

# Workshop: Applying Quantitative Text Analysis to Political Data

Annika Fredén, Karlstad University (Political science)  
Sverker Sikström, Lund University (Psychology)

Université de Montréal, 17 January 2020

# How are you?

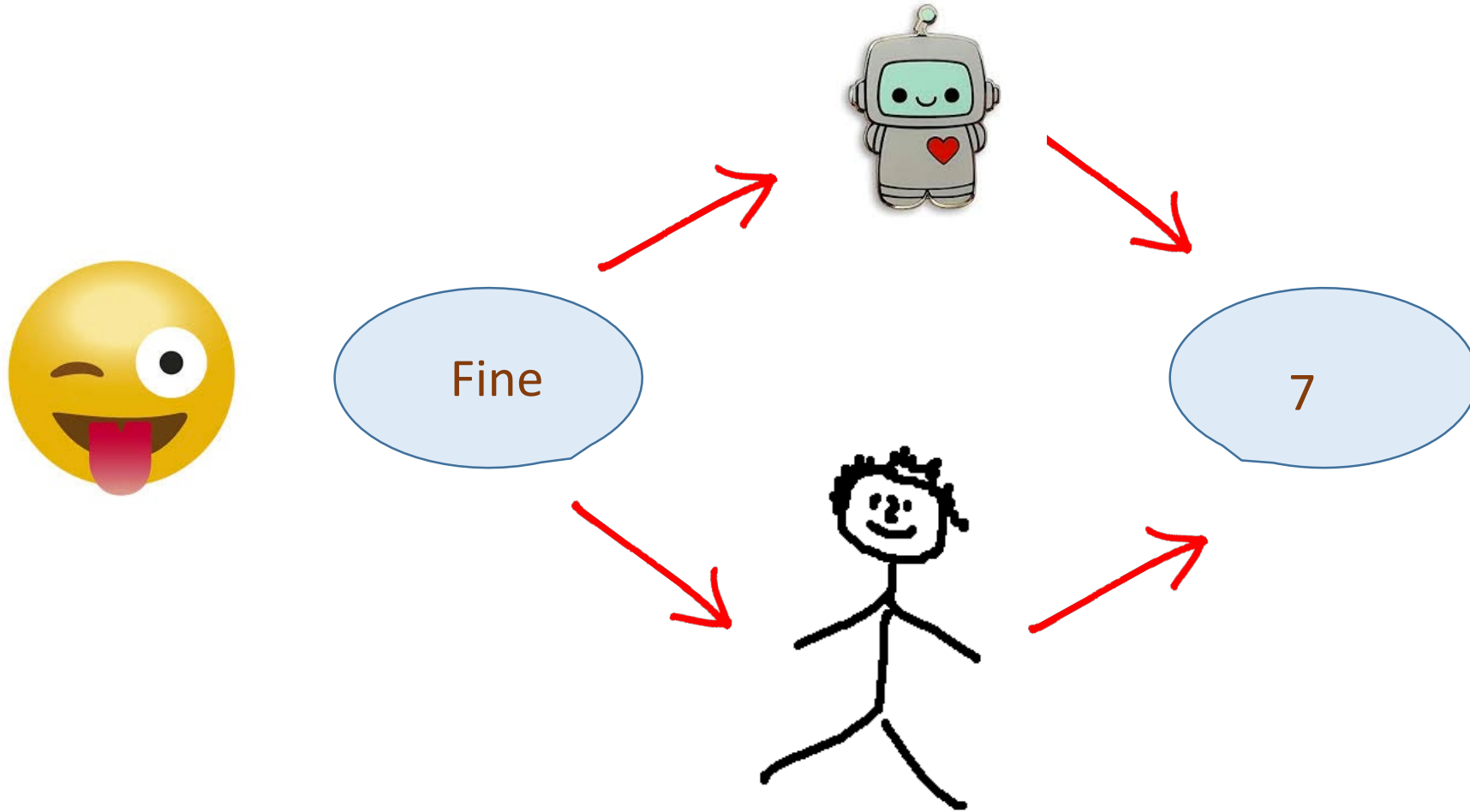


Fine

7



# Who is doing the measuring?





# How are concepts measured? Harmony



My lifestyle allows me to be in harmony

Most aspects of my life are in balance

I am in harmony

I accept the various conditions of my life

I fit in well with my surroundings

1    –    2    –    3    –    4    –    5  
Not at all                      Fully agree



