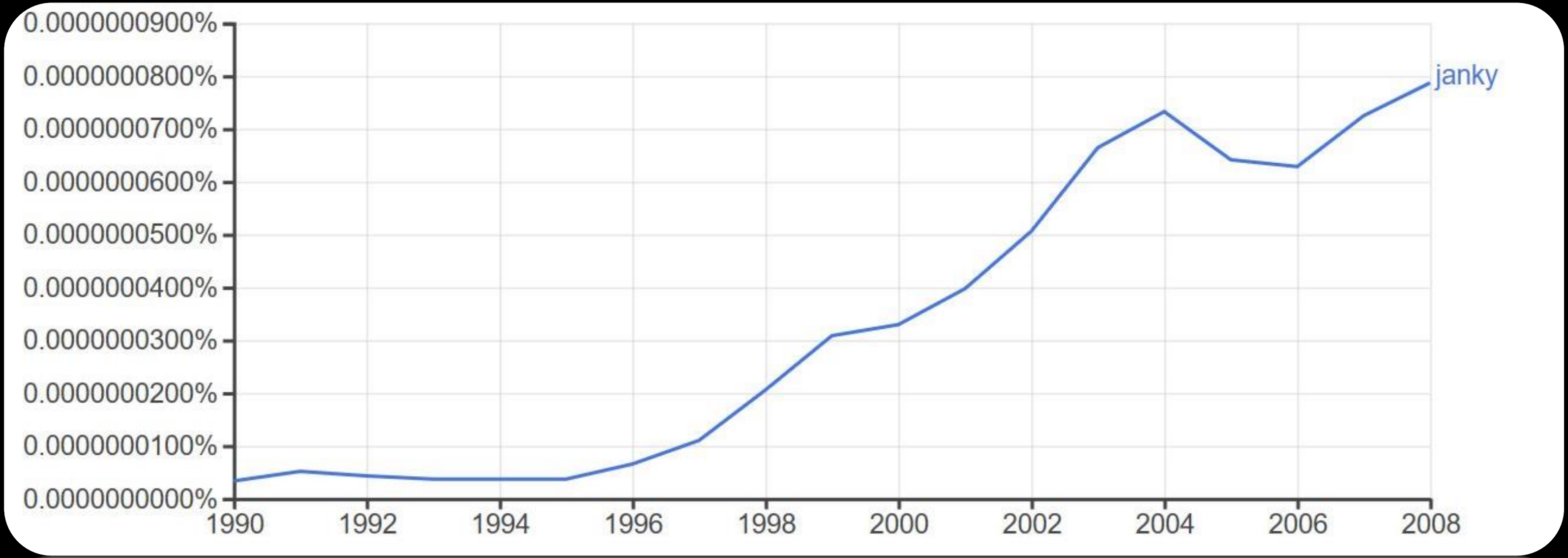
Text analysis

≈ Natural Language Processing, or “*How to do cool stuff with words.*”

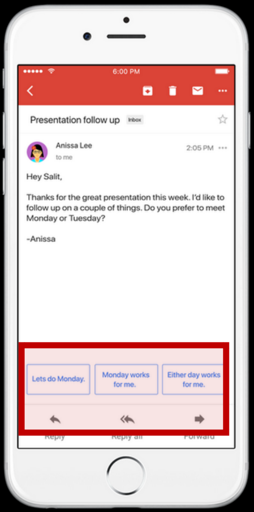


Emily Rae Sabo Data Camp | June 19, 2019

2 objectives for this session:

- ✓ What is NLP /Text Analysis and why would I use it?
- ✓ What tools are out there for me to use?

What is NLP used for?



Presentation follow up

Anissa Lee 2:05 PM
to me

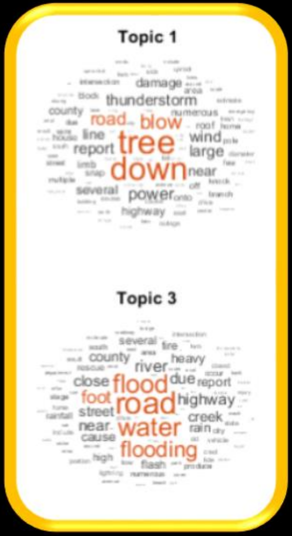
Hey Salit,

Thanks for the great presentation this week. I'd like to follow up on a couple of things. Do you prefer to meet Monday or Tuesday?

