Big Data, Techniques and Platforms
Cours 1/8 : Introduction

Francesca Bugiotti

CentraleSupélec

October 3, 2023

Objectives

- Describe the Big Data landscape
- Identify the high level components in the data science lifecycle and the associated data flow
- Explain the V's of Big Data and why each impacts the collection, monitoring, storage, analysis, and reporting
- Identify Big Data problems

Plan

What, Who, Why, Which

- What is Data Science?

What, Who, Why, Which

- What is Data Science?
- Who is a Data Scientist?

What, Who, Why, Which

- What is Data Science?
- Who is a Data Scientist?
- Why Big Data?

What, Who, Why, Which

- What is Data Science?
- Who is a Data Scientist?
- Why Big Data?
- Which is the relation between Data Science and Big Data?

Data Science

McKinsey - 2013 [2]

- Data science catalyst for economic growth

Data Science

McKinsey - 2013 [2]

- Data science catalyst for economic growth

Two ingredients:

Data Science

McKinsey - 2013 [2]

- Data science catalyst for economic growth

Two ingredients:

- Data
  - Common devices and specialized devices
  - New data sources producing torrent of data

Data Science

McKinsey - 2013 [2]

- Data science catalyst for economic growth

Two ingredients:

- Data
  - Common devices and specialized devices
  - New data sources producing torrent of data
- Cloud Computing
  - Computing anywhere and anytime
  - On-demand Computing

Big Data

Teradata Magazine article, 2011

"Big Data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it within a tolerable elapsed time for its user population."

Big Data

Teradata Magazine article, 2011

"Big Data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it within a tolerable elapsed time for its user population."

The McKinsey Global Institute, 2012

"Big Data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze."

Big Data

Teradata Magazine article, 2011

"Big Data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it within a tolerable elapsed time for its user population."

The McKinsey Global Institute, 2012

"Big Data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze."

Wikipedia, 2016

"Big Data is a collection of data sets so large and complex that it becomes difficult to process them using on-hand database management tools."

When does Data become Big?

When does Data become Big?

How many data in the world?

- 800 Terabytes, 2000
- 160 Exabytes, 2006 (1EB = 10^{18}B)
- 4.5 Zettabytes, 2013 (1ZB = 10^{21}B)
- 44 Zettabytes by 2020

How much is a zettabyte?

- 1,000,000,000,000,000,000,000 bytes
- A stack of 1TB hard disks that is 25,400 km high

How many data in a day?

- 7 TB, Twitter
- 10 TB, Facebook

When does Data become Big?

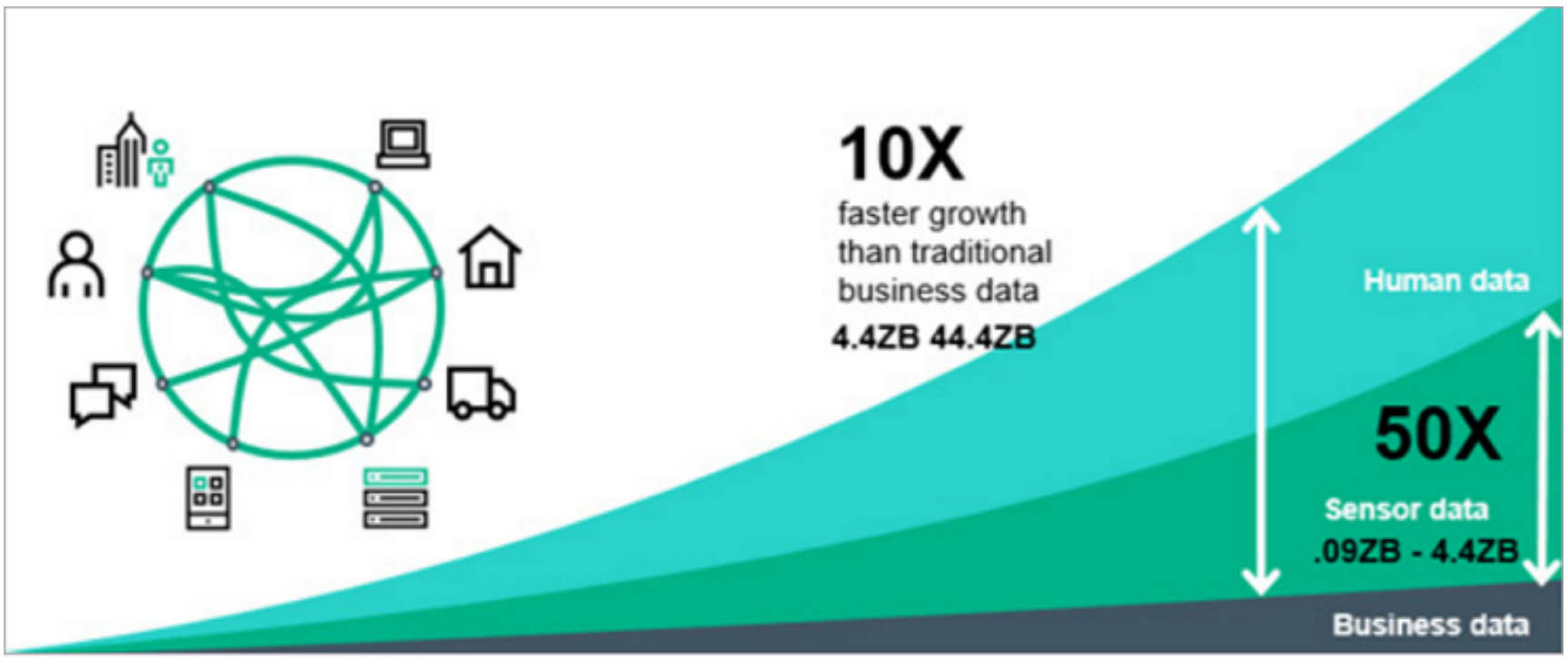
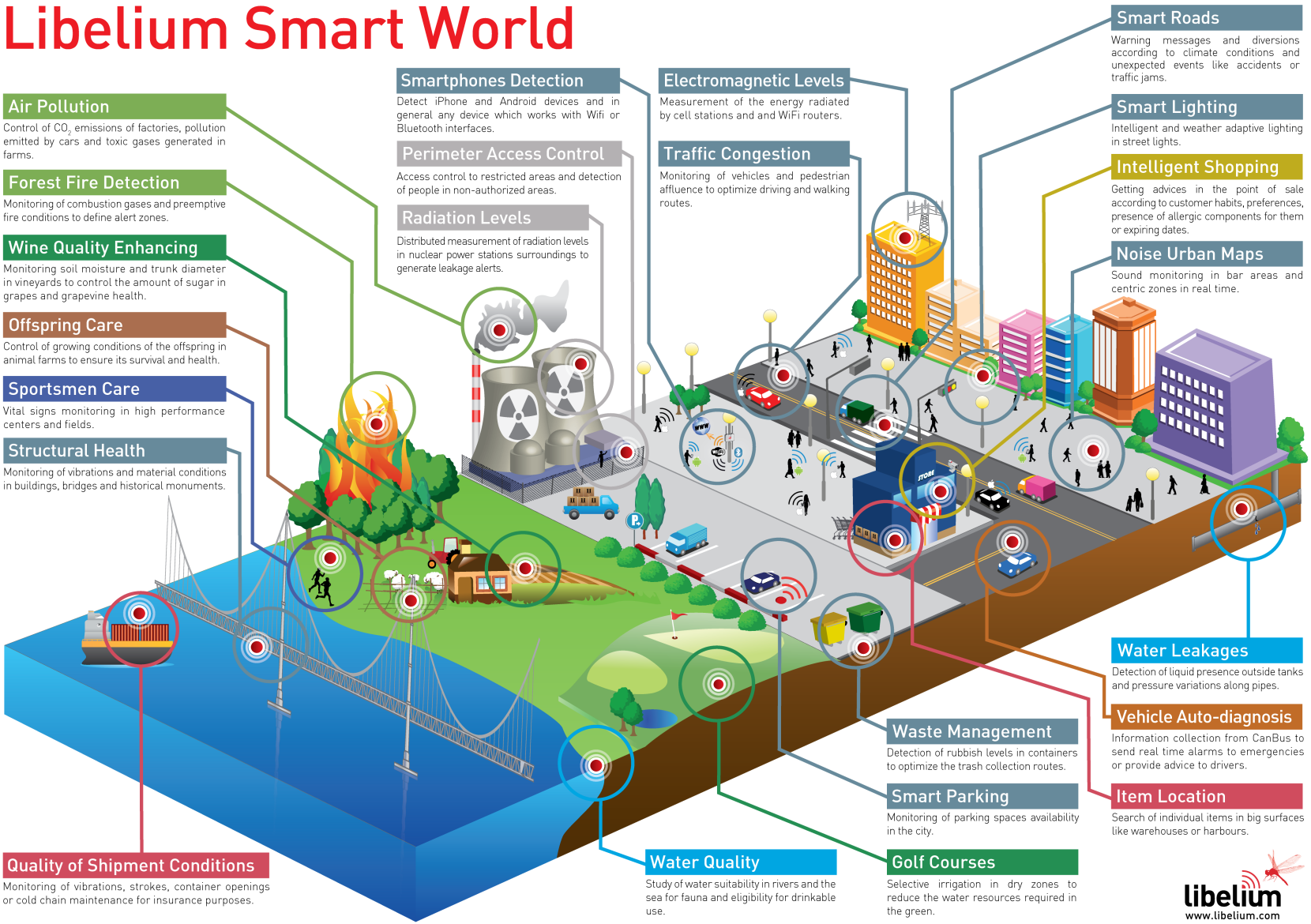


Figure: The Exponential Growth of Data [1]

When does Data become Big?



