Distributed
Natural Language Processing
Systems in Python

Clare Corthell, Founder of Luminant Data
@clarecorthell

Luminant Data

is a Machine Intelligence Consultancy

building more intelligent businesses
with data science strategy and
artificial intelligence technology

luminantdata.com
Clare Corthell, Founder

based in San Francisco

#opensource
data
science
masters

The Open Source Data Science Masters

The most popular curriculum for
learning data science

datasciencemasters.org

Author: Clare Corthell

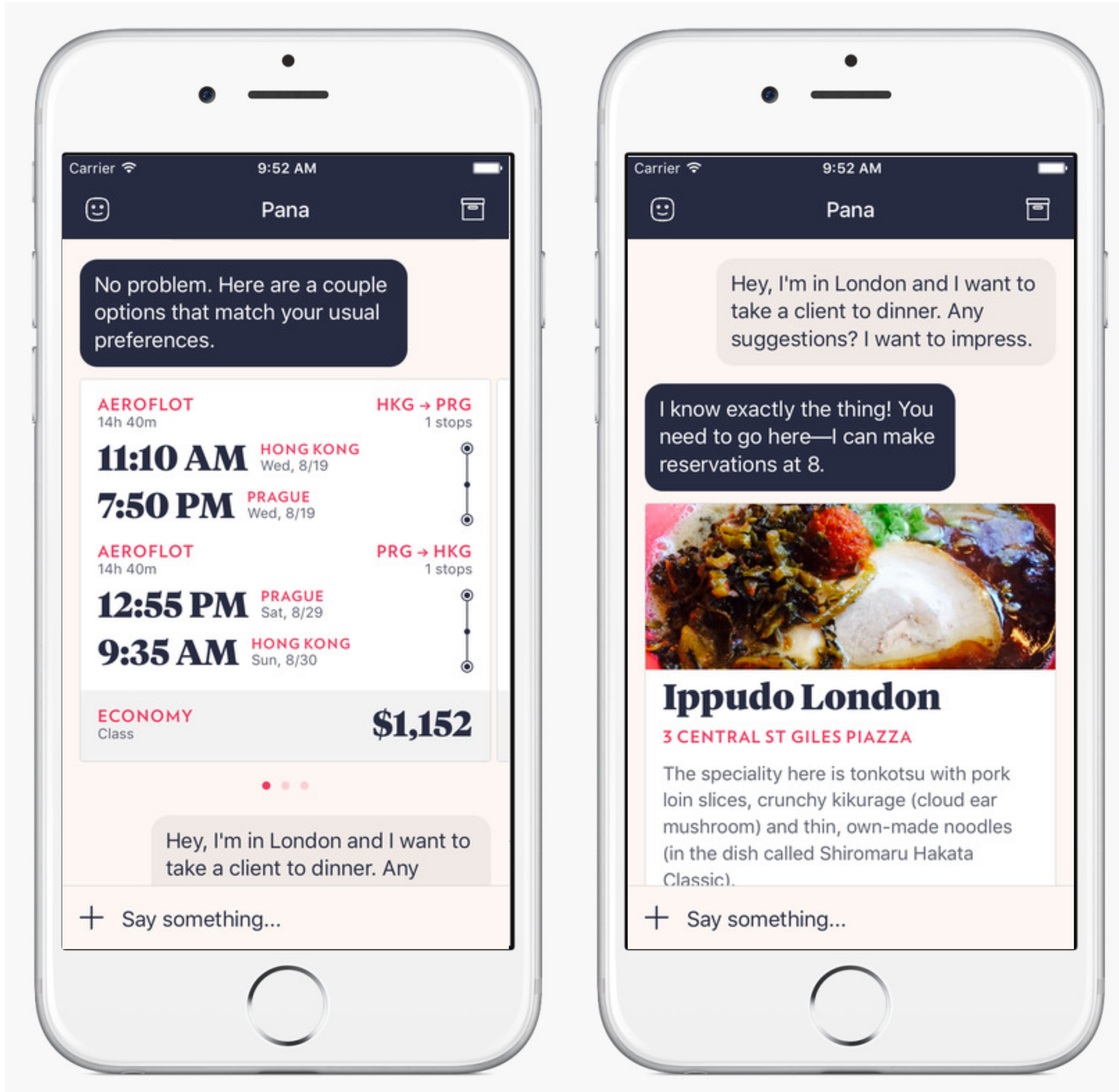
development environment

python 2.7

pip, pandas, iPython, TextBlob, scikit-learn

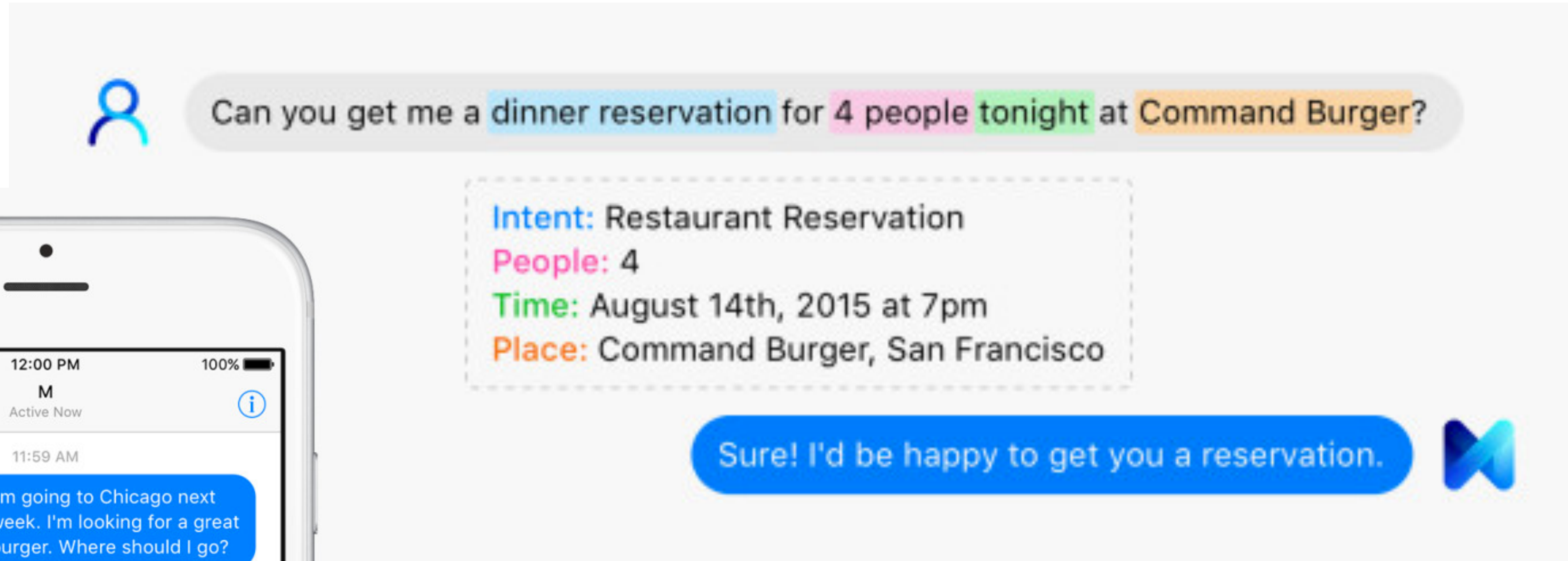
hold on to your seats,
we’re doing NLP end-to-end

