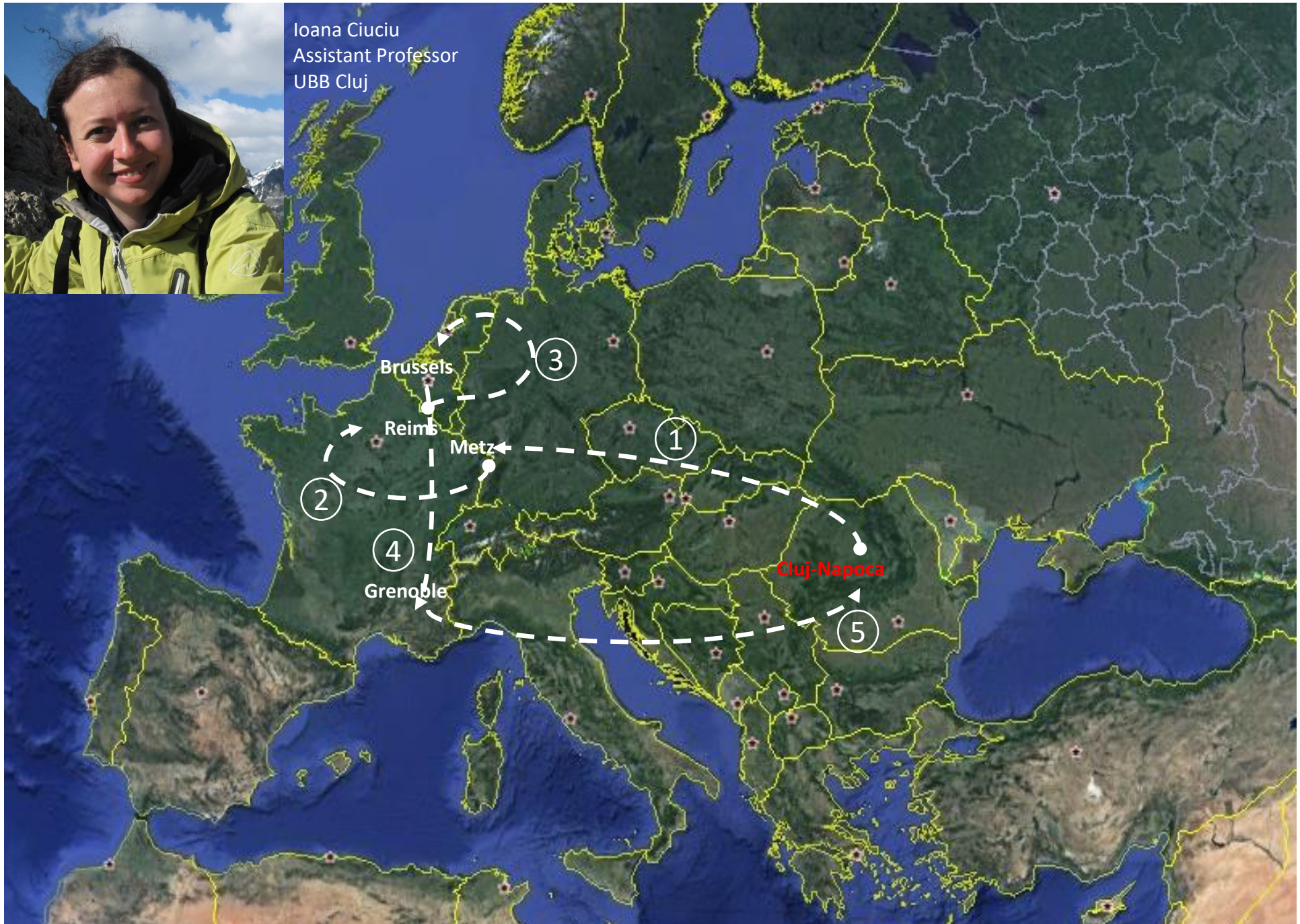
Big Data Processing and Applications

Lecture 1-Introduction to Data Science and Big Data

Ioana Ciuciu
ioana.ciuciu@ubbcluj.ro

Babes-Bolyai University, Cluj-Napoca

03 Oct. 2025


Ioana Ciuciu
Assistant Professor
UBB Cluj
Brussels
Reims
Metz
Grenoble
Cluj-Napoca
1
2
3
4
5

Course structure

Date	Course/Week	Title
03.10.2025	1	Introduction to Data Science and Big Data (Part 1)
10.10.2025	2	Introduction to Data Science and Big Data (Part 2)
17.10.2025	3	Data systems and the lambda architecture for big data
	4	Industrial standards for data mining projects. Big data case studies from industry – invited lecture from Bosch
	5	Lambda architecture: batch layer
	6	Lambda architecture: serving layer
	7	Lambda architecture: speed layer
	8	NoSQL Solutions for Big Data – invited lecturer from UBB
	9	Data Ingestion
	10	Introduction to SPARK – invited lecture from Bosch
	11	Data visualization
	12	Presentation research essays
	13	Presentation research essays + Project Evaluation during Seminar
	14	Presentation research essays + Project Evaluation during Seminar

Slight modifications in the structure are possible

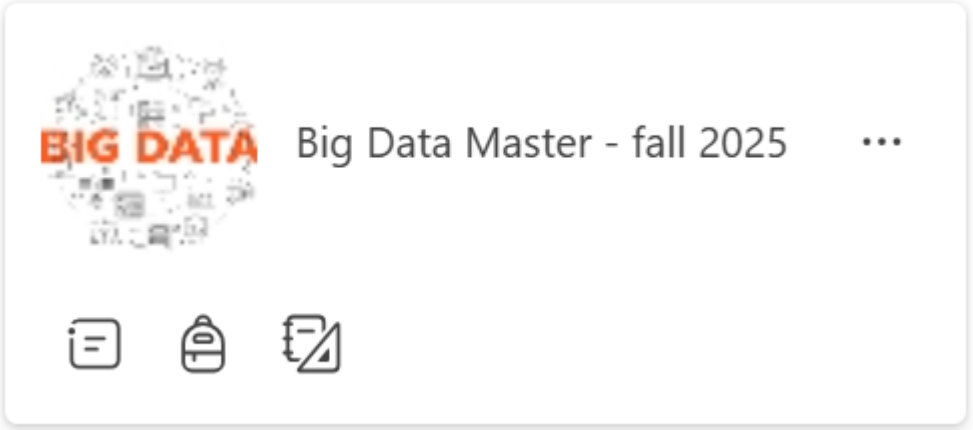
Semester Project

- Team-based (2-5 students with precise roles)
- Multidisciplinary: 3 CS Master Programes (HPC, SI, DS) + Master in Bioinformatics
- Real use cases: collaboration with local IT industry - TBA
- High degree of autonomy in selecting the topic and proposing the solution
- Implementation prototype
- Prototype demo & evaluation – student workshop (last seminar – weeks 13 & 14)
- *Best projects – disseminated in various events (TBD), scientifically disseminated (workshops/conferences/journals) AND/ OR possibility for a dissertation thesis*

Evaluation

- The final grade will be computed as follows:
  - 50% semester project (must be >= 5)
  - 50% research presentation or written exam (must be >= 5)

- Semester project
  - Details available on the course team
    - MS Team: **Big Data Master - fall 2025**
    - Access code: **j1exenq**



Agenda

Foundations of Data Science

- Introduction
- Definitions
- Types of Data
- Data Publishing
- The Data Science Process + CRISP-DM
- Data Science projects: examples
- Data Science Tools

Foundations of Big Data

- What is Big Data and how it is different from traditional data?
- What is big data analytics?
- Main characteristics of Big Data
- Big Data Producers
- Big Data ecosystem: main technological components and tools

Agenda

Foundations of Data Science

- Introduction
- Definitions
- Types of Data
- Data Publishing
- The Data Science Process + CRISP-DM
  Data Science projects: examples
- Data Science Tools

Foundations of Big Data

- What is Big Data and how it is different from traditional data?
- What is big data analytics?
- Main characteristics of Big Data
- Big Data Producers
- Big Data ecosystem: main technological components and tools

INTRODUCTION

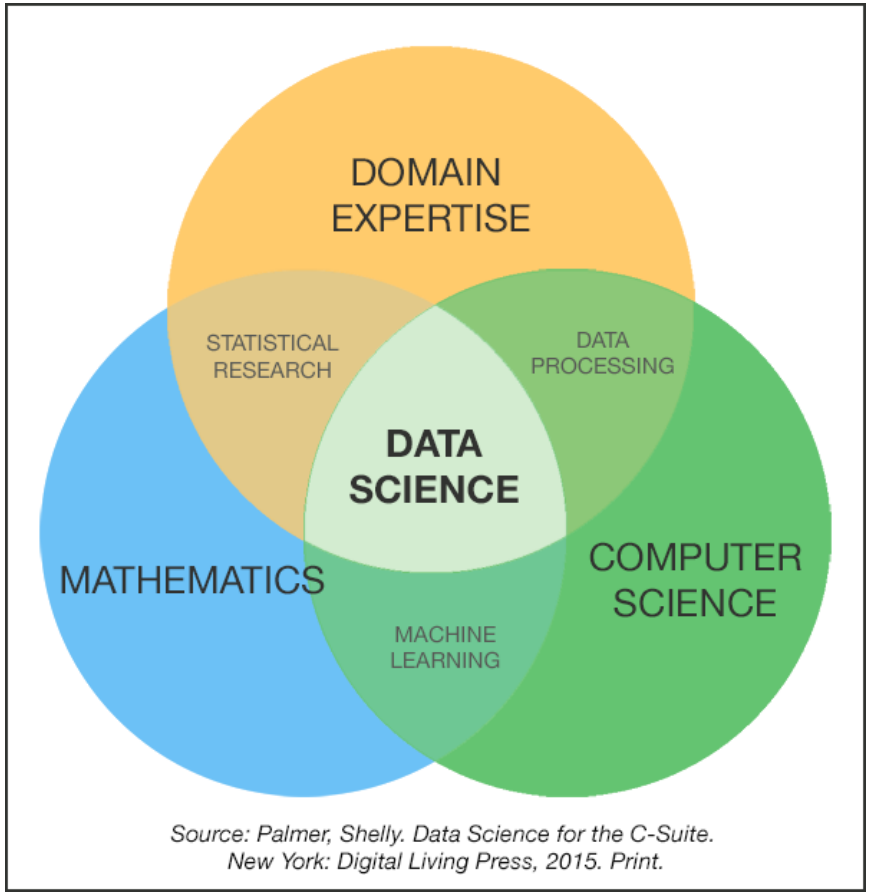
What is Data Science?

- In a Nutshell
  - It is an **interdisciplinary** field with various skills required
    - Computer Science
    - Software development
    - Mathematics and statistics
    - Machine learning
    - Domain specific knowledge
    - + Traditional research skills

- Detailed curriculum represented as a tube map, here:

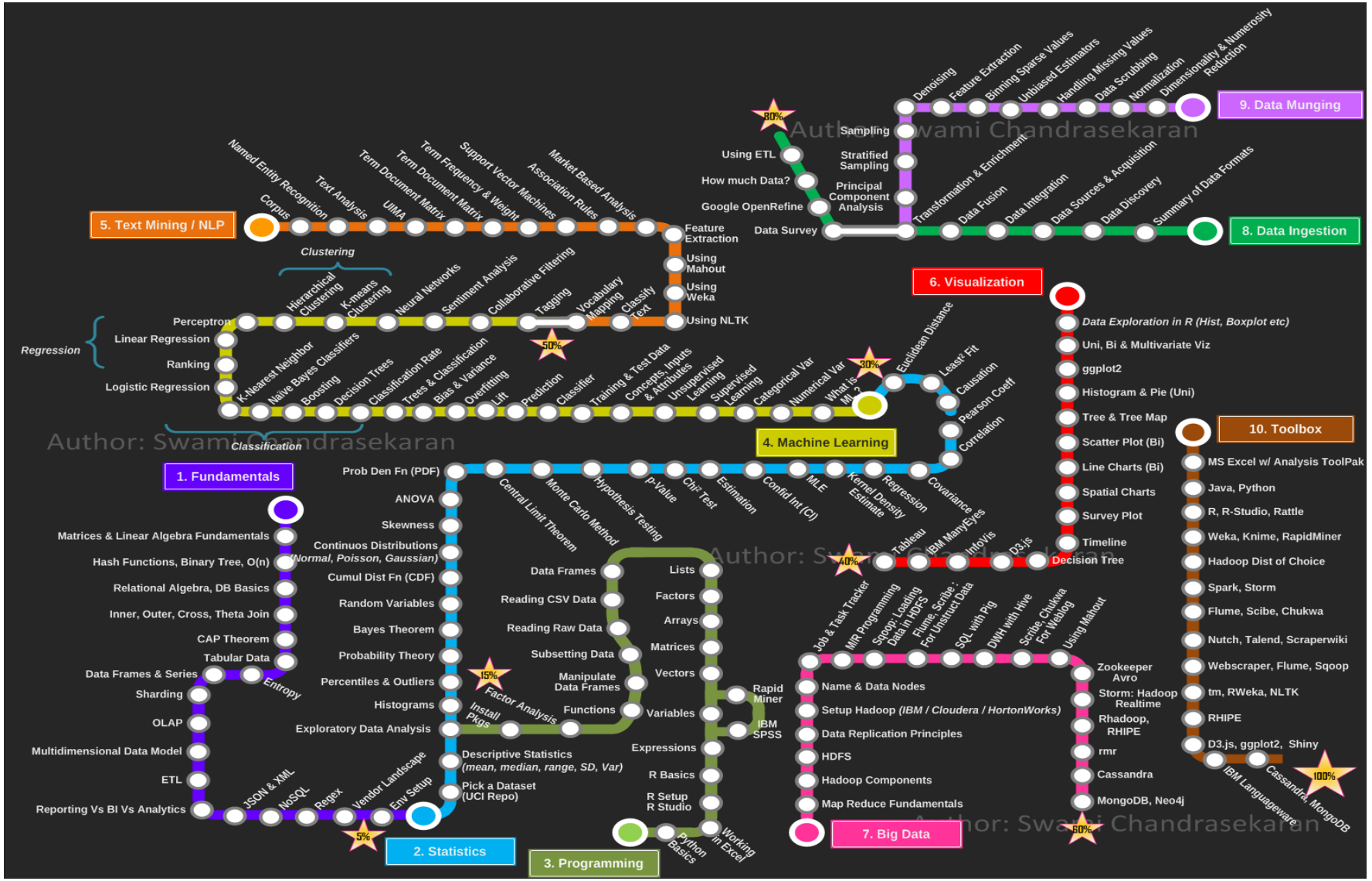
http://nirvacana.com/thoughts/becoming-a-data-scientist/

Data Science competencies



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

The road to data scientist



- Fundamentals
- Statistics
- Programming
- Machine learning
- Text mining
- Visualization
- Big data
- Data munging
- Toolbox

Data Scientist

- Definition by Josh Wills: *"Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician"*

(https://twitter.com/josh_wills/status/1980935121499586 56)

- Data Science roles:

Data Specialist (Data Engineer) + Data Analyst

Definitions

Data – what is data

Data is digital

Definition 1: *The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media* source https://www.lexico.com/definition/data

Definition 2: *In a more technical sense,* ***data*** *are a set of values of qualitative or quantitative variables about one or more persons or objects, while a datum (singular of* ***data****) is a single value of a single variable*

source Wikipedia

Data is not necessarily digital

Data definition: some examples

Lord of the rings	Return of the king	J.R.R. Tolkien	600	50	Yes

What do you consider as being data?

TYPES OF DATA

Types of Data

- Qualitative vs quantitative
- Discrete vs continuous
- Structured vs unstructured
- Open vs closed
- Static vs stream-like
- Big vs bigger

Qualitative vs Quantitative Data

- Qualitative data
  - Data that is non-numerical
  - Expressed in natural language
  - Types
    - **Binomial** – two **mutually exclusive** categories (e.g., heads or tails)
    - **Nominal** – no implicit order or rank to categories (e.g., favourite football team)
    - **Ordinal** – with categories that can be ordered (e.g., January, May, August)

- Quantitative data
  - Numerical data
  - Can be ordered
  - Discrete or continuous

Discrete vs Continuous Data

- Discrete data takes **specific values** from a finite or infinite set
  - Often integers, but can also include any real number which can be defined
  - Each is distinct and there is no grey area in between two discrete values
  - Can be numeric but also categorical (like male/female)
  - E.g., number of cars owned, average salary, etc.
- Continuous data may take **any value**
  - Real numbers
  - Between any two continuous data values there may be an infinite number of others
  - Always essentially numeric
  - E.g., height, weight, GDP
- The distinction is sometimes subject to **conventions** or application scenarios

Practice

- Identify the types of data in the following examples:

  - Number of cars owned
  - Favorite football team
  - Day of the week
  - Height (mm)
  - Height (short/medium/tall)

Courtesy of EDSA

Practice - Answers

- Identify the types of data in the following examples:

  - Number of cars owned – quantitative, discrete
  - Favorite football team – qualitative, nominal
  - Day of the week – qualitative, ordinal
  - Height (mm) – quantitative, continuous
  - Height (short/medium/tall) – qualitative, ordinal

Courtesy of EDSA

Structured vs Unstructured Data

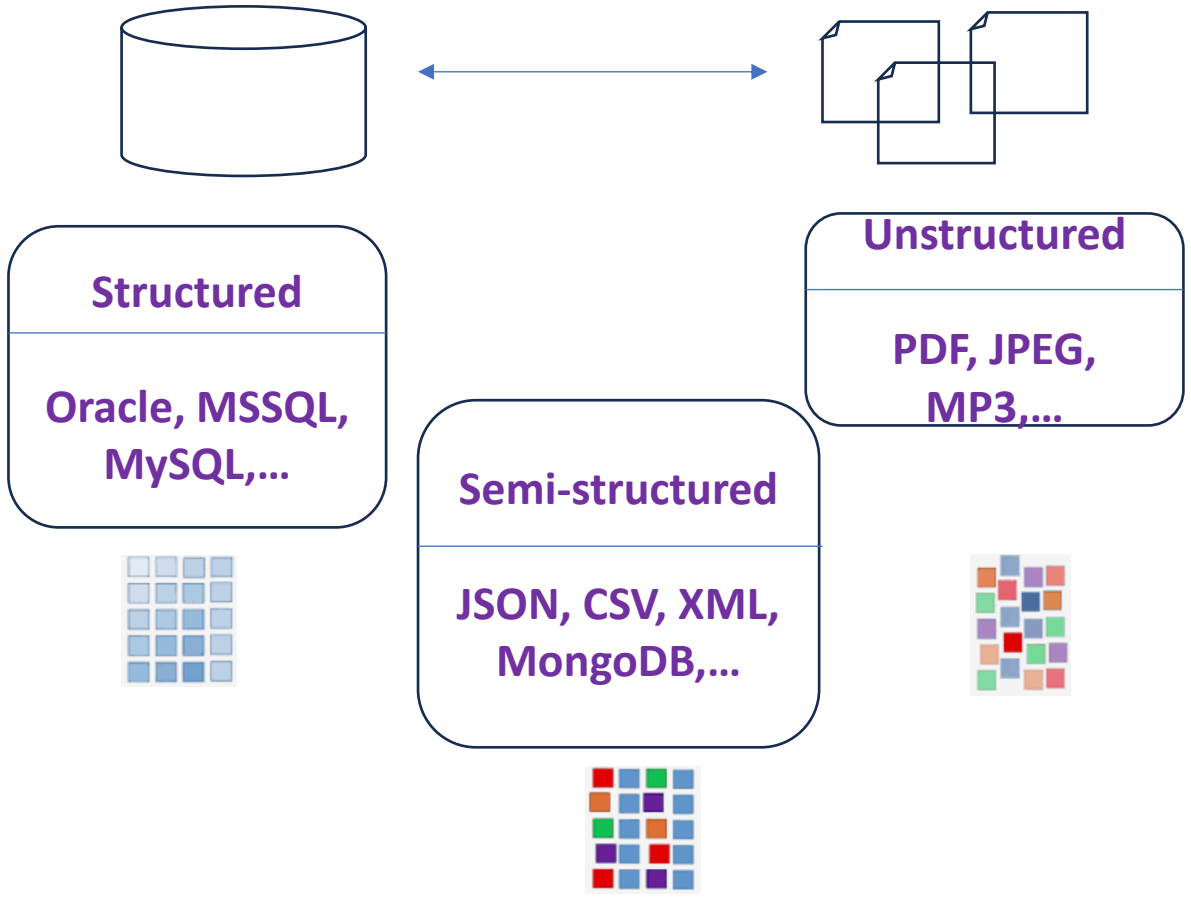
- Structured data
  - Data with a predefined structure (e.g., a relational database)

- Unstructured data
  - Data with no predefined structure
  - With the focus rather on the *content* of the data than the associated metadata (e.g., a blog post, image, video, Tweet, etc.)

- Semi-structured data
  - Allows the data to be stored or represented with its *associated metadata*
  - Formats like JSON allow data to be represented in a flexible manner
  - APIs will often return data in these semi-structured format
  - HTML is another example of representing unstructured data with some structure to enable it to be displayed

Courtesy of EDSA

Structured, semi-structured, unstructured data



Practice

- What types of data are these?
  - A scan of a book
  - Customer order records
  - A cat video
  - A MS Word document containing a report
  - Your bank account statements
  - A Wikipedia page

Courtesy of EDSA

Practice - Answers

- What types of data are these?
  - A scan of a book (Unstructured)
  - Customer order records (Structured)
  - A cat video (Unstructured)
  - A MS Word document containing a report (Unstructured)
  - Your bank account statements (Structured)
  - A Wikipedia page (Semi-structured)

Courtesy of EDSA

DATA PUBLISHING

Data Licenses

- What is “open” data as opposed to shared or closed data?
  - Open/Shared/Closed (https://vimeo.com/125783029)

- Open data definitions
  - “Data that anyone can access, use and share” (The ODI, 2015)
  - Most definitions focus on the possibility to **freely use, reuse** and **redistribute** data

Who Publishes Open Data?

