

# Data Science and Machine Learning

---

Introduction

# Data is: Big!

**Lots of Data** => Lots of Analysis => Lots of Jobs

- 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 :)

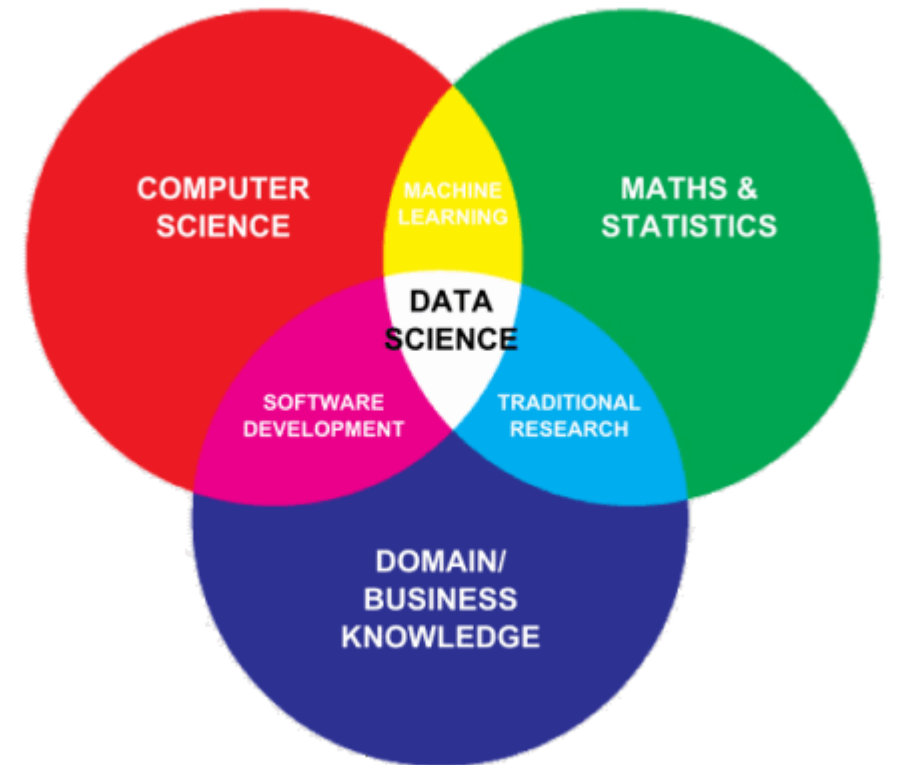
- Everyone wants better predictability, forecasting, customer satisfaction, market differentiation, prevention, great user experience, ...
  - How can I price a particular product?
  - What can I recommend online customers to buy after buying X, Y or Z?
  - How can we discover market segments? group 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**

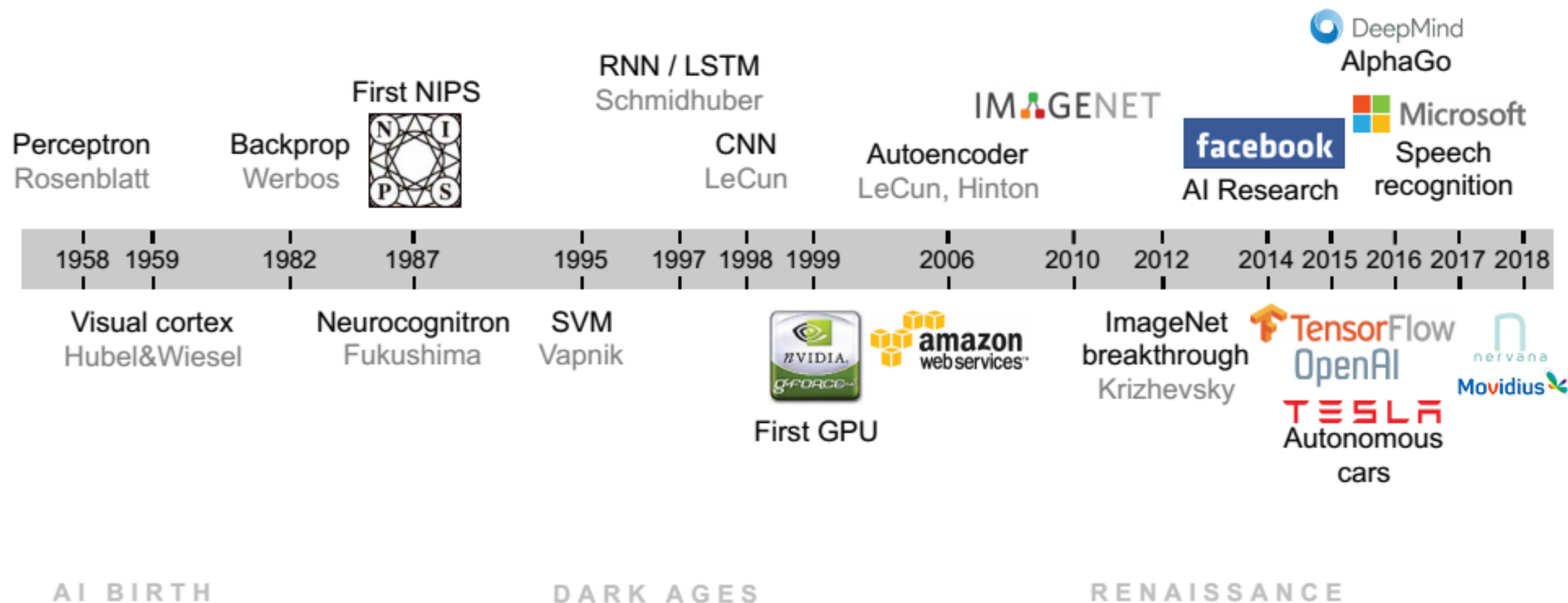
- Statisticians
- Mathematicians
- Computer Scientists in
  - Data mining
  - Artificial Intelligence & Machine Learning
  - Systems Development and Integration
  - Database development
  - Analytics
- Domain Experts
  - Medical experts
  - Geneticists
  - Finance, Business, Economy experts
  - etc.



