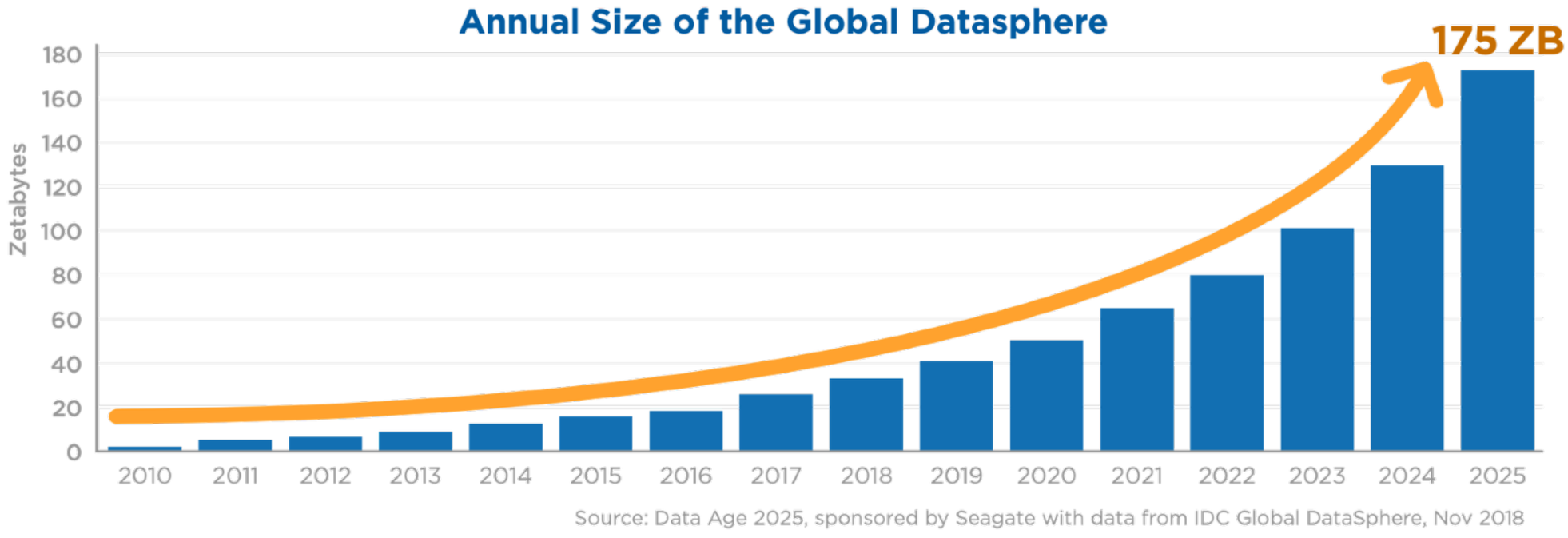
Big Data Overview

CptS 415

Why Big Data?



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

5 Vs of Big Data

Volume

The amount of data: Horrendously Large

Velocity

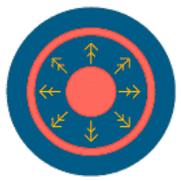
The speed of data entering a solution:

- Real-time or near real-time
- Think of Facebook, Youtube, …

Variety

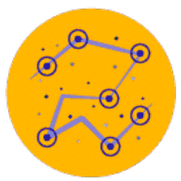
Different Data Sources and Different Types:

- Structured, Unstructured, semi-structured



Veracity

The degree to which the data is accurate, precise, and trusted.



