Data Science

Ubai Sandouk, PhD

2021

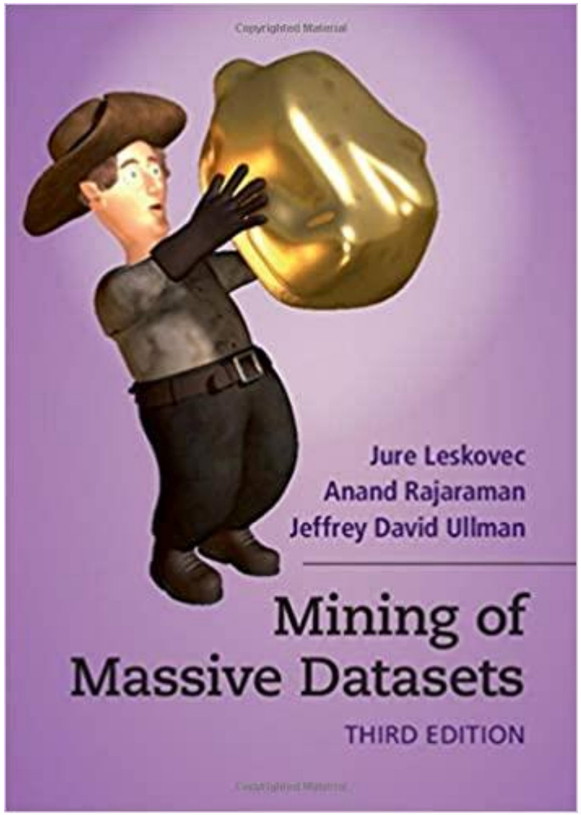
Course structure

- Course Lecturer:
  - Ubai Sandouk
  - ubaisandouk@gmail.com
- Lab Instructor:
  - Eng. Abdulbadee Murad

- Classes:
  - First Semester, 2021.
  - *Theory*: Wednesdays, morning.
    - Focus on Data Science, Data Warehousing, Data Mining
  - *Lab*: Mondays, afternoon.
    - Focus on the correct and optimized working of a database and DBA.

Course structure

- Text book:
  - ***(MMD)*** *Mining of Massive Datasets, 3rd Ed.*, by J. Leskovec, A. Rajaraman, and J. D. Ullman (Cambridge University Press, 2019).
  - http://www.mmds.org



Course structure

- Resources:
  - *(**ADS**) Algorithms for Data Science*, by B. C. H. Steele, John Chandler, and Swarna Reddy. (Springer, 2016) Link
  - *(**IDS1**) Introduction to Data Science (in R)*, by R. A. Irizarry. (CRC Press, 2019) Link
  - *(**IDS2**) Introduction to Data Science (A Python Approach),* by L. Igual, and S. Seguí. (Springer, 2017) *Link*
  - *(**4P**) The Fourth Paradigm*, by T. Hey (Microsoft Research, 2009) Link
  - *(**KDD**) The KDD process for extracting useful knowledge from volumes of data*, by U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. (Communications of the ACM, 1996) Link
  - (***DWT***) *The data warehouse toolkit: the complete guide to dimensional modelling (3rd Ed.),* by R. Kimball, and M. Ross. (John Wiley & Sons, 2013). Link

Course structure

- Evaluation:
  - 70% Exam.
    - Open book.
  - 30% Lab work
    - Homework and Assignments.

PART 0:

The Data Lifecycle

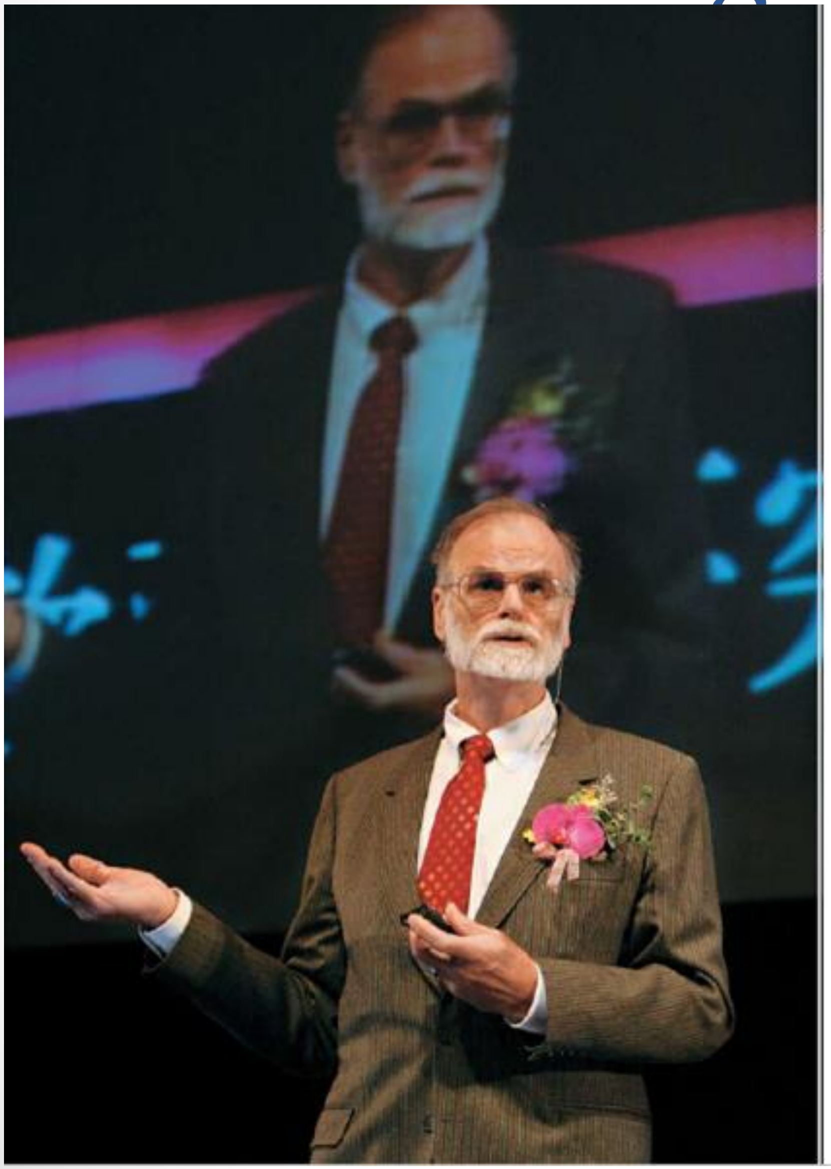
Source: dreamstime

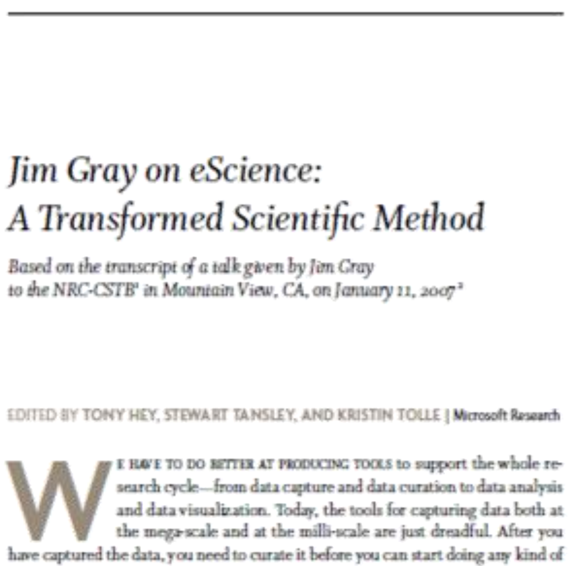
Agenda

- Fourth Paradigm
- Data Flood
- Data Life-Cycle
- Data Quality
- Data Professions
- References

MMD 1, **4P** 1

Fourth Paradigm





Jim Gray on eScience:
A Transformed Scientific Method

Based on the transcript of a talk given by Jim Gray
to the NRC-CSTB[1] in Mountain View, CA, on January 11, 2007[2]

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE | Microsoft Research

WE HAVE TO DO BETTER AT PRODUCING TOOLS to support the whole research cycle—from data capture and data curation to data analysis and data visualization. Today, the tools for capturing data both at the mega-scale and at the milli-scale are just dreadful. After you have captured the data, you need to curate it before you can start doing any kind of

The fourth paradigm is data science!

