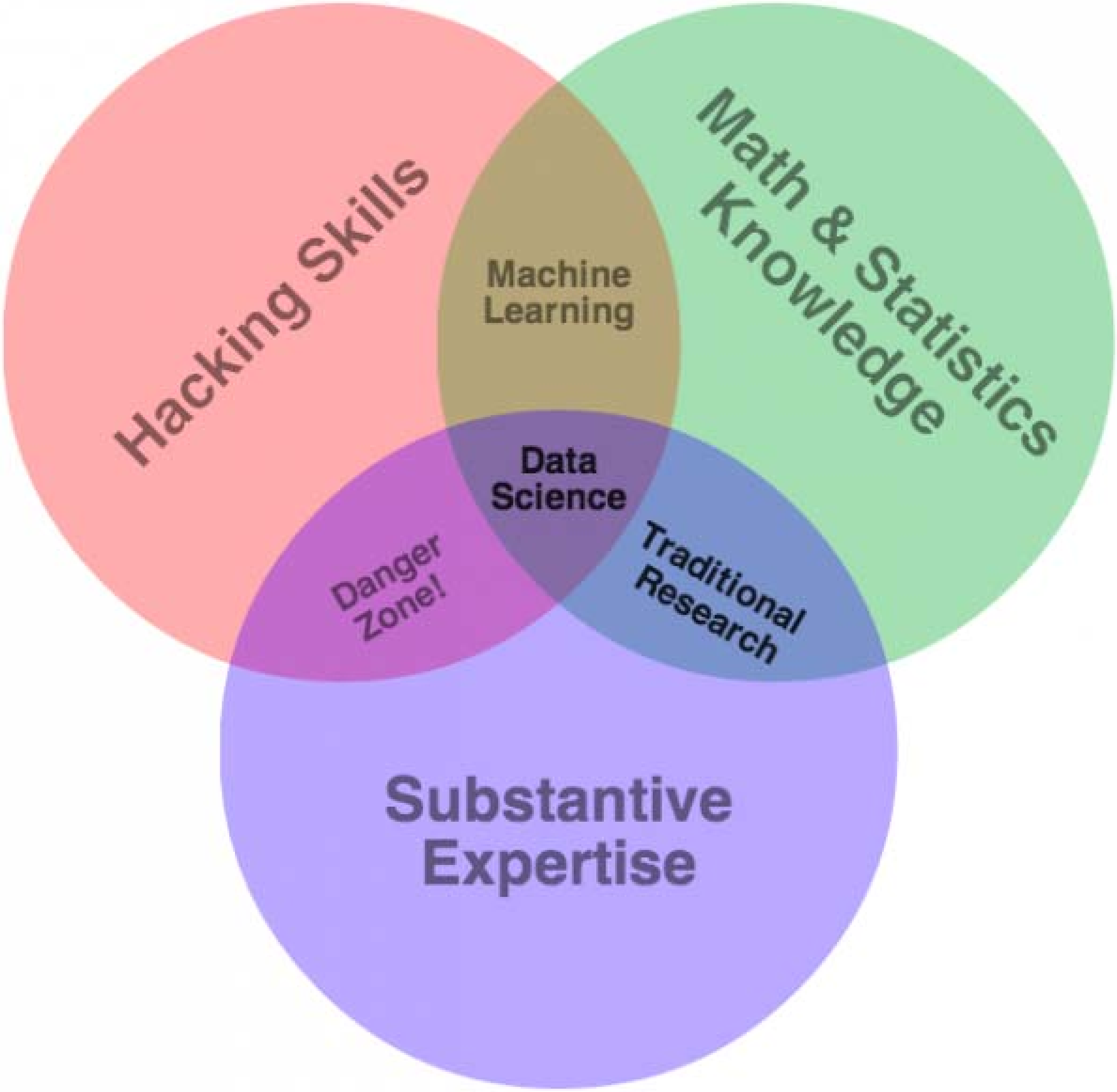