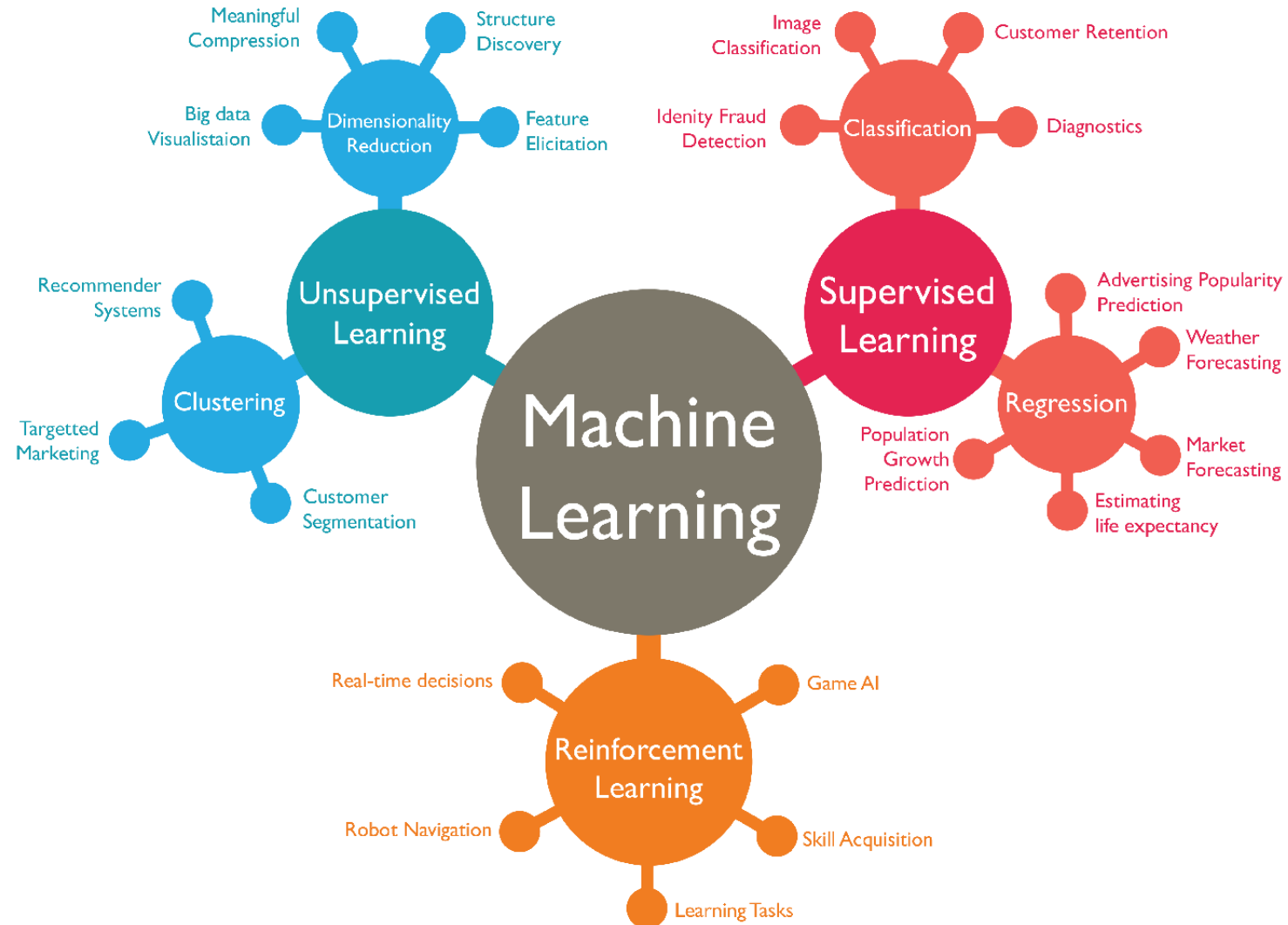
# Data Science is: about the whole processing pipeline to extract information out of data

- Data Scientist understand and care about the whole data pipeline:
  - A data pipeline consists of 3 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

# AI



# AI zoo





# Data Science Success



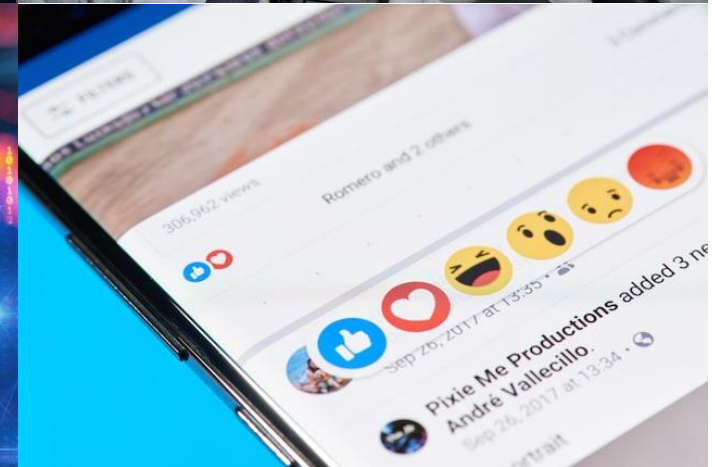
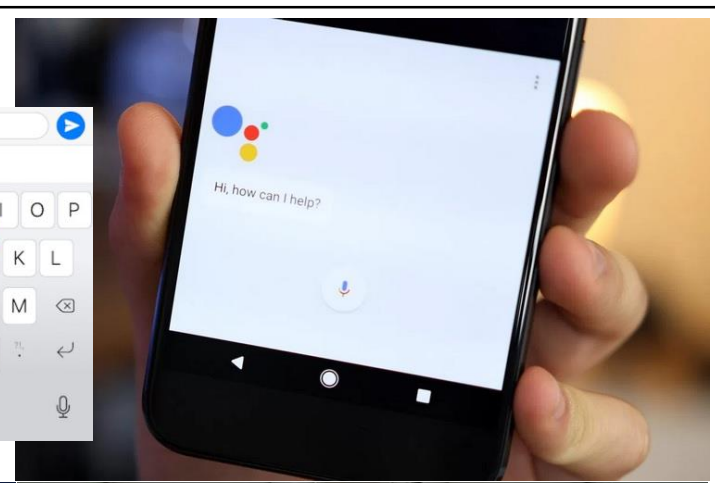
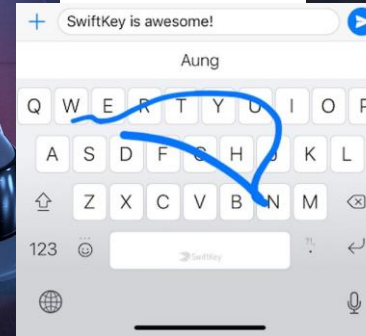
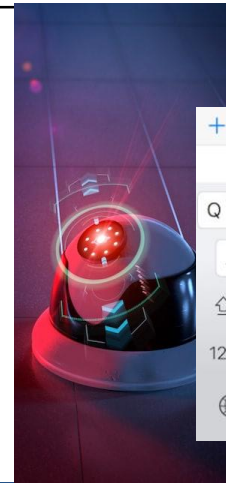
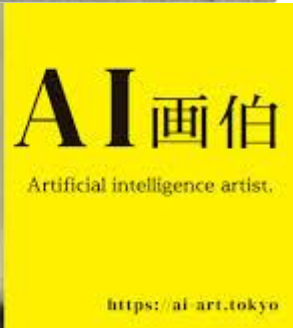
"man in black shirt  
guitar."



"s are playing with  
toy."



"young girl in pink shirt  
swinging on swing."





# Data Science Success



**Mercedes-Benz Autonomous  
Car Interior Concept (Luxury F  
015)**



# Data Science Success

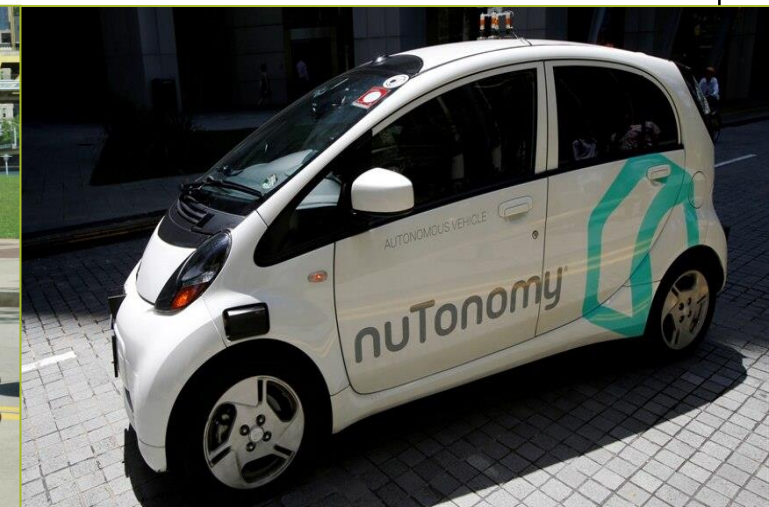


“We really designed the Model S to be a very sophisticated computer on wheels. **Tesla is a software company as much as it is a hardware company.** A huge part of what Tesla is, is a Silicon Valley software company. We view this the same as updating your phone or your laptop.”

