From Theory to Measurement

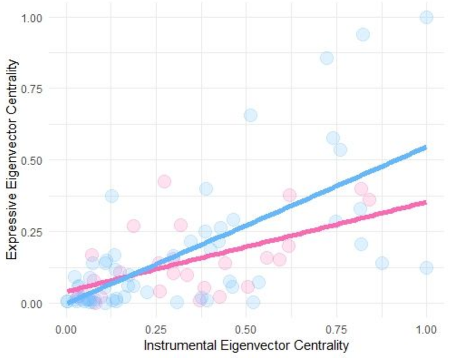
Klint Kanopka
Austin van Loon

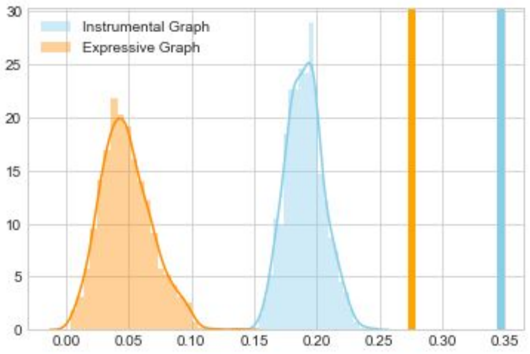
Gender-based Allocative Discrimination in Organizations
(with Kata Mueller-Gastelle)

RQ: How do organizational processes combine with perceptual biases to result in gender-based allocative discrimination?

Data: Enron Email Corpus

Hypothesis: Women end up excluded in either work email networks or social email networks but not both





The Fragility of Scoring Decisions in Automatic Essay Grading Software

RQ: How do automatic essay scoring (AES) systems make scoring decisions and can better understanding these decisions help develop “gaming” strategies?

Data: The Hewlett Foundation AES Kaggle Competition

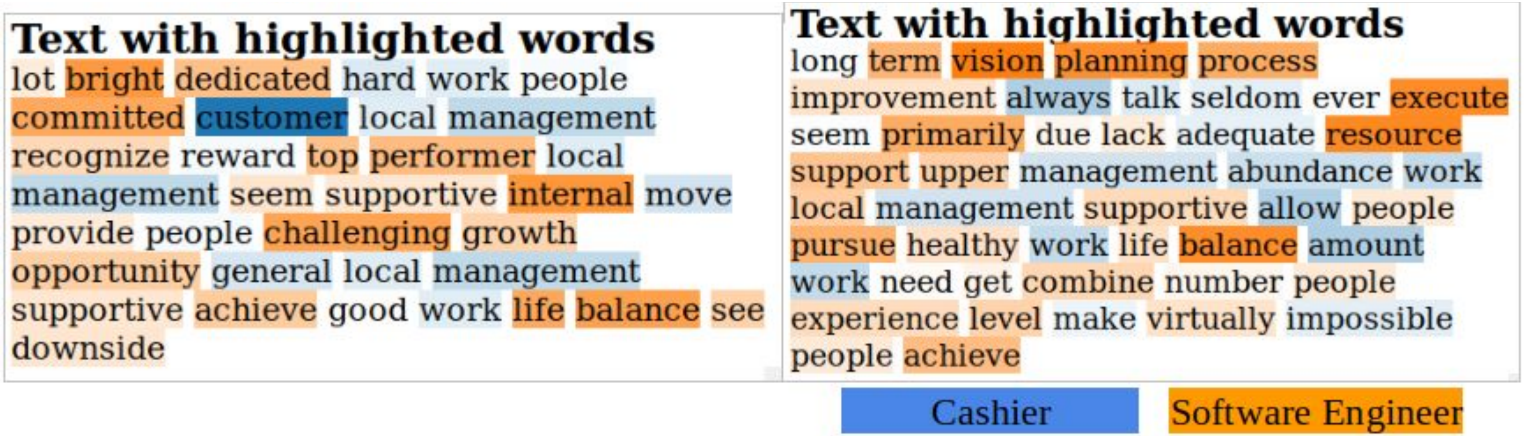
Preliminary Finding: Seemingly high-performing AES systems can make scoring decisions based on single words and randomly inserting these words into low-scoring essays can result in better scores.

“Getting” a Job: Social Position and the Experience and Meaning of Work

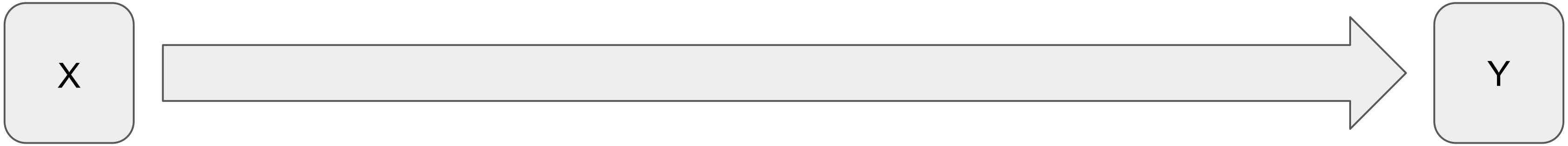
RQ: How do occupations “hang together” with respect to the meaning of work?

