Distributional Semantics Crash Course

September 11, 2018
CSCI 2952C: Computational Semantics
Instructor: Ellie Pavlick
HTA: Arun Drelich
UTA: Jonathan Chang

Agenda

- Quick Background
- Your Discussion Questions
- Step through of VMS/word2vec
- Announcements

Agenda

- Quick Ba[illegible]und
- Your Discussion Que[illegible]
- Step through of VMS/word2vec
- Announcements

!!!

:-D

???

@!#$

…

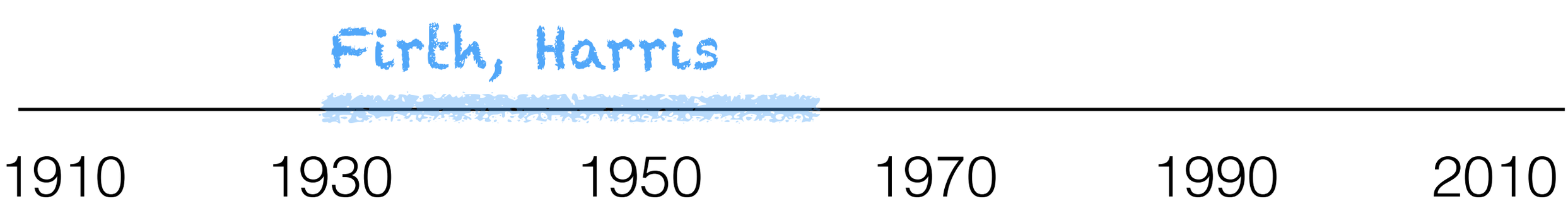
>:(

Constant interruptions from you all

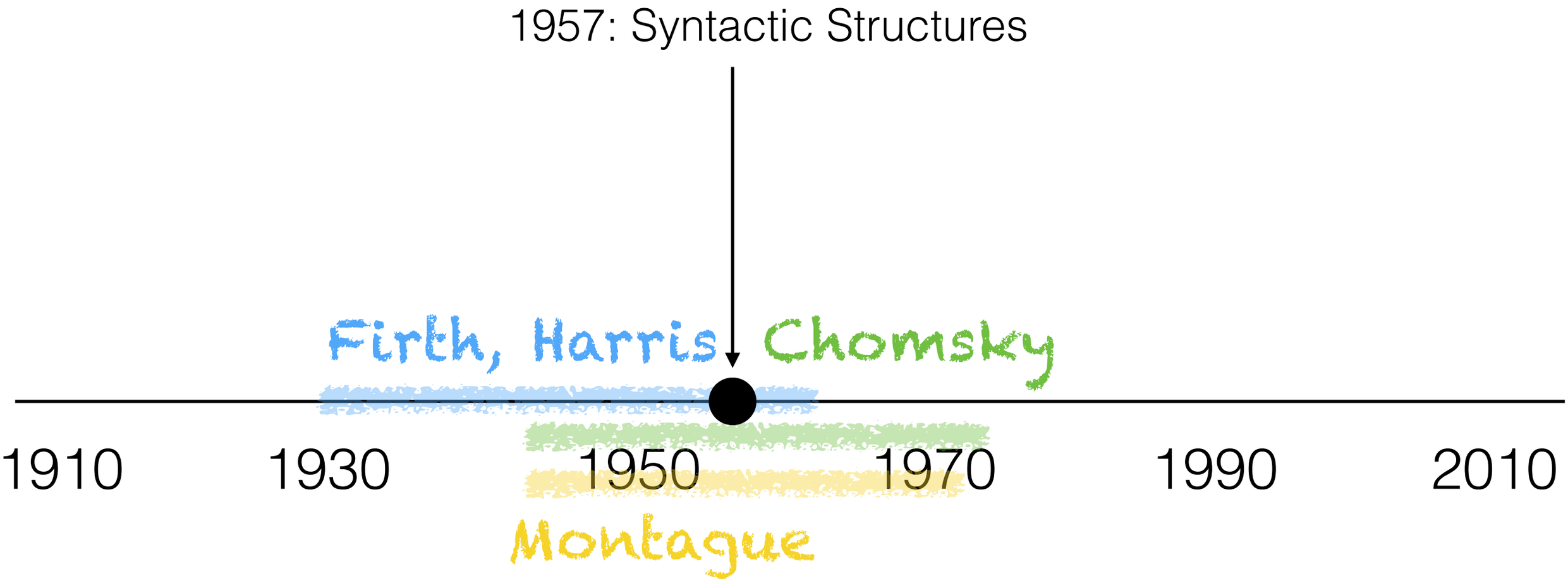
“You shall know a word by the company it keeps!”

