



HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

# Data governance and visualization

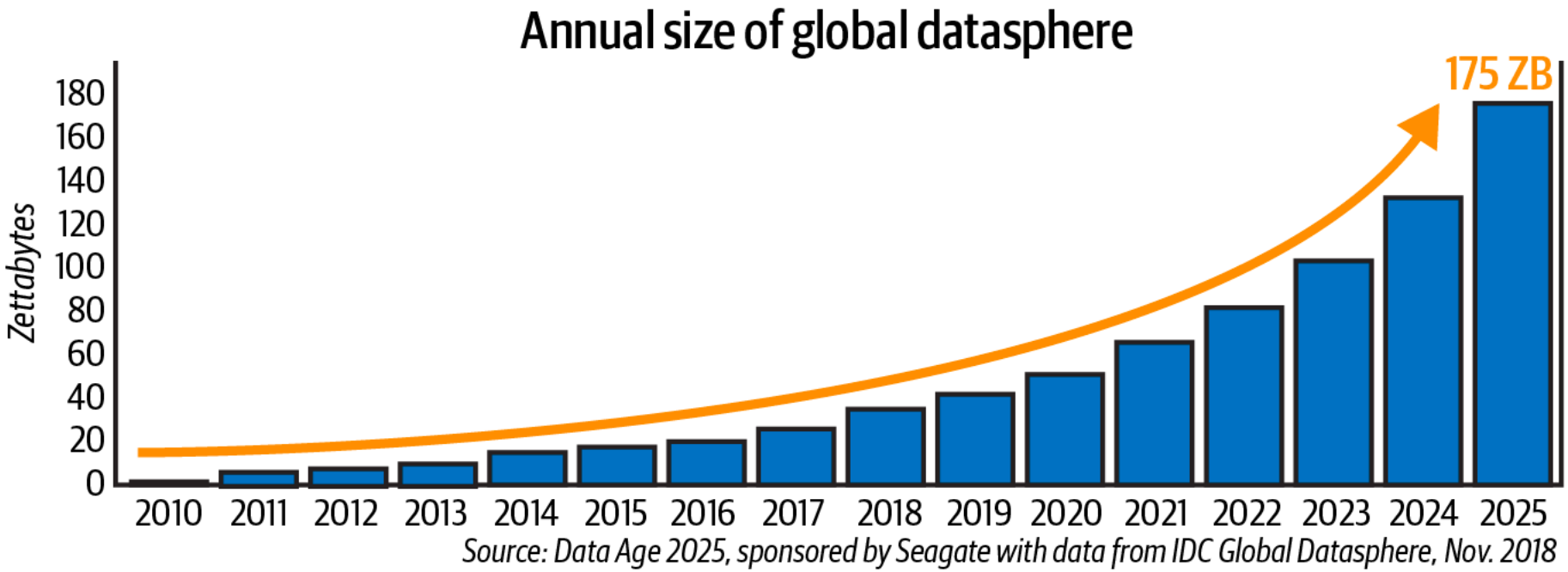
Chapter 1

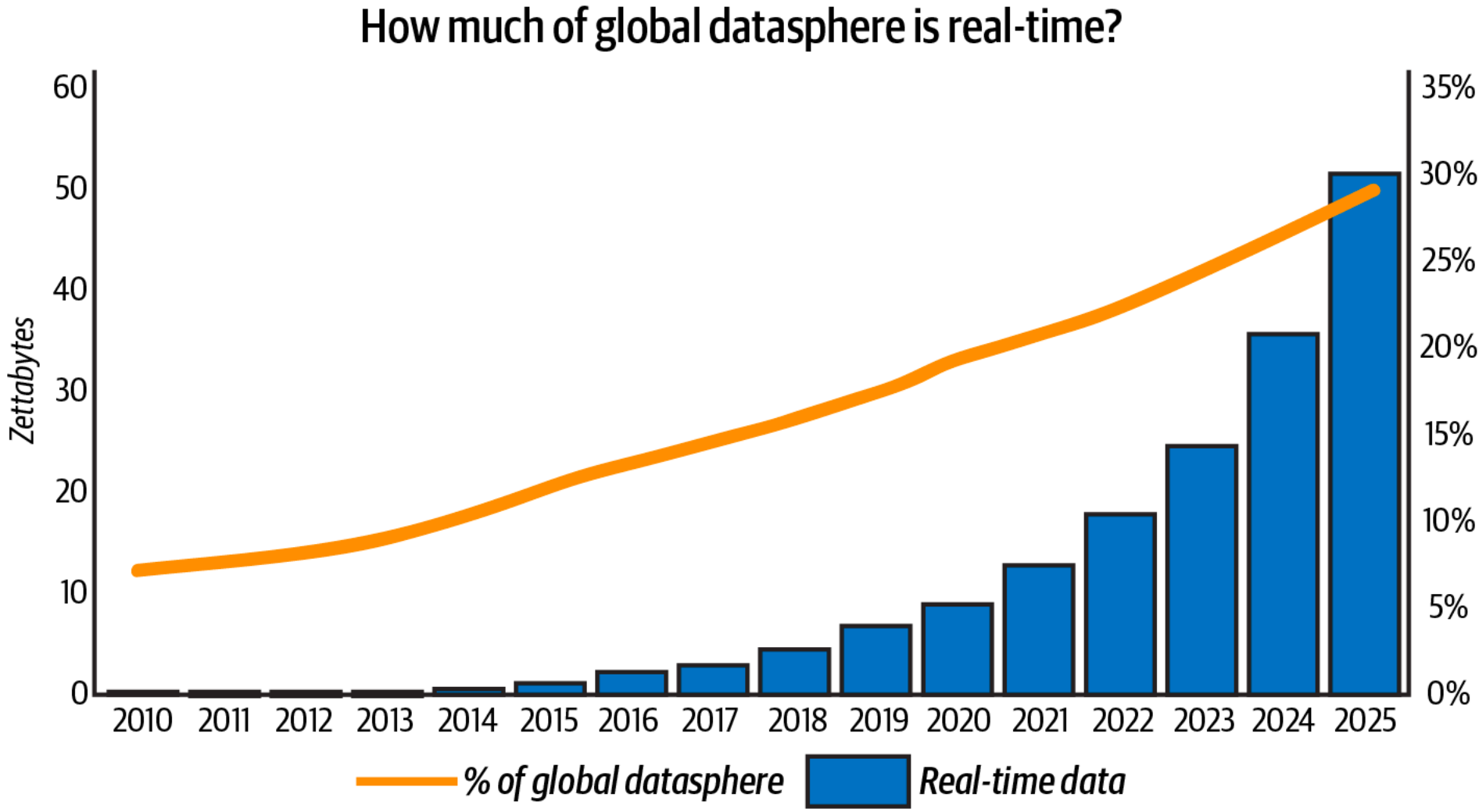
## Introduction to data governance

Viet-Trung Tran

# Data governance is becoming more important

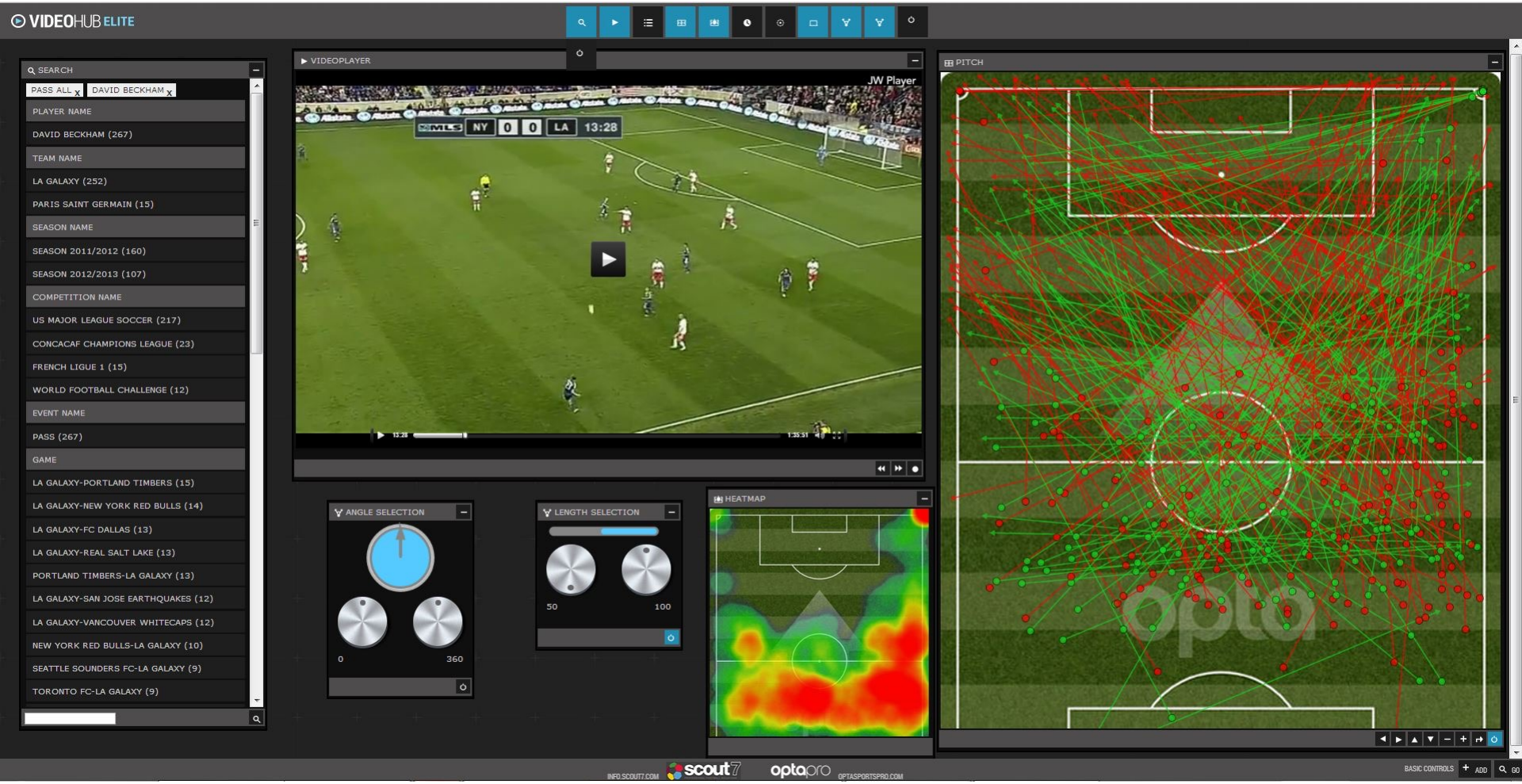
# How big is big data?





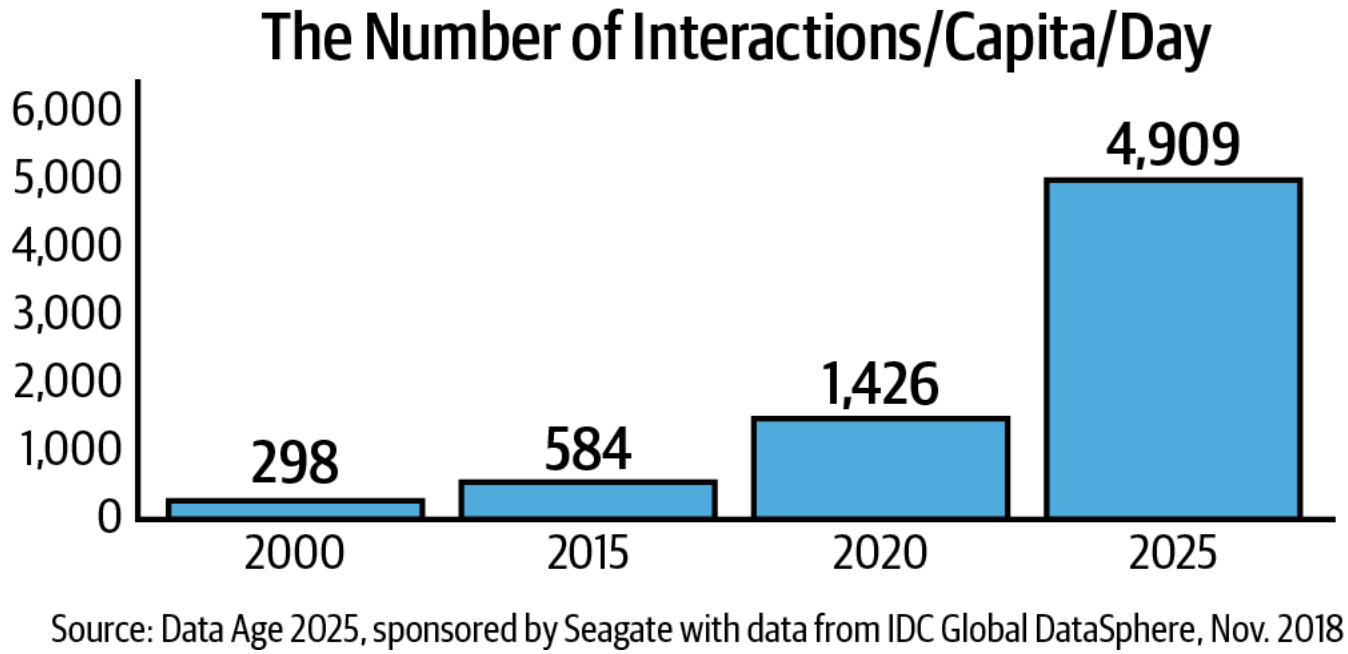

*Source: Data Age 2025, sponsored by Seagate with data from IDC Global Datasphere, Nov. 2018*

