
Department of
Computer Science

# Introduction to Data Science

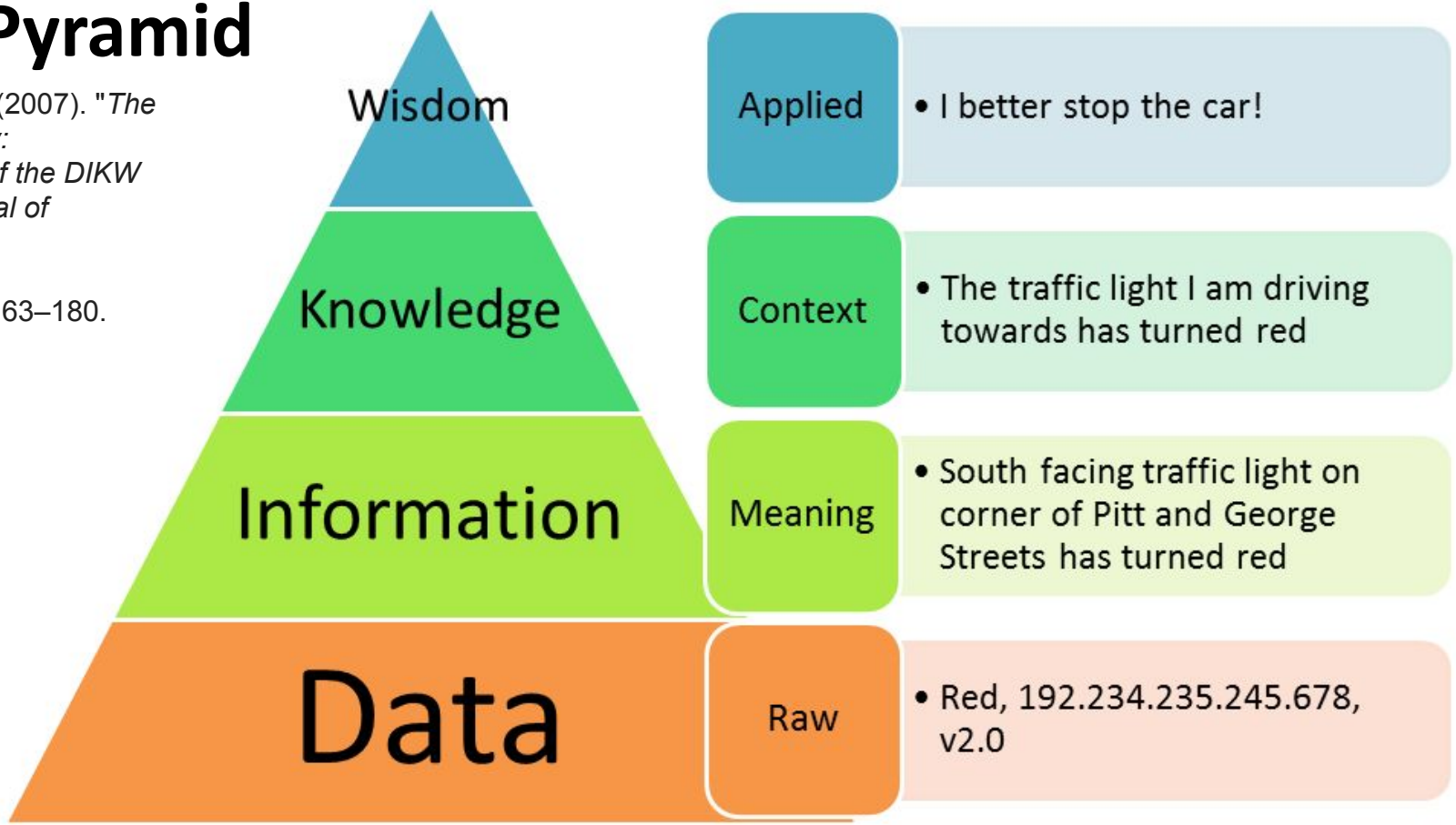
Erick Gomez Nieto, PhD
*emgomez@ucsp.edu.pe*

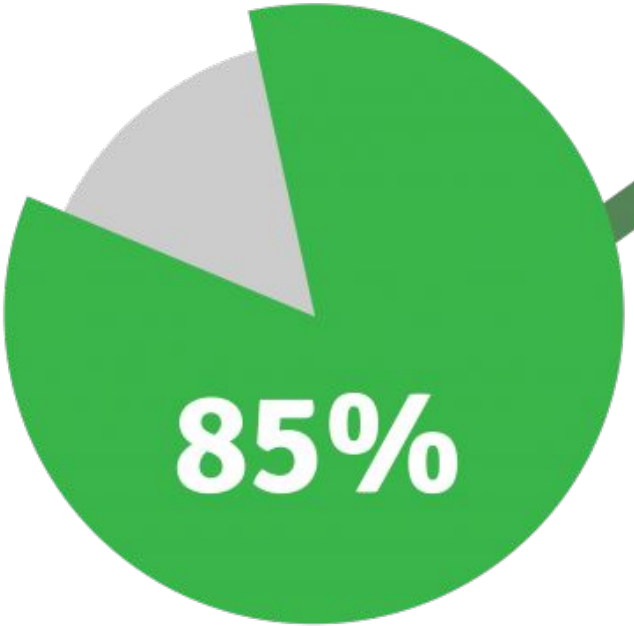
ACM/IEEE Computing Curricula 2020

- Computer Engineering
- ***Computer Science***
- Cybersecurity
- Information Technology
- Software Engineering
- Information Systems
- ***Data Science*** (*under development*)

# DIKW Pyramid

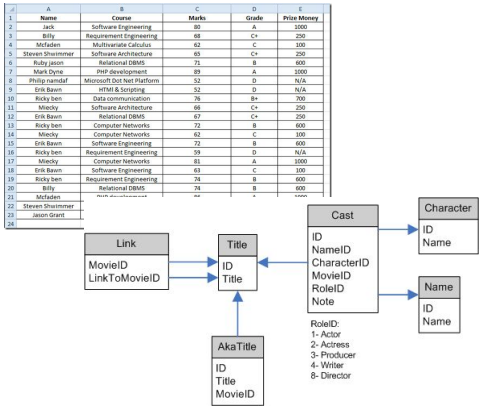
Rowley, Jennifer (2007). "*The wisdom hierarchy: representations of the DIKW hierarchy*". *Journal of Information and Communication Science*. **33** (2): 163–180.






of BUSINESS DATA is
UNSTRUCTURED

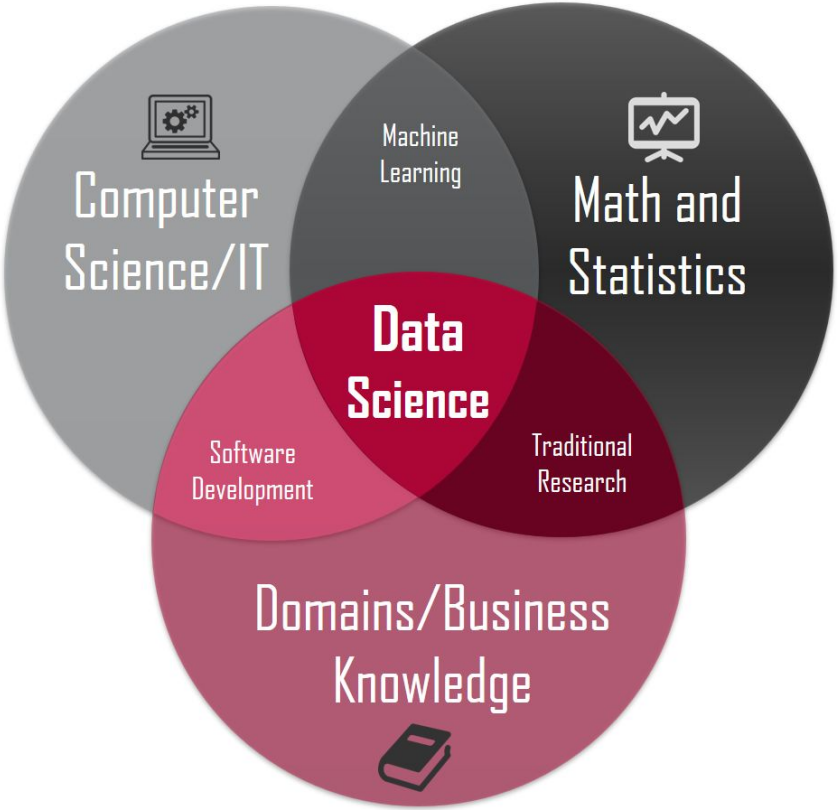
Fuente: IBM (Digital Reasoning website)




