Day 1

Zen of Data

The Data Bootcamp | <DATE GOES HERE>

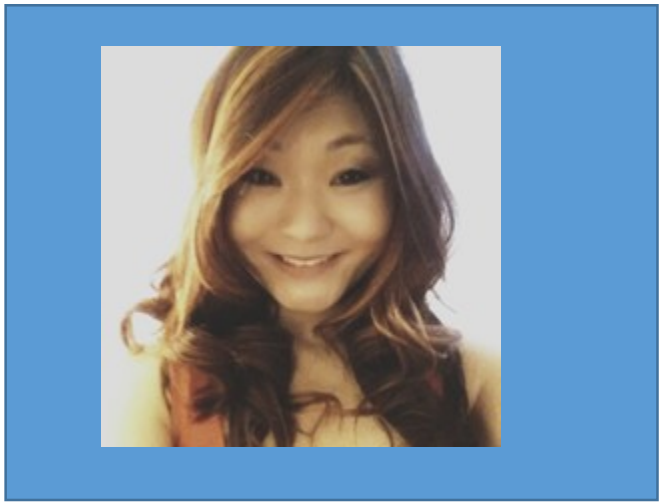
Quick Introductions! (30 seconds)

- Name
- Where From?
- Background (Career, Education, Interests)
- Why are you here?

Instructor = … ?

Candice Chen

- Principal Data Scientist @ Bank Of America
- Career Goal: Use Data to answer Business Questions
- Over 10 years data wrangling, analytics, visualization & data modeling



- Start my career as web developer
- 7+ Health Care
- 4-5 years in Big Data and Machine Learning in Entertainment & Banking

Instructor = … ?

Fun time

- Swimming
- Hotpot
- Reading
  - HBR
  - Economics
  - Fav book
    - Richard Bach: Illusions

Fun Projects

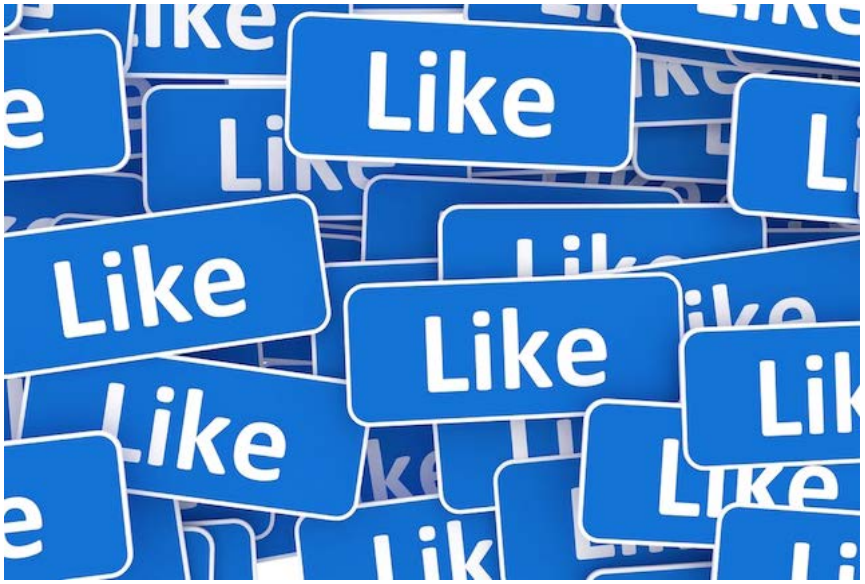
- Lead team to develop analytics as a platform to service Health Plan to streamline medical record retrieval and clinical data analytics such as United, Anthem, Aetna
- In entertainment, increased 3% play rate by optimizing search result, re-design recommendation engine and other advanced analytics and insight
- Banking industry, deliver quantitative insight and assure mortgage risk model quality to pass CCAR regulations

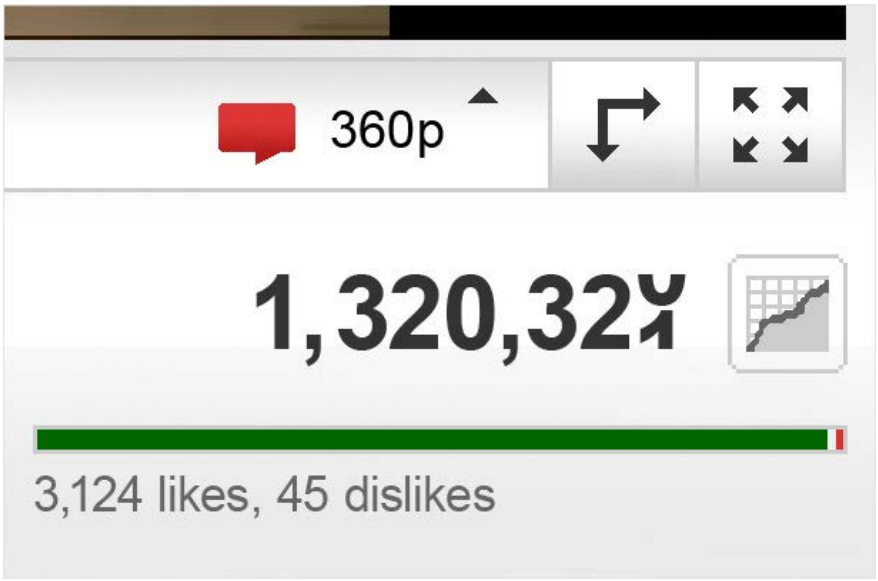
The Rise of Data

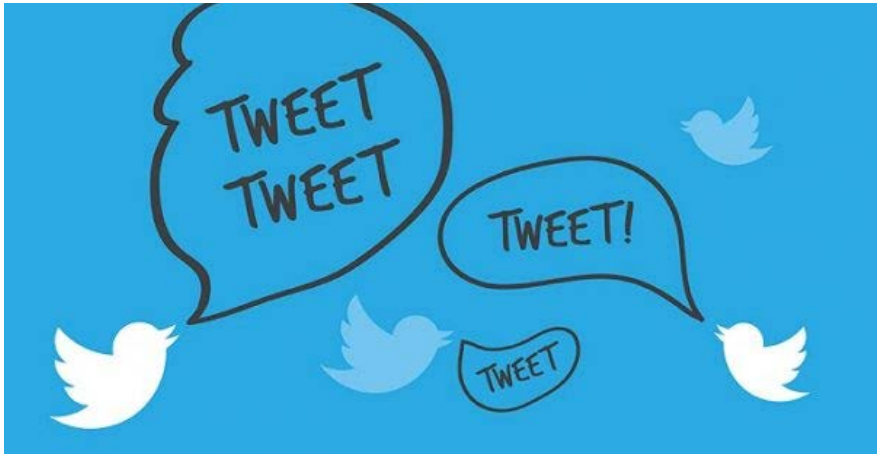
Question:

Why is “data” such a hot skill these days?

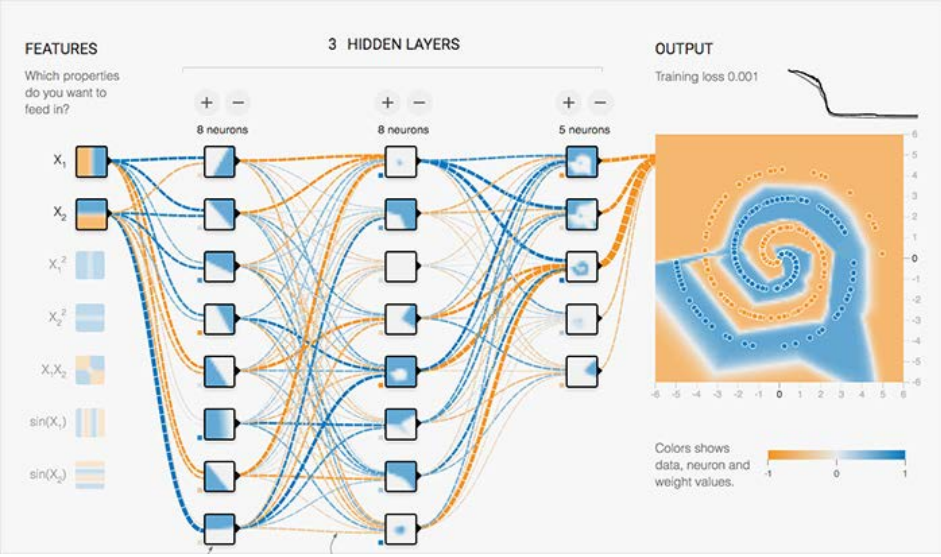
1. Explosive Growth in “Digitized” Data (Creation)

