CMPU 366 · Natural Language Processing

Vector Semantics

29 September 2025

VASSAR COLLEGE

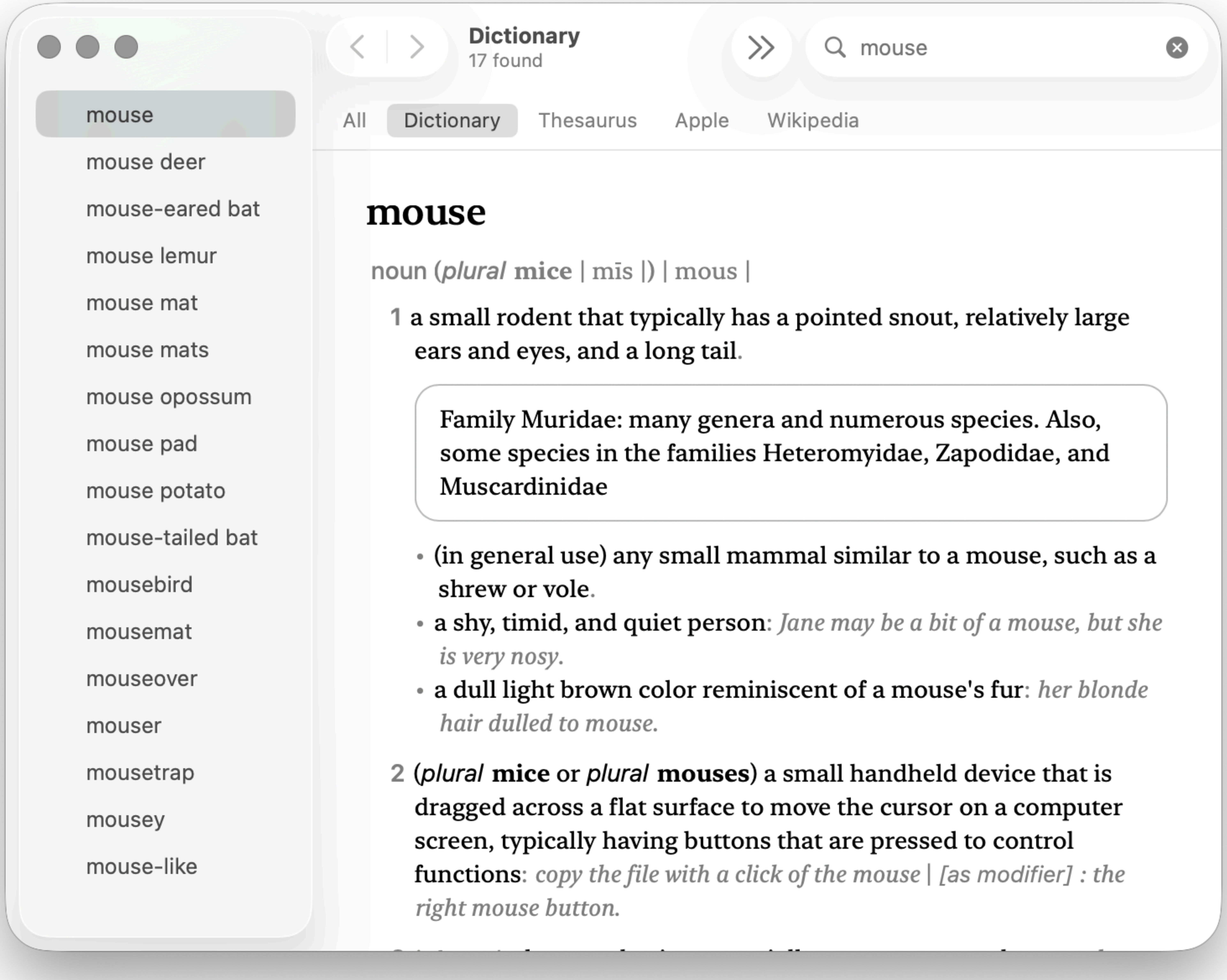
A·D·1861

In the *n*-gram or text classification methods we've seen so far, words are just strings – that's not very satisfactory!

What do words mean?

What do we want from a theory of word meaning?

We can get some ideas from *lexical semantics* – the linguistic study of word meaning.



mouse

noun (*plural* **mice** | mīs |) | mous |

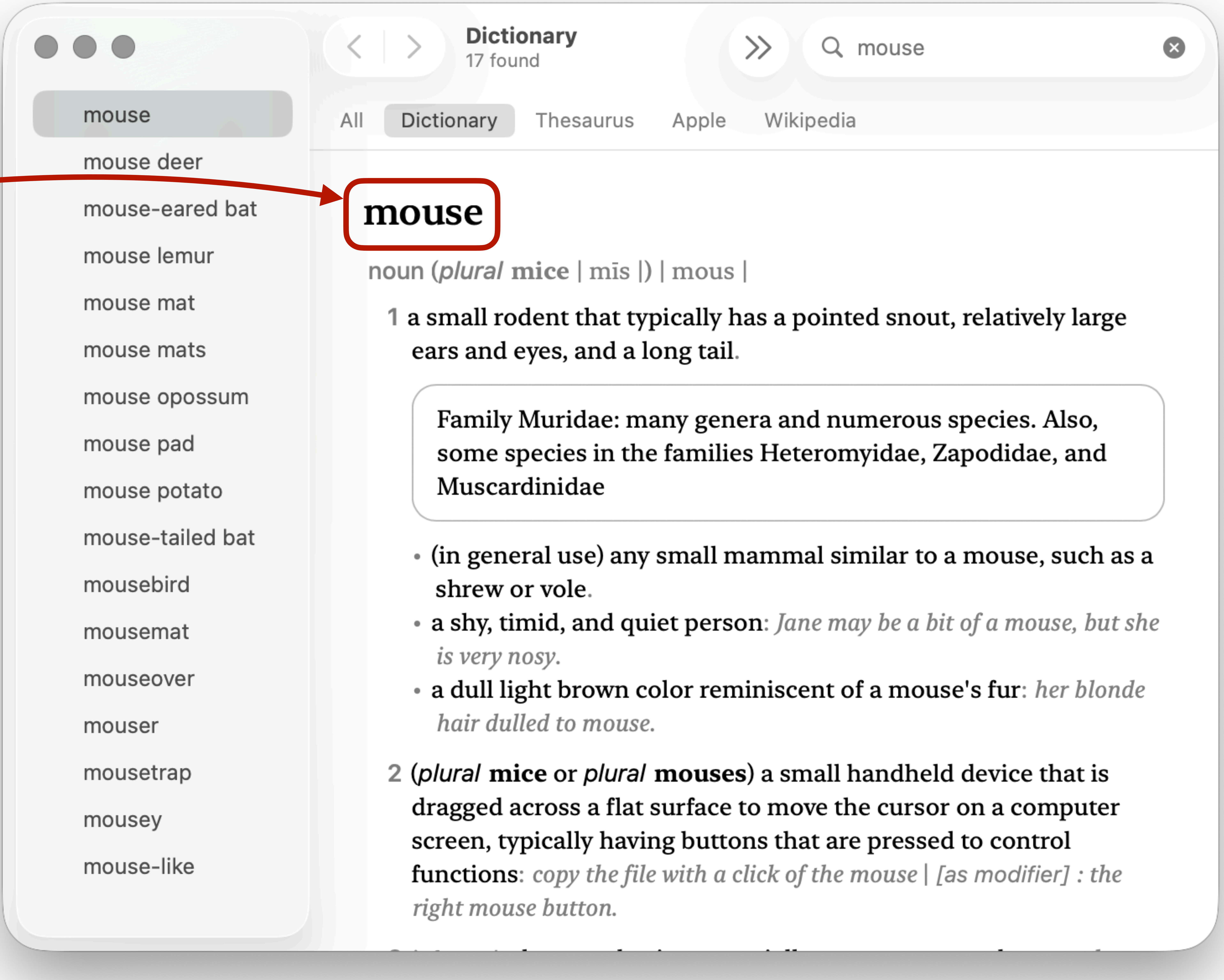
1 a small rodent that typically has a pointed snout, relatively large ears and eyes, and a long tail.

Family Muridae: many genera and numerous species. Also, some species in the families Heteromyidae, Zapodidae, and Muscardinidae

- (in general use) any small mammal similar to a mouse, such as a shrew or vole.
- a shy, timid, and quiet person: *Jane may be a bit of a mouse, but she is very nosy.*
- a dull light brown color reminiscent of a mouse's fur: *her blonde hair dulled to mouse.*

2 (*plural* **mice** or *plural* **mouses**) a small handheld device that is dragged across a flat surface to move the cursor on a computer screen, typically having buttons that are pressed to control functions: *copy the file with a click of the mouse | [as modifier] : the right mouse button.*

Lemma



mouse

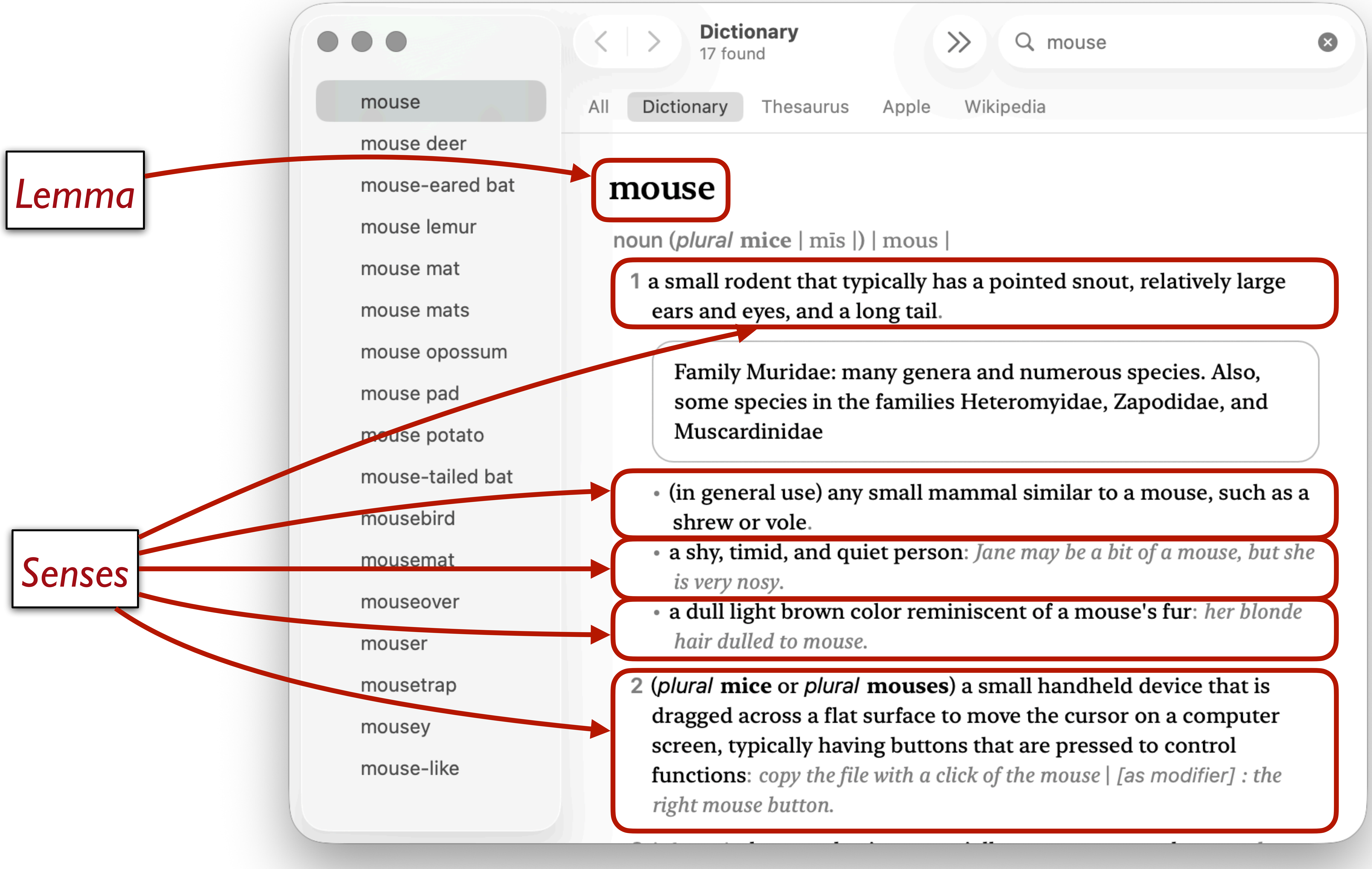
noun (*plural* **mice** | mīs |) | mous |

1 a small rodent that typically has a pointed snout, relatively large ears and eyes, and a long tail.

Family Muridae: many genera and numerous species. Also, some species in the families Heteromyidae, Zapodidae, and Muscardinidae

- (in general use) any small mammal similar to a mouse, such as a shrew or vole.
- a shy, timid, and quiet person: *Jane may be a bit of a mouse, but she is very nosy.*
- a dull light brown color reminiscent of a mouse's fur: *her blonde hair dulled to mouse.*

2 (*plural* **mice** or *plural* **mouses**) a small handheld device that is dragged across a flat surface to move the cursor on a computer screen, typically having buttons that are pressed to control functions: *copy the file with a click of the mouse* | *[as modifier] : the right mouse button.*



Lemma

Senses

mouse

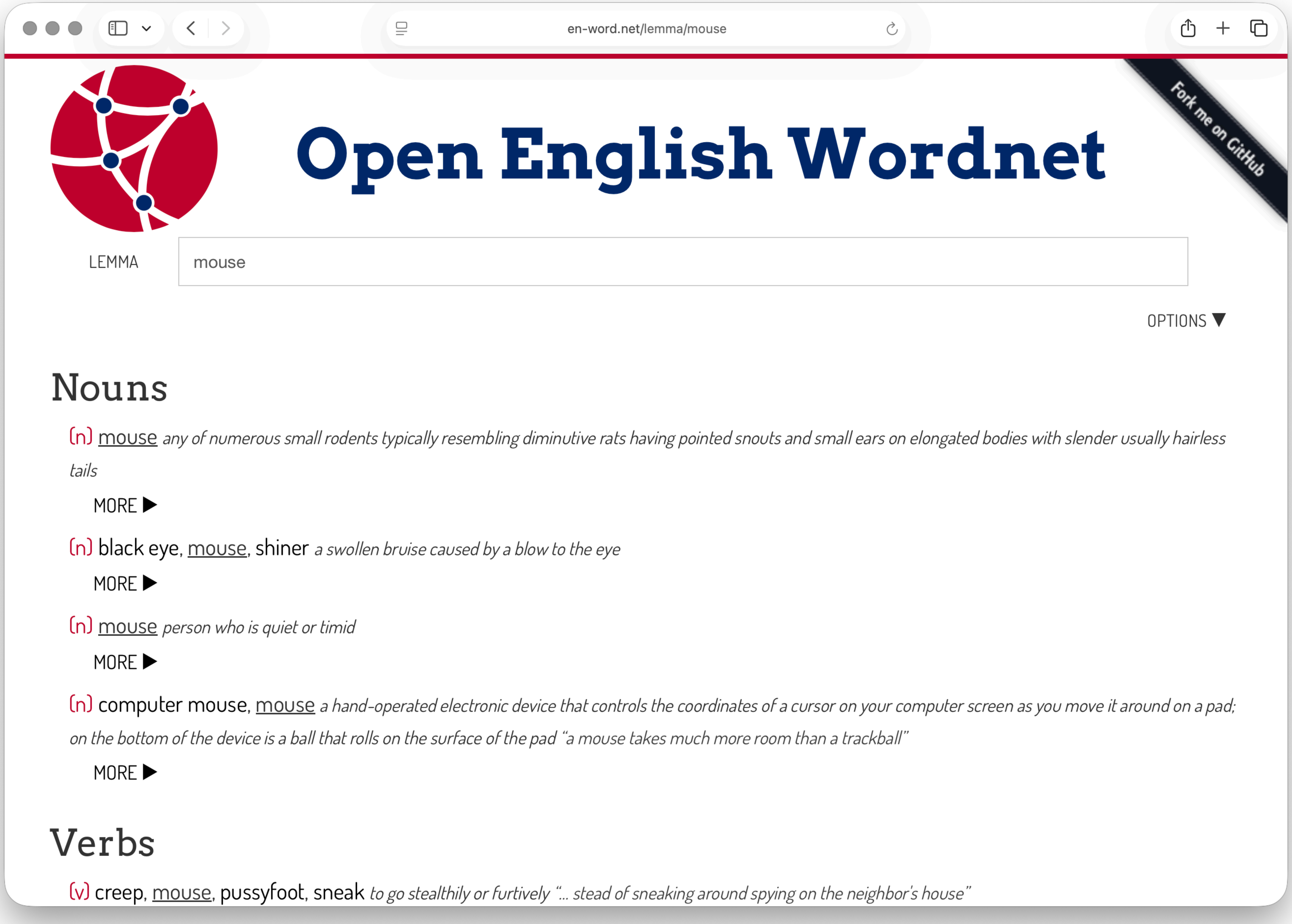
noun (*plural* **mice** | mīs |) | mous |

1 a small rodent that typically has a pointed snout, relatively large ears and eyes, and a long tail.

