

(Advanced) Natural Language Processing

Jacob Andreas / MIT 6.806-6.864 / Spring 2021

An alien broadcast

f84hh4z18da4dzwrzo40hizeb3zm8bbz9e8dzj74z1e0h3z0iz0zded4n42kj814z38h42jehzdels9z
chz18da4dz8iz2708hc0dze5z4bi4184hzdlzj74z3kj27zfk1b8i78d6z6hekfhk3ebf7z06d4mzvvz
o40hizeb3z0d3z5ehc4hz2708hc0dze5z2edieb830j43z6eb3z584b3izfb2m0izd0c43z0zded4n42
kj814z38h42jehze5zj78iz1h8j8i7z8d3kijh80bz2ed6bec4h0j4z05ehcze5z0i14ijeized24zki
43zjezc0a4za4djz2860h4jj4z58bj4hiz70iz20ki43z0z7867f4h24dj064ze5z20d24hz340j7iz0
ced6z0z6hekfze5zmeha4hiz4nfei43zjez8jzceh4zj70dztqo40hiz06ezh4i40h274hizh4fehj43
zj74z0i14ijeiz5814hz2he283eb8j4z8izkdkik0bboh4i8b84djzed24z8jz4dj4hizj74zbkd6izm
8j7z4l4dz1h845z4nfeikh4izjez8jz20ki8d6iocfjecizj70jzi7emzkfz342034izb0j4hzh4i40h

Predictability

f84hh4z18da4dzwrzo40hizeb3zm8bbz9e8dzj74z1e0h3z0iz0zded4n42kj8l4z38h42jehzde1s9z
chz18da4dz8iz2708hc0dze5z4bi4184hzdlzj74z3kj27zfk1b8i78d6z6hekfhk3ebf7z06d4mzvvz
o40hizeb3z0d3z5ehc4hz2708hc0dze5z2edieb830j43z6eb3z584b3izfb2m0izd0c43z0zded4n42

$$p(X_t \mid X_{1:t-1})$$

Can I guess what character is coming next?

Predictability

f84hh4z18da4dzwrzo40hizeb3zm8bbz9e8dzj74z1e0h3z0iz0zded4n42kj814z38h42jehzde1s9z
chz18da4dz8iz2708hc0dze5z4bi4184hzdlzj74z3kj27zfk1b8i78d6z6hekfhk3ebf7z06d4mzvvz
o40hizeb3z0d3z5ehc4hz2708hc0dze5z2edieb830j43z6eb3z584b3izfb2m0izd0c43z0zded4n42

$$p(8 \mid 63b3z)$$

Can I guess what character is coming next?

Predictability

f84hh4z18da4dzwrzo40hizeb3zm8bbz9e8dzj74z1e0h3z0iz0zded4n42kj814z38h42jehzdels9z
chz18da4dz8iz2708hc0dze5z4bi4184hzdlzj74z3kj27zfk1b8i78d6z6hekfhk3ebf7z06d4mzvvz
o40hizeb3z0d3z5ehc4hz2708hc0dze5z2edieb830j43z6eb3z584b3izfb2m0izd0c43z0zded4n42

$$p(8 \mid \underline{63b3z}) = \frac{\#(z \ 8)}{\#(z)}$$

e.g. by counting frequencies?

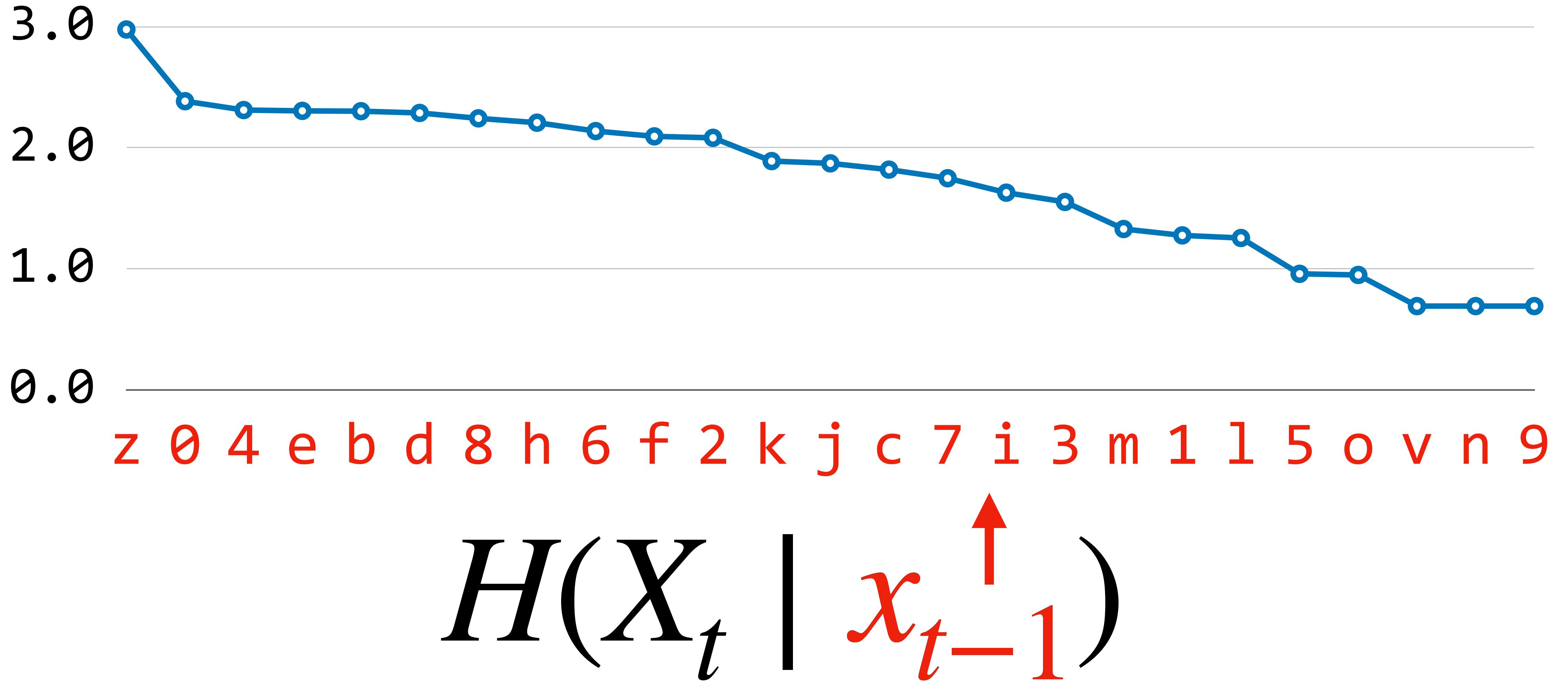
Predictability

f84hh4z18da4dzwrzo40hizeb3zm8bbz9e8dzj74z1e0h3z0iz0zded4n42kj814z38h42jehzde1s9z
chz18da4dz8iz2708hc0dze5z4bi4184hzdlzj74z3kj27zfk1b8i78d6z6hekfhk3ebf7z06d4mzvvz
o40hizeb3z0d3z5ehc4hz2708hc0dze5z2edieb830j43z6eb3z584b3izfb2m0izd0c43z0zded4n42

$$H(X_t \mid z) = - \sum_x p(x_t \mid z) \log p(x_t \mid z)$$

How much uncertainty do I have about the next character,
given that the last one was a z?

Predictability



Predictability

f84hh4z18da4dzwrzo40hizeb3zm8bbz9e8dzj74z1e0h3z0iz0zded4n42kj8l4z38h42jehzdels9z
chz18da4dz8iz2708hc0dze5z4bi4184hzdlzj74z3kj27zfk1b8i78d6z6hekfhk3ebf7z06d4mzvvz
o40hizeb3z0d3z5ehc4hz2708hc0dze5z2edieb830j43z6eb3z584b3izfb2m0izd0c43z0zded4n42

Hypothesis: “islands of local predictability”

Discrete structure?

f84hh4-18da4d-wr-o40hi-eb3-m8bb-9e8d-j74-1e0h3-0i-0-ded4n42kj8l4-38h42jeh-del9-
ch-18da4d-8i-2708hc0d-e5-4bi4184h-dl-j74-3kj27-fk1b8i78d6-6hekfhk3ebf7-06d4m-vv-
o40hi-eb3-0d3-5ehc4h-2708hc0d-e5-2edieb830j43-6eb3-584b3i-fb2m0i-d0c43-0-ded4n42

Hypothesis: “islands of local predictability”

Discrete structure?

f84hh4-18da4d-wr-o40hi-eb3-m8bb-9e8d-j74-1e0h3-0i-0-ded4n42kj814-38h42jeh-dels9-
ch-18da4d-8i-2708hc0d-e5-4bi4184h-dl-j74-3kj27-fk1b8i78d6-6hekfhk3ebf7-06d4m-vv-
o40hi-eb3-0d3-5ehc4h-2708hc0d-e5-2edieb830j43-6eb3-584b3i-fb2m0i-d0c43-0-ded4n42

This segmentation reveals lots of repeated units!

Ordering rules?

1o-j74-kic1-5hec-r9xv-j7hek67-r9yx-j74-0dc2-74b3-0-i4h84i-e5-

m4bb-0jj4d343-0ddk0b-2ed54h4d24i-ik1i4gk4djbo-0-i4h84i-e5-d0j8ed0b-

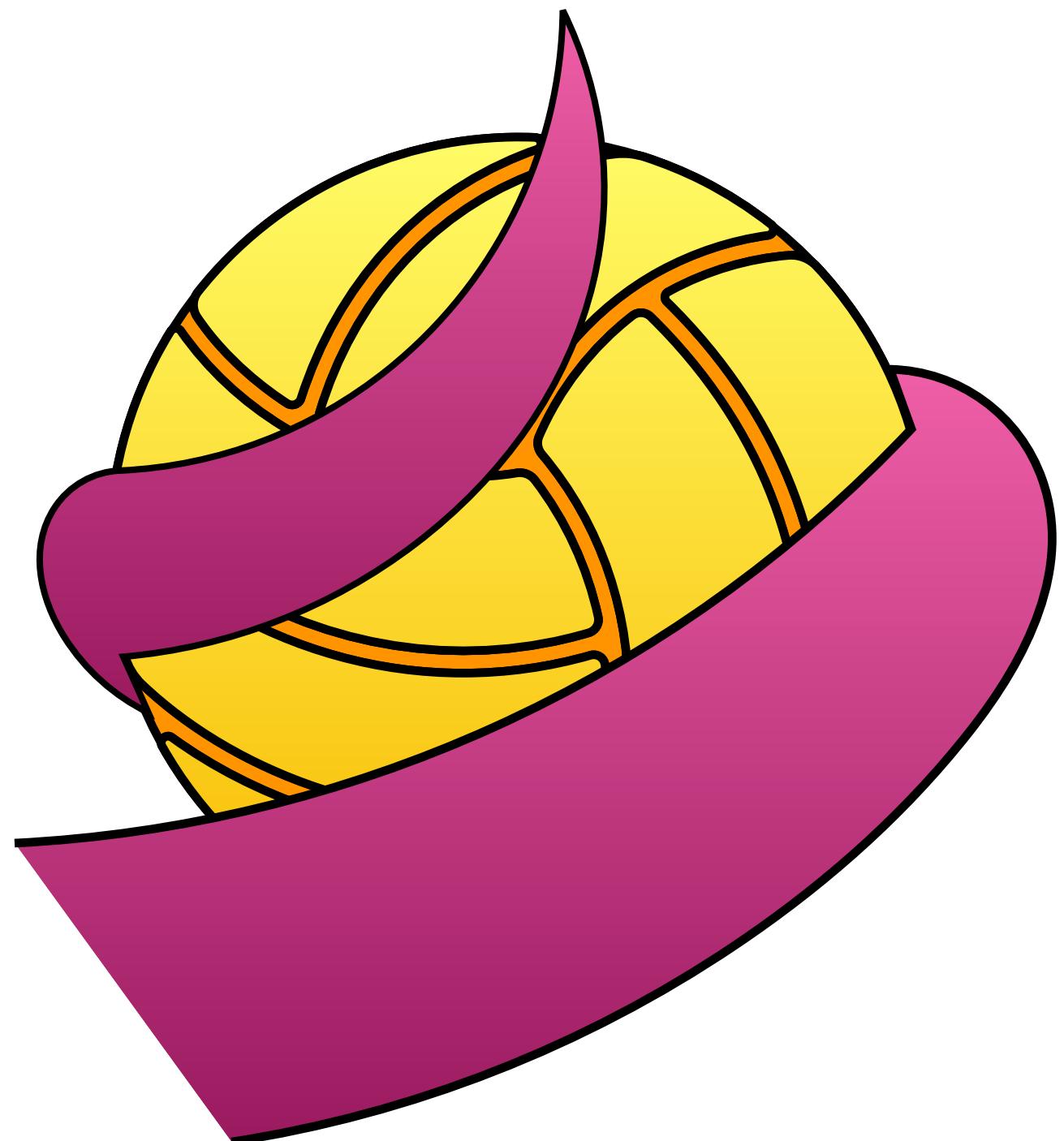
c4jh82-2ed54h4d24i-9e8djbo-ifedieh43-1o-j74-0dc2-j74-ki-c4jh82-0iie280j8ed-

kic0-j74-34f0hjc4dj-e5-2ecc4h24-0d3-j74-d0j8ed0b-8dij8jkj4-e5-ij0d30h3i-0d3-

j427debe6o-d8ij-m4h4-74b3-5hec-r9yx-j7hek67-r99t

Some units seem to occur in similar contexts.

With perceptual information

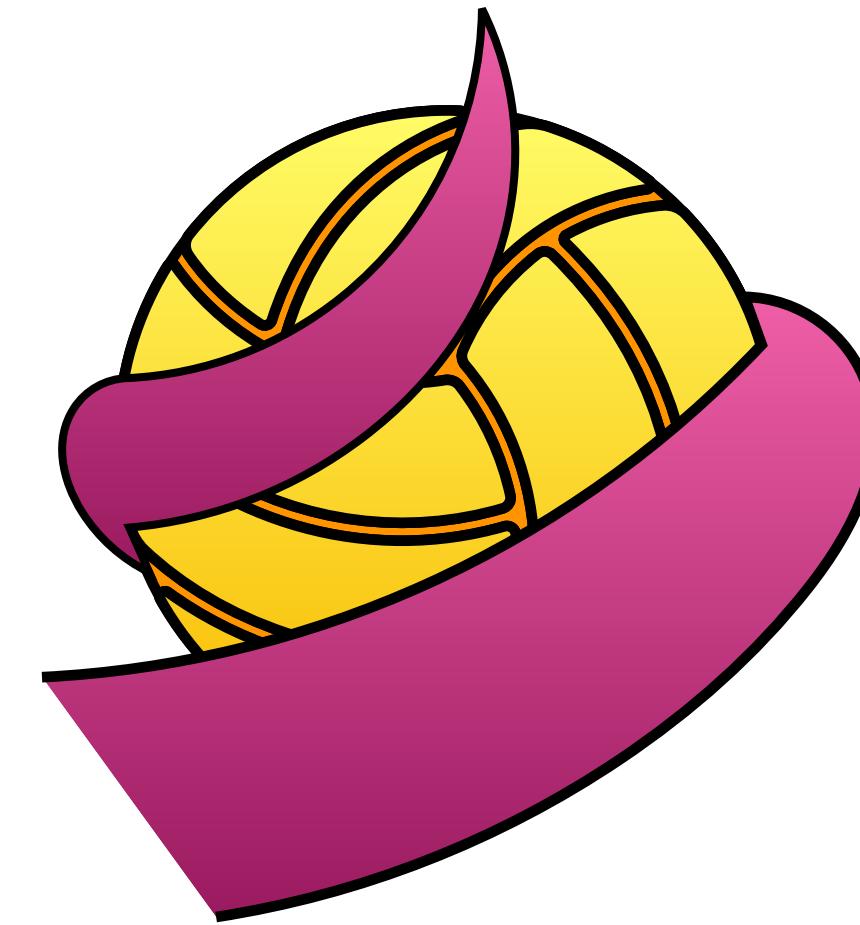


43zjjezc0a4za4djh2860h4jj4z58bj4hiz70iz20
ki43z0z7867f4h24dj064ze5z20d24hz340j7iz0
ced6z0z6hekfze5zmeha4hiz4nfei43zjez8jzce
h4zj70dztqo40hiz06ezh4i40h274hizh4fehj43
zj74z0i14ijeiz5814hz2he283eb8j4z8izkdkik
0bboh4i8b84djzed24z8jz4dj4hizj74zbkd6izm
8j7z4l4dz1h845z4nfeikh4izjez8jz20ki8d6io
cfjecizj70jzi7emzkfz342034izb0j4hzh4i40h

