CSE 4510/5310
BIG DATA

Instructor: Fitzroy Nembhard, Ph.D.

Week 1

Introduction:
What is Big Data?

FLORIDA TECH
FLORIDA'S STEM UNIVERSITY®

Distribution

- All slides included in this class are for the exclusive use of students and instructors associated with Management and Processing of Big Data (CSE 4510/5310) at the Florida Institute of Technology.
- Redistribution of the slides is not permitted without the written consent of the author.

Goals

- To discuss the basics of Big Data
- To formulate a definition of Big Data
- To explore the buzz and interest in Big Data
- To discuss the 5 Vs of Big Data
- To understand Big Data and Bias
- To discuss the average sizes of data being processed
- To understand datafication
- To explore Big Data applications, architectures, tools, and components
- To explore models for processing and analyzing data

1.1 BIG DATA - BACKGROUND

Interest in Big Data

the WHITE HOUSE
PRESIDENT BARACK OBAMA

BRIEFING ROOM | ISSUES | THE ADMINISTRATION | 1600 PENN

Briefing Room

- Your Weekly Address
- Speeches & Remarks
- Press Briefings
- Statements & Releases
- White House Schedule
- Presidential Actions
  - Executive Orders
  - Presidential Memoranda
  - Proclamations
- Legislation
  - Pending Legislation
  - Signed Legislation
  - Vetoed Legislation
- Nominations & Appointments
- Disclosures

The White House

For Immediate Release March 29, 2012

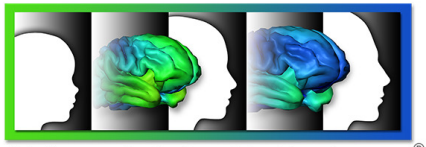
PRESS RELEASE: Obama Administration Unveils "Big Data" Initiative: Announces $200 Million in New R&D Investments

Contact: Rick Weiss 202 456-6037
rweiss@ostp.eop.gov
Lisa-Joy Zgorski 703 292-8311
lisajoy@nsf.gov

Aiming to make the most of the fast-growing volume of digital data, the Obama Administration today announced a "Big Data Research and Development Initiative." By improving our ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help solve some the Nation's most pressing challenges.

To launch the initiative, six Federal departments and agencies today announced more than $200 million in new commitments that, together, promise to greatly improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of digital data.

2012



- The ABCD Data Repository houses all data generated by the Adolescent Brain Cognitive DevelopmentSM (ABCD) Study. The ABCD Study(R) is a prospective longitudinal study starting at the ages of 9-10 and following participants for 10 years. The study includes a diverse sample of nearly 12,000 youth enrolled at 21 research sites across the country. It measures brain development (via structural, task functional, and resting state functional imaging), social, emotional, and cognitive development, mental health, substance use and attitudes, gender identity and sexual health, biospecimens, as well as a variety of physical health, and environmental factors. In addition, various external databases have been linked with ABCD Study data providing information about local conditions for environment, poverty, pollution, school, and policy, as examples. These can be used for providing context when evaluating behavioral and brain development. For more information about the ABCD Study, please visit the ABCD Study website.

2022

FLORIDA TECH

Interest in Big Data

Request for Information on the Federal Big Data Research and Development Strategic Plan Update

A Notice by the National Science Foundation on 08/03/2022

AGENCY:

Networking and Information Technology Research and Development (NITRD) National Coordination Office (NCO), National Science Foundation (NSF).

ACTION:

Request for Information (RFI); extension of comment period.

SUMMARY:

On July 1, 2022, the NITRD NCO and NSF, as part of the NITRD Big Data interagency working group (BD IWG), published in the **Federal Register** a document entitled "Request for

Start Printed Page 47474

Information on the Federal Big Data Research and Development Strategic Plan Update". Through this RFI, the NITRD NCO seeks input from the public, including academia, government, business, and industry groups of all sizes; those directly performing Big Data research and development (R&D); and those directly affected by such R&D, on ways in which the strategic plan should be revised and improved. The public input provided in response to this RFI will assist the NITRD BD IWG in updating the *Federal Big Data Research and Development Strategic Plan*. In response to requests by prospective commenters that they would benefit from additional time to adequately consider and respond to the RFI, the NITRD NCO and NSF have determined that an extension of the comment period until August 17, 2022, is appropriate.

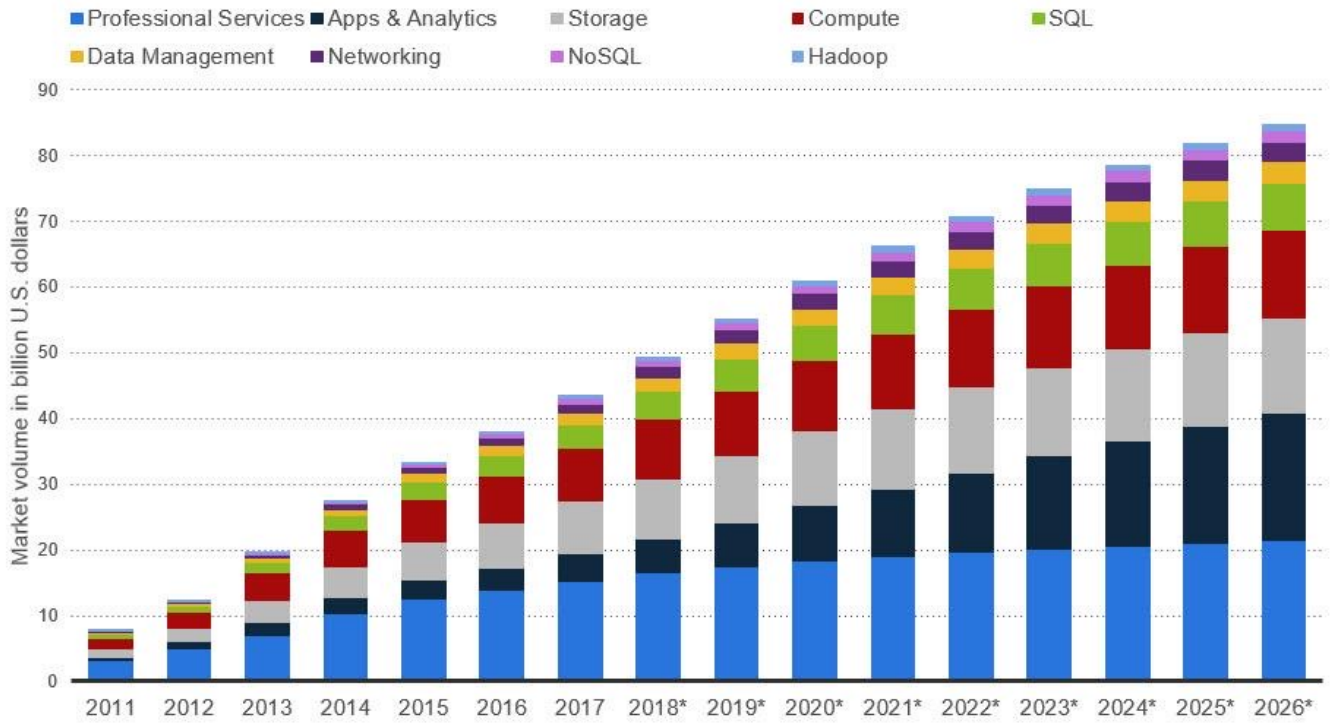
DATES:

The end of the comment period for the document entitled "Request for Information on the Federal Big Data Research and Development Strategic Plan Update", published on July 1, 2022 (87 FR 39567 (/citation/87-FR-39567)), is extended from July 29, 2022, until on or before 11:59 p.m. (ET) August 17, 2022.

Interest in Big Data

Big Data Market Worldwide Segment Revenue Forecast 2011-2026

Big Data Market Forecast Worldwide from 2011 to 2026, by segment (in billion U.S. dollars)



statista

What is Big Data? – New York Times

"Big Data is a vague term, used loosely, if often, these days.

But put simply, the catchall phrase means three things:

First, it is a bundle of technologies.

Second, it is a potential revolution in measurement.

And third, it is a point of view, or philosophy, about how decisions will be—and perhaps should be—made in the future.

— Steve Lohr

The New York Times

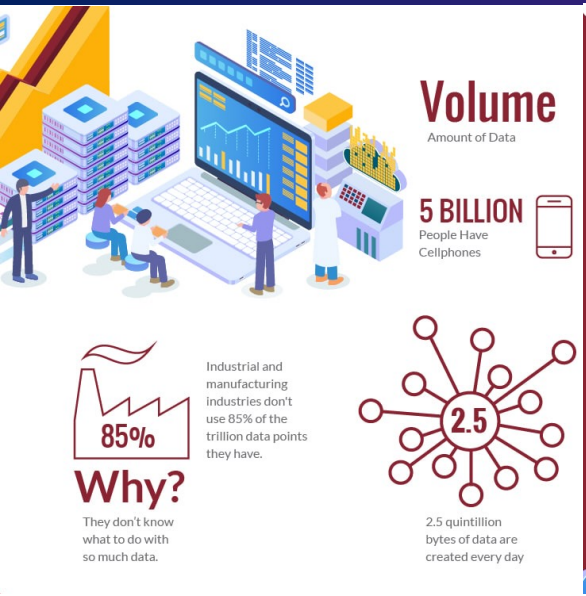
What is Big Data?