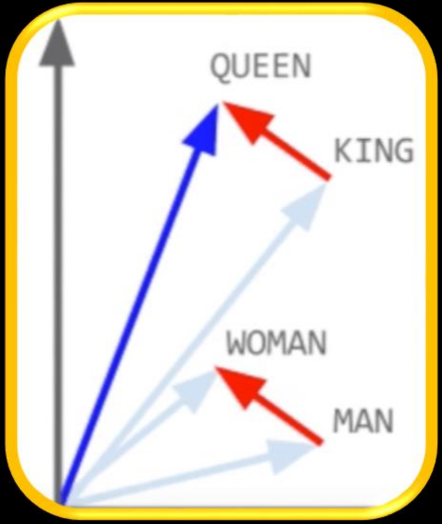
-Anissa

Predicting language

Translating language



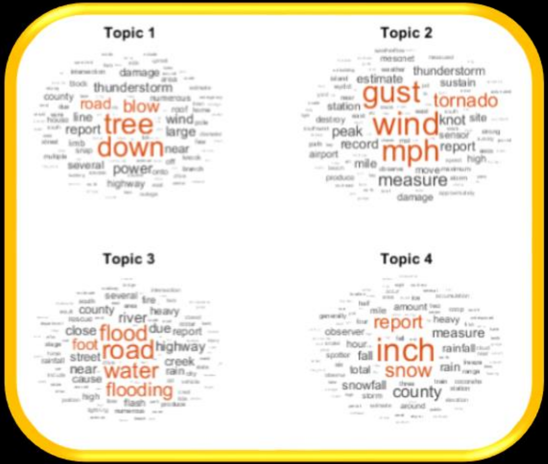
Finding patterns in language



Measuring meaning in language

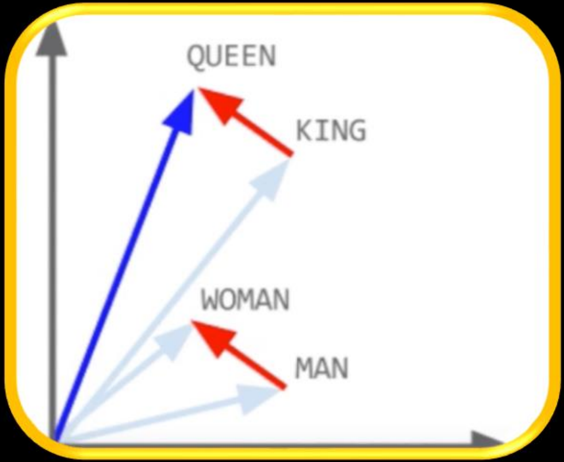
How to apply Text Analysis

Finding patterns in language



