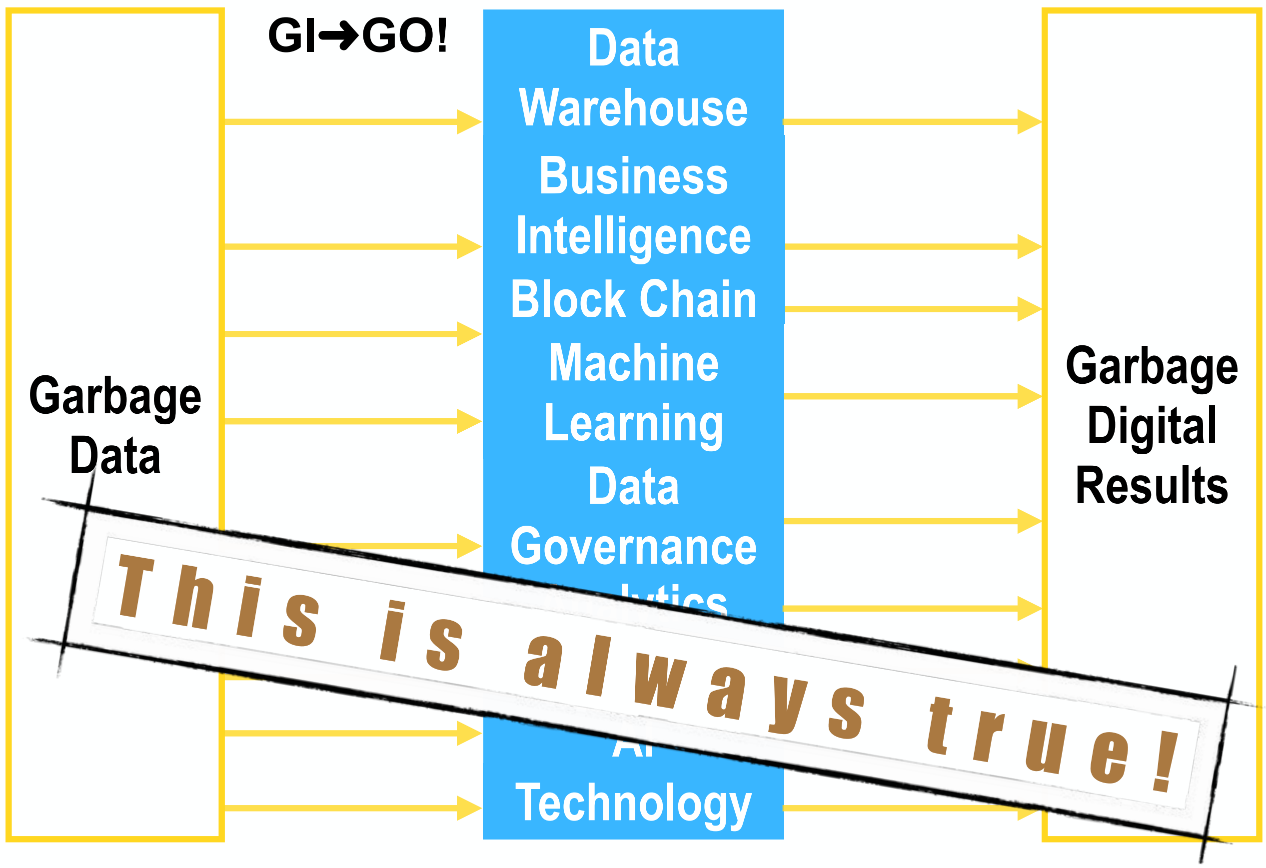


Garbage In → Garbage Out!

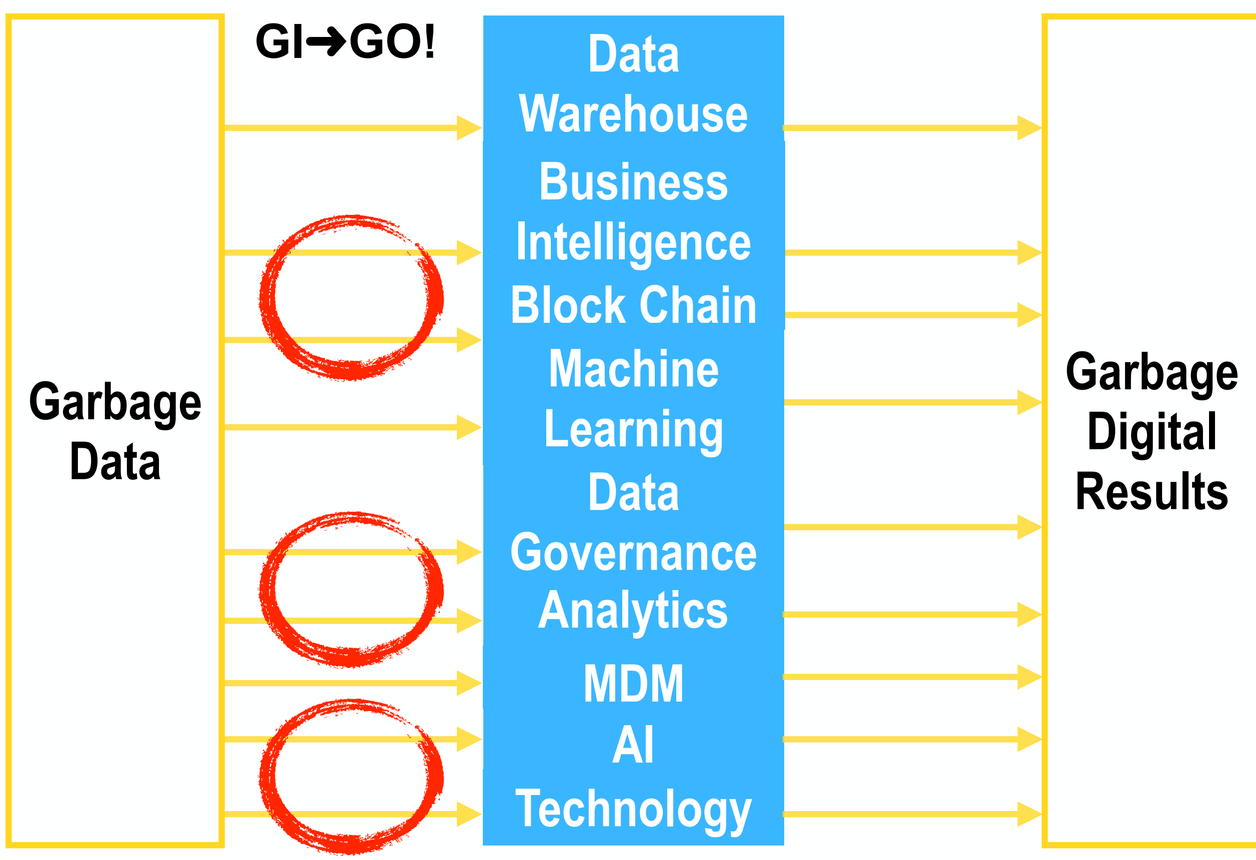
https://www.flashingblinkylights.com/unicorn-poop-emoji-sparkling-led-pins.html

GI➔GO!

