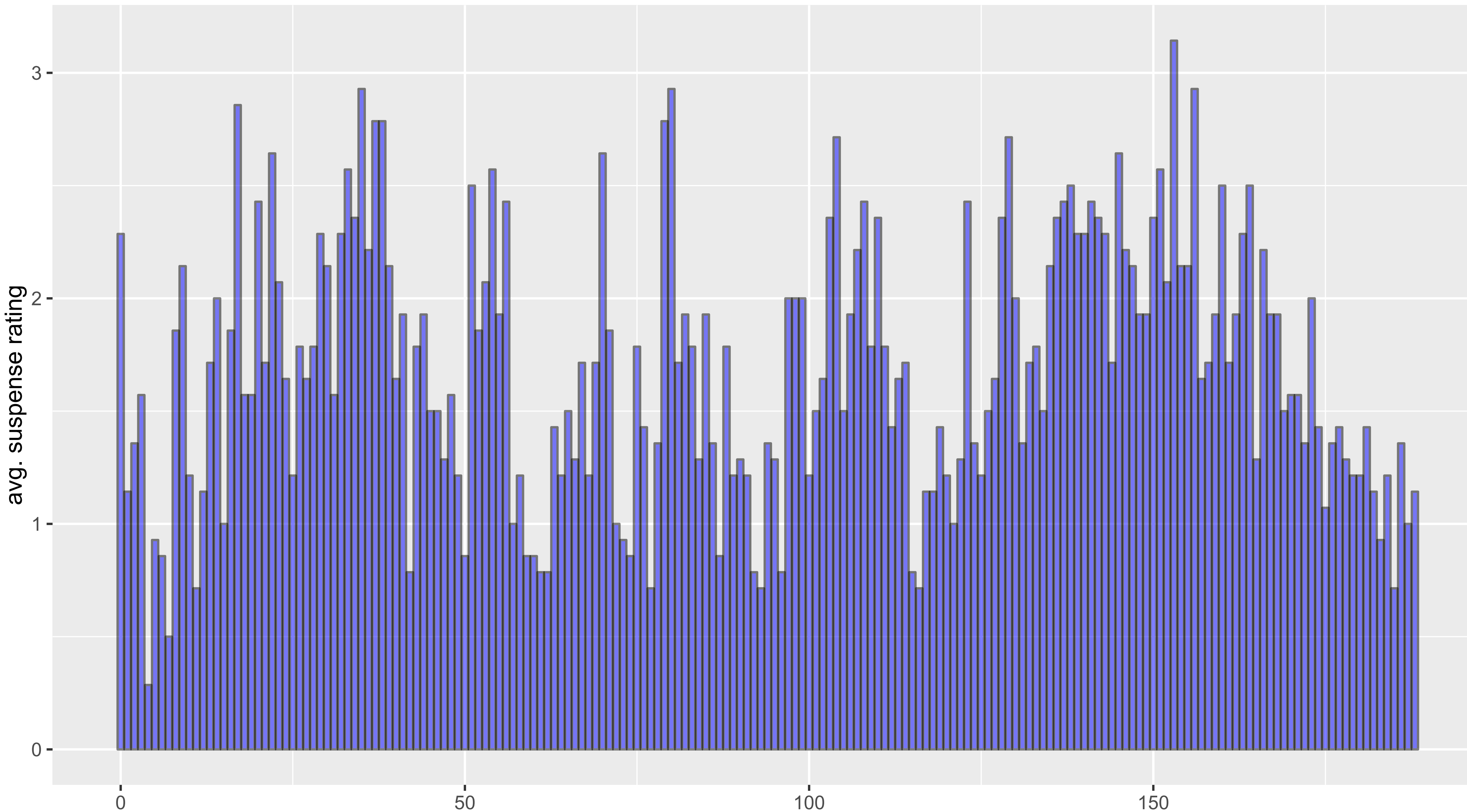
Applied Natural Language Processing

Info 256

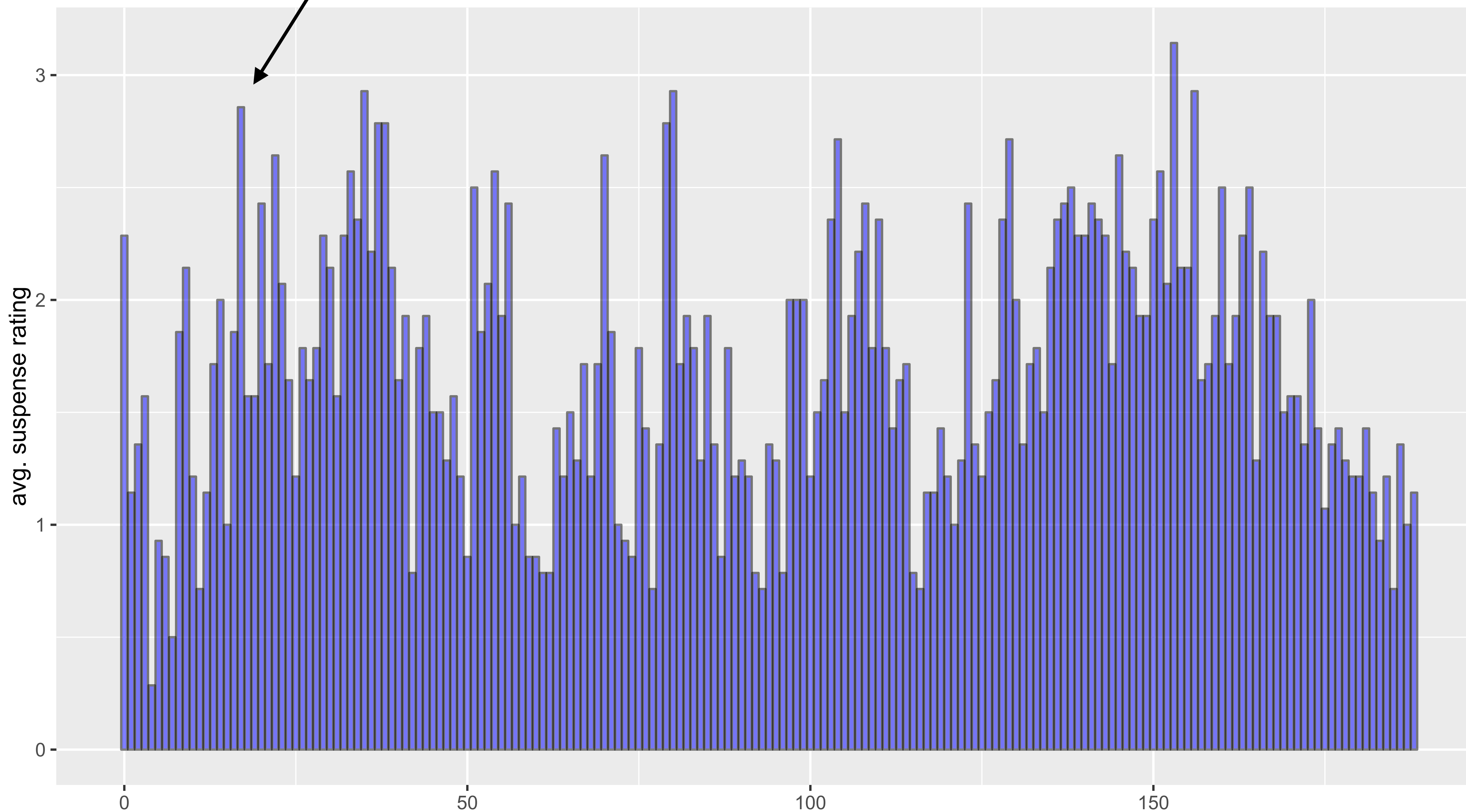
Lecture 17: WordNet 2 (March 21, 2019)

