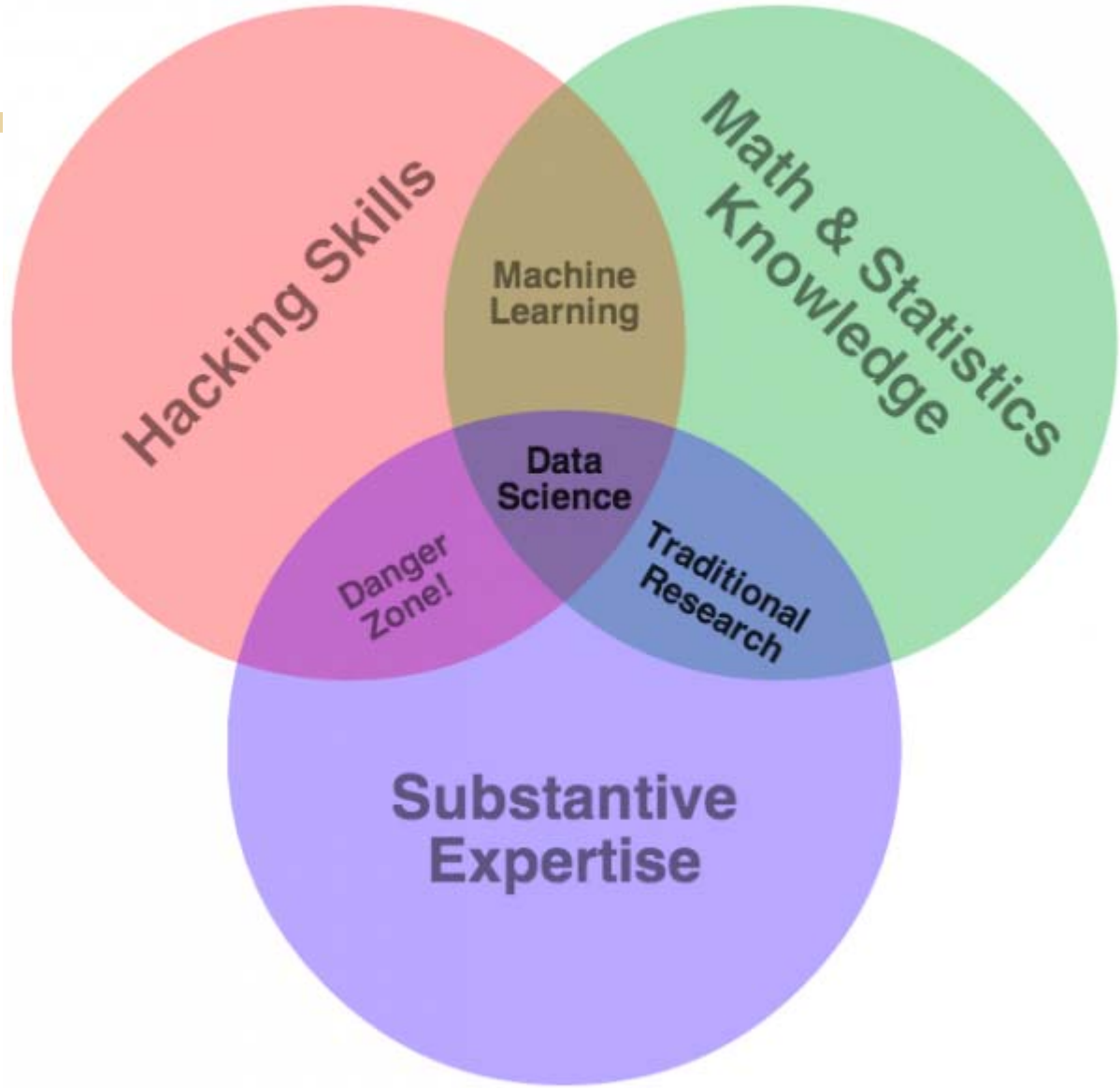
UNIVERSITY of WASHINGTON

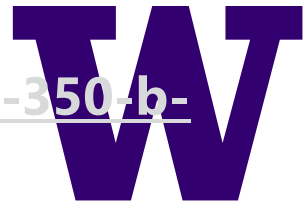
Data Science UW
Methods for Data Analysis

