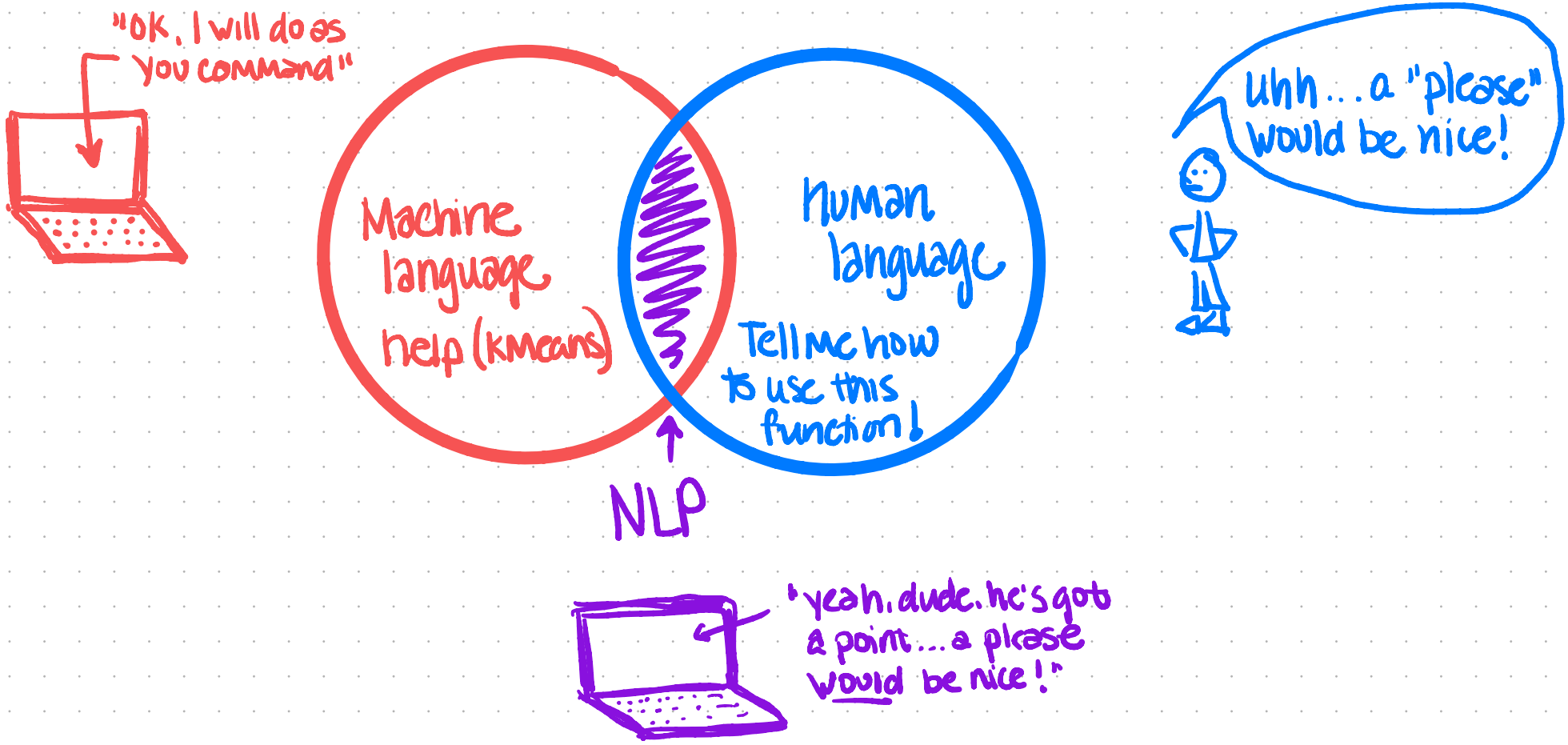
NLP Natural Language Processing

Processing & analyzing large amounts of natural language data via programming.



USE CASES for NLP

Voice of the Customer — analysis of surveys, tweets, other feedback
What drives +/- responses? What is being asked for?
What drives engagement?

