Intro to NLP

Yoav Goldberg

Technion, 2018

About me

- Working with NLP / Text / ML for ~10 years.

- Currently: Faculty at Bar Ilan University
  NLP infrastructure, word similarity, deep learning for language

- Before: Google Research New York
  NLP infrastructure and toolchain, Machine Translation

- Before: Ben Gurion University (msc + phd)
  Hebrew NLP, syntax and parsing, structure prediction

- Before that: computer security / cyber / etc.

This Course

- Natural Language Processing (NLP):
  - Algorithms to **process**, **analyze** and **understand** texts in natural language.
    - Understanding **Structure**
    - Understanding **Meaning**
- Before you solve a problem, you need to understand it.
  - What is natural language? what are the building blocks?

This Course

- Natural Language Processing (NLP):
  - *Algorithms* to **process**, **analyze** and **understand** texts in natural language.
    - Understanding **Structure**
    - Understanding **Meaning**
- *Applications*:
  - Helping people do things with text.
- Before you solve a problem, you need to *understand* it.
  - What is natural language? what are the building blocks?

What

- How to think about working with text data.
- Understand the challenges.
- Understand the linguistic concepts.
- Understand the ML concepts.
- Understand the algorithms.
- Understand where algorithms fail, and how to asses the quality of algorithms.
- Understand what is missing.

What Not

- I assume you know basic statistics / ML.
- I will only briefly mention neural nets / deep learning.
- In many cases I will show the lower-level algorithms, without much details on how their outputs is used---unless you ask me to.

Formalities

- 3 hours a week, given by me.
- No distinction between "Lecture" and "Tirgul" sessions.
- Four programming assignments (40% of the grade)
- Final exam (rest of the grade).

Assignments

- Different from each other, covering different levels of NLP work.
- Can be done in Python or Java (I advocate for Python)
- Its good if you are familiar with Linux/Commandline.
- Assignments require also a **written report** component.

Discussion / Communication

- Discussion about assignments will be done through a Piazza forum. Don't email me with questions, **everything should go through the piazza**. piazza.com/technion.ac.il/spring2018/236299

- Communications with me regarding course, also through piazza (private messages), not via email.

Course

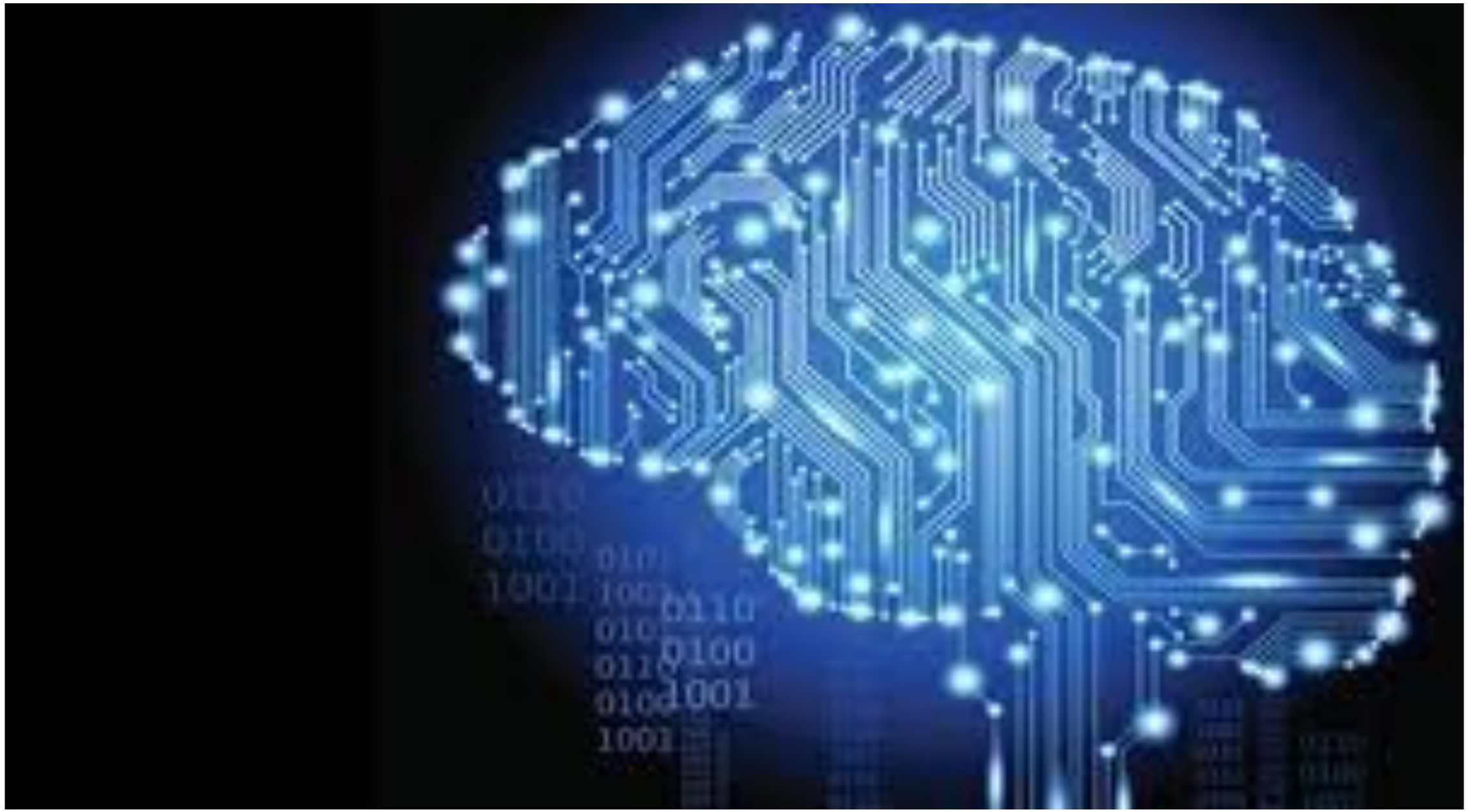
- Very similar, but not identical, to the course we teach at Bar-Ilan.

- Last semester's website: http://www.cs.biu.ac.il/~89-680/

External Materials

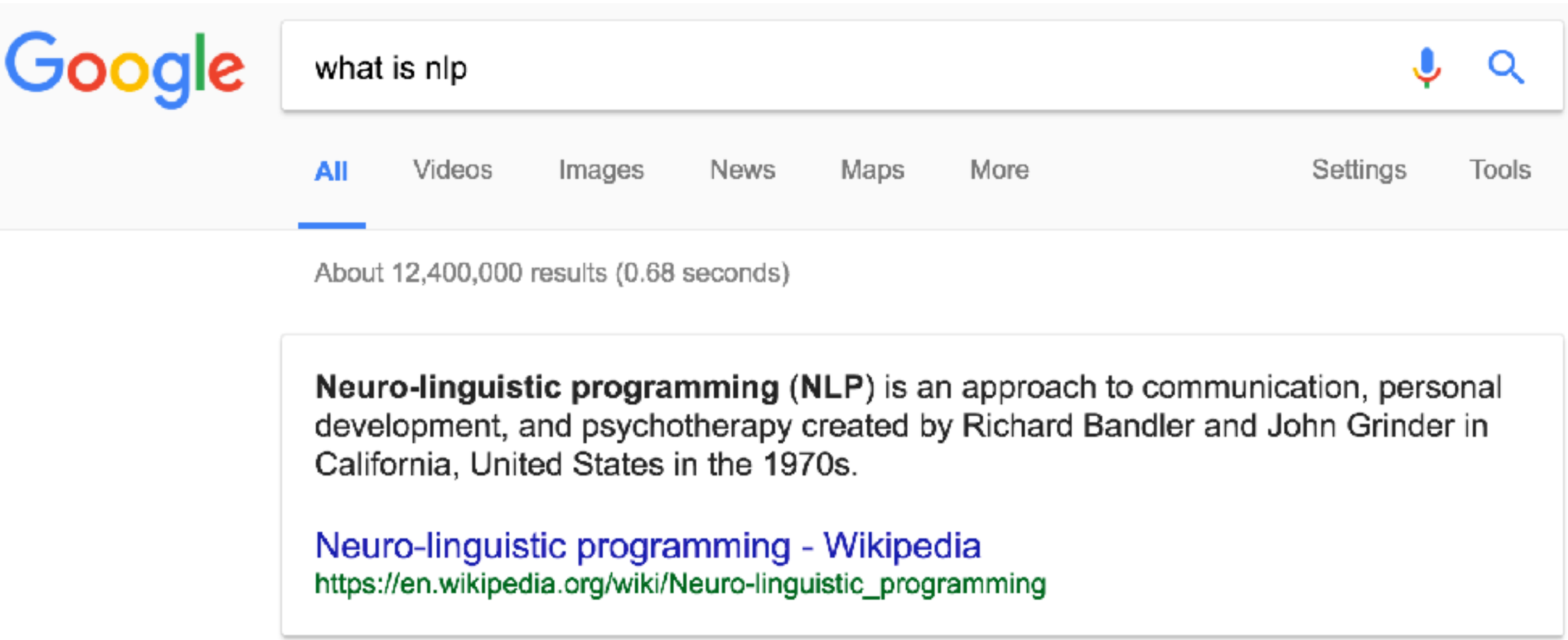
- No books covers the entire class.
- I will post links to related courses/materials on the course website.
- Each class will be accompanied by a list of related papers. **You are encouraged to read them**.

What is NLP



What is NLP

NLP is Everywhere



NLP is Everywhere

Google

what is natural language processing

All Videos Images News Maps More Settings Tools

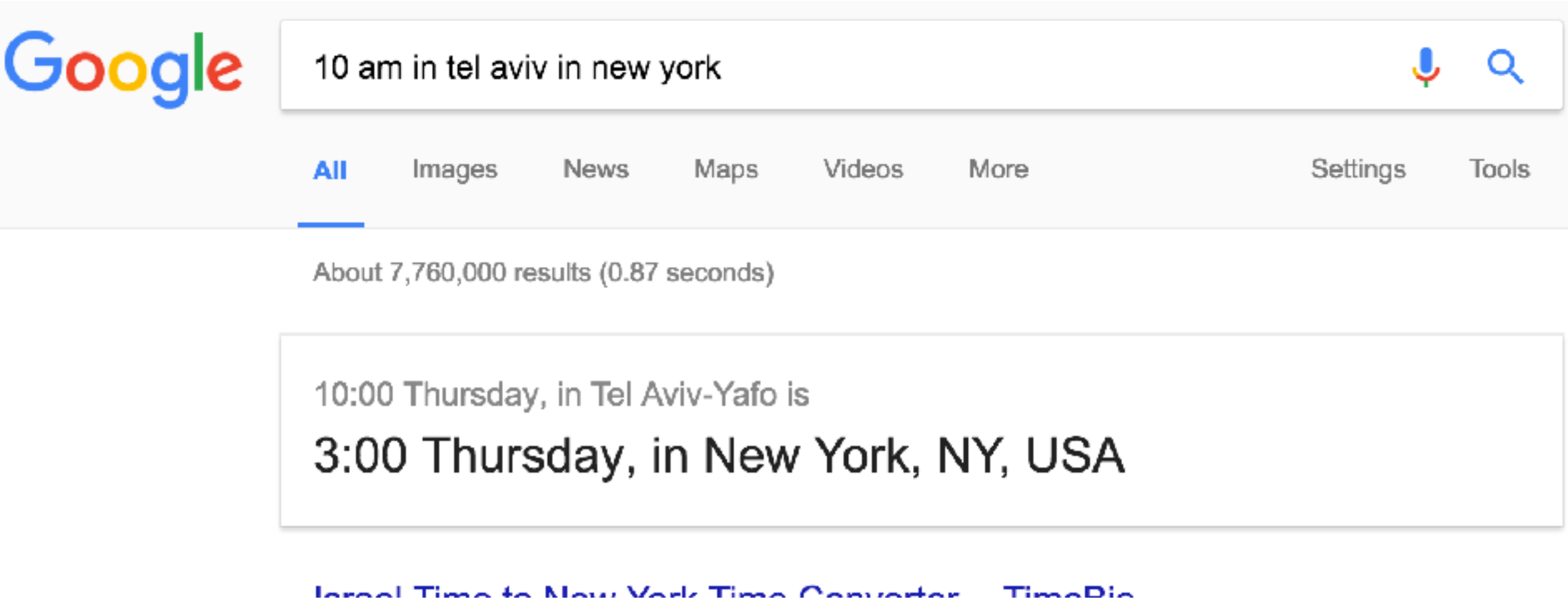
About 8,640,000 results (0.69 seconds)

Natural language processing is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (**natural**) languages. As such, **NLP** is related to the area of human–computer interaction.

Natural language processing - Wikipedia
https://en.wikipedia.org/wiki/**Natural_language_processing**

About this result • Feedback

NLP is Everywhere



NLP is Everywhere

NLP is Everywhere

[redacted] to me

Did you send it before or after our conversation earlier?

Reply

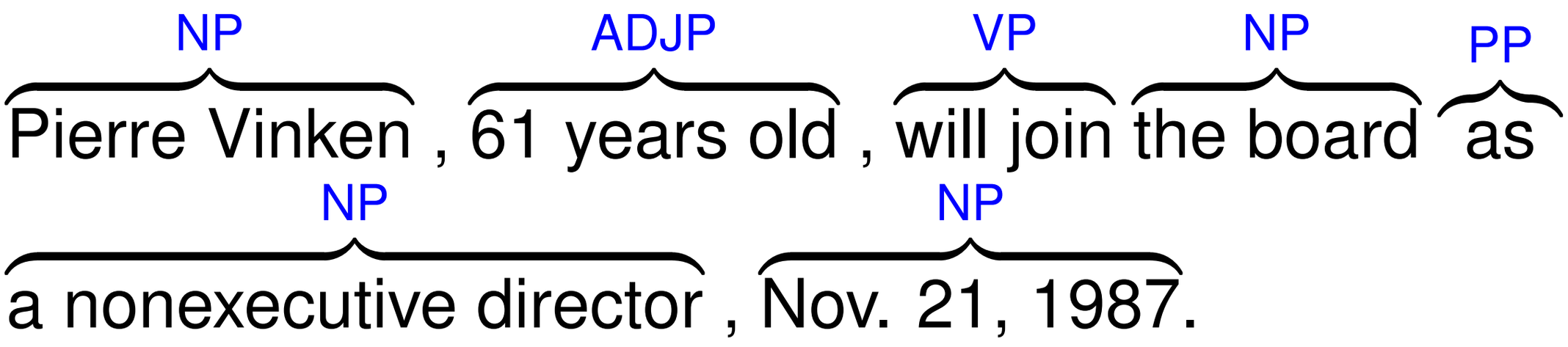
Yes, I sent it. | No, I didn't. | Yes, I did.