- Governments
  - Big push to increase **transparency** in recent years
  - E.g., http://data.gov, http://data.gov.uk
- Businesses
  - Releasing parts of **data catalogues** openly
  - **Innovation**, new ideas and enterprises
  - Unlocks **new value** from their data
  - Or **new businesses** are created through **innovating** with open data

Open Data Sources

- EU http://open-data.europa.eu/en/data
- UK Government http://data.gov.uk
- London Data Store http://data.london.gov.uk
- Universities e.g., http://data.soton.ac.uk
- Music e.g. https://musicbrainz.org/
- Maps e.g., https://www.openstreetmap.org/
- CKAN - registry of open projects http://ckan.org/
- BBC - http://www.bbc.co.uk/things/
- Web observatories http://webscience.org/web-observatory/list-of-web-observatories
- Linked Open Data Cloud http://lod-cloud.net
- European Data Science Academy Dataset Register, https://edsa-project.eu/resources/datasets/
- Etc.

Data Mining Datasets

- http://www.rdatamining.com/resources/data includes:
  - GeoNames geographical database (8 million place names and data)
  - Airport, airline, and route data (6977 airports, 5888 airlines, 59036 routes spanning the globe)
  - GDELT (over 200 million geolocated events from 1979)
  - And many more!

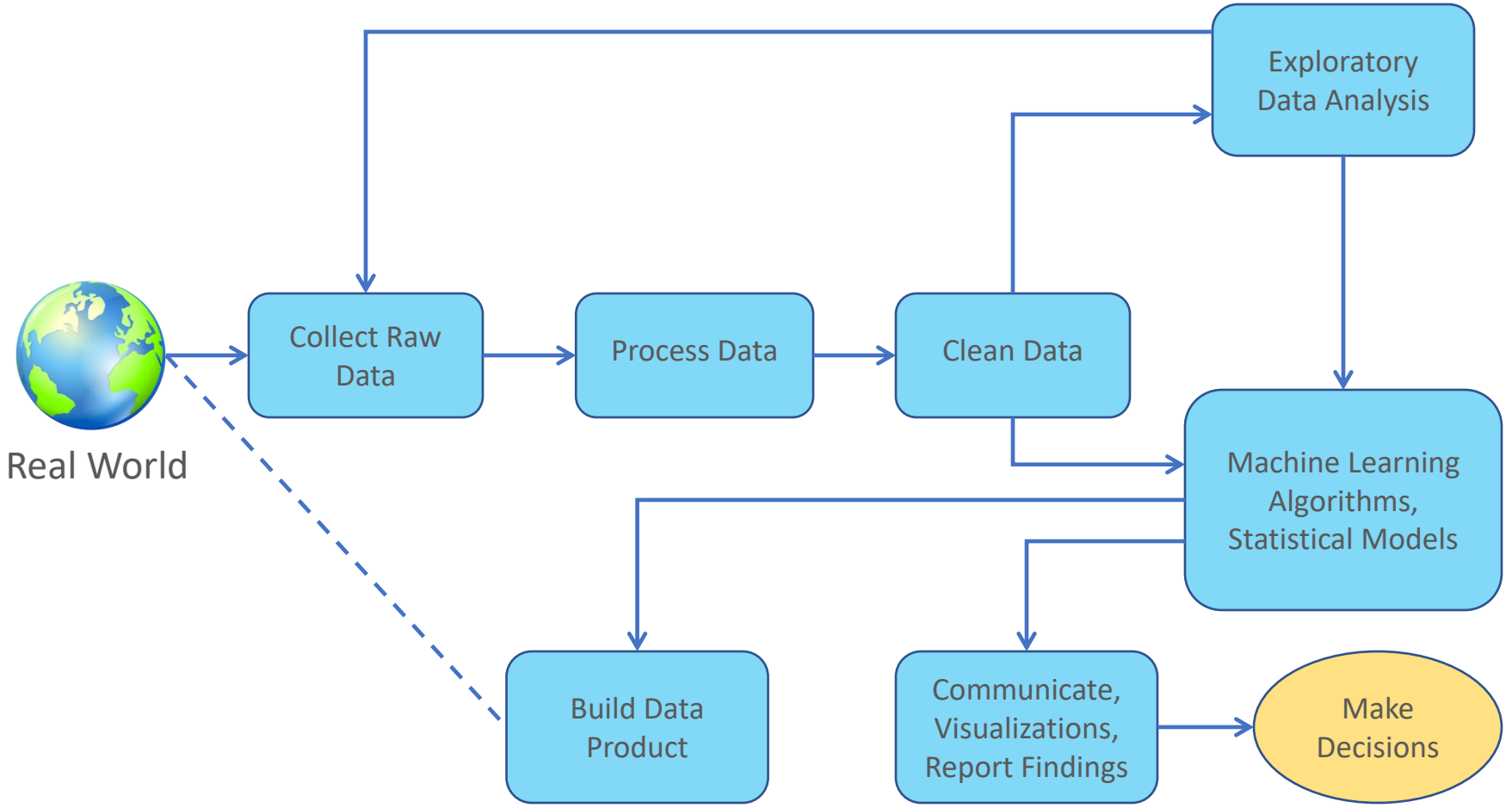
THE DATA SCIENCE PROCESS

Data Scientist at Work

- What does a data scientist do?
  - **Formulates** questions
  - **Discovers data** necessary to produce answers
  - **Plans** analytic techniques to extract insights
  - Creates models and code to carry out simulations or analysis
  - Produces **visualizations** to show insights
  - **Reports findings** back to managers, customers, public, etc.

- Remember these tasks when you implement a big data project!

The Data Science Process



Adapted from O'Neil & Schutt 'Doing Data Science' 1stEdition (2014)

Raw Data Collection

- Various methods, since the data can be anything
  - A sensor can output its readings to be stored
  - Social media interactions could be collected from an API
  - Customer interaction with a service could be recorded in a log for further analysis
  - A page scraper that extracts data from the HTML code

Exploratory Data Analysis

- The process of **exploring** and summarizing the **data**
- The data are **summarized** and the **relationships** between the different variables are explored
- The data are examined in order to identify any **missing data** or **outliers**
  - **These are dealt with in the data cleaning stage**

Data Cleaning

- Preparing data for analysis
  - Identify wrong/missing data and correcting where possible
  - Ensure that data is **consistent**
  - Checking that data is **fit for use**
- Possible issues include
  - **Duplicate** data, e.g., a user with two different accounts
  - **Invalid** data, e.g., email addresses which no longer exist
  - **Inconsistent** data, e.g., a duplicate field which is no longer updated

Data Analysis

- Designing a **model** to **represent** the data
  - Including some form of **algorithm**
  - Using **statistical models** to represent data or predict future values
  - **Machine learning** techniques can be used to gather additional value from the data by making inferences about existing data
  - *The analysis depends on the available data and the type of problem you are trying to solve*
  - Using these techniques in data to gather additional insight is known as **data mining**

Interpreting and Reporting

- Typically involves **visualizing** the results
  - Highlight **critical aspects** of the data (analysis)
- Eases **communication** of the results
  - The results must be accessible to people with a range of backgrounds and levels of expertize
- The report and interpretation are **actionable insights**

Industrial Standards for data science projects

Industrial Standards for Data Science/Big Data Analytics

Why Should There be a Standard Process?

The data science process must be reliable and repeatable - also by people with little data mining background

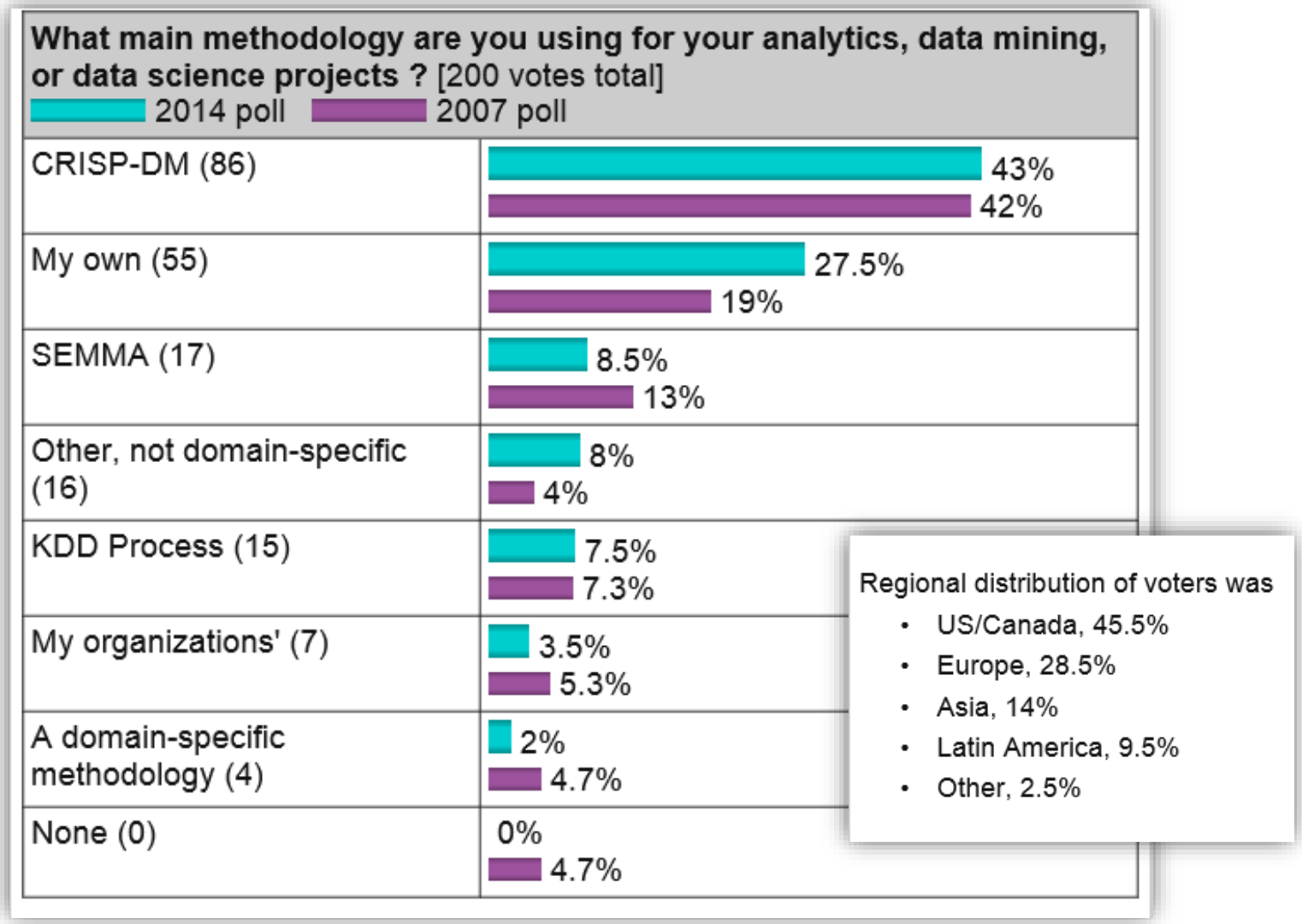
- Framework for recording experience
  - Allows projects to be replicated
- Aid to project planning and management
- “Comfort factor” for new adopters
  - Demonstrates maturity of Data Science

Industrial Standards for Data Science/Big Data Analytics

Poll Results: CRISP-DM still the top methodology

KDnuggets™

CRISP-DM, still the top methodology for analytics, data mining, or data science projects



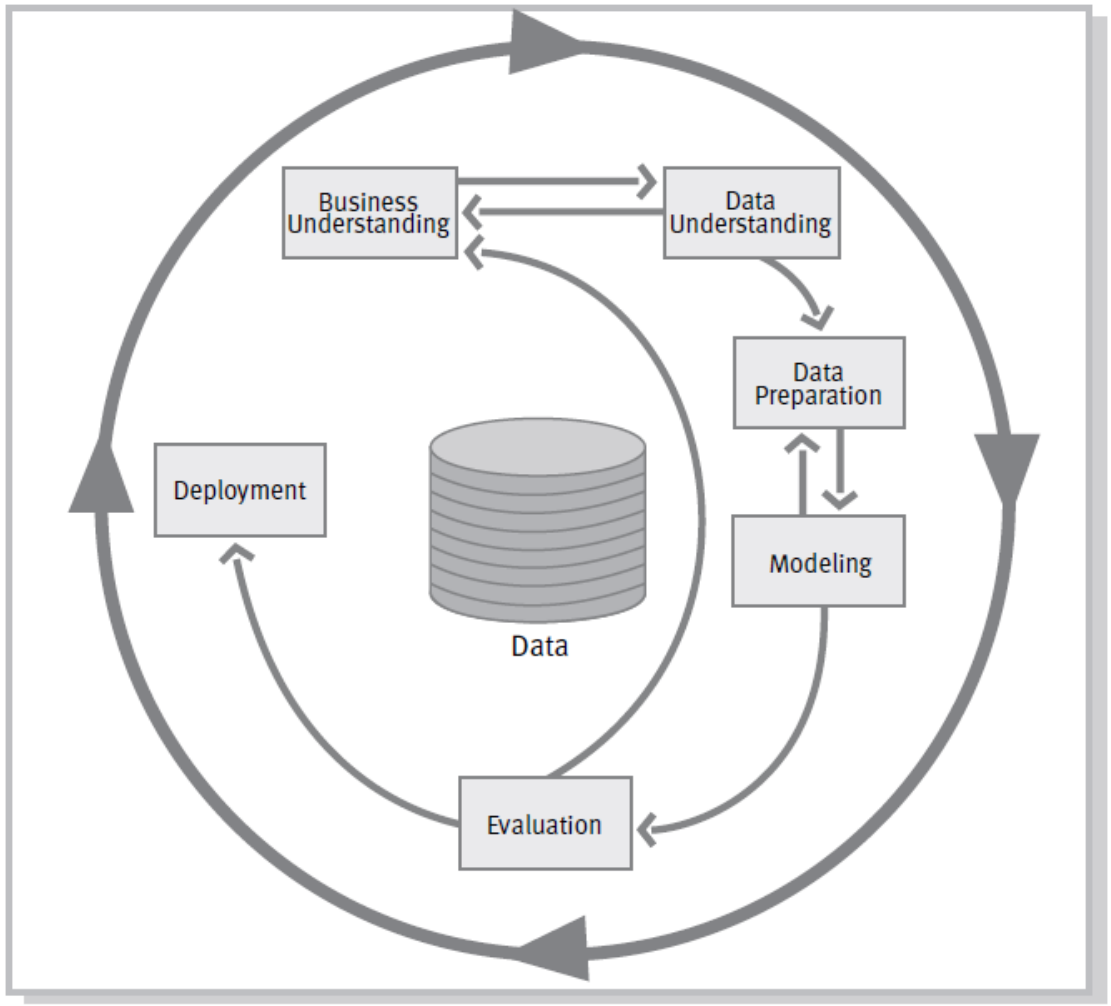
Cross Industry Standard Process for Data Mining
CRISP-DM

- Initiative launched in late 1996 by three “veterans” of data mining market.
  - Daimler Chrysler (then Daimler-Benz), SPSS (now part of IBM) , NCR
- Developed and refined through series of workshops (from 1997-1999)
- Over 300 organization contributed to the process model
- Published CRISP-DM 1.0 (1999):
  - First Version: *https://www.the-modeling-agency.com/crisp-dm.pdf*
  - With Examples: *https://www.ibm.com/docs/it/SS3RA7_18.3.0/pdf/ModelerCRISPDM.pdf*

Industrial Standards for Data Science/Big Data Analytics

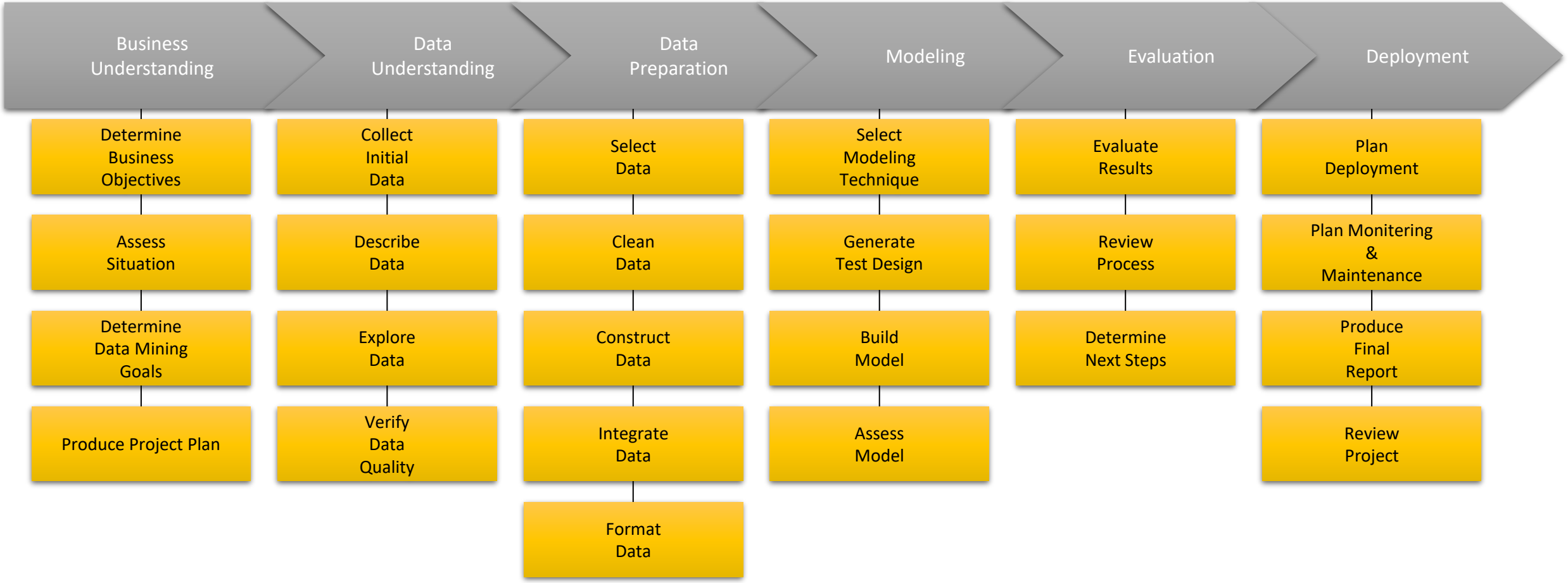
Cross Industry Standard Process for Data Mining

CRISP-DM:



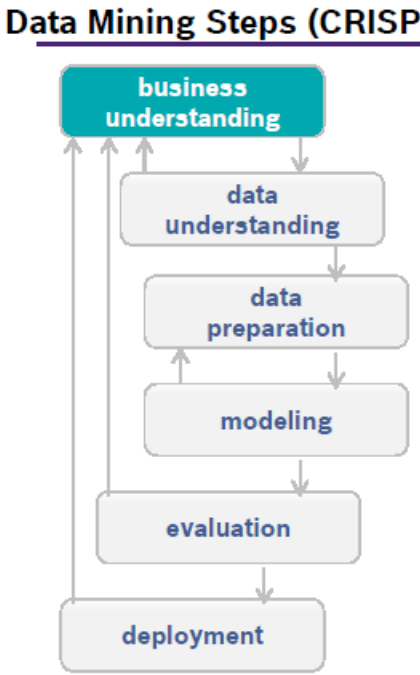
CRISP-DM is a comprehensive data mining methodology and process model that provides anyone- from novices to data mining Experts- with a complete blueprint for conducting a data mining project. CRISP-DM breaks down the life cycle of a data mining project into six phases

Cross Industry Standard for Data Mining
CRISP-DM



CRISP-DM

- Business Understanding



What is the **problem** that data mining should solve for my business?

What are **success criteria** for the data mining activity?

