UNIVERSITY OF CAMBRIDGE

Doubt thy models: rethinking hypothesis testing in NLP

Haim Dubossarsky

University of Stuttgart, IMS

October 2020

hd423@cam.ac.uk

What is a good model ?

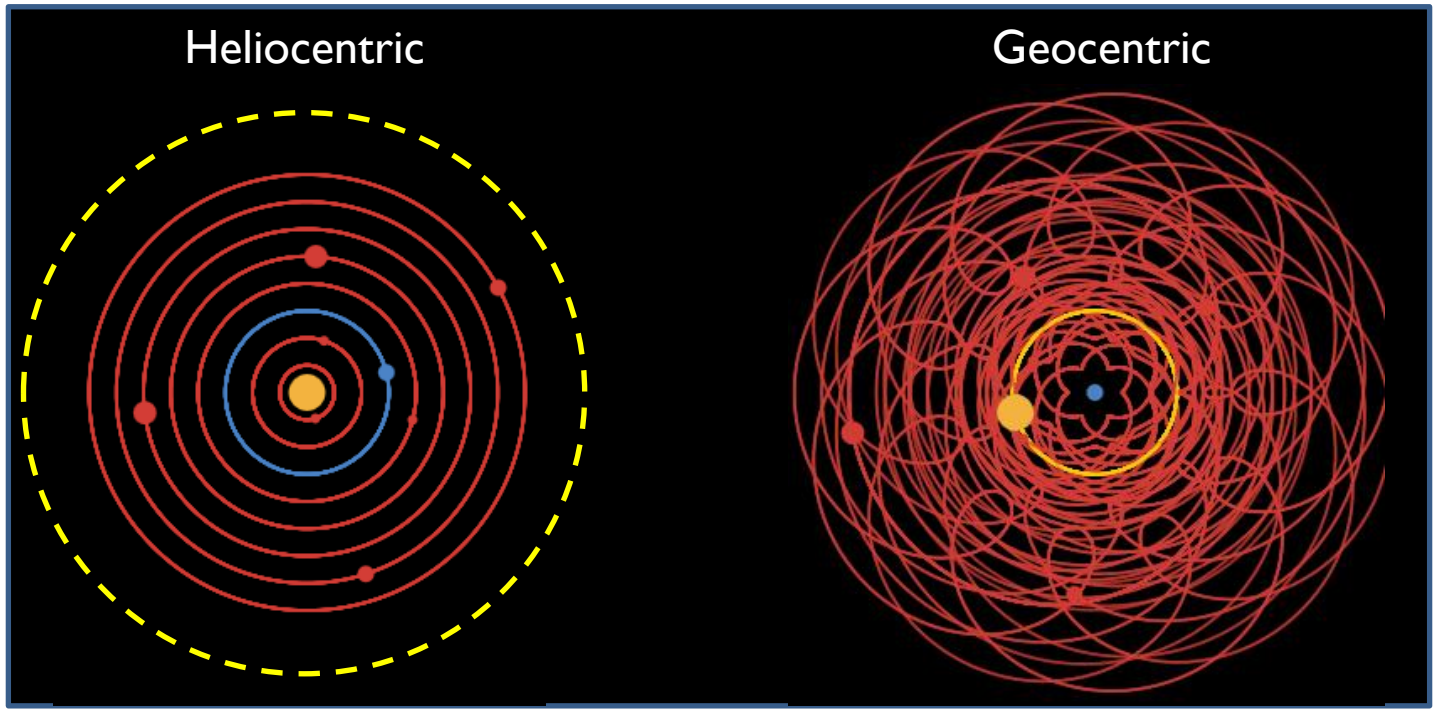


Image taken from Gfycat

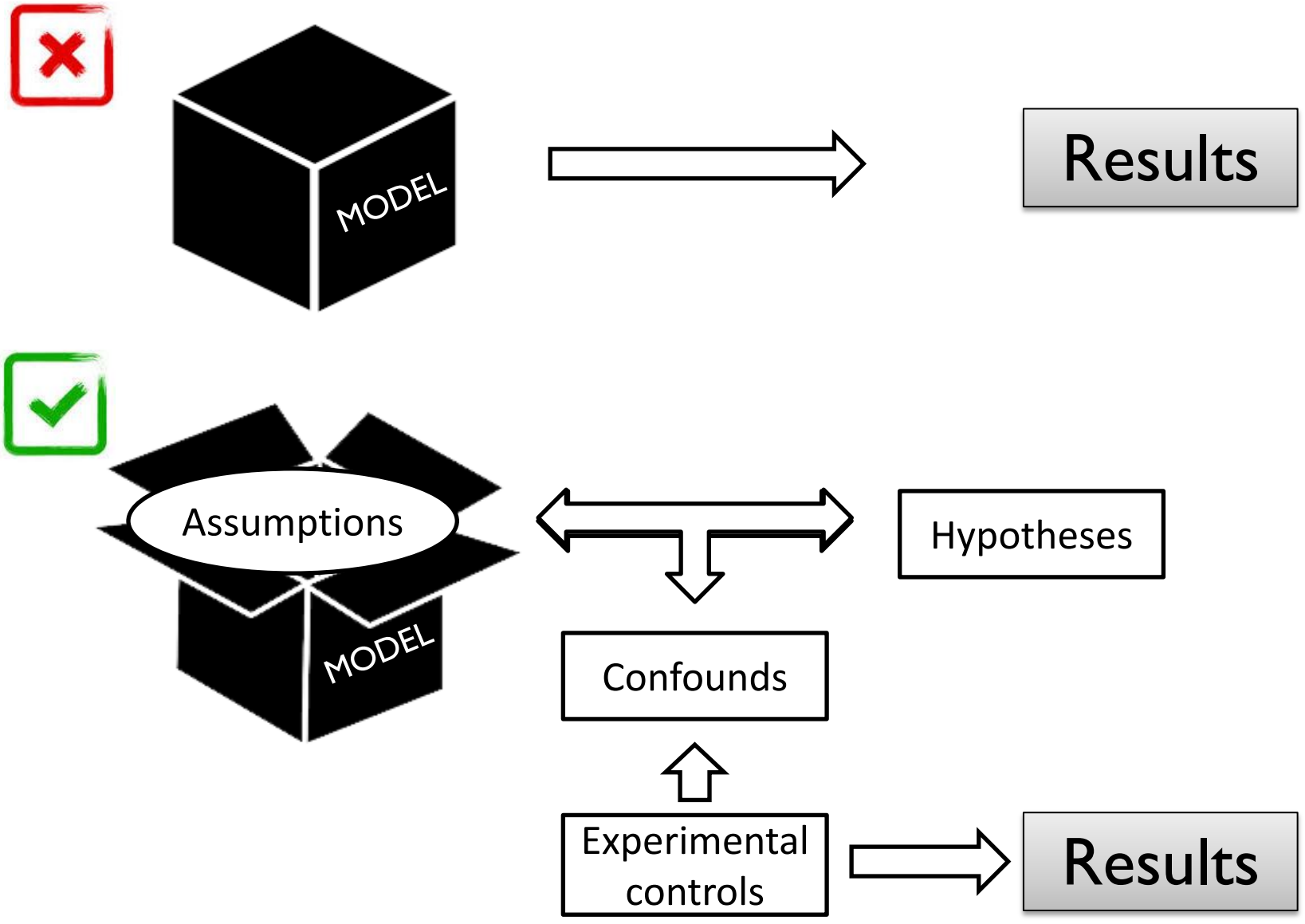
Accuracy?

All models are wrong

Facilitate hypotheses and theories!

All models are wrong, but some are useful.

A different view of models

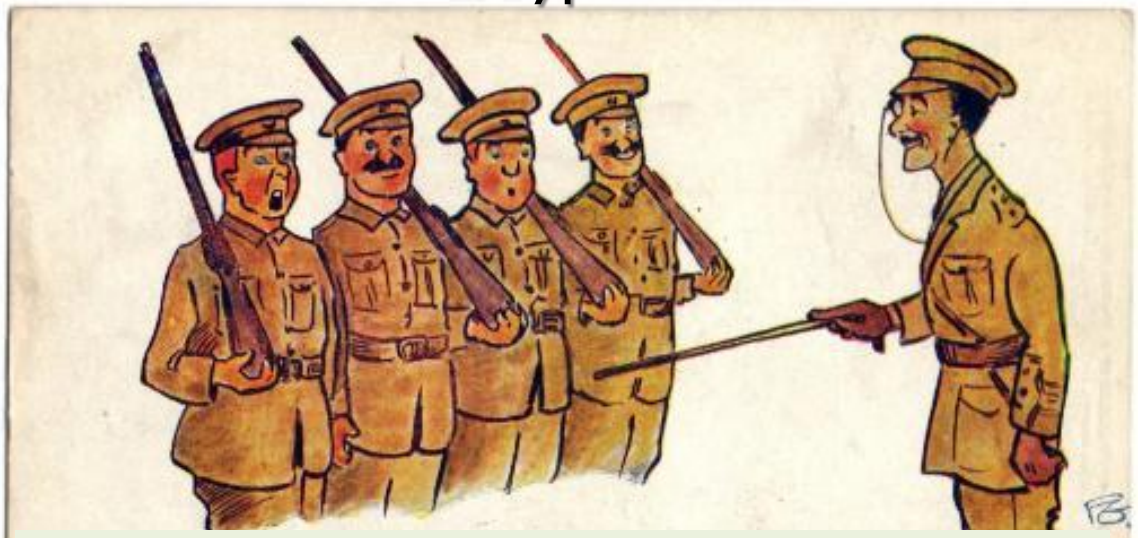


Outline

- ✓ Agenda
- Model validation
- Part I – Problems in semantic change models
- Part II - Working with and improving faulty models
- Conclusions

