Zen of Data

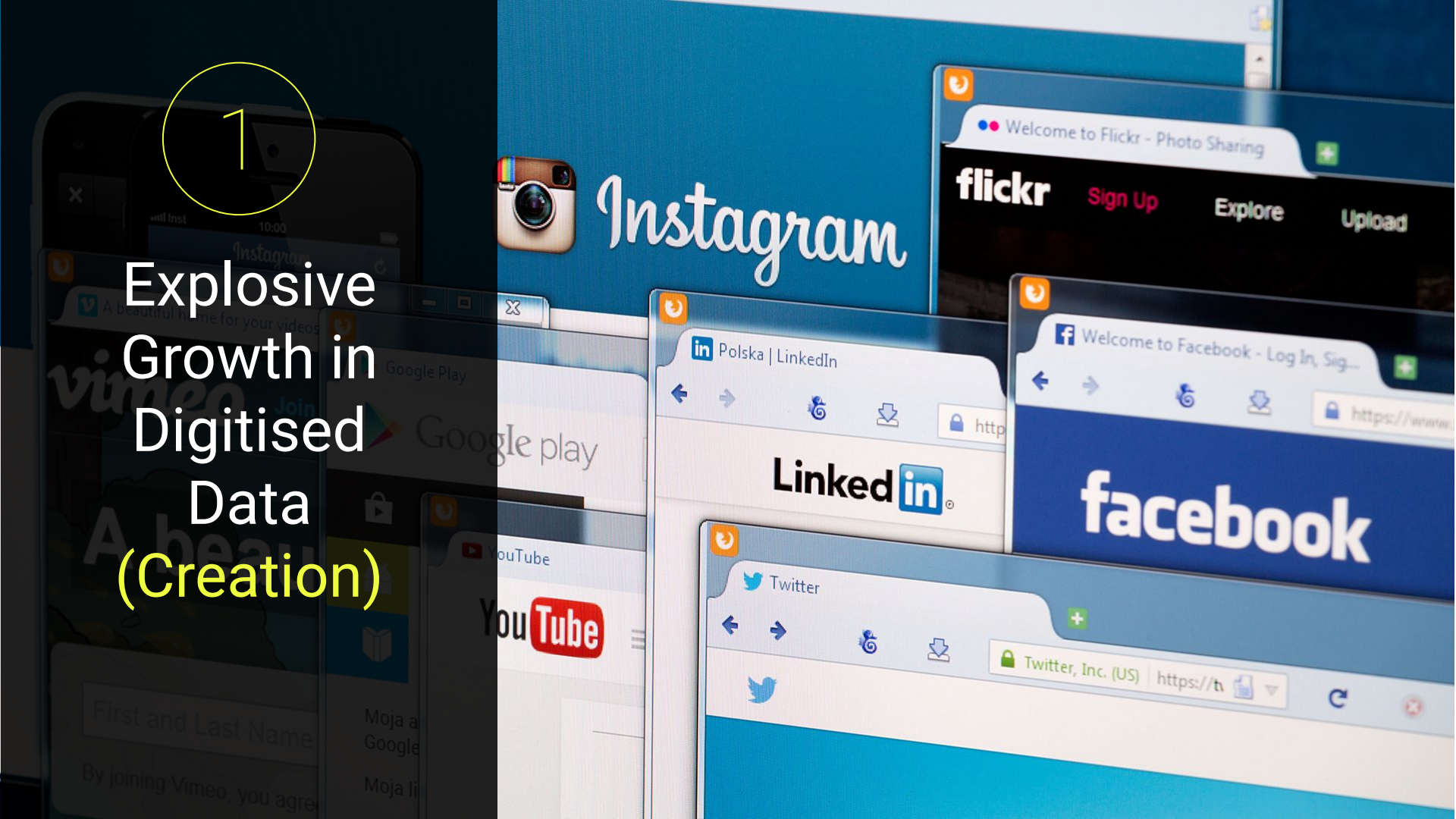
Data Boot Camp

Lesson 1.1


WELCOME

The Rise of Data

Why is **data analytics** such a **hot** skill these days?

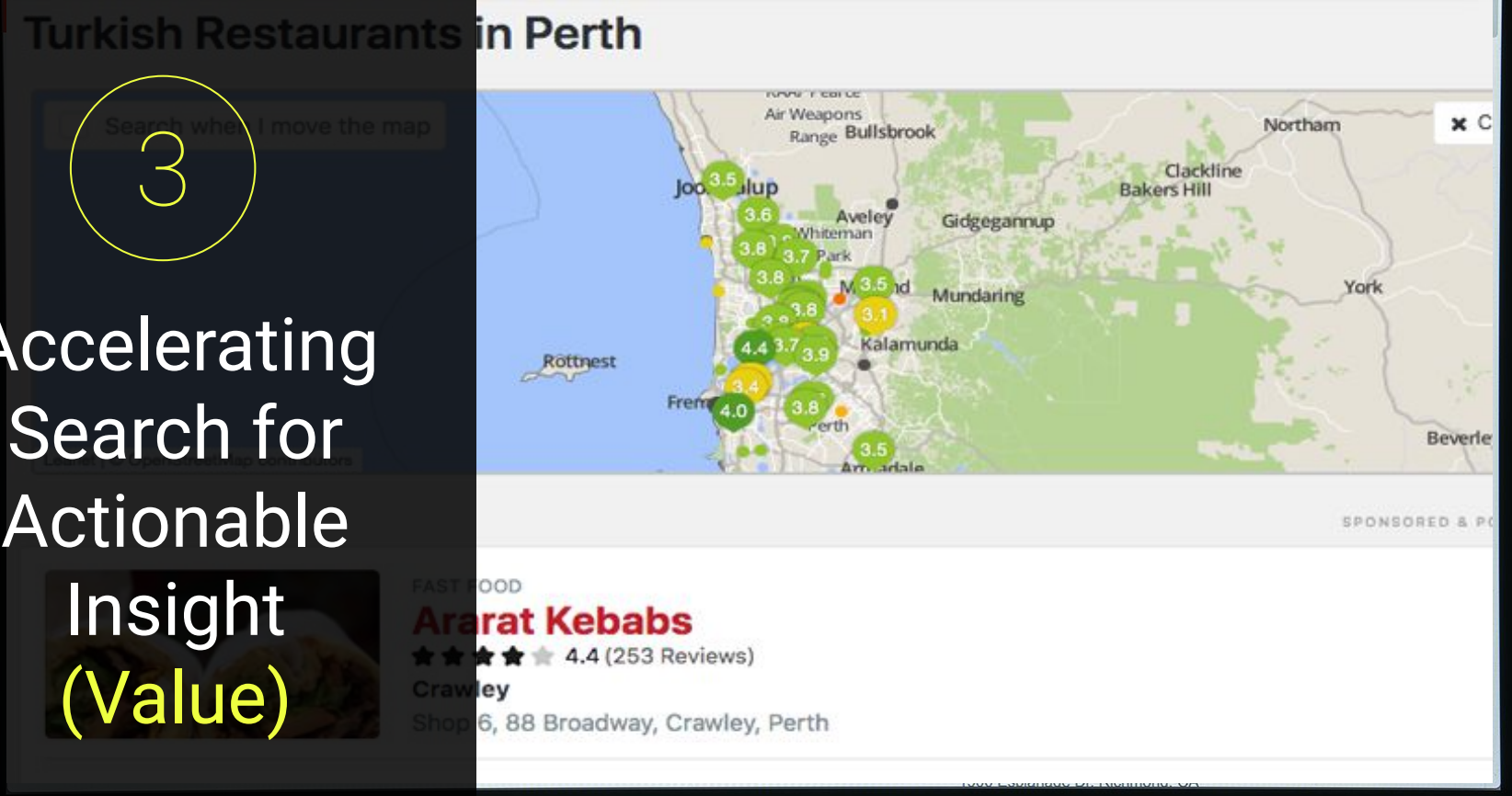

1
Explosive Growth in Digitised Data (Creation)
Instagram
flickr
Sign Up
Explore
Upload
Welcome to Flickr - Photo Sharing
Polska | LinkedIn
LinkedIn
Welcome to Facebook - Log In, Sig...
facebook
Twitter
Twitter, Inc. (US)
YouTube
Google play
vimeo

2

Explosive Growth in Analytic Tools (Synthesis)

3

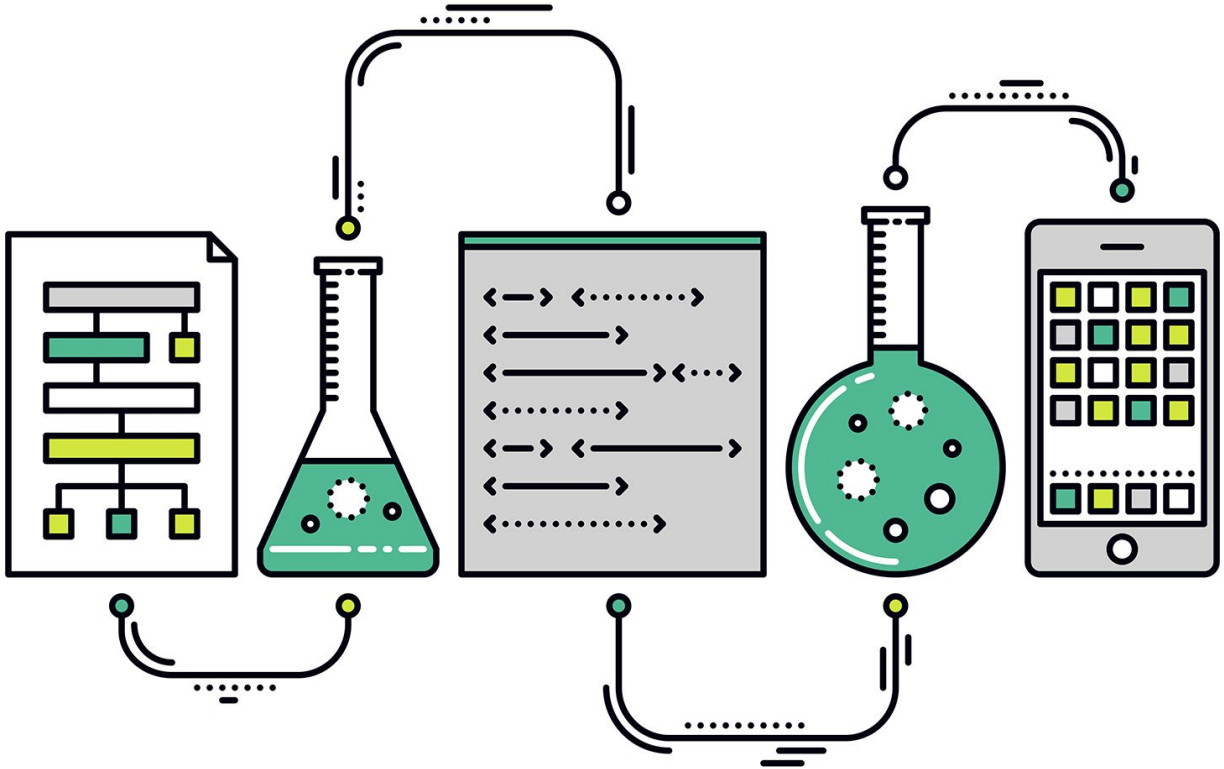
Accelerating Search for Actionable Insight (Value)



What does the term **data science** mean?

Perhaps you are picturing an Excel spreadsheet.