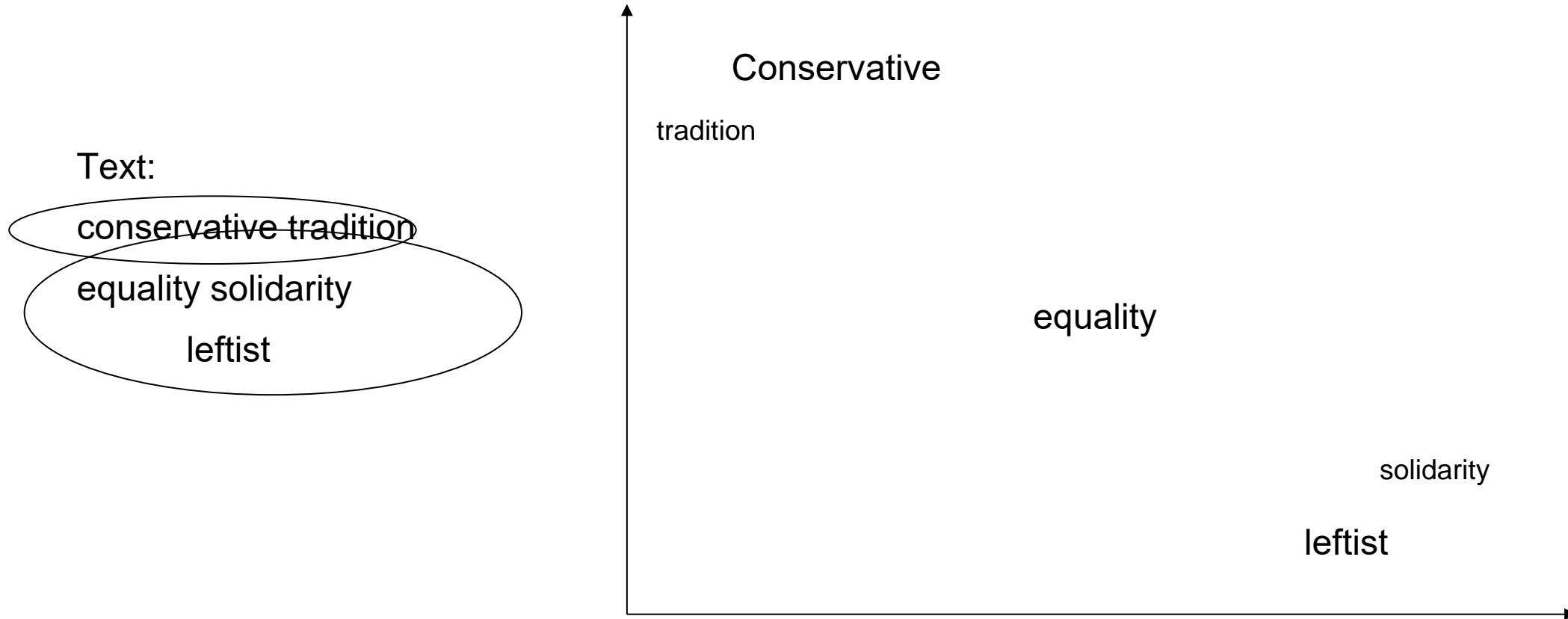
## Starting-Points

- Quantitative textual analysis rapidly growing in pol sci (e.g. Denny and Spirling, 2018; Roberts et al., 2014)
- Still little hands-on knowledge and understanding of the strengths and weaknesses of different methods
- Language and texts about AI and algorithms are often non-transparent and scholars tend to choose between a limited set of methods (e.g. LIWC, LSA, or Structural topic models), we will work with "unpacking" it
- Trial and error exercises with data from psychology and politics

## Dictionary-based vs data-driven

- Two main distinctions: Supervised/Dictionary-based (e.g. LIWC, Pennebaker et al. 2007) vs Unsupervised/Data-driven (e.g. LSA, Landauer and Dumais, 1997)
- Data-driven methods as LSA methods takes more knowledge and sound judgement from the scholar, and are more flexible
- Dictionary-based methods are more transparent, more easily replicated and tested in similar settings