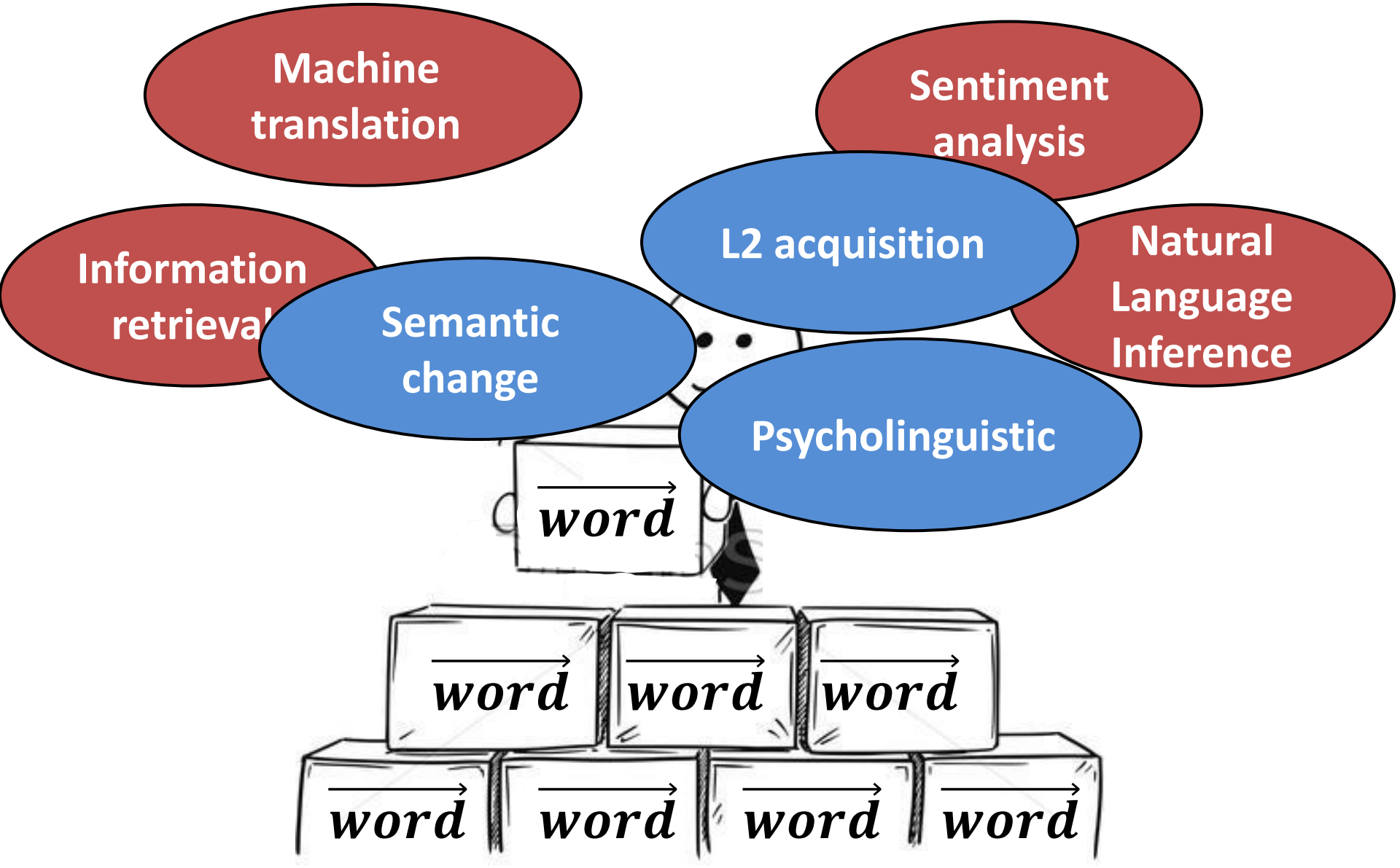
Word meaning representation

The distributional hypothesis



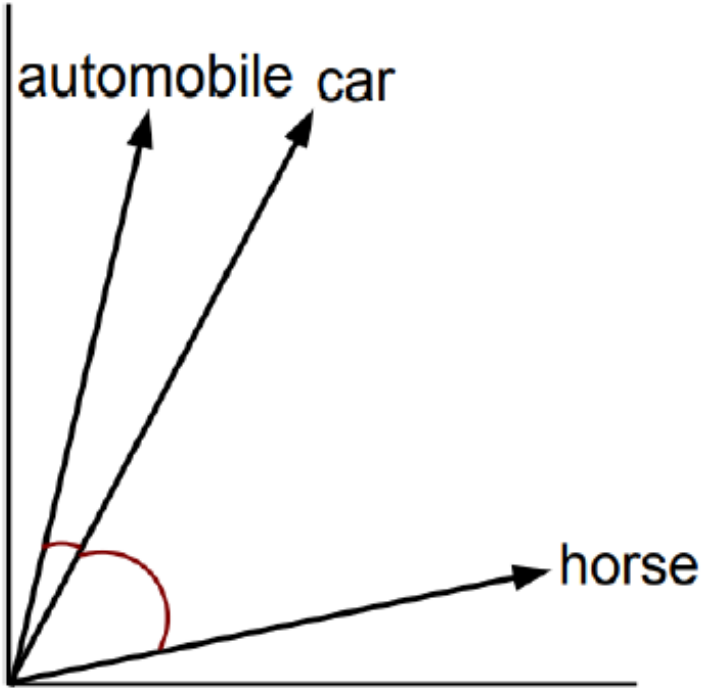
You shall know a word by the company it keeps (Firth, J. R. 1957:11)

Word meaning representation



Model validation (embedding)

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

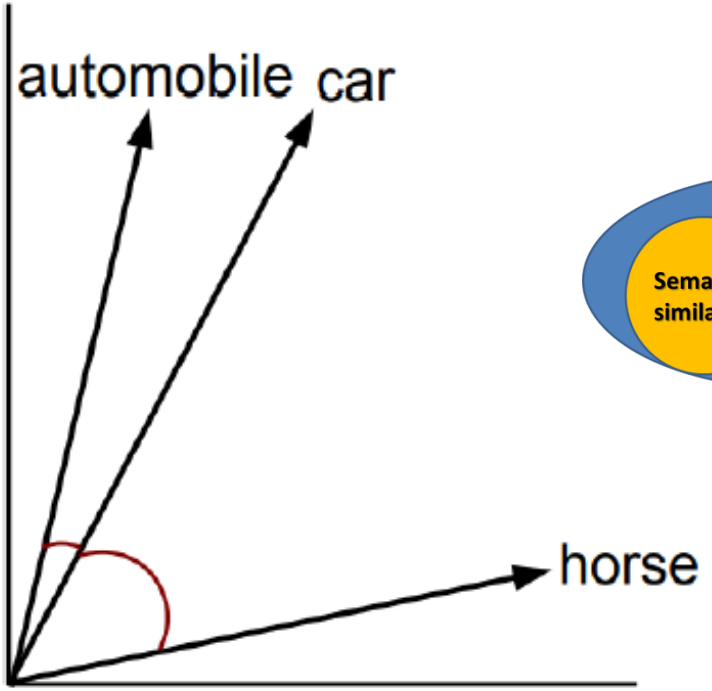


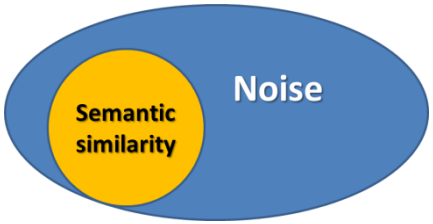
r=.72

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

Model validation (embedding)

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

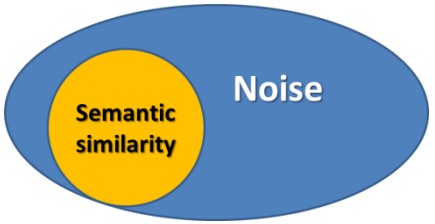
✓ Vectors capture semantic meaning

≠ Vectors capture only semantic meaning

Model validation (embedding)

What is noise?
An artifact.

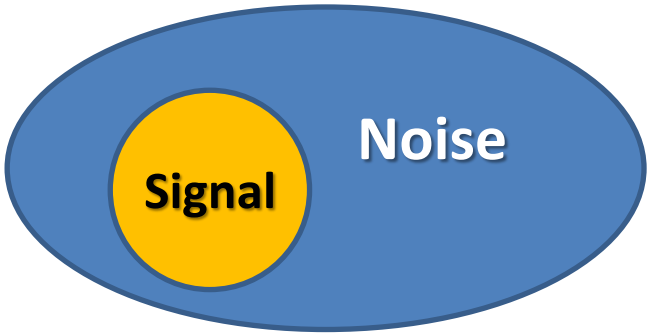
Any unwanted variable that influence our measurment.