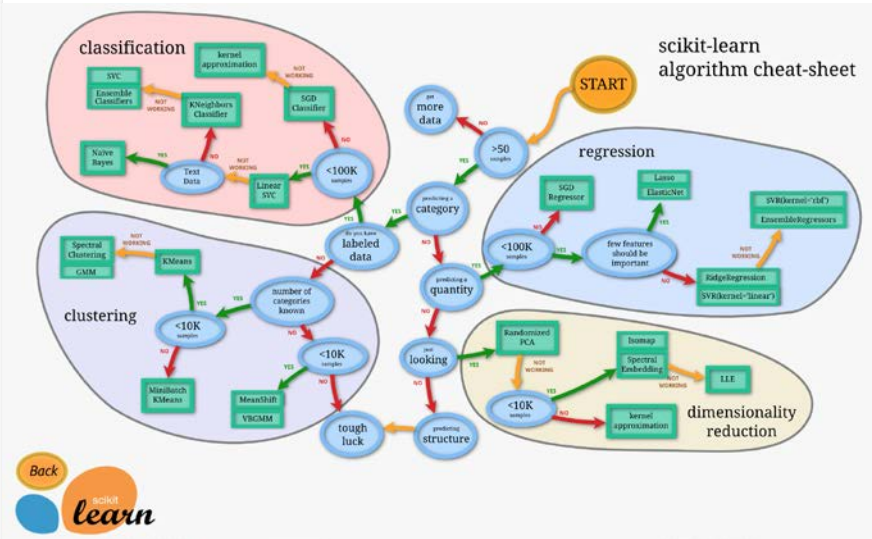


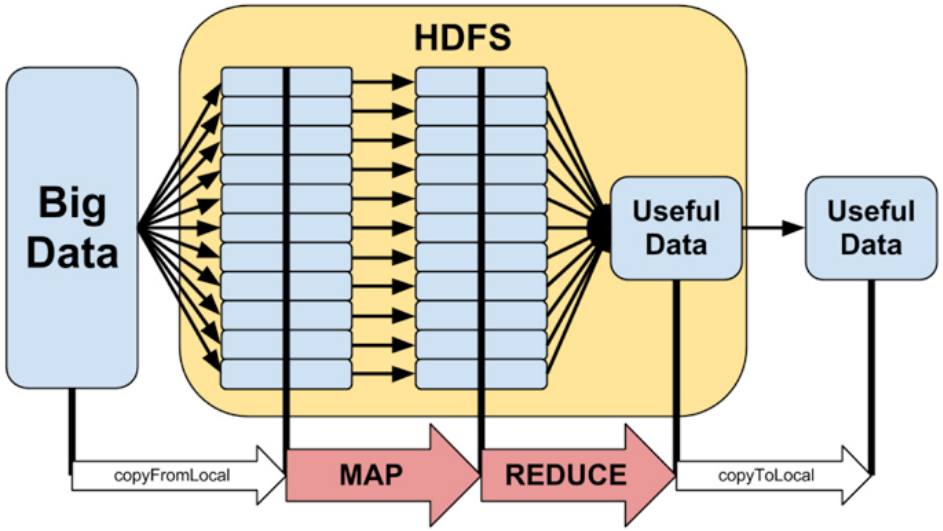


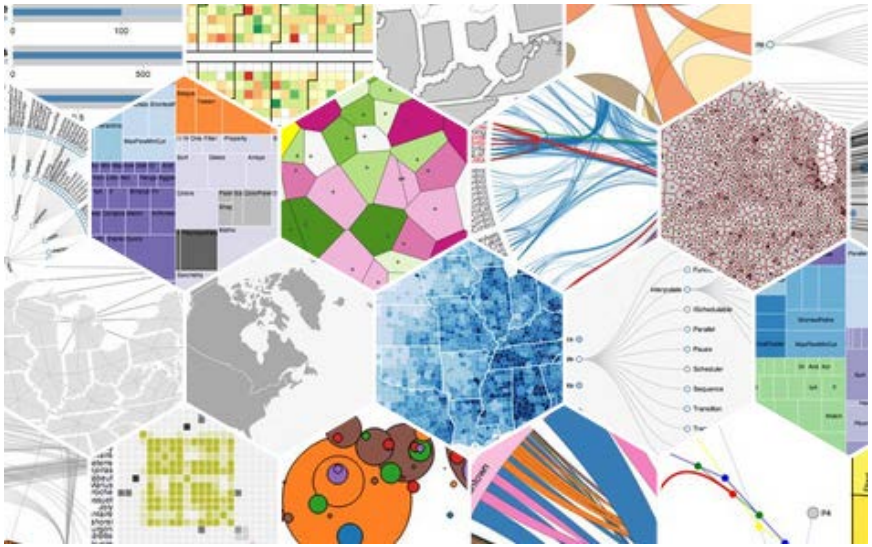


2. Explosive Growth in Analytic Tools (Synthesis)

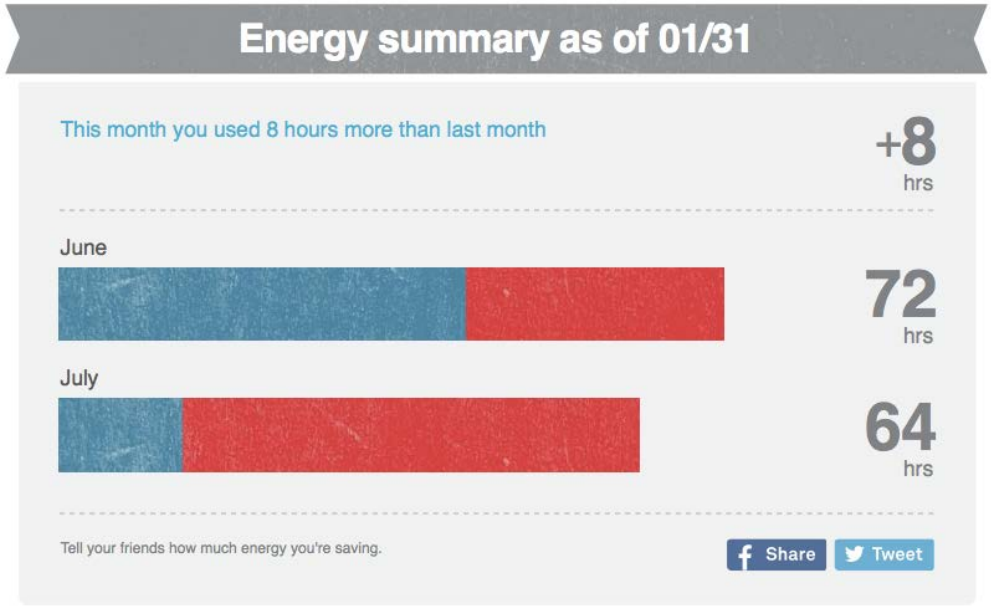


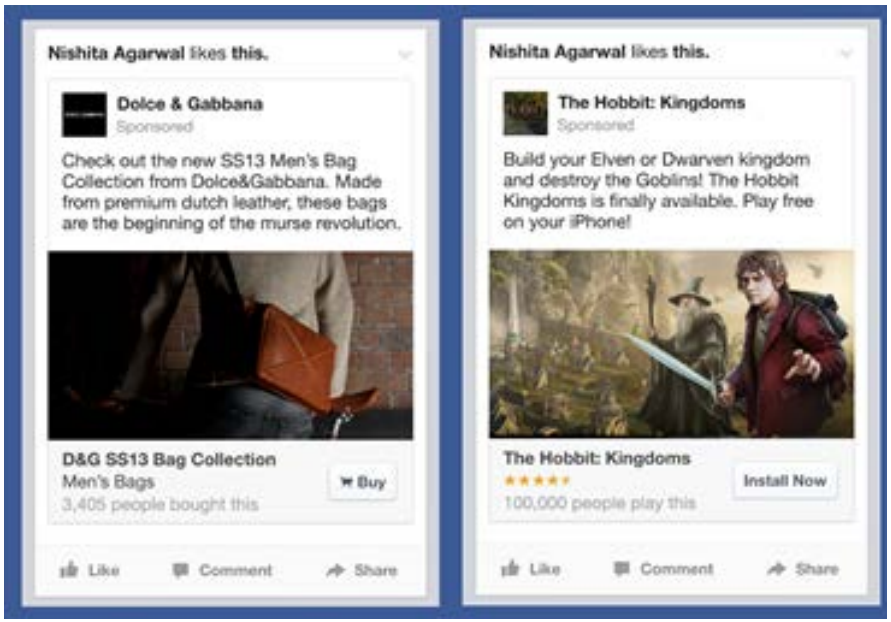






3. Accelerating Search for Actionable Insight (Value)





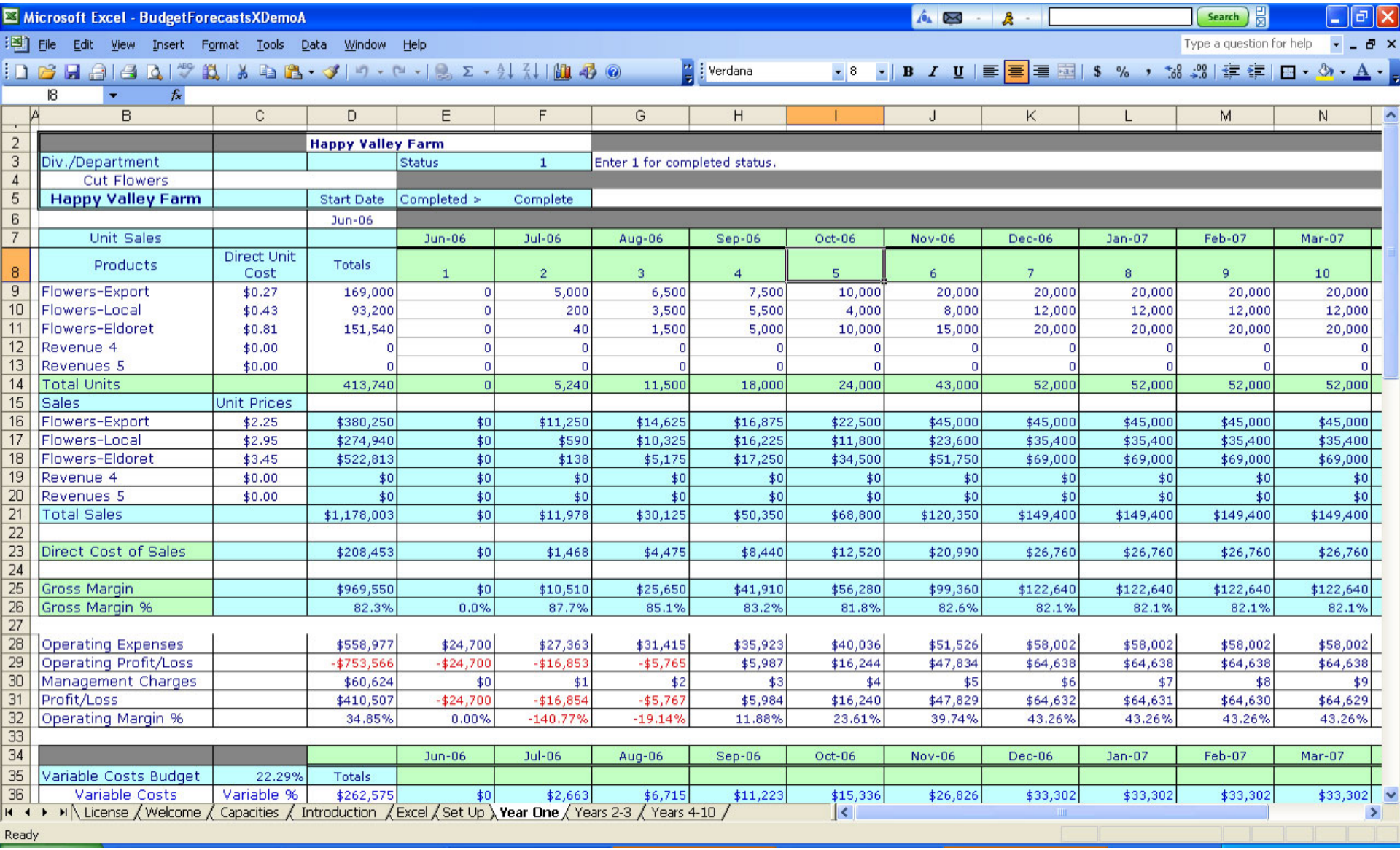
Data Means What?

So… What is **data science**?

First Thoughts…

CI For	Sample Statistic	Margin of Error	Use When
Population mean (μ)	$\bar{x}$	$\pm z^* \frac{\sigma}{\sqrt{n}}$	X is normal, or $n \geq 30$; σ known
Population mean (μ)	$\bar{x}$	$\pm t^*_{n-1} \frac{s}{\sqrt{n}}$	$n < 30$, and/or σ unknown
Population proportion (p)	$\hat{p}$	$\pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$n\hat{p}$, $n(1-\hat{p}) \geq 10$
Difference of two population means ($\mu_1 - \mu_2$)	$\bar{x}_1 - \bar{x}_2$	$\pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	Both normal distributions or $n_1, n_2 \geq 30$; σ_1, σ_2 known
Difference of two population means $\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\pm t^*_{n_1+n_2-2} \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$	$n_1, n_2 < 30$; and/or $\sigma_1 = \sigma_2$ unknown

First Thoughts…



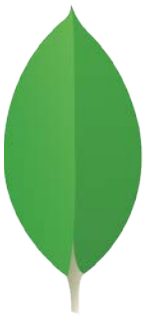
First Thoughts…

SQL

Employees Table

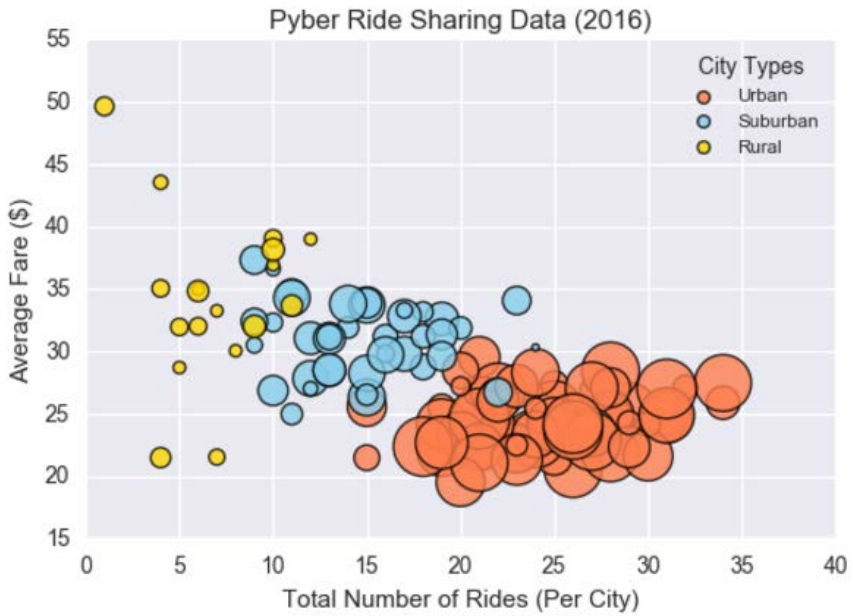
IdNum	LName	FName	JobCode	Salary	Phone
1876	CHIN	JACK	TA1	42400	212/588-5634
1114	GREENWALD	JANICE	ME3	38000	212/588-1092
1556	PENNINGTON	MICHAEL	ME1	29860	718/383-5681
1354	PARKER	MARY	FA3	65800	914/455-2337
1130	WOOD	DEBORAH	PT2	36514	212/587-0013

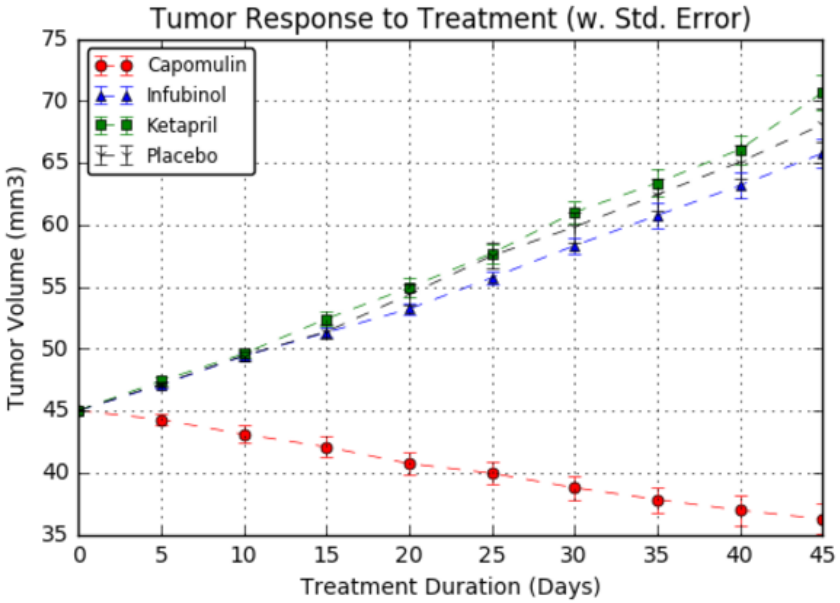
```
{
  "arguments" : { "number" : 10 },
  "url" : "http://localhost:8080/restty-tester/collection",
  "method" : "POST",
  "header" : {
    "Content-Type" : "application/json"
  },
  "body" : [
    {
      "id" : 0,
      "name" : "name 0",
      "description" : "description 0"
    },
    {
      "id" : 1,
      "name" : "name 1",
      "description" : "description 1"
    }
  ],
  "output" : "json"
}
```



mongoDB®

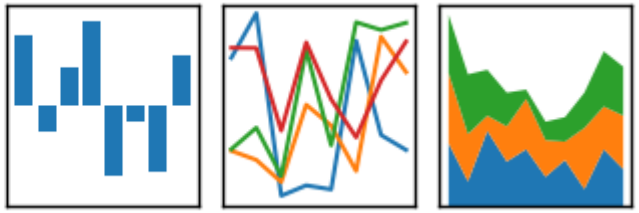
First Thoughts…





pandas

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



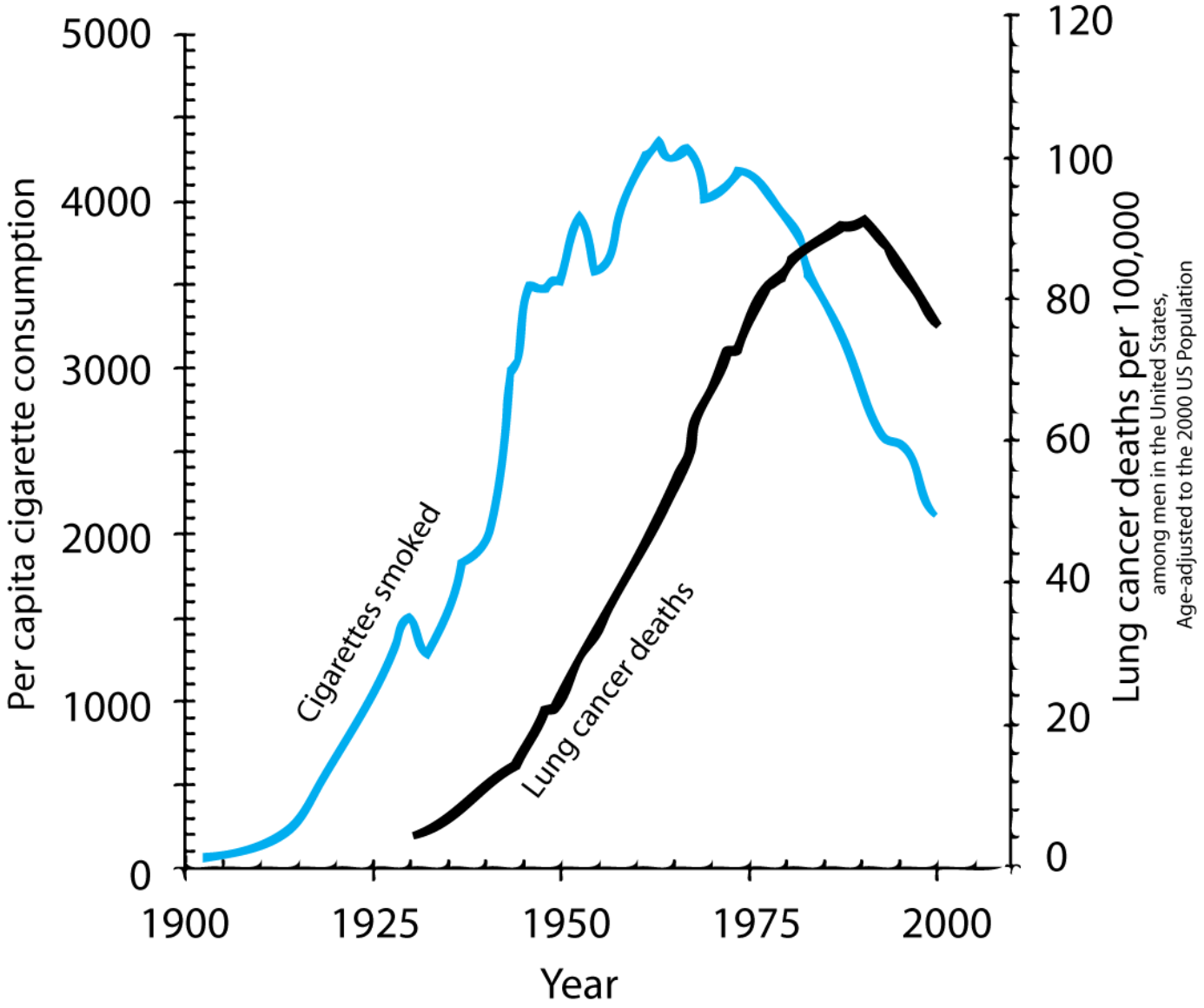
Data Science centers on two things:

