DATA 605 ACTIONABLE VISUALIZATION & ANALYTICS

DR. WESLEY WILLETT

WINTER 2020

UNIVERSITY OF CALGARY

https://tinyurl.com/DATA605-W2020

INSTRUCTOR

DR. WESLEY WILLETT

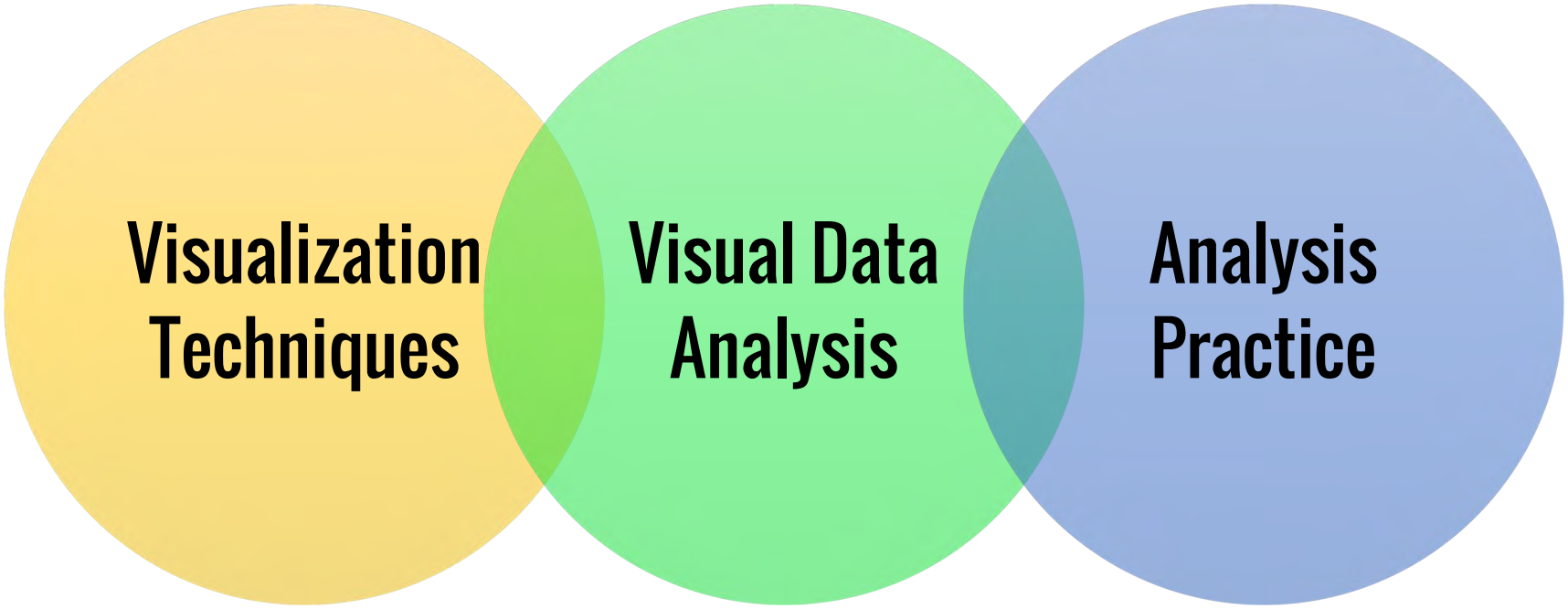
http://dataexperience.cpsc.ucalgary.ca

OFFICE – Math Science 680D

OFFICE HOURS

by appointment

“ACTIONABLE VISUALIZATION & ANALYTICS?”



COURSE GOALS

A DEEPER DIVE INTO **ADVANCED VISUALIZATION TECHNIQUES**

ASSESSING, **CRITIQUING**, AND **DESIGNING** GOOD VISUALIZATIONS

A FOCUS ON PRACTICAL **VISUAL DATA ANALYSIS**

COGNITIVE BIASES, **DATA ETHICS**, AND **GOOD ANALYSIS PRACTICES**

YOU!

QUICK INTROS

Name?
Background and experience?
Any particular interests?

TODAY

Jan 7 — Course Intro

5:00pm - Intros

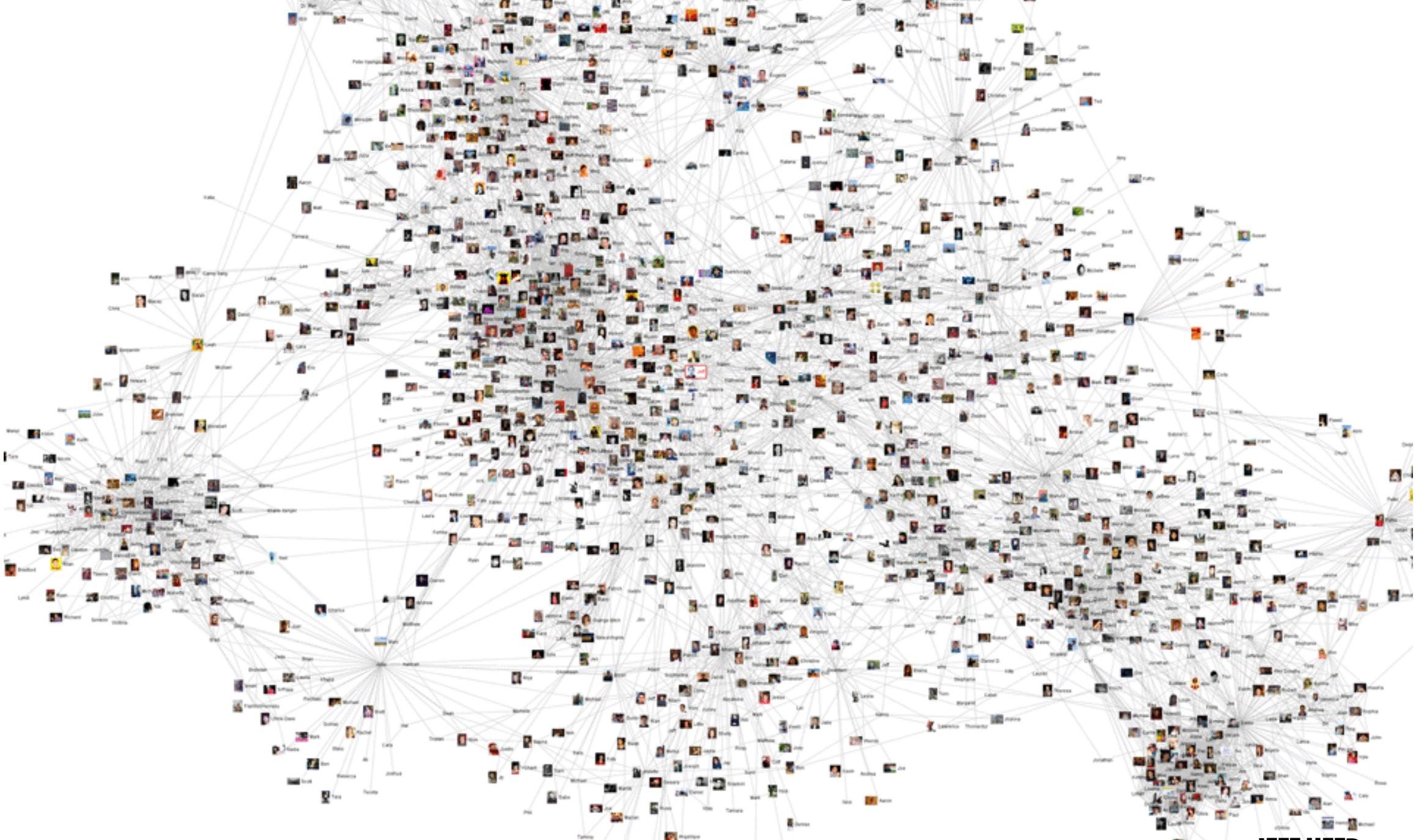
5:15pm - Lecture - Course Intro

6:15pm - Break

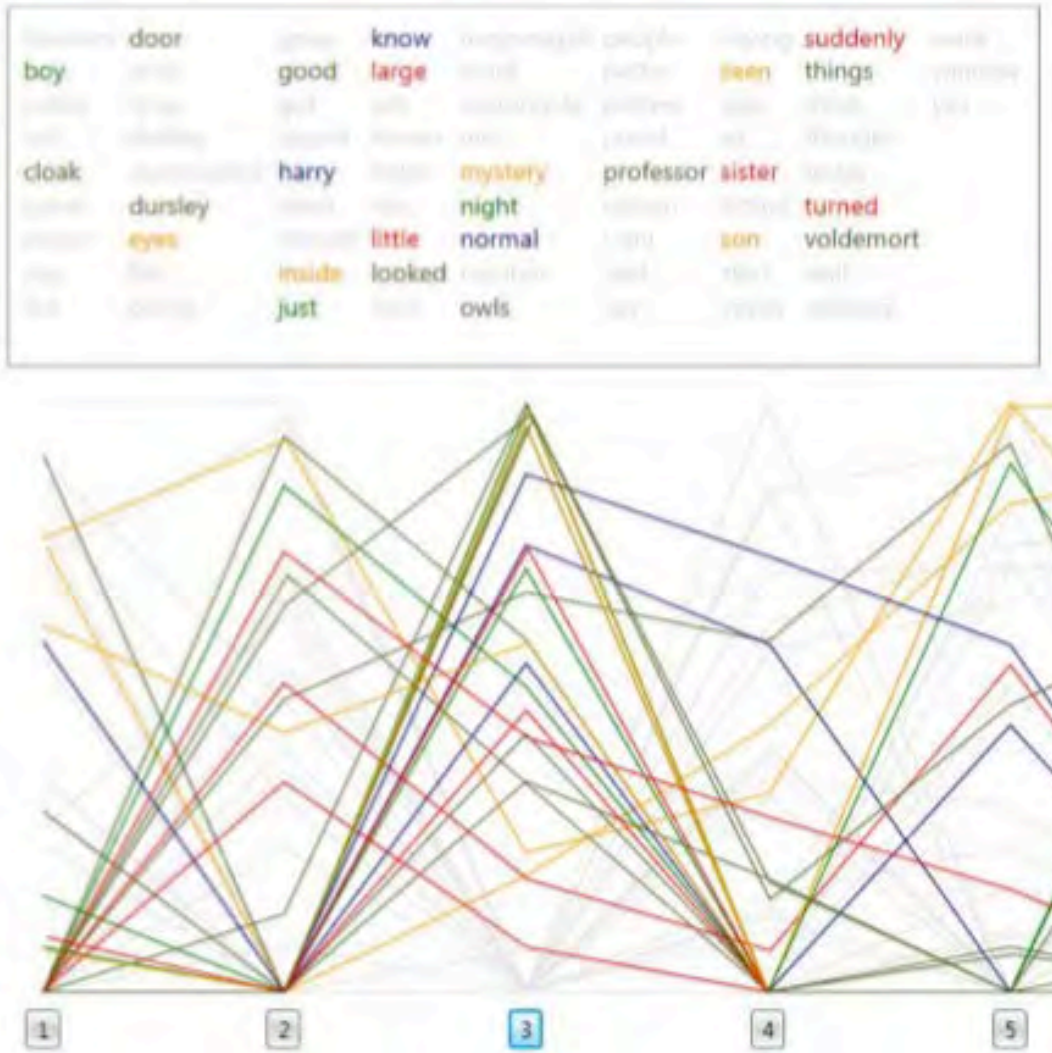
6:30pm - Lecture - Visualization Basics Review

7:30pm - Course Survey

1. MORE ADVANCED VISUALIZATION TECHNIQUES

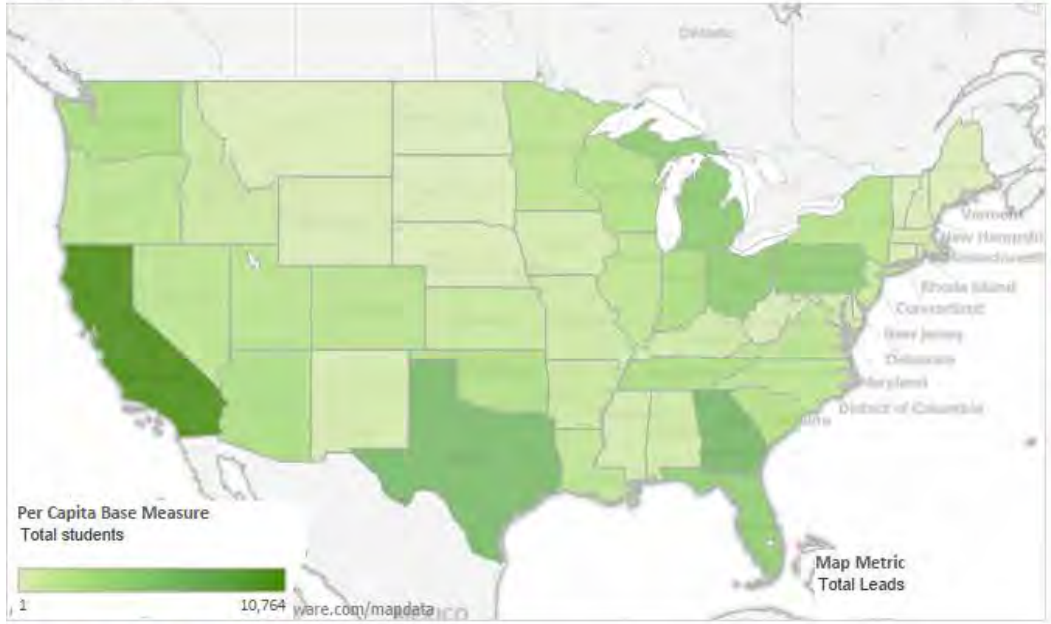


JEFF HEER

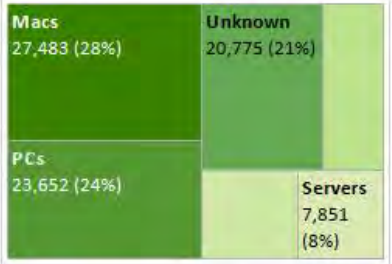


LEE ET AL. 2010

Leads by State

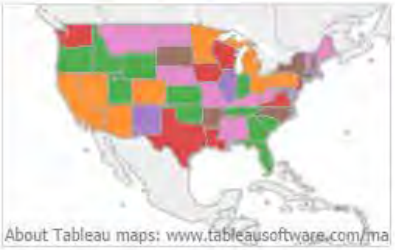


Product Area



Show Classifications Of

State Type



Response Time

Response Time	Convert %	Leads	Converted
< 2 Hrs	6.46%	5,310	343
< 1 Day	4.67%	9,556	446
Later	3.89%	84,134	3,270

Lead Volume Change

	Leads		WoW Change		YoY Change	
	2012	2013	2012	2013	2012	2013
1	4,475	1,933				-57%
2	3,249	1,645	-27%	-15%		-49%
3	1,714	2,035	-47%	24%		19%
4	1,322	4,854	-23%	139%		267%
5	1,476	2,743	12%	-43%		86%
6	5,300	2,643	259%	-4%		-50%
7	3,624	2,420	-32%	-8%		-33%
8	360	1,888	-90%	-22%		424%
9		1,051	-100%	-44%		
10		1,113		6%		
11	1,196	2,639		137%		121%
12	4,418	2,345	269%	-11%		-47%
13	3,990	2,904	-10%	24%		-27%
14	1,155	2,358	-71%	-19%		104%
15		1,809	-100%	-23%		
16		1,086		-40%		
17		1,193		10%		
18		2,941		147%		
19		2,889		-2%		
20		2,616		-9%		
21		3,358		28%		
22		2,554		-24%		
23		1,188		-53%		
24		1,326		12%		
25		2,515		90%		
26		2,411		-4%		
27		2,166		-10%		
28		2,494		15%		
29		1,742		-30%		

