Data Governance and Capitalization in a Big Data Era

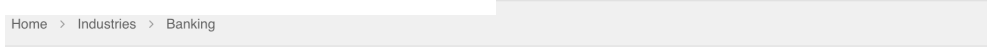
Pieter De Leenheer, PhD

Cofounder and VP, Research and Education at Collibra

MIT ICIQ 2017

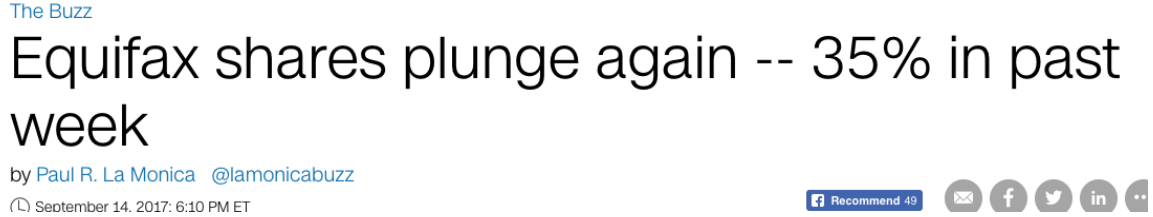
October 6, 2017

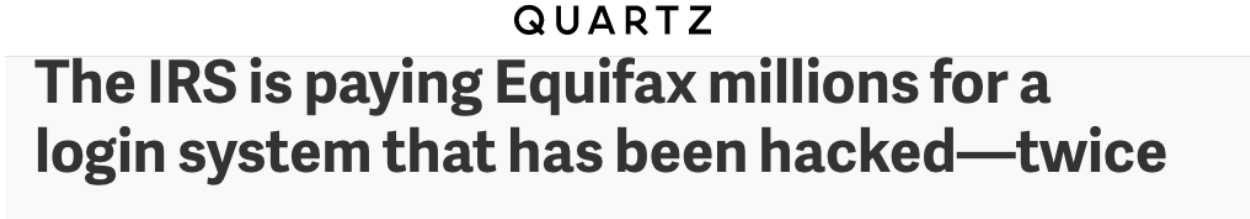
The Daily News



Equifax shares plunge after data breach potentially impacting 143 million Americans

Published: Sept 7, 2017 6:32 p.m. ET

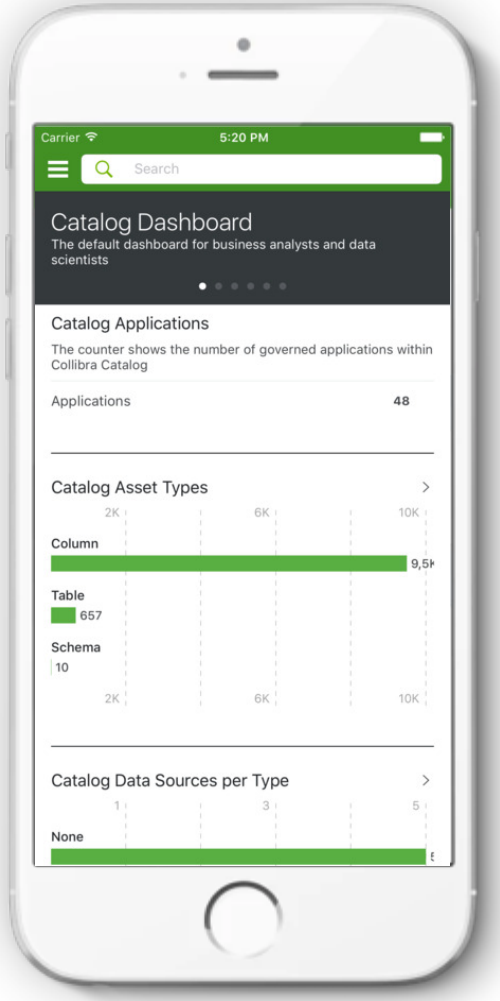




collibra

Administration

- Slide Deck
  - https://www.slideshare.net/pdeleenh/data-gover
- Software
  - Data Governance Center:
    - Inno.collibra.com
    - Username: johnfisher / password: merrimack
  - On the go:
    - Collibra on the App Store
- Collibra University
  - Sign up for free: http://university.collibra.com
- Collibra Blog, e.g.:
  - https://www.collibra.com/blog/unleash-the-data-democracy-5-misconceptions-of-data-governance/



Overview

Intro - A Data Governance Odyssey

- Finding the “man of the book”
- About the company Collibra

Part 1 – The Chief Data Officer Rises

- Digital Darwinism
- The Big Data Bang
- Data Brawls and FUD
- The Chief Data Officer Role Types

Part 2: Data Universe Expands

- Data Value Hierarchies, Networks and Hybrids
- Shift in Data Governance Approaches
- Systems of Record vs. Systems of Engagement
- Challenges:
  - Big Data Analytics
  - Digitalization of Trust
  - Weapons of Math Destruction

Part 3: A Lens on the Data Universe

- A system of record for data
- Use Cases

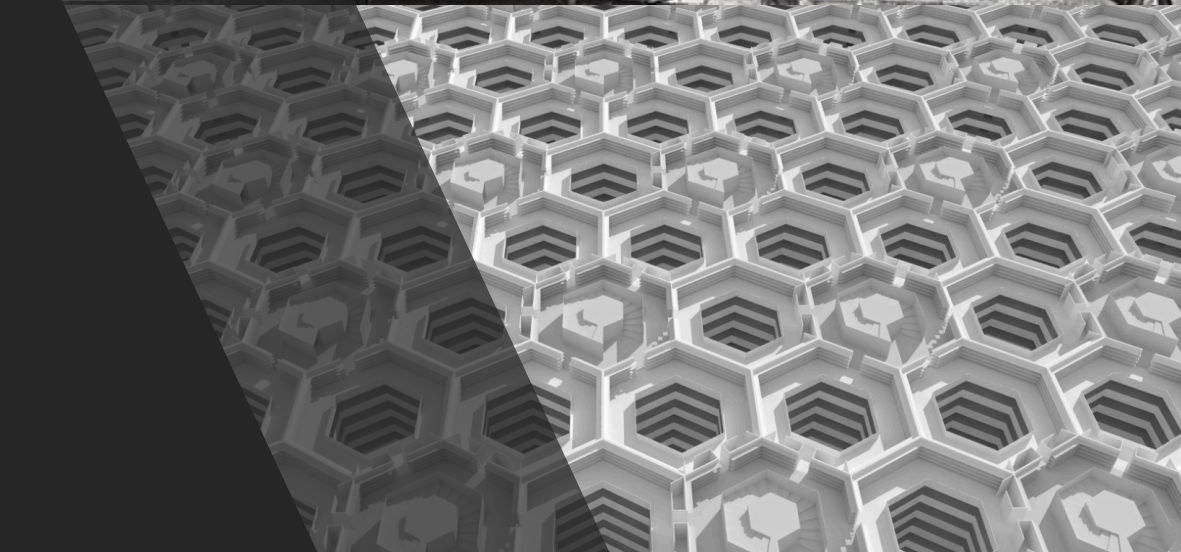
Introduction

2008: A Data Governance Odyssey

JL Borges

Library of Babel

Purifiers
Man of the book



Leading Data Governance Software company

Forbes 2017 THE Cloud100 WORLD'S BEST CLOUD COMPANIES
BESSEMER VENTURE PARTNERS salesforce ventures

