COMP9313: Big Data Management

Introduction to Big Data Management

What is big data?

Dan Ariely

January 6, 2013 at 6:17pm ·

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

Like Comment Share

Tweeted by Prof. Dan Ariely, Duke University

What is big data?

- No standard definition!
- Wikipedia:
  - Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.
- Amazon:
  - Big data can be described in terms of data management challenges that – due to increasing volume, velocity and variety of data – cannot be solved with traditional databases.

What is big data?

Word could which is generated from the top-20 results when search “what is big data” in Google.

What is big data?

- A set of data
- Special characteristics
  - Volume
  - Variety
  - Velocity
  - …
- Traditional methods cannot manage
  - Store
  - Analyse
  - Retrieve
  - Visualization
  - …

That’s why we need this course

Big Data Definitions Have Evolved Rapidly

- 3 V's
  - In a research report by Doug Laney in 2001
  - Volume, Velocity and Variety
- 4 V's
  - In Hadoop – big data tutorial, 2006
  - Veracity
- 5 V's
  - Around 2014
  - Value
- 7 V's, 8 V's, 10 V's, 17 V's, 42 V's, …

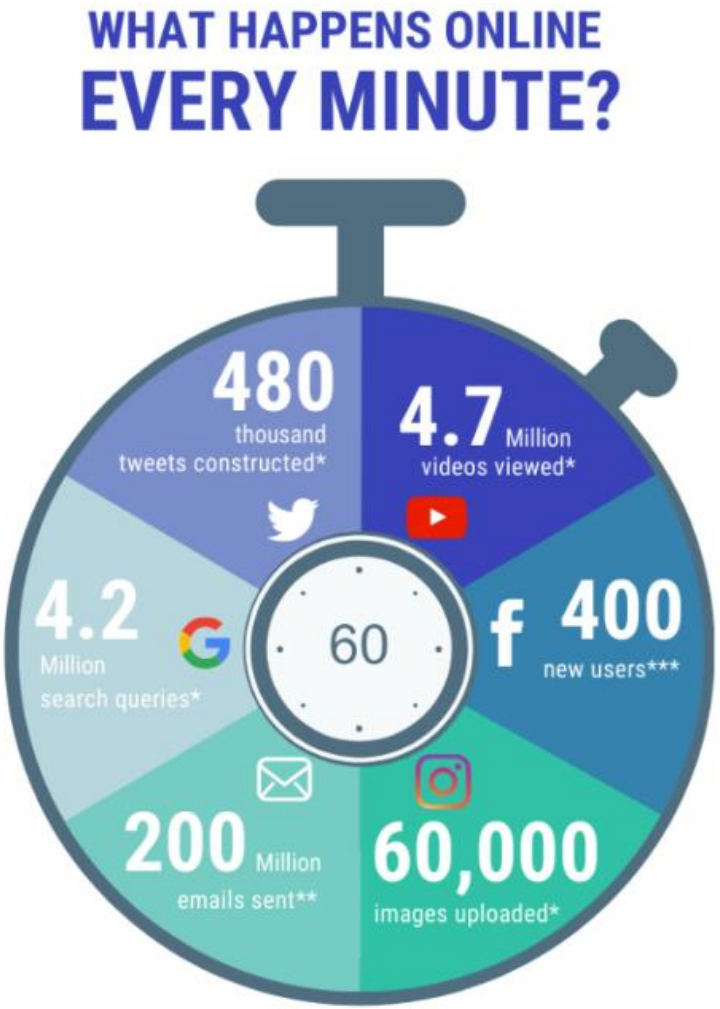
Major Characteristics of Big Data

Volume (Scale)

- Quantity of data being created from all sources
- The fundamental of big data

- 18 Zetabytes (ZB) of data in 2018, will grow to 175 ZB in 2025
  - 1 zettabyte $\approx 10^3$ exabytes $\approx 10^9$ terabytes
  - Source: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

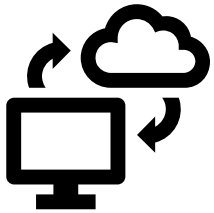
Volume



Source: https://www.nodegraph.se/how-much-data-is-on-the-internet/

Volume – Why Challenging?

Model	RAM	Disk	Data
Macintosh Classic (1990)	1MB – 4MB	0 – 40MB	
Power Mac G4 (2000)	256MB – 1.5GB	20GB – 60GB	5 EB in 2003
iMac (mid 2010)	4GB – 16GB	500GB – 2TB	1 ZB in 2012
iMac (early 2019)	8GB – 64GB	1TB – 3TB	~40 ZB



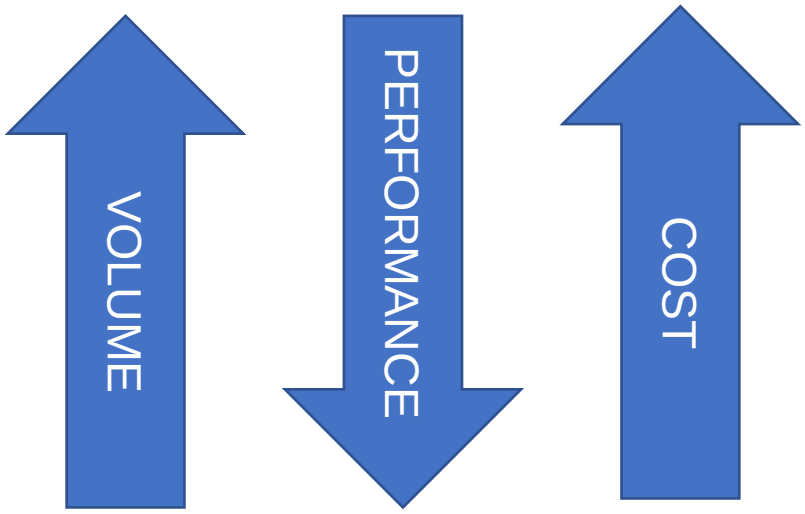
?

1990s 2000s 2010s future

DBMS Storage

Volume – Why challenging?

- Time complexity
  - Sort algorithms: O(N logN)
  - Merge join: O(N logN + M logM)
  - Shortest path: O(V logV + E logV)
  - Nearest neighbor search: O(dN)
  - NP hard problems



Variety (Diversity)