More on Data

It is reductive and not useful to use and analyze single data sources, real value comes from combining multiple data sources

Combining data from different sources

- Machines
  - real-time data, sensor data, trackers, streaming, etc.
- People
  - social media, personal documents, etc.
- Organizations
  - structured knowledge bases, data-warehouses, etc.

More on Data

Precision Medicine

Emerging area of medicine targeted toward an individual person
Goal: analyzing her genetics, her environment, her daily activities to detect or predict a health problem

- Machines
  - Real-time electronic health record systems
  - 200GB to store a genome
  - Fitness devices
- People
  - Fitness devices, social media, etc.
- Organizations
  - Public health databases
  - Knowledge bases

More on Data

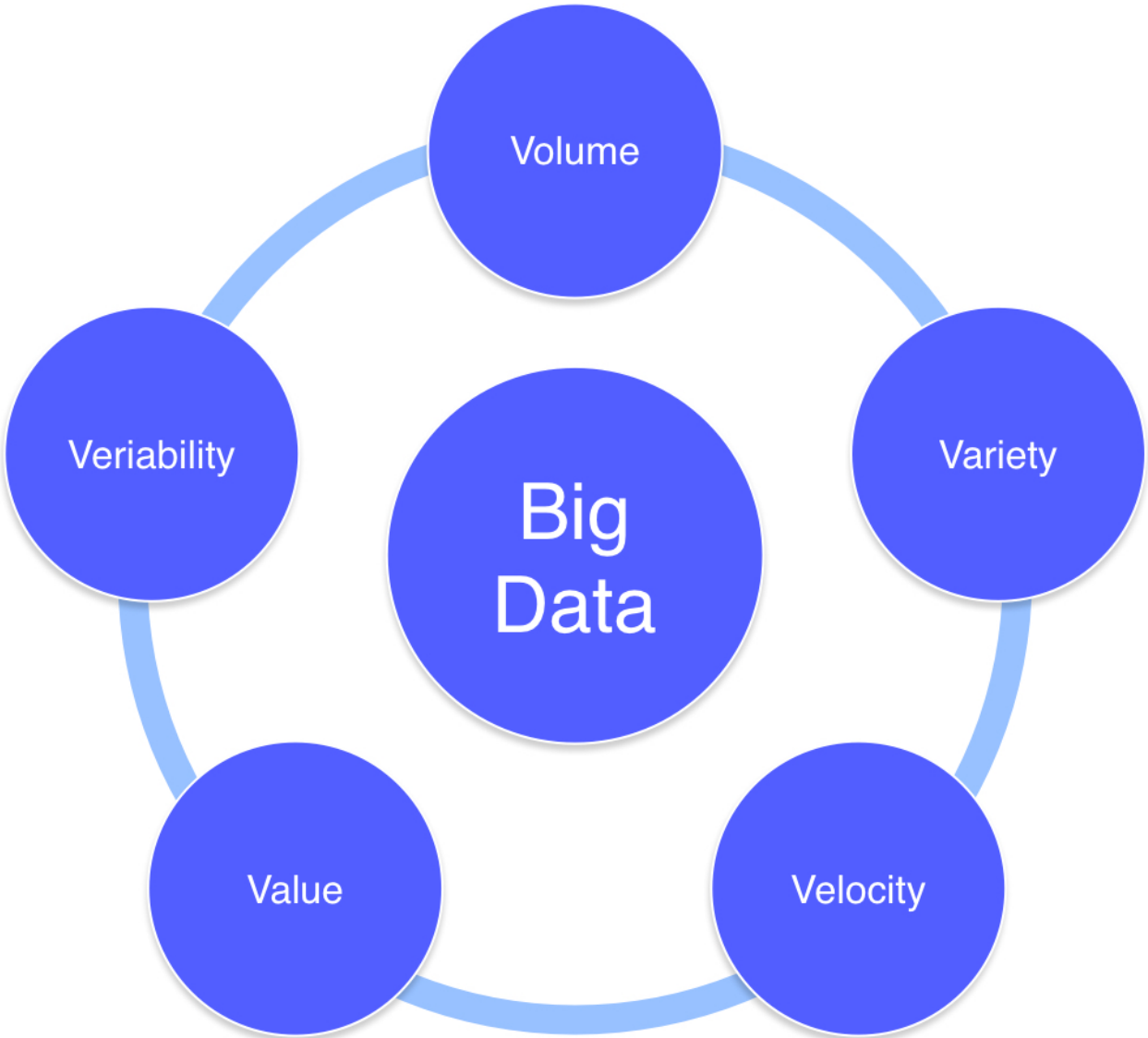
WildFires prediction and Emergency Response

May 2014, in San Diego County: 14 fires burning, total of 26,000 acres destroyed, six people were injured and one person died

- Machines
  - Sensors and satellites, instruments that measure environmental factors such as temperature, humidity, air pressure, images, etc.
- People
  - Data on social media sites such as Twitter, which support photo sharing resources
- Organizations
  - Area maps, vegetation type, presence of fuel, houses, etc.

Plan

Big Data: the 5 Vs



Volume

It is the Big Data dimension that refers to the size of Big Data

Vast amount of data generated every second/minute/hour

- CERN's large hadron collider generates 15 petabytes a year
- Every minute 204 million emails are sent, 200,000 photos are uploaded, and 1.8 million likes are generated on Facebook
- On YouTube, 1.3 million videos are viewed and 72 hours of video are uploaded

Words that start with peta, exa, and yotta

Volume

Challenges

Volume is not just size but also exponential growth

- Store data
  - Data networking
  - Scale data
  - Cost of data storage
- Access data and capture data information
- Process data

Variety

This Big Data dimension refers to the diversity of the data

Increasing different forms of data are collected, stored, and analyzed to solve real world problems

- Image data, text data, network data, geographic maps, computer generated simulation, etc.

Different aspects of variety:

- Semantic variety
  - Age can be a number or a term like infant, teenager, or adult
- Media variety
  - The transcript and the video of the same speech
- Structural variety
  - A table, a graph, a sequence of bytes, etc.
- Availability
  - Data can be accessible continuously or intermittently

Velocity

This Big Data dimension refers to the data production speed

Speed

- Speed at which Big Data is created
- Speed at which data needs to be stored and analyzed

Match the speed of processing with the speed of information generation

- Static data vs data coming from sensors and smart devices
- Real-time processing vs batch processing

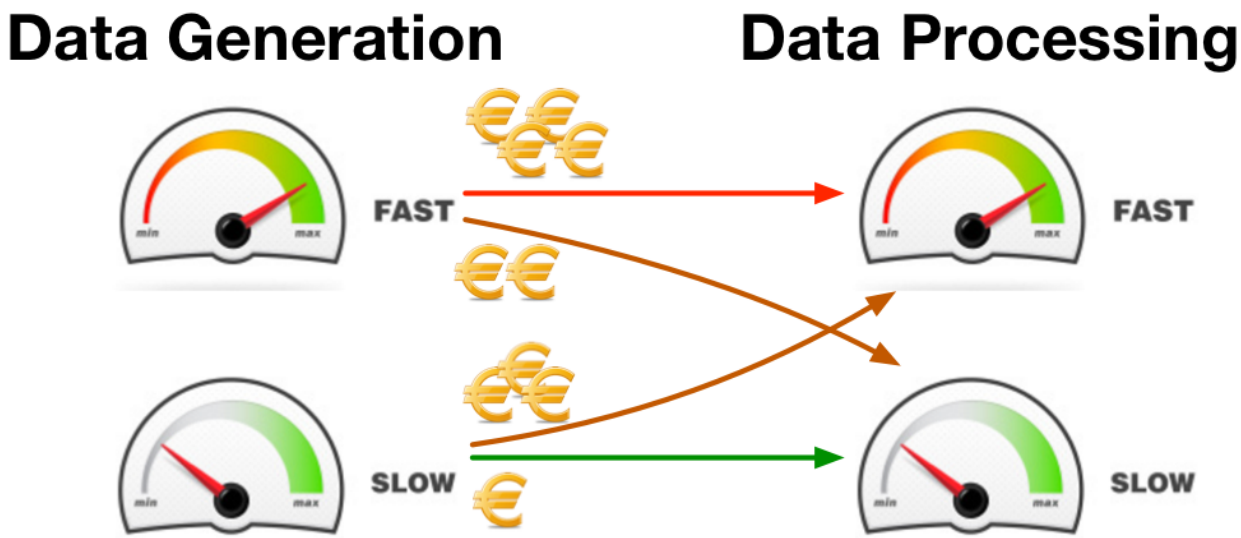
Velocity

Use cases

Information is streaming and needs to be integrated with existing data to produce decisions:

- Emergency response planning in a wildfire
- Deciding trading strategies in real time
- Getting estimates in advertising

Velocity



Veracity

Veracity of Big Data refers to the quality of the data

- Biases
- Noise
- Uncertainty

Multiple factors must be taken into account:

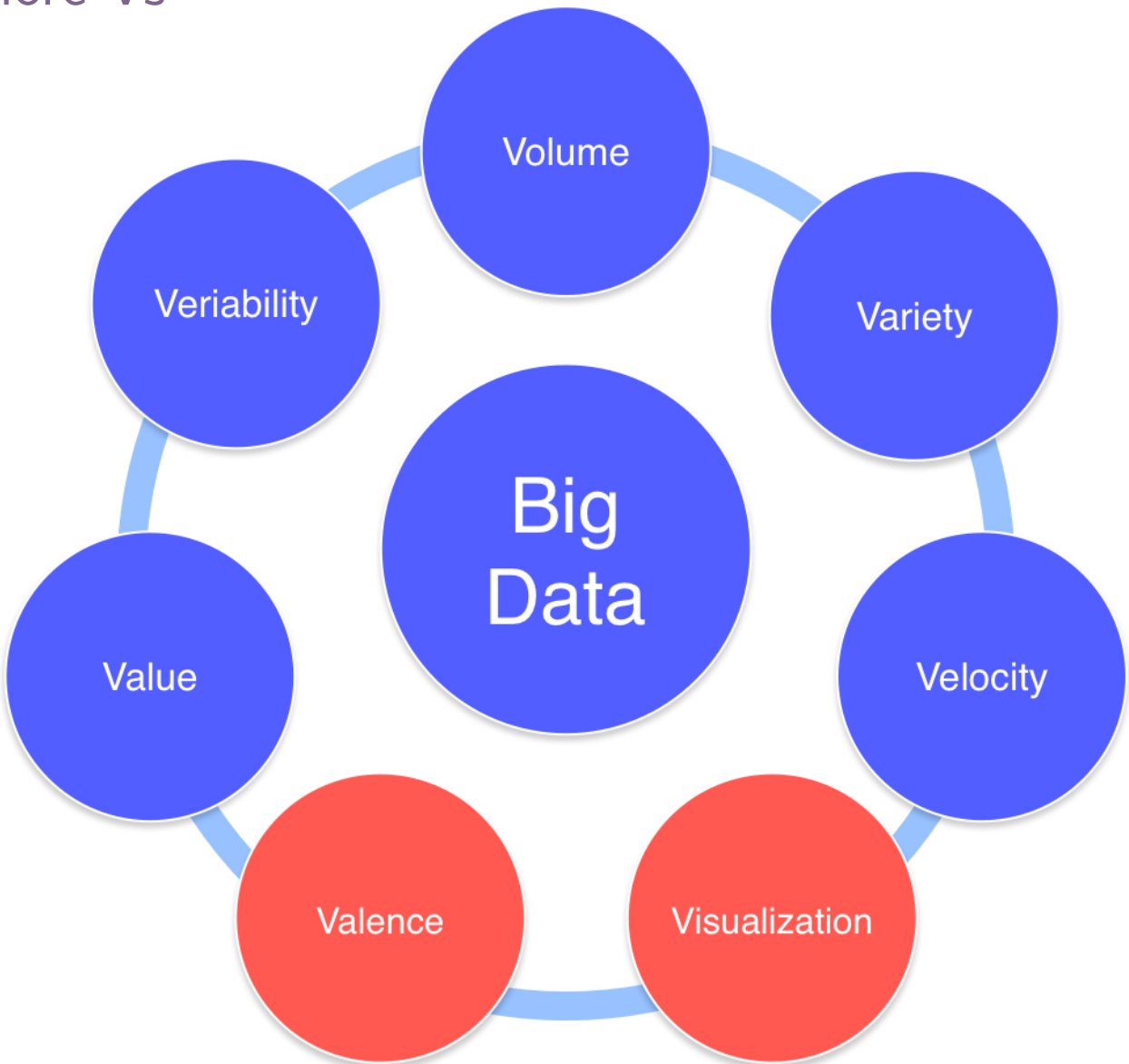
- Trustworthiness or reliability of the data source
- How meaningful the data is with respect to the analysis performed

Value

It is good having access to Big Data but it is useful only if value can be produced

How Big Data benefits your objectives

Big Data: more Vs



Valence

Valence refers to connectedness

Data connectivity increases over time

- Emergent behavior in the whole data set
- Predict how valence of a connected data set may change with time and volume
- Classic analytic critiques very inefficient

Visualization

Visualization-based data discovery methods

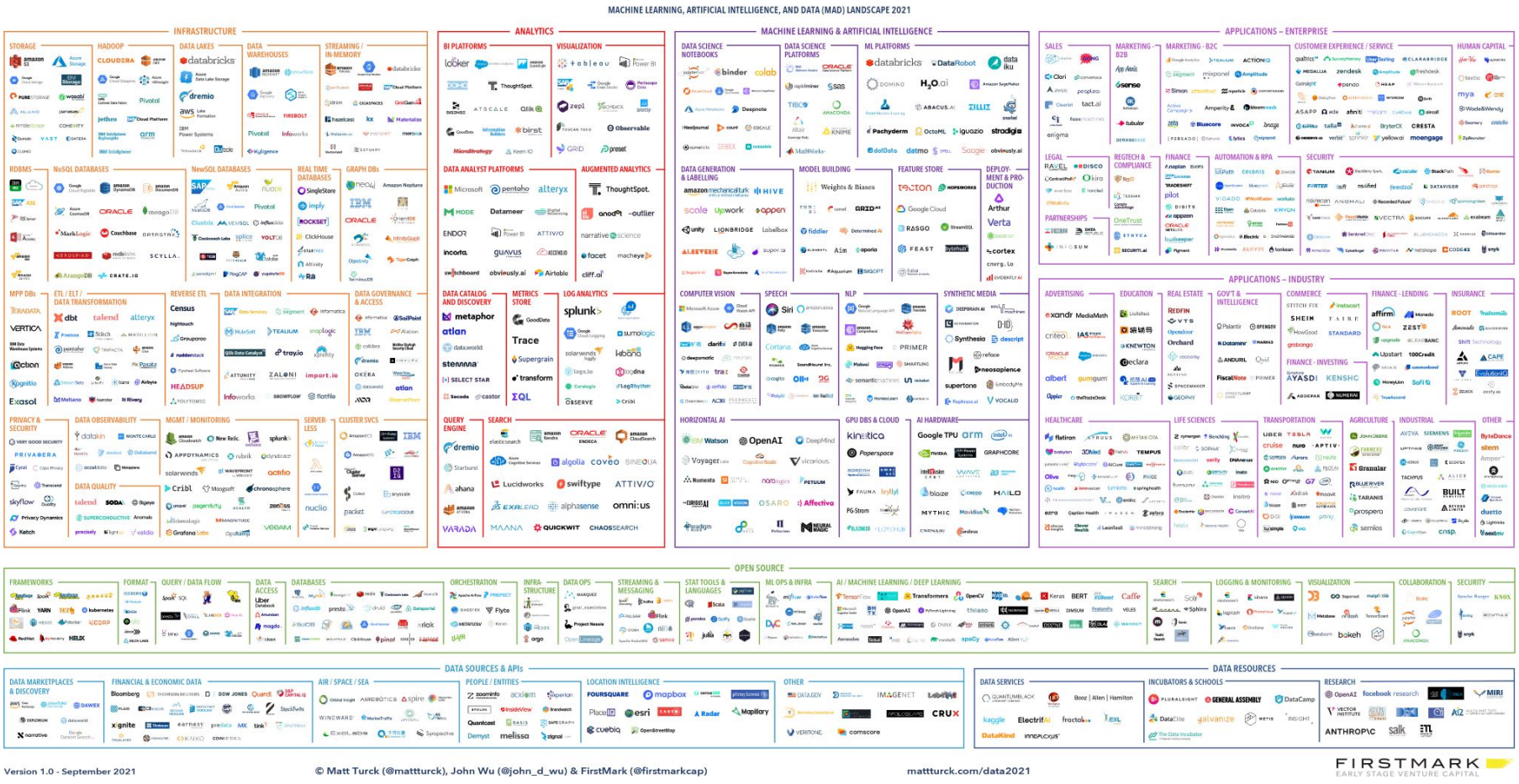
- Tables, diagrams, images, and other intuitive display
- Interactive and animated graphics
- Feature extraction and geometric modeling to greatly reduce data size

Data Science

Two ingredients:

- Data
  - Common devices and specialized devices
  - New data sources producing torrent of data
- Cloud Computing
  - Computing anywhere and anytime
  - On-demand Computing

Technologies



From Data to wisdom

Cliff Stoll

"Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom"

Data

What is Data?

is a set of values of qualitative or quantitative variables [3].