***Estructurados***


***No Estructurados***

# Overview



# MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 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 package 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

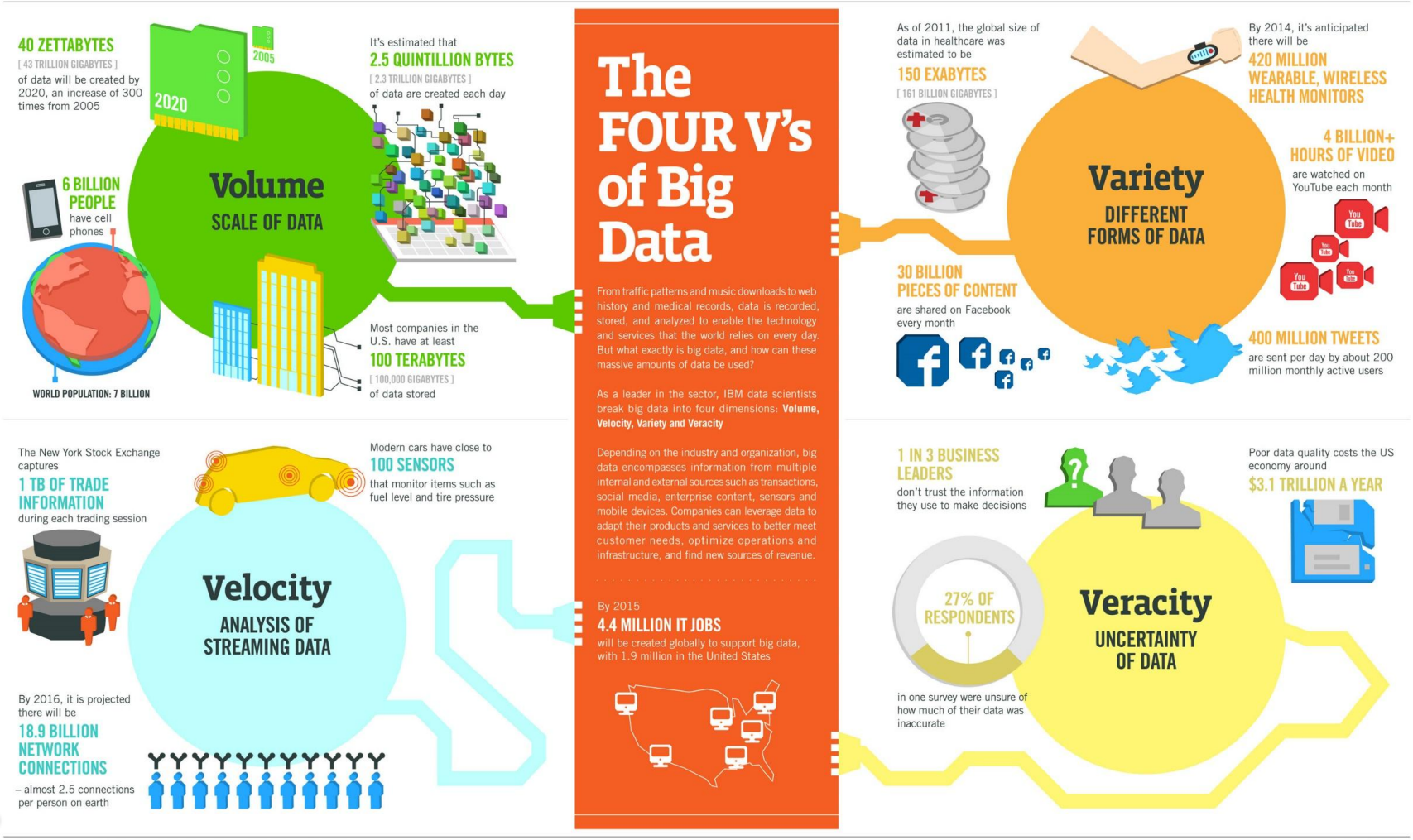
## COMMUNICATION & VISUALIZATION

- 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

# DATA

# BIG DATA

***Big Data*** is a phrase used to mean a massive volume of both **structured** and unstructured **data** that is so large it is difficult to process using traditional **database** and **software** techniques.




2001

# The fifth "V"?

**Big data = the ability to achieve greater Value through insights from superior analytics**

**Case study:** A US-based aircraft engine manufacturer now uses analytics to predict engine events that lead to costly airline disruptions, with 97% accuracy. If this prediction capability had been available in the previous year, it would have saved $63 million.



**2012**

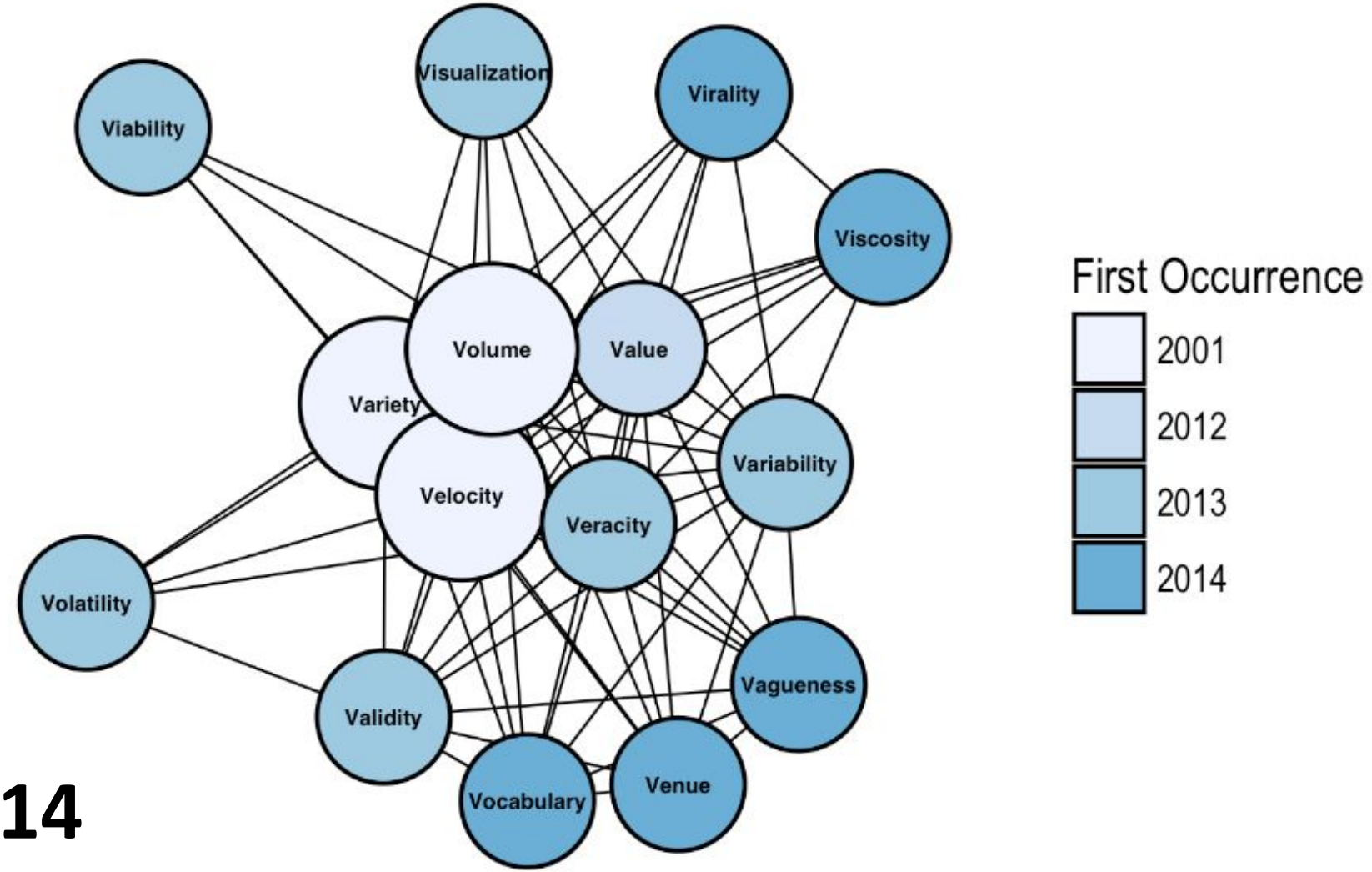
**Volume-based value**: The more comprehensive your 360-degree view of customers and the more historical data you have on them, the more insight you can extract from it all and, all things considered, the better decisions you can make in the process of acquiring, retaining, growing and managing those customer relationships.

