DATA-AI – September 9th, 2024

NATURAL LANGUAGE PROCESSING

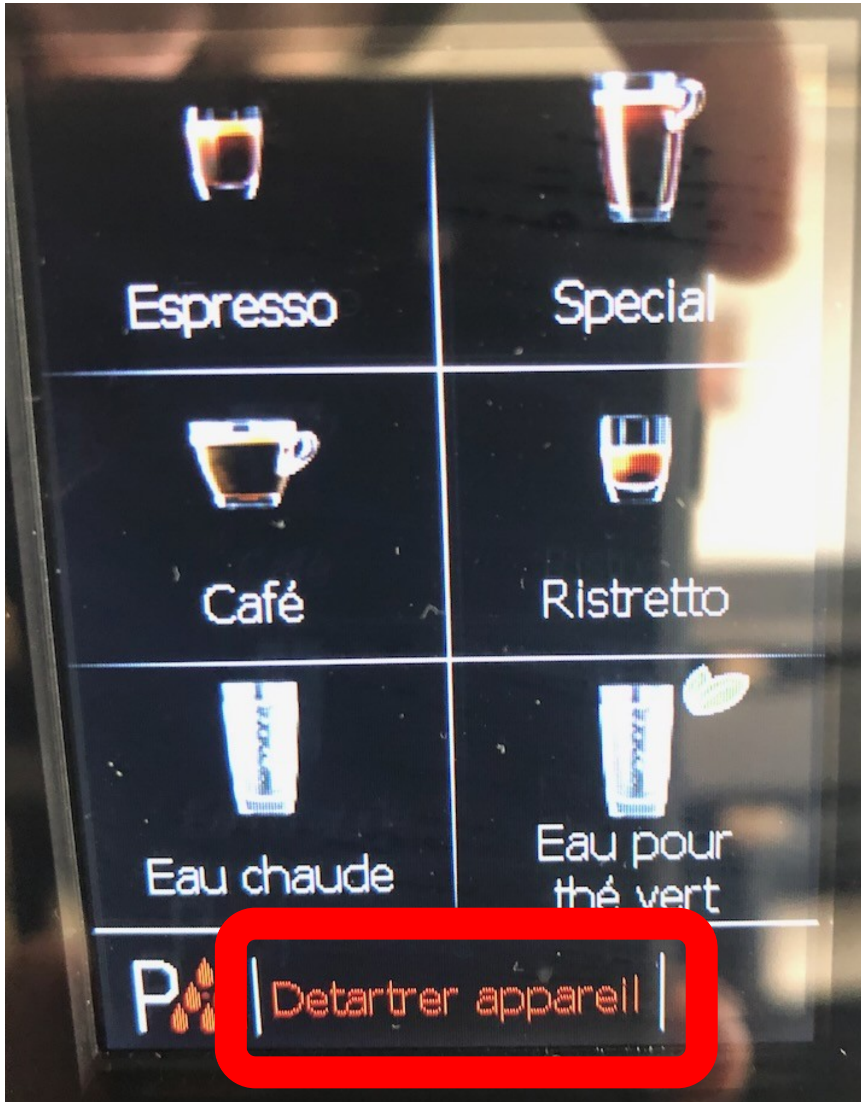
Nils Holzenberger

Télécom Paris

Institut Polytechnique de Paris

Natural Language Processing

- Using language to do something
  - Exchange information
  - Gather information
  - Coordinate actions
  - Convince other people
  - For its own sake
- Humans do it most of the time
  - Your brain is very good at it
  - Children learn it efficiently
- Machines use it too



Natural Language Processing

Some machines are better at it than others…

Introducing ChatGPT

We’ve trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

Natural Language Processing

…but some things are still missing.

Human: The trophy doesn’t fit in the brown suitcase because it is too big. What is too big?

ChatGPT: The trophy is too big to fit in the brown suitcase.

Human: The trophy doesn’t fit in the brown suitcase because it is too small. What is too small?

ChatGPT: If the trophy doesn’t fit in the brown suitcase because it is too small, then the trophy itself is too small to fit in the suitcase.

(This example is from 2022, ChatGPT now answers this correctly.)

What is so hard about it?

- Allophones **Phonetics**
  - “**p**in” *vs* “s**p**in” / [p^h] vs [p]
  - “cana**r**d jaune” *vs* “cana**r**d flottant” / [ʁ] vs [χ] / “Bu**r**e” *vs* “Bu**ch**e”
  - Wikipedia -> Allophone
  - Space between words
- “close to the Sun” **Lexical ambiguity**
  (in the sky, on a picture, within the solar system)
- “In this country a woman gives birth every fifteen minutes. Our job is to find that woman and stop her.” (Groucho Marx) **Semantic ambiguity**
- “She was watching the man with the telescope.” **Syntactic ambiguity**
- **Linguistics** is the science of studying language

Language processing

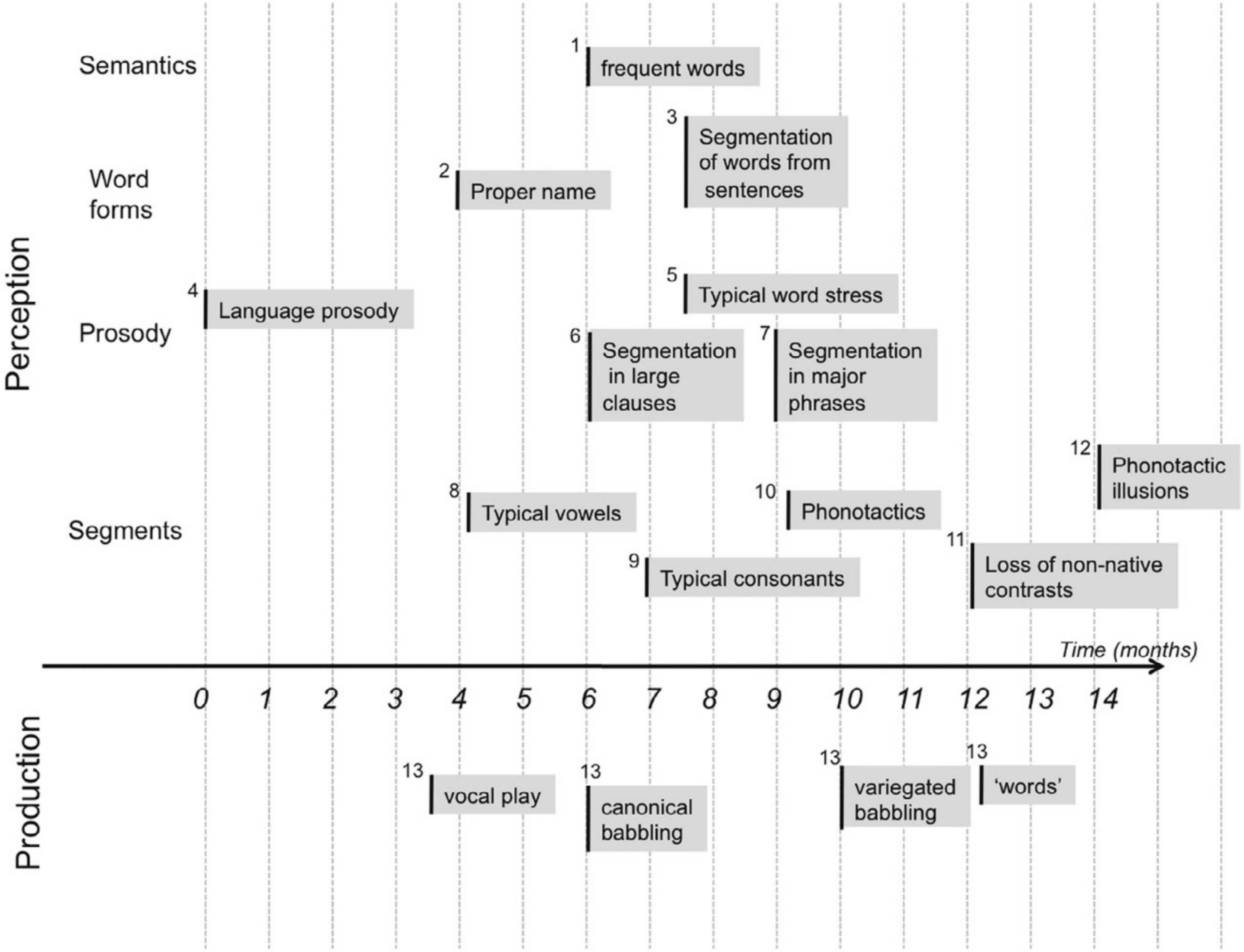
Level	*counter-example*	*example*
Phonology	Bning	Ulbrage
Morphology	Coolth	Warmth
Syntax	House money pay	The … arden eats the
Semantics	The apple …	… t a bike
Relevance	Two year … in the str…	… n the metro.

