MCKNIGHT
CONSULTING GROUP

Data Quality: The ROI of Adding Intelligence to Data

Presented by: William McKnight

#1 Global Influencer in Big Data Thinkers360

President, McKnight Consulting Group

3 X Inc 5000

/in/wmcknight

www.mcknightcg.com
(214) 514-1444

Semarchy

The Unified Data Platform

Master your data to accelerate business outcomes

Steven Lin | Product Marketing Manager

Who We Are

Semarchy

> Easy to use MDM solution allows rapid development of new features. Changes can be implemented, reviewed, tested, and deployed in minutes." - *Global Financial Services Customer*

Master Data Management & Data Integration Leader

Source: Gartner (December 2021)

Gartner Peer Insights Customers' Choice 2022

Gartner peerinsights customers' choice 2020

Gartner peerinsights customers' choice 2019

+330

Clients worldwide across industries & use cases

80%

Deliver functioning MDM solution in under 12 weeks

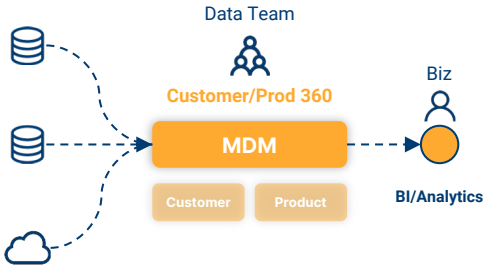
63%

Leverage xDM for multiple use cases or multi-domains

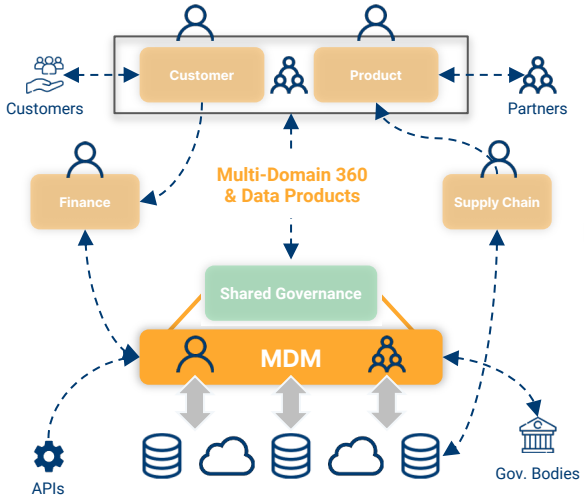
Why is Data Quality so Challenging?

Your data ecosystem and business needs are growing – *fast*.

Data Needs Are Less Like This...



And More Like This...



With Headaches Like These

- What is the value?
- Why start now?
- Where do we start?
- How do we deliver?
- Who can we trust?

Ok...How Should I Get Started Today?

Start small with a single use case/domain with future-ready design

Semarchy

Start Small ---------- **See Results** ---------- ***Scale Fast*** ----------▶

1 Prioritize one use case/domain

Customer 360

2 Align to actual business KPIs

Increase % MoM Revenue Growth

Increase % conversion rate

Increase % data accuracy rate

Increase % data completeness rate

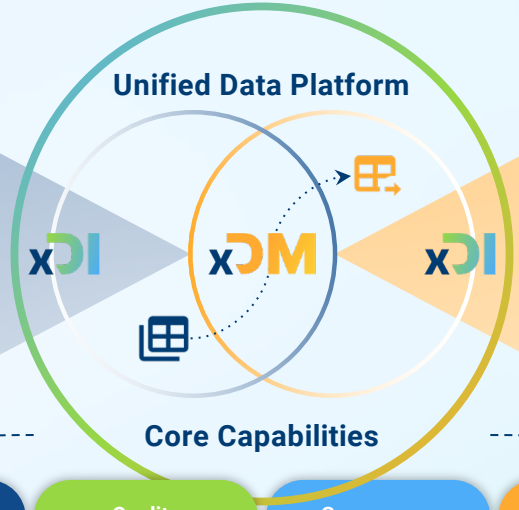
Decrease % duplicate data

3 Design future-ready solution

- **Easily Configurable**
- **Flexible Architecture**
- **Unified Data Platform**

Two Incredible Modules, One Unified Platform

Semarchy's unified data platform **orchestrates (xDI)** and **masters (xDM)** data using a **no-code, business-driven configuration** approach to **rapidly generate custom apps** and **deliver high-quality golden records** across your organization



Ingests Landing Data

From your source systems in real-time or batch processing with any technology

Orchestrates Golden Data

To your target sources on any architecture for your operational or analytical needs

Discovery	Integration	Quality	Governance	Management	Analytics
Profile	ETL/ELT	Cleansing	Workflows	Data Modeling	Data Viz
Pattern Analysis	BPM Tools	Enrich/Blend	Policies	Stewardship	KPI Dash
Charting	Virtualization	Match/Merge	Glossaries	Lineage	Embed
Recommend	APIs	Enforcement	Roles/SSO	Search	Drill-Down

