STAT 209
Data Computing and Visualization

February 5, 2018

Colin Reimer Dawson

Outline

"Data Science"

Intros

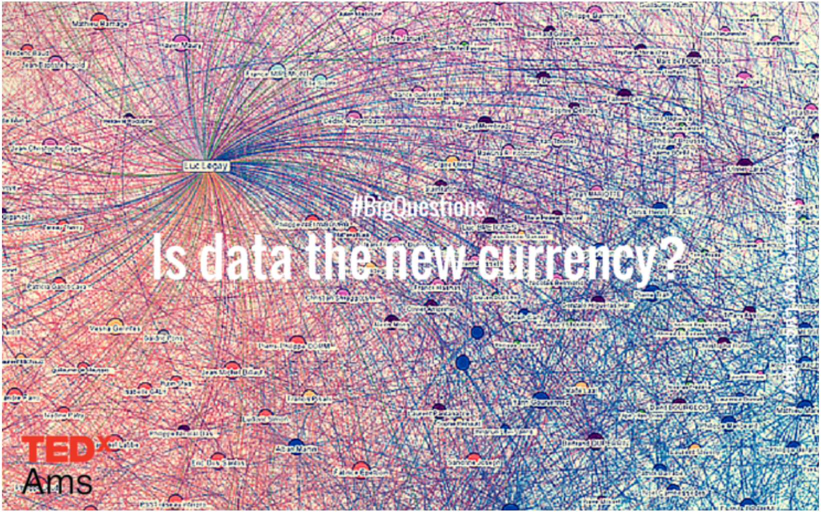
Some Terminology

Course Outline

Data is the new black

"DATA IS THE NEW GOLD"

S
DATA
IS THE NEW OIL
but do you have the resource to refine it?


#BigQuestions
Is data the new currency?
TEDx
Ams

‘DATA’ IS THE NEW CURRENCY

WHAT OPPORTUNITIES ARE YOU MISSING?

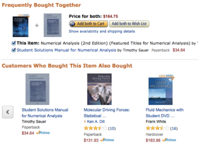
LOTAME®

Data is the New Currency to power Attribution

Prepared by **Evgeny Popov**

Some Cool Things you can do with data

Recommendation Systems



Frequently Bought Together

Price for both: $164.75

Add both to Cart | Add both to Wish List

Show availability and shipping details

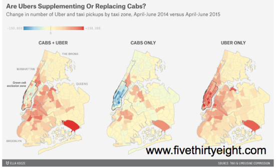
☑ **This item:** Numerical Analysis (2nd Edition) (Featured Titles for Numerical Analysis) by

☑ Student Solutions Manual for Numerical Analysis by Timothy Sauer Paperback $34.64

Customers Who Bought This Item Also Bought

Student Solutions Manual for Numerical Analysis	Molecular Driving Forces: Statistical ...	Fluid Mechanics with Student DVD ...
Timothy Sauer	› Ken A. Dill	Frank White
	★★★½☆ (10)	★★★½☆ (16)
Paperback	Paperback	Hardcover
$34.64 ✓Prime	$131.93 ✓Prime	$183.98 ✓Prime

Data-Driven Journalism

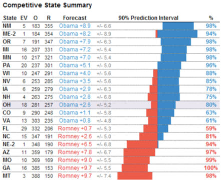


Are Ubers Supplementing Or Replacing Cabs?

Change in number of Uber and taxi pickups by taxi zone, April-June 2014 versus April-June 2015

www.fivethirtyeight.com

SOURCE: TAXI & LIMOUSINE COMMISSION

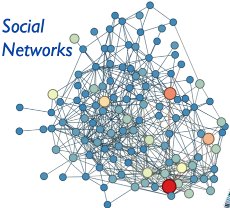


Competitive State Summary

State	EV	O	R	Forecast		90% Prediction Interval	%
NM	5	183	355	Obama +8.9	+/- 6.6		98%
ME-2	1	184	354	Obama +8.2	+/- 8.9		94%
OR	7	191	347	Obama +7.9	+/- 6.3		98%
MI	16	207	331	Obama +7.2	+/- 5.4		98%
MN	10	217	321	Obama +7.0	+/- 5.4		98%
PA	20	237	301	Obama +5.1	+/- 5.0		96%
WI	10	247	291	Obama +4.0	+/- 5.6		88%
NV	6	253	285	Obama +3.5	+/- 5.6		85%
IA	6	259	279	Obama +2.9	+/- 6.3		78%
NH	4	263	275	Obama +2.8	+/- 6.8		75%
OH	18	281	257	Obama +2.6	+/- 5.2		80%
CO	9	290	248	Obama +1.1	+/- 5.8		63%
VA	13	303	235	Obama +0.8	+/- 4.8		61%
FL	29	332	206	Romney +0.7	+/- 5.3		59%
NC	15	347	191	Romney +2.6	+/- 5.0		81%
NE-2	1	348	190	Romney +6.5	+/- 6.8		94%
AZ	11	359	179	Romney +7.8	+/- 6.7		97%
MO	10	369	169	Romney +9.0	+/- 5.5		99%
GA	16	385	153	Romney +9.7	+/- 5.5		100%
MT	3	388	150	Romney +9.7	+/- 7.4		98%

