Machine Learning and Programming in Python

Lecture for Master and PhD students

Chair of Data Science in Economics

Ruhr University Bochum

Summer semester 2024

Lecture 1

Outline

- Logistics
- About this module
- Literature
- Overview: Data Science
- Overview: Machine Learning
- Overview: Programming with Python
- Introduction to Python, Anaconda, Jupyter Notebook

Lecture

- Monday, 08.04.2024, 15:15 – 17:45, HNC 20
- Thursday, 11.04.2024, 14:15 – 16:45, HNC 10
- Friday, 12.04.2024, 14:15 – 16:45, HNC 20
- Monday, 15.04.2024, 15:15 – 17:45, HNC 20
- Thursday, 18.04.2024, 14:15 – 16:45, HZO 30
- Friday, 19.04.2024, 14:15 – 16:45, HGD 30
- Monday, 22.04.2024, 15:15 – 17:45, HNC 20
- Thursday, 25.04.2024, 14:15 – 16:45, HZO 30
- Friday, 26.04.2024, 14:15 – 16:45, HGD 30

approx. 5 minutes break in between - note: a break is a break

- Exam: 29.07.2024
- Information on how and when to register for the exam can be found in Moodle and ecampus/ FlexNow $\Rightarrow$ Note: You will only be able to register for and take part in the exam, if you were formally accepted by me to take part in this course
- Questions about the course, Machine Learning, Programming in Python? Email: dsecon-itc@ruhr-uni-bochum.de
- The slides will be uploaded before the lecture starts
- This is the second time that this lecture is given

Contents:

- 1. Introduction
- 2. Programming in Python (basics, control flow, classes)
- 3. Model complexity (underfitting, overfitting, bias-variance trade-off)
- 4. Regularisation (regression, Ridge, Lasso), Scikit-learn in Python
- 5. Supervised learning (regression, classification, logistic regression, support vector machines), Scikit-learn
- 6. Model evaluation, Scikit-learn
- 7. Decision trees (bagging, random forests, boosting algorithms), Scikit-learn
- 8. Unsupervised learning (k-means clustering, principal component analysis), Scikit-learn
- 9. Deep learning, neural networks, PyTorch, TensorFlow in Python
- 10. Natural language processing (text as data, pre-processing, dictionary methods, supervised learning, naïve Bayes), NLTK in Python

- This is a compact course on Machine Learning and Programming in Python
- You will learn a lot about concepts of Machine Learning, why it is important in the field of Economics, and how to apply the techniques on data sets, using the programming language Python

Some of the data we will be working with:

- a speech by Christine Lagarde (ECB)
- data on beer consumption
- house price data
- Titanic data set
- data on CO2 emissions and economic growth
- a speech by George W. Bush
- hotel reviews

- Prerequisites:
  - Knowledge of Macroeconomics, Microeconomics, Mathematics/ Statistics for Economists
  - Beginner in Python
  - Some previous knowledge of Stata/ R
- Assessment
- Important for the exam: Statistical theory and applications in Python, lecture slides, relevant pages from the textbooks, further material (Jupyter Notebooks, problem sets)
- I like to explain/ visualize concepts on the blackboard ("Tafelbilder") to convey the intuition behind the methods

Learning Objectives:

- Getting to know models and methods from machine learning
- Being able to develop models of machine learning
- Analysis of numerical and text data
- Application of the techniques with the use of the programming language Python
- Being able to understand applications of machine learning and NLP (Natural Language Processing) in Economics

- Programming in Python critical part of the module but this is not a purely programming module!

- Course is on Machine Learning AND Economics, not merely on machine learning.
  - Learn tools & techniques to assist you in empirical economics.
  - Enable you to converse intelligently with data scientists.
  - Encourage you to explore innovative applications in economics.
  - Equip you for a future that will increasingly be driven by machine learning & AI!

Literature:

- James, Witten, Hastie, Tibshirani, Taylor (2023), An Introduction to Statistical Learning, download at
  https : *//hastie.su.domains/ISLP/ISLP_website.pdf*

- Hastie, Tibshirani, Friedman (2017), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Series in Statistics, 2. edition, download at
  https : *//hastie.su.domains/Papers/ESLII.pdf*

- Jurafsky, Martin (2021), Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 3. edition, download at *https* : *//web.stanford.edu/* ∼ *jurafsky/slp*3

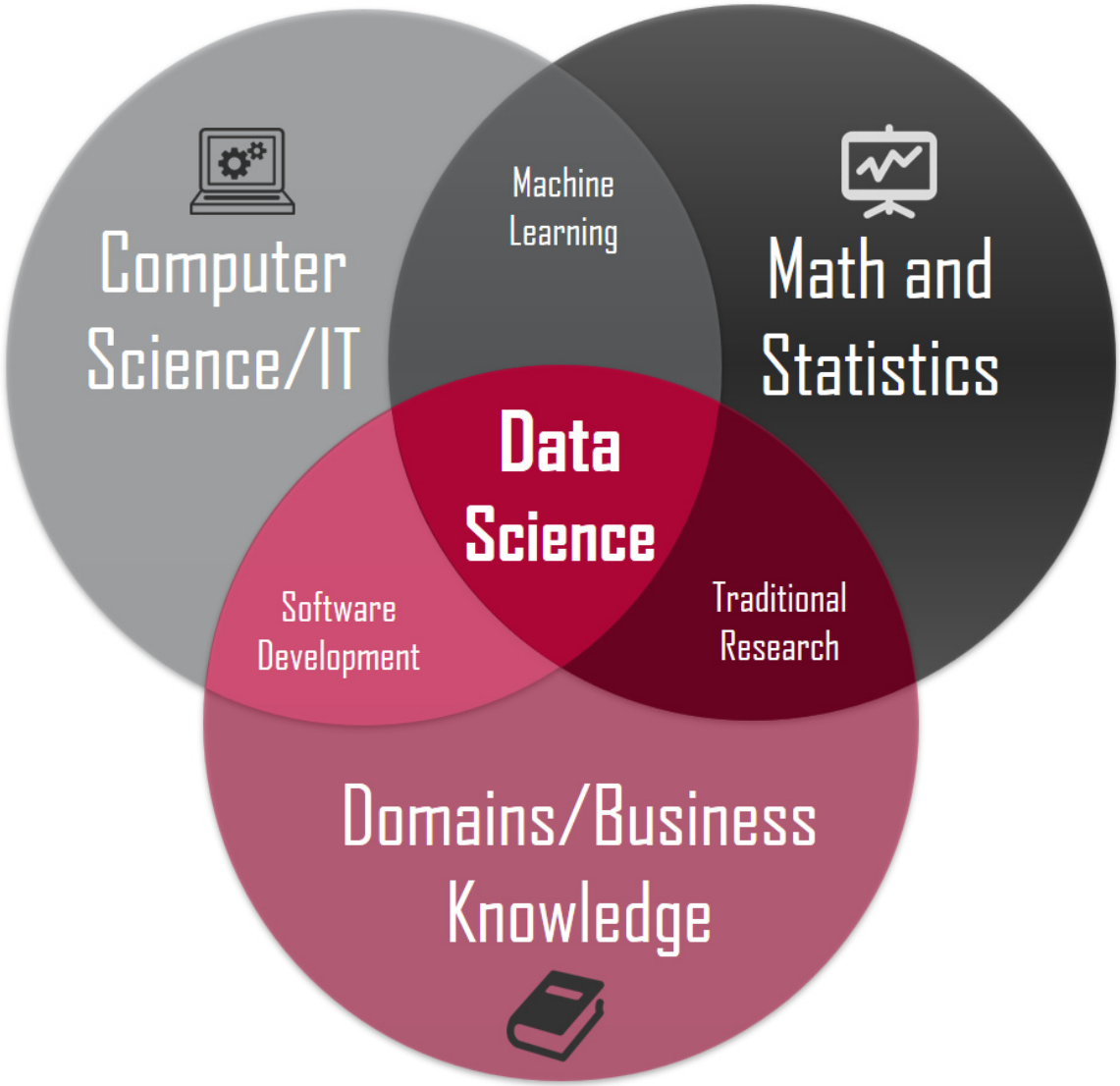
- Grimmer, Roberts, Stewart (2022), Text as Data: A New Framework for Machine Learning and the Social Sciences

