

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 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

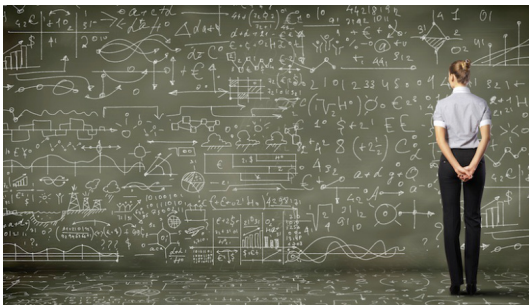
A SIMPLIFIED DATA PROJECT CYCLE

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'

The screenshot shows a Google search results page for the query 'data science textbooks'. At the top, there's a search bar with the query and a magnifying glass icon. Below the search bar are tabs for 'All', 'Shopping', 'News', 'Images', 'Videos', 'More', 'Settings', and 'Tools'. A section titled 'Popular on the web' displays a grid of book covers. The books shown include 'Data Science for Business', 'Practical Data Science with R', 'The Art of R Programming', 'Deep learning', 'Predictive Analytics', 'Storytelling with Data', 'Bayesian Reasoning and Machine Learning', 'Advanced R', and 'Build Your Own Data Science Project'. Below the grid, there are several links to online courses and books. The first link is 'MIT Big Data Short Course - 8-Week Online Course - mit.edu', which includes a brief description and a link to the course. The second link is 'Learn Data Sciences - Become A Data Scientist Today - galvanize.com', which also includes a brief description and a link to the course. The third link is 'Become A Data Scientist - 12-Week Data Science Courses', which includes a brief description and a link to the course. The fourth link is 'Data Scientist Masters Program - 12+ industry-based projects', which includes a brief description and a link to the program. The fifth link is 'What are the best books about data science? - Updated 2018 - Quora'. To the right of the links, there is a section titled 'Shop for data science... on Google' with a 'Sponsored' label. This section displays a grid of book covers with their prices and shipping information. The books shown include 'Algorithms for Data Science', 'R for Data Science', 'Data Science and Big Data', 'Perspectives on Data Science', 'Introduction to Data Science', and 'Developing Analytic Talent'.

data science textbooks

All Shopping News Images Videos More Settings Tools

Popular on the web

Data Science for Business
John Peacock & John Peacock
2013

Practical Data Science with R
John Fox
2016

The Art of R Programming
Norman Matloff
2015

Deep learning
Ian Goodfellow, Yoshua Bengio, and Aaron Courville
2016

Predictive Analytics: The Art of Analyzing Data to Solve Problems
Eric Siegel
2013

Storytelling with Data: A Computer Scientist's Guide to Data Visualization
Cole Nussbaumer Knappert
2015

Bayesian Reasoning and Machine Learning
David Barber
2012

Advanced R
Hadley Wickham
2014

Build Your Own Data Science Project
John Peacock & John Peacock
2013

MIT Big Data Short Course - 8-Week Online Course - mit.edu
[5.0] getsmarter.mit.edu/big-data/analytics
Earn a Certificate From MIT SA+P in Big Data and Social Analytics. Learn More.
Presented Online - Real-World Case Studies - Gain Strategic Advantages - Personalized Support

Learn Data Sciences - Become A Data Scientist Today - galvanize.com
[5.0] www.galvanize.com/New-York/Learn-More - (844) 394-8805
Master Python, Stats, & Machine Learning at Galvanize. Start Your Career Today!
Courses: Math for Data Science, Data Exploration & Stats, Machine Learning 1, Data Processing, D...
Galvanize is the best data science bootcamp - Scott Cronin, Trunk Club

Become A Data Scientist - 12-Week Data Science Courses
[5.0] www.generalassembly.io
4.3 ★★★★★ rating for generalassembly
Learn Python, Git, Unix & More. Join Our Tech Community Today.
Work at a Tech Start-Up - Learn Cutting Edge Skills - Bring Ideas to Life - Advance Your Career

Data Scientist Masters Program - 12+ industry-based projects
[5.0] www.simplilearn.com/Data_Scientist/certificate
Mentorship from industry experts, industry-recommended learning path. Start Now!
High Pass Rate - Instructor led Training - Learn with our Flexi-Pass
Highlights: 50+ In-Demand Skills & Tools, Provides Access to Quality eLearning Content

What are the best books about data science? - Updated 2018 - Quora

Shop for data science... on Google Sponsored

Algorithms for Data Science
\$65.28
Barnes & Noble
Free shipping

R for Data Science by...
\$30.00
St. John's Univer...
Free shipping

Data Science and Big Data...
\$34.06
Textbooks.com
Free shipping

Perspectives on Data Science...
\$56.07
Barnes & Noble
Free shipping

Introduction to Data Science...
\$39.99
The Springer Sh...
Free shipping

Developing Analytic Talent...
\$21.29
Barnes & Noble

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

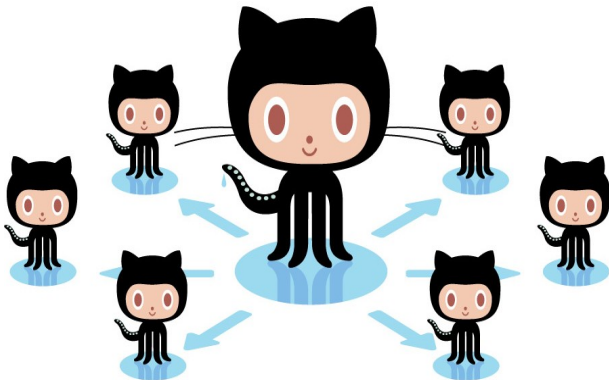
WORKING TOGETHER

- ▶ 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 be presented in a chronological order.

More on this next time.