DATA SCIENCE FOR ALL
SURE, BUT WHO, WHERE, WHEN AND HOW MUCH? OR...
LET’S PUT THE DATA BACK INTO DATA SCIENCE

Richard D. De Veaux
Williams College
IASE Satellite ISI
August, 2019
deveaux@williams.edu
Here’s what keeps me up at night

• Data Science courses — with no “data”
• Our Intro Stats course becoming even less relevant to students’ needs
• Students thinking that the world (or at least the Statistics world) is univariate
• That we are teaching the same course we taught in 1958 — or even 1996
• That we have replaced Math envy with CS envy
WHAT IS DATA SCIENCE?

A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician.
Data science is a method for gleaning insights from structured and unstructured data using approaches ranging from statistical analysis to machine learning.
WHAT IS DATA SCIENCE III?
OUR STUDENTS?

Thomasine lands her dream job — analyst for H&M
Q: How much of their resources should they put online vs. brick and mortar?

H&M's growth in stores vs. overall operating profit growth
AND THEN...
WHAT COURSE(S) ARE WE TALKING ABOUT?

• Intro to Data Science?
• The Intro Course that covers statistical thinking, computing and data curation, architectures and storage is a unicorn.
• What can we cover?
HOW MUCH CODING? — CS ENVY?

WHEN I STARTED PROGRAMMING, WE DIDN'T HAVE ANY OF THESE SISsy "ICONS" AND "WINDOWS."
I CAN PROGRAM...

WHEN I STARTED PROGRAMMING, WE DIDN'T HAVE ANY OF THESE SISY "ICONS" AND "WINDOWS."

ALL WE HAD WERE ZEROS AND ONES -- AND SOMETIMES WE DIDN'T EVEN HAVE ONES.
YOU HAD 0’S?

WHEN I STARTED PROGRAMMING, WE DIDN’T HAVE ANY OF THESE SISY “ICONS” AND “WINDOWS.”

ALL WE HAD WERE ZEROS AND ONES — AND SOMETIMES WE DIDN’T EVEN HAVE ONES.

I WROTE AN ENTIRE DATABASE PROGRAM USING ONLY ZEROS.

YOU HAD ZEROS? WE HAD TO USE THE LETTER “O.”
R VS PYTHON VS JMP (TABLEAU ETC)

• Each has its advantages.
• Eventually, a data science student should see all of these.
• The beginning student?
  Teach the power of Statistics not the mechanics.
• Which to start with?
  Data 8 course.
# Count how many times the names Jim, Tom, and Huck appear in each chapter.

```r
counts = Table().with_columns(
    'Jim', np.char.count(huck_finn.chapters, 'Jim'),
    'Tom', np.char.count(huck_finn.chapters, 'Tom'),
    'Huck', np.char.count(huck_finn.chapters, 'Huck')
)
```

# Plot the cumulative counts:
# how many times in Chapter 1, how many times in Chapters 1 and 2, and so on.

```r
cum_counts = counts.cumsum().with_column('Chapter', np.arange(1, 45, 1))
cum_counts.plot(column='x.ticks=3')
plots.title('Cumulative Number of Times Each Name Appears', y=1.08);
```

In the plot above, the horizontal axis shows chapter numbers and the vertical axis shows how many times each character has been mentioned up to and including that chapter.
WHAT NOT TO TEACH?

3.6 Tutorial: Histogram Construction

```python
path = r'../Data/'  # Set the path to match your data directory.
fileList = os.listdir(path)  # Creates a list of files in path
for filename in fileList:
    try:
        shortYear = int(filename[6:8])
        year = 2000 + shortYear

        fields = functions.fieldDict[shortYear]
        sWt, eWt = fields['weight']
        sBMI, eBMI = fields['bmi']

        file = path+filename
        print(file, sWt, eWt, sBMI, eBMI)
    except(ValueError, KeyError):
        pass
```

EMPOWER STUDENTS

CODAP NHANES

http://datascience.la/introduction-to-data-science-for-high-school-students

ISLE — Carnegie Mellon
WHAT ISN’T DATA SCIENCE?