What is NLP

What is NLP

Pierre Vinken , 61 years old , will join the board as a nonexecutive director , Nov. 21, 1987

⇓ chunking



What is NLP

Named Entity Recognition (NER)

Paris Whitney Hilton (born February 17, 1981) is an American television personality and businesswoman . She is the great-granddaughter of Conrad Hilton , the founder of Hilton Hotels . Born in New York City and raised in both California and New York, Hilton began a modeling career when she signed with Donald Trump's modeling agency .

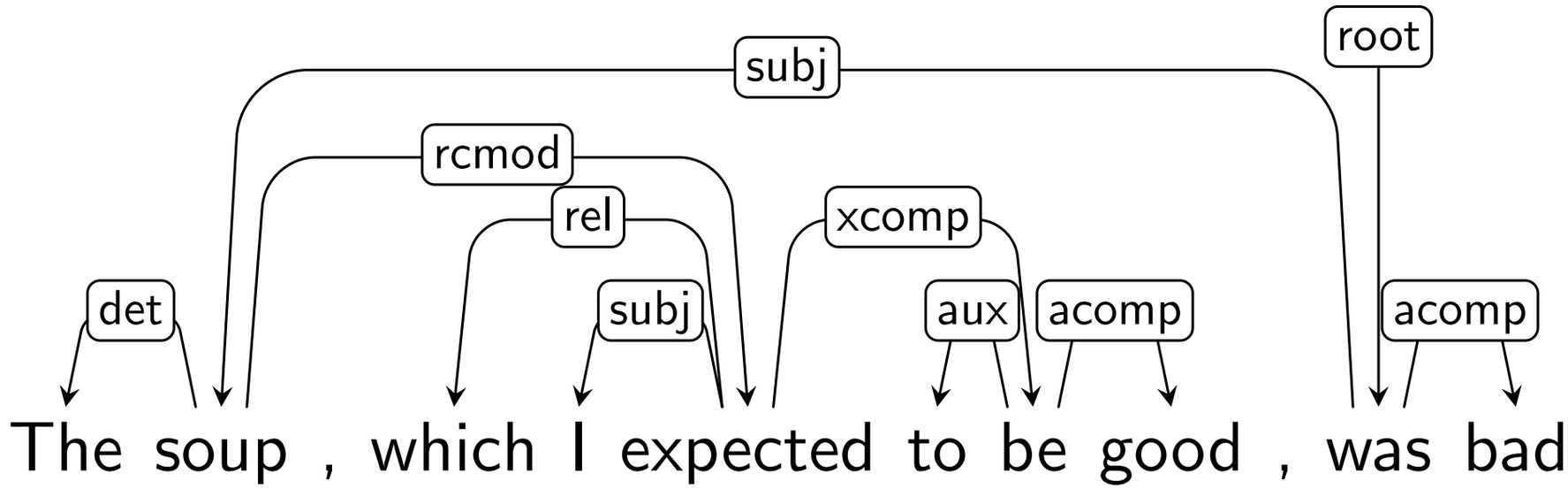
PER: Paris Whitney Hilton
TIME: February 17, 1981
PER: Conrad Hilton
ORG: Hilton Hotels
LOC: New York City
LOC: California
LOC: New York
PER: Hilton
ORG: Donald Trump's modeling agency

What is NLP

sentiment analysis

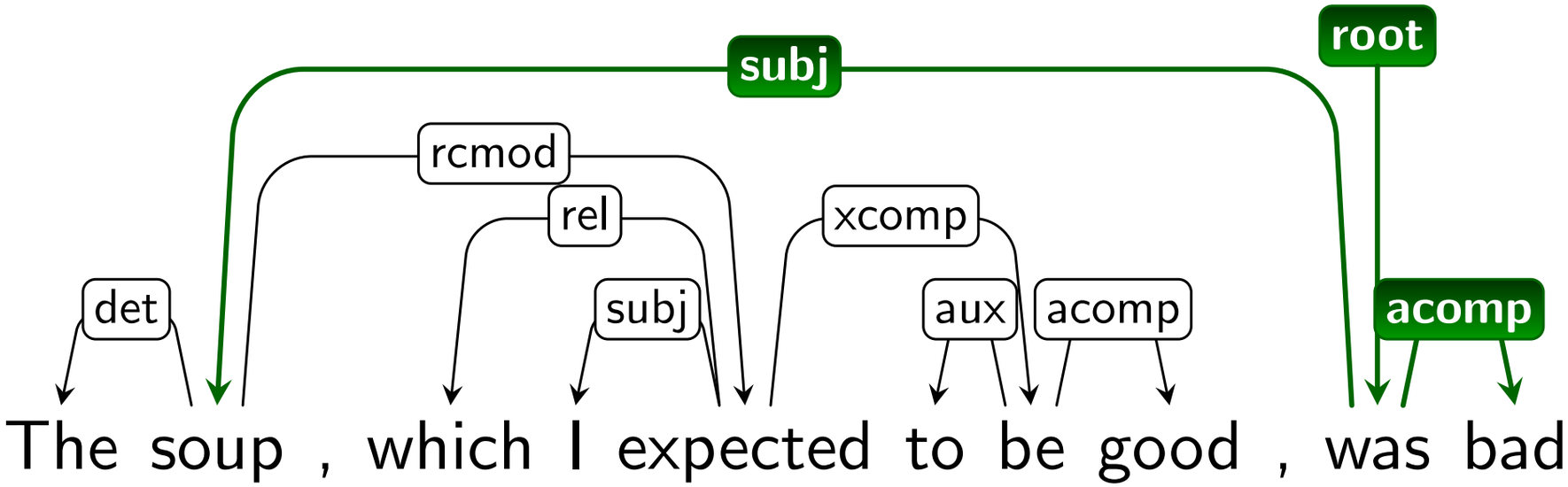
The soup , which I expected to be good , was bad

What is NLP



parsing

Structure of Sentences



parsing helps sentiment

What is NLP

October 14, 2002, 4:00 a.m. PT

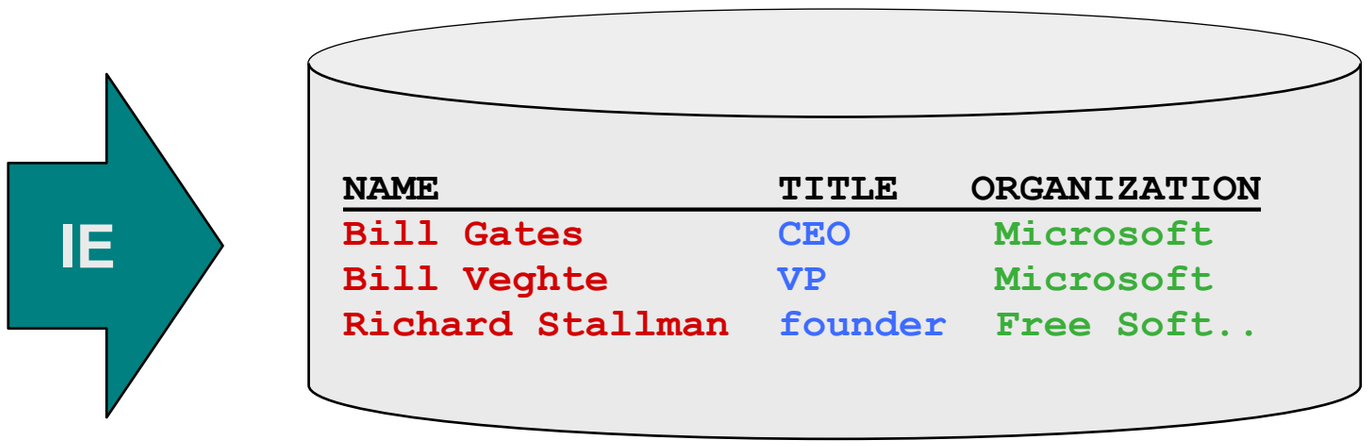
For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

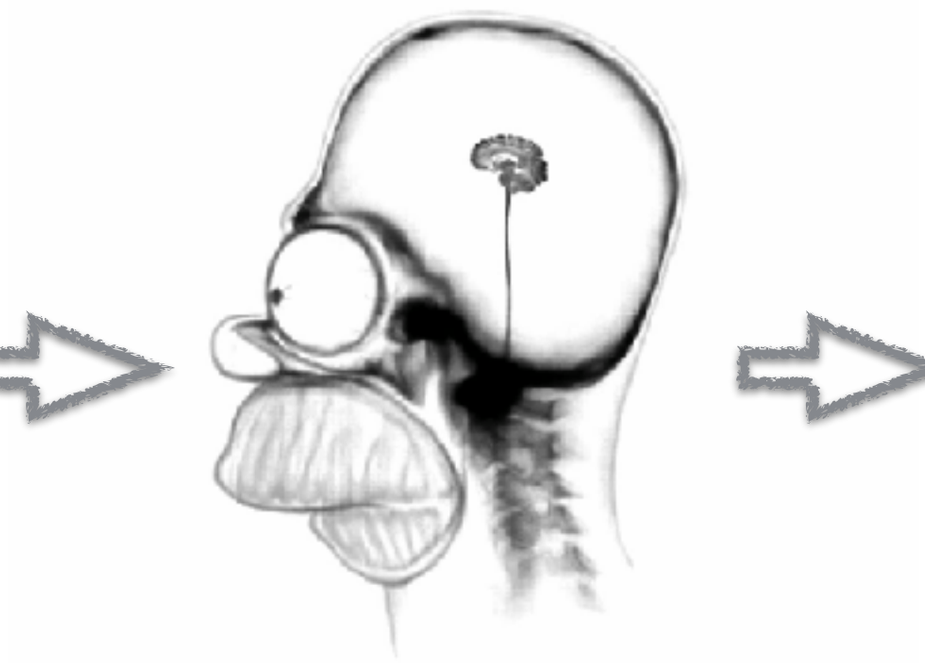
Richard Stallman, founder of the Free Software Foundation, countered saying…

information extraction



NLP In a Nutshell

human language

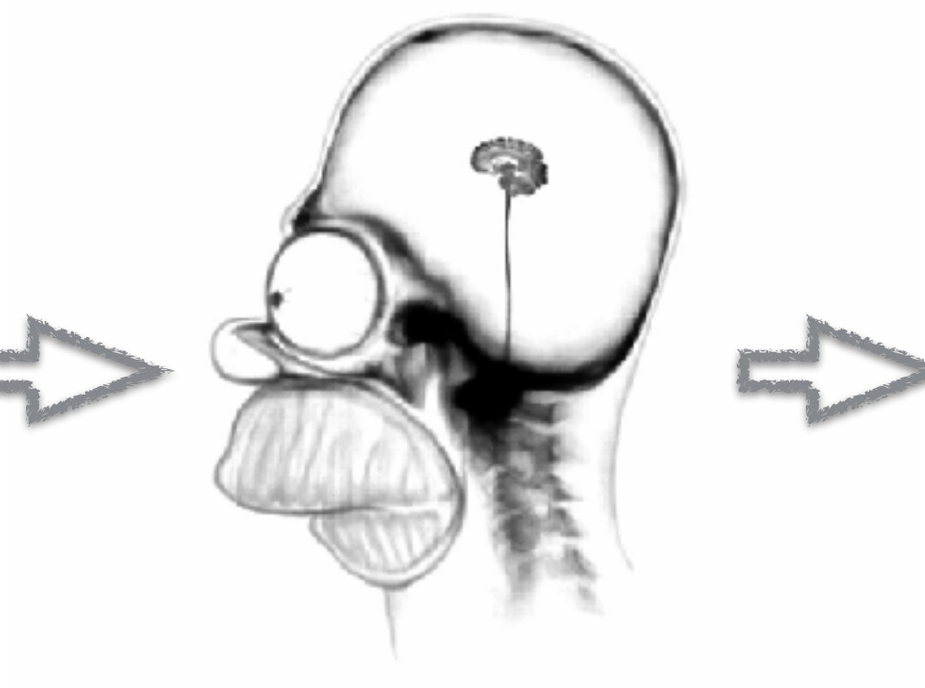


non-trivial useful output

takes as input text in human language
and process it in a way that suggests
an intelligent process was involved

What is NLP

Text in
human language

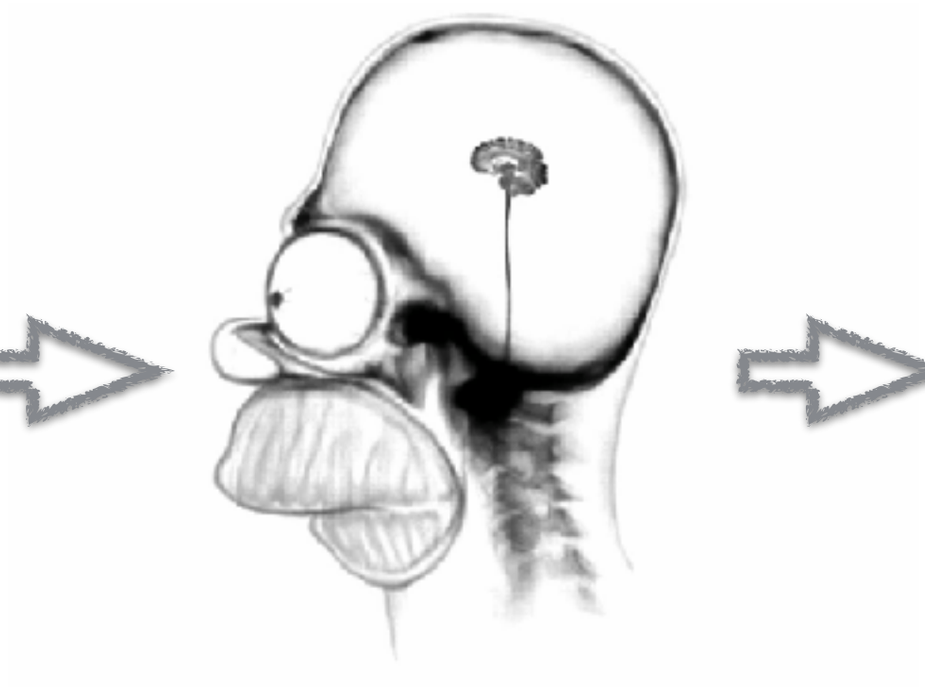


"Meaning"

"insights"