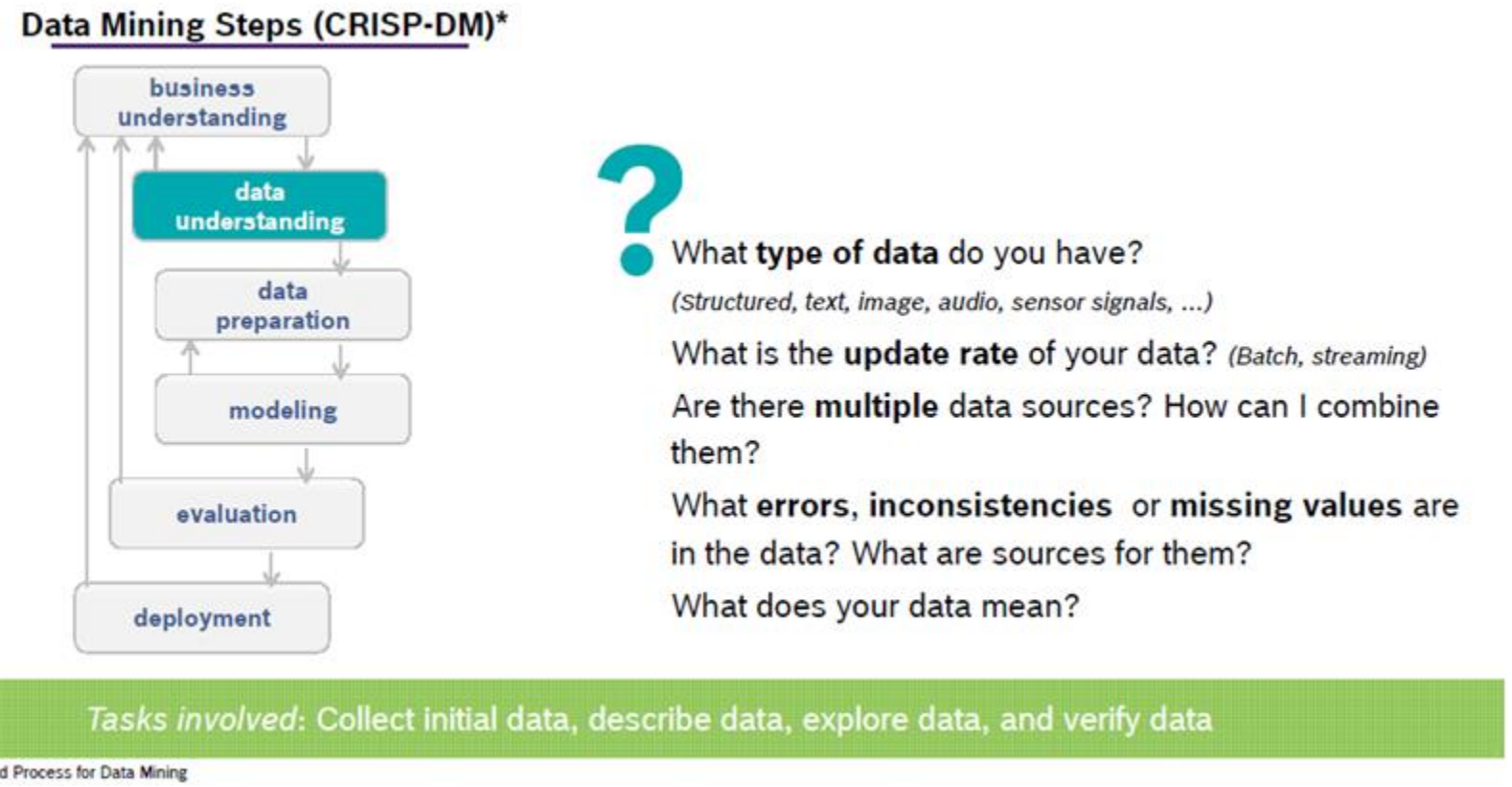
What roles / competencies / collaborations do I need for the project?

Tasks involved: Determine business objectives and requirements, assess situation, determine data mining problem definition and objectives

* Cross-Industry Standard Process for Data Mining

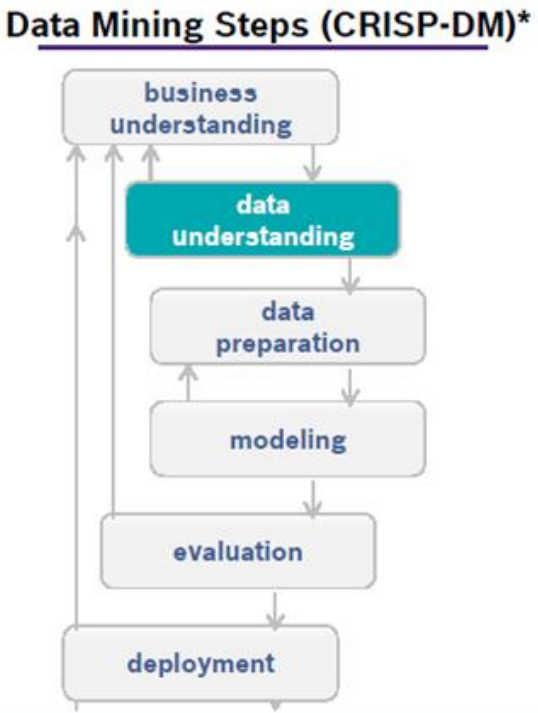
CRISP-DM

- Data Understanding



CRISP-DM

- Data Understanding



Seminar

Student

Attribute	Length	Type	Rules
Name	40	Alpha	At least 2 words
Email Address	50	Mixed	Must contain @
Phone #	10	Numeric	Reject all "555"
Address	30	Mixed	Format - ### alpha
City	20	Alpha	none
State	2	Alpha	Must be a valid state

Data dictionary

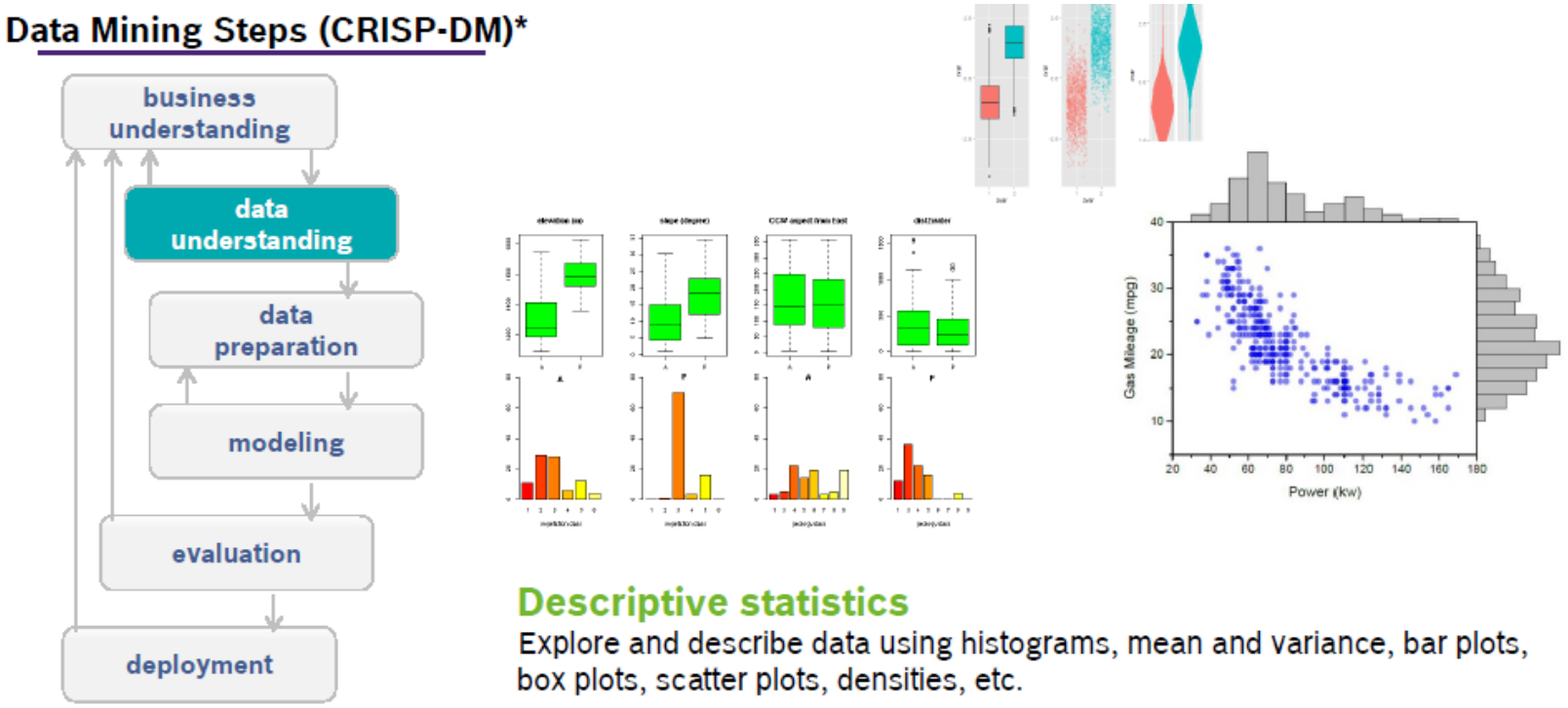
summarizes domain-related and technical information about your data. Documentation of what data you have, what it means and helps to identify inconsistencies and errors.

Tasks involved: Collect initial data, describe data, explore data, and verify data

* Cross-Industry Standard Process for Data Mining

CRISP-DM

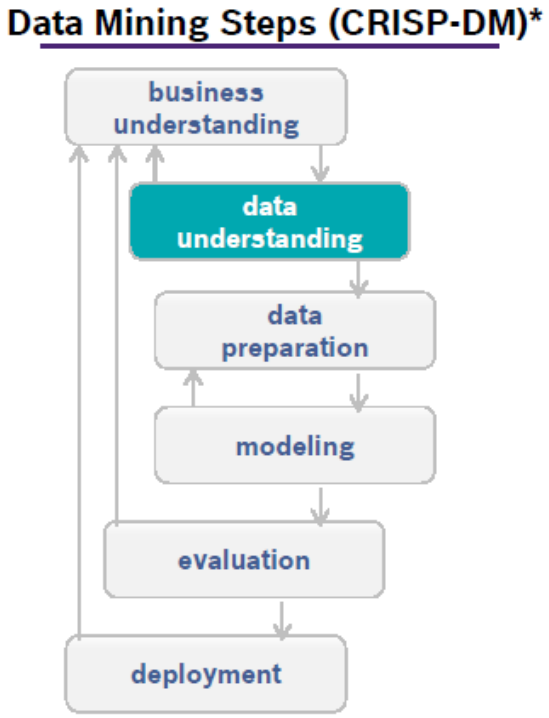
- Data Understanding

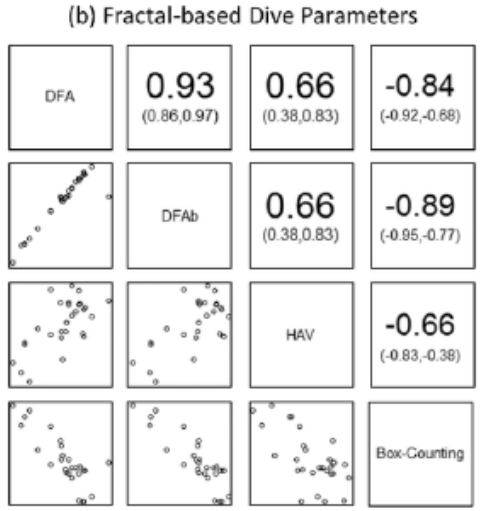


* Cross-Industry Standard Process for Data Mining

CRISP-DM

- Data Understanding





Associations and correlations
between attributes / features can be measured by different correlation coefficients

Pearson's correlation coefficient (linear):

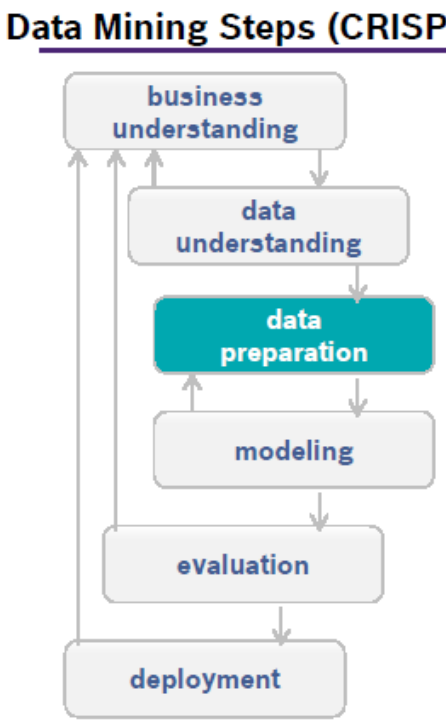
$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

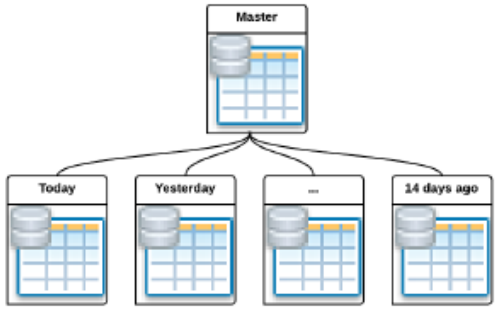
Tasks involved: Collect initial data, describe data, explore data, and verify data

* Cross-Industry Standard Process for Data Mining

CRISP-DM

- Data Preparation





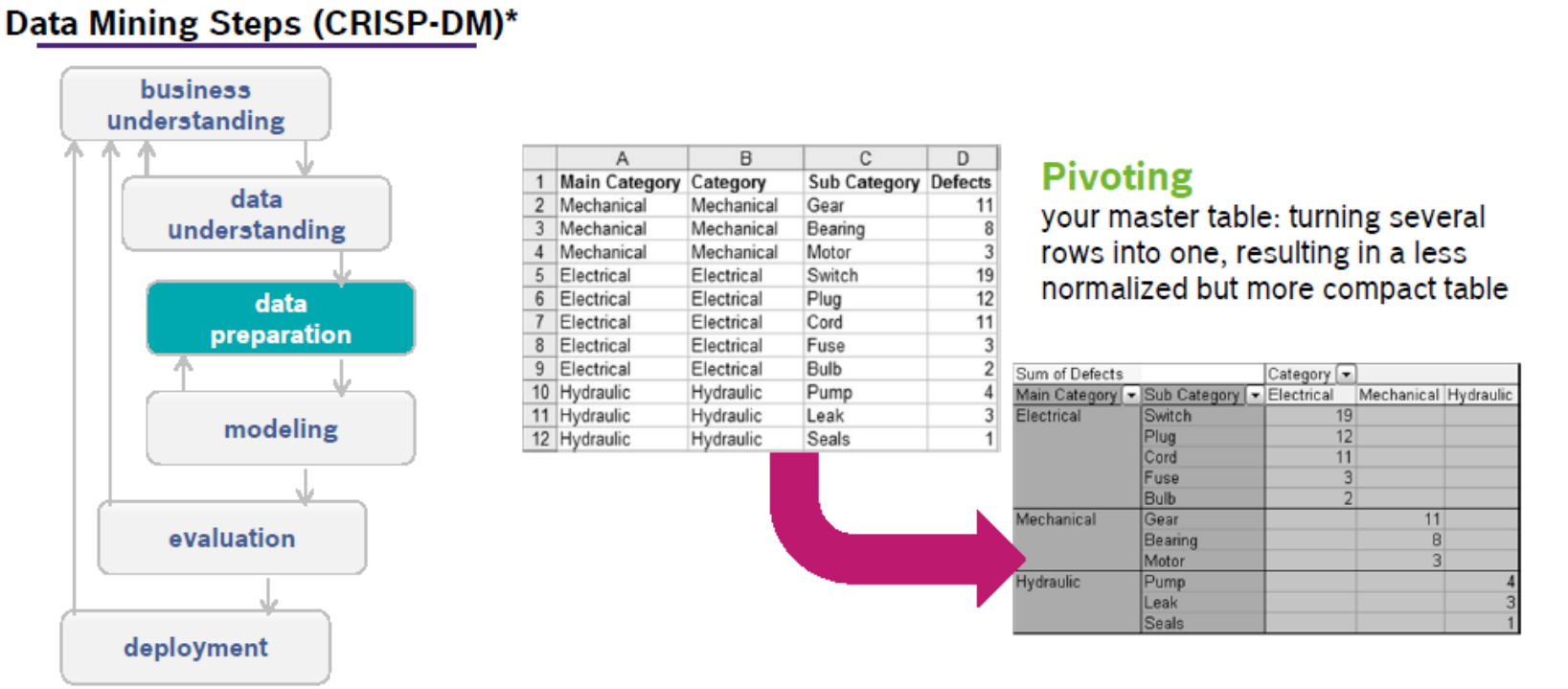
Selecting data I need and gather them in **one master table** – final dataset to feed to modeling

Tasks involved: Select data, clean data, construct data, integrate data, and format data

* Cross-Industry Standard Process for Data Mining

CRISP-DM

- Data Preparation

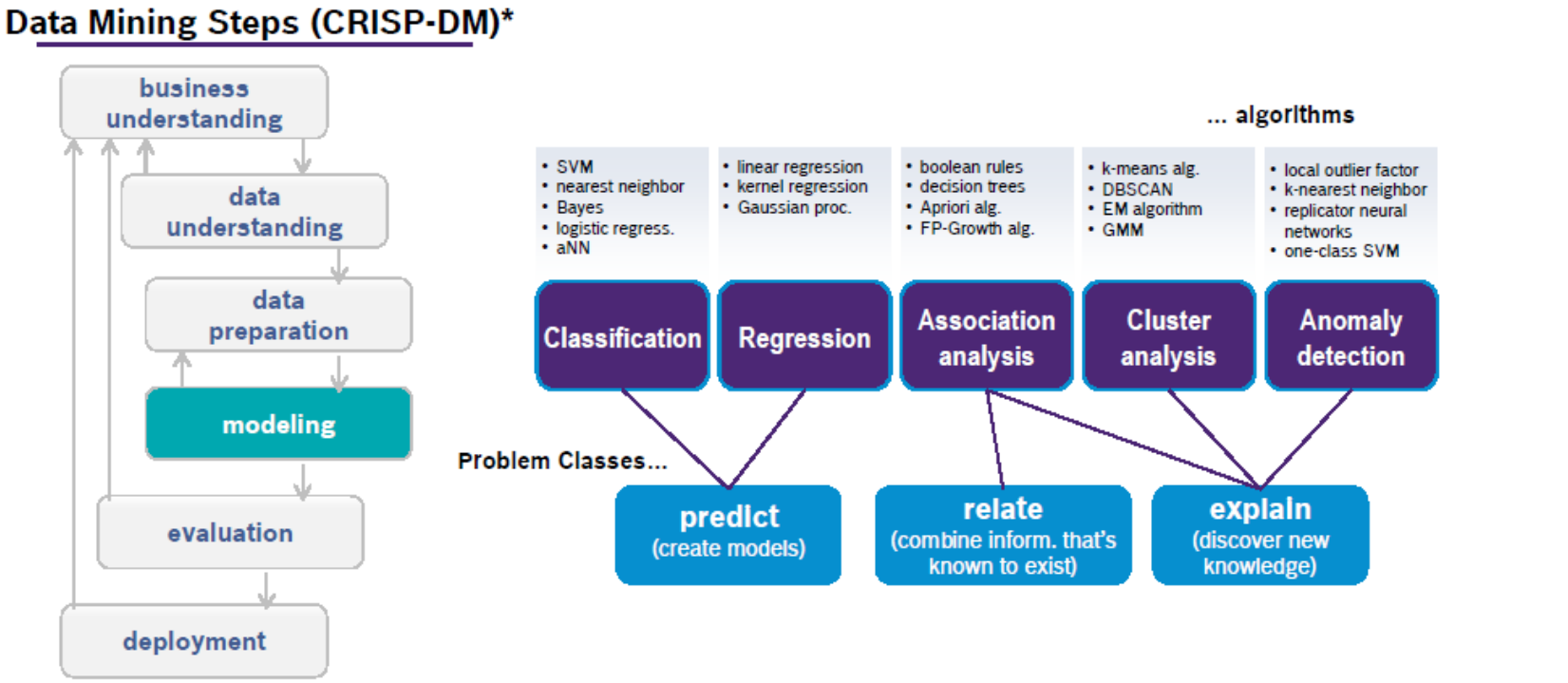


	A	B	C	D
1	Main Category	Category	Sub Category	Defects
2	Mechanical	Mechanical	Gear	11
3	Mechanical	Mechanical	Bearing	8
4	Mechanical	Mechanical	Motor	3
5	Electrical	Electrical	Switch	19
6	Electrical	Electrical	Plug	12
7	Electrical	Electrical	Cord	11
8	Electrical	Electrical	Fuse	3
9	Electrical	Electrical	Bulb	2
10	Hydraulic	Hydraulic	Pump	4
11	Hydraulic	Hydraulic	Leak	3
12	Hydraulic	Hydraulic	Seals	1

Sum of Defects		Category		
Main Category	Sub Category	Electrical	Mechanical	Hydraulic
Electrical	Switch	19		
	Plug	12		
	Cord	11		
	Fuse	3		
	Bulb	2		
Mechanical	Gear		11	
	Bearing		8	
	Motor		3	
Hydraulic	Pump			4
	Leak			3
	Seals			1

CRISP-DM

- Modeling



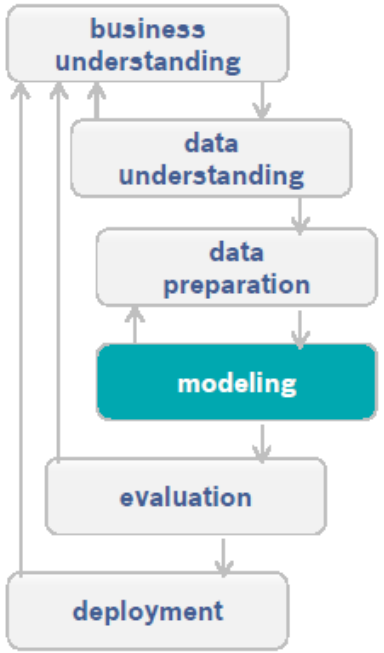
Tasks involved: Select model techniques, generate test design, build model, and asses model

* Cross-Industry Standard Process for Data Mining

CRISP-DM

- Modeling

Data Mining Steps (CRISP-DM)*



Assess a model's **classification performance** using a confusion matrix

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

Sensitivity = TP / (TP+FN)
Specificity = TN / (TN+FP)

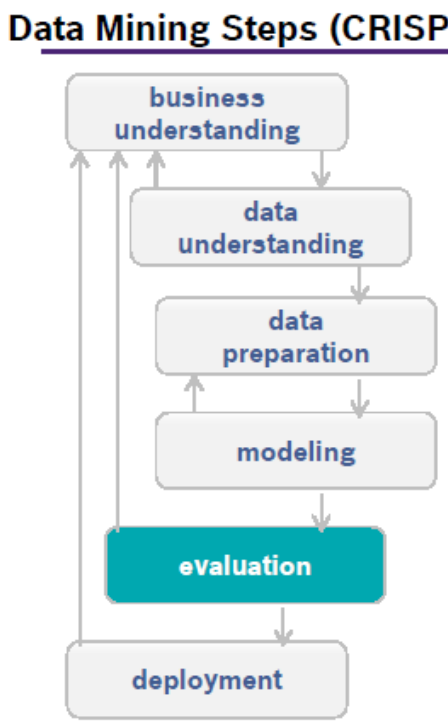
	Spam (Predicted)	Non-Spam (Predicted)	Accuracy
Spam (Actual)	27	6	81.81
Non-Spam (Actual)	10	57	85.07
Overall Accuracy			83.44

Tasks involved: Select model techniques, generate test design, build model, and asses model

* Cross-Industry Standard Process for Data Mining

CRISP-DM

- Evaluation



How to **interpret** the results in terms of the application?

