The AI revolution in scientific research

Source: The Royal Society and The Alan Turing Institute Report

From the t-test to the frontiers of AI

- Scientists aspire to understand the workings of nature, people, and society by:
    - Formulating hypotheses,
    - Designing experiments, and
    - Collecting data

The aim is analyzing and better understanding natural, physical, and social phenomena

From the t-test to the frontiers of AI

- Data collection and analysis is a core element of the scientific method, and scientists have long used statistical techniques to aid their work
  - t-test (Student test) gave researchers a new tool to extract insights from data in order to test the veracity of their hypotheses.
  - Such mathematical frameworks were vital in extracting as much information as possible from data that had often taken significant time and money to generate and collect

From the t-test to the frontiers of AI

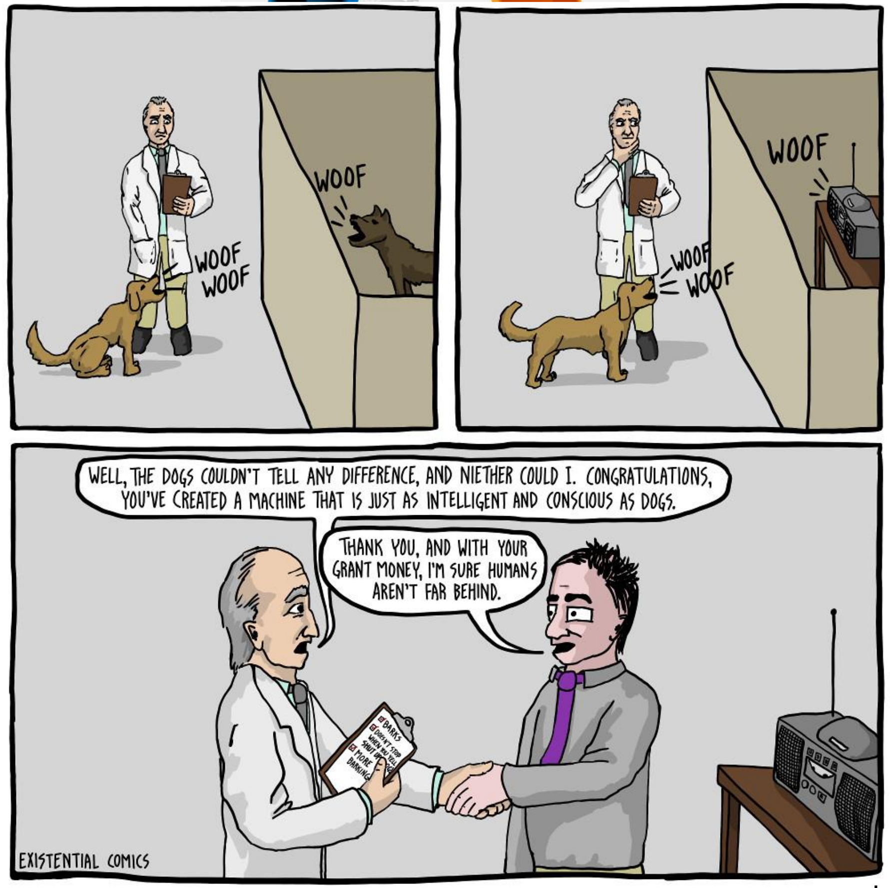
- The application of statistical methods to scientific challenges can be seen throughout history:
  - The analysis by Johannes Kepler of the astronomic measurements of Tycho Brahe in the early seventeenth century led to his formulation of the laws of planetary motion, which subsequently enabled Isaac Newton (and others) to formulate the law of universal gravitation.

From the t-test to the frontiers of AI

- The application of statistical methods to scientific challenges can be seen throughout history:
  - The laboratory at Rothamsted was established as a centre for agricultural research, running continuously monitored experiments from 1856 which are still running to this day. Ronald Fisher – a prominent statistician – was hired to work there in 1919 to direct analysis of these experiments. His work went on to develop the theory of experimental design and lay the groundwork for many fundamental statistical methods that are still in use today.

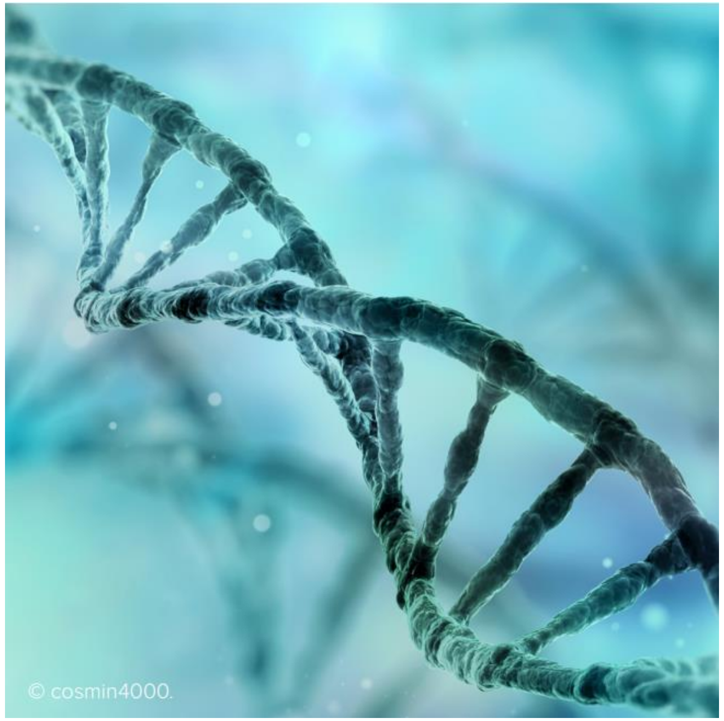
From the t-test to the frontiers of AI

- The development of artificial intelligence (AI) techniques offered additional tools for extracting insights from data.
    - Alan Turing grappled with the idea of machine intelligence.
    - In 1950, he posed the question “can machines think?”, and suggested a test for machine intelligence – subsequently known as the Turing Test – in which a machine might be called intelligent, if its responses to questions could convince a person that it was human



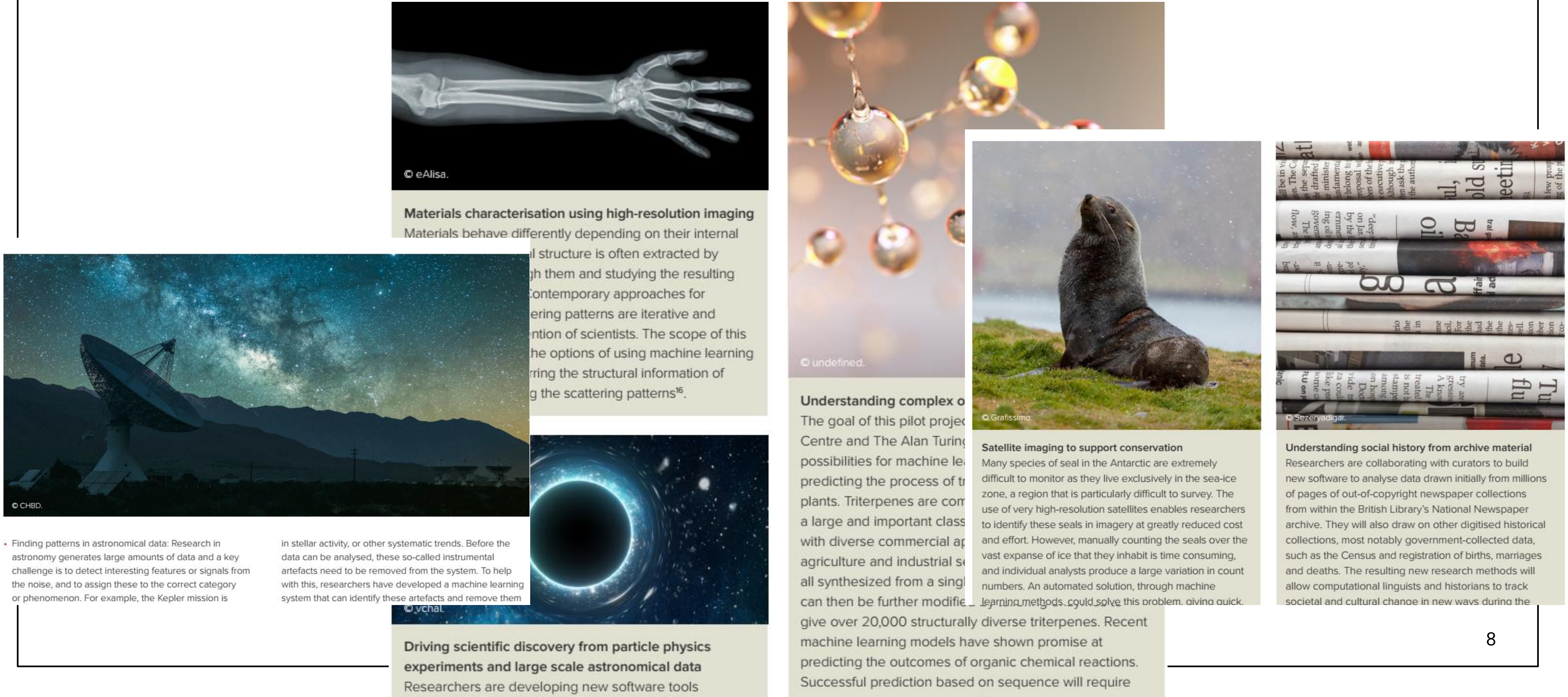
AI as an enabler of scientific discovery

- Using genomic data to predict protein structures:
  - By predicting the protein's shape, scientists can identify proteins that play a role in diseases, improving diagnosis and helping develop new treatments.
  - Determining the shape of a protein from its corresponding genetic sequence – the protein-folding challenge – is a complex task.



- The AlphaFold project at DeepMind has created a deep neural network that predicts the distances between pairs of amino acids and the angles between their bonds, and in so doing produces a highly-accurate prediction of an overall protein structure.

AI as an enabler of scientific discovery



© eAlisa.

