Natural Language Processing

anselm

2. Juni 2018

Quelle

Die meisten Folien sind aus dem Kurs *Natural Language Processing* bei Frank Ferraro an der *University of Maryland, Baltimore County - UMBC* im Herbst 2017:
`https://www.csee.umbc.edu/courses/undergraduate/473/f17/`

Inhalt

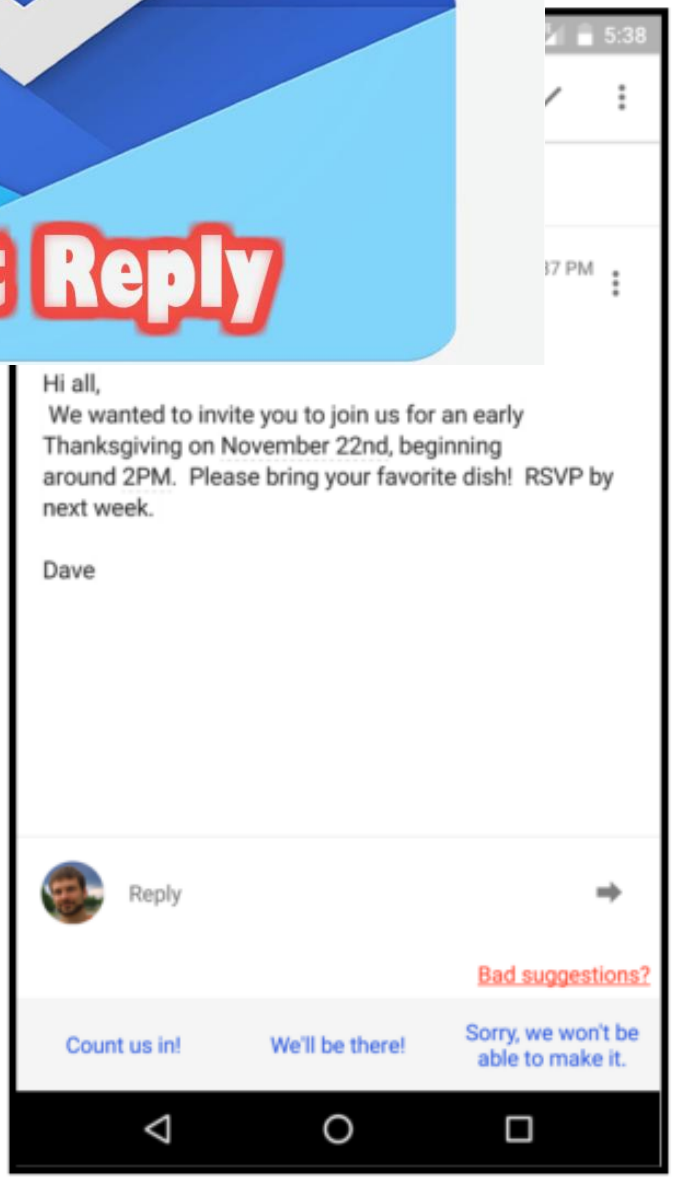
The New York Times Magazine

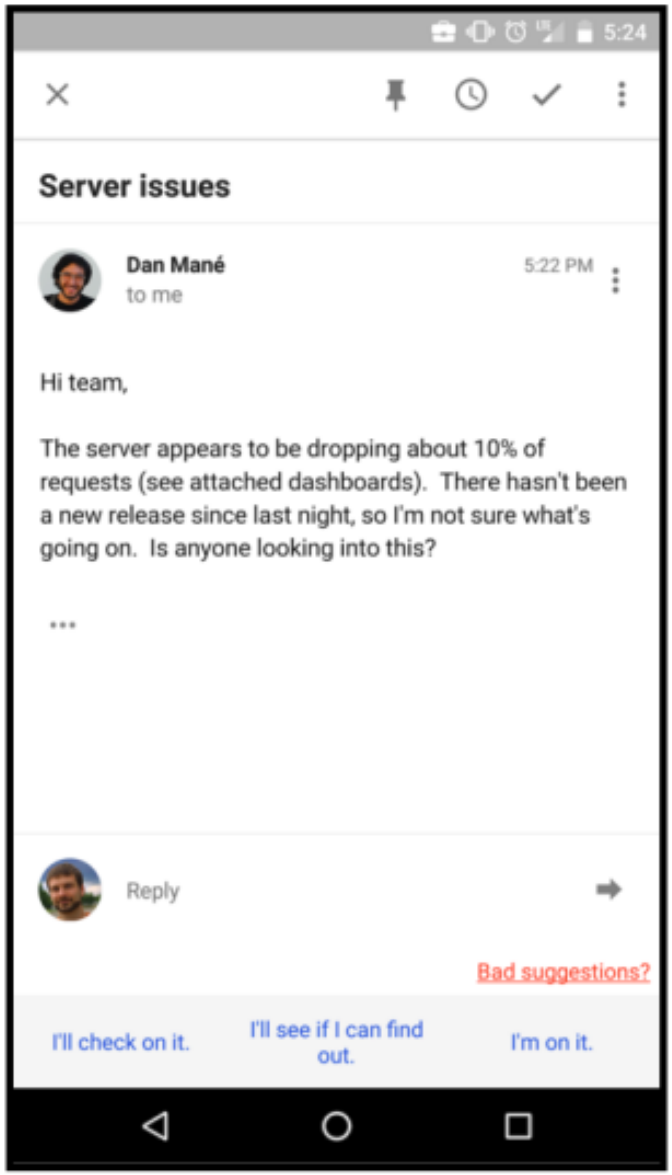
The Great A.I. Awakening

How Google used artificial intelligence to transform Google Translate, one of its more popular services — and how machine learning is poised to reinvent computing itself.

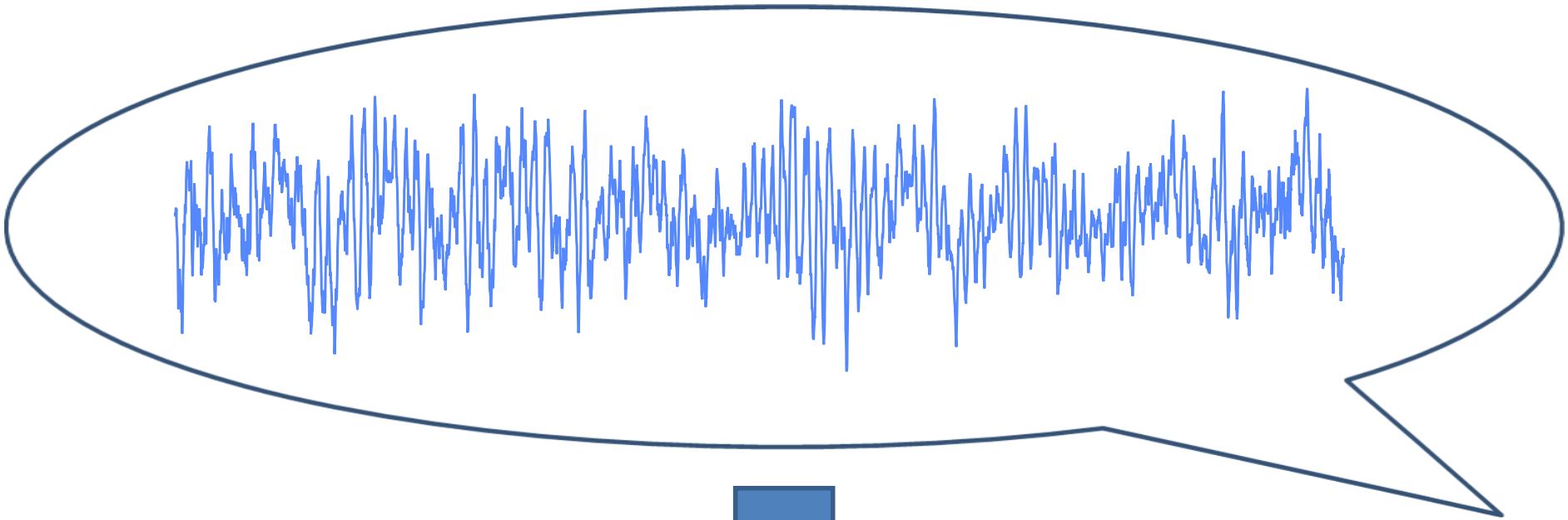
BY GIDEON LEWIS-KRAUS DEC. 14, 2016

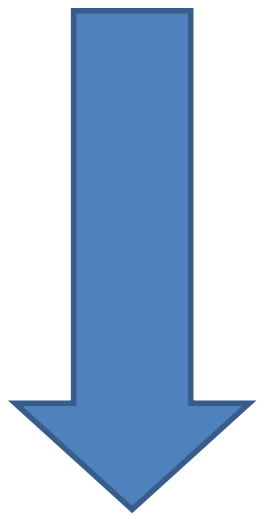
Google Translate





https://cdn.arstechnica.net/wp-content/uploads/2015/11/Screen-Shot-2015-11-02-at-9.11.40-PM-640x543.png





SPORTS

A massive climate change study is canceled ... because of climate change

By **Doug Criss**, CNN
Updated 10:37 AM ET, Tue June 20, 2017

Story highlights

Arctic sea ice has traveled farther south than normal along Newfoundland's northeast coast

An icebreaker has been repeatedly diverted to take part in rescue operations

(CNN) — A $17 million study of climate change in the Canadian Arctic has been nixed for now -- because of climate change.

A team of scientists from the University of Manitoba and four other schools were in the middle of the first leg of a four-year study of how climate change is affecting the areas around the Hudson Bay, the university said in statement. The study, named BaySys, started last month, and the scientists were traveling on the Canadian Research Icebreaker CCGS Amundsen.

But because of warmer temperatures in the Arctic, hazardous sea ice is traveling farther south than usual. The Amundsen, which is part of the Canadian Coast Guard fleet, has been diverted several times because its ice-breaking capabilities have been needed to help out in rescue efforts along Newfoundland's northeast coast. All of the delays and concerns about safety forced

Top stories

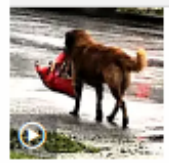


Photo of dog ca viral

North Korea lau amidst US-Sou

Course Goals

Be introduced to some of the core problems and solutions of NLP (big picture)

Course Goals

Be introduced to some of the core problems and solutions of NLP (big picture)

Learn different ways that success and progress can be measured in NLP

Course Goals

Be introduced to some of the core problems and solutions of NLP (big picture)