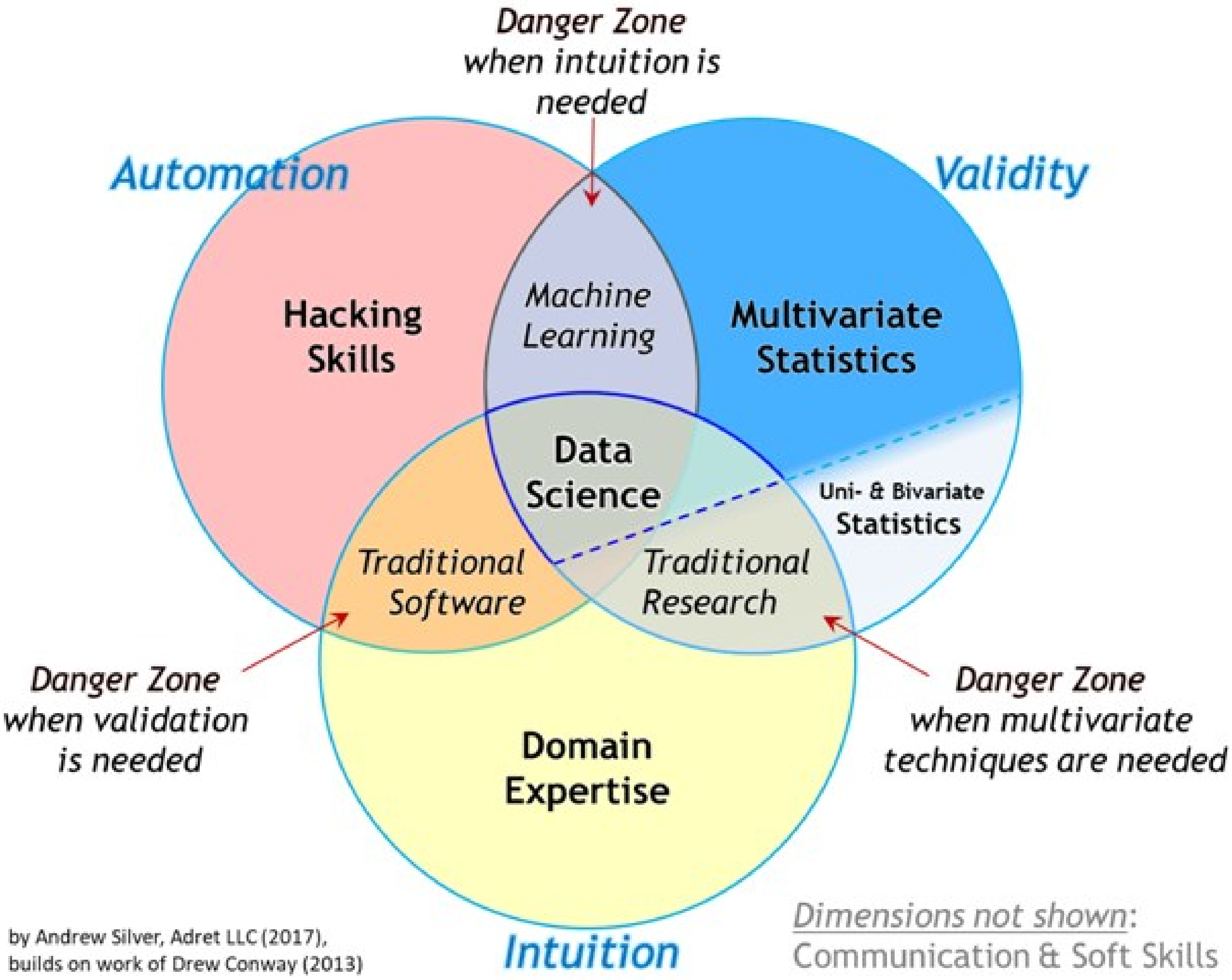
UNIVERSITY *of* WASHINGTON