Materials characterisation using high-resolution imaging
Materials behave differently depending on their internal
l structure is often extracted by
gh them and studying the resulting
Contemporary approaches for
ering patterns are iterative and
ntion of scientists. The scope of this
he options of using machine learning
rring the structural information of
g the scattering patterns[16].

© CHBD.

- Finding patterns in astronomical data: Research in astronomy generates large amounts of data and a key challenge is to detect interesting features or signals from the noise, and to assign these to the correct category or phenomenon. For example, the Kepler mission is

in stellar activity, or other systematic trends. Before the data can be analysed, these so-called instrumental artefacts need to be removed from the system. To help with this, researchers have developed a machine learning system that can identify these artefacts and remove them

© vchal.

Driving scientific discovery from particle physics experiments and large scale astronomical data
Researchers are developing new software tools

© undefined.

Understanding complex o
The goal of this pilot projec
Centre and The Alan Turing
possibilities for machine le
predicting the process of tr
plants. Triterpenes are com
a large and important class
with diverse commercial ap
agriculture and industrial s
all synthesized from a singl
can then be further modifie
give over 20,000 structurally diverse triterpenes. Recent machine learning models have shown promise at predicting the outcomes of organic chemical reactions. Successful prediction based on sequence will require

© Grafissimo.

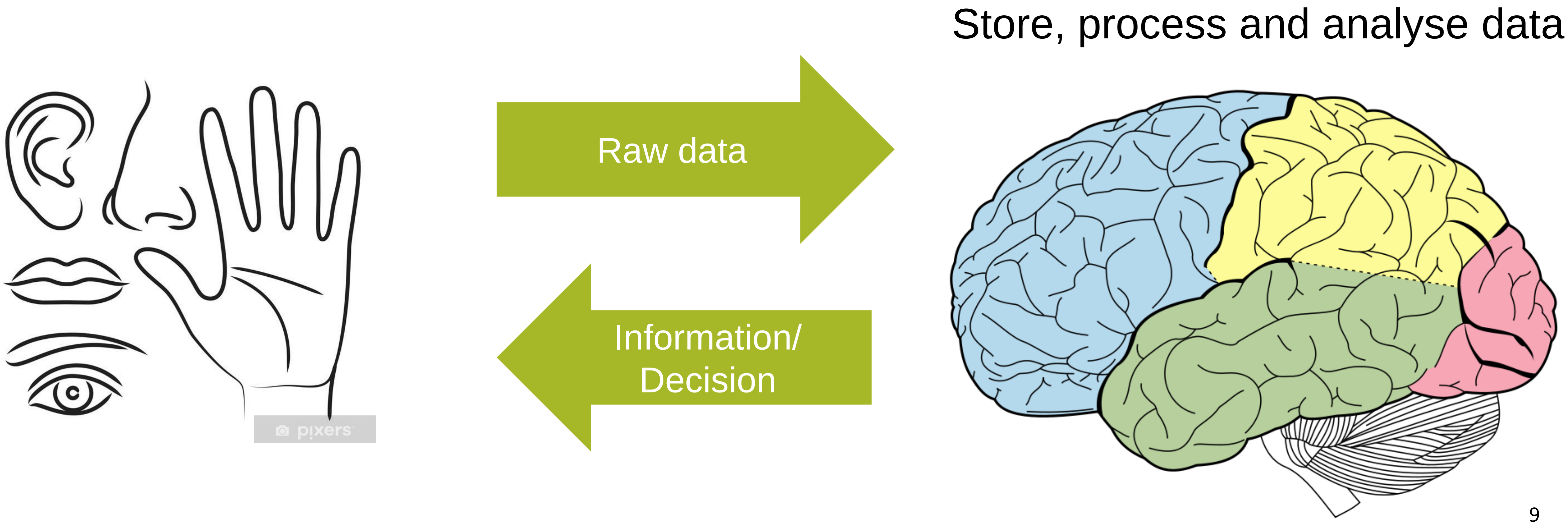
Satellite imaging to support conservation
Many species of seal in the Antarctic are extremely difficult to monitor as they live exclusively in the sea-ice zone, a region that is particularly difficult to survey. The use of very high-resolution satellites enables researchers to identify these seals in imagery at greatly reduced cost and effort. However, manually counting the seals over the vast expanse of ice that they inhabit is time consuming, and individual analysts produce a large variation in count numbers. An automated solution, through machine learning methods, could solve this problem, giving quick,

© Sezeryadigar.

Understanding social history from archive material
Researchers are collaborating with curators to build new software to analyse data drawn initially from millions of pages of out-of-copyright newspaper collections from within the British Library's National Newspaper archive. They will also draw on other digitised historical collections, most notably government-collected data, such as the Census and registration of births, marriages and deaths. The resulting new research methods will allow computational linguists and historians to track societal and cultural change in new ways during the

Research questions to advance the application of AI in science

- Let start with the Human Brain: Information processing



Research questions to advance the application of AI in science

- DATA MANAGEMENT
  - Is there a principled method to decide what data to keep and what to discard, when an experiment or observation produces too much data to store? How will this affect the ability to re-use the data to test alternative theories to the one that informed the filtering decision?

Research questions to advance the application of AI in science

- DATA MANAGEMENT
  - What does ‘open data’ mean in practice where the data sets are just too large, complex and heterogenous for anyone to actually access and understand them in their entirety

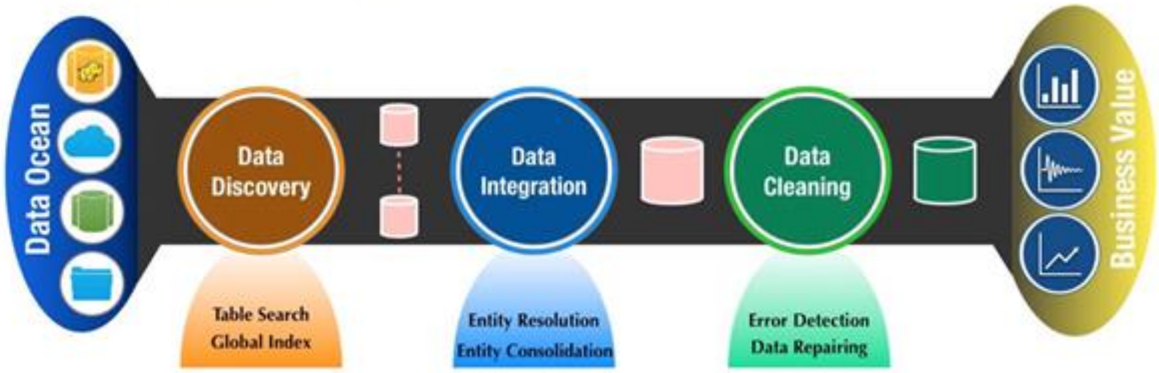
Research questions to advance the application of AI in science

- DATA MANAGEMENT
  - How can scientists search efficiently for rare or unusual events and objects in large and noisy data sets?

Research questions to advance the application of AI in science

- DATA MANAGEMENT

The Data Preparation Pipeline



Or if you want to look at it differently ...

Collection

Cleaning

Integration

Analysis

Research questions to advance the application of AI in science

- AI METHODS AND CAPABILITIES

  - How can machine learning help integrate data from different sources collected under different conditions and for different purposes, in a way that is scientifically valid (for example, integrate observations of the same system taken at different scales)

Research questions to advance the application of AI in science

- AI METHODS AND CAPABILITIES

    - How can researchers re-use data which they have already used to inform theory development, while maintaining the rigour of their work

Research questions to advance the application of AI in science

- AI METHODS AND CAPABILITIES

  - How can AI methods produce results which are transparent as to how they were obtained, and interpretable within the disciplinary context

Research questions to advance the application of AI in science

- AI METHODS AND CAPABILITIES
  - How can research help create more advanced, and more accurate, methods of verifying machine learning systems to increase confidence in their deployment

Research questions to advance the application of AI in science

- INTEGRATING SCIENTIFIC KNOWLEDGE

  - Is there a rigorous way to incorporate existing theory/ knowledge into a machine learning algorithm, to constrain the outcomes to scientifically plausible solutions?

  - How can AI be used to actually discover and create new scientific knowledge and understanding, and not just the classification and detection of statistical patterns?

What is Artificial Intelligence?

What is Artificial Intelligence?

- The expression ‘artificial intelligence’ today is therefore an umbrella term.

- It refers to a suite of technologies that can perform complex tasks when acting in conditions of uncertainty, including visual perception, speech recognition, natural language processing, reasoning, learning from data, and a range of optimisation problems

What is Artificial Intelligence?

Artificial

Not natural

Intelligence

Ability to understand, think, learn and act

What is Artificial Intelligence?

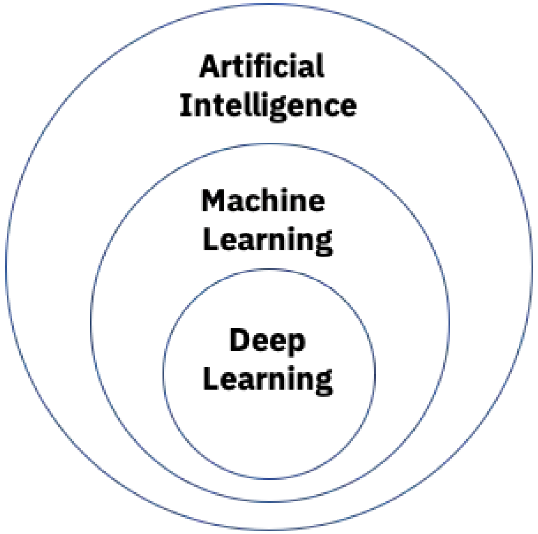
- Artificial Intelligence (AI)
  - Artificial intelligence (AI) is intelligence demonstrated by machines, as opposed to natural intelligence displayed by animals including humans.
  - Leading AI textbooks define the field as the study of "intelligent agents": any system that perceives its environment and takes actions that maximize its chance of achieving its goals.
  - Some popular accounts use the term "artificial intelligence" to describe machines that mimic "cognitive" functions that humans associate with the human mind, such as "learning" and "problem solving", however, this definition is rejected by major AI researchers.

