



ALBERTA MACHINE  
INTELLIGENCE  
INSTITUTE



# NLP Workshop

HackED 2019

Luke Kumar

Machine Learning Scientist @ Amii

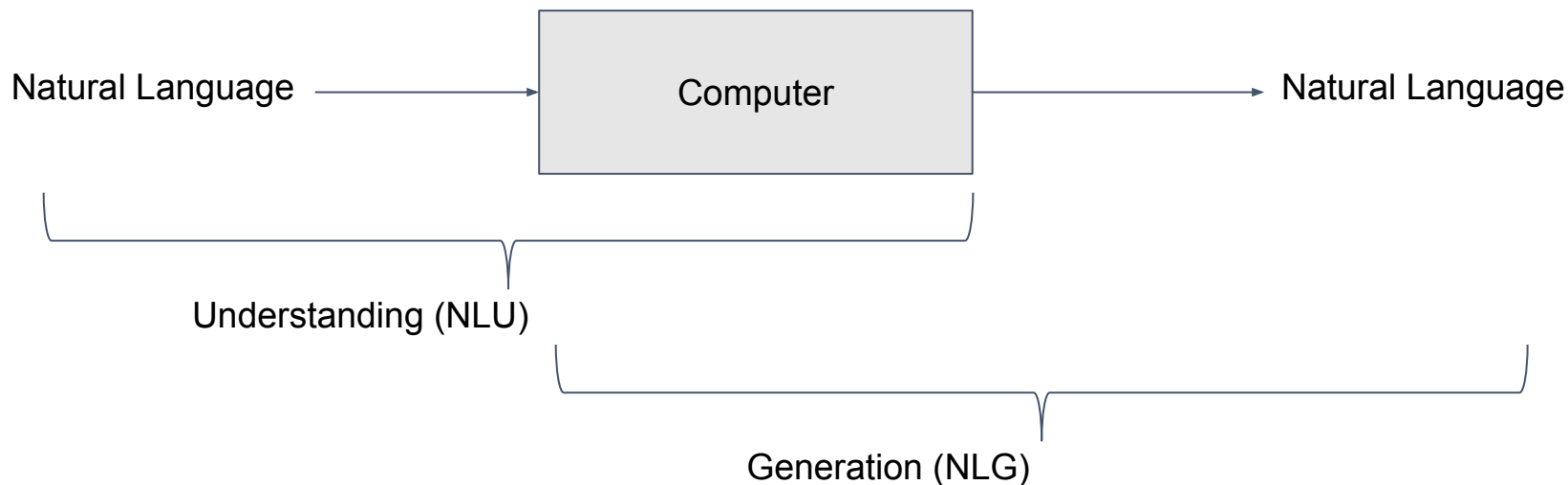
# Agenda

1. NLP & History
2. Why NLP is hard?
3. Symbolic/Classical NLP
4. Statistical/ML NLP
  - Notebook Demos
5. Q&A

Note: Materials & Diagrams from Prof. Regina Barzilay's NLP [Course](#) @ MIT (Advanced NLP)

# Natural language processing

NLP - Building programs that can use NL as input and output



# History of NLP

- First patents for translating machines were applied [mid-1930]
- Alan Turing published his famous article *Computing Machinery and Intelligence* which proposed what is now called the Turing test as a criterion of intelligence [1950]
- Noam Chomsky's Syntactic Structures universal grammar - a rule based system of syntactic structures [1957]
- Watson by IBM [2006]
- Word Embeddings [2013/2014]

Ref: [https://en.wikipedia.org/wiki/History\\_of\\_natural\\_language\\_processing](https://en.wikipedia.org/wiki/History_of_natural_language_processing)

# Why NLP is hard?

- **Ambiguity**

*“Harry loves his mother and Hermione does too”*

- Harry and hermione love their own mothers
- Hermione loves harry’s mother
- Different types of ambiguities:
  - Acoustic (sound)
  - Syntactic (structure)
  - Semantic (meaning)
  - Discourse (multi-clause)
    - “The horse ran up the hill. **It** was very steep. **It** soon got tired”

# How to solve ?

We need the:

1. Knowledge of the language
2. Knowledge of the world

Approaches:

