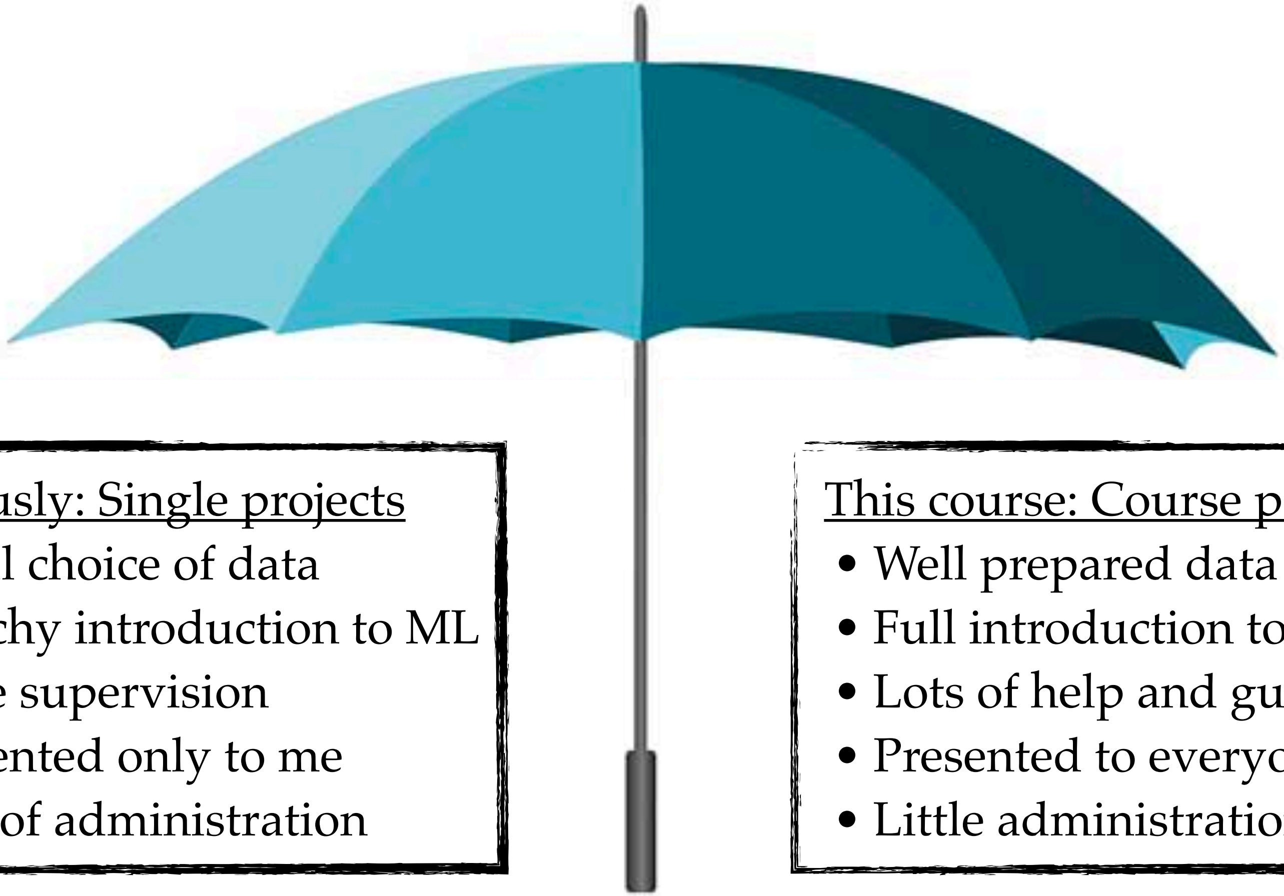
Applied Machine Learning

Course information 2021

Troels C. Petersen, Adriano Agnello,
Zoe Ansari, Carl Johnsen & Vadim Rusakov

This course is (partially) an “umbrella course” for doing projects with ML

An “umbrella course”