Data: ~3 million company reviews from Glassdoor, Inc.

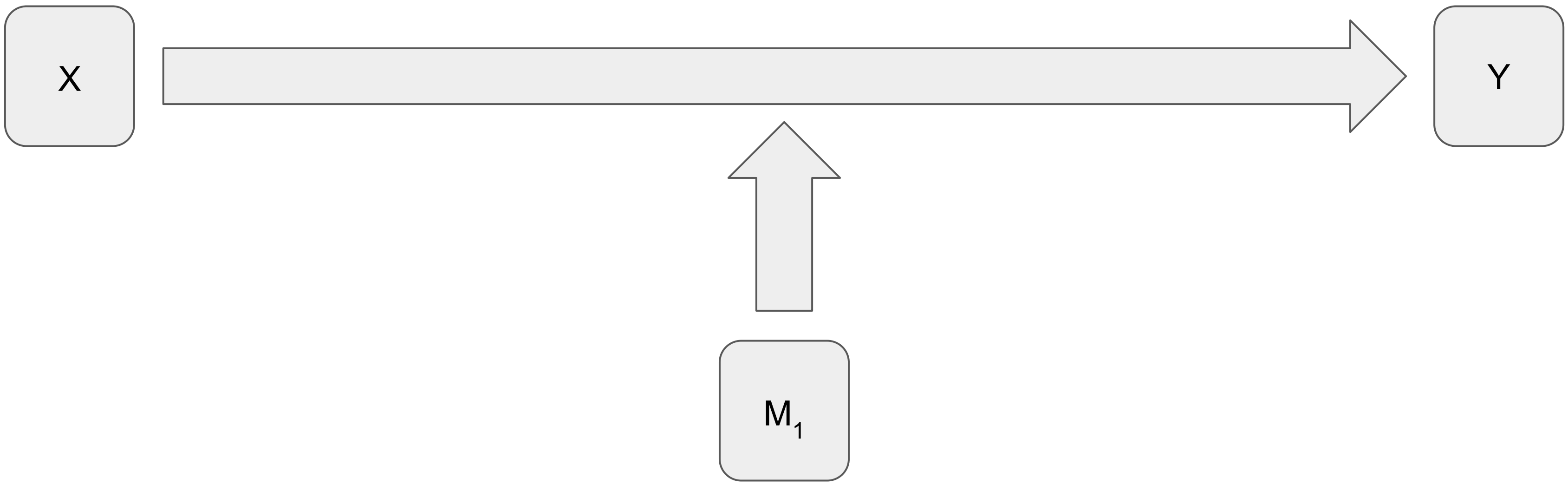
Preliminary analysis: Features that predict the occupation of the reviewer



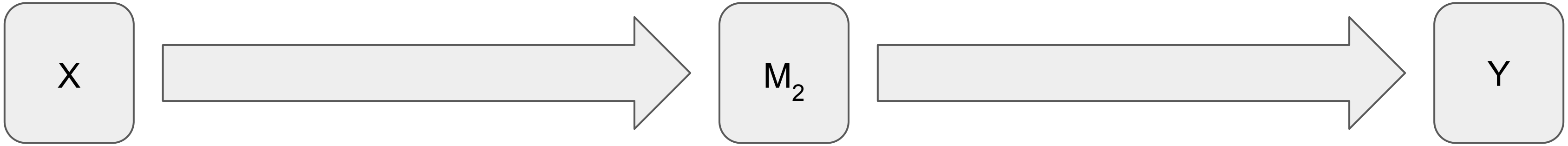
Pragmatic definitions of “theory” and “measurement”



