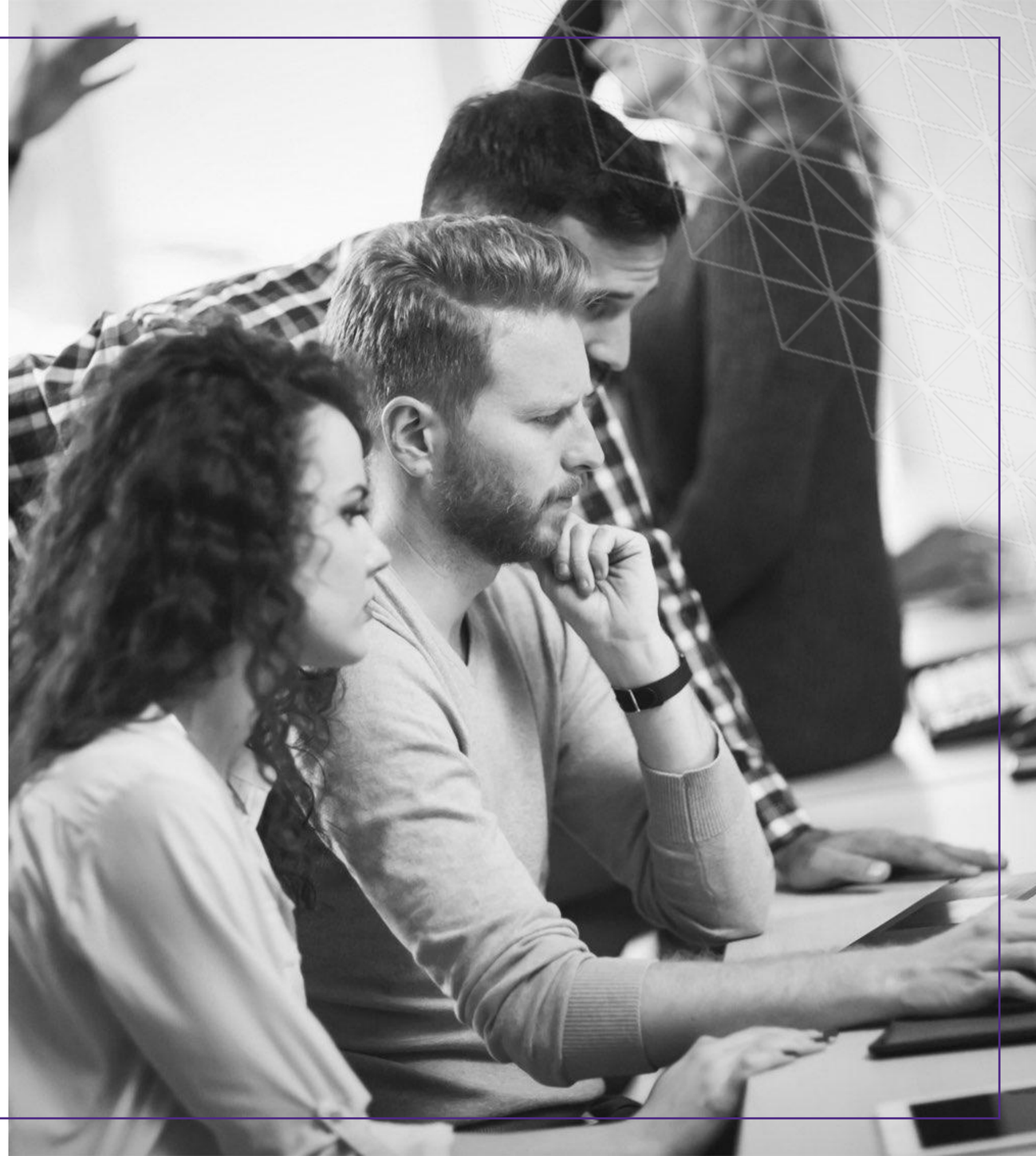
Welcome to the Northwestern Data Science Bootcamp

Wifi Name: Northwestern (if you know your NetID and password) OR Guest

Open:
Slack Desktop App & Add Photo!

Get up and play Bingo!



First Group Project

1. Navigate to www.bootcampspot.com
2. Sign in
3. Click Sessions
4. Find Today’s Date
5. Mark Your Attendance

By the end of this session, you will:

- [ ] Come to know your classmates as the community that you will rely on for collaborative learning.
- [ ] Know the staff who will be providing holistic support throughout the program.
- [ ] Understand the minimum requirements in order to successfully graduate from this boot camp.
- [ ] Be able to list ways to get help and support at your moments of need.

Dartaniel Bliss

Student Success Manager

My goal is to help you successfully complete the program

I enjoy running, camping, shooting pictures, and reading

You can reach me at dbliss@bootcampspot.com

Lauren Jacobs

Program Manager, Northwestern University SPS

"You never lose. Either you win or you learn."

spsbootcamps@northwestern.edu

The School

One of 12 Northwestern University schools, since 1933

Part-time programs for adults

Academically rigorous

The Approach

Access, Impact, Excellence

The Community

Networking opportunities with professional peers and thought leaders

Northwestern

Your turn!

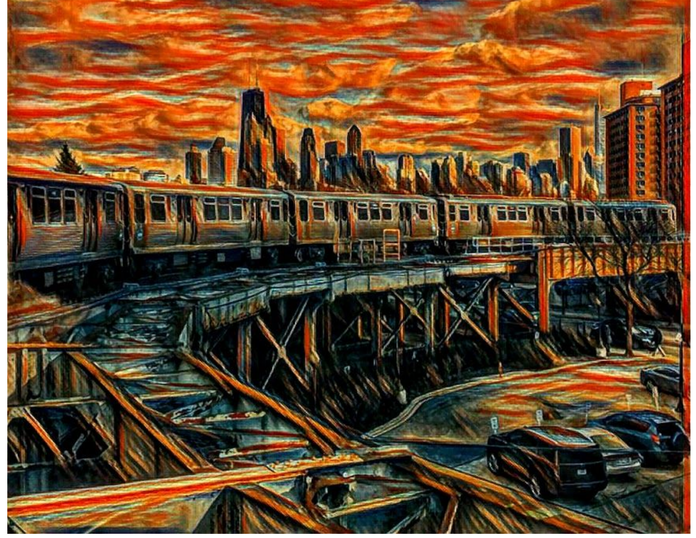
In 30 seconds or less, please share:

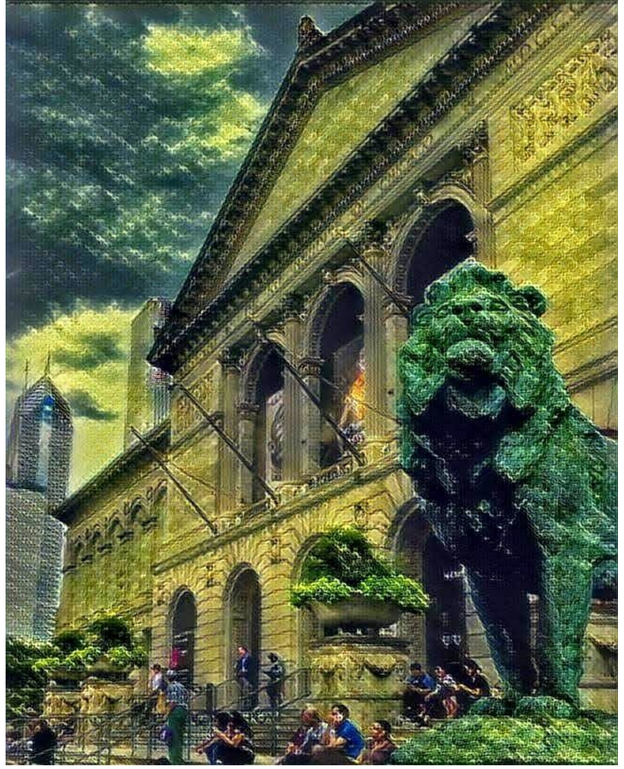
Name

Location (Live and Commuting from)

