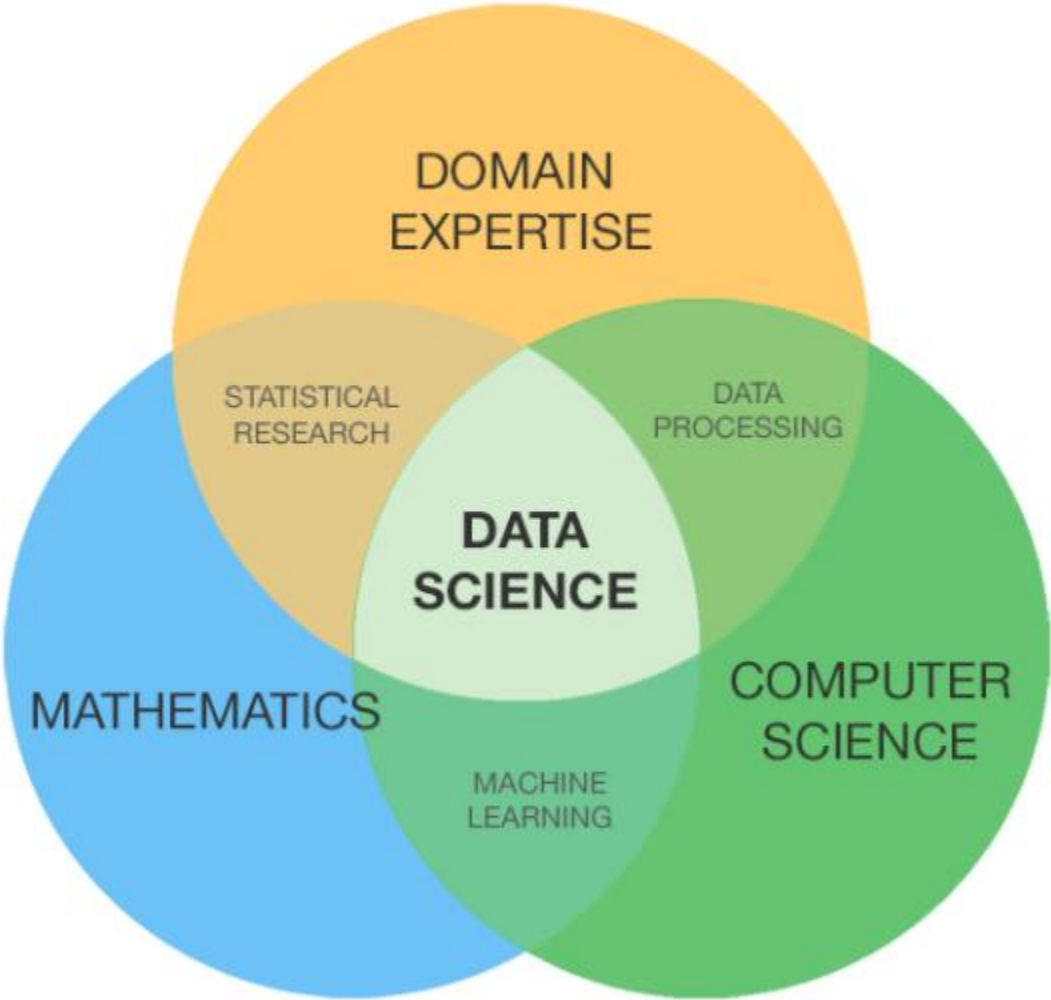


SML201

Introduction to Data Science

Michael Guerzhoy

Data Science



Data Science



Data Science



Dan Ariely

January 6, 2013 at 6:17pm · 



Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

 Like

 Comment

 Share

Data Science and SML201

- Read in and process data for analysis
- Visualize the data to analyze and communicate it
- Use knowledge about where the data came from and how it was collected, as well as data visualization, to formulate mathematical models of the data
- Use models of the data to make predictions about new data
- Use models of the data to make inferences about the data
 - Are the patterns we are seeing in the data indicative of real trends, or are they merely noise?

R and SML201

- Reading in and processing data is more efficient if a computer does it
 - We will do it with R
- Processes/recipes for visualizing and analyzing data can be thought of as algorithms
 - Step-by-step instructions (to a human or a computer) about what to do with the data
 - Will code (some) algorithms in R to
 - Better understand them
 - Be able to code up algorithms that have not yet been implemented

R and SML201

- About half of the class doesn't have any programming background
 - We will not assume you have any programming background
 - We will teach R using a functional approach, which nearly no one in the class has seen
 - (The style of Python, Java, etc. programs is usually much more imperative than functional)

Plan for this semester

- Intro to R
- Data visualization
- Predictive modelling: predicting new data by looking at existing data
 - Using domain knowledge to build models
- Testing statistical hypotheses and confidence intervals: mathematical procedures for determining whether the patterns you see in data are just noise
- Introduction to machine learning
- Case studies

Course organization

- Syllabus to be posted by the end of the week
- (Mandatory) precepts start this week
 - You must work in teams of two
- 3 data analysis projects and 2 problem sets
 - You may work in teams of two, or by yourself
- 2 in-class tests
 - One test during midterm week, one test during class at the end of the semester