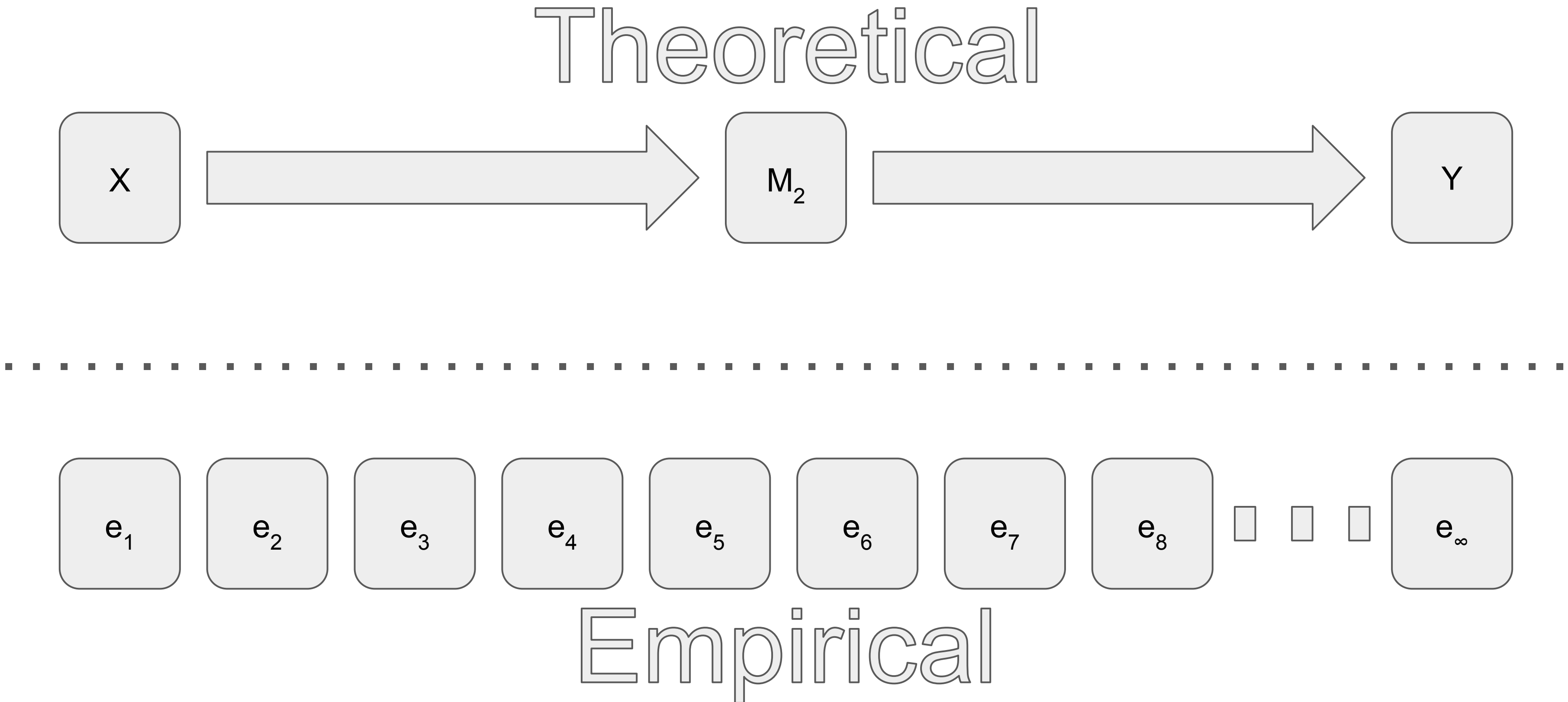
Pragmatic definitions of “theory” and “measurement”



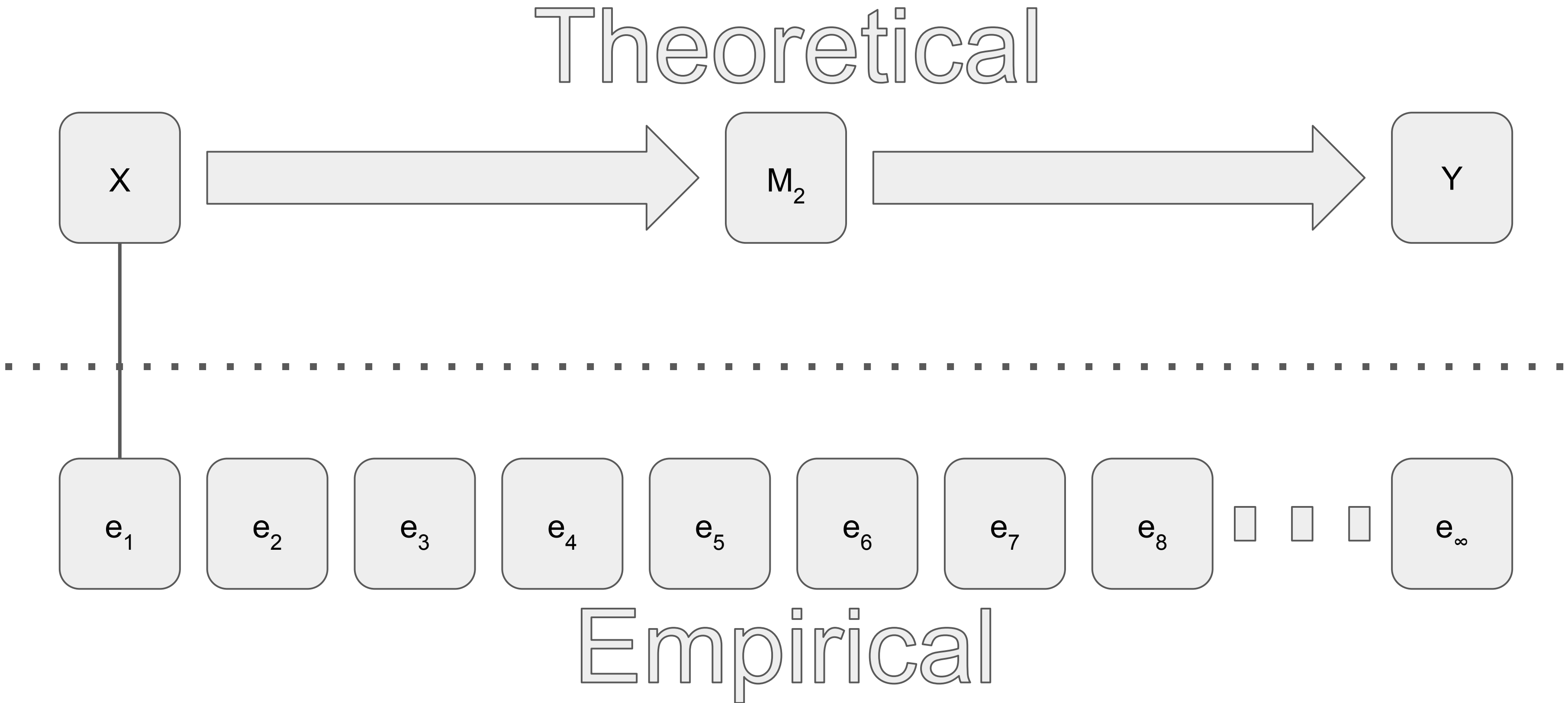
Pragmatic definitions of “theory” and “measurement”



