Reproducible Research with R, The Tidyverse, Notebooks, and Spark

Bob Wakefield
Principal
bob@MassStreet.net
Twitter:
@BobLovesData

Mass Street Analytics

Who is Mass Street?

- Boutique data consultancy
- We work on data problems big or “small”
- We have a focus on helping organizations move from a batch to a real time paradigm.
- Free Big Data training

Mass Street Partnerships and Capability

- Hortonworks Partner
- Confluent Partner
- ARG Back Office

HORTONWORKS PARTNERWORKS CONSULTANT

confluent

ARG® The Better Back Office

Mass Street Analytics

Bob’s Background

- IT professional 16 years
- Currently working as a Data Engineer
- Education
  - BS Business Admin (MIS) from KState
  - MBA (finance concentration) from KU
  - Coursework in Mathematics at Washburn
  - Graduate certificate Data Science from Rockhurst
- Addicted to everything data

Follow Me!

- Personal Twitter: @BobLovesData
- Company Twitter: @MassStreet
- Blog: DataDrivenPerspectives.com
- Website: www.MassStreet.net
- Facebook: @MassStreetAnalyticsLLC

Mass Street
Analytics

KC Learn Big Data Objectives

- Educate people about what you can do with all the new technology surrounding data.
- Grow the big data career field.
- Teach skills not products

Mass Street Analytics

ACM Kansas City

We're looking for a speaker willing to talk in deep detail about data engineering challenges their organization is experiencing.

