CHARACTERIZING THE ECOSYSTEM OF IDEAS IN TEXTS

Chenhao Tan
Department of Computer Science
University of Colorado Boulder
https://chenhaot.com
chenhao@chenhaot.com

TEXT CORPORA ARE UBIQUITOUS

TEXT CORPORA ARE UBIQUITOUS

reddit

The New York Times

CNN

TEXT IS VALUABLE

TEXT IS VALUABLE

- Text allows us to measure a proxy of the world
  - Topics of discussion [Grimmer et al. 2014; Nguyen et al. 2012]
  - Attitude & opinion [Dodds and Danforth 2009; Stuart and Young, 2012]
  - History of ideas [Hall et al. 2008; Uzzi et al. 2013]
  - Bias & polarization [Garg et al. 2018; Niculae et al. 2015; Tan et al. 2018]
- Text allows us to understand social interaction and human behavior
  - Persuasion [Flynn et al. 2017; Tan et al. 2016]
  - Linguistic coordination [DNM et al. 2013; Doyle et al. 2017]
  - Diffusion of norms [Amato et al. 2018; Eisenstein et al. 2014]
  - Mental health [Althoff et al. 2106; Coppersmith et al. 2014]
- Text allows us to understand language

AUTOMATED TEXT ANALYSIS

“QUANTITATIVE METHODS FOR TEXT AMPLIFY RESOURCES AND AUGMENT HUMANS.”

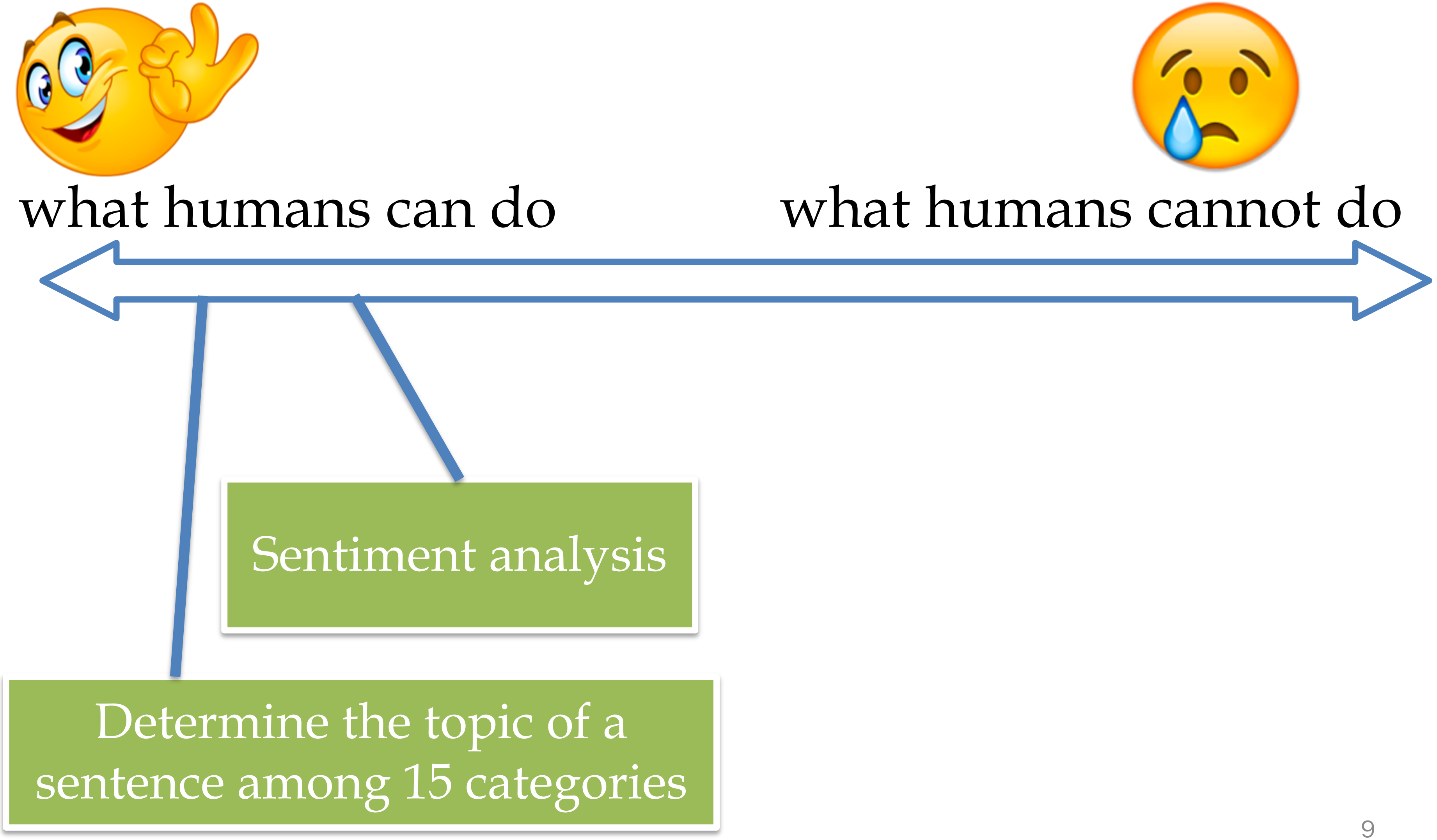
Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts, Grimmer and Stewart 2013.