Introduction and Data Exploration
Lecture 1
Stephen Elston


Hacking Skills
Math & Statistics Knowledge
Substantive Expertise
Machine Learning
Data Science
Danger Zone!
Traditional Research

Course Purpose

- This course focuses on essential concepts
- We are building foundations for your data science skills
- Course Objectives:
  - Become comfortable working with structured and unstructured data.
  - Learn methods to explore and understand data.
  - Understand the core concepts of statistics and probability.
  - Understand and implement various statistical procedures in R.
  - Understanding the mathematical basis of machine learning models.
  - Expand R programming skills to be able to write and test quality code from scratch.
- See syllabus for more information:
  - https://canvas.uw.edu/courses/1105274/pages/datasci-350-b-course-syllabus/edit



Course Requirements and Grading

This course will be graded by attendance, homework, and an individual project.

- Attendance: You MUST attend at least 8 out of 10 classes. **This is a non-negotiable UW requirement**.
- Homework must be completed by the start of the next class. (Assigned weeks 1-8).
  - Returned as a 0,1, or 2.
    - 0 = Not done or a major parts missing.
    - 1 = Completed, but missing or serious errors.
    - 2 = Completed with minor issues. Demonstrates full understanding of subject.
- Individual Project: Due at the start of the last class.
  - Counts as 8 points.

Course Requirements and Grading

There is a total of 24 possible points. (16 pts for hmk + 8 project)

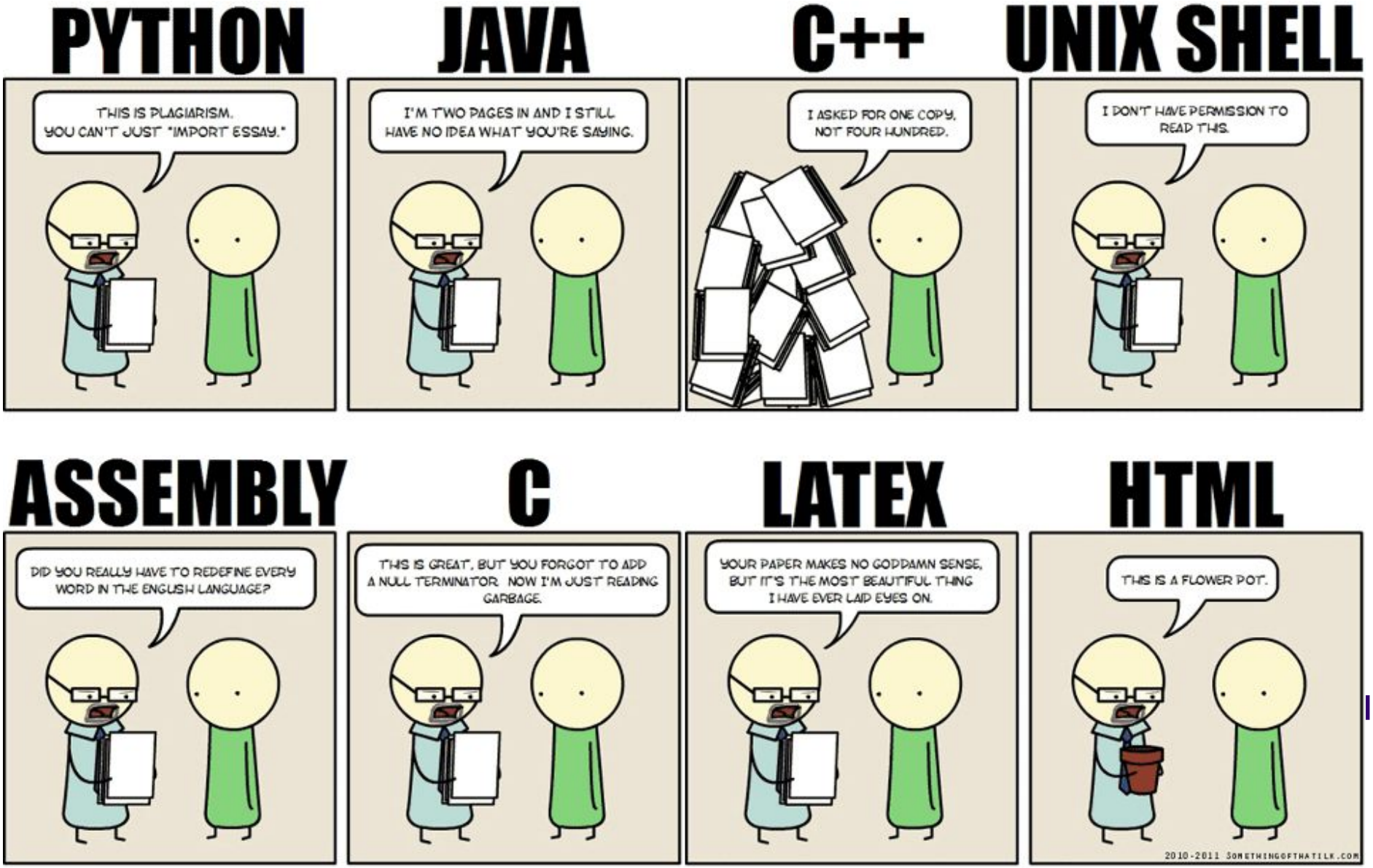
- Must get 18 total points to pass.
- All homework assignments must use good R coding technique
- Results must be presented in a professional style
- The individual project must be production level code.

