Lecture 1: Applied Data Science Introduction

STAT GU4243

Applied Data Science

Cynthia Rush
Columbia University

January 17, 2018

Class today

1. A quick intro about data science, generally.
2. A quick intro to this class, specifically.
3. A discussion of the first project.
4. Some tutorials, if we have time.

What is Data Science?

Data Science

- Data Science represents a new approach to
  * Acquire knowledge,
  * Collect evidence,
  * Form decisions,
  * Make predictions.
- The end points are:

  knowledge, evidence, decisions, and predictions.
- Driven by breakthroughs in technologies.
- Enabling faster solutions to traditional evidence-based practices.
- Creating solutions that would not be otherwise possible.

A simplified data project cycle

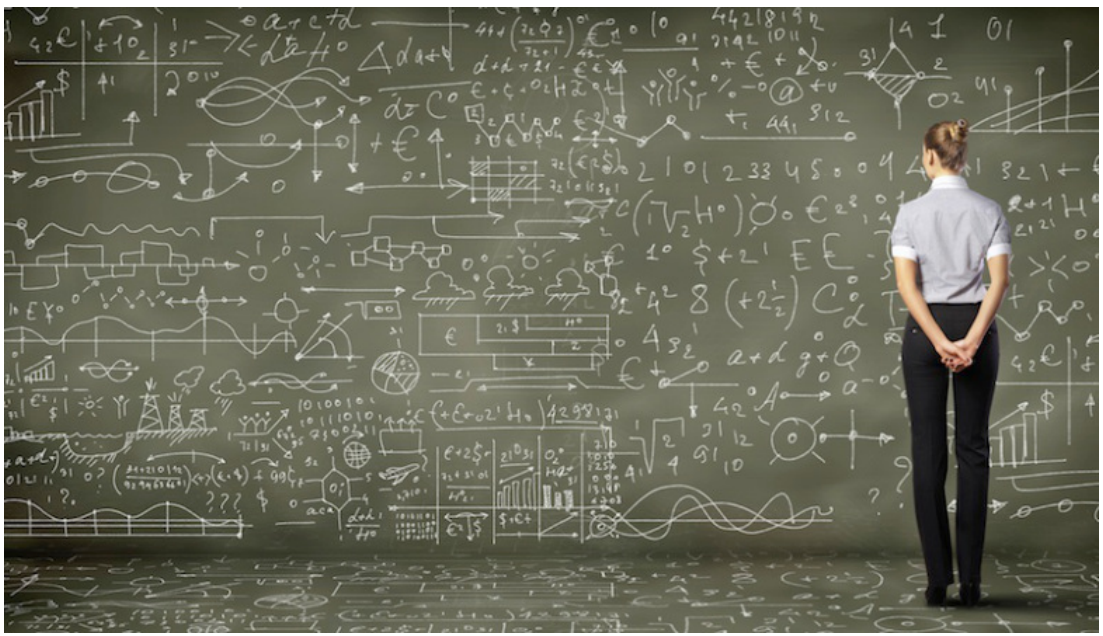
A simplified data project cycle