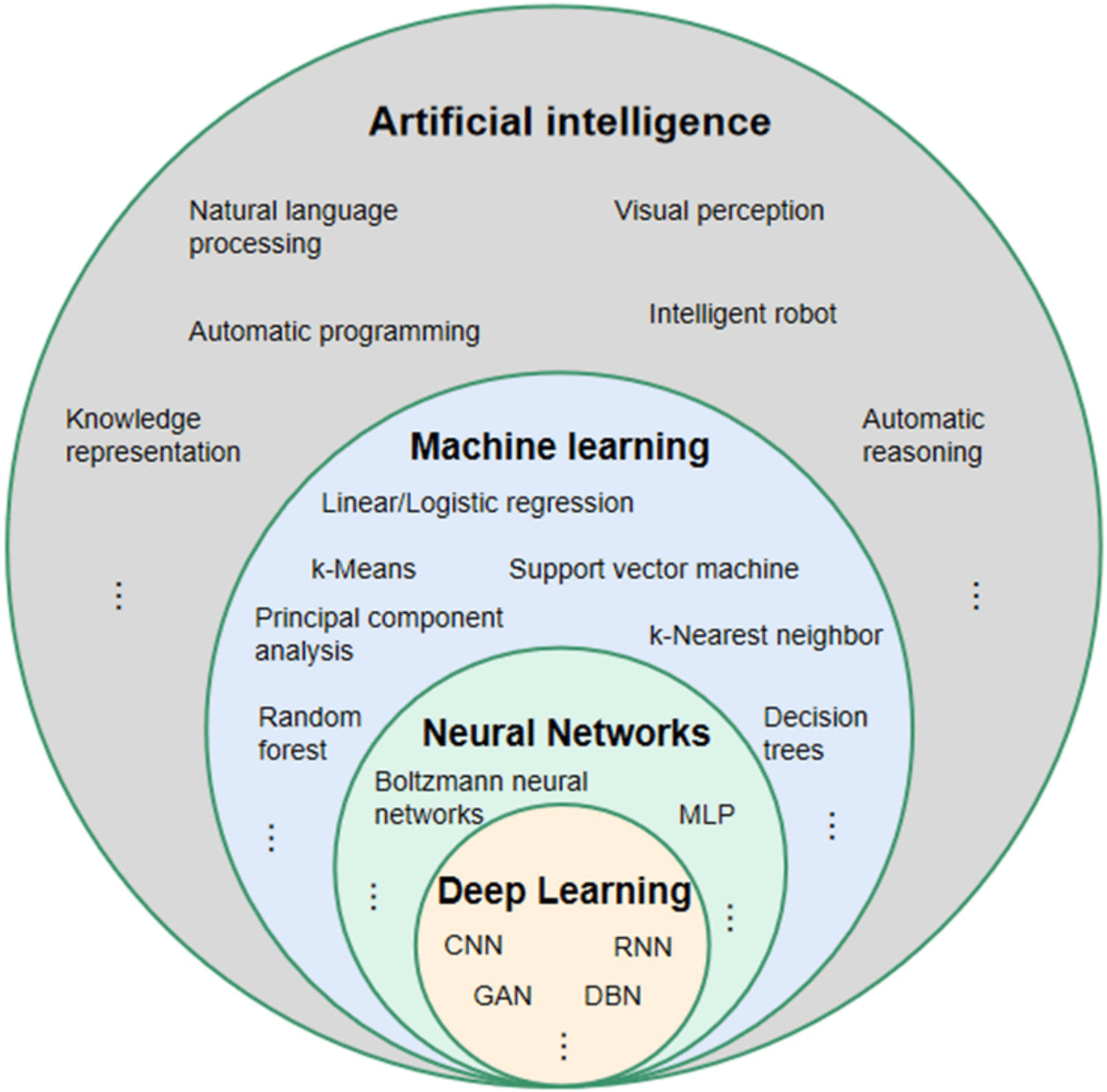
Additional:

- Deitel, P., Deitel, H. (2020), Intro to Python for Computer Science and Data Science, Pearson

- Mueller, Guido (2016), Introduction to Machine Learning with Python, O'Reilly

What is Data Science?


Computer Science/IT
Machine Learning
Math and Statistics
Data Science
Software Development
Traditional Research
Domains/Business Knowledge

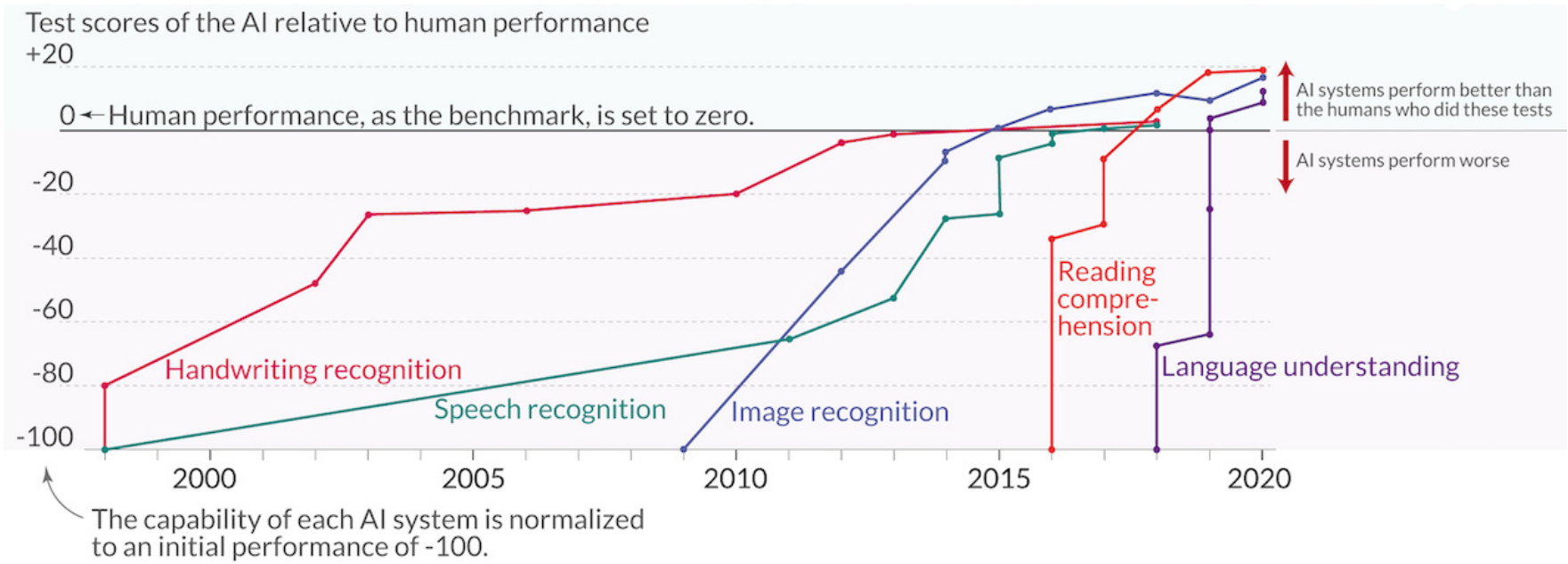

Artificial intelligence
Natural language processing
Visual perception
Automatic programming
Intelligent robot
Knowledge representation
Automatic reasoning
Machine learning
Linear/Logistic regression
k-Means
Support vector machine
Principal component analysis
k-Nearest neighbor
Random forest
Decision trees
Neural Networks
Boltzmann neural networks
MLP
Deep Learning
CNN
RNN
GAN
DBN

What is Artificial Intelligence?

