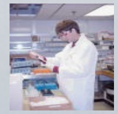





DATA SCIENCE FOR ALL
SURE, BUT WHO, WHERE, WHEN AND HOW MUCH? OR…
LET'S PUT THE DATA BACK INTO DATA SCIENCE

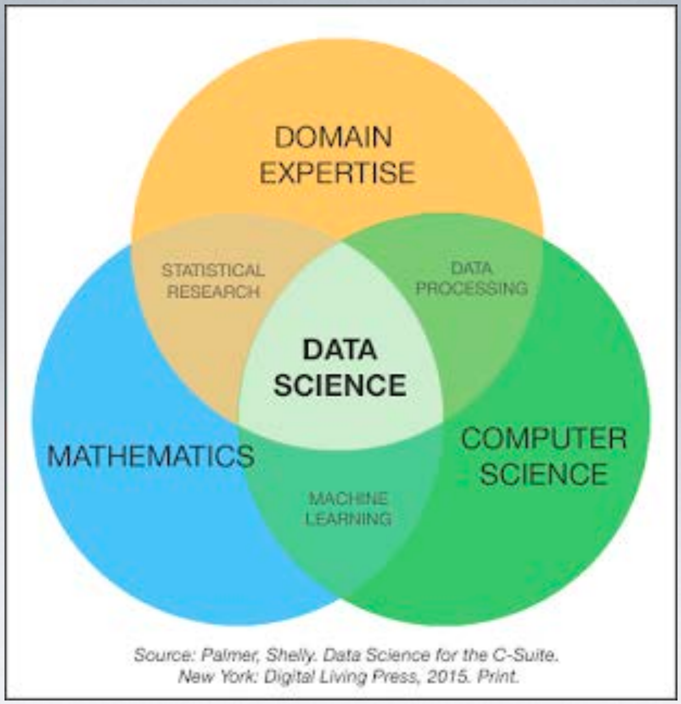
Richard D. De Veaux
Williams College
IASE Satellite ISI
August, 2019
deveaux@williams.edu

HERE'S WHAT KEEPS ME UP AT NIGHT

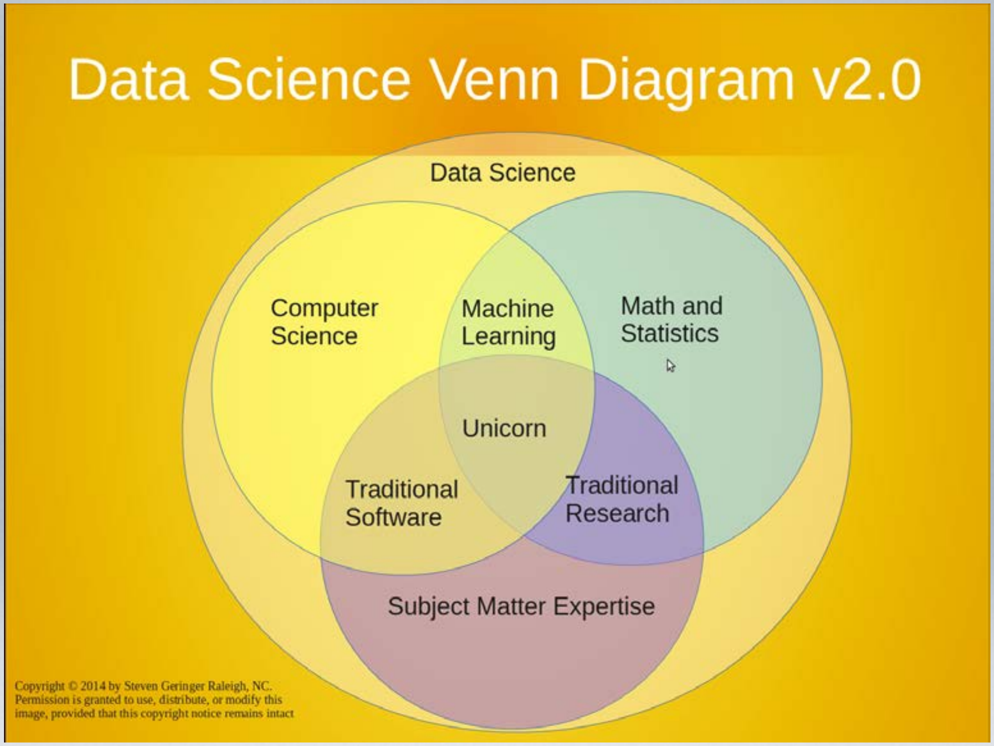
- Data Science courses — with no "data"
- Our Intro Stats course becoming even less relevant to students' needs
- Students thinking that the world (or at least the Statistics world) is univariate
- That we are teaching the same course we taught in 1958 — or even 1996
- That we have replaced Math envy 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

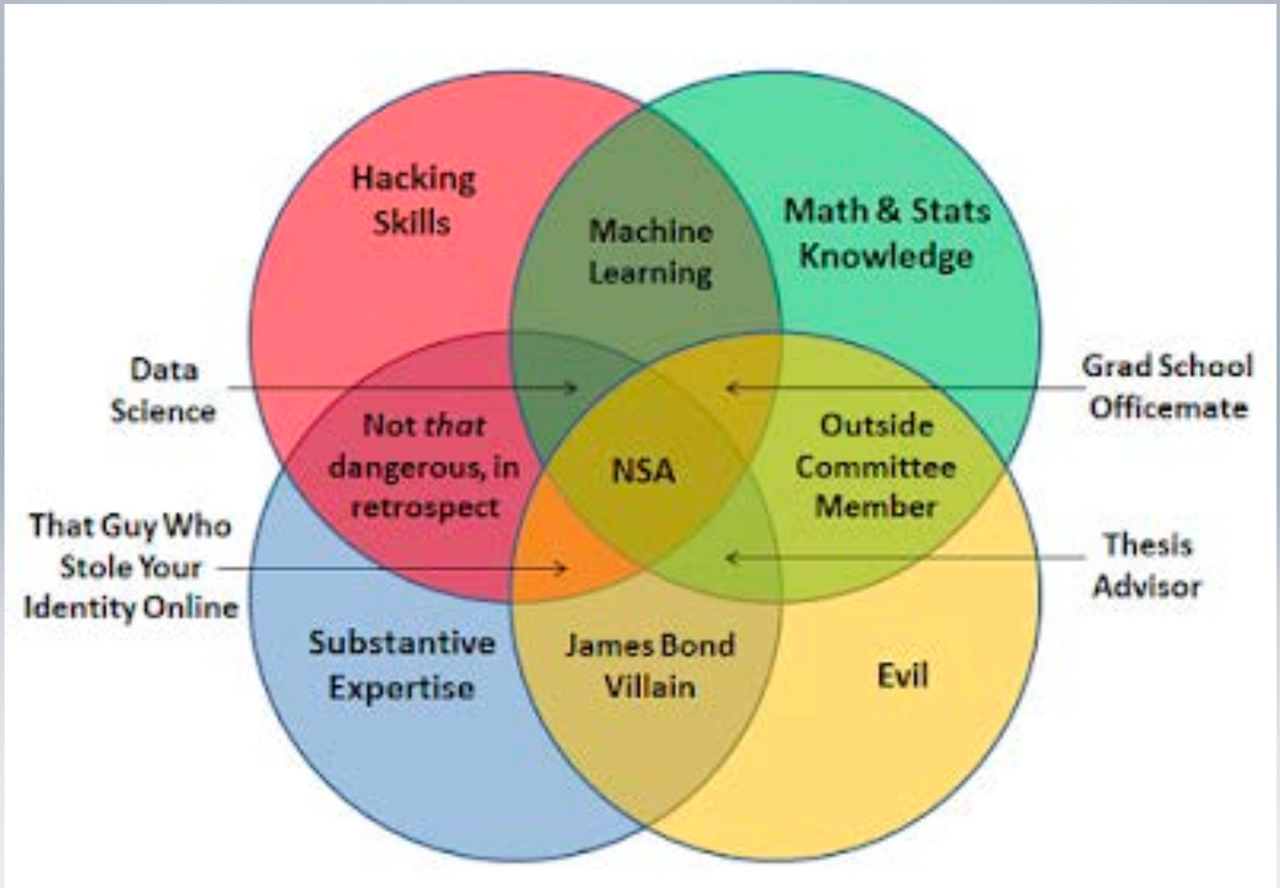


WHAT IS DATA SCIENCE II?



Data science is a method for gleaning insights from structured and unstructured data using approaches ranging from statistical analysis to machine learning.

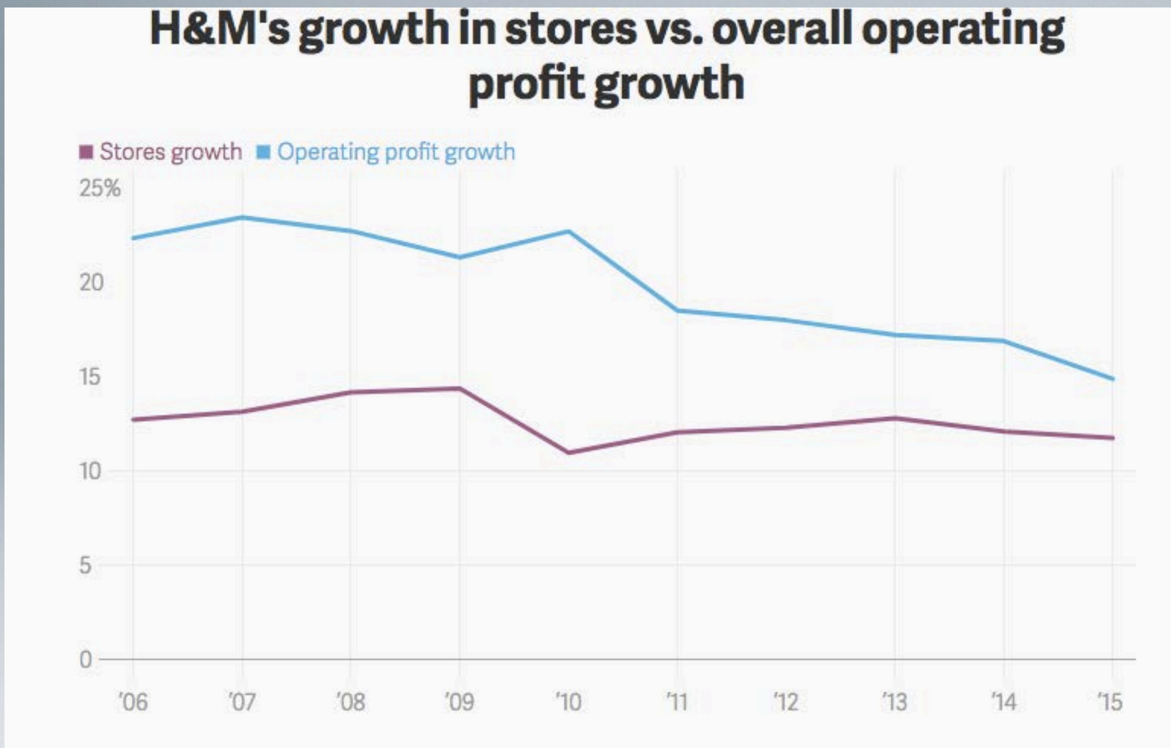
WHAT IS DATA SCIENCE III?



