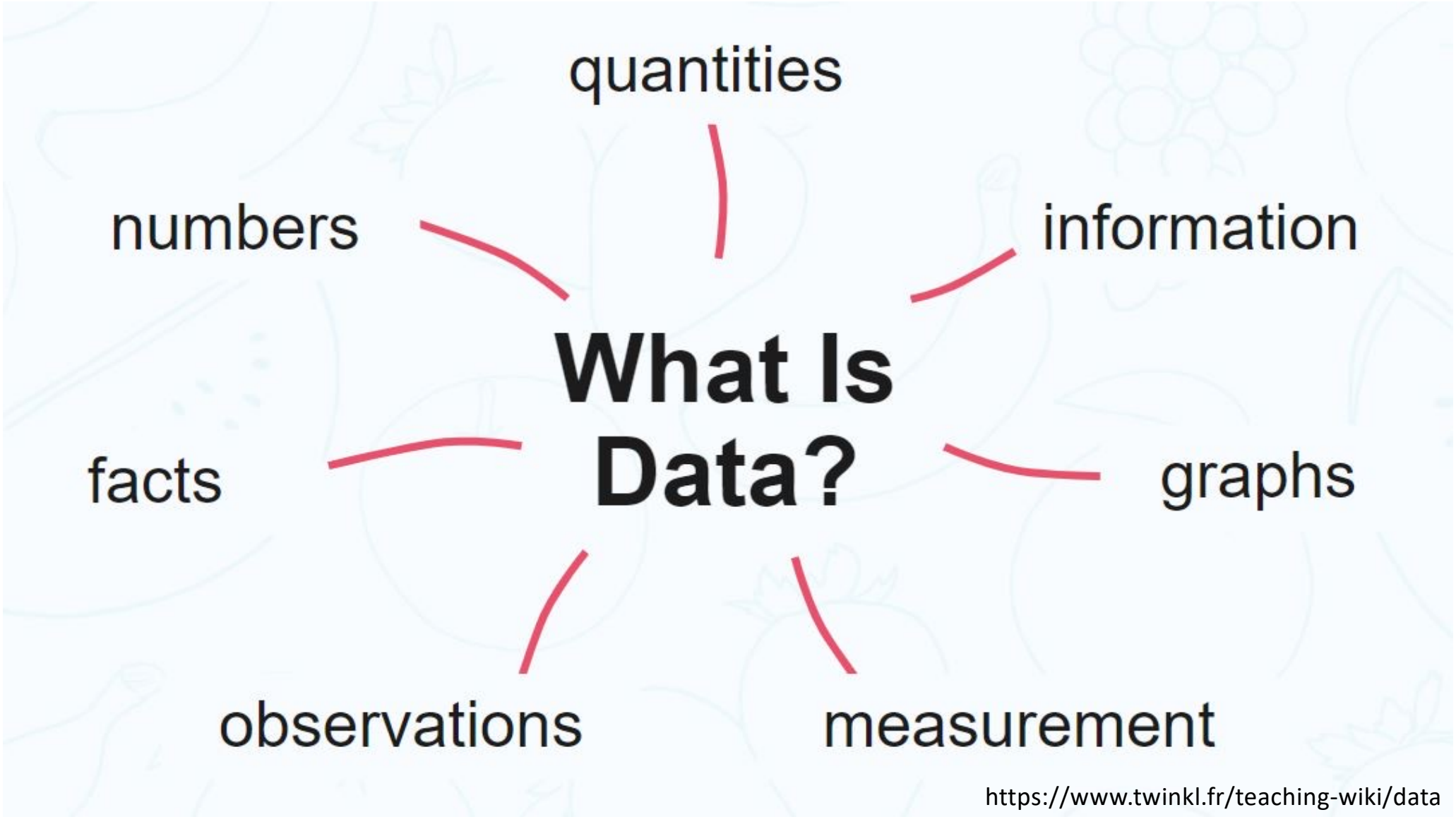
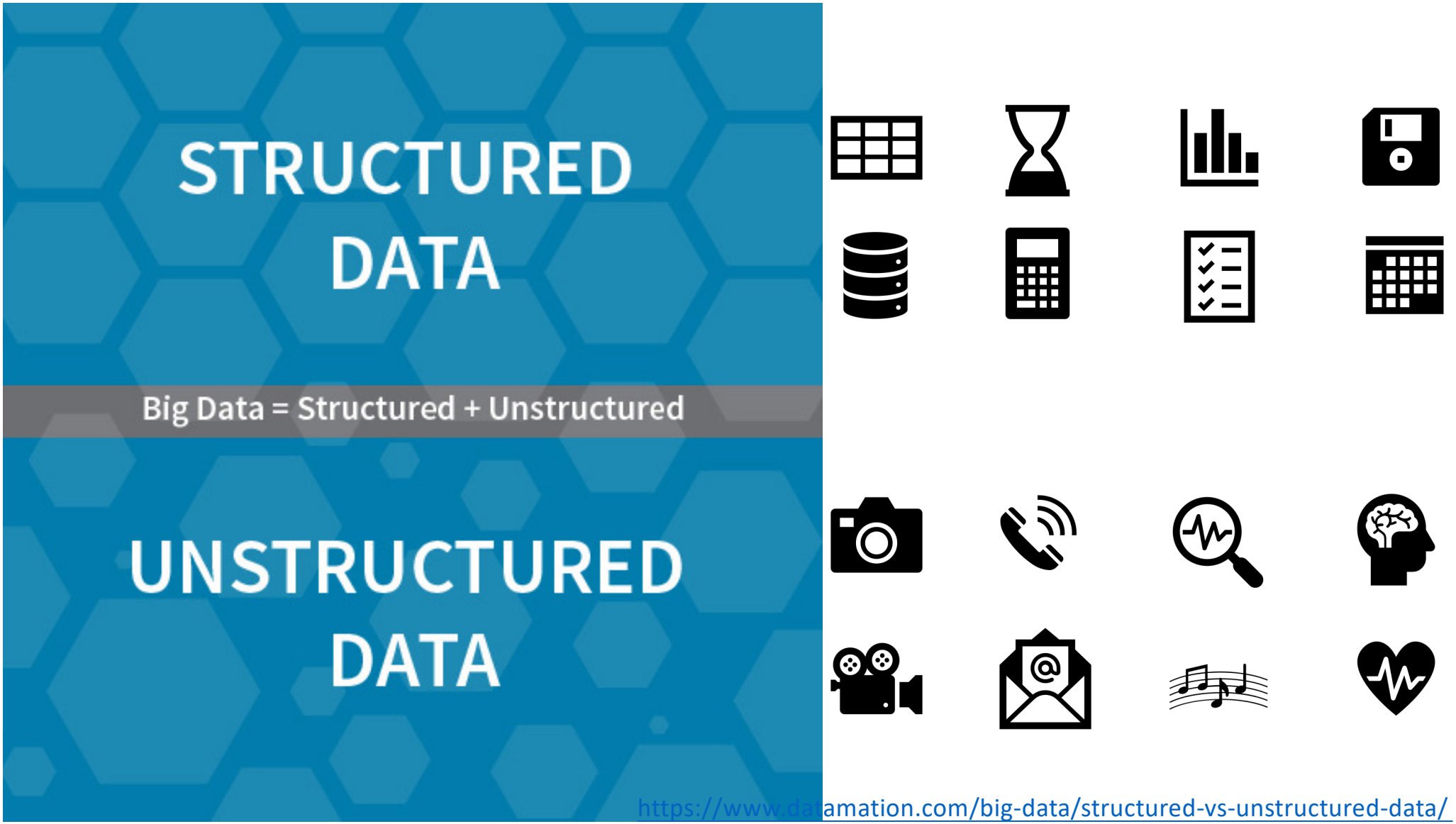
Week 1 Preview