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
- Lewis Carroll, *Jabberwocky*

These levels correspond to cognitive modules

Language acquisition

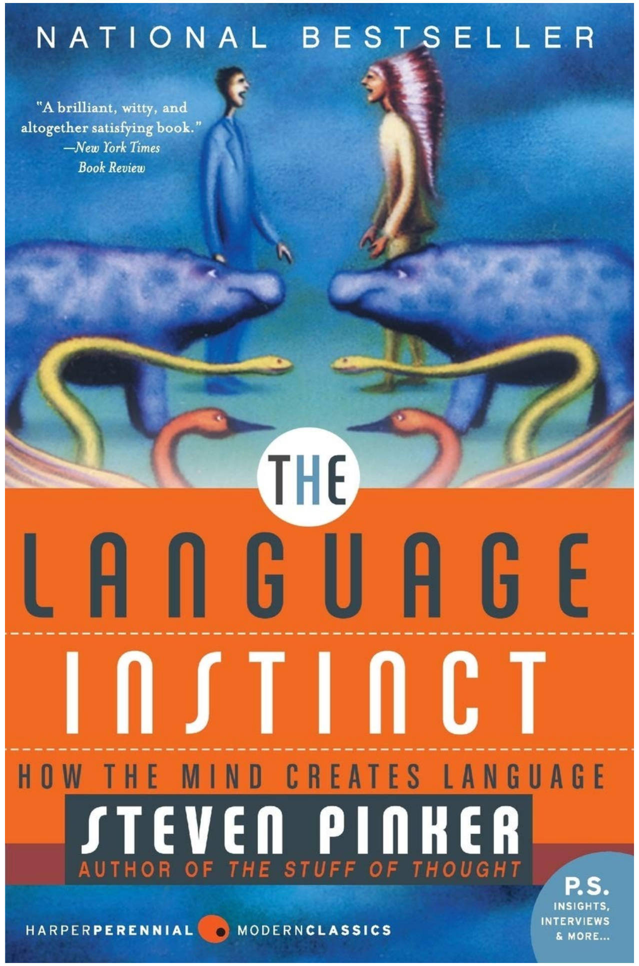
Children acquire language without being explicitly taught



- We know **what** they learn and **when** they learn it, but not **how** they learn it
- 3-10 new words learned daily until adolescence
- Learning guided by structure and rules?

Figure from Emmanuel Dupoux, *Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner*, Cognition 2018

If you want to learn more about linguistics and language acquisition:



Steven Pinker

The Language Instinct

Garden path sentences

Garden path sentences

The government plans to raise taxes are unpopular.

Fat people eat is dangerous.

La maire a marié Pierre et Jeanne y était.

Bei meinem Haus am Stausee müssen die Staubecken gesäubert werden, am besten mit dem Staubtuch.

Can computers deal with language?

- Yes, <u>**if**</u>

True for statistical and symbolic methods

  - the task is well-defined
    how is the program supposed to react to its input?
  - the linguistic phenomena and their variability are abundantly described
    what kind of linguistic processing and what kind of ambiguity are relevant?
- E.g. in speech recognition:
  - What language are we recognizing?
  - What type of speech (spontaneaous, read…)?
  - What are the phonemes in the language?

Natural Language Processing

- Strictly speaking, it’s about getting computers to solve tasks that involve natural language
  - Answering queries posed in natural langage
    What is the capital of states that New York borders? → `SELECT * FROM ...`
  - Extracting information from natural langage
    Founded in 1876, Johns Hopkins was the first... → `founded(jhu, 1876)`
  - Finding documents in a large collection
    e.g. Google search engine
- NLP is mostly engineering, possibly with insights from linguistics

NLP is not linguistics

- Linguistics is the science of studying language
- Language as a natural phenomenon
- What are the (mathematical) tools that best describe/explain language?
- Phonology
  How do languages choose which sounds to use?
- Morphology
  How are words built from sub-units?
- Syntax
  What types of sentences does a given language allow?
- Semantics
  How do humans use language to transmit meaning?

NLP is not psycholinguistics

- Psycholinguistics is the science of how humans process language
- What cognitive abilities are used when…
  - listening to language?
  - reading text?
  - speaking?
- How do infants acquire their first language?
- Evolution of language: how did language emerge in the course of human evolution?

NLP is not computational linguistics

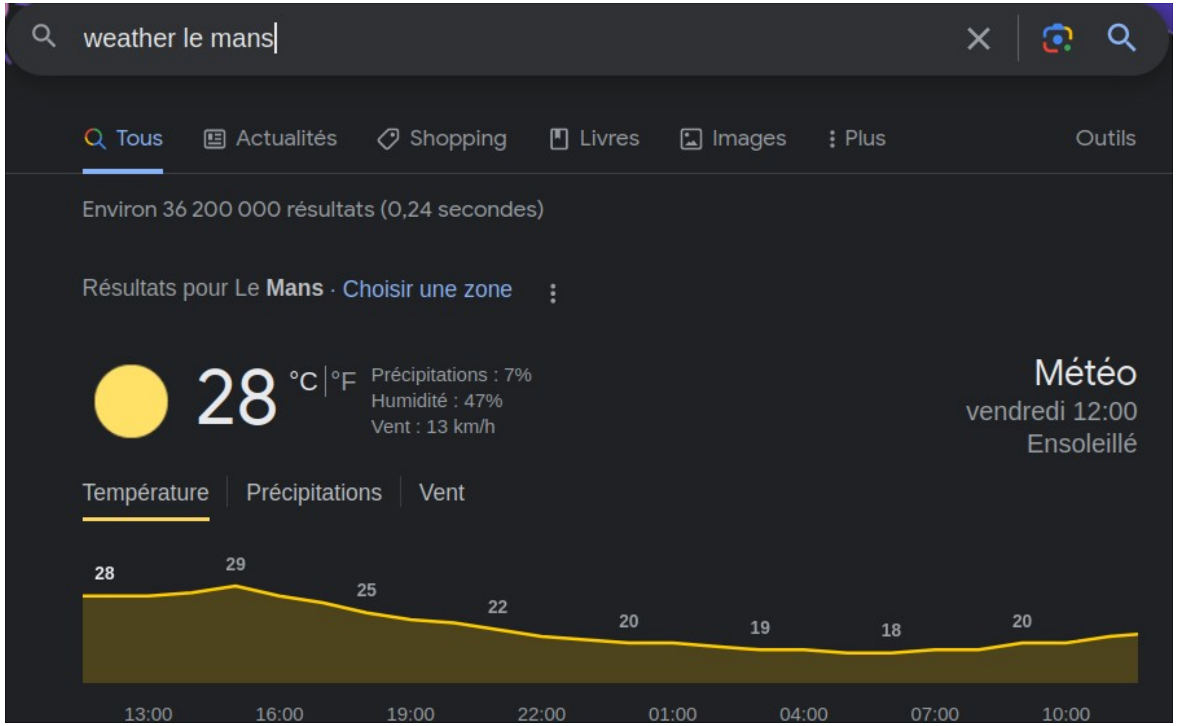
- Computational linguistics uses tools from computer science for research in linguistics
- The goal of (computational) linguistics is to gain insights about language
- The goal of NLP is to accomplish a task that involves language
- Both can involve machine learning and might be published at the same conferences

NLP is not machine learning

- Historically, NLP has increasingly used machine learning as a tool
- NLP can exist without machine learning (and vice-versa)
- NLP and machine learning have inspired one another
  - New ML methods have often been tested on NLP tasks
  - Problems from NLP have helped develop new ML methods

NLP is a popular field of research

- The Association for Computational Linguistics (ACL) was founded in 1962
  - The 1979 ACL conference had a single track, with 29 papers
  - The 2022 conference had 1651 spread across 33 tracks and workshops
- Many major tech applications involve NLP
- New research comes out of universities and tech companies



Adapted from Ryan Cotterell's *Introduction to NLP*

What can NLP do?