“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 Success

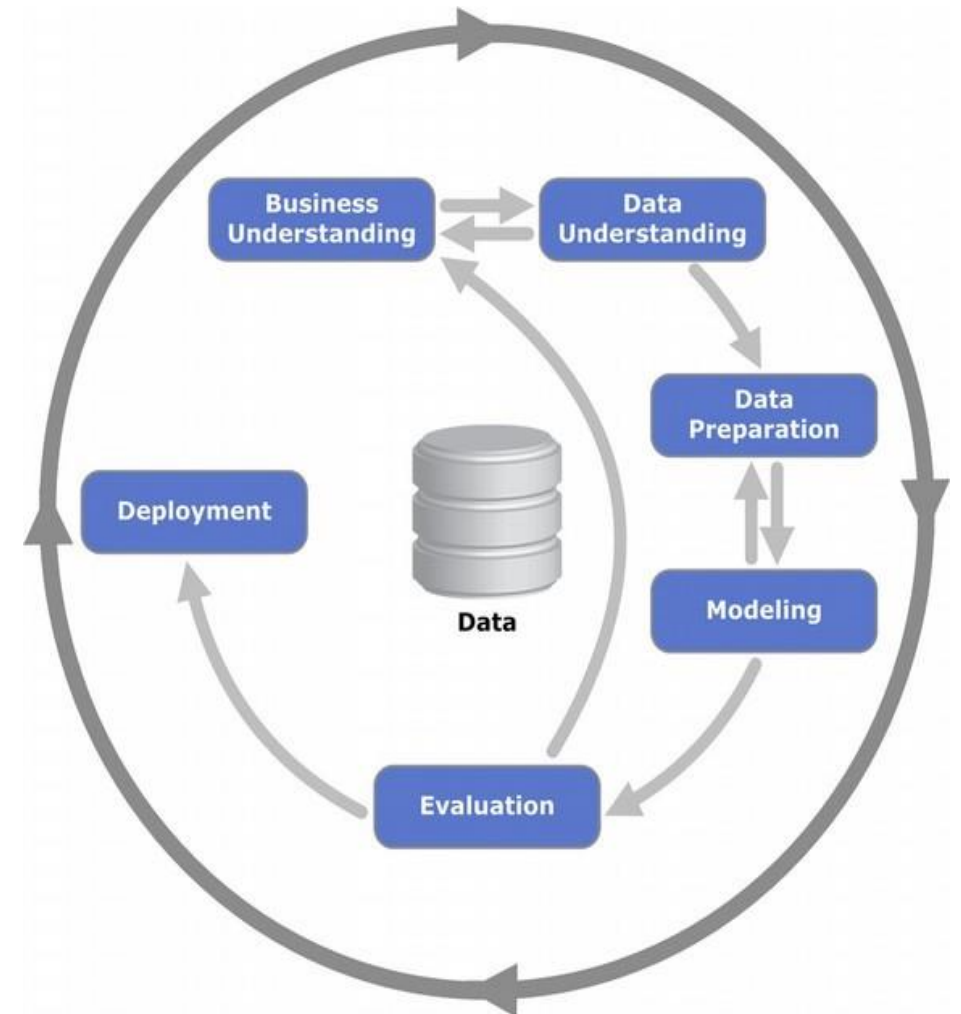


# Data Science Principals

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

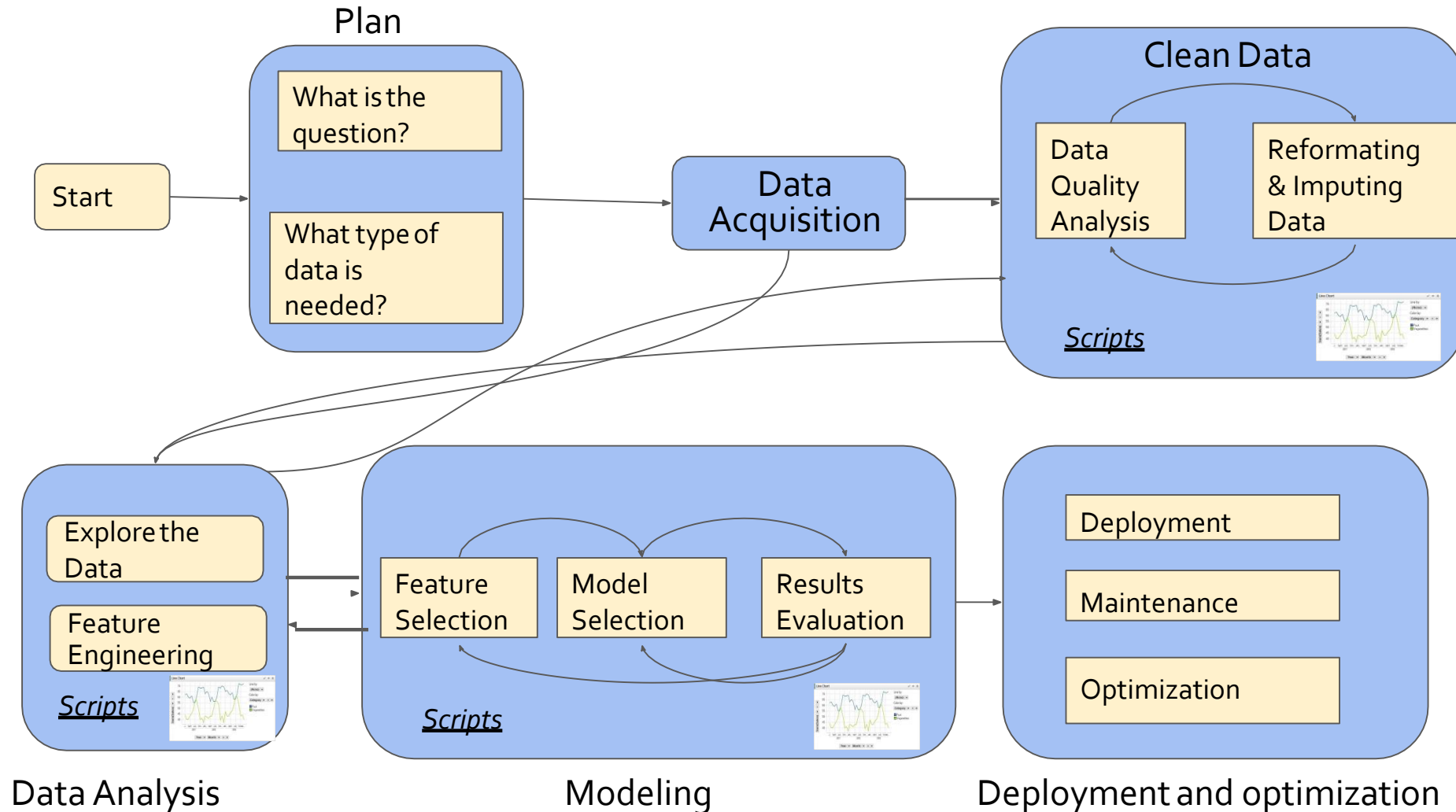
# P1: Data Science is a process

- **Cross-industry standard process for data mining**, known as **CRISP-DM**, is an open standard process model that describes common approaches used by data mining experts.
- It is the most widely-used analytics model