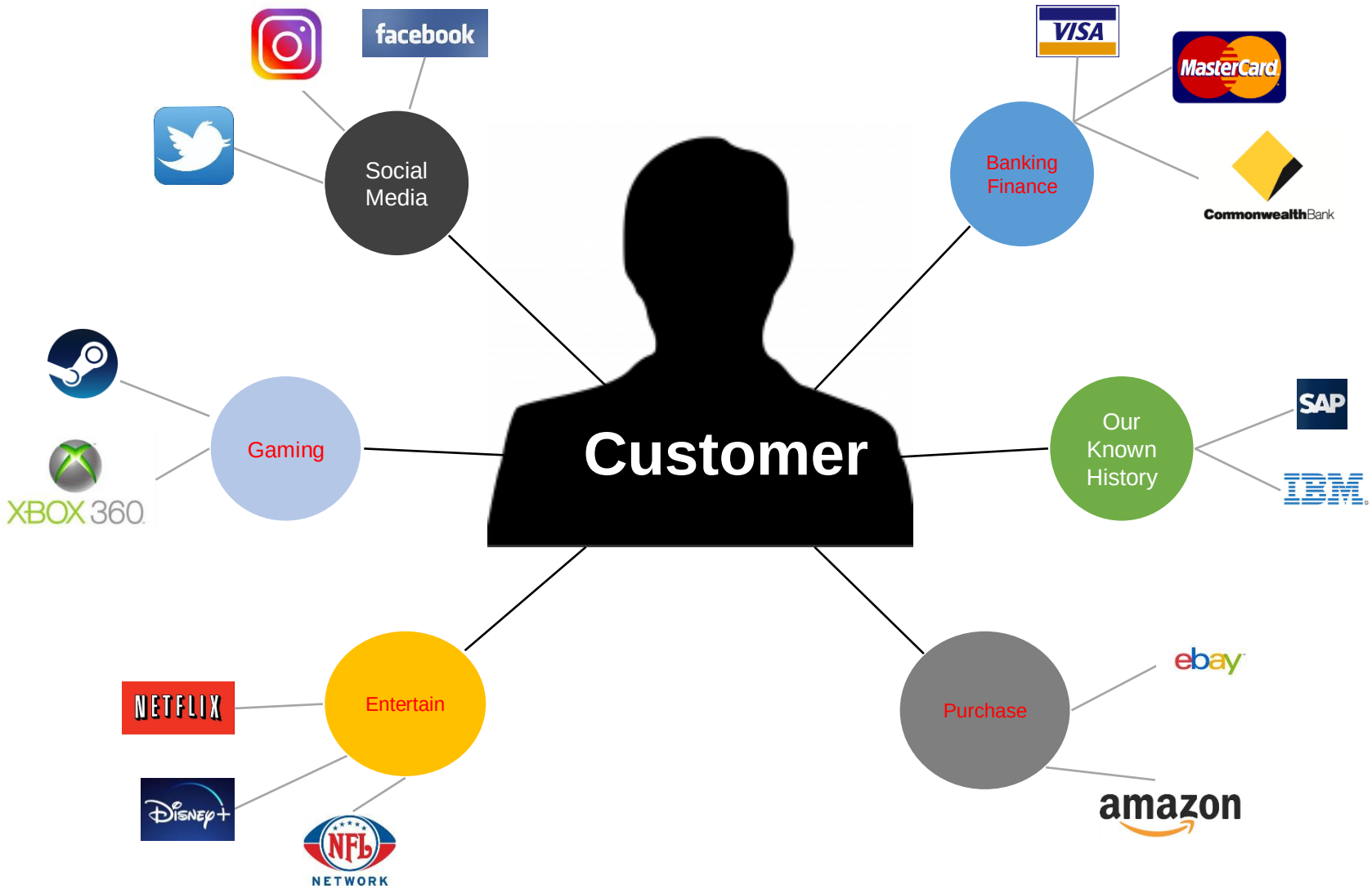
- Different Types
  - Relational data (tables/transactions)
  - Text data (books, reports)
  - Semi-structured data (JSON, XML)
  - Graph data (social network, RDF)
  - Image/video data (Instagram, Youtube)
- Different sources
  - Movie reviews from IMBD and Rotten Tomatoes
  - Product reviews from different provider websites
  - Personal information from different social apps

Variety

- A single application can be generating or collecting multiple types of data
  - Email
  - Webpage

- If we want to extract knowledge, then all the data with different types and sources need to be linked together

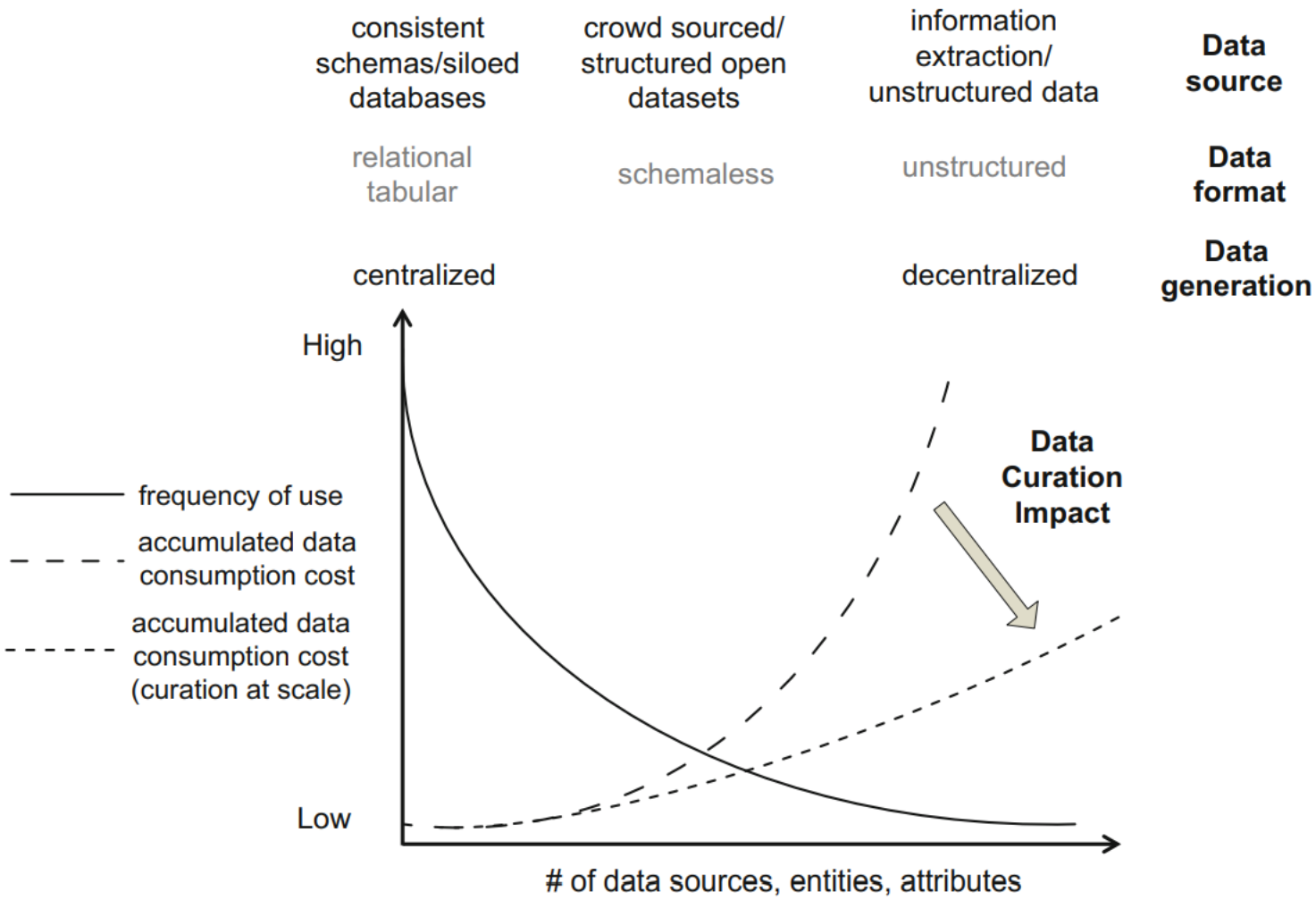
Variety - A Single View to the Customer



Variety – Why Challenging?

- Data integration
  - Heterogeneous
    - Traditional data integration relies on **schema mapping**, the difficulty and time complexity is directed related to the level of heterogenity and data sources
  - Record linkage in variety data
    - needs to identify if two records refer to the same entity. How to make use of different types of data/information from different sources?
- Data curation
  - Organization and integration of data collected from various sources
  - Long tail of data variety

The Long Tail of Data Variety and Data Curation



Source: Curry, E., & Freitas, A. (2014). Coping with the long tail of data variety.