OUR STUDENTS?

Thomasine lands her dream job
— analyst for H&M

FIRST PROJECT



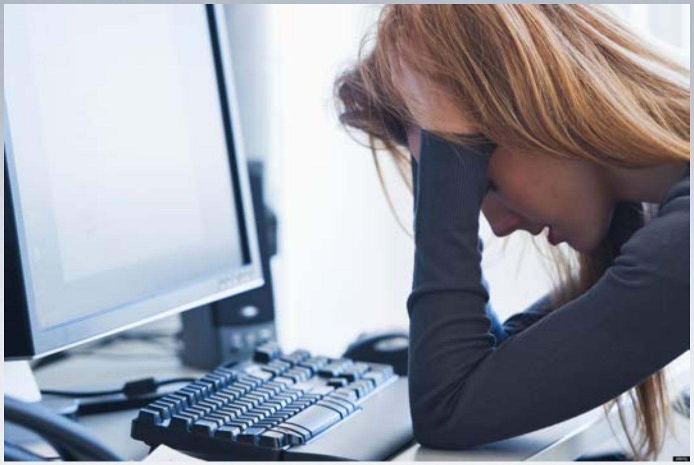
Q: How much of their resources should they put online vs. brick and mortar?

DOWNLOADS THE DATA…

155000	0.41	0	13	18700	0	0	1944	3	1	1.5	7	Yes
86060	0.11	0	0	15000	1	1	840	2	0	1	5	No
120000	0.68	0	31	14000	0	0	1152	4	1	1	7	Yes
153000	0.4	0	33	23300	0	0	2752	4	1	1.5	9	Yes
170000	1.21	0	23	14600	0	0	1662	4	1	1.5	7	Yes
90000	0.83	0	36	22200	0	0	1632	3	0	1.5	6	No
122900	1.94	0	4	21200	0	0	1416	3	0	1.5	6	No
325000	2.29	0	123	12600	0	0	2894	7	0	1	12	No
120000	0.92	0	1	22300	0	0	1624	3	0	2	6	No
85860	8.97	0	13	4800	0	0	704	2	0	1	5	No
97000	0.11	0	153	3100	0	0	1383	3	0	2	6	No
127000	0.14	0	9	300	0	0	1300	3	0	1.5	6	No
89900	0	0	88	2500	0	0	936	3	0	1	6	No
155000	0.13	0	9	300	0	0	1300	3	0	1.5	6	No
253750	2	0	0	49800	0	1	2816	4	1	2.5	9	Yes
60000	0.21	0	82	8500	0	0	924	2	0	1	5	No
87500	0.88	0	17	19400	0	0	1092	3	0	1	6	No
112000	1	0	12	8600	0	0	1056	3	0	1	6	No
104900	0.43	0	21	5600	0	0	1600	3	0	1.5	6	No
148635	0.32	0	1	6200	1	1	1576	3	0	2.5	6	No
150000	0.03	0	24	5100	0	0	2080	3	0	2	7	No
90400	0.36	0	16	5200	0	0	1600	3	0	1.5	6	No
248800	4	0	28	5500	0	0	2224	4	0	3	8	No
135000	1.83	0	126	6000	0	0	1656	3	0	1	6	No
145000	3	0	26	4500	0	0	1170	4	0	1.5	7	No
457000	0.43	1	53	2700	0	0	2461	4	1	2	8	Yes
140000	0.44	0	56	19400	0	1	1544	3	1	1.5	6	Yes
130000	1.24	0	51	24800	0	0	1220	2	2	1	5	Yes
187000	0.46	0	3	15200	0	0	1858	3	1	2.5	7	Yes
229000	0.87	0	9	41100	0	0	2219	3	1	2	7	Yes
227000	1.8	0	201	25500	0	0	1876	3	0	2.5	7	No
179900	0.46	0	1	15200	0	0	2026	4	1	2.5	8	Yes
169900	0.91	0	19	20200	0	1	1671	4	1	3	7	Yes
209900	0.46	0	1	15200	0	0	2060	4	1	2.5	8	Yes
169900	0.59	0	0	17300	0	0	1884	4	1	2.5	8	Yes
293000	7.24	0	43	36600	0	0	2022	4	2	3	8	Yes
245900	0.19	0	0	20700	0	1	2394	4	1	2.5	8	Yes
157000	0.46	0	45	20200	0	0	1390	3	1	1.5	6	Yes
195000	0.41	0	32	27100	0	1	1954	4	0	2.5	8	No
150000	0.78	0	54	24500	0	1	1554	3	1	1.5	6	Yes
234900	0.89	0	9	41600	0	1	1976	3	0	2.5	7	No
279550	1.34	0	0	44400	0	0	2479	4	1	2.5	8	Yes
246500	1	0	0	17100	0	0	2714	4				

```
Municipality  Location  Route  OwnerBuilt  Date.Inspected  SD.FO.Status  Condition  YearInspected  AgeAtInspection
Caroline Town  3.6 MI NW TIOGA CL;RTE 79  79  79 36051035  NYSDOT  1963  10/14/15  N  5.14  2015.7836  52.783562
Caroline Town  .3 MI E JCT SH 79 & CR162  79  79 36051041  NYSDOT  1963  8/19/15  N  6.053  2015.6301  52.630137
Caroline Town  .5 MILE EAST OF BESEMER  BANKS ROAD  County  2008  11/20/14  N  6.172  2014.8849  6.8849315
Caroline Town  .9 MI W SLATERVILLE SPNGS  BOICEVILLE ROAD  County  1942  7/9/15 N  4.604  2015.5178  73.517808
Caroline Town  IN SLATERVILLE SPRINGS  BUFFALO ROAD  County  1993  8/13/15  N  5.795  2015.6137  22.613699
Caroline Town  2 MI N OF SPEEDSVILLE Blackman Hill Rd.  County  1994  11/20/14  N  6.305  2014.8849  20.884932
Caroline Town  4.9 MI SE JCT RTS. 330&79  CENTRAL CHAPEL RD  County  1987  5/14/15  N  4.386  2015.3644  28.364384
Caroline Town  AT GUIDE BOARD CORNERS  CENTRAL CHAPEL RD  County  1966  4/2/15 N  5.583  2015.2493  49.249315
Caroline Town  1 MI SE OF W.SLATERVILLE  CENTRAL CHAPEL RD  County  1966  4/24/15  N  5.614  2015.3096  49.309589
Caroline Town  AT BROOKTONDALE  COOKS CORS-BRK RD  County  1966  6/17/14  N  4.732  2014.4575  48.457534
Caroline Town  1.6 MI SOUTH OF BESEMER  CR113LOUNSBERRYRD  County  2003  6/17/14  N  6.567  2014.4575  11.457534
Caroline Town  .4 MI W SLATERVILLE SPGS.  CREAMERY ROAD County  1977  5/21/15  N  4.644  2015.3836  38.383562
Caroline Town  2.8 MI W SLATERVLLE SPNGS  HARFORD ROAD  County  1977  10/29/14  N  6.314  2014.8247  37.824658
Caroline Town  1 MI SOUTH OF BESEMER  MIDDAUGH ROAD County  1978  4/14/15  N  5.102  2015.2822  37.282192
Caroline Town  .3 MILE S OF SPEEDSVILLE  OLD SEVNTY SIX RD  County  2009  7/16/15  N  6.815  2015.5370  6.5369863
Caroline Town  1.5 MI NW OF SPEEDSVILLE  OLD SEVNTY SIX RD  County  1987  7/16/15  N  4.684  2015.5370  28.536986
Caroline Town  IN SPEEDSVILLE  OLD SEVNTY SIX RD  County  2001  8/6/15 N  6.61  2015.5945  14.594521
Caroline Town  .5 MI S OF WEST SLATERVLE  VALLEY ROAD  County  1966  5/14/15  N  5.591  2015.3644  49.364384
Danby Town  5.6 MI NW TIOGA CL-SH 96B  96B 96B36021057  NYSDOT  1929  11/17/15  N  4.217  2015.8767  86.876712
Danby Town  3.3 mi NW Willseville  96B 3602 1010  NYSDOT  1960  11/3/15  N  5.211  2015.8384  55.838356
Danby Town  1.3 MI NORTH OF W DANBY  BROWN ROAD  Town  1943  11/12/14  N  6.179  2014.8630  71.863014
Danby Town  1.8 MI S BUTTERMILK FALLS  COMFORT ROAD  County  1998  5/7/14 N  5.805  2014.3452  16.345205
Danby Town  3.8 MILES NE OF NEWFIELD  JERSEY HILL ROAD  County  2010  11/12/14  N  6.857  2014.8630  4.8630137
Dryden Town .6 MI NW JCT SH 13 & SH 3  13  13 36033057  NYSDOT  2013  10/27/15  N  6.857  2015.8192  2.8191781
Dryden Town 1.6 MI NE JCT RTS 366 +13  366 366 36011066  NYSDOT  1932  9/18/15  FO  4.547  2015.7123  83.712329
Dryden Town 2.5 MI NE JCT SH 366 & SH  366 366 36011075  NYSDOT  1932  7/15/14  SD  4.516  2014.5342  82.534247
Dryden Town IN ETNA  COUNTY ROAD 109  County  1975  10/22/15  N  4  2015.8055  40.805479
Dryden Town IN ETNA  COUNTY ROAD 109  County  1960  11/17/15  N  4.339  2015.8767  55.876712
Dryden Town 1 MI EAST OF ITHACA  DODGE ROAD  County  1935  5/4/15 SD  3.604  2015.3370  80.336986
Dryden Town 3 MI SE DRYDEN-E LAKE RD  EAST LAKE ROAD County  1999  10/29/15  N  6.436  2015.8247  16.824658
Dryden Town .7 MI SW OF MCLEAN  FALL CREEK ROAD  County  1965  6/12/15  N  5.535  2015.4438  50.443836
Dryden Town 1 MI NE OF FREEVILLE  FALL CREEK ROAD  County  1965  5/23/14  N  4.864  2014.3890  49.389041
Dryden Town AT VARNA  FREESE ROAD  County  1920  9/22/15  SD  3.586  2015.7233  95.723288
Dryden Town 1.3 MI E ITHACA CITY LMTS  GAME FARM ROAD  County  1940  8/11/15  FO  4.426  2015.6082  75.608219
Dryden Town 1.8 MILES SE OF VARNA GENUNG ROAD  County  1940  5/26/15  N  4.719  2015.3973  75.397260
Dryden Town 2.7 MI SE ITHACA CITY LMT  GERMAN CROSS ROAD  County  1983  9/22/15  N  5.567  2015.7233  32.723288
Dryden Town 0.7 MI W OF FREEVILLE  MILL STREET  County  1910  9/3/15 FO  4.345  2015.6712  105.67123
Dryden Town 1.4 MI W JCT SH366 &SH355  PINCKNEY ROAD  County  1990  5/29/14  N  5.88  2014.4055  24.405479
Dryden Town 3.3 MI SE OF VARNA  RINGWOOD ROAD County  2007  4/28/14  N  6.393  2014.3205  7.3205479
Dryden Town 3.5 MI W OF DRYDEN  RINGWOOD ROAD County  1988  5/21/15  N  4.789  2015.3836  27.383562
```

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?

I CAN PROGRAM…

YOU HAD 0'S?

