Big Data:
Challenge and Opportunities

Big Data Society Seminars

https://data-science-group.github.io/BigDataSociety/

Dr. Amin Beheshti

Data Analytics Research Group

Department of Computing

Macquarie University

26 May 2018

Introduction to Big Data

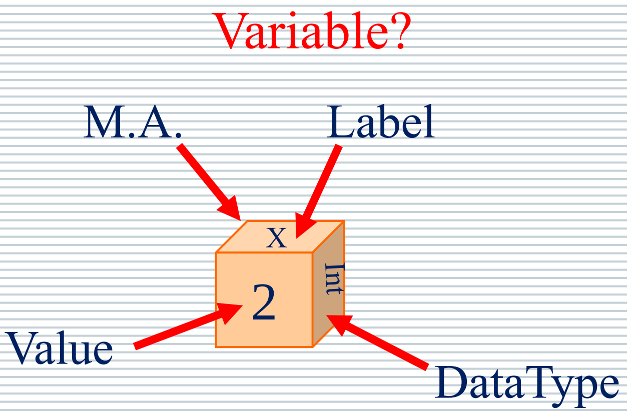
Application

Set of Related **Programs**

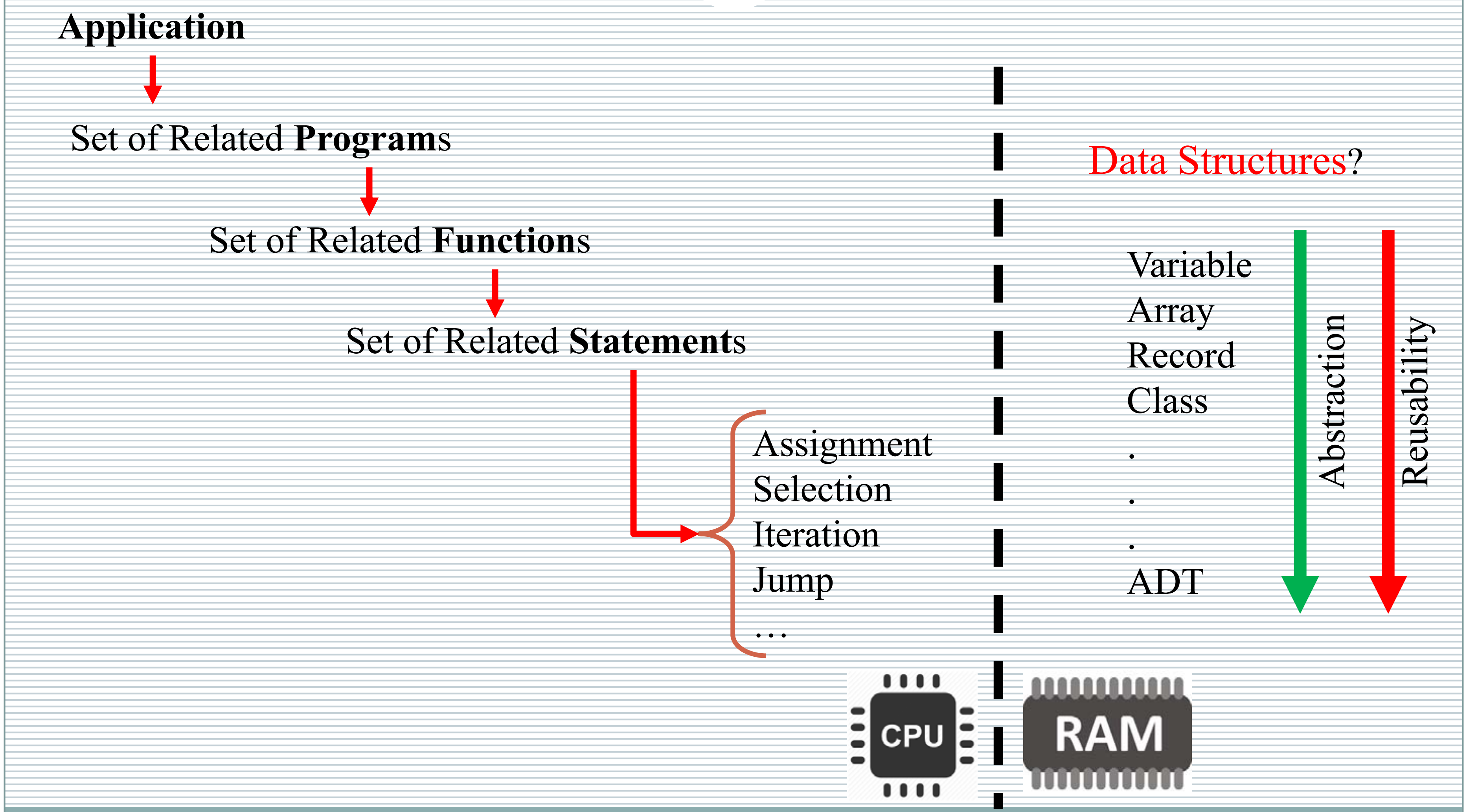
Set of Related **Functions**

Set of Related **Statements**

- Assignment X ← 2
- Selection
- Iteration
- Jump
- …



Introduction to Big Data



Application

Set of Related **Programs**

Set of Related **Functions**

Set of Related **Statements**

Assignment
Selection
Iteration
Jump
…

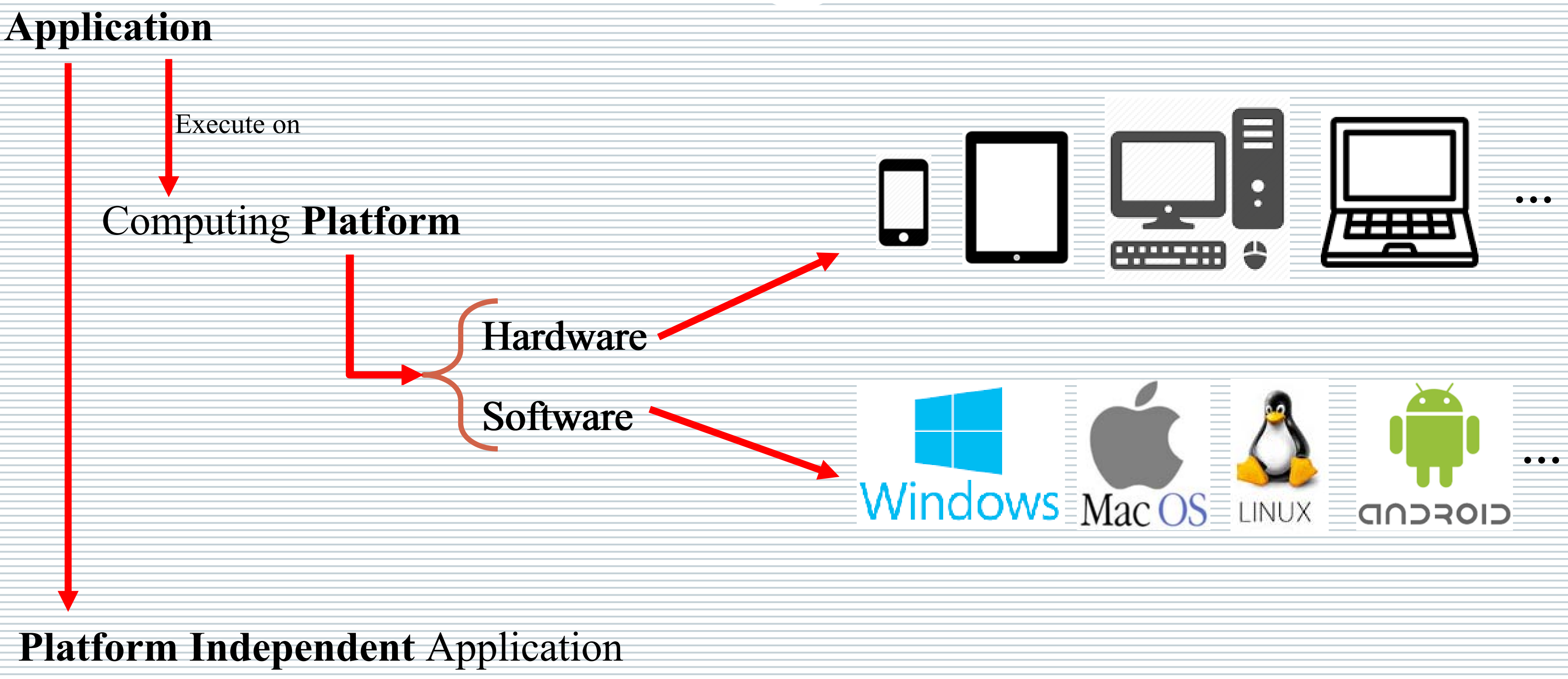
Data Structures?

Variable
Array
Record
Class
.
.
.
ADT

Abstraction

Reusability

Introduction to Big Data



Application

Execute on

Computing **Platform**

Hardware

Software

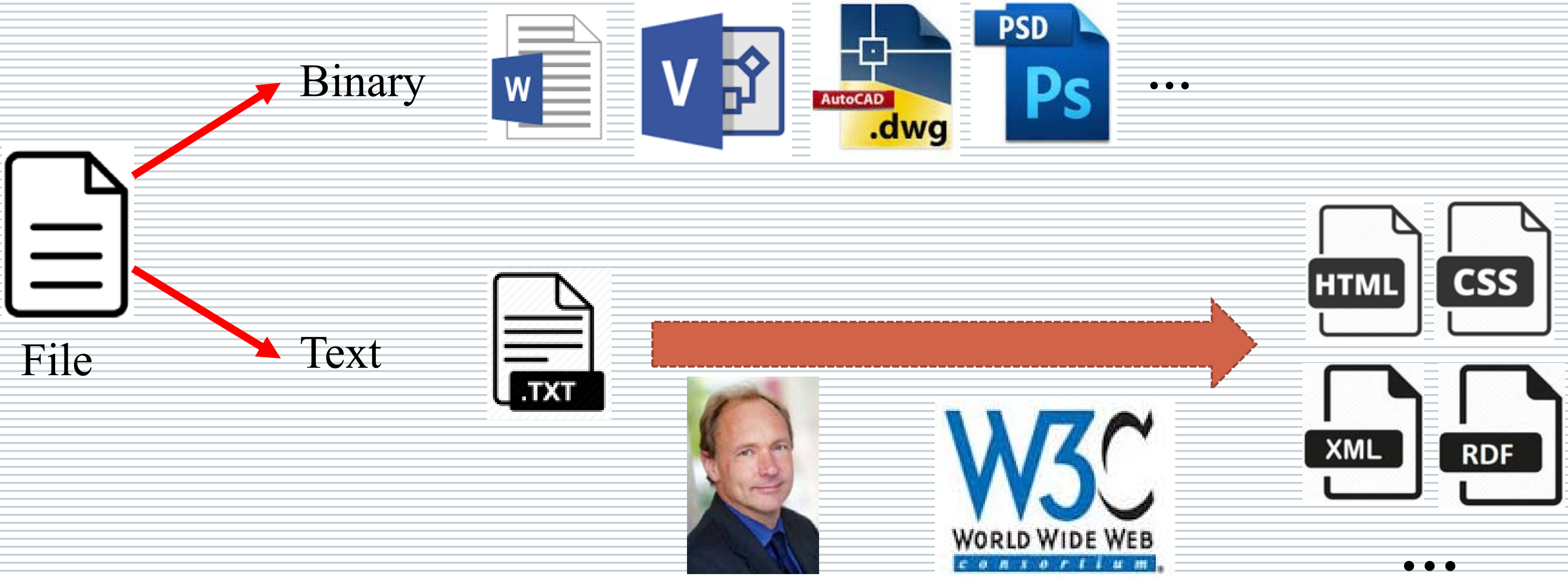
Platform Independent Application

Introduction to Big Data

Application

↓

Platform Independent Application
(e.g. Web Applications)



File

Binary

Text

Tim Berners-Lee

Introduction to Big Data

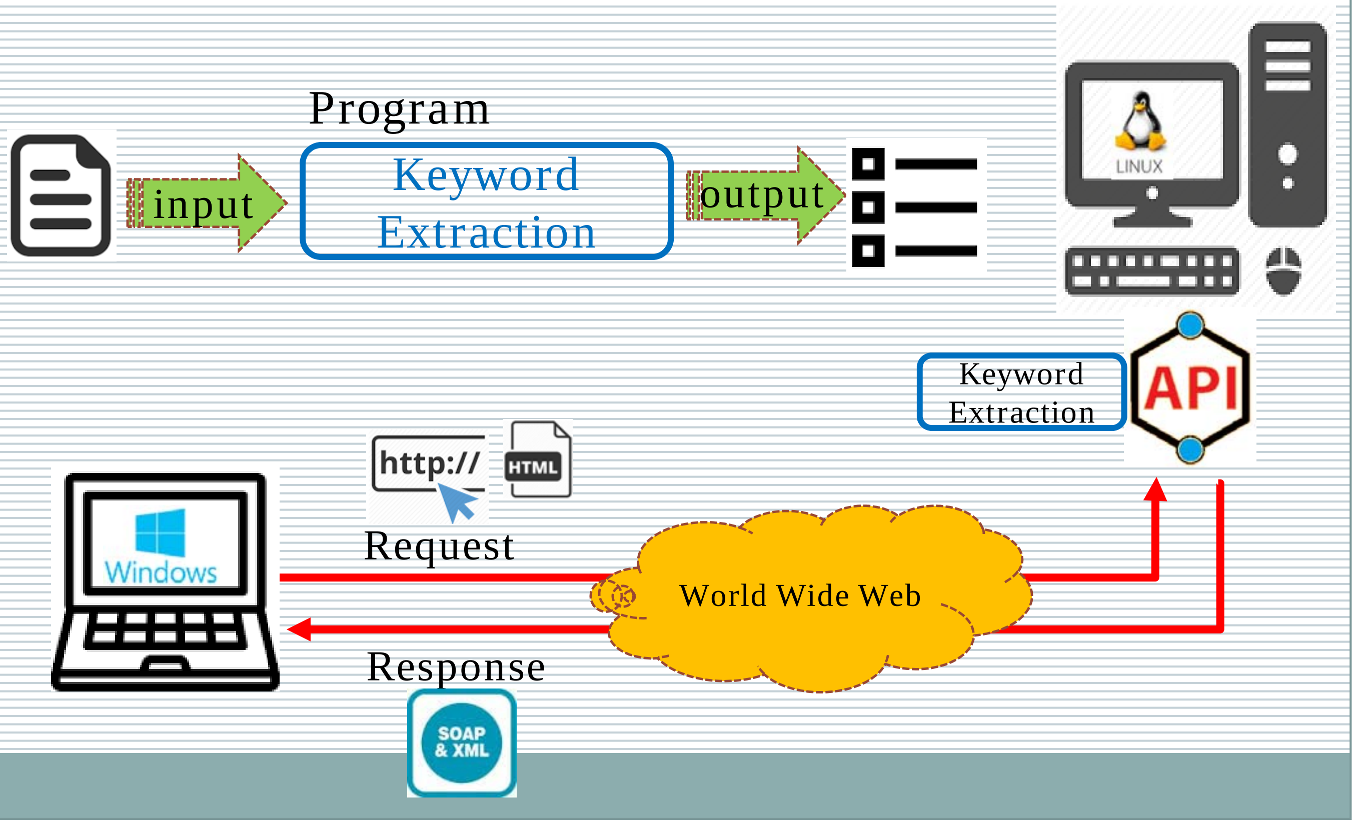
Application

↓

Platform Independent Application
(e.g. Web Applications)

