Data Science

https://github.com/berradais/DNA-2022-2023

Data is Big!

- 2.5 quintillion (10^{18}) bytes of data are generated every day!
- Everything around you collects/generates data (about 87 % of websites)
  - Social media sites
  - Business transactions
  - Location-based data
  - Sensors
  - Digital photos, videos
  - Consumer behavior (online and store transactions)
- More data is publicly available
- Database technology is advancing
- Cloud based & mobile applications are widespread

Source: IBM http://www-01.ibm.com/software/data/bigdata/

If I have data, I will know :)

- Everyone wants better predictability, forecasting, customer satisfaction, market differentiation, prevention, great user experience, …

  - How can I price a particular product?

  - What can I recommend online customers to buy after buying X, Y or Z?

  - How can we discover market segments? group customers into market segments?

  - What customer will buy in the upcoming holiday season? (what to stock?)

  - What is the price point for customer retention for subscriptions?

Data Science is making sense of Data

- Lots of Data => Lots of Analysis => Lots of Jobs

  - Multidisciplinary study of data collections for analysis, prediction, learning and prevention.

  - Utilized in a wide variety of industries.

  - Involves both structured or unstructured data sources.

Data Science is multidisciplinary

- Statisticians
- Mathematicians
- Computer Scientists in
  - Data mining
  - Artificial Intelligence & Machine Learning
  - Systems Development and Integration
  - Database development
  - Analytics
- Domain Experts
  - Medical experts
  - Geneticists
  - Finance, Business, Economy experts
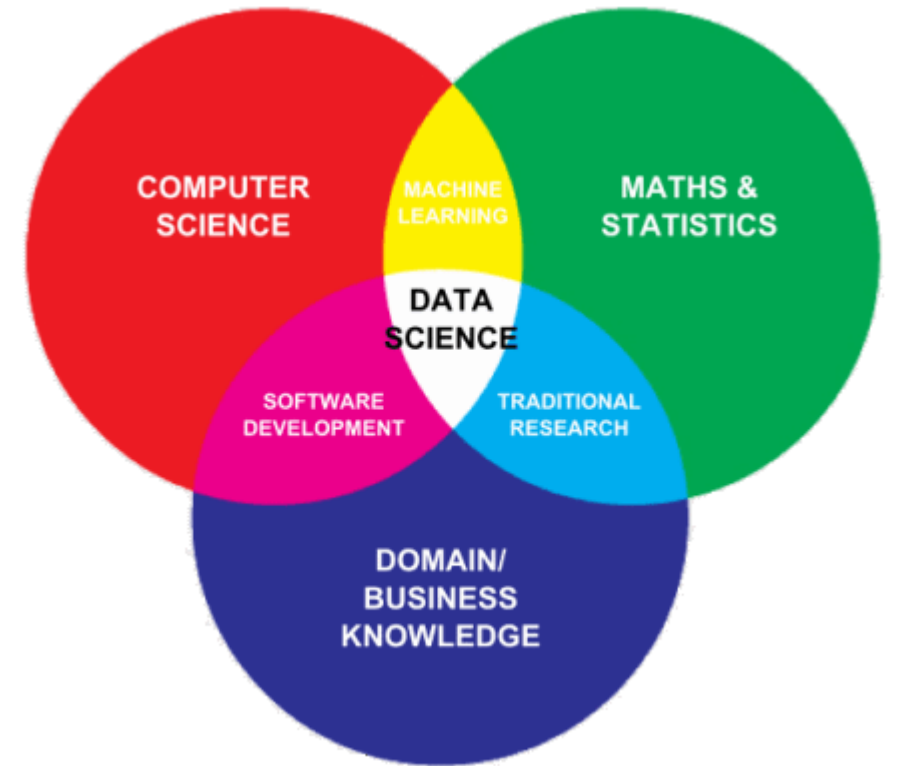
Data Science is about the whole processing pipeline to extract information out of data

Data Science Pipeline

- Three major steps:

  1. Preparing to run a model: gathering, cleaning, integrating, restructuring, transforming, loading, filtering, deleting, combining, merging, verifying, extracting, shaping

  2. Running the model

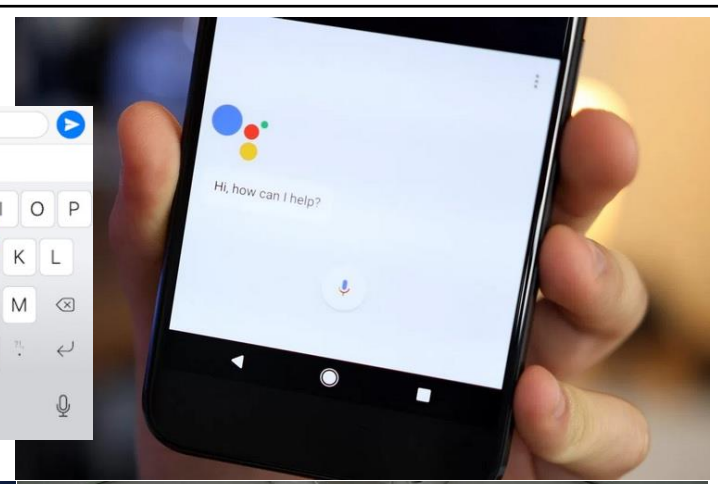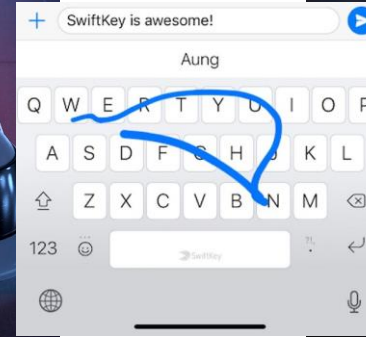 3. Communicating the results
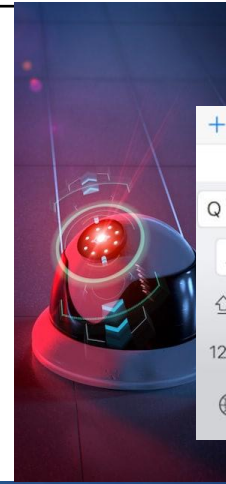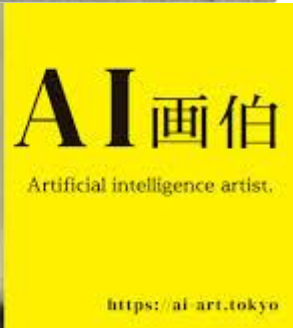
Data Science Success

Data Science Success

"We really designed the Model S to be a very sophisticated computer on wheels. **Tesla is a software company as much as it is a hardware company**. A huge part of what Tesla is, is a Silicon Valley software company. We view this the same as updating your phone or your laptop."

