Reproducible Research

Bringing it all together

Mass Street University

PER EDUCATIONEM PROGRESSUS

Reproducible Research

- Treat data science projects like legit scientific research projects
- You can think of asking a question as developing a hypothesis to prove or disprove
- Your question can be simplistic or as complex as necessary
- Linear regression is the foundation for a lot of algos

Mass Street University

PER EDUCATIONEM PROGRESSUS

Reproducible Research

- Introduction to the topic came from the Not So Standard Deviations Podcast.
- Researches and software engineers approach data science wildly differently.
- Both sides can learn from the other.

Mass Street University

PER EDUCATIONEM PROGRESSUS

It's All Six Degrees of Johns Hopkins' Biostatistics Department

People Associated With Johns Hopkins

People NOT Associated With Johns Hopkins

Not So Standard Deviations Podcast

Jeff Leek

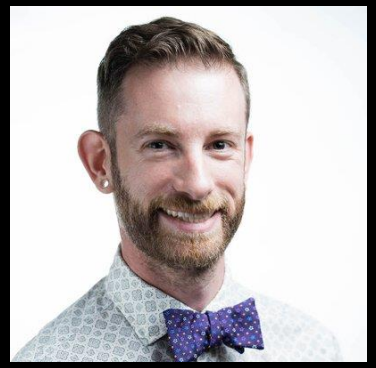
Roger Peng

Hillary Parker

Simply Statistics Website

Tidyverse
Reproducible Research
Literate Statistical Programming
Opinionated Analysis
Data Science as a Science

Brian Caffo



Hadley Wickham

Mass Street University
PER EDUCATIONEM PROGRESSUS

Reproducible Research

- Rise of Python for Data Science/Engineering
- Rise of notebooks (Jupyter, Zeppelin, R Notebook)
- Data Science SaaS (cloud, cloud, and more cloud)
- R got a nice NLP package
- Deep Learn all the things!
- Rise of Spark.

Mass Street University

PER EDUCATIONEM PROGRESSUS

Reproducible Research

- Someone should be able to run your exact analysis and get your result.
- Goal is to reproduce NOT replicate.
  - Reproduce = validate your work
  - Replicate = validate the conclusions of the study
- This is a lot harder than it sounds.
- Reproducibility hasn't been totally figured out.
  - I still struggle with dependencies
  - Build tools for R?

Mass Street University

PER EDUCATIONEM PROGRESSUS

Reproducible Research

- Elements of reproducibility
  1. Analytic data (the Tidy data)
  2. Analytic code
  3. Documentation
  4. Distribution

- Of these, distribution is the trickiest

Reproducible Research

- Literate Statistical Programming
  - Combine your analysis and your code into a single document
  - There are several tools for this
    - Markdown
    - RMarkdown/knitr
    - R Studio
    - Notebooks

Reproducible Research

- A proposed structure of analysis
  - Defining the question
  - Defining the ideal dataset
  - Determining what data you can access
  - Obtaining the data
  - Cleaning the data
  - Exploratory data analysis
  - Statistical prediction/modeling
  - Interpretation/Challenging of results
  - Synthesis and write up
  - Creating reproducible code

Mass Street University

PER EDUCATIONEM PROGRESSUS

Reproducible Research

- Reproducibility Checklist
  - Start with good science
  - Don’t do things by hand
  - Don’t point and click
  - Teach a computer
  - Use version control
  - Keep track of your software environment
  - Don’t save output
  - Set your seed
  - Think about the entire pipeline

Mass Street University

PER EDUCATIONEM PROGRESSUS

Opinionated Analysis Development

- Read Opinionated Analysis Development
- Opinionated analysis = analysis that follows certain practices
- Follows on to the principals of reproducible research
- Lays out a framework for how an analysis should be completed