David Bamman, UC Berkeley

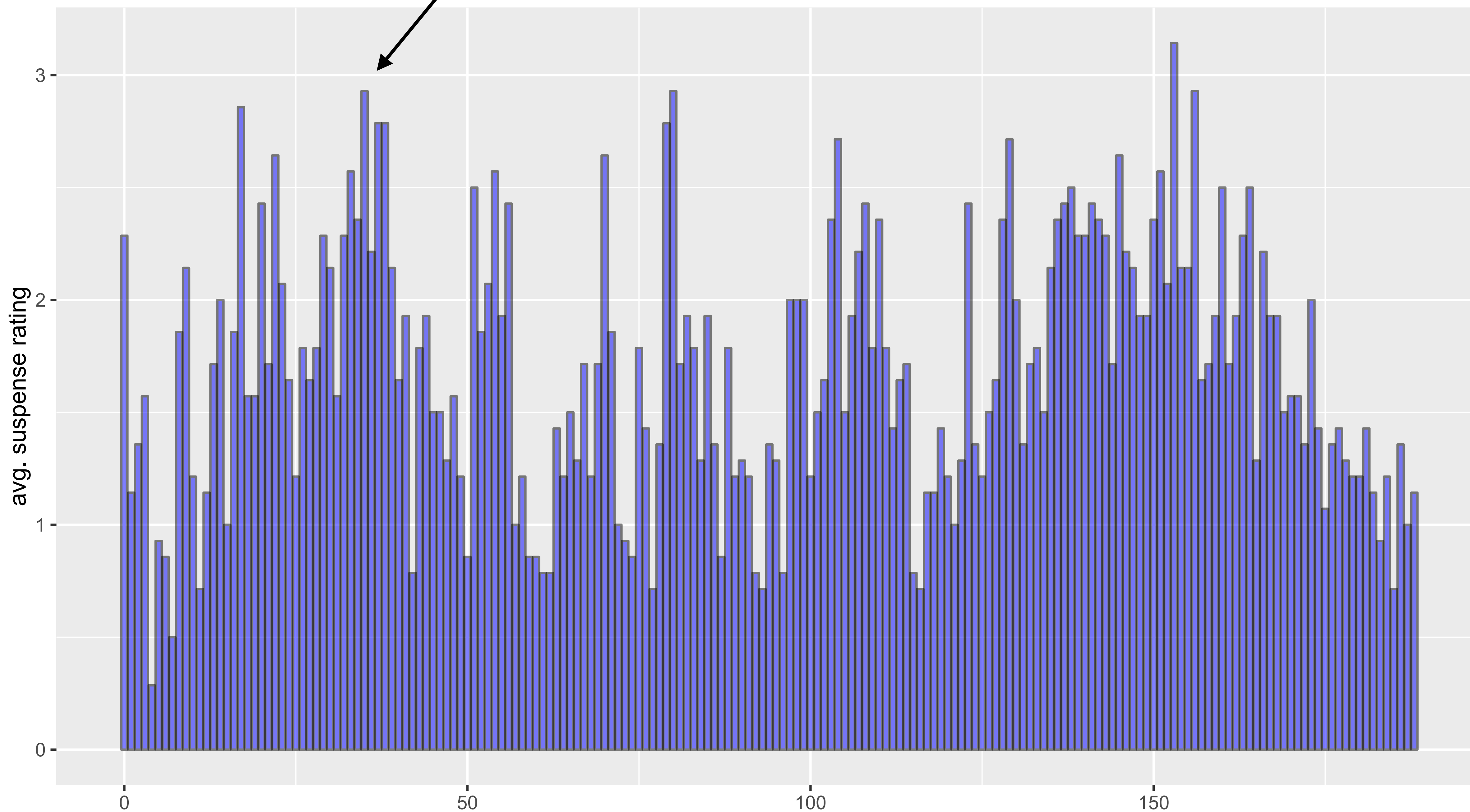
Suspense



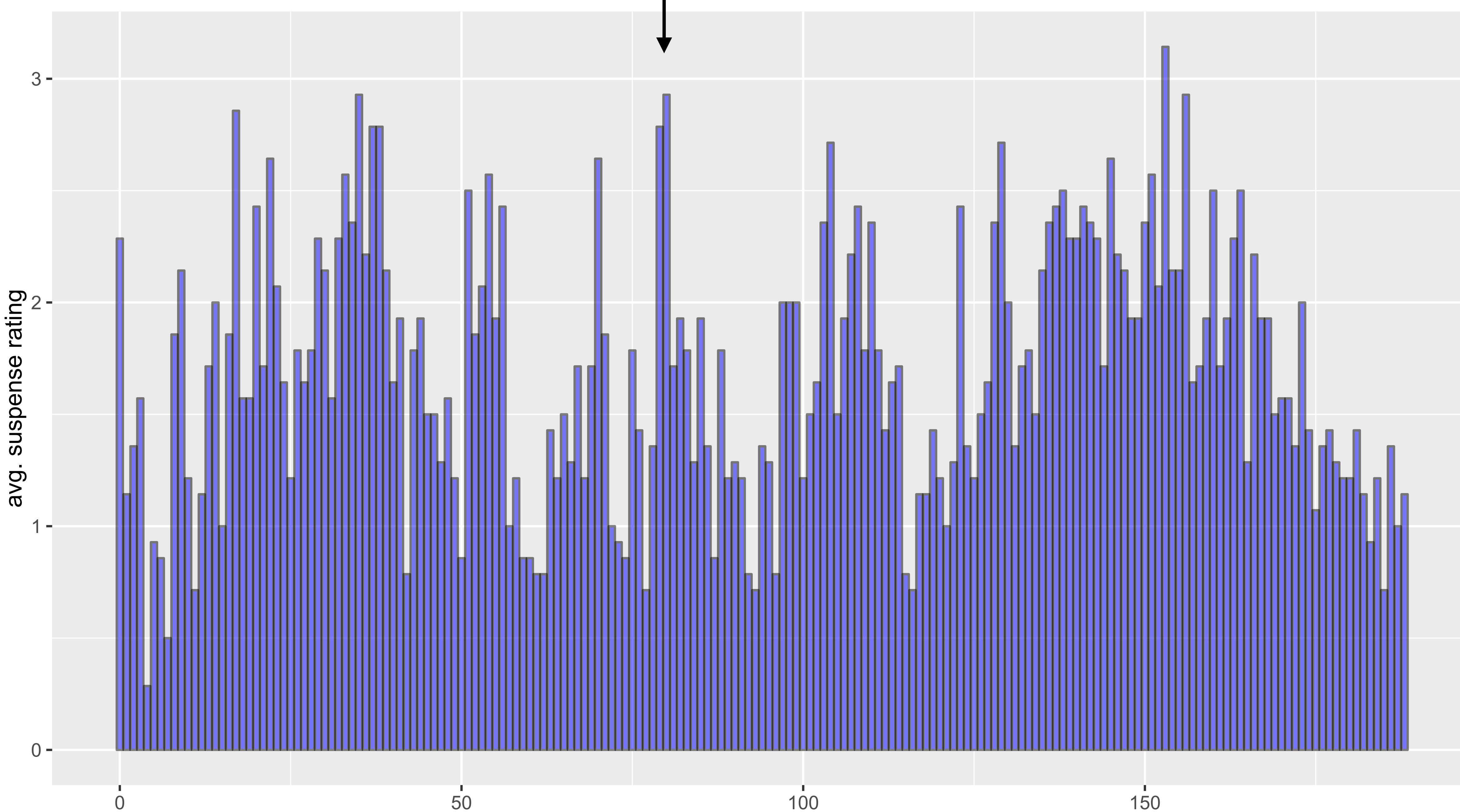
But the yawning policeman saw the thing, the busy crowds in the markets stopped agape, workmen going to their work betimes, milkmen, the drivers of news-carts, dissipation going home jaded and pale, homeless wanderers, sentinels on their beats, and in the country, labourers trudging afield, poachers slinking home, all over the dusky quickening