Truth-Telling
&
Story Telling

Truth-Telling

Data As… Truth-Telling

Unearthing relationships



Data As… Truth-Telling

Making Predictions

Story-Telling

Data As… Story-Telling

Figure 1.2. Data: Congressional Budget Office

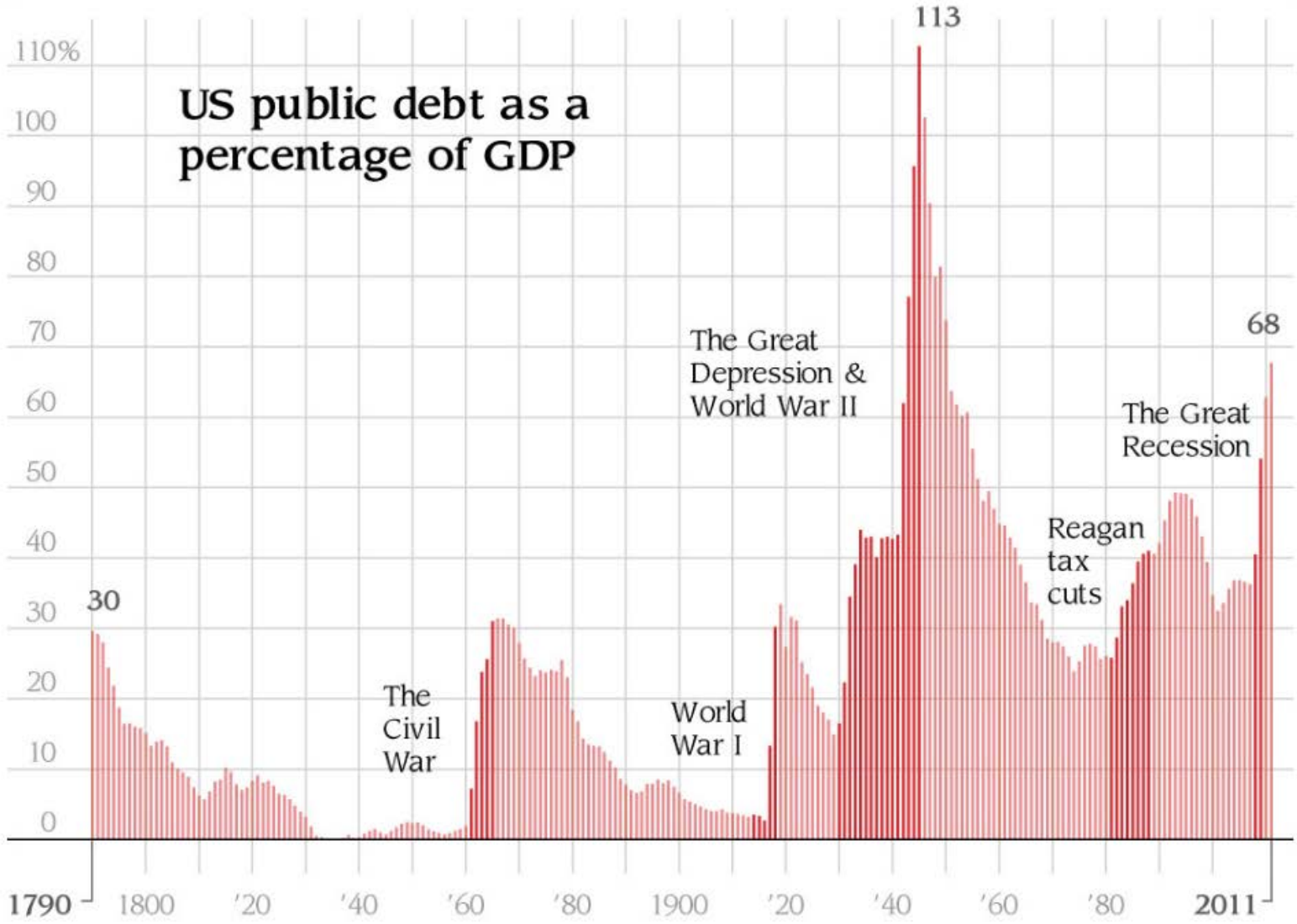
US debt as percentage of gross domestic product, 1790—2011

1790	29.6%	1835	0.0	1880	18.4	1925	21.6	1970	28.0
1791	29.2	1836	0.0	1881	16.8	1926	19.0	1971	28.1
1792	28.0	1837	0.2	1882	14.3	1927	18.0	1972	27.4
1793	24.4	1838	0.6	1883	13.5	1928	17.0	1973	26.0
1794	21.8	1839	0.2	1884	13.3	1929	14.9	1974	23.9
1795	18.7	1840	0.3	1885	13.2	1930	16.5	1975	25.3
1796	16.4	1841	0.8	1886	12.4	1931	22.3	1976	27.5
1797	16.5	1842	1.2	1887	11.2	1932	34.5	1977	27.8
1798	16.0	1843	1.5	1888	10.2	1933	39.1	1978	27.4
1799	15.8	1844	1.0	1889	8.6	1934	44.0	1979	25.6
1800	15.1	1845	0.7	1890	7.8	1935	42.9	1980	26.1
1801	13.3	1846	1.2	1891	7.0	1936	43.0	1981	25.8
1802	13.9	1847	1.7	1892	6.6	1937	40.1	1982	28.7
1803	14.1	1848	2.2	1893	6.8	1938	42.8	1983	33.1
1804	13.2	1849	2.5	1894	7.9	1939	43.0	1984	34.0
1805	10.9	1850	2.3	1895	7.9	1940	42.7	1985	36.4
1806	10.0	1851	2.4	1896	8.5	1941	43.3	1986	39.5

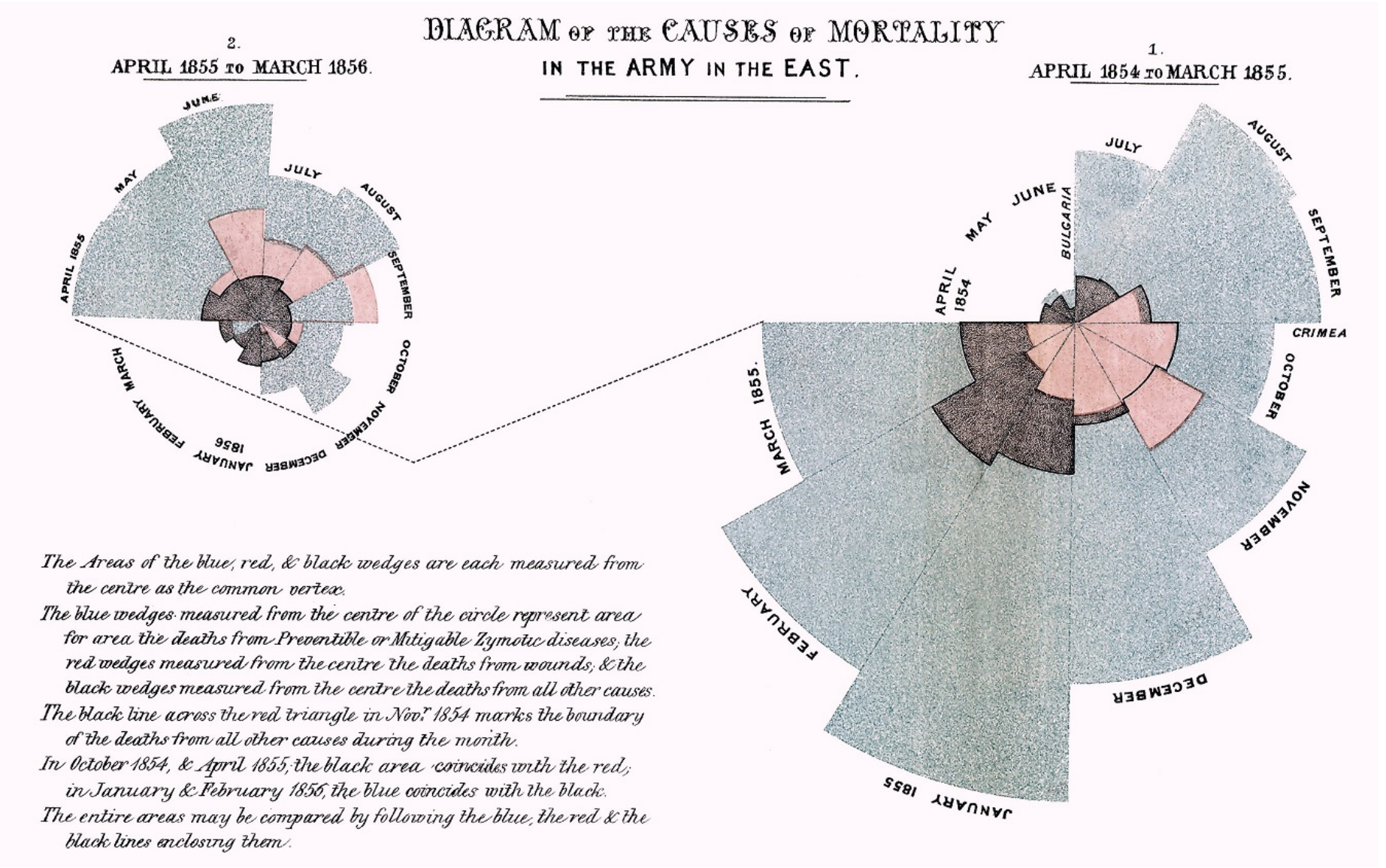
The Great Depression & World War II

Reagan tax cuts

Data As… Story-Telling

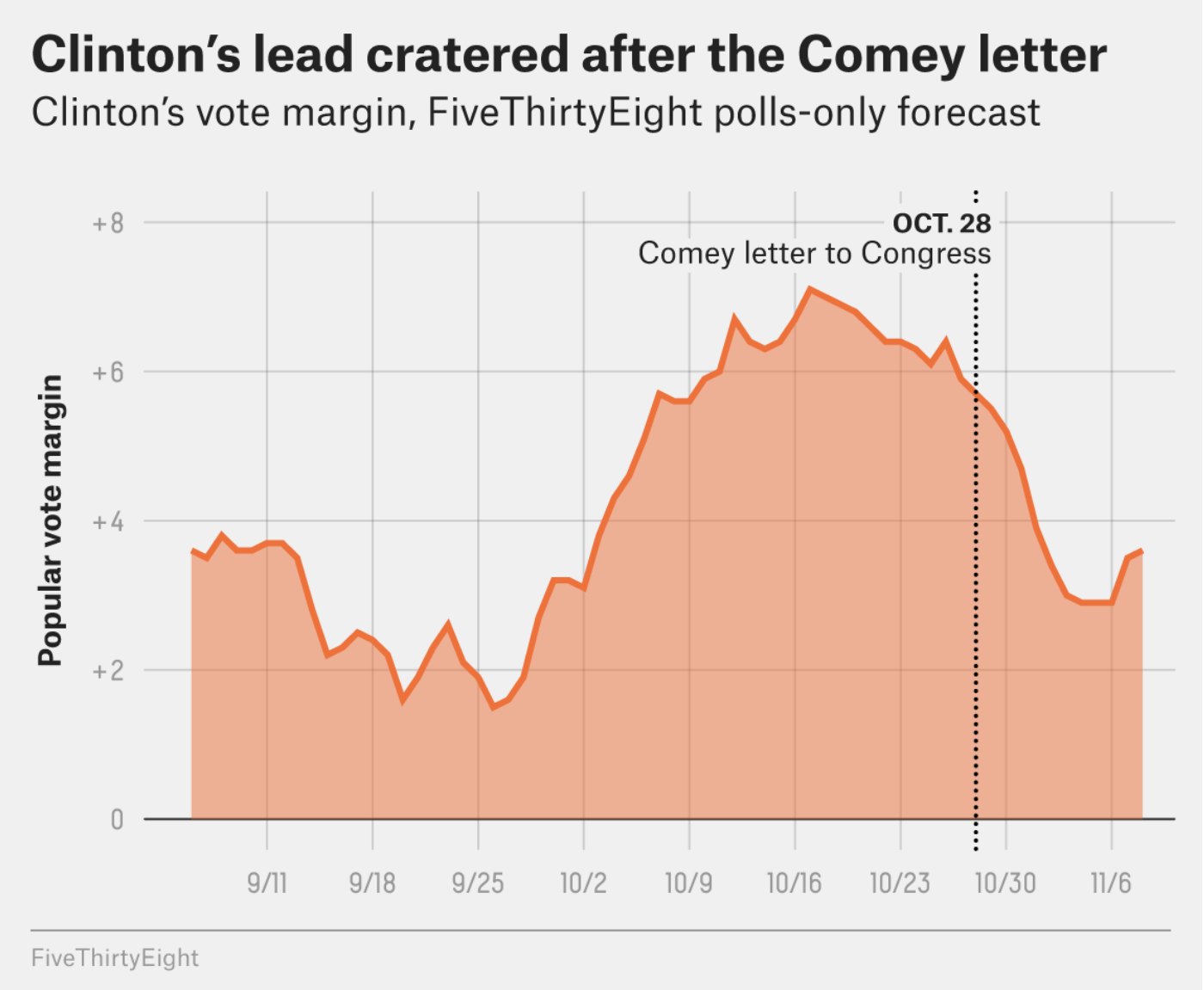


Data As… Story-Telling



Data As… Story-Telling

Data = Drama



Course Overview

Tools for Truths, Skills for Stories…

Our Goals

Truth-Telling **Story-Telling**

Our Means

- Microsoft Excel
- Python
- Pandas
- Tweepy
- VADER
- Matplotlib / Seaborn
- APIs
- Beautiful Soup
- Machine Learning
- MySQL
- MongoDB
- HTML / CSS
- JavaScript
- D3.js
- Leaflet.js / Google Maps
- CartoDB
- Tableau
- Hadoop

Daily Routine

For each class we'll run through the following:

- Set Objectives
- Brief Background Lecture
- Watch Me / Coding Demos
- Code Discussions
- In-Class Exercises
- Project Work

Daily Routine

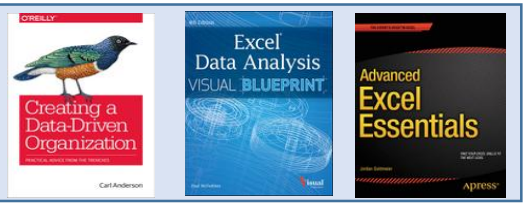
For each class we'll run through the following:

- Set Objectives
- Brief Background Lecture
- Watch Me / Coding Demos
- Code Discussions
- In-Class Exercises
- Project Work

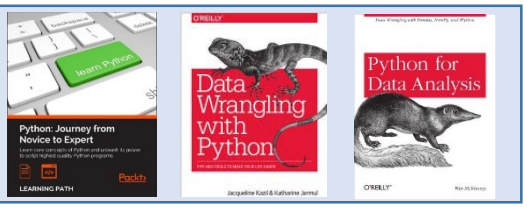
The Super Important Stuff!!!

i.e. Always be doing!

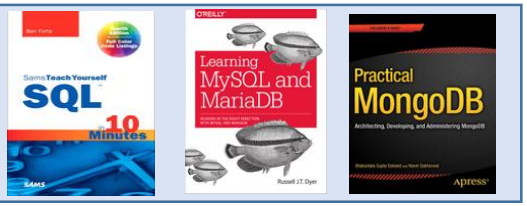
Curriculum At-A-Glance



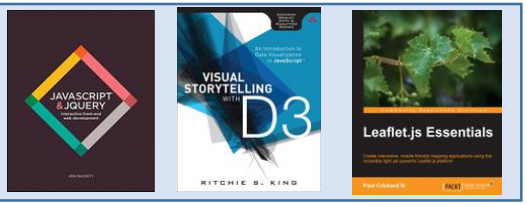
Weeks 1-2: Intro to Data Analytics & Excel Masters
Students begin the course with an introduction to the high-level concepts of data analytics and real-world data crunching with Excel Formulas, Pivot Tables, and Conditional Formatting.



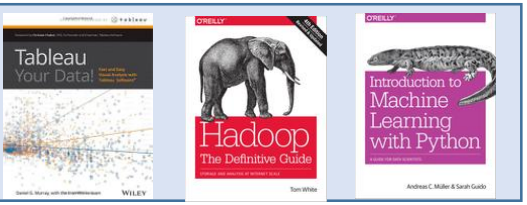
Weeks 3-9: Python Data Analytics and Visualization
Next, students are given a thorough crash-course on Python programming, followed by multiple weeks of data processing using advanced libraries like NumPy, Pandas, Matplotlib, Seaborn, Tweepy, and Beautiful Soup.