# Advanced Data Collection in Sports



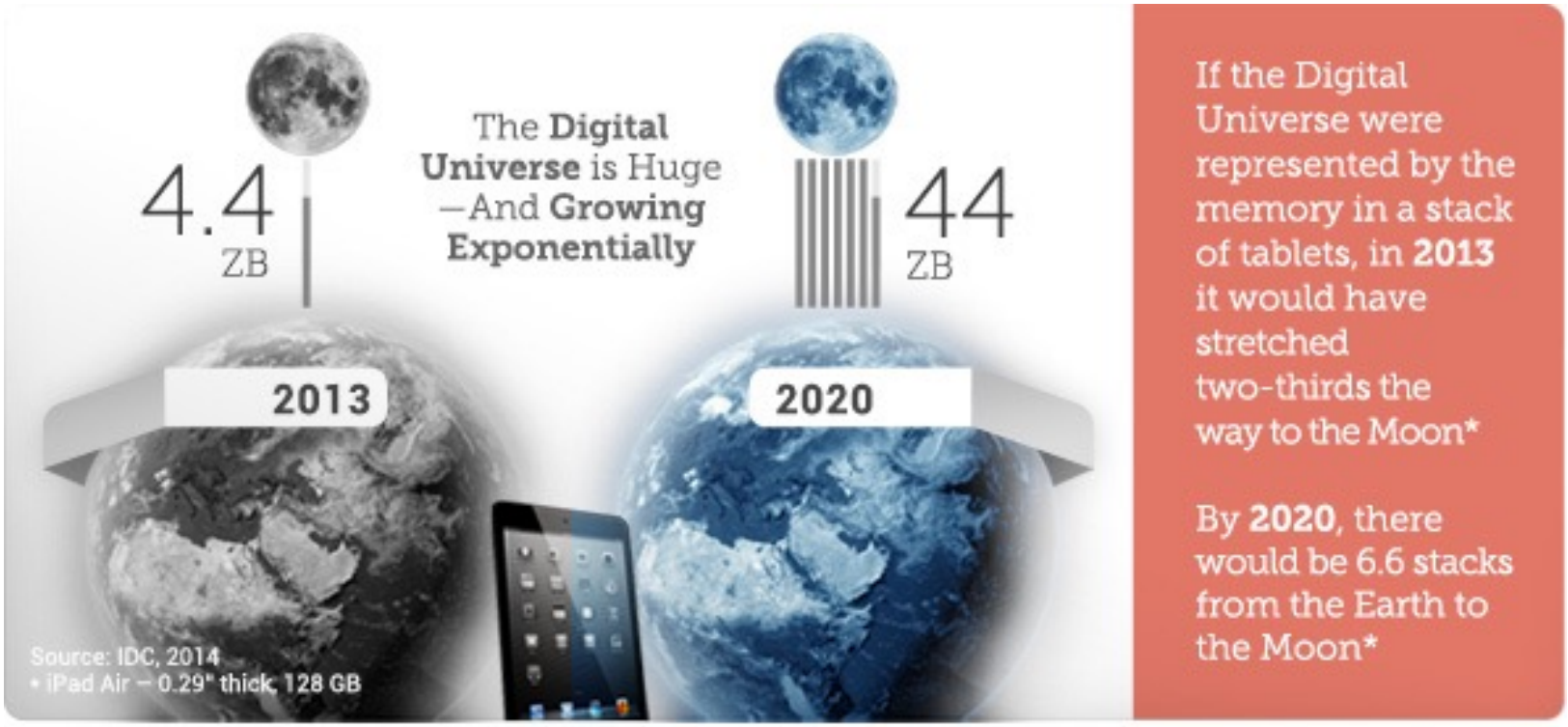
# More Kinds of Data (Including More Sensitive Data) Are Now Being Collected

- One digital interaction every eighteen seconds



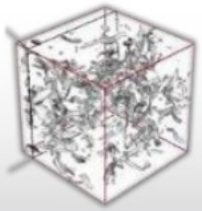

Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov. 2018

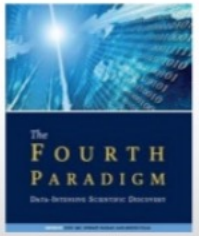
# How big is big data?



# Data science: The 4th paradigm for scientific discovery

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$






| Experimental | Theoretical | Computational | The Fourth Paradigm |
|---|---|---|---|
| Thousand years ago | Last few hundred years | Last few decades | Today and the Future |
| *Description of natural phenomena* | *Newton's laws, Maxwell's equations...* | *Simulation of complex phenomena* | *Unify theory, experiment and simulation with large multidisciplinary Data*<br>*Using data exploration and data mining (from instruments, sensors, humans...)*<br>*Distributed Communities* |

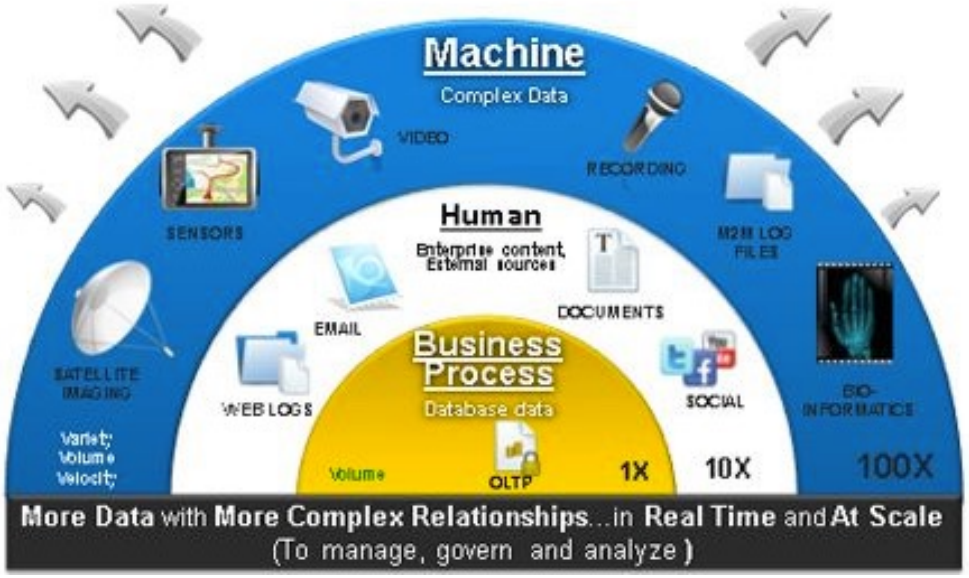
# Big data in 2008

http://www.wired.com/wired/issue/16-07

September 2008

# Big data sources

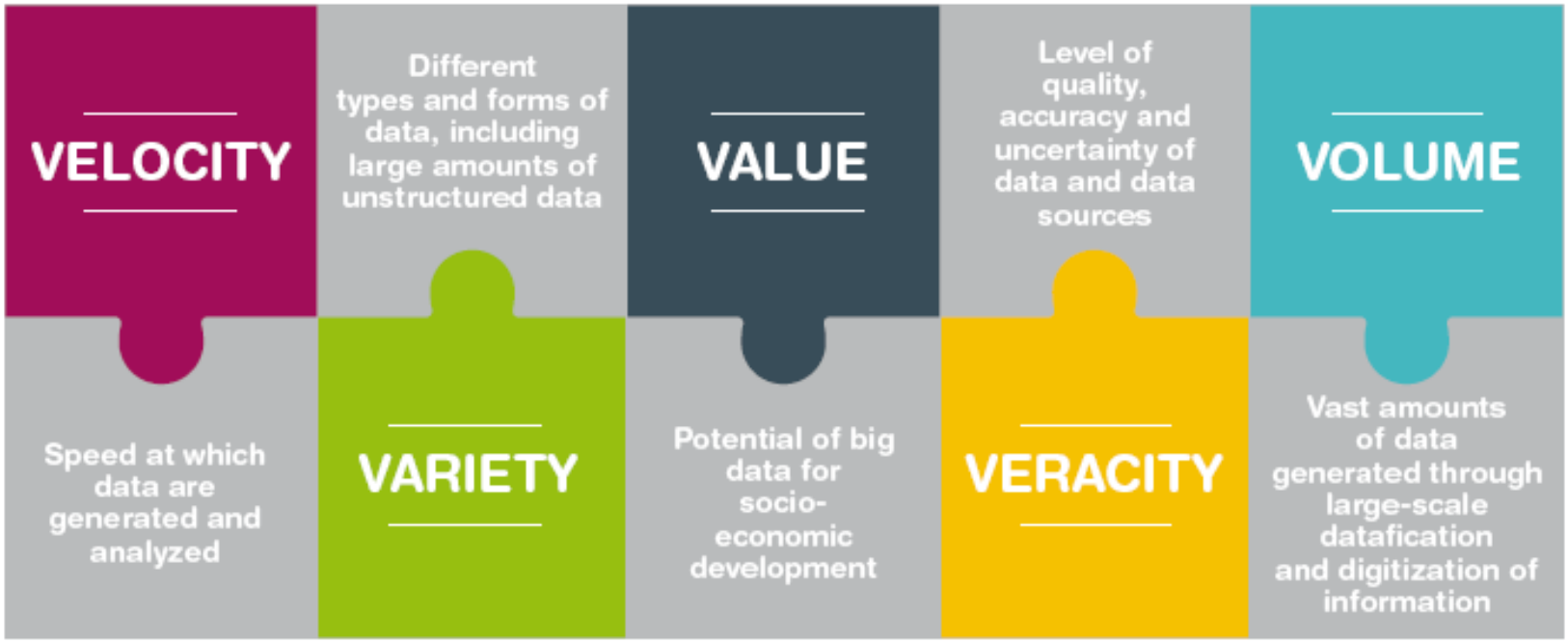
- E-commerce
- Social networks
- Internet of things
- Data-intensive experiments (bioinformatics, quantum physics, etc)



# Data is the new oil



# Big data 5'V



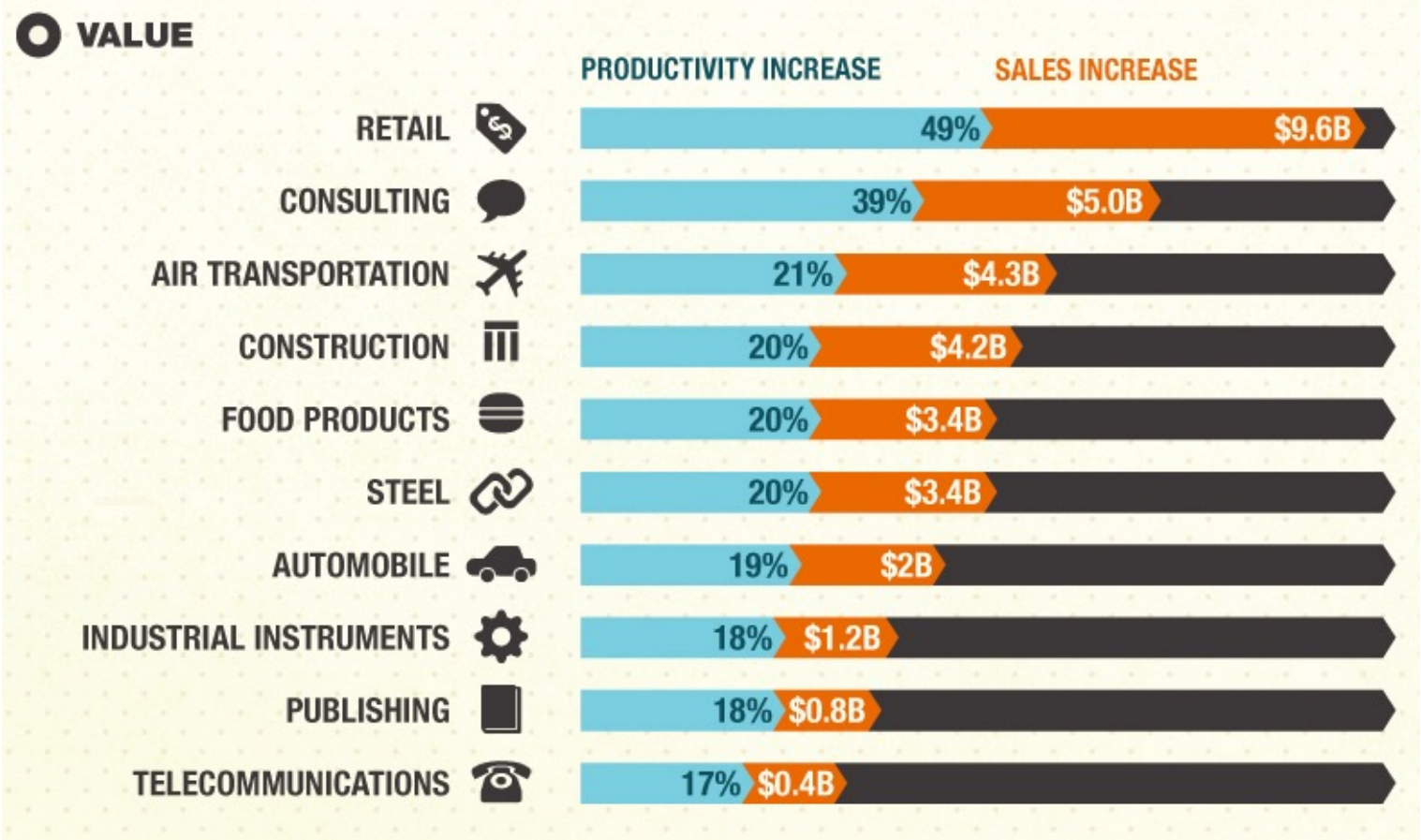

Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them (wikipedia)

# Data value

- ***Data is the most valuable asset in an organisation after its people***
- ***Data is critical to the running of business functions and processes***
- ***Data need constant vigilance and effort to maintain data quality***


Source: sciphilos.info

# Big data – big value




source: wipro.com

# Other facts

- The Number of People Working and/or Viewing the Data Has Grown Exponentially
  - A report by Indeed shows that the demand for data science jobs had jumped 78% between 2015 and 2018.
  - IDC also reports that there are now over five billion people in the world interacting with data, and it projects this number to increase to six billion (nearly 75% of the world's population) in 2025.
  - Companies are obsessed with being able to make "data-driven decisions,"
- New Regulations and Laws Around the Treatment of Data
  - EU's General Data Protection Regulation (GDPR) regulates data, data collection, data access, and data use.
- Ethical Concerns Around the Use of Data
  - 2018. a man was struck and killed by a self-driving car. Who was responsible?
  - 2014, Amazon developed a recruiting tool, however, it was found that the tool discriminated against women.