Background (Career, Education, Interests)

One Fun Fact about yourself





Cody Nicholson

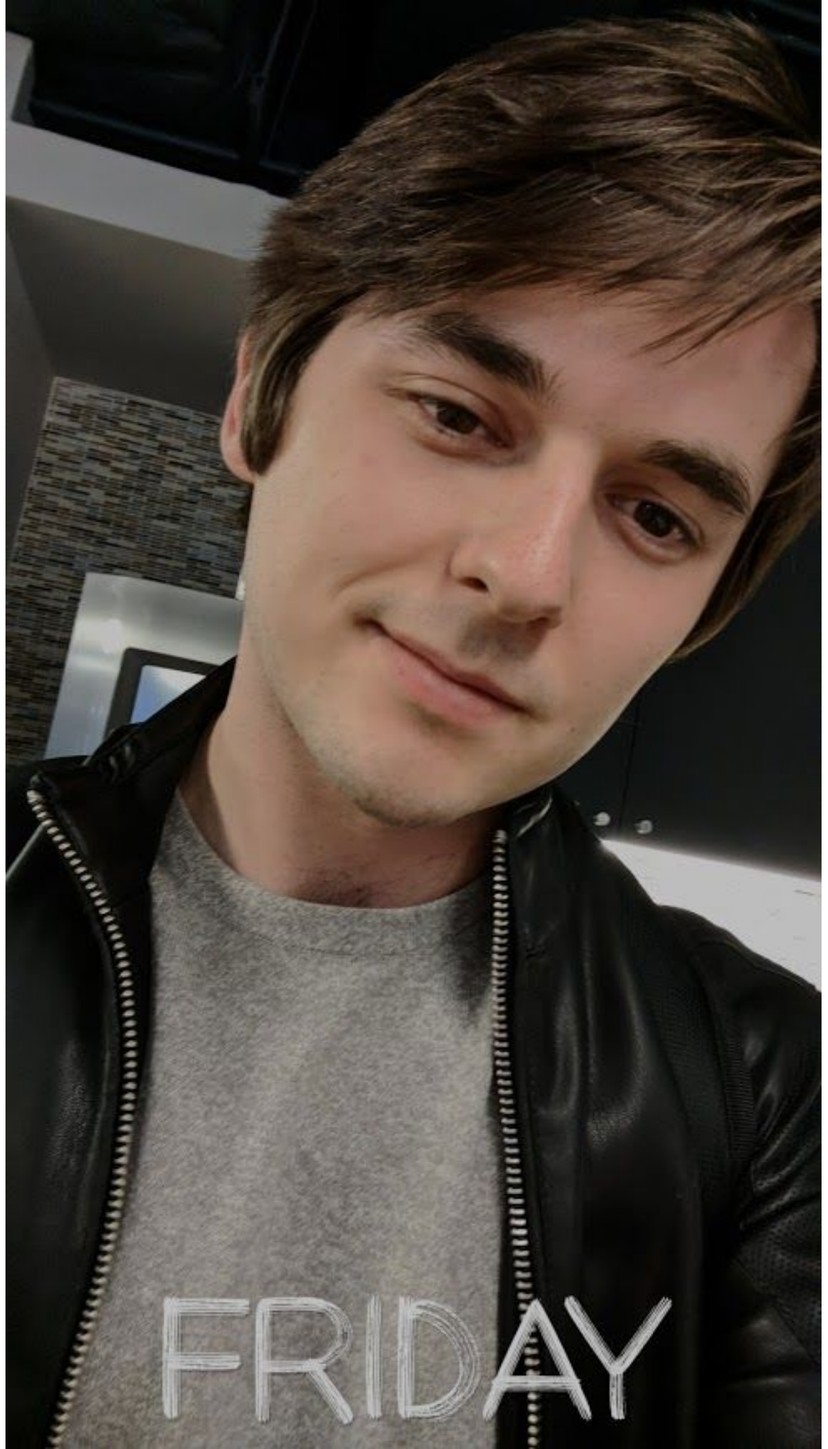
Teaching Assistant, Mon/Wed

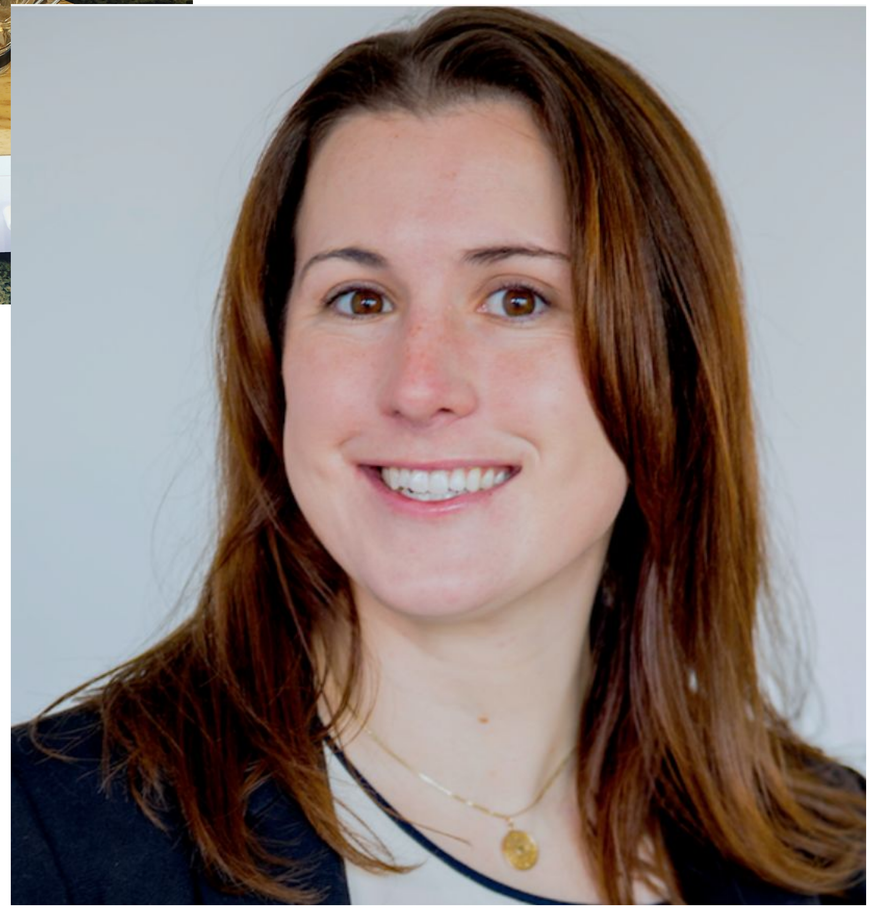
Currently a Software Consultant at Perficient working in the loop.

I used to program self-driving cars until I realized how much I didn't like doing all the math that was involved.

I can show you how to use machine learning to apply cool filters to images you take - like I did to the ones above!

Before studying computer science I was really passionate about doing card tricks. I was really happy when I found out that many of the algorithms I used in card tricks were utilized in software to achieve different goals.





Jennifer Kregor

Teaching Assistant, Mon/Wed

Currently the lead data scientist at Brightfield Group

Graduate of the first or second ? Northwestern cohort.

Worked with the Cambridge Analytica dataset as it was originally intended to be used: to linguistic measures of well-being and health.

Just got the above contraption--a gourmet coffee alarm clock after waiting for four years for the startup to actually make it.

