



ESKWE LABS

Eskwelabs x Shopee Code League

Bash Yumol

Introduction

Who am I?

Name: Albert 'Bash' Yumol

Lives: Manila Philippines

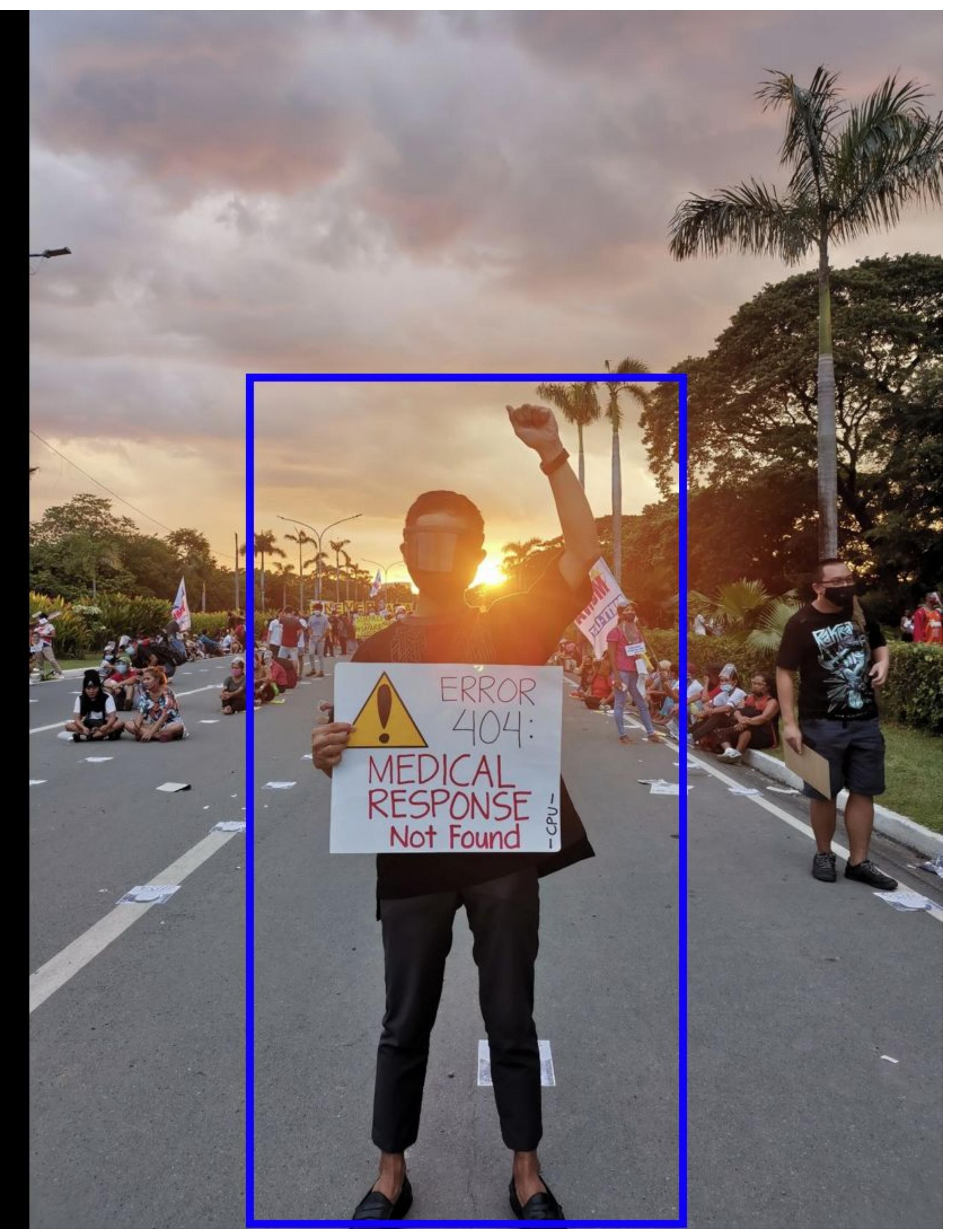
Interests: Physics, AI, Big Data,
Cryptography, IoT, Activism

Occupation: Data Scientist and AI Consultant, EdTech

Connect: <https://www.linkedin.com/in/albertyumol/>

<https://github.com/albertyumol>

<https://albertyumol.github.io/>



Mission



Eskwelabs is an online data upskilling school based in the Philippines driving social mobility in the future of work.



We build data skills for workers and teams through mentor-led project-based upskilling.

For Individuals

90% job placement within 90 days
50% increase in income



For Companies

Build or buy talent
Pivot to mid to high-value work



Source: Job outcome survey reported by students from Cohort I-III, Eskwelabs Data Science program.



eskwelabs.com



Eskwelabs



eskwelabs

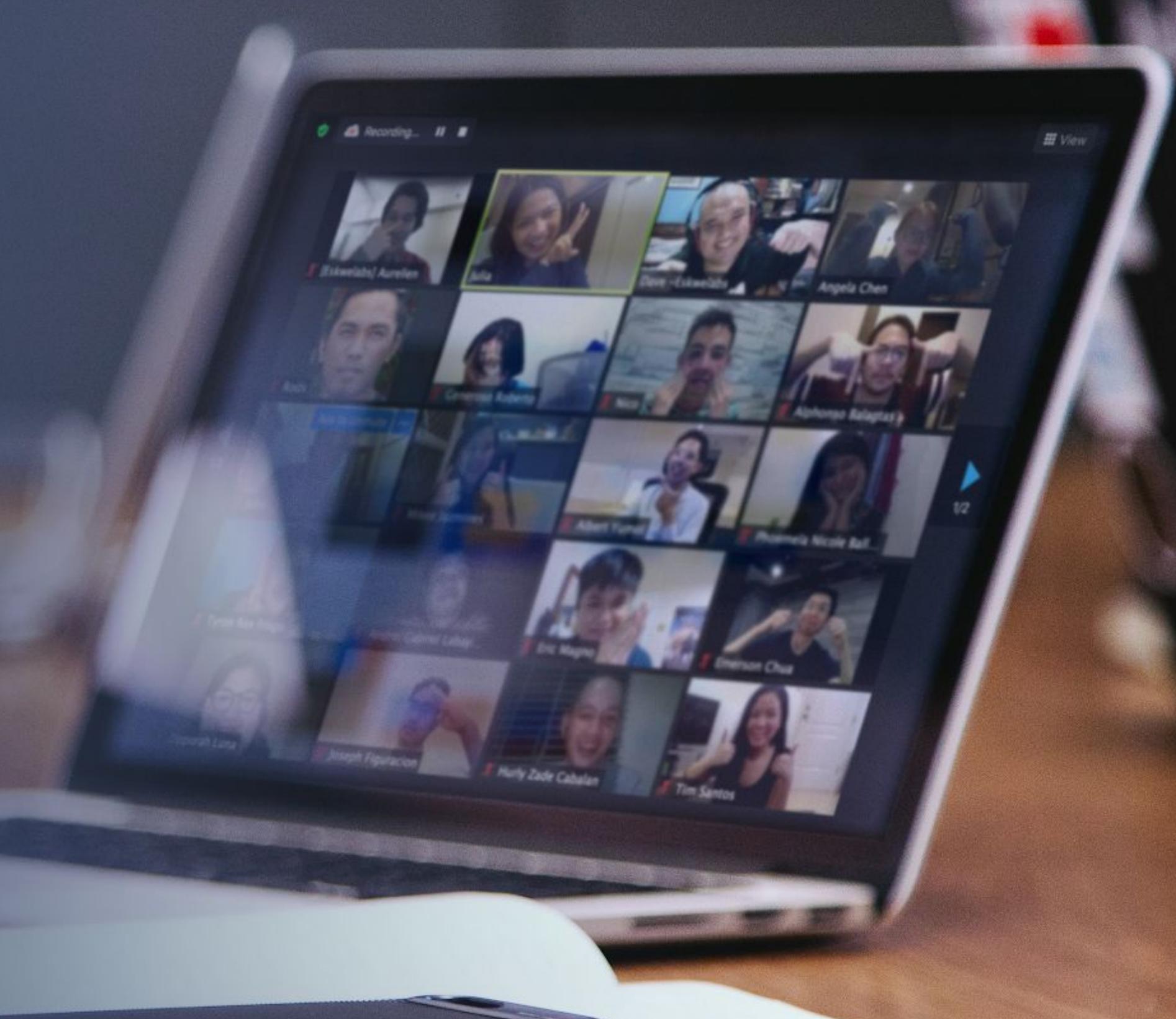


@eskwelabs_ph

DATA CLUB

A virtual upskilling experience as a hands-on laboratory where you are guided by industry mentors to build data projects with friends and add outputs to your portfolio. Lifelong learners at different levels of data proficiency are welcome!

JOIN THE WAITLIST



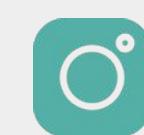
eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

SPRINT TOPICS

www.eskwelabs.com/data-club

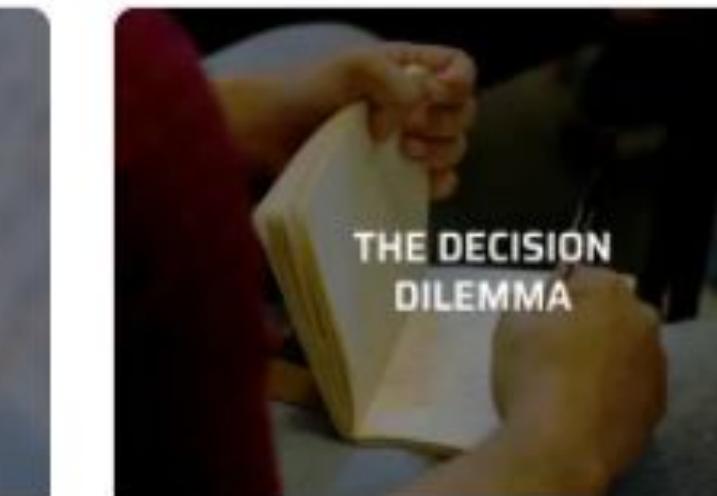


Interactive Data Visualization with PowerBI

Learn how to turn visualization into insights with one of the most powerful tools for data analysis - PowerBI - while building a beautiful, and interactive dashboard to track the latest pandemic developments.

[Beginner](#) [No Code](#) [PowerBI](#)

[READ MORE →](#)



THE DECISION DILEMMA

What-If Analysis and Optimization with Solver in Excel

Make better everyday and business decisions using Excel Solver that optimize allocation of resources.

[Beginner](#) [No Code](#)

[READ MORE →](#)



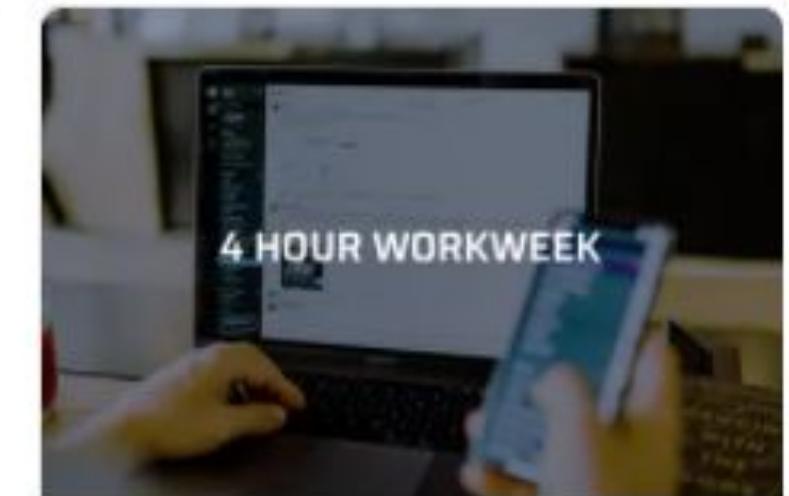
REACHING THE SDGs

Create Exploratory Data Analysis in Python

Tell a story through data on how far the world has progressed on the UN's Global Goals, a universal call to action to end poverty, protect the planet, and ensure that all people enjoy peace and prosperity by 2030.

[Data for Good](#) [Beginner](#) [Python](#)

[READ MORE →](#)



4 HOUR WORKWEEK

Save Time by Automating Work in Excel

Ever wonder how some people manage to get their work done faster? Their secret is working smart by using Excel VBA to automate repeatable tasks.

[Beginner](#) [No Code](#)

[READ MORE →](#)

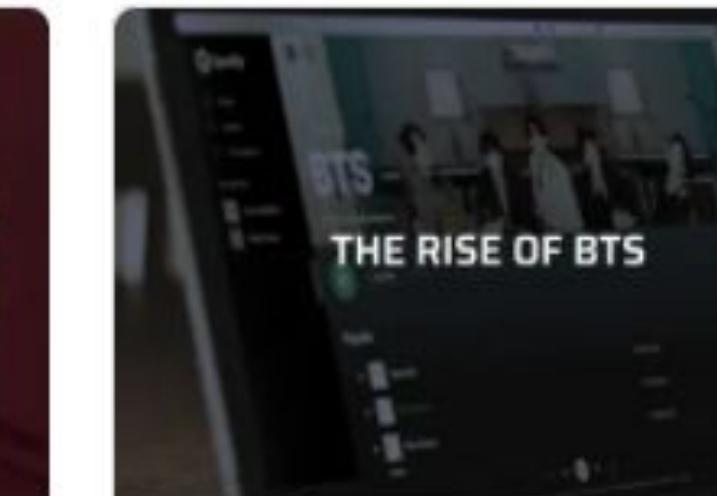


Introduction to Object Recognition

Learn the foundations of computer vision and implement your own object detection algorithm and identify an object of your choice.

[Intermediate](#) [Python](#)

[READ MORE →](#)



The Rise of BTS

Create a bar chart race using Python to visualize how music artist popularity changed over time.

[Beginner](#) [Python](#) [Data Viz](#)

[READ MORE →](#)



THE DIGITAL KRUSTY KRAB

Design Data Strategy for a Fast Food Restaurant

Help craft the data strategy for your favourite fast food chain by understanding how data can serve business goals.

[Beginner](#) [No Code](#)

[READ MORE →](#)



DATA MEETS DON DRAPER

Data meets Don Draper - Customer Segmentation Analysis

The digital economy means customers are online. Help a creative ad agency target the right audiences with digital marketing.

[Beginner](#) [No Code](#) [SQL](#)

[READ MORE →](#)



eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

Start



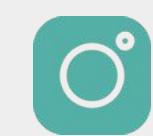
eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

Objectives



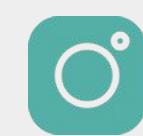
eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

Natural Language Processing

a branch of artificial intelligence that helps machine understand and respond to human language.



eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

Communication With Machines



~50-70s

```
File Edit Edit_Settings Menu Utilities Compilers Test Help
EDIT      BS90.DEVT3.CLIBPAU(TIMMIES) - 01.31          Columns 00001 00
Command ===> █
***** **** Top of Data ****
000001 /* REXX EXEC ****
000002 /*
000003 /* TIMMIES FACTOR - COMPOUND INTEREST CALCULATOR
000004 /*
000005 /* AUTHOR: PAUL GAMBLE
000006 /* DATE: OCT 1/2007
000007 /*
000008 /*
000009 /*
000010 /*
000011 /*
000012 say '*****'
000013 say 'Welcome Coffee drinker.'
000014 say '*****'
000015 DO WHILE DATATYPE(CoffeeAmt) \= 'NUM'
000016   say ""
000017   say "What is the price of your coffee?",
000018     "(e.g. 1.58 = $1.58)"
000019   parse pull CoffeeAmt
000020 END
000021 /*
000022 DO WHILE DATATYPE(CoffeeWk) \= 'NUM'
000023   say ""
000024   say "How many coffees a week do you have?"
000025   parse pull CoffeeWk
000026 END
000027 /*
000028 DO WHILE DATATYPE(Rate) \= 'NUM'
000029   say ""
000030   say "What annual interest rate would you like to see on that money?",
000031     "(e.g. 8 = 8%)"
000032   parse pull Rate
000033 END
000034 Rate = Rate * 0.01 /* CHG TO DECIMAL NUMBER */
```

~80s



today



eskwelabs.com



Eskwelabs



eskwelabs

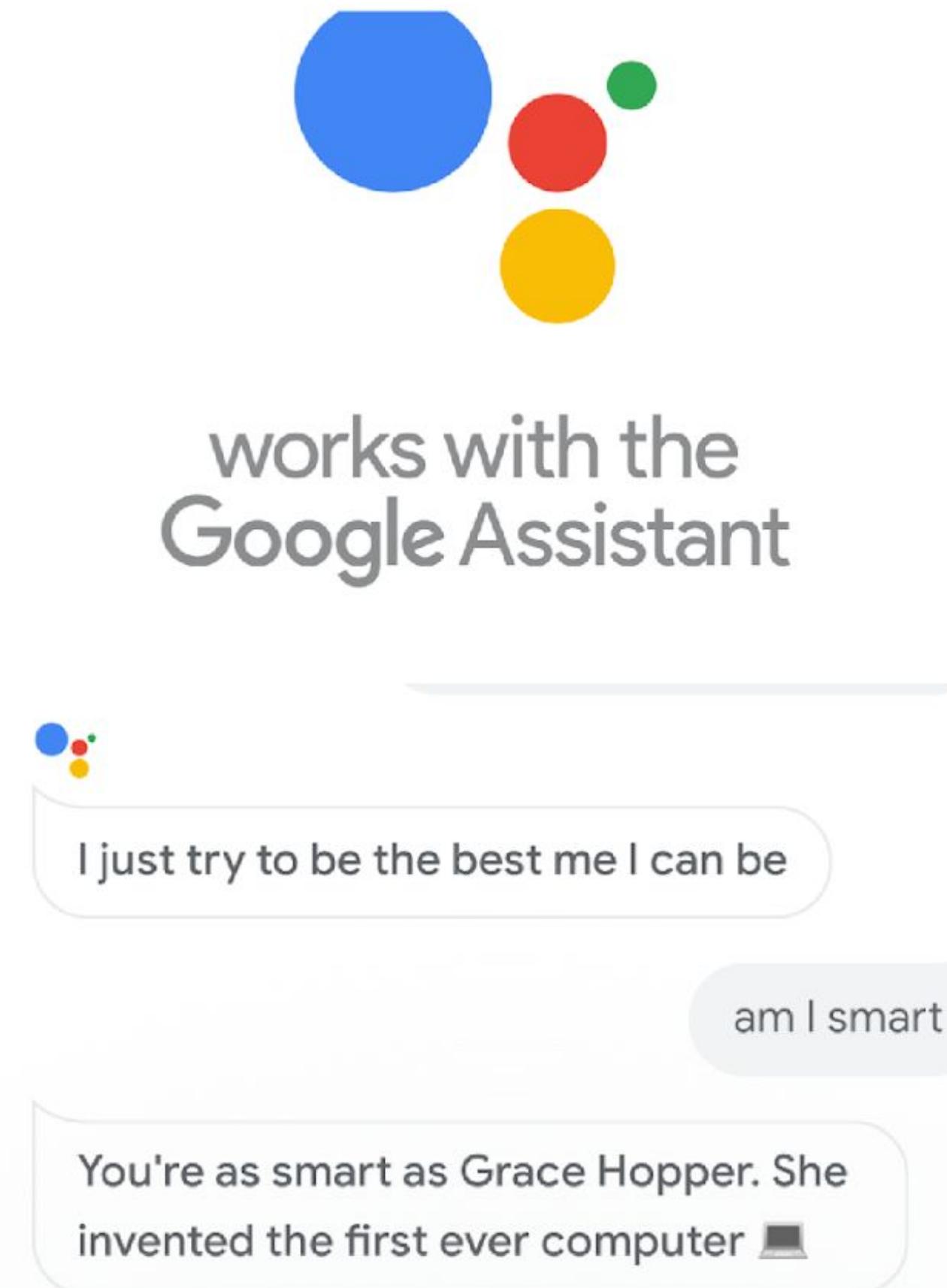


@eskwelabs_ph

Conversational Agents

Conversational agents contain:

- Speech recognition
- Language analysis
- Dialogue processing
- Information retrieval
- Text to speech



eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph



In early 2011, an IBM computing system named Watson competed against the world's best Jeopardy! champions.