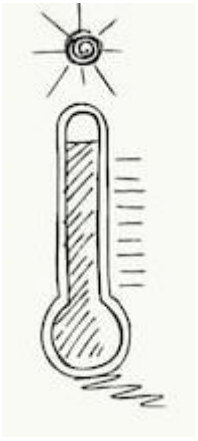


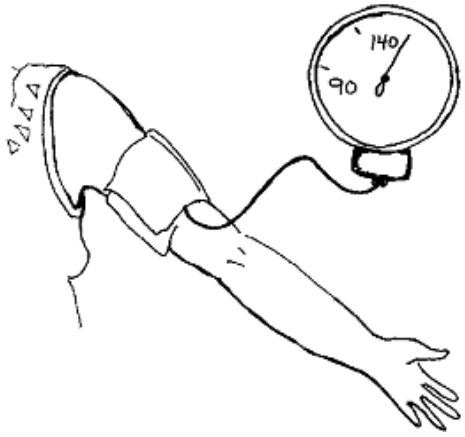
✓ Vectors capture semantic meaning

≠ Vectors capture only semantic meaning

All models are wrong

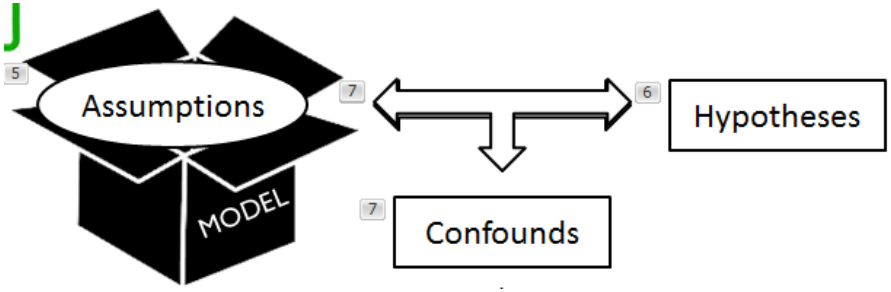






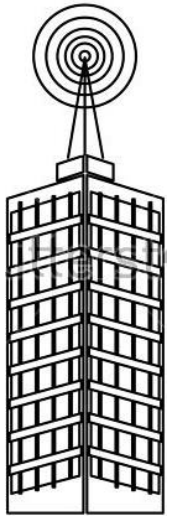
Aspects of wrongness

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



Part I
Problems in semantic change models

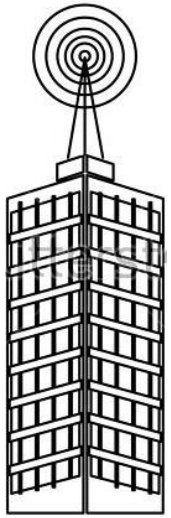
Based on Dubossarsky et al. 2017



Measuring semantic change

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

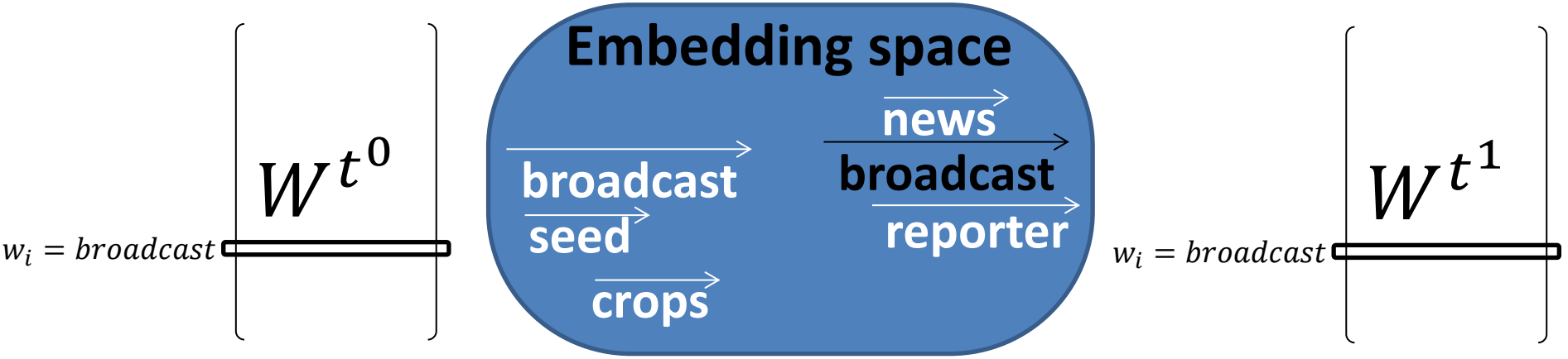
$$\Delta w^{t^0 \rightarrow t^1} = cosDist(w^{t_0}, w^{t_1}) = 1 - \frac{\vec{w}^{t_0} \cdot \vec{w}^{t_1}}{\|\vec{w}^{t_0}\| \cdot \|\vec{w}^{t_1}\|}$$



Measuring semantic change

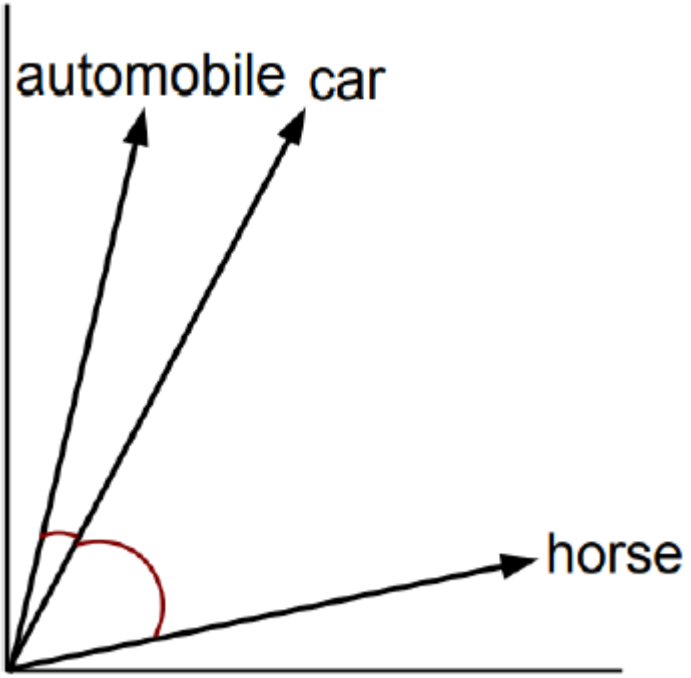
Change to a word's representation* between two time points [word relative to itself]

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



Semantic change validated?

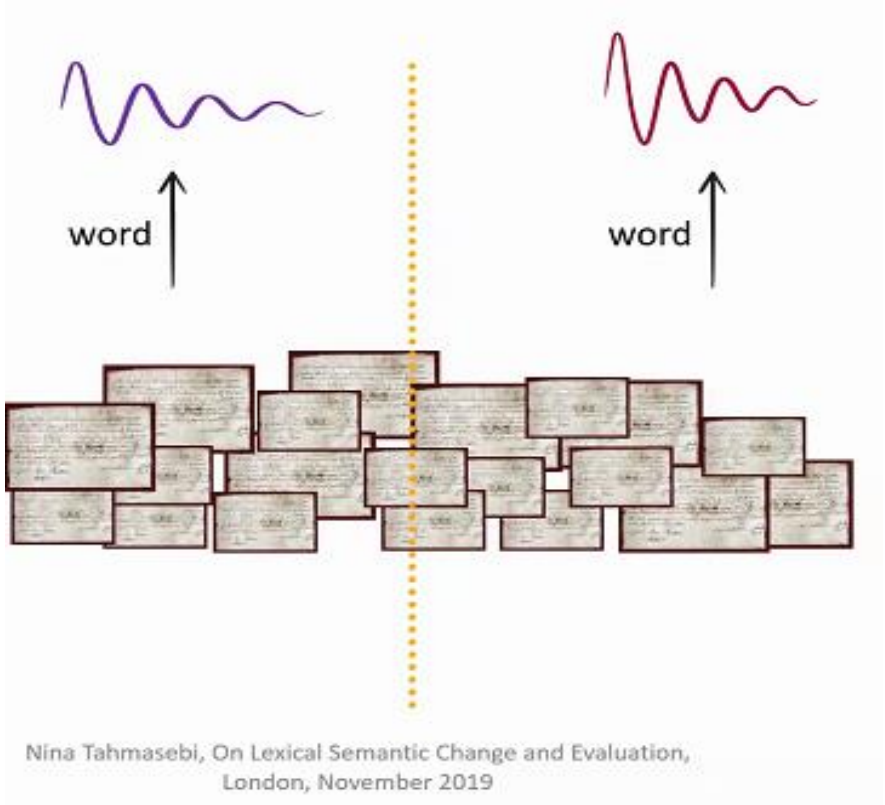
Word 1	Word 2	Human	Embeddin
horse	car	5.9	0.79
book	paper	7.46	0.85
computer	keyboard	7.62	0.79
train	car	6.31	0.5
television	radio	6.77	0.73
drug	abuse	6.85	0.45
bread	butter	6.19	0.65
cucumber	potato	5.92	0.75
doctor	nurse	7	0.84
smart	stupid	5.81	0.6
stock	market	8.08	0.97



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

Semantic change validated?

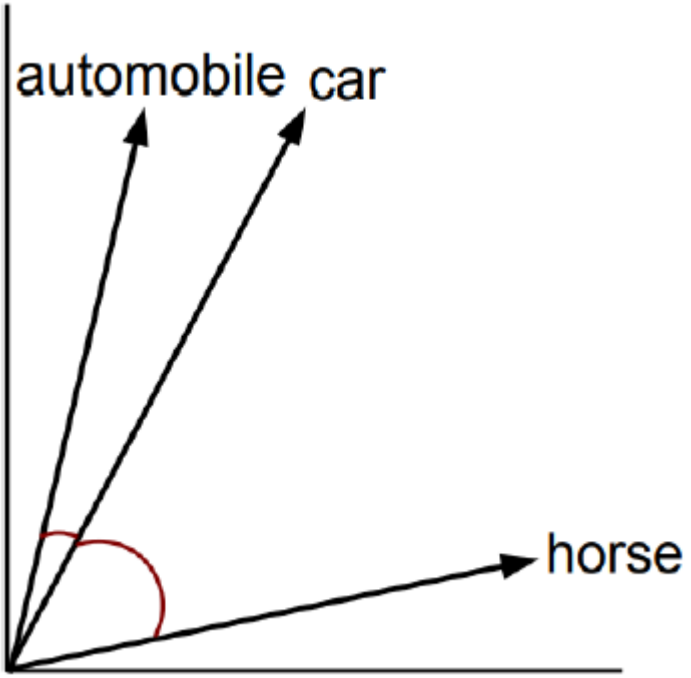
Word 1	Word 2	Human	Embeddin
horse	car	5.9	0.79
book	paper	7.46	0.85
computer	keyboard	7.62	0.79
train	car	6.31	0.5
television	radio	6.77	0.73
drug	abuse	6.85	0.45
bread	butter	6.19	0.65
cucumber	potato	5.92	0.75
doctor	nurse	7	0.84
smart	stupid	5.81	0.6
stock	market	8.08	0.97



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

Semantic change validated?

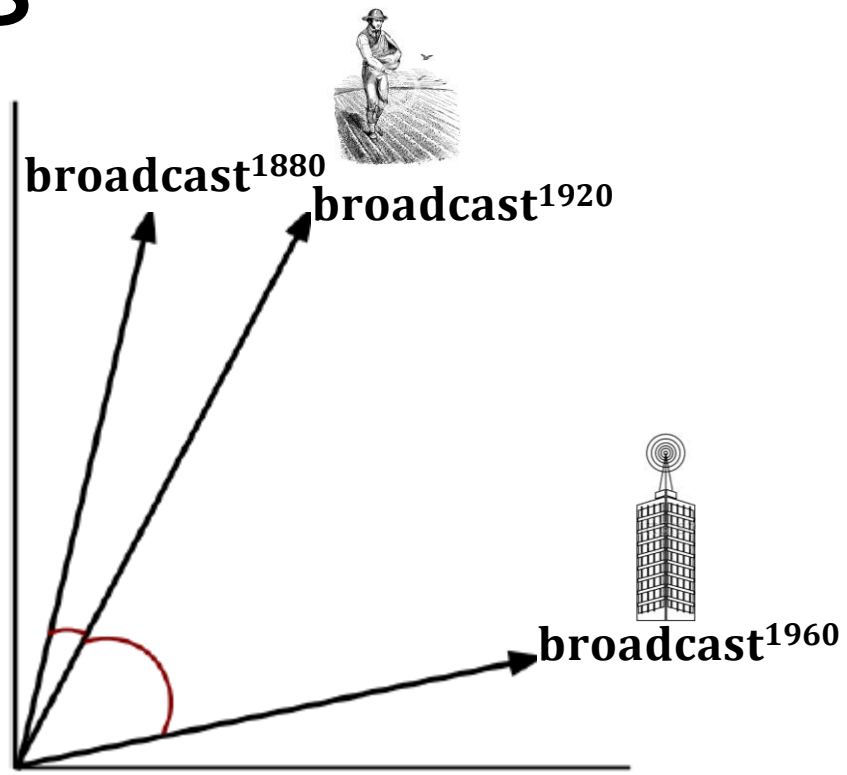
Word 1	Word 2	Human	Embeddin
horse	car	5.9	0.79
book	paper	7.46	0.85
computer	keyboard	7.62	0.79
train	car	6.31	0.5
television	radio	6.77	0.73
drug	abuse	6.85	0.45
bread	butter	6.19	0.65
cucumber	potato	5.92	0.75
doctor	nurse	7	0.84
smart	stupid	5.81	0.6
stock	market	8.08	0.97



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

Semantic change validated?