What is NLP

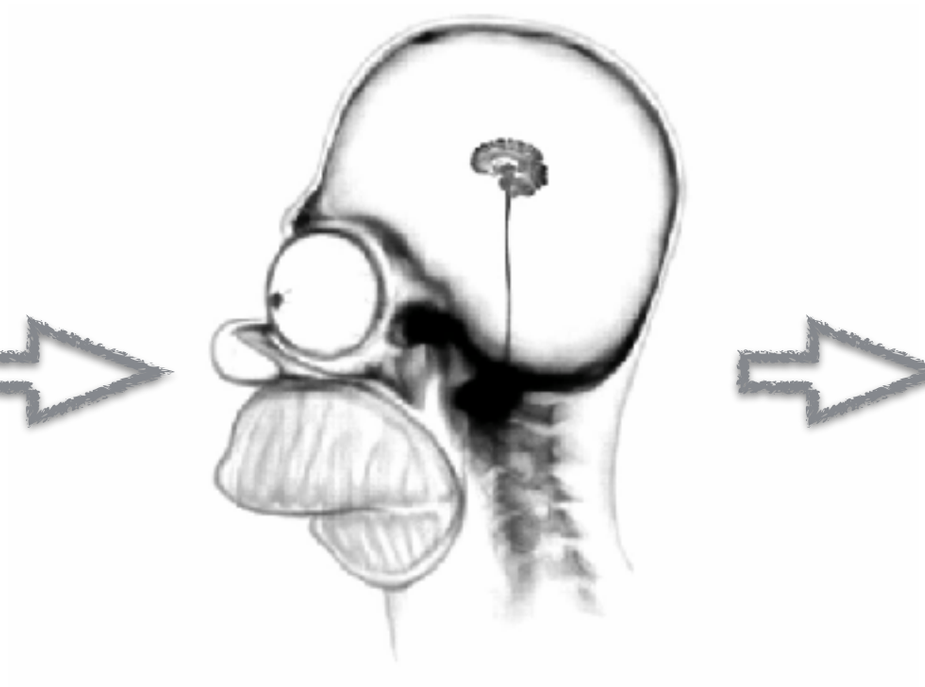
Text in
human language



Text in another
language

What is NLP

structured data
in some format



Text in
human language

(some) NLP Applications

- Search
- Knowledge Discovery
- Translation
- Question Answering
- Ads / market analysis
- Chatbots / personal assistants
- Others? (discuss)

Interdisciplinary Field

- NLP Combines:
  - AI
  - Linguistics
  - Machine Learning / statistics
  - Algorithms (dynamic programming, search)

Related Fields

- Computational Linguistics
    - Using algorithms to understand language.
- Psycholinguistics / Cognitive Science
    - How is language processed by the human brain / by humans?
- Social Linguistics
    - How do humans use and shape language as a social construct?

How do we do NLP?

- 1950 -- ~1990s ---> Write many rules
- 1990s -- ~2000s ---> Corpus based statistics
- 2000s -- ~2014 ---> Supervised machine learning
- 2014 -- today ---> "deep learning"

How do we do NLP?

- 1950 -- ~1990s ---> Write many rules
- 1990s -- ~2000s ---> Corpus based statistics
- 2000s -- ~2014 ---> Supervised machine learning
- 2014 -- today ---> "deep learning"

2021+ --> write rules, aided by machine learning.

How do we do NLP?

- 1950 -- ~1990s ---> Write many rules
- 1990s -- ~2000s ---> Corpus based statistics
- 2000s -- ~2014 --->**Supervised machine learning**
- 2014 -- today ---> "deep learning"

2021+ --> write rules, aided by machine learning.

"Empirical Revolution"

"Whenever I fire a linguists, performance goes up"

(Fred Jelinek, ~1988)

Of course, we must not go overboard and mistakenly conclude that the successes of statistical NLP render linguistics irrelevant (rash statements to this effect have been made in the past, e.g., the notorious remark, “Every time I fire a linguist, my performance goes up”). **The information and insight that linguists, psychologists, and others have gathered about language is invaluable in creating high performance broad domain language understanding systems;** for instance, in the speech recognition setting described above, better understanding of language structure can lead to better language models.” (Lilian Lee, 2001)

Why is language Interesting?

- Invented by humans for communications.
- Learned from experience, in a transparent manner.
- Preserves ideas across space and time.
- Symbolic system.
    - ...also for abstract stuff.
- Language facilitates thinking?
- Tons of human knowledge encoded in language.

Levels of Language Analysis

- Phonetics (how sounds are made: lips, spit, throat)
- Phonology (how sounds can combine)
- Morphology (how words are built)
- Syntax (how words are combined)
- Semantics (the meaning of words/phrases)
- Pragmatics (the true meaning of words/phrases)
- Discourse (structure/meaning across sentences)

Levels of Language Analysis

- Phonetics (how sounds are made: lips, spit, throat)
- Phonology (how sounds can combine)
- Morphology (how words are built) בתיהם
- Syntax (how words are combined)
- Semantics (the meaning of words/phrases)
- Pragmatics (the true meaning of words/phrases)
- Discourse (structure/meaning across sentences)

Levels of Language Analysis

- Phonetics (how sounds are made: lips, spit, throat)
- Phonology (how sounds can combine)
- Morphology (how words are built) בתיהם
- Syntax (how words are combined)
- Semantics (the meaning of words/phrases) [[[מצאתי [מטבע [על הרצפה]
- Pragmatics (the true meaning of words/phrases)
- Discourse (structure/meaning across sentences)

Levels of Language Analysis

- Phonetics (how sounds are made: lips, spit, throat)
- Phonology (how sounds can combine)
- Morphology (how words are built) בתיהם
- Syntax (how words are combined)
[מצאתי [מטבע [על הרצפה]]]
- Semantics (the meaning of words/phrases)
"יש לך שקל?" -> האם יש ברשותך שקל
- Pragmatics (the true meaning of words/phrases)
- Discourse (structure/meaning across sentences)

Levels of Language Analysis

- Phonetics (how sounds are made: lips, spit, throat)
- Phonology (how sounds can combine)
- Morphology (how words are built) בתיהם
- Syntax (how words are combined) [[[מצאתי [מטבע [על הרצפה]
- Semantics (the meaning of words/phrases) "יש לך שקל?" -> האם יש ברשותך שקל
- Pragmatics (the true meaning of words/phrases) "יש לך שקל?" -> תן לי שקל / רוצה שקל?
- Discourse (structure/meaning across sentences)

Levels of Language Analysis

- Phonetics (how sounds are made: lips, spit, throat)
- Phonology (how sounds can combine)
- Morphology (how words are built) בתיהם
- Syntax (how words are combined) [[[מצאתי [מטבע [על הרצפה]]]
- Semantics (the meaning of words/phrases) "יש לך שקל?" -> האם יש ברשותך שקל
- Pragmatics (the true meaning of words/phrases) "יש לך שקל?" -> תן לי שקל / רוצה שקל?
- Discourse (structure/meaning across sentences)

בגלל, ברם, ואכן...

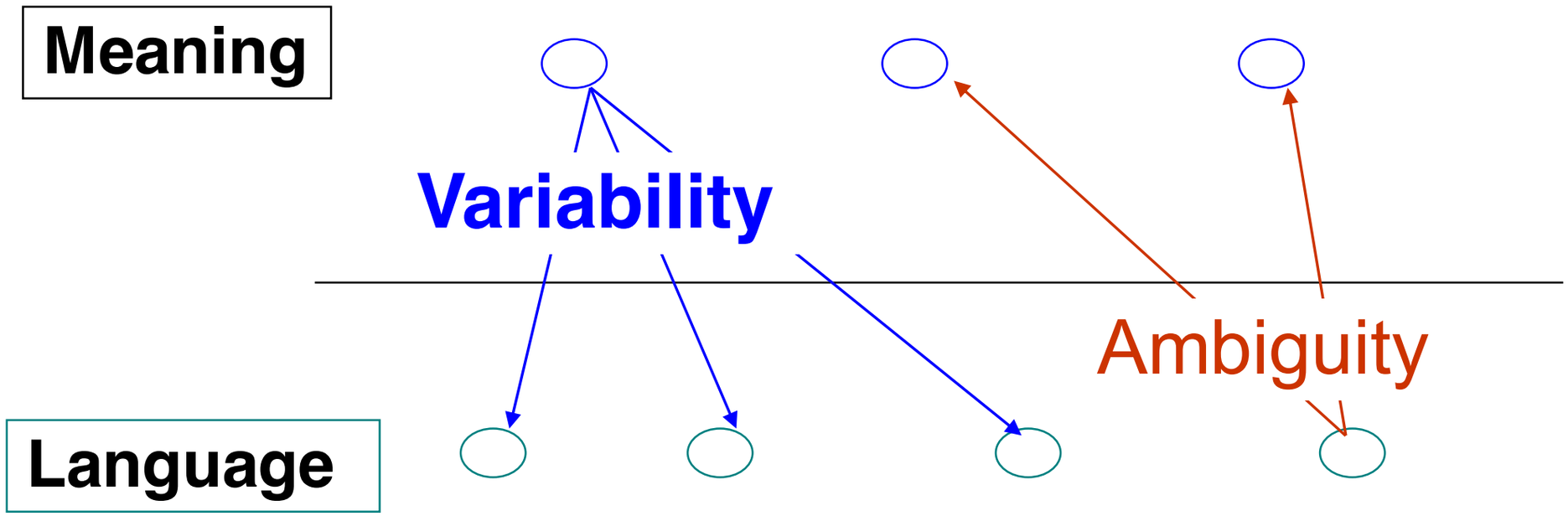
Levels of Language Analysis

- Phonetics (how sounds are made: lips, spit, throat)
- Phonology (how sounds can combine)
- Morphology (how words are built) בתיהם
- Syntax (how words are combined) [[[מצאתי [מטבע [על הרצפה]]]
- Semantics (the meaning of words/phrases) "יש לך שקל?" -> האם יש ברשותך שקל
- Pragmatics (the true meaning of words/phrases) "יש לך שקל?" -> תן לי שקל / רוצה שקל?
- Discourse (structure/meaning across sentences) בגלל, ברם, ואכן... **הוא** אמר ש..

Why is Language Difficult

- Variability
- Ambiguity
- Creativity
- Coverage
- Generalization

Variability and Ambiguity



by Dan Roth

Variability

he acquired it

he purchased it

he bought it

it was bought by him

it was sold to him

she sold it to him

she sold him that

Variability

2m 2ma 2mar 2mara 2maro 2marrow 2mor 2mora 2moro 2morow 2morr 2morro 2morrow 2moz 2mr 2mro 2mrrw 2mrw 2mw tmmrw tmo tmoro tmorrow tmoz tmr tmro tmrow tmrrow tmrrw tmrw tmrww tmw tomaro tomarow tomarro tomarrow tomm tommarow tommarrow tommoro tommorow tommorrow tommorw tommrow tomo tomolo tomoro tomorow tomorro tomorrw tomoz tomrw tomz

Ambiguity

bank

apple

star

Ambiguity

book

book

Ambiguity

book a flight

read a book

Ambiguity

ספר

Ambiguity

הרכבת

Ambiguity

the book was interesting

the movie was interesting

Ambiguity

ספר עזר לרופא שיניים בהוצאת כתר

Ambiguity

שמעתי ש
ספר עזר לרופא שיניים בהוצאת כתר
קניתי

Ambiguity

I ate pizza with olives

I ate pizza with friends

Ambiguity

I saw her duck

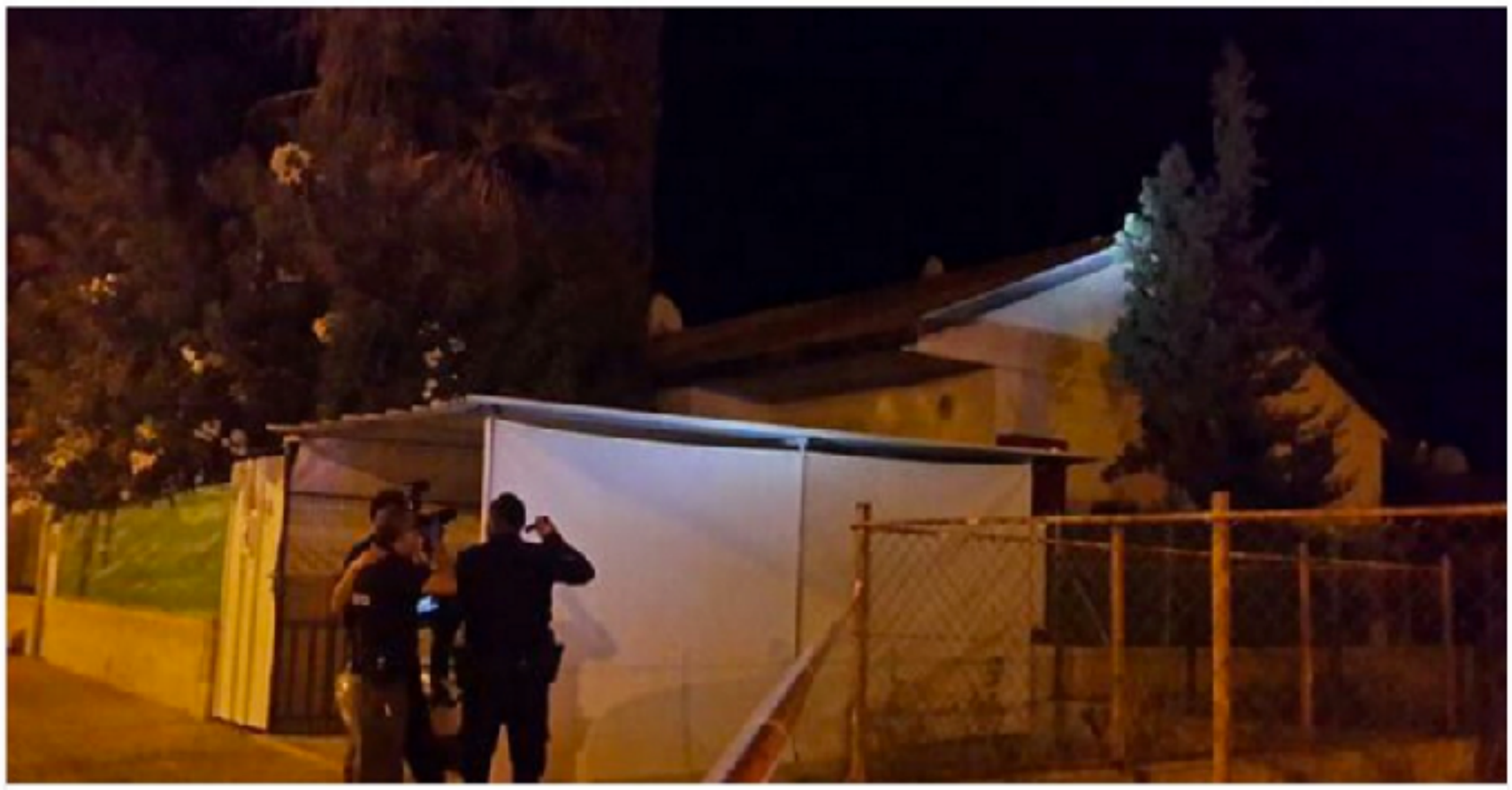
Ambiguity

- Self control is waiting for the movie to start to eat popcorn.