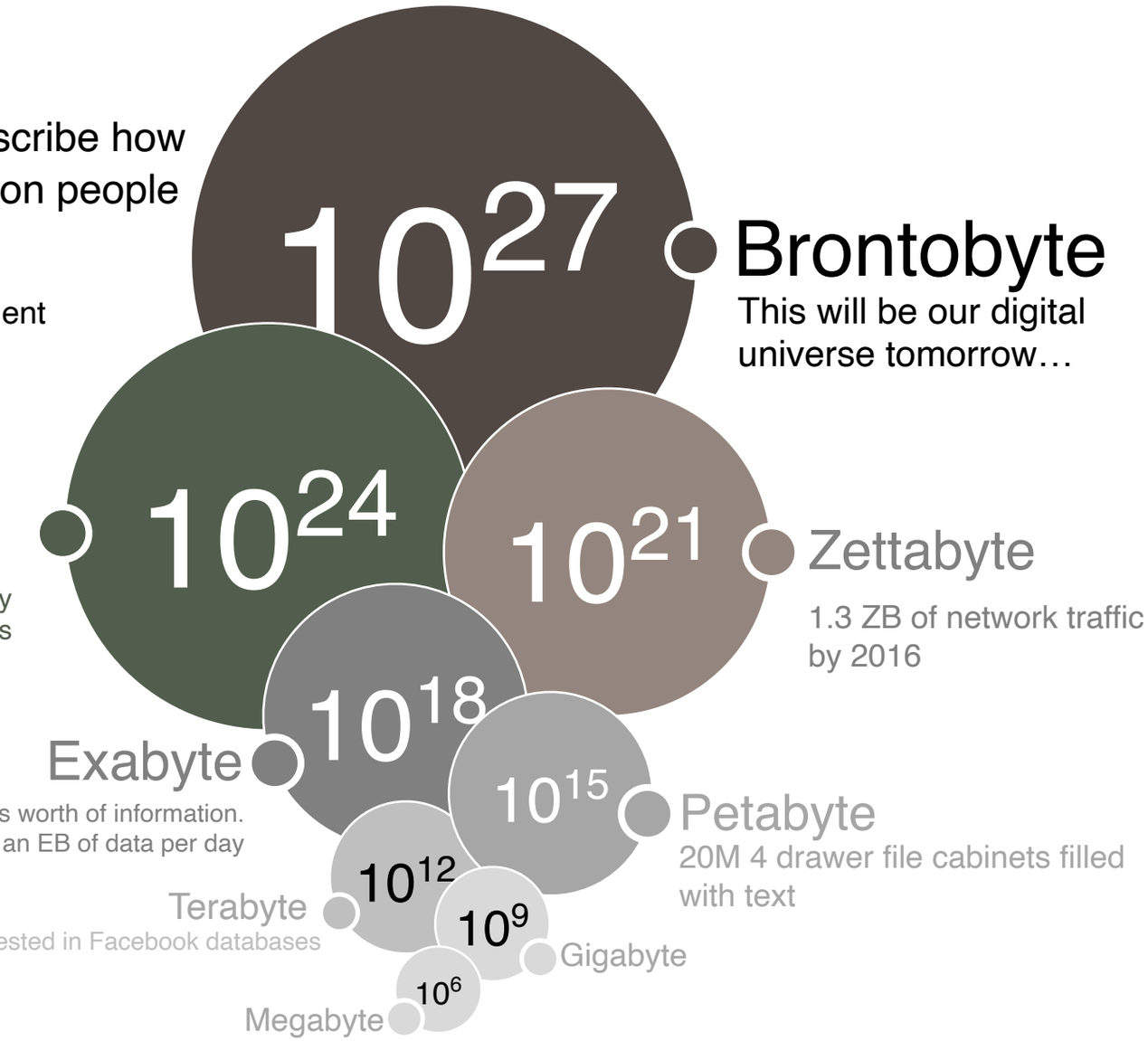
Value

The ability of a solution to extract meaningful information from the data.

5 Vs - Volume

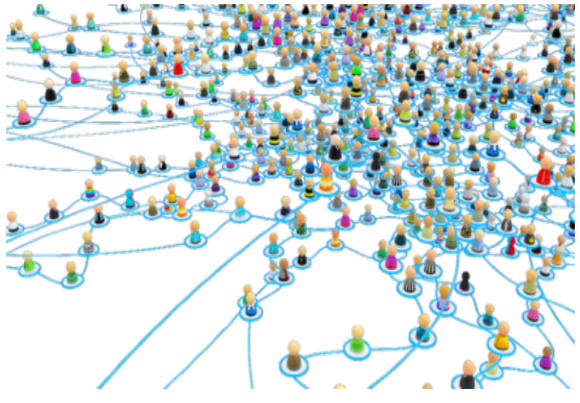
Today data scientist uses Yottabytes to describe how much government data the NSA or FBI have on people altogether.

In the near future, Brontobyte will be the measurement to describe the type of sensor data that will be generated from the IoT (Internet of Things)

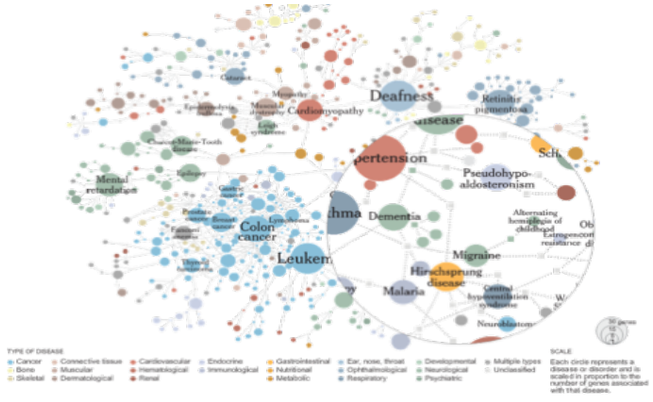


Reference: Defensibly downsizing your data, Hewlett Packard Enterprise, 2016, https://slideplayer.com/slide/12657871/