Family Muridae: many genera and numerous species. Also, some species in the families Heteromyidae, Zapodidae, and Muscardinidae

- (in general use) any small mammal similar to a mouse, such as a shrew or vole.
- a shy, timid, and quiet person: *Jane may be a bit of a mouse, but she is very nosy.*
- a dull light brown color reminiscent of a mouse's fur: *her blonde hair dulled to mouse.*

2 (*plural* **mice** or *plural* **mouses**) a small handheld device that is dragged across a flat surface to move the cursor on a computer screen, typically having buttons that are pressed to control functions: *copy the file with a click of the mouse* | *[as modifier] : the right mouse button.*



Open English Wordnet

LEMMA mouse

OPTIONS ▼

Nouns

(n) mouse *any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails*

MORE ▶

(n) black eye, mouse, shiner *a swollen bruise caused by a blow to the eye*

MORE ▶

(n) mouse *person who is quiet or timid*

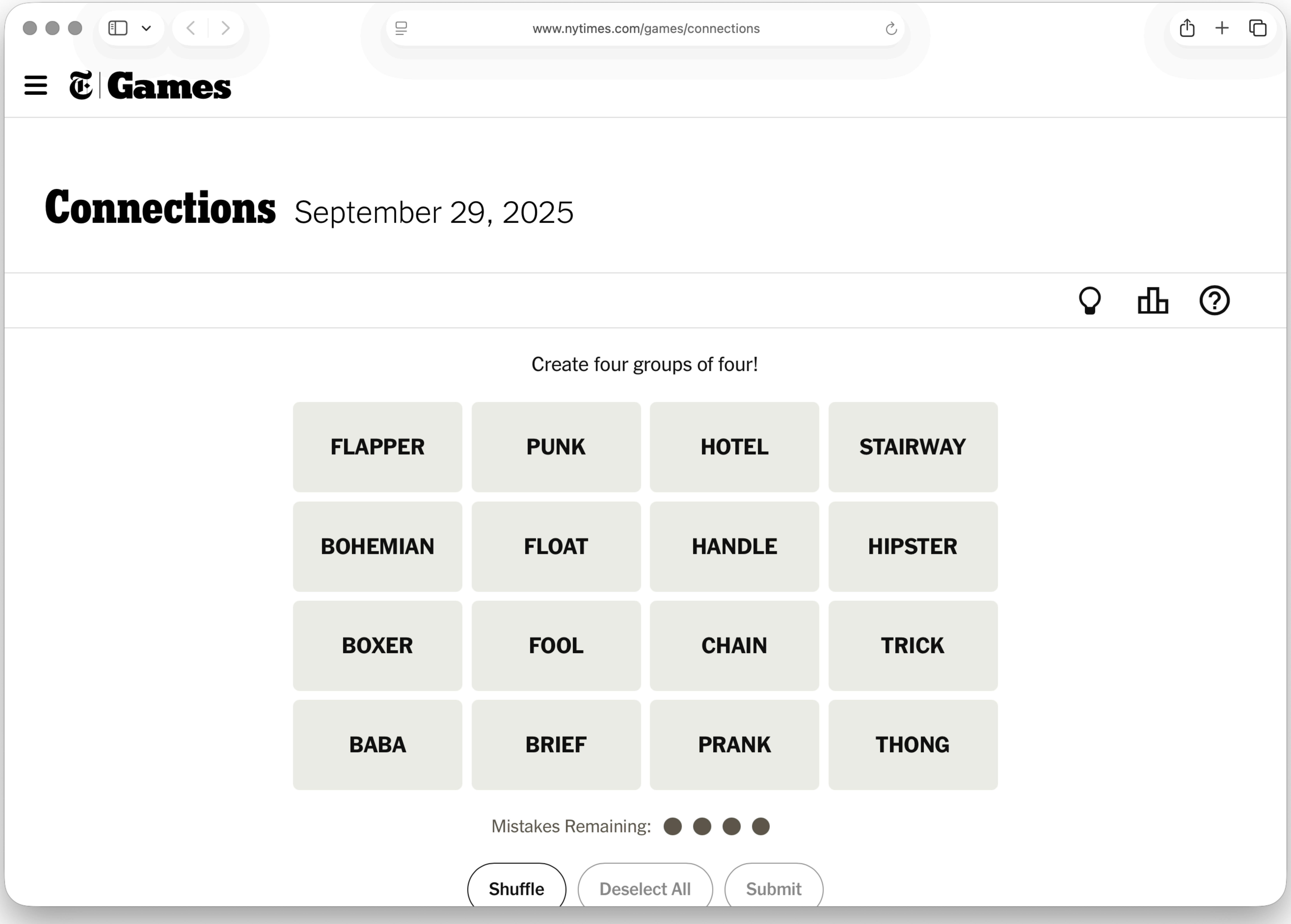
MORE ▶

(n) computer mouse, mouse *a hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad; on the bottom of the device is a ball that rolls on the surface of the pad "a mouse takes much more room than a trackball"*

MORE ▶

Verbs

(v) creep, mouse, pussyfoot, sneak *to go stealthily or furtively "... stead of sneaking around spying on the neighbor's house"*



Connections September 29, 2025

Create four groups of four!

FLAPPER	PUNK	HOTEL	STAIRWAY
BOHEMIAN	FLOAT	HANDLE	HIPSTER
BOXER	FOOL	CHAIN	TRICK
BABA	BRIEF	PRANK	THONG

Mistakes Remaining: ● ● ● ●

Relations between word senses

big / *large*

oculist / *eye-doctor*

car / *automobile*

water / H_2O

draft / *draught*

Relations between word senses

Synonymy

big / *large*

oculist / *eye-doctor*

car / *automobile*

water / *H_2O*

draft / *draught*

Relations between word senses

Synonymy

big / *large*

oculist / *eye-doctor*

car / *automobile*

water / H_2O

draft / *draught*

my big sister ≠ *my large sister*

Relations between word senses

Synonymy

big / large

oculist / eye-doctor

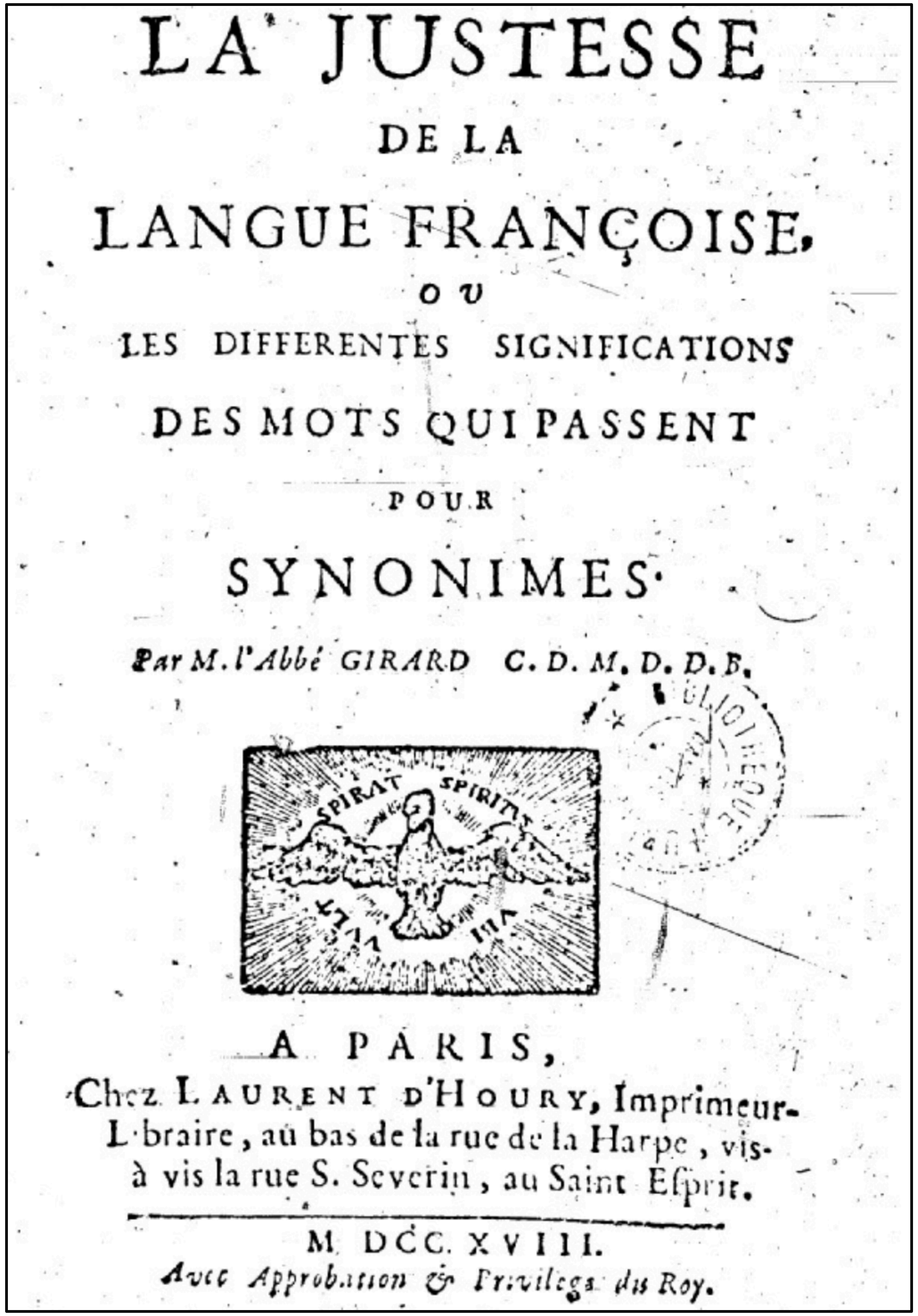
car / automobile

water / H_2O

draft / draught

my big sister ≠ my large sister

Would you use H_2O in a surfing guide?



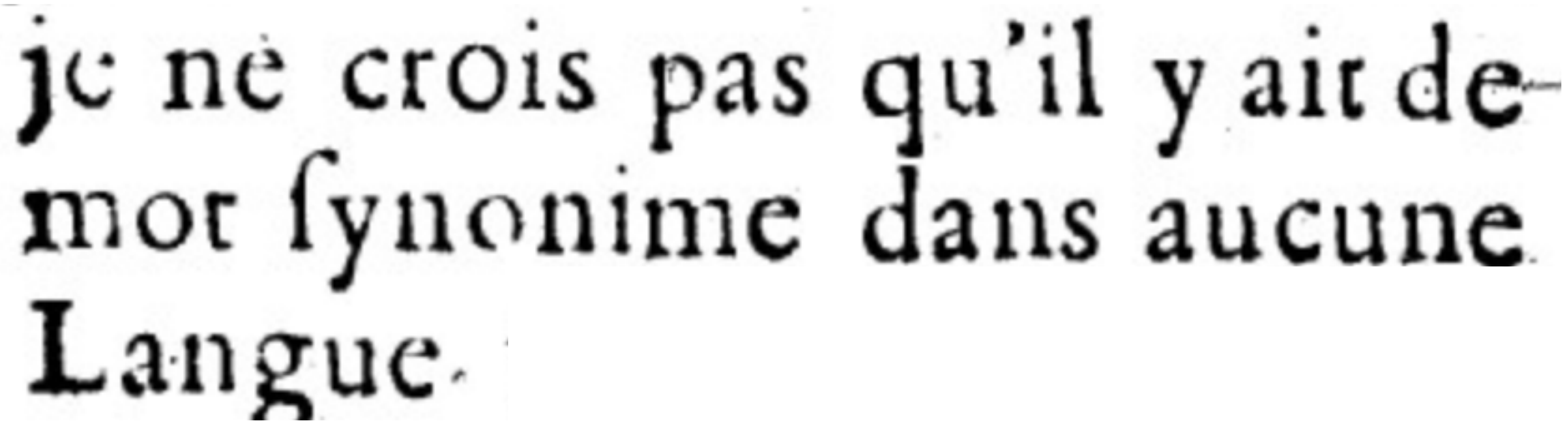
LA JUSTESSE
DE LA
LANGUE FRANÇOISE,
OU
LES DIFFERENTES SIGNIFICATIONS
DES MOTS QUI PASSENT
POUR
SYNONIMES.

Par M. l'Abbé GIRARD C. D. M. D. D. B.

SPIRAT SPIRITUS
UBI VULT

A PARIS,
Chez LAURENT D'HOURY, Imprimeur-Libraire, au bas de la ruë de la Harpe, vis-à vis la ruë S. Severin, au Saint Esprit.

M. DCC. XVIII.
Avec Approbation & Privilege du Roy.



je ne crois pas qu'il y ait de mot ſynonime dans aucune Langue.