1950 Turing test, the first NLP task

A. M. Turing (1950) Computing Machinery and Intelligence. *Mind 49:* 433-460.

COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. The Imitation Game

I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous, If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a

ELIZA

1966 ELIZA chatbot

Could produce intelligible language, but not always relevant content

```
  Welcome to
                    EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                    EE      LL       II       ZZ   AA   AA
                    EEEEE   LL       II     ZZZ    AAAAAAA
                    EE      LL       II    ZZ      AA   AA
                    EEEEEE  LLLLLL  IIII ZZZZZZ    AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

Source: https://en.wikipedia.org/wiki/ELIZA

Prolog

1972 Prolog

- Programming language intended for reasoning with language
- First implementation included a syntax analyzer for French
- Prolog ended up being used mostly for logic programming

TOUT PSYCHIATRE EST UNE PERSONNE.	*Every psychiatrist is a person.*
CHAQUE PERSONNE QU'IL ANALYSE, EST MALADE.	*Every person he analyzes is ill.*
*JACQUES EST UN PSYCHIATRE A *MARSEILLE.	*Jacques is a psychiatrist in Marseilles.*
EST-CE QUE *JACQUES EST UNE PERSONNE?	*Is Jacques a person?*
OU EST *JACQUES?	*Where is Jacques?*
EST-CE QUE *JACQUES EST MALADE?	*Is Jaques ill?*
OUI	*Yes*
A MARSEILLE	*In Marseilles*
JE NE SAIS PAS	*I don't know*

A. Colmerauer *et al*, *Un système de communication homme-machine en Français*, tech report 1973(?)

Boeing’s NLP system

2008 Boeing’s NLP system

- A massive further development over the 1973 Prolog paper
- Meant to solve complex problems stated in natural language

Boeing’s NLP System and the Challenges of Semantic Representation

Peter Clark
Phil Harrison
The Boeing Company (USA)

Boeing’s NLP system

```
(1.1) "An object is thrown with a horizontal speed of 20 m/s
       from a cliff that is 125 m high."
isa(object01,object_n1),
isa(speed01,velocity_n1),
isa(horizontal01,horizontal_a1),
isa(cliff01,cliff_n1),
isa(height01,height_n1),
isa(throw01,throw_v1),
height(cliff01,height01),
value(speed01,[20,m/s_n1]),
mod(speed01,horizontal01),
value(height01,[125,m_n1]),
object(throw01,object01),
"with"(throw01,speed01),
origin(throw01,cliff01).
```

Boeing's NLP system

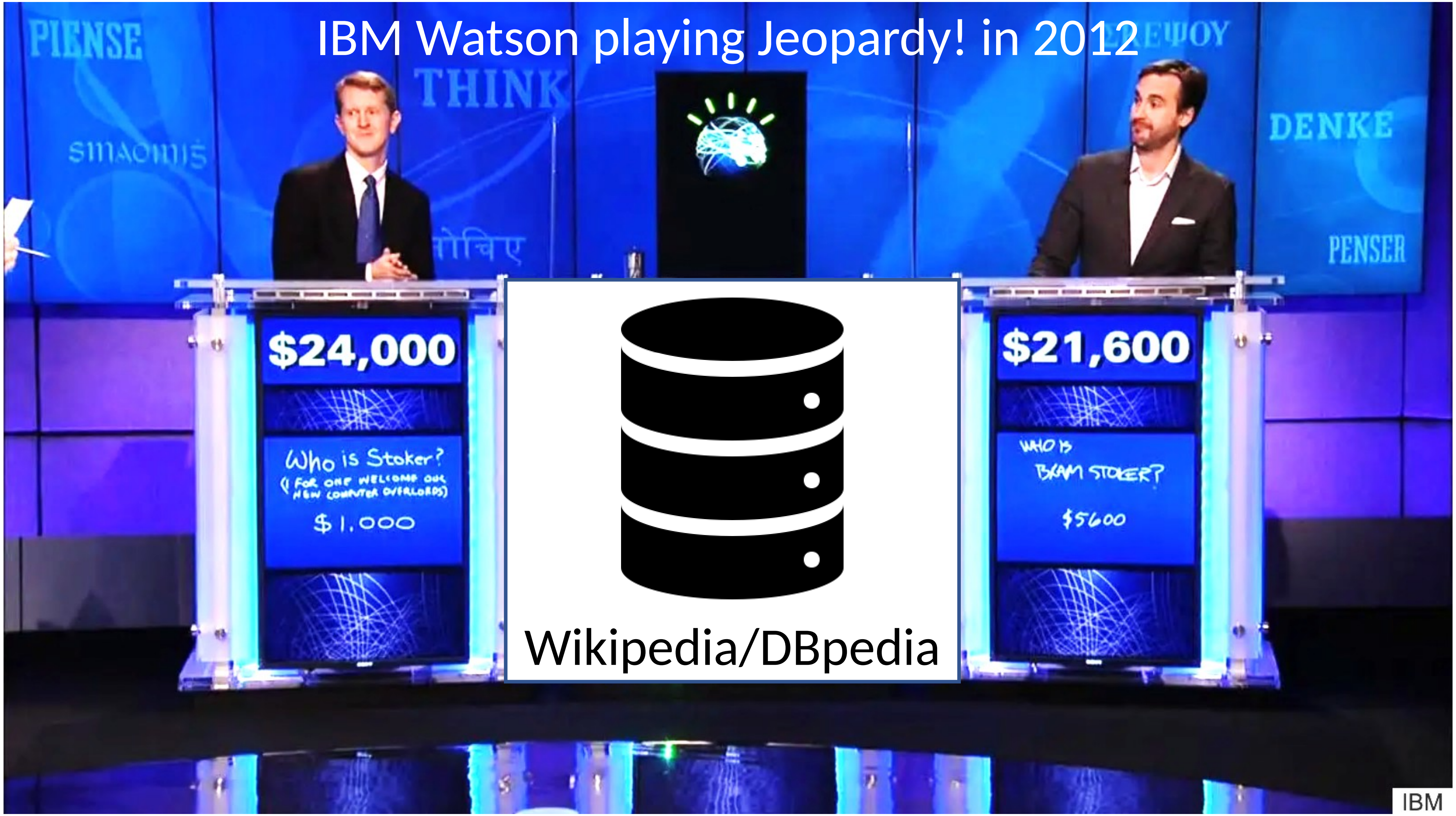
```
(1.2) "The object falls for the height of the cliff."
isa(fall01,fall_v1),
height(cliff01,height01),
agent(fall01,object01),
distance(fall01,height01).
```

This naturally combines with knowledge of physics, expressed as rules