1. Begin with a real world question or problem.

A simplified data project cycle

1. Begin with a real world question or problem.

2. Brainstorm: What data/tools can help?



A simplified data project cycle

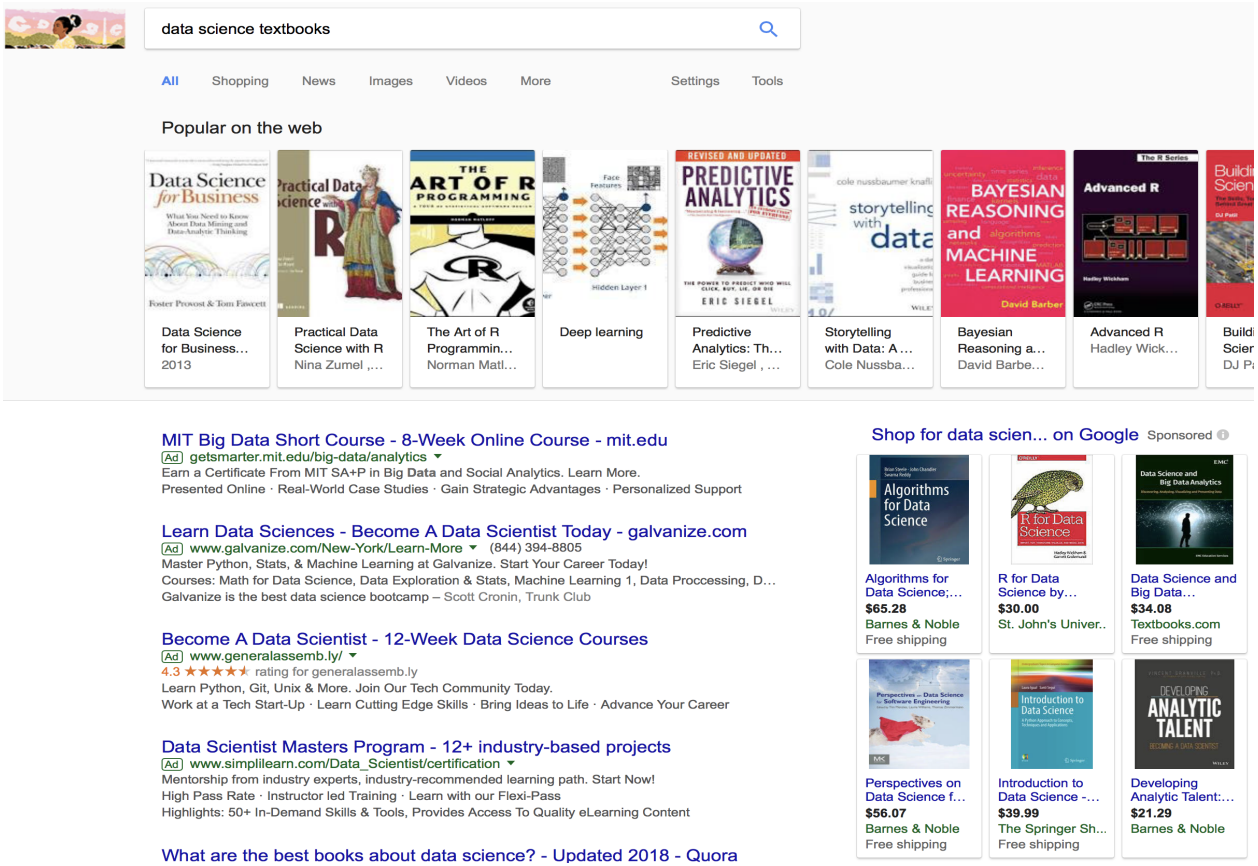
1. Begin with a real world question or problem.

2. Brainstorm: What data/tools can help?

3. Finally, problem solving.

An example – search evaluation

Search results for 'data science textbooks'



1. **Begin with a question:** What DS textbooks should we return?
2. **Brainstorm:** Those with DS in the title? Those other people use? Those for which the publisher pays us?
3. **Problem solving:** How do we find the 'most popular' DS texts?

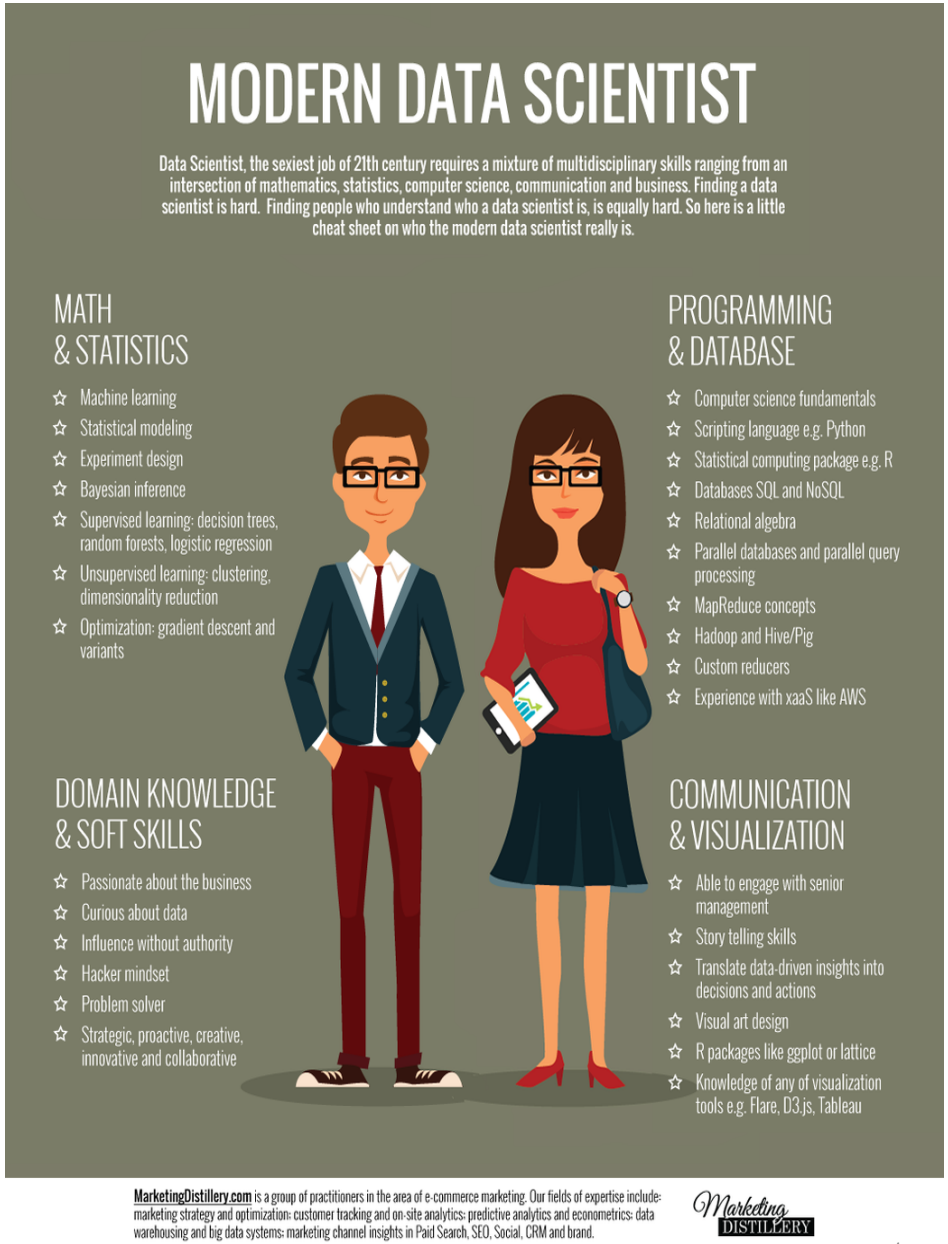
Data Science combines aspects of many disciplines to create meaning from data.