Ambiguity

נעצר עורך דין שניצל ופיתה שופטת וגנב

Ambiguity



וטרינר רצה להחרים טווס - וירה באישה עם מגרפה

ynet.co.il

Ambiguity

- **News headlines**
  - Ban on Nude Dancing on Governor's Desk
  - Teacher Strikes Idle Kids
  - Hospitals Are Sued by 7 Foot Doctors
  - Iraqi Head Seeks Arms
  - Kids Make Nutritious Snacks
  - Local Highschool Dropouts Cut in Half

Ambiguity

חולצה מטיילת בהודו

Ambiguity

fat people eat candy

Ambiguity

fat people eat accumulates

Ambiguity

Levels of Language Analysis

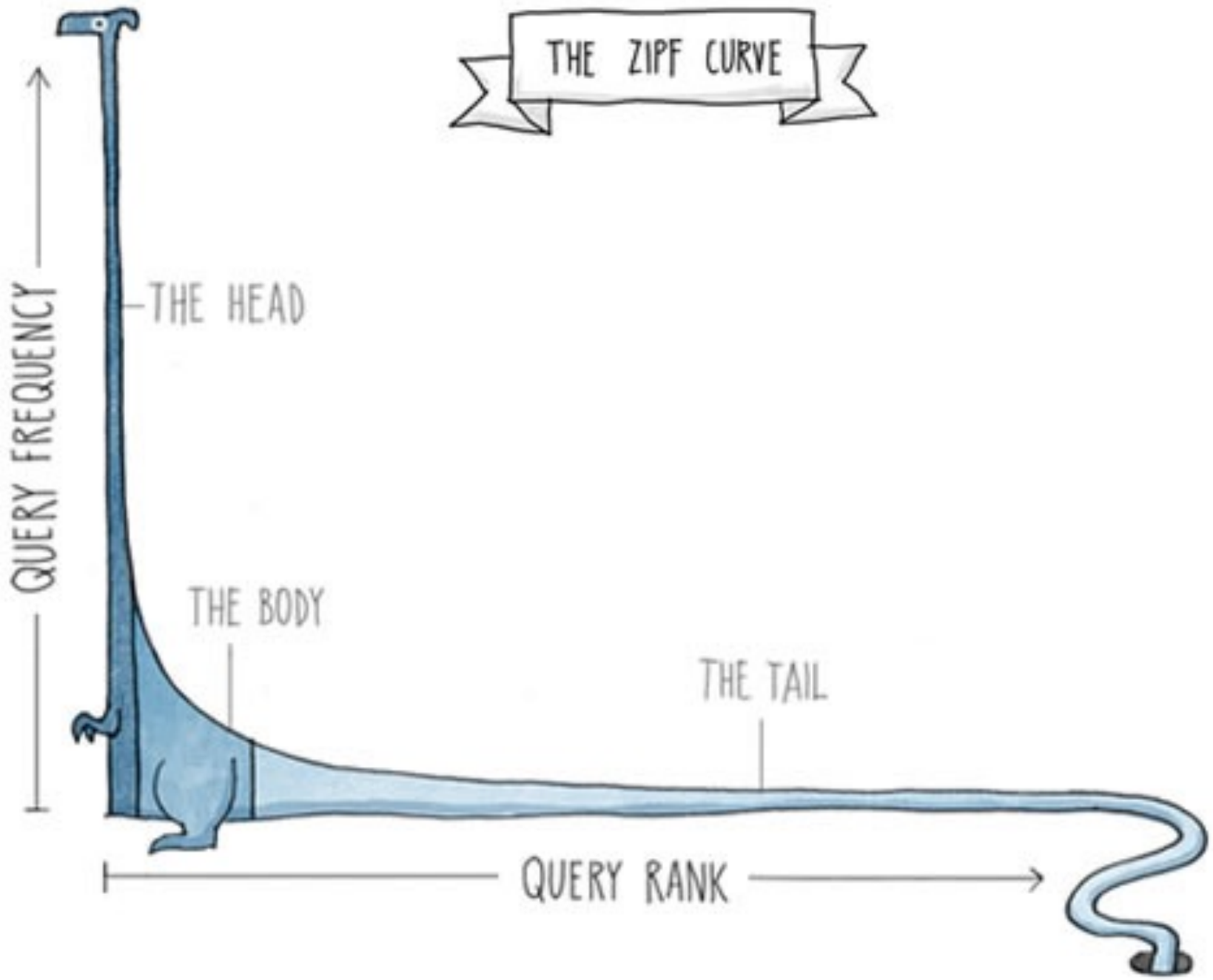
- Phonetics (how sounds are made: lips, spit, throat)
- Phonology (how sounds can combine)
- Morphology (how words are built)
- Syntax (how words are combined)
- Semantics (the meaning of words/phrases)
- Pragmatics (the true meaning of words/phrases)
- Discourse (structure/meaning across sentences)

There is ambiguity on all levels.
(can you think of examples?)

Coverage

- **Zipf law**
  - "the frequency of any word is inversely proportional to its rank in the frequency table"
  - Word frequencies follow a power-law distribution.
  - --> Long tail
  - --> Most "events" will occur only few times, if any.

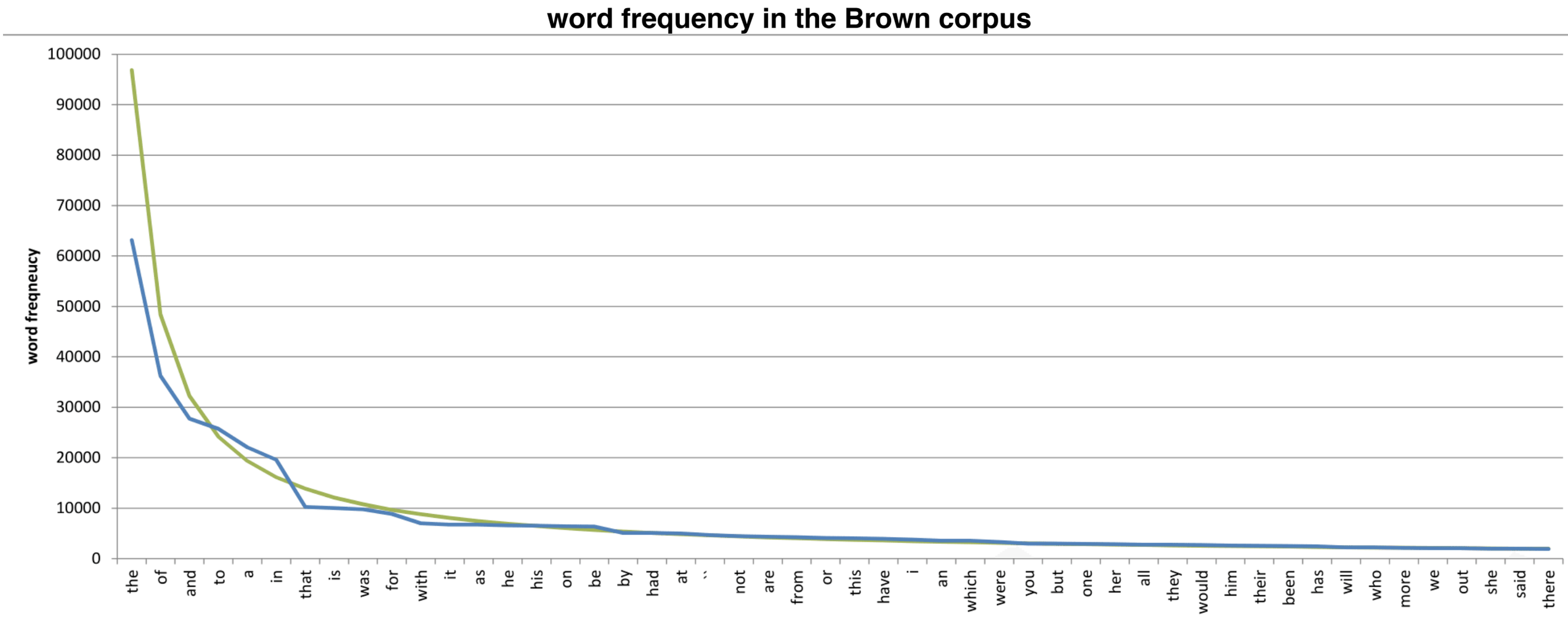
Coverage



Source: smashing magazine

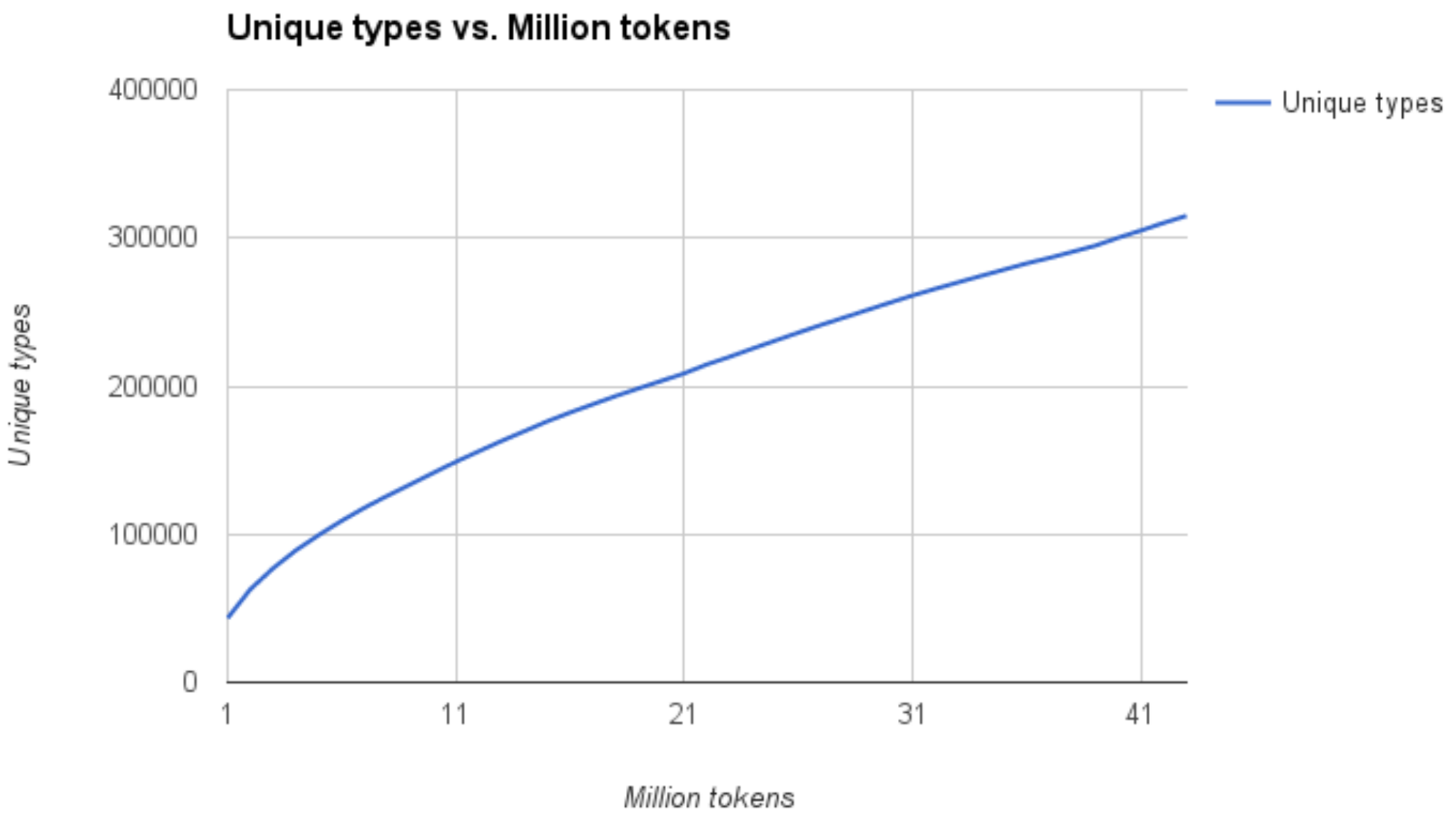
Coverage

- frequency of word is inversely proportional to its rank in the frequency table



Source: http://www.shadycharacters.co.uk/

Coverage


Unique types vs. Million tokens
400000
300000
200000
100000
0
Unique types
1
11
21
31
41
Million tokens
Unique types

Coverage

in a **43M** words text, there are:

- 316,710 unique words
- 144,999 words occur only once
- 42,525 words occur 2 times
- 21,618 words occur 3 times
- 13,306 words occur 4 times
- 9,488 words occur 5 times

- 26,024 words appear >50 times

Coverage

in a **43M** words text, there are:

- 316,710 unique words
- **144,999** words occur **only once**
- 42,525 words occur 2 times
- 21,618 words occur 3 times
- 13,306 words occur 4 times
- 9,488 words occur 5 times

- 26,024 words appear >50 times

how may words occur 0 times?