Summary

Lead Gen Budget	$3,226,785
Leads	99,000
Budget per Lead	$32.59
Converted	4,059
Budget per Conversion	$794.97
Convert %	4.10%

Filters

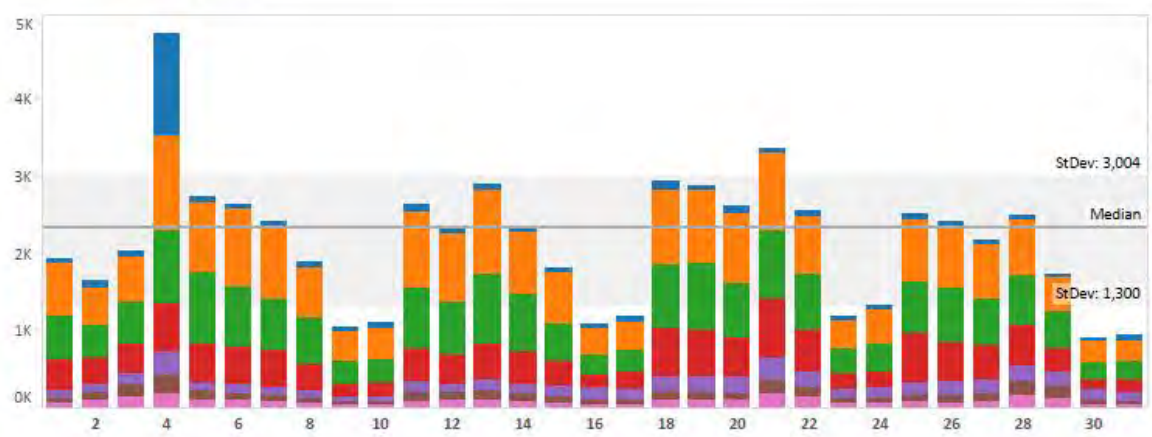
Region

State Type

Lead Source
All

Generated By
All

Breakdown Lead By Day



Measure: Lead

Date Interval: Day

Color By: State Type

- Unspecified
- Type A
- Type B
- Type C
- Type D
- Type E
- Type F

The Upshot

How the Recession Reshaped the Economy, in 255 Charts

By JEREMY ASHKENAS and ALICIA PARLAPIANO Updated: JUNE 6, 2014

Five years since the end of the Great Recession, the economy has finally regained the nine million jobs it lost. But not all industries recovered equally. Each line below shows how the number of jobs has changed for a particular industry over the past 10 years. Scroll down to see how the recession reshaped the nation's job market, industry by industry.

SCROLL

THE THEORY BEHIND DATA VIS

PERCEPTION (HOW WE SEE INFORMATION)

HOW SHOULD WE REPRESENT DATA TO BEST ACHIEVE A PARTICULAR TASK OR GOAL?

2. “CRITICAL THINKING WITH DATA”

(And building competency **actually doing** data analysis.)

Is there anything interesting in this dataset?

Is there a clear difference between these two groups of users?

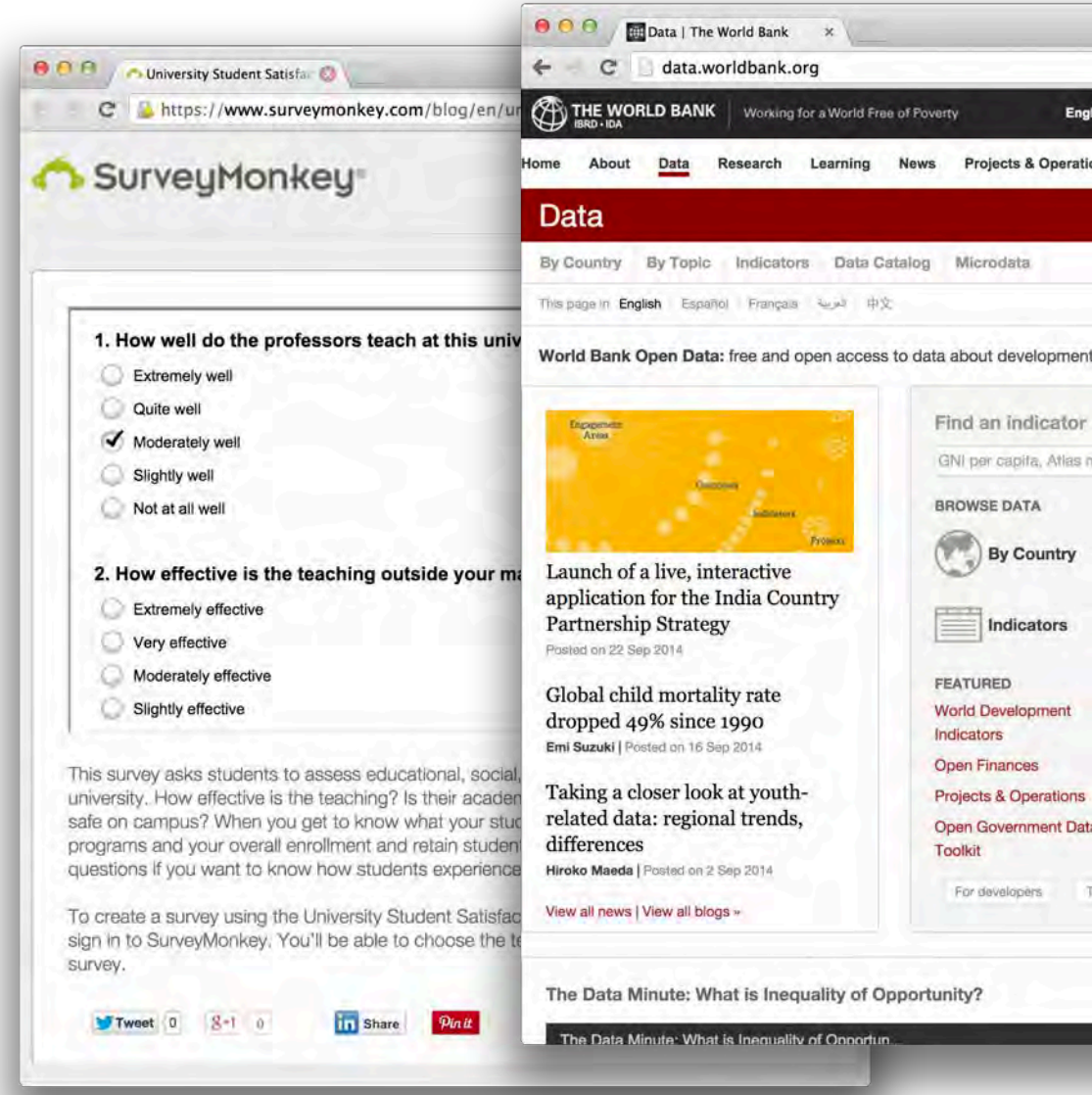
How well do you expect this stock to do in the next month?


DATA COLLECTION
DATA CLEANING
VISUAL ANALYSIS
STATISTICAL ANALYSIS
PRESENTATION

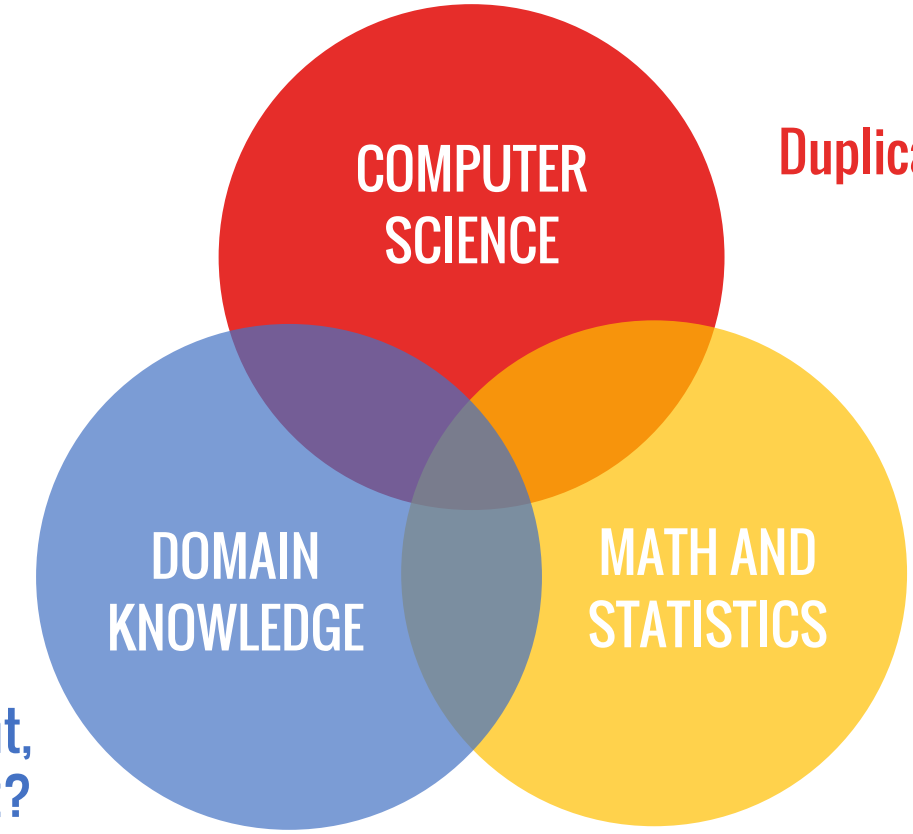

DATA COLLECTION
DATA CLEANING
VISUAL ANALYSIS
STATISTICAL ANALYSIS
PRESENTATION

DATA COLLECTION AND TRANSFORMATION

- Where to find data sets
- Formatting and integrating datasets
- Getting data into tools and ready for use



WHAT DOES IT MEAN TO BE "DIRTY"?



Duplicate values, incorrect formatting, etc.

Data doesn't look right,
Answer doesn't look right?

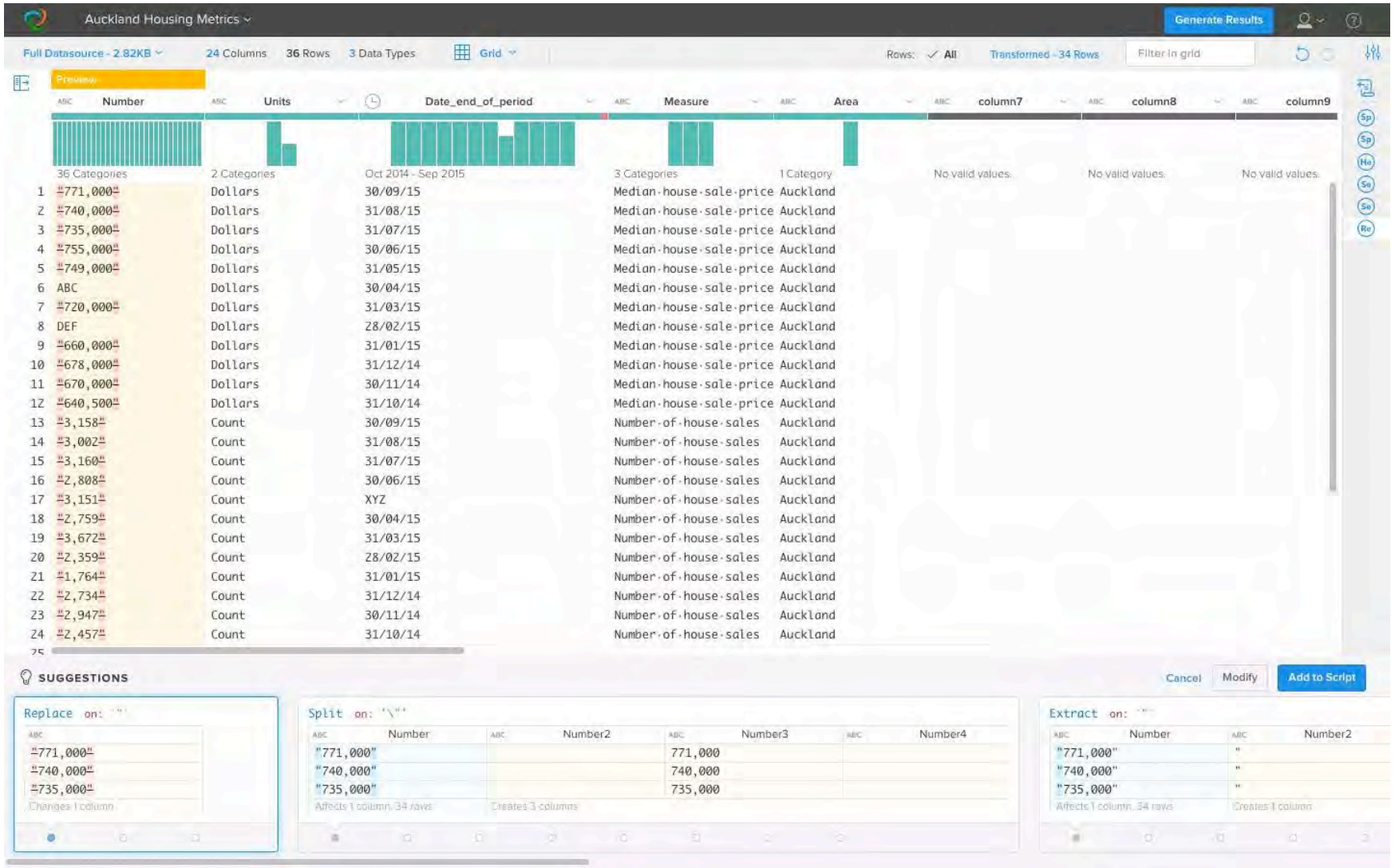
Distortion, selection bias,
Dependence, etc.