• Some elementary coding
• The bits from statistics the don’t require thinking
  • Exploratory Data Analysis
  • Summary Statistics
  • Machine Learning Algorithms

“Nowadays anyone with a laptop and a script can scrape data off the Internet, feed it into an R package, and publish the results. Obviously this isn't data science, but the average citizen isn't going to know the difference.”
THE REAL WORK OF DATA SCIENCE

- Helping to formulate the problem
- Understanding which data to consider and the strengths and limitations in the data
- Determining when new data are needed
- Making clear where the data ends and “intuition” takes over
- Presenting results
- Recognizing that practical decisions involve more than data
LIFE CYCLE OF DATA SCIENCE
HOW DO WE GET THERE?

Curriculum Guidelines for Undergraduate Programs in Data Science*

PARK CITY REPORT

Park City Report identified the following key competencies for a Data Science major.

• Computational and statistical thinking
• Mathematical foundations
• Model building and assessment
• Algorithms and software foundation
• Data curation
• Knowledge transference—communication and responsibility
This ACM Data Science report builds on the Park City work with a heavy orientation toward computer science.

The position of the Task Force is that any Data Science program will have to reflect competencies in mathematics, statistics, and computer science, possibly with different emphases.
CORE COMPETENCIES

• Computing Fundamentals, including Programming, Data Structures, Algorithms, and Software Engineering
• Data Acquisition and Governance
• Data Management, Storage, and Retrieval
• Data Privacy, Security, and Integrity
• Machine Learning
• Data Mining
• Big Data, including Complexity, Distributed Systems, Parallel Computing, and High Performance Computing
• Analysis and Presentation, including Human-Computer Interaction and Visualization
• Professionalism

Other areas of computing may merit attention: sensors and sensor networks, the Internet of Things, vision systems, among others.
Data structures capture common ways in which to store and manipulate data, and they are important in the construction of sophisticated computer programs.

Students are introduced to some of the most important and frequently used data structures: lists, stacks, queues, trees, hash tables, graphs, and files.

Students will be expected to write several programs, ranging from very short programs to more elaborate systems. Emphasis will be placed on the development of clear, modular programs that are easy to read, debug, verify, analyze, and modify.
THREE GROUPS OF STUDENTS

• The usual suspects
  • Our current CS, Stat majors
• Science oriented students
  • Who will use DS
• Everyone else
1. Recognize pervasiveness of uncertainty and basic probability concepts
2. Understand sample and population and appropriateness of data
3. Two types of errors and consequences
4. Basic inference and causation vs. association
PRODUCERS OR CONSUMERS?

How to teach a lay up?
Who's the audience?
  • spectators
  • referees
  • players
    • beginners
    • pros

Roxy Peck
A CAUTIONARY TALE

• 10,700 houses collected from Saratoga NY public records by my former student Candice Corvetti for her senior thesis

Candice M. Corvetti
Principal

Candice joined Berkshire Partners in 2014. Prior to Berkshire, she worked at Madison Dearborn Partners. Candice started her career as an analyst at J.P. Morgan.

Education
Williams College, B.A.
Stanford Graduate School of Business, M.B.A.
## DATA SCIENCE

![Spreadsheet Image]