Weeks 10-12: Deep Dive into Databases
Students then immerse themselves in three-weeks of introductory and advanced work with SQL (MySQL and Postgresql) and noSQL databases (MongoDB).



Weeks 13-17: Web Based Data Visualization
Students then complete a series of rigorous weeks introducing them on the fundamental tools of web-development (HTML, CSS, JavaScript) and advanced libraries useful for data visualization (D3.js, Leaflet.js)



Weeks 20-24: Final Projects & Advanced Topics
Finally, students complete the program by developing a “real-world” data visualization project that applies all they’ve learned. During this time student will be introduced to advanced topics like Tableau, Hadoop, Machine Learning, and R.

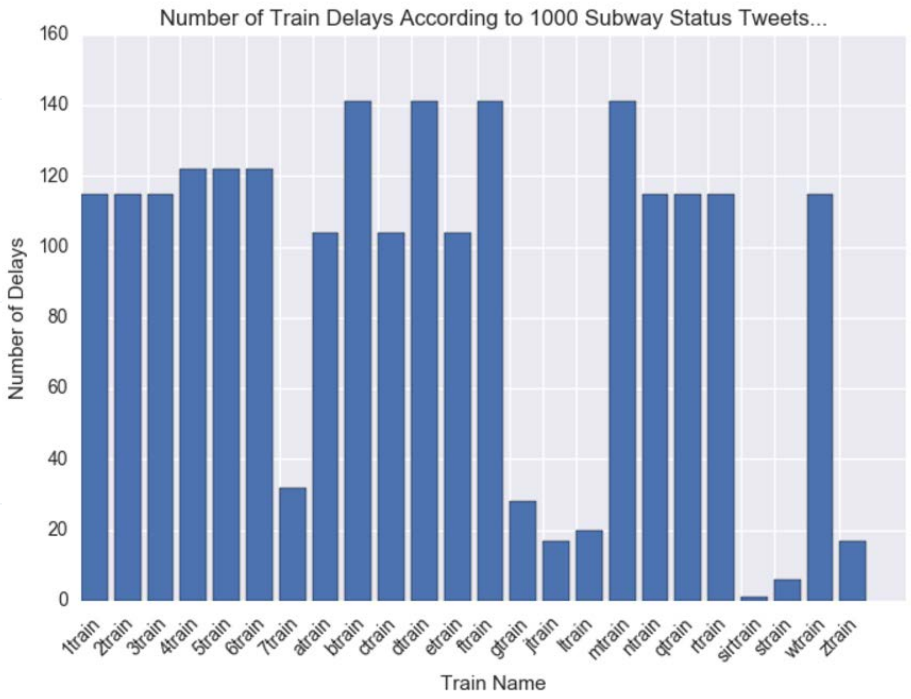
Example Activity: Subway Delays… According to *Twitter*

Subway Status Delays @SubwayStats · 2m
#4train #5train and #6train have delays #NYC subwaystats.com

Subway Status Delays @SubwayStats · 22m
Service change on #4train #5train and #6train #MTA subwaystats.com

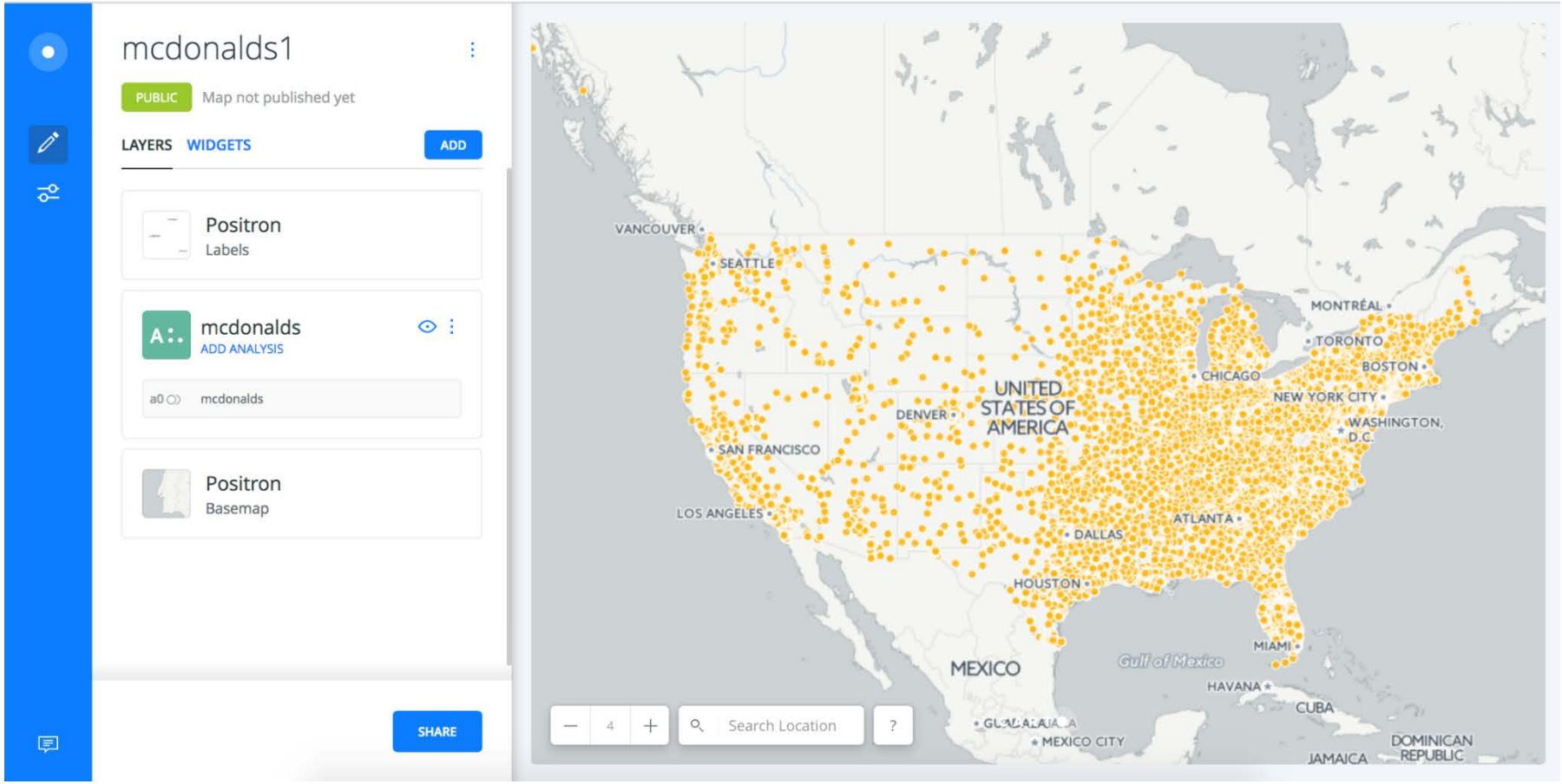
Subway Status Delays @SubwayStats · 5h
#Btrain #Dtrain #Ftrain #Mtrain have planned work ... #NYC subwaystats.com

Subway Status Delays @SubwayStats · 5h
Planned work on #Atrain #Ctrain and #Etrain ... #MTA subwaystats.com



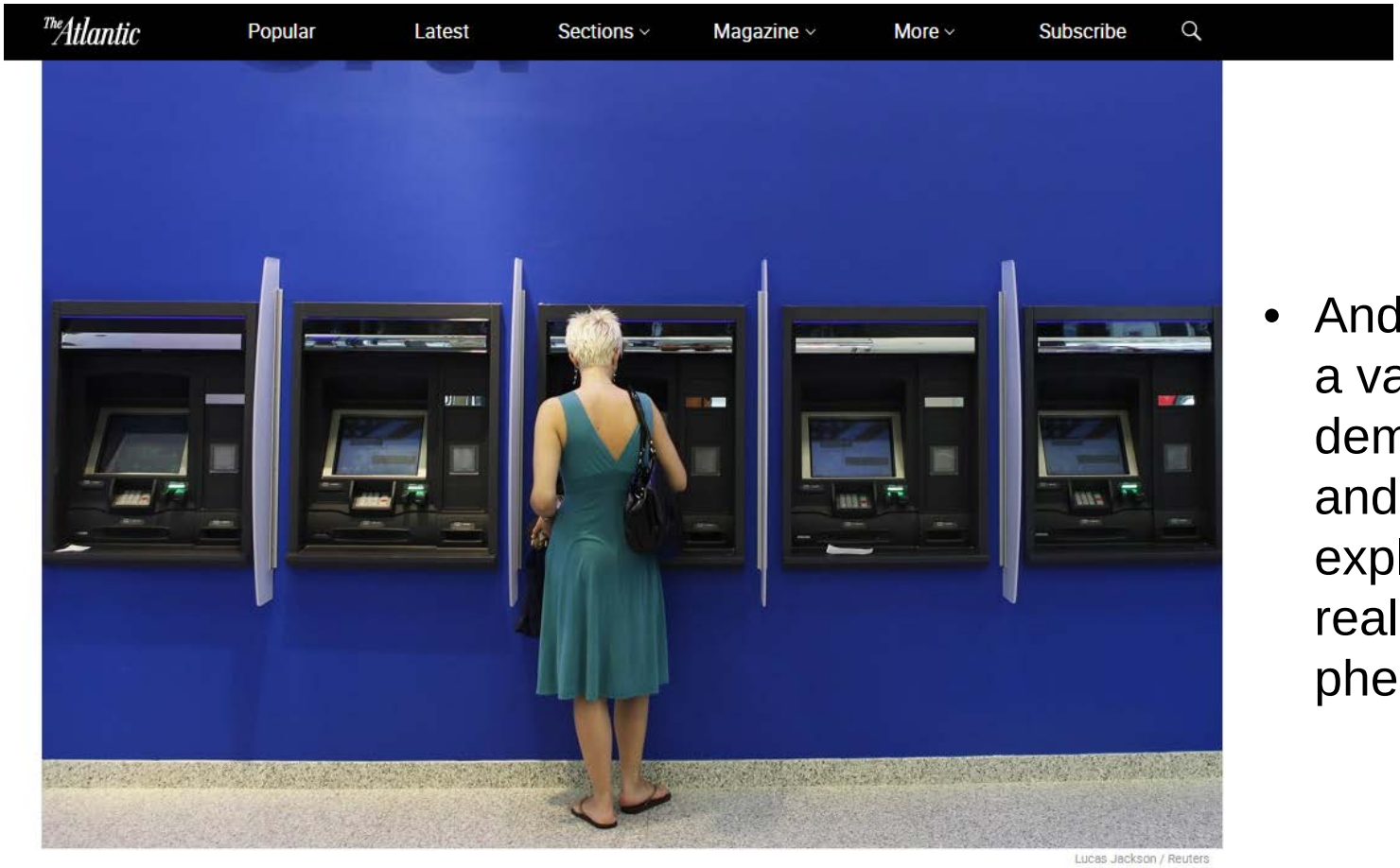
- We'll utilize Twitter Data online to assess which NYC trains are most likely to be delayed. We will be using Python Pandas, Tweepy, Twitter's API, and Matplotlib to convert tweet data into meaningful graphs.

Example Activity: McDonalds Map (Carto)



- We'll be using publicly available data to create interactive dashboards of all McDonalds across the country using easy-to-use visualization tools like Carto.

Example Activity: Banking Deserts



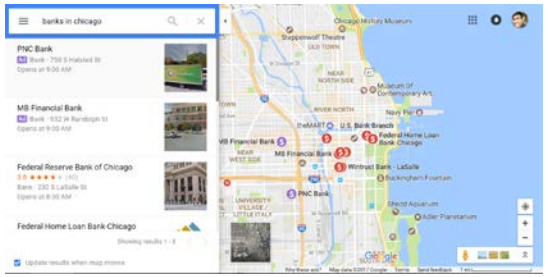
Life in a Banking Desert

Without access to basic financial services, poor and minority communities are more likely to use dangerous, high-cost options.

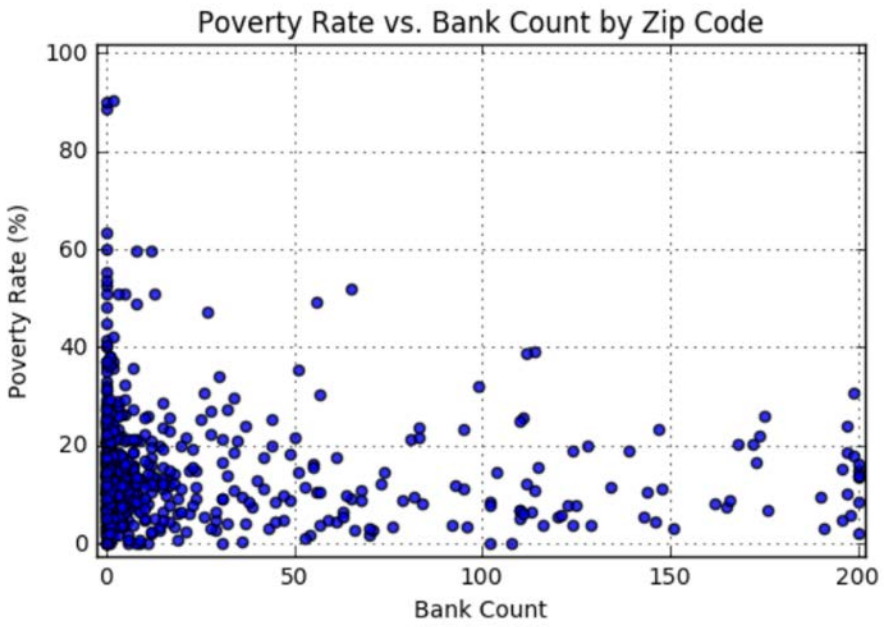
TERRI FRIEDLINE AND MATHIEU DESPARD | MAR 13, 2016 | BUSINESS

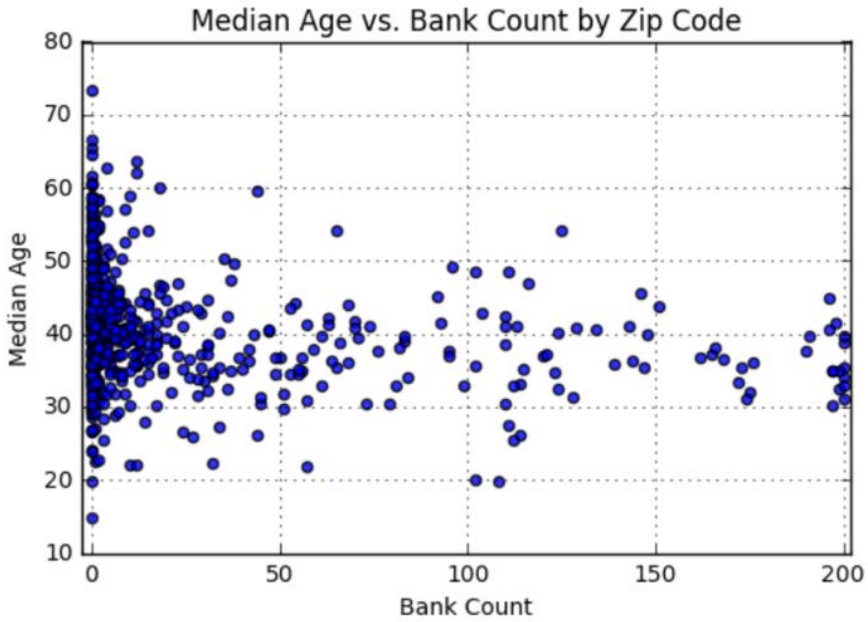
- And we'll be using a variety of public demographic data and APIs to explain many real-world social phenomena.

Example Activity: Banking Deserts



United States™ Census 2010





- We'll be utilizing data from sources like the US Census, Google Maps, and more to dredge out insights on poverty, discrimination, and the impact of changing economies.

