Natural Language Processing

Lecture 05

Dirk Hovy

dirk.hovy@unibocconi.it

@dirk_hovy

Bocconi

Today's Goals

- Know when (and when not) to use **Regular expressions**
- Understand how language/information can be modeled as **probability distributions**
- Understand how information can be quantified with **entropy**
- Learn about **KL-divergence** for the difference between distributions
- Understand **PMI** and see why it can help find collocations

Flexible Matches: Regular Expressions

http://xkcd.com/208/

The promise…

Is it an (Email) Address?

- notMyFault@webmail.com
- smithie123@gmx
- Free stuff@unibocconi.it
- mark_my_words@hotmail;com
- truthOrDare@webmail.in
- look@me@twitter.com
- how2GetAnts@aol.dfdsfgfdsgfd

NAME	@	DOMAIN	.	CODE

Simple Matching

sequence	Matches
`e`	any single occurrence of `e`
`at`	**at**, r**at**, m**at**, s**at**, c**at**, **at**tack, **at**tention, l**at**er

Quantifiers

	Means	Example	Matches
*	0 or more	`cooo*l`	`cool, coool`
+	1 or more	`hello+`	`hello, helloo, hellooooooo`
?	0 or 1	`fr?og`	`fog, frog`

Bocconi

Special Characters

	Means	Example	Matches
.	any single character	.el	eel, Nel, gel
\n	newline character (line break)	\n+	One or more line breaks
\t	a tab stop	\t+	One or more tabs
\d	a single digit [0-9]	B\d	B0, B1, ..., B9
\D	a non-digit	\D.t	' t, But, eat
\w	any alphanumberic character	\w\w\w	Top, WOO, ash, bee, ...
\W	non-alphanumberic character		
\s	a whitespace character		
\S	a non-whitespace character		
****	"Escapes" special characters to match them	.+ \.com	abc.com, united.com
^	the beginning of the input string	^...	First word in line
$	the end of the input string	^\n$	Empty line

Classes

	Means	Example	Matches
[abc]	Match any of a, b, c	`[bcrms]at`	`bat, cat, rat, mat, sat`
[^abc]	Match anything BUT a, b, c	`te[^ ]+s`	`tens, tests, teens, texts, terrors...`
[a-z]	Match any lowercase character	`[a-z][a-z]t`	`act, ant, not, ... wit`
[A-Z]	Match any uppercase character	`[A-Z]...`	`Ahab, Brit, In a, ..., York`
[0-9]	Match any digit	`DIN A[0-9]`	`DIN A0, DIN A1, ..., DIN A9`

Bocconi

Groups

	Means	Example	Matches
(abc)	Match abc	`.(ar).`	`hard, cart, fare, …`
(ab|c)	Match ab OR c	`(ab|c)ate`	`abate, cate`

Bocconi

Matching Addresses

NAME	@	DOMAIN	.	CODE

```
^[A-Za-z0-9_\.-]+@[A-Za-z0-9_\.-]+\.[A-Za-z0-9_][A-Za-z0-9_]+$
```

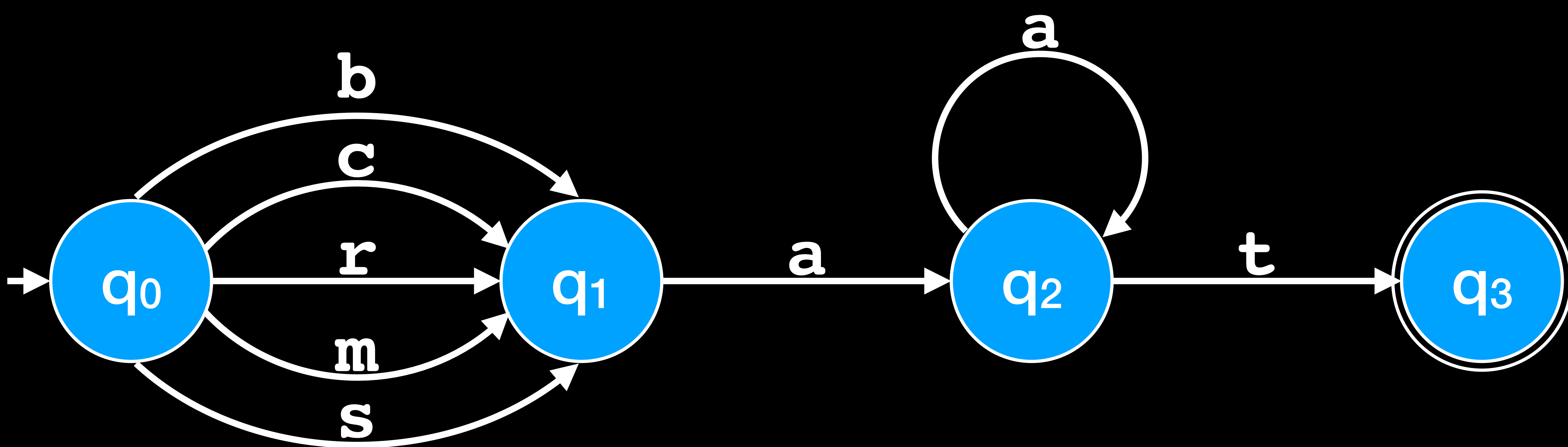
http://xkcd.com/1171/

A (W|w)ord of [Ww]arning

```
(?:(?:\r\n)?[ \t])*(?:(?:(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?
[ \t]))*"(?:(?:\r\n)?[ \t])*)(?:\.(?:(?:\r\n)?[ \t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\
[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[ \t]))*"(?:(?:\r\n)?[ \t])*))*@(?:(?:\r\n)?[ \t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?
[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[ \t])*)(?:\.(?:(?:\r\n)?[ \t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:
(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[ \t])*))*|(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:
\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[ \t]))*"(?:(?:\r\n)?[ \t])*)*\<(?:(?:\r\n)?[ \t])*(?:@(?:
[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[ \t])*)(?:\.(?:(?:
```


```
[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[ \t])*)(?:\.(?:(?:\r\n)?[ \t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:
(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[ \t])*))*|(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:
\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[ \t]))*"(?:(?:\r\n)?[ \t])*)*\<(?:(?:\r\n)?[ \t])*(?:@(?:
[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[ \t])*)(?:\.(?:(?:
\r\n)?[ \t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?
[ \t])*))*(?:,@(?:(?:\r\n)?[ \t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\
\.)*\](?:(?:\r\n)?[ \t])*)(?:\.(?:(?:\r\n)?[ \t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\
[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[ \t])*))*)*:(?:(?:\r\n)?[ \t])*)?(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:
\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[ \t]))*"(?:(?:\r\n)?[ \t])*)(?:\.(?:(?:\r\n)?[ \t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:
\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[ \t]))*"(?:(?:\r\n)?[ \t])*))*@(?:(?:\r\n)?[ \t])*(?:[^()<>@,;:\
\".\[\] \000-\031]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[ \t])*)(?:\.(?:(?:\r\n)?[ \t])*(?:
[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[ \t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[ \t])*))*\>(?:(?:\r\n)?
[ \t])*))*)?;\s*)
```