# P1: Data Science is a process

## A simple workflow from a technical PoV

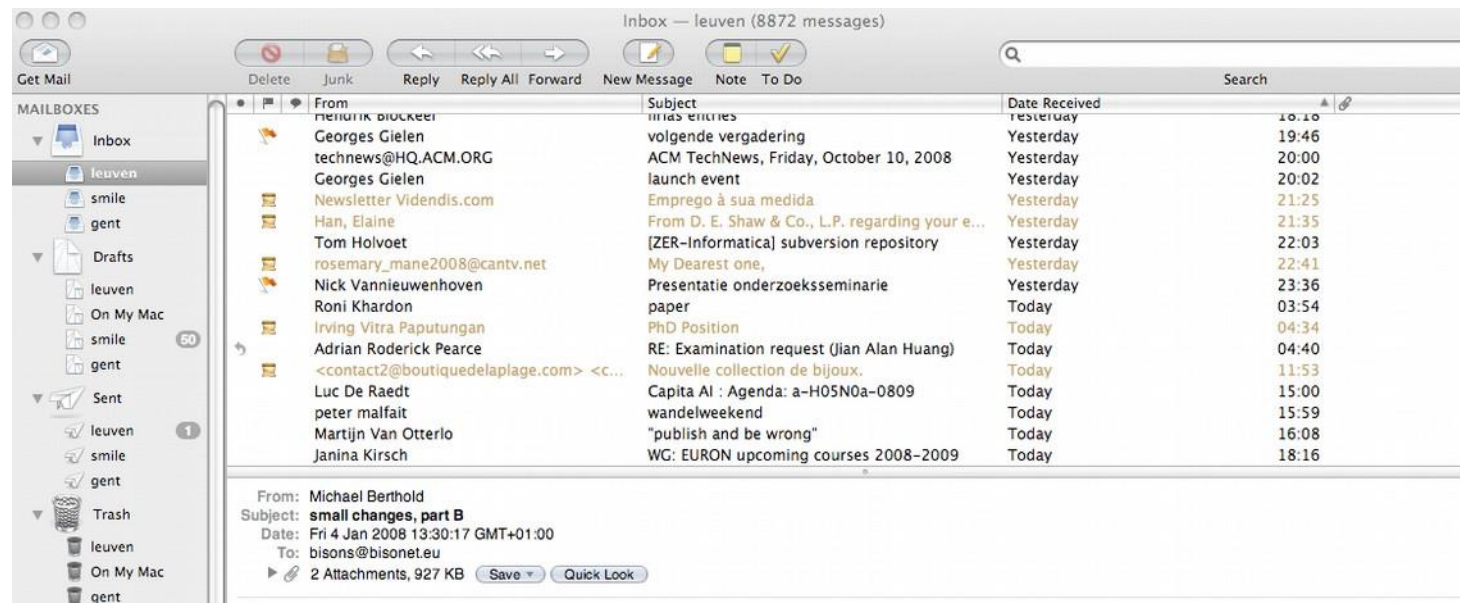




# P1: Data Science is a process

## A simple workflow

### Business understanding



SPAM email reduces productivity, automatically remove it

# P1: Data Science is a process

## A simple workflow

### Data understanding

- Collect messages, in general and from the user, that are spam (negative) and legitimate (positive): acquisition, annotation, ...
- 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: Data Science is a process

## A simple workflow

### Data preparation

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

- Bag of words, TFIDF, Word2Vec

### 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: Data Science is a process

## 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)
- Does it work well? → evaluate

# P1: Data Science is a process

## A simple workflow

### Evaluation

on unseen emails

|              |            | Truth               |                    |
|--------------|------------|---------------------|--------------------|
|              |            | Spam                | Legitimate         |
| Predicted as | Spam       | 150                 | 30 False positives |
|              | Legitimate | 200 False negatives | 720                |

# P2: Machine learning is optimization

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

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

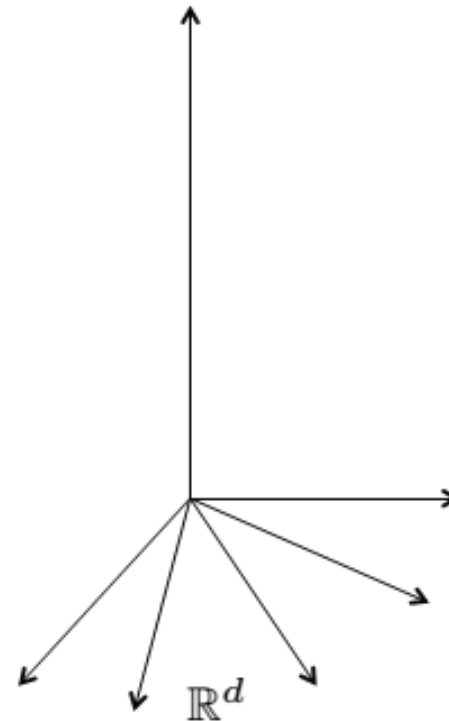
- Training set

$$S = \{(\mathbf{x}_i \in \mathbb{R}^d, y_i = f(\mathbf{x}_i))\}_{i=1}^T$$

- Parametric model  $f_{\Theta}$  of  $f$

**Supervised learning:** find optimal model parameters by minimizing the loss  $\ell$  on the training set

$$\Theta^* = \underset{\Theta}{\operatorname{argmin}} \sum_{i=1}^T \ell(f_{\Theta}(\mathbf{x}_i), y_i)$$