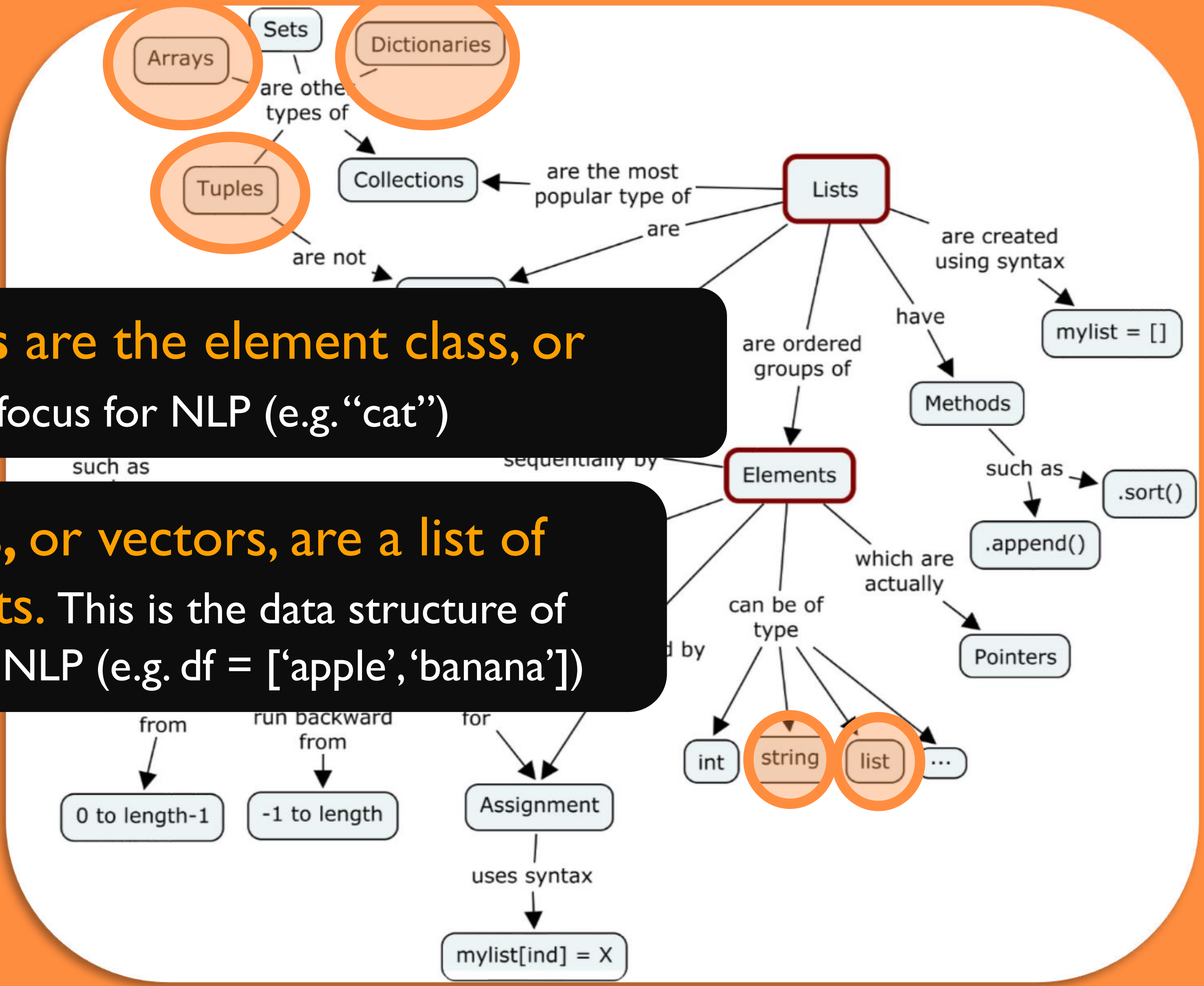
- Change over time with Google Ngram
- Topic Modeling with Gensim, NLTK
- String matching and token extraction with RegEx

Measuring meaning in language



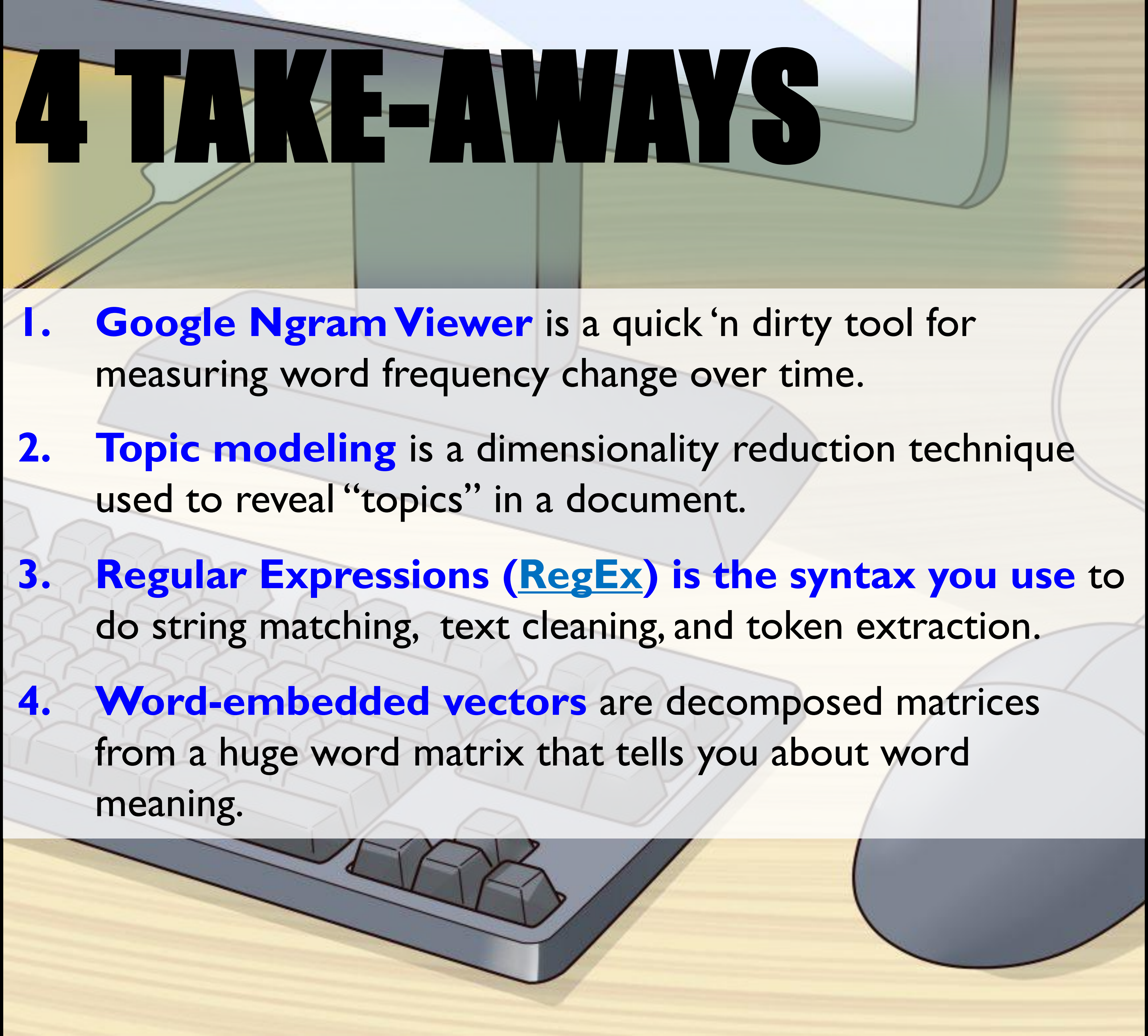
- Vector space modeling with word-embedded vectors like Word2Vec in *Gensim* or GloVe in *SpaCy*