Office Hours and Contact Information

- Contact me at:
  - stephen.elston@quantia.com
- When I'm *usually* available:
  - Off/on for simple things during work. (M-F 8am-5pm PST)
  - Sunday various afternoon/evening times.

Emergency contact: 402-980-3192

Review


PYTHON
THIS IS PLAGIARISM. YOU CAN'T JUST "IMPORT ESSAY."
JAVA
I'M TWO PAGES IN AND I STILL HAVE NO IDEA WHAT YOU'RE SAYING.
C++
I ASKED FOR ONE COPY, NOT FOUR HUNDRED.
UNIX SHELL
I DON'T HAVE PERMISSION TO READ THIS.
ASSEMBLY
DID YOU REALLY HAVE TO REDEFINE EVERY WORD IN THE ENGLISH LANGUAGE?
C
THIS IS GREAT, BUT YOU FORGOT TO ADD A NULL TERMINATOR. NOW I'M JUST READING GARBAGE.
LATEX
YOUR PAPER MAKES NO GODDAMN SENSE, BUT IT'S THE MOST BEAUTIFUL THING I HAVE EVER LAID EYES ON.
HTML
THIS IS A FLOWER POT.
2010-2011 SOMETHINGOFTHATILK.COM

Languages for data science

- Skills every data scientist should have
- SQL is the 'Linqua Franka' of data access
- R – widely used for visualization, statistical analysis, and machine learning
  - We use R in this course
- Python 3 – widely used for visualization, machine learning, big data APIs (e.g. Spark), deep learning APIs
  - Example for visualization: https://github.com/StephenElston/ExploringDataWithPython