DATA MUNGING / DATA WRANGLING

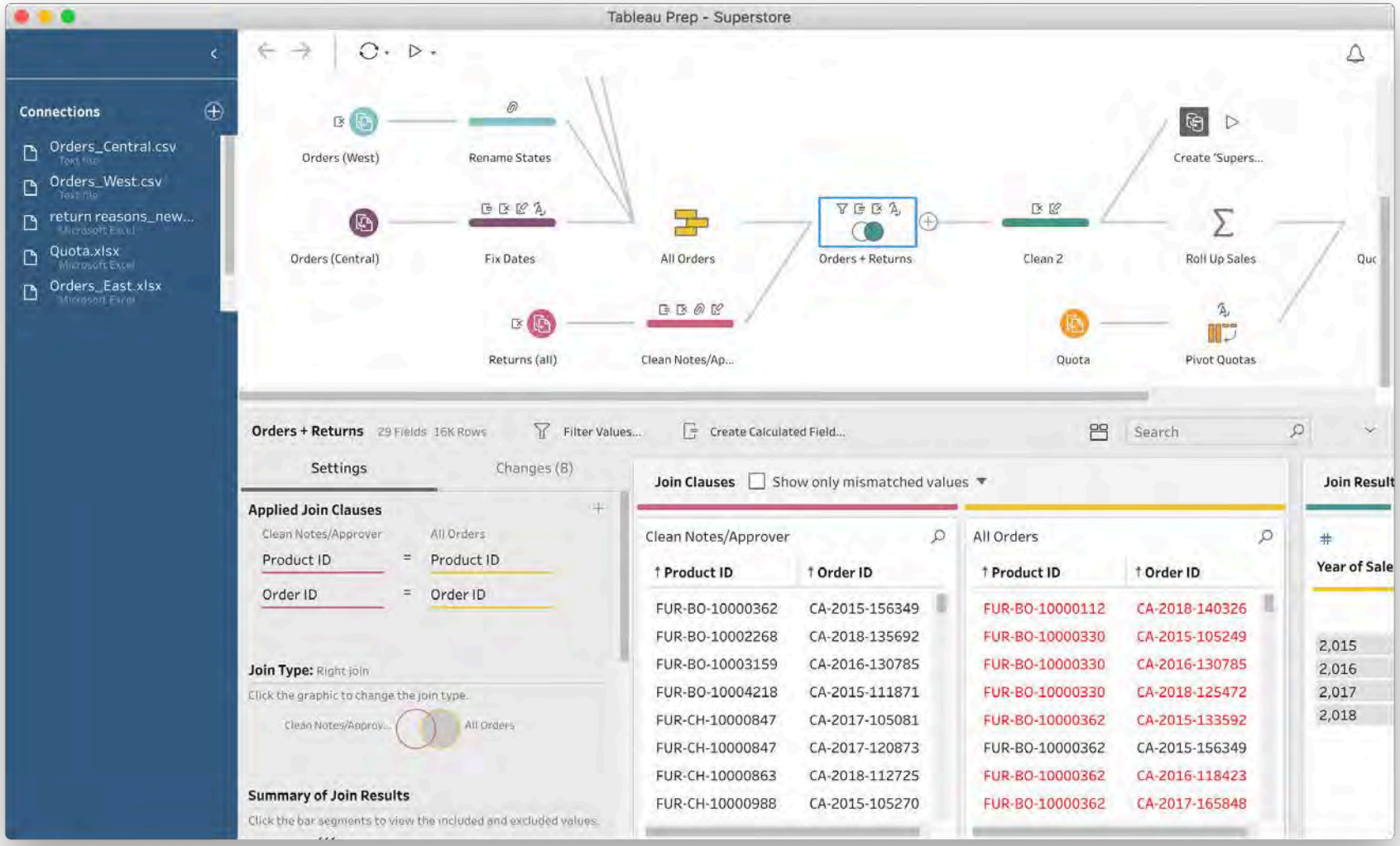
- Reformatting data
- Transforming data
- Profiling
- Correcting and removing errors
- Verifying data
- Supplementing data

TOOLS: BASIC TOOLBOX

TOOLS: WRANGLER



TOOLS: TABLEAU PREP




DATA COLLECTION
DATA CLEANING
VISUAL ANALYSIS
STATISTICAL ANALYSIS
PRESENTATION

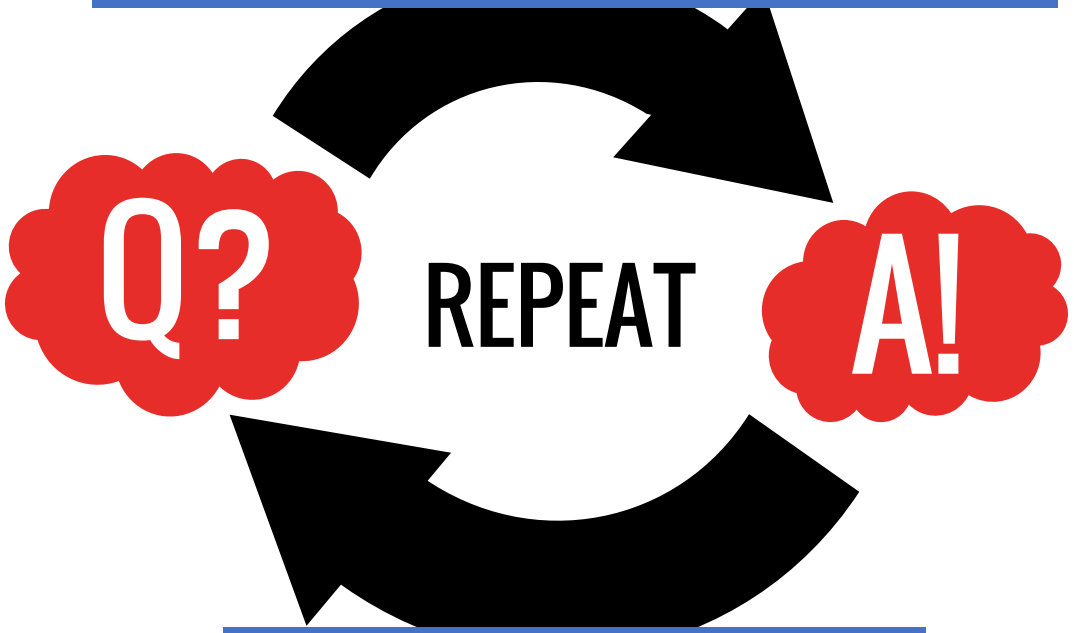
WHY CONDUCT VISUAL ANALYSIS?

(HINT: Usually, because you have questions you need to answer! Even if you don’t know what they are!)

DATA ➡ ANSWERS

ANALYSIS IS A CYCLE

GATHERING DATA,
APPLYING STATISTICAL TOOLS, AND
CONSTRUCTING GRAPHICS TO
ADDRESS QUESTIONS



INSPECT "ANSWERS" AND
ASSESS NEW QUESTIONS

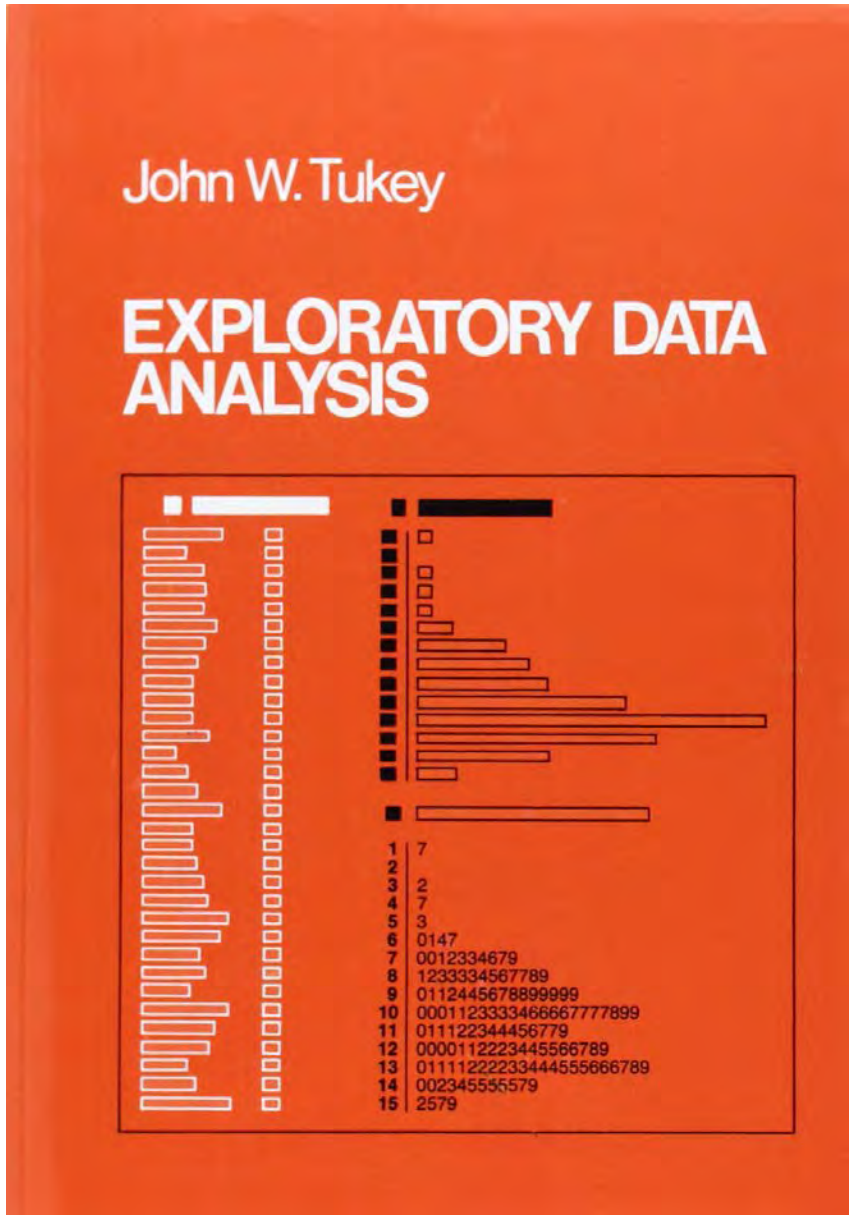
"EXPLORATORY DATA ANALYSIS"

JOHN TUKEY

(IN CONTRAST TO "**CONFIRMATORY**" DATA ANALYSIS)

"**Exploratory data analysis** is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there."

– John Tukey (1979)



DATA ANALYSIS IS ABOUT UNDERSTANDING DATA AND CHECKING ASSUMPTIONS

- IS THE DATA CORRECT?
- DOES IT MATCH OUR PREVIOUS EXPECTATIONS?
- IS THERE A RELATIONSHIP?
  A CORRELATION?
  A TREND?
  ETC.?

A MIX OF METHODS

STATISTICAL APPROACHES

- Summary Statistics
- Correlation Analysis
- Significance Tests

VISUALIZATION APPROACHES

- Visual Exploration
- Comparison
- Visual Inference

Statistical Analysis?

WHY USE BOTH?

Visualization?

BASIC STATS

Measures of Central Tendancy
Mean
Median
Mode
...

Measures of Variability
Min-Max
Standard Deviation
Variance
...

Measures of Relationship/(In)dependence
Correlation
Regression
...

These are easy to compute (sometimes even automatically).

Sometimes you don't want or need to look at all of the data!

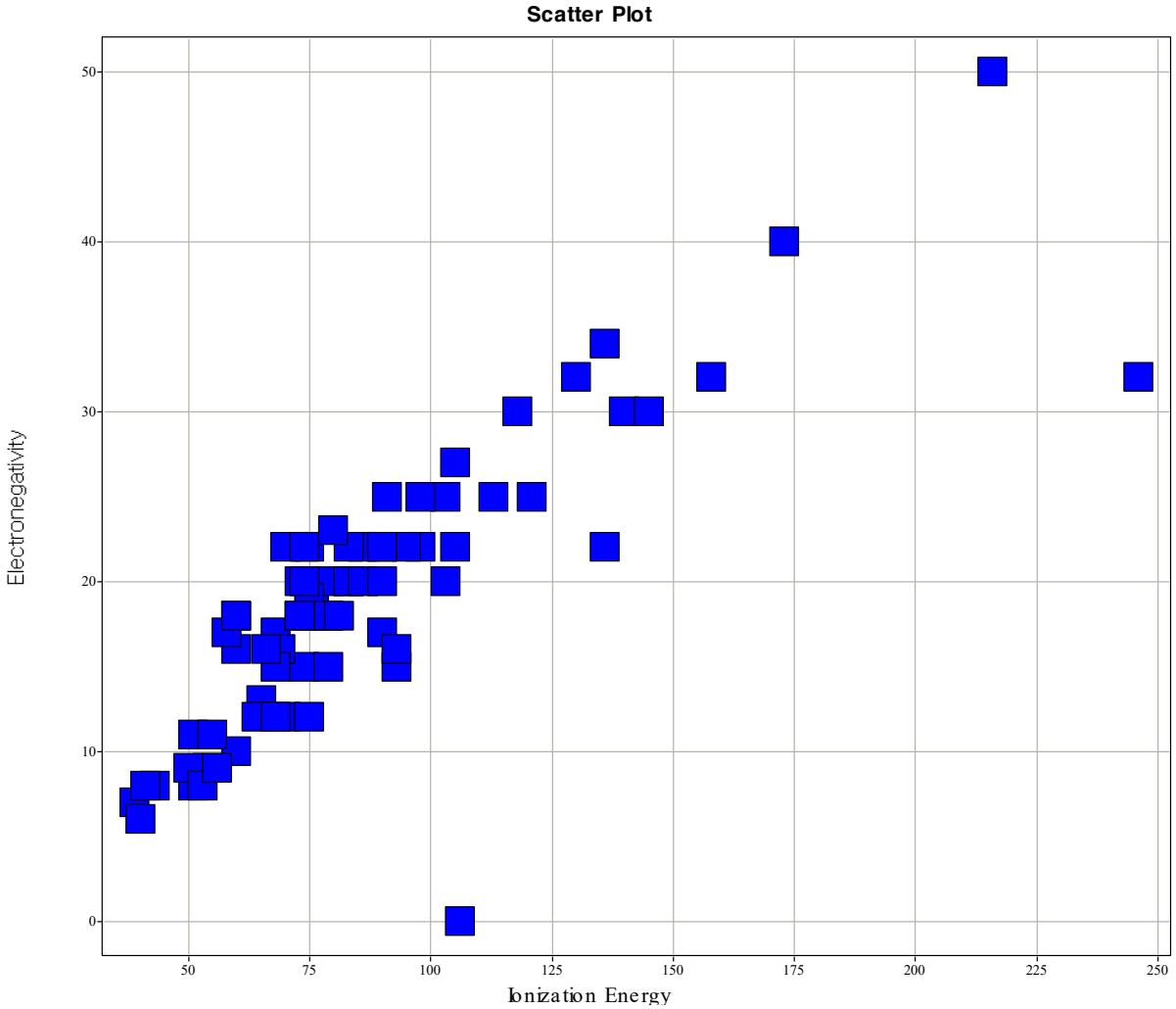
BASIC VIS

	A	B	C	D	E	F	G	H	I	J	K
1	Element	*P1	*P2	Atomic Num	Atomic Mas	Atomic Radi	Ionic Radius	Ionization E	Electronega	*C1	*C2
2	Ac	140	0	89	227	200	126	51	11	62	56
3	Ag	630	80	47	107	144	129	75	18	124	40
4	Al	750	160	13	27	143	67	60	16	28	25
5	Ar	1050	160	18	39	98	154	158	32	176	51
6	As	870	120	33	75	120	72	98	22	115	33
7	At	990	40	85	210	140	76	95	22	119	22
8	Au	630	40	79	197	144	99	91	25	131	22
9	B	750	200	5	10	85	41	83	20	101	8
10	Ba	80	40	56	137	222	149	51	8	46	56
11	Be	80	200	4	9	112	59	93	15	82	15
12	Bi	870	40	83	209	150	117	73	20	140	27
13	Br	990	120	35	79	114	182	118	30	161	44
14	C	810	200	6	12	77	30	113	25	82	1
15	Ca	80	120	20	40	197	114	60	10	70	51
16	Cd	690	80	48	112	151	109	90	17	113	43
17	Cl	990	160	17	35	100	167	130	32	173	47
18	Co	500	120	27	59	125	83	79	18	120	30
19	Cr	320	120	24	52	128	75	68	17	91	28
20	Cs	20	40	55	132	265	181	39	7	7	56
21	Cu	630	120	29	63	128	87	76	19	118	32
22	F	990	200	9	19	72	119	173	40	39	1
23	Fe	440	120	26	55	126	83	79	18	115	32
24	Fr	20	0	87	223	269	194	40	6	1	56
25	Ga	750	120	31	69	135	76	60	18	89	31
26	Ge	810	120	32	72	122	87	79	20	118	33
27	H	20	240	1	1	32	0	136	22	40	1
28	He	1050	240	2	4	31	93	246	32	1	1
29	Hf	200	40	72	178	159	85	70	12	95	44
30	Hg	690	40	80	200	151	116	103	20	147	27
31	I	990	80	53	126	133	206	105	27	153	44
32	In	750	80	49	114	167	94	58	17	93	42
33	Ir	500	40	77	192	136	82	90	22	116	25
34	K	20	120	19	39	227	152	43	8	37	56
35	Kr	1050	120	36	83	112	169	140	30	163	47