country it could be seen--and out at sea by seamen watching for the day--a great white star, come suddenly into the westward sky!



And voice after voice repeated, "It is nearer," and the clicking telegraph took that up, and it trembled along telephone wires, and in a thousand cities grimy compositors fingered the type.

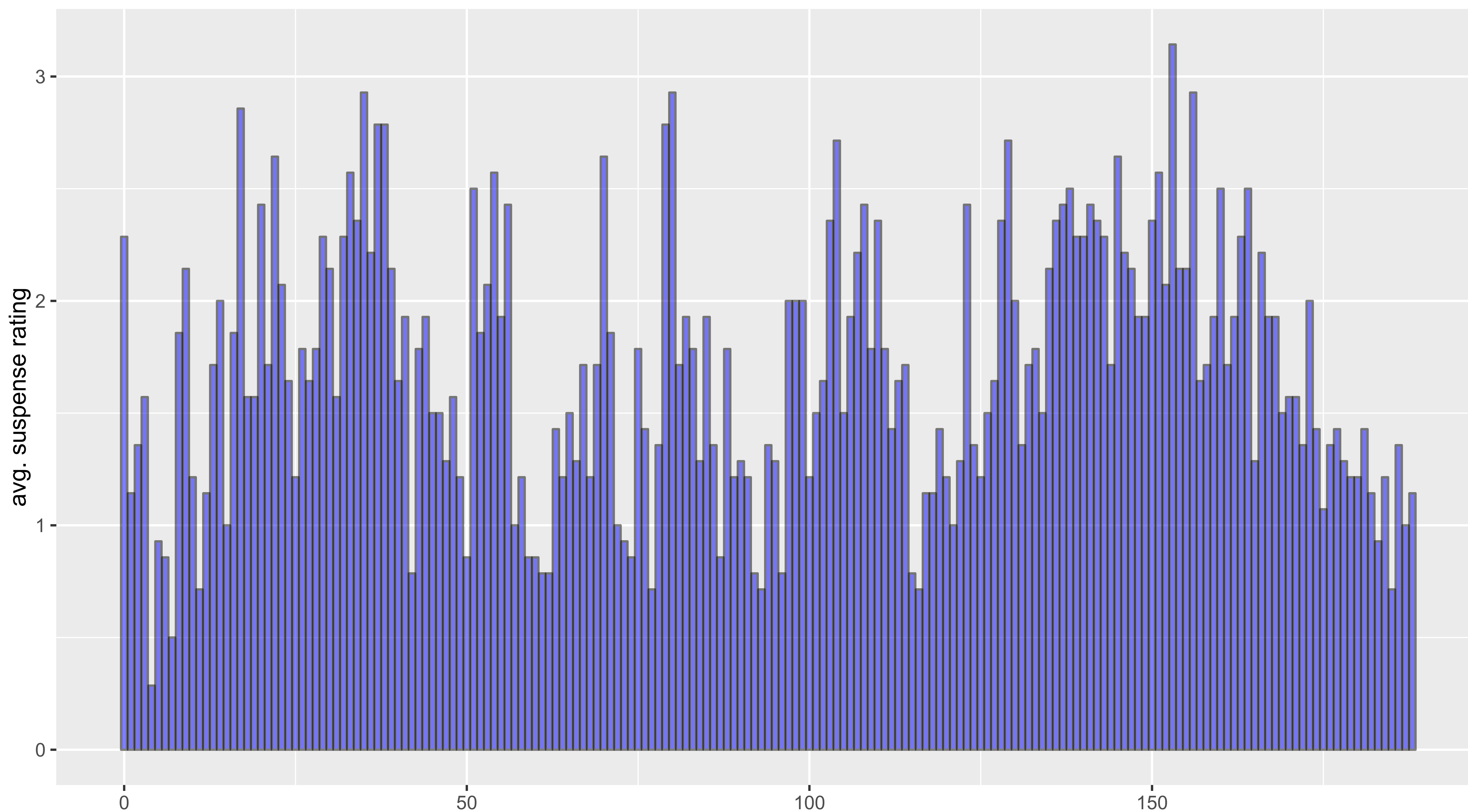


"It would seem, gentlemen, if I may put the thing clearly and briefly, that--Man has lived in vain."