RegEx as Automata



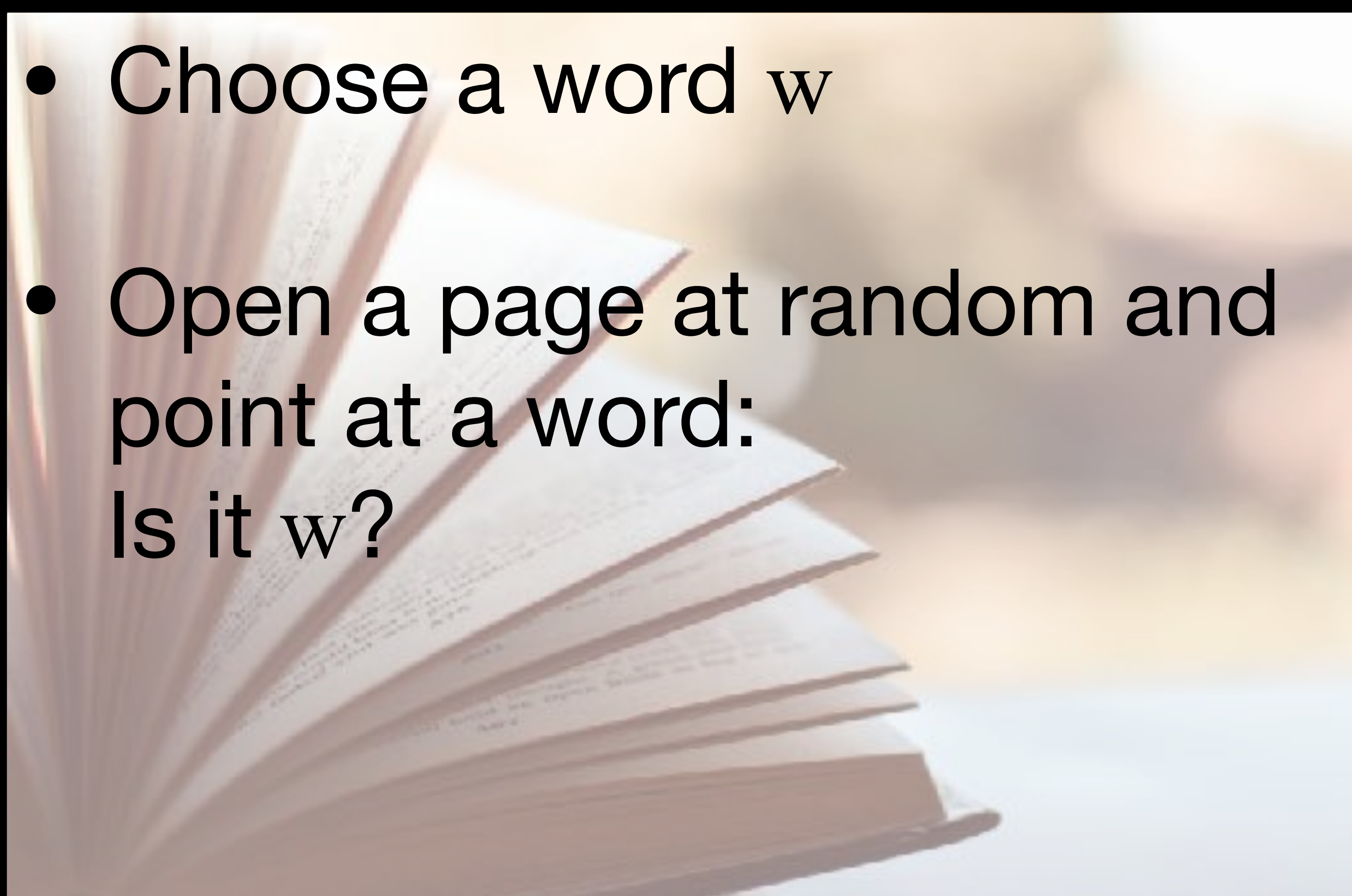
`[bcrms]a+t`

The Probability of Words

Probability of a Word

"It must be recognized that the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term."

—Noam Chomsky (grumpy linguist)



- Choose a word w
- Open a page at random and point at a word:
  Is it w?

HOW OFTEN WE HAVE SEEN W

$$P(w) = \frac{c(w)}{\sum_{v \in V} c(v)}$$

...ALL WORDS

Bocconi

Conditional Probability

We finish each others' ______________

SENTENCES

SANDWICHES

MOVIES

WHAT'S MORE LIKELY?

Conditional Probabilities

*TOTALLY MADE UP NUMBERS

h	w	P(w|h)*
tea with	milk	0,42
	sugar	0,35
	a	0,18
	stevia	0,05
for the	win	0,25
	majority	0,21
	birds	0,15
	…	…

SUM TO 1.0

Where Probabilities Come From

WE NEED A WAY TO ASSIGN

P(WORD | "WE FINISH EACH OTHERS'")

We finish each others' …

HOW OFTEN WE

HAVE SEEN W

$$P(w|h) = \frac{c(w)}{\sum_{h \in V} c(h, w)}$$

…AFTER THE OTHER WORDS

Bocconi

Count in 57m Tweets

MAXIMUM LIKELIHOOD ESTIMATION (MLE)

```
12 The weather today is just
 9 The weather today is so
 9 the weather today is slightly
 8 The weather today is perfect
 5 The weather today is beautiful
 4 The weather today is slightly
 3 the weather today is so
 3 the weather today is perfect
 3 The weather today is nearly
 3 the weather today is bitter
 3 The weather today is absolutely
 2 The weather today is wonderful
 2 The weather today is beyond
 2 The weather today is amazing
 2 The weather today is a
 1 the weather today is worth
 1 the weather today is weird
 1 The weather today is too
 1 the weather today is the
 1 The weather today is that
 1 the weather today is that
 1 The weather today is splendid
 1 THE WEATHER TODAY IS SO
 1 the weather today is simply
 1 The weather today is sickening
 1 The weather today is seriously
 1 The weather today is pretty
 1 the weather today is pretty
 1 The weather today is Perrfff
 1 the weather today is PERFECT
```