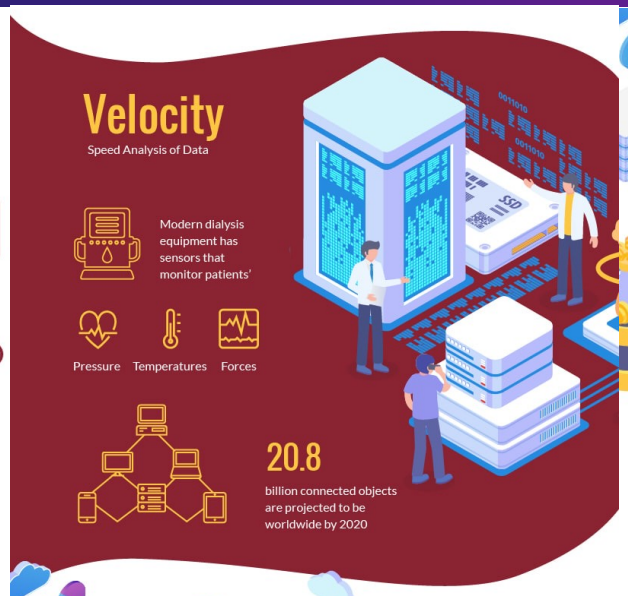
- Many of the methods and techniques we're using—and the challenges we're facing now—are part of the evolution of everything that's come before.
- **"Big" is a moving target** - when the size of the data outstrips the state-of-the-art current computational solutions (in terms of memory, storage, complexity, and processing speed) available to handle it
- **"Big" is when you can't fit it on one machine**
- **Big Data is a cultural phenomenon**. It describes how much data is part of our lives, precipitated by accelerated advances in technology
- **The 4 Vs: Volume, variety, velocity, and value**. Many people are circulating this as a way to characterize Big Data.
- (Note that there are 5, 7, 10 and 42 Vs as well – See https://www.elderresearch.com/blog/42-v-of-big-data)

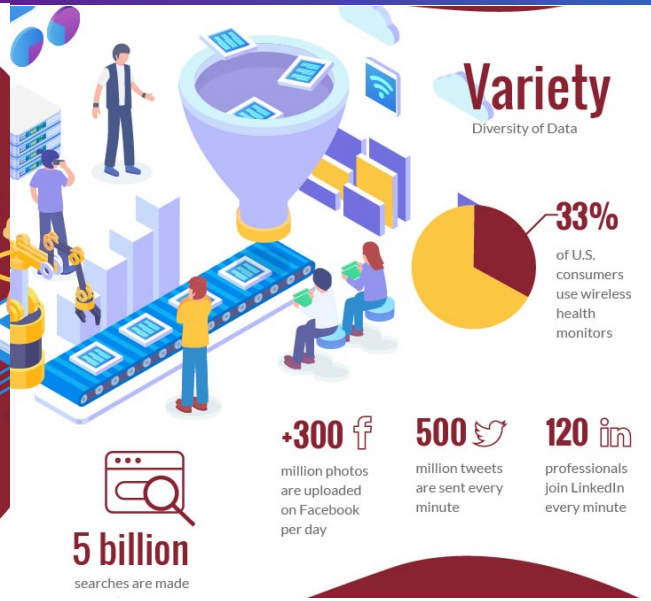
What is Big Data – A Definition

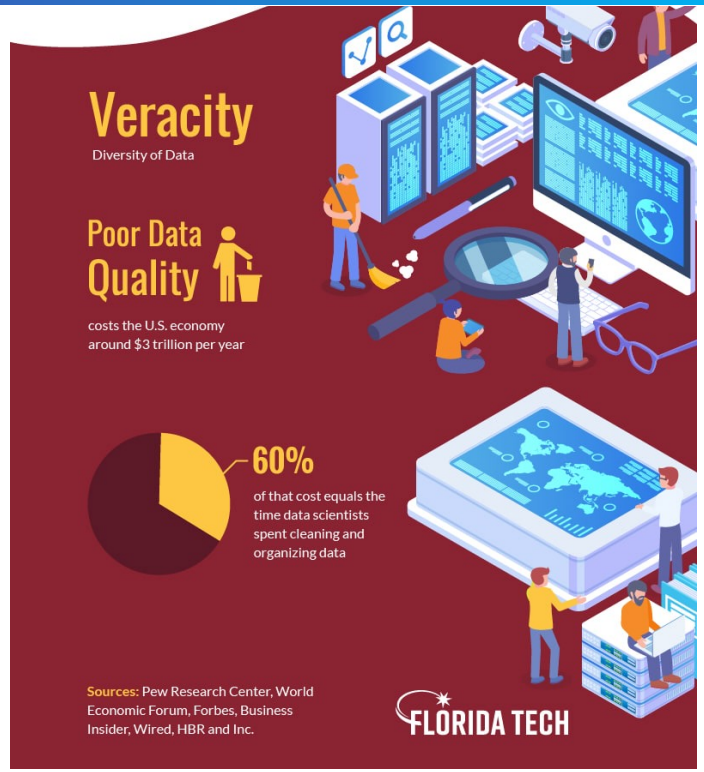
- Big data is an umbrella term for a collection of data sets so large and complex that it becomes difficult to process them using traditional data management tools.

The Four Vs of Big Data



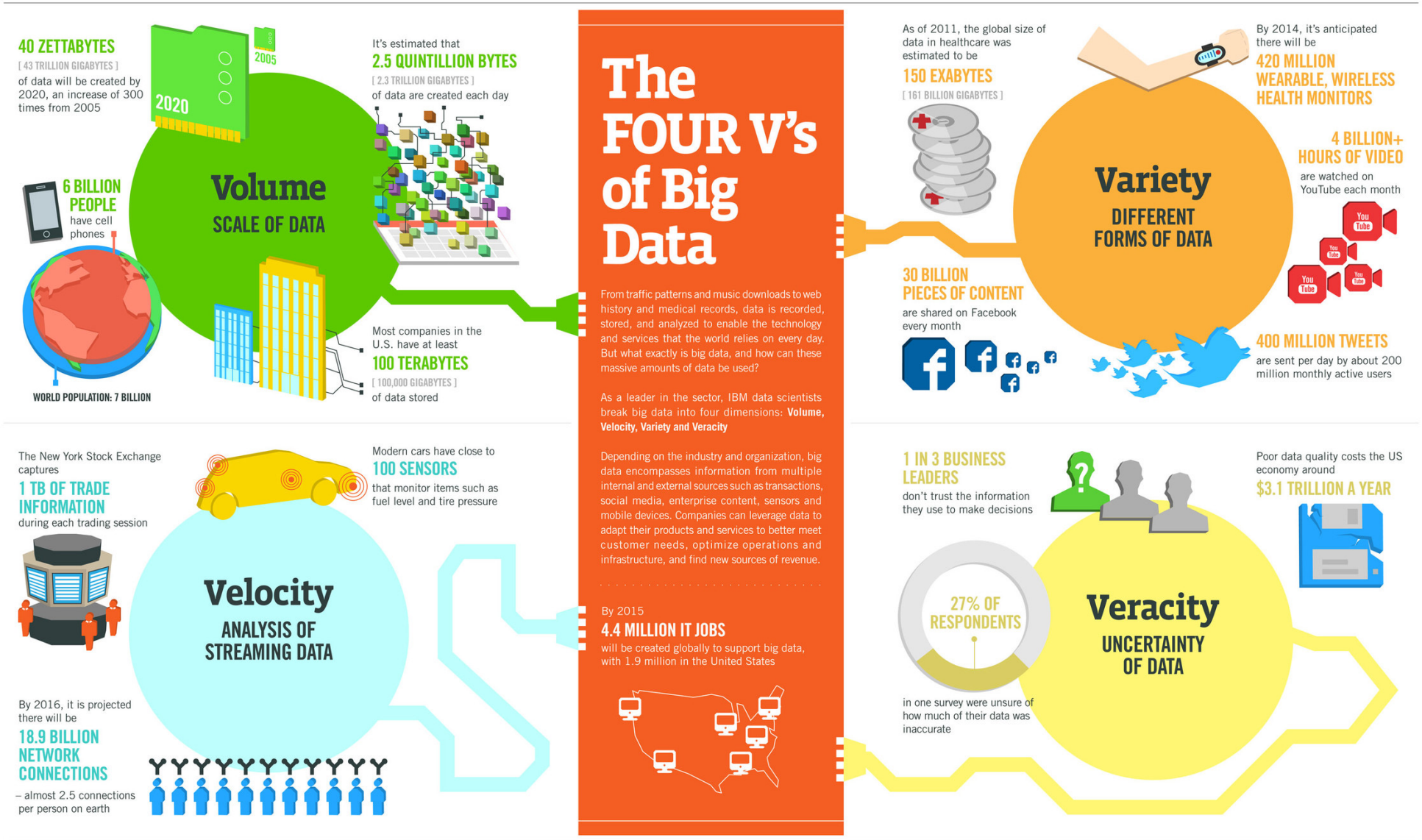






Source: https://www.floridatechonline.com/blog/information-technology/infographic-the-4-vs-of-big-data/

What are the Four (4) Vs of Big Data?



What is Big Data?

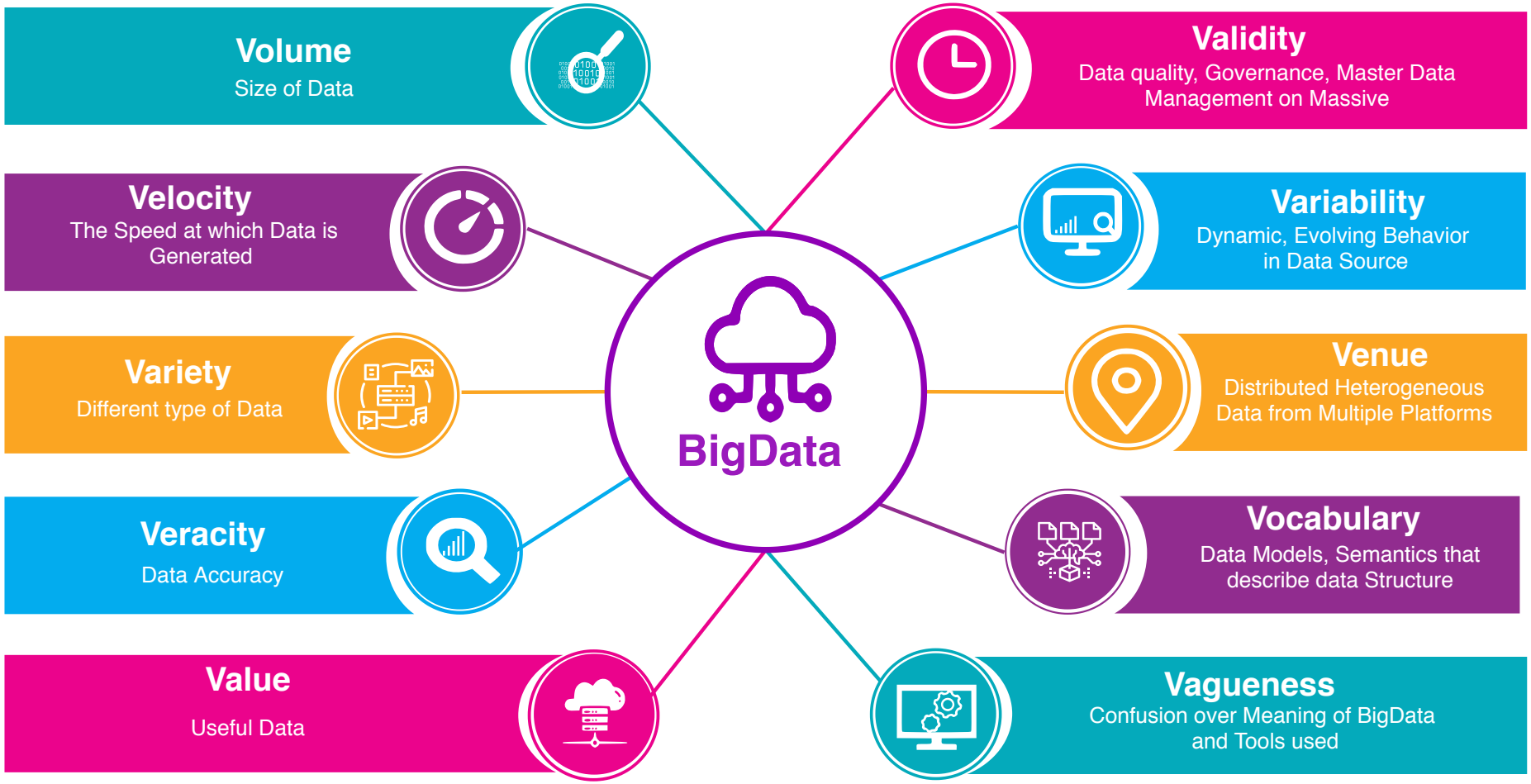


Image recreated from https://www.xenonstack.com/images/blog/10-vs-of-big-data.png