Questions: full grounded meaning

43-je-c0a4-a4dj-2860h4jj4-58bj4hi-70i-20

ki43-0-7867f4h24dj064-e5-20d24h-kdkiko-0



j74-0i14ijei-5814h-2he283eb8j4-8i-kdkiko

bboh4i8b84dj-ed24-8j-4dj4hi-j74-bkd6i-mX



Questions: full grounded meaning

Some units co-occur with features
of the world described by the messages

j74-0i14ijei-5814h-2he283eb8j4-8i-**kdkiko**

bboh4i8b84dj-ed24-8j-4dj4hi-j74-bkd6i-mX



Questions: full grounded meaning

$p(\text{ } \mid X_{1:t})$



$\mid \text{kdkik0})$

j74-0i14ije-5814h-2he283eb8j4-8i-kdkik0

bboh4i8b84dj-ed24-8j-4dj4hi-j74-bkd6i-mX

and help us accurately predict the context (and meaning) of new messages

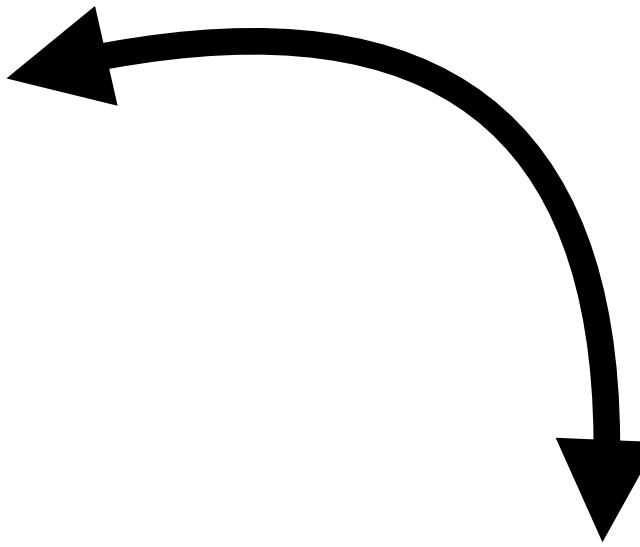


Questions: controlled generation

$p(\text{qxx} - 93\text{ar} \mid \text{take me to your leader})$

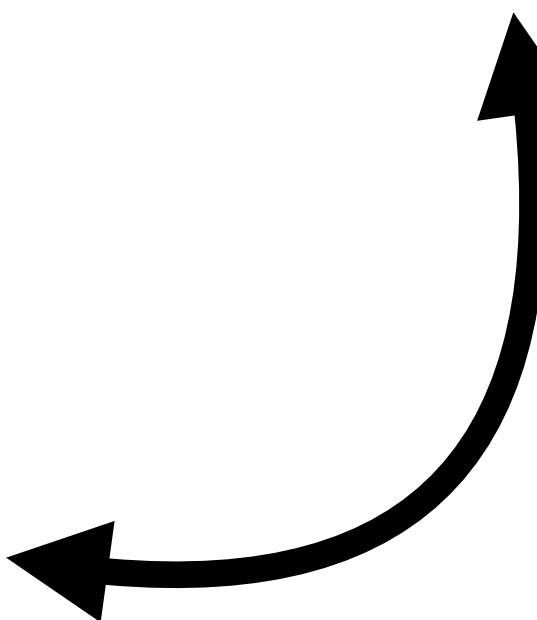
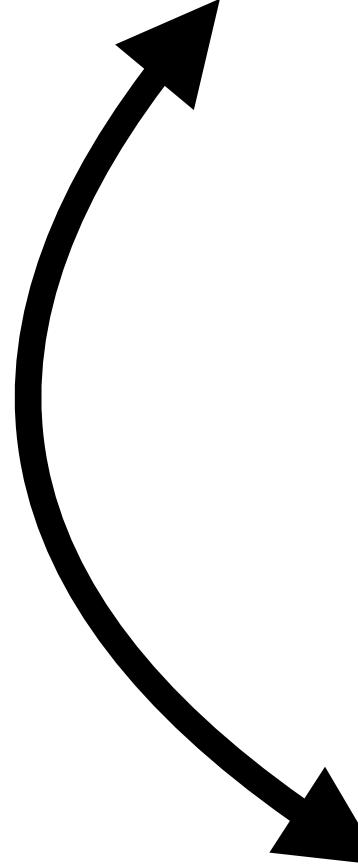
or even to generate new messages based on meanings we want to communicate.

Probabilistic models
of language



Linguistic structure
& “speaker intuition”

meaning & use



f84hh4z18da4dzwrzo40hizeb3zm8bbz9e8dzj74z1e0h3z0iz0zded4n42kj814z38h42jehzde1s9z
chz18da4dz8iz2708hc0dze5z4bi4184hzdlzj74z3kj27zfk1b8i78d6z6hekfhk3ebf7z06d4mzvvz
o40hizeb3z0d3z5ehc4hz2708hc0dze5z2edieb830j43z6eb3z584b3izfb2m0izd0c43z0zded4n42
kj814z38h42jehze5zj78iz1h8j8i7z8d3kijh80bz2ed6bec4h0j4z05ehcze5z0i14ijeized24zki
43zjezc0a4za4djz2860h4jj4z58bj4hiz70iz20ki43z0z7867f4h24dj064ze5z20d24hz340j7iz0
ced6z0z6hekfe5zmeha4hiz4nfei43zjez8jzceh4zj70dztqo40hiz06ezh4i40h274hizh4fehj43
zj74z0i14ijeiz5814hz2he283eb8j4z8izkdkik0bboh4i8b84djzed24z8jz4dj4hizj74zbkd6izm
8j7z4l4dz1h845z4nfeikh4izjez8jz20ki8d6iocfjecizj70jzi7emzkfz342034izb0j4hz4i40h

This is what all datasets look like to NLP models

((Human (language)) (processing))

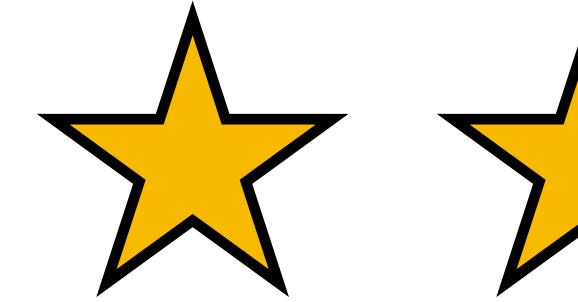
Language as input

Text classification

(input)

This film will ruin your childhood.

(output)



Language as input

Machine translation

(input)

Le programme a été mis en application.

(output)

The program was implemented.

Language as input

Automatic summarization

(input)

(output)

A Maximum-Entropy-Inspired Parser *

Eugene Charniak
Brown Laboratory for Linguistic Information Processing
Department of Computer Science
Brown University, Box 1910, Providence, RI 02912
ec@cs.brown.edu

Abstract

We present a new parser for parsing down to Penn tree-bank style parse trees that achieves 90.1% average precision/recall for sentences of length 40 and less, and 89.5% for sentences of length 100 and less when trained and tested on the previously established [5,9,10,15,17] “standard” sections of the Wall Street Journal tree-bank. This represents a 13% decrease in error rate over the best single-parser results on this corpus [9]. The major technical innovation is the use of a “maximum-entropy-inspired” model for conditioning and smoothing that let us successfully to test and combine many different null hypothesis. We also show how this model can be used to implement a maximum-entropy parser.

is s . Then for any s the parser returns the parse π that maximizes this probability. That is, the parser implements the function

$$\begin{aligned}\arg \max_{\pi} p(\pi | s) &= \arg \max_{\pi} p(\pi, s) \\ &= \arg \max_{\pi} p(\pi).\end{aligned}$$

What fundamentally distinguishes probabilistic generative parsers is how they compute $p(\pi)$, and it is to that topic we turn next.

2 The Generative Model

The model assigns a probability to a parse by summing the probabilities of all the paths

We present a new parser for the Penn Treebank. The parser achieves 90% accuracy using a “maximum-entropy-inspired” model for conditioning and smoothing that let us combine many different conditioning events.

Language as output

Generation from structured data

(input)

Frederick Parker-Rhodes	
Born	21 November 1914 Newington, Yorkshire
Died	2 March 1987 (aged 72)
Residence	UK
Nationality	British
Fields	Mycology , Plant Pathology , Mathematics , Linguistics , Computer Science
Known for	Contributions to computational linguistics , combinatorial physics , bit-string physics , plant pathology , and mycology

(output)

Frederick Parker-Rhodes (21 March 1914 – 21 November 1987) was an English linguist, plant pathologist, computer scientist, mathematician, mystic, and mycologist.

[Lebret et al. 2016]

Language as interface

Task-oriented dialog

What do I have today?

You have five events scheduled. The first is a one-on-one with Anjali.

Can you reschedule that for tomorrow at the same time?

Sure, I've sent an email to let her know.

Can you add a cram session with Nick and his manager? We'll need a room on the 10th floor.

Instruction following



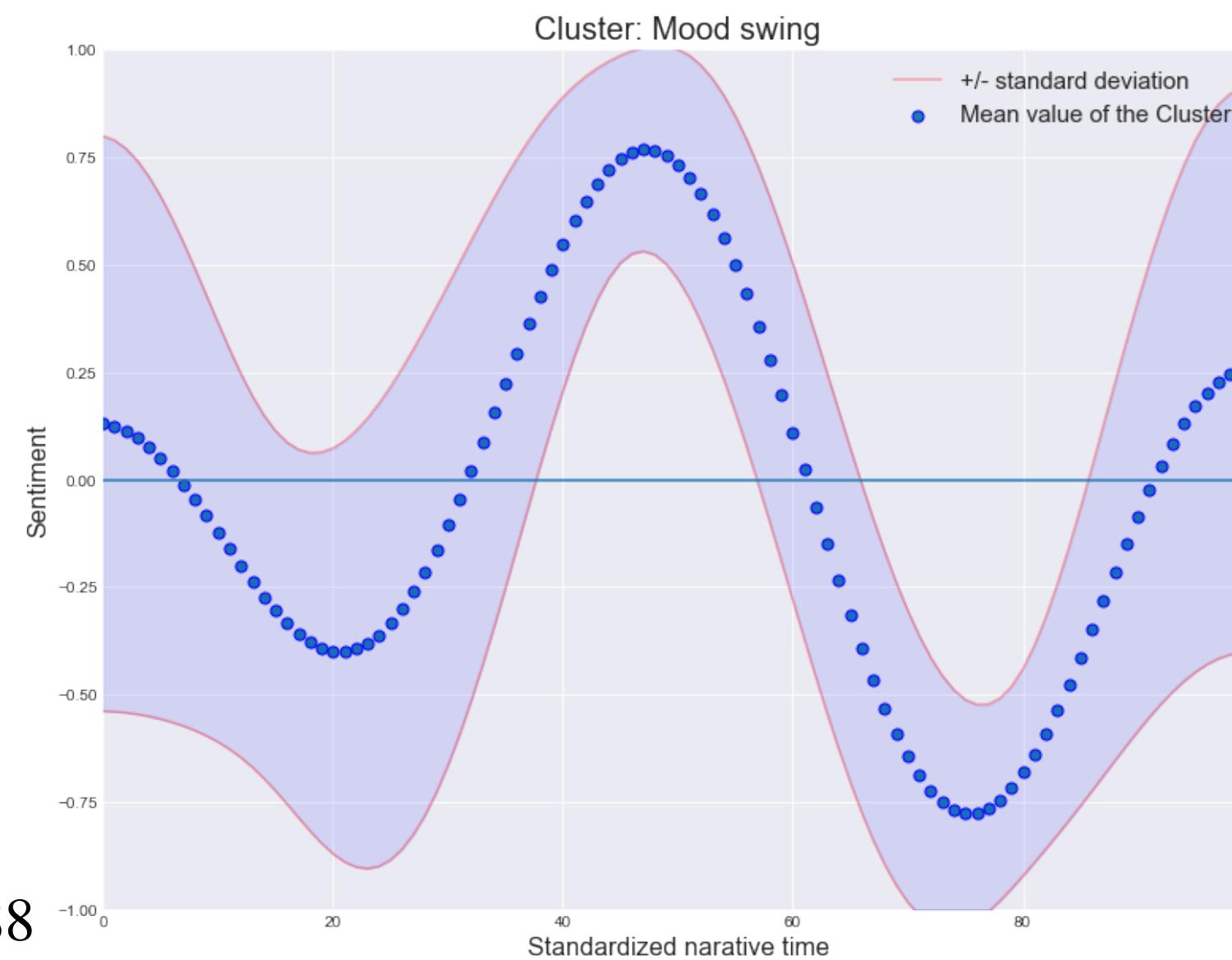
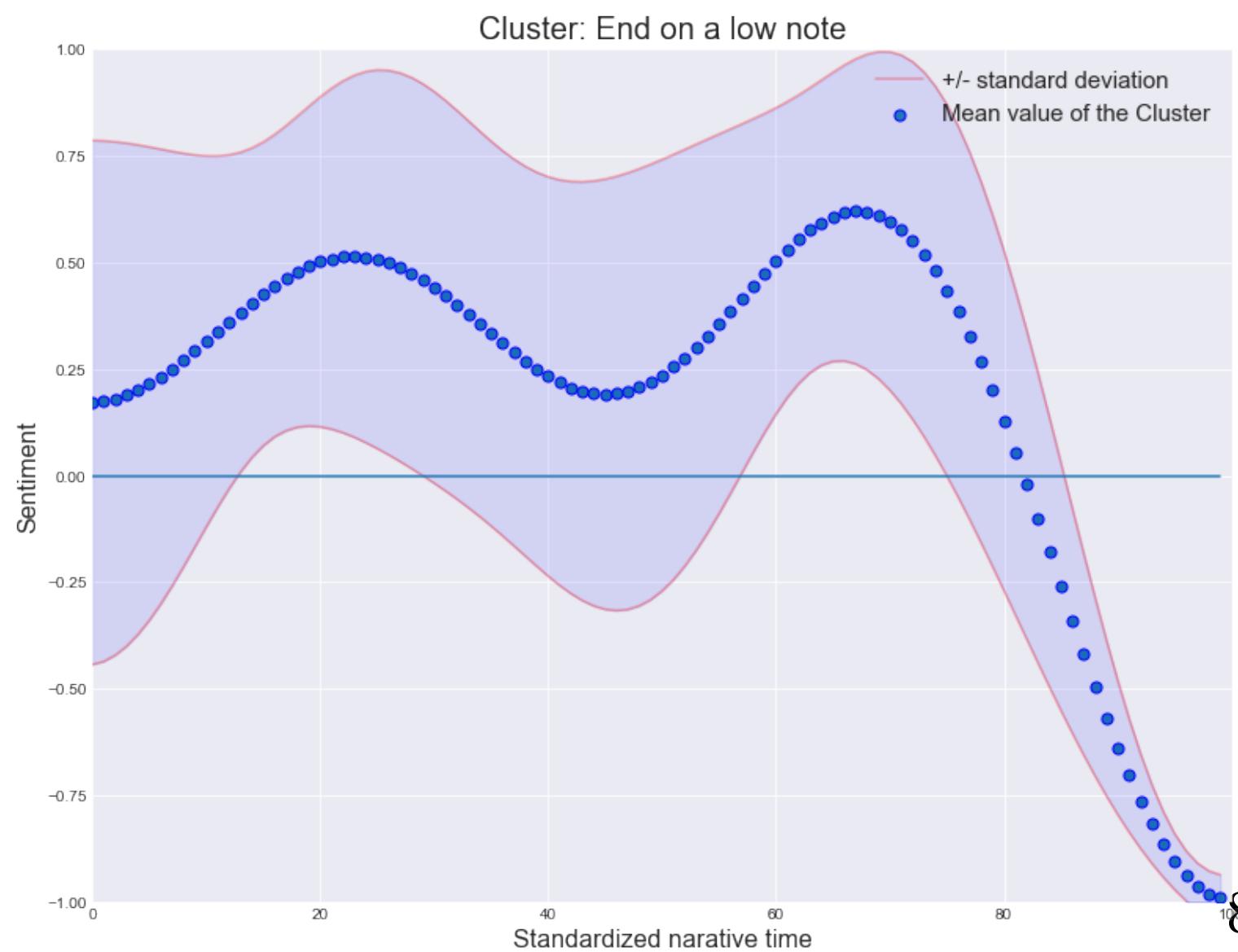
[Tellex et al., 2011]