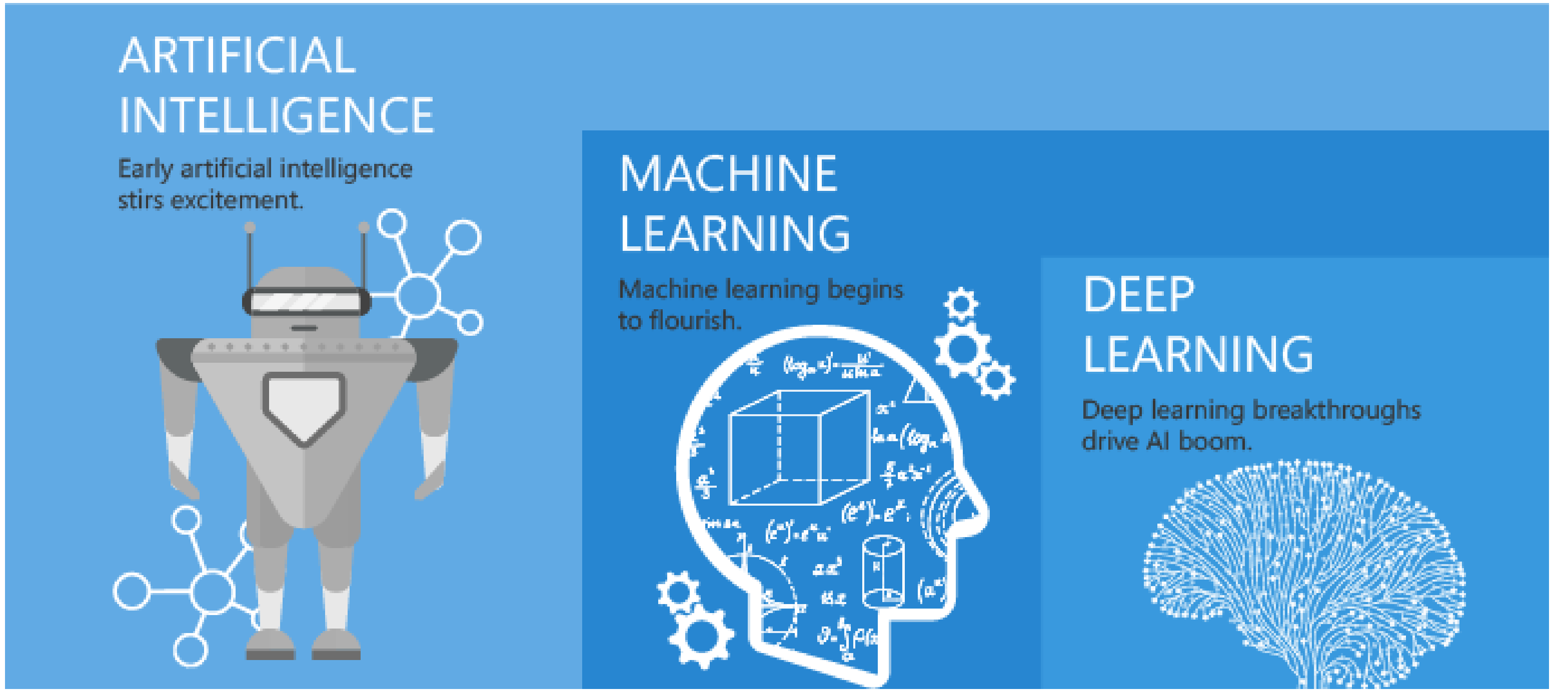
What is Artificial Intelligence?

- Artificial General Intelligence (AGI)
  - AGI is the hypothetical ability of an intelligent agent to understand or learn any intellectual task that a human being can.
  - It is a primary goal of some artificial intelligence research and a common topic in science fiction and futures studies.
  - AGI can also be referred to as strong AI, full AI, or general intelligent action. (Although academic sources reserve the term "strong AI" for computer programs that experience sentience or consciousness.)

Artificial Intelligence

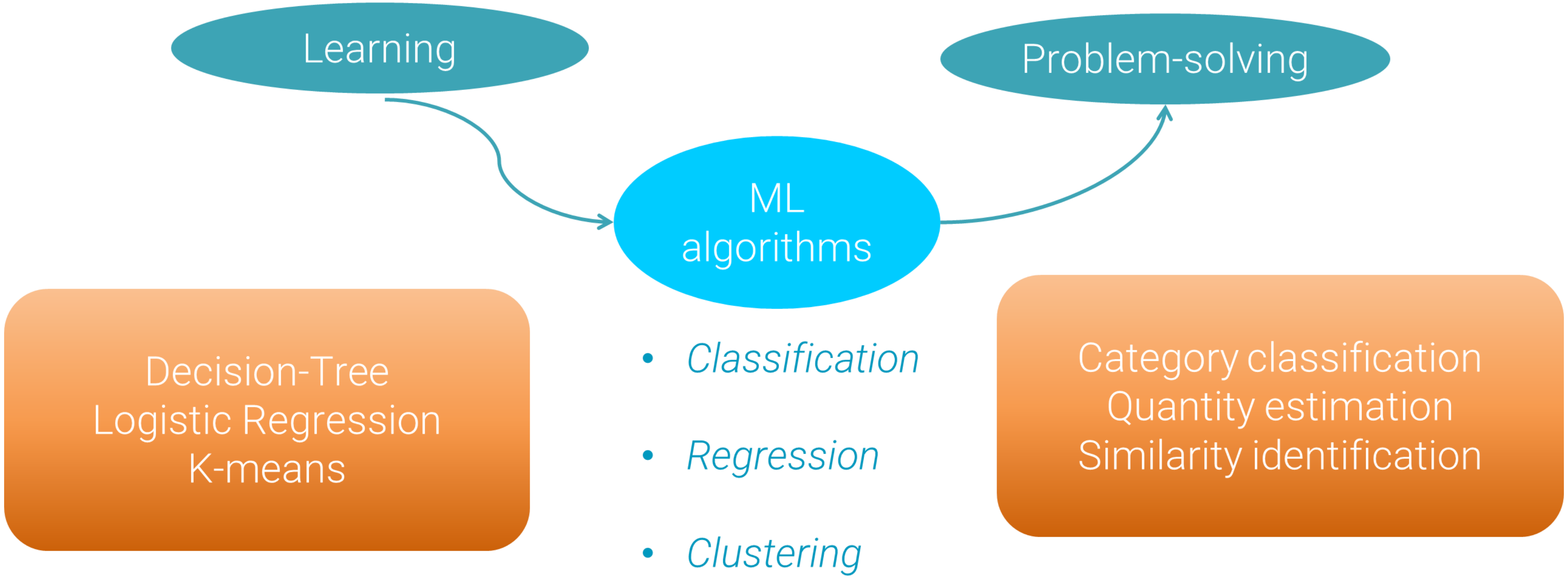
- It’s not the destination, it's the journey





Artificial Intelligence

- Machine Learning (ML)
  - Is a method of achieving AI by using algorithms enabling machines to learn from data how to make decisions.



Data Science and Machine Learning

https://github.com/berradais/light-dna-2020-2021

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?