Larger grew the star, and larger, hotter, and brighter with a terrible swiftness now. The tropical ocean had lost its phosphorescence, and the whirling steam rose in ghostly wreaths from the black waves that plunged incessantly, speckled with storm-tossed ships.

And then came a wonder.



Activity

- 9.annotation/IAAMetrics.ipynb

WordNet

- WordNet encodes human-judged measures of similarity.

- S: (v) **serve**, function (serve a purpose, role, or function) *"The tree stump serves as a table"; "The female students served as a control group"; "This table would serve very well"; "His freedom served him well"; "The table functions as a desk"*
- S: (v) **serve** (do duty or hold offices; serve in a specific function) *"He served as head of the department for three years"; "She served in Congress for two terms"*
- S: (v) **serve** (contribute or conduce to) *"The scandal served to increase his popularity"*
- S: (v) service, **serve** (be used by; as of a utility) *"The sewage plant served the neighboring communities"; "The garage served to shelter his horses"*
- S: (v) **serve**, help (help to some food; help with food or drink) *"I served him three times, and after that he helped himself"*
- S: (v) **serve**, serve up, dish out, dish up, dish (provide (usually but not necessarily food)) *"We serve meals for the homeless"; "She dished out the soup at 8 P.M."; "The entertainers served up a lively show"*
- S: (v) **serve** (devote (part of) one's life or efforts to, as of countries, institutions, or ideas) *"She served the art of music"; "He served the church"; "serve the country"*
- S: (v) **serve**, serve well (promote, benefit, or be useful or beneficial to) *"Art serves commerce"; "Their interests are served"; "The lake serves recreation"; "The President's wisdom has served the country well"*
- S: (v) **serve**, do (spend time in prison or in a labor camp) *"He did six years for embezzlement"*
- S: (v) **serve**, attend to, wait on, attend, assist (work for or be a servant to) *"May I serve you?"; "She attends the old lady in the wheelchair"; "Can you wait on our table, please?"; "Is a salesperson assisting you?"; "The minister served the King for many years"*
- S: (v) **serve**, process, swear out (deliver a warrant or summons to someone) *"He was processed by the sheriff"*
- S: (v) suffice, do, answer, **serve** (be sufficient; be adequate, either in quality or quantity) *"A few words would answer"; "This car suits my purpose well"; "Will $100 do?"; "A `B' grade doesn't suffice to get me into medical school"; "Nothing else will serve"*
- S: (v) **serve** (do military service) *"She served in Vietnam"; "My sons never served, because they are short-sighted"*
- S: (v) **serve**, service (mate with) *"male animals serve the females for breeding purposes"*
- S: (v) **serve** (put the ball into play) *"It was Agassi's turn to serve"*

Synsets

synset	gloss
mark, grade, score	a number or letter indicating quality
scratch, scrape, scar, mark	an indication of damage
bell ringer, bull's eye, mark, home run	something that exactly succeeds in achieving its goal
chump, fool, gull, mark, patsy, fall guy, sucker, soft touch, mug	a person who is gullible and easy to take advantage of
mark, stigma, brand, stain	a symbol of disgrace or infamy

Synsets

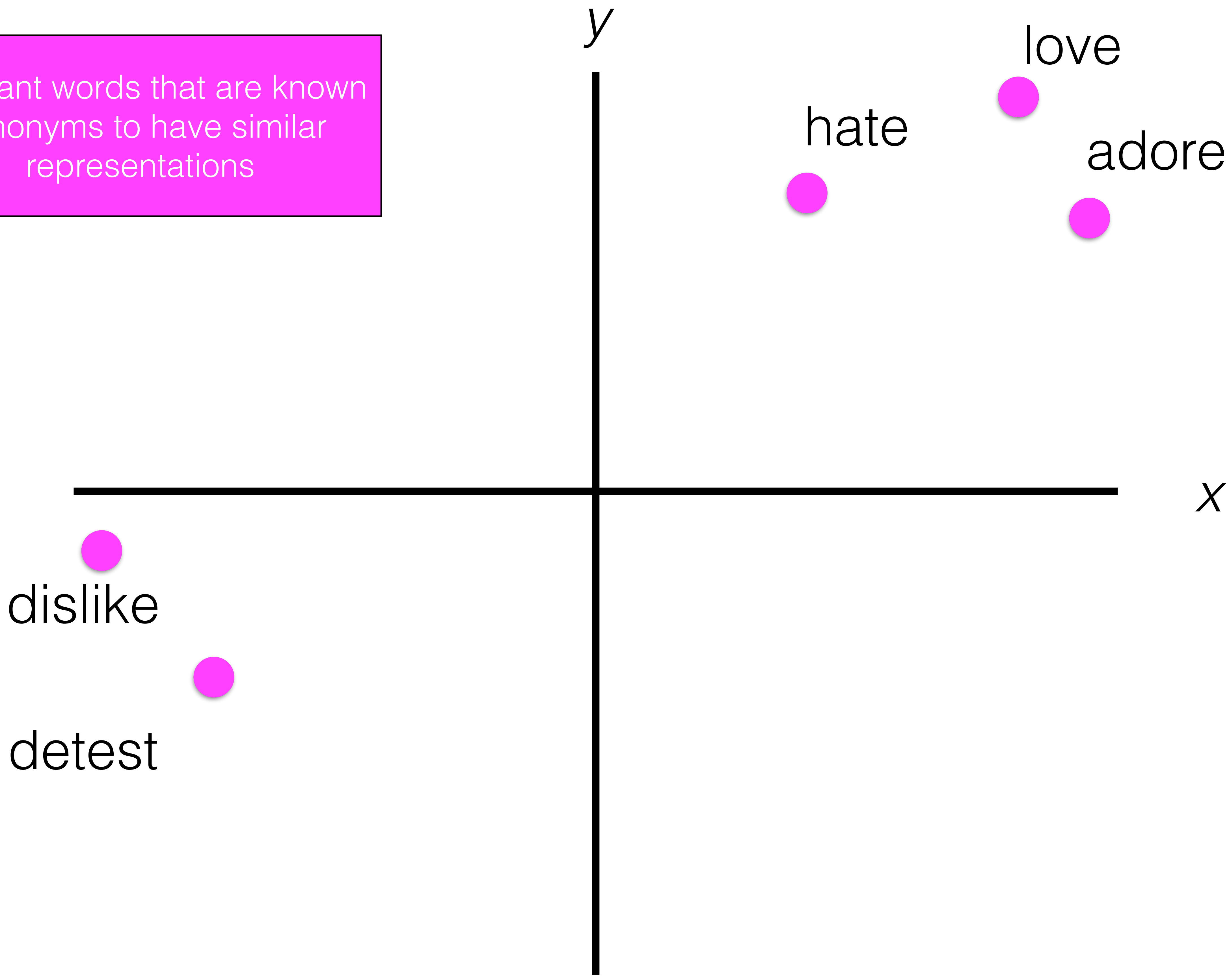
- S: (n) victim, dupe (a person who is tricked or swindled)
  - S: (n) person, individual, someone, somebody, mortal, soul (a human being) *"there was too much for one person to do"*
    - S: (n) organism, being (a living thing that has (or can develop) the ability to act or function independently)
      - S: (n) living thing, animate thing (a living (or once living) entity)
        - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity) *"how big is that part compared to the whole?"; "the team is a unit"*
          - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) *"it was full of rackets, balls and other objects"*
            - S: (n) physical entity (an entity that has physical existence)
              - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Hypernyms of {chump, fool, gull, mark, patsy, fall guy, sucker, soft touch, mug} sunset

