Introduction to Data Science

How to Execute a Machine Learning Project

version April 2021

TU/e Eindhoven University of Technology | Tilburg University | 's-Hertogenbosch | Provincie Noord-Brabant | JADS Jheronimus Academy of Data Science

1

Copyright notice

2

Introduction to data science and how to execute a machine learning project

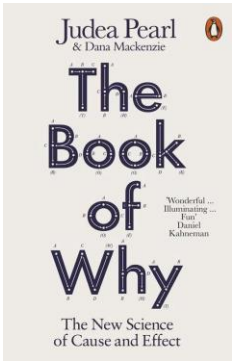
Agenda for today

- Background
- Definition of data science and machine learning
- Executing a data science project with CRISP-DM: Business Understanding

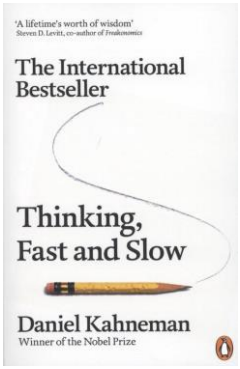
Introduction to data science and machine learning

Background

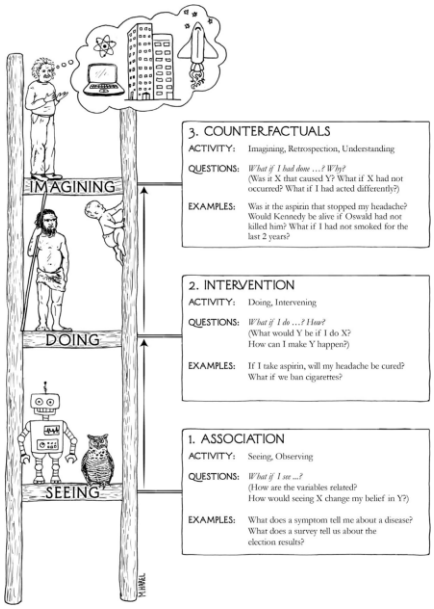
If I were to recommend to books to get a better perspective on data science and machine learning



- What you can and cannot do with data
- Helps you understand difference between correlation and causation



- How humans think and (not use) data to make decisions
- How framing the same data can lead to different preferences



Ladder of causality

- Judea Pearl
  the Book of Why (2019)
- "Data do not understand causes and effects. Humans do."
- Machine Learning sits on the first rung of the ladder

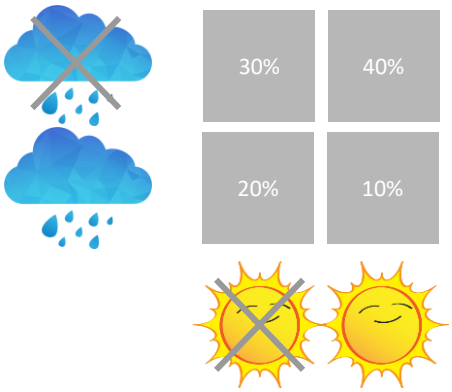
First rung: association a.k.a. correlation

- ‘What if I see …’
- Based on statistics and probabilities
  - Traditional statistics (frequentist approach)
  - Bayesian statistics as basis for calculating conditional probabilities

```
P(sunny and dry)
= P(sunny | dry) * P(dry)
= P(dry | sunny) * P(sunny)
```


Conditional probabilities

- **Given:**
  - P(dry) = 7 / 10
  - P(sunny) = 5 / 10
- **Conditional probabilities:**
  - P(sunny | dry) = 4 / 7
  - P(dry | sunny) = 4 / 5
- **How to calculate P(sunny and dry):**
  - P(sunny | dry) * P(dry) = (4 / 7) * (7 / 10) = 40%
  - P(dry | sunny) * P(sunny)= (4 / 5) * (5 / 10) = 40%



Thomas Bayes (1701 - 1761) and his theorem

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{B}|A) \cdot P(A)}{P(B)}$$

Probability of B occurring given evidence A has already occurred

Probability of A occurring

Probability of A occurring given evidence B has already occurred

Probability of B occurring

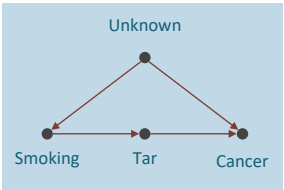
https://en.wikipedia.org/wiki/Thomas_Bayes, https://en.wikipedia.org/wiki/Bayes'_theorem

