Five College Undergraduate Data Science Education

Nicholas Horton
Statistics
Amherst College

Andrew McCallum
Computer Science
UMass

github.com/Amherst-Statistics/DataScienceSoiree2017-02

Five College Undergraduate Data Science Education

Nicholas Horton
Statistics
Amherst College

Andrew McCallum
Computer Science
UMass

Thanks to Mass Mutual for their support of this event and other data science initiatives.

Envisioning the Data Science Discipline

National Academies of Science workshop and roundtable

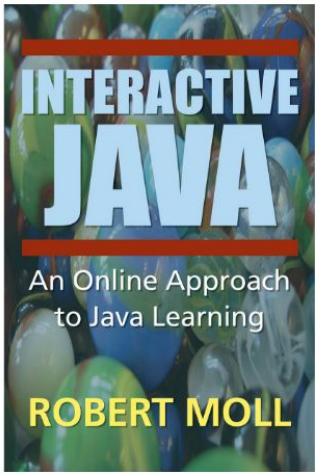
Nicholas Horton
Statistics
Amherst College

Introductory CS at UMass

Neena Thota
Computer Science
UMass

COMPSCI 121: INTRODUCTION TO PROBLEM SOLVING WITH COMPUTERS

- Spring enrollment: 437 students.
- Introduction to object-oriented paradigm with Java programming language.
- No previous programming experience assumed.
- Required class for CS major; also for Informatics, Electrical engineering, and Mathematics majors.
- Weekly Lectures 2 + Lab 1.
- Mid-term and Final exams.
- Future: POGIL



Online Web Learning
University of Massachusetts at Amherst - Amherst, Massachusetts
Computer Science

What Data Science Skills Do Our Students Need? Let's Ask Them!

Amy Wagaman
Amherst College

Let's ask…

- Students who took my multivariate data analysis course (Fall 2009 or Fall 2012)
  - With no prior statistics courses and who took no further statistics courses (from our department)
  - Asked what topics from the course they'd seen elsewhere; how taking the course had impacted them
- The course covered multivariate techniques and their applications without delving deeply into the underyling linear algebra.
- Students undertook course projects where they had to employ at least 2 techniques we'd covered.
- The R software was used.
- We now teach the course with a pre-req of intro stats.

Selected Student Responses

- Two are consultants; one is a research associate
- First consultant
  - “Most relevant for my current job is methods of data visualization for large data sets.”
  - The course was useful for interviews.
- Second consultant
  - “In my job as a consultant I've used clustering, classification and factor analysis.”
  - Having experience computing has been very useful.
- Research associate
  - Has seen lots of Principal components analysis
  - “Learning a statistical program/language relatively early in my college career was particularly valuable for the work I do now”

Summary

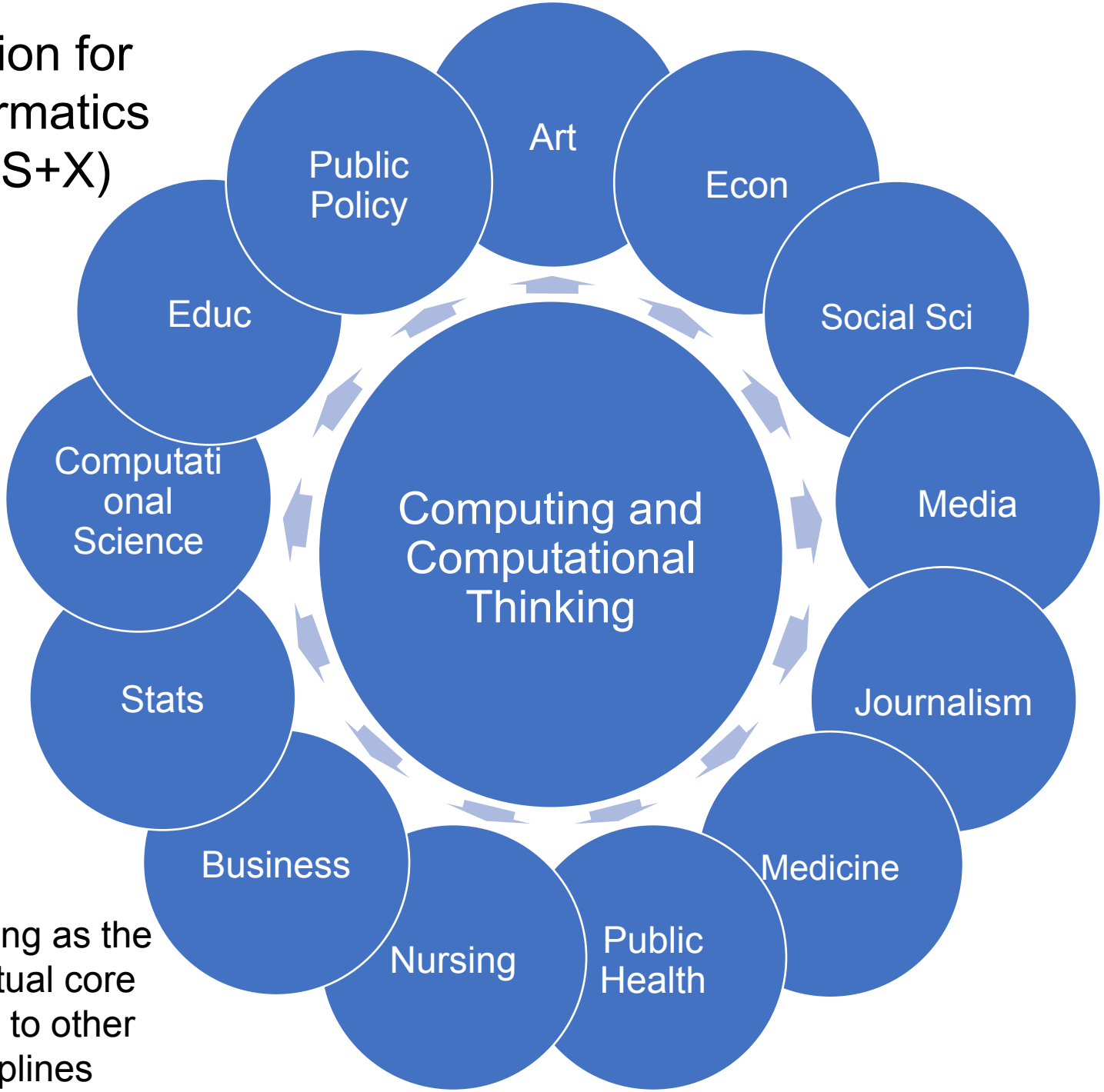
- Many possibilities lie before us
- Wide–range of topics to decide to cover or not
- Computing knowledge is certainly increasing in importance.

- What would your students say?
- See Wagaman (*The American Statistician*, 2016)

Data Science in the Informatics Program at UMass

Ramesh Sitaraman
Computer Science
UMass

Vision for Informatics (CS+X)



Computing as the intellectual core applied to other disciplines

Core

Foundations

- 101 Introduction to Informatics
- 150 A Mathematical Foundation of Informatics

Programming Sequence

- 121 Problem Solving with Computers
- 186 Using Data Structures

Human Factors/Societal

- 290NW A Networked World
- 325 Human Computer Interaction (HCI)

Data Science Track Core Requirements

- 397F: Intro to Data Science
- Stats Requirement: COMPSCI 240 or STATS 240 or Resource Economics 212
- 326: Web Programming (satisfies IE requirement)
- 345: Databases

Data Science Electives (4 courses)

From a menu of allowed courses across campus

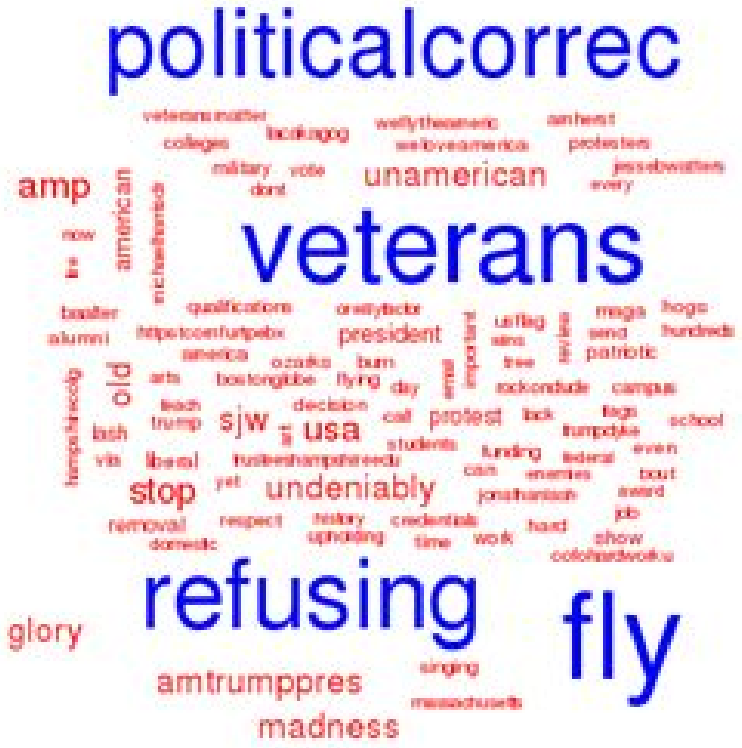
- CS 390MB: Mobile Health Sensing & Monitoring
- STATS 501: Methods of Applied Statistics
- OIM 301: Operations Management
- And, many many more

Students may propose electives for approval

Current and Projected Growth

- Currently 69 info majors, majority in the data science track (49 majors), rest in multimedia
- Excellent diversity (23 women, 4 African Americans, 35 non-caucasian).
- First info/DS student graduated F'16. More in S'17.
- Projected growth: 50 new majors/per year till 200 majors total

Data Science at Hampshire College (Ethan Meyers)



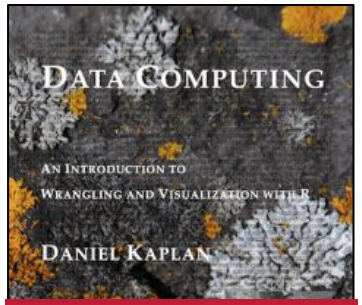
CS 173T: Intro to Data Science tutorial

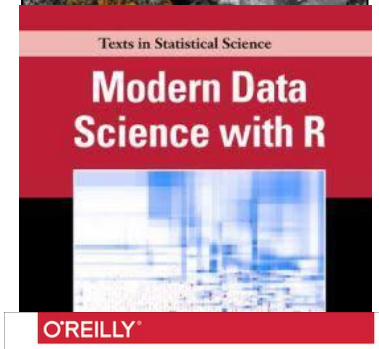
The class:

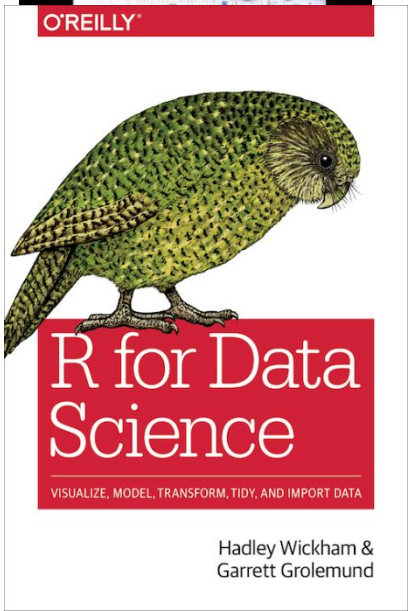
- 8 first year students
  - Most had just a little prior programming experience
- Tutorial: Intro to college + advisor + class
- Met twice a week for an 80 minutes

Topics

- Intro to R
- Data Wrangling
- Data Visualization (grammar of graphics/ggplot)
- Machine learning
- Interactive visualizations (Shiny)
- Text manipulation, data scraping (web APIs)







CS 173T: Intro to Data Science tutorial

In class:

- Explained concepts then had them try it in R

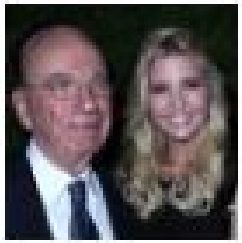
Assignments:

1. Read Data Journalism articles
   - Quote and Reactions and a presentation
2. Weekly worksheets and DataCamp
   - R Markdown worksheets > DataCamp
3. Midterm and final projects
   - 5-10 page Markdown report + presentation

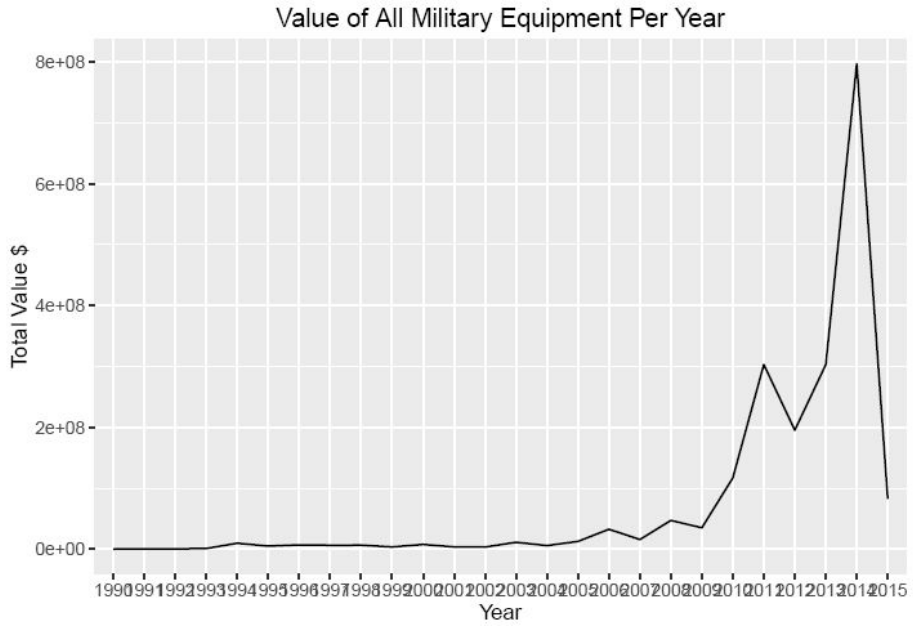
Assignments

A Grump-Murdoch Axis, and What It Means for the World

By JIM RUTENBERG



The ties between Rupert Murdoch and Mr. Grump are undeniably close, writes our media columnist. But what does that



Takeaways

Super fun

So much more to add!

- R-bloggers
  - La Quinta is Spanish for “Next to Denny’s”

Educational outcomes

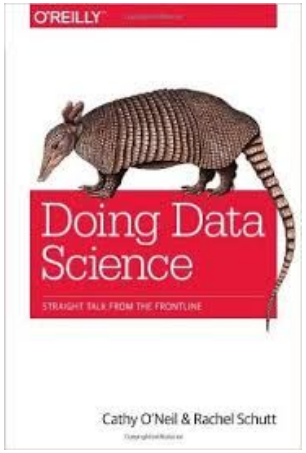
- Students learned how explore data and answer questions using R
- Prerequisite for Intro to Statistics?

Undergraduate CS Data Science Education at UMass

Benjamin Marlin
Computer Science
UMass

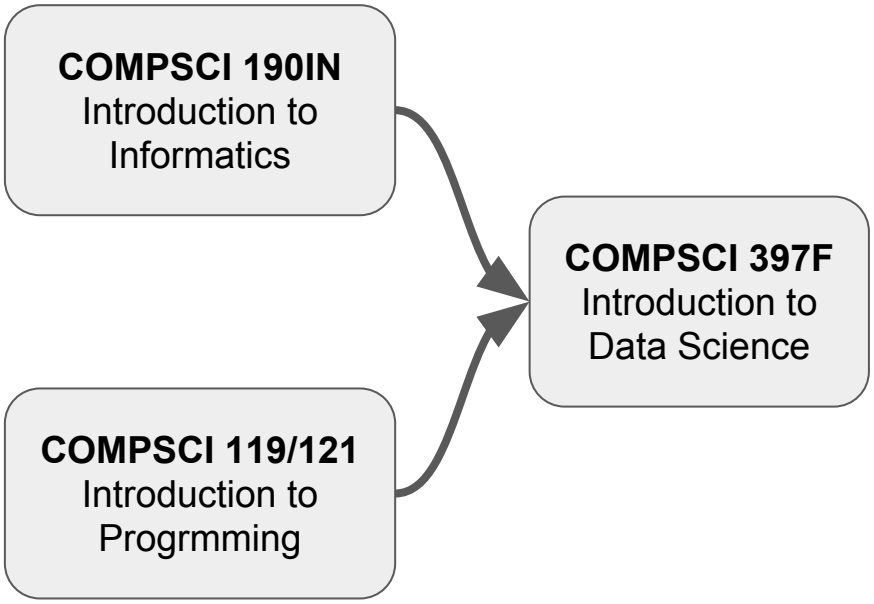
COMPSCI 397F: Introduction to Data Science

- **Program/Level:** Informatics Major, Sophomores/Juniors
- **Topics:** Data acquisition and wrangling, exploratory data analysis (descriptive statistics, clustering), prediction (KNN, naive Bayes, logistic regression, linear regression), data visualization.
- **Focus:** Applications



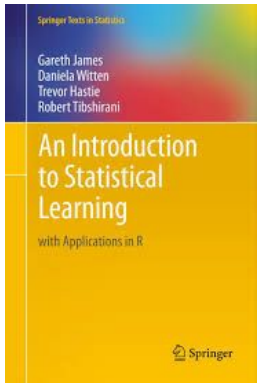
COMPSCI 397F: Introduction to Data Science

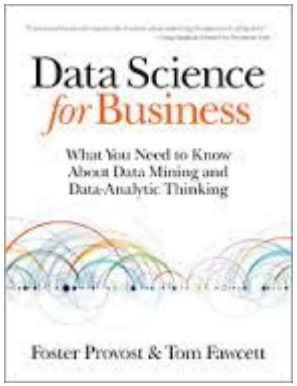
- **Assumed Background:** Programming



COMPSCI 348: Principles of Data Science

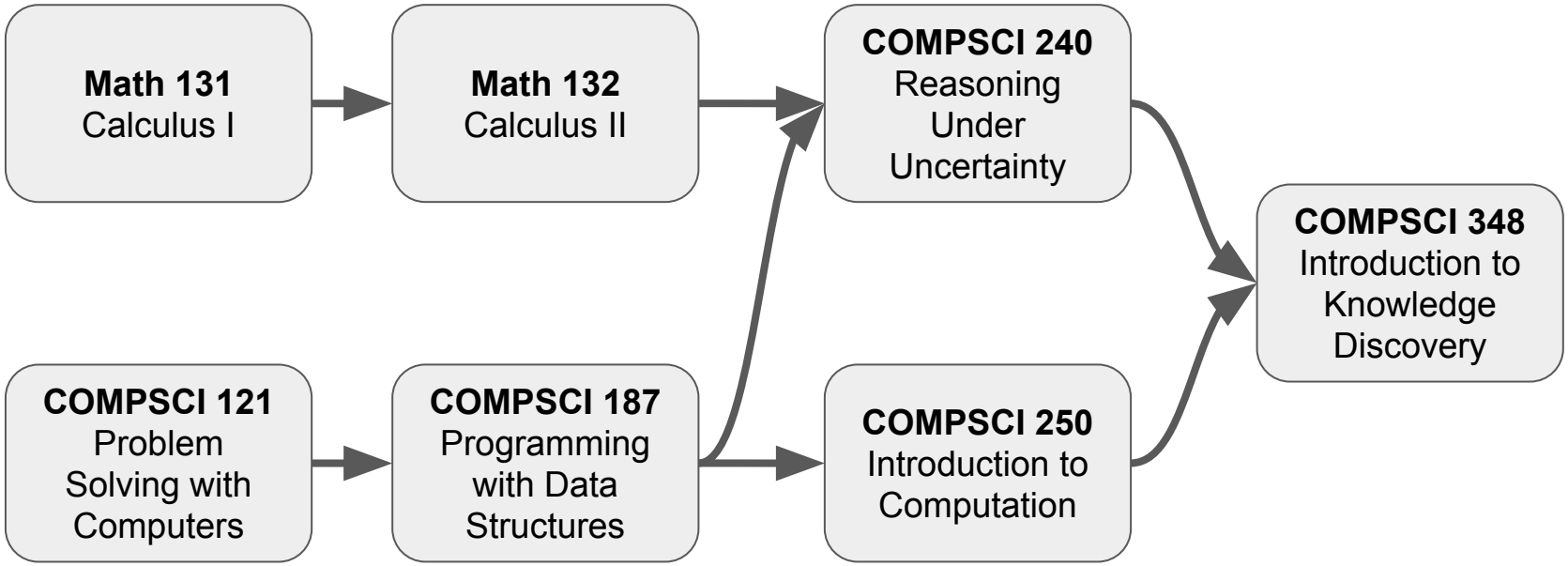
- **Program/Level:** CS Major, Juniors and Seniors
- **Topics:** Exploratory data analysis (descriptive analytics, clustering), prediction (decision trees, linear regression, generalization and overfitting), probabilistic models (Bayesian networks, anomaly detection), causal inference (propensity score methods, causal networks).
- **Focus:** Theory and Applications