— J. R. Firth

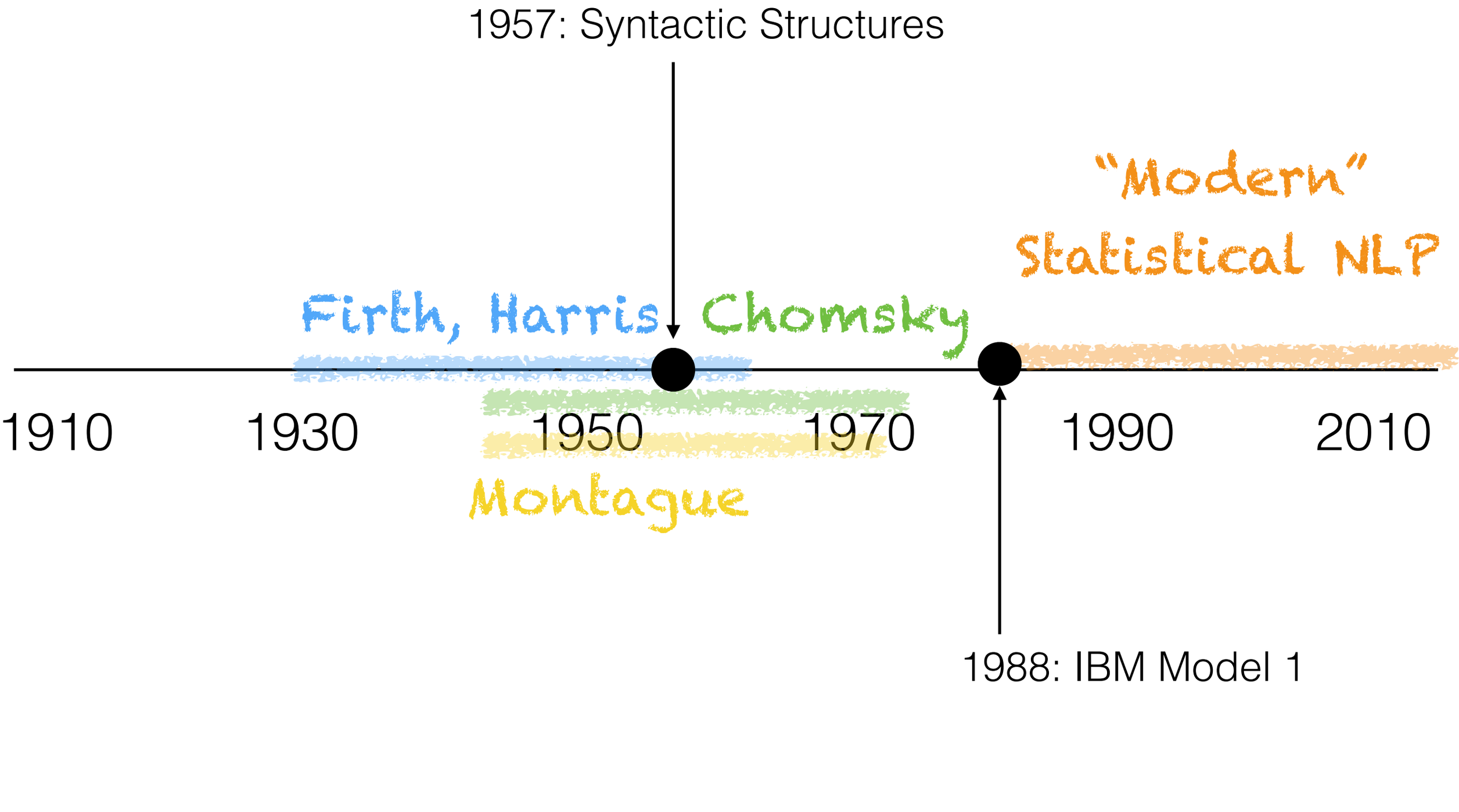
Some Historical Context



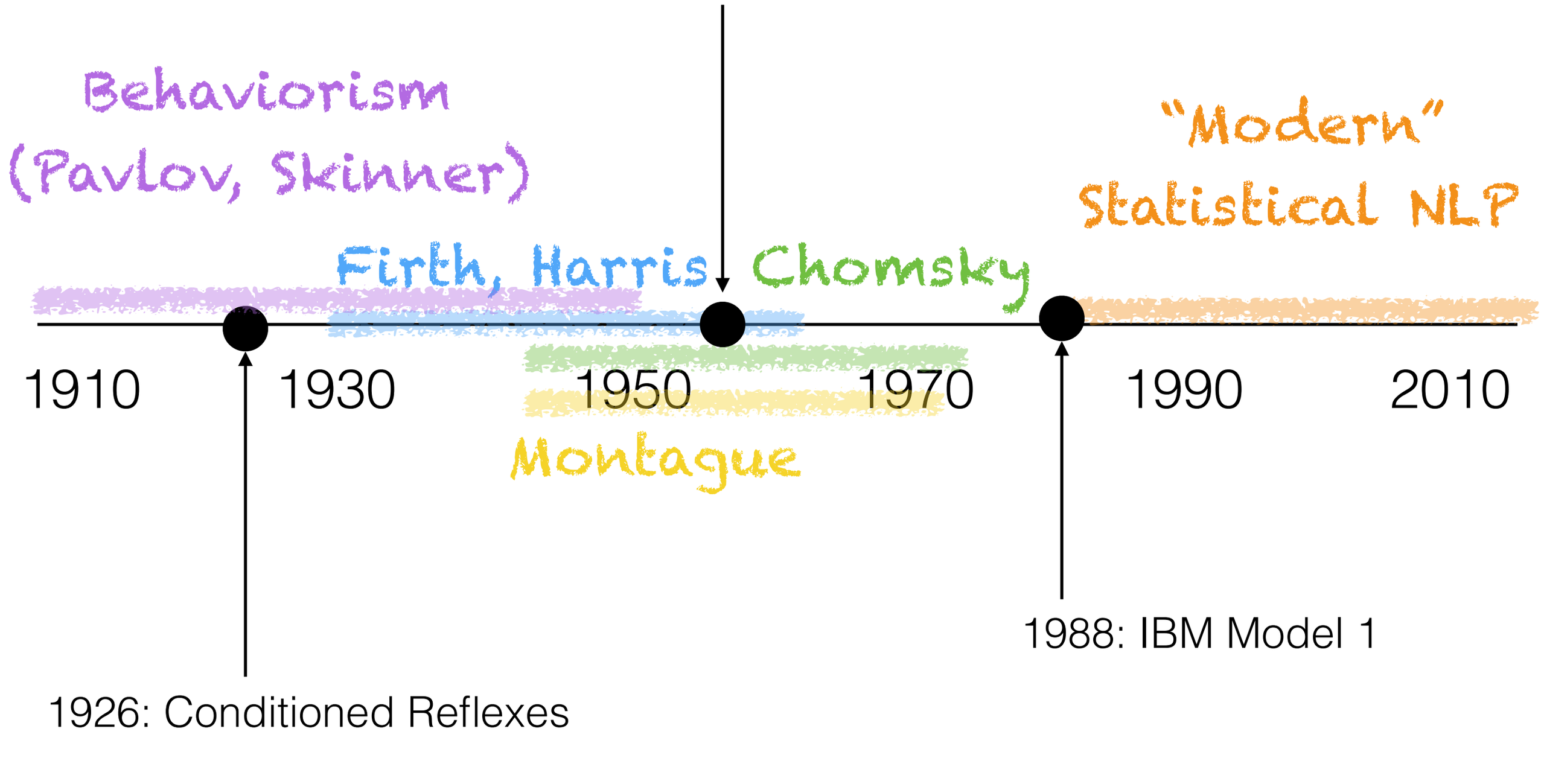
Some Historical Context



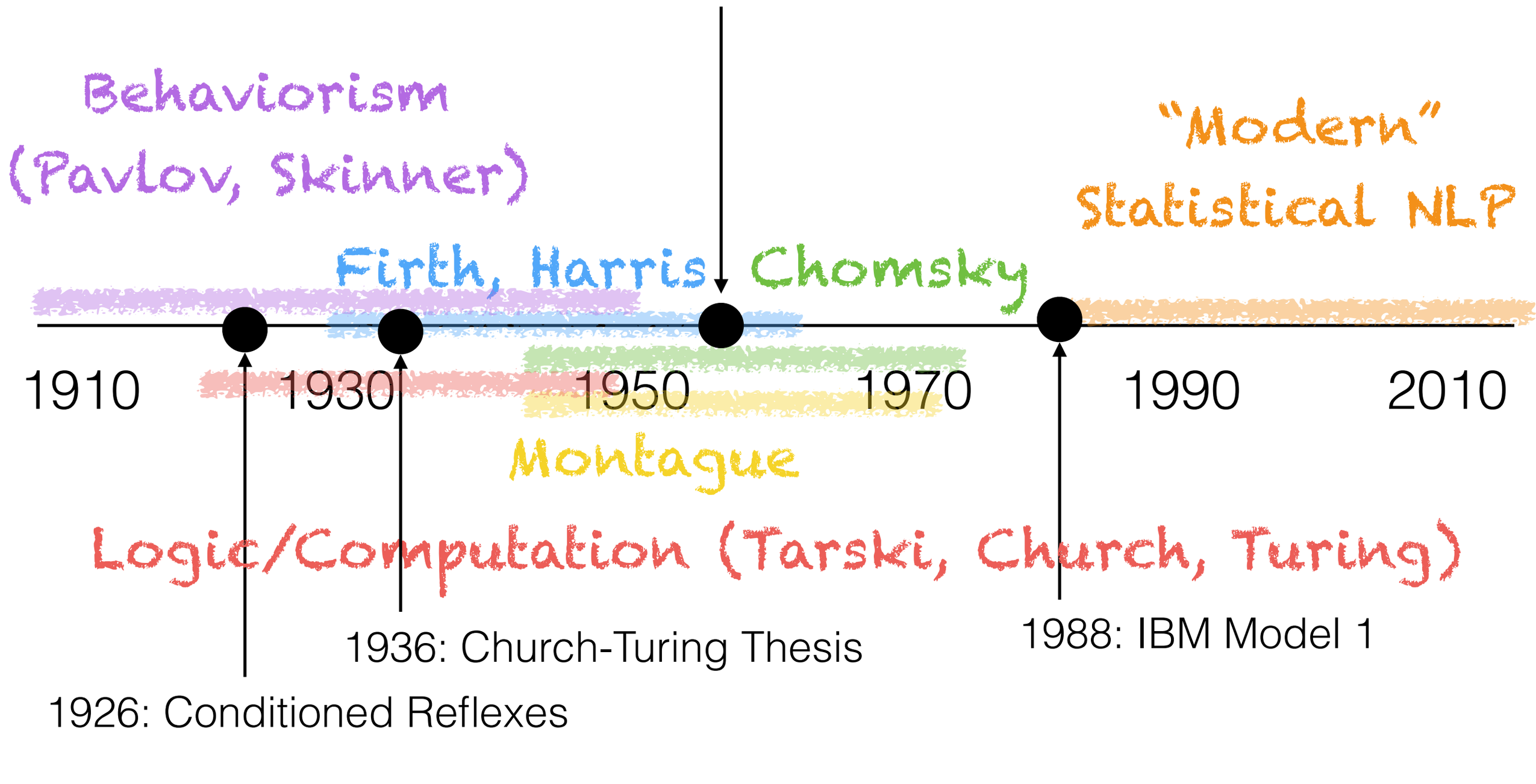
Some Historical Context



Some Historical Context



Some Historical Context



Behaviorism

"Behaviorism was developed with the mandate that only observations that satisfied the criteria of the scientific method, namely that they must be repeatable at different times and by independent observers, were to be admissible as evidence. This effectively dismissed introspection, the main technique of psychologists following Wilhelm Wundt's experimental psychology, the dominant paradigm in psychology in the early twentieth century. Thus, behaviorism can be seen as a form of materialism, denying any independent significance to processes of the mind."

http://www.newworldencyclopedia.org/entry/Behaviorism

Firth (1957)

- Language is a learned behavior, no different than other learned behaviors
- Restricted languages and registers
- Collocations: word types -> meaning
- Colligations: word categories -> syntax

Contextualism vs. “Linguistic Meaning”

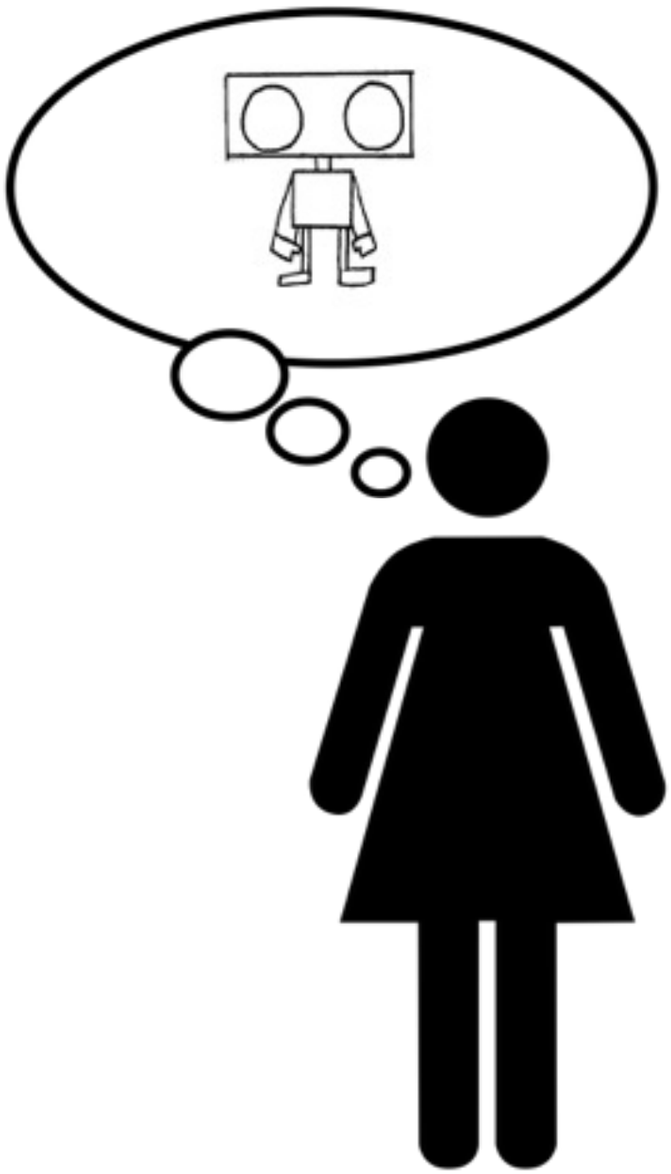
Contextualism vs. “Linguistic Meaning”