Foundations of data science:

- Data engineering
- Software engineering
- Machine learning
- Statistics

Data science skill set

- How to use data to solve problems:
  - Mathematics, Statistics, Machine Learning
- How to *handle* data:
  - Technologies: Python, Java, Hadoop, Spark, etc
- How to work with others: teamwork and collaboration skills
- How to turn data into business intelligence: find value in data
  - Innovation, intellectual curiosity
  - Problem-solving skills
- How to convince others of your results
  - Visualization, story telling
  - Communication skills



How does this course make you a better data scientist?

What this course won't do:

- No formal instruction on statistics/machine learning topics.
- Not intended to be a comprehensive data science bootcamp.

What this course will do:

Project-based learning or learning by doing.

- Problem identification via teamwork and discussion.
- Problem solving by using existing skills or new skills, learn new things "on the job", and learn from your peers.
- Present your codes, your results and your story (try to sell them).
- There will be things I cannot answer but let's learn together.

Project-based learning

Learning objectives

- Become self-directed learners
- Develop our skill set:
  1. Problem-solving skills
  2. Teamwork skills: collaboration, reasoning, and communication
  3. Self-assessment skills
  4. Presentation and critique skills
- Gain 'hands-on' data science experience
- Master the toolkit collected from more tradition classes

Student-centered approach

- I am not here to lecture, but rather to facilitate active learning.
- I will design open-ended challenges, each of which focuses on a slightly different area in data science.
- In each challenge,
  1. Start with information/knowledge we already have (maybe not you but your teammate) about the problem.
  2. Identify knowledge/skills we need to solve the problem.
  3. Articulate the above thinking process in a team and implement an inquiry as a team
- I will provide case studies and tutorials to give guidance during the above processes.