Deploy anywhere: on-prem, cloud or hybrid

MONTE CARLO

Measuring the ROI of Data Quality with Data Observability

Data Quality Best Practices

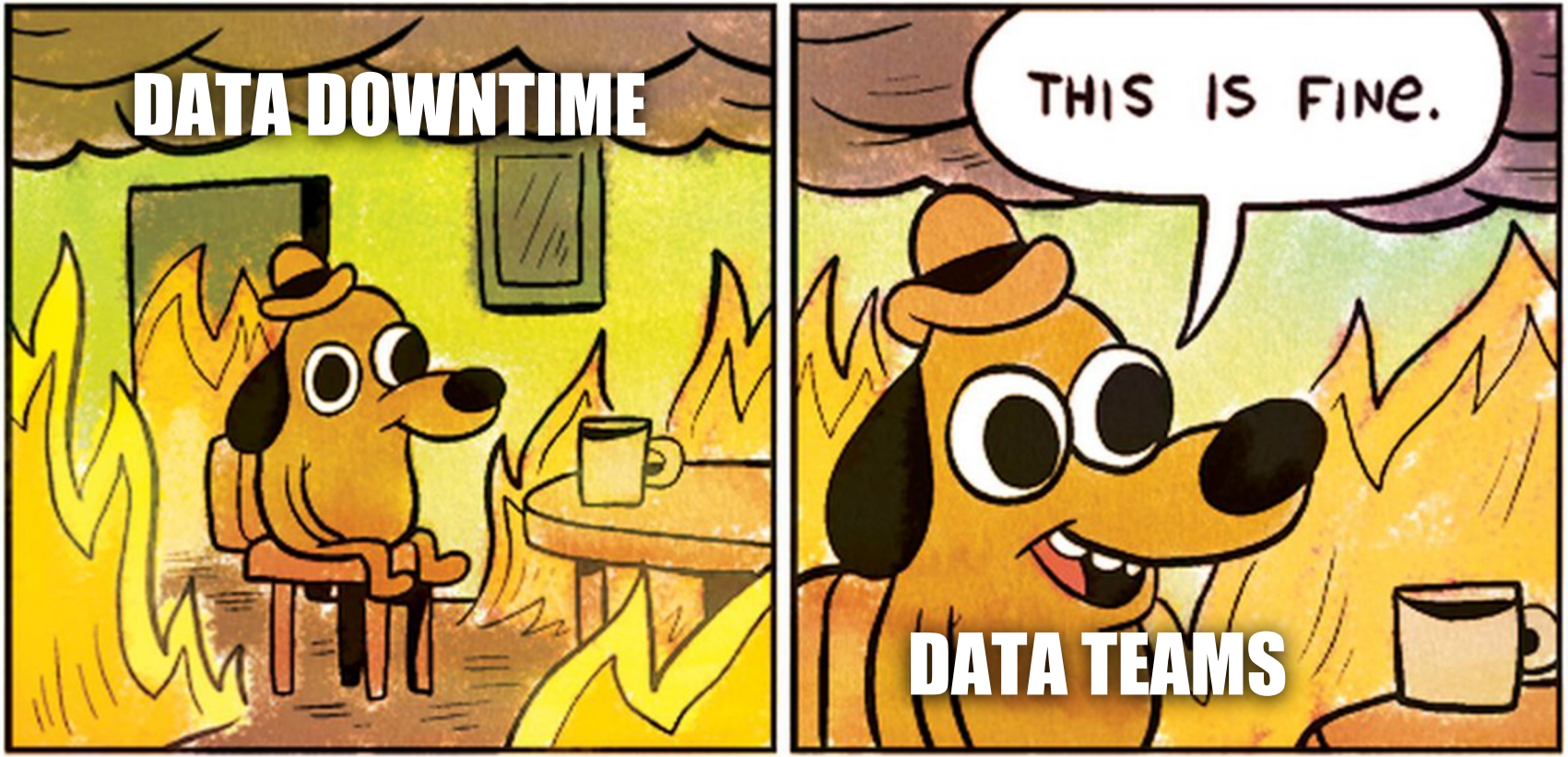
Today's Sponsor

Jesse Miller

Senior Product Manager
at Monte Carlo

MONTE CARLO

What is Data Downtime?


DATA DOWNTIME
THIS IS FINE.
DATA TEAMS

Business Impact from poor data quality

~70

high severity events each year per every 1k tables [1]

30-50%

data engineering time spent on fire drills [2]

80%

data science & analytics teams time spent on collecting, cleaning, and preparing data [3]

1. Benchmark data based on Monte Carlo customer production deployments
2. Monte Carlo market research and customer-reported benchmarks
3. Crowdflower