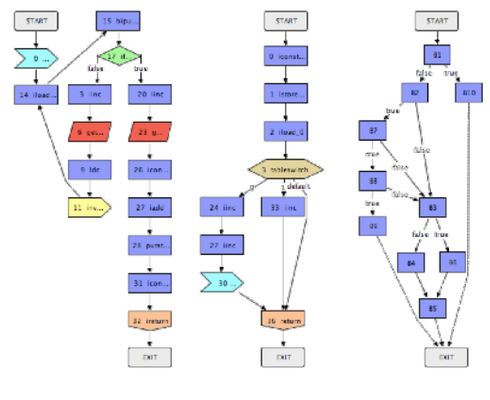
No Data is an Island



Social Network



Knowledge Graph



Control Flow

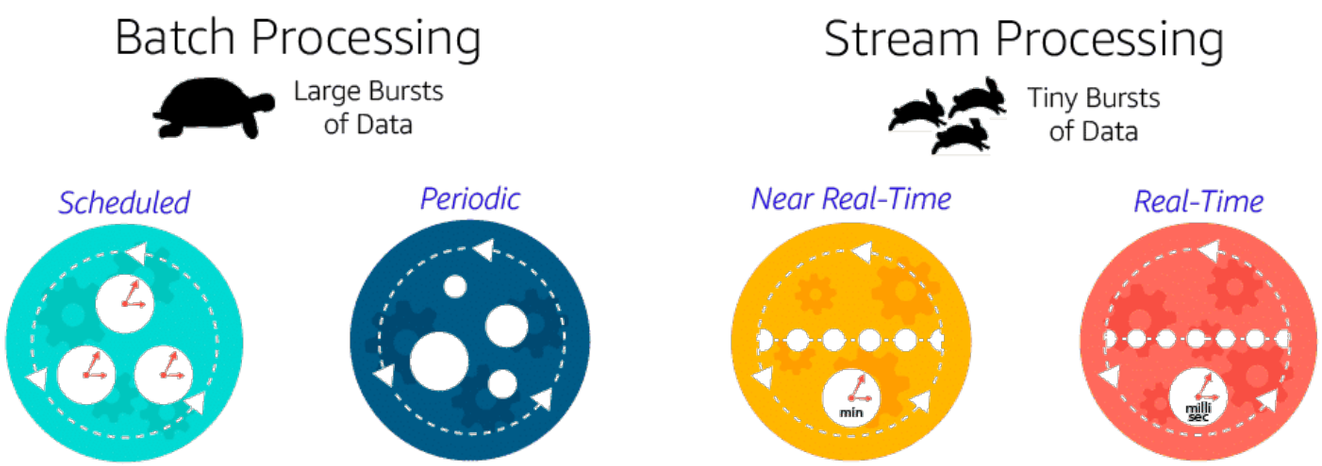
Internet of Things

Brain Network

Challenge

Find a needle in a Haystack

Velocity



Scheduled

- Velocity is very predictable
- Large bursts of data transfer at scheduled intervals

Periodic

- Velocity is less predictable
- The loss of scheduled events can strain the system

Near Real-Time

- Velocity is a huge concern
- Data need to be processed within minutes

Real-Time

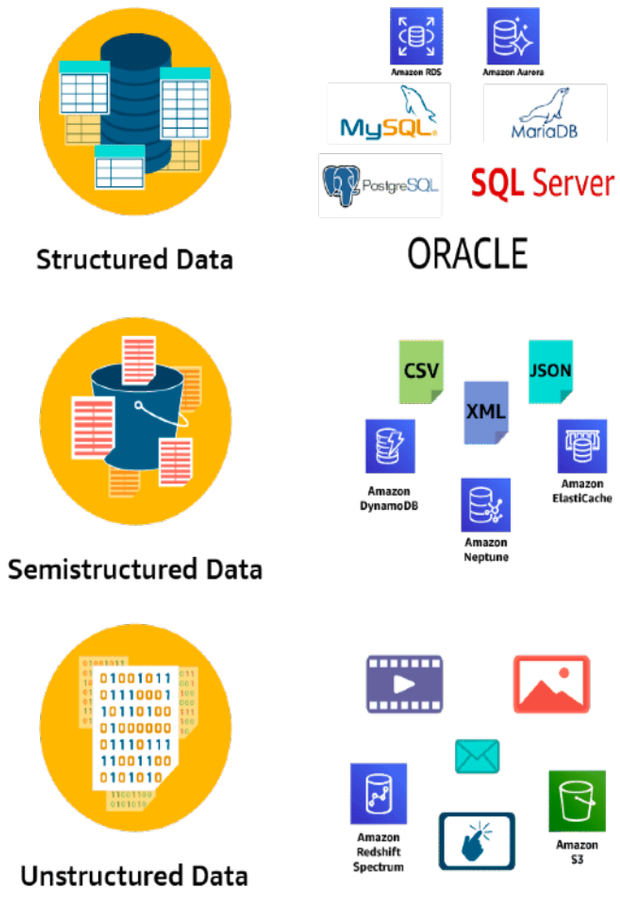
- Velocity is a paramount concern
- Data need to be processed in seconds or sub-seconds

Challenge

Drinking from a Firehose

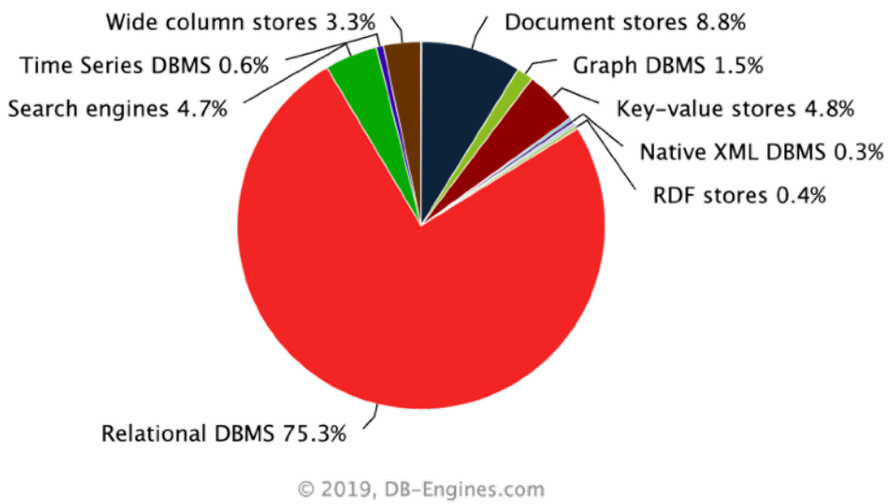
Variety (Complexity)

- There are just as many types of data as there are people…
- Not only data types,
  - numbers, texts, videos, JSON files, tables, etc.
- Data Source Types:
  - Structured: tabular format.
    - database management systems (DBMS).
  - Semi-structured: in the form of elements within a file
    - JSON, XML, CSV, etc.
  - Unstructured: in the form of a file
    - pictures, videos, audios, email content, etc.