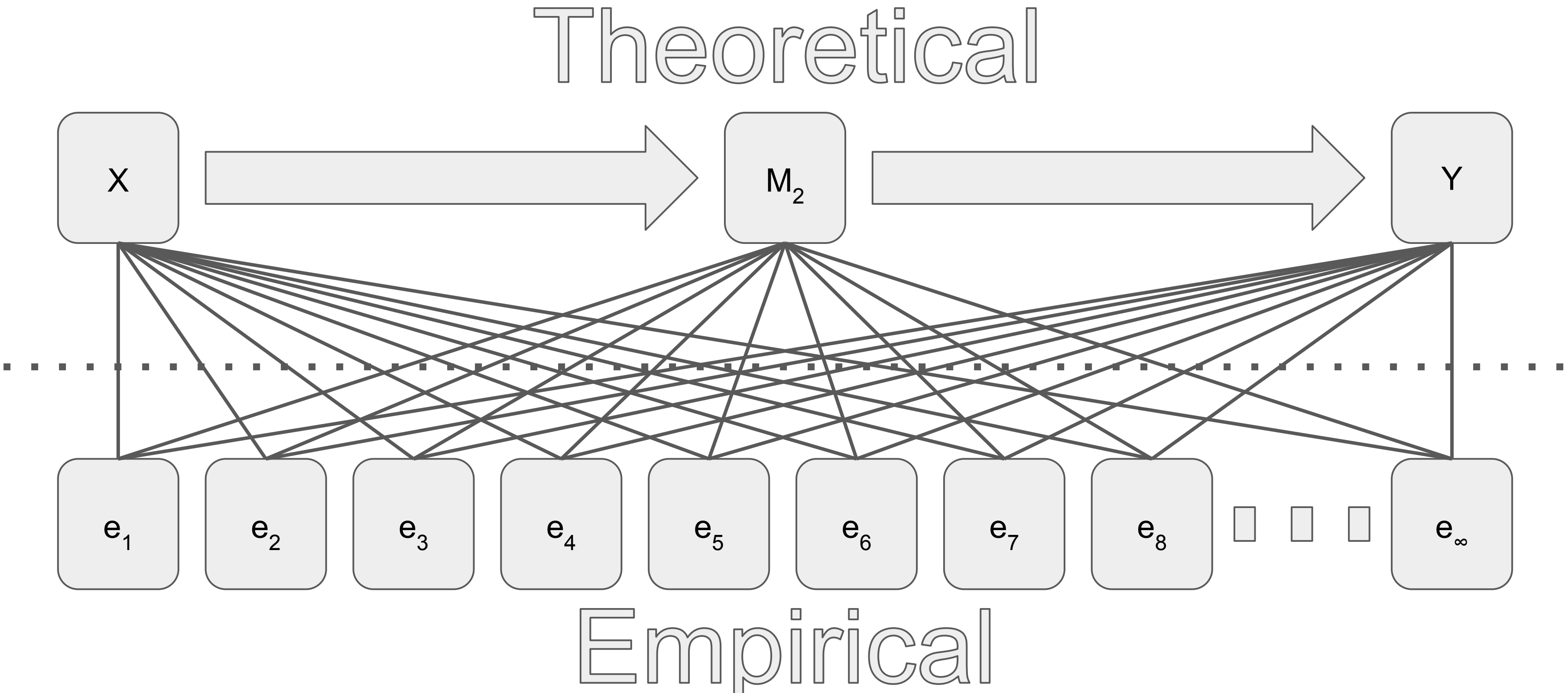
Pragmatic definitions of “theory” and “measurement”



