INFO 550 - Introduction

Steve Pittard wsp@emory.edu

January 25, 2018

Software Engineering Jobs

According to businessinsider.com

Cisco Systems

Brocade Communi… Systems

Apple

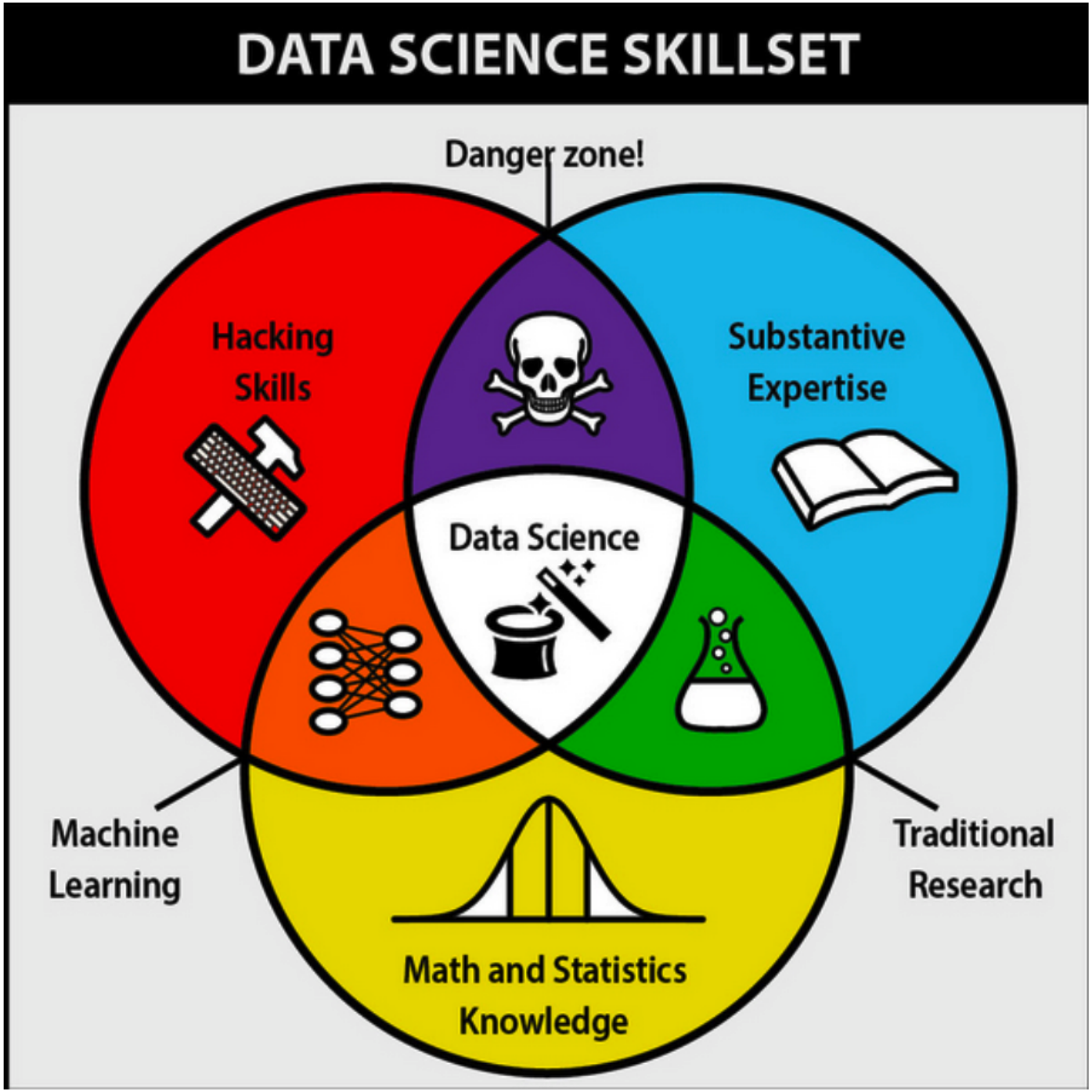
PayPal

Arista Networks

Google

- Cisco: $130,221. Thomson Reuters. ...
- Brocade: $132,036. Glassdoor/Brocade. ...
- Apple: $138,300. Steve Kovach/Tech Insider. ...
- PayPal: $146,795. Thomson Reuters. ...
- Arista: $151,648. NYSE. ...
- Google: $164,683. AP. ...
- LinkedIn: $170,839. Getty Images/Mandel Ngan. ...
- Facebook: $177,014. REUTERS/Robert Galbraith.

