

# EXPLAINING TOPICAL DISTANCES USING WORD EMBEDDINGS

Nils Witt, [ZBW Leibniz Information Centre for Economics](#)  
Christin Seifert and Michael Granitzer, [University of Passau](#)

5 September 2016, Porto Portugal

# INTRODUCTION

# WORD EMBEDDINGS

- Efficient estimation of word representations in vector space, 2013, Mikolov et al.
- Distributed representations of words and phrases and their compositionality, 2013, Mikolov et al.

# WORD EMBEDDINGS

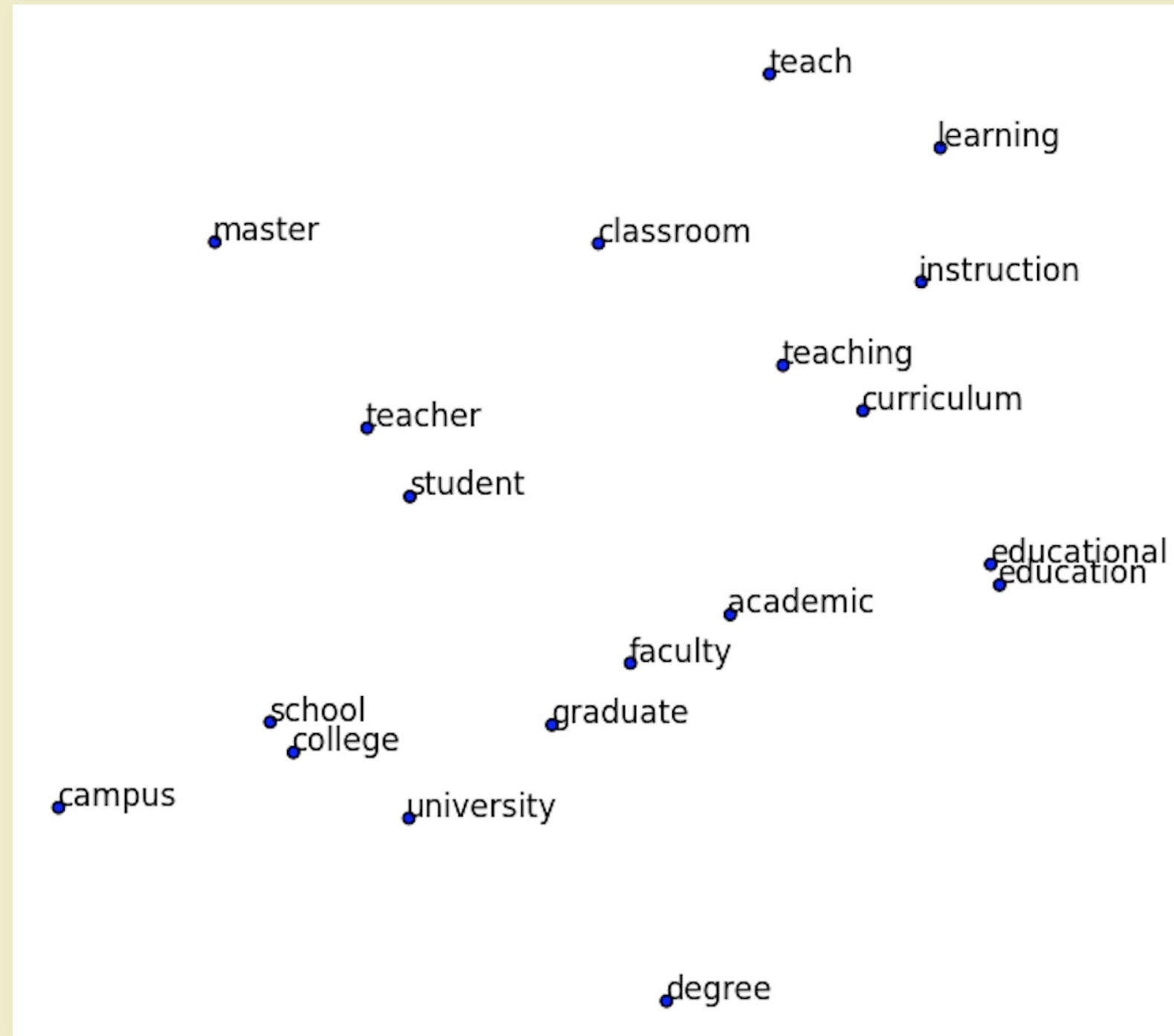
Numerical representations for words

Respect semantic structure