6



eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

Question Answering



- What does “divergent” mean?
- What year was Abraham Lincoln born?
- How many states were in the United States that year?
- How much Chinese silk was exported to England in the end of the 18th century?
- What do scientists think about the ethics of human cloning?





Machine Translation



The screenshot shows the Google Translate interface. On the left, the main window displays the text "我学习深度学习和机器学习" (I study deep learning and machine learning) in Chinese, with its Pinyin transcription "Wǒ xuéxí shēndù xuéxí hé jīqì xuéxí" below it. The detected language is "CHINESE - DETECTED". The target language is set to "ENGLISH". On the right, a separate window titled "Google Translate" shows the "DETECT LANGUAGE" tab selected. It lists various languages in a grid format, with "Detect language" checked. The languages listed include Afrikaans, Albanian, Amharic, Arabic, Armenian, Azerbaijani, Basque, Belarusian, Bengali, Bosnian, Bulgarian, Catalan, Cebuano, Chichewa, Chinese, Corsican, Croatian, Czech, Danish, Dutch, English, Esperanto, Estonian, Filipino, Finnish, French, Frisian, Galician, Georgian, German, Greek, Gujarati, Haitian Creole, Hausa, Hawaiian, Hebrew, Hindi, Hmong, Hungarian, Icelandic, Igbo, Indonesian, Irish, Italian, Japanese, Javanese, Kannada, Kazakh, Khmer, Korean, Kurdish (Kurmanji), Kyrgyz, Lao, Latin, Latvian, Lithuanian, Luxembourgish, Macedonian, Malagasy, Malay, Maltese, Malayalam, Maltese, Sesotho, Shona, Vietnamese, Sinhala, Xhosa, Marathi, Mongolian, Norwegian, Pashto, Nepali, Persian, Polish, Portuguese, Punjabi, Romanian, Russian, Samoan, Scots Gaelic, Serbian, Urdu, Uzbek, Welsh, Yiddish, Slovenian, Yoruba, Zulu, Swedish, and Tajik.

DETECT LANGUAGE	ENGLISH	SPANISH	FRENCH	ENGLISH	SPANISH	ARABIC
<input checked="" type="checkbox"/> Detect language	Czech	Hebrew	Latin	Portuguese	Tajik	
Afrikaans	Danish	Hindi	Latvian	Punjabi	Tamil	
Albanian	Dutch	Hmong	Lithuanian	Romanian	Telugu	
Amharic	English	Hungarian	Luxembourgish	Russian	Thai	
Arabic	Esperanto	Icelandic	Macedonian	Samoan	Turkish	
Armenian	Estonian	Igbo	Malagasy	Scots Gaelic	Ukrainian	
Azerbaijani	Filipino	Indonesian	Malay	Serbian	Urdu	
Basque	Finnish	Irish	Malayalam	Sesotho	Uzbek	
Belarusian	French	Italian	Maltese	Shona	Vietnamese	
Bengali	Frisian	Japanese	Maori	Sindhi	Welsh	
Bosnian	Galician	Javanese	Marathi	Sinhala	Xhosa	
Bulgarian	Georgian	Kannada	Mongolian	Slovak	Yiddish	
Catalan	German	Kazakh	Myanmar (Burmese)	Slovenian	Yoruba	
Cebuano	Greek	Khmer	Nepali	Somali	Zulu	
Chichewa	Gujarati	Korean	Norwegian	Spanish		
Chinese	Haitian Creole	Kurdish (Kurmanji)	Pashto	Sundanese		
Corsican	Hausa	Kyrgyz	Persian	Swahili		
Croatian	Hawaiian	Lao	Polish	Swedish		



Natural Language Processing

Applications

- Machine Translation
- Information Retrieval
- Question Answering
- Dialogue Systems
- Information Extraction
- Summarization
- Sentiment Analysis
- ...

Core Technologies

- Language modeling
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition
- Word sense disambiguation
- Semantic role labeling
- ...

NLP lies at the intersection of computational linguistics and machine learning.



eskwelabs.com



Eskwelabs

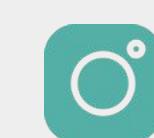
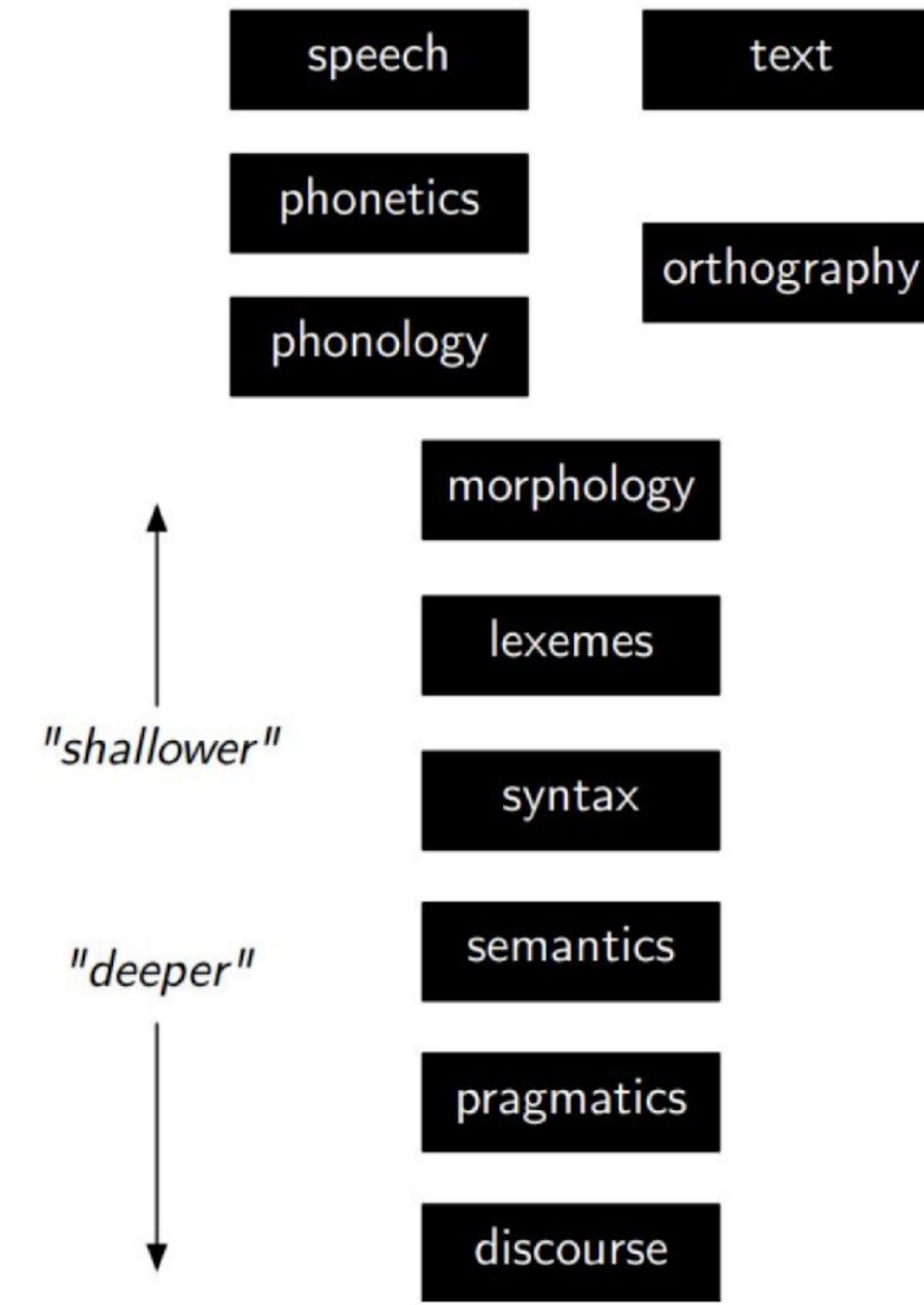


eskwelabs



@eskwelabs_ph

Level Of Linguistic Knowledge



Phonetics, Phonology

■ Pronunciation Modeling

SOUNDS

Th i a si e n



eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

Words

- Language Modeling
- Tokenization
- Spelling correction

WORDS

This is a simple sentence



eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

Morphology

- Morphology analysis
- Tokenization
- Lemmatization

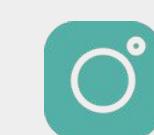
WORDS	This	is	a	simple	sentence
MORPHOLOGY				be 3sg present	



Part of Speech

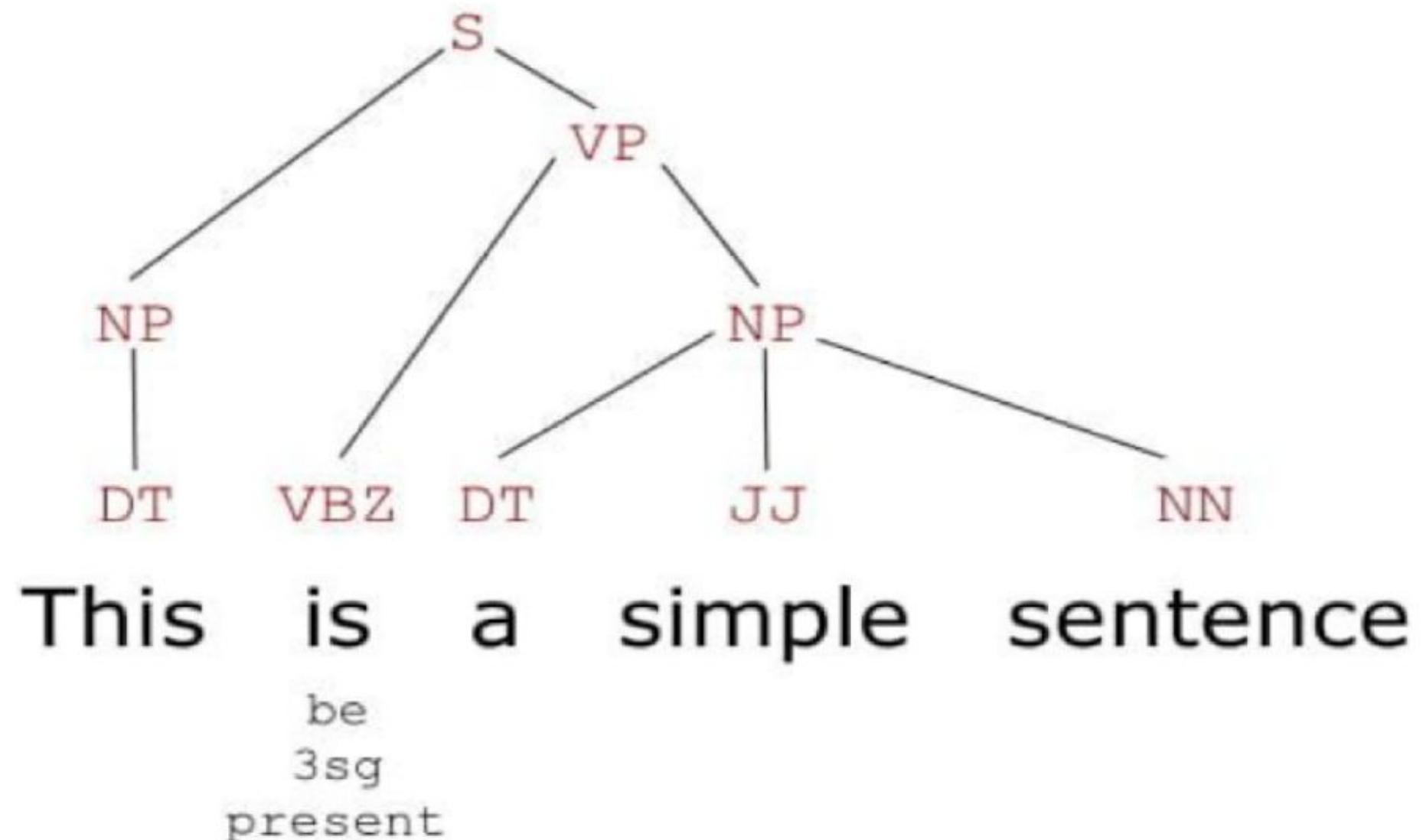
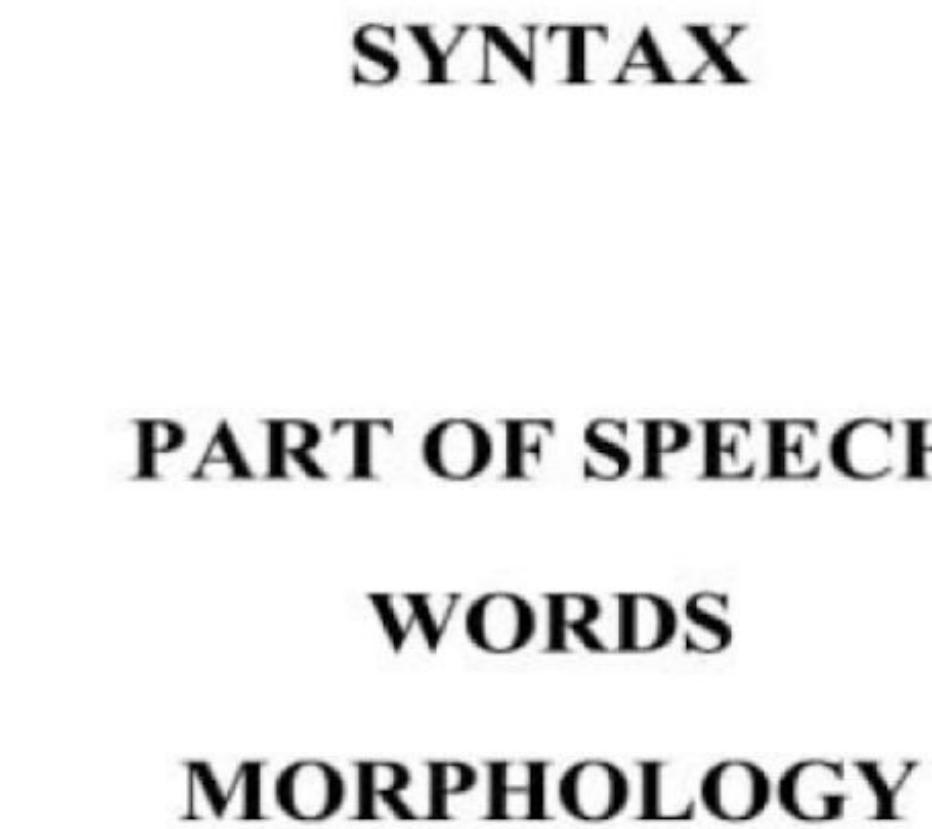
■ Part of speech tagging

PART OF SPEECH	DT	VBZ	DT	JJ	NN
WORDS	This	is	a	simple	sentence
MORPHOLOGY		be 3sg present			



