Course overview

Artificial Intelligence and Machine Learning for SupTech – Lecture 1

Iman van Lelyveld – Michiel Nijhuis

VU Amsterdam

Course overview

1. Why is this course relevant?
2. What is the structure of the course — what can you expect?
3. How is the tooling, infrastructure and the Fintech landscape developing?

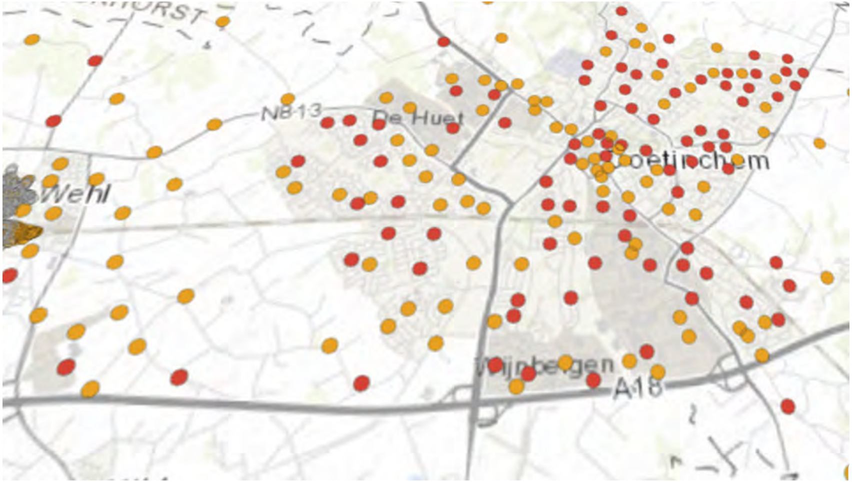
Introducing the instructors - contact details

This course will be taught by Iman van Lelyveld – Michiel Nijhuis

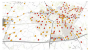
Course coordinator:	Iman van Lelyveld
Email:	iman.van.lelyveld@vu.nl
Office hours:	by appointment
Website:	Personal page

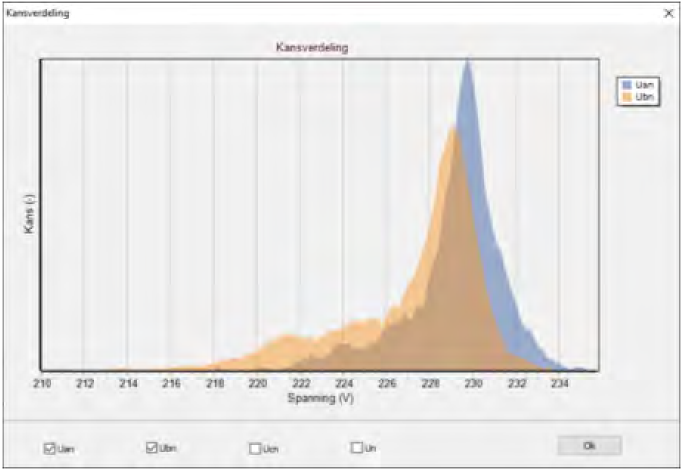
Instructor:	Michiel Nijhuis
Email:	m.nijhuis@dnb.nl
Office hours:	by appointment

Michiel Nijhuis: Researching Algorithms for the Planning of Electrical Networks



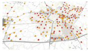
Bringing Machine Learning Algorithms in Production

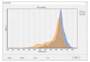




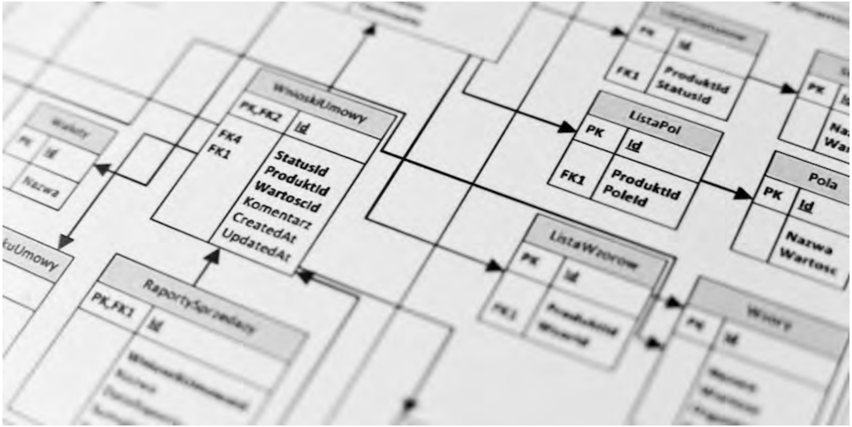
Working with all the money in the world ...

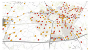
well, the Netherlands, actually

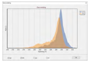




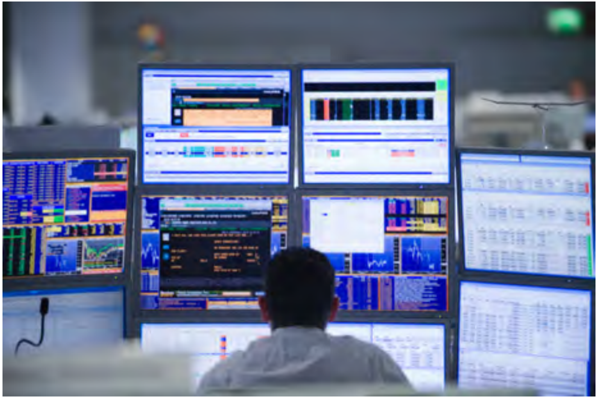
… and interconnectedness of granular data



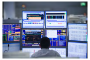




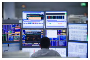
Iman van Lelyveld: Forex trading at Deutsche Bank