[I do not believe that there is a synonymous word in any language]

Relations between word senses

Synonymy

big / *large*

oculist / *eye-doctor*

car / *automobile*

water / H_2O

draft / *draught*

yes / *no*

dark / *light*

hot / *cold*

up / *down*

happy / *sad*

Relations between word senses

Synonymy	*Antonymy*
big / *large*	*yes* / *no*
oculist / *eye-doctor*	*dark* / *light*
car / *automobile*	*hot* / *cold*
water / H_2O	*up* / *down*
draft / *draught*	*happy* / *sad*

Relations between word senses

Synonymy

big / *large*

oculist / *eye-doctor*

car / *automobile*

water / H_2O

draft / *draught*

Antonymy

yes / *no*

dark / *light*

hot / *cold*

up / *down*

happy / *sad*

cat / *dog*

cardiologist / *pulmonologist*

car / *bus*

sheep / *goat*

glass / *mug*

Relations between word senses

Synonymy

big / large

oculist / eye-doctor

car / automobile

water / H_2O

draft / draught

Antonymy

yes / no

dark / light

hot / cold

up / down

happy / sad

Similarity

cat / dog

cardiologist / pulmonologist

car / bus

sheep / goat

glass / mug

Relations between word senses

Synonymy

big / *large*

oculist / *eye-doctor*

car / *automobile*

water / H_2O

draft / *draught*

Antonymy

yes / *no*

dark / *light*

hot / *cold*

up / *down*

happy / *sad*

Similarity

cat / *dog*

cardiologist / *pulmonologist*

car / *bus*

sheep / *goat*

glass / *mug*

coffee / *cup*

waiter / *menu*

farm / *cow*

house / *roof*

theater / *actor*

Relations between word senses

Synonymy

big / large

oculist / eye-doctor

car / automobile

water / H_2O

draft / draught

Antonymy

yes / no

dark / light

hot / cold

up / down

happy / sad

Similarity

cat / dog

cardiologist / pulmonologist

car / bus

sheep / goat

glass / mug

Association

coffee / cup

waiter / menu

farm / cow

house / roof

theater / actor

Semantic field

Words that

- cover a particular semantic domain
- bear structured relations with each other

Examples:

hospitals	*surgeon*, *scalpel*, *nurse*, *anesthetic*, *hospital*
restaurants	*waiter*, *menu*, *plate*, *food*, *chef*
houses	*door*, *roof*, *kitchen*, *family*, *bed*

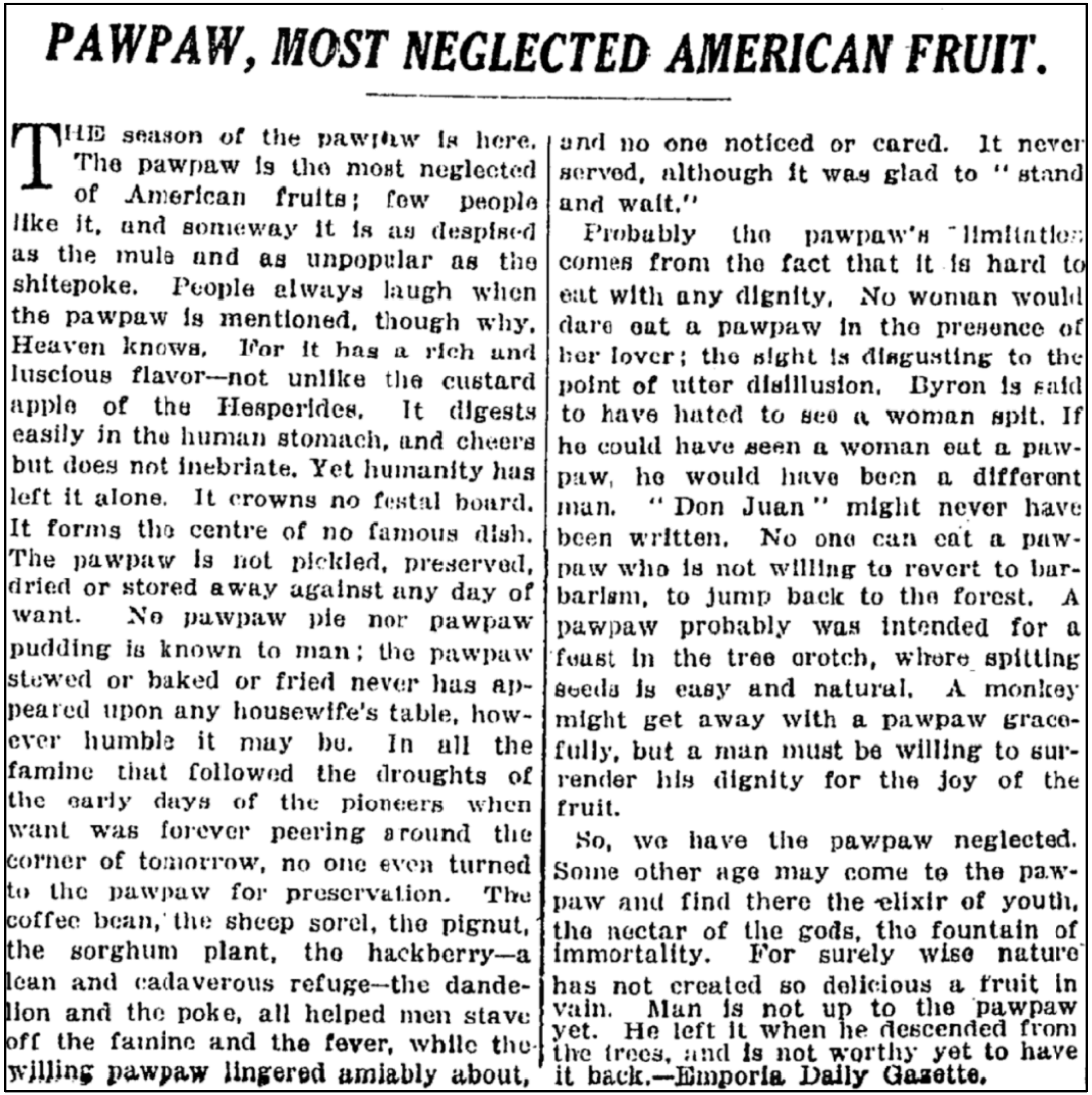
We can also think about aspects of meaning that have to do with a single lemma, e.g., the positive or negative *connotation* of words.

copy	*fake*
replica	*knockoff*
reproduction	*forgery*
positive	*negative*

Can we build a representation of these kinds of word meanings?

Vector semantics

If you haven’t heard the word before, how could you learn what *pawpaw* means?



PAWPAW, MOST NEGLECTED AMERICAN FRUIT.

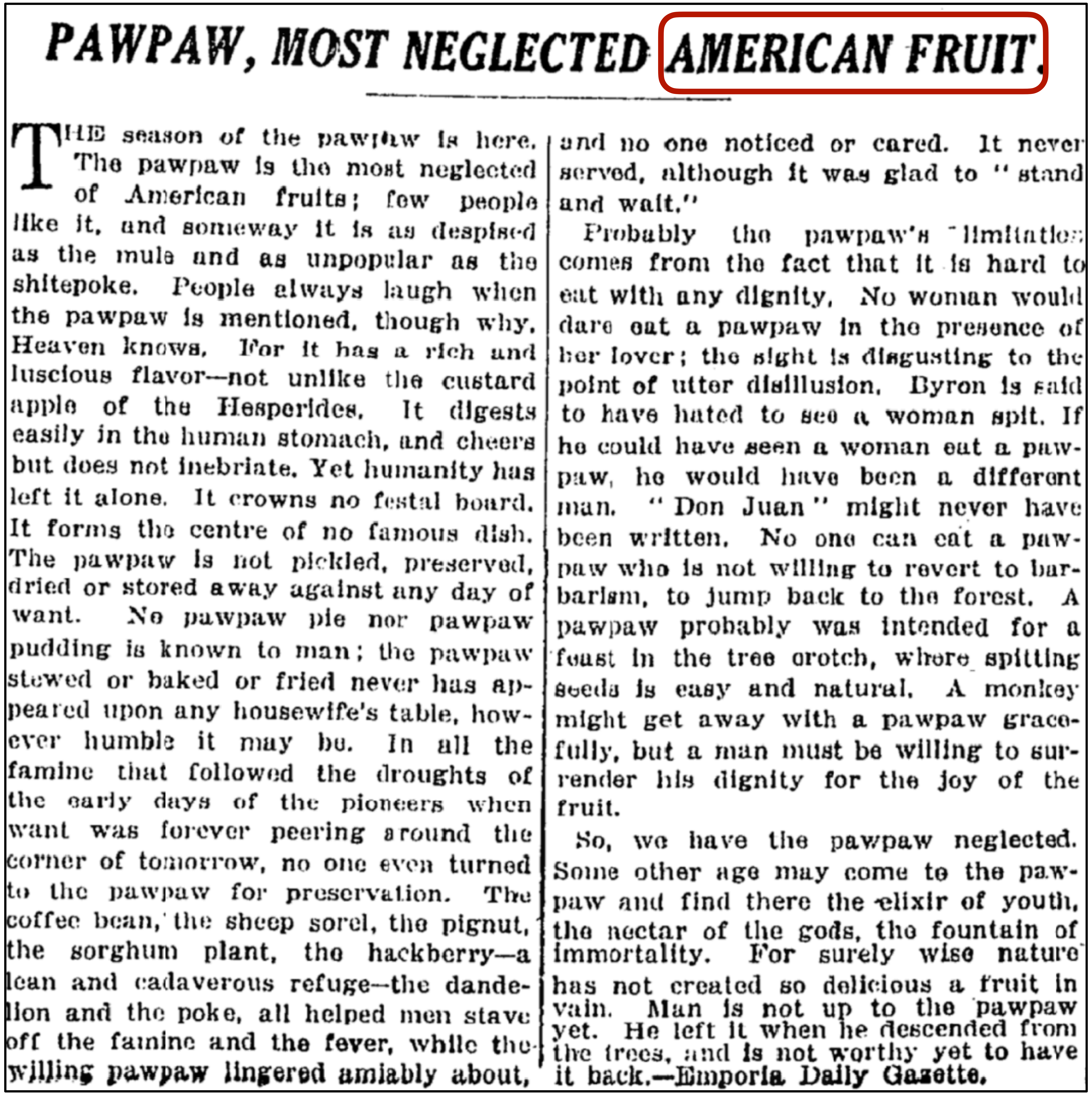
THE season of the pawpaw is here. The pawpaw is the most neglected of American fruits; few people like it, and someway it is as despised as the mule and as unpopular as the shitepoke. People always laugh when the pawpaw is mentioned, though why, Heaven knows. For it has a rich and luscious flavor—not unlike the custard apple of the Hesperides. It digests easily in the human stomach, and cheers but does not inebriate. Yet humanity has left it alone. It crowns no festal board. It forms the centre of no famous dish. The pawpaw is not pickled, preserved, dried or stored away against any day of want. No pawpaw pie nor pawpaw pudding is known to man; the pawpaw stewed or baked or fried never has appeared upon any housewife's table, however humble it may be. In all the famine that followed the droughts of the early days of the pioneers when want was forever peering around the corner of tomorrow, no one even turned to the pawpaw for preservation. The coffee bean, the sheep sorel, the pignut, the sorghum plant, the hackberry—a lean and cadaverous refuge—the dandelion and the poke, all helped men stave off the famine and the fever, while the willing pawpaw lingered amiably about, and no one noticed or cared. It never served, although it was glad to "stand and wait."

Probably the pawpaw's "limitatio comes from the fact that it is hard to eat with any dignity. No woman would dare eat a pawpaw in the presence of her lover; the sight is disgusting to the point of utter disillusion. Byron is said to have hated to see a woman spit. If he could have seen a woman eat a pawpaw, he would have been a different man. "Don Juan" might never have been written. No one can eat a pawpaw who is not willing to revert to barbarism, to jump back to the forest. A pawpaw probably was intended for a feast in the tree crotch, where spitting seeds is easy and natural. A monkey might get away with a pawpaw gracefully, but a man must be willing to surrender his dignity for the joy of the fruit.

So, we have the pawpaw neglected. Some other age may come to the pawpaw and find there the elixir of youth, the nectar of the gods, the fountain of immortality. For surely wise nature has not created so delicious a fruit in vain. Man is not up to the pawpaw yet. He left it when he descended from the trees, and is not worthy yet to have it back.—Emporia Daily Gazette.

The New York Times,

19 November 1922



PAWPAW, MOST NEGLECTED AMERICAN FRUIT.

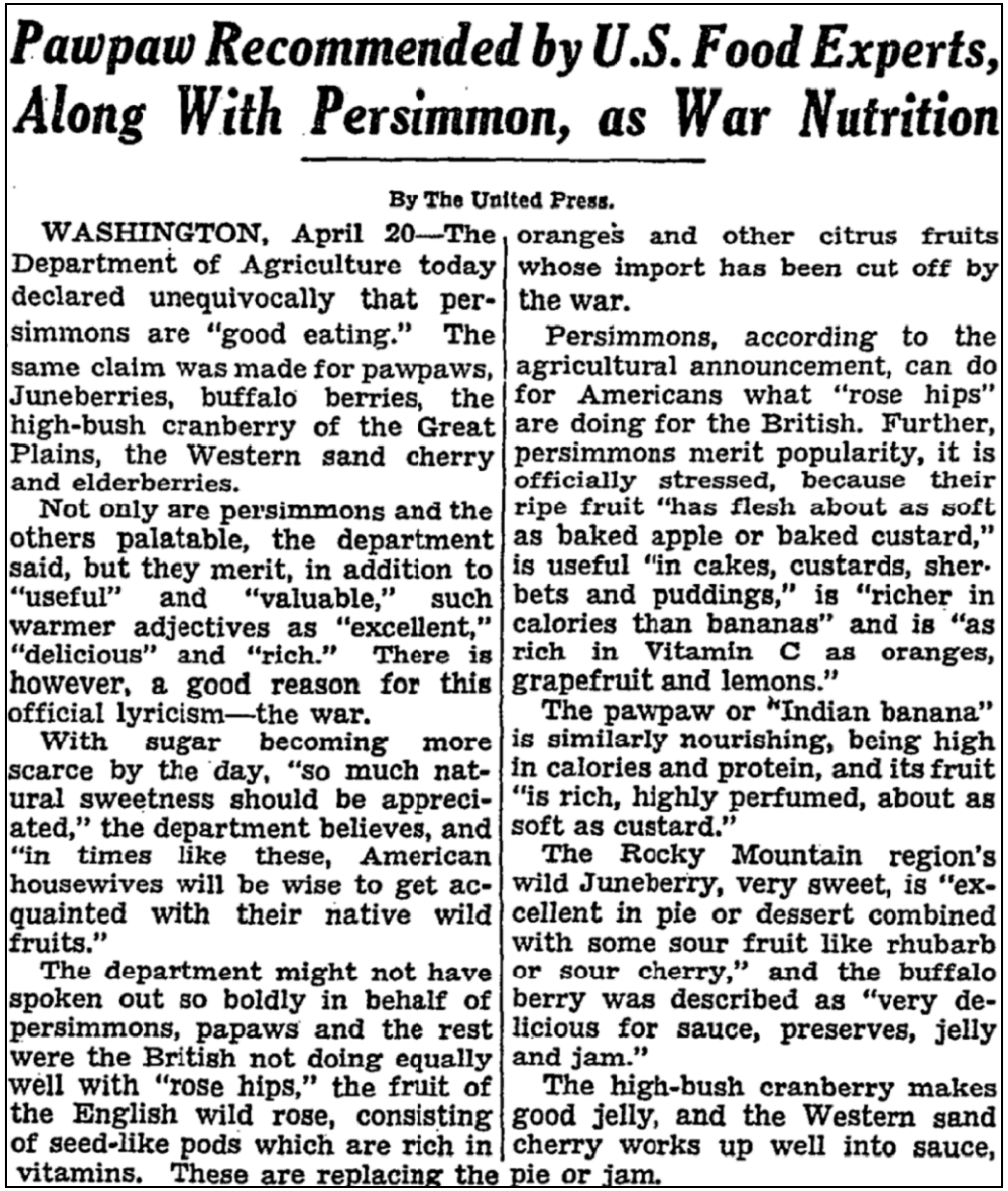
THE season of the pawpaw is here. The pawpaw is the most neglected of American fruits; few people like it, and someway it is as despised as the mule and as unpopular as the shitepoke. People always laugh when the pawpaw is mentioned, though why, Heaven knows. For it has a rich and luscious flavor—not unlike the custard apple of the Hesperides. It digests easily in the human stomach, and cheers but does not inebriate. Yet humanity has left it alone. It crowns no festal board. It forms the centre of no famous dish. The pawpaw is not pickled, preserved, dried or stored away against any day of want. No pawpaw pie nor pawpaw pudding is known to man; the pawpaw stewed or baked or fried never has appeared upon any housewife's table, however humble it may be. In all the famine that followed the droughts of the early days of the pioneers when want was forever peering around the corner of tomorrow, no one even turned to the pawpaw for preservation. The coffee bean, the sheep sorel, the pignut, the sorghum plant, the hackberry—a lean and cadaverous refuge—the dandelion and the poke, all helped men stave off the famine and the fever, while the willing pawpaw lingered amiably about, and no one noticed or cared. It never served, although it was glad to "stand and wait."