A Tiered Definition of Data Sizes

- The significance lies more in the different orders of magnitude rather than hard size limits. For example, on a very powerful computer, small data might be on the order of 10s of gigabytes, but not on the order of terabytes

Dataset type	Size range	Fits in RAM?	Fits on local disk?
Small dataset	Less than 2–4 GB	Yes	Yes
Medium dataset	Less than 2 TB	No	Yes
Large dataset	Greater than 2 TB	No	No

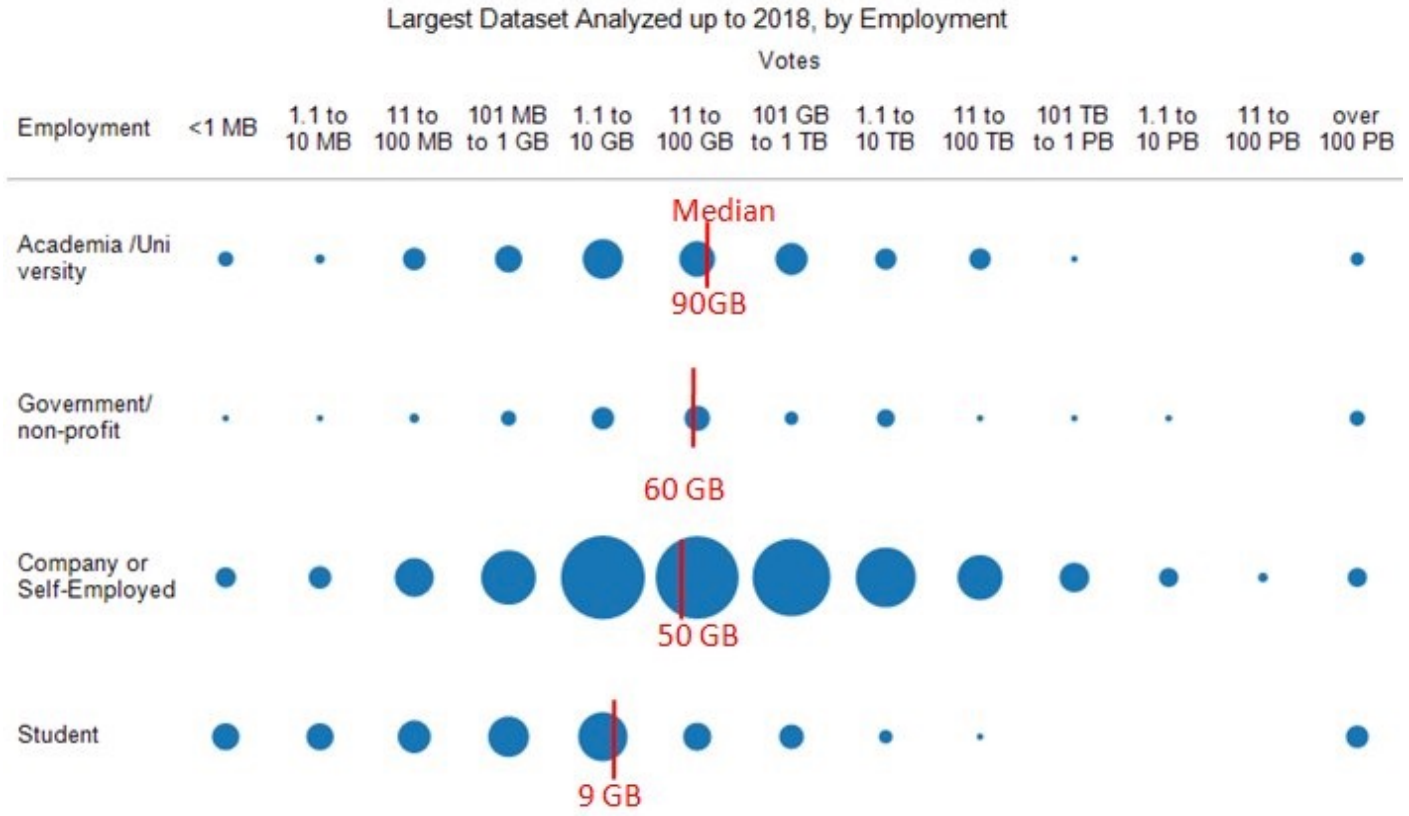
Big Data and Bias

- Even if we have access to all of Facebook's or Google's or Twitter's data corpus, any inferences we make from that data should not be extended to draw conclusions about humans beyond those sets of users, or even those users for any particular day.
- The uncertainty created by the sampling process is part of the ***sampling distribution***. Like that 2010 movie *Inception* with Leonardo DiCaprio, where he's in a dream within a dream within a dream, it's possible to instead think of the complete corpus of a certain kind of data at a company (e.g., emails at BigCorp) as not the population but as a sample

Poll: What size Data are people handling?

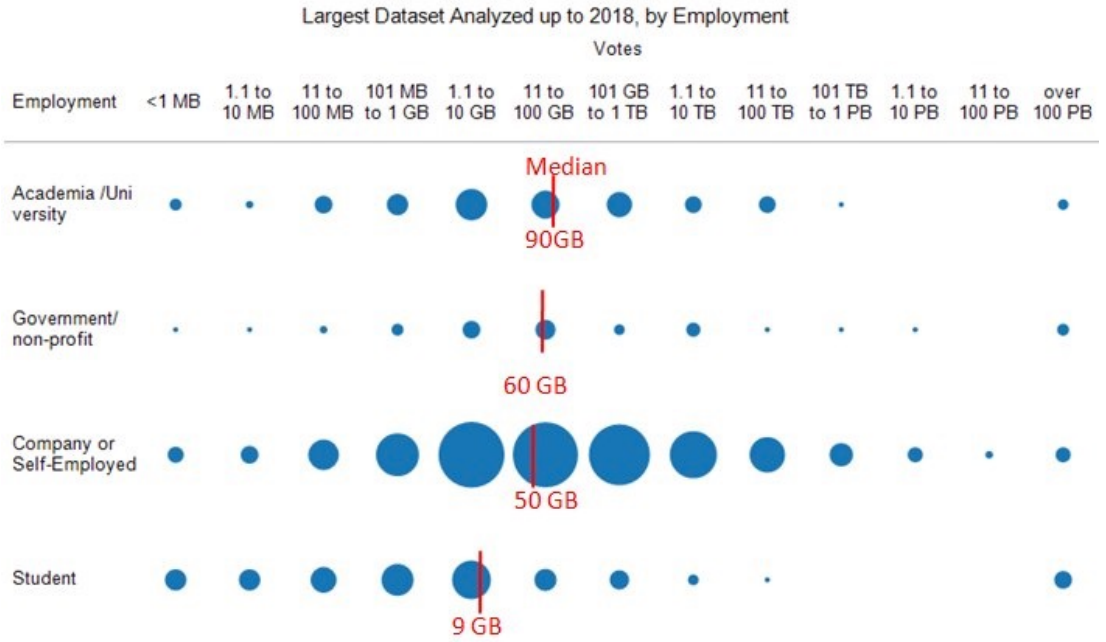
Largest Datasets Analyzed in 2013

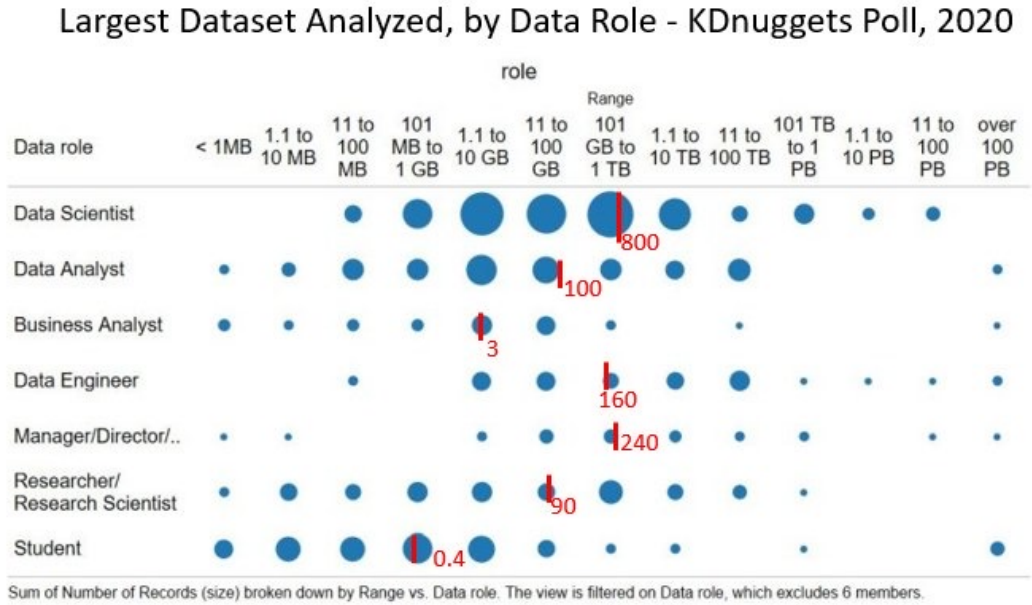
Less than 1 MB (12)	3.7%
1.1 to 10 MB (8)	2.5%
11 to 100 MB (14)	4.3%
101 MB to 1 GB (50)	15.5%
1.1 to 10 GB (59)	18%
11 to 100 GB (52)	16%
101 GB to 1 TB (59)	18%
1.1 to 10 TB (39)	12%
11 to 100 TB (15)	4.7%
101 TB to 1 PB (6)	1.9%
1.1 to 10 PB (2)	0.6%
11 to 100 PB (0)	0%
Over 100 PB (6)	1.9%



https://www.kdnuggets.com/

Poll: What size Data are people handling?