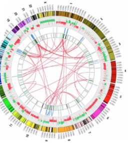
Political Science

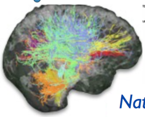
Social Networks



GIS / Development / Public Policy

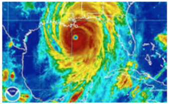
Sports





Natural Sciences

Meteorology / Climate Science

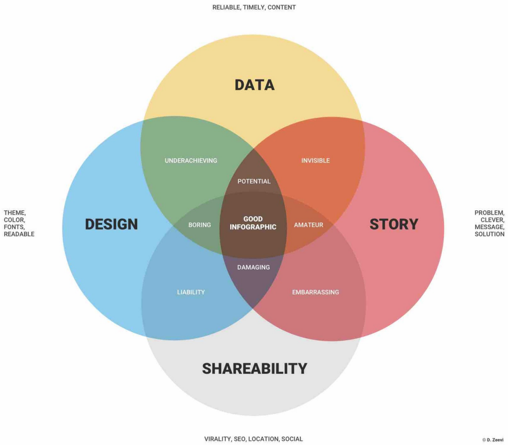


Finance

AMZN (Amazon.com, Inc.) Nasdaq GS
28-Jan-2011
Open 177.95 High 185.00 Low 166.90 Close 171.14 Volume 48.3M Chg -6.28 (-3.54%)
©StockCharts.com
RSI(14) 56.13
AMZN (Weekly) 171.14
MA(50) 144.63
MA(200) 96.26
Volume 48,315,680

Thanks to David Shuman at Macalester College for this slide

DATA VISUALIZATION



Brainstorm

What is the difference between “data” and “information”?

Cases

Cases When we collect data, we write down some measurements or characteristics of our **cases** — the individual “entities”, sometimes called “observational units”, that make up our dataset.

- The people in a survey or research study
- Plots of land in an agricultural experiment
- Days, in a weather dataset

Categorical vs. Quantitative Variables

For each case we record one or more **variables**. One of the most basic distinctions is between **categorical** (or "qualitative") and **quantitative** data.

Categorical: "Qualitative" variable that divides cases into groups

Quantitative: Measures something on a scale; arithmetic makes sense

Data Frames

A standard form for a dataset is a grid, called a **data frame**, where each row is a *case*, and each column is a *variable*.

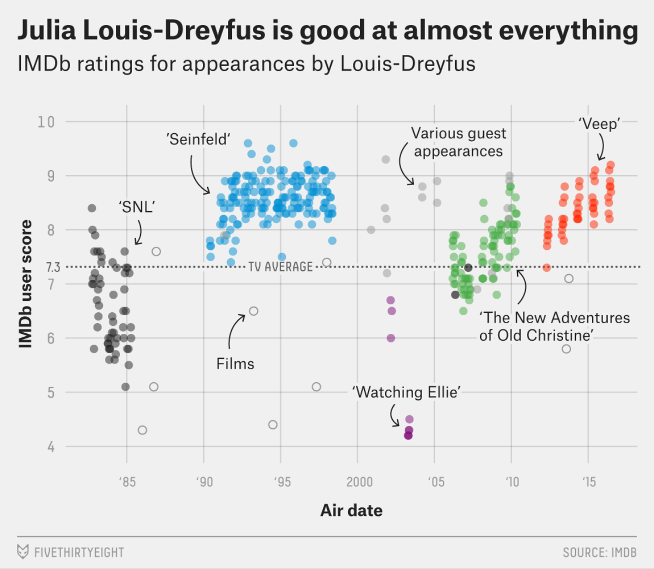
ID	Major	Height
1	Neuroscience	67
2	CS	71
	…	
21	Economics	64