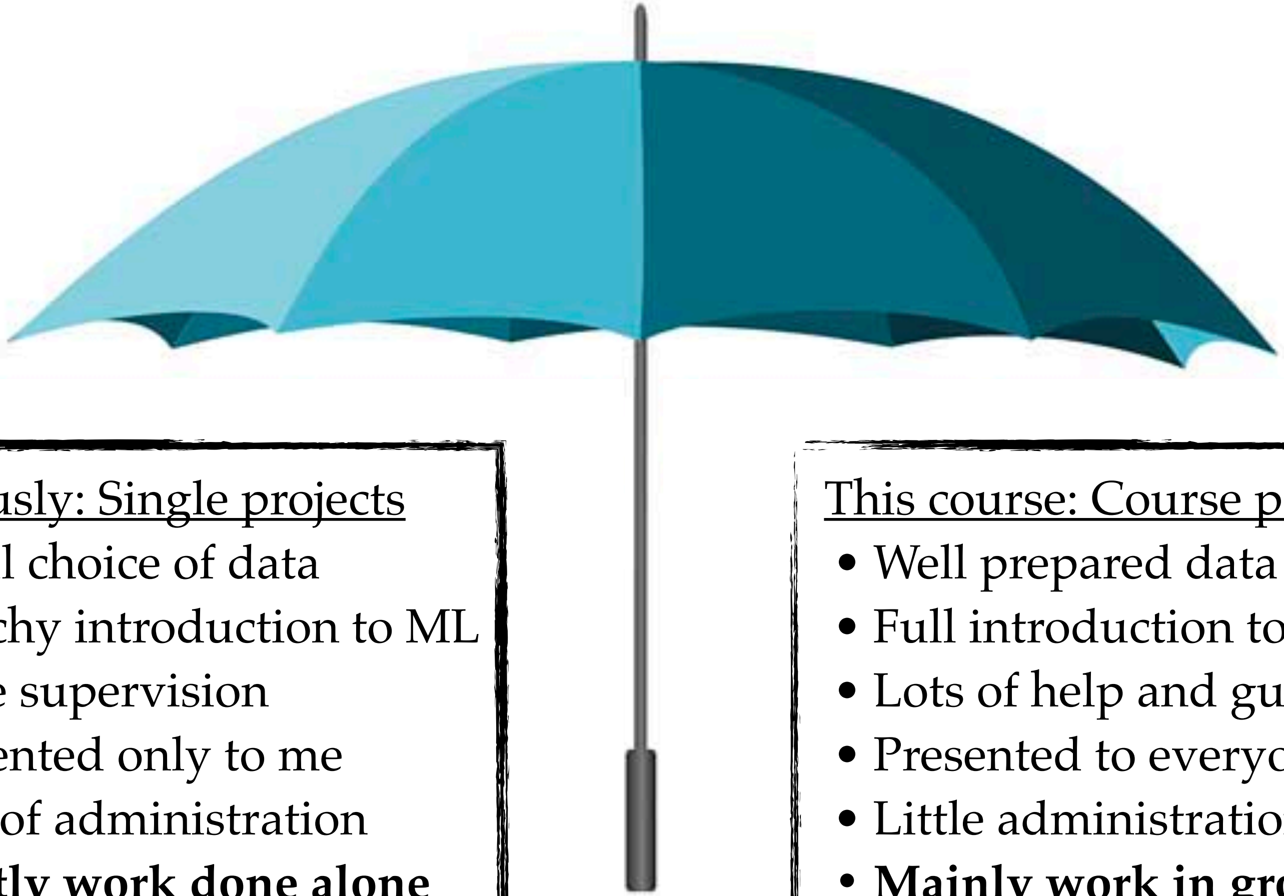
Previously: Single projects

- Small choice of data
- Sketchy introduction to ML
- Little supervision
- Presented only to me
- Lots of administration

This course: Course projects

- Well prepared data cases
- Full introduction to ML
- Lots of help and guidance
- Presented to everyone
- Little administration

An “umbrella course”



Previously: Single projects

- Small choice of data
- Sketchy introduction to ML
- Little supervision
- Presented only to me
- Lots of administration
- **Mostly work done alone**

This course: Course projects

- Well prepared data cases
- Full introduction to ML
- Lots of help and guidance
- Presented to everyone
- Little administration
- **Mainly work in groups**

General words on the course

These are extraordinary times, which call for extra-ordinary measures!

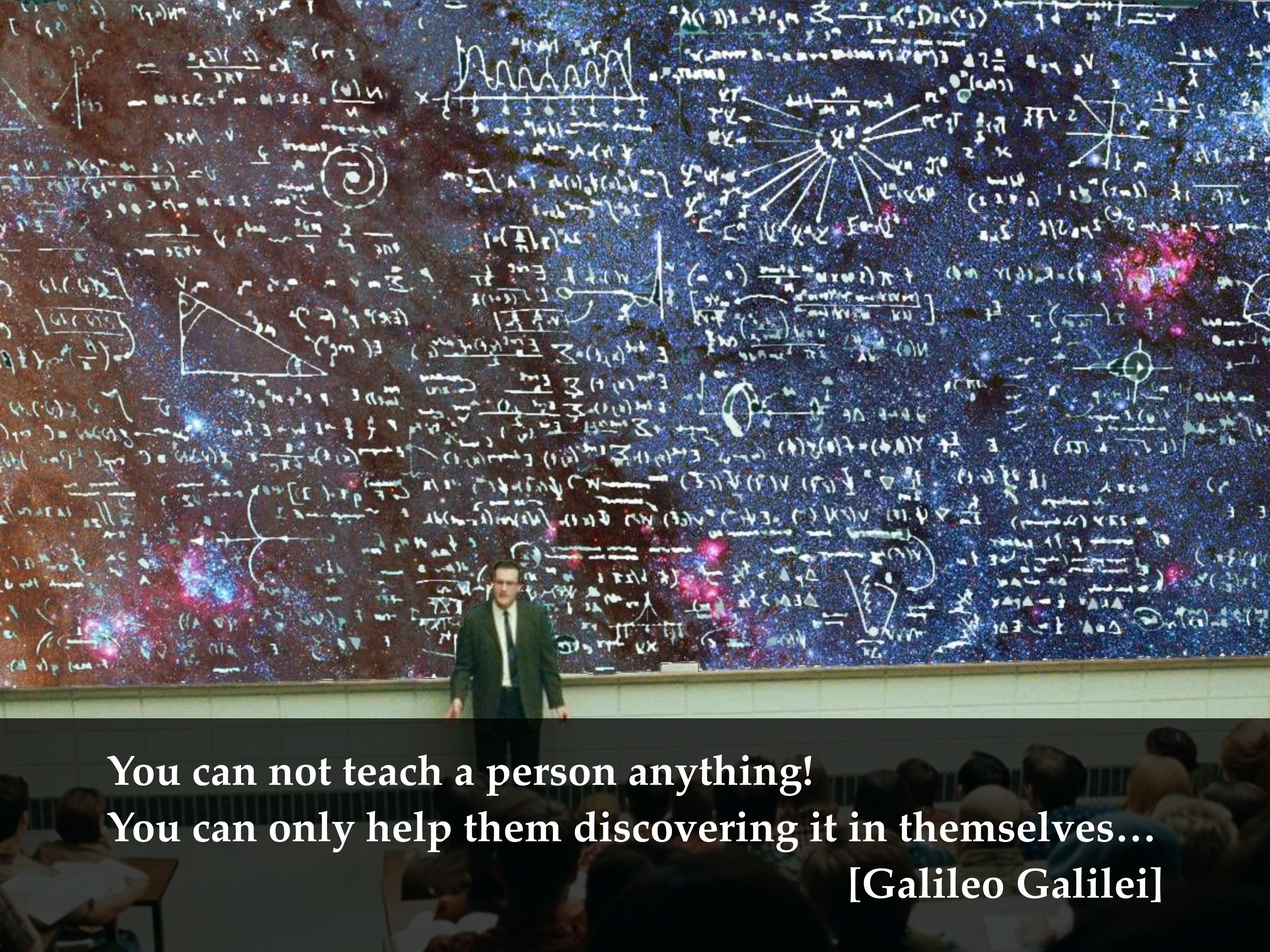
We will (at least to begin with) run the course **online only**, with lectures given via Zoom, followed by exercises which are supervised via Zoom, Slack & your favorit communication platform.

This will require both self-disciplin and dedication to the course work.

We will of course do our best to inspire, help, and promote collaboration, but it is up to you (more than normally), how much you want to learn/benefit from this course.

Course work can/should be done in collaboration with fellow students.

So please make small teams of peers, with whom you can discuss the many details of ML coding and the problems, challenges, and issues involved. This is you best way of discussing with peers, learning most, and not getting stuck.



You can not teach a person anything!
You can only help them discovering it in themselves…
[Galileo Galilei]

Lectures

The lectures will be by Zoom only.