# From Text to Space

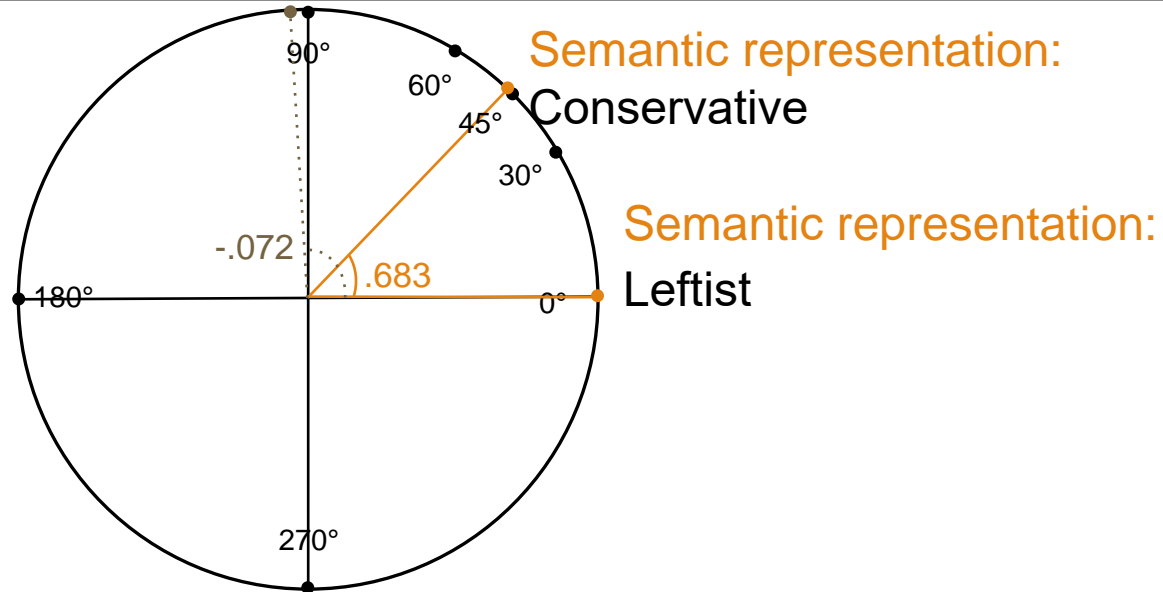


# Semantic similarity

Semantic representation = coordinates of a point in space

The semantic distance between two points can be represented by the cosine angle between them

Semantic representation: Apple



[https://www.rapidtables.com/calc/math/Cos\\_Calculator.htm](https://www.rapidtables.com/calc/math/Cos_Calculator.htm)



## Exercise 1- Test Semantic similarities and p-values in semanticexcel.com

- **Create your own account at semanticexcel.com** (works best from google chrome). There you can find a demo dataset that we will use for some of the exercises.
- **Semantic similarity**  
Fill two neighbor cells with two words (e.g. two political concepts) you would like to measure the semantic similarity between.  
Click on another cell and execute the function "Semantic similarity" via "Functions", referring to the cells where you wrote the words. Is the value what you expected? (Higher value indicates greater similarity.) You can test several words and compare their semantic similarity.
- **Semantic test**  
Fill two columns with words you would like to compare,.The columns should represent two different types of words, e.g. colors in one and political parties in another.  
Click on another cell and execute the function "Semantic test" via "Functions", referring to the two columns. The function generates significance tests and p-values. How do you interpret these? (Low p-value indicates significant differences between the two sets of words.)

# Predict voting from semantic space

Ord	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Voted Conservative	Point biserial correlation
gender equal	0.04	-0.03	0.10	0.00	0.09	0	0.25
socialist	0.06	-0.14	0.02	0.05	0.03	0	0.30
6 hours workday	0.06	-0.05	-0.05	-0.04	0.03	0	0.23
defense	0.03	-0.04	-0.02	-0.01	0.00	1	0.61
leadership	0.05	-0.06	0.00	-0.03	0.01	1	0.80
order	0.06	-0.09	-0.02	-0.01	-0.02	1	0.71
armed forces	0.03	-0.01	-0.02	0.01	0.00	?	0.78
right-wing	0.08	-0.17	0.04	0.07	0.00	?	0.97
tradition	0.07	0.01	-0.02	-0.01	0.10	?	0.76

**Training data**

**Leave out data  
(predicted values)**

## Exercise II: Training

- The **training function** associates words with numbers and can predict numeric outcome from words
- Fill one empty column in semanticexcel with 20 words: 10 *positive* and 10 *negative*
- Evaluate the degree of negativity/positivity of the words you have chosen on a scale from -5 to +5. Put the score of the specific word in the cell just to the right of the word itself, so that you obtain a numeric column besides the words column.
- Click on an empty cell on the right and execute the "training function" via Functions, referring to the two columns with words/numbers . You may also store the values in the column next to the words/numbers using the Advanced Options.
- **Interpret the outcome:** Do the predicted values differ from your evaluations? Which predictions were closer and which were farther from your evaluated score?