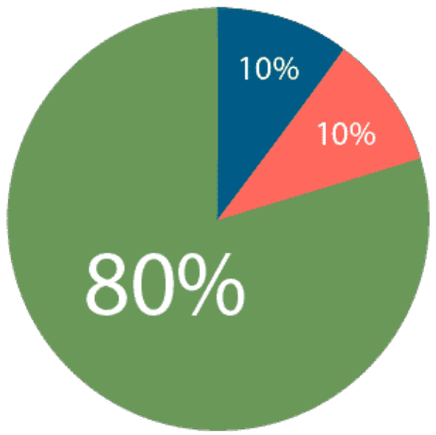


Data Sources

Popularity of DBMS (Jul. 2019)



Percentage of different data sources

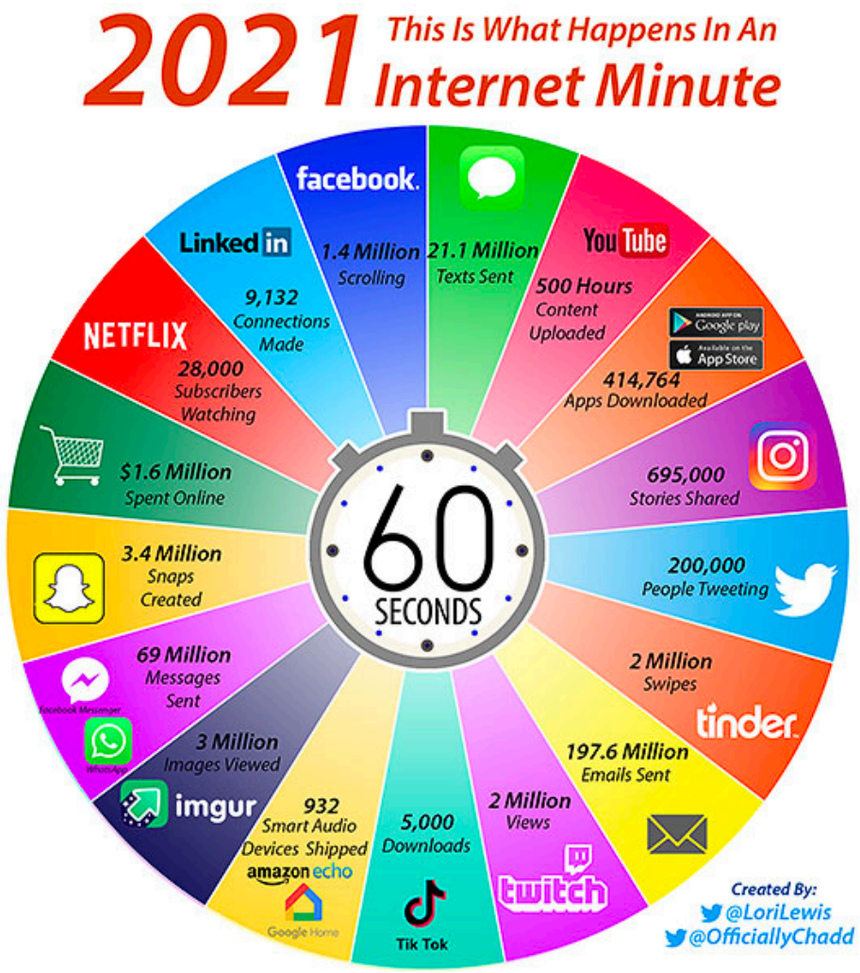


Structured Data - 10% - Tabular Data

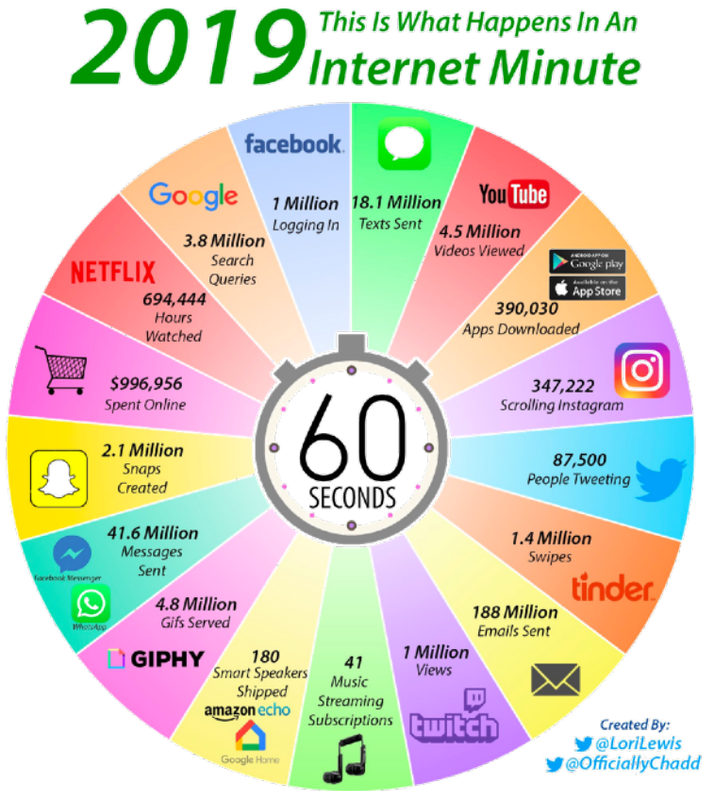
Semistructured Data - 10% - CSV, XML, JSON Files

Unstructured Data - 80% - Everything Else

Internet Data in 60 Seconds



source: https://www.allaccess.com/merge/archive/32972/infographic-what-happens-in-an-internet-minute