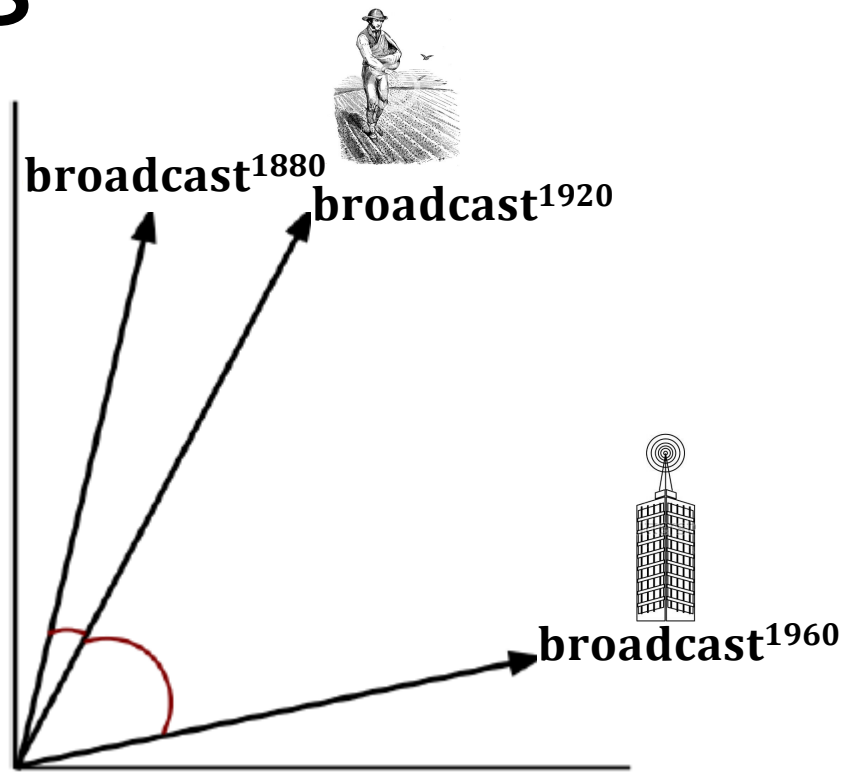
Word 1	Word 2	Human	Embeddin
horse	car	5.9	0.79
book	paper	7.46	0.85
computer	keyboard	7.62	0.79
train	car	6.31	0.5
television	radio	6.77	0.73
drug	abuse	6.85	0.45
bread	butter	6.19	0.65
cucumber	potato	5.92	0.75
doctor	nurse	7	0.84
smart	stupid	5.81	0.6
stock	market	8.08	0.97



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

Semantic change validated?

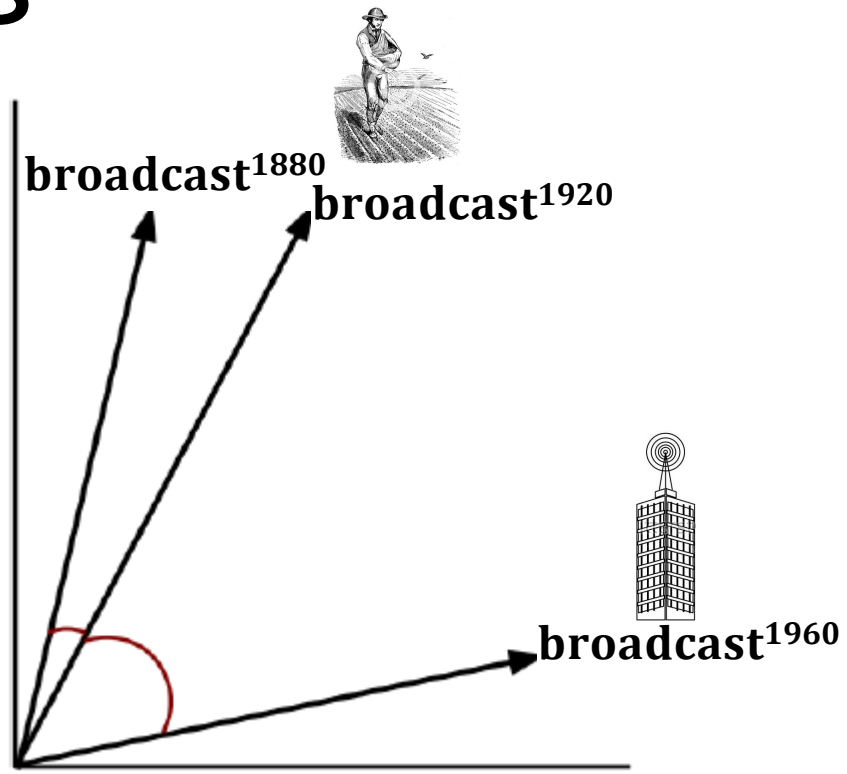
Word 1	Word 2	Human	Embeddin
horse	car	5.9	0.79
book	paper	7.46	0.85
computer	keyboard	7.62	0.79
train	car	6.31	0.5
television	radio	6.77	0.73
drug	abuse	6.85	0.45
bread	butter	6.19	0.65
cucumber	potato	5.92	0.75
doctor	nurse	7	0.84
smart	stupid	5.81	0.6
stock	market	8.08	0.97



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

Semantic change validated?

Word 1	Word 2	Human	Embeddin
horse	car	5.9	0.79
book	paper	7.46	0.85
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

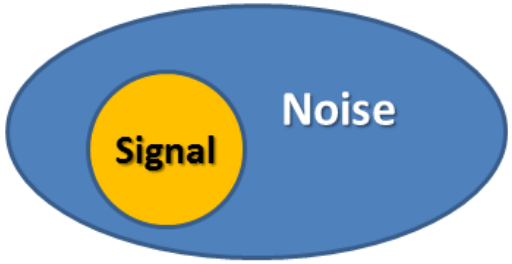


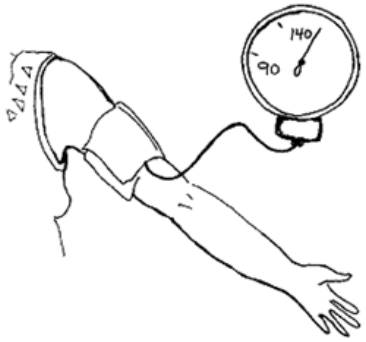
$$cosine\ similarity(w^{t1}, w^{t2}) = \frac{\vec{w}^{t1} \cdot \vec{w}^{t2}}{\|\vec{w}^{t1}\| \cdot \|\vec{w}^{t2}\|}$$

Luckily we have SemEval-2020 (SemEval 2020)

How wrong models are?

All models are wrong

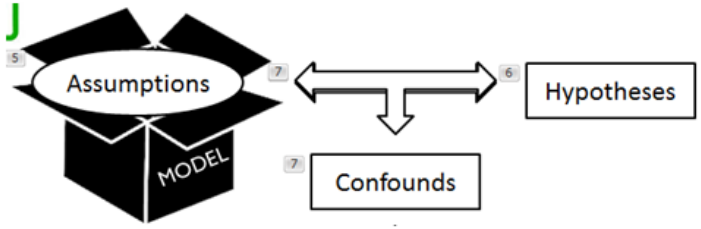




old

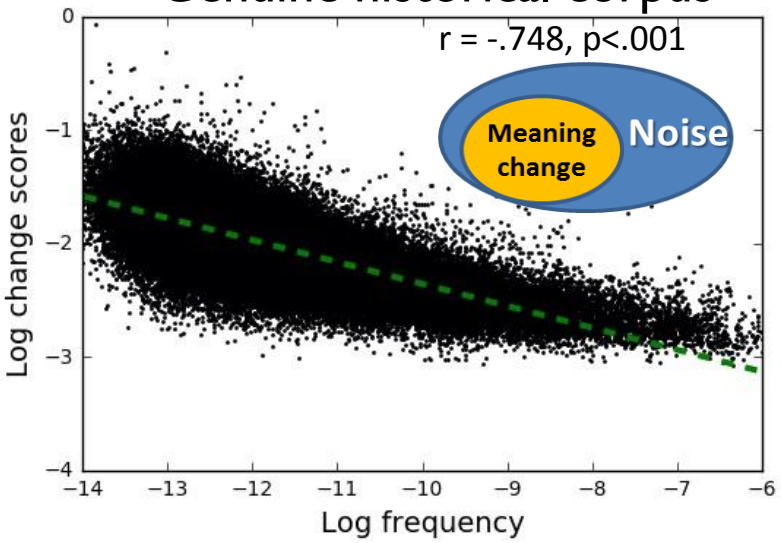
Aspects of wrongness

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

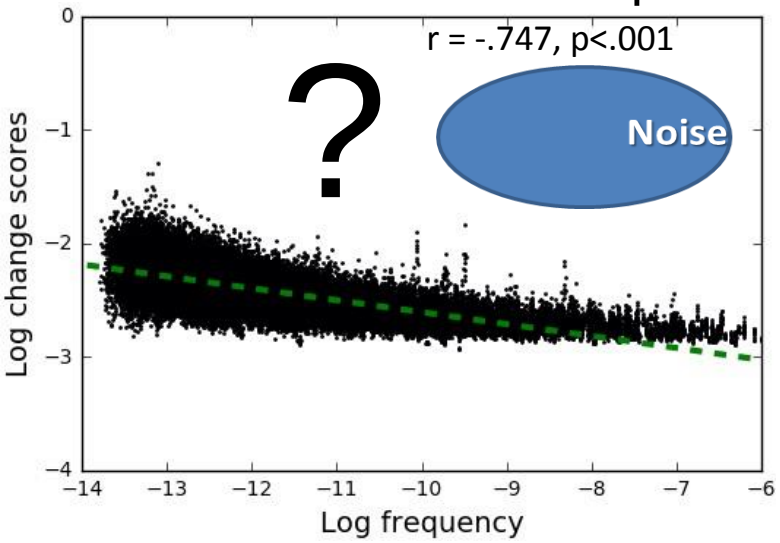


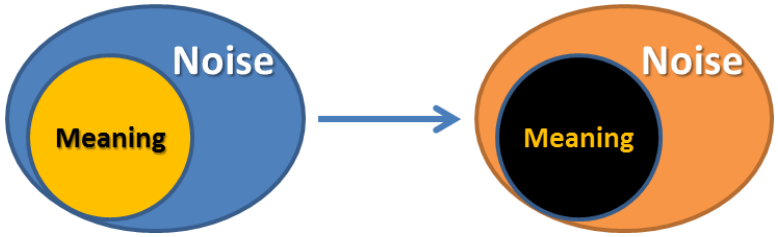
Are they importantly wrong?