Hundreds or thousands dimensional

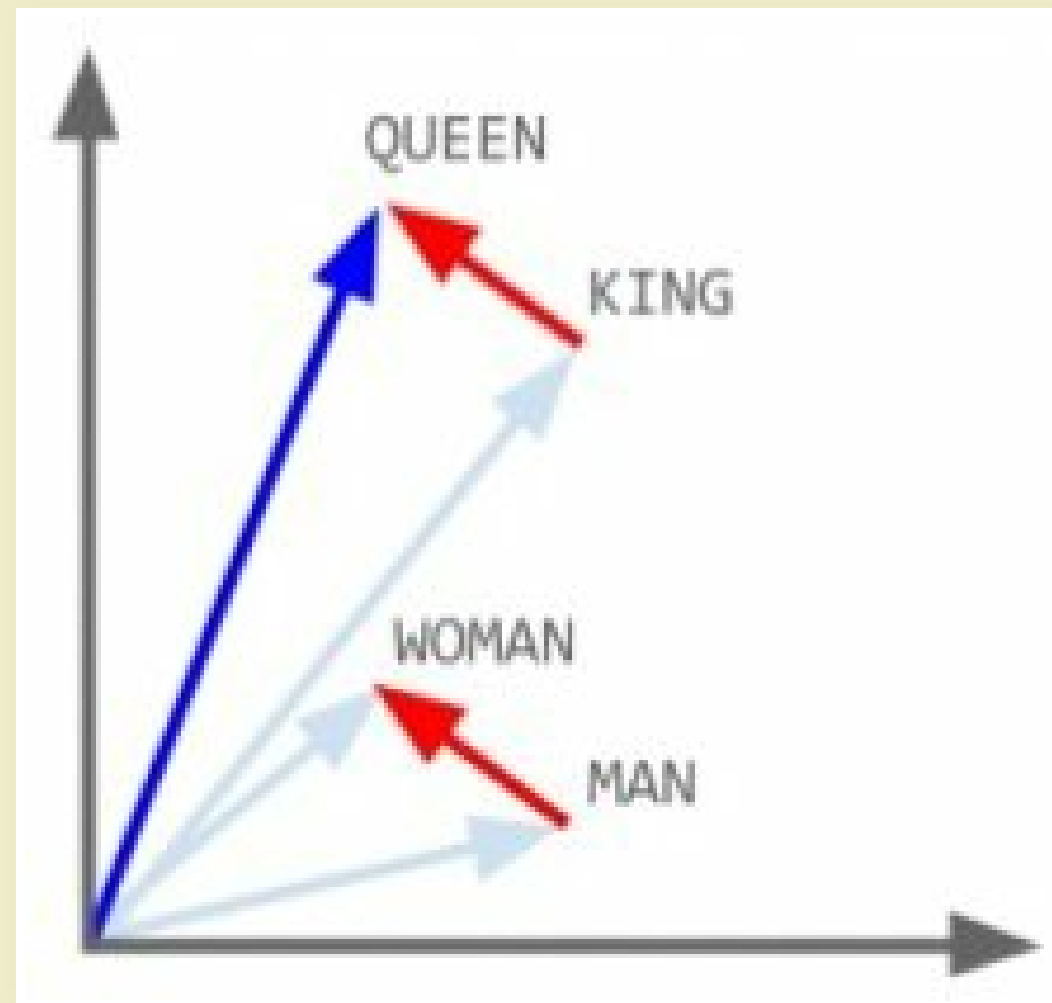
Explaining how it works is beyond the scope of this talk

# WORD2VEC SPACE PROJECTED INTO 2D



# WORD2VEC ARITHMETIC (1/2)

```
# Gender  
vec = model["king"] - model["man"] + model["woman"]  
model.find_nearest(vec)  
>>> "queen"
```



# WORD2VEC ARITHMETIC (2/2)

```
# Composition  
vec = model["human"] + model["robot"]  
model.find_nearest(vec)  
>>> "cyborg"
```

# DOCUMENT EMBEDDINGS

Distributed representations of sentences and documents, 2014,  
Q. V. Le and T. Mikolov