```
(1.4) "What is the duration of the fall?"
isa(fall01,fall_v1),
isa(duration01,duration_n1),
duration(fall01,duration01),
query-for(duration01).
```

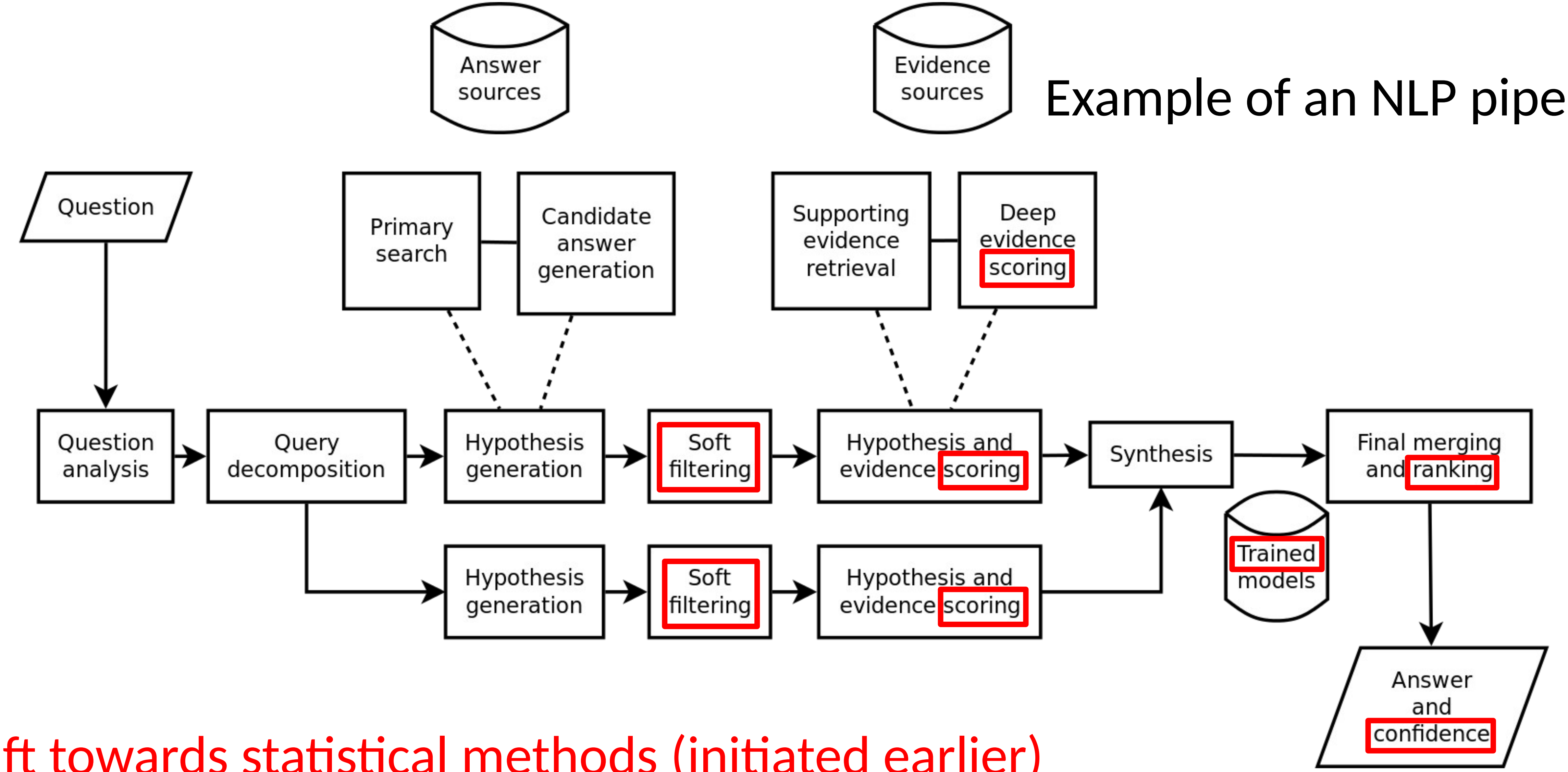
How do we get those rules?
- Lots of research has gone into devising these expert systems
- Human experts are not reliable enough

IBM Watson


IBM Watson playing Jeopardy! in 2012
PIENSE
THINK
DENKE
PENSER
$24,000
Who is Stoker?
(I For one welcome our new computer overlords)
$1,000
$21,600
WHO IS
BRAM STOKER?
$5600
Wikipedia/DBpedia
IBM

IBM Watson

Example of an NLP pipeline



Shift towards statistical methods (initiated earlier)

Google

1996 PageRank and Google search engine

- Usual search engines would count keywords

- Larry Page and Sergey Brin came up with something even better for the web

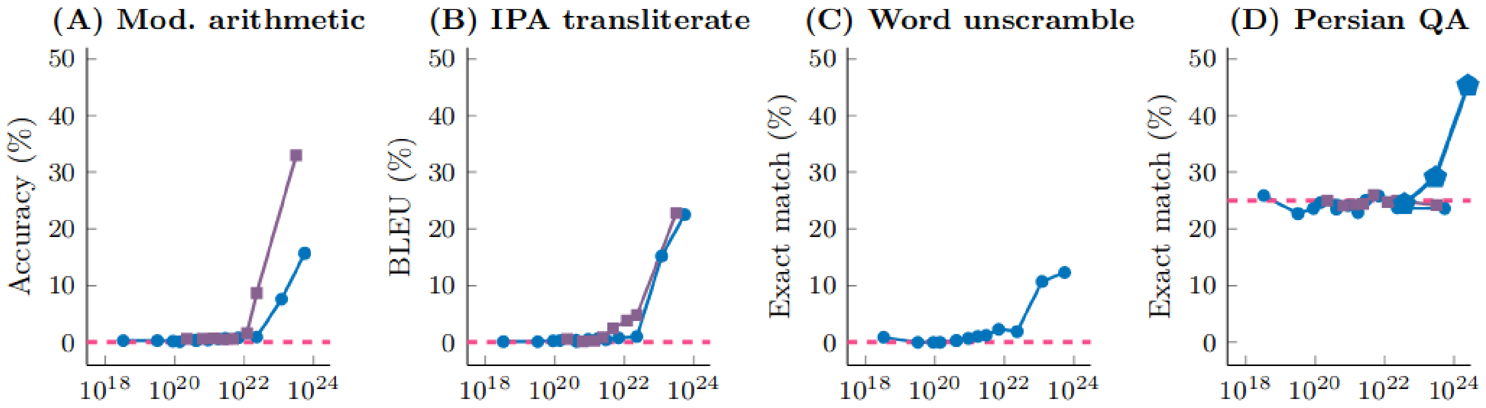
- This is based on their PhD research and relies on statistical methods (as opposed to symbolic methods)

- Statistical methods rely on data, and the internet was large enough for this

Generative Pre-trained Transformers

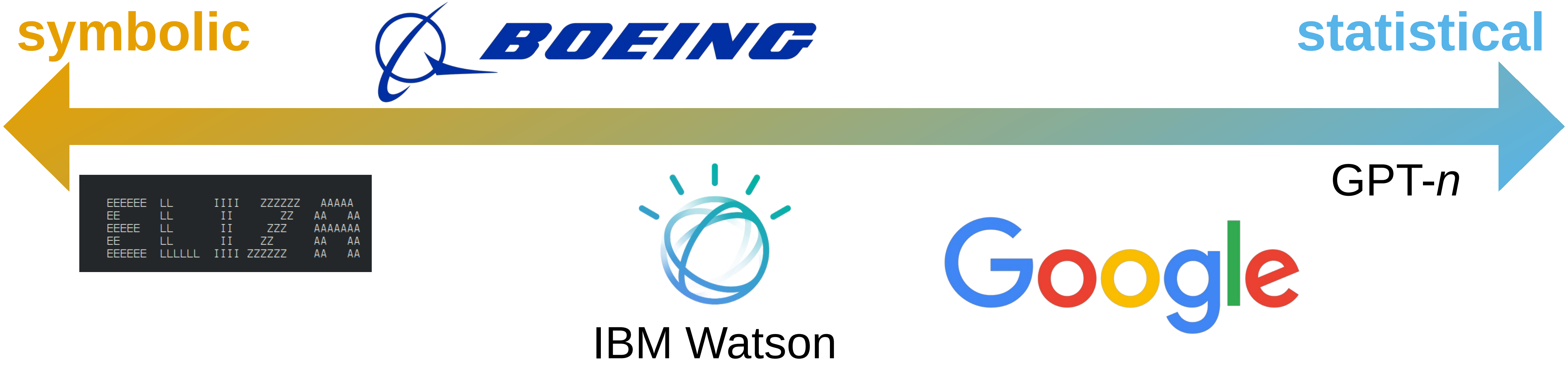
- 2018 GPT-1
- 2019 GPT-2
- 2020 GPT-3
- 2022 GPT-3.5
- 2023 GPT-4

- Newton's equations and their parameters explain the motions of planets
- GPT-*n* models (and their parameters) explain the text found on the internet
- Surprisingly, this class of models can be used to solve NLP problems
- More surprisingly, they have emergent abilities



Wei *et al*, *Emergent Abilities of Large Language Models*, TACL 2022

Symbolic vs statistical NLP



- This is only one way of viewing the evolution of NLP
- Shift enabled by more compute and available data (the Internet)
- Systems on different points of the spectrum coexist

Advantages of symbolic NLP

- Interpretability
  How did you arrive at that conclusion?
- Generalization
  - Conjugate a new verb
  - Acquire new words
- Incorporate domain-specific knowledge
  - Knowledge of physics
  - Specialized vocabulary
- Knowledge can be audited and edited

Advantages of statistical NLP

- Historically a lot of progress has come from statistical models
- More straightforward to model available observations than to keep guessing for new theories
- You can skip feature design (more on this later)
- This doesn't exclude that someday we'll find a theory of language that is useful for NLP
  Tycho Brahe → Johannes Kepler → Isaac Newton

"Anytime a linguist leaves the group the recognition rate goes up" – Fred Jelinek

Probabilities and log-linear models

- Probabilities is the dominant modeling paradigm in current NLP
- Focus on:
    - Probabilities for NLP
    - Log-linear models
    - Parameter estimation
    - Gradient descent

Probabilities

- What is a probability?

    If I toss a coin, what is the probability it lands on tails?

    If I throw two even dice, what is the probability of a double six?

- Frequency of outcome if I toss the same coin 10,000 times?
- Measurement of my belief that the coin will land on tails?

I’m playing poker and my opponent gets a royal flush three rounds in a row. Is my opponent cheating?

- (Money) Odds in a bet

Random variables

- A random variable is a function that maps the outcome of an experiment to a value

Coin-flipping experiment:

X = {“the coin lands on heads” -> X = 1,

“the coin lands on tails” -> X = 0}

Poker game:

Y = {“my opponent cheated” -> Y = 1,

“my opponent did not cheat” -> Y = 0}

Z = {“my opponent is dealt a royal flush” -> Z = 1, …}

- We can reason about the probability of X = 1, noted p(X=1)

What is the proportion of outcomes that would result in X = 1?

Random variables

- Random variables are not random
- Random variables are not variables
- Random variables are functions
- Random variables are deterministic
- The randomness comes from the outcome
