Language
Technology
Lab

UNIVERSITY OF CAMBRIDGE

Semantic Change in the Time of Machine Learning: doing it right!

Haim Dubossarsky

1st International Workshop on Computational Approaches to Historical Language Change

Florence, August 2019

hd423@cam.ac.uk

Congratulations!

+

Historical distributional semantics

Outline

- Problem breakdown
- Working with faulty models
- Case I: Laws of semantic change
- Case II: Comparing models' quality
- Conclusions

Problem breakdown

it’s everywhere,

it’s effects can be felt,

but you cannot see or touch it

-> meaning is the dark matter of language

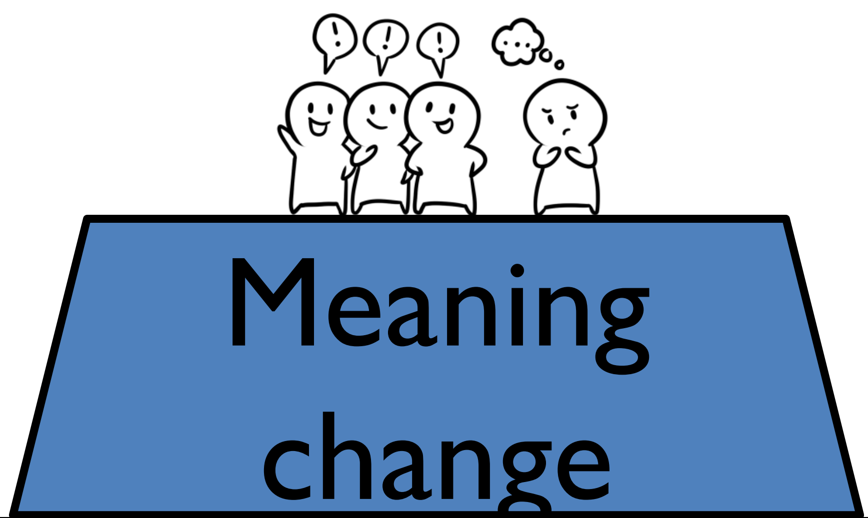
Slide, courtesy of Prof. Dirk Geeraerts

Problem breakdown

it’s everywhere,

it’s effects can be felt,

but you cannot see or touch it



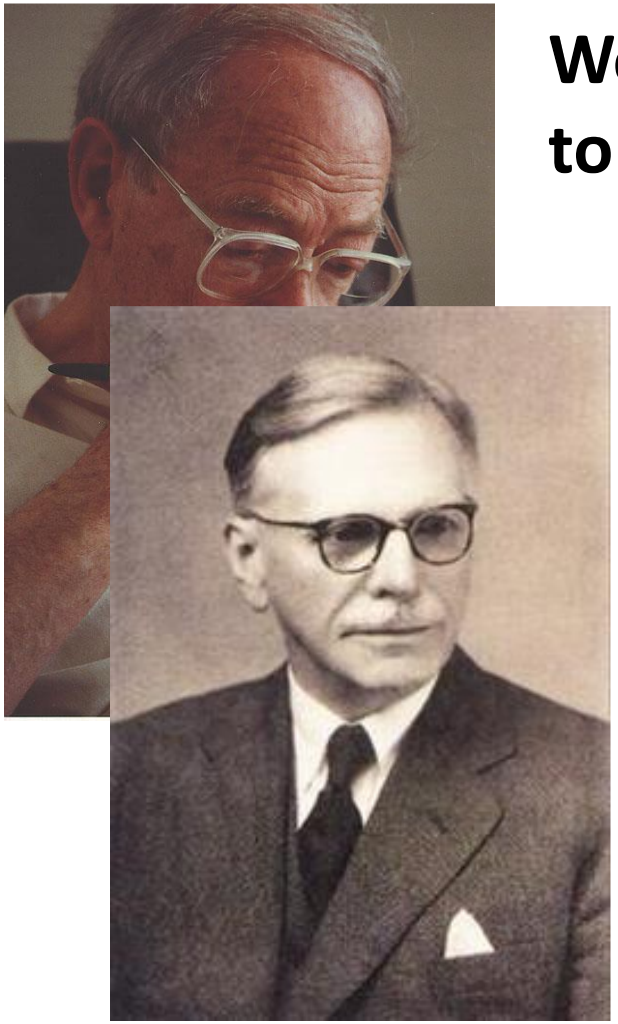
-> meaning is the dark matter of language

Solving this conundrum

1st International Workshop on Computational Approaches to Historical Language Change 2019

Nina Tahmasebi , Lars Borin , Adam Jatowt , Yang Xu

The distributional hypothesis



Words occurring in similar contexts tend to have similar meanings (Z. Harris, 1954)

You shall know a word by the company it keeps (Firth, J. R. 1957:11)

Word embeddings*

* Not a survey

- Could be sparse vectors (counts, PPMI, RI)

$w_j = news$ $\quad w_k = reporter$ $\quad w_l = do$ $\quad w_m = ceiling$

$w_i = broadcast$

	94				56					60				0	

$|V|$

- Or dense vectors (word2vec , FastText, Glove)

$? \quad ?? \quad ???$

$w_i = broadcast$

…	…	…	…	…	…	…	…

$|d|$

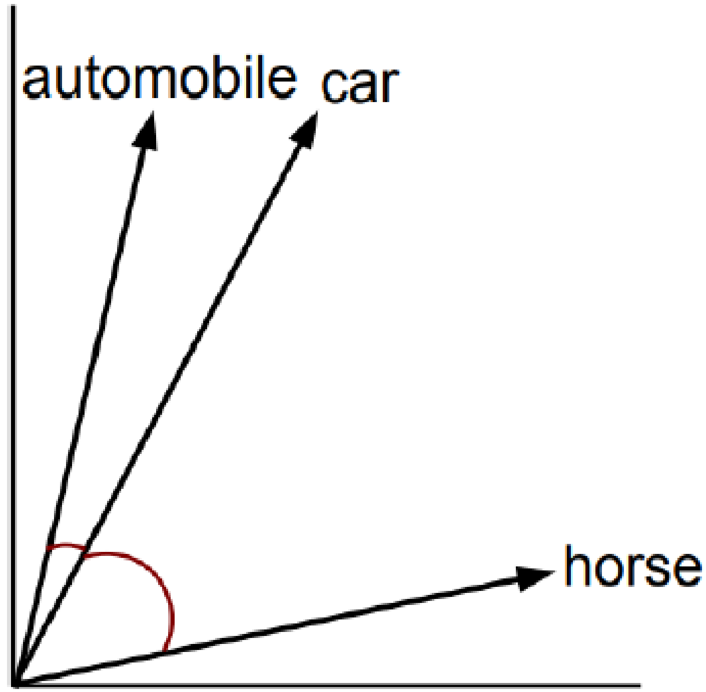
- Or yet contextual embedding (ELMo, Bert)

All <u>define</u> meaning as usage statistics.

Embeddings capture meaning

But how did we come up with that conclusion?

Word 1	Word 2	Human	Embeddin
horse	car	5.9	0.79
book	paper	7.46	0.85
computer	keyboard	7.62	0.79
train	car	6.31	0.5
television	radio	6.77	0.73
drug	abuse	6.85	0.45
bread	butter	6.19	0.65
cucumber	potato	5.92	0.75
doctor	nurse	7	0.84
smart	stupid	5.81	0.6
stock	market	8.08	0.97



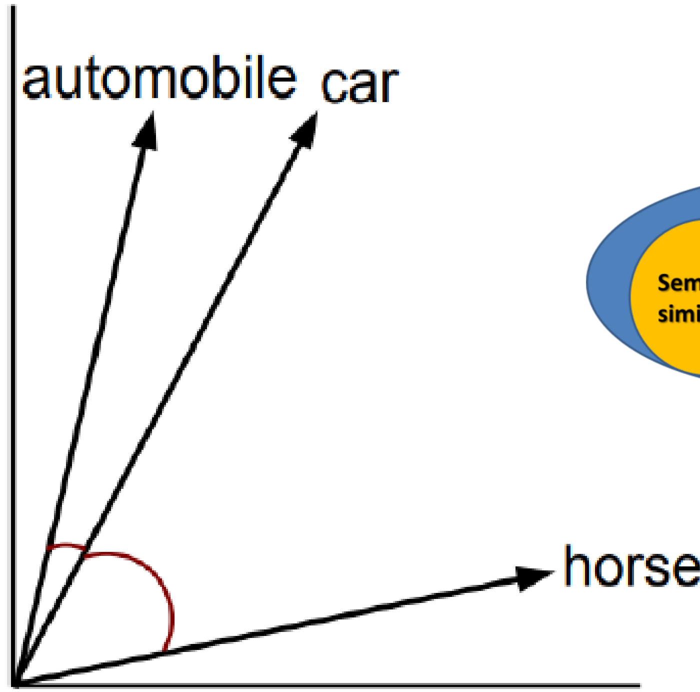
r=.72

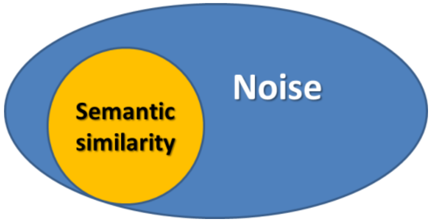
$$cosine\ similarity(w^1, w^2) = \frac{\vec{w}^1 \cdot \vec{w}^2}{\|\vec{w}^1\| \cdot \|\vec{w}^2\|}$$

Embeddings capture meaning

But how did we come up with that conclusion?

Word 1	Word 2	Human	Embeddin
horse	car	5.9	0.79
book	paper	7.46	0.85
computer	keyboard	7.62	0.79
train	car	6.31	0.5
television	radio	6.77	0.73
drug	abuse	6.85	0.45
bread	butter	6.19	0.65
cucumber	potato	5.92	0.75
doctor	nurse	7	0.84
smart	stupid	5.81	0.6
stock	market	8.08	0.97





r=.72

✓ Vectors capture semantic meaning

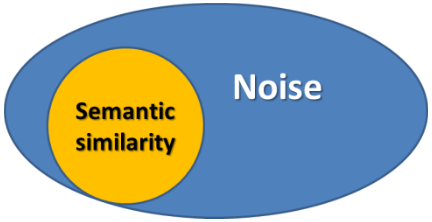
≠ Vectors capture <u>only</u> semantic meaning

Embeddings capture meaning

But how did we come up with that conclusion?

What is noise?
A confound.

Any unwanted variable that influence our metric.



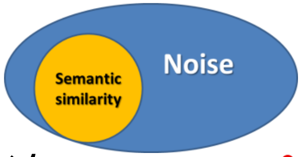
✓ Vectors capture semantic meaning

≠ Vectors capture <u>only</u> semantic meaning

Problem breakdown

Semantic change definition

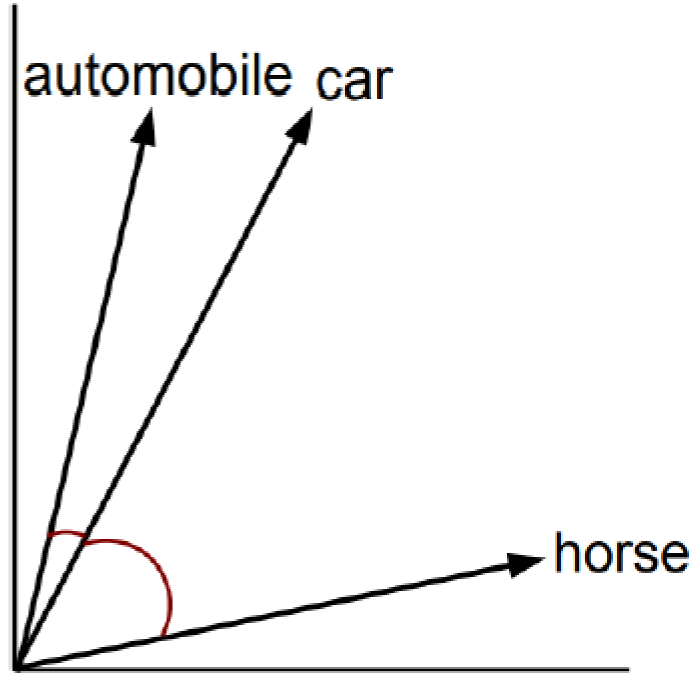
Change to a word's embeddings between two time points [word relative to itself]



$$\Delta w^{t^0 \to t^1} = cosDist(w^{t_0}, w^{t_1}) = 1 - \frac{\vec{w}^{t_0} \cdot \vec{w}^{t_1}}{\|\vec{w}^{t_0}\| \cdot \|\vec{w}^{t_1}\|}$$

Semantic change validated?