After,

- empirical evidence,
- scientific theory, and
- computational science.

Current Climate

- Information systems deal with *information and documents*; *data is* the driving mechanism of all information systems.

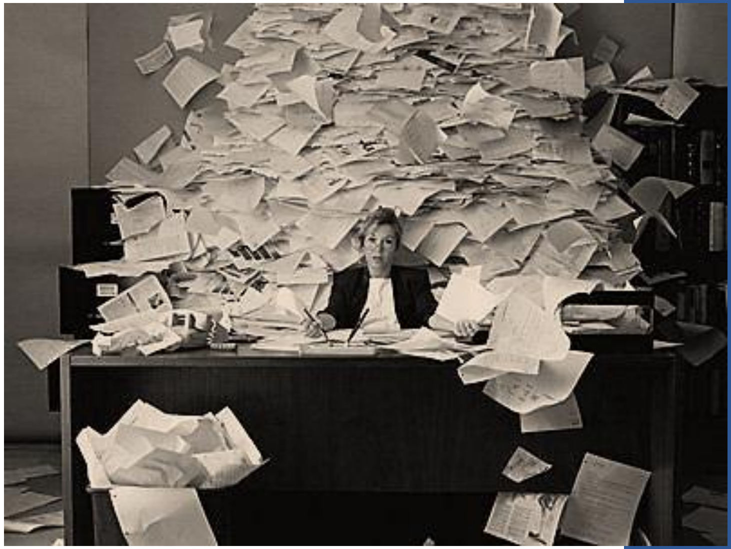
- Success in engineering an information systems is wholly dependent on success in engineering its data.

- Row data is used in all aspects on running a system; including daily transactions, customer profiling, strategy analysis, business performance management, decision making, predictive and descriptive analysis, etc…

- Data-driven information systems are now *in-escapable*.

Current Climate

- Information is everywhere
  - Companies, Businesses, Governments, Schools, etc…
- Information is kept in “documents”
  - Hard copies and/or soft copies
- Information must be maintained
  - Business flow, legal reasons, archive reasons, competitive reasons, decision support, etc…



http://www.businessinsider.com/wheres-that-darn-file-why-document-management-matters-2011-6

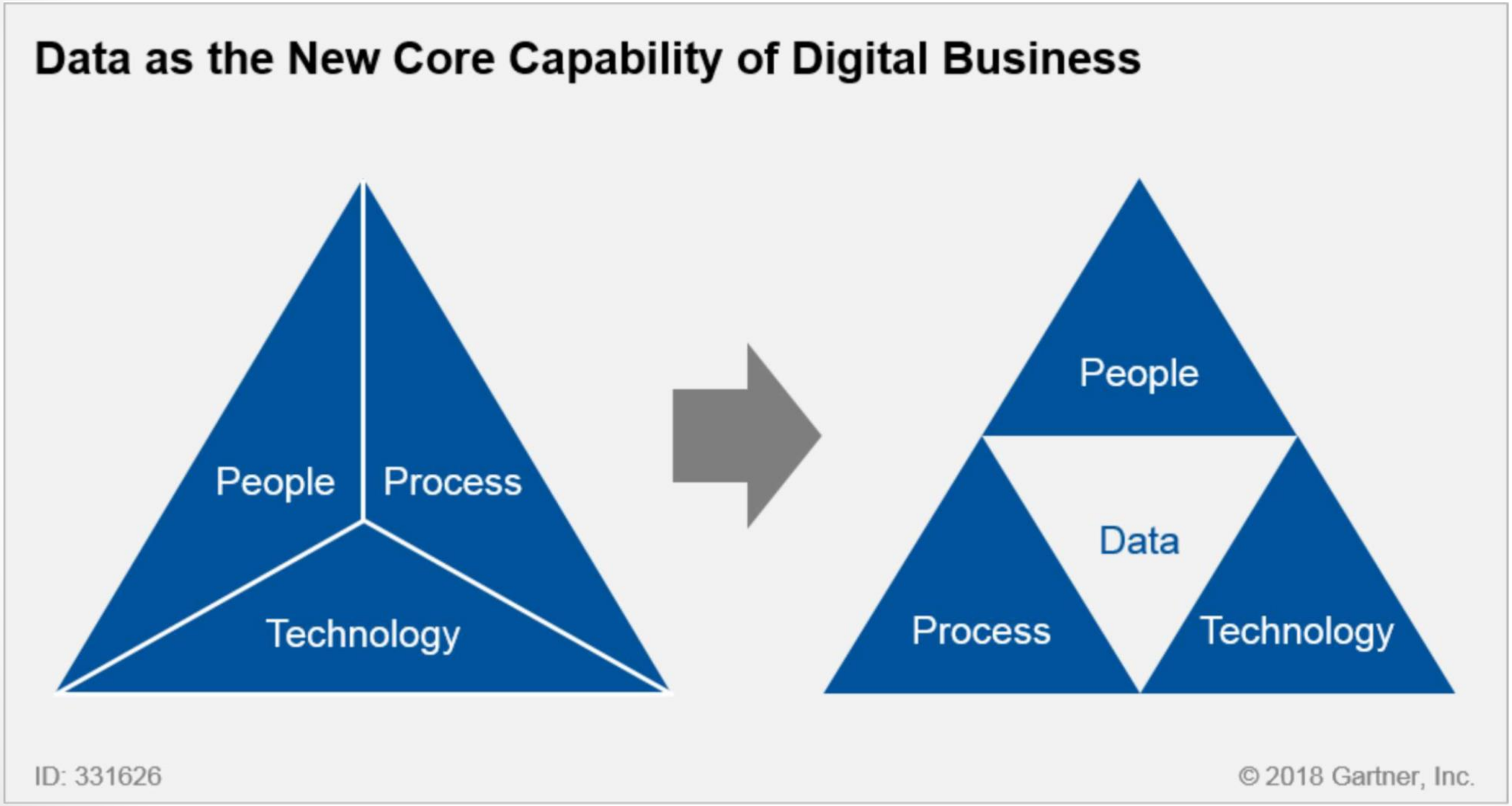
Current Climate

- Information is created, stored, processed, maintained and discarded.
  - “an Enterprise exists only as far as its ability to control information”

- Access, dissemination and use of information nowadays defy human understanding.

- Result, maintaining information in documents is expensive and unreasonable.

Data - Business



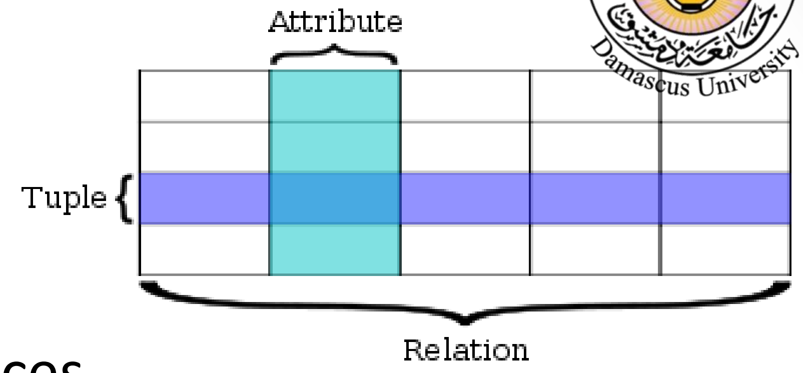
Data as the New Core Capability of Digital Business

ID: 331626

Data

- Collections of objects and their attributes.
- Building blocks of information!
- Can express all types of information when coupled with the *proper context*
- With the digital age, very simple to store and maintain.