Probably the pawpaw's "limitation comes from the fact that it is hard to eat with any dignity. No woman would dare eat a pawpaw in the presence of her lover; the sight is disgusting to the point of utter disillusion. Byron is said to have hated to see a woman spit. If he could have seen a woman eat a pawpaw, he would have been a different man. "Don Juan" might never have been written. No one can eat a pawpaw who is not willing to revert to barbarism, to jump back to the forest. A pawpaw probably was intended for a feast in the tree crotch, where spitting seeds is easy and natural. A monkey might get away with a pawpaw gracefully, but a man must be willing to surrender his dignity for the joy of the fruit.

So, we have the pawpaw neglected. Some other age may come to the pawpaw and find there the elixir of youth, the nectar of the gods, the fountain of immortality. For surely wise nature has not created so delicious a fruit in vain. Man is not up to the pawpaw yet. He left it when he descended from the trees, and is not worthy yet to have it back.—Emporia Daily Gazette.

The New York Times,

19 November 1922



Pawpaw Recommended by U.S. Food Experts, Along With Persimmon, as War Nutrition

By The United Press.

WASHINGTON, April 20—The Department of Agriculture today declared unequivocally that persimmons are "good eating." The same claim was made for pawpaws, Juneberries, buffalo berries, the high-bush cranberry of the Great Plains, the Western sand cherry and elderberries.

Not only are persimmons and the others palatable, the department said, but they merit, in addition to "useful" and "valuable," such warmer adjectives as "excellent," "delicious" and "rich." There is however, a good reason for this official lyricism—the war.

With sugar becoming more scarce by the day, "so much natural sweetness should be appreciated," the department believes, and "in times like these, American housewives will be wise to get acquainted with their native wild fruits."

The department might not have spoken out so boldly in behalf of persimmons, papaws and the rest were the British not doing equally well with "rose hips," the fruit of the English wild rose, consisting of seed-like pods which are rich in vitamins. These are replacing the oranges and other citrus fruits whose import has been cut off by the war.

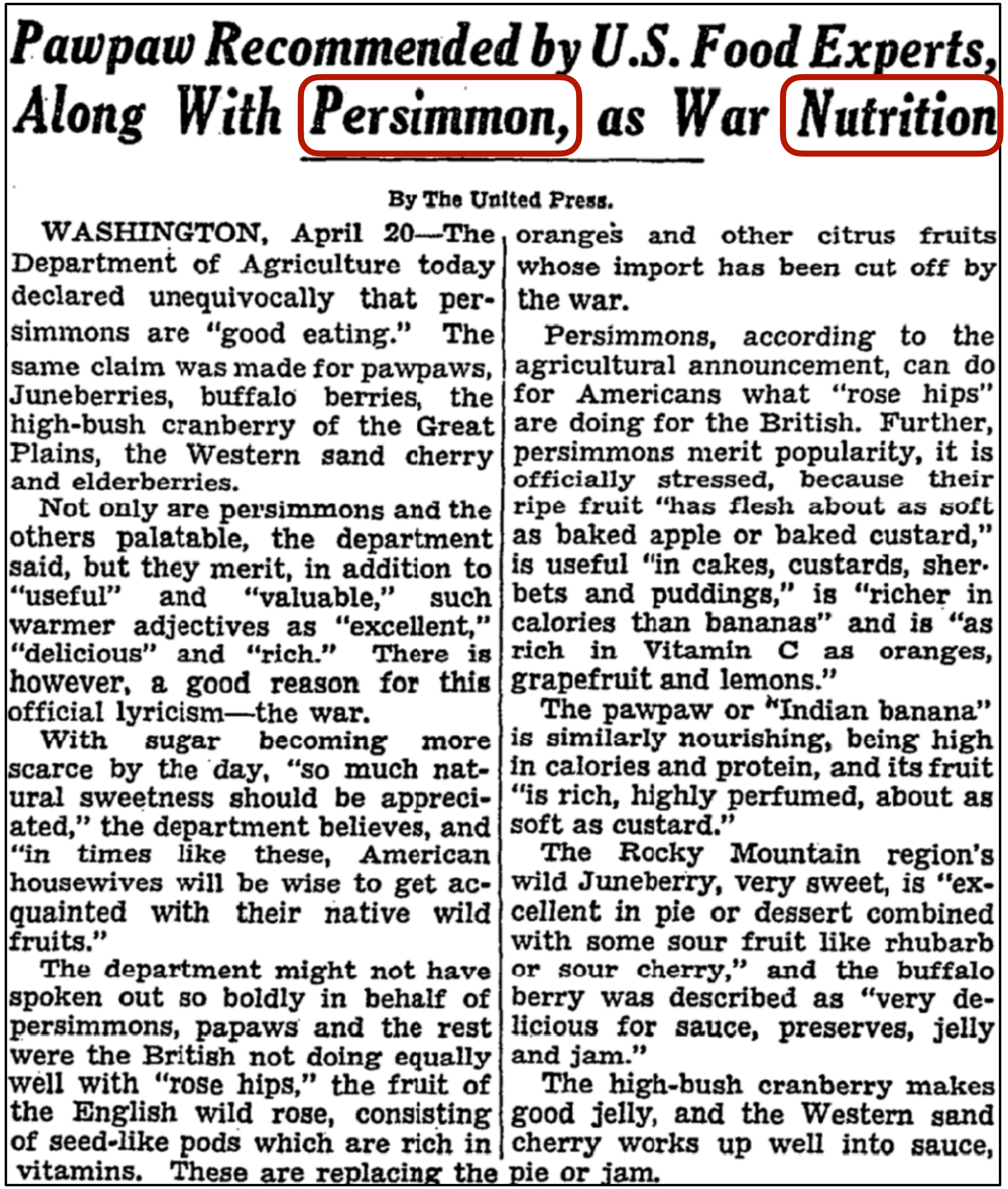
Persimmons, according to the agricultural announcement, can do for Americans what "rose hips" are doing for the British. Further, persimmons merit popularity, it is officially stressed, because their ripe fruit "has flesh about as soft as baked apple or baked custard," is useful "in cakes, custards, sherbets and puddings," is "richer in calories than bananas" and is "as rich in Vitamin C as oranges, grapefruit and lemons."

The pawpaw or "Indian banana" is similarly nourishing, being high in calories and protein, and its fruit "is rich, highly perfumed, about as soft as custard."

The Rocky Mountain region's wild Juneberry, very sweet, is "excellent in pie or dessert combined with some sour fruit like rhubarb or sour cherry," and the buffalo berry was described as "very delicious for sauce, preserves, jelly and jam."

The high-bush cranberry makes good jelly, and the Western sand cherry works up well into sauce, pie or jam.

The New York Times,
21 April 1942



Pawpaw Recommended by U.S. Food Experts, Along With Persimmon, as War Nutrition

By The United Press.

WASHINGTON, April 20—The Department of Agriculture today declared unequivocally that persimmons are "good eating." The same claim was made for pawpaws, Juneberries, buffalo berries, the high-bush cranberry of the Great Plains, the Western sand cherry and elderberries.

Not only are persimmons and the others palatable, the department said, but they merit, in addition to "useful" and "valuable," such warmer adjectives as "excellent," "delicious" and "rich." There is however, a good reason for this official lyricism—the war.

With sugar becoming more scarce by the day, "so much natural sweetness should be appreciated," the department believes, and "in times like these, American housewives will be wise to get acquainted with their native wild fruits."

The department might not have spoken out so boldly in behalf of persimmons, papaws and the rest were the British not doing equally well with "rose hips," the fruit of the English wild rose, consisting of seed-like pods which are rich in vitamins. These are replacing the oranges and other citrus fruits whose import has been cut off by the war.

Persimmons, according to the agricultural announcement, can do for Americans what "rose hips" are doing for the British. Further, persimmons merit popularity, it is officially stressed, because their ripe fruit "has flesh about as soft as baked apple or baked custard," is useful "in cakes, custards, sherbets and puddings," is "richer in calories than bananas" and is "as rich in Vitamin C as oranges, grapefruit and lemons."

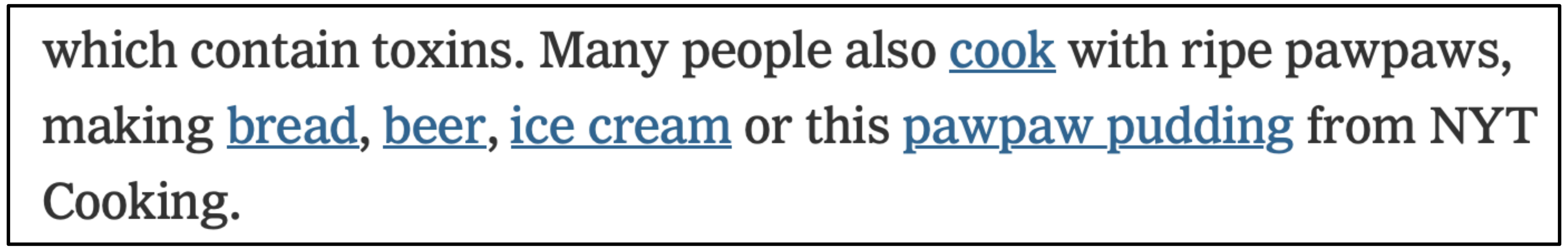
The pawpaw or "Indian banana" is similarly nourishing, being high in calories and protein, and its fruit "is rich, highly perfumed, about as soft as custard."

The Rocky Mountain region's wild Juneberry, very sweet, is "excellent in pie or dessert combined with some sour fruit like rhubarb or sour cherry," and the buffalo berry was described as "very delicious for sauce, preserves, jelly and jam."

The high-bush cranberry makes good jelly, and the Western sand cherry works up well into sauce, pie or jam.

The New York Times,

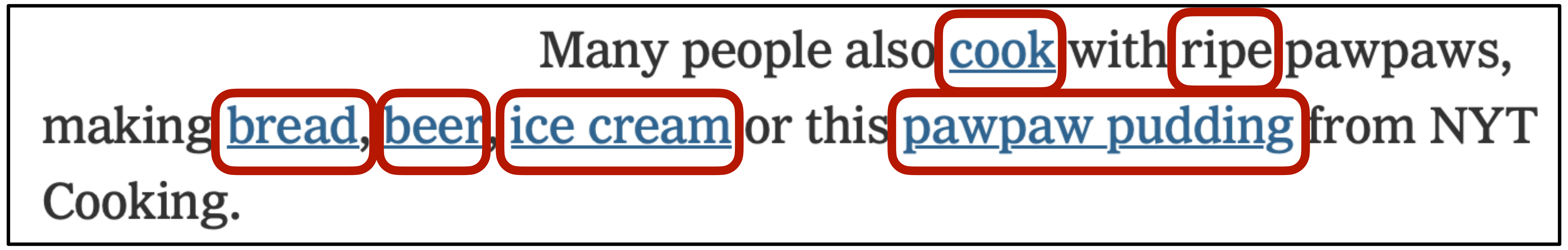
21 April 1942



which contain toxins. Many people also cook with ripe pawpaws, making bread, beer, ice cream or this pawpaw pudding from NYT Cooking.

The New York Times,

19 October 2020



Many people also cook with ripe pawpaws, making bread, beer, ice cream or this pawpaw pudding from NYT Cooking.

The New York Times,
19 October 2020

The pawpaw is also pollinated by flies and other insects rather than by honeybees, said Mr. Dain, and it flowers over several weeks instead of all at once, which ensures that fruit isn't lost to the Northeast's spring frosts.

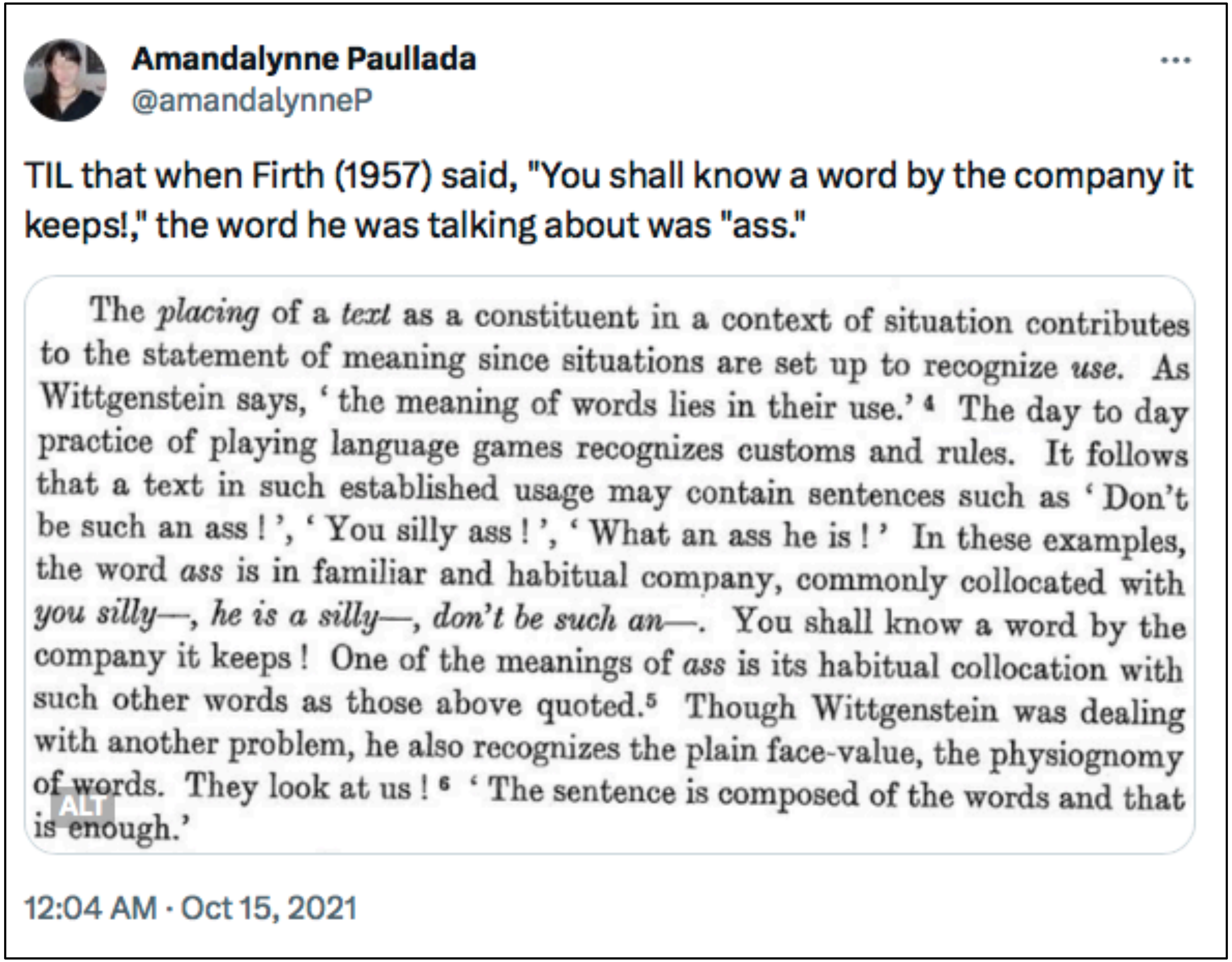
The New York Times,
19 October 2020

This sounds like another sense – the tree rather than the fruit.