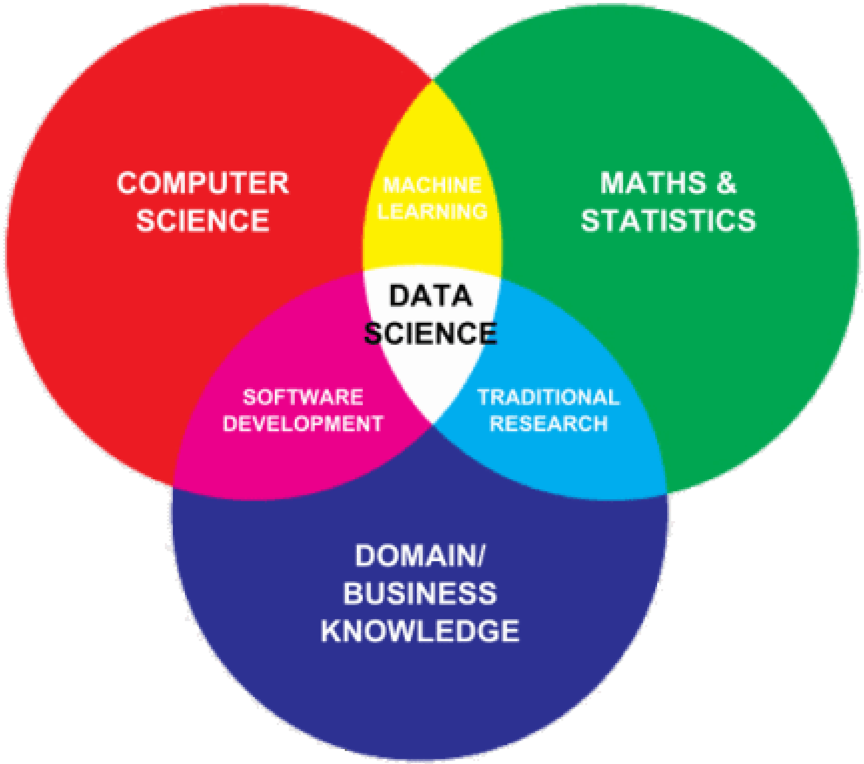
Data Science is: making sense of Data

- 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

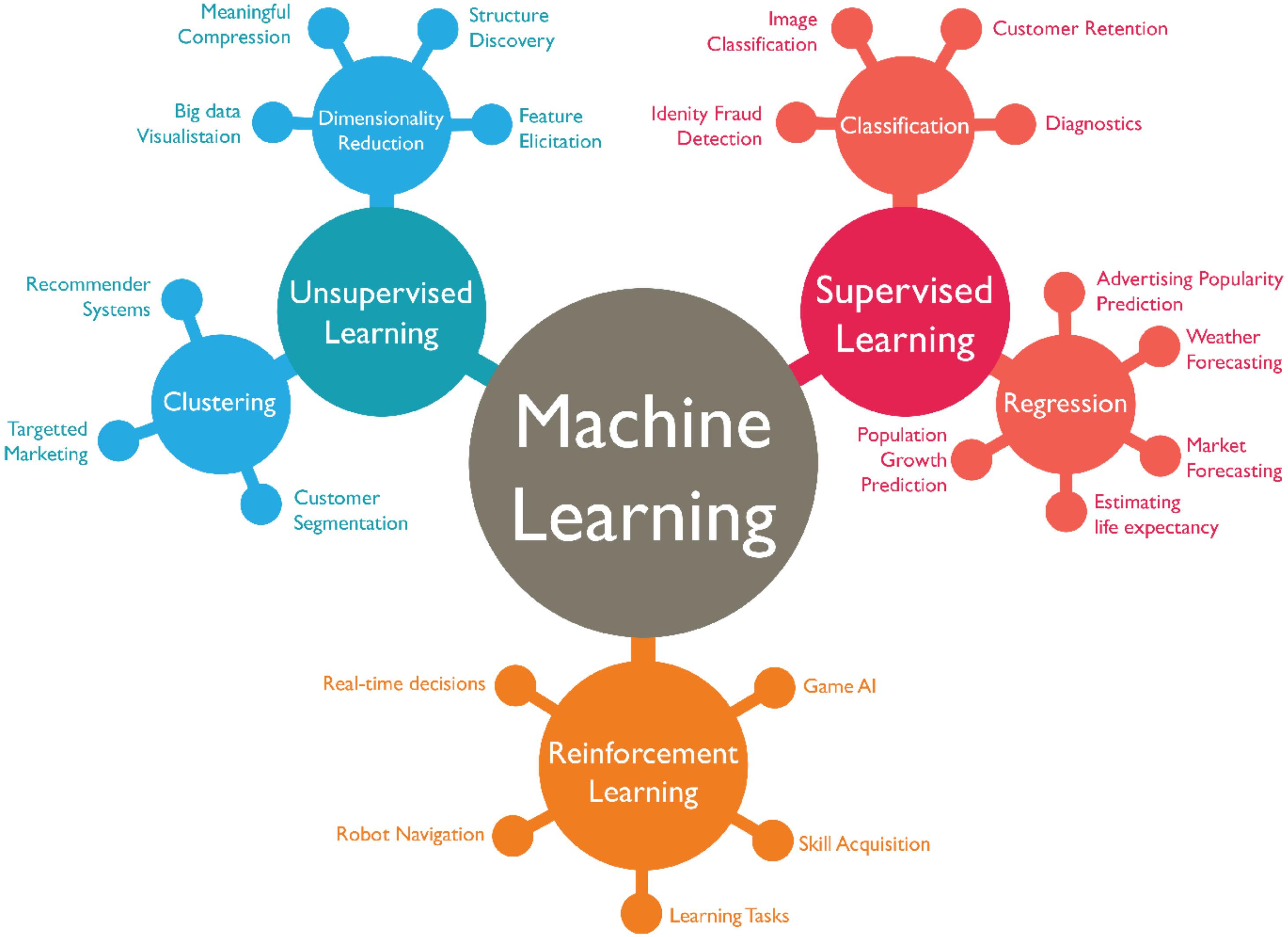
- Statisticians
- Mathematicians
- Computer Scientists in
  - Data mining
  - Artificial Intelligence & Machine Learning
  - Systems Development and Integration
  - Database development
  - Analytics
- Domain Experts
  - Medical experts
  - Geneticists
  - Finance, Business, Economy experts
  - etc.



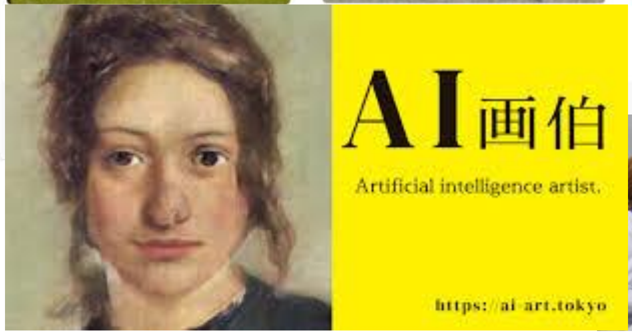
Data Science is: about the whole processing pipeline to extract information out of data

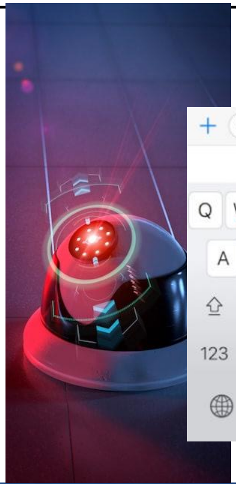
- Data Scientist understand and care about the whole data pipeline:
  - A data pipeline consists of 3 steps:
    1. Preparing to run a model: Gathering, cleaning, integrating, restructuring, transforming, loading, filtering, deleting, combining, merging, verifying, extracting, shaping
    2. Running the model
    3. Communicating the results

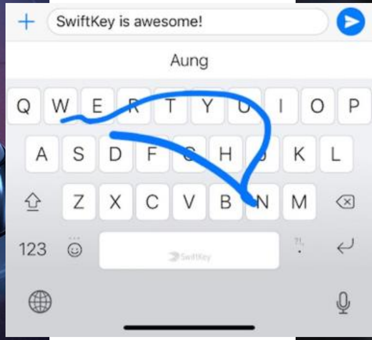
Artificial Intelligence Zoo

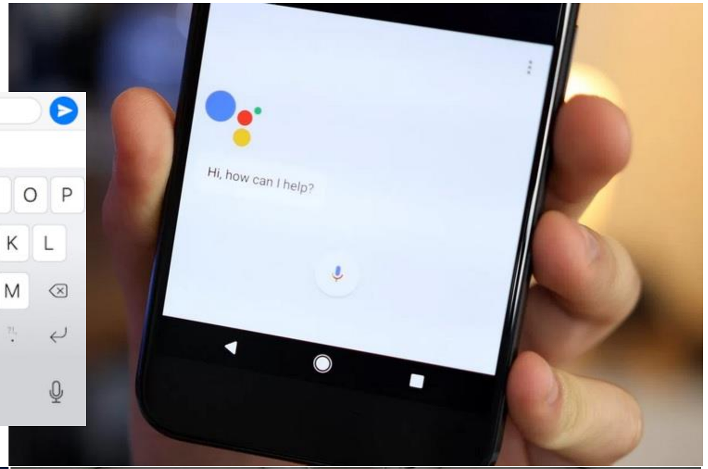


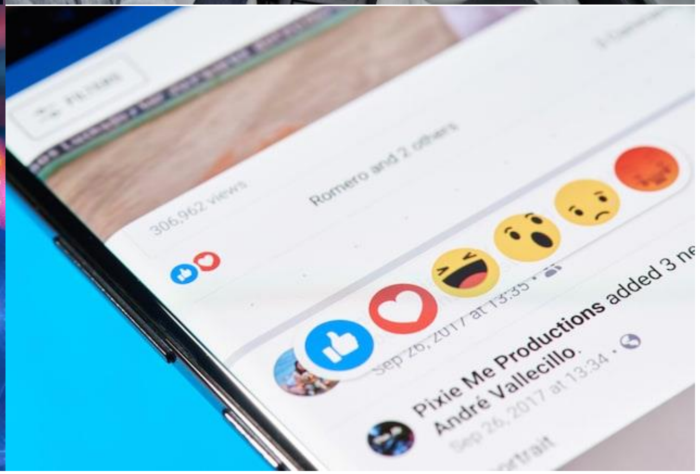
Data Science Success











Data Science Success

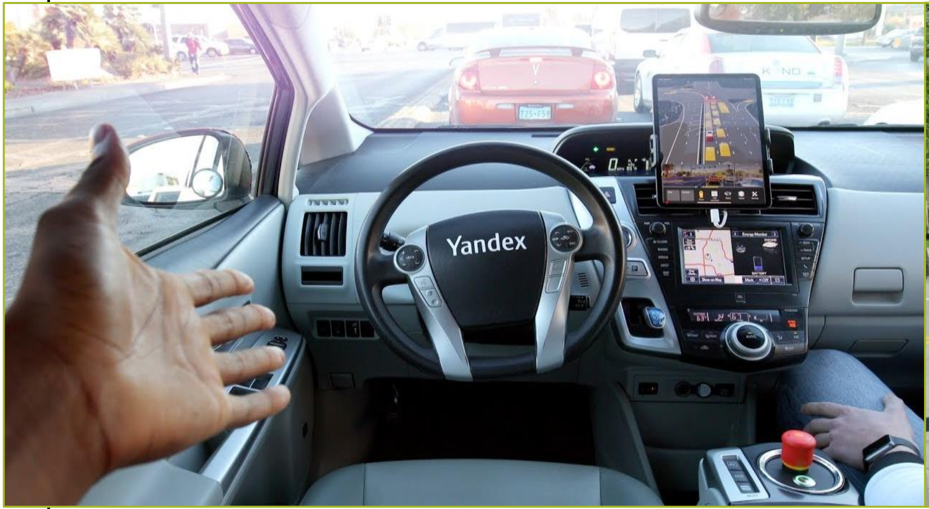
Mercedes-Benz Autonomous Car Interior Concept (Luxury F 015)