1. Symbolic: Code all the rules into a program
2. Statistical: Learn language properties from examples

# NLP in use ...

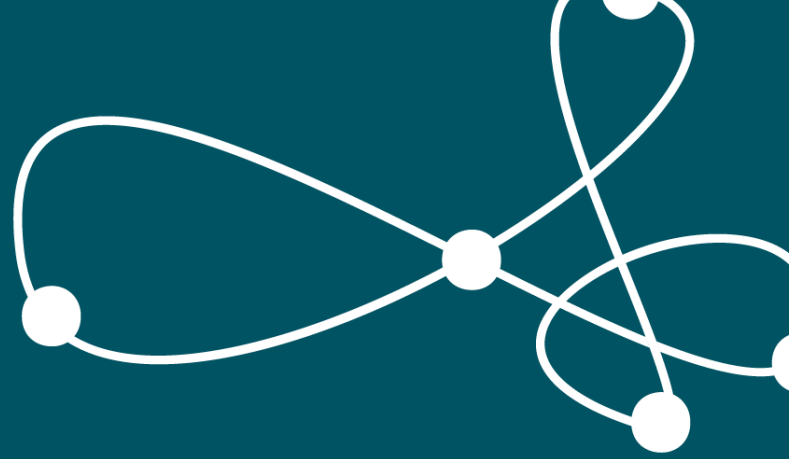
1. Language translation
2. Information extraction
  - a. Search
3. Text summarization
4. Sentiment analysis
5. Text to Speech
  - a. WaveNET
6. Chatbots - Conversational AI
  - a. Alexa
  - b. Google Home

.....

# NLP in research - ACL 2019 tracks

- Dialogue and Interactive Systems
- Discourse and Pragmatics
- Document Analysis
- Generation
- Information Extraction and Text Mining
- Linguistic Theories, Cognitive Modeling and Psycholinguistics
- Machine Learning
- Machine Translation
- Multidisciplinary
- Word-level Semantics
- Multilinguality
- Phonology, Morphology and Word Segmentation
- Question Answering
- Resources and Evaluation
- Sentence-level semantics
- Sentiment Analysis and Argument Mining
- Social Media
- Summarization
- Tagging, Chunking, Syntax and Parsing
- Textual Inference and Other Areas of Semantics
- Vision, Robotics, Multimodal, Grounding and Speech





# Classical/Symbolic NLP

# Topics in symbolic NLP

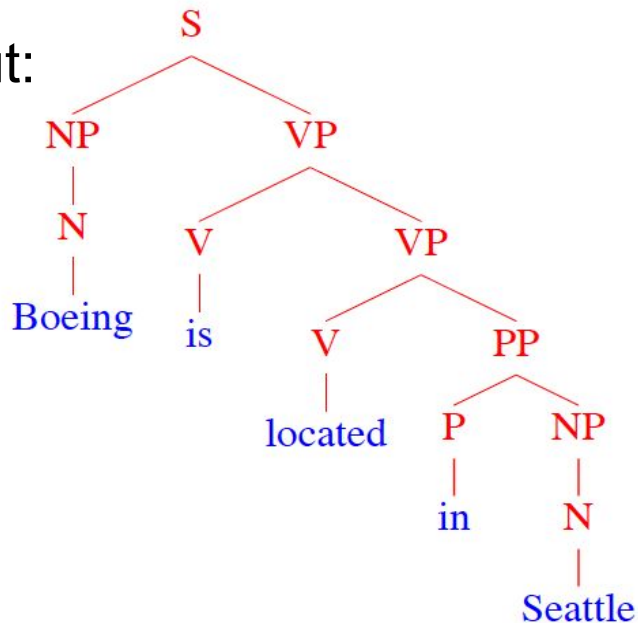
1. Parsing (we discuss)
2. Lexical semantics - Meanings of words
  - a. WordNet - “mother”
3. Stemming and lemmatization
  - a. am, are, is => be
  - b. car, cars, car's, cars' => car
4. Named entity recognition (NER)
  - a. Map text items to proper names (eg: people, location, organization)
5. ....