**Velocity-based value**: The more customer data you can ingest rapidly into your big-data platform and the more questions that a user can pose more rapidly against that data (via queries, reports, dashboards, etc.) within a given time period prior, the more likely you are to make the right decision at the right time to achieve your customer relationship management objectives.

**Variety-based value**: The more varied customer data you have – from the CRM system, social media, call-center logs, etc. – the more nuanced portrait you have on customer profiles, desires and so on, hence the better-informed decisions you can make in engaging with them.

**Veracity-based value**: The more consolidated, conformed, cleansed, consistent current the data you have on customers, the more likely you are to make the right decisions based on the most accurate data.

From: *"Measuring the Business Value of Big Data"* - James Kobielus

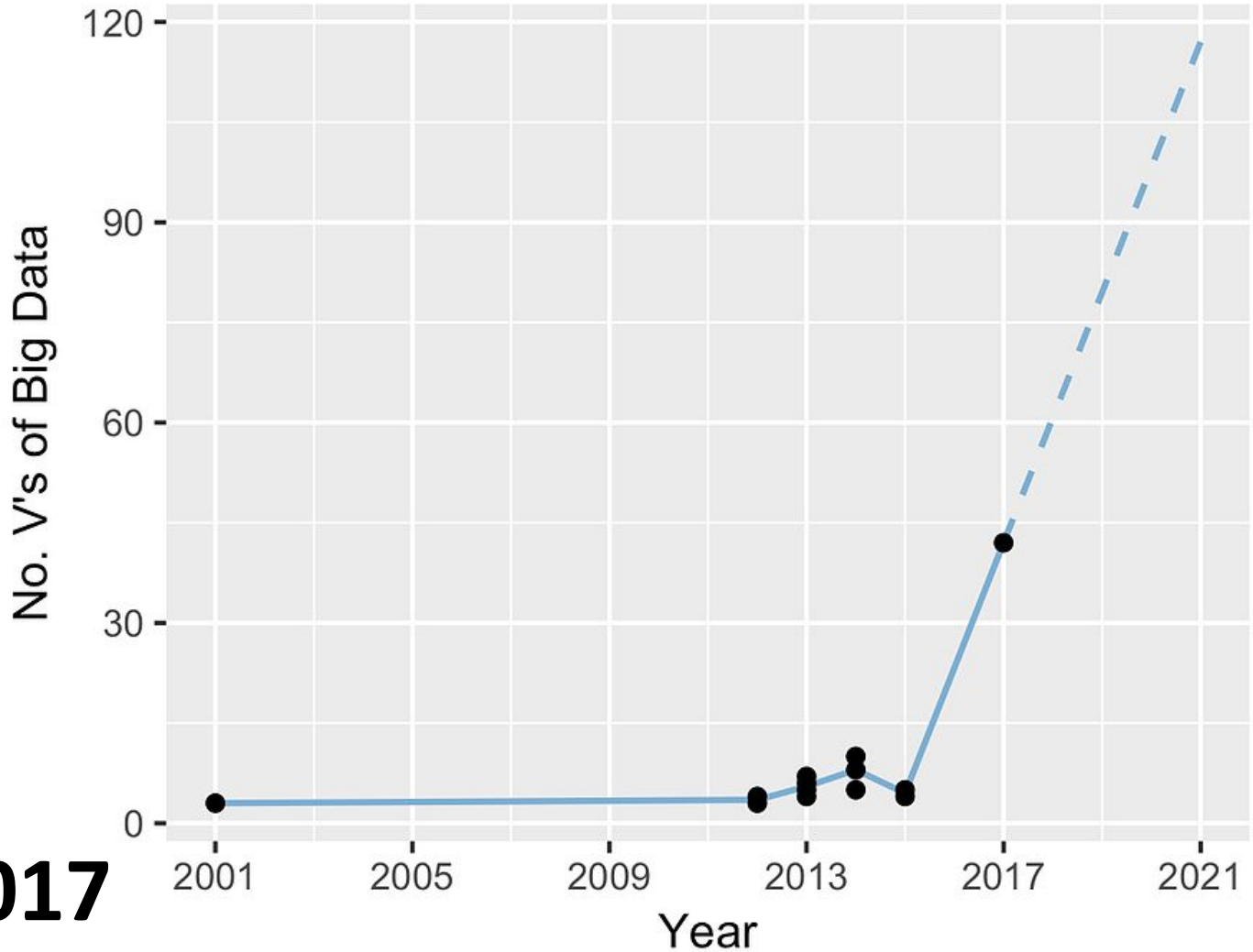



**2014**

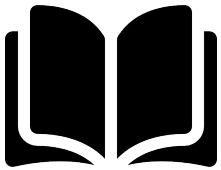
**Vulnerability**
Big data brings new security concerns. After all, a data breach with big data is a big breach. Does anyone remember the infamous AshleyMadison hack in 2015?

**Visualization**
Another characteristic of big data is how challenging it is to visualize.

**Volatility**
How old does your data need to be before it is considered irrelevant, historic, or not useful any longer? How long does data need to be kept for?




**2017**



# The 42 V's of Big Data and Data Science

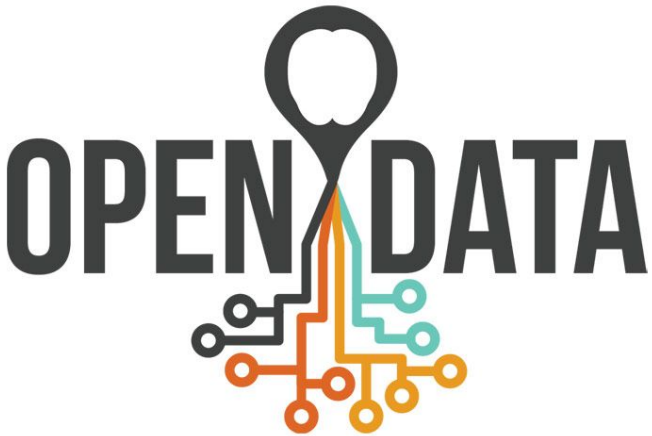
https://www.elderresearch.com/blog/42-v-of-big-data

# Open Data

Open data is **data** that **anyone** can **access**, **use** and **share**. Governments, businesses and individuals can use open data to bring about social, economic and environmental benefits.

Open data becomes usable when made available in a **common, machine-readable format**.

Open data must be licensed. Its licence must permit people to use the data in any way they want, including transforming, combining and sharing it with others, **even commercially**.