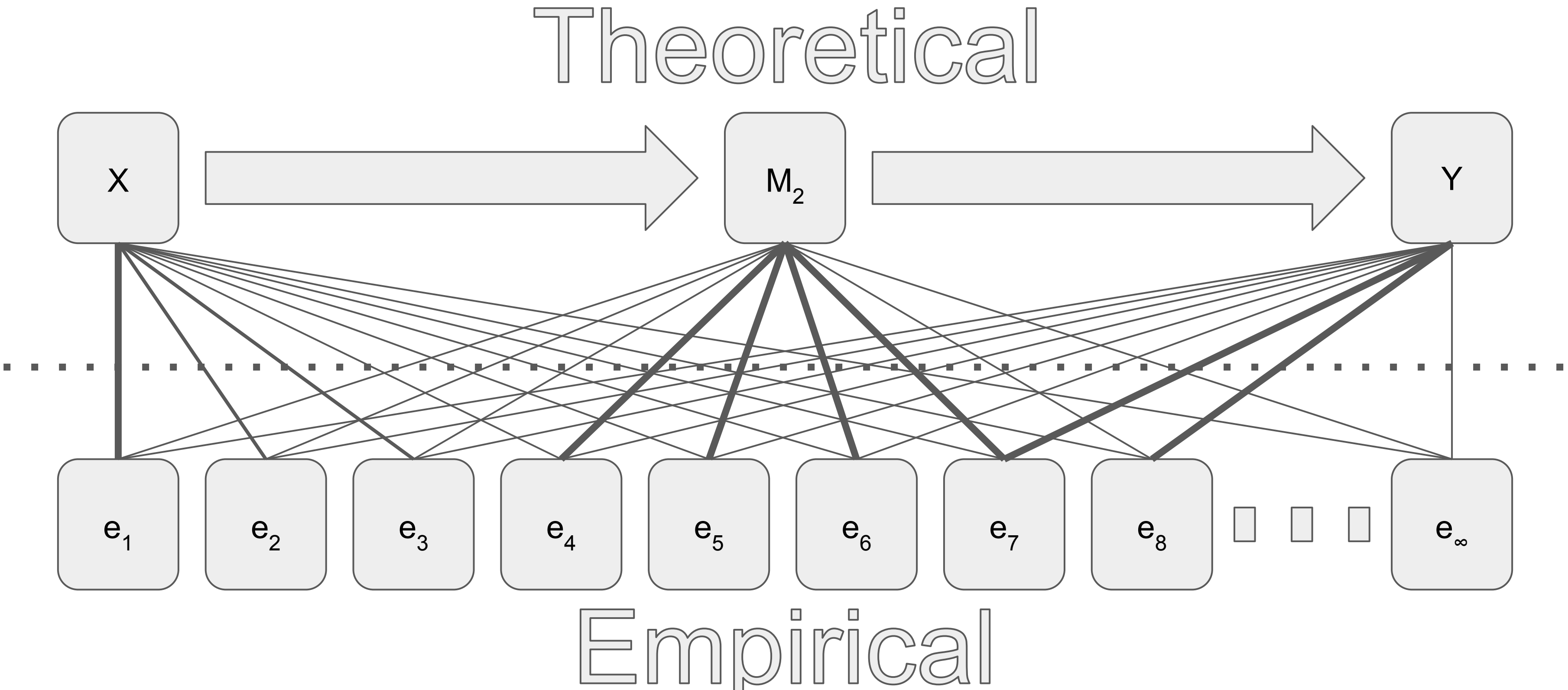
Pragmatic definitions of “theory” and “measurement”



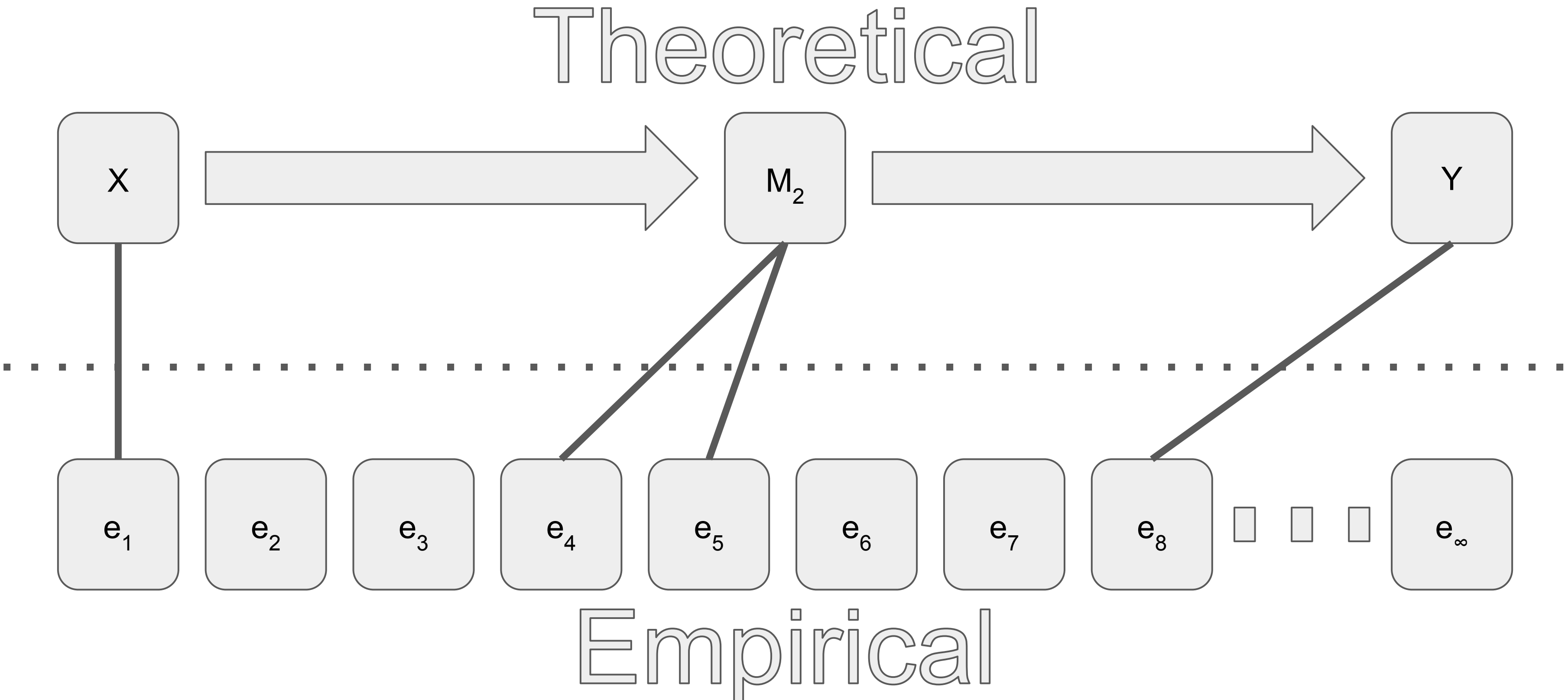
Pragmatic definitions of “theory” and “measurement”