- Marvin Minsky: "Artificial Intelligence is the science of making machines do things that would require intelligence if done by men"

- Stuart Russell and Peter Norvig: "the designing and building of intelligent agents that receive percepts from the environment and take actions that affect that environment"

Language and image recognition capabilities of AI systems have improved rapidly

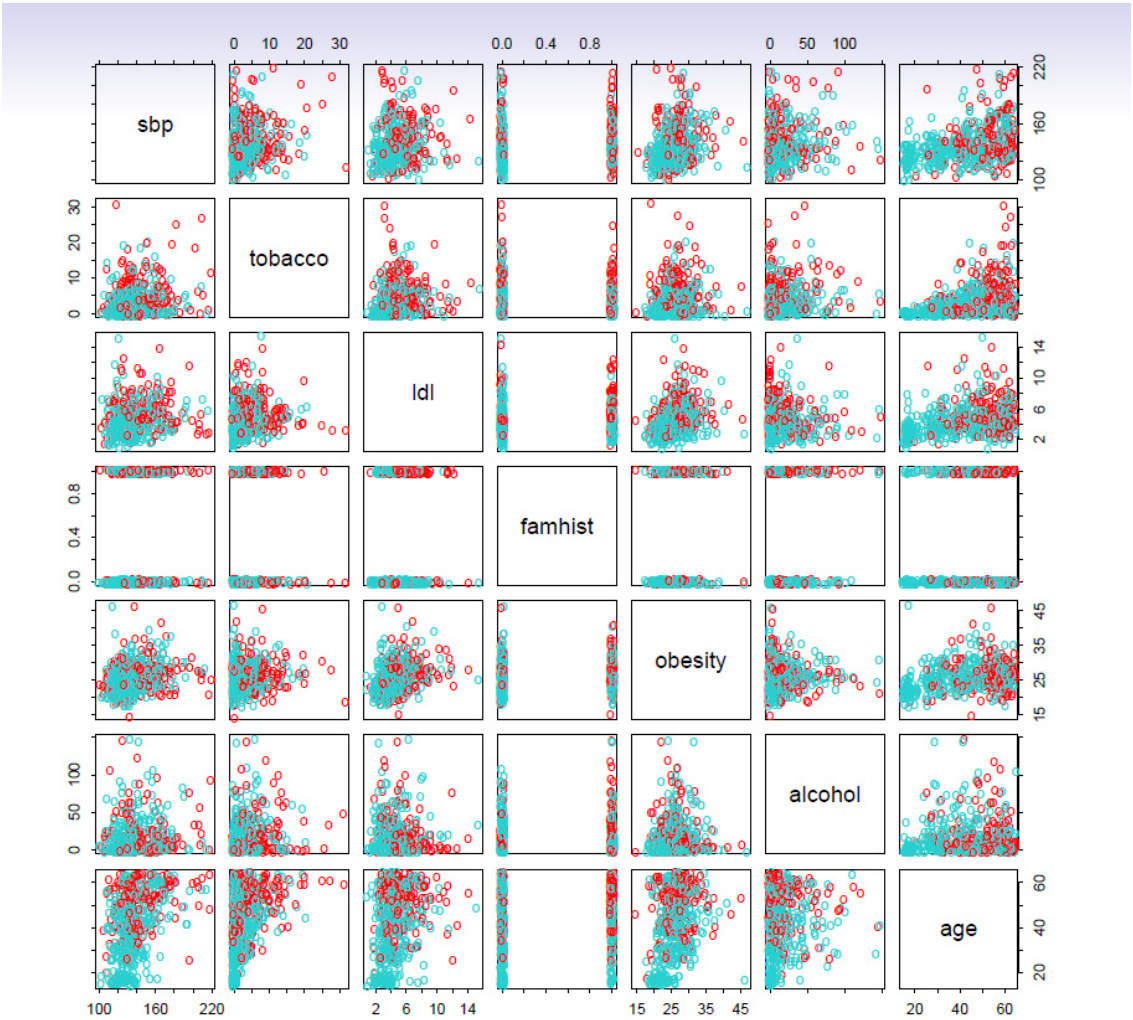


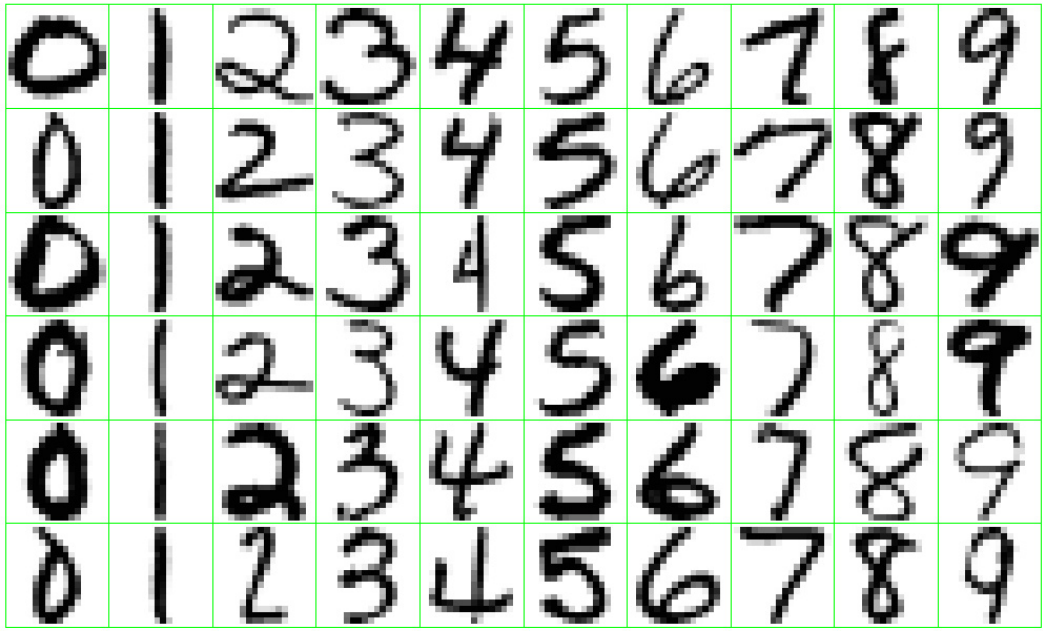
Data source: Kiela et al. (2021) – Dynabench: Rethinking Benchmarking in NLP