WHEN I STARTED PROGRAMMING, WE DIDN'T HAVE ANY OF THESE SISSY "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."
S. Adams © 1992 United Feature Syndicate, Inc.
9-8

R VS PYTHON VS JMP (TABLEAU ETC)

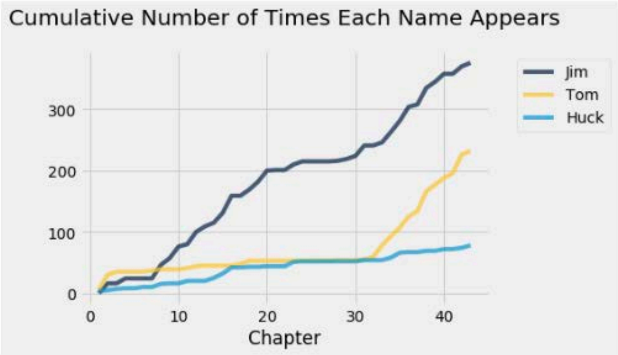
- Each has its advantages.
- Eventually, a data science student should see all of these
- The beginning student?

  Teach the power of Statistics not the mechanics
- Which to start with?

  Data 8 course

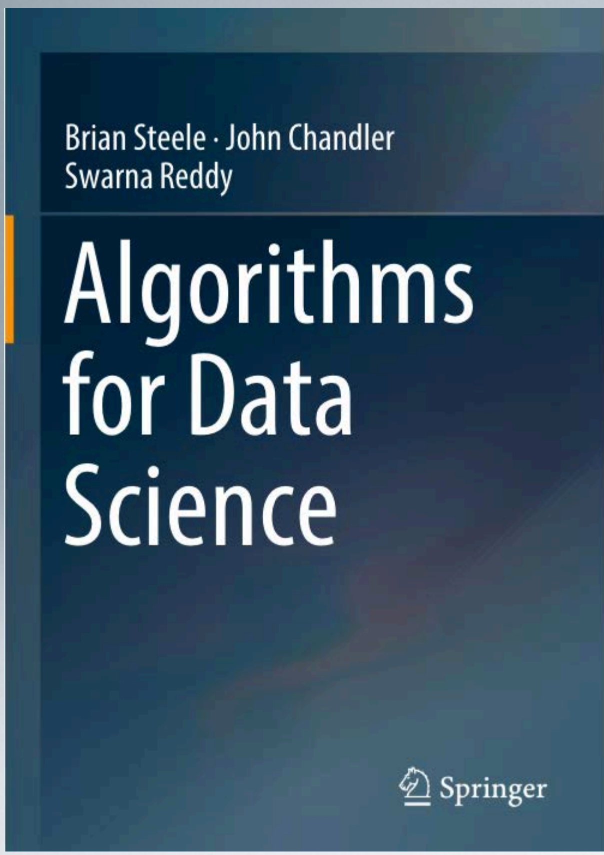
WHAT NOT TO TEACH II?

```
# Count how many times the names Jim, Tom, and Huck appear in each chapter.

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.

cum_counts = counts.cumsum().with_column('Chapter', np.arange(1, 44, 1))
cum_counts.plot(column_for_xticks=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