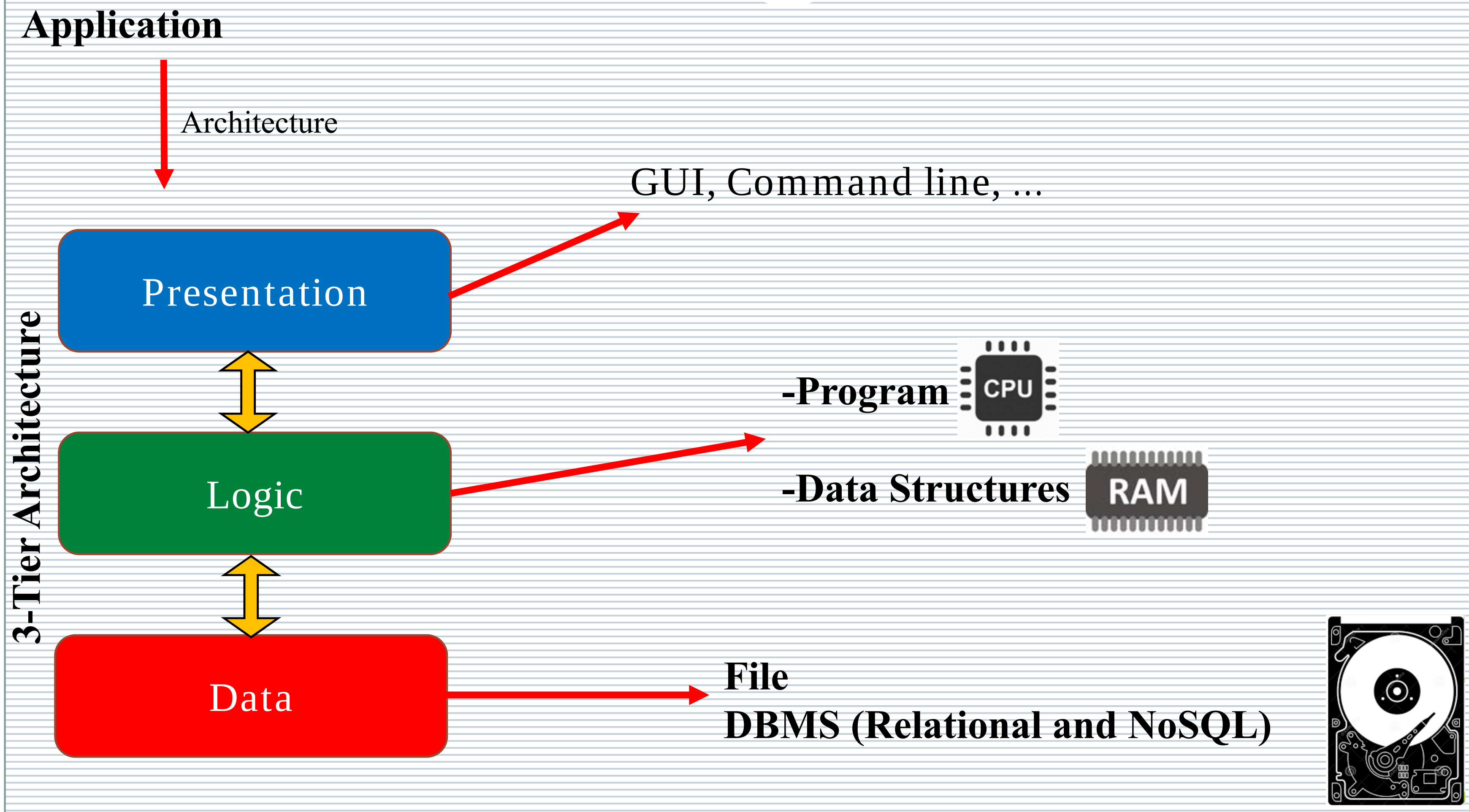
↓

Web Services



-API Engineering
-Microservices

Introduction to Big Data



Application

Architecture

3-Tier Architecture

Presentation

Logic

Data

GUI, Command line, …

-Program

-Data Structures

File

DBMS (Relational and NoSQL)

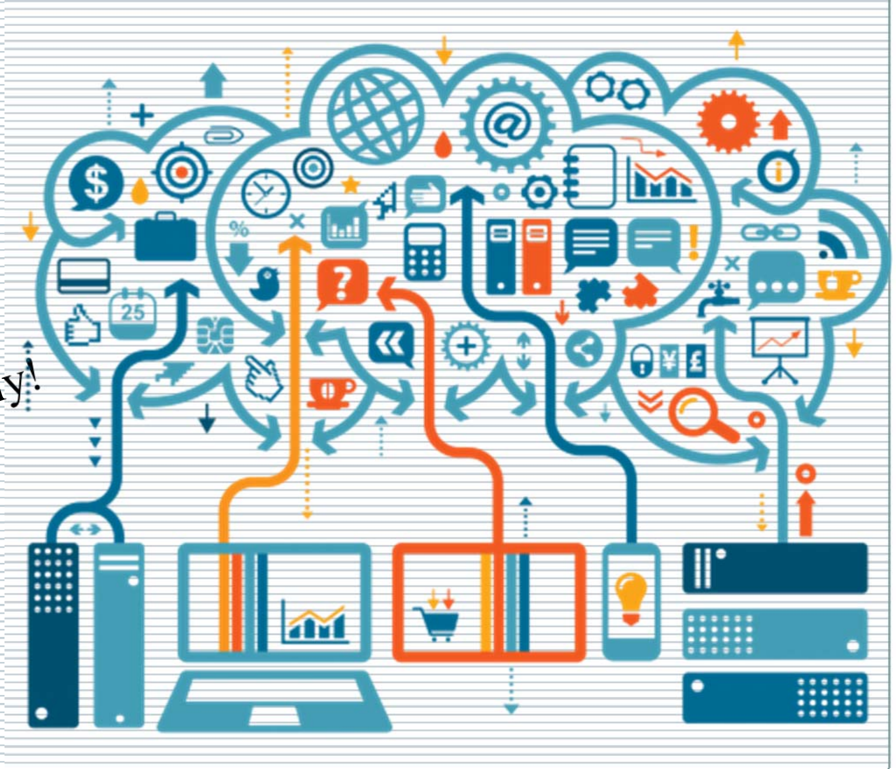
What is Data ?

What is Data ?

Every day, we create **2.5 quintillion** bytes of data.

- posts to social media sites
- sensors used to gather climate information
- digital pictures and videos
- purchase transaction records
- cell phone GPS signals
- …

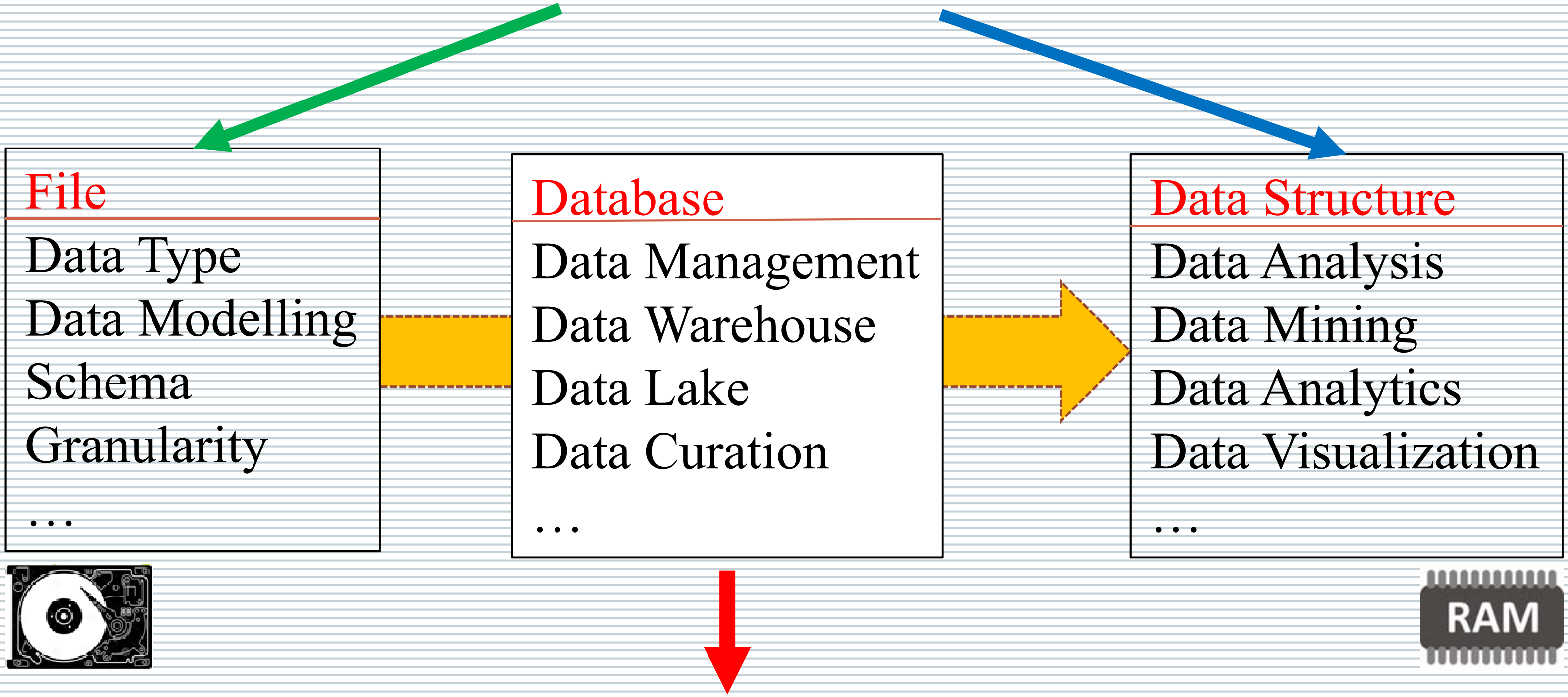
- 500 Million Tweets sent each day!
- 5.75 BILLION Facebook likes every day.
- 3.6 Billion Instagram Likes each day.
- 4.3 BILLION Facebook messages posted daily!
- 6 BILLION daily Google Searches!
- …



https://www-01.ibm.com/; http://www.internetlivestats.com/; http://semeon.com/

What is Data ?

In computing, data is information that has been translated into a form that is efficient for **storage** and/or **processing**.



File

Data Type
Data Modelling
Schema
Granularity
…

Database

Data Management
Data Warehouse
Data Lake
Data Curation
…

Data Structure

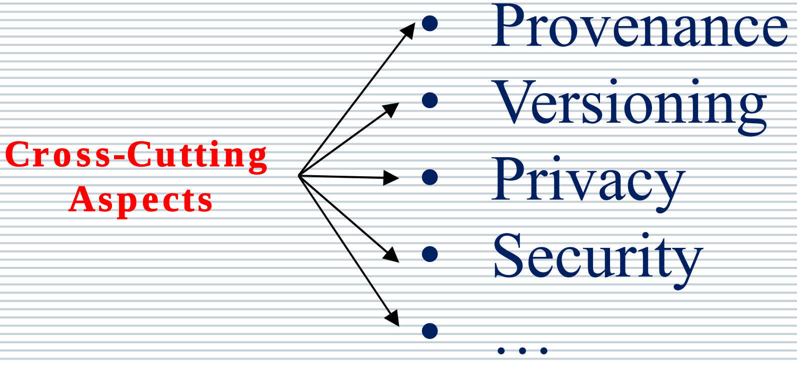
Data Analysis
Data Mining
Data Analytics
Data Visualization
…

Preparing the Data for Processing
(Organizing and Curating)

What is Metadata ?

We are **Tracing** everything:

- What is happening?
- Who is doing that?
- Where it is happening?
- When?
- Why?
- How?
- …



Cross-Cutting Aspects

- Provenance
- Versioning
- Privacy
- Security
- …

- Smart **Phones**, tracks:
  - Our location,
  - Our speed,
  - What apps we are using,
  - What music we listen to,
  - …
- Smart **TVs**, tracks:
  - Channels we watch,
  - Time and duration,
  - Apps we use,
  - …
- Smart **Watches**, tracks:
  - Our health signs,
  - Our activity,
  - Location,
  - …

…

Beheshti et al. "**Enabling the Analysis of Cross-Cutting Aspects in Ad-hoc Processes**", CAiSE Conference (2013)

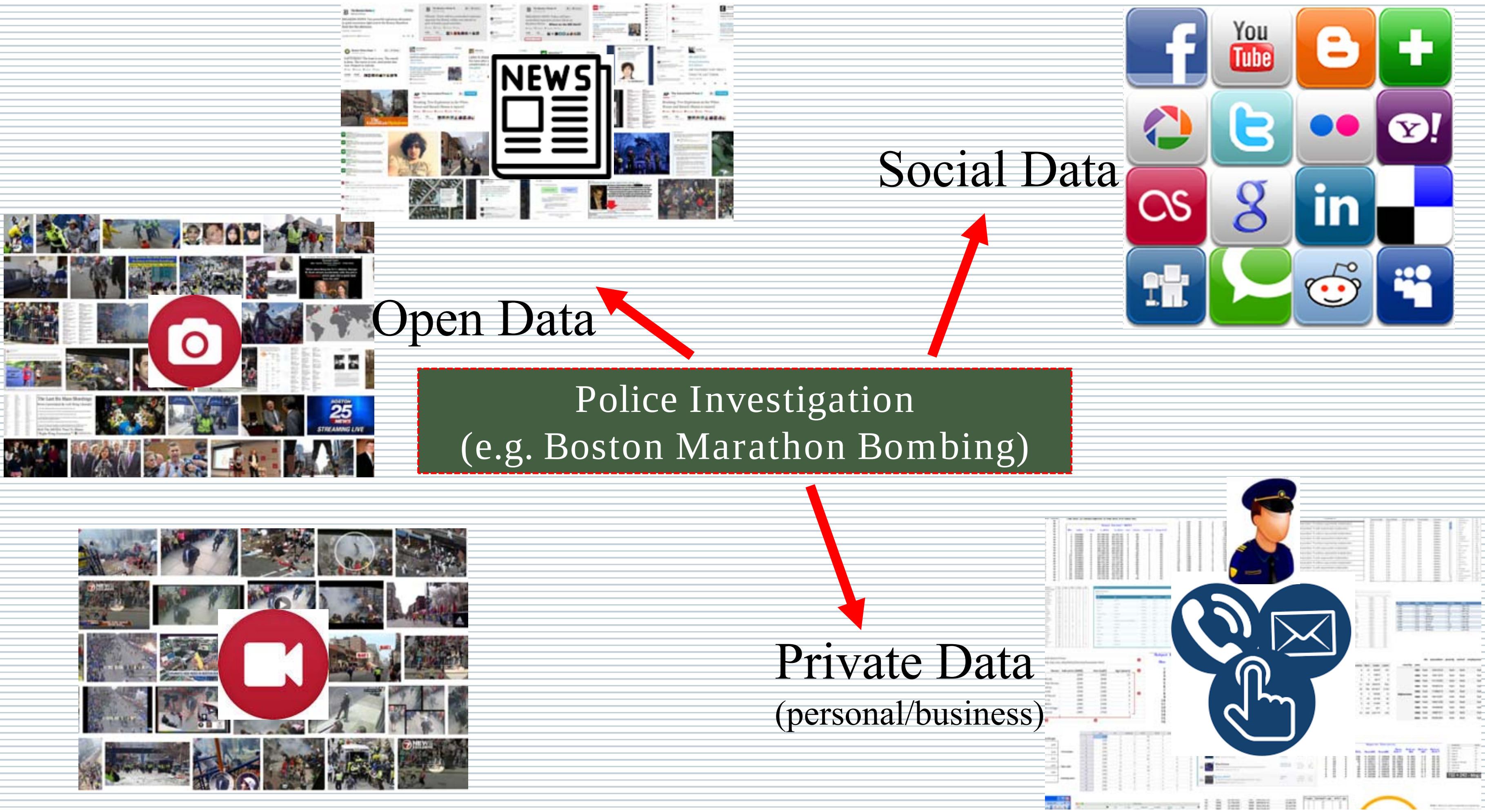
What is Big Data ?