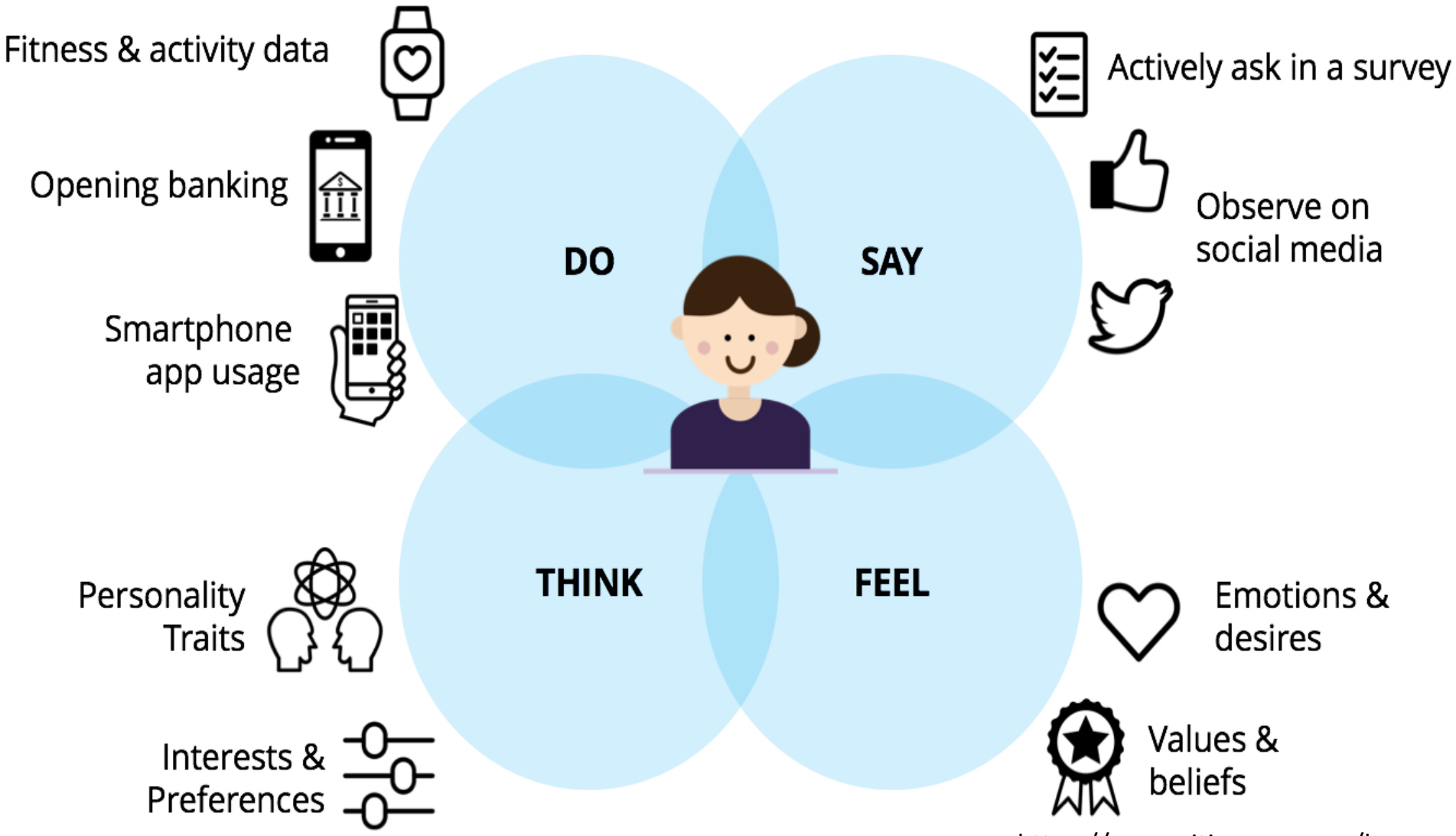
- What is Data Science?
  - definition of data and data science
  - brief history and modern development
- Data Science pipeline overview
  - end-to-end process of a data science project
  - a real data example with baby names
- Tools you will need for this course
  - Python and Jupyter notebook