Coverage

Some 1-count words:

- poetically
- unregulation
- PORTLAND
- timbers
- wildlands
- philologist
- Metallurgic
- 8,553
- baby-killer
- sofia-based

Some 5-count words:

- hologram
- 500.7
- Ben-Gurion
- yahoo
- Nineteenth
- profitabilty
- ballpoint
- interstate-highway
- nested
- algorithm

Coverage

Some 10-count words:

- pivot
- sensitively
- provable
- cathode-ray
- social-democratic
- encyclical
- anti-military
- marinated
- transparently

Some 50-count words:

- gadgets
- treasurers
- narrator
- 349
- Syrians
- Kearns
- drummer
- trophy
- vanishing

Coverage

- No matter how large the training corpus is:
  - there are very likely to be word forms we did not see
  - there will be many word-pairs we didn't see
  - there will be even more word-triplets we didn't see
  - there will be even more sentences we didn't see

Word Clustering

mailman	10000011010111
salesman	10000011011 0000
bookkeeper	10000011011 00010
troubleshooter	10000011011 000110
bouncer	10000011011 000111
technician	10000011011 00100
janitor	10000011011 00101
saleswoman	10000011011 00110
...	
Nike	10110111001001011 00100
Maytag	10110111001001011 10010
Generali	10110111001001011 10011
Gap	10110111001001011 10100
Harley-Davidson	10110111001001011 10101
Enfield	10110111001001011 10110
genus	10110111001001011 10111
Microsoft	10110111001001011 11000
Ventritex	10110111001001011 11001
Tractebel	10110111001001011 11010
Synopsys	10110111001001011 11011
WordPerfect	10110111001001011 11100
....	
John	10111001000000000
Consuelo	10111001000000001
Jeffrey	10111001000000010
Kenneth	10111001000000011 00
Phillip	10111001000000011 010
WILLIAM	10111001000000011 011
Timothy	10111001000000011 10

Creativity

ENTERTAINMENT

Lindsay Lohan Just Went Full Regina George On Ariana Grande's Selfie

Raise your hand if you've ever felt personally victimized by Lindsay Lohan.

11/19/2016 01:27 pm ET | **Updated** 4 days ago

Generalization

- Discrete events --> hard to generalize to unseen events.

Properties of Text Data

- Discrete symbols
- Sequences
- Hierarchies

Properties of Text Data

- **Discrete symbols**
- Sequences
- Hierarchies

- many unseen events
- hard to generalize
- hard to use math
- sparse representations

Properties of Text Data

- Discrete symbols
- **Sequences**
  - order matters
  - sub sequences matter
- Hierarchies

Properties of Text Data

- Discrete symbols
- Sequences
- **Hierarchies**

document --> paragraphs --> sentences
--> phrases --> words --> morphemes

Properties of Text Data

- Discrete symbols
- Sequences
- **Hierarchies**

- which level to look at?
- sequences are not enough

document --> paragraphs --> sentences
--> phrases --> words --> morphemes

Kinds of NLP Applications

- document --> label
- sentence --> label
- sentence(s) --> sentence(s)
- many documents / many sentences (aggregate)
  --> Information/relations (Information extraction 1)
- document / sentence --> information/relations (Information extraction 2)
- sentence --> API command (semantic parsing)
- clustering

document --> label

- language classification
- topic classification
- author classification
- sentiment classification
- interestingness
- relevance

sentence --> label

- mostly the same as in "document --> label" but on a more granular level.
    - (but: shorter text --> harder task)

sentence(s) --> sentence(s)

- translation
- question answering
- simplification
- summarization
- query completion
- email auto-response generation

many documents / sentences (aggregate) --> Information/relations

- Information extraction (1)
- "Table filling"

"bill gates", "founder of", "microsoft"

"bethoven", "born in", "bonn"

What is “Information Extraction”

As a task: **Filling slots in a database from sub-segments of text.**

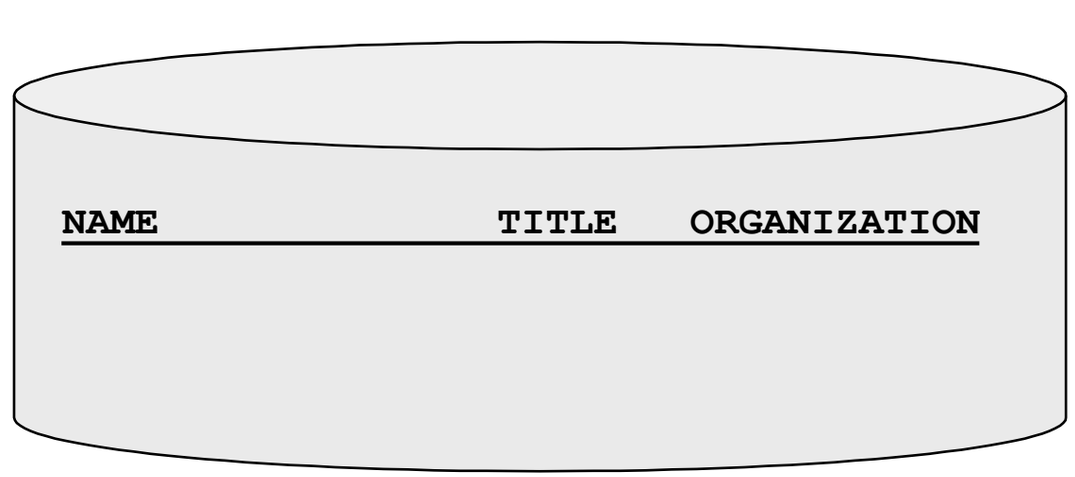
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…



What is “Information Extraction”

As a task: **Filling slots in a database from sub-segments of text.**

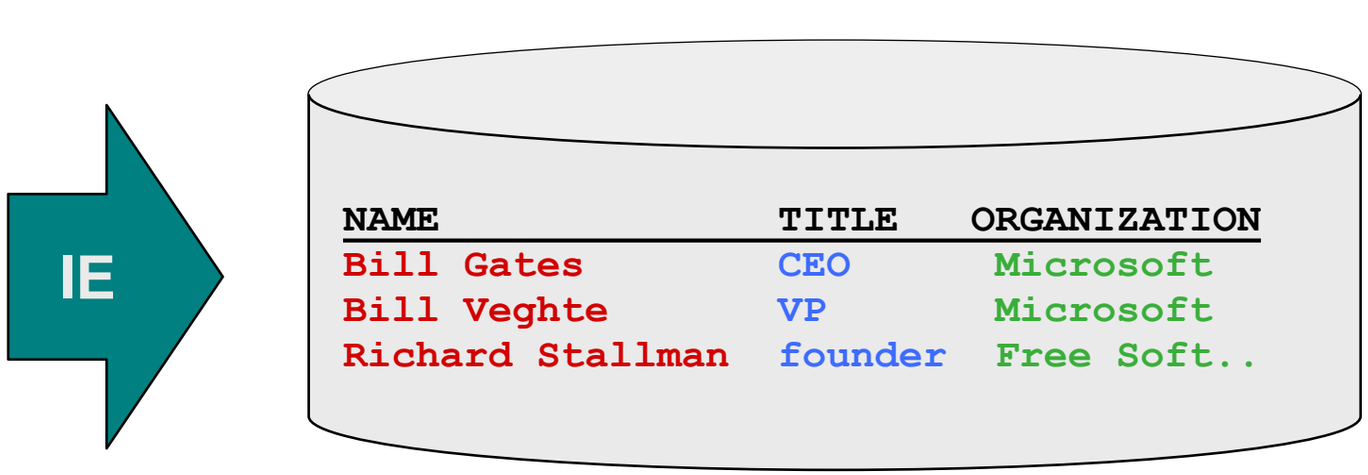
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…



What is “Information Extraction”

As a family of techniques:

Information Extraction =
segmentation + classification + clustering + association

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

aka “named entity extraction”

What is “Information Extraction”

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

What is “Information Extraction”

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates

Microsoft
Gates
Microsoft

Bill Veghte
Microsoft
VP

Richard Stallman
founder
Free Software Foundation

What is “Information Extraction”

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

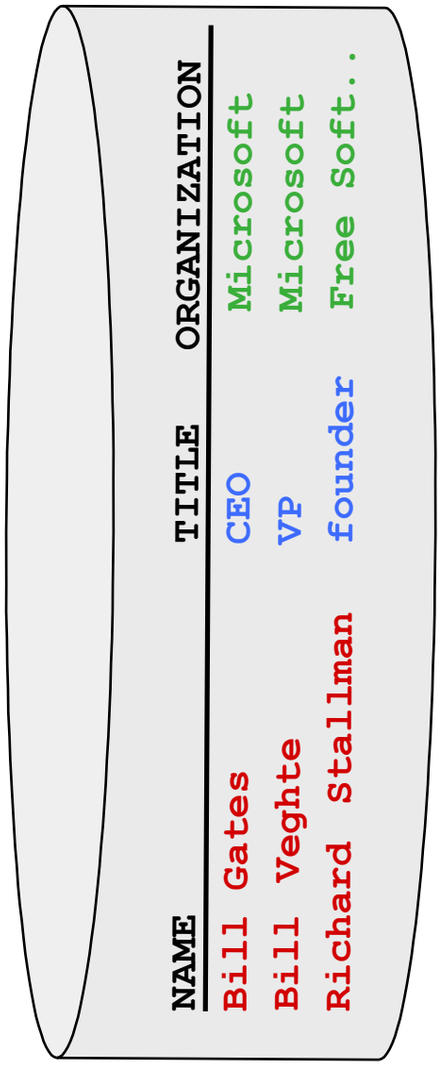
Richard Stallman, founder of the Free Software Foundation, countered saying…

* Microsoft Corporation
CEO
Bill Gates

* Microsoft
Gates
Microsoft

* Bill Veghte
Microsoft
* VP

Richard Stallman
founder
Free Software Foundation



document / sentence --> information/relations

- Information extraction (2)
  - similar to IE1, but harder (cannot rely on redundancy)
  - "facts about the text" / "what is this about"
  - Google's "calendar events from email" feature.

Sentence --> API command (semantic parsing)

- "busses from tel-aviv to petach-tikva on thursday"
- "show me images of dogs"
- "add a meeting with mark on tuesday afternoon"
- "send sms to wife buy milk"
- "when was bethoven born"

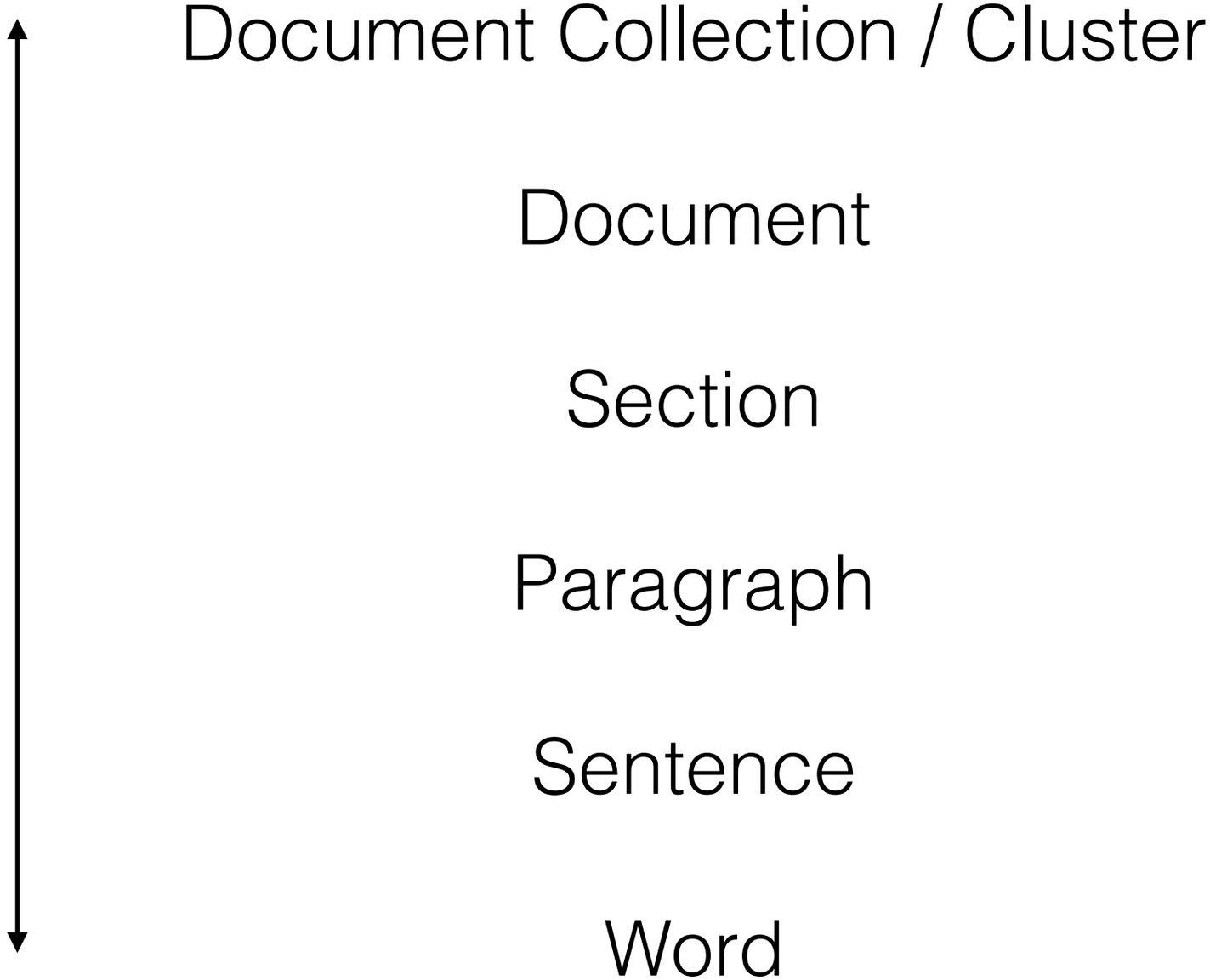
Clustering

- Find similar texts

- Summarize collections of similar texts

the basic units of text processing

What are we processing?



What are we processing?