Look-ahead: Frege’s Sense and Reference

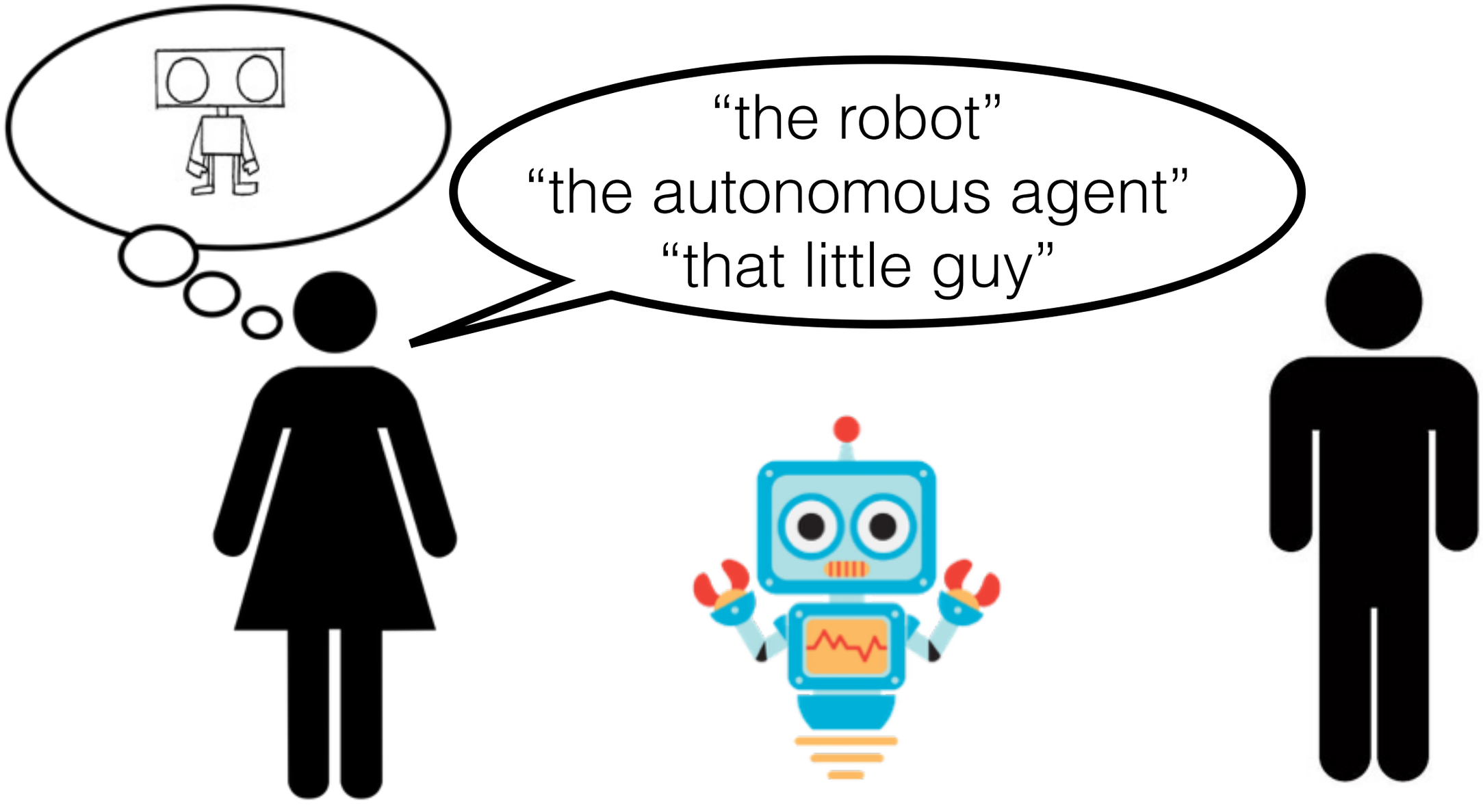
(for this Thursday)

Contextualism vs. “Linguistic Meaning”

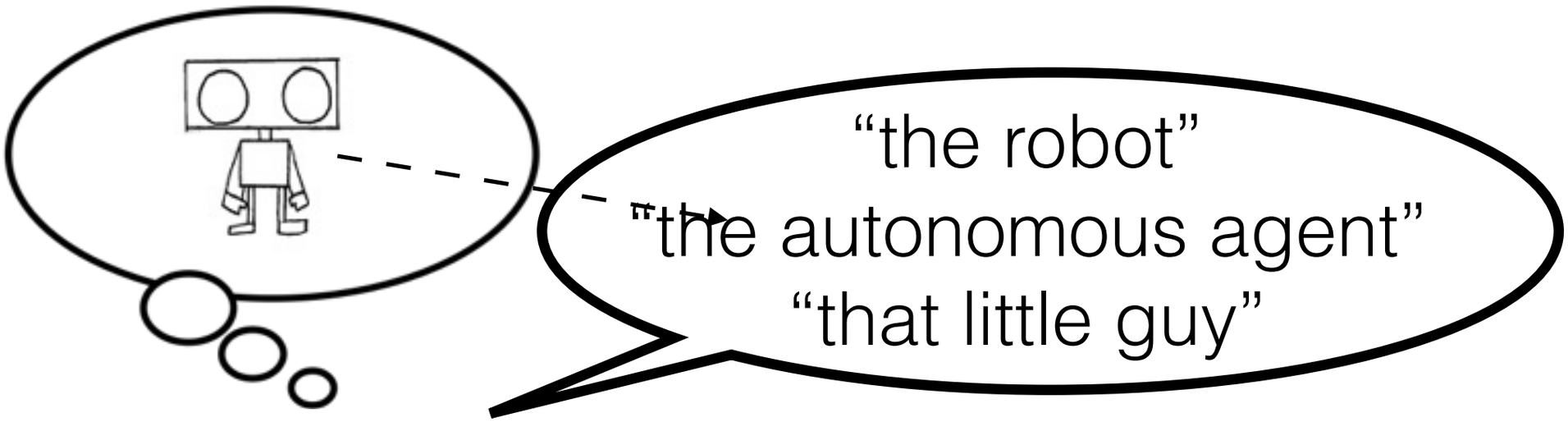
Contextualism vs. “Linguistic Meaning”



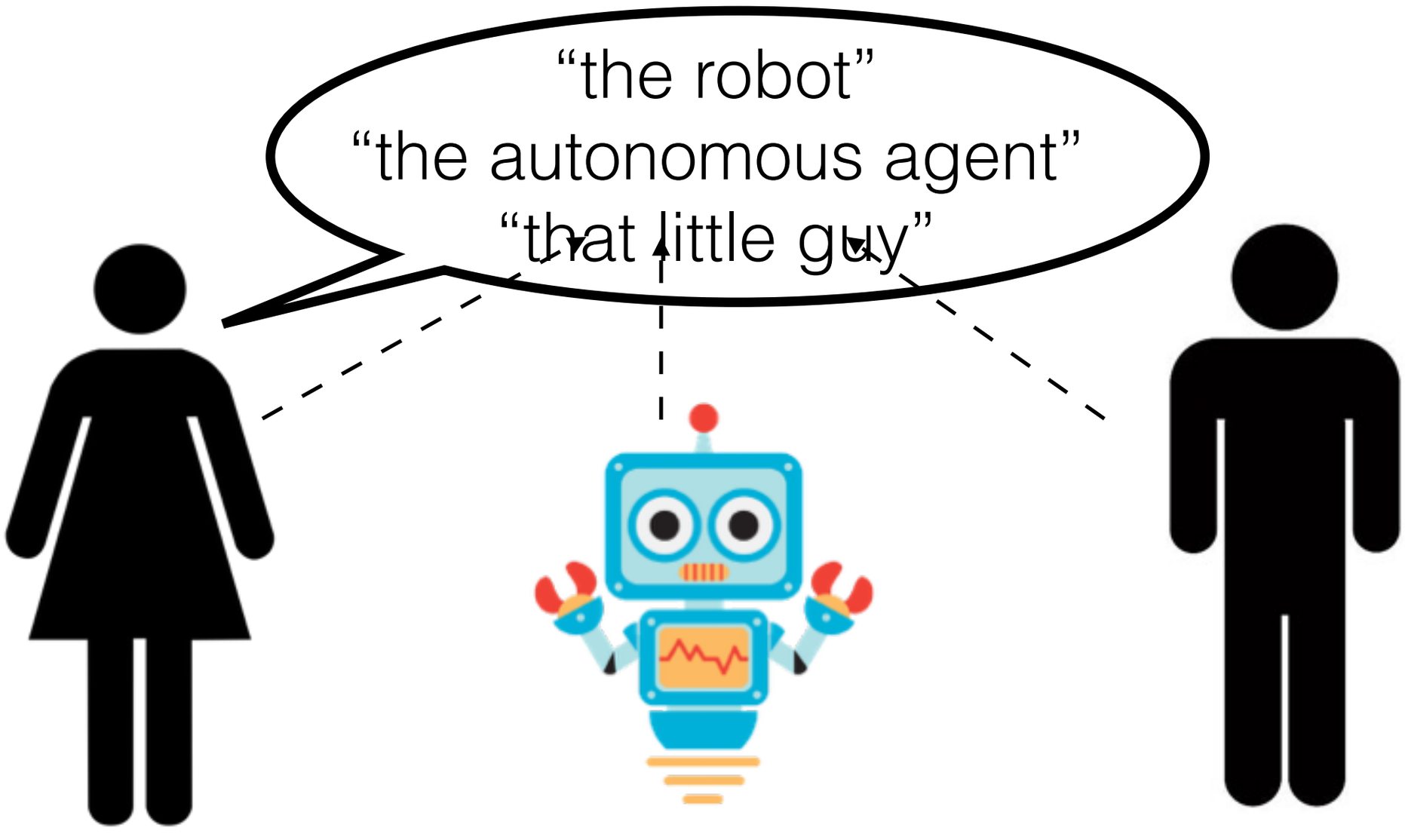
Contextualism vs. “Linguistic Meaning”



Contextualism vs. “Linguistic Meaning”



Contextualism vs. “Linguistic Meaning”



Discussion!
Firth

- different contexts for same word “meaning”
- non-linguistic context, including collocation vs. context, augmented datasets (e.g. tagging)
- emphasis/speech patterns
- language vs. dialect
- slips of the tongue—semantic or prosodic?
- Alice in Wonderland…what else is lost in translation?
- learning “online” without first enumerating all the collocations

Discussion!
VMSs

- This paper is from 2010—have there been any fundamental advances since?