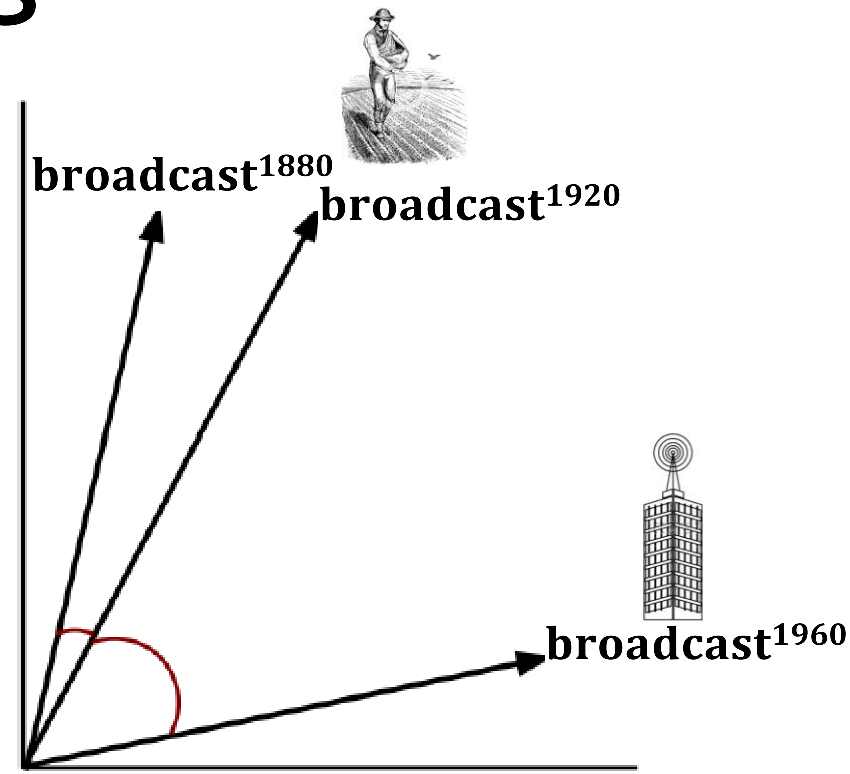
Word 1	Word 2	Human	Embeddin
horse	car	5.9	0.79
book	paper	7.46	0.85
computer	keyboard	7.62	0.79
train	car	6.31	0.5
television	radio	6.77	0.73
drug	abuse	6.85	0.45
bread	butter	6.19	0.65
cucumber	potato	5.92	0.75
doctor	nurse	7	0.84
smart	stupid	5.81	0.6
stock	market	8.08	0.97



$$cosine\ similarity(w^1, w^2) = \frac{\vec{w}^1 \cdot \vec{w}^2}{||\vec{w}^1|| \cdot ||\vec{w}^2||}$$

Semantic change validated?

Word 1	Word 2	Human	Embeddin
horse	car	5.9	0.79
book	paper	7.46	0.85
computer	keyboard	7.62	0.79
train	car	6.31	0.5
television	radio	6.77	0.73
drug	abuse	6.85	0.45
bread	butter	6.19	0.65
cucumber	potato	5.92	0.75
doctor	nurse	7	0.84
smart	stupid	5.81	0.6
stock	market	8.08	0.97

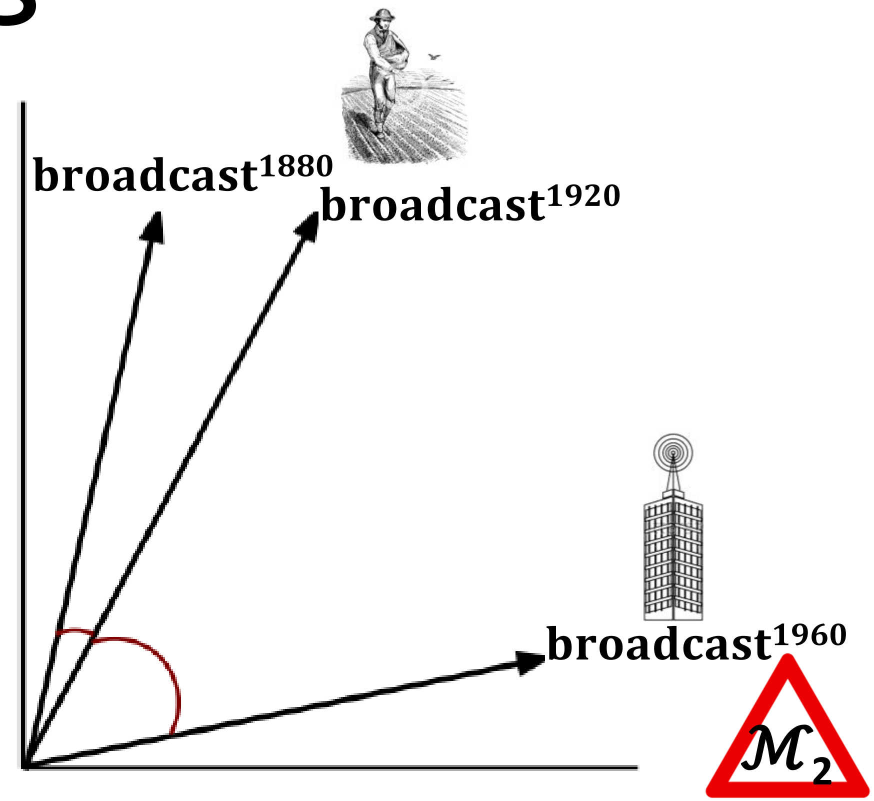


$$cosine\ similarity(w^{t1}, w^{t2}) = \frac{\vec{w}^{t1} \cdot \vec{w}^{t2}}{\|\vec{w}^{t1}\| \cdot \|\vec{w}^{t2}\|}$$

Semantic change validated?

Word 1	Word 2	Human	Embeddin
horse	car	5.9	0.79
book	paper	7.46	0.85
c[illegible]	[illegible]	[illegible]	[illegible]
tr[illegible]	[illegible]	[illegible]	[illegible]
television	radio	6.77	0.73
drug	abuse	6.85	0.45
bread	butter	6.19	0.65
cucumber	potato	5.92	0.75
doctor	nurse	7	0.84
smart	stupid	5.81	0.6
stock	market	8.08	0.97

Gulordava and Baroni (2011)

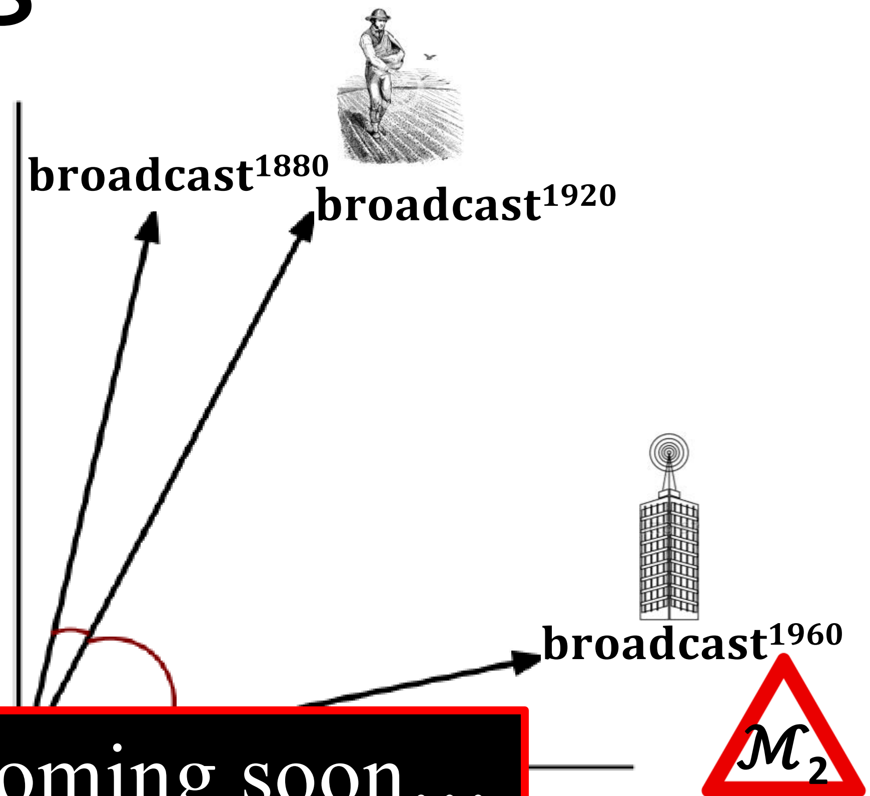


$$cosine\ similarity(w^{t1}, w^{t2}) = \frac{\vec{w}^{t1} \cdot \vec{w}^{t2}}{\|\vec{w}^{t1}\| \cdot \|\vec{w}^{t2}\|}$$

Semantic change validated?