Language as data

Computational social science

sentiment trajectories in Youtube videos

predictiveness of political stance



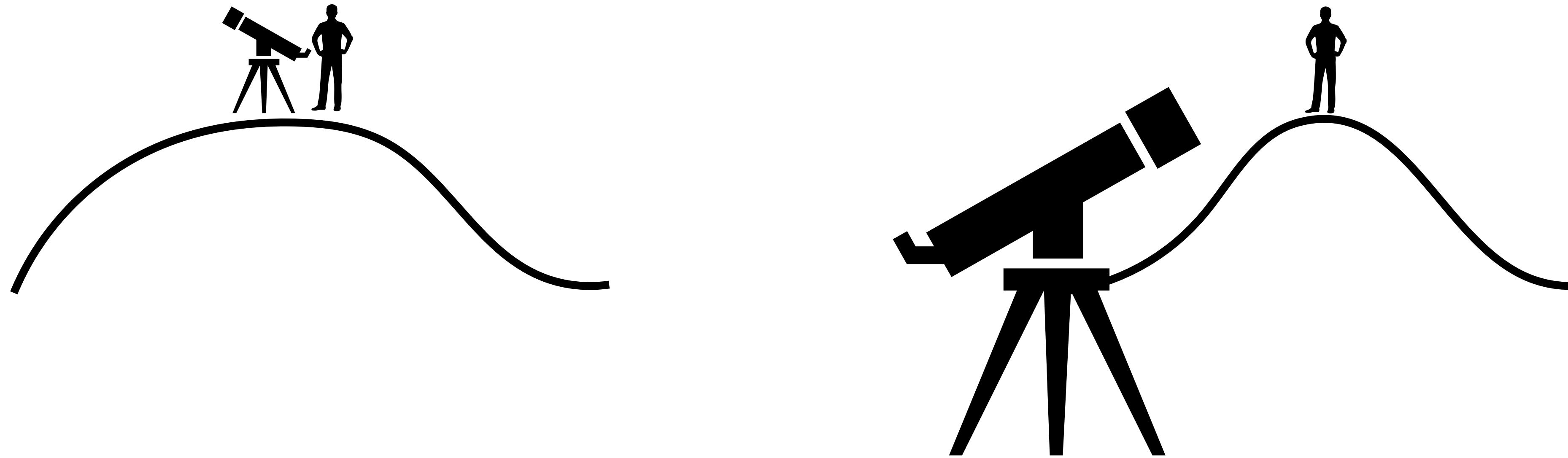
Cluster	Political leaning	
	Left	Right
Rags to riches	-0.96	0.96
Riches to rags	1.09	-1.09
Downhill from here	-3.25*	3.25*
End on a high note	1.28	-1.28
Uphill from here	2.74*	-2.74*
End on a low note	-2.44	2.44
Mood swing	1.13	-1.13

Our toolbox

Probabilistic modeling

I saw the man on the hill with the telescope.

Probabilistic modeling

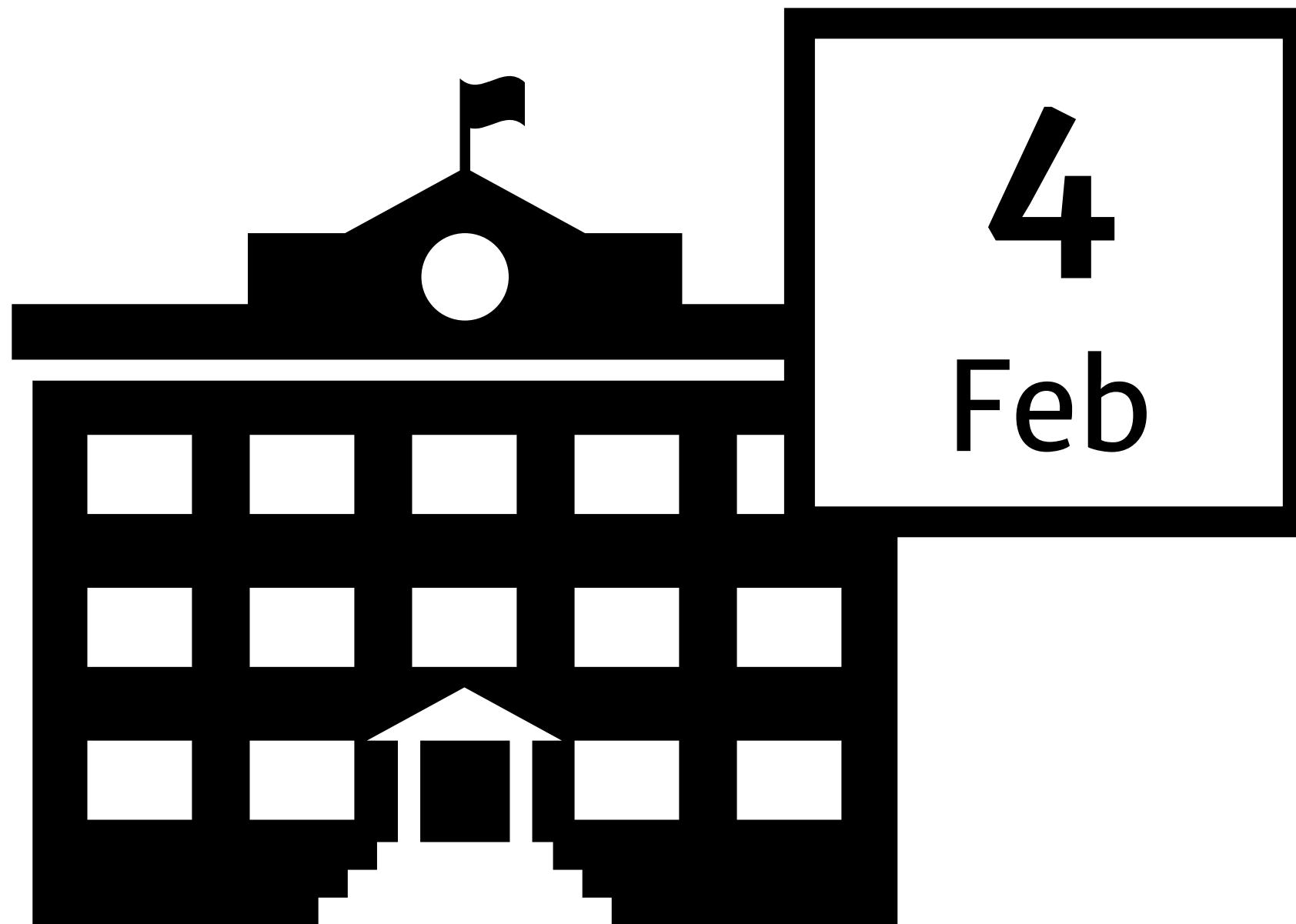


I saw the man on the hill with the telescope.

Probabilistic modeling

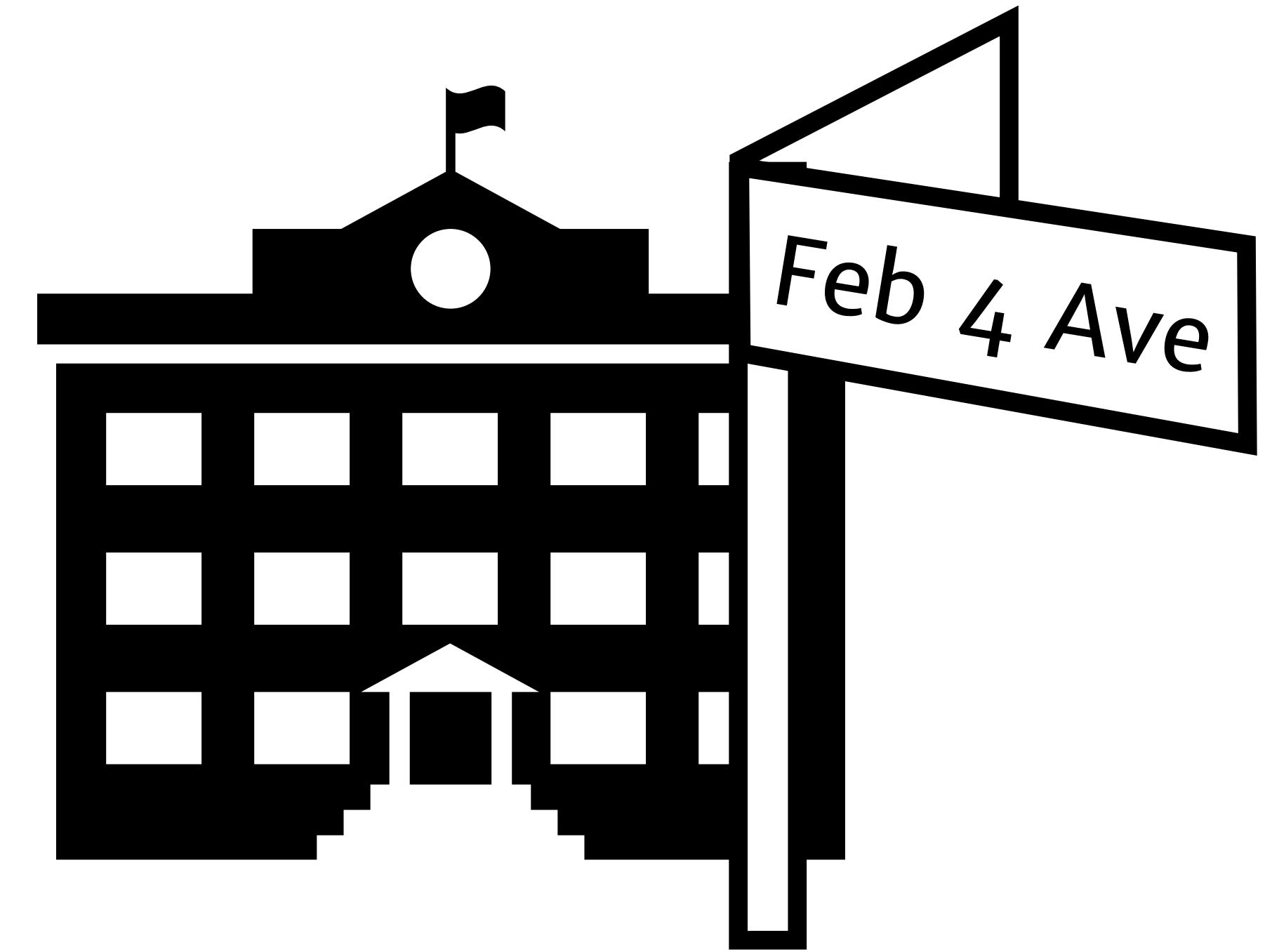
I went to the restaurant on February 4th.

Probabilistic modeling



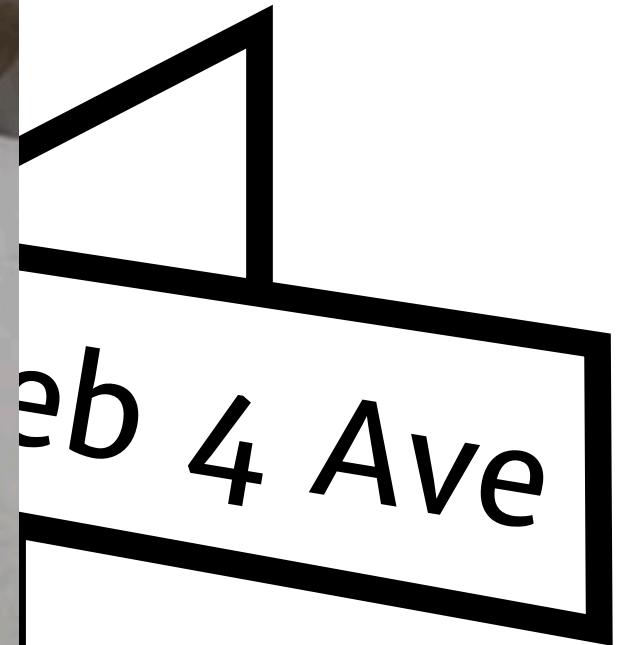
I went to the restaurant on February 4th.

Probabilistic modeling

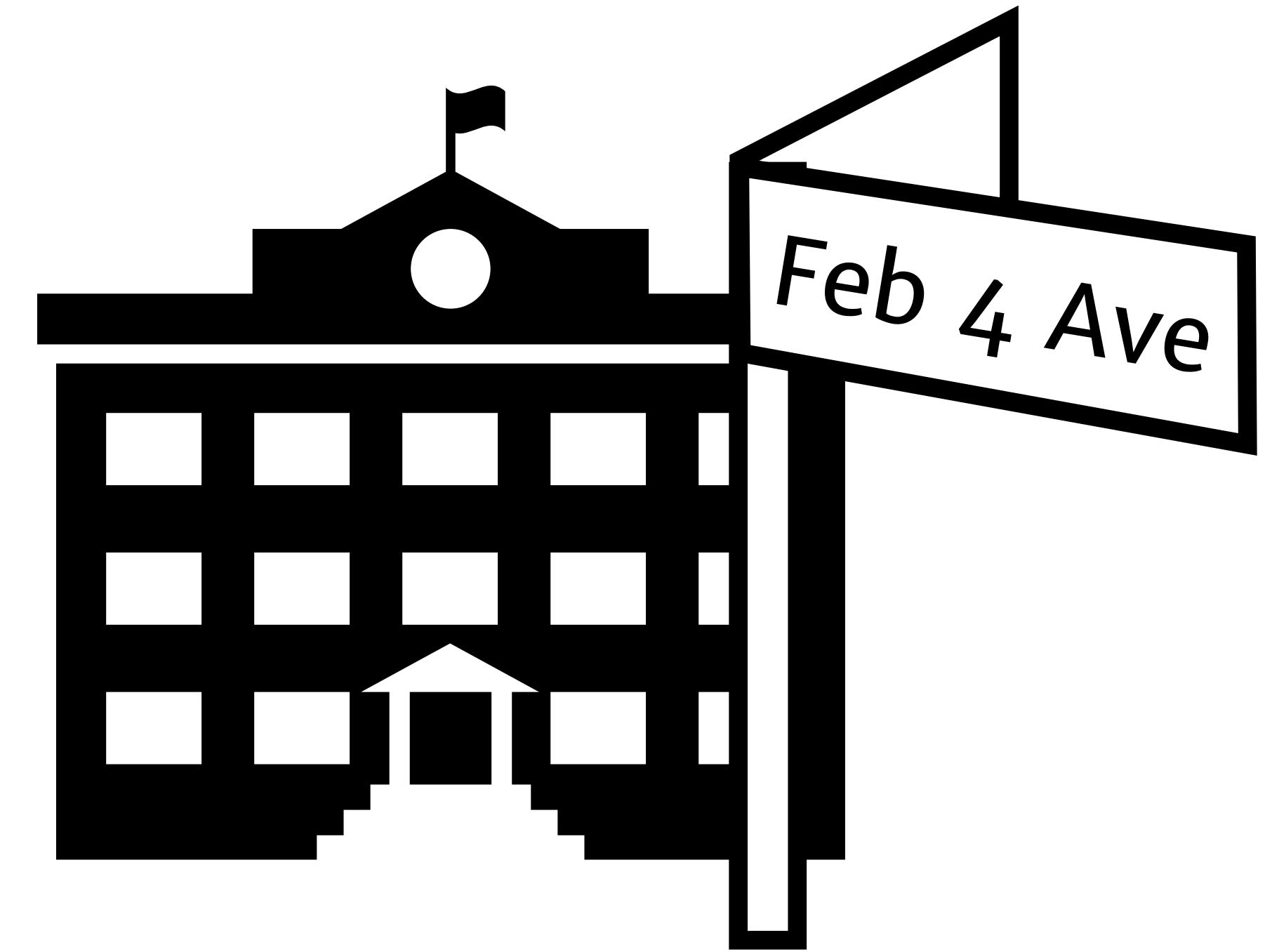
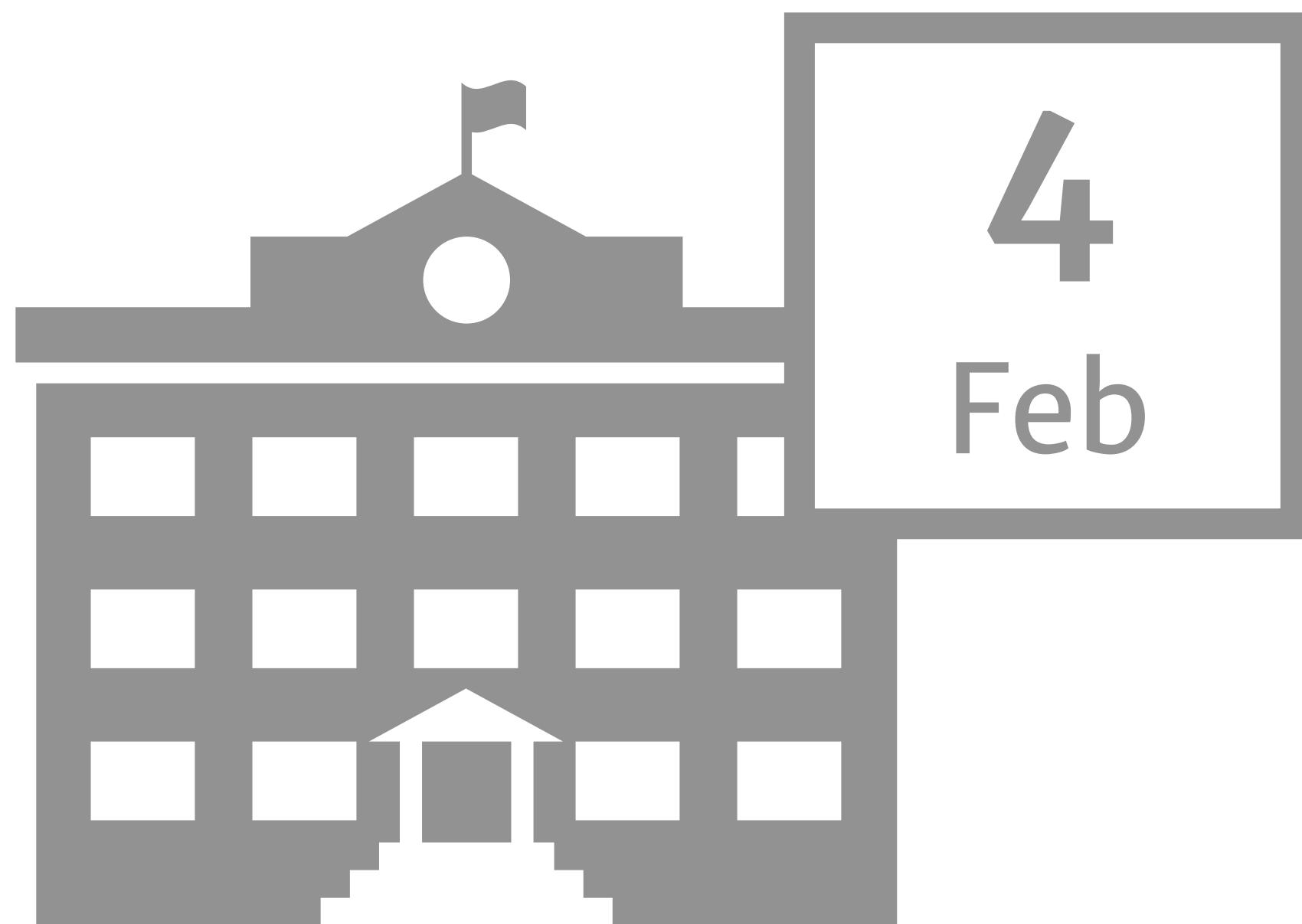