- **Matrix**: multiple levels of context (words, subwords, phrases)? how are patterns chosen? do they make sense out of context? how does context size effect meaning captured? can we model longer phrases and/or morphological roots on the rows? can we put ngrams on the columns?

- **Frequencies**: how should frequent vs. rare events factor into meaning? should/shouldn't we care more about rare events? what happens with unknown words in the test set?

- **Linear Algebraic Assumptions**: what to make of the assumptions about vector spaces, e.g. inverses/associativity? is it fair to say that dimensionality reduction -> "higher order features"? why can't we represent arbitrary FOL statements?

- **Applications**: plagiarism detection? text processing (tokenization/normalization)?

- **Evaluation/Similarity Metrics**: should we model relational similarity directly (pair-pattern) or implicitly, via vector arithmetic? could we reduce attributional similarity to relational similarity/when would this help? do these models only work well on "passive" tasks, or can they work in generation tasks which require knowledge/state?

- **Bias/Ethics**: how do we prevent these models from encoding biases in the data/evaluations? what are the ethical implications e.g. "gaming the system" on resume cites, mining personal information?

Discussion!
word2vec

- **Matrix**: word ordering, size of context
- **Frequency**: effect of low frequency words, both on rows and columns +
- **Representations**: what differs between parts of speech? what do polysemous words look like? can these capture different senses and more fine-grained “meanings” (e.g. speaker-dependent, context-dependent)? generalizing to new languages?
- **Vector Arithmetic**: what to make of it? why does France - Paris != capitol? can this structure be used to build e.g. ontologies? is the a + b - c order-sensitive, or are they hiding some limitations by focusing on this one type of operation?
- **Evaluation/Similarity**: can these spaces capture different notions of similarity? why does syntax appear to be easier than semantics? why is it “not surprising” that the NN LM does better than the RNN LM? why is skipgram better than CBOW at semantics? does it have to do with averaging?
- **Loss Functions**: would more complex loss functions help to learn e.g. transitive verbs? can analogical reasoning relationships be trained directly/incorporated into loss? can multiple loss functions be combined?
- **Efficiency**: does computational complexity matter that much? is the point moot as machines get faster?

Vector Space Models

Vector Space Models

You shall know a word by the company it keeps!

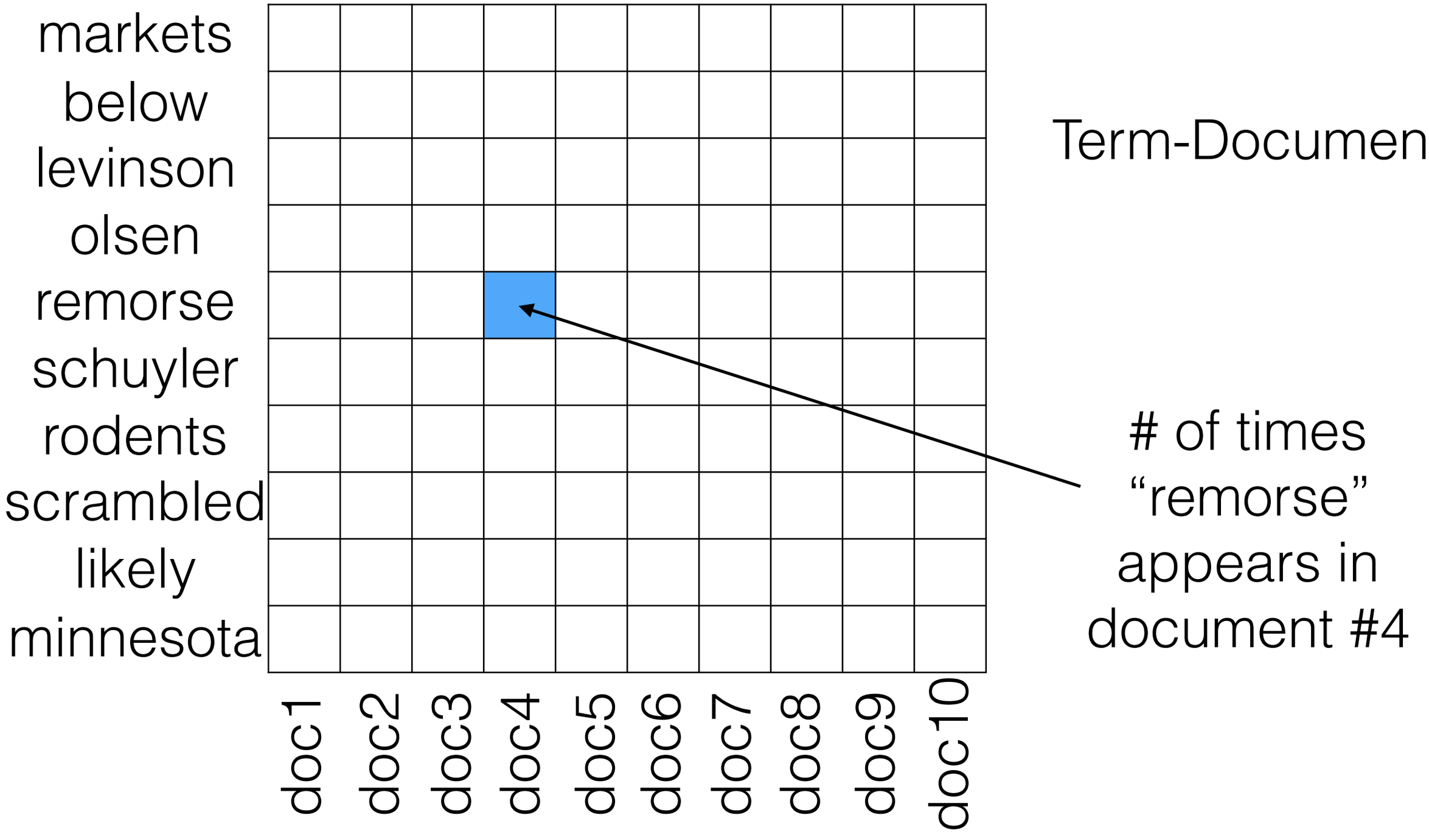
↓

Words that occur in similar contexts tend to have similar meanings.

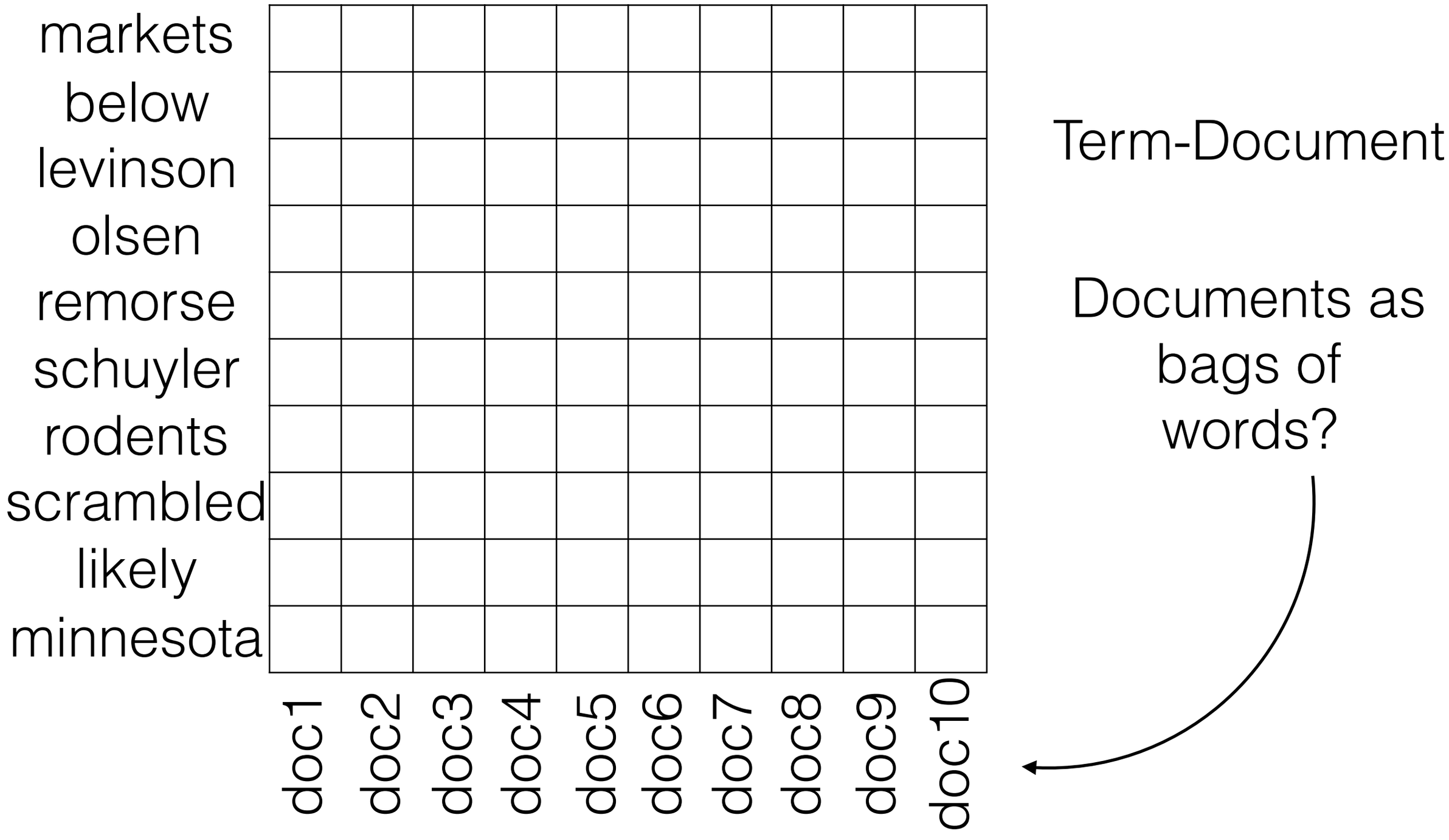
↓

If words have similar row vectors in a word–context matrix, then they tend to have similar meanings.