Word 1	Word 2	Human	Embeddin
horse	car	5.9	0.79
book	paper	7.46	0.85
co[illegible]	[illegible]	[illegible]	[illegible]
tr[illegible]	[illegible]	[illegible]	[illegible]
television	radio	6.77	0.73
drug	abuse	6.85	0.45
bread	butter	6.19	0.65
cucumber	potato	5.92	0.75
doctor	nurse	7	0.84
smart	stupid	5.81	0.6
stock	mark[illegible]	[illegible]	[illegible]

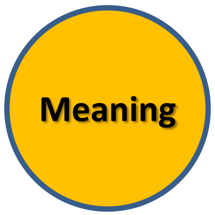
Gulordava and Baroni (2011)

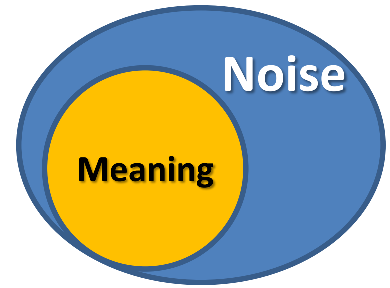


SemEval-2020. Coming soon…

$$cosine\ similarity(w^{t1}, w^{t2}) = \frac{\vec{w}^{t1} \cdot \vec{w}^{t2}}{\|\vec{w}^{t1}\| \cdot \|\vec{w}^{t2}\|}$$

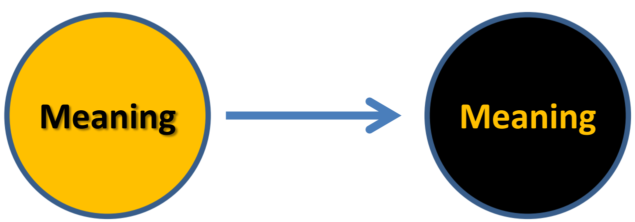
Interim summary

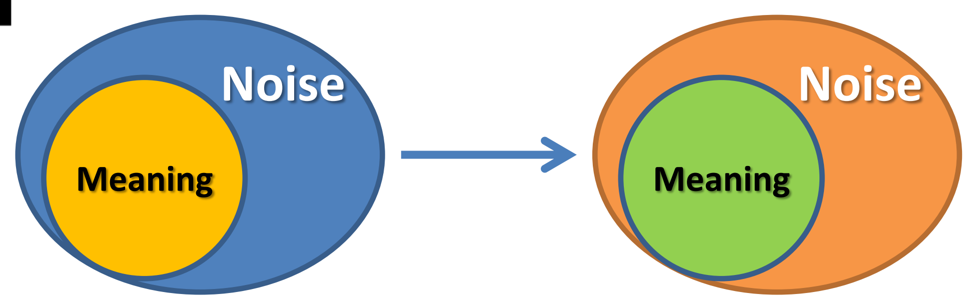




What we think word embeddings capture

What word embeddings REALLY capture

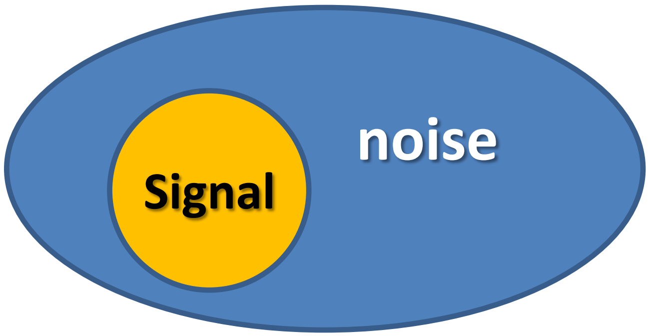


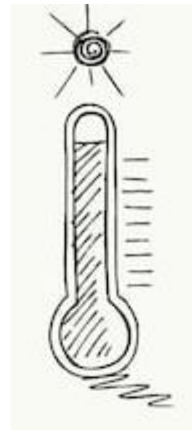


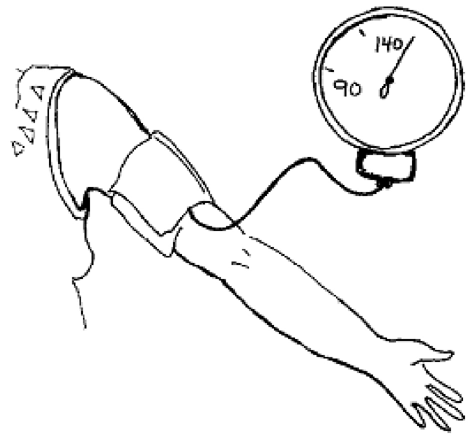
What we think
models of semantic change capture

What
models of semantic change REALLY capture

All models are wrong







1. How wrong are they?
2. Are they importantly wrong?

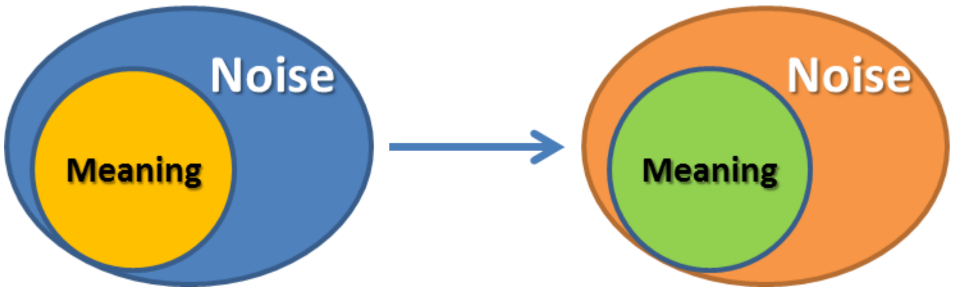
Depends on what do we use these models for

When to worry about noise

1. Word embedding as some proxy for meaning
   - Machine translation, chat bots…
   - Detecting semantic change

Most Changed		Least Changed	
Word	Similarity	Word	Similarity
checked	0.3831	*by*	0.9331
check	0.4073	*than*	0.9327
gay	0.4079	*for*	0.9313
actually	0.4086	*more*	0.9274
supposed	0.4232	*other*	0.9272
guess	0.4233	*an*	0.9268
cell	0.4413	*own*	0.9259
headed	0.4453	*with*	0.9257
ass	0.4549	*down*	0.9252
mail	0.4573	*very*	0.9239

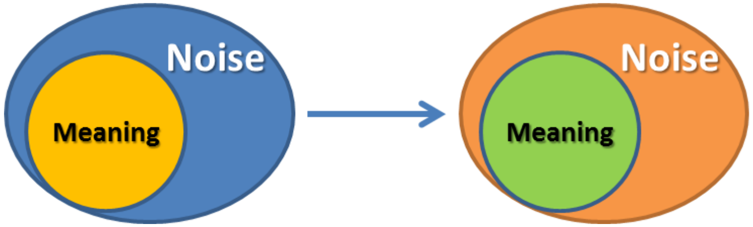
From Kim et al. (2014)



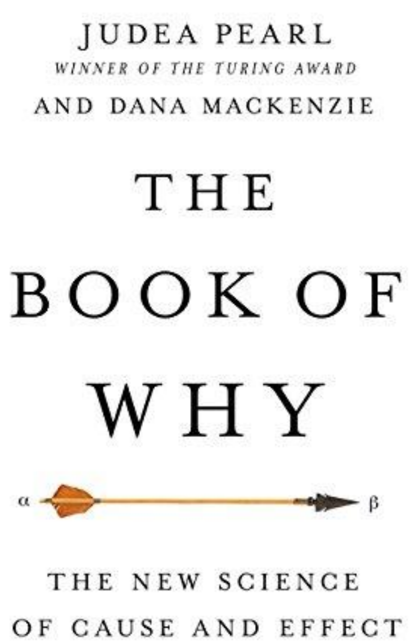
When to worry about noise

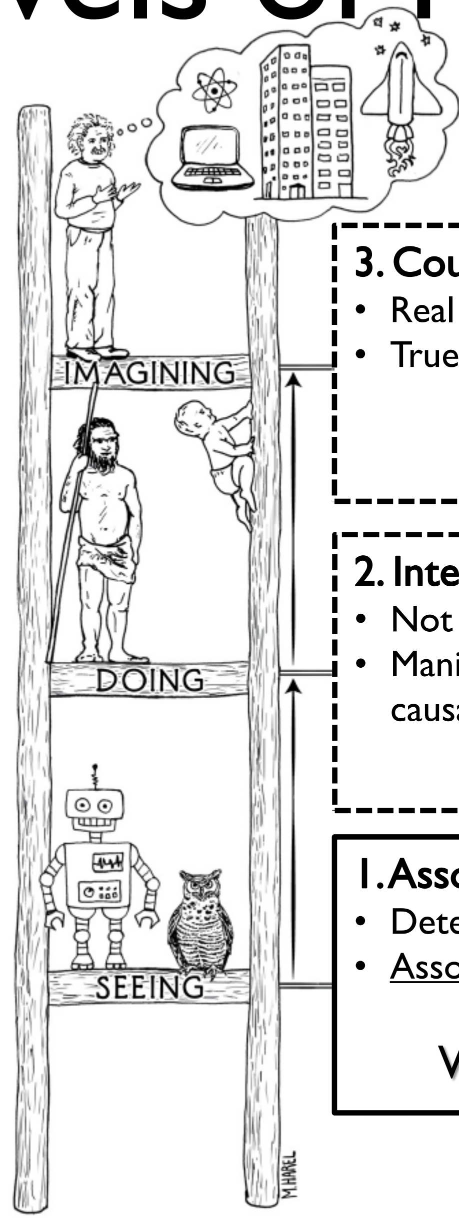
1. Word embedding as some proxy for meaning
   - Machine translation, chat bots…
   - Detecting semantic change

2. Word embedding as <u>the object of study</u>
   - Over interpret differences in embeddings that actually stem from noise.
   - Laws of semantic change



Levels of reasoning





3. Counterfactuals
- Real causation
- True laws of semantic change

What causes what

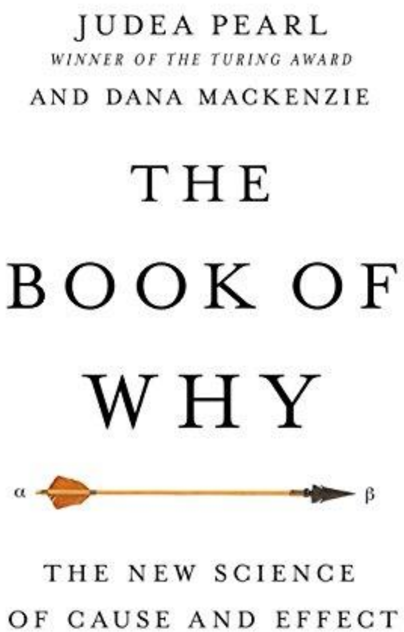
2. Intervention / dissociation
- Not just observing, but changing.
- Manipulating variables to establish causation.