Big Data and Big$^{\text{Big}}$ Metadata

share, comment, review, crowdsource, etc.

Example

What is Big Data ?



Open Data

Social Data

Police Investigation
(e.g. Boston Marathon Bombing)

Private Data
(personal/business)

Beheshti et al. “ProcessAtlas: A scalable and extensible platform for business process analytics”, Software: Practice and Experience, 2018

What is Big Data ?

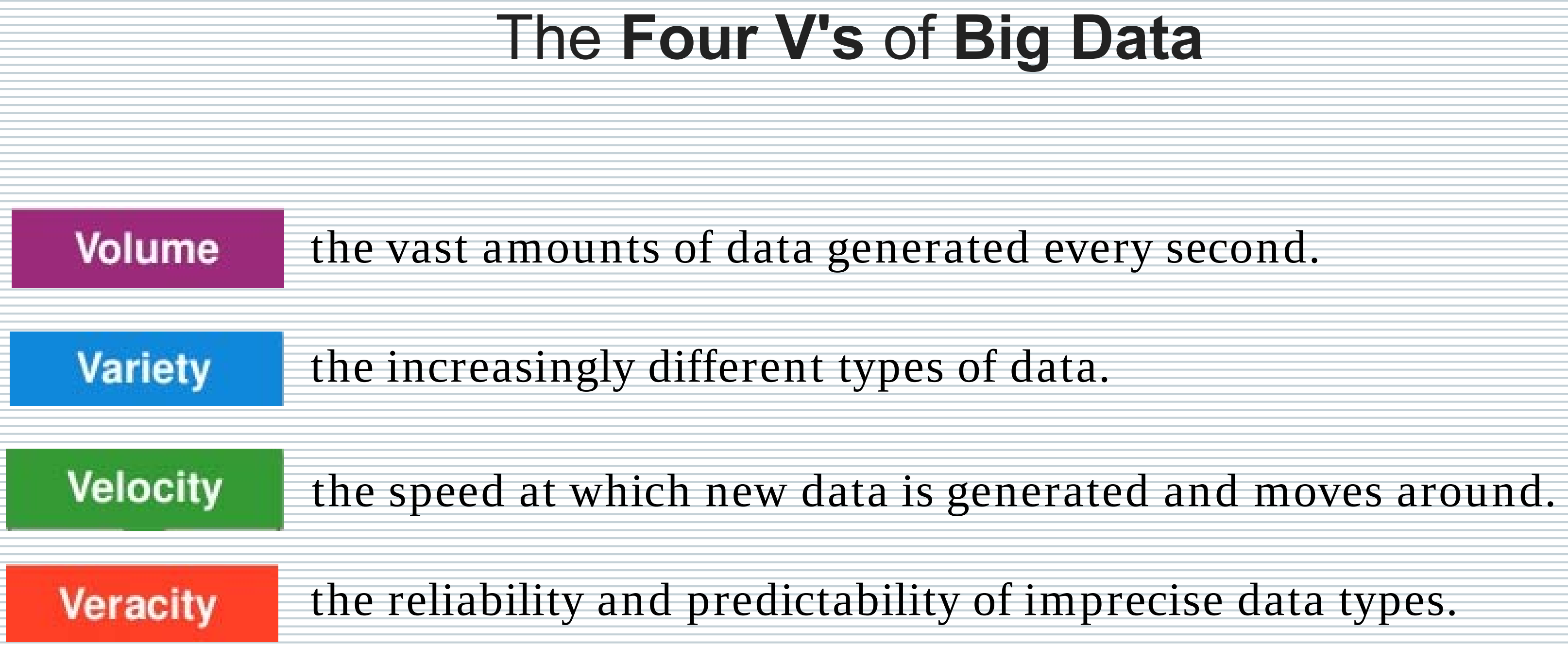
- Big data refers to our ability to collect and analyse the ever expanding amounts of **data** and **meta-data** that we are generating every second!
- Big data can be seen as a massive number of small **data islands** from Private (Personal/Business), Open and Social Data.

Organizing, Curating, Analysing and Presenting this data is *challenging* and of high interest.

Organizing Big Data

Organizing Big data

- How to store vast amount of noisy data (varying from structured entities to unstructured documents) being generated on a continuous basis ?



The **Four V's** of **Big Data**

Volume the vast amounts of data generated every second.

Variety the increasingly different types of data.

Velocity the speed at which new data is generated and moves around.

Veracity the reliability and predictability of imprecise data types.

Big data - Volume

Volume, the quantity of data to be stored, is a key characteristic of Big Data.

How to deal with storing large volume of data ?

Scale Up:

Keep the same number of Systems, but migrating each system to a larger System.

e.g. Changing from a server with 16 CPU cores and 1 TB storage system to a server with 64 CPU cores and a 100 TB storage system.

Scale Out:

When the workload exceeds the capacity of a server, the work load is spread out across a number of servers.

This is also referred to as **Clustering**.

Notice:

It is cheaper to buy ten 100 TB storage systems than it is to buy a single 1 PB storage system

Big data - Velocity

Velocity, refers to the **rate at which new data enters the system** as well as the **rate at which the data must be processed**.

Example:

Past

Present

Amazon used to capture only the **data about the final transaction** of a customer making a purchase!

amazon

Amazon.com

Electronic commerce company

Amazon captures **NOT ONLY** the final transaction **BUT ALSO** every click of the mouse in searching, browsing, comparing, as well as the purchase process.

Instead of capturing 1 event it might capture data on more than 30 events.

30× increase in the velocity of the data.

Big data - Velocity

Velocity, refers to the **rate at which new data enters the system** as well as the **rate at which the data must be processed**.

The velocity of processing can be broken down into: **Stream** and **Feedback Loop** Processing

Stream Processing, requires analysis of the data stream as it enters the system.
(Focus on the INPUT)

Example:
CERN Large Hadron Collider (the largest and most powerful particle accelerator in the world) experiments produce about 600 TB per second of raw data.

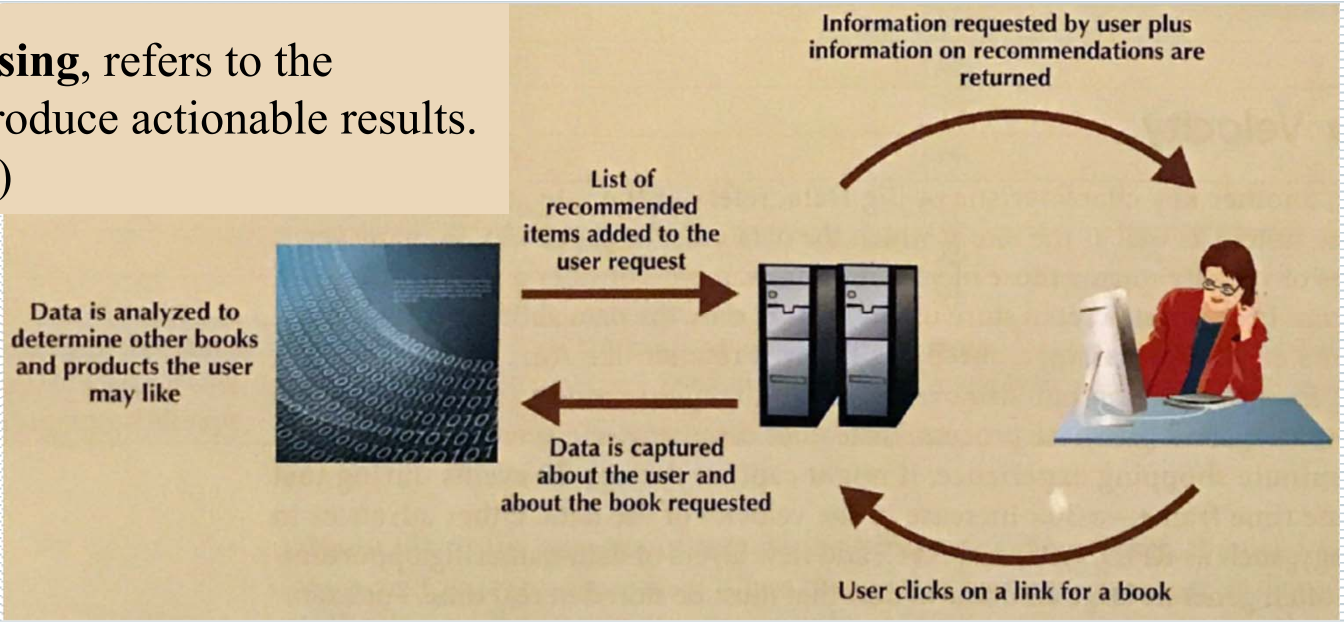
All this data can not be processes, accordingly scientists created algorithms to decide ahead of time which data will be kept; and to **filter the data down** to only about 1 GB per second.

Big data - Velocity

Velocity, refers to the **rate at which new data enters the system** as well as the **rate at which the data must be processed**.

The velocity of processing can be broken down into: **Stream** and **Feedback Loop** Processing

Feedback Loop Processing, refers to the analysis of the data to produce actionable results. (Focus on the OUTPUT)


Information requested by user plus information on recommendations are returned
List of recommended items added to the user request
Data is analyzed to determine other books and products the user may like
Data is captured about the user and about the book requested
User clicks on a link for a book

Big data - Variety

Variety, refers to the vast array of **formats and structures in which the data may be captured**: structured, unstructured and semi-structured.

Structured Data, is data that has been organized to fit a predefined data model.

Unstructured Data, is data that is not organized to fit into a predefined data model.

Semi-structured Data, combines elements of both Structured and Unstructured.

Big data – Veracity

Veracity, refers to the trustworthiness of the data.

Challenge:

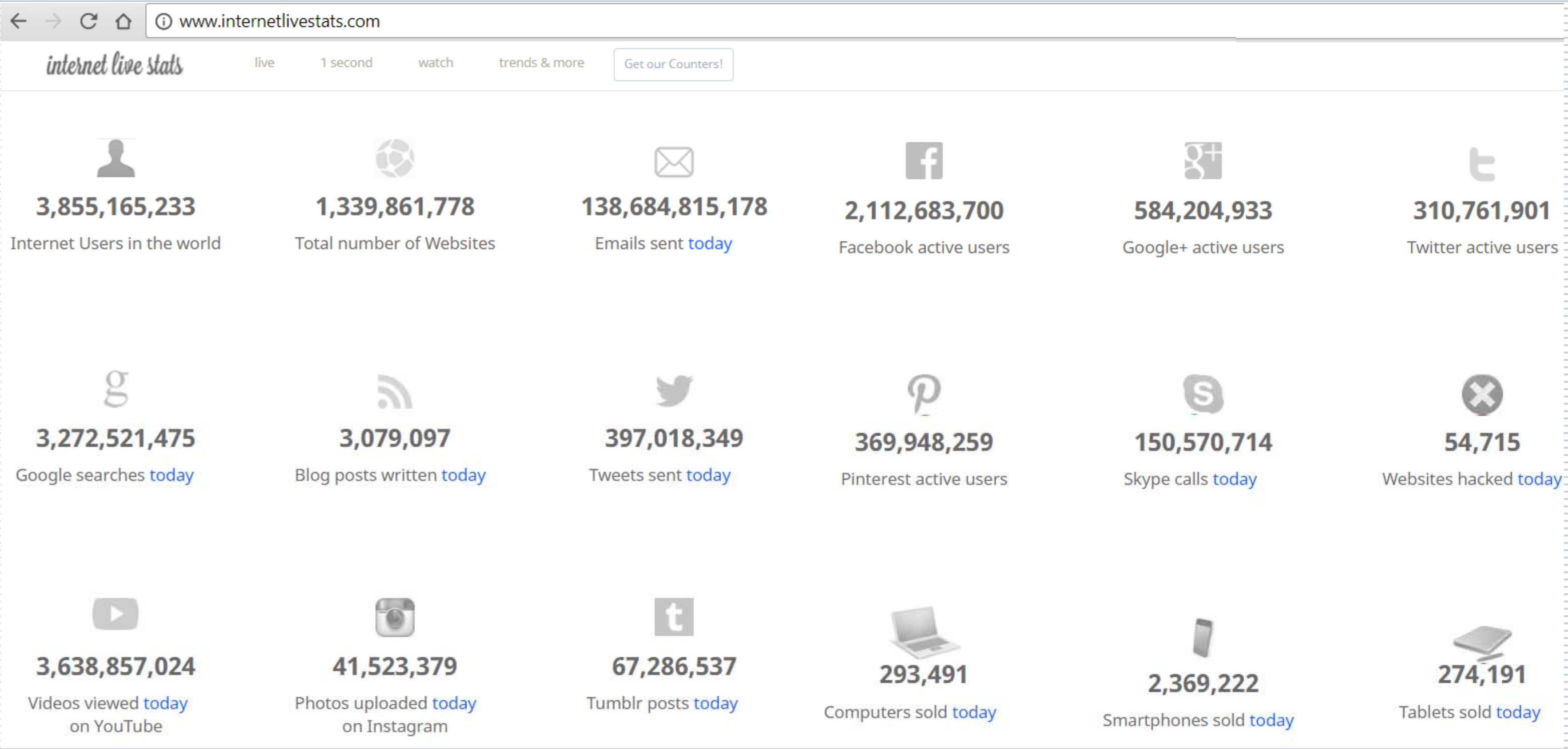
Given the automation of data capture and some parts of the analysis, can decision makers reasonably rely on the accuracy of the data and the information generated from it ?

Example:

Uncertainty about the data can arise from several causes, such as having to capture only selected portions of data due to high velocity! E.g. in CERN

Organizing Big data

How Big is the Big Data ?