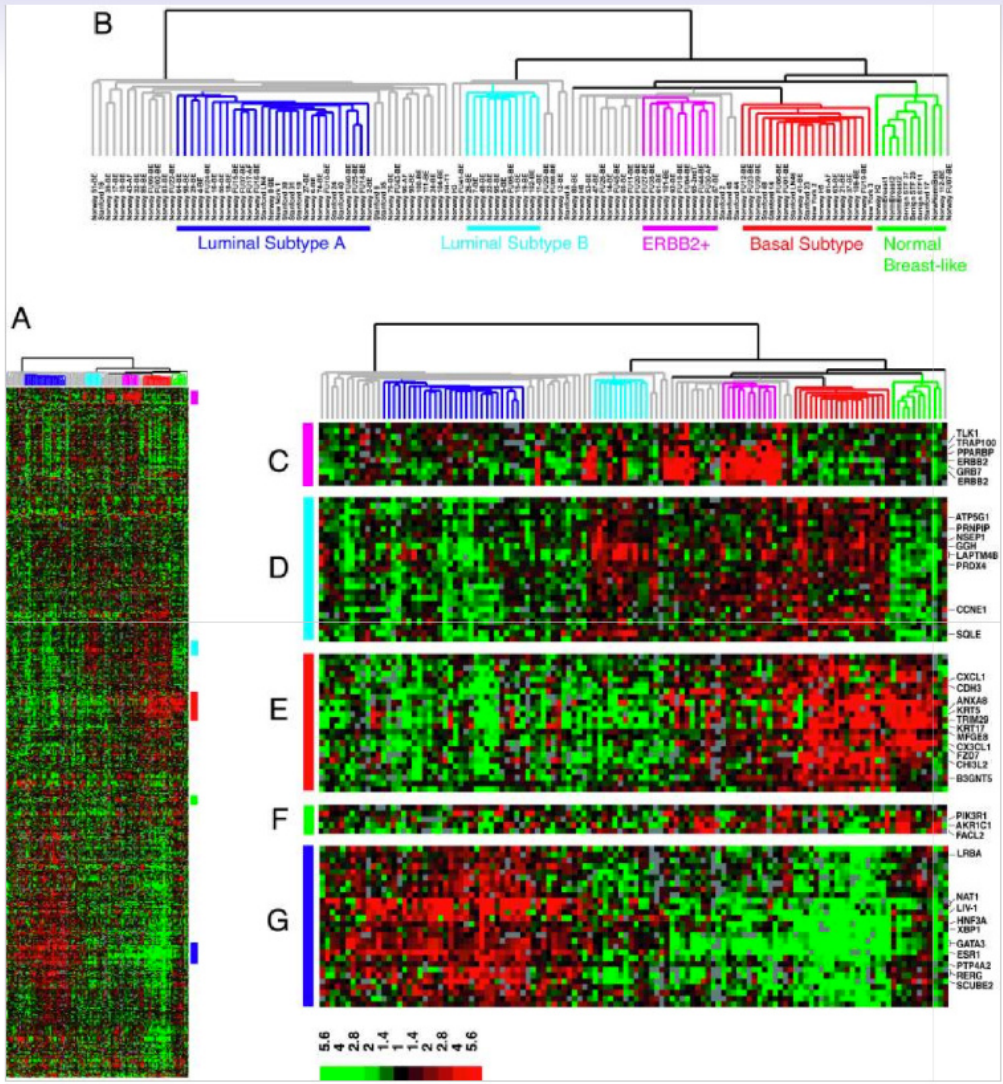
Some problems you may solve with the methods of Data Science, AI, Machine Learning:

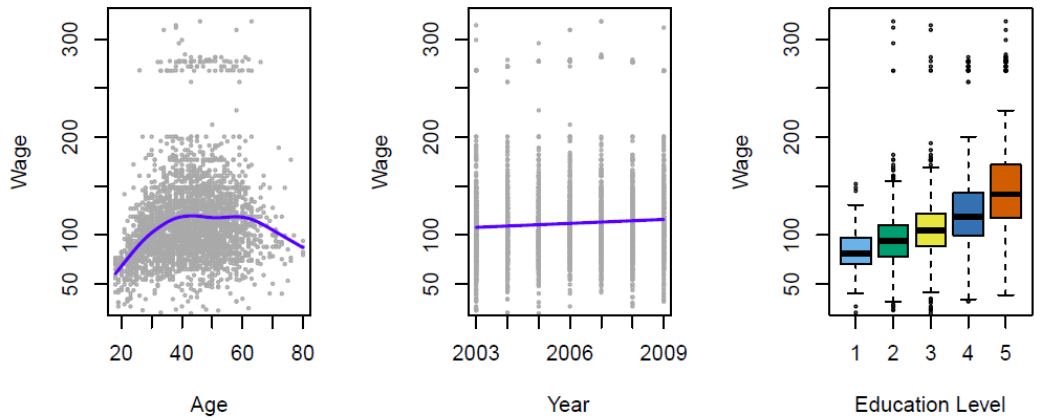
- Identify the risk factors for prostate cancer
- Classify a customer to default on a credit
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements
- Customize an email spam detection system
- Identify the numbers in a handwritten zip code
- Group a tissue sample into one of several cancer groups, based on a gene expression profile
- Establish the relationship between salary and demographic variables in population survey data
- Classify the pixels in a LANDSAT image, by usage


lpsa
lcavol
lweight
age
lbph
svi
lcp
gleason

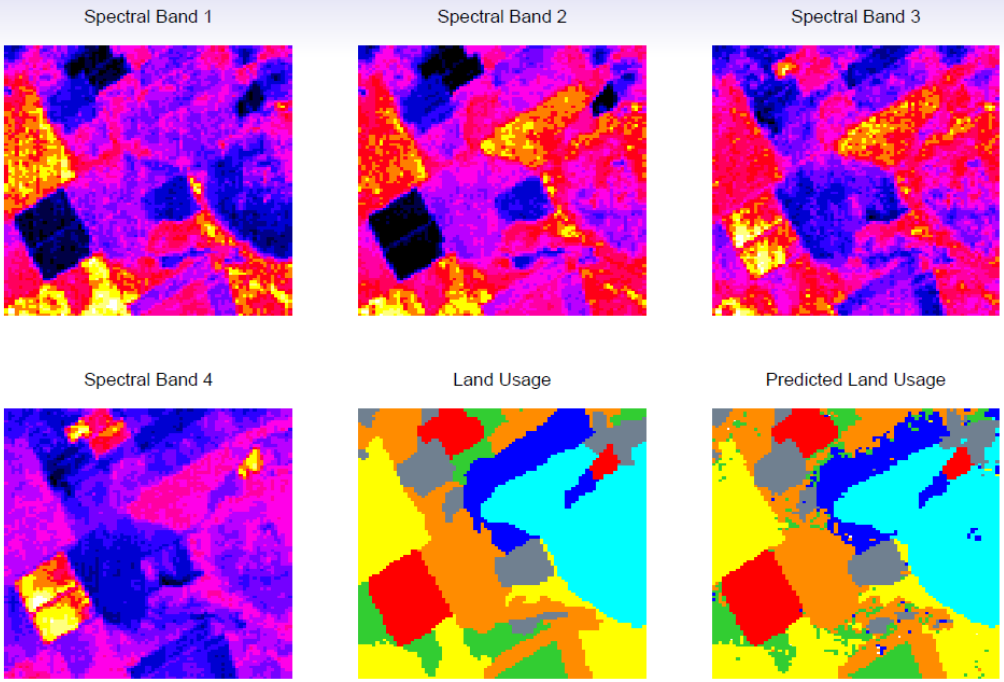

sbp
tobacco
ldl
famhist
obesity
alcohol
age




B
Luminal Subtype A
Luminal Subtype B
ERBB2+
Basal Subtype
Normal Breast-like
A
C
D
E
F
G
TLK1
TRAP100
PPARBP
ERBB2
GRB7
ERBB2
ATP5G1
PRNPIP
NSEP1
GDH
LAPTM4B
PRDX4
CCNE1
SQLE
CXCL1
CDH3
ANXA8
KRT5
TRIM29
KRT17
MFGE8
CX3CL1
FZD7
CH13L2
B3GNT5
PIK3R1
AKR1C1
PACL2
LRBA
NAT1
LIV-1
HNF3A
XBP1
GATA3
ESR1
PTP4A2
RERG
SCUBE2
5.6
4
2.8
2
1.4
1
1.4
2
2.8
4
5.6



Income survey data for males from the central Atlantic region of the USA in 2009.