Travis Taylor

Instructor, Mon/Wed

Data professional for over 10 years

Former roller derby photographer

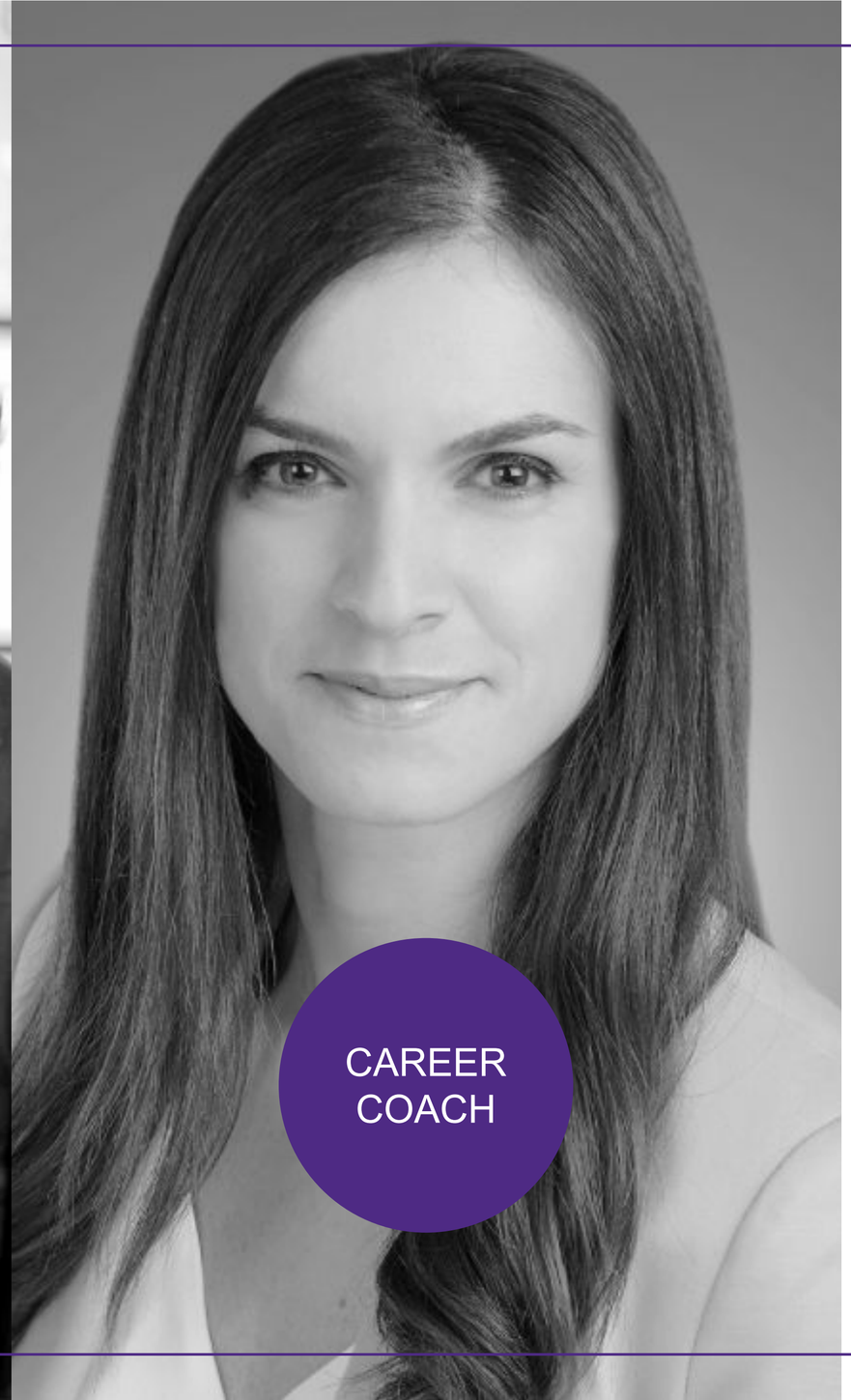
Calculated the damages in a securities fraud case that settled for $2.4 billion

TUE/THUR TEAM

Mike
TA

Ronit
TA

Asif
Instructor



YOUR CAREER SERVICES TEAM

PROFILE COACHES

CAREER COACH

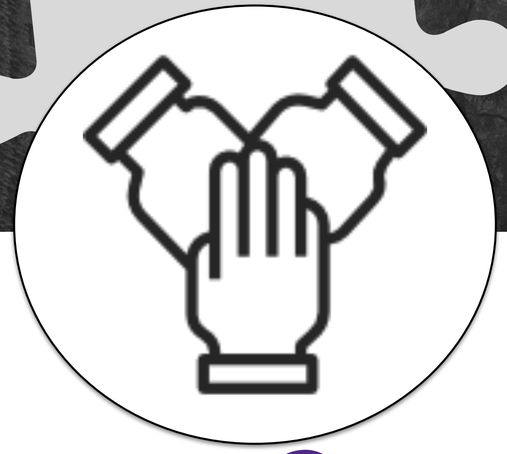
YOU

SUPPORT STRATEGIES

You are responsible for your **success**, but you're not alone!

You

Live Chat on Bootcampspot



Schedule a Check-in with your SSM

Attend Office Hours

Form Study Groups

Activity: Landing Zones

1. Read this, then break into groups. (We’ll direct you)

2. As a group, decide on a meeting time and place for your study group on your assigned day of the week.

3. Post day and location in your Slack and pin.

4. Choose a speaker to share your decision with the class.

Minimum Graduation Requirements:

- [ ] Miss no more than 4 classes
- [ ] Have no more than 2 incomplete homeworks
- [ ] Participate in all projects

Four Absences & Four Remote Sessions

1. Make formal request in BootcampSpot (remote sessions are the same in BCS as formal absences but “remote attendance” is selected)
2. Notify staff of absence/remote status in advance via Slack If remote: remain active during the hours of class (ask/answer questions, send screenshots of class activities in real-time).
3. If formal absence: 2-paragraph summary of class recording, screenshots of completed in-class activities emailed to SSM within 1-week.

Let’s Review What We’ve Covered So Far:

- [x] Come to know your classmates as the community that you will rely on for collaborative learning.
- [x] Know the staff who will be providing holistic support throughout the program.
- [x] Understand the minimum requirements in order to successfully graduate from this bootcamp.
- [x] Be able to list ways to get help and support at your moments of need.

Break

See you back in 15 minutes!

Thought Experiment #1

The Great Debate

It’s time for a group activity!

Form a group of 3-4 students.

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

Imagine…

Your entire bonus rests on answering this next question.

Which do Americans prefer: Italian or Mexican food?

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

Which do Americans prefer: Italian or Mexican food?

Which do **Americans** prefer:

Italian or Mexican food?

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?

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

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.***

Which do Americans prefer:

Italian or Mexican food?

Which do Americans prefer:

Italian or Mexican food?