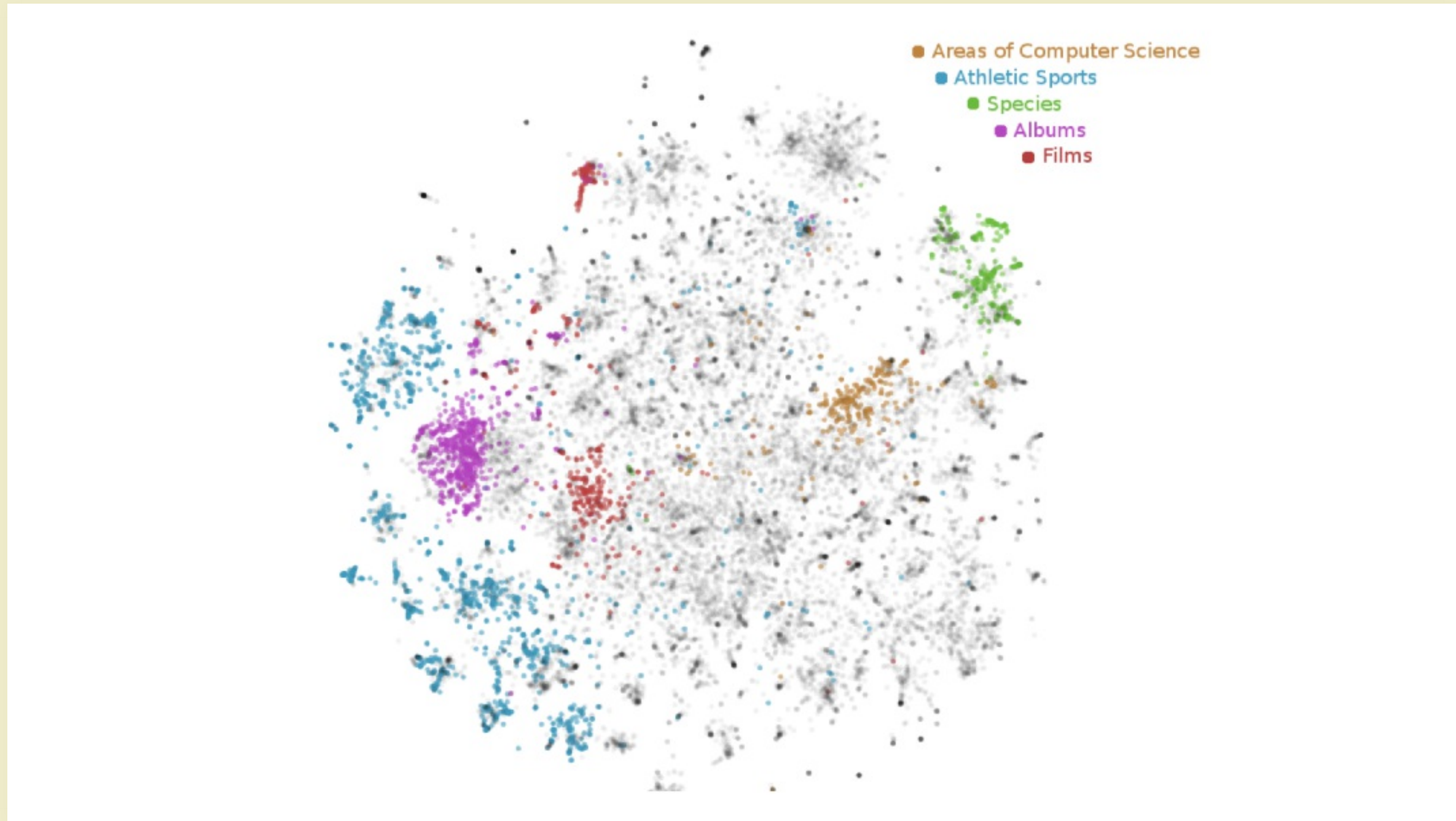


# DOCUMENT EMBEDDINGS

Same idea, applied to larger text pieces

*Allow semantical text comparisons*

# DOC2VEC SPACE PROJECTED INTO 2D



from Dai et al., “Document embedding with paragraph vectors”, 2015

# DOC2VEC ARITHMETIC

```
# Find similar text documents
vec = model["Distributed Representations of Words and " +
"Phrases and their Compositionality"]
model.find_nearest(vec)
>>> "Distributed representations of sentences and documents"
```

```
# Comparison
doc_1 = model["Annual Report 2012"]
doc_2 = model["Annual Report 2013"]
doc_3 = model["Annual Report 1990"]
similarity(doc_1, doc_2)
>>> 0.739
similarity(doc_1, doc_3)
>>> 0.357
```

# COMBINED ARITHMETIC

Document embedding with paragraph vectors, 2014, Dai et al.

```
vec = model.docvecs["Lady Gaga"] - model["American"] + model["Japanese"]
model.docvecs.find_nearest(vec)
>>> "Ayumi Hamasaki"
```

Ayumi Hamasaki is one of the most famous pop singers in Japan  
She also has an album called “Poker Face”, released in 1998

# INITIAL IDEA

Given

```
vec = model.docvecs["Lady Gaga"] - model.docvecs["Ayumi Hamasaki"]
```

Can this be done?

```
model.infer_path(vec) // Can infer_path be implemented?  
>>> {"American": -1, "Japanese": 1}
```