Do the results fulfill the **data mining goal**? Do they contribute to your business objectives?

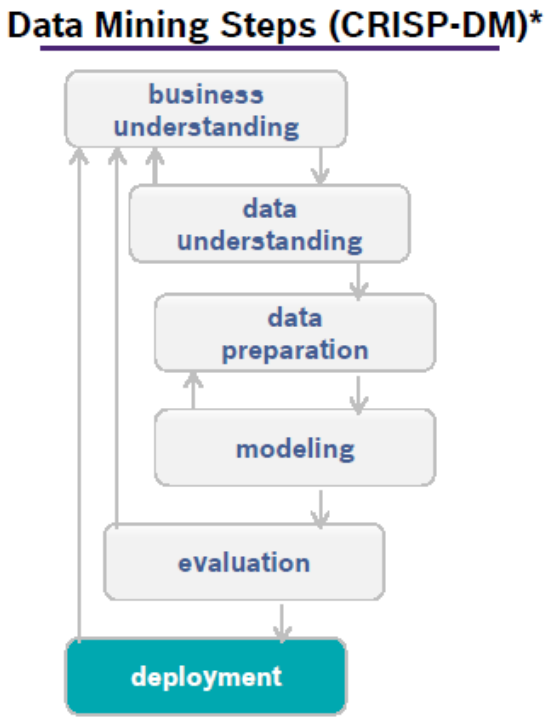
What are findings besides the model? Are these findings related to the **original business goal**?

Tasks involved: Evaluate results, review process, and determine next steps

* Cross-Industry Standard Process for Data Mining

CRISP-DM

- Deployment



What are **deployable results**?

How will **information propagate** to users and decision makers? What is the interface?

What is the **execution model** and the **necessary IT infrastructure**? How will it be maintained?

How will the use of the model be monitored? How to get **user feedback**?

Tasks involved: Plan deployment, plan monitoring and maintenance, produce final report, and review project

* Cross-Industry Standard Process for Data Mining

Industrial Standards for Big Data Analytics
CRISP-DM References

- Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz, (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler) “CRISP-DM 1.0 - Step-by-step data mining guide”

- “The CRISP-DM Model: The New Blueprint for DataMining”, Colin Shearer, JOURNAL of Data Warehousing, Volume 5, Number 4, p. 13-22, 2000

- https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining
- http://www.kdnuggets.com/2016/03/data-science-process-rediscovered.html/2

DATA SCIENCE APPLICATIONS

Smart Home

- Is a residential home equipped with special wiring, sensors and meters, to enable occupants to (remotely) control (and automate) lighting, heating, ventilation. air conditioning, appliances, and security

Source: http://www.householdappliancesworld.com/2016/06/21/rising-interest-iot-smart-home/

Data Science App for Smart Home

Data Science Process	Application
1. Data Collection / data sources	Data comes from (smart) meters and sensors
2. Data Processing	Analyses data to identify ways of improving the efficiency of the home appliances, e.g., energy efficiency
3. Data Analysis (Machine learning, Data mining)	Can predict potential problems with the running of the smart home or make recommendations for the occupants, e.g. when to switch on/off heating in order to save energy
4. Data Product	Helps residential users to personalize the resource consumption in their home in order to improve resource saving while keeping a maximum comfort in the home

Data Science App for Smart Home

1. Data Collection / data sources

- Data Ingestion Simulation

```python
from pyspark.sql import SparkSession
from pyspark.sql.types import StructType, StructField, TimestampType, FloatType, StringType, IntegerType
from pyspark.sql.functions import current_timestamp, rand, expr

spark = SparkSession.builder.appName("SmartHomeDataCollection").getOrCreate()

# Define schema
sensor_schema = StructType([
    StructField("timestamp", TimestampType(), True),
    StructField("sensor_id", StringType(), True),
    StructField("temperature", FloatType(), True),
    StructField("humidity", FloatType(), True),
    StructField("power_consumption", FloatType(), True),
    StructField("light_level", IntegerType(), True),
    StructField("appliance_status", StringType(), True)
])

# Simulate sensor data generation
def generate_sensor_data(num_rows):
    return spark.range(num_rows).select(
        current_timestamp().alias("timestamp"),
        expr("concat('sensor_', floor(rand() * 5))").alias("sensor_id"),
        (rand() * 30 + 15).alias("temperature"),  # Temperature between 15 and 45
        (rand() * 60 + 30).alias("humidity"),  # Humidity between 30 and 90
        (rand() * 100).alias("power_consumption"),
        (rand() * 1000).cast("integer").alias("light_level"),
        expr("CASE WHEN rand() > 0.5 THEN 'ON' ELSE 'OFF' END").alias("appliance_status")
    )

sensor_data = generate_sensor_data(1000)
sensor_data.show(5)
```

Data Science App for Smart Home

2. Data Processing (for Energy Efficiency Analysis)

- Calculating Energy Consumption Patterns

```
from pyspark.sql.functions import hour, dayofweek, avg, sum, col, date_format

# Extract time features
processed_data = sensor_data.withColumn("hour", hour("timestamp")) \
                            .withColumn("day_of_week", dayofweek("timestamp"))\
                            .withColumn("date", date_format("timestamp", "yyyy-MM-dd"))

# Aggregate power consumption by hour and day
hourly_power_consumption = processed_data.groupBy("hour").agg(avg("power_consumption").alias("avg_power"))
daily_power_consumption = processed_data.groupBy("date").agg(sum("power_consumption").alias("total_power"))

hourly_power_consumption.show()
daily_power_consumption.show()

#Aggregate power consumption by appliance status
appliance_power_usage = processed_data.groupBy("appliance_status").agg(avg("power_consumption").alias("avg_power"))
appliance_power_usage.show()

#Find the power consumed during the weekend vs weekdays
weekday_power = processed_data.filter(col("day_of_week").isin([2,3,4,5,6])).agg(sum("power_consumption").alias("weekday_power"))
weekend_power = processed_data.filter(col("day_of_week").isin([1,7])).agg(sum("power_consumption").alias("weekend_power"))

weekday_power.show()
weekend_power.show()
```

Data Science App for Smart Home

2. Data Processing

(for Energy Efficiency Analysis)

```
# Assuming appliances have unique sensor_ids
appliance_power = processed_data.groupBy("sensor_id").agg(avg("power_consumption").alias("avg_power"))

# Identify appliances with unusually high power consumption
threshold = appliance_power.agg(avg("avg_power")).collect()[0][0] * 1.5  # Example threshold
inefficient_appliances = appliance_power.filter(col("avg_power") > threshold)

inefficient_appliances.show()
```

- Identify Inefficient Appliances

Data Science App for Smart Home

3. Data Analysis

(Predictive Maintenance and Recommendations)

- Temperature Prediction (Regression)

```python
from pyspark.ml.feature import VectorAssembler, StandardScaler
from pyspark.ml.regression import LinearRegression
from pyspark.ml.evaluation import RegressionEvaluator

# Prepare features
assembler = VectorAssembler(inputCols=["hour", "humidity", "power_consumption", "light_level"], outputCol="features")
feature_data = assembler.transform(processed_data)

# Scale features
scaler = StandardScaler(inputCol="features", outputCol="scaled_features")
scaled_data = scaler.fit(feature_data).transform(feature_data)

# Split data
train_data, test_data = scaled_data.randomSplit([0.8, 0.2], seed=42)

# Train Linear Regression model
lr = LinearRegression(featuresCol="scaled_features", labelCol="temperature")
model = lr.fit(train_data)

# Evaluate model
predictions = model.transform(test_data)
evaluator = RegressionEvaluator(labelCol="temperature", predictionCol="prediction", metricName="rmse")
rmse = evaluator.evaluate(predictions)
print(f"Temperature Prediction RMSE: {rmse}")

predictions.select("temperature","prediction").show(10)
```

Data Science App for Smart Home

3. Data Analysis

(Predictive Maintenance and Recommendations)

- Anomaly Detection (potential problems)

```python
from pyspark.ml.clustering import KMeans

# KMeans for anomaly detection (e.g., unusual power consumption patterns)
kmeans = KMeans().setK(2).setSeed(42)  # Adjust K based on your data
anomaly_model = kmeans.fit(scaled_data)
anomaly_predictions = anomaly_model.transform(scaled_data)

# Identify potential anomalies
anomaly_predictions.groupBy("prediction").agg(avg("power_consumption")).show()
```

Data Science App for Smart Home

3. Data Analysis

(Predictive Maintenance and Recommendations)

```
# Simple recommendation: Suggest turning off appliances during peak hours
peak_hour_power = hourly_power_consumption.orderBy(col("avg_power").desc()).first()["hour"]

print(f"Recommendation: Consider turning off non-essential appliances during hour {peak_hour_power} to save energy.")
```

- Recommendation System

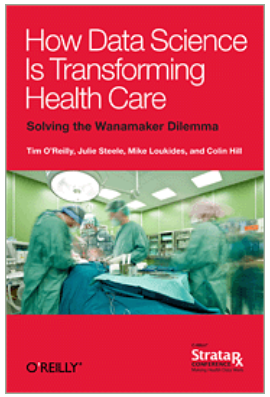
Data Science App for Smart Home

4. Data Product: Personalized Resource Consumption

- Building a User-Friendly Interface
- Integrating with Smart Home Devices

Source: plavno.io

Healthcare



- Patient data (via remote monitoring)
- Clinical data (medical records and images)
- Pharmaceutical R&D data (data from clinical trials)
- Activity and cost data (what diseases and where, cost of treatment, etc.)

http://insidebigdata.com/2013/09/20/free-ebook-data-science-transforming-health-care/

Data Science in Healthcare

Data Science Process	Application
Collecting Data	Health-related data collected through sensors, either in the hospital or outside, e.g., wearable technology
Machine leaning / Data mining	Analyses data to identify e.g., when a patient might be at risk, what are the best medical choices for a given patient, etc.
Product	Notifies patients or hospital if the patient is at risk, provides recommendations regarding a certain medical situation, etc.

Football Analytics

- **Player movements are monitored** with sensors
- Improve team performance with analysis and decision making
  - Spot potential injuries **before they occur**, based on player movement

Source: http://www.bbc.com/news/business-26771259

Data Science Process on Football Analytics

Data Science Process	Application
Collecting data	Uses sensors positioned at strategic locations on the players and on the pitch
Data processing and Machine learning	(sensor) Data is processed in real time. Can predict factors which may lead to players getting injured, or issues that players need to improve
Visualization	Diagram showing the players' movement on a football pitch
Product	An application that advises football teams

Data Science in COVID 19 fighting

Possible applications

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7204193/table/Tab1/?report=objectonly

COVID 19 pandemic as a data science issue

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7144860/

DATA SCIENCE TOOLS

Technologies and Tools

- Depending on the situation (or the problem to solve), different technologies and tools will be used

- Therefore, it is suggested that you know about as many tools and technologies as possible in order to identify which one to use in a particular scenario

Technologies and Tools

- Programming languages
  - Python, Scala, Java, JavaScript, R
- Big data processing
  - Hadoop, HDFS, MapReduce, Spark, Storm, Hive, Pig
- Data management
  - SQL, NoSQL (MongoDB, Cassandra)
- Visualization
  - D3.js, Gephi, Tableau, Shiny, Excel, Gapminder
- Data Analysis
  - R, Python, Excel, Weka, RapidMinder, OpenRefine, SAS, SPSS, Watson Analytics, Open Calais, Matlab

Data Science – take home

- Data science: what is it and what is it used for?
- Main data types
- Data Scientist vs Data Engineer vs Data Analyst
- Data science process/CRISP – DM
- Data Science main tools and real life apps

Foundations of Big Data

Agenda

Foundations of Data Science

- Introduction
- Definitions
- Types of Data
- Data Publishing
- The Data Science Process + CRISP-DM (in a separate course)
- Data Science projects: examples
- Data Science Tools

Foundations of Big Data

- Big Data: Origins
- Big Data: Evolution
- Big Data: Stats, Growth and Facts
- Big Data: Definition and characteristics
- Big Data ecosystem: main technological components and tools
- Big Data Tools
- Big Data: Challenges
- Big Data Market

Big Data: Origins

2013 - Oxford English Dictionary introduced the term “Big Data” for the first time in its dictionary

WHO CAME UP WITH BIG DATA?

Big Data is 'the' thing to be in. Do you know who really came up with Big Data?

Infographic authored by: **Ramesh Dontha**
https://www.**linkedin.com/in/rameshdontha**
Twitter: **rkdontha1**
www.**DigitalTransformationPro**.Com

1944

Wesleyan University Librarian **Fremont Ryder** speculated that 2040 Yale Library will have 200 million volumes because of information explosion

1980

Oxford English Discovery folks discovered that Sociologist **Charles Tilly** was the **first person** to use the term **Big Data** in this sentence in his article.

1990

Peter Denning thought of what's possible: "To build machines that can recognize or predict patterns in data"

1997

Michael Cox and David Ellsworth used the term Big Data for the **first time** in ACM paper.

1998

John Mashey of SGI is **credited with coming up** with the term Big Data and used in a paper in this year.

2000

Francis Diebold referred to Big Data as "the explosion in the quantity (and sometimes, quality) of available and potentially relevant data"

2001

Doug Laney (Meta/Gartner) came up with the **3 'V's** (Volume, Velocity, Variety)

2005

Tim O'Reilly published 'What is Web 2.0?' Roger Mougalas of O'Reilly Media used the term 'Big Data in its modern context '

2005

Hadoop was created by Yahoo! built on top of Google's MapReduce.

2008

Google processed 20 Petabytes of data in single day

2013

4.4 Zettabytes of information was produced by the universe

2016/Present

Businesses are implementing latest **Big Data technologies such as in-memory technologies** to take advantage of Big Data

D T PRO

WHO CAME UP WITH BIG DATA?

Big Data is 'the' thing to be in. Do you know who really came up with Big Data?

Infographic authored by: **Ramesh Dontha**
https://www.**linkedin.com/in/rameshdontha**
Twitter: **rkdontha1**
www.**DigitalTransformationPro**.Com

1944

Wesleyan University Librarian **Fremont Ryder** speculated that 2040 Yale Library will have 200 million volumes because of information explosion

1980

Oxford English Discovery folks discovered that Sociologist **Charles Tilly** was the **first person** to use the term **Big Data** in this sentence in his article.

1990

Peter Denning thought of what's possible: "To build machines that can recognize or predict patterns in data"

1997

Michael Cox and David Ellsworth used the term Big Data for the **first time** in ACM paper.

1998

John Mashey of SGI is **credited with coming up** with the term Big Data and used in a paper in this year.

2000

Francis Diebold referred to Big Data as "the explosion in the quantity (and sometimes, quality) of available and potentially relevant data"

First time Big Data is linked to the way we understand the term today

2001

Doug Laney (Meta/Gartner) came up with the **3 'V's** (Volume, Velocity, Variety)

2005

Tim O'Reilly published 'What is Web 2.0?'
Roger Mougalas of O'Reilly Media used the term 'Big Data' in its modern context

The Big Data Revolution has begun

2005

Hadoop was created by Yahoo! built on top of Google's MapReduce.

2008

Google processed 20 Petabytes of data in single day

2013

4.4 Zettabytes of information was produced by the universe

The term Big Data appears in Oxford Dictionaries

2016/Present

Businesses are implementing latest **Big Data technologies such as in-memory technologies** to take advantage of Big Data

D T PRO

Big Data: Origins

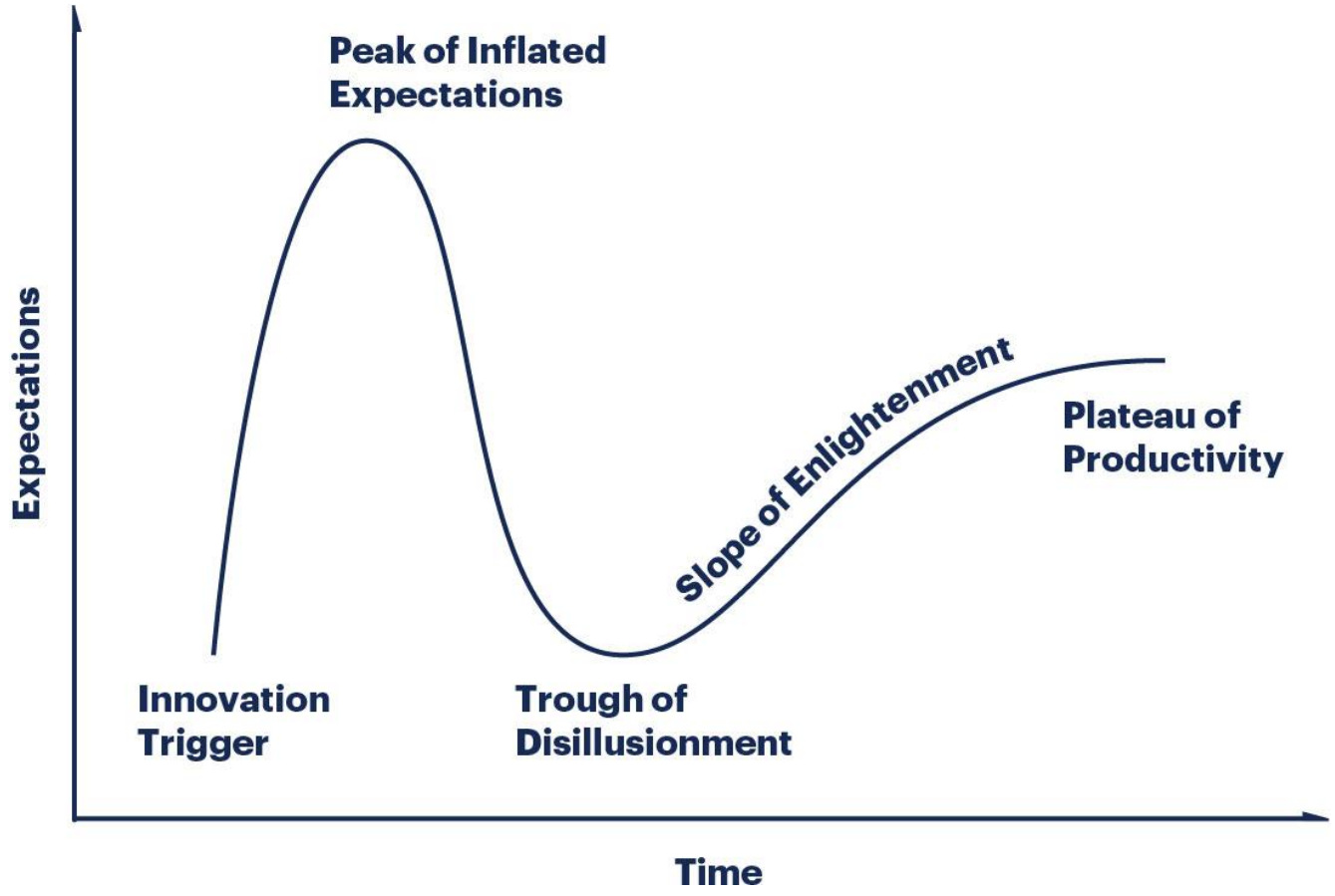
Sources:
The Origins of Big Data
A Very Short History of Big Data by Gil Press of Forbes
The Origins of Big Data: An Etymological detective story
A Short history of Big Data
The OED, Big Data, and Crowdsourcing

Big Data: Evolution

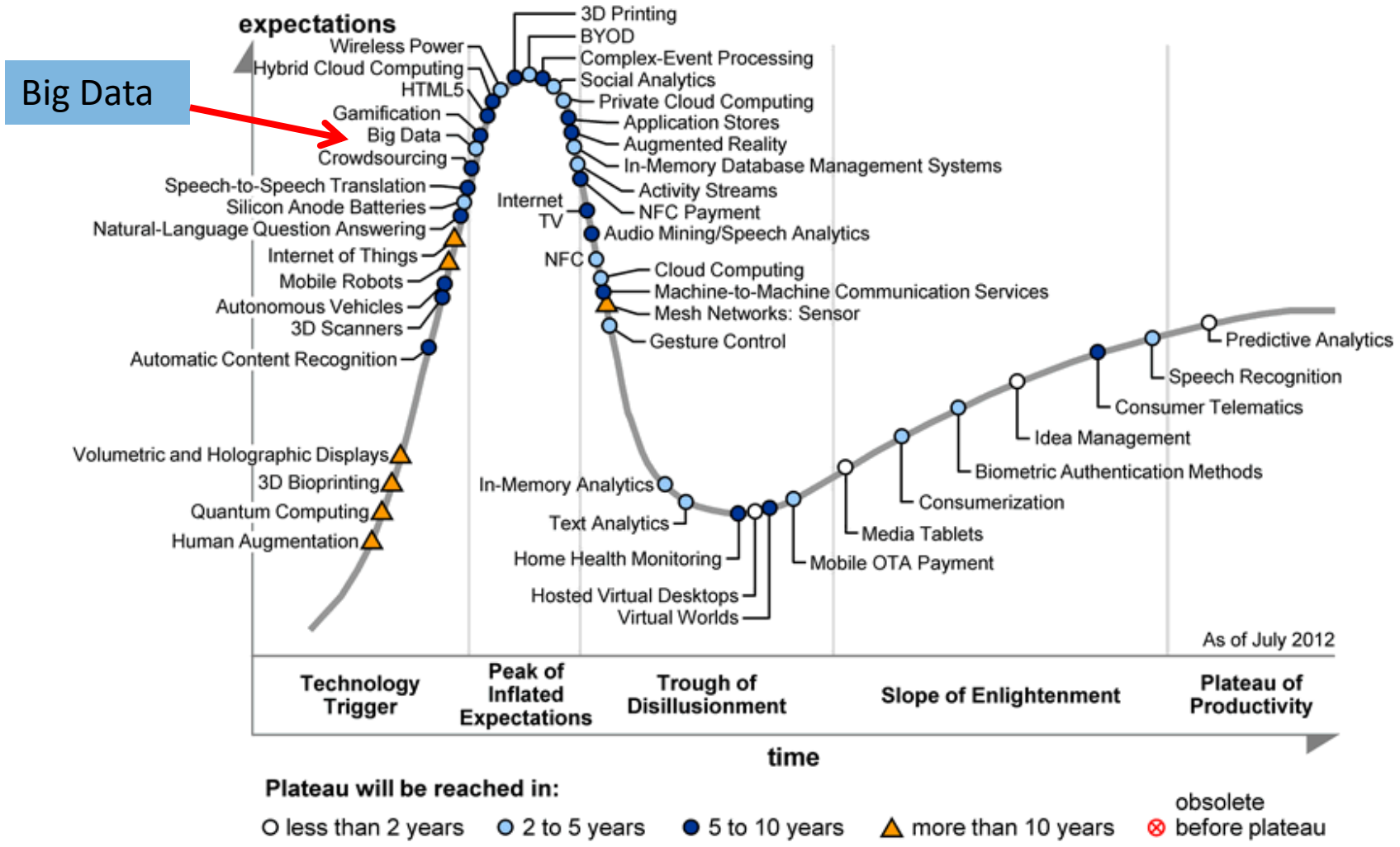
Gartner Hype Cycle

The **Gartner Hype Cycle** is a graphical representation of the perceived value of a technology trend or innovation—and its relative market promotion.

