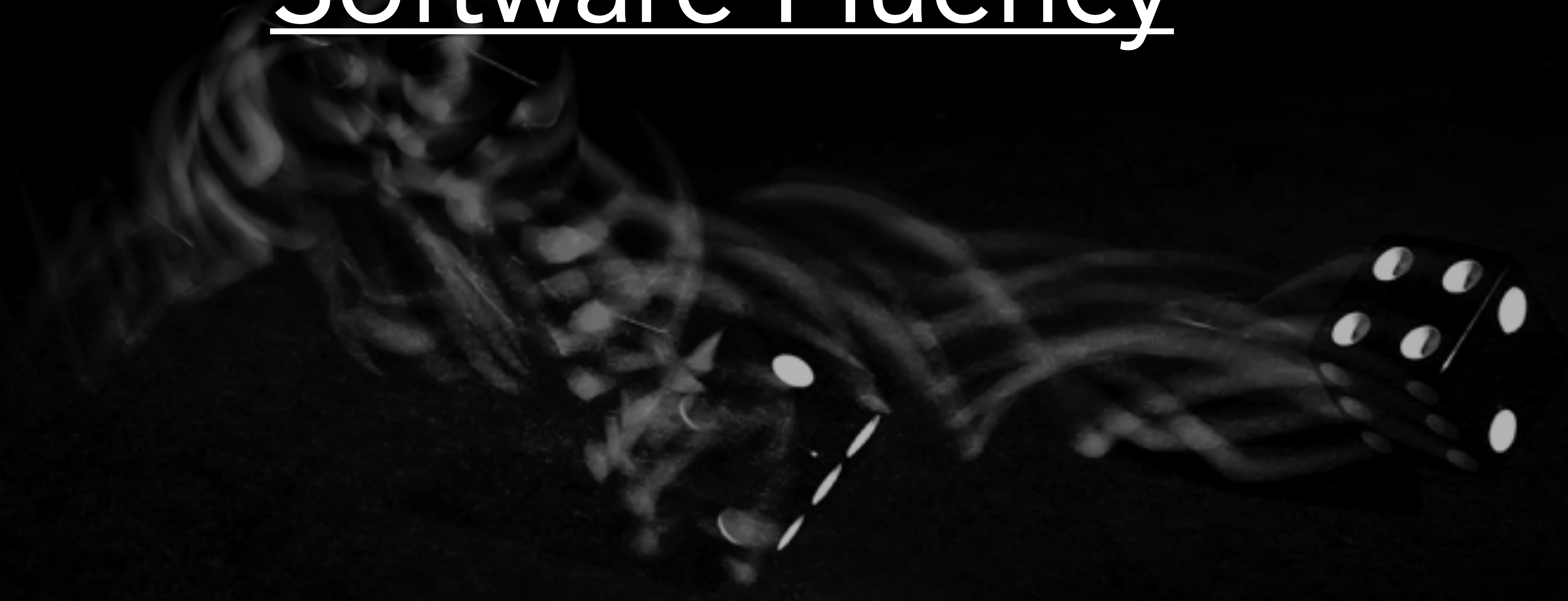
Week 0: Data Handling and Software Fluency



D. Jason Koskinen

koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics

Feb - Apr 2019

Photo by Howard Jackman

Instructors

- Main Teacher: D. Jason Koskinen
- My scientific focus is on experimental neutrino oscillation, where I work on the IceCube neutrino observatory situated at the South Pole

- Teaching Assistant: Jean-Loup Tastet
- Astroparticle and particle physics phenomenology, specifically heavy neutral leptons

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
  - Matplotlib is a conventional choice
- 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)
  - In past years many people have switched to Scikit-learn
- Splines
  - SciPy has an interpolate function and other spline options
- Bayesian Inference Sampling - MultiNest
  - pymultinest, Nestle
- 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 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 Jason and Jean-Loup) 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