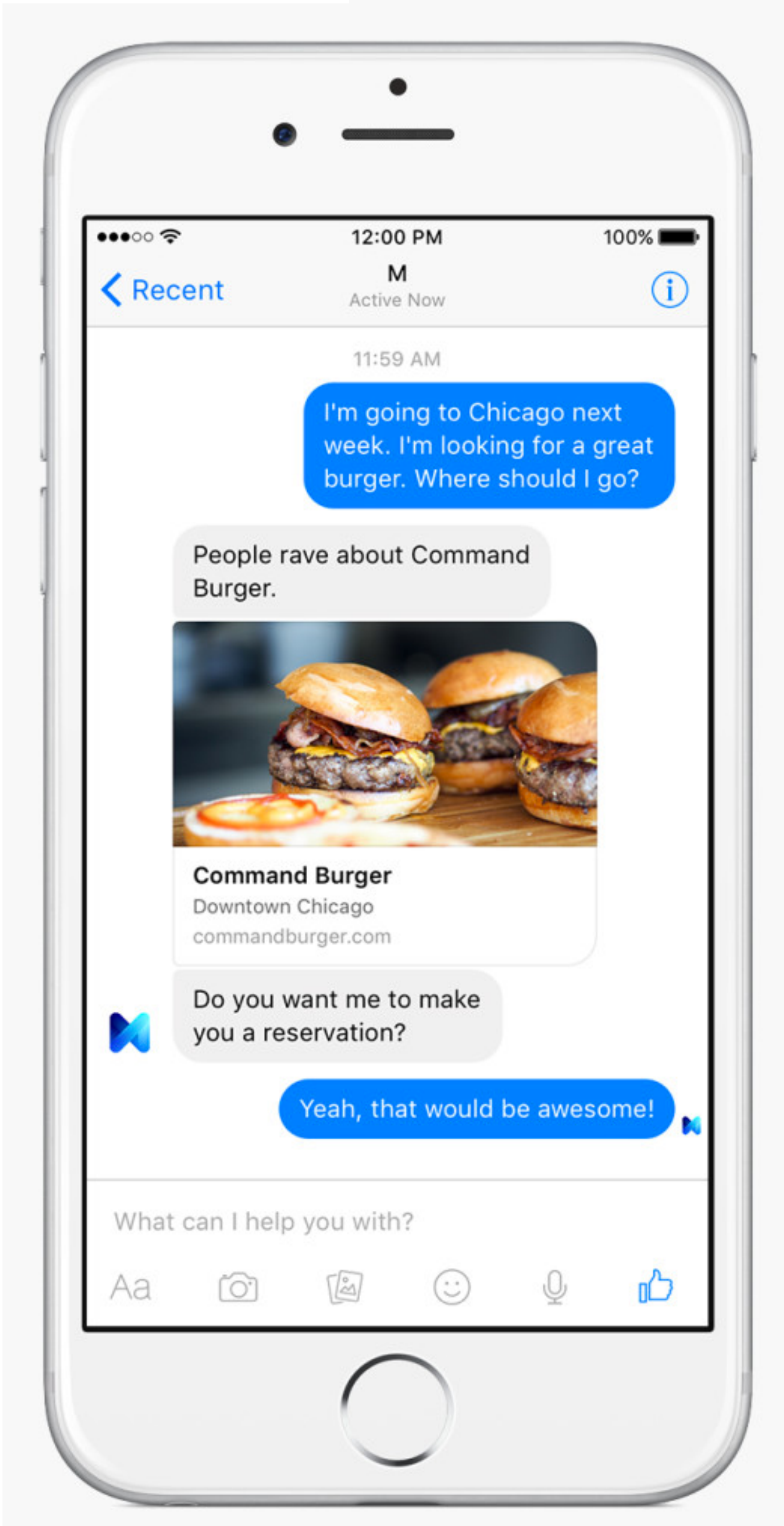
What is natural language processing 4?



how does this artificial assistant **know** what you're talking about?

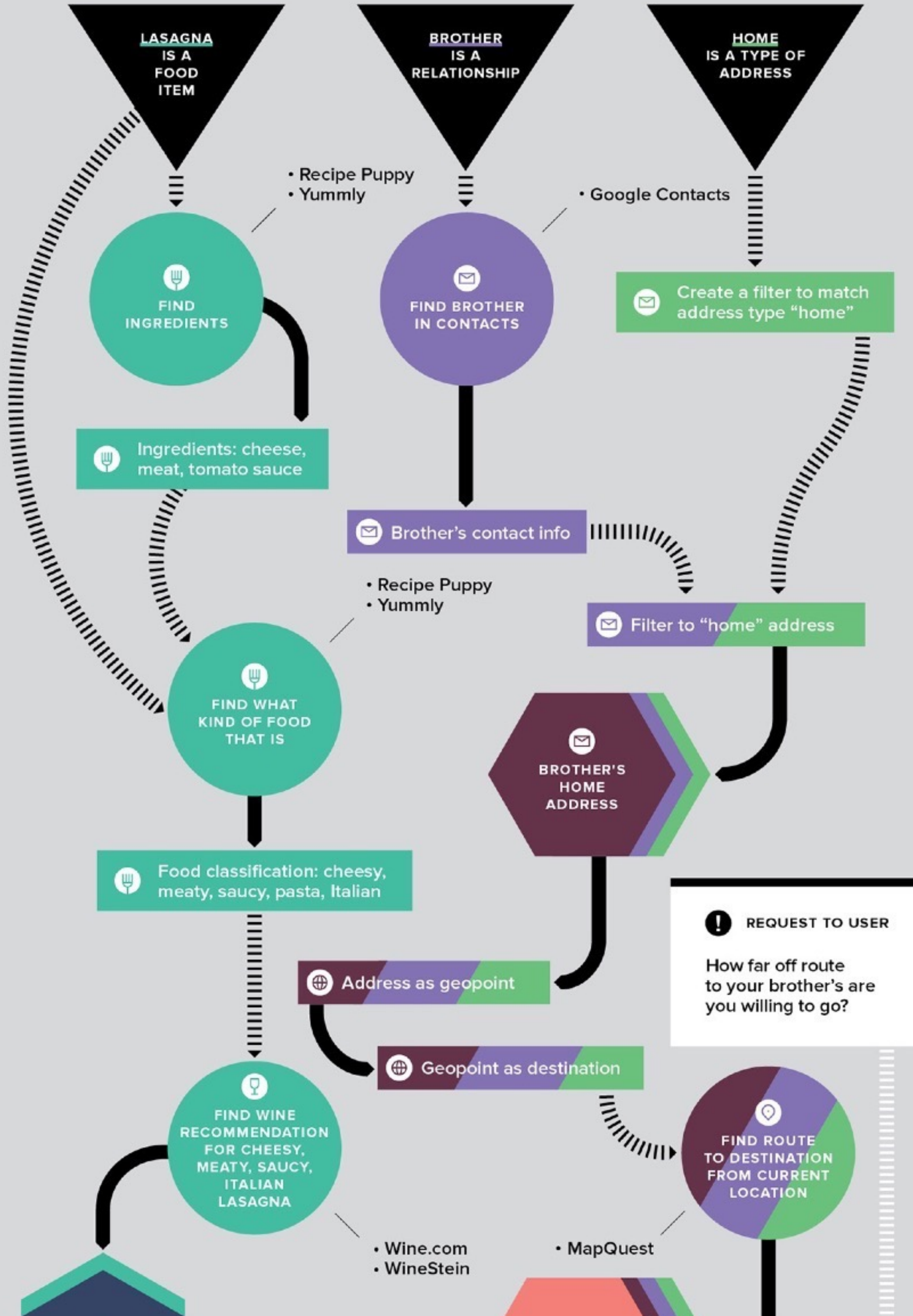
@clarecorthell





It’s simple, right?

On the way to my brother's house, I need to pick up some cheap wine that goes well with lasagna.*



Natural Language Processing is the domain concerned with translating human language into something a computer can process

today —

teaching computers to better
understand human language

Natural Language Processing

Code

1. **Intuition for Text Analysis**
2. **Representations of Text for Computation**

Case Study

3. **Mining the Web for Obscure Words**
4. **NLP in Production**

@clarecorthell

Intuition for Text Analysis

Worked Example:
http://github.com/clarecorthell/nlp_workshop

Representations of Text for Computation

Worked Example:
http://github.com/clarecorthell/nlp_workshop

Case Study

Mining the Web for Obscure Words

Finding new definitions of words for the **Wordnik** dictionary
codename: **Serapis**

wordnik

SERAPIS
SCALABLE
WORD GOBBLER

Important Note:

We combine Natural Language Processing and Machine Learning in this framework

Wordnik's Challenge

Find 1 million new words that are already defined online what does this word mean?

The term “cheeseors” describes flighted globules of intergalactic cheese, known to be the scourge of the asteroid belt.