The **cycle** can help you understand how the perceived value of a given technology evolves over the course of its maturity lifecycle.

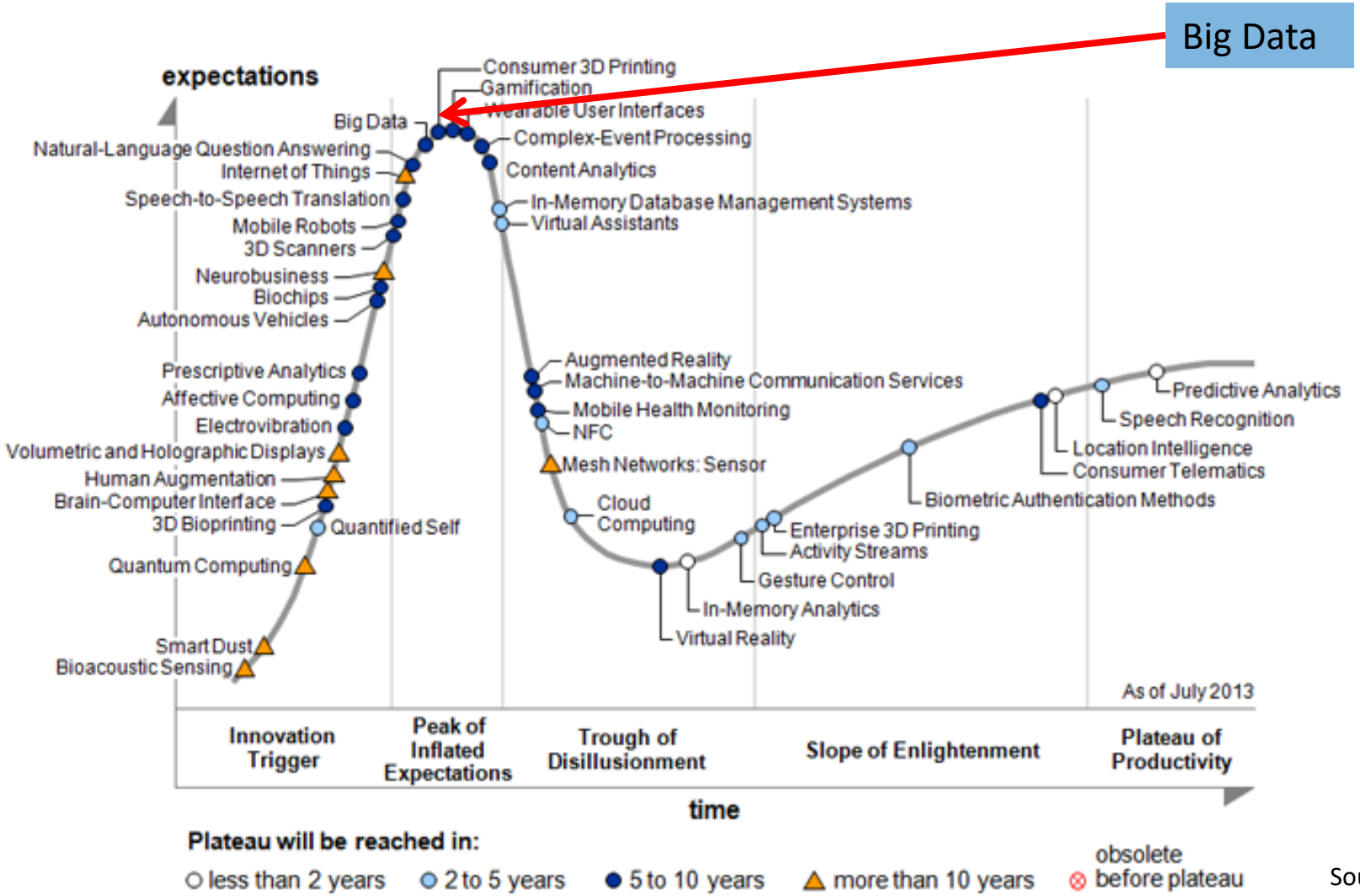


Gartner: Emerging Technologies Hype Cycle 2012



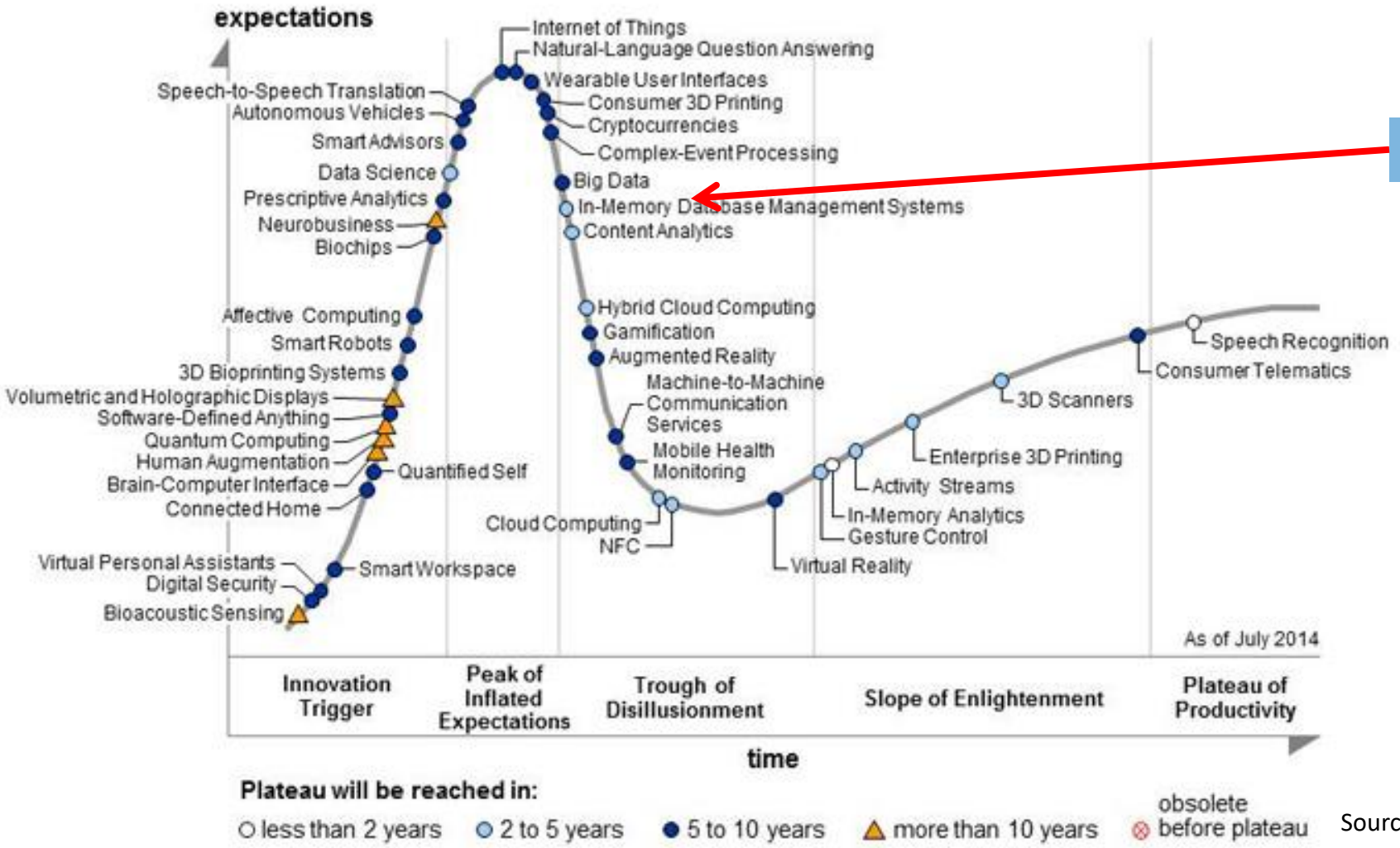
Source: http://www.gartner.com

Gartner: Emerging Technologies Hype Cycle 2013



Source: http://www.gartner.com

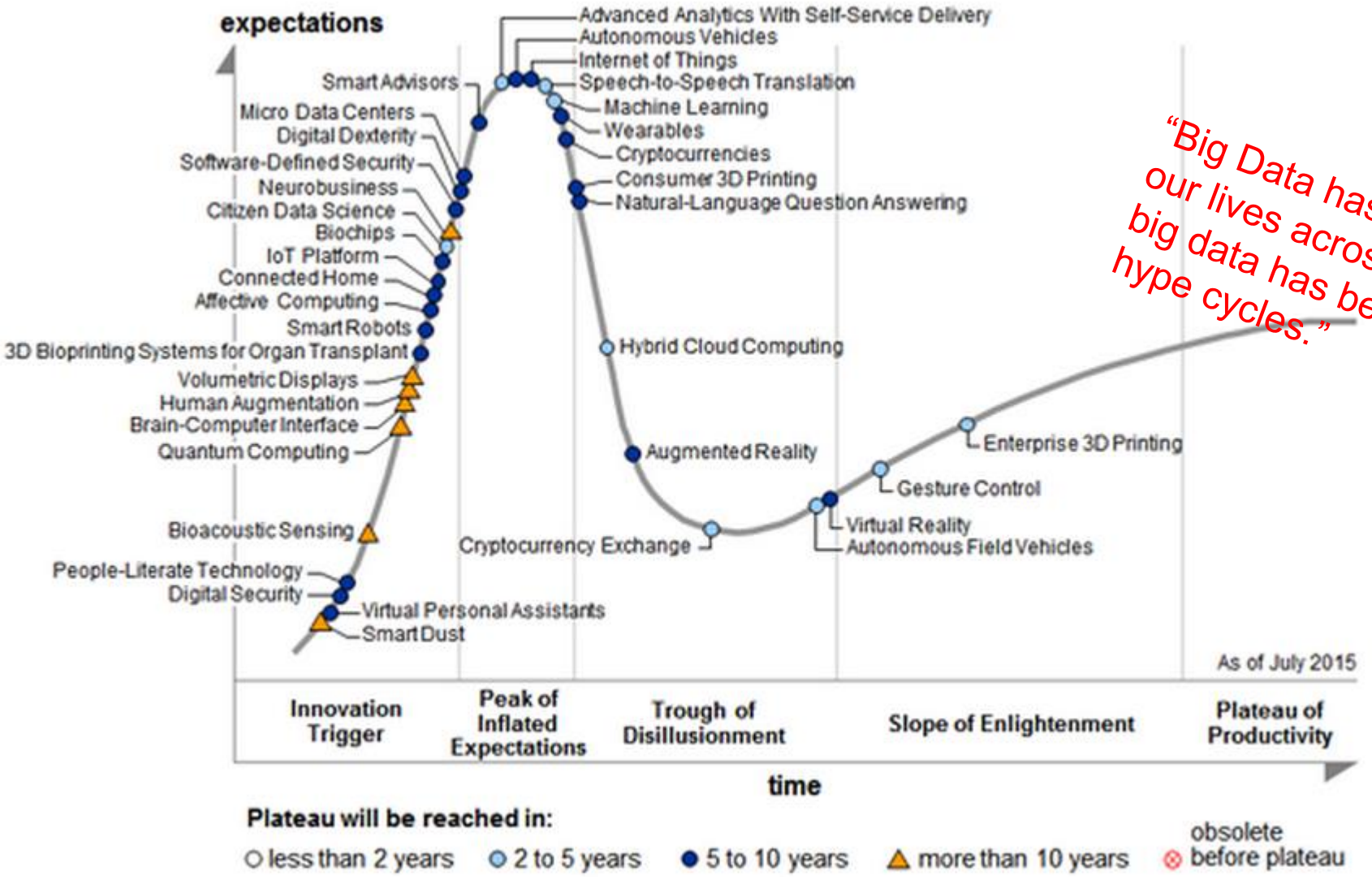
Gartner: Emerging Technologies Hype Cycle 2014



Big Data

Source: http://www.gartner.com

Gartner: Emerging Technologies Hype Cycle 2015



Starting 2015 Big Data is not considered any more a hype by Gartner

"Big Data has become prevalent in our lives across many hype cycles. So big data has become a part of many hype cycles."

Source: http://www.gartner.com

Big Data: Statistics, Growth and Facts

Stats, Growth and Facts

Key Stats:

- The **big data market will reach an estimated value of $103 billion by 2023**
- It's estimated **97.2% of companies are starting to invest in big data technology**
- Every day, **internet users create 2.5 quintillion bytes of data**
- IDC's Digital Universe Study from 2012 found that just **0.5% of data was actually being analyzed**
- It was estimated that by **2021 every person will generate about 1.7 megabytes of data per second**
- Companies like **Netflix leverages Big Data to save US$1 billion per year** on customer retention

Source: https://saasscout.com/big-data-statistics/

Stats, Growth and Facts

1 quintillion = 10^18
1 quintillion = 10^9 billion (or one billion of billions)

Key Stats:

- The **big data market will reach an estimated value of $103 billion by 2023**
- It's estimated **97.2% of companies are starting to invest in big data technology**
- Every day, **internet users create 2.5 quintillion bytes of data**
- IDC's Digital Universe Study from 2012 found that just **0.5% of data was actually being analyzed.**
- It is estimated that by **2021 every person will generate about 1.7 megabytes of data per second**.
- Companies like **Netflix leverages Big Data to save US$1 billion per year** on customer retention

Source: https://saasscout.com/big-data-statistics/

Stats, Growth and Facts

1 quintillion bytes = 1 exabyte = 10^6 TB

Key Stats:

- The **big data market will reach an estimated value of $103 billion by 2023**
- It's estimated **97.2% of companies are starting to invest in big data technology**
- Every day, **internet users create 2.5 quintillion bytes of data**
- IDC's Digital Universe Study from 2012 found that just **0.5% of data was actually being analyzed.**
- It is estimated that by **2021 every person will generate about 1.7 megabytes of data per second**.
- Companies like **Netflix leverages Big Data to save US$1 billion per year** on customer retention

Source: https://saasscout.com/big-data-statistics/

Stats, Growth and Facts

Every day, **internet users create 2.5 quintillion bytes of data**

Why Big Data is growing?

33%	• of the time people spent on social media
16%	• for online TV streaming
16%	• for online movie streaming
13%	• for browsing news
22%	• for all other activities

Facebook
- The active users in 2019 were 2.3 billion

Twitter
- Twitter user send half million tweets every minute

Google
- 40,000 search queries submitted every second

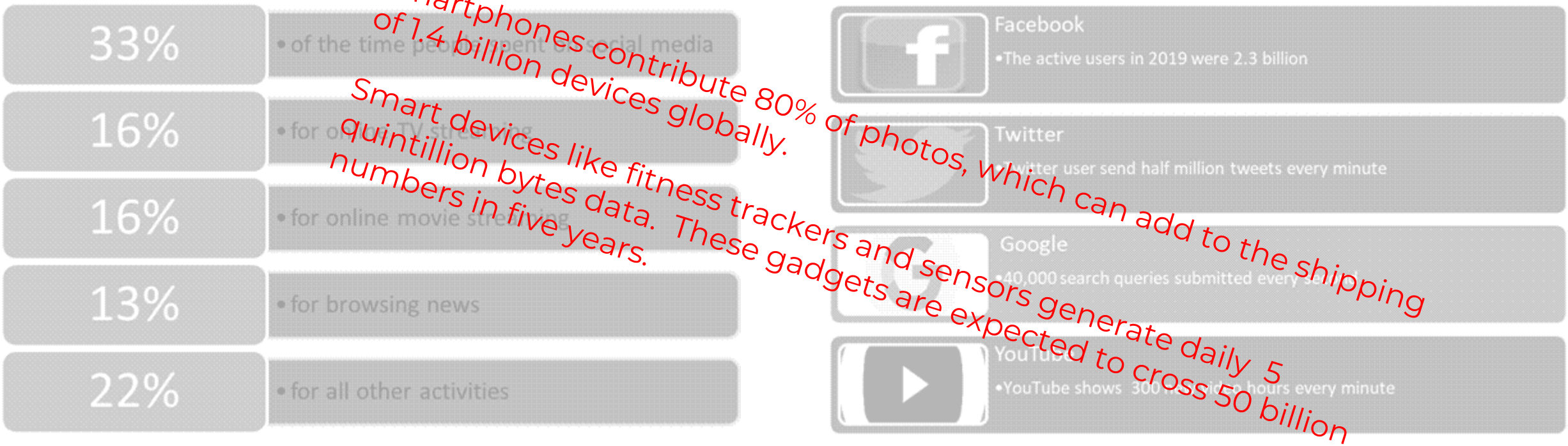
YouTube
- YouTube shows 300 new video hours every minute

Source: https://saasscout.com/big-data-statistics/

Stats, Growth and Facts

Every day, **internet users create 2.5 quintillion bytes of data**

Why Big Data is growing?



Smartphones contribute 80% of photos, which can add to the shipping of 1.4 billion devices globally.

Smart devices like fitness trackers and sensors generate daily 5 quintillion bytes data. These gadgets are expected to cross 50 billion numbers in five years.

Source: https://saasscout.com/big-data-statistics/

BIG DATA: DEFINITION AND CHARACTERISTICS

What is Big Data?