Python's basic elements & data structures



Strings are the element class, or type, of focus for NLP (e.g. "cat")

Arrays, or vectors, are a list of elements. This is the data structure of focus for NLP (e.g. df = ['apple', 'banana'])

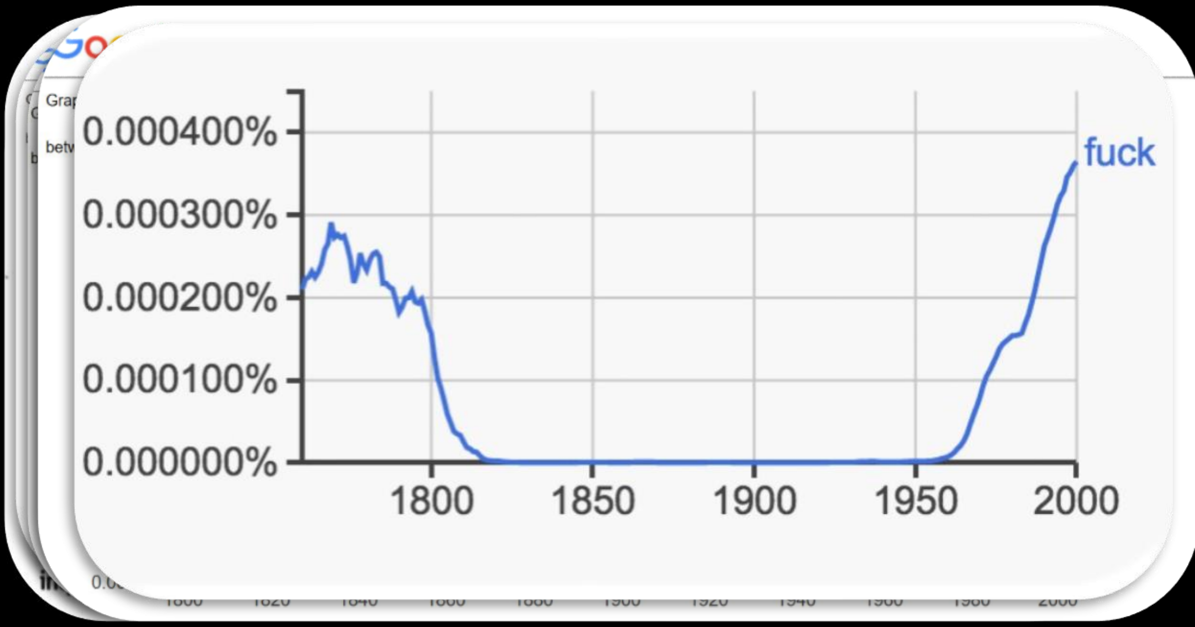


4 TAKE-AWAYS

1. **Google Ngram Viewer** is a quick 'n dirty tool for measuring word frequency change over time.
2. **Topic modeling** is a dimensionality reduction technique used to reveal "topics" in a document.
3. **Regular Expressions (RegEx) is the syntax you use** to do string matching, text cleaning, and token extraction.
4. **Word-embedded vectors** are decomposed matrices from a huge word matrix that tells you about word meaning.