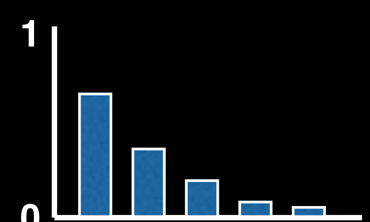
Probability Distributions

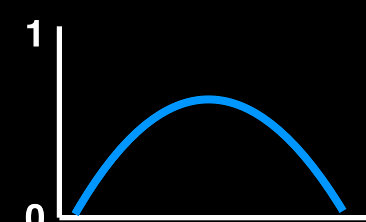
- mathematical way to describe a sample

```
def p(number_on_die):
    return 0.1667
```

$$P(x;N) = \frac{1}{N}$$

- discrete or continuous

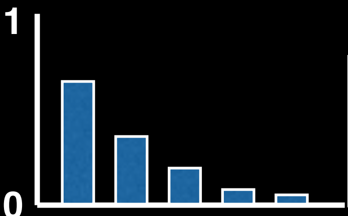




IN NLP:
USUALLY DISCRETE

- define “shape” and properties with parameters

- compute probability for any x

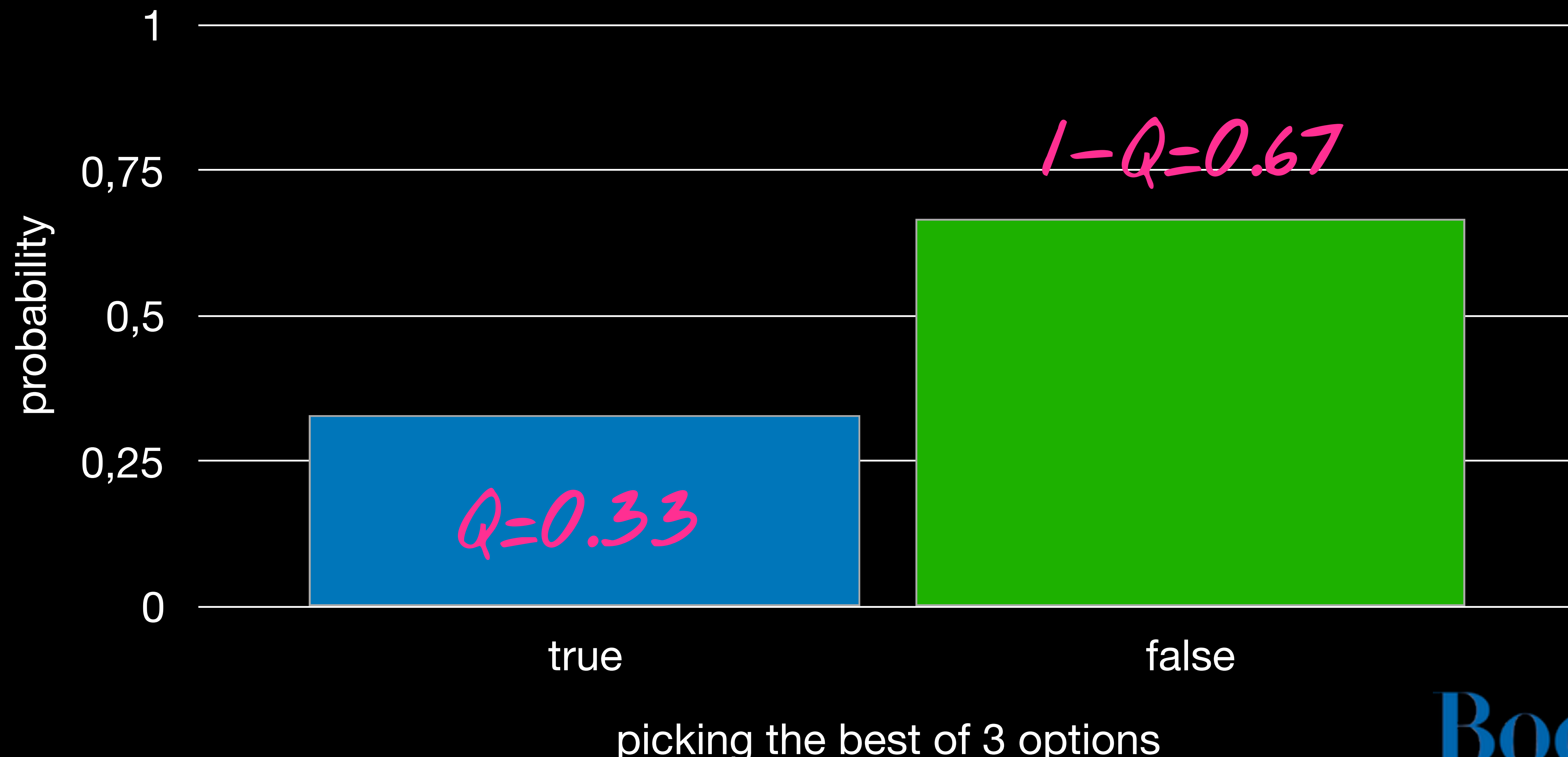


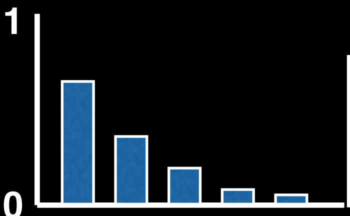
Bernoulli Distribution

Jacob Bernoulli

Parameters: q PROBABILITY OF SUCCESS

Function: $$P(x;q) = \begin{cases} 1-q & \text{if } x = 0 \\ q & \text{if } x = 1 \end{cases}$$ SUMS TO 1.0





Bernoulli Distribution

- has only two outcomes
- the probability of failure is the complement of success
- Examples: binary classification, indicator features

Categorical Distribution

VECTOR WITH ALL PROBABILITIES SUMS TO 1.0

Parameters: θ

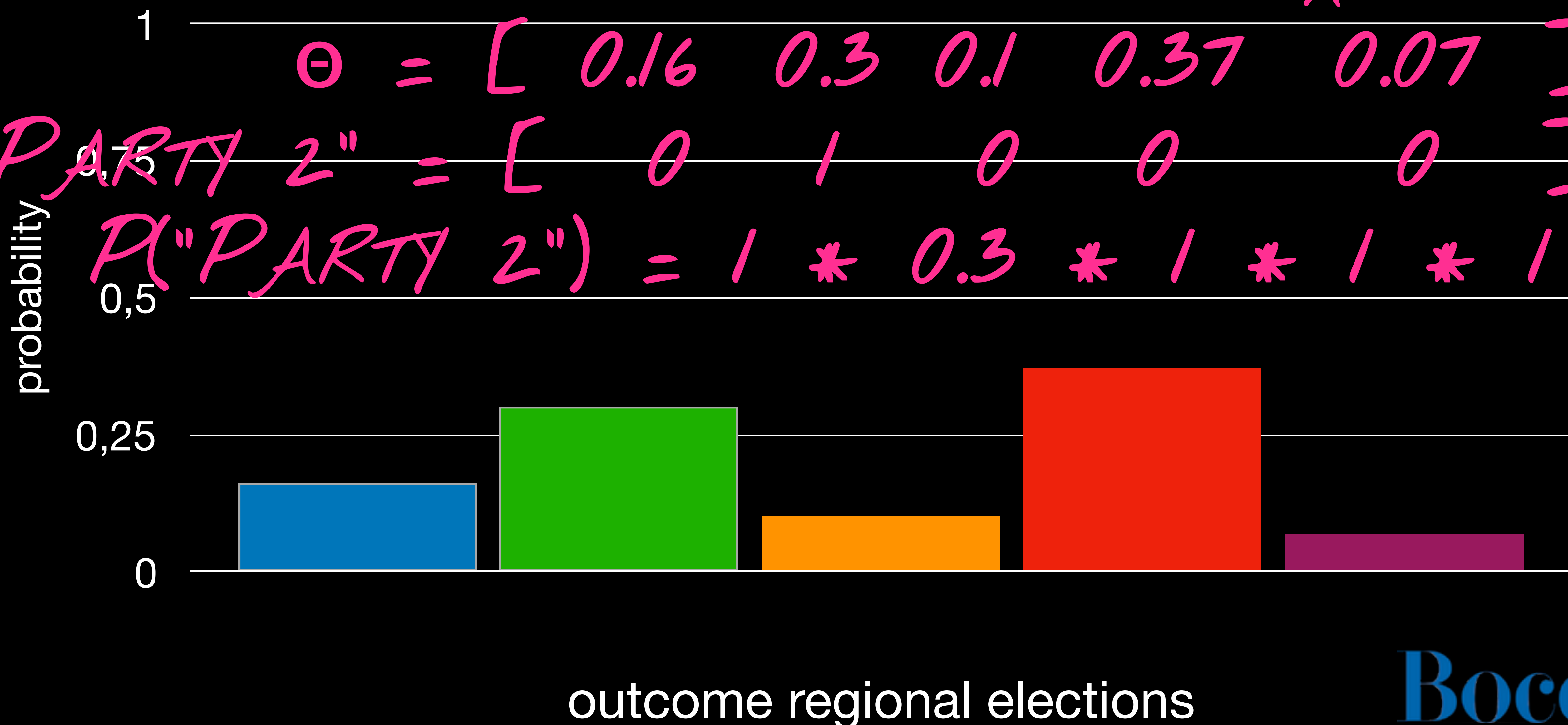
Function: $$P(x;\theta) = \prod_{j=1}^{K} \theta_j^{\mathbb{I}(x_j=1)}$$