+200 Customers

~280 Collibrians

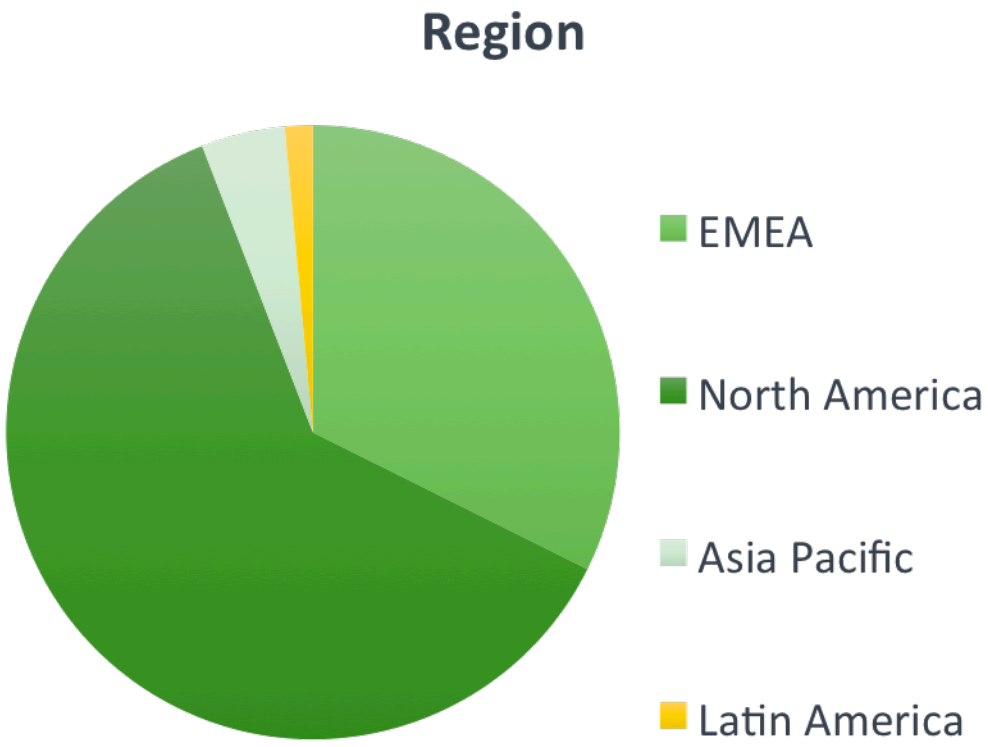
Offices in NYC, Bxl,

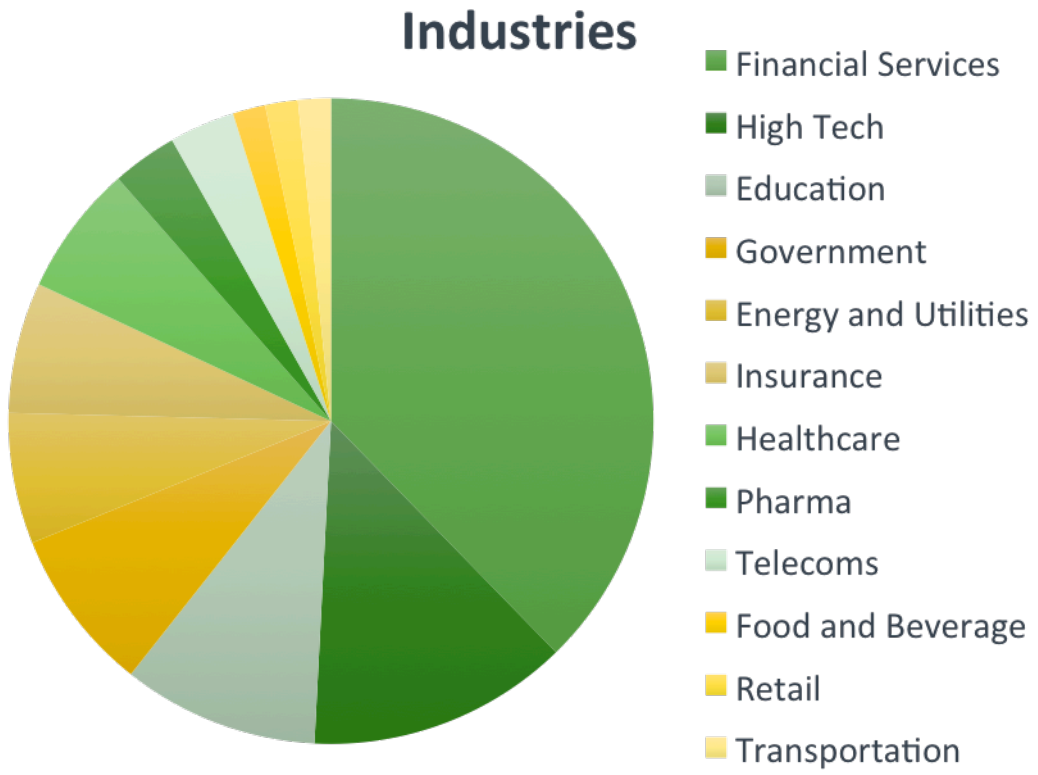
London, Paris, Wroclaw

collibra

Across verticals & geographies

150 Customers **>50% cloud** **6 of 10 largest US banks**



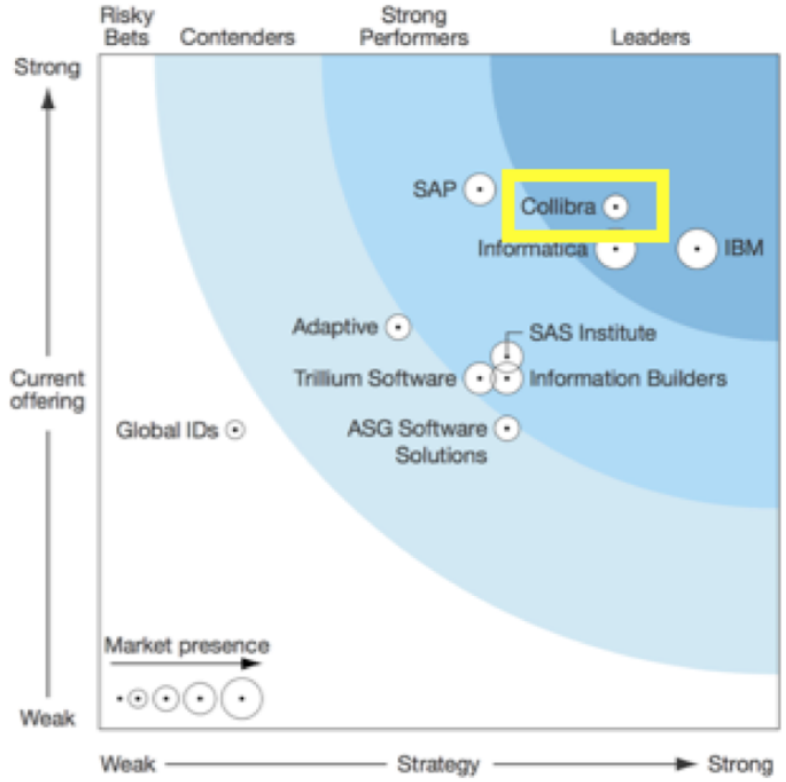


collibra

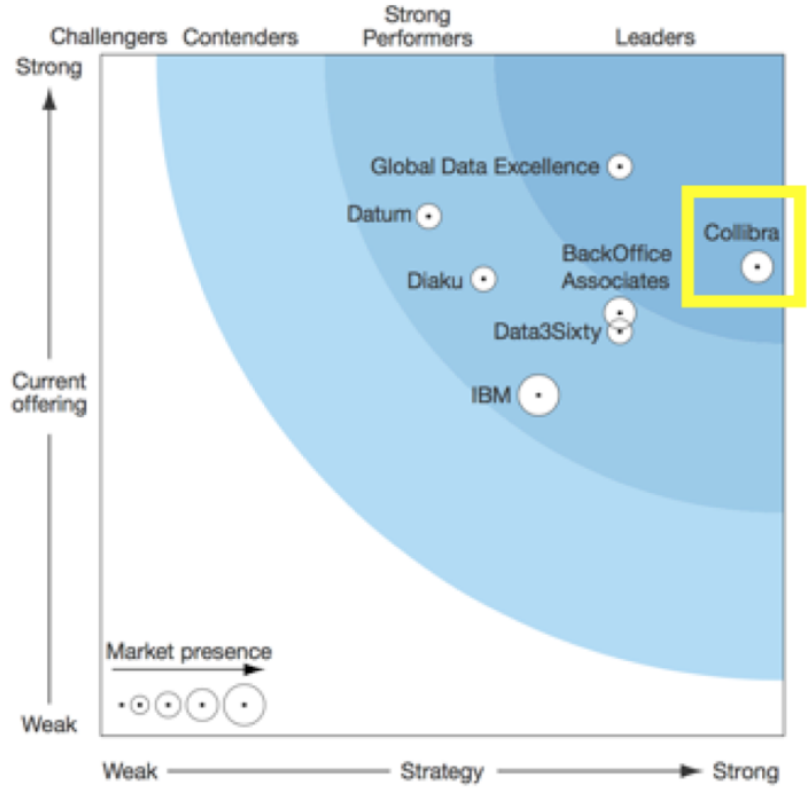
With Blue-chip Customers

collibra

Analysts' Positioning



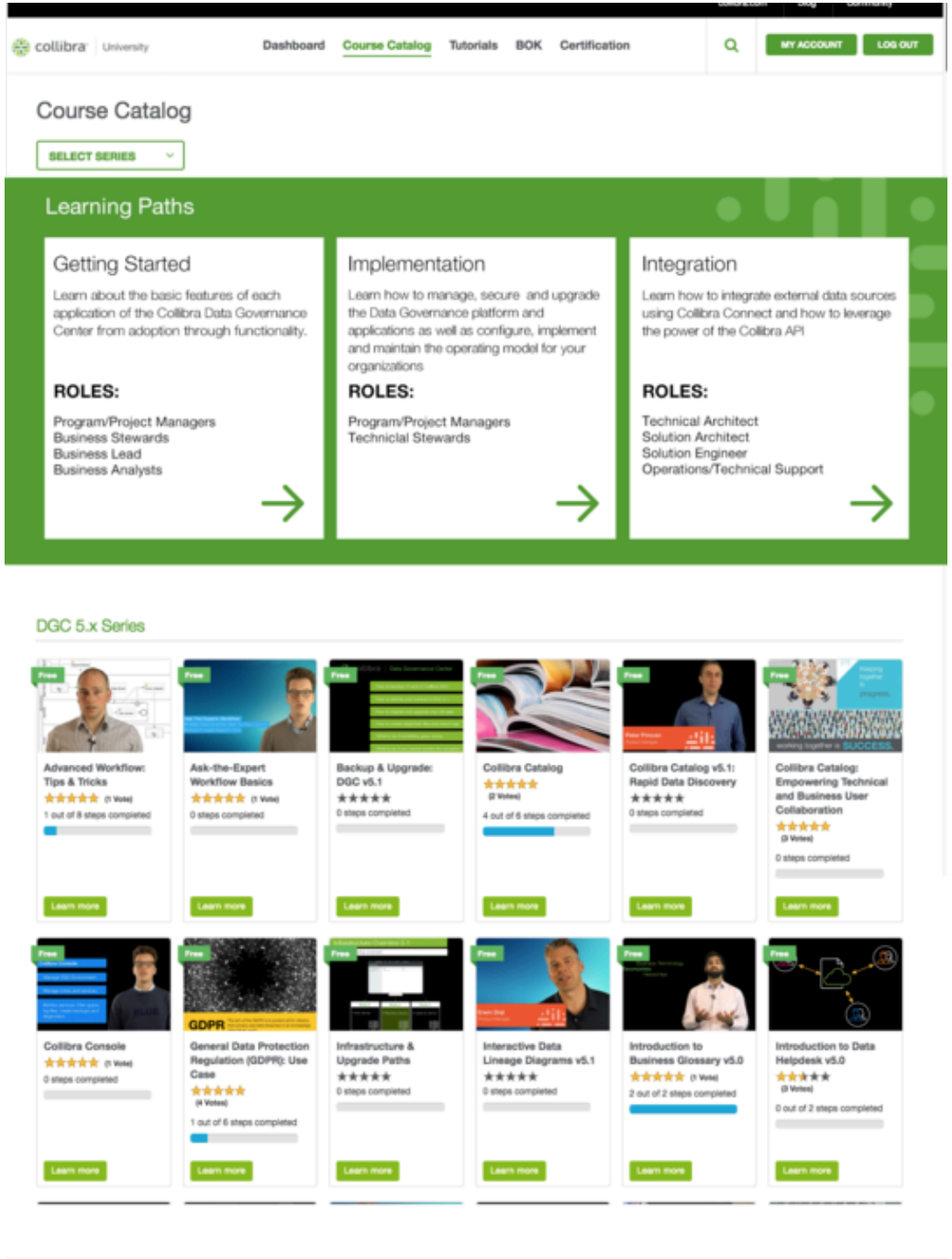
Forrester Wave : Q2 2014



Forrester Wave : Q1 2016

A Fool with a Tool: Collibra Data Education

university.collibra.com



Part 1

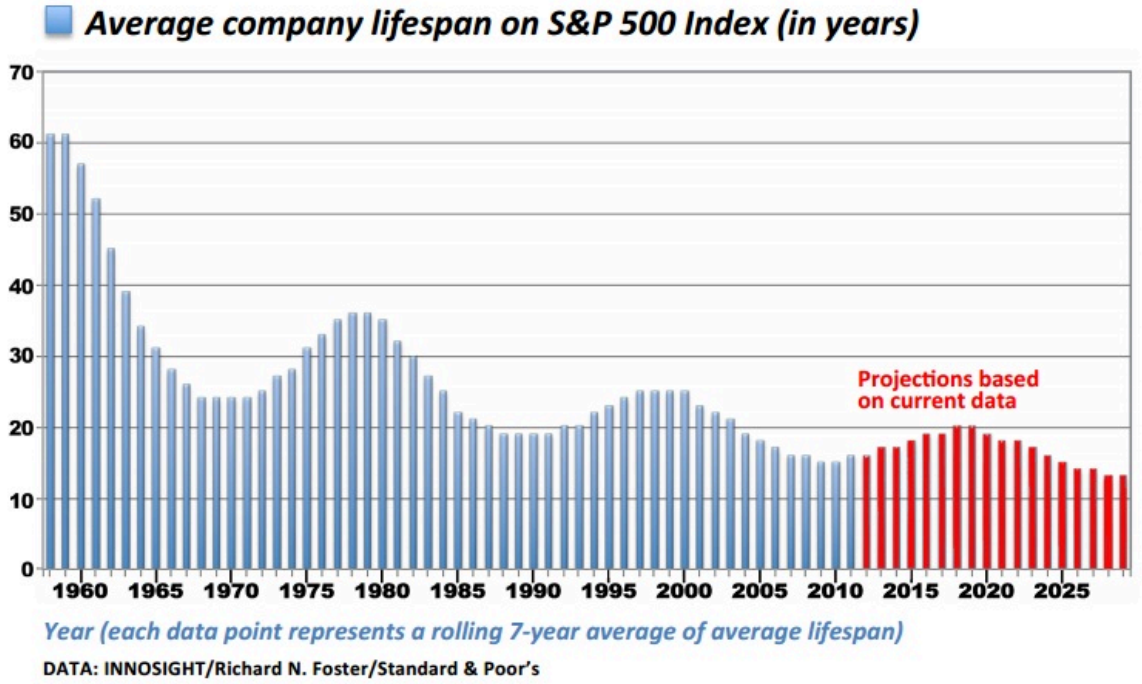
The Chief Data Officer Rises

What do the companies in these groups have in common?

- **Group A**: American Motors, Brown Shoe, Studebaker, Collins Radio, Detroit Steel, Zenith Electronics, and National Sugar Refining.
- **Group B**: Boeing, Campbell Soup, General Motors, Kellogg, Procter and Gamble, Deere, IBM and Whirlpool.
- **Group C:** Facebook, eBay, Home Depot, Microsoft, Office Depot and Target.

Conclusion

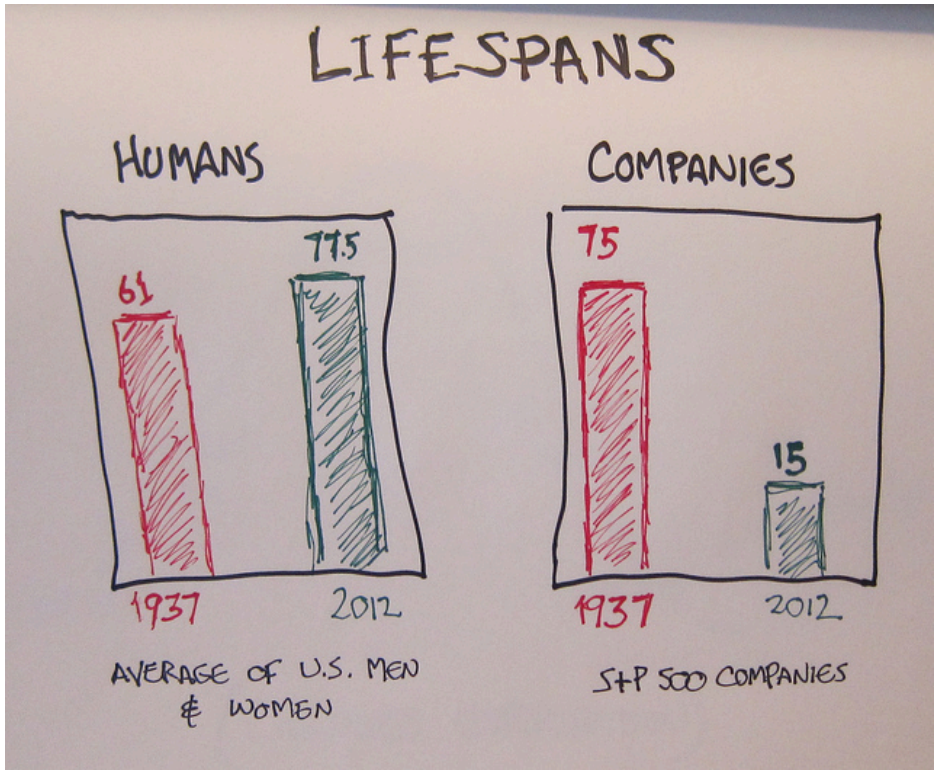
- only 12.2% of the Fortune 500 companies in 1955 were still on the list 59 years later in 2014
- life expectancy of a firm in the Fortune 500
  - 50 years ago : ~ 75 years
  - Today: < 15 years and declining



MIT Technology Review, Sept., 2013

What happened between 1955 and today that caused this ‘creative destruction’?

- Name some compelling events in information technology history
- Order them chronologically
- Try to explain the phenomenon in terms of the events
- E.g.,
  - Invention of the transistor
  - First modern computer
  - Publication of the Internet protocol
  - Launch of the World Wide Web
  - Wikipedia