MONTE CARLO

Business Impact from poor data quality

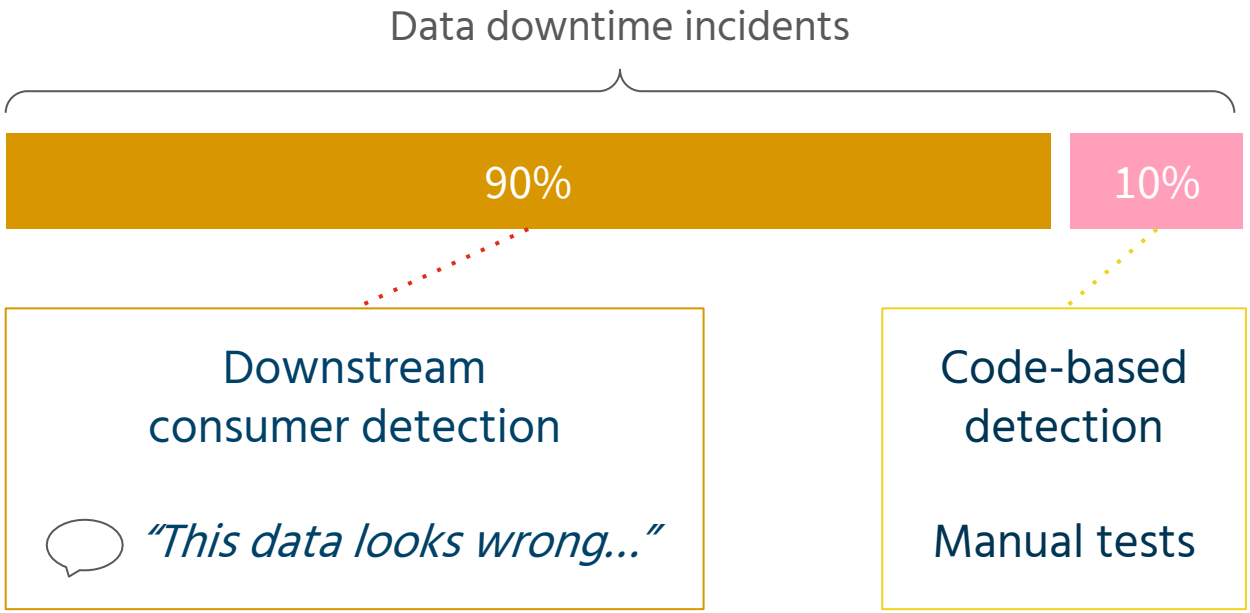
12-27%

avg. annual revenue lost for companies resulting from poor data quality [1]

1. Experian Data Quality; MC research via Wakefield Research Survey

MONTE CARLO

Data quality incidents are detected reactively today



The result?

Days to weeks pass before incidents are detected and resolved

MONTE CARLO

Good news: Data downtime looks similar across companies

- Is this data up-to-date?
- Why does this data size look off?
- Isn't this value suspiciously high?
- Why are there so many nulls?
- Why do we have duplicate IDs?
- What reports will I break with this schema update?
- Why are there 0s on tiles that usually show 100s?

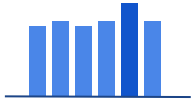
MC MONTE CARLO

The 5 pillars of Data Observability

Freshness

Seeks to understand how **up-to-date** your data tables are as well as the cadence at which your tables are updated.

Volume



Refers to the **completeness of your data** tables and offers insights on the health of your data sources.

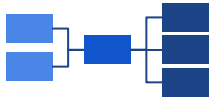
Distribution

Ensures trust in your data by **monitoring the values in your tables** and alerts you if your data falls out of an accepted range.

Schema

Monitors all **schema changes** in your environment, and alerts you to added, removed or updated fields or tables.

Lineage



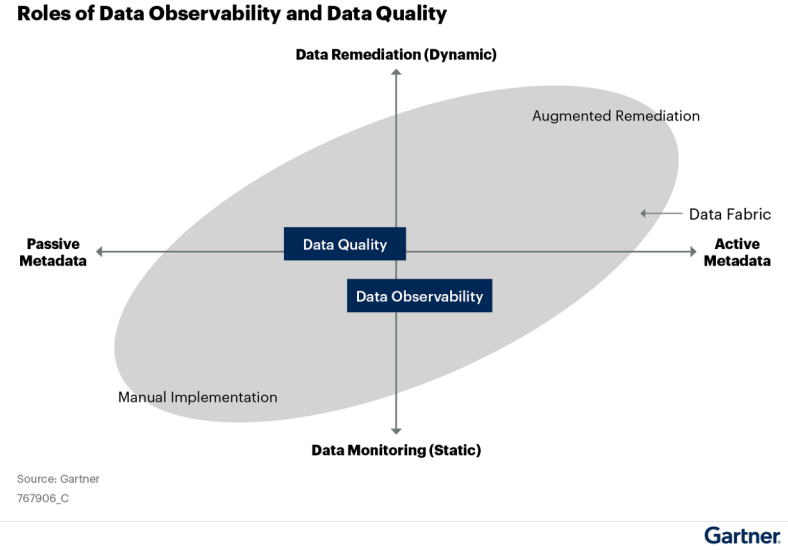
Monitors the **upstream sources and downstream ingestors** of data to highlight what may be impacted by a break in your data.

MONTE CARLO

Complimentary Copy

[New] Innovation Insight: Data Observability Enables Proactive Data Quality

https://info.montecarlodata.com/gartner-innovation-insights-data-observability/



MONTE CARLO

Thank you!

Enterprise Data is Still a Mess

- The proliferation of data sources
- The complexity of data formats
- The lack of data governance
- The push into AI