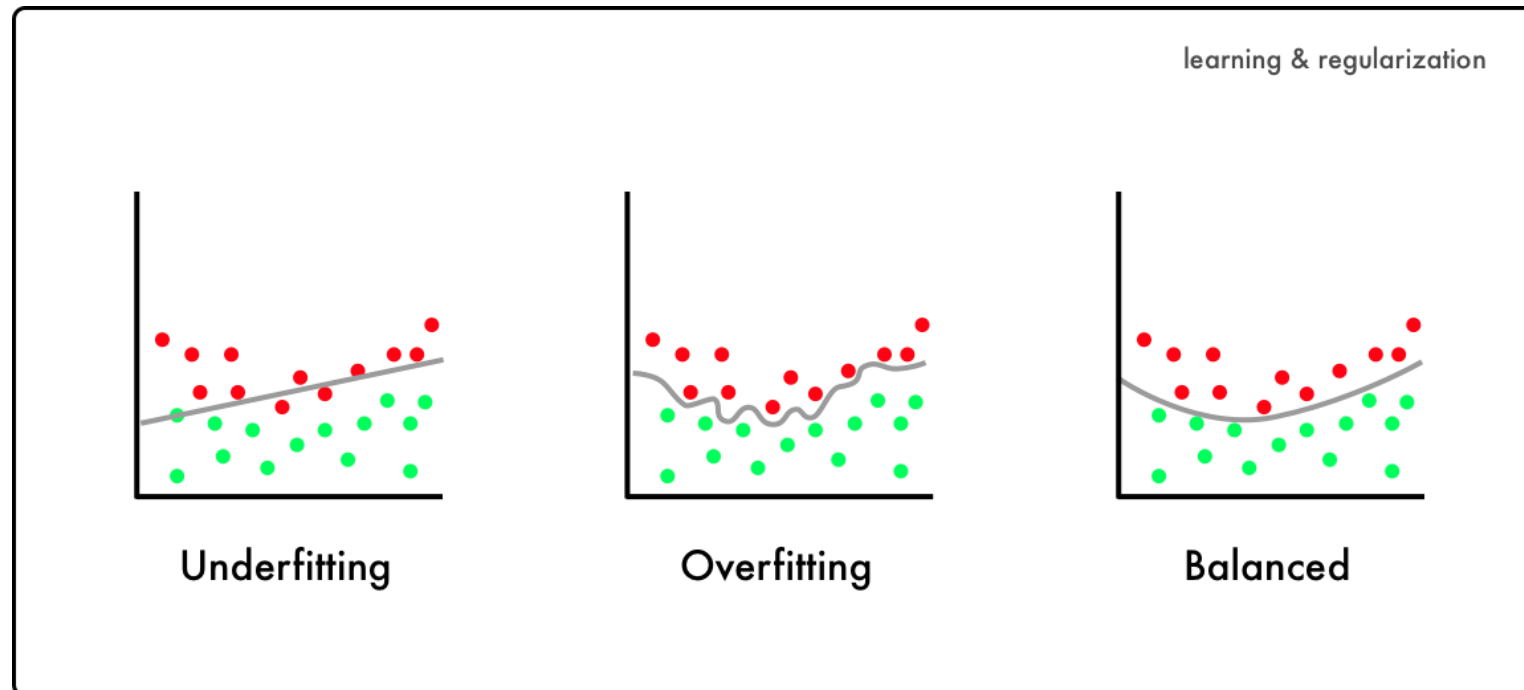


# P2: Machine learning is optimization

when using AI heuristics to find some optimum, you may end up in a local maximum

# P2: Generalizing

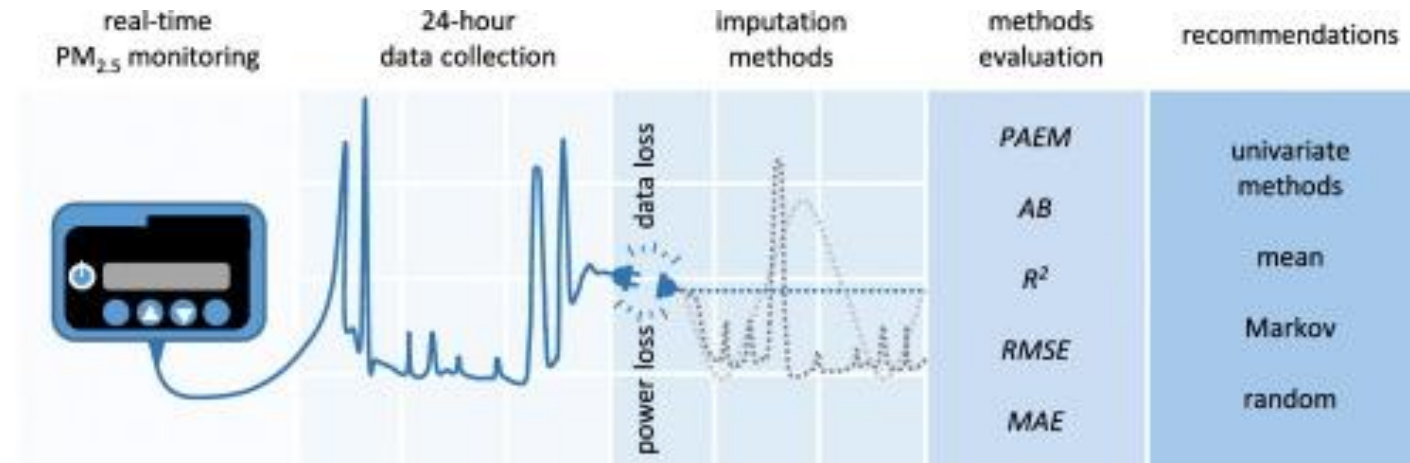
- If you look too hard at a dataset, you will find something, but it might not generalize beyond the data you're looking at (unseen data) = Overfitting





# P3: Missing information

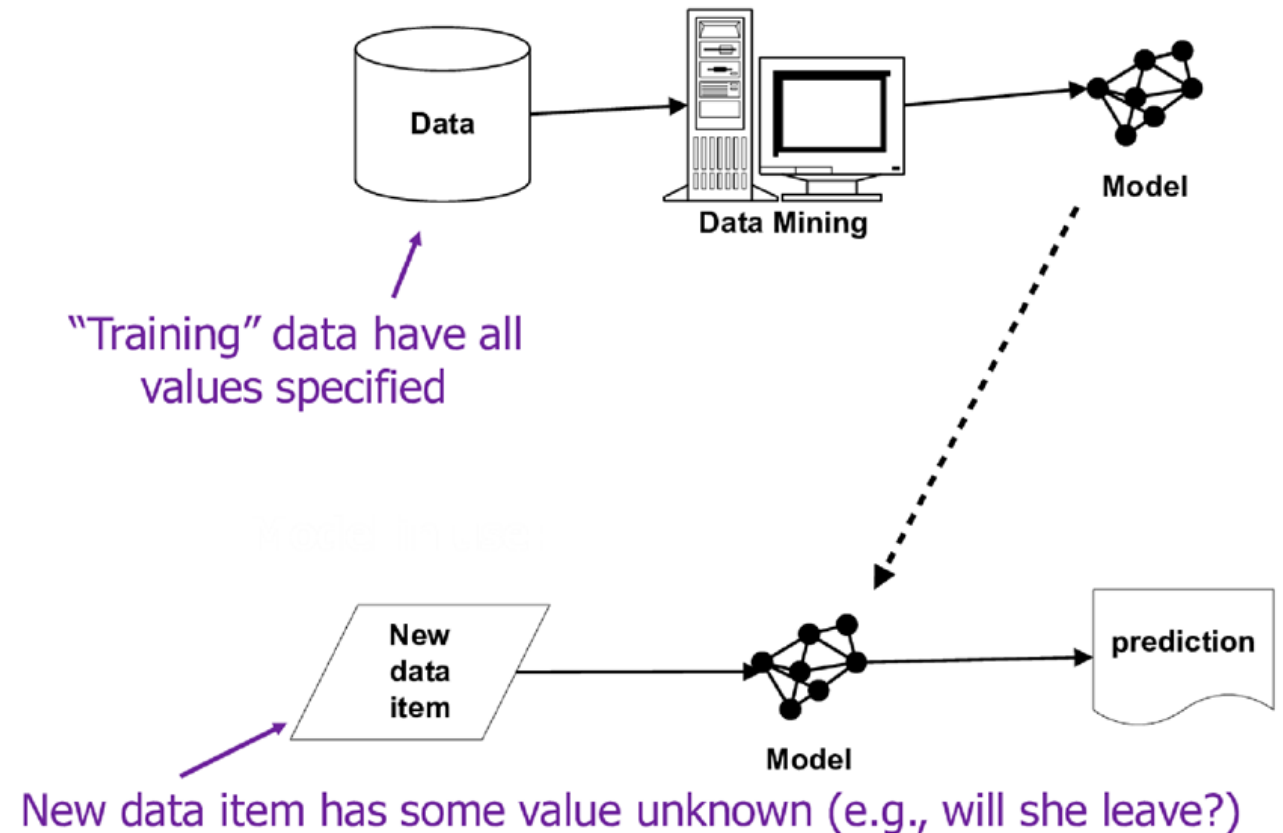
- The impact of missing data on quantitative research can be serious, leading to biased estimates of parameters, loss of information, decreased statistical power, increased standard errors, and weakened generalizability of findings.
  - multiple imputation, maximum likelihood, and expectation-maximization algorithm



# P<sub>4</sub>: Data science needs to be evaluated in the context of operation

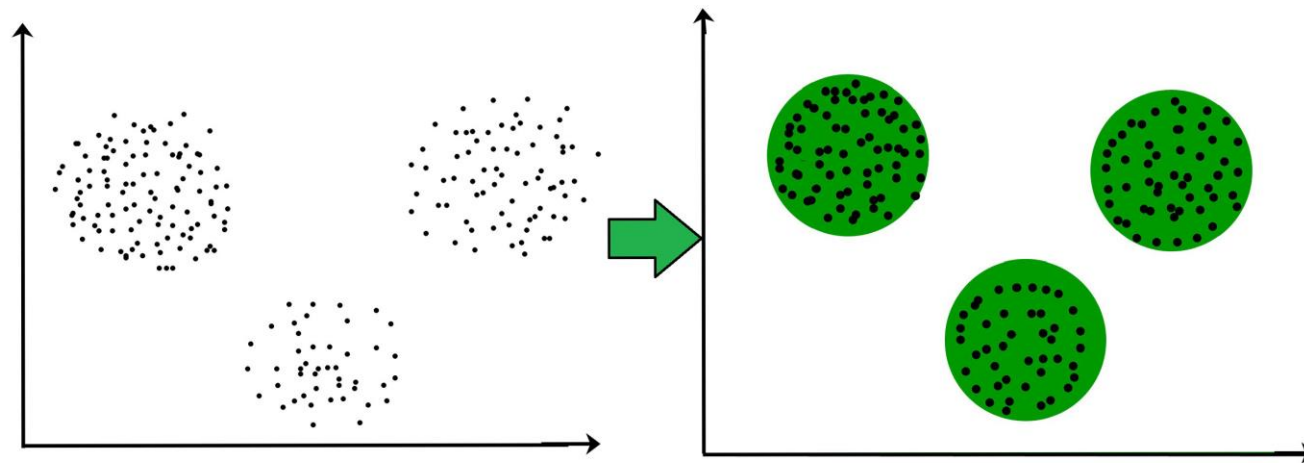
- Training data is not consistent with actual use
  1. Bad samples
  2. Bad features