# Introduction to data governance

# Data governance

- **Data governance is a collection of processes, roles, policies, standards, and metrics that ensure the effective and efficient use of information**
  - for the end-to-end lifecycle of data (collection, storage, use, protection, archiving, and deletion).

Data governance is …

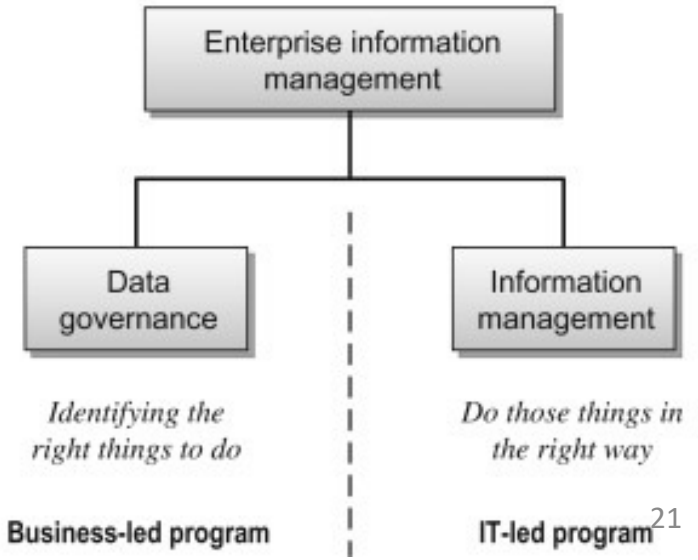
- a set of guidelines for how people behave and make decisions about data

# Important characteristics of DG

| Data governance IS | Data Governance IS NOT |
| --- | --- |
| • More about people and behavior than data<br>• A system that requires and promotes shared agreement<br>• Formal (i.e. written down)<br>• Adds value by supporting institutional mission/goals | • IT's responsibility<br>• Solved by technology<br>• Equally applied across all data assets |

# Data governance vs. data management

- Data management is the technical implementation of data governance.
  - Data governance without implementation is just documentation.
  - Enterprise data management enables the execution and enforcement of policies and processes.
- Data management refers to the management of the full data lifecycle needs of an organization.
  - **Cleansing and standardization**
  - **Masking and encryption**
  - **Archiving and deletion**



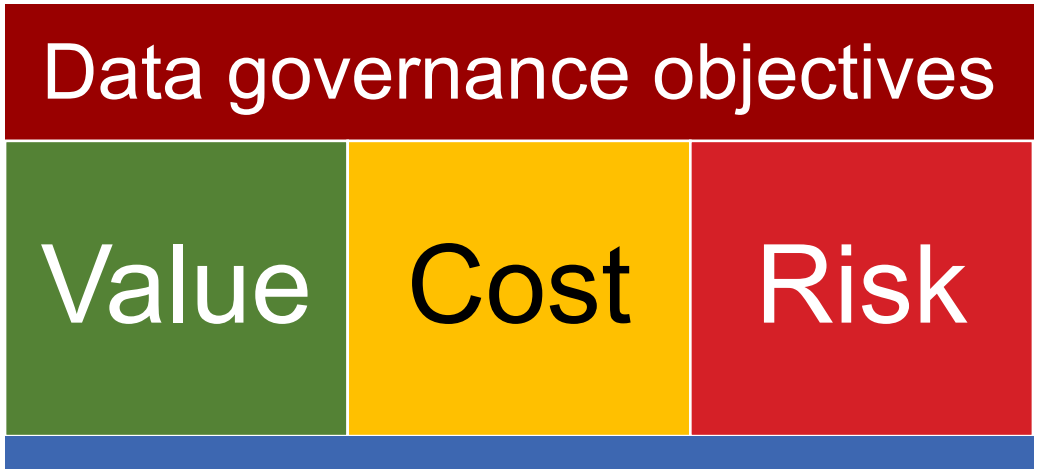
# Data governance vs. data management

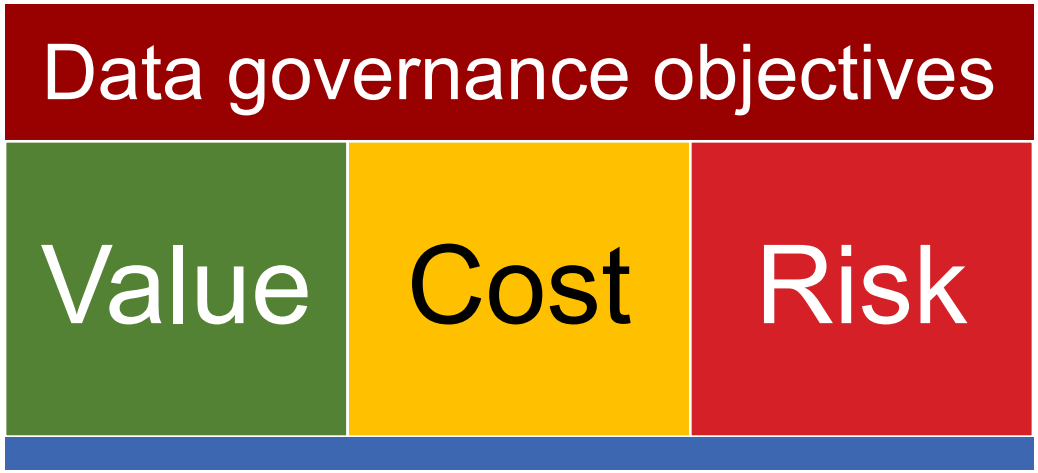
"while data governance and data management are different entities, their goals are the same: create a solid, trustworthy data foundation to empower the smartest people in your enterprise to do their best work."

# Why do we need data governance?

# Data governance objectives

- Everything an organization does should tie to one of three universal value drivers
  - Increase revenue and value
  - Manage cost and complexity
  - Support Risk Management and Compliance efforts and increase confidence






- Value – what could you do that you can't do now?
- Costs – what costs are you incurring because data are not well governed?
- Risks – what risks are you taking because data are not well governed?

# Value: Accelerated decision making

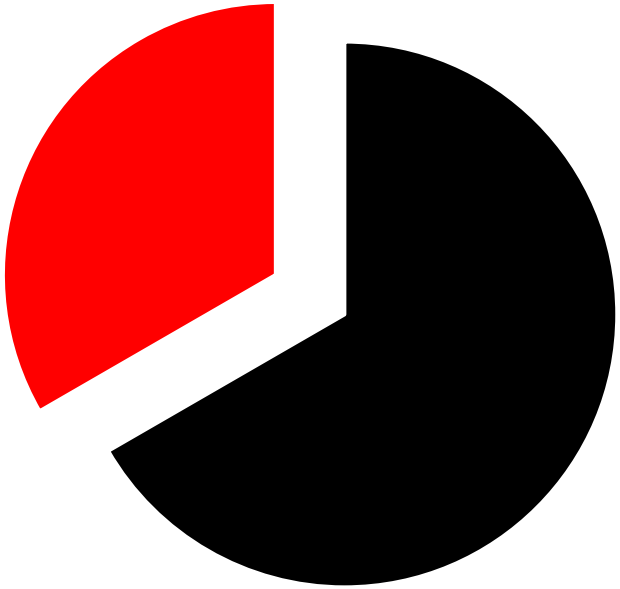
- Improved evidence-based, strategic, and investment decisions by:
  - Quickly acquiring and analyzing large sets of data
  - Decreased reporting errors
  - Easily accessing uniform, reliable data
  - Improved standardization, increasing confidence and transparent communication

# Value: Increased revenue

- Heightened business intelligence and advanced customer analytics drive revenue growth by:
  - Introducing new products
  - Enhancing customer service
  - Optimizing marketing techniques

# Cost control [1]

- A third of Fortune 100 organizations will experience “an information crisis, due to their inability to effectively value, govern and trust their enterprise information.”



Gartner. (2014). “Why data governance matters to your online business,” retrieved August 1, 2016 from http://www.gartner.com/newsroom/id/1898914s-why-data-governance-matters-to-your-online-business/

# Cost control [2]

- Poor data quality costs the US economy $3.1 trillion every year


IBM. (n.d.). “Extracting business value from the 4 V's of big data,” retrieved October 1, 2018 from https://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data

# Cost control [3]

- The average financial impact of poor data quality on businesses is $9.7 million per year. Opportunity costs, loss of reputation and low confidence in data may push these costs higher.


- Forbes (2017). “Poor-quality data imposes costs and risks on businesses,” retrieved October 22, 2018 from https://www.forbes.com/sites/forbespr/2017/05/31/poor-quality-data-imposes-costs-and-risks-on-businesses-says-new-forbes-insights-report

# Manage risk (theft, misuse, data corruption)

- CIO key concerns
  - What are my risk factors, what is my mitigation plan, and what is the potential damage?
- Data governance comes to provide a set of tools, processes, and positions for personnel to manage the risk to data
  - Theft
    - Data is either the product or a key factor in generating value
  - Misuse
    - 2015, AT&T's payout to the FCC after its call center employees disclosed consumers' personal information to third parties for financial gain.
  - Data corruption
    - The risk materializes when deriving operational business conclusions from corrupt (and therefore incorrect) data.

Enterprise Risk Management Approach


Data Governance Supports ERM

- Data Self-Assessment
- Data Risk Reports
- Audit Assessments
- Regulatory Assessments
- Data Issue Process

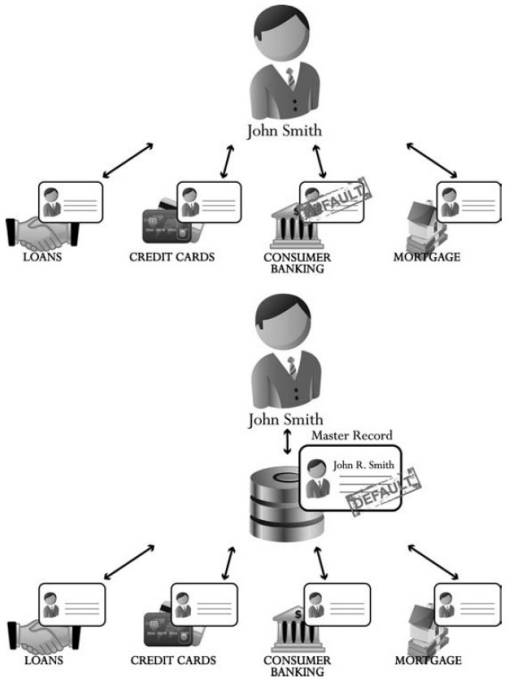
- Data Governance Policy
- Data Standards and Data Quality Framework
- Data Control Inventory

- Clear Data Accountabilities and Expectation
- Data Control Best Practices/Common Design
- Business Rules Develop
- Data Cleansing and Profiling
- Monitoring using DQ dimension (completeness, accuracy, validity checks)
- Reporting and Escalating Data Issue
- Facilitating Remediation of Data Issues

Detective Controls → Corrective Controls → Preventative Controls

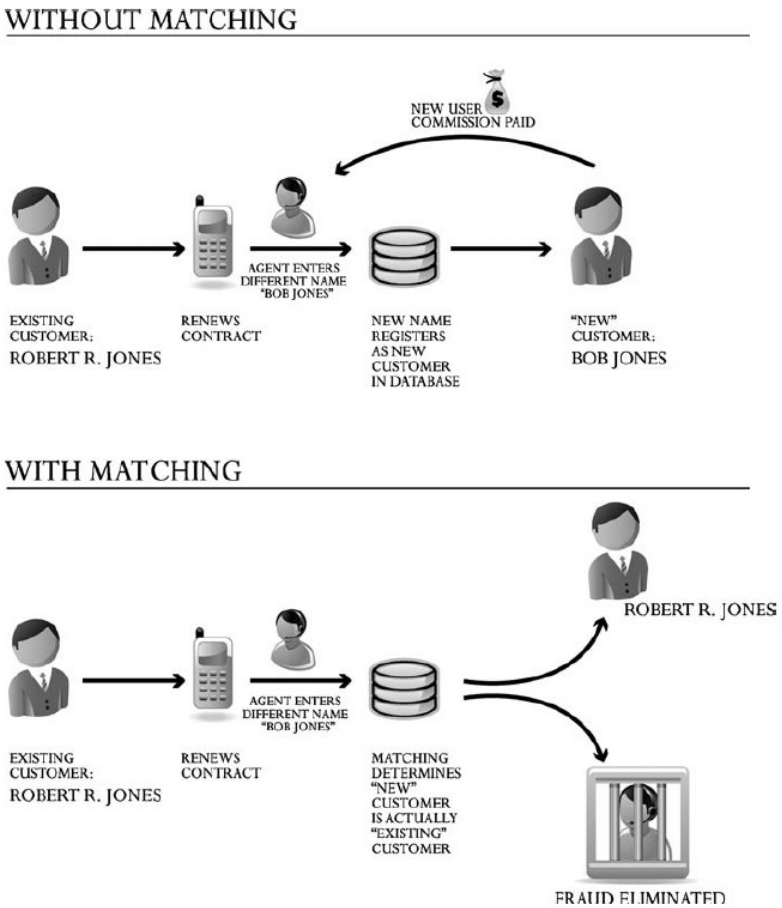
# Risk Mitigation: One version of the truth helps retail bankers manage risk

- Many retail banks have product-oriented risk management systems
- If a customer fails to make a loan, the bank can often take up to several weeks to change the credit limits on credit cards held by the same customer



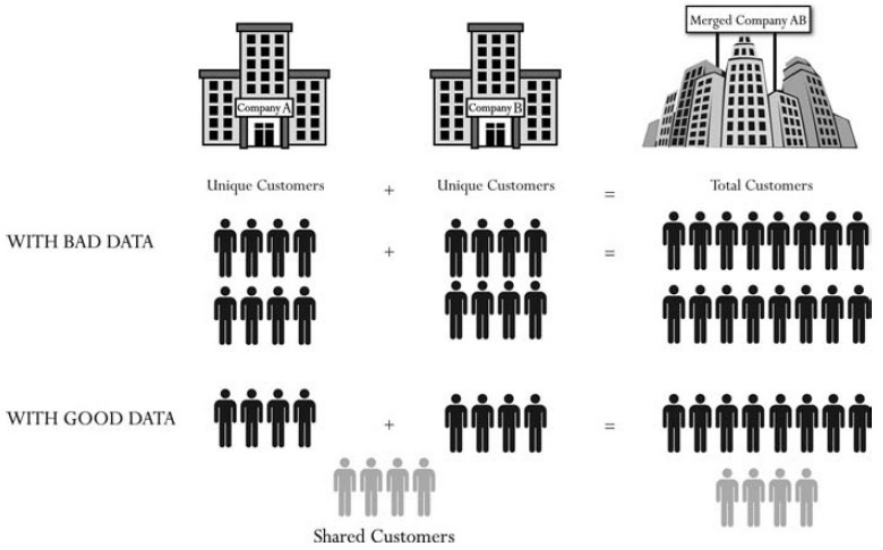
# Fighting fraud with accurate data

- Without matching
  - Mobile sales agents were entering existing customers as new customers by using a slightly different name.
  - Higher commission being paid to the agent.
- With matching
  - The company was able to detect the fraud by reconciling the name with existing customer data already on file.