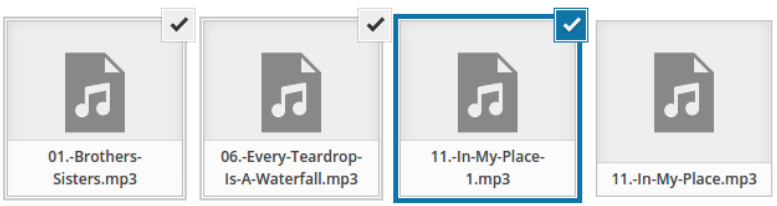
- text in documents
- images
- audio clips
- etc.

Text Data

1.00	1.00	1.00	1.00	10.00	7.00	96.00	96.00	52.00	96.00	96.00	40.00	96.00	96.00	96.00
5.00	3.00	5.00	5.00	10.00	96.00	96.00	96.00	7.00	96.00	96.00	7.00	96.00	96.00	96.00
1.00	3.00	1.00	4.00	10.00	96.00	96.00	96.00	7.00	96.00	96.00	6.00	41.00	96.00	96.00
4.00	4.00	3.00	4.00	9.00	96.00	96.00	96.00	7.00	96.00	96.00	6.00	96.00	96.00	96.00
4.00	4.00	4.00	4.00	7.00	96.00	96.00	96.00	28.00	96.00	96.00	6.00	96.00	96.00	96.00
2.00	4.00	6.00	3.00	10.00	96.00	96.00	96.00	60.00	96.00	96.00	7.00	96.00	96.00	96.00
2.00	2.00	1.00	3.00	10.00	96.00	96.00	96.00	60.00	96.00	96.00	7.00	96.00	96.00	96.00
4.00	5.00	4.00	4.00	10.00	7.00	96.00	96.00	60.00	96.00	96.00	7.00	6.00	96.00	96.00
5.00	4.00	6.00	4.00	10.00	47.00	96.00	96.00	60.00	96.00	96.00	33.00	96.00	96.00	96.00
2.00	1.00	1.00	2.00	7.00	96.00	96.00	96.00	60.00	96.00	96.00	33.00	96.00	96.00	96.00
1.00	4.00	5.00	2.00	10.00	7.00	96.00	96.00	60.00	96.00	96.00	6.00	41.00	96.00	96.00
4.00	5.00	3.00	4.00	10.00	7.00	96.00	96.00	28.00	96.00	96.00	2.00	5.00	96.00	96.00
4.00	3.00	6.00	2.00	10.00	7.00	96.00	96.00	18.00	96.00	96.00	7.00	96.00	96.00	96.00
2.00	2.00	2.00	3.00	10.00	7.00	96.00	96.00	60.00	96.00	96.00	6.00	96.00	96.00	96.00
4.00	3.00	3.00	3.00	10.00	96.00	96.00	96.00	60.00	96.00	96.00	27.00	96.00	96.00	96.00
4.00	4.00	5.00	4.00	7.00	10.00	96.00	96.00	28.00	96.00	96.00	2.00	7.00	96.00	96.00
5.00	5.00	5.00	4.00	7.00	96.00	96.00	96.00	10.00	96.00	96.00	2.00	5.00	96.00	96.00
2.00	2.00	6.00	2.00	7.00	10.00	96.00	96.00	60.00	96.00	96.00	33.00	96.00	96.00	96.00
1.00	4.00	5.00	2.00	10.00	7.00	96.00	96.00	60.00	96.00	96.00	33.00	96.00	96.00	96.00
2.00	3.00	6.00	2.00	10.00	52.00	96.00	96.00	60.00	96.00	96.00	5.00	6.00	96.00	96.00
1.00	4.00	4.00	1.00	10.00	47.00	96.00	96.00	60.00	96.00	96.00	7.00	96.00	96.00	96.00
4.00	4.00	6.00	3.00	3.00	10.00	96.00	96.00	47.00	96.00	96.00	7.00	5.00	96.00	96.00
1.00	5.00	5.00	4.00	10.00	7.00	96.00	96.00	47.00	96.00	96.00	7.00	2.00	96.00	96.00
4.00	4.00	4.00	4.00	10.00	7.00	96.00	96.00	28.00	18.00	96.00	7.00	6.00	96.00	96.00
2.00	2.00	3.00	2.00	10.00	3.00	7.00	96.00	60.00	96.00	6.00	96.00	96.00	96.00	96.00
3.00	6.00	3.00	3.00	10.00	7.00	96.00	96.00	28.00	18.00	96.00	6.00	42.00	96.00	96.00
4.00	5.00	5.00	4.00	10.00	7.00	96.00	96.00	60.00	96.00	96.00	9.00	6.00	96.00	96.00
3.00	3.00	4.00	2.00	7.00	96.00	96.00	96.00	60.00	96.00	96.00	33.00	96.00	96.00	96.00
5.00	4.00	5.00	5.00	7.00	96.00	96.00	96.00	60.00	96.00	96.00	35.00	7.00	96.00	96.00
2.00	1.00	6.00	2.00	10.00	96.00	96.00	96.00	60.00	96.00	96.00	6.00	96.00	96.00	96.00
1.00	1.00	1.00	1.00	10.00	7.00	50.00	96.00	3.00	96.00	96.00	6.00	41.00	96.00	96.00
2.00	2.00	1.00	2.00	10.00	52.00	96.00	96.00	60.00	96.00	96.00	6.00	10.00	96.00	96.00
4.00	5.00	1.00	4.00	10.00	7.00	96.00	96.00	60.00	96.00	96.00	27.00	96.00	96.00	96.00
2.00	4.00	5.00	2.00	10.00	47.00	96.00	96.00	60.00	96.00	96.00	27.00	96.00	96.00	96.00
5.00	5.00	5.00	4.00	10.00	96.00	96.00	96.00	60.00	96.00	96.00	27.00	96.00	96.00	96.00
2.00	3.00	2.00	3.00	10.00	7.00	96.00	96.00	60.00	96.00	96.00	6.00	96.00	96.00	96.00
3.00	4.00	4.00	2.00	32.00	47.00	96.00	96.00	60.00	96.00	96.00	2.00	96.00	96.00	96.00
4.00	3.00	6.00	4.00	10.00	96.00	96.00	96.00	60.00	96.00	96.00	4.00	96.00	96.00	96.00
4.00	2.00	6.00	3.00	1.00	47.00	96.00	96.00	60.00	96.00	96.00	7.00	96.00	96.00	96.00
1.00	3.00	1.00	3.00	1.00	47.00	96.00	96.00	60.00	96.00	96.00	6.00	96.00	96.00	96.00
5.00	3.00	3.00	4.00	10.00	50.00	96.00	96.00	7.00	96.00	96.00	42.00	96.00	96.00	96.00

Image Data

Audio Data


01.-Brothers-Sisters.mp3
06.-Every-Teardrop-Is-A-Waterfall.mp3
11.-In-My-Place-1.mp3
11.-In-My-Place.mp3

Data

Data can be created, collected, processed, analyzed, saved, and stored.

Data becomes information by interpretation

Data

Data can be created, collected, processed, analyzed, saved, and stored.

Data becomes information by interpretation

8.848,56
5.895
4.810

Data

Data can be created, collected, processed, analyzed, saved, and stored.

Data becomes information by interpretation

8.848,56m
5.895m
4.810m

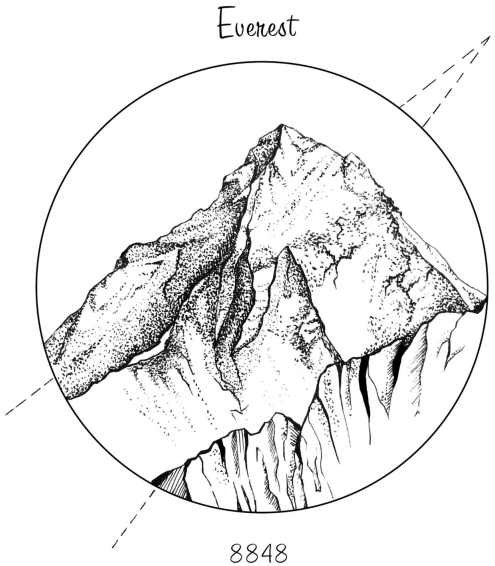
Data

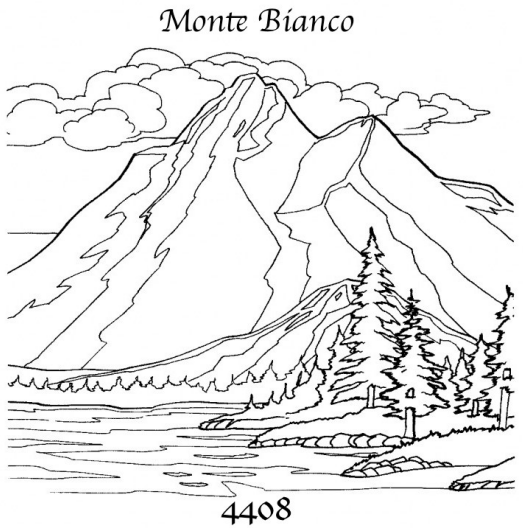
Data can be created, collected, processed, analyzed, saved, and stored.