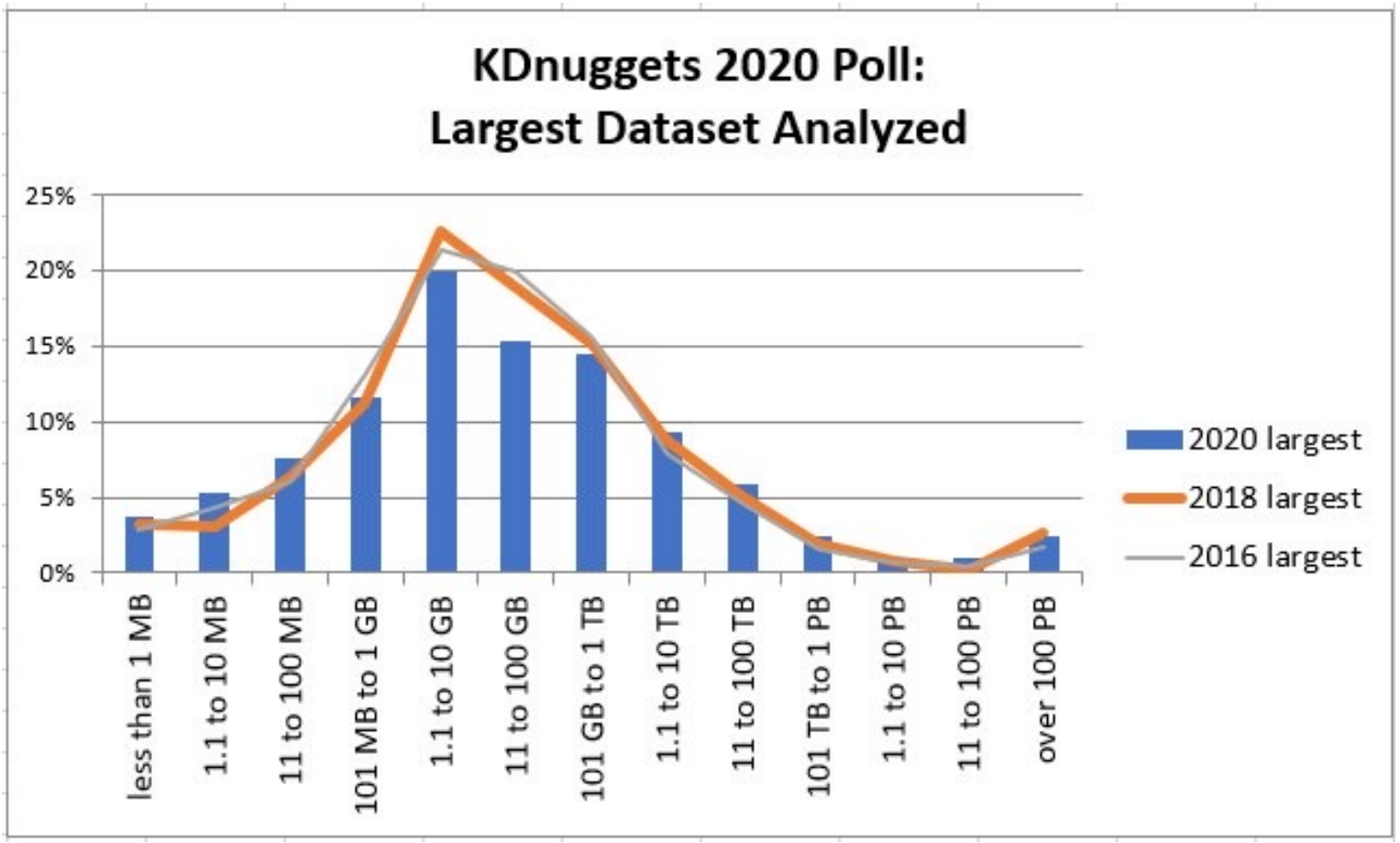




https://www.kdnuggets.com/

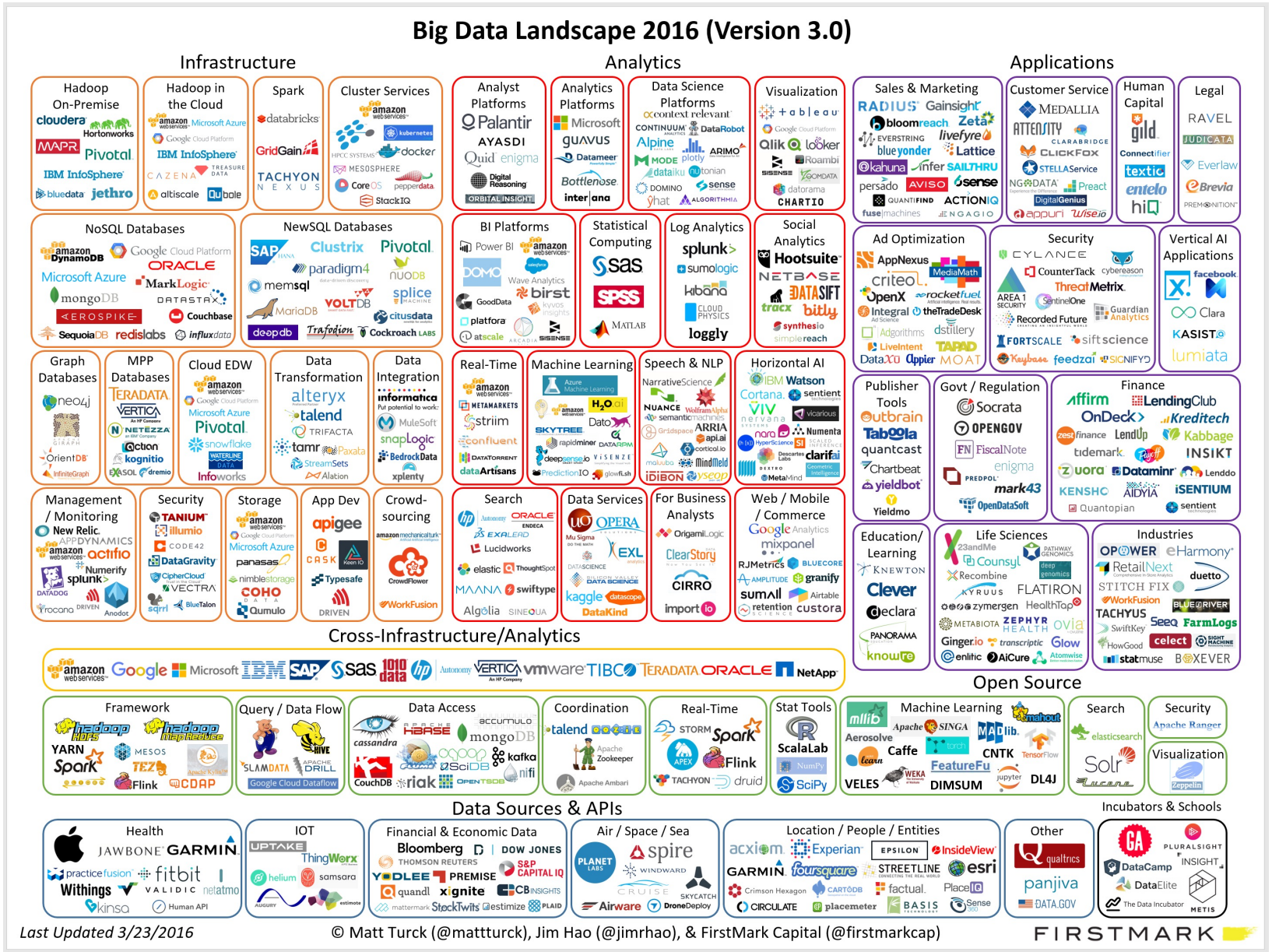
Poll: What size Data are people handling?