Data Science Involves Spreadsheets and Formulas



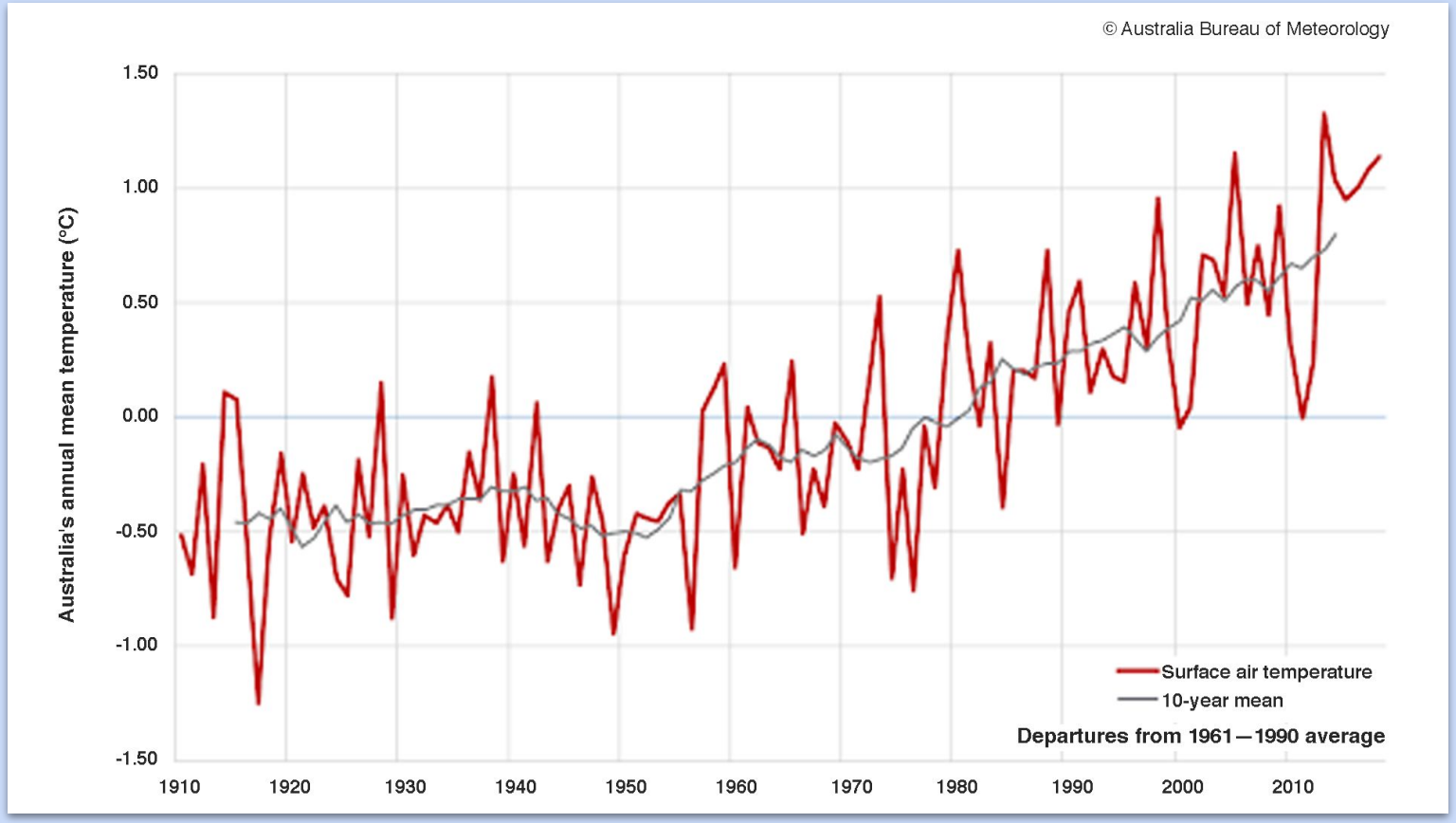
Fundamentally, data science is about **storytelling** and **truth telling.**

Data as Storytelling

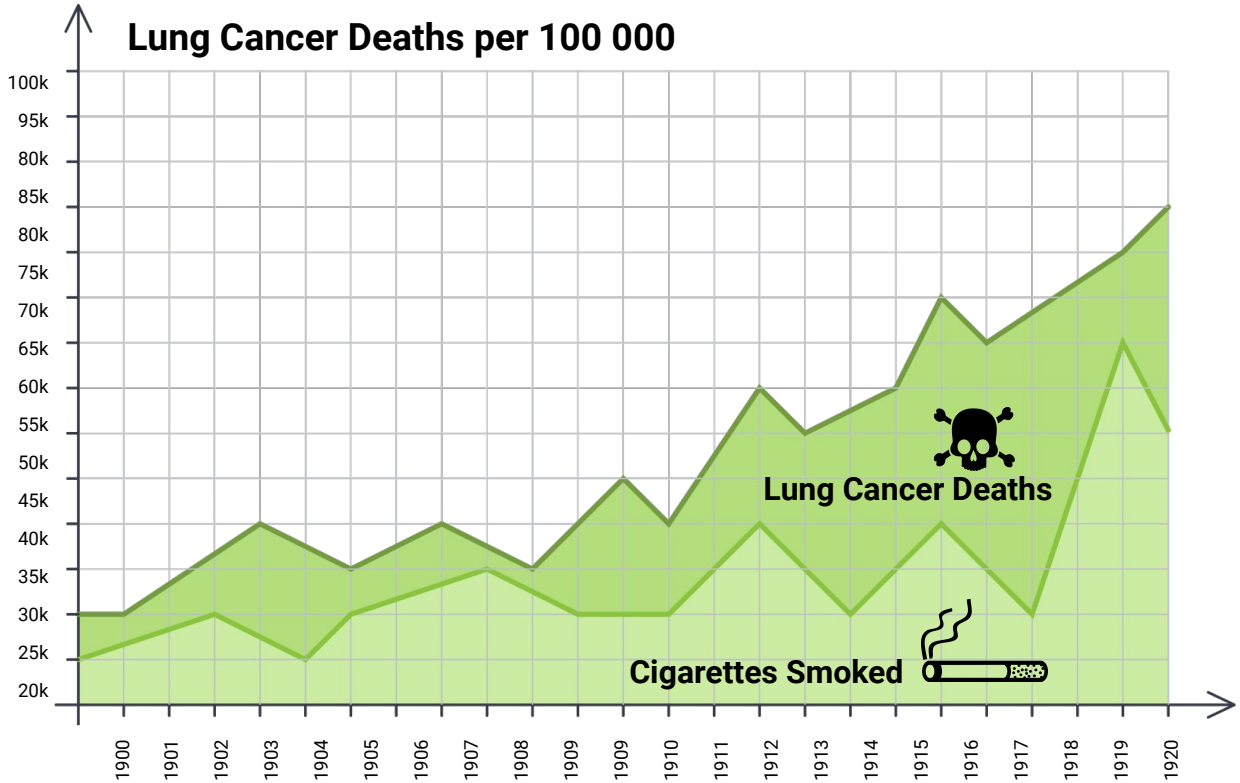
Data as Storytelling

Australia GDP

Data = Drama: Australia's annual mean temperature (°C)