## “Supervised” modeling:



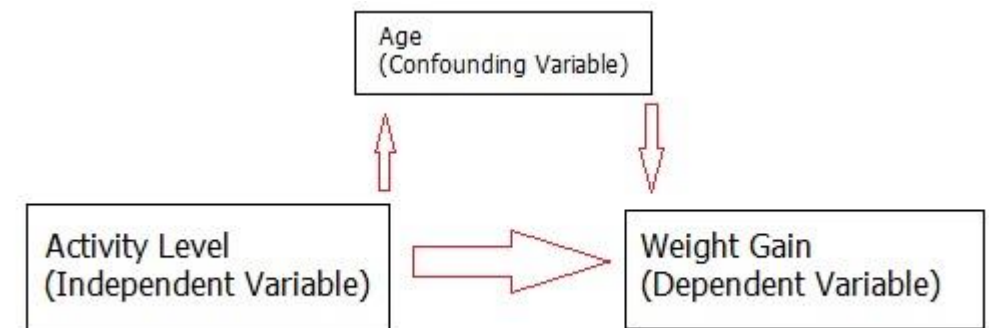
# P5: Entities that are similar on some attributes often are similar on unseen attributes (causality)

- Clustering
- Also optimization, e.g. min. distances to cluster center
- Key concept: distance between objects
  - Euclidean, Manhattan, edit distances (strings), Dynamic time warping (temporal sequences), ...



# P6: Correlation

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



# ML vs Stat

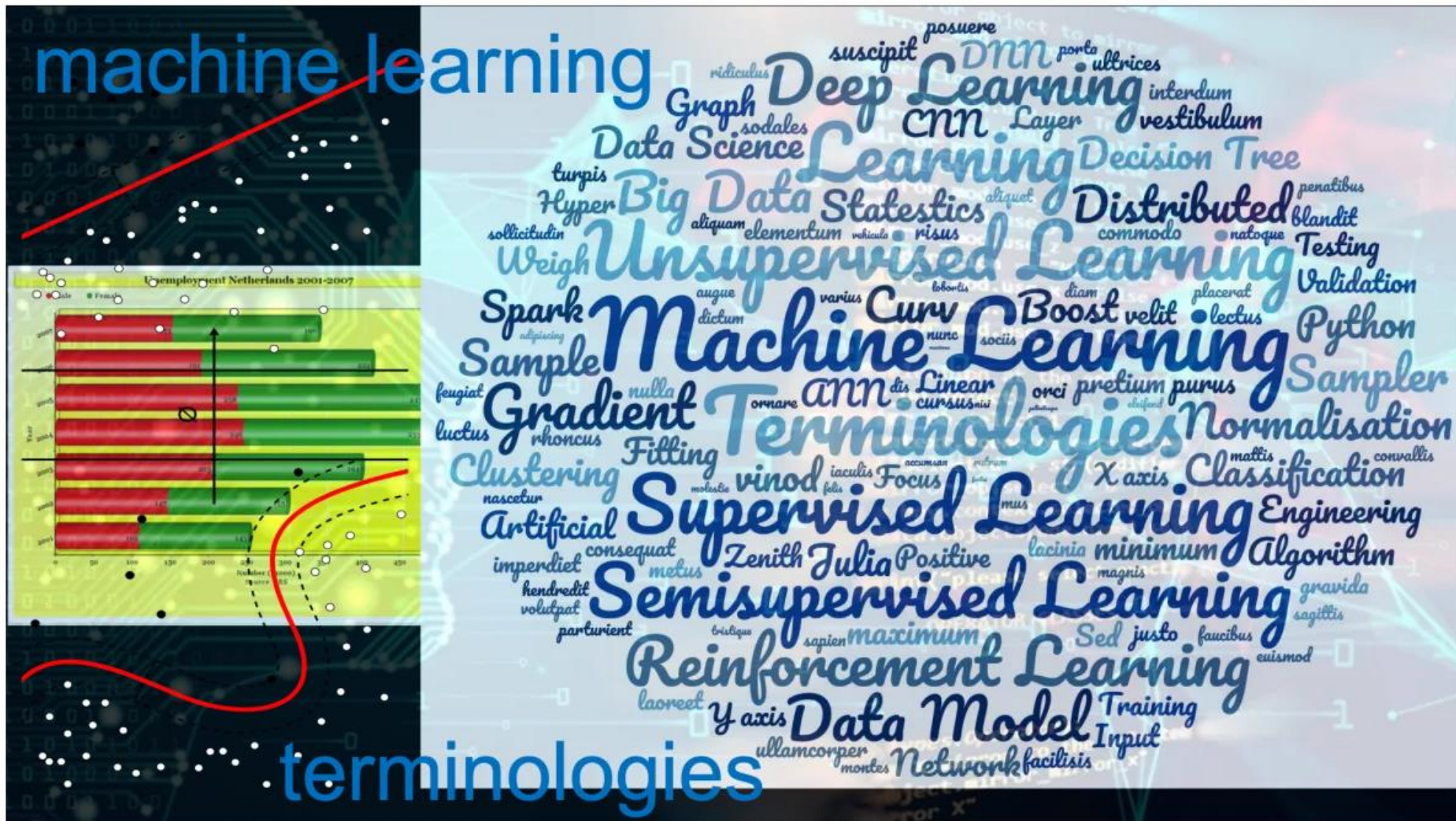
- In his course on statistics, Rob Tibshirani, a statistician who also has a foot in machine learning, provides a glossary that maps terms in statistics to terms in machine learning, reproduced below.

## Glossary

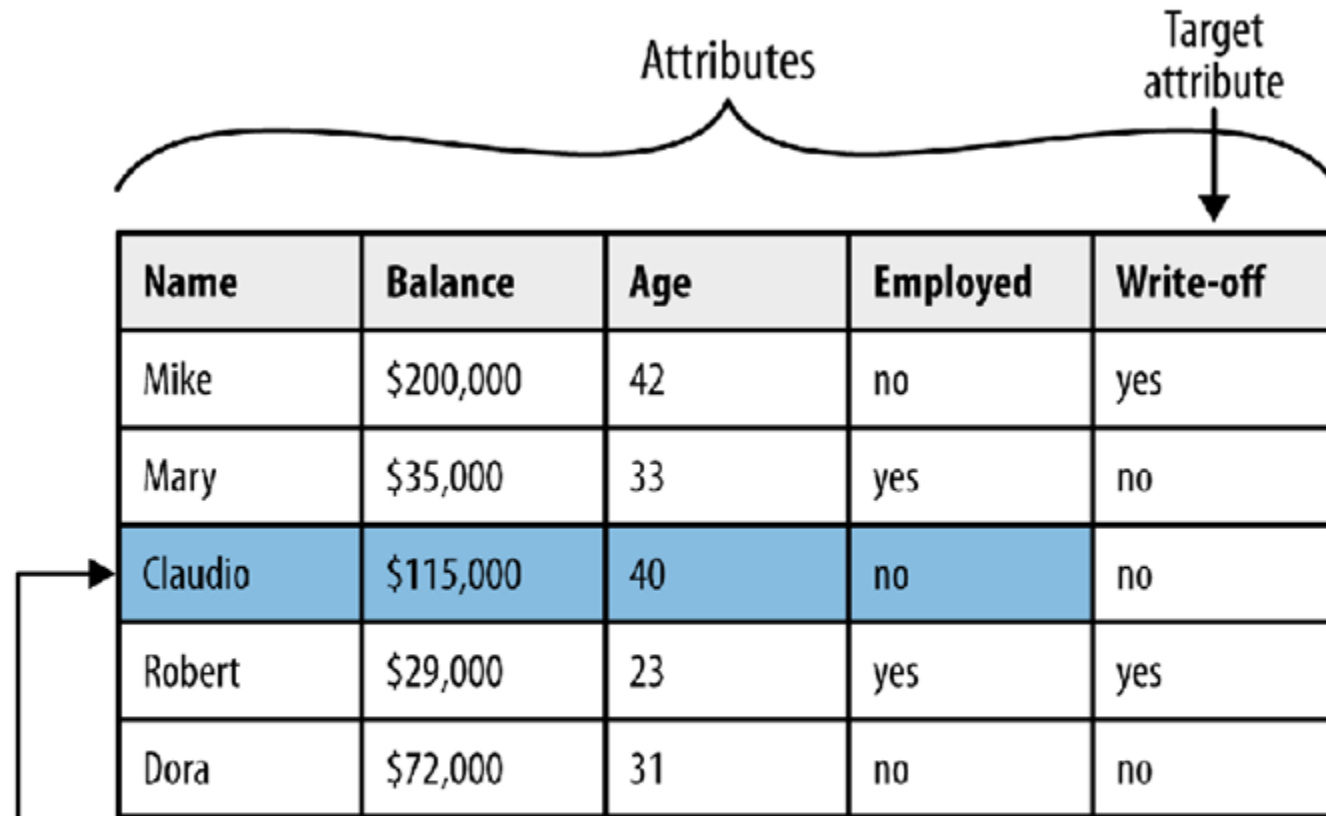
| Machine learning   | Statistics   |
|--|--|
| network, graphs  | model  |
| weights  | parameters   |
| learning   | fitting  |
| generalization   | test set performance                                 |
| supervised learning  | regression/classification                            |
| unsupervised learning  | density estimation, clustering                       |
| large grant = \$1,000,000                                    | large grant= \$50,000                                |
| nice place to have a meeting:<br>Snowbird, Utah, French Alps | nice place to have a meeting:<br>Las Vegas in August |