3,855,165,233
Internet Users in the world

1,339,861,778
Total number of Websites

138,684,815,178
Emails sent today

2,112,683,700
Facebook active users

584,204,933
Google+ active users

310,761,901
Twitter active users

3,272,521,475
Google searches today

3,079,097
Blog posts written today

397,018,349
Tweets sent today

369,948,259
Pinterest active users

150,570,714
Skype calls today

54,715
Websites hacked today

3,638,857,024
Videos viewed today on YouTube

41,523,379
Photos uploaded today on Instagram

67,286,537
Tumblr posts today

293,491
Computers sold today

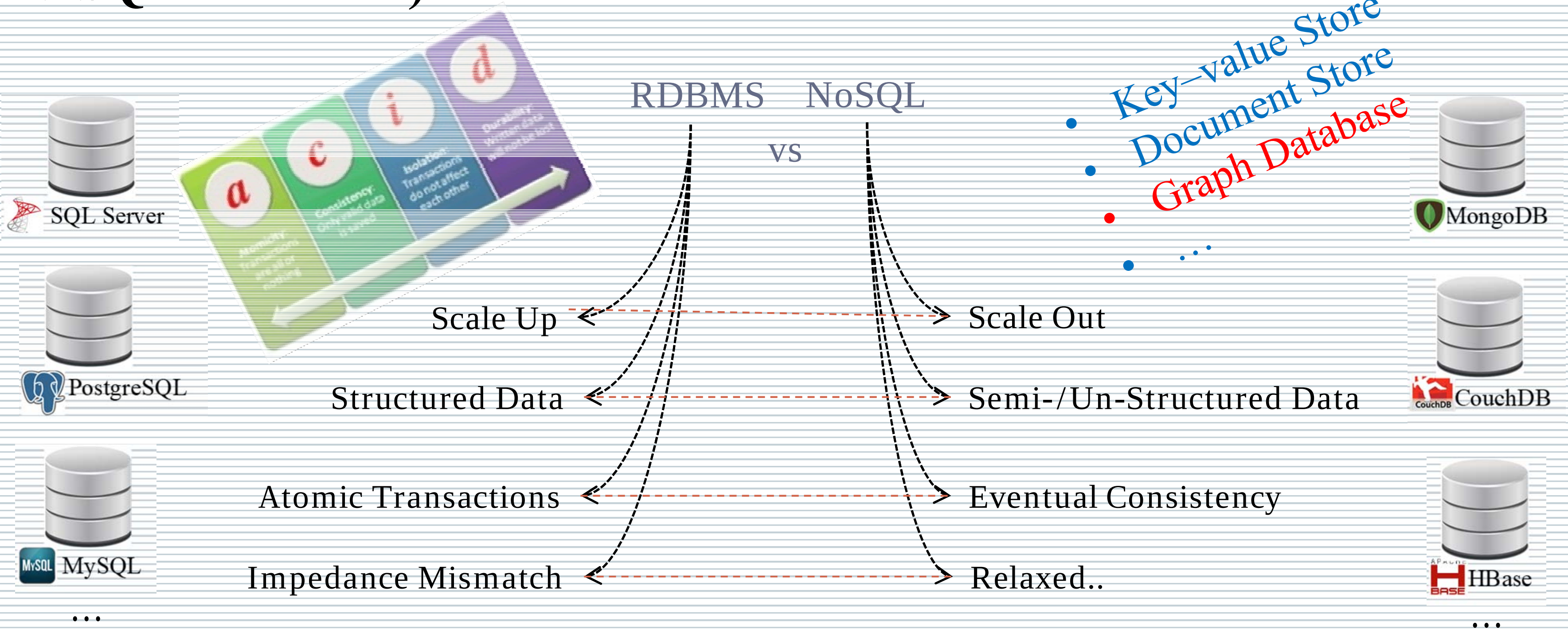
2,369,222
Smartphones sold today

274,191
Tablets sold today

http://www.internetlivestats.com/

Organizing Big data

- How to store vast amount of noisy data (varying from structured entities to unstructured documents) being generated on a continuous basis ?
- **What technology to use for persisting the data (from Relational to NoSQL databases)?**



RDBMS vs NoSQL

Scale Up — Scale Out

Structured Data — Semi-/Un-Structured Data

Atomic Transactions — Eventual Consistency

Impedance Mismatch — Relaxed..

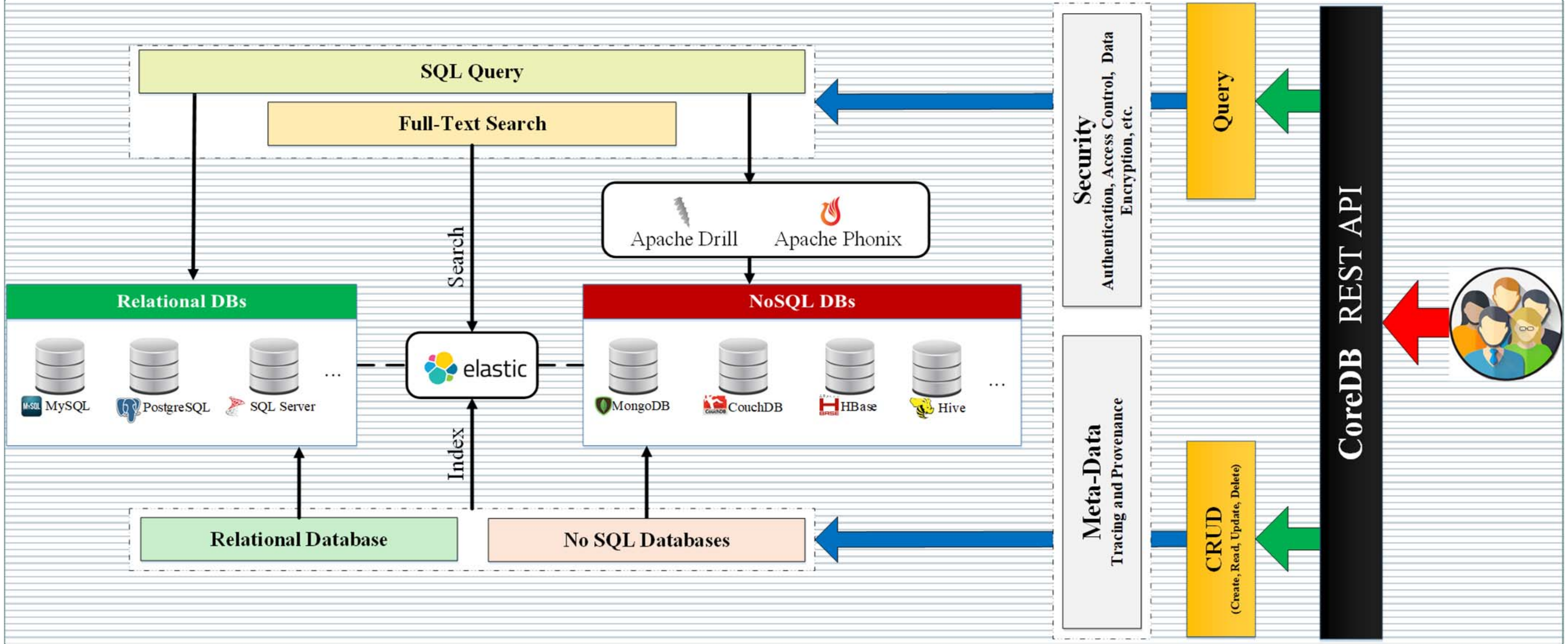
- Key–value Store
- Document Store
- Graph Database
- …

SQL Server
PostgreSQL
MySQL
…
MongoDB
CouchDB
HBase
…
a
c
i
d
Atomicity
Consistency
Isolation: Transactions do not affect each other
Durability: Written data will not be lost

Data Lake

Organizing Big data

A **Data Lake** is a storage repository that holds a vast amount of raw **data** in its native format, including structured, semi-structured, and unstructured **data**.



Beheshti et al., **CoreDB: a Data Lake Service**, https://github.com/unsw-cse-soc/CoreDB

Curating Big Data

Curating Big data

Data Curation is the process of transforming raw data into **Curated Data**.

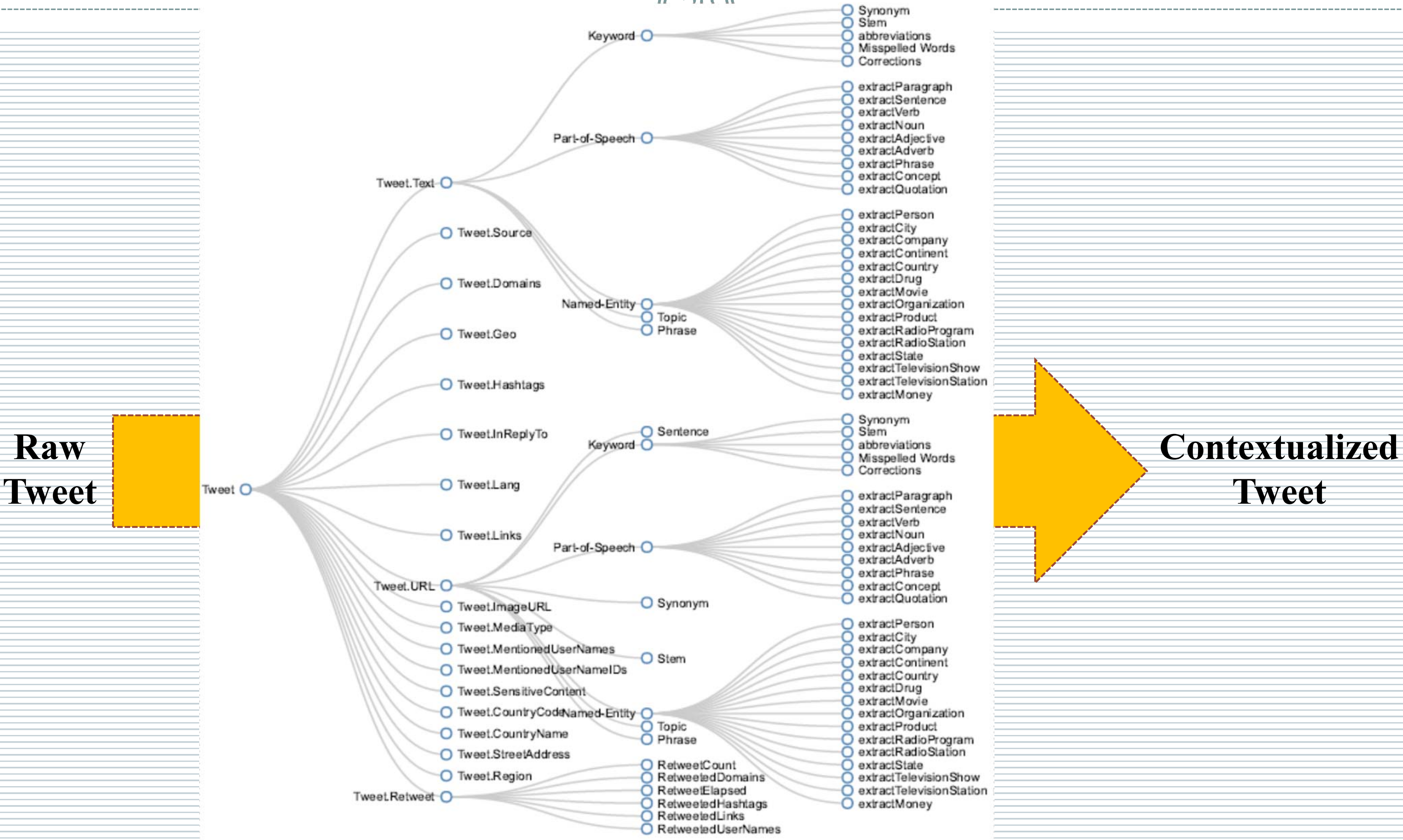
Curated Data is the **Contextualized data and knowledge** that is maintained and made available for use by end-users and applications.

Data curation involves identifying relevant data sources, **extracting** data and knowledge, **cleaning**, **maintaining**, **merging**, **enriching** and **linking** data and knowledge.

Beheshti et al., "On Automating Basic Data Curation Tasks". WWW 2017. **https://github.com/unsw-cse-soc/Data-curation-API**

Example

Curating Big data

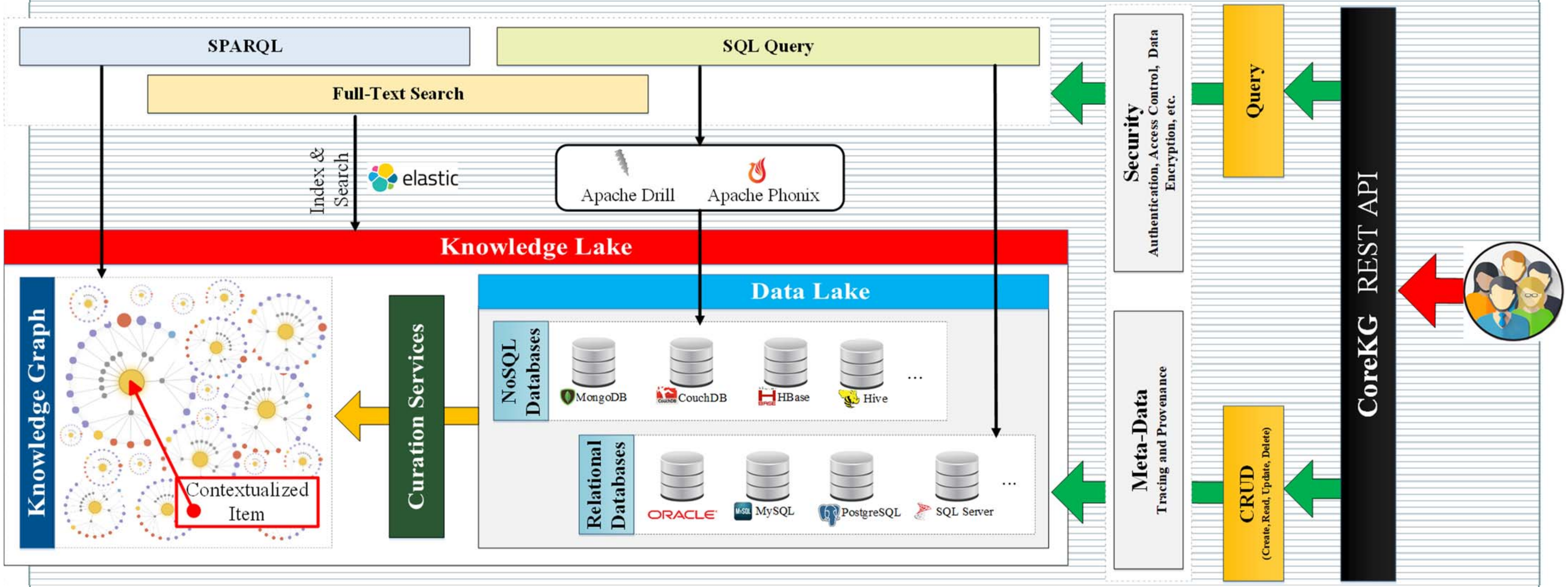


Beheshti et al., “On Automating Basic Data Curation Tasks”. WWW 2017. **https://github.com/unsw-cse-soc/Data-curation-API**

Knowledge Lake

Curating Big data

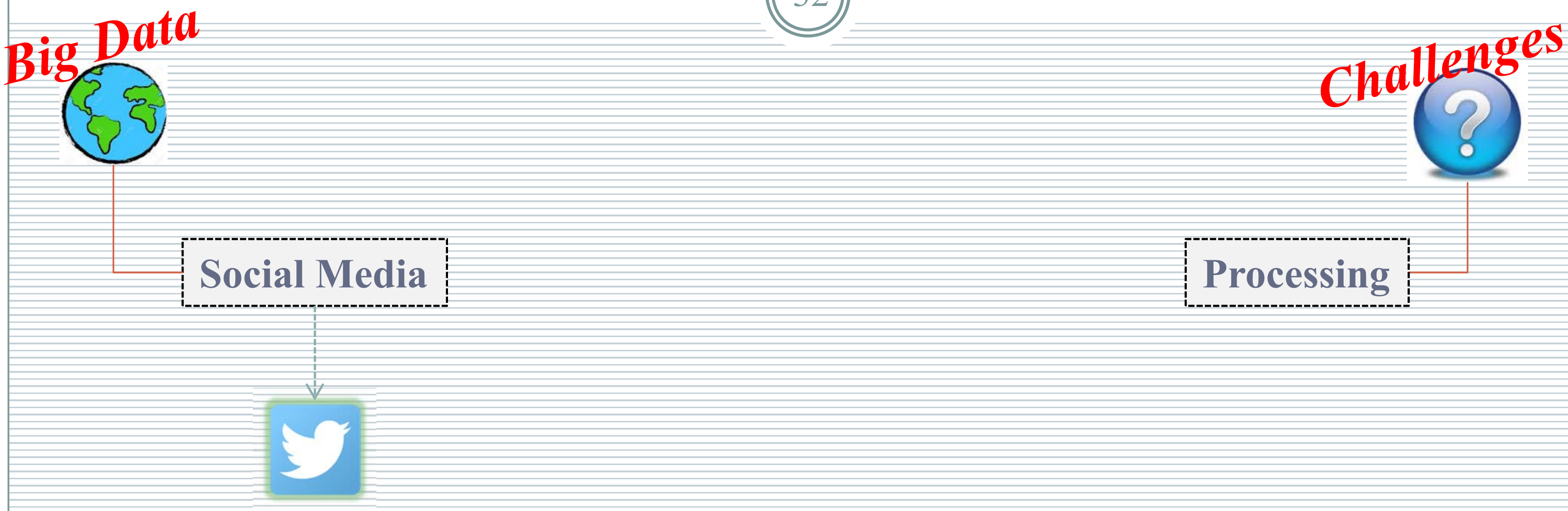
A **Knowledge Lake**, i.e. a contextualized Data Lake, is a centralized repository containing virtually inexhaustible amounts of both data and contextualized data that is readily made available to perform analytical activities.