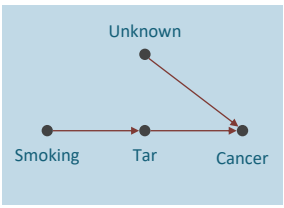
Second rung: intervention

- 'What if I do …'
- Based on causal models

```
P(Y | do(X))
```

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





Third rung: counterfactuals

- 'What if I had …'
- Concept of **potential outcomes** to address fundamental problem that we can only observe one reality
- Causal effect is taken as average difference of all potential outcomes for all individuals

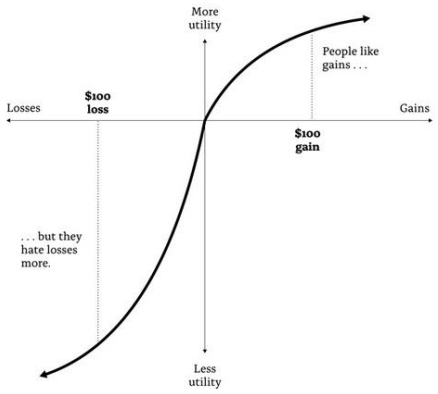
Two possible outcomes

- Outcome if treated: D^1
- Outcome if untreated: D^0

Causal effect

- Individual: $D^1 - D^0$
- Average: $E(D^1 - D^0)$

Framing: perhaps one of the most important concepts for data science and government



- People do not respond 'linearly' to gains and losses
- Depending on how you frame the question (as a loss or gain), people will choose differently

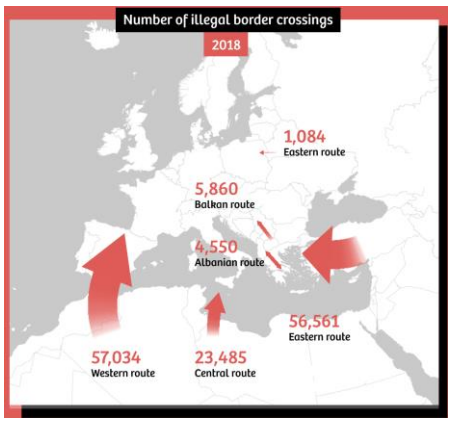
Source: Dorien Julien, All Frames Created Are Not Identical

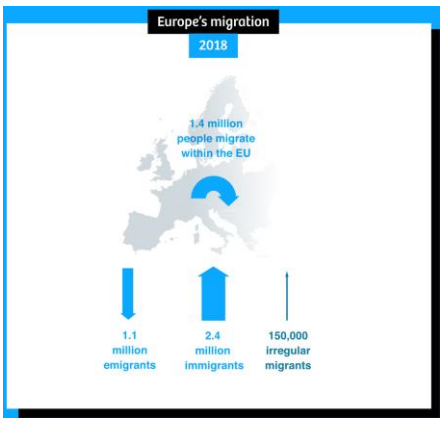
Say we are going to treat 600 people

Framing	Treatment A	Treatment B	
positive	“Saves 200 lives”	"A 33% chance of saving all 600 people, 66% possibility of saving no one."	72% chooses A
negative	“400 people will die”	"A 33% chance that no people will die, 66% probability that all 600 will die."	22% chooses A

https://en.wikipedia.org/wiki/Framing_effect_(psychology)

Framing is everywhere, especially in the media





The Correspondent, How maps in the media make us more negative about migrants

Four villains of decision making

1. **Narrow Framing:** "… the tendency to define our choices too narrowly, to see them in binary terms. We ask, "Should I break up with my partner or not?" instead of "What are the ways I could make this relationship better?"
2. **Confirmation Bias:** "When people have the opportunity to collect information from the world, they are more likely to select information that supports their pre-existing attitudes, beliefs, and actions." We pretend we want the truth, yet all we really want is reassurance.
3. **Short-term Emotion:** "When we've got a difficult decision to make, our feelings churn. We replay the same arguments in our head. We agonize about our circumstances. We change our minds from day to day. If our decision was represented on a spreadsheet, none of the numbers would be changing—there's no new information being added—but it doesn't feel that way in our heads."
4. **Overconfidence:** "People think they know more than they do about how the future will unfold."