# OUTLINE

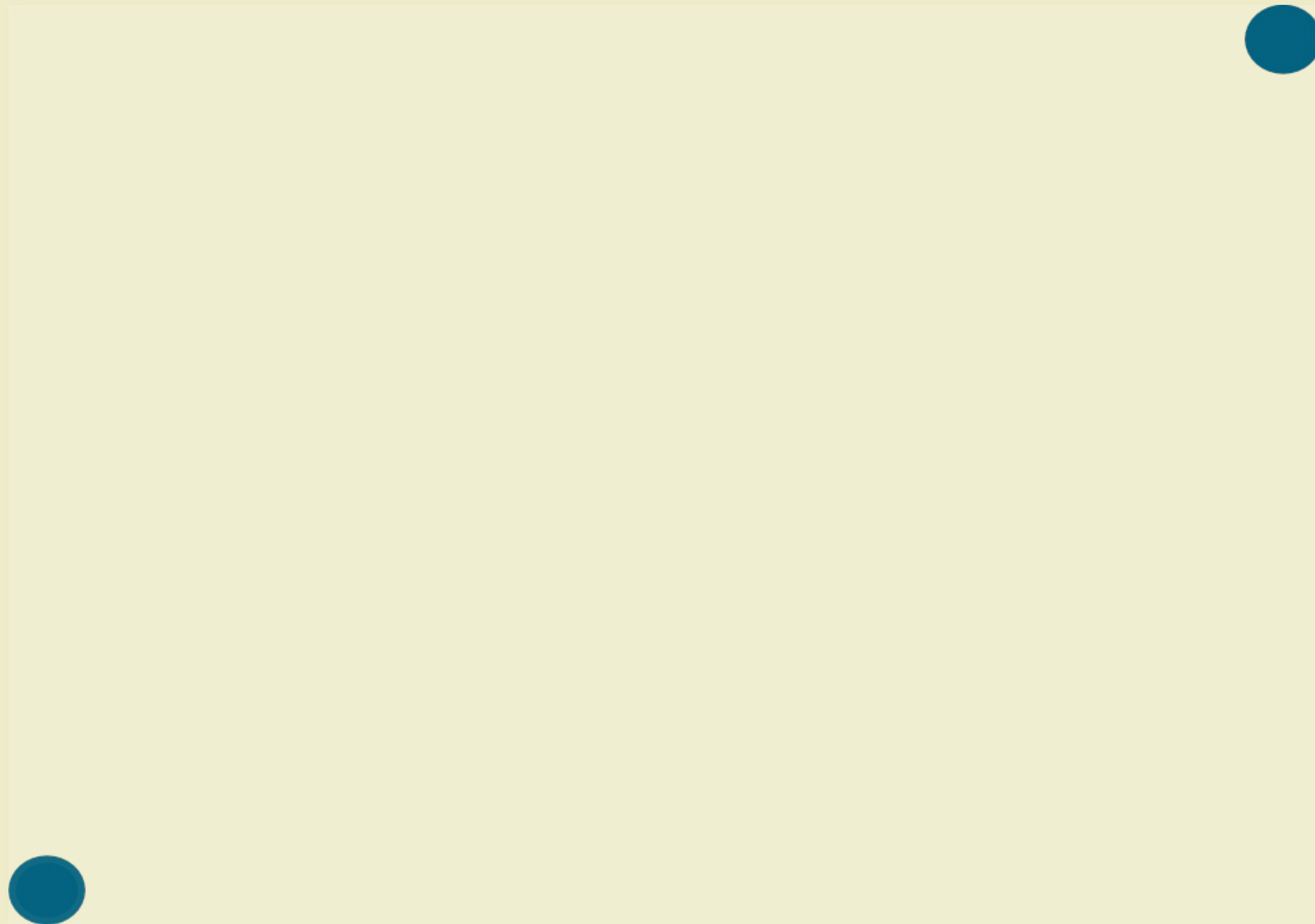
**IDEA**

**EXPERIMENT AND RESULTS**

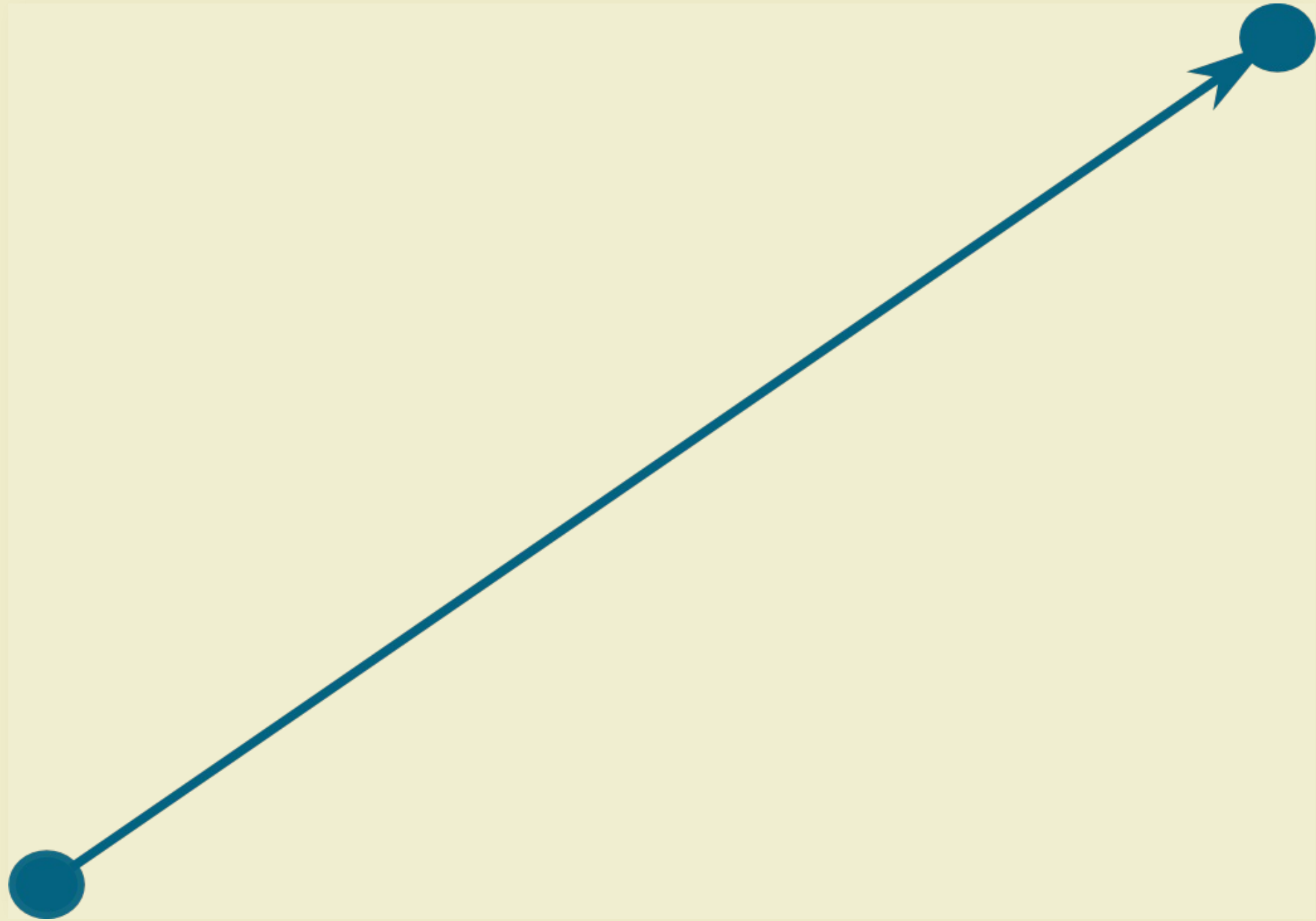
**DISCUSSION**

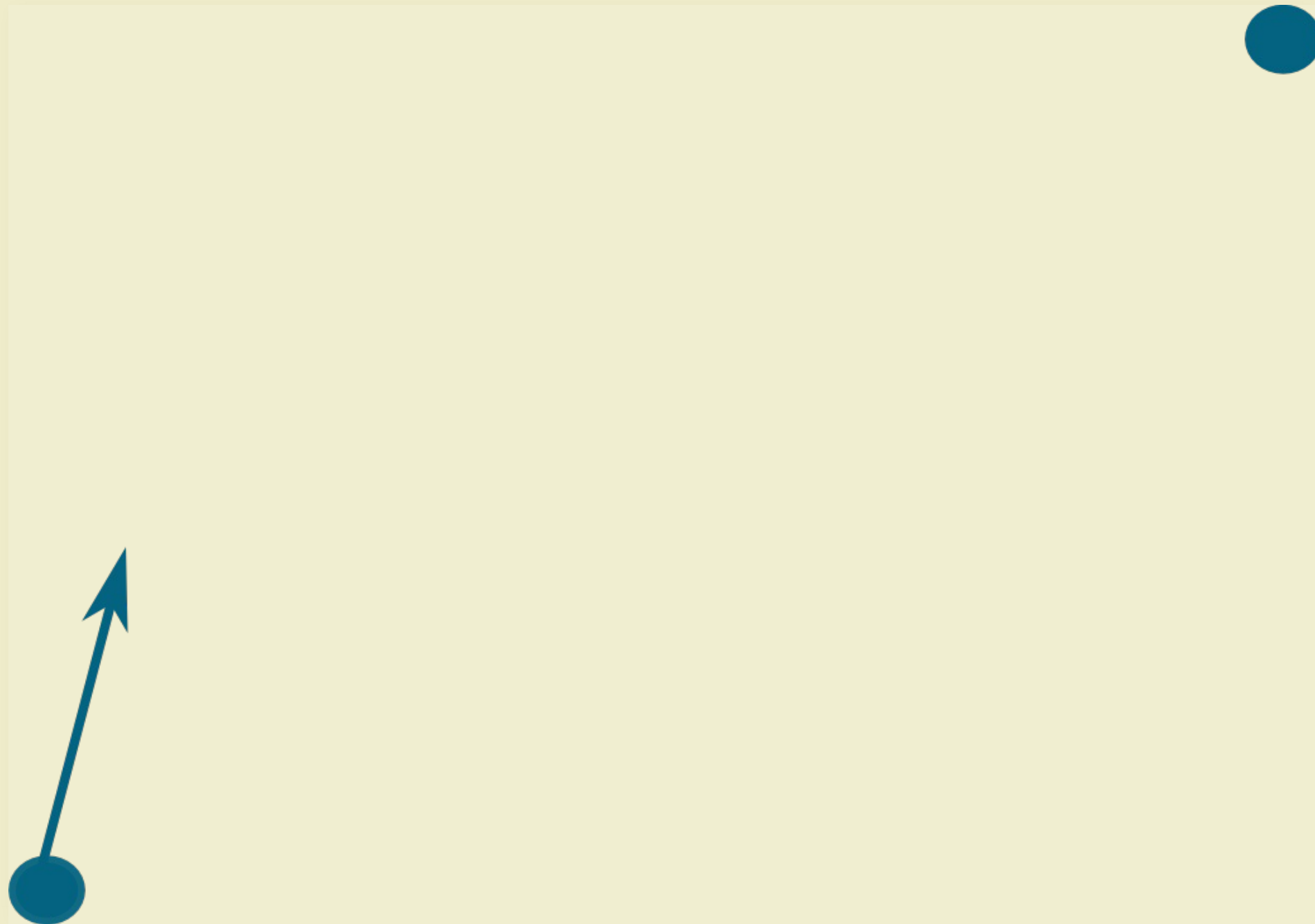