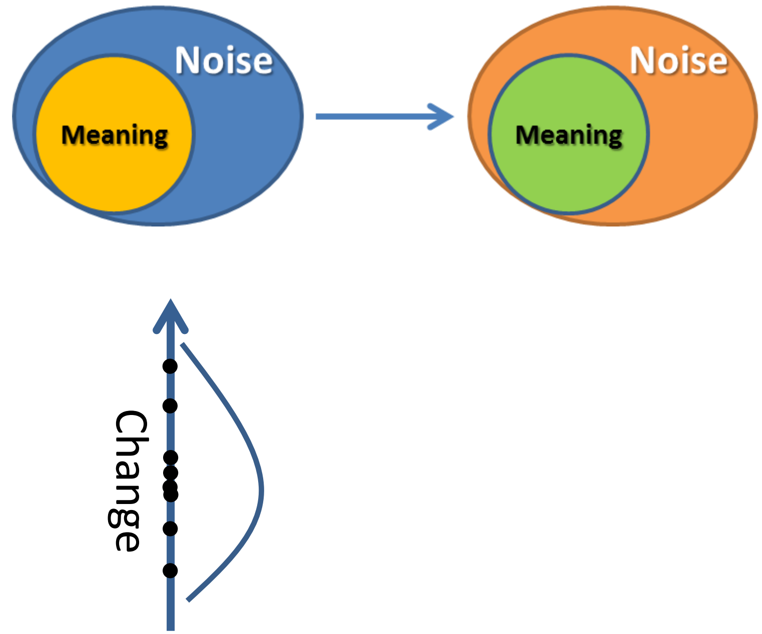
What causes what

1. Association via correlation
- Detecting semantic change
- Associative "laws" of semantic change

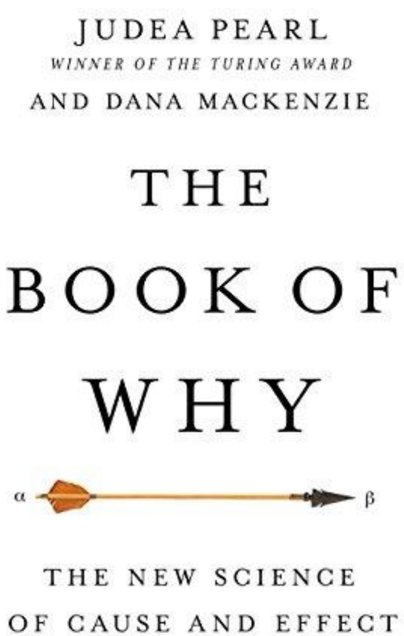
What happens with what

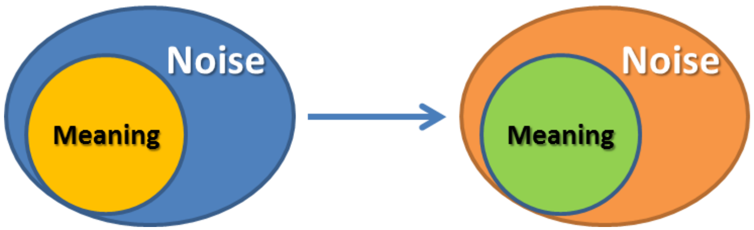
Risks in associative reasoning

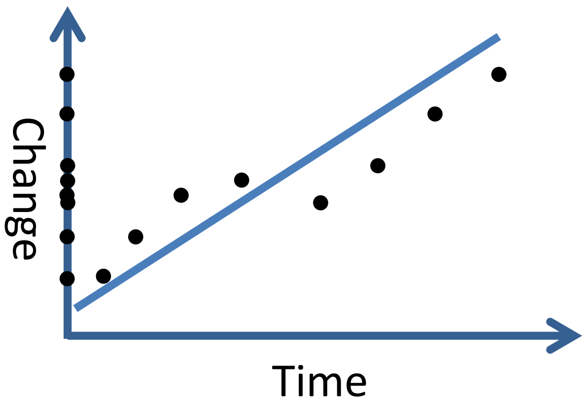




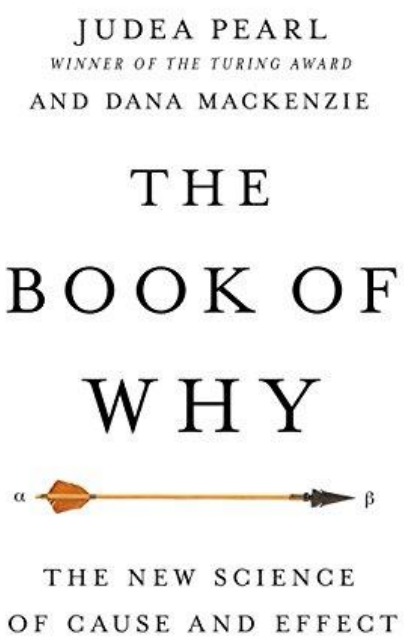
Risks in associative reasoning

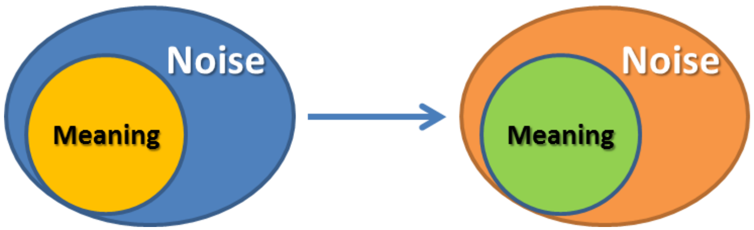


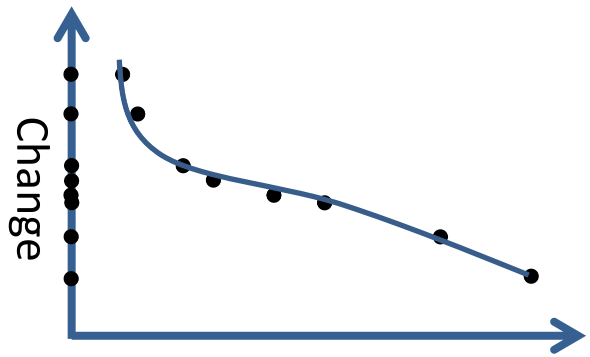




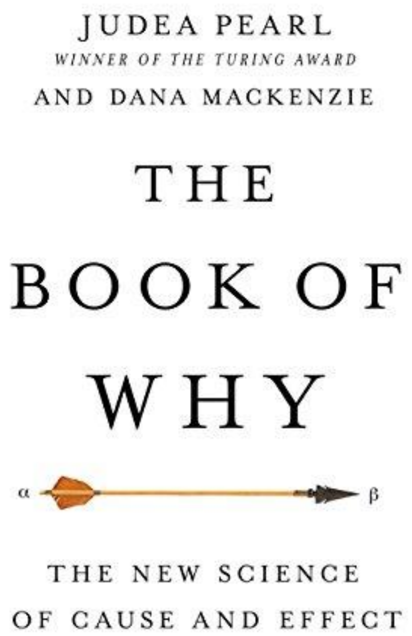
Risks in associative reasoning

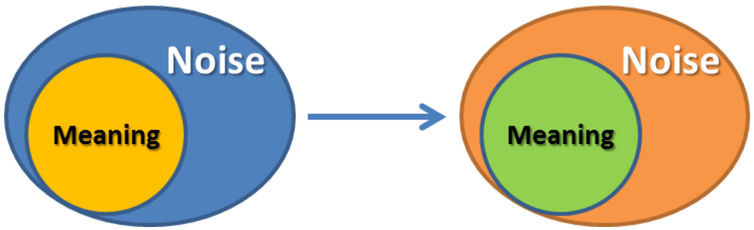


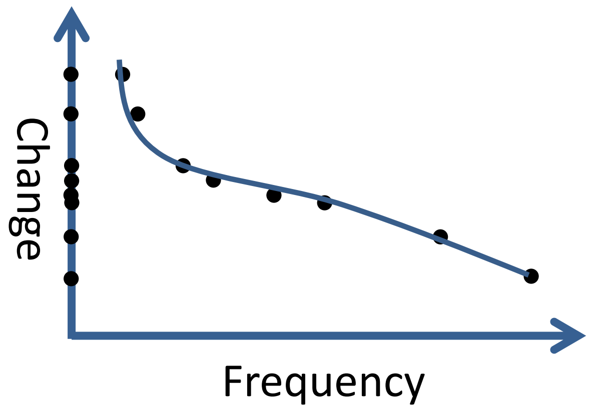


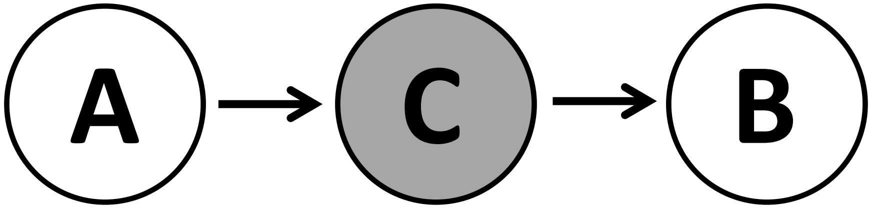


Risks in associative reasoning

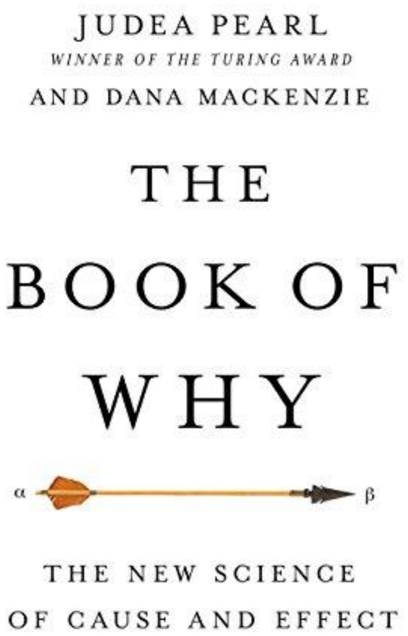