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



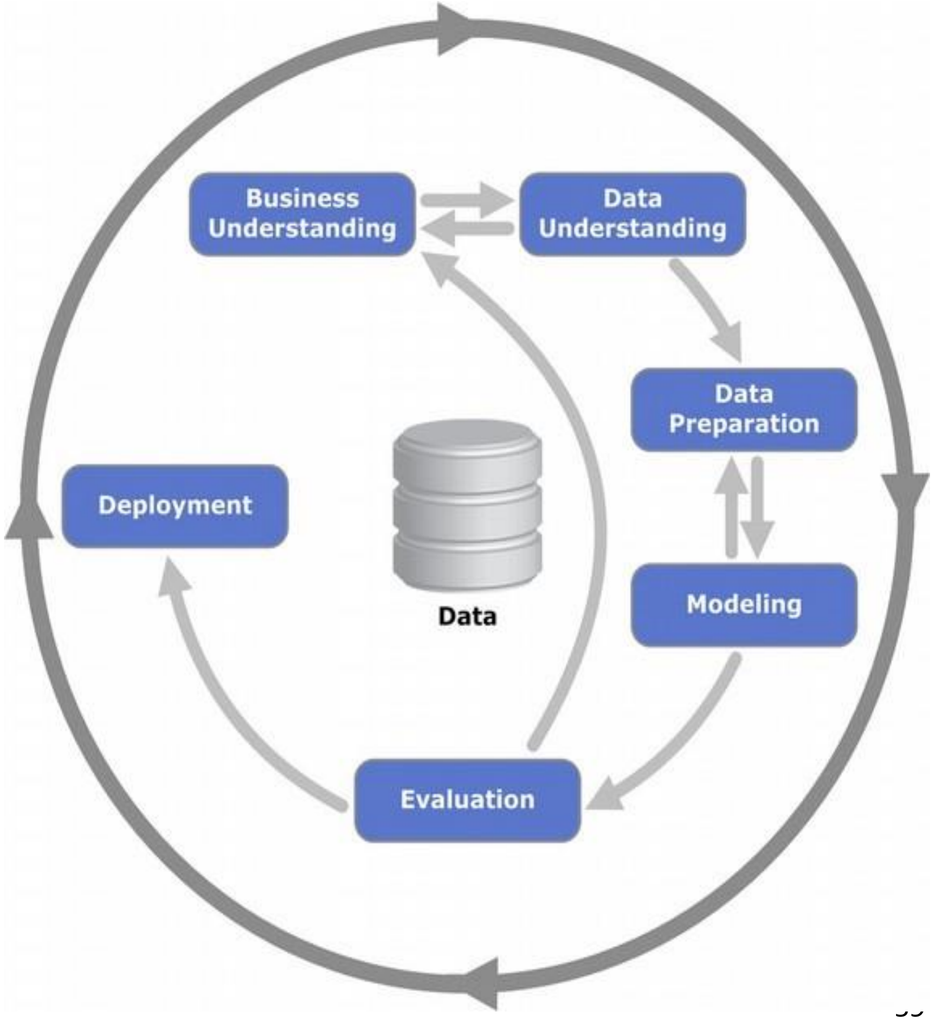
Data Science Success

Data Science Principals

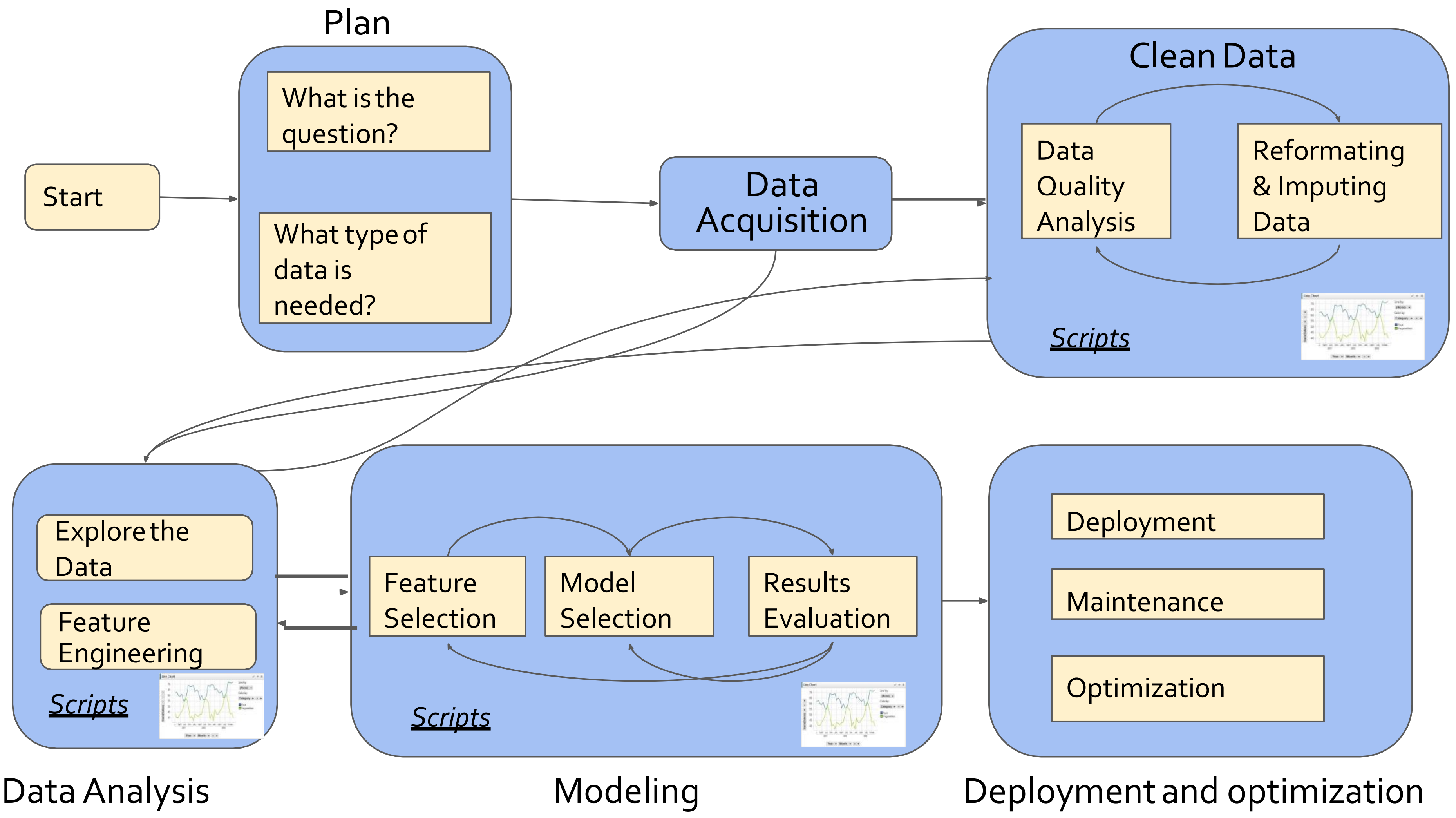
1. Data Science is a process
2. ML is optimization of loss functions
3. ML must generalize to unseen data
4. Evaluate data science in its operational context
5. Similar entities can have similar unseen attributes
6. Correlation, not causation

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

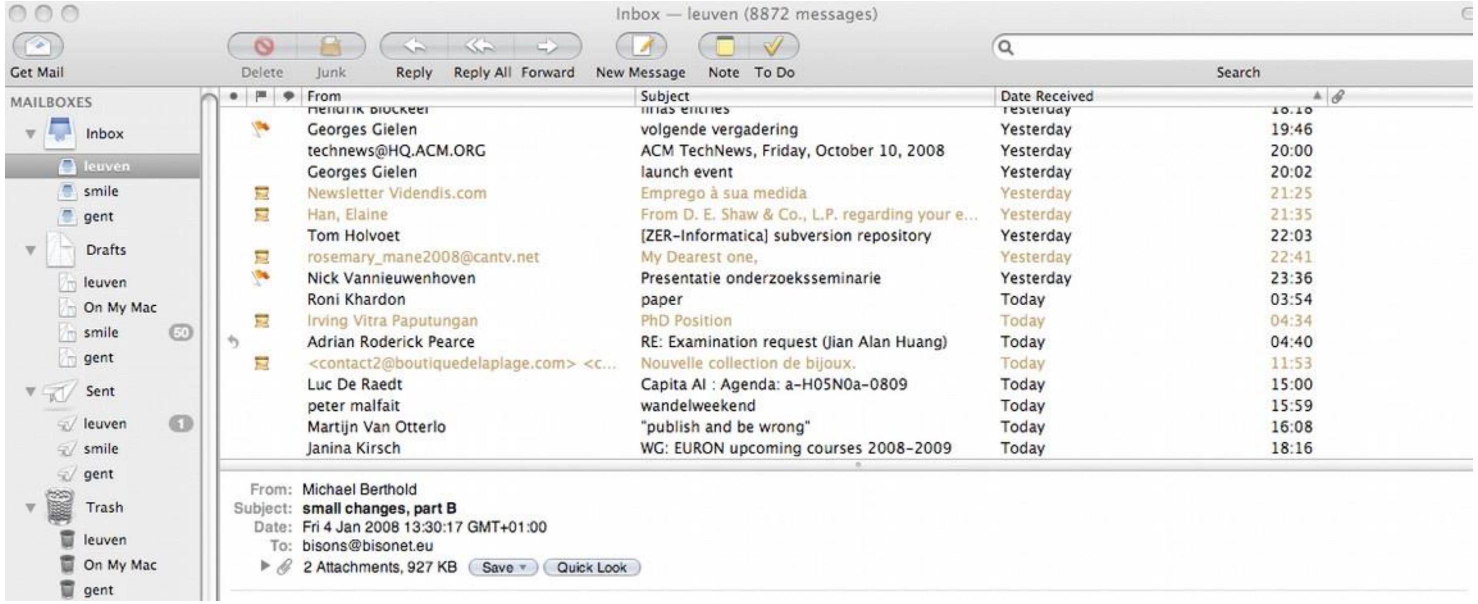


P1: Data Science is a process
A simple workflow from a technical PoV



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, 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: 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, tree-based)
- Does it work well? → evaluate