AUTOMATED TEXT ANALYSIS TO AUGMENT HUMANS

what humans can do

what humans cannot do

AUTOMATED TEXT ANALYSIS TO AUGMENT HUMANS



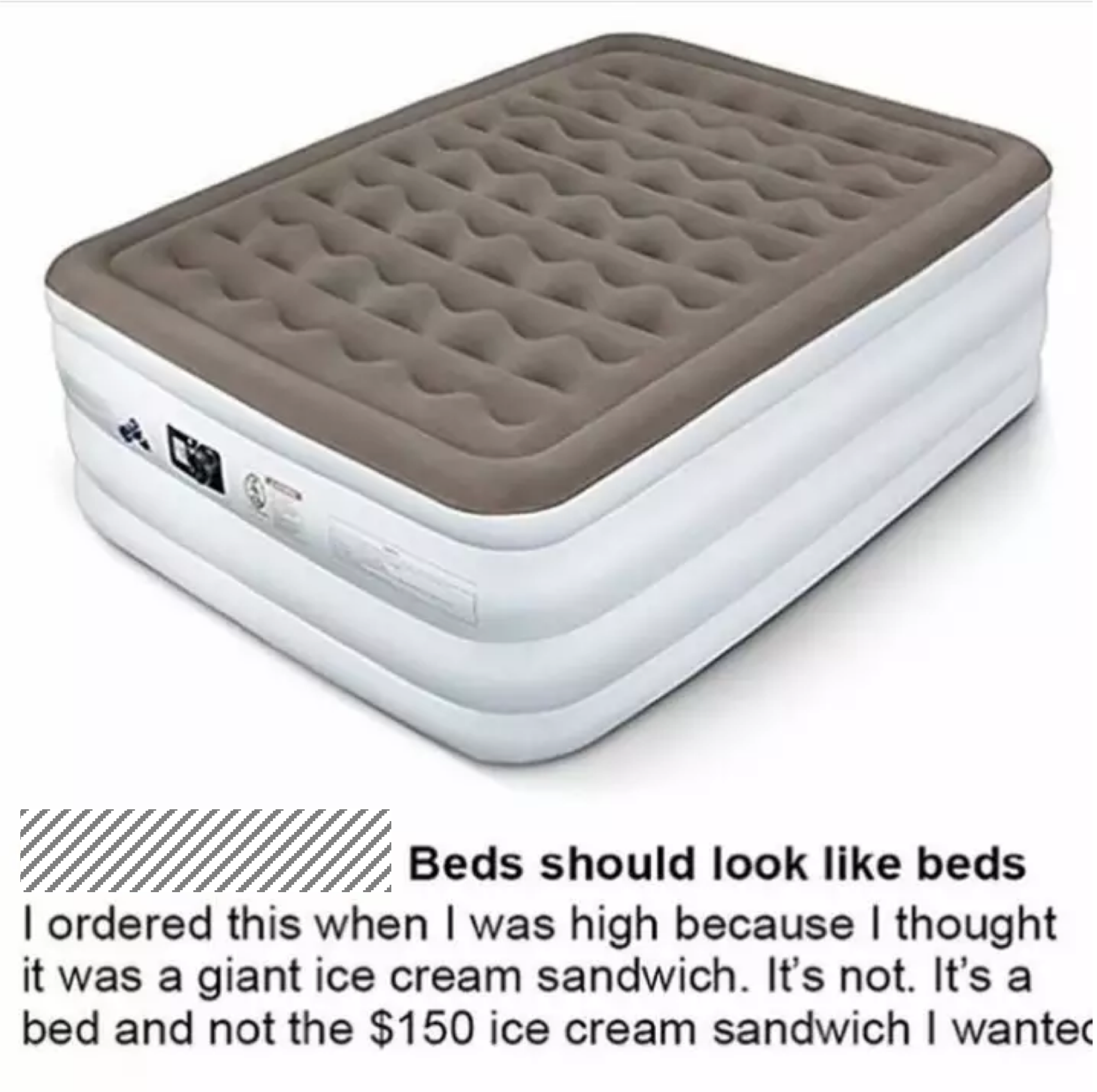
what humans can do

what humans cannot do

Sentiment analysis

Determine the topic of a sentence among 15 categories

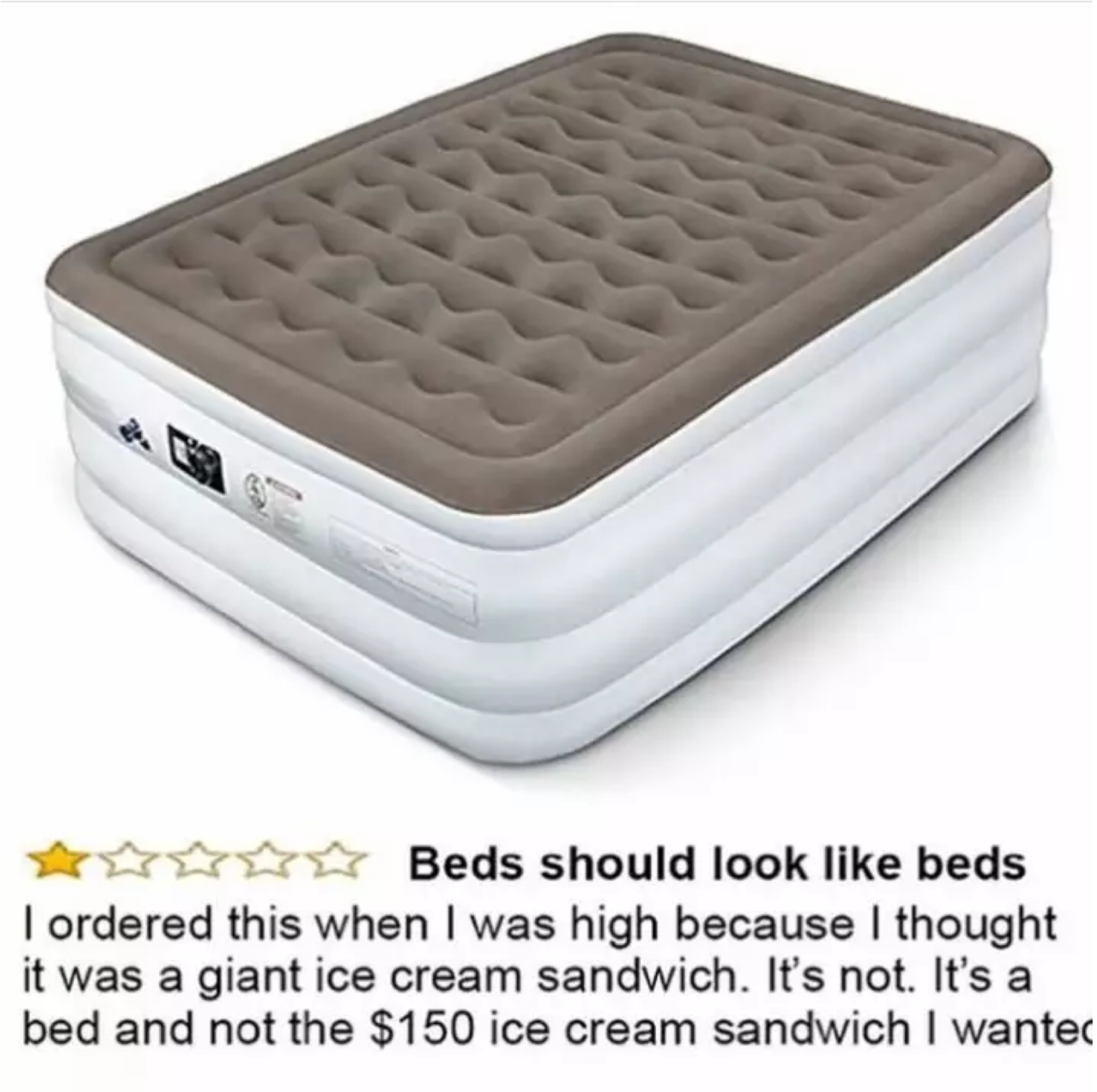
SENTIMENT ANALYSIS: IS THIS REVIEW POSITIVE OR NEGATIVE?



[redacted] **Beds should look like beds**

I ordered this when I was high because I thought it was a giant ice cream sandwich. It's not. It's a bed and not the $150 ice cream sandwich I wantec

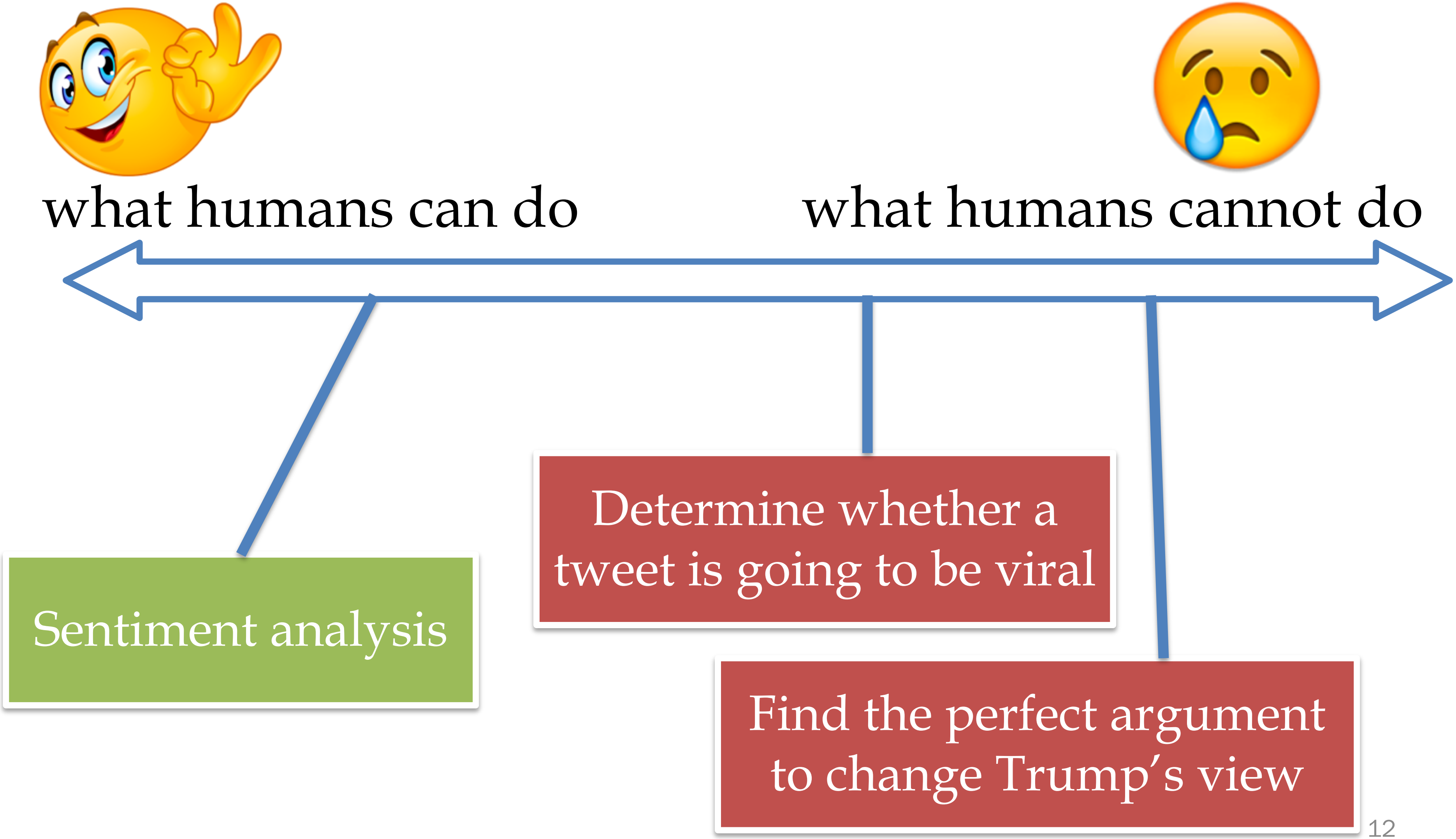
SENTIMENT ANALYSIS: IS THIS REVIEW POSITIVE OR NEGATIVE?



★☆☆☆☆ **Beds should look like beds**

I ordered this when I was high because I thought it was a giant ice cream sandwich. It's not. It's a bed and not the $150 ice cream sandwich I wantec

AUTOMATED TEXT ANALYSIS TO AUGMENT HUMANS



what humans can do

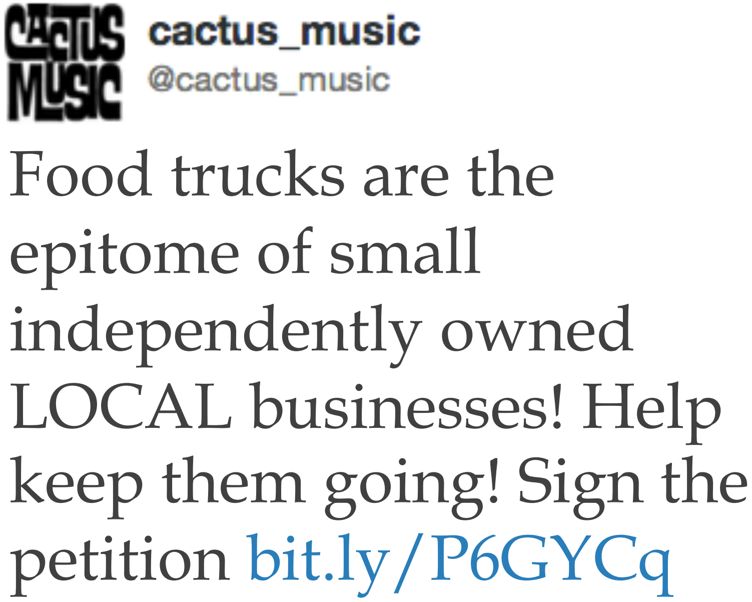
what humans cannot do

Sentiment analysis

Determine whether a tweet is going to be viral

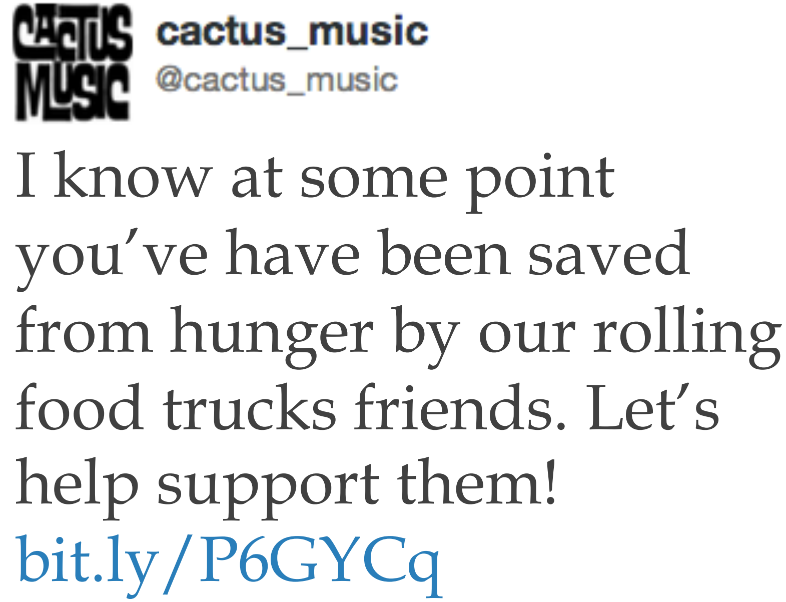
Find the perfect argument to change Trump's view

WHICH TWEET WAS RETWEETED MORE?



cactus_music
@cactus_music

Food trucks are the epitome of small independently owned LOCAL businesses! Help keep them going! Sign the petition bit.ly/P6GYCq

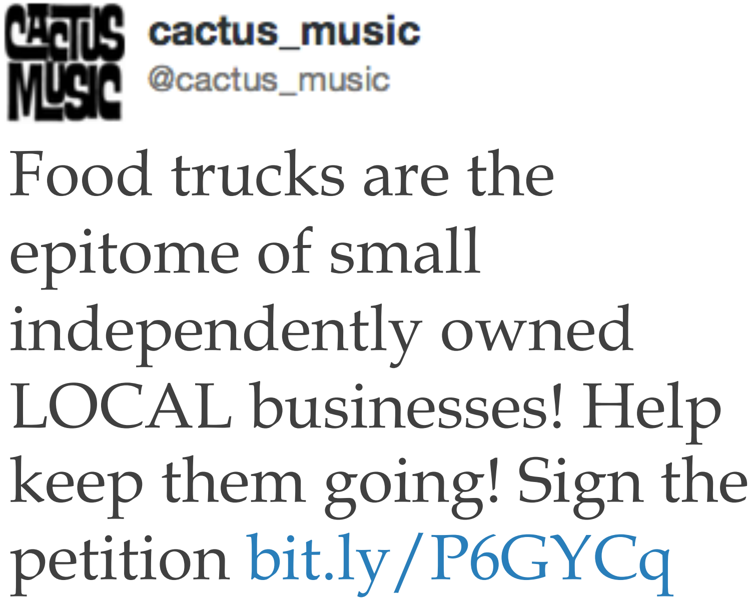


cactus_music
@cactus_music

I know at some point you've have been saved from hunger by our rolling food trucks friends. Let's help support them! bit.ly/P6GYCq