Data Science


DATA SCIENCE SKILLSET
Danger zone!
Hacking Skills
Substantive Expertise
Data Science
Machine Learning
Traditional Research
Math and Statistics Knowledge

Software Engineering

Definition: *Application of Sound Methods and Processes to the Development of Software Products*

- Analyze user need and design, test, and develop software
- Desgin each piece of an application including integration design
- Produce models and workflows to faciliate rapid software development
- Debug modules using a wide variety of inputs
- Test program capabilities and maintain

Goals

INFO 550 is A Class For Writing Software

- It's best to learn software development by doing !
- Write code with an explanatory purpose instead of just for an assignment
- Organize code into modules and track changes using industry tools
- Acquire a knowledge of "best practices" in software development

Prerequisites/Assumptions

Assumptions:

- You have previous programming experience and skill
- You can learn things rapidly
- You like to try things
- You know R and/or willing to learn it
- Knowledge and use of other languages is fine BUT
- Course content will be delivered in R
- All of the concepts we will cover exist in other languages

Class Topics

- Git for software development
- data.table and dplyr for large data managment
- SQL and Databases
- Debugging in R
- Web scraping with rvest
- XML and JSON
- Graphics and Visualization
- Regression and Clustering
- Bag of Words and Sentiment Analysis
- S3 and S4 Objects in R
- Amazon Web Services

The Final Project

- The bulk of your work relates to the Final Project
- There is one homework just to make sure you have the tools in place to proceed
- Propose an analysis and find some interesting data
- Or the other way around
- Can pull together skills learned in other classes

The Final Project

- Must relate to data of substance in content and size
- Can also be based off of existing research papers and/or media reports
- Apply the techniques and approaches you will learn in the class
- Data Sources can be flat files, databases, or APIs
- Generate some driving questions
- Think about questions and things you would want to know
- Ask your colleagues for ideas - ask the instructor for ideas

Some Examples

- Genome Analysis
- Visualizing Flu Data or Geographical Trends
- Employment Rates / Labor Data
- Medicare / Nursing Home Data
- Stock Markets and Economics
- Tax and Income Data
- Weather and Meterology
- Gaming APIs
- Twitter Data

The Final Project

- Propose an idea by Spring Break (early March)
- I prefer students work individually BUT
- I will also consider team efforts
- Meet with instructor early on to talk over ideas
- Use the provided report format to help flesh out ideas

The Final Project - Report Format

A 2-3 page PDF file:

- Use R Markdown, Sweave, or a Notebook
- Your name and project title
- Summary of research questions and results
- Describe Motivations and originating ideas
- Cite the Dataset(s) and Sources
- Methodology
- Results
- Reproducibility

The Final Project - Title and Name

Title and Name:

- Project Title and your Name
- Should have a tag line that summarizes your research
  - "A Model to Predict Nursing Home Violations In The Southeast"
  - "A Dynamic Map of Public Transit Patterns in Boston"
  - "Classification of Influenza Cases Use K-Means Clustering"

The Final Project - Research Questions

Research Questions:

- One to three sentences that articulate each research question
- Provide a summary response for each research question
- Does not need to be detailed - summary only

The Final Project -Motivation and Original Ideas

Motivation and Originating Ideas:

- Provide background and rationale why this is interesting
- Remember - this can be about existing research or topics
- This can be as long as you want

The Final Project - Datasets and Sources

Dataset(s) and Sources:

- Provide citation info for any and all datasets used or considered
- You must provide download links and URLs
- If not then describe how you obtained the data
- If too large then consider putting on Dropbox or Emory Box
- Your cleaning and analysis code must relate to this data
- That is - any and all cleaning and transfofrmation must be reproducible using your code
- Do **NOT** use data that contains confidential or medical data

The Final Project - Methodology

Methodology:

- Describe what you did analytically and why.
- Write as if you were explaining to an individual
- Talk about any obstacles you had programmatically
- Remember you can use the power of R Markdown:
  - Can intermix code with narrative
  - This includes early diagnostic graphs
- Talk about any statistical approaches you took
- Mention any key moments or "a-ha" ideas you had

The Final Project - Results

Results:

- Present Conlusions for each Research Question you Posed
- Provide an interpretration for the lay person
- Results should be supported by analytic results and graphs
- Present any hanging or unanswered questions
- What would you do next if you had more time ?

Source Code in Support of Project

Submitting Your Code:

- You can intermix code within your narrative BUT
- You are required to submit all code to a GitHub account
- The code should be well documented with comments
- The comments should be helpful but not overly detailed
- The code should have a significant of git commits
- Feel free to develop an R Package
- Your Final Report should also be available on git

Final Presentation

You will present the last day of class:

- Present your Final Report to the class
- 10-15 minutes is plenty
- This is "story time"
- Provide personal observations and reactions
- Talk abouit obstacles and challanges
- Answer questions from the class and instructor