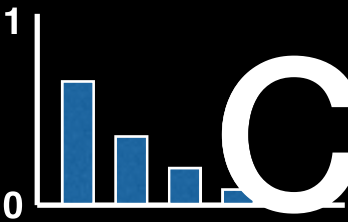
USE VALUE AT VECTOR POSITION FOR X

$\Theta = [\ 0.16 \quad 0.3 \quad 0.1 \quad 0.37 \quad 0.07\ ]$

"PARTY 2" $= [\ 0 \quad 1 \quad 0 \quad 0 \quad 0\ ]$

P("PARTY 2") $= 1 * 0.3 * 1 * 1 * 1$





Categorical Distribution

- has many outcomes (also called *multinomial*)

- if outcomes are numeric, we can compute the expected average value, or **expectation**

$$\mathbb{E}(\mathbf{X}) = \sum_{i=1}^{N} x_i \cdot P(x_i)$$

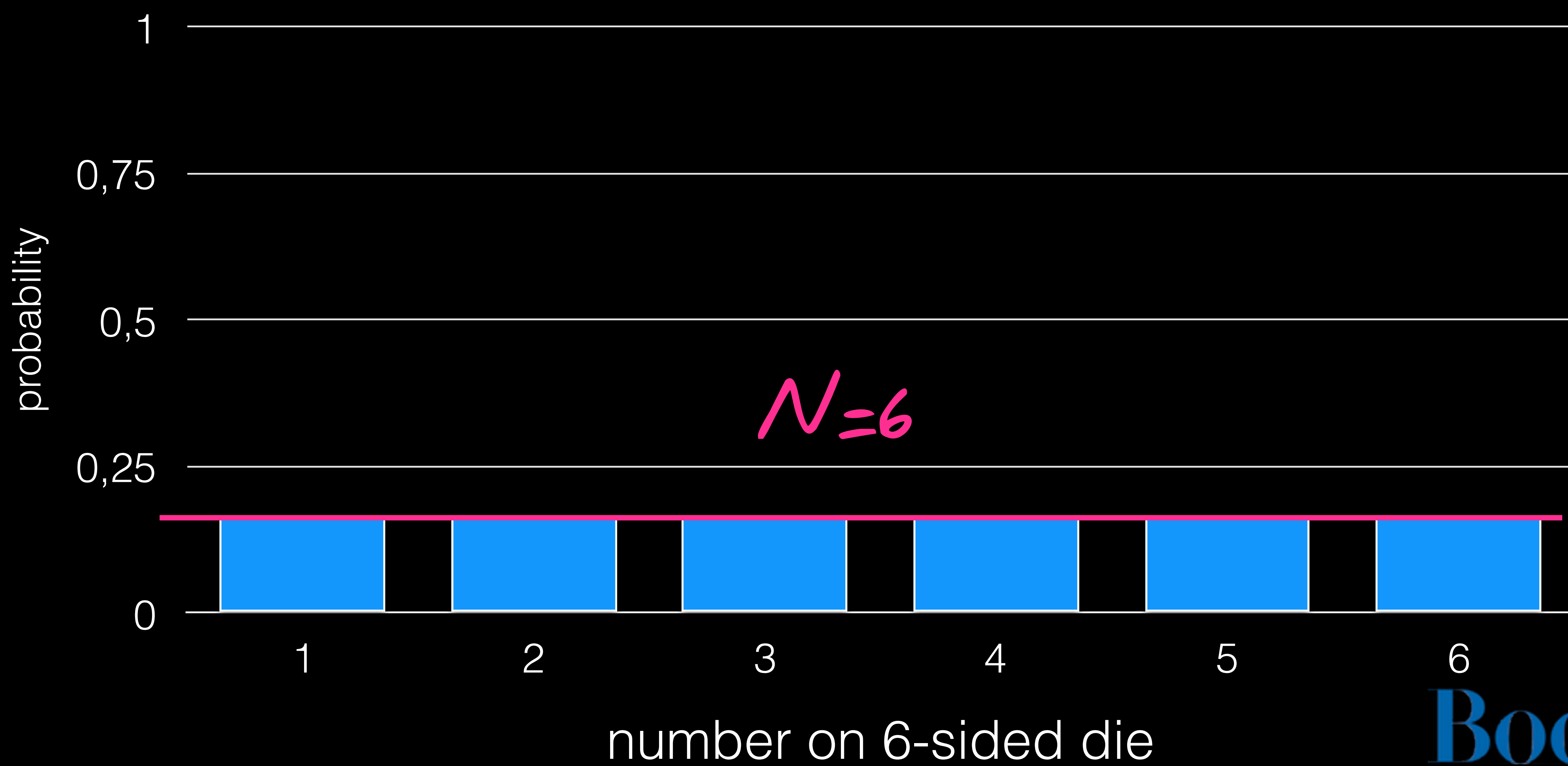
$$\begin{aligned} \text{E(die\_roll)} &= \mathbf{1} \cdot 0.1667 + \mathbf{2} \cdot 0.1667 + \mathbf{3} \cdot 0.1667 \\ &+ \mathbf{4} \cdot 0.1667 + \mathbf{5} \cdot 0.1667 + \mathbf{6} \cdot 0.1667 \\ &= \mathbf{3.5} \end{aligned}$$

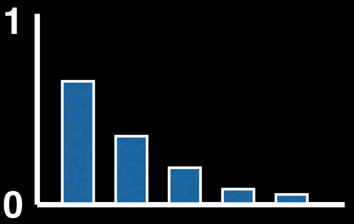
- Examples: word sequences, topics, multi-class labels, etc.