Data as Truth Telling



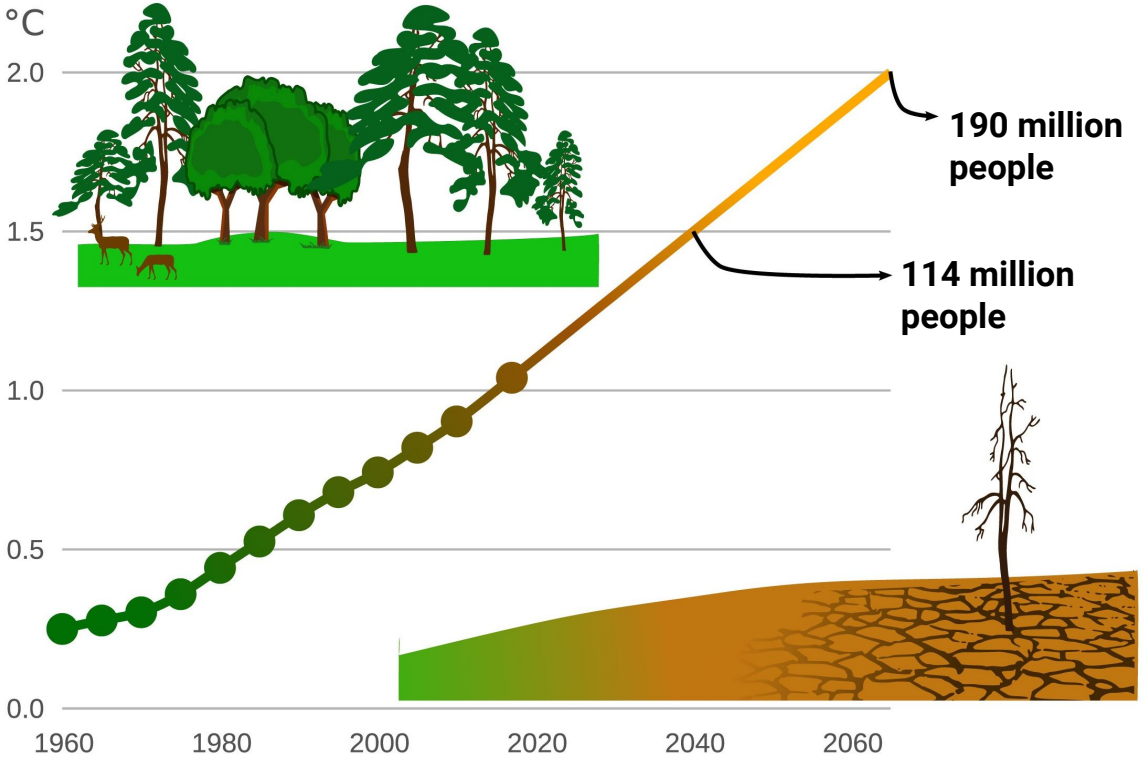
Data as Truth Telling

Unearthing Relationships

Data as Truth Telling

Making Predictions

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

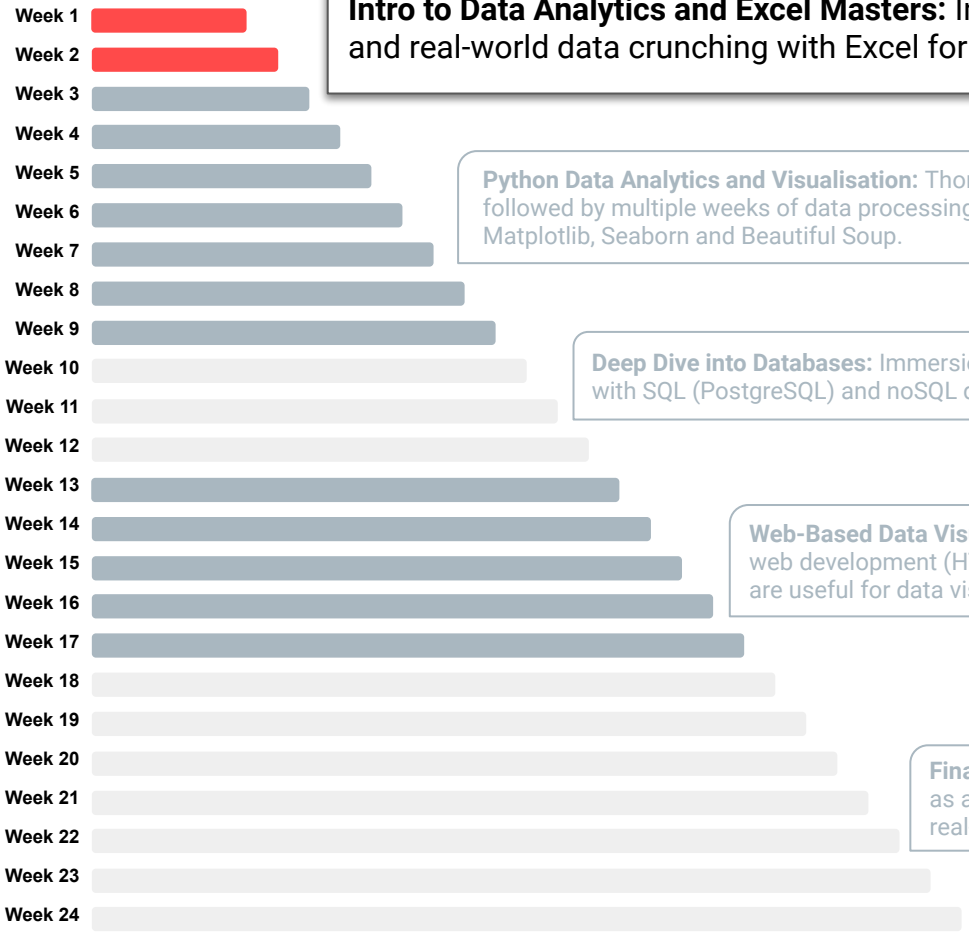
Course Overview

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

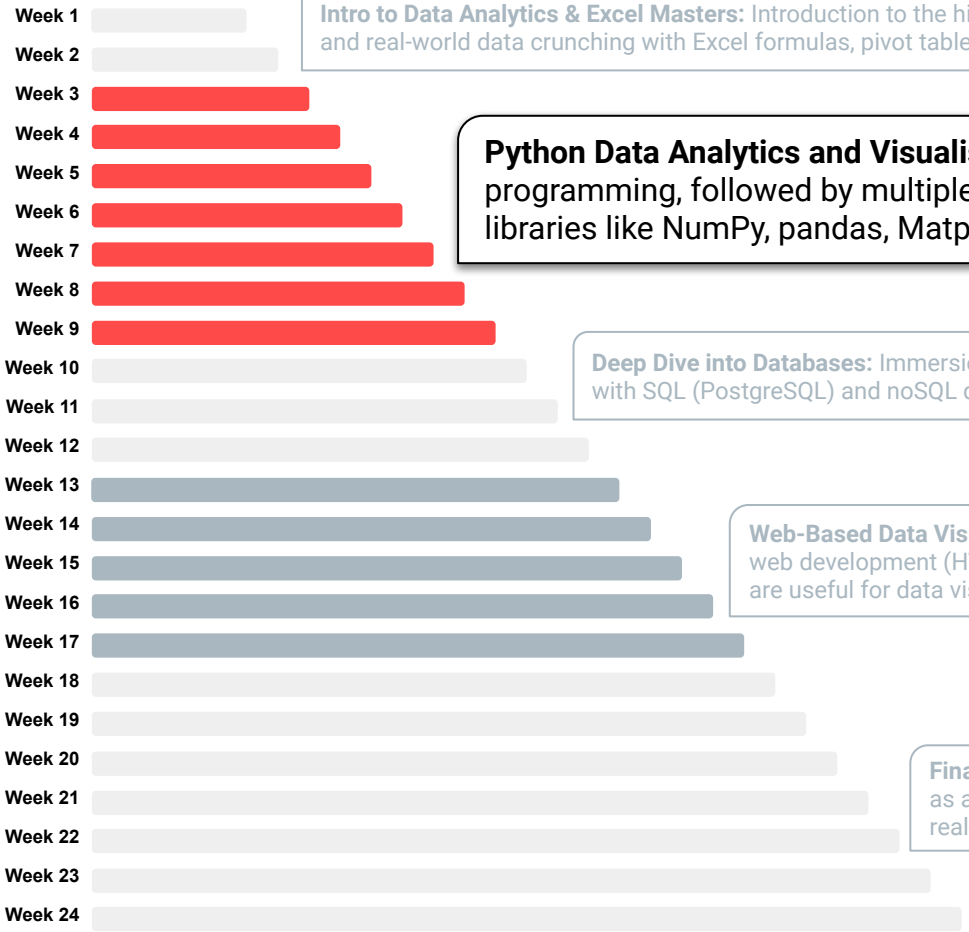
Python Data Analytics and Visualisation: Thorough crash course on Python programming, followed by multiple weeks of data processing using advanced libraries like NumPy, pandas, Matplotlib, Seaborn and Beautiful Soup.

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

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

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

Weeks 3–9



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.

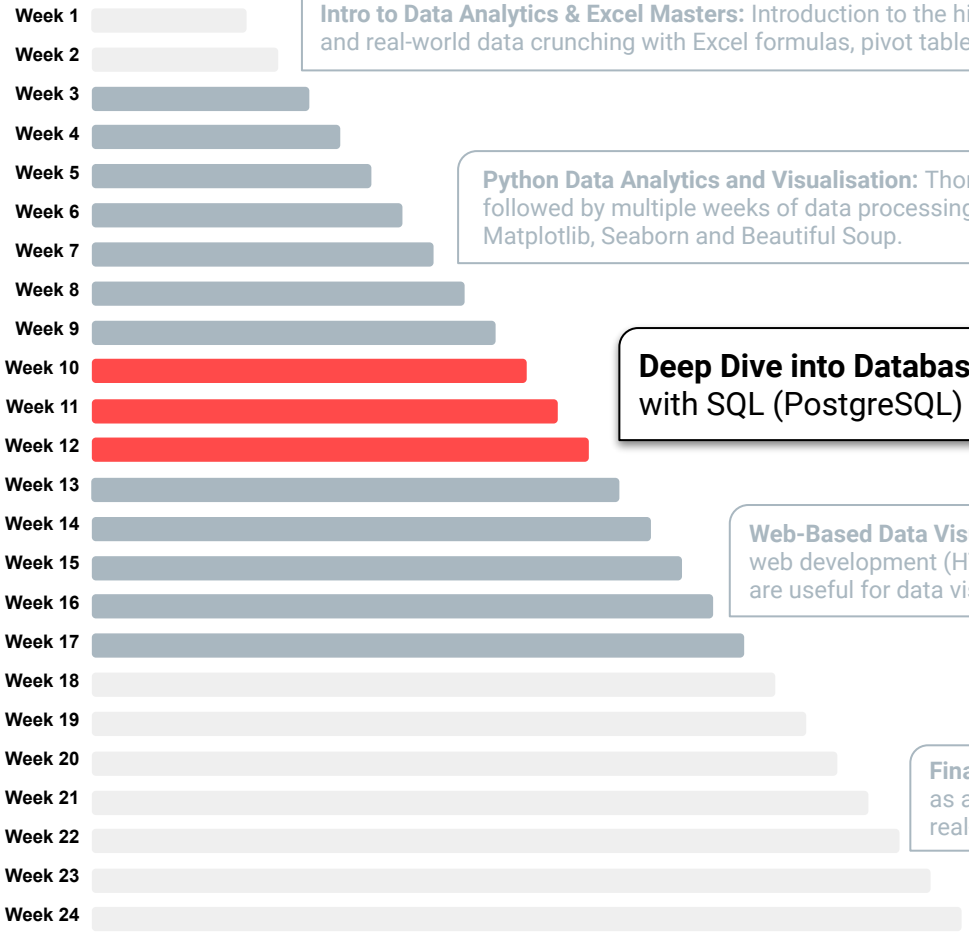
Python Data Analytics and Visualisation: Thorough crash course on Python programming, followed by multiple weeks of data processing using advanced libraries like NumPy, pandas, Matplotlib, Seaborn, and Beautiful Soup.

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

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

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

Weeks 10–12



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.

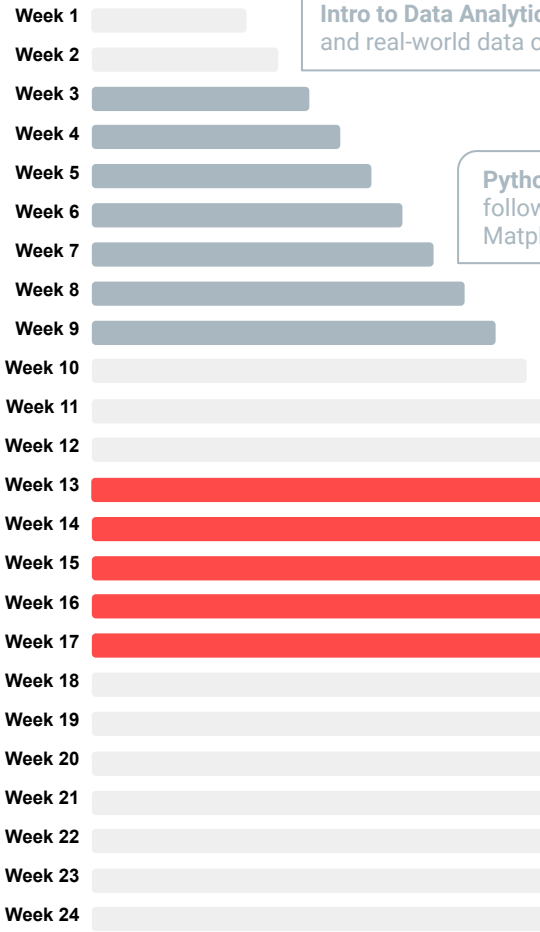
Python Data Analytics and Visualisation: Thorough crash course on Python programming, followed by multiple weeks of data processing using advanced libraries like NumPy, pandas, Matplotlib, Seaborn and Beautiful Soup.

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

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

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

Weeks 13–17



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.

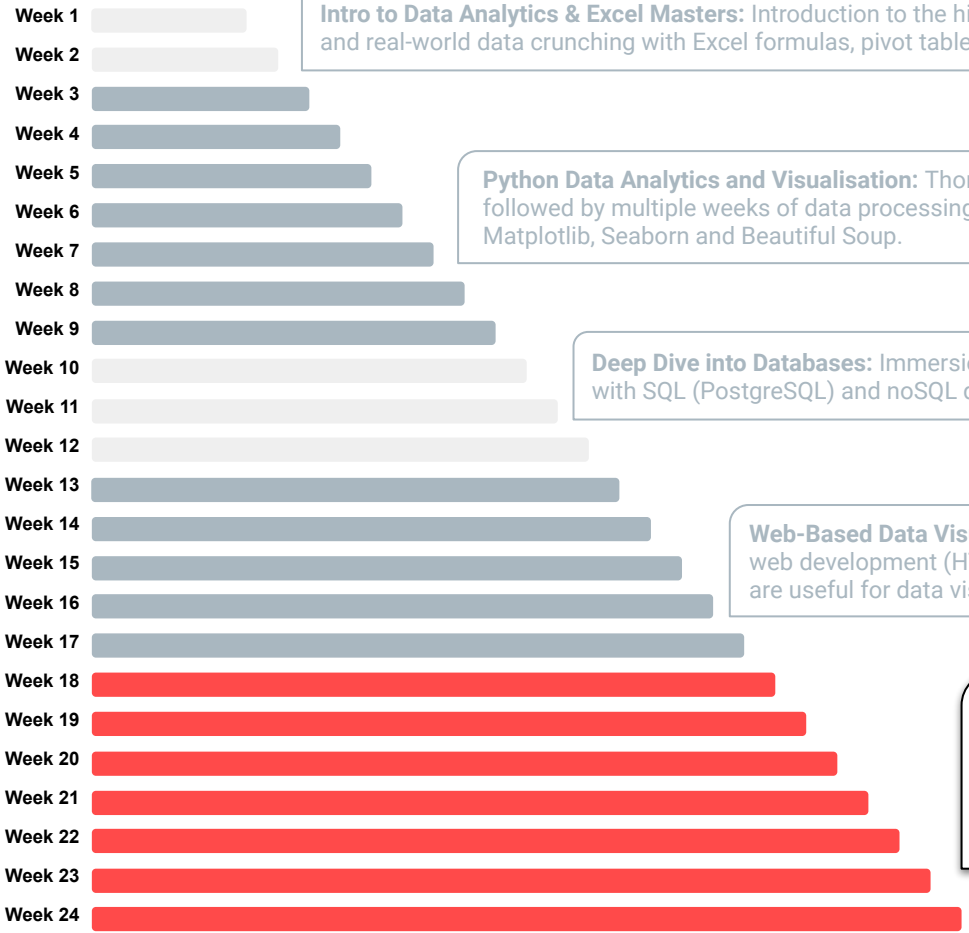
Python Data Analytics and Visualisation: Thorough crash course on Python programming, followed by multiple weeks of data processing using advanced libraries like NumPy, pandas, Matplotlib, Seaborn and Beautiful Soup.