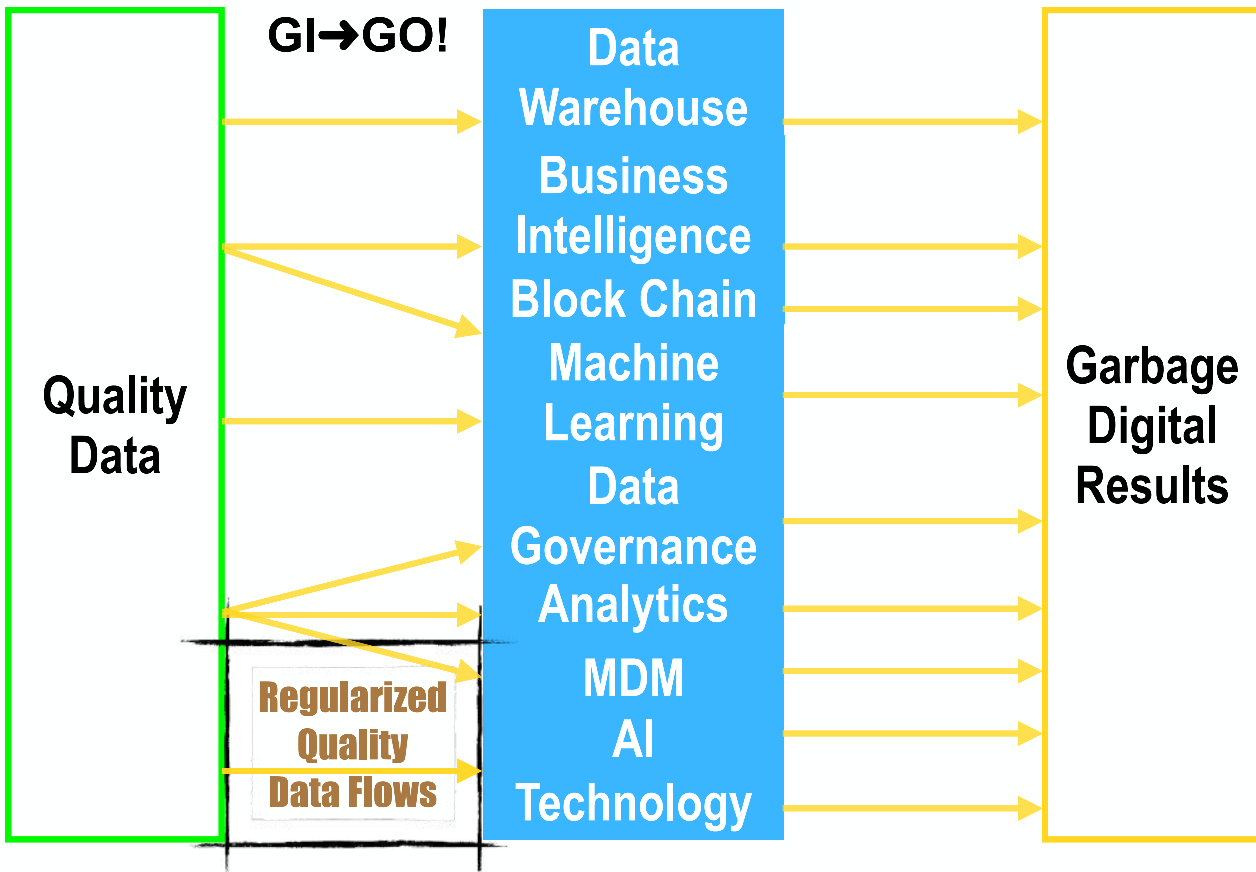


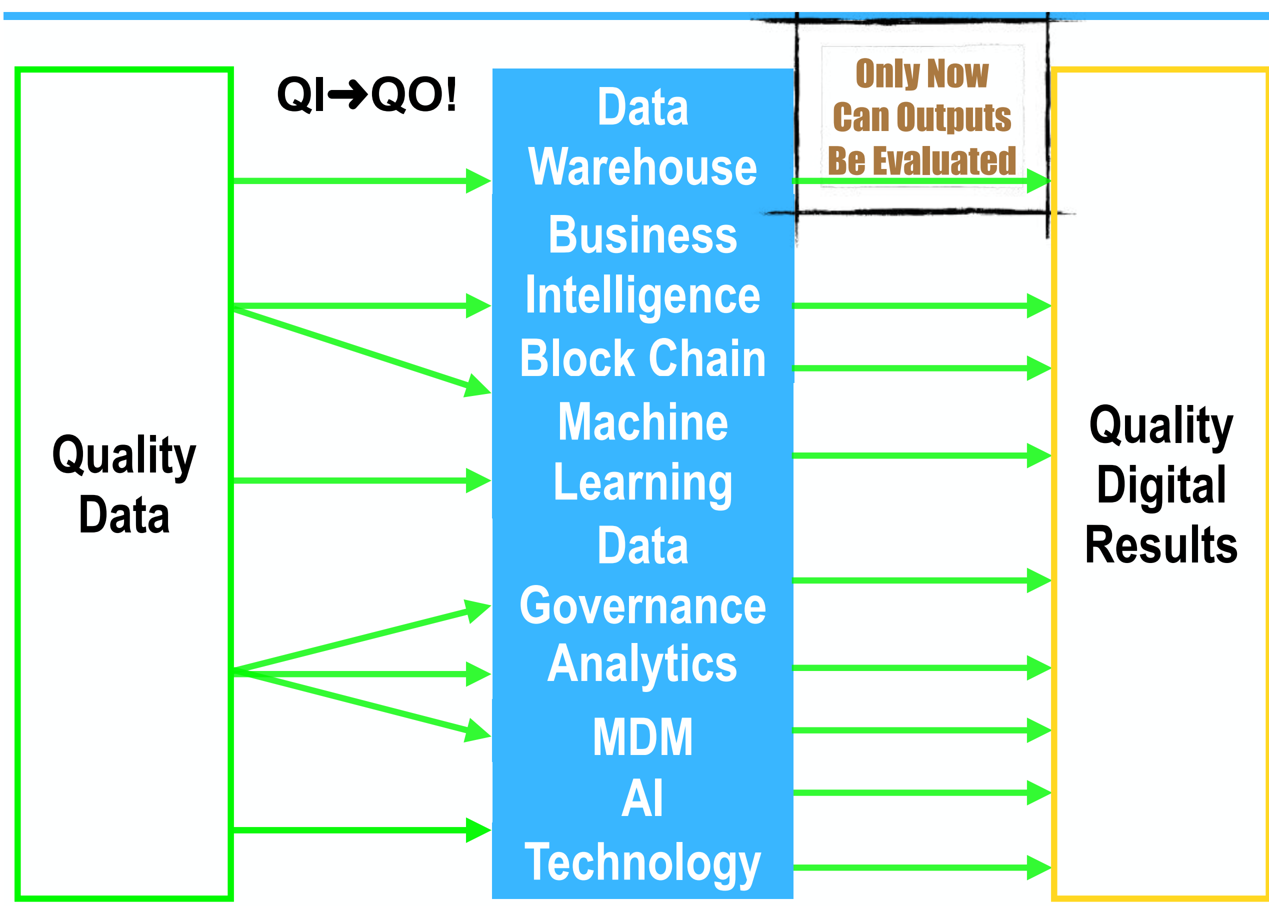


https://anythingawesome.com

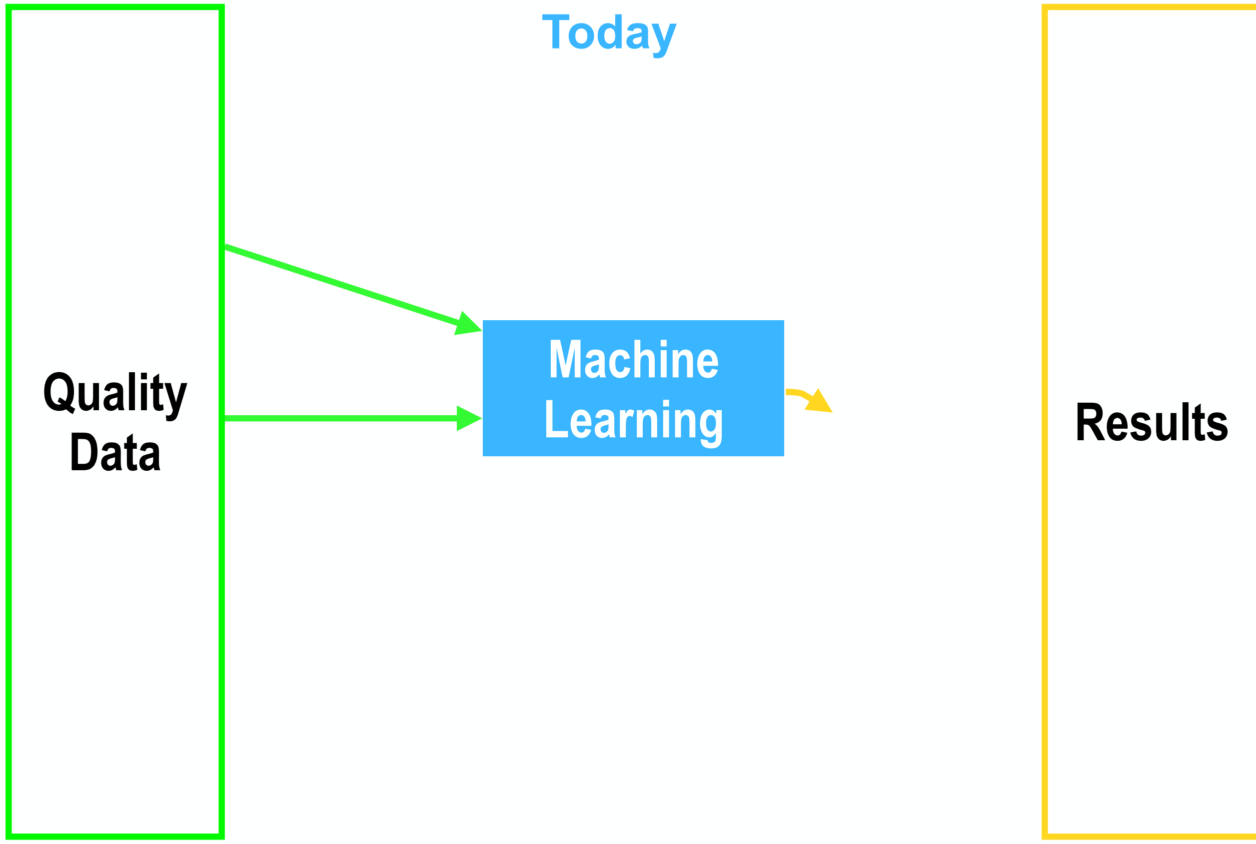


https://anythingawesome.com


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

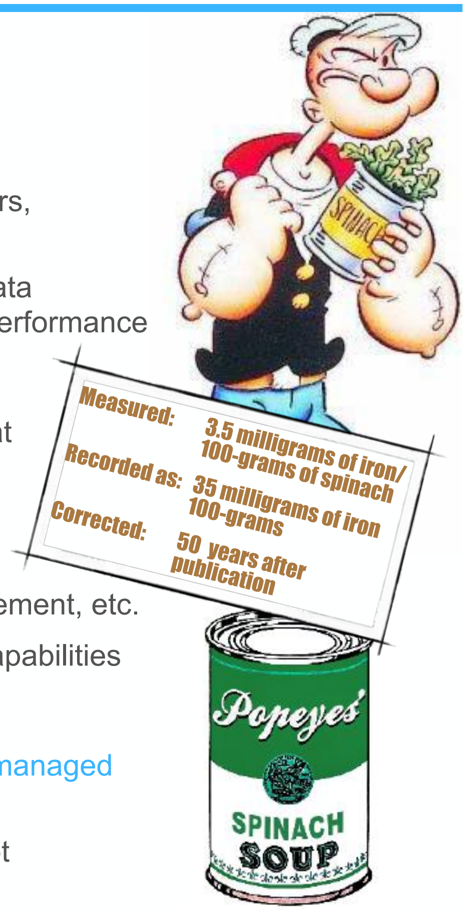

QI➔QO!
Quality Data
Data Warehouse
Business Intelligence
Block Chain
Machine Learning
Data Governance
Analytics
MDM
AI
Technology
Only Now Can Outputs Be Evaluated
Quality Digital Results

Today



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!



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

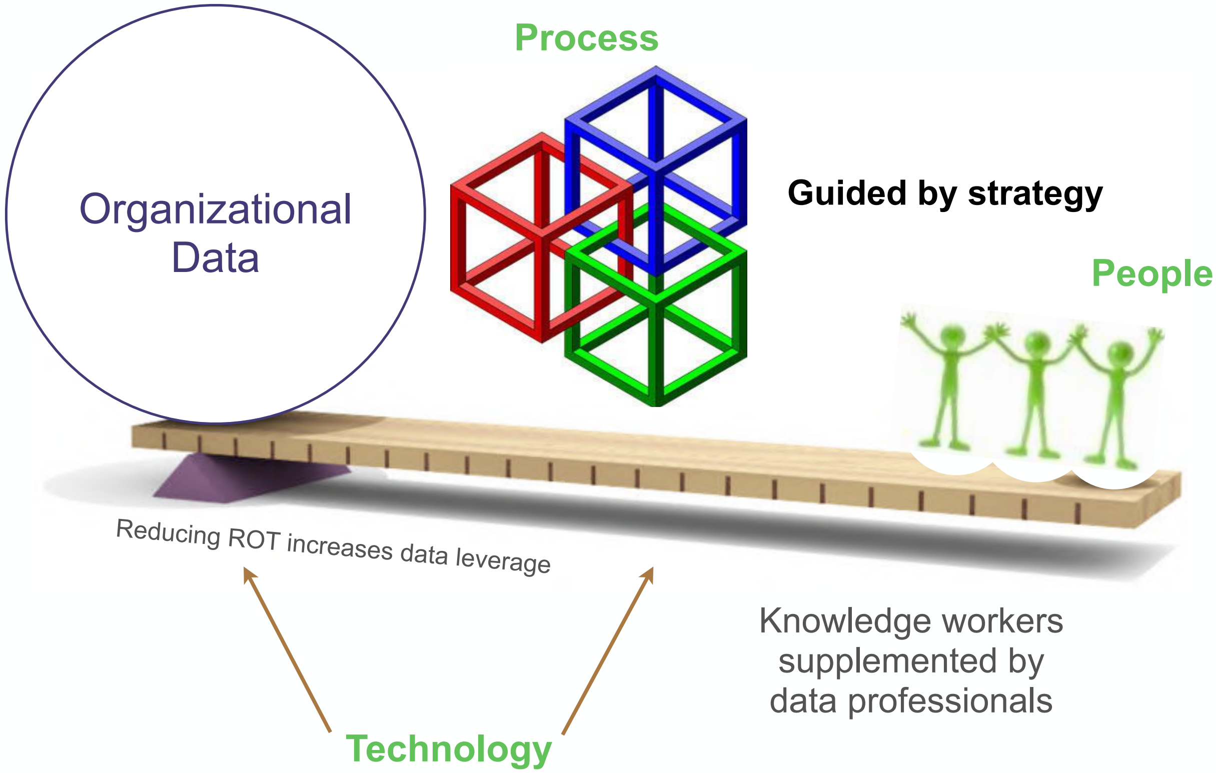
The Princess on the Pea

by
Hans Christian Andersen

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

Data Leverage Is a Multi-Use Concept

- Permits organizations to better manage their data
  - Within the organization, and
  - With organizational data exchange partners
  - In support of the organizational mission
- Leverage
  - Obtained by implementation of data-centric technologies, processes, and human skill sets
  - Focus on the non-ROT data
    - The bigger the organization, the greater potential leverage exists
- Treating data more asset-like simultaneously
  - Lowers organizational IT costs and
  - Increases organizational knowledge worker productivity

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

Events Are Not Always Recognized as Data Quality Challenges?

- IRS-coronavirus payments
- A letter from a bank
- A very expensive, very small data rounding error
- Health data story
- The chocolate story
- Covid-19

SLATE

Search Sign In

News & Politics Culture Technology Business Human Interest

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.

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."

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://anythingawesome.com

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

P.O. Box 247046
Omaha, NE 68124-7046

The SunTrust VISA® Gift Card
Your Gift. Their Choice.

SUNTRUST
www.suntrust.com/giftcard

GIFT CARD NUMBER
4145750100091592

INSTITUTE FOR DATA RESEARCH
501 E FRANKLIN ST STE414
RICHMOND VA 23219-2330

0239036895253015

0000015 0319 0000015 7520 0010 0001 000 0G15 001

We hope you enjoy your SunTrust VISA® gift card. Your gift card value is:

$0

Check your gift card balance online at www.suntrust.com/giftcard or by phone at 1-800-318-0210.

a gift for you a gift for you a gift for you a gift for you

Congratulations! You have just received a prepaid gift card that can be used everywhere the VISA® card is accepted in the United States. Use it at any retail store, restaurant, gas station and grocer. Or, enjoy it to buy books, music as well as go to the movies or a concert. This is the hassle-free gift that fits you perfectly!