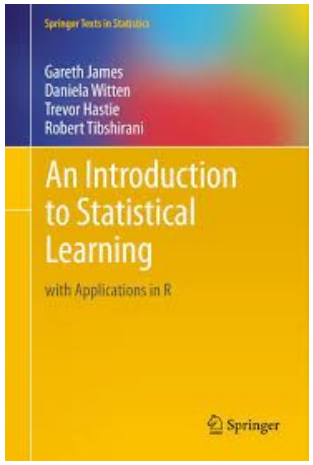
COMPSCI 348: Principles of Data Science

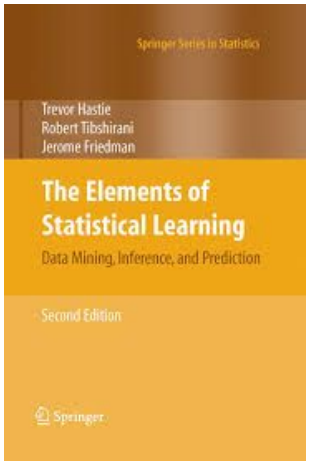
- **Assumed Background:** Intro probability, statistics, programming, and CS theory.



COMPSCI 589: Machine Learning

- **Program/Level:** Seniors CS Major; MS CS; Other MS/PhD
- **Topics:** Broad survey of supervised and unsupervised learning methods. Understanding mathematical and geometric properties of individual models and relationships between models. Issues of model selection, capacity control, design of machine learning experiments, scalability.
- **Focus:** Theory and applications.

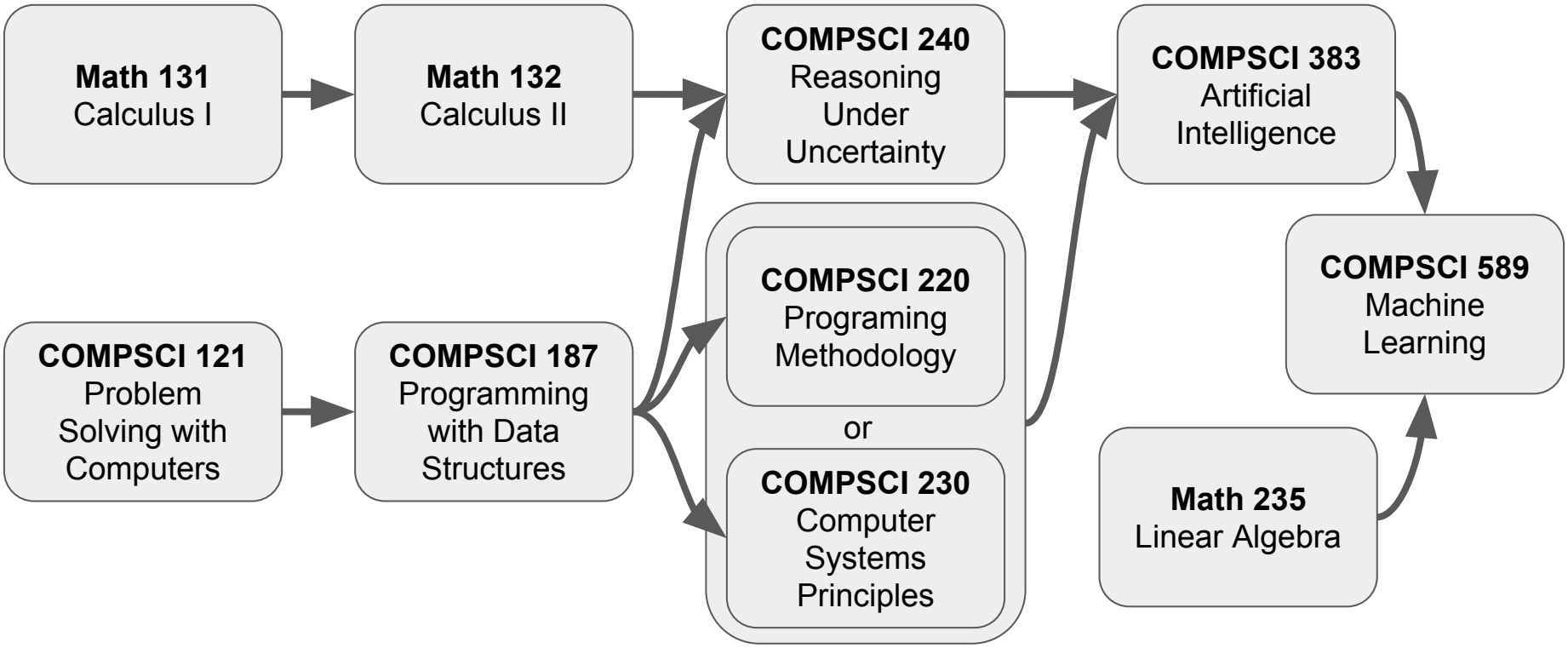




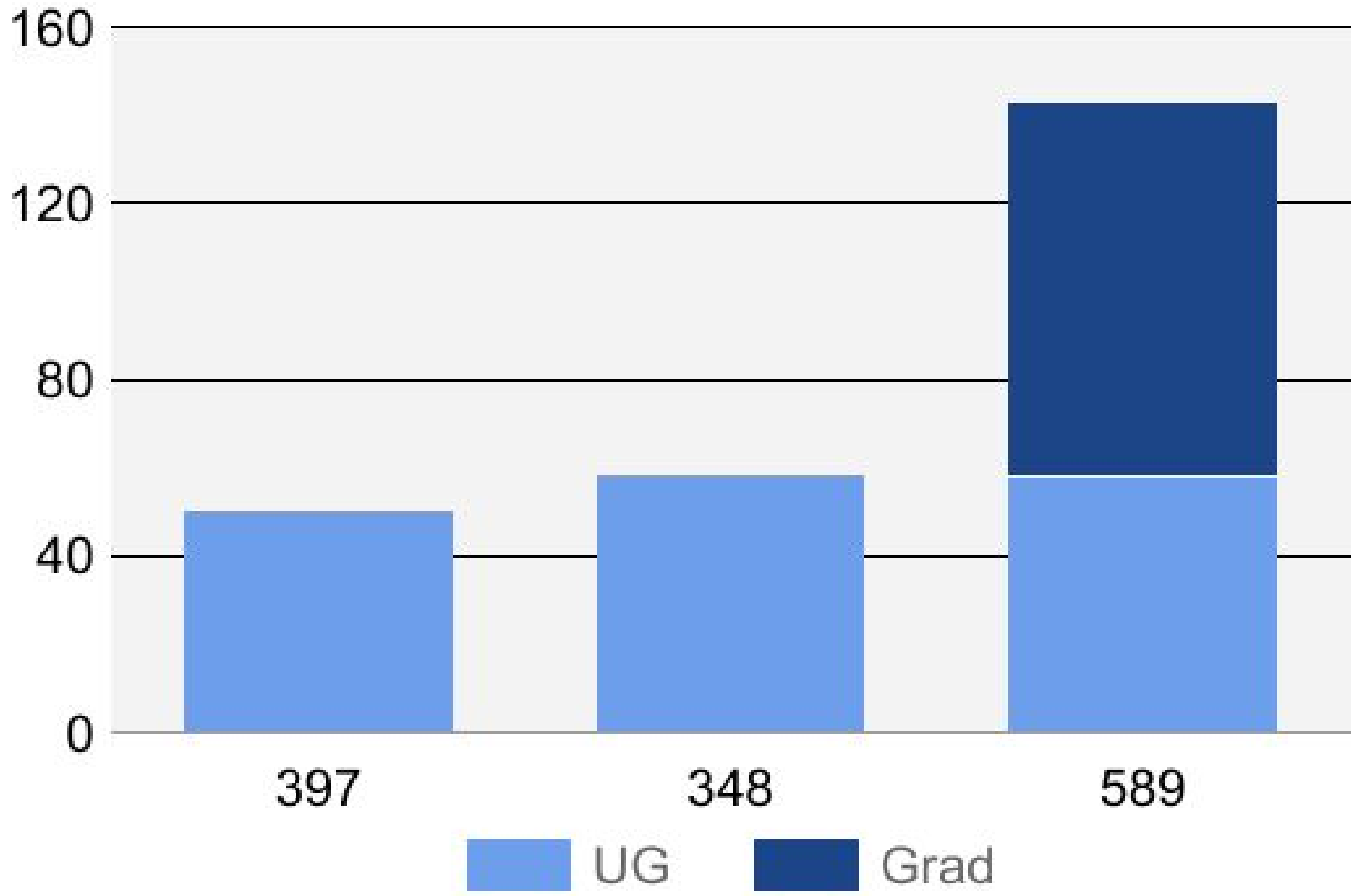
Open Source Materials: https://github.com/mlds-lab/COMPSCI-589

COMPSCI 589: Machine Learning

- **Assumed Background:** Differential calculus, programming, data structures, probability, statistics, programming languages, linear algebra, and AI.