"You shall know a word by the company it keeps!"

J. R. Firth, 1957

Cambridge



Amandalynne Paullada
@amandalynneP

TIL that when Firth (1957) said, "You shall know a word by the company it keeps!," the word he was talking about was "ass."

The *placing* of a *text* as a constituent in a context of situation contributes to the statement of meaning since situations are set up to recognize *use.* As Wittgenstein says, ‘the meaning of words lies in their use.’[4] The day to day practice of playing language games recognizes customs and rules. It follows that a text in such established usage may contain sentences such as ‘Don’t be such an ass!’, ‘You silly ass!’, ‘What an ass he is!’ In these examples, the word *ass* is in familiar and habitual company, commonly collocated with *you silly—, he is a silly—, don’t be such an—.* You shall know a word by the company it keeps! One of the meanings of *ass* is its habitual collocation with such other words as those above quoted.[5] Though Wittgenstein was dealing with another problem, he also recognizes the plain face-value, the physiognomy of words. They look at us![6] ‘The sentence is composed of the words and that is enough.’

12:04 AM · Oct 15, 2021

What words can appear in these contexts?

Word 1

drown, bathroom, shower, fill, fall, lie, electrocute, toilet, whirlpool, iron, gin

Word 2

eat, fall, pick, slice, peel, lie, tree, throw, fruit, pie, bite, crab, grate

Word 3

advocate, overthrow, establish, citizen, ideal, representative, dictatorship, campaign, bastion, freedom

Word 4

spend, enjoy, remember, last, pass, end, die, happen, brighten, relive

What words can appear in these contexts?

bathtub
drown, bathroom, shower, fill, fall, lie, electrocute, toilet, whirlpool, iron, gin

Word 2
eat, fall, pick, slice, peel, lie, tree, throw, fruit, pie, bite, crab, grate

Word 3
advocate, overthrow, establish, citizen, ideal, representative, dictatorship, campaign, bastion, freedom

Word 4
spend, enjoy, remember, last, pass, end, die, happen, brighten, relive

What words can appear in these contexts?

bathtub
drown, bathroom, shower, fill, fall, lie, electrocute, toilet, whirlpool, iron, gin

apple
eat, fall, pick, slice, peel, lie, tree, throw, fruit, pie, bite, crab, grate

Word 3
advocate, overthrow, establish, citizen, ideal, representative, dictatorship, campaign, bastion, freedom

Word 4
spend, enjoy, remember, last, pass, end, die, happen, brighten, relive

What words can appear in these contexts?

bathtub
drown, bathroom, shower, fill, fall, lie, electrocute, toilet, whirlpool, iron, gin

apple
eat, fall, pick, slice, peel, lie, tree, throw, fruit, pie, bite, crab, grate

democracy
advocate, overthrow, establish, citizen, ideal, representative, dictatorship, campaign, bastion, freedom

Word 4
spend, enjoy, remember, last, pass, end, die, happen, brighten, relive

What words can appear in these contexts?

bathtub
drown, bathroom, shower, fill, fall, lie, electrocute, toilet, whirlpool, iron, gin

apple
eat, fall, pick, slice, peel, lie, tree, throw, fruit, pie, bite, crab, grate

democracy
advocate, overthrow, establish, citizen, ideal, representative, dictatorship, campaign, bastion, freedom

day
spend, enjoy, remember, last, pass, end, die, happen, brighten, relive

The *distributional hypothesis*: Similar words appear in similar contexts.

Therefore, we can measure similarity in meaning as similarity in contexts.

One more:

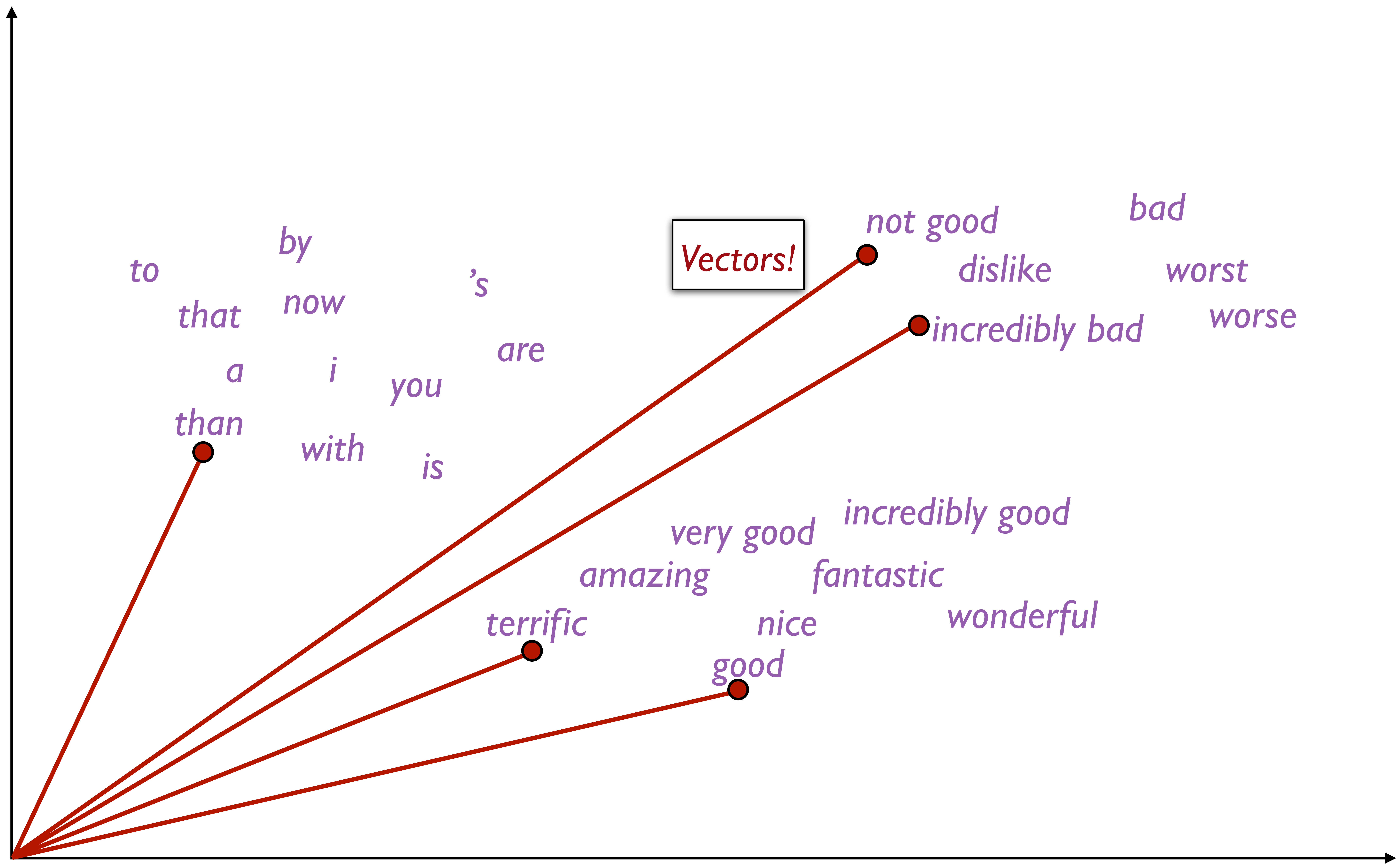
Word 5

eat, paint, peel, last,
apple, fruit, juice,
lemon, blue, grow

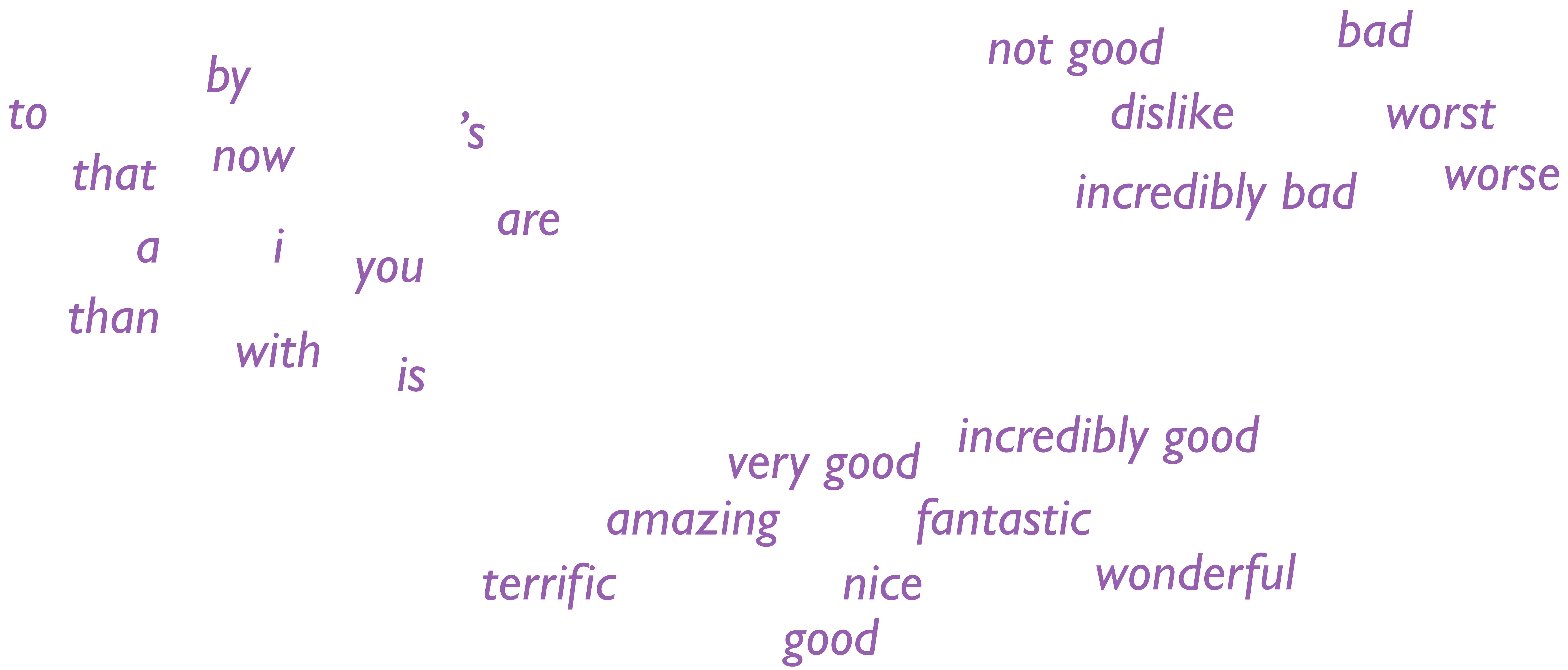
One more:

orange
eat, paint, peel, last,
apple, fruit, juice,
lemon, blue, grow

Vector semantics instantiates the distributional hypothesis by learning the representations of the meanings of words directly from their distributions in texts.


Vectors!
to
by
that
now
’s
are
a
i
you
than
with
is
not good
bad
dislike
worst
incredibly bad
worse
incredibly good
very good
amazing
fantastic
terrific
nice
wonderful
good

Words used in the same context will cluster nearby in the same space.



to
by
that
now
's
are
a
i
you
than
with
is

not good
bad
dislike
worst
incredibly bad
worse

very good
incredibly good
amazing
fantastic
terrific
nice
wonderful
good

The vector representations of words are called *embeddings*.

- In *static embeddings*, each *word type* has a unique embedding.
- In *contextual embeddings*, each *word token* has a unique embedding.
  - This is useful for learning multiple senses.
  - We'll return to contextual embeddings when we study LLMs in a few weeks.

We'll look at two kinds of static embeddings this week:

Simple count embeddings, where words are represented by the counts of nearby words

Word2vec, where representation is created by training a classifier to predict whether a word is likely to appear nearby

Count-based embeddings

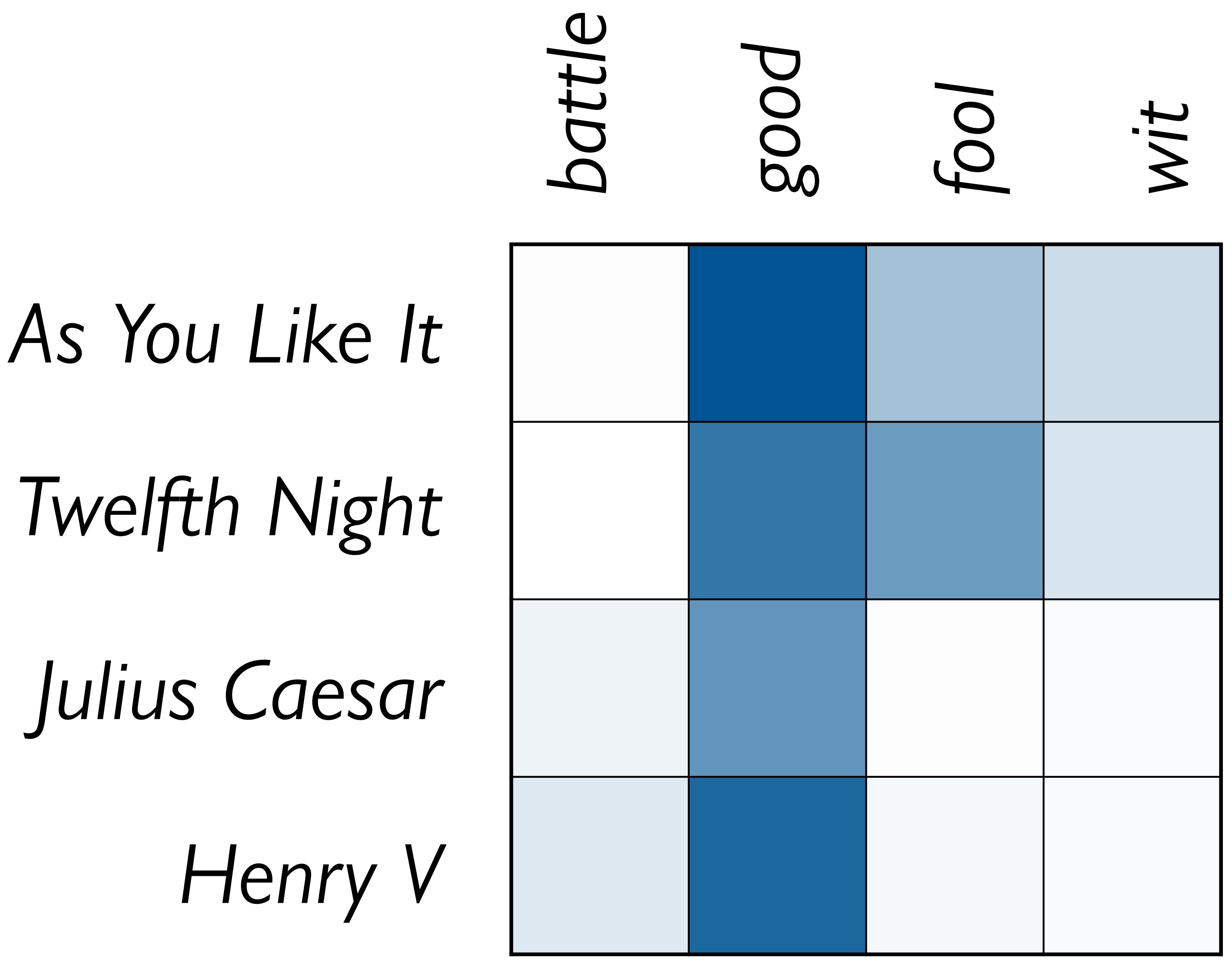
Word vectors

	battle	good	fool	wit
As You Like It	1	114	36	20
Twelfth Night				
Julius Caesar				
Henry V				

Document–word matrix

A document can be represented as the number of times each word occurs in it.