"Full autonomy is really a software limitation: **The hardware exists to create full autonomy, so it's really about developing advanced, narrow AI for the car to operate on**"   **Elon Musk**
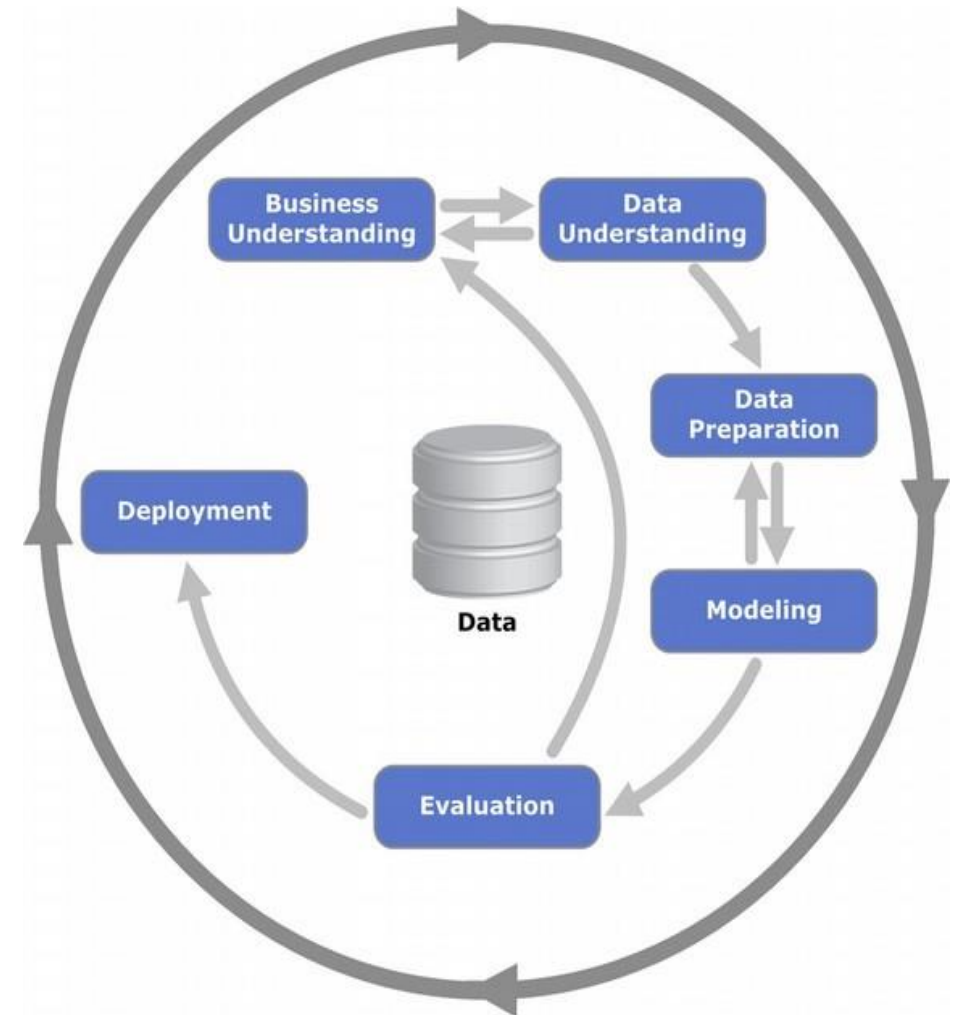
Data Science Success

Data Science Principals

1. Data Science is a process

2. ML is optimization of loss functions

3. ML must generalize to unseen data

4. Evaluate data science in its operational context

5. Similar entities can have similar unseen attributes

6. Correlation, not causation

P1: Data Science is a process

- **Cross-industry standard process for data mining**, known as **CRISP-DM:**
  - An open standard process model that describes common approaches used by data mining experts.

- It is the most widely-used analytics model

P1: A simple workflow from a technical PoV

Plan

Start

What is the question?

What type of data is needed?