Beheshti et al., **CoreKG: a Knowledge Lake Service (VLDB'18)**, https://github.com/unsw-cse-soc/CoreKG

Processing Big Data

Processing Big Data



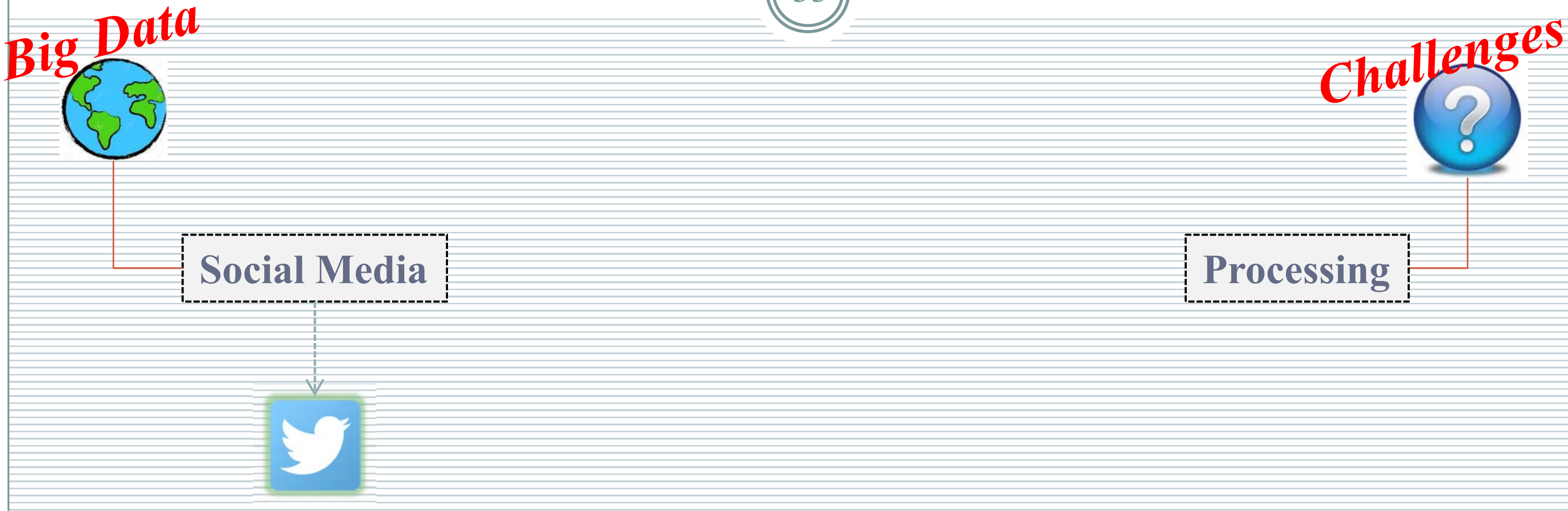
Big Data

Social Media

Challenges

Processing

Processing Big Data

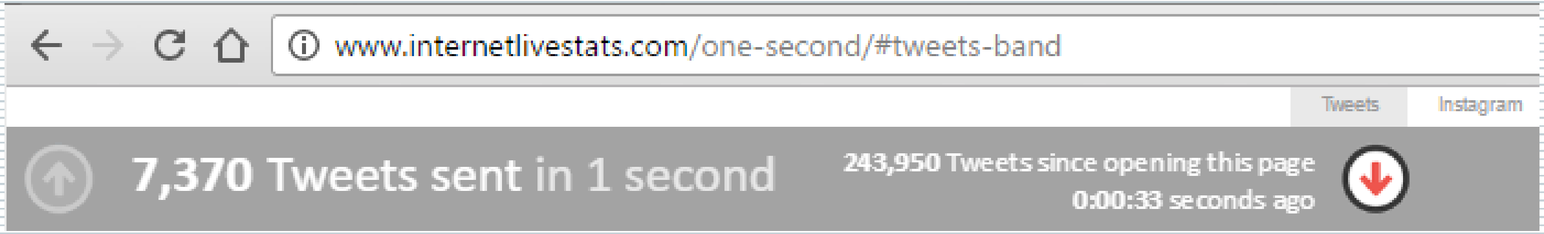


Big Data

Social Media

Challenges

Processing



www.internetlivestats.com/one-second/#tweets-band

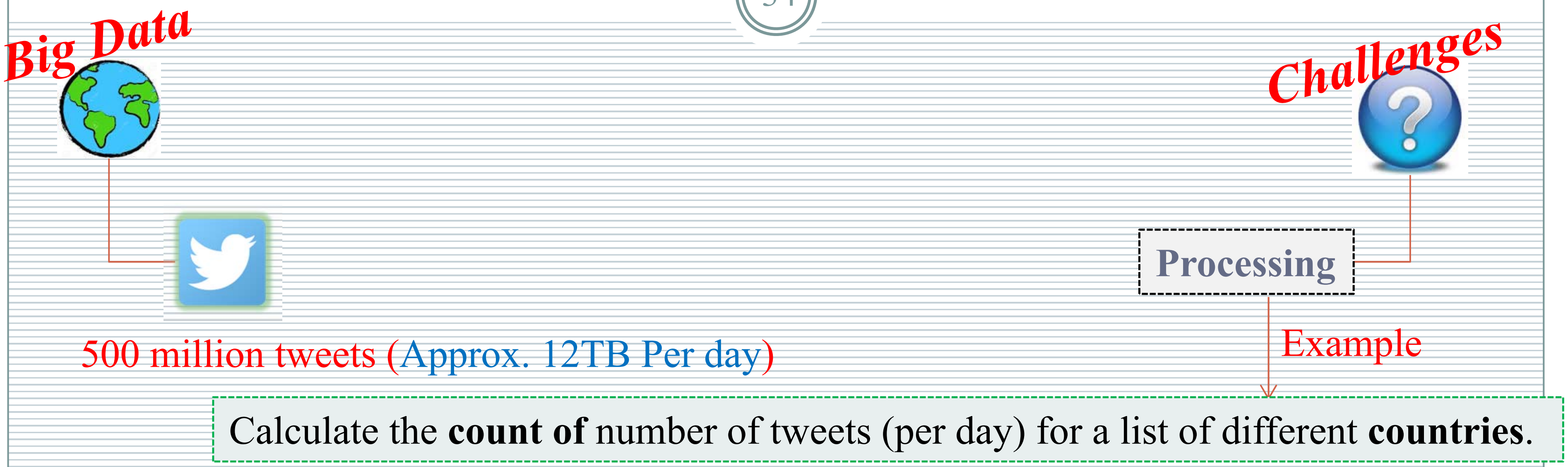
7,370 Tweets sent in 1 second

243,950 Tweets since opening this page
0:00:33 seconds ago

Average one second: 6,000 tweets

Average one Day: 500 million tweets (Approx. 12TB Per day)

Processing Big Data



Big Data

500 million tweets (Approx. 12TB Per day)

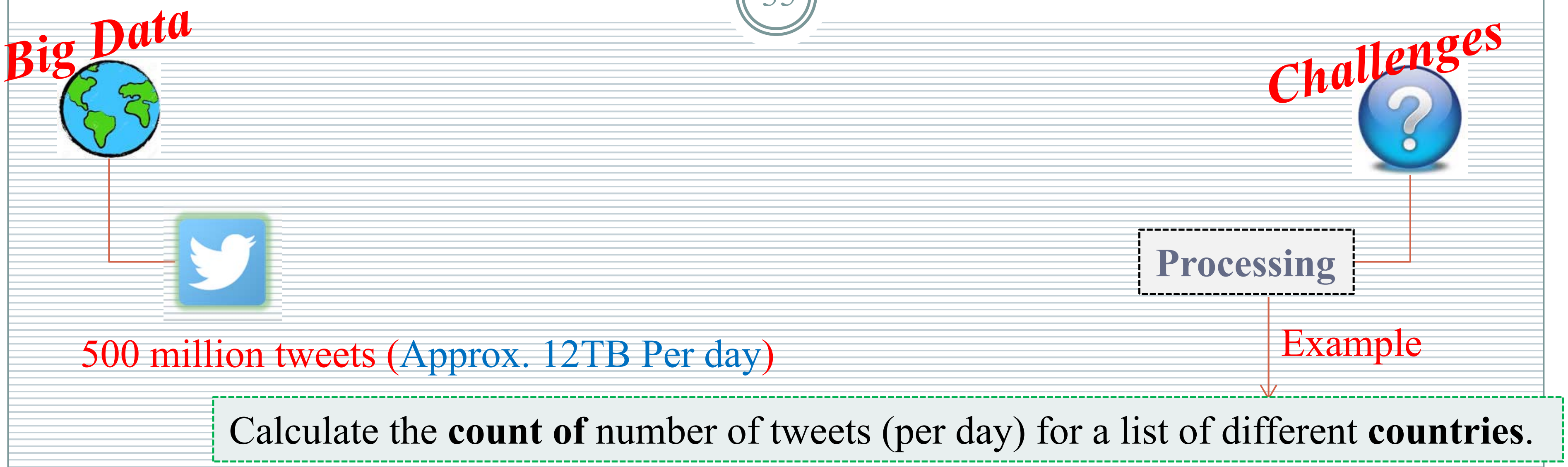
Challenges

Processing

Example

Calculate the **count of** number of tweets (per day) for a list of different **countries**.

Processing Big Data



Big Data

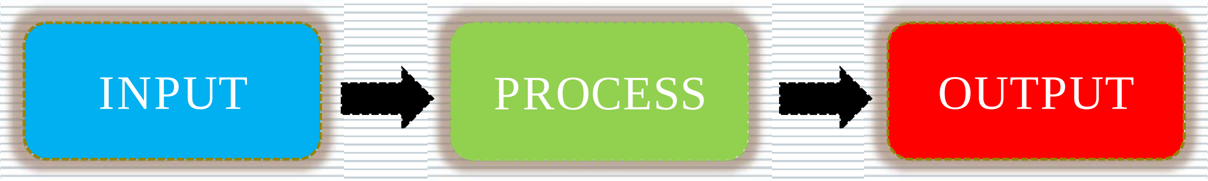
500 million tweets (Approx. 12TB Per day)

Challenges

Processing

Example

Calculate the **count of** number of tweets (per day) for a list of different **countries**.



INPUT → PROCESS → OUTPUT

IPO Model:

An approach in **software engineering** for describing the structure of an **information processing program.**

https://en.wikipedia.org/wiki/IPO_model

Processing Big Data

Big Data

Challenges

Processing

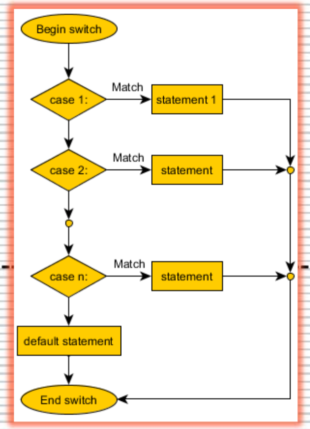
500 million tweets (Approx. 12TB Per day)

Example

Calculate the **count of** number of tweets (per day) for a list of different **countries**.

Traditional Enterprise Systems normally have a centralized server to store and process data.

INPUT ------ Tweet ---- Tweet ---- Tweet ---- Tweet ------>



------> OUTPUT

Process

Processing Big Data

Big Data

500 million tweets (Approx. 12TB Per day)

Challenges

Processing

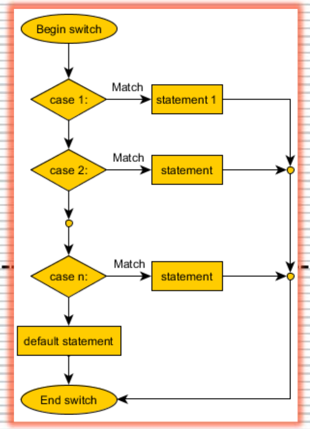
Example

Calculate the **count of** number of tweets (per day) for a list of different **countries**.

1 Machine ..
500 million Tweets ..
How long will the process take?

INPUT

Tweet Tweet Tweet Tweet



OUTPUT

Process

Processing Big Data

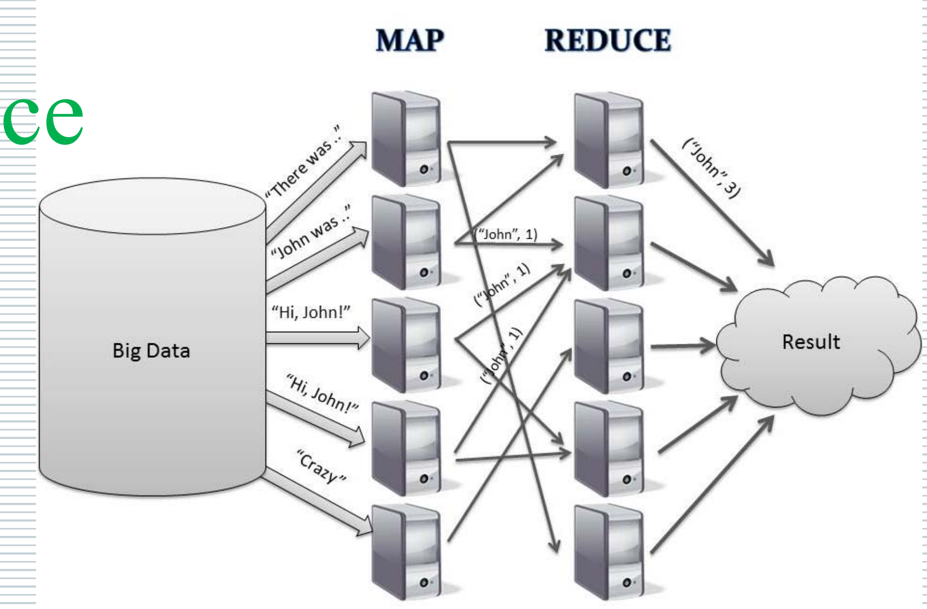
Big Data

Huge amount of data

Challenges

Processing

MapReduce



Processing Big Data:
Cloud computing

Processing Big Data

Big Data

Huge amount of data

Challenges

Processing

Apache Hadoop

Hadoop is an open source framework that uses a simple programming model to enable distributed processing of large data sets on clusters of computers.