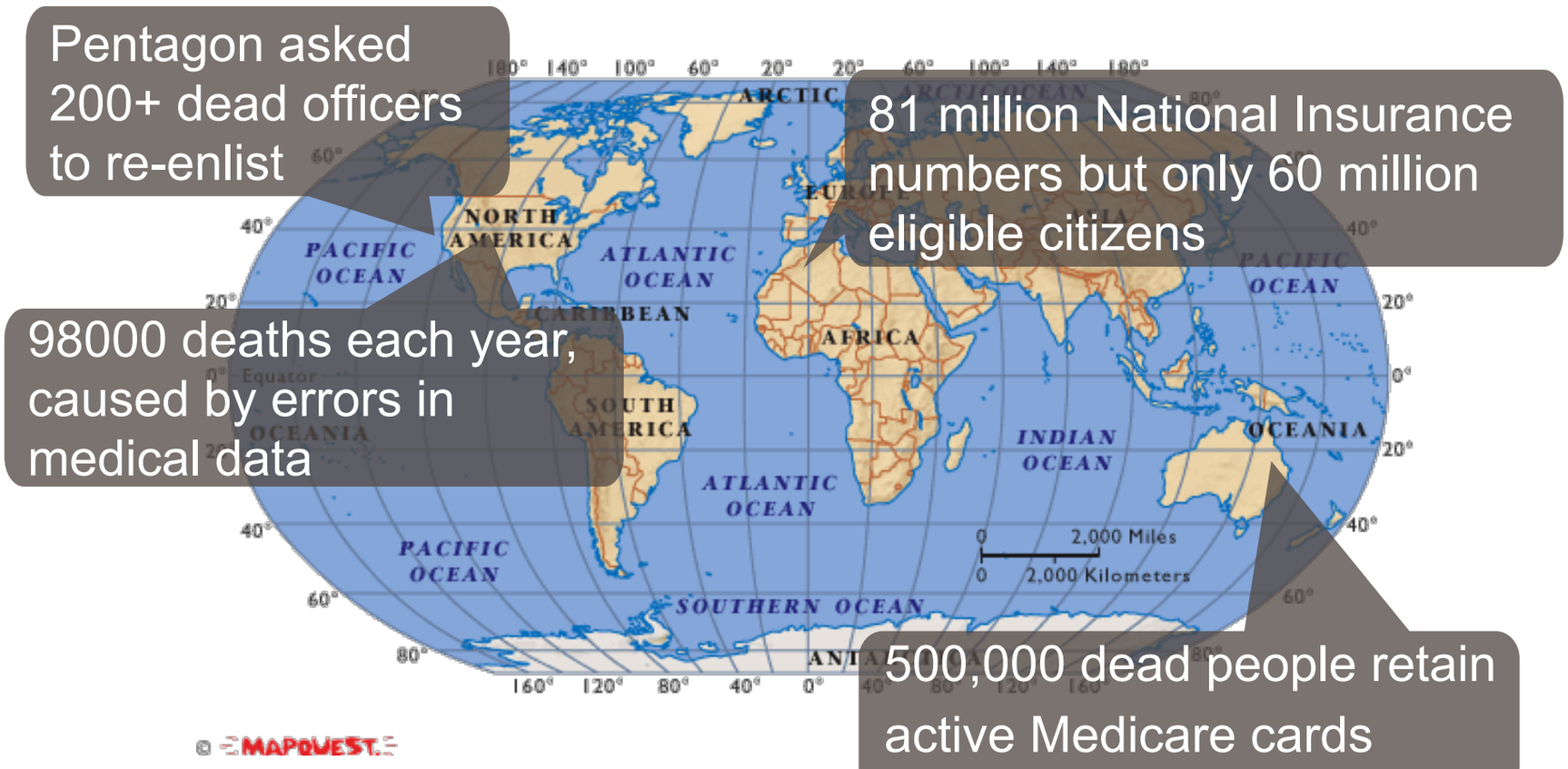


Reference: https://www.visualcapitalist.com/what-happens-in-an-internet-minute-in-2019/

Veracity (Quality & Trust)

- Data = Quantity + Quality
- Can we trust the answers to our queries?
- Dirty data routinely lead to
  - Misleading financial reports
  - Strategic business planning decisions
  - Loss of revenue
  - Loss of credibility and customers
  - Disastrous consequences

Data in Real-Life is Often Dirty



Dirty Data is Costly

- Poor data cost U.S. business $611 billion annually
- Erroneously priced data in retail databases cost US customers $2.5 billion each year
- 1/3 of system development projects were forced to delay or cancel due to poor data quality
- 30%-80% of the development time and budget for data warehousing are for data cleaning