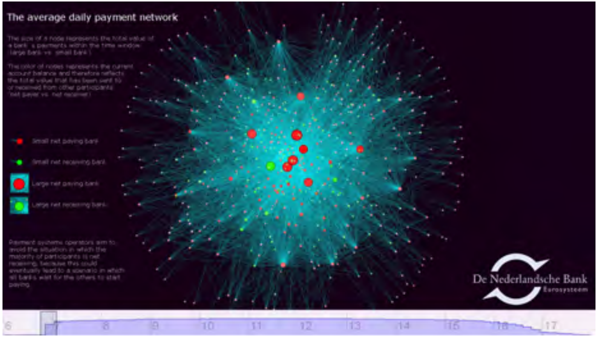


Bank of England in the 2008 crisis



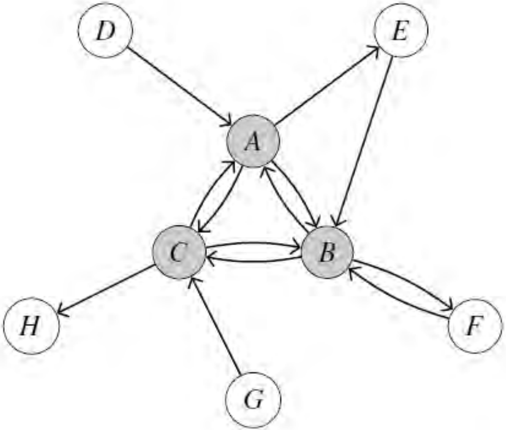
From seeing financial network structure ...



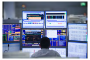


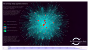
Heijmans et al. (2016)

... to more rigorous tests of structure

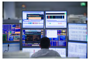


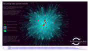
IntVeld2014

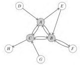




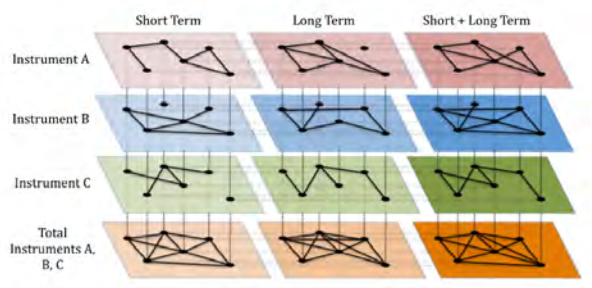
BIS – setting up an International Data Hub



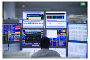


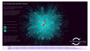


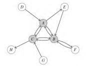
DNB Statistics Division – connecting the dots



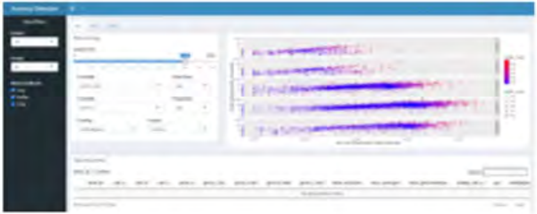
Aldasoro and Faia (2015)

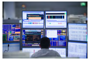


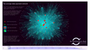


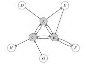


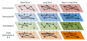
The DNB Data Science Hub





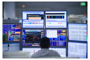


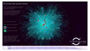


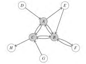


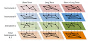
Website – Data Science Hub

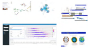
And here we are ...











HOW ABOUT YOU

The intro survey and check-in

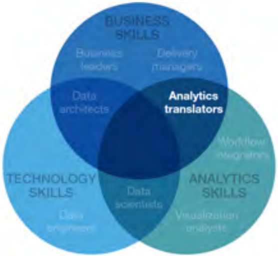
- What are you expecting?
  - How are machine learning models used in practice?
  - What jobs fit Machine Learning?
  - How to apply machine learning to trading?
  - How can machine learning be used for financial research
  - Further improve code complexity
- ... but maybe somethin entirely different:
- Pre workshop survey: voting link at www.menti.com. Code 7968 6494
- Results: result link

What is the objective of the course?

The course goal is not to become experts in the business, technology or analytics ’circle’.

The objective is to:

- Theory
  - provide an overview and understanding of popular Machine Learning (ML) and Artificial Intelligence (AI) techniques
  - understand the opportunities and limitations of these techniques
  - be able to interact with the experts
- Practice
  - work hands-on with ML/AI methods in Python
  - “demystify” the black box of ML/AI
  - prime you so that you can continue to learn by yourself



Source: McKinsey

Outline

Course Material

- The slides will be provided
- Jupyter Notebooks and other relevant material will be posted to Google colab
- A good all-round book is Géron (2019)
  - Other good references are Hilpisch (2018), Raschka (2016), and van der Plas (2016)
  - Most of these authors have put their supporting material on-line
- Where relevant we cite papers and resources in the slides

Course schedule I

Day 1

Lecture 1: Course Overview

- Why is this course relevant?
- What can you expect?
- What we will cover?

Tutorial 1: How to read data, how to use sklearn?

- Getting started with Python and data manipulation.
- How is this different from Excel?
- Read in the data and get to know it.
- Introduction to sklearn: where to find the buttons

Course schedule II

Lecture 2: Introduction to Machine Learning (ML)

- What is ML? What is ML applied to?
- Linear regression from the ML lens.
- The outlines of the ML approach
    - Supervised vs. unsupervised learning
    - Hyperparameters and how to select them
    - Gradient descent and grid search

Course schedule III

Tutorial 2: Regressions versus Classifiers

- Logit as a statistical model vs ML model
- How to find the optimal (hyper)parameters
- A different classifier: Support vector machines
  - Different types of kernels
  - First glimpse: Dangers of overfitting
  - Evaluating performance

Course schedule IV

Lecture 3: Machine Learning – the basics

- Importance of preprocessing your data
- Building up to the workhorse classifier: the logit model
- When is a classifier doing a "good" job?
- Confusion matrix, Receiver Operator Characteristic (ROC)
- What are overfitting, bias and variance?