Data becomes information by interpretation

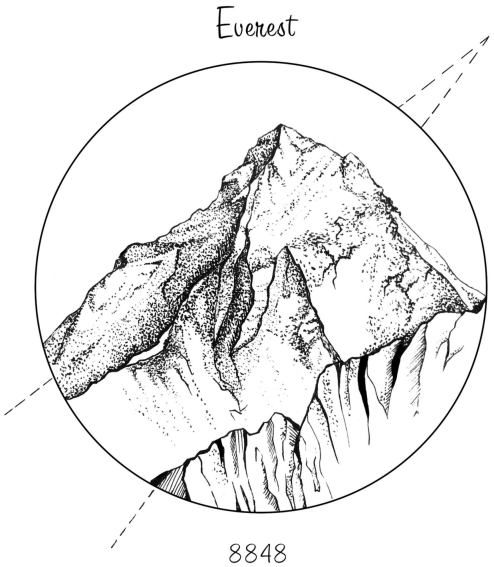
8.848,56m Everest
5.895m Kilimanjaro
4.810m Monte Bianco

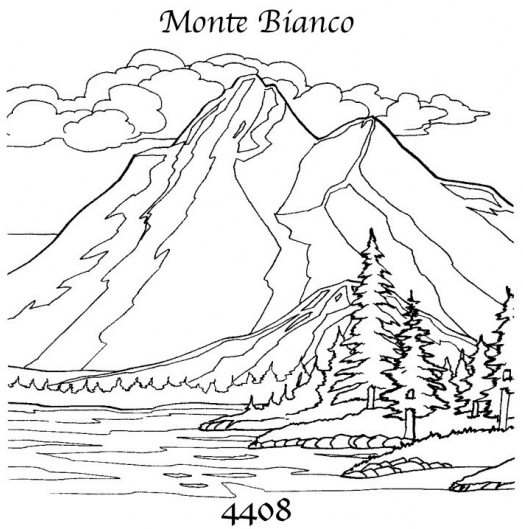
Data becomes information by interpretation



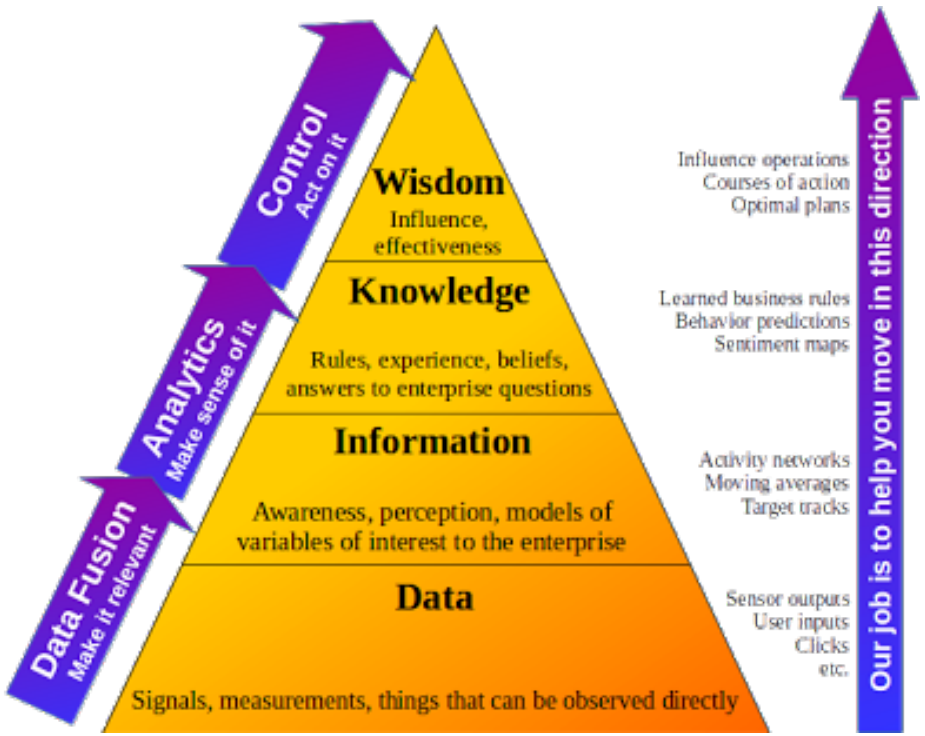


Data becomes information by interpretation





From Data to wisdom



Many tasks to get to wisdom

- Define the objective:
  - What do I want to discover from those data
- Prepare a model
  - Put all the ingredients
- Run the model
  - Technologies
- Communicate the results
  - Insight
  - Action

Plan

Data Science

An introduction to Data Science, Jeffrey Stanton

"Data Science refers to an emerging area of work concerned with the collection, preparation, analysis, visualization, management, and preservation of large collections of information."

Data Science

An introduction to Data Science, Jeffrey Stanton

"Data Science refers to an emerging area of work concerned with the collection, preparation, analysis, visualization, management, and preservation of large collections of information."

Hal Varian, Google Chief Economist NYT, 2009

"Data science is the next sexy job"
"The ability to take data - to be able to understand it, to process it, to extract **value** from it, to visualize it, to communicate it - that's going to be a highly important skill."

Data Science

Mike Driscoll, CEO of meta-markets

"Data science is the civil engineering of data. It is acolytes that possess a practical knowledge of tools & materials, coupled with a theoretical understanding of what's possible."

Plan

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

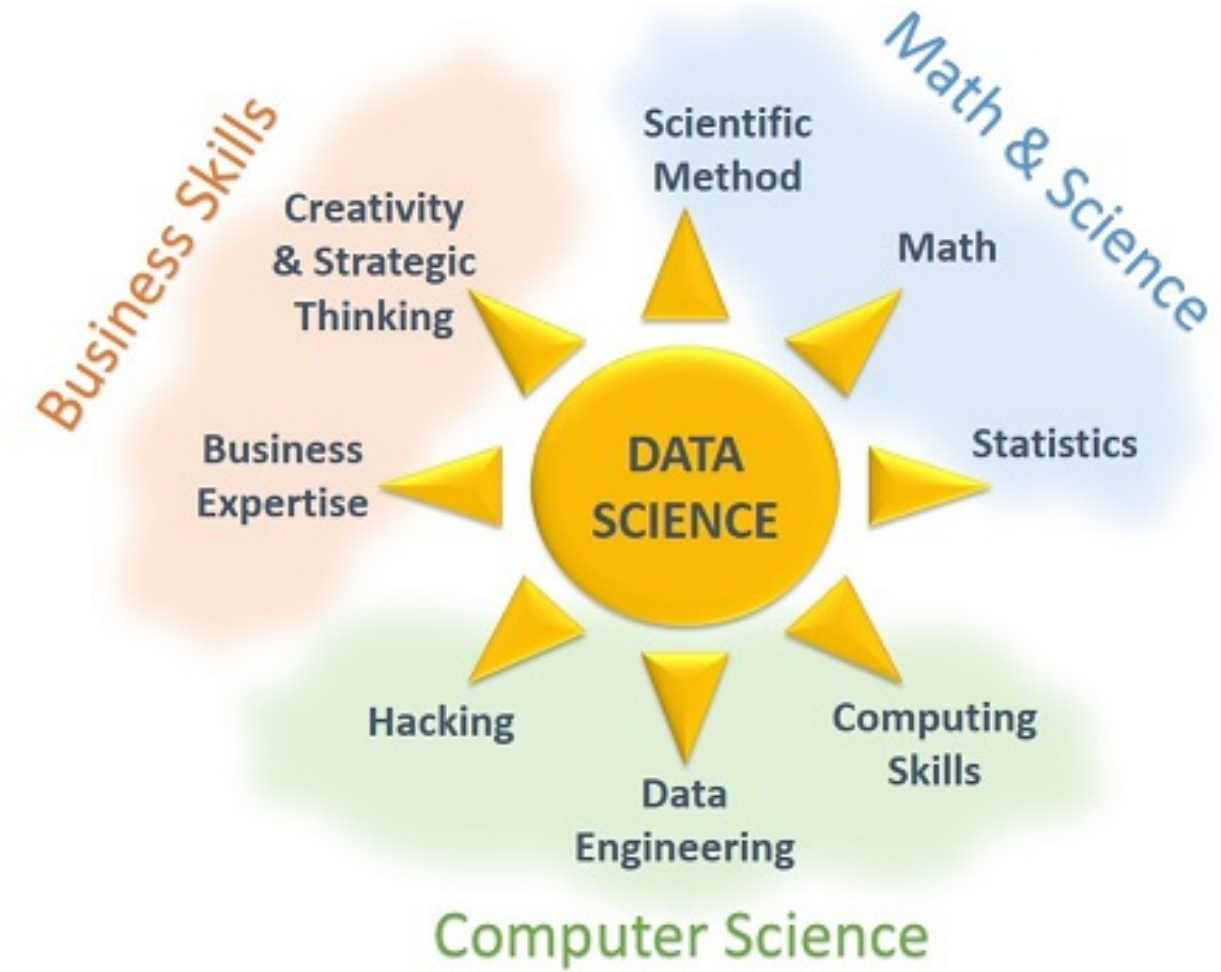
- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Data scientists