Syntax

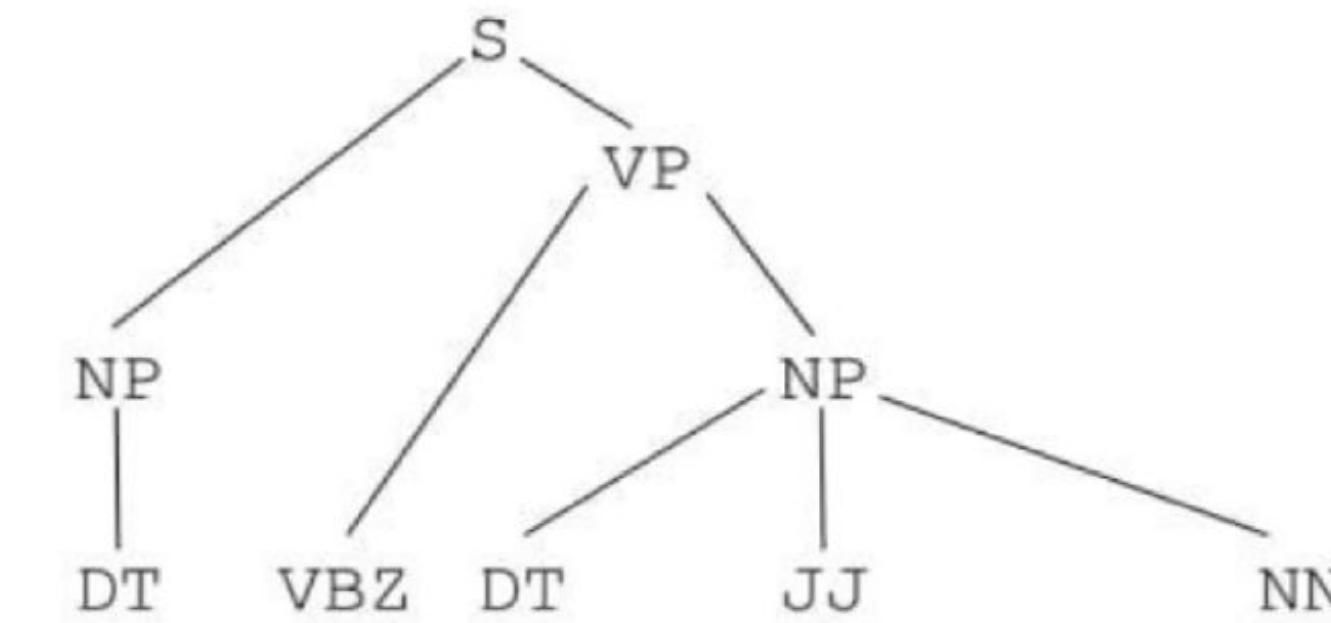
- Syntactic parsing



Semantics

- Named entity recognition
- Word sense disambiguation
- Semantic role labeling

SYNTAX



PART OF SPEECH

WORDS

MORPHOLOGY

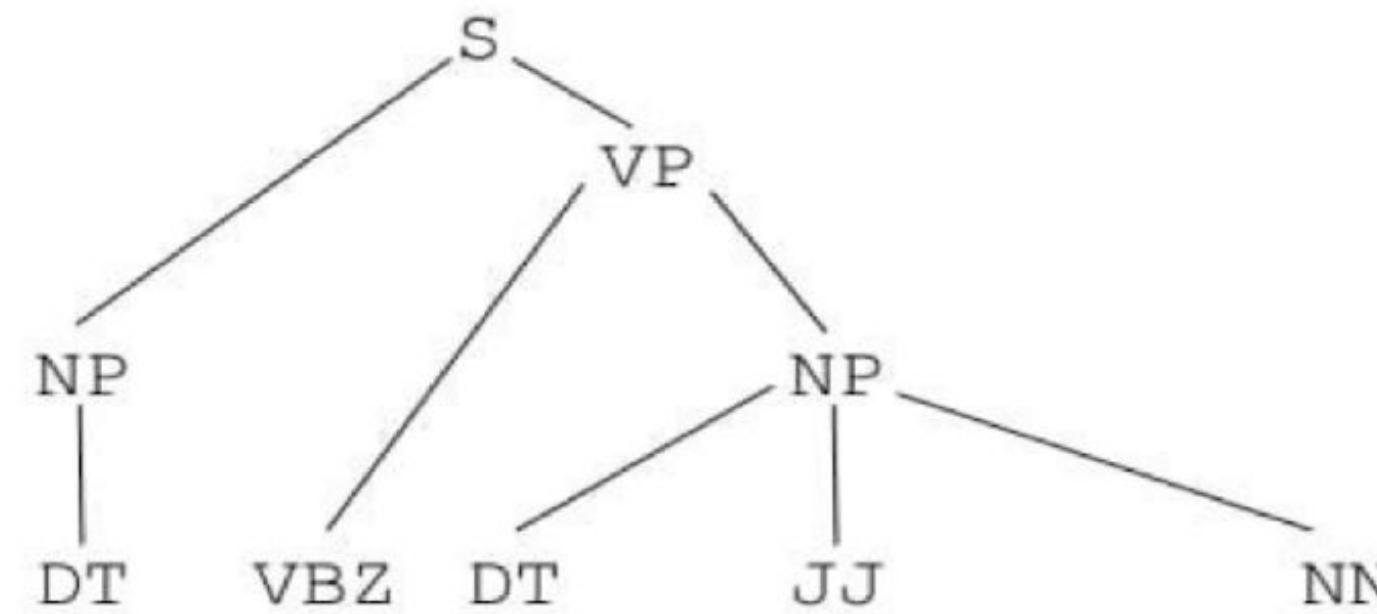
SEMANTICS

This is a simple sentence
be 3sg present SIMPLE1 having few parts SENTENCE1
string of words satisfying the grammatical rules of a language



Discourse

SYNTAX



PART OF SPEECH

WORDS

This is a simple sentence

be
3sg
present SIMPLE1
having
few
parts SENTENCE1
string of words
satisfying the
grammatical rules
of a language

MORPHOLOGY

SEMANTICS

DISCOURSE

But it is an instructive one.

CONTRAST



eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

Where Are We Now?



eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

Where Are We Now?

Baseline mutual information model (Li et al. 2015)

A: Where are you going? (1)

B: I'm going to the restroom. (2)

A: See you later. (3)

B: See you later. (4)

A: See you later. (5)

B: See you later. (6)

...

...

A: how old are you? (1)

B: I'm 16. (2)

A: 16? (3)

B: I don't know what you are talking about. (4)

A: You don't know what you are saying. (5)

B: I don't know what you are talking about . (6)

A: You don't know what you are saying. (7)

...

VS



eskwelabs.com



Eskwelabs

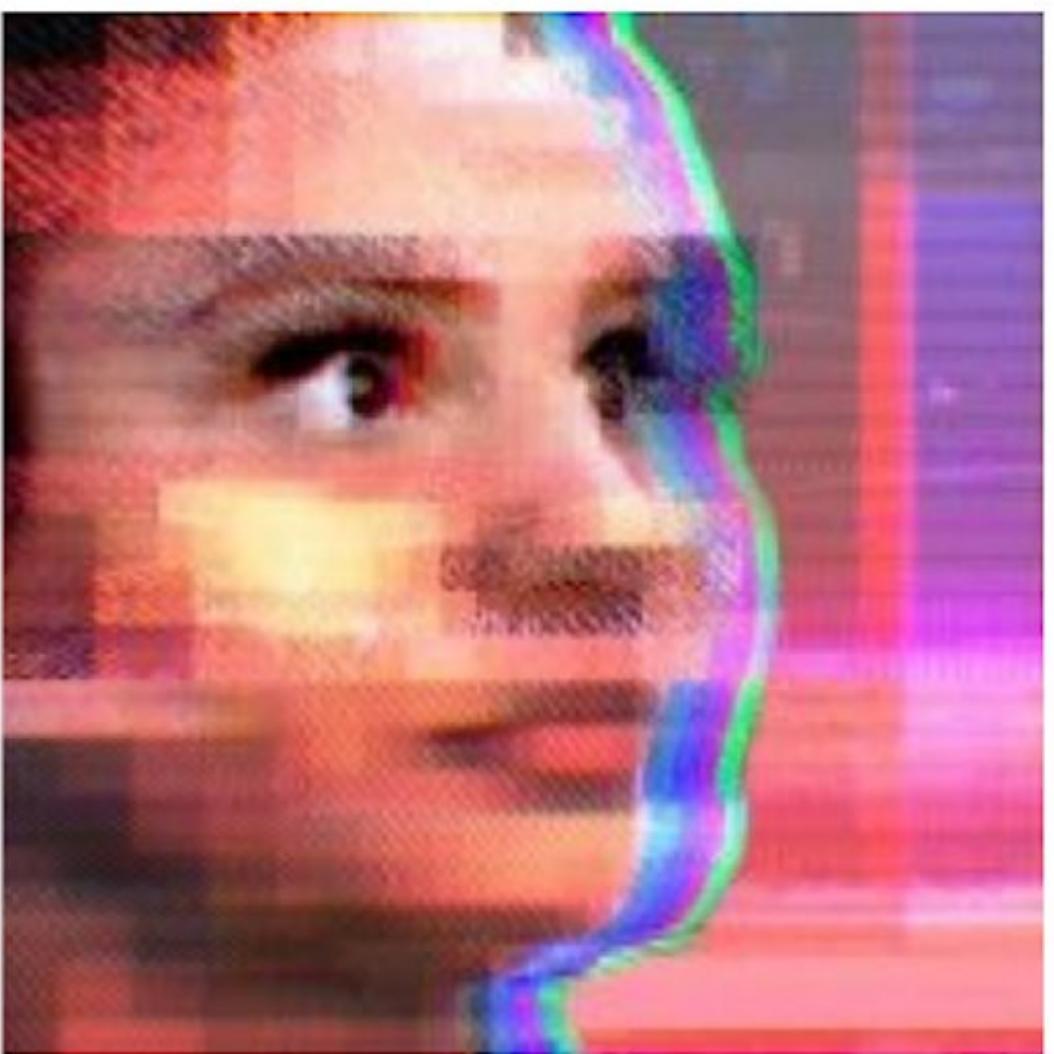


eskwelabs



@eskwelabs_ph

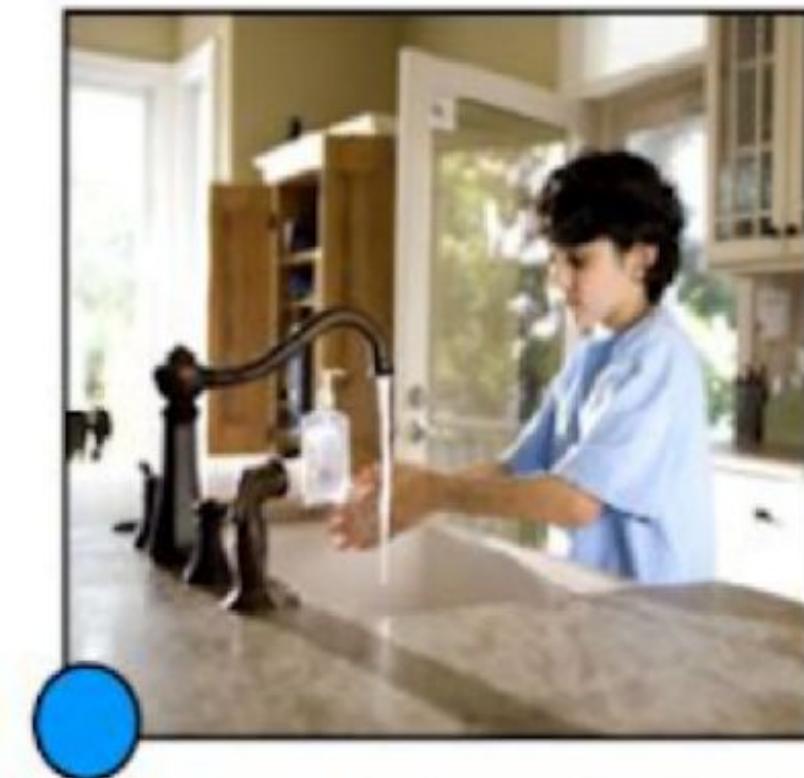
Where Are We Now?



[https://www.theverge.com/2016/3/24/11297050
/tay-microsoft-chatbot-racist](https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist)



woman cooking



man fixing faucet

Zhao, J., Wang, T., Yatskar, M., Ordonez, V and Chang, M.-W. (2017) Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraint. EMNLP



eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

How can computers understand language?

Mapping words and numbers with word embedding

the 0.0897 0.0160 -0.0571 0.0405 -0.0696 ...

and -0.0314 0.0149 -0.0205 0.0557 0.0205 ...

of -0.0063 -0.0253 -0.0338 0.0178 -0.0966 ...

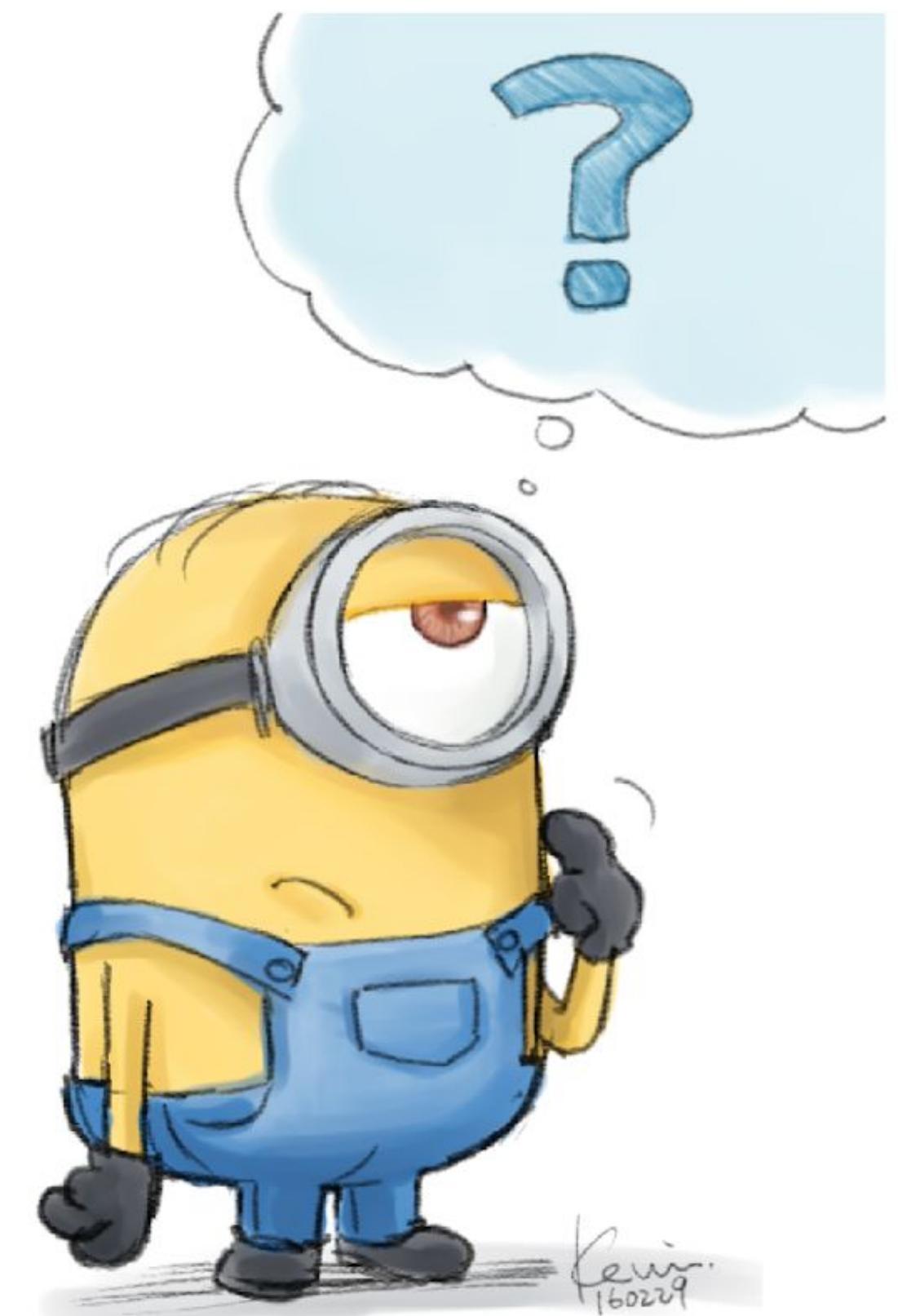
to 0.0495 0.0411 0.0041 0.0309 -0.0044 ...

in -0.0234 -0.0268 -0.0838 0.0386 -0.0321 ...



Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations



eskwelabs.com



Eskwelabs



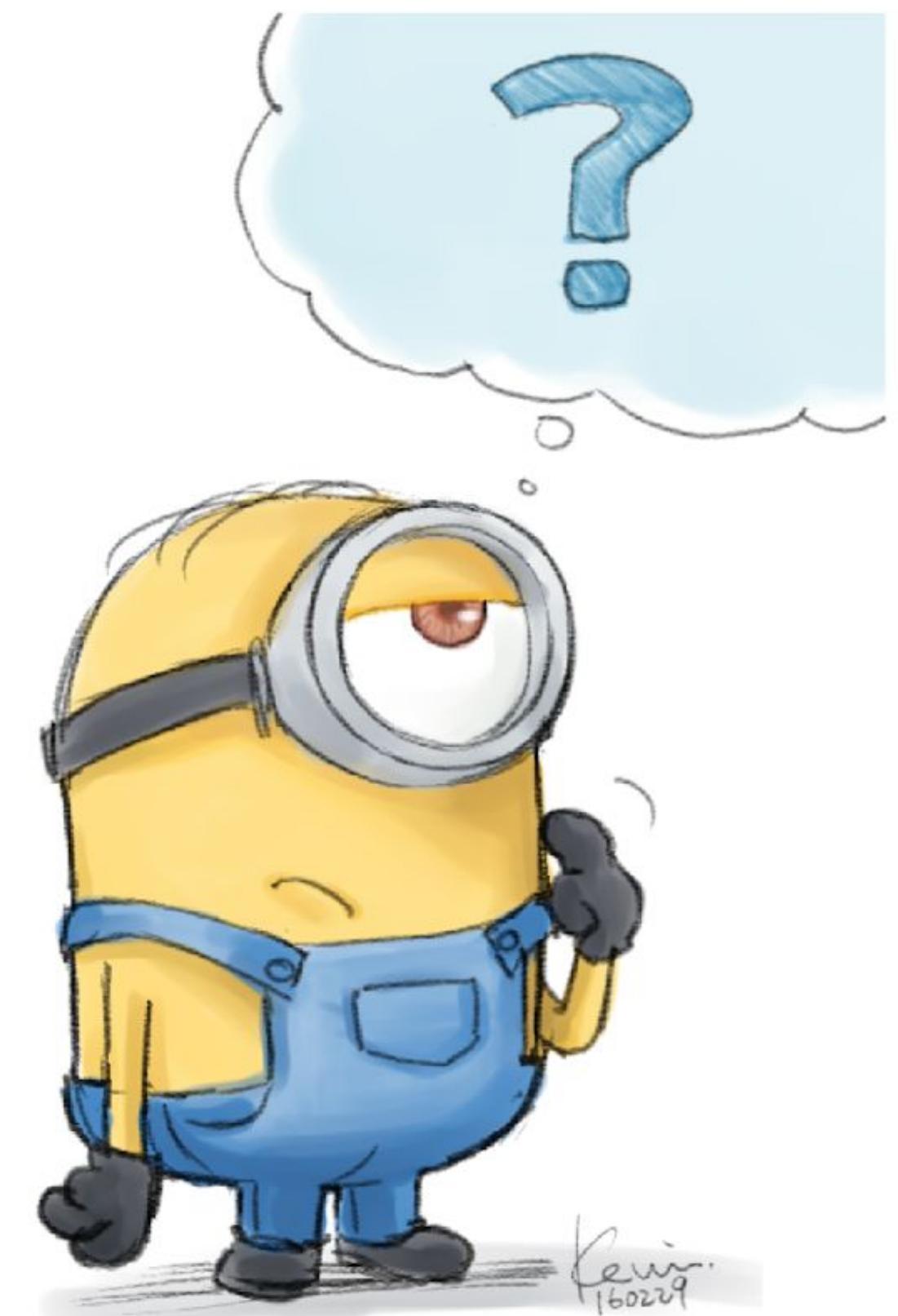
eskwelabs



@eskwelabs_ph

Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations



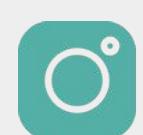
eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

Ambiguity

- Ambiguity at multiple levels
 - Word senses: **bank** (finance or river ?)
 - Part of speech: **chair** (noun or verb ?)
 - Syntactic structure: **I can see a man with a telescope**
 - Multiple: **I made her duck**





“One morning I shot
an elephant in my pajamas”



I made her duck

[SLP2 ch. 1]

- I cooked waterfowl for her
- I cooked waterfowl belonging to her
- I created the (plaster?) duck she owns
- I caused her to quickly lower her head or body
- ...