Data Acquisition

Clean Data

Data Quality Analysis

Reformating & Imputing Data

Scripts

Explore the Data

Feature Engineering

Scripts

Feature Selection

Model Selection

Results Evaluation

Scripts

Deployment

Maintenance

Optimization

Data Analysis

Modeling

Deployment and optimization

13
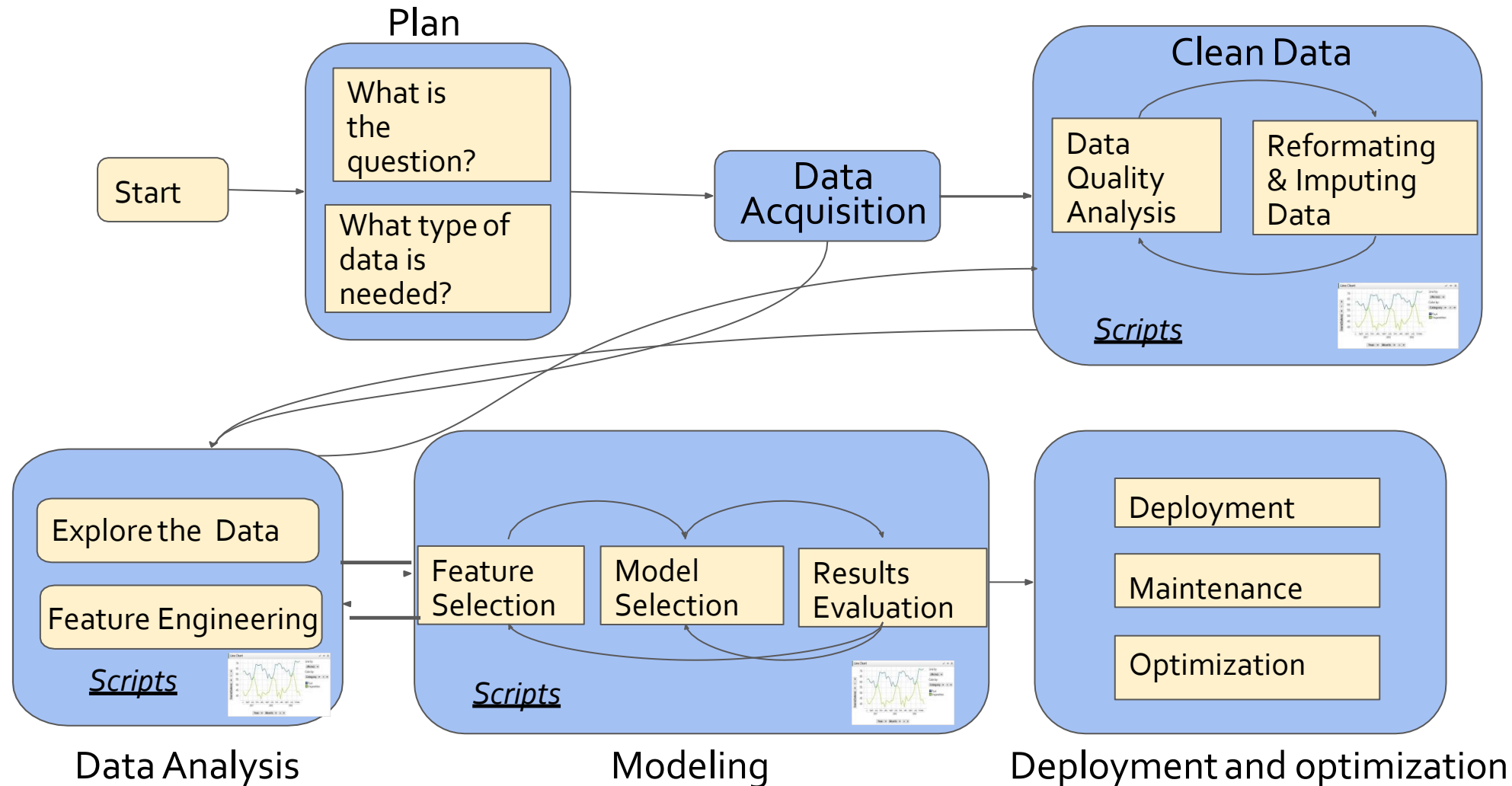
P1: A simple workflow

Business understanding
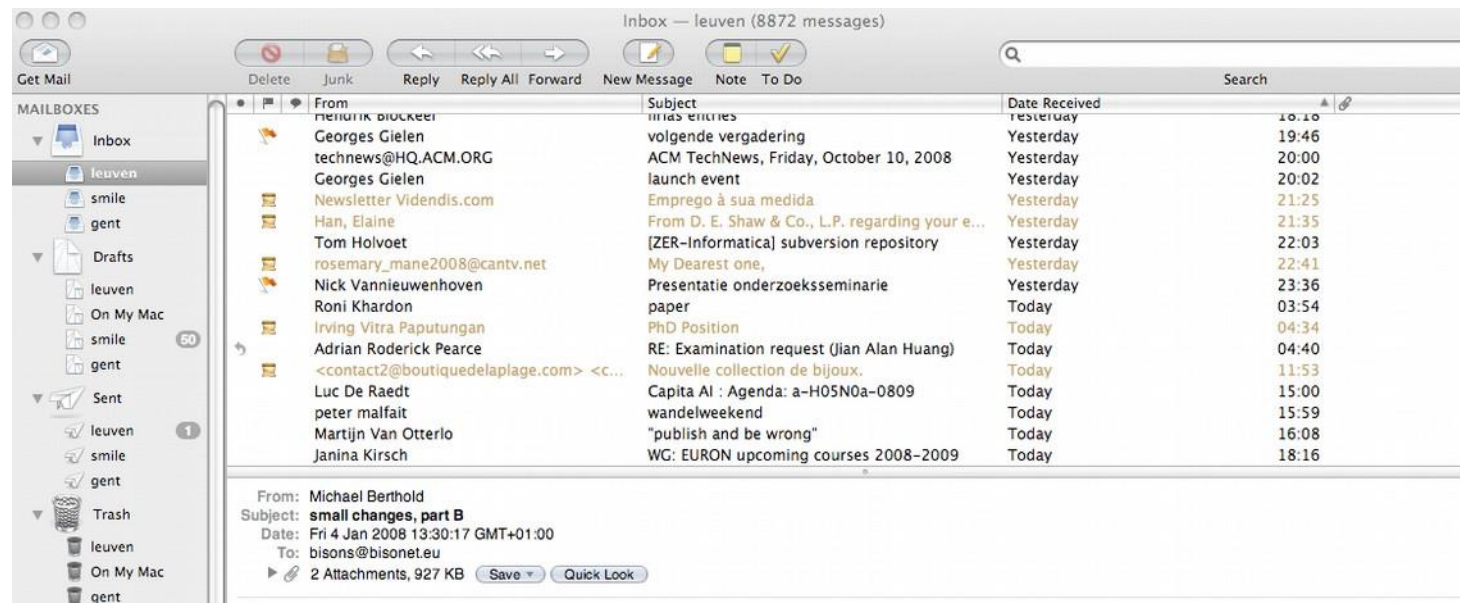
SPAM email reduces productivity, automatically remove it

P1: A simple workflow

Data understanding

- Collect messages, in general and from the user, that are spam (negative) and legitimate (positive): acquisition, annotation, definition of the targer, …

- Given a text message, predict whether it is spam or not

→ text categorization, useful in general

→ we want a function from message to {0,1}

→ is called binary classification problem

P1: A simple workflow

Data preparation

Given a raw text, convert string data into numerical data one

- Bag of words, TFIDF, Word2Vec, Embedding

Text Preprocessing

1. Remove Noisy Data: header, footer, HTML, XML, markup data

2. Tokenization: word, character, and subword (n-gram characters)

3. Normalization: converting all words to lowercases, …

P1: A simple workflow

Modeling

- We could write a rule-based system, such as

  if Title.contains("YOU HAVE WON!!!") then return Spam

- Train a classifier (e.g. naïve bayes, tree-based)

- Does it work well? $\rightarrow$ evaluate

P1: A simple workflow

Evaluation

On unseen emails

		Truth	
		Spam	Legitimate
Predicted as	Spam	150	30 — False positives
	Legitimate	200	720

False negatives

What does mean positive for you (to be spam or to be legitimate)?