- Erin McKean, Wordnik Founder

Free-Range Definition or "FRD"

a sentence that contains and contextually defines a word the natural language challenge:

we're given words without definitions (in Wordnik)

we want FRD sentences for that word (from the internet)

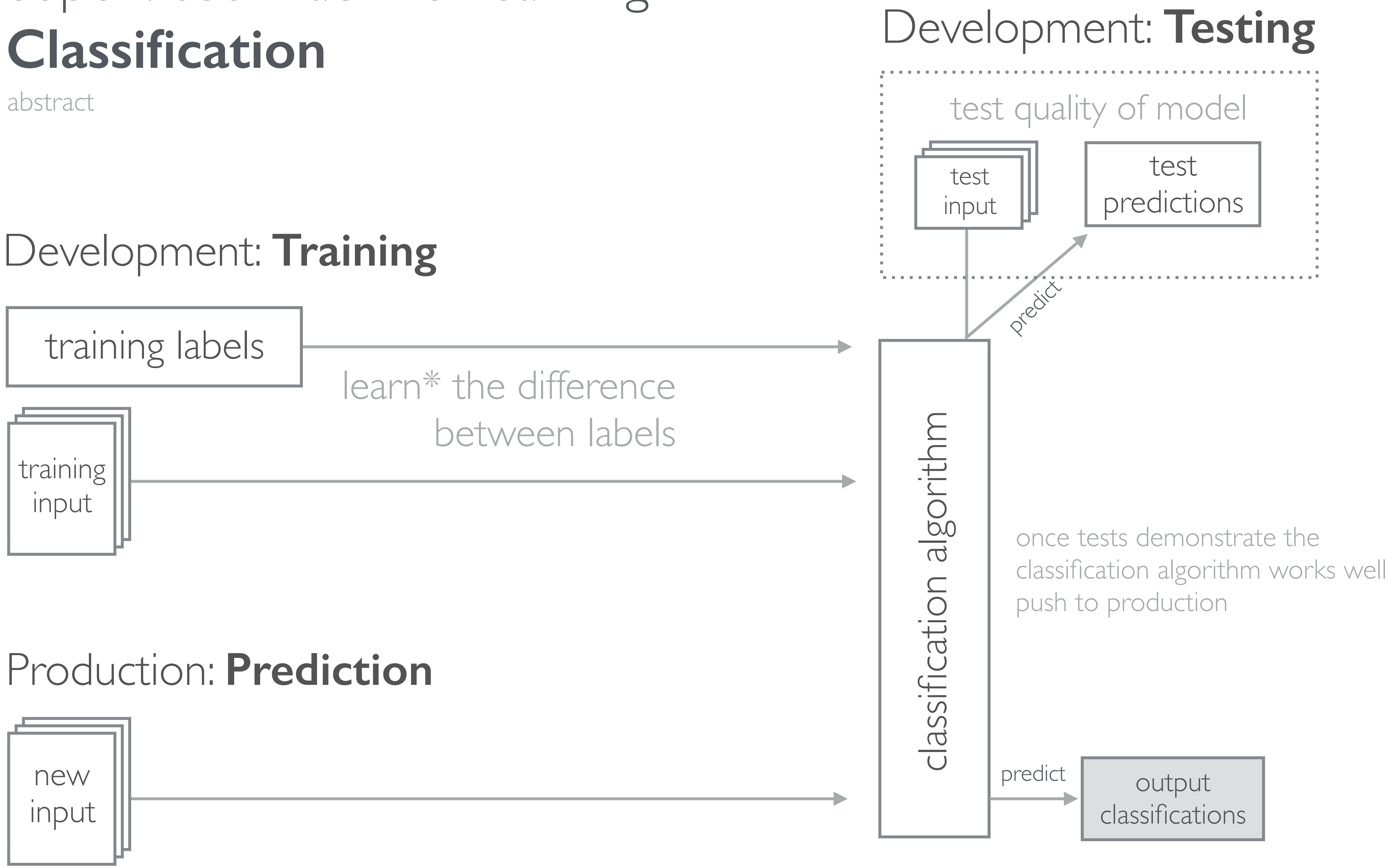
This is a **Supervised Classification Problem**

Given a bunch of sentences containing a given word, we want to know which ones are an FRD and which are not.

P(FRD | Sentence) = [0.0,1.0]

Supervised Machine Learning: **Classification**

abstract



* ***Learning*** means finding the functions that define the difference between training labels, based on training input.

@clarecorthell

To get from word to FRD,
What data do we need?

```
sabatonaeaee
sabatonow
sabatonæ˜¯æ¥è
sabators
sabatos
sabatosays
sabatothrows
sabatoun, sabatyn
sabatoun,+sabatyn
sabatounsabatyn
sabats
sabatt
sabatte
sabattical
sabattus
sabatucci
sabau
sabauda
sabauddin
sabaudo
sabaw
sabawi
sabawoon
sabay
sabaya
sabayon saucecomments
sabayon sauceetymologies
sabayon sauceexamples
sabayon saucestatistics
sabayon+saucecomments
sabayon+saucedefinitions
sabayon+sauceetymologies
sabayon+saucepronunciations
sabayon+saucerelated
sabayone
sabayons
sabayonsauce
sabayonsaucecomments
```

user lookups from Wordnik

real world text data is messy

search the web for

word —> sentence with word


the internet
real world text data is messy
@clarecorthell

“ephemerable”

Google

ephemerable

All Images Maps Videos News More Search tools

About 595 results (0.48 seconds)

Urban Dictionary: ephemerable
www.urbandictionary.com/define.php?term=**ephemerable**
May 24, 2013 - Kindle books are speculated to be **ephemerable** after their mysterious disappearance from users' devices. by yourfatherandfriends May 24, ...

ephemerable
Kindle books are speculated to be ephemerable after their ...

More results from urbandictionary.com »

ephemerable definition | What does ephemerable mean?
definithing.com/**ephemerable**/
ephemerable definition. Having the quality of fleetingness; having the likelihood or ability to become ephemeral Kindle books are speculated to be **ephemerable** ...

#ephemerable hashtag on Twitter
https://twitter.com/hashtag/**ephemerable**
See Tweets about #**ephemerable** on Twitter. See what people are saying and join the conversation.

Ephemerable Web Analysis - Ephemerable.com
ephemerable.com.cutestat.com/
ephemerable.com is 1 year 7 months old. It is a domain having .com extension. This website is estimated worth of $ 8.95 and have a daily income of around ...

ephemerable.com
ephemerable.com.w3snoop.com/
View **ephemerable**.com - Free traffic, earnings, ip, location, rankings report about **ephemerable**.com.

ephemerable.com ∞ ephemerable
ephemerable.com.citedot.com/

Both are messy!
We need to do some pre-processing.