Semantic Search — a search engine that searches using intent & context rather than just key words.

Knowledge Management & Discovery — tag documents, blog posts, etc. based on the topics for ease of recommending, finding, discovering.

Customer support — Route support requests based on needs / topic identified through the text contained in the requests.
Chat bots to solve the common issues before routing to a human.

Security — Phishing emails. national security / terrorism on Social Media, dark web
Cyber bully

Healthcare — New cells of viruses/flu through SM monitoring, Mental health via SM

Virtual Assistants — Siri, Alexa, googlehome

Translation, Speech to text. competitive intel. Spam

Spell check & Autocorrect, Auto completion,

NLP Module Goals & Agenda

- use **REGULAR Expressions** to find/extract/replace/substitute text that matches a pattern.
- Acquire data through **WEB SCRAPING**
- Parse text: **NORMALIZE, TOKENIZE**
  STEM, LEMMATIZE, remove **STOPWORDS**
- Explore text data **WORDCLOUDS, N-GRAMS**
- Create features using **BAG OF WORDS, TF-IDF**
- **PROJECT**

Types of NLP Programs

Sentiment Analysis

Topic Modeling

Text Classification

"Data Science is fun"
- No label - extract topics (LDA)
- tag = DataScience LABEL!

Translation

Google's Crowdsource App → Get labeled data
Also, some Rule-based using dictionaries

Word Embeddings
Document Embeddings

find similar articles, forum question duplication

	Supervised	Unsupervised
Sentiment Analysis	☑	☑
Topic Modeling	☒	☑
Text Classification	☑	☒
Translation	☑	☑
Word Embeddings / Document Embeddings	☒	☑

WRANGLING in NLP

MODULES

Acquire your **CORPUS** of documents 

REQUESTS (get the HTML) → BEAUTIFUL SOUP (parse the HTML)

Convert your corpus to individual **DOCUMENTS**

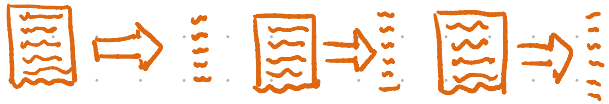
NORMALIZE text

UNICODEDATA
RE

He went to the store to buy ~~A~~ rolls of toilet paper~~.~~ He was really na~~i~~ve to think there would be toilet paper~~.~~

↓

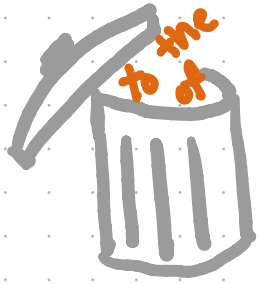
he went to the store to buy rolls of toilet paper he was really
nave to think there would be toilet paper