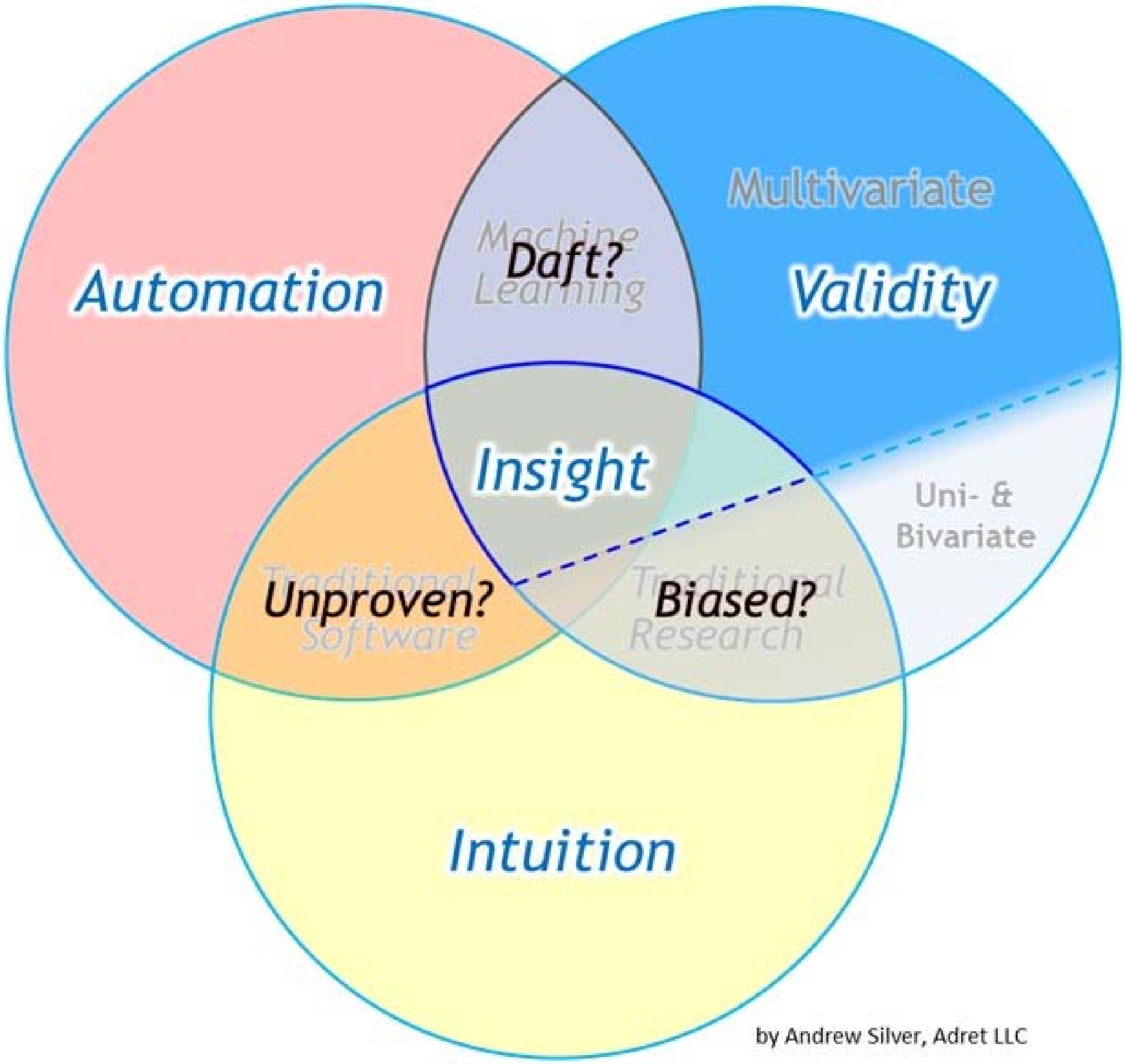
Data Science UW
Methods for Data Analysis