Hilary Mason, chief scientist at `bit.ly`

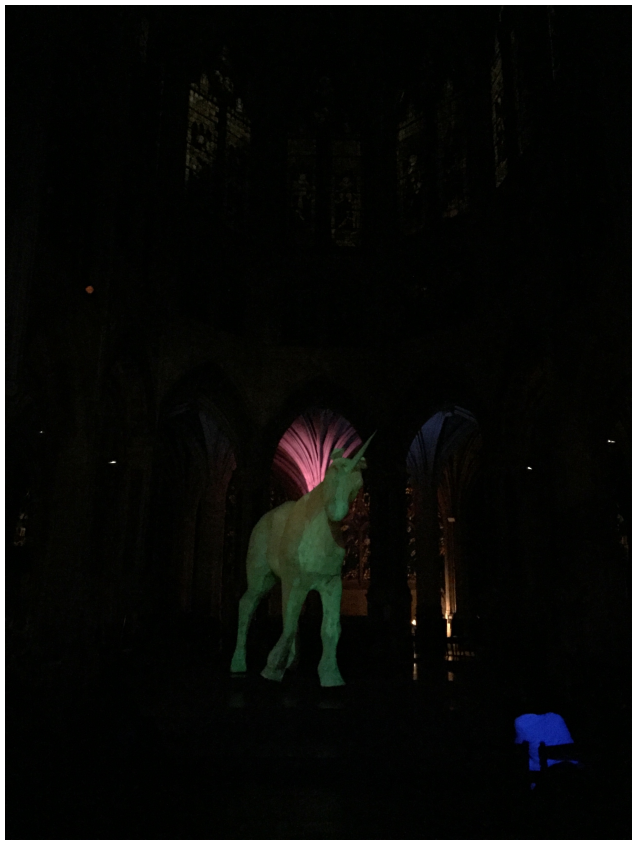
"A data scientist is someone who can obtain, scrub, explore, model, and interpret data, blending hacking, statistics and machine learning. Data scientists not only are adept at working with data, but appreciate data itself as a **first-class product**."

Data scientists



Data scientists

Example




Une apparition
La licorne phosphorescente
ou l'évanescence d'un rêve
Marguerite Lantz
Nuit blanche
Samedi 1 oct 2016
Eglise Saint Séverin 8h-2h
Eclairages : Benjamin Terzi & Paulo Gerard
Réalisation sonore : Frédéric Blondy

Data scientists

Data scientist

Team of people that act like one

A Data science team often comes together to analyze situations, business or scientific cases, which none of the individuals can solve on their own

- Passionate about the story and the meaning behind the data
- They understand the problem and try to solve it
- They look for the right analytical methods to apply
- They have communication skills
- They put together their knowledge

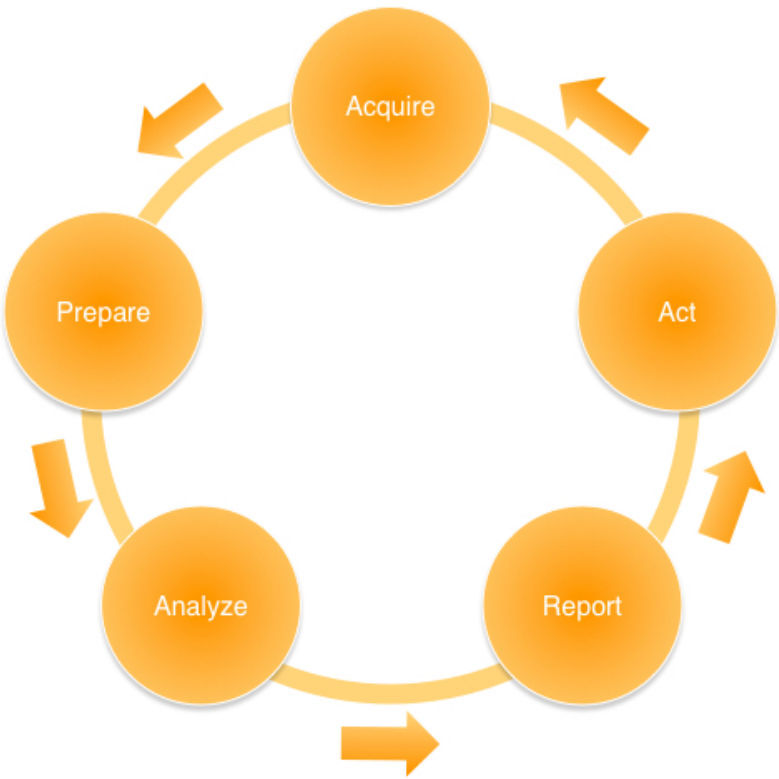
Plan

Data science process

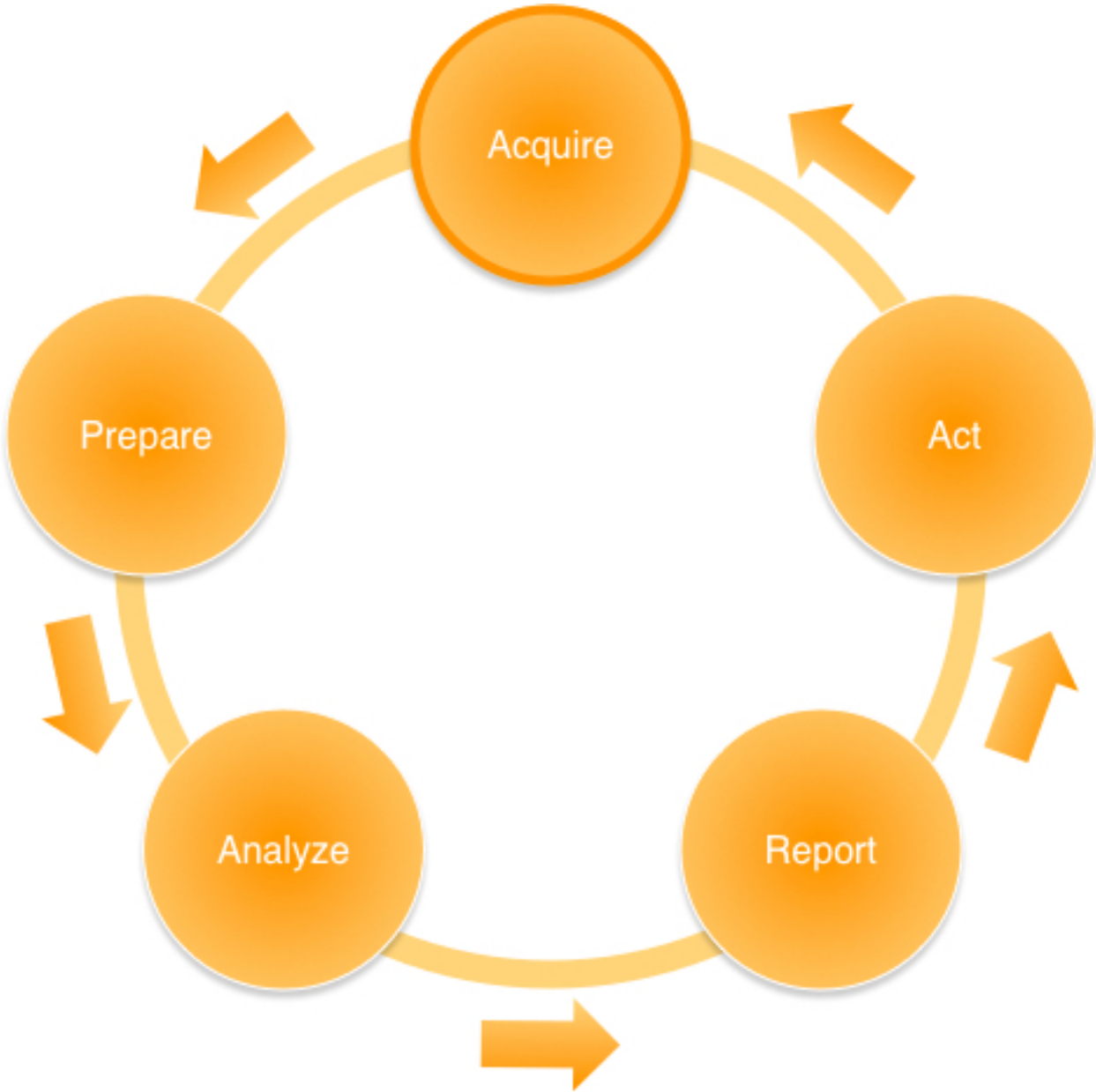
And what is our team going to do?