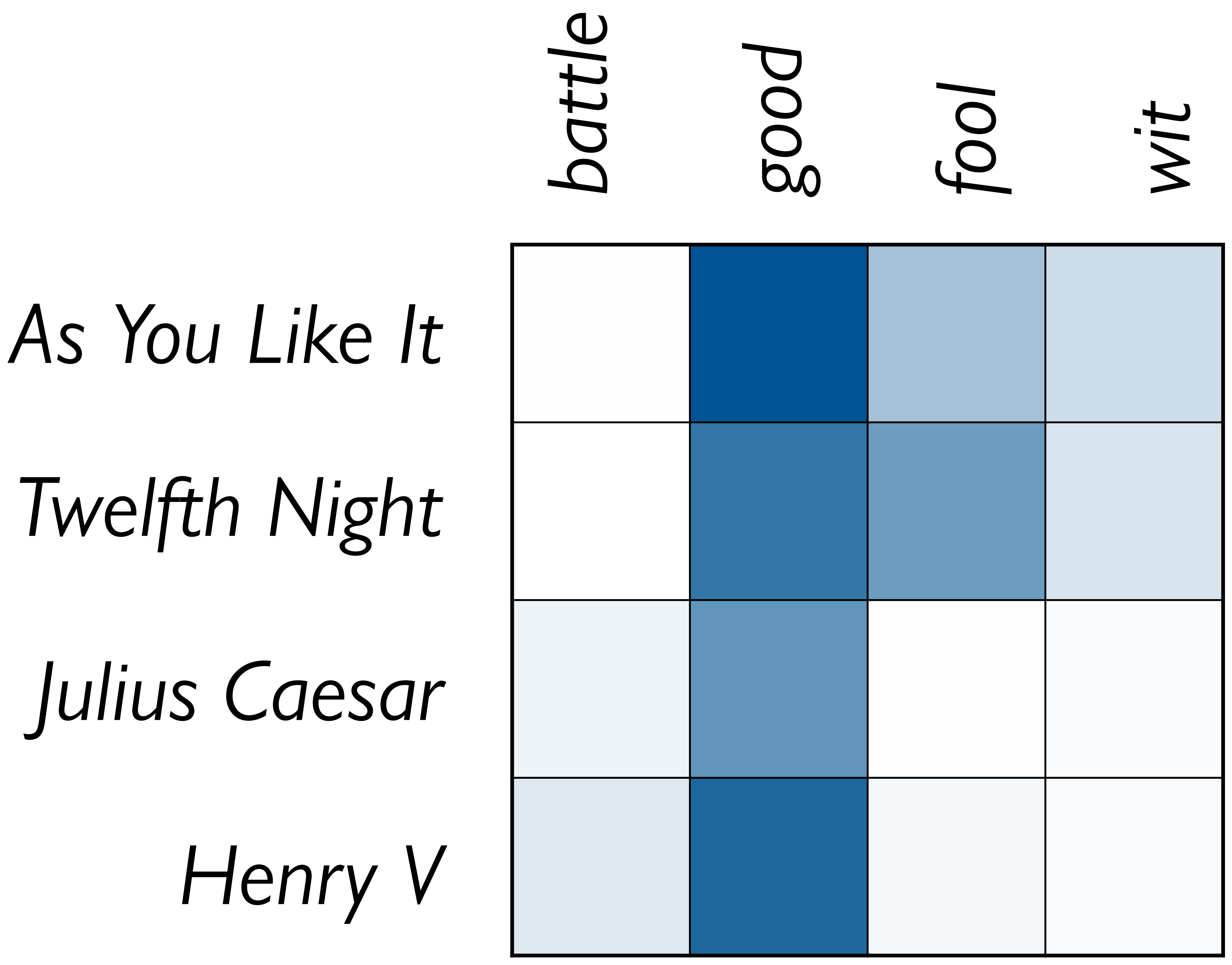
Word vectors



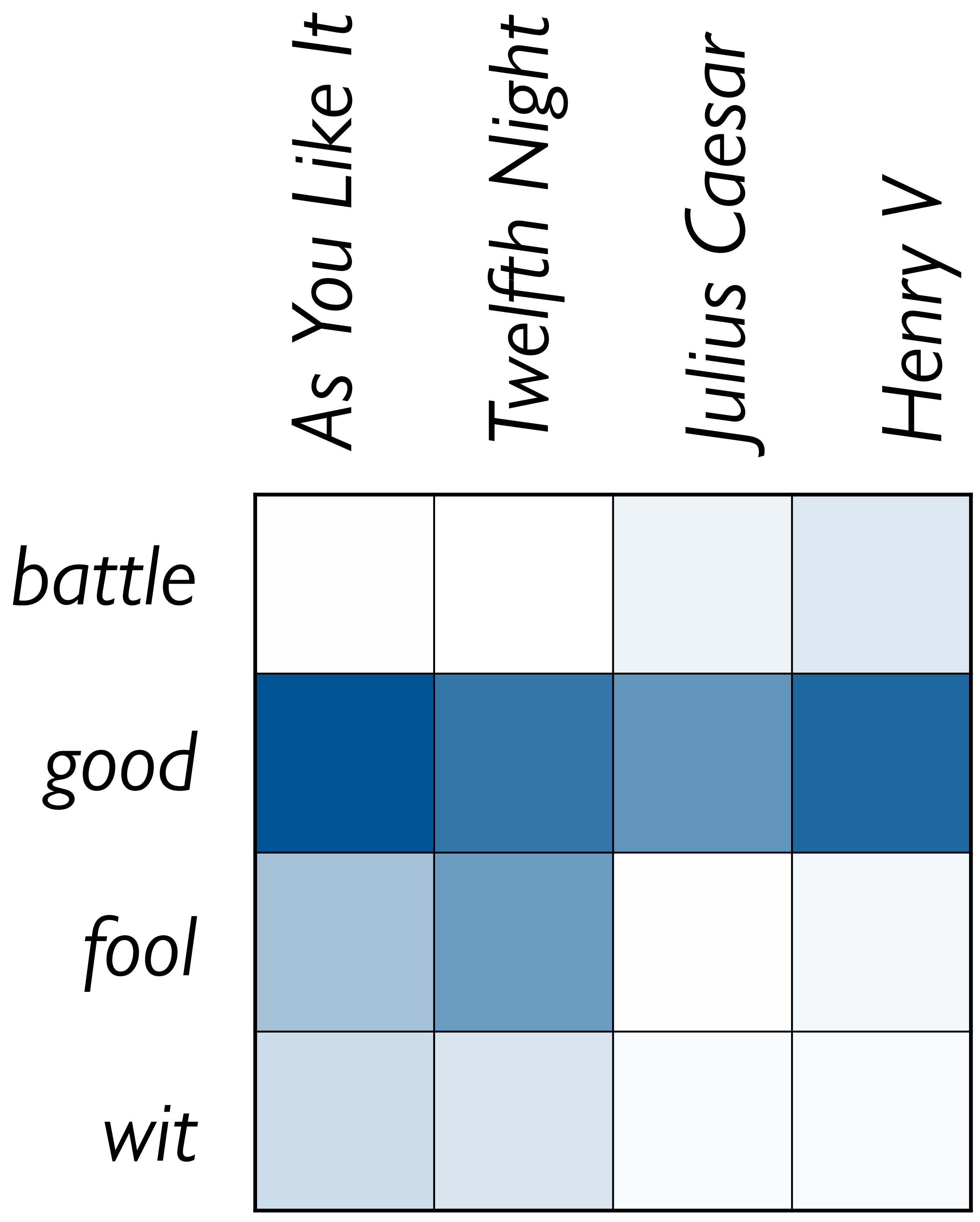
Document–word matrix

A document can be represented as the number of times each word occurs in it.

Word vectors



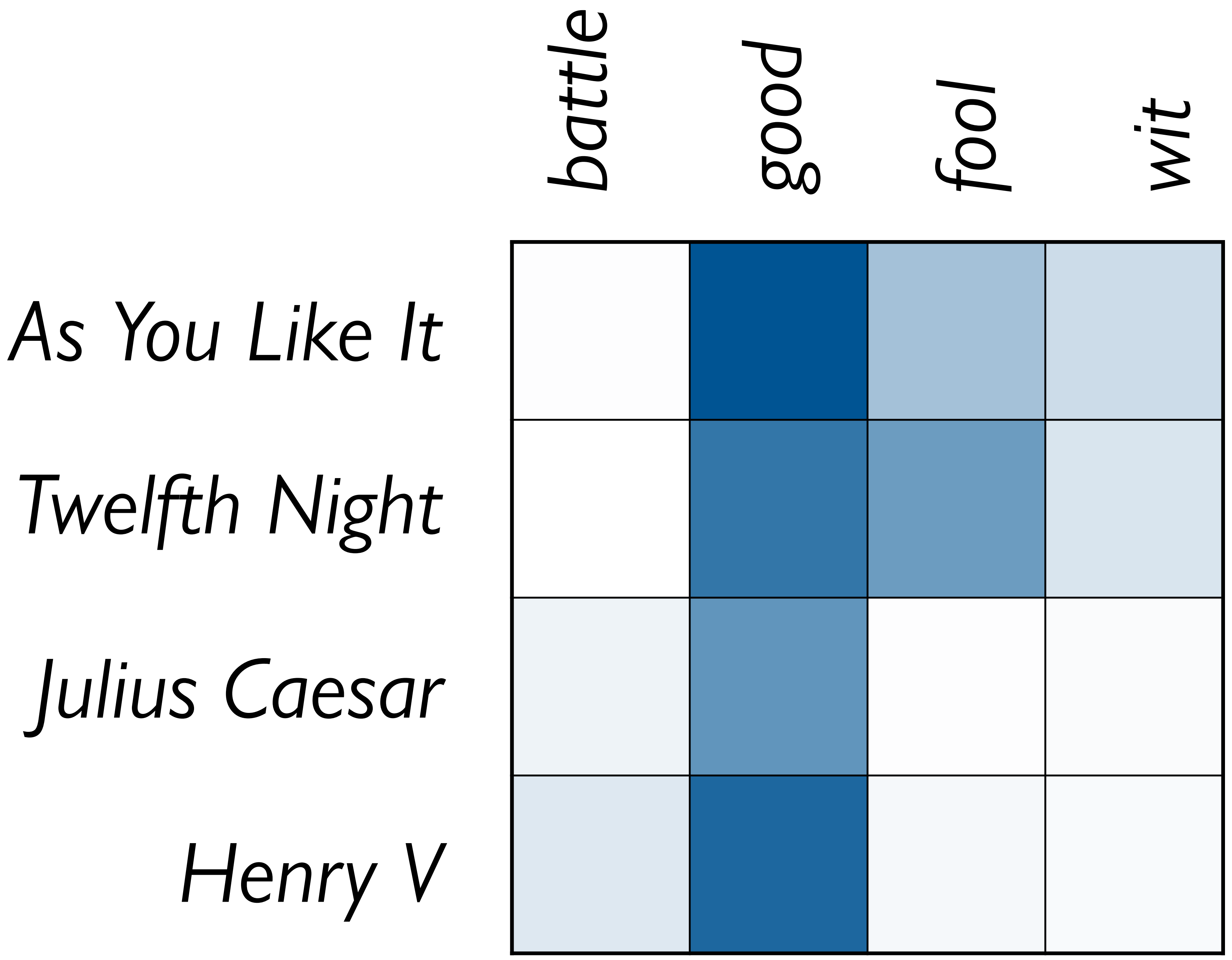
Document–word matrix



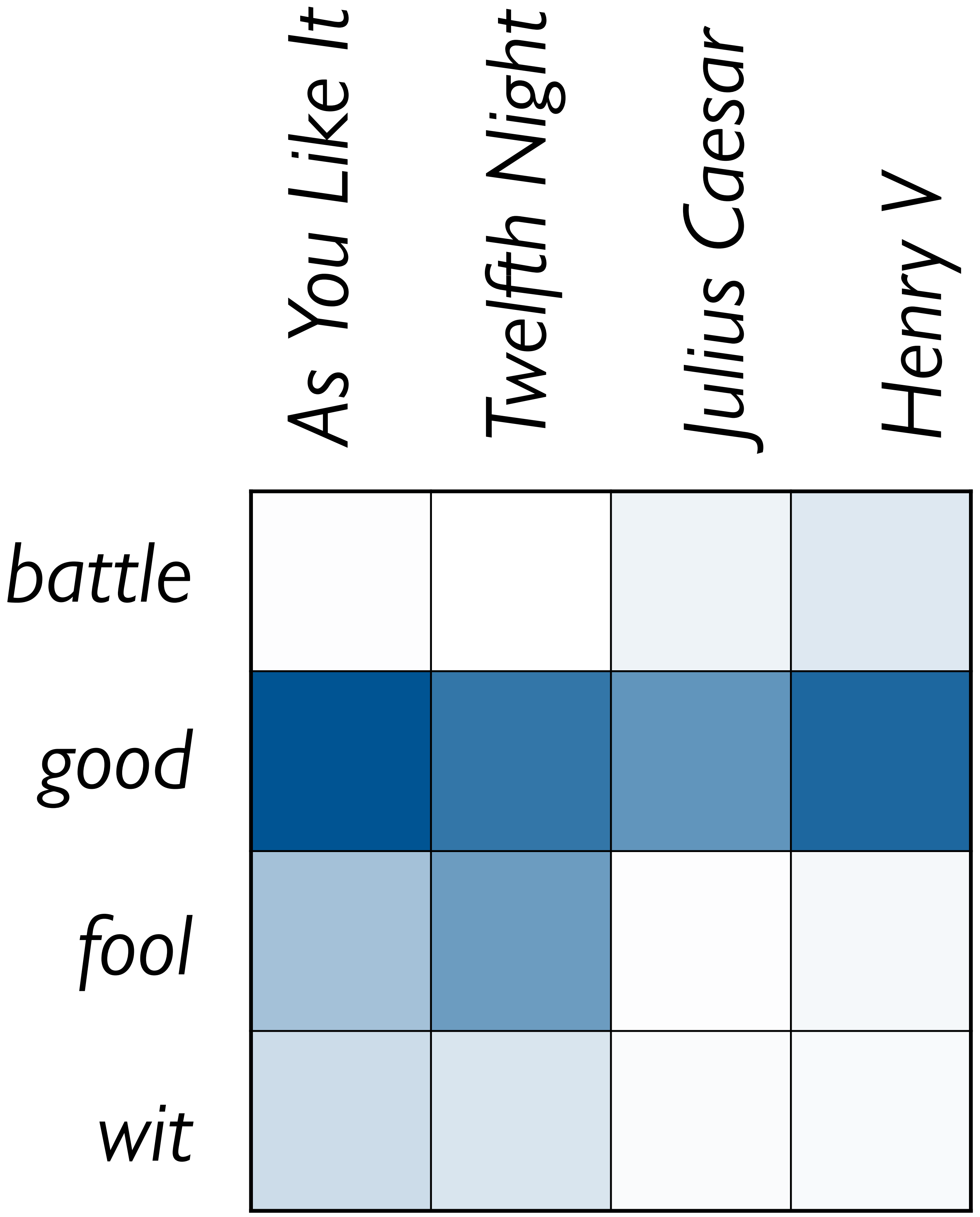
Word–document matrix

A word's meaning can be represented as the documents it appears in!