Introduction and Data Exploration
Lecture 1
Stephen Elston

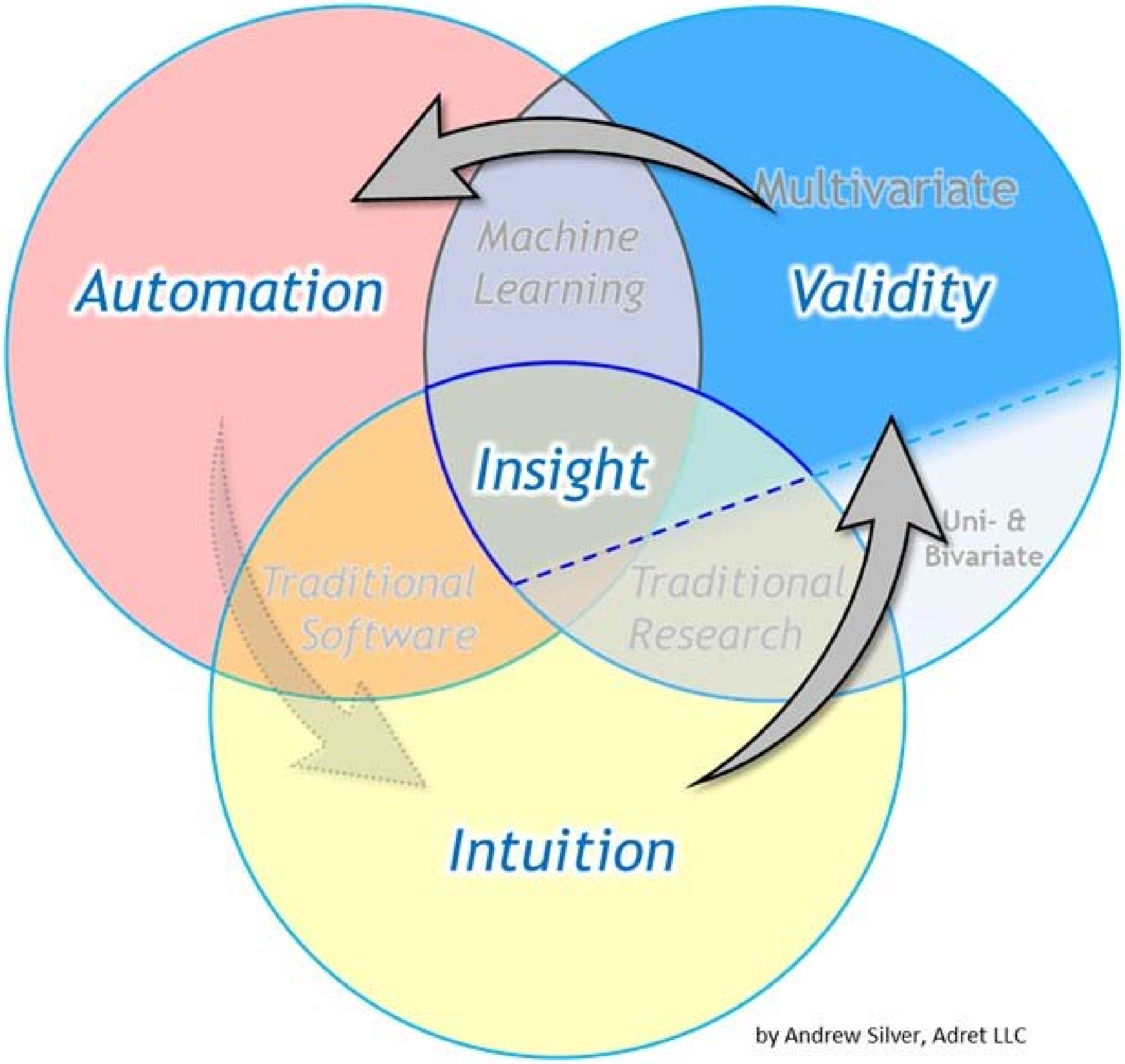

Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise



by Andrew Silver, Adret LLC (2017),
builds on work of Drew Conway (2013)



by Andrew Silver, Adret LLC



by Andrew Silver, Adret LLC

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 Python
  - Understanding the mathematical basis of machine learning models.
  - Expand Python programming skills to be able to write and test quality code from scratch.
- For more information about the course, please see the Canvas home page:
  - https://canvas.uw.edu/courses/1347202

Course Requirements and Grading

- Attendance: You MUST attend at least 8 out of 10 classes. **This is a non-negotiable UW requirement**.
- Need at least 75% cumulative grade to pass course
- Grading is based on:
  - **Quiz** in Canvas most weeks: questions on concepts.
  - **Discussion questions** in Canvas, each week: easy credit!!
  - **Homework** for most modules
  - **Milestone projects** are more substantial to pull concepts together for you
- Pay attention to the due dates. **Late work in be penalized!!**

Course Requirements and Grading

Homework and project guidelines

- All homework assignments must use good Python coding technique
  - Use loops, list comprehensions, functions etc.
  - Don’t just cut and paste code for multiple cases
  - If you have questions about coding, **ask!**
- Results must be presented in a professional style
  - **Presentation of results is a key data science skill**
  - We must be able to understand your conclusions
- The individual project must be complete and code explained (documented)

Office Hours and Contact Information

- Use the forum on Canvas.
  - **Answer other people's questions**
  - **You are responsible for reading the forum!**
  - I will try to read and answer most days
  - Make sure you have your **profile set** to get notifications!!!
  - Make sure your **email is correct** in your profile!
- 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.