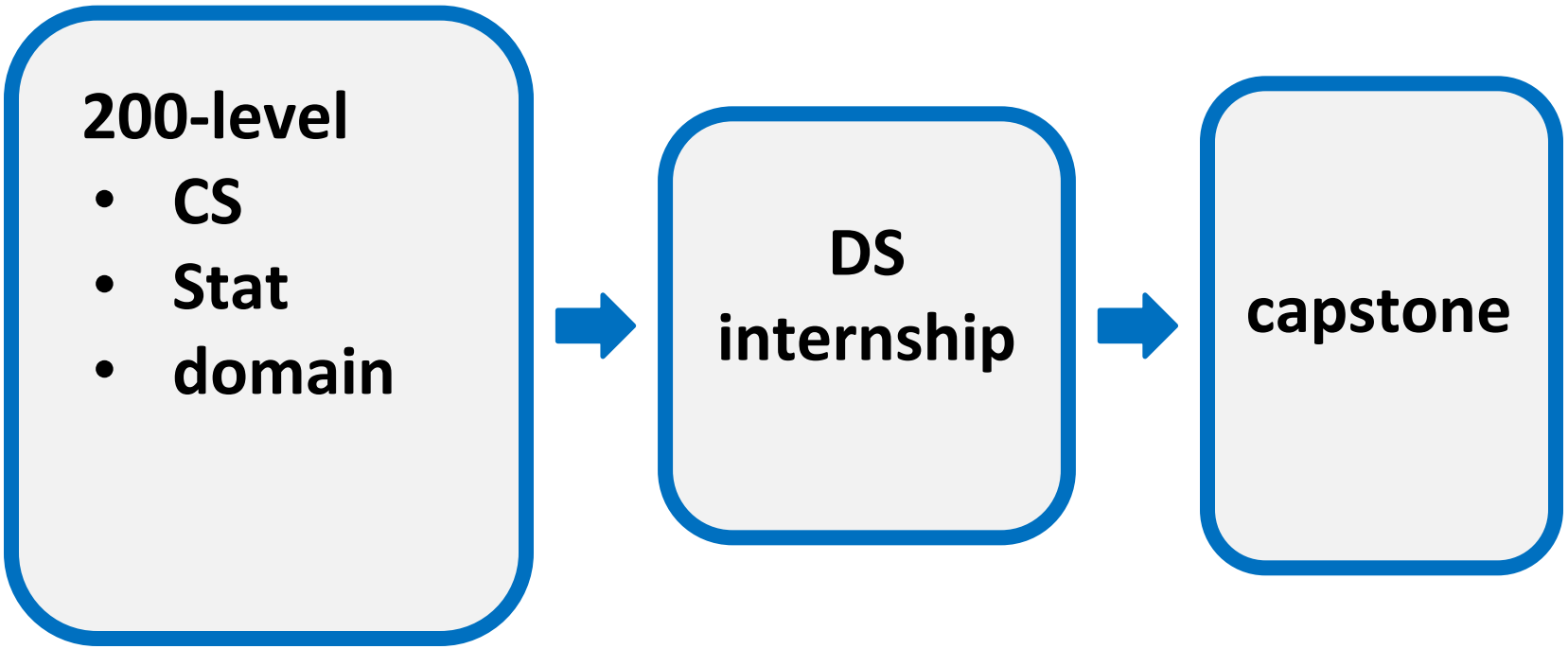
UMass Data Science Course Sizes



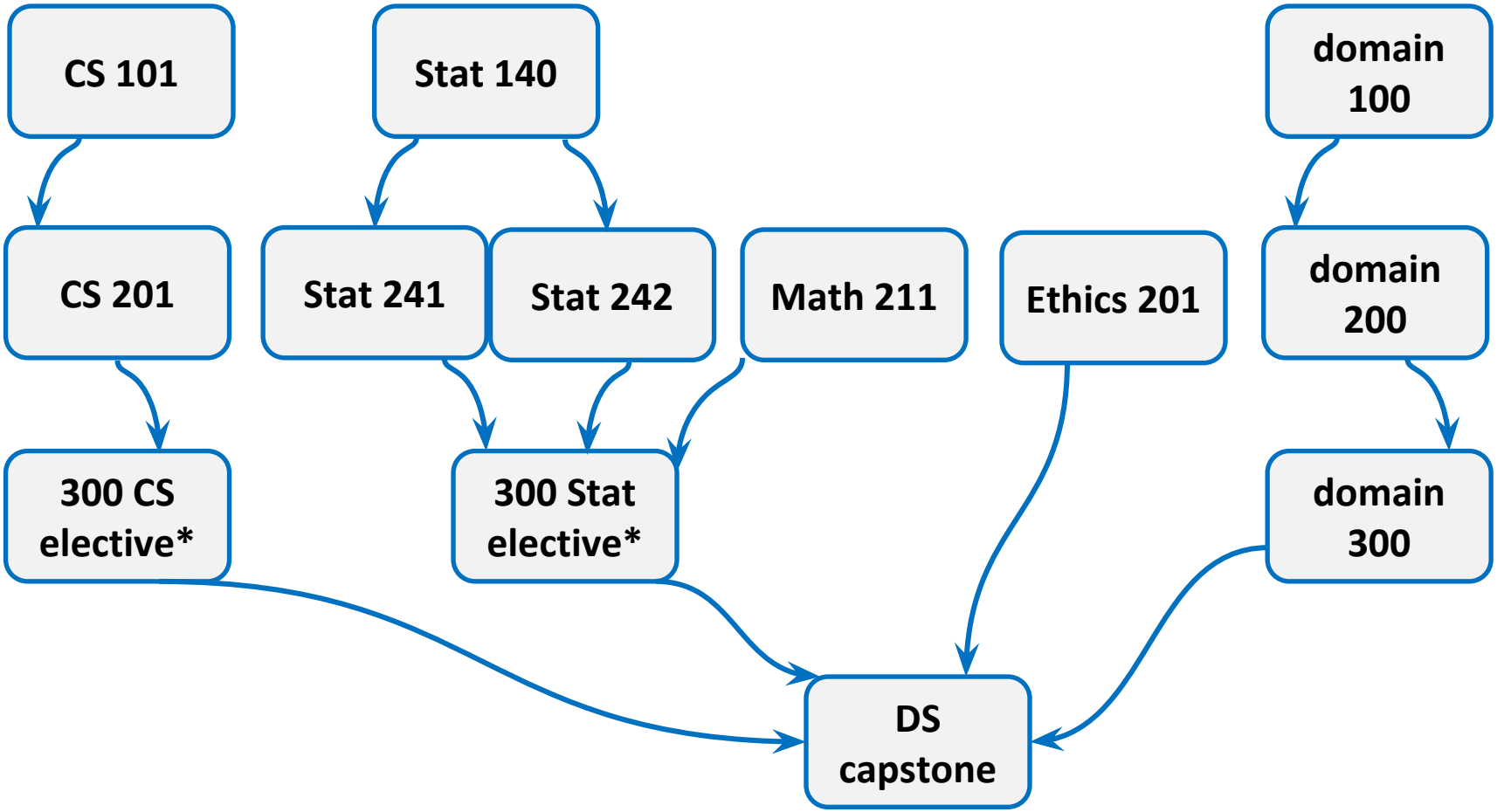
Data Science at Mount Holyoke College

Presenter: Martha Hoopes
Contributors: MHC DS Curriculum Committeee – Tim Chumley, Amber Douglas, Andrea Foulkes, Janice Gifford, Barb Lerner, Tim Malacarne, Eitan Mendelowitz, Steve Schmeiser, Sam Tuttle