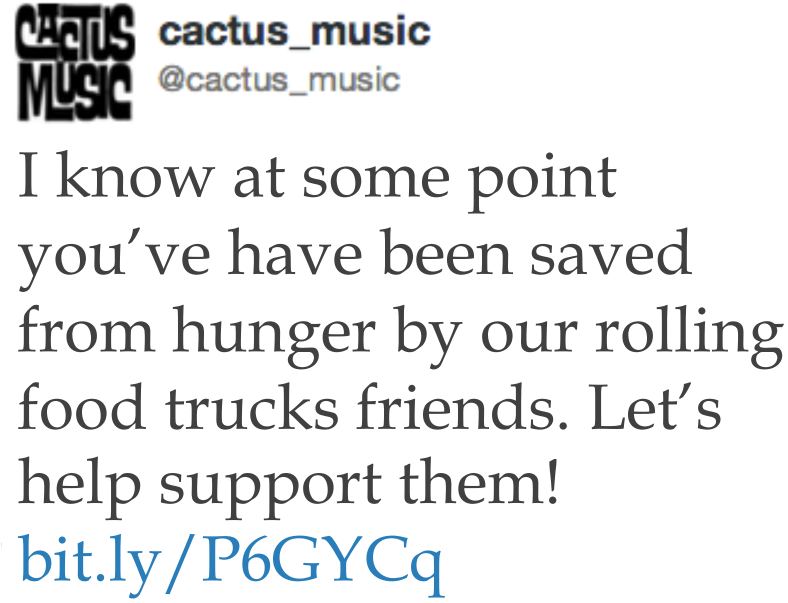
Tan, Pang and Lee 2014

WHICH TWEET WAS RETWEETED MORE?



cactus_music
@cactus_music

Food trucks are the epitome of small independently owned LOCAL businesses! Help keep them going! Sign the petition bit.ly/P6GYCq

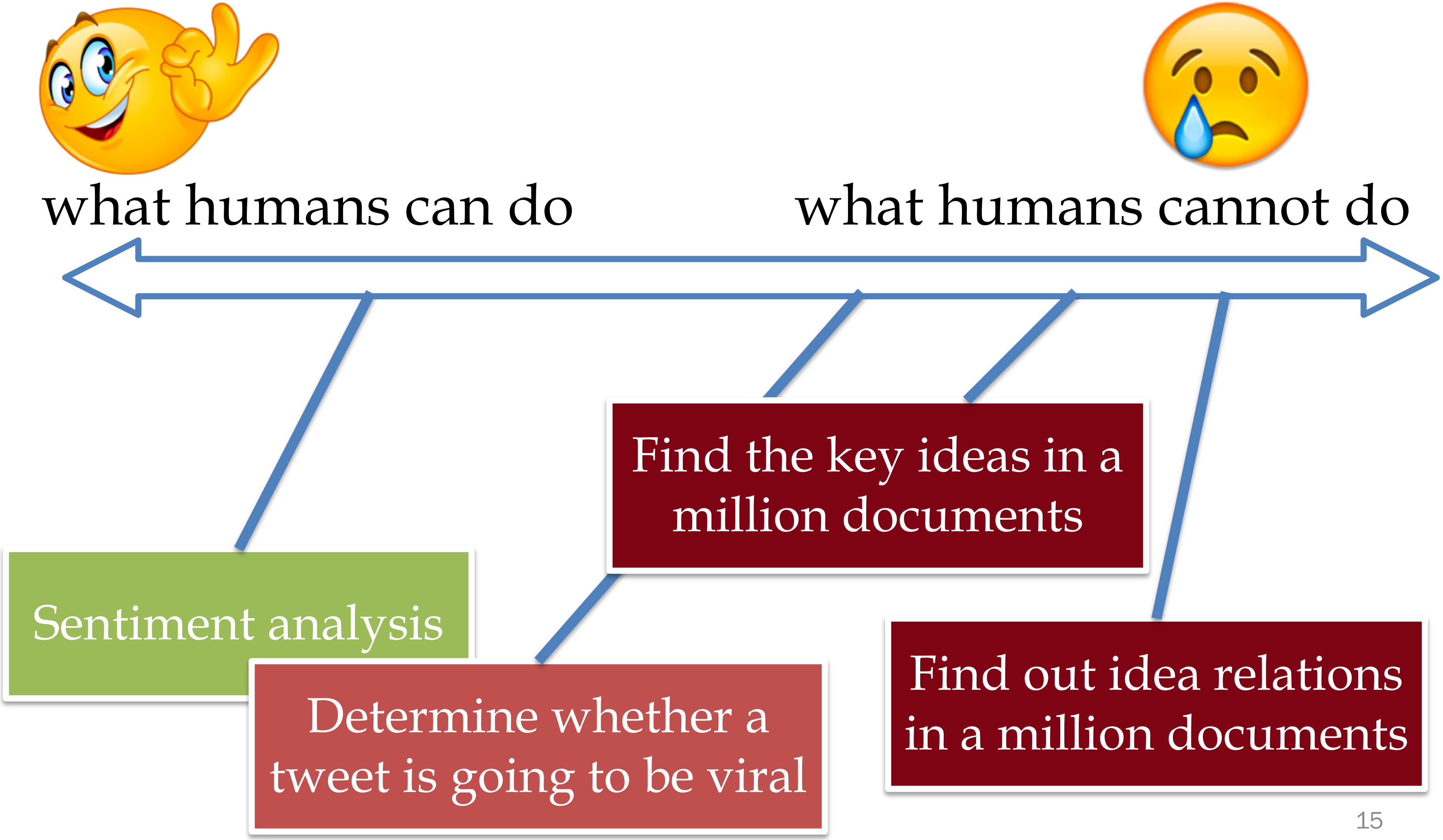


cactus_music
@cactus_music

I know at some point you've have been saved from hunger by our rolling food trucks friends. Let's help support them! bit.ly/P6GYCq

Tan, Pang and Lee 2014

AUTOMATED TEXT ANALYSIS TO AUGMENT HUMANS



what humans can do

what humans cannot do

Sentiment analysis

Determine whether a tweet is going to be viral

Find the key ideas in a million documents

Find out idea relations in a million documents

ANNO MDCCXLII

AUTOMATED TEXT ANALYSIS TO AUGMENT HUMANS

what humans can do

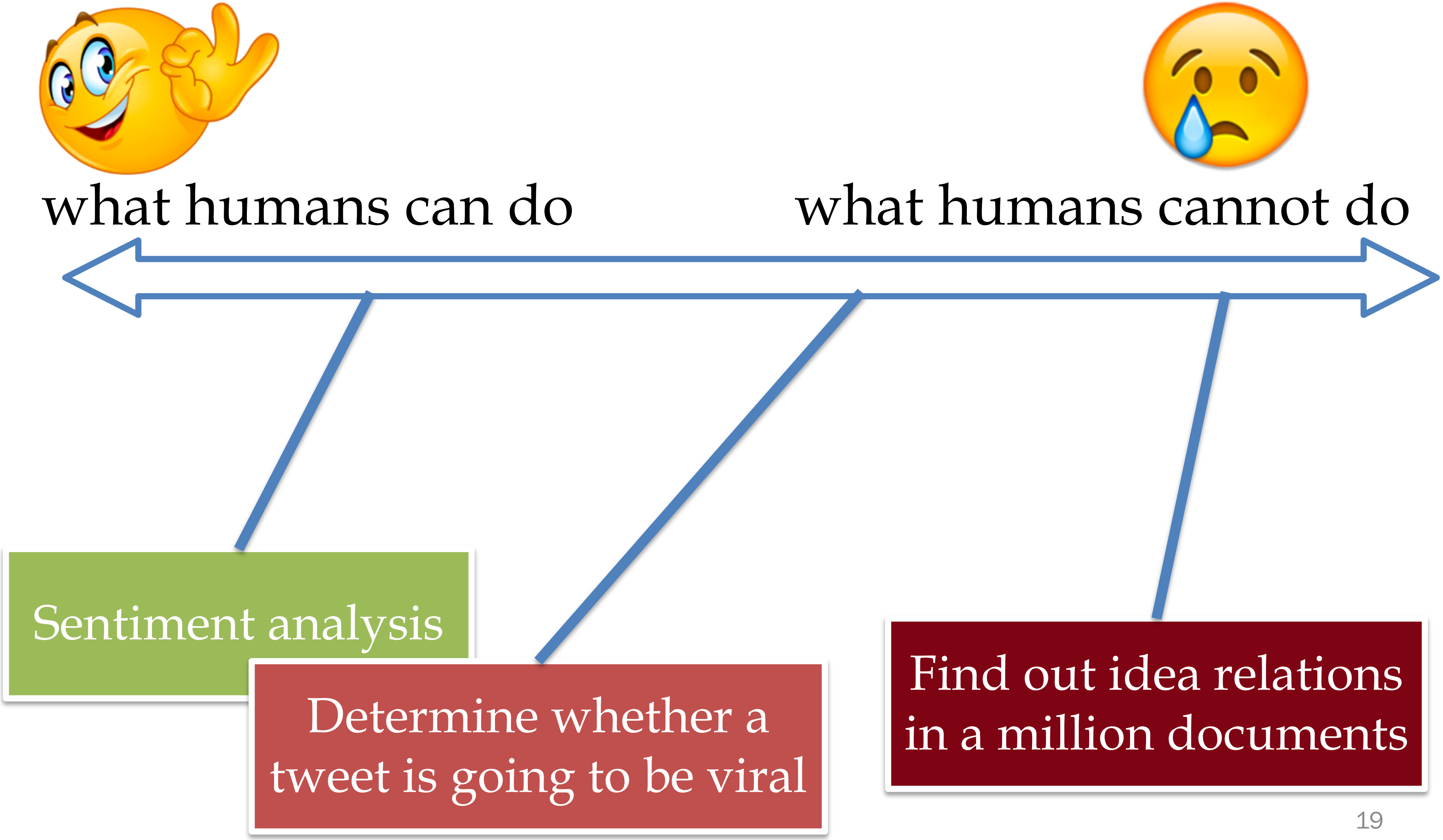
what humans cannot do

- What approach to take
- How to interpret the results
- How to evaluate the results

“THERE IS NO GLOBALLY BEST METHOD FOR AUTOMATED TEXT ANALYSIS.”

Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts, Grimmer and Stewart 2013.