### Table: Data Example

<table>
<thead>
<tr>
<th>SWIS Master</th>
<th>SWIS CD</th>
<th>SWIS TOWN</th>
<th>SWIS SMS</th>
<th>SWIS ID</th>
<th>STAR</th>
<th>ADJ2</th>
<th>ADJ5</th>
<th>LOLL VR</th>
<th>ROLL VR</th>
<th>PLUS</th>
<th>ONE AHER ID</th>
<th>OWNER</th>
<th>TYPE</th>
<th>OWN CODE</th>
<th>OWNER FIRST NAME</th>
<th>OWNER INITIAL NAME</th>
<th>OWNER LAST NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>34</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2005</td>
<td>1</td>
<td>1</td>
<td>2005</td>
<td>1</td>
<td>11556</td>
<td>A</td>
<td>Renes</td>
<td>Bankhead</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>0</td>
<td>1414</td>
<td>0</td>
<td>1</td>
<td>2004</td>
<td>1</td>
<td>1</td>
<td>2005</td>
<td>1</td>
<td>7511</td>
<td>F</td>
<td>Larry</td>
<td>Brown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>0</td>
<td>552</td>
<td>0</td>
<td>1</td>
<td>2004</td>
<td>1</td>
<td>1</td>
<td>2005</td>
<td>1</td>
<td>7533</td>
<td>P</td>
<td>Michael</td>
<td>Delkirk</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>34</td>
<td>0</td>
<td>1172</td>
<td>0</td>
<td>1</td>
<td>2004</td>
<td>1</td>
<td>1</td>
<td>2005</td>
<td>1</td>
<td>7533</td>
<td>P</td>
<td>Dorothy</td>
<td>Harris</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>34</td>
<td>0</td>
<td>322</td>
<td>0</td>
<td>1</td>
<td>2004</td>
<td>1</td>
<td>1</td>
<td>2005</td>
<td>1</td>
<td>7533</td>
<td>P</td>
<td>Gary</td>
<td>Hanes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>40</td>
<td>0</td>
<td>3130</td>
<td>0</td>
<td>1</td>
<td>2004</td>
<td>1</td>
<td>1</td>
<td>2005</td>
<td>1</td>
<td>7533</td>
<td>P</td>
<td>Don</td>
<td>Kline</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>50</td>
<td>1</td>
<td>3145</td>
<td>0</td>
<td>1</td>
<td>2004</td>
<td>1</td>
<td>1</td>
<td>2005</td>
<td>1</td>
<td>7533</td>
<td>P</td>
<td>Robert</td>
<td>Kansas</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>50</td>
<td>89</td>
<td>813</td>
<td>0</td>
<td>1</td>
<td>2004</td>
<td>1</td>
<td>1</td>
<td>2005</td>
<td>1</td>
<td>7533</td>
<td>P</td>
<td>Lolly</td>
<td>Kline</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>50</td>
<td>89</td>
<td>2187</td>
<td>0</td>
<td>1</td>
<td>2004</td>
<td>1</td>
<td>1</td>
<td>2005</td>
<td>1</td>
<td>7533</td>
<td>P</td>
<td>Dan</td>
<td>Larisson</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>50</td>
<td>89</td>
<td>1187</td>
<td>0</td>
<td>1</td>
<td>2004</td>
<td>1</td>
<td>1</td>
<td>2005</td>
<td>1</td>
<td>7533</td>
<td>P</td>
<td>Beth</td>
<td>Leon</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>50</td>
<td>89</td>
<td>1187</td>
<td>0</td>
<td>1</td>
<td>2004</td>
<td>1</td>
<td>1</td>
<td>2005</td>
<td>1</td>
<td>7533</td>
<td>P</td>
<td>Kim</td>
<td>Leon</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>50</td>
<td>89</td>
<td>1187</td>
<td>0</td>
<td>1</td>
<td>2004</td>
<td>1</td>
<td>1</td>
<td>2005</td>
<td>1</td>
<td>7533</td>
<td>P</td>
<td>Mary</td>
<td>Leon</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

*Note: The table continues with more entries.*
HOW MUCH IS A FIREPLACE WORTH?

• A random sample of 1729 houses is now in SaratogaHouses in library(mosaic) in R

Problem: I have a house without a fireplace. My contractor says he can build one for $35,000
START BY LOOKING AT THE DATA

• Difference in means is $65,000
• Contractor can add one for $35,000 — good business decision?
LET’S THINK “STATISTICALLY”

H₀: Means are equal

t = 14.971, df = 1724.7
p-value < 2.2e-16

95 percent confidence interval:
56710.60    73810.61
LET’S THINK RANDOMIZATION BASED

Bootstrap Confidence Interval

diffmeans=do(1000)*diffmean(price~Fireplace,data=resample(SaratogaHouses))
quantile(diffmeans$diffmean,c(0.025,0.975))
hist(diffmeans$diffemean)

95 percent confidence interval:
57113.90.  73642.89
THAT SETTLES IT (!?)