Nexus in Data Science

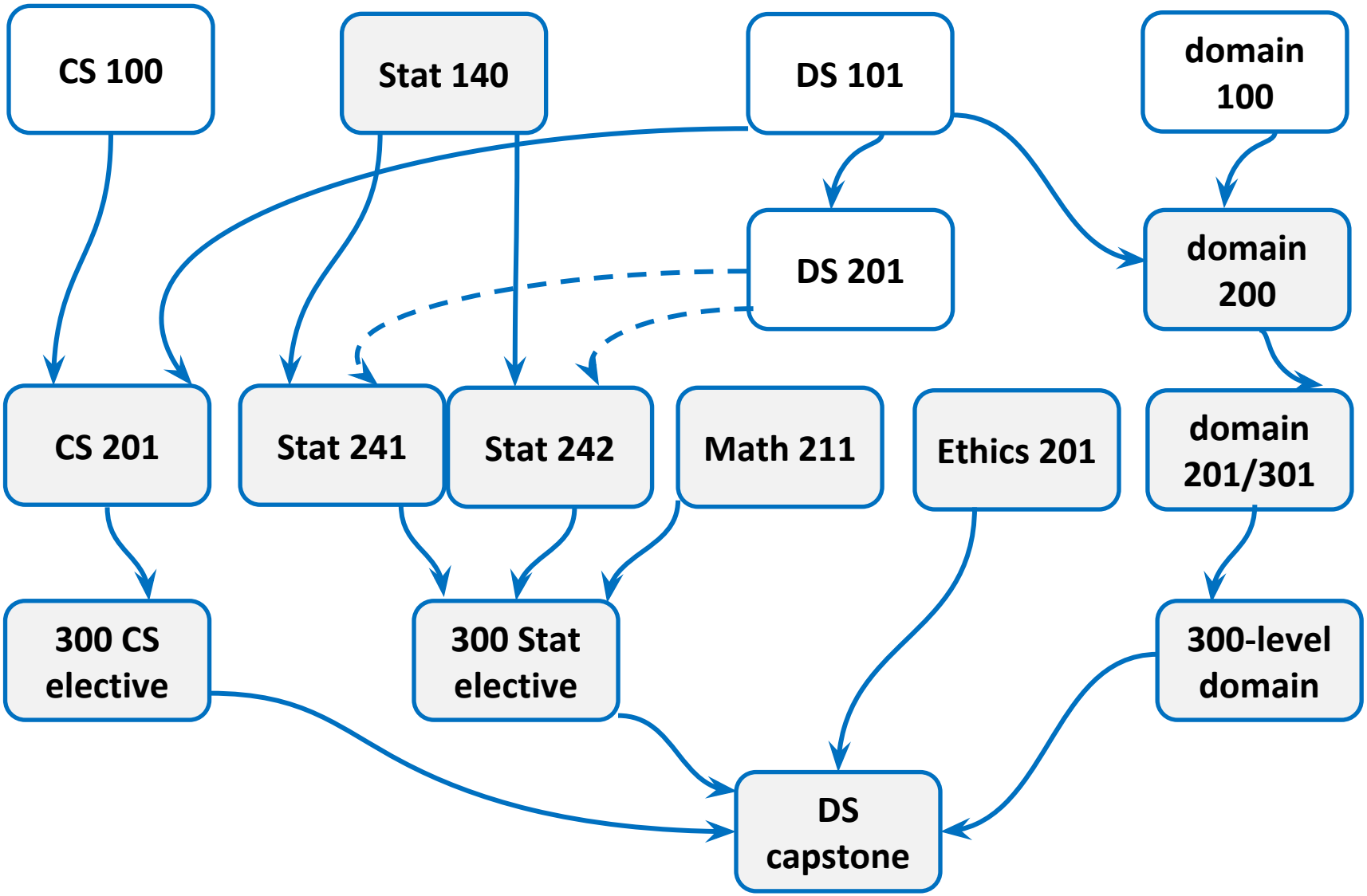


DS major using only existing courses

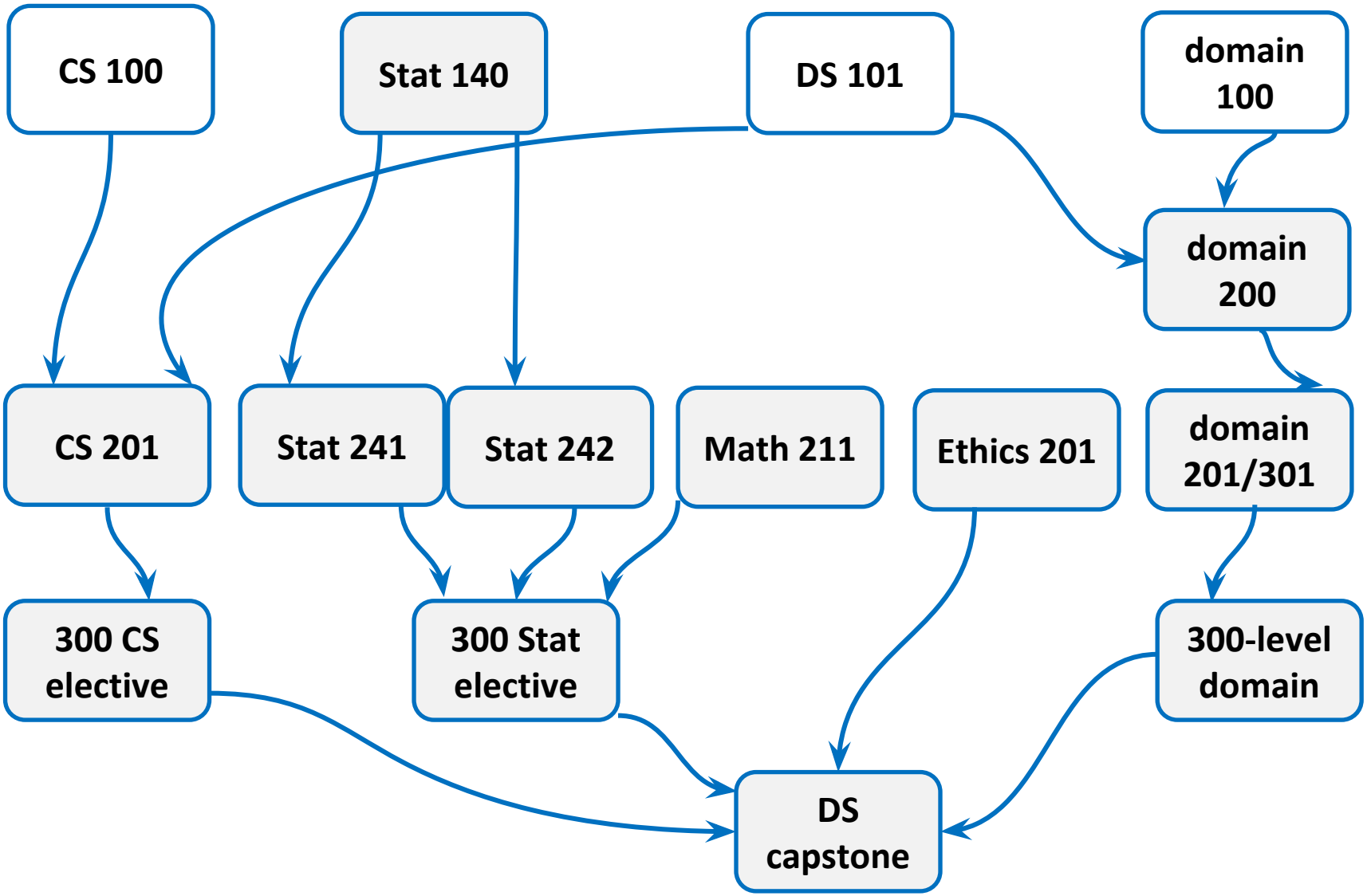


*choose one of these electives

DS major with new courses



DS major with new courses



Essential elements of intro DS courses

"computation thinking"
- comfortable with splitting a big problem into approachable pieces
- comfortable with basic programming constructs (functions, loops, branching)

"inferential thinking“ and “analysis thinking”
- basic statistical concepts (randomization, populations, models, SD, mean, variation, hypothesis testing)
- models, parameterization, and estimation

“data awareness”
- sources, manipulation, management
- ethics of use

“communication thinking”
- visualization
- written and oral presentations and interpretations

proficiency with R and/or python and some DS related libraries

project and data from real world and relevant domain questions

Introductory Data Science at Smith College

Ben Baumer
Statistical and Data Sciences
Smith College

A brief history of data science at Smith

- MTH 292 (Data Science)
  - prereqs: intro stats + programming experience
  - Fall 2013: enrollment 18
  - Fall 2014: enrollment 23
  - *A Data Science Course for Undergraduates: Thinking with Data*, TAS 69 (4), 2015
- SDS 192 (Introduction to Data Science)
  - PKAL/TIDES grant to promote diversity and achievement in STEM
  - prereqs: none -- but willingness to learn to code
  - Spring 2016: enrollment 25
  - Spring 2017: enrollment **89**
- SDS major approved in late spring 2016
  - SDS 192 required for the major

SDS 192: Foundational Skills

Data Visualization

~ 3 weeks

Data Wrangling

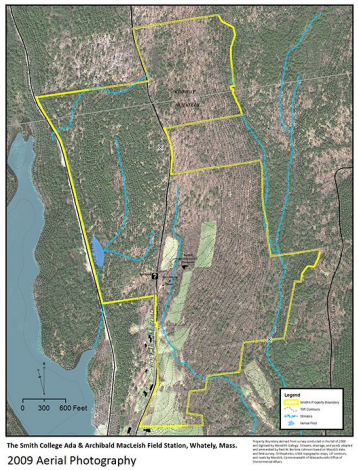
~ 4 weeks

Reproducible Workflow

SDS 192: Liberal Arts Modules

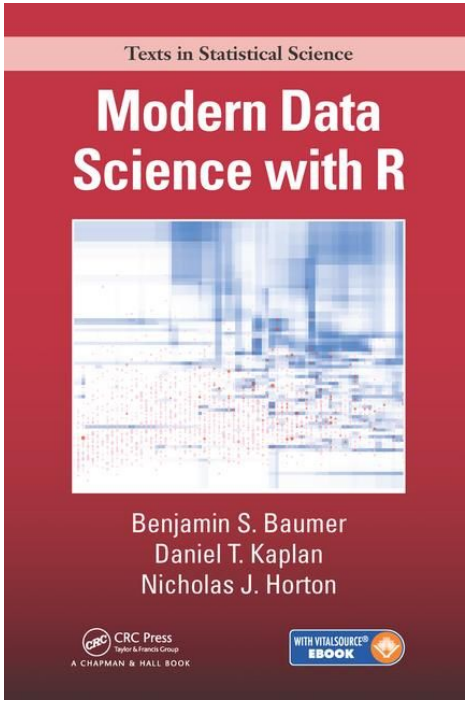
- “Sequels with SQL”
- Alex Keller, Film Studies
- Q:
  - Horrors and subgenres
  - Westerns over time
  - Genre multiplicity

- Space at MacLeish
- Reid Bertone-Johnson, Landscape Studies, MacLeish field station manager
- Q:
  - Optimal 2nd campsite location
  - algorithmic trail difficulty
  - propose new trails



SDS 192: Modern tools

- New book!



