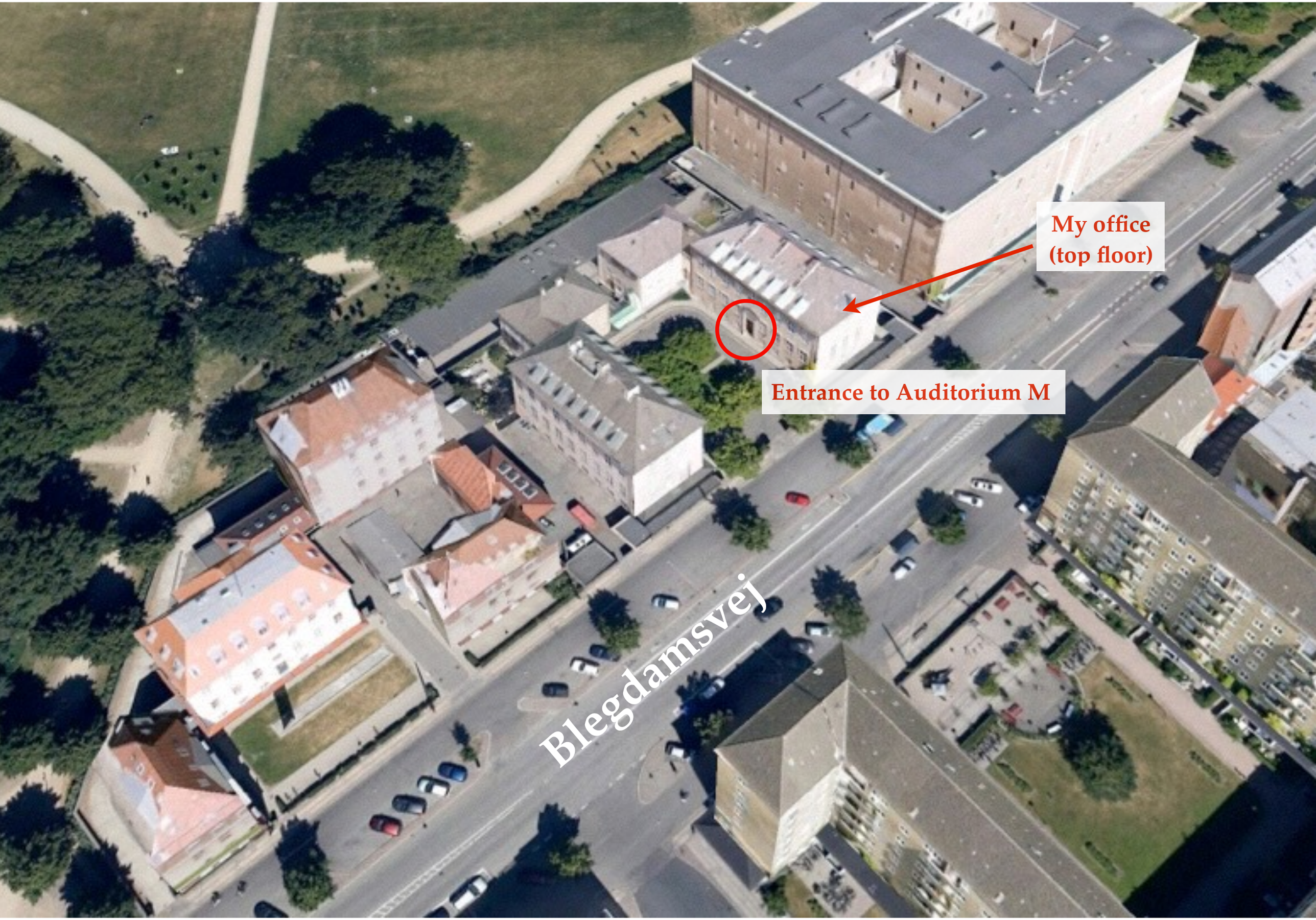
Course Information

D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2016

Photo by Howard Jackman


My office
(top floor)
Entrance to Auditorium M
Blegdamsvej

Times & Locations

From 08:30-09:00 I will be in Aud. M
but I will only be there for discussion and
student Q&A, i.e. no new material until 09:00

Actual
08:00-08:30 Study
08:30-09:00 Study/Q&A
09:00+ Lecture

- Hours
  - Course is in Block A
  - Tuesday 08:00 - 12:00
  - Thursday 08:00 - 12:00 and 13:00 - 17:00
- Location
  - Auditorium M at Blegdamsvej
  - Possibly the adjacent room to Aud. M for doing group work
- In-class Activities
  - ~20-30% of the time will be lecture
  - Most of the time will be practical exercises
    - Finding appropriate software package or function
    - Properly instantiating the relevant statistical method
    - Debugging
    - Documentation, plotting, code clean-up, maybe even in-line comments, etc.