  - Internet startups: FB, Google, etc.
  - data big bang



Data Big Bang

- Phenomenon: connectivity between
  - Social
  - Knowledge
  - Technology
- Draws curiosity
  - Web Science (Pentland, etc)
  - Big Data Native Market Entrants (23andMe, Uber, Inventure)
- Big-date native entrants
  - 23andMe, Uber, Inventure
  - Enter Bottom up, Low-end and disrupt
  - Pure data strategy
  - Serving “data-citizen” Millennials
- +80% unstructured data or ‘dark energy’

Digital Darwinism

- Disruptive evolution* from analogue to digital business models
  - Entrants disrupt from the low-end, bottom-up
  - Selection is driven by network effects
- Direct access to the consumer
  - Minimizing the middleman / transaction cost
  - Through data aggregators like, e.g., Etsy, Uber, Airbnb, TaskRabbit, Spotify, etc.
  - Perhaps cut out middleman entirely though next-gen internet technology, like distributed ledger technology?
- Drive new value from data as
  - Use real-time feedback data to improve the "brand"
  - Generate second revenue streams from "good" data
  - From consumers as well as any "thing" (via IoT, smart grid, smart cities, via smart contracts)
- Paradox of the "Big Shift"
  - Consumers (especially Millenials, born after ~1982) embrace information sharing for almost every aspect of their life: data citizenship
  - Yet outdated institutional structures continue to inhibit organizational information flows
    - Archaic conceptions of data value and data management (see next)
    - Not adapted to complex network-centric business environments with improved cloud, big data, and security capabilities [2]
  - Soon young data citizens will enter into positions in these institutes and may accelerate the big shift