P2: Machine learning is optimization

- Data vectors $\mathbf{x} \in \mathbb{R}^d$
  (e.g. for 512×512 images $d \approx 10^5$)

- Unknown classification functional
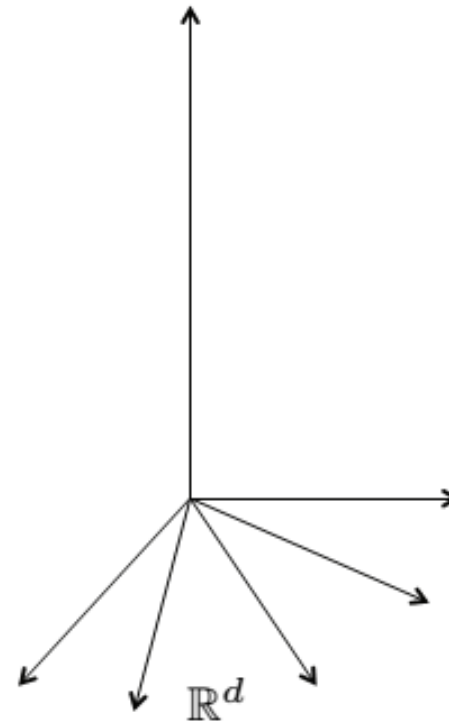 $f : \mathbb{R}^d \to \{1, \ldots, L\}$ in L classes

- Training set
  $$S = \{(\mathbf{x}_i \in \mathbb{R}^d, \, y_i = f(\mathbf{x}_i))\}_{i=1}^T$$
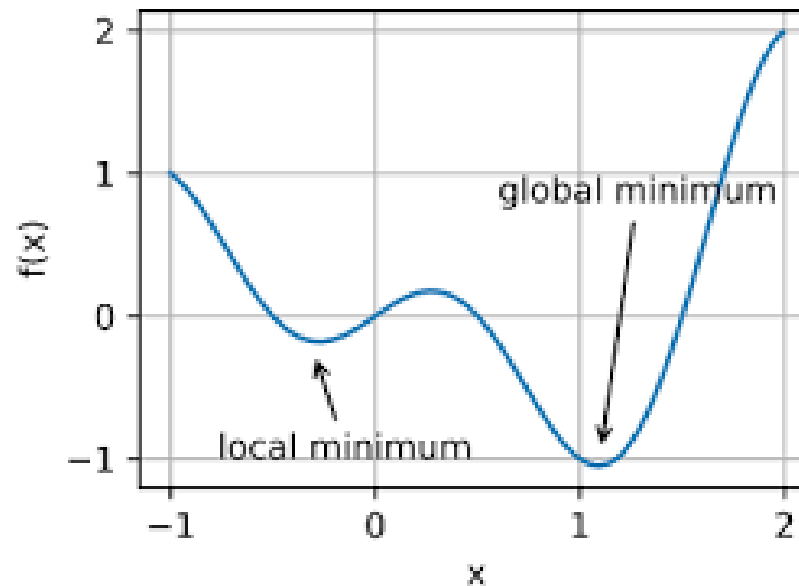
- Parametric model $f_{\boldsymbol{\Theta}}$ of f

Supervised learning: find optimal model parameters by minimizing the loss ℓ on the training set

$$\boldsymbol{\Theta}^* = \operatorname*{argmin}_{\boldsymbol{\Theta}} \sum_{i=1}^T \ell(f_{\boldsymbol{\Theta}}(\mathbf{x}_i), y_i)$$
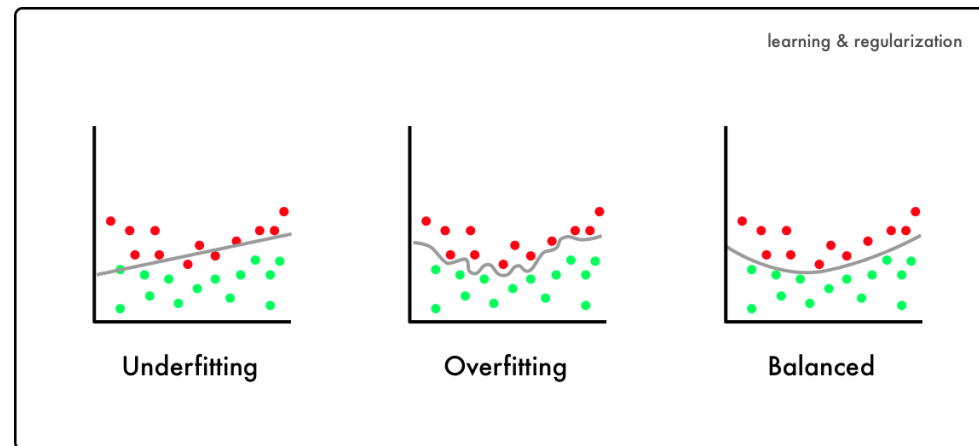
P2: Machine learning is optimization

- Problem: Optimal solutions

    - When using AI heuristics to find some optimum, you may end up in a local minima

P3: Machine learning is generalization

- Generalization
    - If you look too hard at a dataset, you will find something, but it might not generalize beyond the data you're looking at (unseen data) = Overfitting