Tutorial 3: Data preprocessing and assessing model performance

- How to preprocess: standardize your data
- Pros and cons of standardization
- Working with the confusion matrix
  - What if costs are not symmetric?
  - The trade-off between precision and recall

Course schedule V

Day 2

Lecture 4: Fighting the curse of dimensionality

- How to reduce dimensionality?
  - K-Nearest Neighbors (KNN)
  - Principal Components Analysis (PCA)
- Feature selection and regularization
  - How to select the most important features?
  - Examples: RIDGE, LASSO, Elastic net
- Is a “good” model always good? What is external validity?
- Cross-validation and holdouts

Course schedule VI

Tutorial 4: Cross validation applied to LASSO variable selection

- Looking closer at cross validation (CV) and holdouts
- K-fold, Leave-one-out, stratified CV
- Splitting your data into training and testing samples
- How to use CV to tune a LASSO model

Lecture 5: Improving weak learning

- How to grow a decision tree? How to split?
  - Purity measures
- Can Ensemble Classifiers improve weak learners?
  - Bagging and boosting
  - Examples: AdaBoost, XGBoost

Course schedule VII

Tutorial 5: Decision trees and random forests

- Growing your own decision tree
- How deep? How many splits? How big are the leaves?
- From trees to random forests
- Comparing performance with the confusion matrix

Lecture 6: Unsupervised learning and clustering

- Supervised versus unsupervised learning
- What can we do with unsupervised learners?
- K-means, t-SNE, DBSCAN, Gaussian mixtures

Course schedule VIII

Tutorial 6: Finding clusters and neighbors

- Implementing K-means and DBSCAN
- Hierarchical clustering: Bottom-up or Top-down?
- Visual inspection of results

Course schedule IX

Day 3

Lecture 7: Natural Language Processing (NLP)

- What are the main approaches in textual analysis?
- Going beyond simple word counts
- How to extract market sentiment?

Tutorial 7: NLTK and sentiment analysis

- Constructing a bag of words
- Classifying sentiments (positive/negative)
- Example with financial news data

Course schedule X

Lecture 8: Summary

- Discuss some things that can go wrong
  - Survivorship bias, input errors and deceit
  - Fairness and discrimination
- What is the reaction of authorities?

Getting started

For the tutorials we will be using Google Colab in combination with Github

1. Have your own Google account here. This will give you access to Google Colab.
2. The tutorials are hosted on Github. You can requst access by following this link Request access to course folder
3. Select the notebook for the tutorial you want to open and select the button Open in Colab
4. You can now run the notebook from Google Colab

What Math/Stats background do you need?

- For a refresher see a textbook like Greene (2013)
- Partial derivatives, e.g.

$$\frac{\partial(x^3 + y^2 + 1)}{\partial x}$$

- Matrix and vector operations

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \times \begin{bmatrix} 7 \\ 8 \\ 9 \end{bmatrix} = \begin{bmatrix} 1 \times 7 + 2 \times 8 + 3 \times 9 \\ 4 \times 7 + 5 \times 8 + 6 \times 9 \end{bmatrix} = \begin{bmatrix} 50 \\ 122 \end{bmatrix}$$

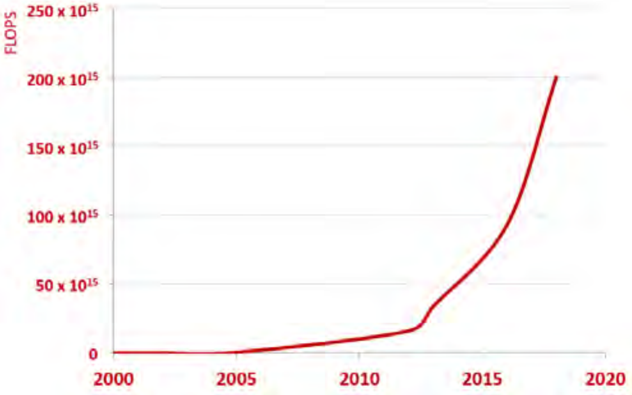
- Vector dot product

$$z = \mathbf{w}^{\mathbf{T}}\mathbf{x} = \sum_{j=0}^{m} \mathbf{w_j x_j}$$

$$\begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \times \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} = 1 \times 4 + 2 \times 5 + 3 \times 6 = 32.$$

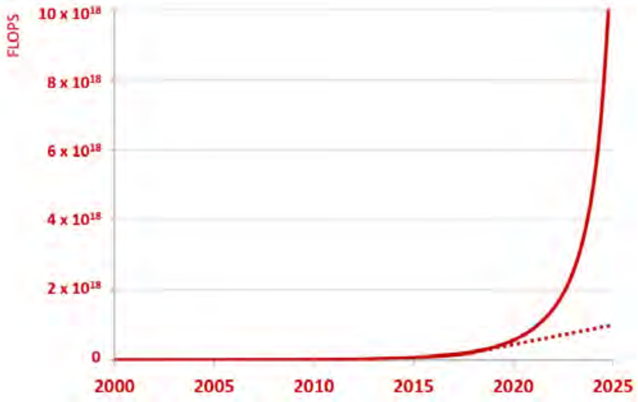
Outline

Computing power increases continuously



200,000,000,000,000,000
floating point operations per second
(or: 20,000 high-end PC's)