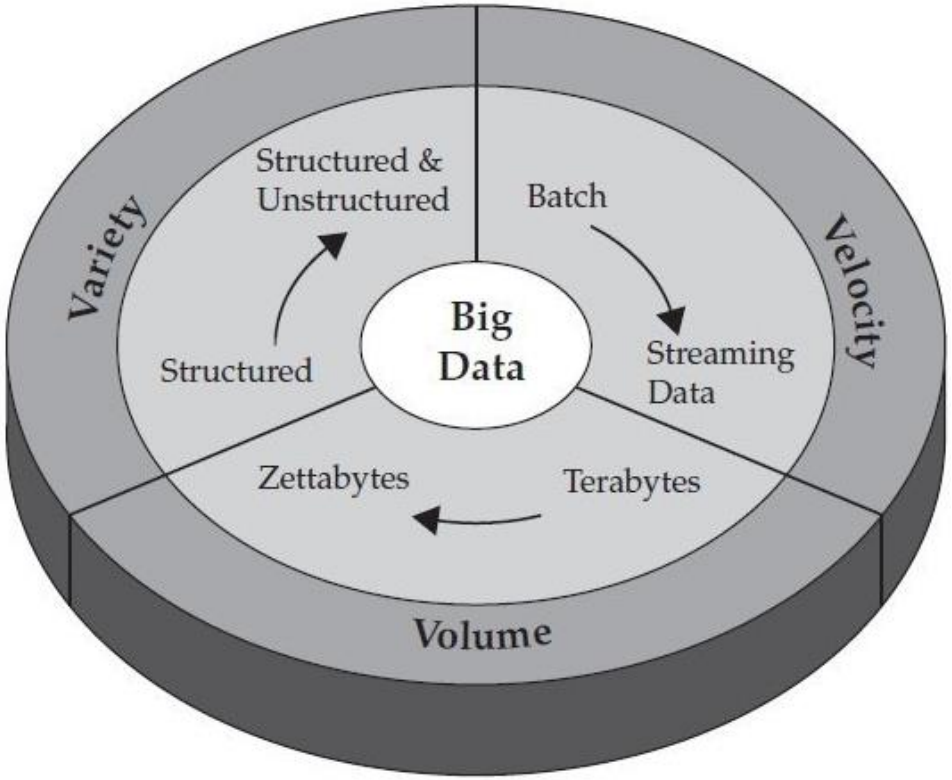
- Definition 1: “Sets of information that are too large or too complex to handle, analyze or use with standard methods” (Oxford Dictionary)

- Definition 2: “Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.” (Gartner)

The V3 Characterization of Big Data

Initial definition of Big Data made use of three specific characteristics:

- **Volume**–The data being used is a large source compared to other datasets
- **Variety**–Data which comes from diverse sources, different formats, generated by people as well as computers
- **Velocity**–Data are generated very quickly, and probably continuously

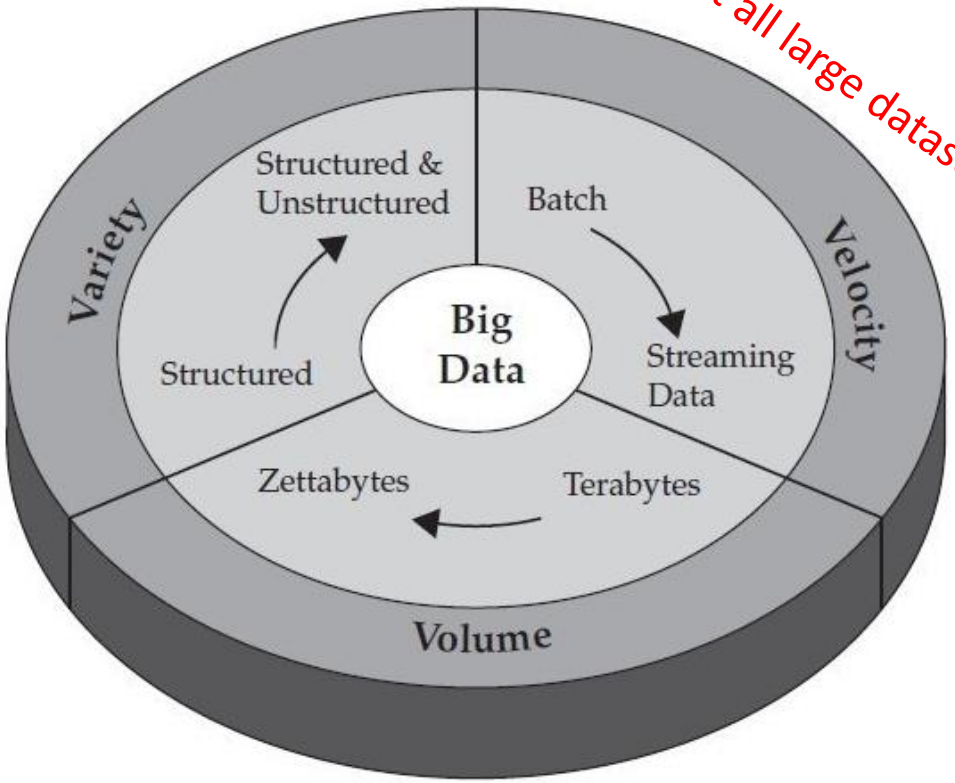


"Understanding Big Data", IBM
https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/Big+Data+University/page/FREE+ebook+-+Understanding+Big+Data

The V3 Characterization of Big Data

Note: Not all large datasets are big data

- When talking about big data, there are three Vs which can be used to categorize its value
  - **Volume**–The data being used is a large source compared to other datasets
  - **Variety**–Data which comes from diverse sources, generated by people as well as computers
  - **Velocity**–Data are generated very quickly, and probably continuously



"Understanding Big Data", IBM
https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/Big+Data+University/page/FREE+ebook+-+Understanding+Big+Data

Big Data Volume: How "Big" is Big Data?

- Big data "size" – a constantly moving target
- Ranges from
  - Terabytes (10^{12} bytes) to Petabytes (10^{15} bytes) to Exabytes (10^{18} bytes) to Zettabytes (10^{21} bytes) of data
  - The range can continue (Yottabyte, Xenottabyte, Shilentnobyte, Domegemegrottebyte)
  - Examples
    - 1 TB – all the X-ray films in a large technological hospital
    - 2 PB – all US academic research libraries
    - 1 EB – all words ever spoken by human beings

(http://highscalability.com/blog/2012/9/11/how-big-is-a-petabyte-exabyte-zettabyte-or-a-yottabyte.html)

Big Data Variety: data type

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data – Social Network, Semantic Web (RDF), …
- Streaming Data – You can only scan the data once
- Images, video

Big Data Variety: data type

- Relational Data (Tables/T
- Text Data (Web)
- Semi-structured Data (XI
- Graph Data – Social Netv
- Streaming Data – You ca
- Images, video

The type of data generated and stored varies by sector[1]

	Video	Image	Audio	Text/ numbers
Banking	Medium	Medium	Medium	High
Insurance	Low	Low	Low	High
Securities and investment services	Low	Low	Low	High
Discrete manufacturing	Medium	Medium	Low	High
Process manufacturing	Medium	Medium	Low	High
Retail	Medium	Low	Low	High
Wholesale	Low	Low	Low	High
Professional services	Medium	Medium	Medium	High
Consumer and recreational services	Medium	Low	Medium	Low
Health care	Low	High	Low	High
Transportation	Low	Low	Low	High
Communications and media[2]	High	Low	High	High
Utilities	Low	Low	Low	High
Construction	Low	Low	Low	Low
Resource industries	Medium	Medium	Low	High
Government	High	Low	High	High
Education	High	Low	High	Low

Penetration: High, Medium, Low

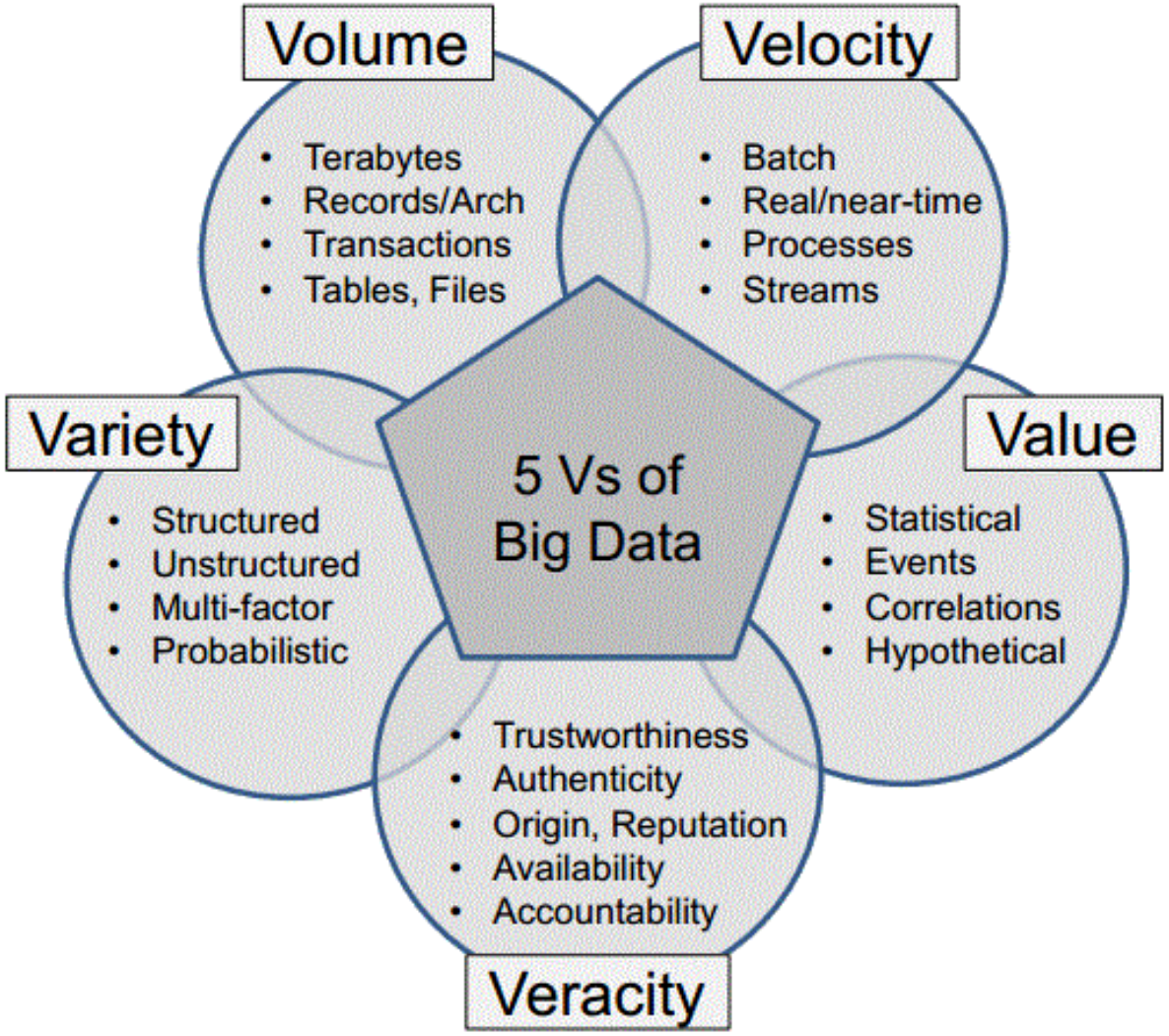
1 We compiled this heat map using units of data (in files or minutes of video) rather than bytes.
2 Video and audio are high in some subsectors.
SOURCE: McKinsey Global Institute analysis

Big Data Velocity

- How long does it take to do something with it or even know it has been recorded?

- e-commerce purchases,
- weather events,
- utility service usage,
- geo-location of people and things,
- server activity

The V5 Characterization of Big Data



Veracity refers to the quality or trustworthiness of the data.

Value refers to the ability to transform data into business value.

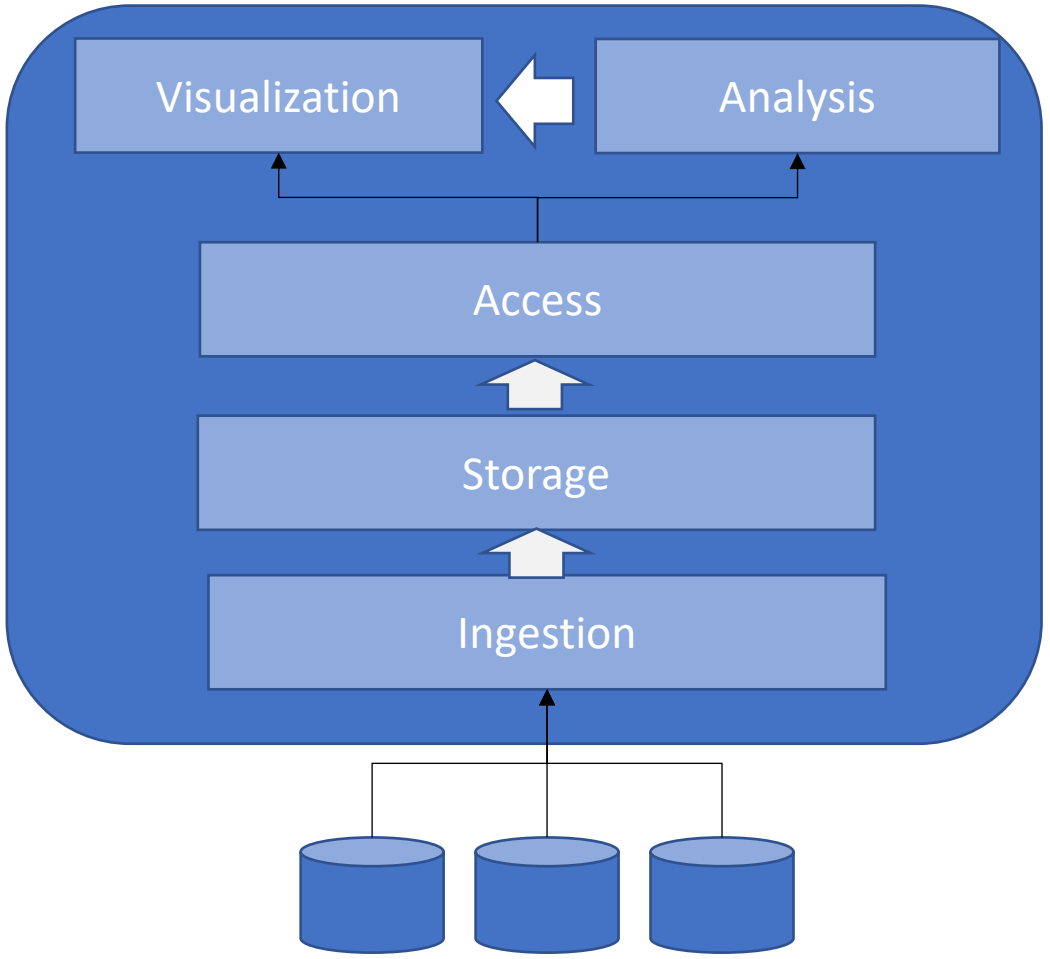
The Extended 3+n Vs of Big Data

- 1. **Volume**
- 2. **Variety**
- 3. **Velocity**
- 4. **Veracity** - If being used for business value, care has to be taken that the data being used are accurate and of good quality. Particularly if coming from many different sources
- 5. **Value**
- 6. **Venue** (location)
- 7. **Vocabulary** (semantics)
- 8. V…
- 9. V…
- … V…

Big Data Challenges

Big Data ecosystem: main technological components and tools

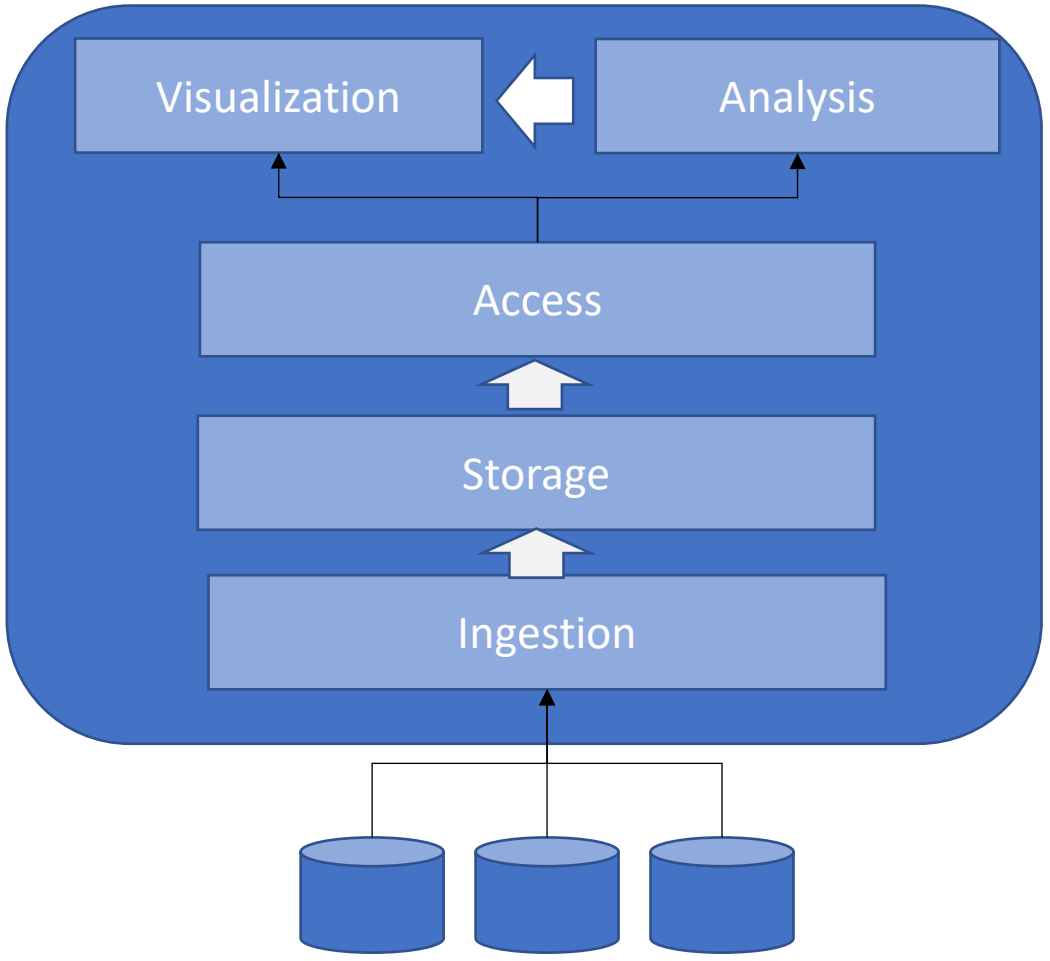
Big Data Ecosystem: main components



Data sources:

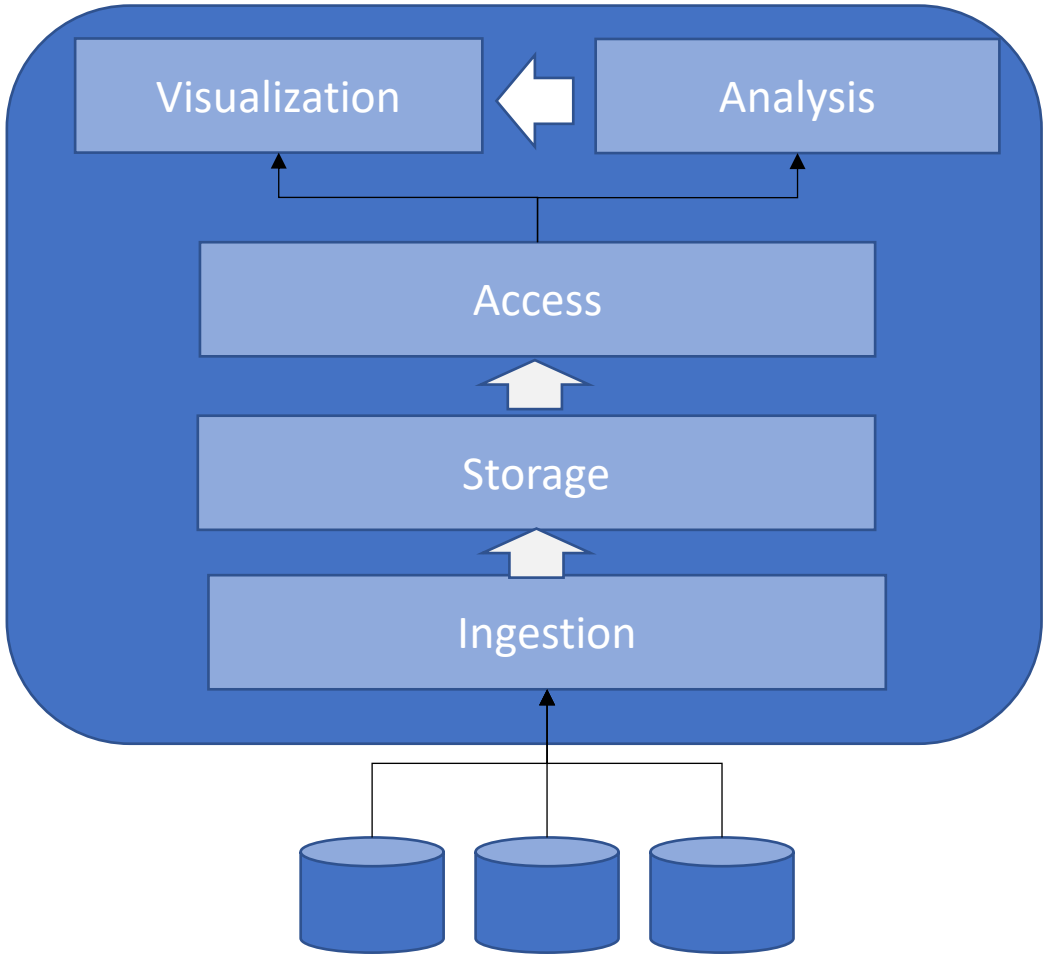
- Existing data storages,
- Sensors,
- Machines,
- Smart devices
- File repositories

Big Data Ecosystem: main components



Data ingestion is the process of obtaining and importing **data** for immediate use or storage in a **file system/database**. **Data** can be streamed in real time or **ingested** in batches.

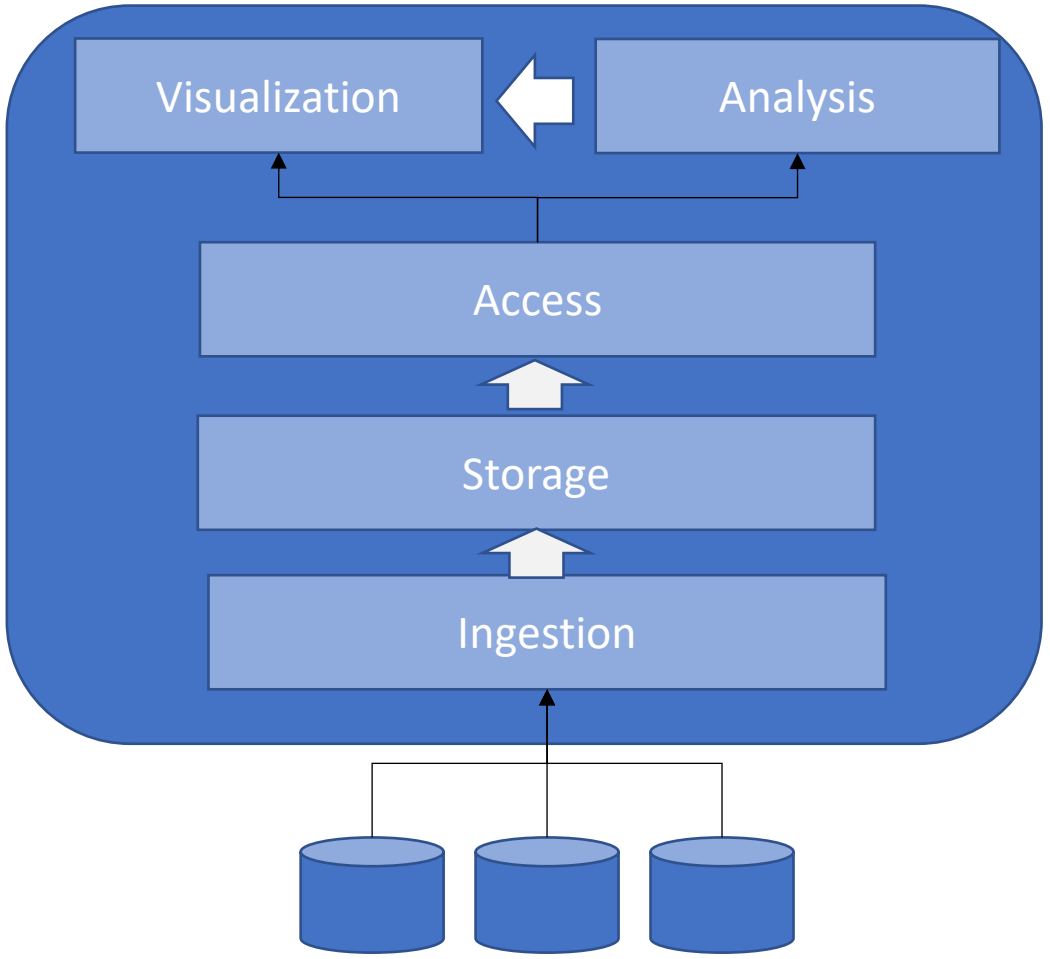
Big Data Ecosystem: main components



Big data storage is a storage infrastructure that is designed specifically to store, manage and retrieve massive amounts of data, or big data. Big data storage enables the storage and sorting of big data in such a way that it can easily be accessed, used and processed by applications and services working on big data. Big data storage is also able to flexibly scale as required.