https://www.twinkl.fr/teaching-wiki/data

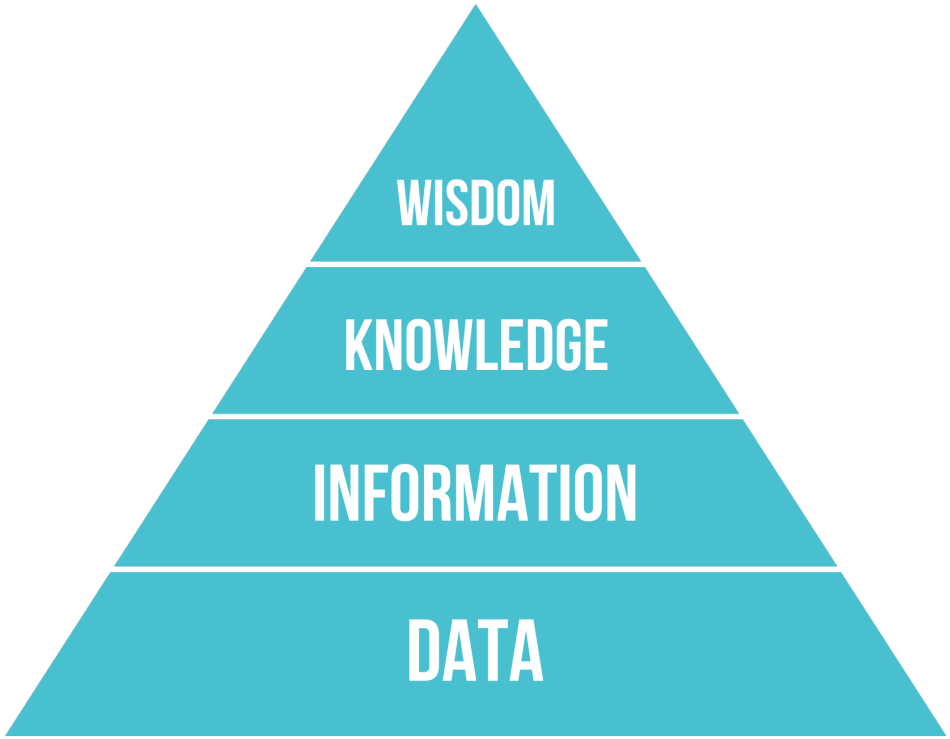


https://www.datamation.com/big-data/structured-vs-unstructured-data/



https://www.citizenme.com/human-data/

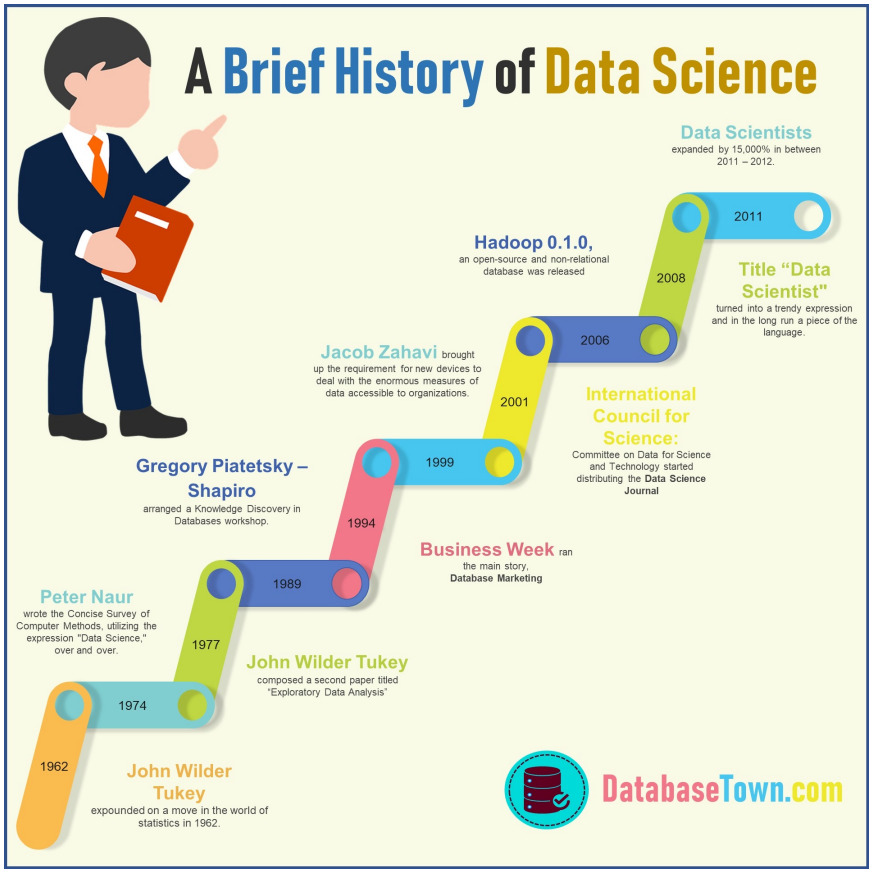
The DIKW pyramid



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

A definition of Data Science

- Draw useful information and actionable knowledge from data using computation, so that we can better understand the world and solve problems!
- Data Science is fundamentally interdisciplinary!
- The fourth paradigm of science:
  - theoretical – experimental –simulation – data-intensive
- Technology trends:
  - hardware industry – software industry - Internet industry – data industry – ?



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

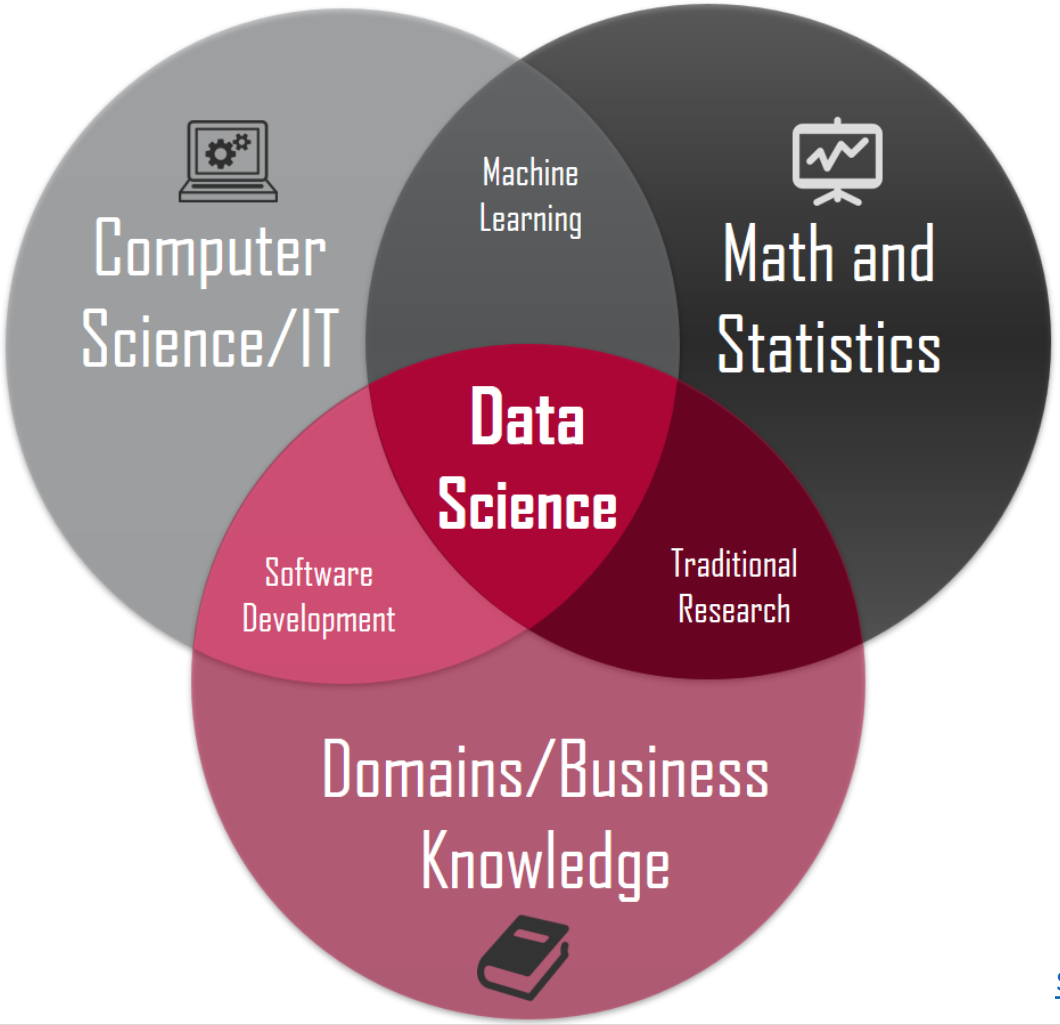
- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