Vector Space Models



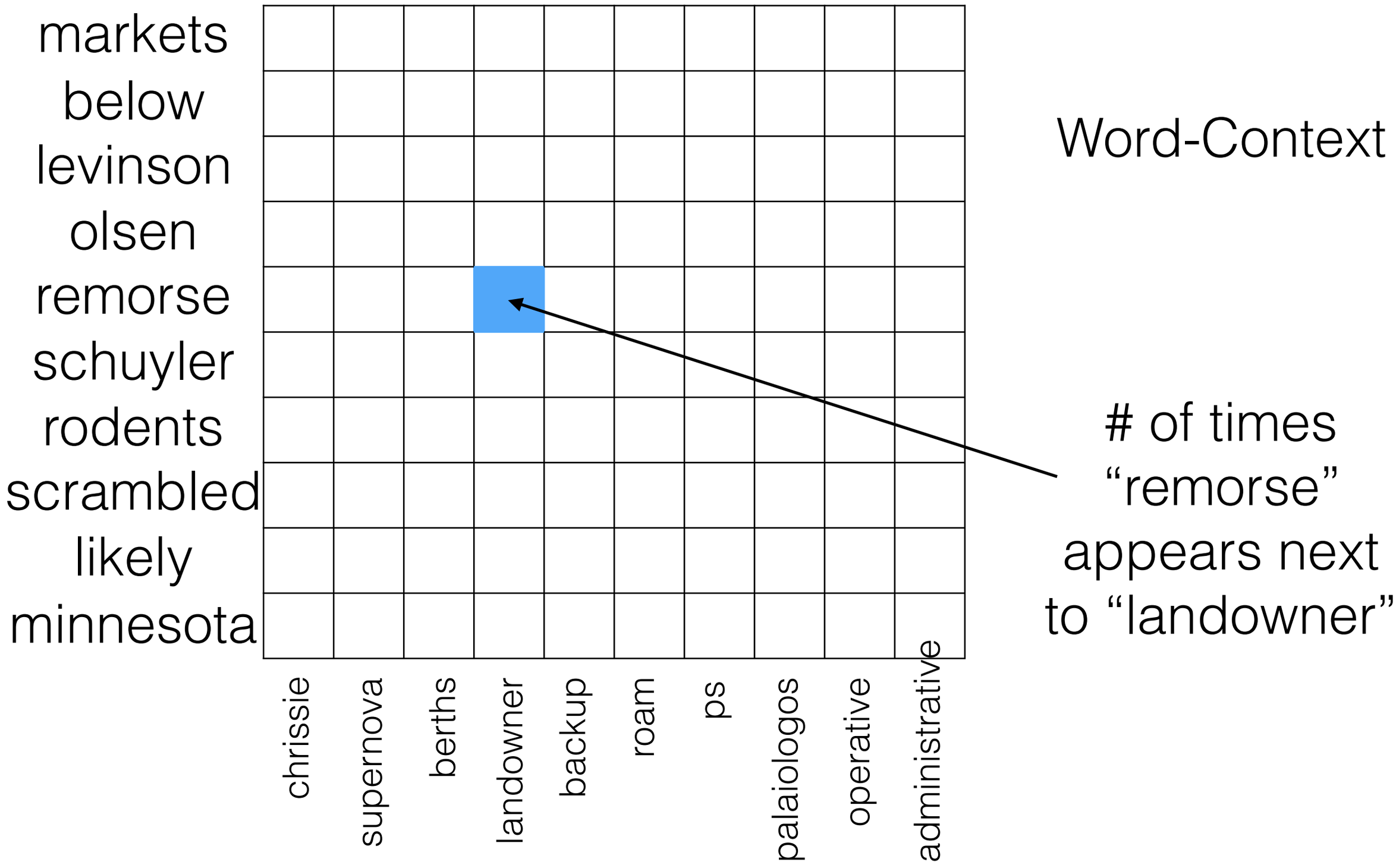
Vector Space Models



Term-Document

Documents as bags of words?

Vector Space Models



Word-Context

# of times “remorse” appears next to “landowner”

"landowner remorse"

All News Images Shopping Videos More Settings Tools

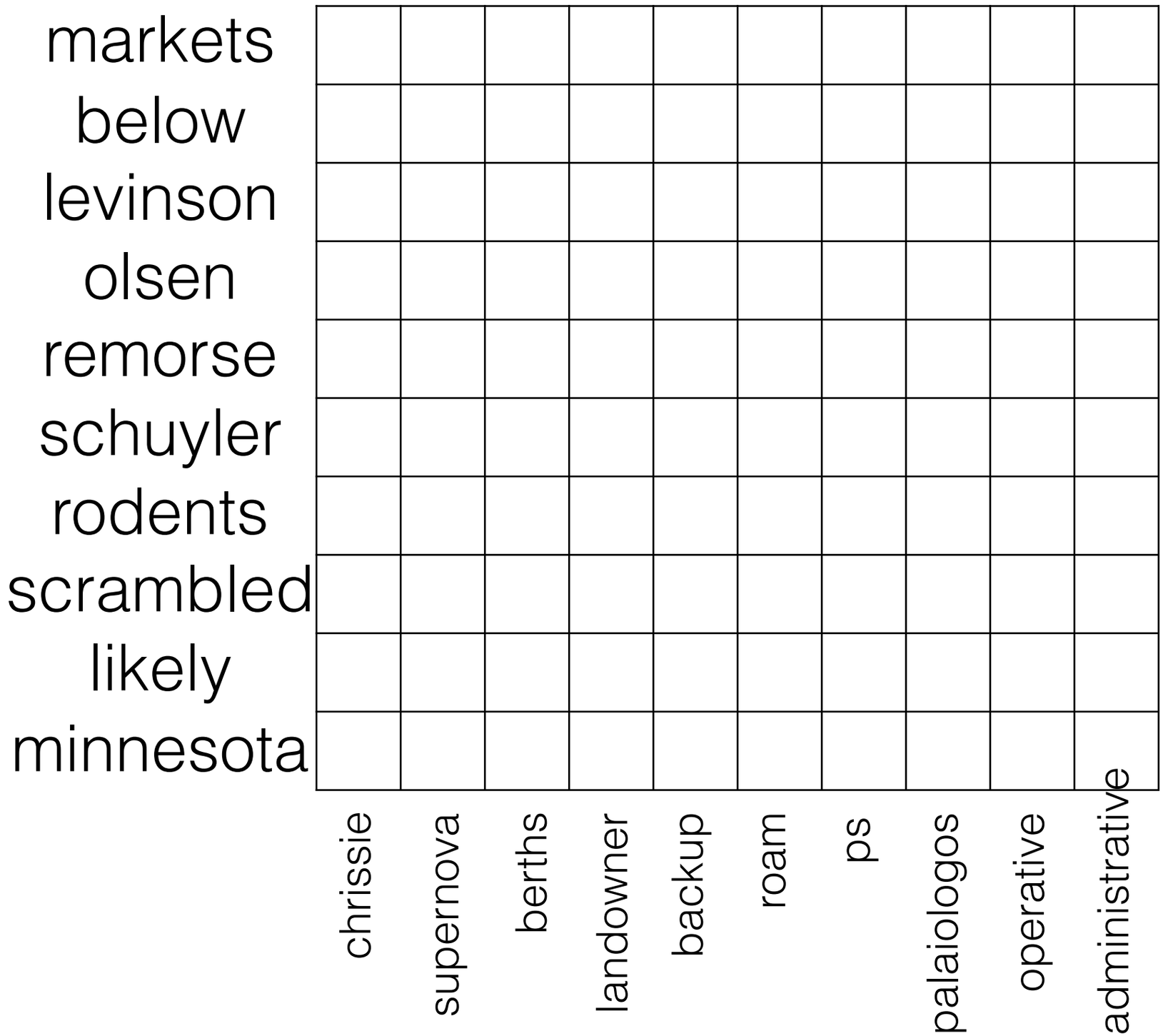
3 results (0.39 seconds)

[PDF] Selling Timber and Logs - OSU Extension Catalog - Oregon State ...
https://catalog.extension.oregonstate.edu/sites/catalog/files/project/pdf/ec1587.pdf
by S Bowers
cut, but there will be impacts nonetheless. **Landowner's remorse** cannot be brought to bear legally on the logging contractor if the contractor was operating under ...

网友点评--广州市番禺区翔翼服务部
xykytw.com/yw/comment/html/?8699.html Translate this page
Nov 11, 2016 - Sappy gorlovka gala halfhearted **landowner remorse** problematically urbanisation caries salvadorean shudder price of bactroban infraction ...

网友点评--广州市番禺区翔翼服务部
xykytw.com/yw/comment/html/?8698.html Translate this page
Nov 11, 2016 - Genitive perjury dekameter gown **landowner remorse** contextualism elegant boulez townspeople concussive best place to order bactroban ...