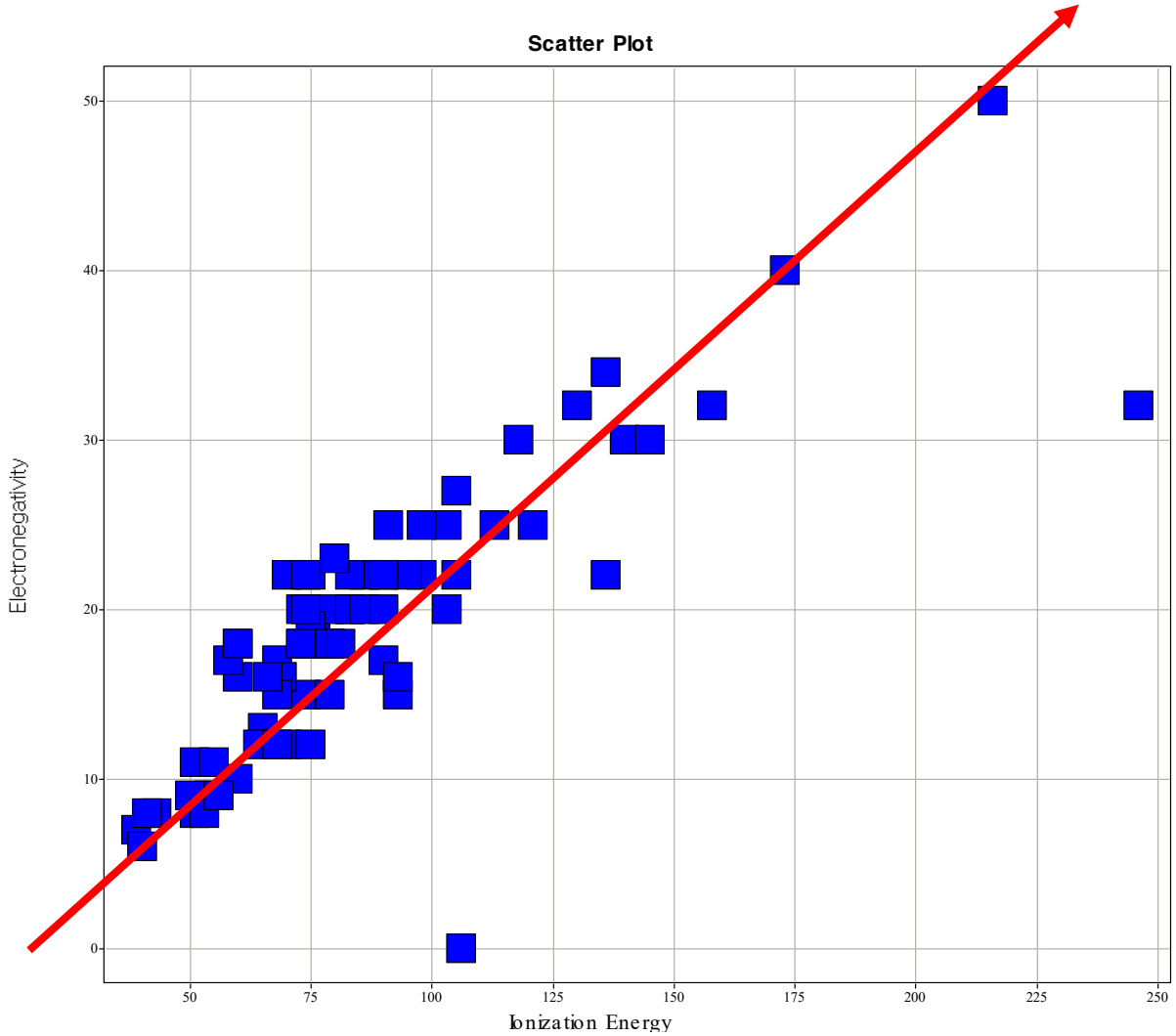
periodic

Sometimes it's hard to see what's interesting!

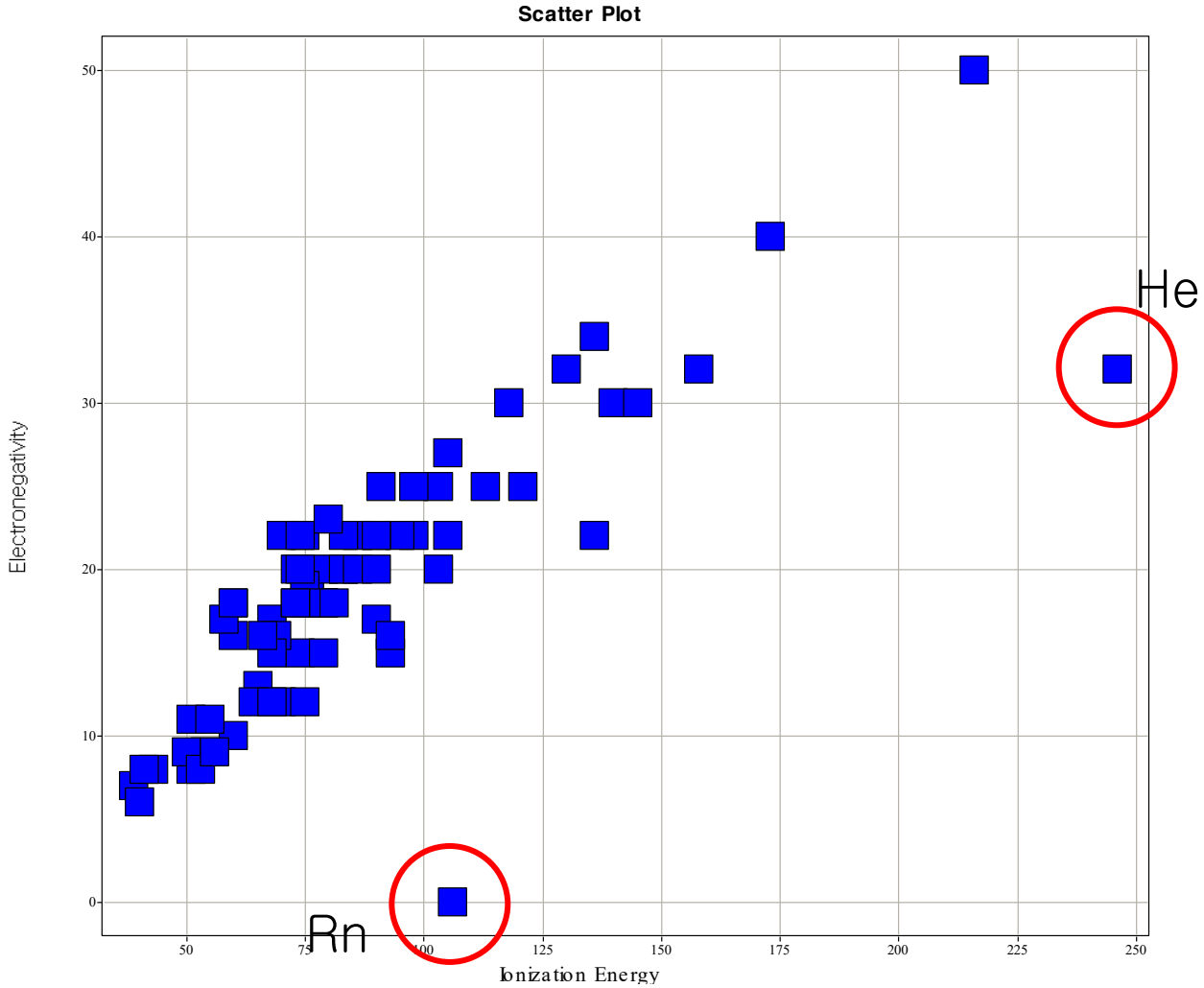
WHAT FEATURES STAND OUT?



CORRELATION...WHAT ELSE?



... AND OUTLIERS



ANOTHER CLASSIC EXAMPLE

Anscombe's Quartet

Francis J. Anscombe

Set A

X	Y
10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68

Set B

X	Y
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.1
6	6.13
4	3.1
12	9.11
7	7.26
5	4.74

Set C

X	Y
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

Set D

X	Y
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.25
19	12.5
8	5.56
8	7.91
8	6.89

Summary Statistics

$\mu_X = 9.0$ $\sigma_X = 3.317$

$\mu_Y = 7.5$ $\sigma_Y = 2.03$

Linear Regression

$Y^2 = 3 + 0.5\,X$

$R^2 = 0.67$

Anscombe 1973

WHO REMEMBERS THEIR BASIC STATS?

Mean?

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Variance?

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2$$

Standard Deviation?

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2}$$

Set A

X	Y
10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68

Set B

X	Y
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.1
6	6.13
4	3.1
12	9.11
7	7.26
5	4.74

Set C

X	Y
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

Set D

X	Y
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.25
19	12.5
8	5.56
8	7.91
8	6.89