AUTOMATED TEXT ANALYSIS TO AUGMENT HUMANS



what humans can do

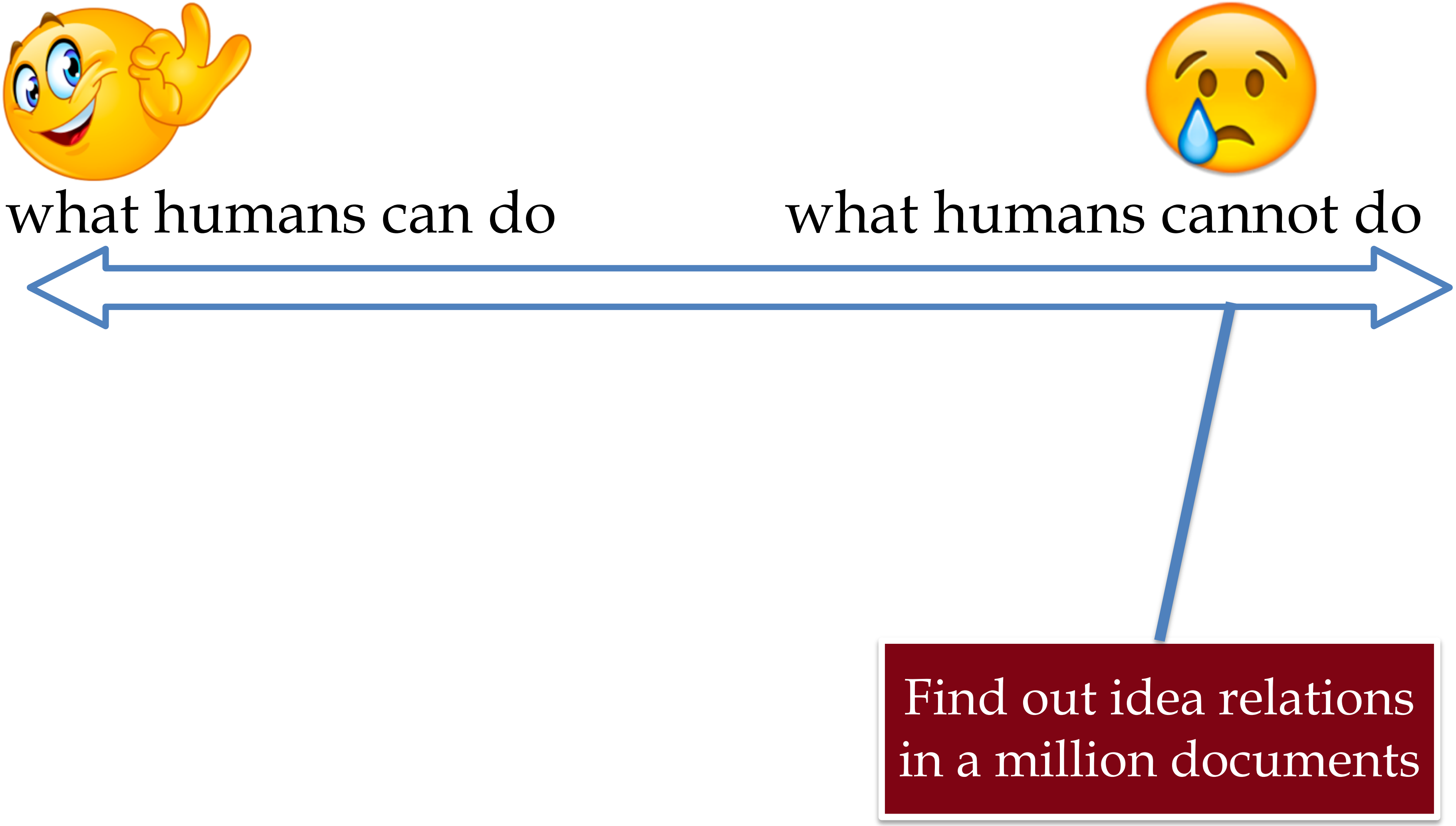
what humans cannot do

Sentiment analysis

Determine whether a tweet is going to be viral

Find out idea relations in a million documents

AUTOMATED TEXT ANALYSIS TO AUGMENT HUMANS



what humans can do

what humans cannot do

Find out idea relations in a million documents

ANNO MDCCXLIII.

Ecosystems of ideas

THE ECOSYSTEM OF IDEAS

- Competition and collaboration
  - Natural selection [Dawkins 1976]
  - Marketplace of ideas [Milton 1644; Mill 1859]

    Tan, Card, Smith, ACL'17
- Evolution of ideas
  - Telephone game

    Tan, Friggeri, Adamic, ICWSM'16

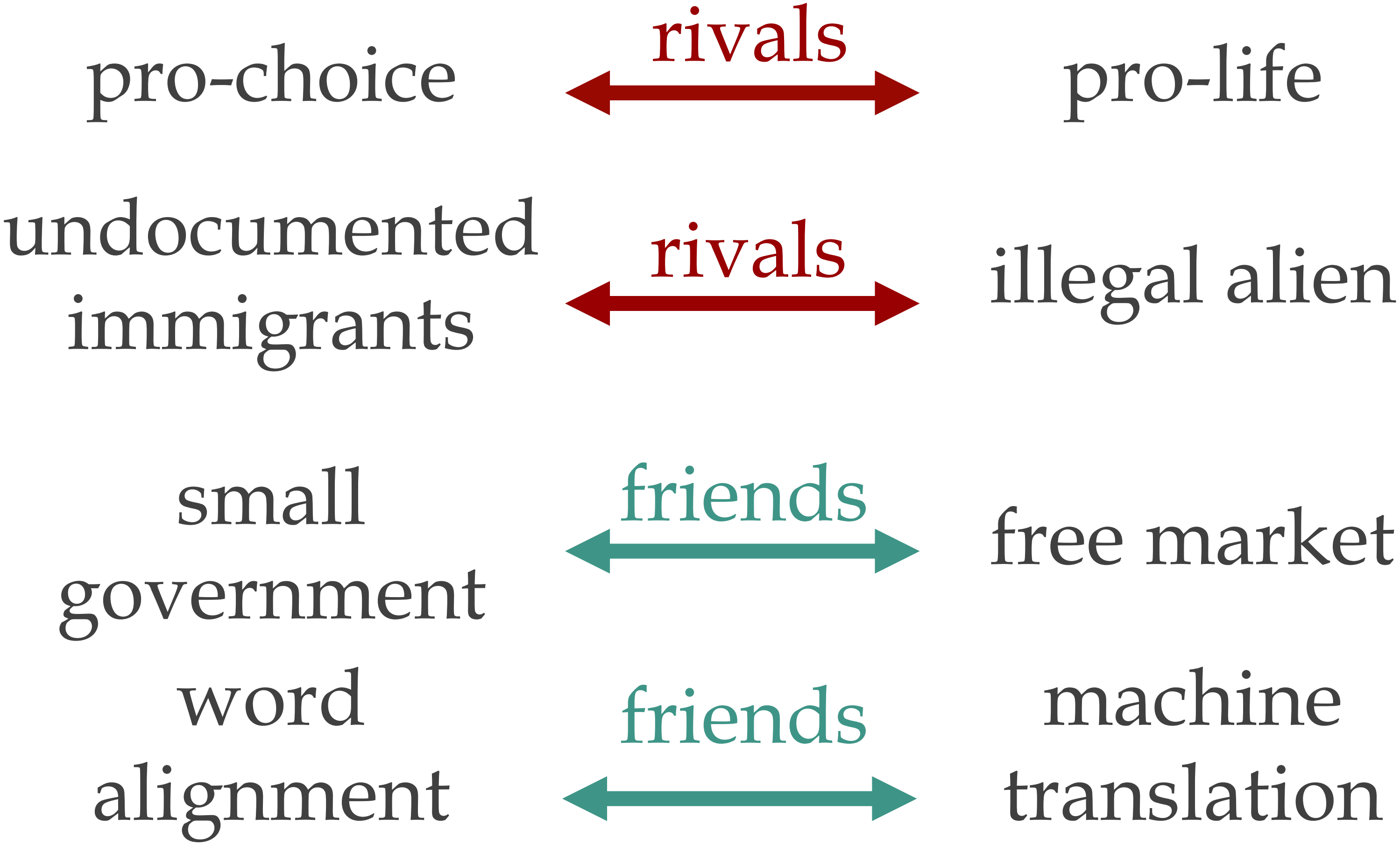
THE ECOSYSTEM OF IDEAS

- Competition and collaboration
  - Natural selection [Dawkins 1976]
  - Marketplace of ideas [Milton 1644; Mill 1859]

    Tan, Card, Smith, ACL'17
- Evolution of ideas
  - Telephone game

    Tan, Friggeri, Adamic, ICWSM'16

RELATIONS BETWEEN IDEAS



pro-choice rivals pro-life

undocumented immigrants rivals illegal alien

small government friends free market

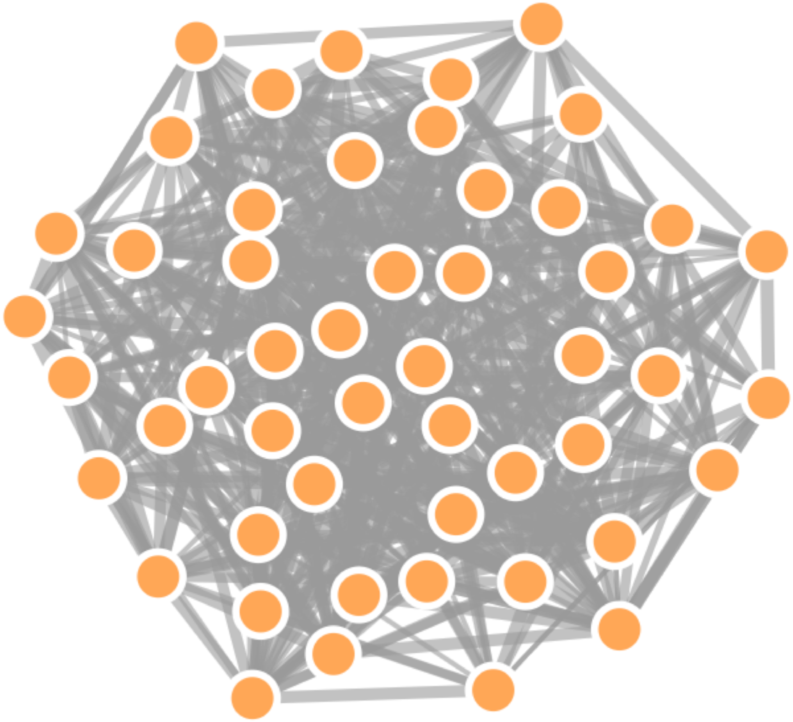
word alignment friends machine translation