We've got a long ways to go…

Helpful Tips

1. Embrace Your Inner Toddler

2. Brace Yourself for Doubt, Challenge, and Confusion

2. Brace Yourself for Doubt, Challenge, and Confusion

"You can't tell whether you're learning something when you're learning it—in fact, **learning feels a lot more like frustration**."

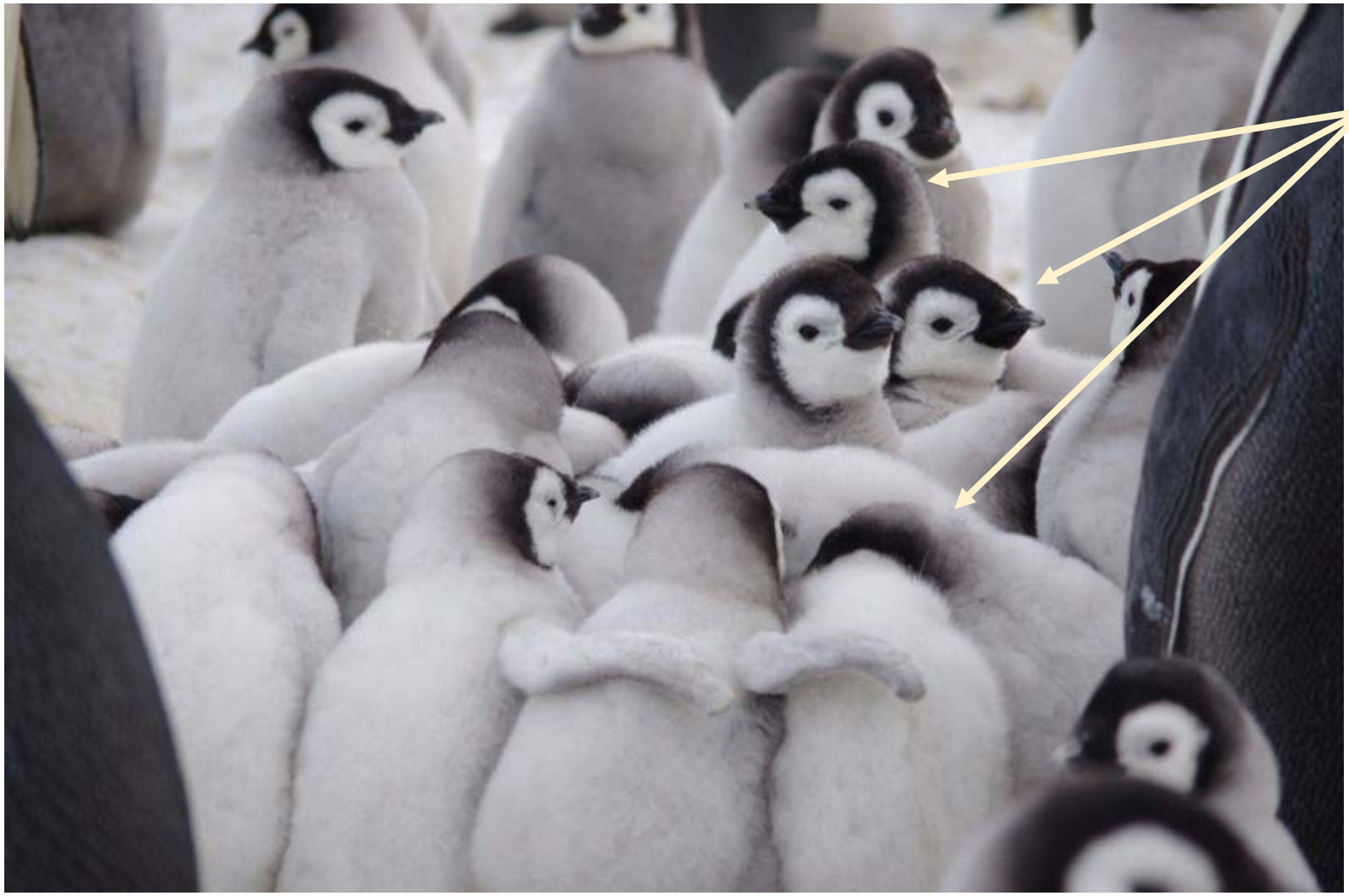
"What I've learned is that during this period of frustration is actually when people improve the most, and their improvements are usually obvious to an outsider. If you feel frustrated while trying to understand new concepts, try to remember that it might not feel like it, but you're probably rapidly expanding your knowledge."

Jeff Dickey, Author of Write Modern Web Apps with the MEAN Stack: Mongo, Express, AngularJS, and Node.JS

3. Relish the Novice Experience

Expect a *lot* of lightbulb moments

4. Form a Community Now!



Your Classmates

5. Put in the Hours (And the Effort)!

There is no “magic pill”. You’ve got to put in the hours!

6. Celebrate Your Successes!

It’s time for a group activity!

Form a group of 3-4 students.

(Psst… They shouldn't be someone right next to you)

Break!

Thought Experiment #1

The Great Debate

Imagine…

Your entire bonus rests on answering this next question.

Or if you want something higher stakes... Imagine
Your entire life depends on answering this question…

The Question…

Which do Americans prefer: Italian or Mexican food?

> YOUR TURN!

Activity*:* The Great Debate **| Suggested Time:** 15 min

Assignment:
With your group develop a complete strategy for answering this question with as much confidence possible. Specifically, answer questions like:

- What data will you attempt to gather?
- What relationships will you be looking for?
- How will you ensure your answer is most likely “true”?

Assumptions:

- You are given 5 hours and a budget of $10 to accomplish this.
- Your answer will be tested by randomly selecting 9 Americans who will each be asked the question – with 0 qualifiers.
- You only have your team.

Be prepared to share! *(P.S. Your answer had better not be: “We Googled it”)*

Thought Experiment #1

The Great Debate (Analyzed)

Step 1:
Decompose the Ask

Step 1: Decompose the “Ask”

Which do Americans prefer: Italian or Mexican food?

Step 1: Decompose the “Ask”

Which do **Americans** prefer:
Italian or Mexican food?

Step 1: Decompose the "Ask"

Which do **Americans** prefer: Italian or Mexican food?

Questions it Raises:

- Who exactly is an American?
- Are Americans just white, forty-year old males?
- Do Americans just live in big cities?
- Are Americans just millennials?

Obviously not. So, how can we get a ***representative*** sample of Americans?

Step 1: Decompose the “Ask”

Which do Americans **prefer**:
Italian or Mexican food?

Step 1: Decompose the "Ask"

Which do Americans **prefer**: Italian or Mexican food?

Questions it Raises:

- How do we define "preference"?
- Do people prefer the foods they eat most frequently?
- Do people prefer the foods they *wish* they could eat if cost was not an issue?
- How uniform is the preference? Is it regionalized? Is it different by demographic?

Inherently, preference is ***subjective.*** We are going to need to make it ***objective.***

Step 1: Decompose the “Ask”

Which do Americans prefer:

Italian or Mexican food?

Step 1: Decompose the “Ask”

Which do Americans prefer:

Italian or Mexican food?

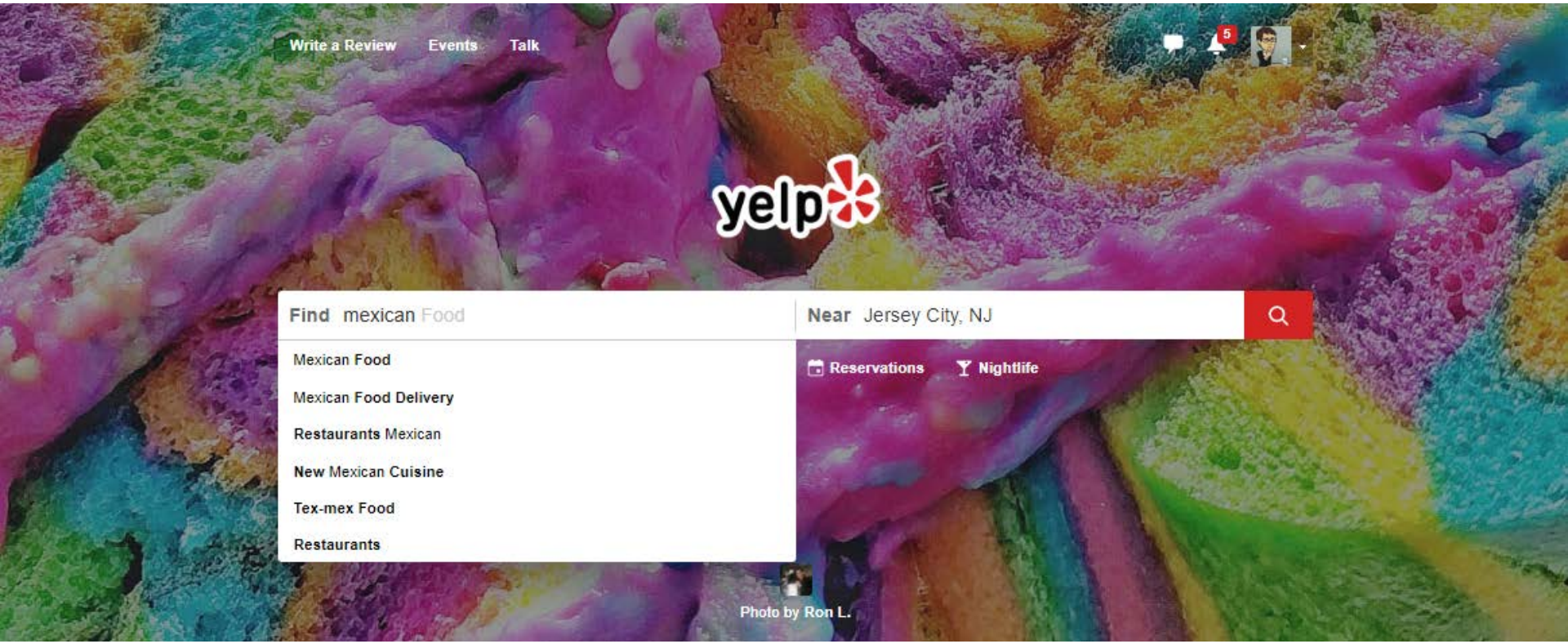
Questions it Raises:

- How do we categorize foods? Is Pizza Italian? Is Taco Bell Mexican?
- How do we categorize food? Does making pasta at home constitute Italian? Or are we just talking about restaurants?
- Are we just talking about “best experiences?” Or are we including poorer renditions of these foods?

These are **broad** categories we are pursuing. We will have to **narrow** the scope.

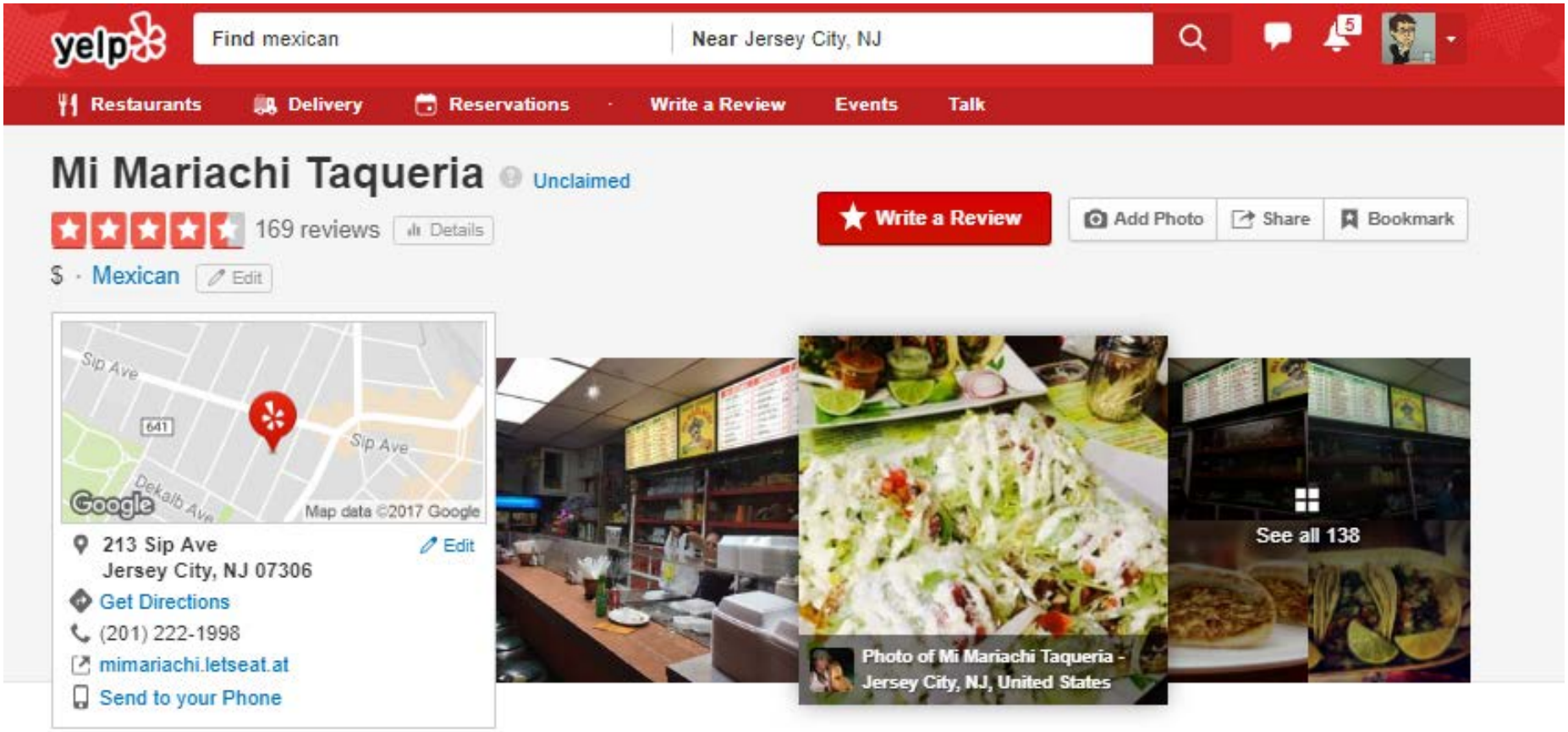
Step 2:
Identify Data Sources

Step 2: Identify Data Sources



- As everyday consumers, we are *regularly* getting a pulse of everyday American food preferences to inform our own decisions. Perhaps we can make use of the same approach?

Step 2: Identify Data Sources



- Accessing a web service like Yelp provides an almost encyclopedic amount of information on the eating preferences of Americans.

Step 2: Identify Data Sources

- **Why poll an audience,** when there already exist enormous databases of information on American food preferences – readily available online?