“Data-driven” is the Holy Grail of business

But Becoming the Fittest is not Easy

74%

of firms say they **want** to be

data-driven

In reality, only

29%

Say they are good at connecting

analytics to action

collibra

Here’s what typically happens

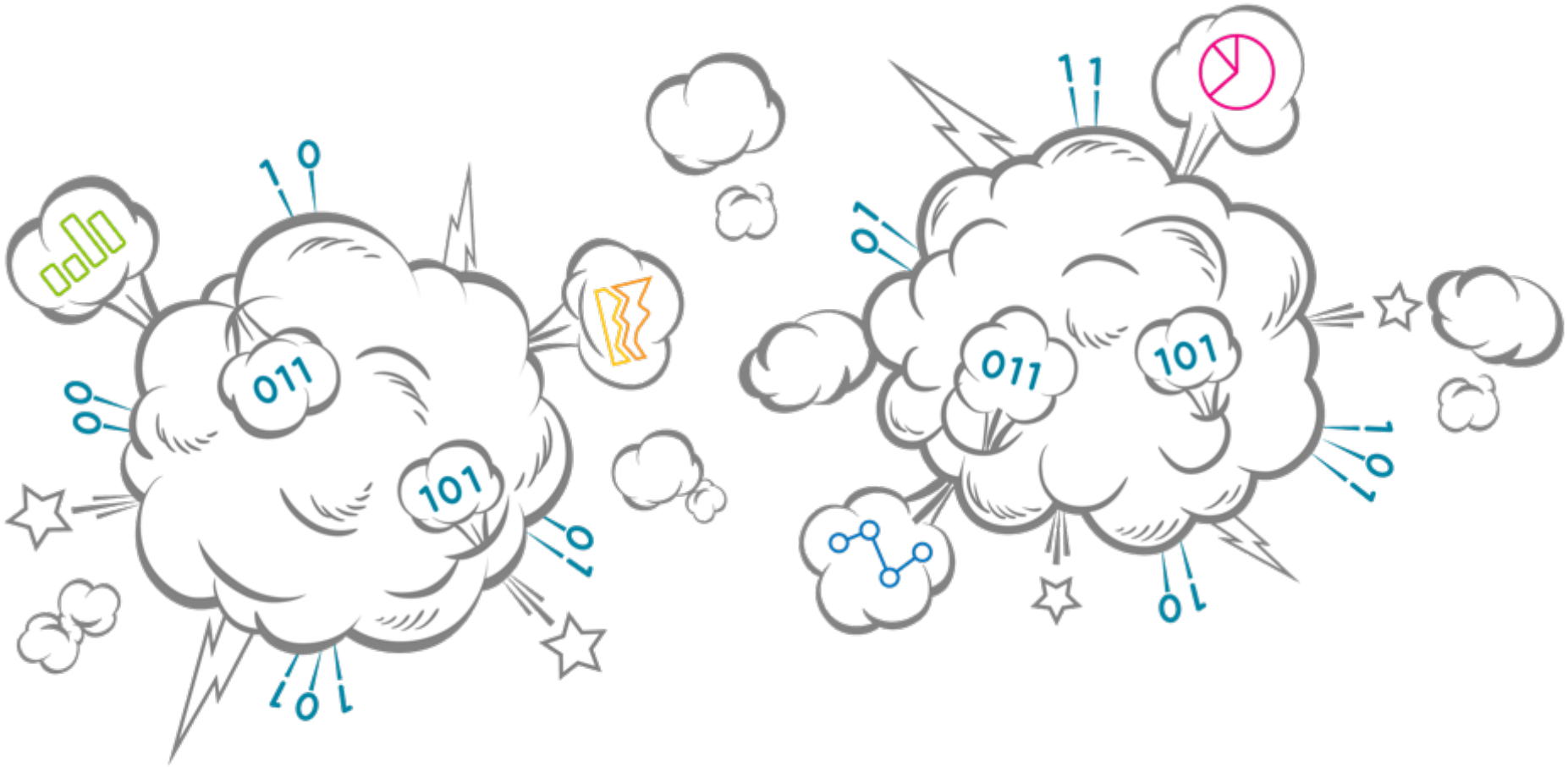
- Hours spent preparing for the meeting
- Collect data from finance, IT and the data lake
- Do your analysis, prepare insights and recommendations
- You’re ready to go

collibra

Same company, same data, different results

collibra

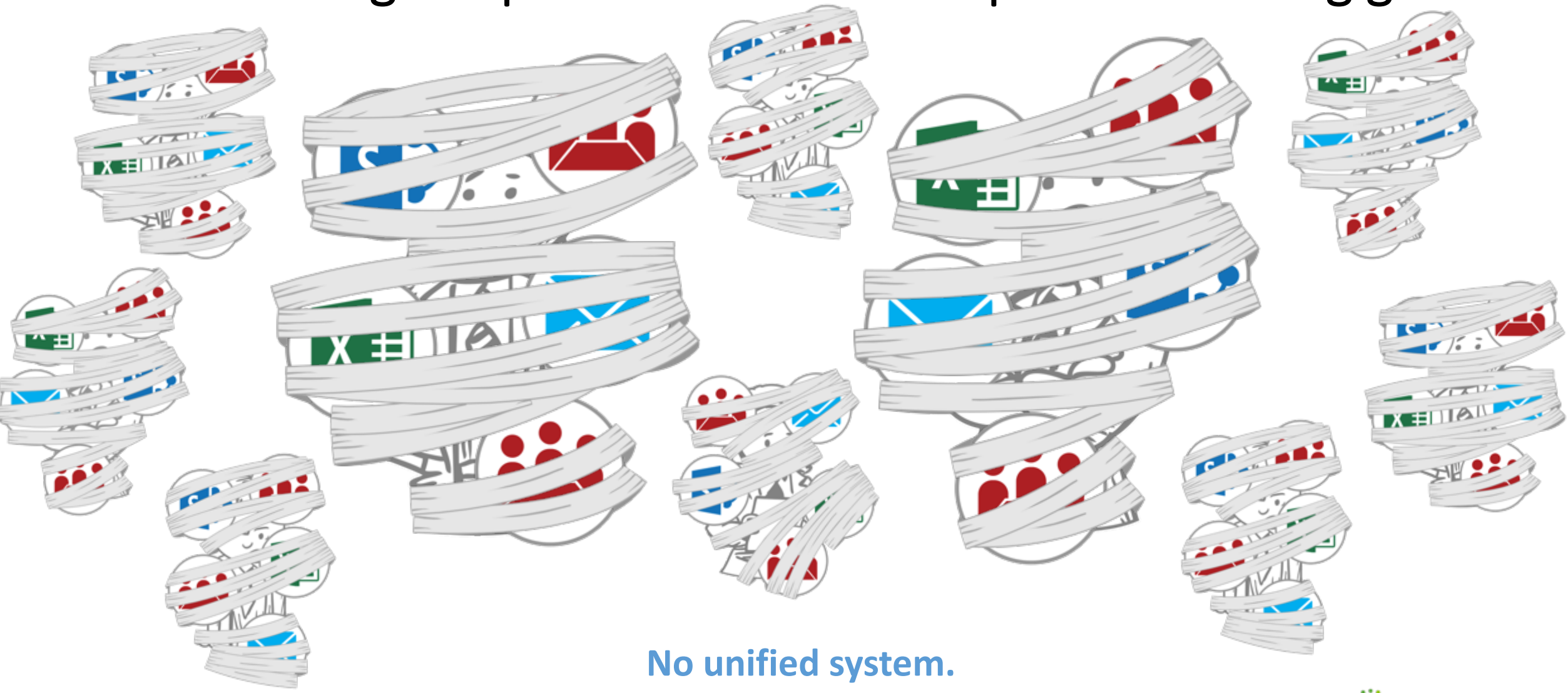
Meetings dissolve into data brawls



You need: trustworthy data, common understanding, complete traceability, transparent data ownership

collibra

Solving the problem with duct tape and chewing gum



No unified system.

collibra

Data Chaos Leads to Data FUD*

I don't trust these numbers

Am I allowed to use this ?

I don't understand this report

Where do I find the data ?

Who can help me with this ?