Languages for data science

Skills every data scientist should have:

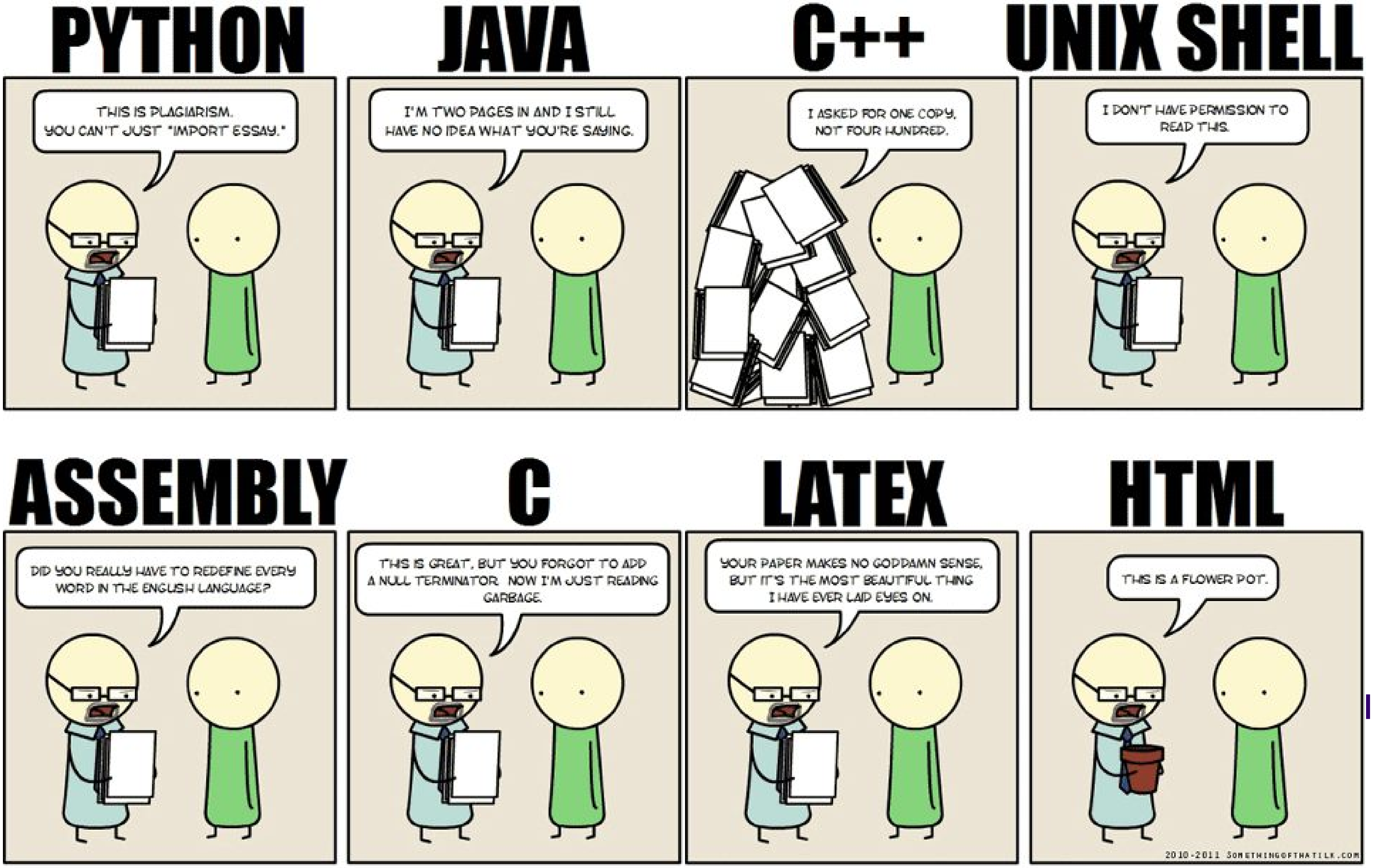
> SQL is the ‘lingua franca’ of data access – Data scientists need to access data!