Channels of Communication

During class time

- Brainstorm
- Ask questions during tutorial

Before and after classes

- Piazza

If you have questions

- Piazza
- As a last resort, email

Group Projects

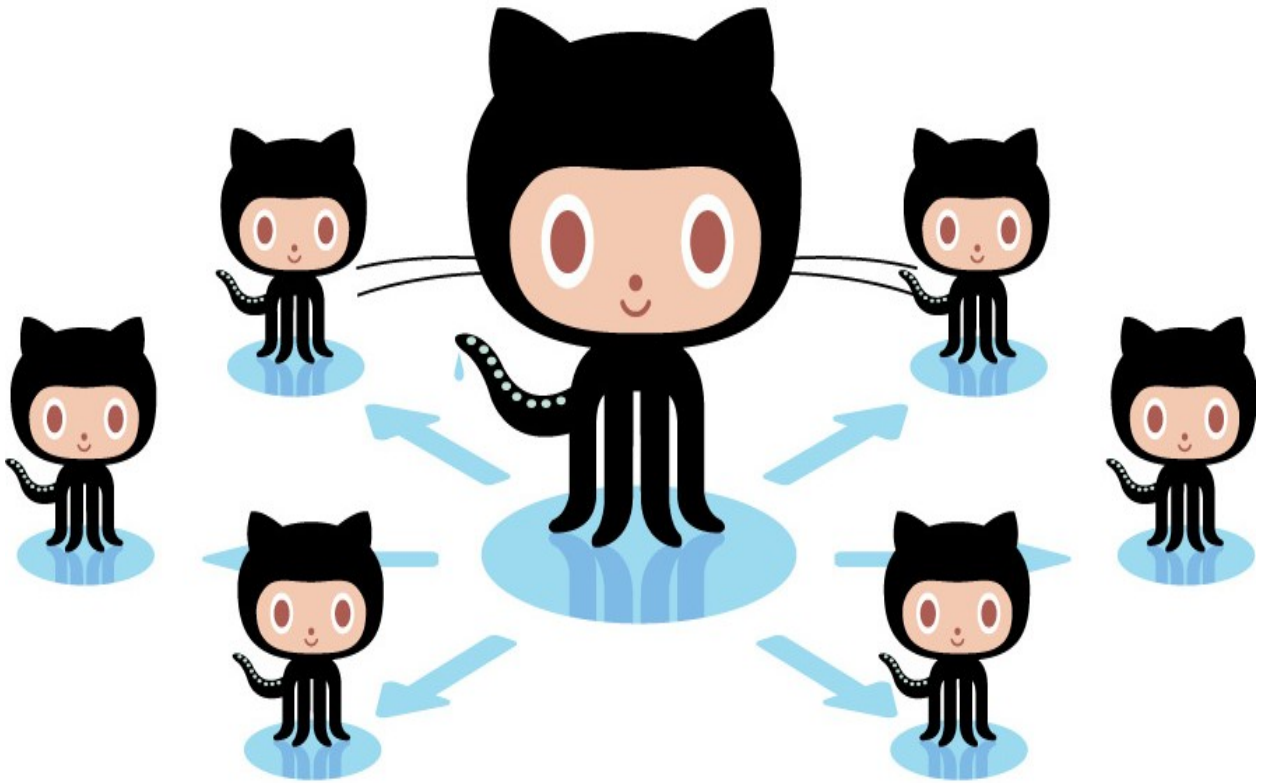
Working together

- You don't have to be in the same room at the same time to work together.
- You will work together in this course in the following ways:
  1. Face-to-face brainstorming
  2. Online discussion in a group forum
  3. Online video chat (say, via Google Hangout) with screen share
  4. GitHub collaboration

Learning on GitHUB

This semester we will use Classroom for GitHub

It allows the instructor to create parallel private repositories for groups to collaborate.



Project assignment

- I will create a starter code folder
- I will create groups with group numbers (off GitHub)
- I will share the group info with students (especially group number) on Piazza
- I will create assignments (private) and set the option for "new set of groups"
- I will send invitation links to students with instructions:
  - First, check whether your teammate already created a team for your group from the "Join an existing group".
  - If you cannot find your group's name (as assigned in the Excel name), please create the team using precisely the name specified in the Excel file.
- The Project name and membership can be managed later but the most important part is we get all the teams/groups set up automatically.
- Everyone from your team should install Git, GitHub Desktop, and use Git with RStudio.

Reproducible data analysis

Improve reproducibility

- Setup project folder
- Documentation
- Project history and source control

Project setup

- Rstudio really makes it easy to keep track of a project.
- First, identify a working folder.
- Inside the working folder, create the following subfolders. data: data used in the analysis. Read only doc: the report or presentation files
  * data: data used in the analysis. Read only.
  * doc: the report or presentation files.
  * figs: contains the figures. Only contains generated files, images used for report should be put in a separate image folder under doc.
  * lib: various files with function definitions and code.
  * output: analysis output, processed datasets, logs, or other processed things. Only contains generated files.

Use GiT for version control

More on this next time.

Use knitr for reproducible data analysis

- knitr is an R package that processes R markdown files.
- An R markdown file follows the markdown syntax and contains R code blocks.
- An R markdown file can be "knitted" into either a html page or PDF document that reproduces a data analysis.
- It shows both the code chunks and the results produced.
- One can also include seamlessly project discussion, method section (with LaTeX support) and results discussion.
- It should be viewed as a data analysis documentation, rather than a report though, as the analysis needs to presented in a chronological order.

More on this next time.