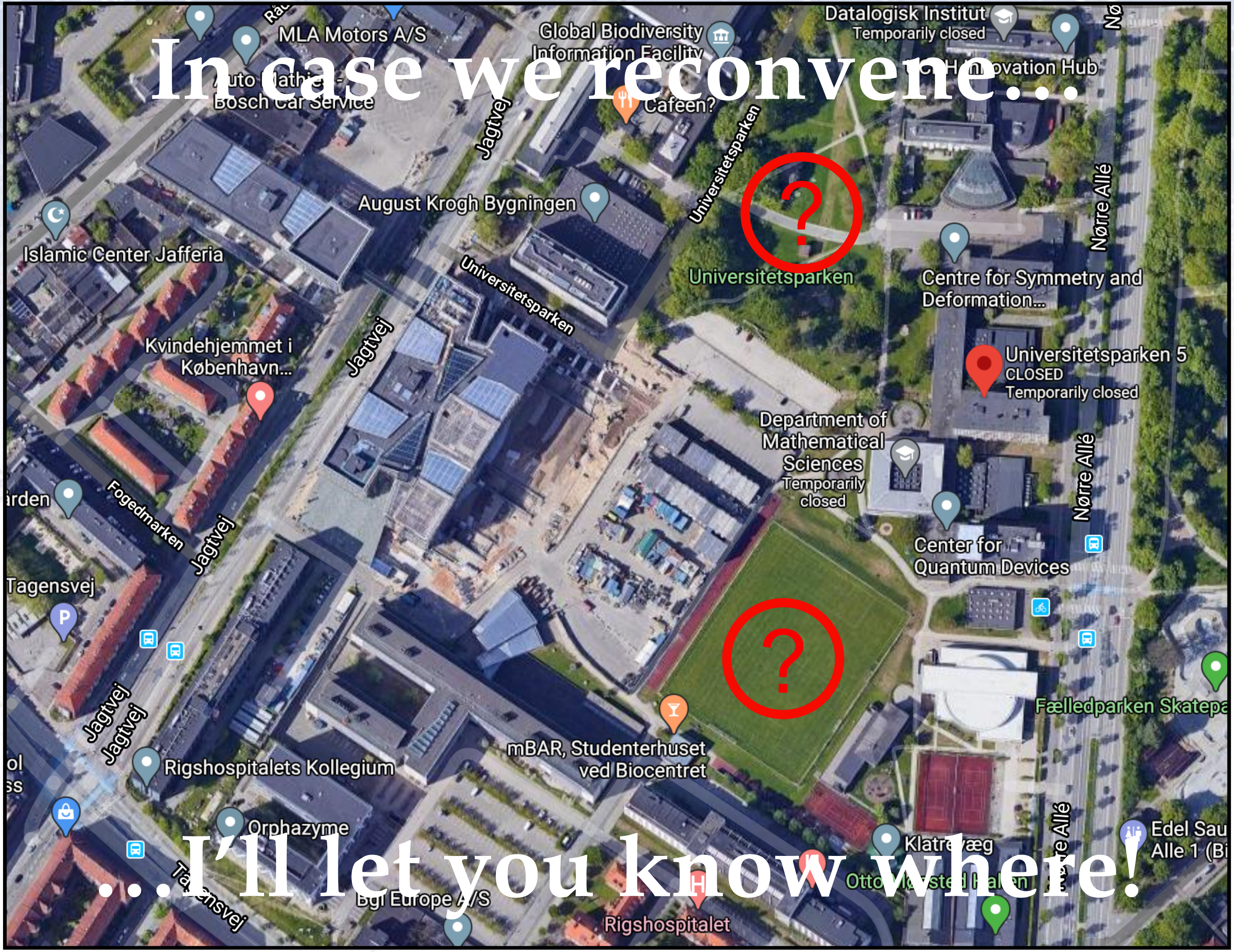
This has some advantages:

- **I will record the lectures (and chat)** and put it on the course webpage.
- You will be able to write questions/comments in the chat during the lectures. You can either write to "everybody" or to the "TA on duty".

I know that Zoom is not the most inspiring source, but I'll do my best to keep the lectures (partially) inspiring and (remotely) entertaining.

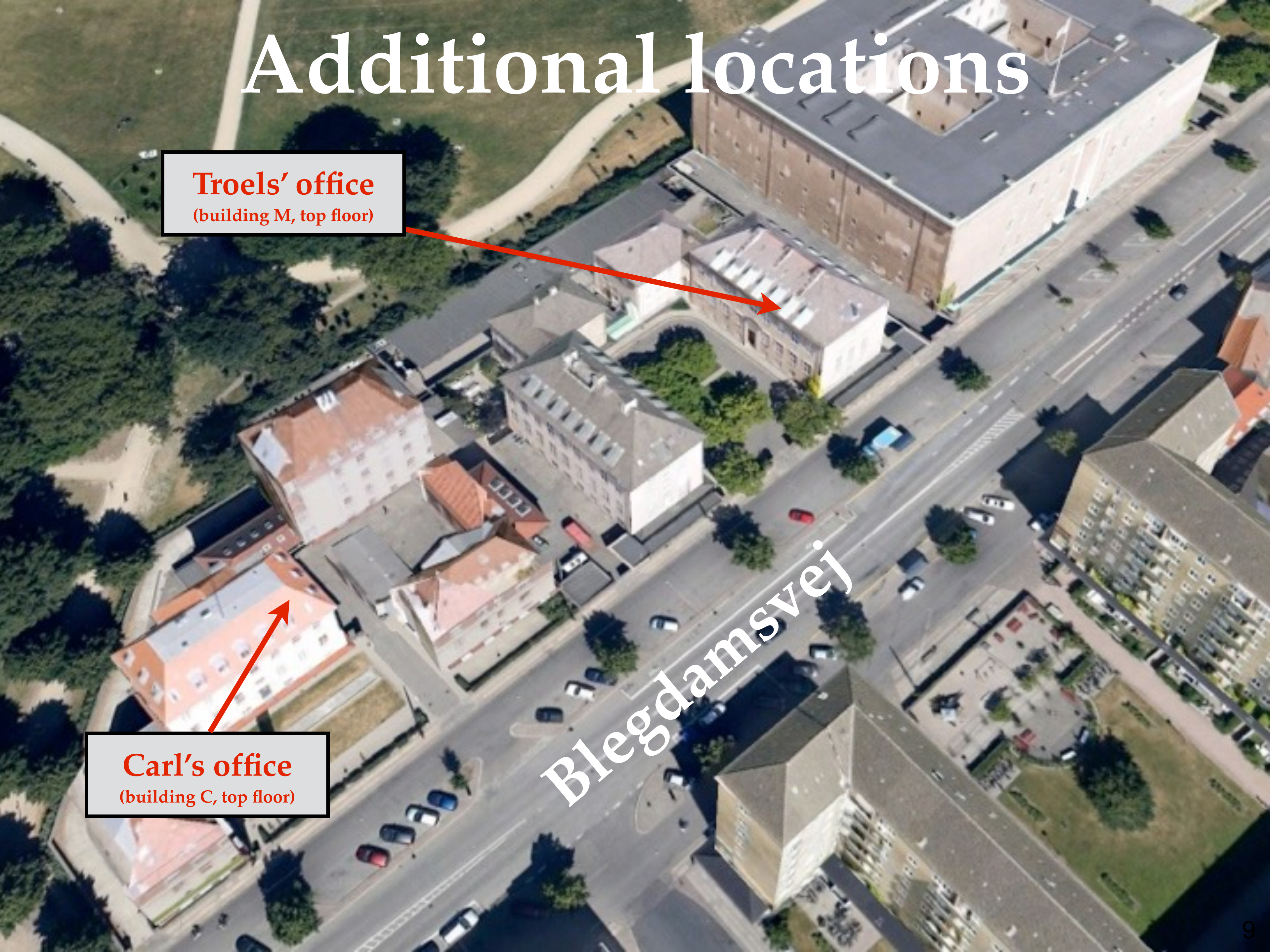
Personally, I hope (and trust) that we will reconvene again (in exercises). But the course will be possible to follow fully online.