https://fs.blog/2013/03/how-to-make-better-choices-in-life-and-work/

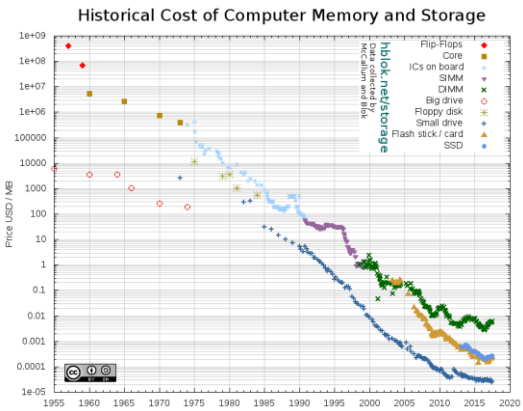
Introduction to Data Science

What is Data Science and Machine Learning?

Why is machine learning booming today?

1. Volume of data available in digital form
2. Widespread availability of cloud-based digital platforms
3. Computing capacity
4. Improvement in algorithms and mathematical models

The digital deluge



1982: ZX Spectrum - 16 KB RAM

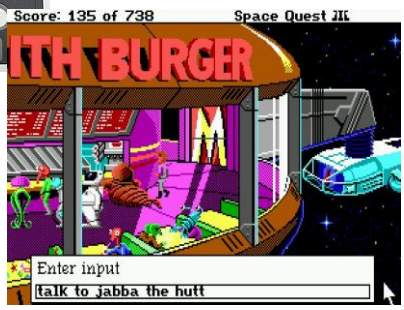
1982: cassette tapes
60KB per 15 minutes

https://hblok.net/blog/posts/2017/12/17/historical-cost-of-computer-memory-and-storage-4/

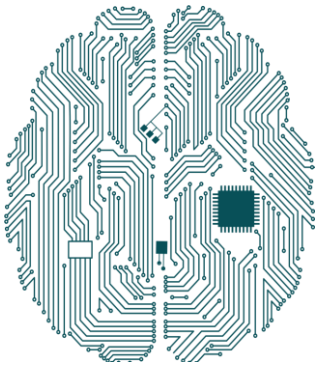
The digital deluge

1991 - x386 PC

1.2 MB on 5 ¼ floppy disk
4 MB of RAM



Data science, AI, machine learning?



- natural language processing
- knowledge representation
- automated reasoning
- machine learning
- computer vision
- robotics

Norvig & Russell: Artificial Intelligence: A Modern Approach (fourth edition 2020)

Types of algorithms

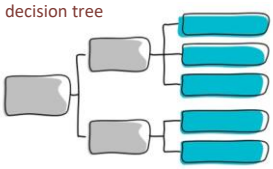
Type	Examples
Time-series forecasting	• Very old, has been around since 1800 • Healthcare, econometrics, astronomy, meteorology
Optimisation & Operations Research	• Optimizing tactics in World War II • Logistics, business process improvement
Simulation	• Markov Chain Monte Carlo method (Manhattan project) • Applied for many problems that we can't solve analytically
Rule-based (if-then-else)	• Used since 1970s with invention of first computers • Expert systems in financial sector (mortgages)
'machine learning'	• Many algorithms invented since 1980s • Mainstream applications in organizations since 2000s

The essence of machine learning: From 2 to 2,000 parameters

Principle		Machine Learning
function	$y = f(x)$	$Y = F(X)$
	$y = ax + b$	matrices
algorithm/ model	linear regression	decision trees, neural networks, support vector machines etc.
parameters	a, b	tens to thousands (hyper)parameters
performance	R^2	Many different information criteria and penalties