- A random variable deterministically maps an outcome to a value

Adapted from Ryan Cotterell's Introduction to NLP

Probabilities in NLP

Outcome space = emails received by Nils

Random variable X = 1 if email is spam, 0 otherwise

X = 1	X = 1	X = 0
Alzheimer : Ces petits signes qui DOIVENT vous alerter ! Dès 50 ans, votre cerveau perd chaque jour un peu plus de son potentiel. Moins vif, moins efficace, moins fiable… jour après jour, votre cerveau RÉGRESSE ! Comme 73% des français, vous laissez passer certains...	Vendre sa voiture rapidement et au meilleur prix ! La vente de votre voiture : étape par étape 1. Évaluation en ligne – Évaluez votre voiture gratuitement ! Complétez les informations de votre véhicule sur notre outil d’évaluation en ligne.	Hi all, sorry for duplicate sending but there will be this seminar today by Petr Kuznetsov When 09-06 starting at 1.30 pm (duration ~1h) Where Telecom Paris, room 4D19 zoom link https://telecom-paris.zoom.us/j/9709503070...

Probabilities in NLP

Outcome space = emails received by Nils

Random variable X = 1 if email is spam, 0 otherwise

X = ?

Colleague,

Submit your manuscript to Systems and Soft Computing and you'll benefit from a 50% discount on the article publishing charge (APC) if your paper is accepted before the end of this year. As an established open access journal, we welcome cutting-edge research papers and short communications on...

We can compute how probable it is that this email is spam, i.e. compute p(X=1)

Why is this better than a rule-based system?

- Examples of spam can be contradictory:
    - Not always clear-cut
    - Noisy judgments
- The system might express degrees of belief
- Probabilities can nicely combine multiple sources of information (e.g. prior knowledge)

Probabilities in NLP

Outcome space = strings of characters

Random variable X = 1 if the sentence is proper English, 0 otherwise

Borislav lab programs	X = 0
Borislav programs in the lab	X = 1
Frederica programs in the lab with Borislav	X = 1
Borislav Borislav Borislav	X = 0

This task can be done entirely without probabilities

In the second lecture you'll see how to do this with probabilities

Adapted from Ryan Cotterell's *Introduction to NLP*

Useful properties

Non-negativity

$$\forall x \in D,\ p(X = x) \geq 0 \qquad \forall x \in D, f(x) \geq 0$$

Sums to 1

$$\sum_{x \in D} p(X = x) = 1 \qquad \int_{x \in D} f(x)\mathrm{d}x = 1$$

Additivity

$$p(\{X \in A\} \cup \{X \in B\}) = p(\{X \in A\}) + p(\{X \in B\})$$

$$\text{if } A \cap B = \emptyset$$

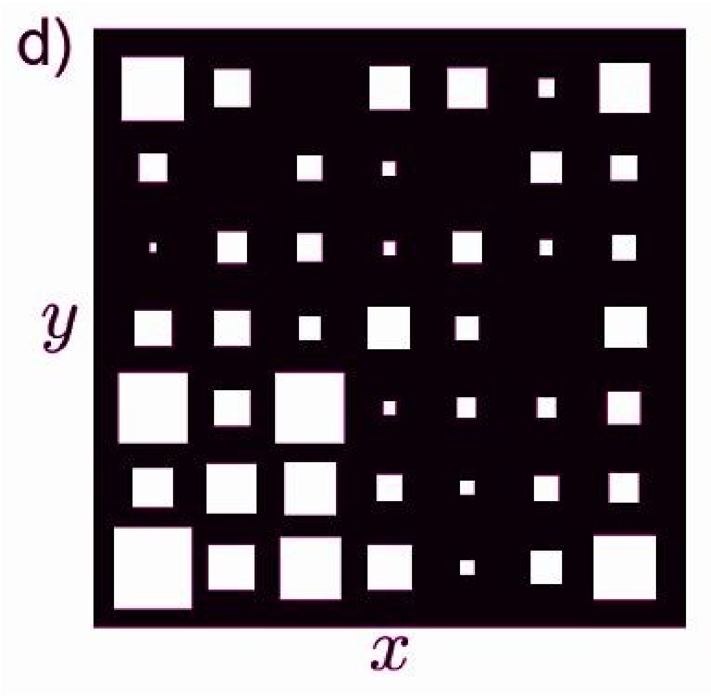
Useful properties

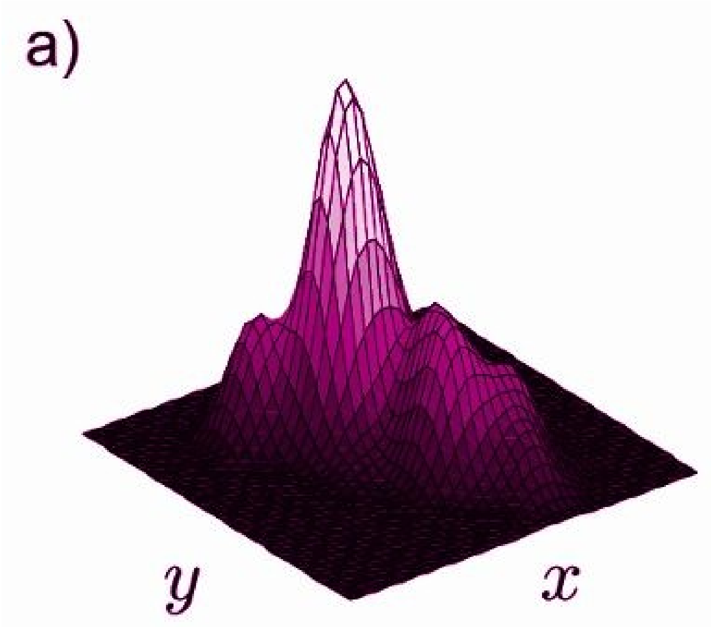
Joint probabilities

"probability that *X*=*x* and *Y*=*y* at the same time"

e.g. *X* is a question, *Y* is an answer

$$p(X=x, Y=y) \stackrel{\text{def}}{=} p(\{X=x\} \cap \{Y=y\})$$



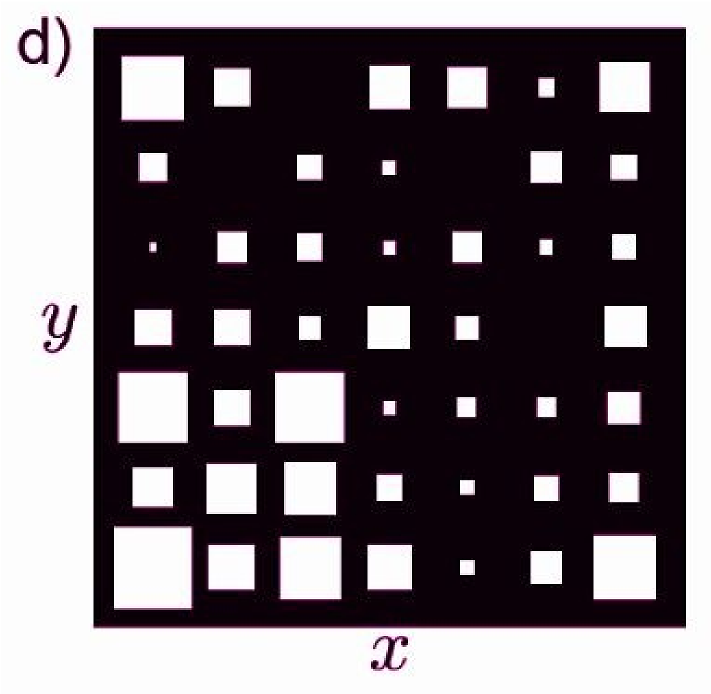


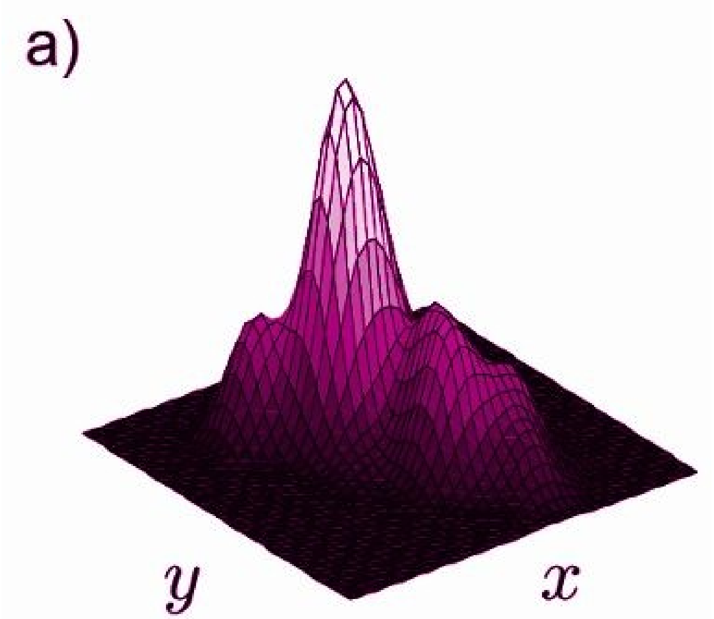
Adapted from Ryan Cotterell's *Introduction to NLP*

Useful properties

Marginalization

$$p(X = x) = \sum_{y \in D_Y} p(X = x, Y = y)$$





Adapted from Ryan Cotterell's *Introduction to NLP*

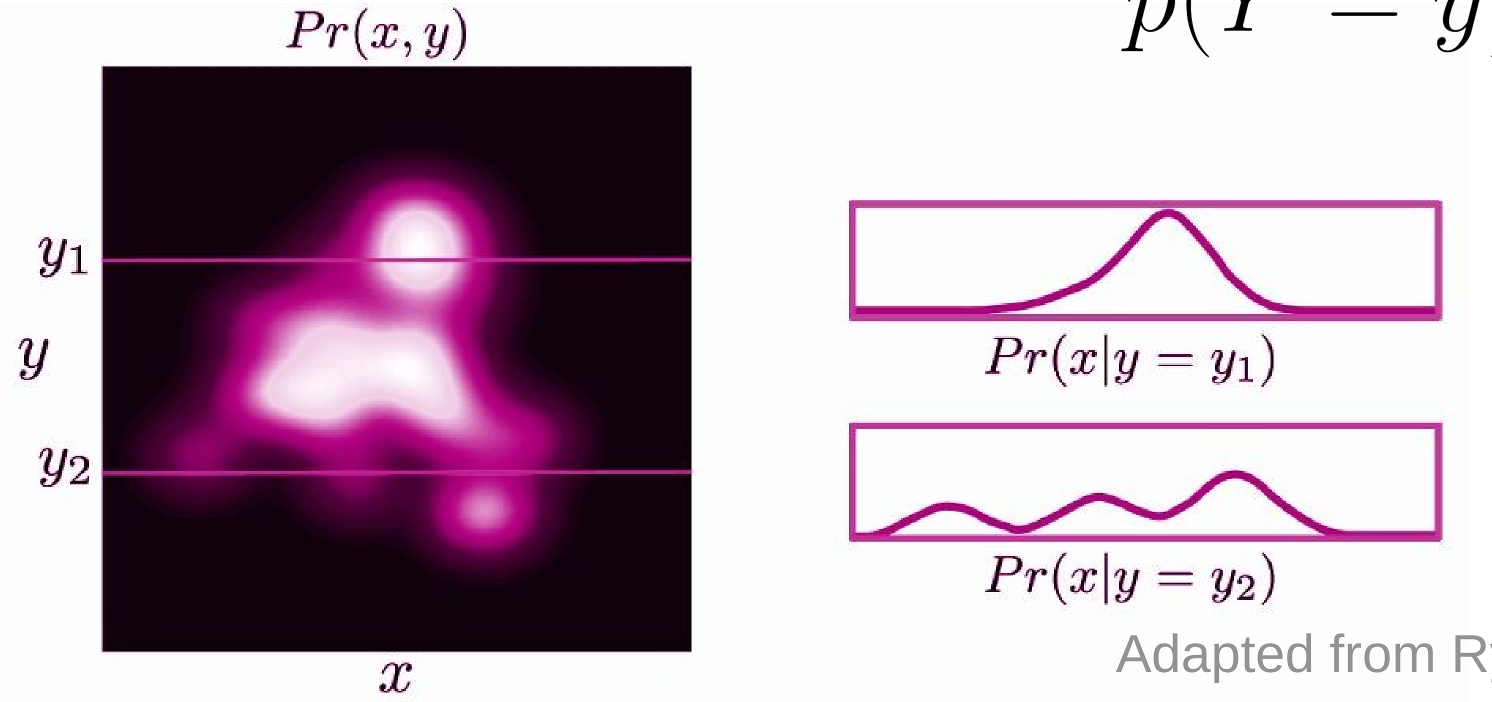
Useful quantities

Conditional probability

"probability of *X*=*x* given *Y*=*y*"

e.g. probability that the answer is *y* given that the question is *x*?

$$p(X = x | Y = y) \stackrel{\text{def}}{=} \frac{p(X = x, Y = y)}{p(Y = y)}$$



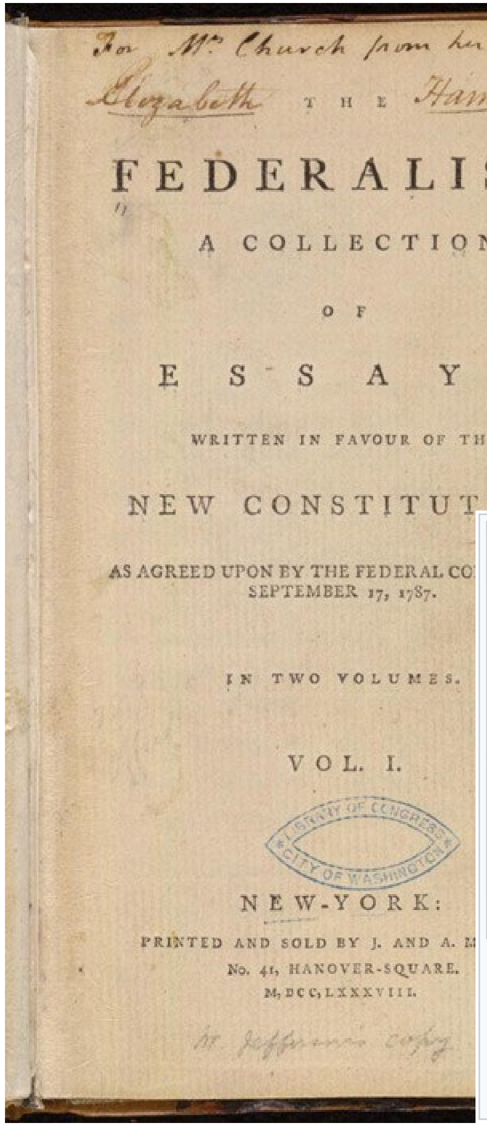
Adapted from Ryan Cotterell's *Introduction to NLP*

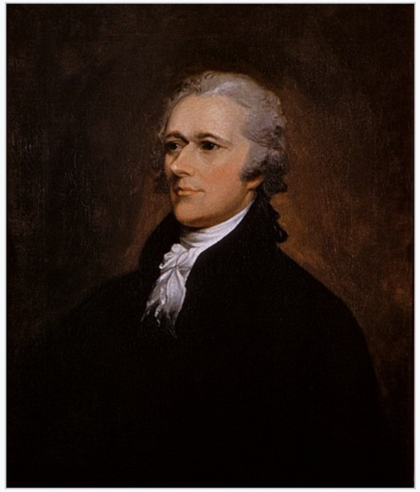
Bayes' rule

Definition of conditional probabilities + marginalization:

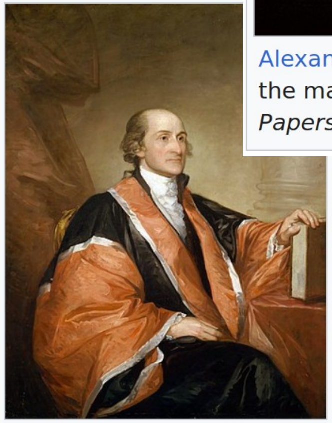
$$p(Y = y|X = x) = \frac{p(X = x|Y = y)p(Y = y)}{\sum_z p(X = x|Y = z)p(Y = z)}$$

Bayes' rule

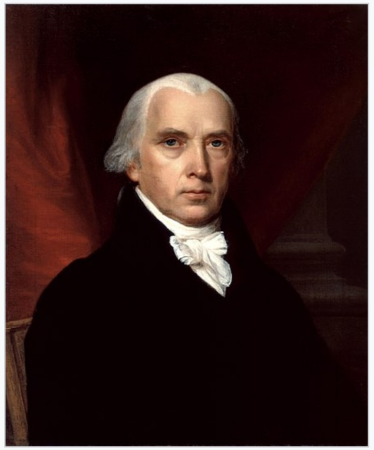

For Mrs Church from her Sister
Elizabeth THE Hamilton
FEDERALIST:
A COLLECTION