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

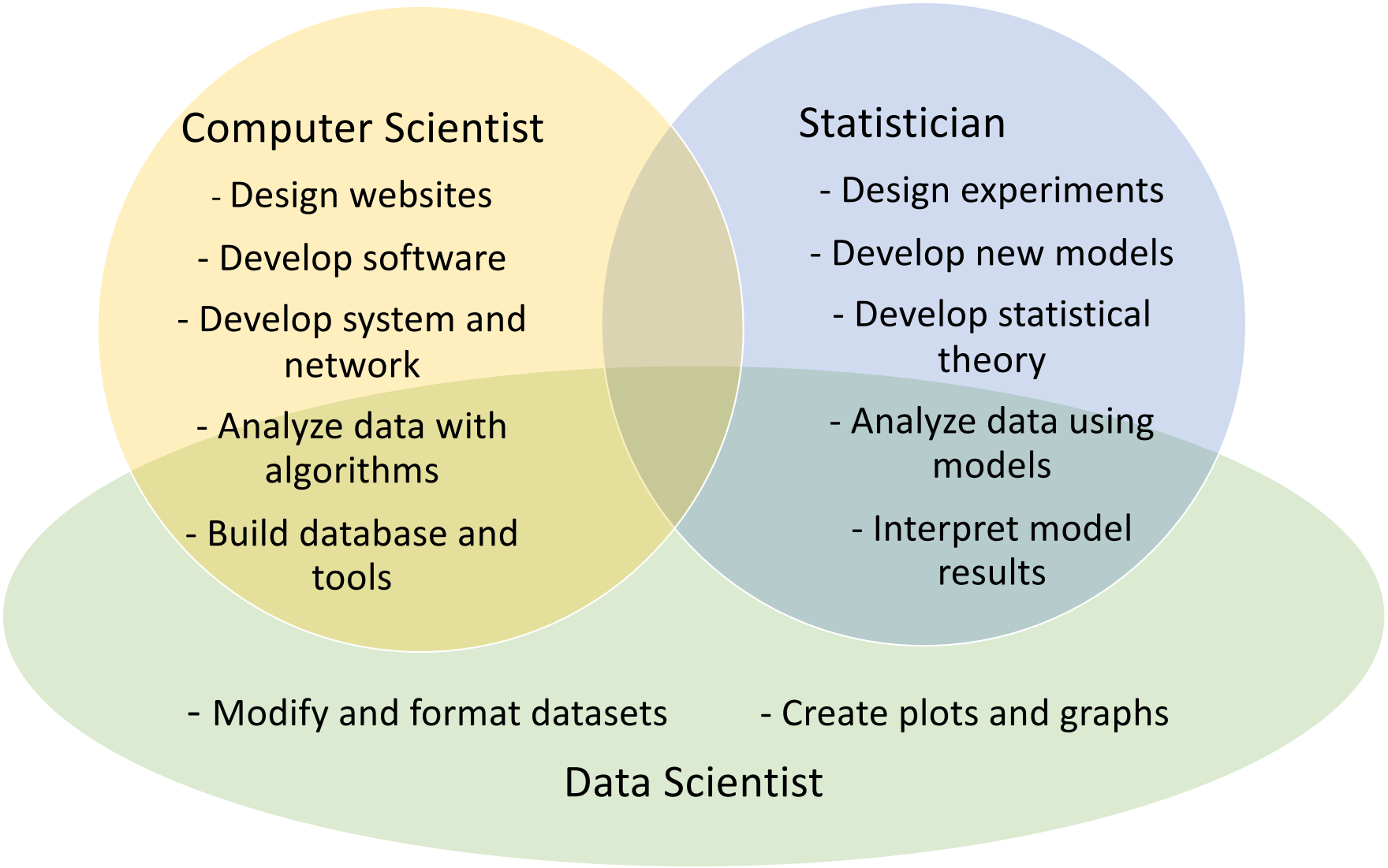
COMMUNICATION & VISUALIZATION

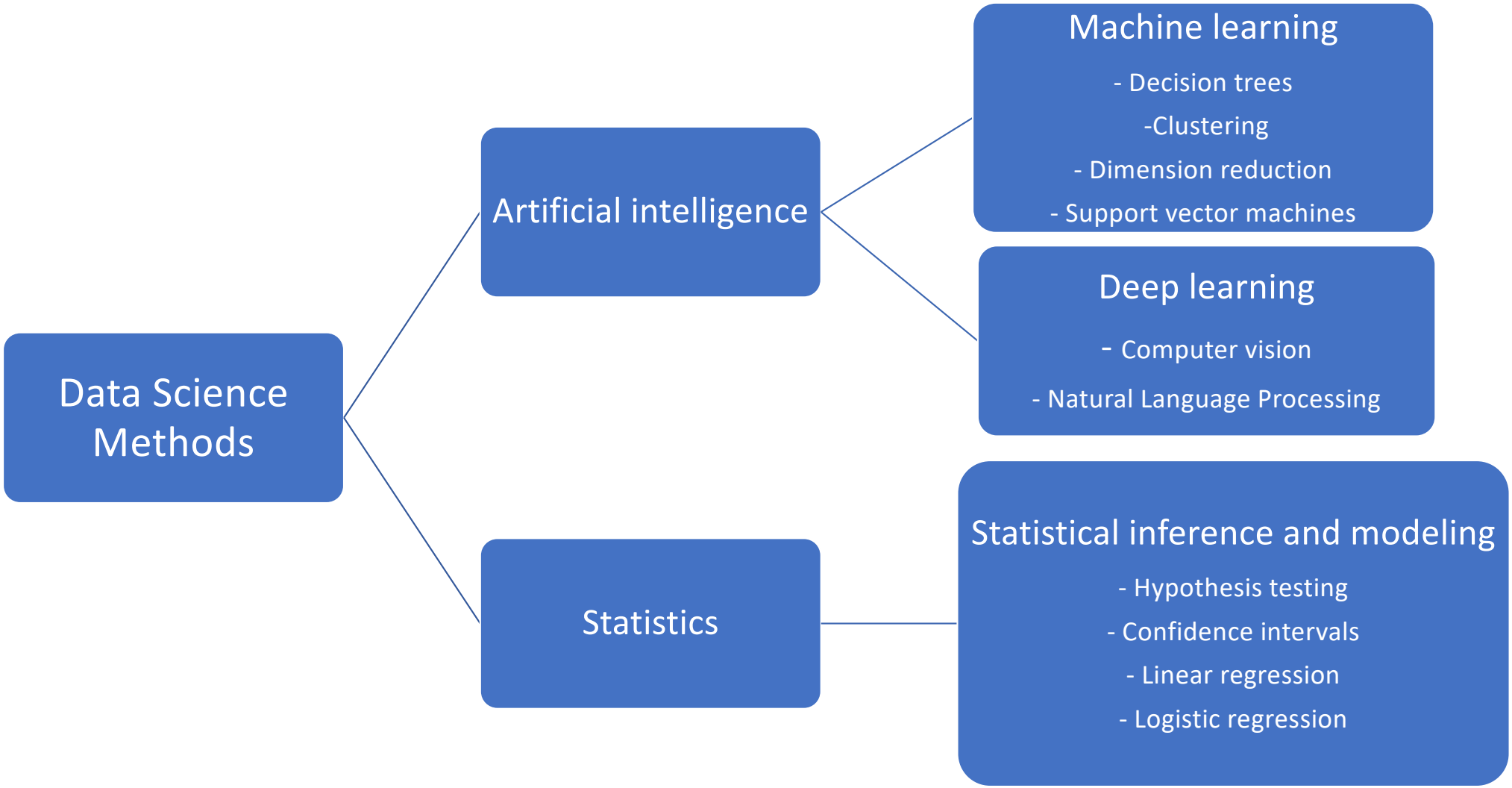
- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau



source

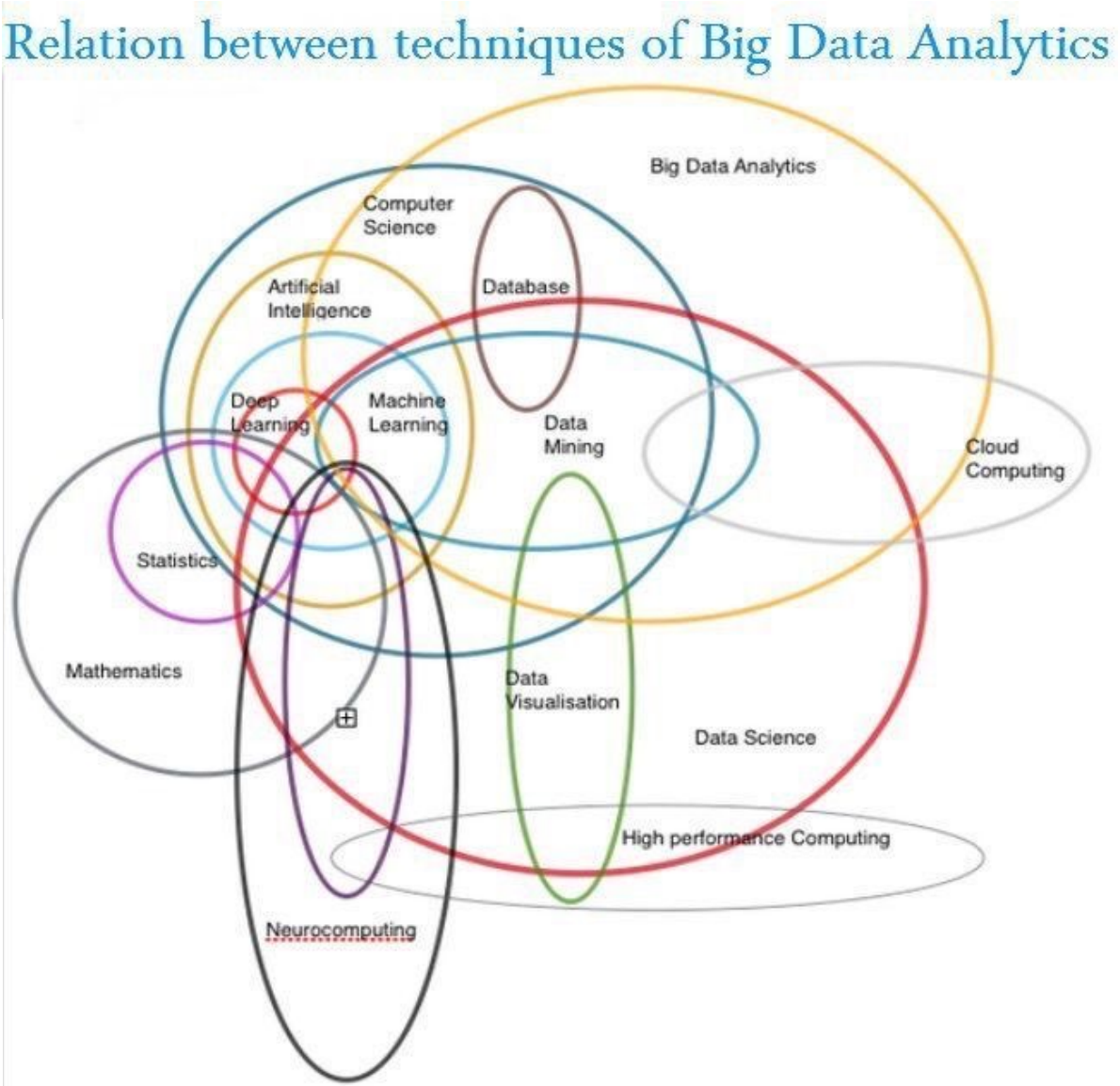
Skills of Data Science


Computer Scientist
- Design websites
- Develop software
- Develop system and network
- Analyze data with algorithms
- Build database and tools
Statistician
- Design experiments
- Develop new models
- Develop statistical theory
- Analyze data using models
- Interpret model results
- Modify and format datasets
- Create plots and graphs
Data Scientist