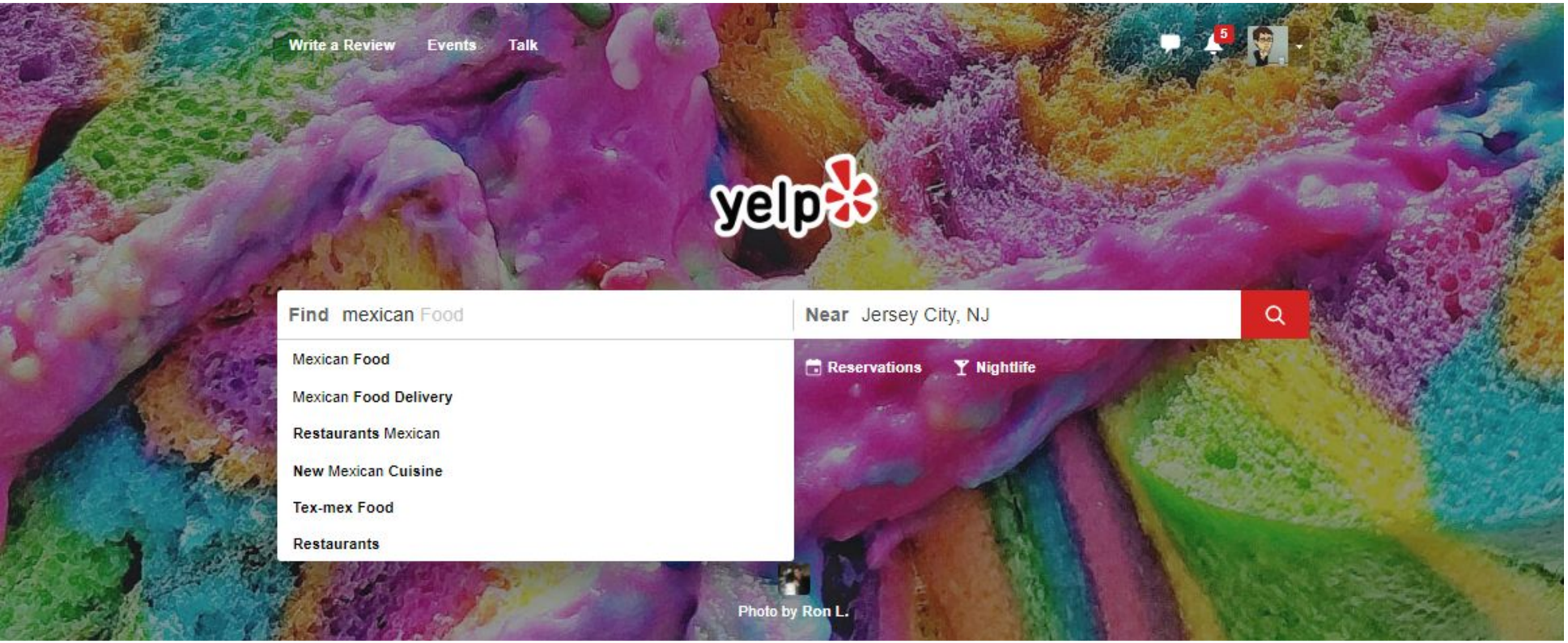
Questions it Raises:

- 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.

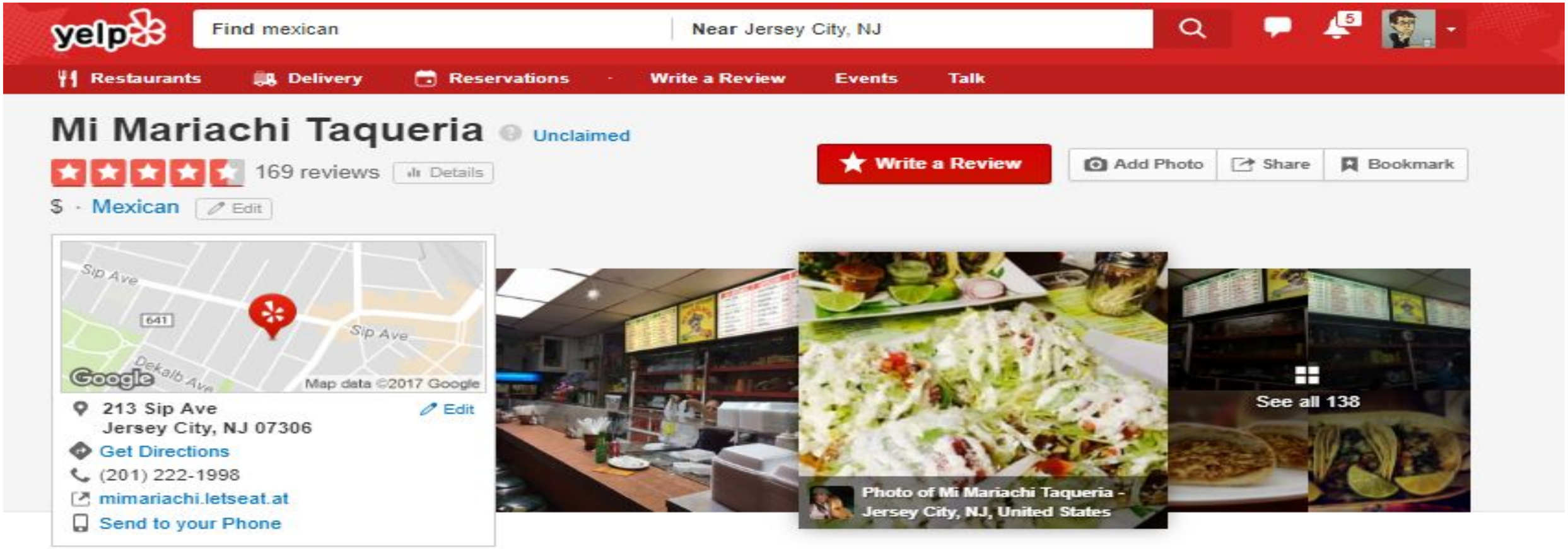
Step 2:
Identify Data Sources

Step 2: Identify Data Sources



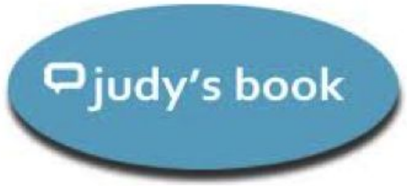
- As everyday consumers, we are *regularly* getting a pulse of everyday American food preferences to inform our own decisions. Perhaps we can make use of the same approach?