```
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
```

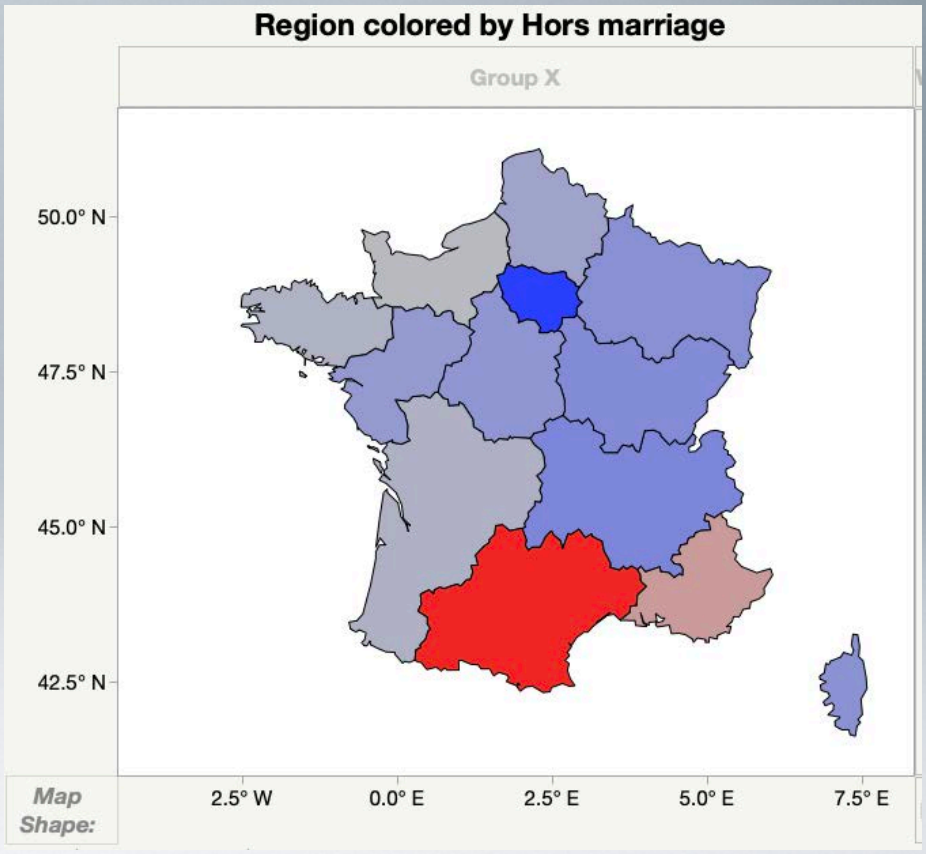
http://www.ams.org/journals/bull/2019-56-01/S0273-0979-2017-01596-0/

EMPOWER STUDENTS

CODAP NHANES

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

ISLE — Carnegie Mellon

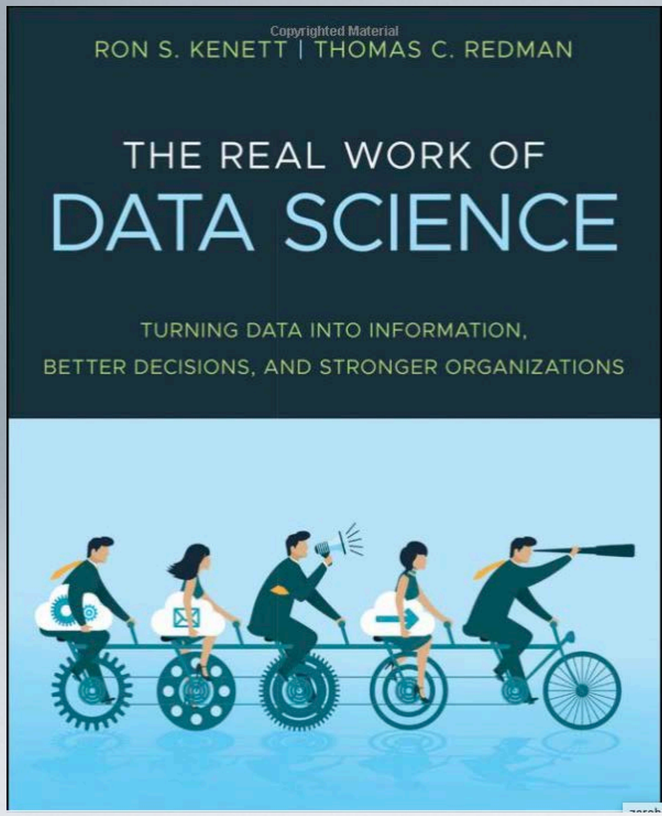


WHAT ISN'T DATA SCIENCE?

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

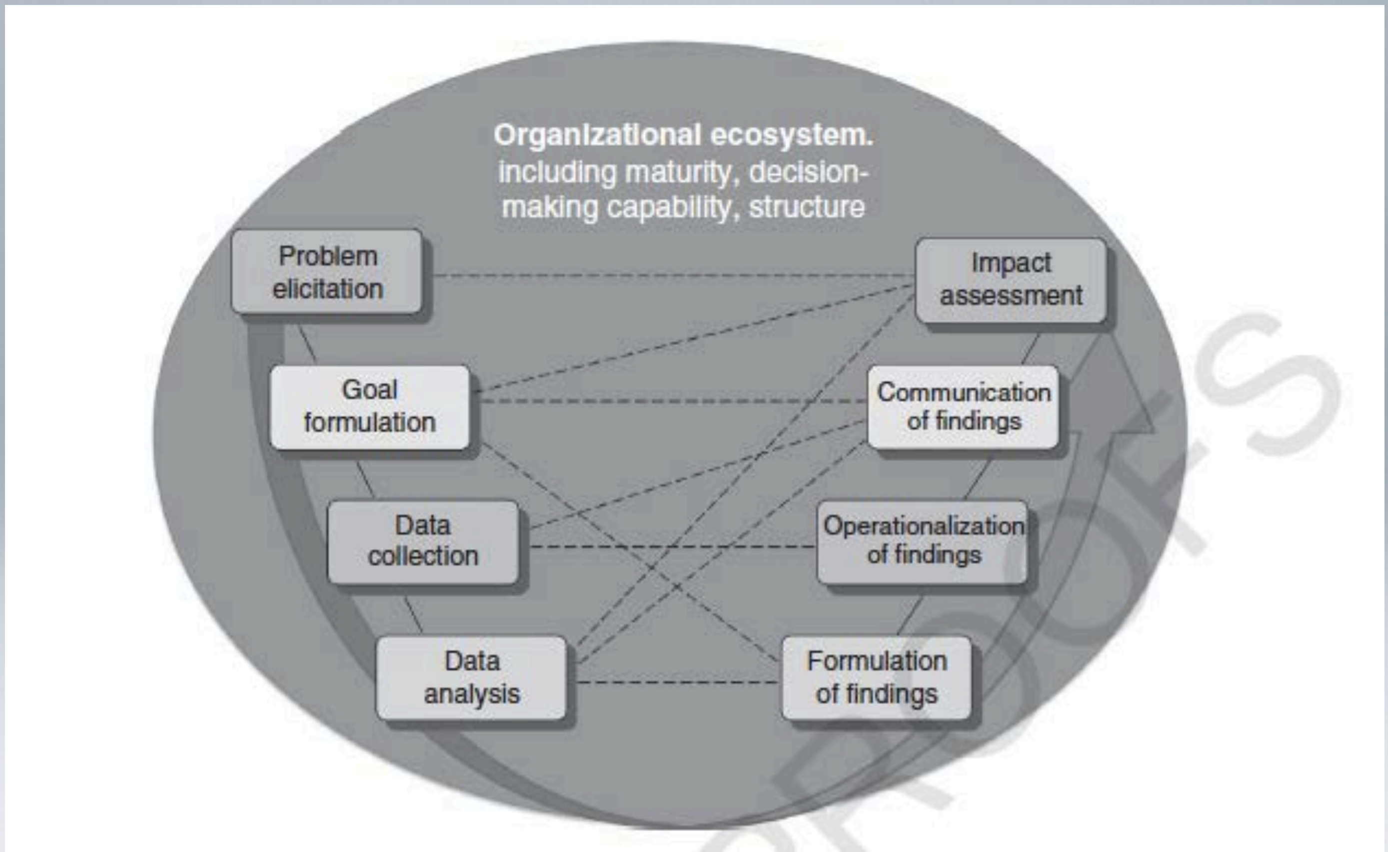
"Nowadays anyone with a laptop and a script can scrape data off the Internet, feed it into an R package, and publish the results. Obviously this isn't data science, but the average citizen isn't going to know the difference."

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*

Richard D. De Veaux,[1] Mahesh Agarwal,[2] Maia Averett,[3] Benjamin S. Baumer,[4] Andrew Bray,[5] Thomas C. Bressoud,[6] Lance Bryant,[7] Lei Z. Cheng,[8] Amanda Francis,[9] Robert Gould,[10] Albert Y. Kim,[11] Matt Kretchmar,[12] Qin Lu,[13] Ann Moskol,[14] Deborah Nolan,[15] Roberto Pelayo,[16] Sean Raleigh,[17] Ricky J. Sethi,[18] Mutiara Sondjaja,[19] Neelesh Tiruviluamala,[20] Paul X. Uhlig,[21] Talitha M. Washington,[22] Curtis L. Wesley,[23] David White,[24] and Ping Ye[25]

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

REBUTTAL? FROM ACM

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 Acquirement 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.

CS 136

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

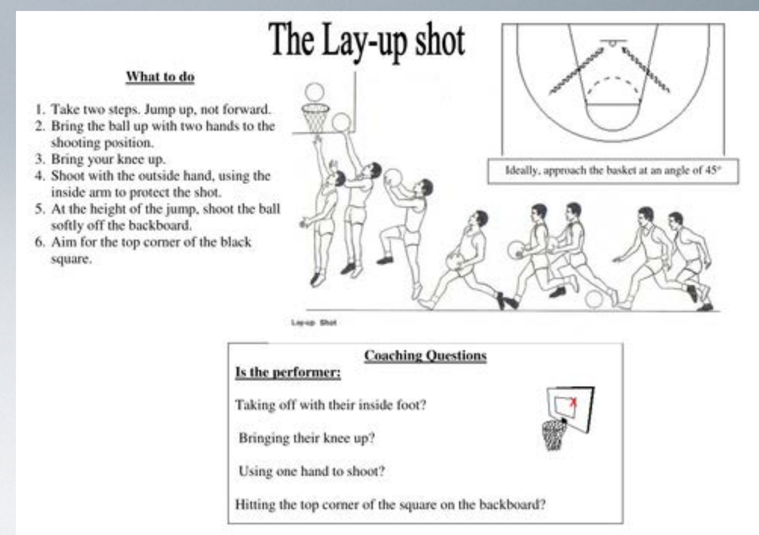
PRODUCERS OR CONSUMERS?

How to teach a lay up?

Who's the audience?

- spectators
- referees
- players
  - beginners
  - pros

Roxy Peck



The Lay-up shot

What to do

1. Take two steps. Jump up, not forward.
2. Bring the ball up with two hands to the shooting position.
3. Bring your knee up.
4. Shoot with the outside hand, using the inside arm to protect the shot.
5. At the height of the jump, shoot the ball softly off the backboard.
6. Aim for the top corner of the black square.

Ideally, approach the basket at an angle of 45°

Lay-up Shot

Coaching Questions

Is the performer:

Taking off with their inside foot?

Bringing their knee up?

Using one hand to shoot?

Hitting the top corner of the square on the backboard?

A CAUTIONARY TALE

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

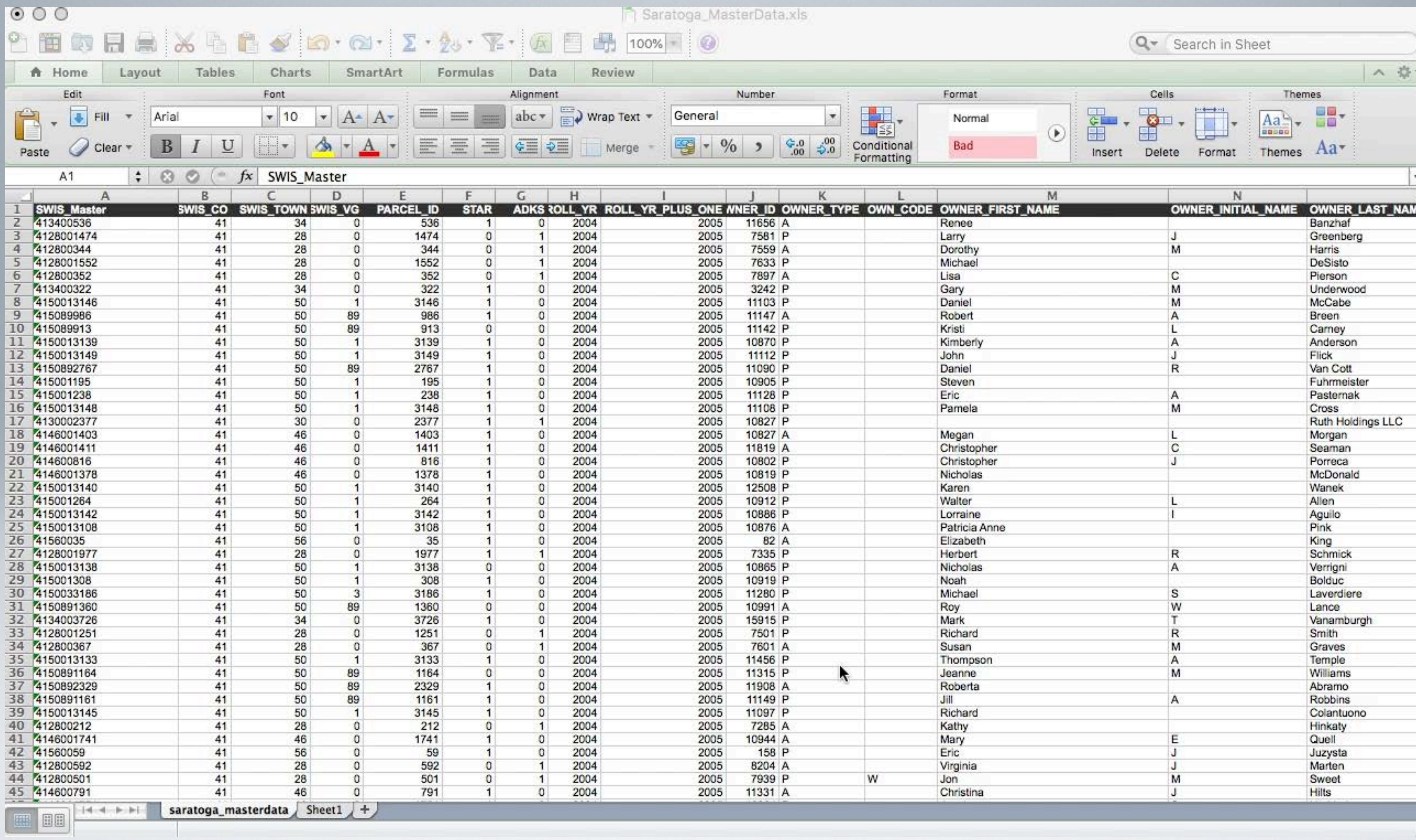
Candice M. Corvetti

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



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	SWIS_Master	SWIS_CO	SWIS_TOWN	SWIS_VG	PARCEL_ID	STAR	ADKS	ROLL_YR	ROLL_YR_PLUS_ONE	WNER_ID	OWNER_TYPE	OWN_CODE	OWNER_FIRST_NAME	OWNER_INITIAL_NAME	OWNER_LAST_NAME
2	413400536	41	34	0	536	1	0	2004	2005	11656	A		Renee		Banzhaf
3	4128001474	41	28	0	1474	0	1	2004	2005	7581	P		Larry	J	Greenberg
4	412800344	41	28	0	344	0	1	2004	2005	7559	A		Dorothy	M	Harris
5	4128001552	41	28	0	1552	0	1	2004	2005	7633	P		Michael		DeSisto
6	412800352	41	28	0	352	0	1	2004	2005	7897	A		Lisa	C	Pierson
7	413400322	41	34	0	322	1	0	2004	2005	3242	P		Gary	M	Underwood
8	4150013146	41	50	1	3146	1	0	2004	2005	11103	P		Daniel	M	McCabe
9	415089986	41	50	89	986	1	0	2004	2005	11147	A		Robert	A	Breen
10	415089913	41	50	89	913	0	0	2004	2005	11142	P		Kristi	L	Carney
11	4150013139	41	50	1	3139	1	0	2004	2005	10870	P		Kimberly	A	Anderson
12	4150013149	41	50	1	3149	1	0	2004	2005	11112	P		John	J	Flick
13	4150892767	41	50	89	2767	1	0	2004	2005	11090	P		Daniel	R	Van Cott
14	415001195	41	50	1	195	1	0	2004	2005	10905	P		Steven		Fuhrmeister
15	415001238	41	50	1	238	1	0	2004	2005	11128	P		Eric	A	Pasternak
16	4150013148	41	50	1	3148	1	0	2004	2005	11108	P		Pamela	M	Cross
17	4130002377	41	30	0	2377	1	1	2004	2005	10827	P				Ruth Holdings LLC
18	4146001403	41	46	0	1403	1	0	2004	2005	10827	A		Megan	L	Morgan
19	4146001411	41	46	0	1411	1	0	2004	2005	11819	A		Christopher	C	Seaman
20	414600816	41	46	0	816	1	0	2004	2005	10802	P		Christopher	J	Porreca
21	4146001378	41	46	0	1378	1	0	2004	2005	10819	P		Nicholas		McDonald
22	4150013140	41	50	1	3140	1	0	2004	2005	12508	P		Karen		Wanek
23	415001264	41	50	1	264	1	0	2004	2005	10912	P		Walter	L	Allen
24	4150013142	41	50	1	3142	1	0	2004	2005	10886	P		Lorraine	I	Aguilo
25	4150013108	41	50	1	3108	1	0	2004	2005	10876	A		Patricia Anne		Pink
26	41560035	41	56	0	35	1	0	2004	2005	82	A		Elizabeth		King
27	4128001977	41	28	0	1977	1	1	2004	2005	7335	P		Herbert	R	Schmick
28	4150013138	41	50	1	3138	0	0	2004	2005	10865	P		Nicholas	A	Verrigni
29	415001308	41	50	1	308	1	0	2004	2005	10919	P		Noah		Bolduc
30	4150033186	41	50	3	3186	1	0	2004	2005	11280	P		Michael	S	Laverdiere
31	4150891360	41	50	89	1360	0	0	2004	2005	10991	A		Roy	W	Lance
32	4134003726	41	34	0	3726	1	0	2004	2005	15915	P		Mark	T	Vanamburgh
33	4128001251	41	28	0	1251	0	1	2004	2005	7501	P		Richard	R	Smith
34	412800367	41	28	0	367	0	1	2004	2005	7601	A		Susan	M	Graves
35	4150013133	41	50	1	3133	1	0	2004	2005	11456	P		Thompson	A	Temple
36	4150891164	41	50	89	1164	0	0	2004	2005	11315	P		Jeanne	M	Williams
37	4150892329	41	50	89	2329	1	0	2004	2005	11908	A		Roberta		Abramo
38	4150891161	41	50	89	1161	1	0	2004	2005	11149	P		Jill	A	Robbins
39	4150013145	41	50	1	3145	1	0	2004	2005	11097	P		Richard		Colantuono
40	412800212	41	28	0	212	0	1	2004	2005	7285	A		Kathy		Hinkaty
41	4146001741	41	46	0	1741	1	0	2004	2005	10944	A		Mary	E	Quell
42	41560059	41	56	0	59	1	0	2004	2005	158	P		Eric	J	Juzysta
43	412800592	41	28	0	592	0	1	2004	2005	8204	A		Virginia	J	Marten
44	412800501	41	28	0	501	0	1	2004	2005	7939	P	W	Jon	M	Sweet
45	414600791	41	46	0	791	1	0	2004	2005	11331	A		Christina	J	Hilts

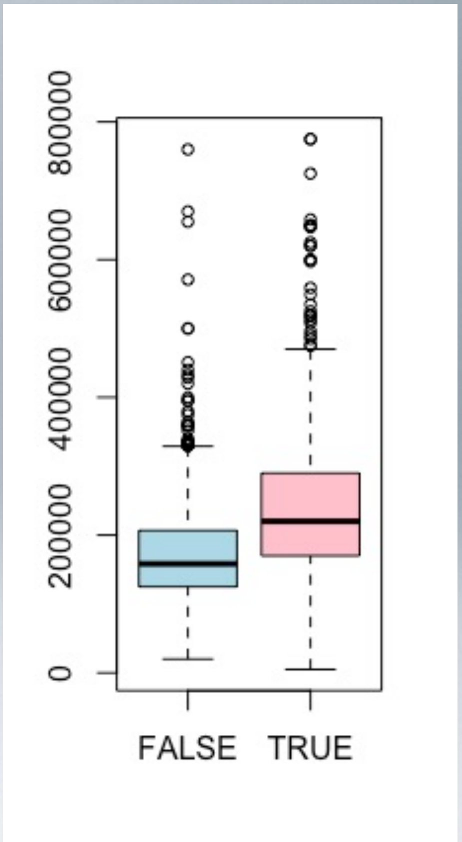
HOW MUCH IS A FIREPLACE WORTH?

- 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"

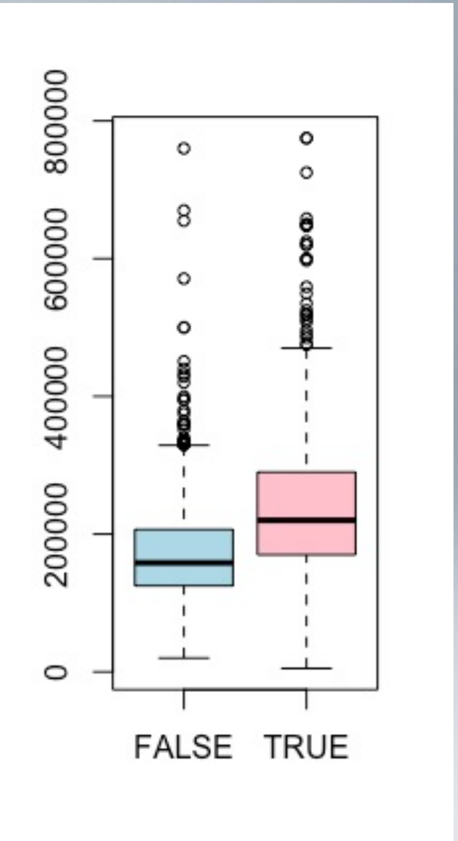
H_0: Means are equal

t = 14.971, df = 1724.7

p-value < 2.2e-16

95 percent confidence interval:

56710.60 73810.61



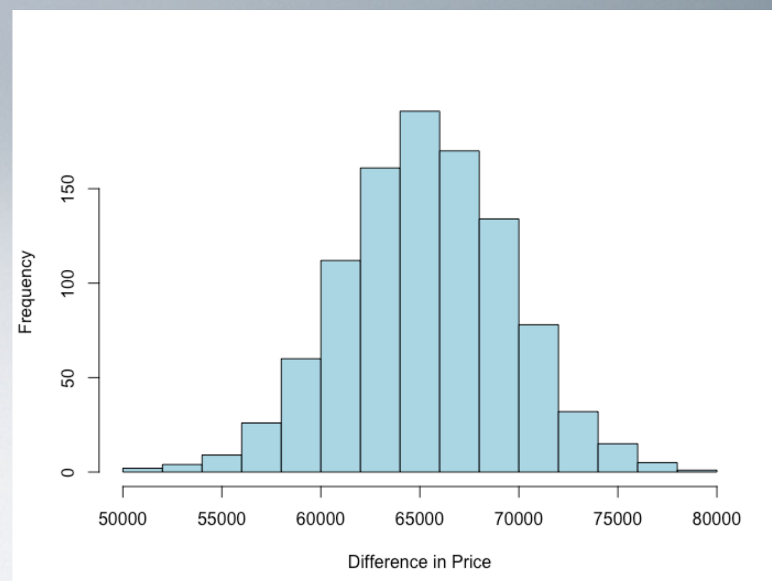
LET'S THINK RANDOMIZATION BASED

Bootstrap Confidence Interval

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

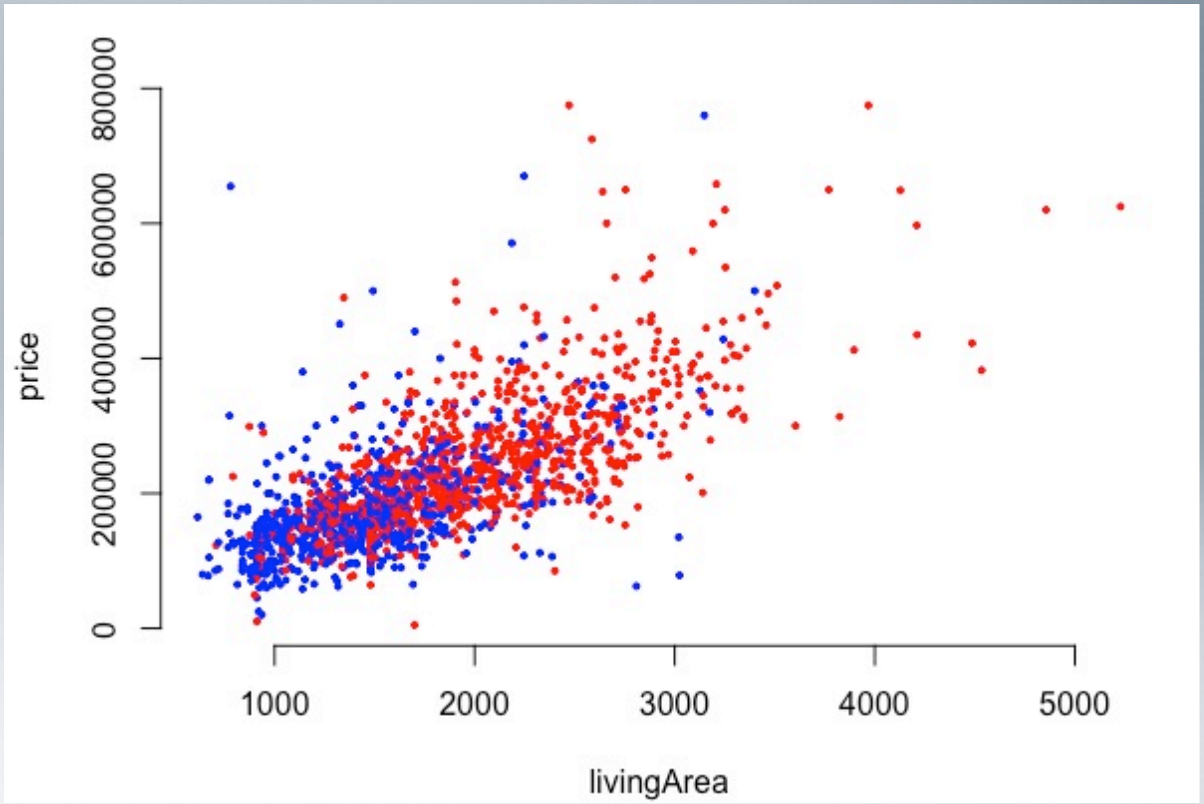
95 percent confidence interval:

57113.90.   73642.89

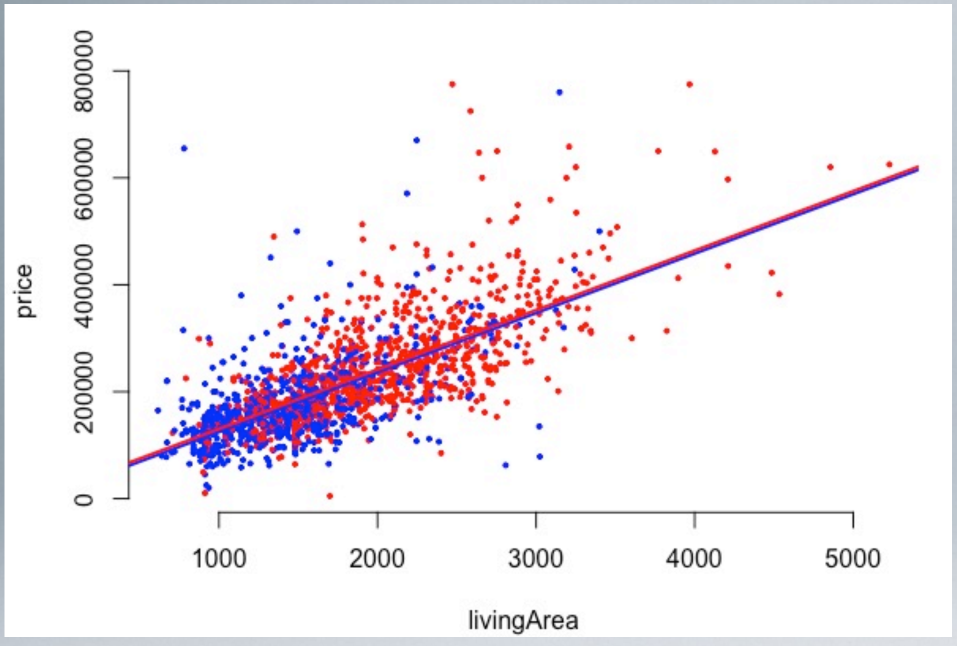


THAT SETTLES IT (!?)

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

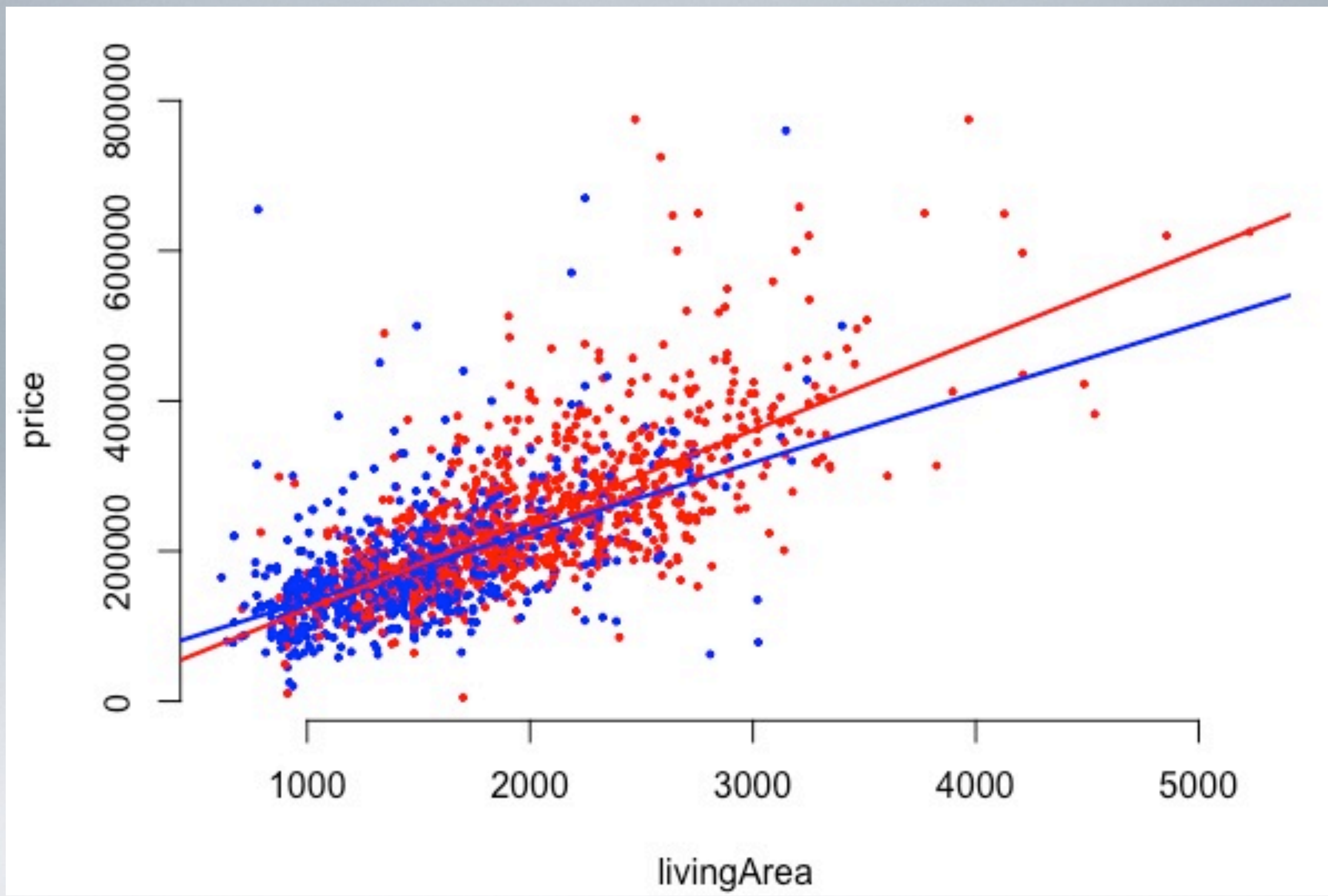


DIFFERENT INTERCEPTS?