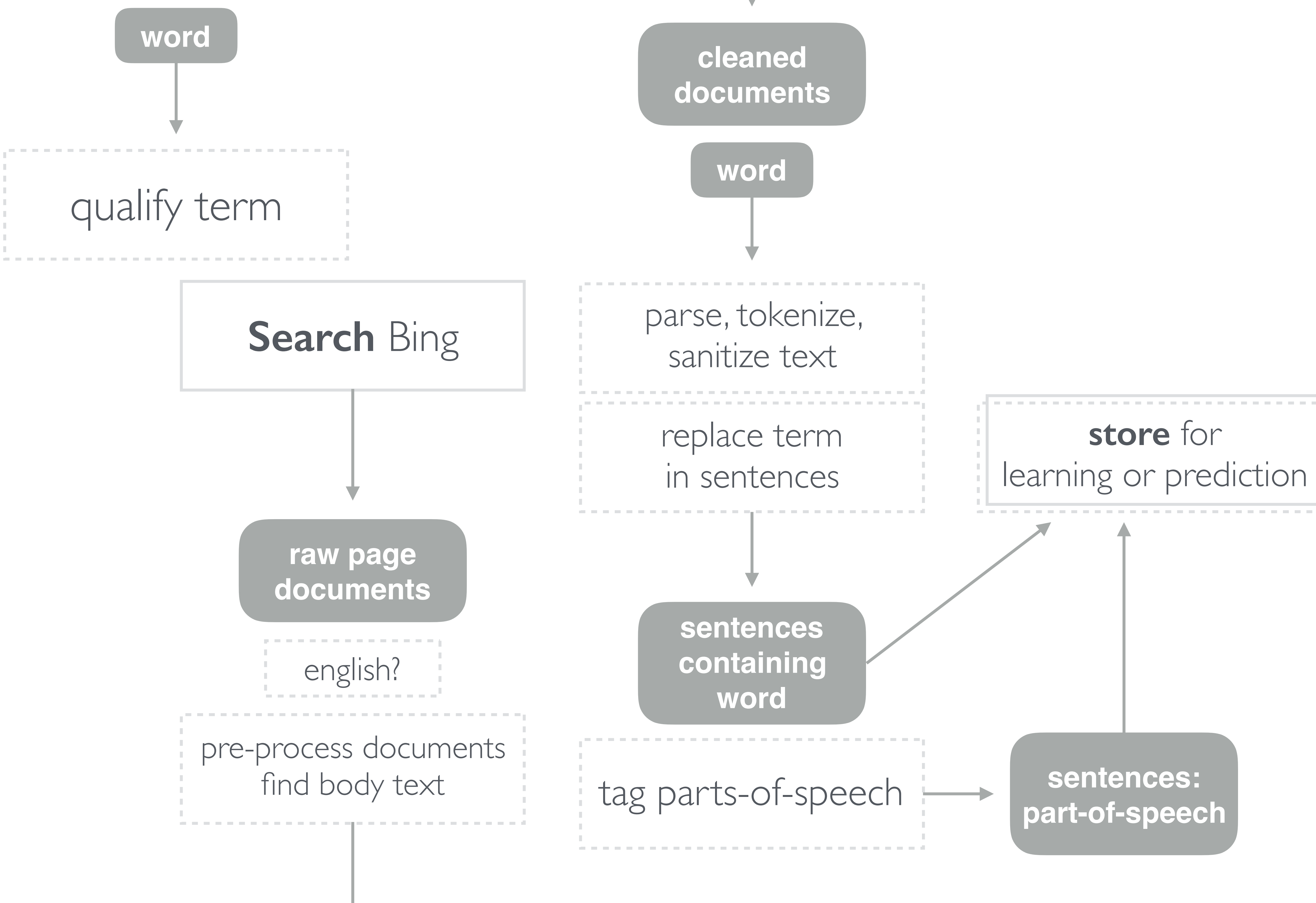
Pre-processing is often custom, determined by:

- the domain of the text
- your goals
- dealing with edge cases

What does the solution look like so far?

Processing Natural Language for Wordnik

@clarecorthell



All with the intent of creating features!

What are **Features**?

individual measurable properties of the thing we’re observing.

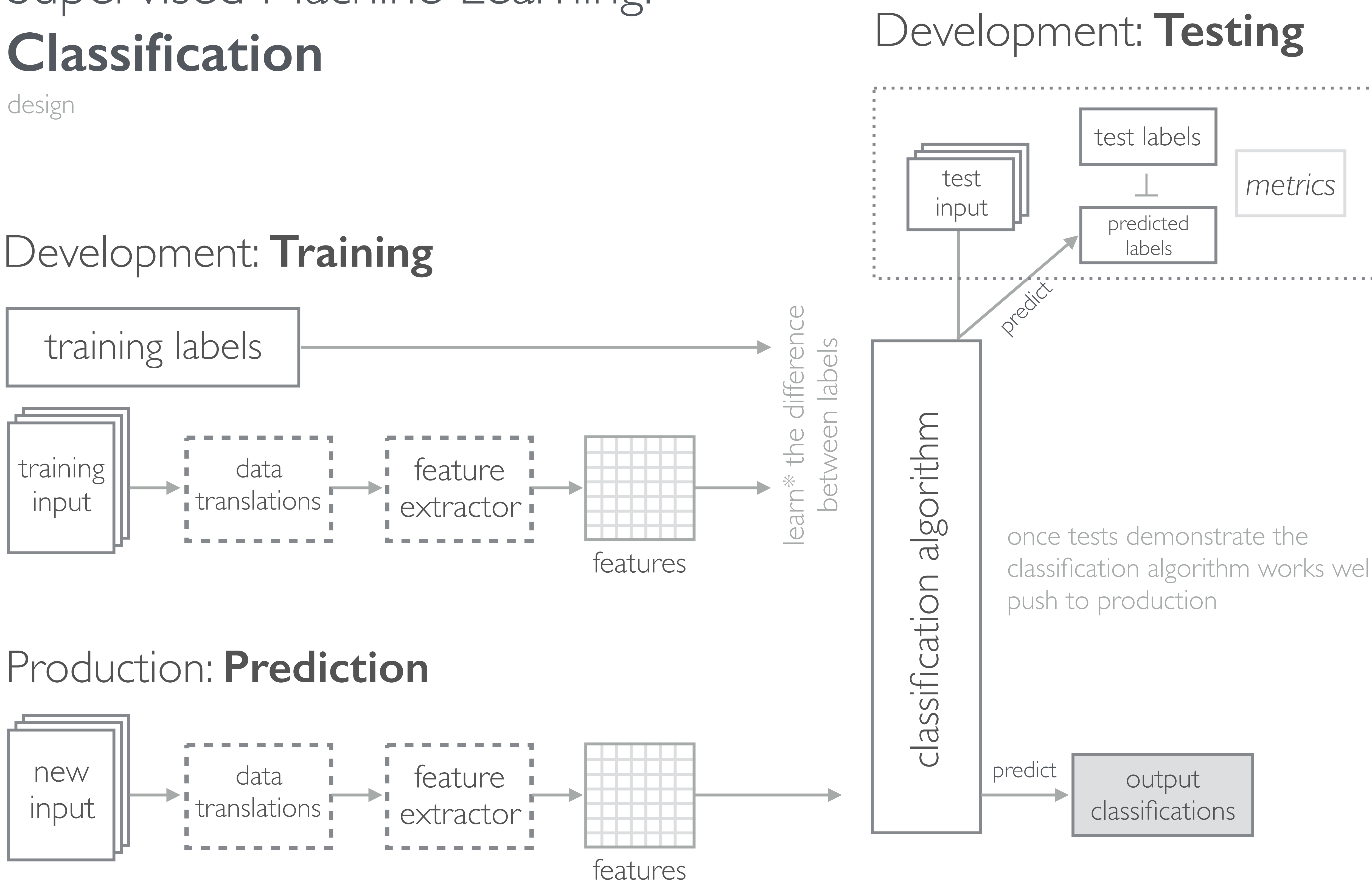
The point of feature development and extraction

is to build derived values from the data
that represent characteristics that identify the data

or differentiate data points from one another
in such a way that the computer can observe that difference

Supervised Machine Learning: **Classification**

design



* ***Learning*** means finding the functions that define the difference between training labels, based on training input.

@clarecorthell

How do we construct features
that will differentiate our examples?

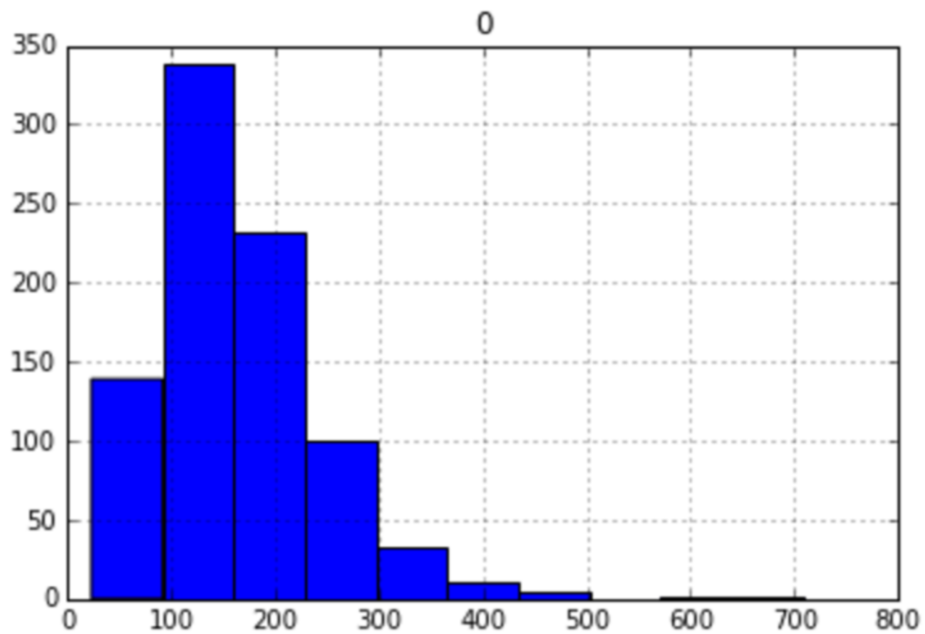
We have to get creative, make guesses, and statistically test them to see what features will work