# Reducing the risk in mergers and acquisitions

- Bad data can lead you to think you have more customers than you really do.
  - There could be shared customers of the companies being merged.
  - A merger won't achieve the financial gains that were expected.



# Risk management

## Most disruptive fraud events by industry

N-iX

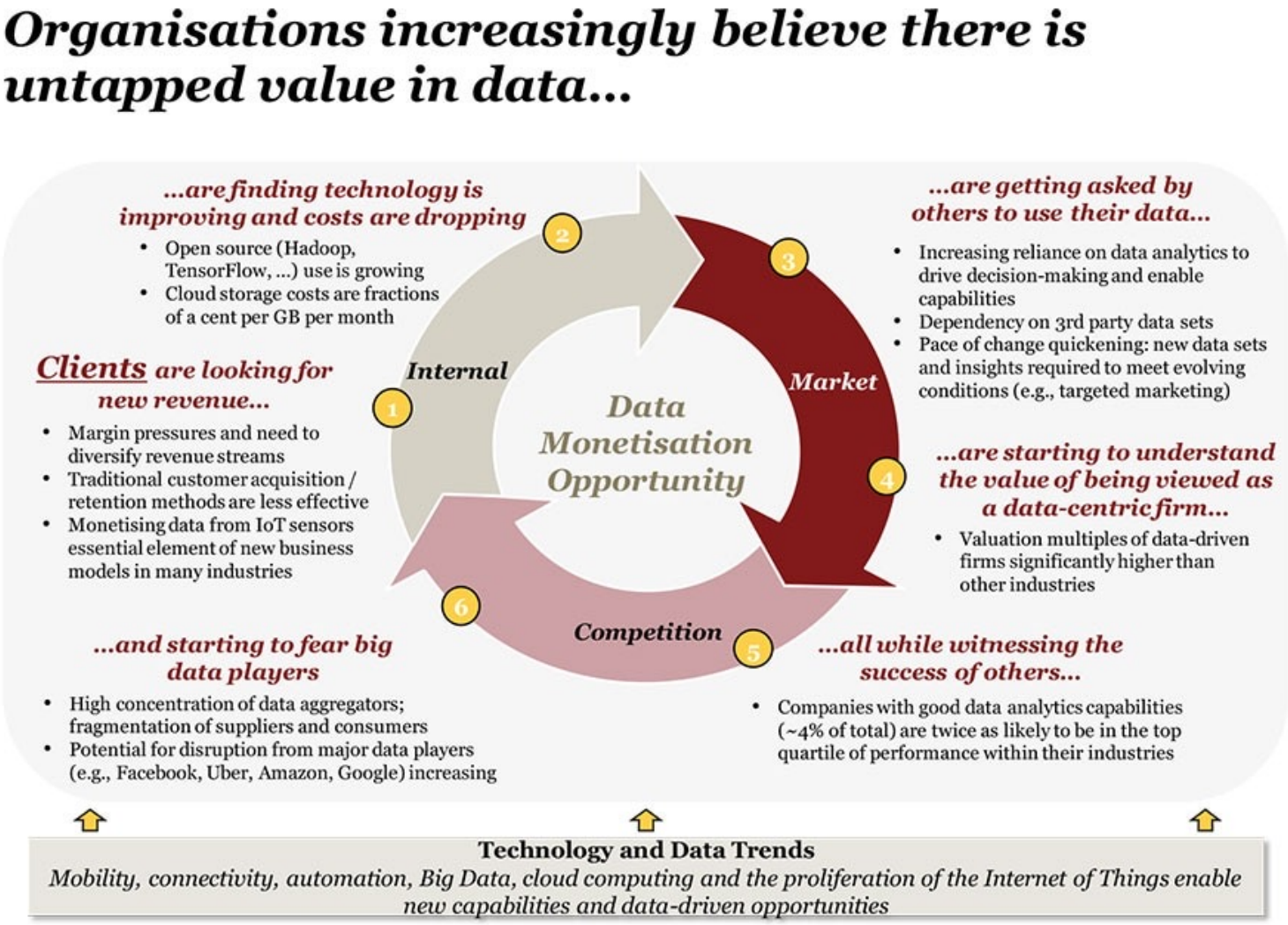
*Source: PwC' 2020 Global Economic Crime and Fraud Survey*

| | Consumer Markets | Energy, Utilities & Resouces | Financial Services | Government & Public Sector | Health Industries | Industrial Products & Manufacturing | Technology, Media & Telecommunications |
|---|---|---|---|---|---|---|---|
| 1 | Customer Fraud **18%** | Bribery and Corruption **17%** | Customer Fraud **27%** | Cybercrime **17%** | Cybercrime **16%** | Asset Misappropriation **21%** | Cybercrime **20%** |
| 2 | Asset Misappropriation **17%** | Asset Misappropriation **16%** | Cybercrime **15%** | Accounting/ Financial Statement Fraud **17%** | Accounting/ Financial Statement Fraud **13%** | Cybercrime **15%** | Accounting/ Financial Statement Fraud **16%** |
| 3 | Cybercrime **16%** | Accounting/ Financial Statement Fraud **13%** | Accounting/ Financial Statement Fraud **14%** | Bribery and Corruption **16%** | Customer Fraud **13%** | Customer Fraud **14%** | Customer Fraud **13%** |

# Regulatory compliance

- Regulations are, in essence, policies that must be adhered to in order to play within the business environment the organization operates in
- Regulation will usually refer to one or more of the following specifics:
  - Fine-grained access control
  - Data retention and data deletion
  - Audit logging
  - Sensitive data classes

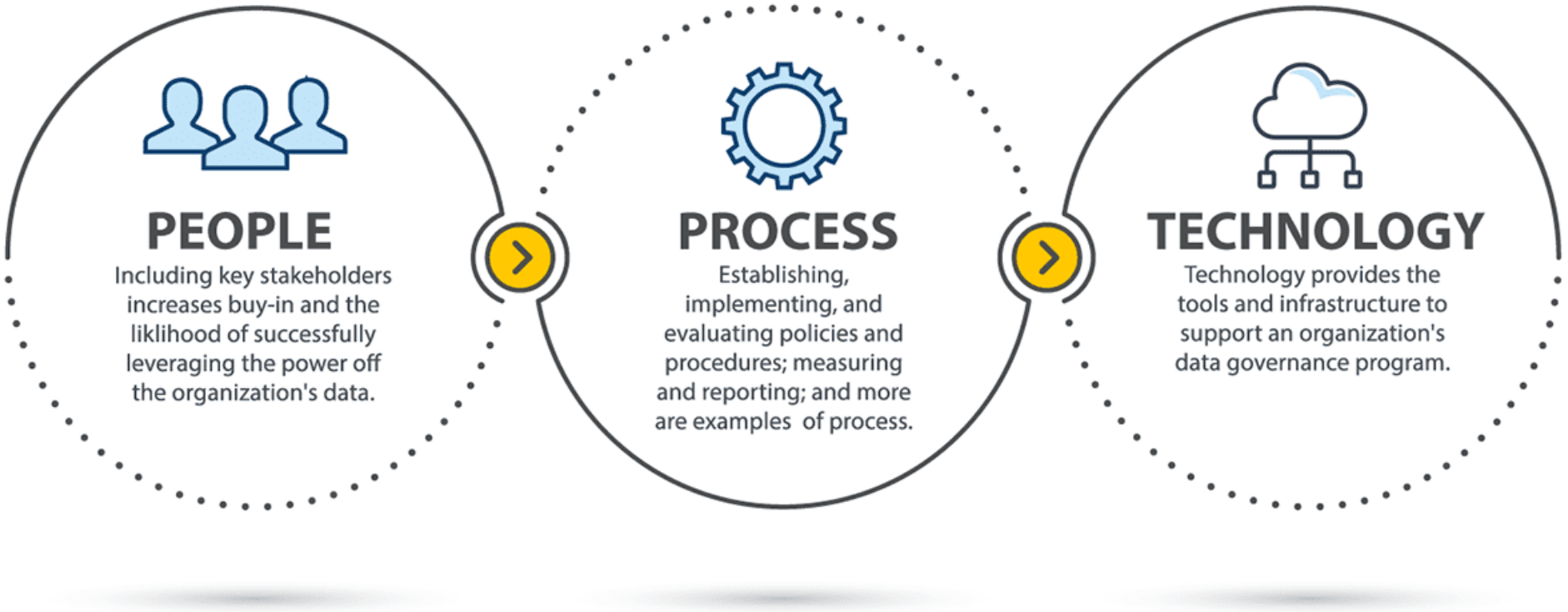
# Data as an asset




*Finding value in data*

# Data governance ingredients

# Data governance ingredients




Data governance embodies three components: the right technology, used by the right people, in the right business process

# The People: Roles and responsibilities

- Who sets success metrics and monitors how well the data governance program is working?
- Who are the data owners?
- Who defines and maintains a business glossary?
- Who creates and maintains policies on access security?
- Who is protecting data privacy for compliance with GDPR and CCPA?
- Who is looking after data quality across all brochures and partner websites?
- Who ensures customer data is consistent across all systems?
- Who is policing external subscription data usage vs the license?
- Who is policing privileged users like DBAs and data scientists?

# Executive Sponsors

- Senior management support is critical to an enterprise-wide activity like data governance.
  - Budget authority and revenue goals
  - If the boss considers it a priority, the staff will, too.
  - Promote collaboration by making it an objective by which employees are measured
- Executives are concerned about how to generate revenue, cut costs, and reduce risk.
  - Keep these key business drivers in mind when building the business case for data governance

# Stakeholders

- Stakeholders are the business owners of data
- They are the people that manage lines of business and functional areas
  - A stakeholder could be the marketing director that is trying to segment customers based on household value
- These stakeholders are often quite attached to their own data silos and need to be convinced that enterprise-wide data is a positive thing.

# Business Experts

- Every department or line of business has a small handful of people that are always consulted for their expertise and knowledge
- Stakeholders are usually very dependent on the business experts for advice as well as execution
- The business experts play a pivotal role by using their expertise to steer the data governance committee
  - Craft the appropriate data definitions and rules.
  - Ensure the data models, the data rules, and the data usage are fit for the needs of their line of business.

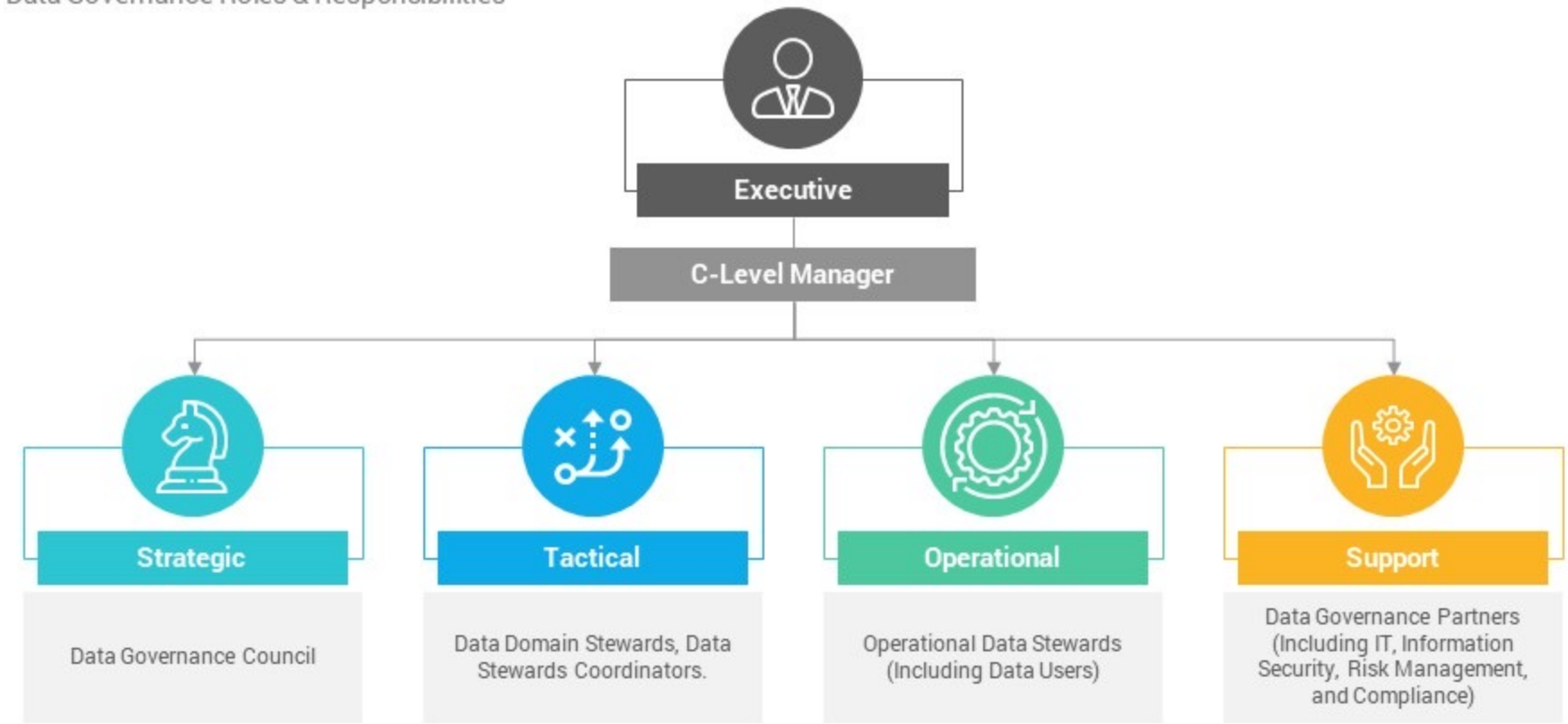
# Data Stewards

- Play a critical role in the collaboration between business and IT, they must be able to speak the language of both groups
  - Representing the best interests of the line of business stakeholder to ensure that data decisions that are made are compatible with the stakeholder needs.
  - Representing the IT experts to ensure that the decisions that are made can be implemented and supported by the technology that supports the functional area.