- Courses typically end with A/B tests
- This summer's course?
- How do we get students to think multivariately?
DIFFERENT INTERCEPTS?

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 13599.164| 4991.695   | 2.724   | 0.00651  |
| livingArea     | 111.218  | 2.968      | 37.476  | < 2e-16  |
| FireplaceTRUE  | 5567.377 | 3716.947   | 1.498   | 0.13436  |

![Scatter plot showing the relationship between living area and price with different intercepts for different categories.](image)
WHAT NOW?
WHAT ABOUT BEDROOMS?

Which one costs more?

8 bedrooms

2 bedrooms
# LINEAR MODEL

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|---------|
| (Intercept)    | 33252    | 9630       | 3.453   | 0.000568|
| bedrooms       | 57196    | 3044       | 18.789  | < 2e-16 *** |
AN EASY QUARTER MILLION $ 

If I chop each bedroom into 4, I’ll have an 8 bedroom house worth > $250,000 more !!!
TWO CORRELATED PREDICTORS

Coefficients:

| Predictor    | Estimate | Std. Error | t value | Pr(>|t|) |
|--------------|----------|------------|---------|----------|
| (Intercept)  | 36667.895| 6610.293   | 5.547   | 3.36e-08 *** |
| livingArea   | 125.405  | 3.527      | 35.555  | < 2e-16 *** |
| bedrooms     | -14196.769| 2675.159   | -5.307  | 1.26e-07 *** |
TWO CORRELATED PREDICTORS
KAGGLE MENTALITY

• Data require sophisticated cleaning and manipulation

• Problem is unclear, but often involves predicting a response more closely than other groups.

Content

Environmental Remediation Sites are areas being remediated under one of DEC's remedial programs, including State Superfund and Brownfield Cleanup. This database contains records of the sites which have been remediated or are being managed under by the agency. All sites listed on the "Registry of Inactive Hazardous Waste Disposal Sites in New York State" are included in this database. The Database also includes the "Registry of Institutional and Engineering Controls in New York State". Each site record includes: Administrative information, including site name, classification, unique site code, site location, and site owner(s). Institutional and Engineering Controls implemented at the site. Wastes known or thought to be disposed at the site.

• Did the analysis solve the problem?
DATA QUALITY (!?)

• Dealing with data quality takes about 80% of data scientists’ time (Wilder-James 2016)

• It is the problem most complained about on Kaggle 2017.

• On average 47% of newly created data records have at least on “critical” error

• 82.5% of all statistics are made up.
1. Introduction to Data Science I

(a) Vision

- A complete alpha-to-omega introduction to data science. Students will engage in the full data workflow, including collaborative data science projects. This class is meant to be a high-level introduction to the spectrum of data science topics, probably best taught in an iterative cycle from initial investigation and data acquisition to the communication of final results

(b) Learning goals

- Exploring and wrangling data
- Writing basic functions and coding
- Summarizing, visualizing, and analyzing data
- Modeling and simulating deterministic and stochastic phenomena
- Presenting the results of a complete project in written, oral, and graphical forms
2. Introduction to Data Science II

(a) Vision

- Exposure to different data types and sources, the process of data curation for the purpose of transforming them to a format suitable for analysis. Introduction to the elementary notions in estimation, prediction and inference. We envision this class to be taught through case studies involving less-manicured data to enhance their computational and analytical abilities.

(b) Learning goals

- Interacting with a variety of data sources including relational databases
- Accessing data via different interfaces
- Building structure from a variety of data forms to enable analysis
- Formulating problems and bringing elementary concepts in estimation, prediction, and inference to bear
- Understanding how the data collection process influences the scope of inference
WHAT ABOUT ETHICS?