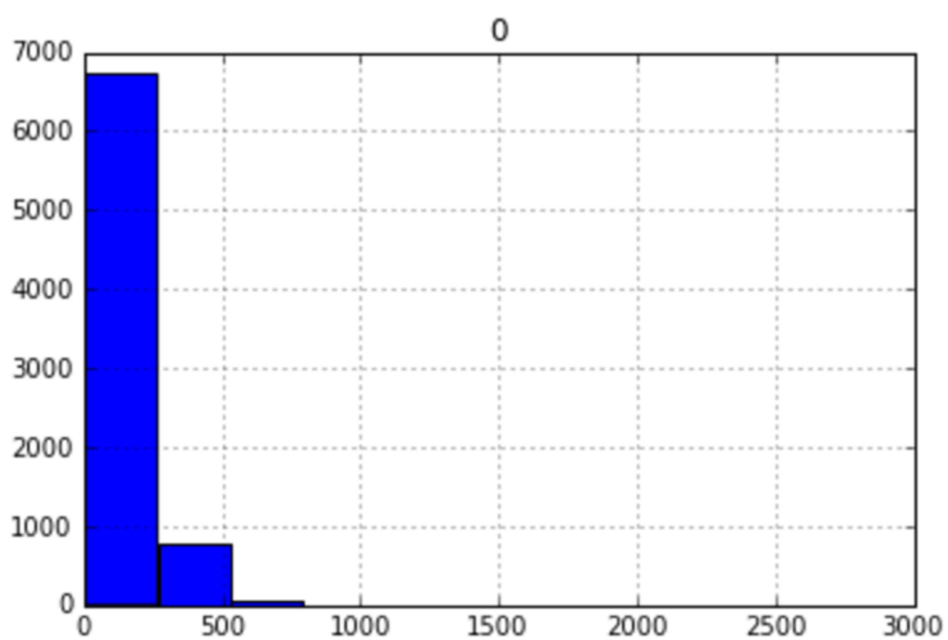
What features do sentences with FRDs have?

The term “cheeseors” describes flighted globules of intergalactic cheese, known to be the scourge of the asteroid belt.

Possible Feature: Length of Sentence

```
In [7]: # Length of Sentence
        pd.DataFrame([len(p['sentence']) for p in positive]).hist()
        pd.DataFrame([len(n['sentence']) for n in negative]).hist()

Out[7]: array([[<matplotlib.axes._subplots.AxesSubplot object at 0x110a17790>]], dtype=object)
```





@clarecorthell

Chi-Squared (chi^2) Selection

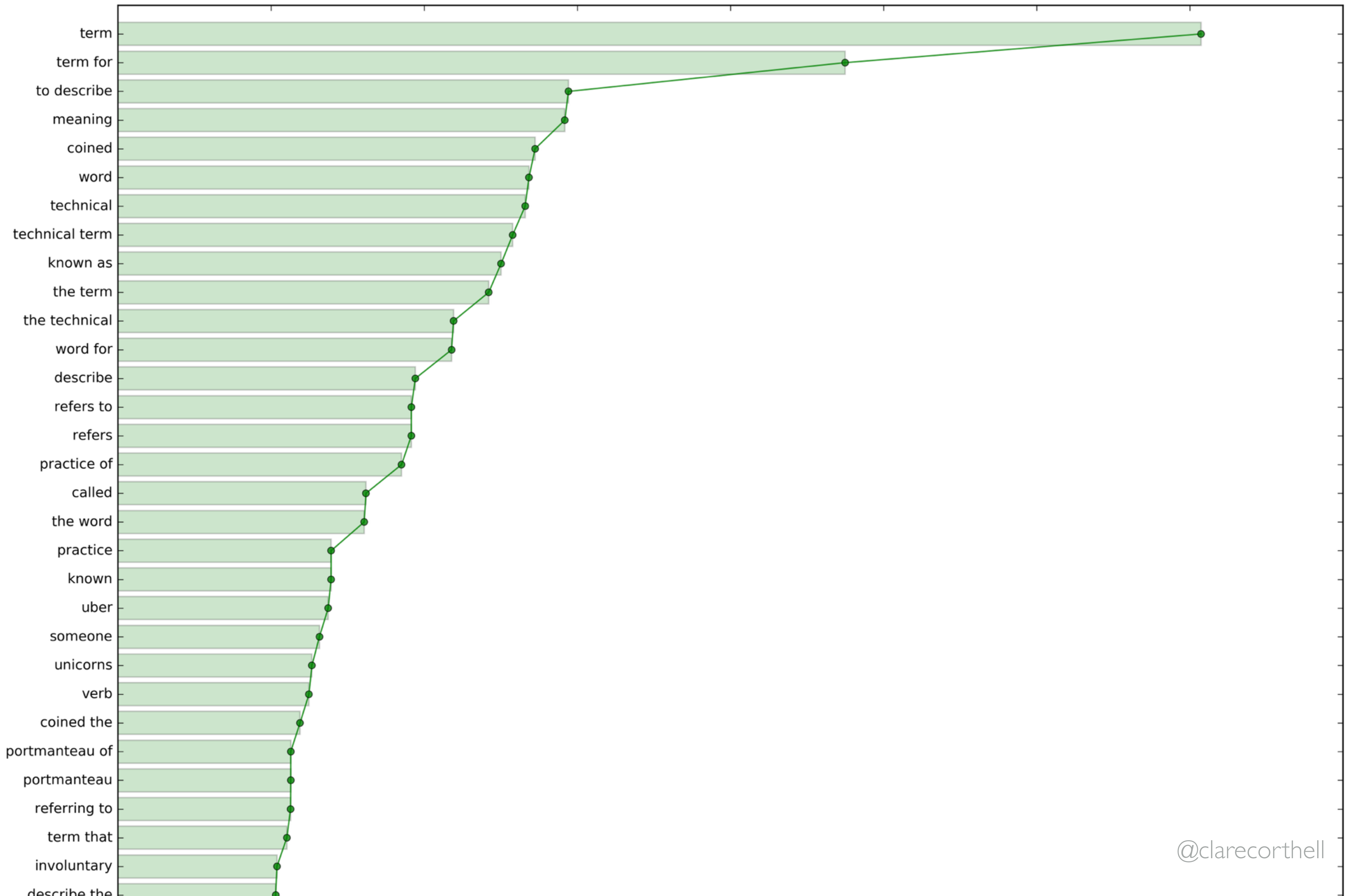
Chi Squared is a statistical test that describes how **independent** two given events are.

In selecting features, the two events are **occurrence of the term** and **occurrence of the class**. If the score is high or significant, it means that the occurrence is dependent on the class.