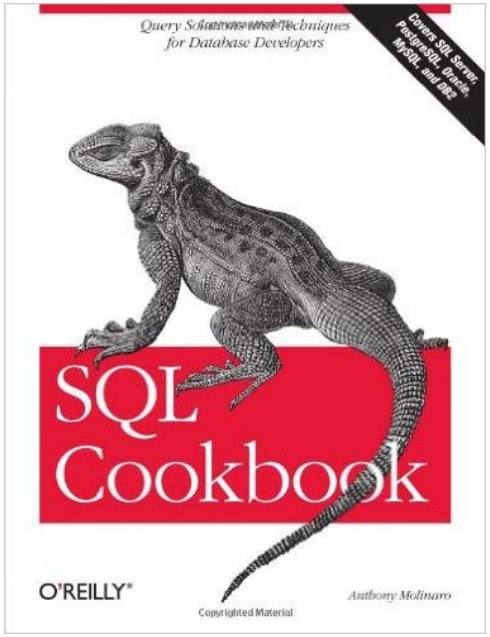
SQL Resources

SQL Tutorial and Resources

http://www.w3schools.com/sql/

Querying with Transact SQL Course, Graeme Malcom

https://www.edx.org/course/querying-transact-sql-microsoft-dat201x-3



Prepare for R Demos

> Install R

https://cran.r-project.org/

-or-

https://mran.revolutionanalytics.com/download/

> Install RStudio

https://www.rstudio.com/products/rstudio/download/

> Install Jupyter and the R kernel: **Follow directions carefully!**

https://www.continuum.io/downloads

https://irkernel.github.io/requirements/

GitHub

> Code, data and slides for this course are in a GitHub repository

https://github.com/StephenElston/DataScience410

> Install Git and GitHub for desk top

https://git-scm.com/download (Links to an external site.)Links to an external site.

https://help.github.com/desktop/guides/getting-started/installing-github-desktop/

- Or, just download the zip files

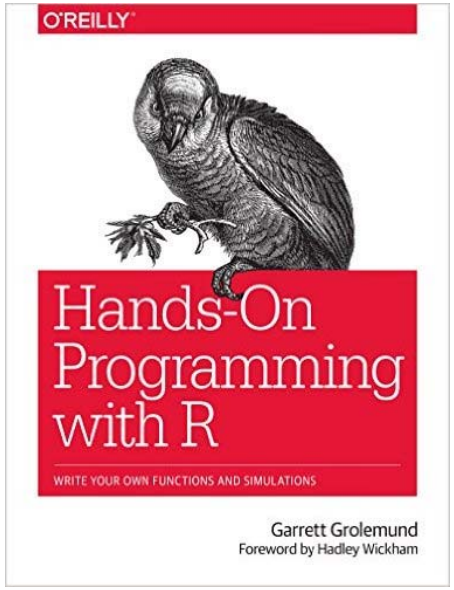
R Review

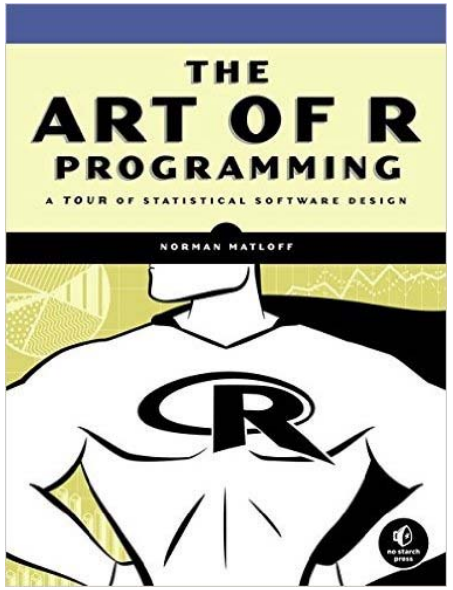
- R resources:
  - R page:
    - http://www.r-project.org/other-docs.html
  - Stackoverflow:
    - http://www.stackoverflow.com
  - 'Little' R intro:
    - http://cran.r-project.org/doc/contrib/Rossiter-RIntro-ITC.pdf
  - Quick R:
    - http://statmethods.net/
  - There are many tutorials available online, e.g.,
    - http://cyclismo.org/tutorial/R/
  - Google's Style Guide:
    - http://google-styleguide.googlecode.com/svn/trunk/google-r-style.html