Velocity (Speed)

- Data is being generated fast, thus need to be
  - stored fast
  - processed fast
  - analysed fast

- Every second
  - **8,991** Tweets sent
  - **994** Instagram photos uploaded
  - **4,683** Skype calls
  - **93,508** GB of Internet traffic
  - **83,165** Google searches
  - **2,915,385** Emails sent

Source: http://www.internetlivestats.com/one-second/

Velocity

- Reason of growth
  - Users:
    - 16 million in 1995 to 3.4 billion in 2016
  - IoT:
    - sensor devices, surveillance cameras
  - Cloud computing:
    - $26.4 billion in 2012 to $260.5 billion in 2020
  - Website:
    - 156 million in 2008 to 1.5 billion in 2019
  - Scientific data:
    - weather data, seismic data

Velocity

- Data is now streaming into the server in real time, in a continuous fashion and the result is only useful if the delay is very short.

- Many application need immediate response
  - Fraud detection
  - Healthcare monitoring
  - Walmart's real-time alerting

Velocity – Why Challenging?

- Batch processing

- Real time processing

- Transmission
  - Transferring data becomes a prominent issue in big data
  - Balancing latency/bandwidth and cost
  - Reliability of data transmission

Veracity (Quality)

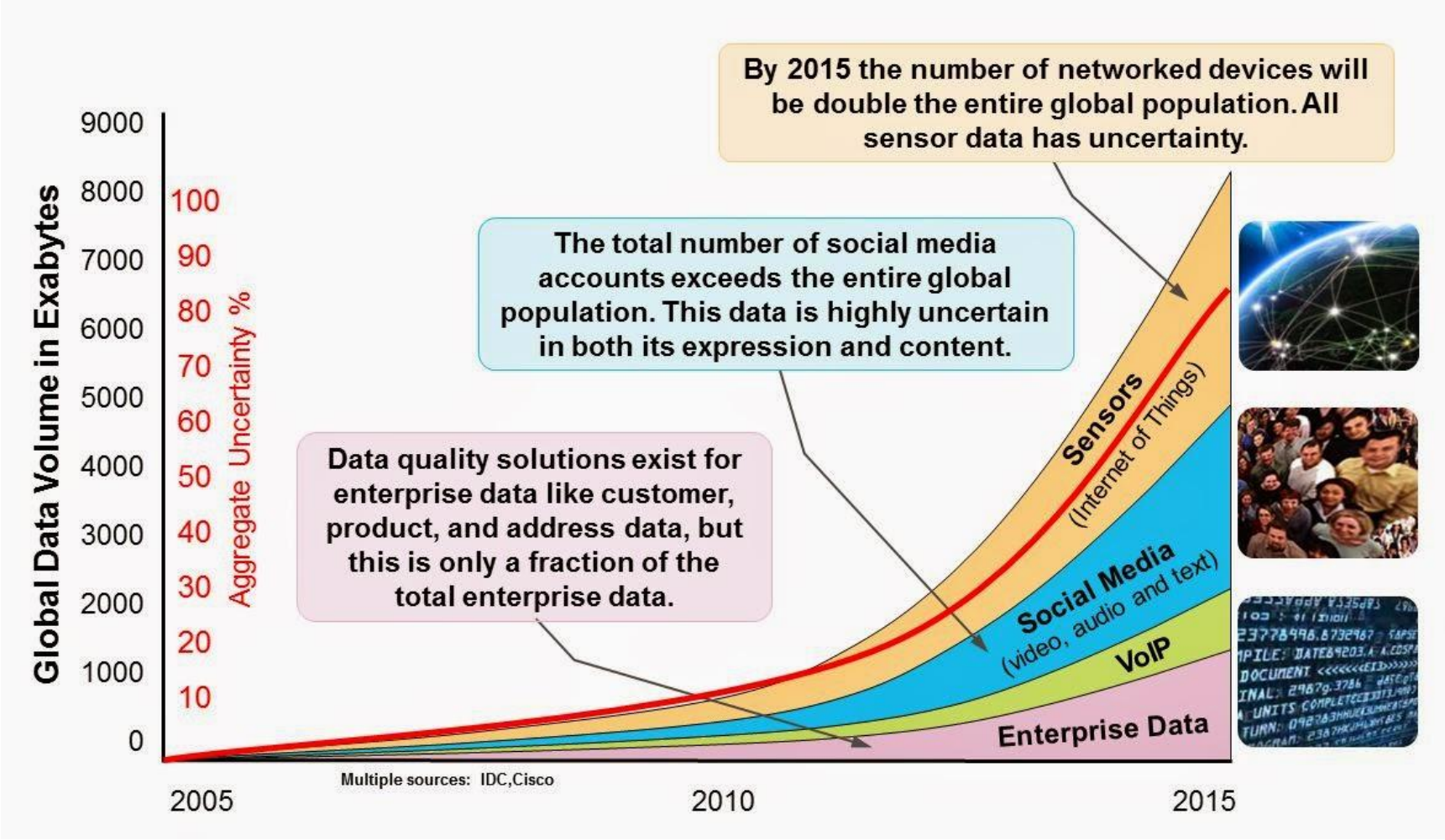
数量 质量

- Data = quantity + quality
  - Some argues that veracity is the most important V in big data
  - 4-th V in big data

- Can we trust the answers to our queries and the prediction result?
  - Dirty data routinely lead to misleading financial reports, strategic business planning decision ⇒ loss of revenue, credibility and customers, disastrous consequences
  - Example: machine learning