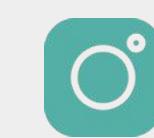
eskwelabs.com



Eskwelabs

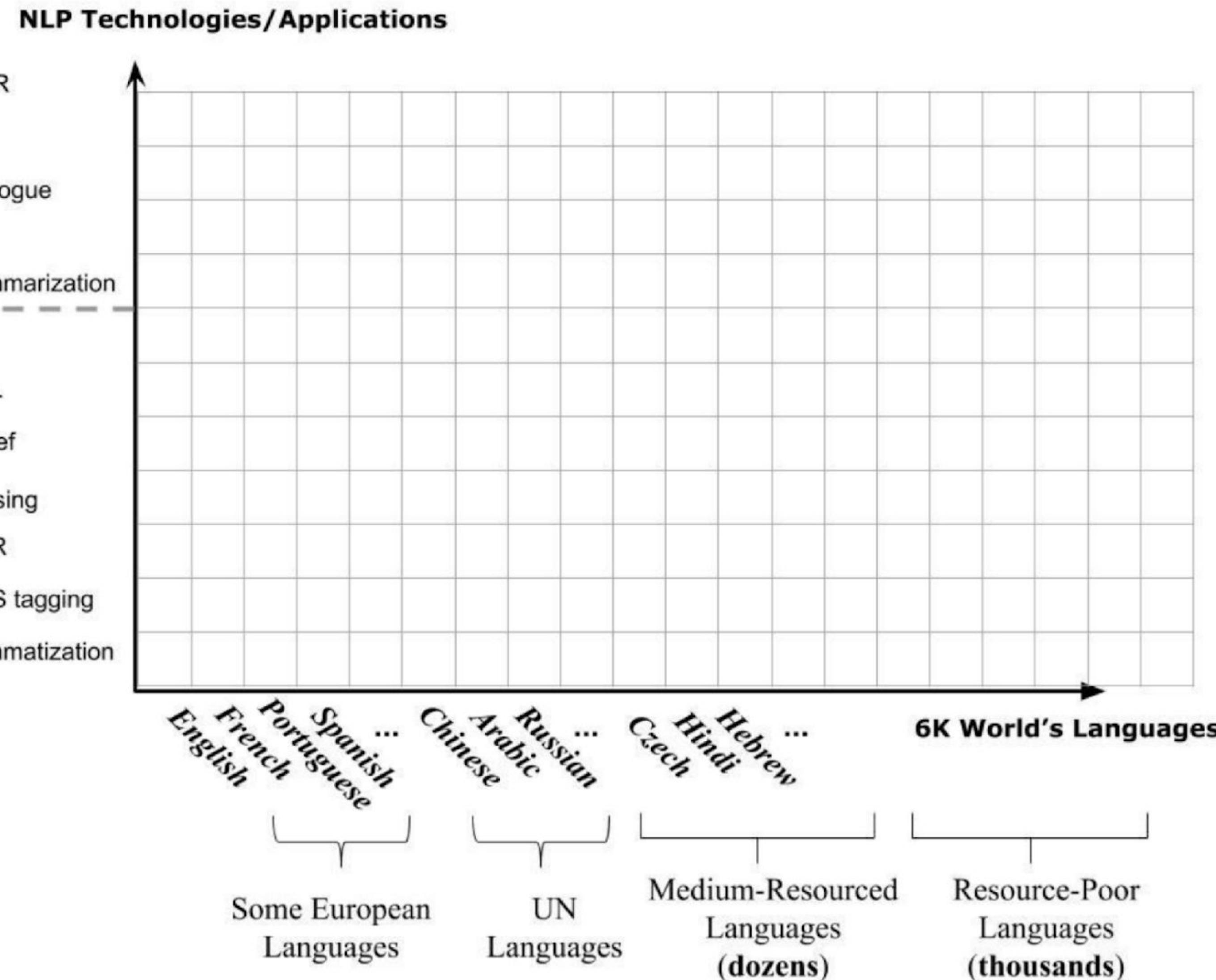


eskwelabs



@eskwelabs_ph

Ambiguity and Scale



eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

The Challenges of “Words”

- Segmenting text into words
- Morphological variation
- Words with multiple meanings: bank, mean
- Domain-specific meanings: latex
- Multiword expressions: make a decision, take out, make up



eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

Part of Speech Tagging

ikr smh he asked fir yo last name

so he can add u on fb lololol



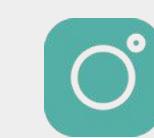
eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

Part of Speech Tagging

I know, right shake my head for your
ikr smh he asked fir yo last name

you Facebook laugh out loud
so he can add u on fb lololol



eskwelabs.com



Eskwelabs



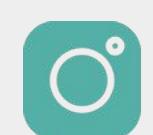
eskwelabs



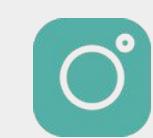
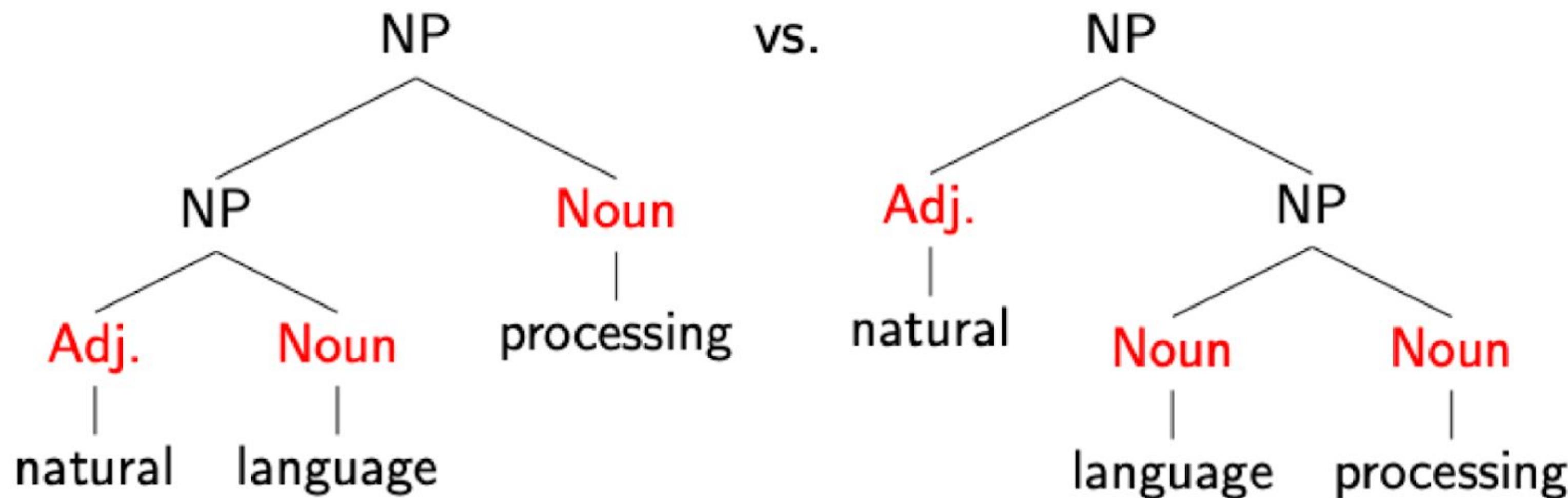
@eskwelabs_ph

Part of Speech Tagging

I know, right	shake my head		for	your			
ikr	smh	he	asked	fir	yo	last	name
!	G	O	V	P	D	A	N
interjection	acronym	pronoun	verb	prep.	det.	adj.	noun
		you		Facebook		laugh out loud	
so	he	can	add	u	on	fb	lololol
P	O	V	V	O	P	^	!
preposition				proper noun			



Syntax



Morphology + Syntax



A ship-shipping
ship, shipping
shipping-ships



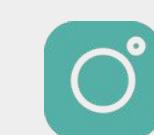
eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

Semantics

- Every fifteen minutes a woman in this country gives birth.



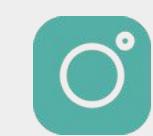
eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

Semantics

- Every fifteen minutes a woman in this country gives birth. Our job is to find this woman, and stop her!

– Groucho Marx



Syntax + Semantics

- We saw the woman with the telescope wrapped in paper.



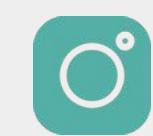
eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

Syntax + Semantics

- We saw the woman with the telescope wrapped in paper.
 - Who has the telescope?
 - Who or what is wrapped in paper?
 - An even of perception, or an assault?



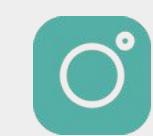
eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

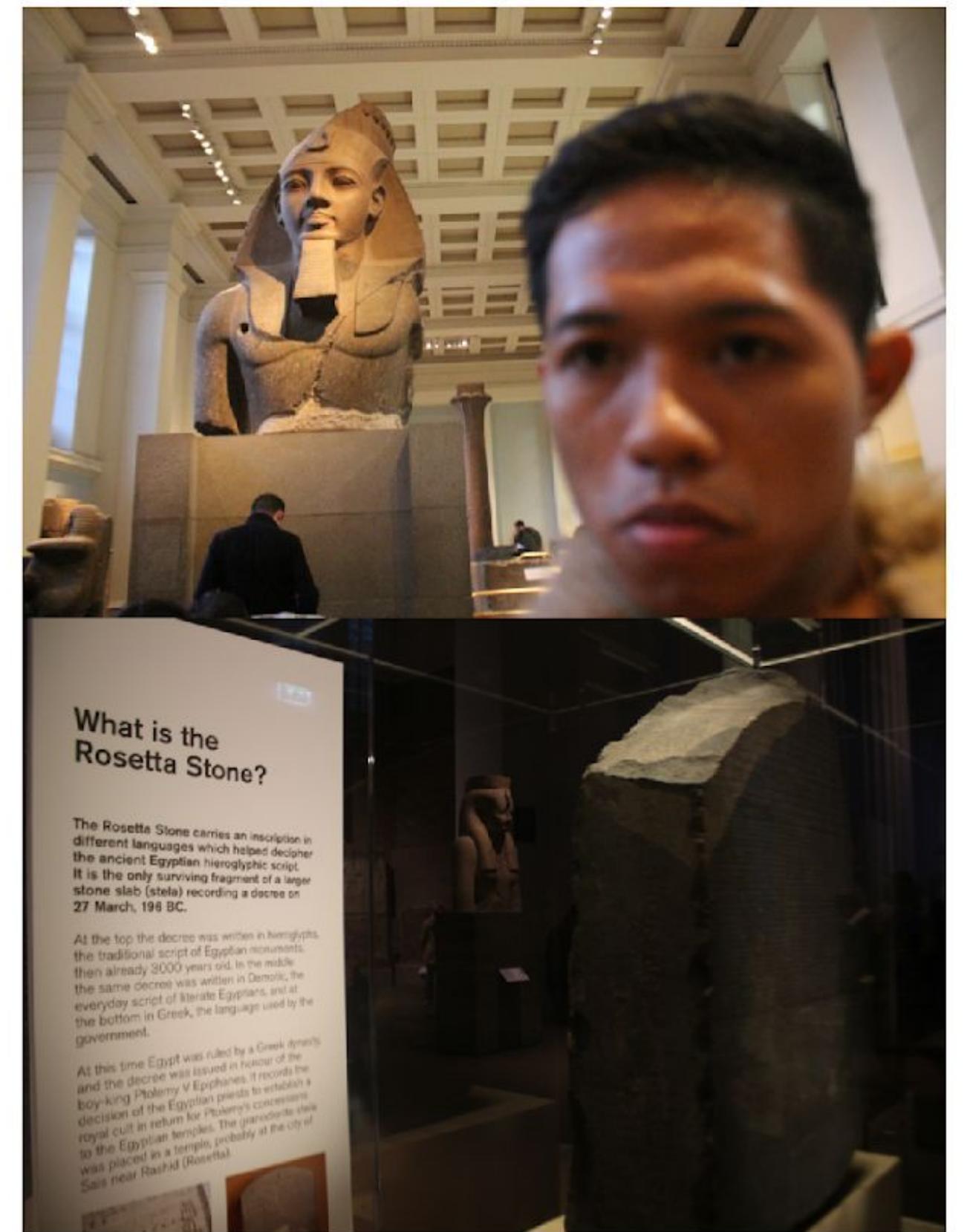
Dealing with Ambiguity

- How can we model ambiguity?
 - Non-probabilistic methods (CKY parsers for syntax) return **all possible analyses**
 - Probabilistic models (HMMs for POS tagging, PCFGs for syntax) and algorithms (Viterbi, probabilistic CKY) return **the best possible analyses**, i.e., the most probable one
- But the “best” analysis is only good if our probabilities are accurate. Where do they come from?



Corpora

- A corpus is a collection of text
 - Often annotated in some way
 - Sometimes just lots of text
- Examples
 - Penn Treebank: 1M words of parsed WSJ
 - Canadian Hansards: 10M+ words of French/English sentences
 - Yelp reviews
 - The Web!



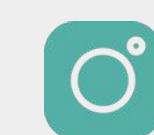
eskwelabs.com



Eskwelabs



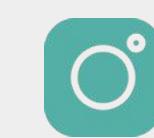
eskwelabs



@eskwelabs_ph

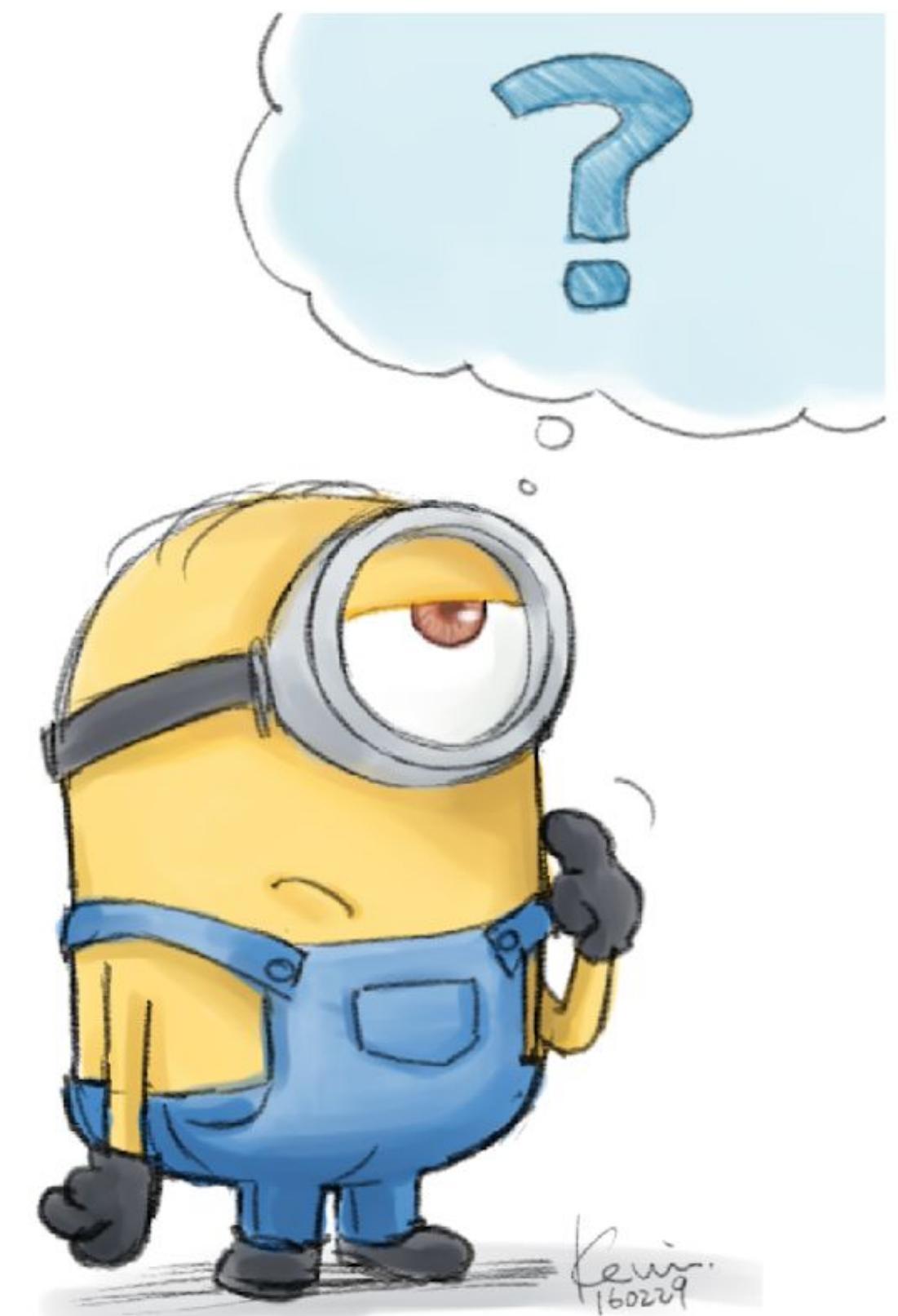
Statistical NLP

- Like most other parts of AI, NLP is dominated by statistical methods
 - Typically more robust than rule-based methods
 - Relevant statistics/probabilities are **learned from data**
 - Normally requires lots of data about any particular phenomenon



Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations



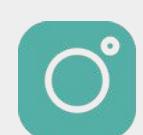
eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