# Parsing - syntactic structure

Input:

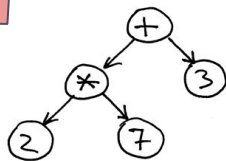
“Boeing is located in Seattle”

Output:



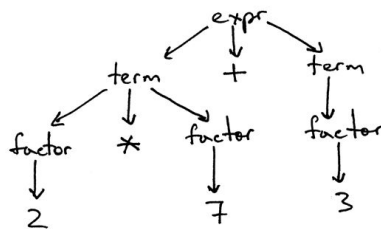
Programming  
Languages

AST



2 \* 7 + 3

Parse tree



# Parse tree

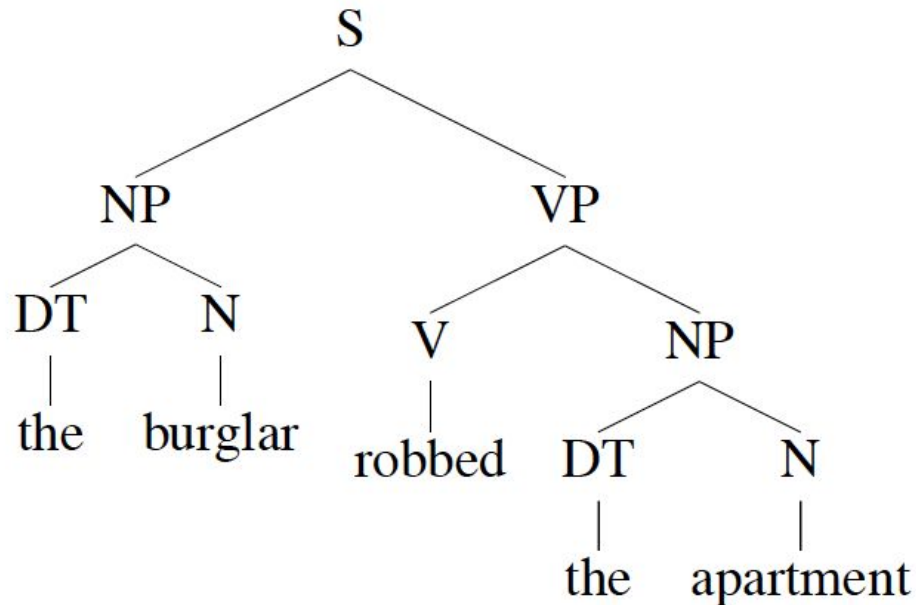
Parts of speech:

1. Words:

- a. N - noun
- b. V - verb
- c. DT - determiner

2. Phrases:

- a. NP - noun phrases
- b. VP - verb phrases
- c. S - Sentence



- 
- The diagram illustrates a full parse tree for the given sentence. The root node is TOP, which branches into S (Sentence) and VP (Verb Phrase). The S node further branches into NP (Noun Phrase), NNPS (Noun Phrase Singular), VBD (Verb, Past Tense), and another VP. This second VP branches into NP, PP (Prepositional Phrase), and another VP. The third VP branches into NP, PP, and another VP. The fourth VP branches into NP, PP, and another VP. The fifth VP branches into NP, PP, and another VP. The sixth VP branches into NP, PP, and another VP. The seventh VP branches into NP, PP, and another VP. The eighth VP branches into NP, PP, and another VP. The ninth VP branches into NP, PP, and another VP. The tenth VP branches into NP, PP, and another VP. The eleventh VP branches into NP, PP, and another VP. The twelfth VP branches into NP, PP, and another VP. The thirteenth VP branches into NP, PP, and another VP. The fourteenth VP branches into NP, PP, and another VP. The fifteenth VP branches into NP, PP, and another VP. The sixteenth VP branches into NP, PP, and another VP. The seventeenth VP branches into NP, PP, and another VP. The eighteenth VP branches into NP, PP, and another VP. The nineteenth VP branches into NP, PP, and another VP. The twentieth VP branches into NP, PP, and another VP. The twenty-first VP branches into NP, PP, and another VP. The twenty-second VP branches into NP, PP, and another VP. The twenty-third VP branches into NP, PP, and another VP. The twenty-fourth VP branches into NP, PP, and another VP. The twenty-fifth VP branches into NP, PP, and another VP. The twenty-sixth VP branches into NP, PP, and another VP. The twenty-seventh VP branches into NP, PP, and another VP. The twenty-eighth VP branches into NP, PP, and another VP. The twenty-ninth VP branches into NP, PP, and another VP. The thirtieth VP branches into NP, PP, and another VP. The thirty-first VP branches into NP, PP, and another VP. The thirty-second VP branches into NP, PP, and another VP. The thirty-third VP branches into NP, PP, and another VP. The thirty-fourth VP branches into NP, PP, and another VP. The thirty-fifth VP branches into NP, PP, and another VP. The thirty-sixth VP branches into NP, PP, and another VP. The thirty-seventh VP branches into NP, PP, and another VP. The thirty-eighth VP branches into NP, PP, and another VP. The thirty-ninth VP branches into NP, PP, and another VP. The fortieth VP branches into NP, PP, and another VP. The forty-first VP branches into NP, PP, and another VP. The forty-second VP branches into NP, PP, and another VP. The forty-third VP branches into NP, PP, and another VP. The forty-fourth VP branches into NP, PP, and another VP. The forty-fifth VP branches into NP, PP, and another VP. The forty-sixth VP branches into NP, PP, and another VP. The forty-seventh VP branches into NP, PP, and another VP. The forty-eighth VP branches into NP, PP, and another VP. The forty-ninth VP branches into NP, PP, and another VP. The fiftieth VP branches into NP, PP, and another VP. The fifty-first VP branches into NP, PP, and another VP. The fifty-second VP branches into NP, PP, and another VP. The fifty-third VP branches into NP, PP, and another VP. The fifty-fourth VP branches into NP, PP, and another VP. The fifty-fifth VP branches into NP, PP, and another VP. The fifty-sixth VP branches into NP, PP, and another VP. The fifty-seventh VP branches into NP, PP, and another VP. The fifty-eighth VP branches into NP, PP, and another VP. The fifty-ninth VP branches into NP, PP, and another VP. The sixtieth VP branches into NP, PP, and another VP. The sixty-first VP branches into NP, PP, and another VP. The sixty-second VP branches into NP, PP, and another VP. The sixty-third VP branches into NP, PP, and another VP. The sixty-fourth VP branches into NP, PP, and another VP. The sixty-fifth VP branches into NP, PP, and another VP. The sixty-sixth VP branches into NP, PP, and another VP. The sixty-seventh VP branches into NP, PP, and another VP. The sixty-eighth VP branches into NP, PP, and another VP. The sixty-ninth VP branches into NP, PP, and another VP. The seventieth VP branches into NP, PP, and another VP. The seventy-first VP branches into NP, PP, and another VP. The seventy-second VP branches into NP, PP, and another VP. The seventy-third VP branches into NP, PP, and another VP. The seventy-fourth VP branches into NP, PP, and another VP. The seventy-fifth VP branches into NP, PP, and another VP. The seventy-sixth VP branches into NP, PP, and another VP. The seventy-seventh VP branches into NP, PP, and another VP. The seventy-eighth VP branches into NP, PP, and another VP. The seventy-ninth VP branches into NP, PP, and another VP. The eightieth VP branches into NP, PP, and another VP. The eighty-first VP branches into NP, PP, and another VP. The eighty-second VP branches into NP, PP, and another VP. The eighty-third VP branches into NP, PP, and another VP. The eighty-fourth VP branches into NP, PP, and another VP. The eighty-fifth VP branches into NP, PP, and another VP. The eighty-sixth VP branches into NP, PP, and another VP. The eighty-seventh VP branches into NP, PP, and another VP. The eighty-eighth VP branches into NP, PP, and another VP. The eighty-ninth VP branches into NP, PP, and another VP. The ninetieth VP branches into NP, PP, and another VP. The ninety-first VP branches into NP, PP, and another VP. The ninety-second VP branches into NP, PP, and another VP. The ninety-third VP branches into NP, PP, and another VP. The ninety-fourth VP branches into NP, PP, and another VP. The ninety-fifth VP branches into NP, PP, and another VP. The ninety-sixth VP branches into NP, PP, and another VP. The ninety-seventh VP branches into NP, PP, and another VP. The ninety-eighth VP branches into NP, PP, and another VP. The ninety-ninth VP branches into NP, PP, and another VP. The hundredth VP branches into NP, PP, and another VP.