State of Data Quality

On average, U.S. companies believe

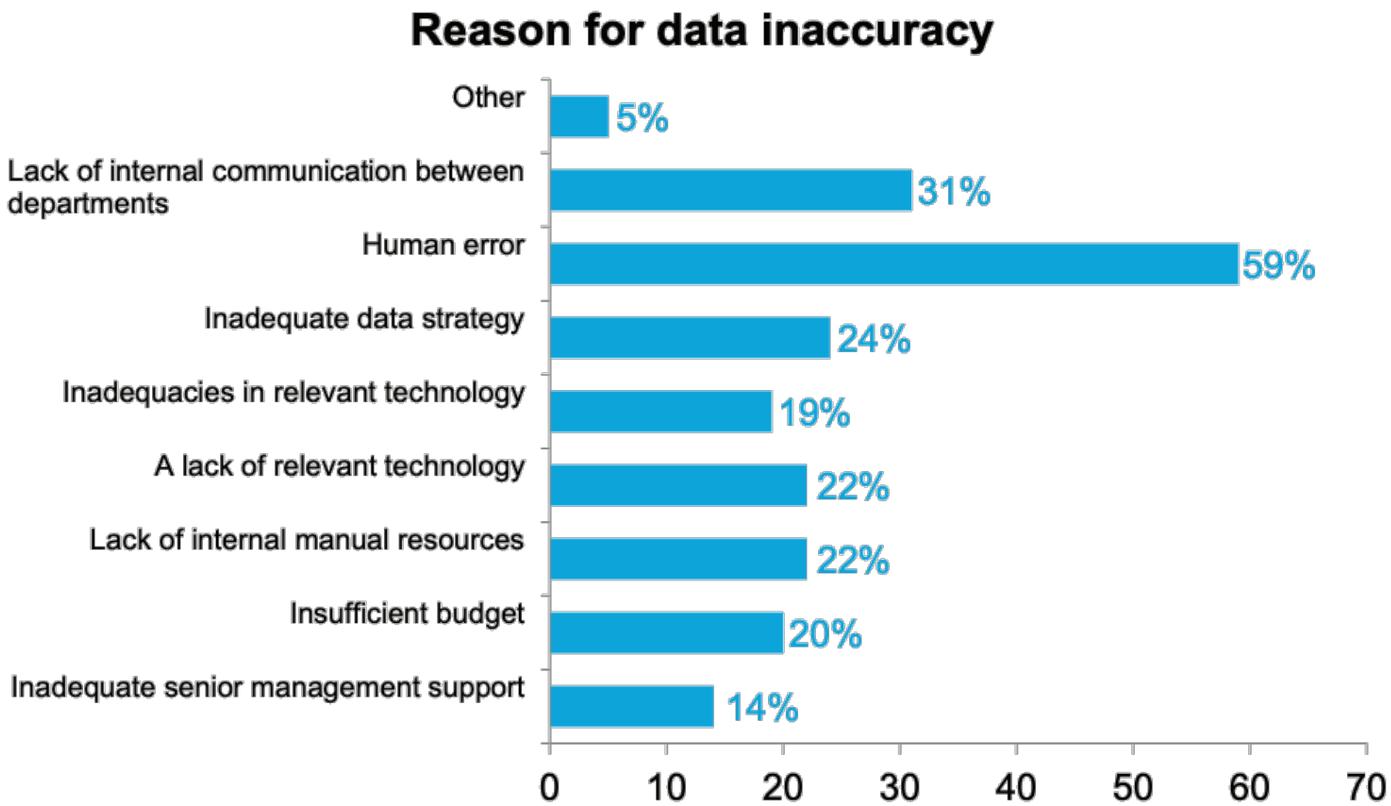
25 percent of their data

is inaccurate.

Common data errors plague

91 percent

of organizations.

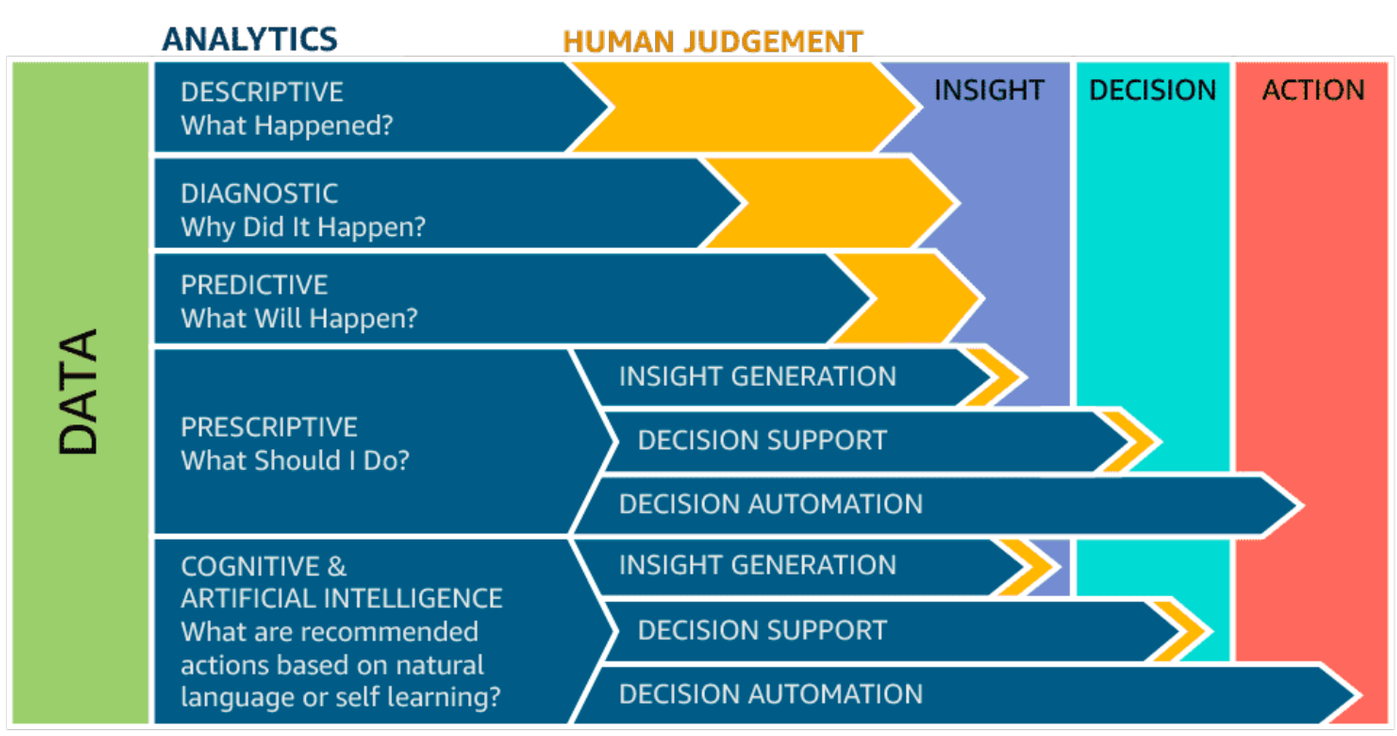


Reference: The state of data quality, an Experian data quality white paper. https://www.experian.com/assets/decision-analytics/white-papers/the%20state%20of%20data%20quality.pdf