Data Infrastructure (IT)

TREND
Exploding volume, velocity, and veracity of data

NEED
Manage data complexity

NEED
Data
Collaboration
Understanding
Discovery
Trust

Data Consumers (Business)

TREND
Increased reliance on analytics and regulatory reporting

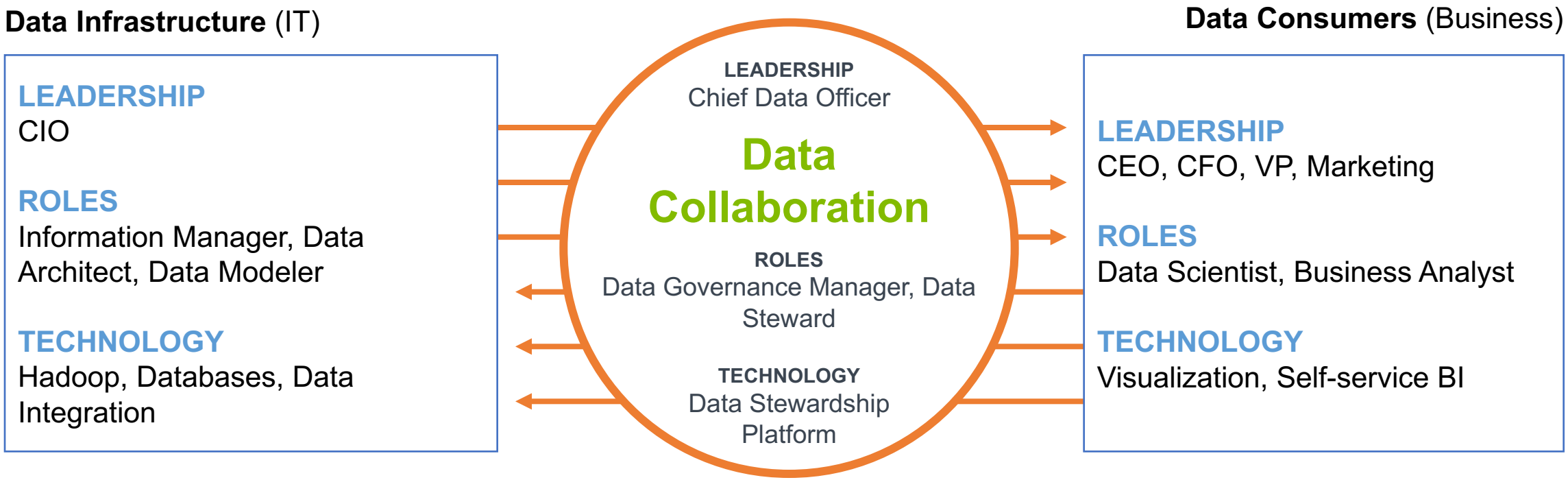
NEED
Trusted data as a business dependency

* fear, uncertainty and doubt

collibra

You Need the Right Level of Control and Trust in Data

Data governance & stewardship provide the right level of control and trust in data

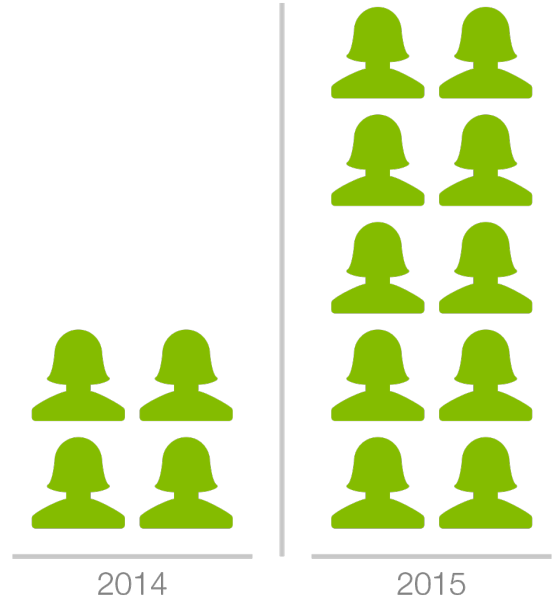


collibra

The Rise of the Chief Data Officer (CDO)

Gartner on Chief Data Officers:

90% of Large Organizations Will Have a Chief Data Officer by 2019



1,000 Chief Data Officers or Chief Analytics Officers Forecast in Large Organizations by the End of 2015, Up From 400 in 2014

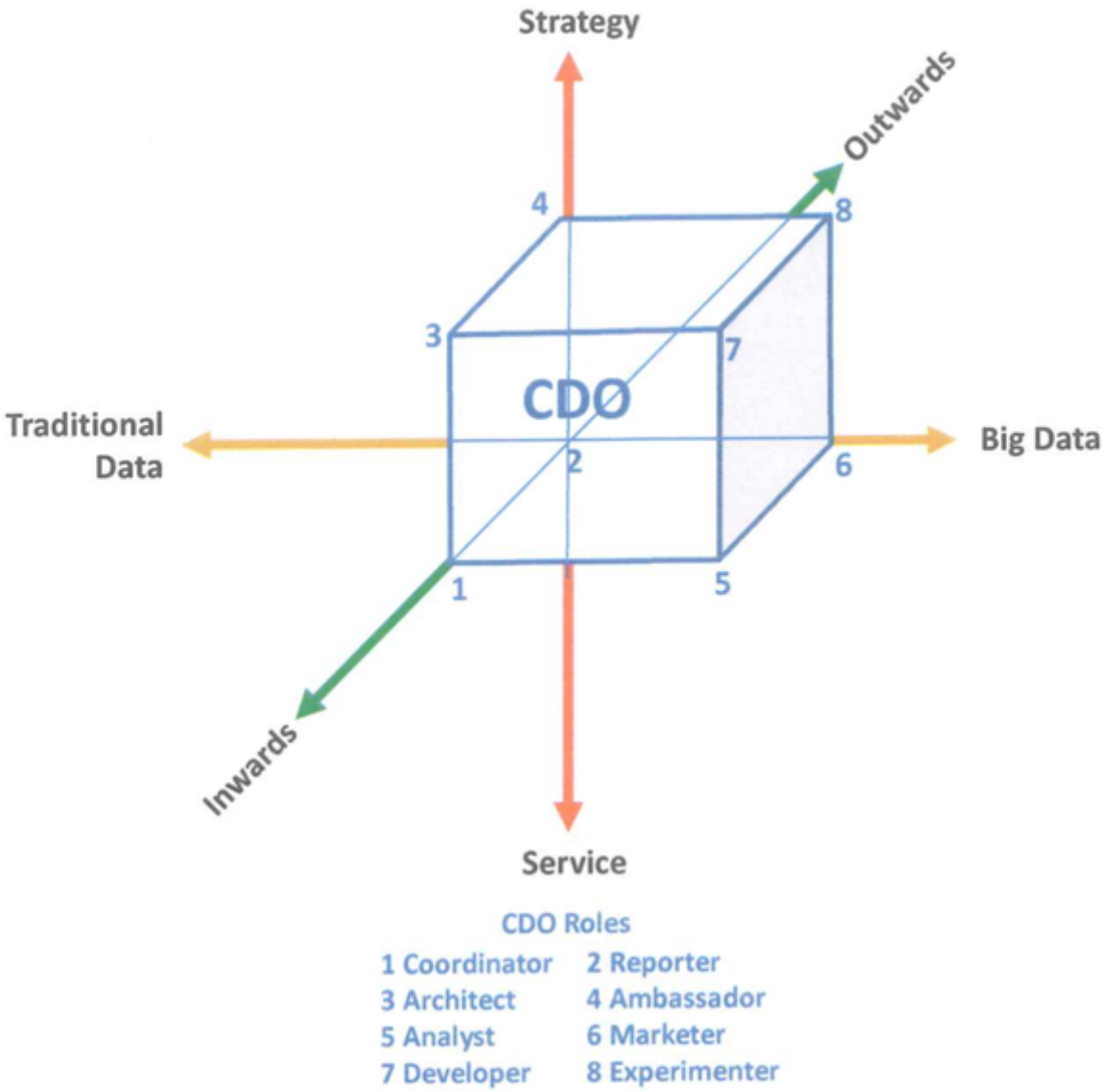
collibra

Role Types for the Chief Data Officer (CDO)

(Lee et al., 2014)

- Dimensions of CDO Roles
  - Collaboration: inwards / outwards
  - Data Space: traditional data / big data
  - Value Impact: service / strategy

- Reporting:
  - 30% to CDO
  - 20% to COO
  - 10% to CFO



CDO Assessment

Join our MIT Sloan CDO Research
https://university.collibra.com/cdo-survey/

Table 1:		
Collaboration Dimension: Inward vs. Outward *High score for #1 and #2 implies inward direction.* *High score for #3 and #4 implies outward direction.*	**Assessment Score (1-7)** **1** Strongly disagree **4** Neutral **7** Strongly agree	Assessment discussion notes Why section: explain reason for Assessment
1. It is critical that our organization uses data effectively for internal business operations.	*3*	*We do this well, thus, not critical at this point.*
2. Our company has the opportunity to significantly improve internal operations.	*3*	*Maintain what we do well.*
3. It is critical that our organization collaborates with other value chain enterprises, such as suppliers, customers, distributors, or competitors.	*6*	*We need to know our suppliers and customers much better.*
4. Our organization's success is critically interlocked with other companies, market changes, external situations or environments.	*7*	*Our procurement can be vastly improved with better understanding our suppliers.*
Data Space Dimension: Traditional Data vs. Big Data *High score for #5 and #6 implies traditional data; High score for #7 and #8 implies Big Data.*		
5. Our organization's transactional data should be more effectively used to address the enterprise's needs	*6*	*We need to know more about aggregated amounts of materials for different suppliers.*
6. It is critical for our organization to use the transactional data in an integrated fashion across different business areas.	*7*	*To negotiate with our suppliers, we need to get all divisions to use the information we have already in a consistent way.*
7. Our company needs to identify opportunities for using big data and data analytics.	*5*	*We may not be there yet to go for this direction.*
8. It is critical for our organization to understand external sources of data, such as social media for engaging customers.	*6*	*Our customers might be ready for new sources in the future and we need to explore and exploit social media.*
Value Impact Dimension: Service vs. Strategy *High score for #9 and #10 implies Service; High score for #11 and #12 implies Strategy.*		
9. Our organization's data efforts should be largely initiated or requested by the enterprise's business units.	*4*	*We do this well.*
10. It is critical for our organization to improve the efficiency of the data service for operation.	*5*	*We can still improve, but we do well on serving data for the internal business units.*
11. Our organization's data efforts should be largely initiated by the need for changes in the way we do business.	*6*	*We can use the data for changing the way we do procurement planning with our global suppliers.*
12. Our organization must achieve its strategic business goals with better data.	*7*	*We must figure out who our best business customers are and set different strategies for different customers.*

Hierarchical Data Management

- Formal
- Traditional data focus
- Inward Focus:
  - Improve Internal/external coordination
  - Understand customer
  - Predict next transaction
- Central Servicing
  - MDM, DWH, DM, Dashboards
  - Tedious Waterfall
  - Comprised by Obsolete Cost assumption
- Elite Consumer Base
  - C-level

Hierarchical Data Governance (1)

- Wikipedia:

  "a set of processes that ensures that important data assets are formally managed throughout the enterprise. Data governance ensures that data can be trusted and that people can be made accountable for any adverse event that happens because of low data quality".

