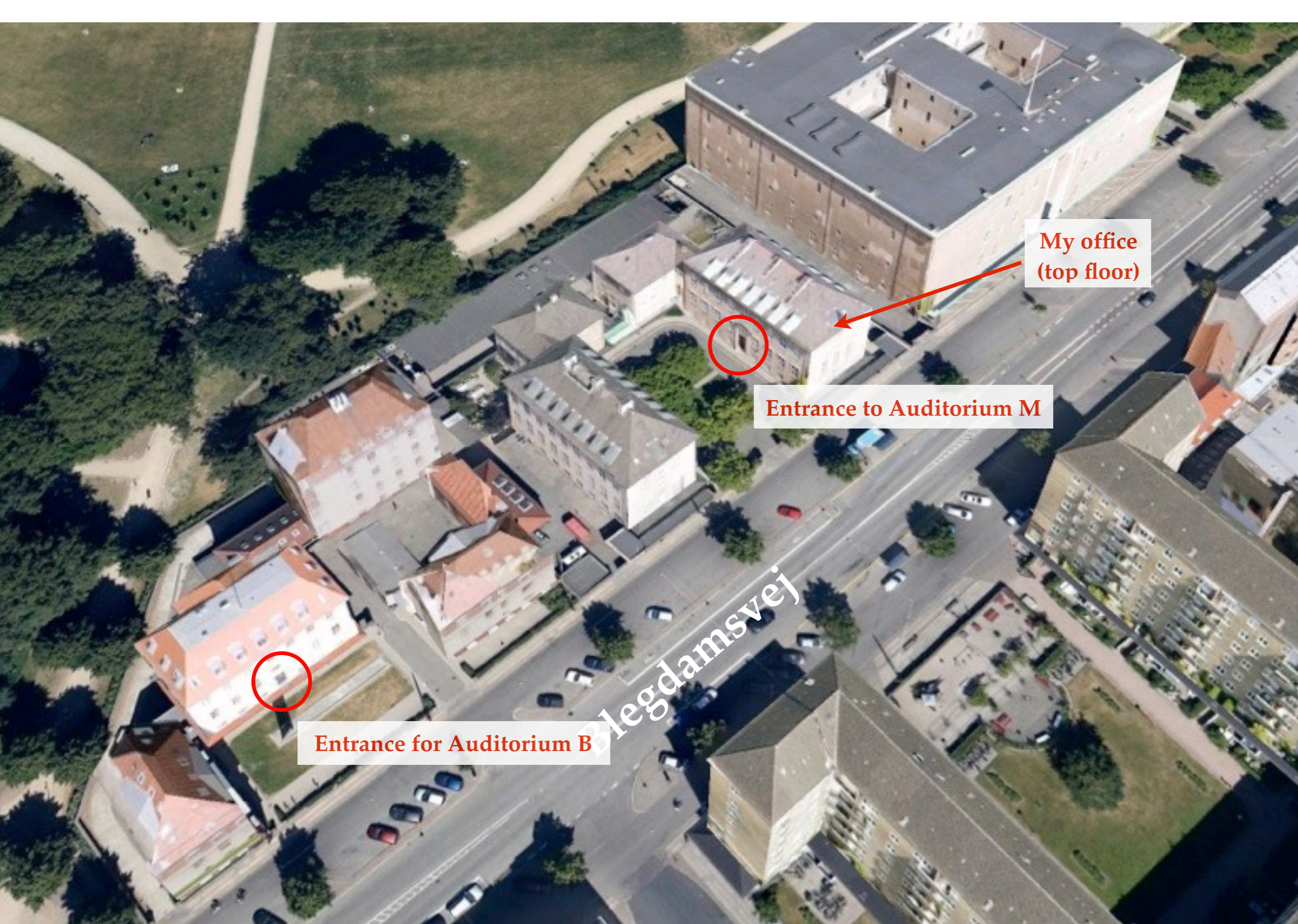


Week 0: Data Handling and Software Fluency

D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2018



My office
(top floor)

Entrance to Auditorium M

Entrance for Auditorium B

Blegdamsvej

Me



- I go by "Jason"
- My scientific focus is on experimental neutrino oscillation, where I work on the IceCube neutrino observatory situated at the South Pole

Software Packages

- Some of the methods we will use in the course will require software packages that include:
 - Minimizers: for example BFGS, MIGRAD, SIMPLEX, etc.
 - Markov Chain Monte Carlo
 - Spline routines for interpolation, including basis splines (b-splines)
 - Multi-Variate Machine Learning: boosted decision trees, neural networks, support vector machines, etc. (we will for sure cover boosted decision trees)
- Other more specialized uses I will let you know about in advance of the lecture
 - MultiNest nested sampling algorithm

More Specifically

- Below I will list the needed packages and some python options
- Plotting
 - I use ROOT from CERN, but it's only common for 'old' people
 - Matplotlib is what the 'cool' kids seem to use
- For Python users, I'm a big fan of "Jupyter" notebooks
 - Combination of both text fields, inline figures/plots display, and executable code
 - Great way to keep things organized
- Minimizer Routines
 - I normally use MINUIT2 (via iminuit)
 - SciPy has a minimize function with a bunch of algorithms and is more common nowadays

More Specifically

- Markov Chain Monte Carlo
 - I have used PyMC, but other packages such as MCMC, emcee, or Nestle look like better tools
- Multi-Variate Analysis (MVA)
 - I used the ROOT software from CERN (TMVA)
- Splines
 - SciPy has an interpolate function and other spline options
- Bayesian Inference Sampling - MultiNest
 - pymultinest
- Even if you're using python, you don't **need** any of the above mentioned *specific* packages, e.g. iminuit.