> R – widely used for visualization, statistical analysis, and machine learning

> Python 3 – widely used for visualization, machine learning, big data APIs (e.g. Spark), deep learning APIs

- We use Python 3 in this course
- Use Anaconda stack for data science

Languages for data science


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

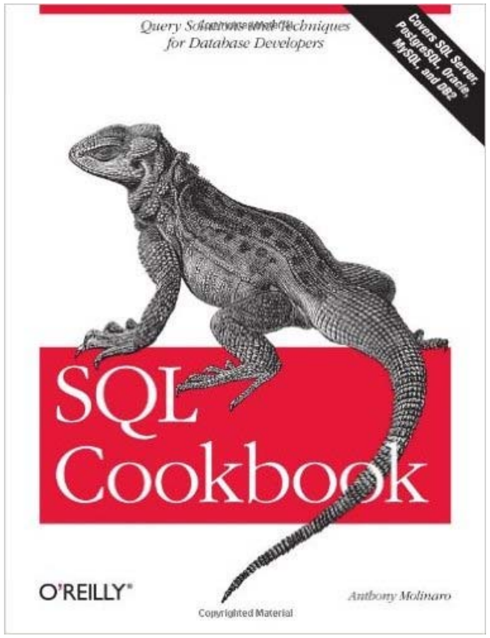
SQL Resources

SQL Tutorial and Resources

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

Querying with Transact SQL Course, Graeme Malcom

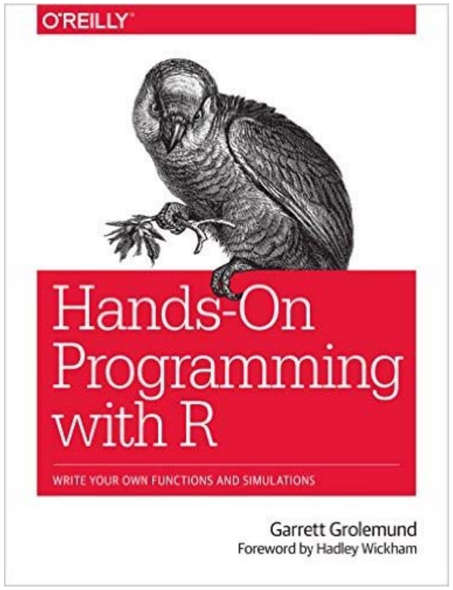
https://www.edx.org/course/querying-transact-sql-microsoft-dat201x-3

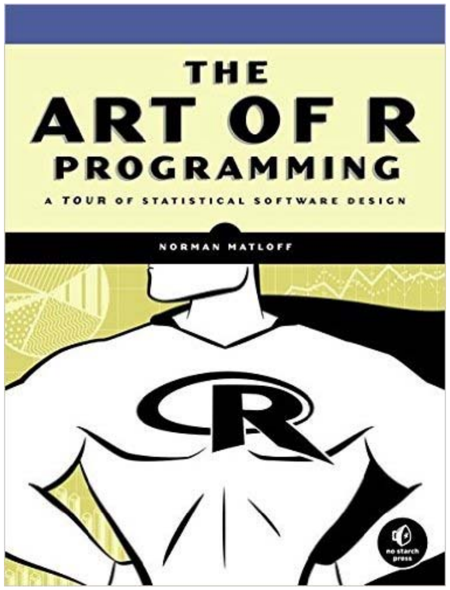

Query Solutions and Techniques
for Database Developers
Covers SQL Server, PostgreSQL, Oracle, MySQL, and DB2
SQL
Cookbook
O'REILLY
Anthony Molinaro
Copyrighted Material

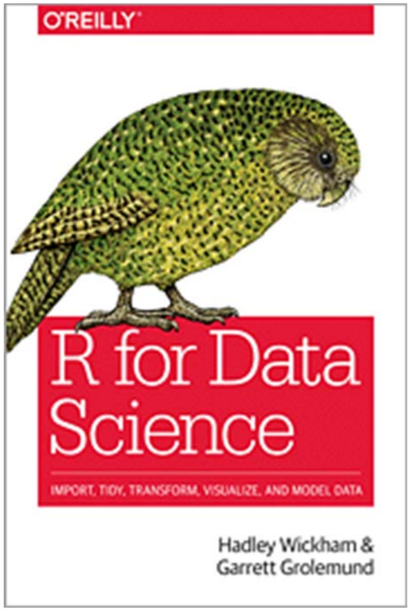
R Data Science Resources

R Inferno, Pat Burns

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


O'REILLY
Hands-On Programming with R
WRITE YOUR OWN FUNCTIONS AND SIMULATIONS
Garrett Grolemund
Foreword by Hadley Wickham