Activate - Go online at www.suntrust.com/giftcard or call 1-800-318-0210 and enter the last 4-digits of the phone number provided by the purchaser of this card.

Salutate - Sign your card before using.

Celebrate - Get what you have always wanted.

Important Information about your SunTrust Visa® Gift Card:

- The SunTrust VISA® Gift Card is welcomed at all merchant locations wherever VISA® debit cards are accepted. Restrictions do apply.*
- Your gift card is valid for at least one (1) year after the date of purchase or until the Card balance is zero, whichever occurs first. The expiration date is shown on your Card.
- For general incuires and to check your balances go to www.suntrust.com/giftcard or contact us at 1 800 318-0210.
- If your card is lost or stolen, it will be replaced with the remaining balance less a $5 replacement fee. To report lost/stolen cards contact us at 1-800-318-0210.
- Use your card soon! The service fee for the card is waived for the first six months. A $2.50 fee will be deducted from the available balance each month thereafter.

Please see the reverse side for frequently asked questions

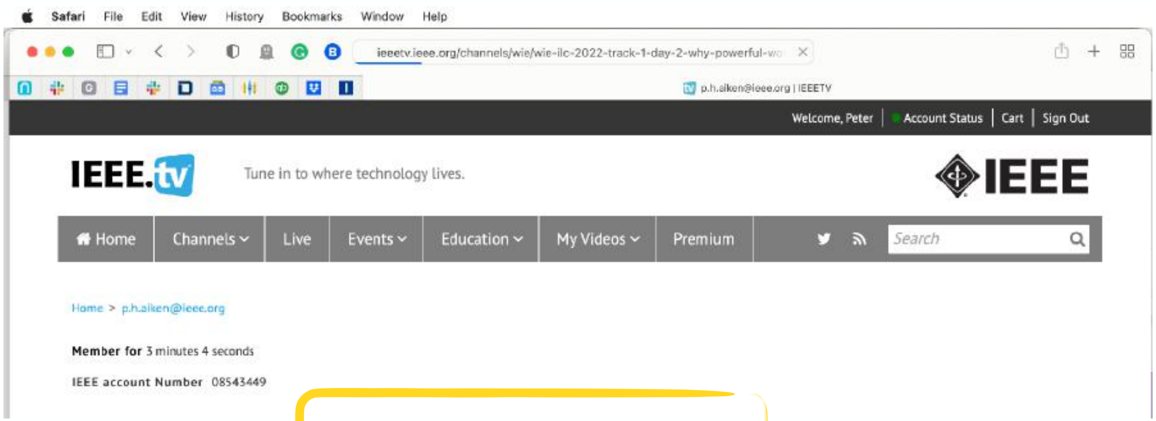
A congratulations letter from another bank

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

https://anythingawesome.com

IEEE Senior Member Status (30+ Years Membership)



The Seattle Times

Port of Seattle

Wednesday, August 5, 2009 - Page updated at 12:00 AM

Permission to reprint or copy this article or photo, other than personal use, must be obtained from The Seattle Times. Call 206-464-3113 or e-mail resale@seattletimes.com with your request.

Small construction mistake at Port of Seattle may cost $1 million

By Bob Young
Seattle Times staff reporter

A small mistake at the Port of Seattle is going to cost a lot, perhaps about $1 million.

DANIEL HOUGHTON / THE SEATTLE TIMES
Tenant SSA Marine took occupancy Monday at Terminal 30, two months later than planned because a trench was built too narrow for the cranes' electrical cable.

- Needed trench for electrical cable 2.52" - delivered 2.5"
- $1M required to rent other facilities while new cable is obtained
- Either rounding or truncation could explain
  - We need to get a summary on all of this," he said. "How did the mistake occur? Who's at fault? What are the damages? And how is money going to be recovered?"

2.52" → 2.5"

Why Britain has 17,000 pregnant men


The Washington Post | Ezra Klein's

WONKBLOG

No free lunch in GOP budget | Senate Dems inching towards | Where does coal go if we don't burn | Is industrial poli back?

Economy | Health Care | Environment | 2012 | Interviews | Graphs | Authors | Fu

Comments

Why Britain has 17,000 pregnant men

Posted by Sarah Kliff at 02:00 PM ET, 04/07/2012

Text Size | Print | E-mail | Reprints | Share | More >

(Waltraud Grubitsch - AFP/Getty Images)

The data seemed, at first glance, like it could be indicative of a medical miracle. Between 2009 and 2010, thousands of British men turned up at hospitals to be treated for many pregnancy-related services, things like obstetric exams and midwife services. All told, there were 17,000 of them.

This research, published as a letter this week in the British Medical Journal, was meant to draw attention to how much data gets entered incorrectly in the country's medical system. These guys weren't turning up at the doctor for pregnancy-related services. Instead, they were at their doctor for procedures that had medical codes similar to those of midwifery and obstetric services. With a misplaced keystroke here or there, an annual physical could become a consultation with a midwife.

Why Using Microsoft's Tool Caused Covid-19 Results To Be Lost

BBC | Home | News | Sport | Reel | More

NEWS ☰ Menu

Excel: Why using Microsoft's tool caused Covid-19 results to be lost

By Leo Kelion
Technology desk editor

5 October

https://www.bbc.com/news/technology-54423988?es_p=12801491

Why Using Microsoft's Tool Caused Covid-19 Results To Be Lost

BBC | Home | News | Sport | Reel | More

NEWS ☰ Menu

Tech

Excel: Why using Microsoft's tool caused Covid-19 results to be lost

By Leo Kelion
Technology desk editor

5 October

- Since 2007 should have been forced to use .xlsx (1,000,000+ rows)
- Used .xls (65,000 rows)
- Additional data was dropped without notification

Under-reported figures

From 25 Sept to 2 Oct

50,786

Cases initially reported by PHE

15,841

Unreported cases, missed due to IT error

8 days of incomplete data

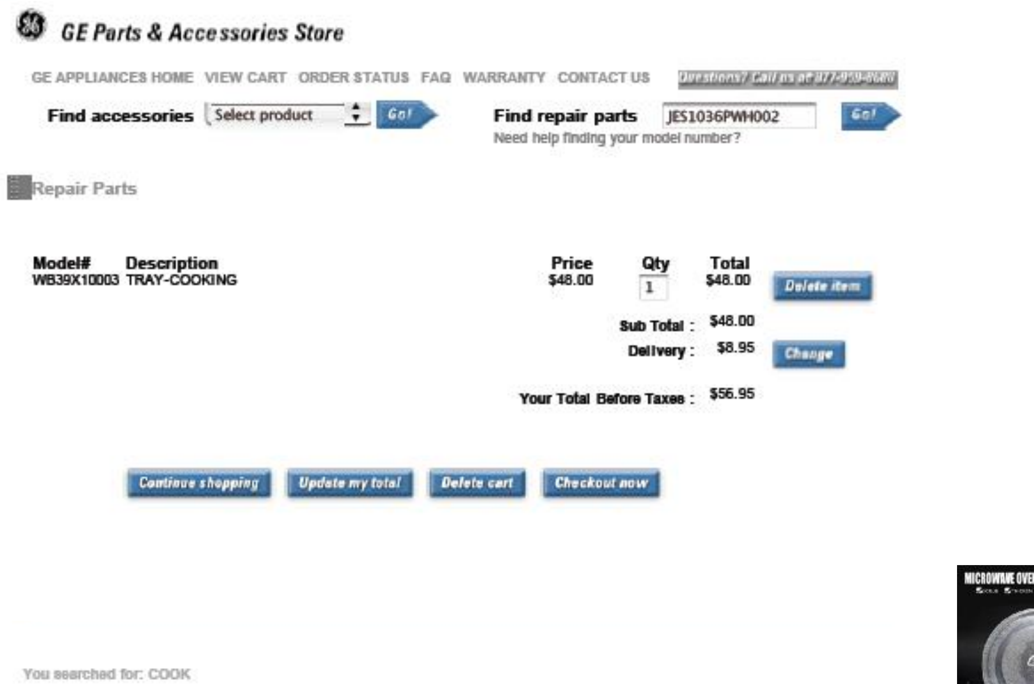
1,980 cases per day, on average, were missed in that time

48 hours Ideal time limit for tracing contacts after positive test

https://www.bbc.com/news/technology-54423988?es_p=12801491

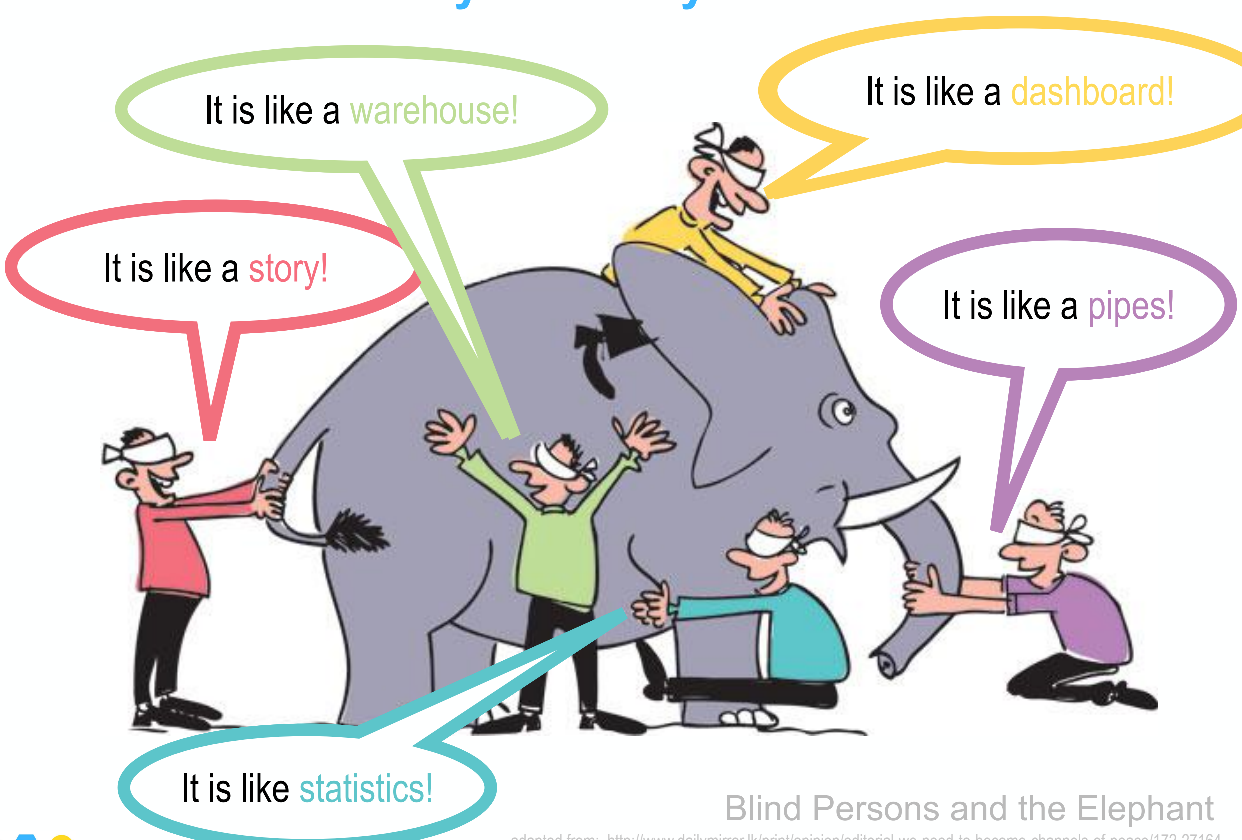
Source: PHE and gov.uk

How To Solve This Data Quality Problem Using Just Tools?



Microwave Oven retail price was $40

Data Is Not Broadly or Widely Understood



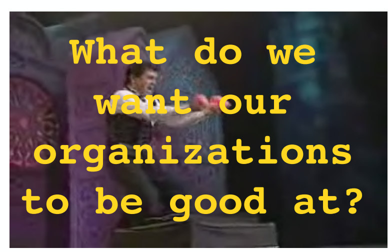
Blind Persons and the Elephant

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

Many DQ Challenges Are Unique and/or Context Specific!