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

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

Final Projects & Advanced Topics: Introduction to Tableau and R as well as advanced topics like Hadoop and Machine Learning. Develop a real-world data visualisation 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.

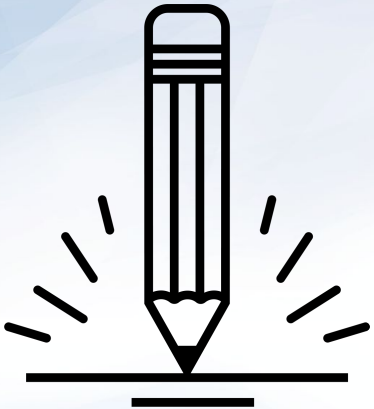
Python Data Analytics and Visualisation: Thorough crash course on Python programming, followed by multiple weeks of data processing using advanced libraries like NumPy, pandas, Matplotlib, Seaborn and Beautiful Soup.

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

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

Final Projects and Advanced Topics: Introduction to Tableau and R, as well as advanced topics like Hadoop and Machine Learning; develop a real-world data visualisation project.

Example Activity



Example Activity:
Banking Deserts

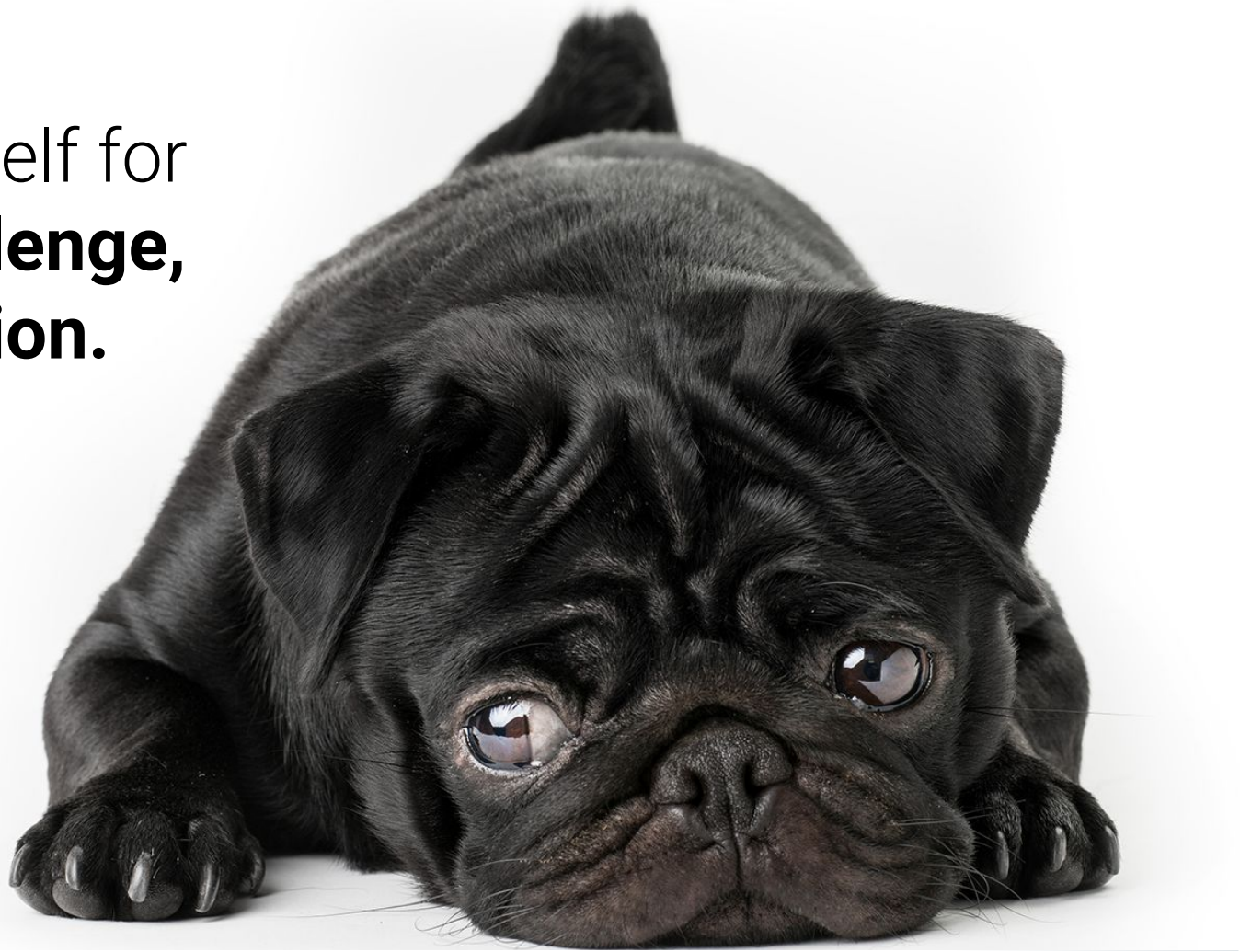
In this activity, you will use a variety of public demographic data and APIs to explain many real-world social phenomena. Utilise data from sources like the Australian Census, Google Maps, and more to find insights on poverty, discrimination, and the impact of changing economies.

Suggested Time:
20 minutes

Helpful Tips


Embrace your
inner toddler.

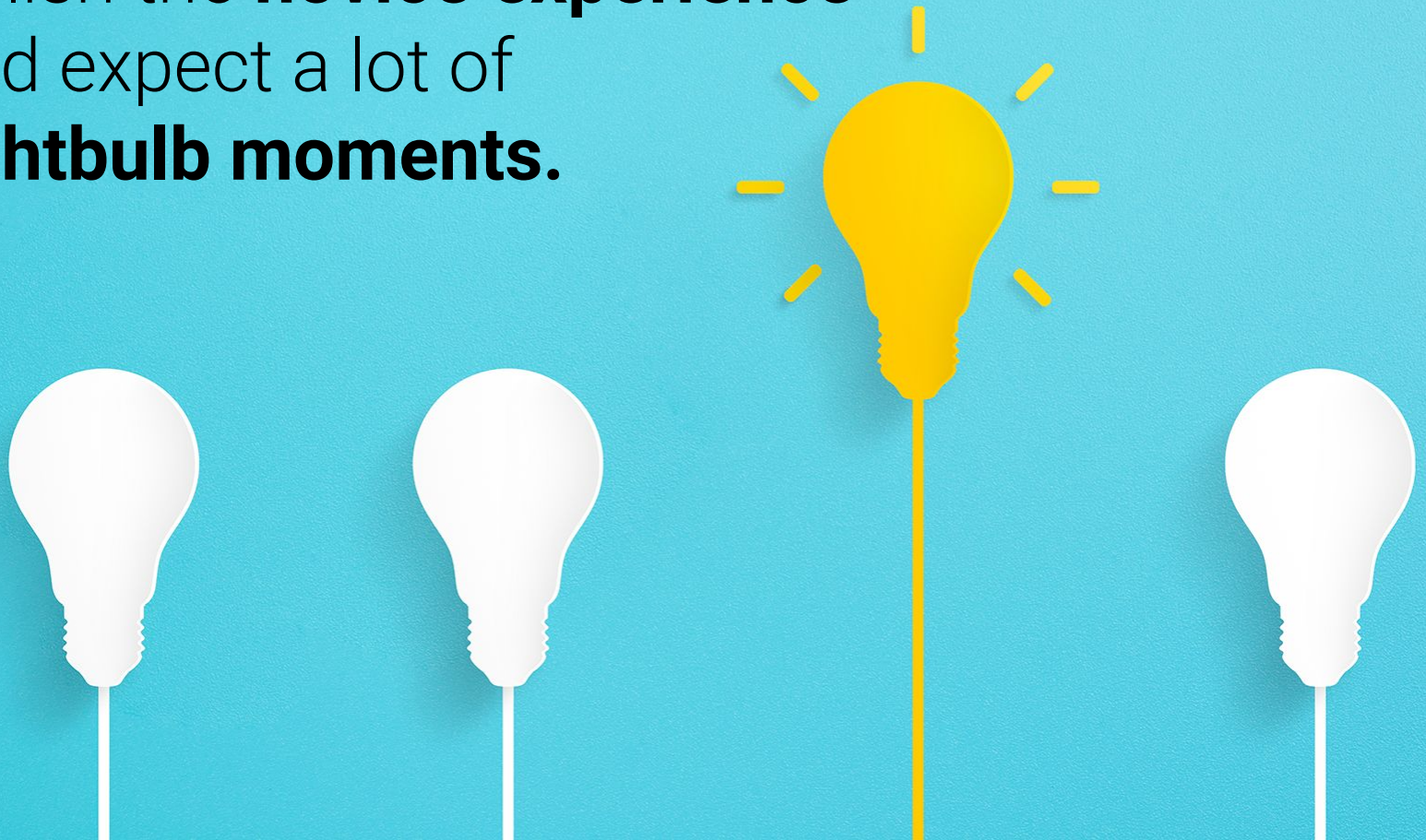
Brace yourself for **doubt, challenge,** and **confusion.**



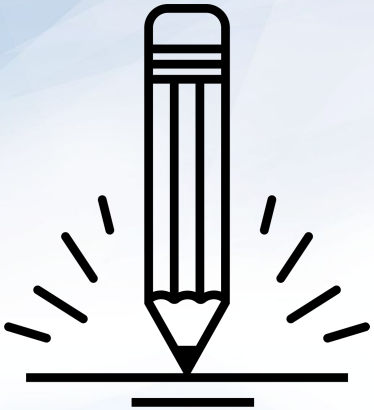
There is no shortcut.
You've got to **put in the hours!**


Form a community
with your classmates.

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



Celebrate your successes!