# Open Repositories

kaggle


UCI Machine Learning Repository

Open Data REPOSITORY

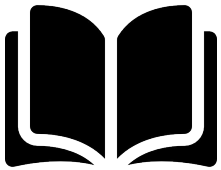
EUROPEAN DATA PORTAL

NYC OPEN DATA

OPEN DATA

opendata .swiss

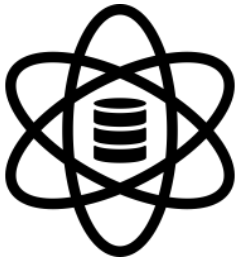
DATA HUB



# OPEN DATA HANDBOOK

https://opendatahandbook.org/

# DATA

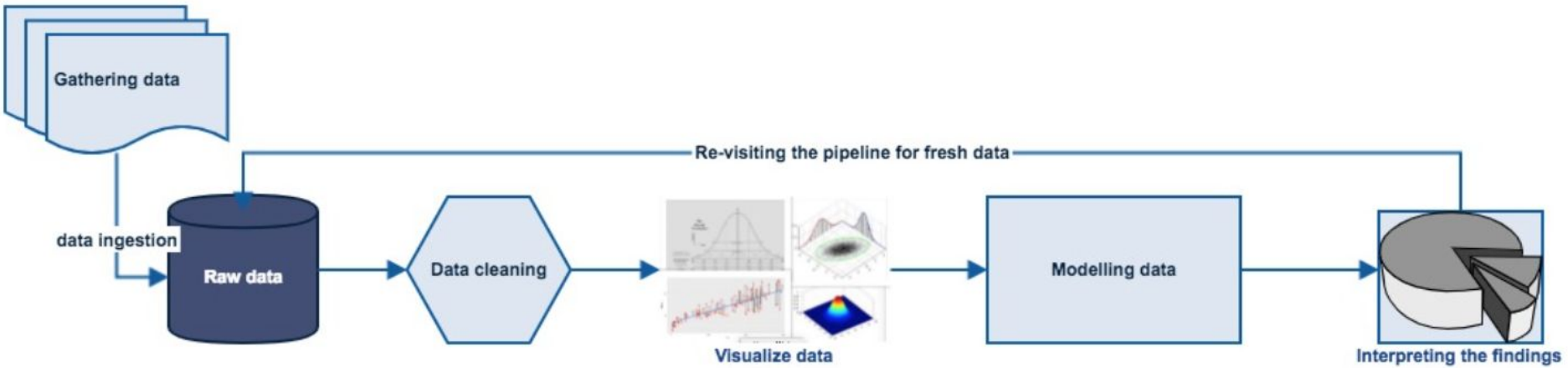


# DATA SCIENCE

**Data science** is a multidisciplinary approach to extracting actionable insights from the large and ever-increasing volumes of data collected and created by today's organizations. **Data science** encompasses preparing data for analysis and processing, performing advanced data analysis, and presenting the results to reveal patterns and enable stakeholders to draw informed conclusions.

*Extracted from **IBM**.com*
(https://www.ibm.com/cloud/learn/data-science-introduction)

# How it works ?




A high-level pipeline to address any data science problem

1. Getting your data.

2. Preparing/cleaning your data.

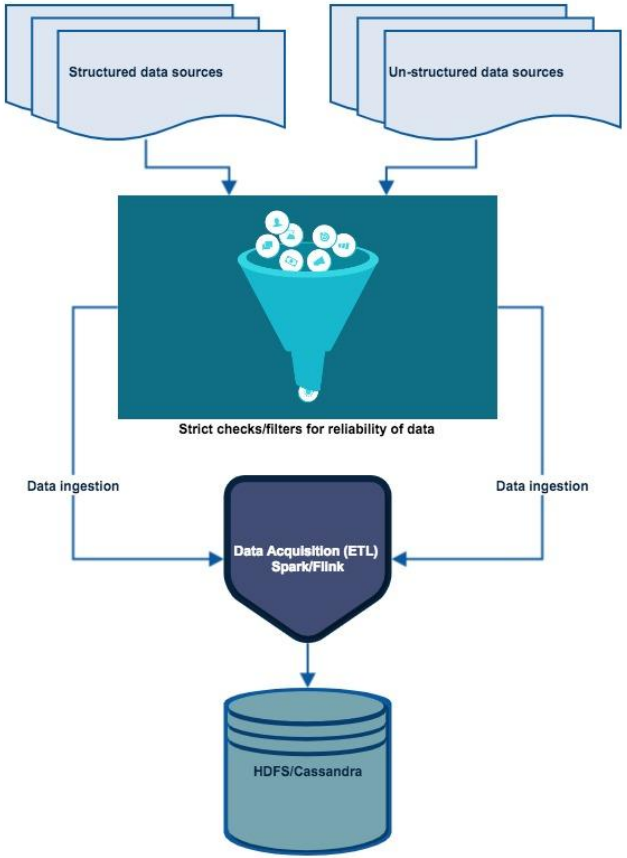
3. Exploration/visualization of data which allows you to find patterns in the numbers.

4. Modeling the data.

5. Interpreting the findings.

6. Re-visiting/updating your model.

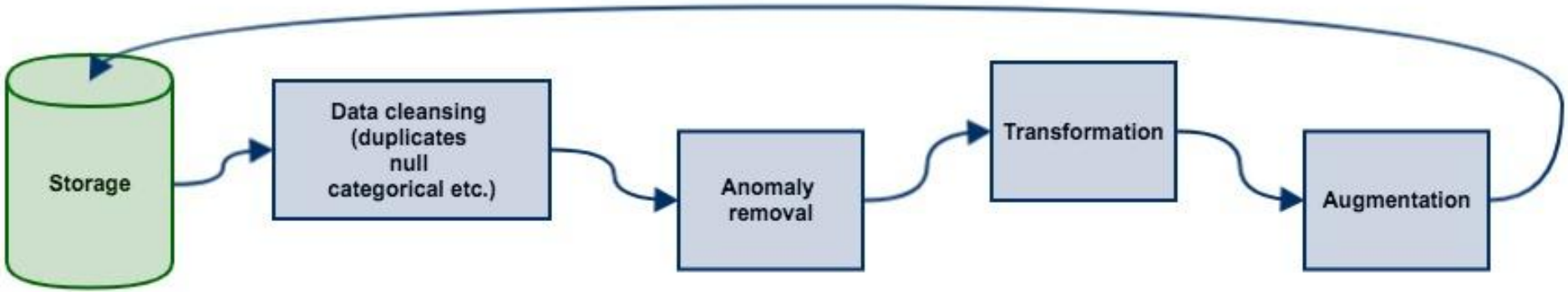
# 1. Getting your data




**Skills Required:**

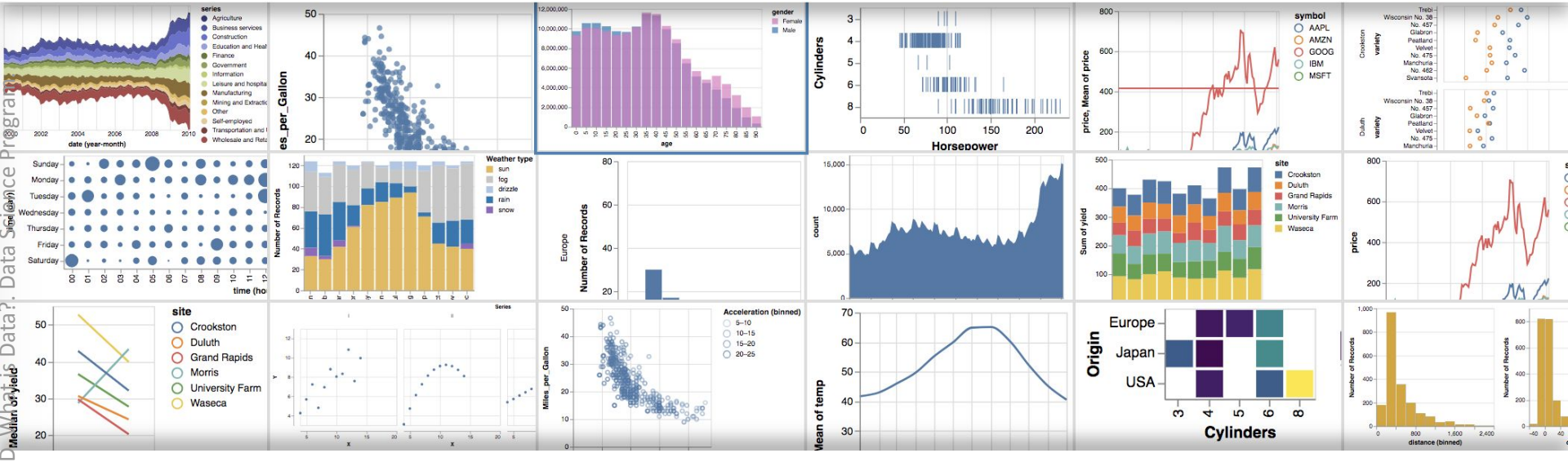
- Distributed Storage: Hadoops, Apache Spark/Flink.
- Database Management: MySQL, PostgresSQL, MongoDB, SQLite.
- Querying Relational Databases.
- Retrieving Unstructured Data: text, videos, audio files, documents.

# 2. Preparing/cleaning your data.




**Skills Required:**
Coding language: Python, R.
Data Modifying Tools: Python libs, Numpy, Pandas, R.
Distributed Processing: Hadoop, Map Reduce/Spark.