Value

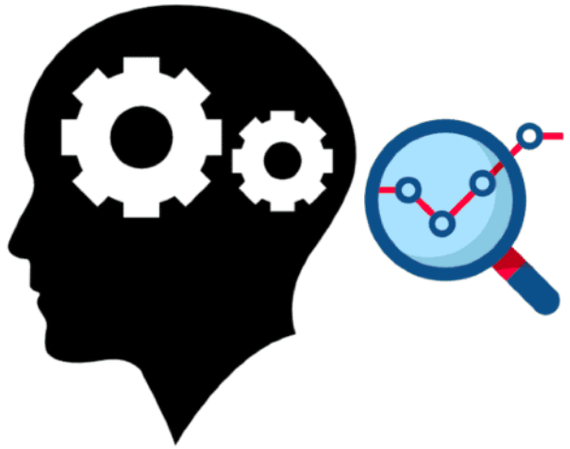
"We believe **data is our oil**, our gold. But having hundreds of millions of terabytes of **data that isn't actionable** really does nothing for me."

- Rob Roy, Chief Digital Officer at Sprint



Case Study: Starbucks

Starbucks uses behavioural analytics to cater to its customers



Starbucks gather a lot of info about their customers' coffee-buying habits from their preferred drinks to what time of day they're usually ordering

The company directs exciting offers and coupons to their customers and ensures to maintain their interest

An Example Scenario

- 15 JSON data files, each about 2.5 GB in size.
- They are placed on a file server once an hour.
- Must be Ingested as soon as they arrive.
- Combine with transactions from the financial dashboard for this same period
- Compare to the recommendations from the marketing engine
- All data is fully cleansed before arrival
- The results must be made available to decision makers within 10 minutes

Volume

Velocity

Variety

Veracity

Value

Big Data Topics

CptS 415

Topic I: Data Models, Storage, Management

- Relational Data Models and DBMS
  - Relational data and relation algebra
  - DBMS: centralized; single processor
  - Relational Databases

- Challenge: How to store and represent Big Data?

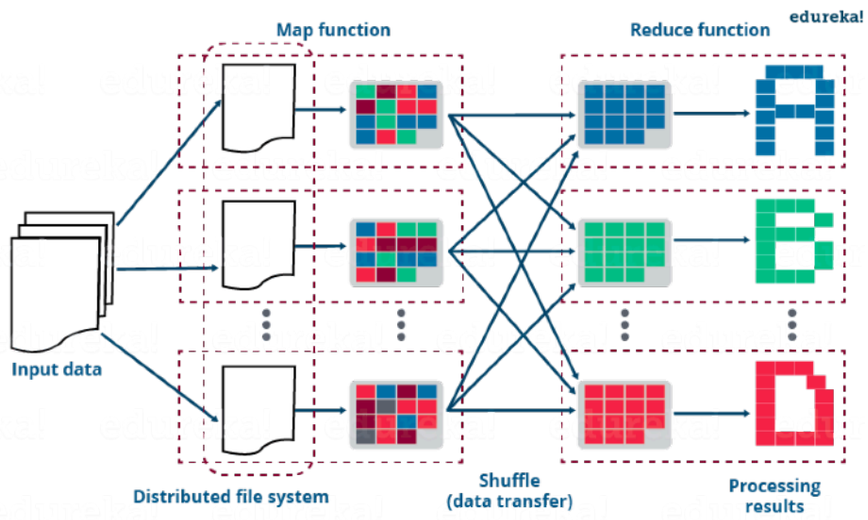
- Beyond Relational Databases
  - Non-relational data, semi-structured data
  - noSQLs, newSQLs, Key-value stores, Column stores, Document stores, …
  - Graph data, Graph Databases

Topic II: Search/Query Big Data

- Popular query languages
  - SQL Fundamentals
  - XML, Xquery and SPARQL
- Challenge: How to find needle in the Big Data haystack?
- Search algorithms: design principles and case study
  - Indexing and Views
  - Exact Vs. Approximate search
  - Compression and summarization
  - Resource bounded search
  - Cope with data streams

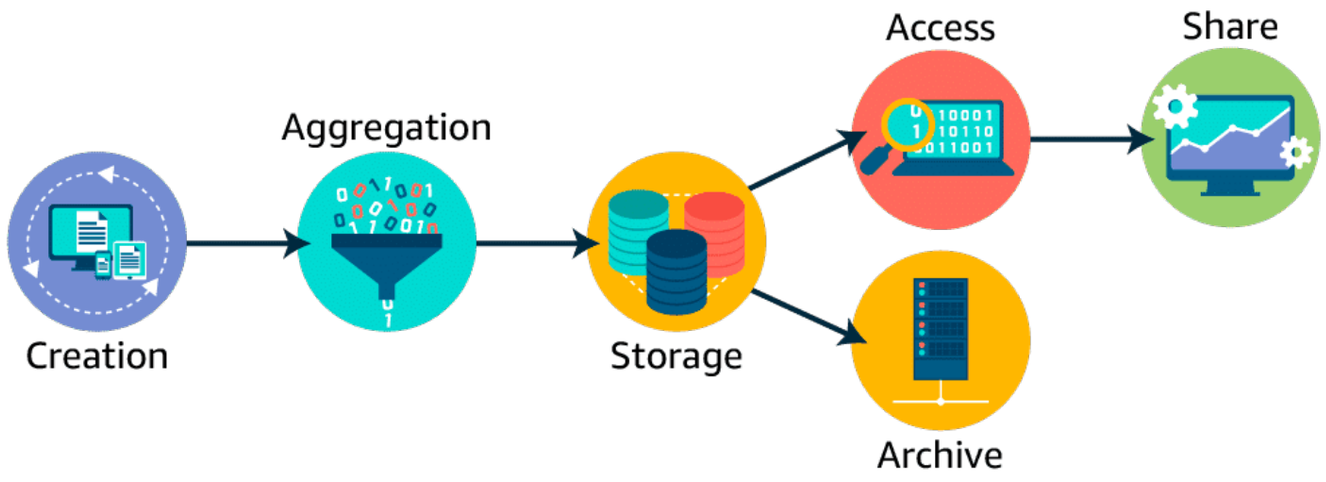
Topic III: Parallel/Distributed Systems

- Recall traditional DBMS:
  - Database: “single” memory, disk
  - DBMS: centralized; single processor
- Question: Can we do better provided with multiple processors/ servers?
- Parallel DBMS: exploring parallelism
  - Improve performance
  - Reliability and availability