Sparsity

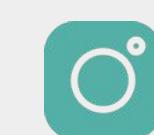
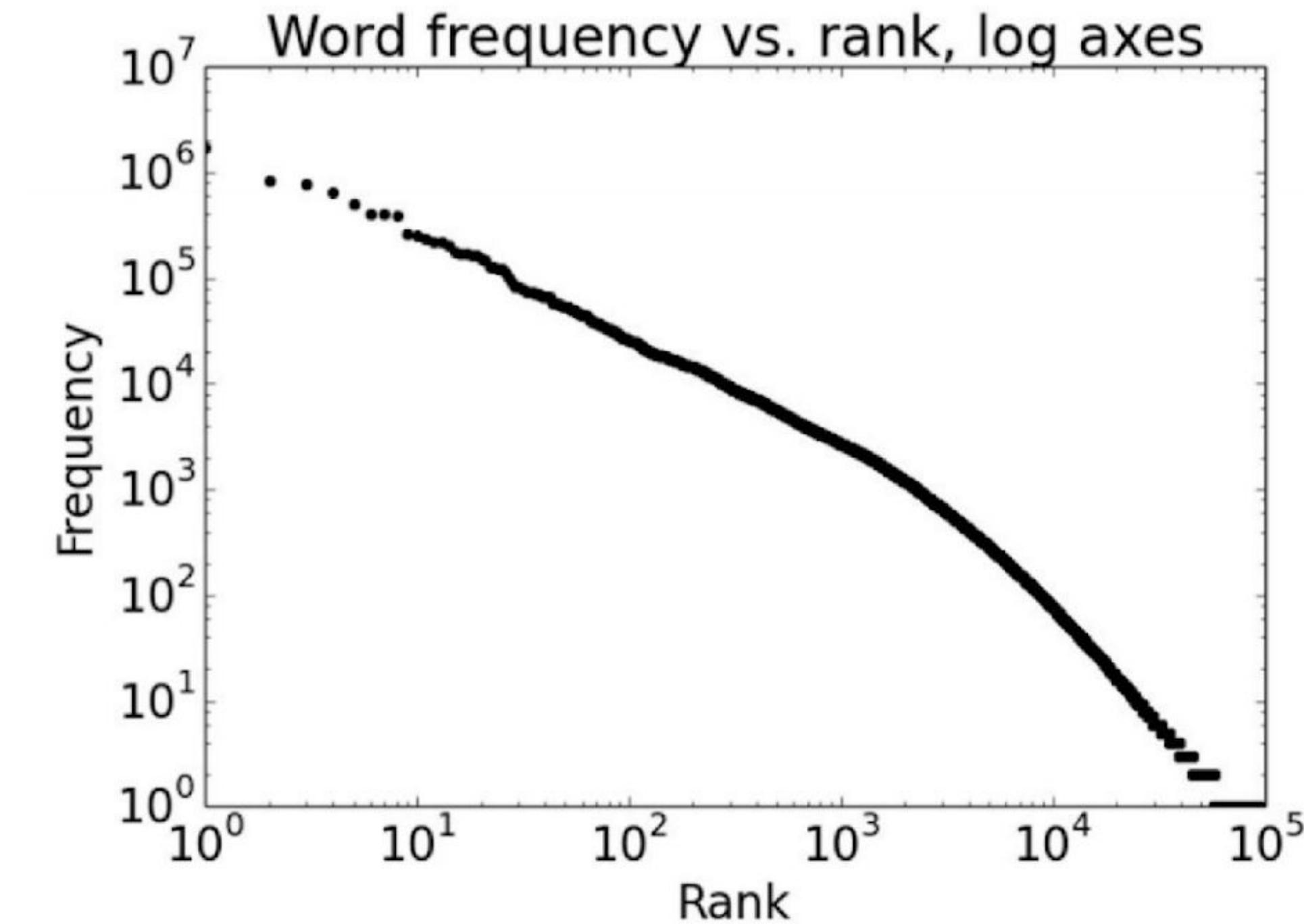
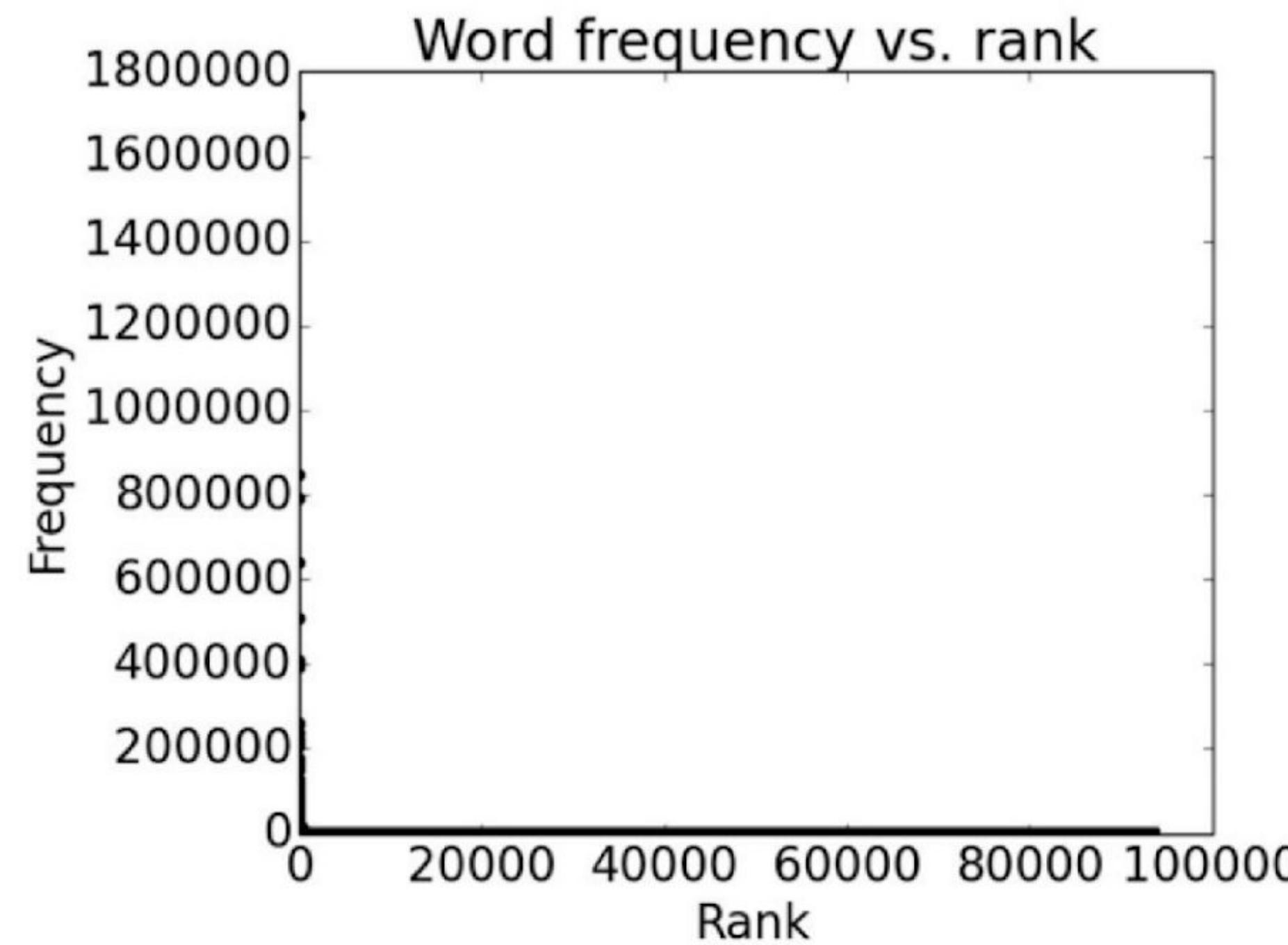
- Sparse data due to **Zipf's Law**
- Example: the frequency of different words in a large text corpus

any word		nouns	
Frequency	Token	Frequency	Token
1,698,599	the	124,598	European
849,256	of	104,325	Mr
793,731	to	92,195	Commission
640,257	and	66,781	President
508,560	in	62,867	Parliament
407,638	that	57,804	Union
400,467	is	53,683	report
394,778	a	53,547	Council
263,040	I	45,842	States



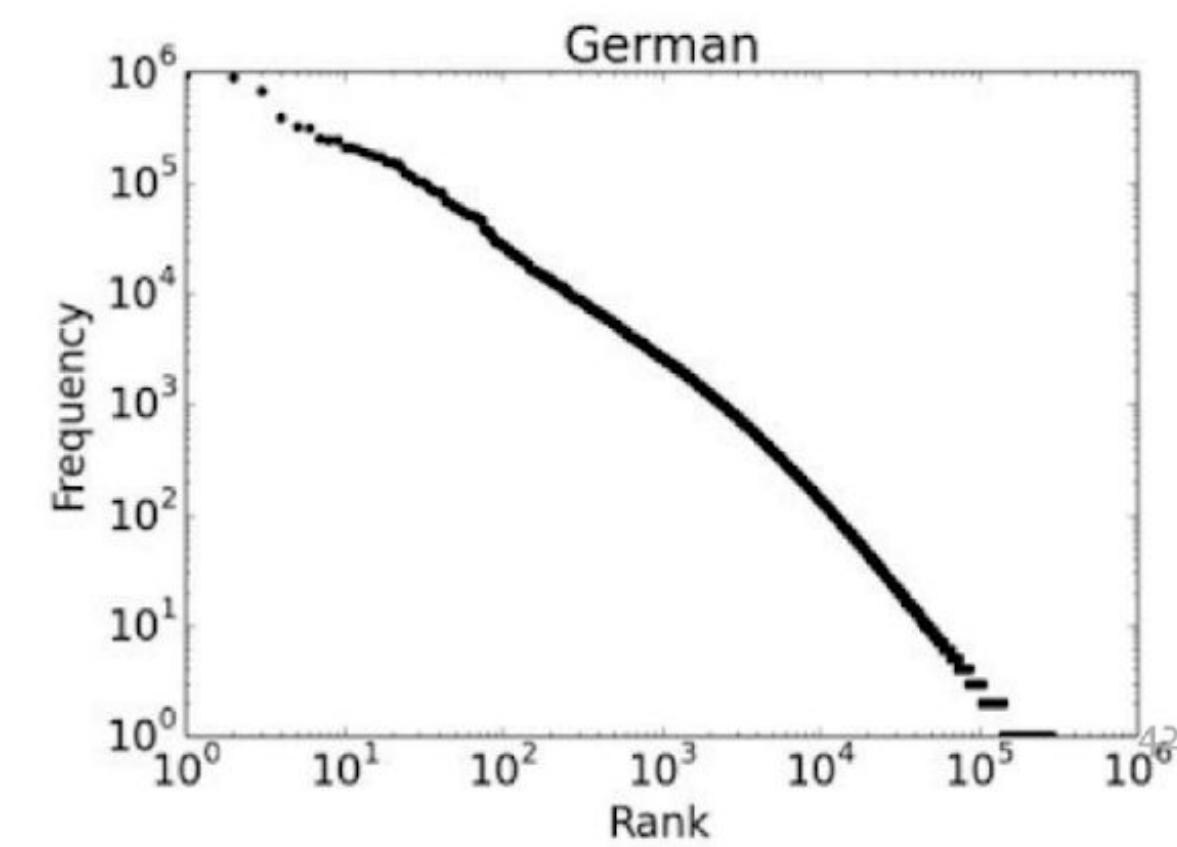
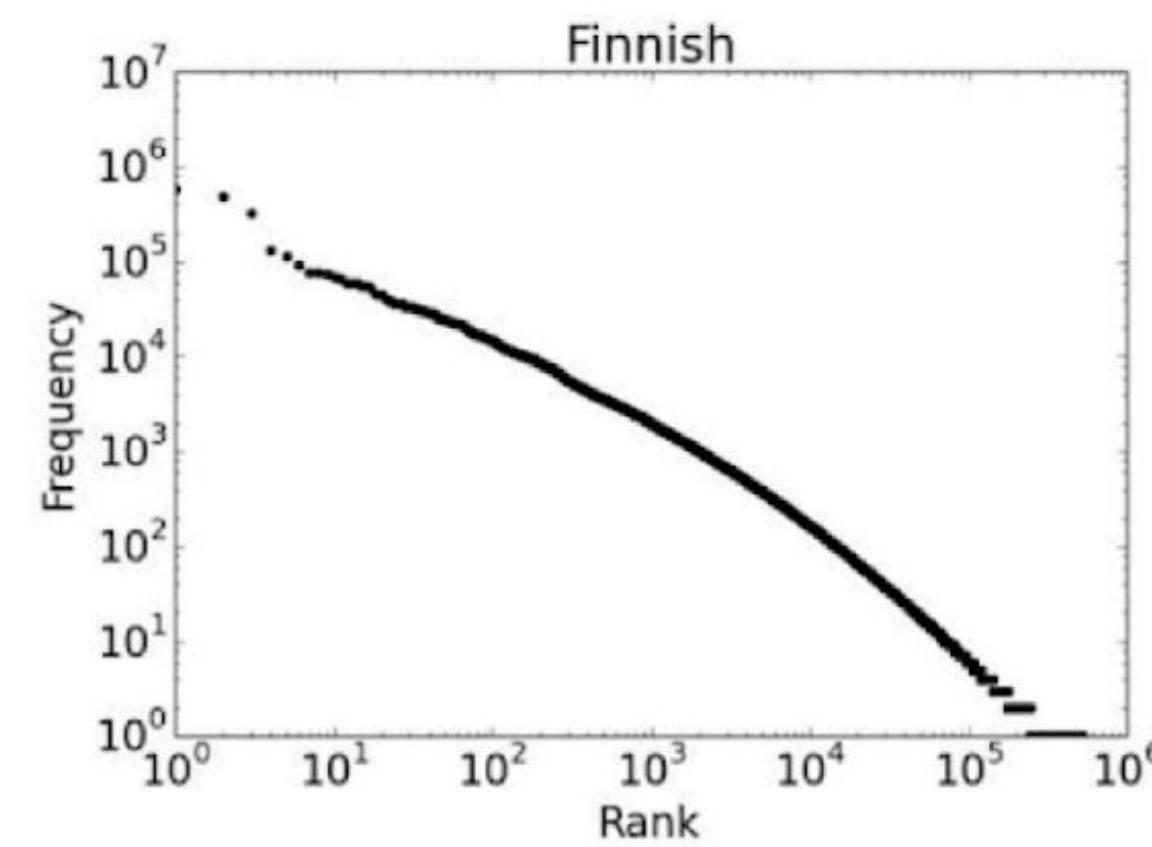
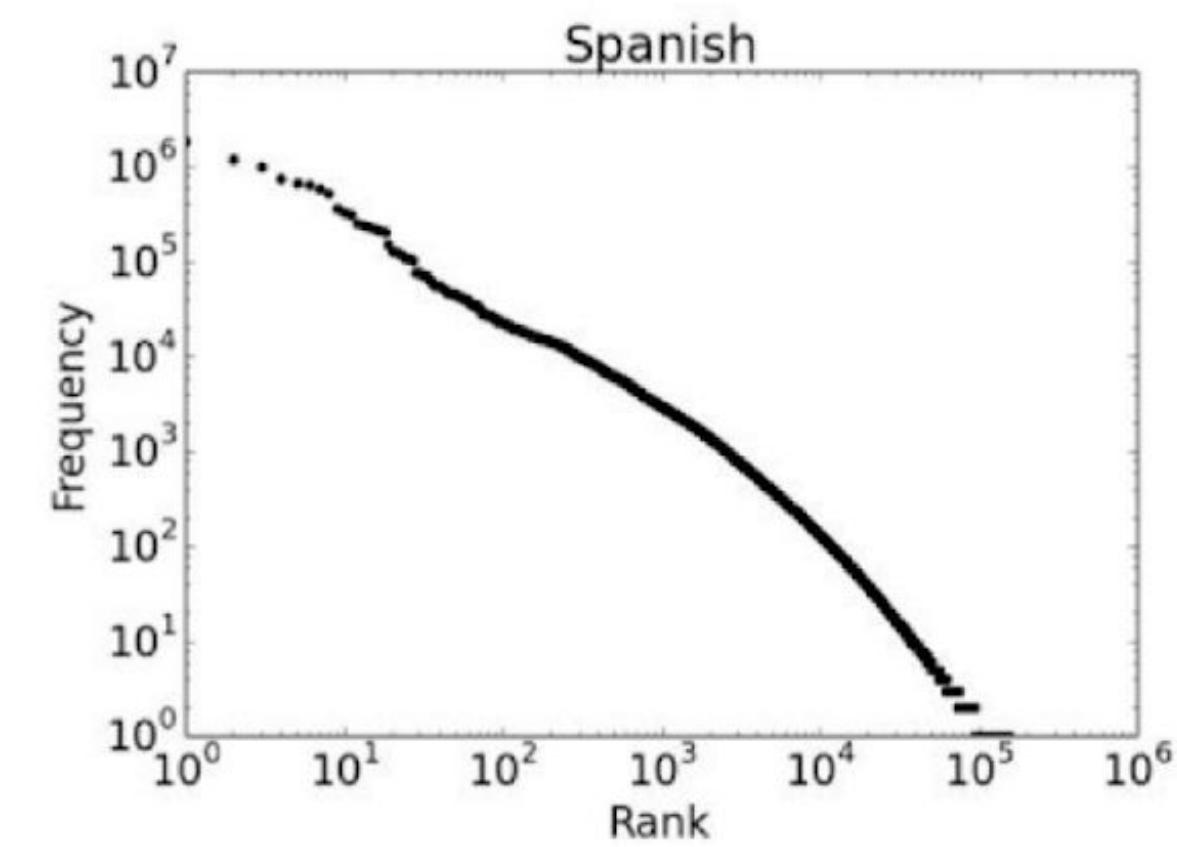
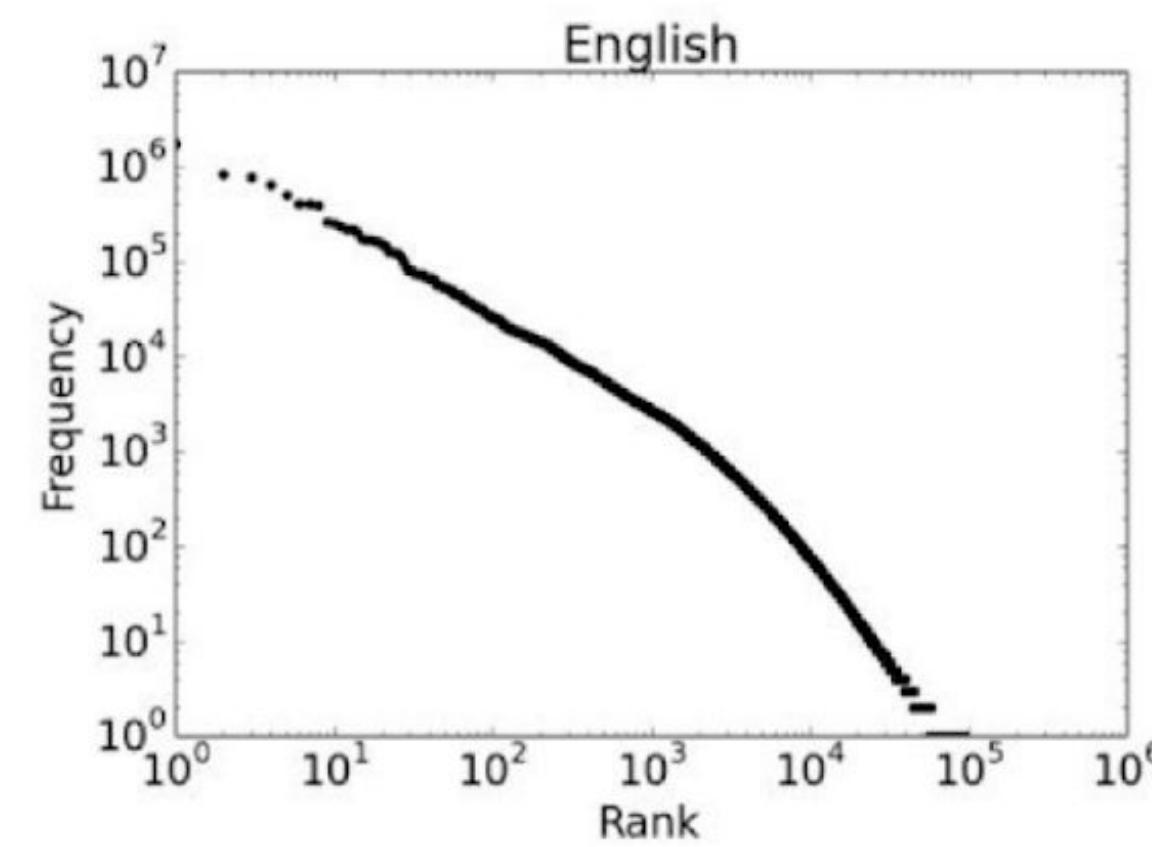
Sparsity

- Order words by frequency. What is the frequency of nth ranked word?