Analytics And Data Science

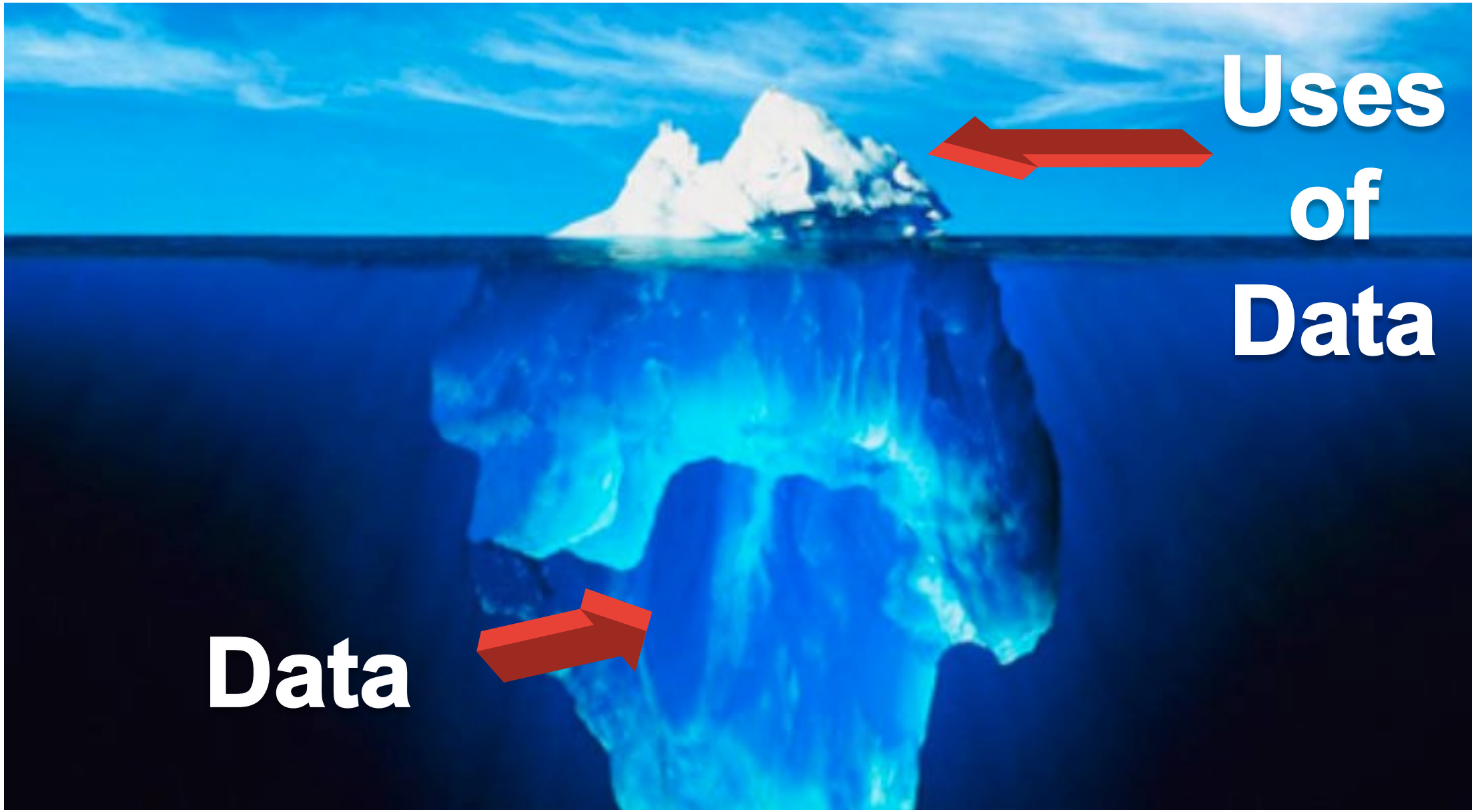
Bad Data Is Sapping Your Team's Productivity