Usage $\in$ {*red soil, cotton, vegetation stubble, mixture, gray soil, damp gray soil*}

Traditionally, economists used (relatively small-sized) data sets from:

- surveys (households, firms, ...)
- experiments (field, lab, ...)
- administrative records (e.g. taxes)
- the census, etc.

Focus has been on causal inference

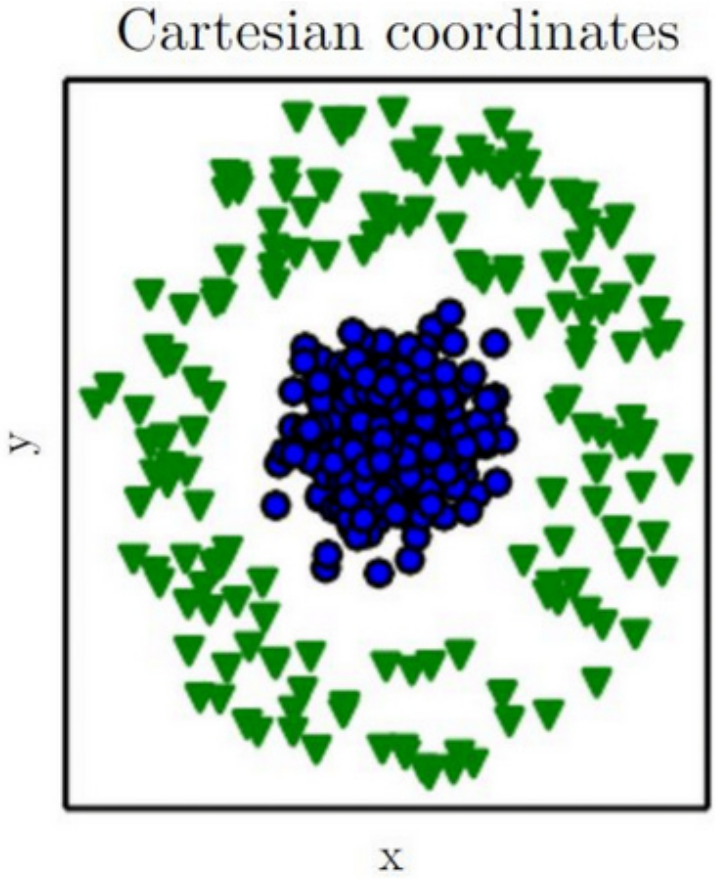
Big Data

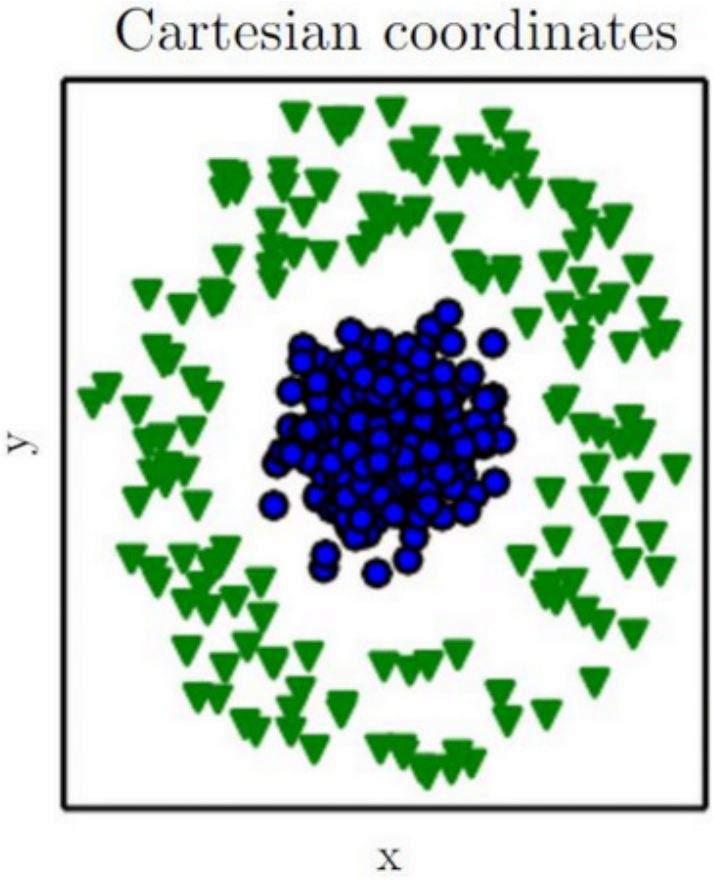
- Volume: Datasets orders of magnitude larger e.g. retail scans, debit/credit card usage, social media posts, web clicks (several GB or TB large)
- Real time: Data streaming in real time - useful for business and policy to analyse and respond quickly
- Variety: Numeric or structured text (standard), unstructured text, images, video, browsing behaviour etc.
- Technology: Feasible to gather, store, access, and manipulate vast datasets

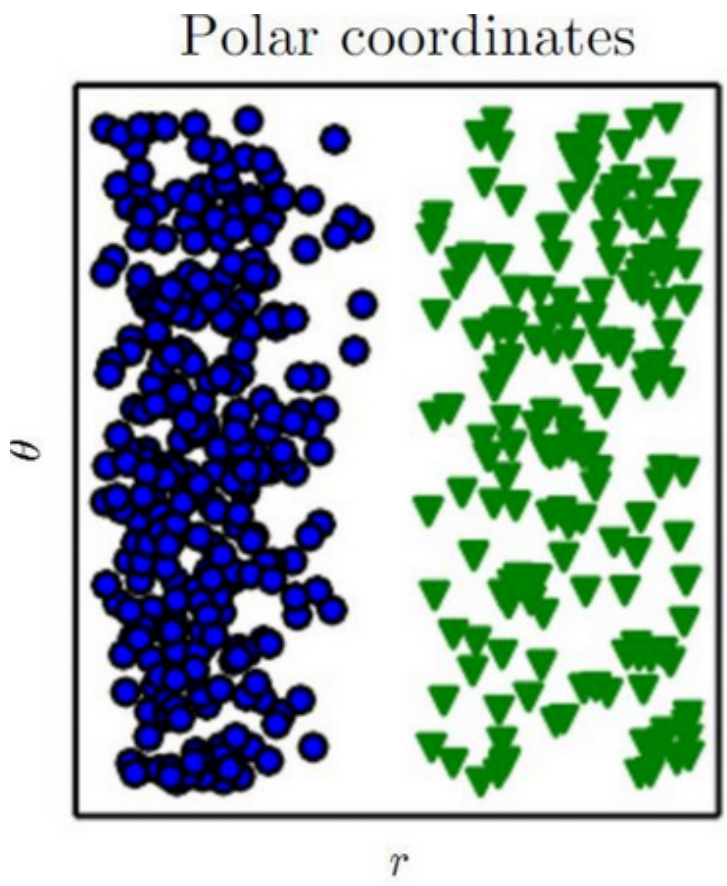
- Big is not the critical thing - what matters is extracting information from data (new, unstructured data)

- Machine learning: a field that develops algorithms designed to be applied to datasets, with the main areas of focus being prediction -such as regression and classification- and clustering or grouping tasks. (Athey (2017), "The Impact of Machine Learning on Economics")

- For the application of such algorithms, data representation plays a crucial role
- Suppose that you want to draw a line that separate the blue circles and the green triangles







- Impossible under cartesian coordinates, trivial under polar coordinates
- Thus, Data Science also requires creativity in terms of data representation

So: What is Data Science?