Learn different ways that success and progress can be measured in NLP

Relate to statistics, machine learning, and linguistics

Implement NLP programs

Natural Language Processing

≈

Computational Linguistics

Natural Language Processing

≈

Computational Linguistics

science focus

computational bio
computational chemistry
computational *X*

build a system to translate
create a QA system

engineering focus

Natural Language Processing

≈

Computational Linguistics

science focus

computational bio
computational chemistry
computational *X*

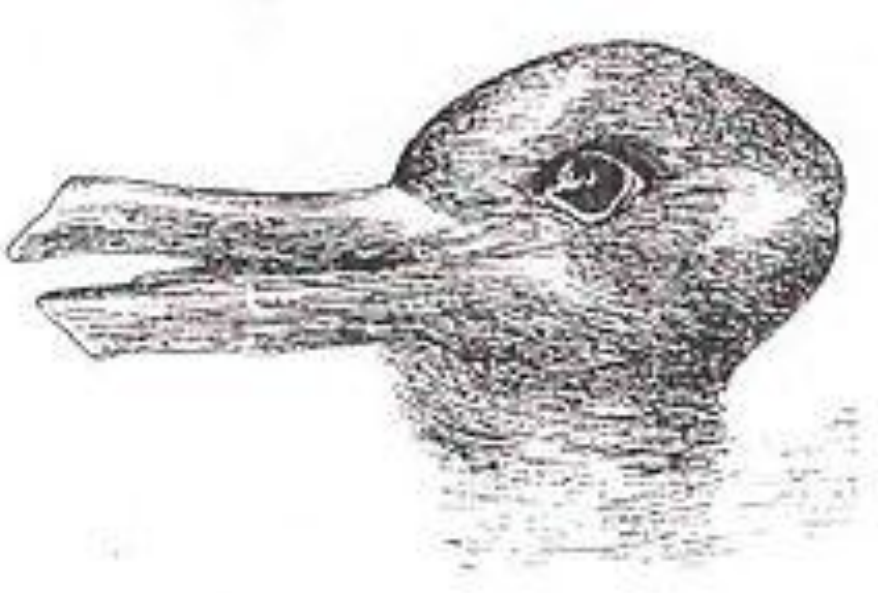
Language is Productive

Troopergate

Watergate ➔ Bridgegate

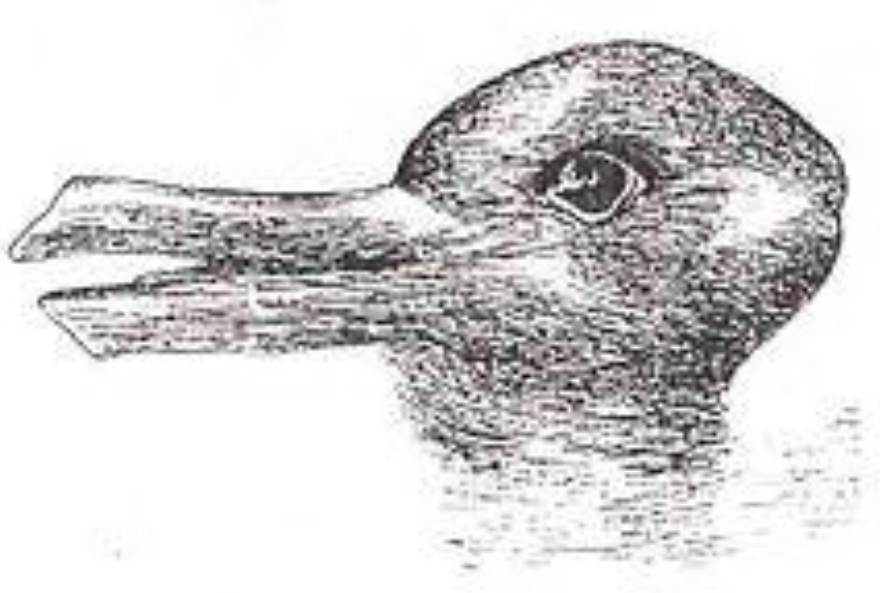
Deflategate

Language is Ambiguous



Ambiguity

Kids Make Nutritious Snacks

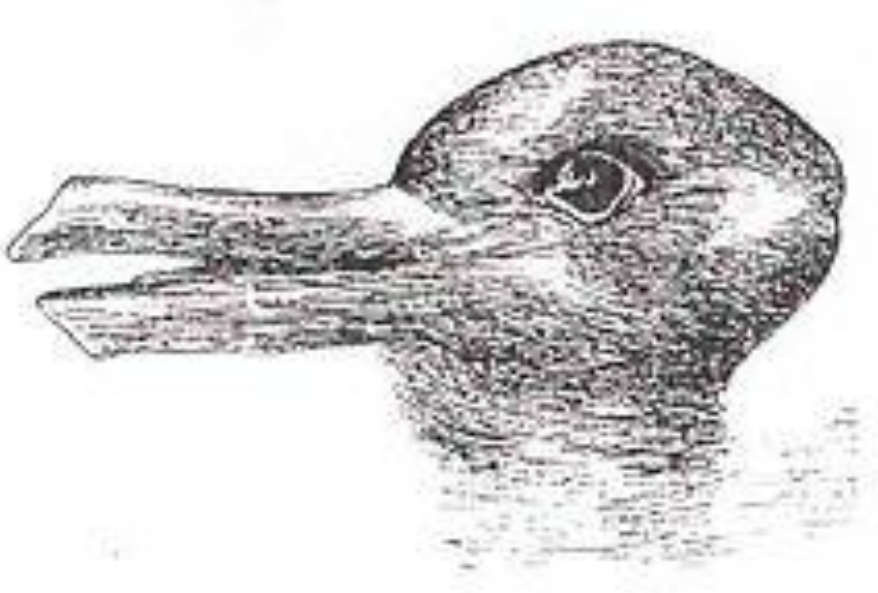


Ambiguity

Kids Make Nutritious Snacks

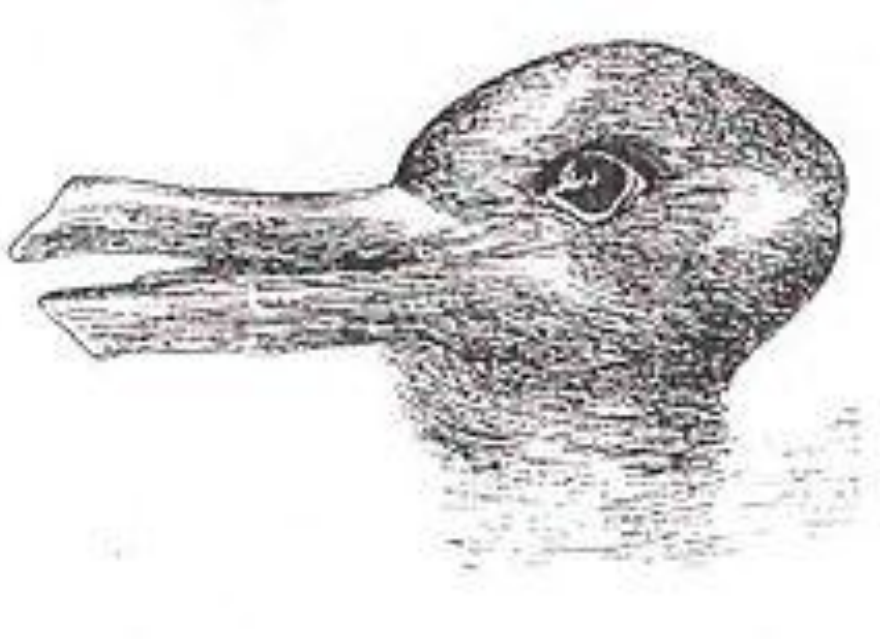
Kids *Prepare* Nutritious Snacks

Kids *Are* Nutritious Snacks



Ambiguity

British Left Waffles on Falkland Islands

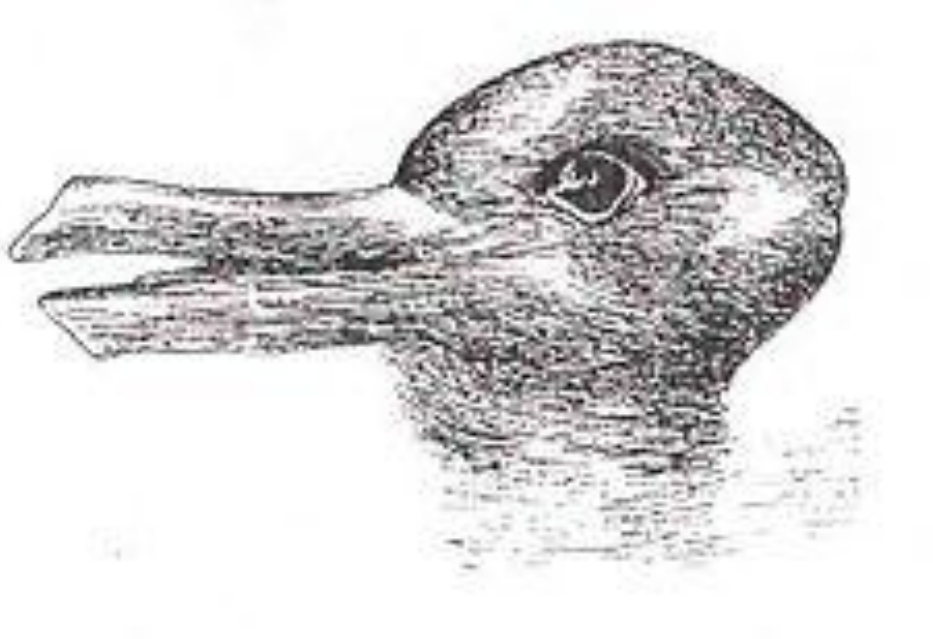


Ambiguity

British Left Waffles on Falkland Islands

British Left Waffles on Falkland Islands

British Left Waffles on Falkland Islands



Part of Speech Tagging

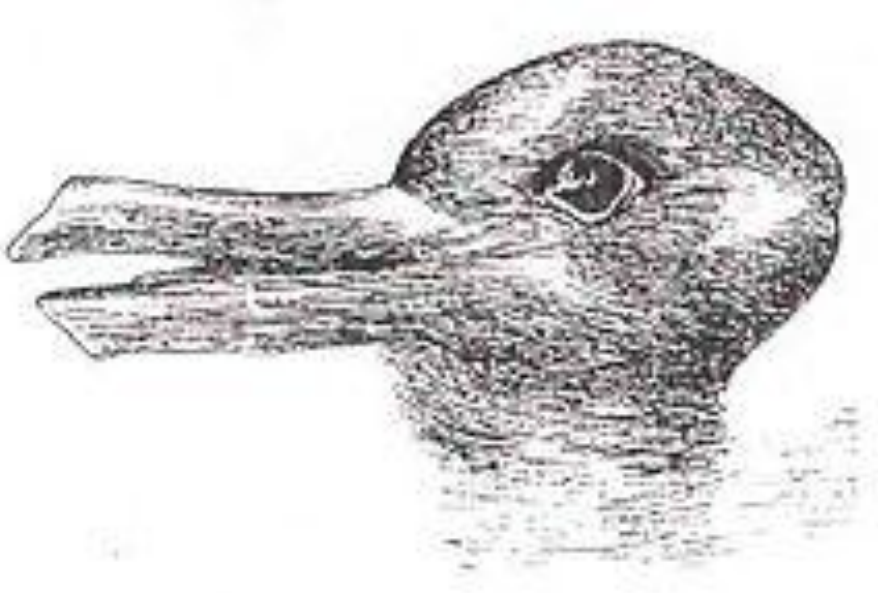
British Left Waffles on Falkland Islands

British Left Waffles on Falkland Islands

Adjective *Noun* *Verb*

British Left Waffles on Falkland Islands

Noun *Verb* *Noun*



Ambiguity

Pat saw Chris with the telescope on the hill.

“Der Alte bedient das boot.”

“The old man the boat.”

Language Can Be Surprising

Garden Path Sentences

Garden Path Sentences

The

Garden Path Sentences

The old

Garden Path Sentences

The old man

Garden Path Sentences

The old man the

Garden Path Sentences

The old man the boat

Garden Path Sentences

The old man the boat .

Garden Path Sentences

The old man the boat .

Garden Path Sentences

The complex houses married and single soldiers and their families.

Garden Path Sentences

The complex houses married and single soldiers and their families.

Discourse Processing through Coreference

I spread the cloth on the table to protect it.

I spread the cloth on the table to display it.

Discourse Processing through Coreference

I spread the cloth on the table to protect **it**.

I spread the cloth on the table to display **it**.

Courtesy Jason Eisner

Discourse Processing through Coreference

I spread the cloth on the table to protect it.

I spread the cloth on the table to display it.

Courtesy Jason Eisner

Discourse Processing

John thought a coffee was good every few hours.

Discourse Processing

John thought a coffee was good every few hours.

(coffee should be consumed very often)

Discourse Processing

John thought a coffee was good every few hours.

(coffee gets cold & stale after a while)

Inhalt

Stochastik

- Wahrscheinlichkeitsfunktion: $p : \Omega \to [0, 1]$
  - $\sum_{i \in \Omega} p(i) = 1$
- Beispiele: Würfel, Münze
- Unabhängige Ereignisse:
  - $p(A \text{ und } B) = p(A) \cdot p(B)$

Beispiel Bedingte Wahrscheinlichkeiten: Würfel

- $p(X \text{ gerade}) =$

Beispiel Bedingte Wahrscheinlichkeiten: Würfel

- $p(X \text{ gerade}) = \frac{1}{2}$, $p(X \geq 4) =$

Beispiel Bedingte Wahrscheinlichkeiten: Würfel

- $p(X \text{ gerade}) = \frac{1}{2}$, $p(X \geq 4) = \frac{1}{2}$
- $p(X \text{ gerade} | X \geq 4) =$

Beispiel Bedingte Wahrscheinlichkeiten: Würfel

- $p(X \text{ gerade}) = \frac{1}{2}$, $p(X \geq 4) = \frac{1}{2}$
- $p(X \text{ gerade} | X \geq 4) = \frac{2}{3}$

Bayes Rule

$$p(X \mid Y) = \frac{p(Y \mid X) * p(X)}{p(Y)}$$

Inhalt

Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today.

score(Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today.)

$$p_\theta(\text{Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today.})$$

p_θ(Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today.)

what's a probability?

$p_\theta($ Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today. $)$

what do we estimate?

Documents? Sentences? Words? Characters?

p_θ(Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today.)

what's a word?

how to deal with morphology and orthography

$$p_\theta(\text{Tree people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of an Shining Path attack today.})$$

how do we estimate robustly?

p_θ(Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of an ISIS attack today.)

how do we generalize?

What Are Words?

Linguists don’t agree

(Human) Language-dependent

White-space separation is a sometimes okay (for written English longform)

What Are Words?

Linguists don't agree

(Human) Language-dependent

White-space separation is a sometimes okay (for written English longform)

Social media? Spoken vs. written? Other languages?

What Are Words?

bat

http://www.freepngimg.com/download/bat/9-2-bat-png-hd.png

What Are Words?

bats

http://www.freepngimg.com/download/bat/9-2-bat-png-hd.png

What Are Words?

Fledermaus

flutter mouse

http://www.freepngimg.com/download/bat/9-2-bat-png-hd.png

What Are Words?

pişirdiler

They cooked it.

What Are Words?

pişmişlermişlerdi

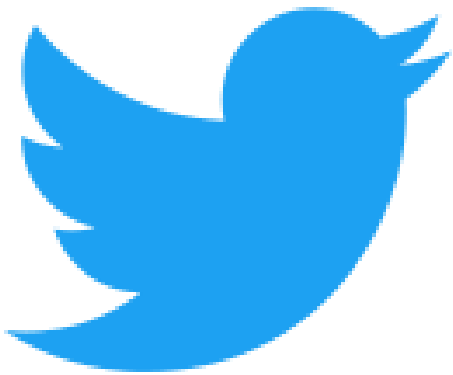
They had it cooked it.

What Are Words?

):

What Are Words?

my leg is hurting nasty):



What Are Words?

add two cups (a pint): bring to a boil

http://www.dummies.com/wp-content/uploads/88513.image4.jpg

What Are Words? Tokens vs. Types

The film got a great opening and the film went on to become a hit .

Type: an element of the vocabulary.

Token: an instance of that type in running text.

How many of each?

What Are Words? Tokens vs. Types

The film got a great opening and the film went on to become a hit .