Me

- I go by “Jason”, although ‘incredible Dr. Koskinen’ will also work
- My scientific focus is on experimental neutrino oscillation, where I work on the IceCube neutrino observatory situated at the South Pole

Computers & Software

- Everyone should have a laptop
- Software specifics are left to the student
  - Suggestions are python, R, or C/C++
    - The more common the language the more likely you can use the internet for help
  - Lectures and examples will be mostly in python using some external packages:
    - ROOT (https://root.cern.ch), note that ROOT is not necessary and can be painful to install
    - SciPy (http://www.scipy.org)
    - NumPy (http://www.numpy.org)
    - PyMC for Markov Chain Monte Carlo (https://pypi.python.org/pypi/pymc)
- I encourage you to find at least one person who is nominally using the same setup, e.g. Windows 10, R 3.2.3
  - I'll post online who's expecting to use certain software

Software, Checklist, & Skills

- Have an installed text editor for writing/editing software
- Have some package for the production of plots and diagrams (ROOT, matplotlib, R, gnuplot, Matlab, etc.)
- Provide:
  - Name, and preferably a photo too so I know whom you are
  - Area of physics research or interest
  - Anticipated software usage
  - Preferably in this google doc (http://tinyurl.com/j5pjvf3) or by email (koskinen@nbi.ku.dk) to me
- I strongly recommend reviewing the undergraduate stats course (http://www.nbi.dk/~petersen/Teaching/AppliedStatistics2015.html)
  - Actually do the exercises, not just scan the material, and see what isn't clear or familiar

Course Material

- NO required text or textbooks. I will cover many topics w/ in-class lectures where the notes will be posted online. But, this may be insufficient in depth or explanation for your personal preference, so students can use…
- The Internet
  - Seriously. Probably the best source for information and help.
- "Statistical Data Analysis" by Glen Cowan
- "Modern Statistical Methods for Astronomy" by Feigelson & Babu
- Journal articles
  - Any that you might find relevant
  - Some posted by me

Student Assessment

- Oral Presentation and 1-page summary (10%)
  - Take topic and/or relevant article for presentation to the class
  - Can be done in groups, but no more than 3 people
- Graded problem sets (15%)
  - Can be done in groups of any size, but must be submitted individually, along with all code (unzipped)
- Project (25%)
  - Larger data analysis project, nominally related to your field of research
  - Can be done in groups (no more than 3 people) with a single 4-6 page written report
- Final Exam (50%)

Student Assessment

- All assessment material will be graded based on the results
  - Code can be sloppy and inefficient and will not affect your grade
  - Obviously, the exception is cheating, e.g. using the 97% of the code of someone else and only changing the comments , variable names, line colors, etc.
- I encourage you to share solutions, efficient code, elegant solutions, etc. for everything other than the final exam
  - If you use a portion of someone else's code (which is fine by me) make in-line acknowledgement in the comments of your code
  - Beware, that if your code starts to look like a collection of only other people's code, it's unlikely that the Final Exam will go well

Final Exam

- 1-day (~24 hour) take home test
- Requires computers, writing/modifying code that has been developed during the course
- **You must work on your own!!!**
  - Along with the answers, the code producing the results must also be submitted
- Here is one topic which is guaranteed to be on the Final Exam. I will generate a joint dataset that is comprised of at least 3, but no more than 20, different underlying distributions, where you will be asked to:
  - Find out how many distinct distributions, their types (gaussian, exponential, linear, step-wise, etc.), and properties of the underlying distributions
  - The data set will have at least thousands of entries, probably multivariate, and I may include 'experimental' systematics or biases

Challenges

- Multiple student backgrounds and multiple topics mean that some students may feel like they would benefit from more challenging material… have no fear
- I have collected some projects/questions from colleagues
- No guarantee that undertaking any advanced challenges will result in a better grade. Similarly, students may earn top marks in the course without ever looking at extra topics
- Potentially pick something on your own and discuss it with me, maybe even put together some lecture material and add it to the course

Expectations

- As graduate students, the days ‘of being spoon fed information from a text book or professor’ are ending
- Software and hardware difficulties and solutions are the sole domain of each student. You can do all the projects on a PlayStation 4 w/ screenshots if you want.
- These are nominally advanced topics
  - Open to suggestions about topics
  - In the absence of unassailable experts, discussion and participation by individuals and groups is important
- Leave the course with at least 1 new tool that you can use in your research