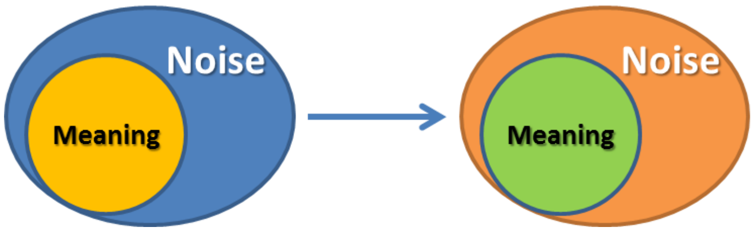




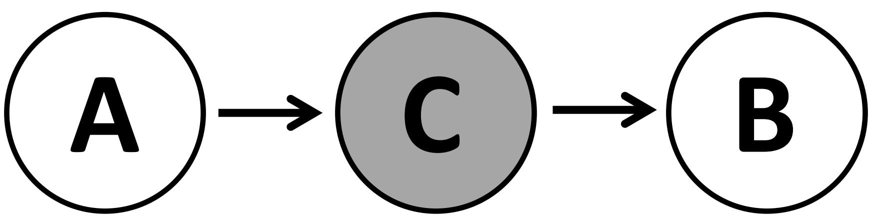


Risks in associative reasoning



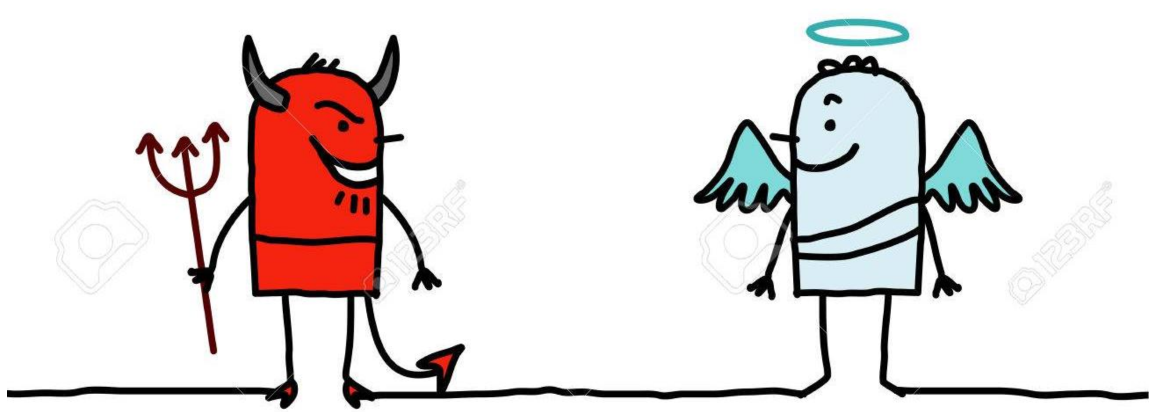


Randomized Controlled Trials

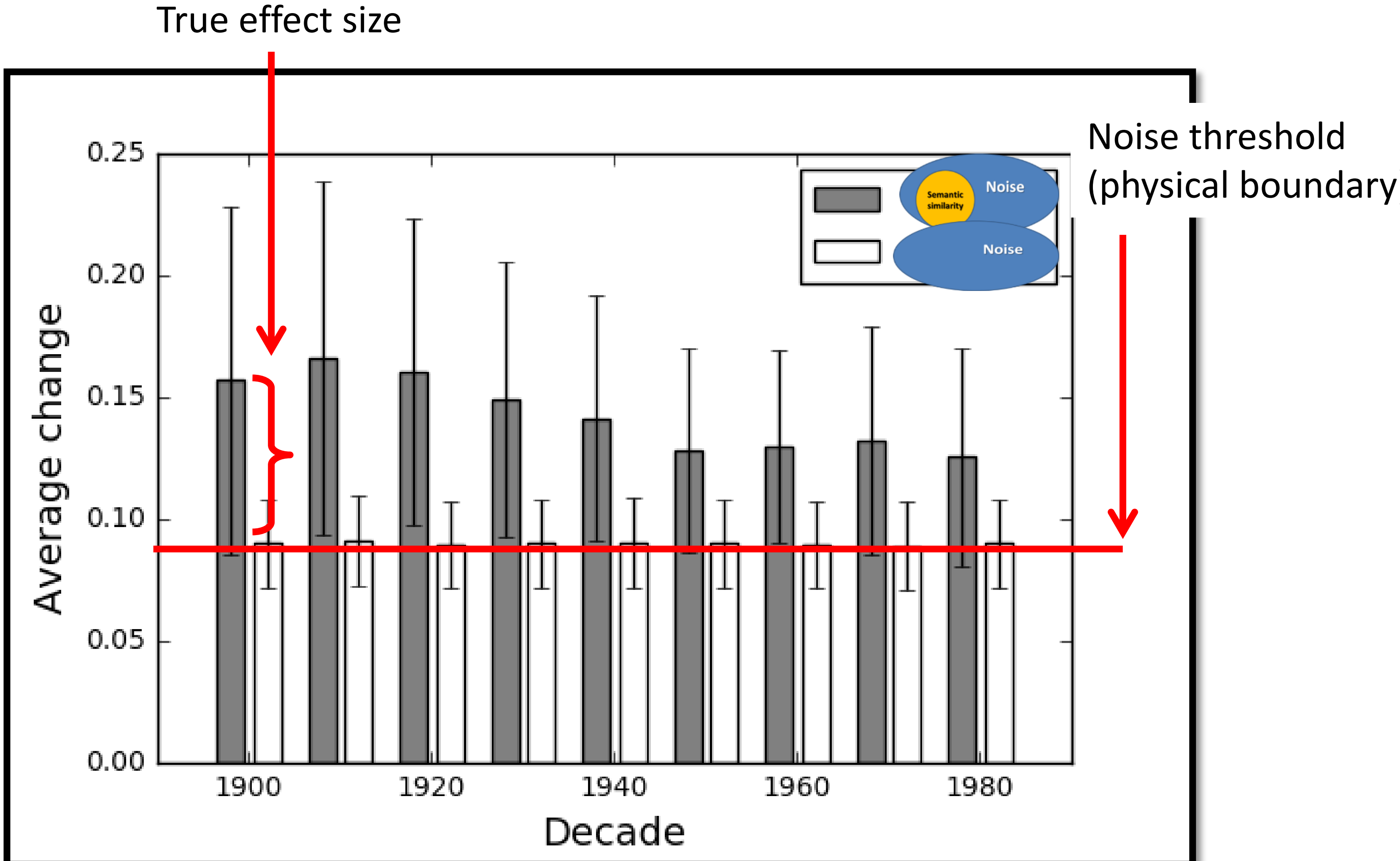


Randomized Controlled Trials Case I

Laws of semantic change



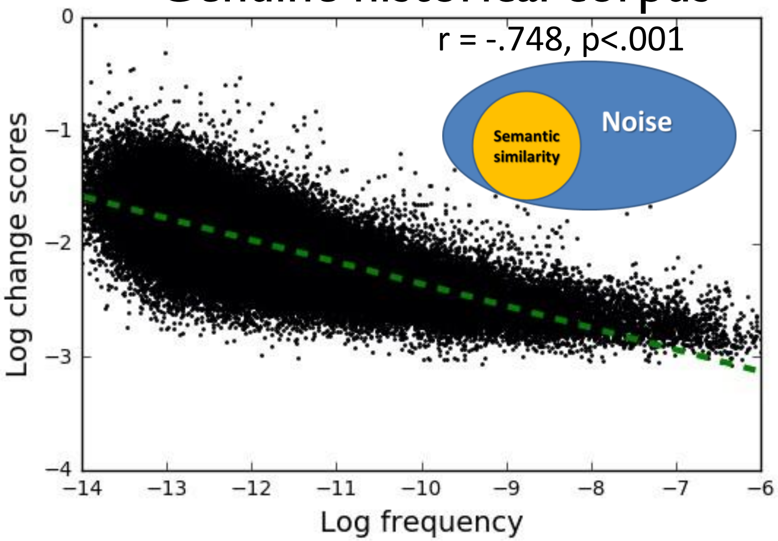
How wrong models are?



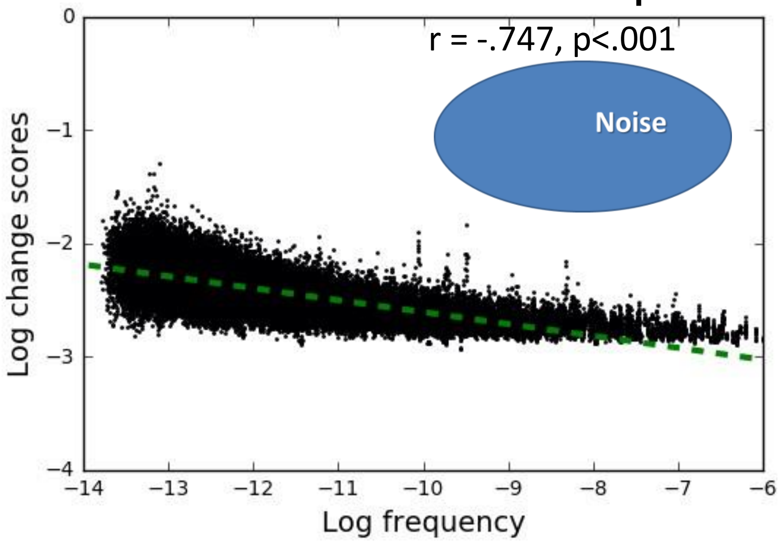
From Dubossarsky et al. (2017)

Are they importantly wrong?

Genuine historical corpus



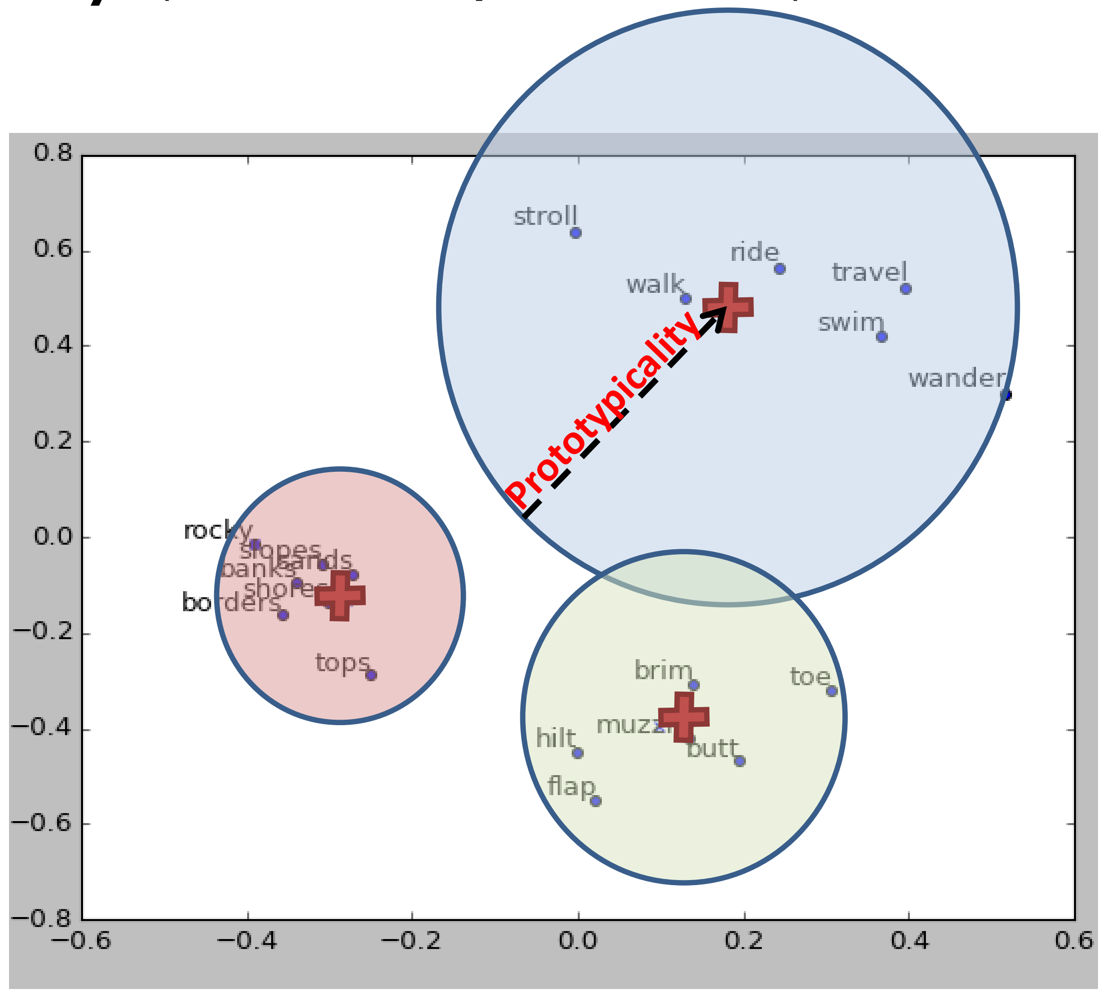
Shuffled historical corpus



Equal effect sizes for the *genuine* historical corpus and the *shuffled* historical corpus (Dubossarsky et al. 2017).

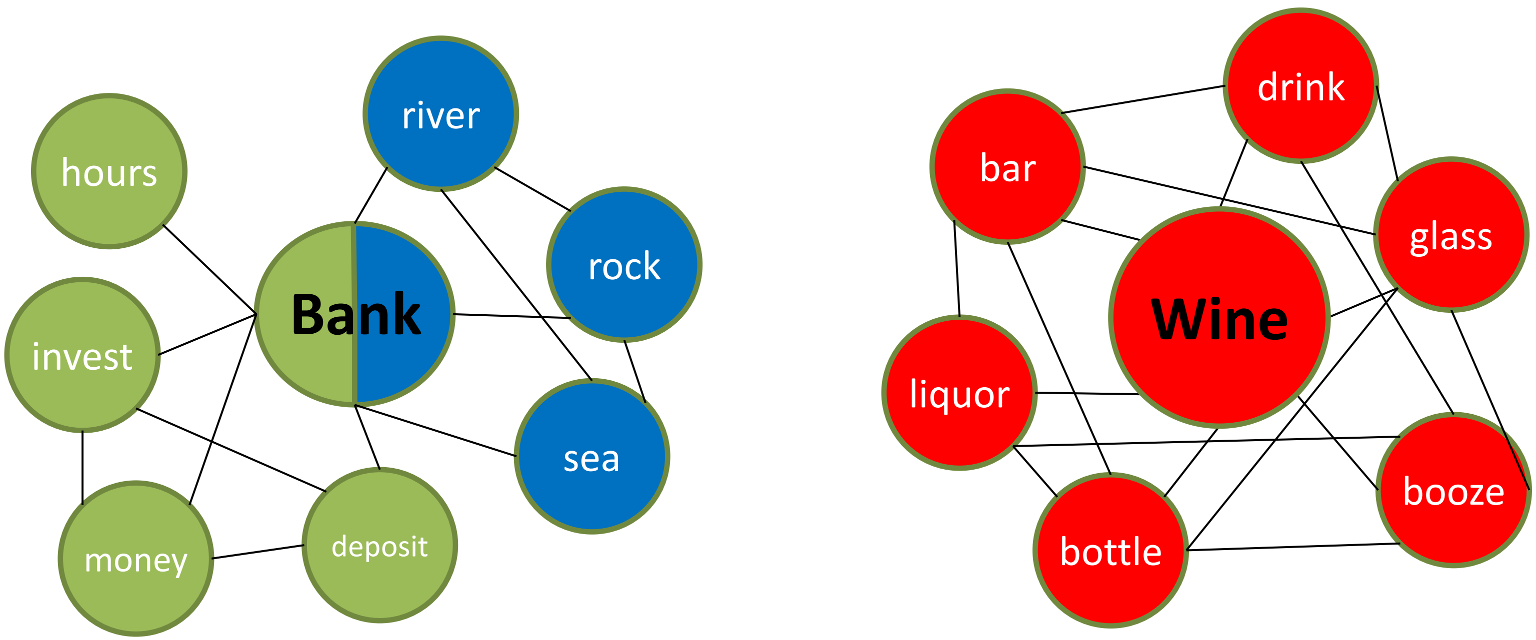
"Laws" of semantic change

- Law of Prototypicality (Dubossarsky et. al. 2015).



