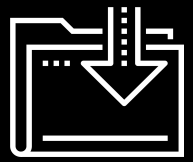
Introduction to Data

Data Boot Camp

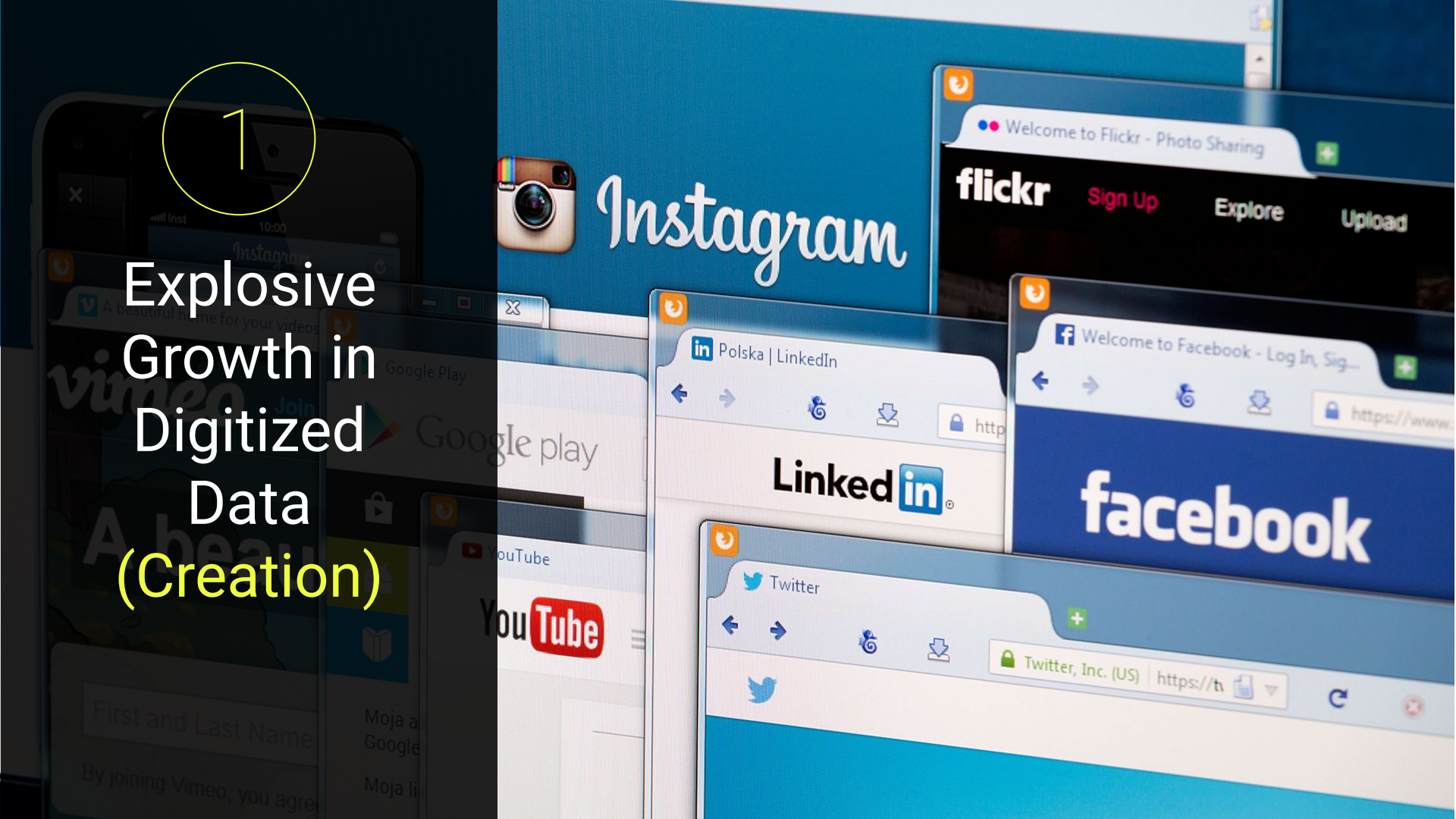
Lesson 1.1




WELCOME

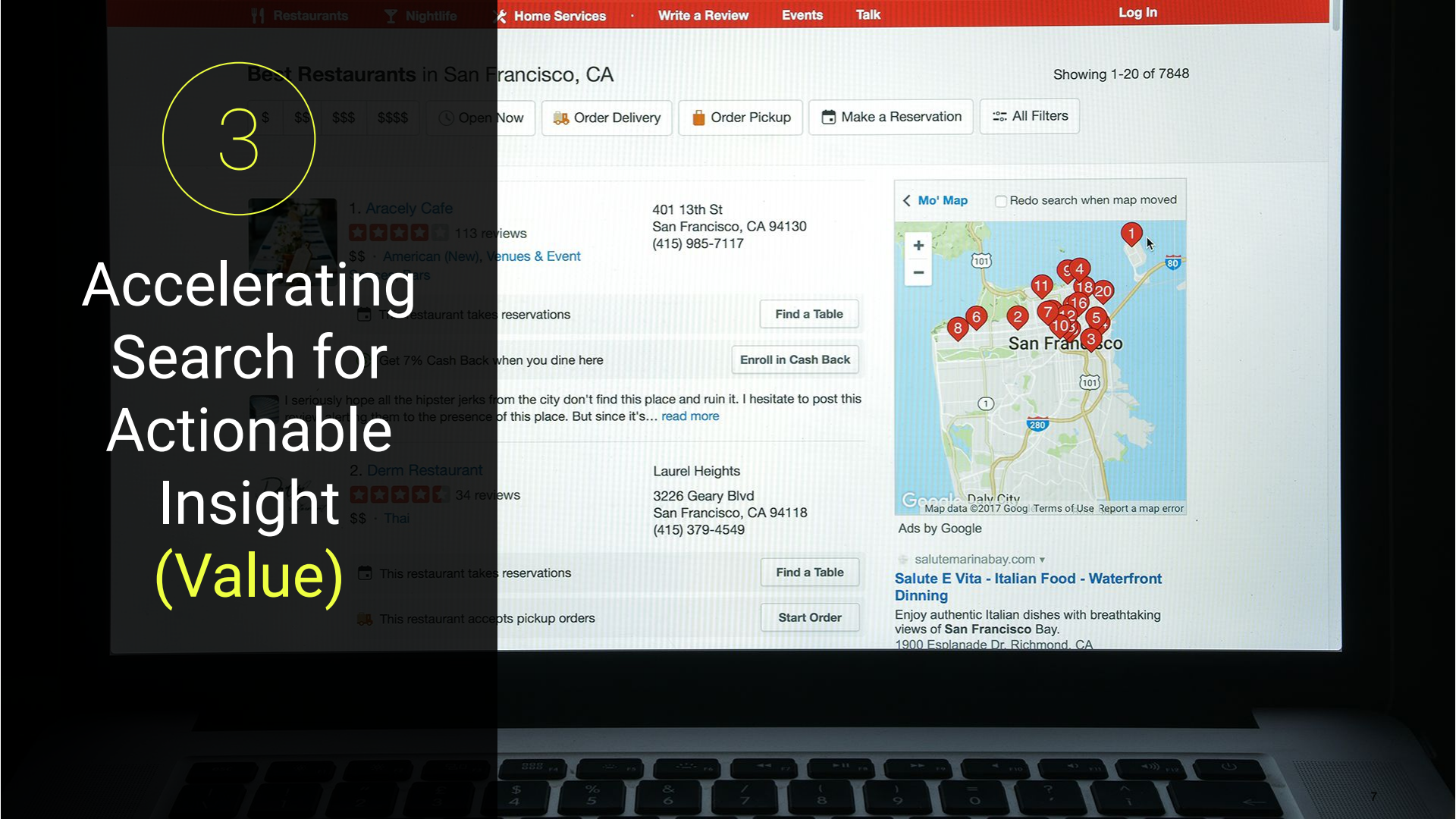
The Rise of Data

Why is data analytics such a hot skill these days?


1
Explosive Growth in Digitized Data (Creation)
Instagram
Welcome to Flickr - Photo Sharing
flickr
Sign Up
Explore
Upload
Polska | LinkedIn
LinkedIn
Welcome to Facebook - Log In, Sig...
facebook
Twitter
Twitter, Inc. (US)
YouTube
Google play
vimeo

2

Explosive Growth in Analytic Tools (Synthesis)



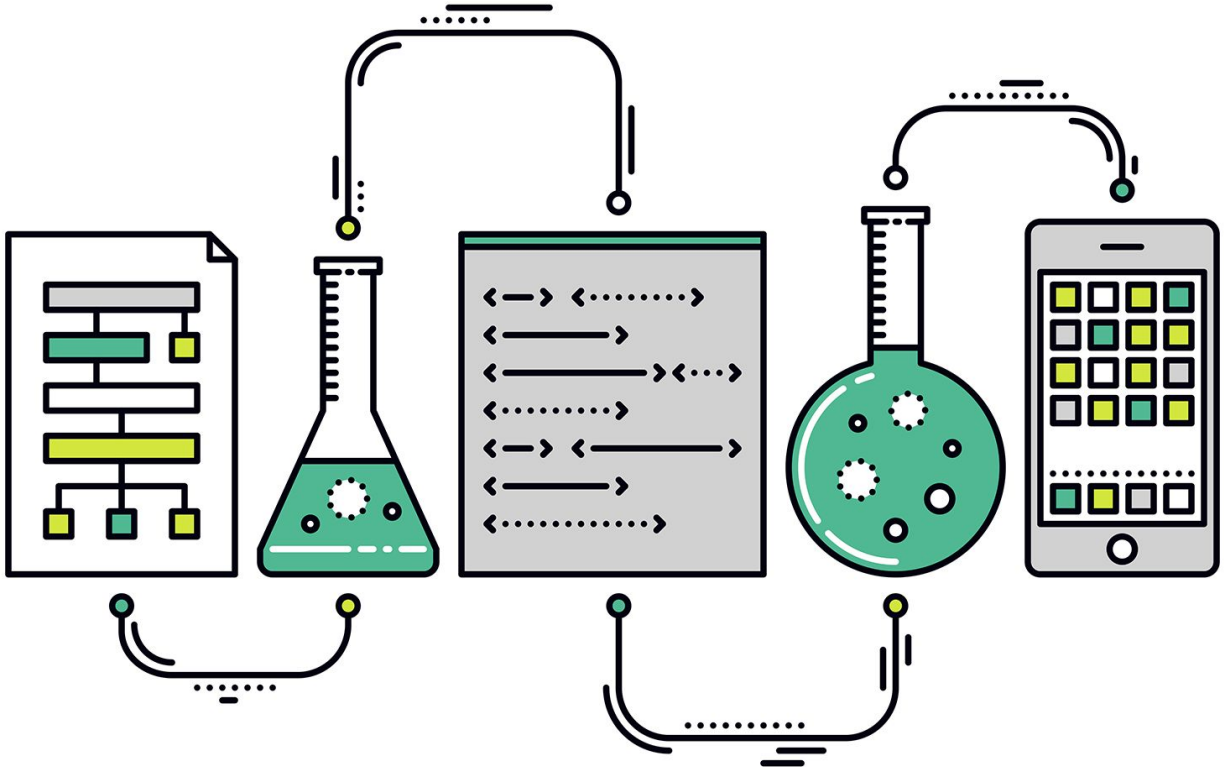
3

Accelerating Search for Actionable Insight (Value)

What does the term data science mean?