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

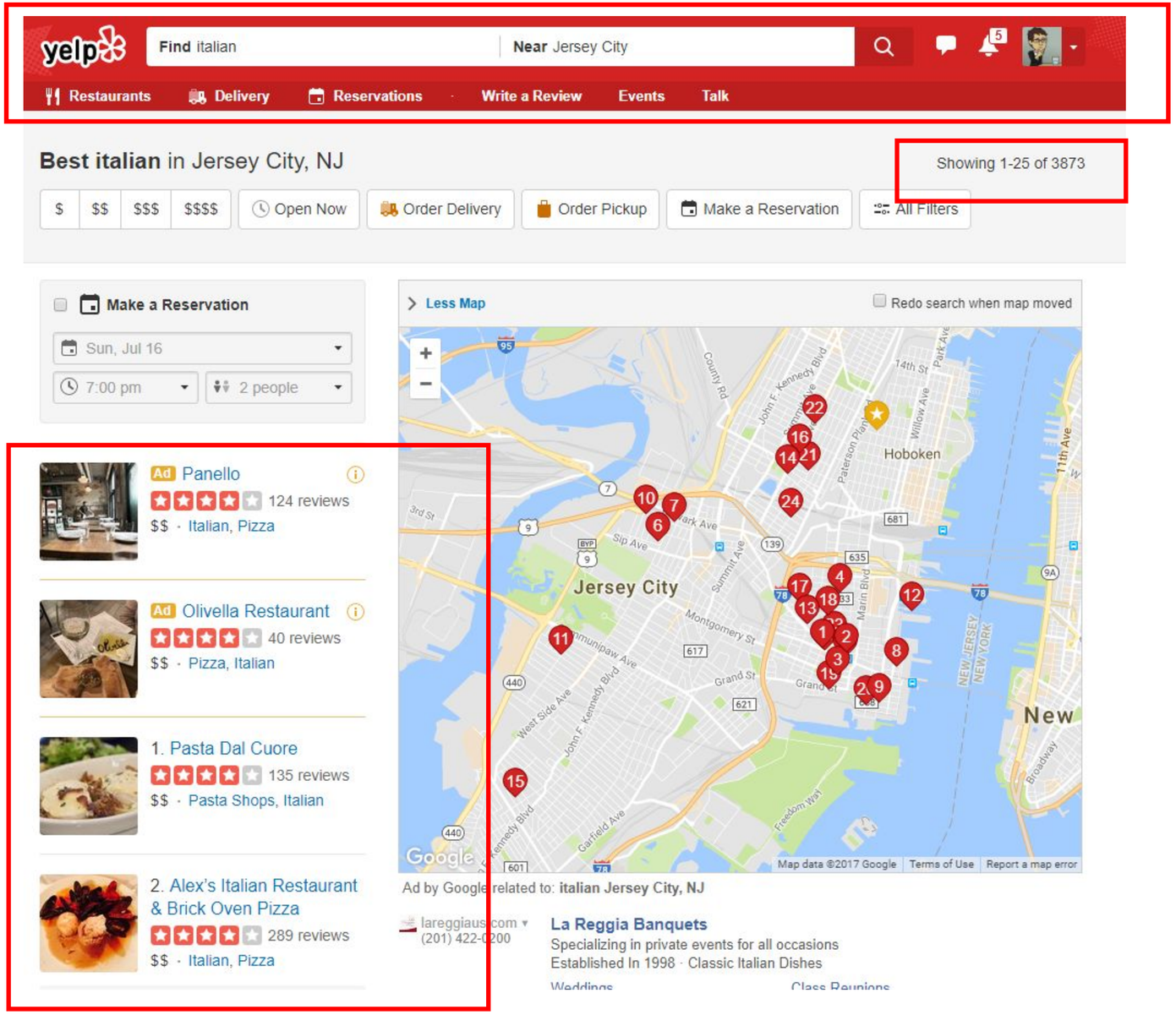
Food Type

Location

And *LOTS* of Data!!

Review Counts

Ratings



Thank you Yelp!!!!

Step 3:
Define Strategy and Metrics

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		Restaurant
Restaurant	VS	Restaurant
Restaurant		Restaurant
Restaurant		Restaurant

Tucson, AZ

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

Washington, DC

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

Omaha, NE

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

San Diego, CA

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

Atlanta, GA

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

Repeat this analysis for as many cities as possible…

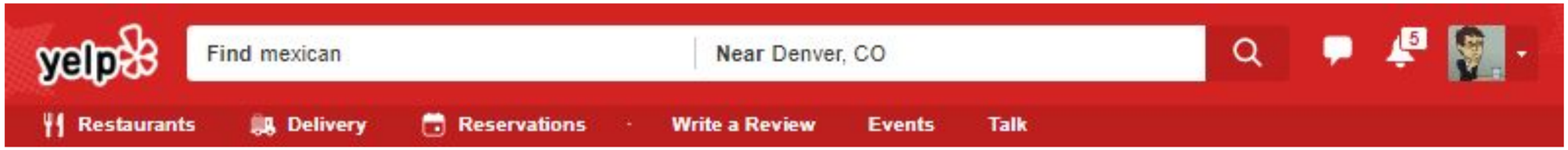
Step 4:
Build Data Retrieval Plan

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

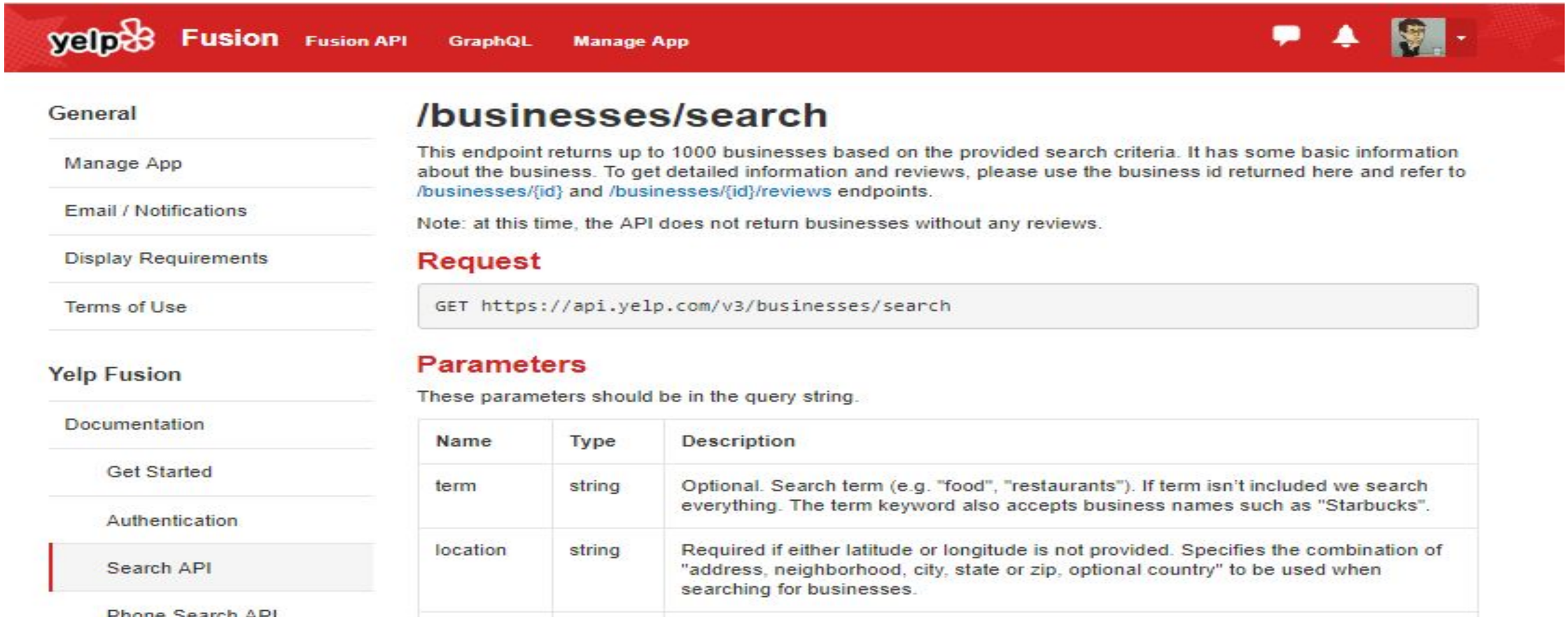
Step 4: Data Retrieval Plan



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)

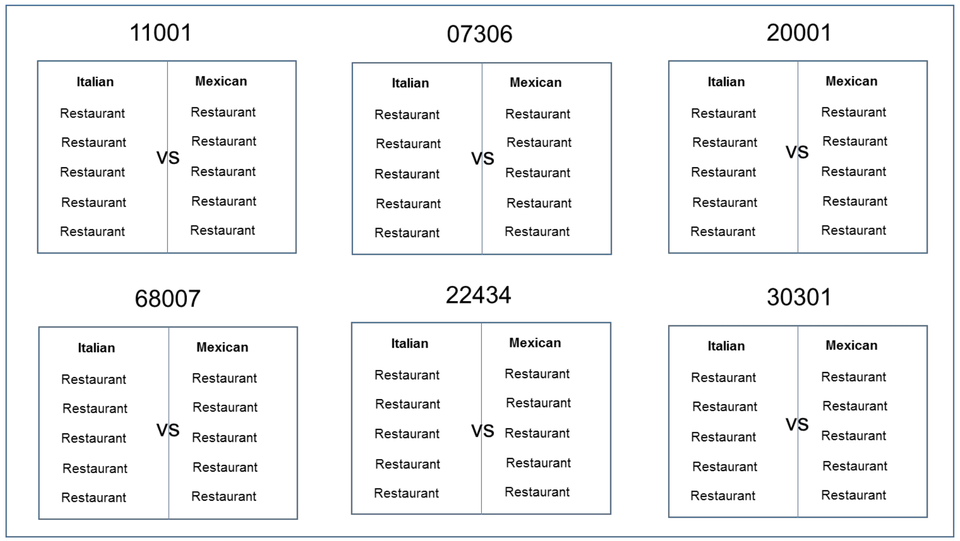
Thank You Yelp!

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.