Data Science Methods
Artificial intelligence
Machine learning
- Decision trees
-Clustering
- Dimension reduction
- Support vector machines
Deep learning
- Computer vision
- Natural Language Processing
Statistics
Statistical inference and modeling
- Hypothesis testing
- Confidence intervals
- Linear regression
- Logistic regression

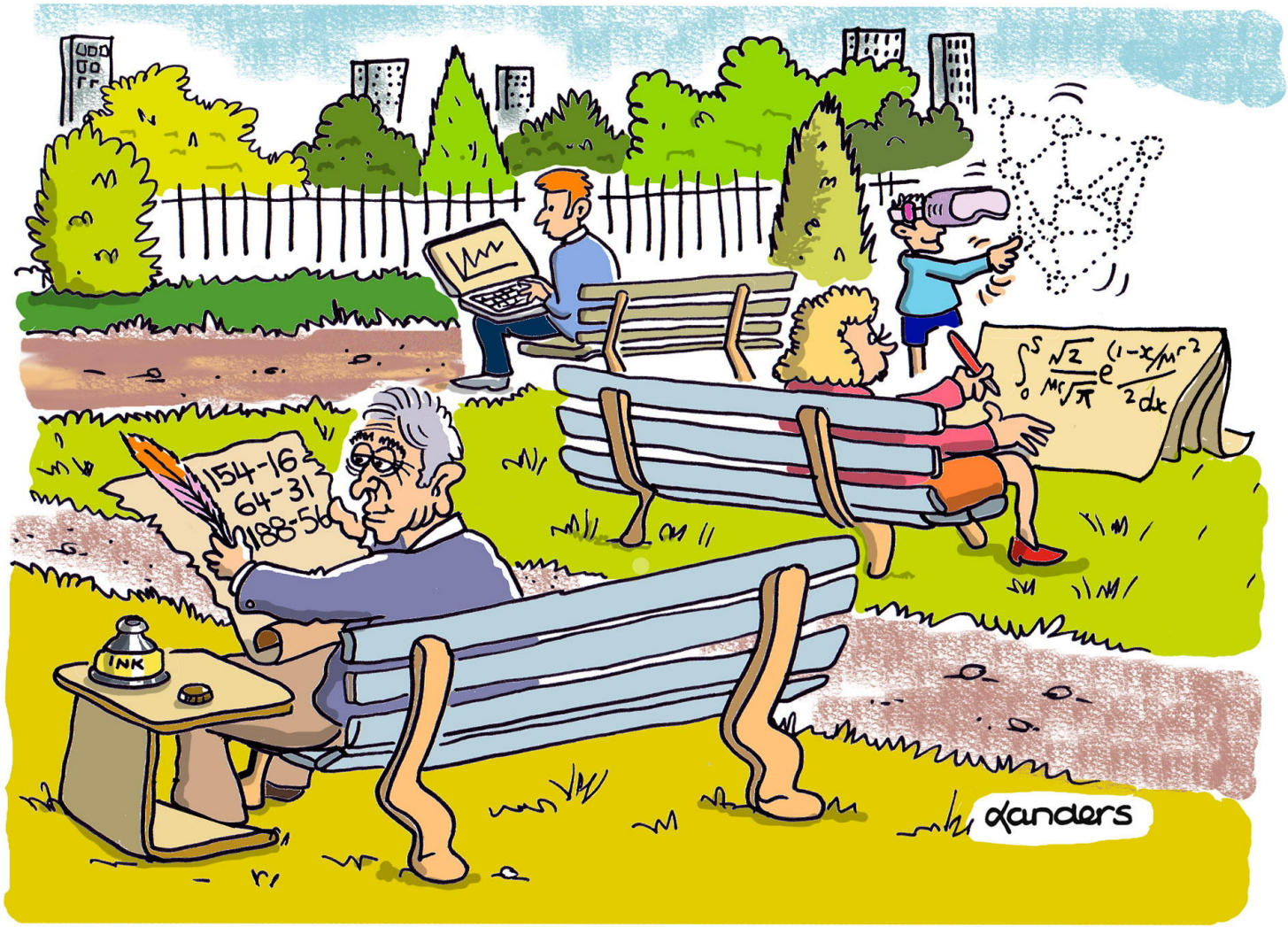
When data gets BIG…

source




J.P.Morgan
Meta
salesforce
Adobe
aws
Johnson & Johnson
IBM
VISA
MasterCard
Deloitte.
intel
ORACLE
airbnb
Honeywell
citibank
RSM
Google
Microsoft
CapitalOne
accenture
verizon
AT&T
Walmart
Save money. Live better.
comcast
twitter
PayPal
pwc
xfinity
Roche
TikTok
TESLA
BOEING

Job Title	Description
Data engineers	Data engineers specialize in data gathering and storage. Data engineers extract, transform, and load datasets for later analysis.
Data scientists	Data scientists gather data, transform data, and use models and algorithms to extract meaningful insights from datasets.
Data analysts	Data analysts work with industry experts to analyze datasets and create visualizations. Data analysts use some data science models, but tend to use data visualization and summary more than modeling.
Business intelligence analysts	Business intelligence analysts specialize in data related to financial and market transactions. Data analysts and business intelligence analysts are similar roles, but the term business intelligence is more common in business and finance.
Machine learning engineers	Machine learning engineers specialize in machine learning models instead of statistical models. Machine learning engineers often focus on the implementation and development of a model rather than selection and interpretation.