Types

- The
- film
- got
- a
- great
- opening
- and
- the
- went
- on
- to
- become
- hit
- .

Tokens

- The
- film
- got
- a
- great
- opening
- and
- the
- film
- went
- on
- to
- become
- a
- hit
- .

What Are Words? Tokens vs. Types

The film got a great opening and the film went on to become a hit .

Types

- The
- film
- got
- a
- great
- opening
- and
- the
- went
- on
- to
- become
- hit
- .

Tokens

- The
- film
- got
- a
- great
- opening
- and
- the
- ~~film~~
- went
- on
- to
- become
- ~~a~~
- hit
- .

Some Issues with Tokenization

mph, MPH, M.D.

MD, M.D.

Baltimore's mayor

I'm, won't

state-of-the-art

San Francisco

CaSE inSensitive?

Replace all letters with lower case version

Can be useful for information retrieval (IR), machine translation, language modeling

cat vs Cat (there are other ways to signify beginning)

CaSE inSensitive?

Replace all letters with lower case version

Can be useful for information retrieval (IR), machine translation, language modeling

cat vs Cat (there are other ways to signify beginning)

But… case **can** be useful

Sentiment analysis, machine translation, information extraction

US vs us

cat $\stackrel{?}{=}$ cats

Lemma: same stem, part of speech, rough word sense

cat and cats: same lemma

Word form: the fully inflected surface form

cat and cats: different word forms

Lemmatization

Reduce inflections or variant forms to base form

am, are, is $\rightarrow$ *be*

car, cars, car's, cars' $\rightarrow$ *car*

the boy's cars are different colors $\rightarrow$

the boy car be different color

Inhalt

"The Unreasonable Effectiveness of Recurrent Neural Networks"

"The Unreasonable Effectiveness of Recurrent Neural Networks"

"The Unreasonable Effectiveness of Character-level Language Models"

"The Unreasonable Effectiveness of Recurrent Neural Networks"

"The Unreasonable Effectiveness of Character-level Language Models" (and why RNNs are still cool)

Simple Count-Based

$$p(\text{item})$$

Simple Count-Based

"proportional to"

$$p(\text{item}) \propto count(\text{item})$$

Simple Count-Based

"proportional to"

$$p(\text{item}) \propto count(\text{item})$$

$$= \frac{count(\text{item})}{\sum count(\text{other item})}$$

Simple Count-Based

"proportional to"

$$p(\text{item}) \propto count(\text{item})$$

$$= \frac{count(\text{item})}{\underbrace{\sum count(\text{other item})}_{\text{constant}}}$$

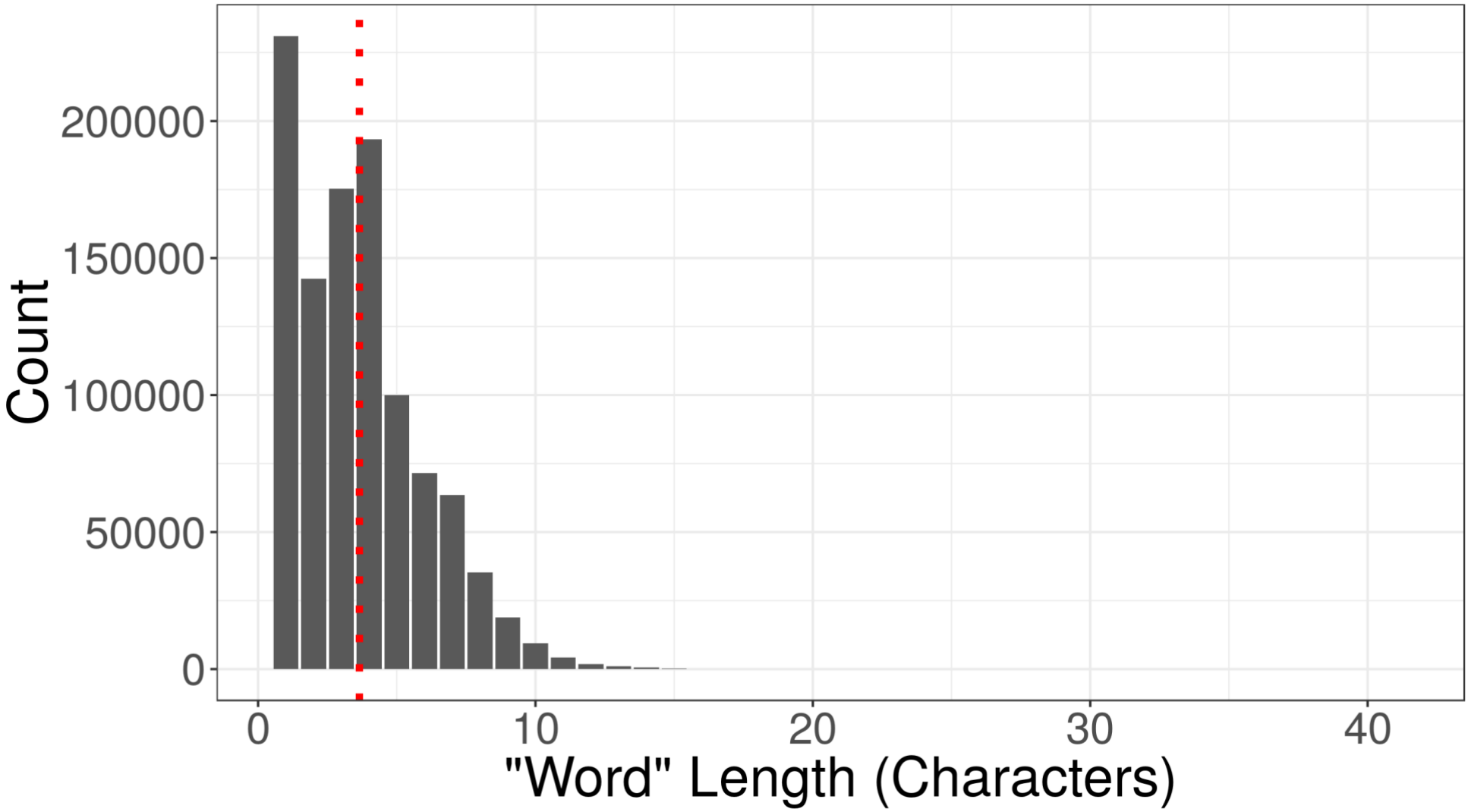
Simple Count-Based

$$p(\text{item}) \propto count(\text{item})$$

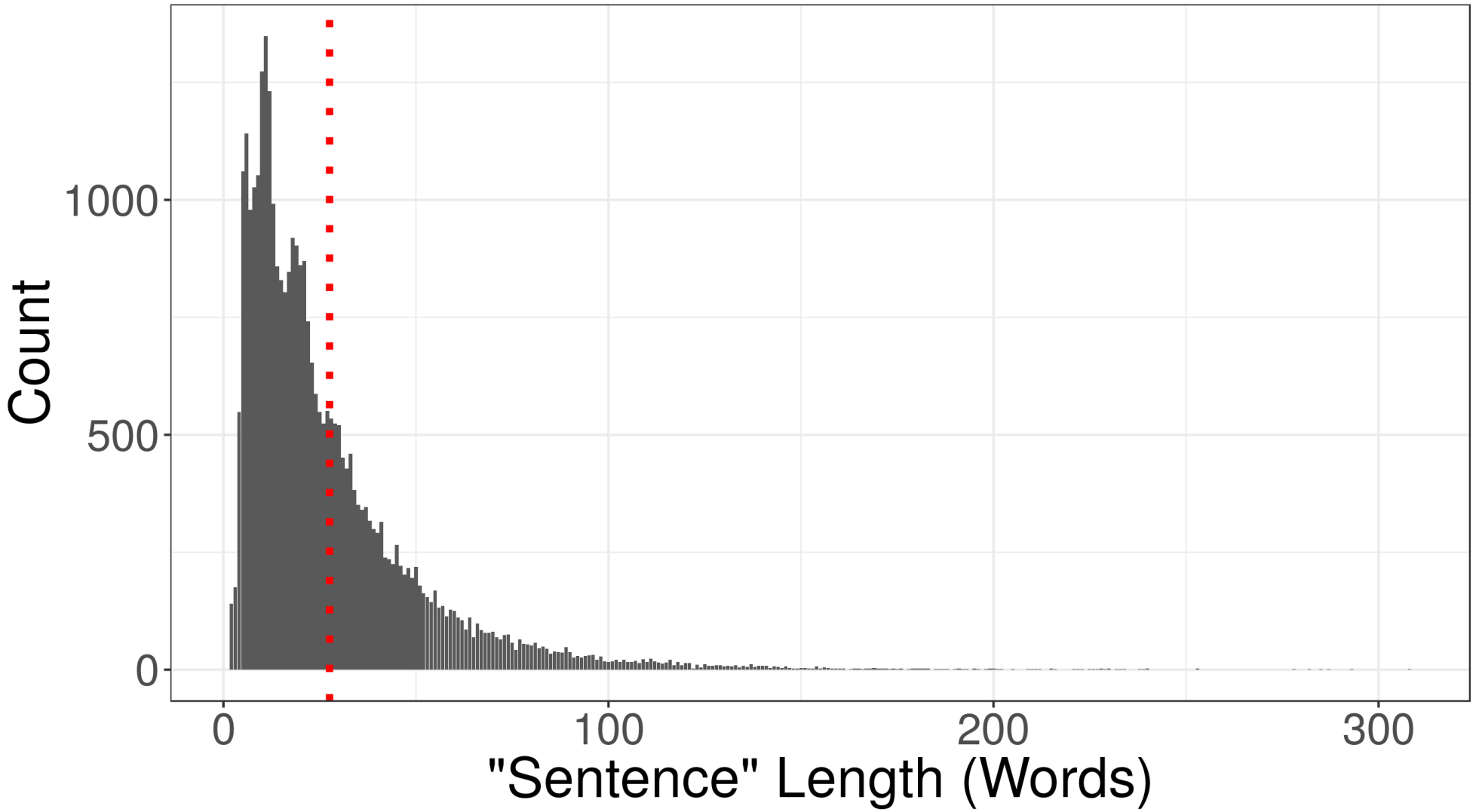
sequence of characters → pseudo-words

sequence of words → pseudo-phrases

Shakespearian Sequences of Characters



Shakespearian Sequences of Words



Novel Words, Novel Sentences

"Colorless green ideas sleep furiously" – Chomsky (1957)

Let's observe and record all sentences with our big, bad supercomputer

Red ideas? Read ideas?

Probability Chain Rule

$$p(x_1, x_2) = p(x_1)p(x_2 \mid x_1)$$

Bayes rule

Probability Chain Rule

$$p(x_1, x_2, \ldots, x_S) =$$
$$p(x_1)p(x_2 \,|x_1)p(x_3 \mid x_1, x_2) \cdots p(x_S|x_1, \ldots, x_i)$$

Probability Chain Rule

$$p(x_1, x_2, \ldots, x_S) =$$

$$p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \cdots p(x_S \mid x_1, \ldots, x_i) =$$

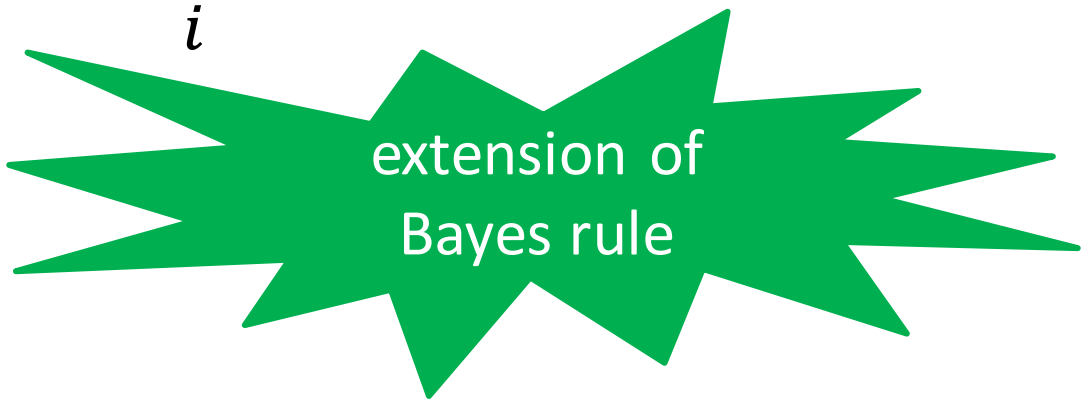
$$\prod_{i}^{S} p(x_i \mid x_1, \ldots, x_{i-1})$$

Probability Chain Rule

$$p(x_1, x_2, \ldots, x_S) =$$

$$p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \cdots p(x_S \mid x_1, \ldots, x_i) =$$

$$\prod_{i}^{S} p(x_i \mid x_1, \ldots, x_{i-1})$$



N-Grams

Maintaining an entire inventory over sentences could be too much to ask

Store “smaller” pieces?

p(Colorless green ideas sleep furiously)

N-Grams

Maintaining an entire *joint* inventory over sentences could be too much to ask

Store “smaller” pieces?

p(Colorless green ideas sleep furiously) =
p(Colorless) *

N-Grams

Maintaining an entire *joint* inventory over sentences could be too much to ask

Store “smaller” pieces?

p(Colorless green ideas sleep furiously) =
p(Colorless) *
p(green | Colorless) *

N-Grams

Maintaining an entire *joint* inventory over sentences could be too much to ask

Store “smaller” pieces?