learning & regularization

Underfitting   Overfitting   Balanced

P3: Machine learning is generalization

- The impact of missing data on quantitative research can be serious, leading to biased estimates of parameters, loss of information, decreased statistical power, increased standard errors, and weakened generalizability of findings.
  - multiple imputation, maximum likelihood, and expectation-maximization algorithm
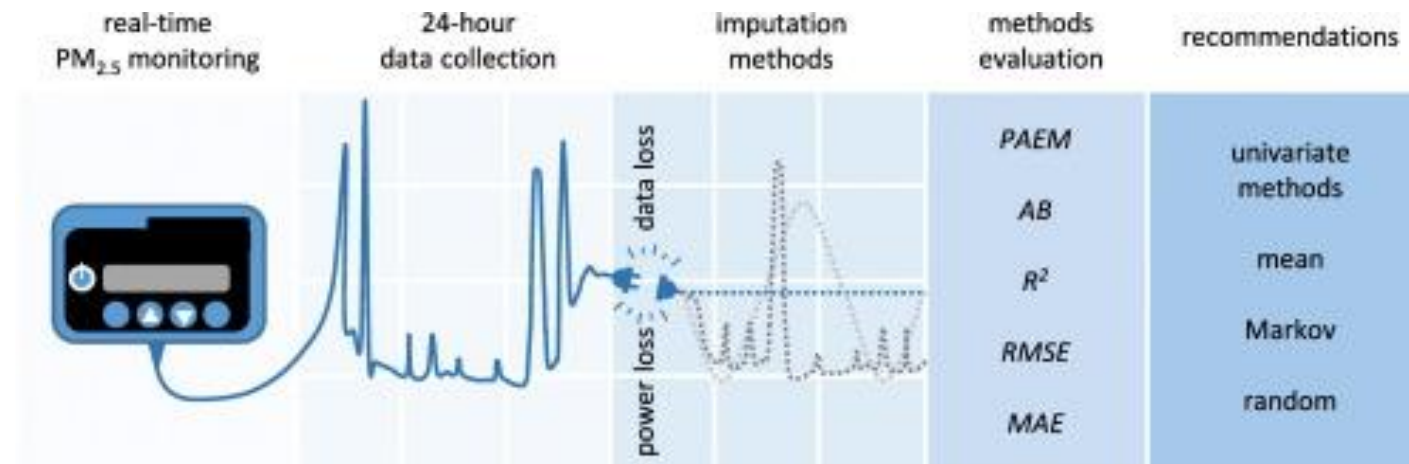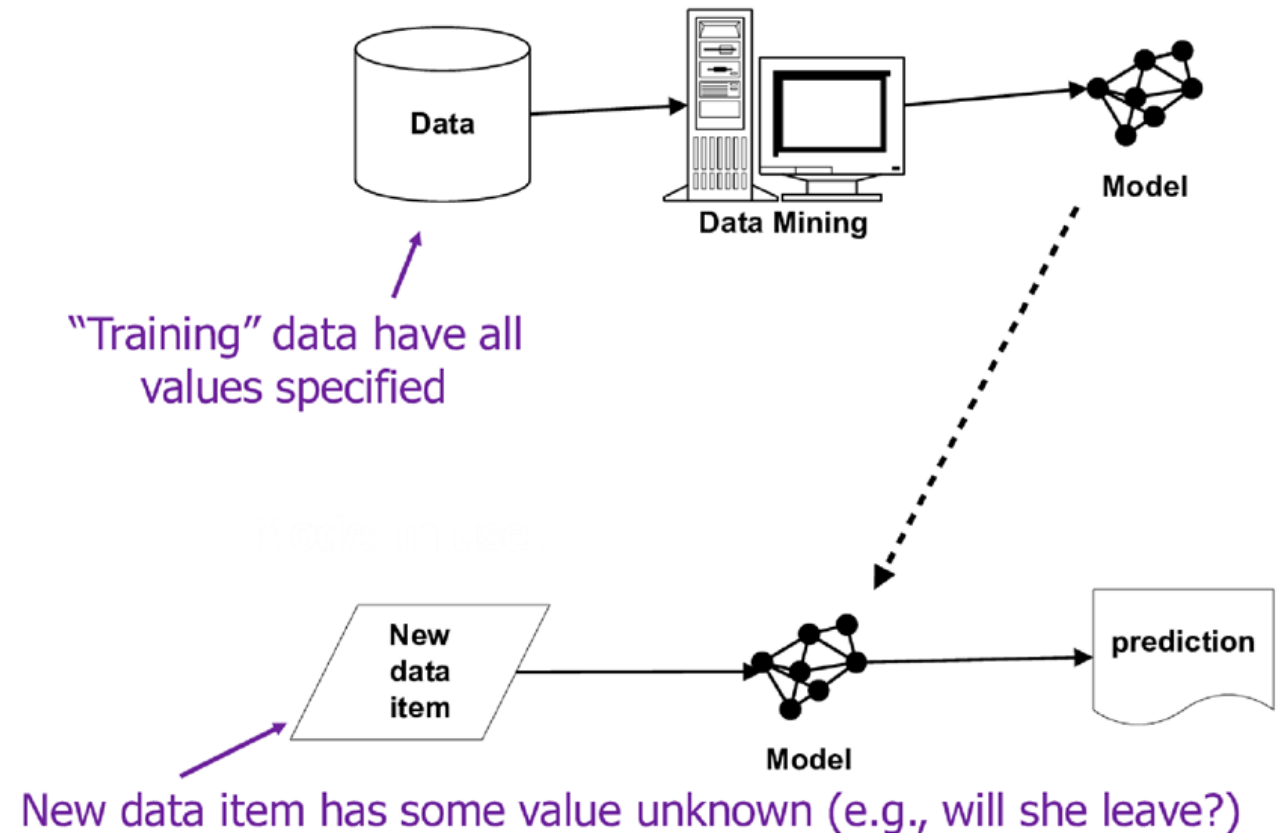
P4: Data science needs to be evaluated in the context of operation

- Training data is not consistent with actual use:

  1. Bad samples

  2. Bad features

"Supervised" modeling:

Data

Data Mining

Model

"Training" data have all values specified

New data item

Model

prediction

New data item has some value unknown (e.g., will she leave?)

23

P5: Entities that are similar on some attributes often are similar on unseen attributes

- Clustering and optimization (e.g. min. distances to cluster center)

- Key concept: <u>distance</u> between objects
  - Euclidean, Manhattan, edit distances (strings), Dynamic time warping (temporal sequences), …
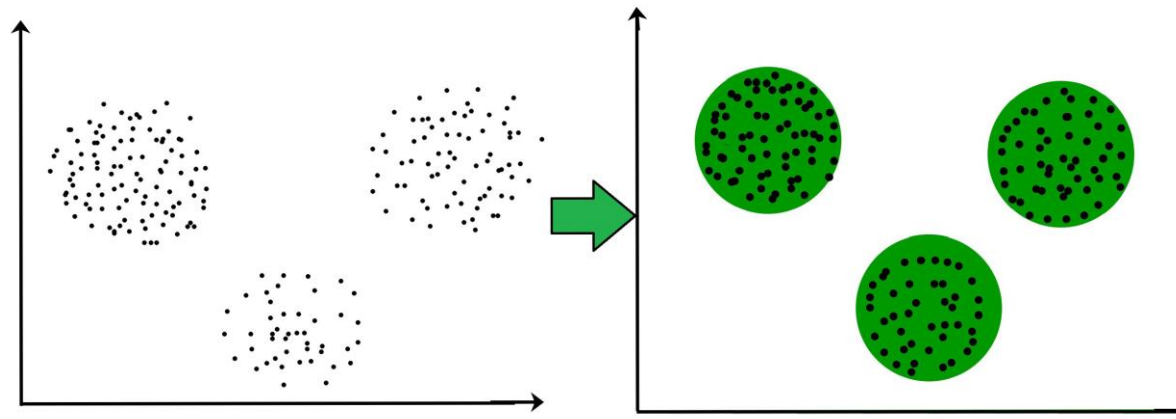
P6: Correlation

- To draw <u>causal</u> conclusions, one must pay very close attention to the presence of (possibly unseen) <u>confounding factors</u>