Software and Data Handling

- As a precursor to doing computer aided statistics, the first problem set will focus on data handling, parsing text, writing code, and simple presentation
- Exercises will focus on USA college basketball statistics from the 2014 Ken Pomeroy Basketball page at <http://kenpom.com/index.php?y=2014>
 - The content is largely **irrelevant** and was chosen due to some *fairly evil...* I mean *interesting* features
- This will be potentially time-consuming
 - It took me ~4 hours to originally produce all the results
 - Had I stored/handled the data in a different format it would have gone much quicker
 - Could take as little as 15 min.

First Assignment

- Conceptually this is a simple assignment
 - No advanced or even difficult statistical methods or analyses
- The goal of the first assignment is to assess how well people can load, analyze, and plot data
 - Essentially a plotting and data throughput exercise
 - But, there are some interesting data features
- Words of advice for the following problem set
 - Don't be overly reliant on spreadsheets
 - Don't assume that the input data (or format) is stable between years for exercises 2 and 3
- There are some known (at least by me) ambiguities in the exercises. If you come across what you perceive is an ambiguity, detail it in your write-up.

Problem Set Submission

- The submission is:
 - A write-up as a PDF document, which includes any plots, diagrams, tables, pictures, and explanations
 - In a separate “file”, submit all code used to derive the results
 - Tarball, zipped directory, lots of individual files w/ self-explanatory titles, etc.
 - Include data files
- Material is marked on a 10-point scale
 - 9+ is very good
 - 8-9 is pretty good
 - 7-8 is okay
 - 6-7 is acceptable
 - 5-6 subpar
 - 4-5 inadequate
 - <4 reflects serious omissions and/or deficiencies

Assignment

Exercise 1

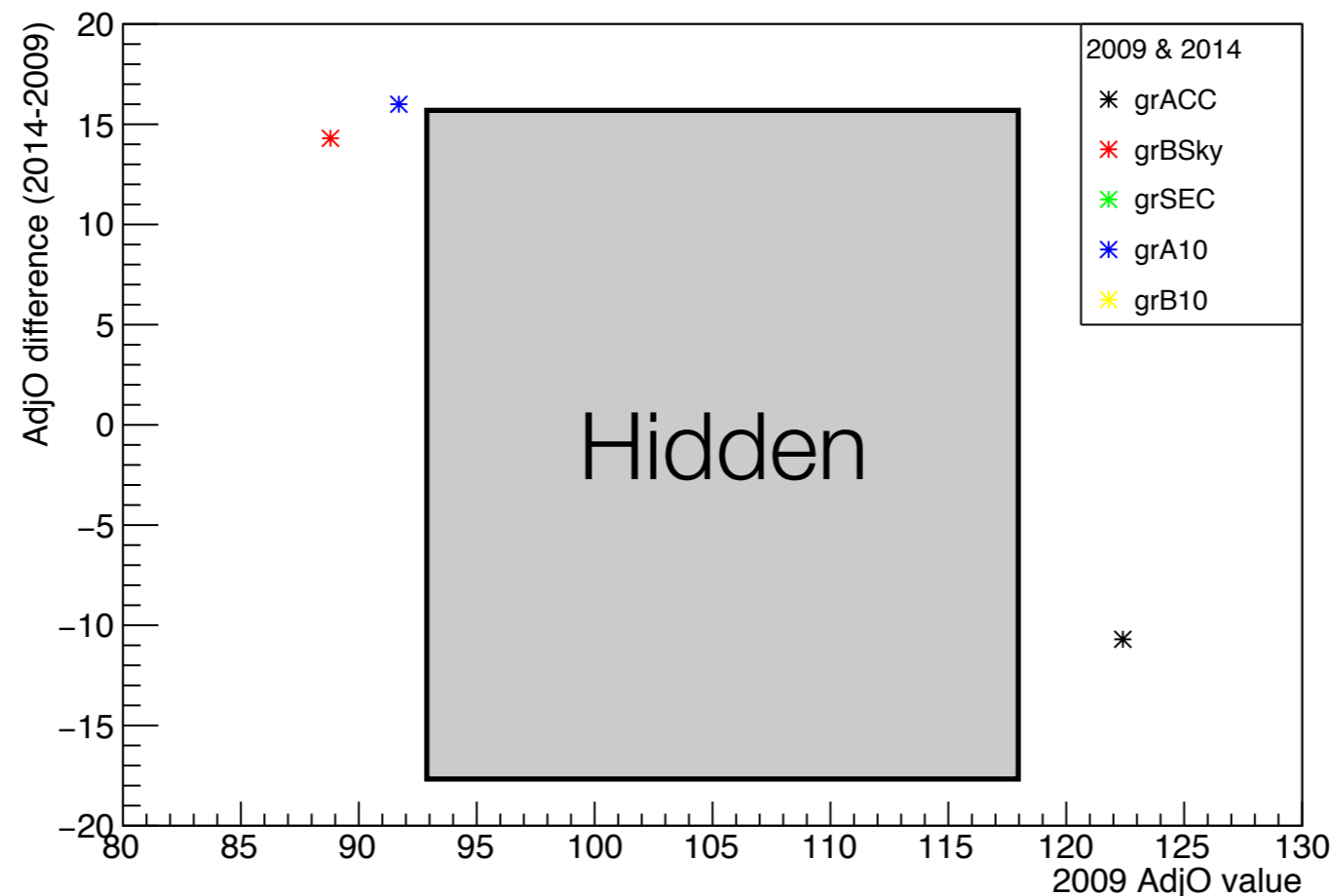
- Take the 2014 Ken Pomeroy data related to NCAA College Basketball analytics from <http://kenpom.com/index.php?y=2014>
- On a single plot draw histograms of:
 - The Adjusted Defense "AdjD" for all the teams in the 5 conferences (ACC, SEC, B10, BSky, and A10)
 - Different colors for each conference and add a legend
 - Save as a PDF