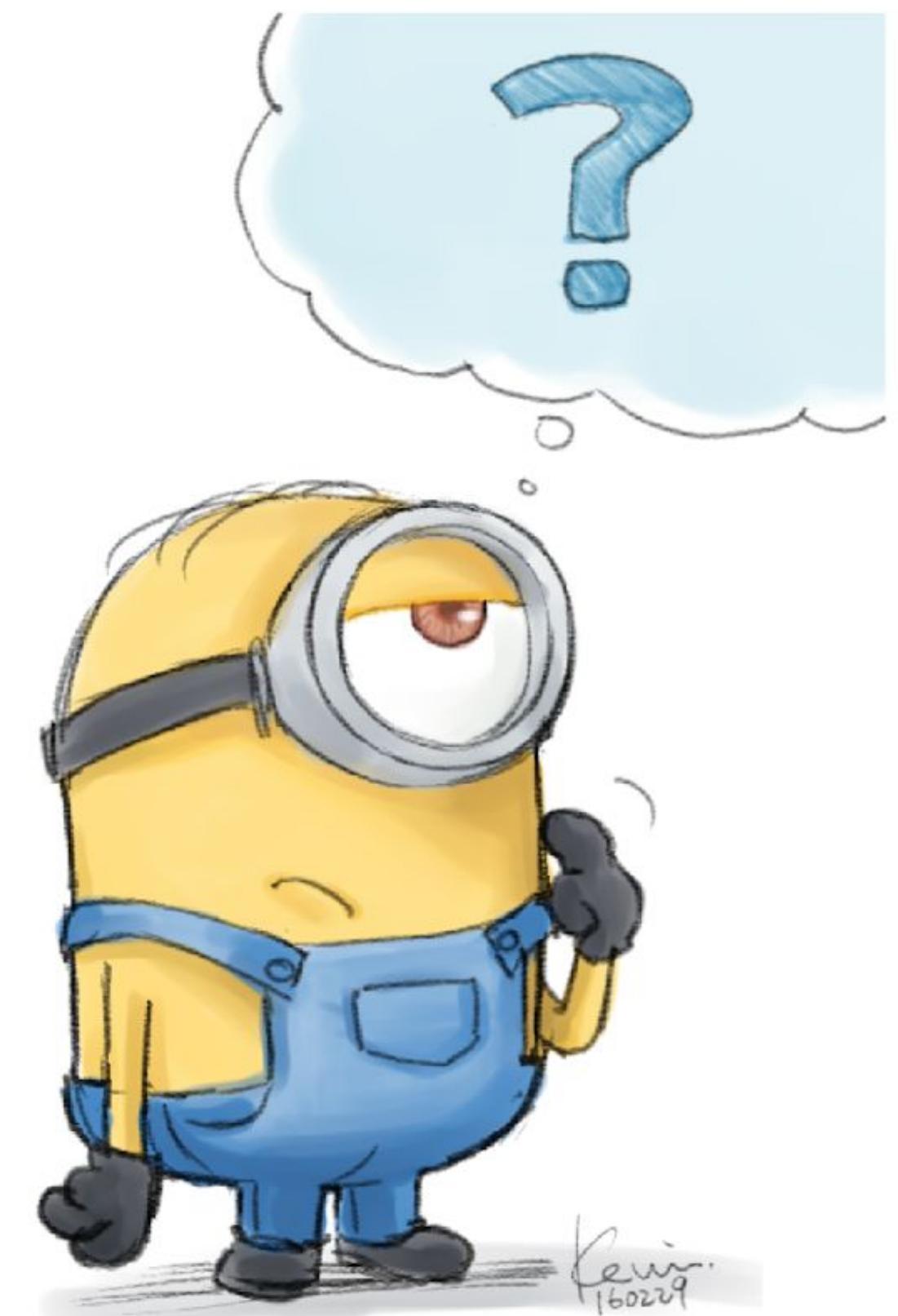
Sparsity

- Regardless of how large our corpus is, there will be a lot of infrequent words
- This means we need to find clever ways to estimate probabilities for things we have rarely or never seen



Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations



eskwelabs.com



Eskwelabs



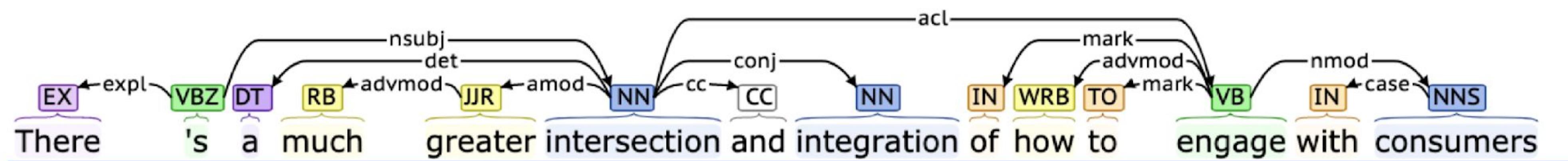
eskwelabs



@eskwelabs_ph

Variation

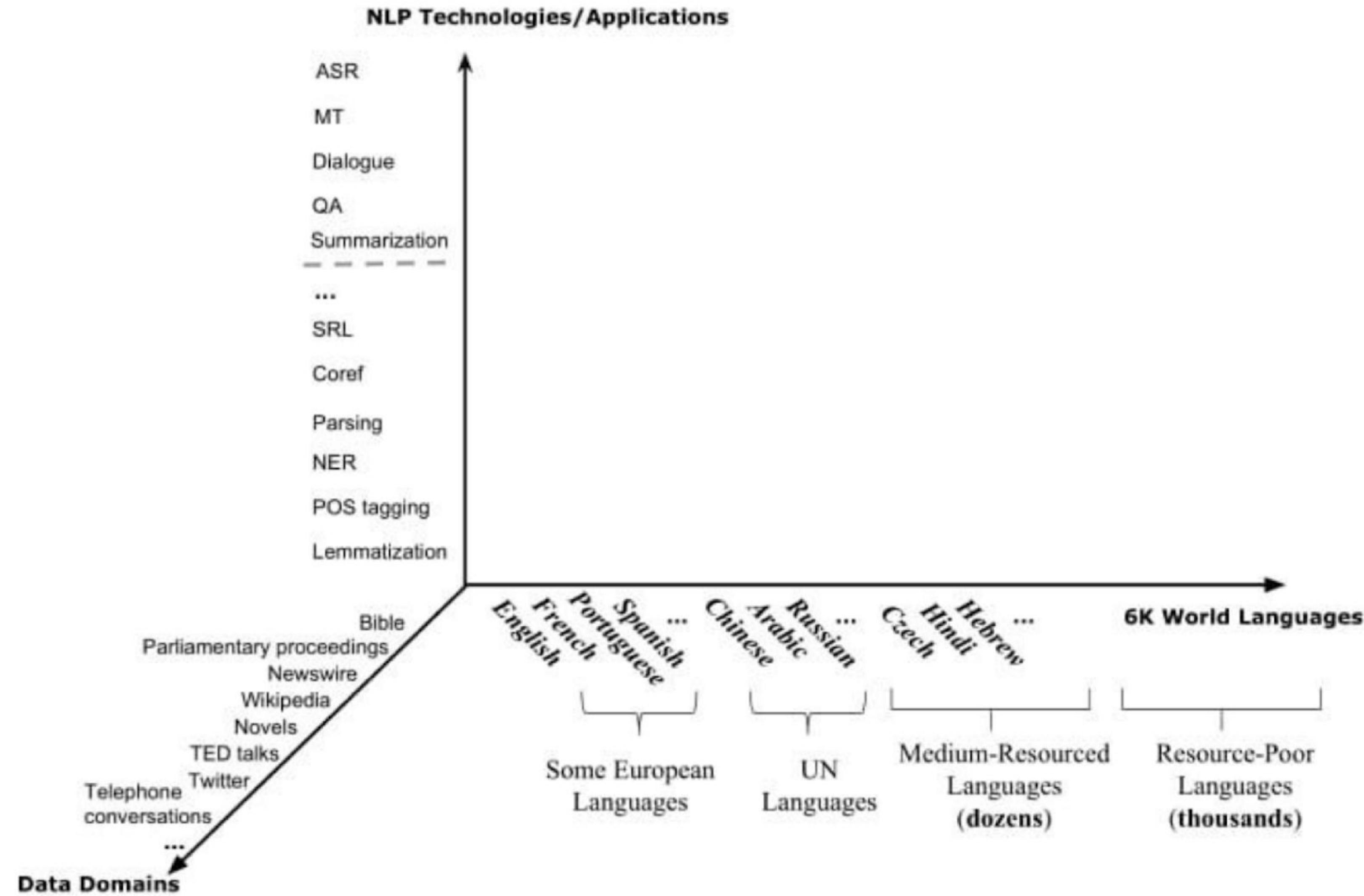
- Suppose we train a part of speech tagger or a parser on the **Wall Street Journal**



- What will happen if we try to use this tagger/parser for **social media**?
 - "ikr smh he asked fir yo last name so he can add u on fb lololol"



Variation



Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations



eskwelabs.com



Eskwelabs



eskwelabs

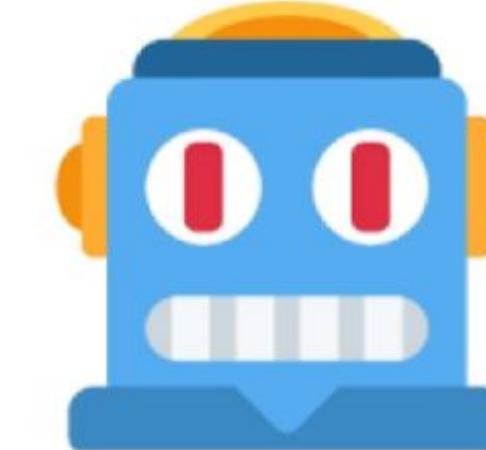


@eskwelabs_ph

Expressivity

- Not only can one form have different meanings (ambiguity) but the same meaning can be expressed with different forms:
 - *She gave the book to Tom* vs. *She gave Tom the book*
 - *Some kids popped by* vs. *A few children visited*
 - *Is that window still open?* vs. *Please close the window*

Please be quiet.
The talk will
begin shortly.



Shut up! The
talk is starting!



eskwelabs.com



Eskwelabs



eskwelabs

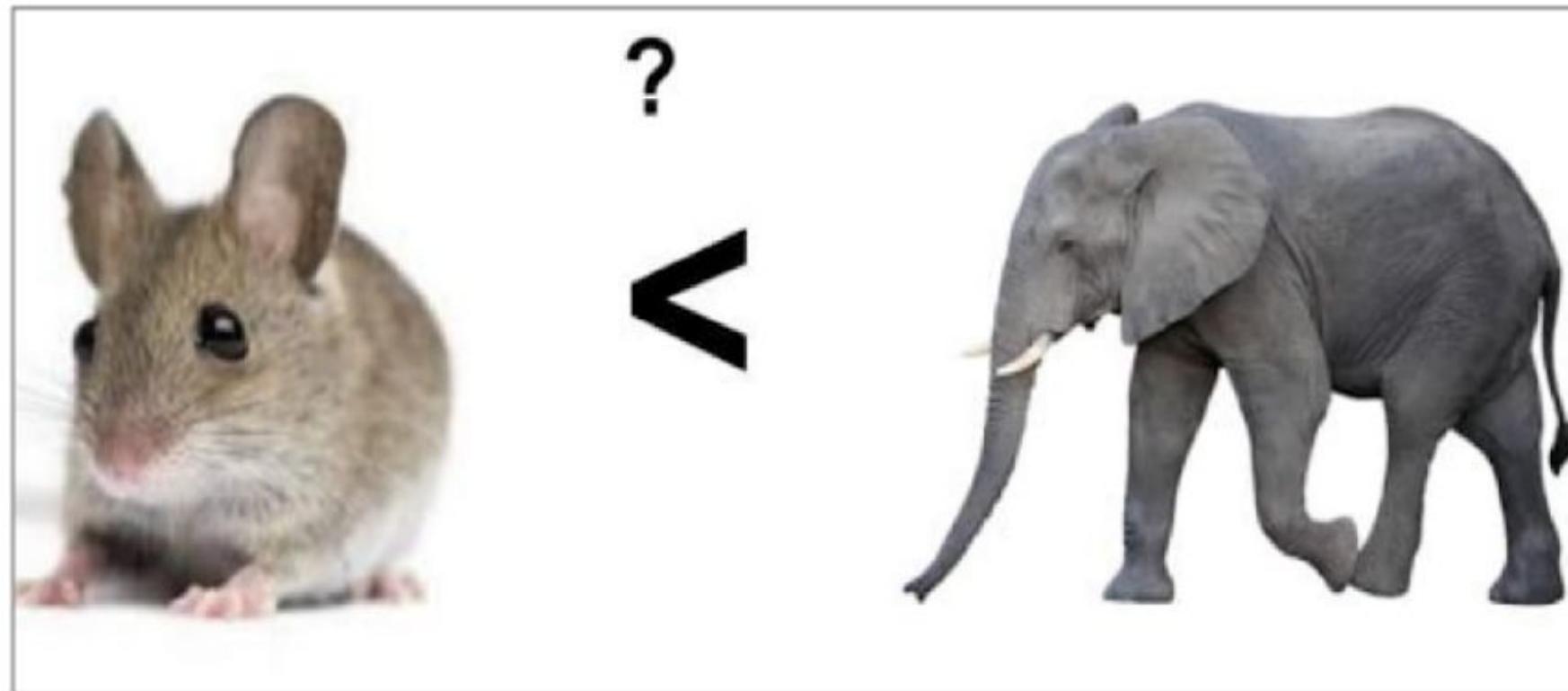


@eskwelabs_ph

Unmodeled Variables



“Drink this milk”



World knowledge

I dropped the glass on the floor and it broke

I dropped the hammer on the glass and it broke



eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

Unmodeled Representation

Very difficult to capture what is **R**, since we don't even know how to represent the knowledge a human has/needs:

- What is the “meaning” of a word or sentence?
- How to model context?
- Other general knowledge?



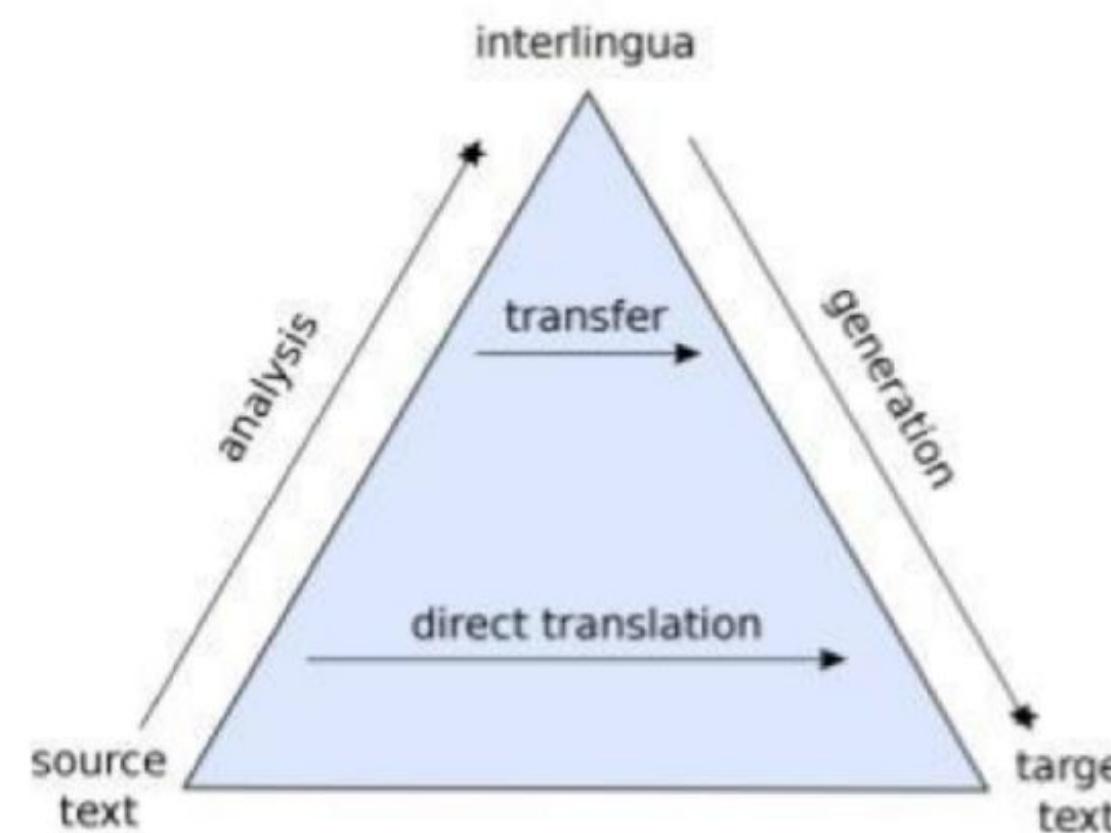
Desiderata for NLP Models

- Sensitivity to a wide range of phenomena and constraints in human language
- Generality across languages, modalities, genres, styles
- Strong formal guarantees (e.g., convergence, statistical efficiency, consistency)
- High accuracy when judged against expert annotations or test data
- Ethical