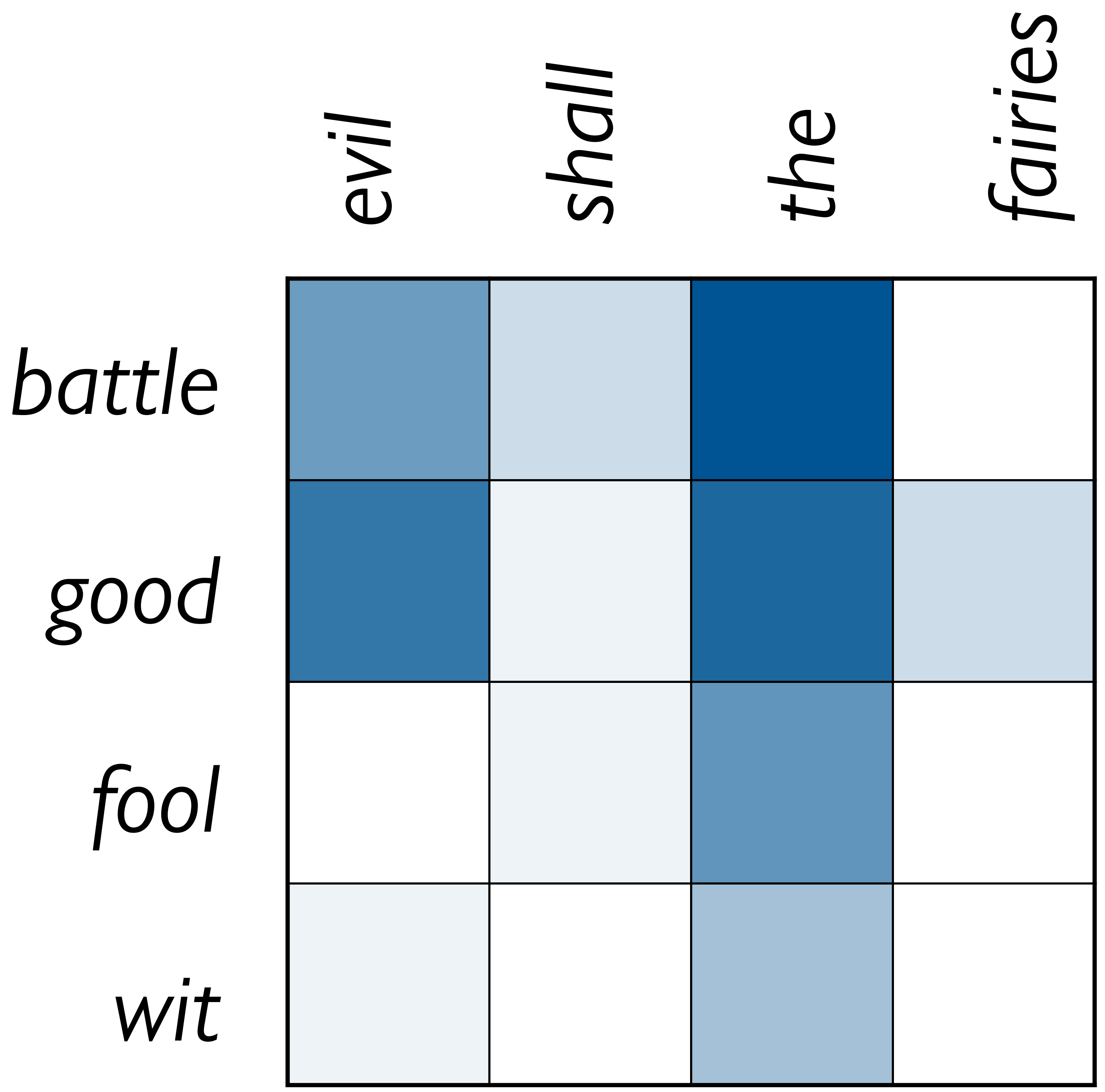
Word vectors



Document–word matrix

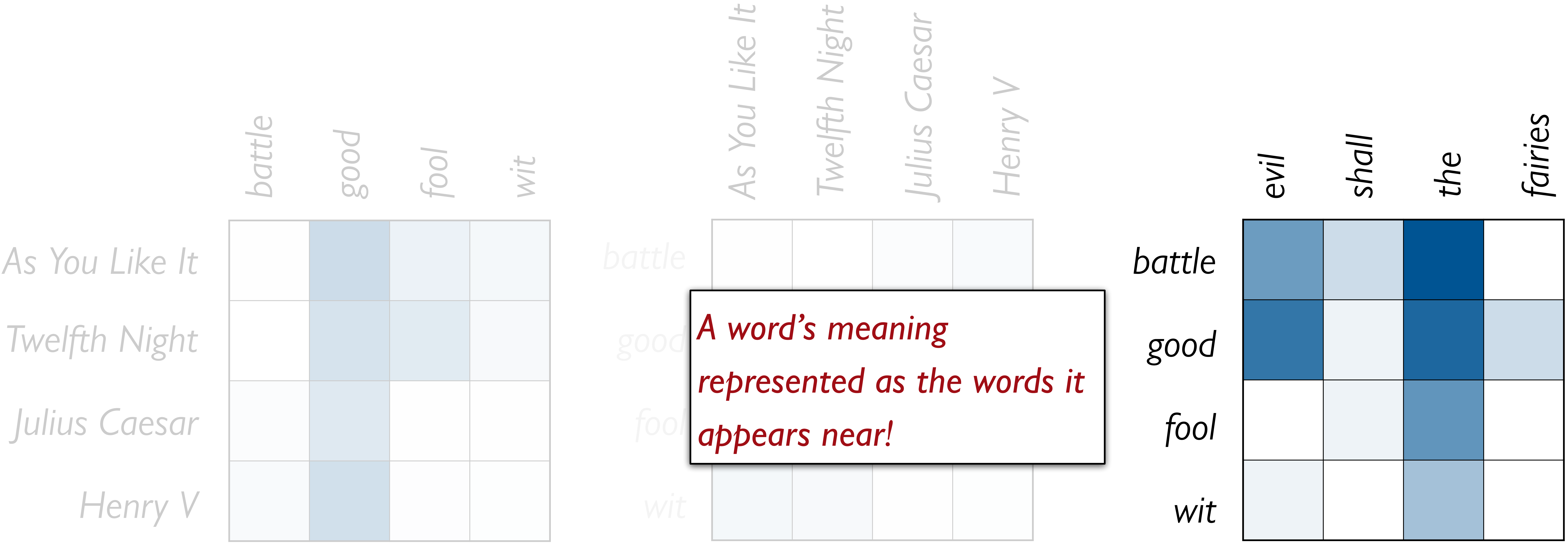


Word–document matrix



Word–context matrix

Word vectors



Document–word matrix

Word–document matrix

Word–context matrix

A word's meaning represented as the words it appears near!

If a corpus can be split into short documents (e.g., tweets or headlines), you might use the entire document as the context.

Otherwise, we might choose a fixed ***context window*** of L tokens to either side, e.g., $L = 2$:

…lemon , a tablespoon of apricot jam , a pinch…

c_1 c_2 w c_3 c_4

target

They picked up red apples that had fallen to the ground.

Eating apples is healthy.

She ate a red apple.

Pick an apple.

They picked up red apples that had fallen to the ground.

Eating apples is healthy.

She ate a red apple.

Pick an apple.

They picked up red apples that had fallen to the ground.

Eating apples is healthy.

She ate a red apple.

Pick an apple.

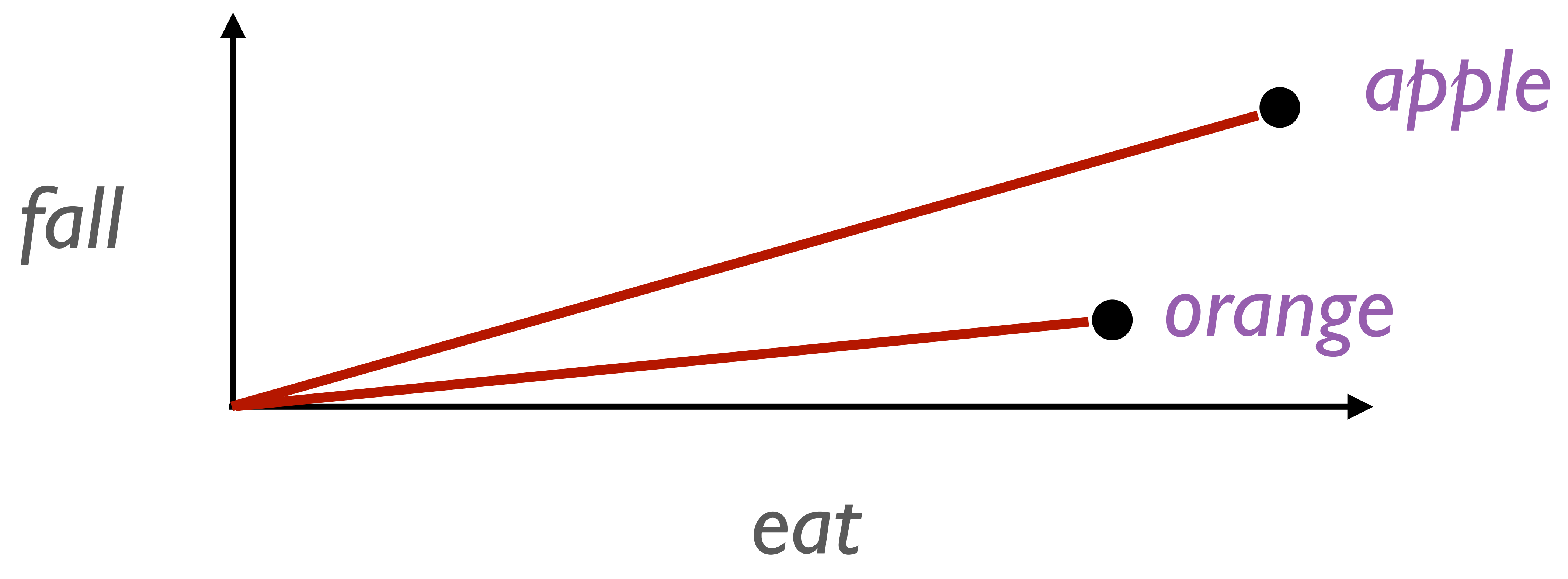
	a	*be*	*eat*	*fall*	*have*	*healthy*	*pick*	*red*	*that*	*up*
apple	1	2	1	1	1	1	2	2	1	1

If we count context words in a larger corpus, we get counts like these:

	eat	*fall*	*ripe*	*slice*	*peel*	*tree*	*throw*	*fruit*	*pie*	*bite*	*crab*
apple	794	244	47	221	208	160	145	156	109	104	88

If we count context words in a larger corpus, we get counts like these:

	eat	*fall*	*ripe*	*slice*	*peel*	*tree*	*throw*	*fruit*	*pie*	*bite*	*crab*
apple	794	244	47	221	208	160	145	156	109	104	88
orange	265	22	25	62	220	64	74	111	4	4	8



Every context word becomes a dimension.

	eat	*fall*	*ripe*	*slice*	*peel*	*tree*	*throw*	*fruit*	*pie*	*bite*	*crab*
apple	794	244	47	221	208	160	145	156	109	104	88
orange	265	22	25	62	220	64	74	111	4	4	8

Given a word–context matrix, we can use each row as an embedding.

These embeddings are *sparse*, meaning they have a lot of zeros (when the word in the row never co-occurs with the word in the column).

What does this get us?

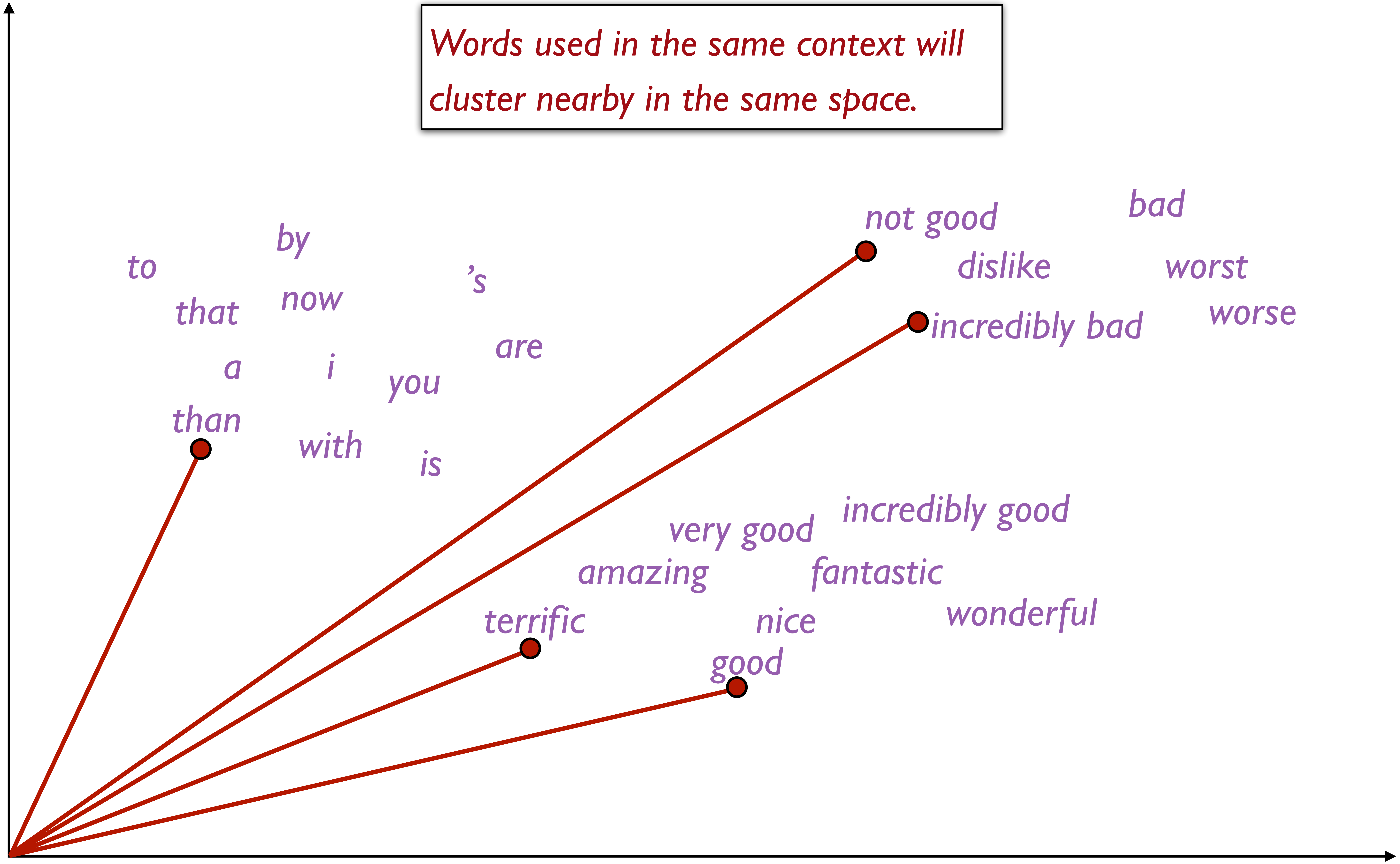
For tasks like sentiment analysis,

using words, we require the *same* word to be in training and test;

using embeddings, it's ok if *similar* words occurred!