- Different definitions
- It is about how we manage the existing –and growing- information, in a very interconnected world
- Also, it has to do with understanding our environment: Society, economy, consumption habits, machinery functioning...
- Data is around us: E-mails, social networks, surveys, banking details. . . .

- $\Rightarrow$ Data science has to do with the multiple methodologies conceived to manage the existing information in a changing world, using the data to extract plausible and reasonable conclusions

We use data for many purposes:

- It allows us to describe situations in multiple frameworks (e.g.: the average productivity of workers in a particular sector; social security affiliations, etc.).
- Detect anomalous events: By using past information –creating trends, growth rates. . .-, we can interpret information and classify it. (e.g.: sharp declines of stock prices; suspect logins in bank accounts, etc.).
- Ex-post evaluations: Understanding the causes and the roots of many events (e.g.: The crash of 2007/2008).

- Data collection has become in a powerful weapon for anyone interested in a field of expertise.
- Your personal information is constantly being recorded, saved, used and interpreted. And this, sometimes, is scary.
- Imagine that you buy a car. What kind of information are you going to provide to the brand in question and how this will be used?

How does Data Science work?

How does Data Science work?

- 1. We download/digitise the data
- 2. We clean the information, check for typos, remove columns, include rows (or vice versa), cluster the individuals, identify strings, floats etc., double-check the absence of entry errors, organisational mistakes, missing data...
- 3. Descriptive analysis, visualisation
- 4. Analysis, prediction, inference, evaluation.

The Workflow in Data Science

- 1. We need to well-define our question.
  - How is my heart rate when I do running? How many financial transfers in Madrid are fraudulent? What is the probability of having a car accident in the French coast during summer?
- 2. Some initial data allowing us to reach preliminary conclusions.
  - You can use some past recordings of your heart frequency / Banking transfer records / Average number of car accidents in French Riviera during 2018-19.
- 3. Upcoming new sets of data, so we can use new algorithms or run stochastic processes.
  - New data will allow you to establish comparisons ex-ante vs ex-post.

Who works in Data Science?

- Everyone with interest in understanding complex processes involving information generation.
- "Traditionally", four types of jobs have been identified:
  - Data engineer: Work on the architecture of information and build the "structure" to store and keep all the data generated.
  - Data analyst: Economists could be placed here. Clean, summarise and interpret the data.
  - Data scientist: Also a job for economists, with strong statistical skills though. They make experiments and run codes to set up automatised inference process.
  - Machine learning scientist: They focus on the last part of the analytic process. They are interested in forecasting with large datasets, so create complex machine learning algorithms to do this.

- Nowadays, Data Science is a discipline that can be learned by any profile: It is used by economists, sociologists, historians, engineers, physicians, etc.

- The volume of information is growing in every field, so that's what explains the boom: both private and public sector must understand and interpret real world.

What is Machine Learning?

Machine Learning:

- Focus on prediction.
- May (or may not) be concerned with insights or causal inference.
- Systematic and transparent model selection.
- Can handle large numbers of covariates.

Econometrics:

- Look for patterns and insight.
- Causal inference is usually a major goal.
- Estimate one model and focus on significance/ confidence intervals of key parameters.

Terminology

- Features: Predictors, inputs, regressors, covariates, independent variables (X)
- Output: Outcome, response, target, label, class, dependent variable (y)
- Model: Learner, classifier, regression equation; e.g some function of the features that predicts outcome
- Training Data: Sample used for building the model
- Test Data: Sample used for testing the model (out of sample)
- Loss Function: Cost function, fit (usually the mean squared error MSE (error or residual sum of squares RSS))

Statistics and Machine Learning - Differences in Terminology

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

Comparision of Machine Learning and Statistics, Glossary by R. Tibshirani

The Supervised Learning Problem

- Outcome measurement Y (also called dependent variable, response, target)
- Vector of p predictor measurements X (also called inputs, regressors, covariates, features, independent variables)
- In the regression problem, Y is quantitative (e.g price, blood pressure)
- In the classification problem, Y is qualitative (survived/died, digit 0-9, cancer class of tissue sample)
- We have training data $(x_1, y_1), \ldots, (x_N, y_N)$. These are observations (examples, instances) of these measurements

Objectives

On the basis of the training data we would like to:

- Make accurate predictions
- Understand which inputs affect the outcome, and how

With the test data we will:

- Assess the quality of our predictions and inferences

Supervised Learning

- $\Rightarrow$ Learn a model from labeled training data to make predictions in unseen or future data.

- Say, goal is to learn a model to detect spam email.
- Training data is a sample of emails labeled (by humans) as spam or not.
- Classification task: label is a category (binary or multiple).
- Regression task: label is a continuous variable e.g. house price, stock value etc.

- Goal is to predict y given k features, $X = [x_1, x_2, x_3, ..., x_k]$
- Posit that $y = f(X) + \epsilon$
  - ϵ is random noise in data (cannot be eliminated).
  - f is unknown but we would like to approximate it as well as possible.
  - Different ML algorithms take different approaches to 'learning' f
- Say we train a model $\hat{f}$. We predict $\hat{y} = \hat{f}(X)$.
- How good an approximation is $\hat{f}$? Are $\hat{y}$ close to the true y?
  - Loss (or cost) function: $J(y, \hat{y})$
  - A common loss function is $J = \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- ML picks $\hat{f}$ to minimise the loss function.
- **Minimisation** problem solved using **Gradient descent**, an optimisation algorithm.

Unsupervised Learning

- No outcome variable, just a set of predictors (features) measured on a set of samples
- objective is more fuzzy — find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation
- difficult to know how well your are doing
- different from supervised learning, but can be useful as a pre-processing step for supervised learning

Unsupervised Learning

- $\Rightarrow$ Learn a model to explore unlabeled data and discover hidden patterns.
- Say, goal for a firm is to identify distinct customer groups (segments).
- Members within a segment have similar tastes and spending.
- Clustering.

Machine Learning workflow:

- Prep the raw data:
  - Quantify the various features (sometimes normalised to range [0,1] or standard normal).
  - Split into training and test samples; set aside the test sample.
- Train models
  - Training data often split into training and validation subsamples; allows us to evaluate how models generalise before final evaluation on the test sample.
  - Estimate models using training data.
- Evaluate on the test data and select a model.
  - Pick a performance metric e.g. classification accuracy.
  - Select the best performing model.

Ethical questions: ML as substitute for human decisions?