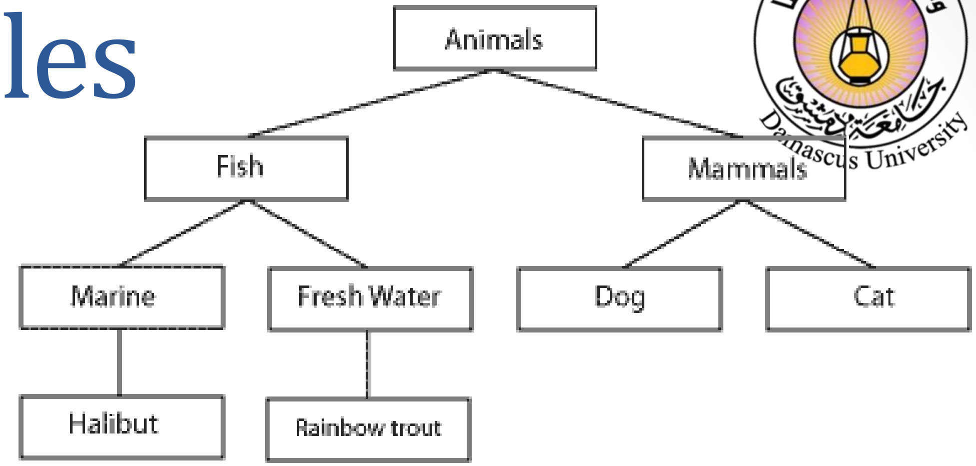
Data - Examples



- Relational Databases
  - Logical model: ERD
  - Physical model: tables, indices
  - Query language: SQL
- Not representing the real world.
- Only one type of data: the relation.
  - Conversion burden
- Homogeneous data.
- Other RDBMs limitations: changes to schema, short-lived transaction, etc…

Data - Examples

- Hierarchical Data
  - Logical model: Tree
  - Inheritance,
    IS-A relations, etc..



Animals

Fish

Mammals

Marine

Fresh Water

Dog

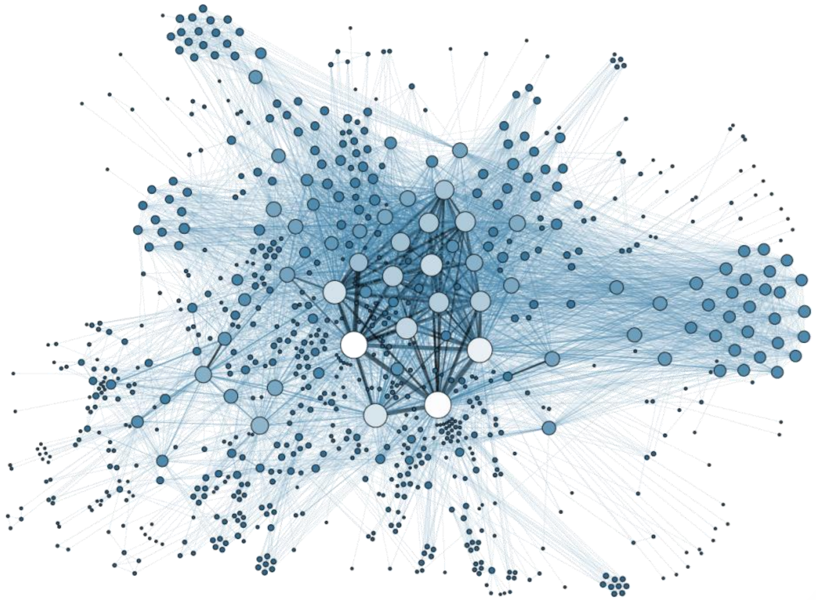
Cat

Halibut

Rainbow trout

http://mistercameron.com/2006/03/storing-hierarchical-data-in-a-database-part-1/

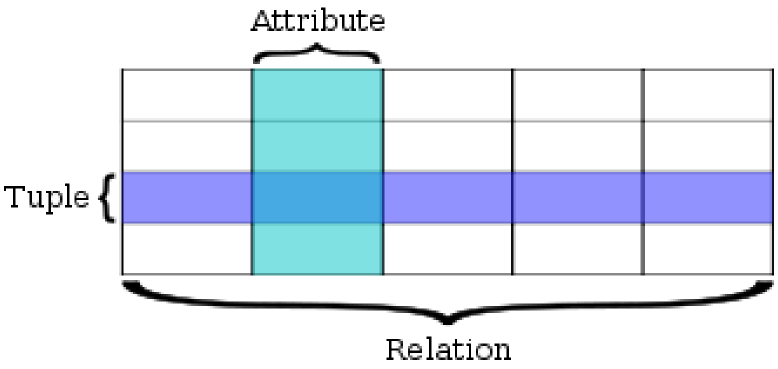
- Networked Data
  - Logical model:
    Graph, network
  - Recommendation,
    similarity, etc…

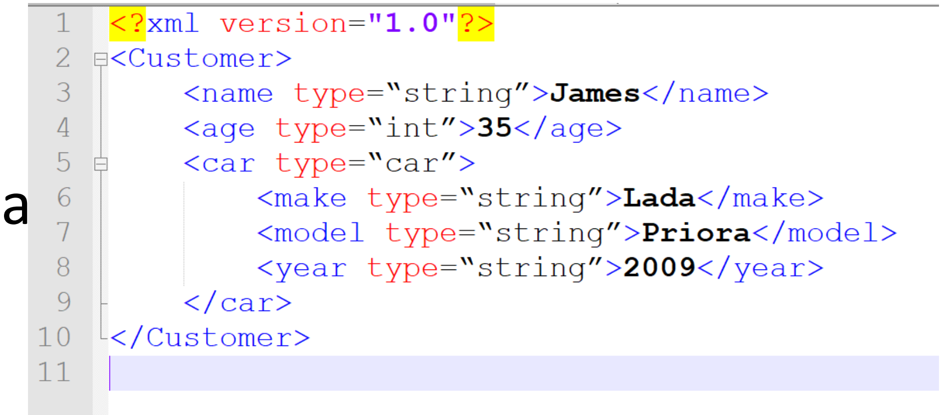


https://commons.wikimedia.org/wiki/File:Social_Network_Analysis_Visualization.png

Data - Types

- Three types of data:
  - Structured data
    - Often in databases
  - Semi-structured data
    - Often in XML files
  - Unstructured data
    - Often in the wild





```
<?xml version="1.0"?>
<Customer>
    <name type="string">James</name>
    <age type="int">35</age>
    <car type="car">
        <make type="string">Lada</make>
        <model type="string">Priora</model>
        <year type="string">2009</year>
    </car>
</Customer>

```

https://www.semanticscholar.org/paper/The-German-Traffic-Sign-Recognition-Benchmark%3A-A-Stallkamp-Schlipsing/22fe619996b59c09cb73be40103a123d2e328111

Data

- Data is everywhere!
  - Much more prevalent than documents

- Data must be maintained for proper business flow

- Data must be kept for future reference.