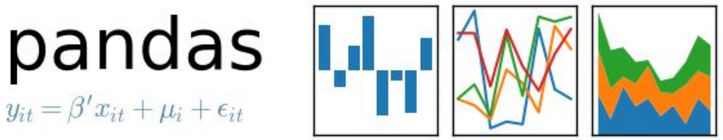
Step 5:
Retrieve the Data

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")
```

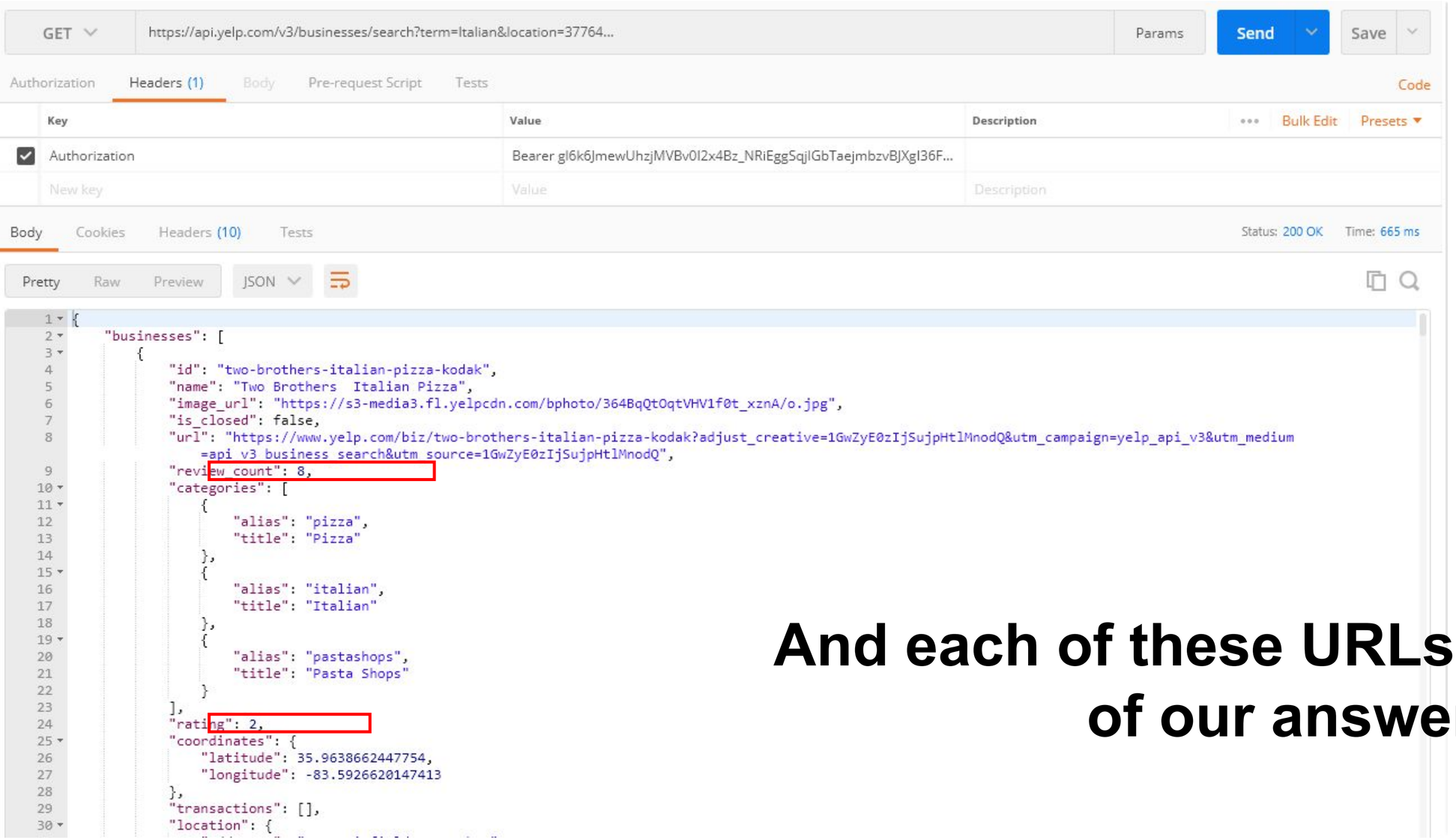
This funky code…

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)

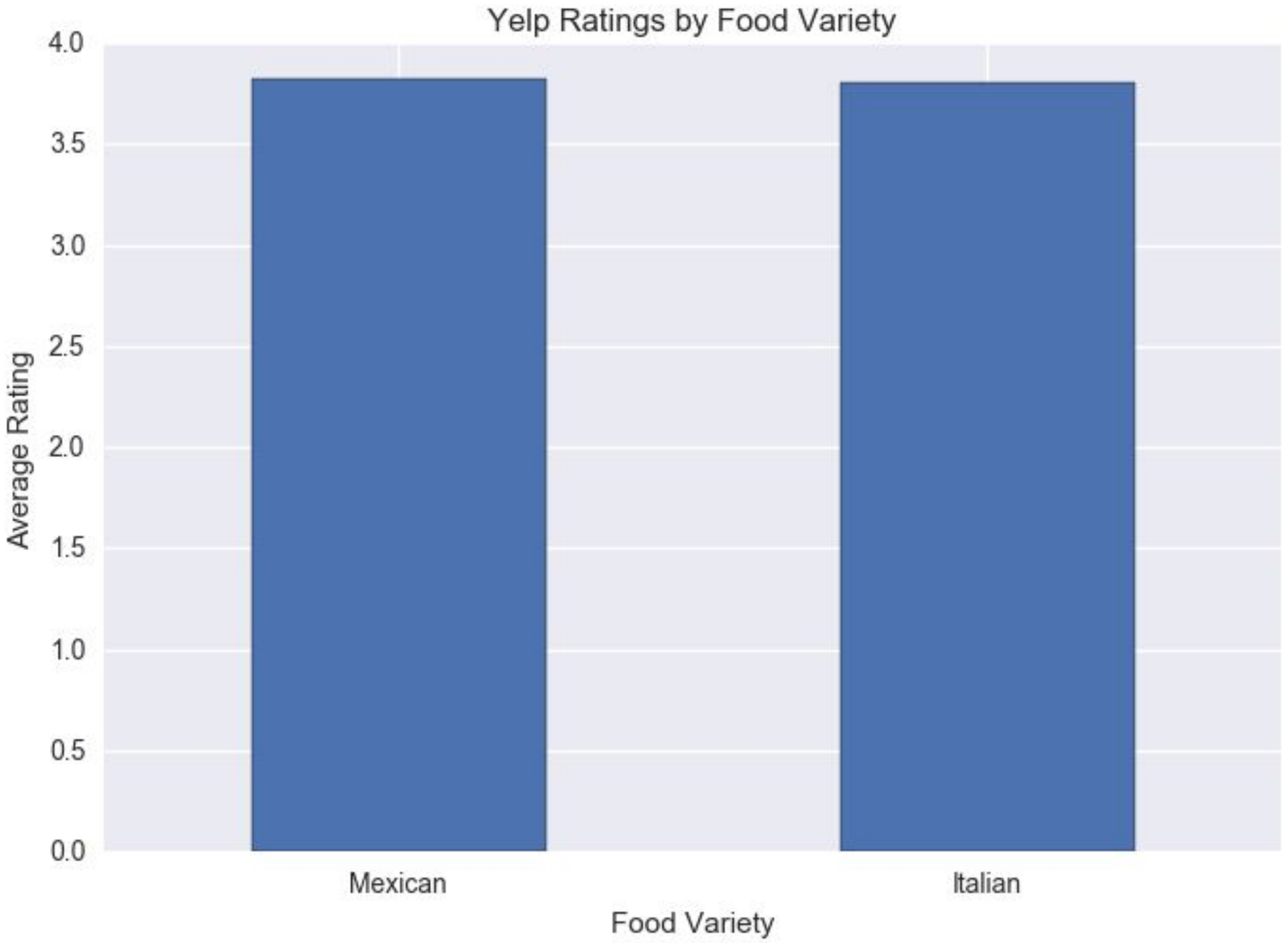
Display Summary of Results