Document Collection / Cluster

Document

Section

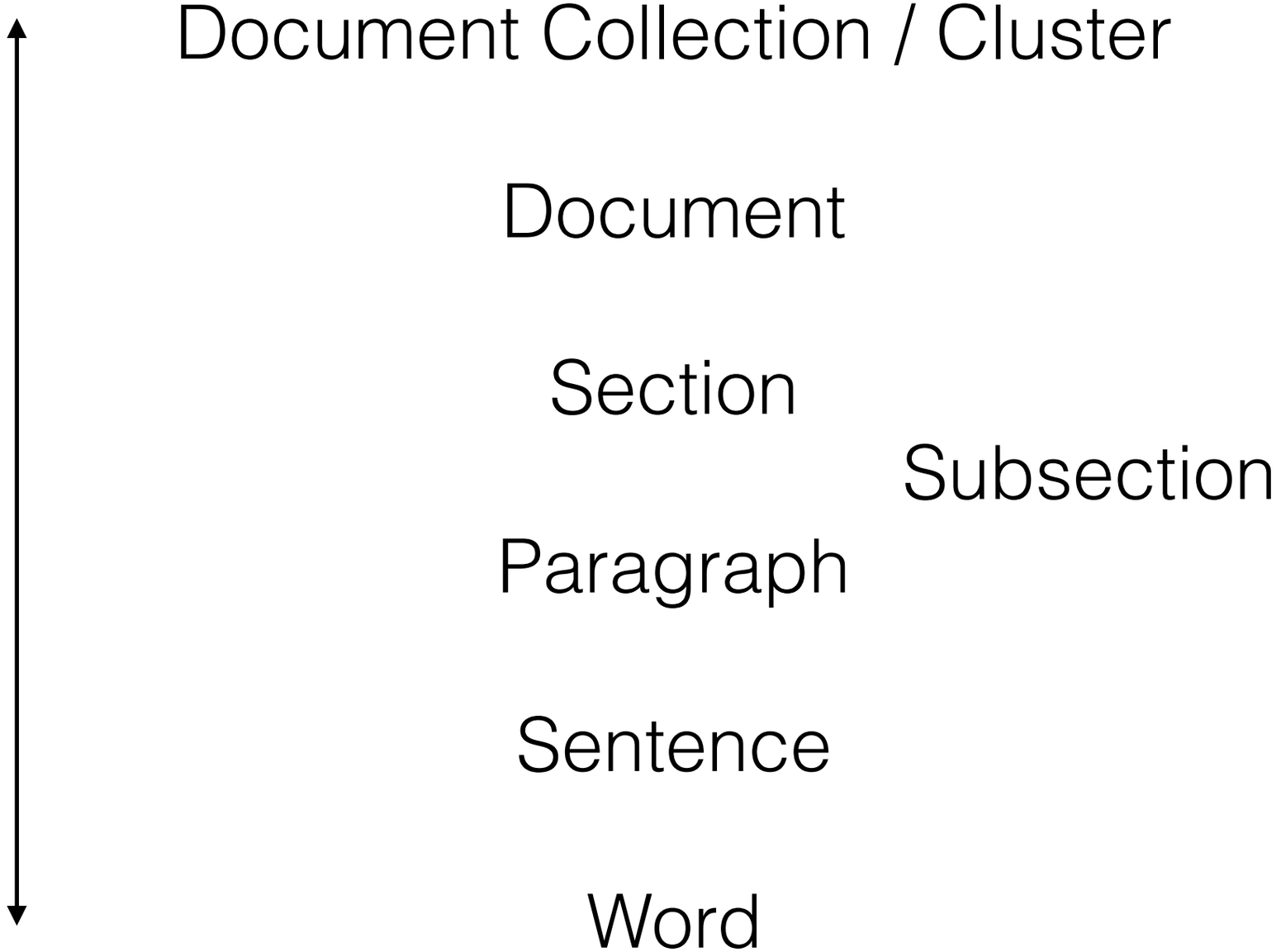
Paragraph

Sentence

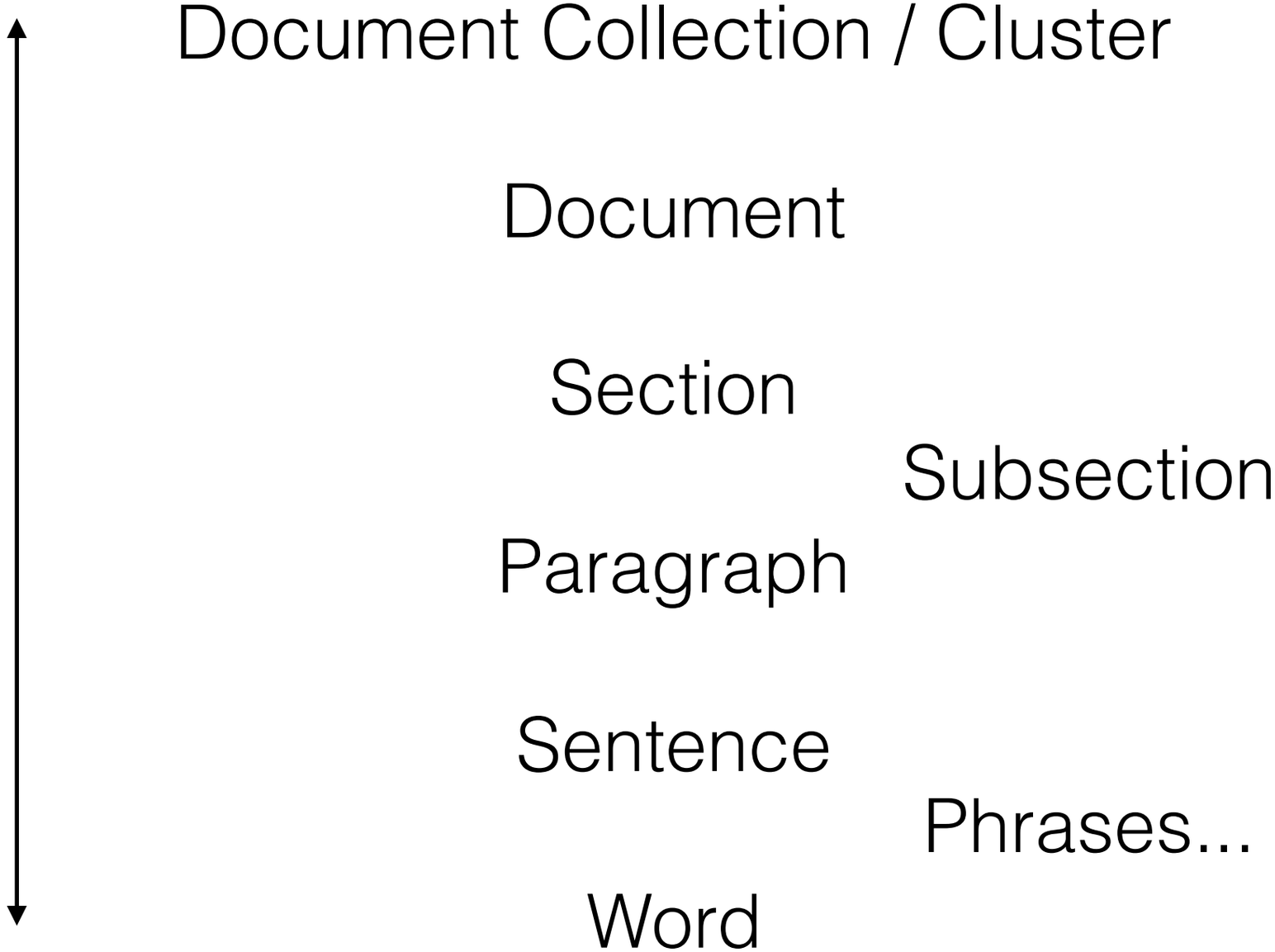
Word

by date
by topic
by web domain
...

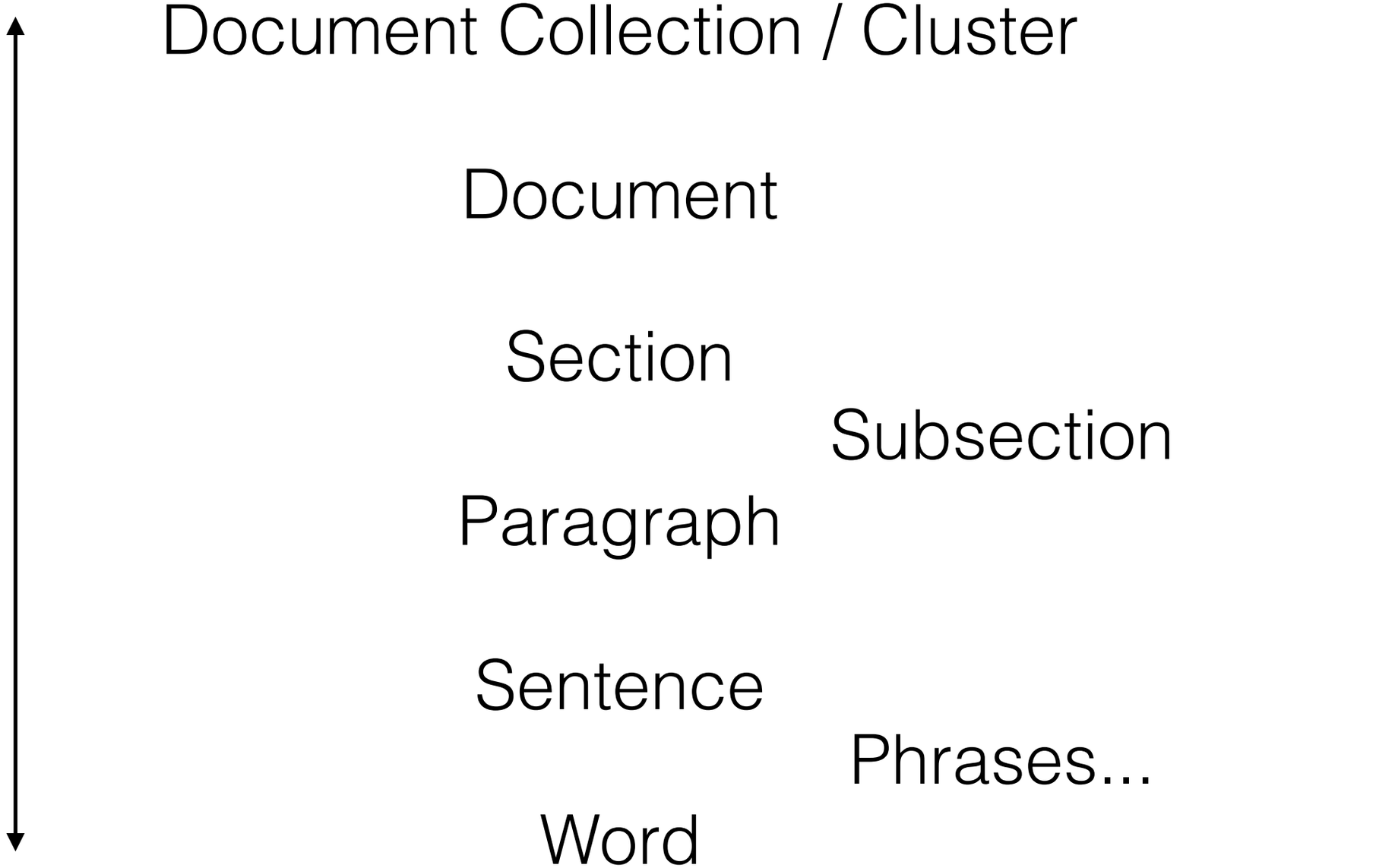
What are we processing?



What are we processing?



What are we processing?



What are we processing?

Document Collection / Cluster

Document

Section

Paragraph

Sentence

Word

What is a document?

edition.cnn.com/2016/06/21/opinions/brexit-campaign-disaffection-wismayer...

CNN Regions + International Edition + menu

British lawmakers clash ahead of Brexit vote 05:57

Story highlights

Britons will vote Thursday on whether their country should remain in the European Union

Henry Wismayer: Urge to leave has caught the wind of native disaffection

Editor's Note: *Henry Wismayer is a writer and commentator based in London. You can follow him @henrywismayer. The views expressed are his own.*

(CNN) — For any external observer curious to understand the issues shaping the debate over Britain's membership of the European Union, last Wednesday's edition of The Sun, the country's most popular daily newspaper, produced a revealing metaphor.

A day after the Rupert Murdoch-owned tabloid declared its support for the "Leave" campaign, its front-page went full-hysteria. A "biblical swarm" of super-moths has invaded Britain from mainland Europe, it exclaimed without a trace of irony, "and is set to annihilate our cabbages."

The prominence given to these ghastly foreign moths, accompanied on Page 2 with a helpful diagram depicting a battalion of Lepidoptera hovering over the English Channel, tells you all you need to know about Britain's EU referendum debate, where fear and frustration have come to overshadow the biggest political decision in a generation.

Top stories

Rio thief: How Olympic tourists can protect phone

Chicago's deadly Father's Day weekend: 12 murders, 54 shootings

edition.cnn.com/2016/06/21/opinions/brexit-campaign-disaffection-wismayer...

CNN Regions + International Edition + menu

British lawmakers clash ahead of Brexit vote 05:57

Story highlights

Britons will vote Thursday on whether their country should remain in the European Union

Henry Wismayer: Urge to leave has caught the wind of native disaffection

Editor's Note: *Henry Wismayer is a writer and commentator based in London. You can follow him @henrywismayer. The views expressed are his own.*

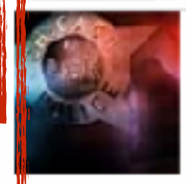
(CNN) — For any external observer curious to understand the issues shaping the debate over Britain's membership of the European Union, last Wednesday's edition of The Sun, the country's most popular daily newspaper, produced a revealing metaphor.

A day after the Rupert Murdoch-owned tabloid declared its support for the "Leave" campaign, its front-page went full-hysteria. A "biblical swarm" of super-moths has invaded Britain from mainland Europe, it exclaimed without a trace of irony, "and is set to annihilate our cabbages."

The prominence given to these ghastly foreign moths, accompanied on Page 2 with a helpful diagram depicting a battalion of Lepidoptera hovering over the English Channel, tells you all you need to know about Britain's EU referendum debate, where fear and frustration have come to overshadow the biggest political decision in a generation.

Top stories

Rio thief: How Olympic tourists can protect phone



Chicago's deadly Father's Day weekend: 12 murders, 54 shootings

edition.cnn.com/2016/06/21/opinions/brexit-campaign-disaffection-wismayer...

CNN Regions + International Edition + menu

British lawmakers clash ahead of Brexit vote 05:57

Story highlights

Britons will vote Thursday on whether their country should remain in the European Union

Henry Wismayer: Urge to leave has caught the wind of native disaffection

Editor's Note: *Henry Wismayer is a writer and commentator based in London. You can follow him @henrywismayer. The views expressed are his own.*

(CNN) — For any external observer curious to understand the issues shaping the debate over Britain's membership of the European Union, last Wednesday's edition of The Sun, the country's most popular daily newspaper, produced a revealing metaphor.

A day after the Rupert Murdoch-owned tabloid declared its support for the "Leave" campaign, its front-page went full-hysteria. A "biblical swarm" of super-moths has invaded Britain from mainland Europe, it exclaimed without a trace of irony, "and is set to annihilate our cabbages."

The prominence given to these ghastly foreign moths, accompanied on Page 2 with a helpful diagram depicting a battalion of Lepidoptera hovering over the English Channel, tells you all you need to know about Britain's EU referendum debate, where fear and frustration have come to overshadow the biggest political decision in a generation.