I went to the restaurant on February 4th.

Probabilistic modeling

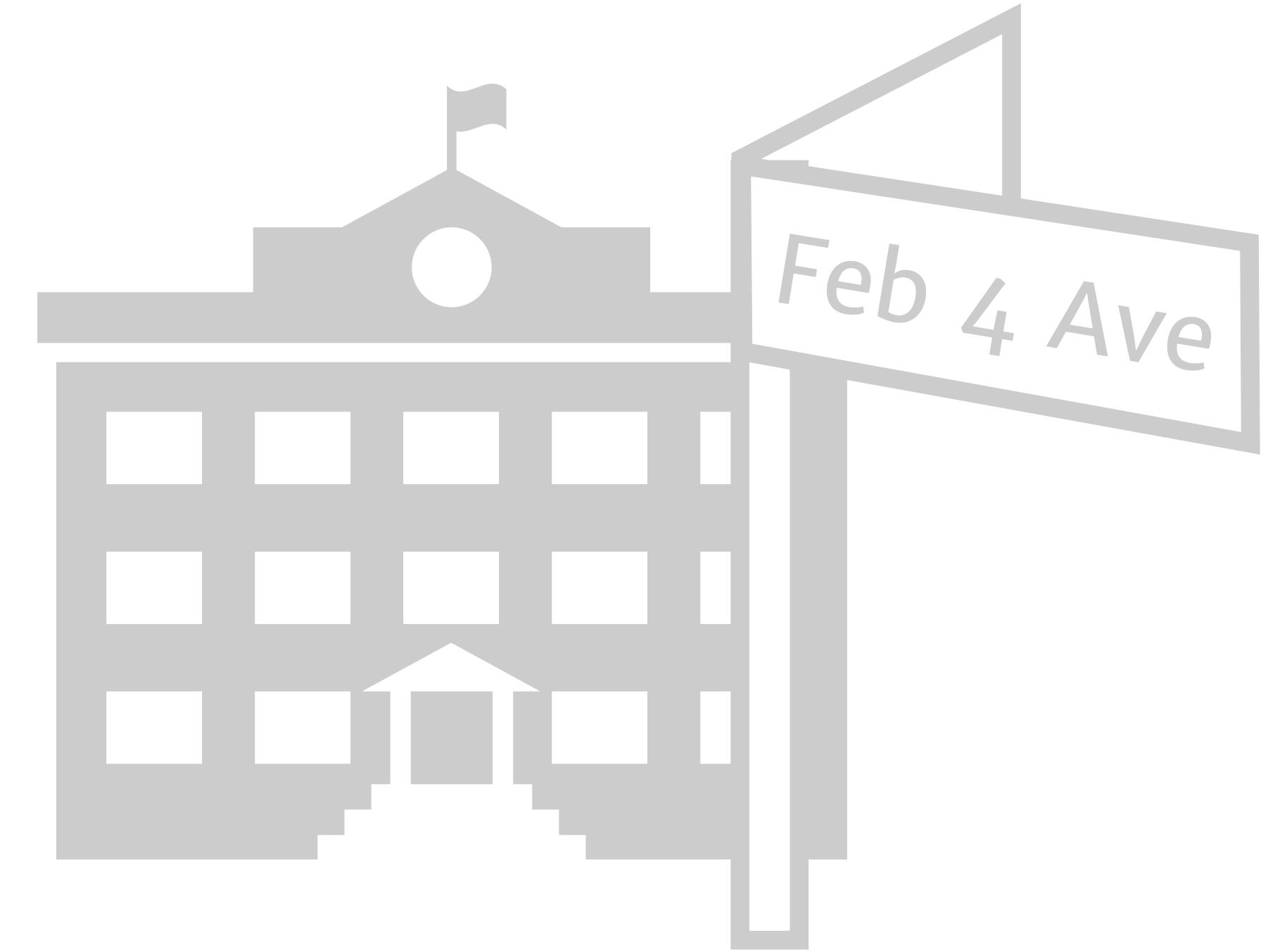
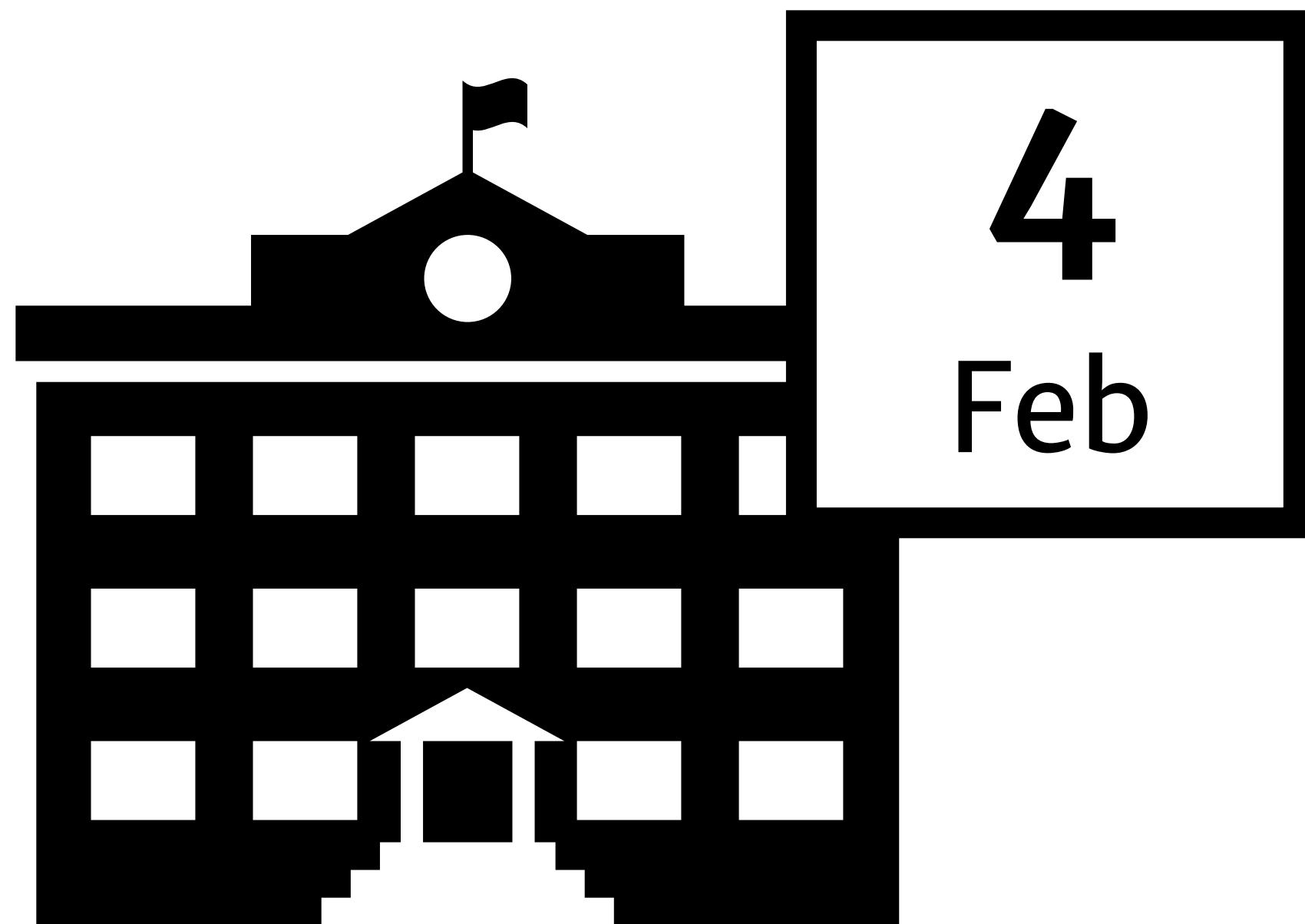


Probabilistic modeling



I went to the restaurant on February 4th.

Probabilistic modeling



I went to the restaurant on February 4th.

Probabilistic modeling

$p(\text{a quick brown fox} \mid \text{jumps over the lazy dog})$

$p(\text{jumps over the lazy dog} \mid \text{a quick brown fox})$

$p(\text{a quick brown fox})$

$p(\text{jumps over the lazy dog})$

We need to predict which interpretations
are **allowed**, and which are **most likely**.

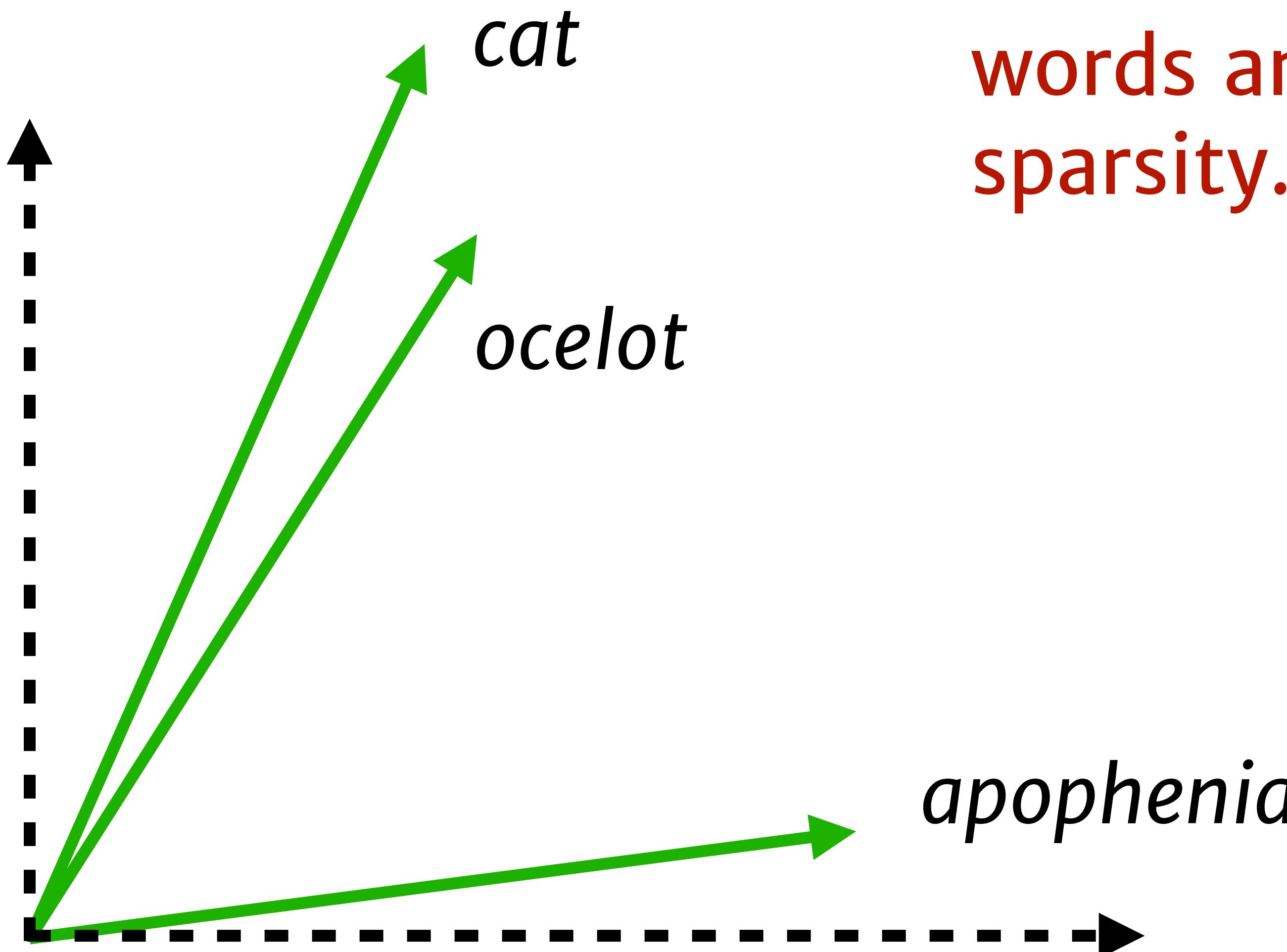
Machine learning

$$p_{\theta}(\text{a quick brown fox}) \propto \exp\{\theta^T(f(\text{a}) + f(\text{quick}) + \dots)\}$$

$$\theta^* = \operatorname{argmin}_{\theta} \sum_{\text{sentence}} -\log(p_{\theta}(\text{sentence}))$$

We need to estimate these probability distributions from corpus data.

Representation learning



We need to share information across words and tasks to handle with data sparsity.

Linguistics

I went to the restaurant on February 4th.

on-1 ?
on-5

We need to constrain the space of **interesting prediction problems** (and relevant features ?) and **form hypotheses about model behavior**.

Admin

Admin

(please read the syllabus)

Will lectures be recorded?