Data Knowledge Is Too Little and Too Informal

- Data management happens 'pretty well' at the workgroup level
  - Defining characteristic of a workgroup
  - Without guidance (strategy), what are the chances that all workgroups are pulling toward the same objectives?
  - Consider the time spent attempting informal practices
- Data chaff becomes sand in the machinery
  - Preventing smooth interoperation and exchanges
  - Losses due to lots of little data cuts have been difficult to account for
- Organizations and individuals lack data quality
  - Knowledge
  - Skills
  - Data Engineering (*How?*)
  - Data Strategy (*Why this as opposed to that?*)



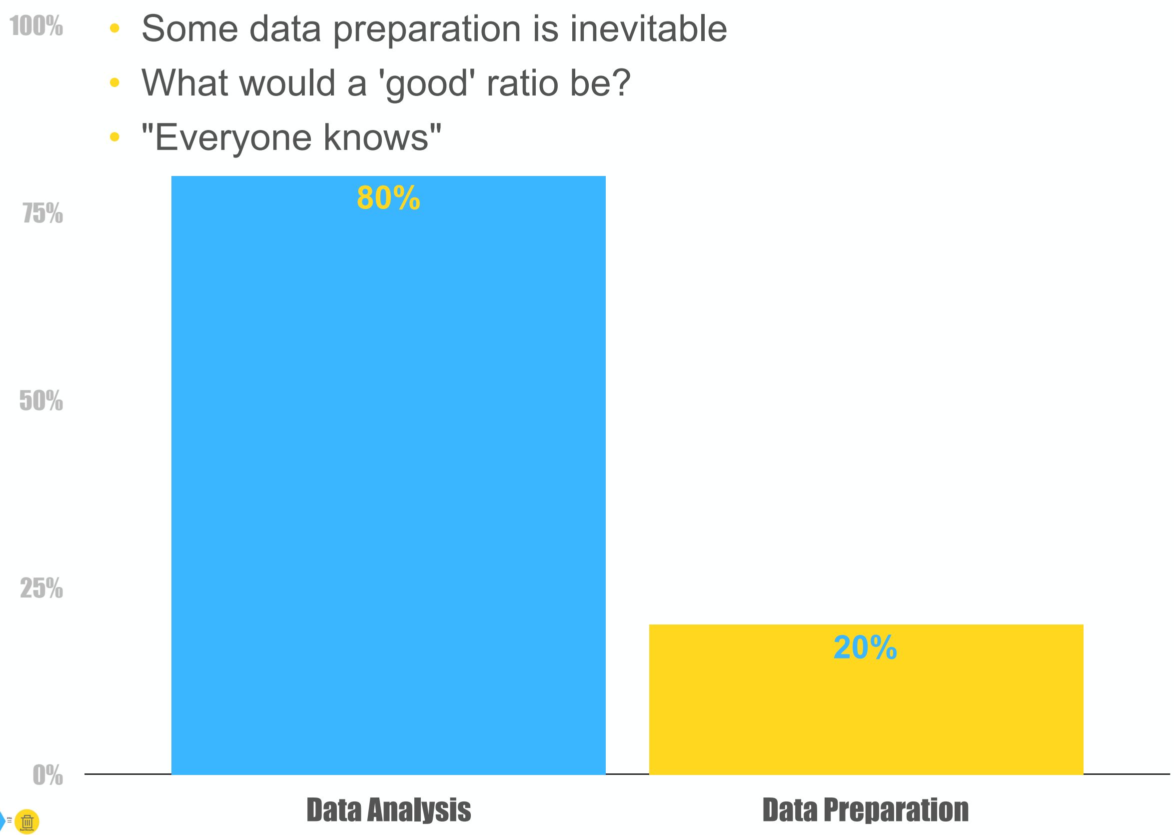
Wally Easton Playing Piano
https://www.youtube.com/watch?v=NNbPxSvlI-Q

DQ Effort Pattern

- Change date order field
- Sequencing/ordering information
- Correcting factual errors
- ...

Everyone Wants To Do Better Data Analysis …

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



All organizational challenges have data roots

Burning Bridge

- 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

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)

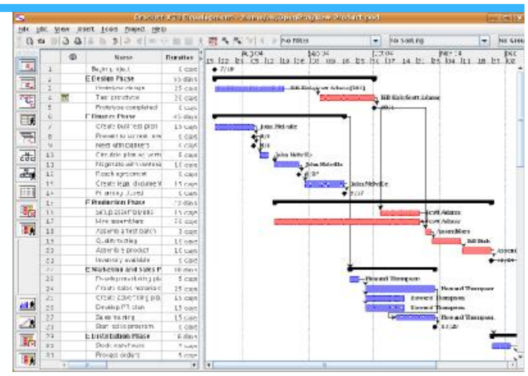
Differences Between Programs and Projects

- Programs are Ongoing, Projects End
  - Managing a program involves long term strategic planning and continuous process improvement is not required of a project
- Programs are Tied to the Financial Calendar
  - Program managers are often responsible for delivering results tied to the organization's financial calendar
- Program Management is Governance Intensive
  - Programs are governed by a senior board that provides direction, oversight, and control while projects tend to be less governance-intensive
- Programs Have Greater Scope of Financial Management
  - Projects typic ... ight forward budget and project financial managemen ... program planning, managemen
- Program Ch ... Capability
  - Projects employ a formal change ... level, change management requires executive leadership ... change is driven more by an organization's strategy and is subject to market conditions and changing business goals

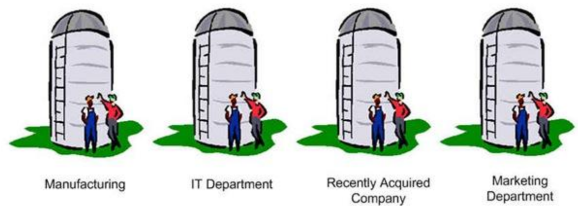
Your data quality program must last at least as long as your HR program!

Adapted from http://top.idownloadnew.com/program_vs_project/ and http://management.simplicable.com/management/new/program-management-vs-project-management

Data Quality Is Not a Project



- Durable asset
  - An asset that has a usable life more than one year
- Reasonable project deliverables
  - 90 day increments
  - Data evolution is measured in years
- Data
  - Evolves - it is not created
  - Significantly more stable



- Readymade data architectural components
  - Prerequisite to agile development
- Only alternative is to create additional data siloes!

What Does It Mean "Data Quality Program"?

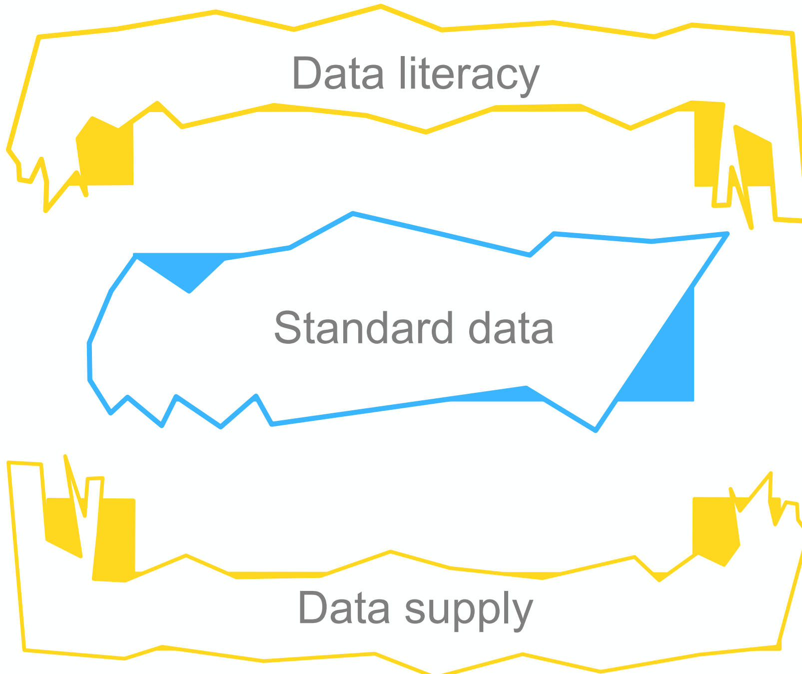
- Ongoing commitment
  - Permits evolutionary improvement of the approach
- Governance
  - Senior level coordination, direction, and control
- Executive leadership capabilities
  - Change and risk management
- Data quality approach inherits (above)
  - Budget, strategic priorities
  - Senior level attention and improving topical facility
  - Reasonable timelines/expectations

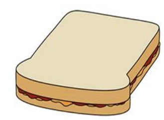
https://blog.ducenit.com/data-quality-management

NYC Bridge Maintenance

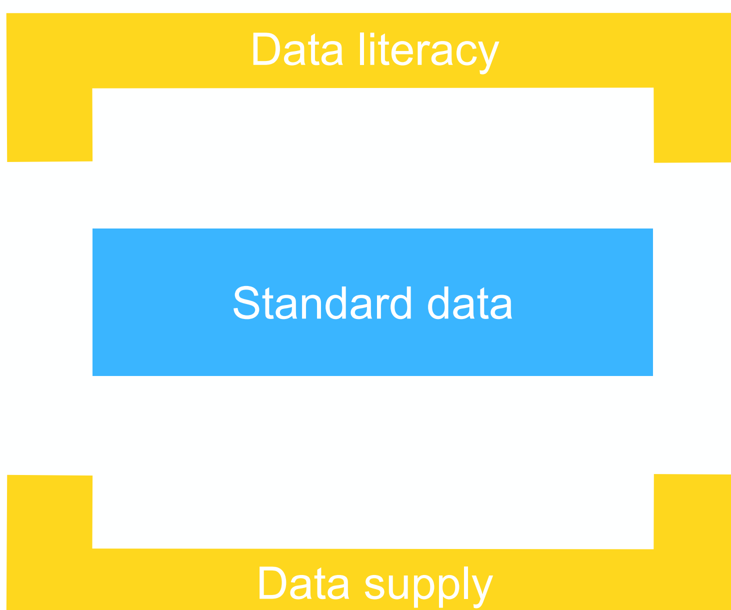
- Painting is continuous and optimized
- As is much of the remaining maintenance
- Ensures a knowledgeable workforce

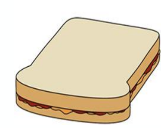
Leverage Point - High Performance Automation





Leverage Point - High Performance Automation





Leverage Point - High Performance Automation

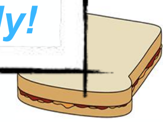
This cannot happen without engineering and architecture!

Data literacy

Standard data

Data supply

Quality engineering/ architecture work products do not happen accidentally!



Leverage Point - High Performance Automation

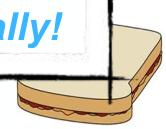
This cannot happen without data engineering and architecture!

Data literacy

Standard data

Data supply

Quality data engineering/ architecture work products do not happen accidentally!