Perhaps you are picturing an Excel spreadsheet.

Data Science Involves Spreadsheets and Formulas



Fundamentally, data science is about storytelling and truth-telling.

Data as Storytelling

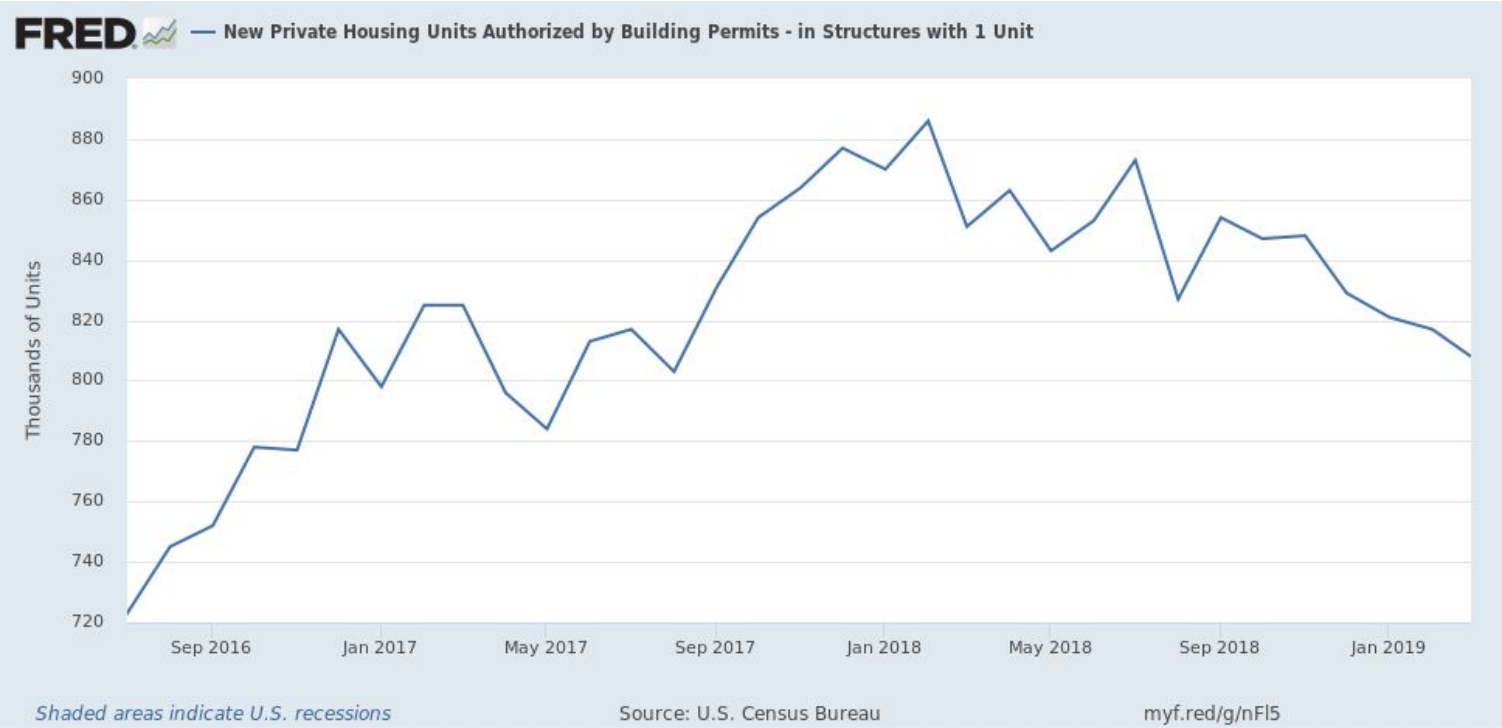
Data as Storytelling

U.S. Debt as Percentage of Gross Domestic Product, 1790–2011										
1790	29,6%	1835	0.0	1880	18.4	1925	21.6	The Great Depression & World War II	1970	28.0	Reagan Tax Cuts
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	26.4	
1806	10.0	1851	2.4	1896	8.5	1941	43.3		1986	38.5	

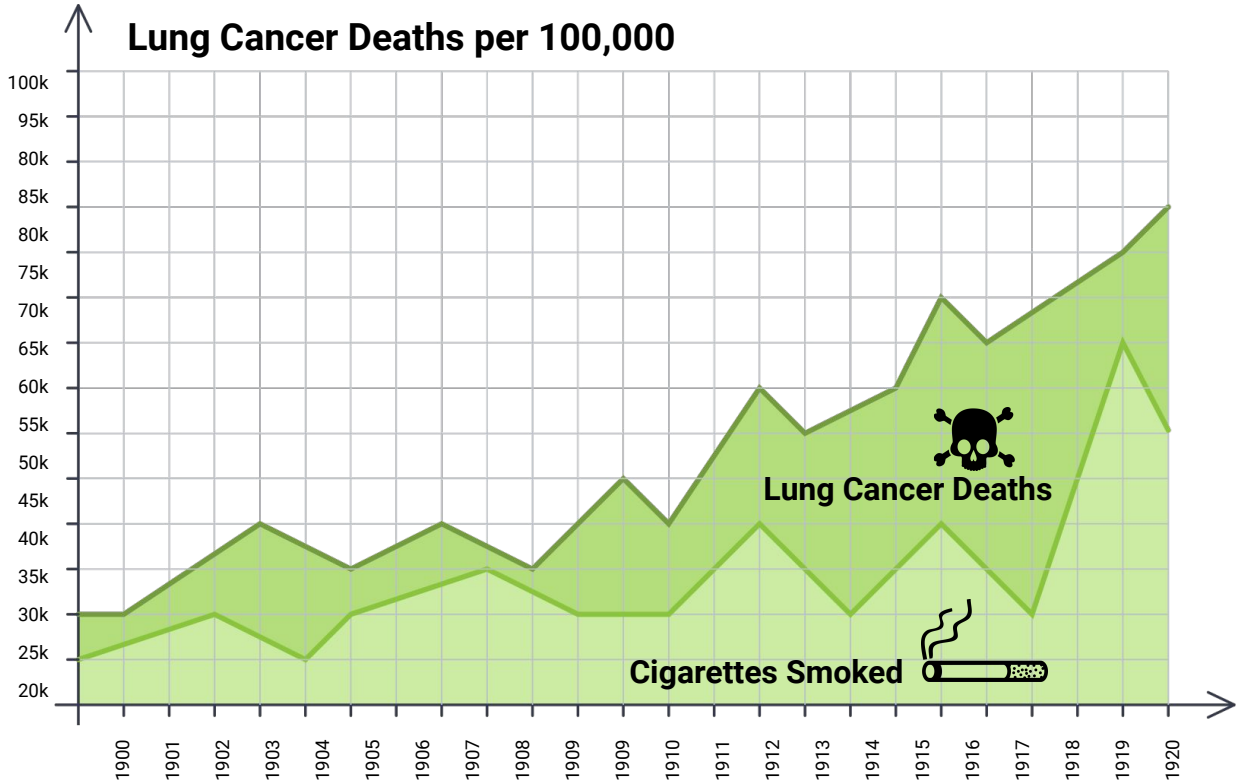
Data: Congressional Budget Office

Data = Drama

New Housing Construction: Making A Bottom, At Close To Recessionary Levels



Data as Truth-Telling



Data as Truth-Telling

Unearthing Relationships

Data as Truth-Telling

Making Predictions

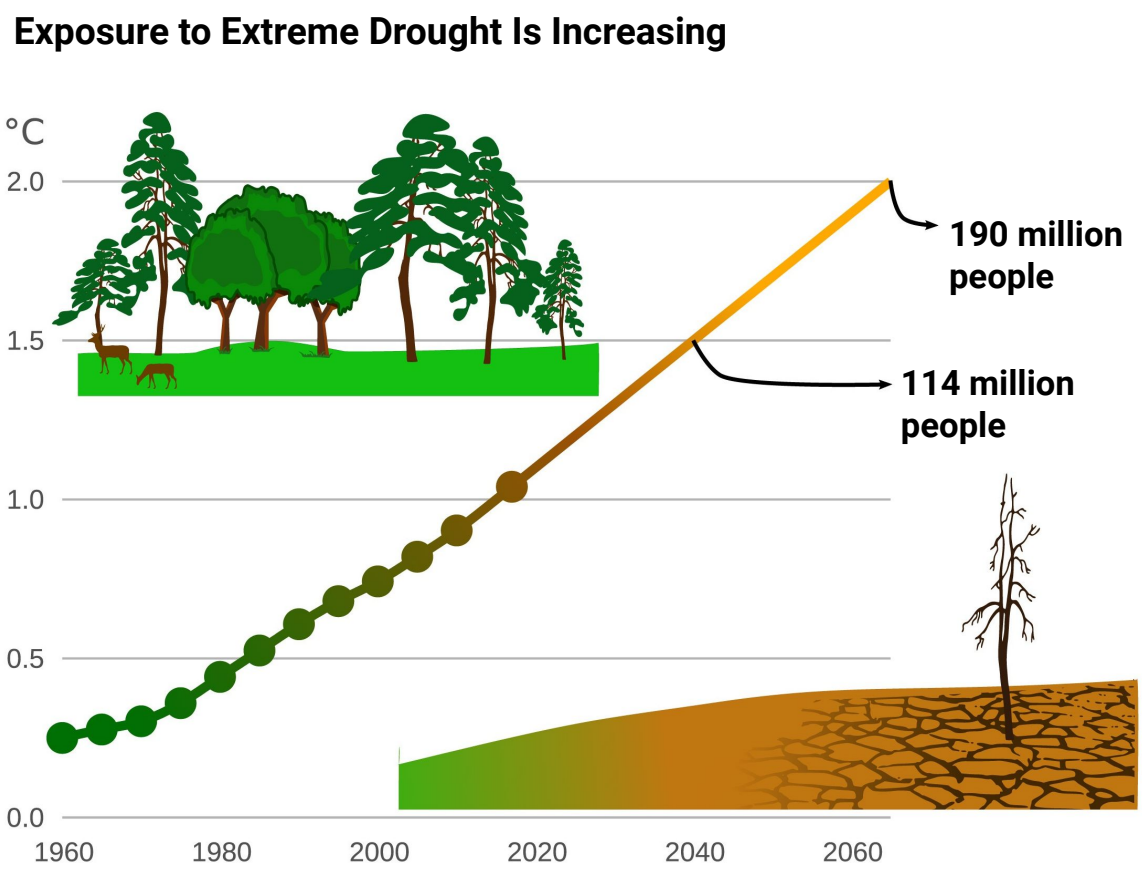


Image Source: [Exposure to Extreme Drought is Increasing]