Group Activity:

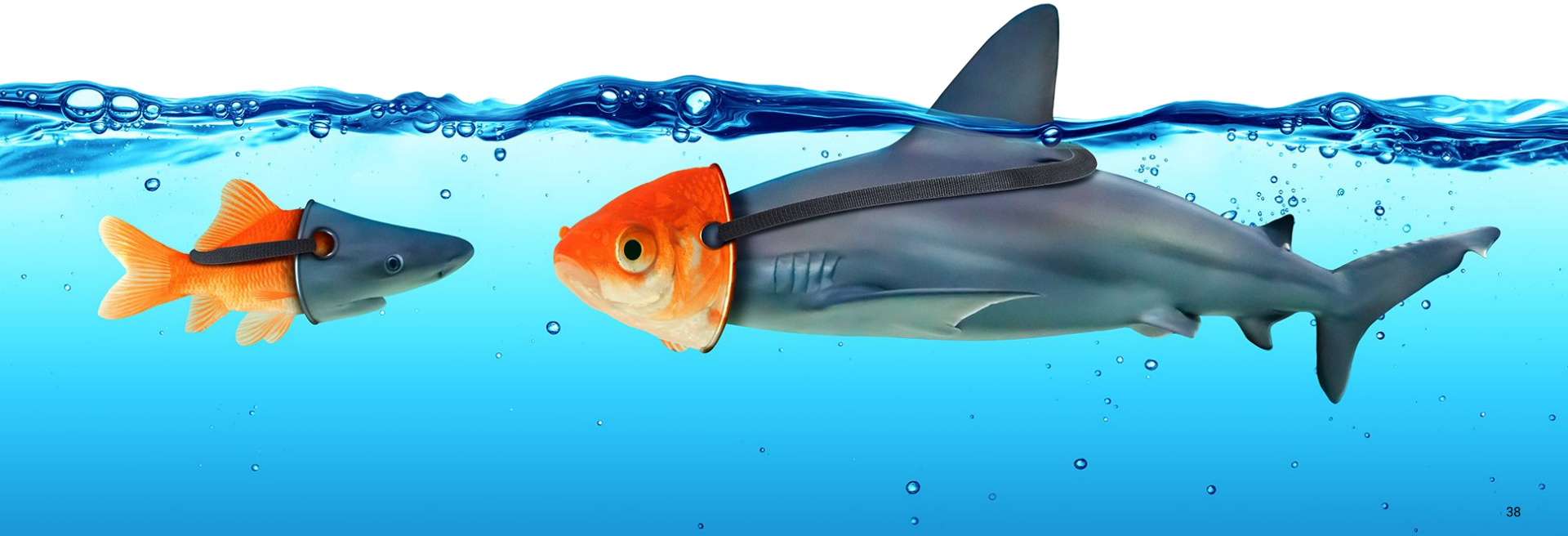
Form groups of 3 or 4 people. Introduce yourself to your group.

Don’t be shy!

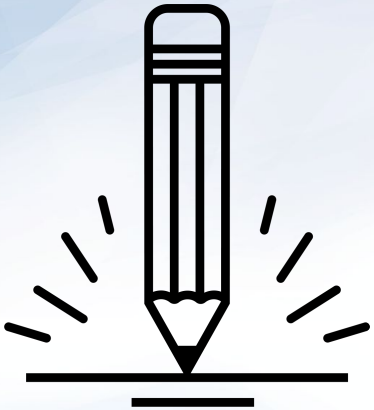
Suggested Time:
10 minutes

Two Truths and a Lie!

Please slack out your answer in the class slack channel




Break



Group Activity: The Great Debate

Rejoin the same group from the previous activity. Together, ponder the following question ...

Suggested Time:
20 minutes

Group Activity: The Great Debate

Which do Australians prefer: Italian or Thai 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 nine Australians who will each be asked the question—with 0 qualifiers.

You only have your team.

Suggested Time: 20 minutes

The Great Debate (Analysed)

Step 1: Decompose the ‘Ask’

Step 1: Decompose the ‘Ask’

Which do **Australians** prefer:
Italian or Thai food?

Step 1: Decompose the ‘Ask’

Which do **Australians** prefer: Italian or Thai food?

- Who exactly is an **Australian**?
- Are **Australians** just homeowners?
- Do **Australians** just live in big cities?
- Are **Australians** just millennials?

How can we get a representative sample of Australians?

Step 1: Decompose the ‘Ask’

Which do Australians **prefer**:
Italian or Thai food?

Step 1: Decompose the ‘Ask’

Which do Australians **prefer**: Italian or Thai 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 regionalised? 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 Australians prefer:

Italian or Thai food?

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

Step 1: Decompose the ‘Ask’

Which do Australians prefer: **Italian or Thai food**?

01

How do we categorise foods? Is pizza Italian? Is Bing Boy Thai?

02

How do we categorise 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 renditions of these foods?

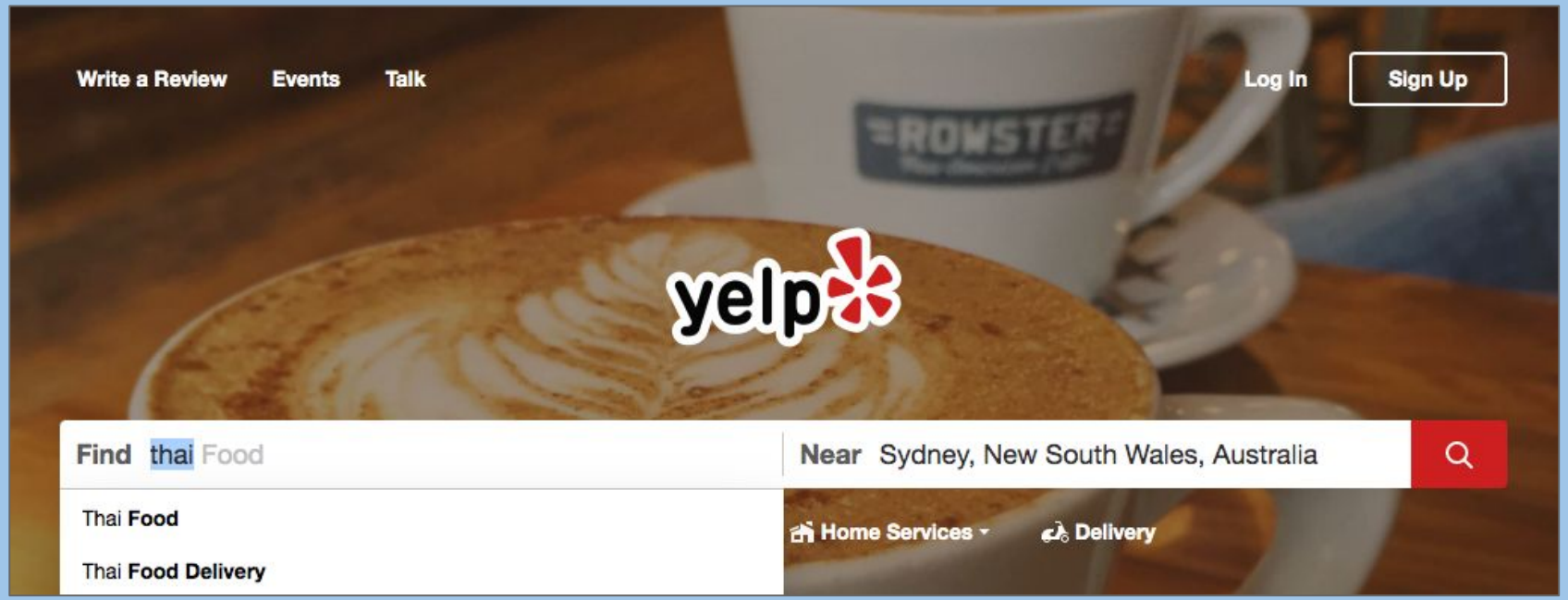
Step 2: Identify Data Sources

Step 2: Identify Data Sources

Why poll an audience when there already exist enormous databases of information about Australians' food preferences readily available online?

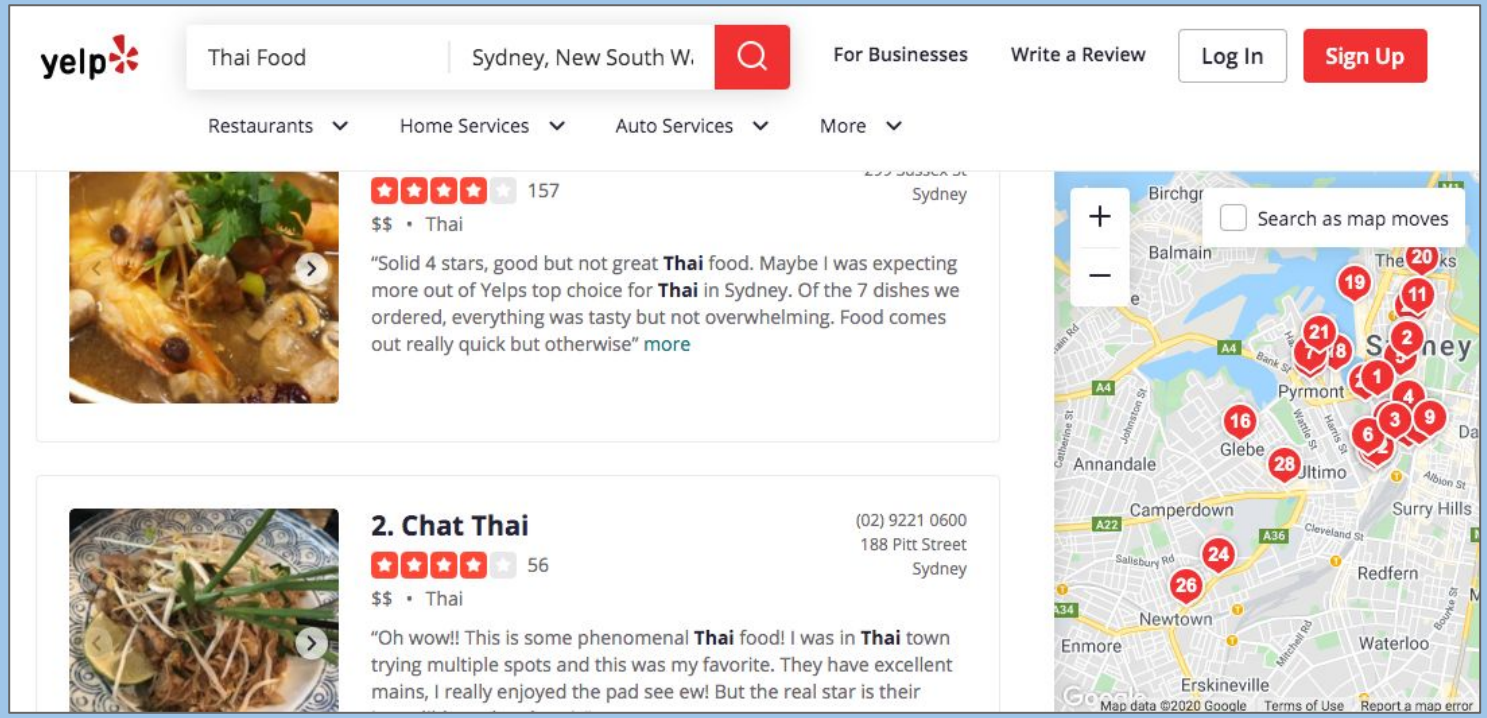
Step 2: Identify Data Sources

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



Step 2: Identify Data Sources

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



Step 3: Define Strategy and Metrics

Step 3: Define Strategy and Metrics

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

Australians:

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

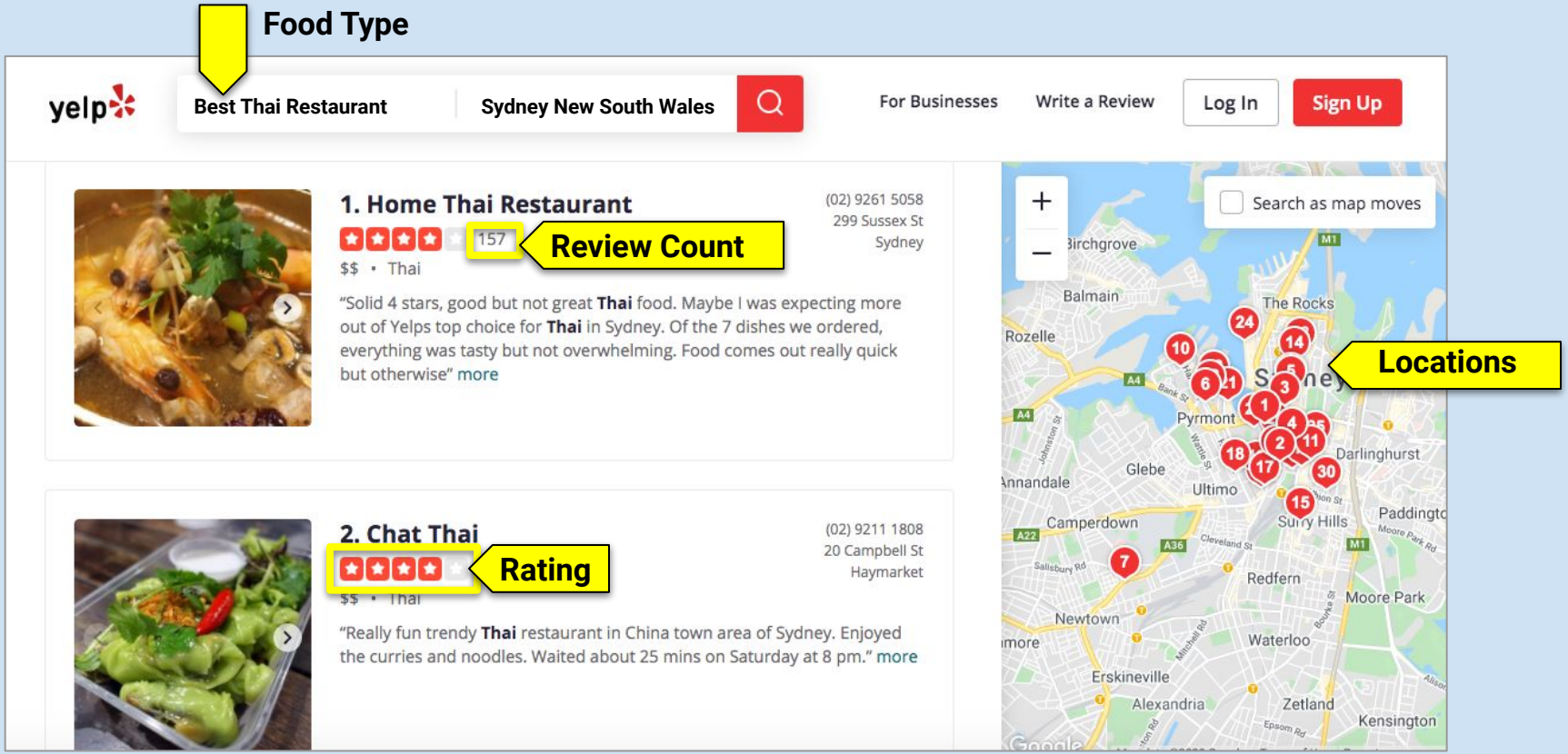
Preference:

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

Italian and Thai Food:

- Top 20 Italian and Thai restaurants in every city

Step 3: Define Strategy and Metrics



Step 3: Define Strategy and Metrics

Repeat this analysis for as many cities as possible.

Melbourne, VIC	
Italian	**Thai**
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

VS

Sydney, NSW	
Italian	**Thai**
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

VS

Brisbane, QLD	
Italian	**Thai**
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

VS

Adelaide, SA	
Italian	**Thai**
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

VS

Darwin, NT	
Italian	**Thai**
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

VS

Perth, WA	
Italian	**Thai**
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

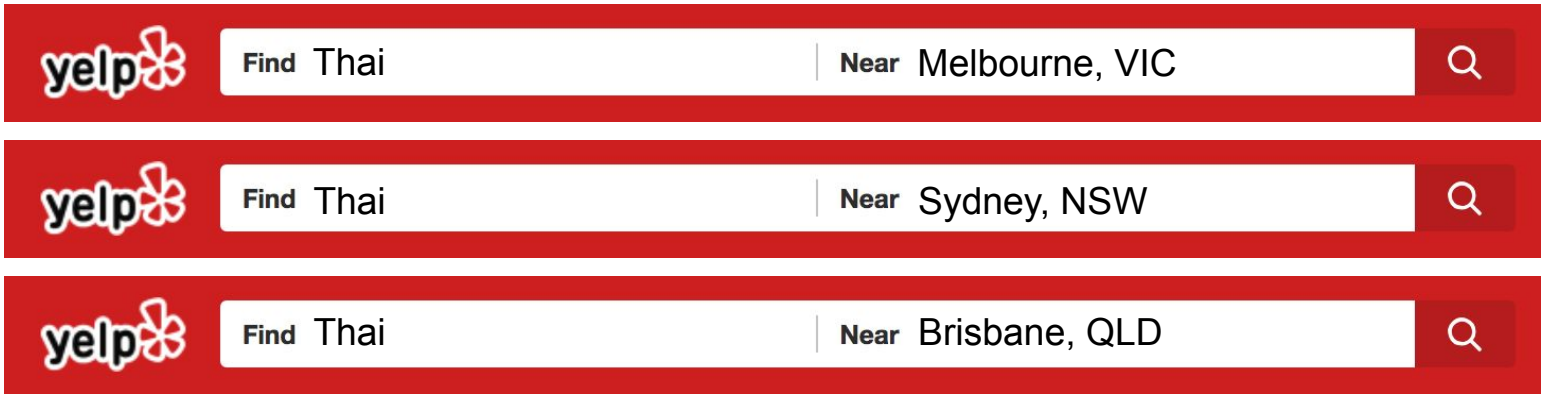
VS

Step 4: Build Data Retrieval Plan

Step 4: Build Data Retrieval Plan

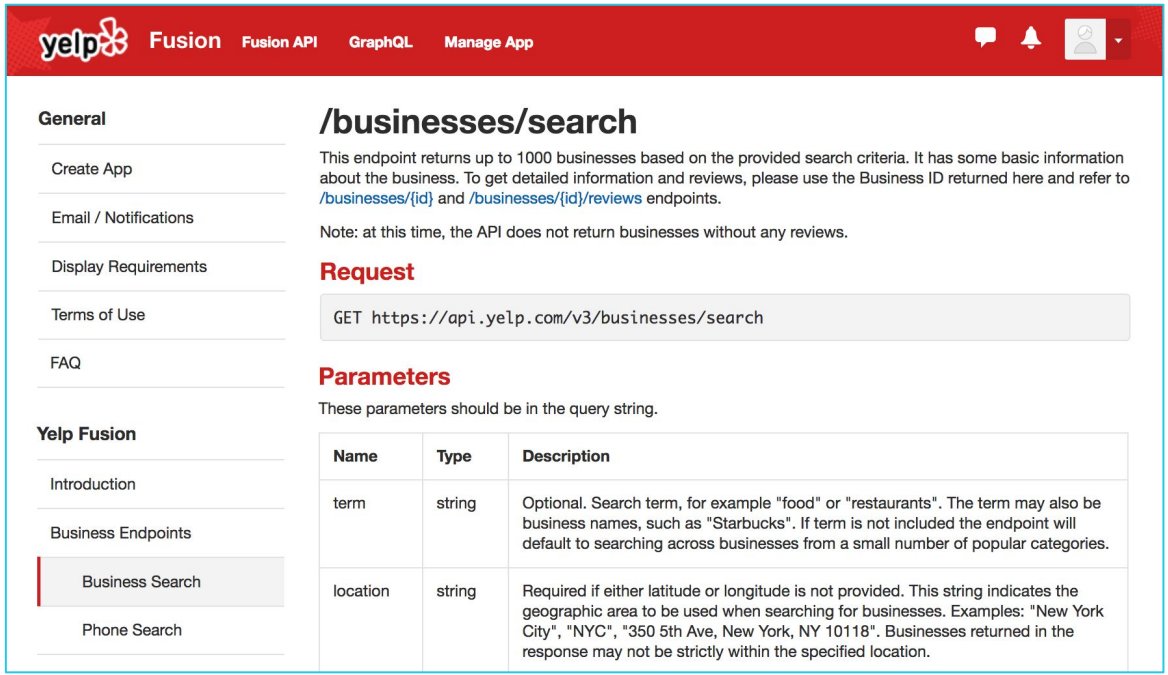
We could retrieve this data by brute force, but it would be:

- 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!


```
{'id': 'WNpF7jhHH9U12t4dhuDlEA',
 'alias': 'pure-thai-berala-berala',
 'name': 'Pure Thai Berala',
 'image_url': 'https://s3-media1.fl.yelpcdn.com/bphoto/mlkp03PLFa4gVIlWj82cXg/o.jpg',
 'is_closed': False,
 'url': 'https://www.yelp.com.au/biz/pure-thai-berala-berala?adjust_creative=1GwZyE0zIjSujp
m=api_v3_business_search&utm_source=1GwZyE0zIjSujpHtlMnodQ',
 'review_count': 3,
 'categories': [{'alias': 'thai', 'title': 'Thai'}],
 'rating': 3.5,
 'coordinates': {'latitude': -33.871492, 'longitude': 151.031695},
 'transactions': [],
 'location': {'address1': '160 Woodburn Rd',
  'address2': '',
  'address3': '',
  'city': 'Berala',
  'zip_code': '2141',
  'country': 'AU',
  'state': 'NSW',
  'display_address': ['160 Woodburn Rd', 'Berala New South Wales 2141']},
 'phone': '+61296492882',
 'display_phone': '(02) 9649 2882',
 'distance': 9435.618278117496}
```


Step 4: Build Data Retrieval Plan

We will build a Python script to randomly select over 700 postcodes from the Australian Census, and then acquire review data from the top 20 Thai and Italian restaurants for each postcode using the Yelp API.