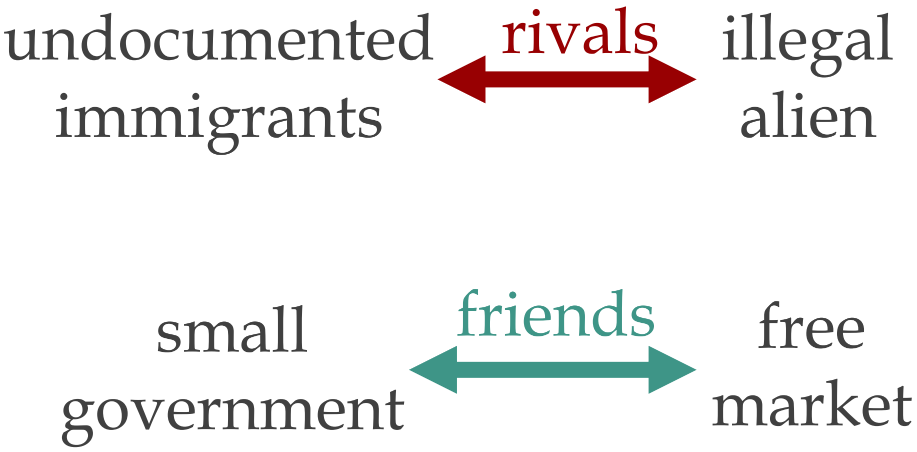
Chong and Druckman, 2007; Dawkins 1976; Entman, 1993; Gitlin, 1980; Lakoff, 2014; Milton 1964

MAIN CONTRIBUTIONS

First **quantitative** framework to systematically describe relations between ideas

Demonstrate **effective explorations** with this framework on a wide range of datasets





undocumented immigrants

rivals

illegal alien

small government

friends

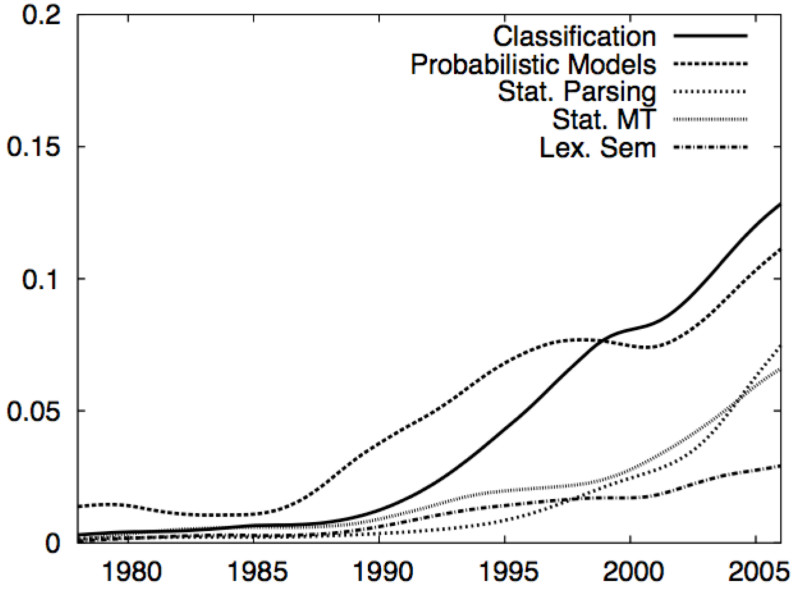
free market

USING TEXT TO TRACE IDEAS

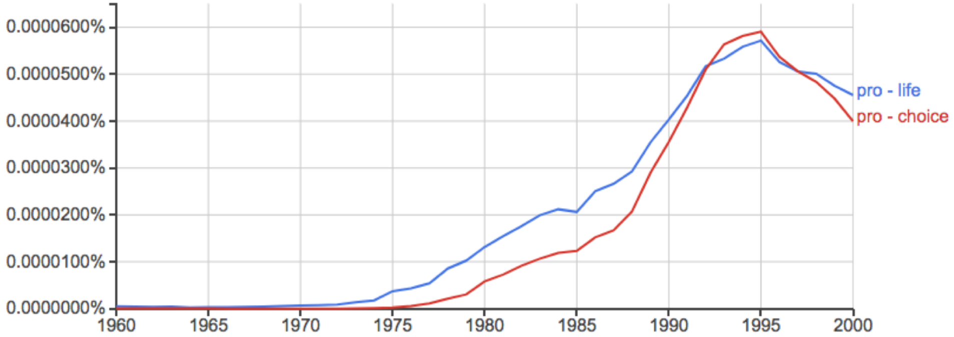
Our focus is on relations between ideas.

We will use standard approaches

- Topics from latent Dirichlet allocation (Blei et al. 2003)
- Keywords (Monroe et al. 2008)



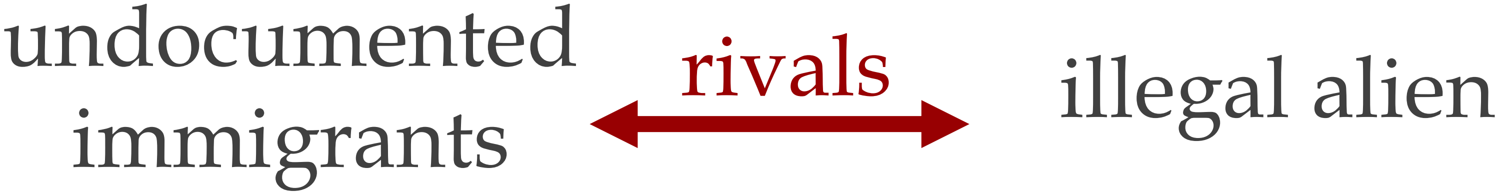
Hall et al. 2008



Culturomics, Michel et al. 2011

QUANTITATIVELY DESCRIBE RELATIONS BETWEEN IDEAS

- Given a corpus of documents over time, each document consists of a set of ideas



undocumented immigrants

rivals

illegal alien

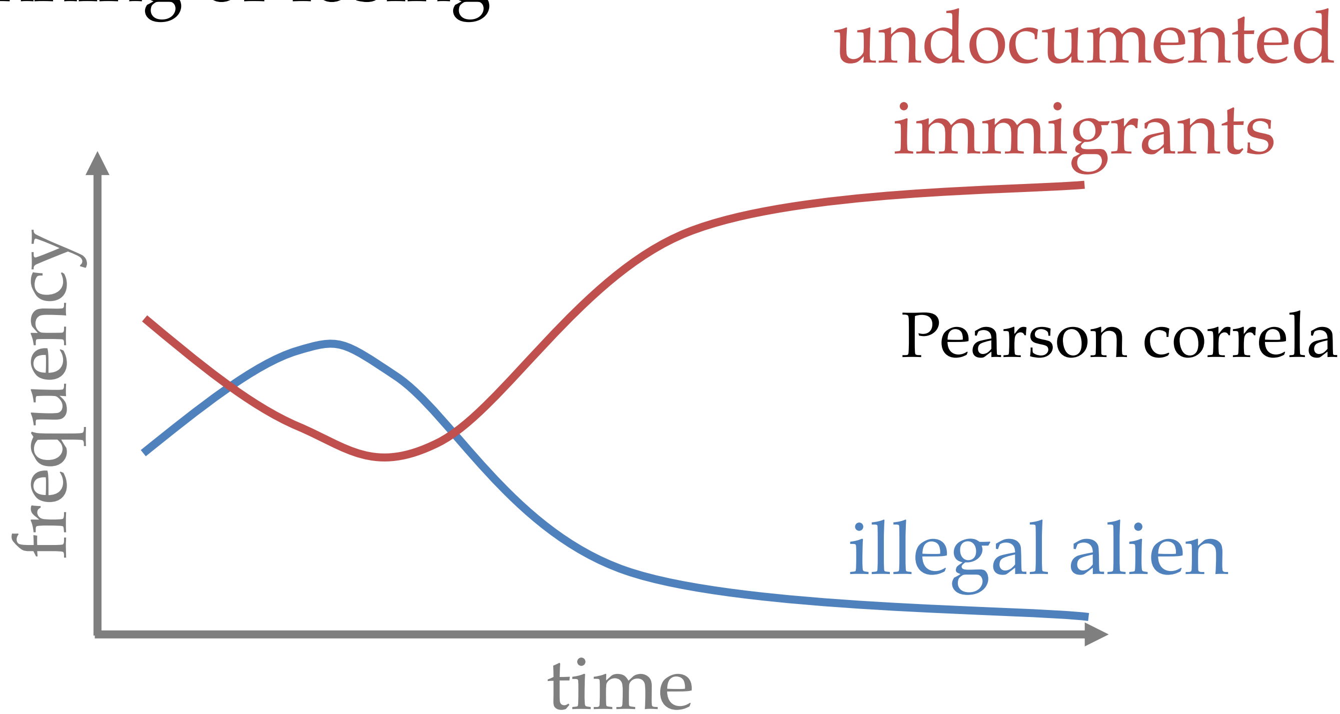
Cooccurrence

Pointwise mutual information [Church and Hanks 1990]

Rarely cooccur

QUANTITATIVELY DESCRIBE RELATIONS BETWEEN IDEAS

- Given a corpus of documents over time, each document consists of a set of ideas
  - Cooccurrence does not capture which is winning or losing



Pearson correlation

QUANTITATIVELY DESCRIBE RELATIONS BETWEEN IDEAS

- Given a corpus of documents over time, each document consists of a set of ideas