- Collaborative Projects

GitHub

- Intrateam Communication

slack

- Online learning

DataCamp

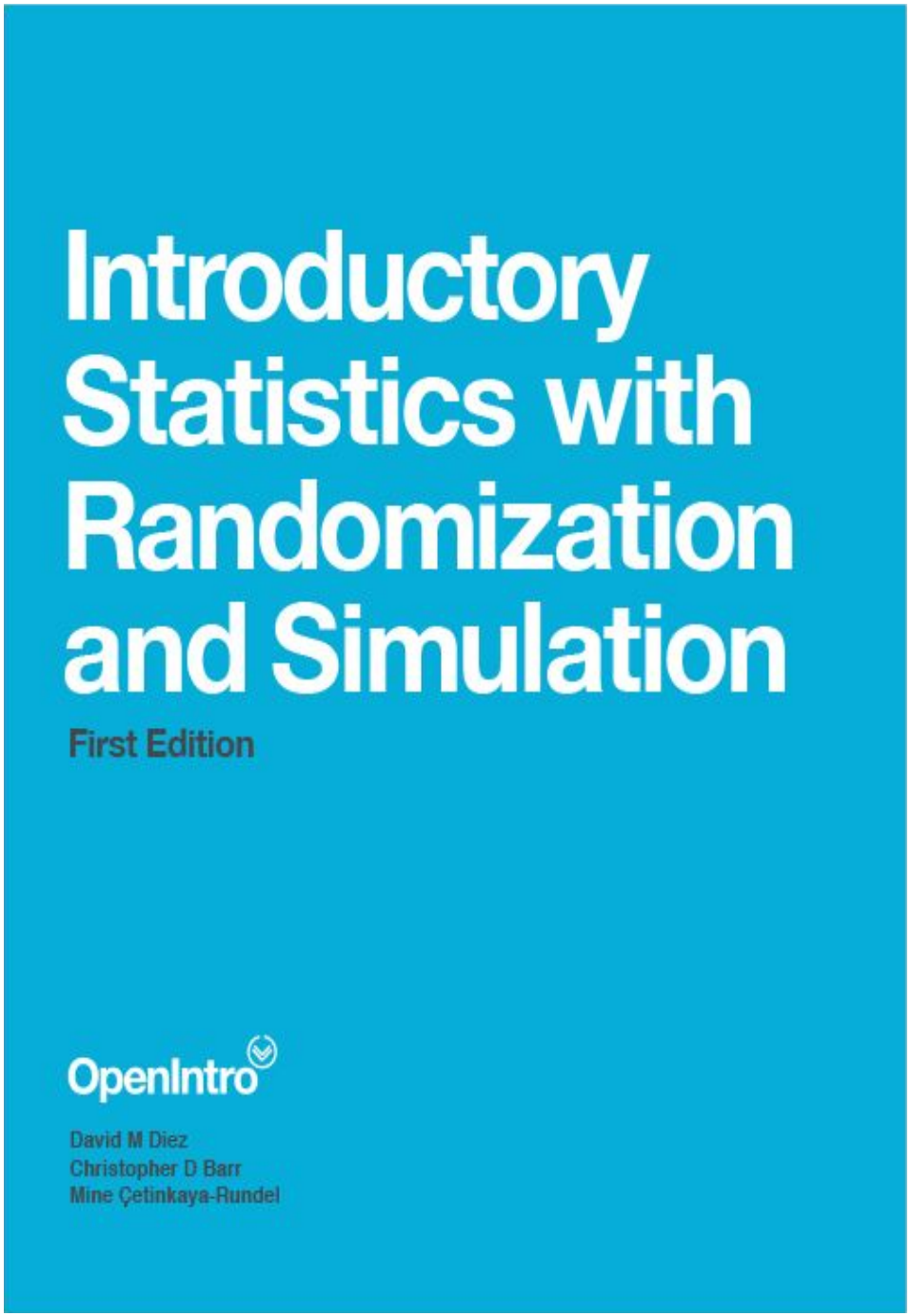
https://beanumber.github.io/sds192/

Introductory Statistics at Smith College

Amelia McNamara
Statistics and Data Sciences
Smith College

Textbook: Introductory Statistics with **Randomization** and **Simulation**

Available as PDF, tablet-friendly PDF, source (GitHub), paper book ($9 on Amazon)



Integrated computation:

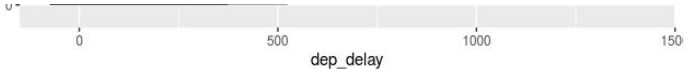
- Weekly labs
- R and RStudio
- RMarkdown assignments

“Flavors” of labs

- `base` (deprecated?)
- `tidyverse` (`ggplot2`, `dplyr`, etc)
- `mosaic`

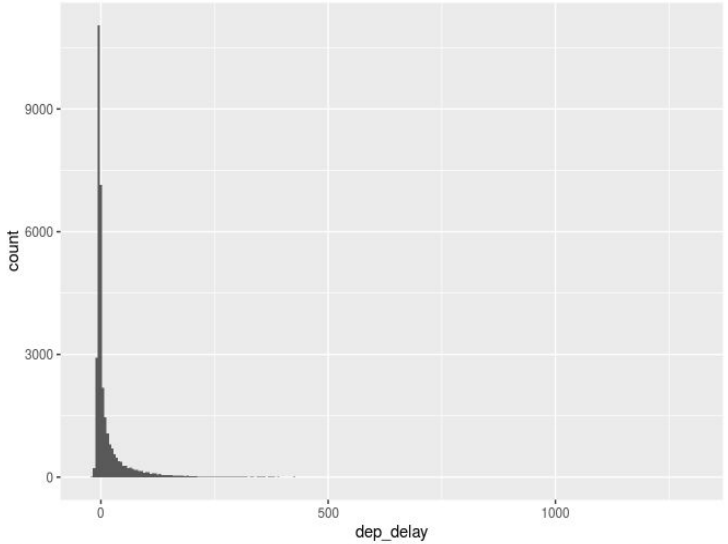
Some convergence

Exercise 1:
Exercise 2:
Exercise 3:
Exercise 4:
Exercise 5:
Exercise 6:
Exercise 7:
Exercise 8:
Exercise 9:



Hide

```
qplot(x = dep_delay, data = nycflights, geom = "histogram", binwidth=5)
```



When the binwidth is 15, we see that there are some flights that leave early (a negative binwidth). And with a binwidth of 5 minute increments, we see more detail and with a bin width of 150, we see much less detail. The shape of the data is skewed right and has a mode of 0-5 minutes delayed.

Exercise 2:

Create a new data frame that includes flights headed to SFO in February, and save this data frame as sfo_feb_flights. How many flights meet these criteria?

Hide

```
sfo_feb_flights<-nycflights %>%
  filter(dest=="SFO", month==2)

nrow(sfo_feb_flights)
```

```
## [1] 68
```

68 flights meet these criteria, I found this by using the nrow() command.

Final project

- Groups of ~3 students
- Student-chosen data
- Typically multiple regression
- Data cleaning

Contextualizing Ebola Frenzy

some Smith students

ABSTRACT

The panic from the onset of an Ebola epidemic in West Africa has had reverberations worldwide, but especially in the American media. Does this apparent surge of media attention on Ebola reflect an unusual response to pandemic disease, or is this response expected based on the characteristics of the disease itself? This study sought to analyze media attention surrounding Ebola by comparing its associated media response, proxied by scaled interest from Google Trends, to an expected range of interest. Media attention was broken down into two timescales - the month of November 2014 and the past year - to clarify discrepancies between media response in the short and long term. To generate the expected range of interest, a multiple linear regression model was built to predict proxied media response as a function of common characteristics of pandemics. This model predicted that media response to Ebola did not fall outside of this expected range; however, shortcomings in data collection and violations of assumptions for inference from multiple linear regression render these results inconclusive.

INTRODUCTION

Since the outbreak of Ebola in the summer of 2014, what was once an obscure disease has turned into media frenzy on the tips of everyone's tongues. Suddenly panic was in the air as people made apocalyptic preparations. The public feared that a virus of which they had only recently become aware could leach into their water, into their air, invading their bodies and dissolving their organs into sludge. As the news reports catalogued the growing death tolls thousands of miles and an ocean away, hysteria spread on American soil. This is certainly neither the first nor the last time that pandemic disease will be accompanied by an equally contagious anxiety. So is the attention surrounding Ebola a normal response warranted by the characteristics of the disease itself, or does this trend in media response indicate a more extreme reaction?

In this observational study, we used the characteristics for 23 pandemic diseases to build a model predicting the media response to a disease as proxied by scaled data from Google Trends. We analyzed the media response to Ebola both in the short term (the month of November 2014) and in the long term (the previous year) to elucidate any trends or discrepancies determining media response over time.

DATA

The observational units of this study are pandemic diseases. These diseases represent a sample of all pandemic diseases, based on availability of data, and are not restricted to those in any particular area of the world. However, some of these diseases do not have global spread. The causes of these diseases are of either bacterial or viral nature. While the features that characterize diseases are complex, we limit our characterization to a few parameters based on the availability of data. The data was derived from various credible online sources, primarily from a project called Microbe-scope, the Center of Disease Control, the World Health Organization, and the National Center for Biotechnology Information.

```
mydata <- read.table("StatsGroupProjectNewData.csv", header=TRUE, sep=",")
head(mydata, n=5)
```

```
##           Disease CaseFatality ReproductionNum YearlyFatal YearlyFatalUS
## 1 Bird Flu (H5N1)       0.6000            1.00          18             0
## 2         Cholera       0.0163            2.13       94000             0
## 3      Diphtheria       0.0750            6.50        5000             1
## 4          E.coli       0.0400            1.15      260000           100
## 5     Hepatitis B       0.0075            4.04      600000          3000
##   YearlyFatalAfrica Cure Prevent SearchNov SearchYear
## 1                 3    0       1         8         47
## 2             55000    1       0        28        440
## 3              1700    0       0        37       1797
## 4            103000    1       1        32        540
## 5             18600    1       0       168       2807
```

A portion of our data is shown above. Note that we do not include Ebola in this dataset since its inclusion would artificially improve the model's fit to its parameters.

The case fatality rate is a percentage of the total cases for a disease which resulted in death.The average basic reproduction number (R_0) describes the contagiousness of the disease; it is the number of cases one case generates on average over the course of its infectious period, in an otherwise uninfected population. Thus a greater R_0 value implies greater potential for transmission. The yearly fatalities of each disease, the number of deaths, were found for three different populations pertinent to the analysis of Ebola: worldwide, in the United States, and in Africa. The availability of a cure or vaccine describes the accessible means to successfully treat and prevent a given disease. These treatments must not only have been developed, but also be approved by the FDA and available for distribution. Thus, a drug may have been developed for a certain virus, but if it is not available in the US, it was not considered for the purposes of this study. Both of these are categorical variables, which take binary values 0 (cure or vaccine is available) or 1 (cure or vaccine is not available). This distinction is based on the assumption that the unavailability of a vaccine or cure would cause greater media attention.

Guidelines for Undergraduate Data Science Programs

Albert Y. Kim
Middlebury College
http://www.amstat.org/asa/files/pdfs/EDU-DataScienceGuidelines.pdf

UMass
Center for Data Science

~30 faculty in UMass CICS
~250 faculty across the Five Colleges
~20 industry partners

Andrew McCallum
Computer Science
UMass CS

UMass *Center for Data Science*
Industry Outreach (related to students)

Regular industry presentations on campus

weekly “Machine Learning & Friends” lunchtime seminar series

Google, IBM, Microsoft, Twitter, Amazon, NVIDIA,...

Industry Mentors for “Practicum” Independent Study projects

Semester long projects in team of 4-5 students.

Oracle, Lexalytics, Meta, … Burning Glass, … MITRE, Twitter,...

Summer internship match-making help

Especially helpful for inexperienced students (and companies!)

Data Science Career Fair

with student project/research posters

MassMutual Development Program

~15 MassMutual employees in UMass CS MS Data Science program

UMass *Center for Data Science*
Education Outreach

Facilitating education coordination between UMass departments:

Facilitating coordination between UMass departments:
CS, Stats, Biostat, Eng, Management, Social Sciences, Natural Sciences

New CS *MS Concentration in Data Science*

Double number of data science MS courses:

Machine Learning, DS Systems, DS Algorithms, Visualization

Triple number of students in the MS program

New CS+Stats *Certificate in Statistical & Computational Data Science*

Data Science "clubs" for undergrads and MS students.