3001	
Italian	**Thai**
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

2009	
Italian	**Thai**
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

4000	
Italian	**Thai**
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

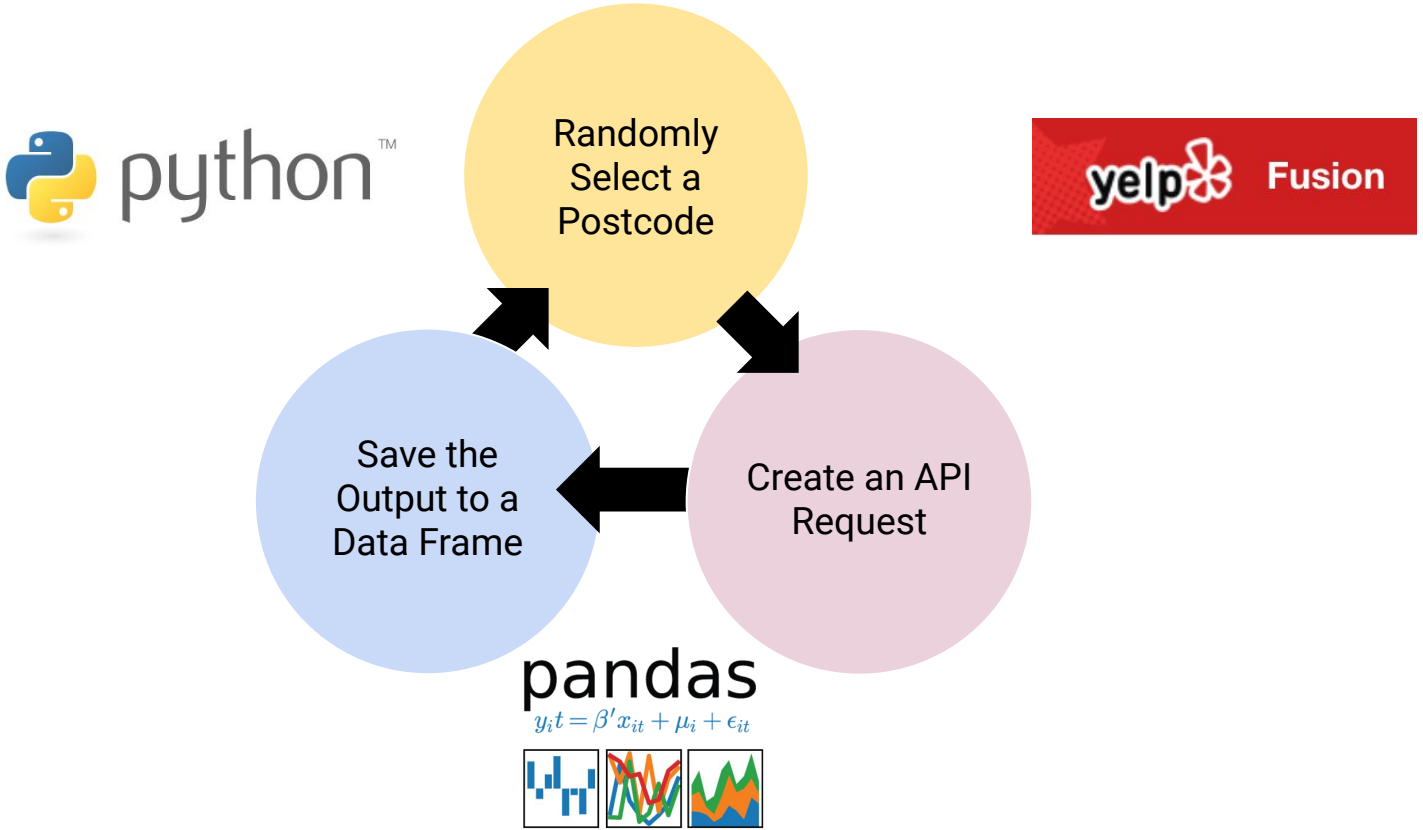
5012	
Italian	**Thai**
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

0810	
Italian	**Thai**
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

6001	
Italian	**Thai**
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant
Restaurant	Restaurant

Step 5: Retrieve the Data

Pulling with Python



Pulling with Python

```
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_thai["businesses"]:
        thai_review_count = thai_review_count + business["review_count"]
        thai_weighted_review = thai_weighted_review + (business["review_count"] * business["rating"])

    # Append the data to the appropriate column of the data frames
    italian_data = italian_data.append({
        'Postal Code': row["Postal Code"],
        'Italian Review Count':italian_review_count,
        'Italian Average Rating':(italian_weighted_review / italian_review_count),
        'Italian Weighted Rating':italian_weighted_review},ignore_index=True)

    thai_data = thai_data.append({
        'Postal Code': row["Postal Code"],
        'Thai Review Count':thai_review_count,
        'Thai Average Rating':(thai_weighted_review / thai_review_count),
        'Thai Weighted Rating':thai_weighted_review},ignore_index=True)

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

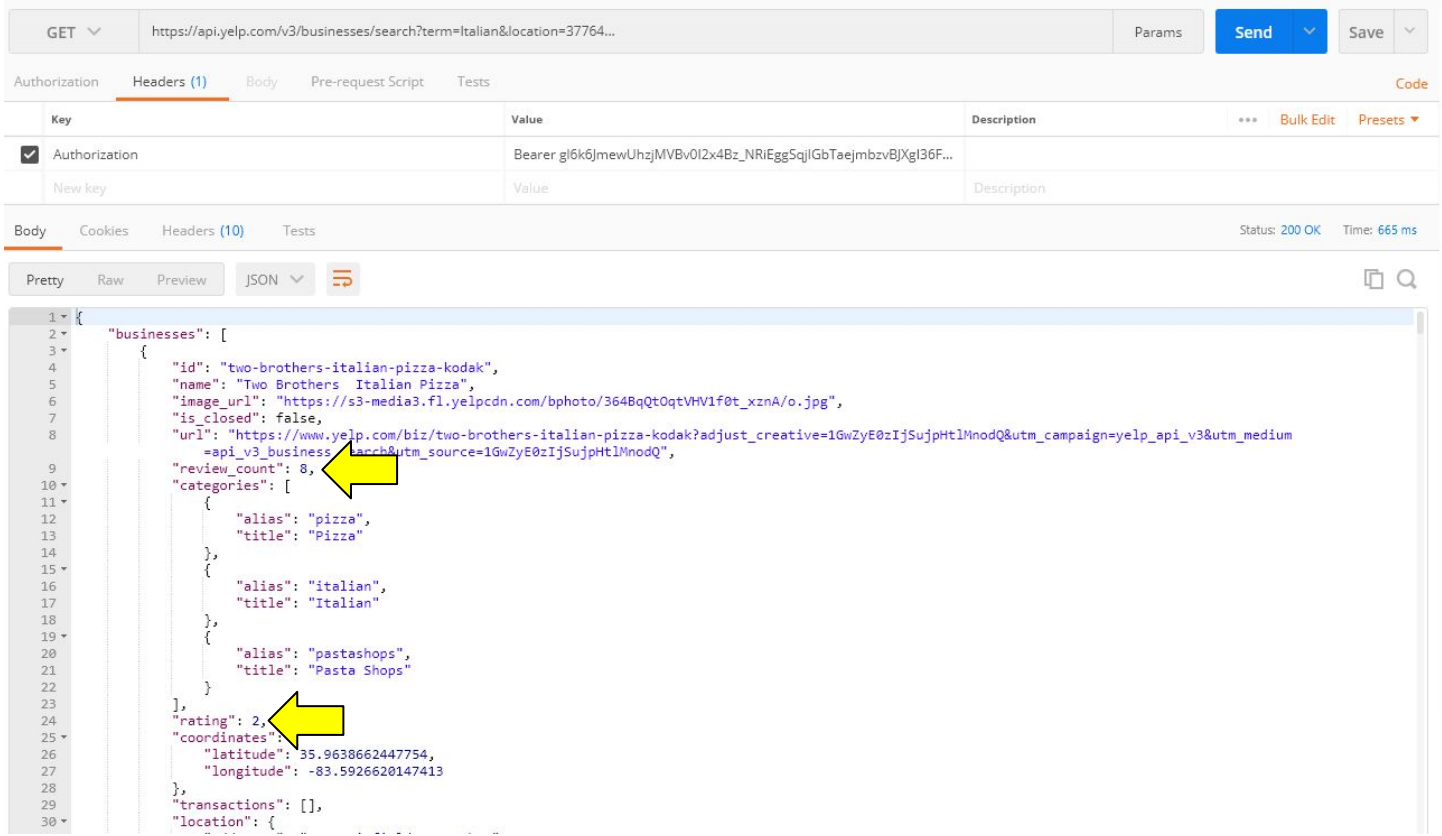
This funky code...

Pulling with Python

```
1
https://api.yelp.com/v3/businesses/serch?term-Italian&location=2055
https://api.yelp.com/v3/businesses/serch?term-Thai&location=2055
2
https://api.yelp.com/v3/businesses/serch?term-Italian&location=2059
https://api.yelp.com/v3/businesses/serch?term-Thai&location=2059
3
https://api.yelp.com/v3/businesses/serch?term-Italian&location=2060
https://api.yelp.com/v3/businesses/serch?term-Thai&location=2060
4
https://api.yelp.com/v3/businesses/serch?term-Italian&location=2000
https://api.yelp.com/v3/businesses/serch?term-Thai&location=2000
5
https://api.yelp.com/v3/businesses/serch?term-Italian&location=2001
https://api.yelp.com/v3/businesses/serch?term-Thai&location=2001
6
https://api.yelp.com/v3/businesses/serch?term-Italian&location=2020
https://api.yelp.com/v3/businesses/serch?term-Thai&location=2020
7
https://api.yelp.com/v3/businesses/serch?term-Italian&location=2129
https://api.yelp.com/v3/businesses/serch?term-Thai&location=2129
```

...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 intrinsically 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(thai_data, italian_data, on="Postal Code")
combined_data.head( )
```

	Postcode	Thai Review Count	Thai Average Rating	Thai Weighted Rating	Italian Review Count	Italian Average Rating	Italian Weighted Rating
0	0801	97	4.1134	399	63	3.78571	238.5
1	4000	256	4.11133	1052.2	266	3.81955	1016
2	5012	378	3.64286	1377	66	3.2197	212.5
3	3001	222	4.16892	925.5	420	3.77857	1587
4	2009	2842	3.94053	11199	2829	3.92824	11113

Step 7: Analyse for Trends

Analyse for Trends (Table)