13

# Grammars

1. Context-Free Grammars (CFG)
  - a. Chomsky Normal Form (CNF)
2. Probabilistic CFG

$N = \{S, NP, VP, PP, DT, Vi, Vt, NN, IN\}$

$S = S$

$\Sigma = \{\text{sleeps, saw, man, woman, telescope, the, with, in}\}$

$R =$

$S$	$\Rightarrow$	$NP$	$VP$
$VP$	$\Rightarrow$	$Vi$	
$VP$	$\Rightarrow$	$Vt$	$NP$
$VP$	$\Rightarrow$	$VP$	$PP$
$NP$	$\Rightarrow$	$DT$	$NN$
$NP$	$\Rightarrow$	$NP$	$PP$
$PP$	$\Rightarrow$	$IN$	$NP$

$Vi$	$\Rightarrow$	sleeps
$Vt$	$\Rightarrow$	saw
$NN$	$\Rightarrow$	man
$NN$	$\Rightarrow$	woman
$NN$	$\Rightarrow$	telescope
$DT$	$\Rightarrow$	the
$IN$	$\Rightarrow$	with
$IN$	$\Rightarrow$	in

# CFG - “the man sleeps”

## Derivation

S

NP VP

DT NN VP

the NN VP

the man VP

the man Vi

*the man sleeps*

## Rules

S → NP VP

NP → DT NN

DT → the

NN → man

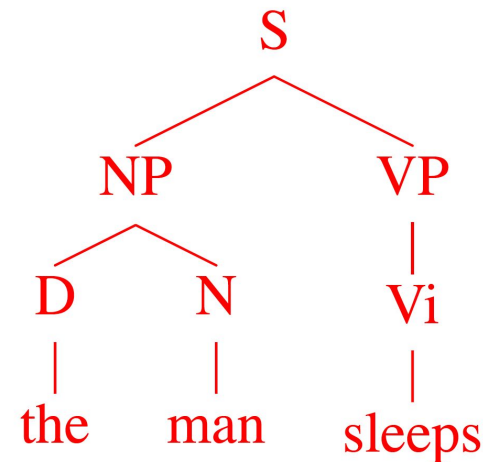
VP → Vi

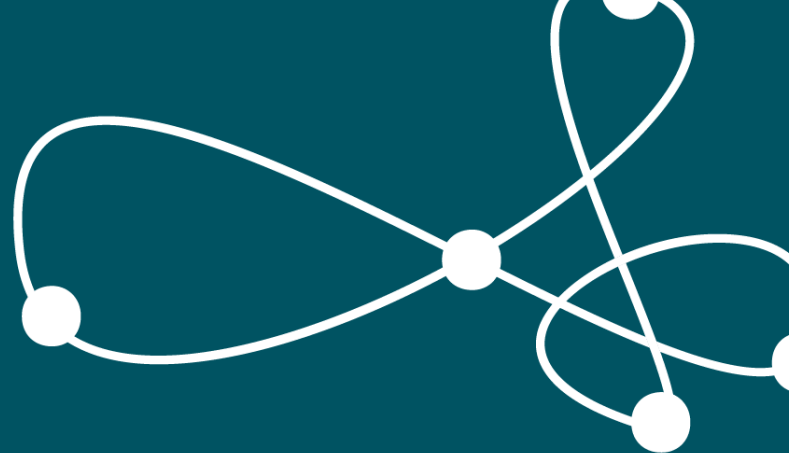
Vi → sleeps

$$R =$$

S	⇒	NP	VP
VP	⇒	Vi	
VP	⇒	Vt	NP
VP	⇒	VP	PP
NP	⇒	DT	NN
NP	⇒	NP	PP
PP	⇒	IN	NP

Vi	⇒	sleeps
Vt	⇒	saw
NN	⇒	man
NN	⇒	woman
NN	⇒	telescope
DT	⇒	the
IN	⇒	with
IN	⇒	in





# Statistical NLP

ML applications in NLP



# Topics in statistical NLP (we discuss)

- Sentiment analysis
  - Positive vs negative polarity
- Language model
  - Probability distribution of a natural language
- Word embedding
  - Representing words as numerical vectors
- Topic model
  - Categorizing document collections

# Other topics

- Statistical Machine Translation (SMT)
- Tagging (eg: POS tagging)
  - Hidden Markov Models (HMMs)
  - Conditional Random Forests (CRFs)
- Recurrent neural networks
  - Sequence to Sequence tasks (eg: Translation)
- Duplicate detection
- Other embeddings
  - Document2vector
  - Character2vector
- Conversational AI
- .....