Bonfire Collapse

http://www.tamu.edu/bonfire-commission/reports/Packer_Report.pdf

Data Quality Engineering

Program Overview

Getting Data Quality Right
Engineering Success Stories

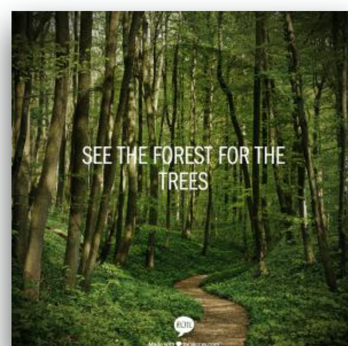
- Approaching Data Quality
  - Cloud considerations
  - Data quality attributes
  - Structural versus practice-related challenges
  - Digitization depends on quality data
  - Definitions
  - Must be built on leverage
  - Data quality examples
  - Causes can be difficult to discern
  - High quality data requires architecture/ engineering

- What do we need to get better at?
  - Systems thinking
  - Not looking at data quality in isolation
  - Developing repeatable capabilities/core data quality expertise
  - PDCA
- How do we get better?
  - Refocus the request around business outcomes
  - Get good at munging
  - Strategy
  - Investment characteristics
  - Conversations
  - Leadership
  - Programatic focus
  - Team development
  - Tangible ROI
- Takeaways and Q&A

Systems Thinking

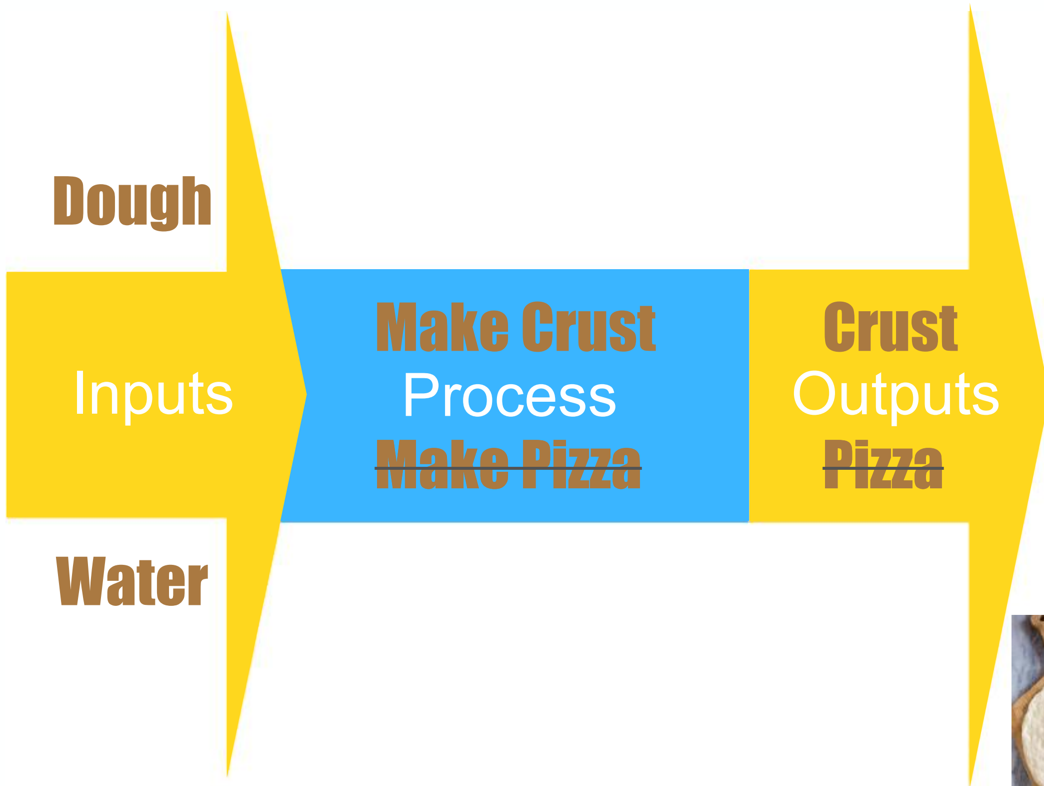
- A framework that is based on the belief that the component parts of a system can best be understood in the context of relationships with other systems, rather than in isolation.
- The only way to fully understand why a problem or element occurs and persists is to understand the part in relation to 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

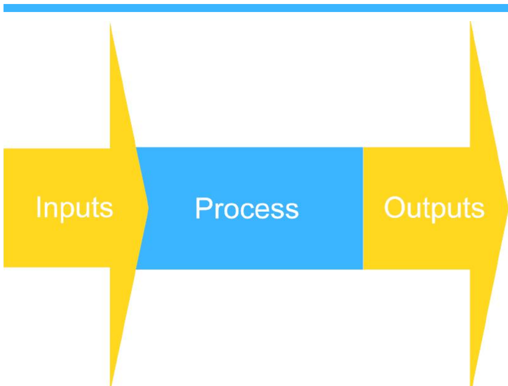


http://victorianscandal.wordpress.com/picturesque/rachel-olshausen/

Input ➜ Process ➜ Output Diagram



Data Steward Quality Responsibilities

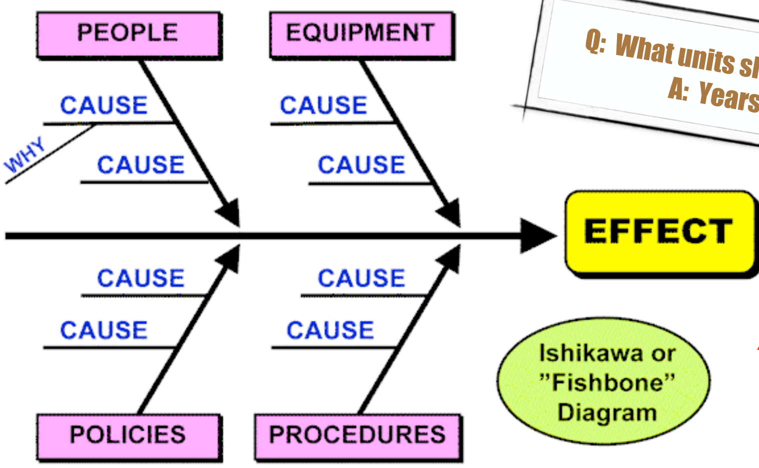


- Inputs
  - From where, do each of these my responsible data items come?
  - Why are they produced?
  - *What level of quality is required by 'my processes?'*
- Process
  - What business processes use the data within my fiduciary responsibility?
  - For what business purpose do they use each data item?
  - *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?*

How long will these challenges take to correct?

- *Why is infant mortality so high?*
  Malnourished mothers
- *Why are mothers malnourished?*
  Substandard biology educations in high school
- *Why do are biology programs substandard?*
  Poor education of high school biology teachers
- *Why do we have poor biology teacher education?*
  Biology profession unaware of consequences

Q: What units should be used to measure progress?
A: Years minimally - likely decades!



Asking "why" repeatedly!

Interdependencies

Data Governance

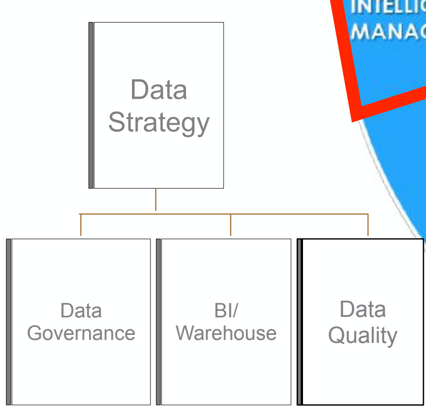
Data Quality

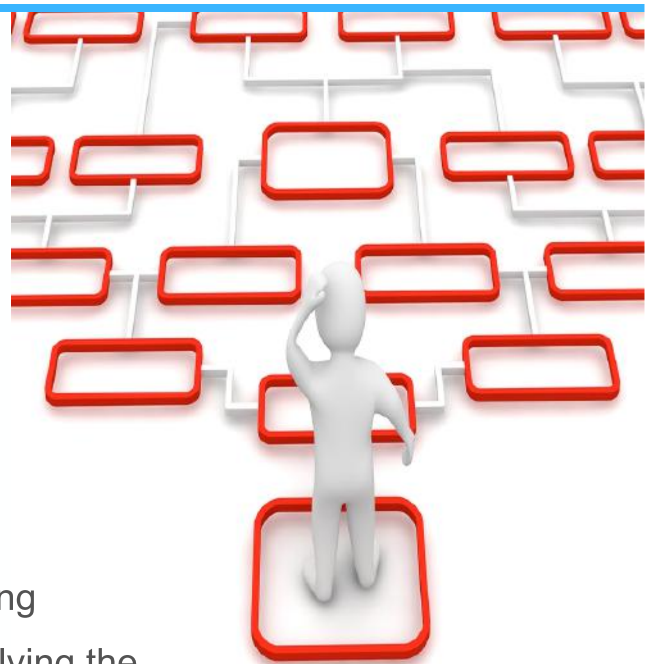
CRM (SalesForce)

Data Management Body of Knowledge (DM BoK V2)

Practice Areas

Perfecting operations in 3 data management practice areas





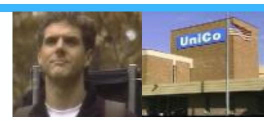
Programmatic Data Quality Engineering

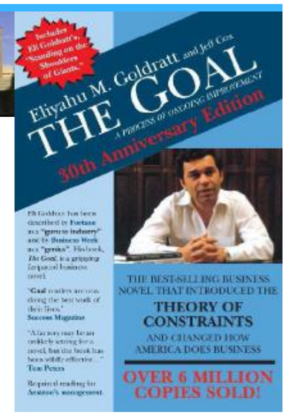
1. Allow the form of the problem to guide the form of the solution
2. Provide a means of decomposing the problem
3. Feature a variety of tools simplifying system understanding
4. Offer a set of strategies for evolving the design of a programmatic solution
5. Provide criteria for evaluating the quality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

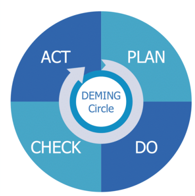
https://anythingawesome.com

Theory of constraints (TOC)

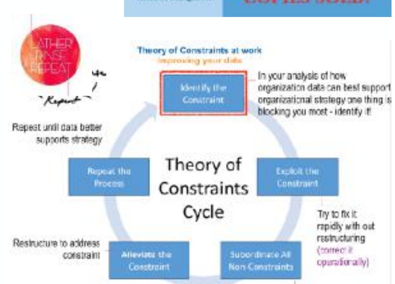
From Wikipedia, the free encyclopedia

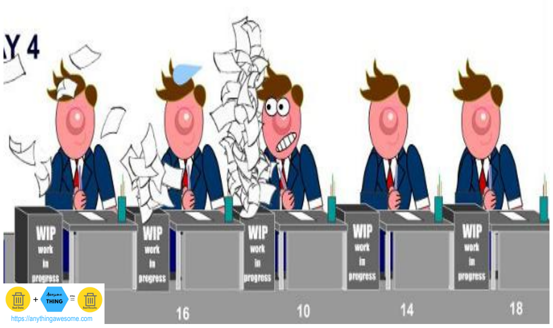






- A management paradigm that views any manageable system as being limited in achieving more of its goals by a small number of constraints (Eliyahu M. Goldratt)
- There is always at least one constraint, and TOC uses a focusing process to identify the constraint and restructure the rest of the organization to address it





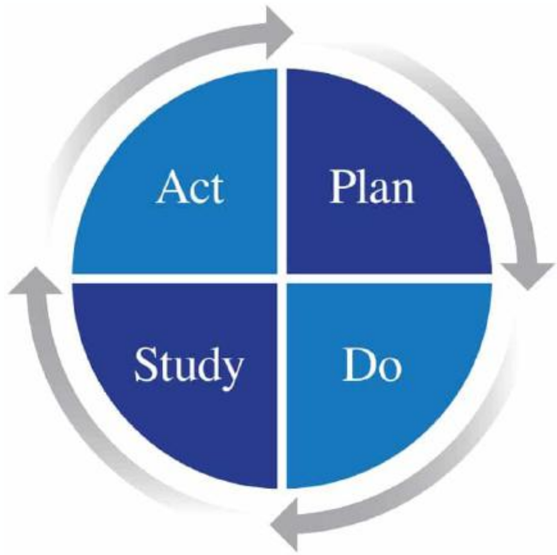
- TOC adopts the common idiom "a chain is no stronger than its weakest link," processes, organizations, etc., are vulnerable because the weakest component can damage or break them or at least adversely affect the outcome