Lectures will be recorded!

Prereq: probability

$p(\text{a quick brown fox} \mid \text{jumps over the lazy dog})$

$p(\text{jumps over the lazy dog} \mid \text{a quick brown fox})$

$p(\text{a quick brown fox})$

$p(\text{jumps over the lazy dog})$

Prereq: intro ML

$$p_{\theta}(\text{a quick brown fox}) \propto \exp\{\theta^T(f(\text{a}) + f(\text{quick}) + \dots)\}$$

$$\theta^* = \operatorname{argmin}_{\theta} \sum_{\text{sentence}} -\log(p_{\theta}(\text{sentence}))$$

Prereq: algorithms & discrete math

a

quick

brown

fox w_i

a

bright

purple w_{i-1}

fox

the

fluorescent

indigo

badger

$$f(w_i) = \sum_{w_{i-1}} f(w_{i-1}) \cdot g(w_{i-1}, w_i)$$

w_{i-1}

Do I need 6.036???

You should be comfortable **implementing, debugging, and reasoning about ML models trained via gradient descent.**

Do I need 6.046???

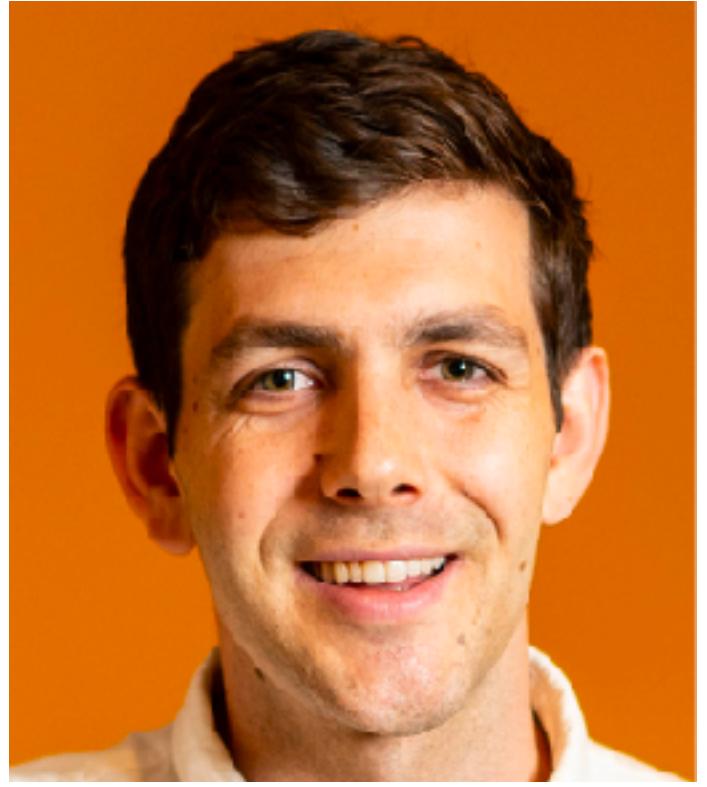
You should be comfortable **implementing, debugging, and reasoning about dynamic programming algorithms** that are more complicated than anything you saw in 6.006.

Getting caught up

Released today: “HWo” (ungraded)

Exercises in probability, machine learning, and algorithms.
Discussed in recitations this Friday!

Course staff



Jacob Andreas



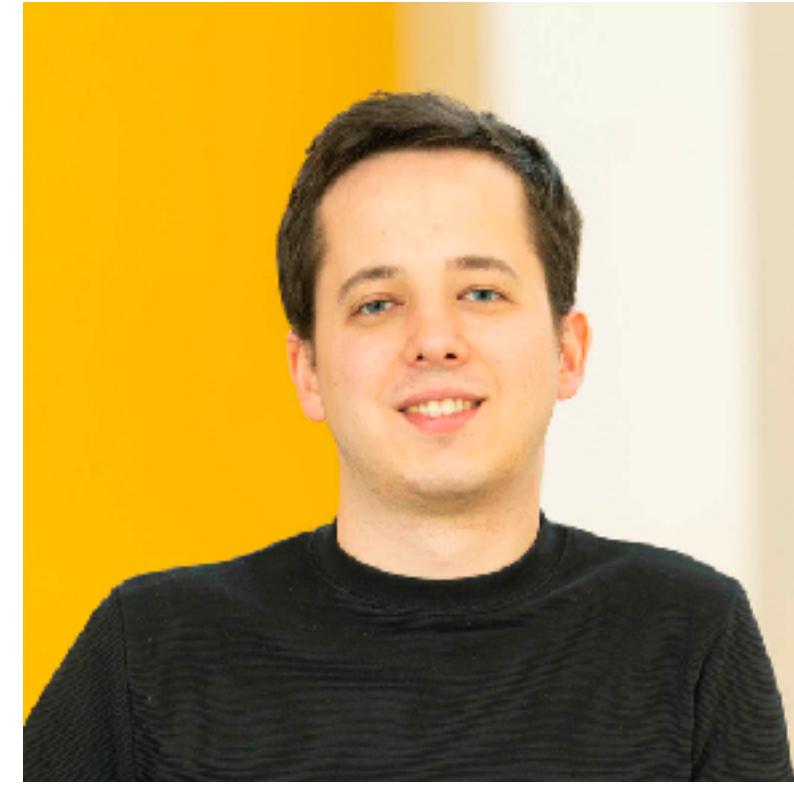
Jim Glass



Abby Bertics



Dylan Doblar



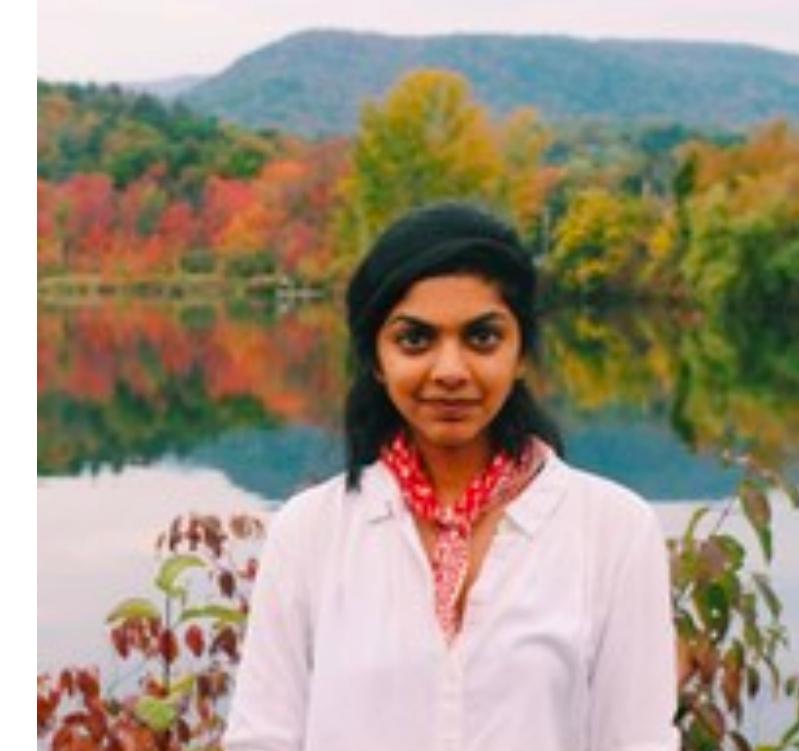
Ekin Akyürek



Evan Hernandez



Hongyin Luo



Harini Suresh



Pranav Krishna



Wei Fang

Asking questions

Type your question in Zoom chat; one of the TAs will insert a question break at an appropriate moment and **ask you to say it out loud.**

Course outline

Feb 16 - Mar 18: modeling sequences

Mar 25 - Apr 15: syntax, semantics, and speech

Apr 22 - May 13: special topics

May 18 - May 20: project presentations

Structure of the course

- Four homework assignments
(6.864: extra questions)
- Final group project
(6.864: extra pages)

Homework assignments

- paper, coding, and *peer grading*
- coding: we'll provide **pytorch** notebooks in Google colab, but you're free to submit whatever you want

Homework assignments

Collaboration policy:

We encourage you to work together,
but final writeup and code must be
your own!

Final projects

- Implement a {previously published, new} model for a {standard benchmark, new task}
- Groups of 2–4 people
- 3 submissions: proposal, update, final report
- Dates & details on syllabus

Late work policy

- Due dates will be posted on Canvas
- Due at noon
- 10% off for every day late
- Late final projects will not be accepted!

Recitations

2x on Fridays, times on syllabus

Recorded!

No need to pre-register, and
attendance is not required

Office Hours

**Monday / Wednesday / Friday
(times & signup forms on syllabus)**

Not recorded.

Course website

<https://canvas.mit.edu/courses/7503>

Detailed syllabus, assignments, slides, recordings.

Piazza

<https://piazza.com/class/kjytwuixga4me>

Discussions for homework, class content.

Getting help

This is a challenging class (and a challenging semester)!

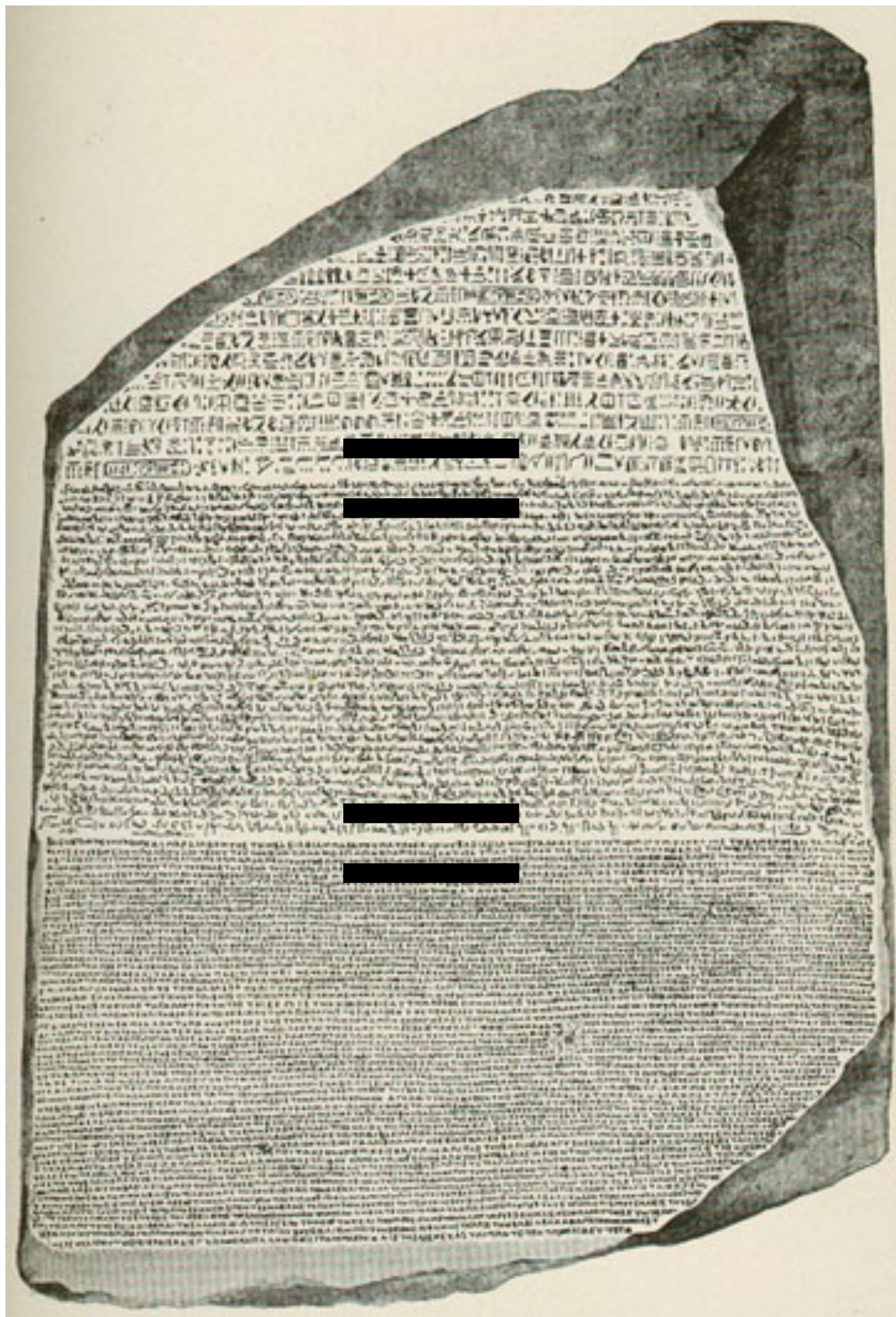
Don't be shy about posting on Piazza, coming to OH,

Talk to S³ if you need specific accommodations
(HW extensions, ...)

If you just feel like you're struggling, email me or Jim
(addresses on syllabus).

Preview

Unsupervised translation



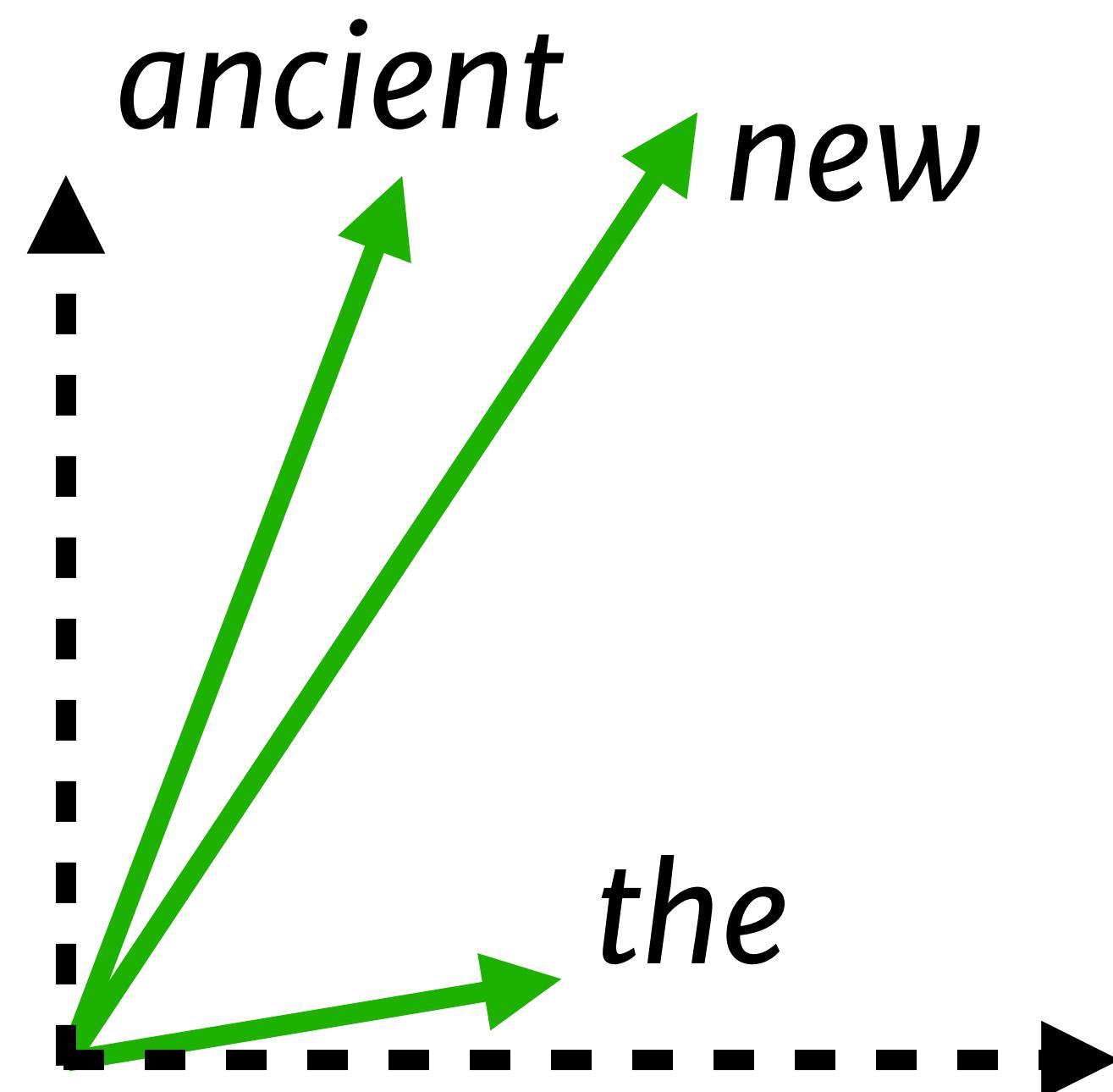
Two households, both alike in dignity, In fair Verona,
where we lay our scene, From ancient grudge break to
new mutiny, Where civil blood makes civil hands
unclean. From forth the fatal loins of these two foes A
pair of star-cross'd lovers take their life; whose misadv