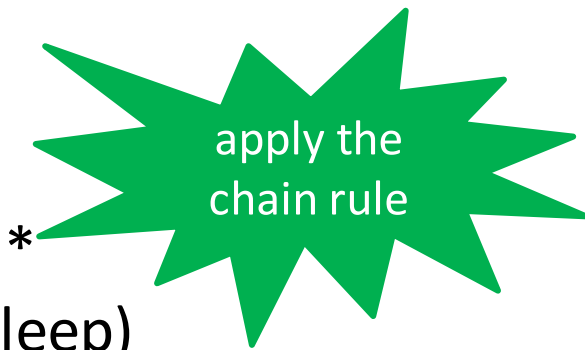
p(Colorless green ideas sleep furiously) =
p(Colorless) *
p(green | Colorless) *
p(ideas | Colorless green) *
p(sleep | Colorless green ideas) *
p(furiously | Colorless green ideas sleep)

N-Grams

Maintaining an entire *joint* inventory over sentences could be too much to ask

Store "smaller" pieces?

p(Colorless green ideas sleep furiously) =
p(Colorless) *
p(green | Colorless) *
p(ideas | Colorless green) *
p(sleep | Colorless green ideas) *
p(furiously | Colorless green ideas sleep)

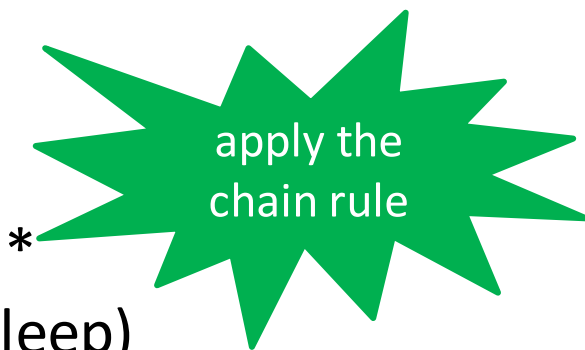


N-Grams

Maintaining an entire *joint* inventory over sentences could be too much to ask

Store “smaller” pieces?

p(Colorless green ideas sleep furiously) =
p(Colorless) *
p(green | Colorless) *
p(ideas | Colorless green) *
p(sleep | Colorless green ideas) *
p(furiously | Colorless green ideas sleep)



N-Grams

p(furiously | Colorless green ideas sleep)

How much does “Colorless” influence the choice of “furiously?”

N-Grams

p(furiously | Colorless green ideas sleep)

How much does “Colorless” influence the choice of “furiously?”

Remove history and contextual info

N-Grams

p(furiously | Colorless green ideas sleep)

How much does “Colorless” influence the choice of “furiously?”

Remove history and contextual info

p(furiously | Colorless green ideas sleep) ≈
p(furiously | ~~Colorless green~~ ideas sleep)

N-Grams

p(furiously | Colorless green ideas sleep)

How much does “Colorless” influence the choice of “furiously?”

Remove history and contextual info

p(furiously | Colorless green ideas sleep) ≈
p(furiously | ideas sleep)

N-Grams

p(Colorless green ideas sleep furiously) =
p(Colorless) *
p(green | Colorless) *
p(ideas | Colorless green) *
p(sleep | Colorless green ideas) *
p(furiously | Colorless green ideas sleep)

N-Grams

p(Colorless green ideas sleep furiously) =
p(Colorless) *
p(green | Colorless) *
p(ideas | Colorless green) *
p(sleep | ~~Colorless~~ green ideas) *
p(furiously | ~~Colorless green~~ ideas sleep)

Trigrams

p(Colorless green ideas sleep furiously) =
p(Colorless) *
p(green | Colorless) *
p(ideas | Colorless green) *
p(sleep | green ideas) *
p(furiously | ideas sleep)

Trigrams

p(Colorless green ideas sleep furiously) =
p(Colorless) *
p(green | Colorless) *
p(ideas | Colorless green) *
p(sleep | green ideas) *
p(furiously | ideas sleep)

Trigrams

p(Colorless green ideas sleep furiously) =
p(Colorless | *<BOS> <BOS>*) *
p(green | *<BOS>* Colorless) *
p(ideas | Colorless green) *
p(sleep | green ideas) *
p(furiously | ideas sleep)

Consistent notation: Pad the left with <BOS> (beginning of sentence) symbols

Trigrams

p(Colorless green ideas sleep furiously) =
p(Colorless | *<BOS> <BOS>*) *
p(green | *<BOS>* Colorless) *
p(ideas | Colorless green) *
p(sleep | green ideas) *
p(furiously | ideas sleep) *
p(*<EOS>* | sleep furiously)

Consistent notation: Pad the left with <BOS> (beginning of sentence) symbols
Fully proper distribution: Pad the right with a single <EOS> symbol

N-Gram Terminology

n	Commonly called	History Size (Markov order)	Example
1	unigram	0	p(furiously)

N-Gram Terminology

n	Commonly called	History Size (Markov order)	Example
1	unigram	0	p(furiously)
2	bigram	1	p(furiously | sleep)

N-Gram Terminology

n	Commonly called	History Size (Markov order)	Example
1	unigram	0	p(furiously)
2	bigram	1	p(furiously | sleep)
3	trigram (3-gram)	2	p(furiously | ideas sleep)

N-Gram Terminology

n	Commonly called	History Size (Markov order)	Example
1	unigram	0	p(furiously)
2	bigram	1	p(furiously | sleep)
3	trigram (3-gram)	2	p(furiously | ideas sleep)
4	4-gram	3	p(furiously | green ideas sleep)
n	n-gram	n-1	$p(w_i \mid w_{i-n+1} \ldots w_{i-1})$

N-Gram Probability

$$p(w_1, w_2, w_3, \cdots, w_S) =$$

$$\prod_{i=1}^{S} p(w_i | w_{i-N+1}, \cdots, w_{i-1})$$

Count-Based N-Grams (Unigrams)

$$p(\text{item}) \propto count(\text{item})$$

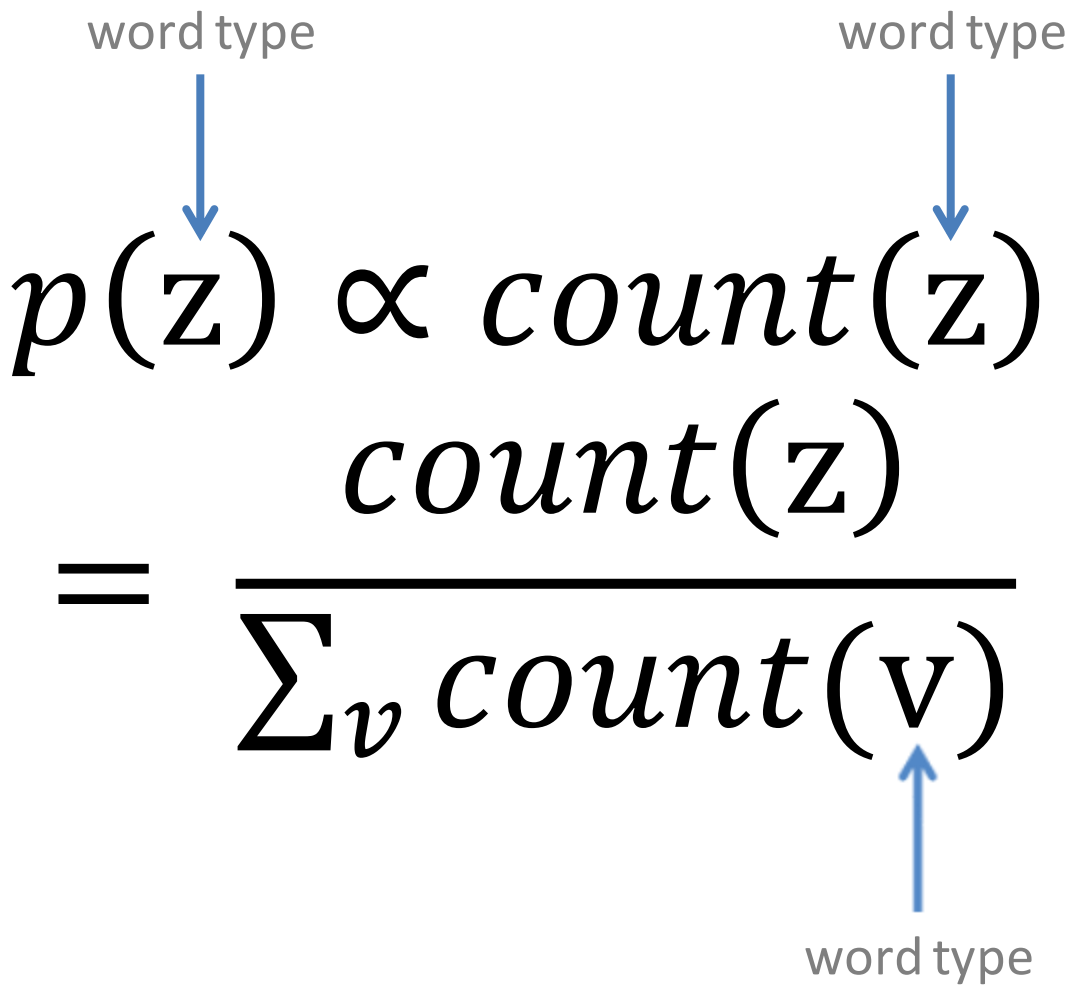
Count-Based N-Grams (Unigrams)

$$p(\mathrm{z}) \propto count(\mathrm{z})$$

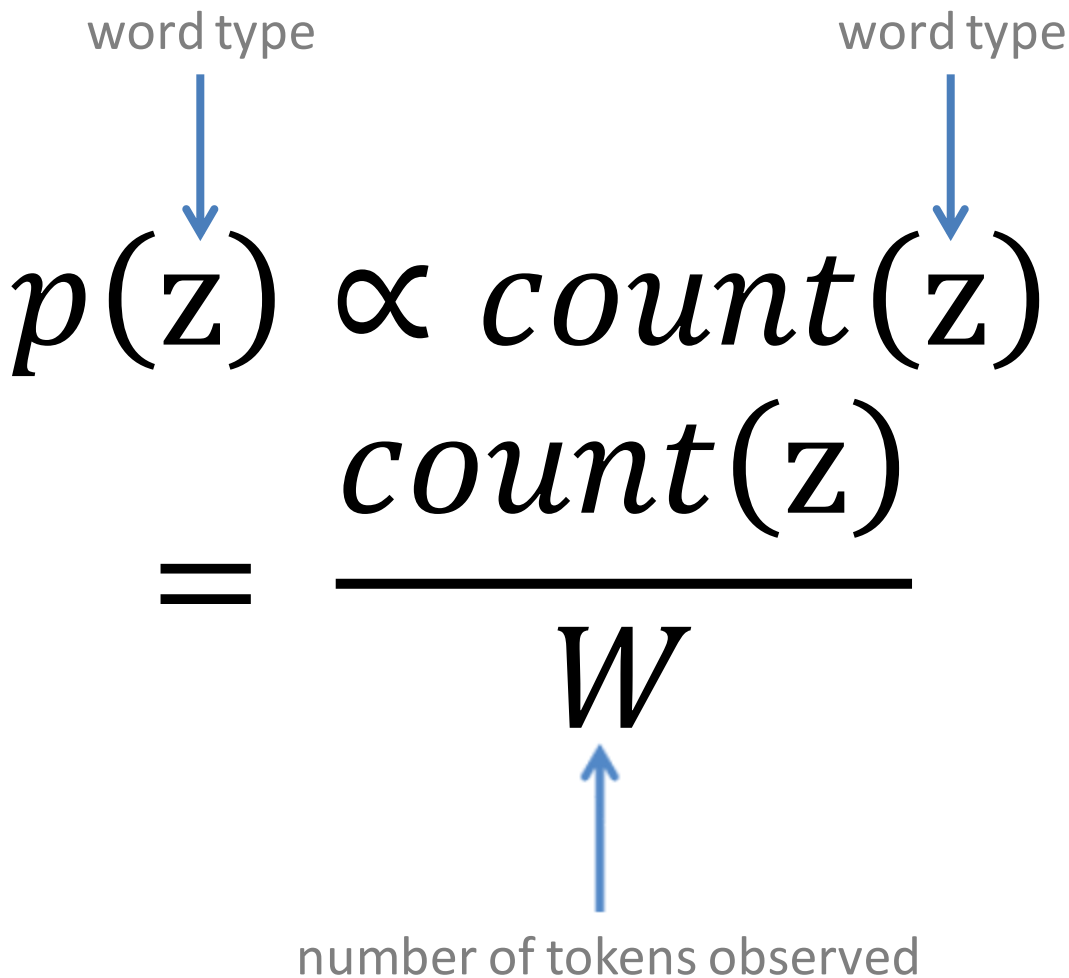
Count-Based N-Grams (Unigrams)

$$p(\mathrm{z}) \propto count(\mathrm{z})$$
$$= \frac{count(\mathrm{z})}{\sum_v count(\mathrm{v})}$$

Count-Based N-Grams (Unigrams)



Count-Based N-Grams (Unigrams)



Count-Based N-Grams (Trigrams)

$$p(\mathrm{z}|\mathrm{x},\mathrm{y}) \propto count(\mathrm{x},\mathrm{y},\mathrm{z})$$

implicitly conditioning
z on x and y

Count-Based N-Grams (Trigrams)

$$p(\mathrm{z}|\mathrm{x},\mathrm{y}) \propto count(\mathrm{x},\mathrm{y},\mathrm{z})$$
$$= \frac{count(\mathrm{x},\mathrm{y},\mathrm{z})}{\sum_v count(\mathrm{x},\mathrm{y},\mathrm{v})}$$