Computing Hardware

~500-core cluster (old), ~2000-core cluster (new), ~500 GPU cluster

Future: **Short courses for cross-department education**.

Future: **Freshman-level, non-prereq data science course, together with Stats**.

Berkeley's Data 8 for the Five Colleges

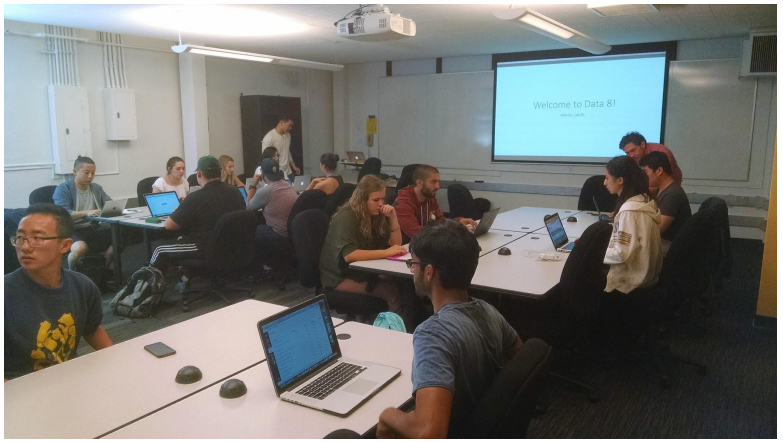
GOALS:
Statistical and Computational Literacy for all undergrads, with zero pre-recs.
Support for many other majors interested in building on data science.
Enticement into the Statistics and Computer Science majors.

Patrick Flaherty
Statistics
UMass

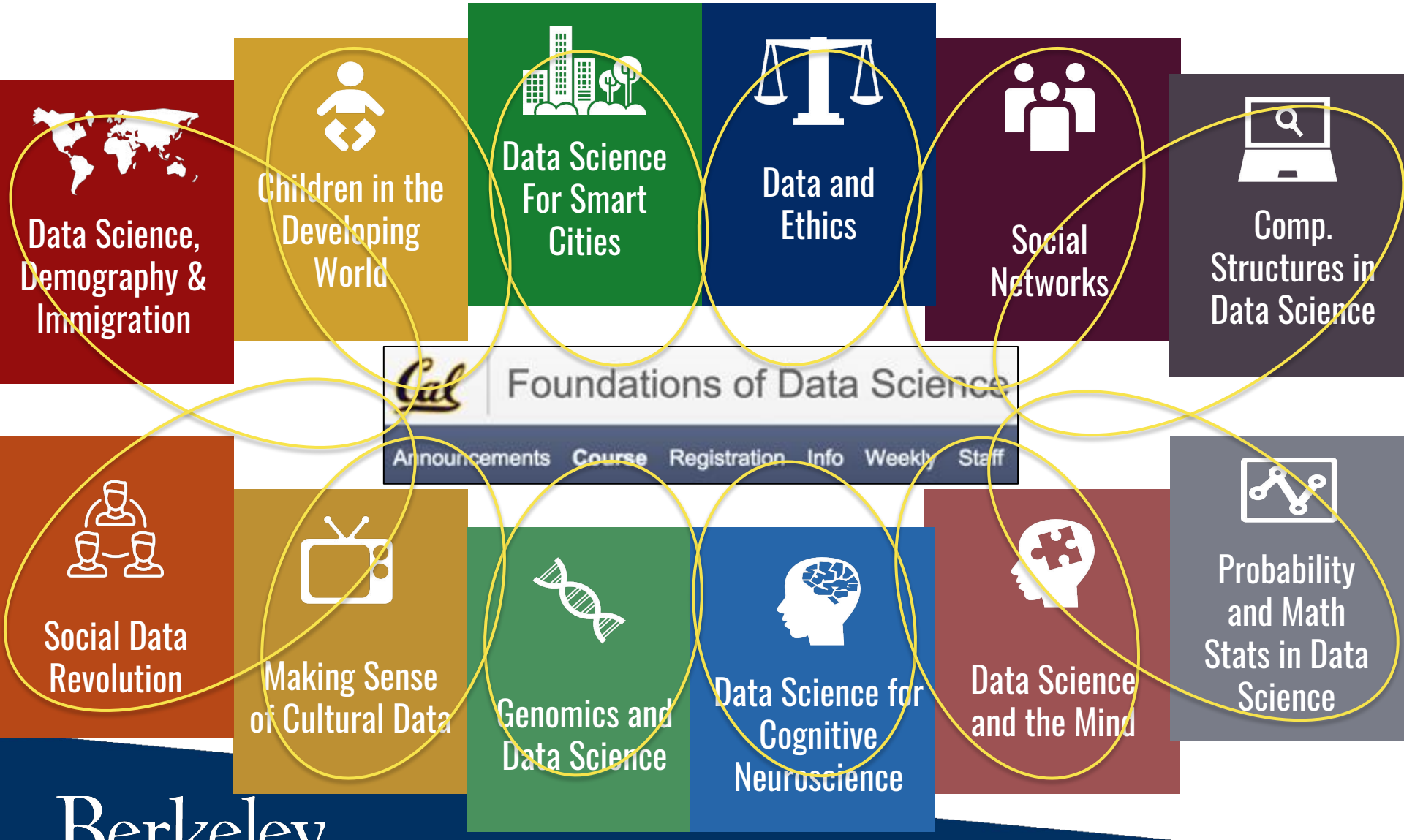
Andrew McCallum
Computer Science
UMass CS

Foundations of Data Science - “Data 8”

- Designed for 1st and 2nd-year students of **any** major
  - 4 units, MWF lecture + 2-hour W-F lab
  - No CS or stats experience required
- Students gain hands-on experience with *real data* while learning computing and stats concepts + programming
- Strong support (OHs, HKN/CS Mentors tutoring, Data Scholars program, laptop loans, etc.)
- 500+ students per semester & growing
- M-Tu Connectors



Data Science Connectors



Berkeley
UNIVERSITY OF CALIFORNIA

Syllabus

Cause & Effect

Expressions & data types

Tables, rows, charts, histograms

Functions, groups, joins, iteration

Privacy

Randomness, sampling, variability

Hypothesis testing, error probs

Confidence intervals & testing

Center & spread

Normal distribution, sample means

Correlation

Linear regression, least squares

Classification, features

Optimization

Comparing samples, A/B testing

Causality

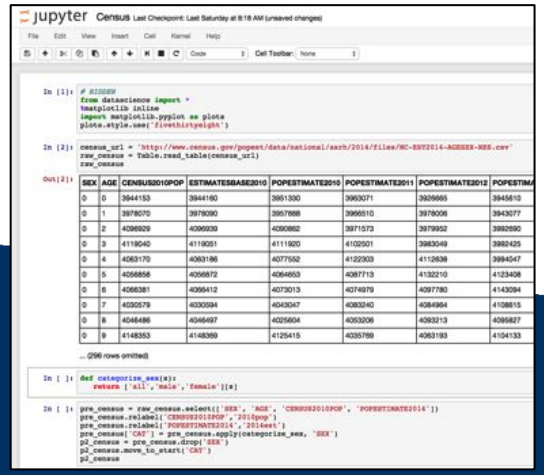
Decisions

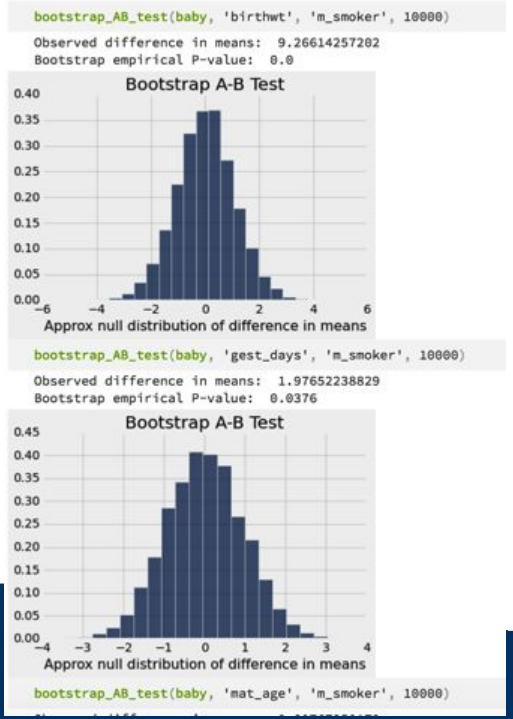
Coding in Python (in browser, with Jupiter). 10 Labs. 3 Projects.

Data8 - Concepts and Computing

- Fundamental co-mingling of CS & Stat concepts on real data
  - Learn computing concepts by doing interesting things on data
  - Learn advanced statistical concepts by observing what's interesting
  - Codify understanding of concepts symbolically
- "Explorations" in visualization, privacy, personalized medicine,…
- Entirely cloud-based computing environment built on Python Jupyter notebooks plus UCB datascience Tables

http://data8.org





Tables (https://github.com/dsten/datascience)

ordered collection of labeled columns of anything

dict, record,tuple

label

values

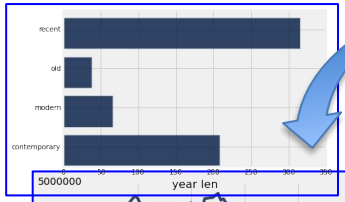
select, where, take, drop, group
stats, bin
sample
pivot,
pivot_bin

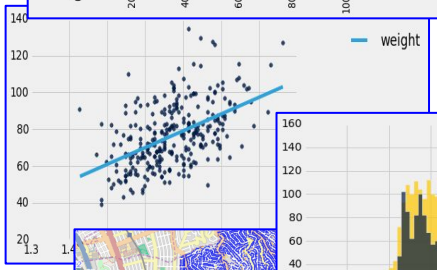
split

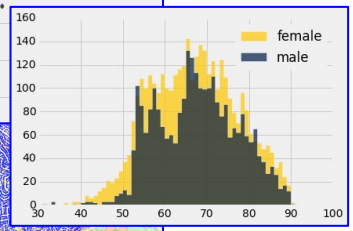
join

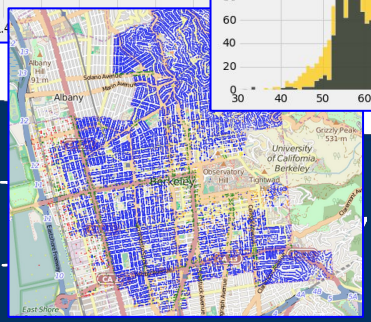
T['label']