Genuine historical corpus



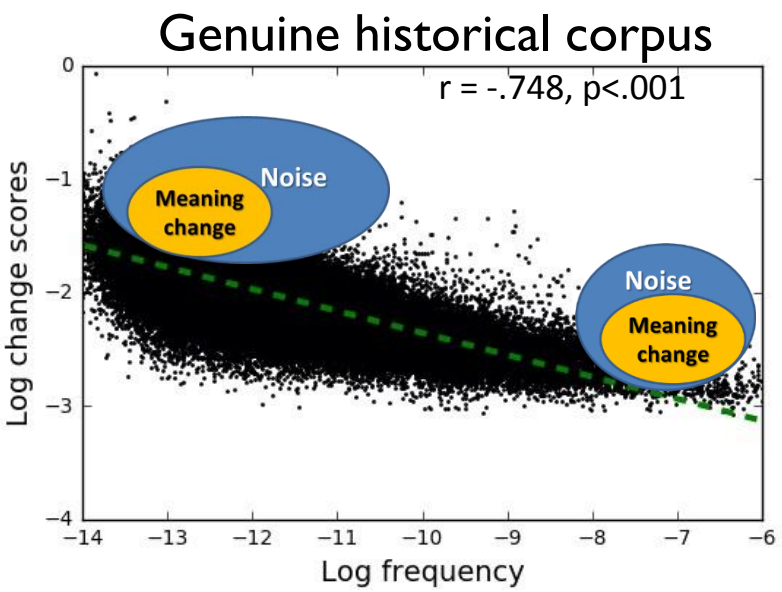
Shuffled historical corpus

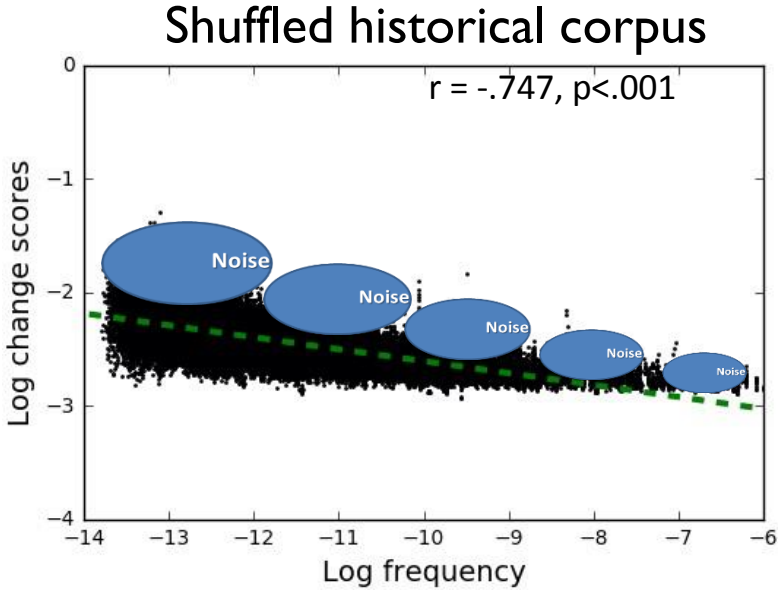


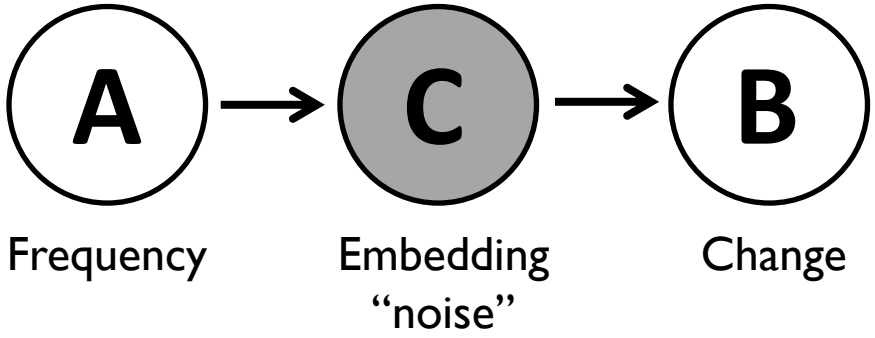


t^0 t^1

The artefact is a confound

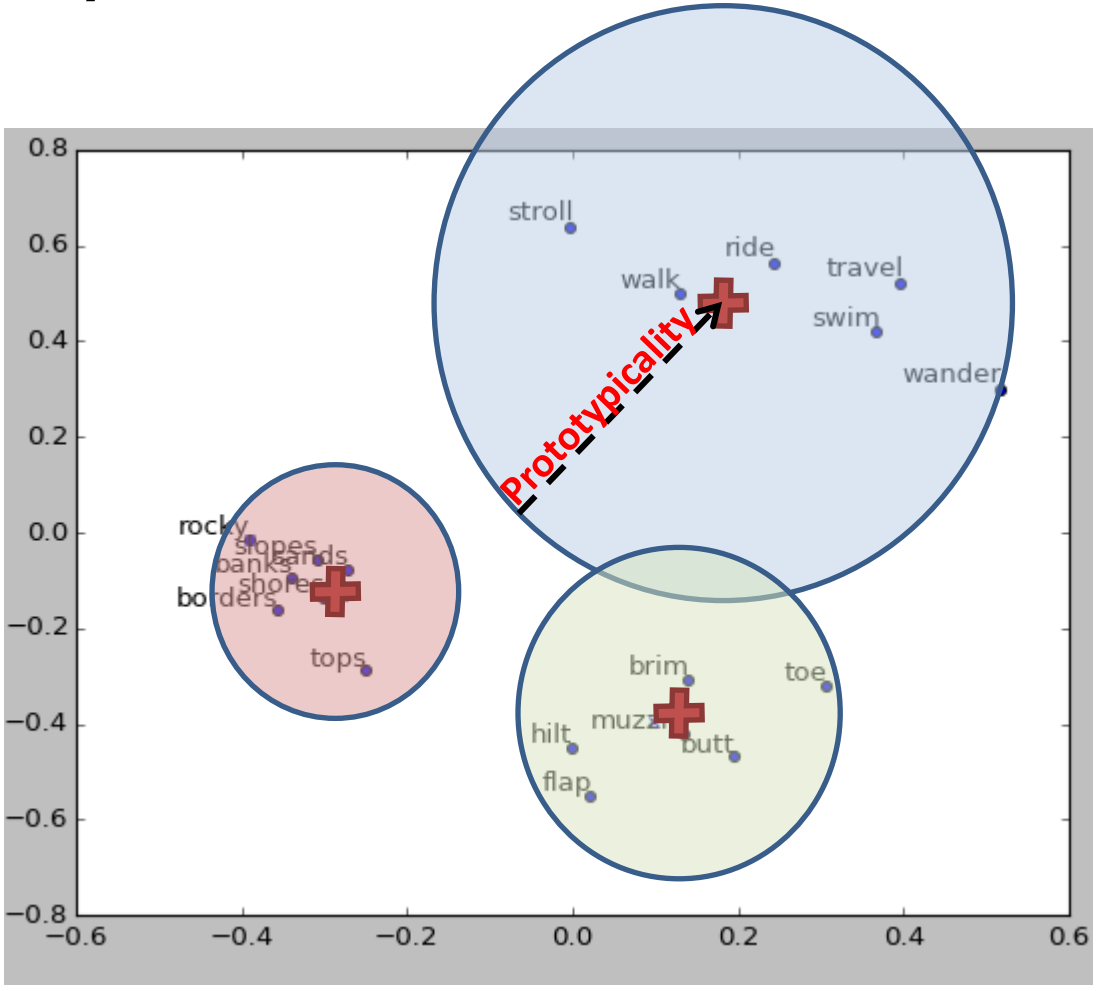






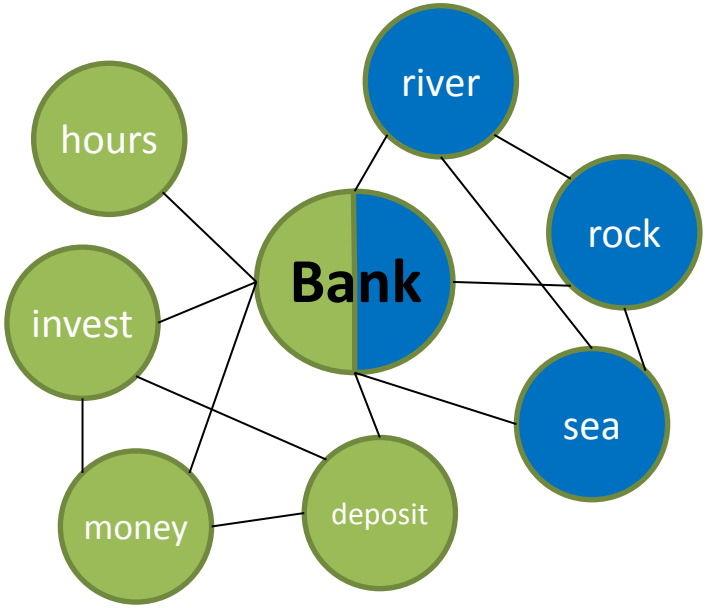
"Laws" of semantic change

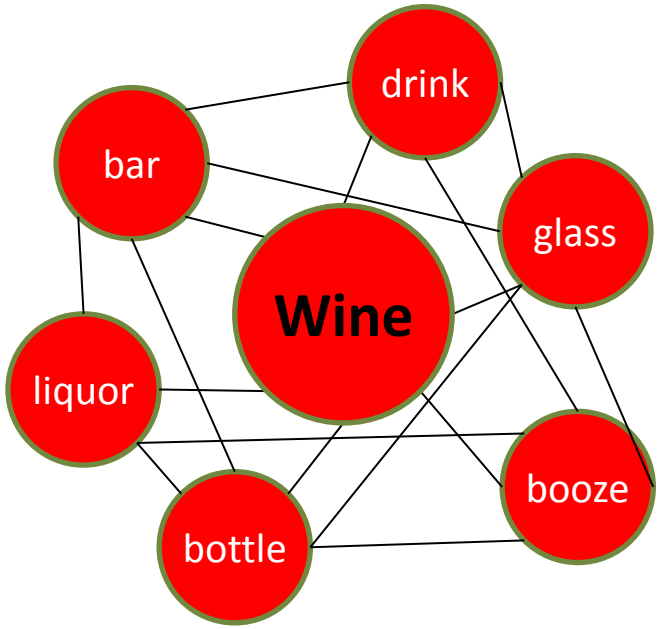
- Law of Prototypicality (Dubossarsky et. al. 2015).