Data as Truth-Telling

Stating Significance

Course Overview

Tools for Truths, Skills for Stories:

Our Goals:

Truth-telling

Storytelling

Our Means:

Microsoft Excel
Python
pandas
Matplotlib/Seaborn
APIs
Beautiful Soup
Machine Learning

SQL
MongoDB
HTML/CSS
JavaScript
D3.js
Leaflet.js/Google Maps
Tableau
Hadoop

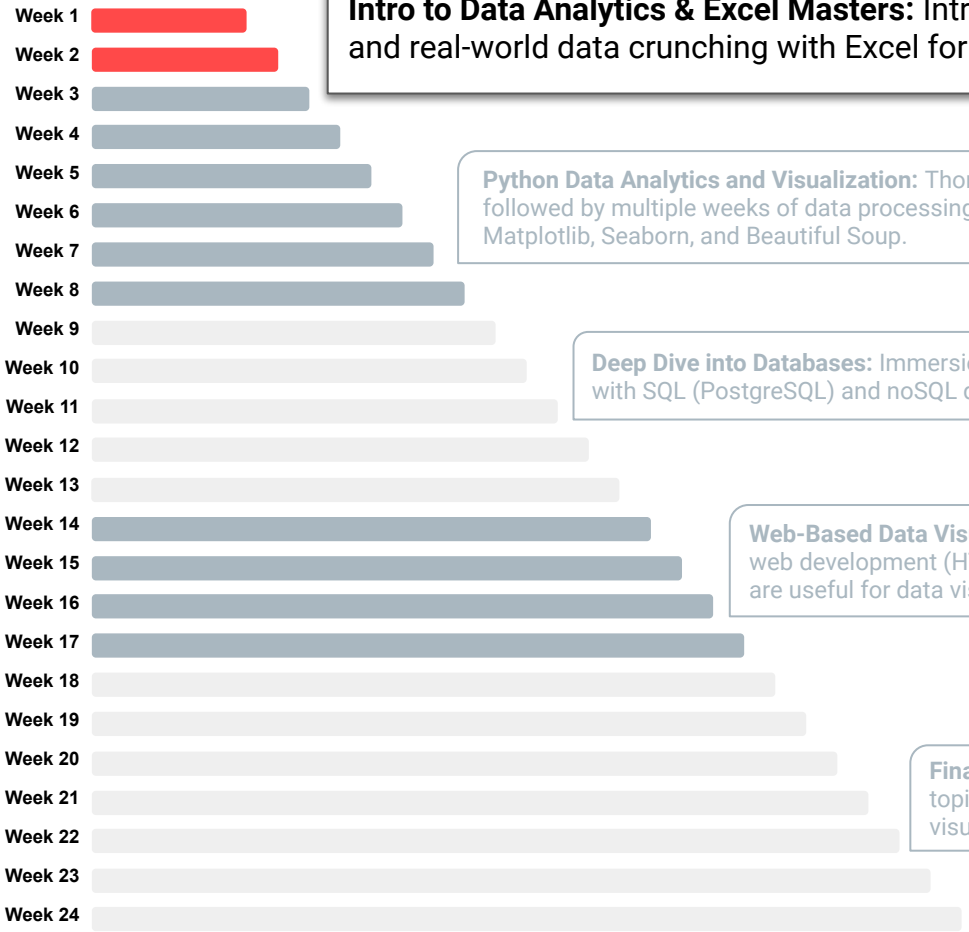
Course Overview

Each class will include the following:

- Overview of Lesson Topics
- Instructor Lecture
- Instructor Demonstration
- Class Discussions
- In-Class Activities
- Project Work

Weekly Breakdown by Subject

Weeks 1–2



Intro to Data Analytics & Excel Masters: Introduction to the high-level concepts of data analytics and real-world data crunching with Excel formulas, pivot tables, and conditional formatting.

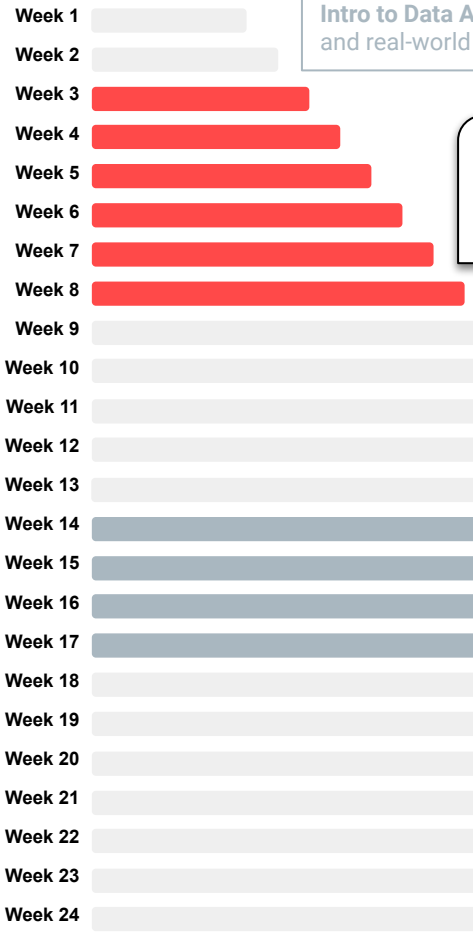
Python Data Analytics and Visualization: Thorough crash course on Python programming, followed by multiple weeks of data processing using advanced libraries like NumPy, pandas, Matplotlib, Seaborn, and Beautiful Soup.

Deep Dive into Databases: Immersion into introductory and advanced work with SQL (PostgreSQL) and noSQL databases (MongoDB).

Web-Based Data Visualization: Introduction to the fundamental tools of web development (HTML, CSS, JavaScript) and advanced libraries that are useful for data visualization (D3.js, Leaflet.js), and R.

Final Projects & Advanced Topics: Introduction to Tableau and advanced topics like Hadoop and Machine Learning. Develop a real-world data visualization project.

Weeks 3–8



Intro to Data Analytics & Excel Masters: Introduction to the high-level concepts of data analytics and real-world data crunching with Excel formulas, pivot tables, and conditional formatting.

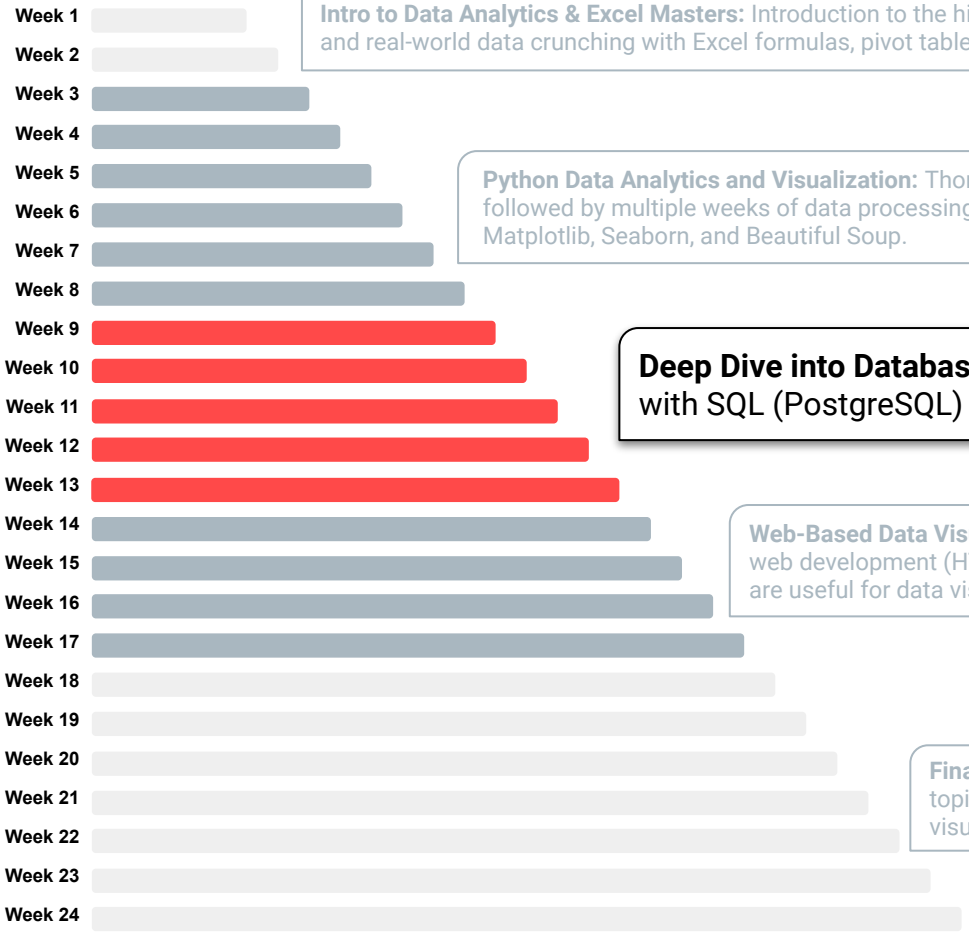
Python Data Analytics and Visualization: Thorough crash course on Python programming, followed by multiple weeks of data processing using advanced libraries like NumPy, pandas, Matplotlib, Seaborn, and Beautiful Soup.

Deep Dive into Databases: Immersion into introductory and advanced work with SQL (PostgreSQL) and noSQL databases (MongoDB).

Web-Based Data Visualization: Introduction to the fundamental tools of web development (HTML, CSS, JavaScript) and advanced libraries that are useful for data visualization (D3.js, Leaflet.js), and R.

Final Projects & Advanced Topics: Introduction to Tableau and advanced topics like Hadoop and Machine Learning. Develop a real-world data visualization project.

Weeks 9–13



Intro to Data Analytics & Excel Masters: Introduction to the high-level concepts of data analytics and real-world data crunching with Excel formulas, pivot tables, and conditional formatting.