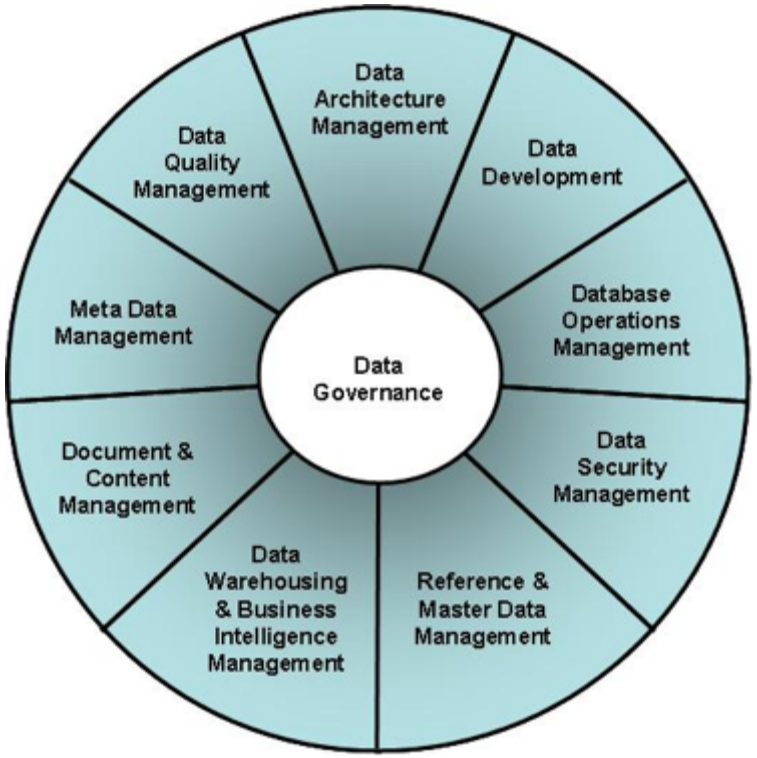
- three objectives of data governance: (i) maximize data trust and (ii) remediate low data quality by (iii) holding people accountable.
- Clearly inspired by Total (Data) Quality Management
- Problems
  - Suggest 'policing' rather than 'empowerment'
  - DQ => Trust versus DQ + Usage + Provenance => Trust

Hierarchical Data Governance (2)

- The bigger picture proposed by DMBOK:

  “The exercise of authority, control and shared decision making (planning, monitoring and enforcement) over the management of data assets”

- Problems:
  - Restricted to the 9 Data Management Organizational functions
  - Focus on ”coordinator” (inward / traditional / service) CDO Role
  - Assumes command-and-control
  - Unclear:
    - How to establish a nuanced notion of Trust, Transparency and Participation?
    - How to allow grass rooting, i.e. local or allied data initiatives?



Data Value Hierarchies, Networks & Hybrids

Digitalization of Physical, Consumerization of Tech, Decentralization, Digitalization of Trust

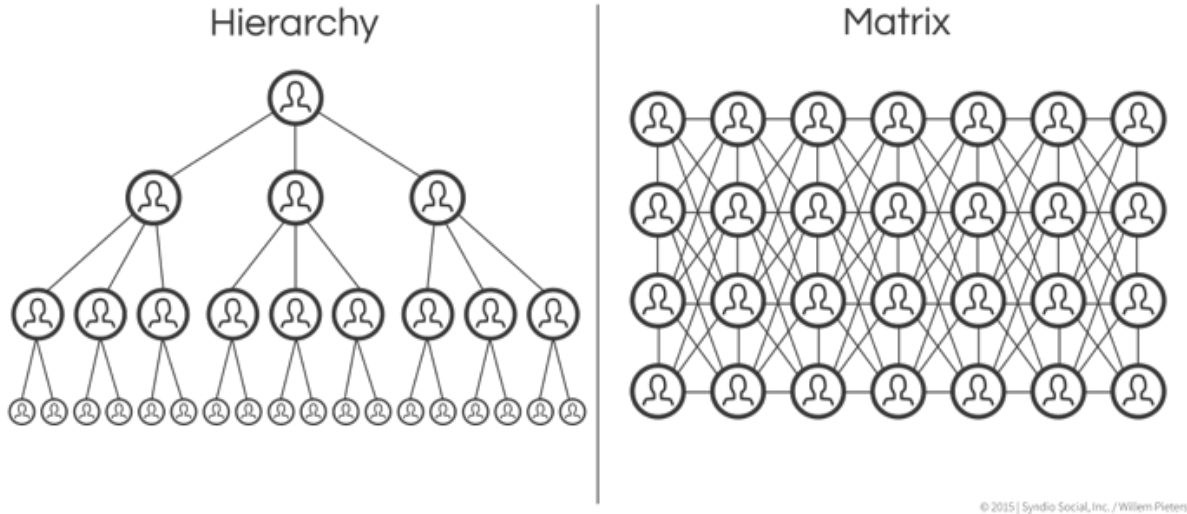
Hierarchy	**Network**	Network peers provide ideas, feedback but also service (uber driver analogy data scientist)
Product Ownership	Service Access	Example: Uber doesn't own. It only dispatches information about rolling material to riders and focus over lifetime value retention. *Data analogy:* access to data more important than owning as cost of IS is marginal and replaced by data value appreciation
Passive resources (material, goods)	Active resources (data, consumer)	
Value-in-exchange	Value-in-use	
Acquisition	Retention	Example: Saas, Netflix, Costco, etc. *Data analogy:* From formal roles and responsibilities to support internal process to social capital based trust
Process and function	Social capital	
Provider push	Consumer pull	Example: Feedback, mods on games, user participation, A/b testing etc. *Data analogy*: data helpdesk

Part 2

The Data Universe Expands

Shift in Data Governance Approaches

- Digital forces pose gigantic risk as well as opportunity on organizations, balance needed between:
- Hierarchical data governance (system of record)
  - CDO as a Coordinator: Inward-oriented / Traditional Data / Service
  - Defensive: Risk-driven
  - Scarcity: Few consumers, few producers
  - Compromises on old obsolete cost assumptions of digital power
  - Use of digital optimizes to some extent
  - Not scalable for big data by larger 'data scientist' populations
- Networked data governance (systems of engagement)
  - CDO as an Experimenter: Outward / Big Data / Strategy
  - Offensive: Value-driven
  - Abundance
  - Many Producers*(Data Democratization*)
    - Eliminate Breadlines
    - Consumerization of BI and cheap digital power
    - Many serve many
    - Supports customer
  - Many Consumers *(Data Amazonification*)
    - Access, SLA, Trust, Secure Cloud, etc



System of Record vs. System of Engagement (AAIM 2017)

System of Record

- Purpose: control and regulate
- Top-down design around discrete pieces of information (“records”)
- Decomposition in ‘black boxes’
- Presumes ‘big picture’
- Examples: SFDC, Workday, ServiceNow, Atlassian

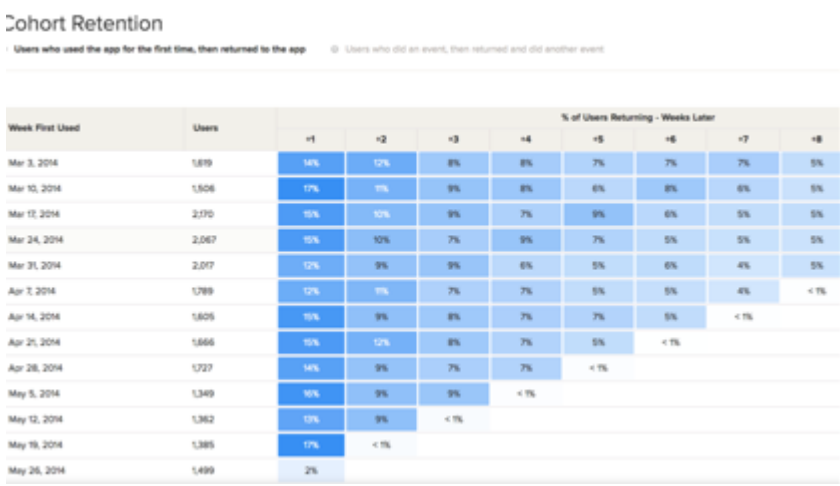
System of Engagement

- Purpose: innovate
- Bottom-up, decentralized, incorporate technologies which encourage peer interactions, leveraged by cloud technologies
- Seed Model
- Emergence of complex system
- Examples: Slack, Confluence