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

https://anythingawesome.com

The DQE Cycle

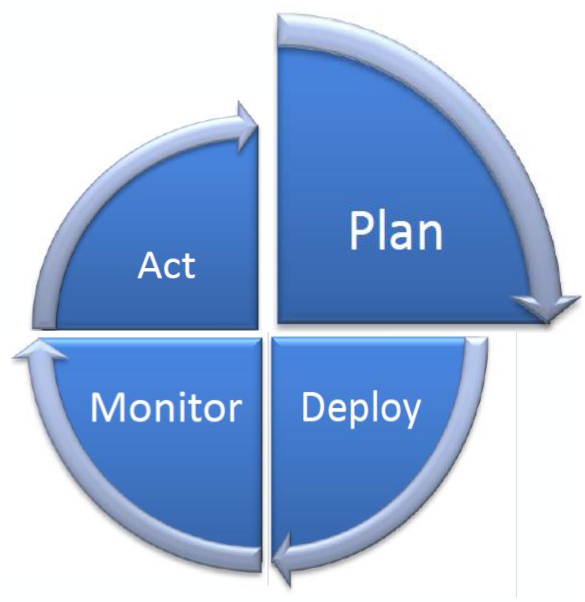
- Deming cycle
- "Plan-do-study-act" or "plan-do-check-act"
  - Identifying data issues that are critical to the achievement of business objectives
  - Defining business requirements for data quality
  - Identifying key data quality dimensions
  - Defining business rules critical to ensuring high quality data



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

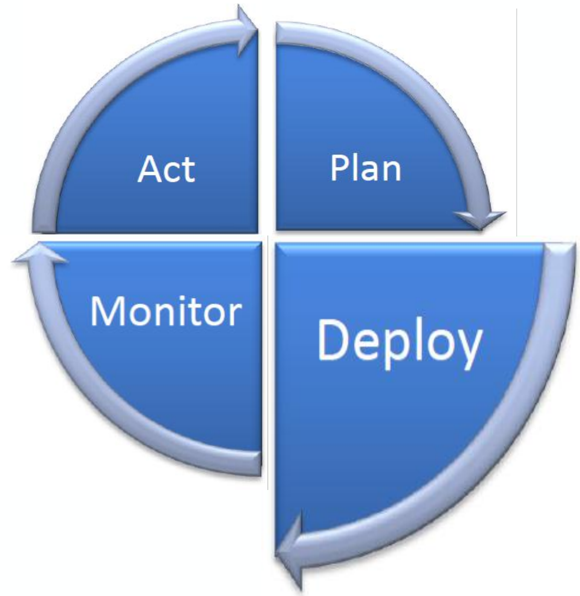
The DQE Cycle: (1) Plan

- Plan for the assessment of the current state and identification of key metrics for measuring quality
- The data quality engineering team assesses the scope of known issues
  - Determining cost and impact
  - Evaluating alternatives for addressing them



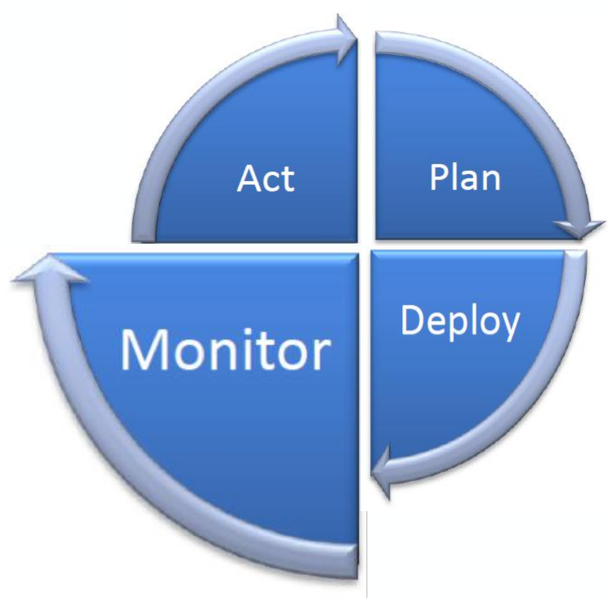
The DQE Cycle: (2) Deploy

- 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



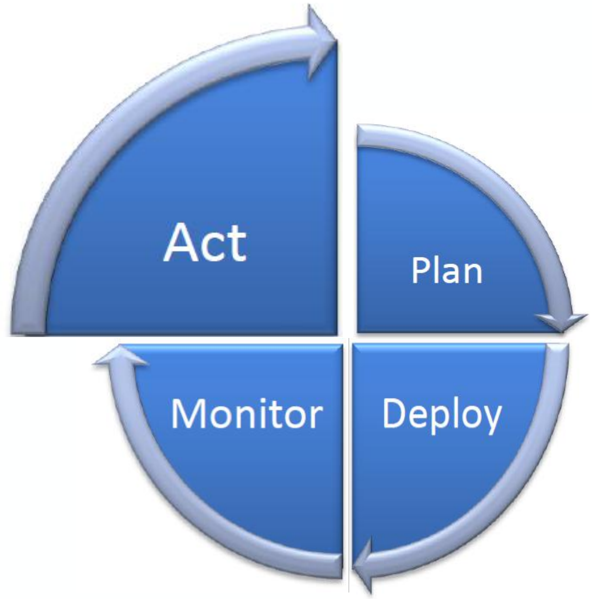
The DQE Cycle: (3) Monitor

- 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

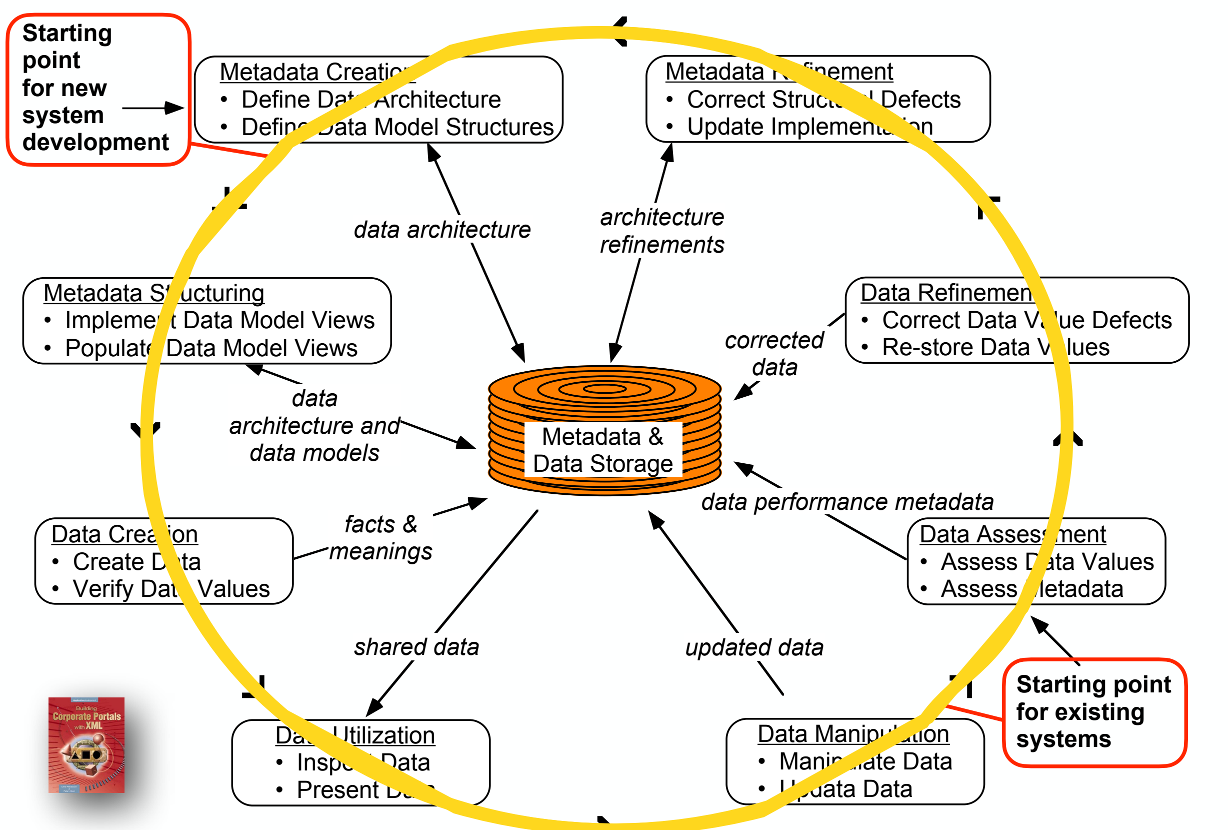


The DQE Cycle: (4) Act

- 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



Extended Data Life Cycle Model With Metadata Sources and Uses



Program Overview

Getting Data Quality Right
Engineering Success Stories

- Approaching Data Quality
  - Cloud considerations
  - Data quality attributes
  - Structural versus practice-related challenges
  - Digitization depends on quality data
  - Definitions
  - Must be built on leverage
  - Data quality examples
  - Causes can be difficult to discern
  - High quality data requires architecture/ engineering

- What do we need to get better at?
  - Systems thinking
  - Not looking at data quality in isolation
  - Developing repeatable capabilities/core data quality expertise
  - PDCA
- How do we get better?
  - Refocus the request around business outcomes
  - Get good at munging
  - Strategy
  - Investment characteristics
  - Conversations
  - Leadership
  - Programatic focus
  - Team development
  - Tangible ROI
- Takeaways and Q&A

Program Overview

Getting Data Quality Right
Engineering Success Stories

- Approaching Data Quality
  - Cloud considerations
  - Data quality attributes
  - Structural versus practice-related challenges
  - Digitization depends on quality data
  - Definitions
  - Must be built on leverage
  - Data quality examples
  - Causes can be difficult to discern
  - High quality data requires architecture/ engineering

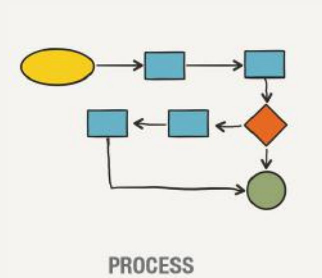
- What do we need to get better at?
  - Systems thinking
  - Not looking at data quality in isolation
  - Developing repeatable capabilities/core data quality expertise
  - PDCA
- How do we get better?
  - Refocus the request around business outcomes
  - Get good at munging
  - Strategy
  - Investment characteristics
  - Conversations
  - Leadership
  - Programatic focus
  - Team development
  - Tangible ROI
- Takeaways and Q&A

What Is Strategy?

strat·e·gy
/ˈstratəjē/
noun

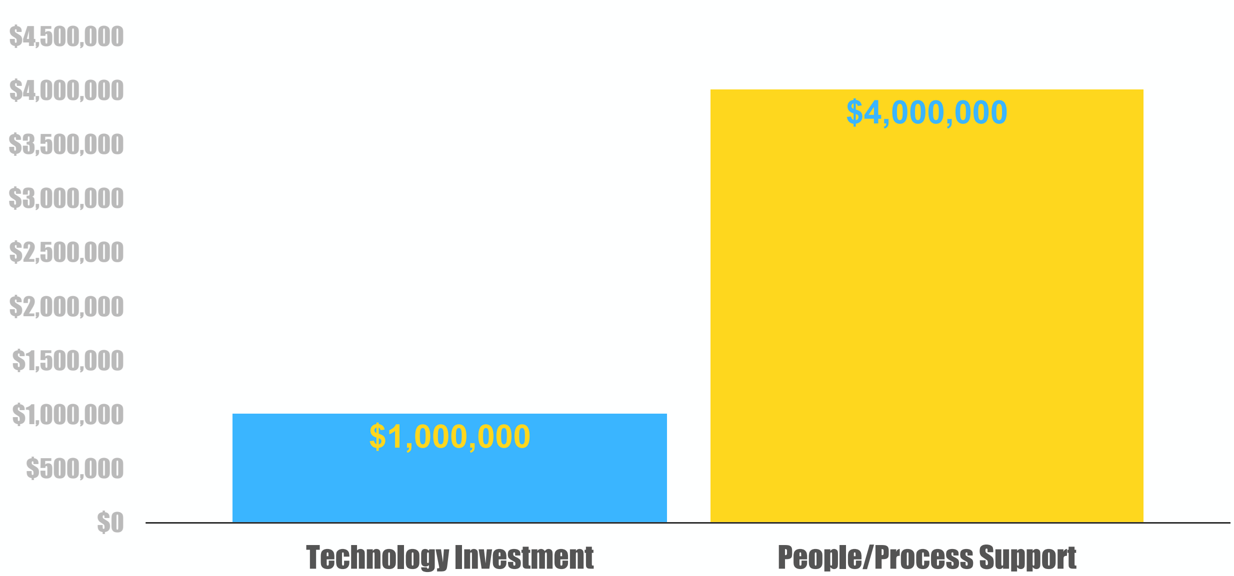
1. a plan of action or policy designed to achieve a major or overall aim.
   "time to develop a coherent economic strategy"
   synonyms: master plan, grand design, game plan, plan (of action), action plan, policy, program; More