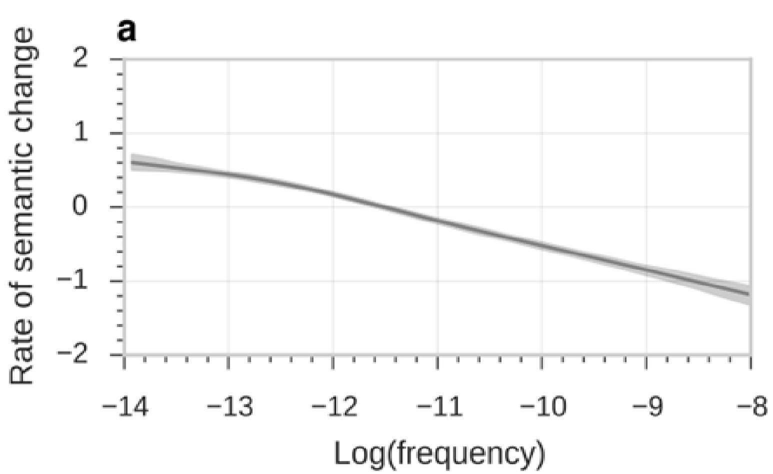
"Laws" of semantic change

- Law of Prototypicality (Dubossarsky et. al. 2015).
- Law of Innovation (Polysemy, Hamilton et. al. 2016).

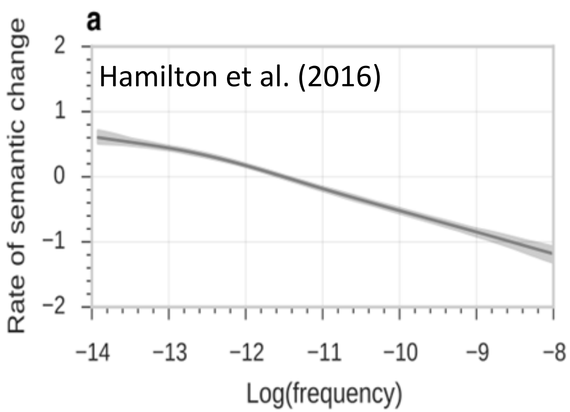


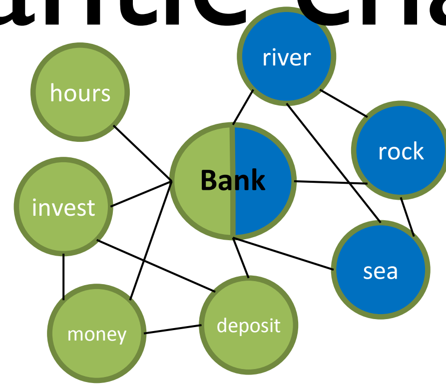
"Laws" of semantic change

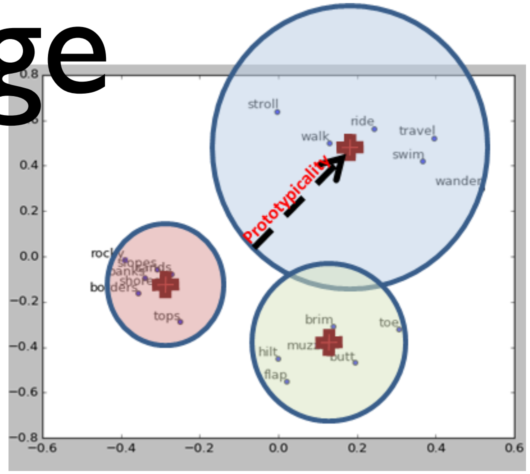
- Law of Prototypicality (Dubossarsky et. al. 2015).
- Law of Innovation (Polysemy, Hamilton et. al. 2016).
- Law of Conformity (Frequency, Hamilton et. al. 2016).



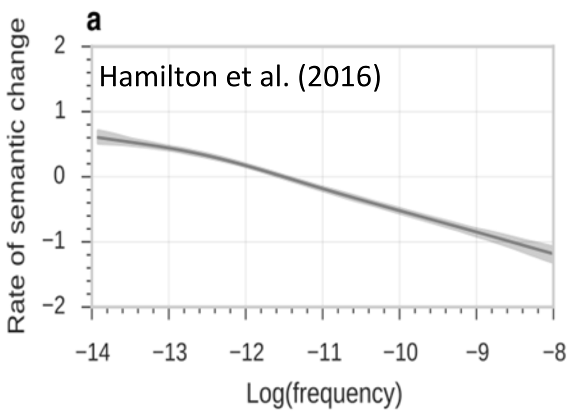
Associative laws of semantic change

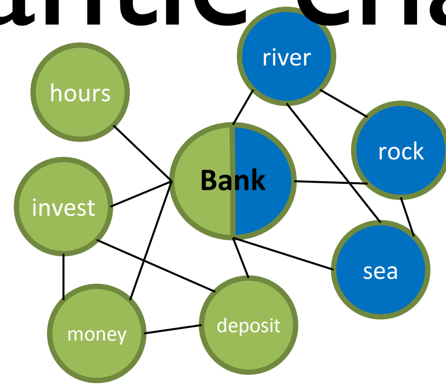


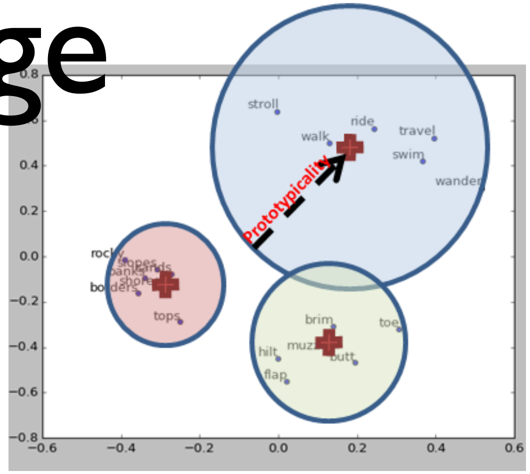




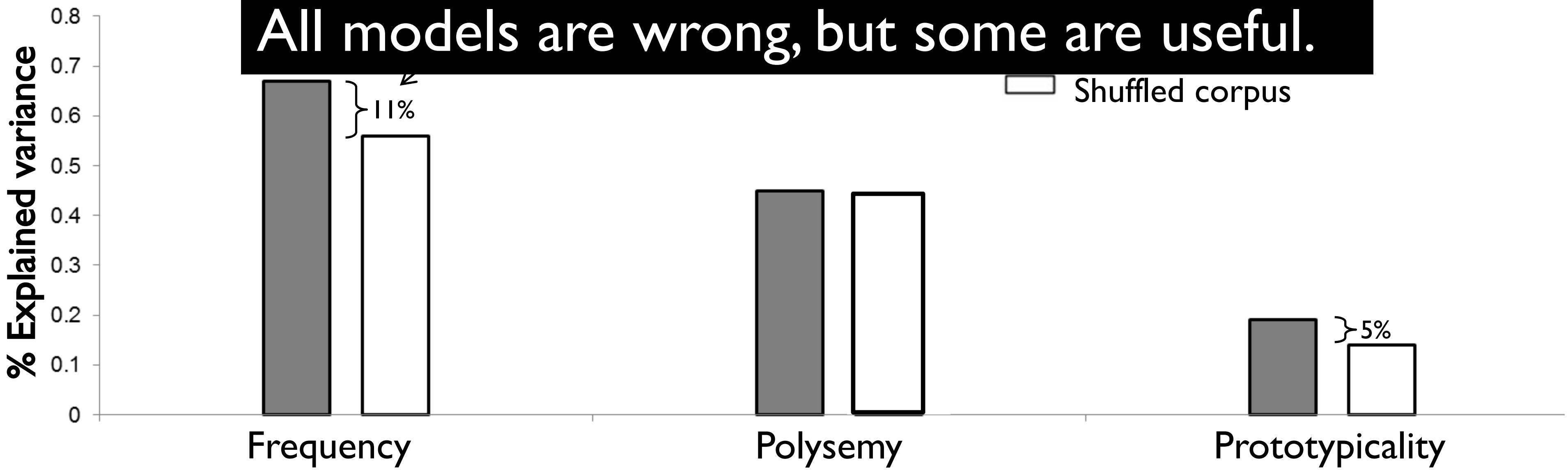
Associative laws of semantic change





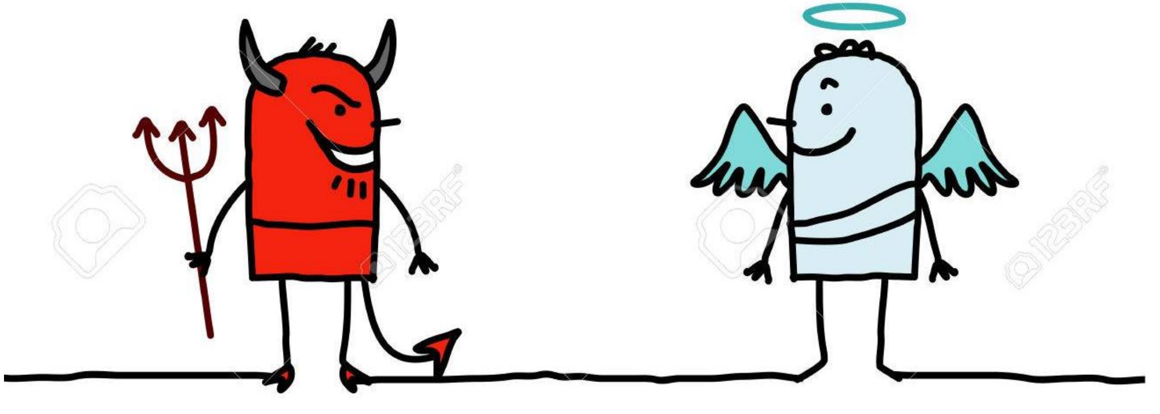


All models are wrong, but some are useful.