P1: Data Science is a process
A simple workflow

<u>Evaluation</u>

on unseen emails

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

P2: Machine learning is optimization

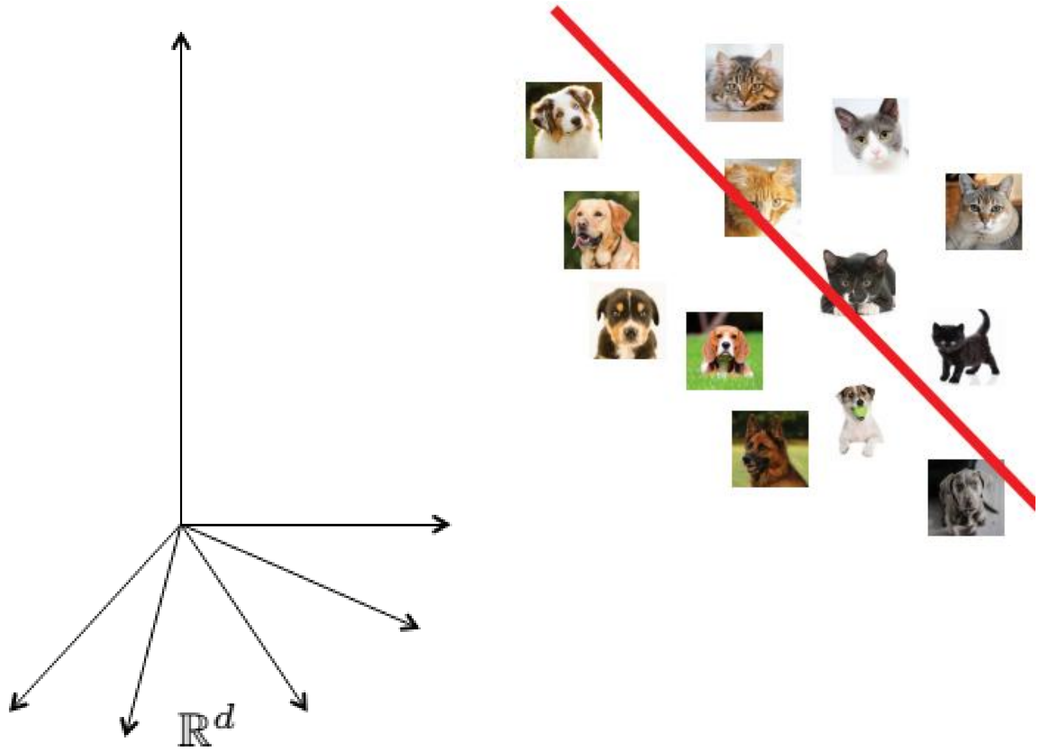
- Data vectors $\mathbf{x} \in \mathbb{R}^d$
  (e.g. for 512×512 images $d \approx 10^5$)

- 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

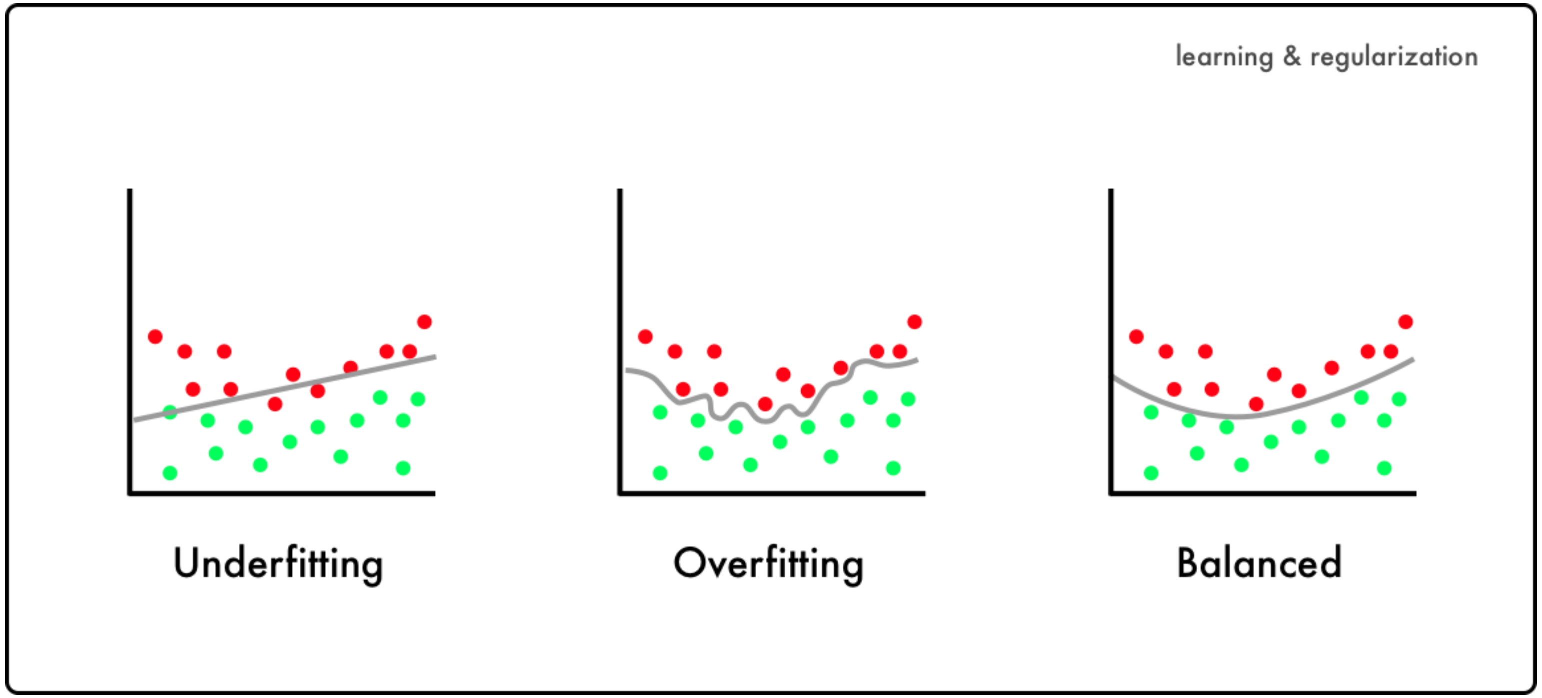
$$\boldsymbol{\Theta}^* = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \sum_{i=1}^{T} \ell(f_{\boldsymbol{\Theta}}(\mathbf{x}_i), y_i)$$