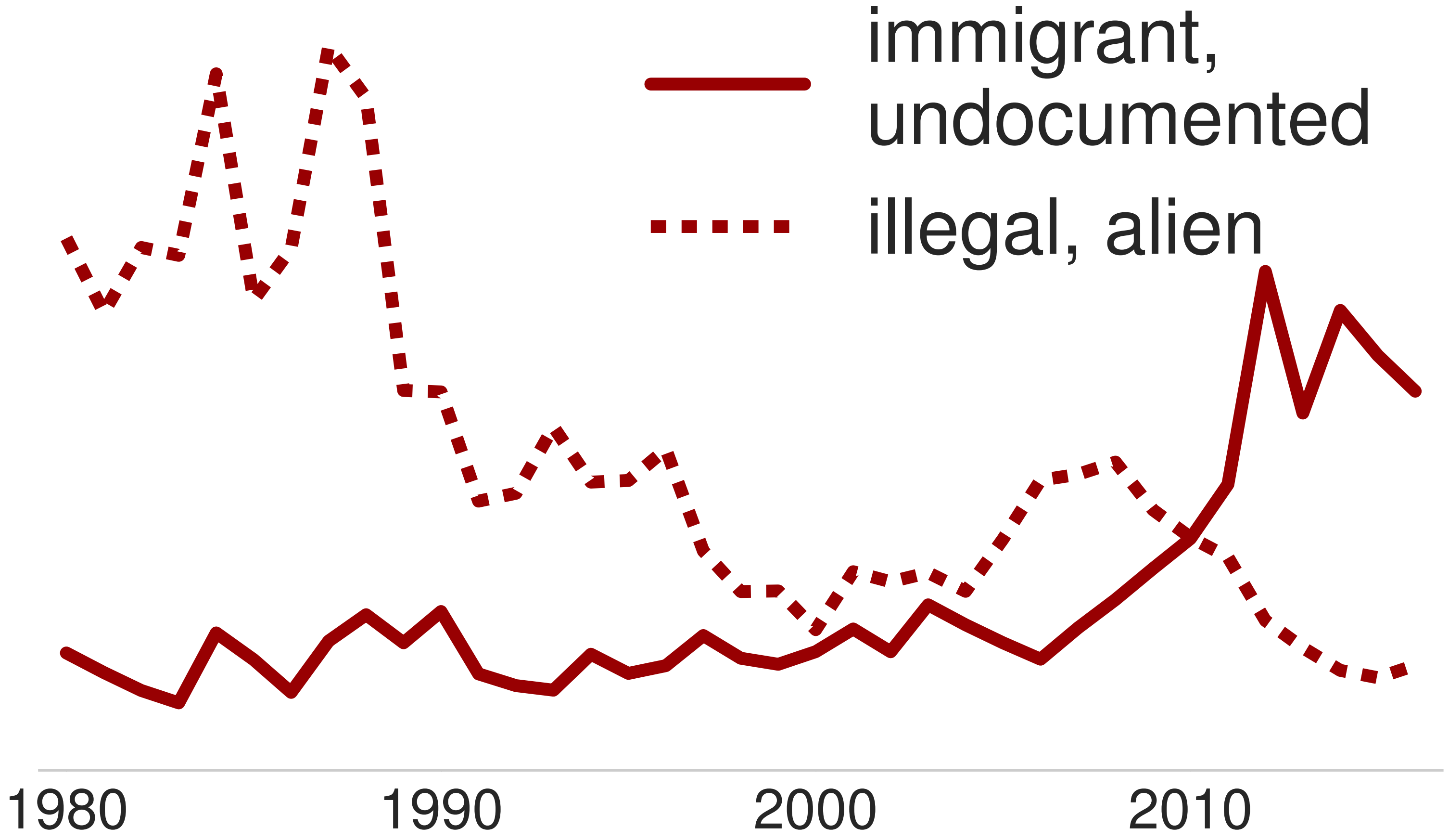
Cooccurrence & Prevalence correlation

Within-document

Across-document

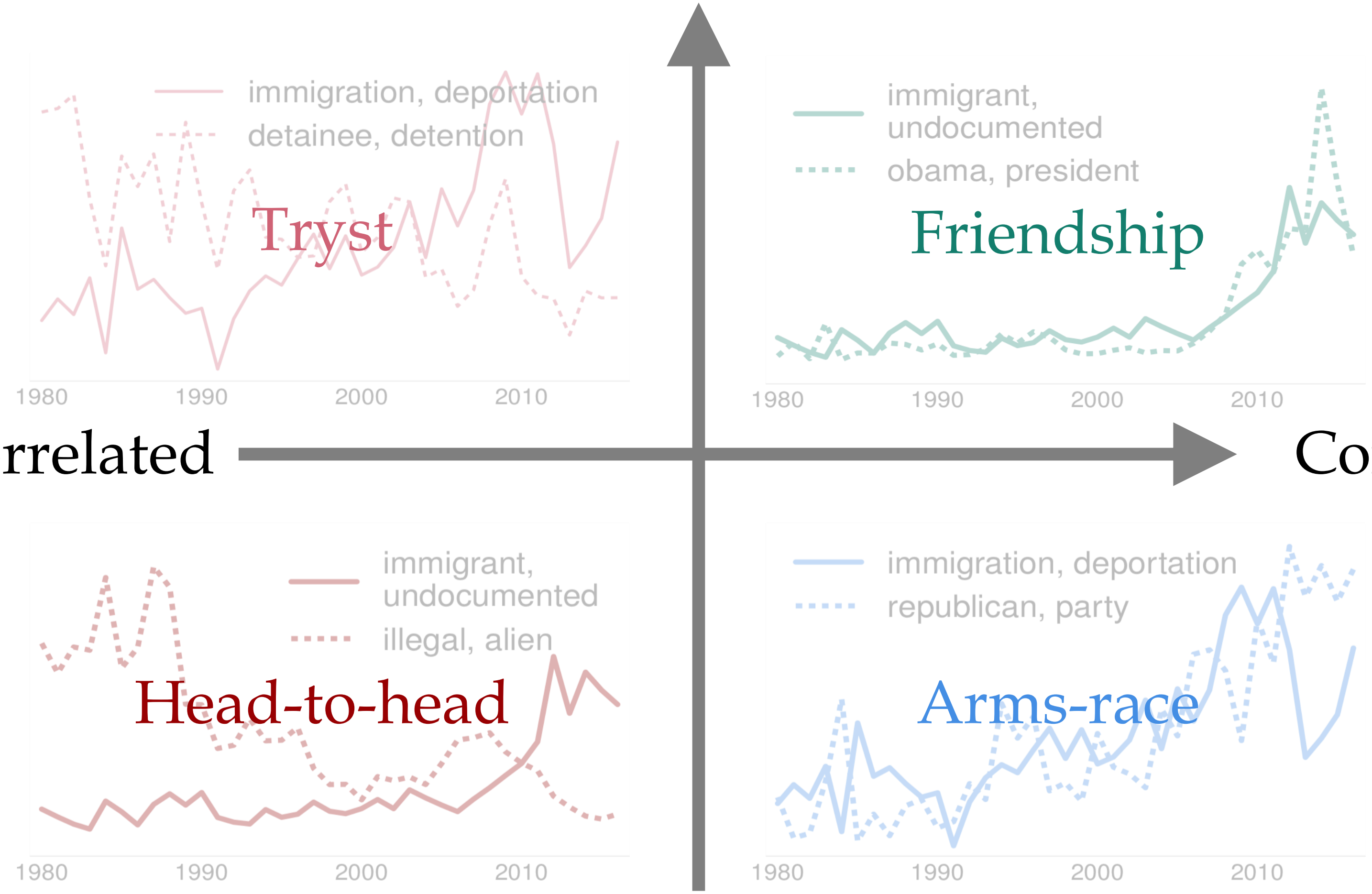
HEAD-TO-HEAD

(ANTI-CORRELATED, RARELY COOCCUR)