Computing word similarity

Words used in the same context will cluster nearby in the same space.

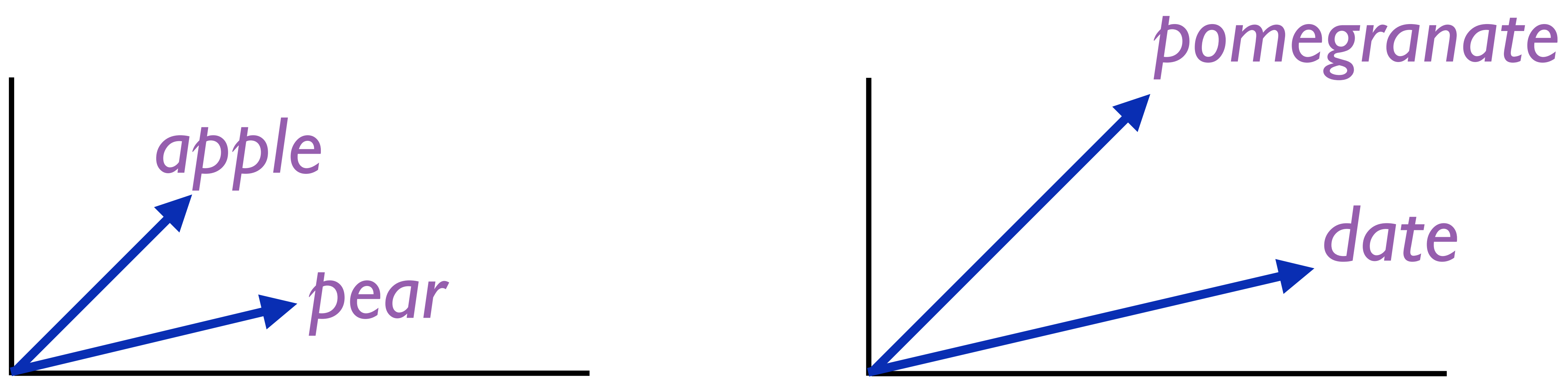


The dot product between two vectors

$$\mathbf{v} \cdot \mathbf{w} = \sum_{i=1}^{N} v_i w_i = v_1 w_1 + v_2 w_2 + \cdots v_N w_N$$

is a scalar that can be a useful similarity metric.

Problem: The longer the vectors are, the higher the dot product is

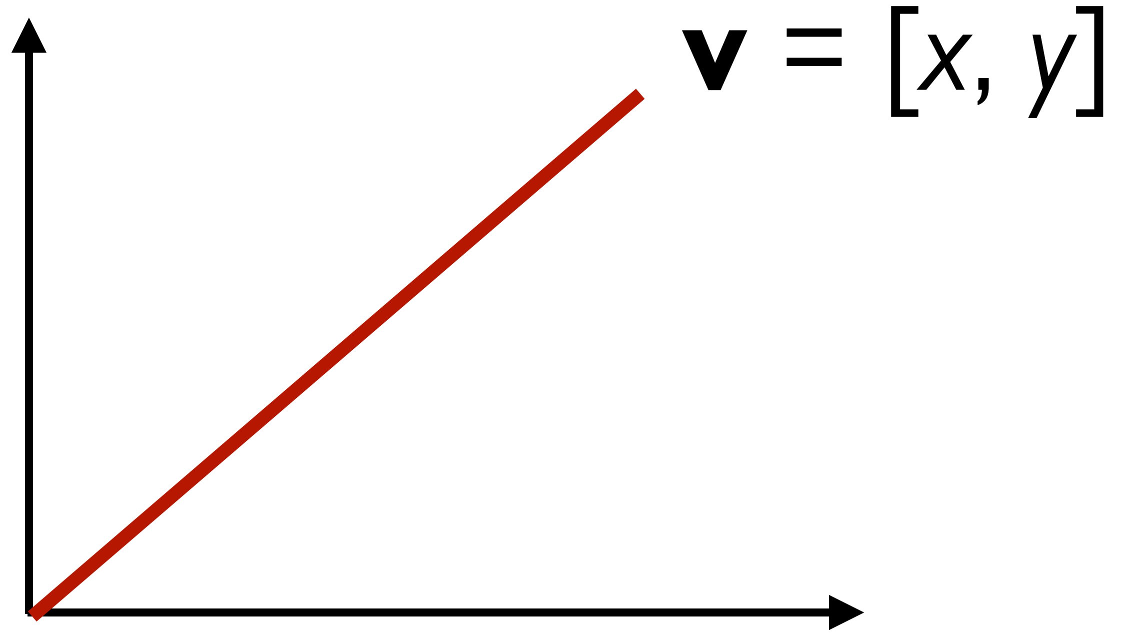


Higher dot product!

Solution: Normalize by vector length.

For a 2-dimensional vector

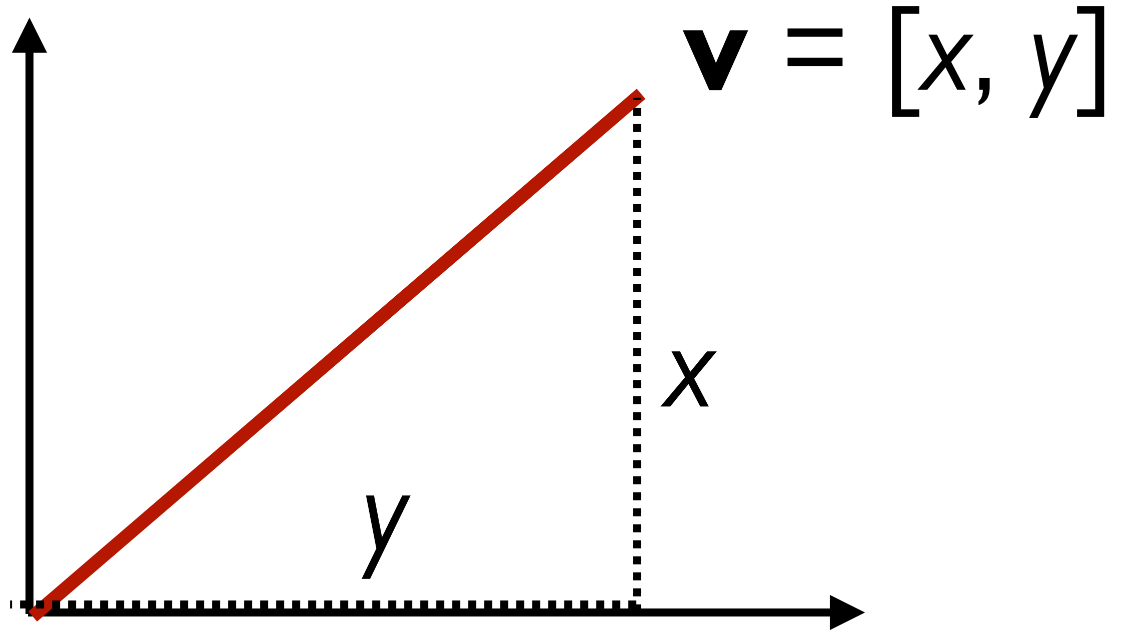
$$|\mathbf{v}| = \sqrt{x^2 + y^2}$$



Solution: Normalize by vector length.

For a 2-dimensional vector

$$|\mathbf{v}| = \sqrt{x^2 + y^2}$$



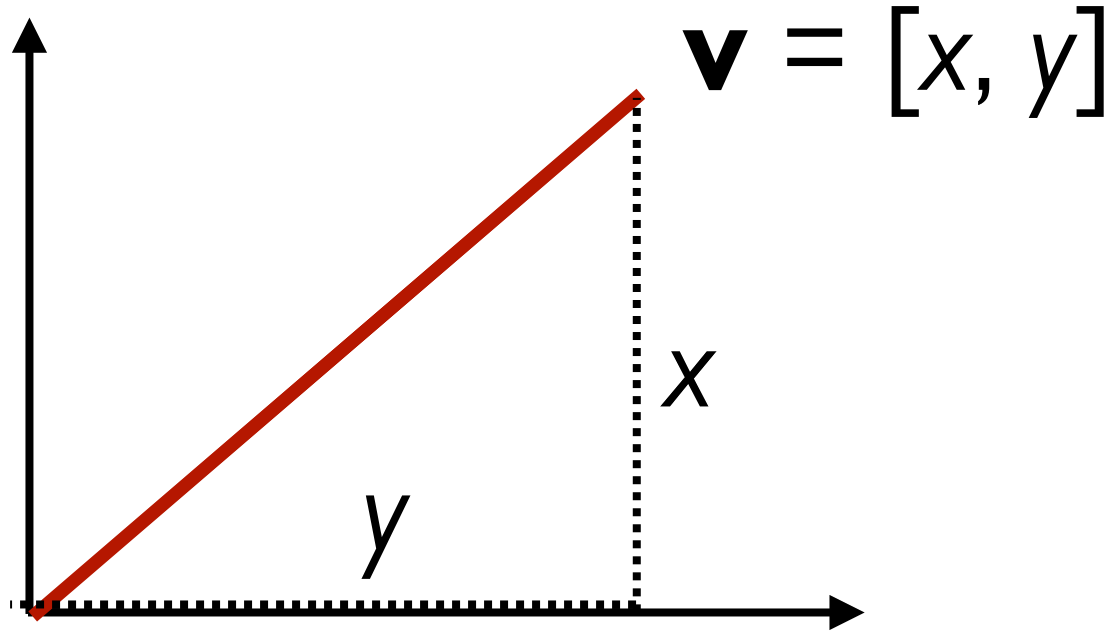
Solution: Normalize by vector length.

For a 2-dimensional vector

$$|\mathbf{v}| = \sqrt{x^2 + y^2}$$

For an N-dimensional vector

$$|\mathbf{v}| = \sqrt{\sum_{i=1}^{N} v_i^2}$$



Solution: Normalize by vector length.

$$|\mathbf{v}| = \sqrt{\sum_{i=1}^{N} v_i^2}$$

$$\frac{\mathbf{v} \cdot \mathbf{w}}{|\mathbf{v}||\mathbf{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i-=1}^{N} w_i^2}}$$

Solution: Normalize by vector length.

$$|\mathbf{v}| = \sqrt{\sum_{i=1}^{N} v_i^2}$$

$$\cos(\mathbf{v}, \mathbf{w}) = \frac{\mathbf{v} \cdot \mathbf{w}}{|\mathbf{v}||\mathbf{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i-=1}^{N} w_i^2}}$$

Surprise! It’s just the cosine of the angle between the vectors!

Cosine as a similarity metric

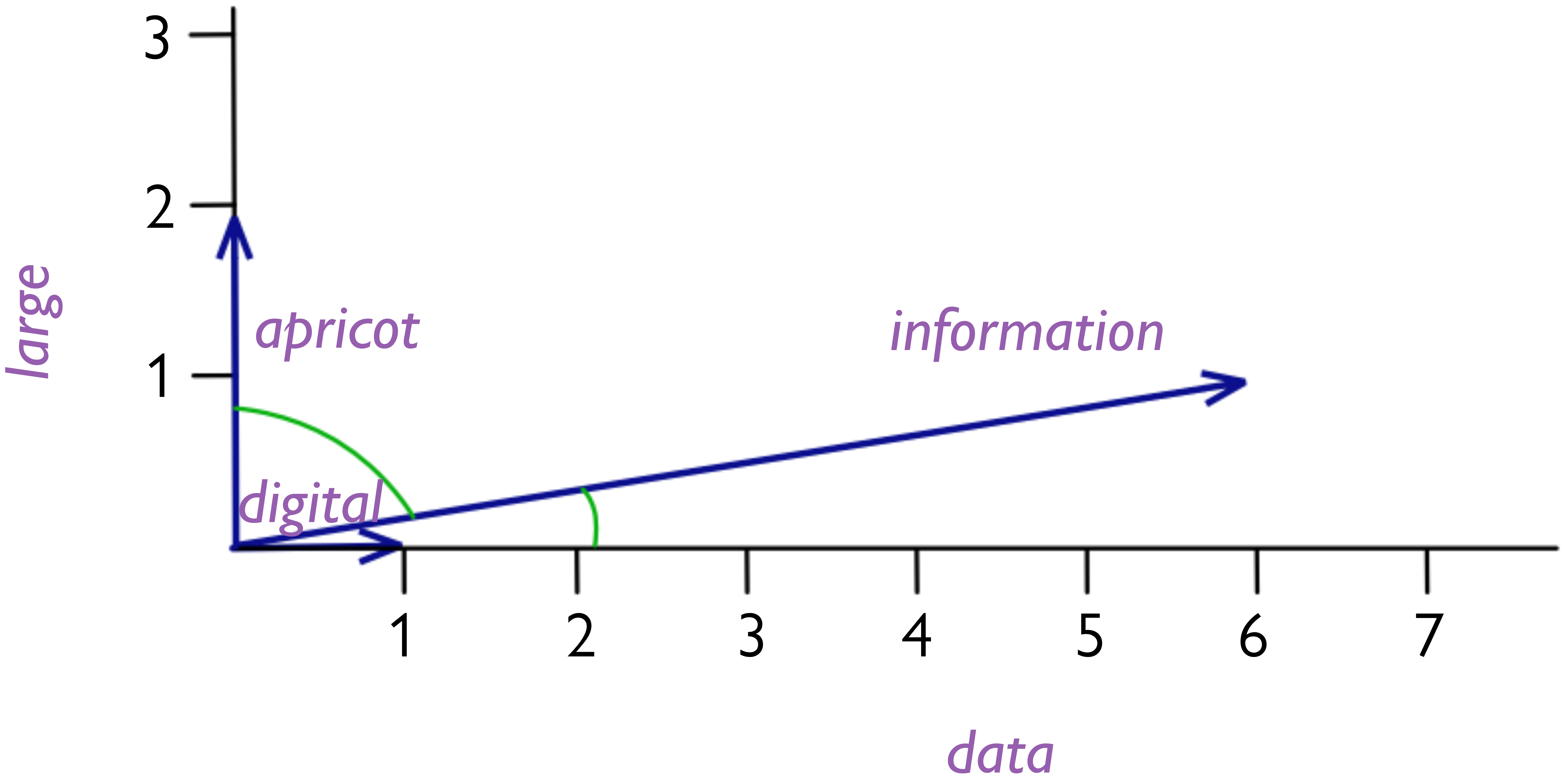
−1: Vectors point in opposite directions

+1: Vectors point the same direction

0: Vectors are orthogonal

But since raw frequency values are non-negative, the cosine for term–term matrix vectors ranges from 0–1.

Cosine as a similarity metric



What now?

Every modern NLP algorithm uses embeddings as the representation of word meaning.

But rather than simple count embeddings, which give us sparse vectors, they usually use embeddings learned with a classifier, which give us dense vectors.

We'll see the most popular technique for this – Word2vec – next class!

Acknowledgments

This class incorporates material from:

Jurafsky & Martin, *Speech and Language Processing*, 3rd ed. draft

Carolyn Anderson, Wellesley College

Katie Keith, Williams College