```
Coefficients:
               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
```

WHAT NOW?



price
800000
600000
400000
200000
0
1000
2000
3000
4000
5000
livingArea

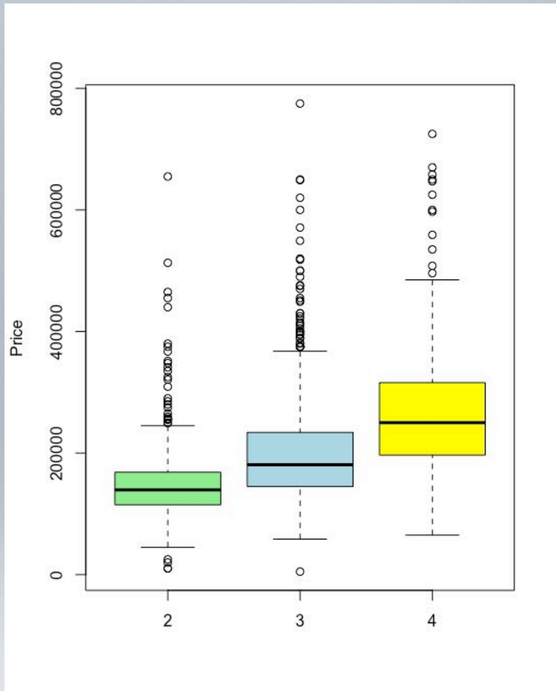
WHAT ABOUT BEDROOMS?

Which one costs more?

8 bedrooms

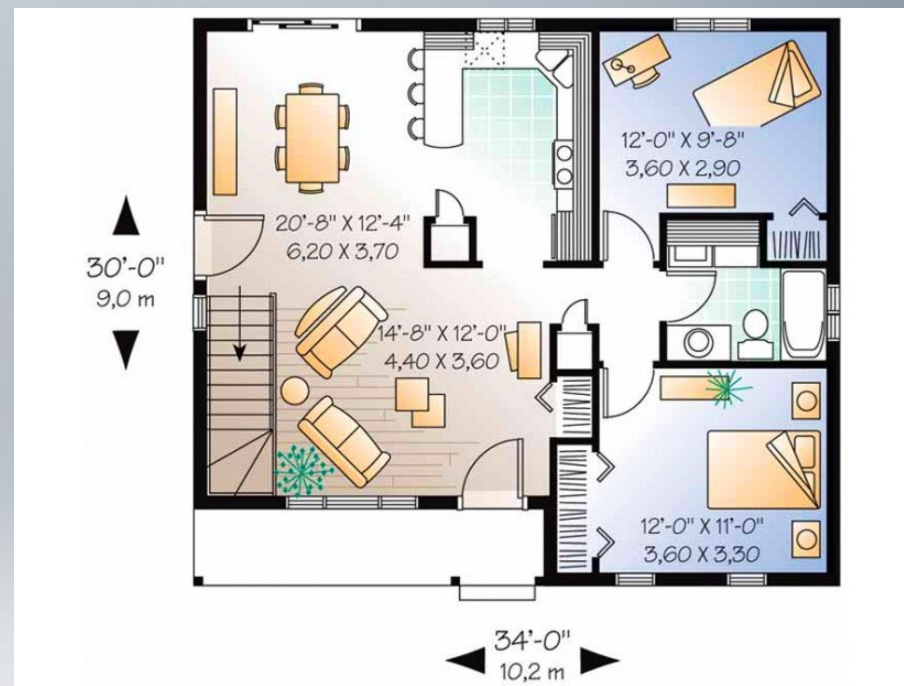
2 bedrooms

LINEAR MODEL