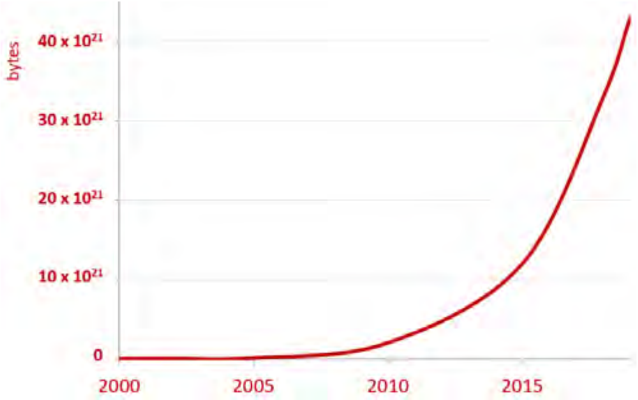
Computing power increases continuously



10,000,000,000,000,000,000

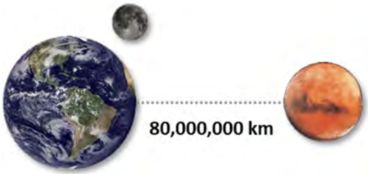
~~200,000,000,000,000,000~~

We collect ever more data

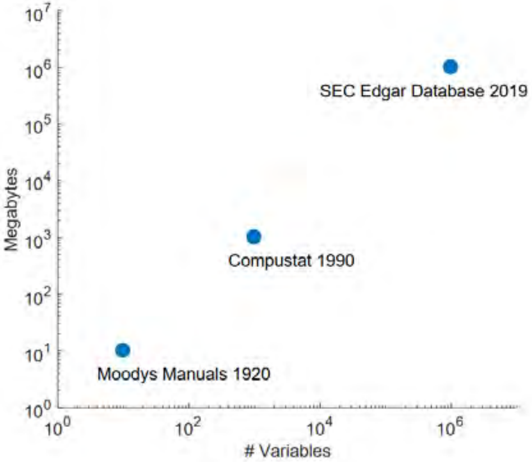


Annual in 2005 $\equiv$ daily in 2020

Or, on CD-rom

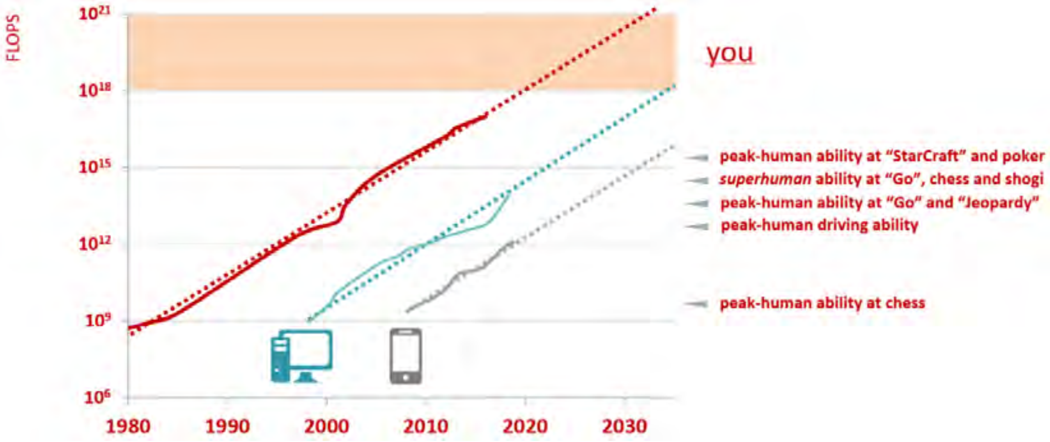


More data available to extend the set of x_t in asset pricing



Source: Nagel (2019)

CPU power + algos + data has resulted in astounding progress



Human activity becomes more digitised

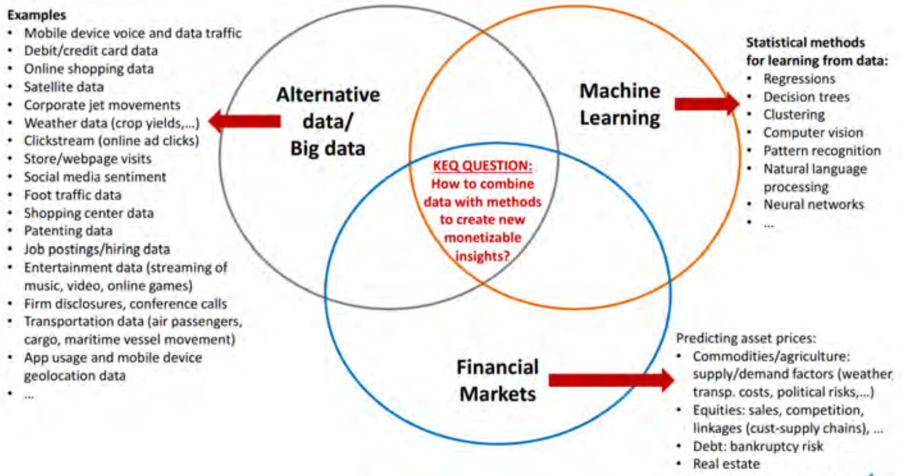
ML is everywhere …

Changes in personal interests or in population characteristics (adaptive news access)

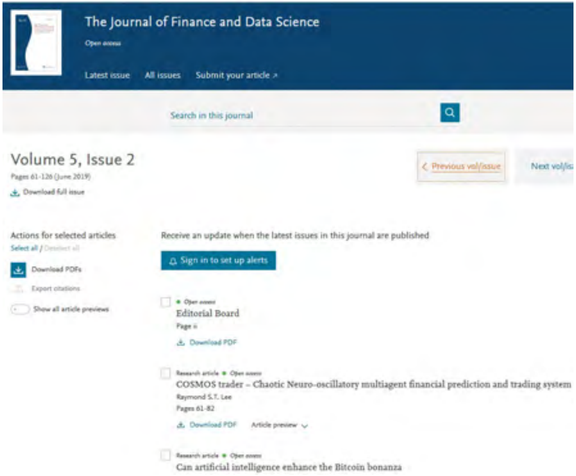
Adversary activities (avoiding spam filters; credit card fraud)

Changes in population characteristics (credit scoring)

… and has many applications in finance



Even has it's own journal!