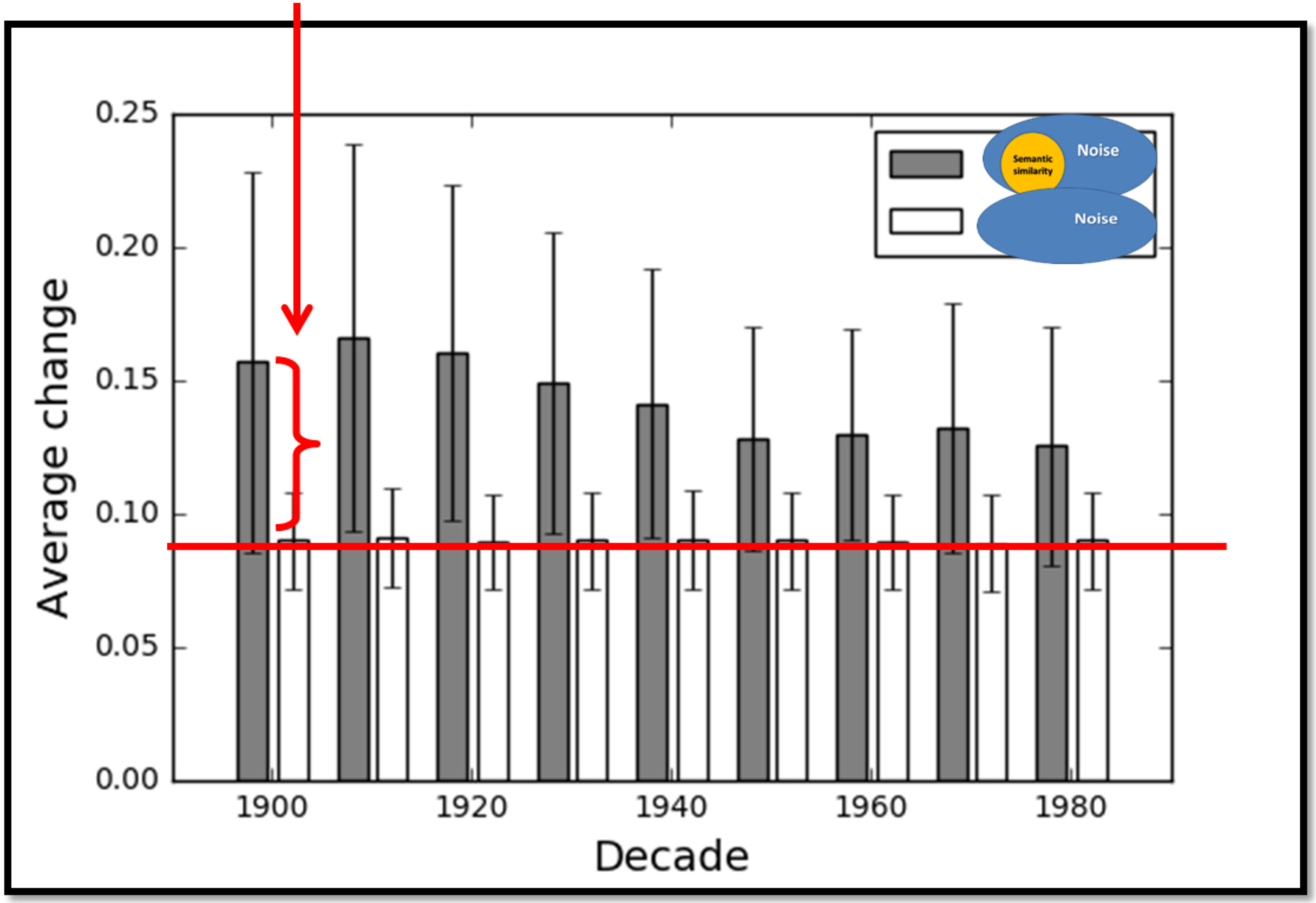
Randomized Controlled Trials
Case II

General framework to compare models' noise levels and quality

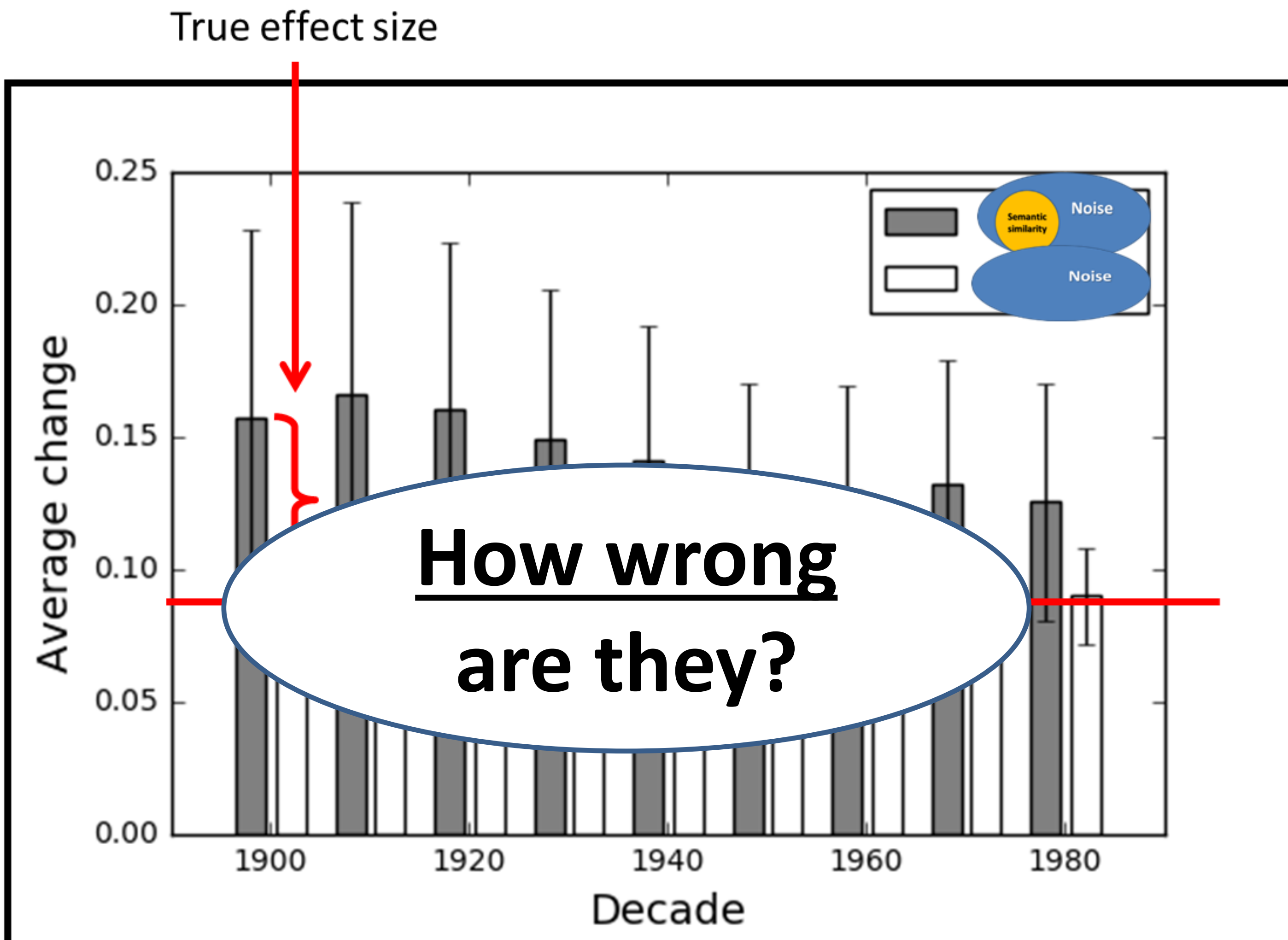


Evaluate noise levels

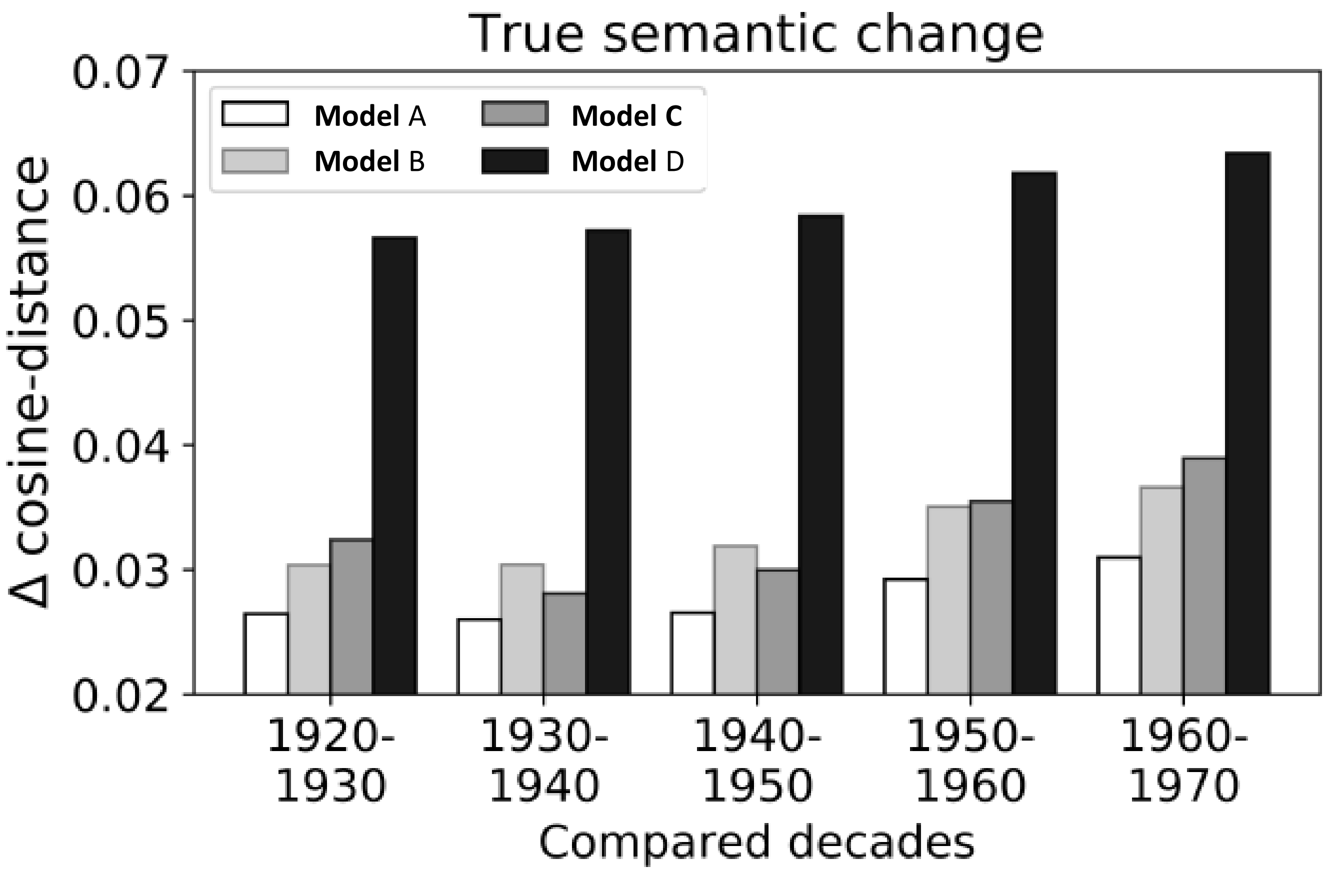
True effect size



Evaluate noise levels

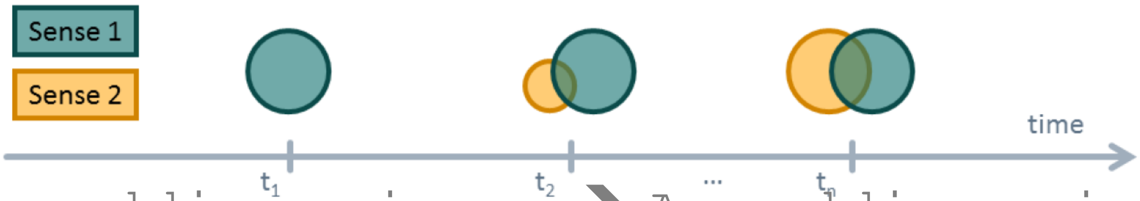


Evaluate noise levels



(Dubossarsky et al. 2019)