In case we reconvene…

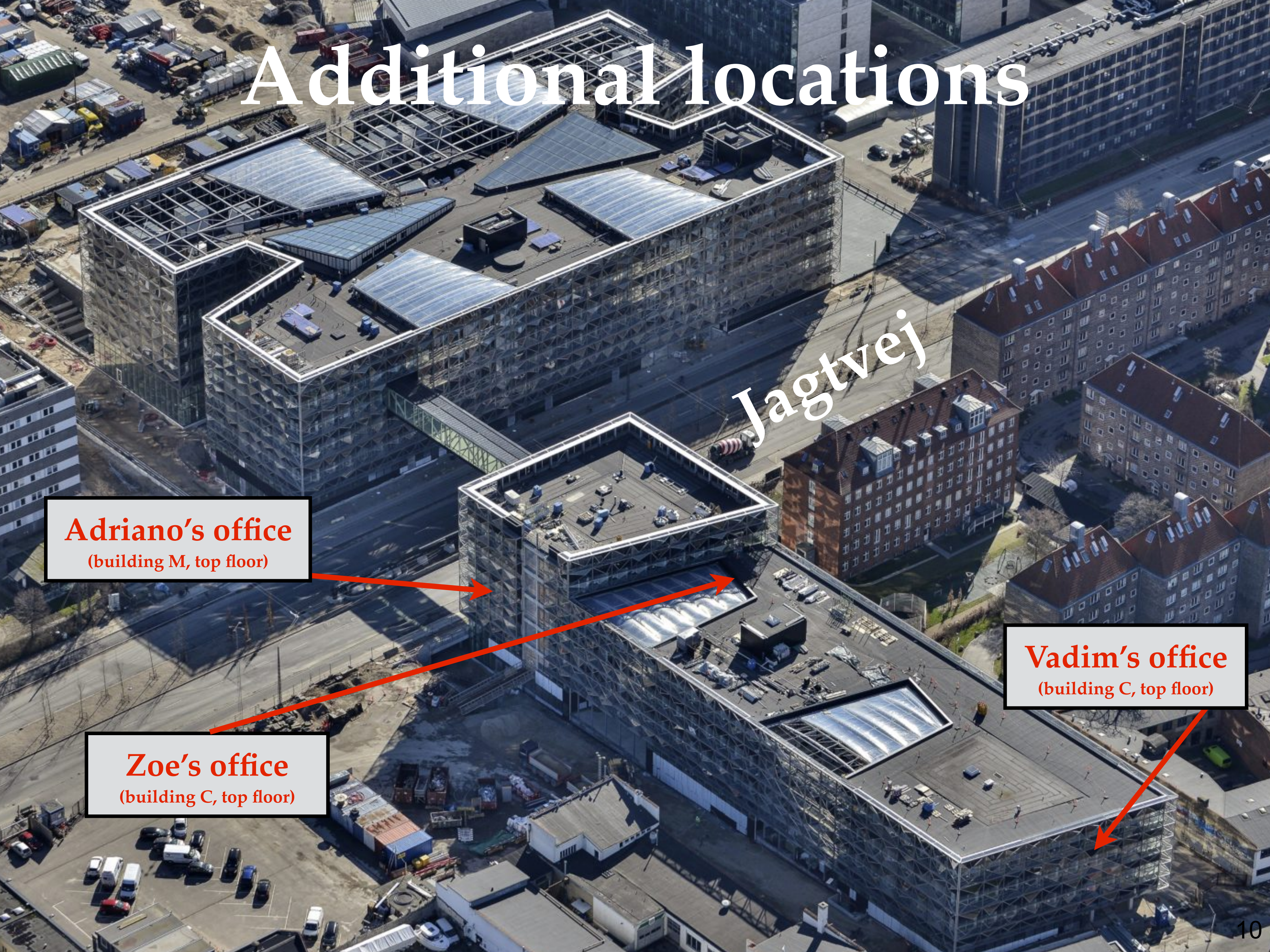


…I'll let you know where!

Additional locations



Additional locations



Adriano's office
(building M, top floor)

Zoe's office
(building C, top floor)

Vadim's office
(building C, top floor)

Computers and software

We will program in Python. You may **choose as you wish**, but we highly recommend Python. We will only provide data, code snippets (in Python), and occasional code/solutions for inspiration.

We suggest that you use Jupyter Notebook, and run everything on your own laptop, possibly with ERDA as a backup. We also recommend that you use GitHub.

Data files will typically be provided in CSV and/or HDF5 format, but others might be used.

We will be using many additional Python packages, introduced along the way, and surely you have your own favourites. Use them happily but knowingly.

Projects / Exam

This course is to some extend an umbrella course for projects using ML.

We will be doing two projects:

- An initial "small project" on **common data** (2 weeks - 40% of your grade).
- A larger "final project" on **data of your choice** (3 weeks - 60% of your grade).

The **small project** will be the basic applications of ML (classification, regression, and clustering) to a data set, and we will evaluate your (algorithm's) performance on a test set.