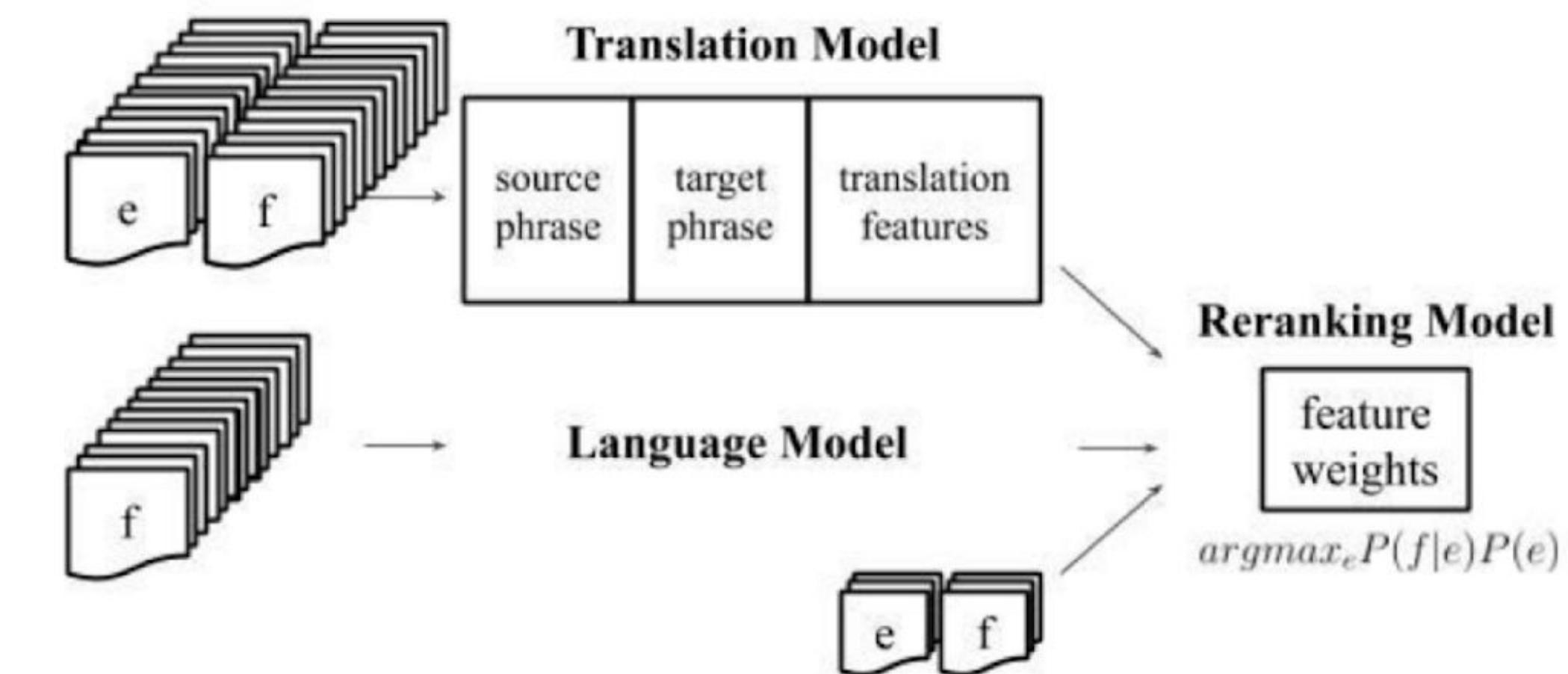
Symbolic and Probabilistic NLP

Logic-based/Rule-based NLP



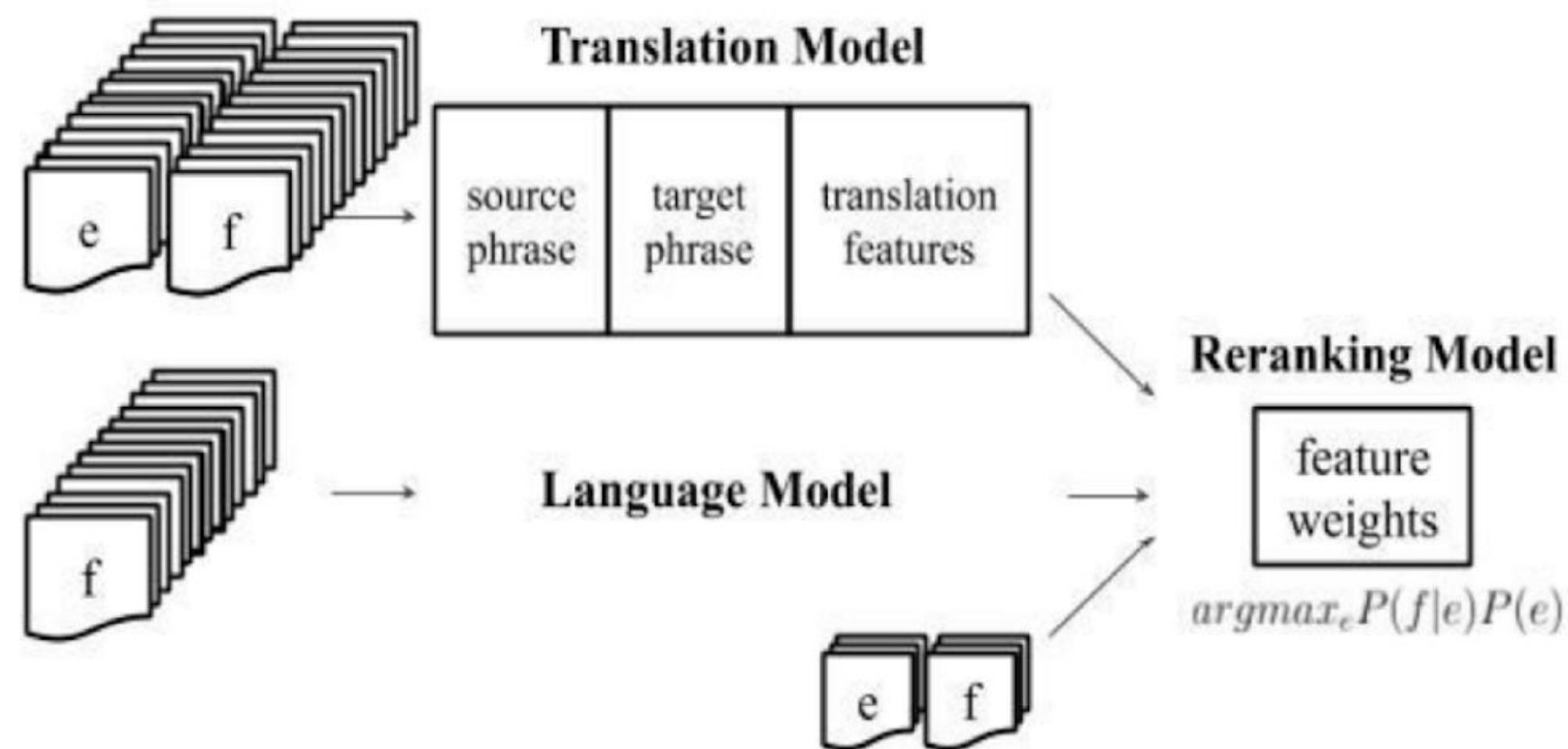
~90s

Statistical NLP

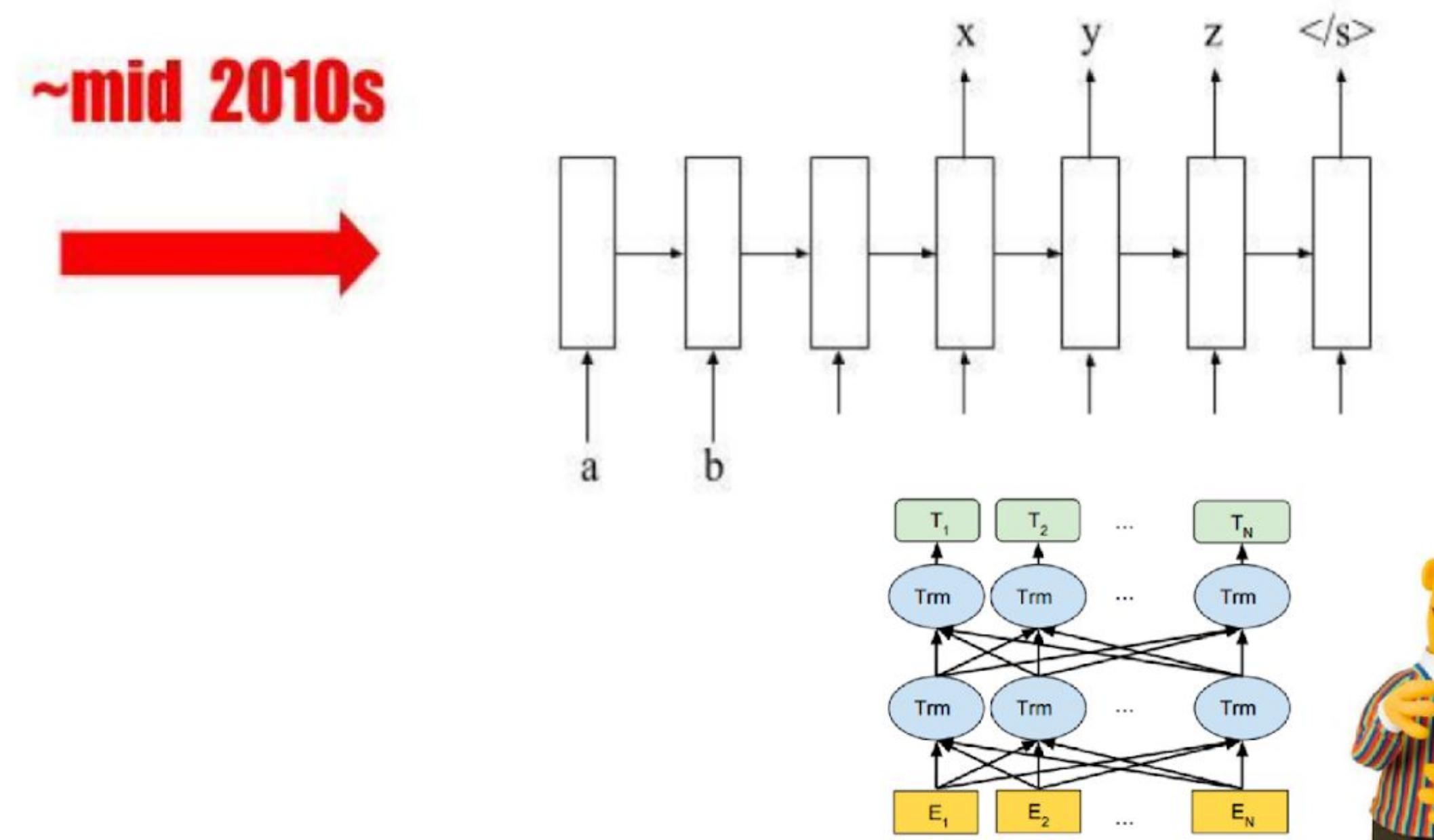


Probabilistic and Connectionist NLP

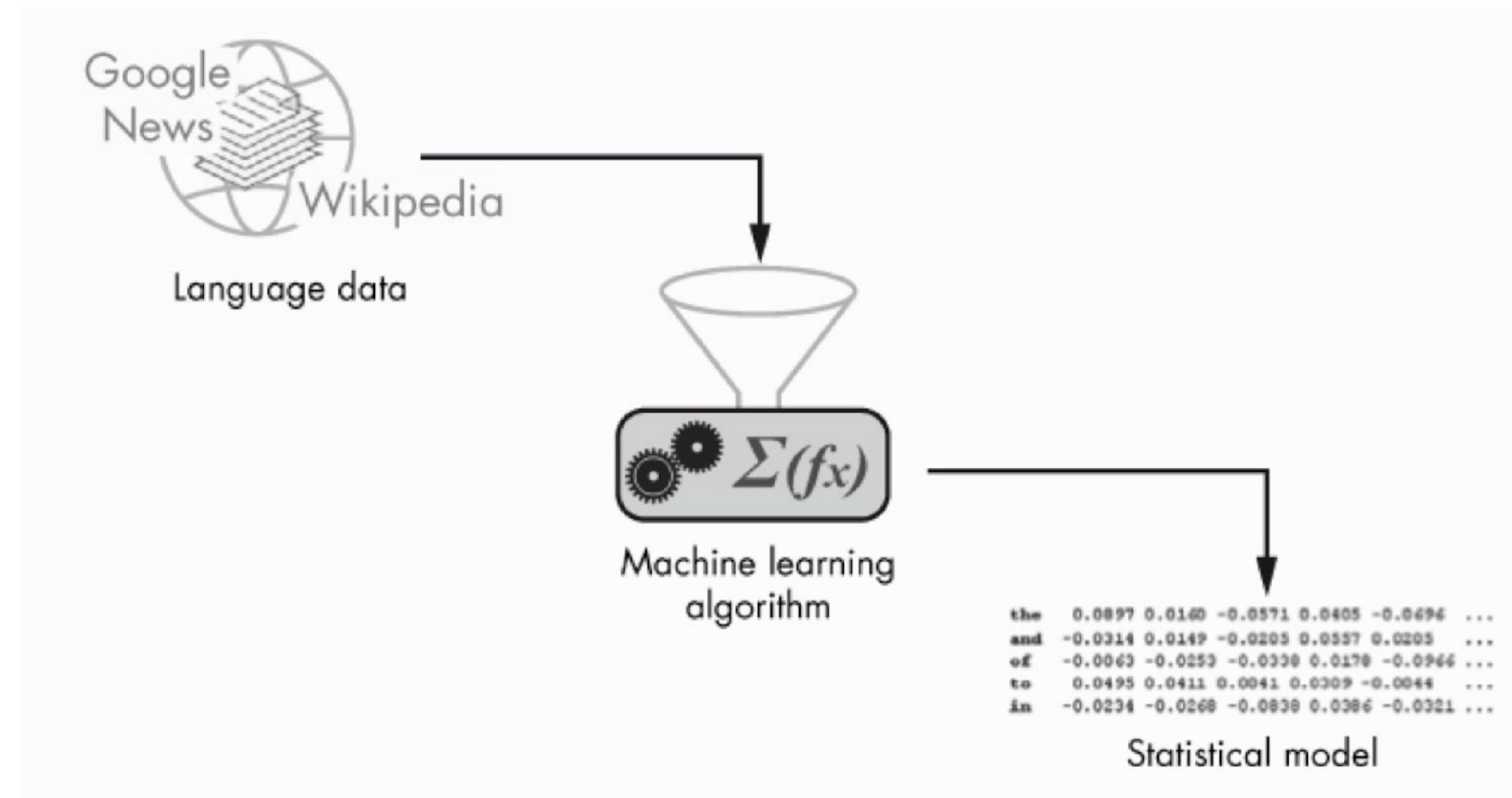
Engineered Features/Representations



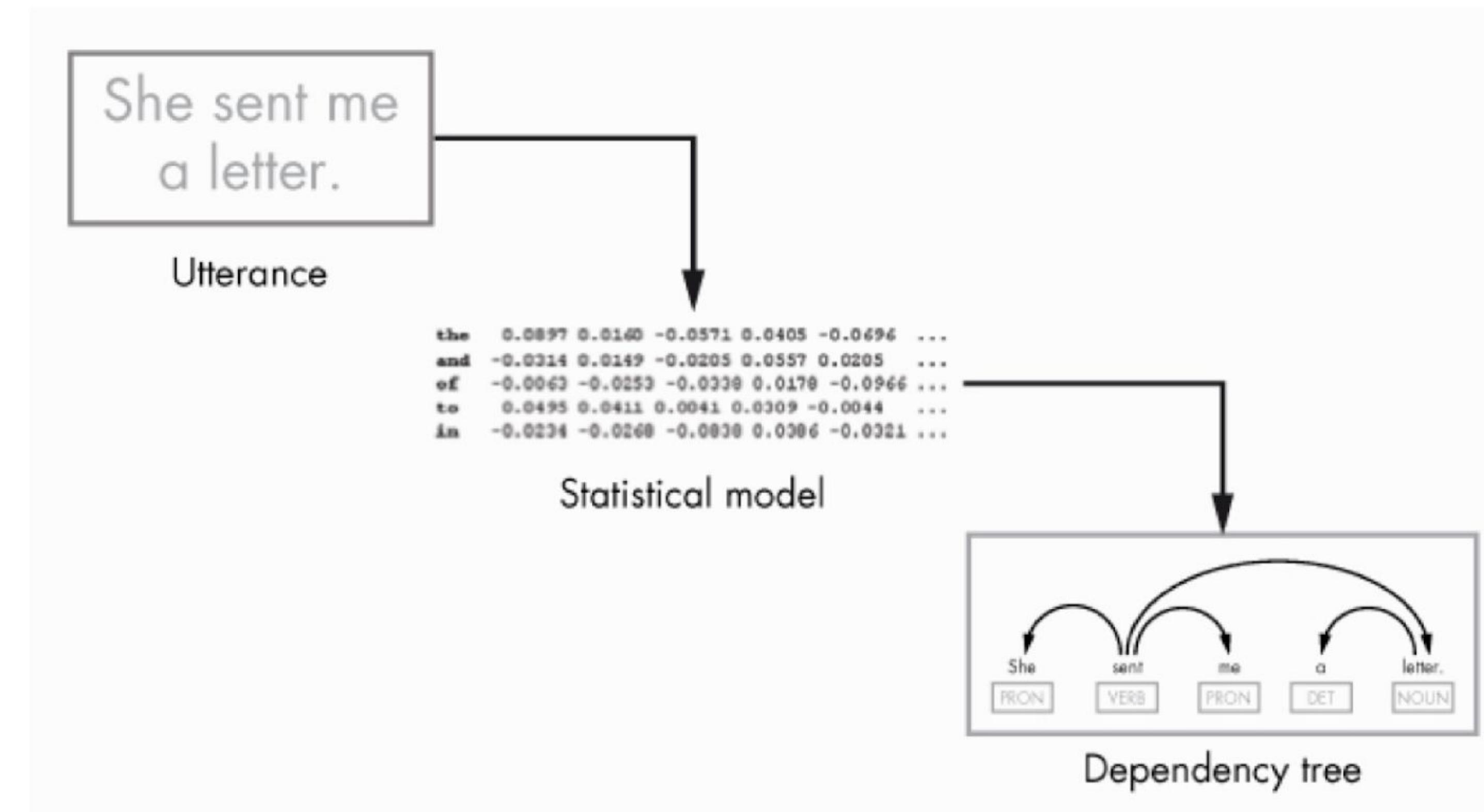
Learned Features/Representations



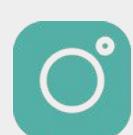
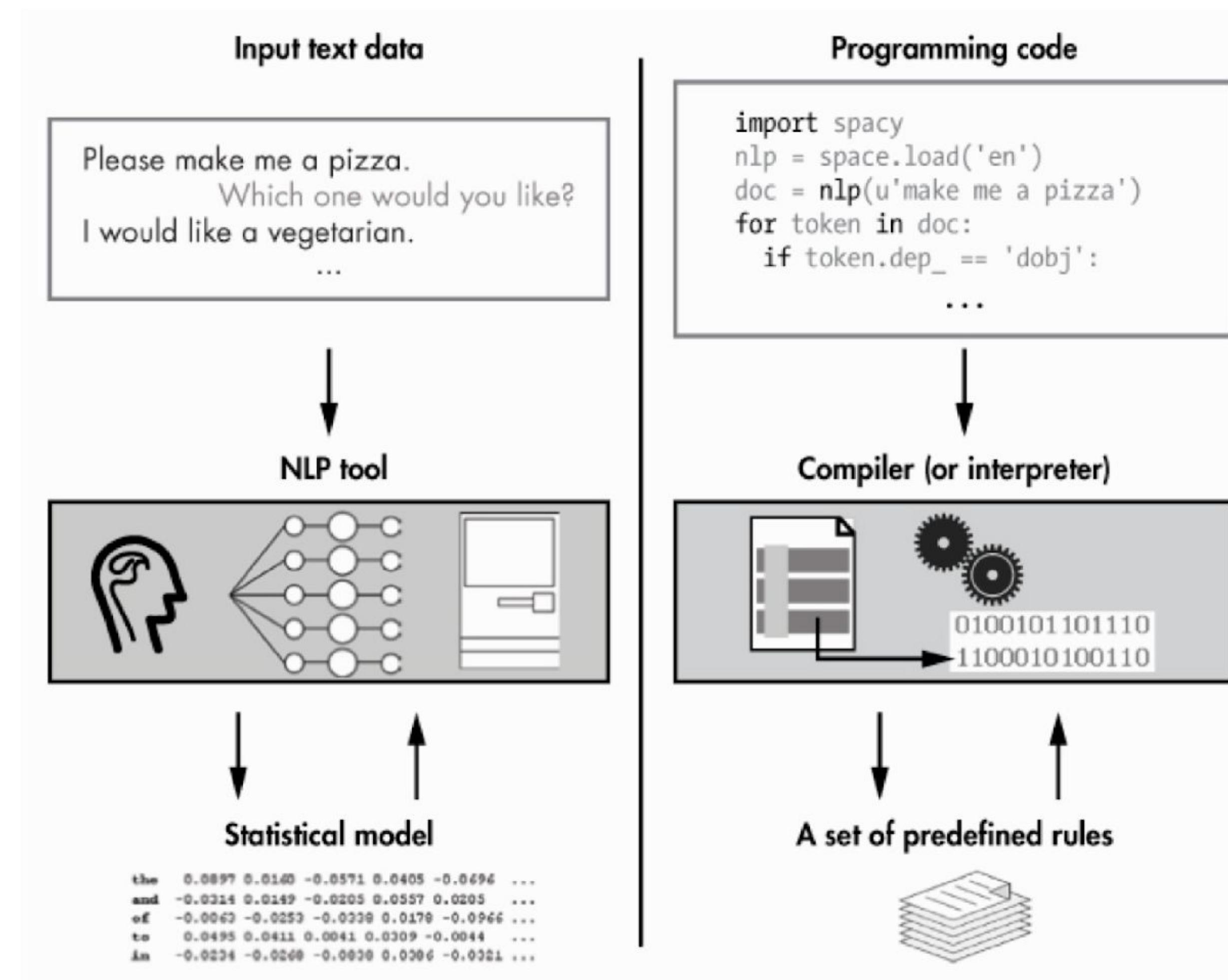
Machine Learning Pipeline for NLP



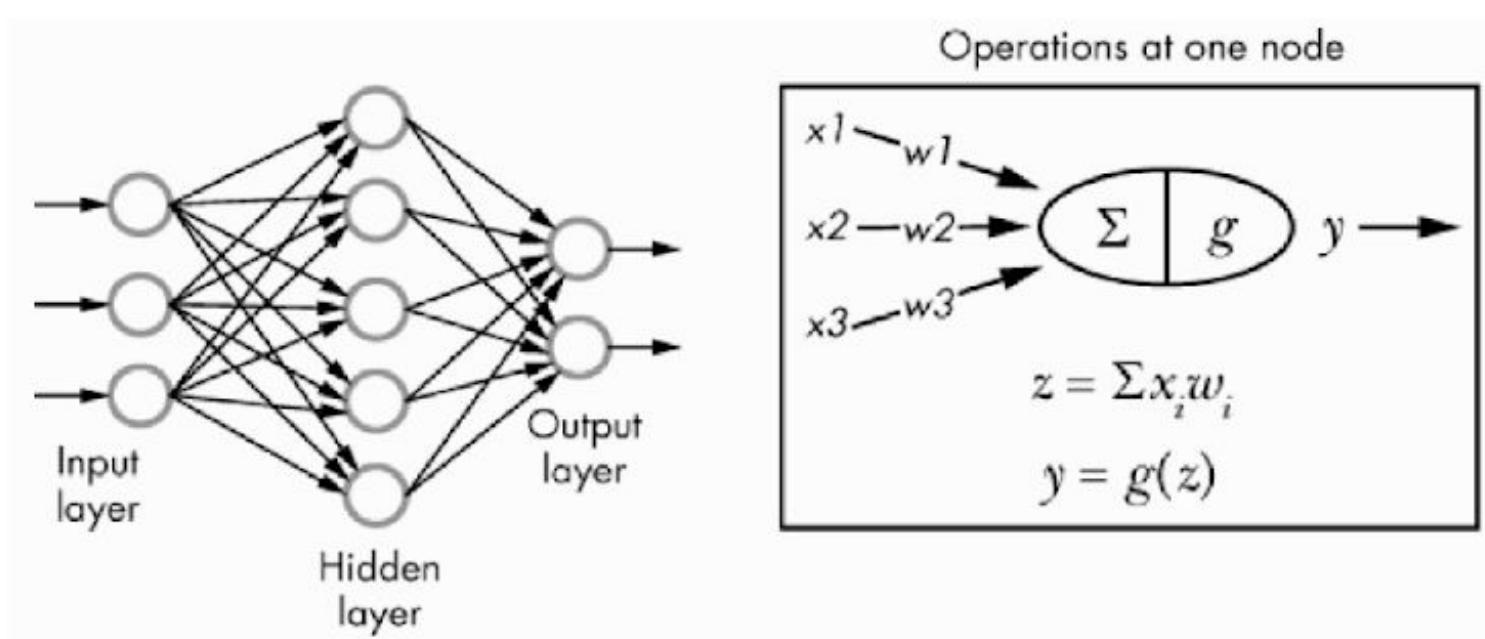
Machine Learning uses Statistical Models



Machine Learning vs Standard Programming



State-of-the-Art Neural Networks



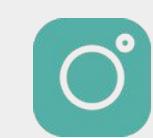
eskwelabs.com



Eskwelabs

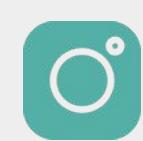
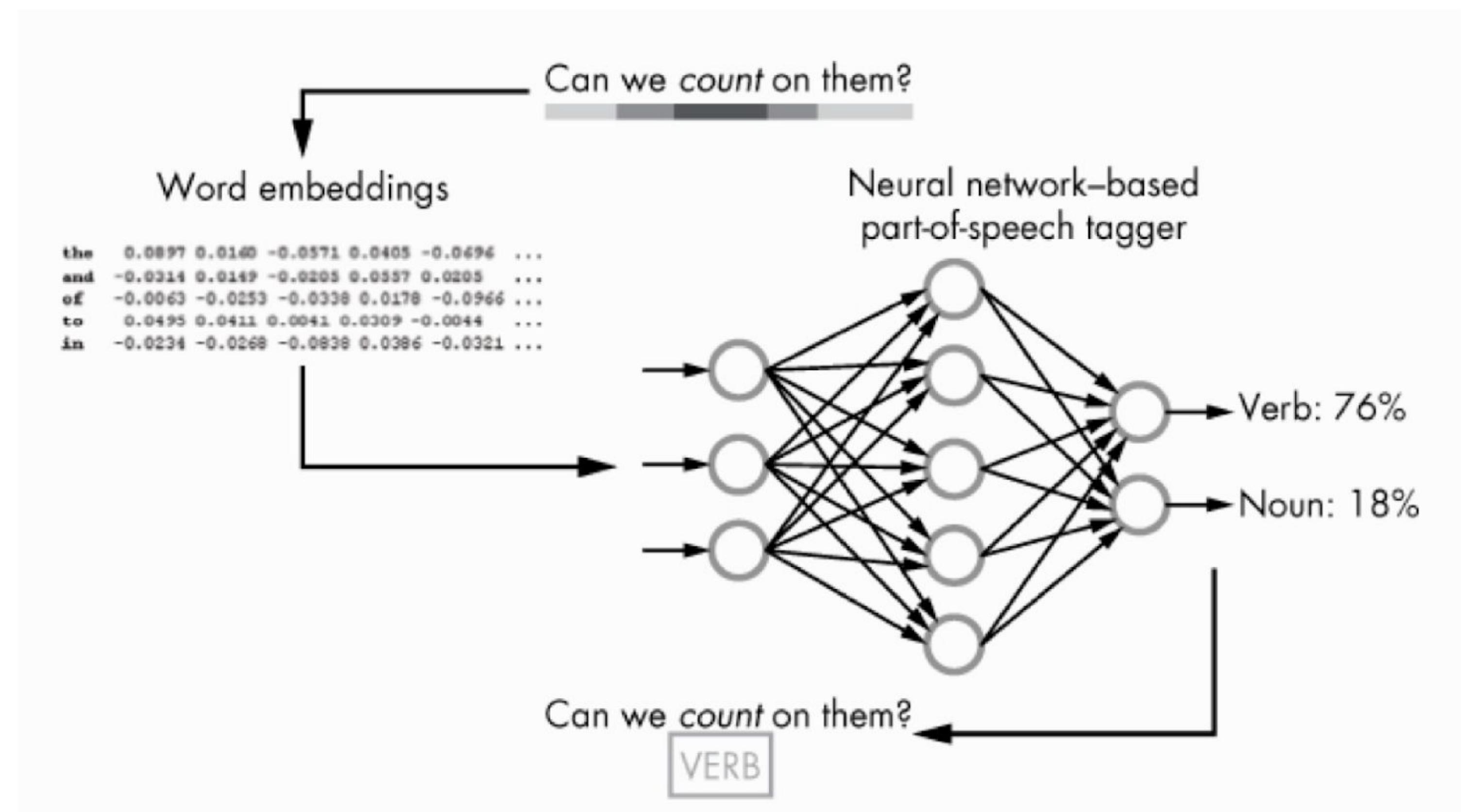
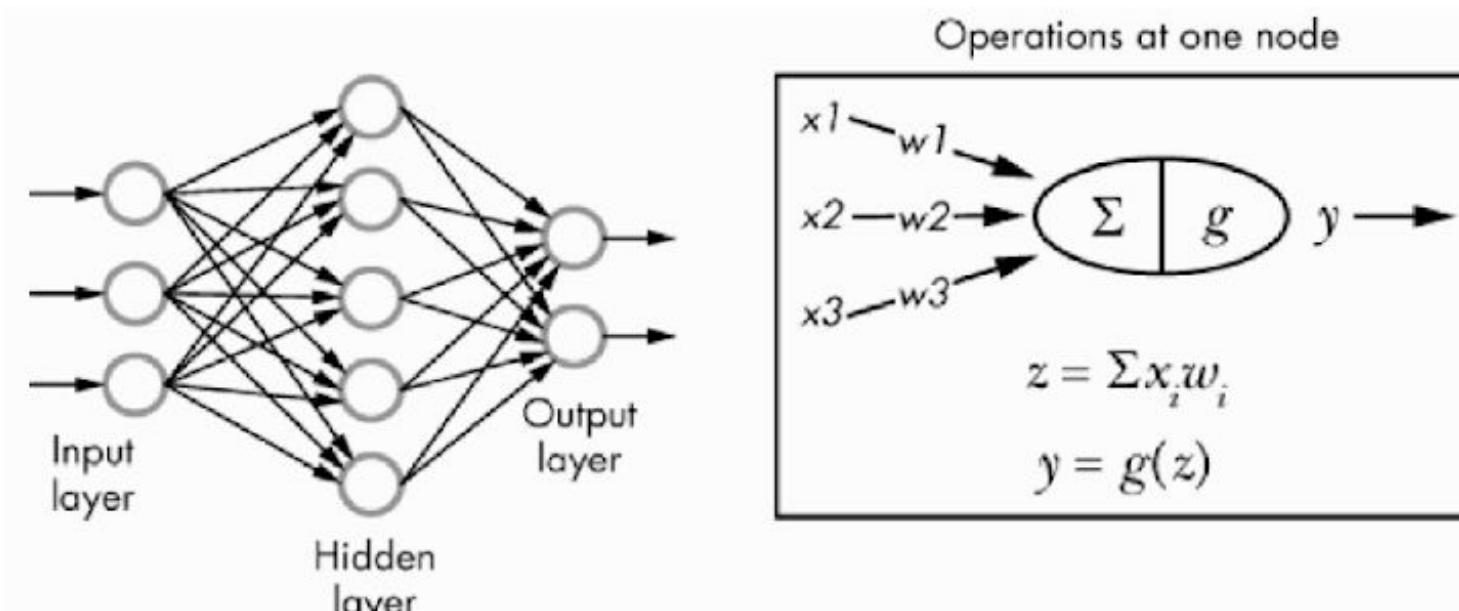


eskwelabs

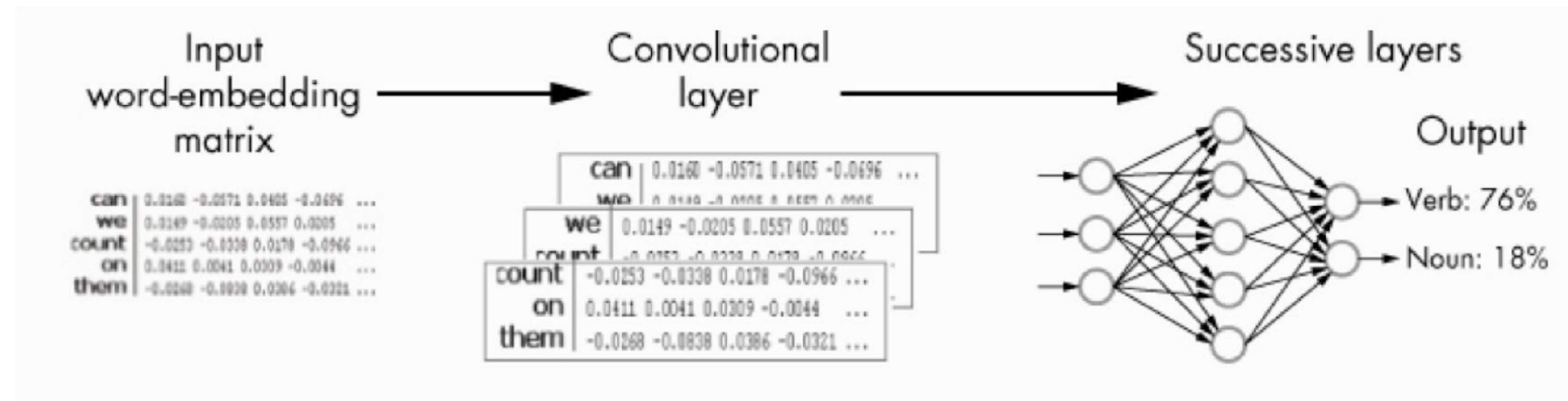


@eskwelabs_ph

State-of-the-Art Neural Networks



State-of-the-Art Neural Networks



NLP vs. Linguistics

- NLP must contend with NL data as found in the world
- NLP ≈ computational linguistics
- Linguistics has begun to use tools originating in NLP!



eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

Fields with Connections to NLP

- Machine learning
- Linguistics (including psycho-, socio-, descriptive, and theoretical)