"Laws" of semantic change

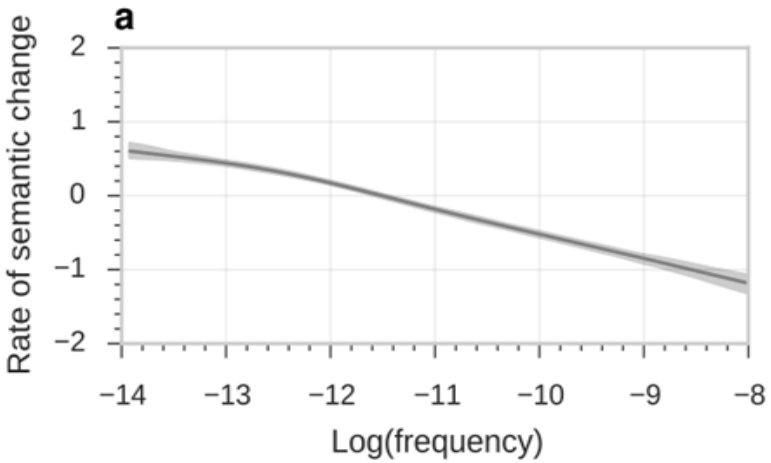
- Law of Prototypicality (Dubossarsky et. al. 2015).
- Law of Innovation (Polysemy, Hamilton et. al. 2016).



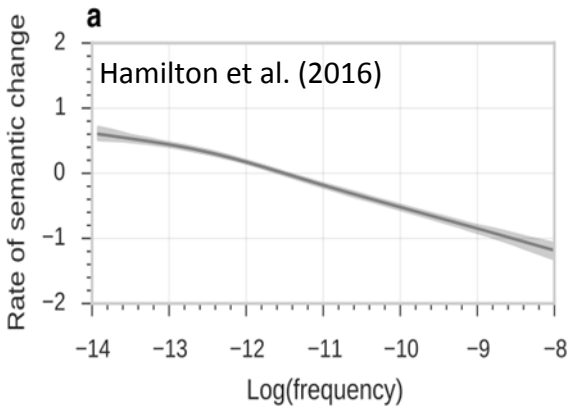


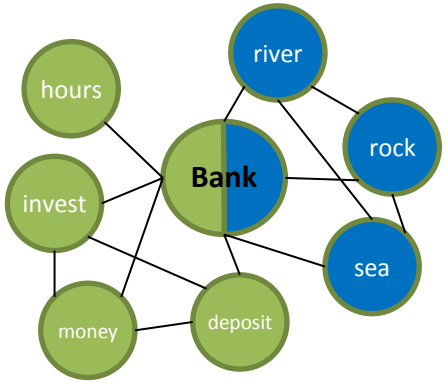
"Laws" of semantic change

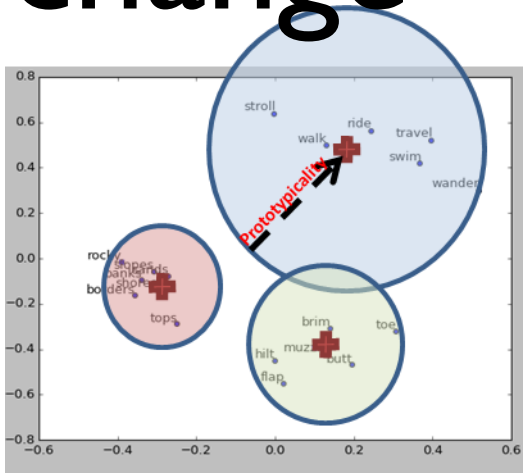
- Law of Prototypicality (Dubossarsky et. al. 2015).
- Law of Innovation (Polysemy, Hamilton et. al. 2016).
- Law of Conformity (Frequency, Hamilton et. al. 2016).



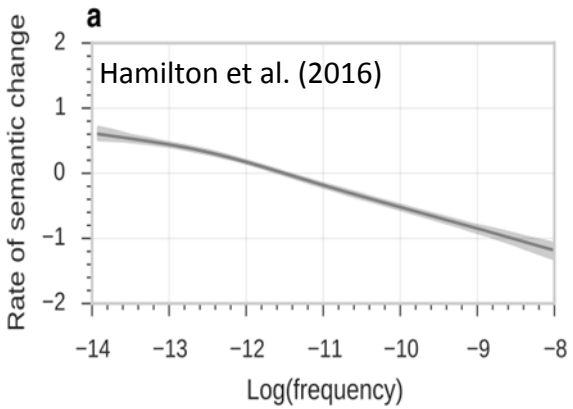
"Laws" of semantic change

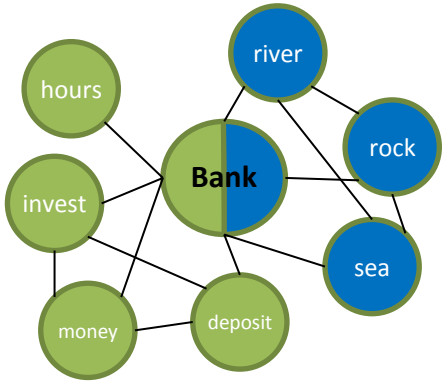


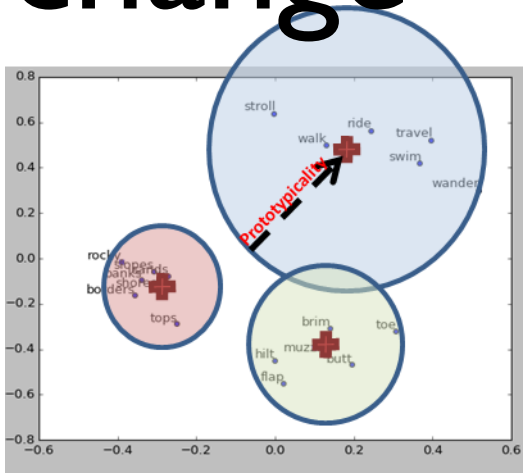




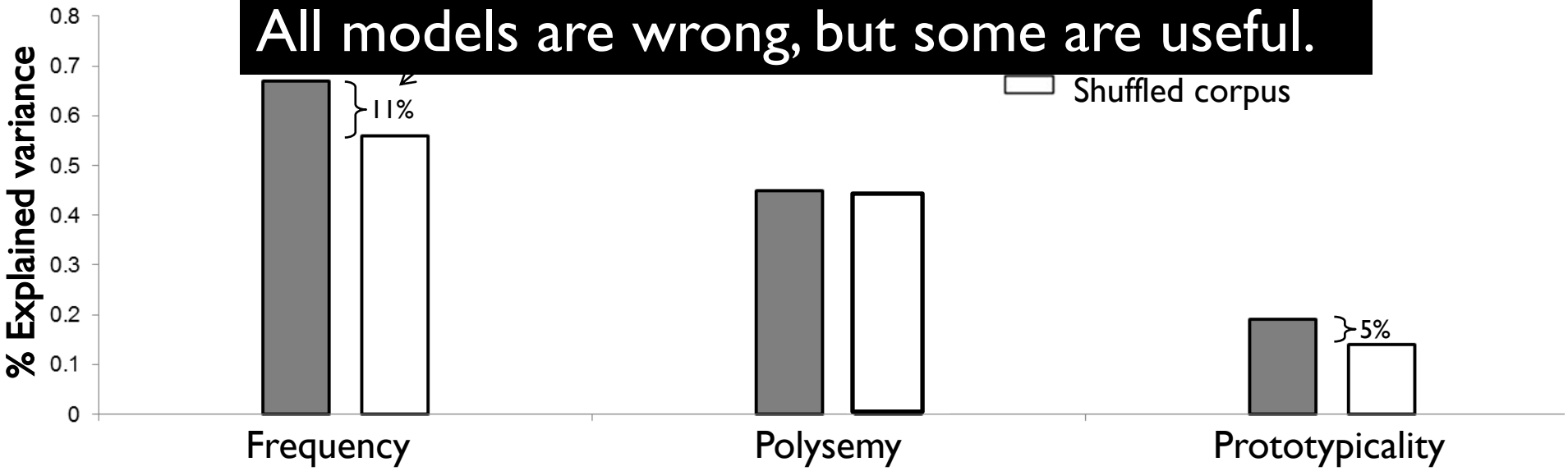
"Laws" of semantic change





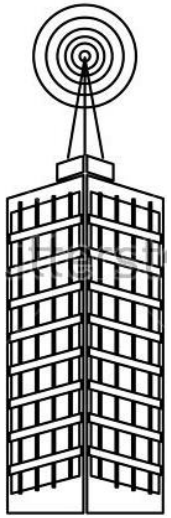


All models are wrong, but some are useful.

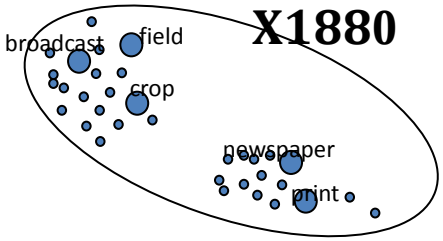


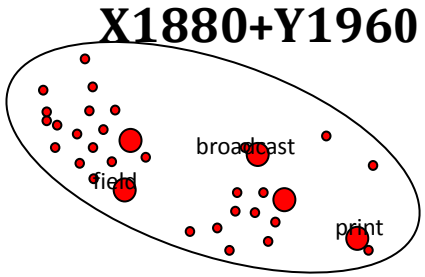
Part II
Working with and improving faulty models

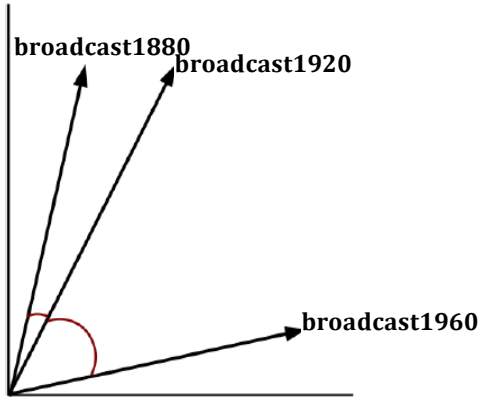
based on Dubossarsky el al. (2019)