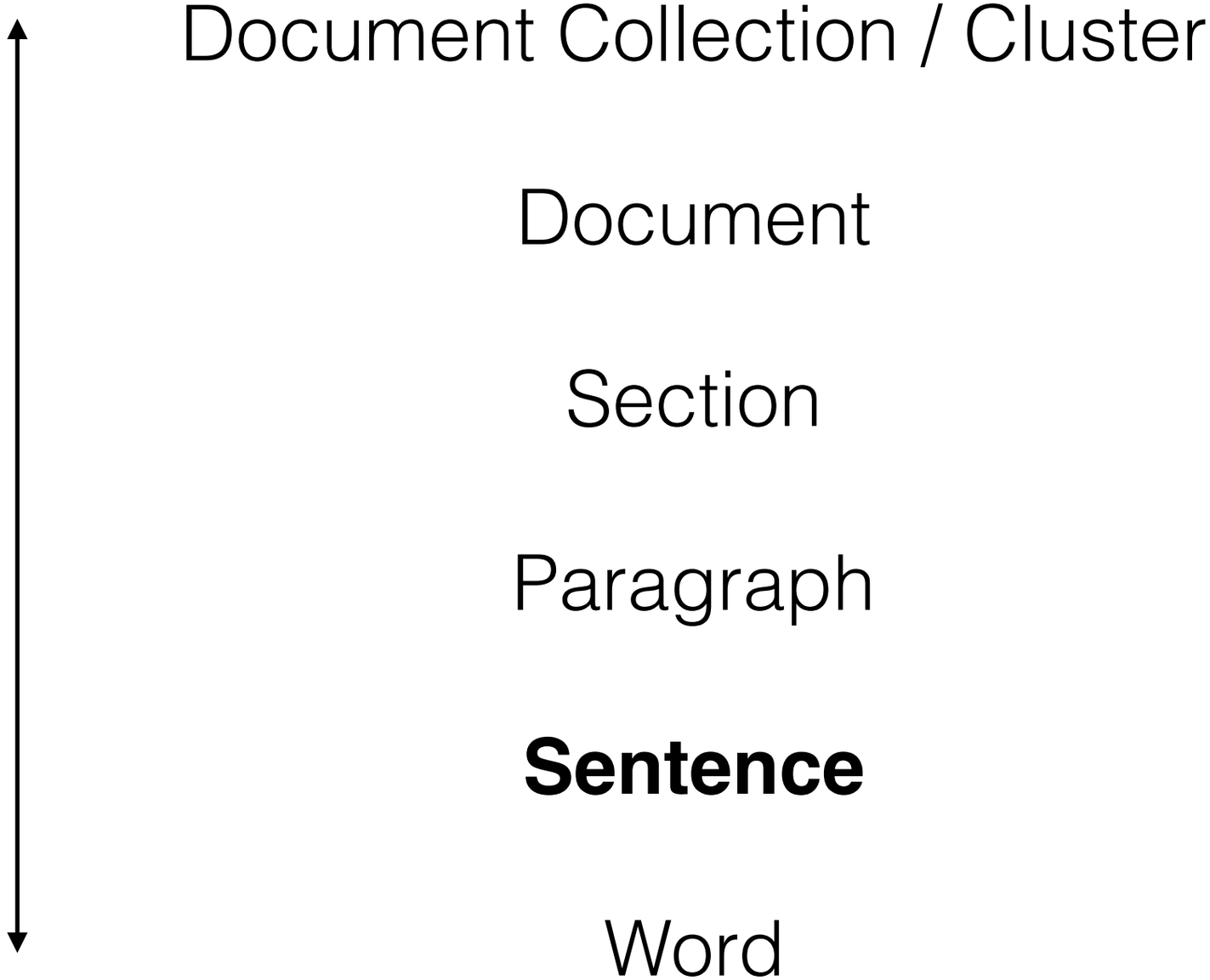
Top stories

Rio thief: How Olympic tourists can protect phone

Chicago's deadly Father's Day weekend: 12 murders, 54 shootings

and maybe each paragraph is a document?

What are we processing?



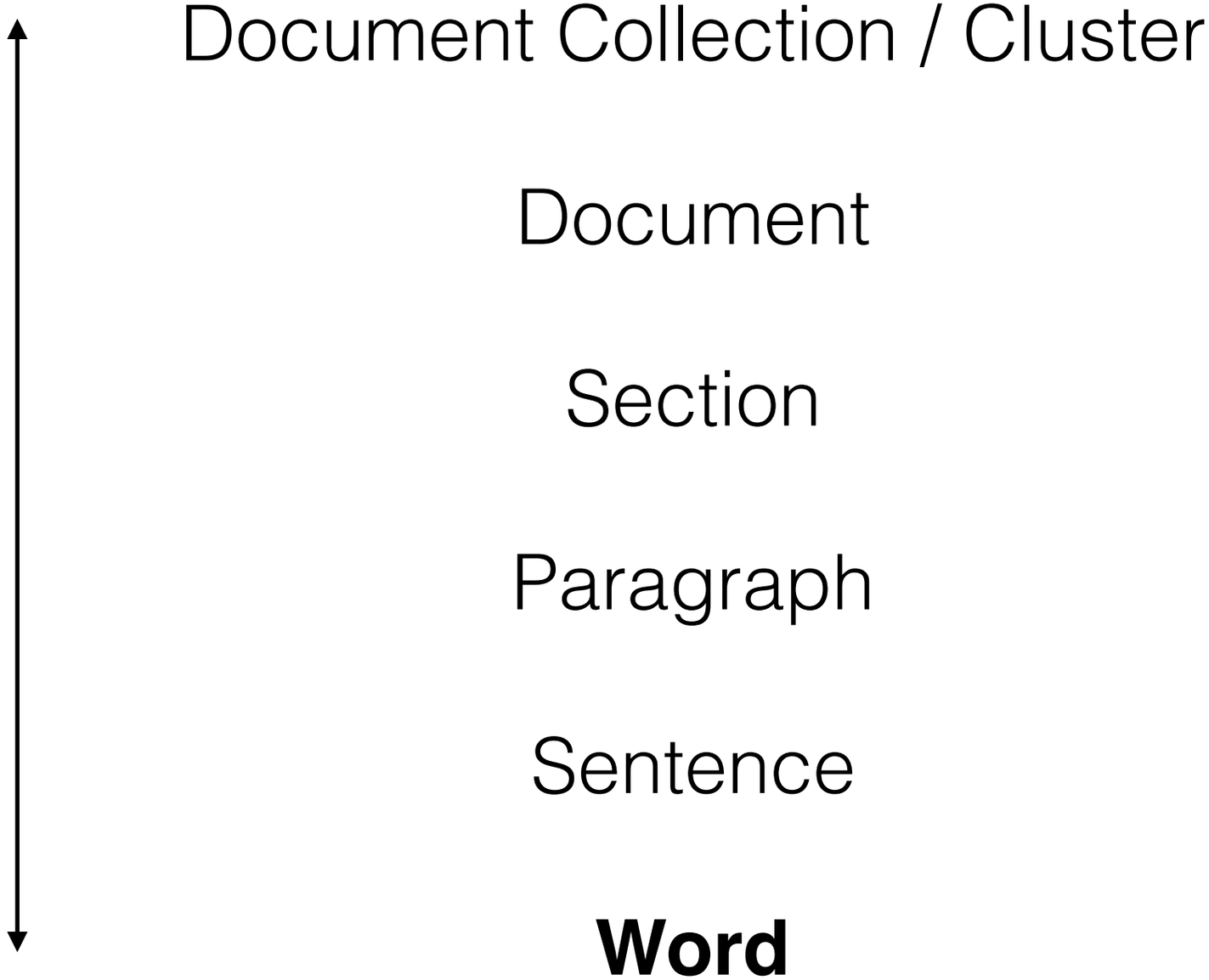
What is a sentence?

```
sentences = text.split(".")?
```

What is a sentence?

(we'll get back to this)

What are we processing?



What is a word?

What is a word?

sequence of characters

basic unit of meaning?

white-space tokenized?

What is a word?

doesn't

John's

unlucky

A Turkish word

uygarlaştıramadıklarımızdanmışsınızcasına

uygar_laş_tır_ama_dık_lar_ımız_dan_mış_sınız_casına

“as if you are among those whom we were not able to civilize (=cause to become civilized)”

uygar: *civilized*

_laş: *become*

_tır: *cause somebody to do something*

_ama: *not able*

_dık: past participle

_lar: plural

_ımız: 1st person plural possessive (our)

_dan: *among* (ablative case)

_mış: past

_sınız: 2nd person plural (you)

_casına: *as if* (forms an adverb from a verb)

K. Oflazer pc to J&M

by Julia Hockenmaier

What is a word?

שלהם

What is a word?

שלהם

שלו

What is a word?

שלהם

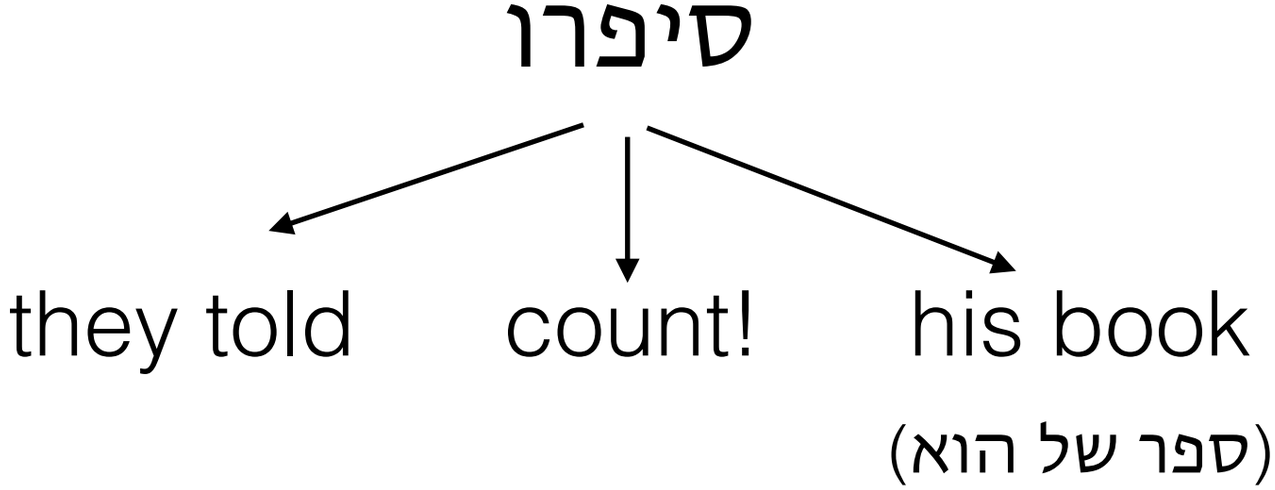
שלו

איתם

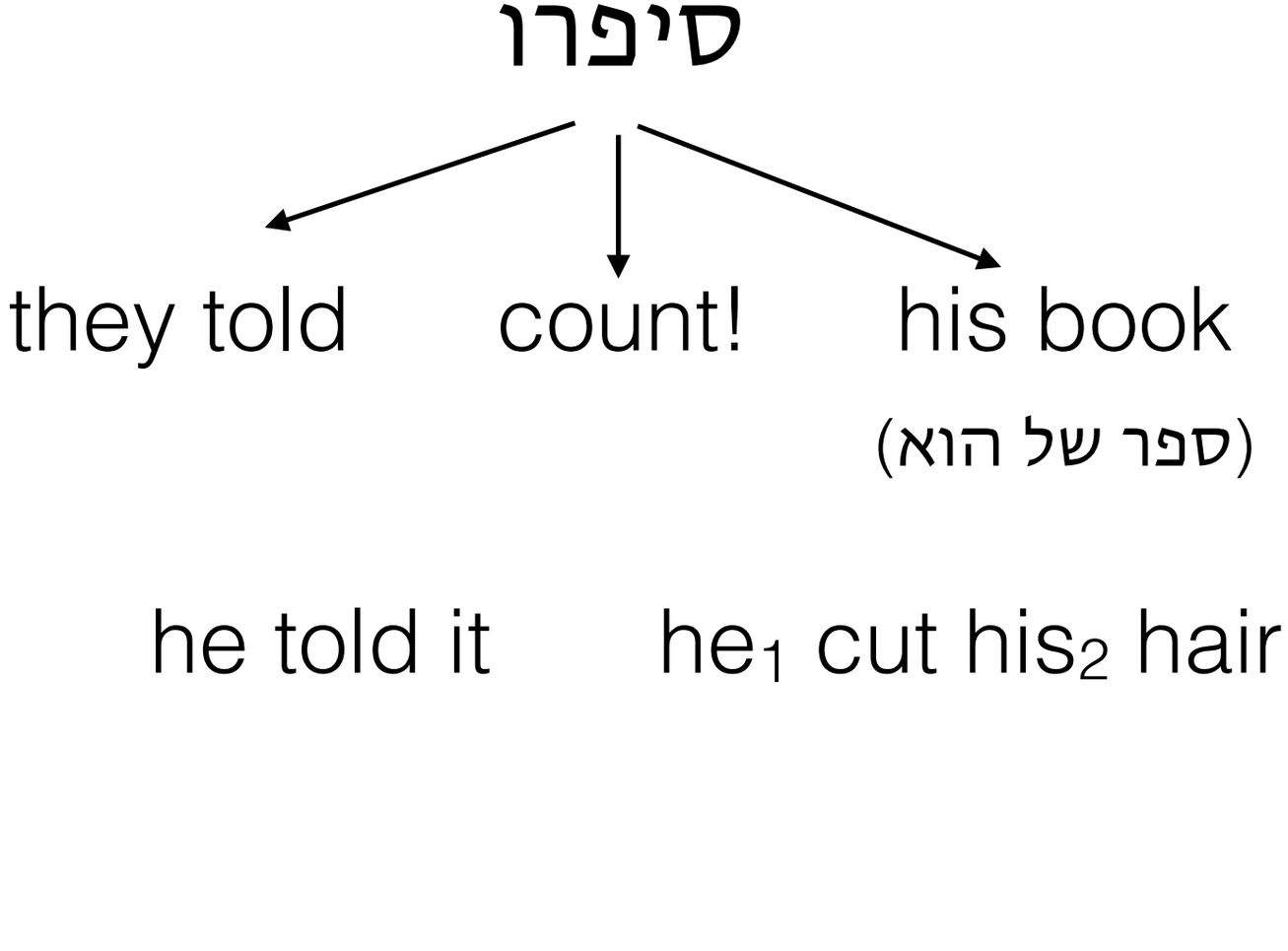
What is a word?

סיפרו

What is a word?



What is a word?



What is a word?

מסין
לסין
בסין
וסין
שסין
כשסין
ומסין
ולסין

What is a word?

בצלם

What is a word?