Pragmatic definitions of “theory” and “measurement”

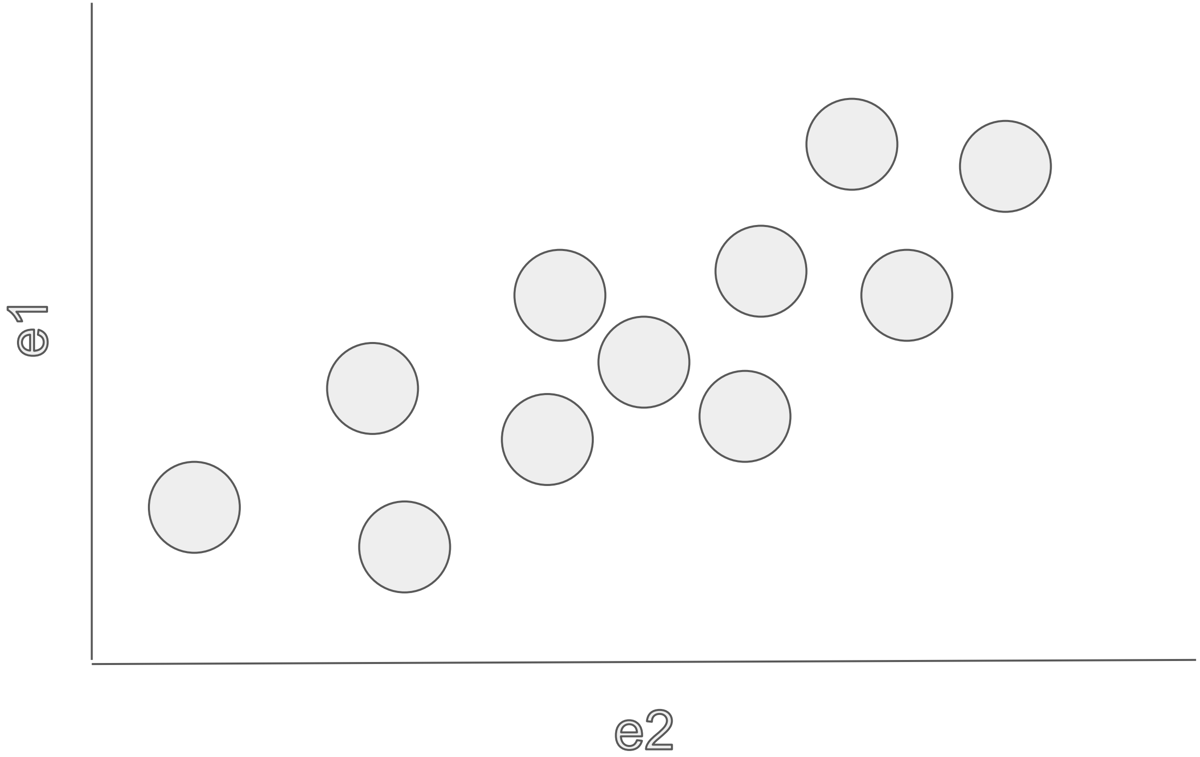


Pragmatic definitions of “theory” and “measurement”

