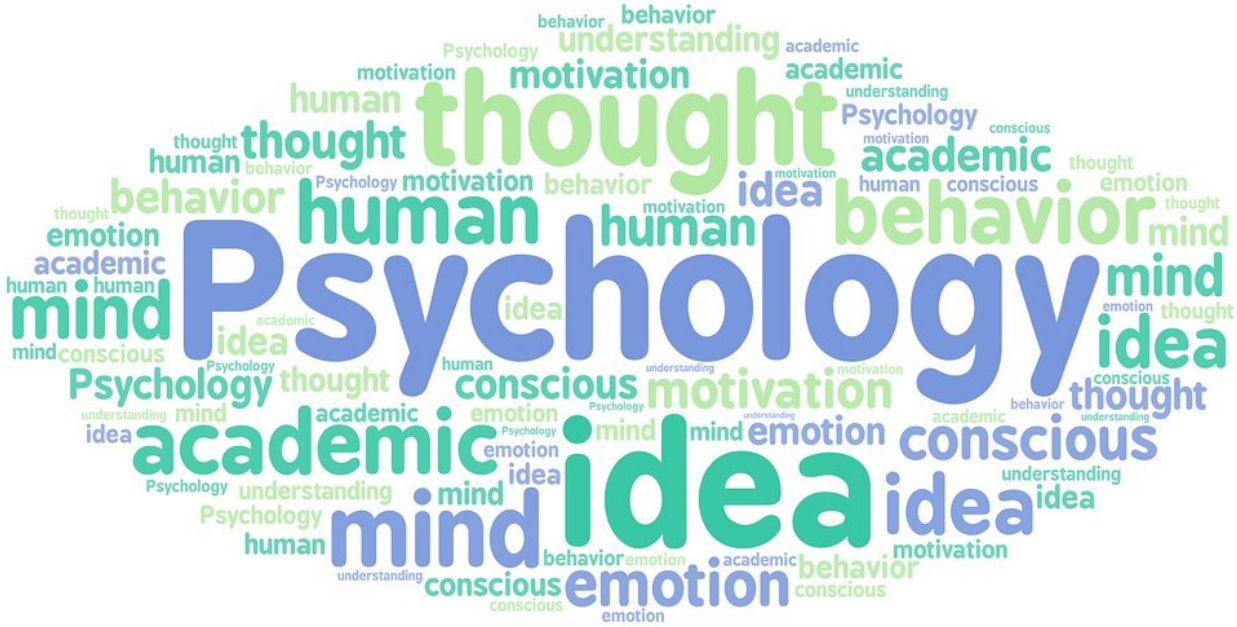
Introduction to Reproducible Science: *Day 3*

Nelson Roque, PhD



Scan for
Slides & Code
available on GitHub

Agenda: Day 3

- Text mining
  - **Skill 1: word and bigram frequency analysis**
  - **Skill 2: generating wordclouds**
  - **Skill 3: sentiment analysis**

- Interacting with APIs and JSON data
  - **Skill 4: querying API for results and data aggregation**

- **Closing Discussion & Q/A**

What is NLP?

- Examples of NLP tasks to make sense of text/voice data include
  - Speech Recognition
    - Speech to text
  - Part of speech tagging
    - identifies ‘make’ as a verb in ‘I can make a paper plane,’
    - and as a noun in ‘What make of car do you own?’
  - Word sense disambiguation
    - E.g., distinguish the meaning of the verb 'make' in ‘make the grade’ (achieve) vs. ‘make a bet’ (place)
  - Named entity recognition
    - E.g., Identifying Florida as a location
  - Co-reference resolution
    - identifying if and when two words refer to the same entity
  - Sentiment Analysis
    - extract subjective qualities — attitudes, emotions, sarcasm, confusion, suspicion
  - Natural language generation
    - Creating human language based on structured information

What is NLP?