- What to keep?... *everything?*

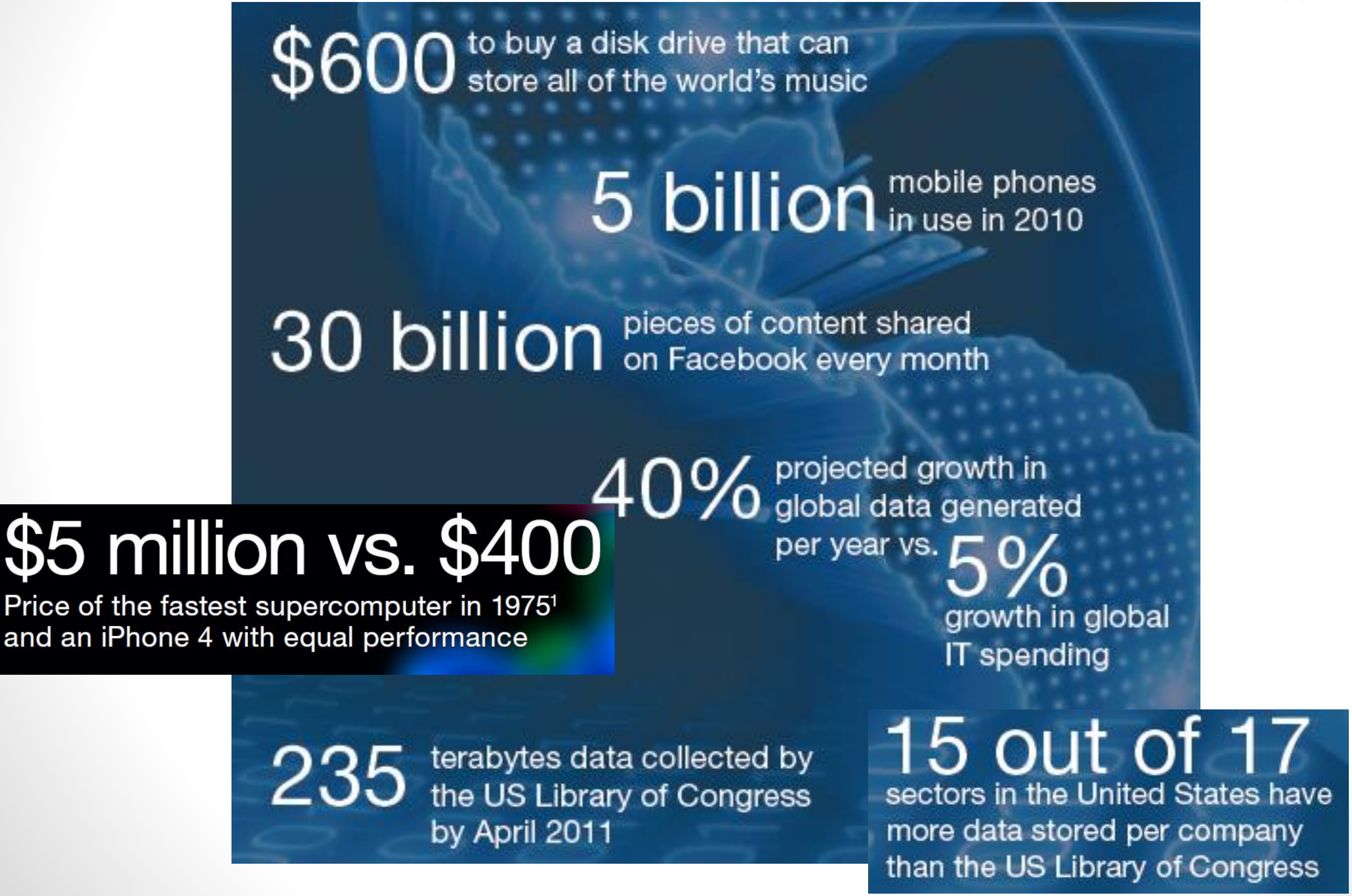
Data - Flood

- World’s data doubling every 2 years (Moore’s Law?)
  http://www.emc.com/about/news/press/2011/20110628-01.htm
- Global Data was estimated at 8 Zettabyte in 2015 (popular media)

> “According to computer giant IBM, 2.5 Exabytes [...] of data was generated every day in 2012. That's big by anyone's standards. 75% of data is unstructured, coming from sources such as text, voice and video” (2014)
> http://www.bbc.com/news/business-26383058

> “[referring to amazon] We are talking about 50 billion to 100 billion pieces of information a day” (2016)
> https://www.cnbc.com/2016/04/26/its-google-vs-amazon-to-create-the-biggest-data-base-in-history.html

Data - Flood



$600 to buy a disk drive that can store all of the world's music

5 billion mobile phones in use in 2010

30 billion pieces of content shared on Facebook every month

40% projected growth in global data generated per year vs. 5% growth in global IT spending

$5 million vs. $400
Price of the fastest supercomputer in 1975[1] and an iPhone 4 with equal performance

235 terabytes data collected by the US Library of Congress by April 2011

15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress

Data - Flood

- Important and unimportant data are lumped together.
- Data retrieval becomes burdensome with massive amounts of useless data to search.
- What to do to avoid drowning in digital waste?
- *We need to be **Data Literate**!*

Data - Literacy

- The ability to read and communicate data in context!
  - Data sources and constructs,
  - Analytical methods, and
  - Techniques applied.

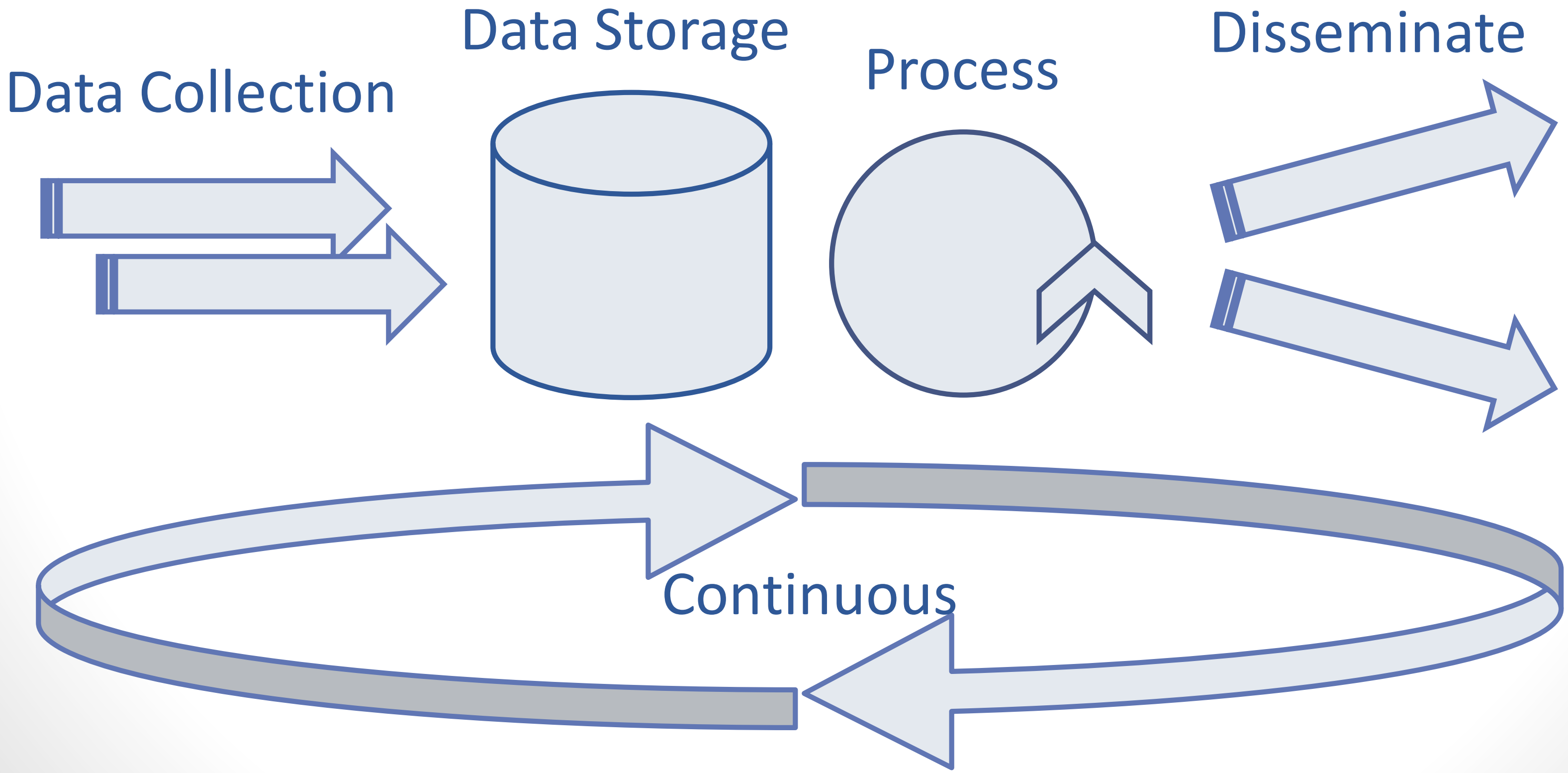
- The ability to describe use cases, applications and resulting value.
- As organizations become more data-driven, poor data literacy will become an inhibitor to growth.