by Thomas C. Redman

November 30, 2022

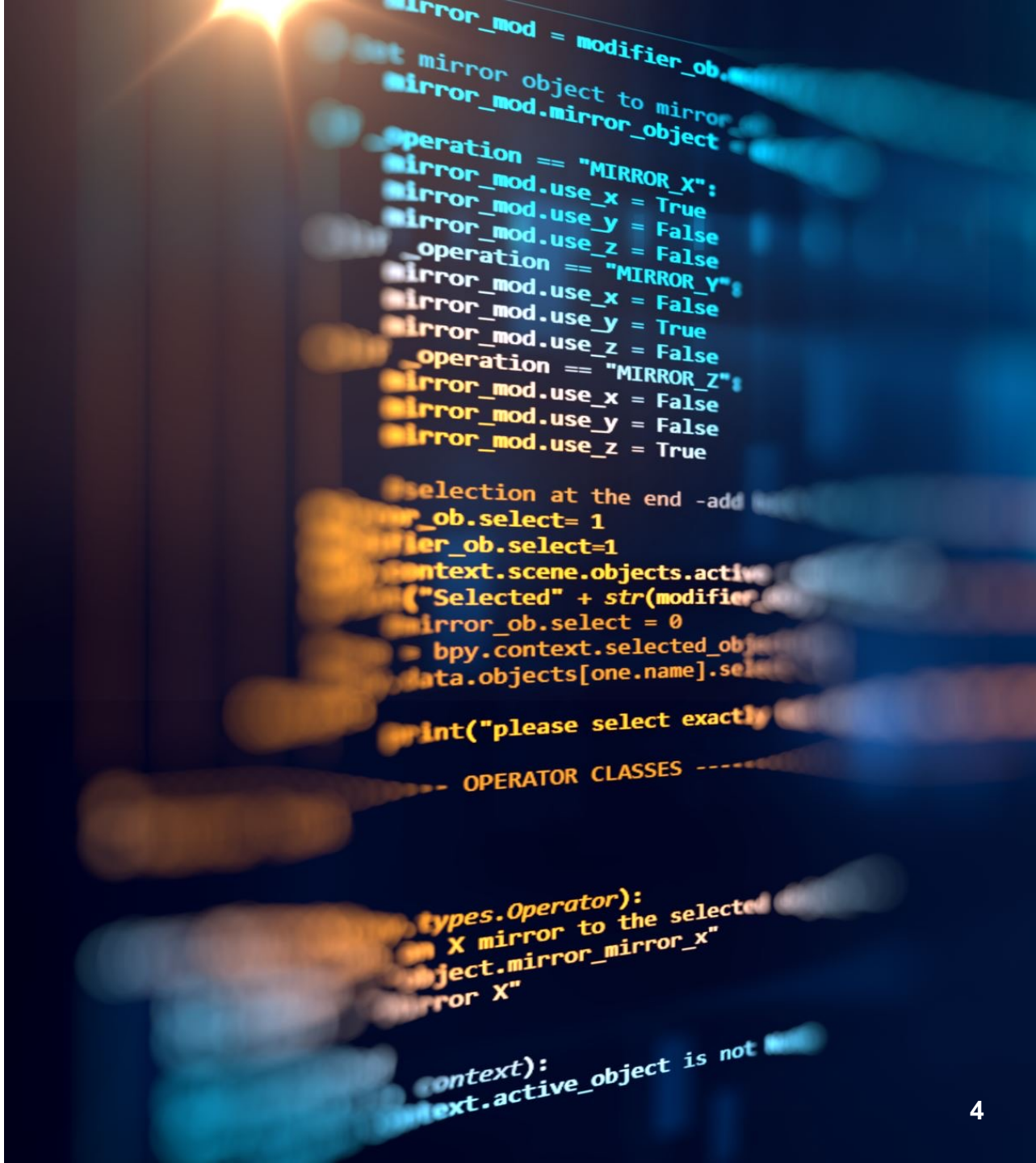
Westend61/Getty Images

Summary. Data science was supposed to create a new productivity boom. But, for many companies, that boom never arrived. What's gone wrong? While companies have invested in data tools, much of the data that's fed into these systems is low quality — with mislabeled, missing, or incorrect information, which in turn creates more work, and more... **more**


Uses of Data
Data

CDO Focus

- Focus on Hypershared Artifacts
  - Data Warehouse
  - Data Lake
  - Master Data Management
- Data Architecture
- Data Innovation
- Data Governance and Data Quality



Data Quality is Essential to Business Success

- “Correct” data is a widespread need
- Yet, data quality lacks consistent definition
  - » You can’t improve what you can’t measure
  - » Tangible benefits accrue from improved efficacy of the applications using the data

Most People Don't Care About It Until They Do

- Usually not considered critical path
- You must be an advocate
- Cite downside risk

Consider these business imperatives

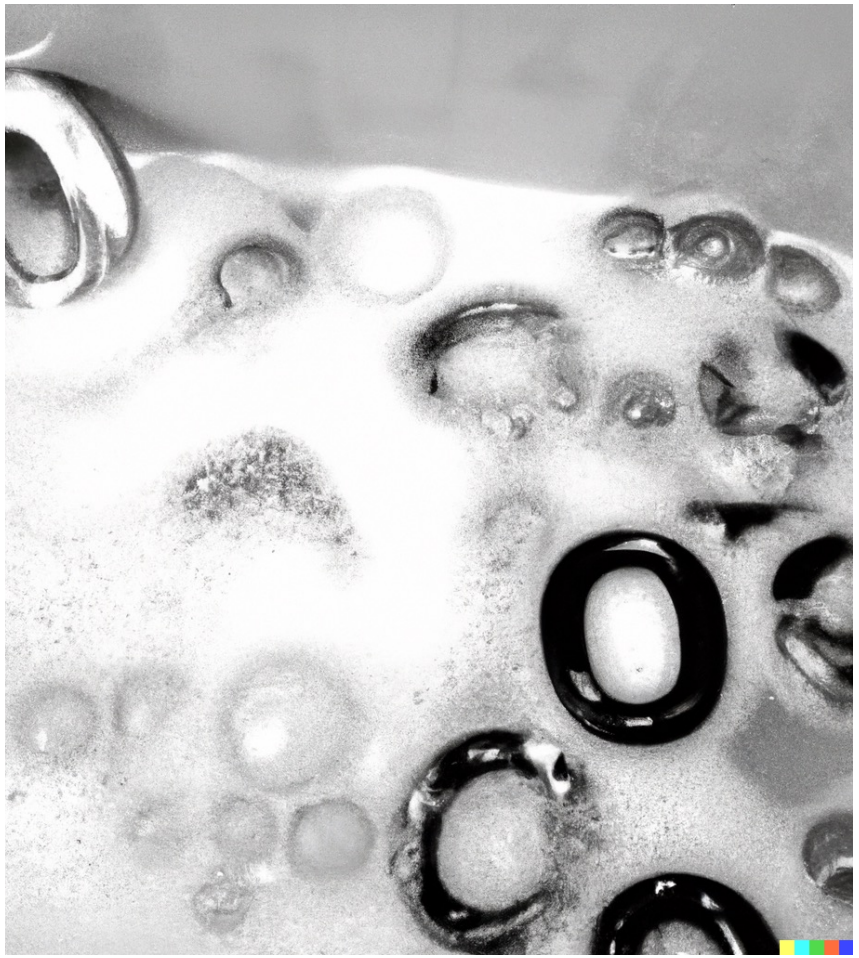
- Information-based in-store and contact center cross- and up-selling (NEEDS clean customer and product data)
- Credit card fraud detection (NEEDS clean customer and transaction data)
- Supply chain efficiencies and just-in-time production capabilities (NEEDS clean product and location data)
- Predictive churn management (NEEDS clean customer and transaction data)
- Many others