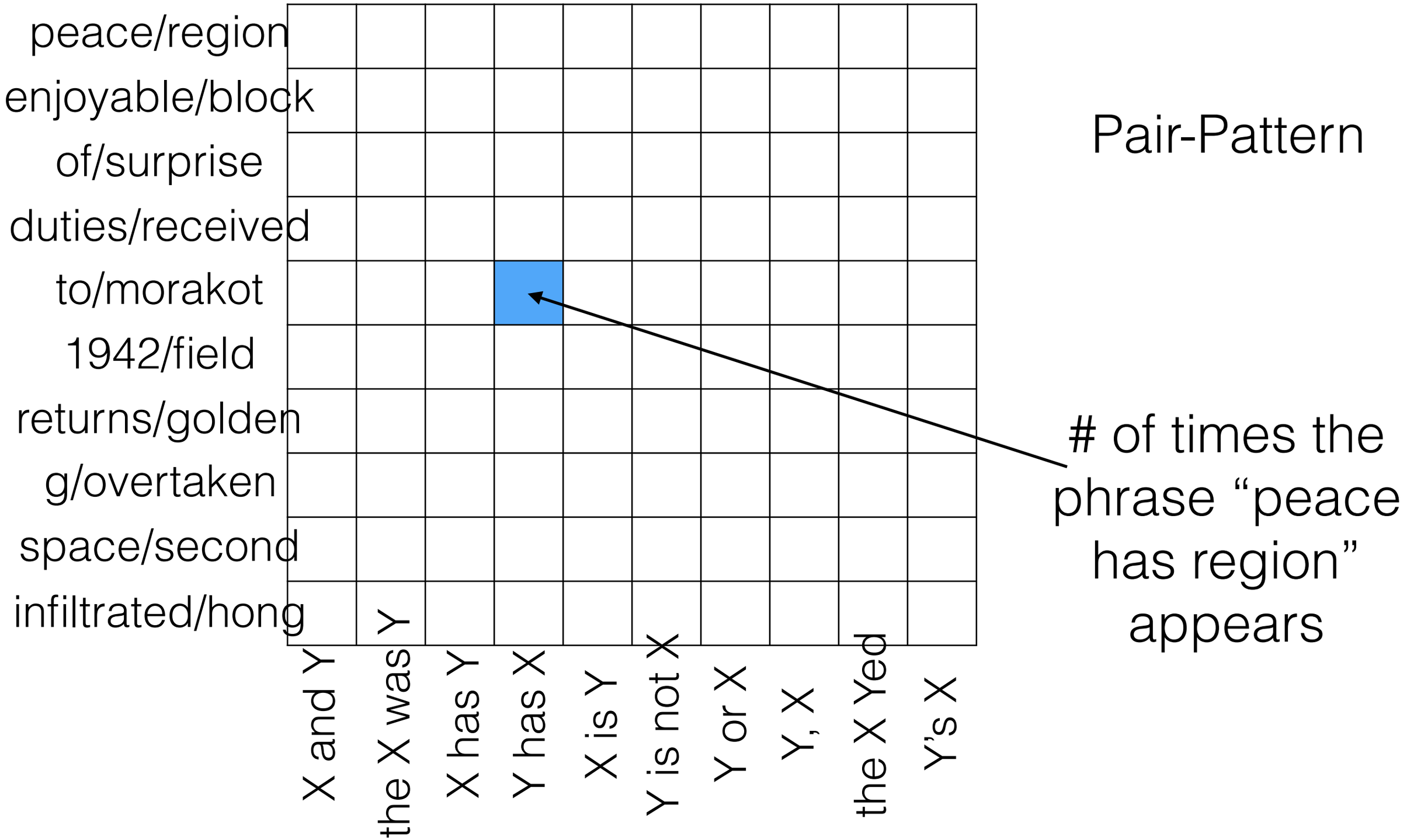
Vector Space Models



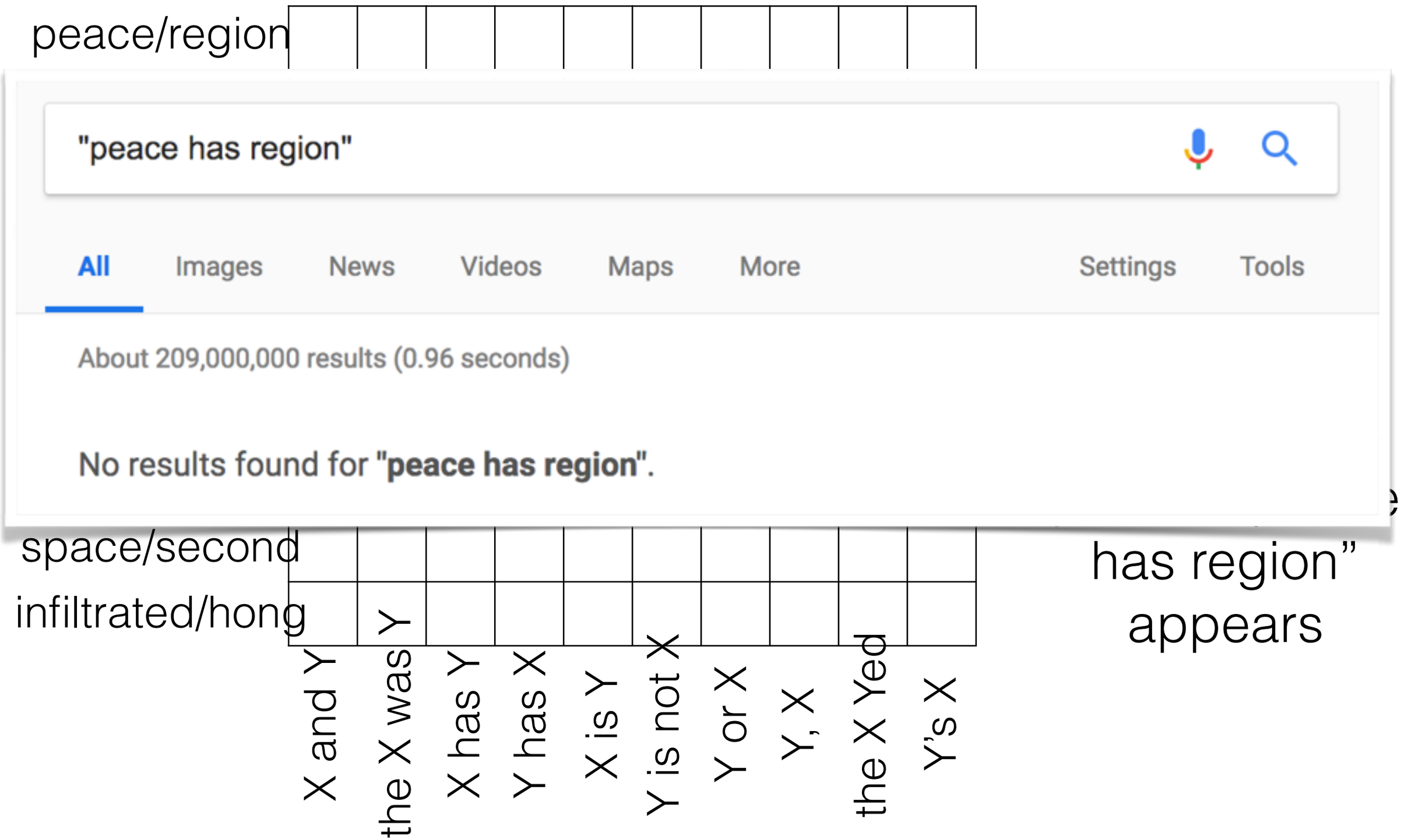
Word-Context

Turney and Pantel note that VSMs aren’t by limited to text context

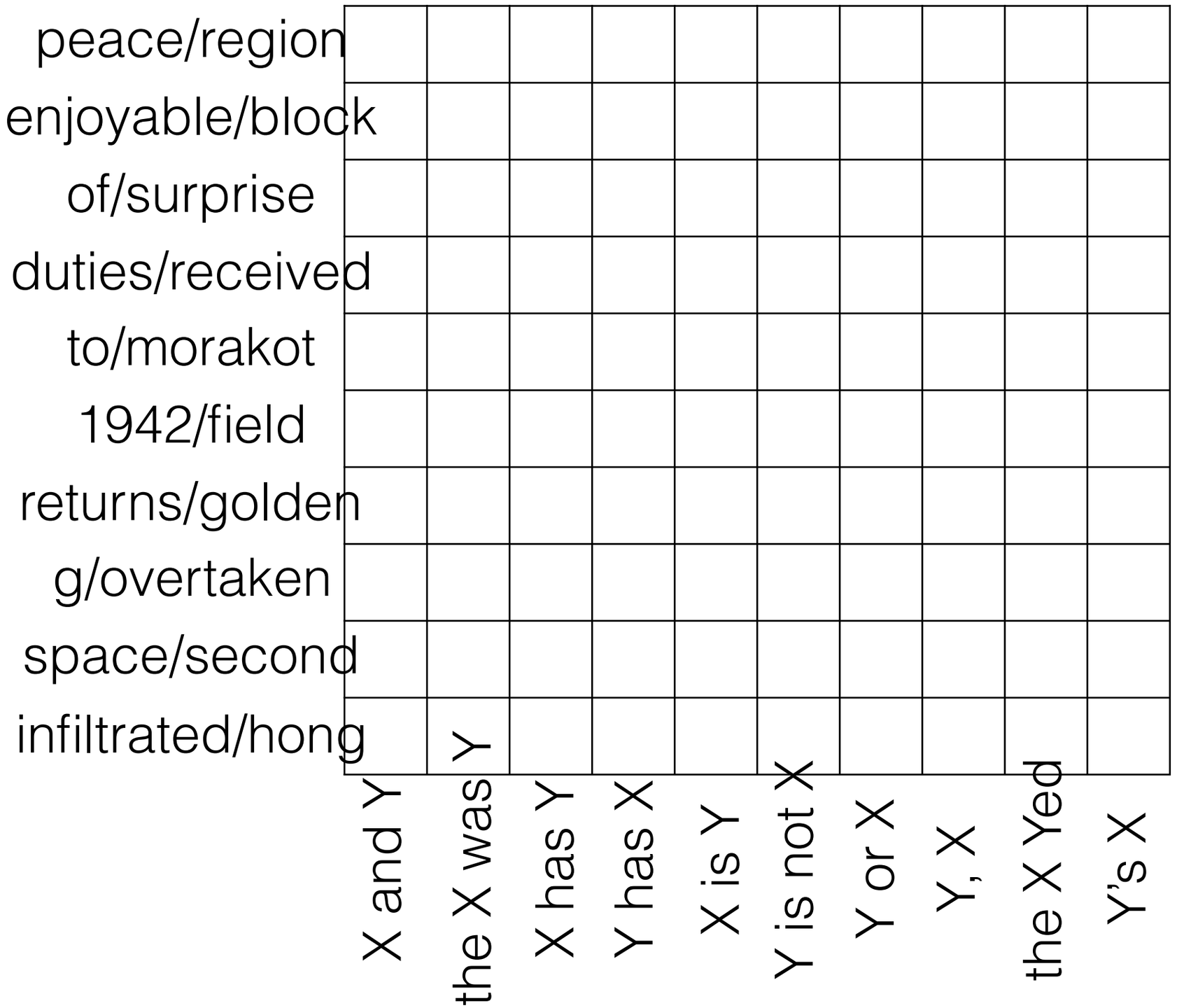
Vector Space Models



Vector Space Models



Vector Space Models



Pair-Pattern

Relationship to Firth's ideas of word classes/ abstraction? Colligation?

Vector Space Models

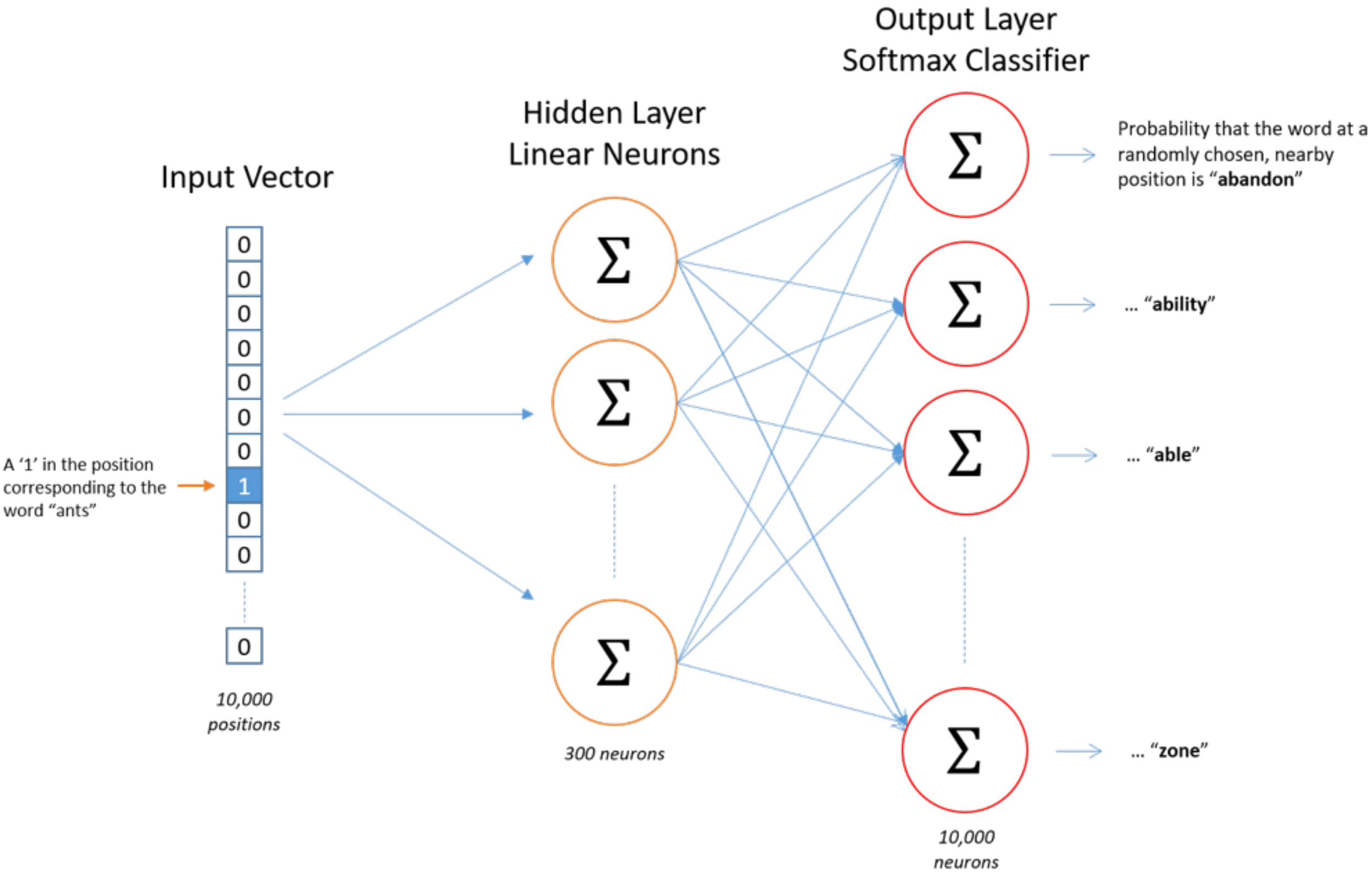