# IT Experts

- Working closely with the data stewards, the ITexperts will help integrate the decisions made by business into the IT architecture that runs the business
  - Integration of the business requirements into IT systems
  - Building and maintaining an IT architecture that supports the business
  - Ensuring that IT infrastructure meets the service requirements of the business in terms of access, response time, and availability
  - Implementation of policies for privacy and security of the applications and databases

# Data Governance Roles & Responsibilities



# Data Governance Body

# Executive level: DG Steering Committee

- Support, sponsorship, and approves of DG program.
- Communicates expectations and requirement of DG program.
- Identifying people in their part of the organization for **Data Governance Council**.
- No specific day-to-day or monthly data governance activities …

# Strategic level: Data Governance Council

- Make decisions at a strategic level
- Set data policy, data role framework, methods, priorities, tools, etc.
- Identify and approve of pivotal data governance roles including cross-enterprise domain stewards and coordinators.
- Resolves escalated issues at strategic level

# Tactical level: Data Domain Stewards

- Responsible for ‘enterprise’ management of a domain of data.
- Involved/facilitator in cross business unit resolution of data definition, production and usage issues.
- Responsible for escalating well-documented issues to the strategic level.
- Responsible for documenting data classification rules, compliance rules, business rules for data in their domain (may delegate this).
- Responsible for making certain the rules are communicated to all stakeholders of data in that domain (may delegate).
- Responsible for participating in tactical groups (with other domain stewards, steward coordinators, and operational stewards) for finite periods of time to address specific issues and projects related to their domain and business unit.

# Tactical level: Data Steward Coordinators

- Act as the point communications person
  - for distributing rules and regulations per domain of data to the operational stewards in their business unit (and making certain that the operational data stewards understand the rules & risks).
  - for their business unit to document and communicate issues pertaining to specific domains of data to the proper **Data Domain Steward**.
- Identify the operational stewards of data per domain for their business unit.
- The Data Steward Coordinator typically has no decision-making authority but plays a pivotal role in data governance and data stewardship success.

# Operational level: Operational Data Stewards

- Data Definers
  - Defining the data that will be used by the organization, how that data will be used, and how that data will be managed.
  - Participate in creating/reviewing/approving data definitions.
  - Participate in the integrity and quality of data definition.
- Data Producers
  - Producing, creating, updating, deleting, retiring, archiving the data that will be managed.
- Data Users
  - Using data to perform their job and processes.
  - Following the rules associated with identifying and classifying data access levels.
- Identifying and documenting regulatory and legal/risk issues.
- Supporting / sharing knowledge with other stewards.
- Communicating new / changed business requirements to individuals who may be impacted and can influence change.
- Communicating concerns, issues, and problems with data to the individuals that can influence change.

# Support level: Data Governance Office

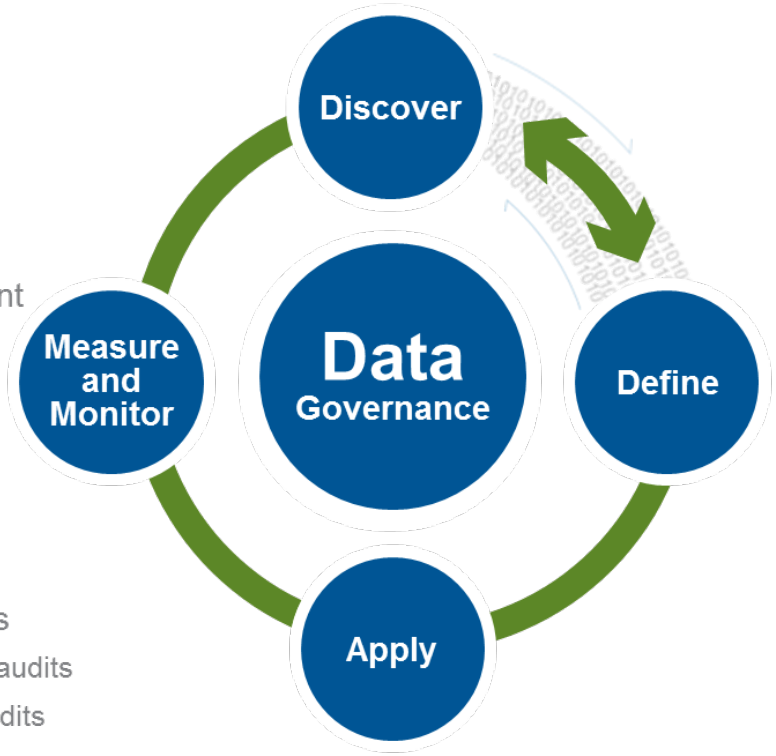
- Participate in Data governance program development.
- Architect solution & framework.
- Assist in administering the program.
  - Facilitate the Data Governance Council meetings.
  - Report results to Data Governance Council.
- Participate in the development & delivery of data governance policies, standards, guidelines, and procedures.
- Assist in defining data quality metrics for periodic release.
- Support data quality issue analysis & remediation for “strategic” data.
- Conduct audits to ensure components are in place for improving the program.

# The Process [1]

- Diverse companies, diverse needs and approaches to Data Governance
  - Virtually every company will be going out and empowering their workers with a certain set of tools, and the big difference in how much value is received from that will be how much the company steps back and really thinks through their business processes . . . thinking through how their business can change, how their project management, their customer feedback, their planning cycles can be quite different than they ever were before. —BILL GATES, MICROSOFT CHAIRMAN AND FOUNDER

# The process [2]

**Discover**
- Data discovery
- Data profiling
- Data inventories
- Process inventories
- CRUD analysis
- Capabilities assessment




**Define**
- Business glossary creation
- Data classifications
- Data relationships
- Reference data
- Business rules
- Data governance policies
- Other dependent policies
- Key Performance Indicators

**Measure and Monitor**
- Proactive monitoring
- Operational dashboards
  - Reactive operational DQ audits
  - Dashboard monitoring/audits
- Data lineage analysis
- Program performance
- Business value/ROI

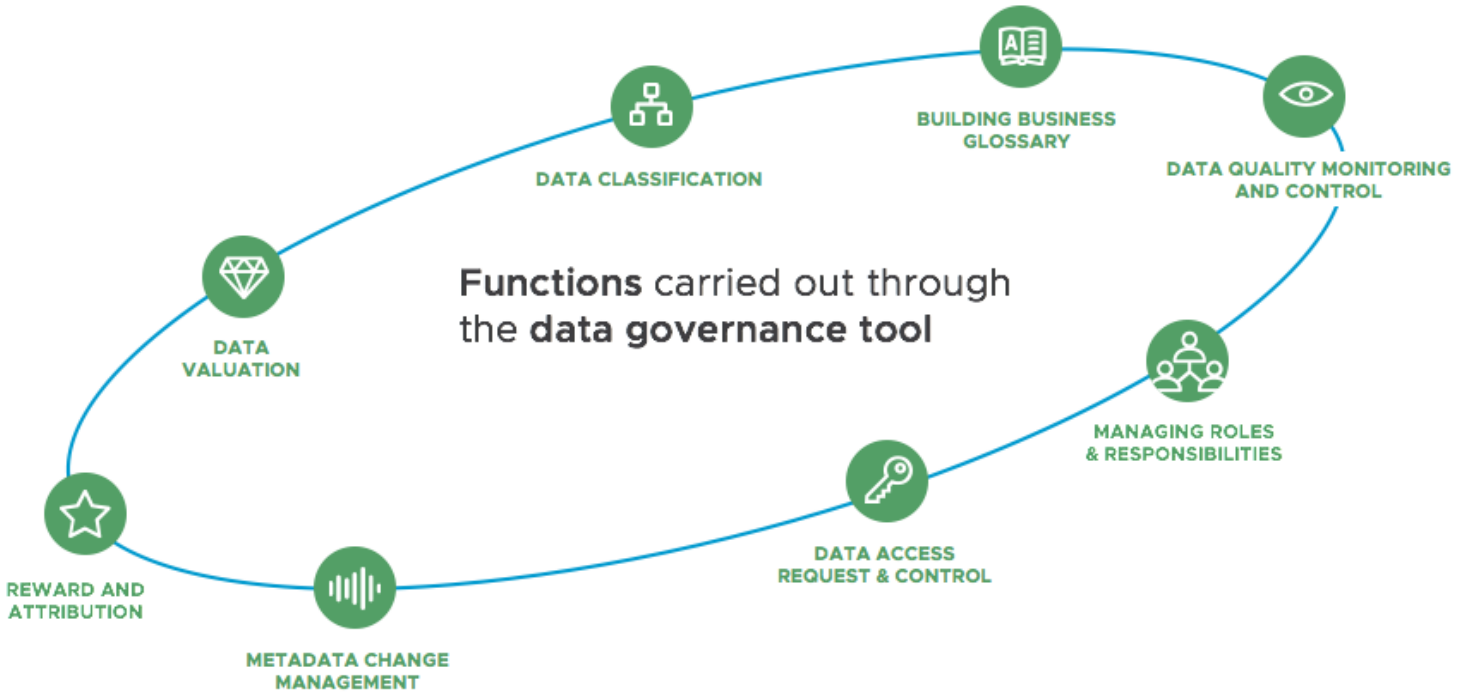
**Apply**
- Automated rules
- Manual rules
- End to end workflows
- Business/IT collaboration

# The process [2]

- Before, IT department is responsible for the collection and management of data.
  - But what is considered good, clean, usable data from a technical standpoint may not be complete, accurate, and timely information for the business user.
- A documented, repeatable process that is adhered to throughout the enterprise will ensure consistent data across your organization.

# The technology/tools [1]

- Aids in the process of creating and maintaining a structured set of policies, procedures, and protocols that control how an organization's data is stored, used, and managed.



# The technology/tools [2]

- Some of the key features to look for in a data governance tool include:
  - **Discovering, capturing, and cataloging data**
    - The catalog serves as a bird's eye view of each data entity, its profile, relationships, lineage, and the business glossary (with the decided common terminology).
  - **Data and metadata management**
    - Encapsulates the data integration application and controls the data lifecycle and tracking each data pipeline
  - **Data ownership and stewardship capabilities**
    - Enables both owners and stewards to do their jobs.
  - **Self-service tools**
    - Essential for organizations whose data governance goals are aligned more toward the business team.
    - These tools must provide an intuitive and clutter-free representation of all data, with reporting and alerting capabilities rolled into it.
    - A self-service station allows for consistent and clear decision-making.