Perplexity

Lower is better : lower perplexity ➔ less surprised

More outcomes ➔
More surprised

Fewer outcomes ➔
Less surprised

Perplexity

Lower is better : lower perplexity ➔ less surprised

n-gram history (n-1 items)

$$\text{perplexity} = \exp\left(\frac{-1}{M} \sum_{i=1}^{M} \log p(w_i \mid h_i)\right)$$

Perplexity

Lower is better : lower perplexity ➔ less surprised

$$\text{perplexity} = \exp\left(\frac{-1}{M} \sum_{i=1}^{M} \log \underbrace{p(w_i \mid h_i)}_{\geq 0,\ \leq 1:\ \text{higher}}\right)$$

Perplexity

Lower is better : lower perplexity ➔ less surprised

$$\text{perplexity} = \exp\left(\frac{-1}{M}\sum_{i=1}^{M}\log p(w_i \mid h_i)\right)$$

≤ 0: higher

≥ 0, ≤ 1: higher

Perplexity

Lower is better : lower perplexity ➔ less surprised

$$\text{perplexity} = \exp\left(\frac{-1}{M}\sum_{i=1}^{M}\log p(w_i \mid h_i)\right)$$

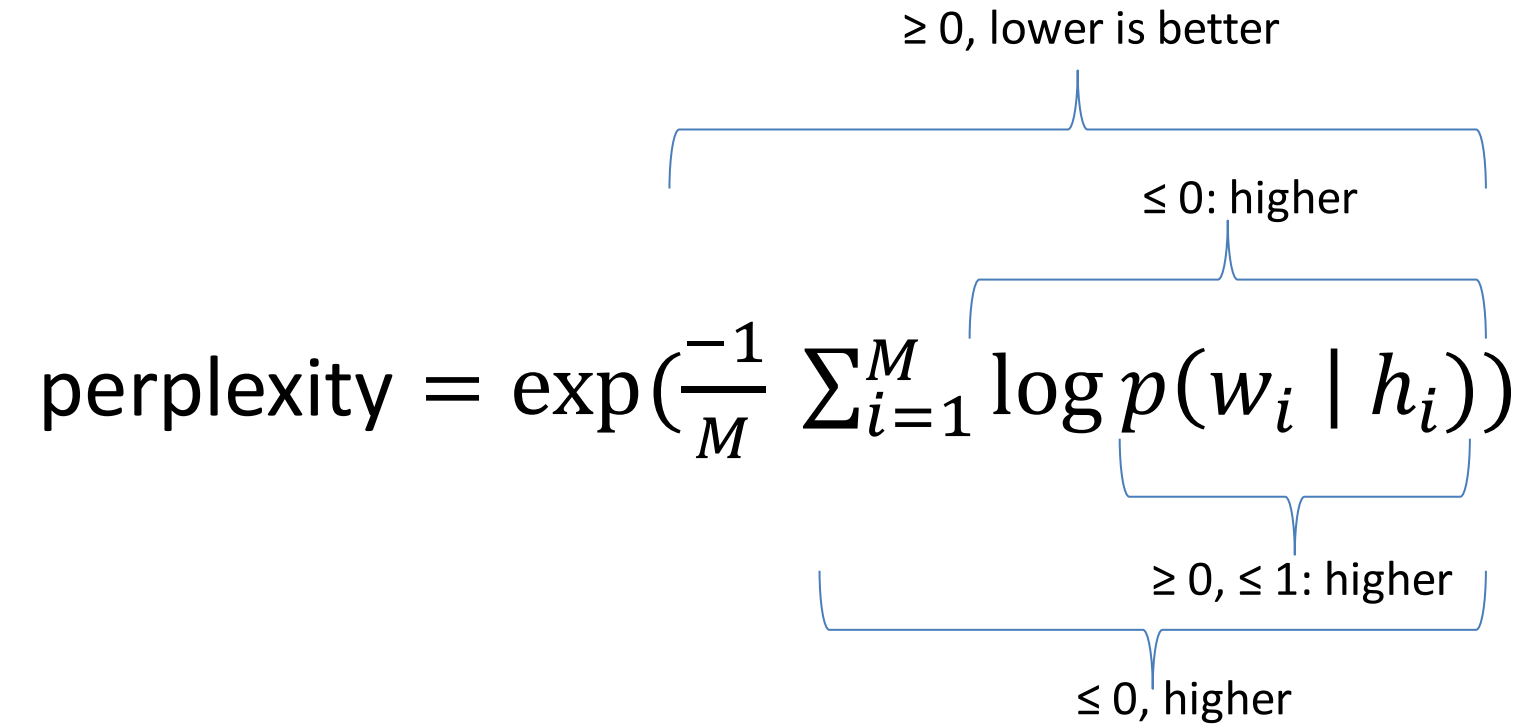
≤ 0: higher

≥ 0, ≤ 1: higher

≤ 0, higher

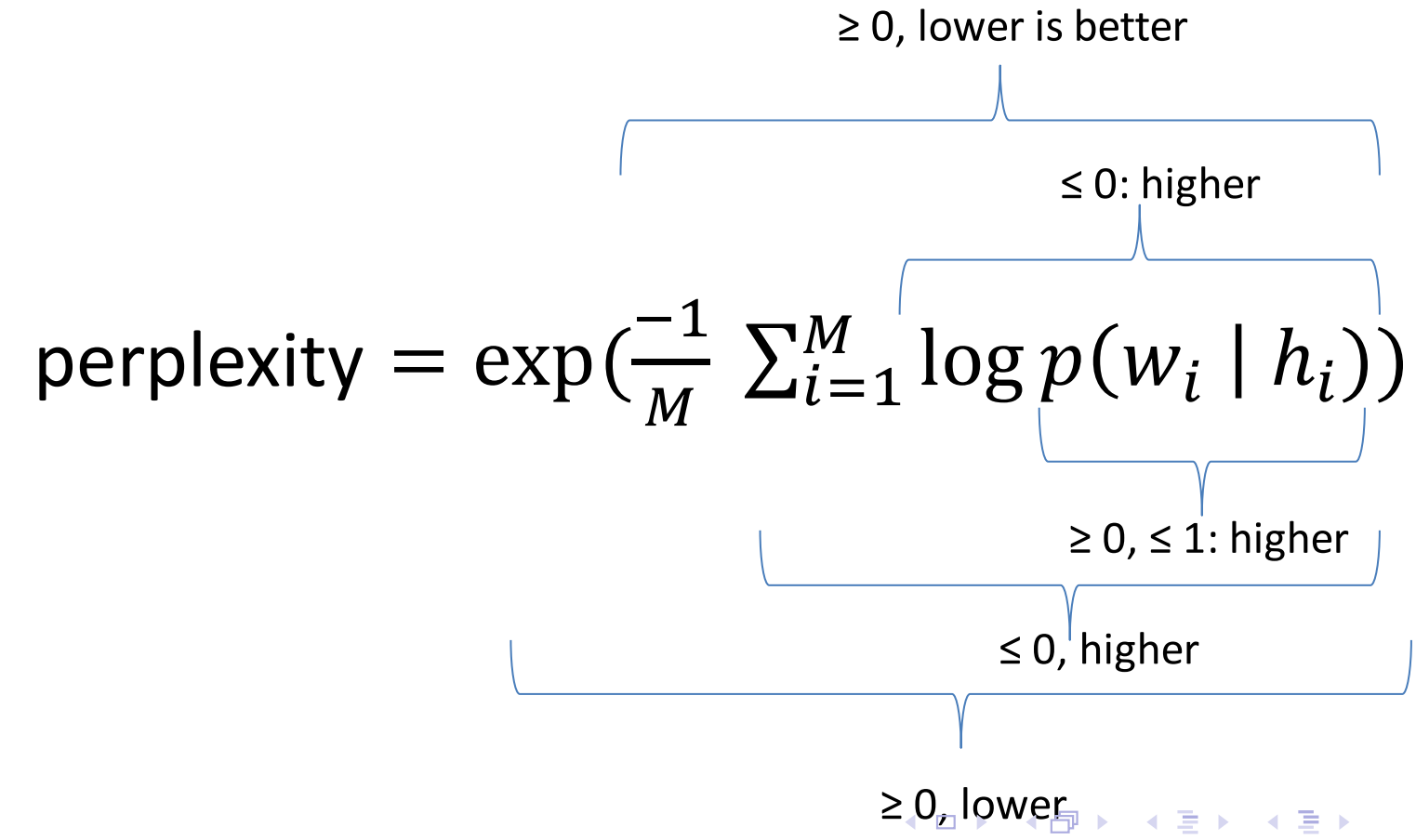
Perplexity

Lower is better : lower perplexity ➔ less surprised



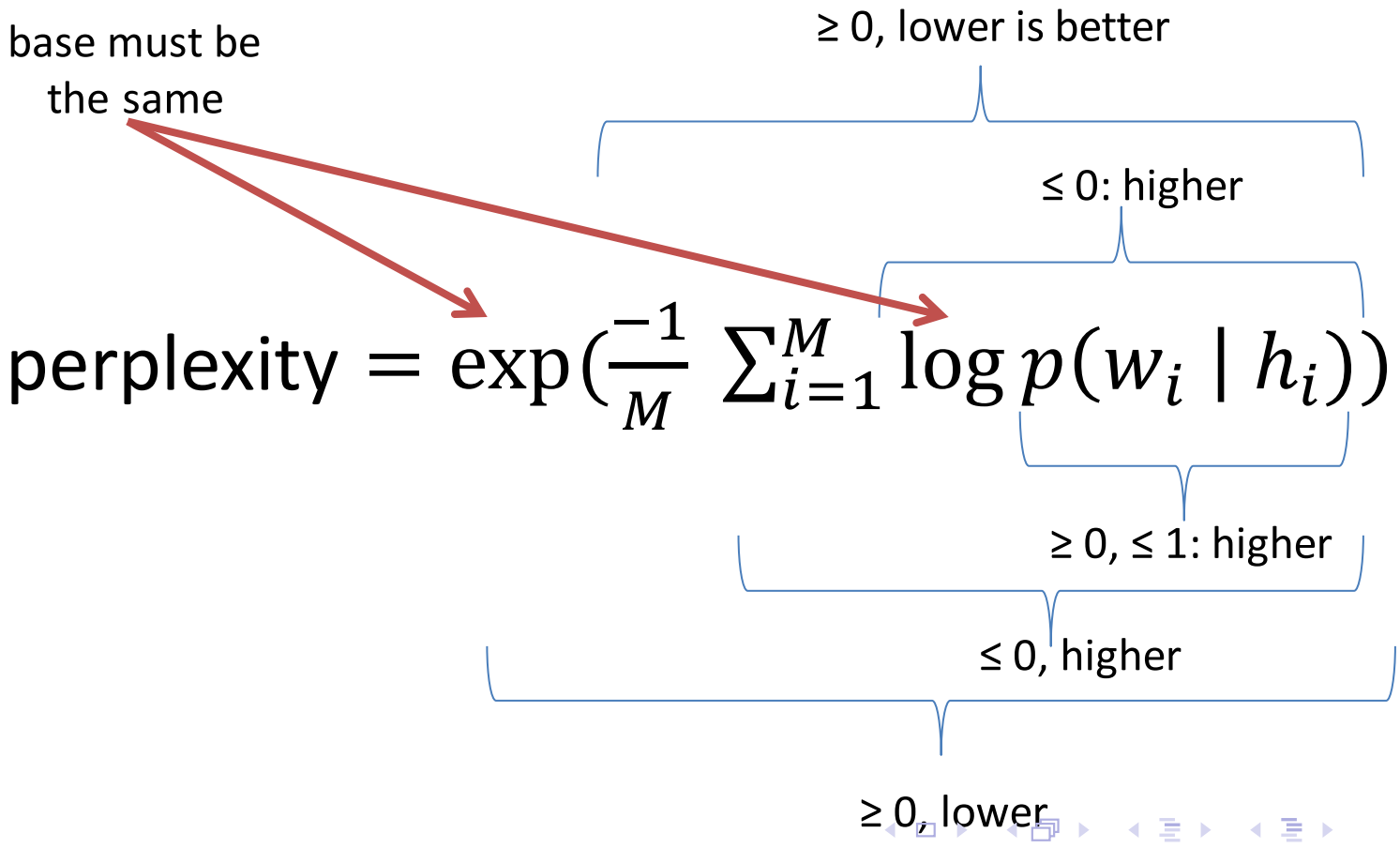
Perplexity

Lower is better : lower perplexity ➔ less surprised



Perplexity

Lower is better : lower perplexity ➔ less surprised



Perplexity

Lower is better : lower perplexity ➔ less surprised

$$\text{perplexity} = \exp\left(\frac{-1}{M} \sum_{i=1}^{M} \log p(w_i \mid h_i)\right)$$

$$= \sqrt[M]{\prod_{i=1} \frac{1}{p(w_i \mid h_i)}}$$

weighted geometric average

Outline

Probability review

Words

Defining Language Models

Breaking & Fixing Language Models

Maximum Likelihood Estimates

$$p(\text{item}) \propto count(\text{item})$$

Maximizes the likelihood of the training set

Do different corpora look the same?