- *We need to study the **Data Lifecycle**!*
  - *Where does it come from, how it is used, where does it go afterwards?*

Smarterwithgartner

Data Lifecycle

- We need to study the Data Lifecycle!
  - Where does it come from, how it is used, where does it go afterwards?



Data Lifecycle - Aspects

- Objectives for DLM: prioritize data and process
- Document retention: Based on what is important, identify important data and plan to maintain it.
- Minimalism: “Data should be discarded unless there is a good business and/or legal reason to retain it”
- Information Security: Protecting own information is fundamental for the survival of any organisation
- Authenticity of own records
- Retrievability of data
- Distribution Control: especially in the digital age!

Data Lifecycle - Collection

- Data collection methods are many.
- No one method is inherently better than another.
- Primary data sources include:
    - Interviews, Experimentation, Surveys, User shares own experiences, User photos, media, Gadgets info, Statistical analysis of samples, etc…
- Secondary data sources include:
    - Reviews, Perceived life-style, Inferred tags/labels, etc…

- Be critical about what to collect. Different data has different value for the business
- *Practice minimalism*! On an organisation level.

Data Lifecycle - Storage

- Data is usually stored in datasets in a digital format.
- No one format is inherently better than another.
- Data storage must be non-volatile and must provide abilities of maintenance, access and retrieval of data.

- Protecting data is fundamental to the survival of an organisation.
- *Security procedures* must be implemented in order to enforce data consistency (more on this later).

Data Lifecycle - Process

- Data is processed in order to support the work flow of an organisation.
- Can be seen as a multi-level type of process.
- What can we get from data?

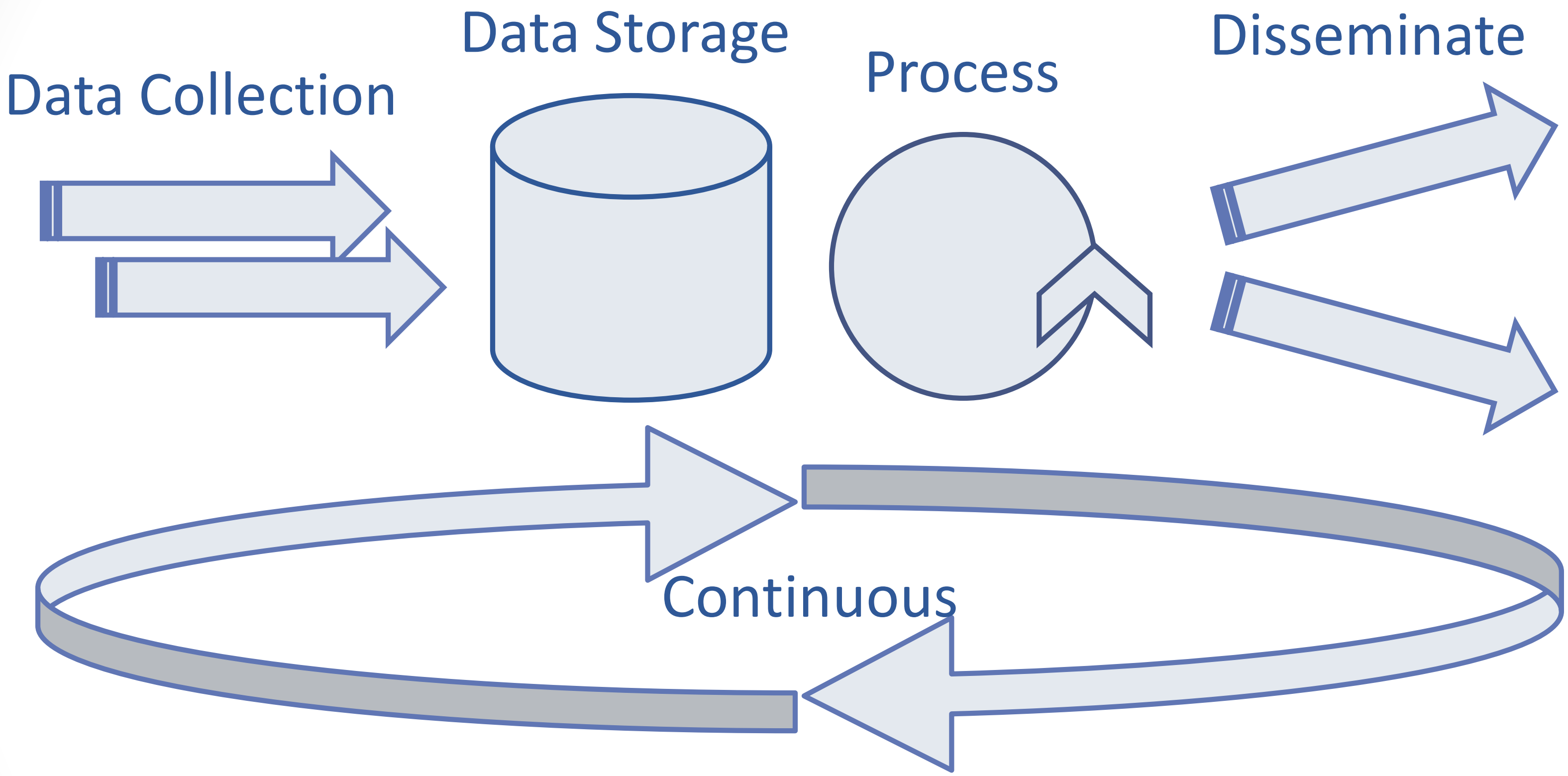
- Security procedures are also implemented in order to avoid unauthorised access to data.

Data Lifecycle - Dissemination

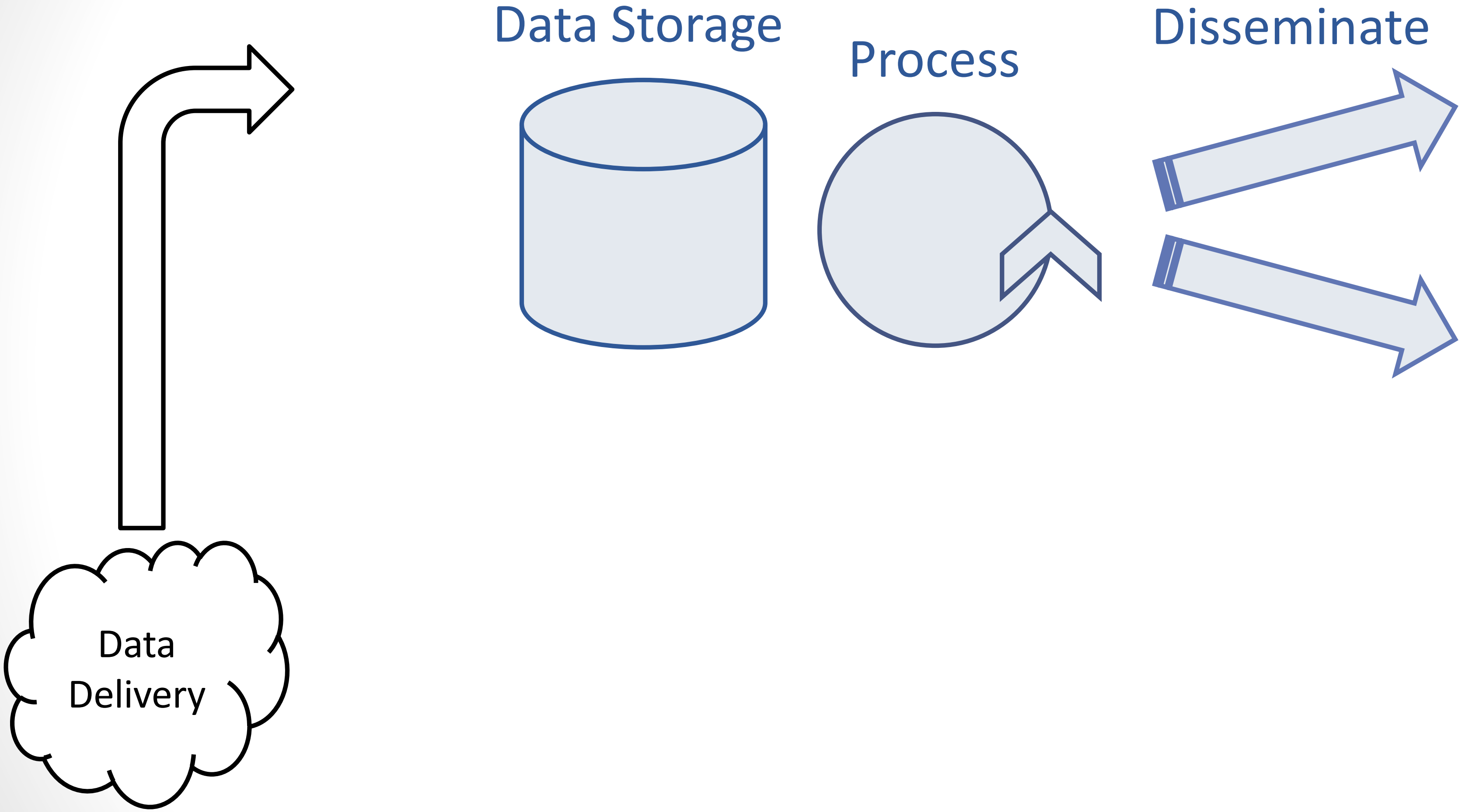
- Data is generally never destroyed.
  - Rather disseminated to applications or tools which use the data.
- Digital communication, reporting tools, BI tools, Knowledge creation, etc…
- Rarely, upon certain requests, data *is* discarded.