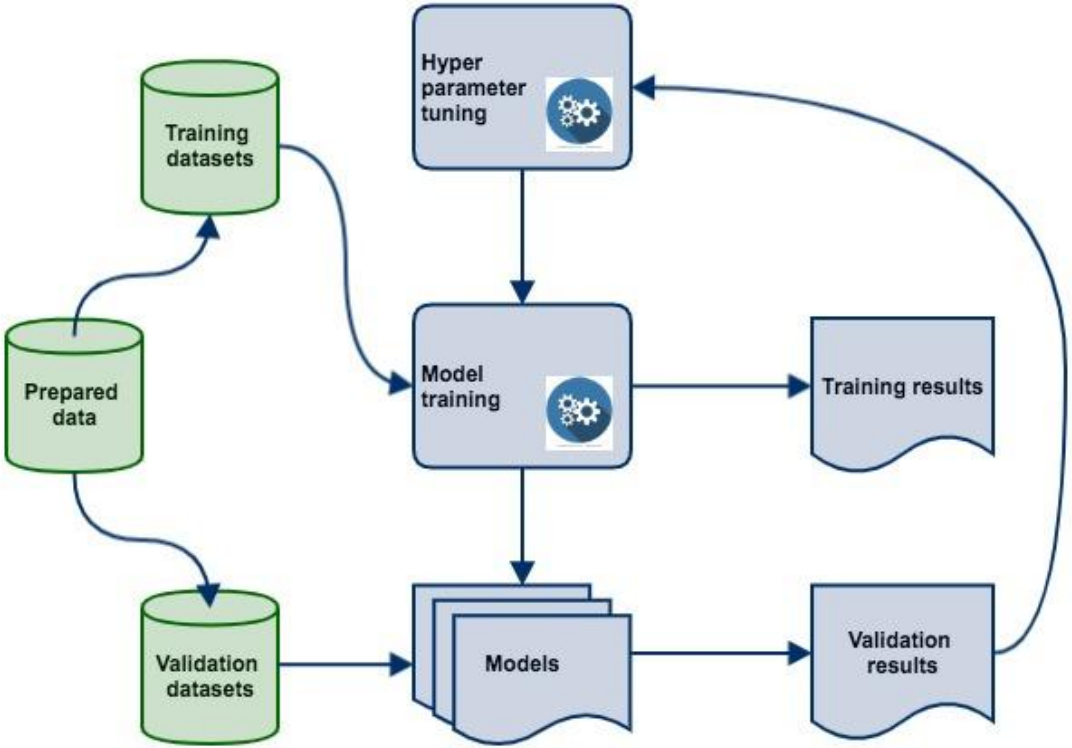
# 3. Exploration/Visualization of Data




**Skills Required:**

- Python: NumPy, Matplotlib, Pandas, SciPy.
- R: GGplot2, Dplyr.
- Statistics: Random sampling, Inferential.
- Data Visualization: Tableau.

# 4. Modeling the Data (Machine Learning)




**Skills Required:**

- Machine Learning: Supervised/Unsupervised algorithms.
- Evaluation methods.
- Machine Learning Libraries: Python (Sci-kit Learn, NumPy).
- Linear algebra and Multivariate Calculus.

# 5. Interpreting the Data

Interpreting the data is more like communicating your findings to the interested parties. If you can't explain your findings to someone believe me, whatever you have done is of no use. Hence, this step becomes very crucial.

The objective of this step is to first identify the business insight and then correlate it to your data findings. You might need to involve domain experts in correlating the findings with business problems. Domain experts can help you in visualizing your findings according to the business dimensions which will also aid in communicating facts to a non-technical audience.

**Skills required:**

- Business domain knowledge.
- Data visualization tools: Tableau, D3.js, Matplotlib, ggplot2, Seaborn.
- Communication: Presenting/speaking and reporting/writing.

# Data Types

# Data/set types+semantics
# Tasks

- Data abstraction
  - Data types
    - categorical, ordinal, quantitative
  - Dataset types
    - Tables
    - Networks/graph (trees)
    - Text / logs
    - Fields
    - Static file vs. dynamic stream
  - Attribute + dataset semantics
    - Spatial vs. non-spatial
    - Temporal vs. non-temporal
    - Keys vs. values
    - Continuous vs. discrete
    - Topology vs. geometry
  - Derived attributes / spaces
- Task abstraction

- Numeric, symbolic (or mix)
- Scalar, vector, or complex structure
- Various units
- Discrete or continuous
- Spatial, quantity, category, temporal, relational, structural
- Dense or sparse
- Ordered or non-ordered
- Disjoint or overlapping
- Binary, enumerated, multilevel
- Independent or dependent
- Multidimensional
- ...

# Introduction - Types of data

## Numeric data

- Discrete (integers)
- Continuous (real)

## Categorical data

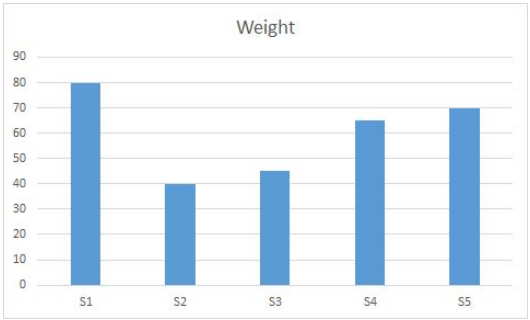
- **Ranked:** Low, medium, high
- **Unordered**: Grass, leaves, paths, urban, waste, woods

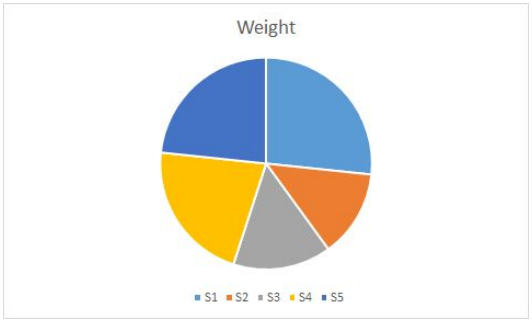
# Multidimensional Data

**Data** which related to more than **two dimensions**.

**[1D]**

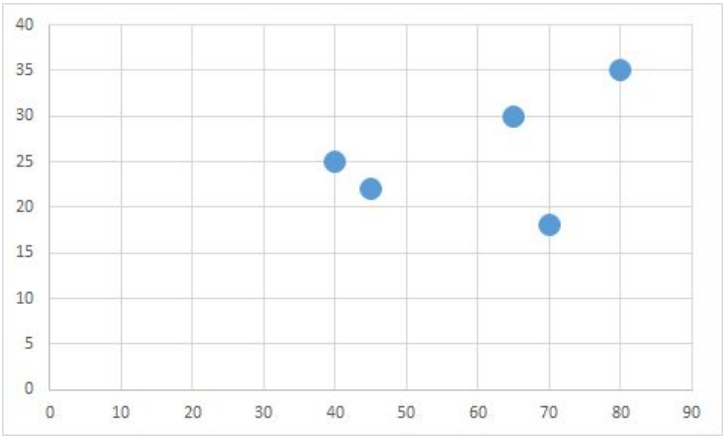
| | Weight |
|---|---|
| S1 | 80 |
| S2 | 40 |
| S3 | 45 |
| S4 | 65 |
| S5 | 70 |