```
Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)    33252       9630   3.453  0.000568
bedrooms       57196       3044  18.789  < 2e-16 ***
```

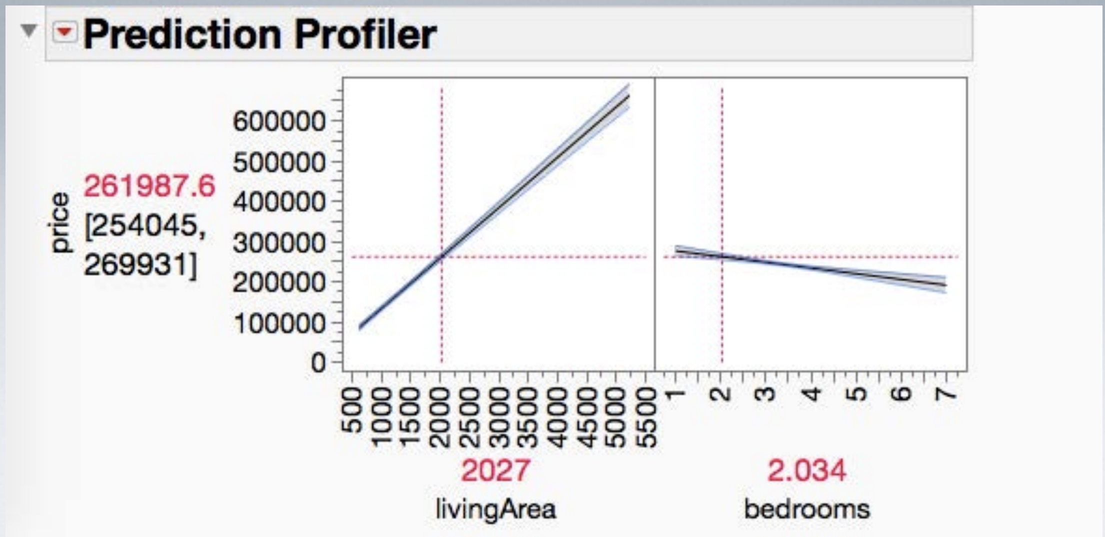
AN EASY QUARTER MILLION $



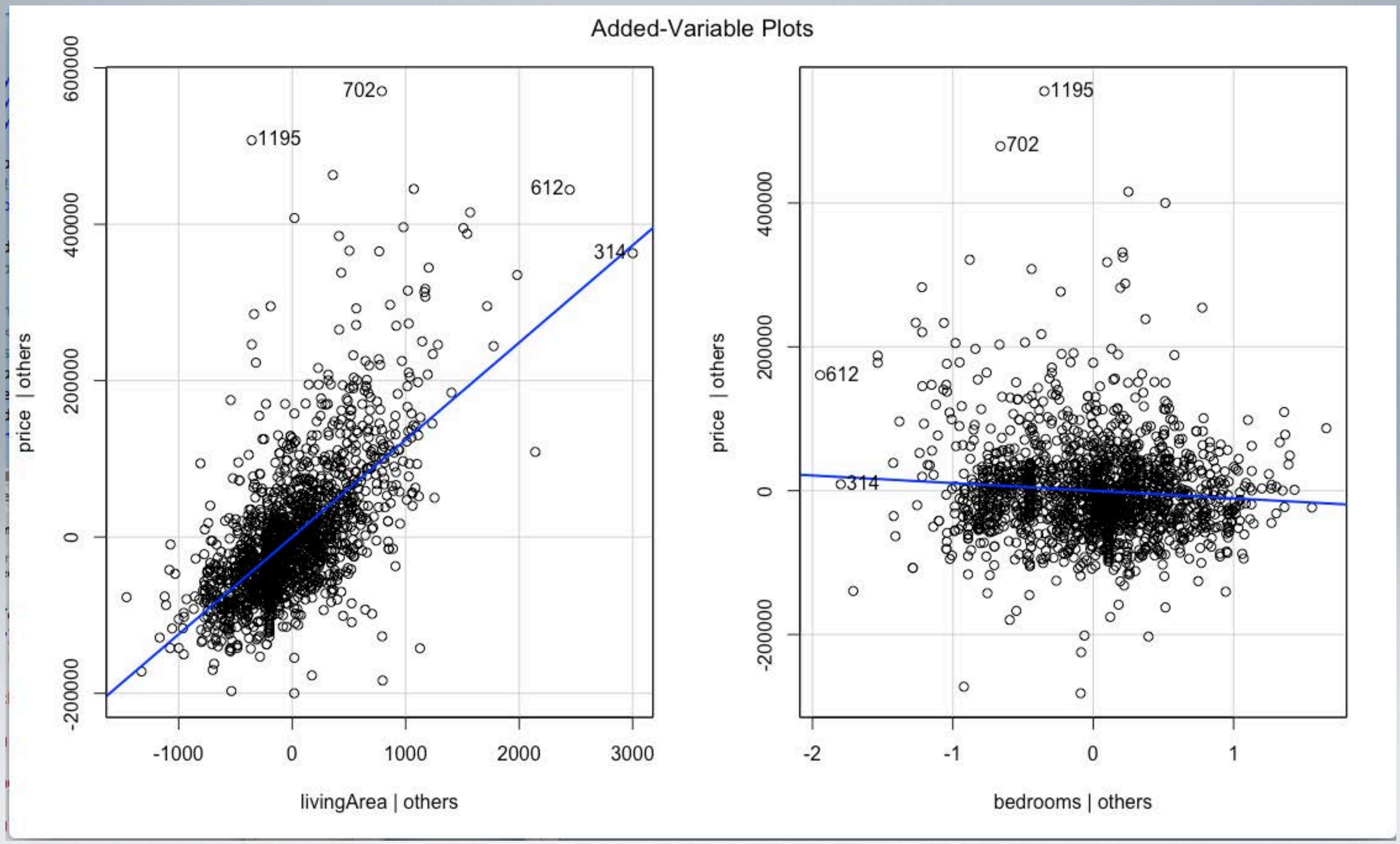
If I chop each bedroom into 4, I'll have an 8 bedroom house worth > $250,000 more !!!

TWO CORRELATED PREDICTORS

```
Coefficients:
               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.