≠

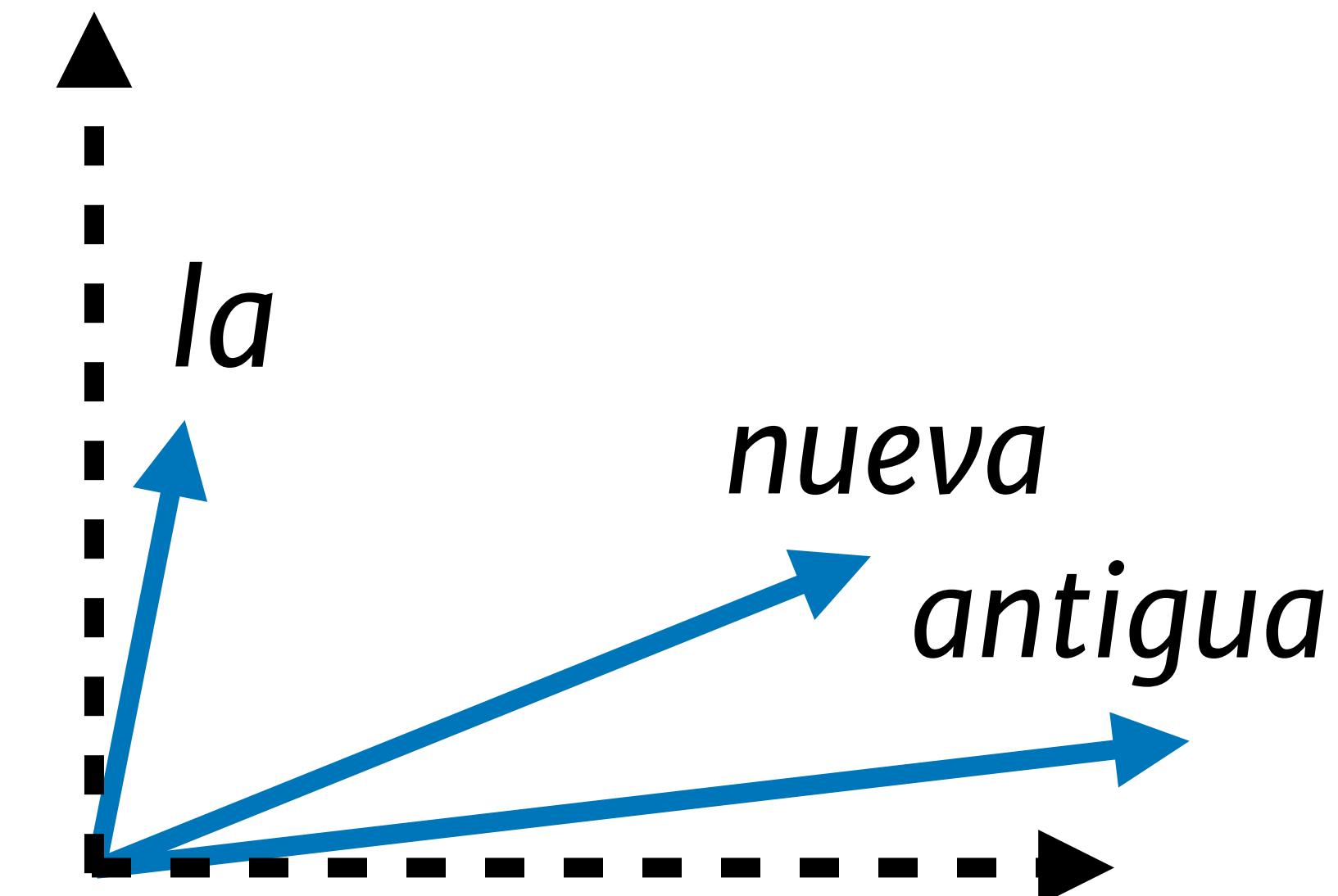
Desocupado lector: sin juramento me podrás creer que quisiera que este libro, como hijo del entendimiento, fuera el más hermoso, el más gallardo y más discreto que pudiera imaginarse. Pero no he podido yo contravenir al orden de naturaleza, que en ella cada cos

Learning word representations

Two households, both alike in dignity, In fair Verona,
where we lay our scene, From ancient grudge break to
new mutiny, Where civil blood makes civil hands
unclean. From forth the fatal loins of these two foes A
pair of star-cross'd lovers take their life; whose misadv



Desocupado lector: sin jura
quisiera que este libro, co
fuerá el más hermoso, el n
que pudiera imaginarse.
contravenir al orden de natu



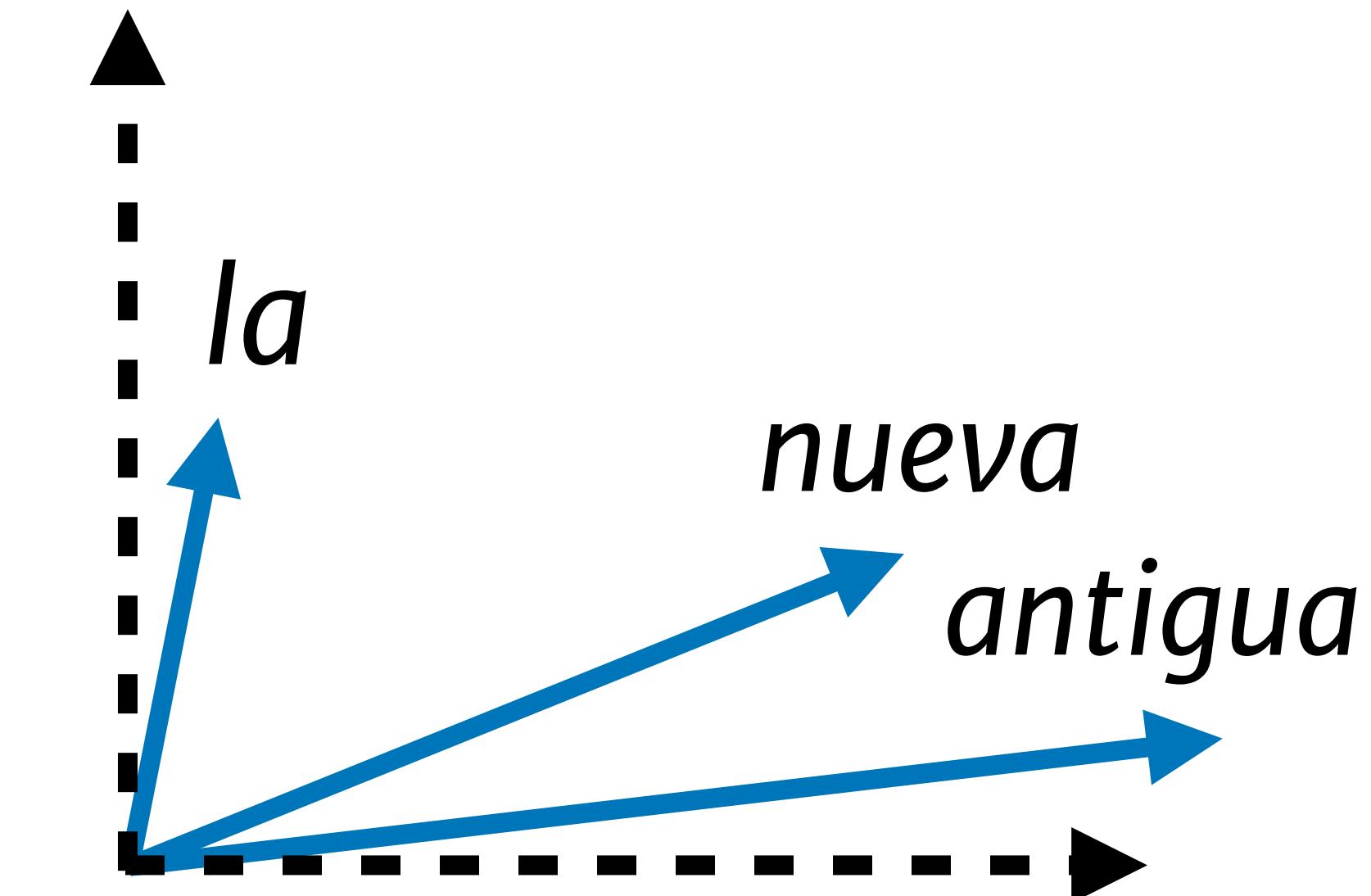
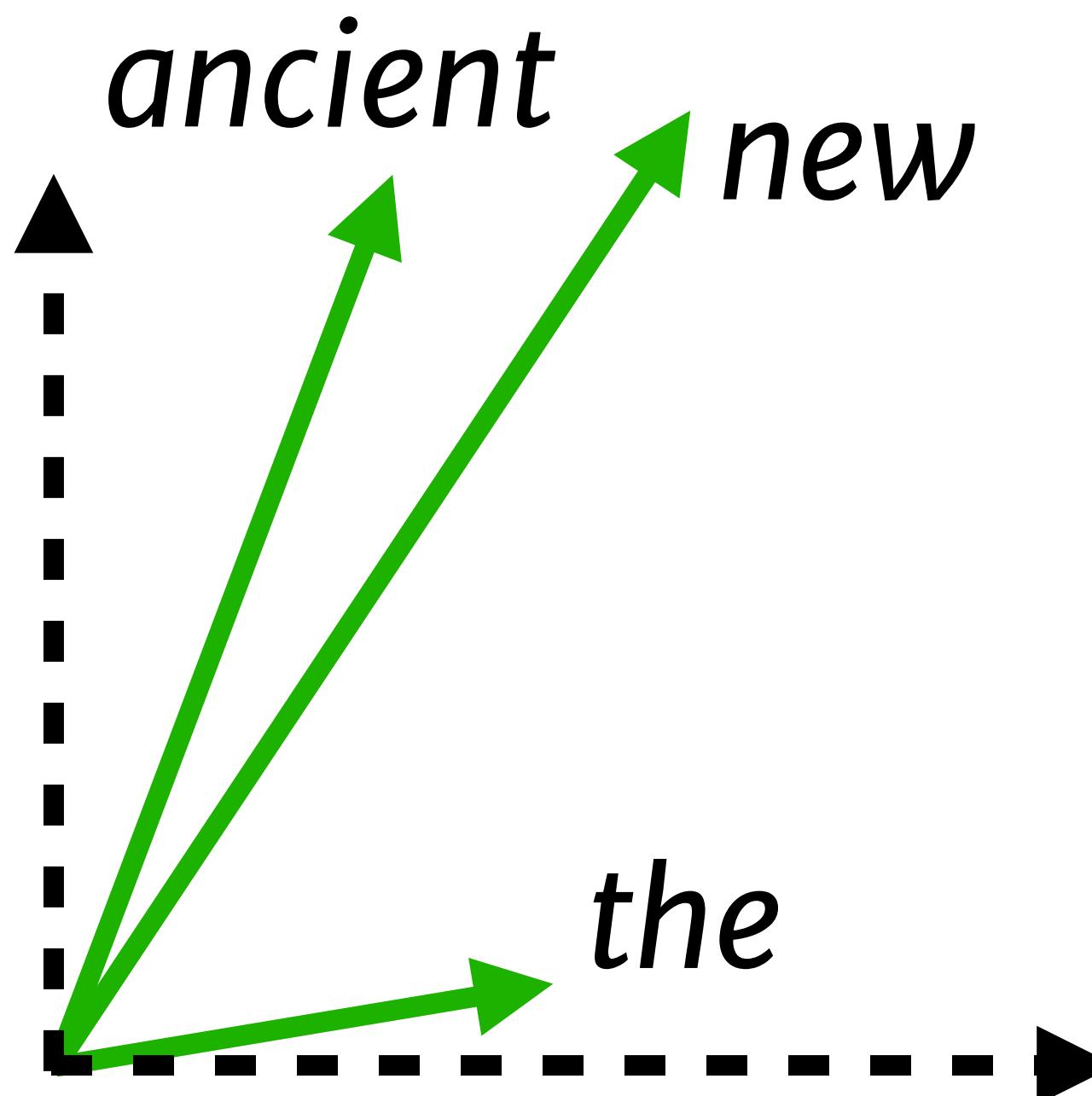
Learning word representations