Step 2: Identify Data Sources

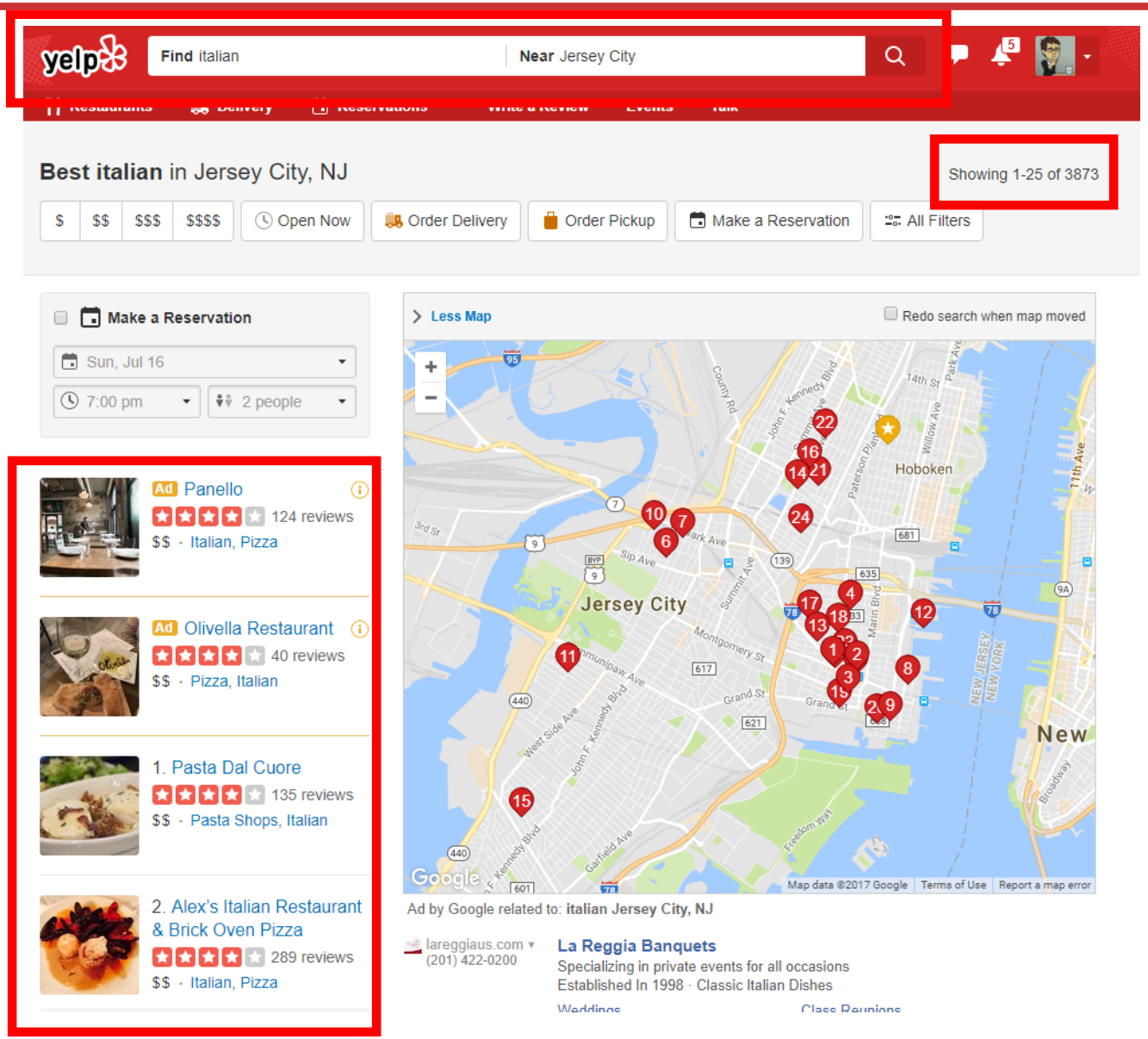
Food Type

Location

And *LOTS* of Data!!

Review Counts

Ratings



Thank you Yelp!!!!

Step 3:
Define Strategy and Metrics

Step 3: Define Strategy and Metrics

Here we created a blueprint for what we're targeting:

Americans:

- Ideally we need thousands of records from Americans in hundreds of different cities. (Large samples)

Preference:

- Number of Yelp Reviews (More = Preference)
- Average Aggregated Ratings (Higher = Preference)

Italian and Mexican Food:

- Top 20 Italian and Mexican restaurants in every city.

Step 3: Define Strategy and Metrics

New York

Italian		Mexican
Restaurant		Restaurant
Restaurant	VS	Restaurant
Restaurant		Restaurant
Restaurant		Restaurant
Restaurant		Restaurant

Tucson, AZ

Italian		Mexican
Restaurant		Restaurant
Restaurant	VS	Restaurant
Restaurant		Restaurant
Restaurant		Restaurant
Restaurant		Restaurant

Washington, DC

Italian		Mexican
Restaurant		Restaurant
Restaurant	VS	Restaurant
Restaurant		Restaurant
Restaurant		Restaurant
Restaurant		Restaurant

Omaha, NE

Italian		Mexican
Restaurant		Restaurant
Restaurant	VS	Restaurant
Restaurant		Restaurant
Restaurant		Restaurant
Restaurant		Restaurant

San Diego, CA

Italian		Mexican
Restaurant		Restaurant
Restaurant	VS	Restaurant
Restaurant		Restaurant
Restaurant		Restaurant
Restaurant		Restaurant

Atlanta, GA

Italian		Mexican
Restaurant		Restaurant
Restaurant	VS	Restaurant
Restaurant		Restaurant
Restaurant		Restaurant
Restaurant		Restaurant

Repeat this analysis for as many cities as possible…

Step 4:
Build Data Retrieval Plan

Step 4: Data Retrieval Plan

We *could* brute force our way to retrieve this data, but…

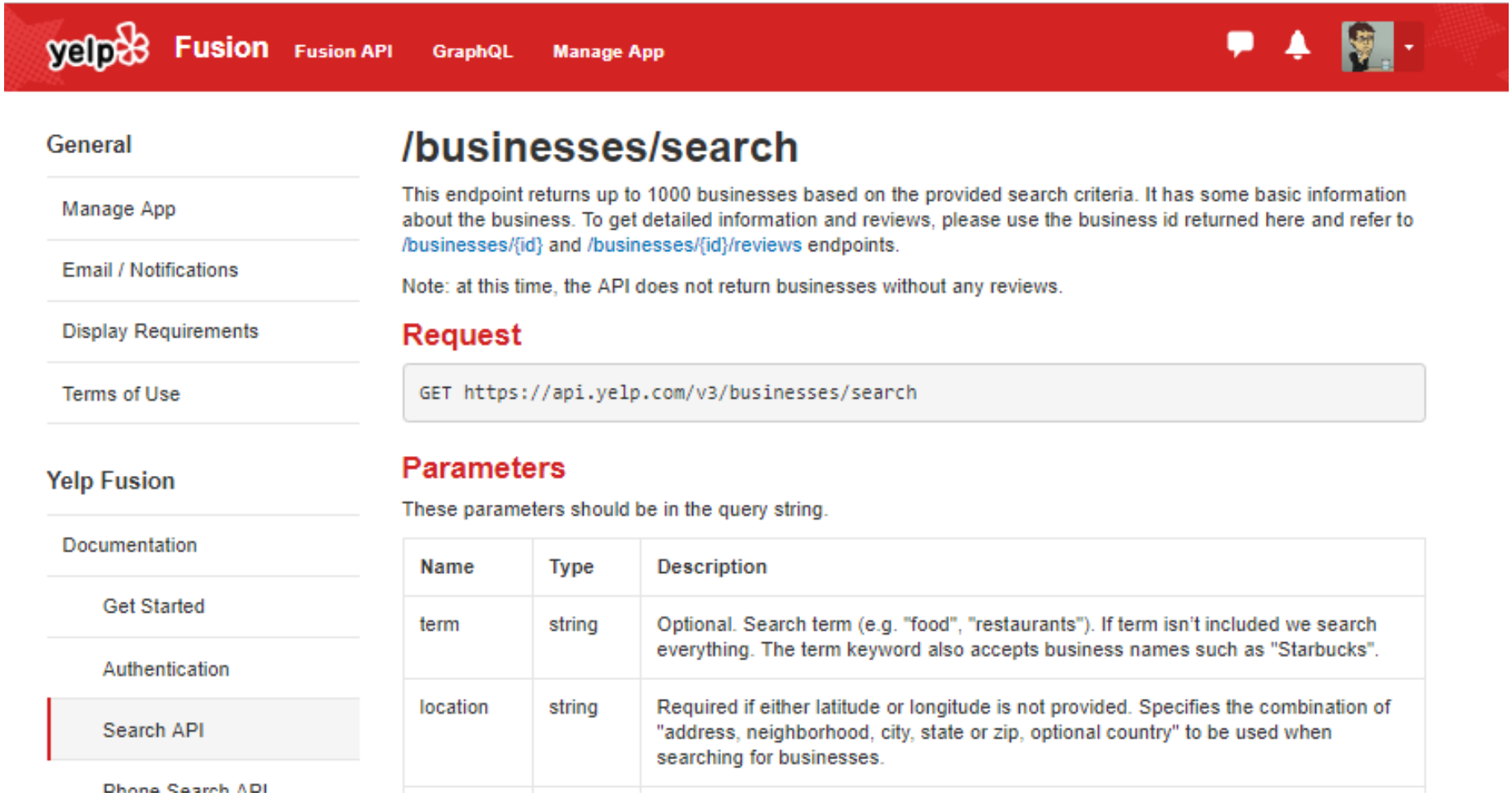
- It would be extremely time consuming
- It would be skewed by our city familiarity
- It would be manually labor intensive

Step 4: Data Retrieval Plan

Basically…

Not in a million years.

Thank You Yelp!



Thankfully, we can take advantage of the **Yelp Fusion API** to **programmatically** run our queries. (#ThankGodForProgramming)

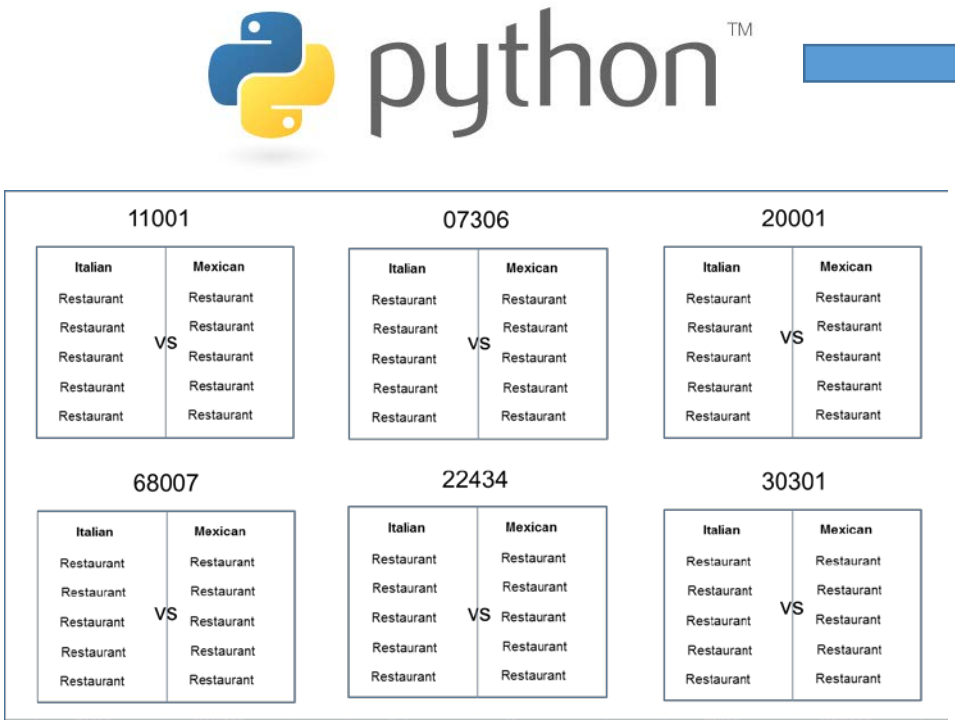
Thank You Yelp!

Thankfully, we can take advantage of the **Yelp Fusion API** to **programmatically** run our queries. (#ThankGodForProgramming)

Response Body


```
{
  "total": 8228,
  "businesses": [
    {
      "rating": 4,
      "price": "$",
      "phone": "+14152520800",
      "id": "four-barrel-coffee-san-francisco",
      "is_closed": false,
      "categories": [
        {
          "alias": "coffee",
          "title": "Coffee & Tea"
        }
      ],
      "review_count": 1738,
      "name": "Four Barrel Coffee",
      "url": "https://www.yelp.com/biz/four-barrel-coffee-san-francisco",
      "coordinates": {
        "latitude": 37.7670169511878,
        "longitude": -122.42184275
      },
      "image_url": "http://s3-media2.fl.yelpcdn.com/bphoto/MmgtASP3l_t4tPCL1iAsCg/o.jpg",
      "location": {
        "city": "San Francisco",
        "country": "US",
        "address2": "",
        "address3": "",
        "state": "CA",
        "address1": "375 Valencia St",
        "zip_code": "94103"
      },
      "distance": 1604.23,
      "transactions": ["pickup", "delivery"]
    },
    // ...
  ],
  "region": {
    "center": {
      "latitude": 37.767413217936834,
      "longitude": -122.42820739746094
    }
  }
}
```

Step 4: Build Data Retrieval Plan



We will build a Python script to randomly select over 700 zip codes from the US Census and then acquire review data from the top 20 Mexican and Italian restaurants for each zip codes using the Yelp API.

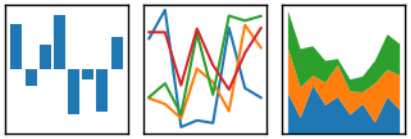
Step 5:
Retrieve the Data

Pulling with Python

Randomly Select a Zip Code

Create an API Request

Save the Output to a Data Frame



Pulling with Python

```
# Use Try-Except to handle errors
try:

    # Loop through all records to calculate the review count and weighted review value
    for business in yelp_reviews_italian["businesses"]:

        italian_review_count = italian_review_count + business["review_count"]
        italian_weighted_review = italian_weighted_review + business["review_count"] * business["rating"]

    for business in yelp_reviews_mexican["businesses"]:
        mexican_review_count = mexican_review_count + business["review_count"]
        mexican_weighted_review = mexican_weighted_review + business["review_count"] * business["rating"]

    # Append the data to the appropriate column of the data frames
    italian_data.set_value(index, "Zip Code", row["Zipcode"])
    italian_data.set_value(index, "Italian Review Count", italian_review_count)
    italian_data.set_value(index, "Italian Average Rating", italian_weighted_review / italian_review_count)
    italian_data.set_value(index, "Italian Weighted Rating", italian_weighted_review)

    mexican_data.set_value(index, "Zip Code", row["Zipcode"])
    mexican_data.set_value(index, "Mexican Review Count", mexican_review_count)
    mexican_data.set_value(index, "Mexican Average Rating", mexican_weighted_review / mexican_review_count)
    mexican_data.set_value(index, "Mexican Weighted Rating", mexican_weighted_review)

except:
    print("Uh oh")
```