DATA SCIENCE I

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

DATA SCIENCE II

2. Introduction to Data Science II

(a) Vision

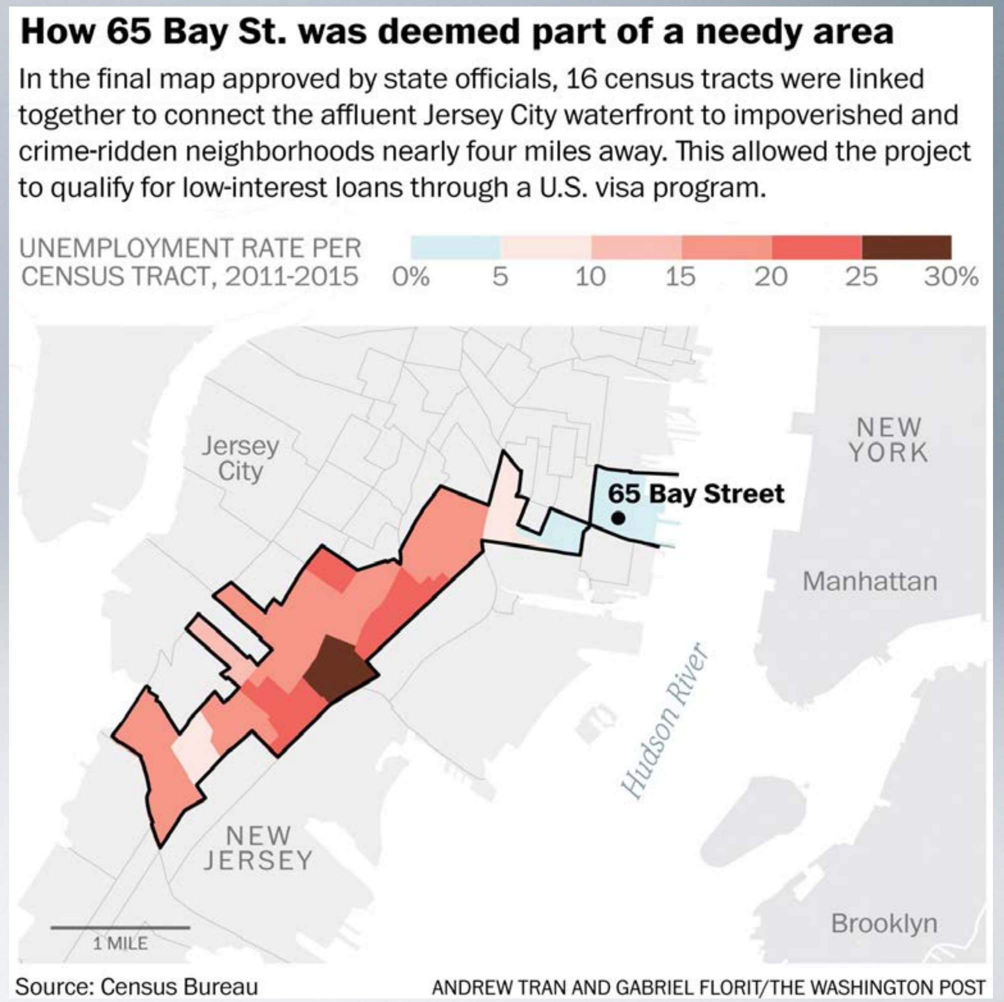
- Exposure to different data types and sources, the process of data curation for the purpose of transforming them to a format suitable for analysis. Introduction to the elementary notions in estimation, prediction and inference. We envision this class to be taught through case studies involving less-manicured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

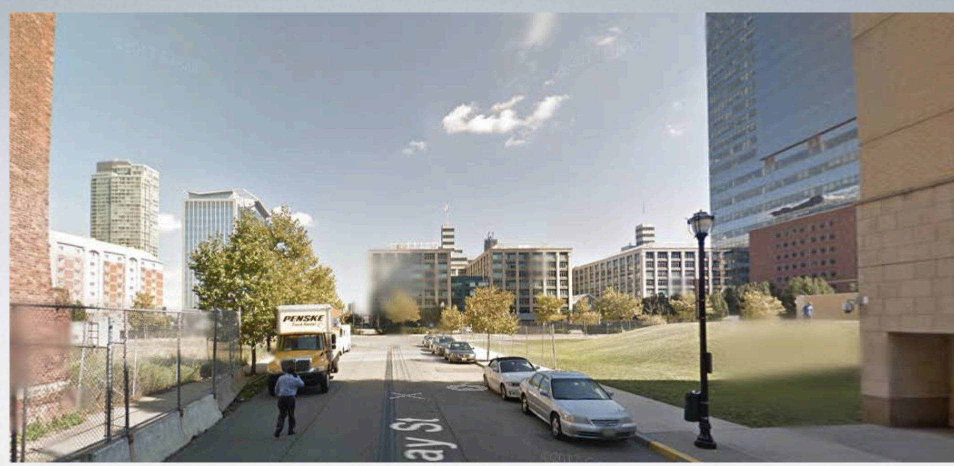
- There are many important ethical issues when dealing with data and models.
- Some are not as obvious as this:



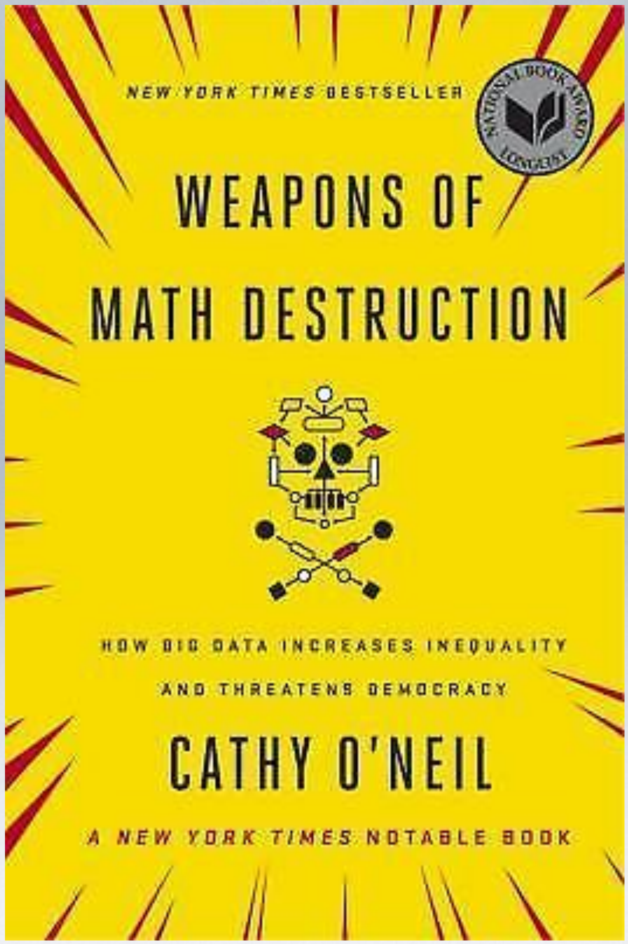
WHAT ABOUT ETHICS?

- Red Area (30% unemployed)

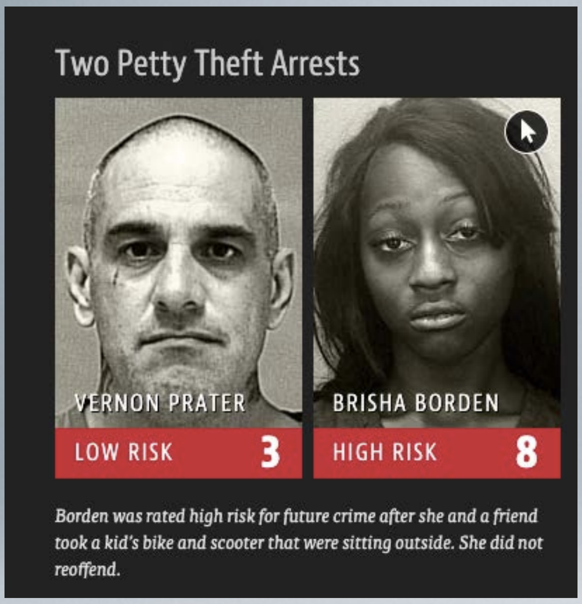
- 65 Bay Street



WHAT ABOUT ETHICS?