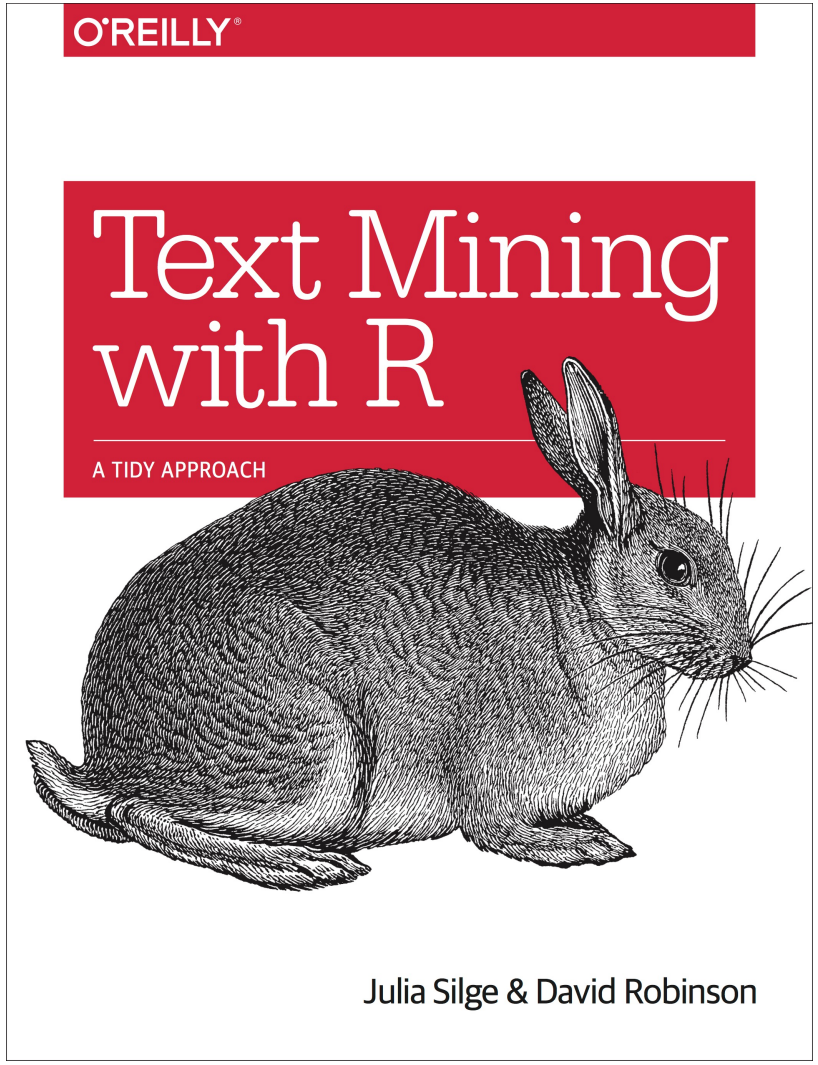
- Methods from various disciplines to enable computers to understand human language in both written and verbal forms:
  - computer science,
  - artificial intelligence,
  - linguistics,
  - data science

What is Text Mining?

- Analyzing text data for key topics, trends, and hidden relationships
  - Word frequency analysis
  - Wordclouds
  - Topic modeling

- Especially important since, 80% of the world's data is in an unstructured format