Mass Street Analytics

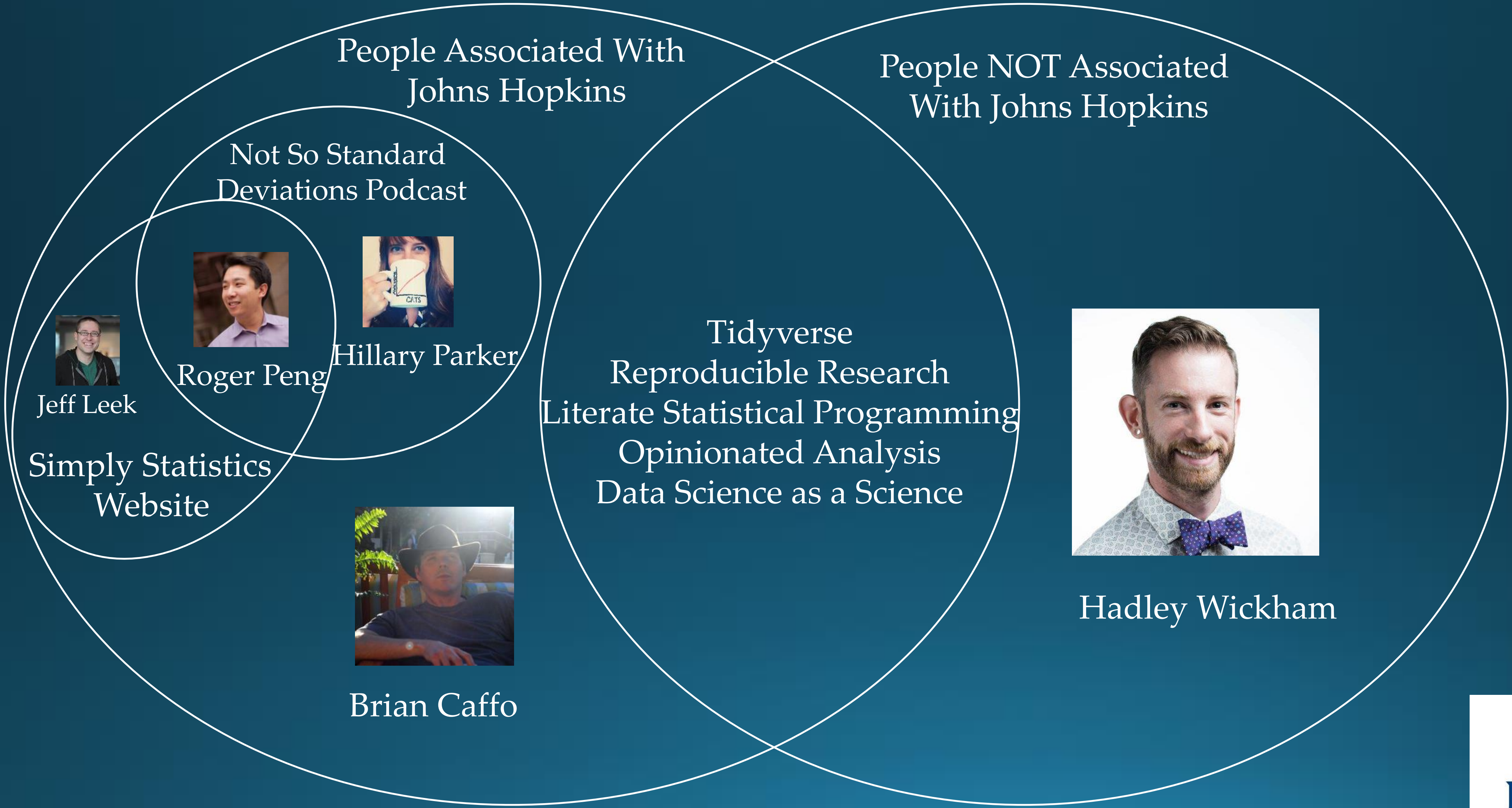
This Evening’s Learning Objectives

- Tonight we’ll cover the following topics
  - Tidyverse
  - Reproducible Research
  - Literate Statistical Programming
  - Opinionated Analysis
  - Data Science as a Science
- We’ll also cover the following tools
  - Git
  - Jupyter Notebook
  - Knitr/Rmarkdown, Markdown
  - S3
  - SparklyR

All Material Can Be Downloaded from GitHub

MassStreetAnalytics/Reproducible-Research

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



People Associated With Johns Hopkins

Not So Standard Deviations Podcast

Simply Statistics Website

Jeff Leek

Roger Peng

Hillary Parker

Brian Caffo

People NOT Associated With Johns Hopkins

Hadley Wickham

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

Mass Street Analytics

Motivations For This Evenings Discussion

- My investment application is moving into a new phase of work.
- I don't have a PhD. (sad face)
- My need to smack down trolls on FB.
  - Demand more from the internet.
  - Showing your work should be the new standard

Mass Street Analytics

Changes to the Career Field in the Past 15 Months

- 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.

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.

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?

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

*From “Report Writing for Data Science in R”

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

*From “Report Writing for Data Science in R”

Opinionated Analysis Development

- Read Opinionated Analysis Development
  - Link in references
- 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

Mass Street Analytics

Tidy Data

- Three rules that make data tidy:
  - Each variable must have its own column.
  - Each observation must have its own row.
  - Each value must have its own cell.
- No you're not crazy. Yes that's third normal form.
- I don't have to deal with this issue often if ever.

Mass Street Analytics

Tidyverse

- Used to be called the Hadleyverse
- An ecosystem of packages designed with common APIs and a shared philosophy
- Helps you get your data tidy
- Also assumes that your data is tidy

Mass Street Analytics

Part II: Tools

- Git
  - Modern Source Control
  - Code repositories: GitHub and Bitbucket
  - GUI: Sourcetree

- Rmarkdown/knitr
  - Appears to be strictly an R Studio thing

- Pandoc Markdown/Jupyter
  - Julia, Python, R
  - Rebranded Ipython

Mass Street Analytics

Part II: Tools

- SparklyR/Databricks
  - SparklyR provides an R API for Spark with dplyr

- AWS/S3
  - Helps solve the problem of accessibility to data
  - Can be annoying to manage

- Tidyverse
  - A set of packages that makes working with data easier

Mass Street Analytics