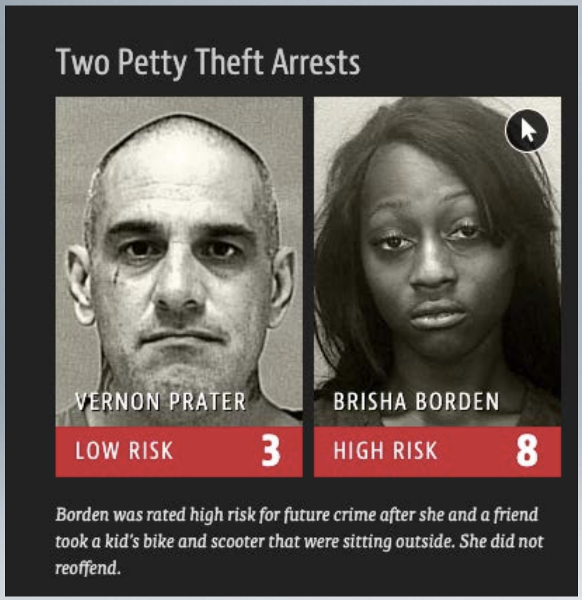


RECIDIVISM ALGORITHM



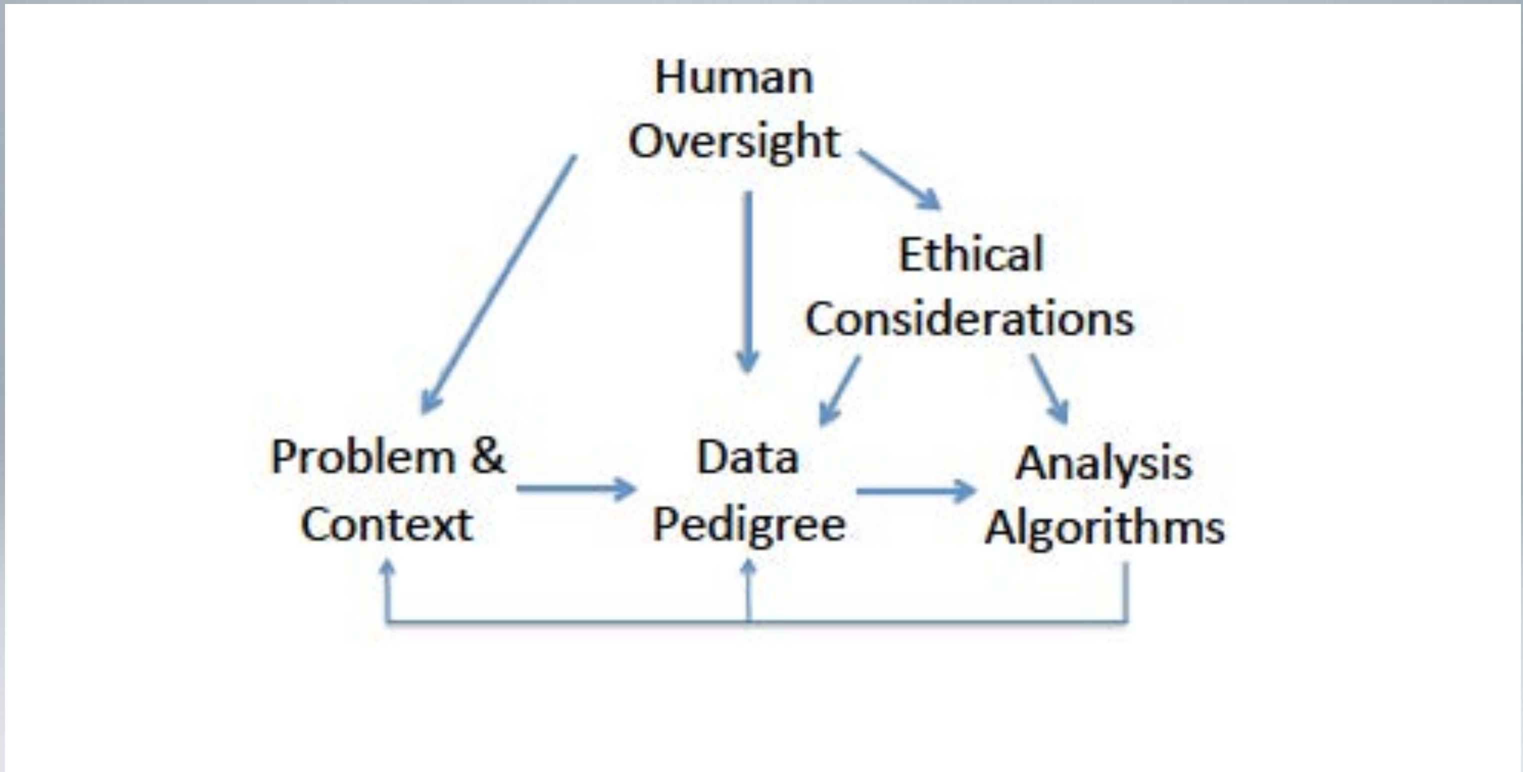
In forecasting who would re-offend, the algorithm correctly predicted recidivism for black and white defendants at roughly the same rate (59 percent for white defendants, and 63 percent for black defendants) but made mistakes in very different ways.

RECIDIVISM ALGORITHM

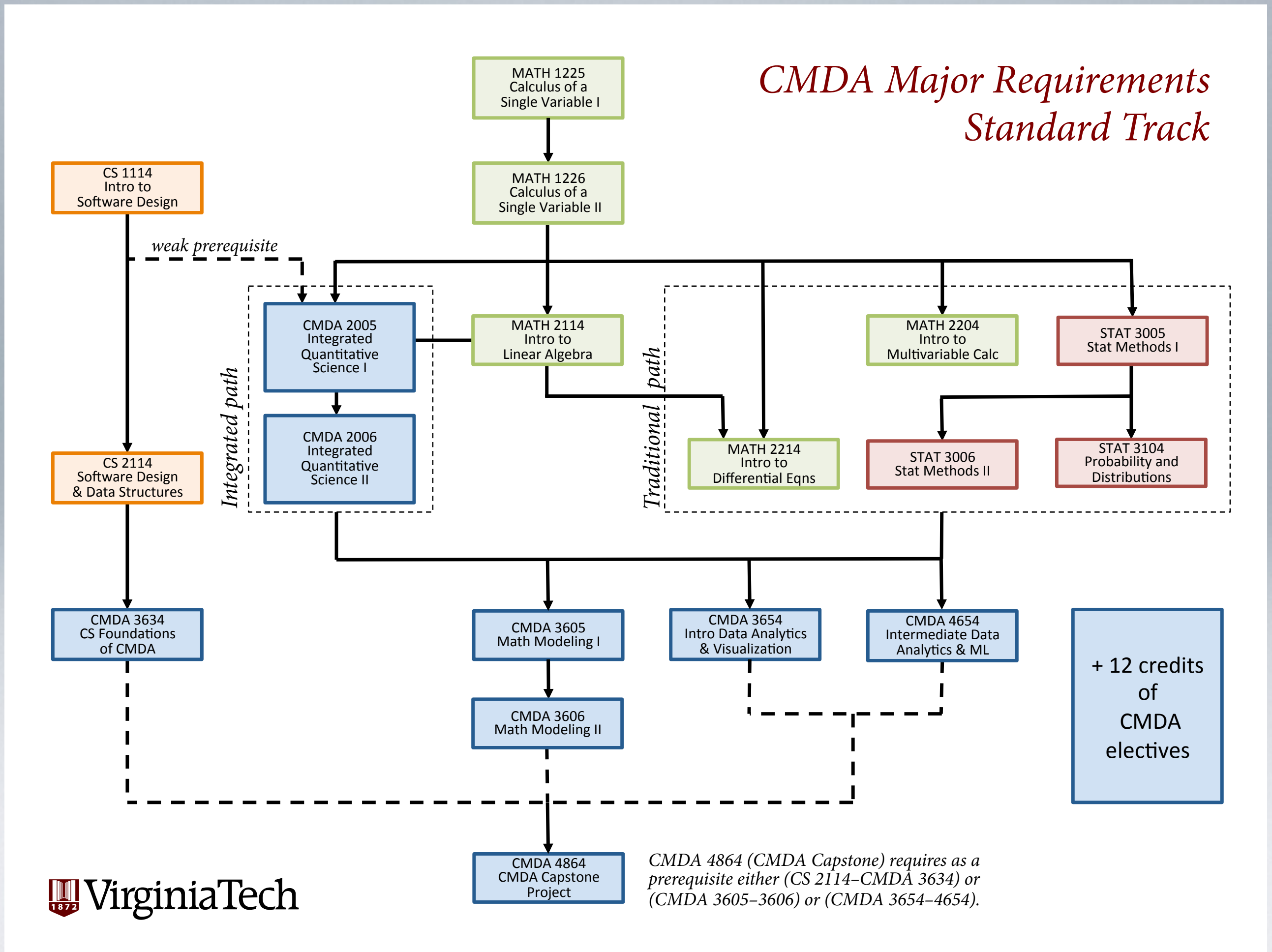


- Black defendants who not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45 percent vs. 23 percent).
- White defendants who re-offended within the next two years were mistakenly labeled low risk almost twice as often 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



PUT IT TOGETHER?

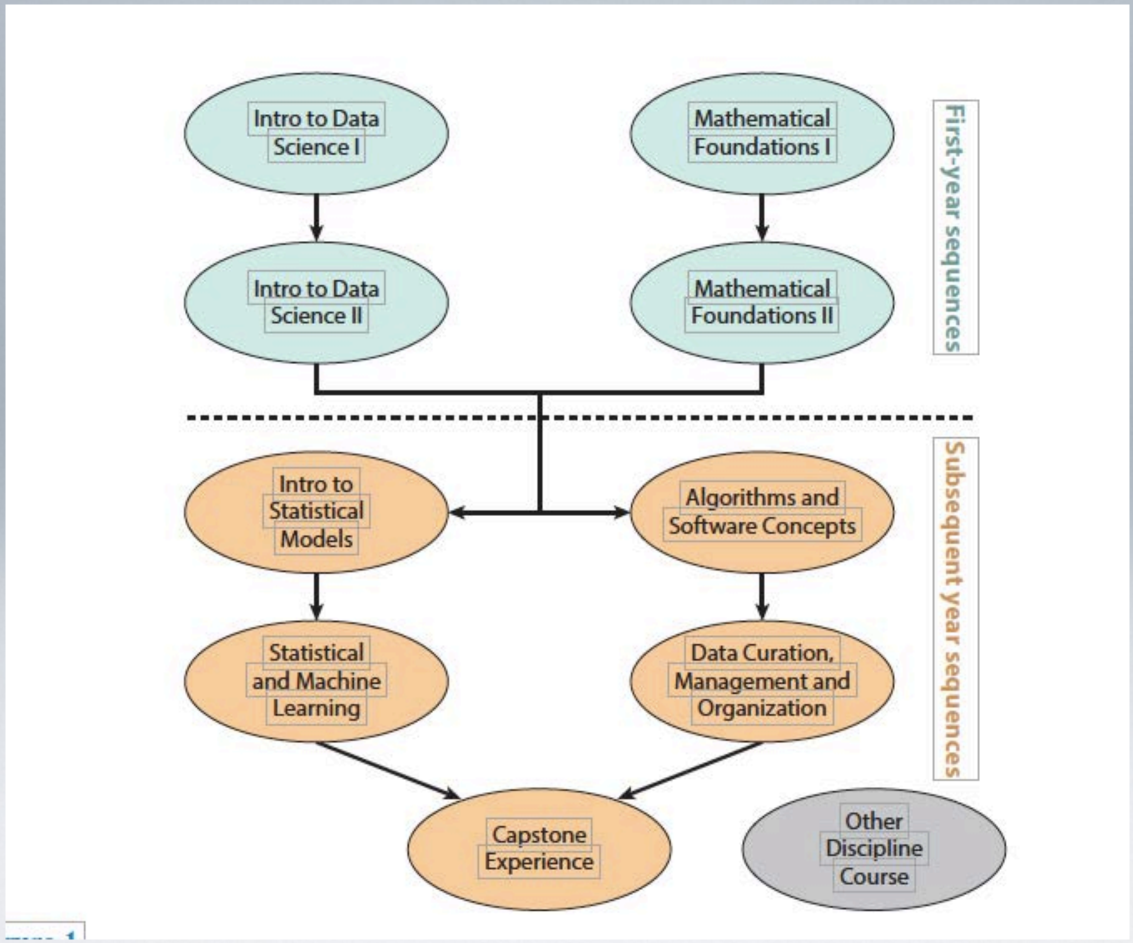


THE PATH FROM HERE?

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)
9. Capstone: Stat 442 - Statistical Learning or CS 374 - Machine Learning
10. Pre-approved domain elective (in something other than CS/Math/Stat?)

THE PATH?



Intro to Data Science I
Mathematical Foundations I
Intro to Data Science II
Mathematical Foundations II
First-year sequences
Intro to Statistical Models
Algorithms and Software Concepts
Statistical and Machine Learning
Data Curation, Management and Organization
Subsequent year sequences
Capstone Experience
Other Discipline Course

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