CS102: Big Data

Tools and Techniques, Discoveries and Pitfalls

Spring 2017

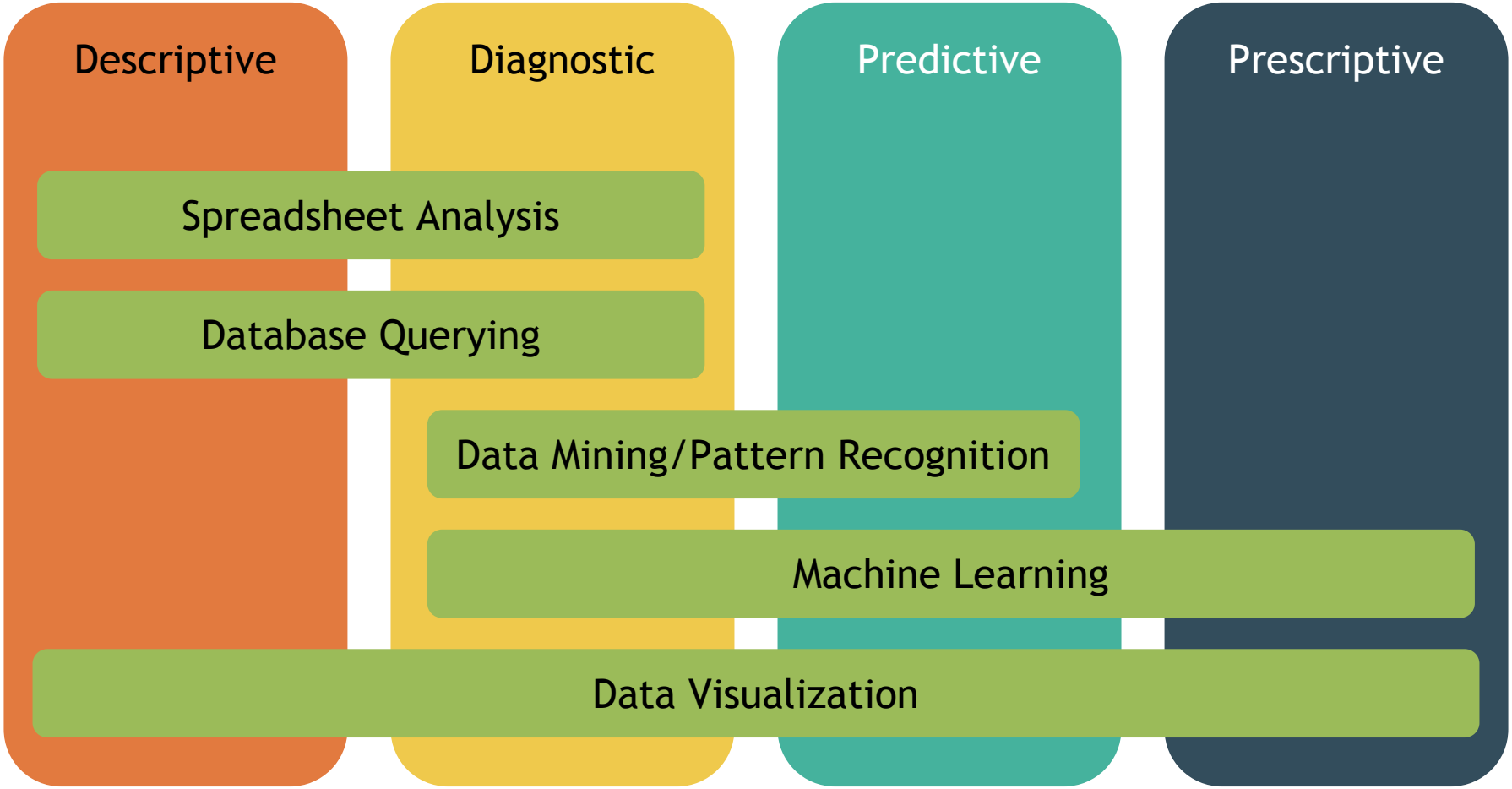
Ethan Chan, Lisa Wang

Lecture 2: Spreadsheet Data Analysis and Visualization

Announcements

- Make a name card for next class!
- Class Attendance on Google Sheet
  - Email us if you can't access the link
- Homework 0 is due this Sun Apr 9
- Homework 1 released, due Sun Apr 16
  - Have fun on Kickstarter Dataset!
  - 5000 kickstarter projects from 2017

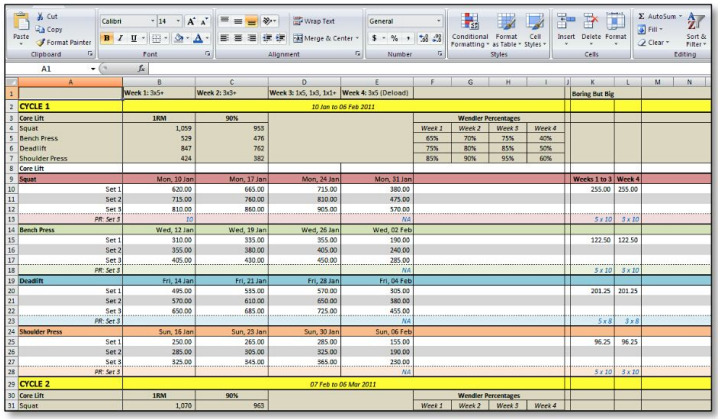
Tools & Techniques



Why Spreadsheets?