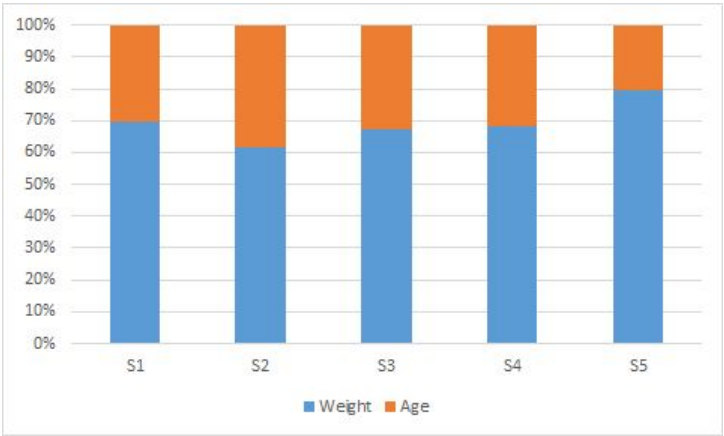


# [2D]

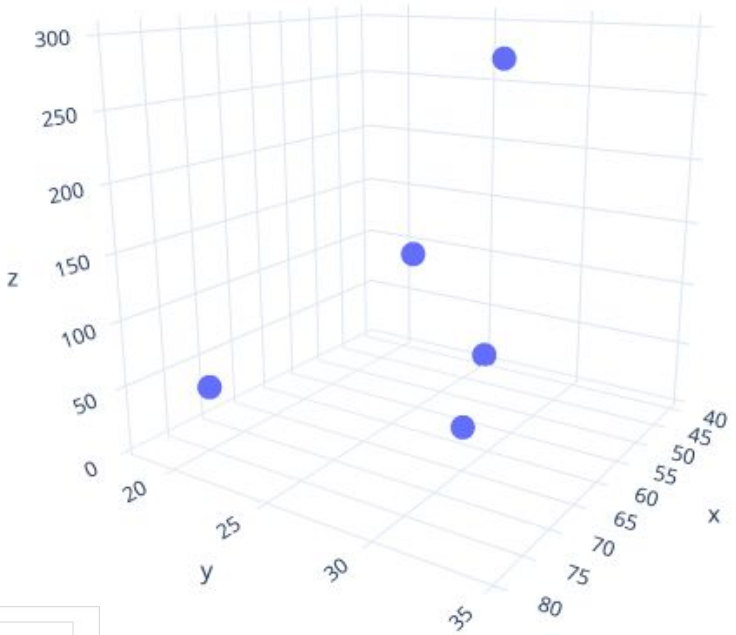
| | Weight | Age |
|---|---|---|
| S1 | 80 | 35 |
| S2 | 40 | 25 |
| S3 | 45 | 22 |
| S4 | 65 | 30 |
| S5 | 70 | 18 |

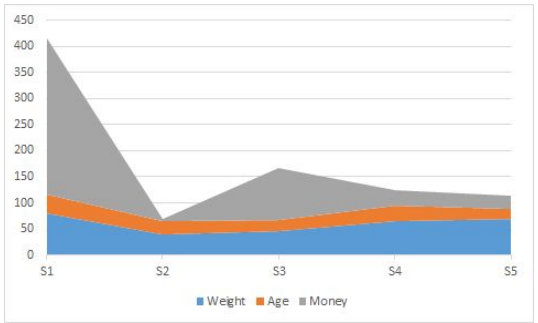


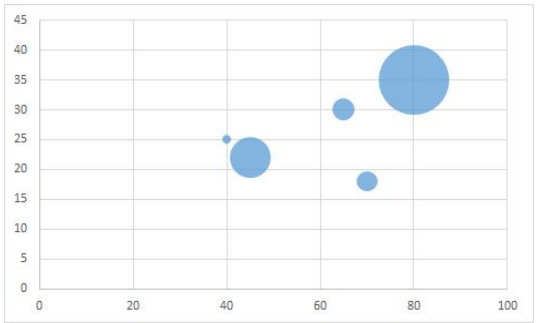



# [3D]

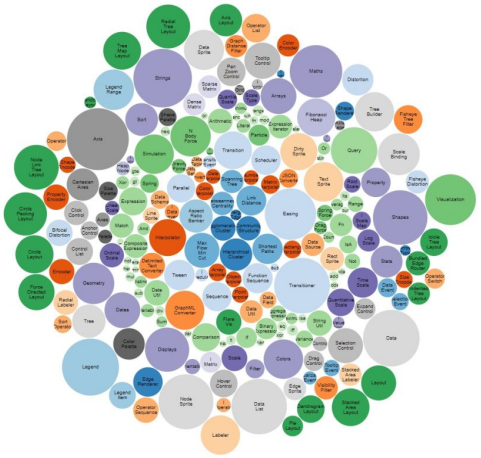
| | Weight | Age | Money |
|---|---|---|---|
| S1 | 80 | 35 | 300 |
| S2 | 40 | 25 | 5 |
| S3 | 45 | 22 | 100 |
| S4 | 65 | 30 | 30 |
| S5 | 70 | 18 | 25 |

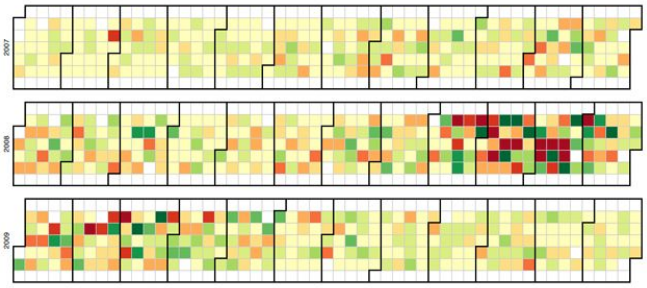


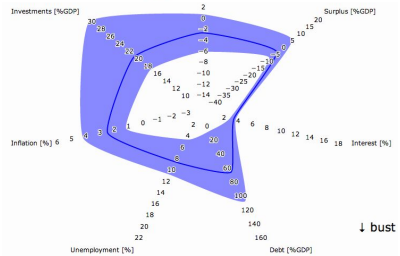
# [4D]<= ?