**CONCLUSIONS**

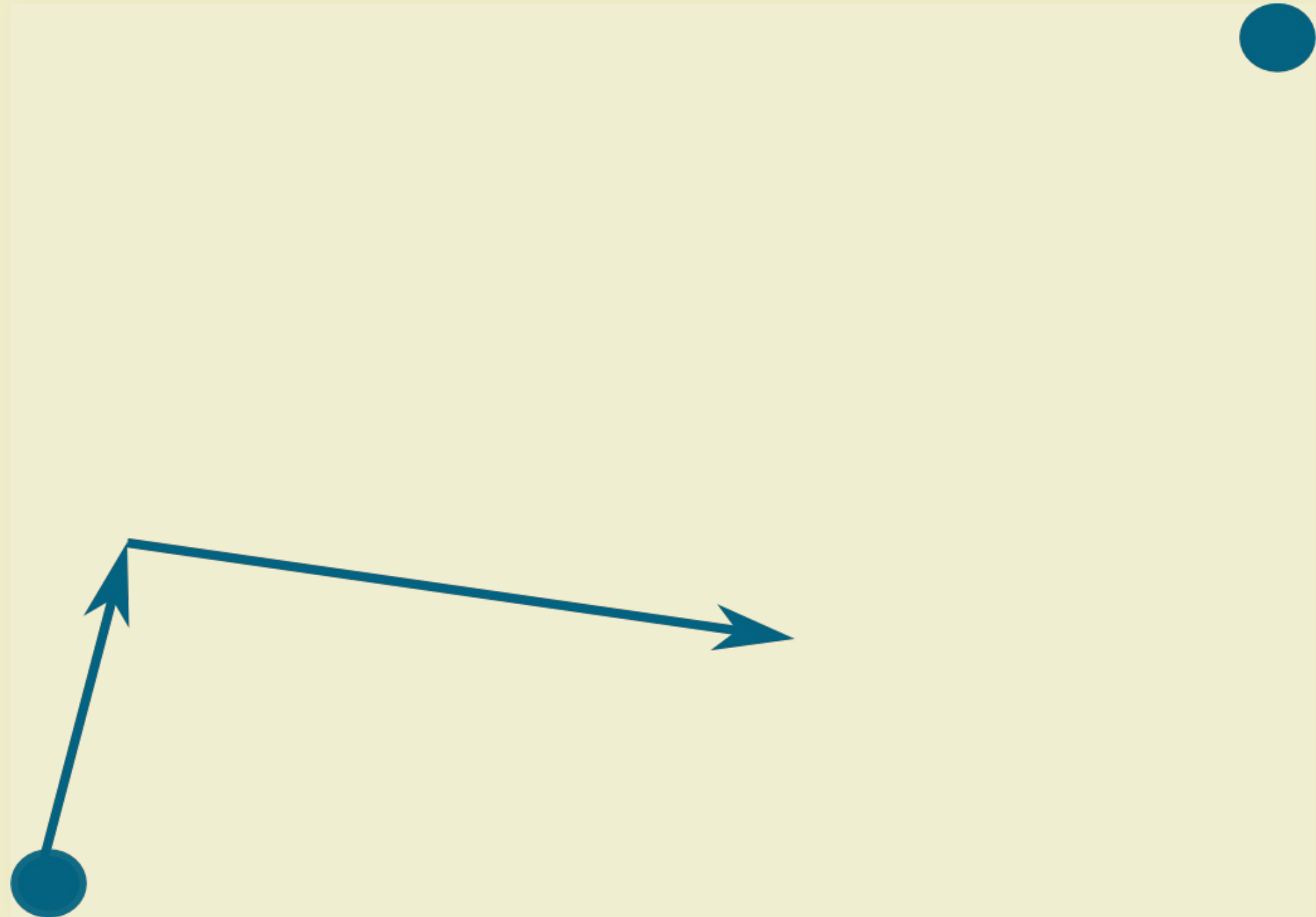
**IDEA**

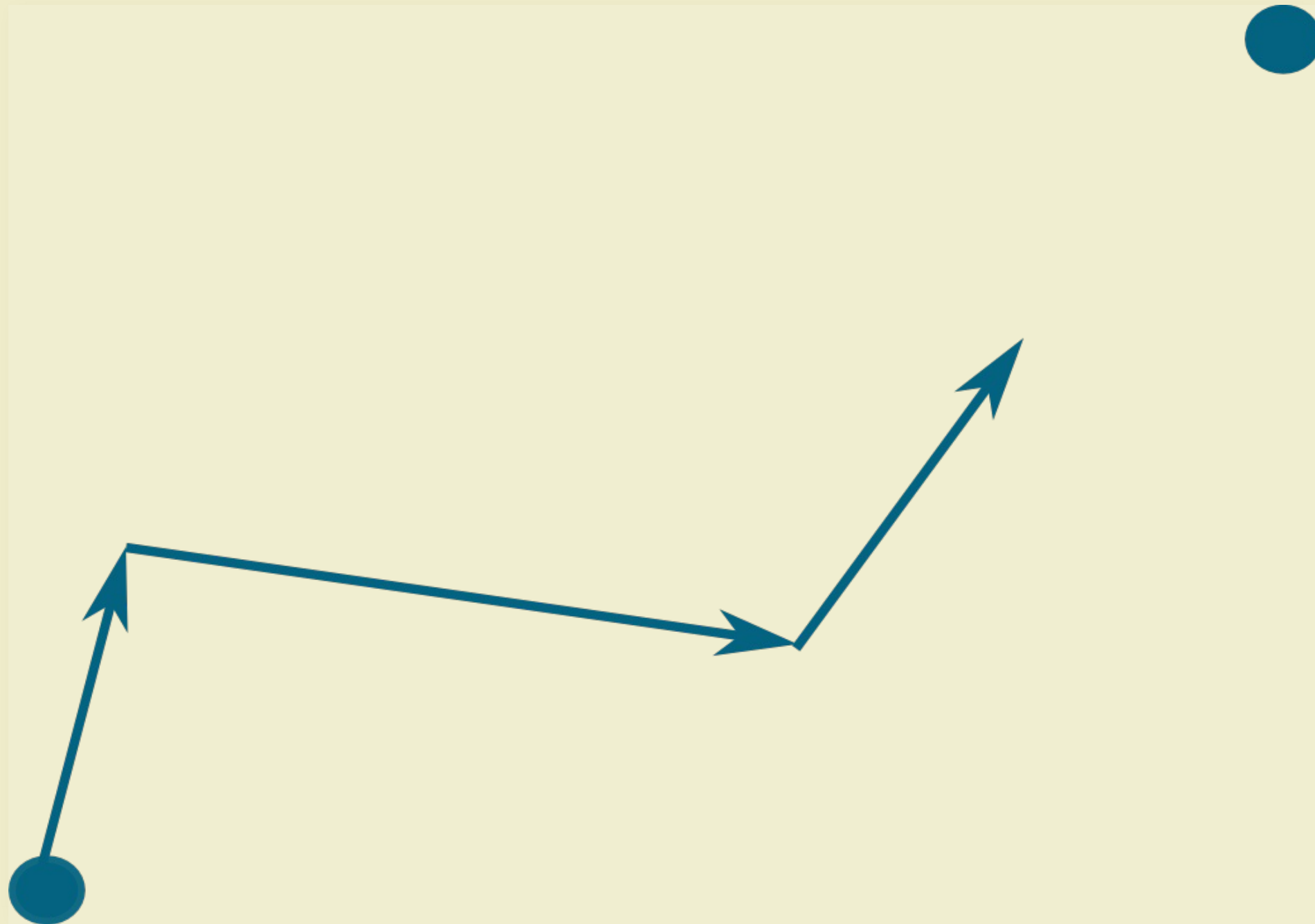


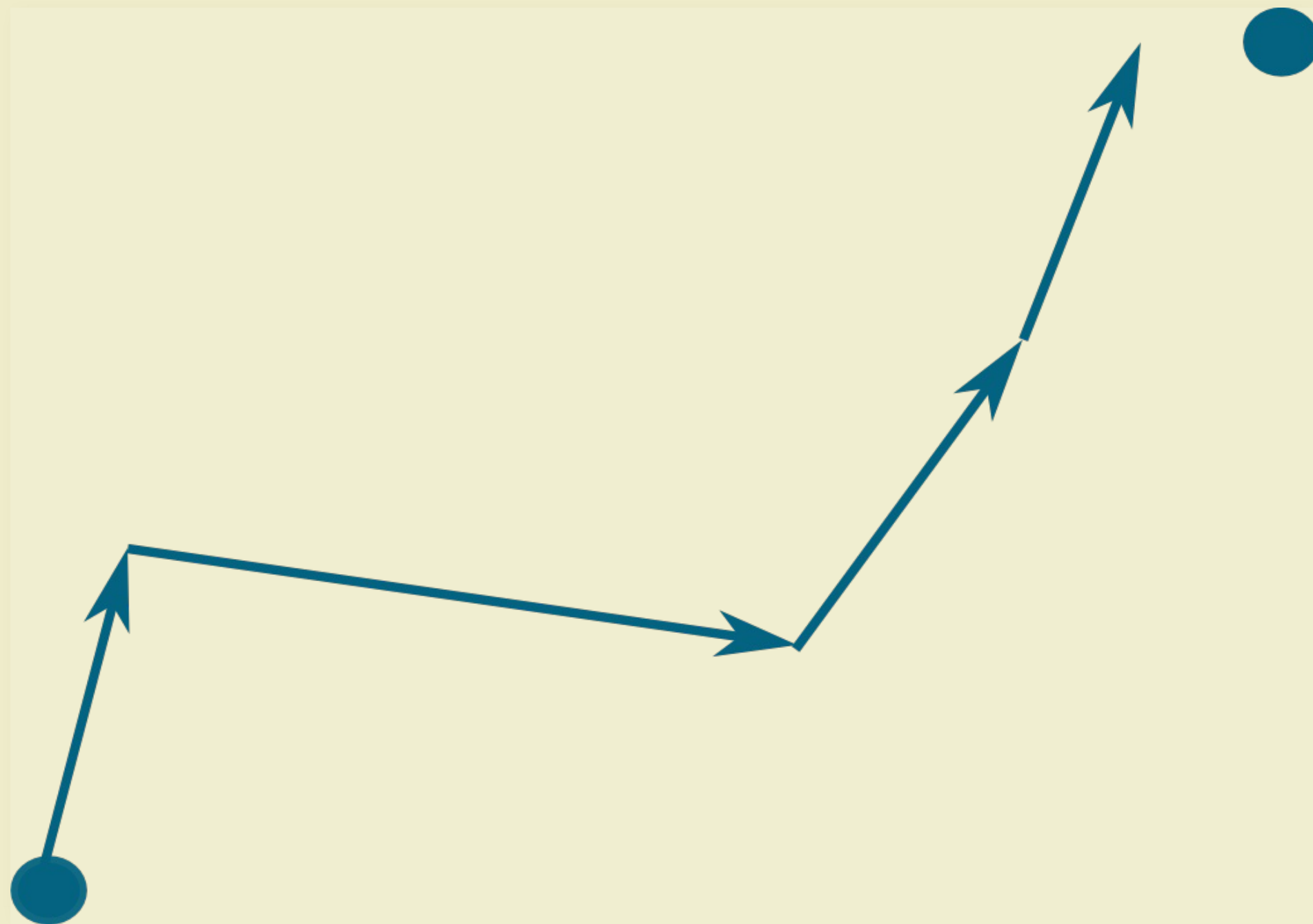












# FORMALIZATION

The goal is to find the smallest set of words  $W$   
( $W \subseteq \text{Vocabulary}$ ) such that:

$$\text{dist}(dv(Y), dv(X) + \sum_{i=0}^{|W|} (-1)^{n_i} wv(W_i)) < \epsilon$$

$$n = \begin{cases} 1 & \text{if } \text{dist}(dv(Y), dv(X) + wv(W_i)) > \\ & \text{dist}(dv(Y), dv(X) - wv(W_i)) \\ 0 & \text{otherwise} \end{cases}$$