Uniform Distribution

Parameters: N NUMBER OF EVENTS

Function: $P(x; N) = \frac{1}{N}$ SUMS TO 1.0





Uniform Distribution

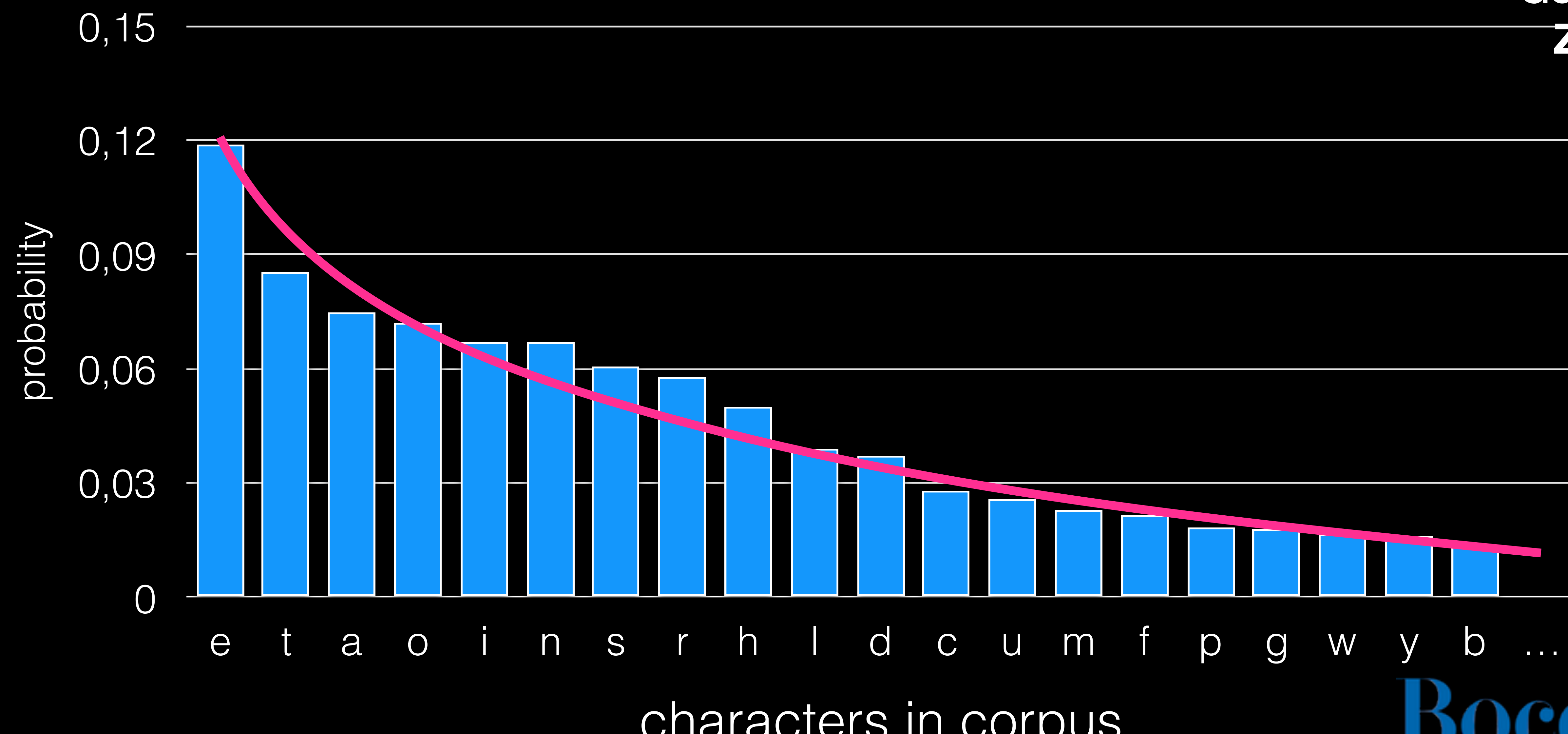
- special case of discrete distros (Bernoulli, categorical)
- all outcomes are equally likely, so it's hardest to predict
- Examples: fair coin toss, die roll