- Current use derived from military
  - a pattern in a stream of decisions
    [Henry Mintzberg]



Data Investment Evaluation

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



Compare the Utility of Data Quality Conversation Topics

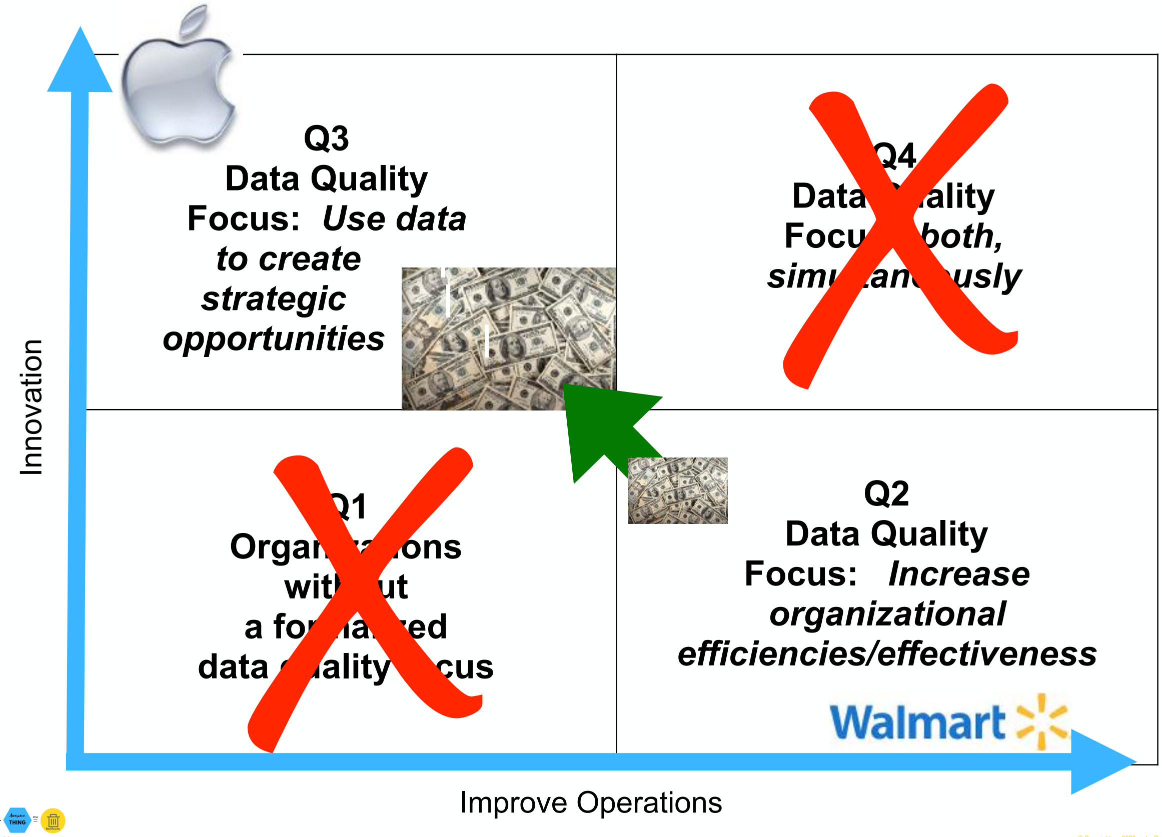
Engineers say:	Business wants to hear:
Clean some data	Decrease the number of undeliverable targeted marketing ads
Reorganize the database	Increase the ability of the salesforce to perform their own analyses
Develop a taxonomy	Create a common vocabulary for the organization
Optimize a query	Shaved 1 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

CDO Agenda

The CDOs goal is to better manage data as an organizational asset in support of the organizational mission!

- Inventory Data -> uncovering assets & decreasing ROT
- Develop the first version of an organizational data strategy
- Monetize your organization's data

Initially Pick One or the Other but Not Both



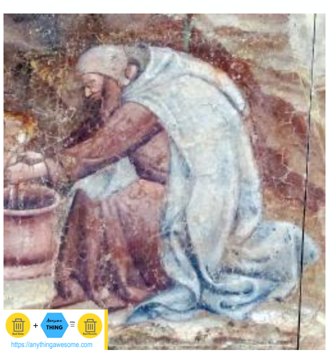
Improving Data Quality During System Migration

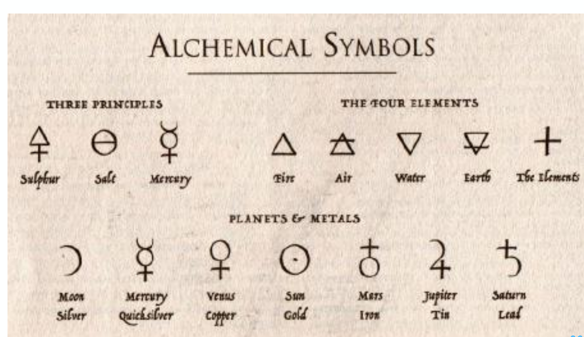
- Challenge
  - Millions of NSN/SKUs maintained in a catalog
  - Key and other data stored in clear text/comment fields
  - Original suggestion was manual approach to text extraction
  - Left the data structuring problem unsolved
- Solution
  - Proprietary, improvable text extraction process
  - Converted non-tabular data into tabular data
  - Saved a minimum of $5 million
  - Literally person centuries of work

Munge

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





Hidden Data Factories

Department A

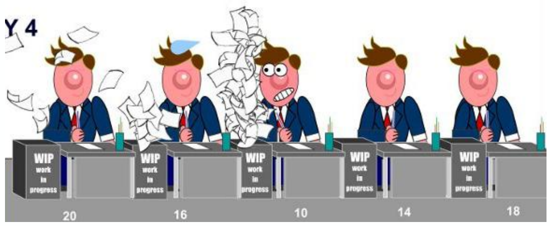
Work products are delivered to

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

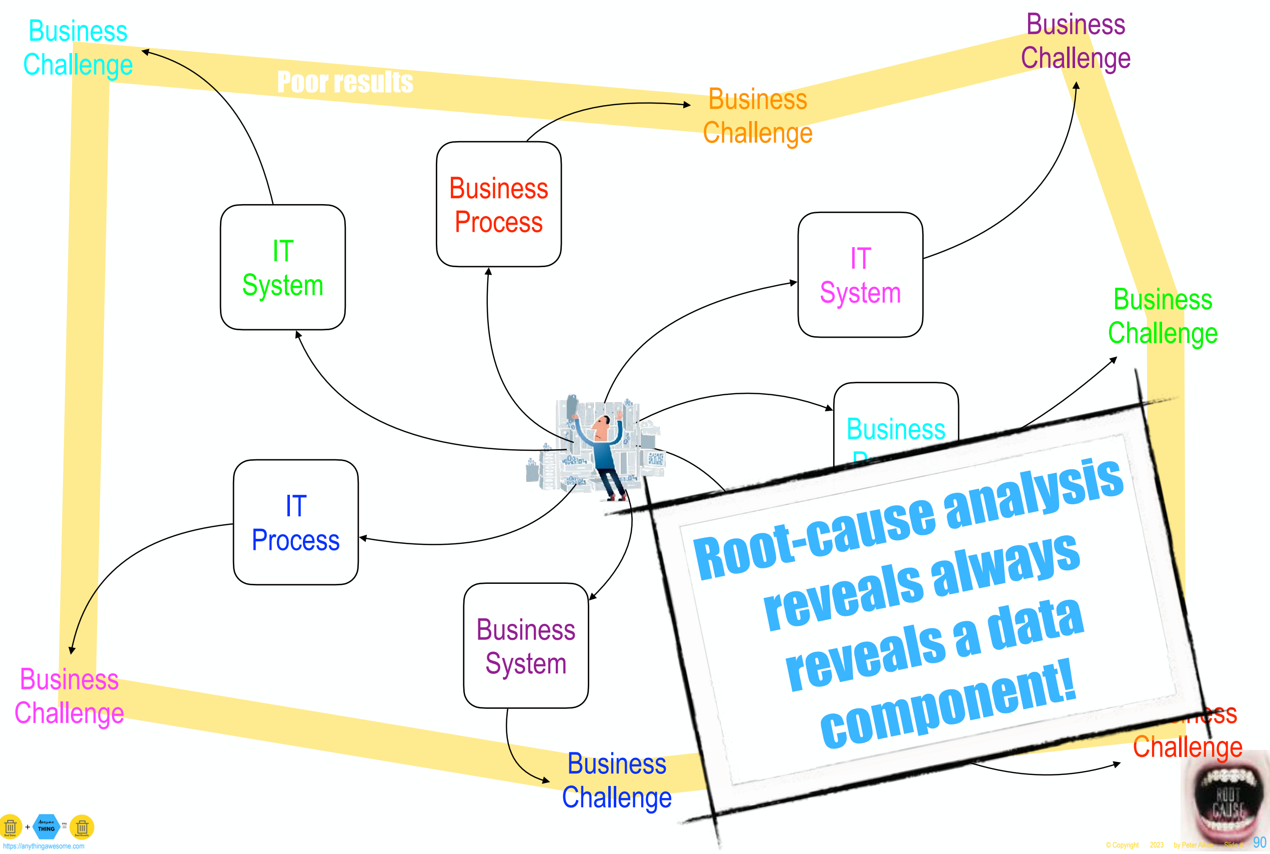


https://en.wikipedia.org/wiki/Theory_of_constraints

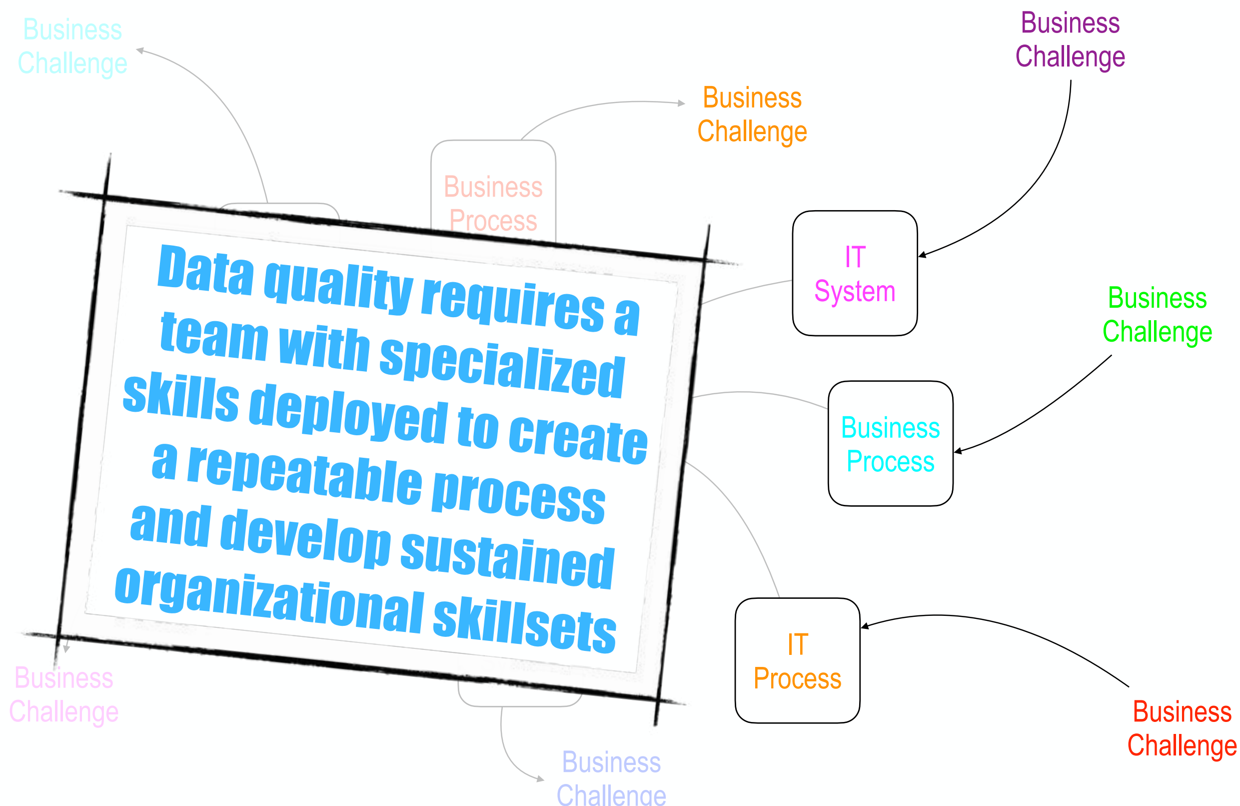
https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year