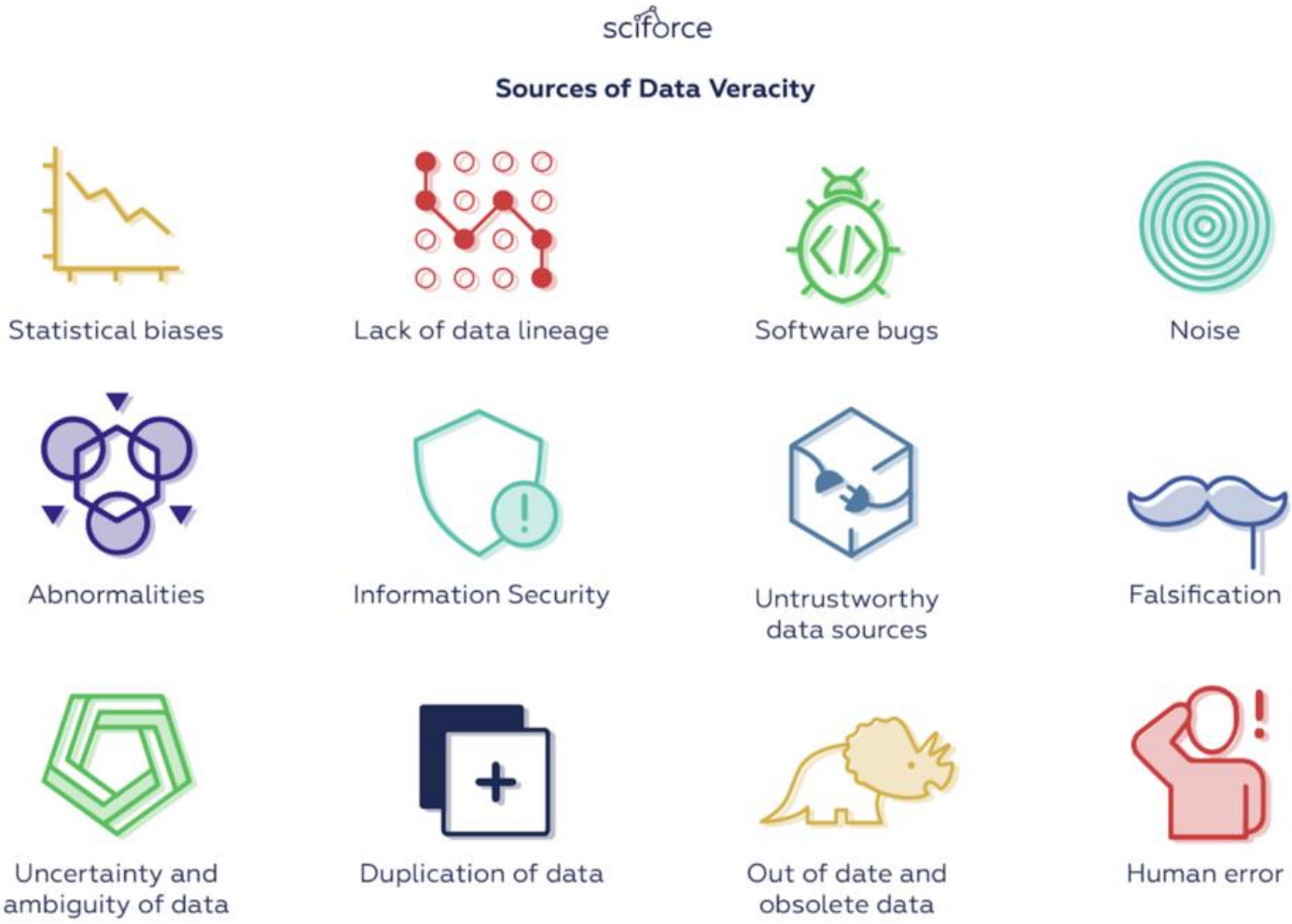
真实性

Veracity



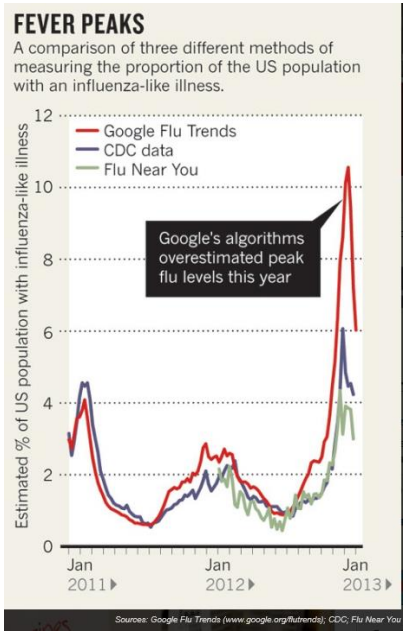
Source: IBM

真实性
Veracity – Where the Uncertainties Come From



Veracity – Why challenging?

- Easy to occur
  - Due to other Vs
- Huge effect to downstream applications
  - E.g., Google Flu Trends
- Difficult to control
  - Identify errors
  - Handle errors
    - correction
    - eliminate the effects



Source

可变性

Variability

Variety:
same entity,
different data

Variability:
same data,
different meaning

Variability

- Meaning of data changing all the time
  - This is a great experience!
  - Great, it totally ruined my day!

- Requires us to have a deeper understanding of the data
  - E.g., make use of the context of the data

Visibility

- Visualization is the most straightforward way to view data
  - Benefits of data visualization

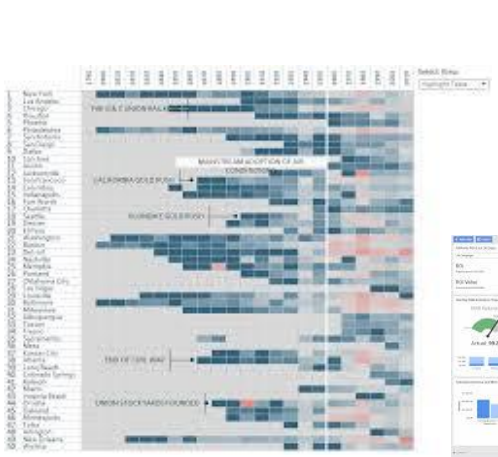
Benefits	Percentages (%)
Improved decision-making	77
Better ad-hoc data analysis	43
Improved collaboration/information sharing	41
Provide self-service capabilities to end users	36
Increased return on investment (ROI)	34
Time savings	20
Reduced burden on IT	15

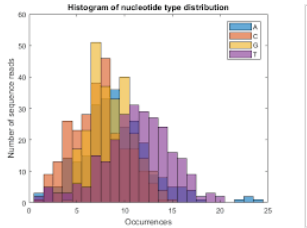
Source: V. Sucharitha, S.R. Subash and P. Prakash , Visualization of Big Data: Its Tools and Challenges

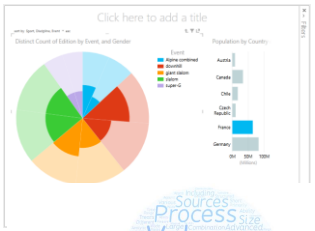
Visibility

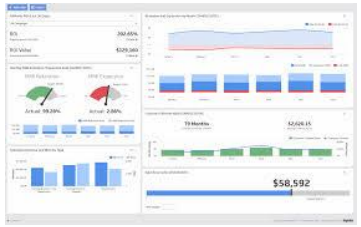
- How to capture and properly present the characteristics of data

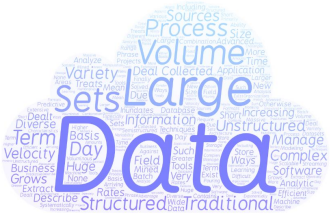
- Simple graphs are only the tip of the iceberg.
- Common general types of data visualization:
  - Charts
  - Tables
  - Graphs
  - Maps
  - Infographics
  - Dashboards











Visibility – Why challenging?

- Choose the most suitable way to present data
  - Characteristics of data
  - Purpose of presentation

- Difficulty of data visualization
  - High dimensional data
  - Unstructured data
  - Scalability
  - Dynamics