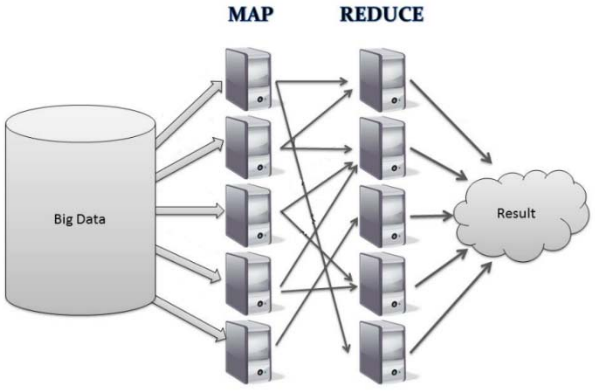
Apache Hadoop solution:

- Distributed File System (HDFS)
- **MapReduce**
- Pig
- HCatalog

Apache Spark!

Who Use Hadoop?

- Amazon
- Facebook
- Google
- IBM
- New York Times
- Yahoo!
- …



http://hadoop.apache.org/

Processing Big Data

Big Data

Huge amount of data

Challenges

Processing

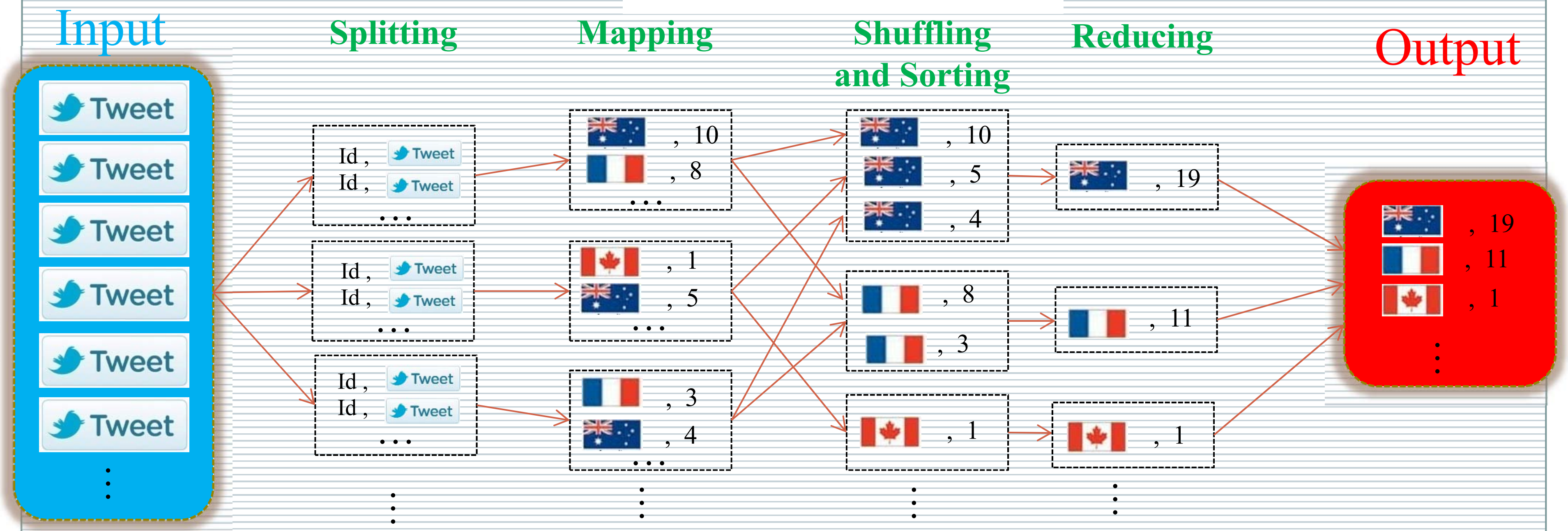
MapReduce

- Is a software framework for **Processing Large Datasets**.
- Provides **Scalability** in a *distributed* fashion over *several machines*.
- Divides the Input into small parts and **MAP** them to many machines.
- Collects the results from each machine and **REDUCE** them to form the Output.

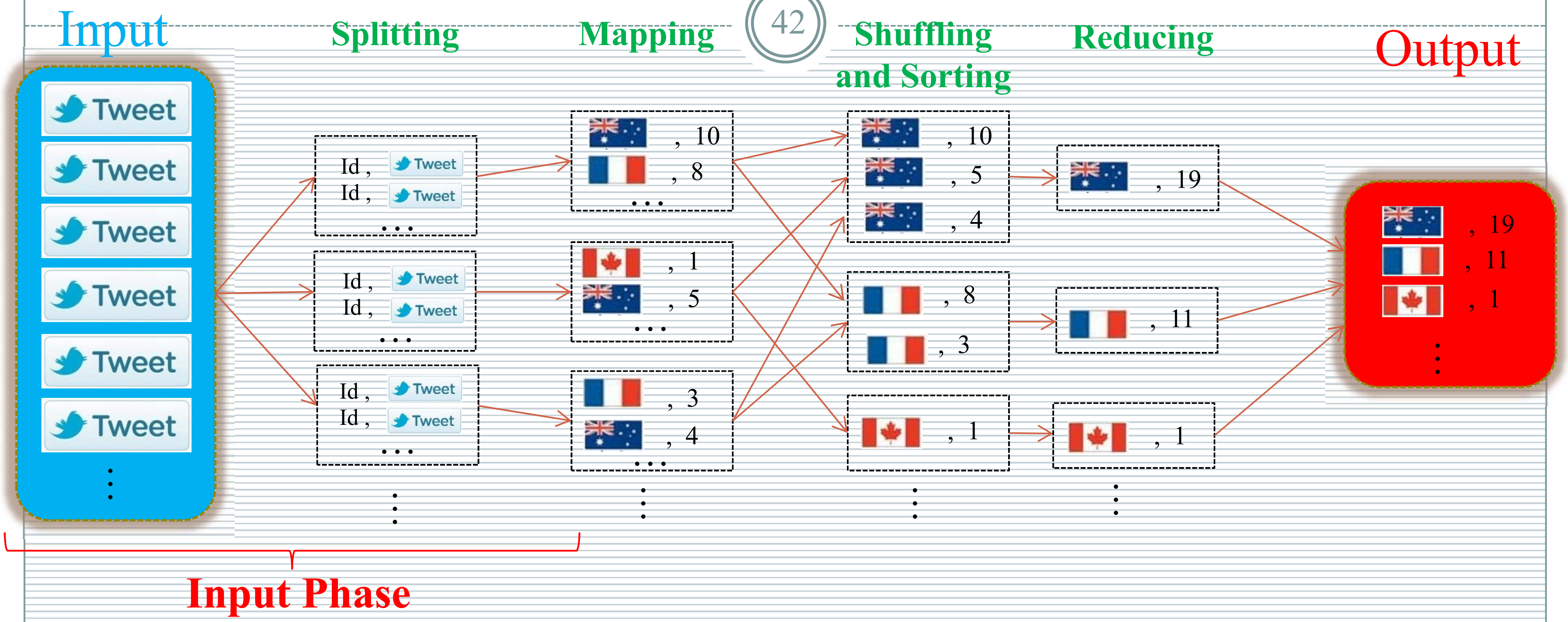
Processing Big Data

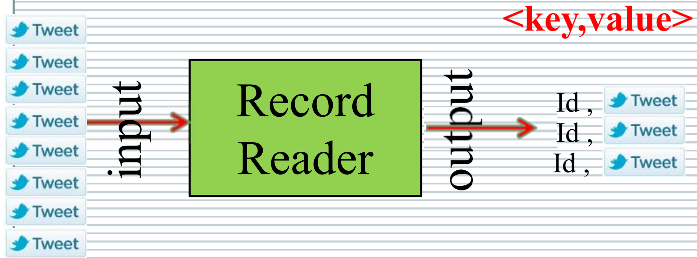
Big Data
Huge amount of data
Solutions
Challenges
Processing

Calculate the **count of** number of tweets (per day) for a list of different **countries**.

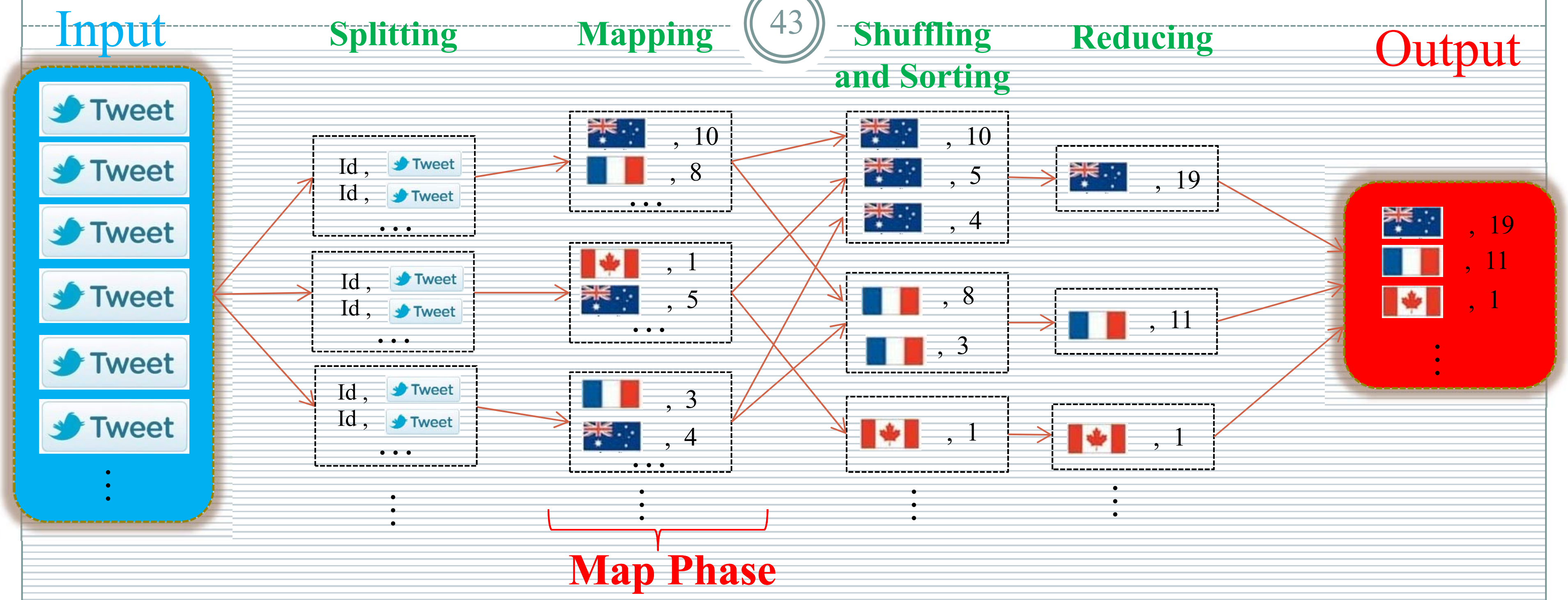


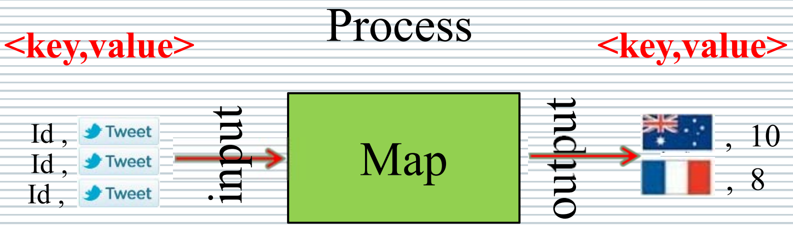
Processing Big Data





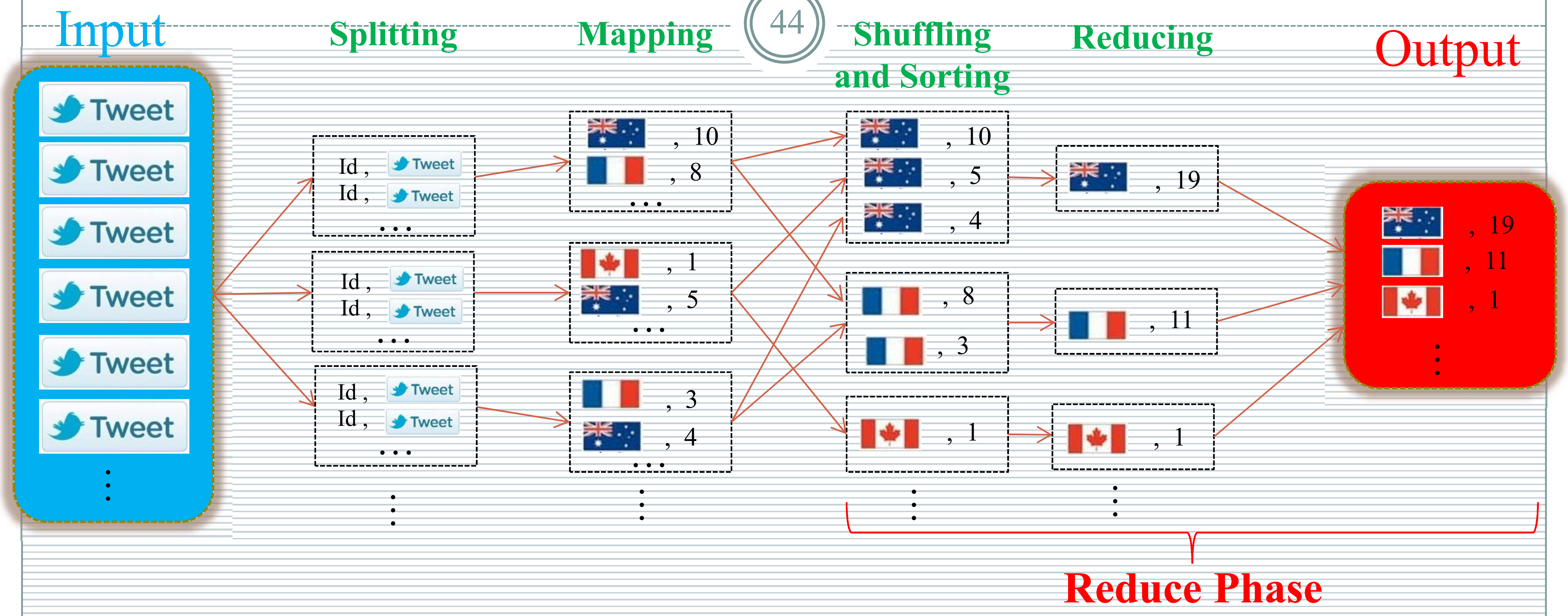
Processing Big Data





Map is a user-defined **function**.

Processing Big Data



Shuffling:

- is the process of transferring data from the mappers to the reducers, using HTTP
- it can start even before the map phase has finished, to save some time.