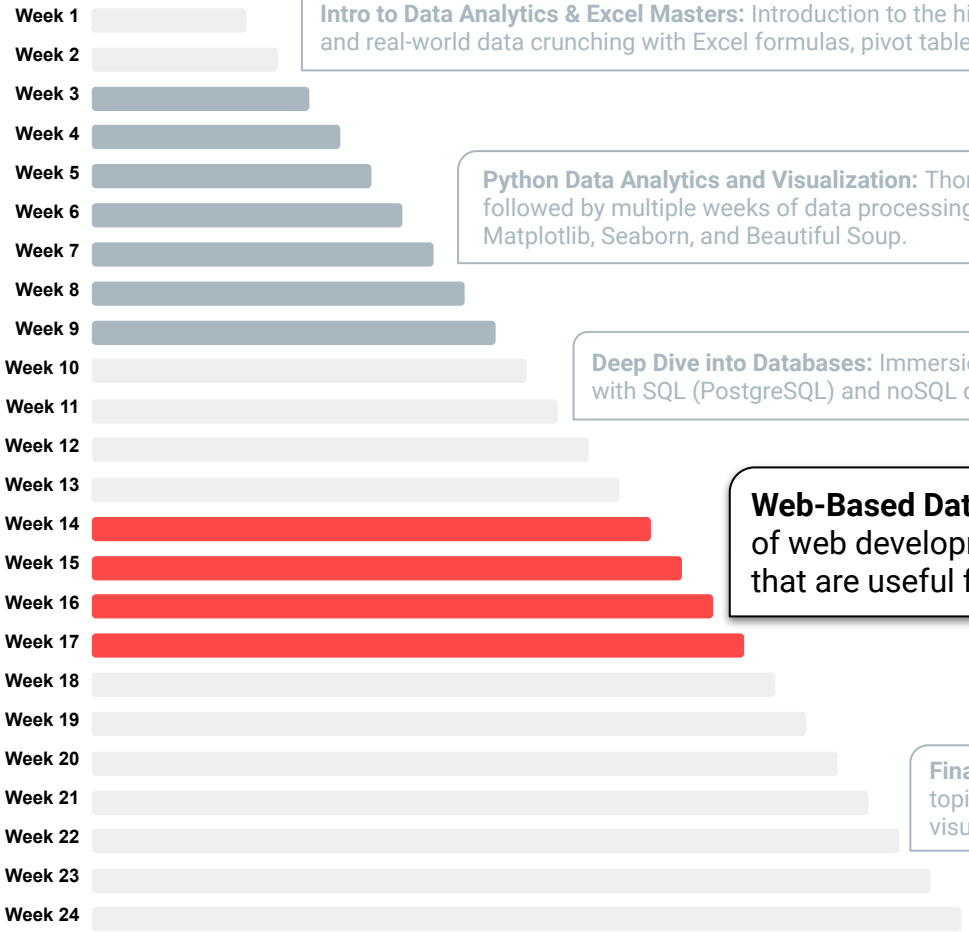
Python Data Analytics and Visualization: Thorough crash course on Python programming, followed by multiple weeks of data processing using advanced libraries like NumPy, pandas, Matplotlib, Seaborn, and Beautiful Soup.

Deep Dive into Databases: Immersion into introductory and advanced work with SQL (PostgreSQL) and noSQL databases (MongoDB).

Web-Based Data Visualization: Introduction to the fundamental tools of web development (HTML, CSS, JavaScript) and advanced libraries that are useful for data visualization (D3.js, Leaflet.js), and R.

Final Projects & Advanced Topics: Introduction to Tableau and advanced topics like Hadoop and Machine Learning. Develop a real-world data visualization project.

Weeks 14–17



Intro to Data Analytics & Excel Masters: Introduction to the high-level concepts of data analytics and real-world data crunching with Excel formulas, pivot tables, and conditional formatting.

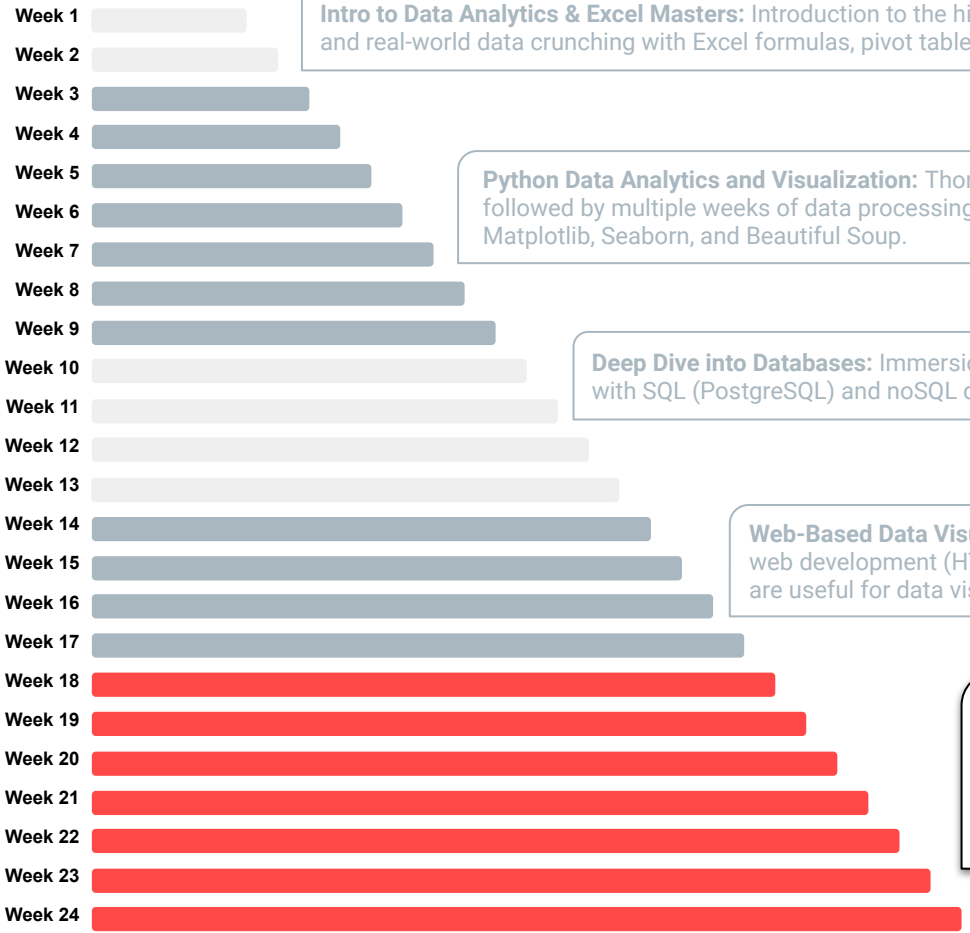
Python Data Analytics and Visualization: Thorough crash course on Python programming, followed by multiple weeks of data processing using advanced libraries like NumPy, pandas, Matplotlib, Seaborn, and Beautiful Soup.

Deep Dive into Databases: Immersion into introductory and advanced work with SQL (PostgreSQL) and noSQL databases (MongoDB).

Web-Based Data Visualization: Introduction to the fundamental tools of web development (HTML, CSS, JavaScript) and advanced libraries that are useful for data visualization (D3.js, Leaflet.js), and R.

Final Projects & Advanced Topics: Introduction to Tableau and advanced topics like Hadoop and Machine Learning. Develop a real-world data visualization project.

Weeks 18–24



Intro to Data Analytics & Excel Masters: Introduction to the high-level concepts of data analytics and real-world data crunching with Excel formulas, pivot tables, and conditional formatting.

Python Data Analytics and Visualization: Thorough crash course on Python programming, followed by multiple weeks of data processing using advanced libraries like NumPy, pandas, Matplotlib, Seaborn, and Beautiful Soup.

Deep Dive into Databases: Immersion into introductory and advanced work with SQL (PostgreSQL) and noSQL databases (MongoDB).

Web-Based Data Visualization: Introduction to the fundamental tools of web development (HTML, CSS, JavaScript) and advanced libraries that are useful for data visualization (D3.js, Leaflet.js), and R.

Final Projects & Advanced Topics: Introduction to Tableau and advanced topics like Hadoop and Machine Learning. Develop a real-world data visualization project.

Example Activity

Example Activity: Banking Deserts

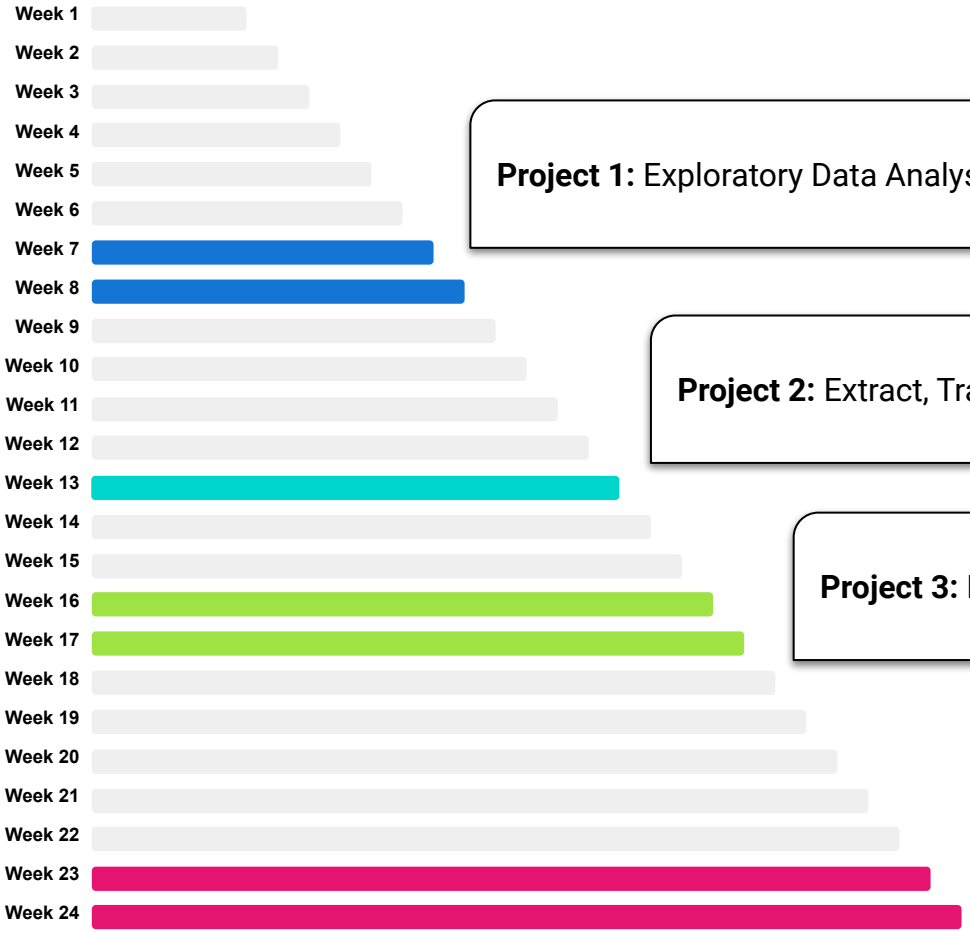
In this activity, you will use a variety of public demographic data and APIs to explain many real-world social phenomena. Utilize data from sources like the U.S. Census, Google Maps, and more to find insights on poverty, discrimination, and the impact of changing economies.

Suggested Time:

20 minutes

Projects

Projects



Project 1: Exploratory Data Analysis

Project 2: Extract, Transform & Load

Project 3: Data Visualizations

Final Project: Machine Learning

Project Specialization Tracks

Tackle each project with a specialized track!