How to measure changes in word frequency over time?

Google Ngram Viewer



- The founding tool of "culturomics"
- Advantages vs. limitations?
- Share one way you could imagine using this in your research.
- Go and play!
  - https://books.google.com/ngrams
  - https://books.google.com/ngrams/info

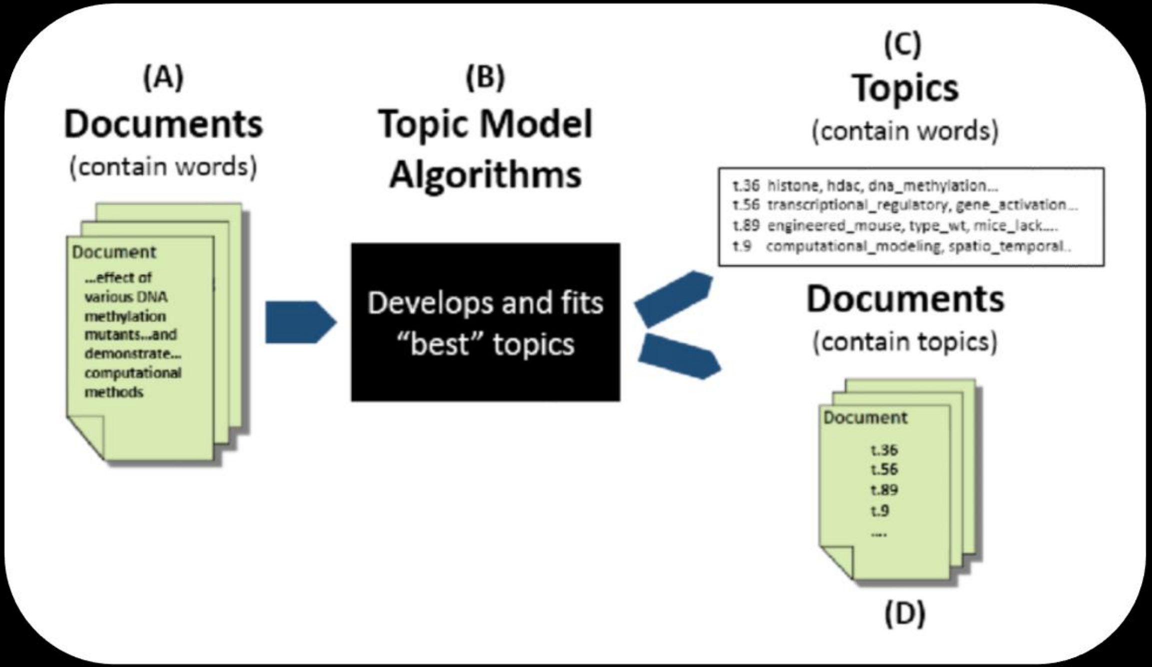
What is Topic Modeling?

It is an **unsupervised approach** used for finding and observing the bunch of words (called "topics") in large clusters of texts."
Bansal (2016)

Click here for a good starter on Topic Modeling in Python with NLTK and Gensim

- It's a dimensionality reduction technique used to discover the hidden or abract "topics" that occur in a document or collection of documents.

- Techniques you may have heard of before: LSA (Latent Semantic Analysis) and LDA (Latent Dirichlet Allocation)



(A)
Documents
(contain words)

Document
...effect of various DNA methylation mutants...and demonstrate... computational methods

(B)
Topic Model Algorithms

Develops and fits "best" topics

(C)
Topics
(contain words)

t.36 histone, hdac, dna_methylation...
t.56 transcriptional_regulatory, gene_activation...
t.89 engineered_mouse, type_wt, mice_lack....
t.9 computational_modeling, spatio_temporal..

Documents
(contain topics)

Document
t.36
t.56
t.89
t.9
....

(D)

What are Regular Expressions, or RegEx?