Sorting:

- Sorting saves time for the reducer…

Example

Processing Big Data

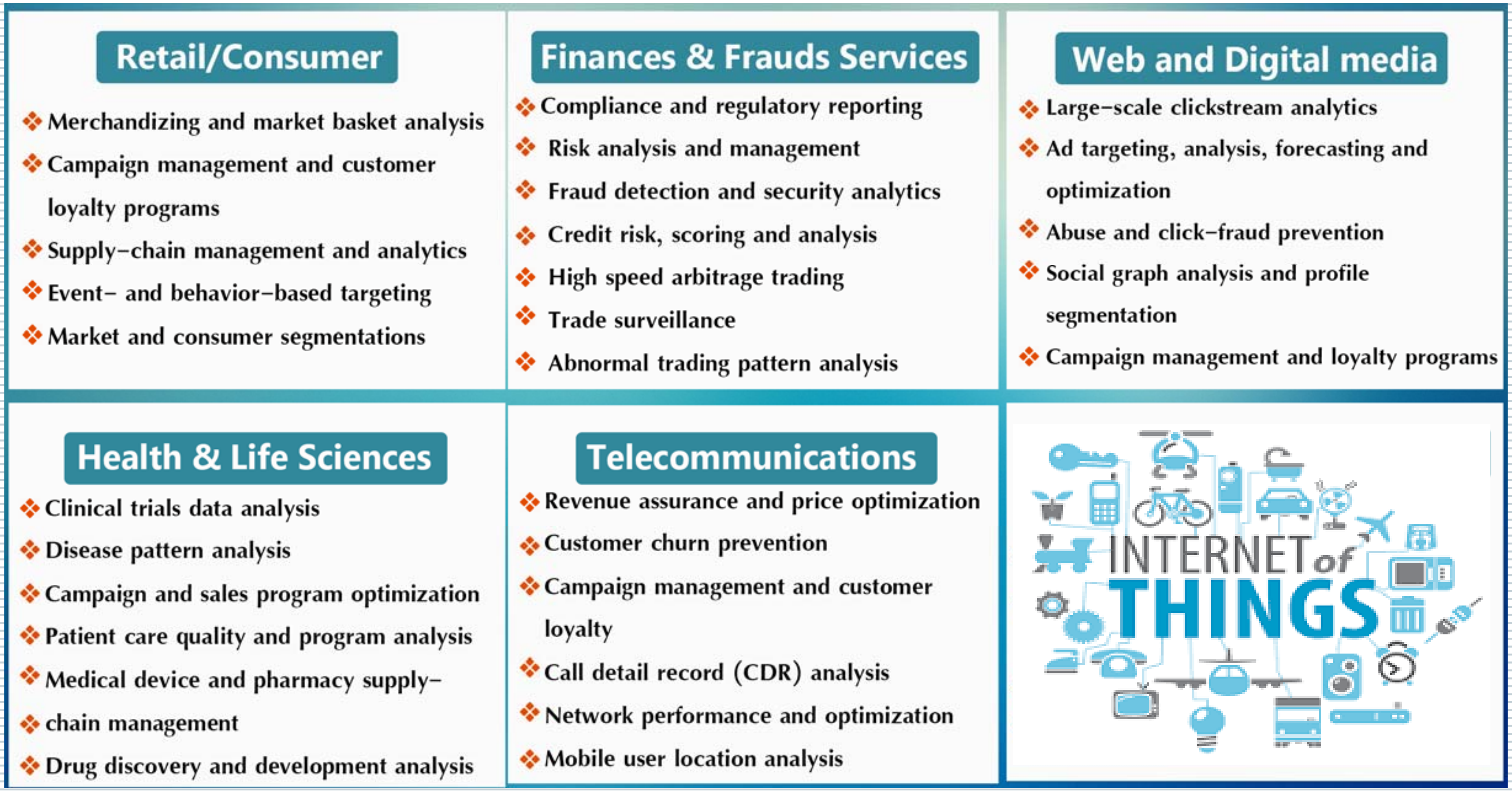
MapReduce,
Example Program

```
package hadoop;
import java.io..
import org.apache.hadoop..

public class ProcessUnits
{
   //Mapper Class
     public static class E_Emapper..
      {
          //Map function
      }

   //Reducer Class
     public static class E_Emapper..
      {
          //Reduce function
      }

   //Main function
     public static void main(String args[])throws Exception
      {
          //Create and Run the job
      }
}
```

Big Data Analytics

Big Data Analytics

Analytics is used to gain insights from data in order to make better decisions, using mathematical or scientific methods.



Retail/Consumer

- Merchandizing and market basket analysis
- Campaign management and customer loyalty programs
- Supply–chain management and analytics
- Event– and behavior–based targeting
- Market and consumer segmentations

Finances & Frauds Services

- Compliance and regulatory reporting
- Risk analysis and management
- Fraud detection and security analytics
- Credit risk, scoring and analysis
- High speed arbitrage trading
- Trade surveillance
- Abnormal trading pattern analysis

Web and Digital media

- Large–scale clickstream analytics
- Ad targeting, analysis, forecasting and optimization
- Abuse and click–fraud prevention
- Social graph analysis and profile segmentation
- Campaign management and loyalty programs

Health & Life Sciences

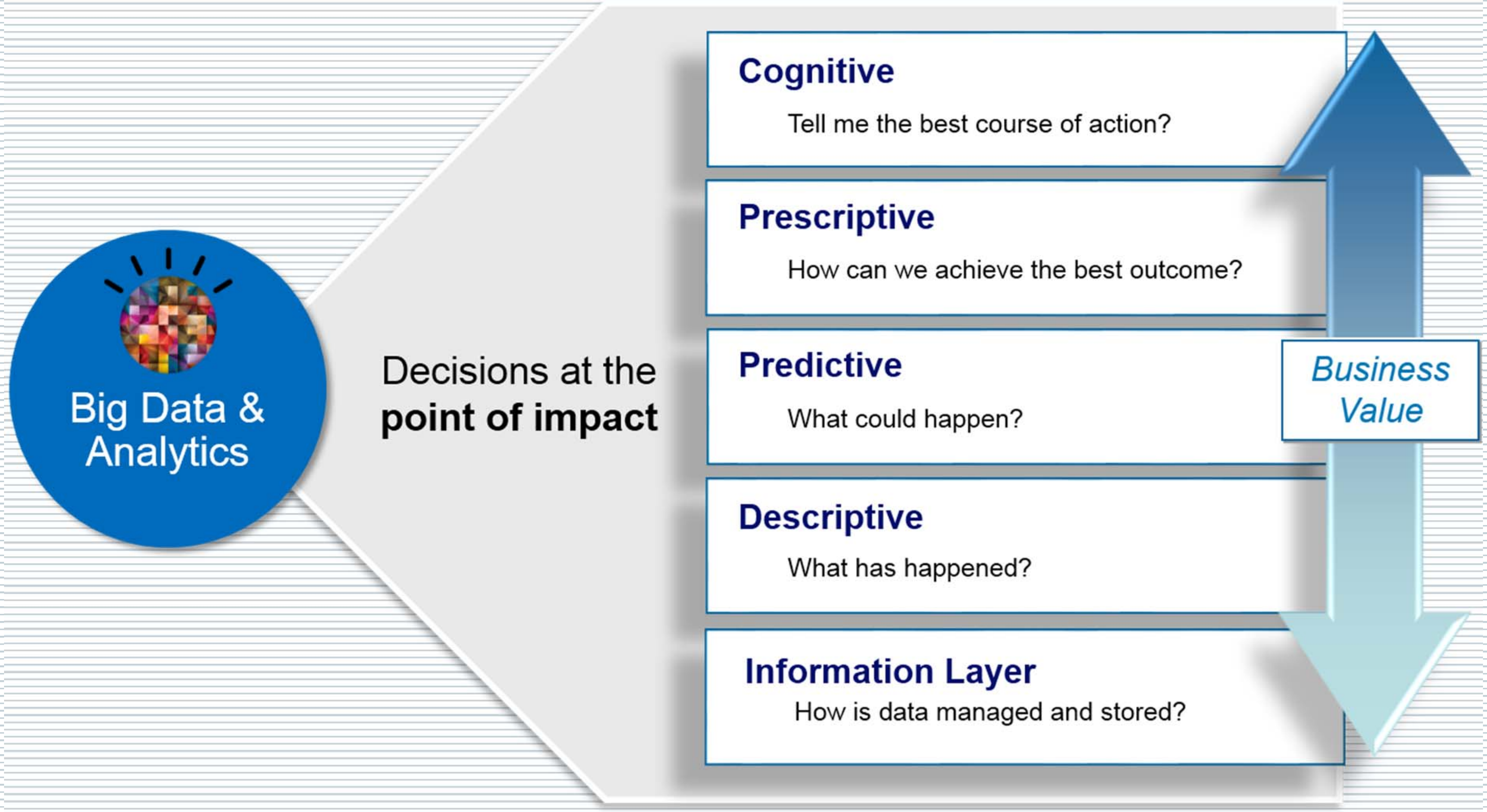
- Clinical trials data analysis
- Disease pattern analysis
- Campaign and sales program optimization
- Patient care quality and program analysis
- Medical device and pharmacy supply–
- chain management
- Drug discovery and development analysis

Telecommunications

- Revenue assurance and price optimization
- Customer churn prevention
- Campaign management and customer loyalty
- Call detail record (CDR) analysis
- Network performance and optimization
- Mobile user location analysis

https://www.greycampus.com/opencampus/big-data-developer/applications-of-big-data

Big Data Analytics



Big Data & Analytics

Decisions at the **point of impact**

Cognitive
Tell me the best course of action?

Prescriptive
How can we achieve the best outcome?

Predictive
What could happen?

Descriptive
What has happened?

Information Layer
How is data managed and stored?

Business Value

https://www.ibm.com/power/solutions/bigdata-analytics

Example

Big Data Analytics

OLAP, is an approach to answering multi-dimensional analytical queries swiftly.

Problem:

- extension of existing OLAP techniques to analysis of graphs is not straightforward.
- key business insights remain hidden in the interactions among objects.

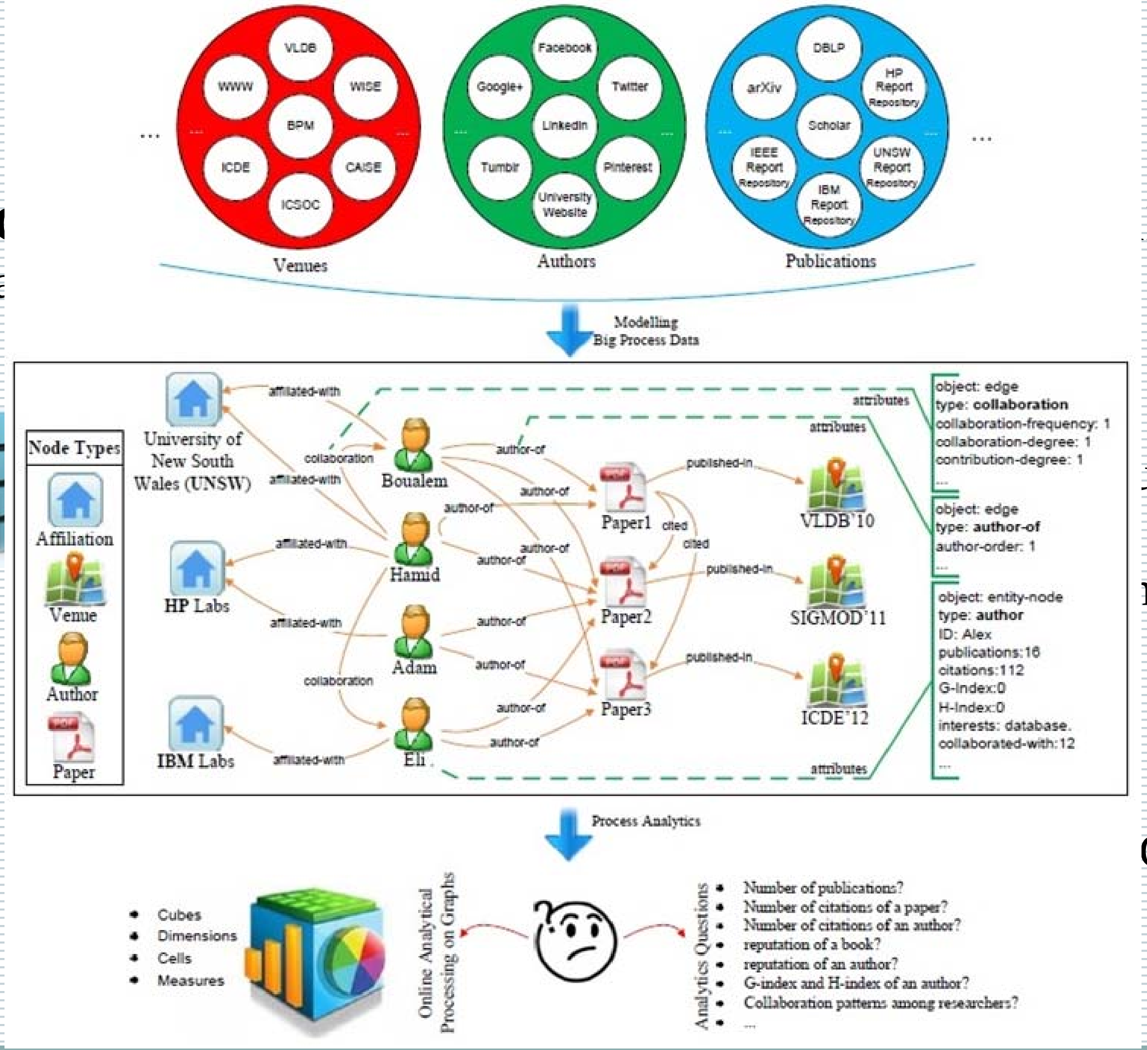
Solution:

- On-Line Analytical Processing on Graphs

Beheshti et al., “Scalable Graph-based OLAP Analytics over Process Execution Data”, **DAPD** Journal (2016).

Example

Big Data Analytics



C

l

a

ıes to analysis

ı in the

Graphs

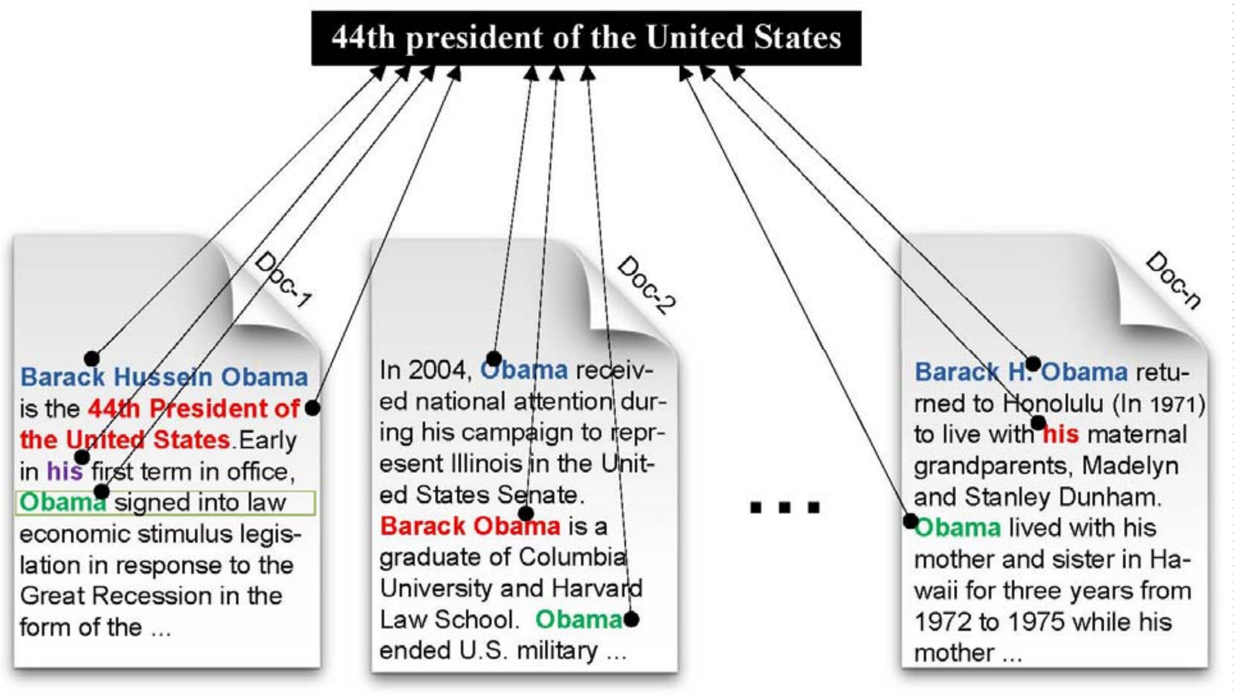
Beheshti et al., "Scalable Graph-based OLAP Analytics over Process Execution Data", **DAPD** Journal (2016).