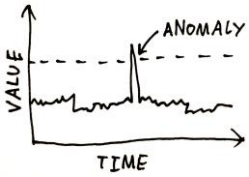
All algorithms need (human) input

decision tree



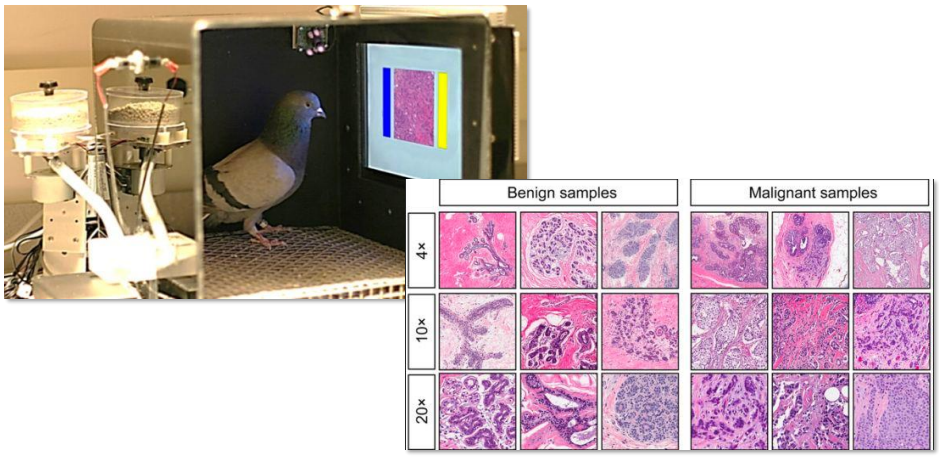
rules-based	**machine learning**
decision rules	training data

time-series



threshold values	auto-correlations

… but how 'smart' is machine learning, really?



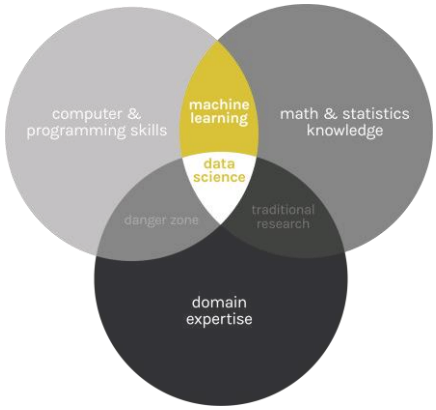
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0141357

A more practical look at data science

> "Data science is a multidisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data."

> Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead."

> Our definition and practical focus:
- multidisciplinary approach
- extracting value from structured and unstructured data
- supporting better decisions



https://en.wikipedia.org/wiki/Data_science, https://en.wikipedia.org/wiki/Machine_learning

The three components of machine learning

Representation	Evaluation	Optimization
Instances - K-nearest neighbor - Support vector machines Hyperplanes - Naive Bayes - Logistic regression Decision trees Sets of rules - Propositional rules - Logic programs Neural networks Graphical models - Bayesian networks - Conditional random field	Accuracy/Error rate Precision and recall Squared error Likelihood Posterior probability Information gain K-L divergence Cost/Utility Margin	Combinatorial optimization - Greedy search - Beam search - Branch-and-bound Continuous optimization Unconstrained - Gradient descent - Conjugate gradient - Quasi-Newton methods Constrained - Linear programming - Quadratic programming

Pedro Domingos, A Few Useful Things To Know About Machine Learning (2012)

Branches of machine learning

Supervised Machine Learning
when the outcome Y is known and machine learning is used to relate input variables to this known outcome Y

Unsupervised Machine Learning
where there is no outcome Y (or it is not used in the analysis), and machine learning is used to search for clusters of comparable instances

Reinforcement Learning
where machine learning is concerned with agency, e.g. learning to take actions depending on the 'state of the world' such that some kind of reward is optimized

for an interactive 'tree' of machine learning, see https://kumu.io/jads/tree-of-machine-learning-algorithms