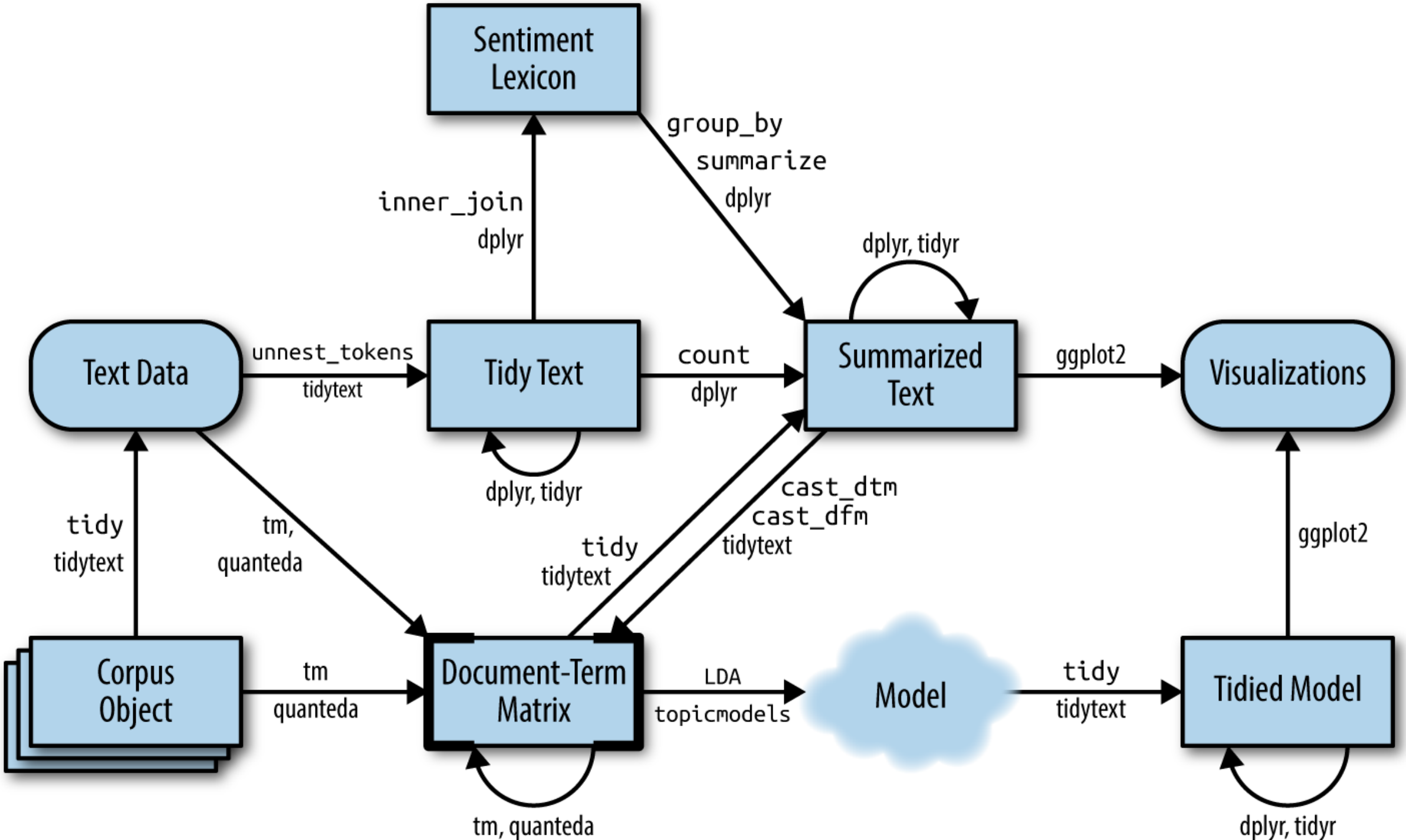
- Consider the data processing implications of your data source
  - Correcting misspellings
  - Correcting incorrect transcriptions (from audio/video)



https://www.tidytextmining.com

Terminology

- Bigrams
  - Pairs of words
- Ngrams
  - A set of words with length of N
- Bag of words
  - Count frequency of words in document
- tf-idf (term frequency - inverse document frequency)
  - Metric for determining how important a term is in a document
  - Determine
    - two documents are similar by comparing their TF-IDF vector using cosine similarity.
    - Keywords to summarize an article
- Stemming
  - Extract root words (like: like, liking, likely)
- Stop words
  - Words like ‘a’, ‘the’ – frequent but not always useful for NLP tasks
  - Lexicons: onix, SMART, snowball


Sentiment Lexicon
group_by
summarize
dplyr
inner_join
dplyr
dplyr, tidyr
Text Data
unnest_tokens
tidytext
Tidy Text
count
dplyr
Summarized Text
ggplot2
Visualizations
dplyr, tidyr
cast_dtm
cast_dfm
tidytext
tidy
tidytext
tm,
quanteda
tidy
tidytext
ggplot2
Corpus Object
tm
quanteda
Document-Term Matrix
LDA
topicmodels
Model
tidy
tidytext
Tidied Model
tm, quanteda
dplyr, tidyr

Latent Dirichlet allocation (LDA)

- One of the most common algorithms for topic modeling guided by two principles:
  - **Every document is a mixture of topics.**
    - We imagine that each document may contain words from several topics in particular proportions.
    - For example, in a two-topic model we could say “Document 1 is 90% topic A and 10% topic B, while Document 2 is 30% topic A and 70% topic B.”
  - **Every topic is a mixture of words.**
    - For example, we could imagine a two-topic model of American news, with one topic for “politics” and one for “entertainment.”
    - The most common words in the politics topic might be “President”, “Congress”, and “government”, while the entertainment topic may be made up of words such as “movies”, “television”, and “actor”.
    - Importantly, words can be shared between topics; a word like “budget” might appear in both equally.
- LDA is a mathematical method for estimating both of these at the same time: finding the mixture of words that is associated with each topic, while also determining the mixture of topics that describes each document.