Spreadsheets

Most familiar use is for data presentation



Formulas handy for sales, budgets and other numeric summaries

I4 =SUM(B4:H4)

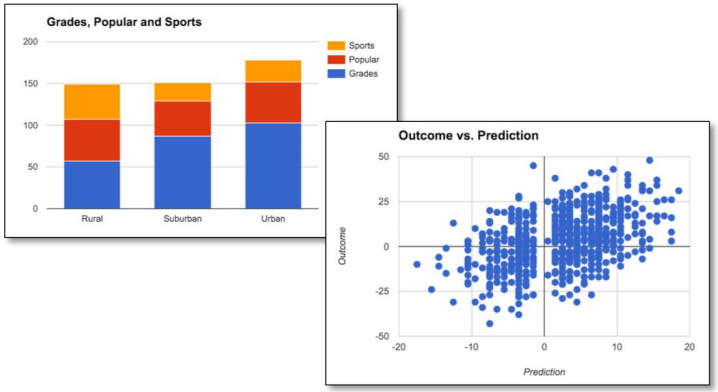
	A	B	C	D	E	F	G	H	I
1									
2									
3		Jan	Feb	Mar	Apr	May	Jun	Jul	Total
4	Person 1	620	768	251	811	664	304	27	**3445**
5	Person 2	1	928	595	214	317	470	360	**2885**
6	Person 3	707	481	849	255	548	550	518	**3908**
7	Person 4	235	110	357	730	739	265	36	**2472**
8	Person 5	610	508	353	952	643	16	738	**3820**
9	Person 6	425	648	740	162	865	332	786	**3958**
10	Person 7	695	751	111	675	736	407	6	**3381**
11	Person 8	326	449	80	612	779	1000	341	**3587**
12	Person 9	981	540	509	860	92	631	900	**4513**
13	**Total**	**4600**	**5183**	**3845**	**5271**	**5383**	**3975**	**3712**	**31969**
14									

Spreadsheets

Convenient tool for analysis of structured data

Data Visualization

	A	B	C	D	E	F	G
1	Year	Week	Home	HomeScore	Away	AwayScore	Prediction
2	1998	1	Green_Bay	38	Detroit	19	9.5
3	1998	1	Chicago	23	Jacksonville	24	-8.5
4	1998	1	Minnesota	31	Tampa Bay	7	3.5
5	1998	1	St._Louis	17	New_Orleans	24	3.5
6	1998	1	Cincinnati	14	Tennessee	23	1.5
7	1998	1	Baltimore	13	Pittsburgh	20	-3.5
8	1998	1	Carolina	14	Atlanta	19	4.5
9	1998	1	NY_Giants	31	Washington	24	2.5
10	1998	1	Philadelphia	0	Seattle	38	-3.5
11	1998	1	San_Diego	16	Buffalo	14	1.5
12	1998	1	San_Francisco	36	NY_Jets	30	7.5
13	1998	1	Dallas	38	Arizona	10	5.5
14	1998	1	Indianapolis	15	Miami	24	-3.5
15	1998	1	Kansas_City	28	Oakland	8	7.5
16	1998	1	Denver	27	New_England	21	7.5
17	1998	2	Tennessee	7	San_Diego	13	7.5
18	1998	2	Green_Bay	23	Tampa Bay	15	7.5
19	1998	2	New_Orleans	19	Carolina	14	-3.5
20	1998	2	St._Louis	31	Minnesota	38	-7.5
21	1998	2	Miami	13	Buffalo	7	7.5
22	1998	2	Jacksonville	21	Kansas_City	16	1.5
23	1998	2	NY_Jets	10	Baltimore	24	3.5
24	1998	2	Pittsburgh	17	Chicago	12	11.5
25	1998	2	Atlanta	17	Philadelphia	12	8.5
26	1998	2	Detroit	28	Cincinnati	34	6.5
27	1998	2	Oakland	20	NY_Giants	17	1.5
28	1998	2	Seattle	33	Arizona	14	7.5
29	1998	2	Denver	42	Dallas	23	7.5
30	1998	2	New_England	29	Indianapolis	0	8.5
31	1998	2	Washington	10	San_Francisco	45	-4.5
32	1998	3	Kansas_City	23	San_Diego	7	9.5
33	1998	3	Minnesota	29	Detroit	6	5.5
34	1998	3	Buffalo	33	St._Louis	34	4.5
35	1998	3	Cincinnati	6	Green_Bay	13	-7.5
36	1998	3	Miami	21	Pittsburgh	0	1.5



Spreadsheets

A surprisingly large fraction of the world's structured data is managed and manipulated in spreadsheets

Microsoft Excel is dominant tool

- Many features
- Proprietary and expensive

Google Sheets

- Open and free
- Fewer features, but catching up

"No, Smith, that's NOT why they're called 'spreadsheets'."

