# Applied Machine Learning & Big Data

## Course information 2020

Troels C. Petersen, Adriano Agnello,
Brian Vinter, Zoe Ansari, Carl Johnsen

# Applied Machine Learning & Big Data

## Course information 2020

Troels C. Petersen | Adriano Agnello | Brian Vinter | Zoe Ansari | Carl-Johannes Johnsen

Troels C. Petersen, Adriano Agnello, Brian Vinter, Zoe Ansari, Carl Johnsen

# 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.***

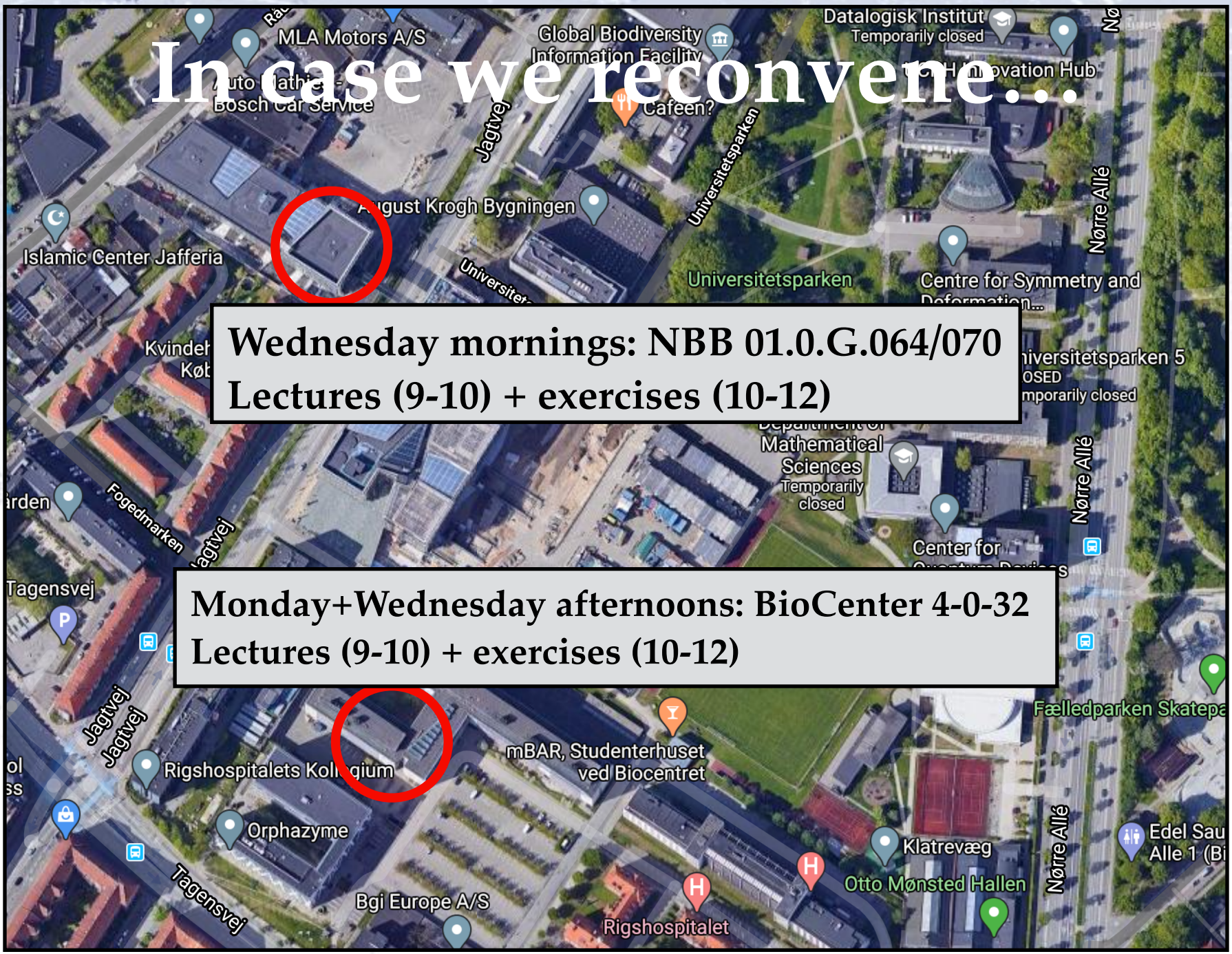
So please make small teams of peers, with whom you can discuss the many details of ML coding and the problems, challenges, and issues involved. This is you best way of discussing with peers, learning most, and not getting stuck.

# General words on the course

It may be, that the whole course will be run virtually, and that we'll never get to meet in person!

Personally, I hope (and trust) that this will not be the case. But if you for "some reason" will not be able to attend in person, if/when we reconvene, then we will of course let you continue to follow the course, and do everything in order to make this as easy as possible.

# In case we reconvene...



# Additional locations

# Computers and software

We will program (mainly) in Python, but in fact **the choice is your own**. We will only provide data, code snippets (in Python) and inspiration.

We suggest that you use Jupyter Notebook, possibly on ERDA, but make sure that you can also run everything on your own laptop, so that you are not dependent on internet and ERDA access. 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).