Possible Features: Length, Location, Position in Sentence

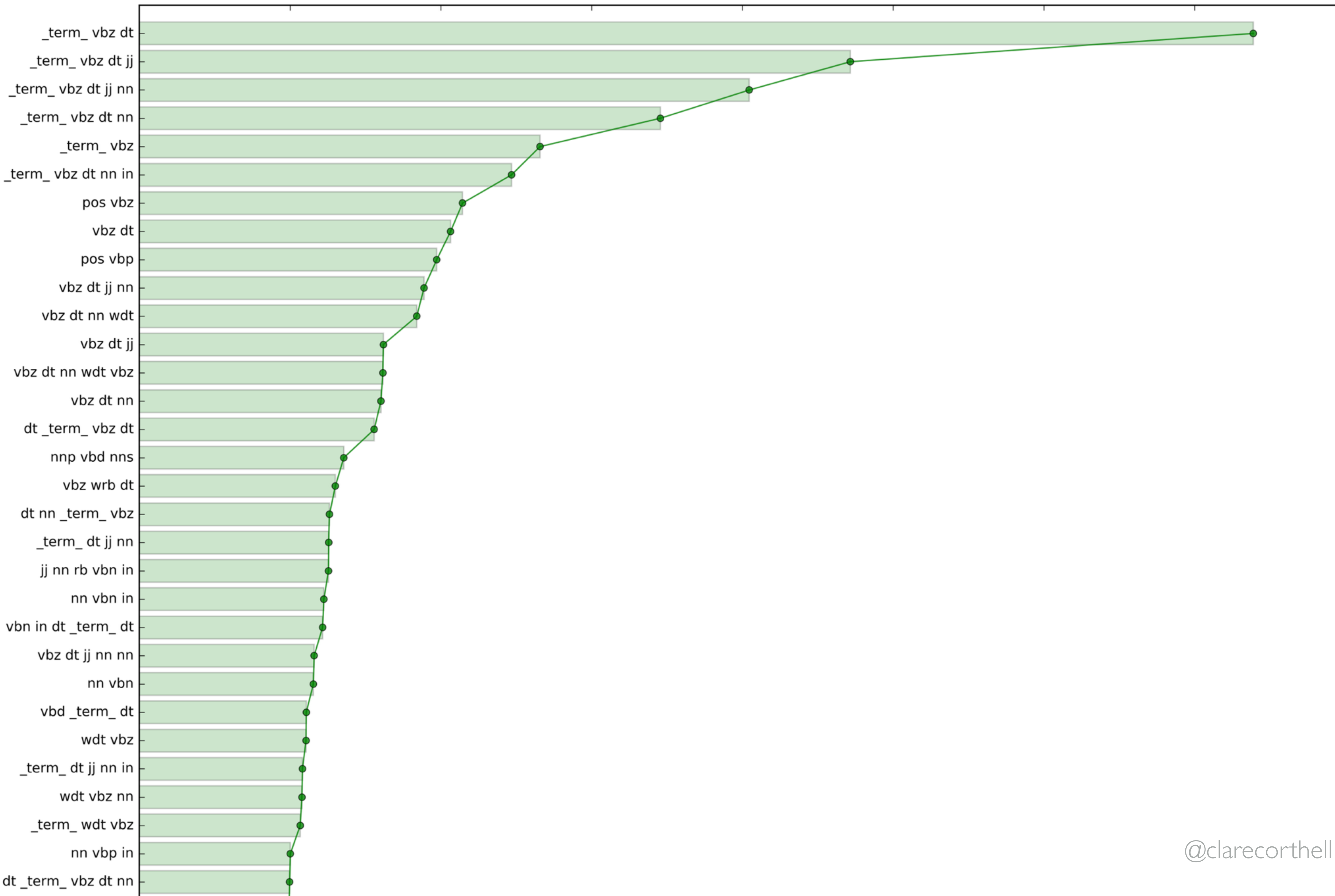
```
In [28]: ch2 = SelectKBest(chi2, k=2)
         X_train = ch2.fit_transform(x_train, y_train)
         X_test = ch2.transform(x_test)
         print zip(ch2.pvalues_, dv.get_feature_names())

[(0.0, 'len'), (0.0, 'location'), (3.7239735847433497e-05, 'position')]
```

highest chi^2 feature scoring for tokens



@clarecorthell

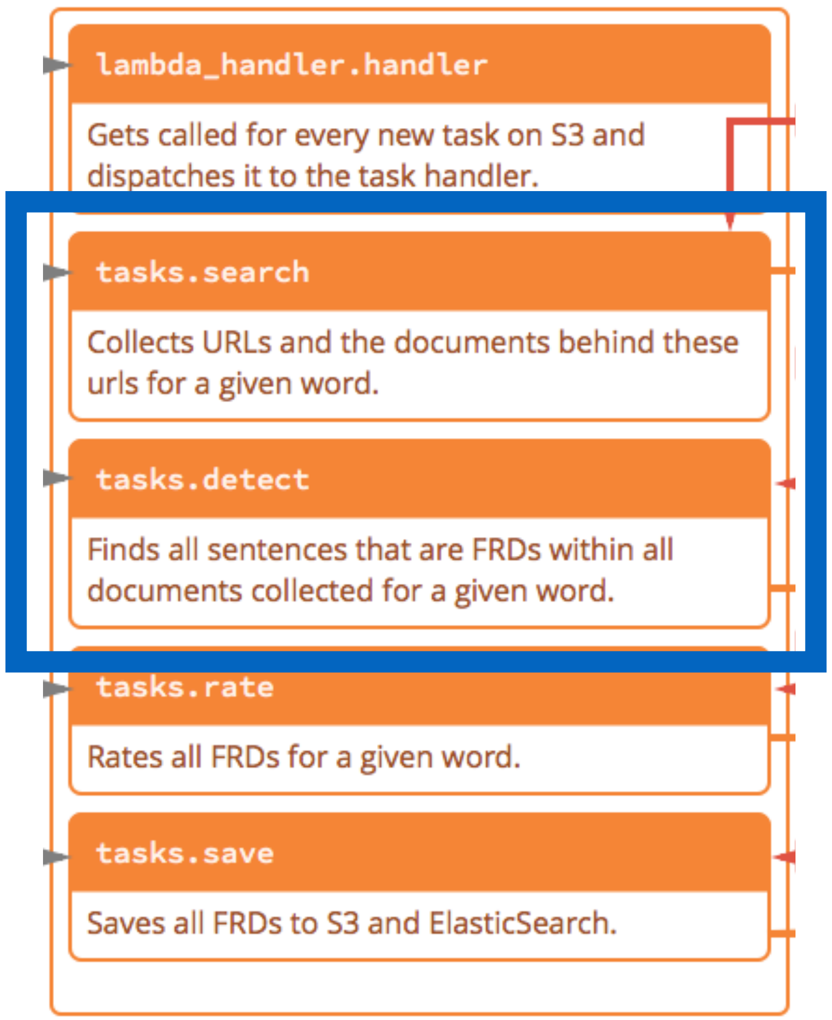
highest chi^2 feature scoring for parts of speech



Final features for FRDs:

token context around the term (tf-idf)
POS tag context around the term

Final NLP Modules

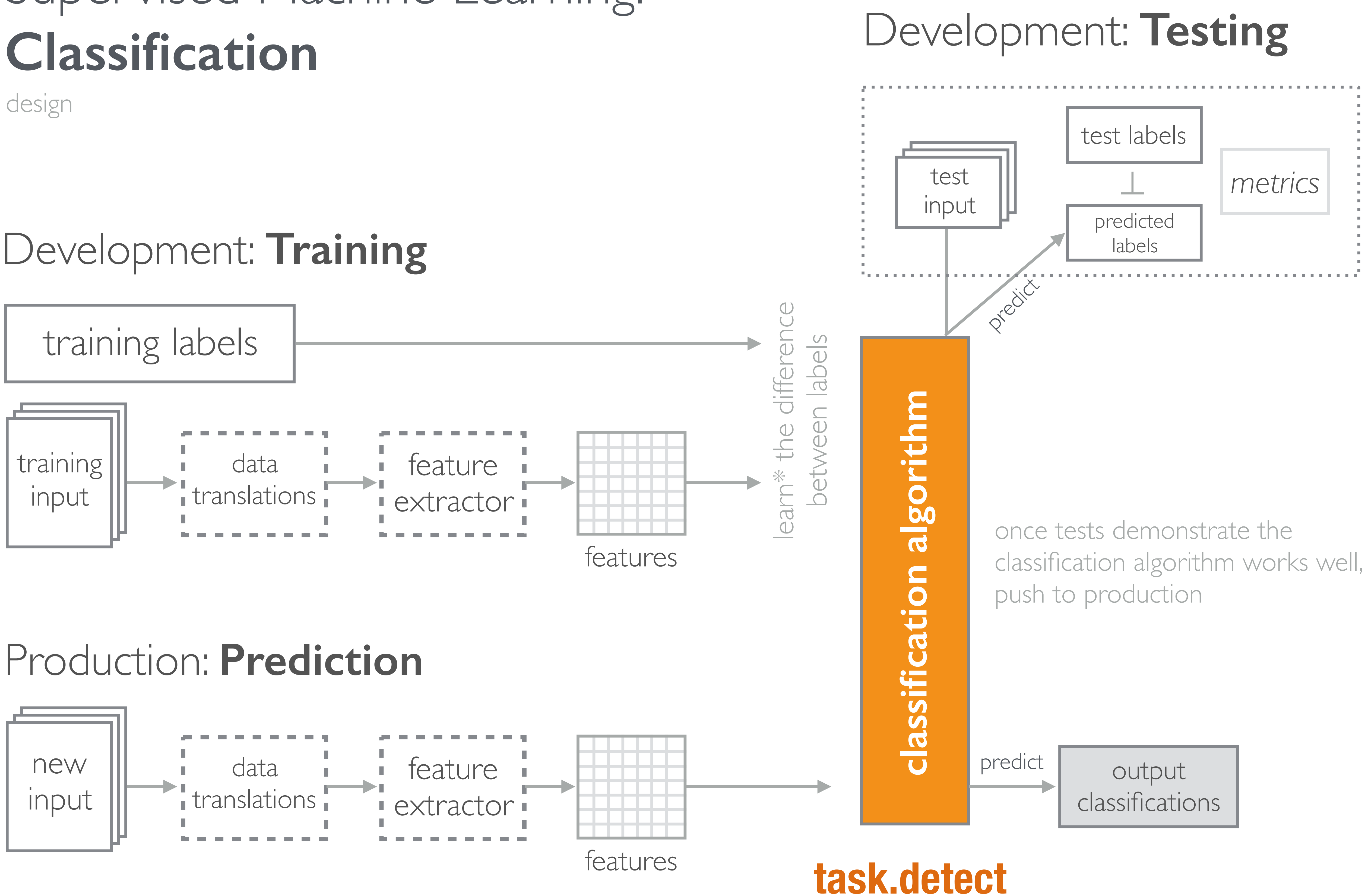


task.search collects URLs and the documents behind those URLs for a given word

task.detect finds all sentences that are FRDs within all documents collected for a given word

Supervised Machine Learning: **Classification**

design



* *Learning* means finding the functions that define the difference between training labels, based on training input.

@clarecorthell

Keys to Success in Supervised Learning:

- **knowing what an FRD is**
- **knowing how to encode the FRD for computation**
- **using the right data sources**
- **choosing the right features**

Ready to put it all together?

wordnik

SERAPIS
SCALABLE
WORD GOBBLER

NLP in Production

Lessons and Patterns from the distributed Wordnik system
codename: **Serapis**

SERAPIS
SCALABLE
WORD GOBBLER

Serapis is the Graeco-Egyptian god of knowledge, education —
— and adding a million words into the dictionary.

Q: Can’t we just run the same thing on a server?

we want to find definitions for **1 million words**

Q: Can't we just run the same thing on a server?

we want to find definitions for **1 million words**

which means we make **1 million search requests**

Q: Can't we just run the same thing on a server?

we want to find definitions for **1 million words**

which means we make **1 million search requests**

returning 40 results each for **40 million search results**

Q: Can't we just run the same thing on a server?

we want to find definitions for **1 million words**

which means we make **1 million search requests**

returning 40 results each for **40 million search results**

if we follow those links to get the pages, that's **40 million page requests**

Q: Can't we just run the same thing on a server?

we want to find definitions for **1 million words**

which means we make **1 million search requests**

returning 40 results each for **40 million search results**

if we follow those links to get the pages, that's **40 million page requests**

each page request takes on average 2 sec, so **40,000,000 pages * 2 sec**

Q: Can't we just run the **same thing on a server**?

we want to find definitions for **1 million words**

which means we make **1 million search requests**

returning 40 results each for **40 million search results**

if we follow those links to get the pages, that's **40 million page requests**

each page request takes on average 2 sec, so **40,000,000 pages * 2 sec**

In series, it would take **5 Years** to get all the documents we want

Q: Can’t we just run the same thing on a server?

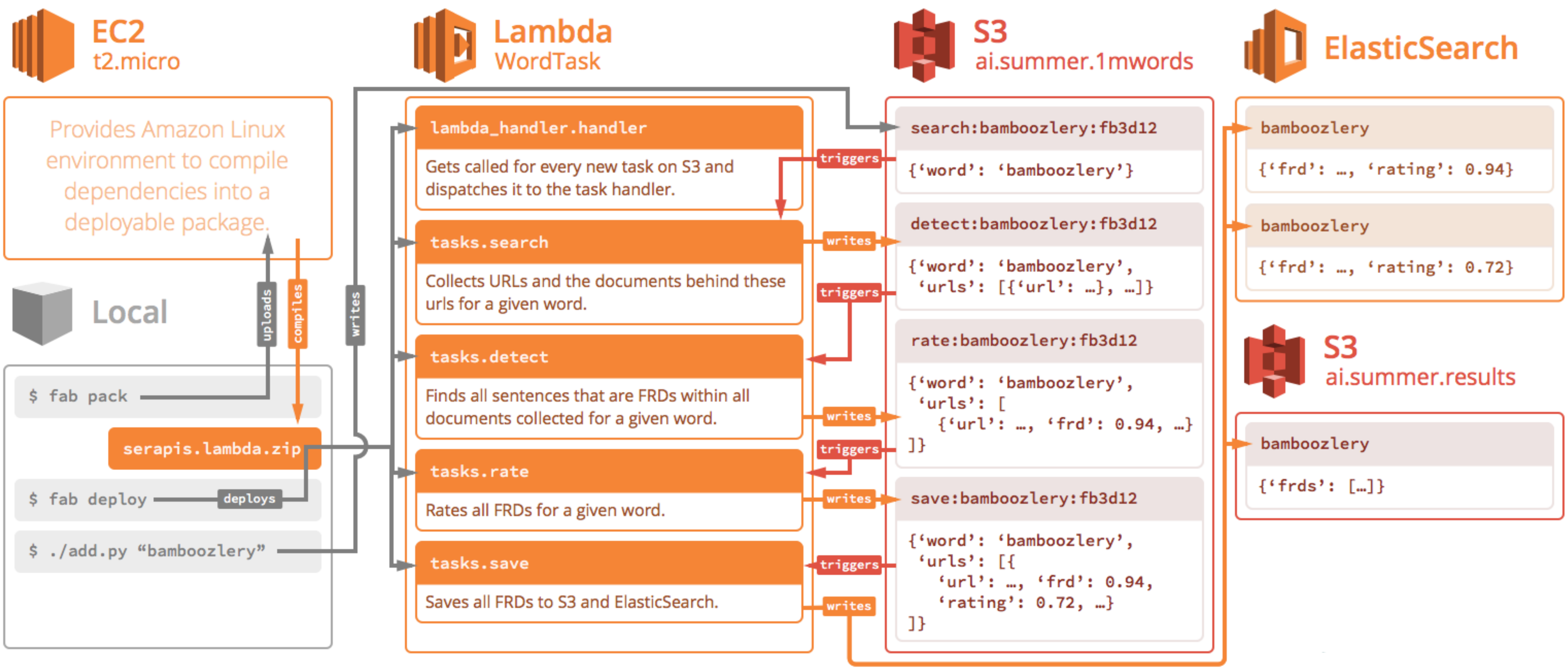
Scale and Page Requests

If we want to get search results for 1 million words, it would take us 5 years to get all the documents if we processed everything sequentially. We need to parallelize.

A: Nope.

SERAPIS
INFRASTRUCTURE
SCHEMATICS

Distributed Infrastructure for data collection, natural language processing, and machine learning