- follow a process to collaborate and communicate around
- discuss costs, timeline, planning, expectation, purposes, etc.

Five distinct steps



Acquire



Acquire

Obtain the source material before analyzing or acting on it

Identify suitable data related to the problem:

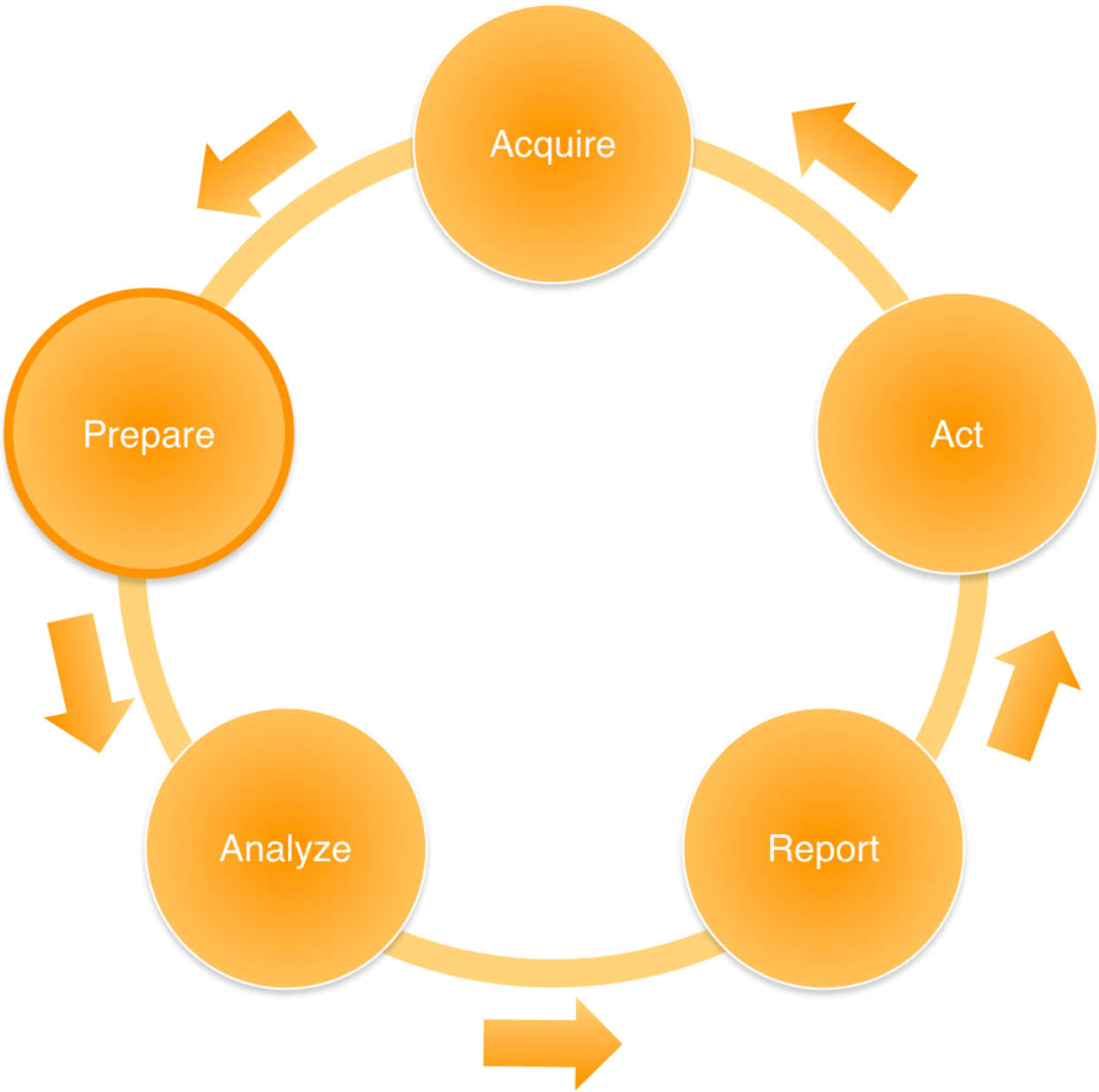
- Data comes from many places
  - Local
  - Remote
- Many varieties
  - Structured
  - Unstructured

Acquire

Data Sources

- Standard Relational data sources
- Text files
- NoSQL databases
- Web
- Remote data
- Various sources

Prepare



Prepare

Explore

Preliminary investigation in order to understand the specific characteristics of the data
Usage of graphs, visualization tools, statistical analysis

- Mean
- Median
- Mode
- Range

It is necessary to understand the nature and the complexity of the dataset

Prepare

Pre-process

The raw data is never in the format that you need to perform analysis on:

- Data cleaning
- Data transformation

Data cleaning

- Inconsistencies
- Duplication
- Missing values
- Invalid data
- Outliers

Prepare

Pre-process

The raw data is never in the format that you need to perform analysis on:

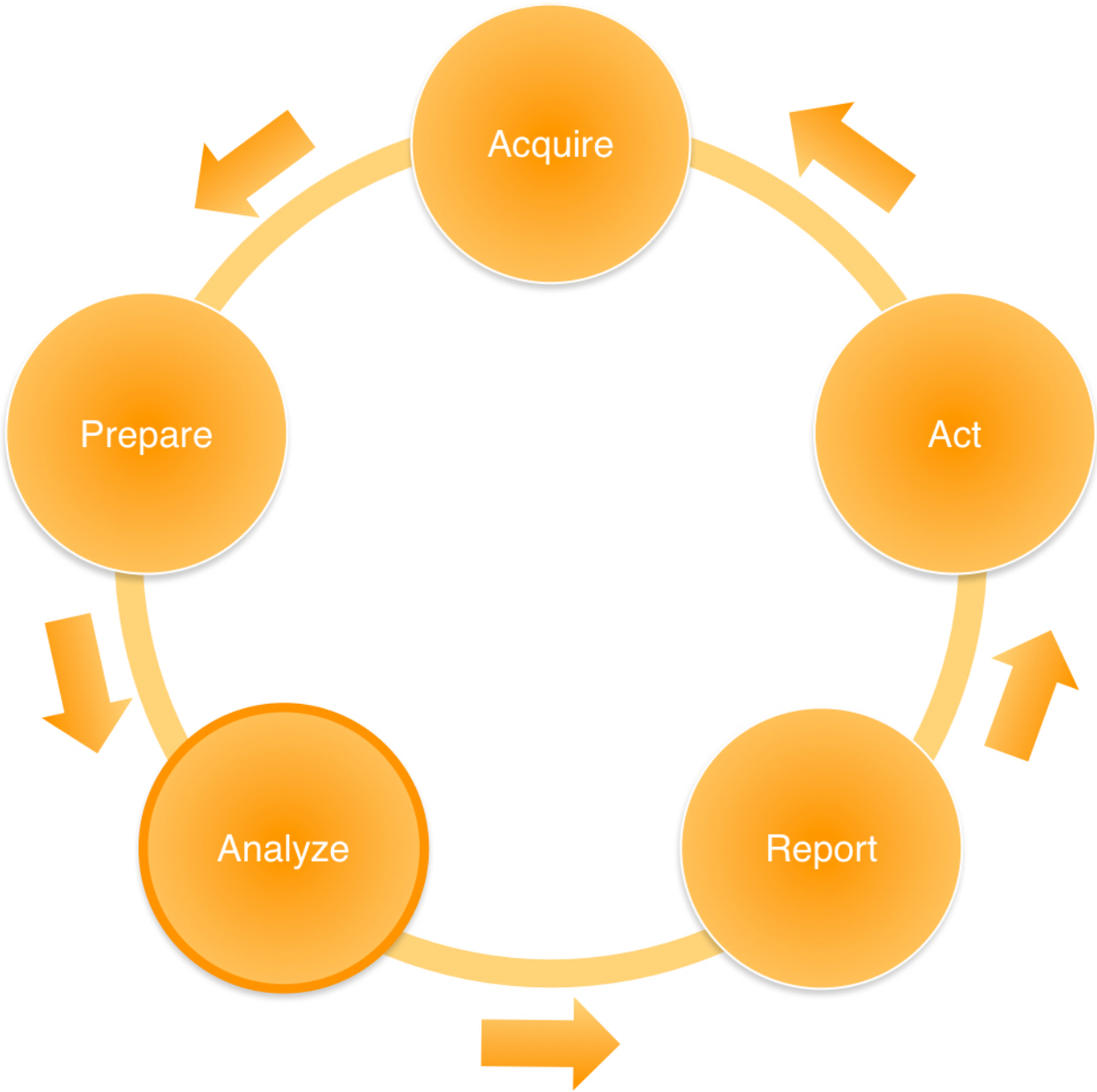
- Data cleaning
- Data transformation

Data transformation

transform the raw data to make it suitable for analysis

- Scale
- Transform
- Feature selection
- Dimensionality reduction
- Data manipulation