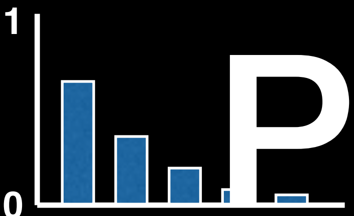
Power-Law Distribution

George Zipf

Parameters: k STEEPNESS OF CURVE

Function: $P(x;k) = kx^{-(k+1)}$





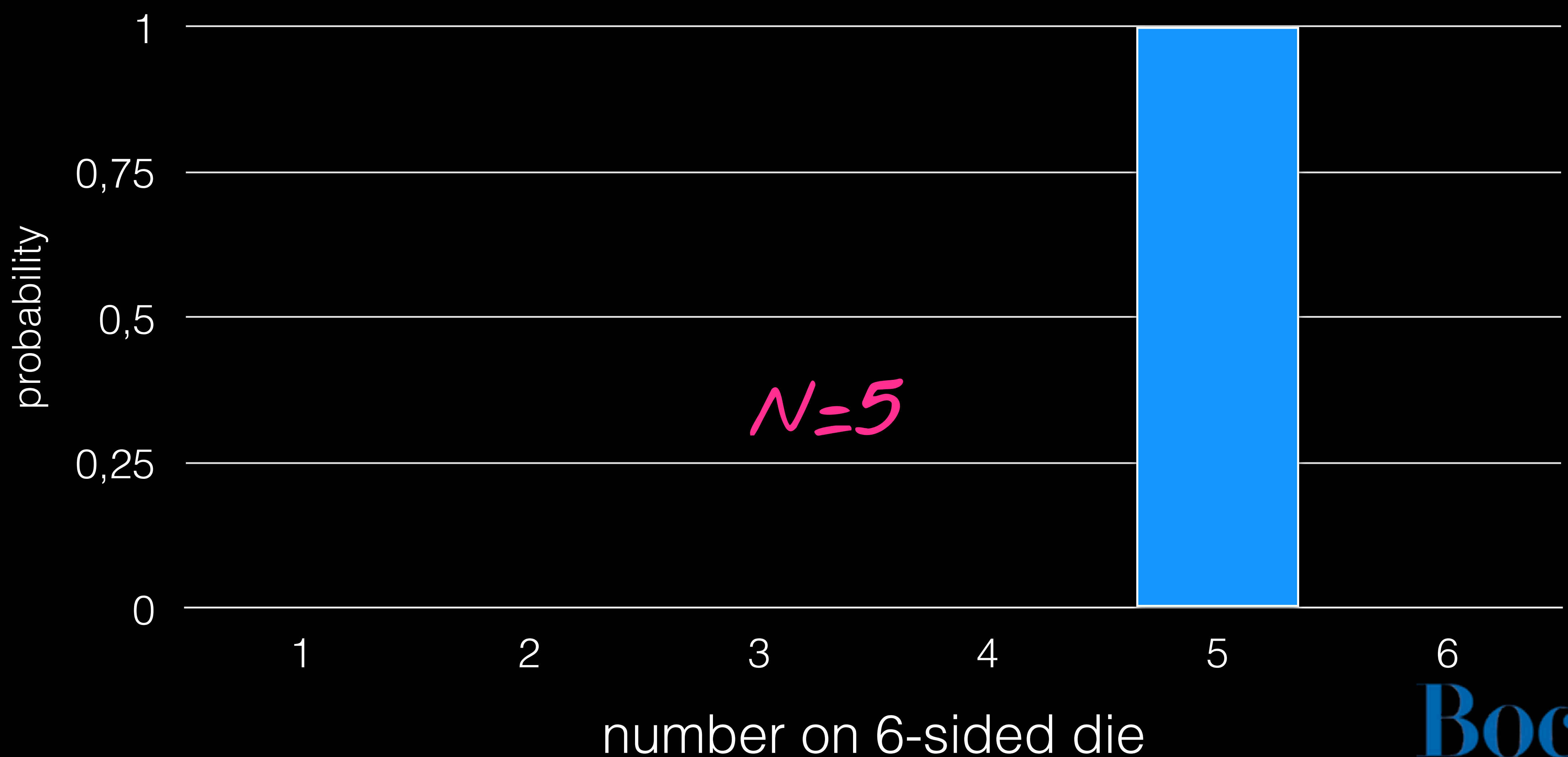
Power-Law Distribution

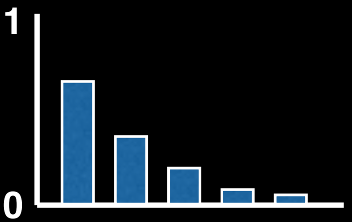
- has many outcomes
- most frequent outcome is *k* times more likely than second most frequent, which is *k* times more likely than third most frequent, etc.
- top N outcomes account for majority of observations
- easy to predict outcome of a random draw
- has a "long tail" of rare outcomes
- mean and median are very different!
- Examples: frequency of words, sounds, or letters in a language, city sizes, wealth distribution. "rich get richer" effect

One-Hot Distribution

Parameters: n ONLY TRUE EVENT

Function: ***P(x; n) = 1 if x=n; else 0***





One-Hot Distribution

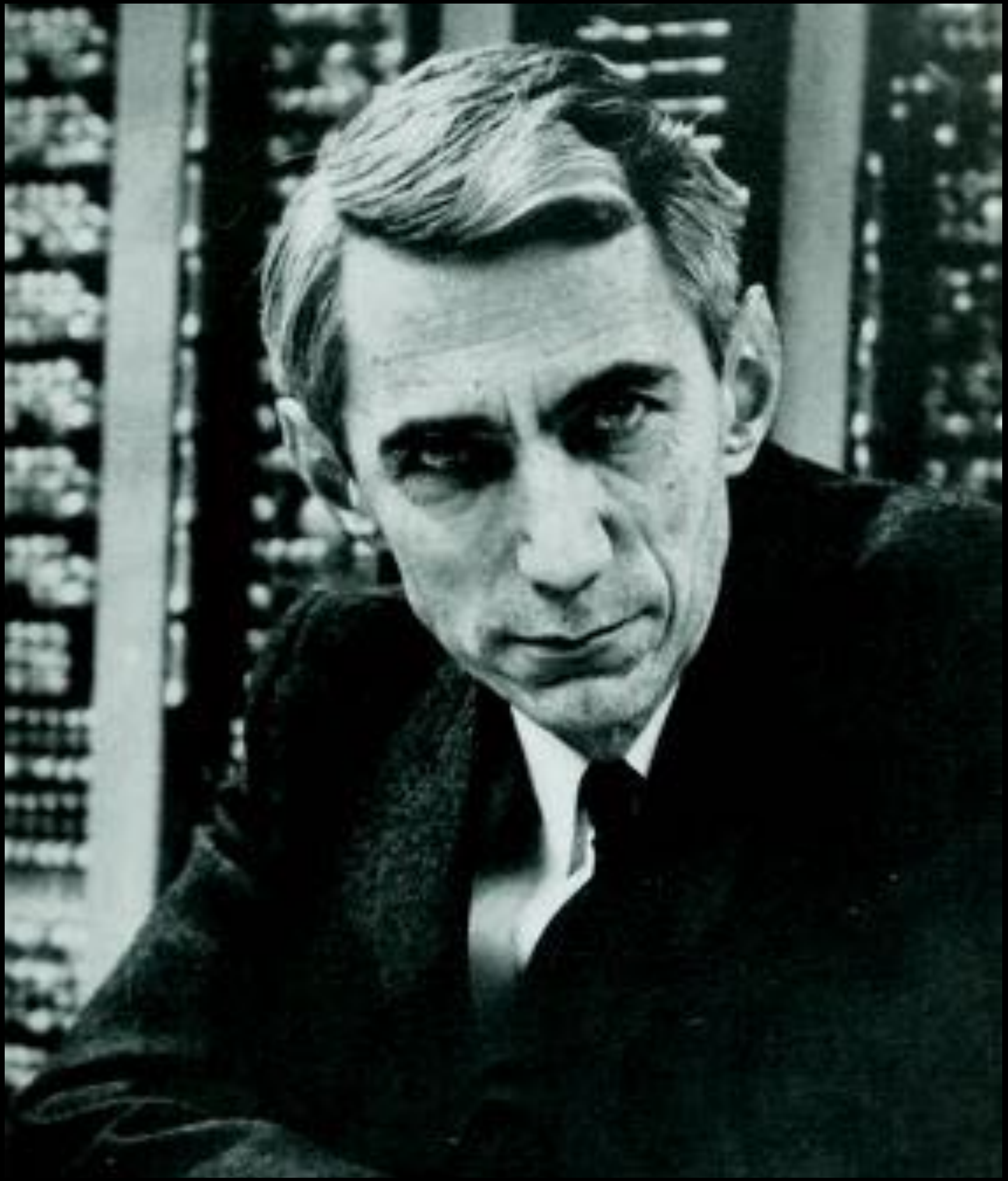
- special case of discrete distros (Bernoulli, categorical)
- easy way to represent a single truth (neural networks)
- Examples: correct answer in multi-class problems

How Skewed are We?
Entropy

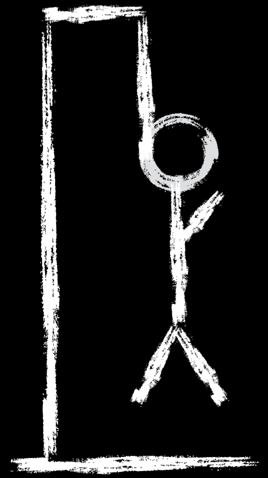
Claude Shannon

1916-2001

- His master's thesis founded a new field: digital circuits
- Invented entropy to quantify language – and a flame-throwing trumpet
- Enabled NLP, cryptography, modern computers…
- Died of Alzheimer's, oblivious to his own inventions' impact