WordNet

- WordNet encodes human-judged measures of similarity.
- Learn distributed representations of words that respect WordNet similarities (Faruqui et al. 2015)

We want words that are known synonyms to have similar representations

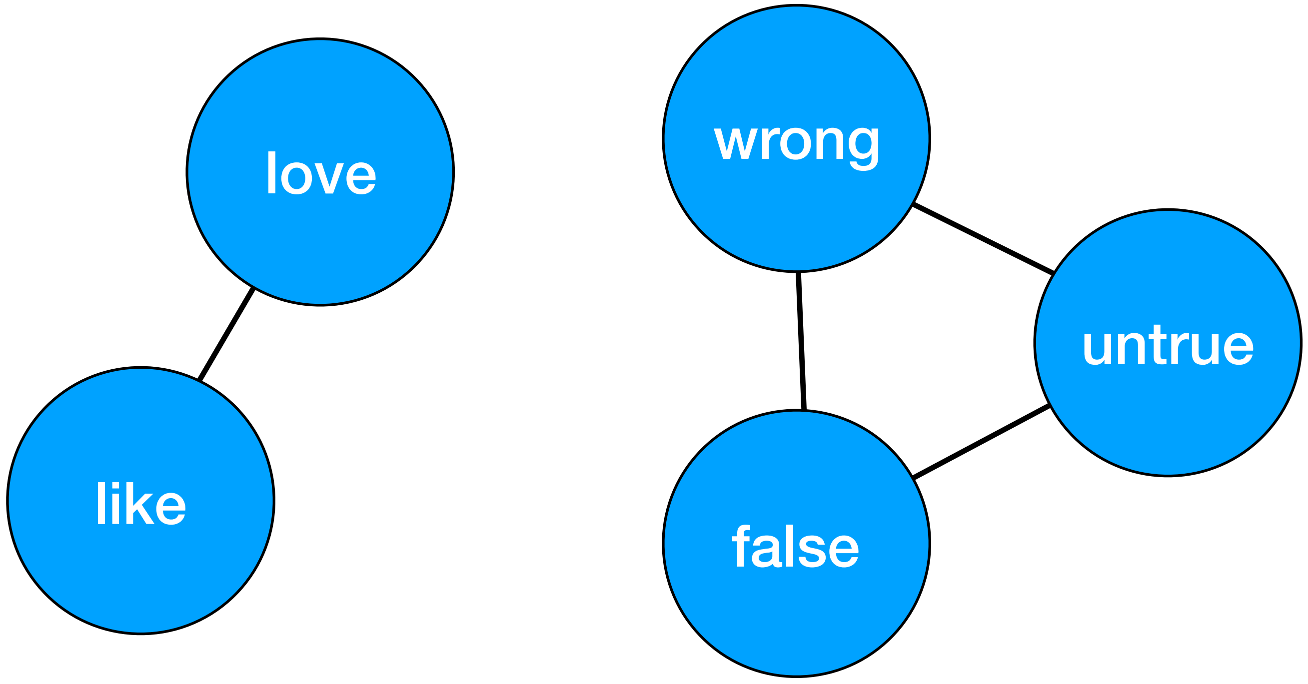


Retrofitting

- Start out with pre-trained word embeddings from any source.