Topic IV: Data Quality, Security and Ethics

- Dealing with Data Veracity:
  - Error Detection
  - Data Cleansing
  - Data Integrity
- Data Privacy and Security

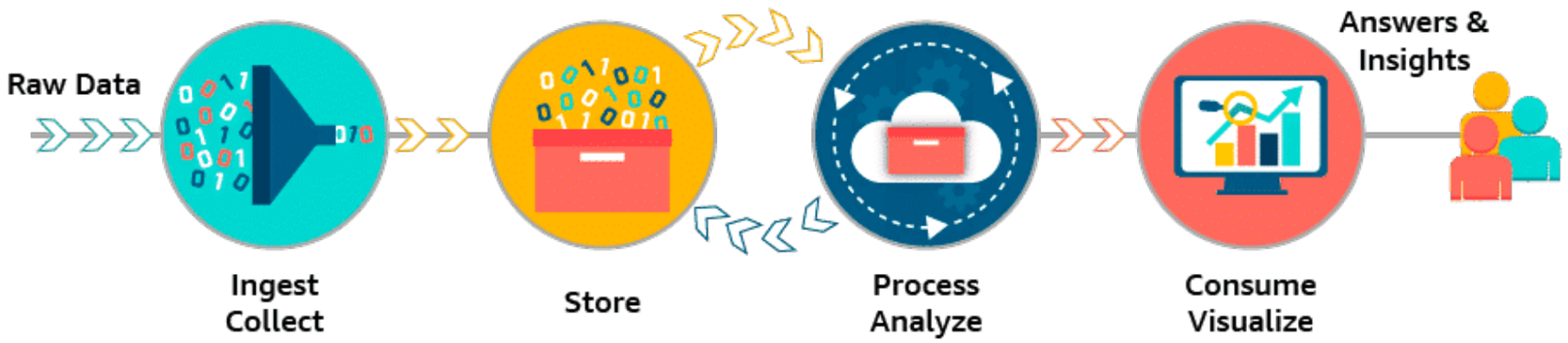


Topic V: Data Analytics and Visualization

- Data Analytics
  - Descriptive
  - Diagnostic
  - Predictive
  - Prescriptive
  - Cognitive and Artificial Intelligence

- Report and Visualization

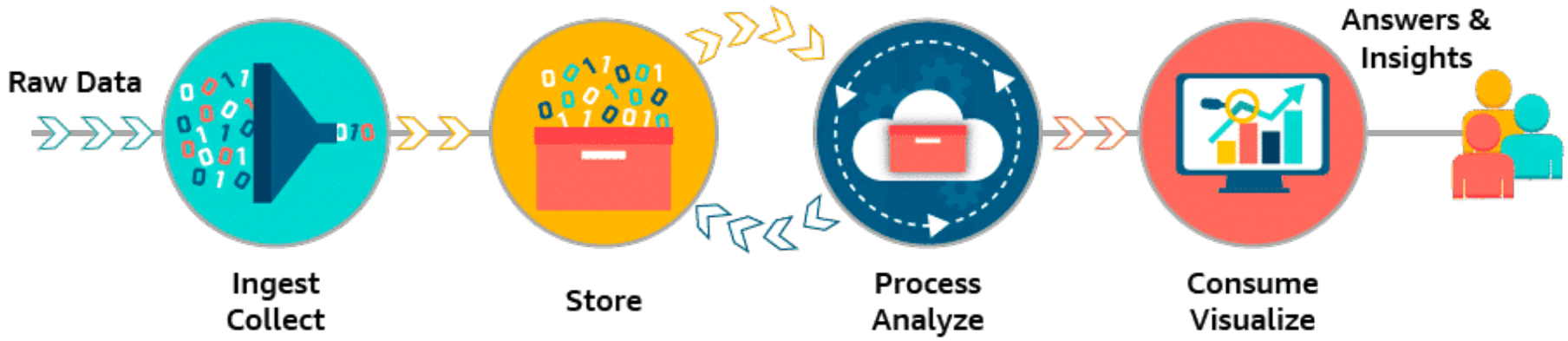
Typical Big Data Solution



- Ingest/Collect
  - Collect raw data from various sources.
  - A good data analysis solution ingest a wide variety of data.
- Store
  - Secure, Scalable and Durable
  - Structured, Unstructured and Semi-structured
- Process/Analysis
  - Transform the data to make it consumable
  - Sorting, aggregating, joining, or applying business logics
  - Store the result
- Consume/Visualize
  - Query
  - Business Intelligence (BI) Tools

CptS 415: Putting Things Together

Topic IV: Data Quality, Security and Ethics



Topic I: Data Models, Storage, Management (CptS 451)

Topic II: Search/Query Big Data (CptS 315, 570, 575)

Topic III: Distributed/Parallel Data Analysis (CptS 411)

Topic V: Data Analytics, Visualization (215)

Skillset for Data Analytics

Basic Programming

Statistical and quantitative Analysis

Data Warehousing

SQL

NoSQL

Data Visualization

Specific Business Knowledge

Computational Frameworks