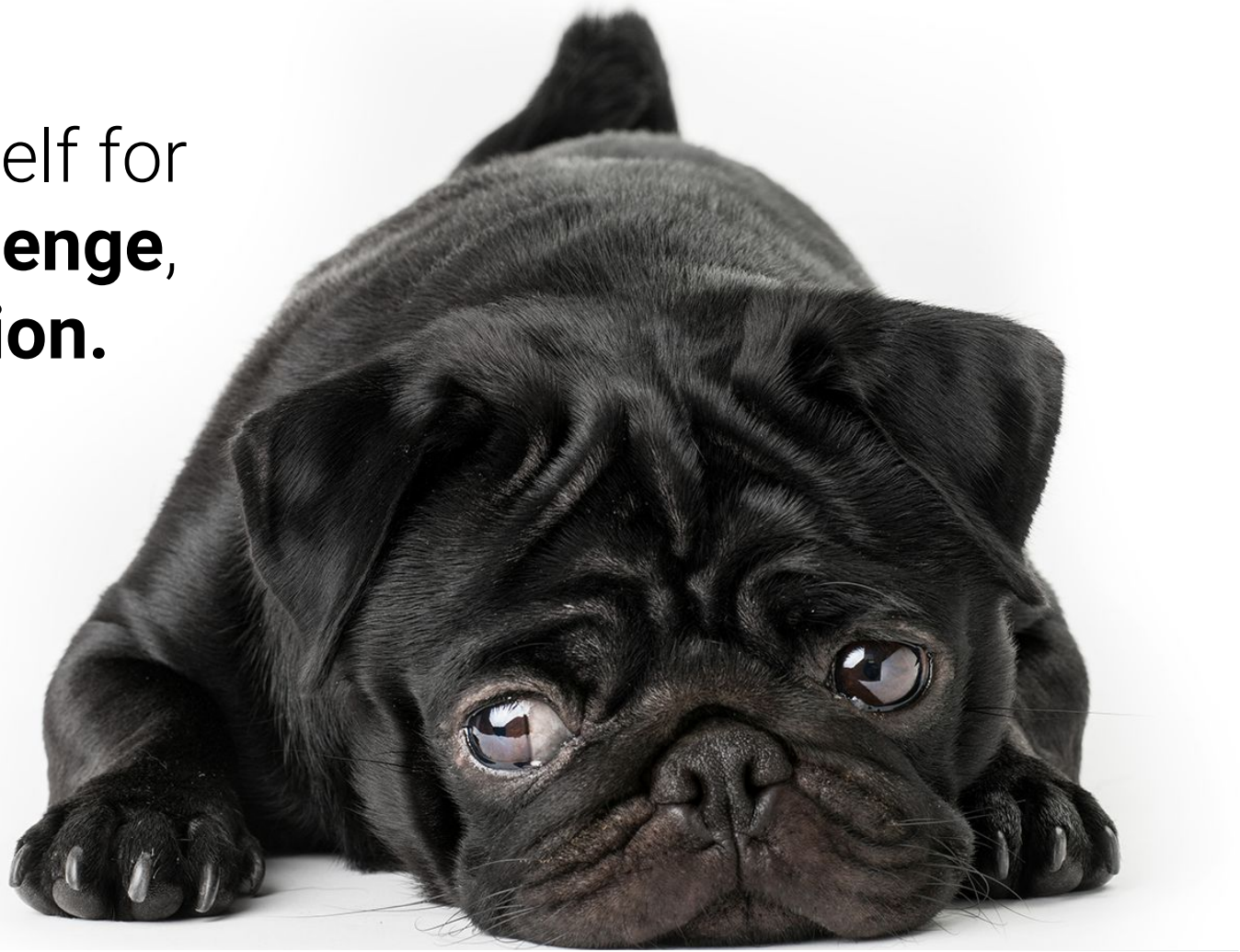
Choose your focus:

- Finance
- Healthcare
- Custom

Helpful Tips


Embrace your
inner toddler.

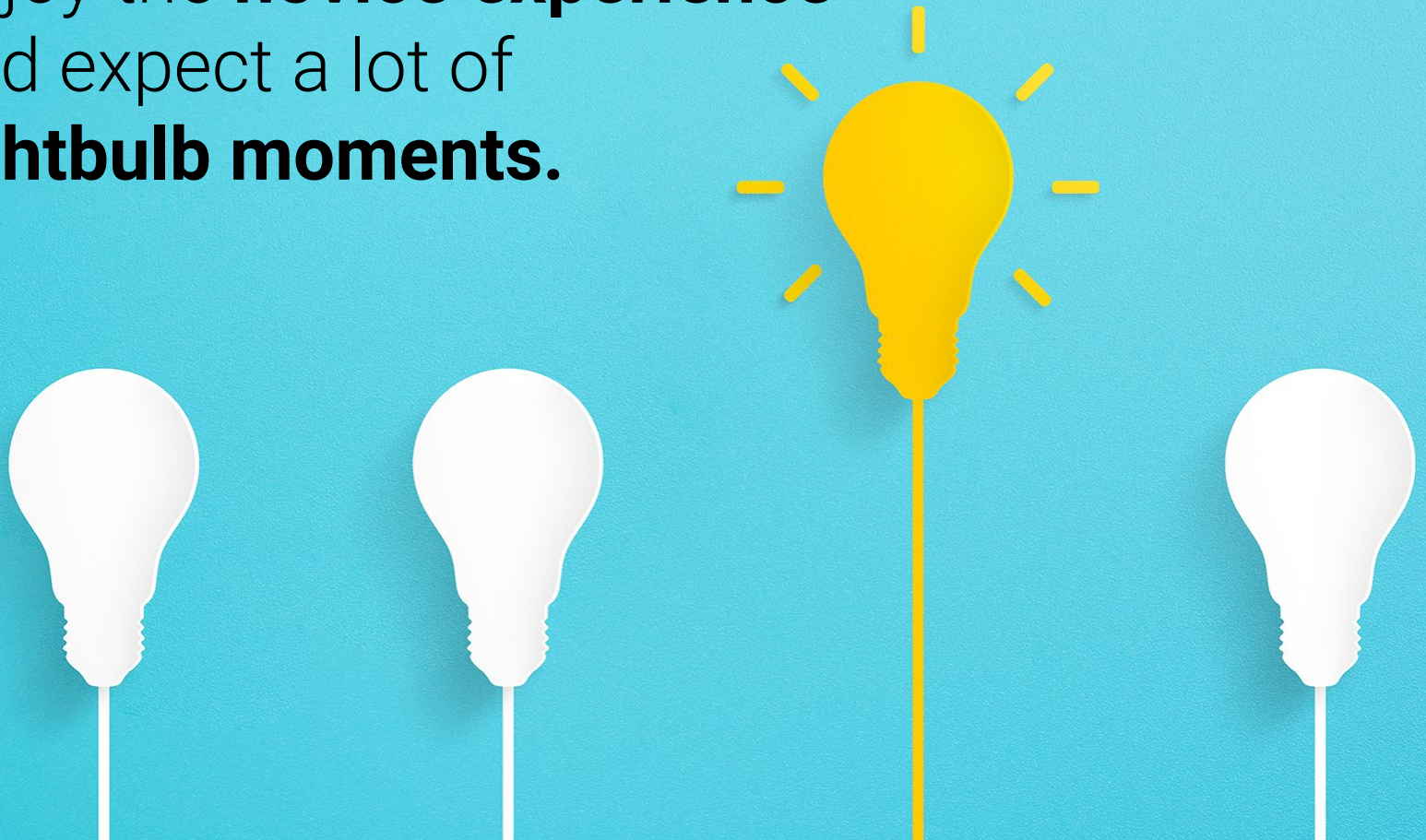
Brace yourself for **doubt**, **challenge**, and **confusion.**



There is no shortcut.
You've got to **put in the hours!**


Form a community
with your classmates.

Enjoy the **novice experience** and expect a lot of **lightbulb moments.**




Celebrate your successes!

Group Activity:

Form groups of 3 or 4 people. Introduce yourself to your group.

Get up from your seats and walk around.

Don’t be shy!

Suggested Time:

10 minutes


Break

Group Activity: The Great Debate

Find the group you formed before the break.
Together, ponder the following question.

Suggested Time:

20 minutes

Group Activity: The Great Debate

Which food do Americans prefer:
Italian or Mexican food?

Group Activity: The Great Debate

With your group, develop a 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.

The Great Debate (Analyzed)

Step 1: Decompose the “Ask”

Step 1: Decompose the “Ask”

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

Step 1: Decompose the “Ask”

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

Who exactly is an **American**?

Are **Americans** just homeowners?

Do **Americans** just live in big cities?

Are **Americans** just millennials?

How can we get a representative sampl 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?

- 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?

Italian and Mexican are broad categories to pursue. We will have to narrow the scope.

Step 1: Decompose the “Ask”

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

01

How do we categorize foods? Is pizza Italian? Is Taco Bell Mexican?

02

How do we categorize food? Does making pasta at home constitute Italian? Or are we just talking about restaurants?

03

Are we just talking about “best experiences”?
Or are we including poorer versions of these foods?

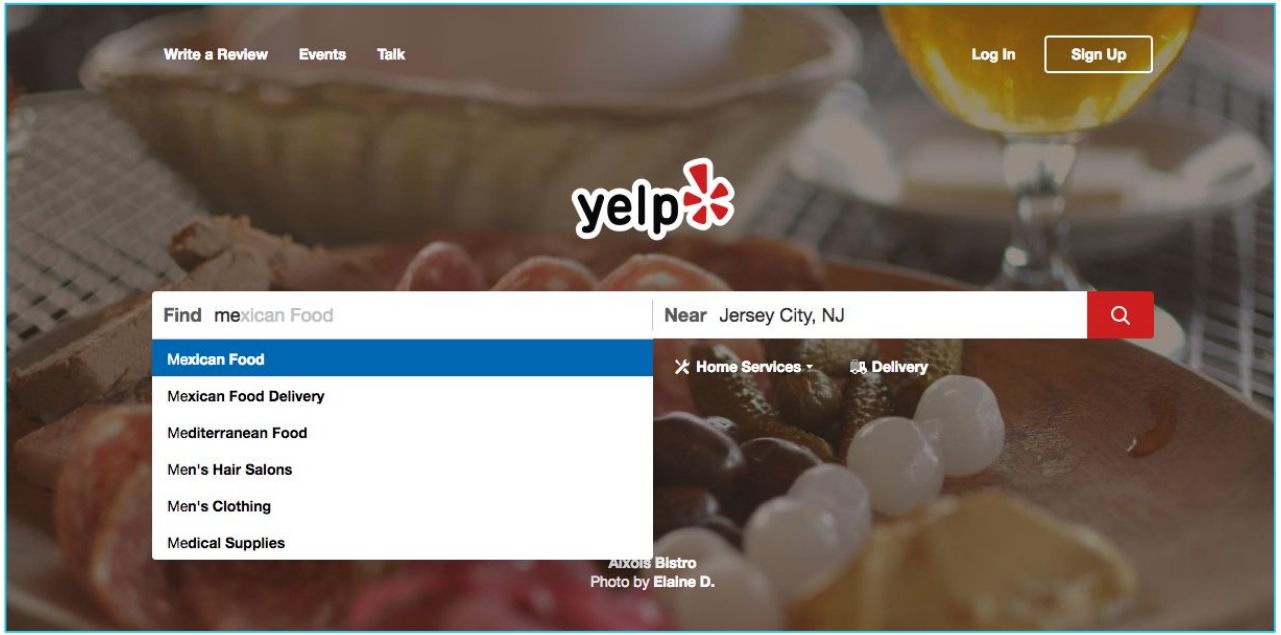
Step 2: Identify Data Sources

Step 2: Identify Data Sources

Why poll an audience when there are already enormous databases of information about Americans' food preferences—readily available online?