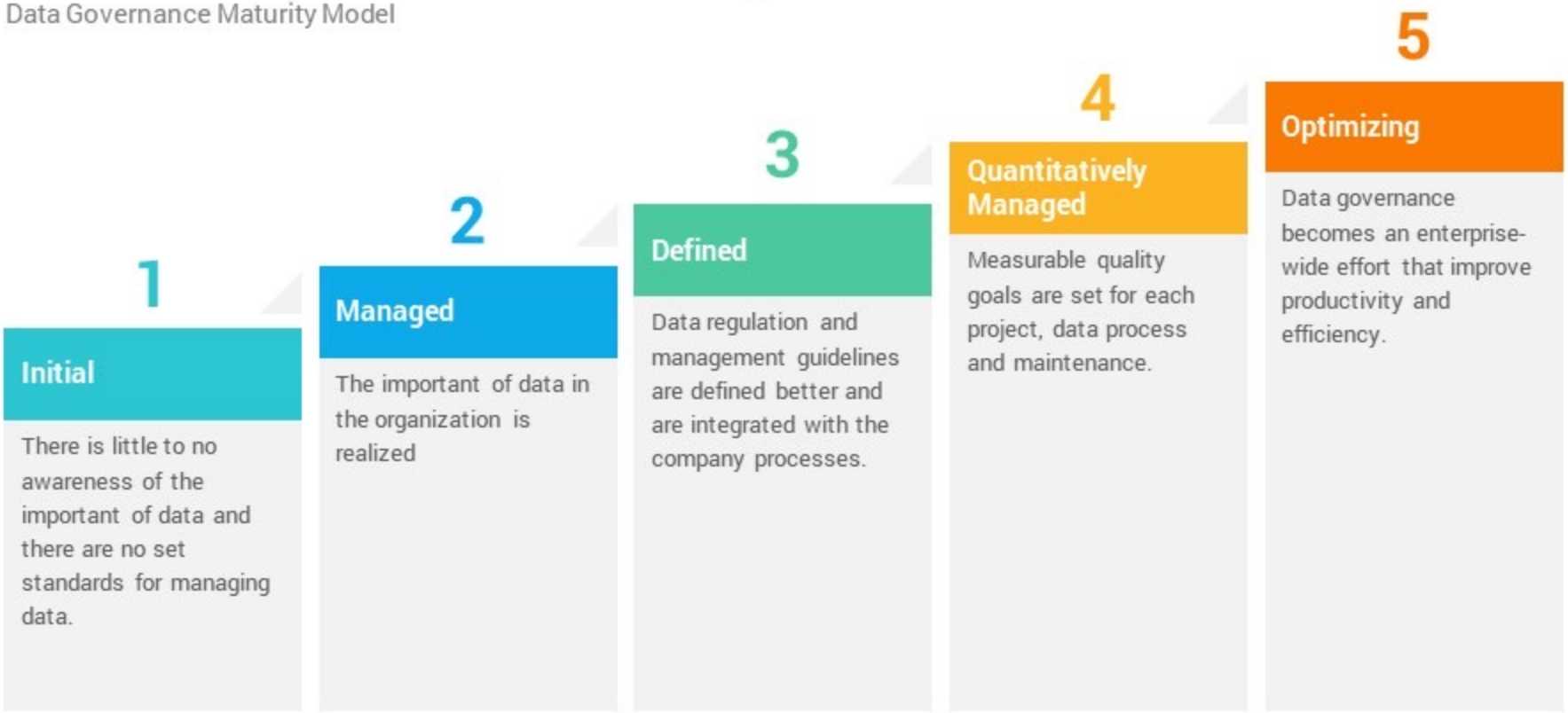
# The technology/tools [3]

- Some of the key features to look for in a data governance tool include:
  - **Data lineage automation**
    - Data lineage tracks the origin of each data entity, the changes that it went through, and its movement within the system. It helps with tracing and spotting any errors flagged by the system.
  - **Business glossary**
    - The starting point of every data governance plan is the creation of common data definitions and formats. Creating a common glossary of business terms helps maintain consistency.
  - **Compatibility with existing systems**
    - This means that the tool picked by your organization must be flexible and customizable.
  - **Compliance audit-ready**
    - must provide for external and internal audits, especially if compliance is one of the key goals of governance
  - **Policy management**
    - include configuration and management of policy controls. Once the controls have been set up, they are expected to automatically enforce policy management.

# Maturity Models

# Data Governance Maturity Model

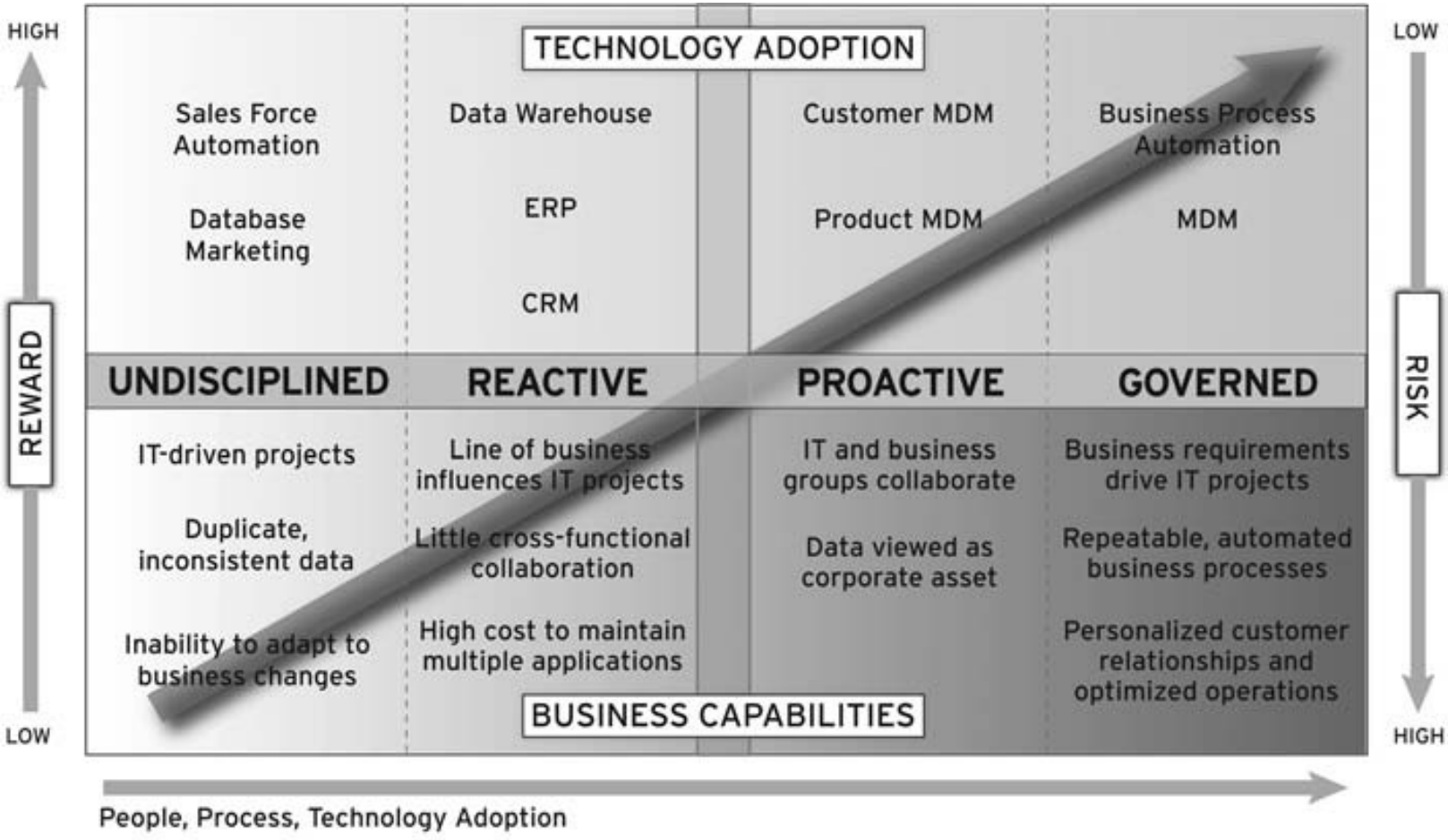
Data Governance Maturity Model





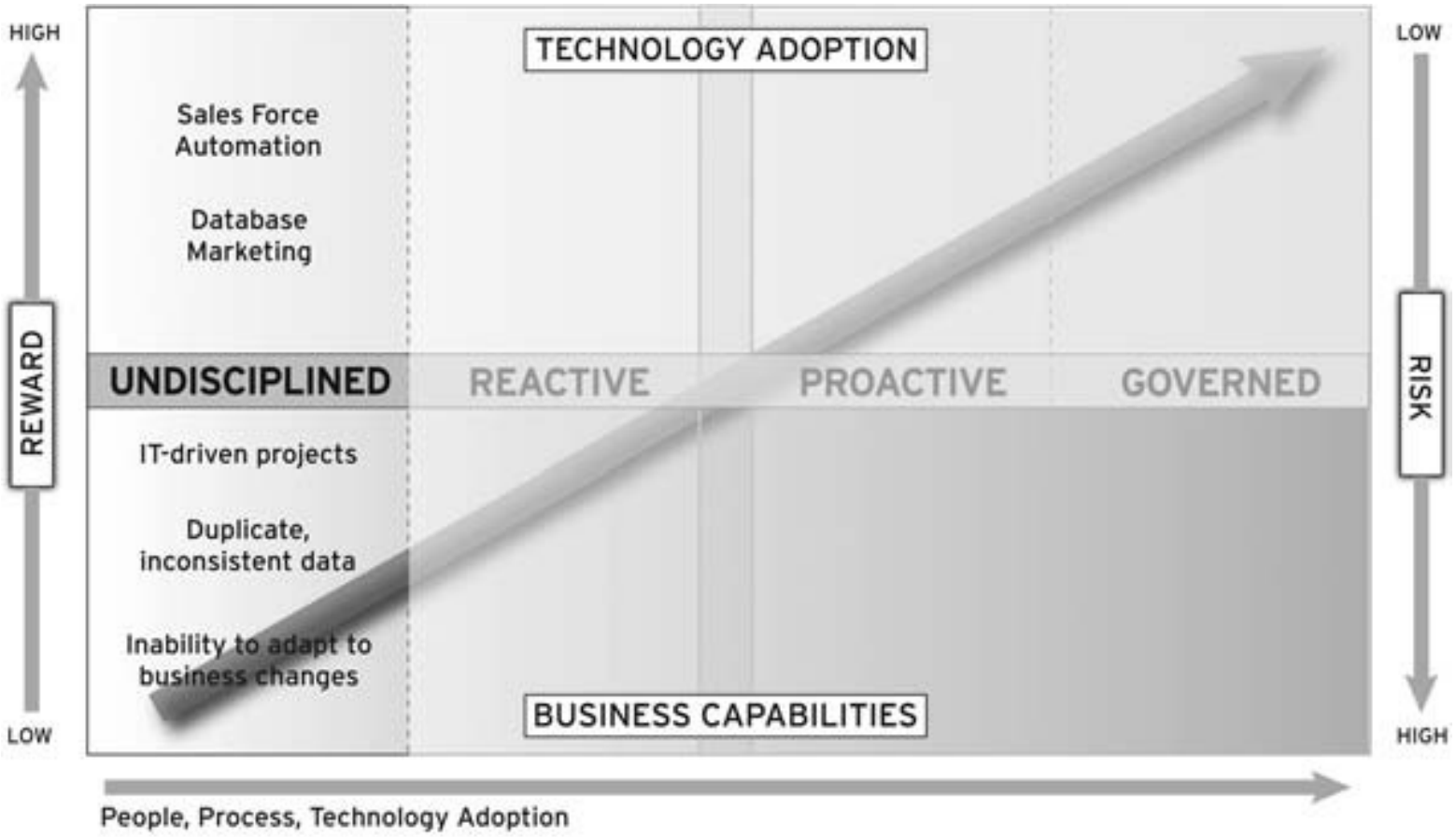

slidesalad

# The data governance maturity model



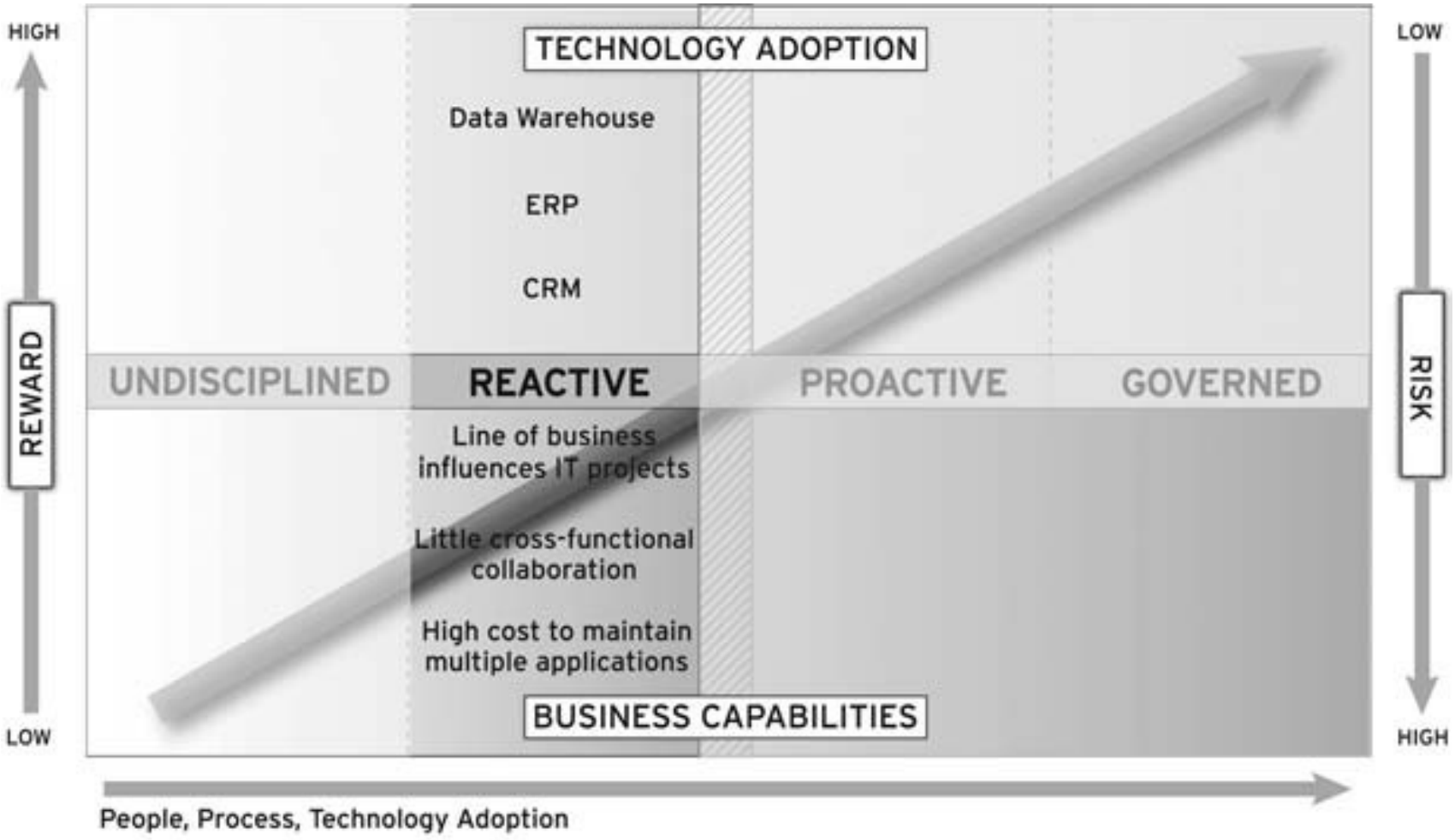
# Undisciplined organizations: Disasters waiting to happen