OF
ESSAYS,
WRITTEN IN FAVOUR OF THE
NEW CONSTITUTION,
AS AGREED UPON BY THE FEDERAL CO
SEPTEMBER 17, 1787.
IN TWO VOLUMES.
VOL. I.
NEW-YORK:
PRINTED AND SOLD BY J. AND A. M
No. 41, HANOVER-SQUARE.
M,DCC,LXXXVIII.



Alexander Hamilton, author of the majority of *The Federalist Papers*

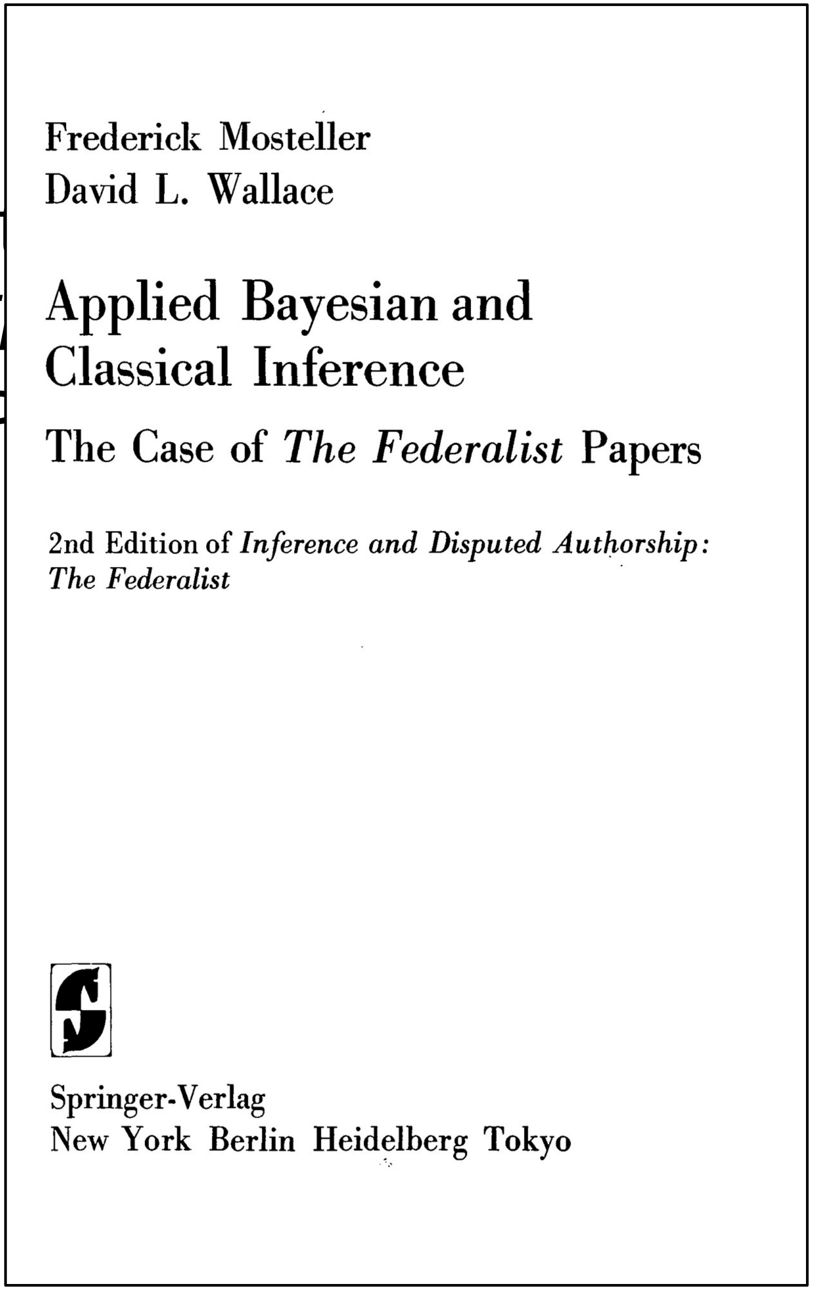


John Jay, author of five of *The Federalist Papers*, later became the first chief justice of the United States



James Madison, Hamilton's major collaborator, later fourth president of the United States (1809-1817)

Who wrote t
essays in *Th*
Federalist P


Frederick Mosteller
David L. Wallace

Applied Bayesian and
Classical Inference

The Case of *The Federalist* Papers

2nd Edition of *Inference and Disputed Authorship:*
The Federalist

Springer-Verlag
New York Berlin Heidelberg Tokyo

Bayes' rule

Who wrote the essays in *The Federalist Papers*?

Experiment: pick an essay from *The Federalist Papers*

A = author of the essay

X = text of the essay

We want to know $p(A = \text{``Hamilton''} | X = x)$

We can estimate $p(X = x | A = \text{``Hamilton''})$

by picking at random something Hamilton wrote

Prior probability

Likelihood

$$\boxed{p(A = \text{``H''} | X = x)} = \frac{\boxed{p(X = x | A = \text{``H''})}\boxed{p(A = \text{``H''})}}{\boxed{p(X = x | A = \text{``H''})}\boxed{p(A = \text{``H''})} + \boxed{p(X = x | A = \text{``M''})}\boxed{p(A = \text{``M''})}}$$

Posterior probability

(John Jay left out for space)

Bayes’ rule

Frederick Mosteller
David L. Wallace

Applied Bayesian and Classical Inference

The Case of *The Federalist* Papers

2nd Edition of *Inference and Disputed Authorship: The Federalist*

Springer-Verlag
New York Berlin Heidelberg Tokyo

Expectation

Let *X* be a random variable with real values, e.g.

- the number of words in an email
- whether an email contains the word "proofreading"

$$\mathbb{E}[X] \stackrel{\text{def}}{=} \sum_x x p(x)$$

$$\mathbb{E}[f(X)] = \sum_x f(x) p(x)$$

$$\mathbb{E}[\alpha X + \beta] = \alpha \mathbb{E}[X] + \beta \qquad \mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

Probabilities in NLP

Step 1. Express the quantities of interest as random variables.

e.g. spam classification:

Experiment = I receive an email

X = the email I receive (it's a string)

Y = 1 if the email is spam, 0 otherwise

$\boxed{p(y|x)}$ "Given that I received email *x*, is it spam?"

$p(y)$ "How probable is it that an email I receive should be spam?"

$p(x)$ "How probable is it that I should receive email *x*?"

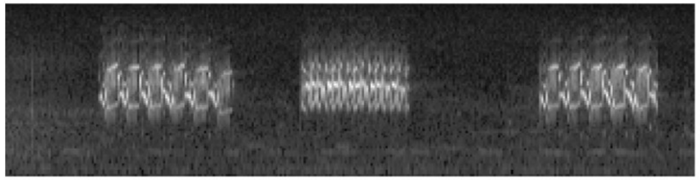
$p(x|y)$ "How probable is it that I should receive email *x*?, assuming that it's spam/not spam?"

Step 2. Compute relevant probabilities.

Probabilities in NLP

How do I compute these probabilities?
Answer: log-linear modeling

$$\boxed{p(y|x)}$$

Task	X	Y
Spam classification	Colleague, Submit your manuscript to Systems and ...	0 (not spam) or 1 (spam)
Sentiment analysis	Gregory Potter ★★★★★ **Having 3 colanders in different sizes is a great option.** Reviewed in the United States on March 22, 2023 **Verified Purchase** The best thing about these colanders is they are stackable, eliminating Also, we'll made!	negative, neutral or positive
BirdNET	 (a) Input spectrogram	

Source: S. Kahl et al, BirdNET: A deep learning solution for avian diversity monitoring, Ecological Informatics 2020

Probabilities as counts

We have examples of emails that are or are not spam

X	*Y*
Alzheimer : Ces petits signes qui DOIVENT vous alerter ! Dès 50 ans, votre cerveau perd chaque jour un peu plus de son potentiel...	1
Vendre sa voiture rapidement et au meilleur prix ! La vente de votre voiture : étape par étape...	1
Hi all, sorry for duplicate sending but there will be this seminar today by Petr Kuznetsov...	0

$$p(y|x) \stackrel{?}{=} \frac{\mathrm{count}(x, y)}{\mathrm{count}(x)}$$