- Machine models exploit correlation, NOT causality
  - Very tempting to inspect model and see "what causes things to be true/false"
  - E.g. coefficients of linear regression
    - $Y = 20*X_1 - 12*X_2 + 300*X_3 + 99*X_4 - 299*X_5$
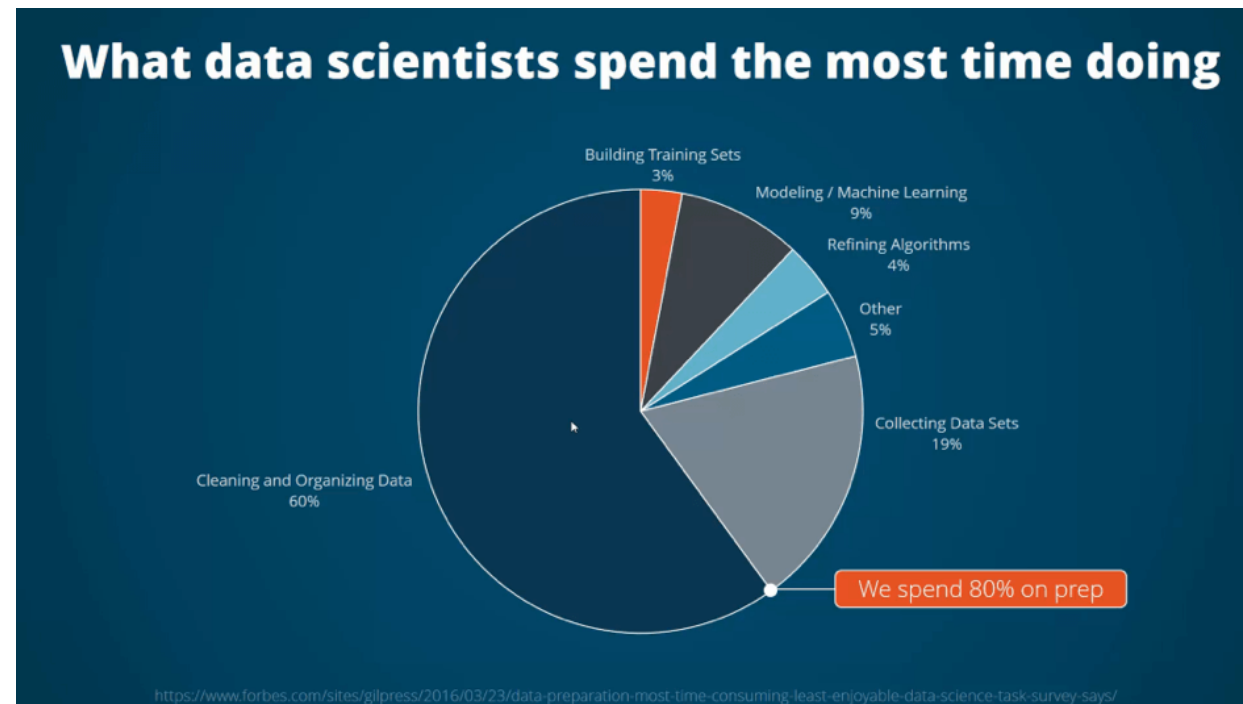   - Which feature has most impact?
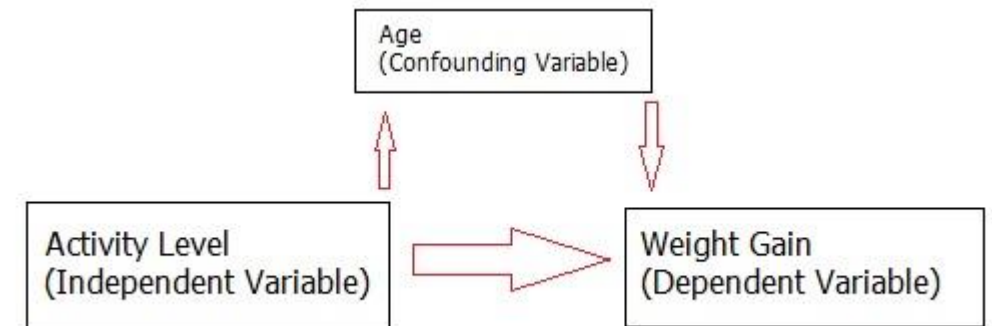
Big Problem: Curse of dimensionality

- **Data sparcity:** we need more traning samples

- **Distance concentration:** proximity or similarity of the samples may not be qualitatively relevant

a) 1D - 4 regions b) 2D - 16 regions c) 3D - 64 regions
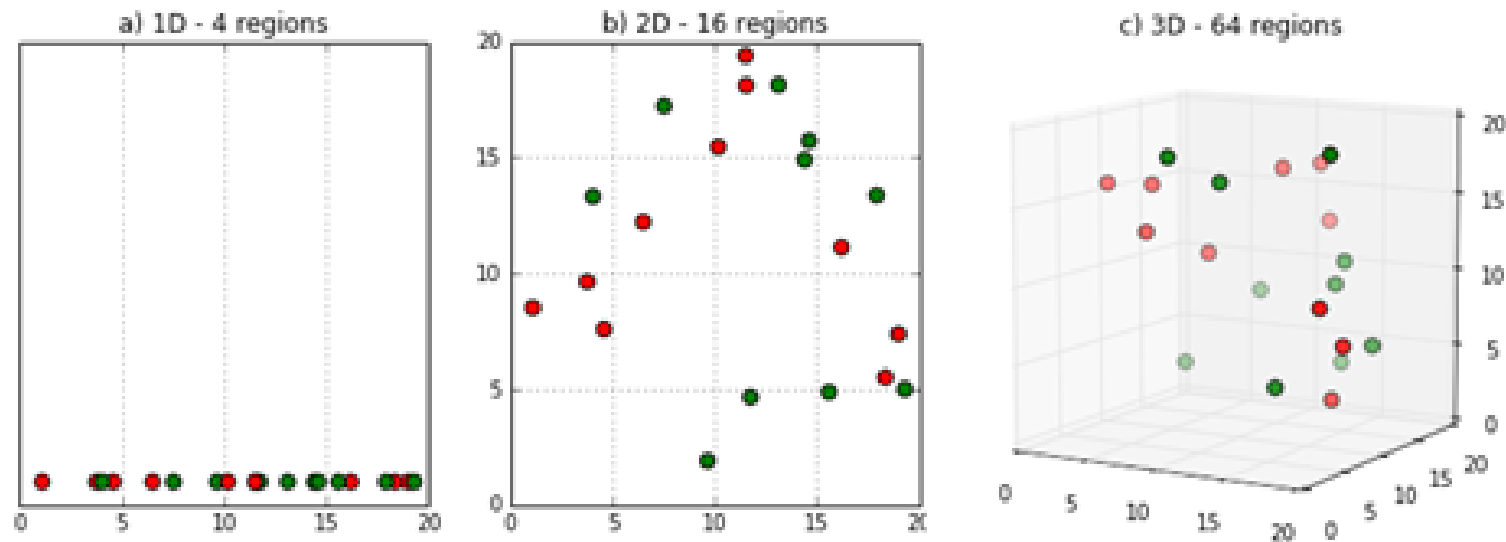
What do data scientists spend the most time doing?