https://www.tidytextmining.com/topicmodeling.html

Sentiment Analysis: Lexicons

- **bing**
  - “positive”/“negative” classification
- **AFINN**
  - score between -5 (most negative) and 5 (most positive)
- **loughran**
  - “positive”/“negative”/“litigious”/“uncertainty”/“constraining”/“superflous” classification
- **nrc**
  - binary “yes”/“no” for categories positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust

Skill 1: Word and bigram frequency analysis

Skill 2: Generating wordclouds

Skill 3: Sentiment analysis

The Data

- For this example, we will work with real and fake product review data from the following open science resource: https://osf.io/tyue9

- Salminen, J., Kandpal, C., Kamel, A. M., Jung, S., & Jansen, B. J. (2022). Creating and detecting fake reviews of online products. Journal of Retailing and Consumer Services, 64,102771. https://doi.org/10.1016/j.jretconser.2021.102771

Let’s jump into R

Scan the QR code to view the code we will work with

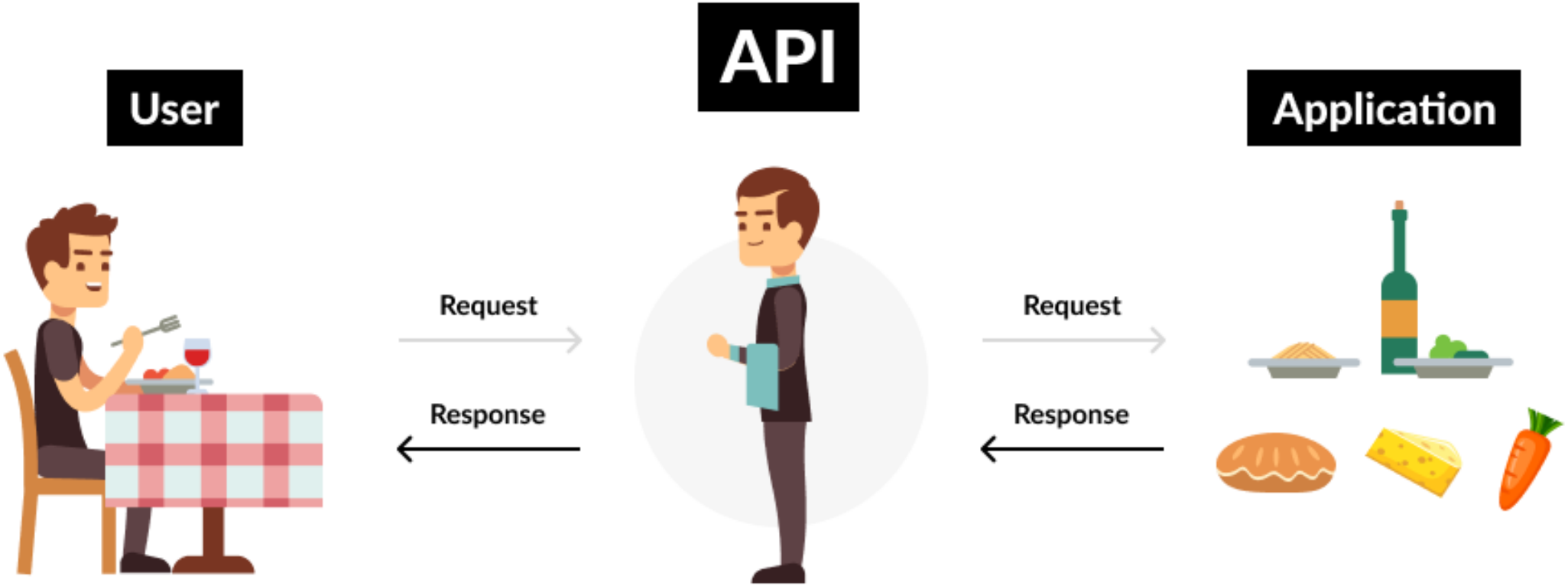
Skill 4: Working with APIs

What is an API?