Value

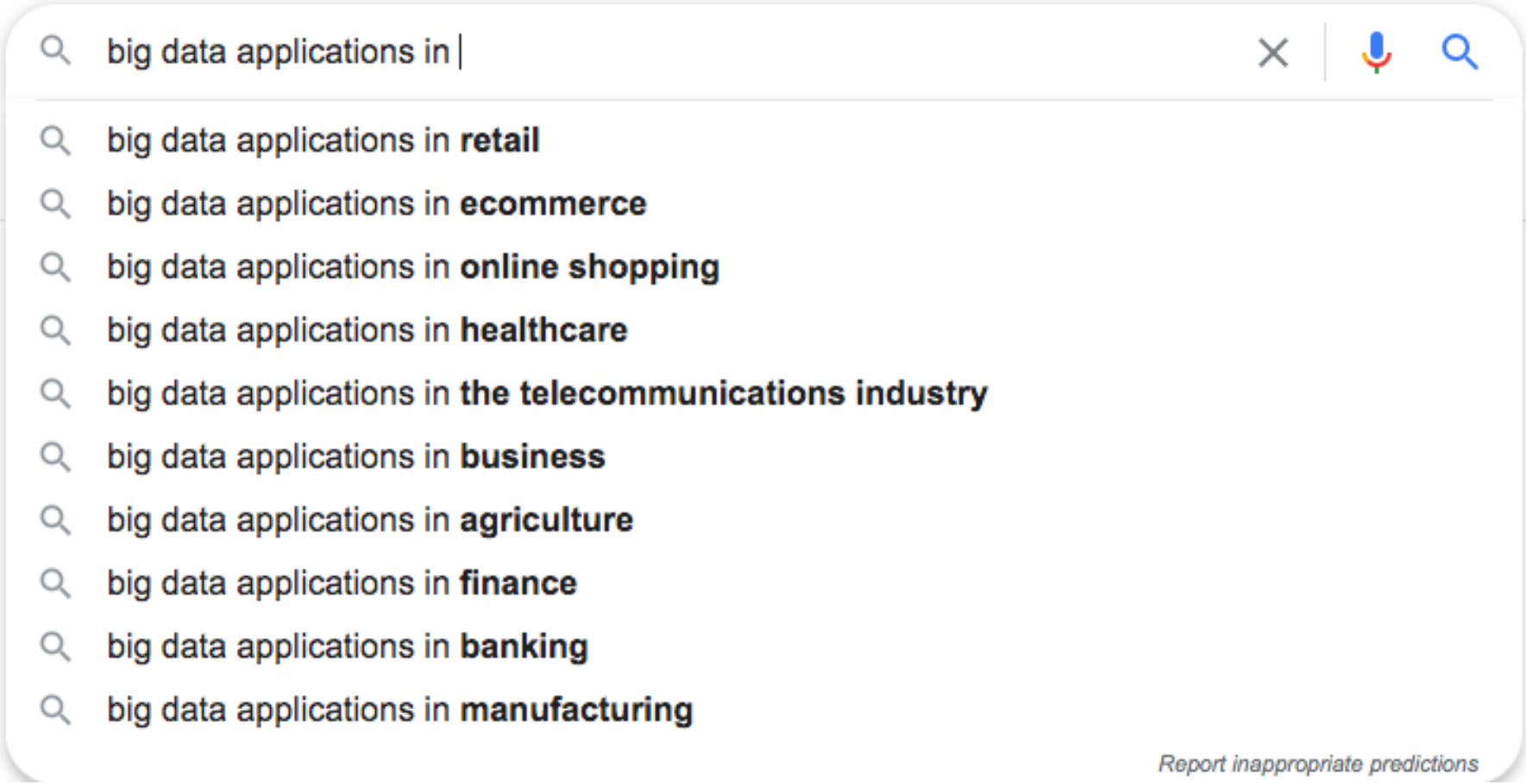
- Big data is meaningless if it does not provide value toward some meaningful goal
- Value from other Vs
  - Volume
  - Variety
  - Velocity
  - …
- Value from applications of big data

Summary of 7 V's in Big Data

- Fundamental V's
  - Volume
  - Variety
  - Velocity
- Characteristics/difficulties
  - Veracity
  - Variability
- Tools
  - Visibility
- Objective
  - Value

- And many other V's …

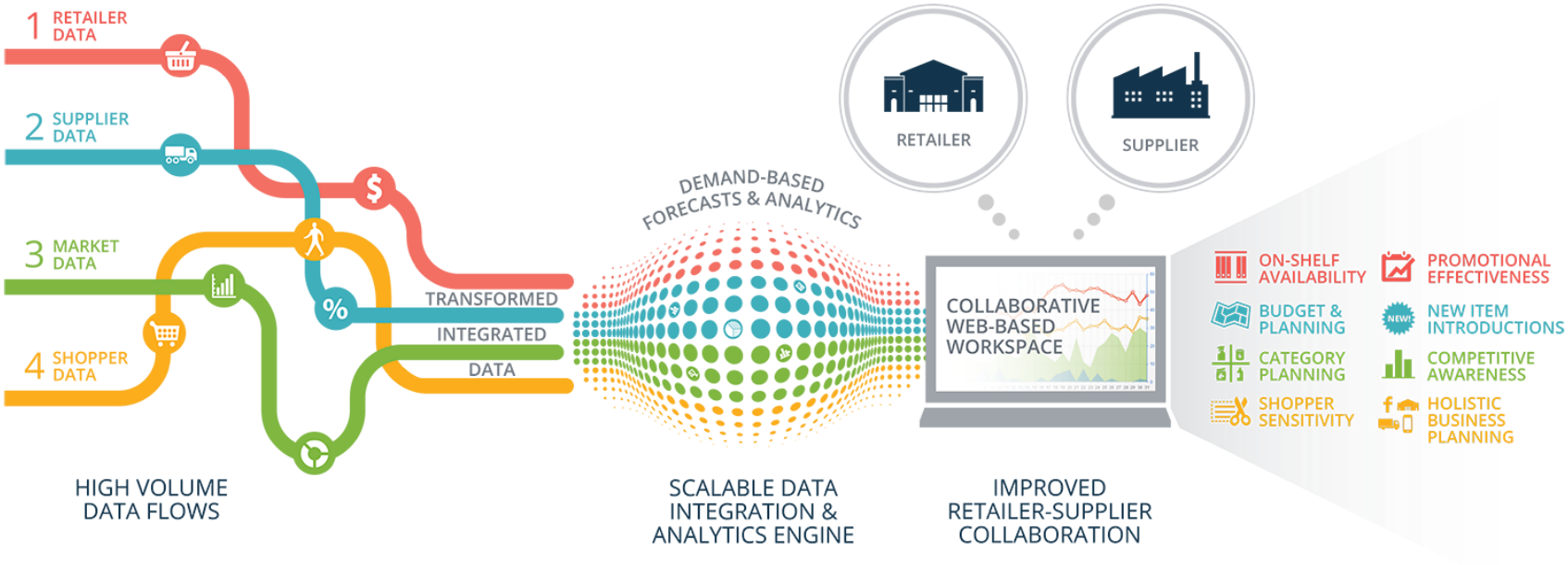
Big Data Applications



Source: google.com

Big Data in Retail

- Retailer:
  - Adjust the price
  - Improve shopping experience
- Supplier:
  - Adjust the supply chain/stock range



source

Big Data in Entertainment

- Predict audience interests
- Understand the customer churn
- Suggest related videos
- Advertisement target

Source

Big Data in National Security

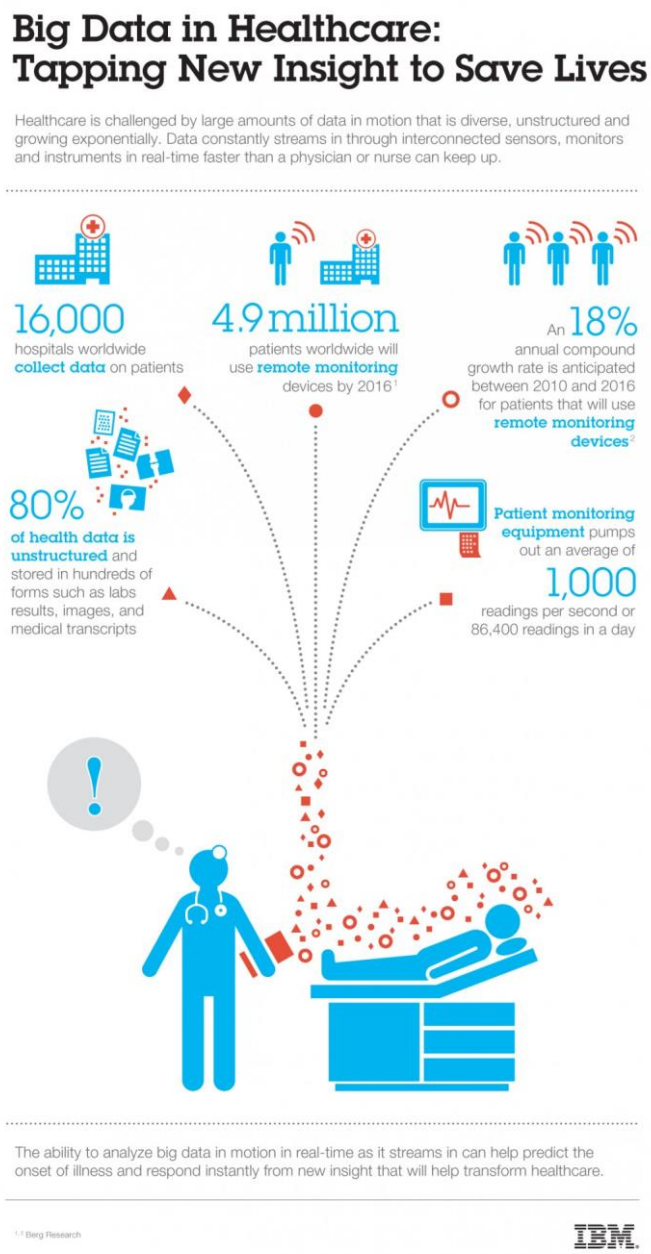
- Integrate shared information
- Entity recognition and tracking
- Monitor, predict and prevent terrorist attacks