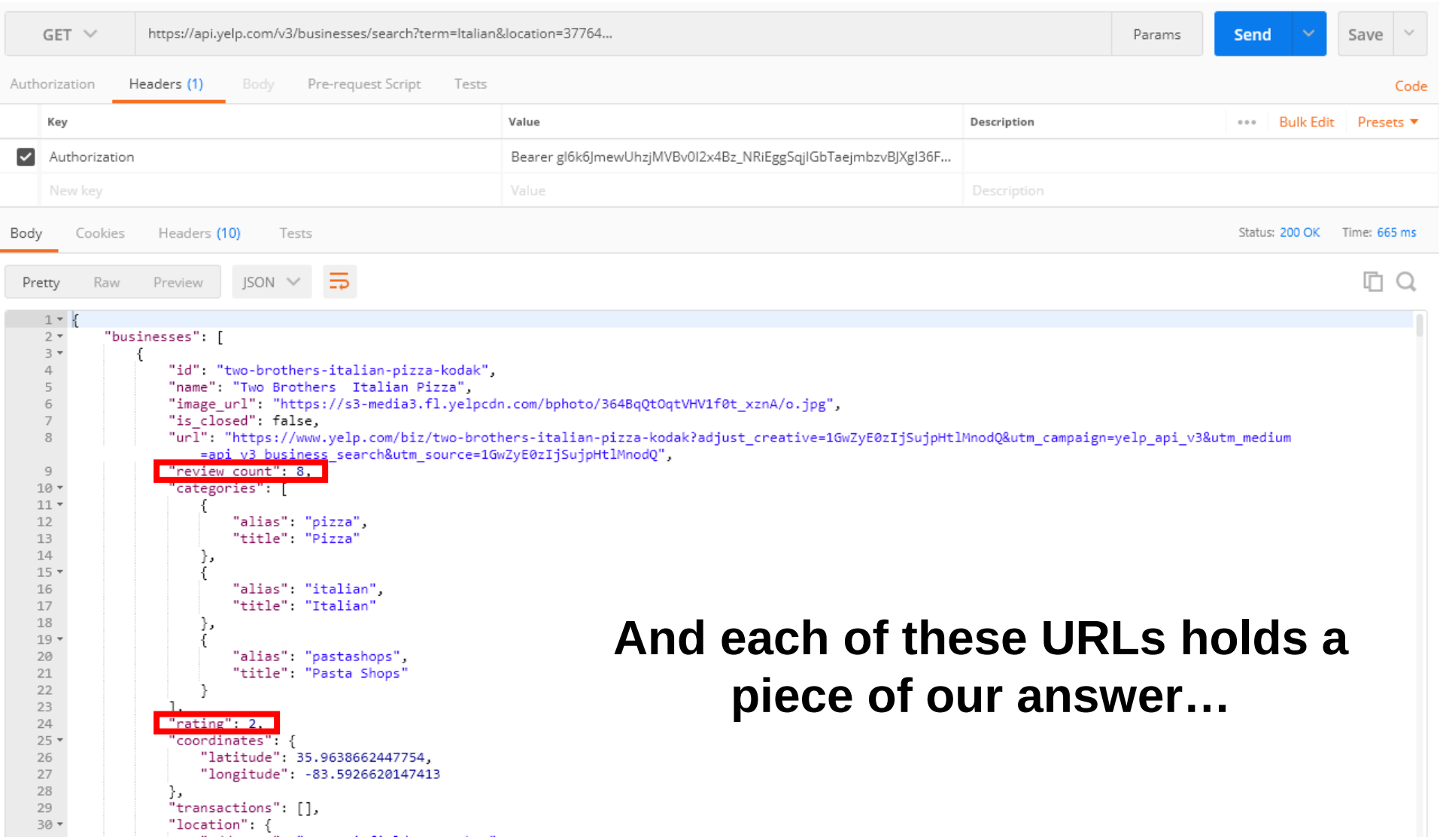
This funky code…

Pulling with Python

```
1
https://api.yelp.com/v3/businesses/search?term=Italian&location=76556
https://api.yelp.com/v3/businesses/search?term=Mexican&location=76556
2
https://api.yelp.com/v3/businesses/search?term=Italian&location=72039
https://api.yelp.com/v3/businesses/search?term=Mexican&location=72039
3
https://api.yelp.com/v3/businesses/search?term=Italian&location=61606
https://api.yelp.com/v3/businesses/search?term=Mexican&location=61606
4
https://api.yelp.com/v3/businesses/search?term=Italian&location=47232
https://api.yelp.com/v3/businesses/search?term=Mexican&location=47232
5
https://api.yelp.com/v3/businesses/search?term=Italian&location=60565
https://api.yelp.com/v3/businesses/search?term=Mexican&location=60565
6
https://api.yelp.com/v3/businesses/search?term=Italian&location=20634
https://api.yelp.com/v3/businesses/search?term=Mexican&location=20634
7
https://api.yelp.com/v3/businesses/search?term=Italian&location=71046
https://api.yelp.com/v3/businesses/search?term=Mexican&location=71046
```

Will make all these URLs

Pulling with Python



And each of these URLs holds a piece of our answer…

Step 6:
Assemble and Clean the Data

Cleaning with Pandas

```
# Combine DataFrames into a single DataFrame
combined_data = pd.merge(mexican_data, italian_data, on="Zip Code")
combined_data.head()
```

	Zip Code	Mexican Review Count	Mexican Average Rating	Mexican Weighted Rating	Italian Review Count	Italian Average Rating	Italian Weighted Rating
0	76556	97	4.1134	399	63	3.78571	238.5
1	72039	256	4.11133	1052.5	266	3.81955	1016
2	61606	378	3.64286	1377	66	3.2197	212.5
3	47232	222	4.16892	925.5	420	3.77857	1587
4	60565	2842	3.94053	11199	2829	3.92824	11113

No data comes out intrinsically the way you want it to.
In our case, we needed multiple steps to aggregate the data along our channels of interest.

Step 7:
Analyze for Trends

Analyze for Trends (Table)

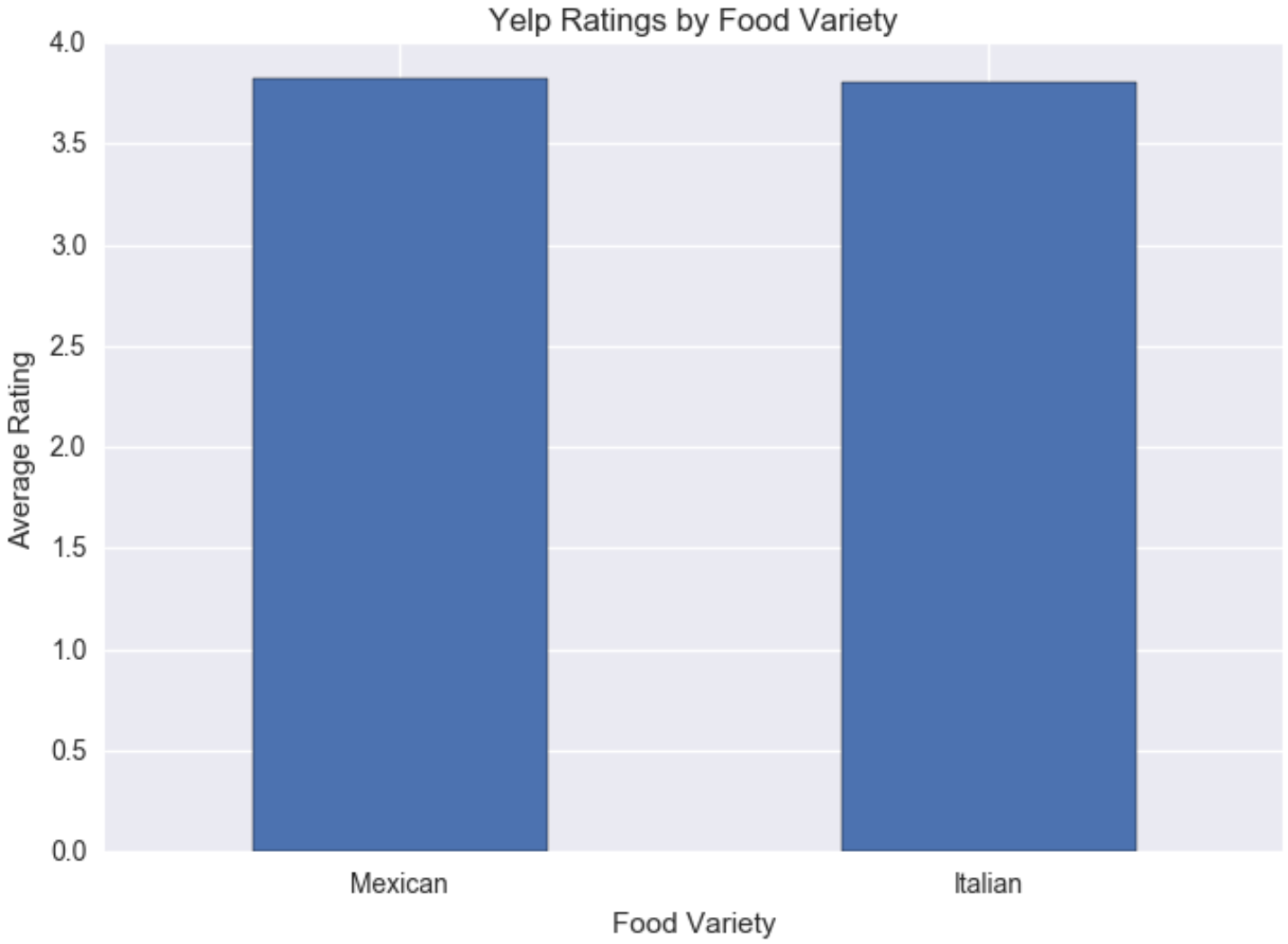
Display Summary of Results

```
# Model 1: Head-to-Head Review Counts
italian_summary = pd.DataFrame({"Review Counts": italian_data["Italian Review Count"].sum(),
                                "Rating Average": italian_data["Italian Average Rating"].mean(),
                                "Review Count Wins": combined_data["Review Count Wins"].value_counts()["Italian"],
                                "Rating Wins": combined_data["Rating Wins"].value_counts()["Italian"]}, index=["Itali

mexican_summary = pd.DataFrame({"Review Counts": mexican_data["Mexican Review Count"].sum(),
                                "Rating Average": mexican_data["Mexican Average Rating"].mean(),
                                "Review Count Wins": combined_data["Review Count Wins"].value_counts()["Mexican"],
                                "Rating Wins": combined_data["Rating Wins"].value_counts()["Mexican"]}, index=["Mexic

final_summary = pd.concat([mexican_summary, italian_summary])
final_summary
```

	Rating Average	Rating Wins	Review Count Wins	Review Counts
Mexican	3.826588	273	220	476889
Italian	3.806869	245	298	573733

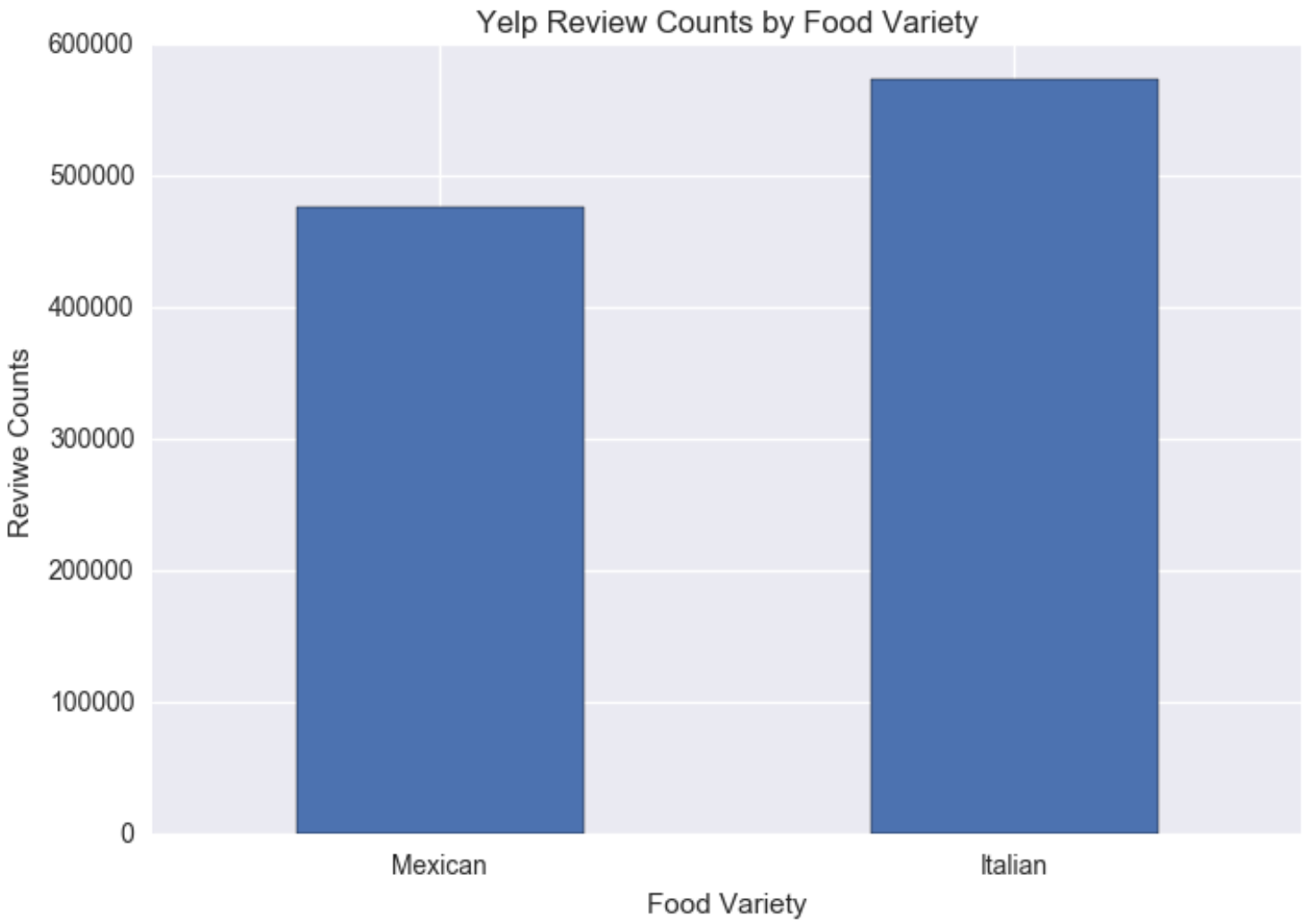
Ugh..
It's Close.

Analyze for Trends (Ratings)



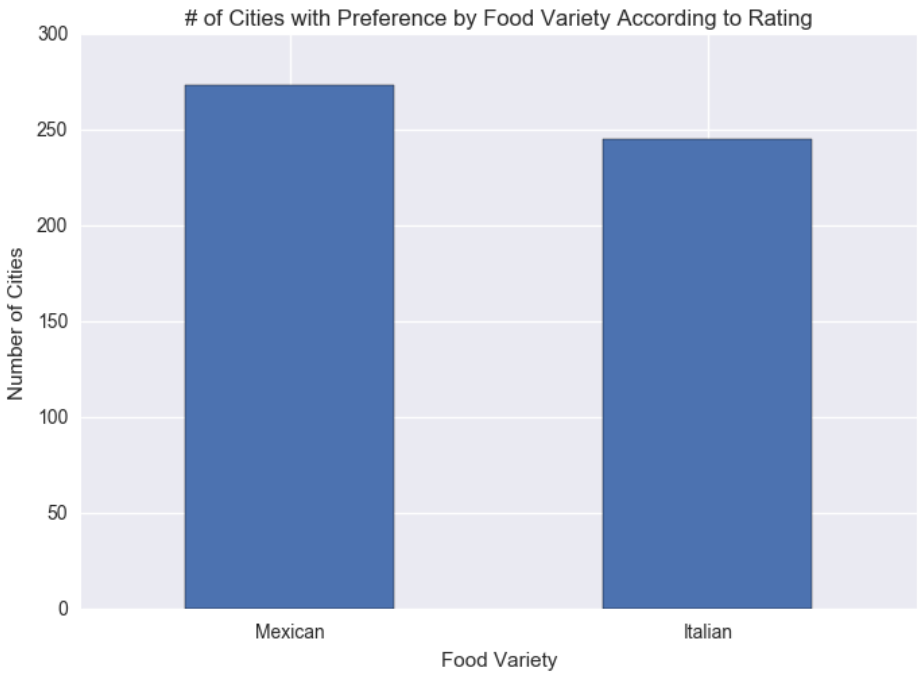
Yelpers Rate Italian and Mexican relatively **equally**.