- Characteristics of an Undisciplined Organization
  - Think locally, act locally
  - Few defined data rules and policies
  - Redundant data found in different sources
  - Little or no executive oversight
- Technology Adoption
  - Tactical applications to solve very specific problems: for example, sales force automation or database marketing
  - Very localized data management technology implemented within the tactical applications, if at all
- Business Capabilities
  - IT-driven projects
  - Duplicate, inconsistent data
  - Inability to adapt to business changes



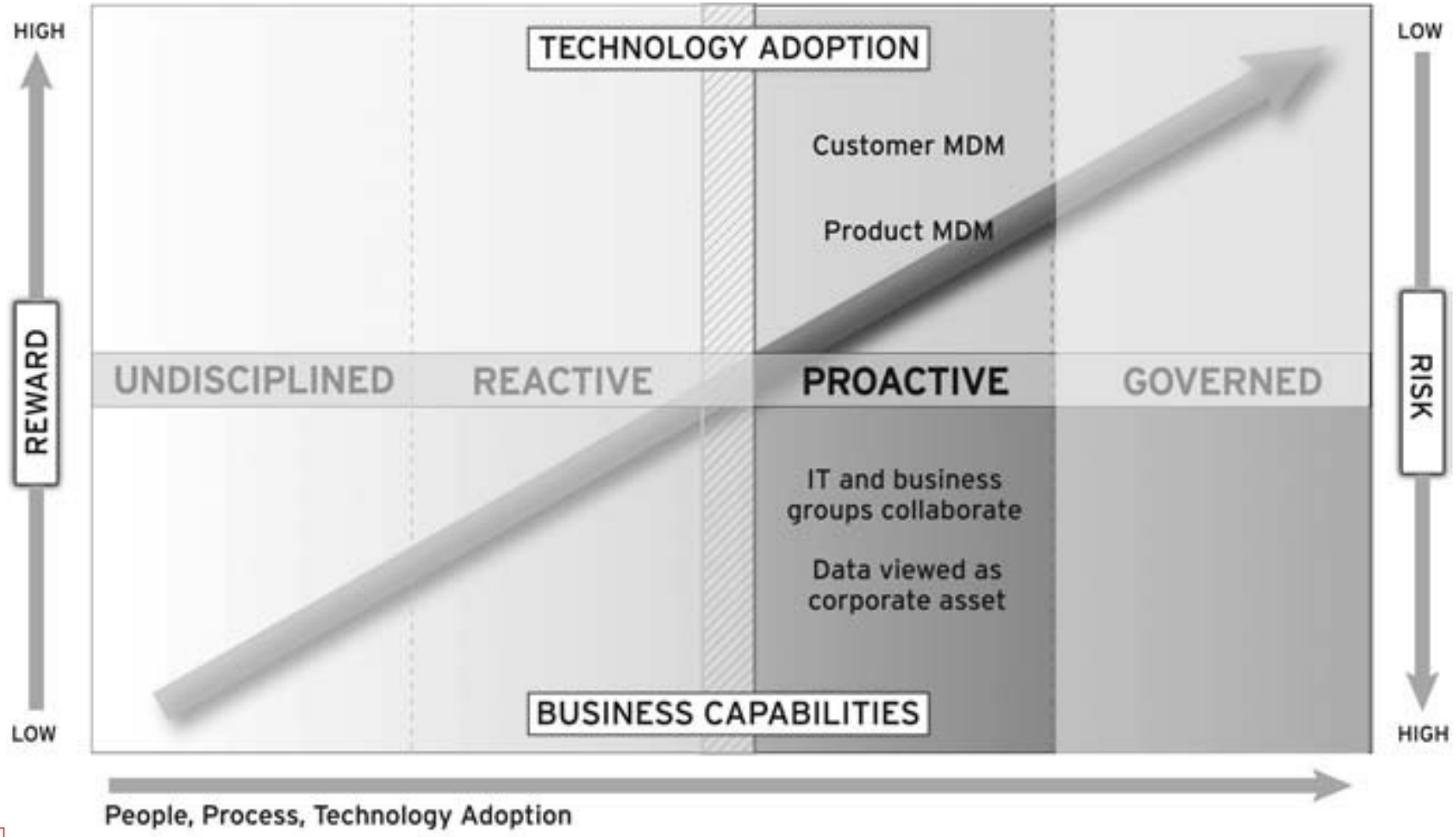
# Reactive Organizations: Trying to get beyond crisis mode

- Characteristics of a Reactive Organization
  - Think globally, act locally
  - Presence of data management technology, but with limited data quality deployment
  - Siloed data leading to many views of what should be the same data
  - Awareness of data problems only after a crisis occurs
- Technology Adoption
  - Data warehouse
  - Enterprise resource planning (ERP)
  - Customer relationship management (CRM)
  - Data integration tools
- Business Capabilities
  - Line of business influences IT projects
  - Little cross-functional collaboration
  - High cost to maintain multiple applications



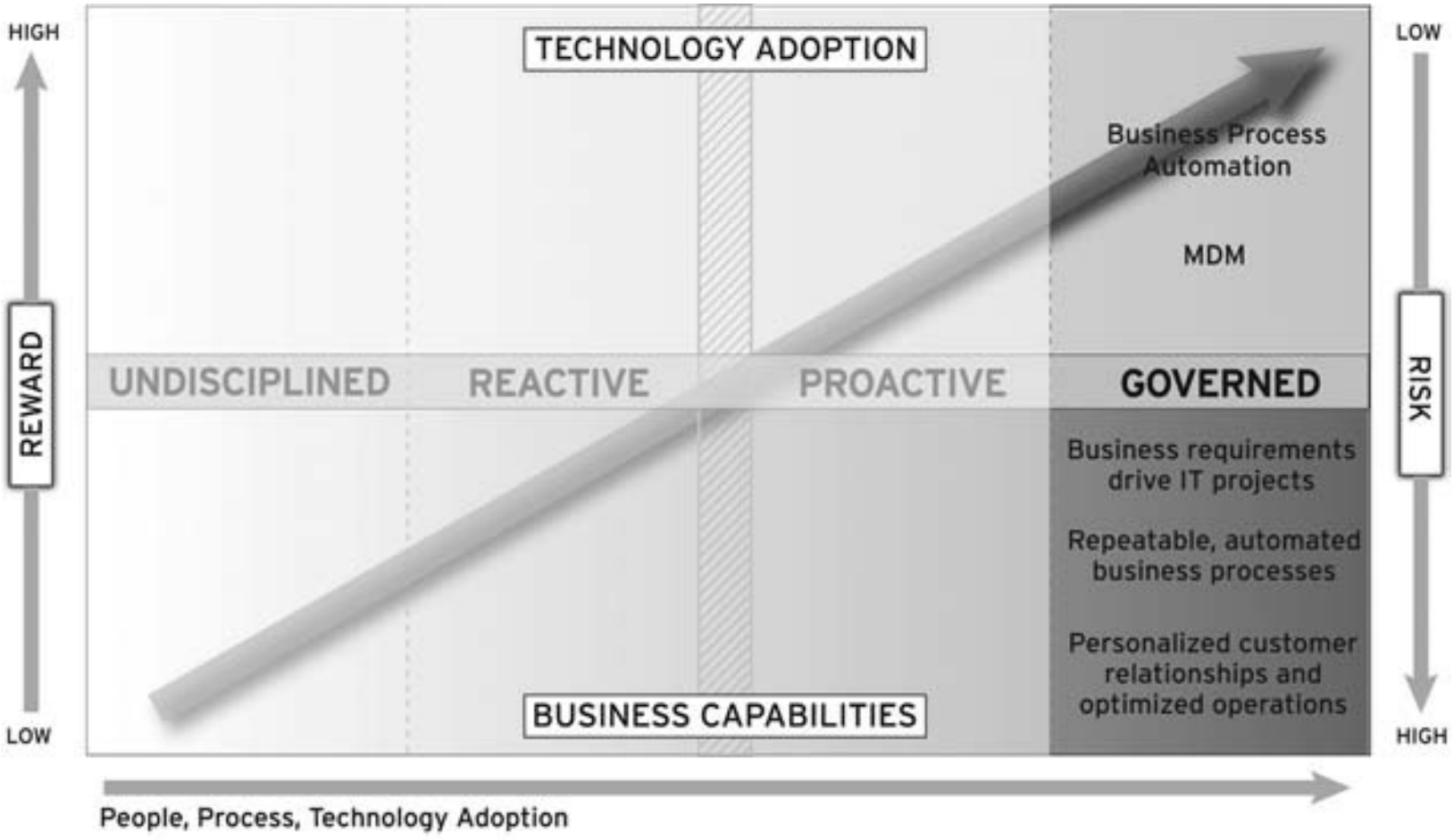
# Proactive organizations: Reducing risk, avoiding uncertainty

- Characteristics of a Proactive Organization
  - Think globally, act collectively
  - Mastered use of enterprise resource planning (ERP), customer relationship management (CRM), and data warehouse technology
  - Executives who view data as a strategic asset
- Technology Adoption
  - Customer master data management (MDM)
  - Product MDM
  - Employing enterprise-wide data definitions and business rules
  - Enabling service-oriented architecture (SOA) architecture for cross organization data consistency
- Business Capabilities
  - IT and business groups collaborate
  - Enterprise view of certain domains
  - Data viewed as a corporate asset



# Governed Organizations: Trust in data pays multiple benefits

- Characteristics of a Governed Organization
  - Think globally, act globally
  - Unified data governance strategy
  - Comfortable incorporating external data without fear of corrupting existing, internal data
  - Executive sponsorship
- Technology Adoption
  - Business process automation
  - Master data management (MDM)
- Business Capabilities
  - Business requirements drive IT projects
  - Repeatable, automated business processes
  - Personalized customer relationships and optimized operations



# Implementing a Governance framework

# Universal Data Governance Principles

1. Integrity

   Data Governance participants will practice integrity with their dealings with each other; they will be truthful and forthcoming when discussing drivers, constraints, options, and impacts for data-related decisions.

2. Transparency

   Data Governance and Stewardship processes will exhibit transparency; it should be clear to all participants and auditors how and when data-related decisions and controls were introduced into the processes.

3. Auditability

   Data-related decisions, processes, and controls subject to Data Governance will be auditable; they will be accompanied by documentation to support compliance-based and operational auditing requirements.

4. Accountability

   Data Governance will define accountabilities for cross-functional data-related decisions, processes, and controls.

5. Stewardship

   Data Governance will define accountabilities for stewardship activities that are the responsibilities of individual contributors, as well as accountabilities for groups of Data Stewards.

6. Checks-and-Balances

   Data Governance will define accountabilities in a manner that introduces checks-and-balances between business and technology teams as well as between those who create/collect information, those who manage it, those who use it, and those who introduce standards and compliance requirements.

7. Standardization

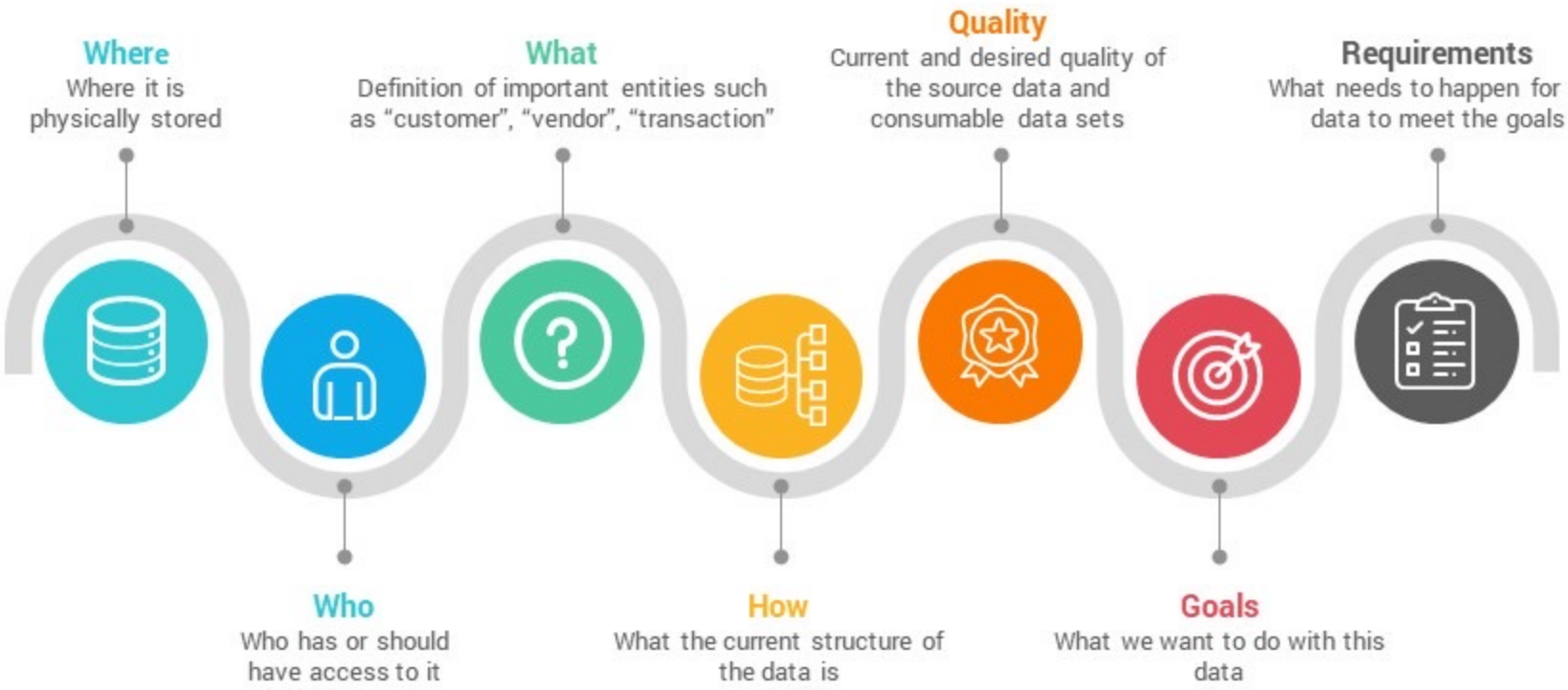
   Data Governance will introduce and support standardization of enterprise data.