These have failed or underperformed because of incomplete, incorrect, inconsistent data

Investments in Data Quality

- Investments Yield “Cleaner” Data
- Business objectives cannot be met without quality data in support
  - Data Quality Returns are in the improved efficacy of projects targeting business objectives
- Data Quality should be an integral part of most projects

The Benefit Of Clean Data Is Not Enough



- ROI
- Strategic Benefit
- Lower TCO

Data Quality Should Have a Value Proposition To Project(s)

- Improved Decision Making
- Increased Efficiency
- Reduced Risk
- Improved Customer Satisfaction
- Increased Profitability
- Enhanced Security
- Improved Compliance


Data Governance

A Good Data Governance Program Keeps Business Interest In Data Quality

Accountable Data Governance

- Data governance is a very malleable term
- Without also establishing accountability and tangible delivery, establishing data governance is not helpful
- Those data governance organizations that are thriving deliver to the organization both in support of projects and as a horizontal organizational function
- Successfully done, it is actually quite a facilitator role
- Data Governance needs to strongly align itself with those data stores that have high leverage in the organization
- Also at the organizational policy-making level, there are global data retention policies, documenting data definitions and providing data lineage

Without a Basis In Data Quality, The Counterparty Has No Idea What You're Talking About

- Listen actively and attentively
- Be open and honest
- Be patient and understanding
- Communicate regularly
- Set clear expectations
- Compromise
- Be flexible

Communication is Real Work

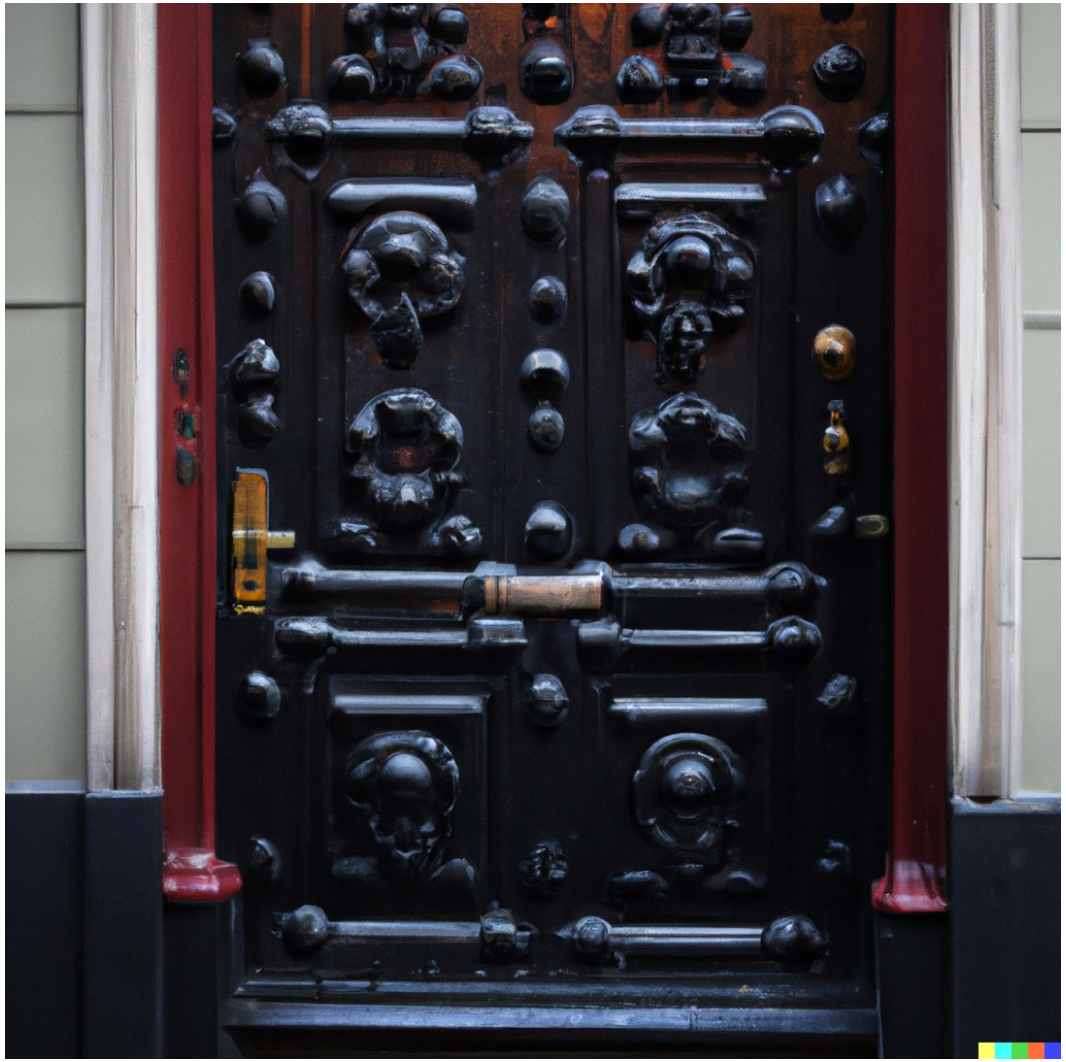
- Governance Meetings
- Decisions
- Actions
- Timing