P2: Machine learning is optimization

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

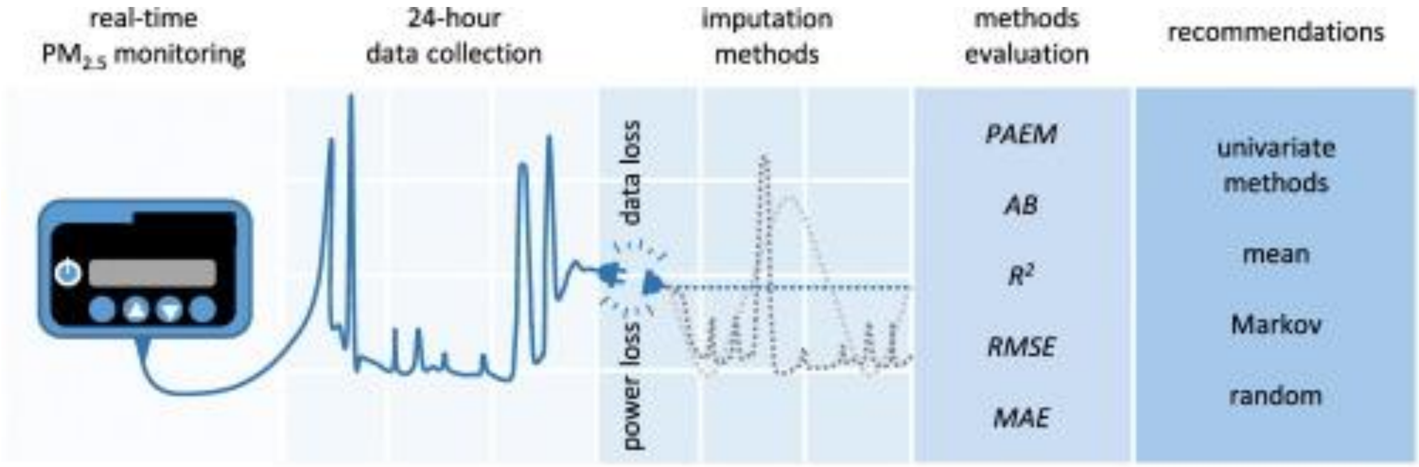
P2: Generalizing

- 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



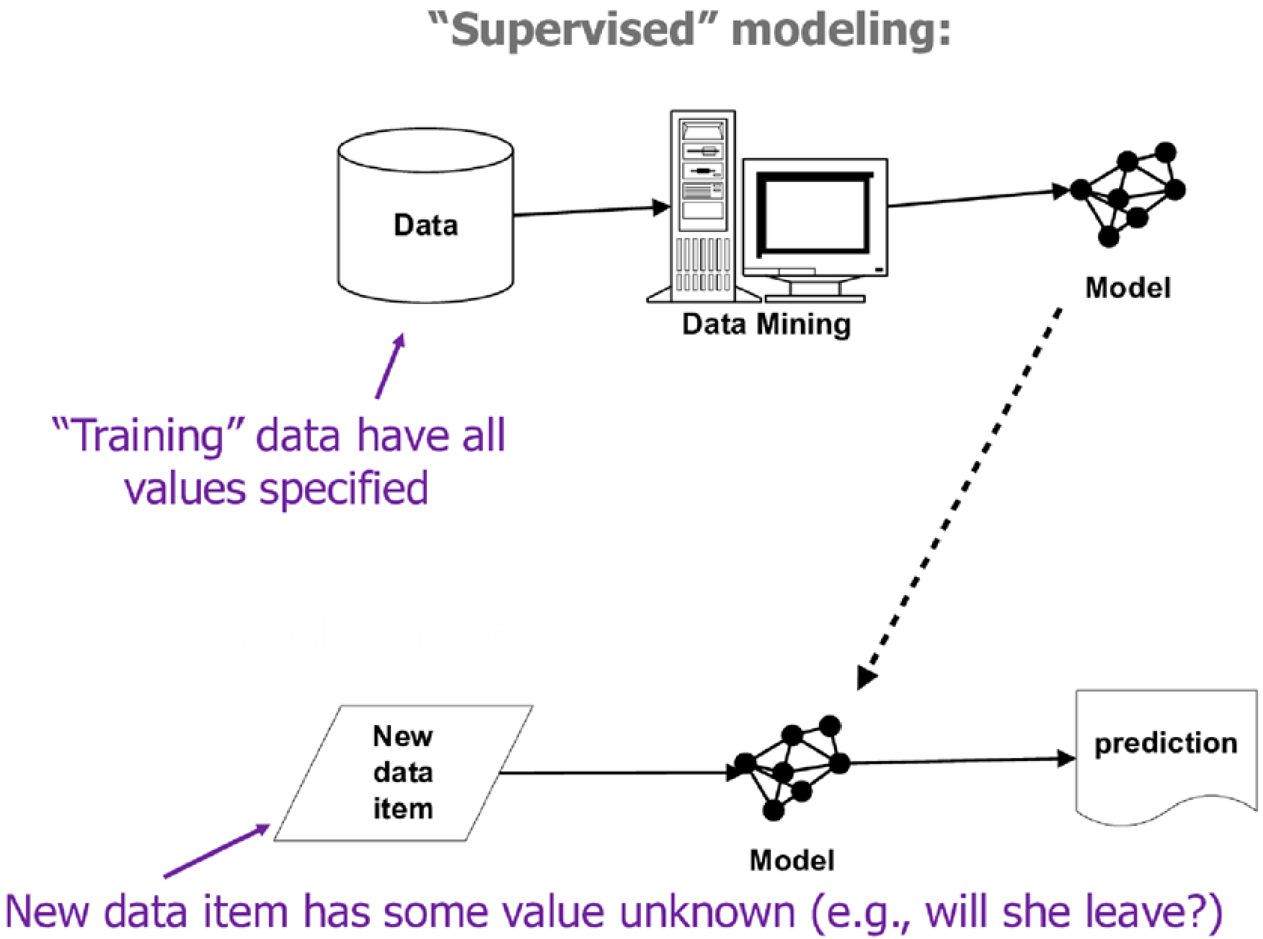
P3: Missing information

- 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



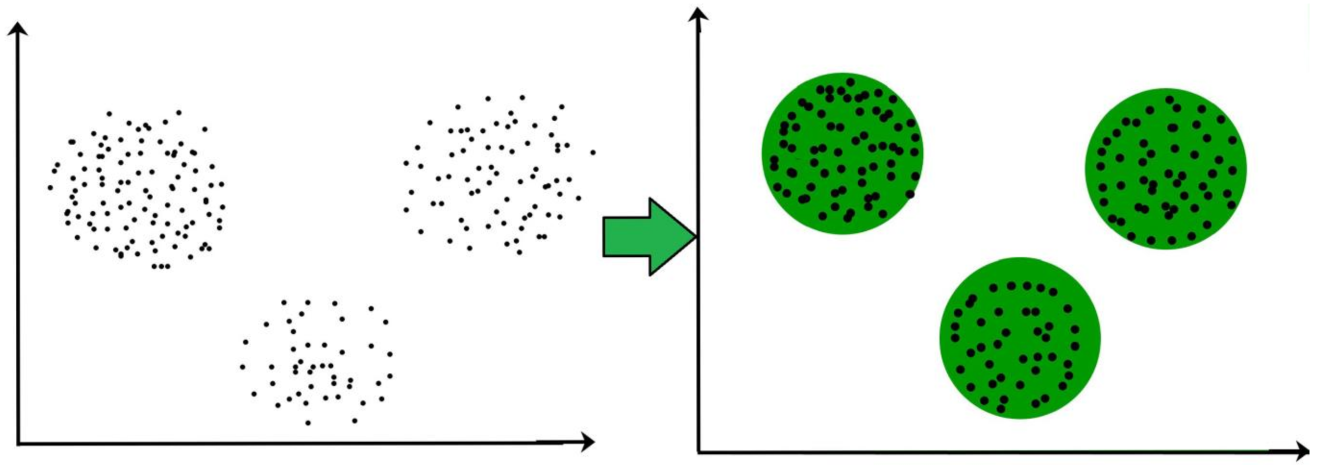
P4: Data science needs to be evaluated in the context of operation

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



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

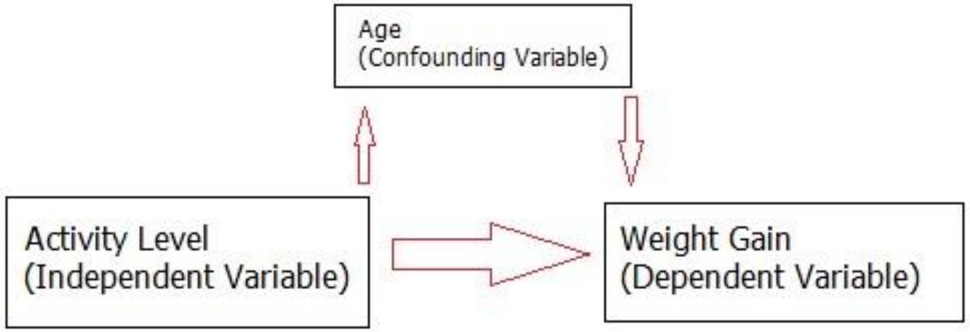
- Clustering
- Also optimization, e.g. min. distances to cluster center
- Key concept: <u>distance</u> between objects
  - Euclidean, Manhattan, edit distances (strings), Dynamic time warping (temporal sequences), ...



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
    - $Y = 20*X1 - 12*X2 + 300*X3 + 99*X4 - 299*X5$
    - Which feature has most impact?



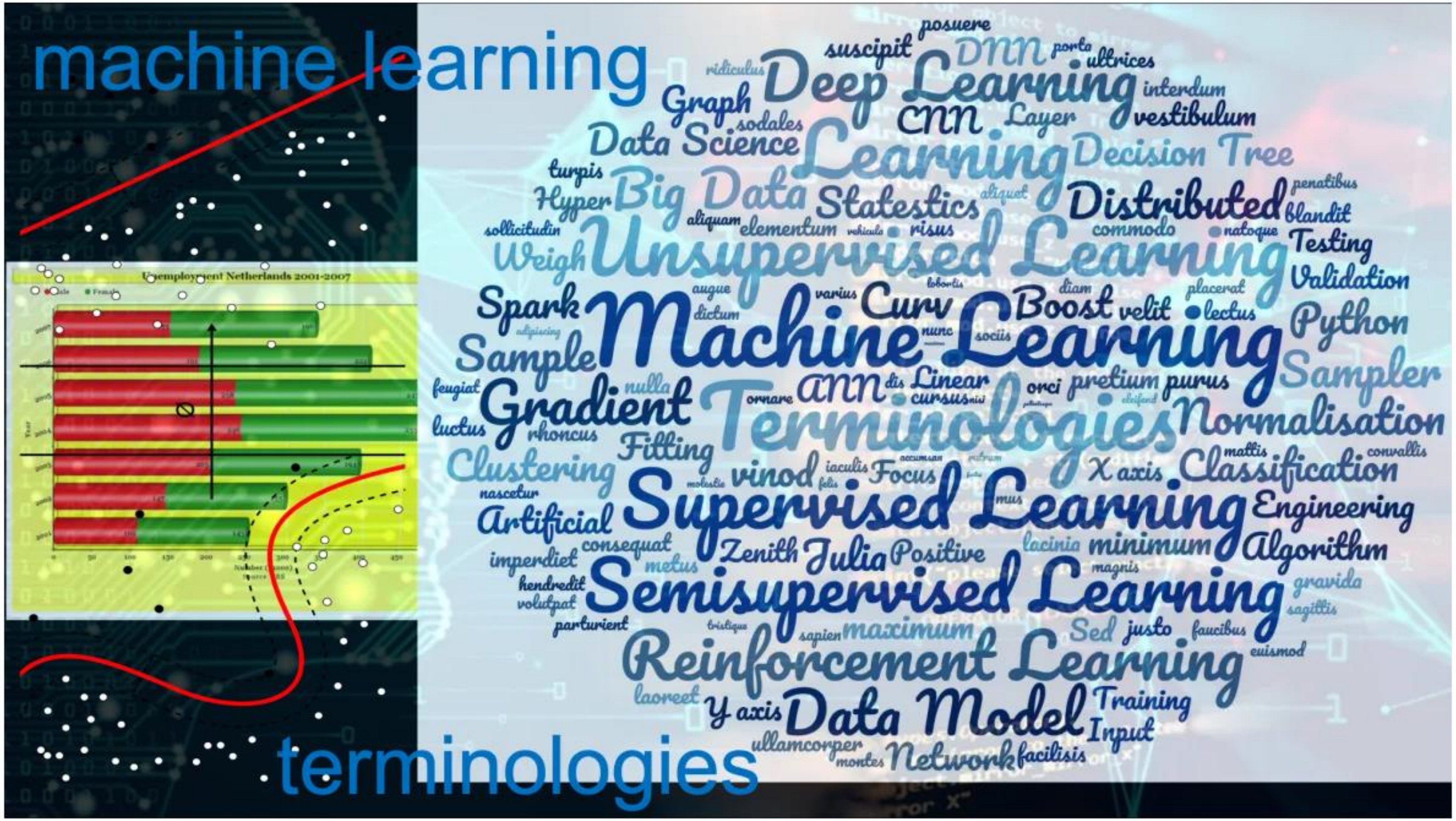
ML vs Stat

- 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: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

ML terminology



ML Terminology

Attributes

Target attribute

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.