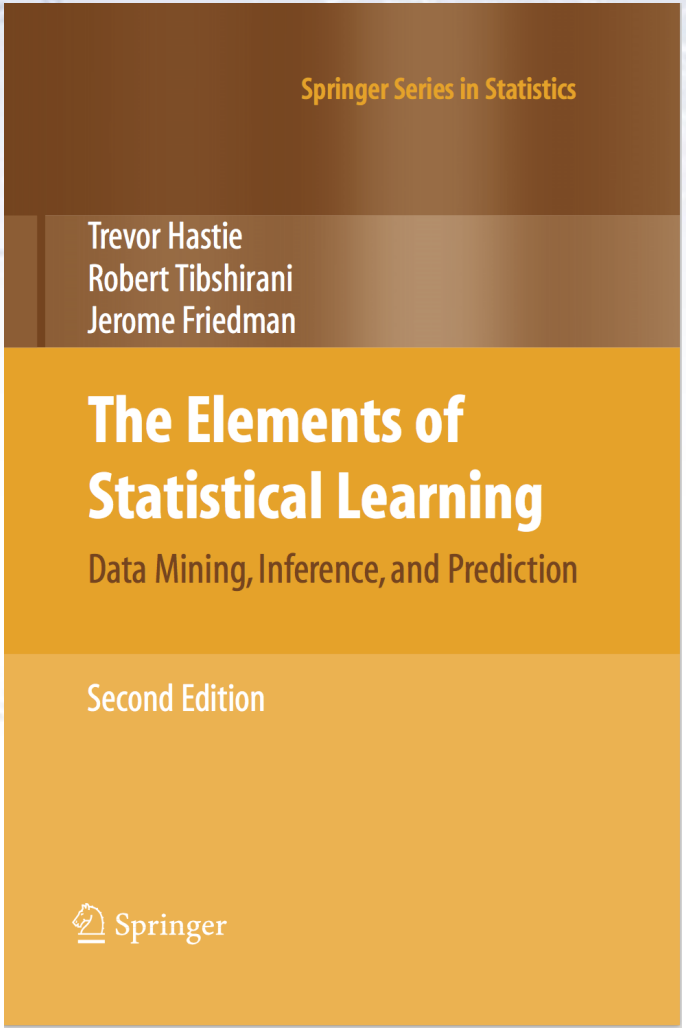
# Literature

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).



# 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.

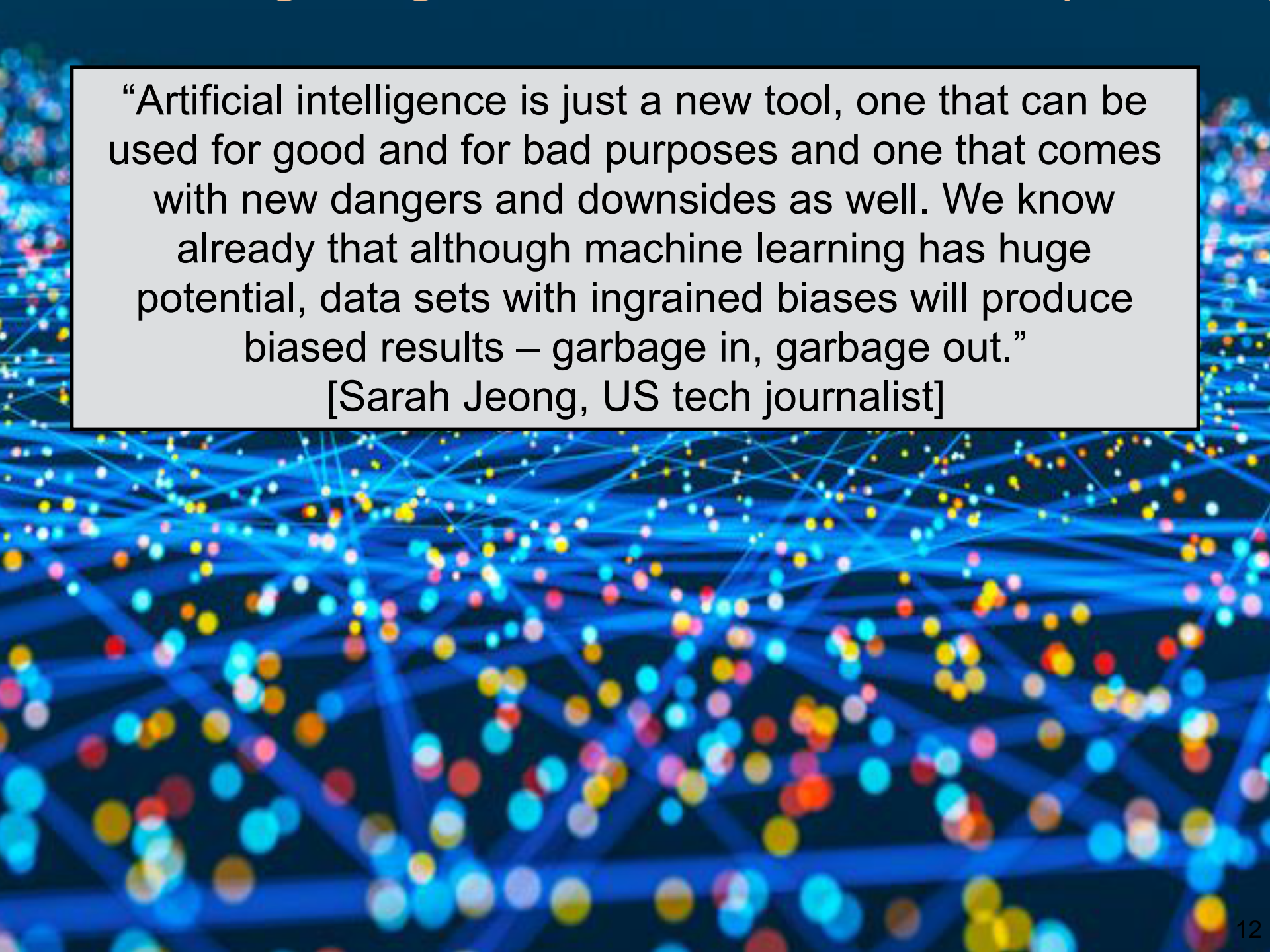
# Problems?

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]