Summary Statistics

$\mu_X = 9.0$ $\sigma_X = 3.317$

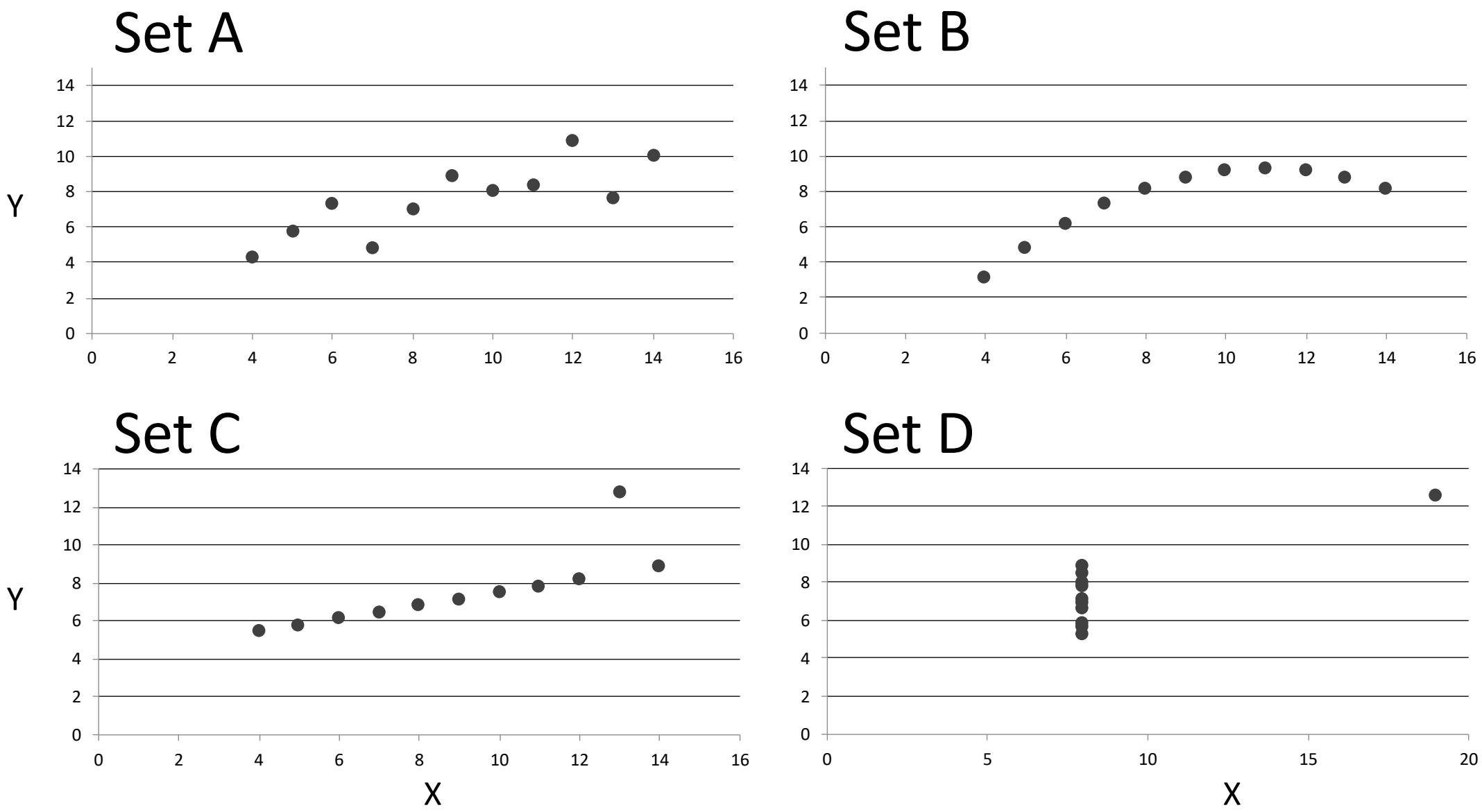
$\mu_Y = 7.5$ $\sigma_Y = 2.03$

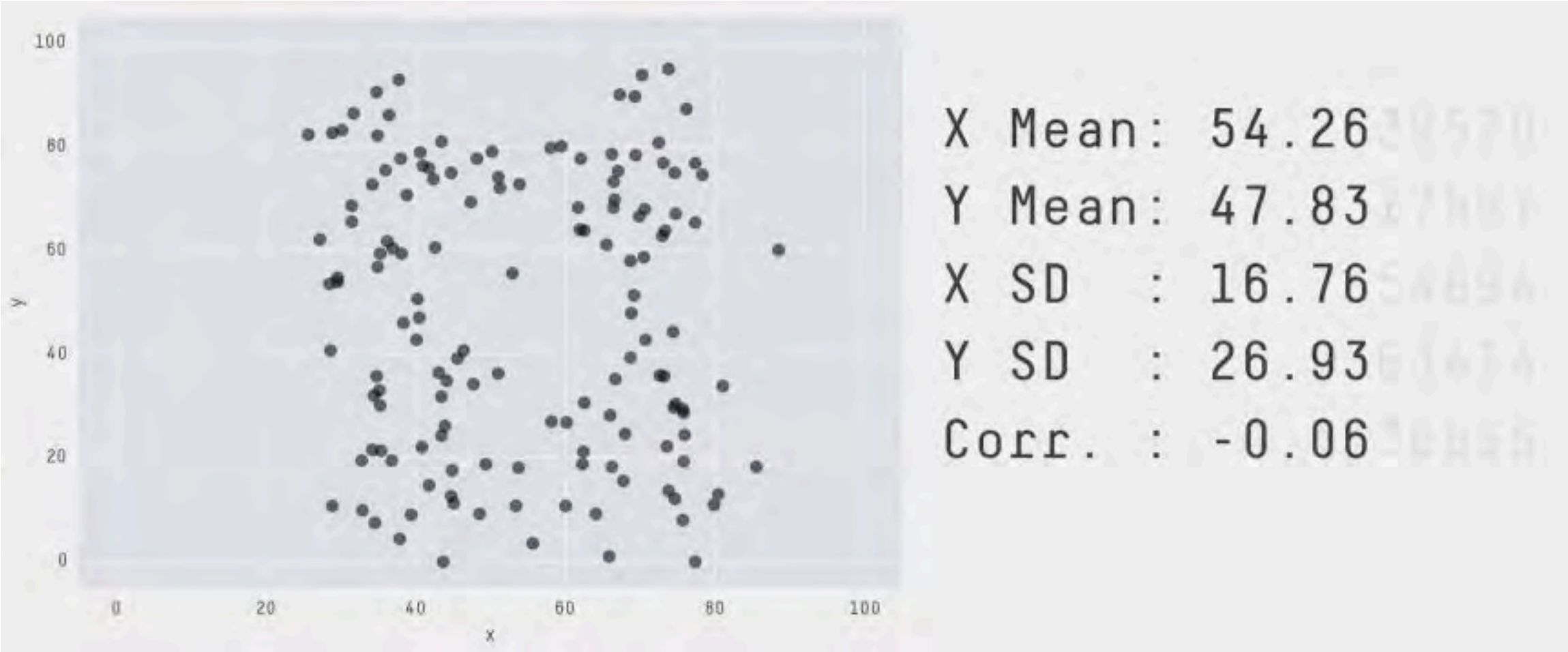
Linear Regression

$Y^2 = 3 + 0.5\,X$

$R^2 = 0.67$

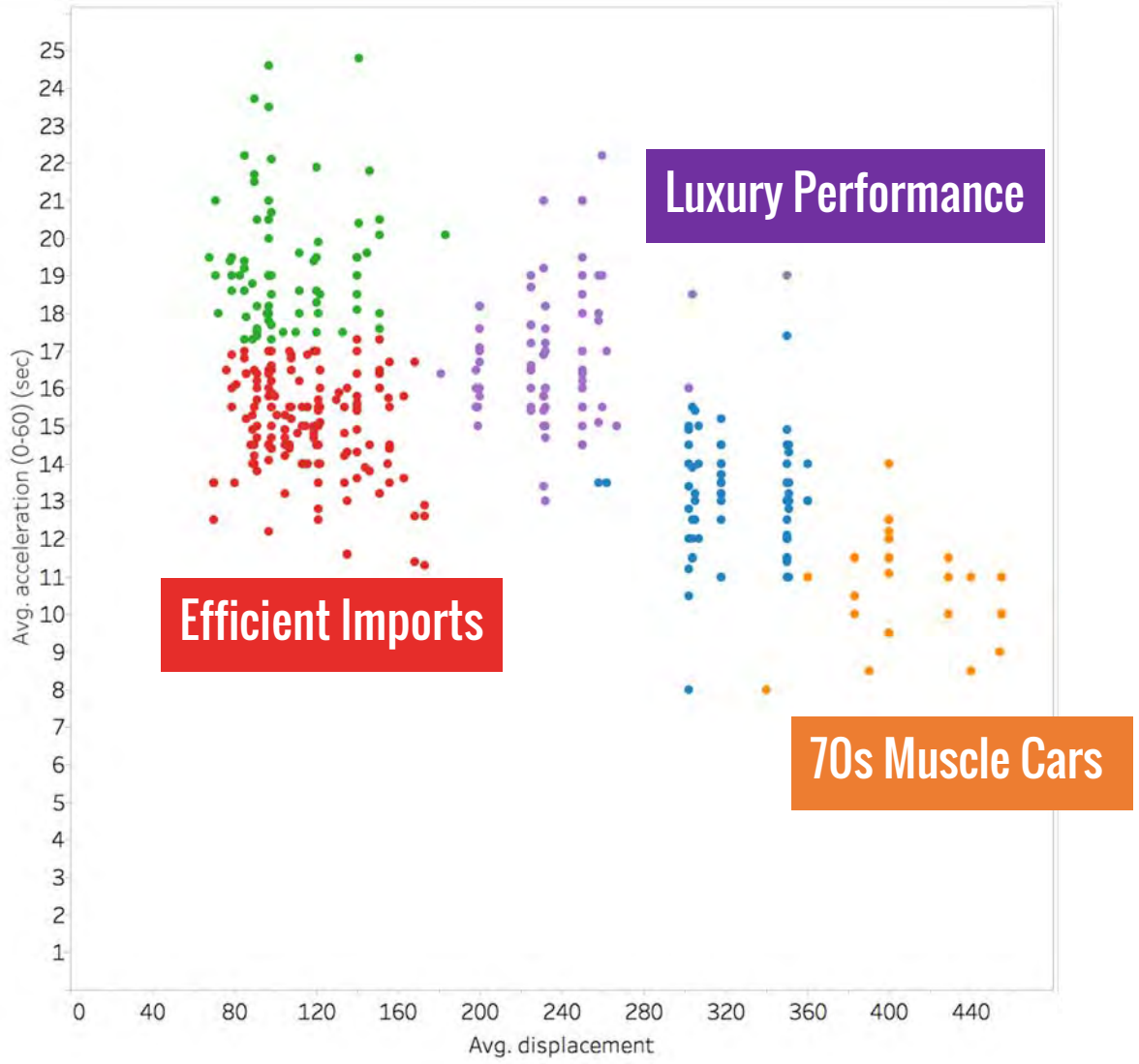
Anscombe 1973


Set A
Set B
Set C
Set D
Y
X
0
2
4
6
8
10
12
14
16
5
15
20

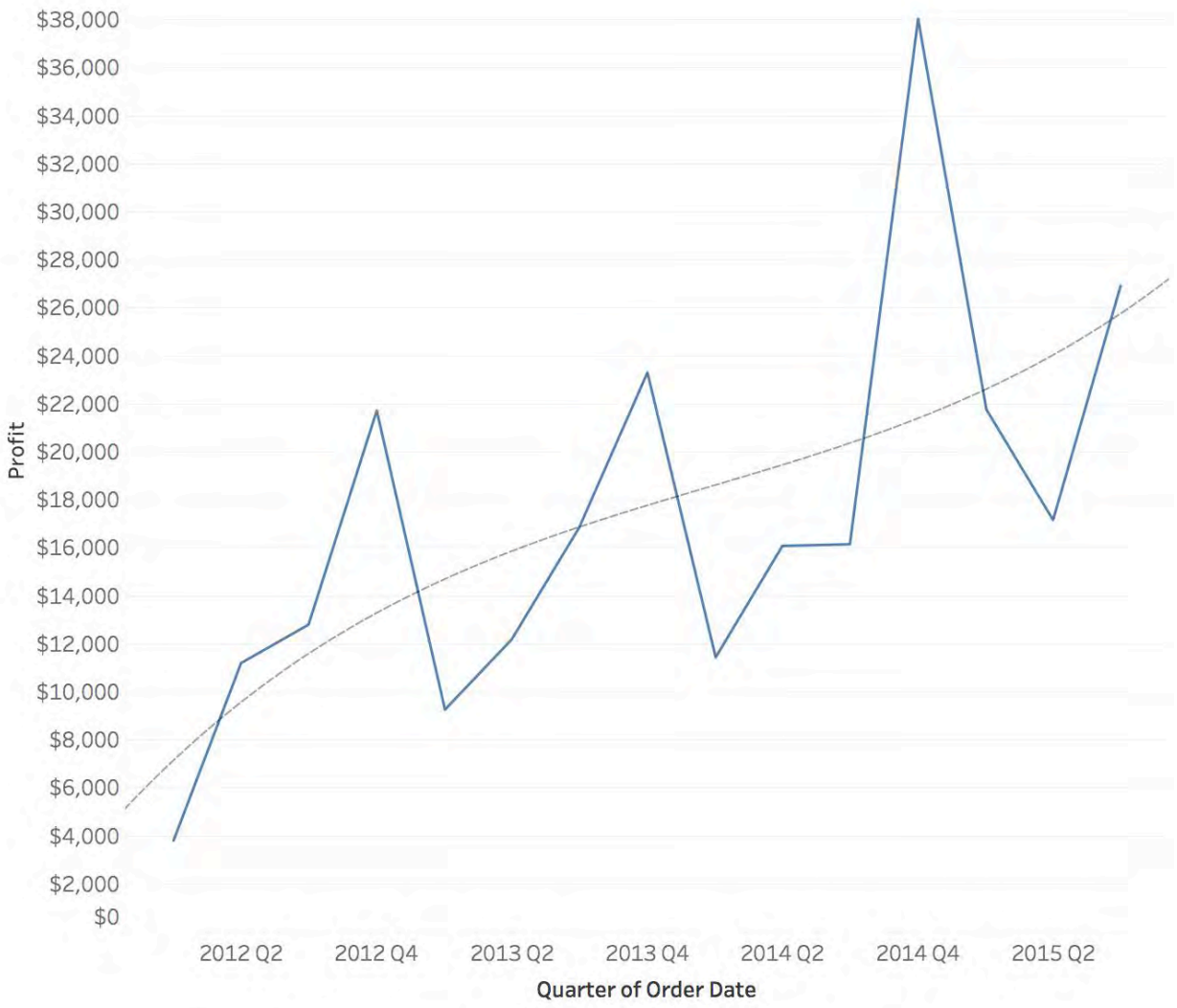


@JustinMatejka / @albertocairo

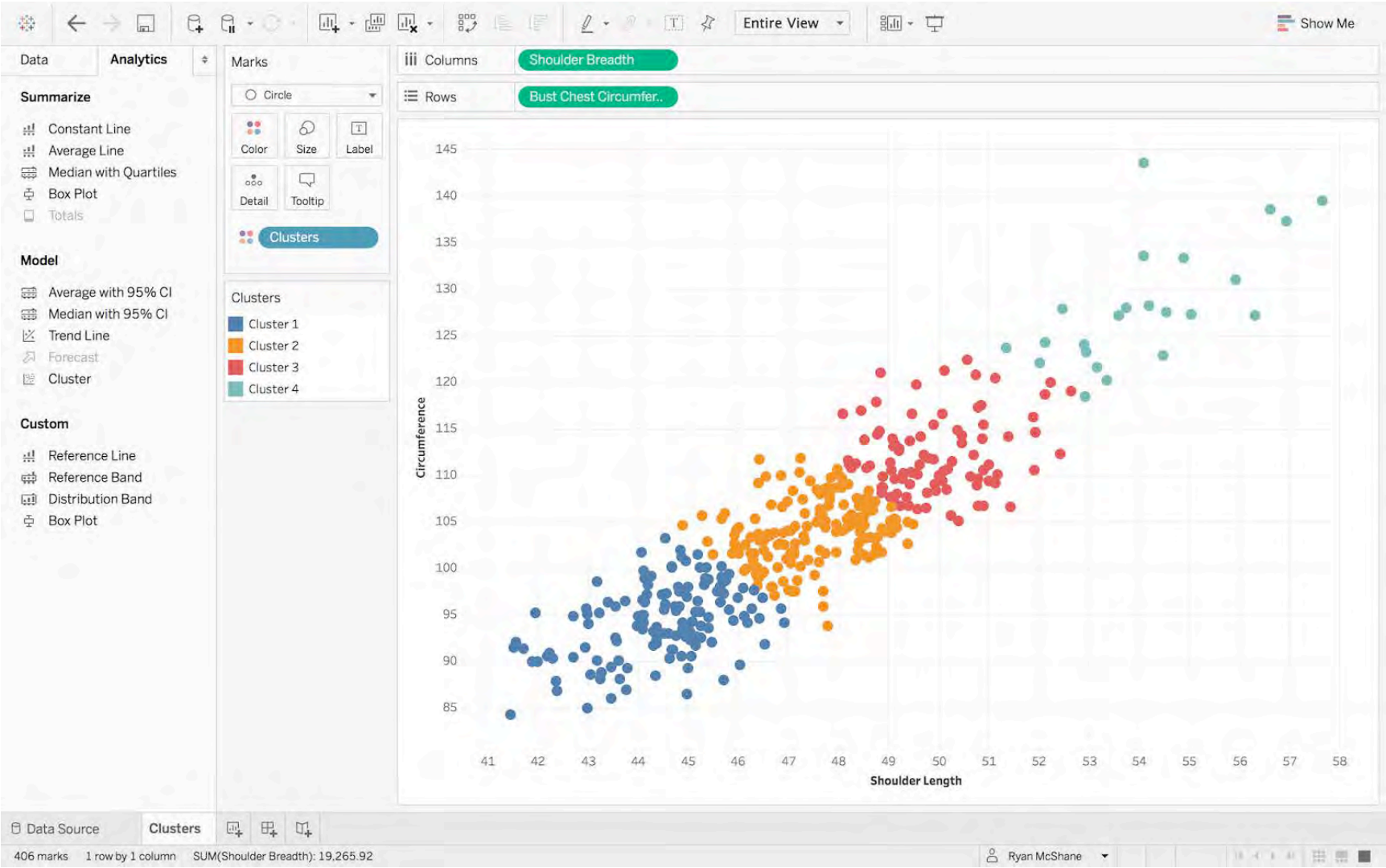
UNDERSTANDING CLUSTERING AND CLASSIFICATIPON