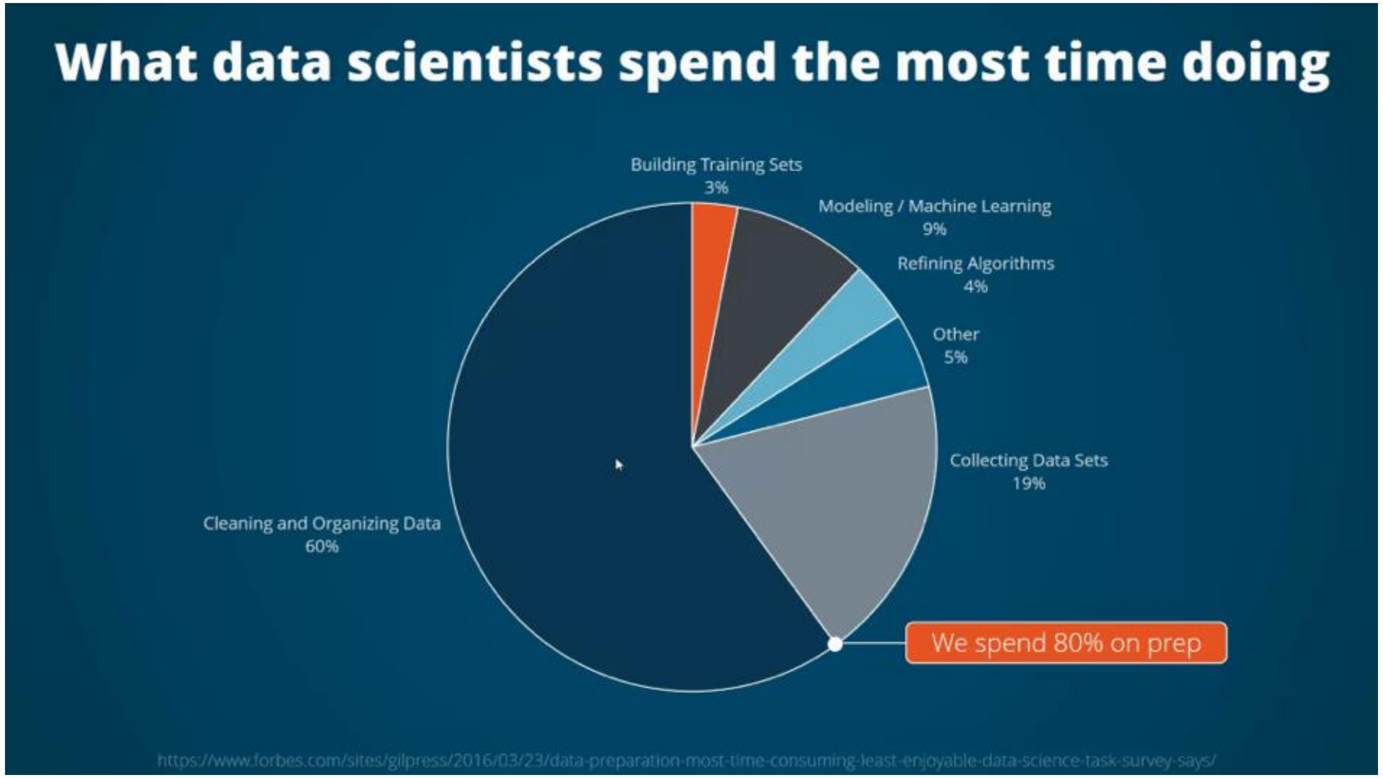
ML Terminology

- 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

- **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.

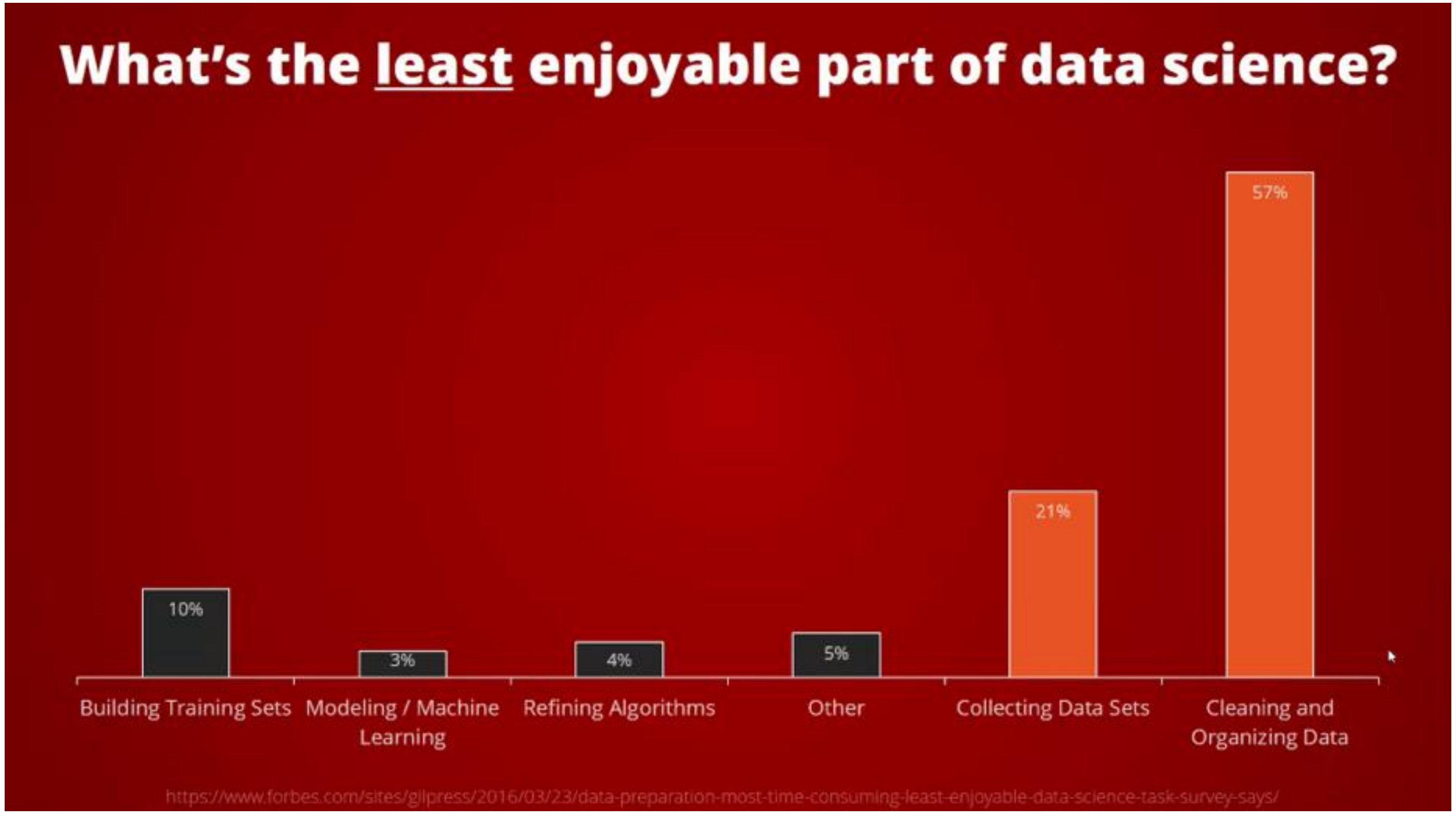
So, we spend more than 80% on data preparation. That's probably because we love it so much, right? Wrong.



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

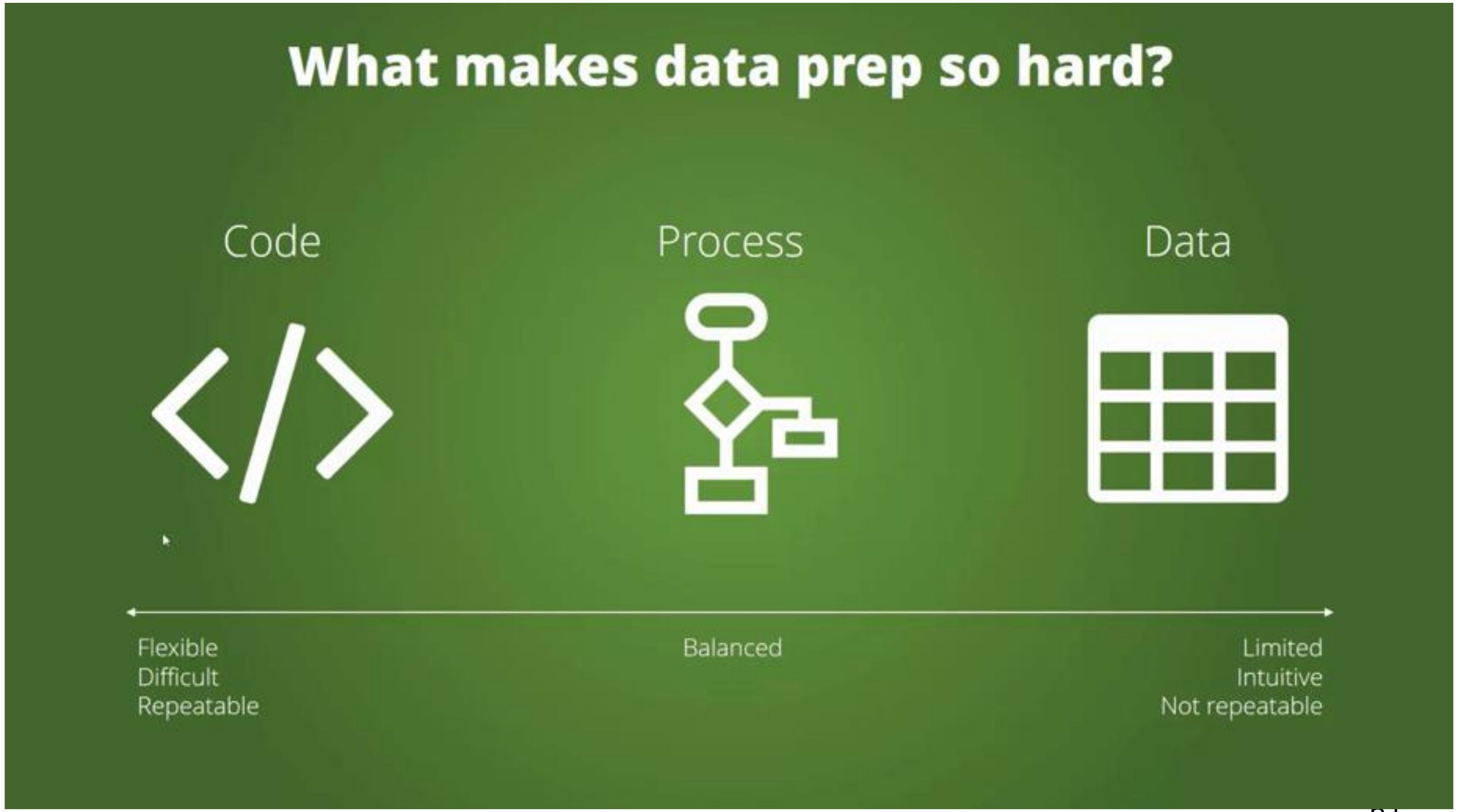
- **Why data preparation is so important?**

- 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**