- **A**pplication **P**rogramming **I**nterface
  - Method to communicate with the data layer of an application to accomplish business / research function, e.g.,
    - Store and query data
    - Update settings
    - Login/logout
    - Retrieve resources (e.g., images)

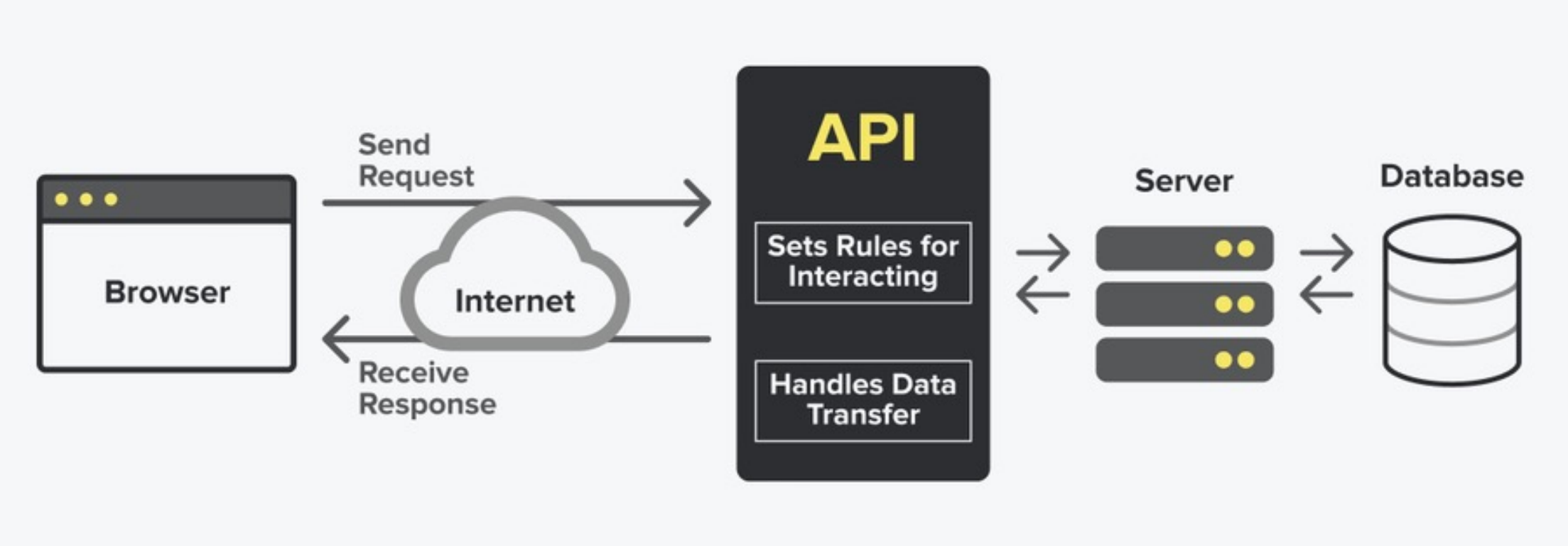
  - APIs make the modern world go round
    - From airlines, to your groceries, to functionality in almost every mobile app, APIs are charged with securely and efficiently shipping data back and forth between entities and their consumers.

What is an API?



https://www.mindtree.com/insights/blog/rise-apis-financial-services-key-future

What is an API?



https://snipcart.com/blog/integrating-apis-introduction

Jeff Bezos' API Mandate at Amazon

1. All teams will henceforth expose their data and functionality through service interfaces.
2. Teams must communicate with each other through these interfaces.
3. There will be no other form of interprocess communication allowed: no direct linking, no direct reads of another team's data store, no shared-memory model, no back-doors whatsoever. The only communication allowed is via service interface calls over the network.
4. It doesn't matter what technology they use.
5. All service interfaces, without exception, must be designed from the ground up to be externalizable. That is to say, the team must plan and design to be able to expose the interface to developers in the outside world. No exceptions.
6. Anyone who doesn't do this will be fired.
7. Thank you; have a nice day!

https://nordicapis.com/the-bezos-api-mandate-amazons-manifesto-for-externalization