8. Change Management

   Data Governance will support proactive and reactive Change Management activities for reference data values and the structure/use of master data and metadata.

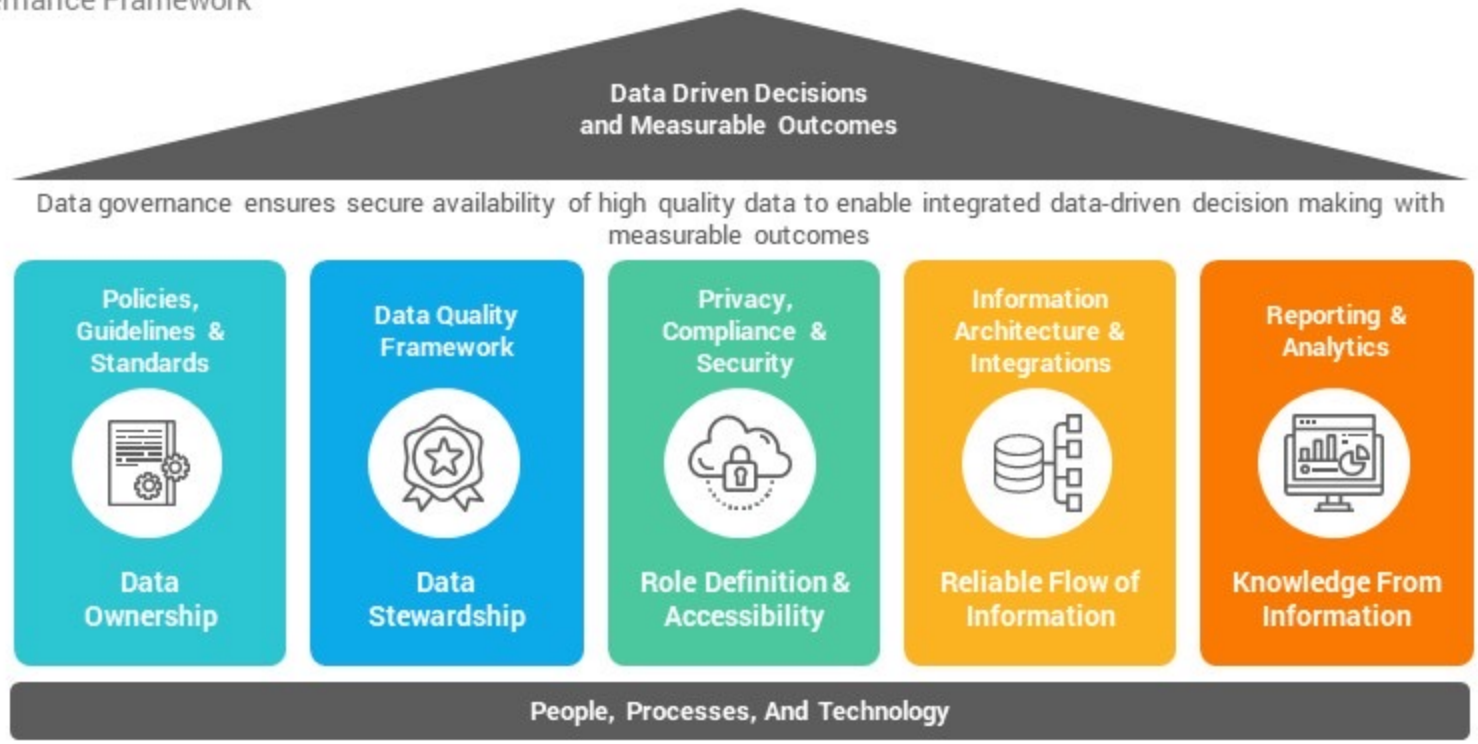
# Data Governance Strategy

Data Governance Strategy



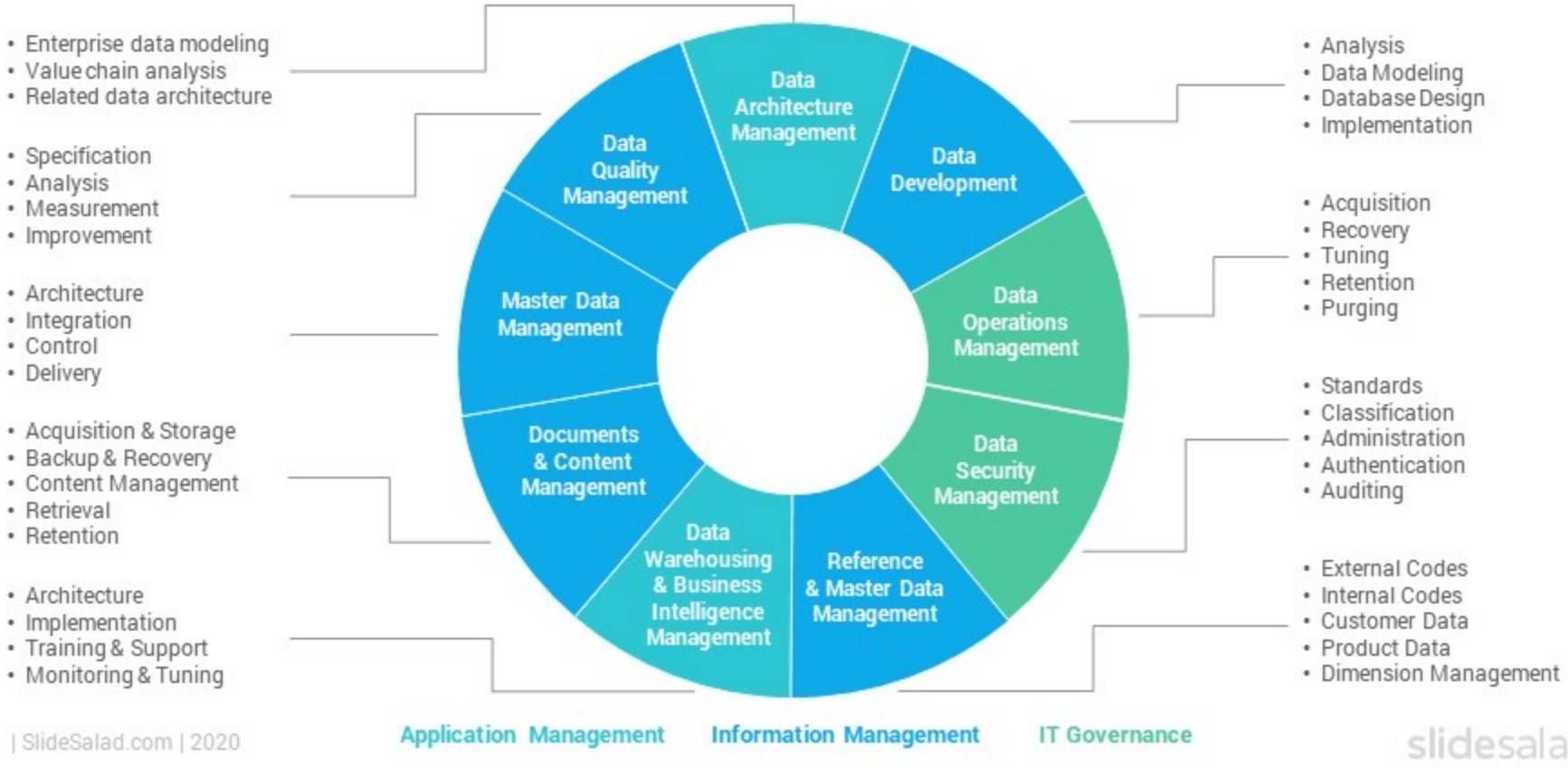
# Data Governance Framework

Data Governance Framework



# Scope Of Data Governance – Data Management

Scope Of Data Governance - Data-management

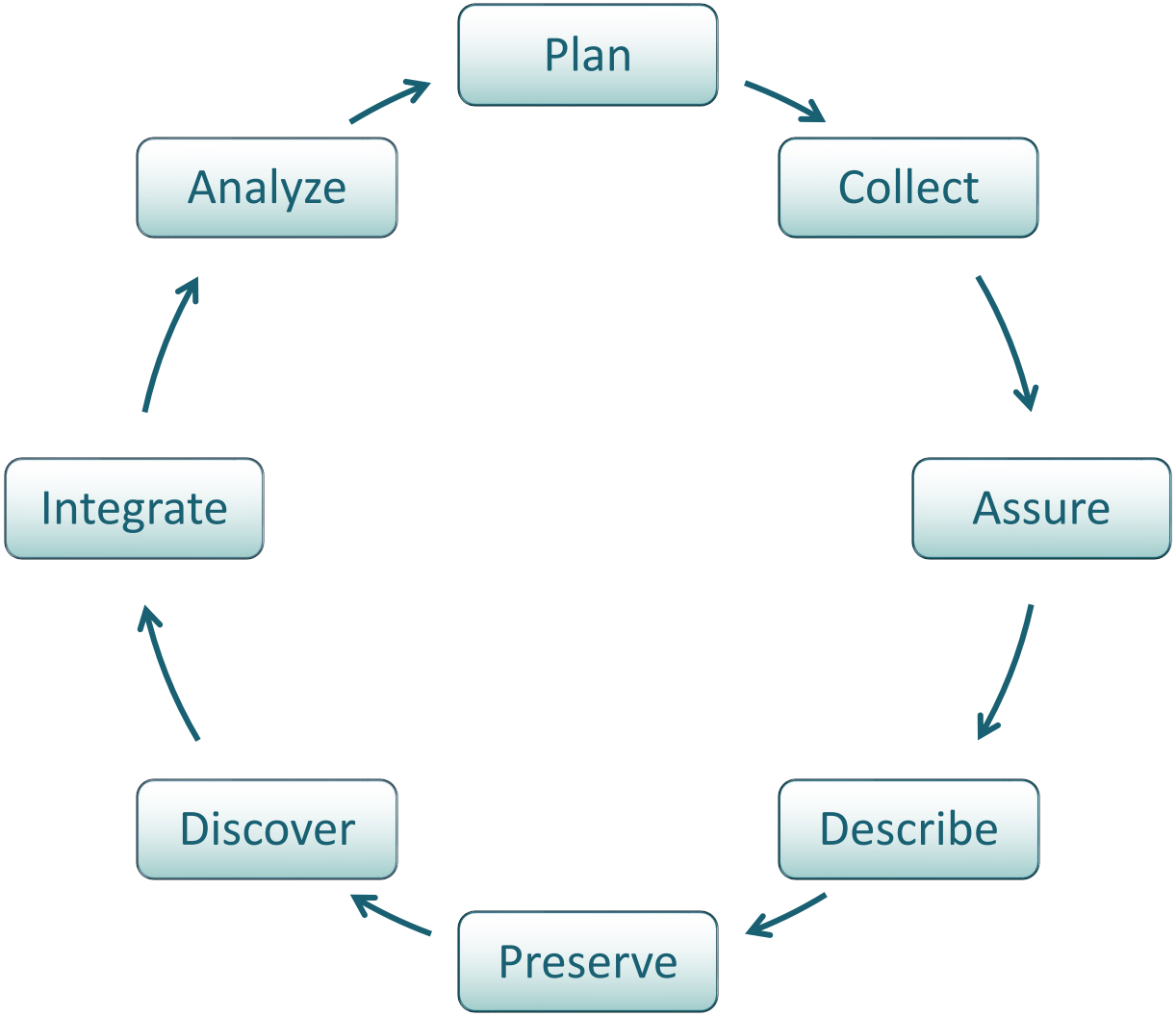



3 | SlideSalad.com | 2020

slidesalad

# Data life cycle management

# Data Life Cycle



# Planning

- Consider data management before you collect data
  - What kind of data will be collected?
  - Which methods will be used (sensors, samples, etc.)?
  - What data formats/standards are appropriate?
  - How will the data be used?
  - How will you share the data?
  - Will your methods satisfy
    - Funding requirements
    - Policies for access, sharing, reuse
  - Budget – most of the time tihis is overlooked!
- Output
  - Formal document

# Collect

- What are some ways that we produce data?
- Experiments, observations, samples,
- Varying frequency, temporal and spatial coverage
- Data collection includes data entry
  - Transcribing notebooks into digital forms
  - Automated processing of data into a database

# Assure

- Strategies for preventing errors from entering datasets
  - Standard data entry forms
  - Pre-specification of formats, units, etc.
- Activities to ensure quality during collection
  - Standard field and laboratory procedures
  - Automated rannge checks for sensor data
- Activities to clean collected data
  - Common to sensor data streams
  - Dependent upon variable and sensor
  - Graphical and statistical summaries

# Describe

- Metadata
  - What metadata are needed?
  - What format for the metadata?
- Documentation and reporting of data
- Contextual details
  - What is it critical to know about the data?
- Description of temporal and spatial details, instruments/sensors, methods, units, files, etc.

# Preserve

- How are you preserving your data?
  - What will be preserved
  - Where will it be preserved
  - Backup, version control?
- Policies for access, sharing, and reuse

- Does Your Office Look Like This?
- What are the potential problems?
- What are some potential solutions?

# Discover

- Most data are not easily discoverable
  - Encapsulated in databases or files
  - Formats not compatible with web indexing technologies
- Conditions for effective data discovery
  - Highly curated data, well described via structured metadata
  - Standards for data and metadata formats

# Integrate & analyze

- Integration
  - Combining data from different sources
  - Creating a unifying view of the data
  - Overcoming heterogeneity
- Analysis
  - To find out insightful values from data

# Takeaways

- Data governance is more about people than data
- Process and written documents are essential
  - Leadership support
  - Broad-based consultation, including faculty
  - Opportunity for consultation
  - Representation
- Software can help, but it won't fix broken processes or organizations
- Starting data governance is hard work; sustaining it is harder

ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

# Thank you for your attention!!!