Measuring Data Quality

Only a Methodological Approach Will Work

- Repeatable Process
- Progressive Improvement
- New Data
- Requirements Change

The Causes Of Poor Data Quality Keep Coming In The Front Door



Data Quality Improvement Program

- Define the quality expectations
- Profile data
- Measure data quality improvement options
- Select the best option
- Improve the quality of data and improve the business

Data Quality Rule Categories

- Align business processes with data-driven insights
- Data-Driven Decision Making
- Referential Integrity
- Uniqueness
- Cardinality
- Subtype/Supertype constructs
- Value reasonableness
- Consistency
- Formatting
- Data derivation
- Completeness
- Correctness

Four Actions to Perform for Data Quality

- Screen Data Entry
- Add Cross-Checking
- Quarantine Data (transfer of DQ)
- Report on Quality Violations (still requires additional action to fix DQ)
- Change or Repair Incorrect Data to Conform to DQ

Yes, Data Quality Can Be Automated

- Data Profiling can be used to automate data quality checks by identifying data inconsistencies and outliers.
- Data Cleansing: Automated data cleansing tools can be used to identify and correct errors in data sets.
- Data Validation: Automated data validation tools can be used to check data for accuracy and completeness.
- Data Standardization: Automated data standardization tools can be used to ensure data is in the correct format.
- Data Integration: Automated data integration tools can be used to ensure data is properly combined and consistent across sources.

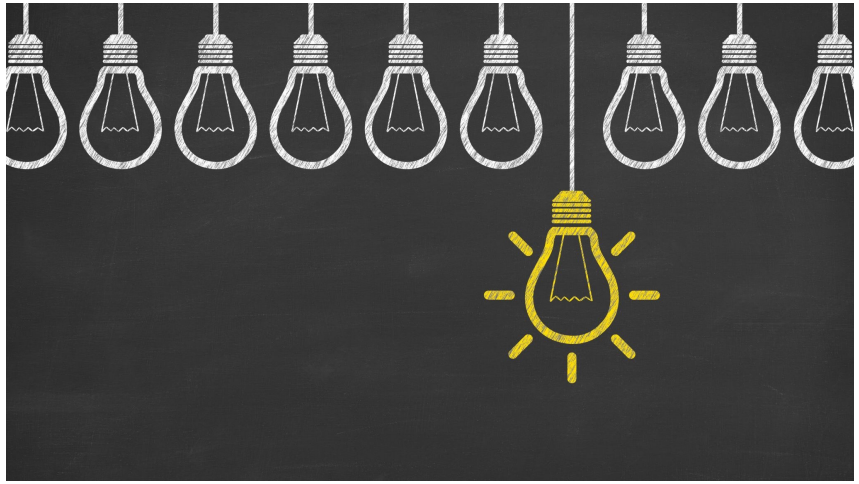
Put Quality Data in a Leveragable Platform

- Data Warehouse
- Data Lake
- Master Data Management
- Data Hub

Every Project Needs a Focus On Data Quality

- Clean data is the key to unlocking the power of many business processes, including:
  - Information-based in-store and contact center cross- and up-selling (NEEDS clean customer and product data)
  - Credit card fraud detection (NEEDS clean customer and transaction data)
  - Supply chain efficiencies and just-in-time production capabilities (NEEDS clean product and location data)
  - Predictive churn management (NEEDS clean customer and transaction data)
- Having clean customer, product, transaction, and location data is essential for these projects to be successful.

Data Quality Will Never Be Perfect



1. Appropriate
2. Suitable
3. Adequate
4. Proper
5. Satisfactory
6. Acceptable
7. Relevant
8. Serviceable
9. Functional
10. Usable
11. **Fit-for-purpose**

Data Quality Scoring

- Scoring defines how well your data meets business expectations

$$\frac{\text{Adherence}}{\text{Possibilities}}$$

Multiple prorated rules used to determine overall system score

Cost to the Enterprise of Poor Data Quality

- One-off DQ repeated remediations
- Poor/Failed Enterprise Initiatives
- Misguided roadmaps
- Compliance cost
- $x per data record attributed to:
  - Failed outreach
  - Losing customers
  - Storage space and effort with duplicate records
  - Incorrect marketing segmentation and personalization
- Cost expansion
  - On average corporate data grows at 40% per year

How can Improved Data Quality Improve Initiatives?

Example: Targeted Marketing

1. Fewer bad contact data
2. Improved customer segmentation
3. Higher Marketing Initiative ROI

Case Example: Retailer Marketing to Prospects

Higher scores means more contact gets through and better targeting.