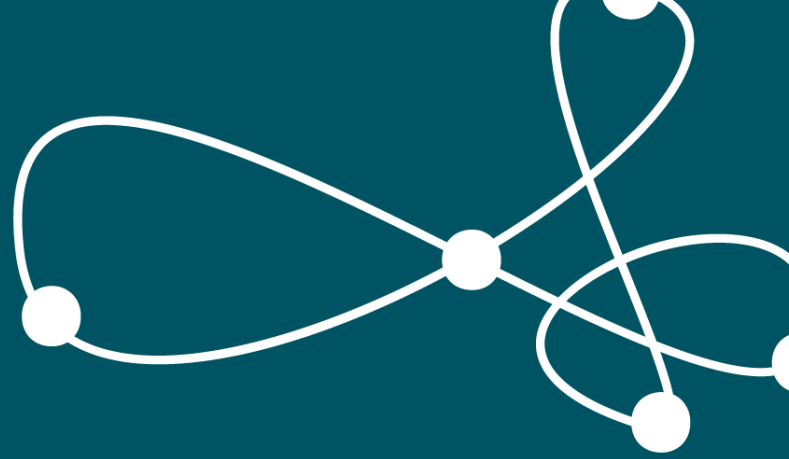
# Text to numeric representation

- Bag of words
  - a. Absence and presence
  - b. Frequency
  - c. Term frequency inverse document frequency (TFiDF)
- N-grams
  - a. Unigrams = Bag of words
- Embeddings
  - a. Word embeddings
  - b. Document embeddings
  - c. Character embeddings

# Sentiment analysis

- AKA Opinion mining
  - Polarity: positive or negative
- Sentiment Classification
  - [Polarity Data 2.0 - Movie reviews](#)
- Sentiment Lexicons
  - [Subjectivity Lexicon](#)
  - [Bing Liu Opinion lexicon](#)
  - [SentiWordNet](#)

Ref: <https://web.stanford.edu/class/cs124/lec/sentiment.pdf>



# Demo

Sentiment\_Analysis Notebook

# Language model (Autoregressive models)

- Learn a probability distribution

$$\sum_{x \in \mathcal{V}^*} \hat{P}(x) = 1, \quad \hat{P}(x) \geq 0 \text{ for all } x \in \mathcal{V}^*$$

$$\hat{P}(\text{the}) = 10^{-12}$$

$$\hat{P}(\text{the fan}) = 10^{-8}$$

$$\hat{P}(\text{the fan saw Beckham}) = 2 \times 10^{-8}$$

$$\hat{P}(\text{the fan saw saw}) = 10^{-15}$$

# Language model - n-grams

- Trigram (triplets) Model

$$P(w_i \mid w_{i-2}, w_{i-1})$$

*eg:*

$$P(\text{"well"} \mid \text{"all"}, \text{"is"}) = \frac{\text{Count}(\text{all, is, well})}{\text{Count}(\text{all, is})}$$

$$P(w_1, w_2, \dots, w_{T-1}, w_T) = \prod_{t=1}^T P(w_t | w_{t-1}, w_{t-2}, \dots, w_1)$$