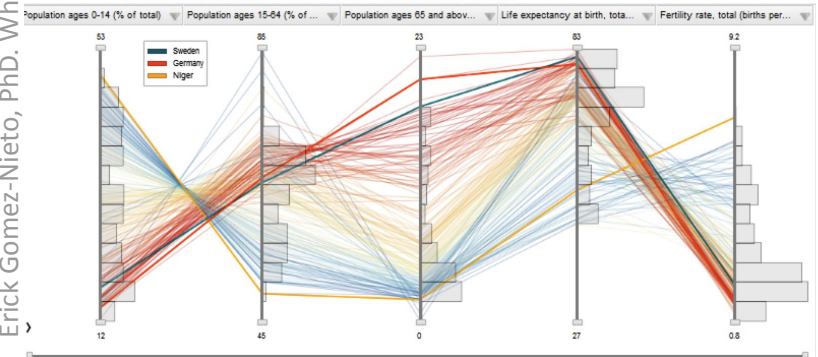


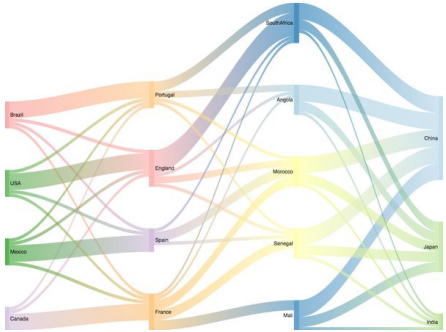


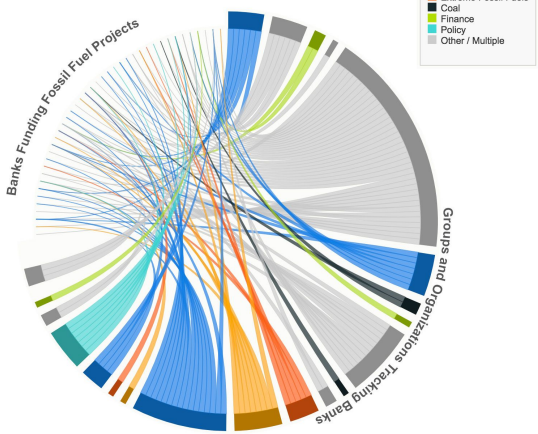
Treemaps

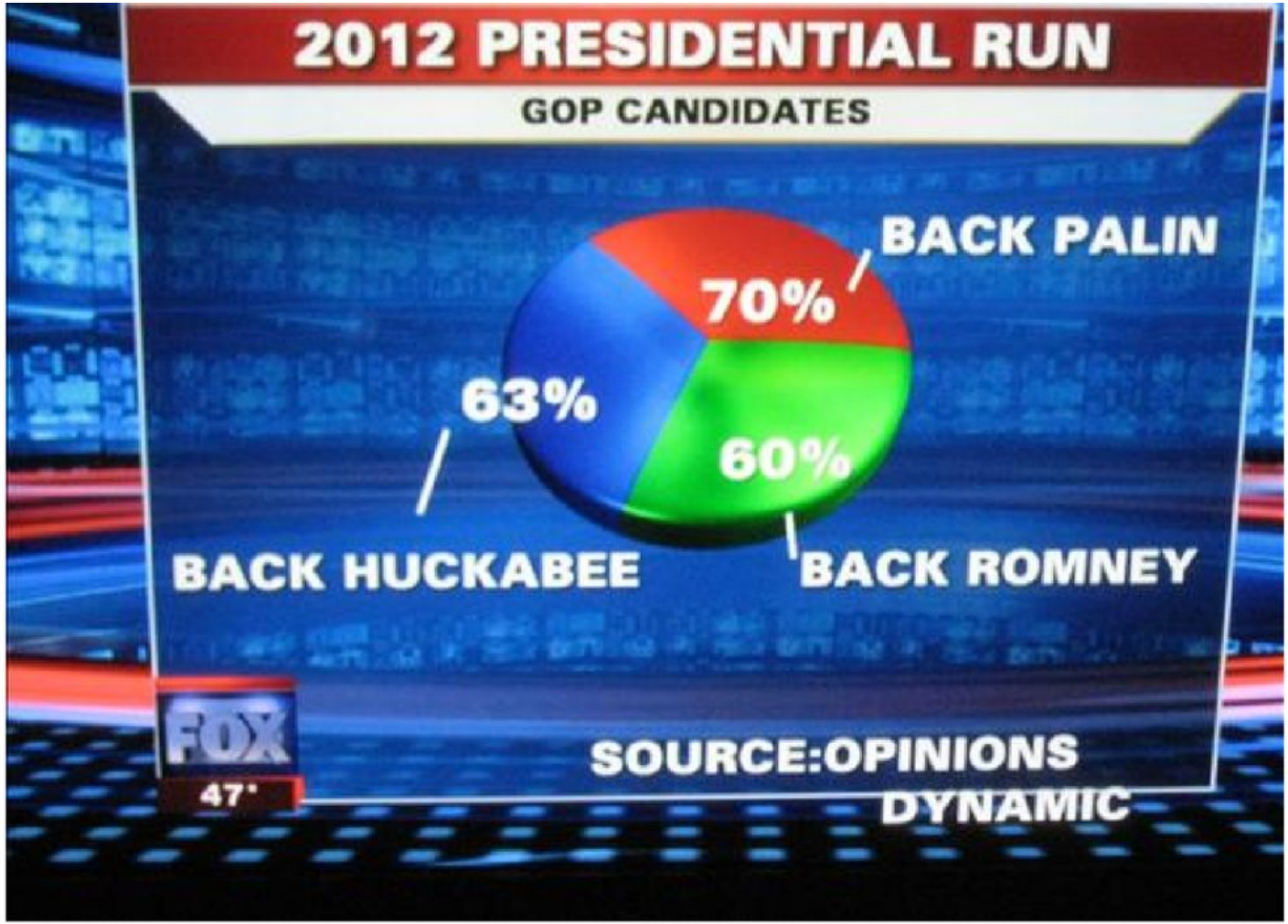
# [4D]<= ?

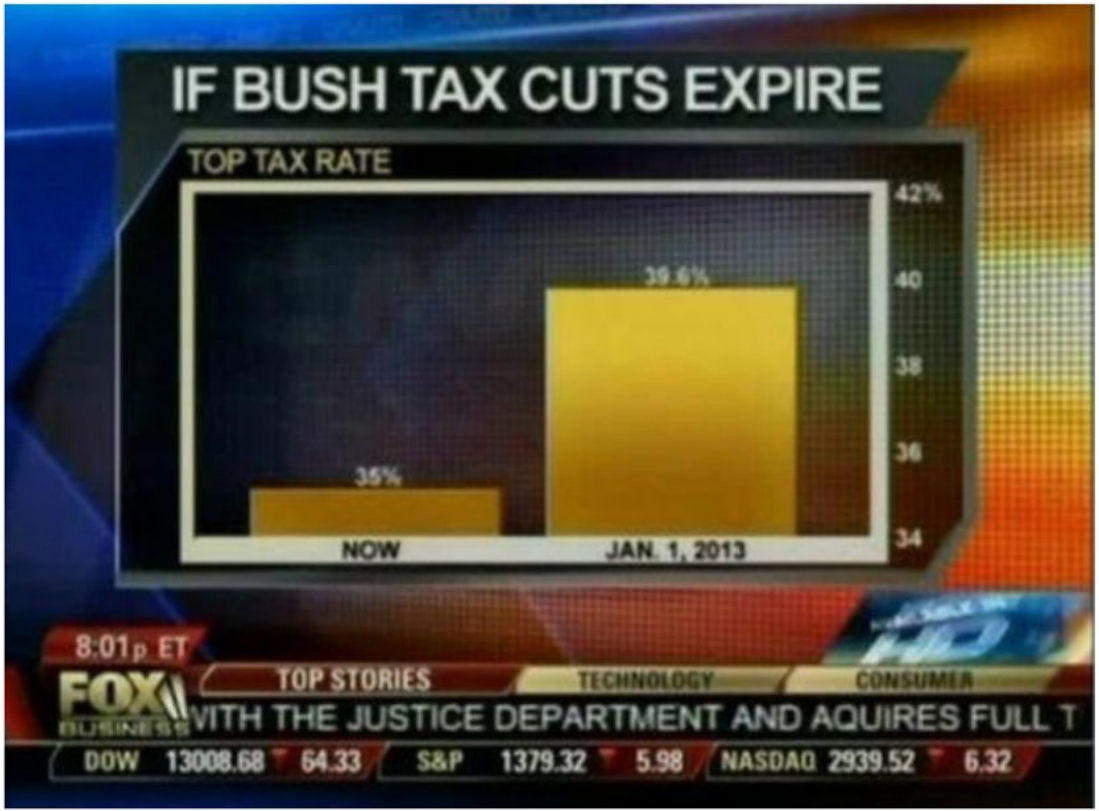


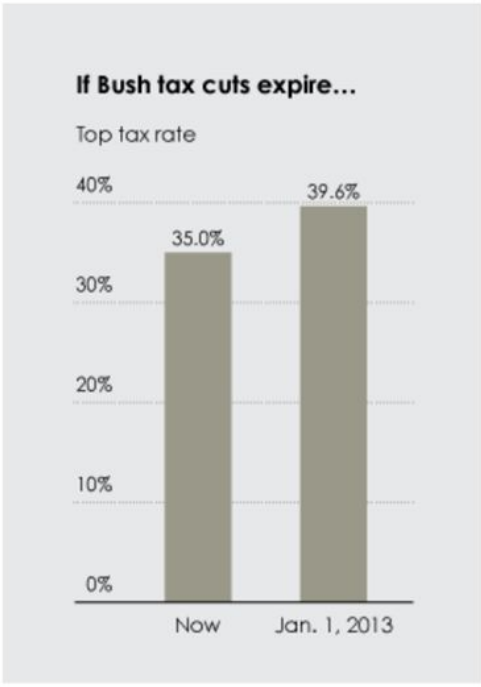



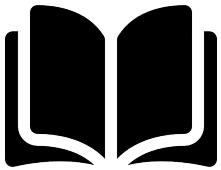











# Visualization Course Material

https://sites.google.com/site/erickgomeznieto/teaching/tcg2017

# Introduction - Types of data

Most techniques are designed to handle only **numeric** data.

Some other handle just **categorical** data.

Real data can contain **mixed multidimensional features**.

| Type1 | Type2 | HP | Attack | Defense | SpAtk | SpDef | Speed | isLegendary | Color | EggGroup1 | EggGroup2 | Height_m | Weight_kg | Body_Style |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Psychic | Psychic | 100 | 100 | 100 | 100 | 100 | 100 | False | Pink | Undiscovere | Undiscovere | 0.41 | 4 | bipedaltaile |
| Normal | Normal | 30 | 56 | 35 | 25 | 35 | 72 | False | Purple | Field | Field | 0.3 | 3.5 | quadruped |
| Electric | Steel | 50 | 60 | 95 | 120 | 70 | 70 | False | Grey | Mineral | Mineral | 0.99 | 60 | multiplebod |
| Fire | Fire | 90 | 110 | 80 | 100 | 80 | 95 | False | Brown | Field | Field | 1.91 | 155 | quadruped |
| Fire | Fire | 65 | 100 | 70 | 80 | 80 | 105 | False | Yellow | Field | Field | 1.7 | 95 | quadruped |
| Normal | Normal | 105 | 95 | 80 | 40 | 80 | 90 | False | Brown | Monster | Monster | 2.21 | 80 | bipedaltaile |
| Normal | Normal | 250 | 5 | 5 | 35 | 105 | 50 | False | Pink | Fairy | Fairy | 1.09 | 34.6 | bipedaltaile |
| Rock | Water | 35 | 40 | 100 | 90 | 55 | 35 | False | Blue | Water1 | Water3 | 0.41 | 7.5 | severallimbs |
| Poison | Ground | 90 | 92 | 87 | 75 | 85 | 76 | False | Blue | Undiscovere | Undiscovere | 1.3 | 60 | bipedaltaile |
| Fighting | Fighting | 80 | 100 | 70 | 50 | 60 | 45 | False | Grey | HumanLike | HumanLike | 1.5 | 70.5 | bipedaltaille |
| Normal | Normal | 90 | 55 | 75 | 60 | 75 | 30 | False | Pink | Monster | Monster | 1.19 | 65.5 | bipedaltaile |
| Water | Ice | 130 | 85 | 80 | 85 | 95 | 60 | False | Blue | Monster | Water1 | 2.49 | 220 | withfins |
| Grass | Poison | 80 | 82 | 83 | 100 | 100 | 80 | False | Green | Monster | Grass | 2.01 | 100 | quadruped |
| Normal | Flying | 40 | 45 | 40 | 35 | 35 | 56 | False | Brown | Flying | Flying | 0.3 | 1.8 | twowings |
| Normal | Flying | 60 | 110 | 70 | 60 | 60 | 100 | False | Brown | Flying | Flying | 1.8 | 85.2 | headlegs |