- KDnuggets Poll (562 participants): Largest Dataset Analyzed, 2020, 2018, 2016



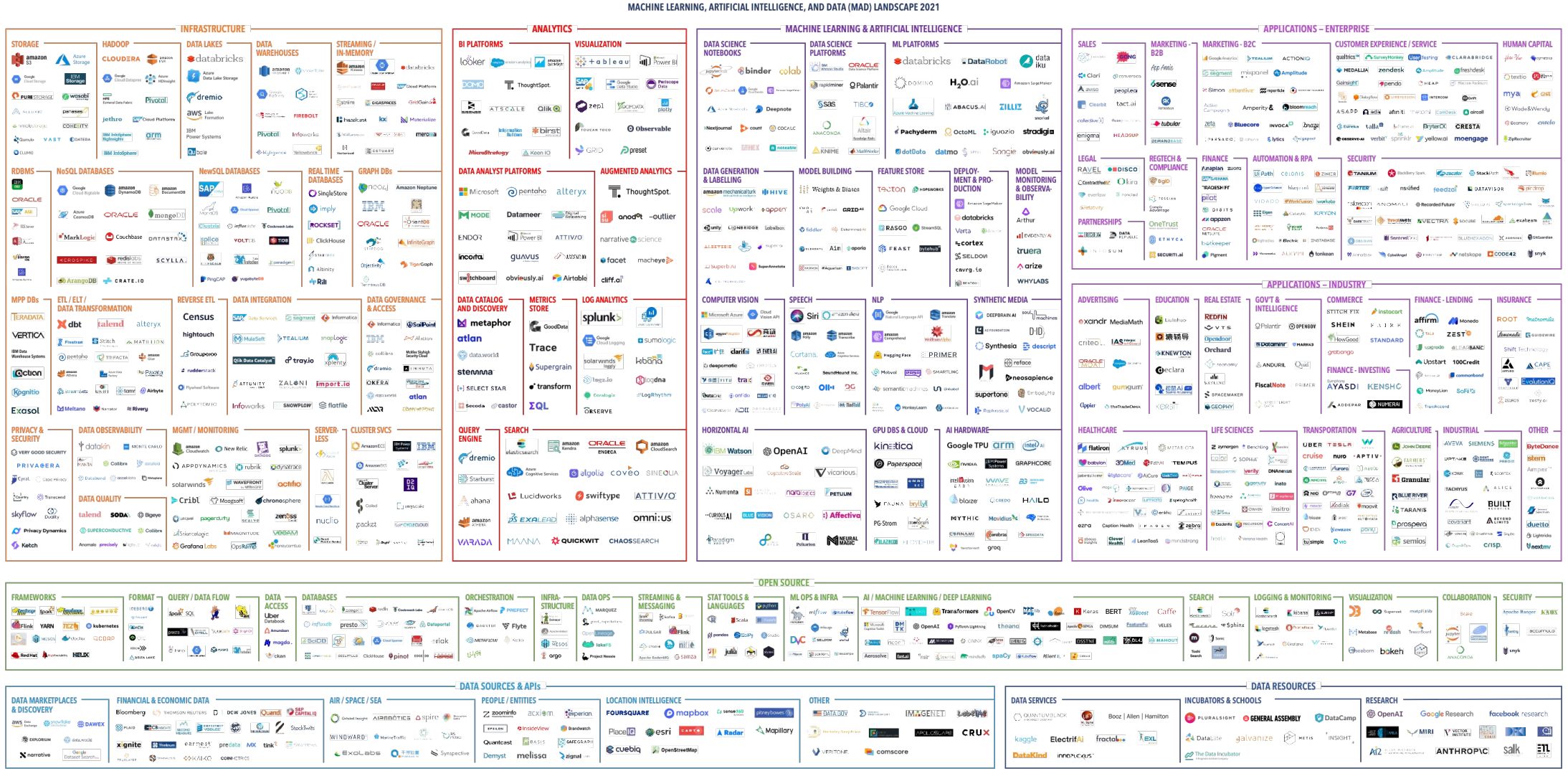
https://www.kdnuggets.com/

Big Data Landscape 2016



FLORIDA TECH

Big Data Landscape 2021



What is Datafication?

- Datafication as a process of “taking all aspects of life and turning them into data.”
- We quantify friendships with “likes”: it’s the way we do everything, online or otherwise
- Google’s augmented-reality glasses datafy the gaze.
- Twitter datafies stray thoughts.
- LinkedIn datafies professional networks.

Types of Big Data

- **Structured Data** owns a dedicated data model, It also has a well-defined structure, it follows a consistent order and it is designed in such a way that it can be **easily accessed** and used by a person or a computer. Structured data is usually stored in well-defined columns and also Databases.
- **Semi-Structured Data** can be considered as another form of Structured Data. It inherits a few properties of Structured Data, but the major part of this kind of data fails to have a definite structure and also, it does not obey the formal structure of data models such as an RDBMS.
- **Unstructured Data** is completely a different type of which neither has a structure nor obeys to follow the formal structural rules of data models. It does not even have a consistent format and it found to be varying all the time. But rarely it may have information related to data and time.

Source: Edureka

Types of Big Data

- A strong data scientist needs to be versatile and comfortable with dealing a variety of types of data, including:
  - Traditional: numerical, categorical, or binary
  - Text: emails, tweets, New York Times articles
  - Records: user-level data, timestamped event data, json-formatted log files
  - Geo-based location data: (e.g. NYC housing data)
  - Network
  - Sensor data
  - Images

Big Data Versus Big Computation

- Full scans (e.g., log processing)
- Range scans
- Point lookups
- Iterations
- Joins (self, binary, or multiway)
- Proximity queries
- Closures and graph traversals

Big Data Applications

- Web search
- Marketing and advertising
- Data cleaning
- Knowledge base
- Information retrieval
- Internet of Things (IoT)
- Visualization
- Behavioral studies

Publicly Available Datasets

- Data.gov
- Institutions: UCR Star (https://star.cs.ucr.edu)
- Google Dataset Search
  - https://toolbox.google.com/datasetsearch
- Twitter Streaming API
- Yahoo! Webscope (http://webscope.sandbox.yahoo.com/)
- Yahoo Stocks
- GDELT (http://www.gdeltproject.org/)
- Other APIs: Google, Facebook, Instagram, Yelp, Reddit, Census API

Big Data Architectures

- **A big data architecture is designed to handle the ingestion, processing, and analysis of data that is too large or complex for traditional database systems**
- Most big data architectures include some or all of the following components

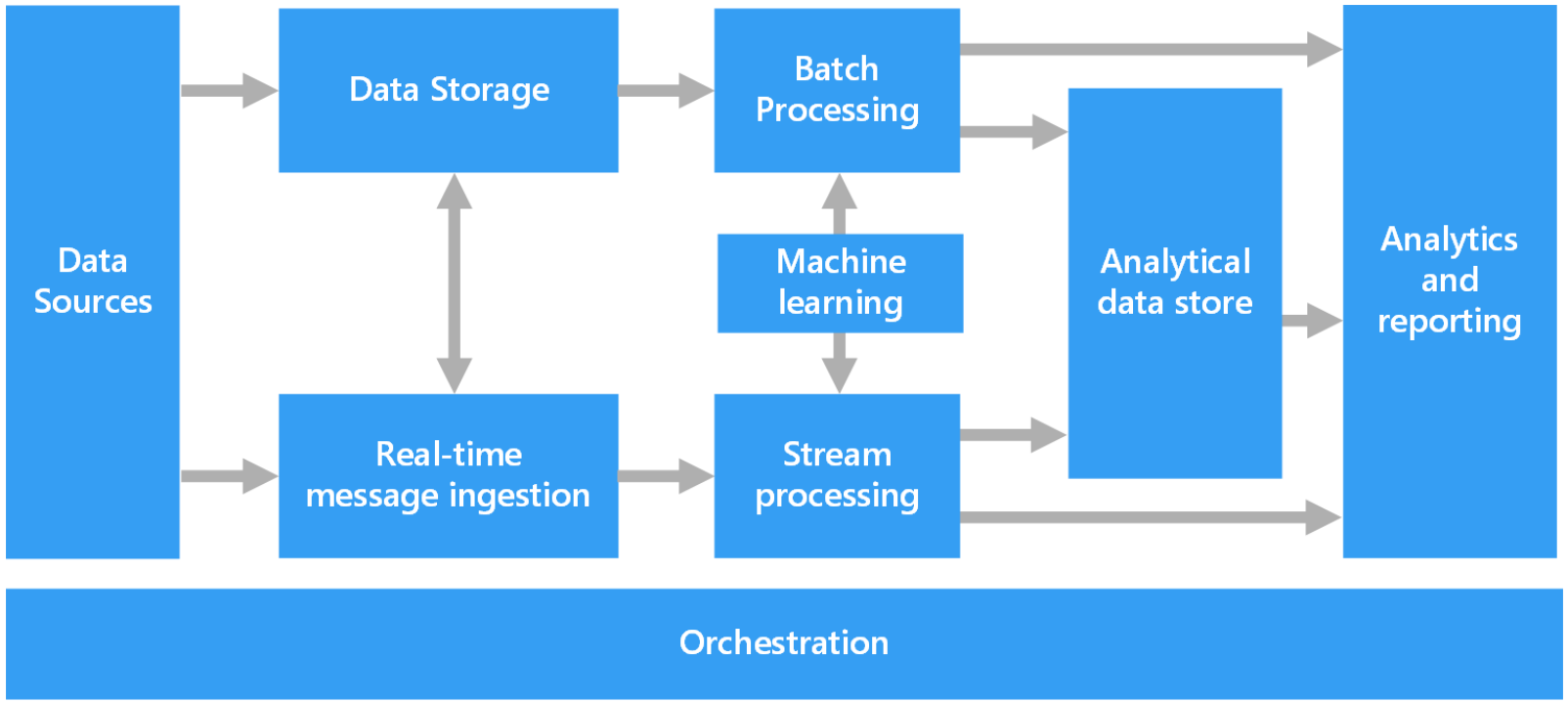


Image source: https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/

Big Data Architectures (Cont'd)

- **Non-relational data structures**
  - Large non-relational databases like Hadoop
    - In Hadoop's Distributed File System (HDFS), data is stored as 'key and data-value' combinations
  - Google BigFile (now Google File System)
  - NoSQL is emerging as a popular language to access and manage non-relational databases
  - There is a matching Data Warehousing system called Hive along with its own PigSQL language
- **Massively parallel computing**: Given the size of data, it is useful to divide and conquer the problem quickly using multiple processors simultaneously
- **Unstructured Information Management Architecture (UIMA)**. This is one of elements in the "secret sauce" behind IBM' Watson's system that reads massive amounts of data, and organizes for just-in-time processing

Top 5 Big Data Architectures

- **Streaming** – Allows ingestion (and possibly analytics) of mission-critical, real-time data that can come at you in manic spurts.
- **General (or specific) purpose 'batch' cluster** – Provides generalized storage and compute capabilities in an extensible, cost-effective cluster which may perform any and all of the functions of the other four architectures.
- **NoSQL engines** – Gives architects the ability to handle the "Three V's" -- high velocity, high volume, or the high variety/variability of the underlying data.
- **Enterprise data warehouse (EDW)** – Lets an organization maintain a separate database for years of historical data and run various long-running analytics on that data.
- **In-place analytics** – Allows users to leave their data "in place" in a low-cost storage engine and run performant, ad-hoc queries against that data without creation of a separate, expensive "cluster."