DQ Score	Prospects Reached	Return on Marketing	Avg. profit	Return (prospects X ROM X avg. profit)	investment *	ROI (return-investment/ investment) *
95	110000	0.04	250	1100000	400000	175.00%
90	105000	0.03	250	787500	390000	101.92%
85	100000	0.025	250	625000	380000	64.47%

*considering only outreach costs, not considering DQ improvements

Cost/Benefit of Adding Data Quality Efforts

DQ Score	investment *	ROI (return-investment/ investment) *	additional cost to improve DQ to the level	Revised Project ROI (return-(investment+DQ cost)/ investment+ DQ cost)
99	410000	198.02%	300000	72.10%
98	400000	175.00%	200000	83.33%
95	390000	101.92%	75000	69.35%
90**	380000	64.47%	0	64.47%

Best Value Proposition

*without considering costing for DQ improvements **Default score – will happen without data quality improvements

Case Example: Where was the Data Quality investment to go?

- Reverse-append existing customers for addresses
- Reverse-append existing customers for demographics
- Purchase additional prospects
- Use of multiple and different third-party data providers (and corresponding de-duplication of results)

Other Examples

- Insurance: Screening erroneous data entry reduces fake claims, inflated claim amounts, multiple claims for the same incident, misrepresentation of policy coverage, policyholder impersonation, claims for ineligible vehicles, all of which could lead to unnecessary expenses
  - DQ can create up-front alerts for proper internal routing which creates efficiencies
- Healthcare: Cross-checking diagnosis and treatment plans to past, successful plans reduces inaccurate diagnosis, improper treatment plans, delayed treatment plans, all of which lead to increased claims cost
- AI Applications: Inaccurate or non-representative data leads to biased and inaccurate results which lead to subpar application ROI and potentially compliance costs
  - Increased computation leads to increased expenses

If you can materially (i.e., by 1%) improve any of these (types of) items, and you do a reasonable job with data quality, DQ will more than pay for itself

Miscellaneous

When to Consider Using a DQ Tool

- Complex data relationships across multiple systems
- 100,000+ attributes, 2,500 entities
- 1,000+ rules to implement
- Egregious data cleansing and transformation @ 3% data
- Data quality assessment requirements multiple times per day
- 7+ years of data to assess
- 25+ data stewards

Big Data Collection Systems

- Exactly once versus At Least Once
- At Least Once
  - Guarantees Order of Delivery
    - Apache Kafka/Amazon MSK
    - Kinesis Data Streams
  - Does not Guarantee Order of Delivery
    - SQS in Standard Mode
    - Kinesis Data Firehose
- Exactly Once with Guaranteed Order
  - SQS in FIFO Mode
  - Dynamo DB Streams

Data Catalogs in the Data Stack

- Data Catalogs serve as metadata store for all services including data integration, prep/transformation, data lake, DW, ML
- Identifies relationships
- Identifies data pipelines
- Serves preferences in data set selection
- Documents all data sets (including connection info)

- **Data Discovery**
- **Data Security**
- **Data Lineage**
- **Data Analysis**
- **Increased Collaboration**

Streaming Data Data Quality

- Streaming data generates a large amount of data in real time, making it extra difficult to assess and maintain data quality.
- Streaming data can come from diverse sources and have different formats, structures, and semantics, posing challenges in data harmonization and standardization.
- Data quality issues need to be identified and resolved quickly in streaming scenarios to ensure the accuracy and reliability of real-time analytics and decision-making.

So…

- Continuously monitor streaming data to identify anomalies, outliers, and deviations from expected patterns, indicating potential data quality issues.
- Implement real-time data validation rules and cleansing techniques to detect and correct errors, inconsistencies, and missing values on the fly.
- Supplement streaming data with additional information from external sources to enhance its completeness and context, enabling more accurate analysis and decision-making.
- Track the origin, transformations, and usage of streaming data to ensure transparency, traceability, and auditability of data quality processes.

Data Lineage

- Lineage collects and organizes knowledge of data structure, operational paths, and business outcomes or analyses.
  - Automated methods gather structure maps and process details, then stays aware of changes to that structure. This is one of the most complicated tasks of Data Cartography.
  - Stewards start the process of adding terminology, enhanced descriptive information, and business “signposts” to fill out the original
  - Wider audiences supervised by stewards continue the process to use and add relevant information into the system: technical support, audit specialists, IT practitioners, etc.
  - High-cost project implementation teams and compliance management can leverage accurate information to do their jobs efficiently and add governance layers onto the lineage.

- Graphical Representation
- Impact Analysis
- Root Cause Analysis
- Extend to Non-Standard or Custom Sources
- General Accessibility to Lineage and Metadata

Data Quality Subsumed into Data Observability

- Data Quality
- Data Freshness
- Data Volume
- Schema Change
- Data Lineage
- FinOps

Scale detection

Leverage ML to generate explainable and adaptive DQ rules

Scale architecture

Scan large and diverse databases, files and streaming data

Scale adoption

Empower users with a unified scoring system, lineage and personal alerts

Recommendations

- Data quality can and should have a value proposition
  - Data quality is never an accident
  - Consider data quality when considering applications
- Measure the level of your data quality
- Data Quality is a business-driven imperative
- Data Quality is becoming part of Data Observability
- Take care of the People Issues Associated with Data Quality
- Establish the Value Proposition for Data Quality in Project Improvement
- Run Data Quality as an Ongoing Process

MCKNIGHT
CONSULTING GROUP

Data Quality: The ROI of Adding Intelligence to Data

Presented by: William McKnight

#1 Global Influencer in Big Data Thinkers360

President, McKnight Consulting Group

3 X Inc 5000

/in/wmcknight

www.mcknightcg.com
(214) 514-1444