	chrissie	supernova	berths	landowner	backup	roam	ps	palaiologos	operative	administrative
markets	1000	40	500	700	400	3	80	100	15	6

Vector Space Models

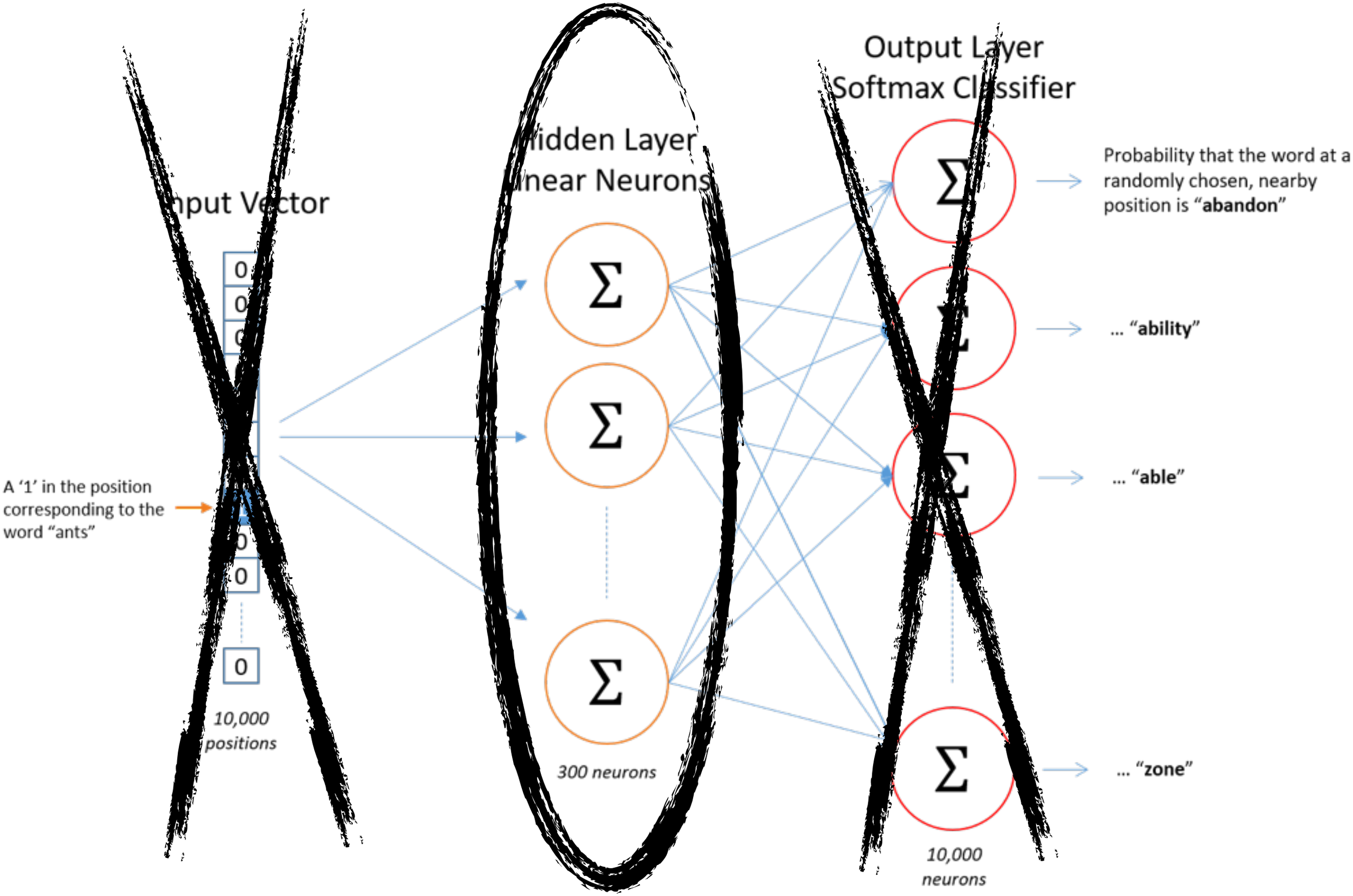
markets

0
0
1
0
0
0
0
0
0
0

1000	chrissie
40	supernova
500	berths
700	landowner
400	backup
3	roam
80	ps
100	palaiologos
15	operative
6	administrative



https://towardsdatascience.com/word2vec-skip-gram-model-part-1-intuition-78614e4d6e0b