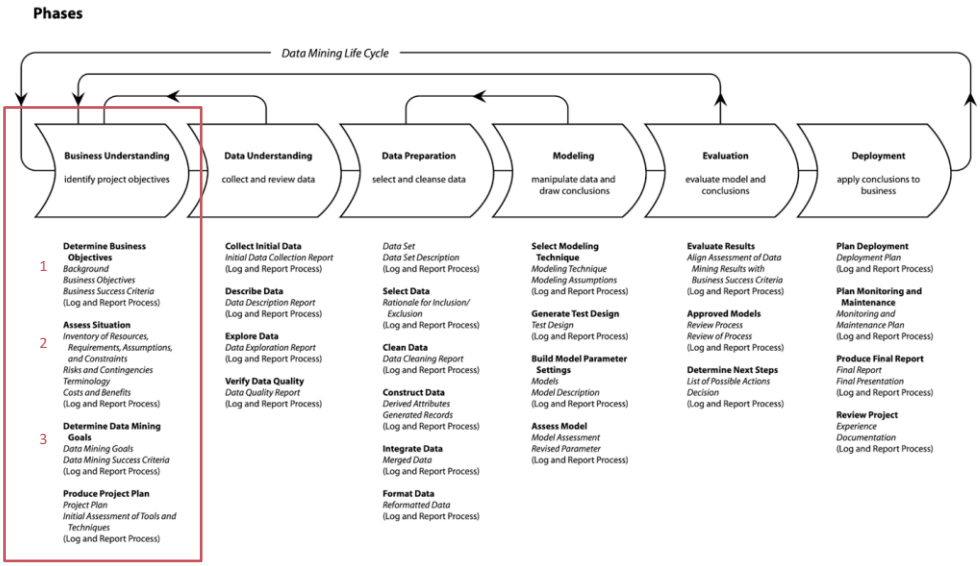
27

Introduction to Data Science

Executing Data Science Projects with CRISP-DM: Business Understanding

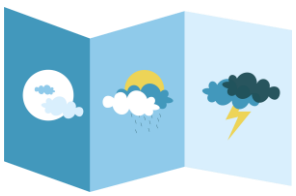
28

Executing a data science project with CRISP-DM



Be aware of difference between prediction and effect problems

Prediction problem

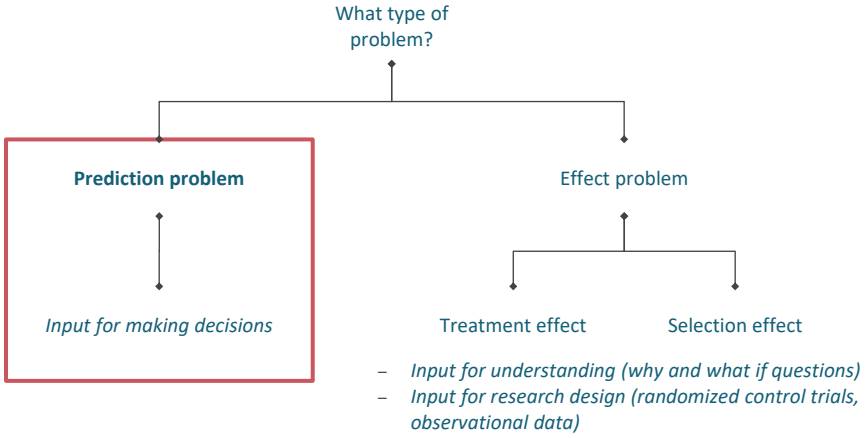


How is the world going to present itself (so that we can adequately *react* to it)?

Effect problem

How can we *change* a certain outcome, behaviour? (so that we can create the best outcome)

Machine learning can *only* be used for prediction problems



- *Input for understanding (why and what if questions)*
- *Input for research design (randomized control trials, observational data)*

Prediction vs effect problems

Can the 'thing' you are interested in be found literally in the available data? (prediction/classification) … or is it actually a change with regard to the available data? (effect)

- How to make employees more productive?
- How to raise revenue?
- How to reduce costs?
- Which clients are expected to return the next year?
- When do we expect equipment to break down?

1. Determine business objectives

- In your first project: start simple!
- Solve an actual problem (not just 'generating insights') , e.g. where can new insights improve the decision making process
- Potential of the project:
  - Large improvement potential (an undesirable status quo)
  - Impacting a large number of clients, patients, etc
  - High impact per client, patient, etc.
- Very clear definition of Y / high quality of Y (the thing you want to predict)
  - Some outcomes are inherently difficult to define (e.g. treatment dropout)
  - Was the outcome accurately measured / determined by experts?