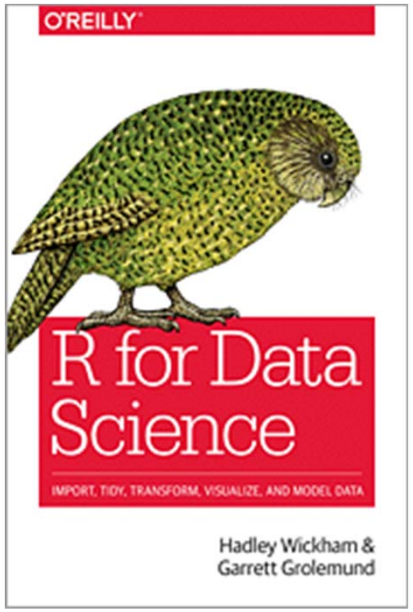
More R Resources

R Inferno, Pat Burns

http://www.burns-stat.com/pages/Tutor/R_inferno.pdf







Presentation and story telling

Important part of data science

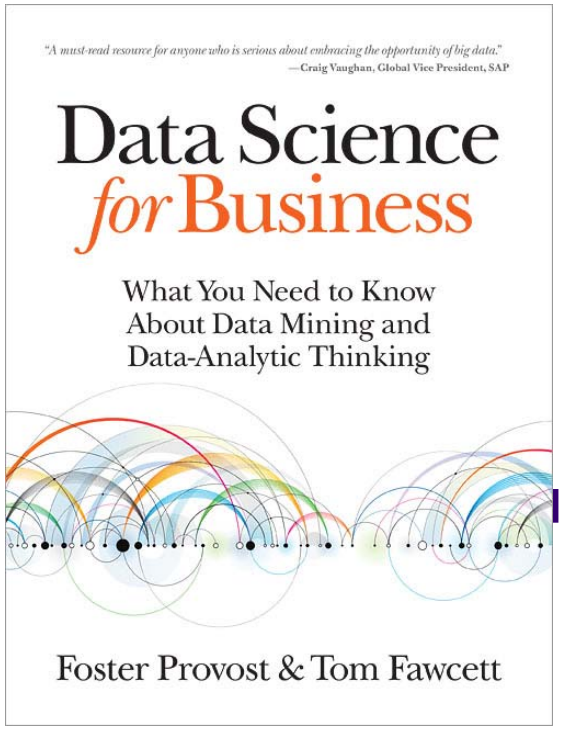
- Data science must have **impact**
- Results **only** have impact if they are understood
- Need to **'tell the story'**
- **Draw clear conclusion**
- Evidence supports conclusion

Presenting results is hard!

Data analytic thinking

Thinking about problems using objective analysis of data

- Define problem in terms of the business impact
- Review available data sources
- Explore the data
- Try various models
- Actionable results generate value
- Support recommendations with data and analysis
- Define metrics of success



Tips for story telling

Make the story clear

- Occam's Razor
- You will only hold attention for a short time
- Don't distract your audience
- Start with your conclusion
- Support your conclusion with evidence
- Few words = greater impact!

Don’t obfuscate your message!

Short and simple has business impact

- Minimize discussion of methodology and technical detail
- Clear charts
  - Label axis
  - Minimize over-plotting
  - Simplify
- Short simple tables
  - Label rows and columns
  - Highlight key point
  - Minimal rows and columns

Assignment

Homework 1:

- Use visualization and summary statistical methods to explore energy efficiency data set.
- Data on over 750 buildings.
- Energy efficiency of building measured as **heating load** or **cooling load**.

Assignment

Don’t panic!!:

- Exercise is deliberately open-ended.
- Exploration of a new data set is open-ended
- Expect exploration to be iterative
  - Try several ideas before you find truly interesting relationships.
  - The real-world is hard to understand!!

Assignment

You must submit:

- **ONE R-script.**
  - Your R script must be clear and concise. Use comments to explain the operation of the code.
- **ONE document outlining your 3 key points.** This document must include the following for each of your three points:
  - A clear statement explaining your points and supporting evidence for your conclusions.
  - Plots and tables presented as evidence to support your conclusion
  - Your document can be a Word document, HTML, or PDF document.

Assignment

Example conclusion:

The heating load of buildings depends on ... Evidence for this relationship can be seen by ... in the figure and by noting .... in the table above.

Summary

- Data Science is at the intersection of
  - Technology, including programming: SQL, R, Python, etc.
  - Math, probability, and statistics
  - Domain knowledge
- Presentation of results is a core skill
- Iterative exploration of the data with visualization
  - Understand the relationships in the data
  - Use multiple views of data