Break documents into **TOKENS** 

NLTK.tokenize.toktok

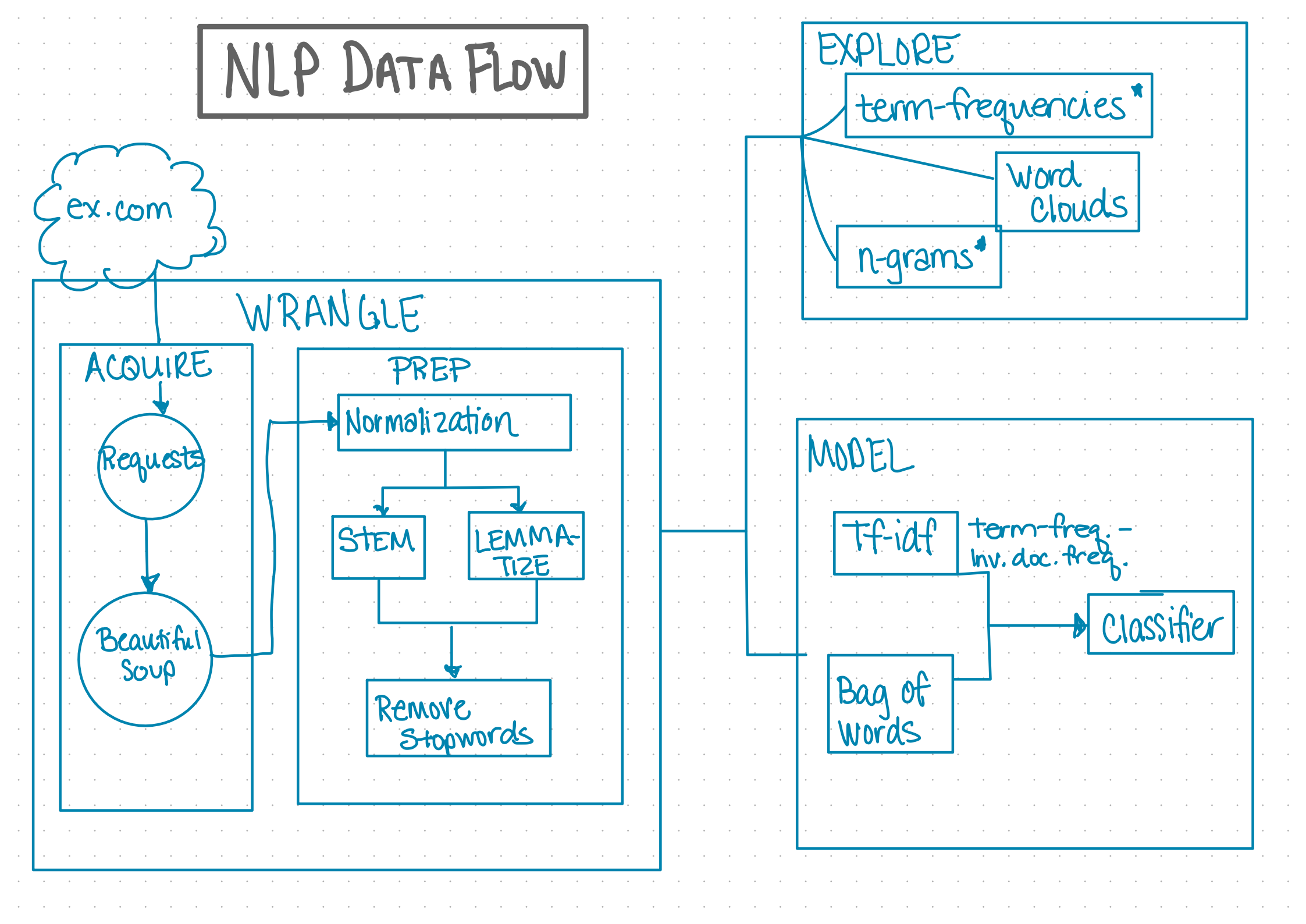
Create **STEMS** or **LEMMAS** went → go rolls → roll was → is

NLTK.porter
NLTK.stem

Remove **STOPWORDS** 

NLTK.stopwords

EXPLORE


NLP DATA FLOW
ex.com
WRANGLE
ACQUIRE
Requests
Beautiful Soup
PREP
Normalization
STEM
LEMMA-TIZE
Remove Stopwords
EXPLORE
term-frequencies*
word clouds
n-grams*
MODEL
Tf-idf
term-freq. - inv. doc. freq.
Classifier
Bag of words

NLP Methods

	PROS	CONS
Rule-based	Explicit Logic, Human Readable	Time to dev, Maintain, doesn't Scale well.
Statistical (e.g. Bayesian)	Automatable, Easy to develop	Doesn't generalize well & not flexible. doesn't capture semantics
Vector Methods (e.g. word2vec)	Uses common and easily accessible ML algorithms	Doesn't capture document structure, or the "Big Picture" well.
Deep Learning / Neural Nets	Good results, little Manual effort.	Black box, Risk of discrimination, biased decisions Made es result, unexplainable, high computational cost.

SOURCE: textician.com

NLP Vocabulary

ENTITY any word or series of words that consistently refers to the same thing.

CORPUS body of docs, entire sample using to analyze

DOCUMENT - single observation

STEM - removing prefixes, suffixes, to the core string. (easing → eas)

LEMMA - the base word. (easing → ease)

TOKEN - single linguistic unit

CORPUS → DOCUMENT → TOKEN