MAKING SENSE OF MODELING AND PREDICTION



TOOLS: TABLEAU

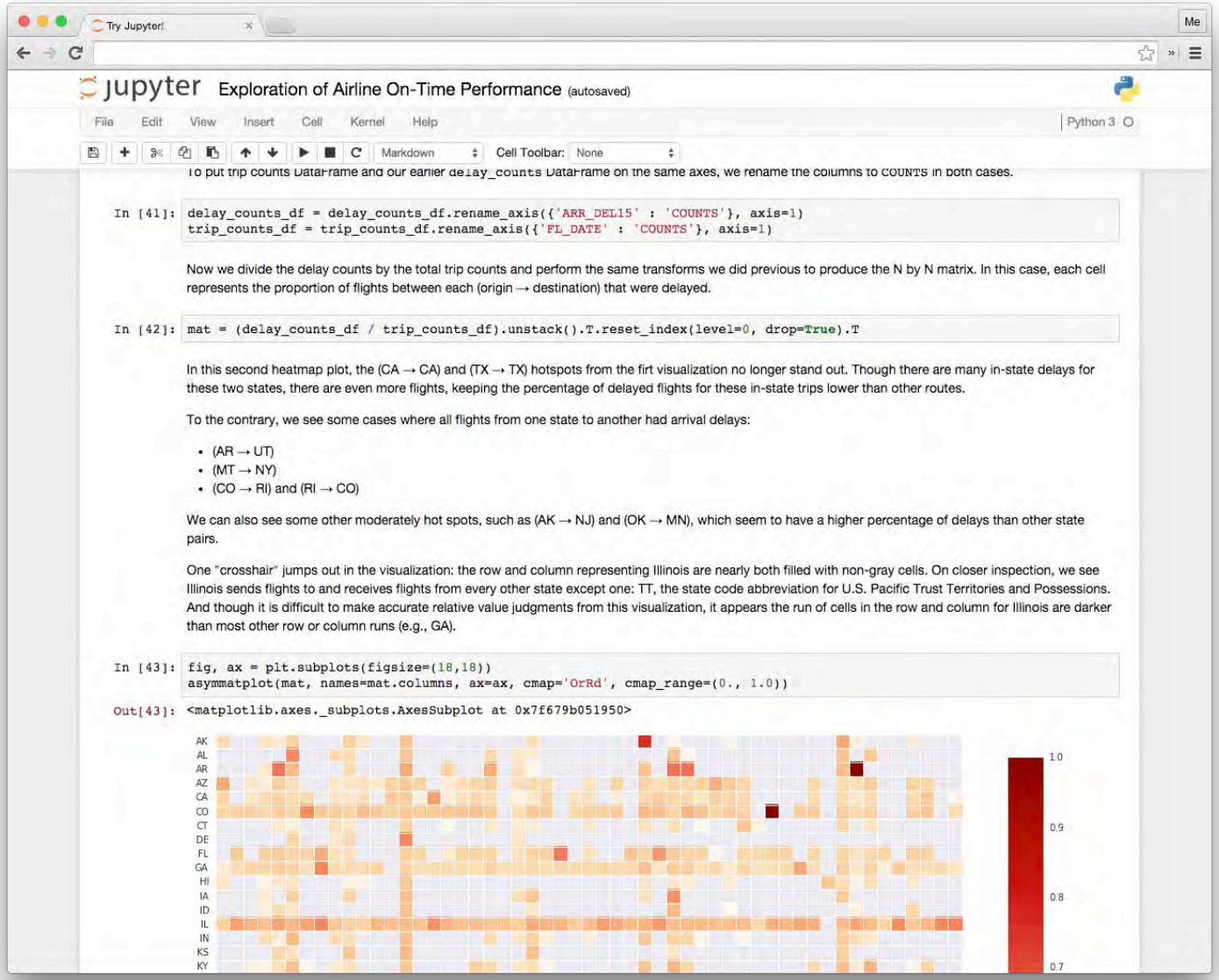


TOOLS: JUPYTER + PYTHON / PANDAS

jupyter

pandas

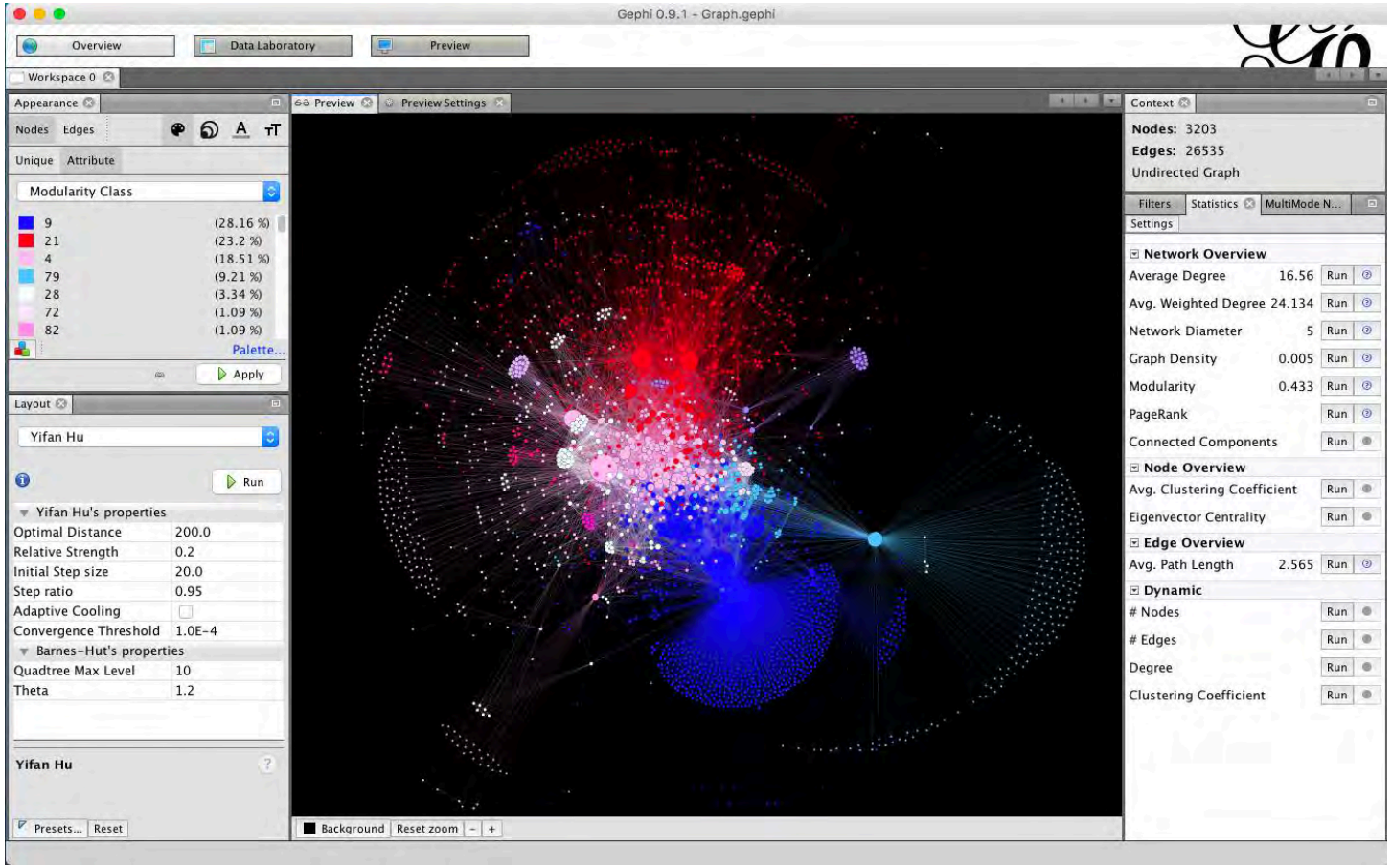
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



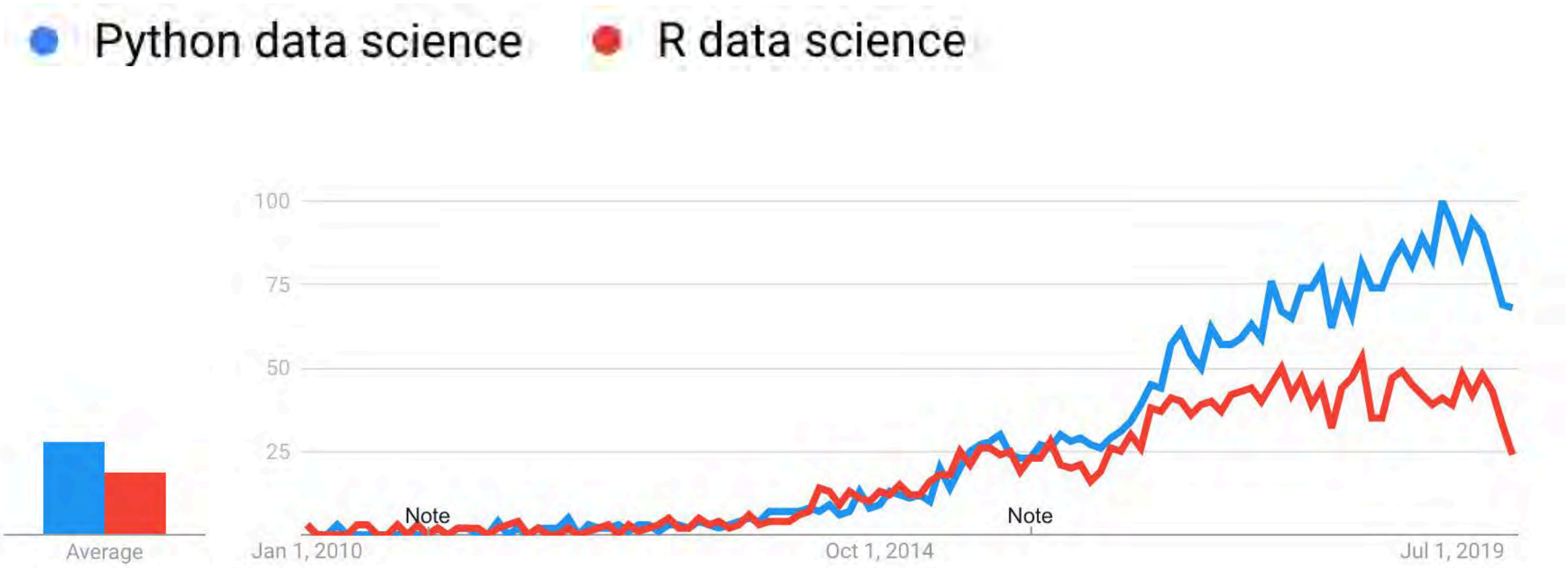
TOOLS: GEPHI

Gephi

- www.gephi.org



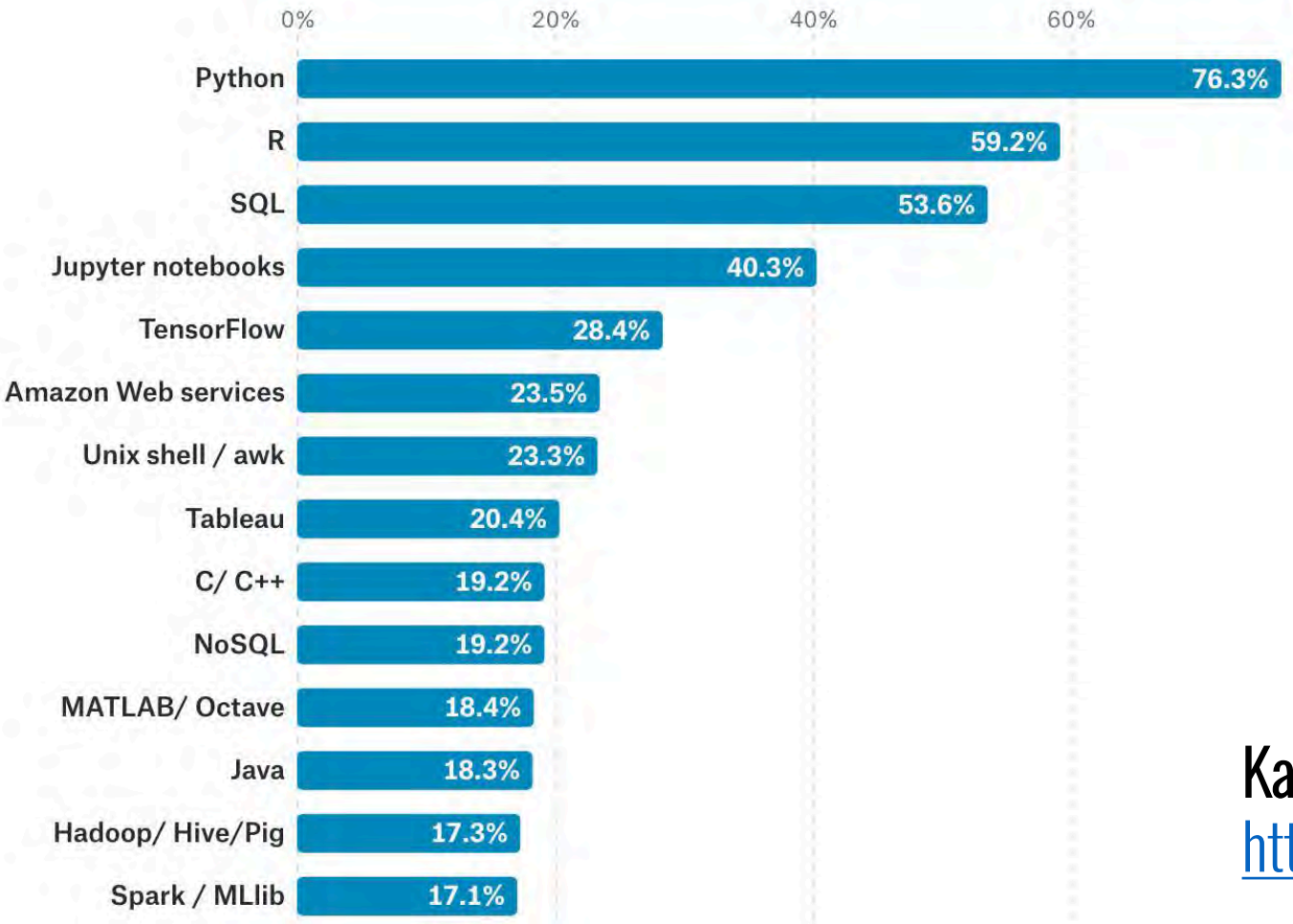
WHY NOT R?


Python data science
R data science
100
75
50
25
Note
Note
Average
Jan 1, 2010
Oct 1, 2014
Jul 1, 2019

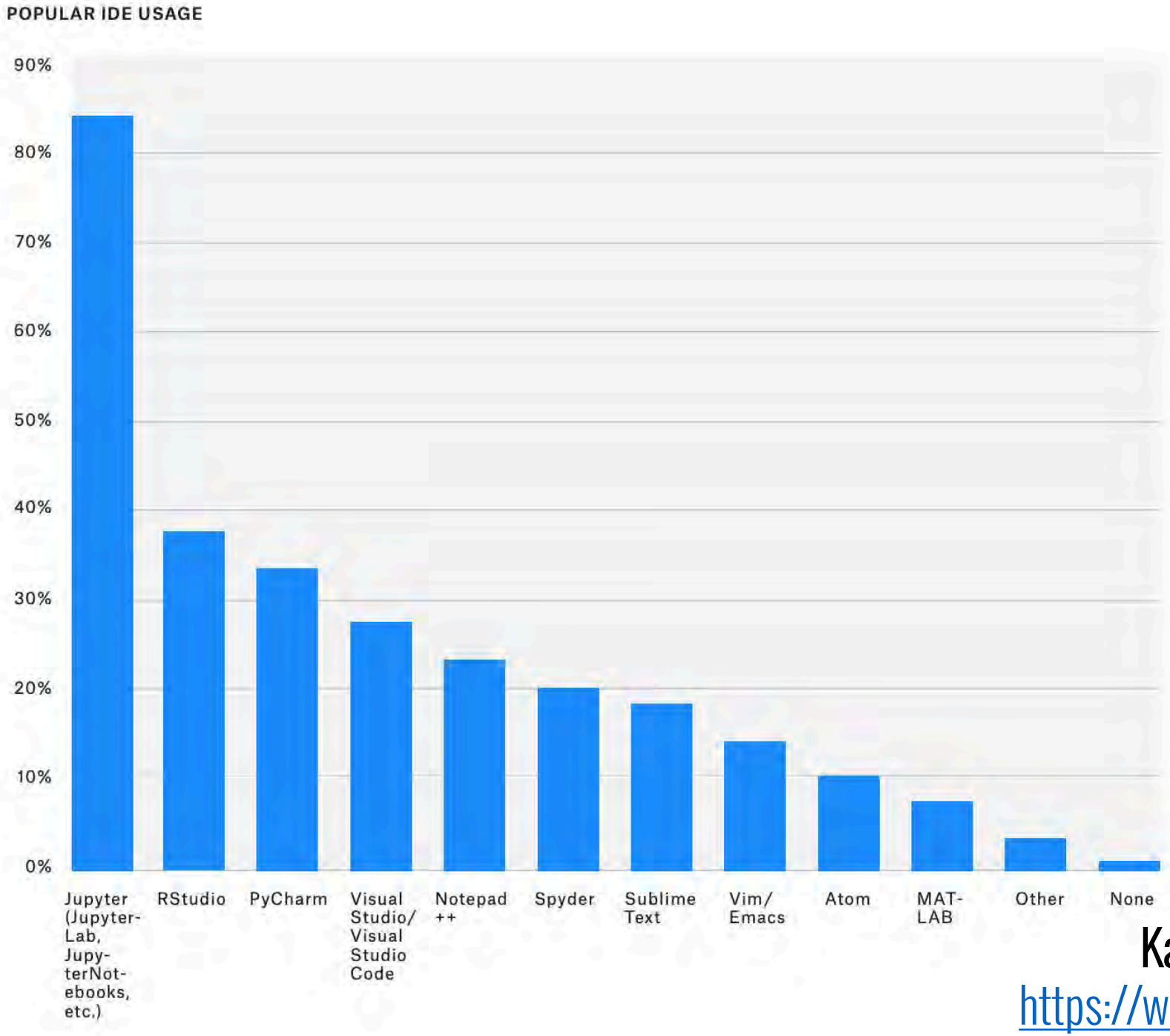
What tools are used at work?

Python was the most commonly used data analysis tool across employed data scientists overall, but more Statisticians are still loyal to R.