Big Data in Science

- Physics
  - The large hadron collider in CERN collect 5 trillion bits of data every second
- Chemistry
  - Extract information from patents
  - Predict the property of compounds
- Biology
  - UK's project alone will sequence 100,000 human genomes producing more than 20 petabytes of data
  - Also helps a lot in medicine domain

Big Data in Healthcare

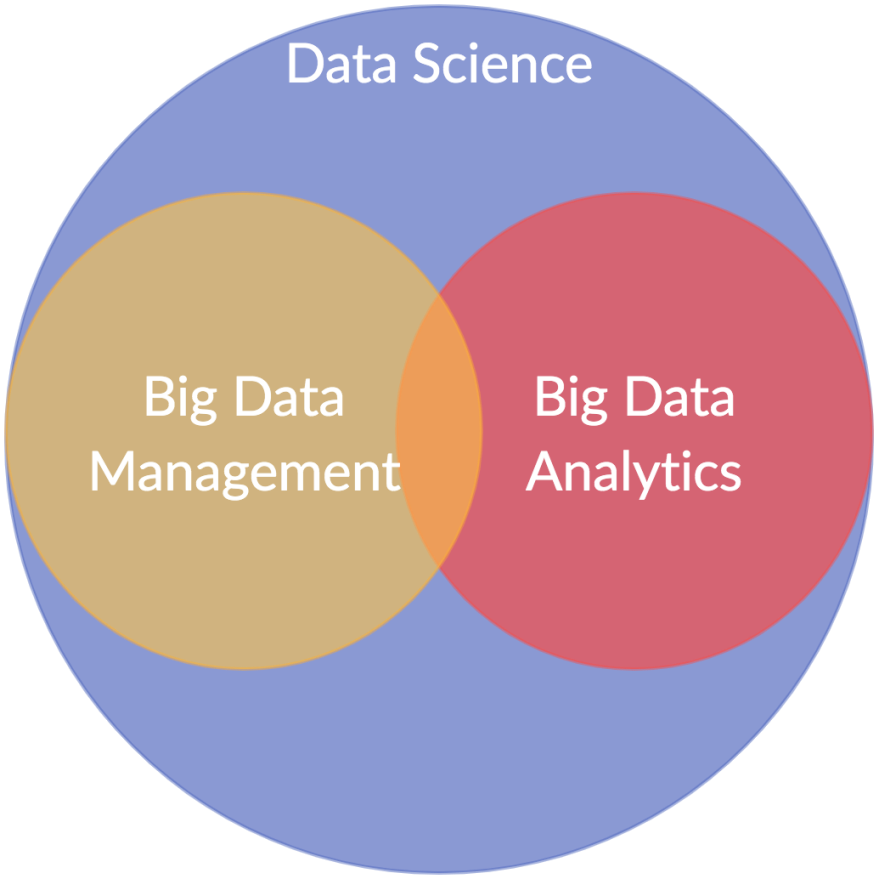
- Diagnostics
  - Data mining and analysis
- Preventative medicine
  - Prevent disease or risk assessment
- Population health
  - Disease trend
  - Pandemics



Source

Introduction to Big Data Management

- Big data management
  - Acquisition 获得
  - Storage
  - Preparation
  - Visualization
- Big data analytics
  - Analysis
  - Prediction
  - Decision making
- Gray (orange?) areas
  - E.g., index construction
- Data science



Example

index	Data Type	Query Type	Accuracy
Binary Search Tree	Sorted keys (1D)	Existence	Exact
B-Tree	Sorted keys (1D)	Range search + NNS + Existence	Exact
Voronoi Diagram	2D	Nearest neighbor search	Exact
R-tree	Multiple dimension	Range search + NNS	Exact
Product Quantization	High dimension	NNS	Approximate
LSH	High dimension	NNS + Range search	Approximate
Bloom Filtering	Any	Existence	Approximate
Count Min	Any	Counting	Approximate

Many other dimensions 尺寸；规模

- Disk-oriented or memory oriented 面向磁盘或内存方向
- Scalability 可扩展性；可伸缩性
- Approx. with or without worst case guarantee

It is meaningful to build indexes only when we know what we need!

获得数据，数据采集

Data Acquisition

- Application oriented
  - Identify data that is relevant to your problem

综合的；广泛的

- Comprehensive
  - Leaving out even a small amount of important data can lead to incorrect conclusions
- Handle data 处理数据
  - from different sources
  - with different types
  - with different velocities

Data Acquisition

- Data in relational databases
  - Structured data
  - Access by SQL
- Data in text files and excel spreadsheets
  - Unstructured or structured data
  - Access by scripting languages (e.g., python, perl)
- Data from website
  - Semi-structured data (e.g., XML) and unstructured data (e.g., image)
  - Access
    - Web socket services
    - REST
    - Crawler

Data Acquisition

- Scientific data
  - E.g., physics experiments, genome data
  - Structured, semi-structured, unstructured
  - Access by specially designed software
- Graph data
  - E.g., knowledge graphs, social networks
  - Access by specially designed programs
  - Difficult to handle (e.g., graph isomorphism problem)
- …

Hybrid in Real Applications

- Usually need to acquire data from multiple resources
- E.g., COVID-19 Map from JHU
  - WHO, CDC, …
    - Structured data (tables)
  - Media reports and Social media (e.g., DXY)
    - Unstructured text data
  - Acquire data from website
  - Extract information from text/tables

Data Storage

- Big data storage is challenging
  - Data Volumes are massive
  - Reliability of Storing PBs of data is challenging
  - All kinds of failures: Disk/Hardware/Network Failures
  - Probability of failures simply increase with the number of machines …