Shannon Game

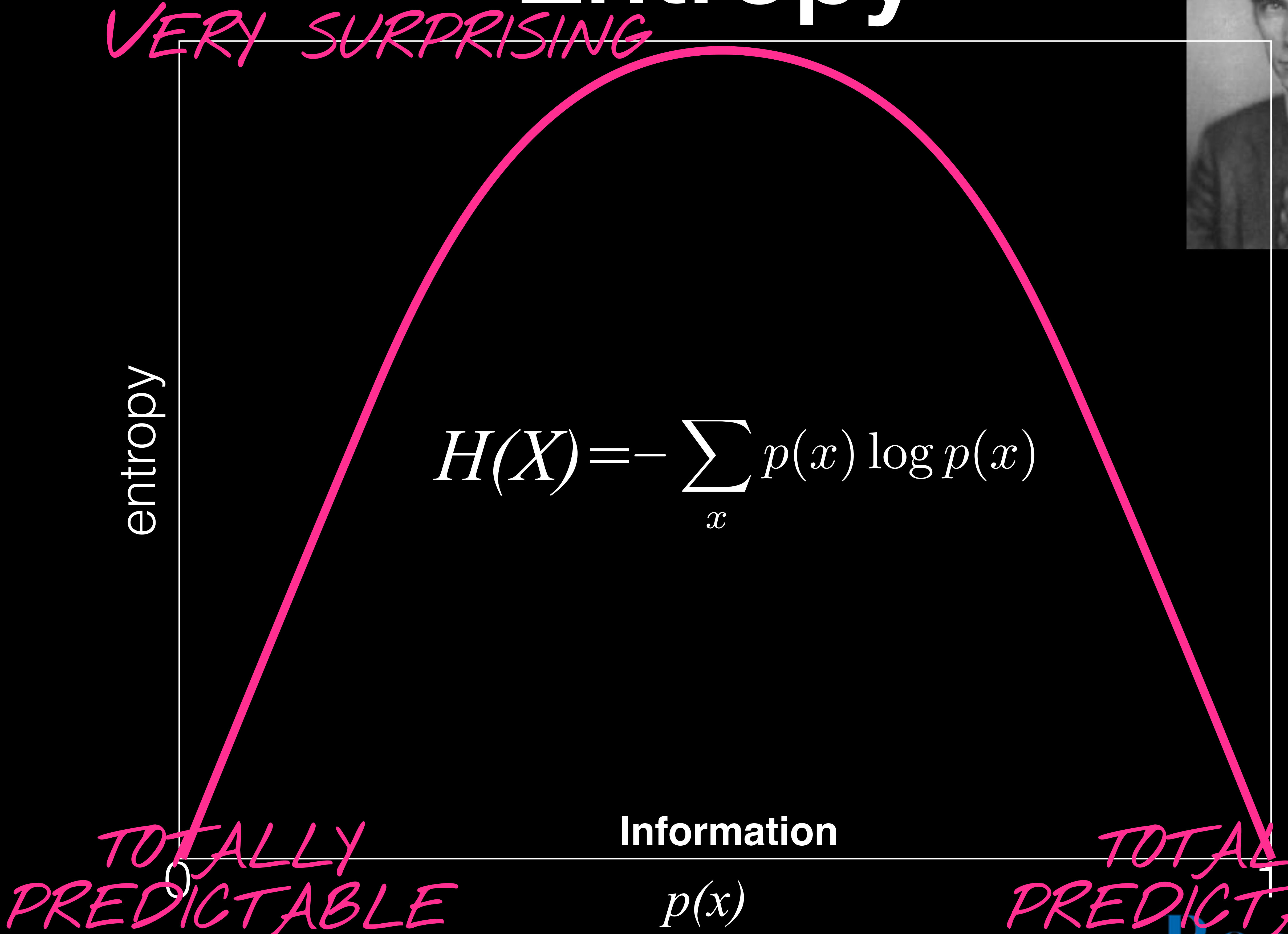


WHAT'S THE NEXT WORD?

The	house			
A	friend			
Then	dog			
If	car			
When	water	with	water	?
My	**hovercraft**	**of**	**eels**	.
He	pants	to	air	!
You	God			
I	word	…	…	
…	…			

HOW MUCH MORE, CLAUDE?