https://www.globalknowledge.com/

Big Data Tools

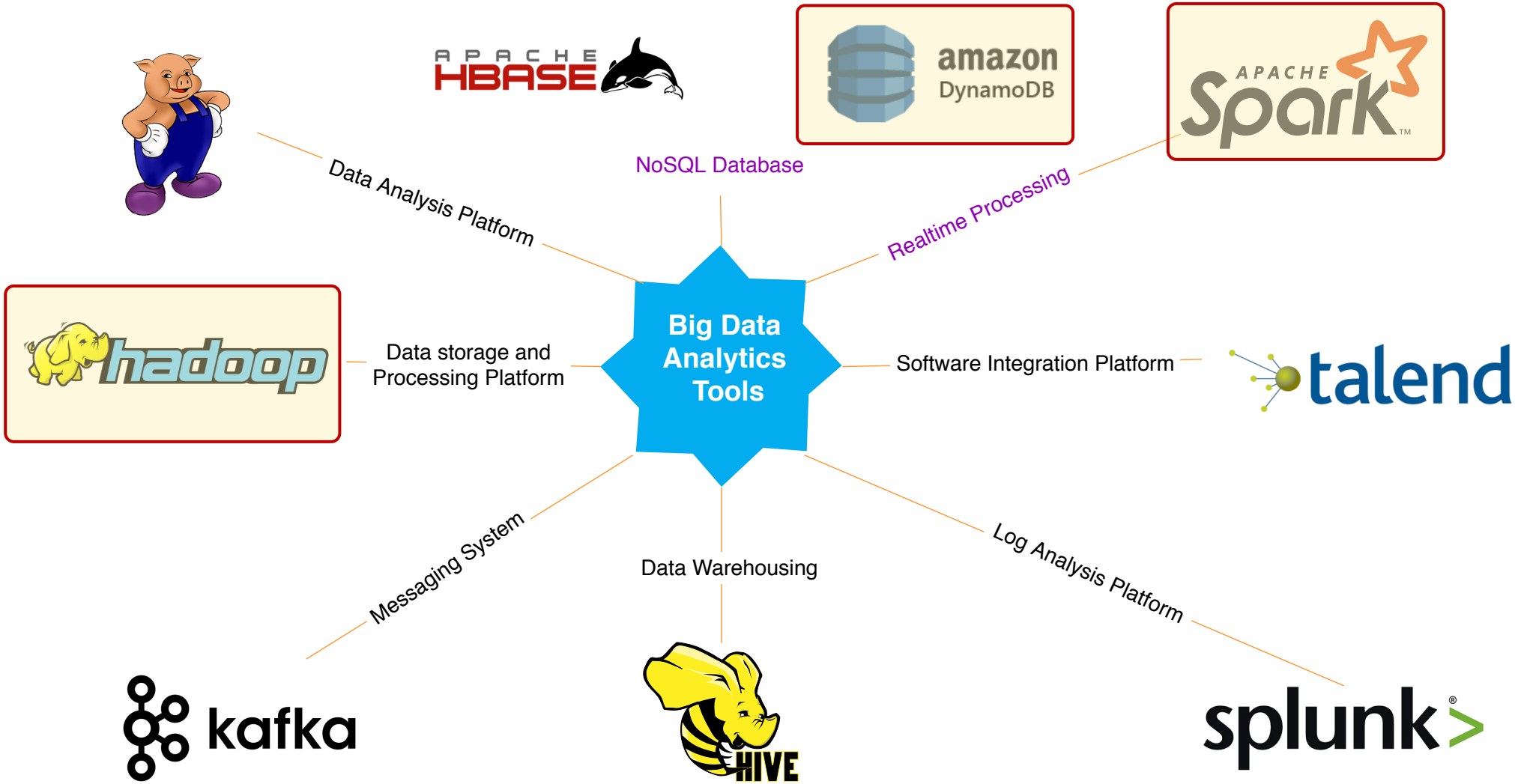


Image Recreated from: https://www.houseofbots.com/

Big Data Tools – AWS-Centered

Area	Open Source	AWS Services
Batch Ingest	Apache Flume StreamSets Apache Sqoop	AWS Glue AWS Data Pipeline AWS Internet of Things (IoT) services
Stream Ingest	Apache Flume StreamSets	Amazon Kinesis Data Firehose
Persistent Storage	Hadoop Distributed File System (HDFS) Relational database management system (RDBMS)	Amazon Simple Storage Service (Amazon S3) Amazon EMR Amazon Relational Database Service (Amazon RDS)
Transient Storage	Apache Kafka	Amazon Kinesis
Batch Processing	Apache Hive Apache Flink Apache Spark MapReduce PostgreSQL	Amazon Redshift AWS Glue AWS Data Pipeline Amazon RDS
Stream Processing	Apache Flink Apache Spark Apache Beam	Amazon Kinesis Data Analytics
Clients or Data Applications	Apache Superset (BI)	Amazon QuickSight
Visualize	SAS Tableau TIBCO tools	Amazon QuickSight Amazon Elasticsearch Service (Amazon ES) and Kibana AWS IoT Analytics Amazon Kinesis Data Analytics Amazon Elastic Compute Cloud (Amazon EC2) Jupyter notebooks

1.2 BIG DATA COMPONENTS

Storage of Big Data

- Data is growing faster than Moore's Law
- Too much data to fit on a single machine
- Partitioning
- Replication
- Fault-tolerance

Hadoop Distributed File System (HDFS)

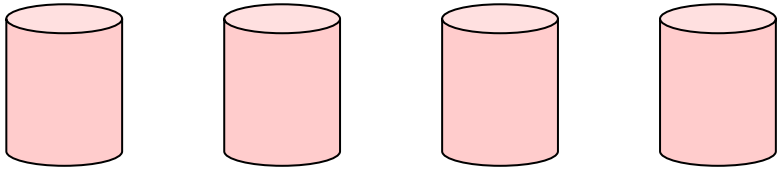
- The most widely used distributed file system
- Fixed-sized partitioning
- 3-way replication
- Write-once read-many
- See also: GFS, Amazon S3, Azure Blob Store

Indexing

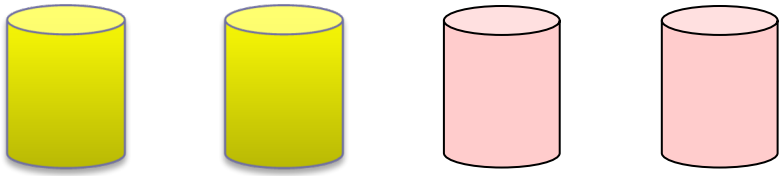
- Data-aware organization
- Global Index partitions the records into blocks
- Local Indexes organize the records in a partition
- Challenges:
- Big volume
- HDFS limitation
- New programming paradigms
- Ad-hoc indexes

Fault Tolerance

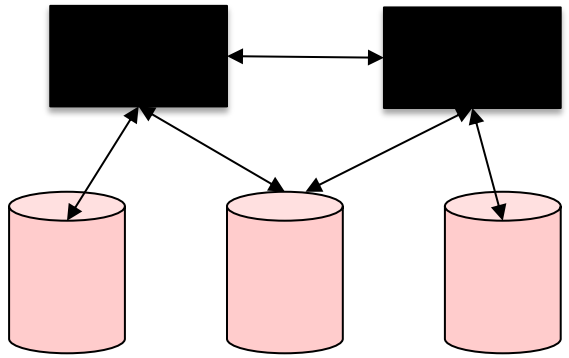
- Replication



- Redundancy

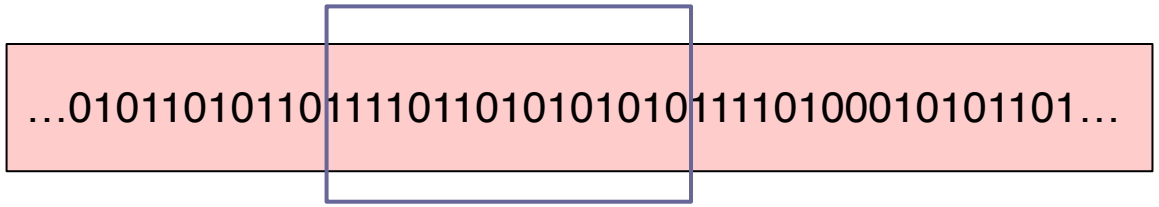


- Multiple masters



Streaming

- Sub-second latency for queries
- One scan over the data
- (Partial) preprocessing
- Continuous queries
- Eviction strategies
- In-memory indexes



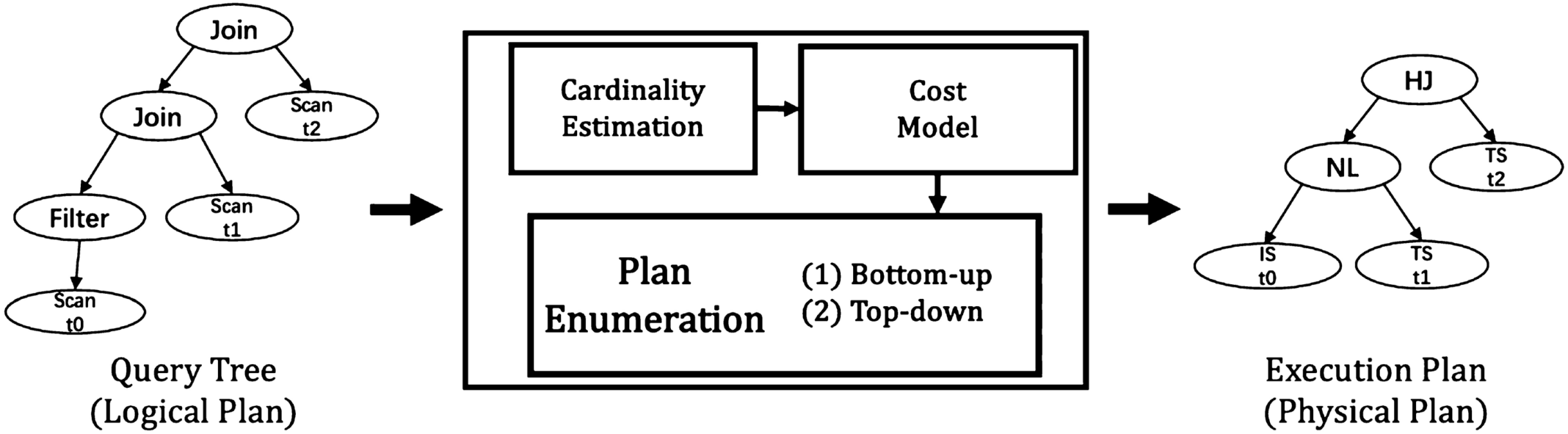
Processing Window

Task Execution

- MapReduce
  - Map-Shuffle- Reduce (Example: Hadoop)
- Resilient Distributed Datasets (RDD)
  - Directed-Acyclic-Graph (DAG) – (example: Dask)
  - In-memory processing
  - Resiliency through lineages (Example: Spark)
- Hyracks
- Stragglers (Example: AsterixDB)
- Load balance