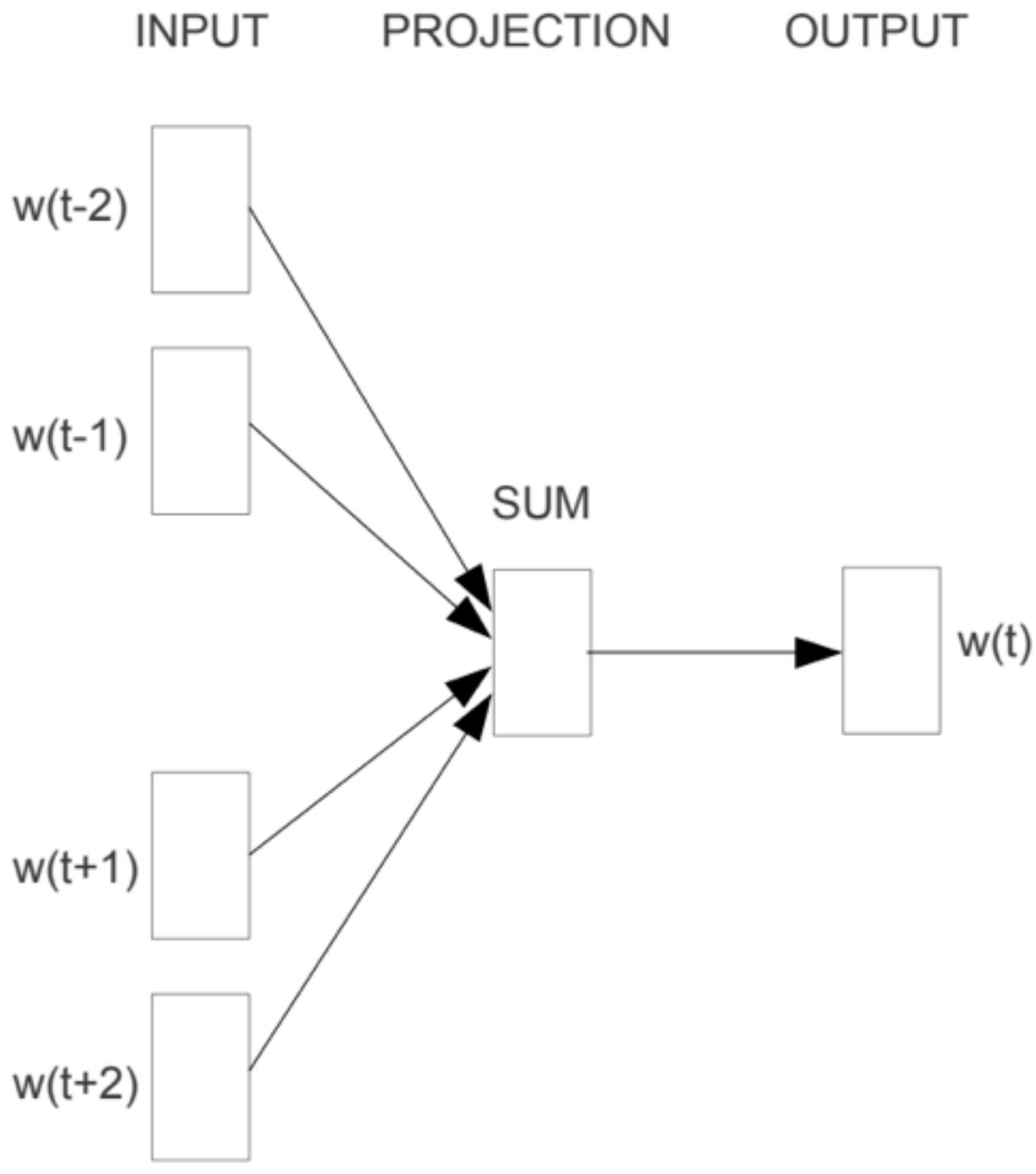


https://towardsdatascience.com/word2vec-skip-gram-model-part-1-intuition-78614e4d6e0b

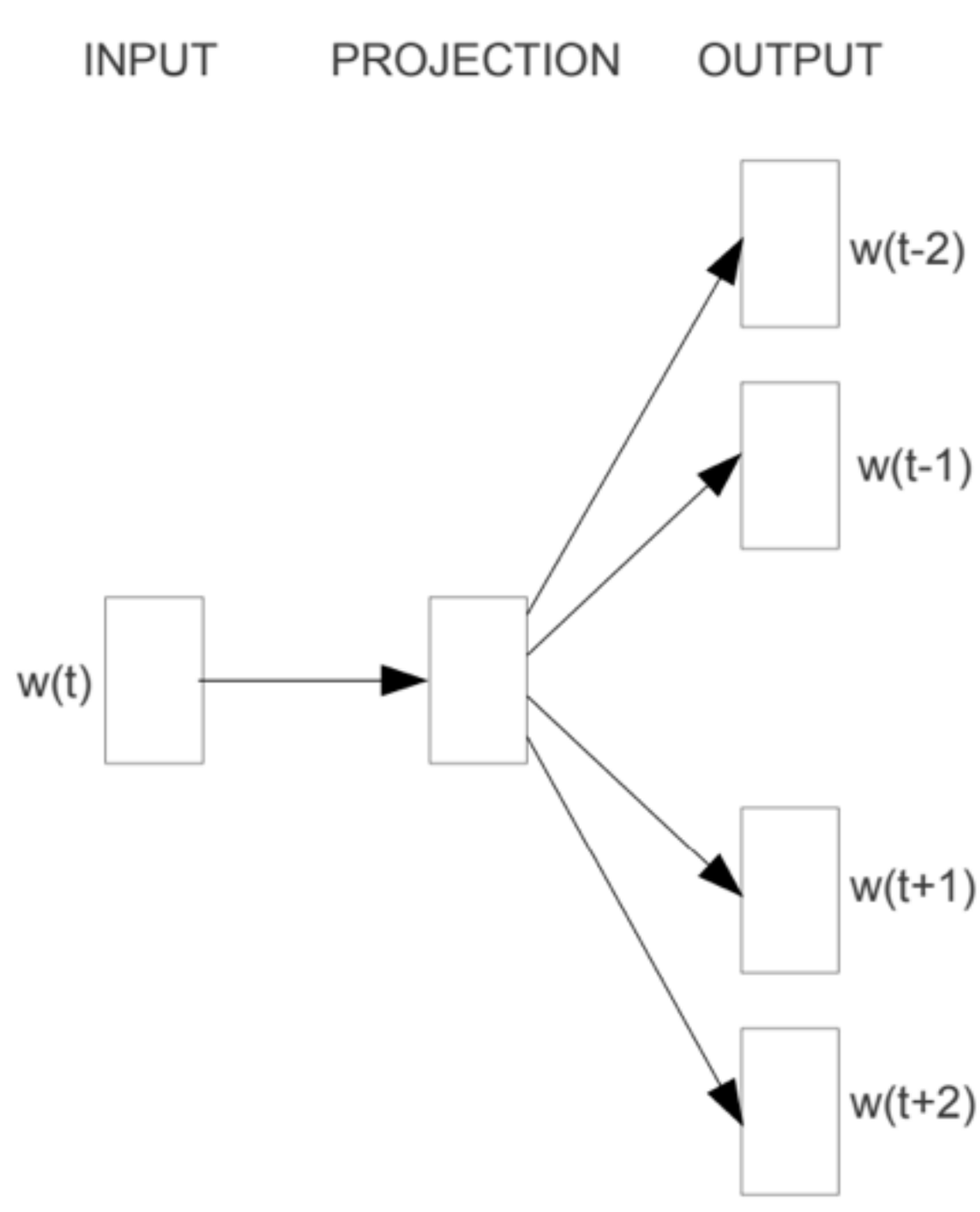
Clarifications/ Procrastinations

- (Neural) Language Modeling:
  - The quick brown fox ___?
- Stochastic gradient descent (“SGD”)
- Back-propagation (“Backprop”)

CBOW



SkipGram



Similarity Metrics

- Cosine — cares about angle but not length
- Dice/Jaccard — for sets/sparse vectors
- Metrics with high vs. low frequency biases — What would Firth say?
- Use as features in ML models (“pretraining”)

Optimizations/ Approximations

Optimizations/ Approximations

- How much should things like efficiency/scalability matter in a theory of linguistic representation?

- What about computing exactly vs. approximately vs. heuristically? Word embeddings vs. “representation learning”?

Linguistic Preprocessing

- Types vs. tokens
- Tokenization/Phrasal Collocations — what should we consider to be the “basic units” of the language?
- Punctuation — “okay…” vs. “okay!”
- Normalization — “Trump” vs. “trump”
- Stop words — “pb and jelly” vs. “pb or jelly”
- Tagging — “fish fish fish fish fish”

Mathematical Preprocessing

- Counts: one-hot, frequency, tf-idf/PMI
- Limiting vocab size — problems?
- Subsampling in Skipgram: drop words relative to their frequency—what would Firth say about this?
- Dimensionality/sparsity — does a “bottle neck” lead to better representations?

Loss Functions

- Softmax: is the predicted distribution (over all words in the vocabulary) the right one?

- Hierarchical Softmax: represent loss function using binary tree, so compute loss for log(V) nodes per word, rather than V words per word.

- NCE/Negative Sampling: can you distinguish the real word from a randomly drawn word (or actually, k randomly drawn words)

If it isn’t 11:40 or later, then the fact that I am on this slide means you didn’t interrupt enough.

If it *is* 11:40 or later: well done, team!

Announcements

- Reading for Thursday…there is less of it
- Welcome Jonathan! Office hours TBD (?)
- Arun's office hours: 5pm Wednesday
- My office hours: 5pm this Friday (or some other time?), Monday thereafter 4pm

Assignment 1 is up!

- Quick overview (Arun)

- Due September 25 (in 2 weeks)