```
# 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

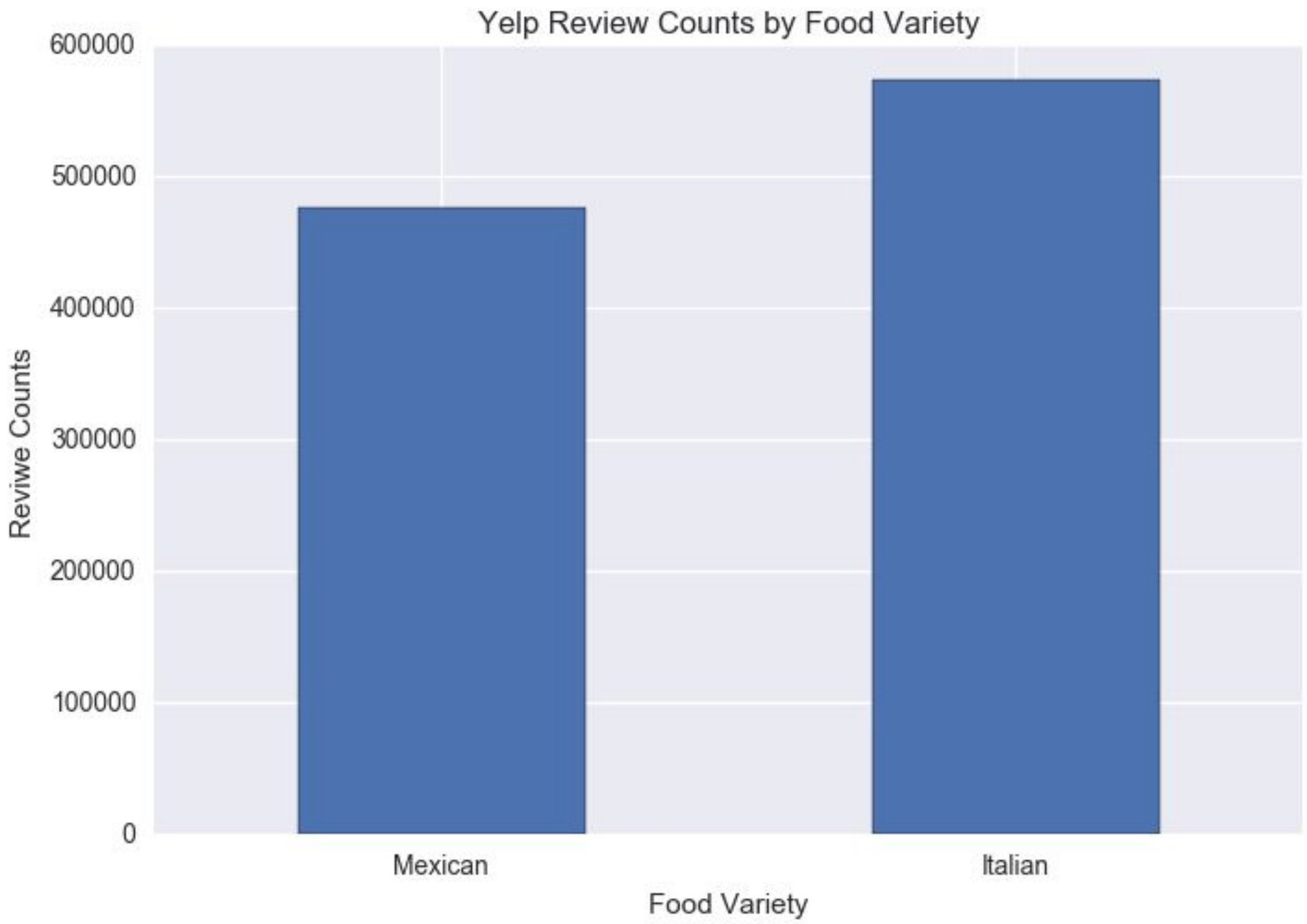
Ugh..
It's Close.

Analyze for Trends (Ratings)



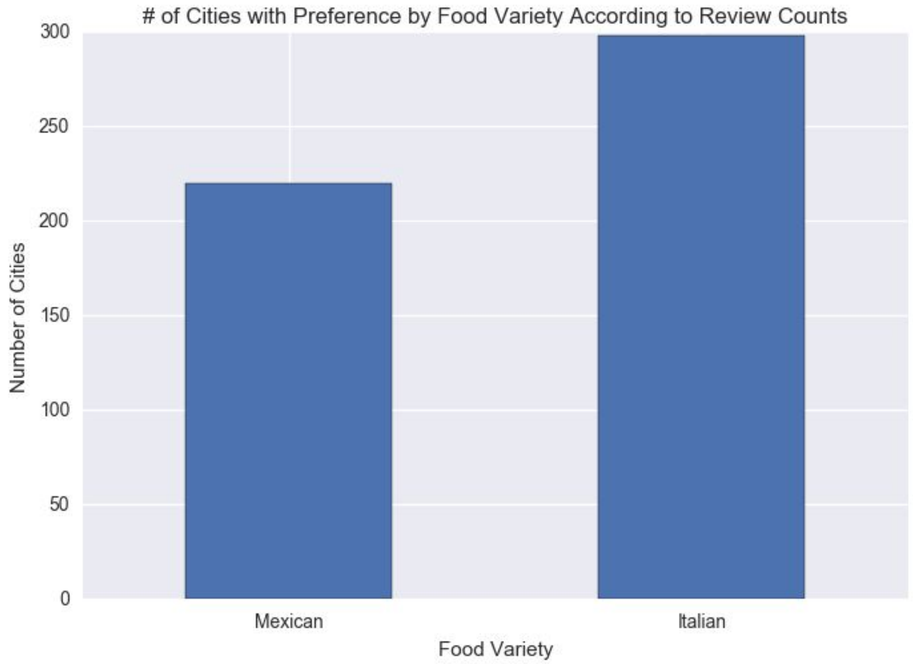
Yelpers Rate Italian and Mexican relatively **equally**.

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)