Consideration	Systems of Record—Enterprise Content Management	Systems of Engagement—Social Business Systems
Focus	Transactions	Interactions
Governance	Command & Control	Collaboration
Core Elements	Facts, Dates, Commitments	Insights, Ideas, Nuances
Value	Single Source of the Truth	Open Forum for Discovery & Dialog
Performance Standard	Accuracy & Completeness	Immediacy and Accessibility
Content	Authored	Communal
Primary Record Type	Documents (Text, Graphics)	“Conversations” (Text-based, Images, Audio, Video)
Searchability	Easy	Hard
Usability	User gets trained on system and has access to follow-on support	User “knows” system from consumer experience
Accessibility	Regulated & Contained	Ad Hoc & Open
Retention	Permanent	Transient

Big Data Analytics Challenges

- Where everybody has data scientists: predict next transaction is not competitive anymore
- from 'predict next transaction' to life-long relation building and value creation
  - reduce search and navigation for customer with better apps
  - crowd sourcing to cross compare with and learn from other customers (Opower, INRIX, zillow)
- get trust from customer through branded non-intrusive apps: personal health monitoring, Nest
- Retention analysis example



Digitalization of Trust Challenges

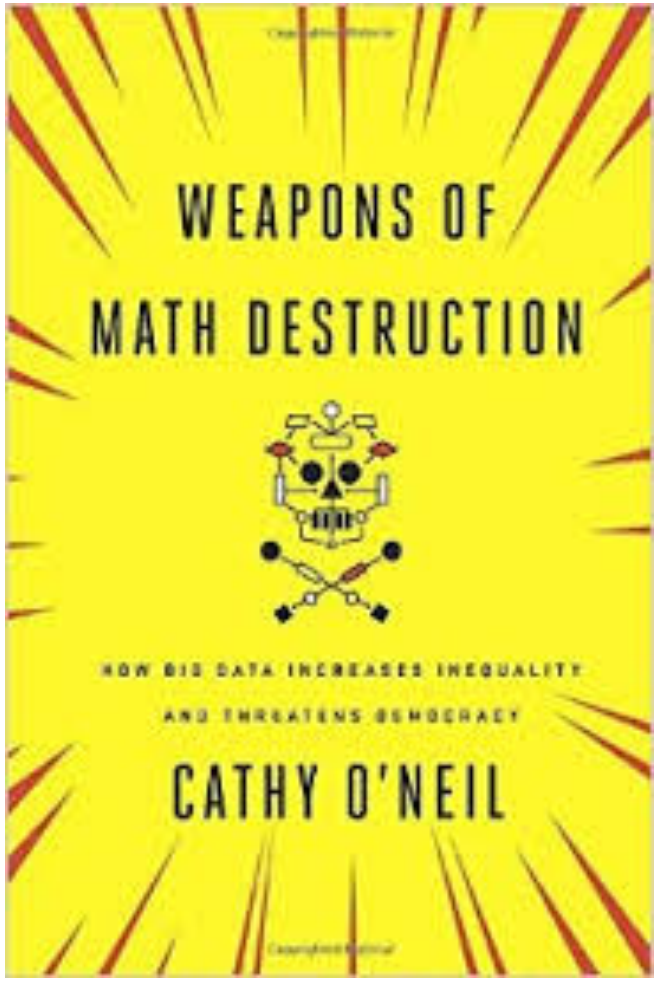
- In Hierarchical Data Governance, trust is
  - established by a centrally sanctioned competence center
  - Or external appointed trustees with formal roles: steward, owners, architects
- In Networked Data Governance, trust is more complicated:
  - Authenticity: is the data factual or opinioned?
  - Intention: does this data have good intentions? Can I use it without peril? Hidden privacy concerns I should be aware of?
  - Assess expertise or quality: are people involved skilled or certified stewards?
  - Is it accurately representing our business reality, i.e. customer base?
  - Is it complete and up to date?
  - Has it be certified through standard process?

Danger of the old paradigm models

- Weapons of Math Destruction (WMD) are models

- Threaten to destabilize
  - Equality
  - Democracy

- Traits of WMDs
  - Opaque
  - Unregulated
  - Uncontestable
  - …hence : **ungoverned**



Preliminary Conclusions

- Digital forces have digitally empowered individuals in the organization
- Hybrid data governance approach should combine
  - Top-down governance of critical data assets to enhance internal coordination
  - Networked peer-driven empowerment to drive ‘serendipity’
  - On a shared platform
- Key challenges are:
  - Digitalization of trust with focus on social capital
  - Big data analytics that drives life-time value for customer
  - Data Valuation based on Usage
  - Legacy of oblique, unregulated and incontestable models
  - Recognize CDO Leadership and Role transition

Part 3

A Lens on the Data Universe

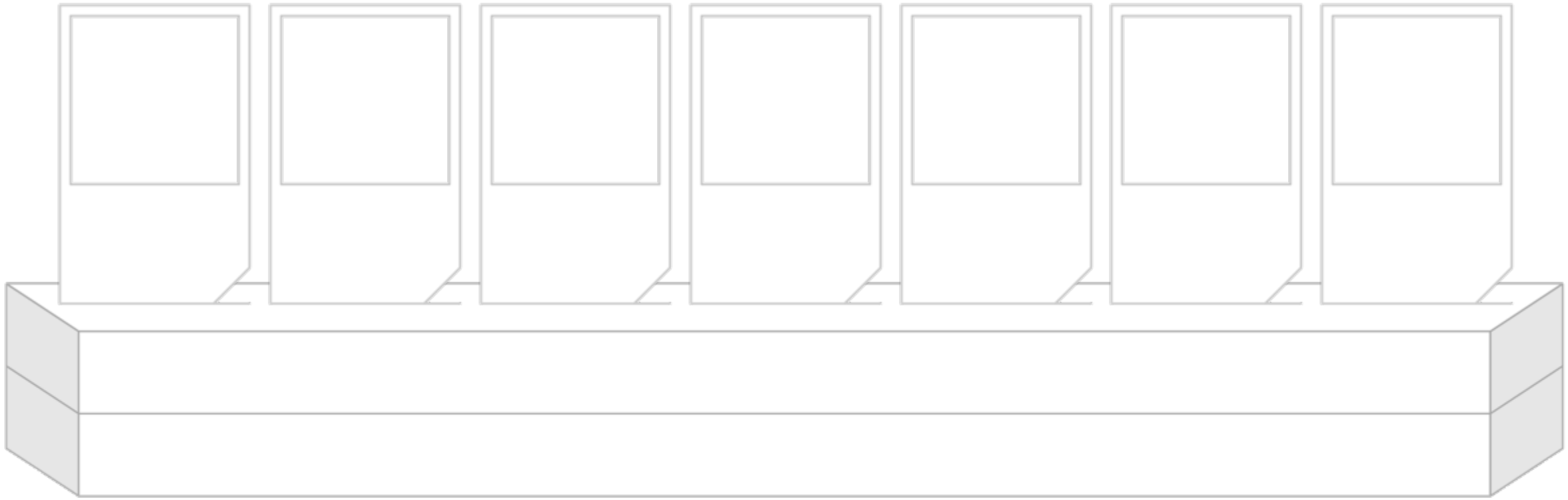
The System of Record for Data Assets

the authoritative source of information for any given data asset *used by* (hence *valuable* for) the organization