RELATIONS BETWEEN IDEAS

Always cooccur



Tryst

Friendship

Anti-correlated

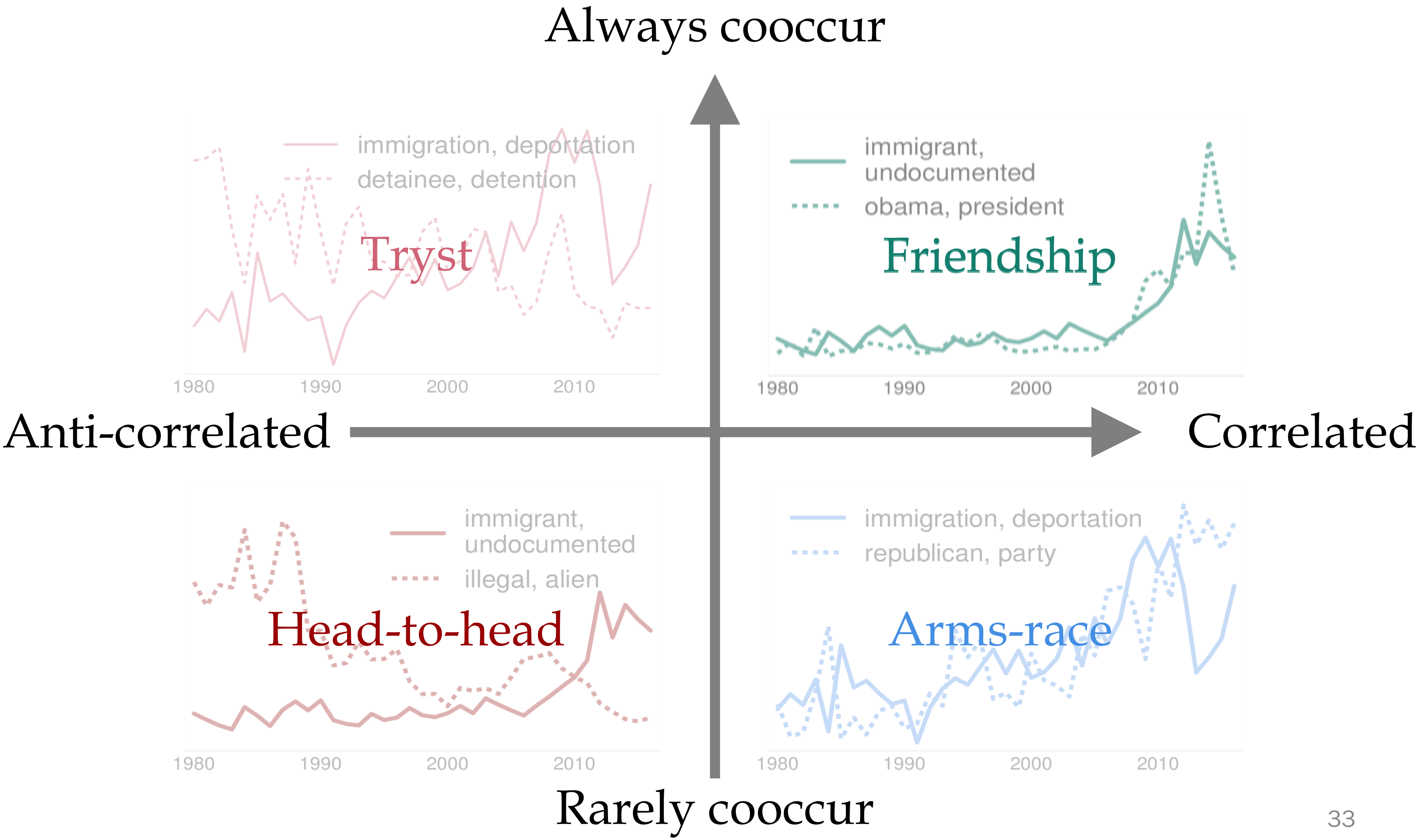
Correlated

Head-to-head

Arms-race

Rarely cooccur

RELATIONS BETWEEN IDEAS



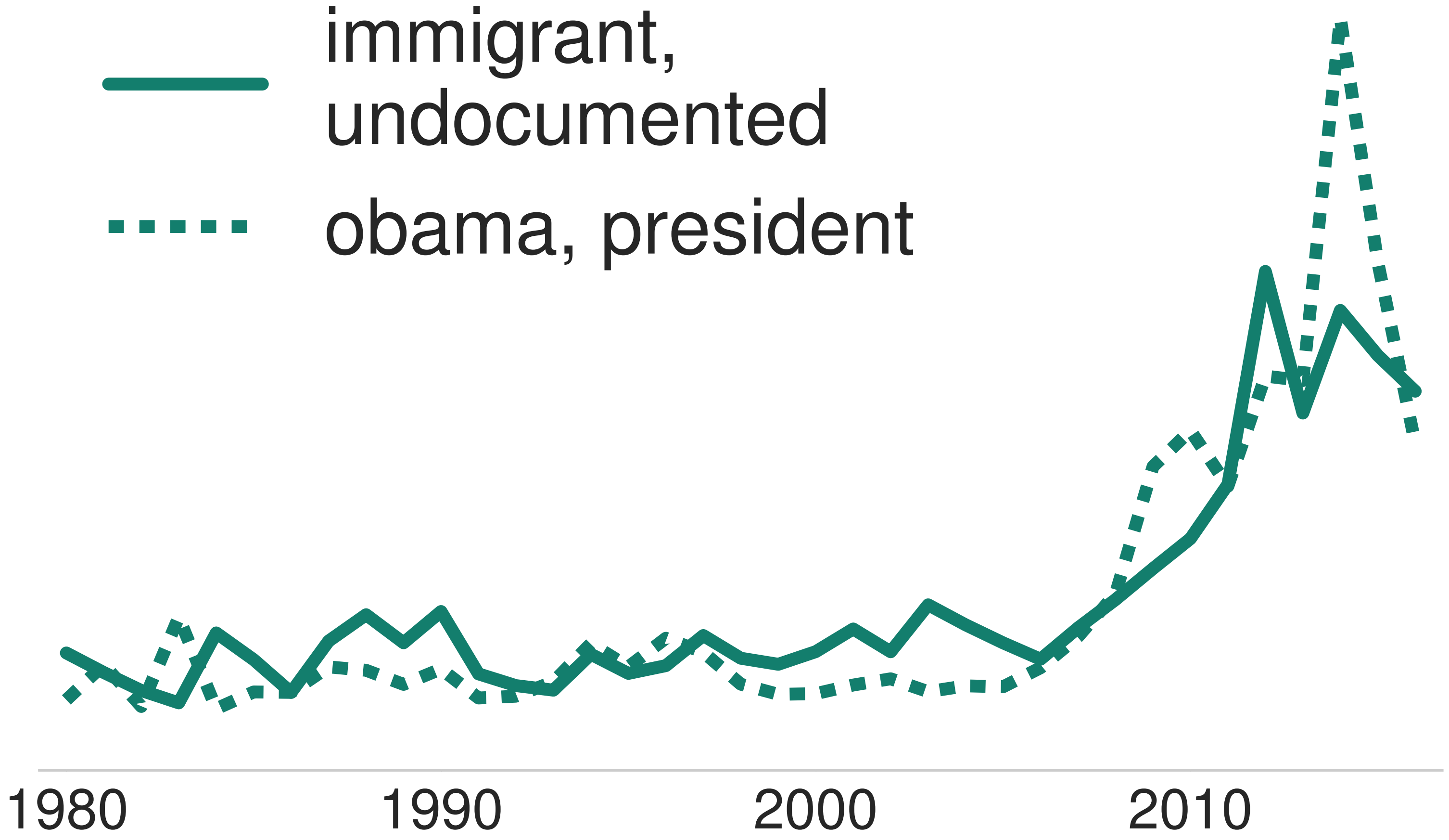
Always cooccur

Anti-correlated

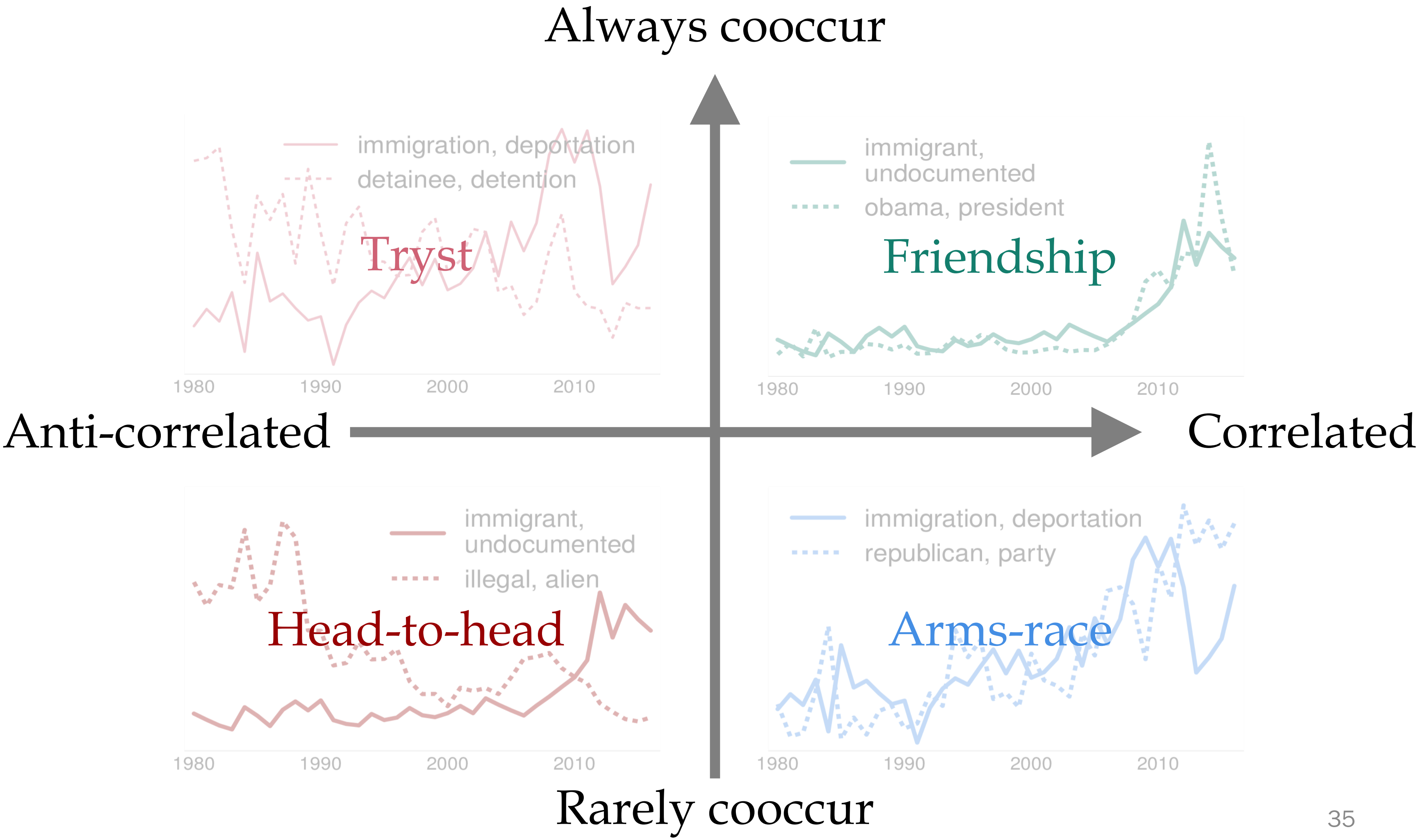
Correlated

Rarely cooccur

FRIENDSHIP
(CORRELATED, LIKELY TO COOCCUR)