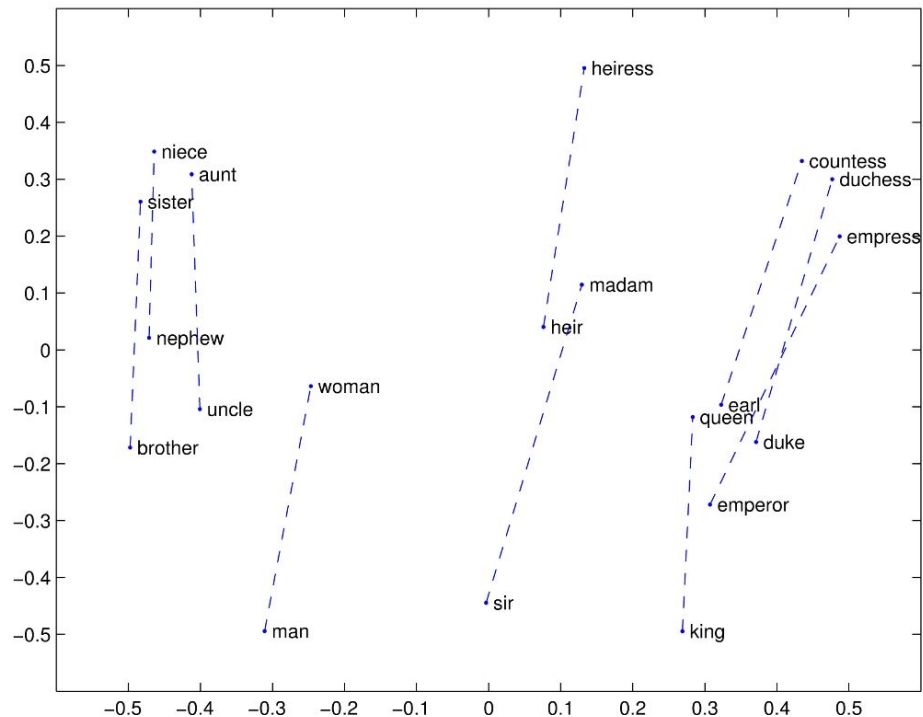
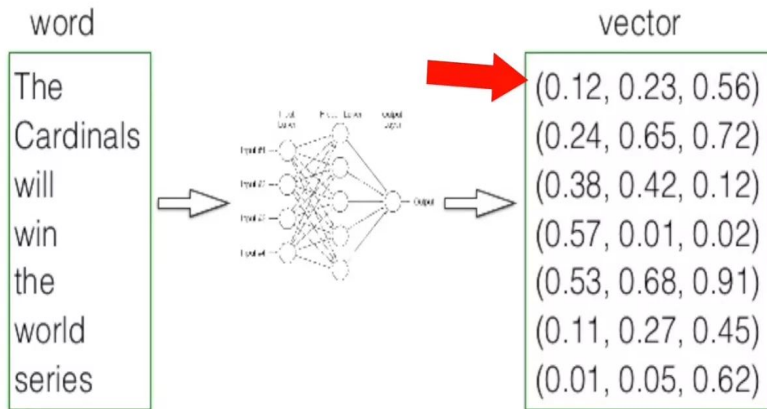
<b>the</b>	cat	sat	on	the	mat	$P(w_1)$
the	<b>cat</b>	sat	on	the	mat	$P(w_2   w_1)$
the	cat	<b>sat</b>	on	the	mat	$P(w_3   w_2, w_1)$
the	cat	sat	<b>on</b>	the	mat	$P(w_4   w_3, w_2, w_1)$
the	cat	sat	on	<b>the</b>	mat	$P(w_5   w_4, w_3, w_2, w_1)$
the	cat	sat	on	the	<b>mat</b>	$P(w_6   w_5, w_4, w_3, w_2, w_1)$

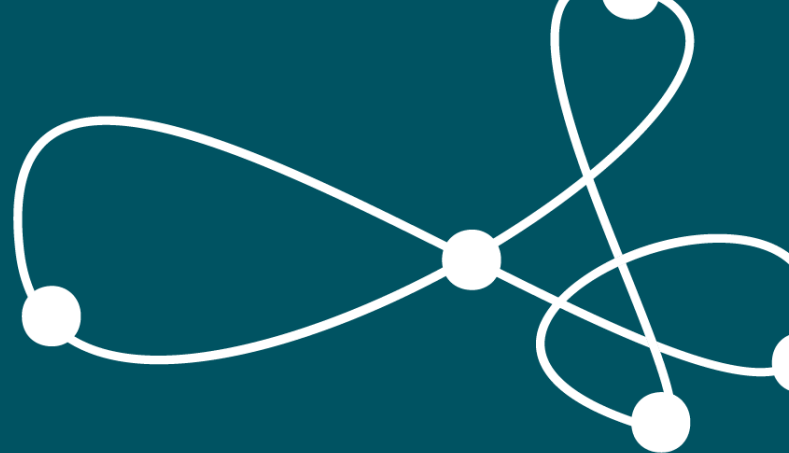


# Word embeddings

- Representations of NL
  - a. Bag of words
  - b. N-grams
- Embedding vectors - *Map words into vectors of real values*
  - a. Embeddings are influenced by the context
  - b. Embeddings try to capture the meaning using the context
- How embeddings are learnt:
  - a. Language Model
    - Word2Vec (Google)
  - b. Co-occurrence Matrix
    - [GloVe \(Stanford\)](#)