Break

Data Lifecycle



Over … Lifecycle



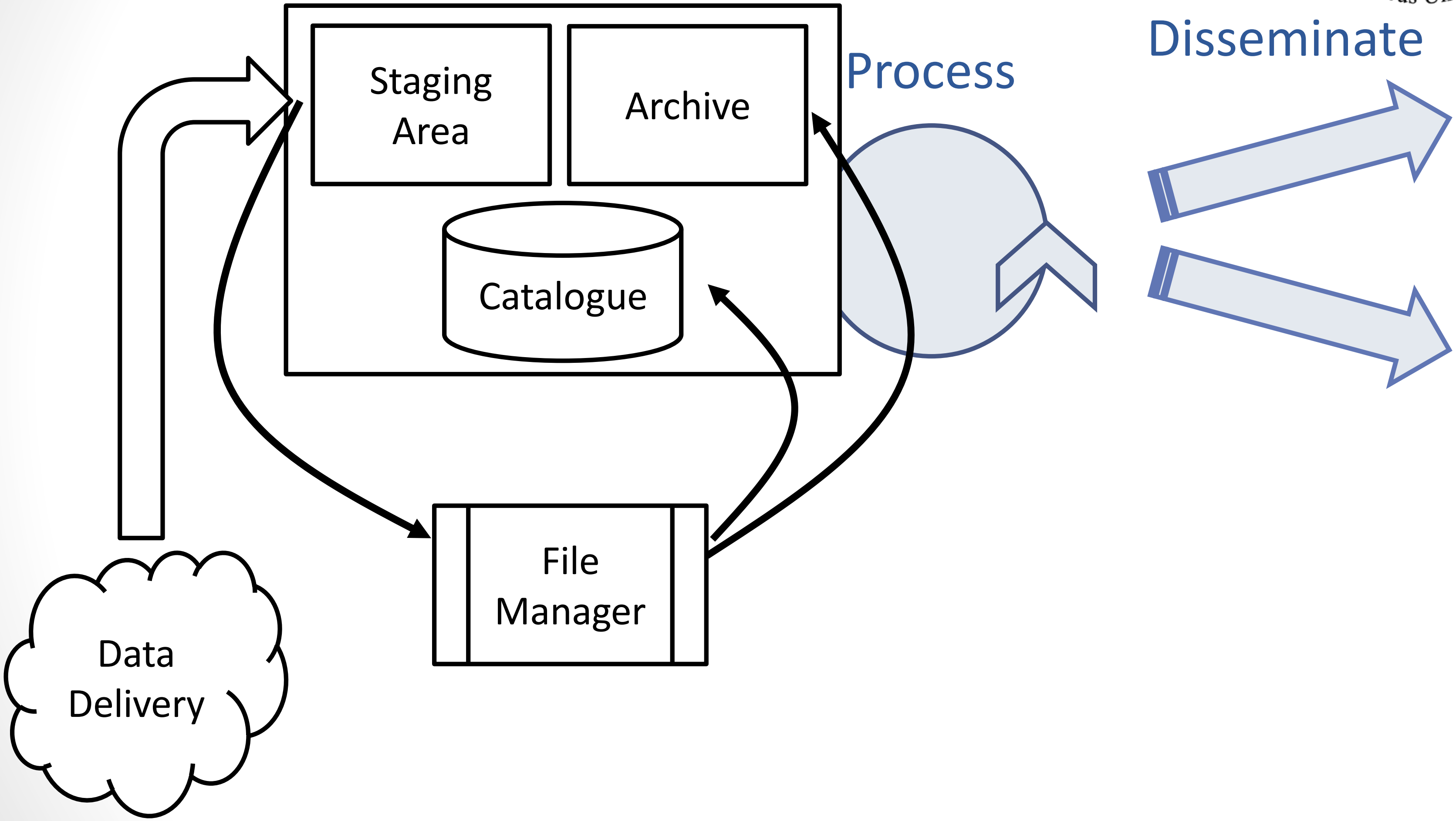
Data Storage

Process

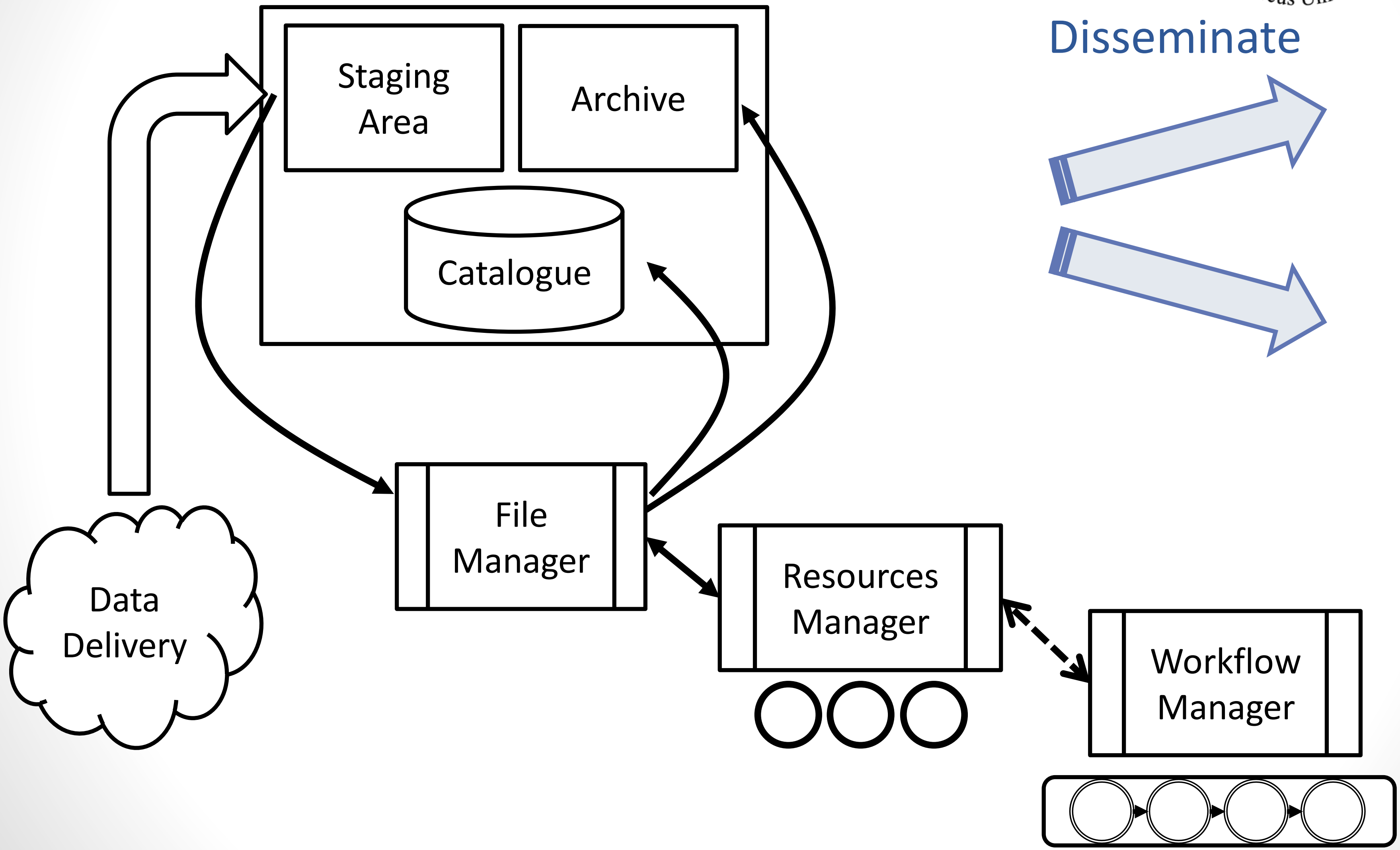
Disseminate

Data Delivery

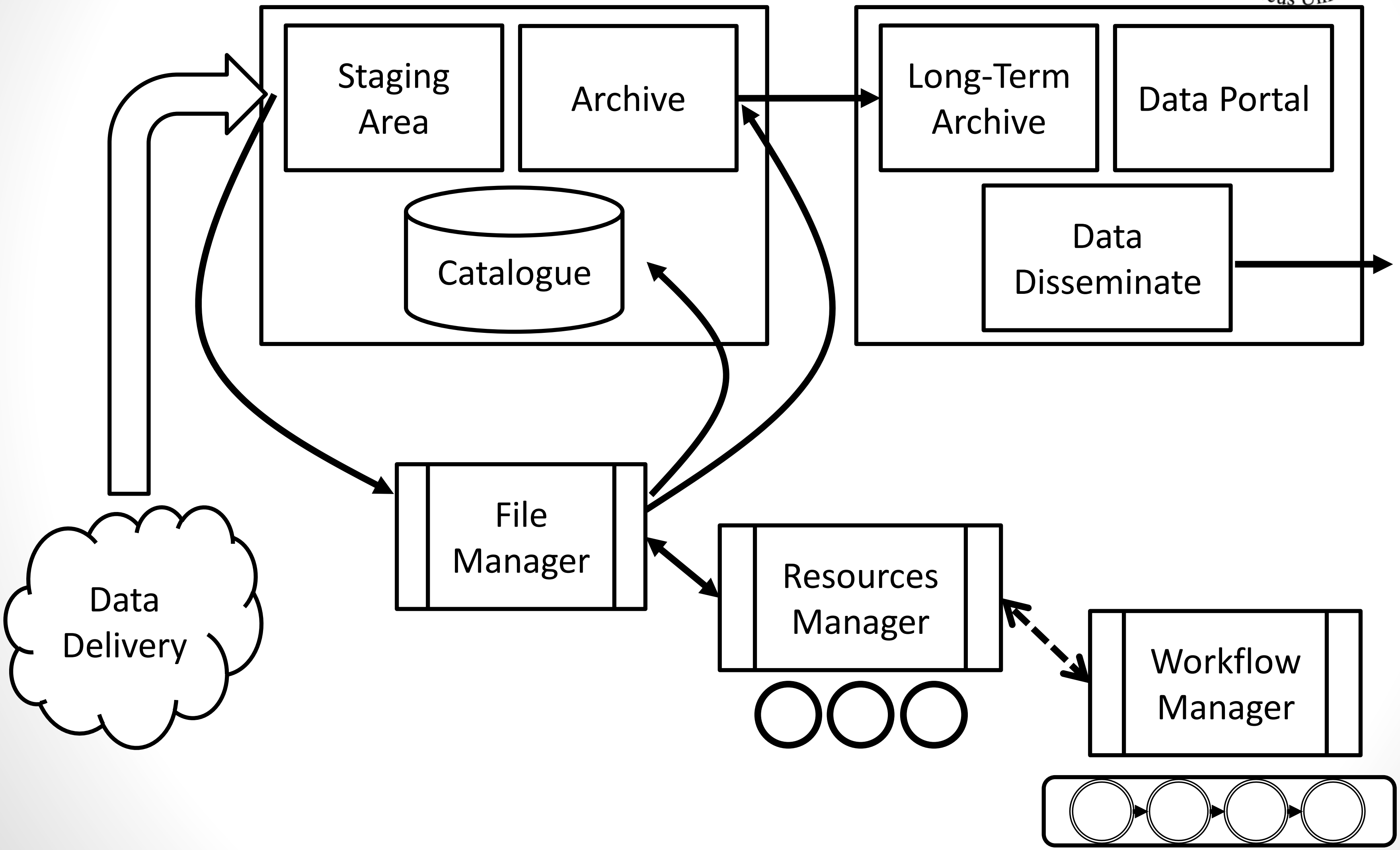
Overall … cycle



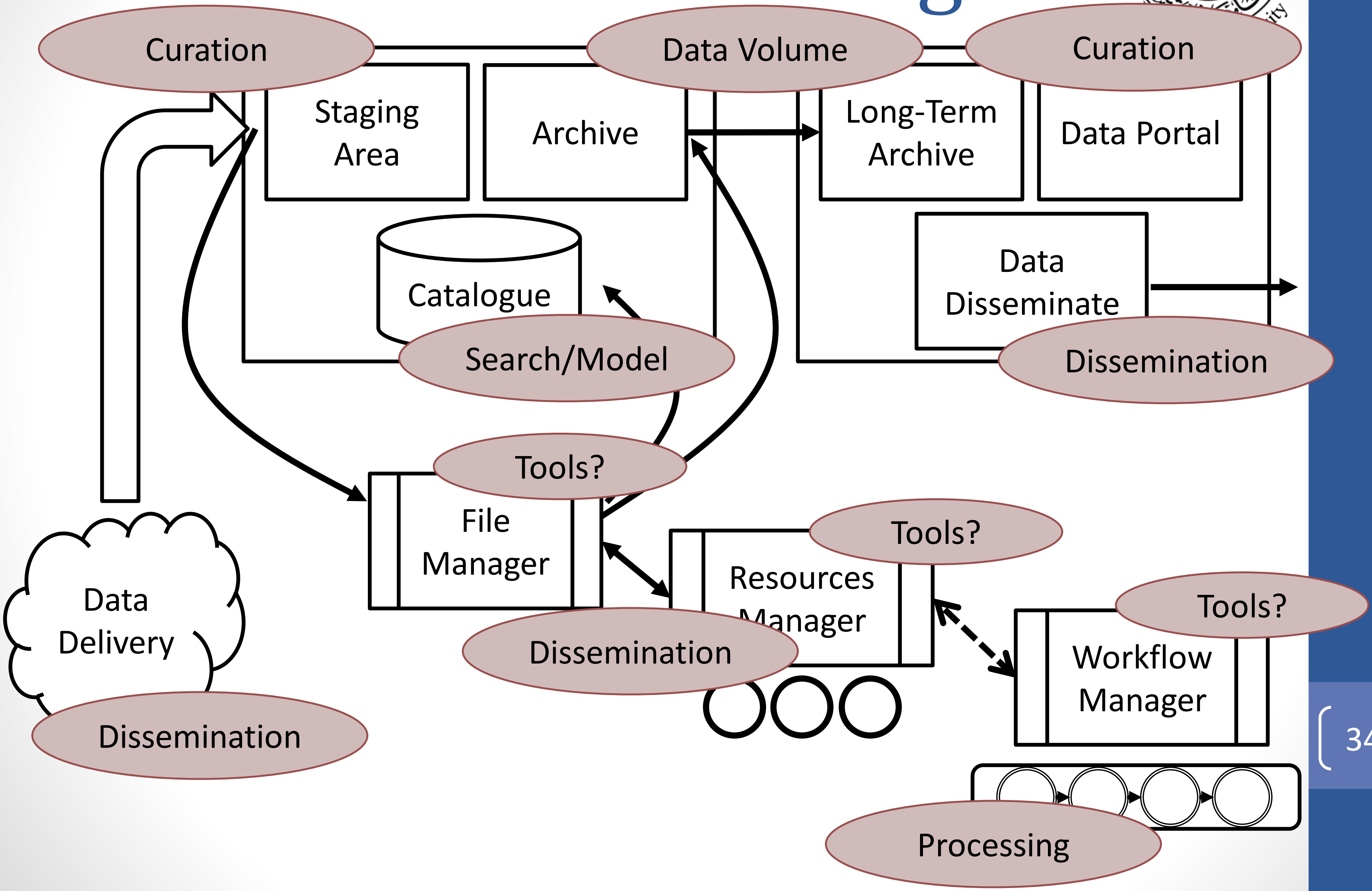
Overall Arch …



Overall Architecture



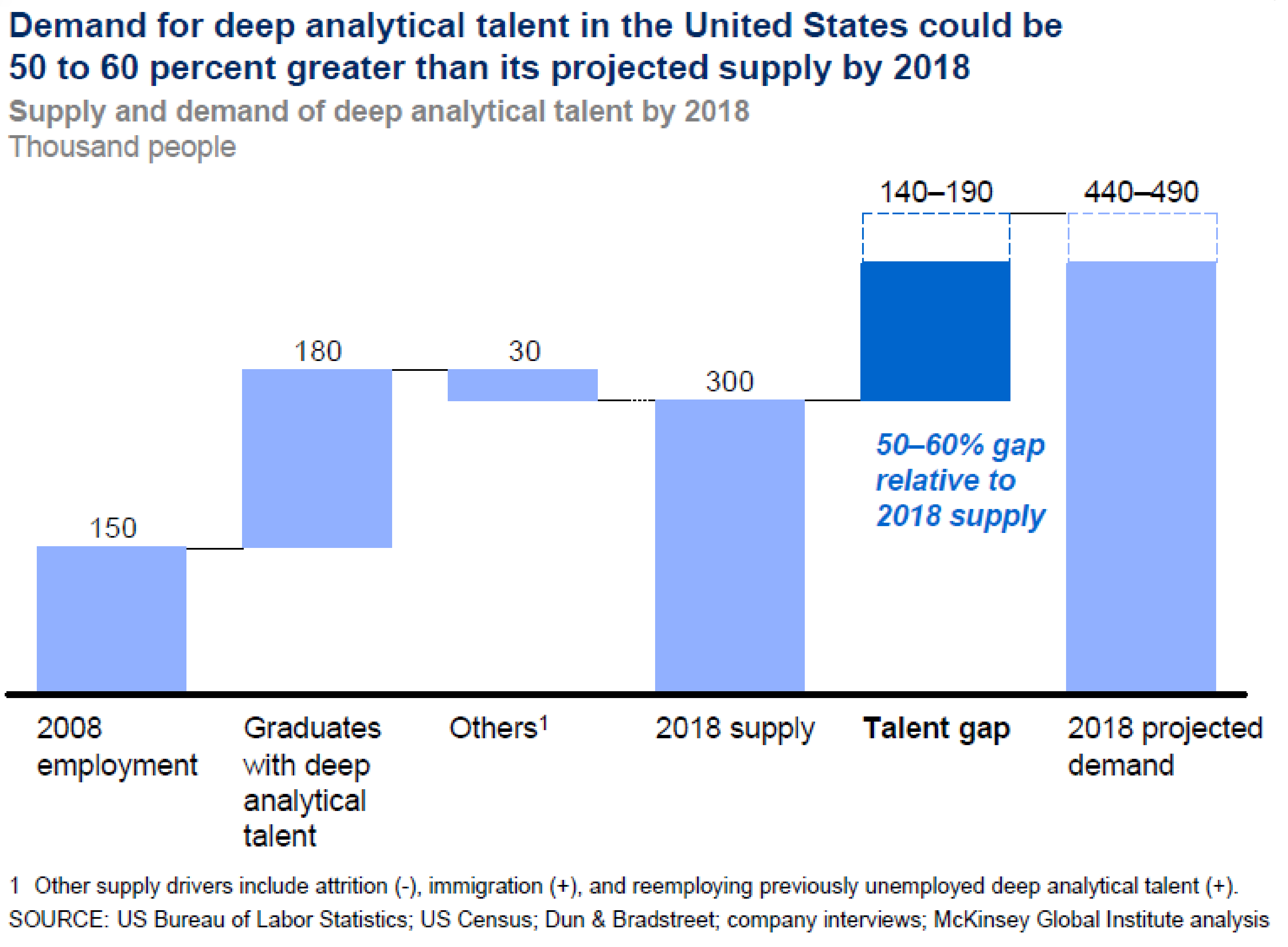
Architecture - Challenges



Architecture - Challenges

- Total Volume
  - Transfer/Process/In memory/Archive/etc…
- Data Dissemination
  - Protocols/APIs/Security/etc…
- Data Curation
  - Stand alone data/completeness vs. availability/timeliness/etc…
- Tools
  - Open Source vs. closed source/Availability and Agility/etc…
- Search and IR
  - Open/Guided/Constrained Queries
- Processing
  - Pipelining/Clouds/Grids/Clusters/etc…

Data in Business



Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018

Supply and demand of deep analytical talent by 2018

Thousand people

[1] Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).

SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey Global Institute analysis