Synthetic semantic change

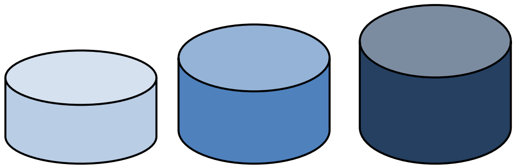


1. A wedding ring → A wedding ring [100%]
   No bracelet!
2. A wedding ring → A wedding ring [100%]
   An arm bracelet → An arm ring [25%]
3. A wedding ring → A wedding ring [100%]
   An arm bracelet → An arm ring [50%]

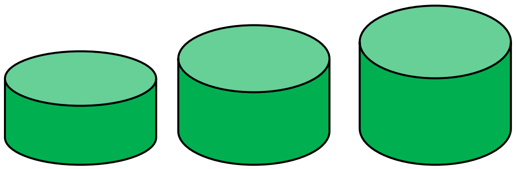
……

4. A wedding ring → A wedding ring [100%]
   An arm bracelet → An arm ring [100%]

Synthetic semantic change



Synthetic change words



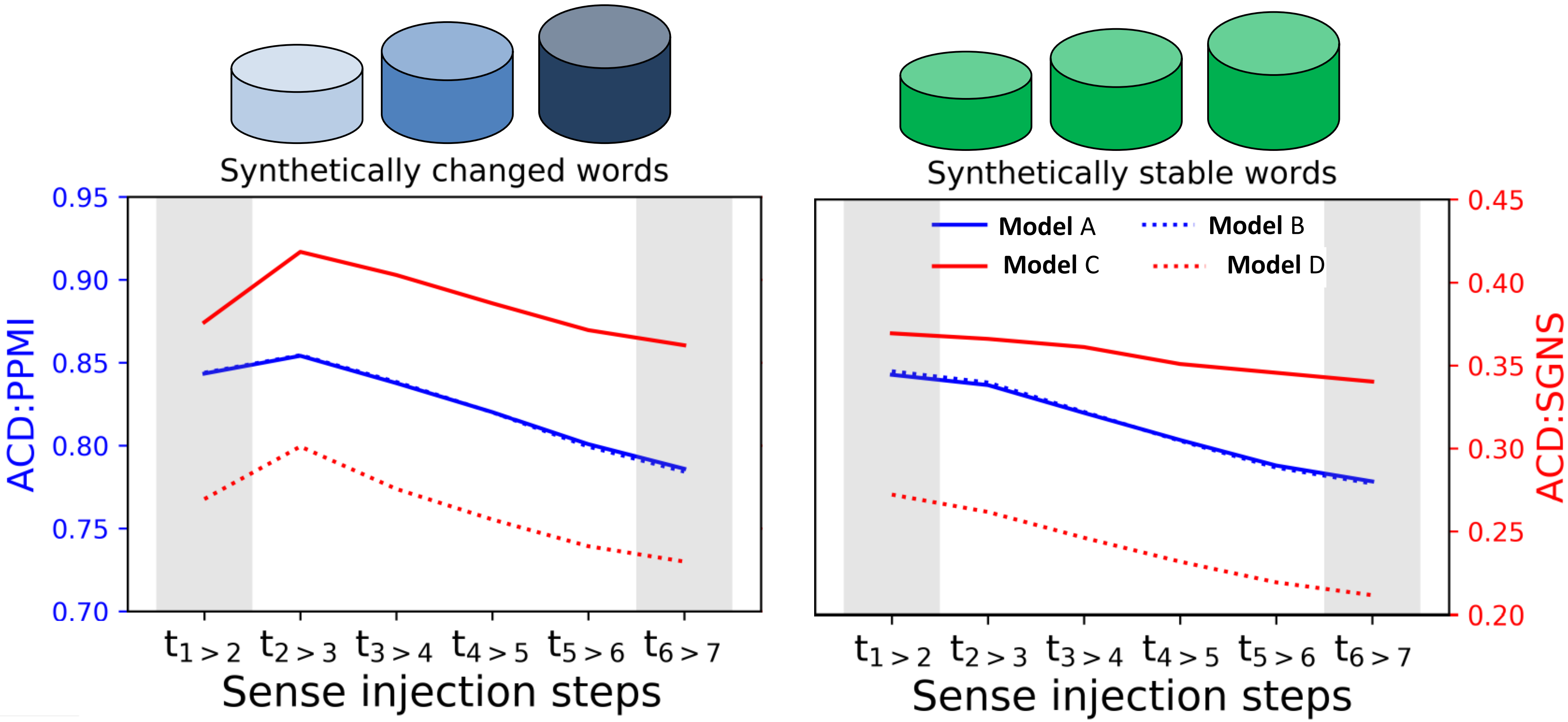
Synthetic stable words

1. A wedding ring → A wedding ring [100%]
   No bracelet!
2. A wedding ring → A wedding ring [100%]
   An arm bracelet → An arm ring [25%]
3. A wedding ring → A wedding ring [100%]
   An arm bracelet → An arm ring [50%]

……

4. A wedding ring → A wedding ring [100%]
   An arm bracelet → An arm ring [100%]

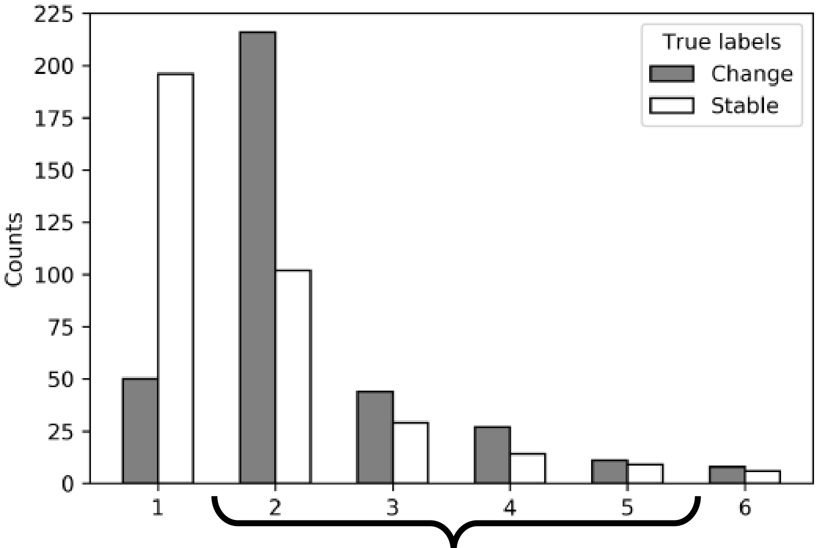
Synthetic semantic change



Evaluate model sensitivity

Are they <u>importantly</u> <u>wrong?</u>

Evaluate model sensitivity



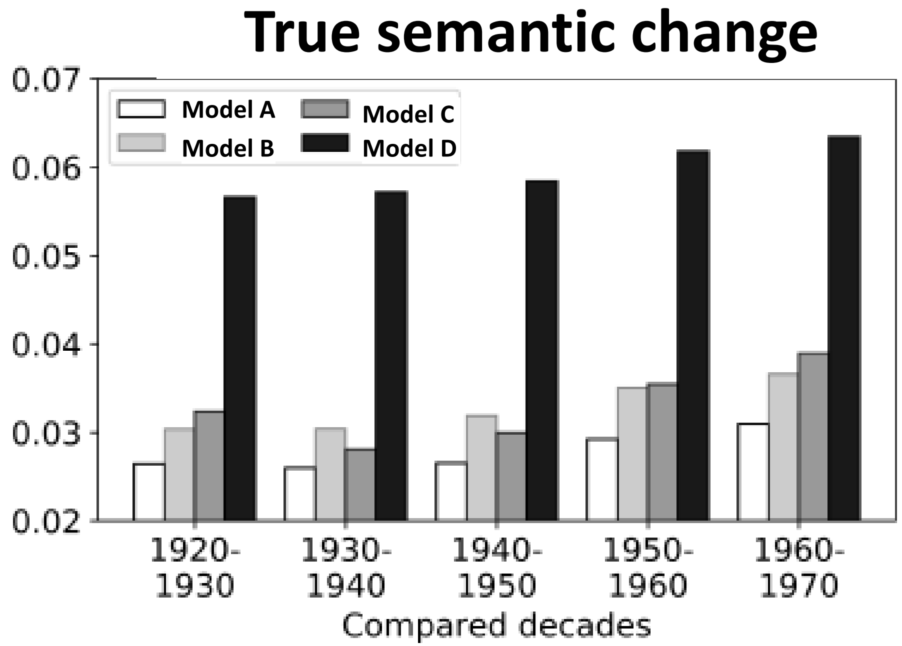
Synthetic change

Naïve classifier

```
if 2=<peak_position=<5:
        semantic_change = True
else:
        semantic_change = False
```

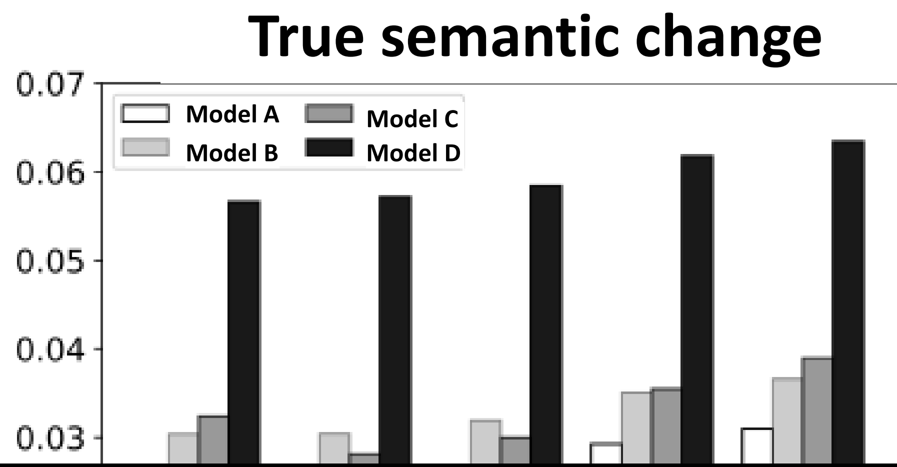
	Model A	**Model** B	**Model** C	**Model** D
accuracy	0.65	0.66	0.59	**0.70**
F1-score	0.69	0.69	0.67	**0.74**

Evaluate model sensitivity



	Model A	**Model** B	**Model** C	**Model** D
accuracy	0.65	0.66	0.59	**0.70**
F1-score	0.69	0.69	0.67	**0.74**

Evaluate model sensitivity



All models are wrong, but some are useful.
And some are more useful than other!

	Model A	**Model** B	**Model** C	**Model** D
accuracy	0.65	0.66	0.59	**0.70**
F1-score	0.69	0.69	0.67	**0.74**

Conclusions

<u>Test your models!</u>

- Use randomized control tests to evaluate levels of noise and alleviate confounds in models.
- Simulate the phenomenon you are investigating.
- Test models' performance on simulated data.
- Not limited to word embedding!

Doing it right

+

Historical distributional semantics

Doing it right

Historical distributional semantics

Credits

- Dubossarsky et al. 2015:

  Chris Dyer, Yulia Tsvetkov and Eitan Grossman

- Dubossarsky et al. 2017:

  Eitan Grossman and Daphna Weinshall

- Dubossarsky et al. 2019:

  Simon Hengchen - University of Helsinki

  Nina Tahmasebi - University of Gothenburg

  Dominik Schlechtweg - University of Stuttgart

Thank you!

SemEval-2020. Coming soon…

hd423@cam.ac.uk