1. **Work through one tutorial:** https://regexone.com/ https://www.tutorialspoint.com/python/python_reg_expressions.htm

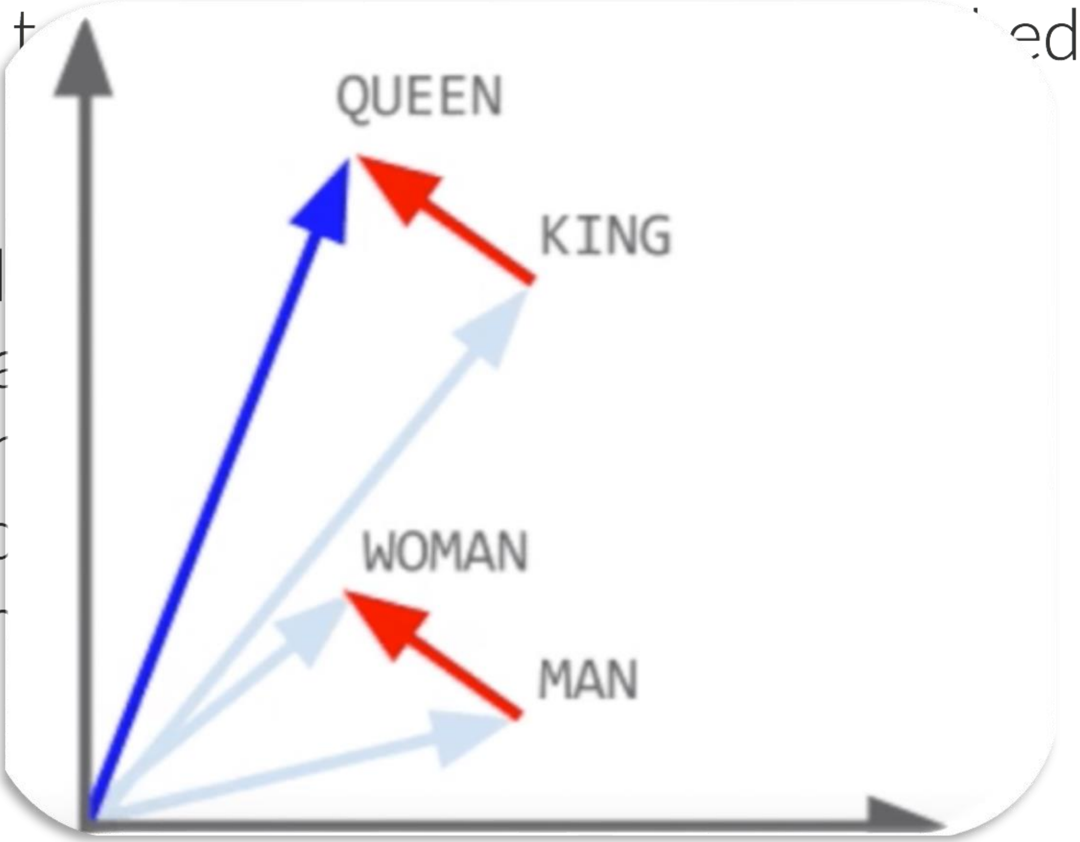
2. Then, open Jupyter, create your own mini-corpus (~20 words) and **write RegEx code to match a string from your corpus.**

WHENEVER I LEARN A NEW SKILL I CONCOCT ELABORATE FANTASY SCENARIOS WHERE IT LETS ME SAVE THE DAY.
OH NO! THE KILLER MUST HAVE FOLLOWED HER ON VACATION!
BUT TO FIND THEM WE'D HAVE TO SEA THROUGH 200 MB OF EMAILS LOOKING FO SOMETHING FORMATTED LIKE AN ADDRESS!
IT'S HOPELESS!
EVERYBODY STAND BACK.
I KNOW REGULAR EXPRESSIONS.
PERL!
TAP TAP

Pro-tip reminders: Be computational *and* creative in your approach. There are an infinite number of ways to accomplish a string matching task!