# Word embeddings cont.





# Demo

Word\_Embeddings Notebook

# Topic models

- A statistical model to learn intrinsic topics in a collection of documents
- Several models are proposed
  - LDA - Latent Dirichlet Allocation (Probabilistic)
  - LSA - Latent Semantic Analysis (SVD)
  - PLSA - Probabilistic Latent Semantic Analysis (Probabilistic)
- Helps to cluster a collection of documents
  - Soft clustering
  - Mostly interpretable (Probabilistic models)

# LDA

## Topics

gene 0.04  
dna 0.02  
genetic 0.01  
...

life 0.02  
evolve 0.01  
organism 0.01  
...

brain 0.04  
neuron 0.02  
nerve 0.01  
...

data 0.02  
number 0.02  
computer 0.01  
...

## Documents

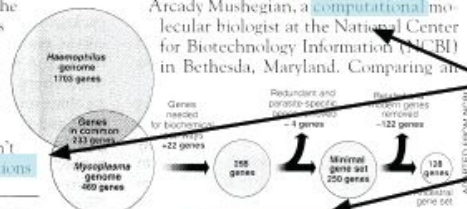
## Topic proportions and assignments

### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson at Uppsala University in Sweden. "We arrived at the 800 number. But coming up with a consensus answer may be more than just a **simple numbers** game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



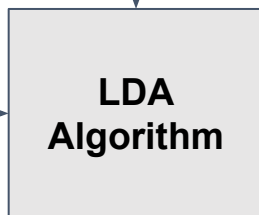
\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

# LDA cont.

1. K - number of topics
2. Prior document-topic distri.
3. Prior topic-word distri.



*Topics*

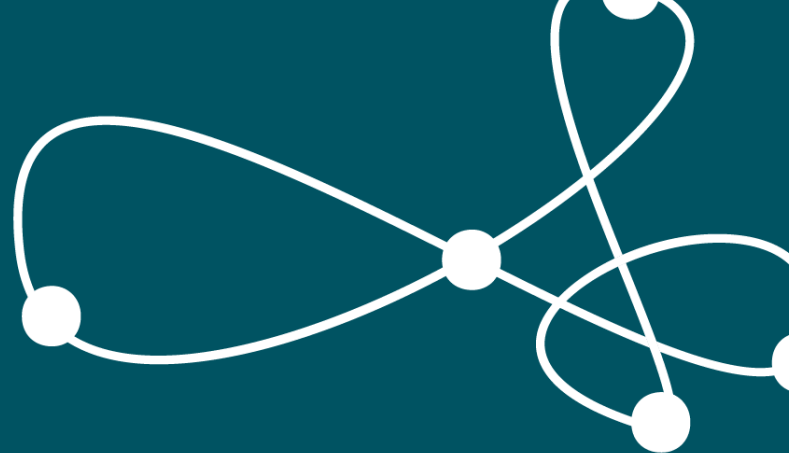
gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

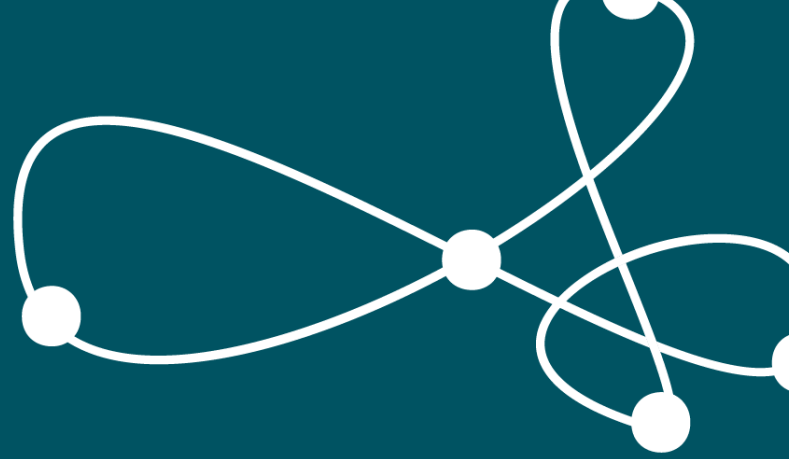
data	0.02
number	0.02
computer	0.01
...	

Topic -1 = 0.25  
Topic -2 = 0.33  
Topic -3 = 0.001  
....



# Demo

Topic\_Models Notebook



# Questions ?

[amii.ca](http://amii.ca)