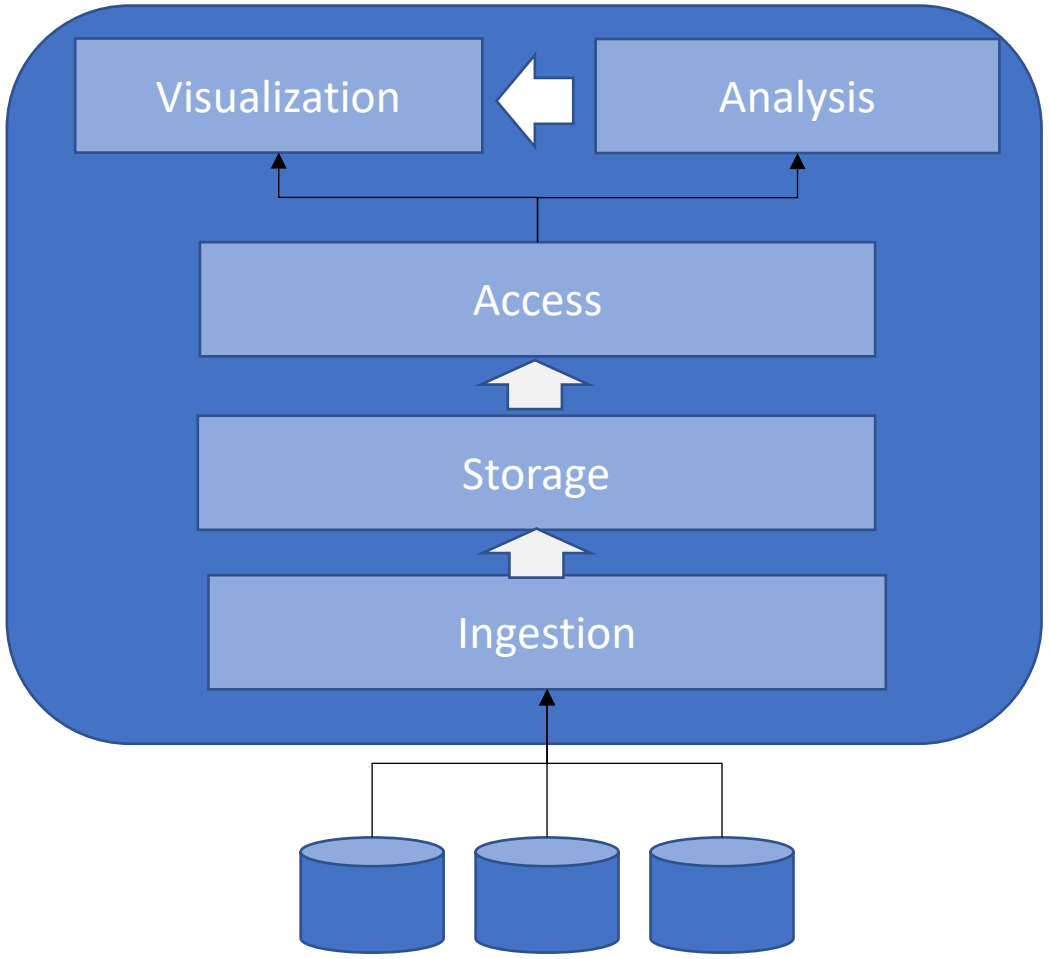
Source: https://www.techopedia.com/definition/29473/big-data-storage

Big Data Ecosystem: main components



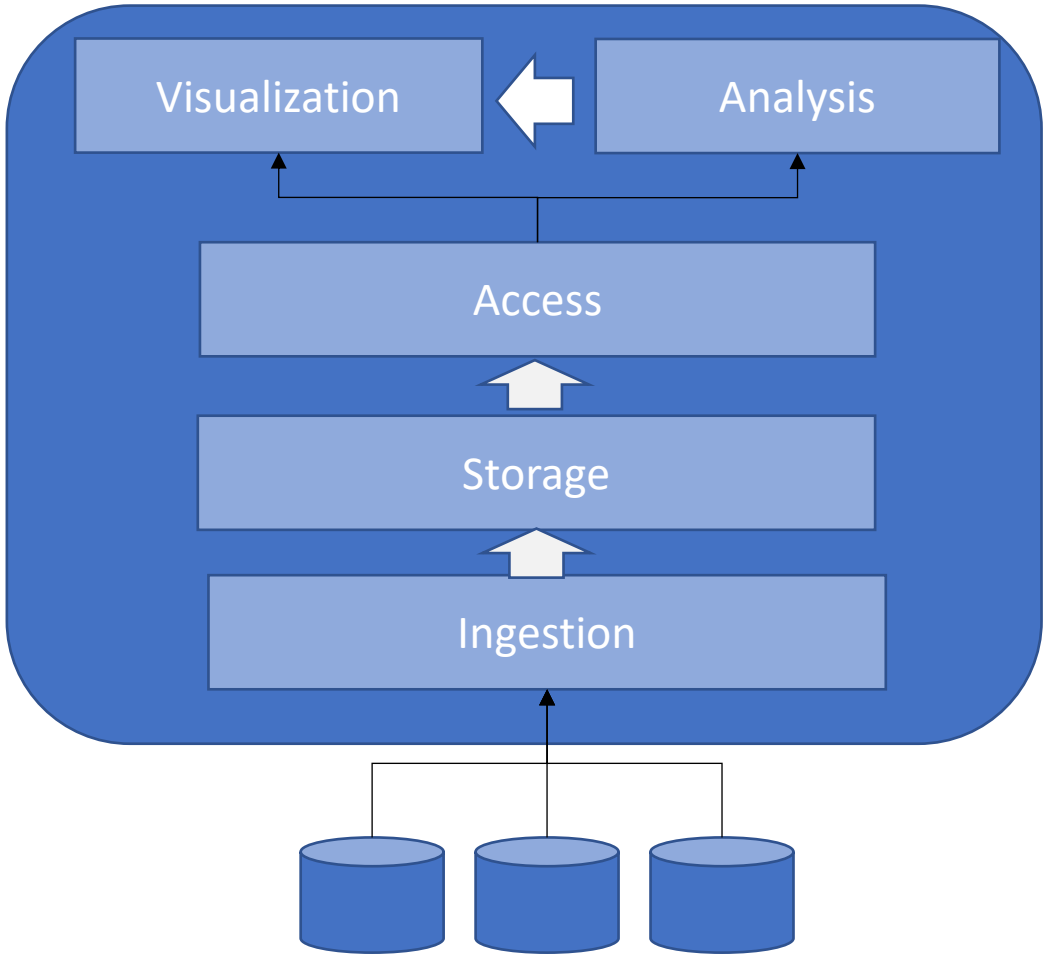
Big Data Access enables fast data integration, aggregation and retrieval (or creates and enables access to customized data views from the master data stored in the storage layer)

Big Data Ecosystem: main components



Big Data Analysis facilitates the analysis of customized data views provided via the access layer and also enables the automatization of decisions made based on the results of analysis.

Big Data Ecosystem: main components



Big Data Visualization facilitates visual inspection of customized data views but also the results of the analysis. It is usually the interface between the data ecosystem and the end-user.

Big Data Ecosystems: A High Level Overview of Existing Technologies

Pythian's CTO Alex Gorbachev’ overview of Big Data Ecosystems.

https://www.youtube.com/watch?v=aH-Ixpo4MSA&ab_channel=Pythian

BIG DATA TOOLS

Big Data Tools - A Landscape

Source: https://datafloq.com/big-data-open-source-tools/os-home/

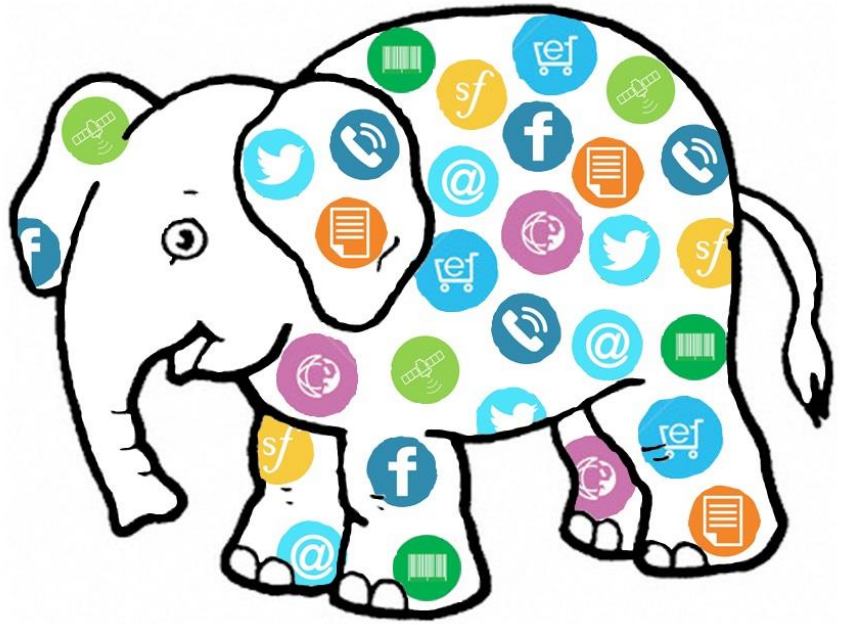
Tools Typically Used in Big Data Scenarios

- **Where processing is hosted**
  - Distributed Servers / Cloud (e.g. Amazon EC2)
- **Where data is stored**
  - Distributed Storage (e.g. Hadoop, Amazon S3)
- **What the programming model is**
  - Distributed Processing (e.g. MapReduce)
- **How data is stored & indexed**
  - High-performance schema-free databases (e.g. MongoDB)
- **What operations are performed on data**
  - Analytic / Semantic Processing / Visualization

BIG DATA MARKET

Key Sources for Big Data

- Data can either be created by people (e.g., social networks) or generated by machines (e.g., sensor data, satellite imagery, purchase transactions records, etc.)
- Key Big Data Sources:
  - People
  - Organizations
  - Sensors & computer systems
- Big data covers many sectors



Source:
http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data

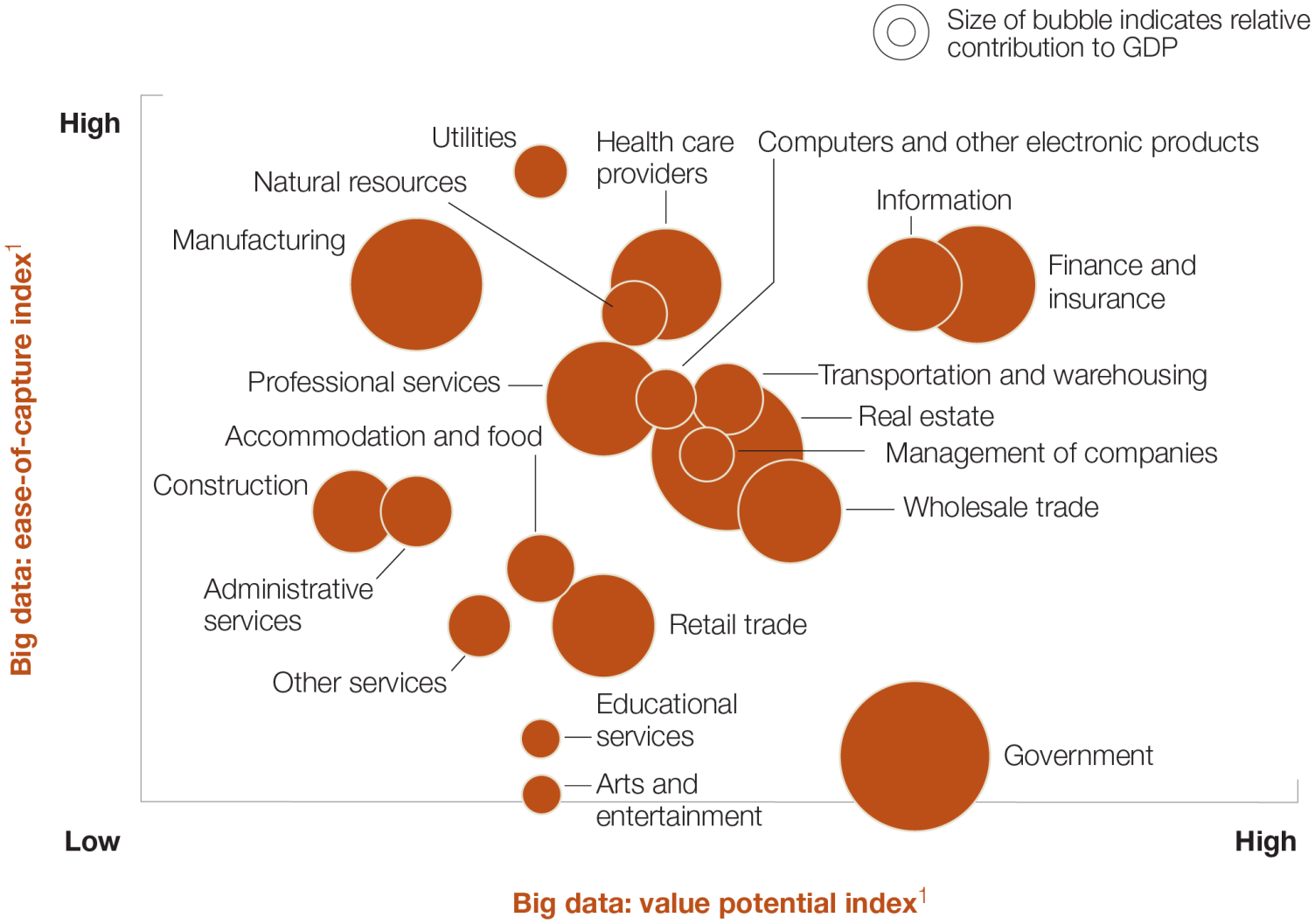
Key Big Data Sources - a possible classification

Key Big Data Sources:

- People (social networks)
  - Social networks
  - Personal documents
  - Pictures (Instagram, Flickr etc.)
  - Videos (Youtube etc.)
  - Internet searches
  - Mobile data
  - E-mail
- Organizations (process-mediated data)
  - Data produced by public agencies (medical records)
  - Data produced by businesses (commercial transactions, banking/stock records, e-commerce, credit cards)
- Sensors and computer systems (Internet of Things data)
  - Fixed sensors (Home automation, weather/pollution sensors, traffic sensors, scientific sensors, security/surveillance videos/images)
  - Mobile sensors (mobile phones, cars, satellite images)
  - Logs
  - Web logs

Source:
http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data

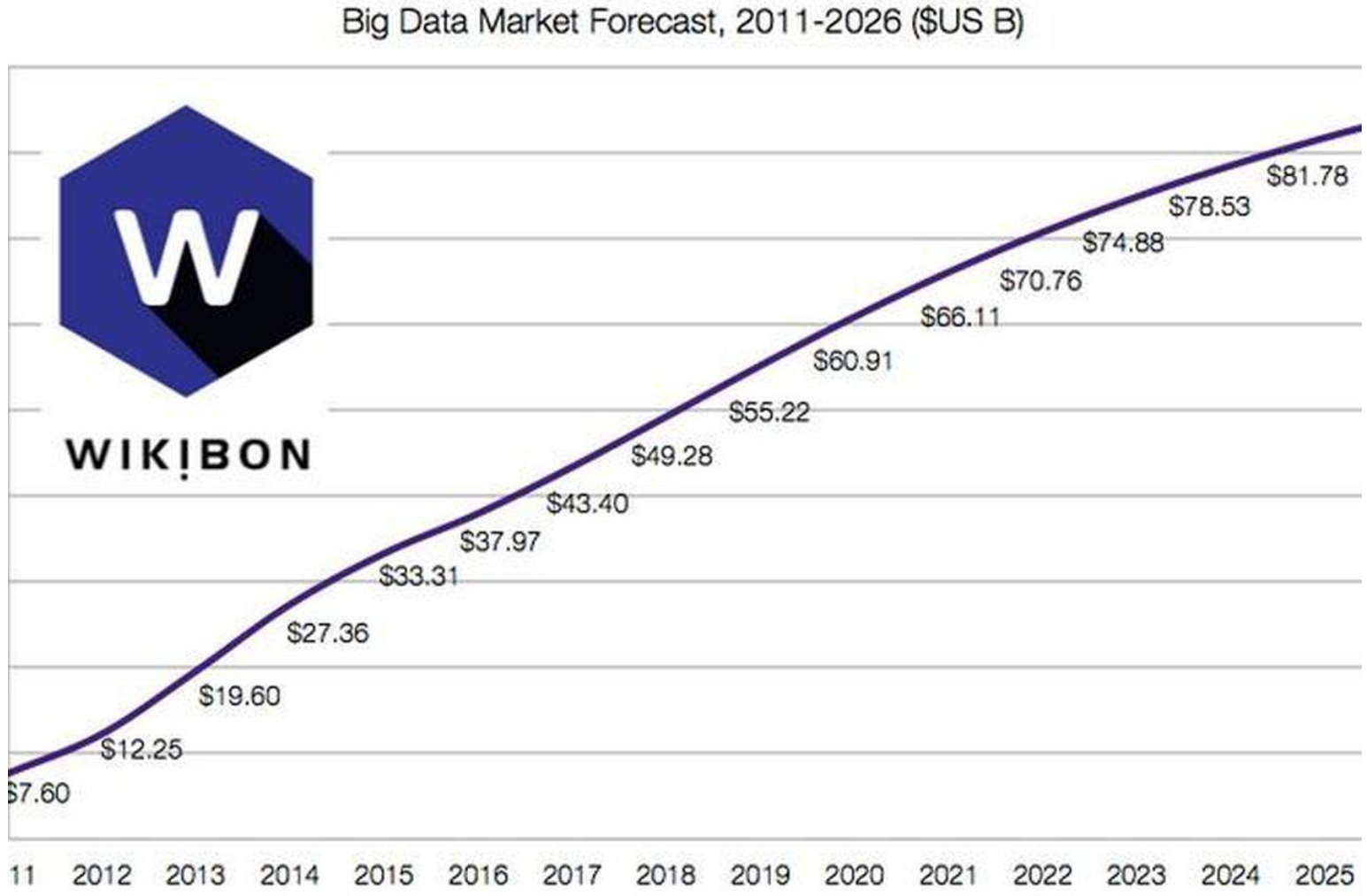
Big Data Sectors



[1]For detailed explication of metrics, see appendix in McKinsey Global Institute full report *Big data: The next frontier for innovation, competition, and productivity*, available free of charge online at mckinsey.com/mgi.

Source: US Bureau of Labor Statistics; McKinsey Global Institute analysis

Big Data Market Forecast (2011-2026)



http://wikibon.org/

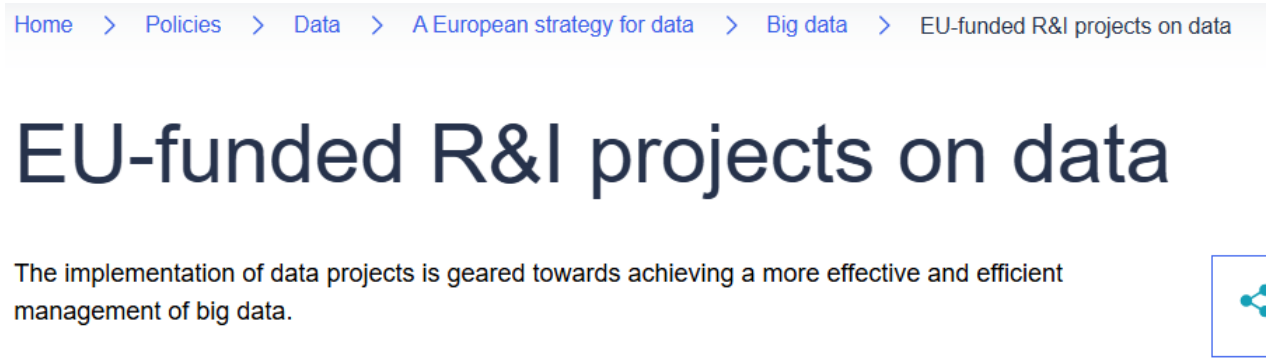
BIG DATA INITIATIVES

BIG DATA INITIATIVES

- EU funded

BIG DATA INITIATIVES

- EU funded



Home > Policies > Data > A European strategy for data > Big data > EU-funded R&I projects on data

EU-funded R&I projects on data

The implementation of data projects is geared towards achieving a more effective and efficient management of big data.