The many generations of data science

The data science pipeline

- a high-level description of the data science workflow

Once upon a time, there's a boy named Data

[A Beginner's Guide to the Data Science Pipeline]

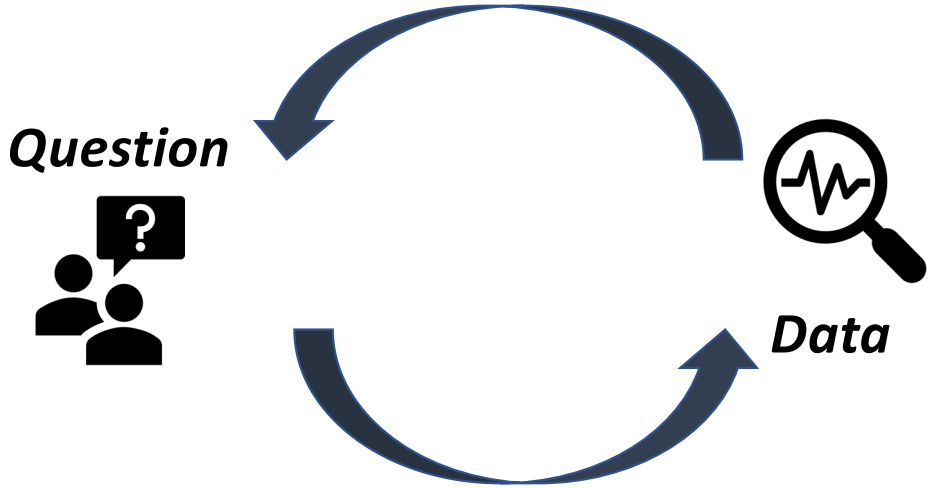
DS Pipeline is **OSEMN**

- **O** → Obtaining data
- **S** → Scrubbing / Cleaning data
- **E** → Exploring / Visualizing
- **M** → Modeling / Predicting
- **N** → iNterpreting / Communicating

Obtaining data

Objective:

- The feedback loop between:



Skills Required:

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

Scrubbing and cleaning

Objective:

- “Clean” data

Skills Required:

- Scripting language: Python, R, SQL
- Data Wrangling Tools: Python Pandas, R tidyverse
- Distributed Processing: Hadoop, Map Reduce / Spark

Exploring

Objective:

- Get to know your data

Skills Required:

- Inferential Statistics
- Data Visualization
- Feature Engineering

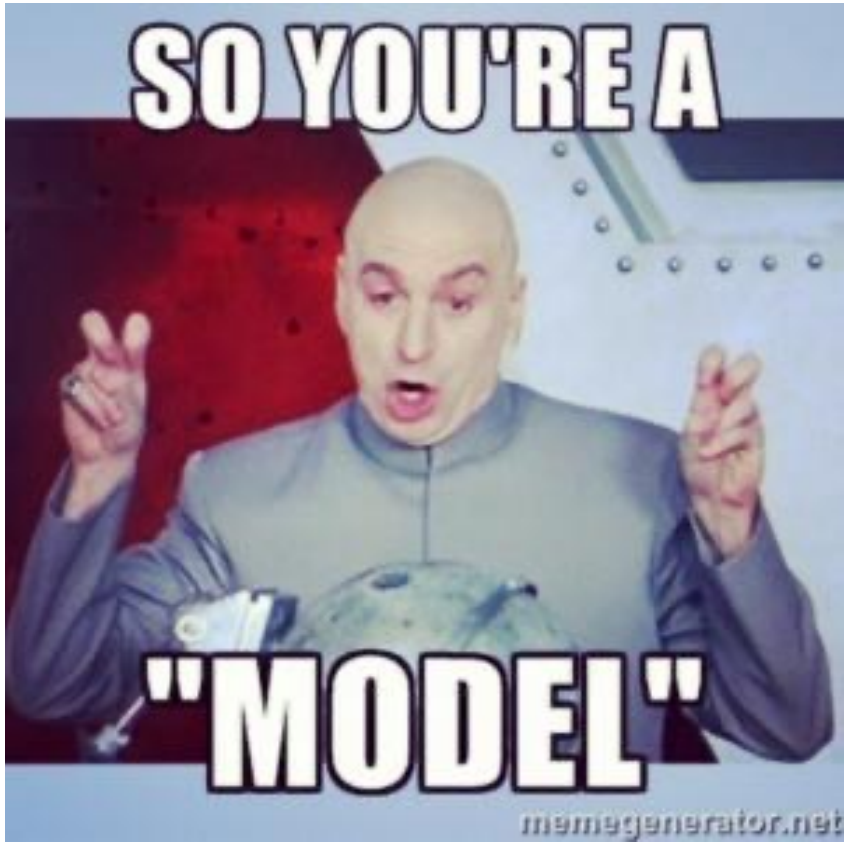
Modeling

Objective:

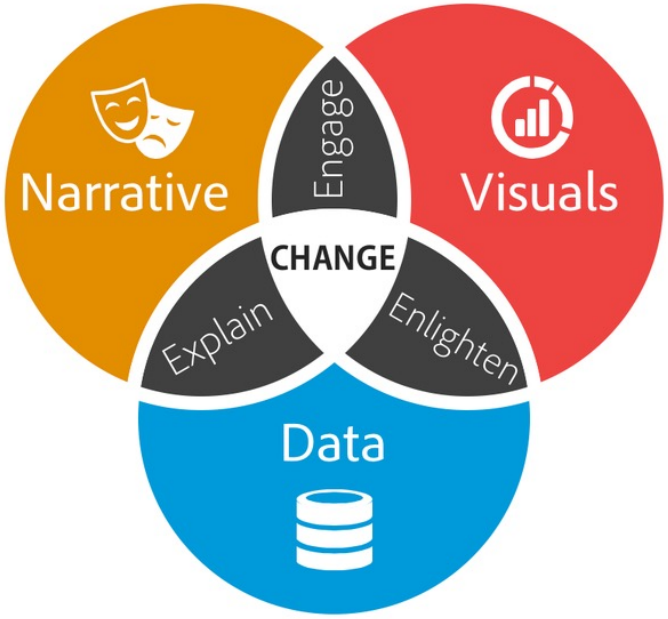
- Prediction / decision making
- Data driven discovery

Skills Required:

- Statistical models
- Machine learning models
- Linear algebra & multivariate Calculus



iNterpretation



Objective:

- Data storytelling

Skills Required:

- Business Domain Knowledge
- Data Visualization Tools
- Communication: Presentation skills & Report writing