Issue: pretty much every *x* is unique
- Need to store many counts
- Might consistently predict 0

Probabilities as scores

We have examples of emails that are or are not spam

X	*Y*
Alzheimer : Ces petits signes qui DOIVENT vous alerter ! Dès 50 ans, votre cerveau perd chaque jour un peu plus de son potentiel...	1
Vendre sa voiture rapidement et au meilleur prix ! La vente de votre voiture : étape par étape...	1
Hi all, sorry for duplicate sending but there will be this seminar today by Petr Kuznetsov...	0

$$p(y|x) \stackrel{?}{=} \mathrm{score}(x, y)$$

Measure how compatible *x* and *y* are

score needs to be designed so that it pays attention to those features or properties or characteristics of *x* relevant to predict *y*

Probabilities as scores

score needs to be designed so that it pays attention to those features of *x* relevant to predict *y:*

- are certain words CAPITALIZED
- does the email contain lots of exclamation marks?!!
- does the email address me with my first name

Make sure to verify the axioms of probabilities:

$$p(y|x) \stackrel{\text{def}}{=} \frac{1}{Z(x)} \exp(\text{score}(x, y))$$

$$Z(x) \stackrel{\text{def}}{=} \sum_{y'} \exp(\text{score}(x, y'))$$

Probabilities as scores

score needs to be designed so that it pays attention to those features of *x* relevant to predict *y:*

- are certain words CAPITALIZED
- does the email contain lots of exclamation marks?!!
- does the email address me with my first name

$$\underbrace{s(x, y)}_{\text{score}} = \sum_{i=1}^{n} \underbrace{\theta_i}_{\text{weight of feature number } i} \underbrace{f_i(x, y)}_{\text{feature number } i}$$

Features

Examples of features:

$$f_1(x, \text{spam}) = 1 \text{ if any word in } x \text{ is capitalized, else } 0$$
$$f_1(x, \text{not spam}) = 0$$

$$f_2(x, \text{spam}) = \text{number of exclamation marks in } x$$
$$f_2(x, \text{not spam}) = 0$$

$$f_3(x, \text{spam}) = 0$$
$$f_3(x, \text{not spam}) = \text{number of exclamation marks in } x$$

$$f_4(x, \text{not spam}) = 1 \text{ if first paragraph of } x \text{ contains "Nils"}$$
$$f_4(x, \text{spam}) = 0$$

Feature functions

$$f(x,y) = \begin{pmatrix} \text{has\_capitalized\_word}(x), & y = \text{spam} \\ \text{has\_capitalized\_word}(x), & y = \text{not spam} \\ \text{count}(x, \text{``!''}), & y = \text{spam} \\ \text{count}(x, \text{``!''}), & y = \text{not spam} \\ \text{contains}(x, \text{``Nils''}), & y = \text{spam} \\ \text{contains}(x, \text{``Nils''}), & y = \text{not spam} \\ \ldots & \end{pmatrix}$$

f maps a document to a real-valued vector

Log-linear model

feature function

$$p(y|x) = \frac{1}{Z(x)} \exp(\theta \cdot f(x, y))$$

normalization factor

feature vector

parameters of the log-linear model

$$f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^n \qquad \theta \in \mathbb{R}^n$$

$$\left.\begin{array}{l} \mathcal{X} = \text{set of strings} \\ \mathcal{Y} = \text{set of labels} \end{array}\right\} \text{usually}$$

Setting the parameters

$$p(y|x) = \frac{1}{Z(x)} \exp(\theta \cdot f(x, y))$$

- Obviously this depends on the choice of the feature function *f*
- Let's assume *f* provides all the relevant information to the problem
- ...so that if we could set the parameters properly, we'd solve the classification problem

Measuring performance

- What does it mean to solve the classification problem?
- How do we know we have the right parameters?
- Presumably *p* would track the real probabilities
- We don’t know the real probabilities, but we have a set of samples drawn from the real probabilities
- Maybe *p* can explain our samples

$$\text{Samples } S = \{\boxed{(x_i, y_i)}\}_{i=1,2,\ldots,N}$$

an email, and whether it is spam or not

Log-likelihood

$$\text{Samples } S = \{(x_i, y_i)\}_{i=1,2,\ldots,N}$$