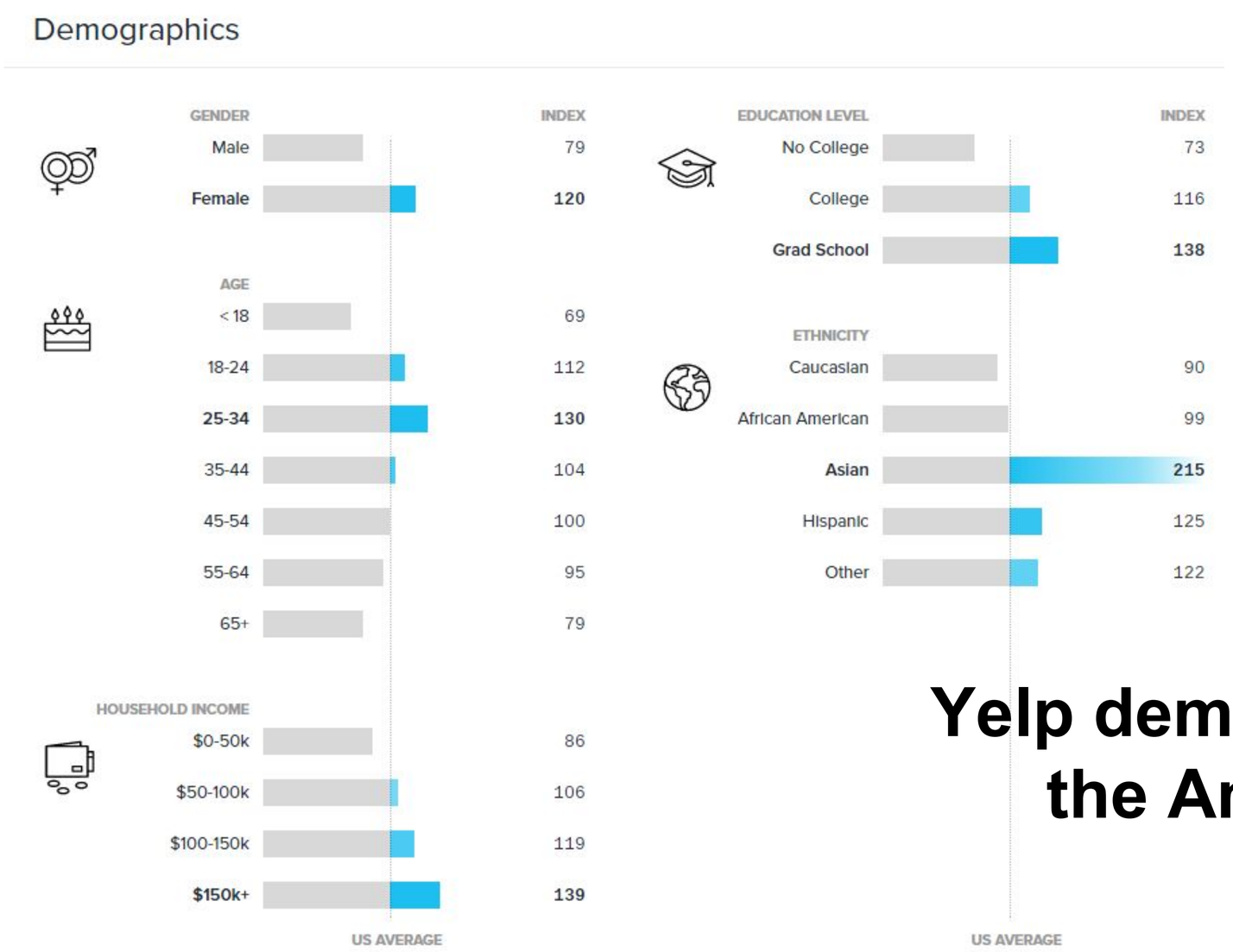
Metric	Italian	Mexican	p-Value (T-Test)
Average Rating	3.806	3.826	0.284
Review Counts	**573k**	476k	0.057

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.*

Step 8:
Acknowledge Limitations

Limitations in Analysis



Yelp demographics may not match the American demographic.

Restaurant experiences do not equate to home cooked meals.

Limitations in Analysis

“Fine” dining effect?

Step 9:
Make the Call

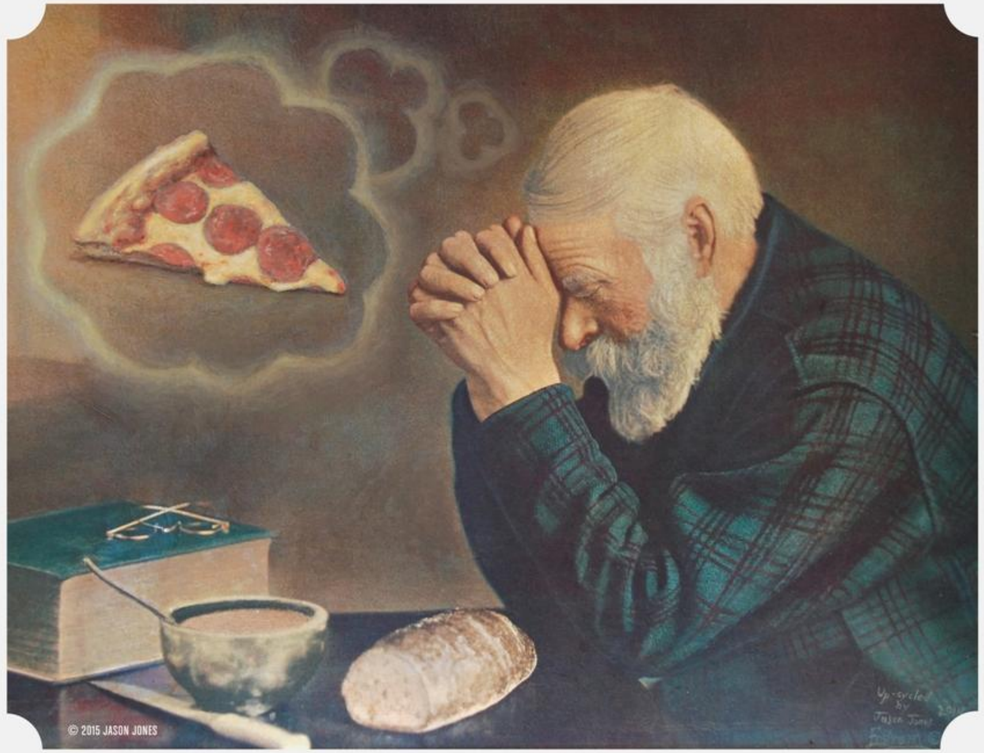
The “Proper” Conclusion:

“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:

Italian.

(But it’s going to be close…)



Activity*:* Analyze the Code **| Suggested Time:** 10 min

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 as you can.

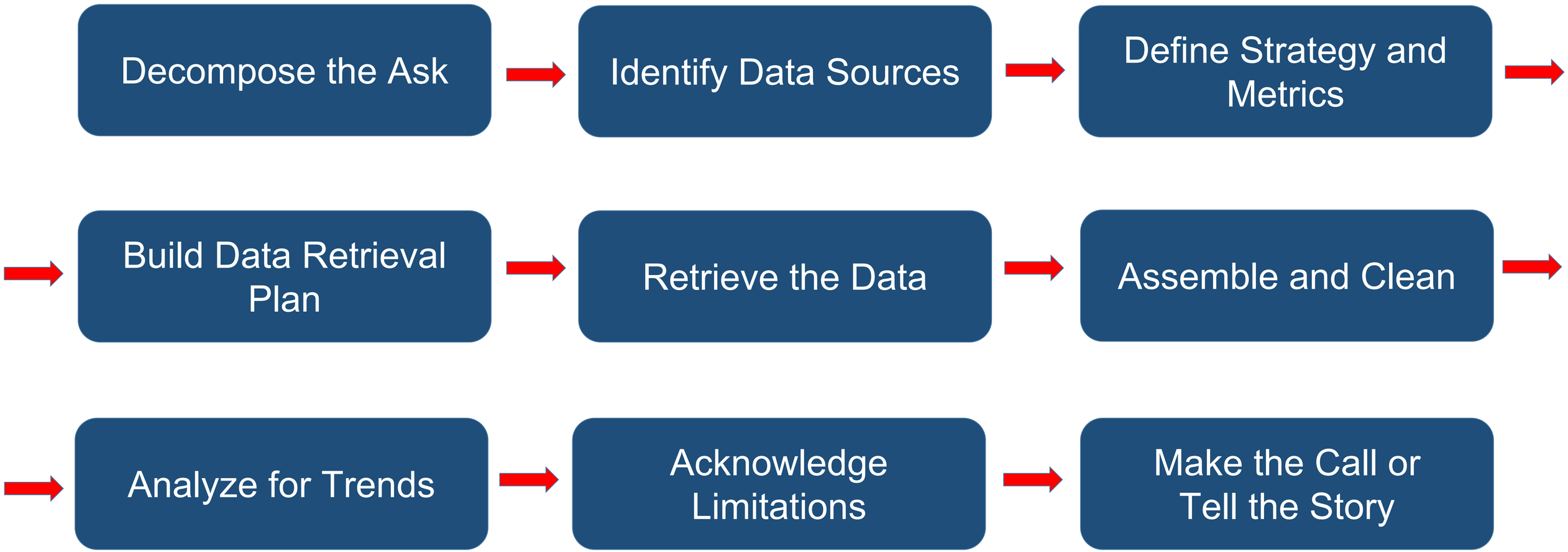
- If you are new to coding, your goal should be to understand what a single line does.

- 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.

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

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

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