80% of the work as machine learning experts and data scientists is preparing the data.

What data scientists spend the most time doing

- Building Training Sets 3%
- Modeling / Machine Learning 9%
- Refining Algorithms 4%
- Other 5%
- Collecting Data Sets 19%
- Cleaning and Organizing Data 60%

We spend 80% on prep

https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

What's the least enjoyable part of data science?

57% said it's cleaning and organizing data : we spend 80% of our time there, and we don't even enjoy it that much. I mean, that's horrible, but it's reality.

What's the <u>least</u> enjoyable part of data science?

Building Training Sets	Modeling / Machine Learning	Refining Algorithms	Other	Collecting Data Sets	Cleaning and Organizing Data
10%	3%	4%	5%	21%	57%

https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/
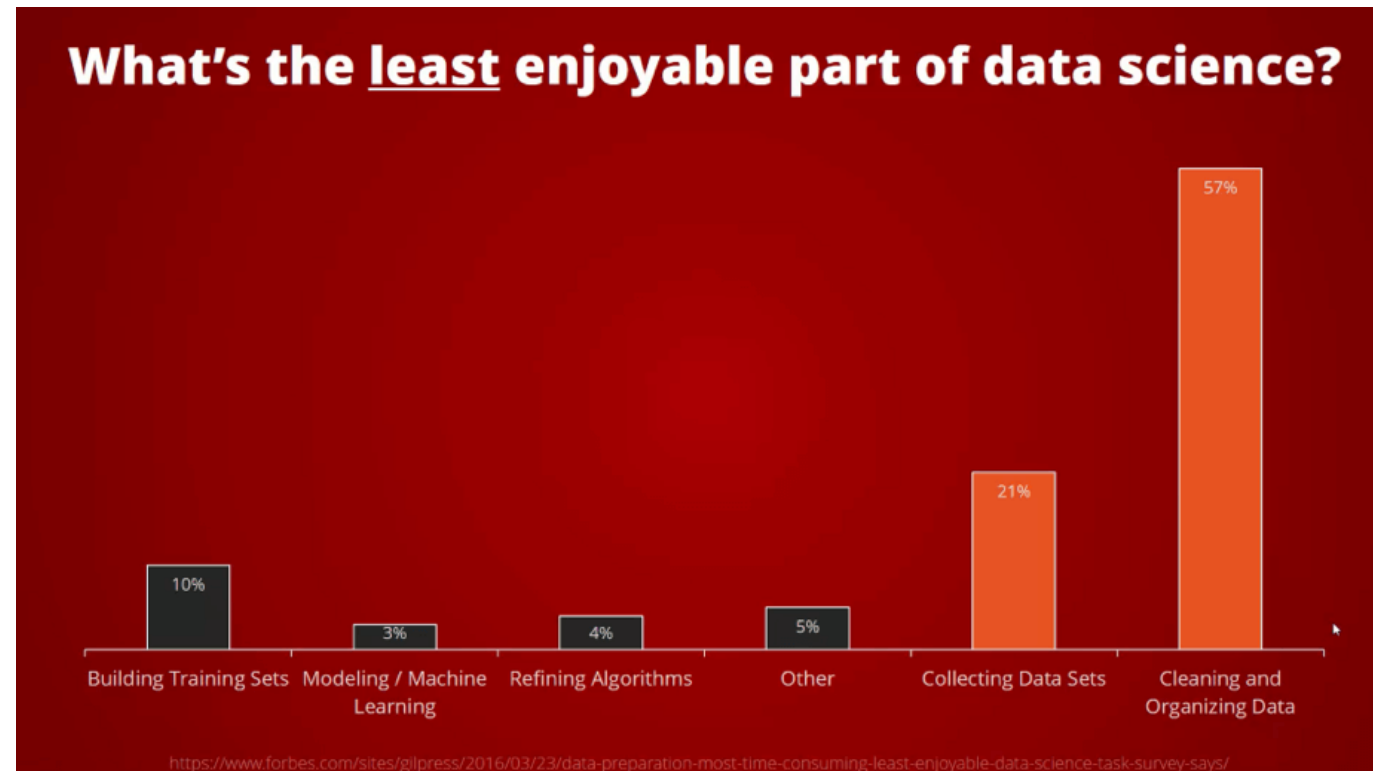
Why data preparation is so important?

- Data preparation is a multi-step process that involves data collection, cleaning & preprocessing, feature engineering, and labeling.

- These steps play an important role in the overall quality of your machine learning model, as they build on each other to ensure a model performs to expectations.

What makes data preparation so difficult?