2. Assess situation

- Is the right type of data available for your project?
- Is enough data available and/or is the data rich (wide) enough?
- Is there enough variation in your outcome Y / is Y not too unbalanced?
- Are there any legal or ethical objections regarding the use of the data?
- If necessary, is it possible to combine data files?

3. Determine data science goals

- Are your outcomes actionable?
- Is the definition of Y in line with what the user is interested in?
  - Does it help the user to do a better job?
- In what domain will the project outcomes contribute?
  - Supporting administrative processes?
  - Supporting expert judgement?
  - Etc.

3. Determine data science goals: actionable insights (1 of 2)

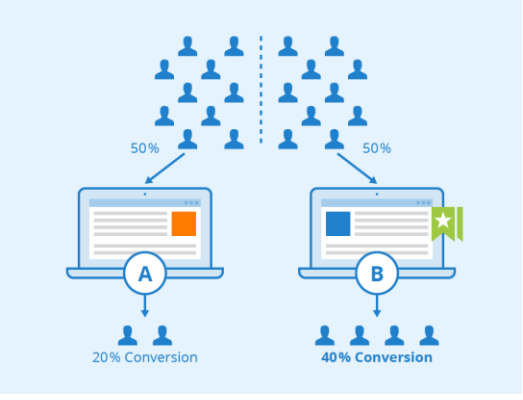
- Examples when it is not obvious how to act on your project outcomes:
  - Knowing the probability that a citizen will require social support
  - Knowing the probability that a client will cancel a subscription
- Examples when it is 'obvious' how to act on your project outcomes:
  - Situations where you know how the world will change when you make different decisions (due to common sense, or understanding underlying mechanisms of the problem)
  - Situations where effect research is available
  - Situations where 'stepped approaches' are applied
  - Situations where monitoring is already in place, but is improved through better use of data

3. Determine data science goals: actionable insights (2 of 2)

- Balancing between what is practically useful and statistically desirable
  - Two relevant, and sometimes competing, interests
- From a statistical perspective, a project is more manageable when:
  - The outcome is defined into two classes
  - Classes are balanced
  - Lots of data / information is available
- From a practical perspective:
  - At first, there might be an interest in multiple classes (but there might be flexibility in this)
  - Not all available information might be easily obtained for new instances (e.g. clients or transactions)

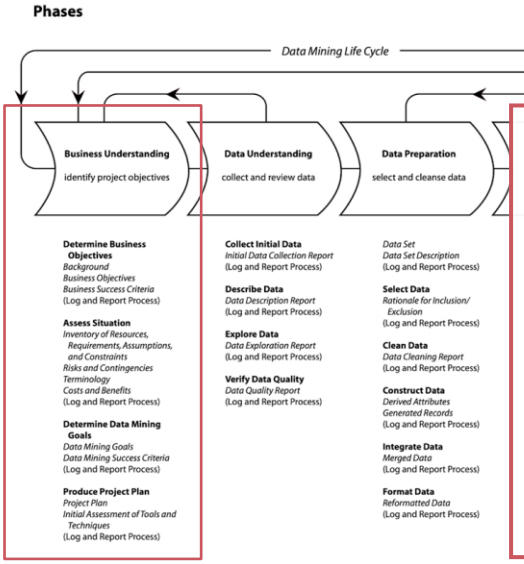
Beyond prediction: experimenting with A/B testing for effect problems

- With websites with enough visitors, A/B testing can be a good option to learn how to improve outcomes by adjusting the process
- In theory, you could assign random treatments in real-world settings (hospital, GP practice) but this often has ethical implications



Source: https://www.seobility.net/en/wiki/AB_Testing

Concluding remarks on Business Understanding



- A project will take a lot of work, so be very careful in the project you select
- Next to considering whether a project is feasible, be very clear as to whether your results can indeed lead to improvement, or whether it will require a follow-up project to make an impact
- In general, carefully consider:
  - Prerequisites (necessary conditions)
  - Project scoping
  - How to use the outcomes of your project