Example

Big Data Analytics

Big Data Analytics benefits from:

- NLP and Machine Learning
  - Pattern recognition, Extraction, Classification, Enrichment, Linking, Similarity, etc.



Barack Hussein Obama is the 44th President of the United States.Early in his first term in office, Obama signed into law economic stimulus legislation in response to the Great Recession in the form of the ...

In 2004, Obama received national attention during his campaign to represent Illinois in the United States Senate. Barack Obama is a graduate of Columbia University and Harvard Law School. Obama ended U.S. military ...

Barack H. Obama returned to Honolulu (In 1971) to live with his maternal grandparents, Madelyn and Stanley Dunham. Obama lived with his mother and sister in Hawaii for three years from 1972 to 1975 while his mother ...

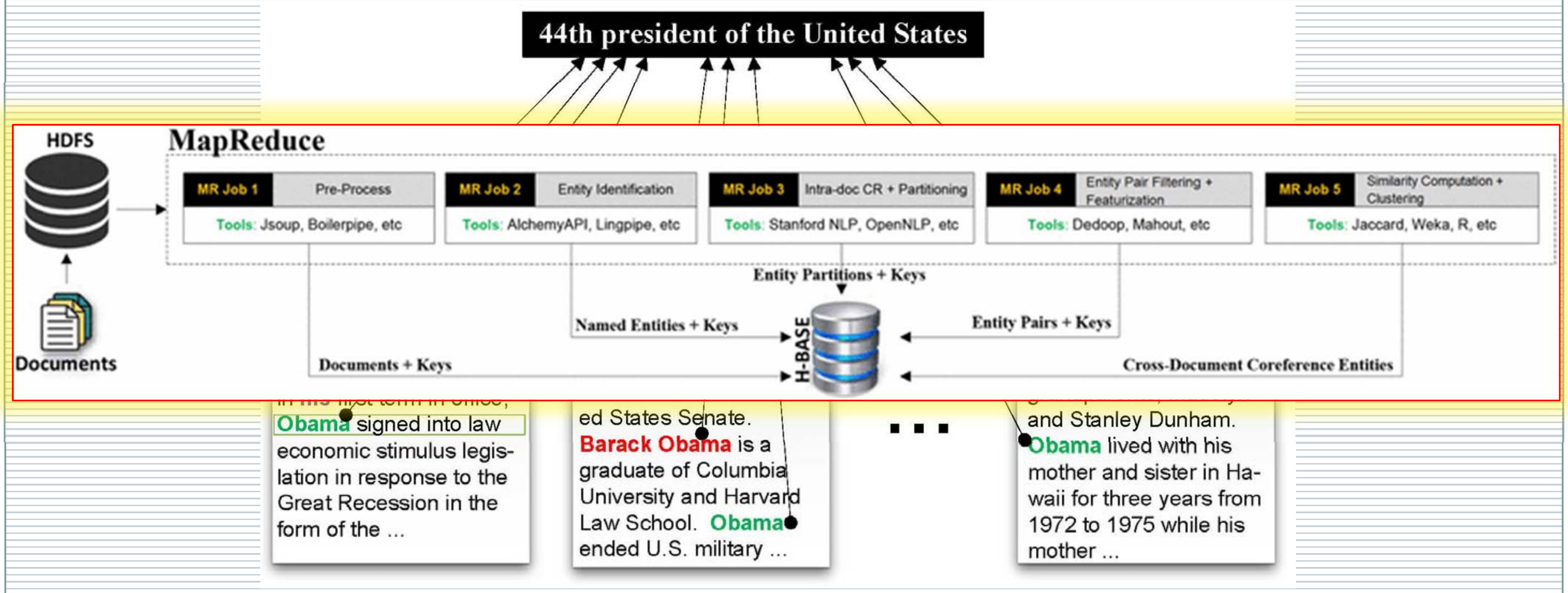
Beheshti et al., “A Systematic Review and Comparative Analysis of Cross-Document Coreference Resolution Methods and Tools”, Computing Journal, 2017.

Example

Big Data Analytics

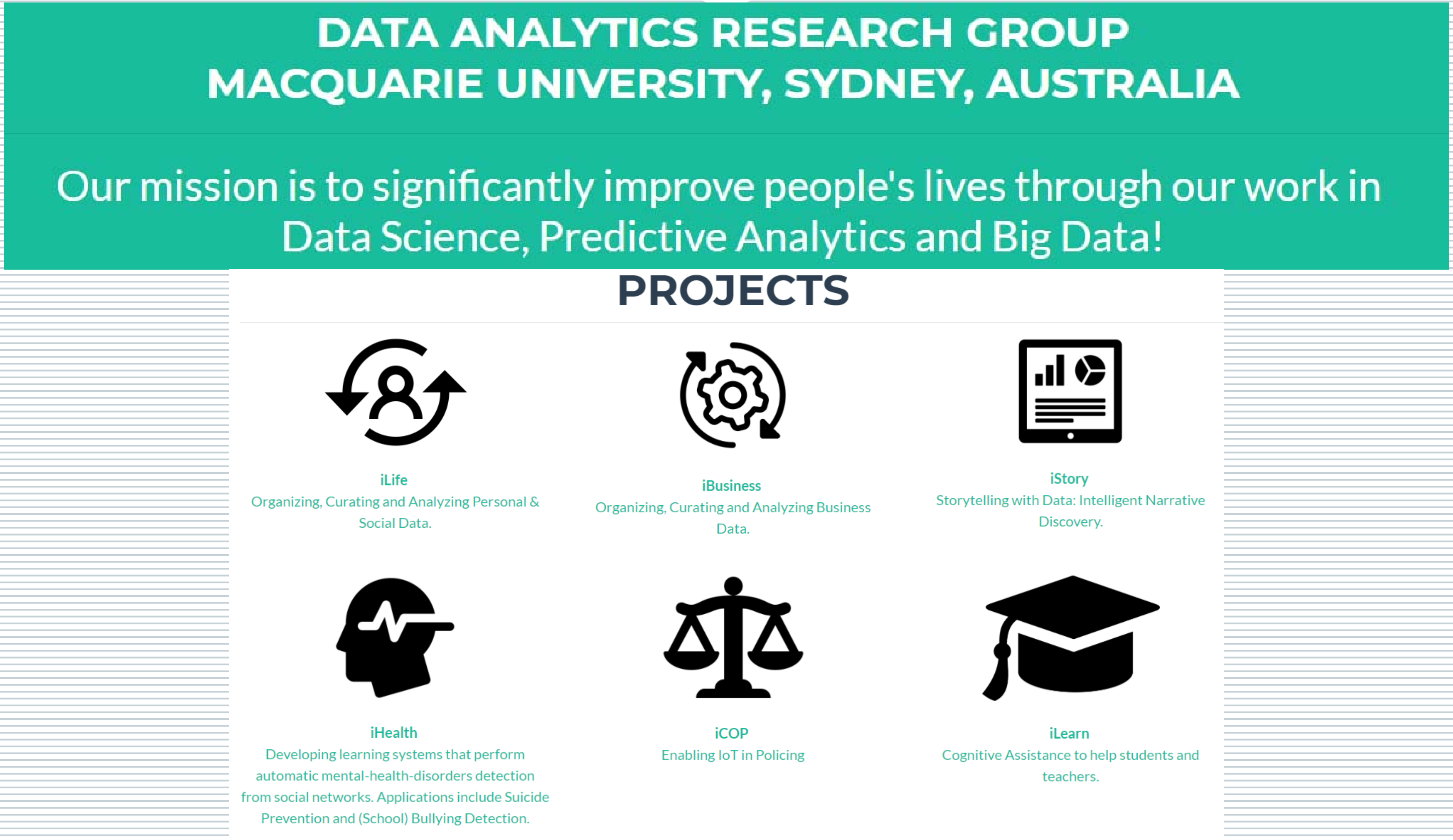
Big Data Analytics benefits from:

- NLP and Machine Learning
  - Pattern recognition, Extraction, Classification, Enrichment, Linking, Similarity, etc.


44th president of the United States
HDFS
MapReduce
MR Job 1
Pre-Process
Tools: Jsoup, Boilerpipe, etc
MR Job 2
Entity Identification
Tools: AlchemyAPI, Lingpipe, etc
MR Job 3
Intra-doc CR + Partitioning
Tools: Stanford NLP, OpenNLP, etc
MR Job 4
Entity Pair Filtering + Featurization
Tools: Dedoop, Mahout, etc
MR Job 5
Similarity Computation + Clustering
Tools: Jaccard, Weka, R, etc
Entity Partitions + Keys
Named Entities + Keys
Entity Pairs + Keys
H-BASE
Documents
Documents + Keys
Cross-Document Coreference Entities
Obama signed into law economic stimulus legislation in response to the Great Recession in the form of the ...
ed States Senate. Barack Obama is a graduate of Columbia University and Harvard Law School. Obama ended U.S. military ...
...
and Stanley Dunham. Obama lived with his mother and sister in Hawaii for three years from 1972 to 1975 while his mother ...

Beheshti et al., “A Systematic Review and Comparative Analysis of Cross-Document Coreference Resolution Methods and Tools”, Computing Journal, 2017.

Big Data Applications



DATA ANALYTICS RESEARCH GROUP
MACQUARIE UNIVERSITY, SYDNEY, AUSTRALIA

Our mission is to significantly improve people's lives through our work in Data Science, Predictive Analytics and Big Data!

PROJECTS

iLife
Organizing, Curating and Analyzing Personal & Social Data.

iBusiness
Organizing, Curating and Analyzing Business Data.

iStory
Storytelling with Data: Intelligent Narrative Discovery.

iHealth
Developing learning systems that perform automatic mental-health-disorders detection from social networks. Applications include Suicide Prevention and (School) Bullying Detection.

iCOP
Enabling IoT in Policing

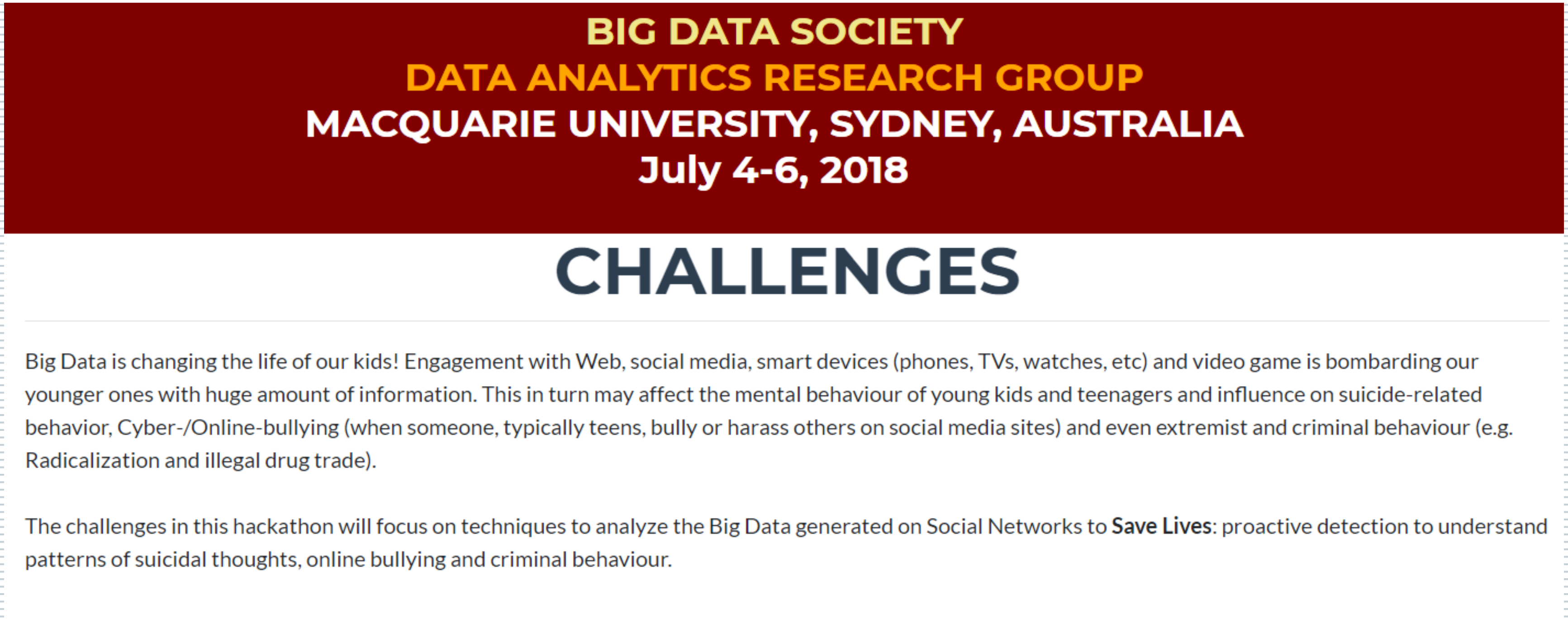
iLearn
Cognitive Assistance to help students and teachers.

https://data-science-group.github.io/

Big Data Applications

BigDataSOC:
https://data-science-group.github.io/BigDataSociety/

Hackathon:
https://data-science-group.github.io/BigDataSociety/Hackathon/2018-07/index.html



BIG DATA SOCIETY
DATA ANALYTICS RESEARCH GROUP
MACQUARIE UNIVERSITY, SYDNEY, AUSTRALIA
July 4-6, 2018

CHALLENGES

Big Data is changing the life of our kids! Engagement with Web, social media, smart devices (phones, TVs, watches, etc) and video game is bombarding our younger ones with huge amount of information. This in turn may affect the mental behaviour of young kids and teenagers and influence on suicide-related behavior, Cyber-/Online-bullying (when someone, typically teens, bully or harass others on social media sites) and even extremist and criminal behaviour (e.g. Radicalization and illegal drug trade).

The challenges in this hackathon will focus on techniques to analyze the Big Data generated on Social Networks to **Save Lives**: proactive detection to understand patterns of suicidal thoughts, online bullying and criminal behaviour.

Big Data: Opportunities

- **Varieties of Data**
  - Text
  - Social Media
  - Networks
  - Multimedia
  - Machine Data
  - Sensors
- **Curation**
  - Include tasks for data creation, maintenance, and management, together with the capacity to add value to data (e.g. extraction. Enrichment, linking, etc)
- **Integration**
  - Integrating enterprise/public data
  - Linked Data and Knowledge Graphs

- **Big Data Performance**
  - In memory
  - New Benchmarks and Architecture
- **Analytics**
  - Summarizing
  - Querying
  - Analyzing
  - Data Mining
  - Machine Learning
  - Deep Learning
  - Cognitive Computing
- **User Experience**
  - Cognitive Assistants
    - Automation and intelligent guidance
  - Visualizing with Analytics
  - Interacting with Analytics
  - Storytelling

Book: Beheshti et al. , "**Process Analytics**: Concepts and techniques for querying and analysing big process data", Springer, 2016.

Summary

- Why Big Data is different from **past** Very Large Datasets? Metadata, Potentially related Data Islands…
- Having the ability to analyse Big Data is of limited value if users cannot understand the analysis.
- How can the industry and academia collaborate towards solving Big Data challenges!!
- What is big today maybe not be big tomorrow!
- **COMP336 – Big Data**
  - http://unitguides.mq.edu.au/unit_offerings/88983/unit_guide

Questions ?