Step 2: Identify Data Sources

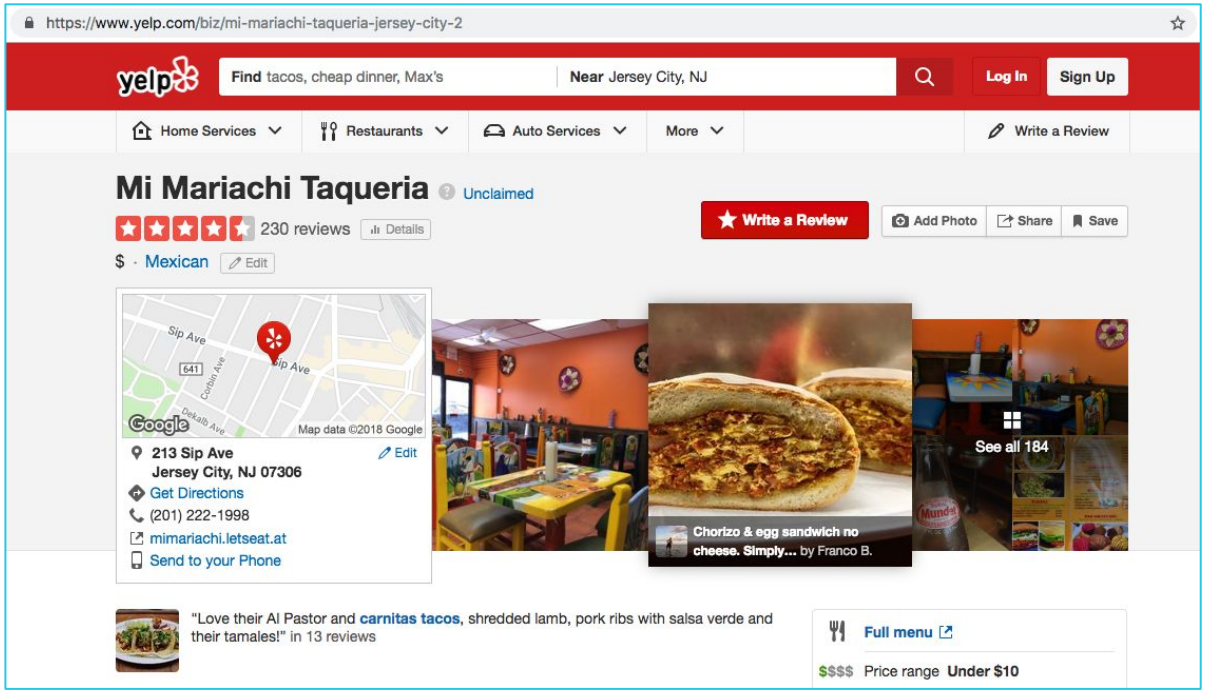
As everyday consumers, we are **regularly** getting a pulse of everyday American food preferences to inform our own decisions. Perhaps we can make use of the same approach.



www.yelp.com

Step 2a: Identify Data Sources

Web services like Yelp provide an almost-encyclopedic amount of information about the eating preferences of Americans.



Step 3: Define Strategy and Metrics

Step 3: Define Strategy and Metrics

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

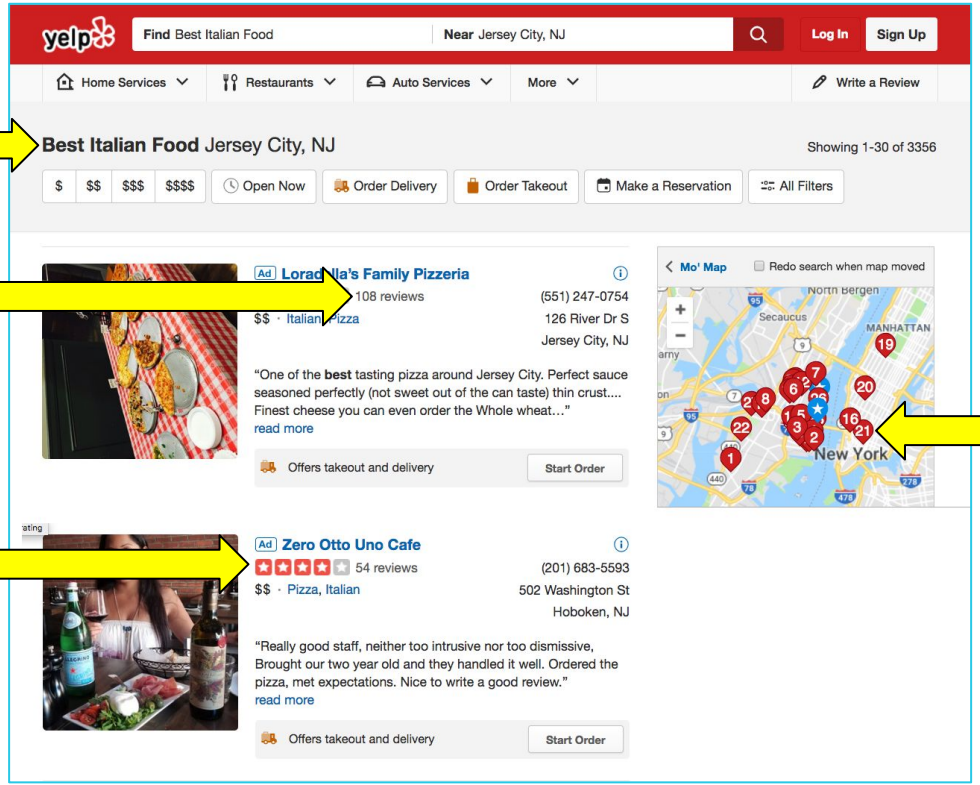
Americans	• Ideally, we need thousands of records from Americans in hundreds of cities. (Large samples)
Preference	• Number of Yelp reviews (More = Preference) • Average aggregated ratings (Higher = Preference)
Italian and Mexican Food	• Top-20 Italian restaurants and top-20 Mexican restaurants in every city

Step 3: Define Strategy and Metrics

Food Type

Review Count

Rating



Locations

Lots of Data!

Step 3: Define Strategy and Metrics

Repeat this analysis for as many cities as possible.

New York, NY	
Italian	**Mexican**
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

VS.

Tucson, AZ	
Italian	**Mexican**
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

VS.

Washington, D.C.	
Italian	**Mexican**
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

VS.

Omaha, NE	
Italian	**Mexican**
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

VS.

San Diego, CA	
Italian	**Mexican**
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

VS.

Atlanta, GA	
Italian	**Mexican**
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

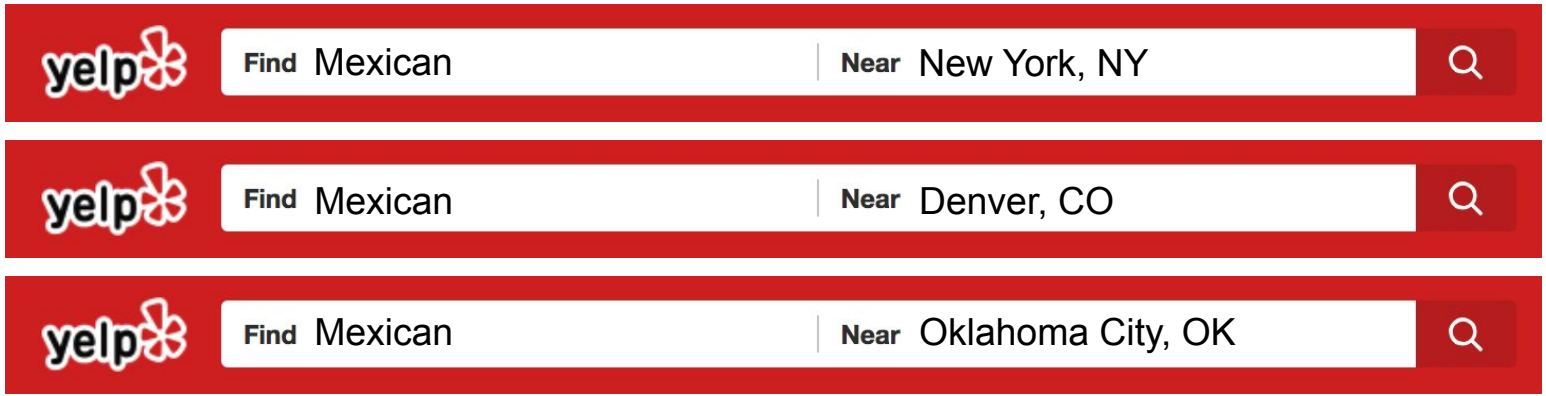
VS.