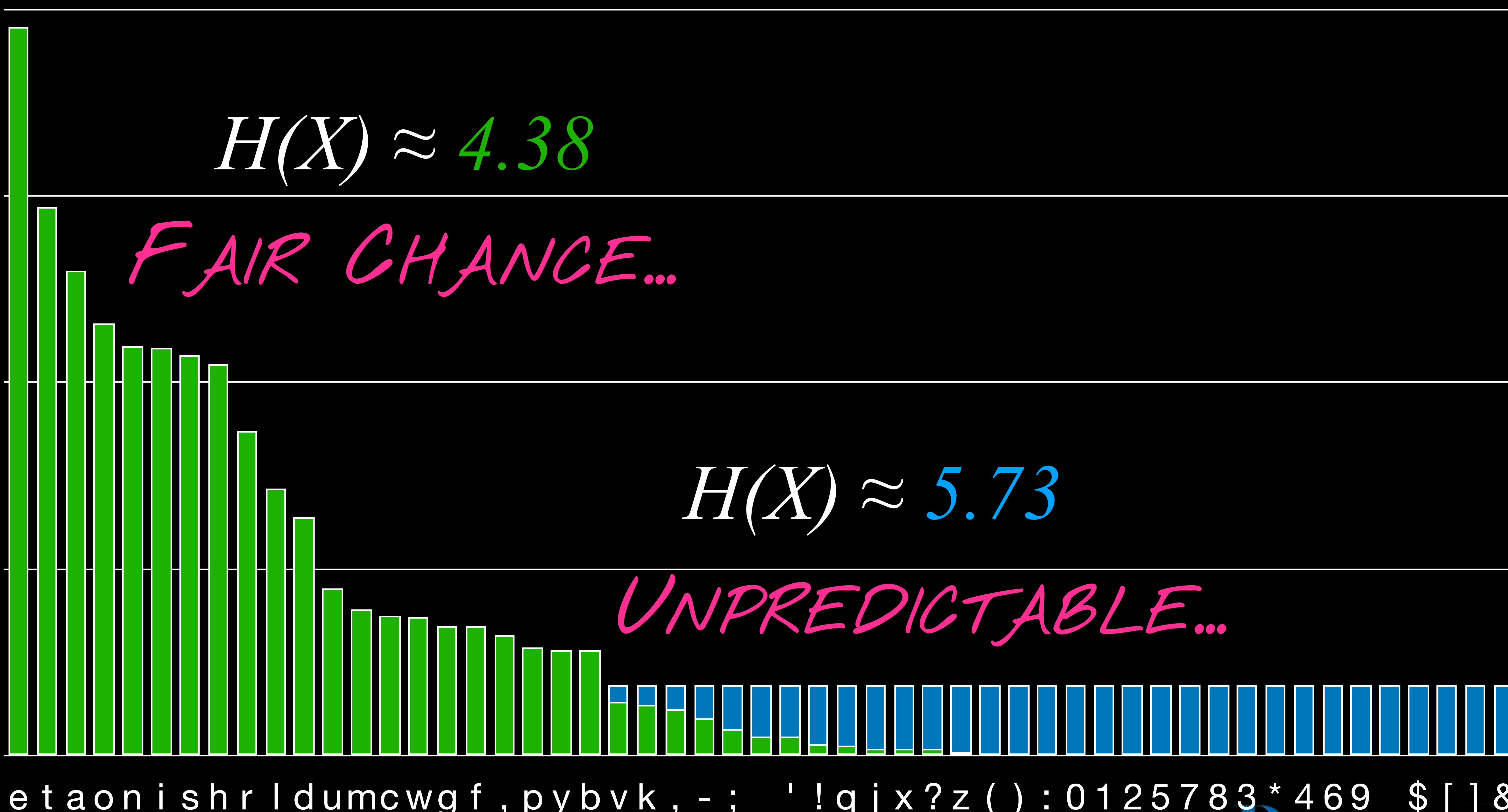
Entropy



$$H(X) = -\sum_x p(x) \log p(x)$$

Entropy in Use

WHAT'S THE NEXT LETTER?

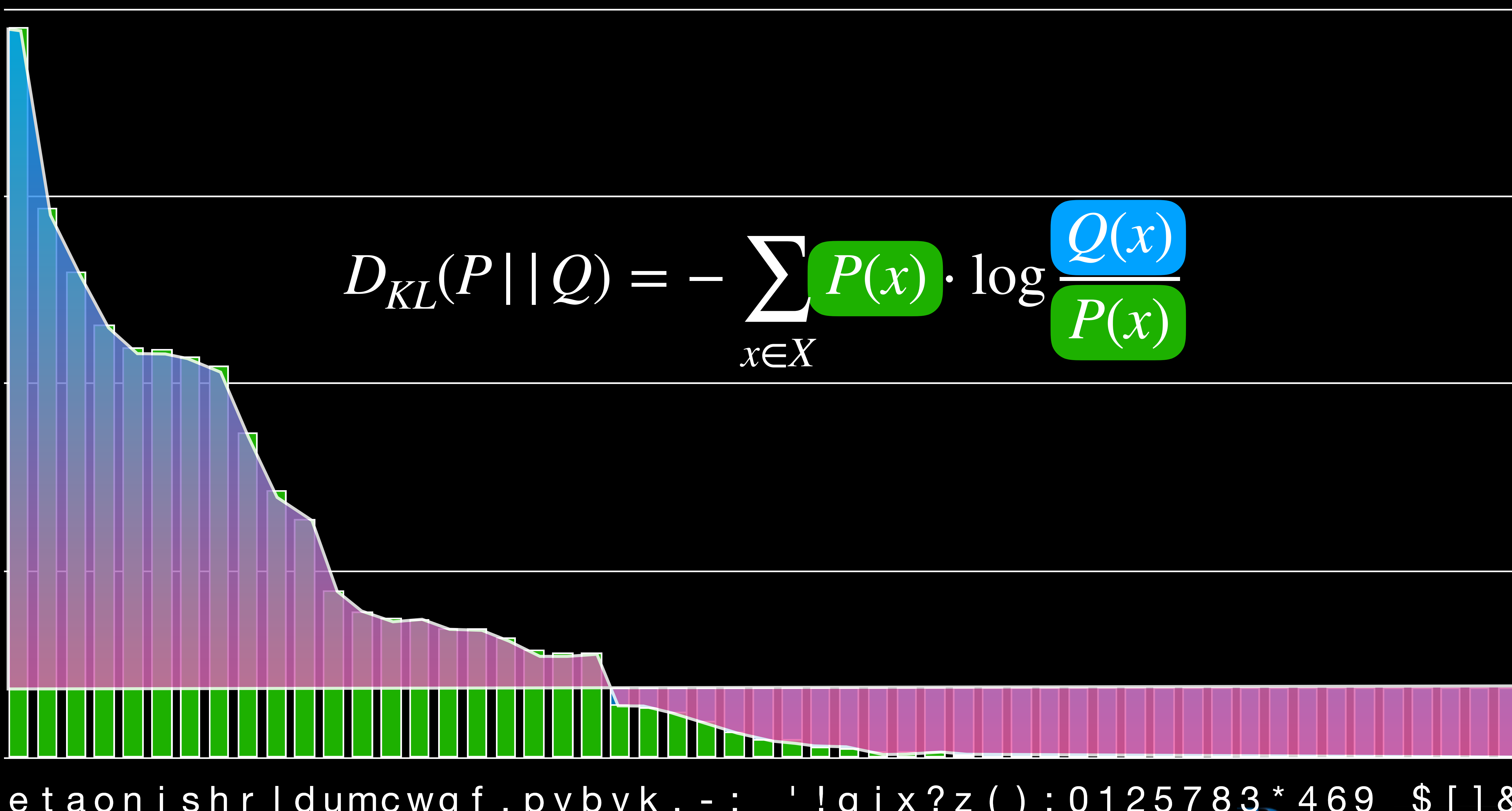


What's the Difference: Kullback-Leibler Divergence

Entropy in Use

"Relative Entropy"

$$D_{KL}(P\,||\,Q) = -\sum_{x \in X} P(x) \cdot \log \frac{Q(x)}{P(x)}$$



Frequent Company: Pointwise Mutual Information

Some are not like the Others

FREQUENT BIGRAMS

frying pan

social media

and thought

New York

of the

COLLOCATIONS

Mutual Informativity

How well can we guess the blank?

social _____

and _____

_____ media

_____ the

Pointwise Mutual Information

CHANCE OF SEEING THEM TOGETHER

$$PMI(x, y) = \log \frac{P(x, y)}{P(x)P(y)}$$

...SEEING EITHER

x	y	c(x)	c(y)	c(xy)	P(x)	P(y)	P(x, y)	PMI(x; y)
moby	dick	83	83	82	0.0003	0.0003	0.0003	**3.48**
captain	ahab	327	511	61	0.0013	0.0020	0.0002	**1.97**
white	whale	280	1150	106	0.0011	0.0045	0.0004	**1.93**
under	the	119	14175	45	0.0005	0.0553	0.0002	**0.83**
is	a	1690	4636	110	0.0066	0.0181	0.0004	**0.56**

c(X) = 256,149

c(XY) = 256,148

Bocconi

Wrapping up

Take home points

- **Regular expressions** allow us to search for flexible patterns
- Word sequences can be seen as discrete **probability distributions**
- **Entropy** allows us to quantify how surprising/predictable an outcome is
- **KL-divergence** tells us how different two distributions are
- **PMI** tells us how likely one word is to occur with/without another to find **collocations**