Python Data Science Resources

Numpy: https://numpy.org/

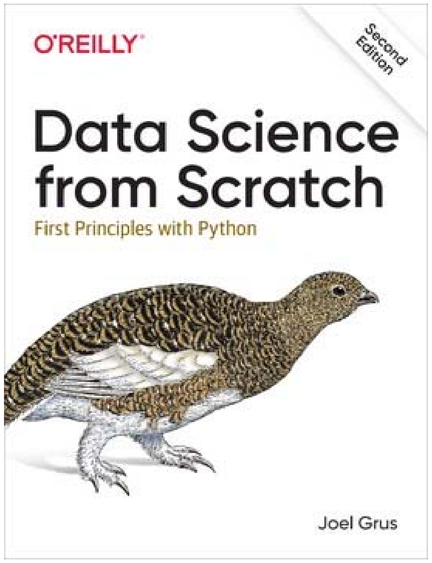
Matplotlib: https://matplotlib.org/

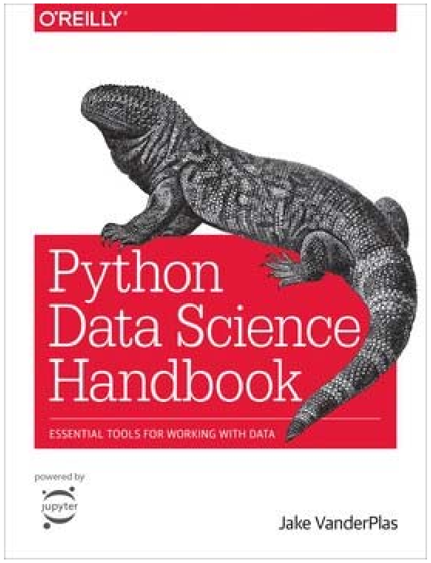
Pandas: https://pandas.pydata.org/

Statsmodels: https://www.statsmodels.org/stable/index.html

Seaborn: https://seaborn.pydata.org/

Scikitlearn: https://scikit-learn.org/stable/




O'REILLY
Python Data Science Handbook
ESSENTIAL TOOLS FOR WORKING WITH DATA
powered by jupyter
Jake VanderPlas

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

Presentation and story telling

Important part of data science

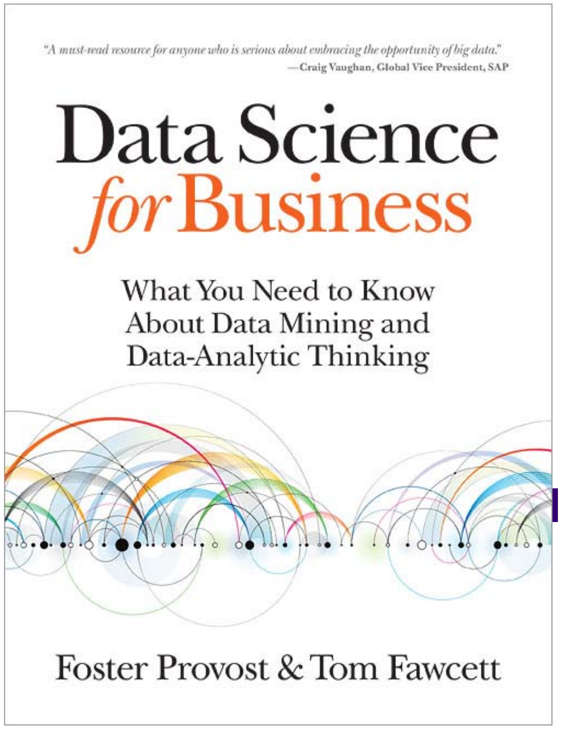
> Data science must have **impact**
> Results **only** have impact if they are understood
> Need to **‘tell the story’**
> **Draw clear conclusion**
> Evidence supports conclusion

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



"A must-read resource for anyone who is serious about embracing the opportunity of big data."
—Craig Vaughan, Global Vice President, SAP

Data Science *for* Business

What You Need to Know
About Data Mining and
Data-Analytic Thinking

Foster Provost & Tom Fawcett

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 Jupyter notebook.**
  - Your Python code must be clear and concise
  - You must explain what you are doing in text!
  - Your conclusions must be clearly written and supported by evidence from your analysis

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 – the topic of this course
  - 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