Precepts

- Each precept assignment is worth 2%
- You must work in teams of two
- You get full points if you make your best effort
 - This is Princeton, I expect everyone to make their best effort
- Both team members should understand the solution completely
 - Multiple models that I saw that worked well
 - A model that *doesn't* work well: two friends, with one friend having a lot more background and doing all the work
 - Switch partners
 - Or have the friend with more experience act as a tutor (that could work well for both team members)

About me

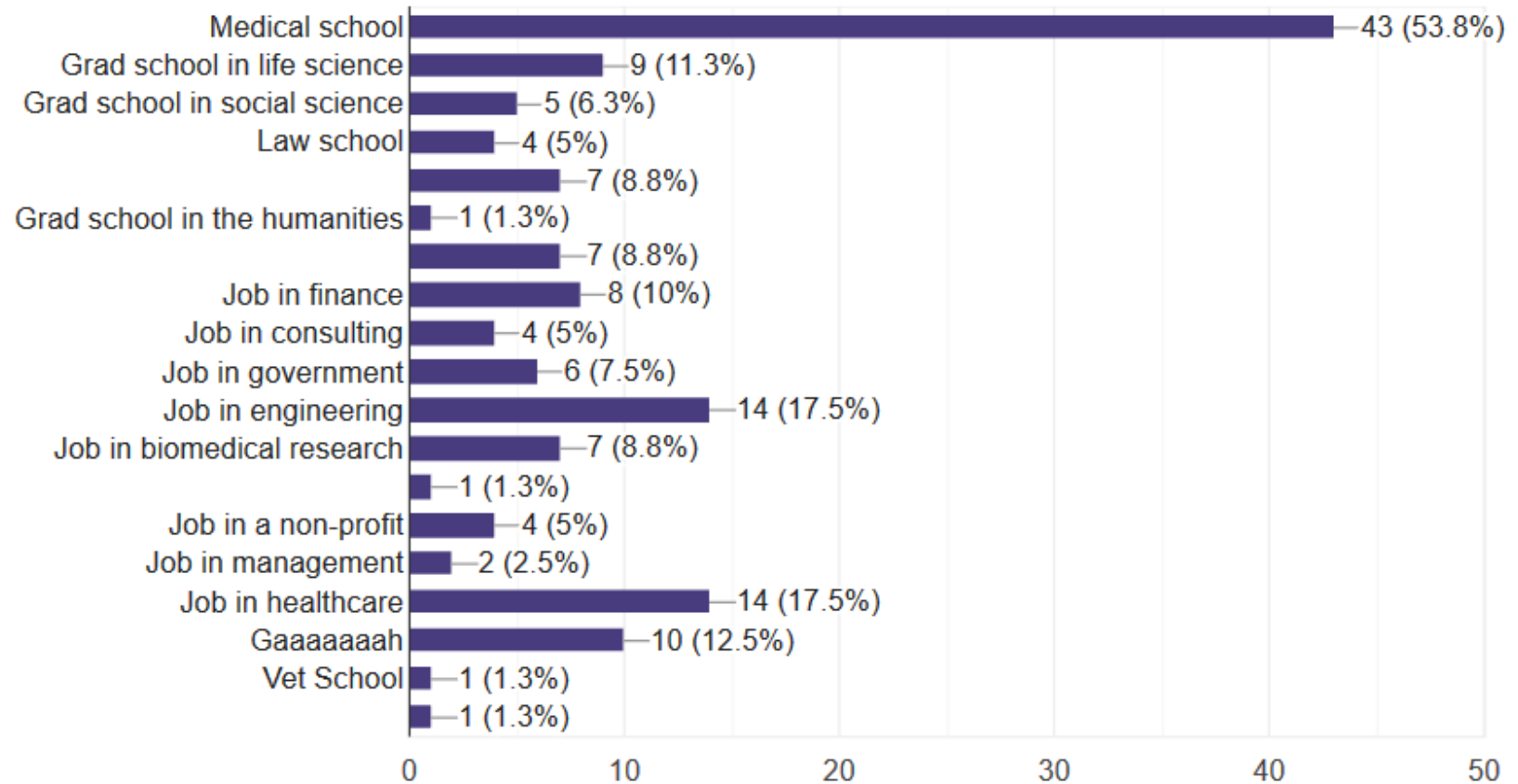
- My last name is spelled “Guerzhoy” and pronounced “ger-ZHOY”
- Started as a lecturer in CSML at Princeton last year
- Also an affiliate scientist at St. Michael’s Hospital in Toronto
 - Using patient data to predict adverse outcomes (death, transfer to ICU) in real time
 - Analyzing CT images of the cervical spine to diagnose fractures
 - Automatically extracting information from dictated clinical notes
- Love data

About you

- Fill out the survey if you haven't yet!
<https://forms.gle/P1EcY29BYu2XqtXu8>

What are your plans for after college?

80 responses



How do locomotives know where to go? They go through lots of training :)

whenever I try to tell a chemistry joke, there's always no reaction

My love life at Princeton University

What do you do with a musician who can't play their instrument?

You give them two sticks and put them at the back of the orchestra.

My love life

My GPA.

What is your favorite joke?

New to Programming?