Analyze

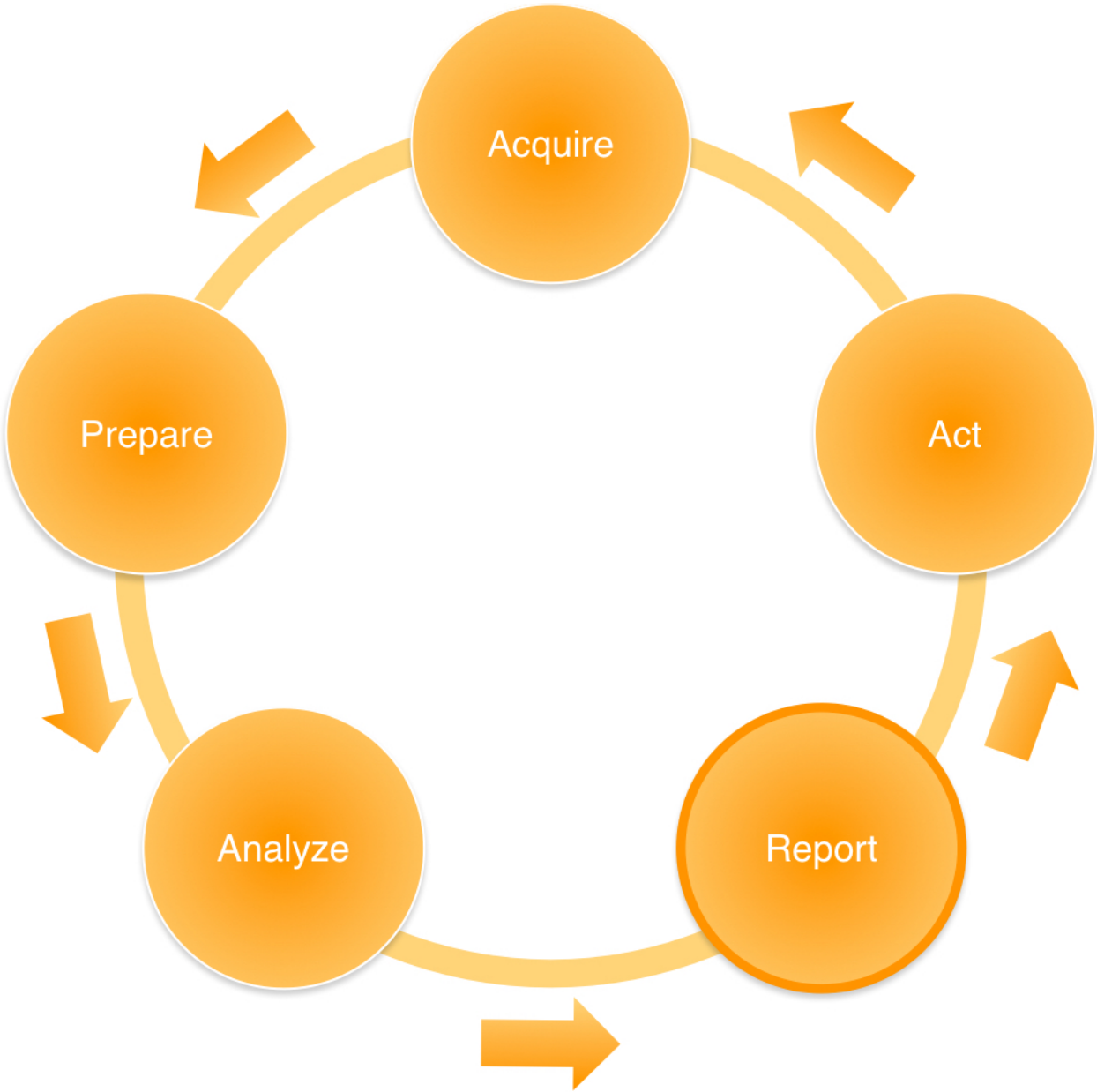


Analyze

The input data is used by the analysis technique to build a model

- Classification
  - Predict the category of the input data
    - Weather as being sunny, rainy, windy, or cloudy
- Regression
  - Predict a numeric value
    - Predict the price of a stock
- Clustering
  - Organize similar items into groups
    - Identifying areas of similar topography
- Association Analysis
  - Capture associations within items or events: when items or events occur together
    - In supermarkets discover connections between two seemingly unrelated products

Report



Report

Reporting the insights gained from the analysis

- All findings must be presented so that informed decisions can be made
- Visualization is an important tool in presenting the results

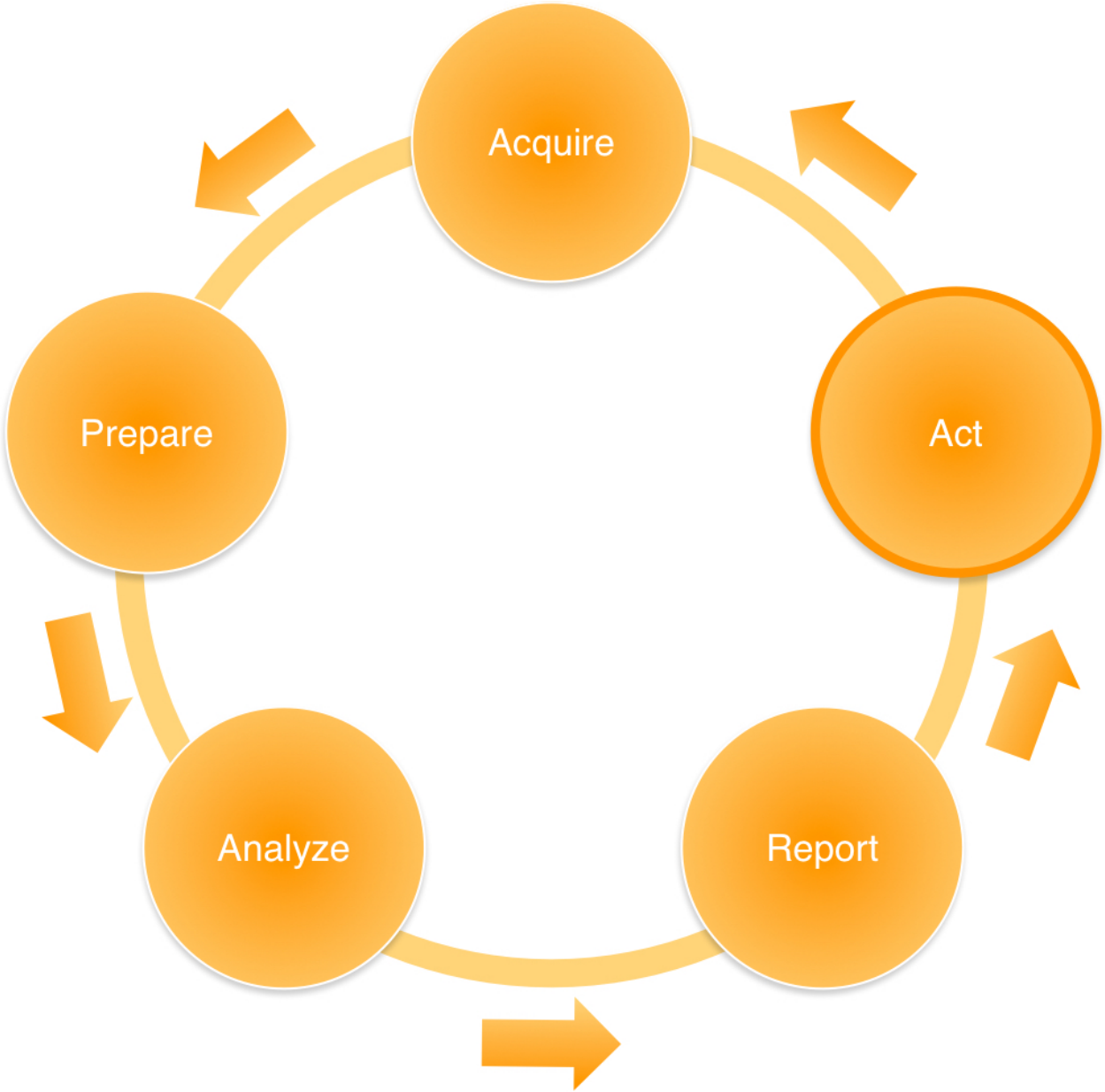
Visualization

Plotting data

Visualization tools

- R: a software package for general data analysis
- Python packages to support data analysis and graphics
- D3: a JavaScript library for producing interactive web based visualizations
- Leaflet: a lightweight JavaScript library to create interactive maps
- etc.

Act



Act

Action or actions should be taken, based on the insights gained

Possible actions

- something to be changed
- something to be added
- something to be implemented
- change something in the input and restart the process

Big Data

"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

References I


The Exponential Growth of Data.
https://insidebigdata.com/2017/02/16/the-exponential-growth-of-data.
Accessed 2017.


Game changers: Five opportunities for US growth and renewal.

http://www.mckinsey.com/global-themes/americas/us-game-changers.
Accessed 2016.


Wikipedia.
http://wikipedia.com.
Accessed 2017.