Source: Journal of Finance and Data Science

Outline

Opinion **Artificial intelligence**

Why Google thinks we need to regulate AI

Companies cannot just build new technology and let market forces decide how it will be used

SUNDAR PICHAI

+ Add to myFT

Outline

Intelligent computers have held promise for a long time ...

In 1958 the New York Times reported that the Perceptron, an early AI machine developed at Cornell University with military money, was "the embryo of an electronic computer that [the American Navy] expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence".

(Economist, May 14th 2015)

More on Artificial Intelligence (AI)

> *"Every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."*
>
> *(Dartmouth Artificial Intelligence Conference (1956))*

AI can refer to anything from a computer program playing a game of chess, to a voice-recognition system like Amazon's Alexa interpreting and responding to speech.

The technology can broadly be categorized into three groups:

- narrow AI: skilled at one specific task, eg IBM's Deep Blue (beat chess grand master Garry Kasparov – 1996), or Google DeepMind's AlphaGo (Go master Lee Sedol – 2016).
- artificial general intelligence (AGI) $\equiv$ human-level
- superintelligent AI

  > *"an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills"*

  In other words, it's when machines have outsmarted us.

More Machine Learning (ML) definitions ...

- "Machine learning research is part of research on artificial intelligence, seeking to provide knowledge to computers through data, observations and interacting with the world. That acquired knowledge allows computers to correctly generalize to new settings."
- "A well-posed learning problem: A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."
- "Machine learning is the science of getting computers to act without being explicitly programmed."
- "Machine learning algorithms can figure out how to perform important tasks by generalizing from examples."

Machine learning = Statistics + Programming!

What is the difference, really ...

Whatever your definitions, environment used is largely irrelevant

Fitting Elastic Net regression in:

R

```
eNet <- glmnet(X, Y, alpha = 0.5, lambda = myLambdas)
```

MATLAB

```
eNet = lasso(X, Y, 'Alpha', 0.5, 'Lambda', myLambdas)
```

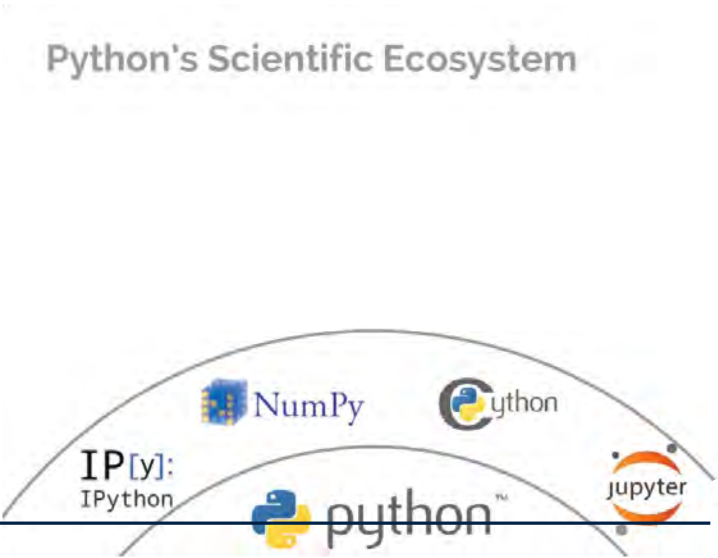
PYTHON/scikitLearn:

```
eNet = ElasticNet(alpha = myLambda, l1 ratio= 0.5)
eNet.fit(X, Y)
```

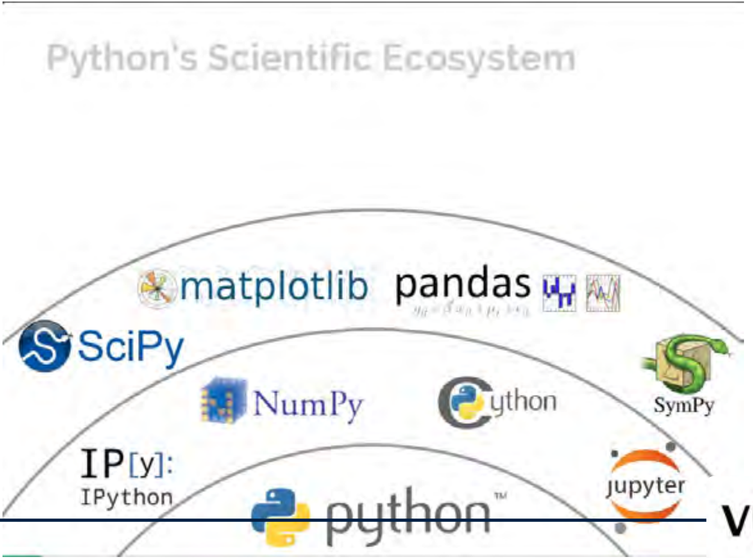
The crucial thing is to know what elastic net, alpha and lambda mean!