The **final project** will be your main task, and can be the application of ML on anything that you like. You will all be presenting your results to each other, so that also others may learn from what you did (and didn't).

You can find much more information about both projects on the course webpage (so please go and read it at least once!):

- Small project (to be submitted individually).
- Final project (to be submitted in groups).

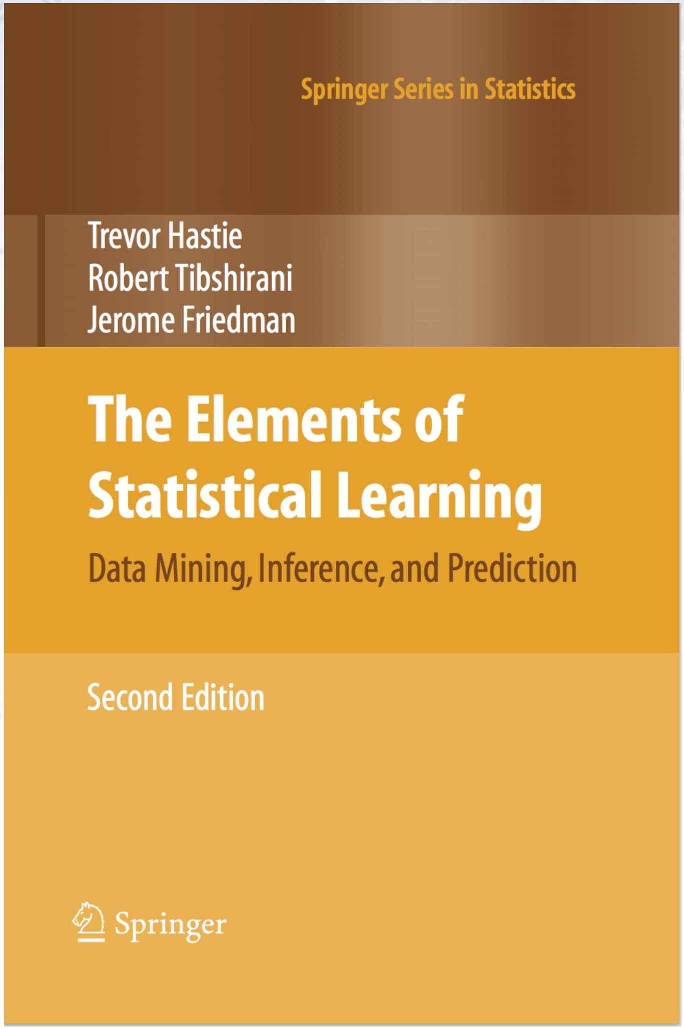
Literature

The main literature will be slides, notes, blogs, and links! However, we also wanted you to have a few more "solid" places to read comprehensively about ML.

"The Elements of Statistical Learning" (TEOSL) is a good read in PDF (though at times rather mathematical), and especially chapter 2 is a good introduction.

"Deep Learning" by Ian Goodfellow et al. in HTML is also good, and Chapter 5 of Part I gives a great overview of ML and its ingredients.

"Pattern Recognition and Machine Learning" by Christopher M. Bishop is also recommended, but it is not available on the web (for free).



Springer Series in Statistics

Trevor Hastie
Robert Tibshirani
Jerome Friedman

The Elements of Statistical Learning

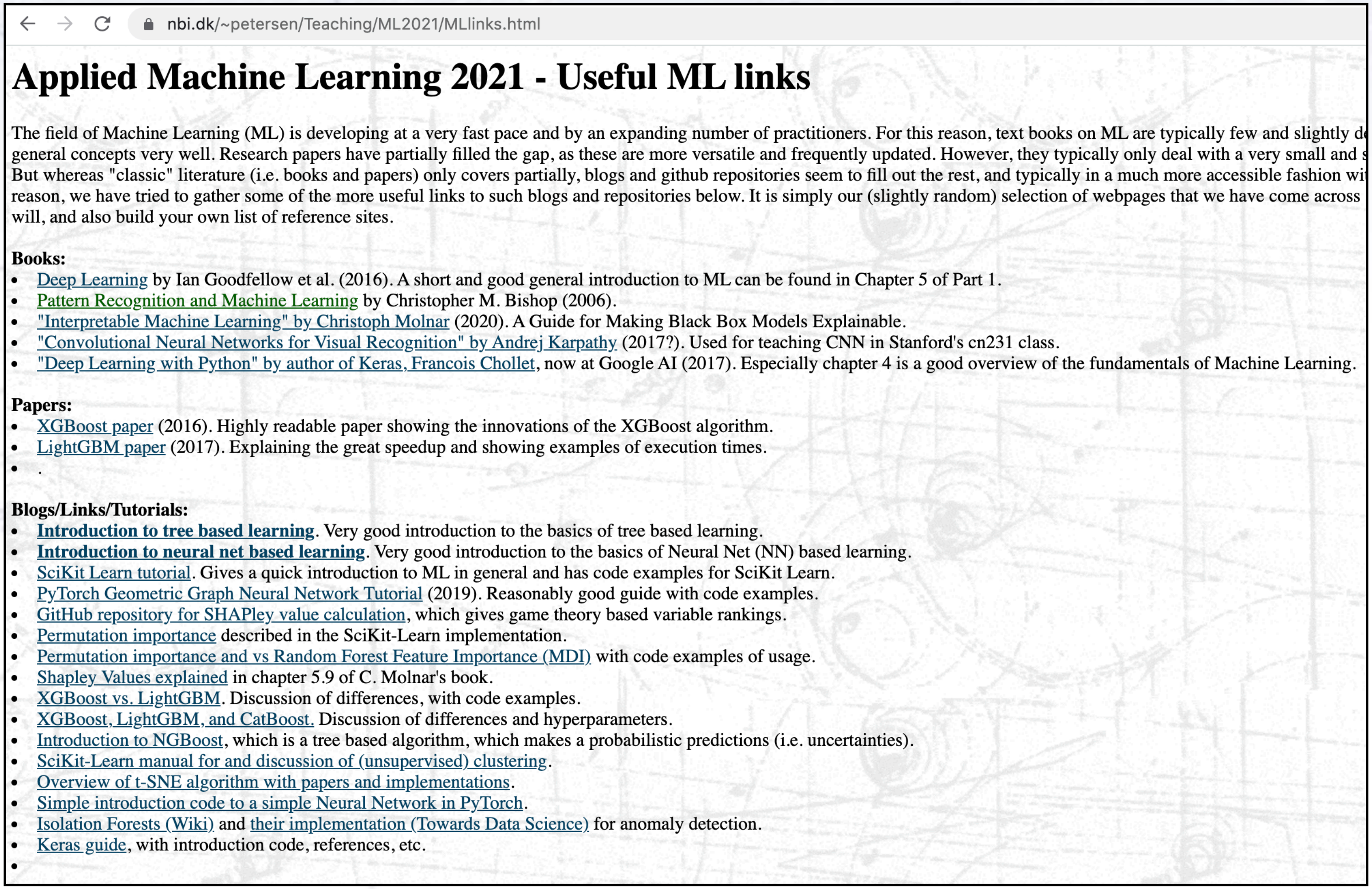
Data Mining, Inference, and Prediction

Second Edition

Springer

Blogs as literature

In ML, blogs/articles/tutorials are a very common (and great) source of literature on ML. For this reason, we've made a list of links that we find good:



Applied Machine Learning 2021 - Useful ML links

The field of Machine Learning (ML) is developing at a very fast pace and by an expanding number of practitioners. For this reason, text books on ML are typically few and slightly d
general concepts very well. Research papers have partially filled the gap, as these are more versatile and frequently updated. However, they typically only deal with a very small and s
But whereas "classic" literature (i.e. books and papers) only covers partially, blogs and github repositories seem to fill out the rest, and typically in a much more accessible fashion wi
reason, we have tried to gather some of the more useful links to such blogs and repositories below. It is simply our (slightly random) selection of webpages that we have come across
will, and also build your own list of reference sites.

Books:
- Deep Learning by Ian Goodfellow et al. (2016). A short and good general introduction to ML can be found in Chapter 5 of Part 1.
- Pattern Recognition and Machine Learning by Christopher M. Bishop (2006).
- "Interpretable Machine Learning" by Christoph Molnar (2020). A Guide for Making Black Box Models Explainable.
- "Convolutional Neural Networks for Visual Recognition" by Andrej Karpathy (2017?). Used for teaching CNN in Stanford's cn231 class.
- "Deep Learning with Python" by author of Keras, Francois Chollet, now at Google AI (2017). Especially chapter 4 is a good overview of the fundamentals of Machine Learning.

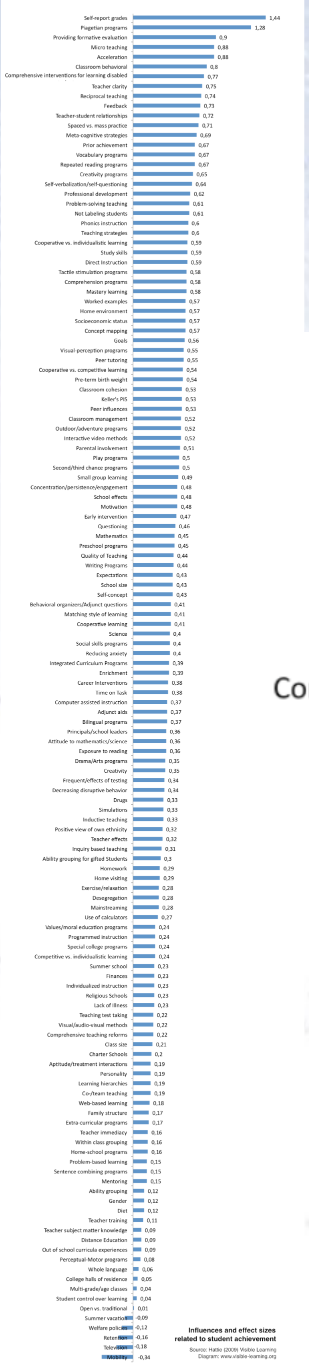
Papers:
- XGBoost paper (2016). Highly readable paper showing the innovations of the XGBoost algorithm.
- LightGBM paper (2017). Explaining the great speedup and showing examples of execution times.
- .

Blogs/Links/Tutorials:
- **Introduction to tree based learning**. Very good introduction to the basics of tree based learning.
- **Introduction to neural net based learning**. Very good introduction to the basics of Neural Net (NN) based learning.
- SciKit Learn tutorial. Gives a quick introduction to ML in general and has code examples for SciKit Learn.
- PyTorch Geometric Graph Neural Network Tutorial (2019). Reasonably good guide with code examples.
- GitHub repository for SHAPley value calculation, which gives game theory based variable rankings.
- Permutation importance described in the SciKit-Learn implementation.
- Permutation importance and vs Random Forest Feature Importance (MDI) with code examples of usage.
- Shapley Values explained in chapter 5.9 of C. Molnar's book.
- XGBoost vs. LightGBM. Discussion of differences, with code examples.
- XGBoost, LightGBM, and CatBoost. Discussion of differences and hyperparameters.
- Introduction to NGBoost, which is a tree based algorithm, which makes a probabilistic predictions (i.e. uncertainties).
- SciKit-Learn manual for and discussion of (unsupervised) clustering.
- Overview of t-SNE algorithm with papers and implementations.
- Simple introduction code to a simple Neural Network in PyTorch.
- Isolation Forests (Wiki) and their implementation (Towards Data Science) for anomaly detection.
- Keras guide, with introduction code, references, etc.
-

What influences learning?