1o-j74-kic1-**5hec**-**r9xv**-**j7hek67**-**r9yx**-j74-0dc2-74b3-0-i4h84i-e5-

m4bb-0jj4d343-0ddk0b-2ed54h4d24i-ik1i4gk4djbo-0-i4h84i-e5-d0j8ed0b-kic0-

j74-34f0hjc4dj-e5-2ecc4h24-0d3-j74-d0j8ed0b-8dij8jkj4-e5-ij0d30h3i-0d3-

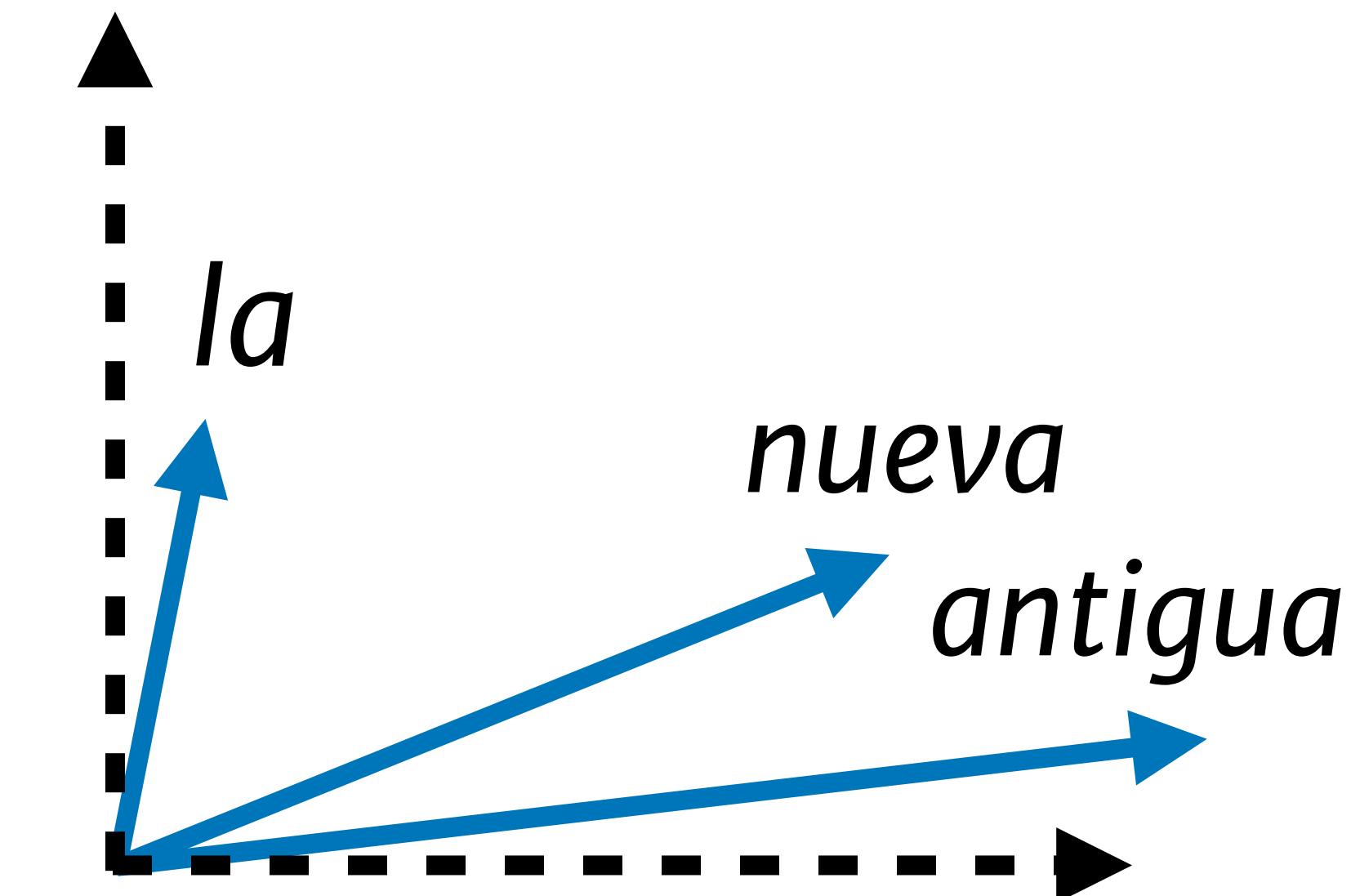
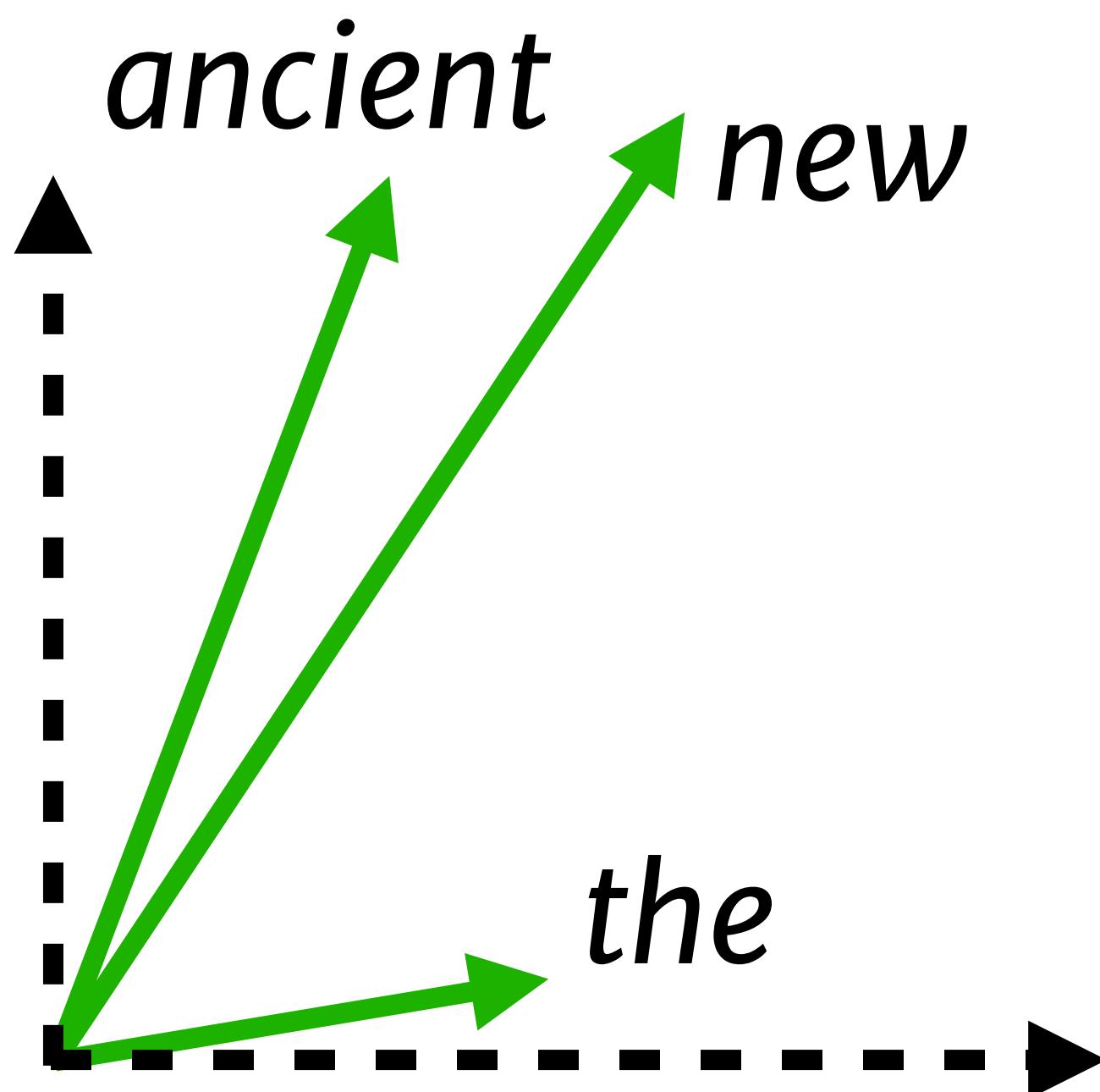
j427debe6o-d8ij-m4h4-74b3-**5hec**-**r9yx**-**j7hek67**-**r99t**



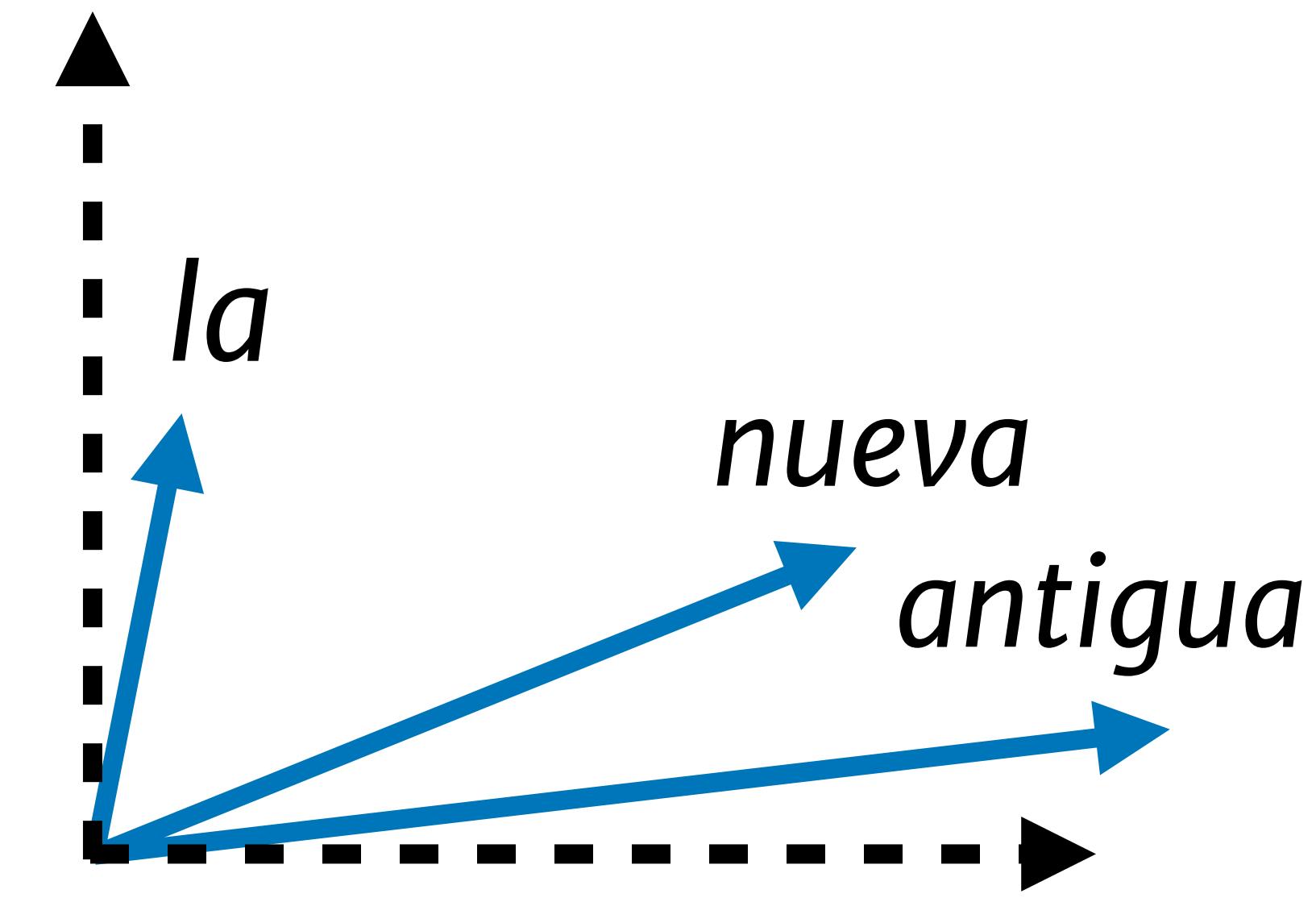
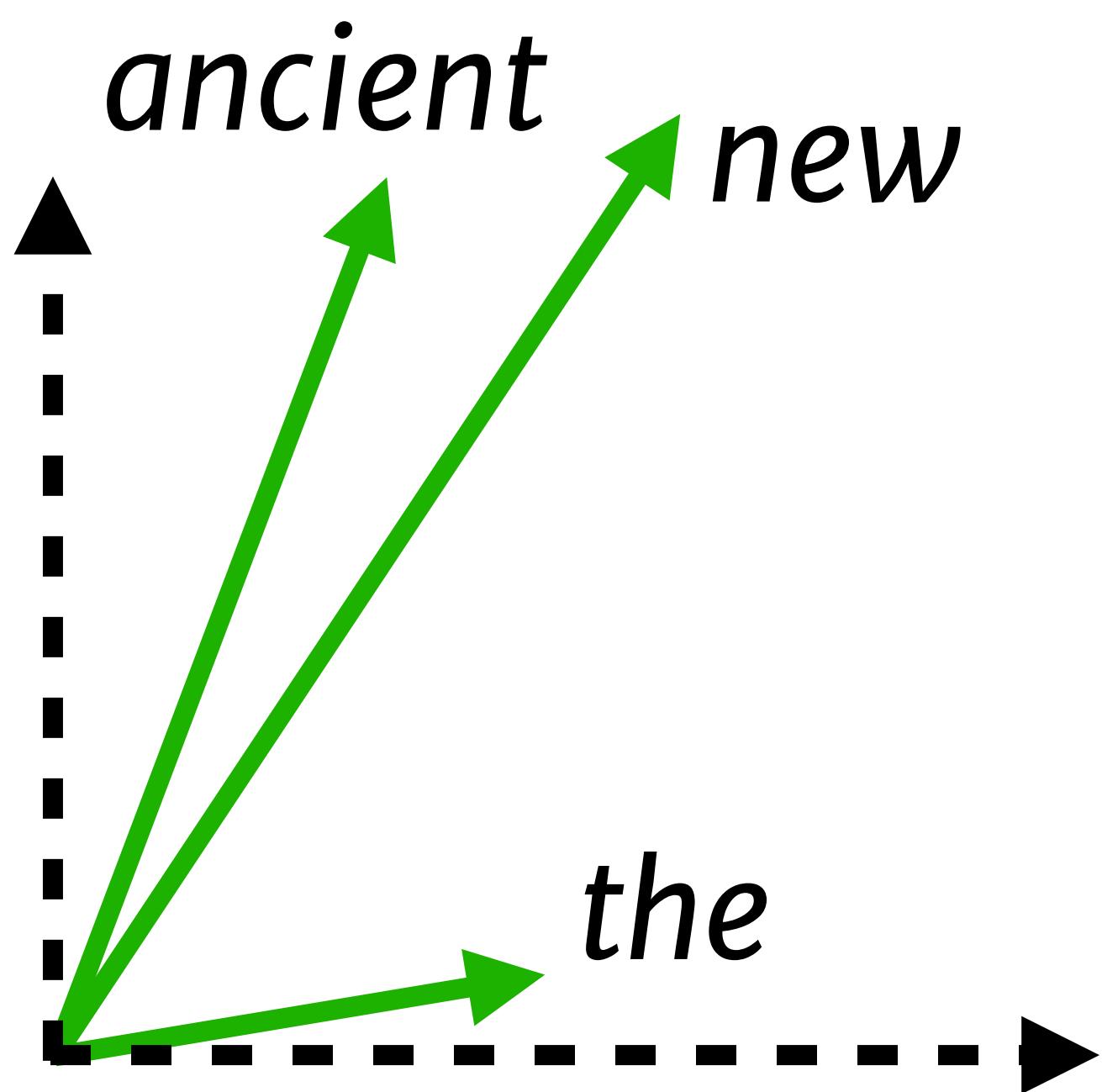
Learning word representations

Two households, both alike in dignity, In fair Verona,
where we lay our scene, From ancient grudge break to
new mutiny, Where civil blood makes civil hands
unclean. From forth the fatal loins of these **two foes** A
pair of star-cross'd lovers take their life; whose misadv

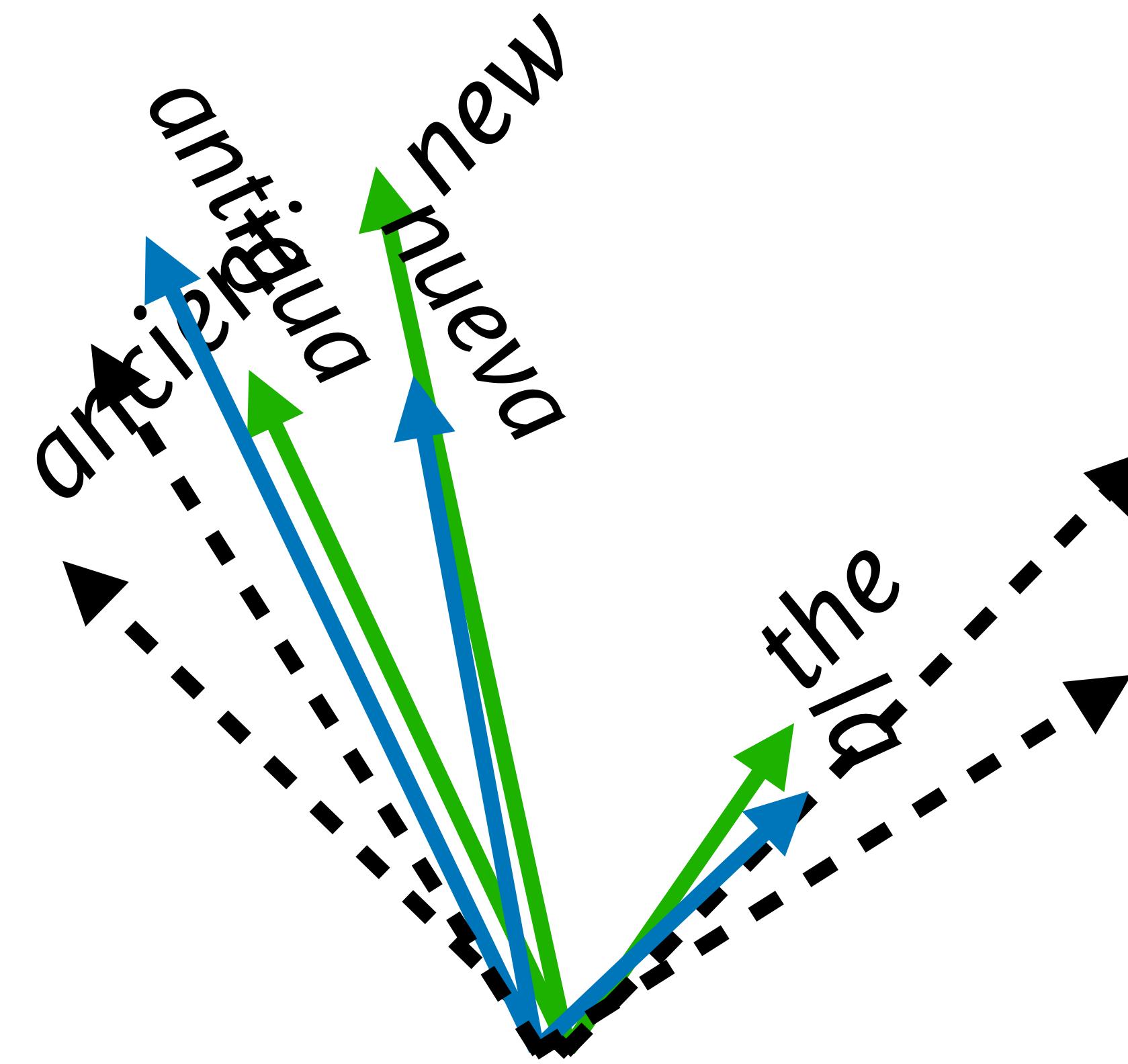
Desocupado lector: sin jura
quisiera que este libro, coi
fuerá el más hermoso, el n
que pudiera imaginarse.
contravenir al orden de natu