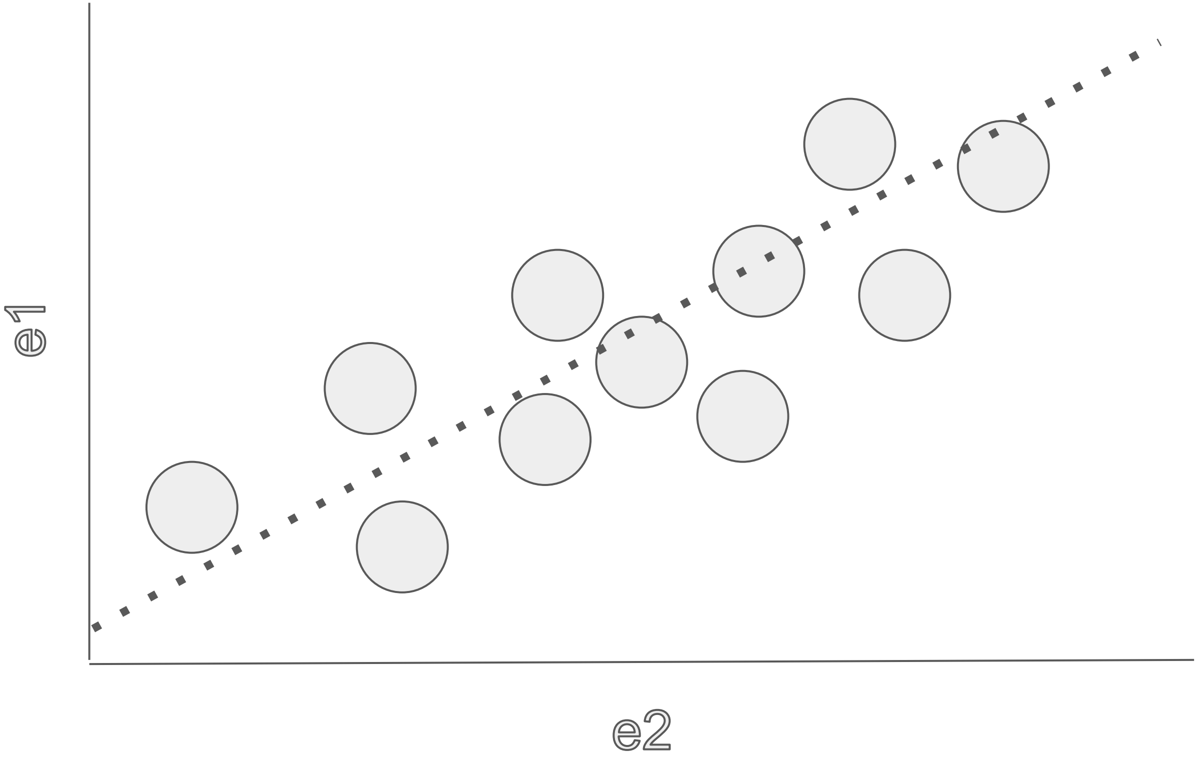


The problem setting

The problem setting



The problem setting



The problem setting

OBSERVATION	e_1	e_2
0	60,000	45,000
1	80,000	60,000
2	45,000	61,000
...	...	...
N	120,000	90,000

Where does text fit in?

Text as predictor (*X*)

- How does the type of moral language used by a politician affect the share of votes they receive from different religious groups in primary elections?
- How does repeating words and using novel words affect the rate at which children learn language? Can “speaking in topics” moderate this?

Text as outcome (*Y*)

- How does threatening the status of whites affect how they describe welfare programs?
- How does the gender makeup of servers affect a restaurants’ reviews on Yelp?

The problem with text...

Assume your data are tweets that all have a length of 30 words

Assume all your tweets only use the one-thousand most popular words in the English language

What would a perfect-fidelity representation of the data look like?

The problem with text...

OBSERVATION	**Tweet 1** ($w_1, w_1, \ldots, w_1, w_1$)	**Tweet 2** ($w_1, w_1, \ldots, w_1, w_2$)	**Tweet 3** ($w_1, w_1, \ldots, w_1, w_3$)	**...**	**Tweet 1000^{30}** ($w_{1000}, w_{1000}, \ldots, w_{1000}, w_{1000}$)
0	0	0	0	...	0
1	0	0	0	...	0
2	0	0	0	...	0
...	...	...	....	...	...
N	0	0	0	...	0

The problem with text… (dimensions = $|W|^P$)

OBSERVATION	**Tweet 1** $(w_1, w_1, \ldots, w_1, w_1)$	**Tweet 2** $(w_1, w_1, \ldots, w_1, w_2)$	**Tweet 3** $(w_1, w_1, \ldots, w_1, w_3)$	**...**	**Tweet 1000^{30}** $(w_{1000}, w_{1000}, \ldots, w_{1000}, w_{1000})$
0	0	0	0	...	0
1	0	0	0	...	0
2	0	0	0	...	0
...	...	...	....	...	...
N	0	0	0	...	0

The problem with text...

Need to reduce the dimensions of data

- Computational constraint
- Inability to make inference (matrix too sparse)

Need figure out ways to extract the meaningful similarities and differences between texts

Some solutions

Idea 1: let’s assume the meaning of one word is independent of the meaning of other words and that word order doesn’t matter (bag of words)

Some solutions

OBSERVATION	w_1	w_2	w_3	...	w_{1000}
0	1	3	0	...	1
1	1	0	1	...	0
2	0	0	0	...	4
...	...	...	....	...	...
N	1	0	0	...	0

Some solutions (dimensions = |W|)

OBSERVATION	w_1	w_2	w_3	...	w_{1000}
0	1	0	0	...	1
1	1	0	3	...	0
2	0	0	0	...	1
...	...	...	....	...	...
N	1	0	0	...	2

Some tweets...

- This is another pathetic @CNN spin-job. The State Dept has been a thorn in the side of Republican Presidents since Reagan. Pompeo rocks.
- Anyhoo I'm not a food critic. I love food. I love to cook it. I don't go out much. I love home. The list is an honest list places that we love when I DO get dressed, and doesn't mean more!
- A great evening last night in Kentucky and Mississippi for the Republican Party with 13 BIG WINS, including a Governorship in Mississippi. Congratulations to everyone!