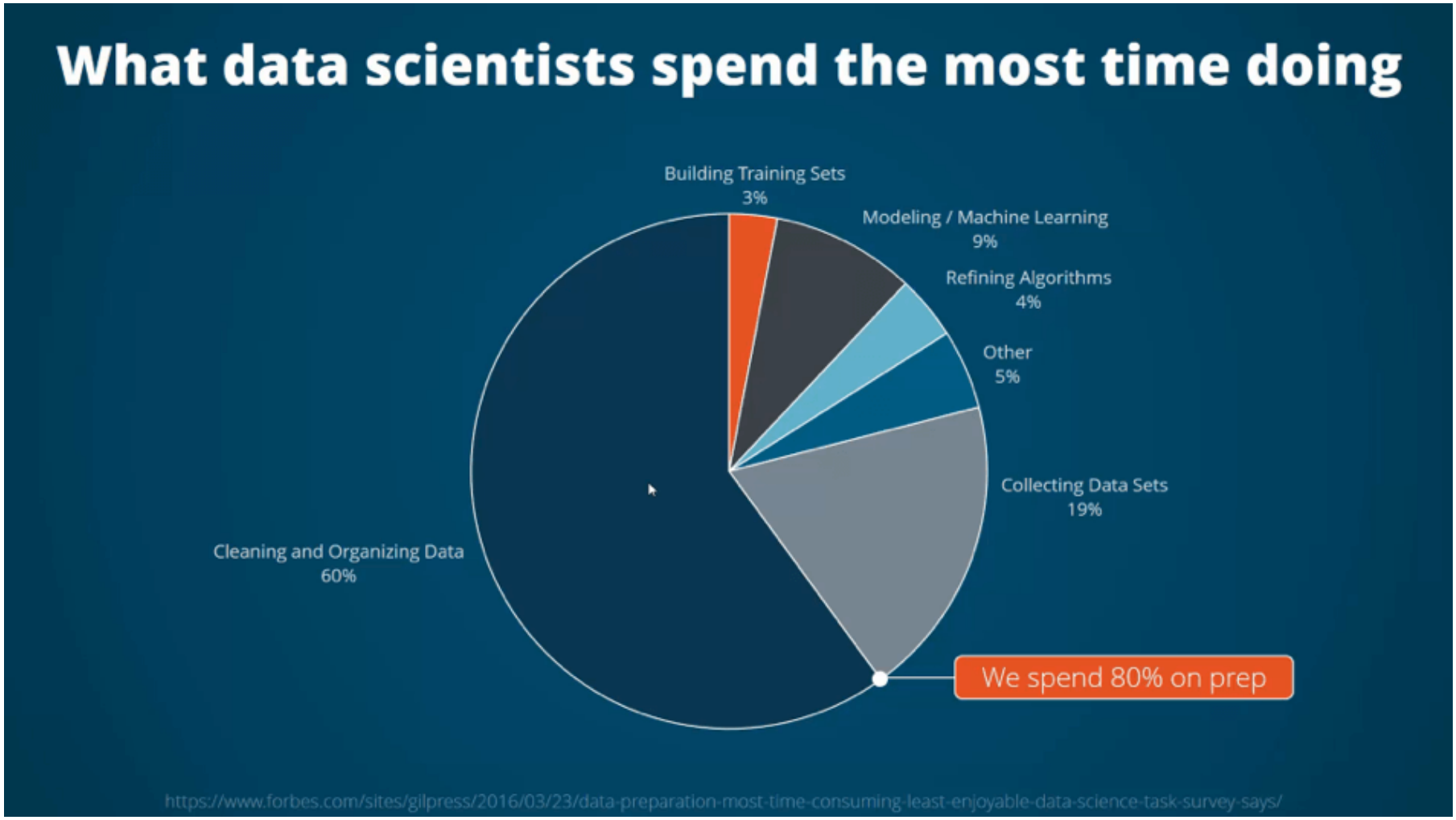
- You don't want to find a needle in your big data haystack.

Data Storage

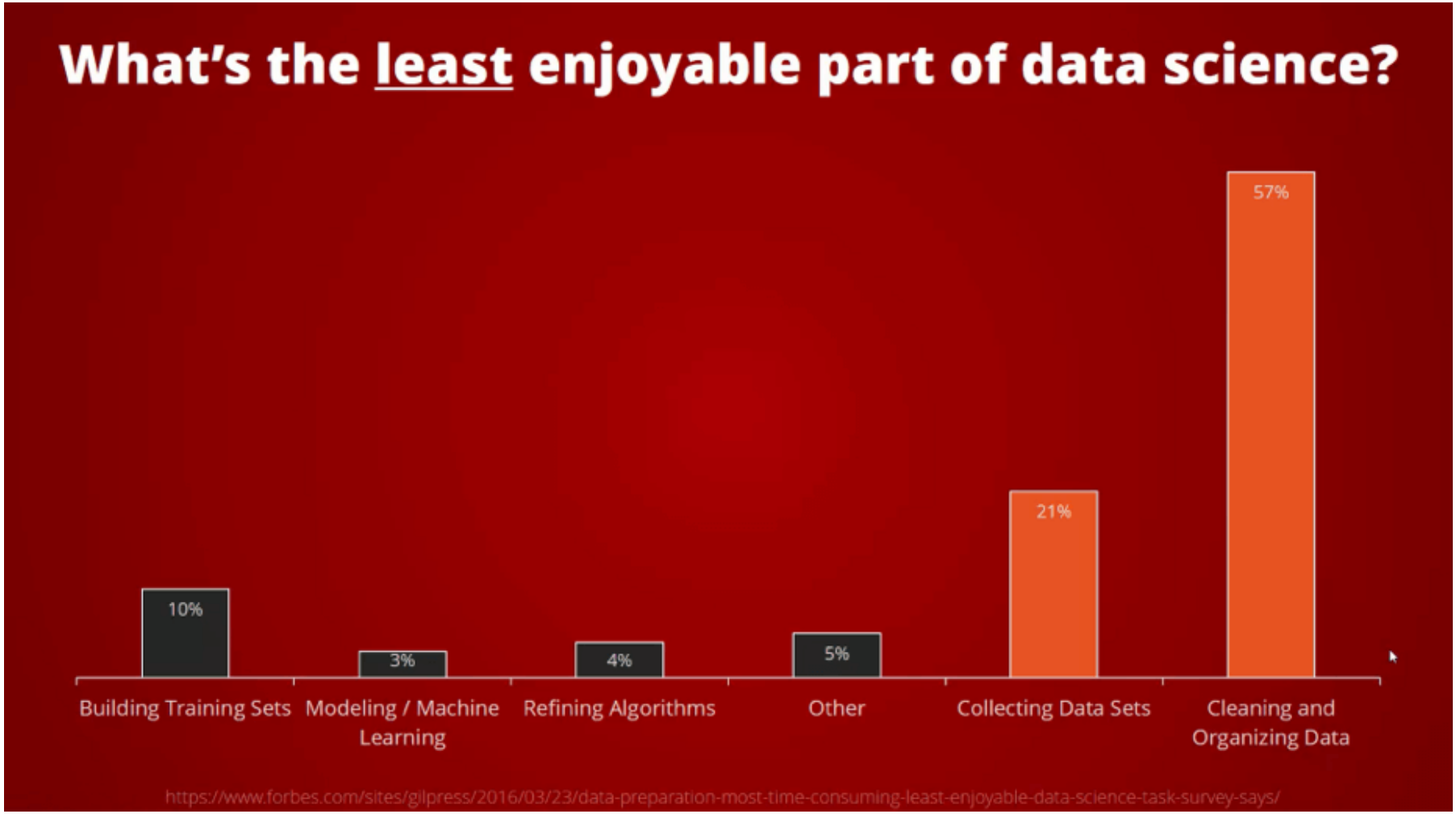
- Traditional way (e.g., RDBMS)
  - Designed for structured data
  - Disk-oriented

- Big data era
  - RDBMS
    - SAP HANA
  - NoSQL
    - HBase, Hive, MongoDB
  - Distributed file systems
    - HDFS
  - …