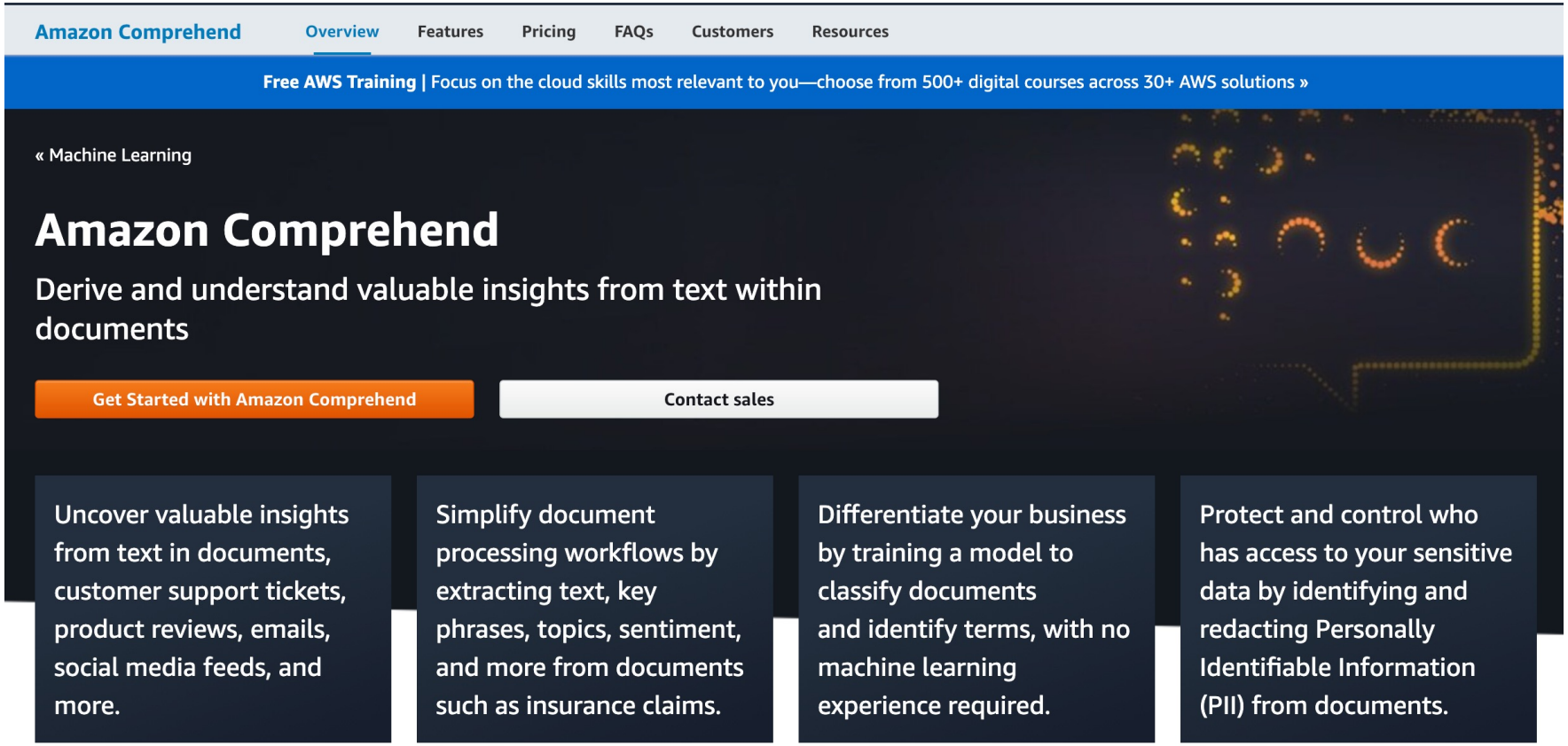
Special Considerations

- Authentication
- Rate-limiting
- Query by Offsets, and limits

Let’s jump into R

Scan the QR code to view the code we will work with

NLP as a Service (NLPaaS)



Thank you!

- Closing Discussion

- Q&A
  - What other topics could be helpful in supporting your research?
    - Working with image data?
    - Video data?
    - Geographic data + geocoding?
    - Emotion detection?
    - Machine learning?

You are not alone

thank you

nelson.roque@ucf.edu

nelsonroque.com

https://github.com/nelsonroque