Some solutions

Idea 1: let’s assume the meaning of one word is independent of the meaning of other words and that word order doesn’t matter (bag of words)

Idea 2: let’s assume words fall into *a priori* classes and that the prevalence of those classes in the text captures the relevant information

Some solutions

OBSERVATION	Prevalence of words in class s_1	Prevalence of words in class s_2	...	Prevalence of words in class s_{70}
0	1	1	...	0
1	5	0	...	0
2	0	2	...	4
...	...	...	...	...
N	0	0	...	8

Some solutions (dimensions = |S|)

OBSERVATION	Prevalence of words in class s_1	Prevalence of words in class s_2	...	Prevalence of words in class s_{70}
0	1	1	...	0
1	5	0	...	0
2	0	2	...	4
...	...	...	...	...
N	0	0	...	8

Some tweets...

- This is another pathetic @CNN spin-job. The State Dept has been a thorn in the side of Republican Presidents since Reagan. Pompeo rocks.
- Anyhoo I’m not a food critic. I love food. I love to cook it. I don’t go out much. I love home. The list is an honest list places that we love when I DO get dressed, and doesn’t mean more!
- A great evening last night in Kentucky and Mississippi for the Republican Party with 13 BIG WINS, including a Governorship in Mississippi. Congratulations to everyone!

Some solutions

Idea 1: let’s assume the meaning of one word is independent of the meaning of other words and that word order doesn’t matter (bag of words)

Idea 2: let’s assume words fall into *a priori* classes and that the prevalence of those classes in the text captures the relevant information

Idea 3: let’s assume a supervised machine learning algorithm we train makes an unbiased estimate of the relevant category of the text

Some solutions

OBSERVATION	Tweet categorized as $c = 0$	Tweet categorized as $c = 1$	...	Tweet categorized as $c = 5$
0	1	0	...	0
1	0	1	...	0
2	0	0	...	1
...	...	...	...	...
N	0	1	...	0

Some solutions (dimensions = *C*)

OBSERVATION	**Tweet categorized as *c* = 0**	**Tweet categorized as *c* = 1**	**...**	**Tweet categorized as *c* = 5**
0	1	0	...	0
1	0	1	...	0
2	0	0	...	1
...	...	...	...	...
N	0	1	...	0

Some tweets...

- This is another pathetic @CNN spin-job. The State Dept has been a thorn in the side of Republican Presidents since Reagan. Pompeo rocks.
- Anyhoo I’m not a food critic. I love food. I love to cook it. I don’t go out much. I love home. The list is an honest list places that we love when I DO get dressed, and doesn’t mean more!
- A great evening last night in Kentucky and Mississippi for the Republican Party with 13 BIG WINS, including a Governorship in Mississippi. Congratulations to everyone!

Some solutions

Idea 1: let's assume the meaning of one word is independent of the meaning of other words and that word order doesn't matter (bag of words)

Idea 2: let's assume words fall into *a priori* classes and that the prevalence of those classes in the text captures the relevant information

Idea 3: let's assume a supervised machine learning algorithm we train makes an unbiased estimate of the relevant category of the text

Idea 4: let's assume the relevant information concerning words is well captured by a given vector embedding and that the document is well represented by some transformation of the vectors representing the words in that document

Some solutions

If each word has a vector, $w \in \mathbf{R}^d$, we might represent an entire review (like "It was a good burger") as:

$$R_1 = (w_{it} + w_{was} + w_a + w_{good} + w_{burger})/5$$

In this case, we used '5' as a (naive) way to normalize for length.

Some solutions

OBSERVATION	d_1	d_2	d_3	...	d_{50}
0	0.719	0.233	0.154	...	0.642
1	0.331	0.321	0.111	...	0.482
2	0.296	0.691	0.982	...	0.201
...	...	...	....	...	...
N	0.998	0.571	0.339	...	0.001

Some solutions (dimensions = *D*)

OBSERVATION	d_1	d_2	d_3	...	d_{50}
0	0.719	0.233	0.154	...	0.642
1	0.331	0.321	0.111	...	0.482
2	0.296	0.691	0.982	...	0.201
...	...	...	....	...	...
N	0.998	0.571	0.339	...	0.001

Some tweets...

- This is another pathetic @CNN spin-job. The State Dept has been a thorn in the side of Republican Presidents since Reagan. Pompeo rocks.
- Anyhoo I’m not a food critic. I love food. I love to cook it. I don’t go out much. I love home. The list is an honest list places that we love when I DO get dressed, and doesn’t mean more!
- A great evening last night in Kentucky and Mississippi for the Republican Party with 13 BIG WINS, including a Governorship in Mississippi. Congratulations to everyone!

The thing about text...

In general, we want to find the happy medium point of fidelity (depends on question and corpus)

Make sure your measure corresponds to theoretical construct and as little other stuff as possible

A lot of text analysis methods are just different ways of extracting potentially useful information (stemming, TF-IDF, dictionaries, sentiment analysis, topic models, word embeddings)

KSS Survey:

shorturl.at/goqO1