Company Size | Industry | Job Title



Kaggle Data Science & ML Survey (2017)
https://www.kaggle.com/surveys/2017



Kaggle Data Science & ML Survey (2019)

https://www.kaggle.com/c/kaggle-survey-2019/

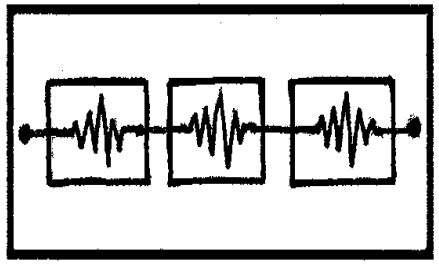

DATA COLLECTION
DATA CLEANING
VISUAL ANALYSIS
STATISTICAL ANALYSIS
PRESENTATION

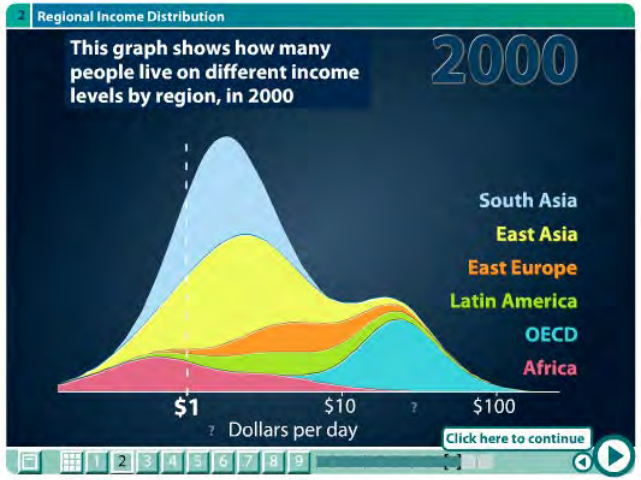
PRESENTATION

- Once you've analyzed a dataset - **how can you make it accessible to others?**
- What visualizations, text, etc. can you use to clearly **communicate your findings?**
- Can you **help others explore the data?**

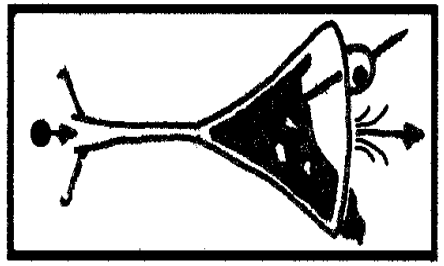
STORYTELLING WITH DATA

Interactive Slide Show

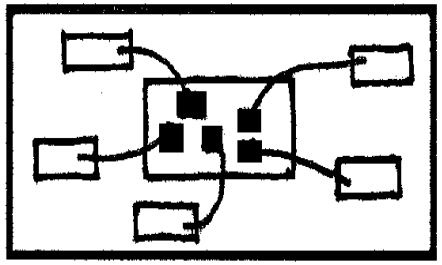


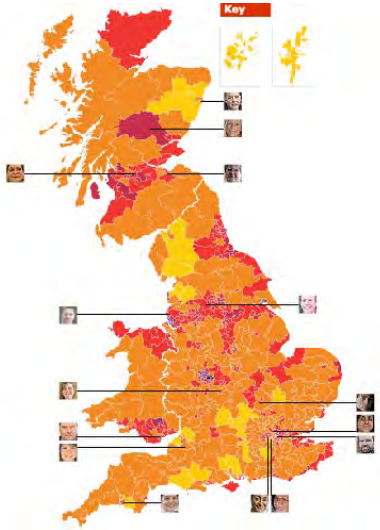


Martini Glass

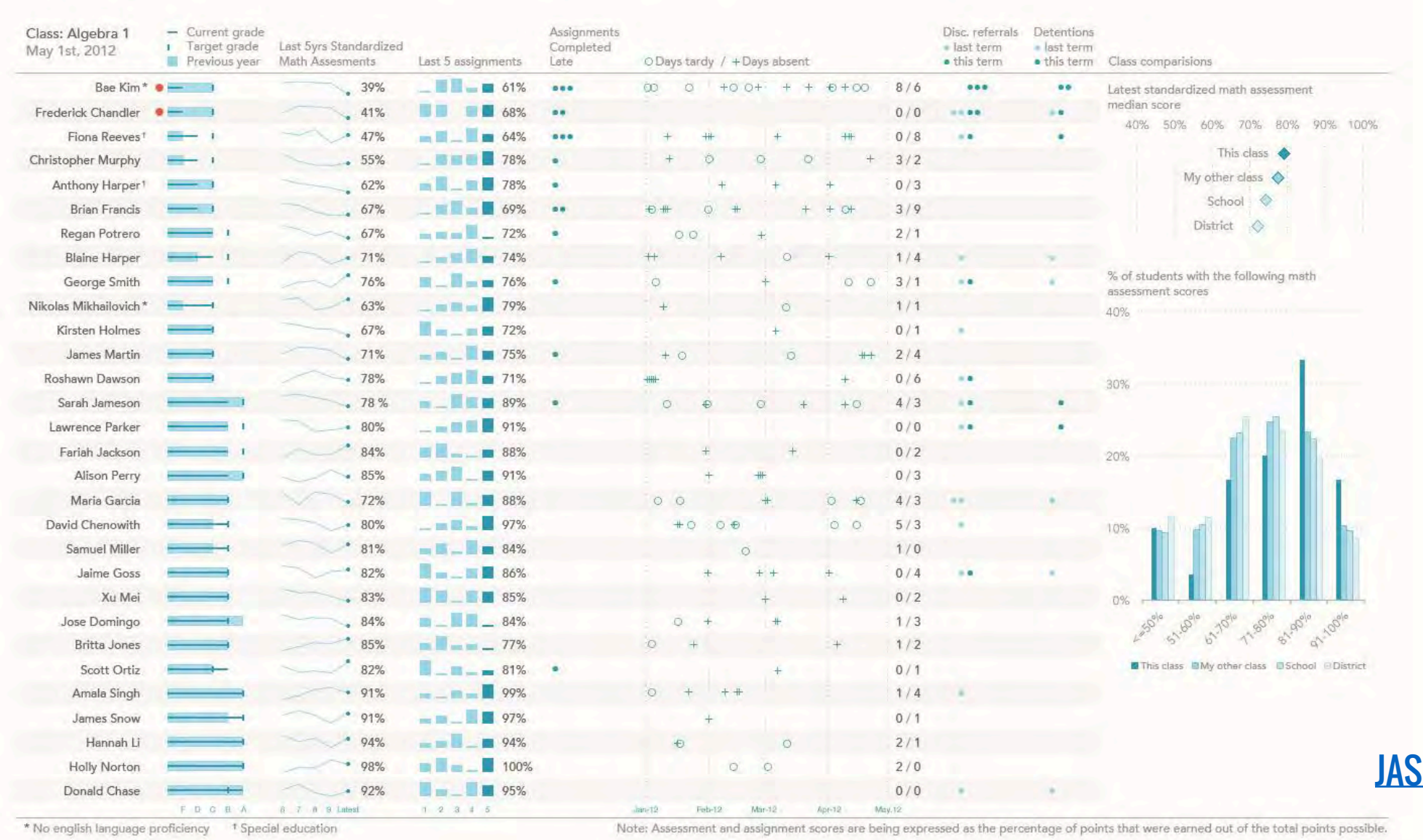


Drill-Down Story



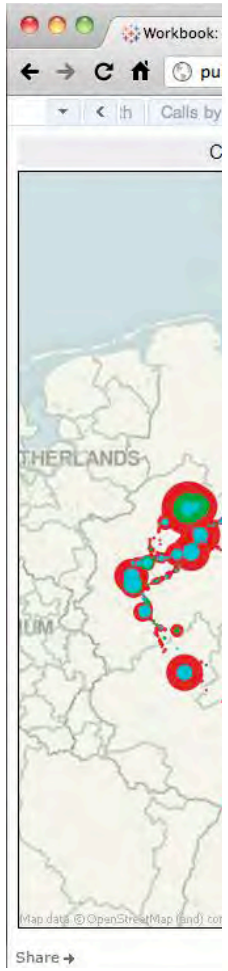


DESIGNING DASHBOARDS AND INTERACTIVES



JASON LOCKWOOD

TOOLS: TABLEAU



OF COURSE, IT'S NEVER THAT STRAIGHTFORWARD

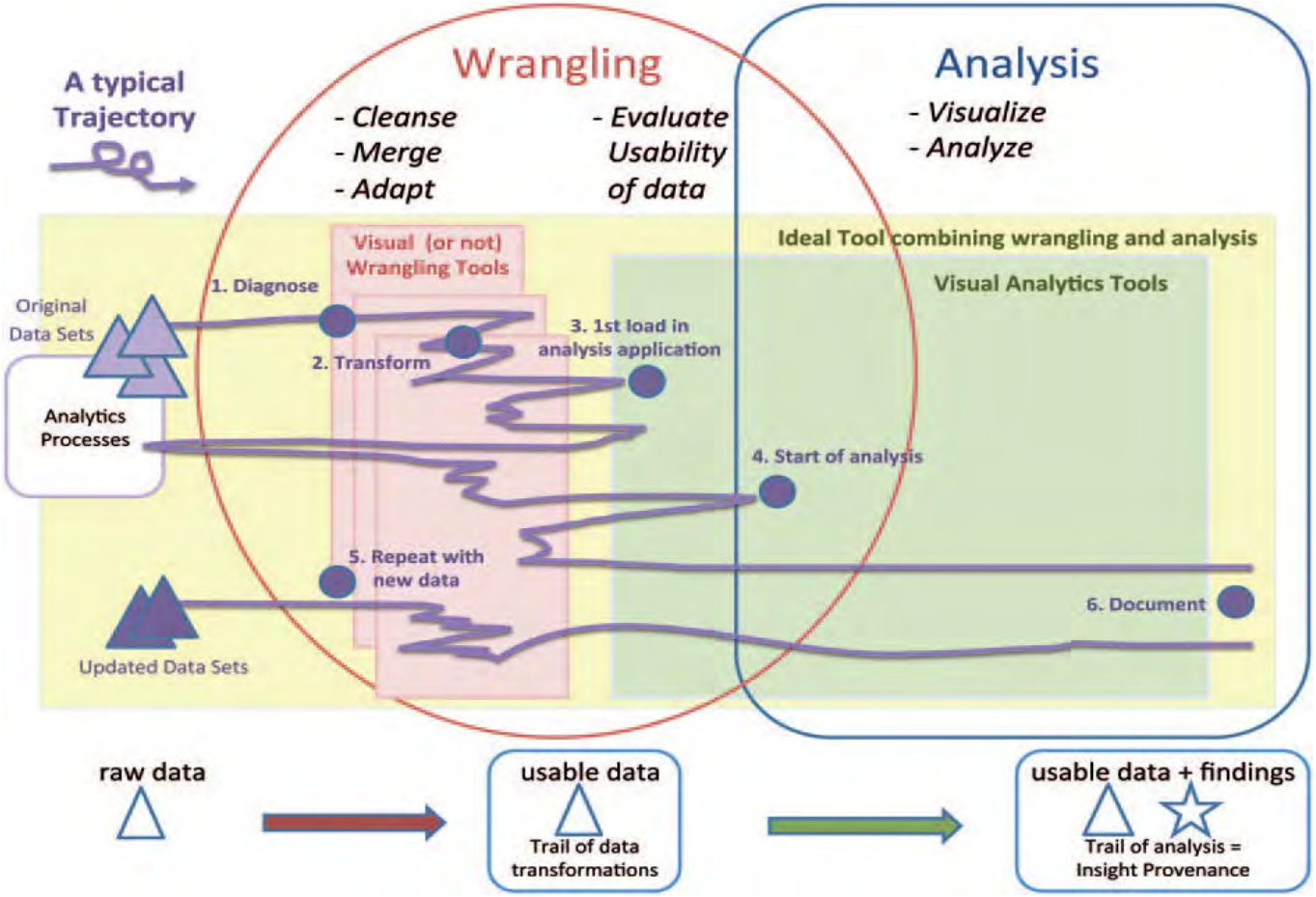
DATA COLLECTION → DATA CLEANING → VISUAL ANALYSIS → STATISTICAL ANALYSIS → PRESENTATION

(And you won't always see the same terminology.)

Ben Fry's Model

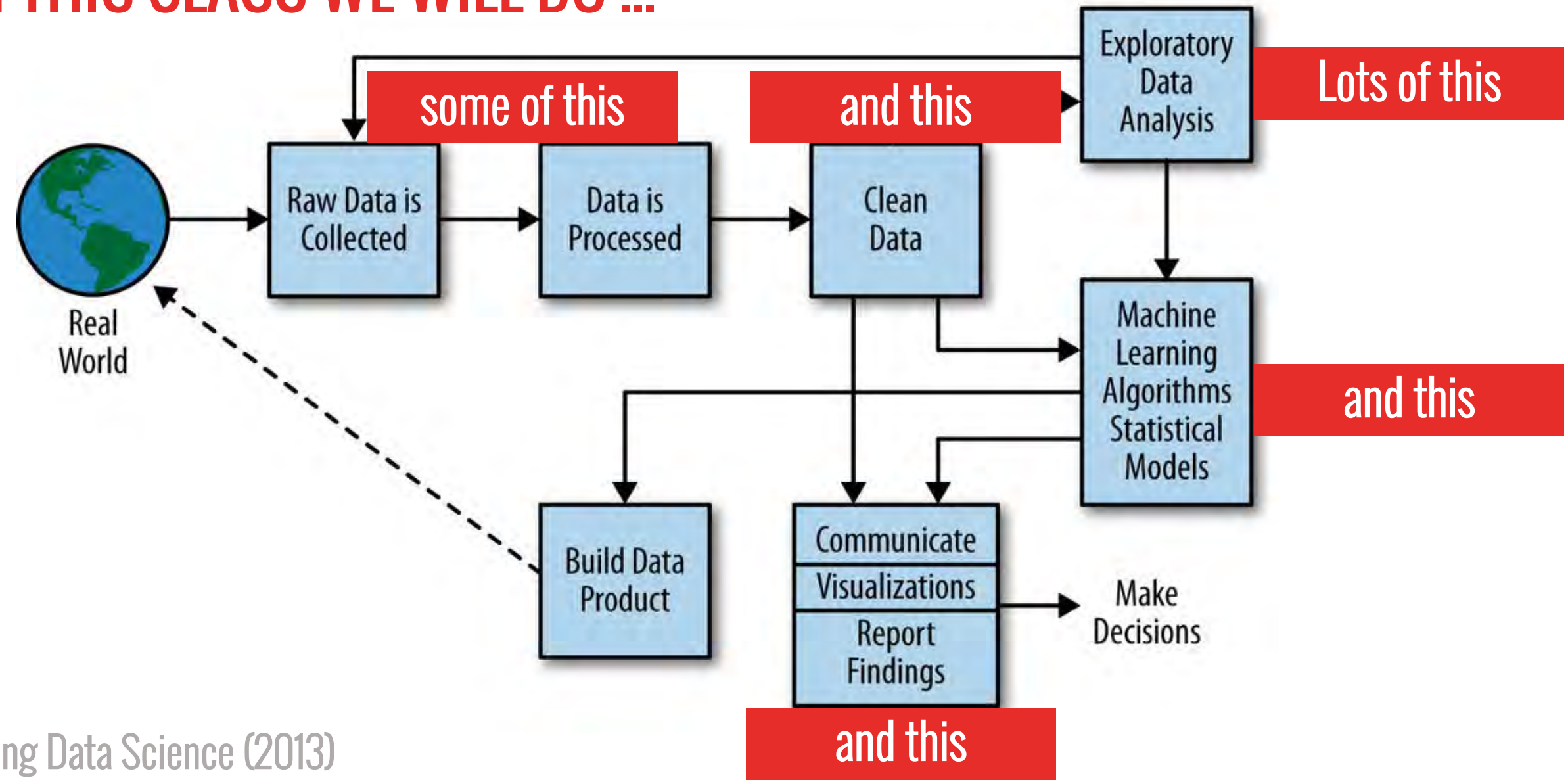
1. Acquire
2. Parse
3. Filter
4. Mine
5. Represent
6. Refine
7. Interact

ANALYSIS TRAJECTORIES



Kandel et al. 2011

IN THIS CLASS WE WILL DO ...