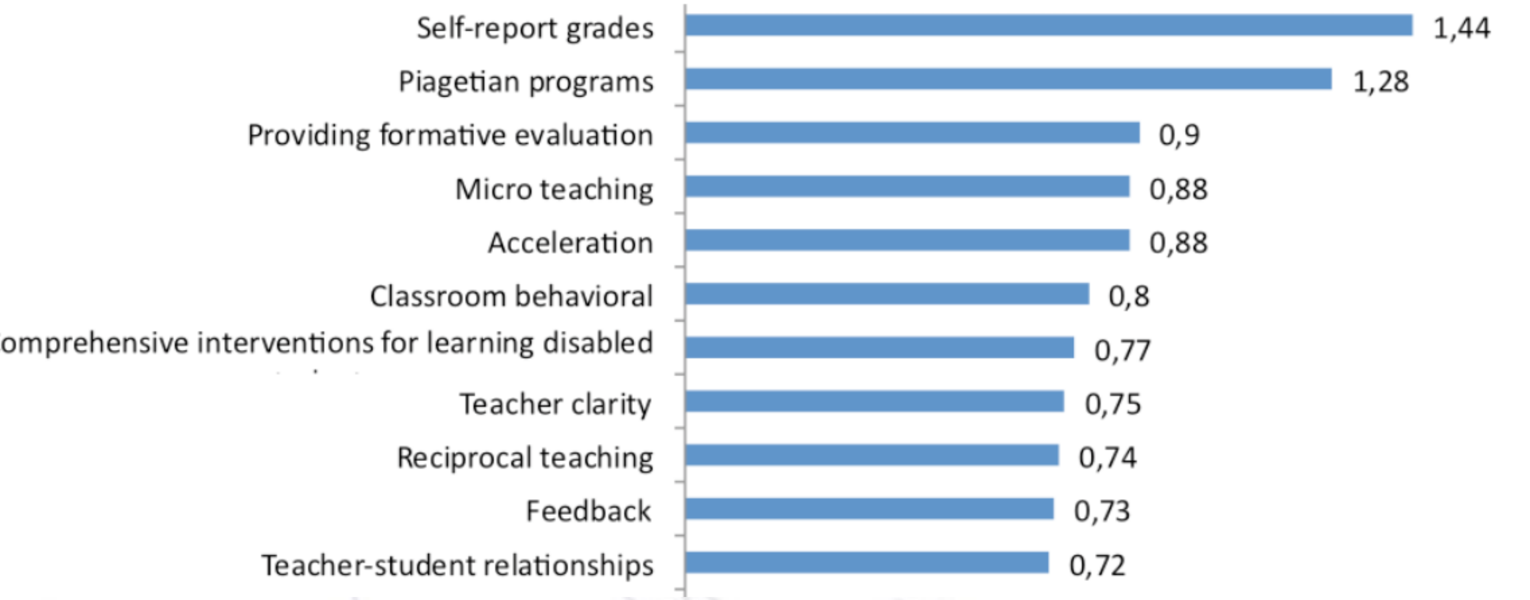
What influences learning?

There are studies of this, one result shown below:



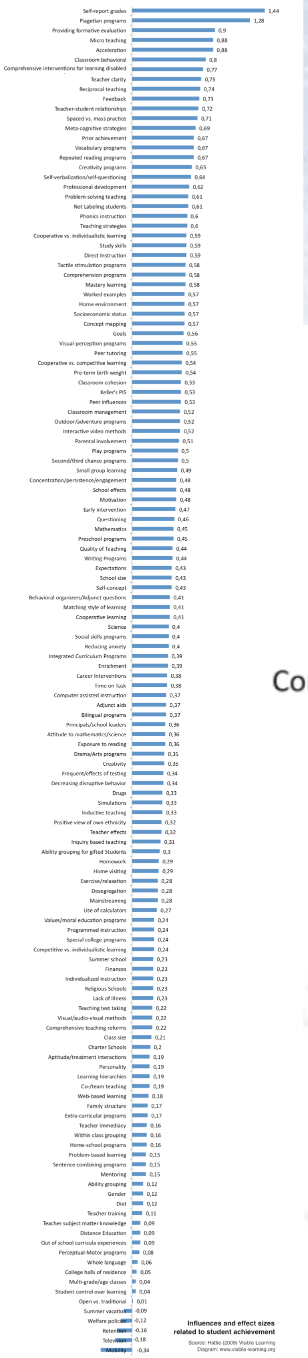
Influences and effect sizes related to student achievement

Source: Hattie (2009) Visible Learning
Diagram: www.visible-learning.org



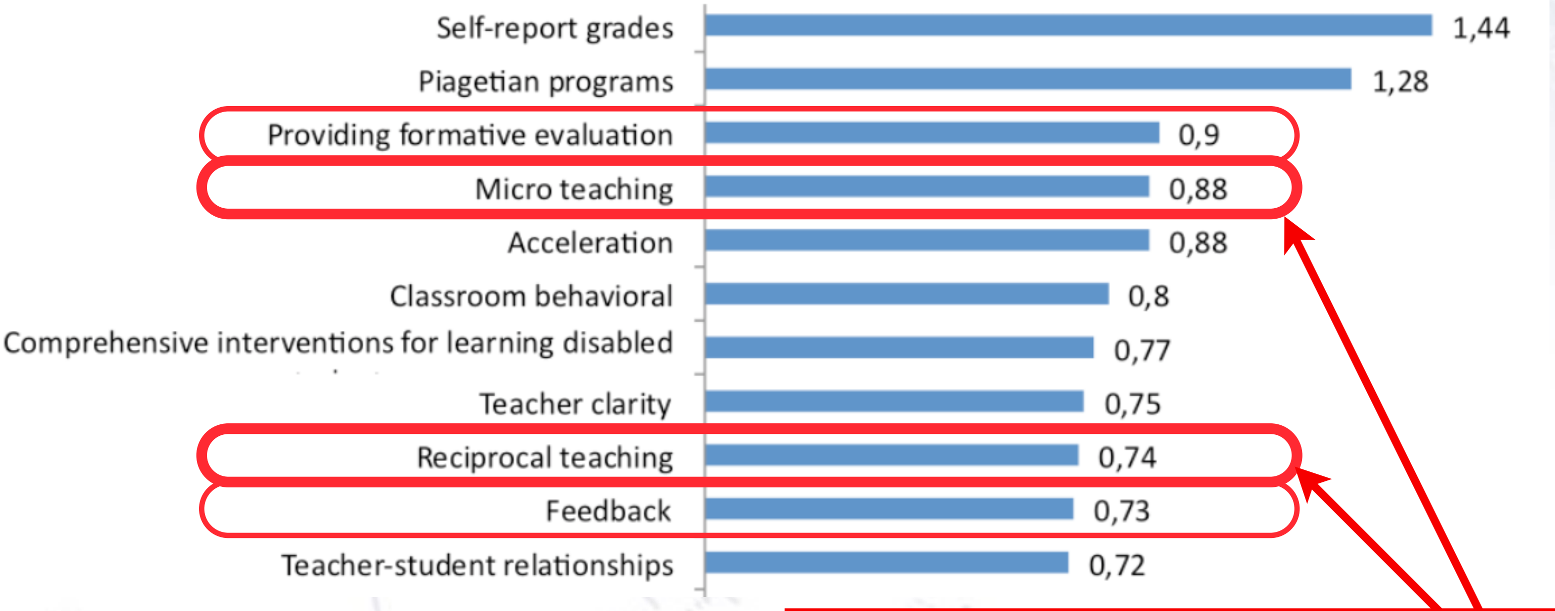
What influences learning?