# ML terminology



# ML Terminology



| Name    | Balance   | Age | Employed | Write-off |
|---------|-----------|-----|----------|-----------|
| Mike    | \$200,000 | 42  | no       | yes       |
| Mary    | \$35,000  | 33  | yes      | no        |
| Claudio | \$115,000 | 40  | no       | no        |
| Robert  | \$29,000  | 23  | yes      | yes       |
| Dora    | \$72,000  | 31  | no       | no        |

This is one row (example).

Feature vector is: **<Claudio,115000,40,no>**

Class label (value of Target attribute) is **no**

# ML Terminology

- Attribute (field, variable, feature)
  - A quantity describing an instance. An attribute has a domain defined by the attribute type, which denotes the values that can be taken by an attribute. The following domain types are common:
    - Categorical: A finite number of discrete values. The type nominal denotes that there is no ordering between the values, such as last names and colors. The type ordinal denotes that there is an ordering, such as in an attribute taking on the values low, medium, or high.
    - Continuous (quantitative): Commonly, subset of real numbers, where there is a measurable difference between the possible values. Integers are usually treated as continuous in practical problems.



# ML Terminology

- A feature is the specification of an attribute and its value.
  - For example, color is an attribute. ``Color is blue" is a feature of an example.
  - Many transformations to the attribute set leave the feature set unchanged (for example, regrouping attribute values or transforming multi-valued attributes to binary attributes).
  - Some authors use feature as a synonym for attribute (e.g., in feature-subset selection).
- Data set: A schema and a set of instances matching the schema. Generally, no ordering on instances is assumed. Most machine learning work uses a single fixed-format table.

# ML Terminology

- Instance (example, case, record): A single object of the world from which a model will be learned, or on which a model will be used (e.g., for prediction). In most machine learning work, instances are described by feature vectors; some work uses more complex representations (e.g., containing relations between instances or between parts of instances).

# Data preparation problem

- **The Problem with Data Science**

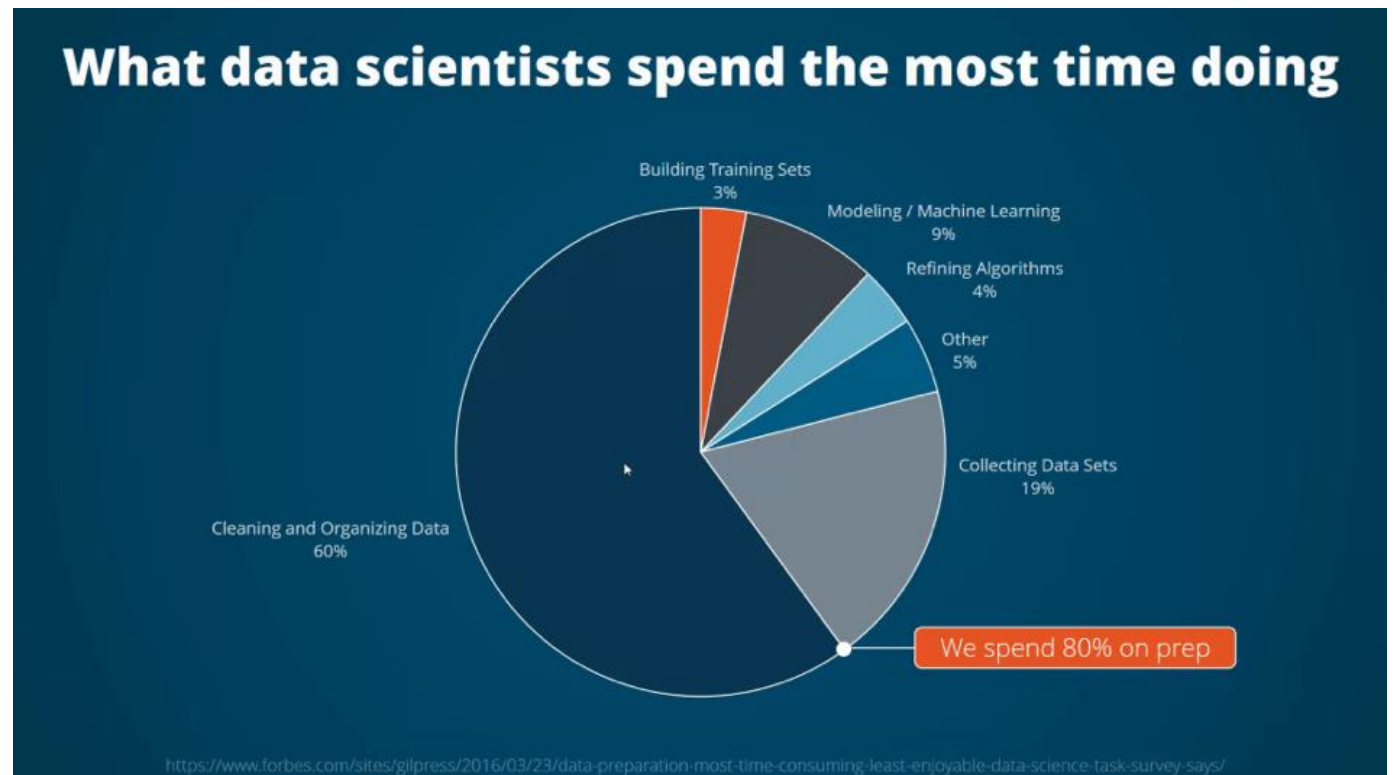
- Most organizations still face a data scientist bottleneck.
  - People want to do data science, they want to use machine learning models in production, but they can't achieve this because they don't have the required resources.
  - While colleges and others try to create as many data scientists as possible, it's just not happening fast enough and it's hard to hire them.
- The potential solution comes with two different flavors
  - The first flavor – let's empower more people to do the work data scientists do. So, that means we need to make this a little bit simpler because data science, in general, and machine learning is just too hard. If data preparation is such an important part of data science, we need to make that simpler as well.
  - The other flavor – let's make sure that the resources we currently have, or we are going to hire in the future, are more productive.

**So now, how does data preparation play an important role here?**

# Data preparation problem

- What do data scientists spend the most time doing?
- We all keep saying that 80% of our work as machine learning experts and data scientists is preparing the data.