Example: Borrower, X variables (age, education, income), y variable (credit default)

- Some types of data might not be available to ML algorithm. $\rightarrow$ behaviour of borrower
- Available data is generated by human activity and is one-sided $\rightarrow$ Only observe flight or crime among bailed defendants $\rightarrow$ Don't have 'labels' for the jailed defendants
- Humans might have different (or more complex) payoff functions $\rightarrow$ loan officers might not be lending to most profitable customers $\rightarrow$ focussed on short-run default instead of long-run outcomes $\rightarrow$ ML predictions not met

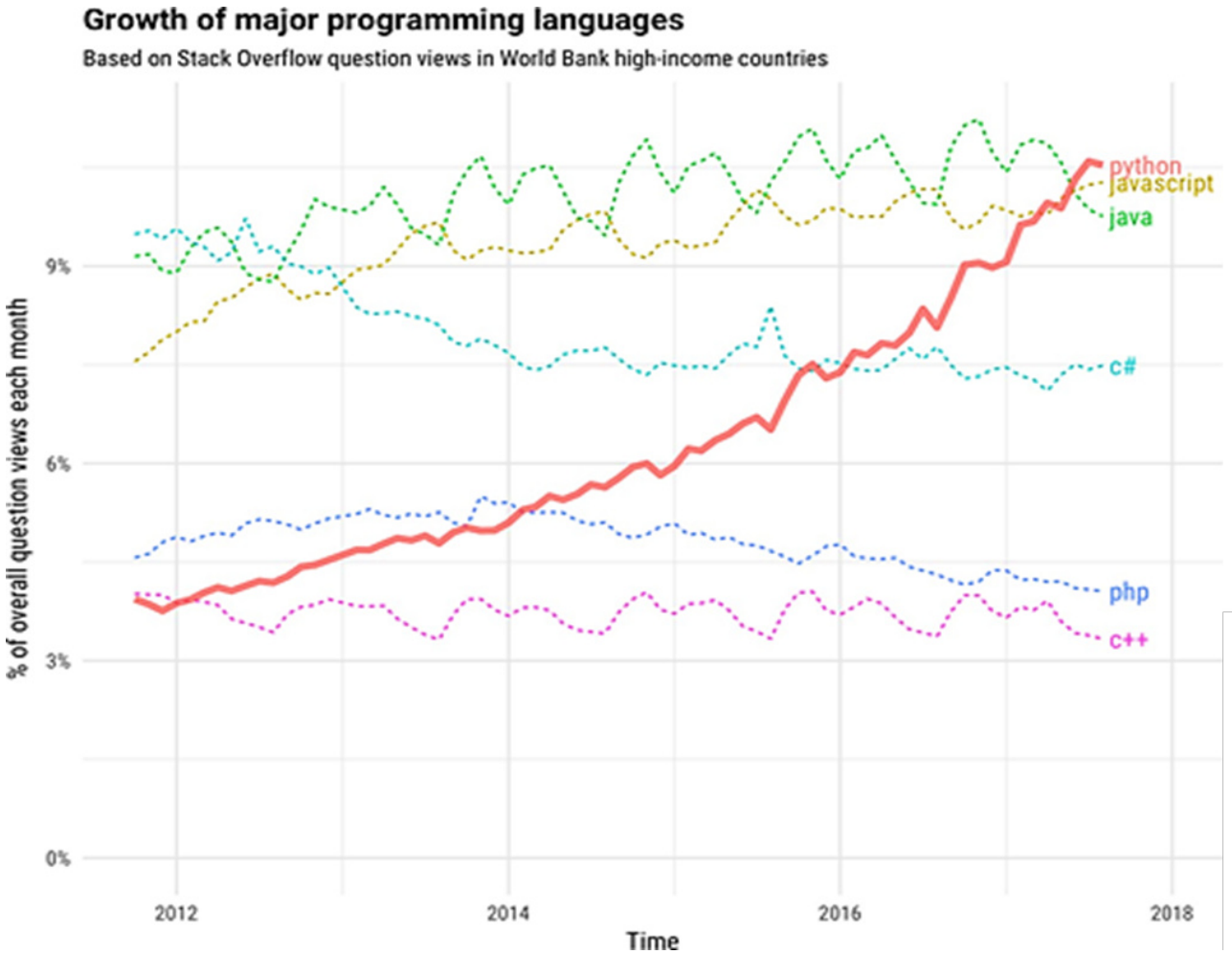
Philosophy of Machine Learning

- It is important to understand the ideas behind the various techniques, in order to know how and when to use them
- One has to understand the simpler methods first, in order to grasp the more sophisticated ones
- It is important to accurately assess the performance of a method, to know how well or how badly it is working [simpler methods often perform as well as fancier ones!]

- This is an exciting research area, having important applications in science, industry and finance

What is Python?

- A very popular programming language.
- Highly versatile - can handle almost any ML task.
- Free and comes with a vast range of "add-ons"; being improved almost daily.
- Countless online resources to learn Python. Find below a Wikilist of Python tutorials for beginners: https://wiki.python.org/moin/BeginnersGuide/Programmers.
- Useful resource on ML and Python: https://jakevdp.github.io/PythonDataScienceHandbook/


Growth of major programming languages
Based on Stack Overflow question views in World Bank high-income countries
% of overall question views each month
9%
6%
3%
0%
2012
2014
2016
2018
Time
python
javascript
java
c#
php
c++

R versus Python

- R mainly used by Statisticians, Python by Computer Science etc. $\Rightarrow$ but now convergence across disciplines
- R is a very popular programming language and free. You can access R routines via Python
- Python is more versatile, you can do all the econometrics in Python and many other tasks. It is one of the main languages of machine learning. And Python is free

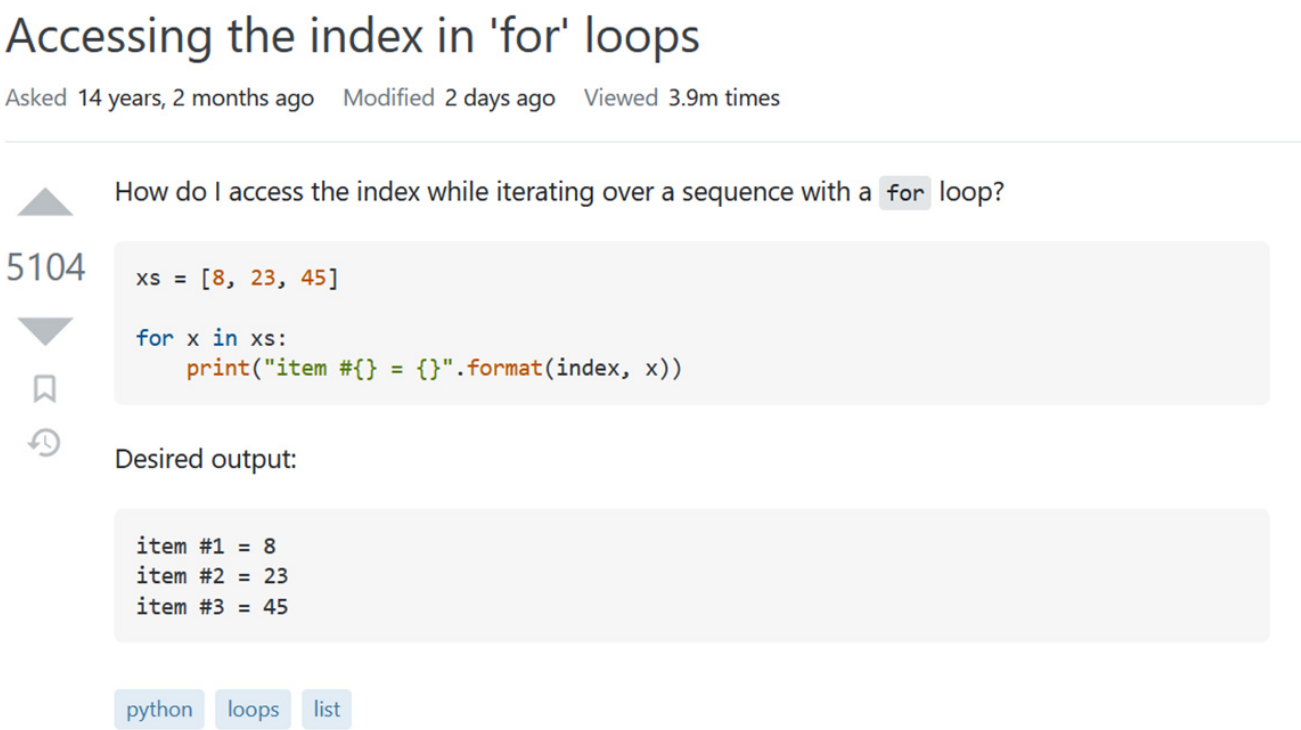
- Many available programming interfaces/ integrated development environments (IDEs) but we will use the Anaconda distribution.
  - Freely available at https://www.anaconda.com for Mac and Windows
  - One of the most popular Python distributions for machine learning
  - Has almost all key Python libraries and tools pre-installed
  - Mainly utilise Jupyter Notebook but may use other platforms (Spyder, PyCharm, (Google Colab) ...)
- You can directly run Python from terminal on Mac or the command-line from Windows