There are studies of this, one result shown below:



Influences and effect sizes related to student achievement

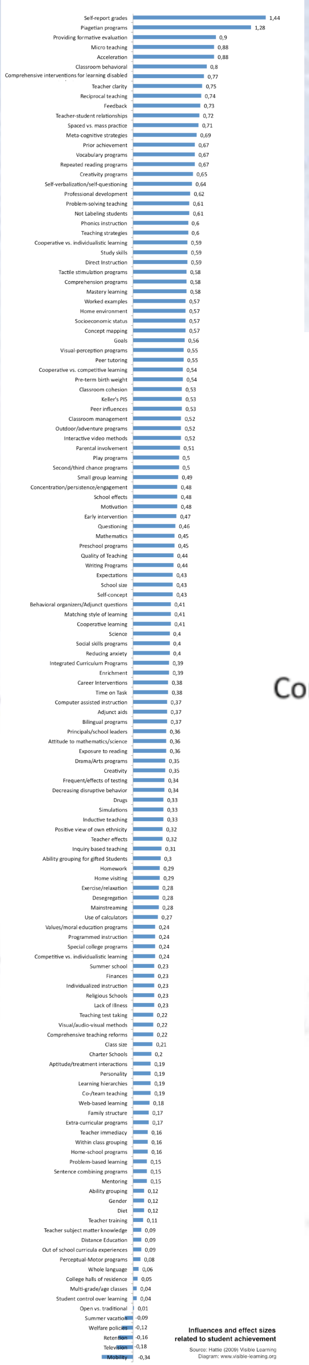
Source: Hattie (2009) Visible Learning
Diagram: www.visible-learning.org



Learning lead by / among students is among the **most effective ways**!

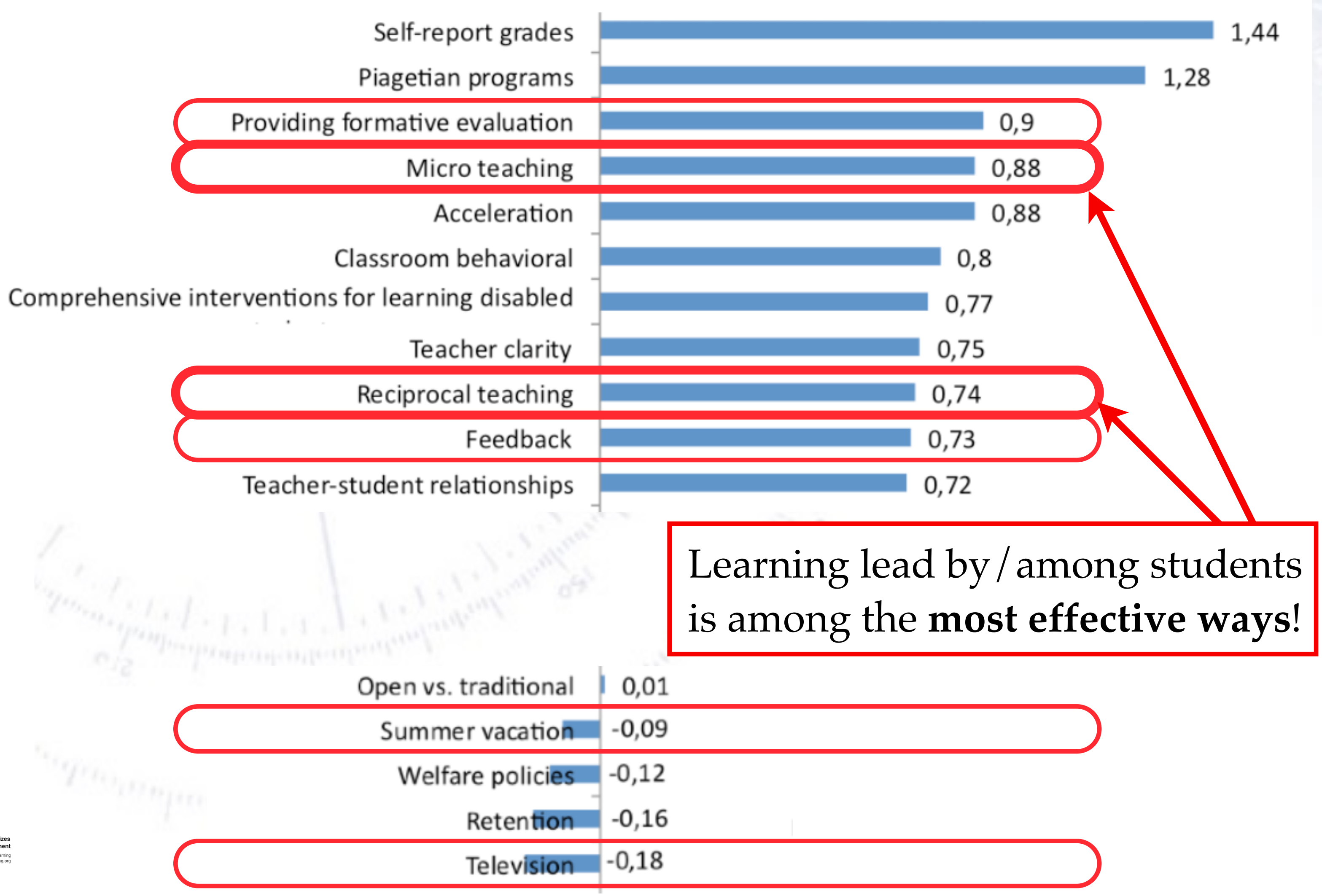
What influences learning?

There are studies of this, one result shown below:



Influences and effect sizes related to student achievement

Source: Hattie (2009) Visible Learning
Diagram: www.visible-learning.org



Learning lead by / among students is among the **most effective ways**!

Expectations

We want (read: insist) this course to be useful to all of you! Therefore, please give us feedback (the earlier the better), if you have anything to add/suggest/criticise/alter.

However, it is also a VERY independent course in the sense that it is up to YOU, how much you get out of it. Consider it rather a project than a course!

The aim is to make you knowledgable about the basics of Machine Learning, and being able to apply it to (suitable) data.

Problems?

If you experience problems in relation to the course, whatever their origin and nature, then write us!

We may not be able to do anything about it, but if we don’t know about your problems, then I most certainly can not do anything about them.

We consider ourselves fairly large, as long as I feel that this largeness is met by sincerity and will.

But… you need to write us in the first place! That is your responsibility.

“Artificial intelligence is just a new tool, one that can be used for good and for bad purposes and one that comes with new dangers and downsides as well. We know already that although machine learning has huge potential, data sets with ingrained biases will produce biased results – garbage in, garbage out.”

[Sarah Jeong, US tech journalist]