Poor Data Quality Manifests as Multifaceted Organizational Challenges

Poor Data Quality Manifests as Multifaceted Organizational Challenges



Consistency Encourages Quality Analysis



Determining Diminishing Returns

Week #	Unmatched Items (% Total)	Ignorable Items (% Total)	Items Matched (% Total)
1	31.47%	1.34%	N/A
2	21.22%	6.97%	N/A
3	20.66%	7.49%	N/A
4	32.48%	11.99%	55.53%
…	…	…	…
14	9.02%	22.62%	68.36%
15	9.06%	22.62%	68.33%
16	9.53%	22.62%	67.85%
17	9.5%	22.62%	67.88%
18	7.46%	22.62%	69.92%

Quantifying Benefits: Original Plan

Time needed to review all NSNs once over the life of the project:	
NSNs	2,000,000
Average time to review & cleanse (in minutes)	5
Total Time (in minutes)	10,000,000
Time available per resource over a one year period of time:	
Work weeks in a year	48
Work days in a week	5
Work hours in a day	7.5
Work minutes in a day	450
Total work minutes/year	108,000
Person years required to cleanse each NSN once prior to migration:	
Minutes needed	10,000,000
Minutes available person/year	108,000
Total Person-Years	92.6
Resource Cost to cleanse NSN's prior to migration:	
Avg salary for SME year (not including overhead)	$60,000.00
Projected years required to cleanse/total DLA person years saved	93
Total cost to cleanse/Total DLA savings to cleanse NSN's:	**$5.5 million**

Quantifying Benefits: Revised Plan

Time needed to review all NSNs once over the life of the project:	
NSNs	**150,000**
Average time to review & cleanse (in minutes)	5
Total Time (in minutes)	**750,000**
Time available per resource over a one year period of time:	
Work weeks in a year	48
Work days in a week	5
Work hours in a day	7.5
Work minutes in a day	450
Total work minutes/year	108,000
Person years required to cleanse each NSN once prior to migration:	
Minutes needed	**750,000**
Minutes available person/year	108,000
Total Person-Years	**7**
Resource Cost to cleanse NSN's prior to migration:	
Avg salary for SME year (not including overhead)	$60,000.00
Projected years required to cleanse/total DLA person years saved	**7**
Total cost to cleanse/Total DLA savings to cleanse NSN's:	**$420,000**

Quantifying Benefits: Social Engineering

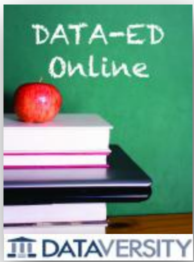
Time needed to review all NSNs once over the life of the project:	
NSNs	2,000,000
Average time to review & cleanse (in minutes)	5
Total Time (in minutes)	10,000,000
Time available per resource over a one year period of time:	
Work weeks in a year	48
Work days in a week	5
Work hours in a day	7.5
Work minutes in a day	450
Total work minutes/year	108,000
Person years required to cleanse each NSN once prior to migration:	
Minutes needed	10,000,000
Minutes available person/year	108,000
Total Person-Years	92.6
Resource Cost to cleanse NSN's prior to migration:	
Avg salary for SME year (not including overhead)	$60,000.00
Projected years required to cleanse/total DLA person years saved	93
Total cost to cleanse/Total DLA savings to cleanse NSN's:	**$5.5 million**

A Musical Analogy That Works for Both Practice and Storytelling

Program Overview

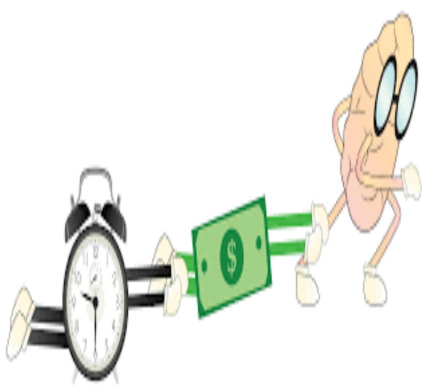
Getting Data Quality Right
Engineering Success Stories

- Approaching Data Quality
  - Cloud considerations
  - Data quality attributes
  - Structural versus practice-related challenges
  - Digitization depends on quality data
  - Definitions
  - Must be built on leverage
  - Data quality examples
  - Causes can be difficult to discern
  - High quality data requires architecture/ engineering

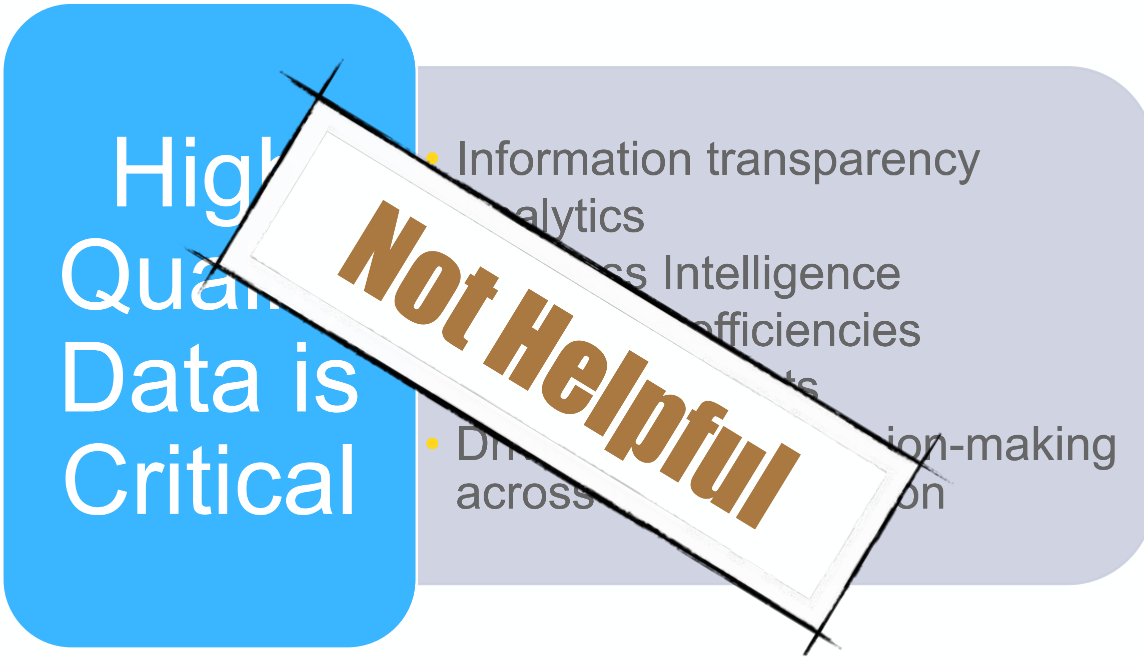


- What do we need to get better at?
  - Systems thinking
  - Not looking at data quality in isolation
  - Developing repeatable capabilities/core data quality expertise
  - PDCA
- How do we get better?
  - Refocus the request around business outcomes
  - Get good at munging
  - Strategy
  - Investment characteristics
  - Conversations
  - Leadership
  - Programatic focus
  - Team development
  - Tangible ROI
- Takeaways and Q&A

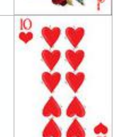
Comprehension by others is critical!



- If others do not understand what you do then you are perceived with a cost bias
- If others understand what you do then you can be perceived with a *value* bias

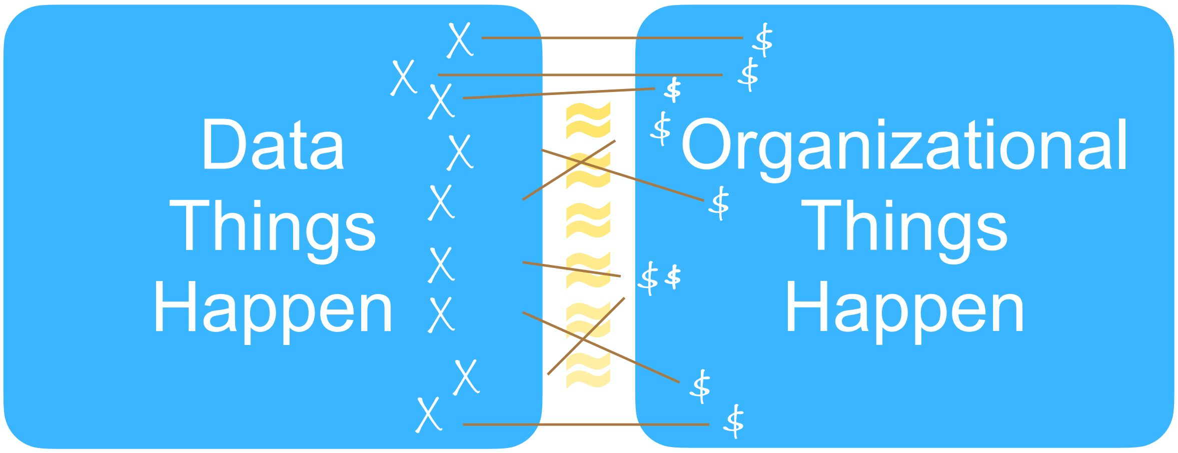


Winning Cards for data quality program success

1	The project needs to be small	Projects should not be allowed to begin unless the data requirements for the entire project are **verified**	
2	The product owner or sponsor or executor must be skilled in data	Few in IT have the requisite data skills and knowledge	
3	The process must be agile-ready	Agile is a construction technique/data requires more planning before construction	
4	The team must be highly skilled in both the data quality processes and technology	Few teams have requisite levels of data skills	
5	The organization must be highly skilled at emotional maturity	Few organizations understand data stuff	

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

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.

Upcoming Events

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

Strategy is Where Data Architecture and Data Governance Collide
10 October 2023

What's in Your Data Warehouse?
14 November 2023

Data Governance & Information Quality CONFERENCE

Data Management Best Practices
12 December 2023

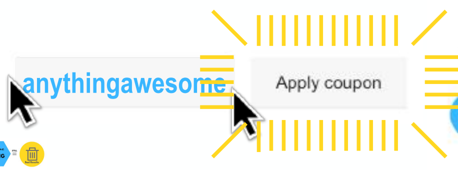
Brought to you by:

DATAVERSITY

[Clicking any webinar title will link directly to the registration page]

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"



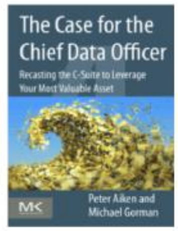
Bad Data + Awesome THING STILL = Bad Results

https://anythingawesome.com

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

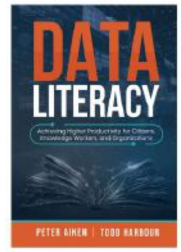
(The Chinese Translation Title is: Chief Data Officer Combat)

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



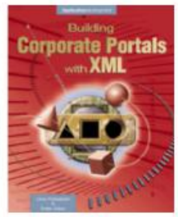
Data Literacy: Achieving Higher Productivity for Citizens, Knowledge Workers, and Organizations

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

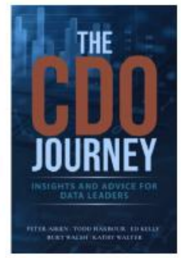


Building Corporate Portals with SML

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?

Executive Data Literacy Training?

Collaboration?

Mentoring?

Bad Data + Awesome THING STILL = Bad Results

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