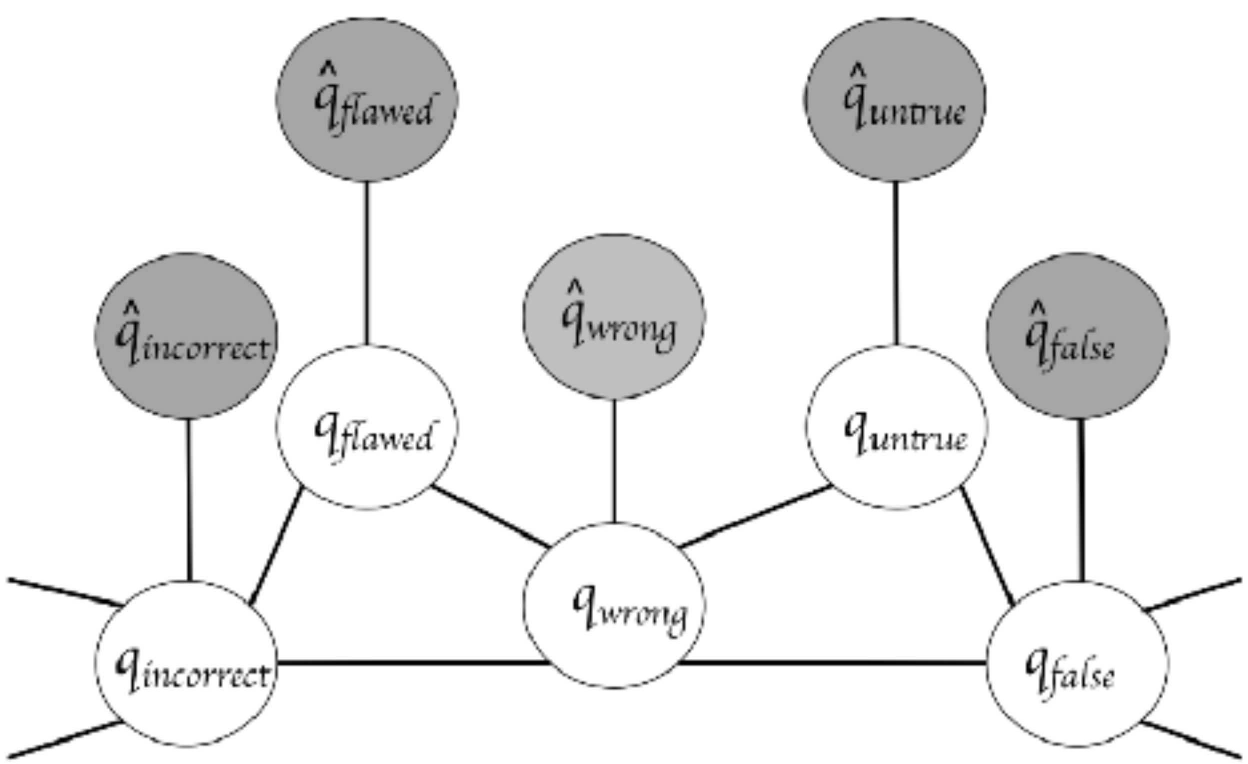
the	likes	wrong	untrue	…
4.1	4.2	0.1	0.12	
-0.9	-0.7	0.3	0.28	

- Define an undirected graph from sets of synomyms



Retrofitting

- Learn new embedding e_{wrong} that is simultaneously close to the original embedding $\hat{e}_{wrong}$ and close to all of its synonyms in WordNet (e_{flawed} $e_{untrue,}$ $e_{incorrect,}$etc).



Retrofitting

- Learn new embedding e_{wrong} that is simultaneously close to the original embedding $\hat{e}_{wrong}$ and close to all of its synonyms in WordNet (e_{flawed} $e_{untrue,}$ $e_{incorrect,}$etc).

We want to minimize this function

$$\Psi(Q) = \sum_{i=1}^{N} \left[\alpha_i \left\| q_i - \hat{q}_i \right\|^2 + \sum_{(i,j)\in E} \beta_{ij} \left\| q_i - q_j \right\|^2 \right]$$

Distance between new embedding and old one

Distance between new embedding and synonyms

Activity

- 10.wordnet/Retrofitting

Semcor

- Semcor: 200K+ words from Brown corpus tagged with Wordnet senses.
    - http://web.eecs.umich.edu/~mihalcea/downloads/semcor/semcor3.0.tar.gz

original	It urged that the city take steps to remedy this problem
lemma sense	It urge^{1} that the city^{2} take^{1} step^{1} to remedy^{1} this problem^{2}
synset number	It $\text{urge}^{2:32:00}$ that the $\text{city}^{1:15:01}$ $\text{take}^{2:41:04}$ $\text{step}^{1:04:02}$ to $\text{remedy}^{2:30:00}$ this $\text{problem}^{1:10:00}$

WordNet

- WordNet encodes human-judged measures of similarity. Learn distributed representations of words that respect WordNet similarities (Faruqui et al. 2015)

- By indexing word senses, we can build annotated resources on top of it for word sense disambiguation.

"All-word" WSD

"Only$_{\text{only1}}$ a relative$_{\text{relative1}}$ handful$_{\text{handful1}}$ of such$_{\text{such0}}$ reports$_{\text{report3}}$ was received$_{\text{receive2}}$"

- For all content words in a sentence, resolve each token to its sense in an fixed sense inventory (e.g., WordNet).

WSD

- Dictionary methods (Lesk)
- Supervised (machine learning)
- Semi-supervised (Bootstrapping)

Dictionary methods

- Predict the sense a given token that has the highest overlap between the token's context and sense's dictionary gloss.

Dictionary methods

bank^1	Gloss:	a financial institution that accepts deposits and channels the money into lending activities
	Examples:	"he cashed a check at the bank", "that bank holds the mortgage on my home"
bank^2	Gloss:	sloping land (especially the slope beside a body of water)
	Examples:	"they pulled the canoe up on the bank", "he sat on the bank of the river and watched the currents"

"The boat washed up on the river bank."

Lesk Algorithm

function SIMPLIFIED LESK(*word*, *sentence*) **returns** best sense of *word*

best-sense ← most frequent sense for *word*
max-overlap ← 0
context ← set of words in *sentence*
for each *sense* **in** senses of *word* **do**
  signature ← set of words in the gloss and examples of *sense*
  overlap ← COMPUTEOVERLAP(*signature*, *context*)
  if *overlap* > *max-overlap* **then**
    max-overlap ← *overlap*
    best-sense ← *sense*
end
return(*best-sense*)

Lesk Algorithm

- Extension (Basile et al. 2014): measure similarity between gloss g = {g_1, … g_G} and context c = {c_1, …, c_C} as cosine similarity between sum of distributed representations

$$\cos\left(\sum_{i=1}^{G} g_i, \sum_{i=1}^{C} c_i\right)$$

Supervised WSD

- We have labeled training data; let’s learn from it.
  - Decision trees (Yarowsky 1994)
  - Naive Bayes, log-linear classifiers, support vector machines (Zhong and Ng 2010)
  - Bidirectional LSTM (Raganato et al. 2017)

Supervised WSD

- Collocational: words in specific positions before/after the target word to be disambiguated

- Bag-of-words: words in window around target (without encoding specific position)

feature
w_{i-1} = fish
w_{i-2}= fish
w_{i+1} = fish
w_{i+2} = fish
word in context = fish
…