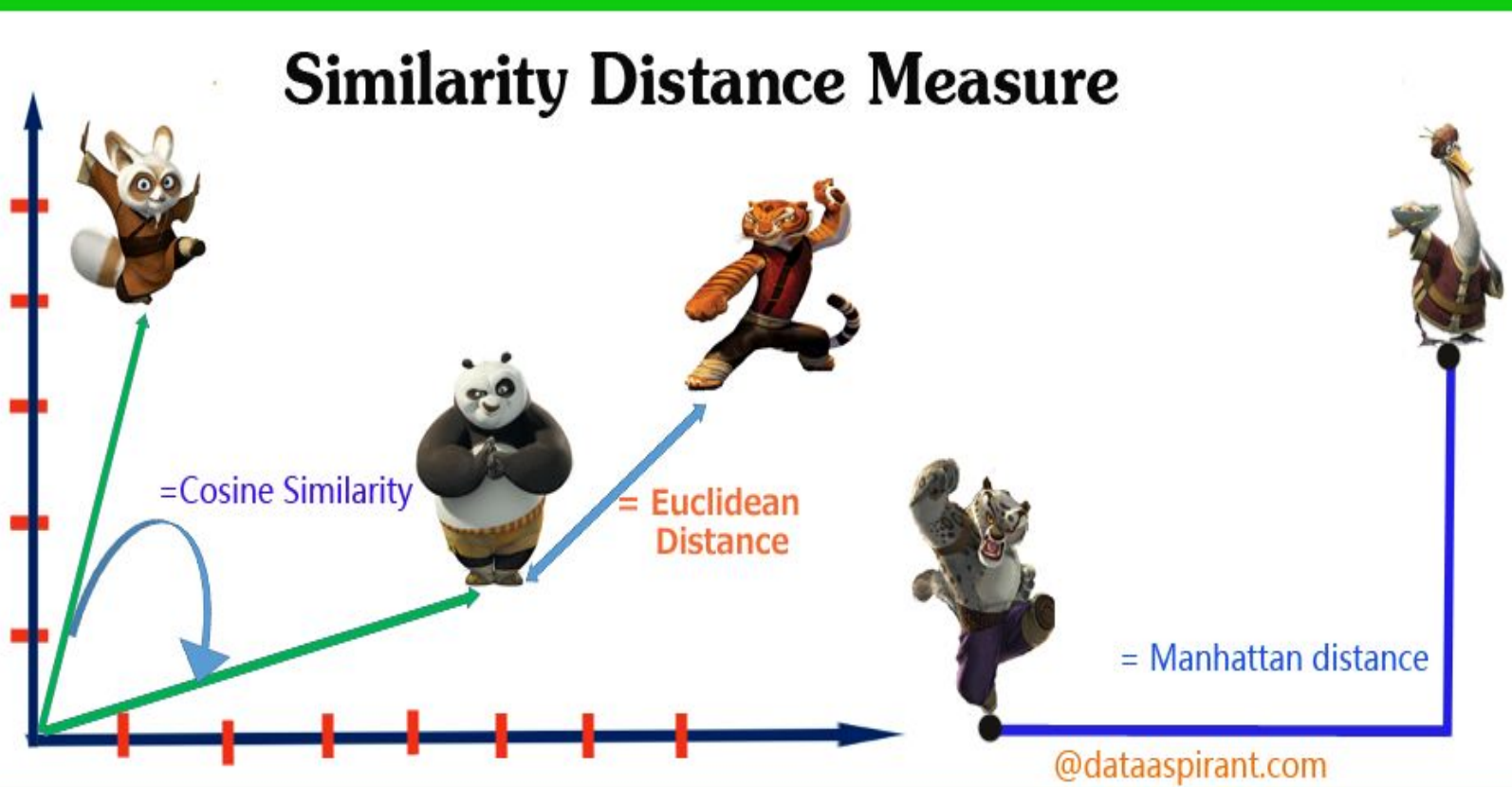
# Distances (similarity)

# Euclidean Distance

Euclidean

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

@dataaspirant.com



# Jaccard Similarity

Set A = {  }

Set B = {  }

|A| = 4 |B| = 5 @dataaspirant.com

Union(A,B) = {  }

Intersection (A,B) = {  }

| Union (A,B) | = 7

| Intersection (A,B)| = 2

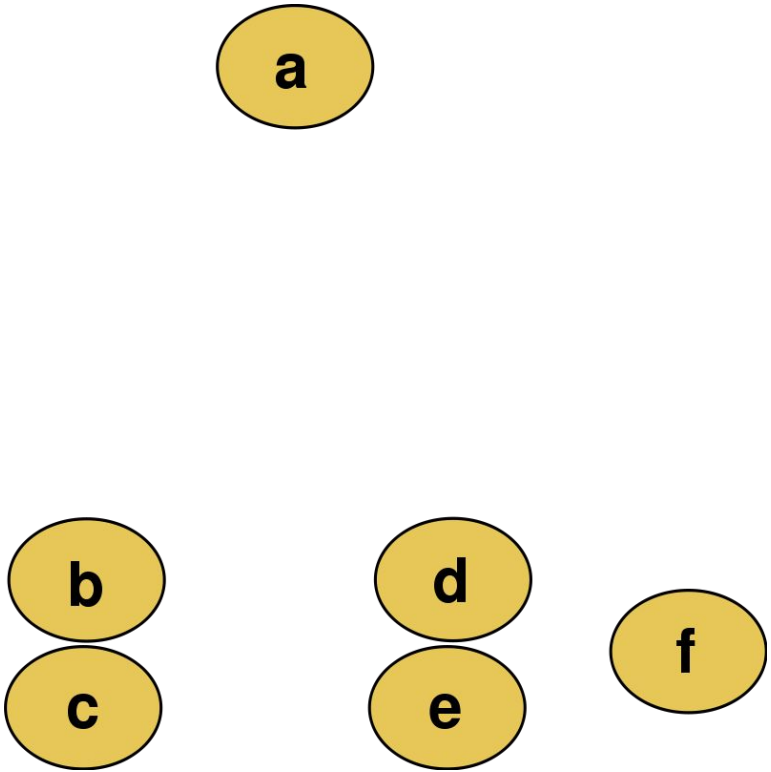
Jaccard Similarity J (A,B) = | Intersection (A,B) | / | Union (A,B) |

= 2 / 7

= 0.286

@dataaspirant.com

# Distance Matrix




| | a | b | c | d | e | f |
|---|---|---|---|---|---|---|
| a | 0 | 184 | 222 | 177 | 216 | 231 |
| b | 184 | 0 | 45 | 123 | 128 | 200 |
| c | 222 | 45 | 0 | 129 | 121 | 203 |
| d | 177 | 123 | 129 | 0 | 46 | 83 |
| e | 216 | 128 | 121 | 46 | 0 | 83 |
| f | 231 | 200 | 203 | 83 | 83 | 0 |

| | a | b | c | d | e | f |
|---|---|---|---|---|---|---|
| **a** | 0 | 184 | 222 | 177 | 216 | 231 |
| **b** | 184 | 0 | 45 | 123 | 128 | 200 |
| **c** | 222 | 45 | 0 | 129 | 121 | 203 |
| **d** | 177 | 123 | 129 | 0 | 46 | 83 |
| **e** | 216 | 128 | 121 | 46 | 0 | 83 |
| **f** | 231 | 200 | 203 | 83 | 83 | 0 |