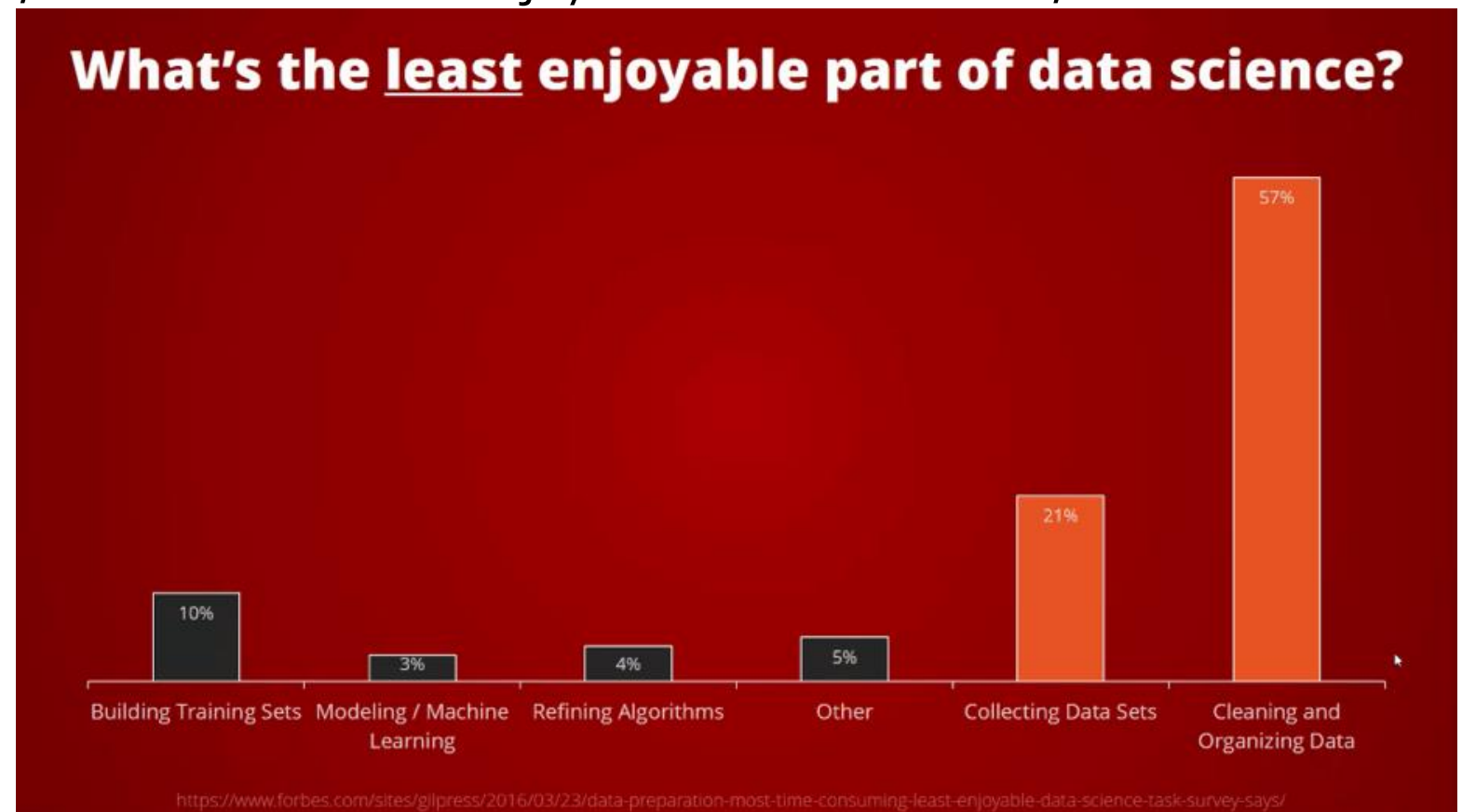
So, we spend more than 80% on data preparation. That's probably because we love it so much, right? Wrong.



# Data preparation problem

- **What's the least enjoyable part of data science?**
- 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.

57% said it's cleaning and organizing data



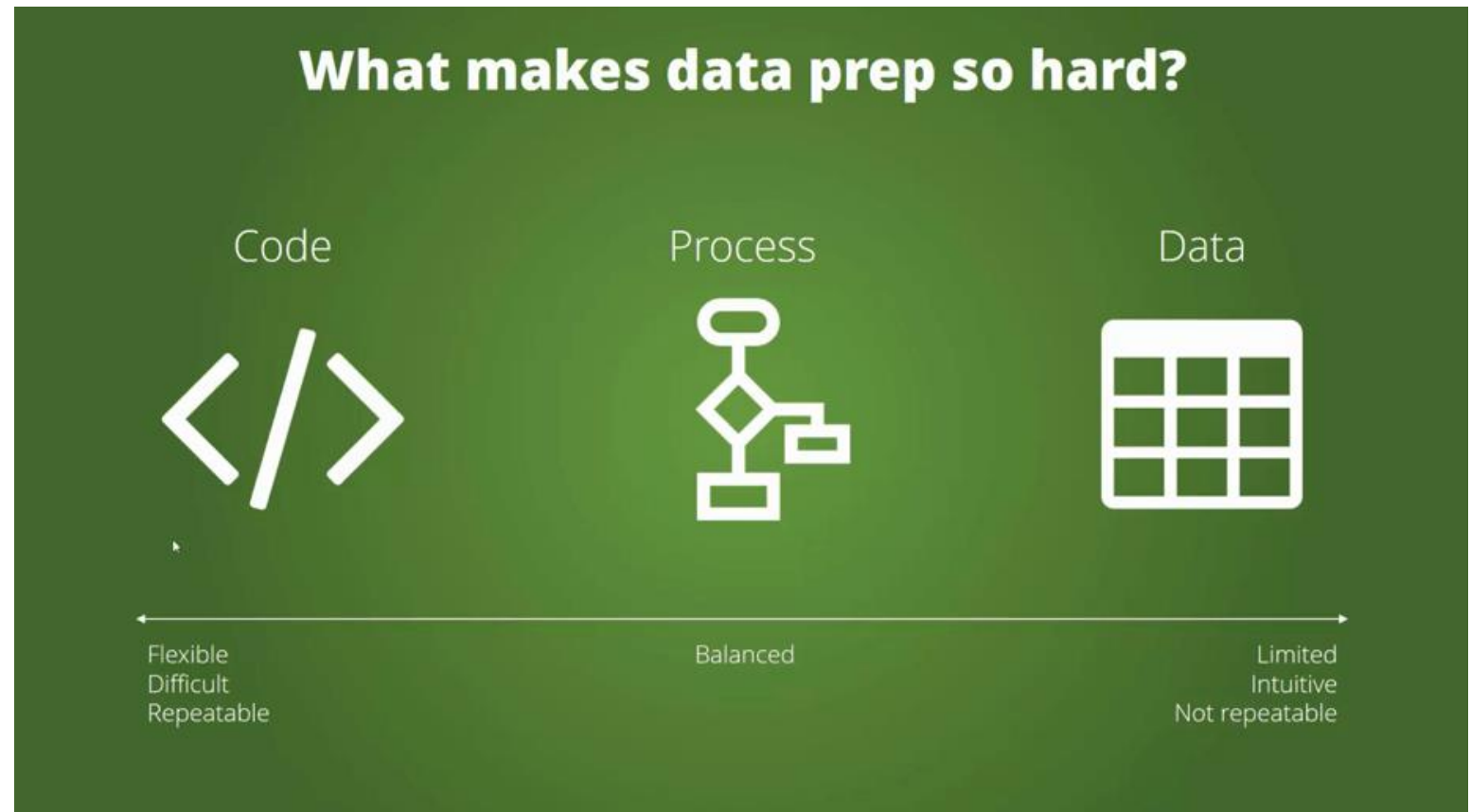
# Data preparation problem

- **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.

# Data preparation problem

- What makes data preparation so difficult?

1. Code-based approach to data science: Python, R
2. Process-based approach to data science: orange, rapidminer
3. Data-centric approach to data science: Excel



# Data preparation problem

- The path to be a data scientist