Data in Business

- **Data Management Professional** (data developer) is an IT role responsible for supporting data and infrastructure.
  - Must know technologies such as Apache Hadoop, Spark, SQL, RDBMS, etc…

- **Data Scientist** is concerned with the entirety of data aspects, regardless of technology and tools. Preforms predictive analysis and analyses results.
  - Must be fluent in statistics, some high-level languages and communication skills.

Source: NetCom Learning

Data in Business

- **Data Engineer** designs and implements infrastructure.
  - Must be up-to-date on data technology and complex data concepts.
- **Business Analyst** understands data and presents in-depth analysis including reports, dashboards and BI.
  - Must have high-level analytical skills and communication skills.
- **Machine learning Practitioner** leverages data in order to uncover knowledge.
  - Must know statistics, algebra, multiple programming languages, and machine learning algorithms.

Source: NetCom Learning

DataOps

- **DevOps**: Customer-value-driven approach to deliver solutions using agile methods, collaboration and automation.
- ***DataOps****: Improves the flow of data to points of consumption in the business. It operationalizes data pipelines and workflow orchestration to specific consumer use cases.*

Only 1 in 10 organizations are able to get 75% or more of their AI model prototypes into production!

DataOps

- Use DataOps to link “Can we do this?” to “How do we provide an optimized, governed data-driven product.

- Use DataOps to support data’s use as a business enabler.
  - Drive collaboration with the business-unit stakeholders.

Data Storage + Management + Analysis
Big Data + Data Quality + Data Mining

References

Assessed Reading:

- Data Life Cycle Management. 2005. http://corporate.findlaw.com/law-library/data-life-cycle-management.html
- G. Paul, R. Copple. Dealing with data. Business Law Today. 2005. http://www.copplelaw.com/uploads/BLTMarchAprData.pdf
- T. Connolly, C. Begg. Database Systems: A Practical Approach to Design, Implementation, and Management 6th edition. 2015.

Up next

- Big Data* – Data Warehousing
- Big Data – MapReduce Model