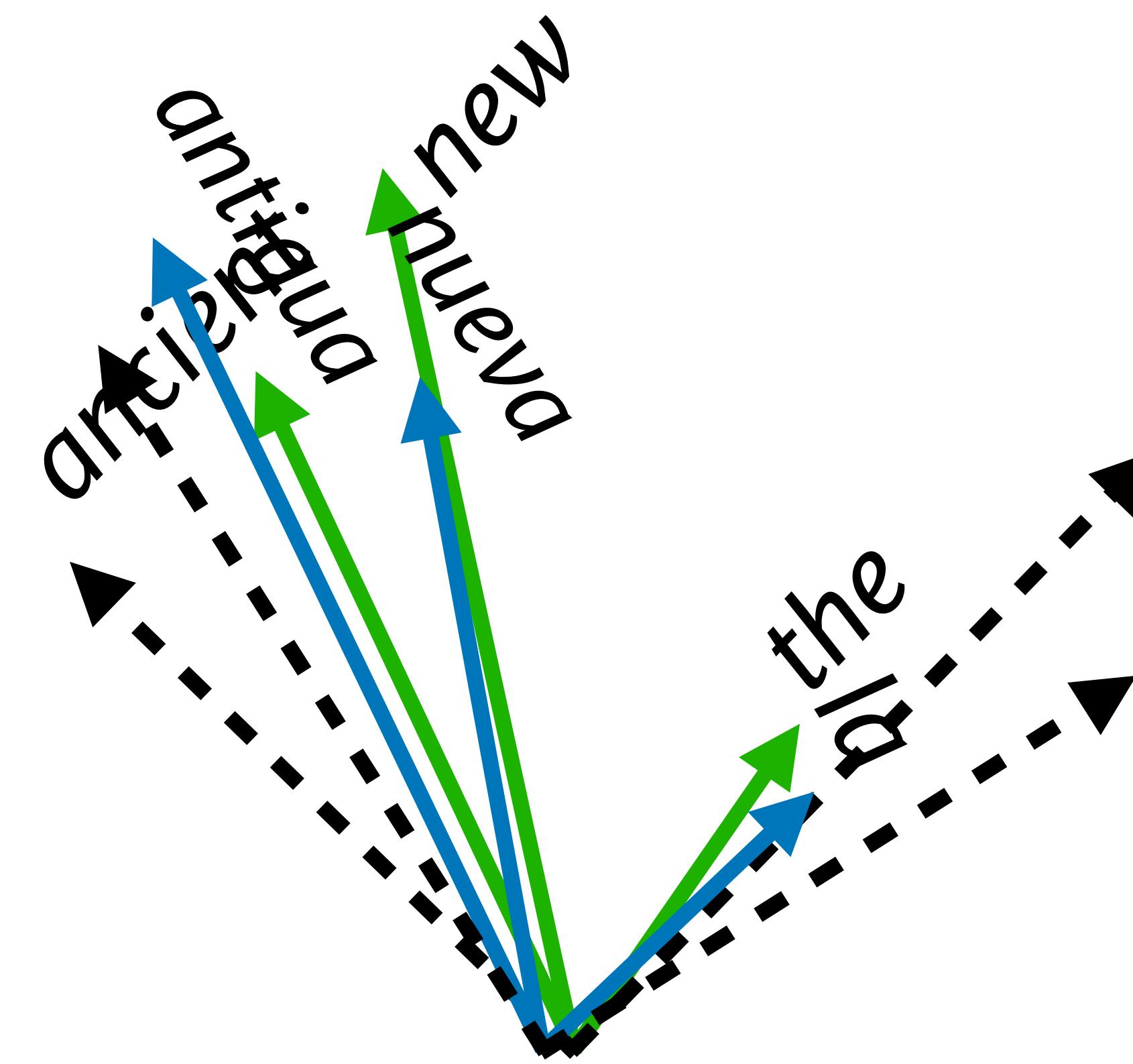
Aligning representations across languages



Aligning representations across languages



Aligning representations across languages



Aligning representations across languages

new
nueva

antigua

ancient

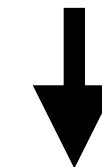
the
la

Translating words

Desocupado lector: sin juramento me podrás creer que
idle reader without oath me able believe that

Denoising

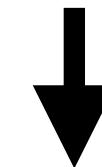
Two households, both alike in dignity, In fair Verona,
where we lay our scene, From ancient grudge break to
new mutiny, Where civil blood makes civil hands uncl



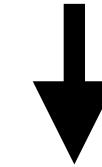
households **Two**, both alike **dignity**, **fair** In Verona,
where **lay** our scene, From ancient grudge **vacation** to
new mutiny, Where **blood** civil **five** makes civil hands

Denoising

Two households, both alike in dignity, In fair Verona,
where we lay our scene, From ancient grudge break to
new mutiny, Where civil blood makes civil hands uncl



households **Two**, both alike **dignity**, **fair** In Verona,
where **lay** our scene, From ancient grudge **vacation** to
new mutiny, Where **blood** civil **five** makes civil hands



Two households, both alike in dignity, In fair Verona,
where we lay our scene, From ancient grudge break to
new mutiny, Where civil blood makes civil hands uncl

$p(\text{original sentence} \mid \text{corrupted sentence})$

Translating sentences

Desocupado lector: sin juramento me podrás creer que

idle reader: without oath me able believe that

idle reader, doubtless you can believe me that

Language to code



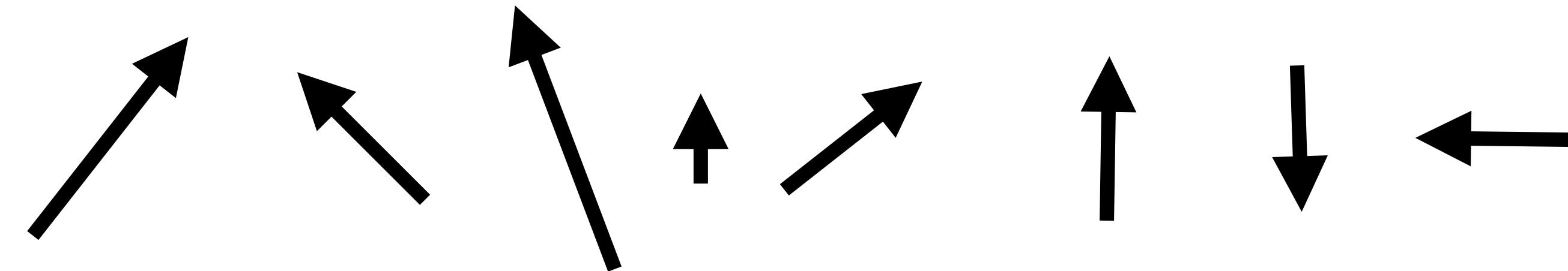
Language to code



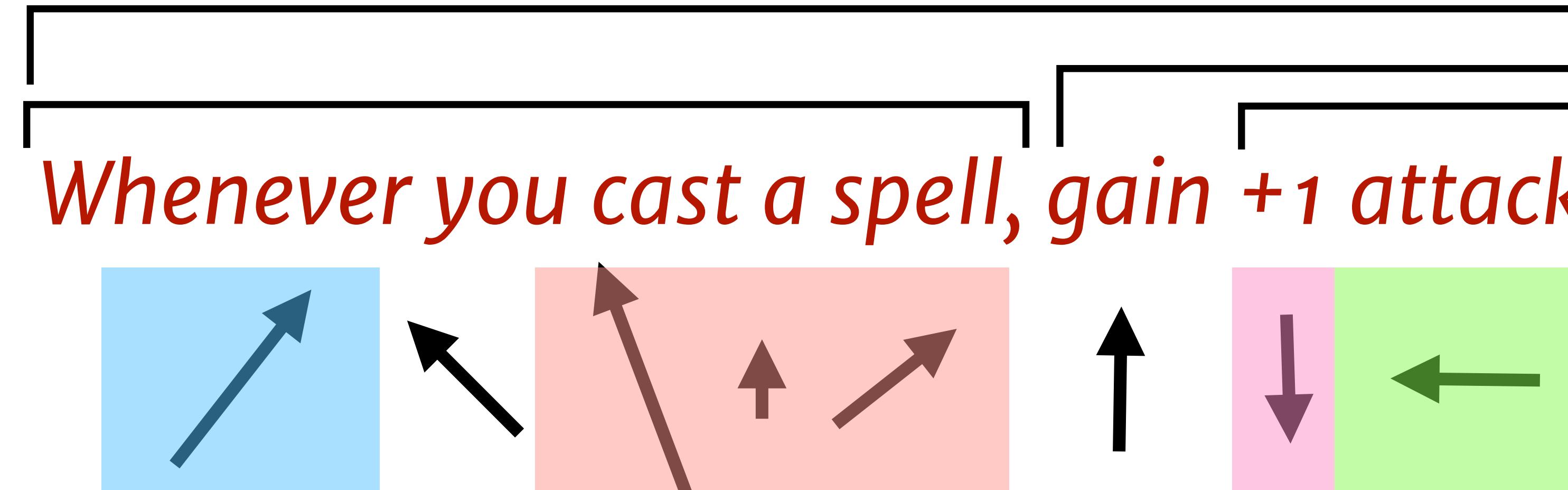
```
return Minion (  
    1, 3, effects=[  
        Effect(  
            SpellCast(),  
            ActionTag(  
                Give(ChangeAttack(1)),  
                SelfSelector())  
        )  
    ] )
```

Learning sentence representations

Whenever you cast a spell, gain +1 attack



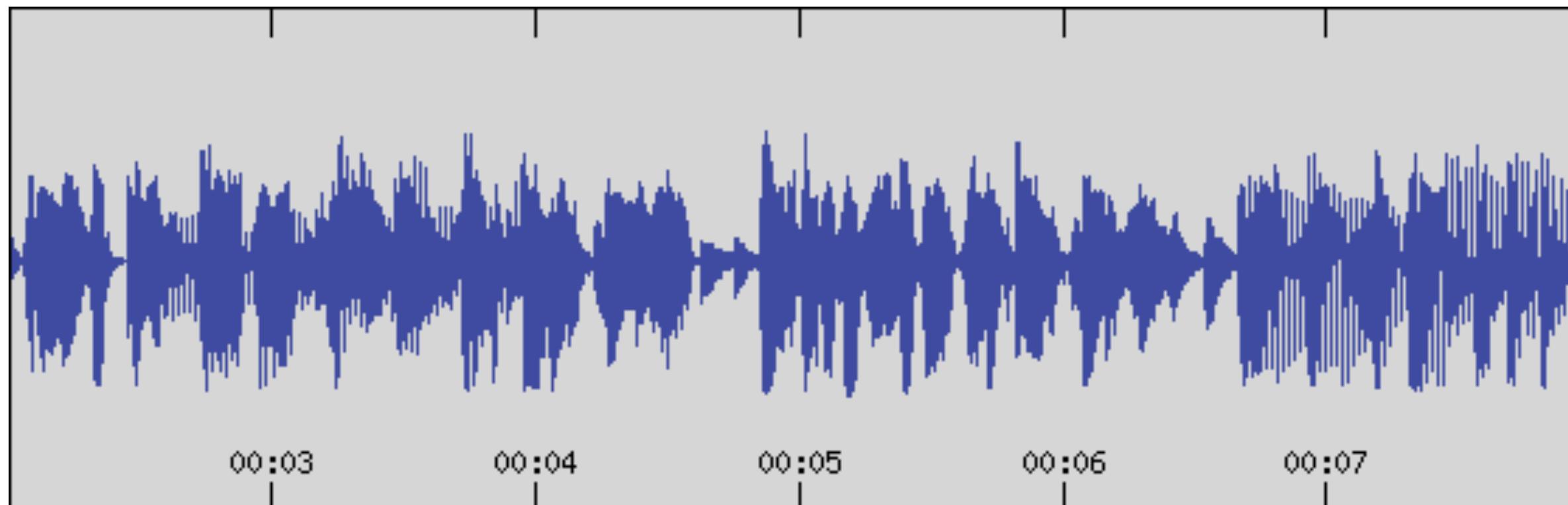
Predicting structured outputs



```
return Minion(  
    1, 3, effects=[  
        Effect(  
            SpellCast(),  
            ActionTag(  
                Give(ChangeAttack( ...  
            )
```

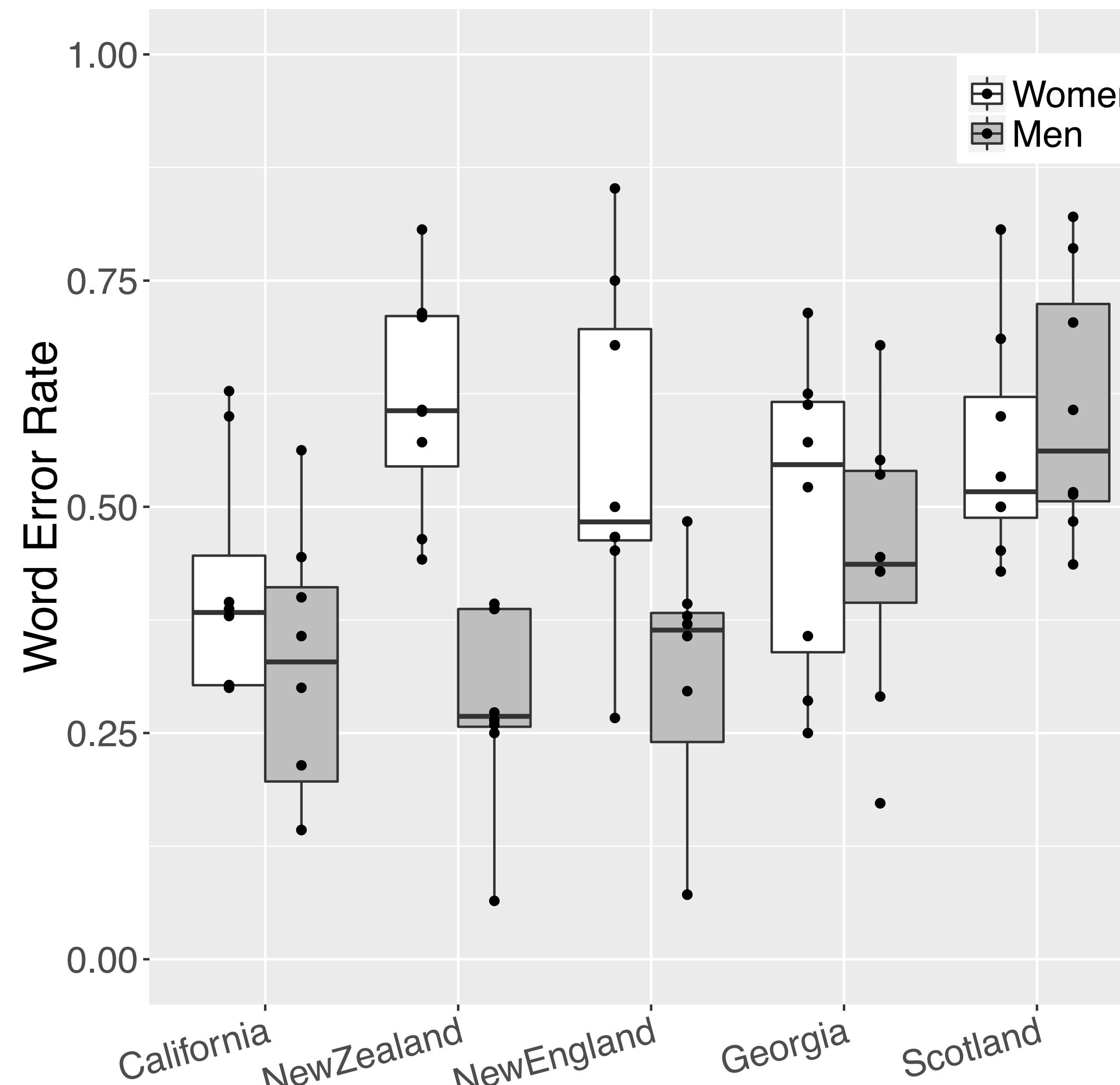
Disparate model accuracy

Speech recognition:



*It's hard to wreck a
nice beach.*

Disparate model accuracy



Building fair datasets & models

Modeling

Inductive bias favoring particular groups

Genuine difficulty of underlying prediction problem

Data collection

Design decisions from researchers

Social expectations of annotators

This semester:

Machine learning approaches to **interpreting**, **generating** and **analyzing** human languages.

Next class: text classification