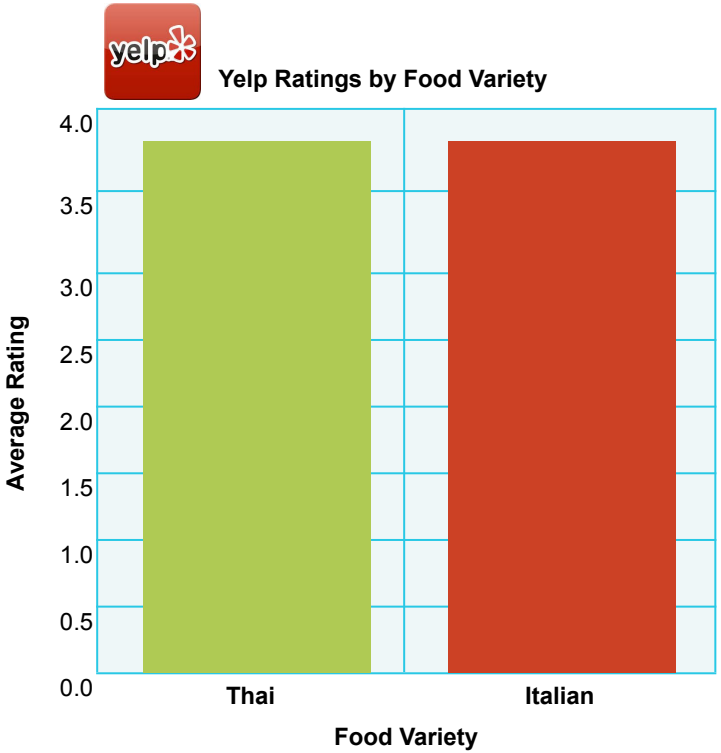
It’s Close:

```
# 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=["Italian"])

thai_summary = pd.DataFrame({"Review Counts": thai_data["Thai Review Count"].sum(),
                                "Rating Average": thai_data["Thai Average Rating"].mean(),
                                "Review Count Wins": combined_data["Review Count Wins"].value_counts()["Thai"],
                                "Rating Wins": combined_data["Rating Wins"].value_counts()["Thai"]}, index=["Thai"])

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

	Review Counts	Rating Average	Review Count Wins	Rating Wins
Thai	2229761.0	3.987034	347	461
Italian	2670762.0	3.964877	600	486

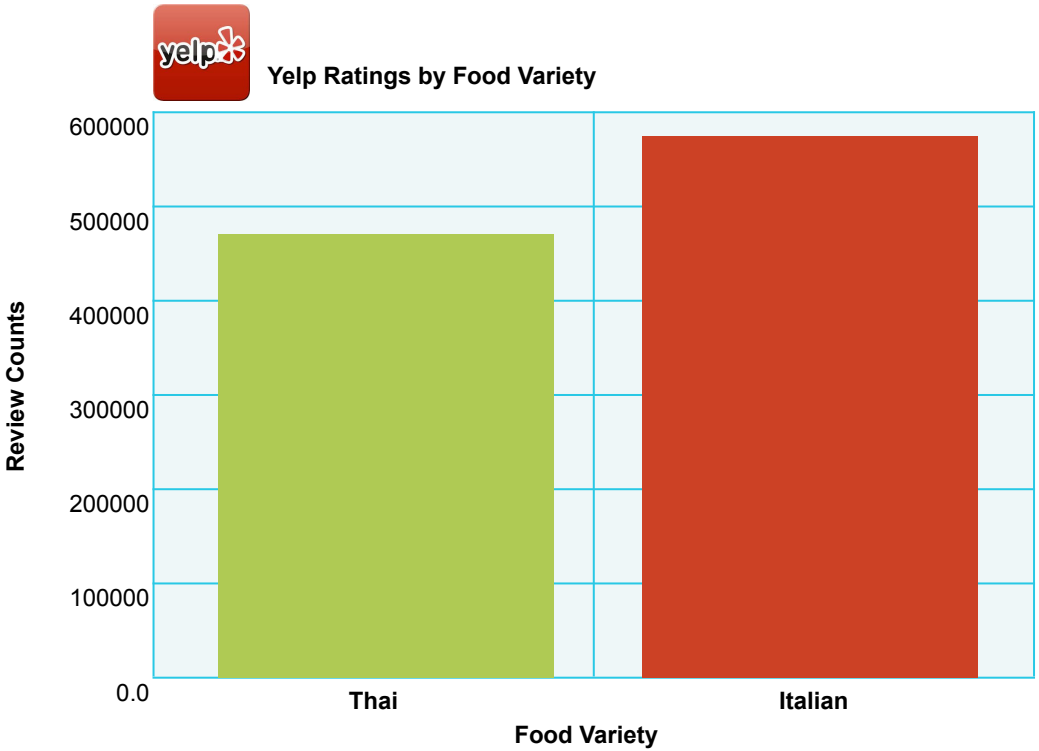
Analyse for Trends (Ratings)

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



Analyse for Trends (Ratings)

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



Analyse for Trends (Statistical Analysis)

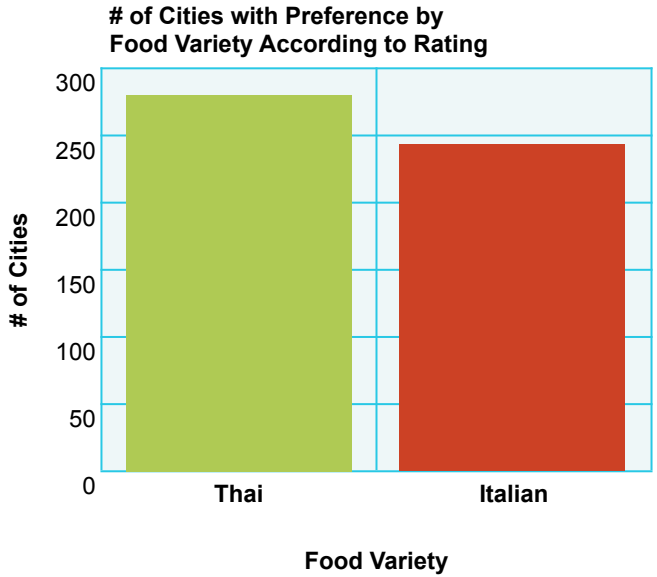
Because of how close the numbers appear, we utilised a Student's t-test to quickly assess if the perceived differences are not statistically significant but could be considered substantial.

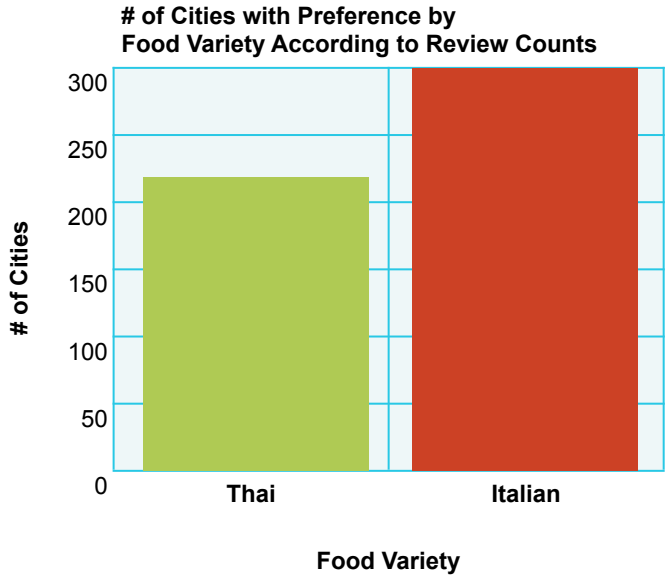
Metric	Italian	Thai	p-Value (t-test)
Average Rating	3.98	3.96	0.47
Review Counts	2.67M	2.29M	0.00023

The difference in review count is **statistically significant**.
But the difference in average rating is **not statistically significant**.

Analyse for Trends (Winner Take All)

Just for kicks, 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.

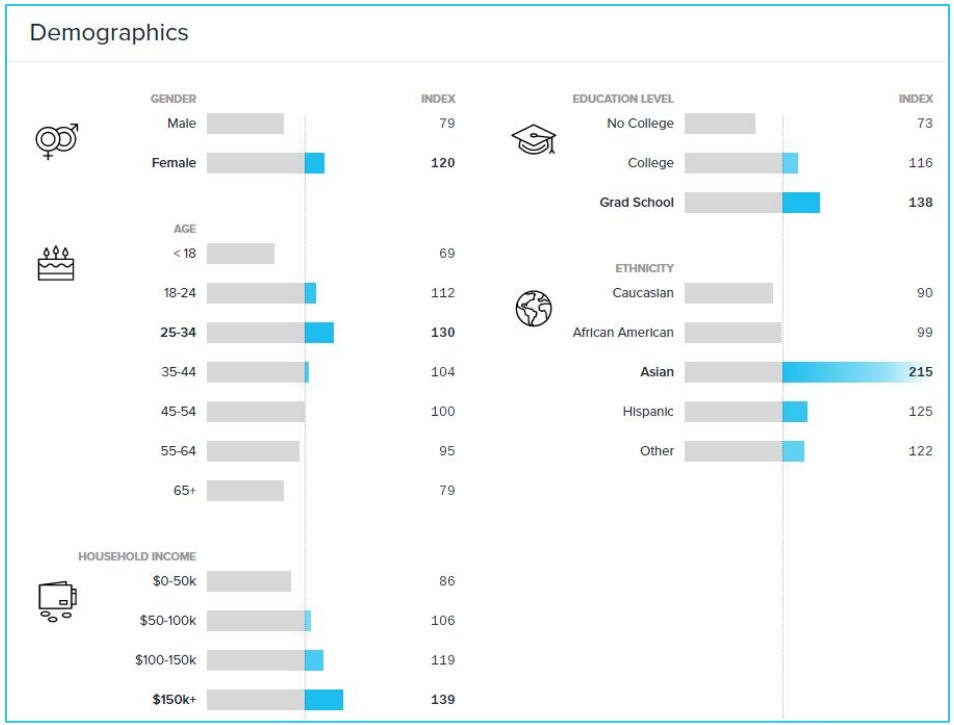




Step 8: Acknowledge Limitations

Limitations of Analysis

Yelp demographics may not match the Australian 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 Australians’ preferences for Italian and Thai food are similar in nature. As a whole, Australians rate Thai and Italian restaurants at statistically similar scores (avg. score: ~3.9, p-value: 0.47). However, there exists substantial evidence that Australians write more reviews of Italian restaurants than Thai restaurants (+400k, p-value: 2.34e-05).

This may indicate there is an increased interest in visiting Italian restaurants at an experiential level, or it may merely suggest that Yelp users enjoy writing reviews of Italian restaurants more than Thai restaurants.

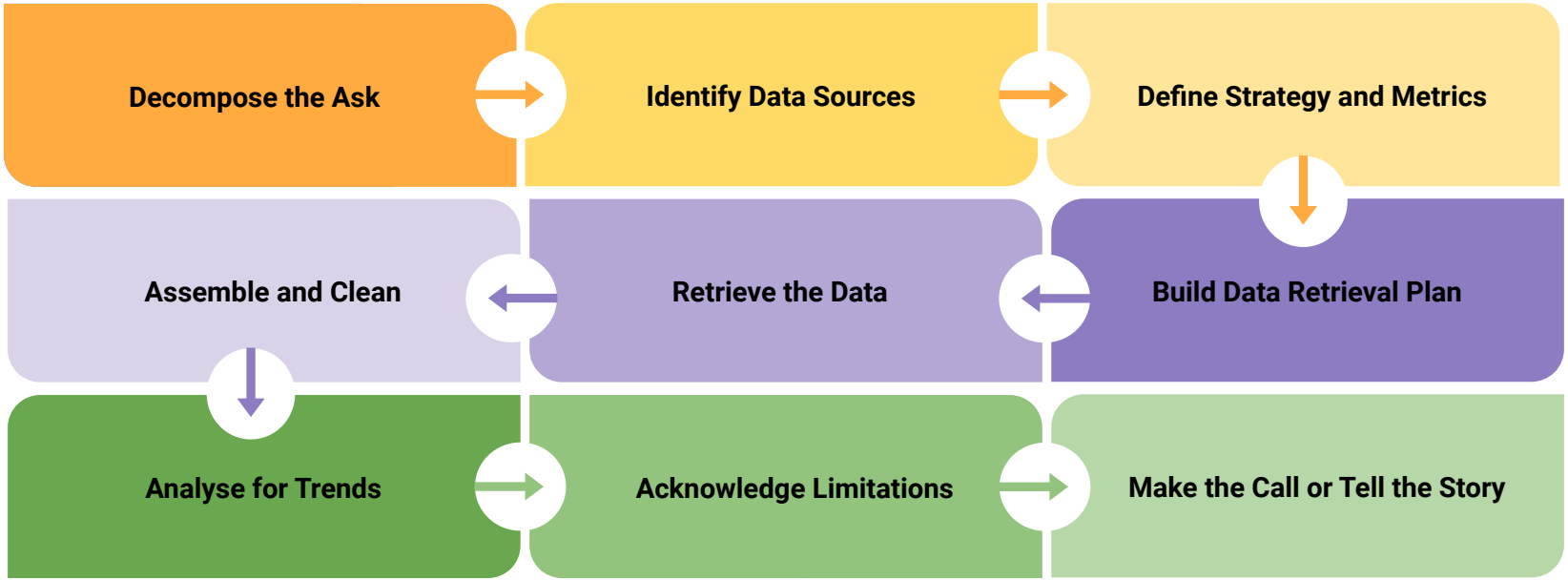
Making the Call

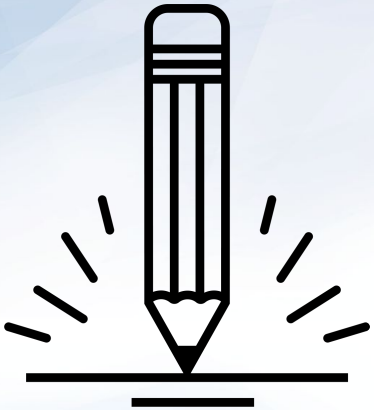
The ‘Let’s Be Real’ Conclusion: Italian (but it’s going to be close)

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 by which you would identify which neighborhoods in our city are seeing signs of gentrification.

Suggested Time:
13 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 visualisation 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?

Suggested Time: 13 minutes

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 are actively looking at it.

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

04 Figure out where class videos will be posted.

Questions?