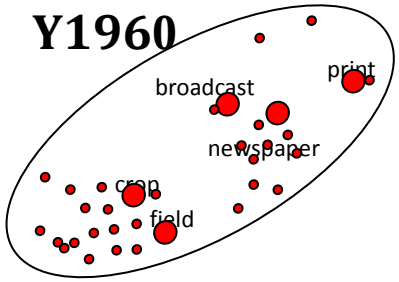


Temporal Referencing





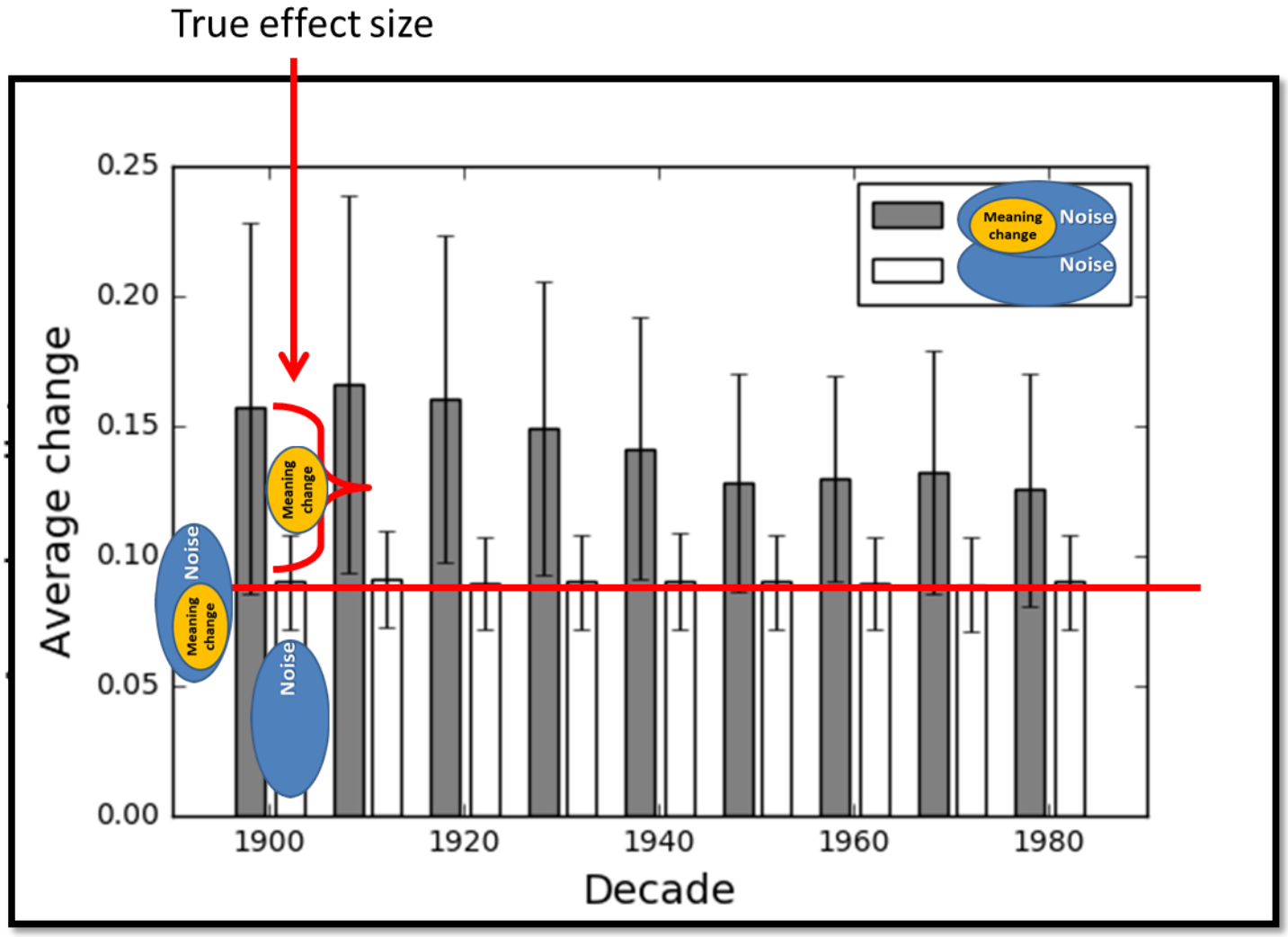




Example

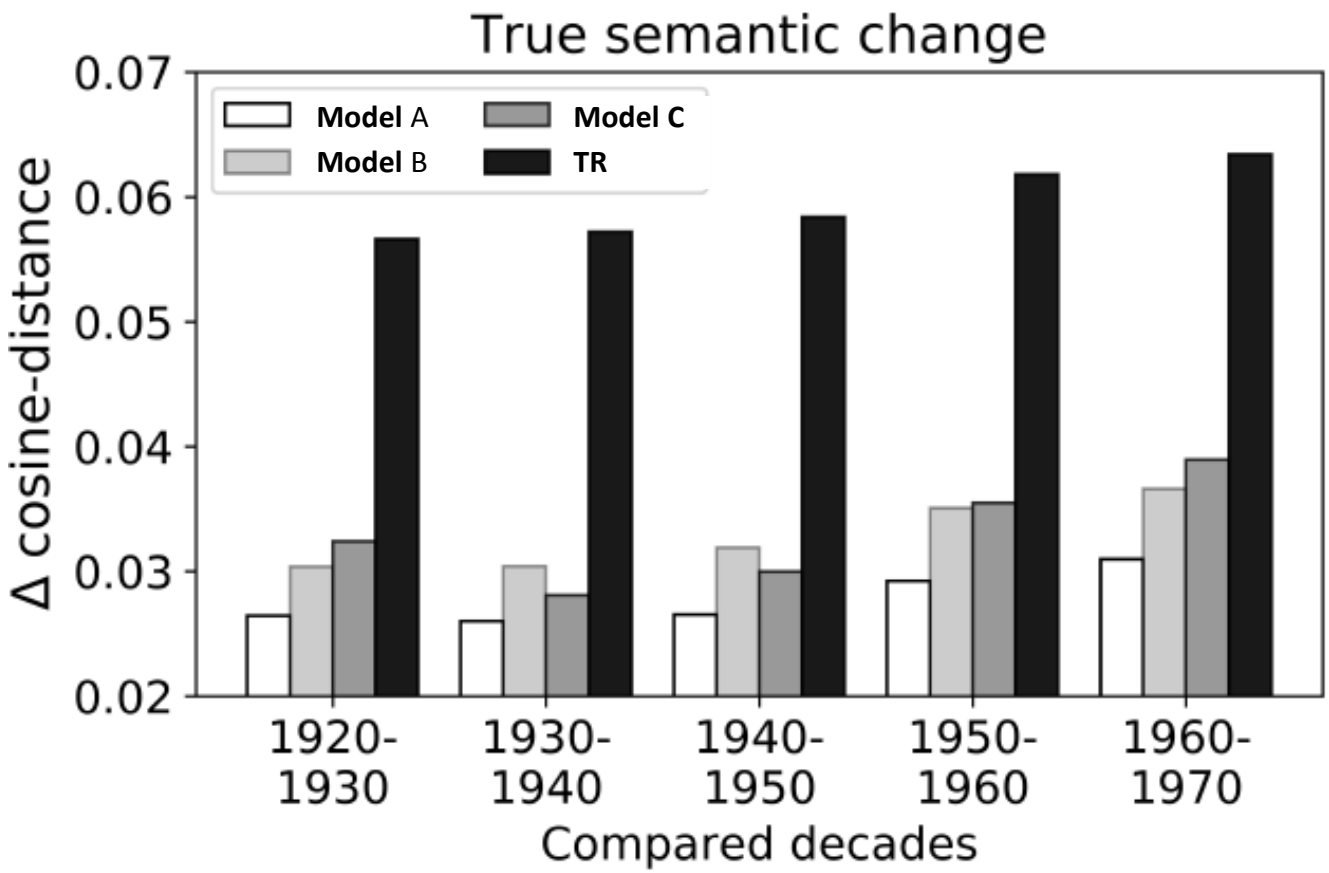
Silken cauliflowers sown broadcast[1870] over the land.
The dramatic broadcast[1970] stunned the nation.

Evaluate noise levels



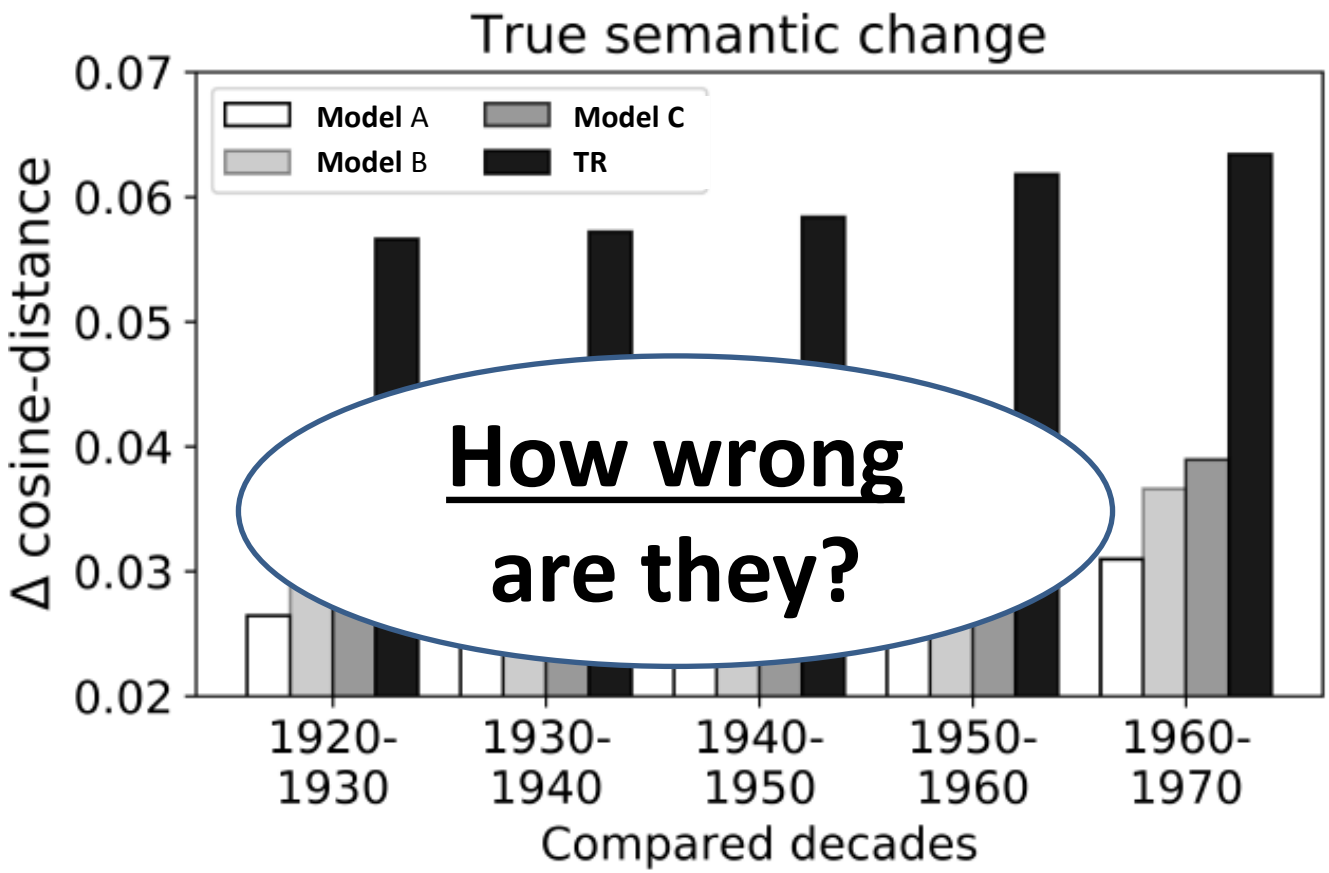
Evaluate noise levels

Meaning change



Evaluate noise levels

Meaning change



How wrong are they?