# EXPERIMENT AND RESULTS

# ECONSTOR16 CORPUS

- Based on ZBW's open-access repository Econstor
- >100,000 documents from economics
- meta data: author, title, number of citations, domain expert assigned keywords
- 77% documents English , 19% German, 4% 20 different languages





# ALGORITHMIC PROCEDURE (1/2)

Distance measure: Cosine similarity

100 and 600 dimensional document embeddings

Semi-automatic procedure

# ALGORITHMIC PROCEDURE (2/2)

Given: Doc2Vec model, Vocabulary and two Documents

Compute difference

Find word with highest similarity

Add/subtract word and document

Repeat

# EXAMPLE: SAME AUTHOR

	<b>Author</b>	<b>Title</b>	<b>Expert assigned Keywords</b>
X	Hans-Werner Sinn	Pareto Optimality in the Extraction of Fossile Fuels and the Greenhouse Effect	global warming, resource extraction, Pareto optimality
Y	Hans-Werner Sinn	EU Enlargement and the Future of the Welfare State	EU expansion, migration, labour market, welfare state

# EU Enlargement and the Future of the Welfare State

+

Iteration	Vector added	Cos Similarity @600D
0		-0.01
1	<i>stock</i>	0.04
2	<i>industrial</i>	0.11
3	<i>employee</i>	0.12
4	<i>fuel</i>	0.15
5	<i>diesel</i>	0.19
6	<i>non-statistical</i>	0.20

≈

Pareto Optimality in the Extraction of Fossil Fuels and the Greenhouse Effect

# EXAMPLE: DISTANT DOCUMENTS

	Author	Title	Expert assigned Keywords
X	Hendrik Hagedorn	In search of the marginal entrepreneur: Benchmarking regulatory frameworks in their effect on entrepreneurship	Benchmarking method, entrepreneurship, incentives, regulation
Y	John Hartwick	Mining Gold for the Currency during the Pax Roman	Gold coinage, Roman money, roman empire

# In search of the marginal entrepreneur: Benchmarking regulatory frameworks in their effect on entrepreneurship

+

Iteration	Vector added	Cos Similarity @100D
0		-0.44
1	<i>job</i>	-0.14
2	<i>carbon</i>	-0.12
3	<i>empire</i>	0.22
4	<i>goldsmith</i>	0.36
5	<i>country</i>	0.52
6	<i>interest</i>	0.61

≈

## Mining Gold for the Currency during the Pax Roman

# TRANSFER RESULTS

100D → 600D

Iteration	Vector added	Cos Sim @600D	Cos Sim @100D
0		-0.03	-0.44
1	<i>job</i>	0.01	-0.14
2	<i>carbon</i>	0.03	-0.12
3	<i>empire</i>	0.10	0.22
4	<i>goldsmith</i>	0.007	0.36
5	<i>country</i>	0.007	0.52
6	<i>interest</i>	0.007	0.61

# TRANSFER RESULTS

600D → 100D

Iteration	Vector added	Cos Sim @600D	Cos Sim @100D
0		-0.01	-0.44
1	<i>stock</i>	0.04	-0.30
2	<i>industrial</i>	0.11	-0.22
3	<i>employee</i>	0.12	-0.24
4	<i>fuel</i>	0.15	-0.20
5	<i>diesel</i>	0.19	-0.22
6	<i>non-statistical</i>	0.20	-0.18



# DISCUSSION

# RESULTS

- Although high initial distance, quick convergence @ 100D
- Higher dimensionality leads to slower convergence
- Most words are meaningful (like *empire*, *goldsmith*)
- Some words aren't obviously meaningful (like *carbon*, *country*)
- *goldsmith* isn't mentioned in either documents
- Results produced in different dimensionalities are not similar

# OPTIMIZATIONS AND FUTURE WORK

- Clearing the corpus
- Comprehensive evaluation of the procedure
- n-grams with  $n > 1$
- tf-idf weighting to retrieve more specific words
- Speed: Restricting vocabulary size
- Influence of the embedding dimensionality on the quality of results needs to be investigated

# APPLICATION (1/2)

On Wikipedia corpus:

Compare **Enrico Fermi** to **James Clerk Maxwell**

...formulate the classical theory of electromagnetic radiation...

Maxwell's religious beliefs and related activities...

...was a Scottish scientist...

# APPLICATION (2/2)

On Econstor16 corpus:

I want to read paper X which builds on paper Y.  
Which parts of X should I focus on?

Paragraph 4


Paragraph 7

# CONCLUSIONS

# CONCLUSIONS

- A method explaining topical difference between documents was presented
- Intuitively suitable words are found by the method
- Optimizations to increase the accuracy suggested
- A new corpus was introduced:  
<https://github.com/n-witt/EconstorCorpus>

# END

 /n-witt

n.witt@zbw.eu