Ethics in Data Science

Data ownership & privacy

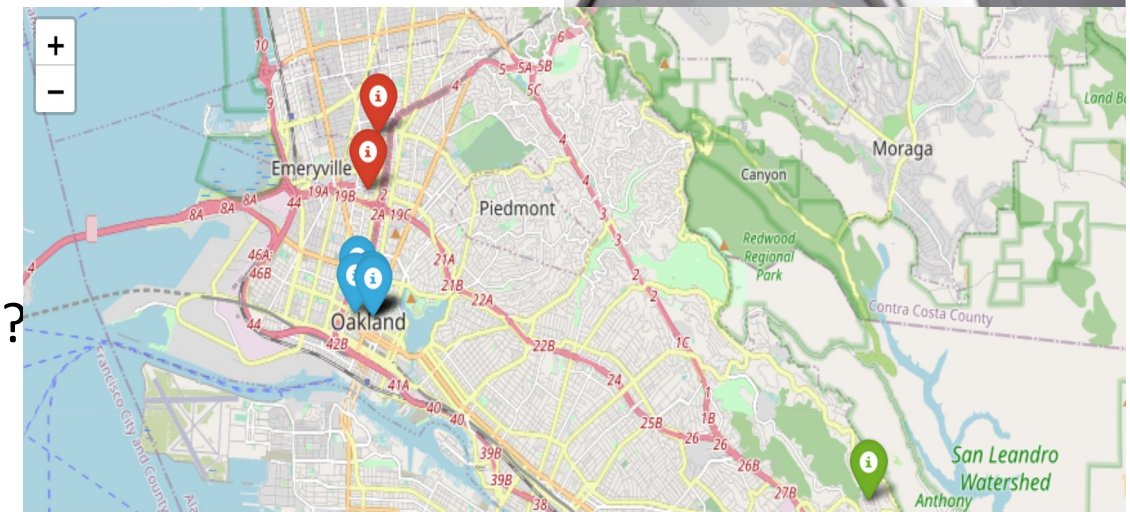
Algorithmic bias & fairness

Environmental impacts

Data Privacy

- Automated license plate readers
  - Oakland public records site
  - license plate, time, location
  - about 2.7 million records

- What can we find?
  - home address
  - work address
  - spend weekend somewhere?



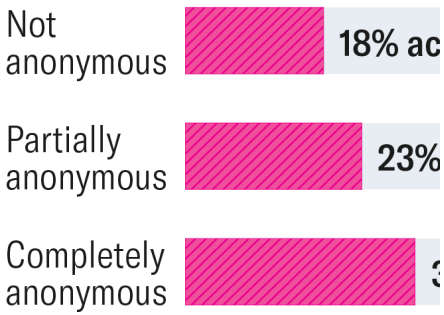
Data Privacy Laws & Regulations

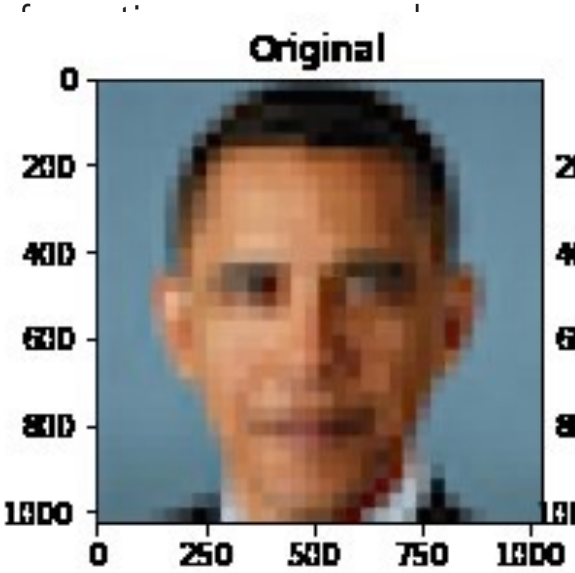
- The Family Educational Rights and Privacy Act (**FERPA**, 1974)
  - rules about access and release of students' information and education records.
- The Health Insurance Portability and Accountability Act (**HIPAA**, 1996)
  - regulates how patient health information is stored and shared.
- The European Union passed the General Data Protection Regulation (**GDPR**) in 2016.
  - governs how companies store, process, and use an individual's personal data.

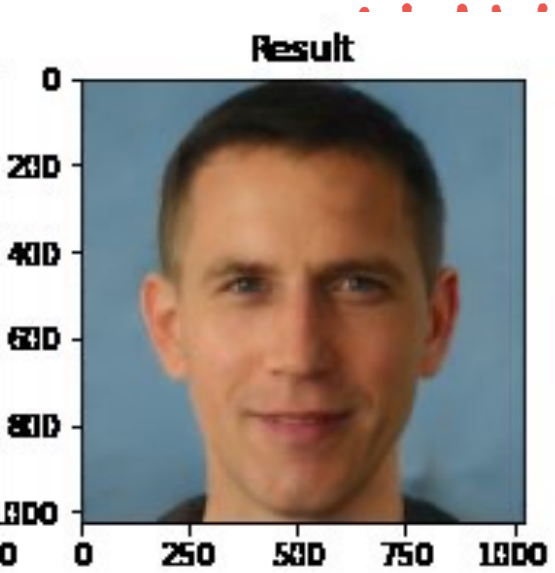
Algorithmic bias & fairness

How Anonymizing Applications Helps Women in Science

When gender-identifying i
from applications for rese
Hubble Space Telescope
were significantly more li





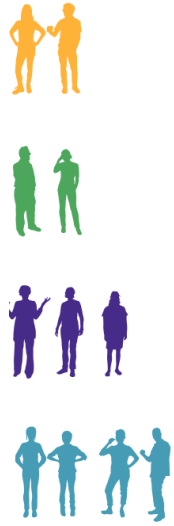


Note: On average, women represented 23% of submitted applications.
Source: Stefanie K. Johnson and Jessica F. Kirk

HBR

Applicants of color denied at higher rates

To illustrate the odds of denial that our analysis revealed, we calculated how many people of each race/ethnic group would likely be denied if 100 similarly qualified applicants from each group applied for mortgages in the United States



9 Black applicants denied

Assignment for this week

- Install Python 3 with the standard modules from an Anaconda installation (lab assignment for this week)

- We will run the Demo01 notebook in class
  - Download and make sure you know how to open a jupyter notebook

- Survey 0: All About You