	Dev	Test Datasets				Concatenation of All Test Datasets				
	SE07	**SE2**	**SE3**	**SE13**	**SE15**	**Nouns**	**Verbs**	**Adj.**	**Adv.**	**All**
BLSTM	61.8	71.4	68.8	65.6	69.2	70.2	56.3	75.2	**84.4**	68.9
BLSTM + att.	62.4	71.4	**70.2**	66.4	70.8	71.0	**58.4**	75.2	83.5	69.7
BLSTM + att. + LEX	63.7	**72.0**	69.4	66.4	**72.4**	**71.6**	57.1	**75.6**	83.2	**69.9**
BLSTM + att. + LEX + POS	**64.8**	**72.0**	69.1	**66.9**	71.5	71.5	57.5	75.0	83.8	**69.9**
Seq2Seq	60.9	68.5	67.9	65.3	67.0	68.7	54.5	74.0	81.2	67.3
Seq2Seq + att.	62.9	69.9	69.6	65.6	67.7	69.5	57.2	74.5	81.8	68.4
Seq2Seq + att. + LEX	64.6	70.6	67.8	66.5	68.7	70.4	55.7	73.3	82.9	68.5
Seq2Seq + att. + LEX + POS	63.1	70.1	68.5	66.5	69.2	70.1	55.2	75.1	84.4	68.6
IMS	61.3	70.9	69.3	65.3	69.5	70.5	55.8	75.6	82.9	68.9
IMS+emb	**62.6**	**72.2**	**70.4**	65.9	71.5	**71.9**	56.6	**75.9**	**84.7**	**70.1**
Context2Vec	61.3	71.8	69.1	65.6	**71.9**	71.2	**57.4**	75.2	82.7	69.6
Lesk_{ext}+emb	⋆56.7	63.0	63.7	66.2	64.6	70.0	51.1	51.7	80.6	64.2
UKB_{gloss} w2w	42.9	63.5	55.4	⋆62.9	63.3	64.9	41.4	69.5	69.7	61.1
Babelfy	51.6	⋆67.0	63.5	**66.4**	70.3	68.9	50.7	73.2	79.8	66.4
MFS	54.5	65.6	⋆66.0	63.8	⋆67.1	67.7	49.8	73.1	80.5	65.5

Raganato et al. 2017

One sense per discourse

- If a word appears multiple times in a document, it’s usually with the same sense. (Gale et al. 1992)
    - Articles about financial banks don’t usually talk about river banks.

Supervised learning

But WSD is not POS tagging

Semi-supervised WSD

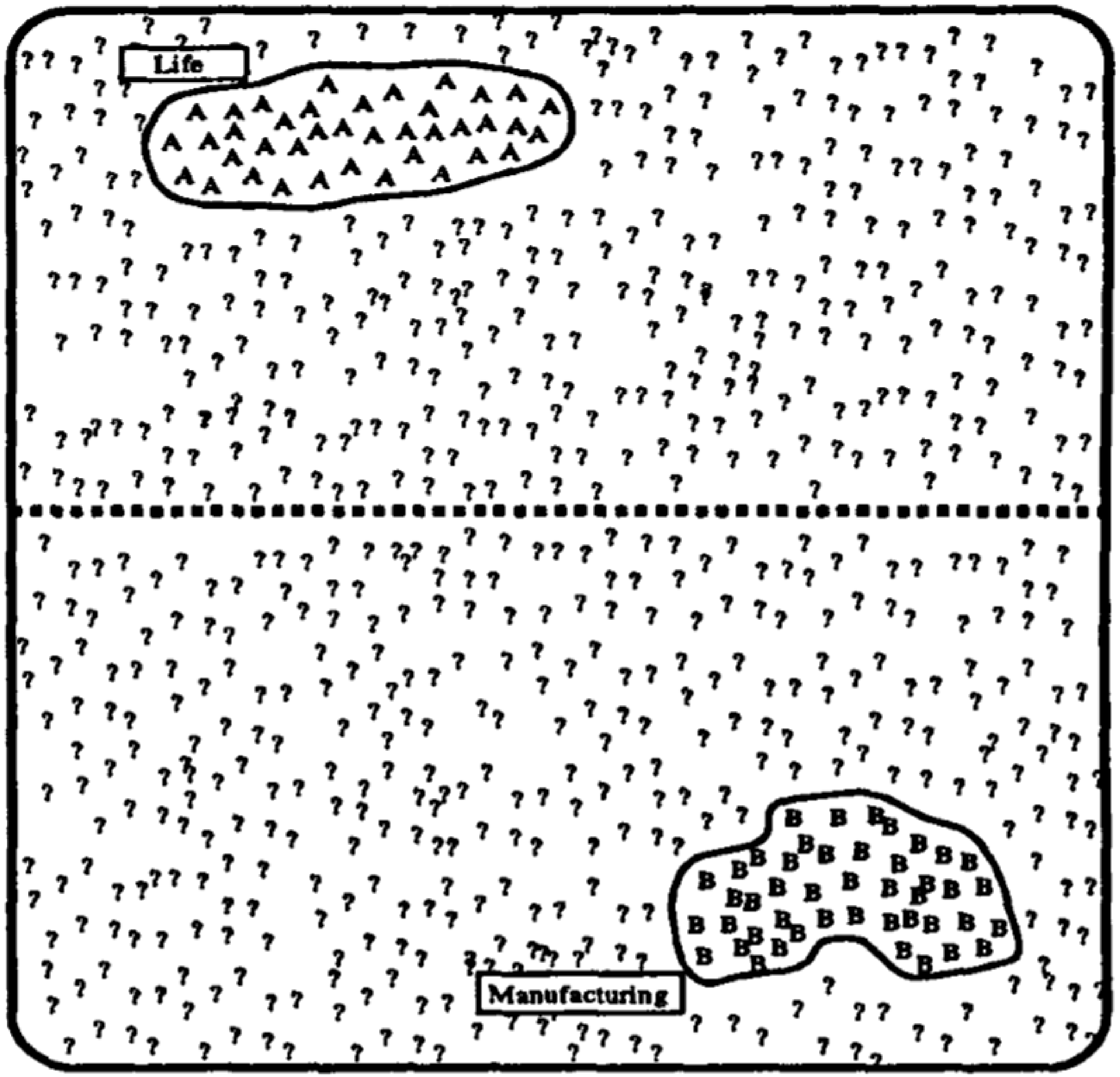
1. Produce seeds (dictionary definitions, single defining collocate, or label common collocates)