Doing Data Science (2013)
Cathy O'Neil & Rachel Schutt

...WITH AN EMPHASIS ON

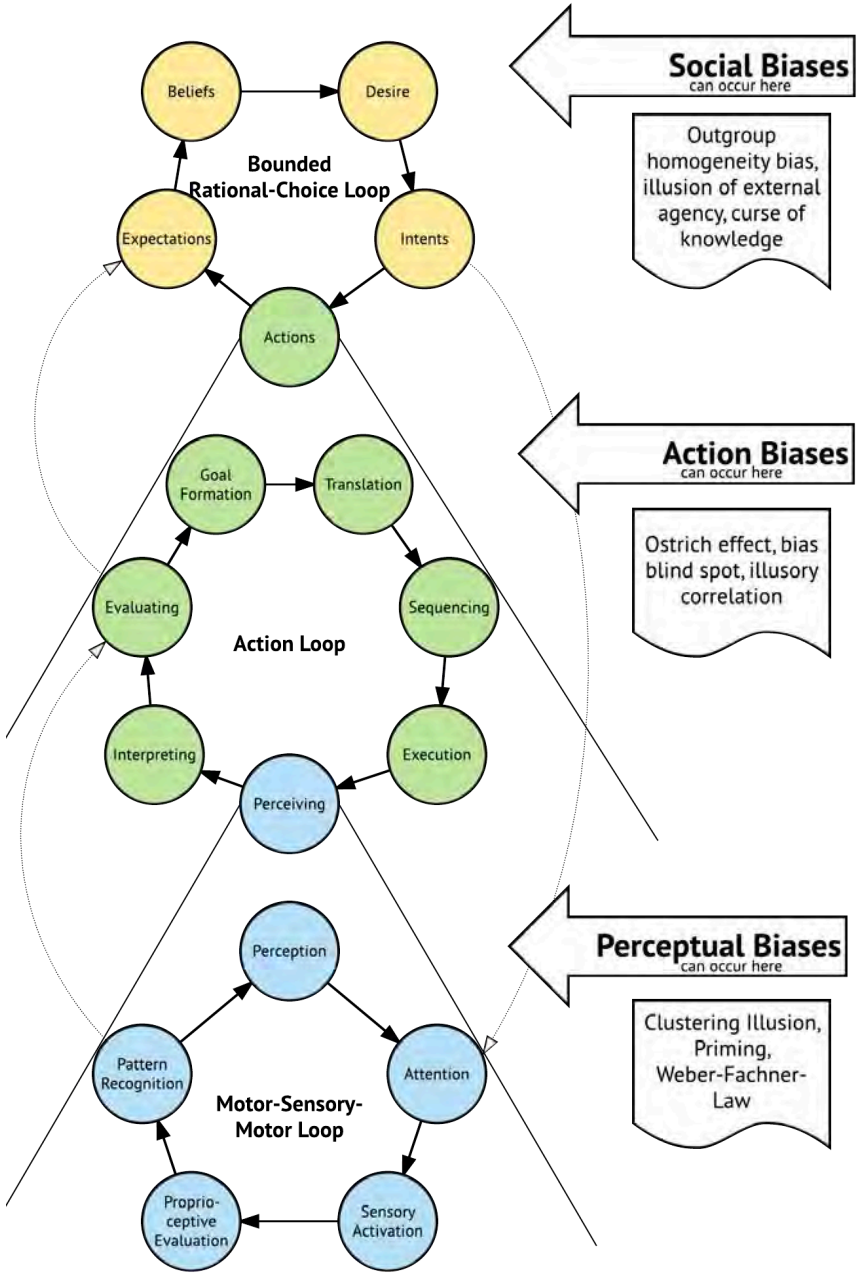
USING VISUALIZATIONS EFFECTIVELY

BUILDING COMFORT WITH VISUAL DATA ANALYSIS TOOLS

...and dealing with

COGNITIVE BIASES

Becoming more aware of all of the ways we can **fool ourselves** and **misinterpret data.**



VALDEZ, ZIEFLE, SEDLMAIR 2017

COURSE LOGISTICS

WHAT YOU'LL BE DOING

60% Practical In-Class Analysis Assignments (In Teams)

15% Analysis Case Study Presentation (Individually)

20% Final Project

ANALYSIS ASSIGNMENTS ("DATATHONS")

THURSDAY (MOST WEEKS)

Assigned at the beginning of class

↓

THURSDAY LECTURE

analysis time and peer/instructor feedback

↓

TUESDAY

10-minute group presentations and discussion at the start of class

COURSE FORMAT (LOOSELY)

TUESDAY

5:00pm-5:30pm "PRESENTATIONS"

5:30pm-6:15pm "LECTURE"

7:00pm-7:45pm "SOFTWARE TUTORIALS"

THURSDAY

5:00pm-6:00pm "LECTURE"

6:00pm-7:45pm "DATATHON"

NEW DATASET

TEAM ANALYSIS

PRESENTATIONS AND PEER-FEEDBACK

FINAL PROJECT

A more polished "**data story**" that builds on one of the in-class analyses.

DATA STORY CASE STUDIES

In-class presentations in which you **dissect**, **examine**, and **critique** real-world analyses and data stories.

ThePudding

Analyzing the Gender Representation of 34,476 Comic Book Characters

By **Amanda Shendruk**

Female characters appear in superhero comics less often than males — but when they *are* included, how are they depicted?

Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance

An open-source exploration of the city's neighborhoods, nightlife, airport traffic, and more, through the lens of publicly available taxi and Uber data

The New York City Taxi & Limousine Commission has released a staggeringly detailed historical dataset covering over 1.1 billion individual taxi trips in the city from January 2009 through June 2015. Taken as a whole, the detailed trip-level data is more than just a vast list of taxi pickup and drop off coordinates: it's a story of New York. How bad is the rush hour traffic from Midtown to JFK? Where does the Bridge and Tunnel

THIS COURSE IS AN EXPERIMENT

There are many visualization and analysis topics we can choose to focus on!

I want feedback!

SOFTWARE

INSTALL BEFORE THURSDAY!

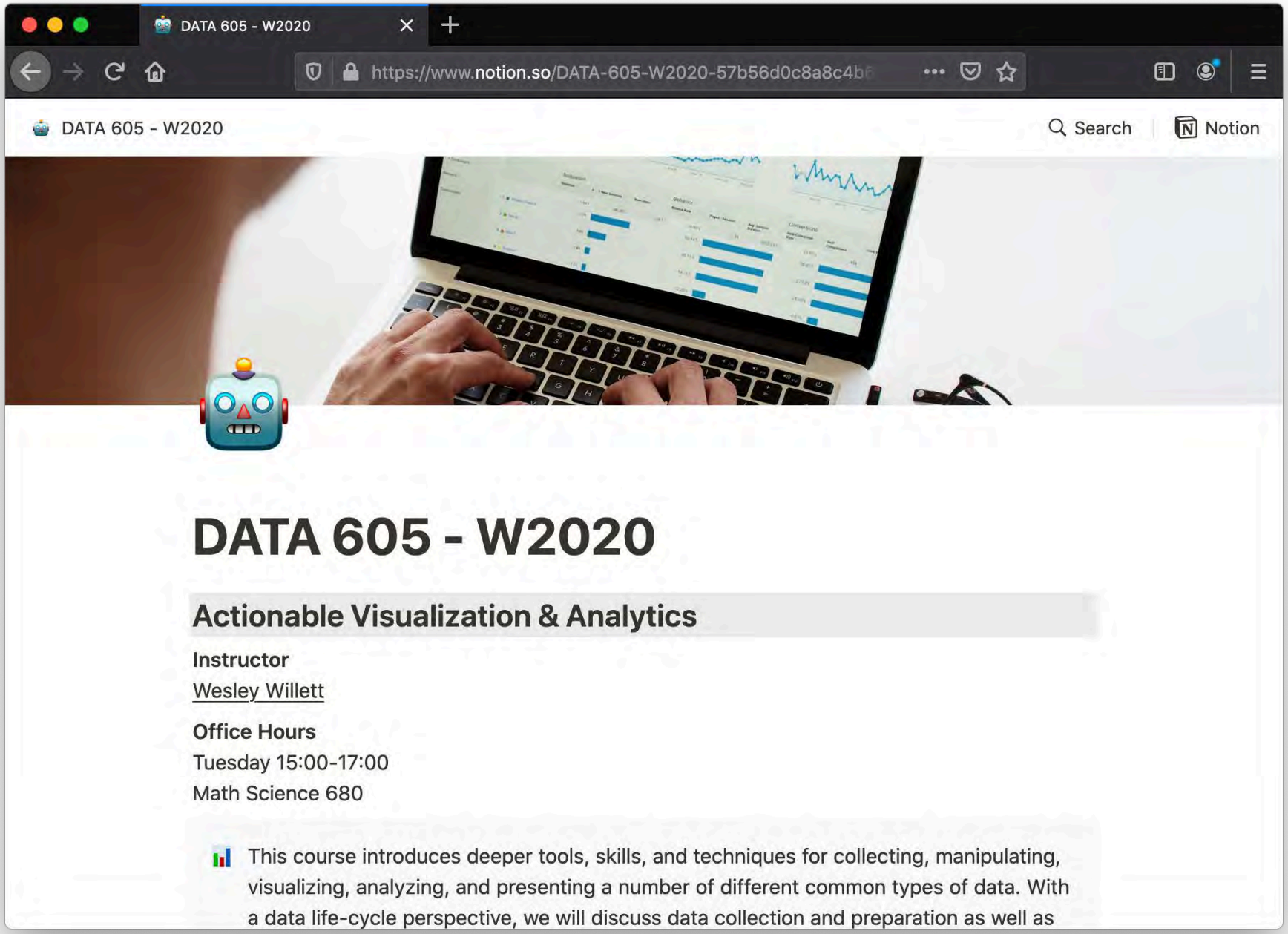
- EXCEL
  Basic Data Management/Manipulation
- TABLEAU
  Exploratory Analysis, Visualization, and Presentation

IN A FEW WEEKS

- JUPYTER / PYTHON
  Statistical Analysis
- GEPHI
  Network Analysis

CLASSROOM POLICIES

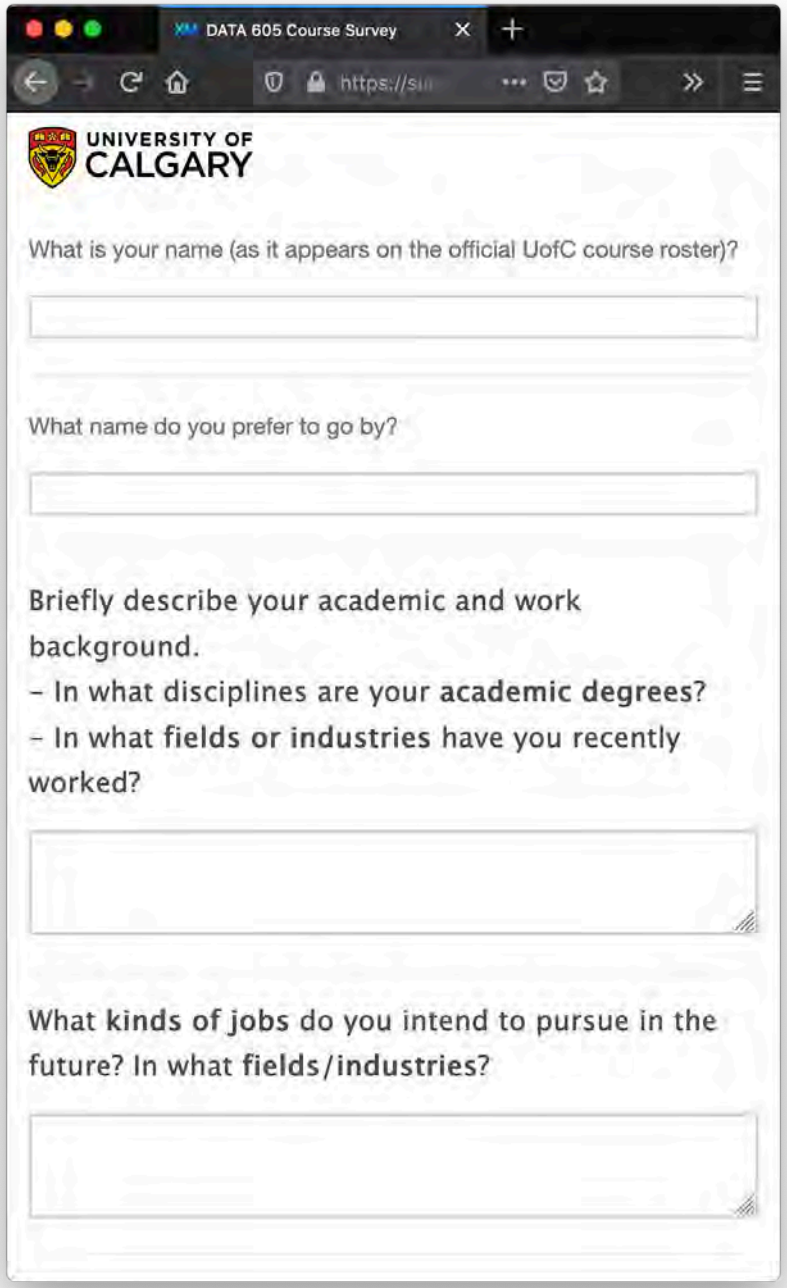
- Laptops are okay, **but stay on-task**.
- **You should** interrupt and ask questions.



https://tinyurl.com/
DATA605-W2020

COURSE SURVEY

Complete before you leave today!



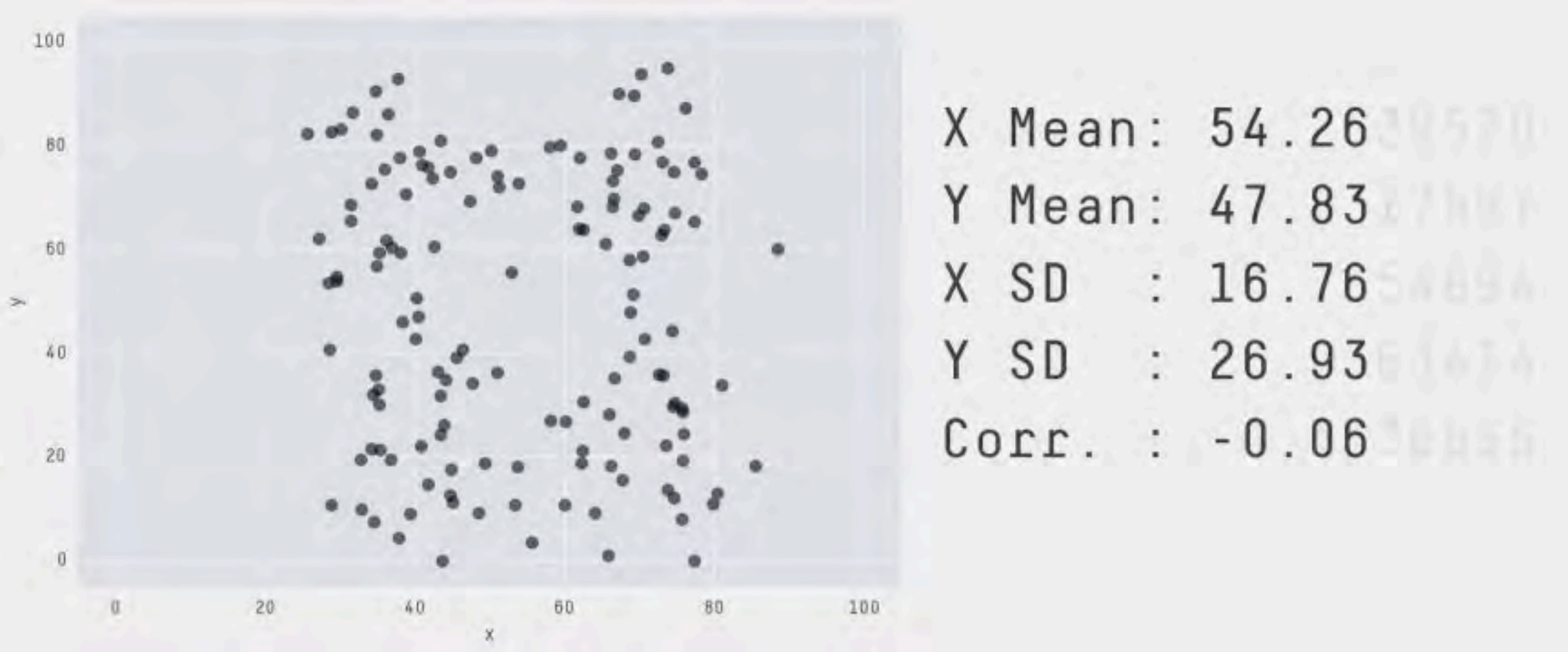
THURSDAY

- **Intro to Tableau**
  (Come with it installed.)

- Installation Instructions & Keys on Course Webpage

ACTIVITY

DIY Anscombe



@JustinMatejka / @albertocairo

1. FORM GROUPS OF 4
2. CREATE YOUR OWN ANSCOMBE-STYLE QUARTET

http://www.wjwillett.net/misc/drawmydata/ → goo.gl/PYwegs

Together, create 4 unique charts that all have (roughly) the same:

- Mean X
- Mean Y
- Standard Deviation X
- Standard Deviation Y
- Correlation

HINTS:

- Think about what each of the summary statistics shows.
- Create a few simple charts first, then try something more adventurous.

(When you're done – Keep your quartet of charts up to present. Also, save a screenshot.)