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%]

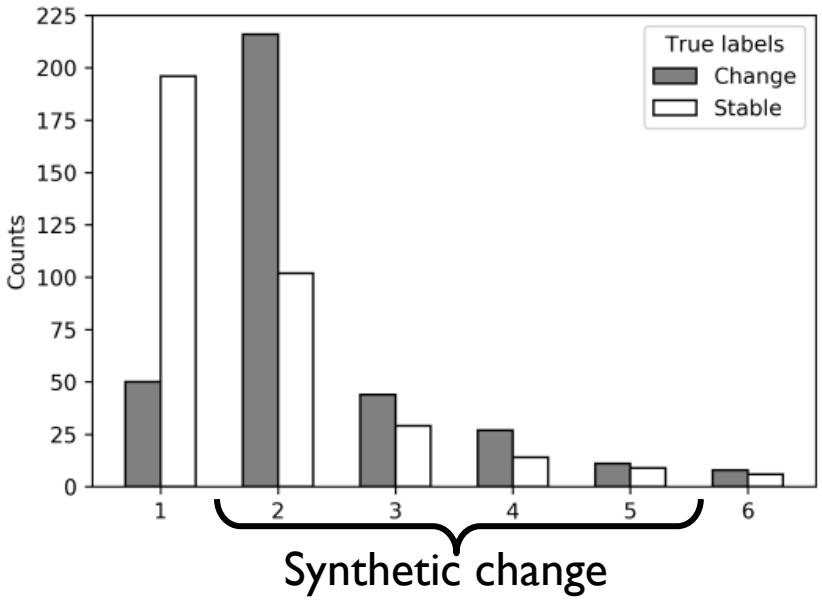
.....

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

Evaluate model sensitivity

Are they importantly wrong?

Evaluate model sensitivity

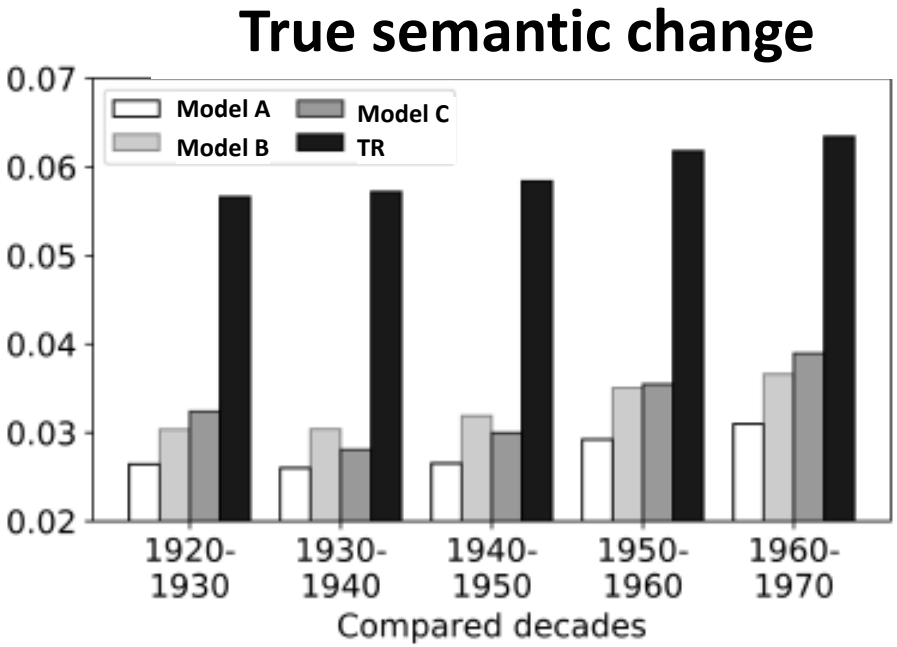


Naïve classifier

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

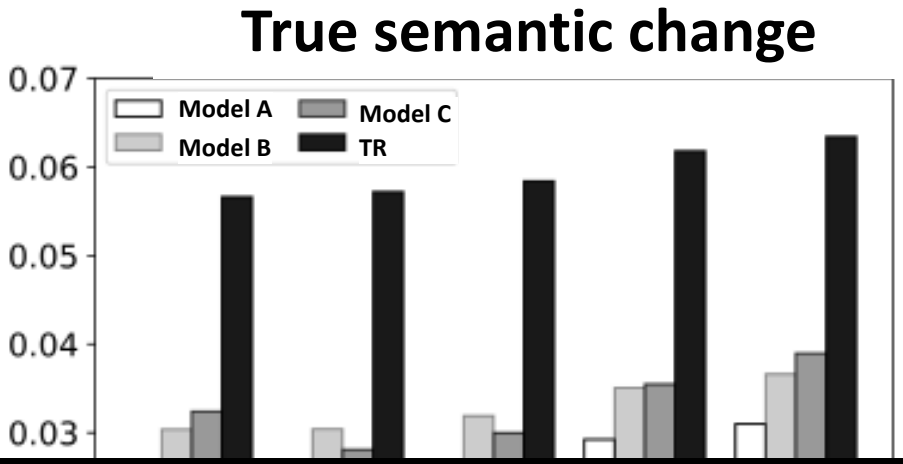
	Model A	**Model** B	**Model** C	**TR**
accuracy	0.65	0.66	0.59	**0.70**
F1-score	0.69	0.69	0.67	**0.74**

Evaluate model sensitivity



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

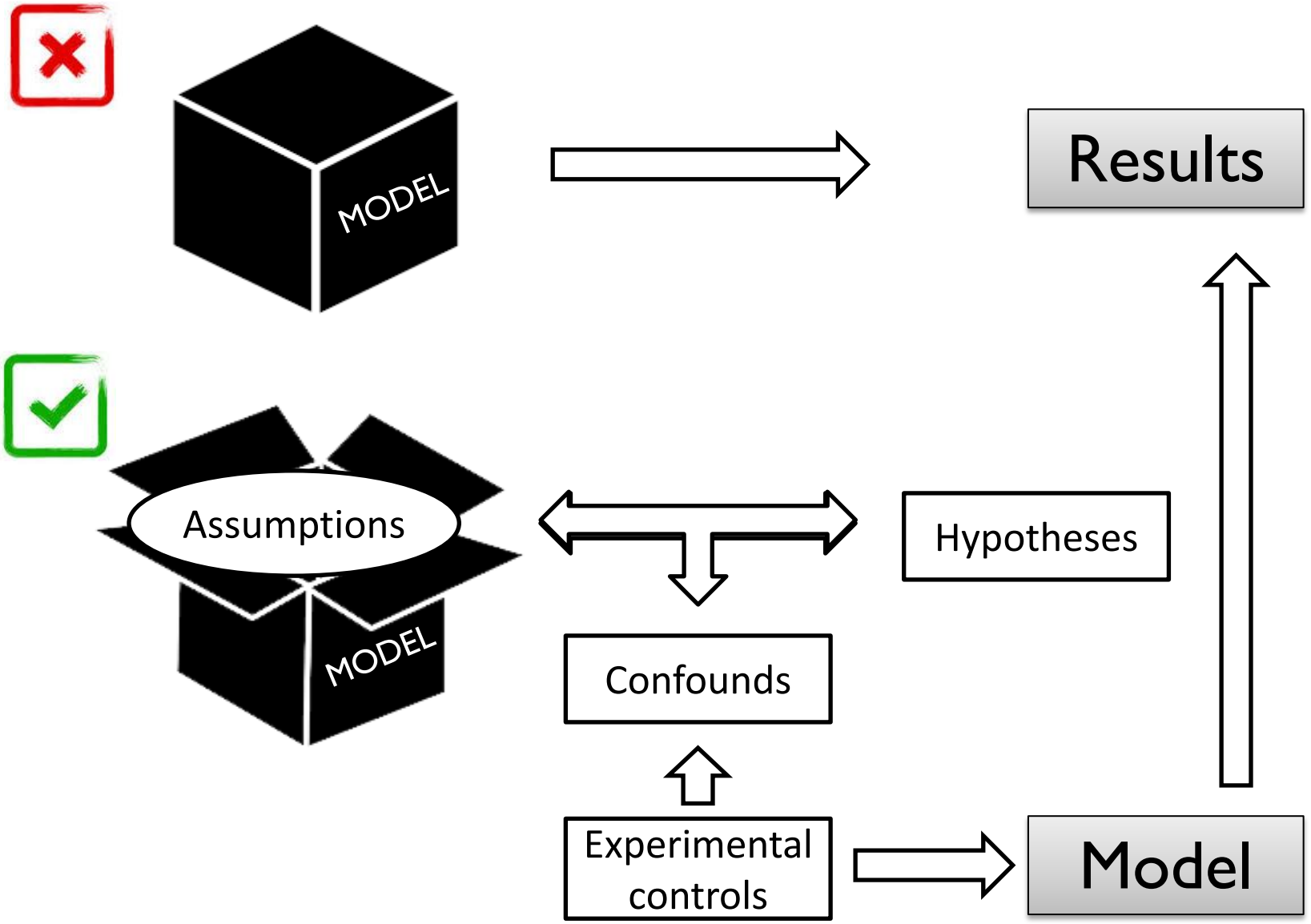
Evaluate model sensitivity



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

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

A different view of models



Conclusions

- All models are wrong!

Conclusions

- All models are wrong!

- But are they importantly wrong?

- Be AWARE of the underlying assumptions of the models and test them.
  - We may get to wrong conclusions.
  - It may guide us in developing better models!

Credits

- Part I - my PhD supervisors:
  Daphna Weinshall and Eitan Grossman

- Part II – my collaborators:
  Simon Hengchen - University of Gothenburg
  Nina Tahmasebi - University of Gothenburg
  Dominik Schlechtweg - University of Stuttgart

hd423@cam.ac.uk

Thank you!

hd423@cam.ac.uk