2. Repeat until convergence:

    1. Train supervised classifier on labeled examples

    2. Label all examples, and keep labels for high-confidence instances

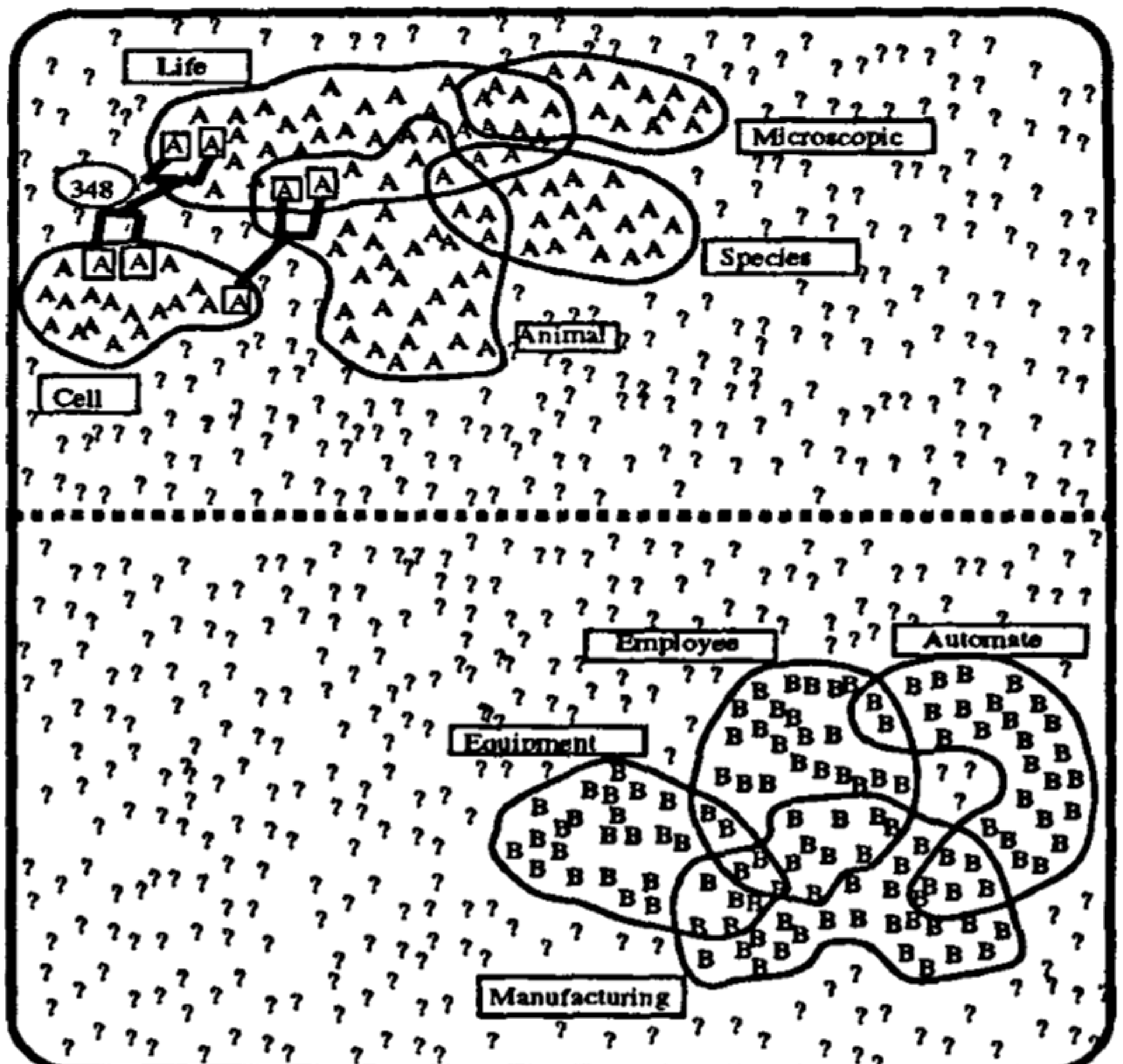
Semi-supervised WSD



“Plant”

A = SENSE-A training example
B = SENSE-B training example
? = currently unclassified training example
Life = Set of training examples containing the collocation “life”.

Semi-supervised WSD



"Plant"

A = SENSE-A training example
B = SENSE-B training example
? = currently unclassified training example
Life = Set of training examples containing the collocation "life".

Evaluation

- Annotated data; cross-validation.
    - Semcor
    - Ontonotes
- Semeval/Senseval competitions

Hyponymy

	… Entity
… Entity	Artifact
Animal	Instrumentality
Vertebrate	Conveyance
Mammal	Wheeled vehicle
Ungulate	Self-propelled vehicle
Equine	Motor vehicle
Horse	Car

NOUNS			
SUPERSENSE	NOUNS DENOTING	SUPERSENSE	NOUNS DENOTING
act	acts or actions	object	natural objects (not man-made)
animal	animals	quantity	quantities and units of measure
artifact	man-made objects	phenomenon	natural phenomena
attribute	attributes of people and objects	plant	plants
body	body parts	possession	possession and transfer of possession
cognition	cognitive processes and contents	process	natural processes
communication	communicative processes and contents	person	people
event	natural events	relation	relations between people or things or ideas
feeling	feelings and emotions	shape	two and three dimensional shapes
food	foods and drinks	state	stable states of affairs
group	groupings of people or objects	substance	substances
location	spatial position	time	time and temporal relations
motive	goals	Tops	abstract terms for unique beginners
VERBS			
SUPERSENSE	VERBS OF	SUPERSENSE	VERBS OF
body	grooming, dressing and bodily care	emotion	feeling
change	size, temperature change, intensifying	motion	walking, flying, swimming
cognition	thinking, judging, analyzing, doubting	perception	seeing, hearing, feeling
communication	telling, asking, ordering, singing	possession	buying, selling, owning
competition	fighting, athletic activities	social	political and social activities and events
consumption	eating and drinking	stative	being, having, spatial relations
contact	touching, hitting, tying, digging	weather	raining, snowing, thawing, thundering
creation	sewing, baking, painting, performing		

Ciarmita and Altun 2003

Supersense tagging

artifact artifact motion time group

The station wagons arrived at noon, a long shining line

motion location location

that coursed through the west campus.

Noun supersenses (Ciarmita and Altun 2003)

Supersense tagging

- Ciarameta and Altun (2006). Trained on data from Semcor (Miller et al. 1993); Brown corpus annotated with WordNet synset labels

- Token-level predictor – each instance of a word has its own supersense tag.

- Maximum-entropy Markov Model (MEMM) trained with averaged perceptron. Features for: word token identity, part-of-speech tag, word shape, previous label + supersense for most frequent synset for word.

- In-domain accuracy: 77.1 F score (cf. 66 F MFS baseline)

Data

- Semcor: 200K+ words tagged with Wordnet senses.
  http://www.cse.unt.edu/~rada/downloads.html#semcor
- WordNet
  https://wordnet.princeton.edu/wordnet/download/

Activity

- 10.wordnet/Lesk_TODO.ipynb