Low(er) bias, high(er) variance

For large data: can actually do reasonably well

n = 1

, , land of in , a teachers The , wilds the and gave a Etienne any

two beginning without probably heavily that other useless the the

a different . the able mines , unload into in foreign the the be either other Britain finally avoiding , for of have the cure , the Gutenberg-tm ; of being can as country in authority deviates as d seldom and They employed about from business marshal materials than in , they

n = 2

These varied with it to the civil wars , therefore , it did not for the company had the East India , the mechanical , the sum which were by barter , vol. i , and , conveniencies of all made to purchase a council of landlords , constitute a sum as an argument , having thus forced abroad , however , and influence in the one , or banker , will there was encouraged and more common trade to corrupt , profit , it ; but a master does not , twelfth year the consent that of volunteers and *[…]*

, the other hand , it certainly it very earnestly entreat both nations .

In opulent nations in a revenue of four parts of production .

n = 3

His employer , if silver was regulated according to the temporary and occasional event .

What goods could bear the expense of defending themselves , than in the value of different sorts of goods , and placed at a much greater , there have been the effects of self-deception , this attention , but a very important ones , and which , having become of less than they ever were in this agreement for keeping up the business of weighing .

After food , clothes , and a few months longer credit than is wanted , there must be sufficient to keep by him , are of such colonies to surmount .

They facilitated the acquisition of the empire , both from the rents of land and labour of those pedantic pieces of silver which he can afford to take from the duty upon every quarter which they have a more equable distribution of employment .

n = 4

To buy in one market , in order to have it ; but the 8th of George III .

The tendency of some of the great lords , gradually encouraged their villains to make upon the prices of corn , cattle , poultry , etc .

Though it may , perhaps , in the mean time , that part of the governments of New England , the market , trade cannot always be transported to so great a number of seamen , not inferior to those of other European nations from any direct trade to America .

The farmer makes his profit by parting with it .

But the government of that country below what it is in itself necessarily slow , uncertain , liable to be interrupted by the weather .

0s Are Not Your (Language Model's) Friend

$$p(\text{item}) \propto count(\text{item}) = 0 \rightarrow$$
$$p(\text{item}) = 0$$

0s Are Not Your (Language Model's) Friend

$$p(\text{item}) \propto count(\text{item}) = 0 \rightarrow$$
$$p(\text{item}) = 0$$

0 probability → item is *impossible*

0s annihilate: x*y*z*0 = 0

Language is creative:

new words keep appearing

existing words could appear in known contexts

How much do you trust your data?

Add-λ estimation

Laplace smoothing,
Lidstone smoothing

Pretend we saw each word λ more times than we did

Add λ to all the counts

Add-λ estimation

Laplace smoothing,
Lidstone smoothing

Pretend we saw each word λ more times than we did

$$p(\mathrm{z}) \propto count(\mathrm{z}) + \lambda$$

Add λ to all the counts

Add-λ estimation

Laplace smoothing,
Lidstone smoothing

Pretend we saw each word λ more times than we did

Add λ to all the counts

$$p(\mathrm{z}) \propto count(\mathrm{z}) + \lambda$$
$$= \frac{count(\mathrm{z}) + \lambda}{\sum_v (count(\mathrm{v}) + \lambda)}$$

Add-λ estimation

Laplace smoothing,
Lidstone smoothing

Pretend we saw each word λ more times than we did

$$p(\mathrm{z}) \propto count(\mathrm{z}) + \lambda$$
$$= \frac{count(\mathrm{z}) + \lambda}{W + V\lambda}$$

Add λ to all the counts

Backoff and Interpolation

Sometimes it helps to use **less** context

- condition on less context for contexts you haven't learned much

Backoff and Interpolation

Sometimes it helps to use **less** context

condition on less context for contexts you haven't learned much about

Backoff:

use trigram if you have good evidence

otherwise bigram, otherwise unigram

Backoff and Interpolation

Sometimes it helps to use **less** context

- condition on less context for contexts you haven't learned much about

Backoff:

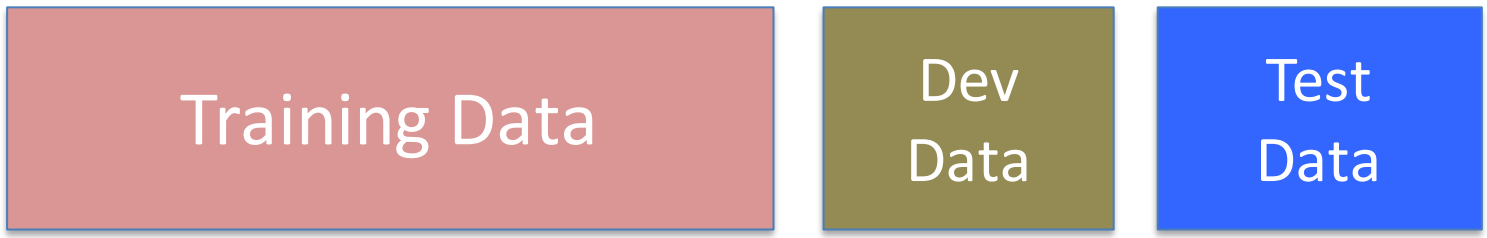
- use trigram if you have good evidence
- otherwise bigram, otherwise unigram

Interpolation:

- mix (average) unigram, bigram, trigram

Setting Hyperparameters

Use a **development** corpus



Choose λs to maximize the probability of dev data:

- Fix the N-gram probabilities (on the training data)
- Then search for λs that give largest probability to held-out set:

Add-λ N-Grams (Unigrams)

The film got a great opening and the film went on to become a hit .

Word (Type)	Raw Count	Norm	Prob.	Add-λ Count	Add-λ Norm.	Add-λ Prob.
The	1	16	1/16			
film	2		1/8			
got	1		1/16			
a	2		1/8			
great	1		1/16			
opening	1		1/16			
and	1		1/16			
the	1		1/16			
went	1		1/16			
on	1		1/16			
to	1		1/16			
become	1		1/16			
hit	1		1/16			
.	1		1/16			

Add-1 N-Grams (Unigrams)

The film got a great opening and the film went on to become a hit .

Word (Type)	Raw Count	Norm	Prob.	Add-1 Count	Add-1 Norm.	Add-1 Prob.
The	1	16	1/16	2		
film	2		1/8	3		
got	1		1/16	2		
a	2		1/8	3		
great	1		1/16	2		
opening	1		1/16	2		
and	1		1/16	2		
the	1		1/16	2		
went	1		1/16	2		
on	1		1/16	2		
to	1		1/16	2		
become	1		1/16	2		
hit	1		1/16	2		
.	1		1/16	2		

Add-1 N-Grams (Unigrams)

The film got a great opening and the film went on to become a hit .

Word (Type)	Raw Count	Norm	Prob.	Add-1 Count	Add-1 Norm.	Add-1 Prob.
The	1	16	1/16	2	16 + 14*1 = 30	
film	2		1/8	3		
got	1		1/16	2		
a	2		1/8	3		
great	1		1/16	2		
opening	1		1/16	2		
and	1		1/16	2		
the	1		1/16	2		
went	1		1/16	2		
on	1		1/16	2		
to	1		1/16	2		
become	1		1/16	2		
hit	1		1/16	2		
.	1		1/16	2		

Bayes Rule → NLP Applications

$$p(X \mid Y) = \frac{p(Y \mid X) * p(X)}{p(Y)}$$

likelihood

prior probability

posterior probability

marginal likelihood (probability)

Text Classification

Assigning subject categories, topics, or genres

Spam detection

Authorship identification

Age/gender identification

Language Identification

Sentiment analysis

…

Text Classification

Assigning subject categories, topics, or genres

Spam detection

Authorship identification

Age/gender identification

Language Identification

Sentiment analysis

…

Input:

a document

a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$

Output: a predicted class $c \in C$

Text Classification: Hand-coded Rules?

Assigning subject categories, topics, or genres

Spam detection

Authorship identification

Age/gender identification

Language Identification

Sentiment analysis

…

Rules based on combinations of words or other features

spam: black-list-address OR (“dollars” AND“have been selected”)

Accuracy can be high

If rules carefully refined by expert

Building and maintaining these rules is expensive

Text Classification: Supervised Machine Learning

Assigning subject categories, topics, or genres

Spam detection

Authorship identification

Age/gender identification

Language Identification

Sentiment analysis

…

Input:

a document d

a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$

A training set of m hand-labeled documents $(d_1,c_1),\ldots,(d_m,c_m)$

Output:

a learned classifier $\gamma:d \rightarrow c$

Text Classification: Supervised Machine Learning

Assigning subject categories, topics, or genres

Spam detection

Authorship identification

Age/gender identification

Language Identification

Sentiment analysis

…

Input:

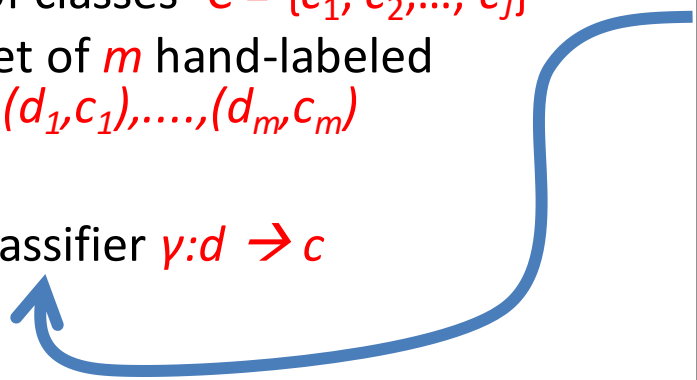
a document d

a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$

A training set of m hand-labeled documents $(d_1,c_1),\ldots,(d_m,c_m)$

Output:

a learned classifier $\gamma: d \rightarrow c$



Naïve Bayes

Logistic regression

Support-vector machines

k-Nearest Neighbors

…

Text Classification: Supervised Machine Learning

Assigning subject categories, topics, or genres

Spam detection

Authorship identification

Age/gender identification

Language Identification

Sentiment analysis

…

Input:

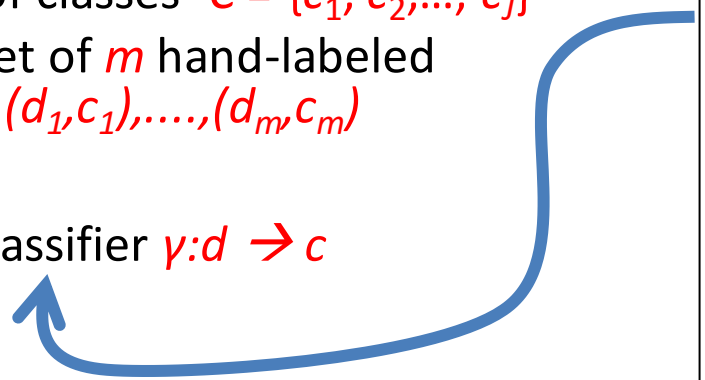
a document d

a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$

A training set of m hand-labeled documents $(d_1,c_1),\ldots,(d_m,c_m)$

Output:

a learned classifier $\gamma: d \rightarrow c$



Naïve Bayes

Logistic regression

Support-vector machines

k-Nearest Neighbors

…

Probabilistic Text Classification

Assigning subject categories, topics, or genres

Spam detection

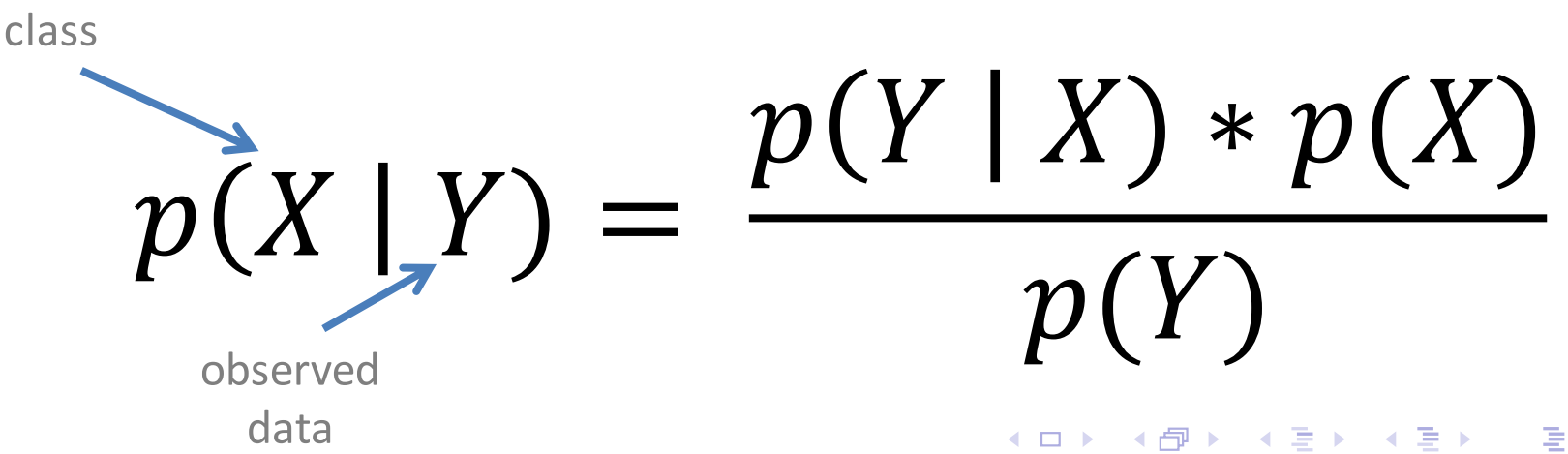
Authorship identification

Age/gender identification

Language Identification

Sentiment analysis

…



Probabilistic Text Classification

Assigning subject categories, topics, or genres

Spam detection

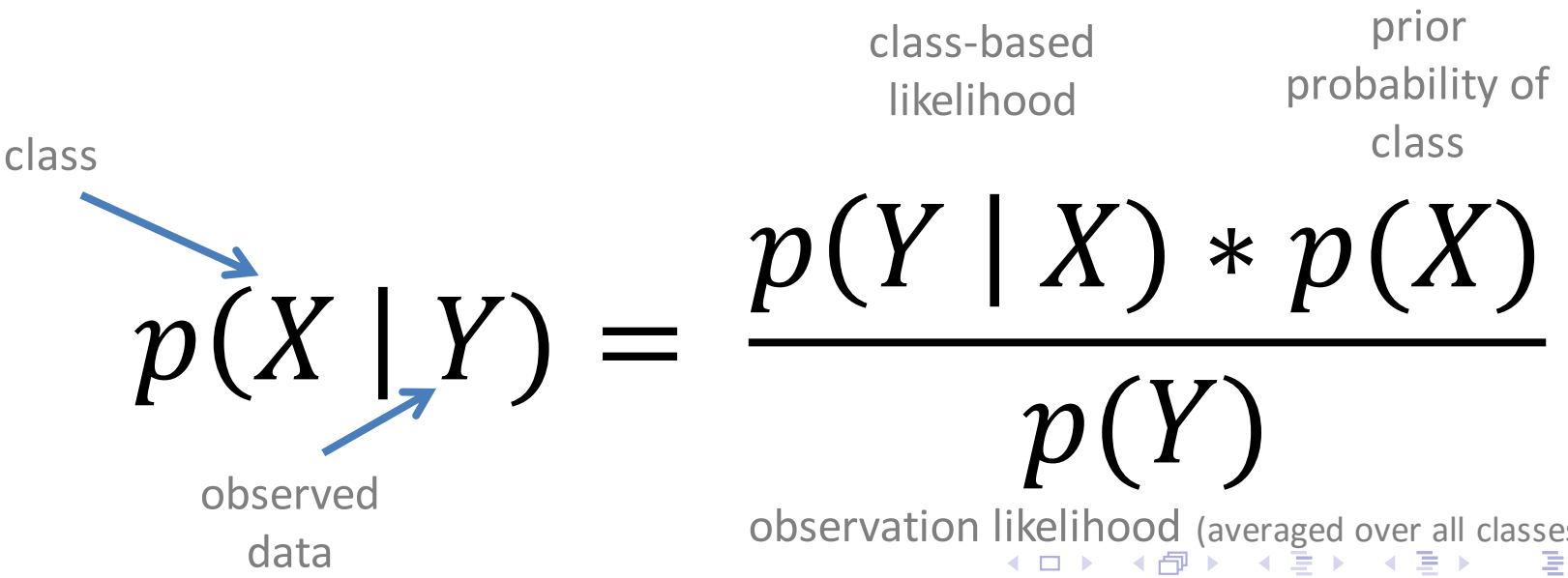
Authorship identification

Age/gender identification

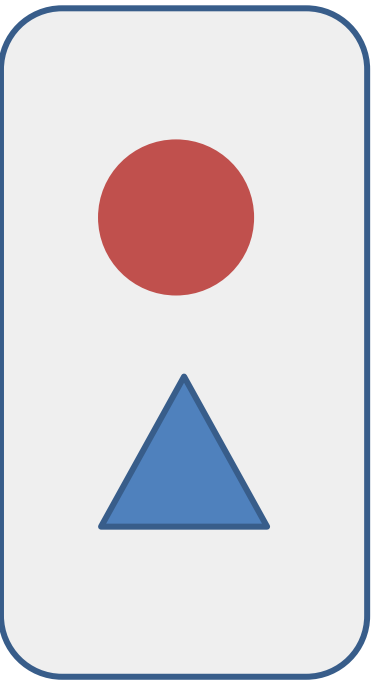
Language Identification

Sentiment analysis

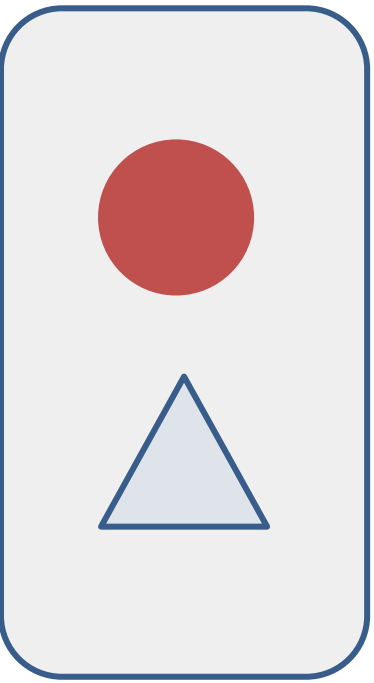
…



Noisy Channel Model

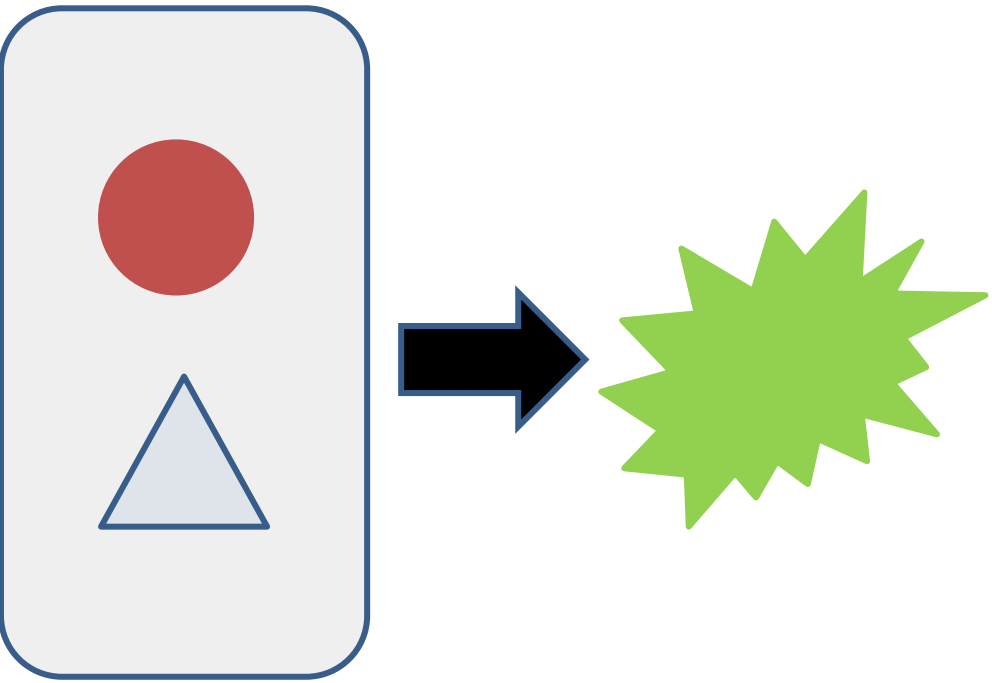


Noisy Channel Model



what I want to tell you *"sports"*

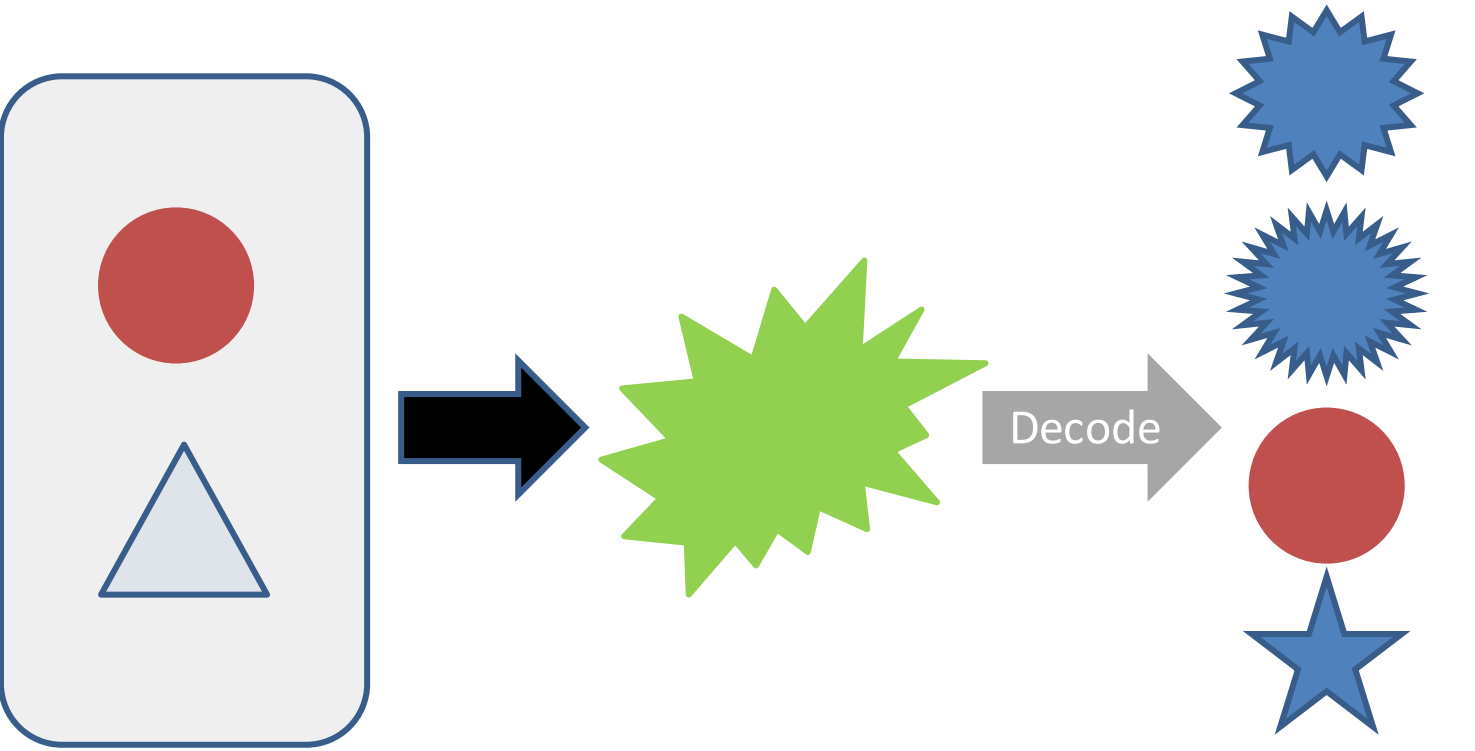
Noisy Channel Model



what I want to tell you *"sports"*

what you actually see *"The Os lost again…"*

Noisy Channel Model

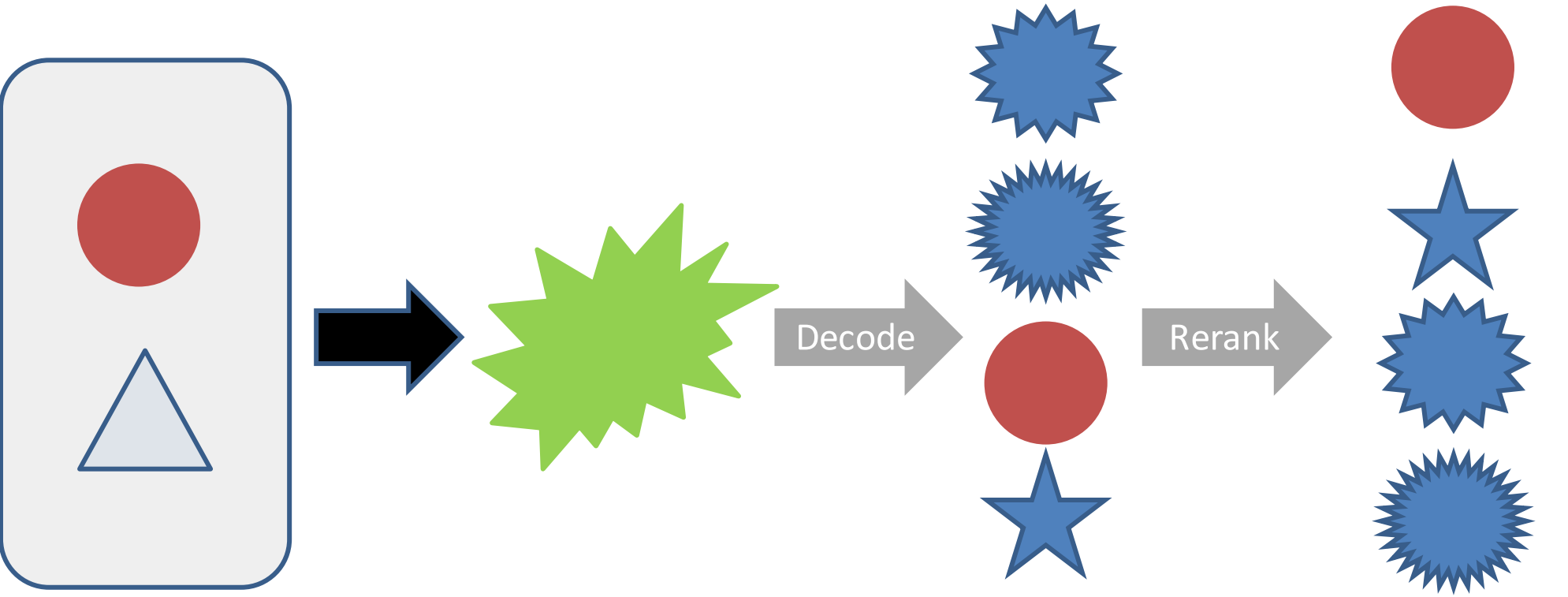


what I want to tell you
"sports"

what you actually see
"The Os lost again…"

hypothesized intent
"sad stories"
"sports"