בצלם

בצל של הם

ב צל של הם

ב צלם

What is a word?

בצלם

בצל של הם

ב צל של הם

ב צלם

in a photographer
when taking a photo
in shape of

What is a word?

我开始写小说 = 我 开始 写 小说

I start(ed) writing novel(s)

by Julia Hockenmaier

whitespace is not enough

What is a word?

במהירות

בעקבות

What is a word?

ice cream

web site

New York

New York-Based

What is a word?

gave up

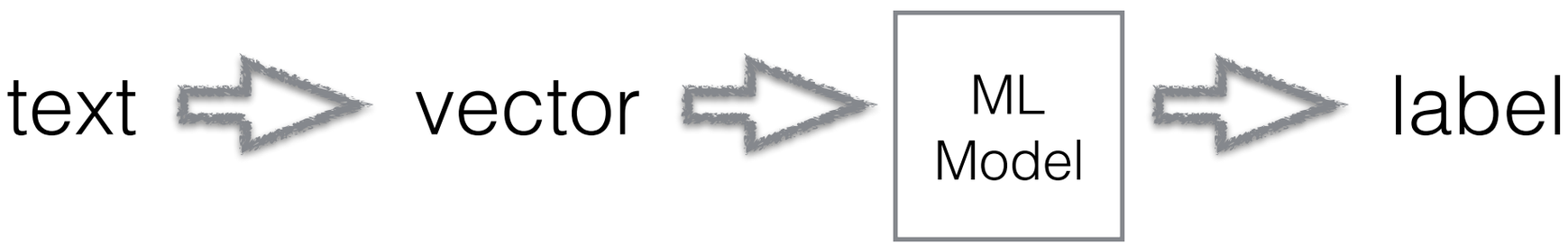
made sense

took a picture

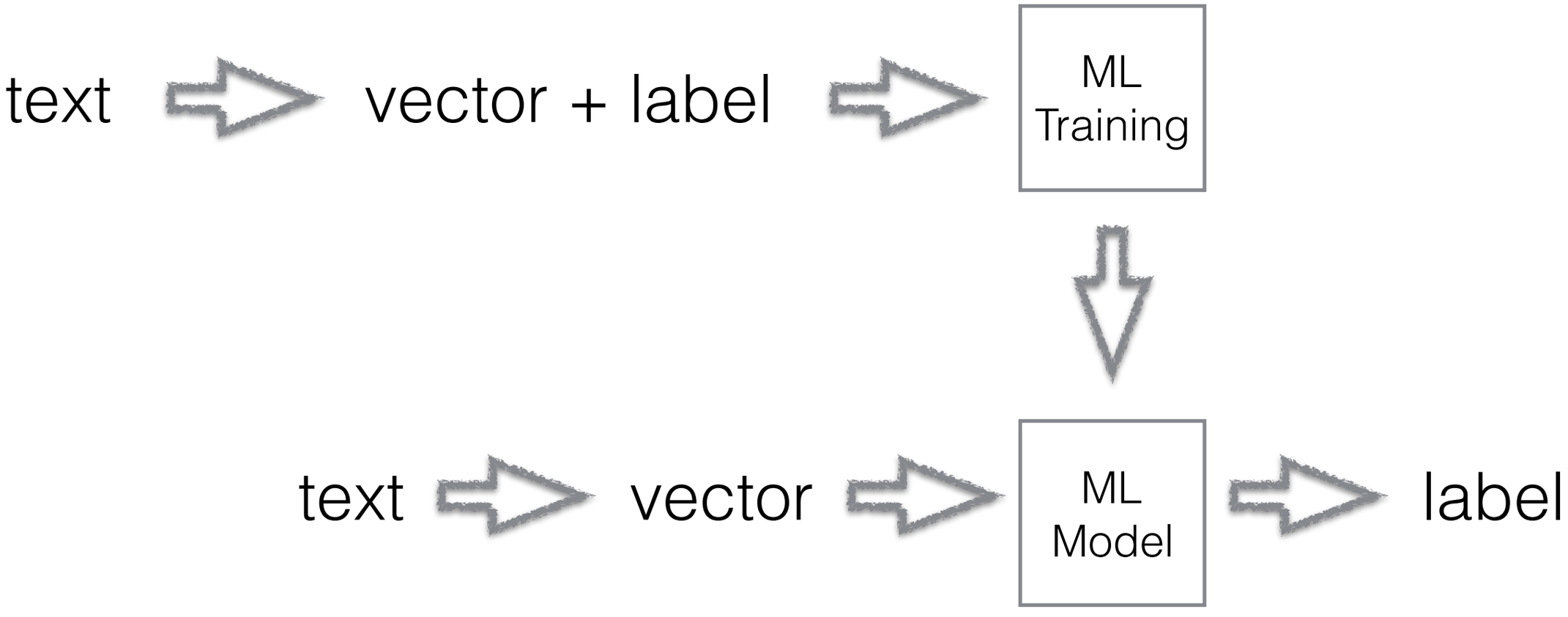
took apart

took the toy apart

Basic Structure of an NLP Algorithm

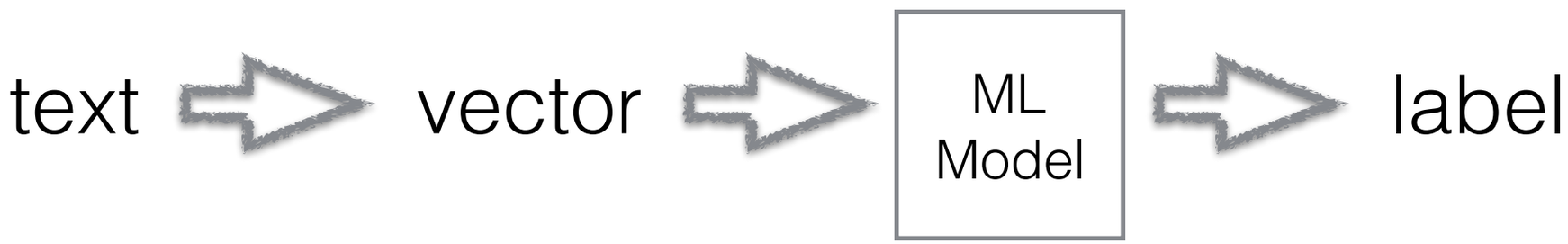


Basic Structure of an NLP Algorithm



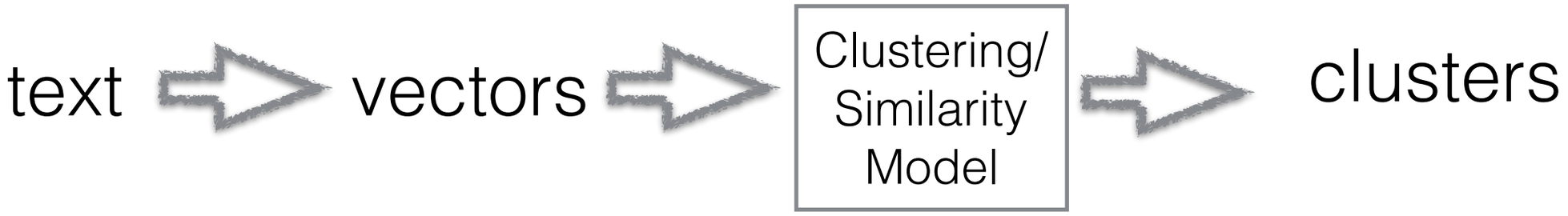
Basic Structure of an NLP Algorithm

Supervised:

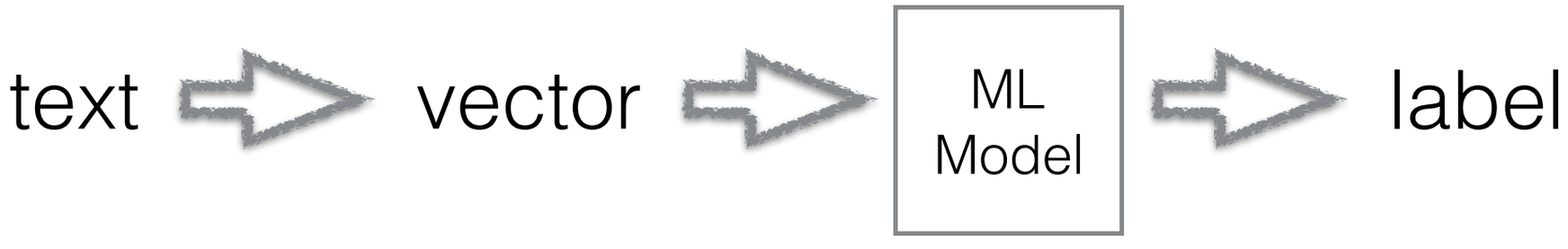


Basic Structure of an NLP Algorithm

Unsupervised:

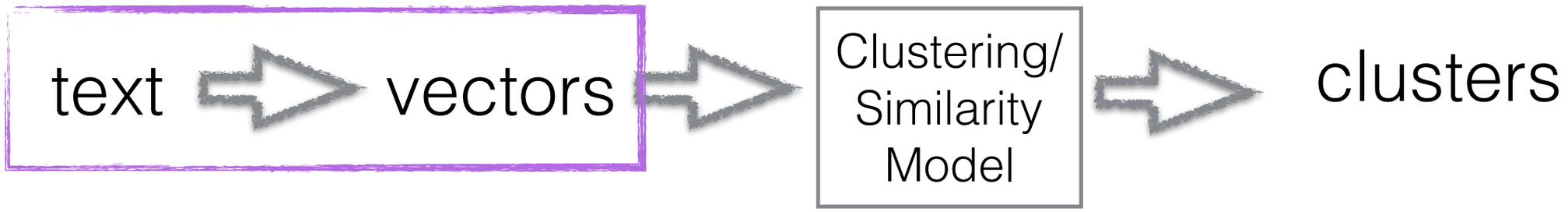


Supervised:

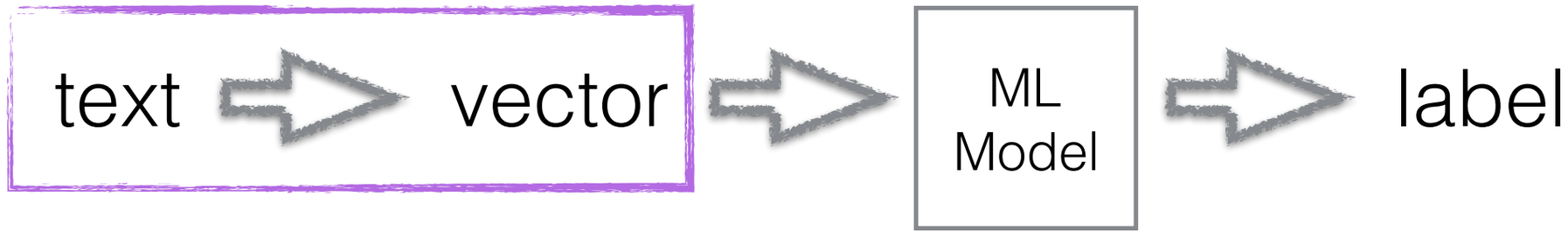


Basic Structure of an NLP Algorithm

Unsupervised:



Supervised:



this is not trivial!

Case Study:
Sentence Boundary Detection.

What is a sentence?

```
sentences = text.split(".")?
```

What is a sentence?

A **graphic adventure game** is a form of adventure game. They are distinct from text adventures. Whereas a player must actively observe using commands such as "look" in a text-based adventure, graphic adventures revolutionized gameplay by making use of visual human perception. Eventually, the text parser interface associated with older interactive fiction games was phased out in favor of a point-and-click interface, i.e., a game where the player interacts with the game environment and objects using an on-screen cursor. In many of these games, the mouse pointer is context sensitive in that it applies different actions to different objects.

What is a sentence?

```
sentences = text.split(". ")?
```

What is a sentence?

Georg "Mr. George" Svendsen (19 March 1894 – 1966) was a Norwegian journalist and crime novelist.

He was born in Eidanger, and started his journalistic career in Bratsberg-Demokraten before moving on to Demokraten where he was a subeditor. In 1921 he was hired in Fremtidenand replaced in Demokraten by Evald O. Solbakken. In 1931 he was hired in Arbeiderbladet. Under the pen name "Mr. George" he became known for his humorous articles in the newspaper. At his death he was also called "the last of the three great criminal and police reporters in Oslo", together with Fridtjof Knutsen and Axel Kielland. He was also known for practising journalism as a trade in itself, and not as a part of a political career. He retired in 1964, and died in 1966.

He released the criminal novels Mannen med ljåen (1942), Ridderne av øksen (1945) and Den hvite streken (1946), and translated the book S.S. Murder by Quentin Patrick as Mord ombord in 1945. He released several historical books: Rørleggernes Fagforenings historie gjennem 50 år (1934), Telefonmontørenes Forening Oslo, gjennem 50 år (1939), Norsk nærings- og nydelsesmiddelarbeiderforbund: 25-års beretning (1948), De tause vitner: av rettskjemiker Ch. Bruffs memoarer (1949, with Fridtjof Knudsen) and Elektriske montørers fagforening gjennom 50 år (1949).

Sentence Boundary Detection

- How would you solve this?

(discuss)

Sentence Boundary Detection

- Worth reading:

Sentence Boundary Detection and the Problem with the U.S.

Dan Gillick
Computer Science Division
University of California, Berkeley
dgillick@cs.berkeley.edu

What is a sentence?

It was a bad time. (it is always a bad time.)

What is a sentence?

In a quiet voice, he said
"this will not work. I am quitting",
and the he left the room.

www.randomwebsite.com

How often do you find yourself on the internet looking at the same boring pages?
You know there is something out there but you don't know where to look

Trust me, how bad could it be?

randomlink

Click above for a new random website
Links will open in a new window

Disclaimer: The websites you are forwarded to may contain naughty stuff.
here is no known "porn" but sites change hands often and sometimes they change ethical direction.
f you see something that looks like porn or is otherwise unacceptable let us know and we'll decide
if it should be removed or if you're just a crank who likes to complain.

Don't leave the house with your fly open. Make sure you've got up-to-date virus protection.

Colophon & Warning & Contact

Want to add your own links? After reading the warning type the url in the box below

add

If you see something that shouldn't be here let us know

random website dot com 2001-2011

"I'd like to quit thinking of the present, like right now,
as some minor insignificant preamble to something else."

Sentence Boundary Detection

- Perhaps the most basic task.
- Non-trivial.
- Need to consider corpus, features, annotation procedure, biases.
- Need to think of your use-cases, choose your sentence definition, methods, and trade-offs.

That's it for today.