Deconstructing Visualizations

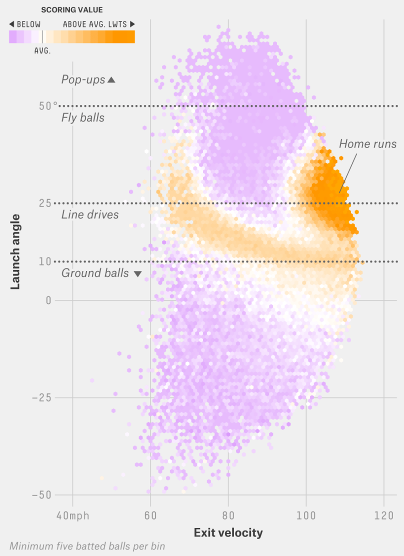
For each of the following visualizations:

1. What are the cases (think “rows” of a dataset)?
2. What variables are depicted (think “columns” of a dataset)?
3. What graphical element (position, color, etc.) is used to encode each variable?

Julia Louis-Dreyfuss



Baseball Hits



The sweet spot

Scoring value (LWTS) of batted balls based on launch angle and speed off the bat, 2015 MLB

Minimum five batted balls per bin

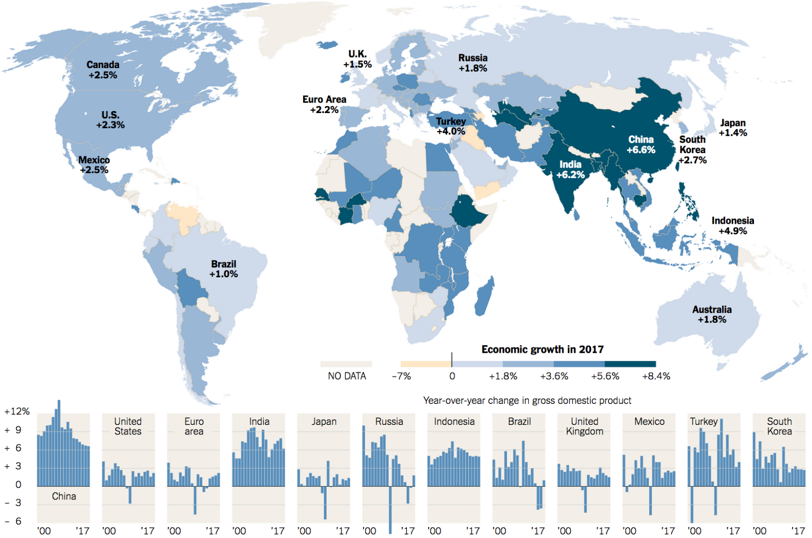
FIVETHIRTYEIGHT

SOURCE: BASEBALL SAVANT

Global Economic Growth

Growth Across the Globe

For the first time since the financial crisis a decade ago, all of the world's major economies are growing.



Some figures are estimates

Source: The Conference Board; Bureau of Labor Statistics | By Karl Russell

Course Outline

- Part I: Basic Visualization (about 3 weeks)
- Part II: Data “Wrangling” (about 4 weeks)
- Part III: Dealing with “large” datasets (about 3 weeks)
- Part IV: Visualizing data with complex structure (spatial and text data) (about 2 weeks)

On the web

- Course Website: `http://colindawson.net/stat209`
  - Syllabus, schedule, homework, slides, code, etc., there
- Blackboard: only for things that need a login/password protection
  - HW Solutions (when applicable)
  - Electronic submission of (some) assignments
- Slack: `stat209s2018.slack.com`, or download the app to computer/mobile device
  - Convenient one-stop place for all course-related electronic communication
- DataCamp
  - Many interactive tutorials to learn/practice computing tools
  - First couple of homework assignments there
- GitHub (later): Good way to track code changes/share code

Graded Components

Course grade based on:

- Homework sets (15%)
- Labs (15%; completion/effort only)
- Four group visualization projects, one for each major "unit" (15% each)
- Participation/engagement (10%)
- Scheduled final exam day used for presentations of project 4

See the syllabus for Honor Code guidelines

Structure of Class

- About half lecture/full group activities, half labs in small groups
  - Some labs are guided exercises to learn new tools
  - Final lab in each unit will ask you to "reverse engineer" a specific visualization from the web
- First three projects:
  - In class "workshop day", to work out the kinks together
  - In class "short presentation day", for feedback on a draft
  - Final writeup due a few days after that
- Project 4: longer presentation of a polished version during finals week

This Week

- First homework is to complete a DataCamp chapter on using RStudio
- Wednesday: Lab 1 to get comfortable with R/RStudio
- Friday: Start on basic elements of visualization