Horizon Europe Programme

The **WORLD LEADING DATA AND COMPUTING TECHNOLOGIES 2023 – 2024** portfolio includes projects from the HORIZON EUROPE WP 2023-2024, 7. Digital, Industry and Space, following topics:

2024

Topic: HORIZON-CL4-2024-DATA-01-01: AI-driven data operations and compliance technologies (AI, data and robotics partnership) (IA) – (4 projects)

2023

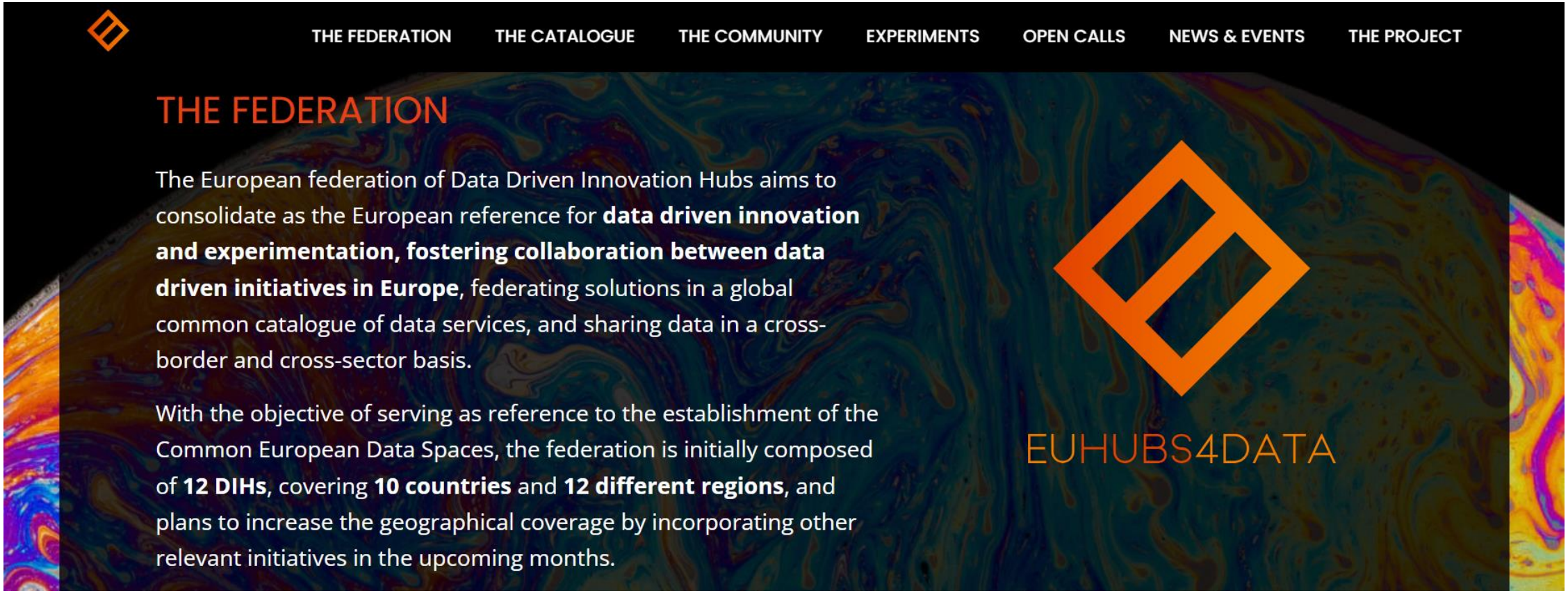
Topic: HORIZON-CL4-2023-DATA-01-02: Integration of data life cycle, architectures and standards for complex data cycles and/or human factors, language (AI, data and robotics partnership) (RIA) – (5 projects)

In addition, 3 projects under the call **A HUMAN-CENTRED AND ETHICAL DEVELOPMENT OF DIGITAL AND INDUSTRIAL TECHNOLOGIES** (Destination 6), are included:

Topic: HORIZON-CL4-2023-HUMAN-01-01: Efficient trustworthy AI - making the best of data (AI

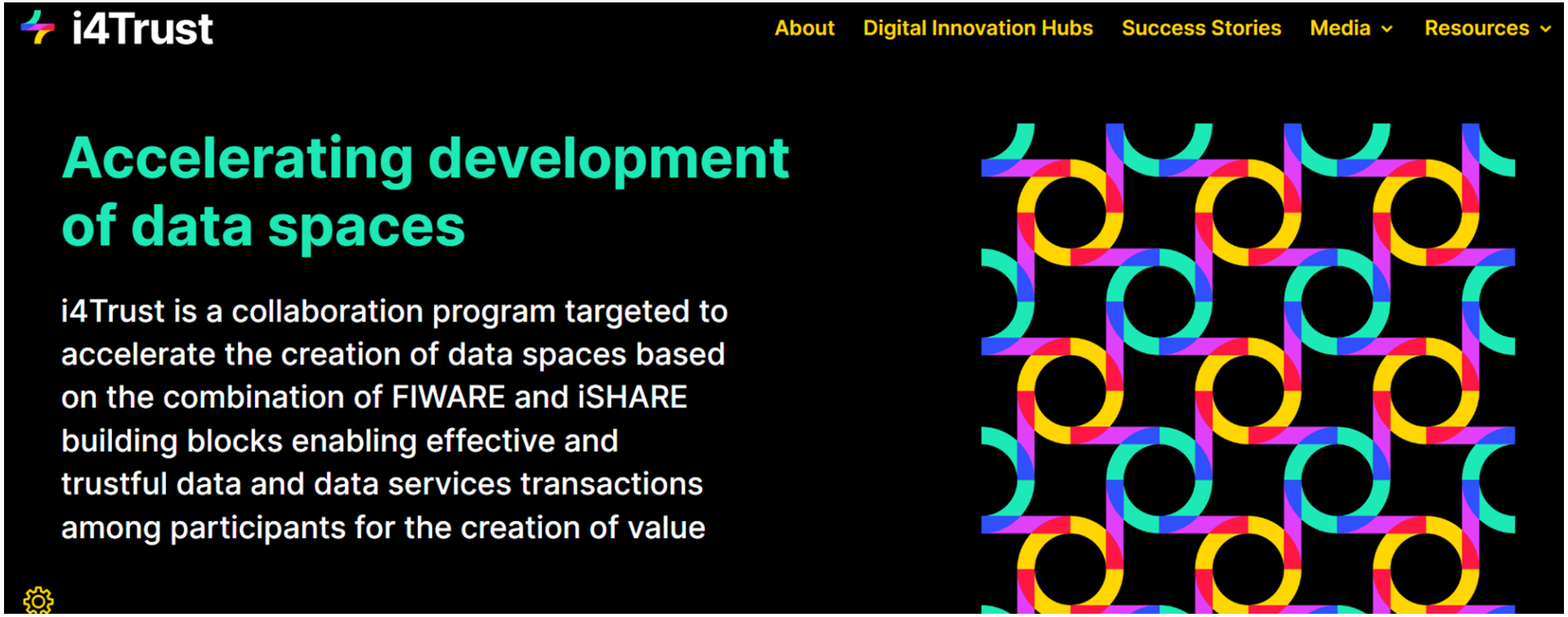
Relevant Big Data EU Projects & Initiatives

- European Federation of Data Driven Innovation Hubs (EUHubs4Data), https://euhubs4data.eu/



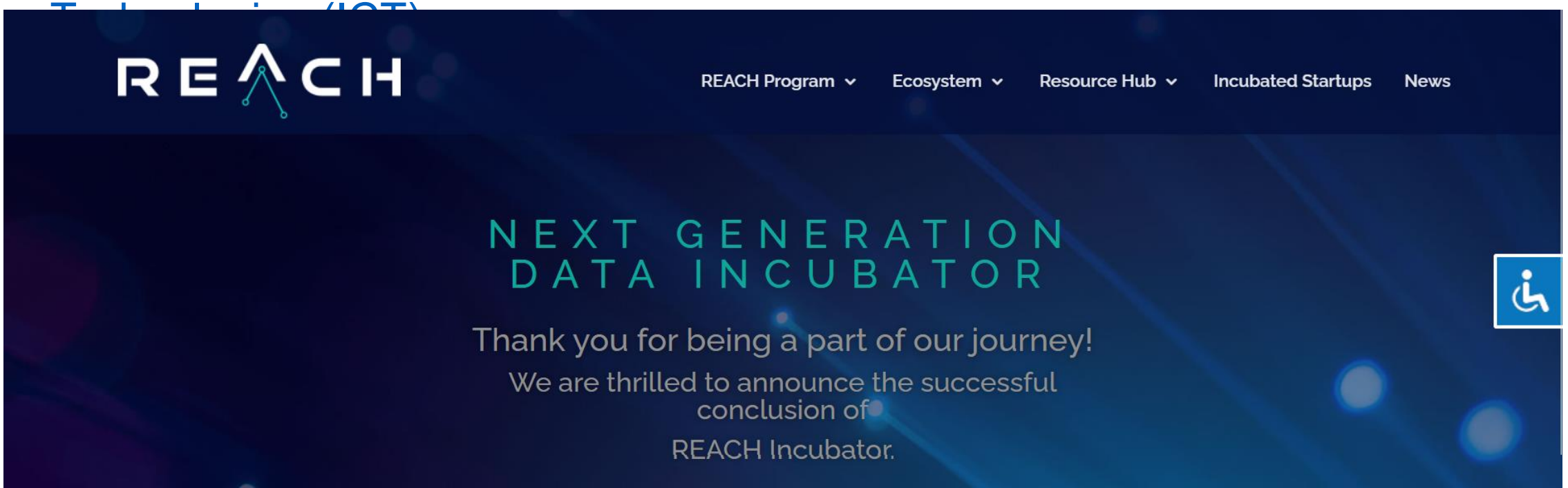
Relevant Big Data EU Projects & Initiatives

- Incubator of Trusted B2B Data Sharing ecosystems of collaborating SMEs linked to Digital Innovation Hubs (i4Trust), https://i4trust.org/



Relevant Big Data EU Projects & Initiatives

- REACH EuRopEAn incubator for trusted and secure data value CHains, https://www.reach-incubator.eu/
- Programme: H2020-EU.2.1.1. - INDUSTRIAL LEADERSHIP - Leadership in enabling and industrial technologies - Information and Communication

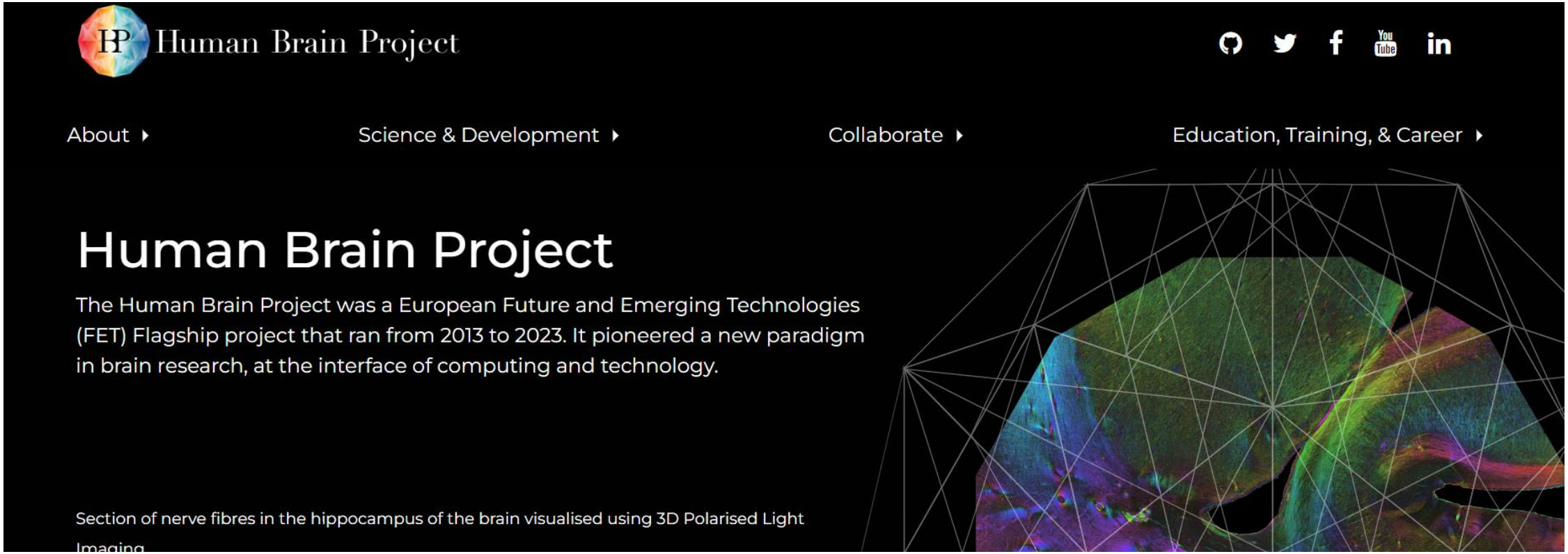


Relevant Big Data EU Projects & Initiatives

- FARE (https://cordis.europa.eu/project/id/853566)

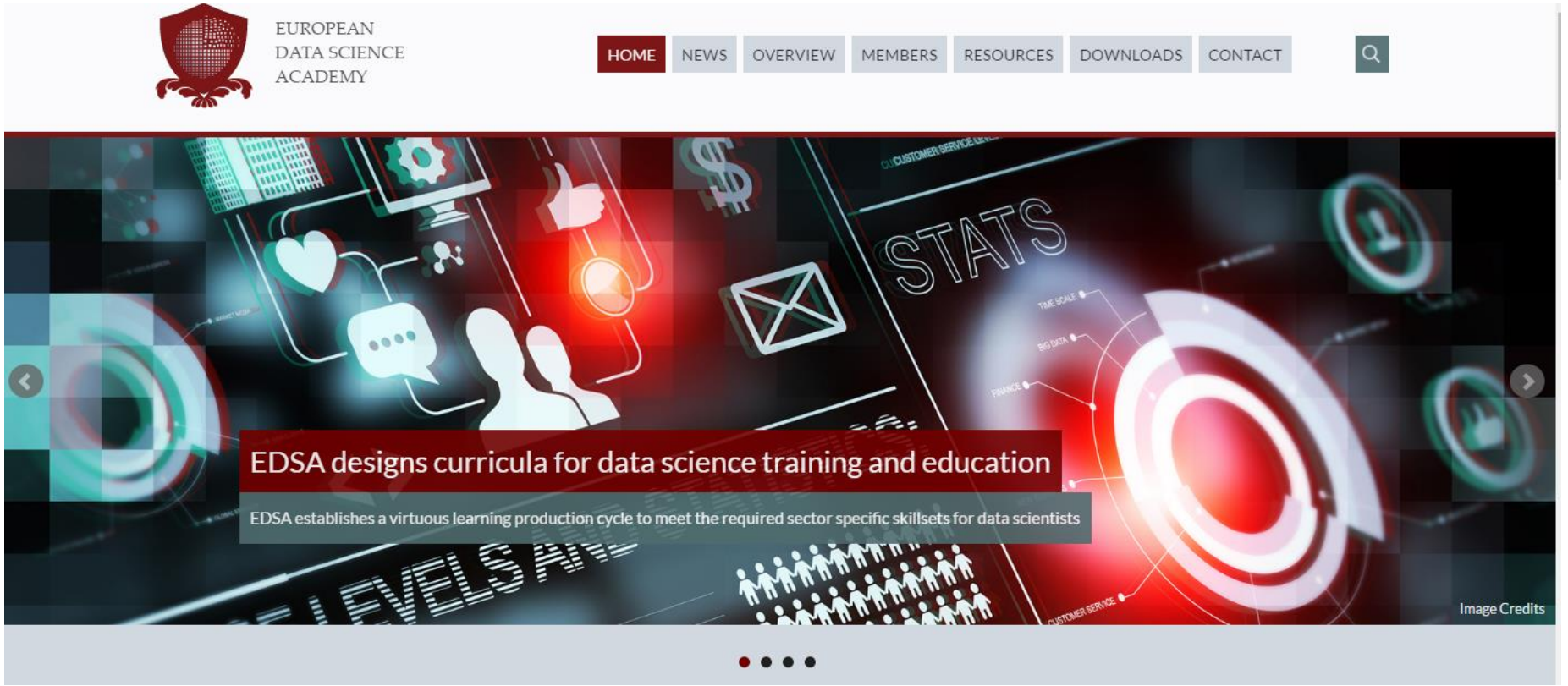
Relevant Big Data EU Projects & Initiatives

- Human Brain Project, https://www.humanbrainproject.eu/en/



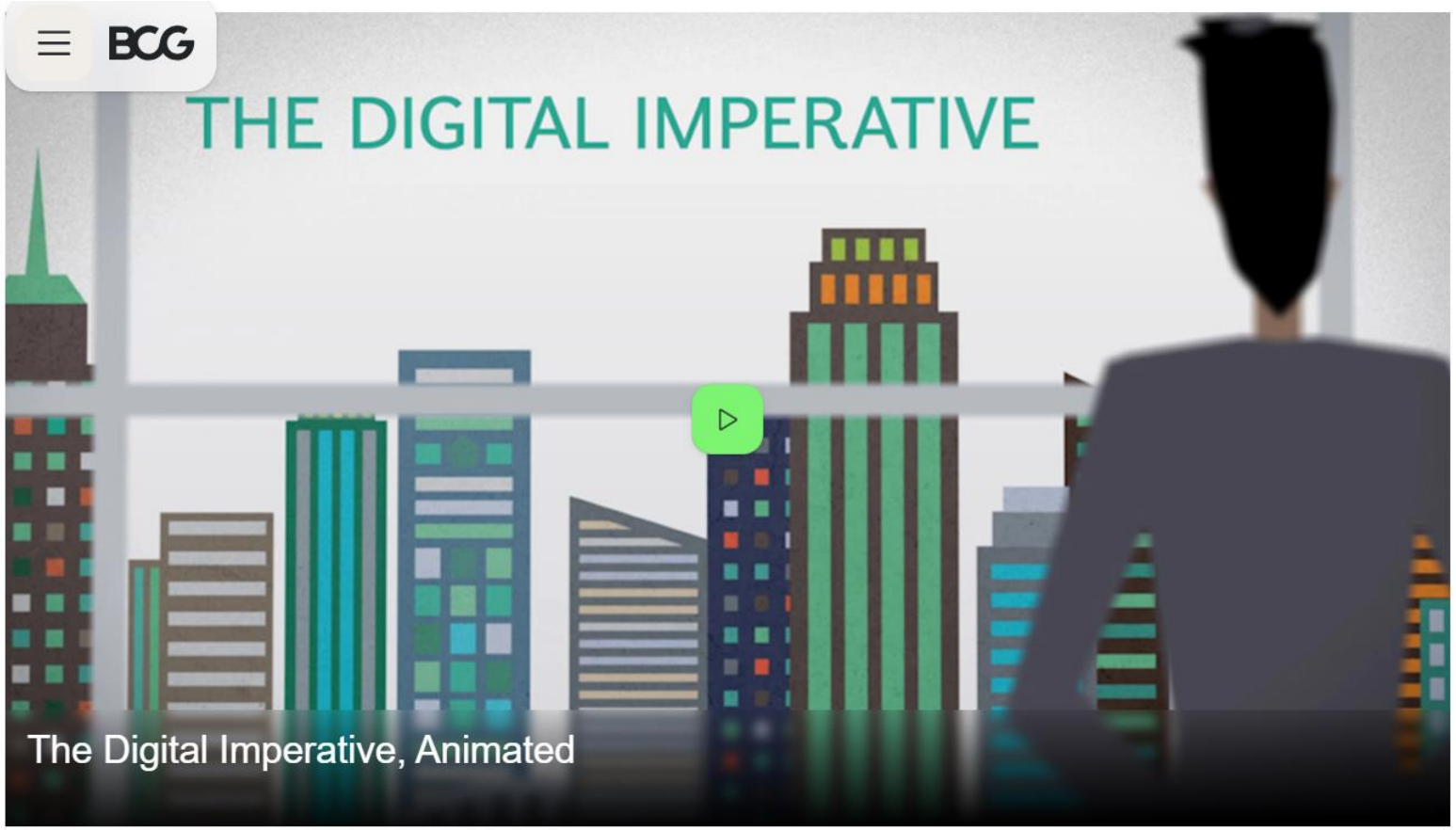
Relevant Big Data EU Projects & Initiatives

- European Data Science Academy, https://edsa-project.eu/



The role of Big Data in Accelerating the Digitalization of Industries

https://www.bcg.com/publications/2015/digital-imperative



Foundations of Big Data: take home

- Definition and main characteristics (3V / 5V)
- Why Big Data is important?
- Main Big Data producers
- Big Data Ecosystem
- Big Data Tools or Implementation of a Big Data Ecosystem

Next course

- Course 3 - Data systems and the lambda architecture for big data

References

- Part of the slides are adapted from EDSA learning materials, https://edsa-project.eu/resources/courses
- Cielen, D., Meysman, A.D.B., & Ali, M. (2016). Introducing Data Science. Big Data, machine learning, and more, using Python tools. Manning Publications
- O'Neil and Schutt, *Doing Data Science -Straight Talk from the Frontline*, O'Reilly, 2014 (Chapter 1)
- https://www.kaggle.com/wiki/DataScienceUseCases
- http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram
- Russell Jurney, *Agile Data Science*, O'Reilly, 2013 (Chapters 1 and 2)
- Thomas Davenport, *Big Data @ Work*, Harvard Business Review Press, 2014
- Jeffrey Stanton, Syracuse University, An Introduction to Data Science https://docs.google.com/file/d/0B6iefdnF22XQeVZDSkxjZ0Z5VUE/edit?pli=1
- Five Big Data Challenges, SAS

https://www.sas.com/resources/asset/five-**big**-**data**-**challenges**-article.pdf

- Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, 2011
- Cosmin Lazar, Standard processes for data science & AI, Invited lecture for "Big Data Processing and Applications", UBB