We choose Python because of the ecosystem

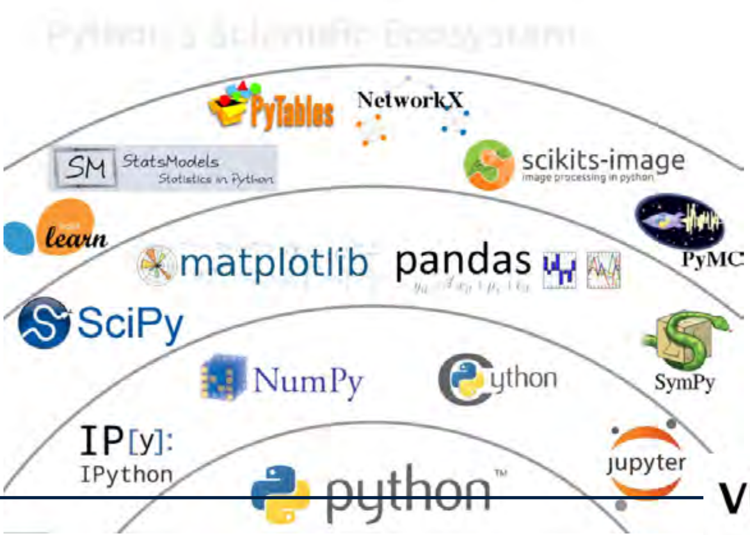
We choose Python because of the ecosystem



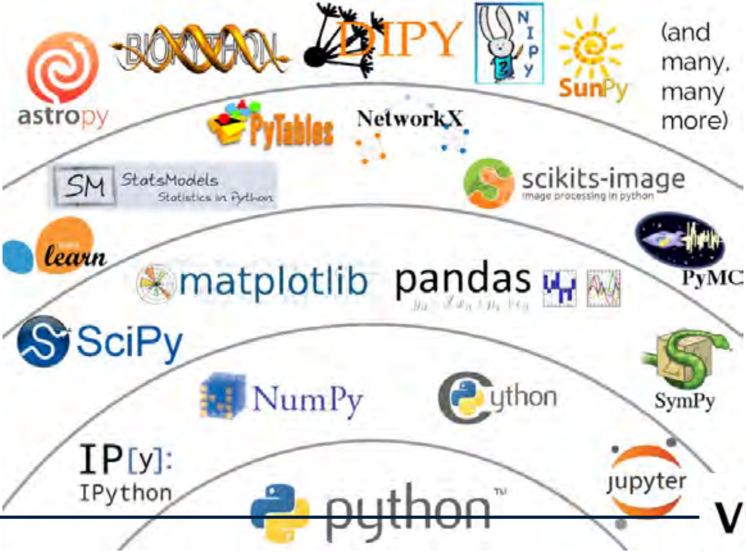
We choose Python because of the ecosystem



We choose Python because of the ecosystem

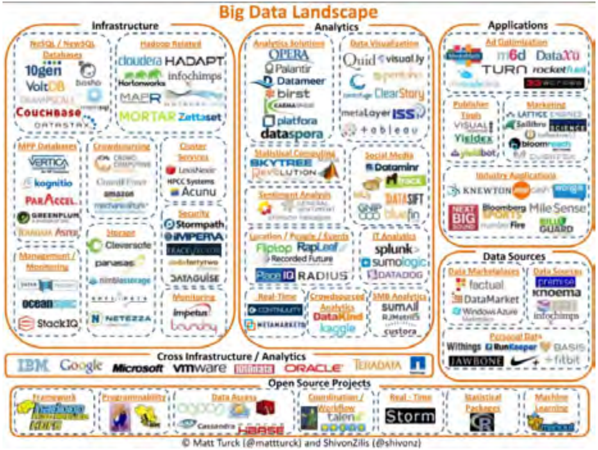


We choose Python because of the ecosystem



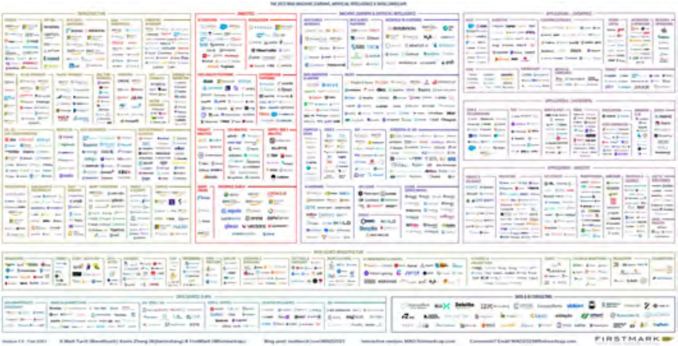
Outline

Tooling is evolving fast ...



2012

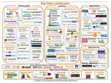
Tooling is evolving fast ...



2023

View in full size: click here.

2012



2016

2020

2021

… as well as the Fintech landscape

2015

… as well as the Fintech landscape

2021

Source: Holand Fintech

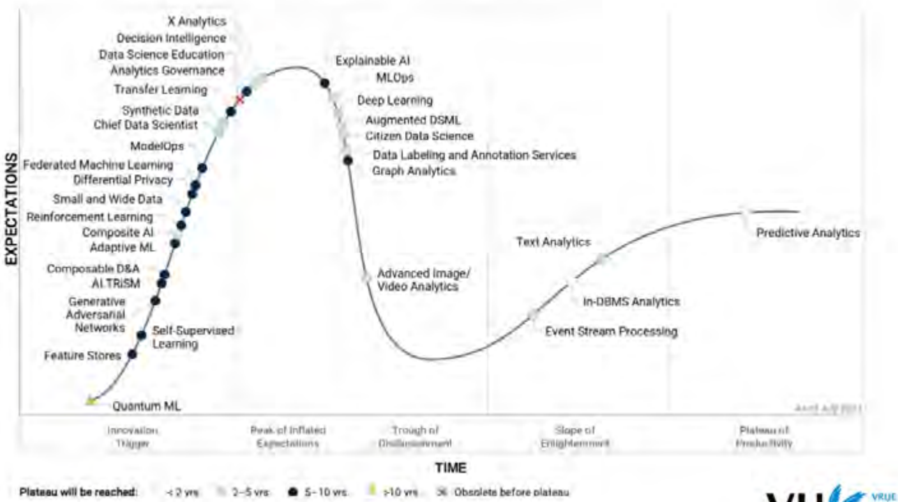
2015

2016

2017

2018

Data Science Hype Cycle – 2021



Summary

In this lecture we covered:

1. Introductions and housekeeping
2. Why is this course relevant and how does it fit in the Honors programme
3. A first discussion of how ML is different
4. Sketching the landscape

Some definitions

- Artificial intelligence (AI): A broad discipline with the goal of creating intelligent machines, as opposed to humans' and animals' natural intelligence. It is a catch all term that nonetheless captures the long term ambition of the field to build machines that emulate and then exceed the full range of human cognition.
- Machine learning (ML): A subset of AI that often uses statistical techniques to give machines the ability to "learn" from data without being explicitly given the instructions for how to do so. This process is known as "training" a "model" using a learning "algorithm" that progressively improves model performance on a specific task.
- Deep learning (DL): An area of ML that attempts to mimic the activity in layers of neurons in the brain to learn how to recognise complex patterns in data. The "deep" in deep learning refers to the large number of layers of neurons in contemporary ML models to achieve performance gains.
- Reinforcement learning (RL): An area of ML concerned with developing software agents that learn goal-oriented behavior by trial and error in an environment that provides rewards or penalties in response to the agent's actions (calle towards achieving that goal.

What if you haven't had enough?

- This course is too short to cover everything
- Luckily, there are a number of very good free courses out there
  - Google's crash course on Machine Learning
  - An excellent full length MOOC by Andrew Ng on Coursera
  - Courses from Harvard, IBM and Microsoft at EDX
  - Also see Udemy although you have to pay a little bit
  - coding for economists
- Check out the use cases at Kaggle
- Or if you are in for a game: Code Combat
- Please let us know if you have come across others!

What about softcopy resources?

- There are an insane amount of blogs, repo's out there. Here is wildly incomplete list
  - Some really good links about ML an Econ on Dario Sansone's webpage
  - Economics and Data Science resources
  - For code, see the website paperswithcode.com
  - A nice book on interpretability as a website or as a Pact book Molnar (2019)
  - Great book (in progress) Coding for economists with great Python examples.
  - Very nice cheatsheets for folks coming from other languages: Stata, R, Matlab, Excel
  - Resources that go with the Python for Finance by Hilpisch
  - More Quant resources at O'Reilly
  - Cheatsheets for folks coming from other languages: Stata. Still looking for R, Matlab, Excel
- Please let us know if you have come across others!

What about hardcopy resources?

- There are also numerous books out there, some of them free and all of them come with a lot of supplementary material
  - McKinney, W. (2022). Python for Data Analysis, 3rd edition. https://wesmckinney.com/book/
  - Géron, A. (2019). Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly. Retrieved April 21, 2019, from https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/
  - Wentworth, P., Elkner, J., Downey, A. B., & Meyers, C. (2019). How to think like a computer scientist: Learning with Python [arXiv: 1011.1669v3 ISSN: 1098-6596].
  - Raschka, S., & Mirajalili, V. (2017). Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow. Retrieved July 27, 2019, from https://www.packtpub.com/big-data-and-business-intelligence/python-machine-learning-second-edition
  - James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R.
  - Hastie, T., Tibshirani, R., & Friedman, J. (2017). The Elements of Statistical Learning: Data Mining, Inference, and Prediction [arXiv: 1011.1669v3 ISSN: 03436993].
  - Efron, B., & Hastie, T. (2017). Computer Age Statistical Inference.
  - Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking (O'Reilly, Ed.).
  - Sweigart, A. (2015). Automating the Boring Stuff with Python. https://automatetheboringstuff.com/

What about your tech stack

This course is not about research methods but if you want take things serious you should invest in the following technologies

- Organise your thoughts: OneNote, Mindmap, Gingko App
- Organise your references: Zotero, Mendeley, Jabref. You might find this tedious but it will pay off by the time you have to find that one article or get the reference right.
- Use LaTeX. Overleaf makes it really easy to start
- Pick your favourite editor – preferably suitable for multiple languages and LaTeX. VSCode, Sublime Text, PyCharm.
  - VS Code $\neq$ VS. Has an active market place and intergrates well with the Microsoft stack
- Check out Open metric for a much more thorough list

Things are moving fast so please let us know if you have come across interesting tools!

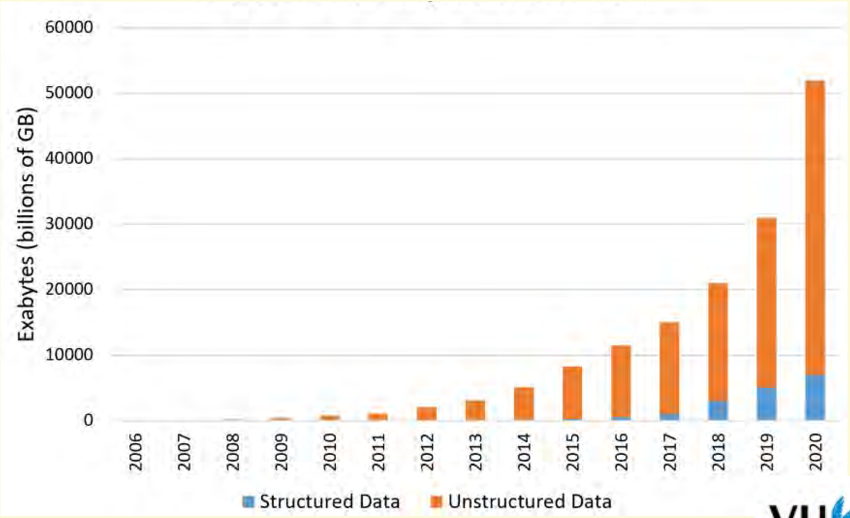
Where to find data?