Step 4: Build Data Retrieval Plan

Step 4: Build Data Retrieval Plan

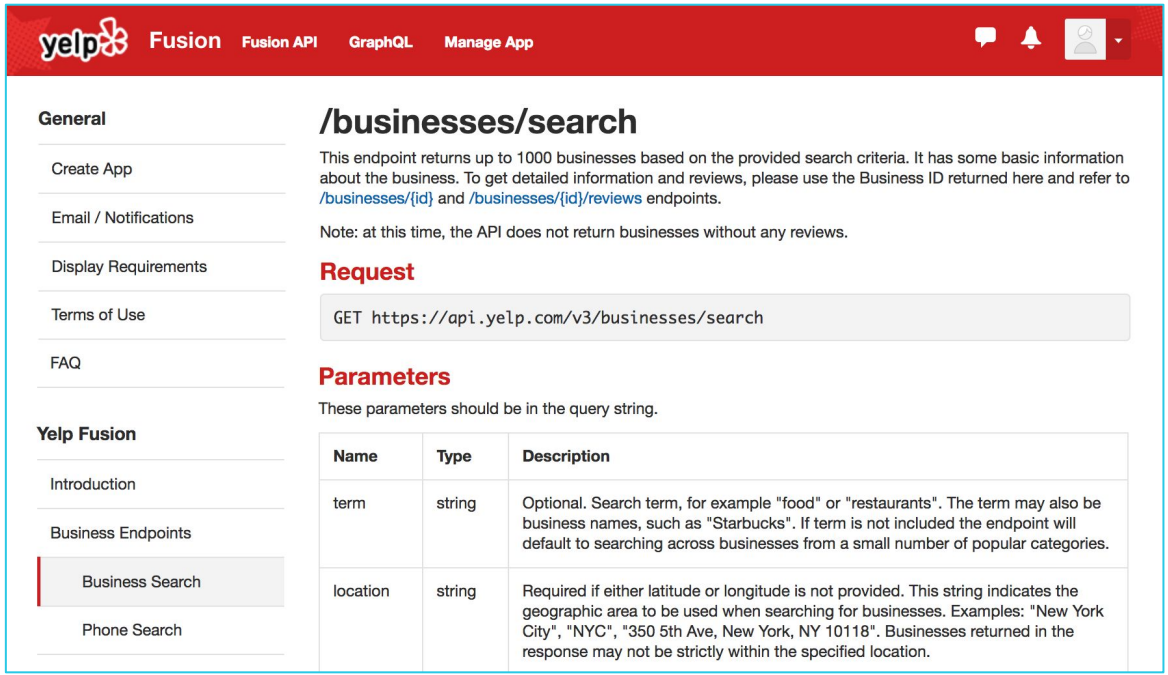
We could retrieve this data by brute force, but it would be:

- Extremely time-consuming
- Skewed by our city familiarity
- Labor intensive



Thank You, Yelp!

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



Thank You, Yelp!

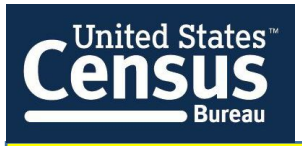
Response Body


```
{
  "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
    }
  }
}
```


yelp Fusion

Step 4: Build Data Retrieval Plan

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



11101	
Italian	Mexican
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

07360	
Italian	Mexican
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

20001	
Italian	Mexican
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

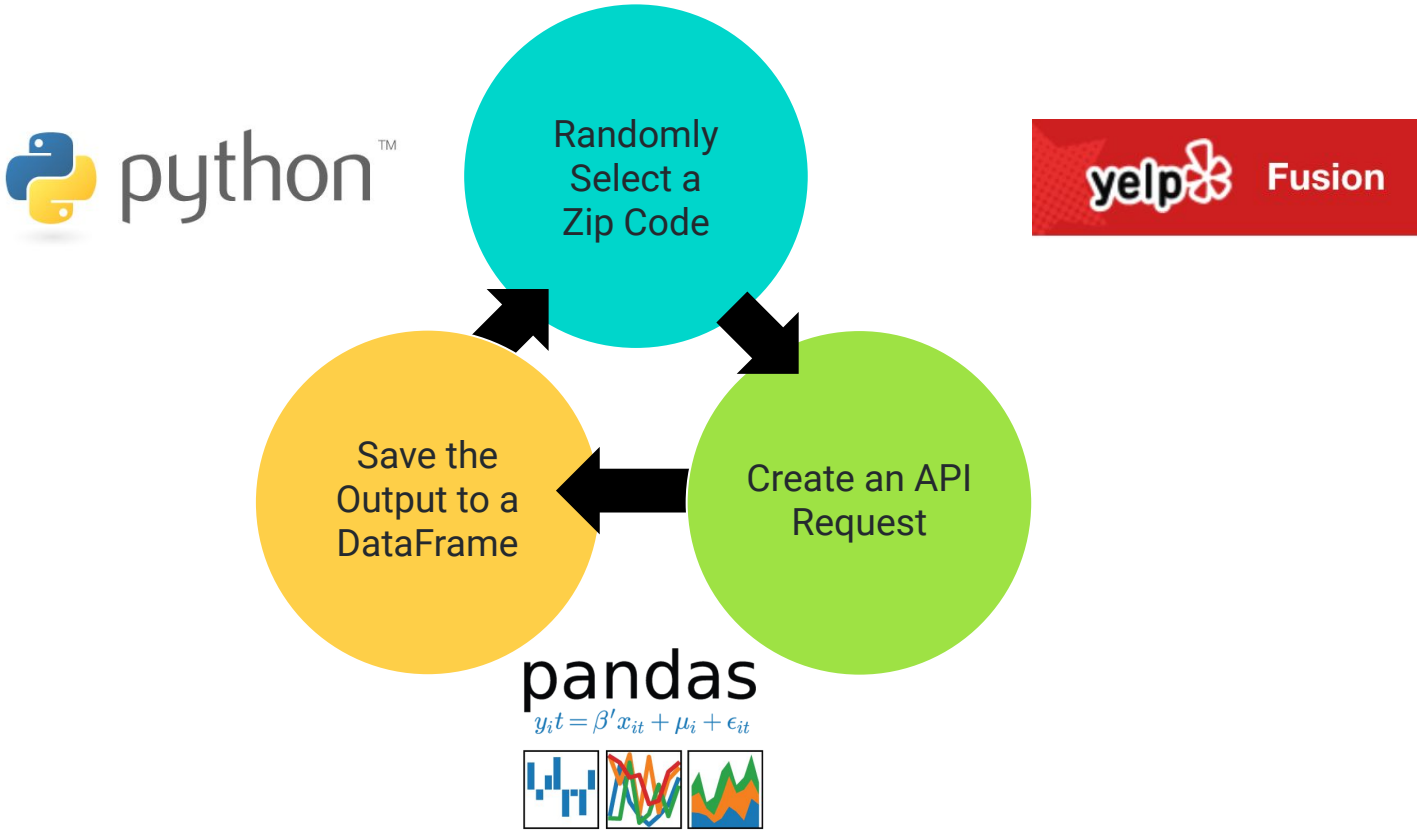
68007	
Italian	Mexican
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

22434	
Italian	Mexican
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

30301	
Italian	Mexican
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

Step 5: Retrieve the Data

Pulling with Python



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")
```

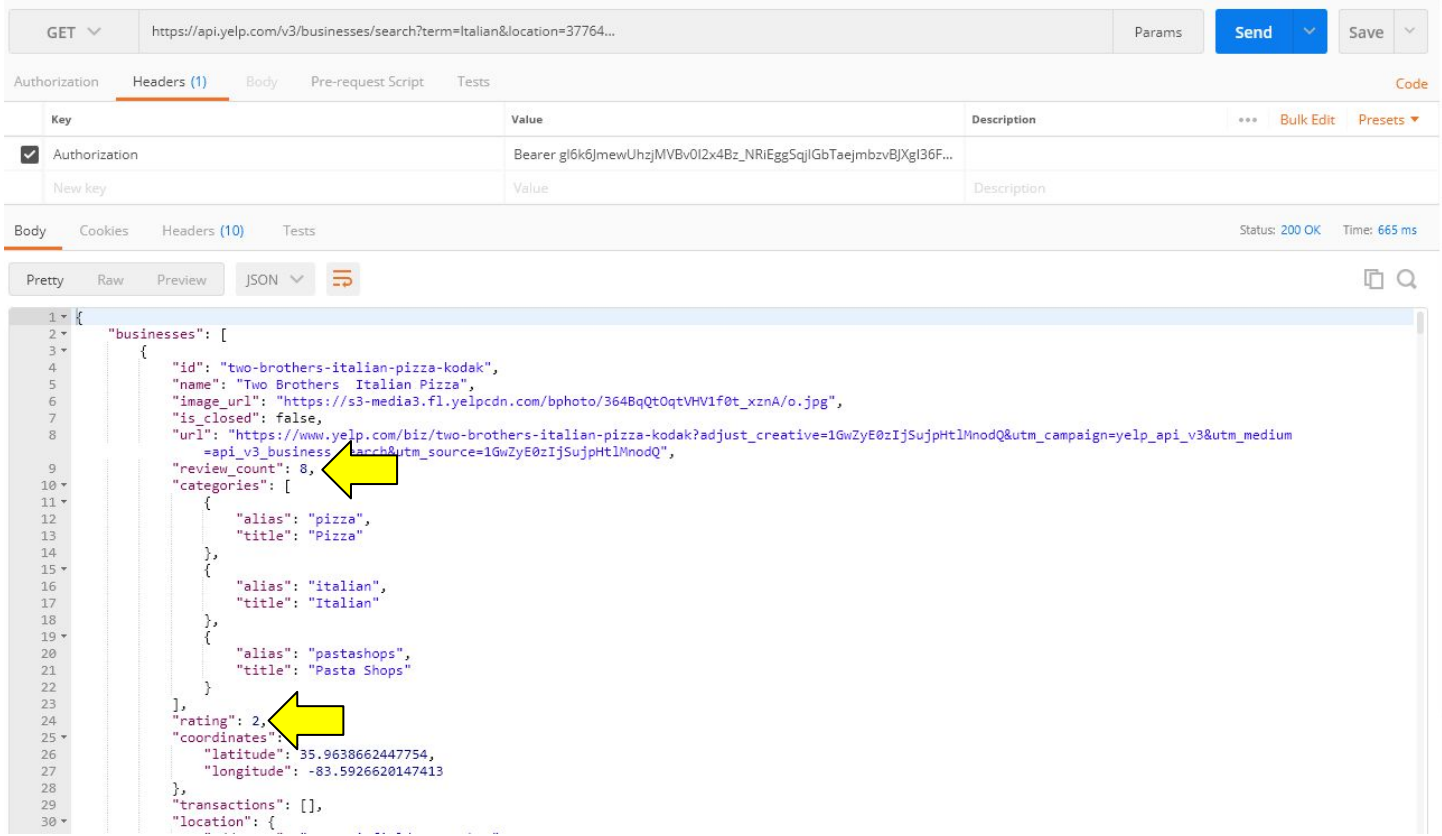
This funky code...

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 of these URLs.

Pulling with Python



Each of these URLs holds a piece of our answer.

Step 6: Assemble and Clean the Data

Cleaning with Pandas

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

```
# 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	1052.2	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

Step 7: Analyze for Trends

Analyze for Trends (Table)

It’s close:

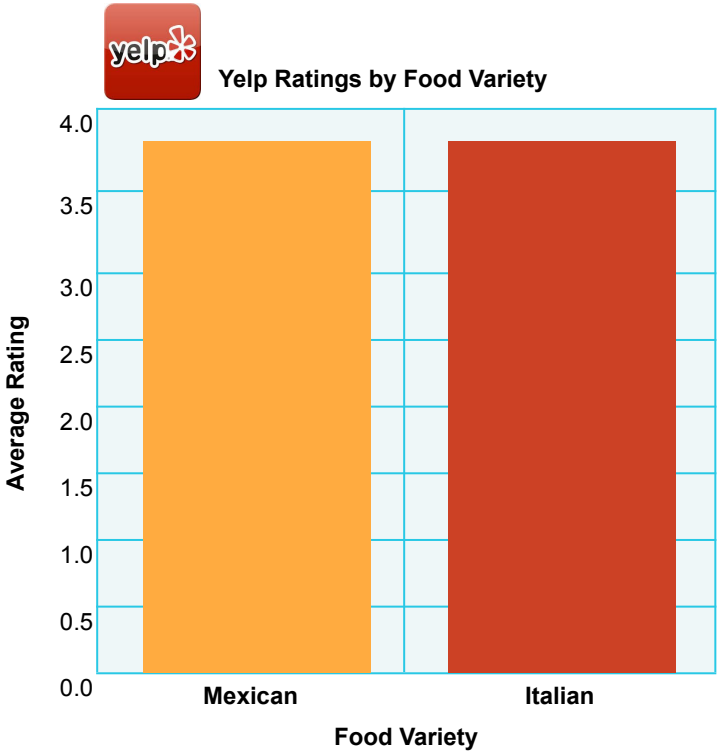
Display Summary of Results

```
# 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

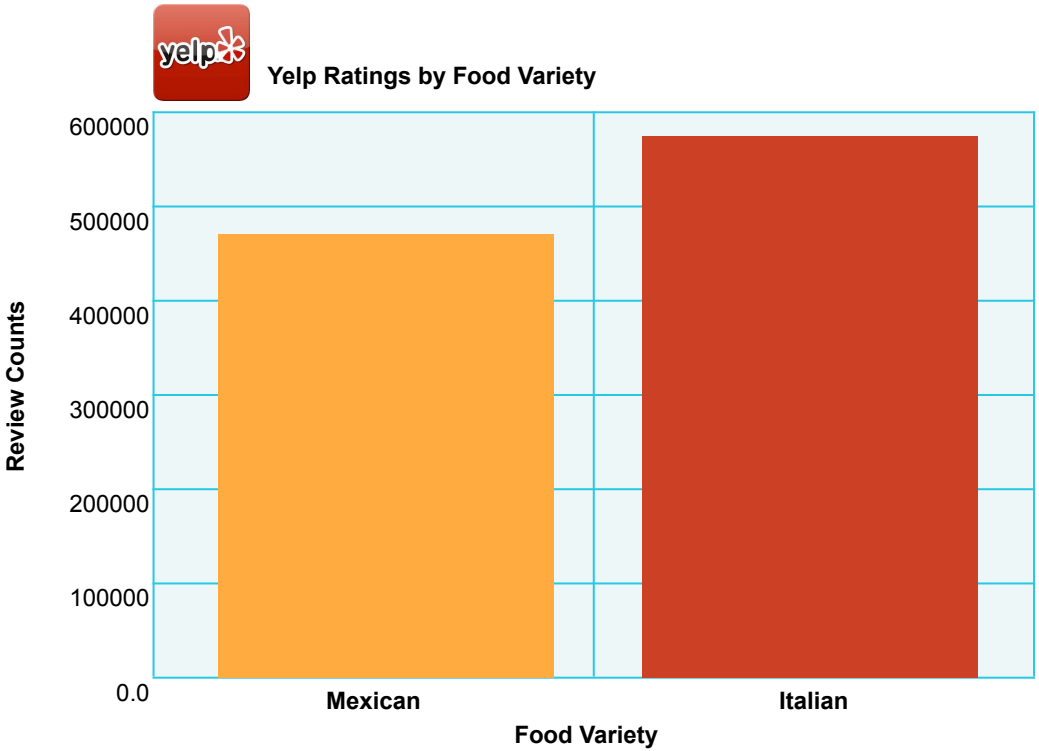
Analyze for Trends (Ratings)

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



Analyze for Trends (Ratings)

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



Analyze for Trends (Statistical Analysis)

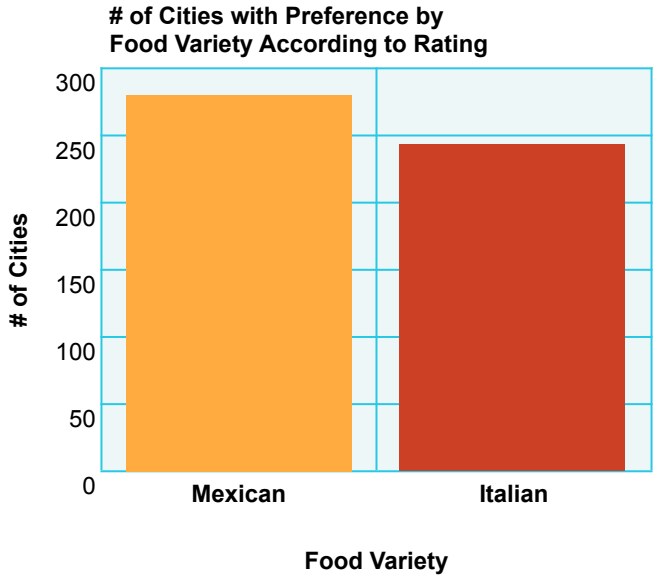
We have an intuitive sense that the numbers are close, but to quantify our intuition, we use a Student's t-test. After performing the t-test, we can quantifiably state that the differences are not statistically significant

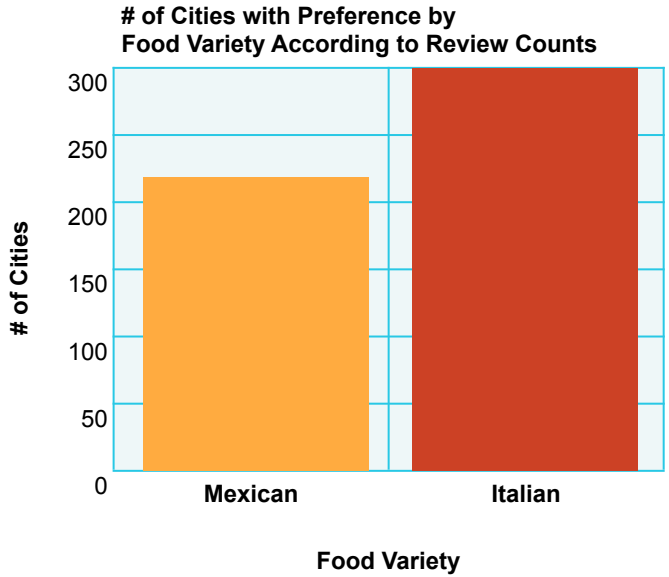
Metric	Italian	Mexican	p-value (t-test)
Average Rating	3.806	3.826	0.284
Review Counts	573k	476k	0.057

The difference in review counts is **not statistically significant**.

Analyze for Trends (Winner Take All)

Just for fun, let's throw in an analysis that aggregates the data from all cities using a winner-take-all approach.
It's sort of a wash.





Step 8: Acknowledge Limitations

Limitations of Analysis

Yelp demographics may not match the American demographic.

Limitations of Analysis

Restaurant experiences do not equate to home-cooked meals.

Limitations of Analysis

Fine-dining effect?

Step 9: Make the Call

Making the Call

The “Proper” Conclusion:
Based on our analysis, it’s clear that Americans’ preferences 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 is substantial evidence that Americans write more reviews of Italian restaurants than Mexican restaurants (+96k, p-value: 0.057).

This may indicate there is an increased interest in visiting Italian restaurants “for the experience.” Or it may merely suggest that Yelp users enjoy writing reviews of Italian restaurants more than Mexican restaurants.

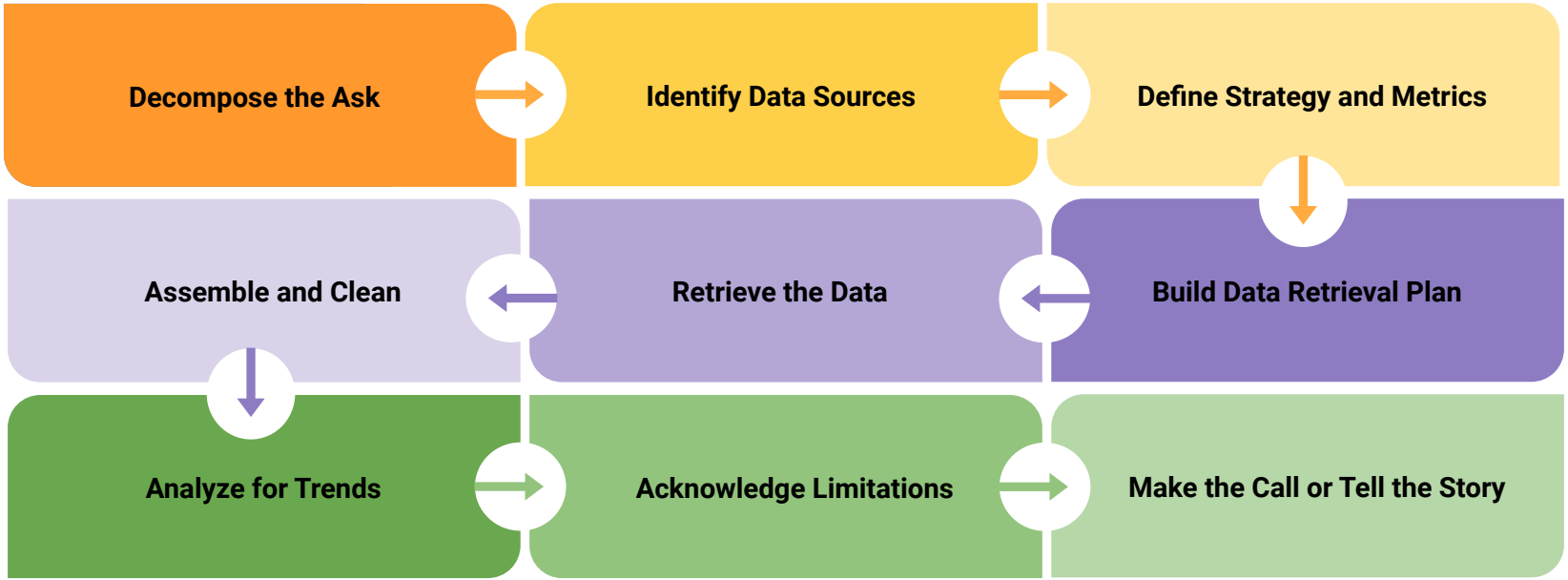
Making the Call

The “Let’s Be Real” Conclusion: Italian (but it’s going to be close).

An Analytics Paradigm

Analytics Paradigm

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



Group Activity: Predicting Gentrification

Using the analytics paradigm as a framework, outline a strategy to identify which neighborhoods in our city are seeing signs of gentrification.

Suggested Time:

15 minutes

Group Activity: Predicting Gentrification

Specifically, how would you answer these 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?

Time's Up! Let's Review.

Prepare for Next Class

By next class:

01 Make certain that you have Microsoft Excel installed.

02 Make certain that you have Slack installed and you are actively checking it.

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

04 Figure out where class videos will be posted.

Questions?