Exercise 2

- Take the 2014 and 2009 Ken Pomeroy data related to NCAA College Basketball analytics
- Calculate the difference in "AdjO" for all the teams in the 5 conferences from Exercise 1:
 - 2014 minus 2009 as a function of the 2009 AdjO value
 - Plot the data as a graph with a data point for each team entry being the same conference color as for the previous histogram in Exercise 1
- Calculate the difference in "AdjO" for all the teams with data in both 2009 and 2014
 - Median and mean for each of the 5 conferences
 - Median and mean for teams that were not in the 5 conferences

Exercise 2 cont.

- Calculate the difference in "AdjO" for the teams in each of the 5 conferences from Exercise 1:
 - Between 2009 and 2014 as a function of the 2009 AdjO value
 - Plot the data as a graph with a data point for each team entry being the same conference color as for the previous histograms



*Be mindful that this plot is an example and is not guaranteed to be accurate

Exercise 2 cont.

- Calculate the difference in "AdjO" for all the teams with data in both 2009 and 2014
 - Median and mean for each of the 5 conferences
 - Median and mean for teams that were not in the 5 conferences

```
AdjO median difference (2014-2009) for conference ACC w/ teams is
The mean is -0. [redacted]
AdjO median difference (2014-2009) for conference BSky w/ teams is
The mean is 3. [redacted]
AdjO median difference (2014-2009) for conference SEC w/ teams is
The mean is 2. [redacted]
AdjO median difference (2014-2009) for conference A10 w/ teams is
The mean is 4. [redacted]
AdjO median difference (2014-2009) for conference B10 w/ teams is
The mean is 2. [redacted]

AdjO median for [redacted] non-conference teams was 2. [redacted]
The mean is 3. [redacted]
```

Note: there is no guarantee that the first significant digit on this screen is correct.

Exercise 3

- Take the 2014 and 2009 Ken Pomeroy data related to NCAA College Basketball analytics
- Redo Exercises 1 and 2, while now adding the “BE” conference to the previous list of 5 conferences
 - For those who have written robust code, this should be trivial
 - It is likely to be much harder for those whom...
 - Parse some data in by hand
 - Only wrote code that requires the exact data format specific to the team names, conferences, AdjO/AdjD position, etc.

Problem Set Submission

- Due by Wednesday Feb. 14 at 08:30 CET
- In a single email to me (koskinen@nbi.ku.dk):
 - Submit the results, plots, numbers, text, etc. by email in a **single** PDF document
 - The submitted PDF document should not contain any code
 - In a separate file include the code, however terrible, broken, crashing, unpretty, or uncommented in the same email
 - Unless you parse directly from the internet HTML, also include the data files you actually used. Sometimes files can change, so please supply the one you are actually using.

Exercise 4 (Extra 1 pt.)

- One of the most important observations in astronomy was recently made with the coincident observation of gravitational waves in addition to photons across a wide range of wavelengths from a binary neutron star merger
- There is an author list at <http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2018/data/authors-acknowledgements-v5.pdf>
 - How many unique authors are there in that list?
 - If there was one single author list in alphabetical order (instead of being grouped by experimental collaboration), what author is the mid-point
 - Who is at the location $(\text{total authors})/2$. Potentially there are two authors depending on whether the total number of authors is an odd or even number.