Spreadsheets

Learning Goals

- Comfortably analyze data with spreadsheets
  - Basic Spreadsheet Operations
  - Basic Data Operations
  - Pivot Tables
- Visualize Data w Google Sheets & rawgraphs.io
- Work on real data from Yelp!

Basic Spreadsheet Operations

- Importing Data
  - **.csv (comma seperated values)**
  - .tsv (tab seperated values)
- Exporting
- Inserting Rows or Columns
- Formulas

Basic Data Operations

- Sorting
  - What are the top 10 ratings?
- Filtering
  - Find all reviews with ratings greater than 2.
- Aggregration
  - Find the highest rating
- Aggregration with Condition
  - Number of reviews above 3 stars
- Grouped Aggregation
  - Find the average rating of each Restaurant
- Joining
  - Merging 2 spreadsheets [Restaurants, Reviews]

Limitations of Spreadsheets

Limitations

Data type

- Only on structured data

Data size

- Google sheets: 400,000 cells

Mechanics

- Header rows, empty cells, strange behaviors, ...

Some analyses are difficult

- E.g., 2 restaurants closest to each other (easy in SQL)

Traceability

Tracibility

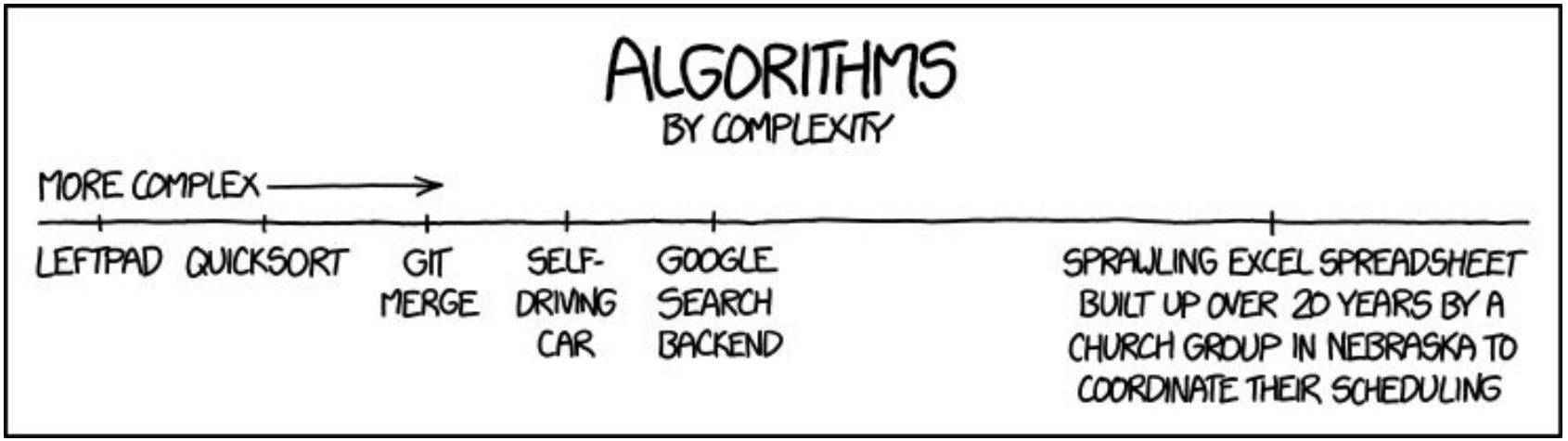
"There it is! I've isolated the origin of the firm's demise."

Hands On Class

Today is a hands on class!

1. Open up a new spreadsheet in Google Sheets
2. Download datasets from https://web.stanford.edu/class/cs102/datasets.htm
   - *YelpRestaurantsSample.csv*
   - *YelpReviewsSample.csv*
3. File > Import > Upload > Select "YelpReviewsSample.csv" > Insert new sheet(s)
4. Repeat (3) for "YelpRestaurantsSample.csv"
5. Let's go!

Note: Lecture slides + notes posted on the website will go through all that the steps that are covered today.


ALGORITHMS
BY COMPLEXITY
MORE COMPLEX
LEFTPAD
QUICKSORT
GIT MERGE
SELF-DRIVING CAR
GOOGLE SEARCH BACKEND
SPRAWLING EXCEL SPREADSHEET BUILT UP OVER 20 YEARS BY A CHURCH GROUP IN NEBRASKA TO COORDINATE THEIR SCHEDULING


THE MBA GUY
I PUT TOGETHER A SPREADSHEET THAT MIGHT INTEREST YOU.
www.dilbert.com scottadams@aol.com
OW! OW! IT'S SO BORING, IT HURTS MY HEAD!
MY BRAIN IS TRYING TO ESCAPE THROUGH MY EAR!
I GET THIS A LOT.
©2009 Scott Adams, Inc./Dist. by UFS, Inc.

Please give us feedback here:
http://bit.ly/cs102_feedback