1. Code-based approach to data science: Python, ..
2. Process-based approach to data science: orange, rapidminer
3. Data-centric approach to data science: Excel
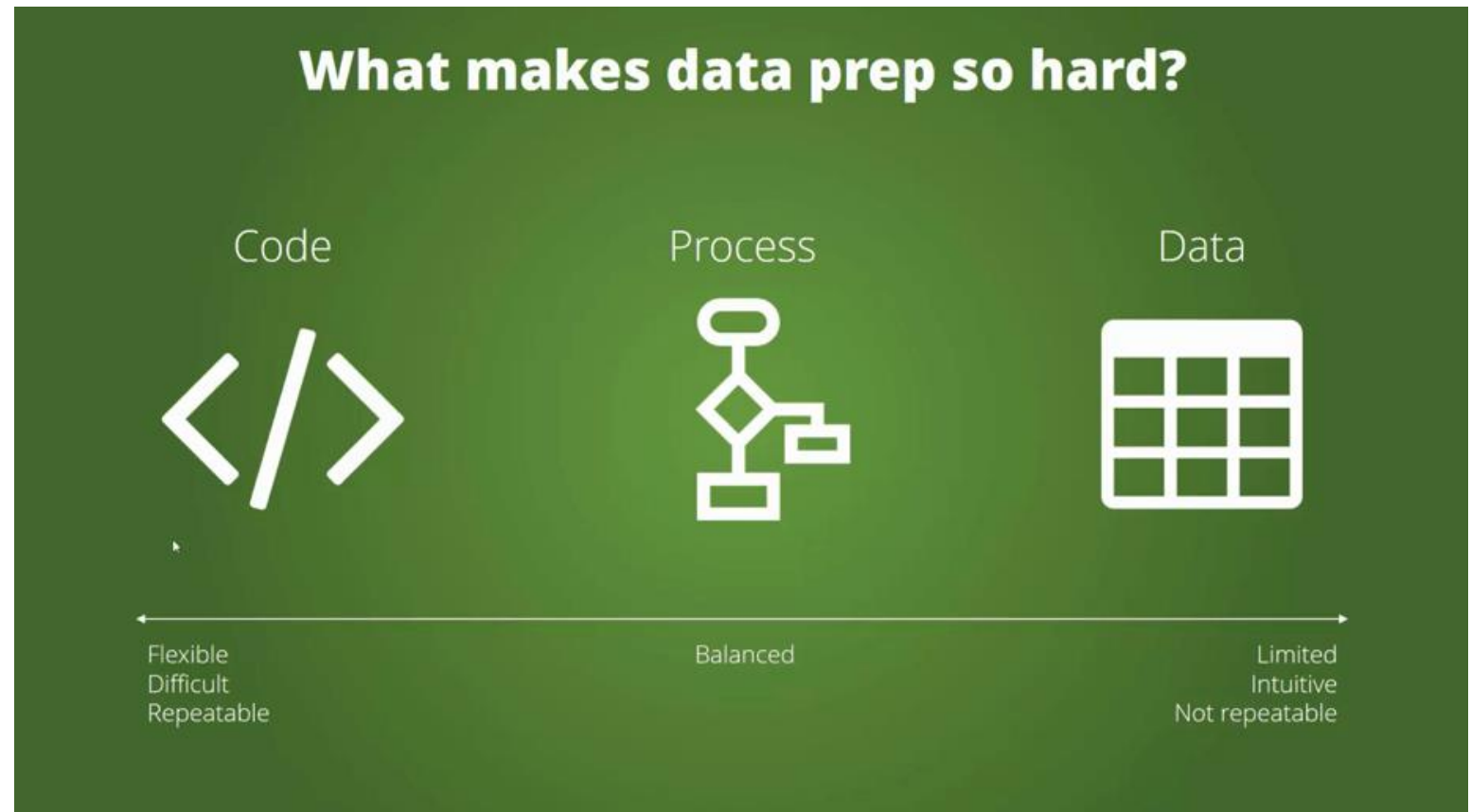
The path to be a data scientist
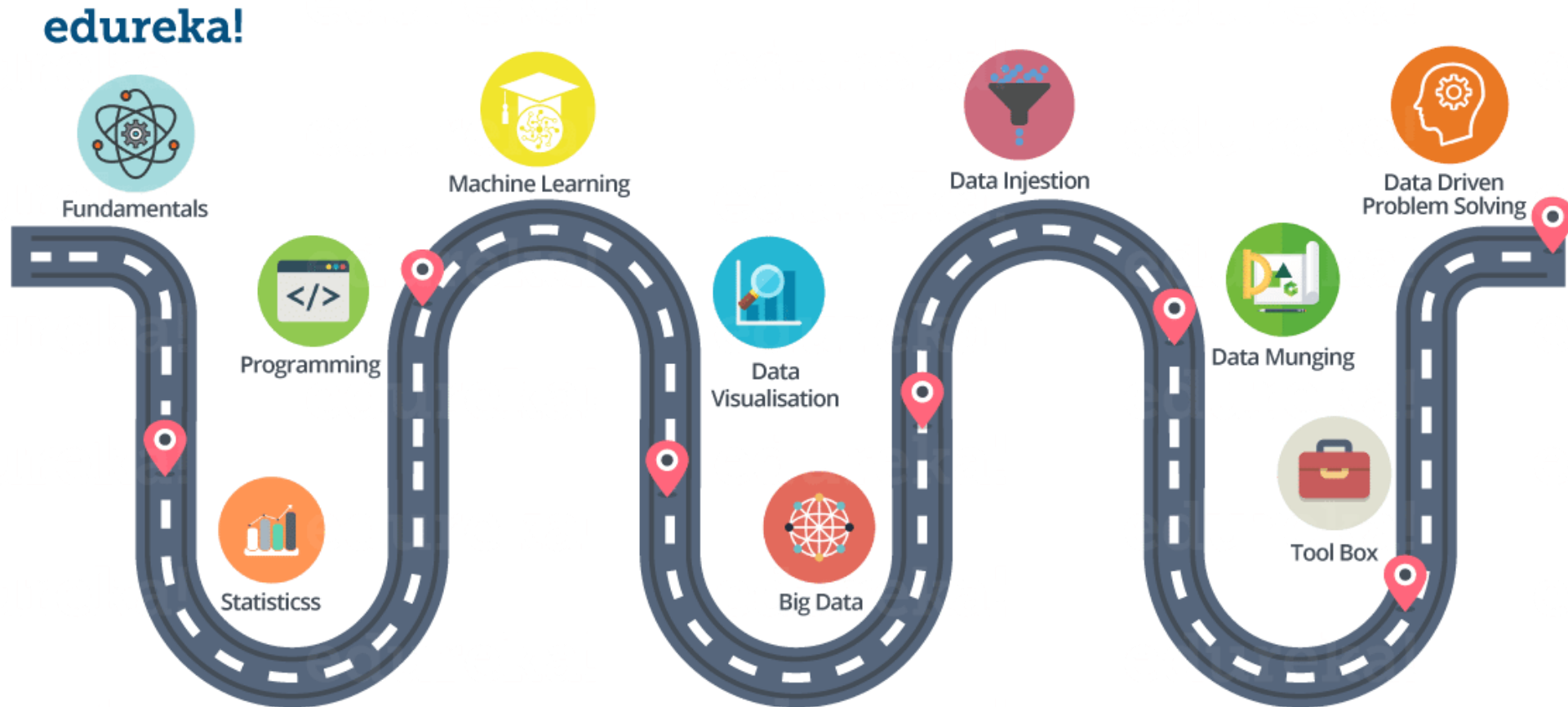
How said AutoML and AutoDS?

- Replacing Data Scientists?
- AI Literacy ?

Literacy is the ability to read, write, speak and listen in a way that lets us communicate effectively and make sense of the world.

How said AutoML and AutoDS?

- AI Literacy ?

Photo by Max Duzij on Unsplash

- Only highly educated people can program new AI applications
- Power only with the large IT companies

- In an age of limited resources, the need for efficient use gets more important

- **AutoML contributes to AI literacy!**

Marius Lindauer