• There are many important ethical issues when dealing with data and models.
• Some are not as obvious as this:

How 65 Bay St. was deemed part of a needy area
In the final map approved by state officials, 16 census tracts were linked together to connect the affluent Jersey City waterfront to impoverished and crime-ridden neighborhoods nearly four miles away. This allowed the project to qualify for low-interest loans through a U.S. visa program.

Source: Census Bureau
ANDREW TRAN AND GABRIEL FLORIT/THE WASHINGTON POST
WHAT ABOUT ETHICS?

• Red Area (30% unemployed)

• 65 Bay Street
WHAT ABOUT ETHICS?
In forecasting who would re-offend, the algorithm correctly predicted recidivism for black and white defendants at roughly the same rate (59 percent for white defendants, and 63 percent for black defendants) but made mistakes in very different ways.
• Black defendants who not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45 percent vs. 23 percent).

• White defendants who re-offended within the next two years were mistakenly labeled low risk almost twice as often as black re-offenders (48 percent vs. 28 percent).

• Even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 45 percent more likely to be assigned higher risk scores than white defendants.

• The violent recidivism analysis also showed that even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 77 percent more likely to be assigned higher risk scores than white defendants.
PUT IT TOGETHER
CMDA Major Requirements

Standard Track

**Integrated Path**
- CS 1114 - Intro to Software Design
- CS 2114 - Software Design & Data Structures
- CMDA 2005 - Integrated Quantitative Science I
- MATH 2114 - Intro to Linear Algebra
- CMDA 2006 - Integrated Quantitative Science II
- CMDA 3634 - CS Foundations of CMDA
- CMDA 3654 - Intro Data Analytics & Visualization
- CMDA 4864 - CMDA Capstone Project

**Traditional Path**
- MATH 1225 - Calculus of a Single Variable I
- MATH 1226 - Calculus of a Single Variable II
- MATH 2114 - Intro to Linear Algebra
- STAT 3006 - Stat Methods II
- STAT 3104 - Probability and Distributions
- STAT 3005 - Stat Methods I

**Weak Prerequisite**
- MATH 2214 - Intro to Differential Equations
- MATH 2204 - Intro to Multivariable Calc

**Electives**
- +12 credits of CMDA electives

CMDA 4864 (CMDA Capstone) requires as a prerequisite either (CS 2114-CMDA 3634) or (CMDA 3605-3606) or (CMDA 3654-4654).
THE PATH FROM HERE?

Goal: Create a major/curriculum for data science using existing resources (faculty, courses, etc).

1. Math 150 - Multivariable Calculus
2. Math 200 - Discrete Math
3. Math 250 - Linear Algebra
4. Stat 161 or 201 - Introductory Statistics
5. Stat 202 - Intro to Statistical Modeling
6. CS 134 - Intro to Computer Science
7. CS 136 - Data Structures
8. CS 256 - Algorithms (or maybe 237 - Computer Organization)
10. Pre-approved domain elective (in something other than CS/Math/Stat?)
THE PATH?
SUMMARY

• Where are we in Data Science?
• Make the Intro Stats course more relevant to Data Science
  • Don’t give up control to other disciplines
• Evolution from Existing Statistics Courses to Data Science Curriculum
• Put the Data back into Data Science !!
WITH APOLOGIES TO DAVID, HOFFMAN, AND LIVINGSTON

Data science, Data science
Night and day it’s Data science
Some say it’s just statistics
Some say that it’s comp science
But what we’re really scared of
Is losing all our clients
So maybe we should join them
And just form some grand alliance
Data science data science
Data science
THANK YOU!

Data science, data science
Night and day it’s data science
Now think about the Russians
And imagine our reliance
As we’re putting neural networks
Into every damned appliance
To imagine this disaster
Doesn’t take much rocket science
Data science data science
Night and day it’s data science
Data science