- Cognitive science
- Information theory
- Logic
- Data science
- Political science
- Psychology
- Economics
- Education



eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

Today's Applications

- Conversational agents
- Information extraction and question answering
- Machine translation
- Opinion and sentiment analysis
- Social media analysis
- Visual understanding
- Essay evaluation
- Mining legal, medical, or scholarly literature



eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

Factors Changing NLP Landscape

1. Increases in computing power
2. The rise of the web, then the social web
3. Advances in machine learning
4. Advances in understanding of language in social context



eskwelabs.com



Eskwelabs

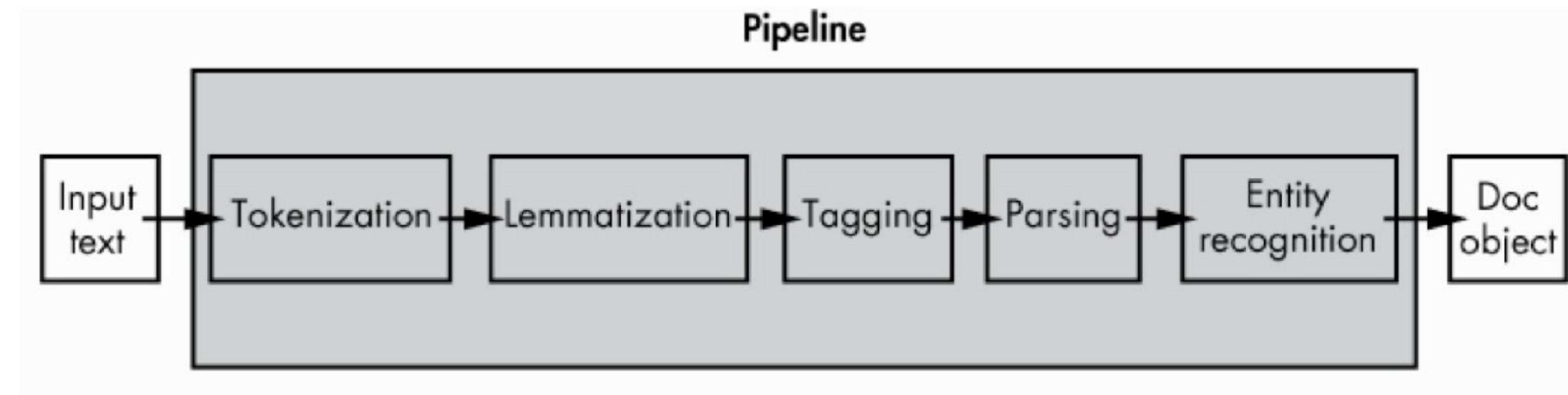


eskwelabs



@eskwelabs_ph

Basic NLP Operations with SpaCy



Code Time



eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

References



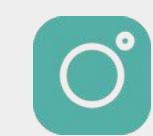
eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

<https://web.stanford.edu/~jurafsky/slp3/>

<https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf>

Natural Language Processing with Python and Spacy: Yuli Vasiliev



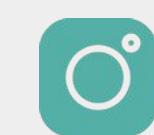
eskwelabs.com



Eskwelabs



eskwelabs



@eskwelabs_ph

THANK YOU!



ESKWE LABS



Q&A