Noisy Channel Model



what I want to tell you *"sports"*

what you actually see *"The Os lost again…"*

hypothesized intent *"sad stories"* *"sports"*

reweight according to what's likely *"sports"*

Noisy Channel

Machine translation

Speech-to-text

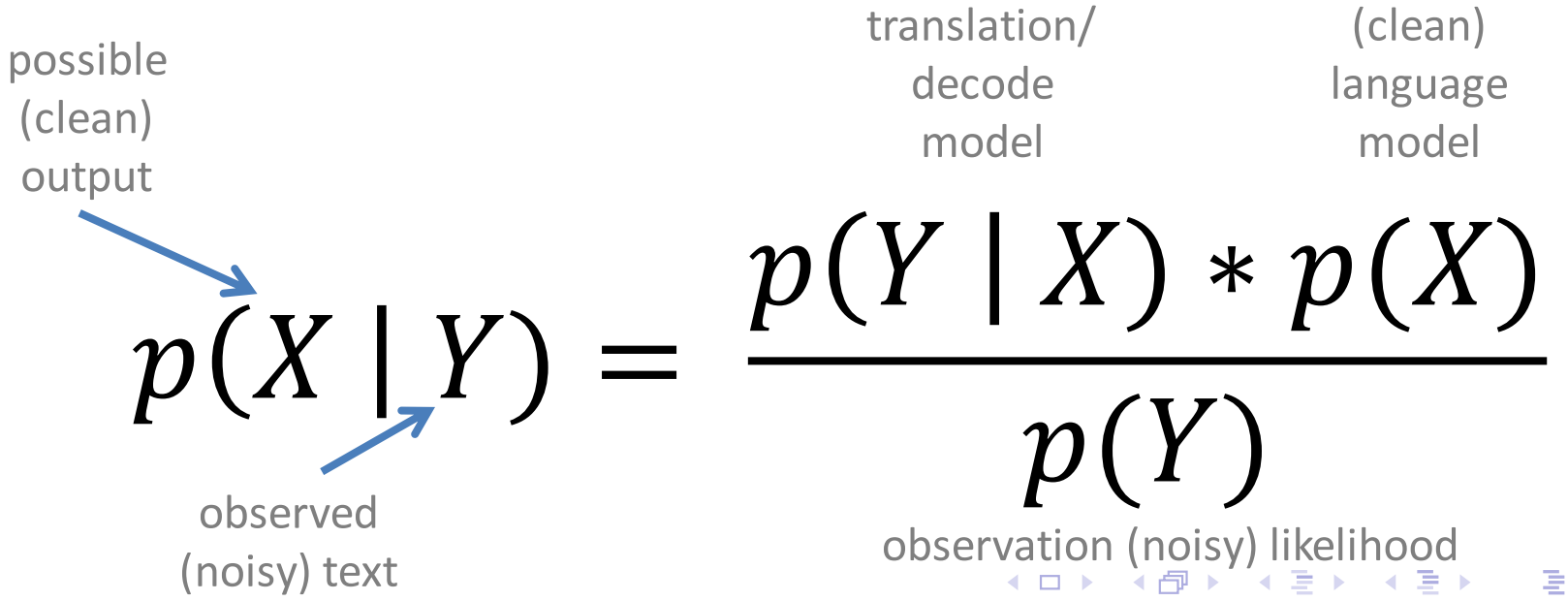
Spelling correction

Text normalization

Part-of-speech tagging

Morphological analysis

…



Noisy Channel

Machine translation

Speech-to-text

Spelling correction

Text normalization

Part-of-speech tagging

Morphological analysis

…

translation/ decode model

(clean) language model

possible (clean) output

$$p(X \mid Y) = \frac{p(Y \mid X) * p(X)}{p(Y)}$$

observed (noisy) text

observation (noisy) likelihood

Language Model

Use any of the language modeling algorithms we've learned

Unigram, bigram, trigram

Add-λ, interpolation, backoff

(Later: Maxent, RNNs, hierarchical Bayesian LMs, …)

Noisy Channel

$$\text{argmax}_X p(X \mid Y)$$

Noisy Channel

$$\text{argmax}_X \frac{p(Y \mid X) * p(X)}{p(Y)}$$

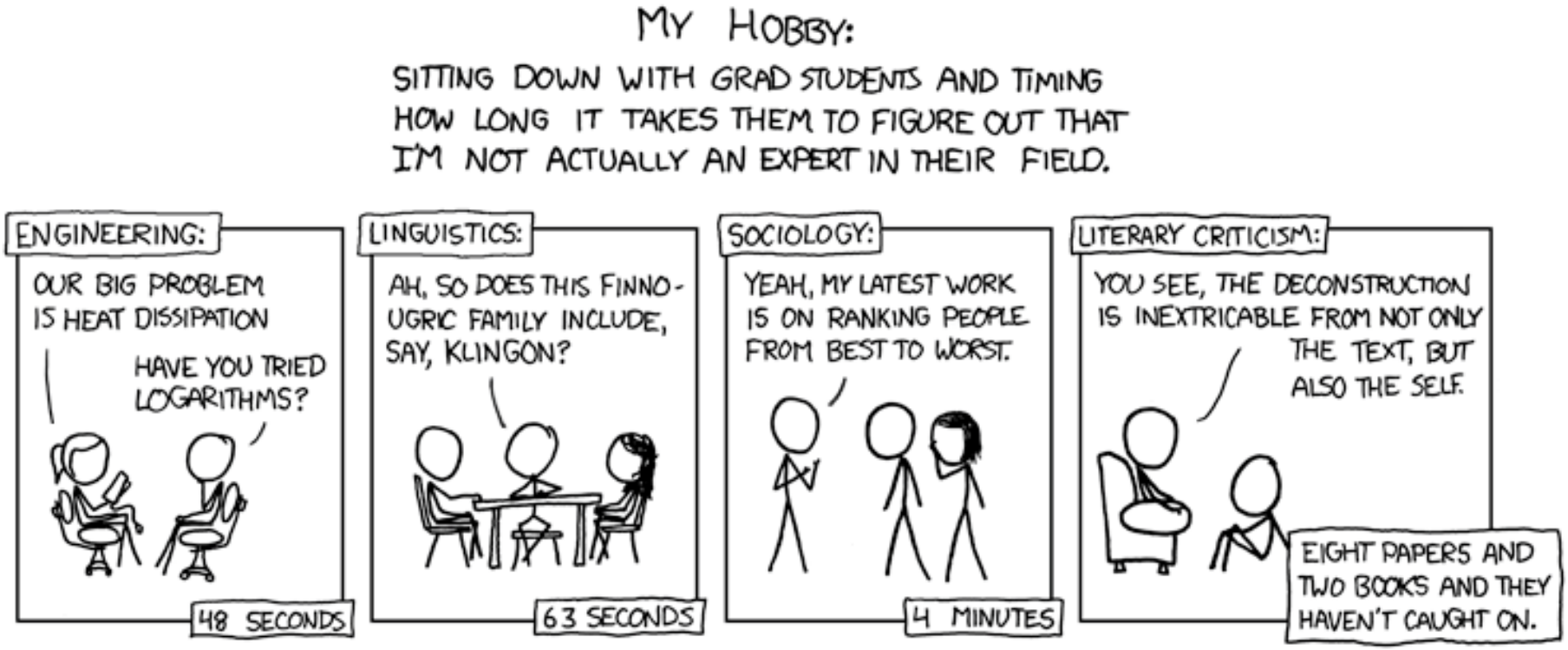
Noisy Channel

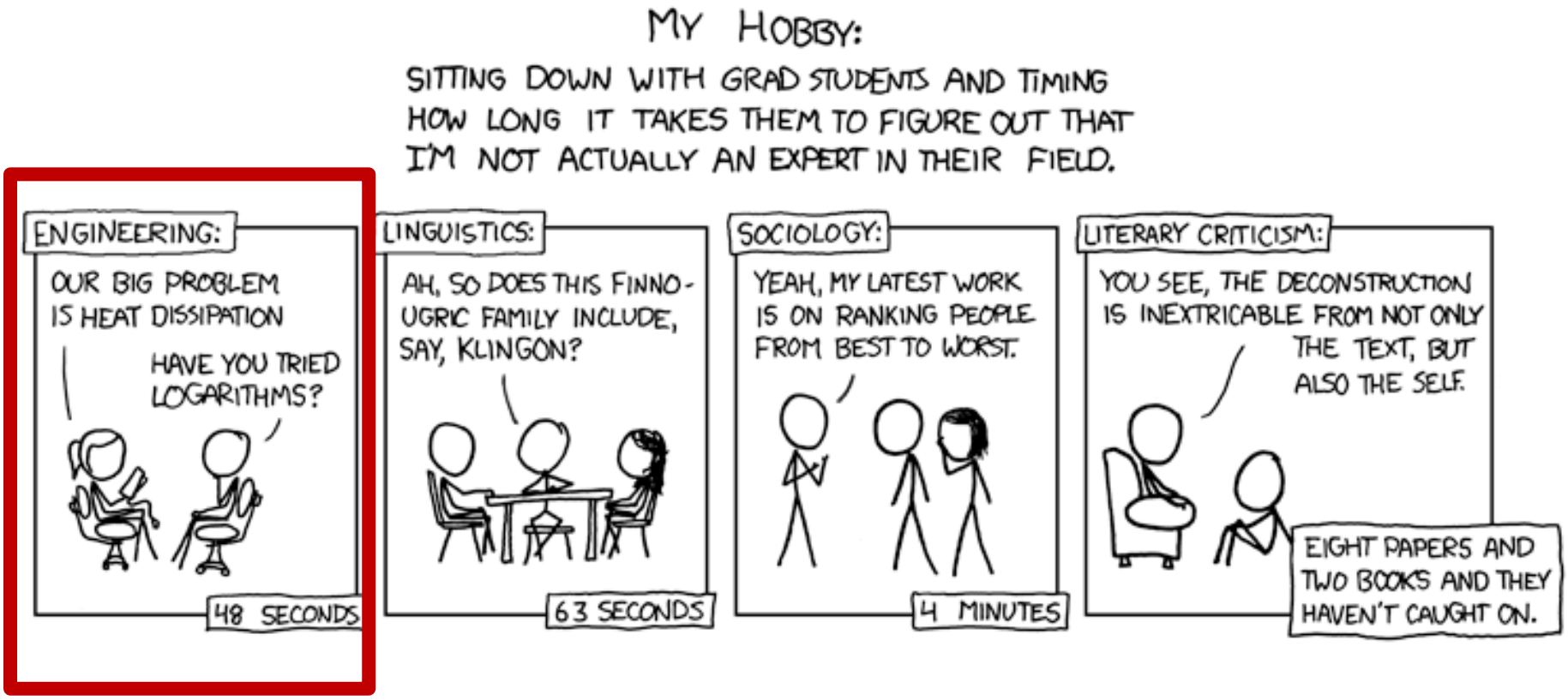
$$\text{argmax}_X \frac{p(Y \mid X) * p(X)}{p(Y)}$$

constant with respect to X

Noisy Channel

$$\text{argmax}_X p(Y \mid X) * p(X)$$


MY HOBBY:
SITTING DOWN WITH GRAD STUDENTS AND TIMING HOW LONG IT TAKES THEM TO FIGURE OUT THAT I'M NOT ACTUALLY AN EXPERT IN THEIR FIELD.
ENGINEERING:
OUR BIG PROBLEM IS HEAT DISSIPATION
HAVE YOU TRIED LOGARITHMS?
48 SECONDS
LINGUISTICS:
AH, SO DOES THIS FINNO-UGRIC FAMILY INCLUDE, SAY, KLINGON?
63 SECONDS
SOCIOLOGY:
YEAH, MY LATEST WORK IS ON RANKING PEOPLE FROM BEST TO WORST.
4 MINUTES
LITERARY CRITICISM:
YOU SEE, THE DECONSTRUCTION IS INEXTRICABLE FROM NOT ONLY THE TEXT, BUT ALSO THE SELF.
EIGHT PAPERS AND TWO BOOKS AND THEY HAVEN'T CAUGHT ON.


MY HOBBY:
SITTING DOWN WITH GRAD STUDENTS AND TIMING HOW LONG IT TAKES THEM TO FIGURE OUT THAT I'M NOT ACTUALLY AN EXPERT IN THEIR FIELD.
ENGINEERING:
OUR BIG PROBLEM IS HEAT DISSIPATION
HAVE YOU TRIED LOGARITHMS?
48 SECONDS
LINGUISTICS:
AH, SO DOES THIS FINNO-UGRIC FAMILY INCLUDE, SAY, KLINGON?
63 SECONDS
SOCIOLOGY:
YEAH, MY LATEST WORK IS ON RANKING PEOPLE FROM BEST TO WORST.
4 MINUTES
LITERARY CRITICISM:
YOU SEE, THE DECONSTRUCTION IS INEXTRICABLE FROM NOT ONLY THE TEXT, BUT ALSO THE SELF.
EIGHT PAPERS AND TWO BOOKS AND THEY HAVEN'T CAUGHT ON.