Objectives of teaching Python in this module:

- Basic understanding of programming in Python.
- More in-depth understanding of working with ML-relevant Python packages.
- Foundation for more advanced study of Python and its applications.
- Data and machine learning libraries: Numpy, Pandas, Scipy, Statsmodels, Matplotlib, and Seaborn, Scikit-learn, PyTorch, TensorFlow

- Hard work and practice - like learning any new language.
- Be proactive and search for solutions.
- Countless online resources:
  - Explore and bookmark a few that you find most useful, but try not to waste too much time.
  - Do one or more online tutorials (many free ones).
- Usually, many alternative ways of coding for the same objective:
  - Annotate your code (helps other readers and also yourself later).
  - Less is better (if 2 lines of code do the trick, why have 10?).

"Stack Overflow is a question and answer site for professional and enthusiast programmers"



Accessing the index in 'for' loops

Asked 14 years, 2 months ago Modified 2 days ago Viewed 3.9m times

5104

How do I access the index while iterating over a sequence with a `for` loop?

```
xs = [8, 23, 45]

for x in xs:
    print("item #{} = {}".format(index, x))
```

Desired output:

```
item #1 = 8
item #2 = 23
item #3 = 45
```

python loops list

Git

- Open-source version control system for software code
- Efficiently tracking and managing changes to code
- Branching - developers duplicate part of source code and modify it
- Merging - modified code merged into source
- These changes are tracked and reversible

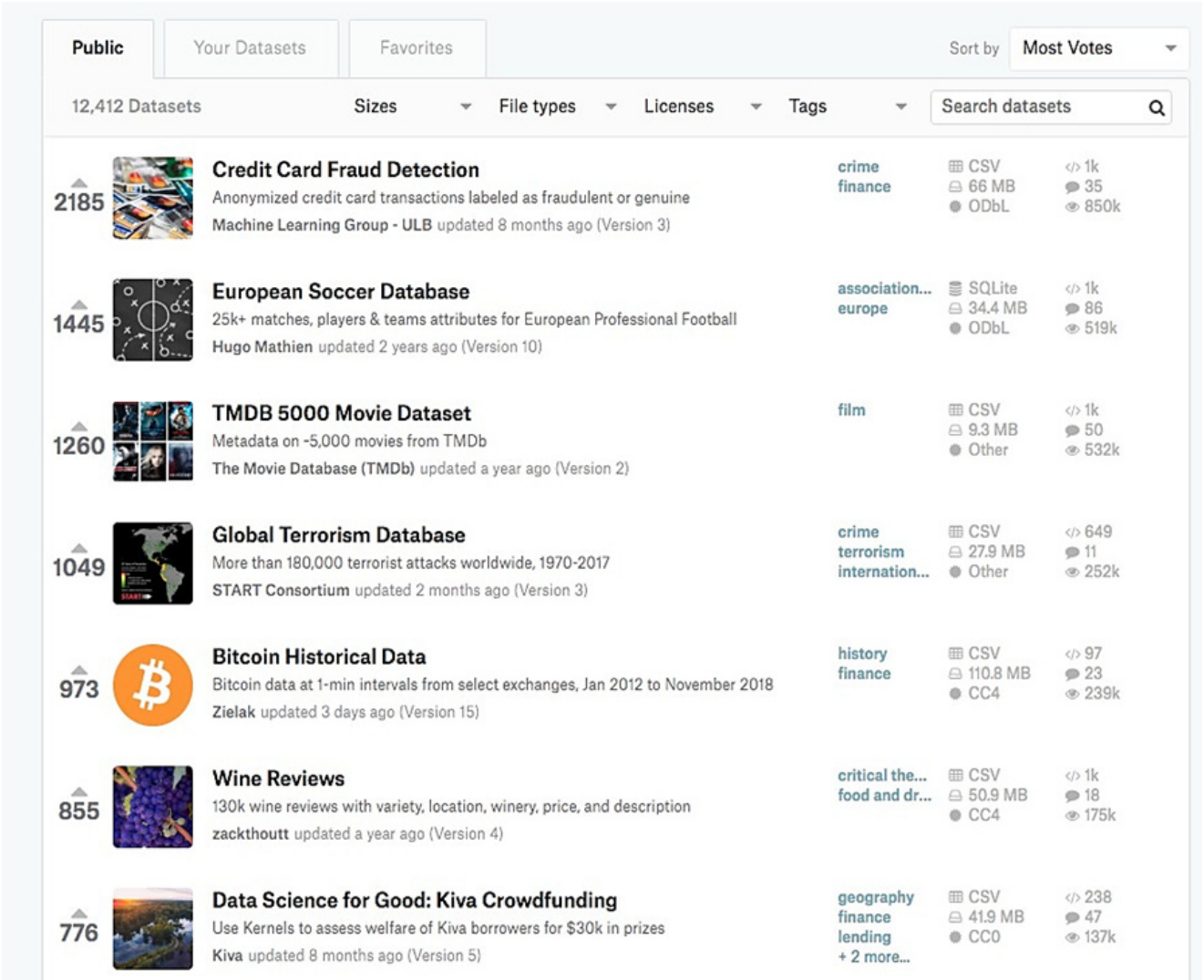
GitHub

- Online platform to host and manage code
- Share code and collaborate with other developers
- Over 83 million users worldwide (in 2022)

Kaggle

- https://www.kaggle.com
- Online community of data science professionals and enthusiasts. Owned by Google.
- Famous for its competitions - firms post machine learning problems and best algorithms win cash prizes.
- Offers short online training courses
- Data sets and code available.
- Recruitment.

Kaggle data sets



Jupyter Notebook

- Tool for interactively developing and presenting data science projects.
- Code and its output in the same document.
- Can include code, text, charts, equations, audio, video etc.
- Accommodates many different programming languages but Python most commonly used.

- Launch Anaconda and then launch the Jupyter Notebook application.
- Opens a new tab (in your default browser) that is the 'Dashboard' - lists folders in the Jupyter start-up directory.
- The address bar may show something like http://localhost:8888/tree - not a remote website but a 'server' on your own local machine.
- Jupyter Notebook and Dashboard run like web apps and are platform-independent.