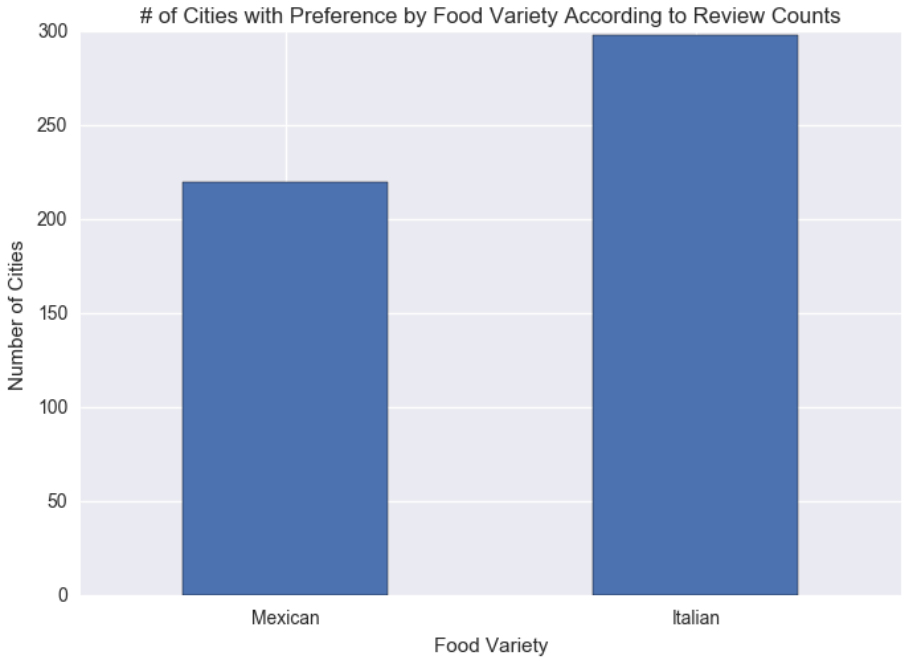
Analyze for Trends (Review Counts)



But… Yelpers seem to significantly **review more Italian** restaurants.

Analyze for Trends (Winner Take All)





Just for kicks I threw in an analysis to ask based on aggregating the data along cities using a “Winner-Take-All” approach.

It’s sort of a wash.

Analyze for Trends (Statistical Analysis)

Metric	Italian	Mexican	p-Value (T-Test)
Average Rating	3.806	3.826	0.284
Review Counts	**573k**	476k	0.057

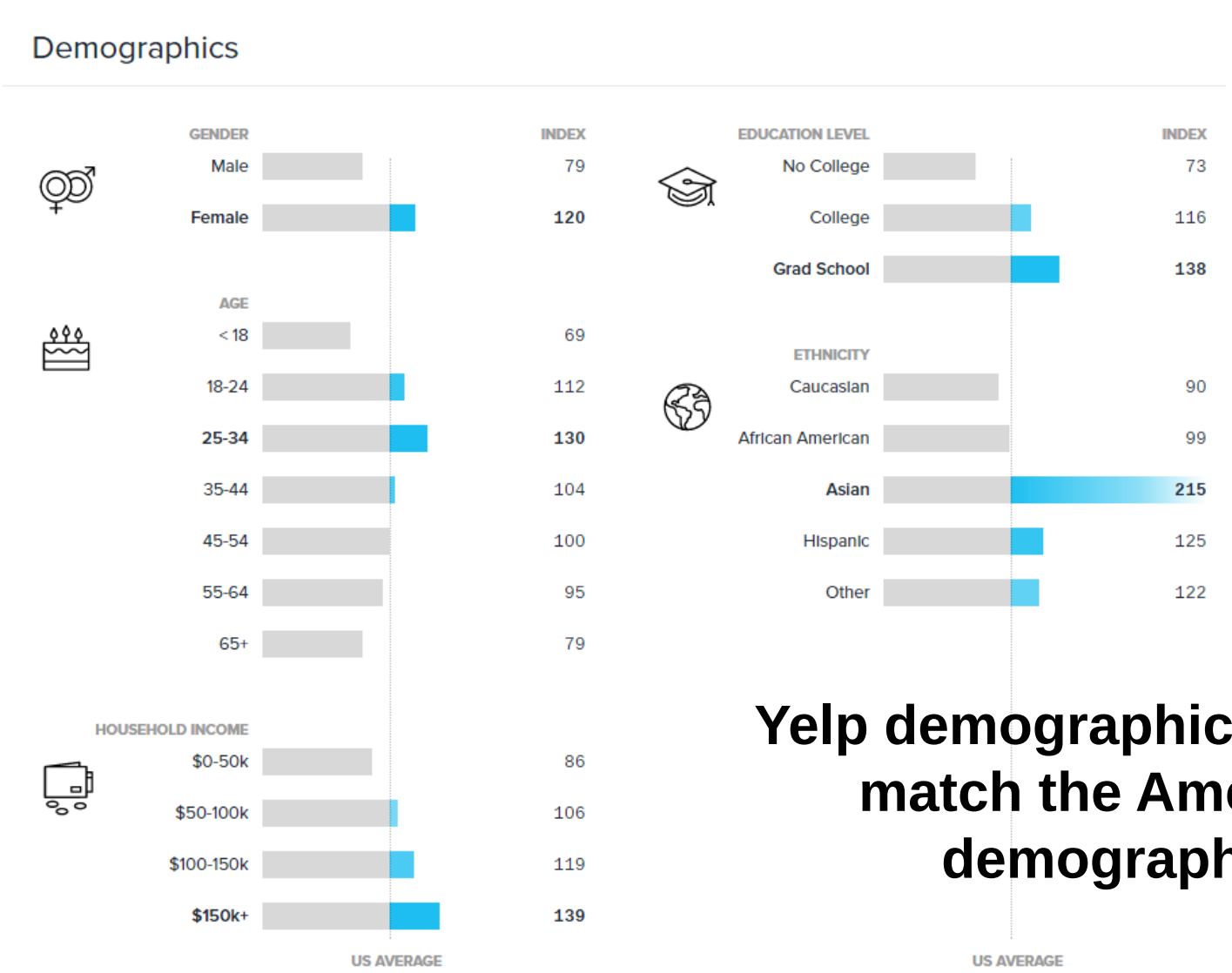
Because of how close the numbers appear, we utilized a Student's T-Test to quickly asses if the perceived differences are statistically significant.

*The difference in **review count is statistically** significant.*

*Note: Depending the type of hypothesis you are making, the smaller the p-value, the more convincing it is. The p-value is related to the effect size, So smaller p-values correspond to larger effect sizes; of course they are more convincing!

Step 8:
Acknowledge Limitations

Limitations in Analysis



Yelp demographics may not match the American demographic.

Limitations in Analysis

Restaurant experiences do not equate to home cooked meals.

Limitations in Analysis

“Fine” dining effect?

Step 9:
Make the Call

Making the Call

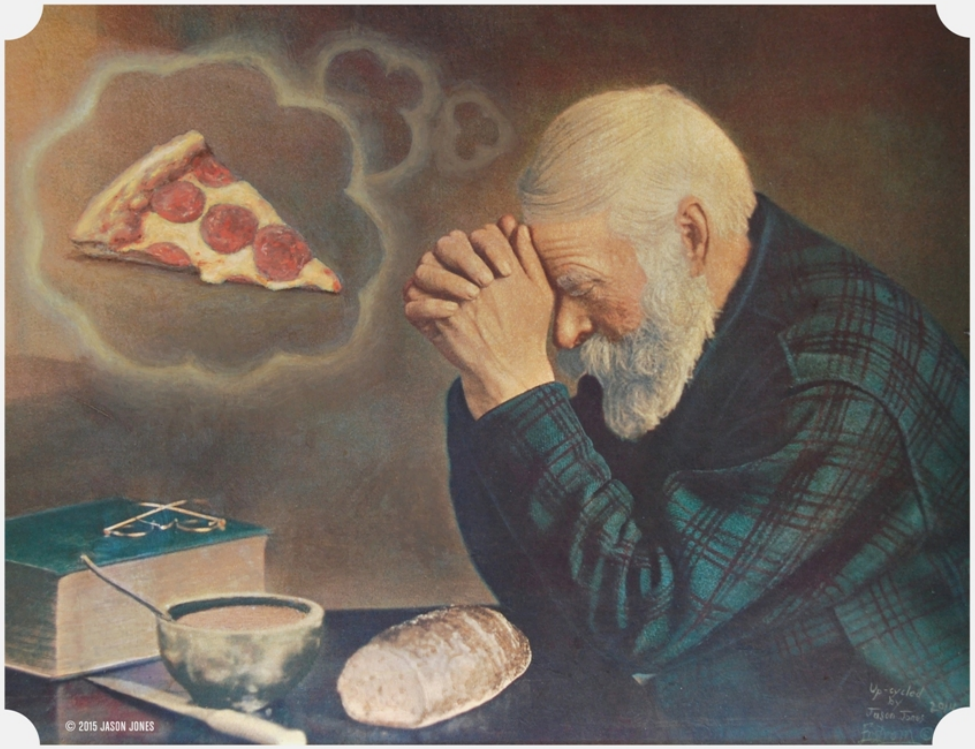
The “Proper” Conclusion:

“Based on our analysis, it is clear that the American preference for Italian and Mexican food are similar in nature. As a whole, Americans rate Mexican and Italian restaurants at statistically similar scores (Avg. score: 3.8, p-value: 0.285). However, there exists statistically significant evidence that Americans write more reviews of Italian restaurants than Mexican (+96k, p-value: 0.057). This may indicate that there is an increased interest in visiting Italian restaurants at an experiential level. However, it may also merely suggest that Yelp users enjoy writing reviews on Italian restaurants more than Mexican restaurants.”

The “Let’s Be Real” Conclusion:

Mexican

(But it’s going to be close…)



Assignment:

Take a few moments to analyze the code provided over slack. Talk to the people in your group and try to dissect as much (or as little) as you can.

- If you are new to coding, your goal should be to understand what a single line does.

- If you are not new to coding, your goal should be to understand the overall flow of activities.

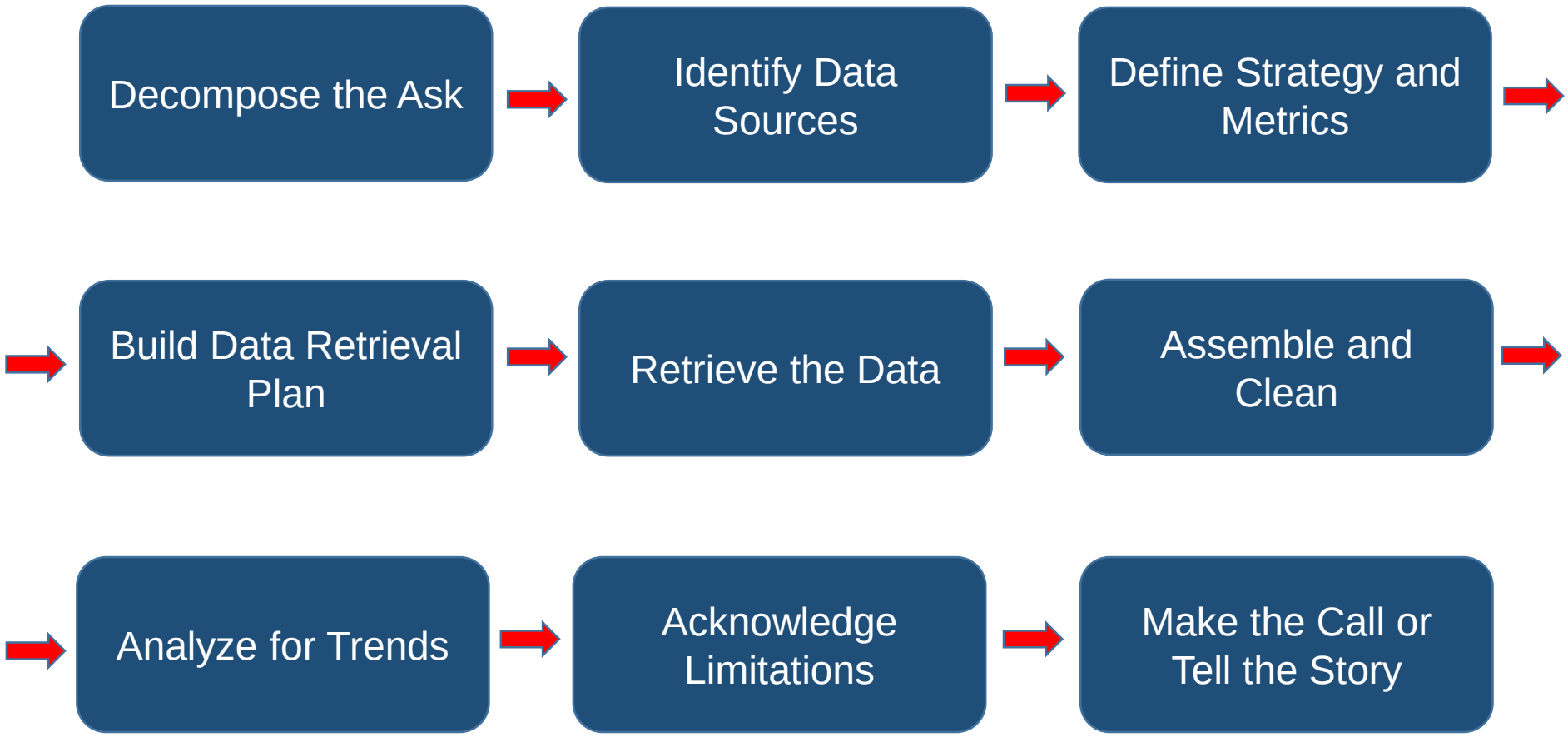
Take a moment to explain what you've learned to the people around you.

Freaked Out a Bit?

It’s okay!

An Analytics Paradigm

Analytics Paradigm



Regardless of type or industry, this paradigm provides a repeatable pathway for effective data problem solving.

Thought Experiment #2

Predicting Gentrification

Let’s Talk Gentrification

> YOUR TURN! **Activity***:* Predicting Gentrification **| Suggested Time:** 10 min

Assignment:
Using the Analytics Paradigm as a framework, outline a strategy by which you would identify which neighborhoods in our city are seeing signs of gentrification?

Specifically, how would you answer the questions:

- What observable signs can we detect to suggest gentrification is happening?
- What means can we use to determine how long the trend has been happening?
- What “proxies” might we use to identify gentrification in non-obvious ways?
- How might you create a visualization of this data to best “tell the story”?

Pay special attention to details like:

- What data will you use to build your “model”?
- How will you retrieve the data?
- What does your final “story” look like?

Prepare for Next Class

Prepare for Next Class

By Next Class:

1. Make certain that you have Microsoft Excel installed.

2. Make certain that you have Slack installed and are actively looking at it.

3. Figure out where the Git repository for our class is.

4. Figure out where class videos will be posted.

Homework #1

Homework #1 - Introduction

id	name	blurb	goal	pledged	state	disable_communication	country	currency
0	GIRLS STATE a new musical comedy TV project	In this new TV show "All Politics is Vocal" as high school girls campaign, sing and cheer to be elected Governor of their summer camp.	8500	11633	successful	FALSE	US	USD
1	FannibalFest Fan Convention	A Hannibal TV Show Fan Convention and Art Collective	10275	14653	successful	FALSE	US	USD
2	Charlie teaser completion	Completion fund for post-production for teaser of British crime/drama tv series about a girl who sells morals for	500	525	successful	FALSE	GB	GBP
3	Unsure/Positive: A Dramedy Series About Life with HIV	We already produced the "very" beginning of this story. Help us to see it	10000	10390	successful	FALSE	US	USD
4	Party Monsters	19th centuryâ¡™s most notorious literary characters, out of step with the times, find comradery as roommates in modern day Los Angeles.	44000	54116.28	successful	FALSE	US	USD
5	Terry Matthews to be the NEXT star on the Network Television	The BBQ Daddy will be Filming the 1st episode of the Next Hit series to come to Network Television "Bailout My	3999	4390	successful	FALSE	US	USD

More information to come on Wednesday!

Questions / Discussion