Numpy array





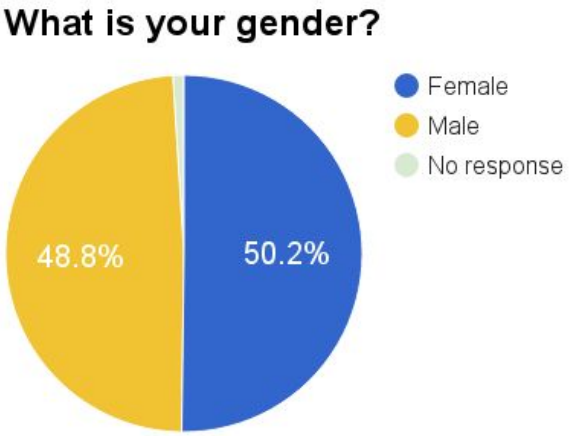


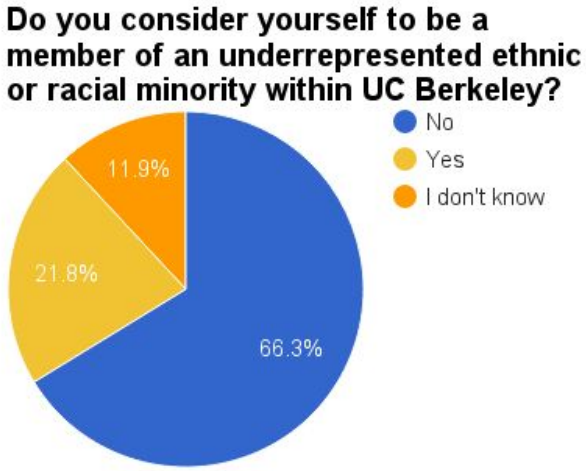


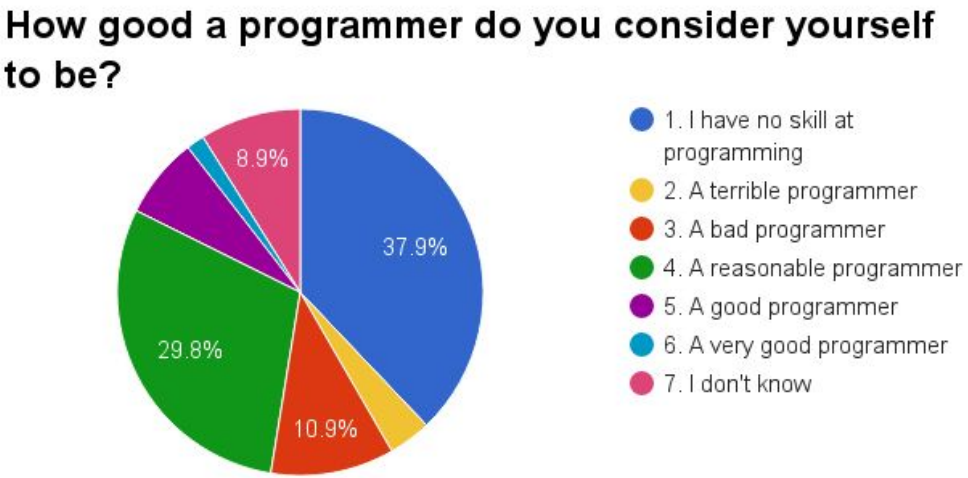
- A single, simple, powerful data structure for all
- Inspired by Excel, SQL, R, Pandas, Numpy, …

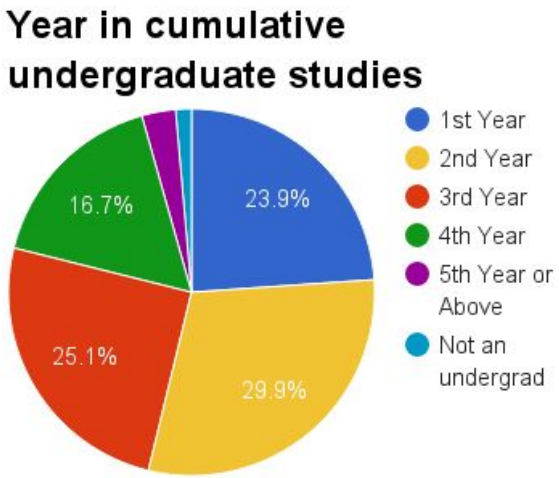
Data 8 Demographics, Fall 2016

- 517 students

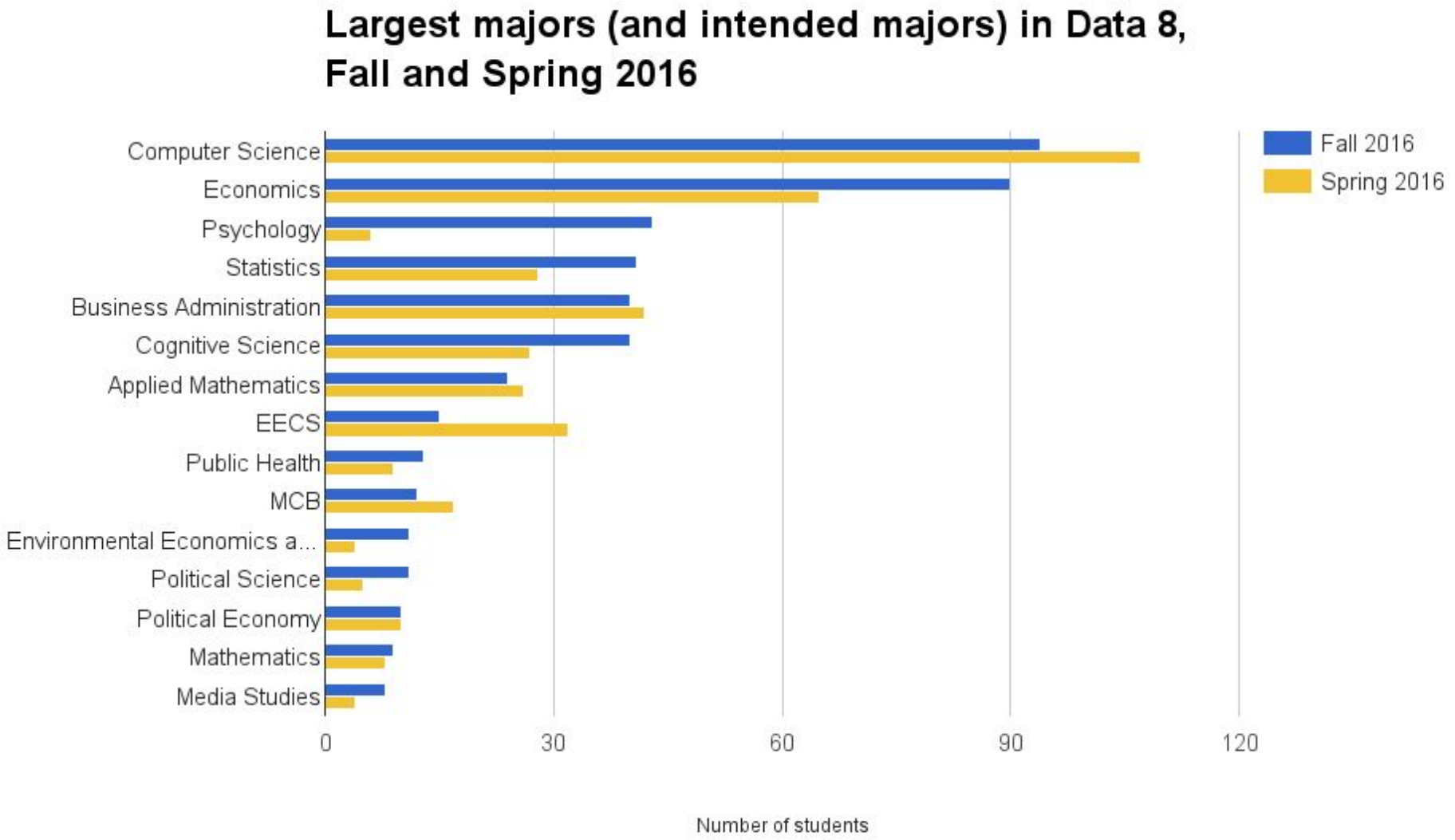








Majors composition of Data 8 (of 59 Fa16)



Bringing “Data 8” to UMass & the Five Colleges

Collaboration between CS and Statistics. Cross-listed course.
Support from both departments.

Would form part of the Informatics major.

Computing resources for large-scale Jupiter installation available from CS.

Targeting Spring 2018. Proposal to faculty senate before May 2017.

Connectors? Outreach in progress, but will follow our initial offering.

Questions: Interest in “Data 8” from Smith, Mt. Holyoke, Amherst, Hampshire?
Your students might take “Data 8” at UMass?
You would offer your own “Data 8”?
Connectors across the Five Colleges?

Five College Undergraduate Data Science Education

Discussion

Nicholas Horton
Statistics
Amherst College

Andrew McCallum
Computer Science
UMass

Some questions to consider...

1. What should be our goals in coordinating across the Five Colleges?
2. How can we facilitate more cross-department coordination in data science?
3. What is the [level/nature] of demand for data science at your institution?
4. Should data science be taught at the freshman level?

5. What's an optimal sequence or combination of courses that link research in a discipline to tools for data science?
6. What are the key concepts from traditional (statistics and computer science) fields that we want students to take away from the first course(s) in this pathway?
7. How much more advanced to anticipate undergraduates will be compared to current graduates by introducing concepts earlier?
8. How does field-specific knowledge fit into data science education?
9. What mathematical foundations should students get before entering introductory data science courses?

Next Steps...

More discussion across the Five Colleges and across departments.
(Open, friendly discussion among Stats, CS, InfoSci, and application areas was very important at Berkeley.)

Launch “Data 8” analogue at UMass as a CS+Informatics+Stats joint effort.
Coordinate with other Five Colleges as they wish.

Gather related course syllabi from across the Five Colleges in Nick’s GitHub repo.

...