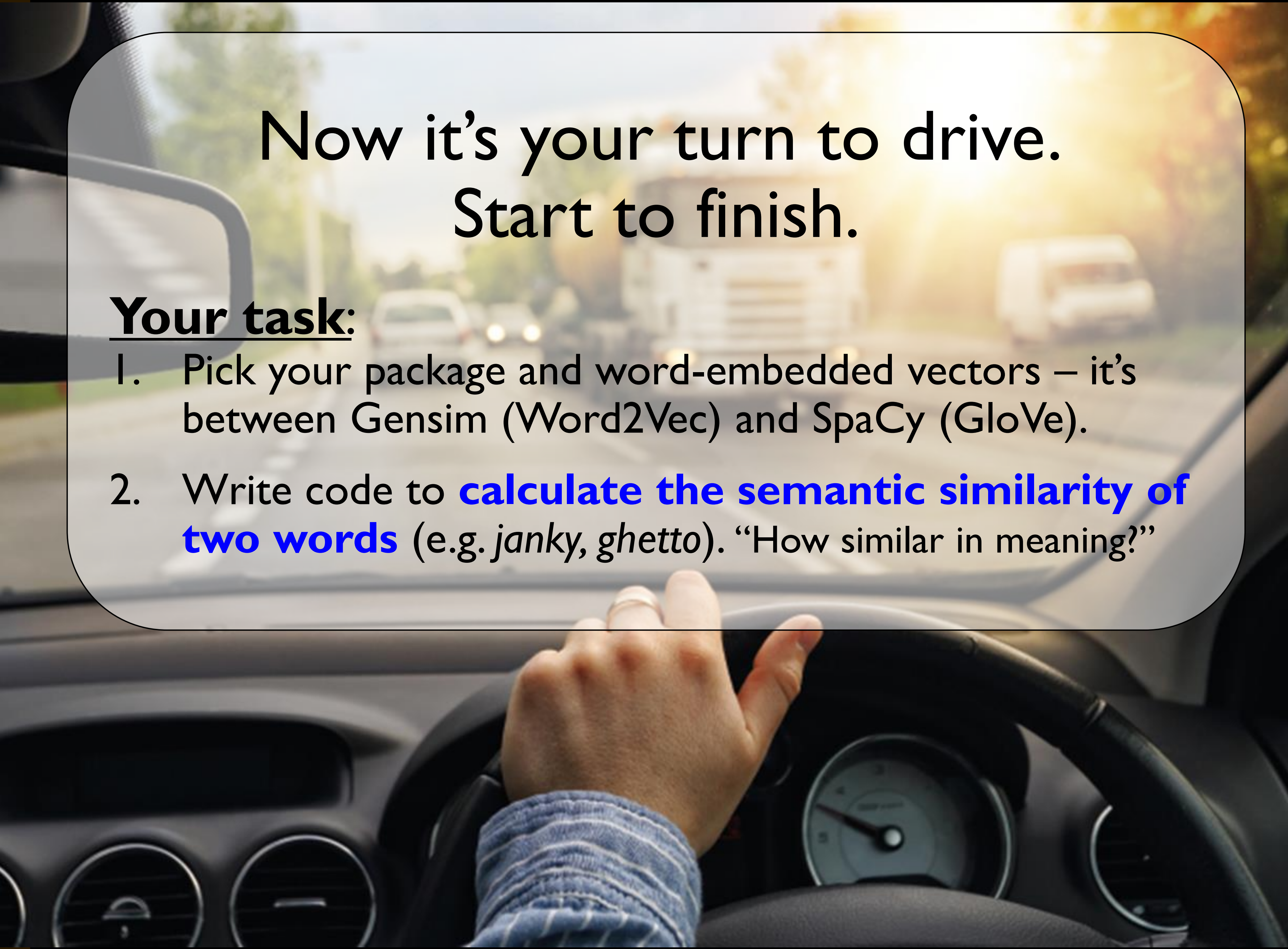
Vector Space Modeling, Word-embedded vectors & Cosine Similarity