Application: Developing new semantic tools for measuring the relationships between party sympathy, social networks and vote choice *Swedish Research Council Project 2018-2021*

- Aim: Create different (and more valid) measures of party preference from leader descriptions, party descriptions, and issue preferences and associate these with behaviors (party choice). Elaborate on cognitive tendencies (e.g. which quality is described first or last, if concrete beats abstract, etc.)
- Data collection strategy: Collect words with as much meaning and as little noise as possible ("three words" or "two words")
- Cases so far: 2018 Swedish General election (Parties, Candidates, Issues) and Project Europe including 15 countries (with Blais and Bol): Leadership, Prime Ministers, Issues, and Emotions

# Collection of data - "three words" Swedish National Election 2018

## Lore Citizen Campaign Panel, University of Gothenburg

### Treatment:

### Party leaders described first

**GÖTEBORGS UNIVERSITET**

Vad representerar följande partiledare/språkrör för dig?  
Ange upp till tre beskrivande ord eller lämna fältet tomt om du inte känner till partiledaren/språkröret.

	Ord 1	Ord 2	Ord 3
Jan Björklund (L)	<input type="text"/>	<input type="text"/>	<input type="text"/>
Ulf Kristersson (M)	<input type="text"/>	<input type="text"/>	<input type="text"/>
Ebba Busch Thor (KD)	<input type="text"/>	<input type="text"/>	<input type="text"/>
Isabella Lövin (MP)	<input type="text"/>	<input type="text"/>	<input type="text"/>



**Medborgarpanelen**

**GÖTEBORGS UNIVERSITET**

Vad representerar följande partier för dig?  
Ange upp till tre beskrivande ord eller lämna fältet tomt om du inte känner till partiet.

	Ord 1	Ord 2	Ord 3
Liberalema	<input type="text"/>	<input type="text"/>	<input type="text"/>
Moderaterna	<input type="text"/>	<input type="text"/>	<input type="text"/>
Kristdemokraterna	<input type="text"/>	<input type="text"/>	<input type="text"/>
Miljöpartiet	<input type="text"/>	<input type="text"/>	<input type="text"/>



**Medborgarpanelen**

## Tentative Results: Word descriptions of Christian Democrats Younger vs Older participants (LSA)

Low x-axis  
[ $r=0.21, p=0.0000$ ]

High x-axis  
[r=0.21, p=0.0000]

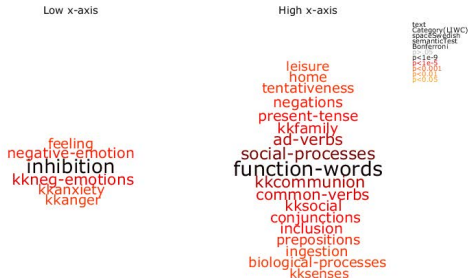


text  
words  
space  
Swedish  
semantic  
test  
Bonferroni

# Tentative Results

## Word descriptions of Christian Democrats

### Younger vs Older participants (LIWC)



### Exercise III:

#### Plot function - comparing LIWC and LSA

- Use one of the word columns from the demo dataset, concerning satisfaction.
- Use plot function from Functions, where you may choose from different types of plots: LIWC or words.
- Make one plot with LIWC, and another with LSA, associating a word column with a numeric outcome
- Interpret the results: Do they look like you expected?



## Read More

Denny, M.J and Spirling, A. 2018. Text Preprocessing for Unsupervised Learning: Why it Matters, When it Misleads, and What to do about it. *Political Analysis* 26, 168-189.

Landauer, T. K., & Dumais, S. T. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211-240.

Roberts, M. E., B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. Kushner Gadarian, B. Albertson and D. Rand. 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science* 58(4), 1064–1082.

Sikström, S., Kjell, O. N. E., & Kjell, K. 2018. Semantic Excel: An Introduction to a User-Friendly Online Software Application for Statistical Analyses of Text Data. <https://doi.org/10.31234/osf.io/z9chp>

Thanks to Alexander Rangfält for research assistance