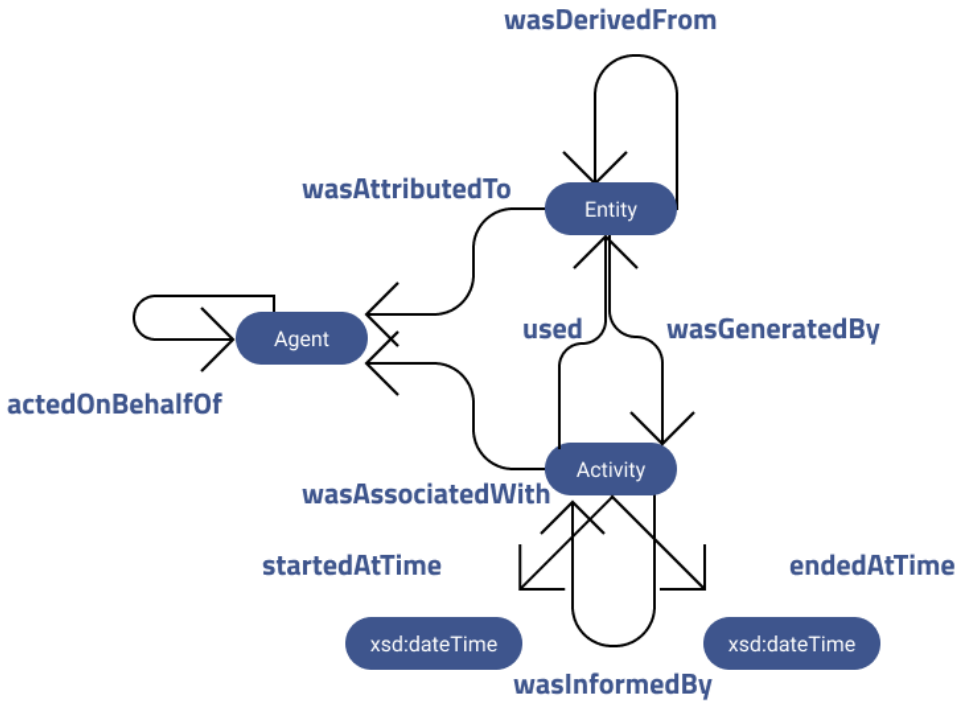
Query Optimization

- Finding the most efficient query plan
  - e.g., grouped aggregation
- Cost model (CPU – Disk – Network)



Provenance

- Debugging in distributed systems is challenging and sometimes painful
- Keeping track of transformations on each record
- Data provenance refers to the description of the origin, creation and propagation process of data.
- Data provenance is the lineage and derivation of the data. It stores ownership and process history about data objects.



Big Graphs

- Motivated by social networks
- Billions of nodes and trillions of edges
- Tens of thousands of insertions per second
- Complex queries with graph traversals

1.3 PROCESSING DATA

Related Fields

Data mining - to extract meaningful insights from data

Business intelligence - covers data analysis that relies heavily on aggregation, focusing mainly on business inform

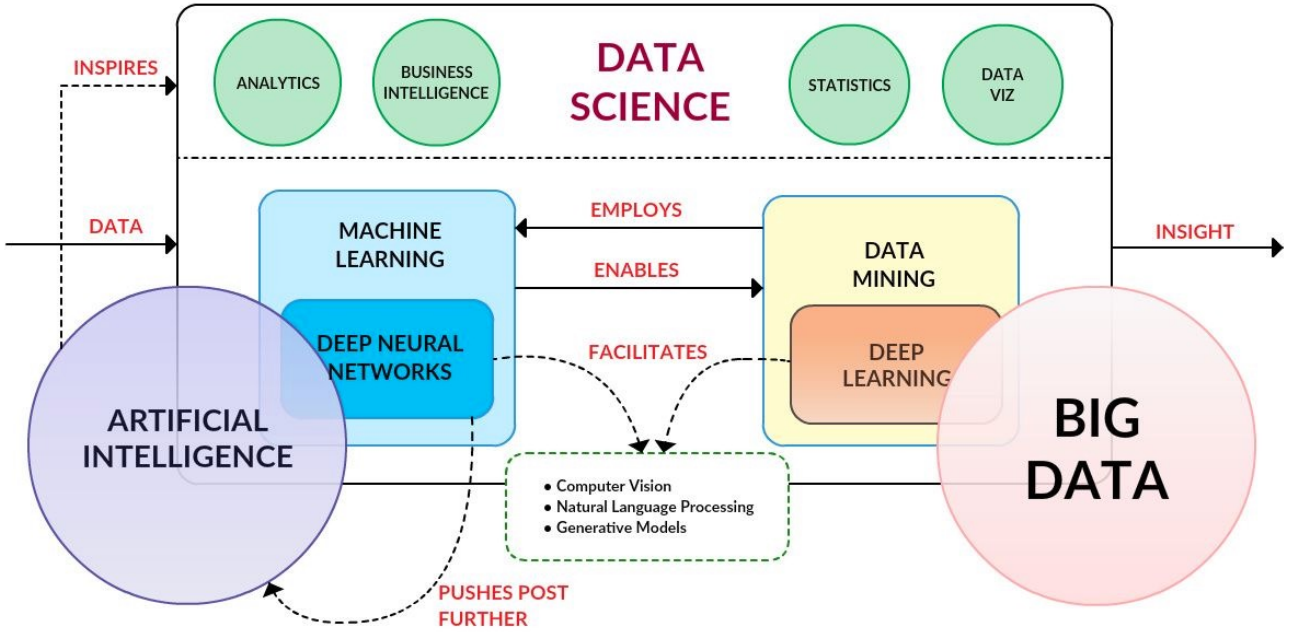
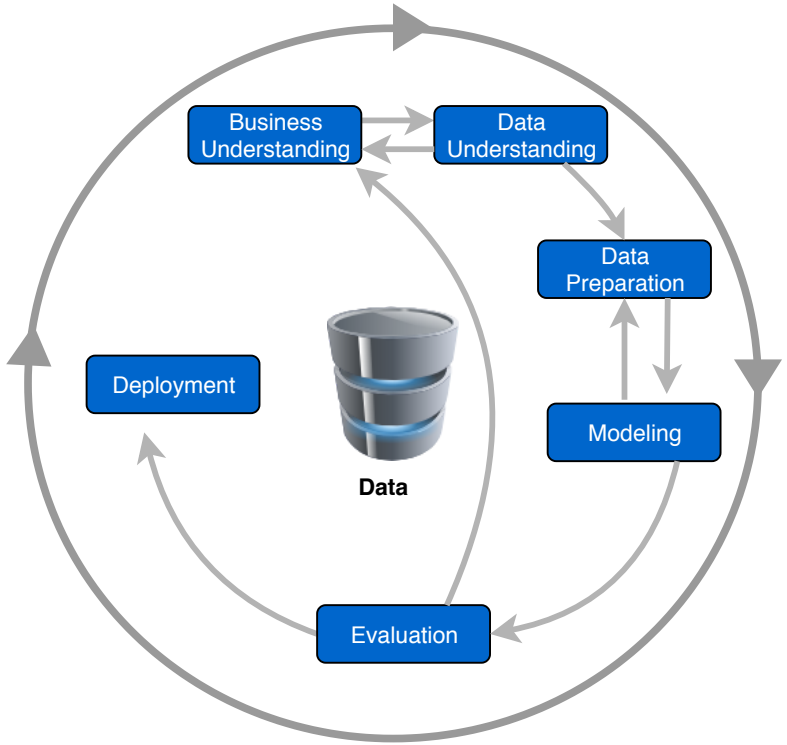