Try t edding



Vocabul

Man, woma
girl, prir
princess, c
king, mor



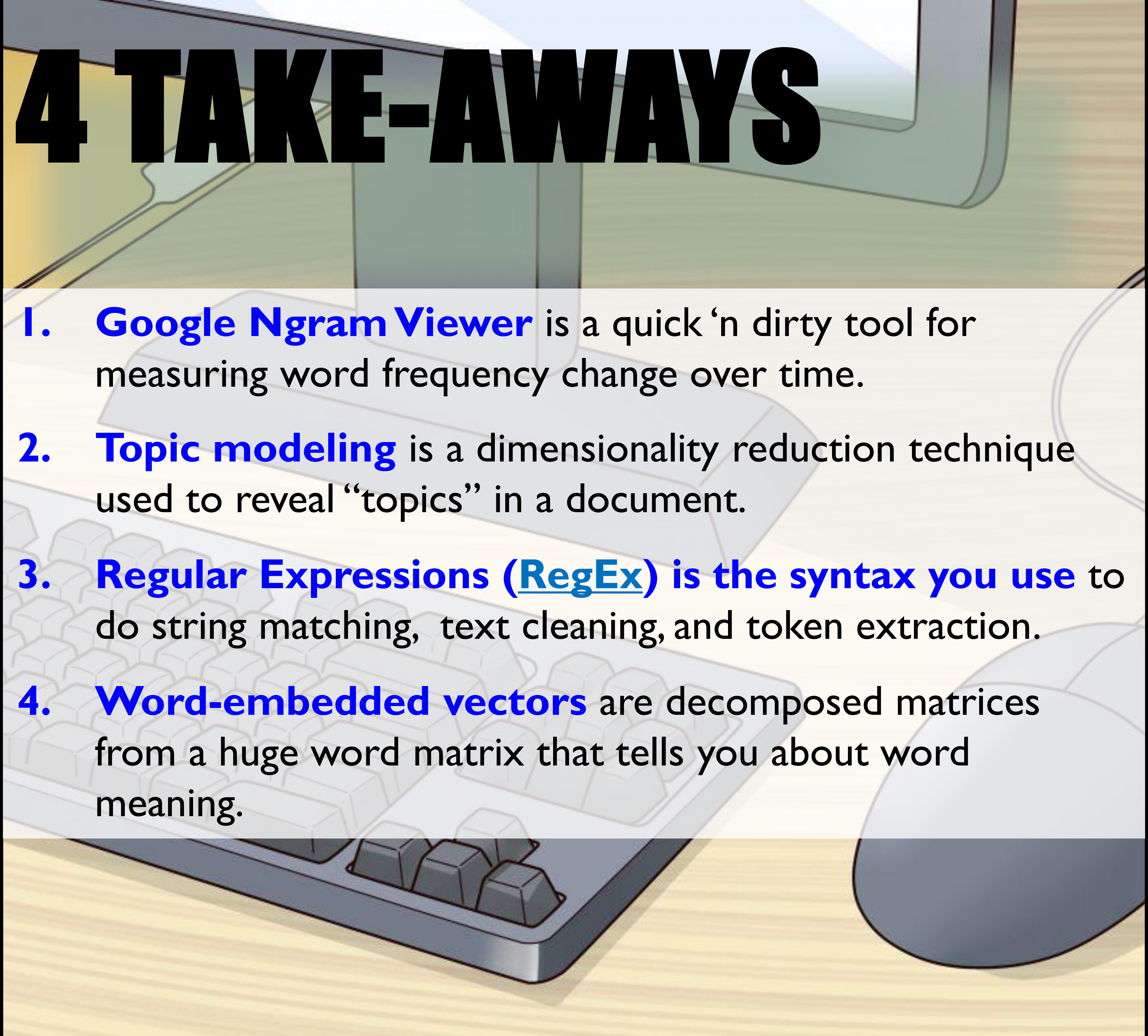
Now it's your turn to drive.
Start to finish.

<u>Your task</u>:

1. Pick your package and word-embedded vectors – it's between Gensim (Word2Vec) and SpaCy (GloVe).

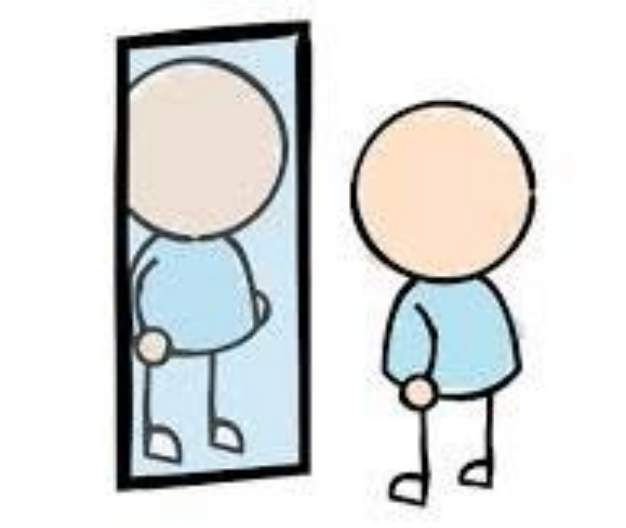
2. Write code to **calculate the semantic similarity of two words** (e.g. *janky, ghetto*). "How similar in meaning?"

4 TAKE-AWAYS



1. **Google Ngram Viewer** is a quick 'n dirty tool for measuring word frequency change over time.
2. **Topic modeling** is a dimensionality reduction technique used to reveal "topics" in a document.
3. **Regular Expressions (RegEx) is the syntax you use** to do string matching, text cleaning, and token extraction.
4. **Word-embedded vectors** are decomposed matrices from a huge word matrix that tells you about word meaning.

CHECK-IN:



1. So far, what is the most insightful thing you've learned during camp?
2. What is the one thing that's still the muddiest for you?

Thank you!

Come to a FREE Nerd Nite talk I'm doing about linguistics on Thursday, June 20th at LIVE, 7pm:

The 13 Things You Need to Know about Language.

Emily Rae Sabo
@StandupLinguist

LANGUAGE MATTERS
The University of Michigan