- Most financial data sets are under lock and key but some of it is free

Awesome list	Overview of public data sets, mostly trade. Opencorporates
DB.nomics	Amazing portal with data from BIS, ECB, OECD and many others. It has an API from within Python
WRDS	Wharton Database, See Leonard Wolk's lectures
Yahoo finance	Various API's available (e.g., Rapidapi)
Mockaroo	If you need to make up a test data set with realistic features
Draw my data	Generate a data set with some properties
Kolanovic and Krishnamachari (2017)	has a great chapter "Handbook of Alternative Data"
COVID travel data	
Synthetic data (JP Morgan)	useful to generate realistic (but fake) data
Google trend and financial time series	
Google Trend and unemployment	Nice example of what to do with Google Trend
re3data	Database of data sets

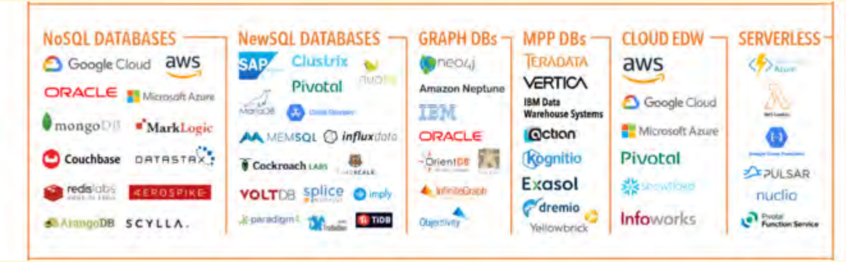
- See the api-packages-overview.xls with an overview of available APIs on De Nederlandsche Bank Github. Let us know if you have comments or a data_science@dnb.nl

The Cambrian explosion … of data



What types of database are out there

- Relational vs Non-relational databases
  - Relational engines require consistency at the end of each transaction, whereas non-relational engines only require that the database be consistent at some point in time (i.e., eventually)



SWOT of relational databases

Relational Database Management Systems (RDMS)

- strictly enforce type (eg numeric, string, etc)
- example: SAS

strong	layout data is known in advance but use not yet
weak	variable data

Source: Perkins et al. (2018)

SWOT of non-relational databases

Key-value (KV)

- Simple keys to values
- example: key-value stores

strong (horizontally) scalable and fast for unrelated data (eg users' session data)
weak no index hence only basic CRUD (**C**reate, **R**ead, **U**pdate, **D**elete)

Source: Perkins et al. (2018)

SWOT of non-relational databases

Key-value (KV)

- Simple keys to values
- example: key-value stores

strong	(horizontally) scalable and fast for unrelated data (eg users' session data)
weak	no index hence only basic CRUD (**C**reate, **R**ead, **U**pdate, **D**elete)

Columnar

- Everything in a column, rows are not kept together, nice versioning
- example: Monet db

strong	indexing web pages
weak	it would be best if you know query structure in advance

Source: Perkins et al. (2018)

… and some more

Document

- any number of fields with unlimited nesting, JavaScript Object Notation (JSON)
- example: MongoDB, CouchDB, document stores

strong highly variable data, good match with object-oriented languages
weak verbose

Graph

- focus on the relation between nodes than on the actual node information
- example: Neo4J, graph databases

strong social network queries
weak comparing node info between 2 nodes $\rightarrow$ this is then stored elsewhere

Source: Perkins et al. (2018)

Where to find handy tidbits?

- Holidays can be found using the Holidays PyPi package
- ISO codes for countries, currencies etc. can be found with the pycountry PyPi package
- Testing out your regular expression regex101 site
- Papers with code
- Compute distance distance between locations (API) at https://developer.here.com
- The list is endless ... please let me know if you find other interesting ones.

References I

- Aldasoro, I., & Faia, E. (2015). Systemic Loops and Liquidity Regulation. *CEPR Discussion Paper*, (10918).
- Efron, B., & Hastie, T. (2017). *Computer Age Statistical Inference*.
- Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly. Retrieved April 21, 2019, from https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/
- Greene, W. H. (2013). *Econometric Analysis* [Publication Title: Econometric Analysis]. MacMillan.
- Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* [arXiv: 1011.1669v3 ISSN: 03436993].
- Heijmans, R., Heuver, R., Levallois, C., & van Lelyveld, I. (2016). Dynamic visualization of large financial networks. *Journal of Network Theory in Finance*, *2*(2), 57–79.

References II

- Hilpisch, Y. (2018). Python for Finance: Mastering Data-Driven Finance. O'Reilly. Retrieved March 1, 2019, from http://shop.oreilly.com/product/0636920117728.do
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R.
- Kolanovic, M., & Krishnamachari, R. T. (2017). Big Data and AI Strategies - Machine Learning and Alternative Data Approach to Investing Quantitative and Derivatives Strategy. J.P. Morgan report.
- McKinney, W. (2022). Python for Data Analysis, 3rd edition. https://wesmckinney.com/book/
- Molnar, C. (2019). Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book/cite.html
- Nagel, S. (2019). Asset Pricing and Machine Learning - Lecture 1. Princeton Lectures in Finance.

References III

Perkins, L., Redmond, E., & Wilson, J. R. (2018). Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movem The Pragmatic Bookshelf. Retrieved April 20, 2019, from http://www.pragprog.com.

Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic t (O'Reilly, Ed.).

Raschka, S. (2016). Python Machine Learning.

Raschka, S., & Mirajalili, V. (2017). Python machine learning: Machine learning and deep learning with Python, scikit-learn, Retrieved July 27, 2019, from https://www.packtpub.com/big-data-and-business-intelligence/python-machine-learning-second-edition

Sweigart, A. (2015). Automating the Boring Stuff with Python. https://automatetheboringstuff.com/

References IV

- van der Plas, J. (2016). Python Data Science Handbook. O'Reilly. Retrieved October 21, 2019, from https://jakevdp.github.io/PythonDataScienceHandbook/
- Wentworth, P., Elkner, J., Downey, A. B., & Meyers, C. (2019). How to think like a computer scientist: Learning with Python [arXiv: 1011.1669v3 ISSN: 1098-6596].