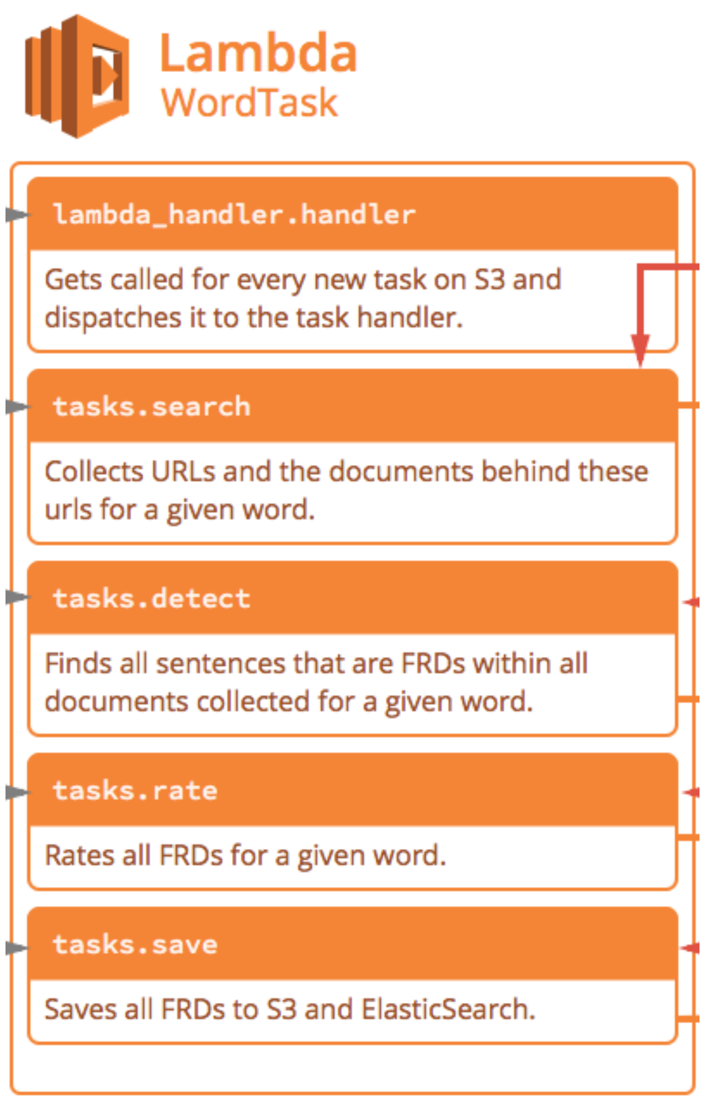
resizable compute

compute management service

scalable storage

index

@clarecorthell



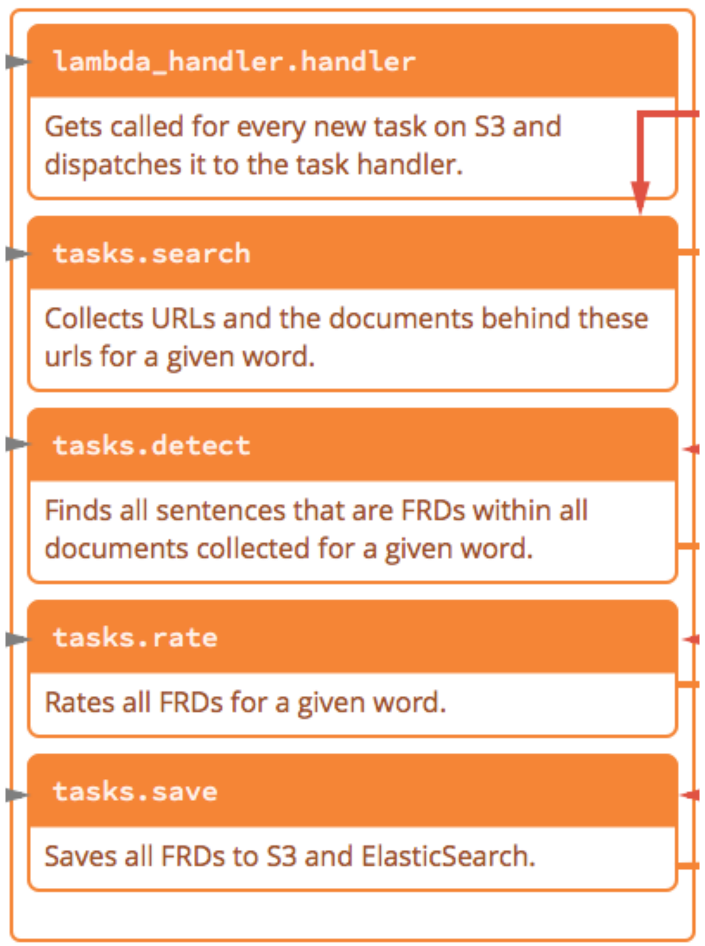
S3
ai.summer.1mwords

ElasticSearch

AWS Lambda is a compute service that runs your code in response to events and automatically manages the underlying compute resources for you.

The promise of Lambda is that you don't have to worry about infrastructure, rather you set a task and a trigger, such as a change to a document in an S3 bucket. Lambda takes over scaling out the resources to complete all the tasks.

Lambda
WordTask



S3
ai.summer.1mwords

ElasticSearch

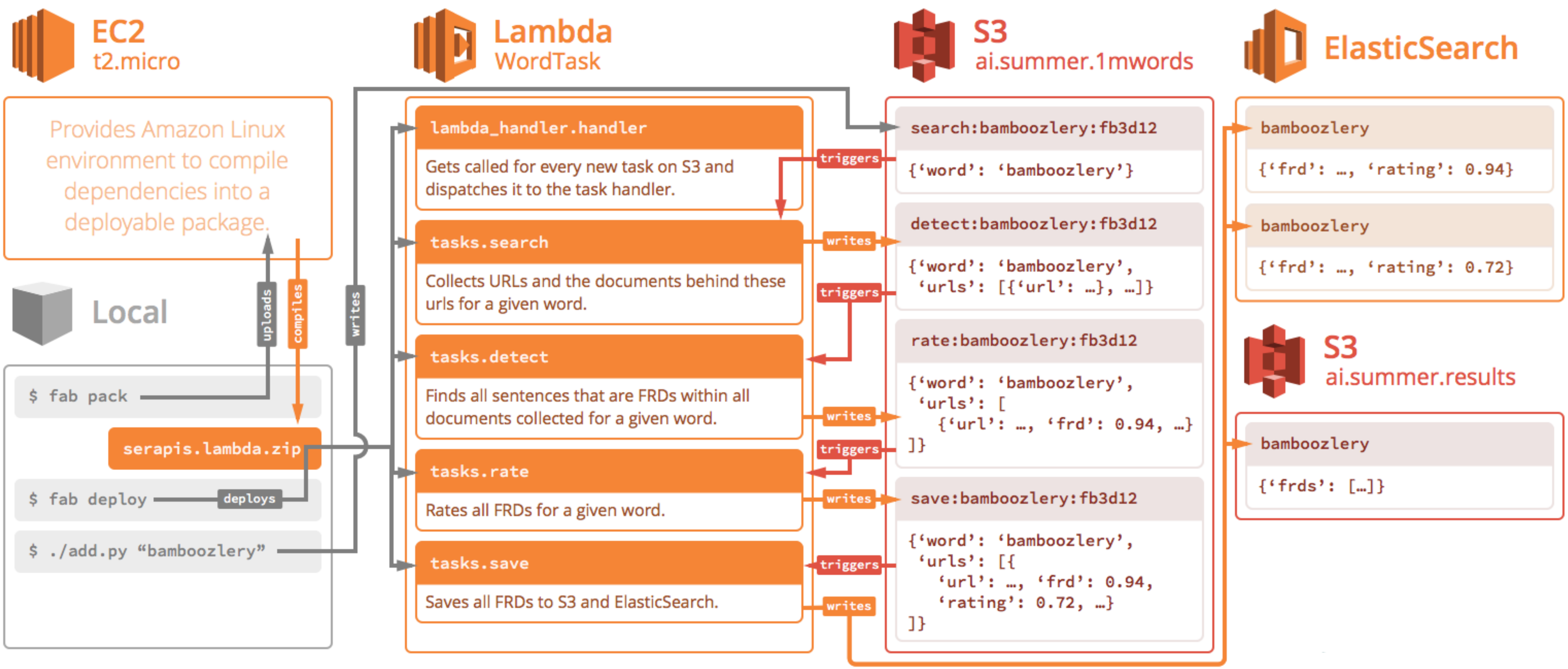
1. **Start an EC2 instance** with the Amazon Linux AMI
2. **Build shared libraries** from source on EC2
3. **Create a virtualenv** with python dependencies
4. **Write a python handler function** to respond to a given event and do its work (process text, etc)
5. **Bundle** the virtualenv, your code and the binary libs into a zip file
6. **Publish the zip file** to AWS Lambda

Voila.

@clarecorthell

SERAPIS INFRASTRUCTURE SCHEMATICS

Distributed Infrastructure for data collection, natural language processing, and machine learning



A Few New Words from FRDs

It has developed a mechanism to 'dye' very small bitcoin transactions (called **'bitcoin dust'**) by adding extra data to them so that they can represent bonds, shares or units of precious metals.

'oxt weekend,' in other words, means 'not this coming weekend but the one after.'

This summer, a new, trendier one, emerged: **NATU**, for Netflix, Airbnb, Tesla and Uber.

Thank You!

Clare Corthell
clare@luminantdata.com
@clarecorthell