Know where the data comes from

Know what the data means

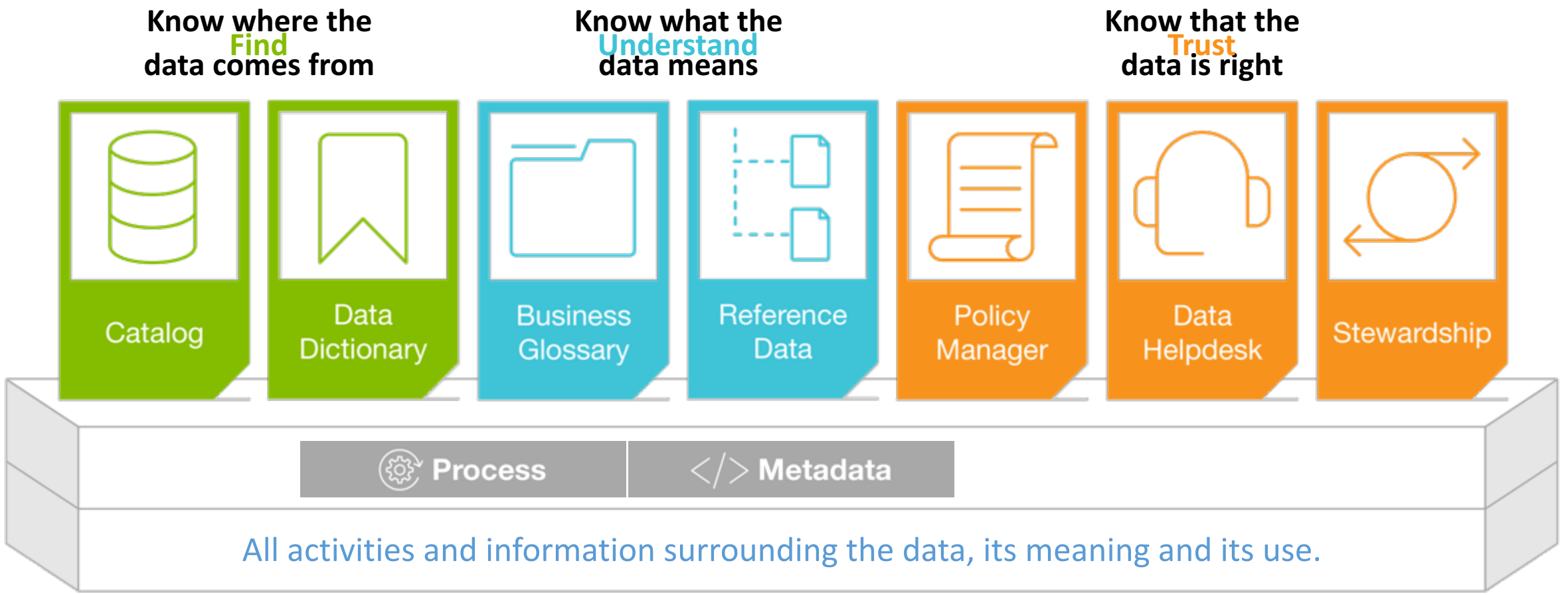
Know that the data is right



collibra

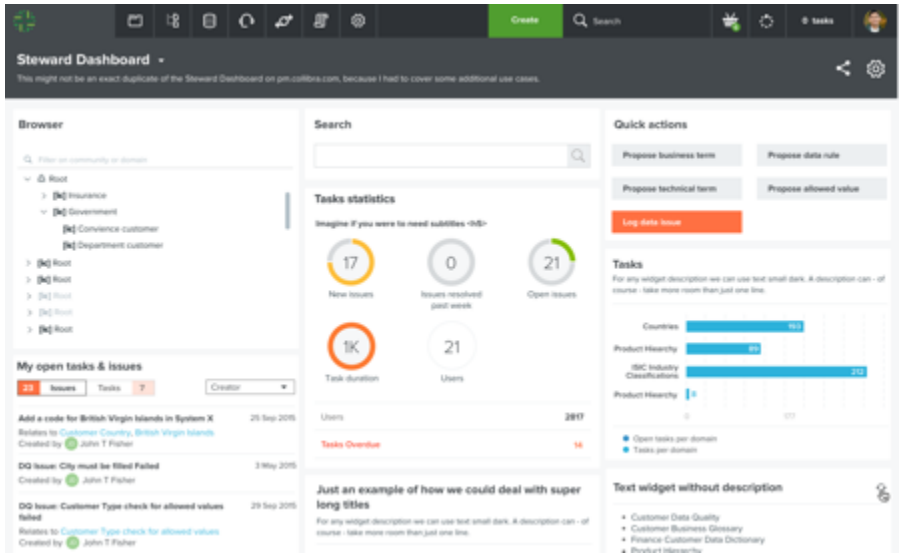
The System of Record for Data Assets

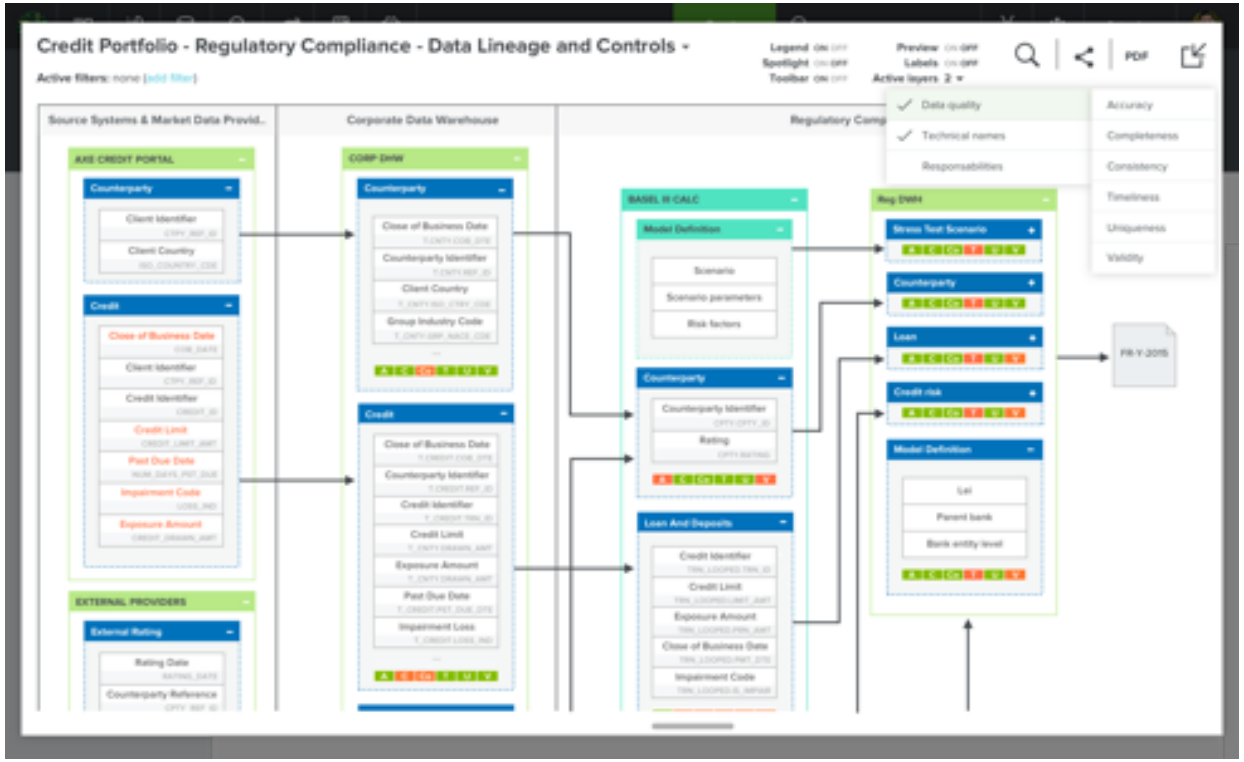
the authoritative source of information for any given data asset *used by* (hence *valuable* for) the organization



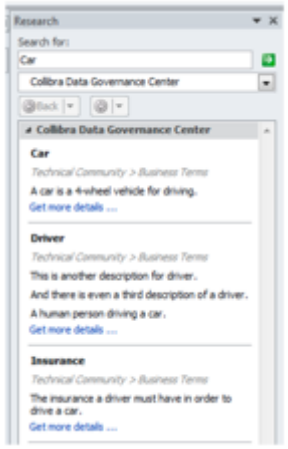
collibra

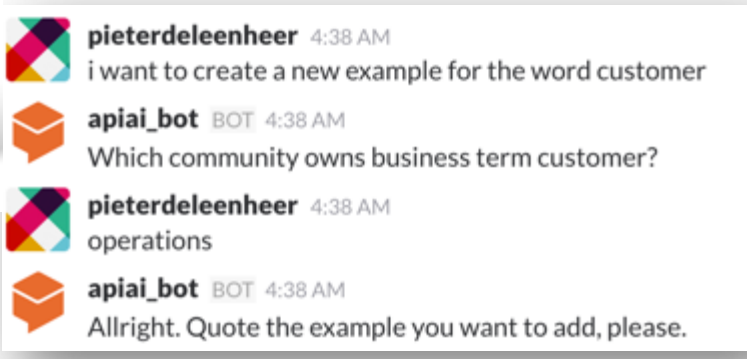
Demo


Steward Dashboard
Browser
Search
Quick actions
Tasks statistics
Tasks
My open tasks & issues
Just an example of how we could deal with super long titles
Text widget without description


Credit Portfolio - Regulatory Compliance - Data Lineage and Controls
Source Systems & Market Data Provid...
Corporate Data Warehouse
Regulatory Comp...
AXE CREDIT PORTAL
CORP DWH
BASEL III CALC
EXTERNAL PROVIDERS


Quality Dashboard
Asset Quality
Customer Domain


Research
Search for:
Car
Collibra Data Governance Center
Car
Driver
Insurance


pieterdeleenheer 4:38 AM
i want to create a new example for the word customer
apiai_bot BOT 4:38 AM
Which community owns business term customer?
pieterdeleenheer 4:38 AM
operations
apiai_bot BOT 4:38 AM
Allright. Quote the example you want to add, please.

Recommended Reading

- Books:
  - O'Neil, C.: Weapons of Math Destruction
  - Franks, B.: Taming the Big Data Tidal Wave
  - Sundararajan, A.: The Sharing Economy
  - Pentland, S.: Social Physics: How Good Ideas Spread
  - Zittrain, J.: The Future of the Internet
  - Tunguz, T.; Bien, F. (2016) Winning with Data
- Articles:
  - Lee et al. (2014) A Cubic Framework for the Chief Data Officer: Succeeding in a World of Big Data. MIS Quarterly Executive 13:1
  - AAIM, Systems of Engagement and the Future of Enterprise IT (2017)
  - http://mitiq.mit.edu/IQIS/Documents/CDOIQS_201177/Papers/05_01_7A-1_Laney.pdf
  - http://si.deis.unical.it/zumpano/2004-2005/PSI/lezione2/ValueOfInformation.pdf
  - http://dupress.deloitte.com/dup-us-en/topics/emerging-technologies/the-burdens-of-the-past.html
- Blog Posts
  - https://www.collibra.com/blog/unleash-the-data-democracy-5-misconceptions-of-data-governance/
  - https://www.collibra.com/blog/the-rise-of-the-chief-data-officer-cdo/
  - https://www.collibra.com/blog/blognew-years-resolution/
  - https://www.collibra.com/blog/data-lineage-diagrams-paradigm-shift-information-architects/