$$p(S) = p((x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N))$$

The samples were
drawn independently →

Log-likelihood

$$\text{Samples } S = \{(x_i, y_i)\}_{i=1,2,\ldots,N}$$

$$\begin{aligned} p(S) &= p((x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)) \\ &= \prod_{i=1}^{N} p(x_i, y_i) \\ &= \prod_{i=1}^{N} p(y_i | x_i) \prod_{i=1}^{N} p(x_i) \end{aligned}$$

The samples were drawn independently

Log-likelihood

$$\text{Samples } S = \{(x_i, y_i)\}_{i=1,2,\ldots,N}$$

$$\log p(S) = \textstyle\sum_{i=1}^{N} \log p(y_i|x_i) + \displaystyle\sum_{i=1}^{N} \log p(x_i)$$

our log-linear model can define this

does not depend on our model's parameters

Log-likelihood

$$\text{Samples } S = \{(x_i, y_i)\}_{i=1,2,\ldots,N}$$

Maximizing $p(S)$ with our parameters is like maximizing

$$L(\theta) \stackrel{\text{def}}{=} \sum_{i=1}^{N} \log p(y_i | x_i, \theta)$$

the log-likelihood of the data

probability defined by our log-linear model

parameters of the log-linear model

Log-likelihood

$$\text{Samples } S = \{(x_i, y_i)\}_{i=1,2,\ldots,N}$$

Minimize the negative log-likelihood

$$NLL(\theta) \stackrel{\text{def}}{=} -L(\theta) = \sum_{i=1}^{N} -\log p(y_i|x_i,\theta)$$

$$NLL(\theta) = \sum_{i=1}^{N} -\theta \cdot f(x_i, y_i) + \log(\sum_{y' \in \mathcal{Y}} \exp(\theta \cdot f(x_i, y')))$$

The smaller the negative log-likelihood, the better our parameters

The negative log-likelihood is our loss function

Loss function

- There are lots of possibilities to define loss functions
- The NLL is a very common one for statistical models

Desirable property of a loss function	**Does the NLL have that property?**
Continuity (a change in parameters is noticeable)	Yes
Convexity (any local minimum is a global minimum)	Yes
Closed-form solution for minimum	No
Differentiability (see next slide)	Yes

Gradient descent

- How do we find parameters that minimize the NLL?

- In absence of a closed-form solution, use a typical method from optimization: gradient descent

- The gradient of NLL tells us what change in the parameters will most increase the NLL

$$\frac{\partial NLL(\theta)}{\partial \theta_k} = \sum_{i=1}^{N} [\underbrace{-f_k(x_i, y_i)}_{\text{observed}} + \sum_{y' \in \mathcal{Y}} \underbrace{p(y'|x_i, \theta) f_k(x_i, y')}_{\text{expected}}]$$

Gradient descent

$$\theta_k \leftarrow \theta_k - \lambda \frac{\partial NLL(\theta)}{\partial \theta_k}$$

$$\mathbb{E}[f_k(X,Y)|X = x_i]$$

$$\theta_k \leftarrow \theta_k + \lambda [\boxed{\sum_{i=1}^{N} f_k(x_i, y_i)} - \boxed{\sum_{i=1}^{N} \boxed{\sum_{y' \in \mathcal{Y}} p(y'|x_i,\theta) f_k(x_i,y')}}]$$

observed feature count

expected feature count

Gradient descent

$$\theta_k \leftarrow \theta_k + \lambda \Delta_k$$

$$\Delta_k = \sum_{i=1}^{N} f_k(x_i, y_i) - \sum_{i=1}^{N} \sum_{y' \in \mathcal{Y}} p(y' | x_i, \theta) f_k(x_i, y')$$

- This update rule can be intuitive: if our model’s expectation is higher than the true count, the corresponding weight is decreased, which also decreases the expectation; and vice-versa

- Visual explanation: https://www.cs.jhu.edu/~jason/tutorials/loglin/#1

Gradient descent

- For log-linear models, gradient descent will find the global minimum
- What happens there?

$$\forall k, \ \frac{\partial NLL(\theta)}{\partial \theta_k} = 0$$

$$\forall k, \ \sum_{i=1}^{N} f_k(x_i, y_i) = \sum_{i=1}^{N} \sum_{y' \in \mathcal{Y}} p(y'|x_i, \theta) f_k(x_i, y')$$

- True counts and expected counts match up
- This is not a closed form solution nor a perfect solution

More NLP courses

- **CSC_5AI12_TP** *Natural Language Processing*
Matthieu Labeau

- **APM_5AI27_TP** *Large-scale Generative Models for NLP and Speech Processing*
Nils Holzenberger and Mehwish Alam

- **APM_4AI12_TP** *Machine Learning for Text Mining*
Matthieu Labeau

- **APM_5AI29_TP** *Language Models and Structured Data*
Mehwish Alam