Image: https://imgur.com/

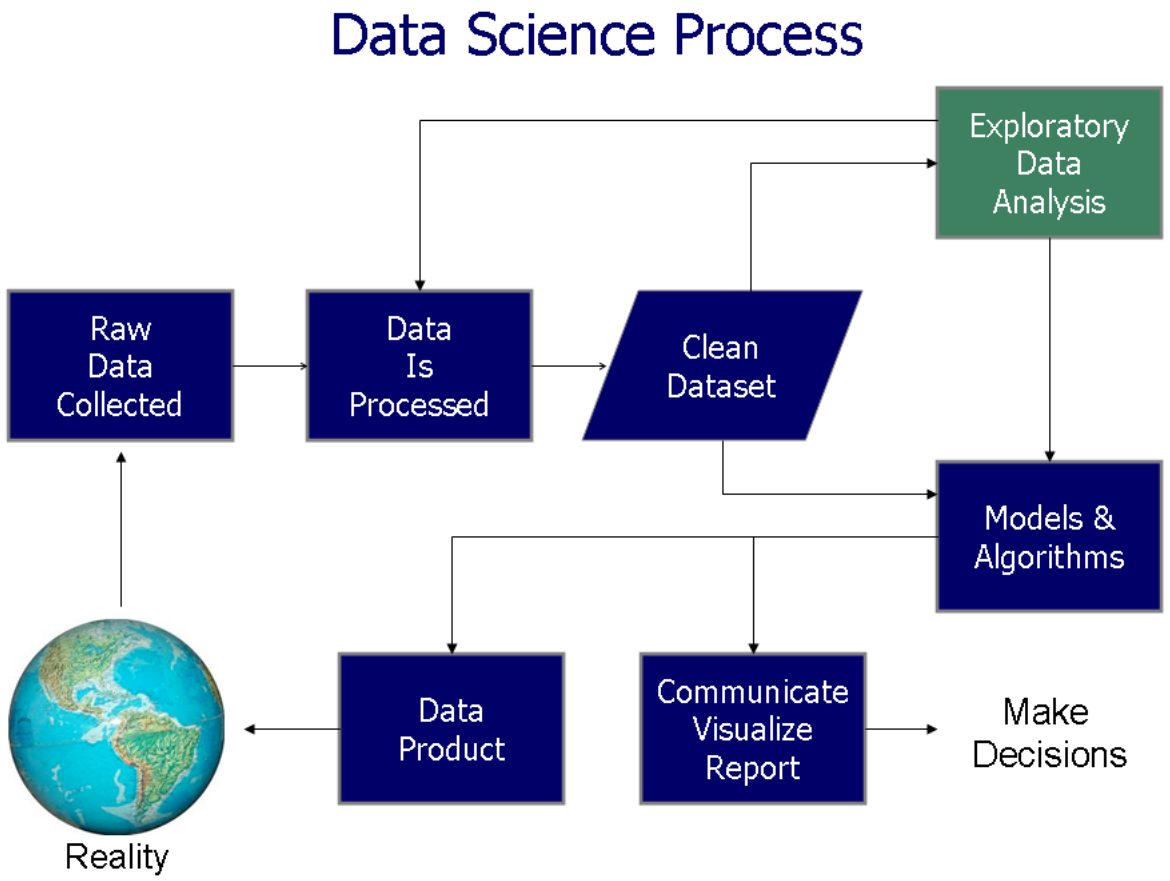
Related Fields

Data mining

CRISP-DM – Cross Industry Standard Process for Data Mining

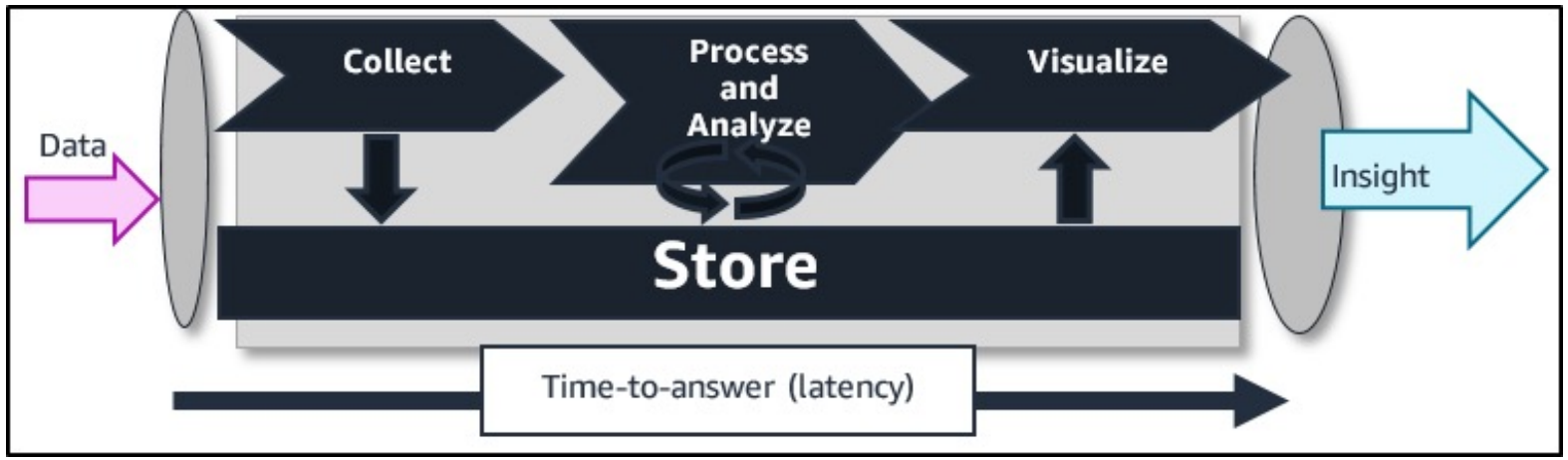


Doing Data Science

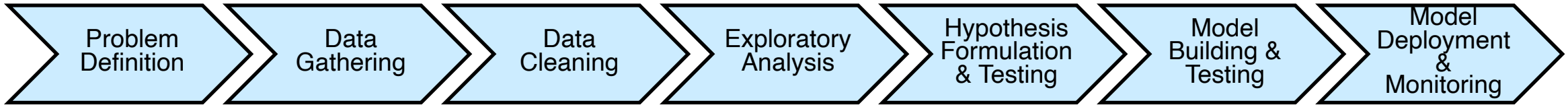


Doing Data Science, by Schutt & O'Neil (2013)

Big Data Pipeline

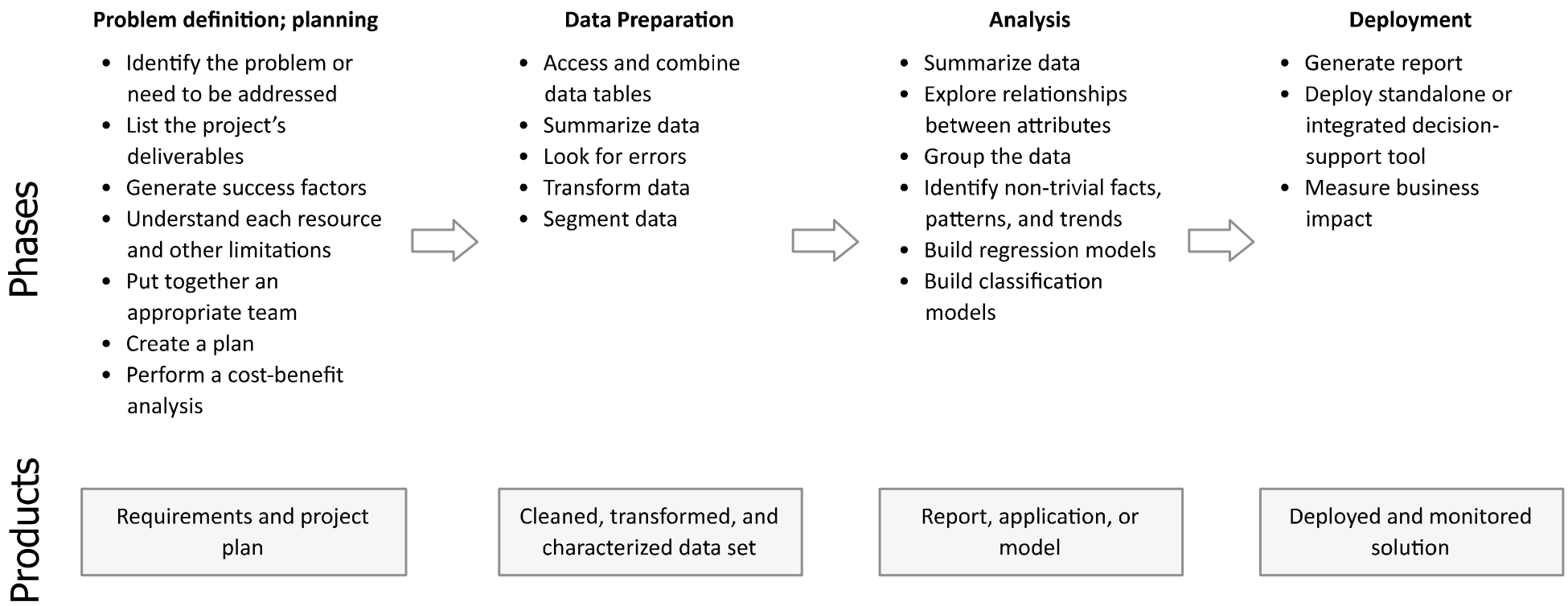


Data Processing Workflow



Data Analysis/Processing

- Steps to consider in developing a data analysis or data mining project.

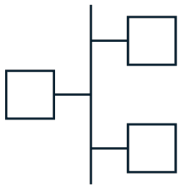


Making Sense of Data, Glenn Myatt et al.

Data Processing

Five steps can turn process data into smart data.

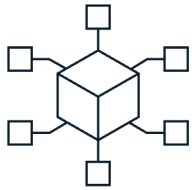
1



Define the process

Label process steps using plant schematics or engineering drawings, identify critical sensing and their limits

2



Enrich the data

Remove nonstandard operating regimes, address sensor calibrations, and build a high-quality dataset

3

Reduce the dimensionality

Leverage engineering formulas to intelligently combine sensor data

4

Apply machine learning

Overlay advanced-analytics models on engineered data to capture stochastic variability

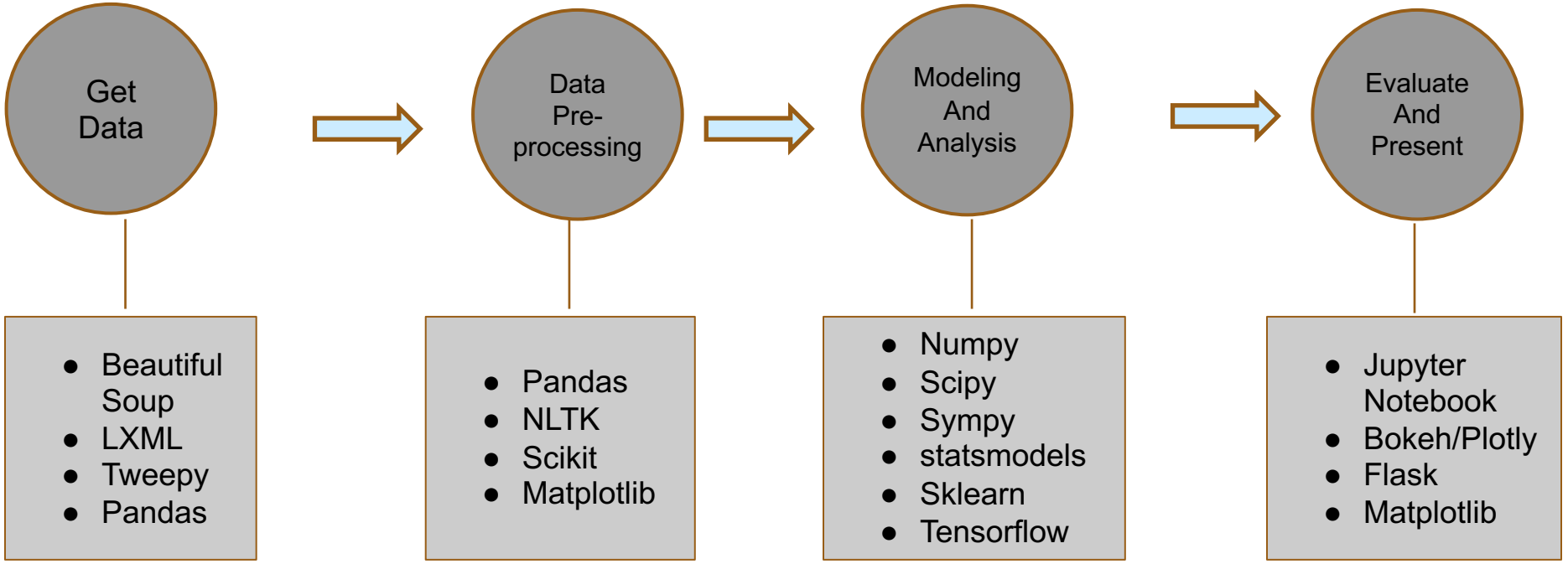
5

Implement and validate the models

Check for causality, ensure the results are physical, and push for insights to be implemented

McKinsey & Company

Data Processing with Python



Slide courtesy of Steven Skiena, Stony Brook University

The End