Data Preparation



Data Preparation



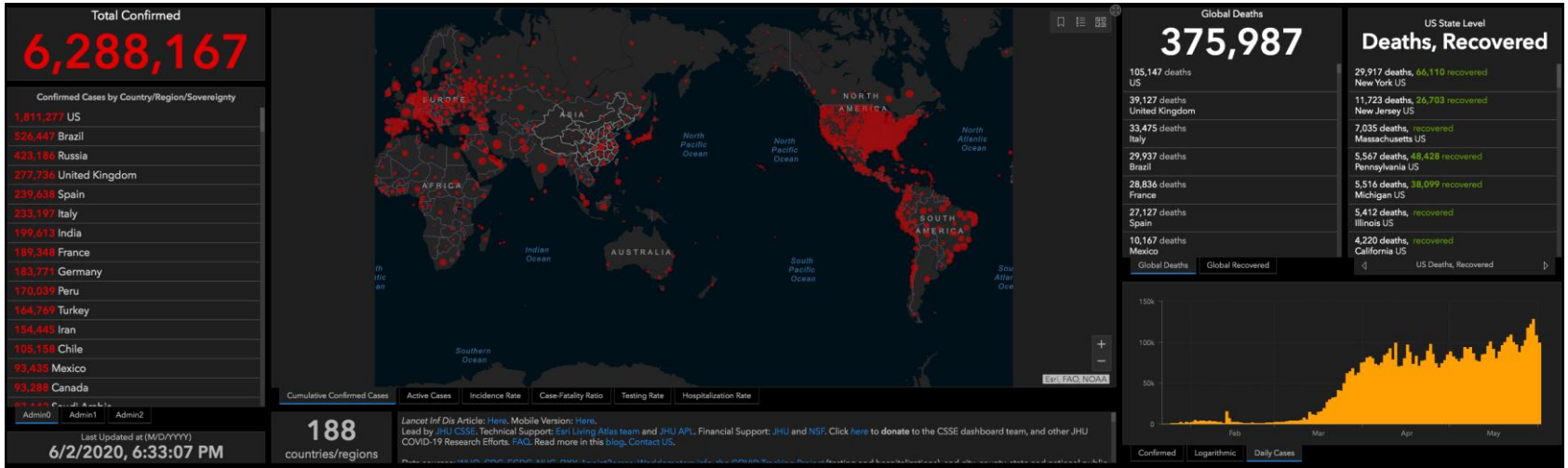
Data preparation

- Two-step data preparation process
- Data Exploration
  - understand your data
- Data pre-processing
  - Data cleansing
    - Veracity
  - Data Integration
    - Variety

Data Exploration

- Explore
  - Trends
  - Correlations
  - Outliers
  - Statistics
    - Mean, Mode, Median, Standard deviation, Range
- Visualization also helps data exploration

Source



数据清洗 Data Cleansing

- Dirty data types
  - Miss values/records
  - Invalid data 不正确的数据
  - Inconsistency 不一致；不一致性
  - Duplicate 复制；复印
  - Outliers 异常值

- Data cleansing requires data understanding and domain knowledge

Data Integration

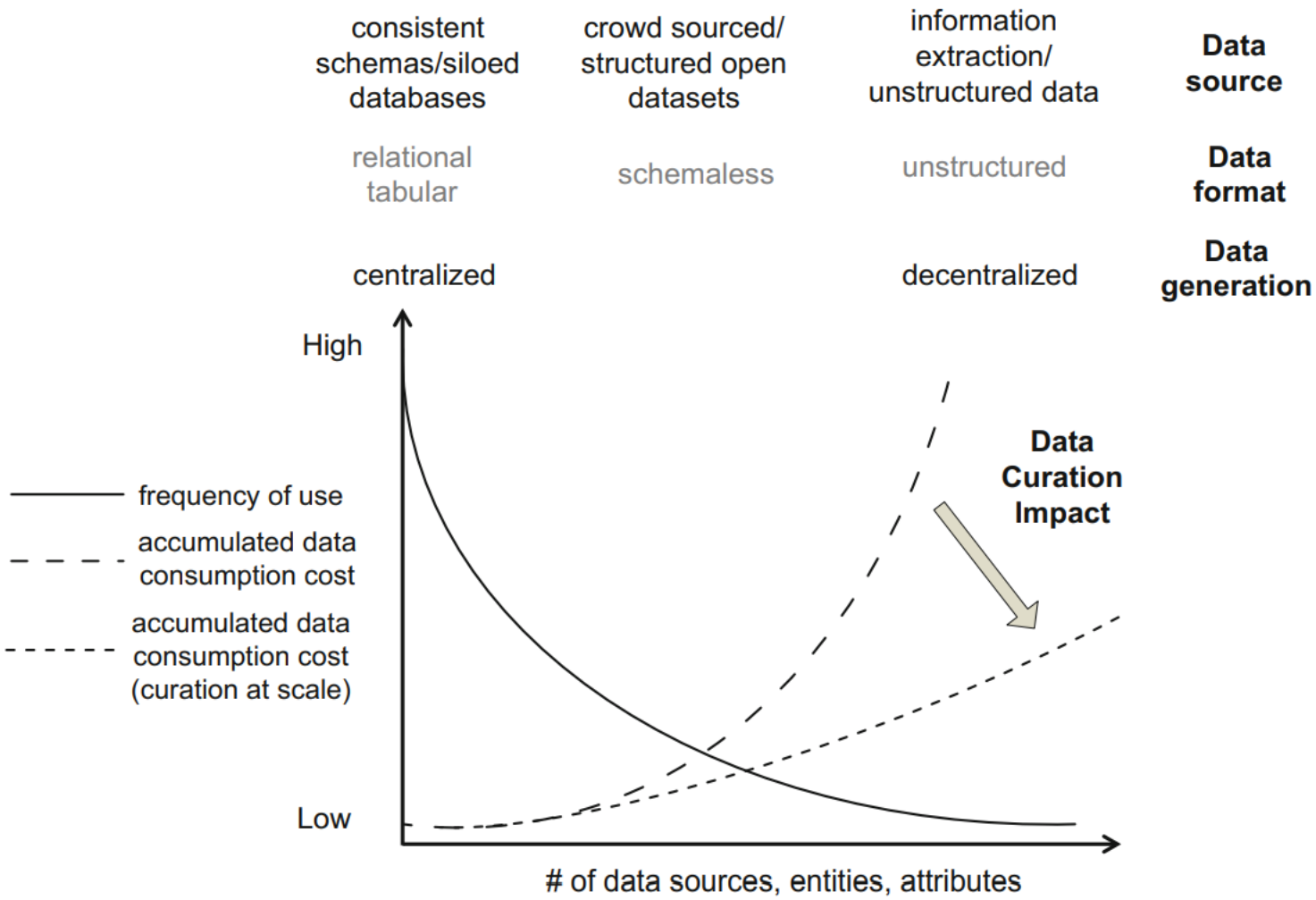
合并

- Merge data from multiple, complex and heterogenous resources.
  - To perform a unified view of data
- Mature field in traditional databases
  - Schema mapping
    - Variety
  - Record linkage
    - Identify if two records refers to same entity
    - Variety, velocity
  - Data fusion
    - Resolving conflicts
    - veracity

Data Curation

- Data curation includes all the processes needed for principled and controlled data creation, maintenance, and management, together with the capacity to add value to data.
  - Analogy to an art curator…
  - make decisions regarding what data to collect,
  - oversee data care and documentation (metadata)
  - conduct research based on the collection
  - data-driven decision making
  - ensure proper packaging of data for reuse
  - share that data with the public
  - …

The Long Tail of Data Variety and Data Curation



Source: Curry, E., & Freitas, A. (2014). Coping with the long tail of data variety.