RELATIONS BETWEEN IDEAS



Always cooccur

Anti-correlated

Correlated

Rarely cooccur

Tryst

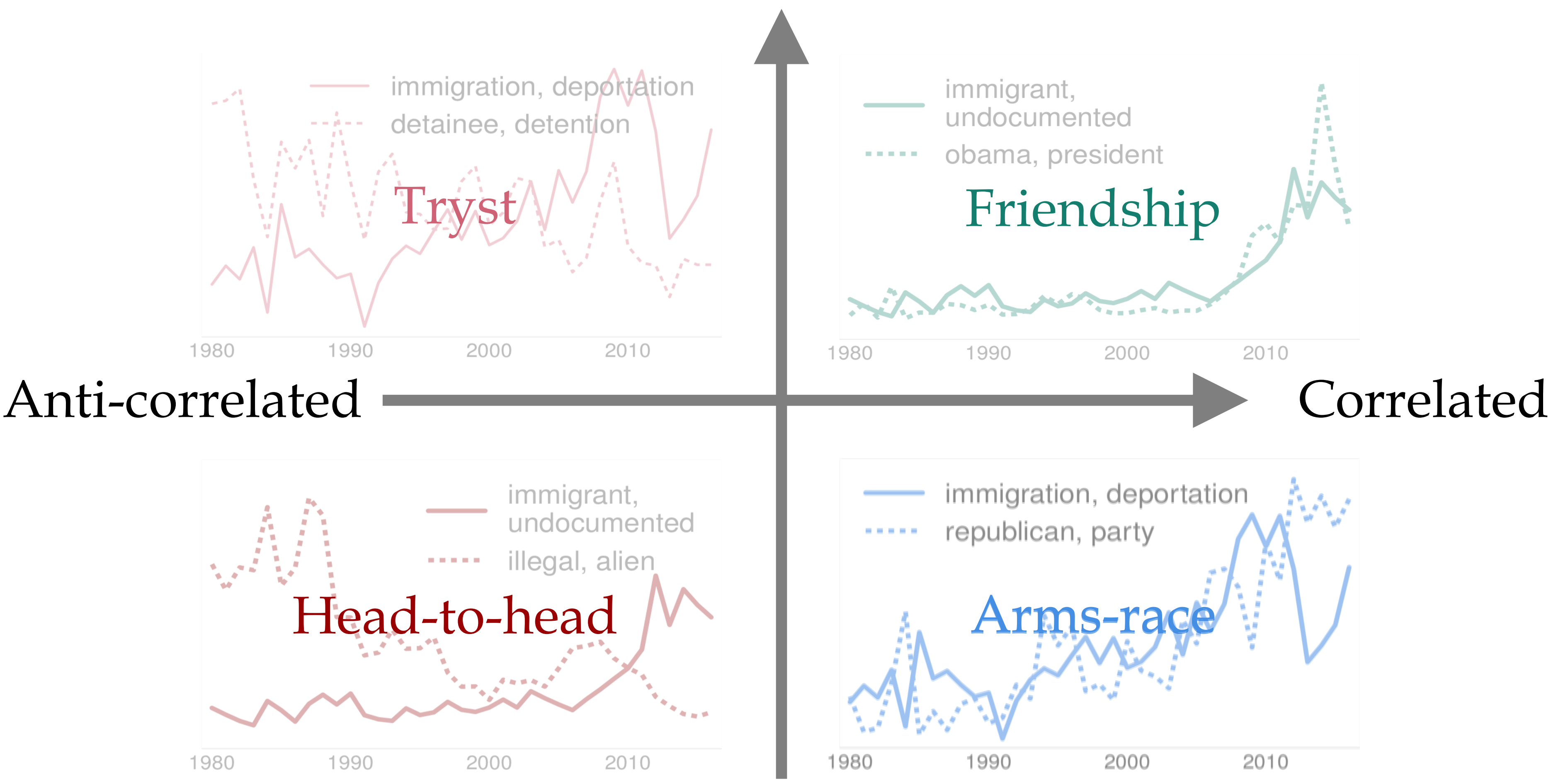
Friendship

Head-to-head

Arms-race

RELATIONS BETWEEN IDEAS

Always cooccur



Anti-correlated

Correlated

Tryst

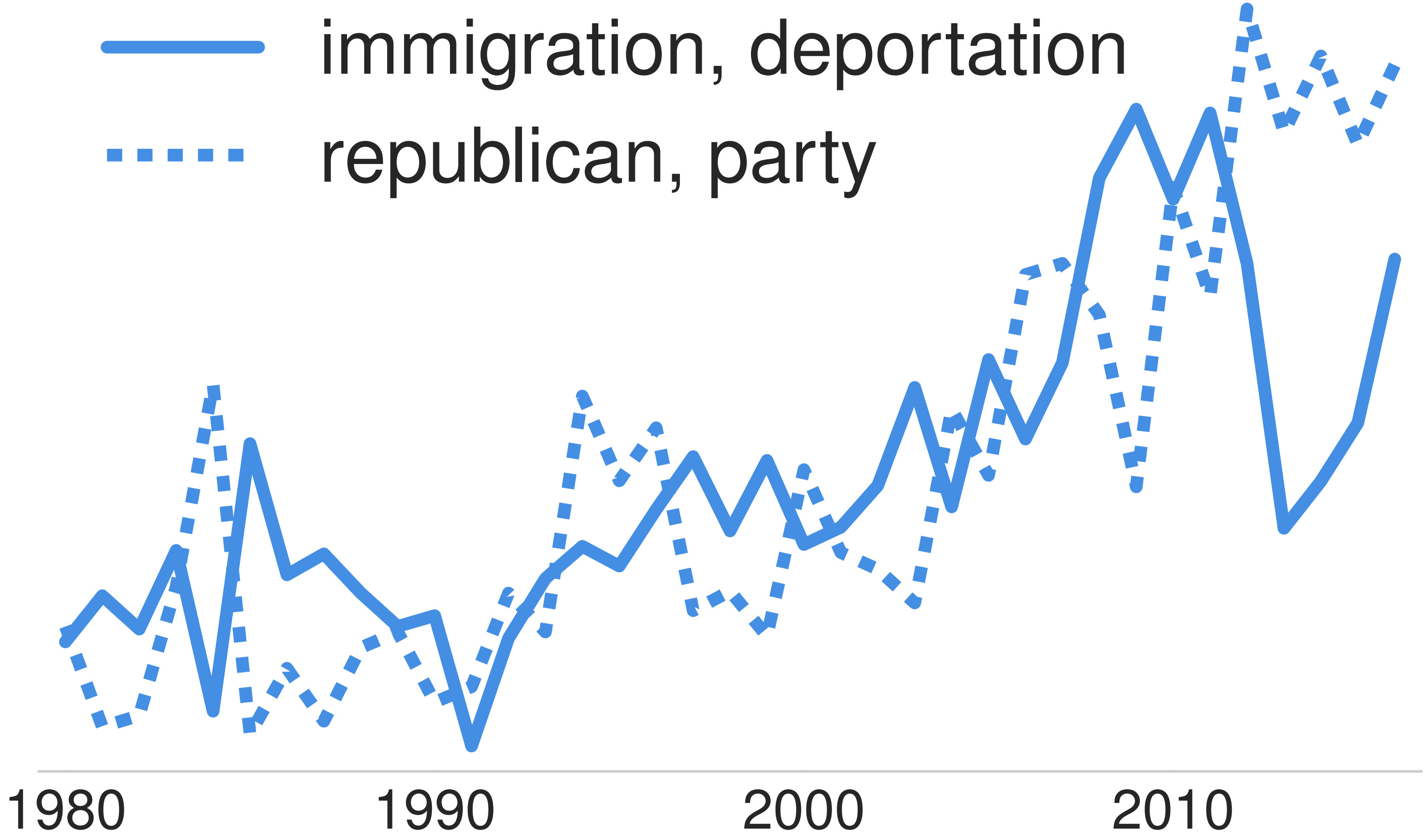
Friendship

Head-to-head

Arms-race

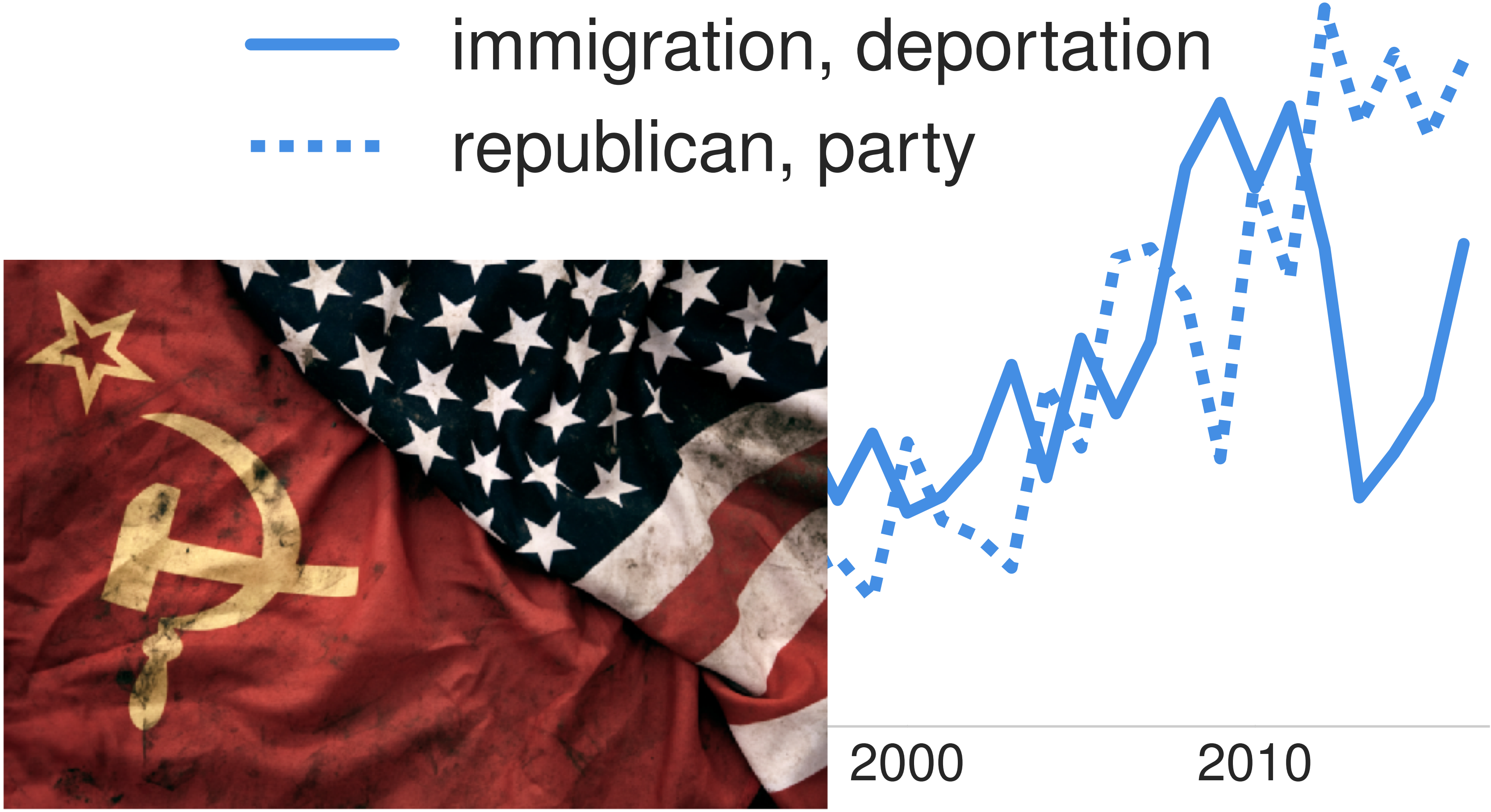
Rarely cooccur

ARMS-RACE
(CORRELATED, RARELY COOCCUR)



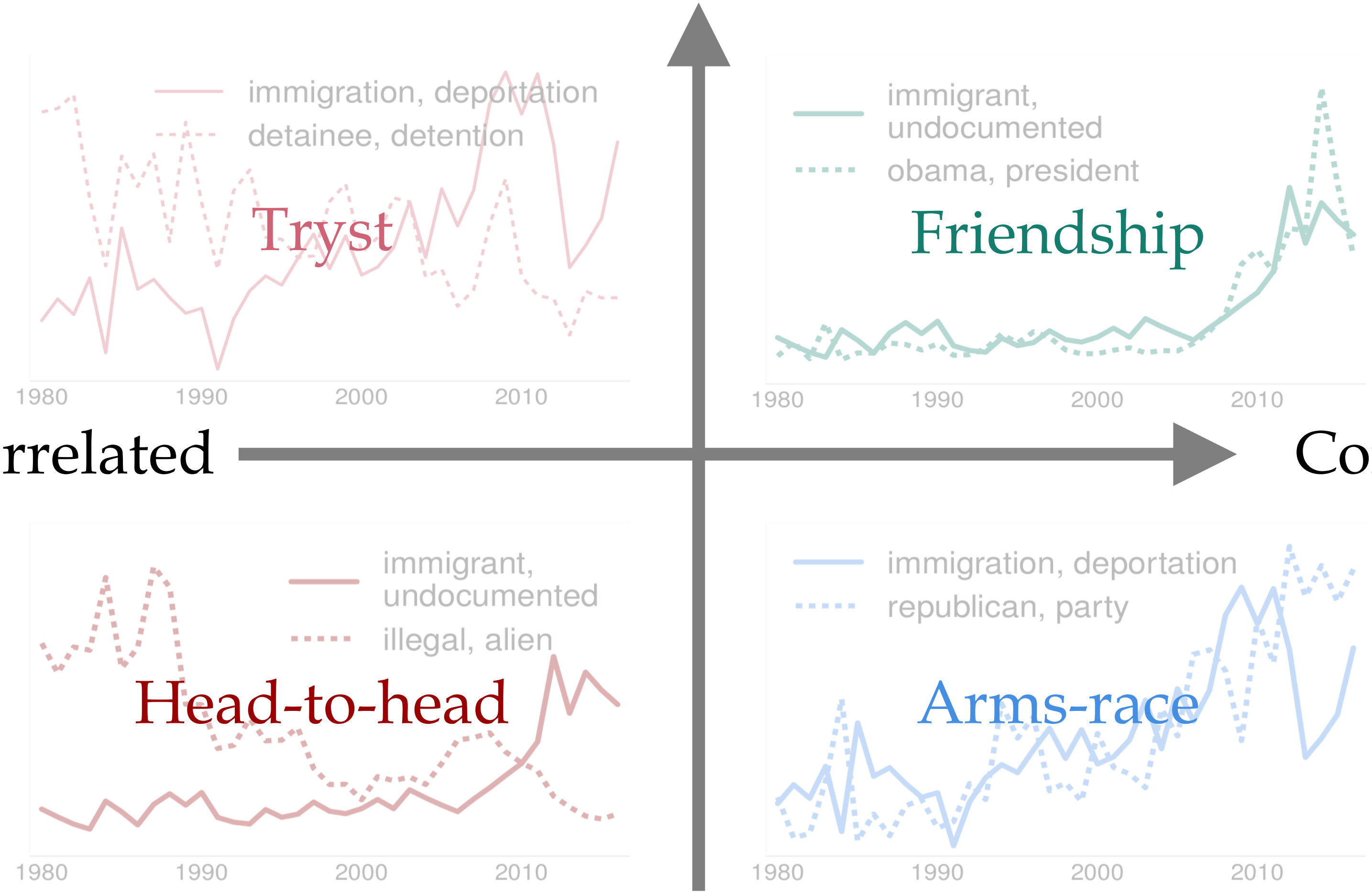
ARMS-RACE

(CORRELATED, RARELY COOCCUR)



RELATIONS BETWEEN IDEAS

Always cooccur



Tryst

Friendship

Anti-correlated

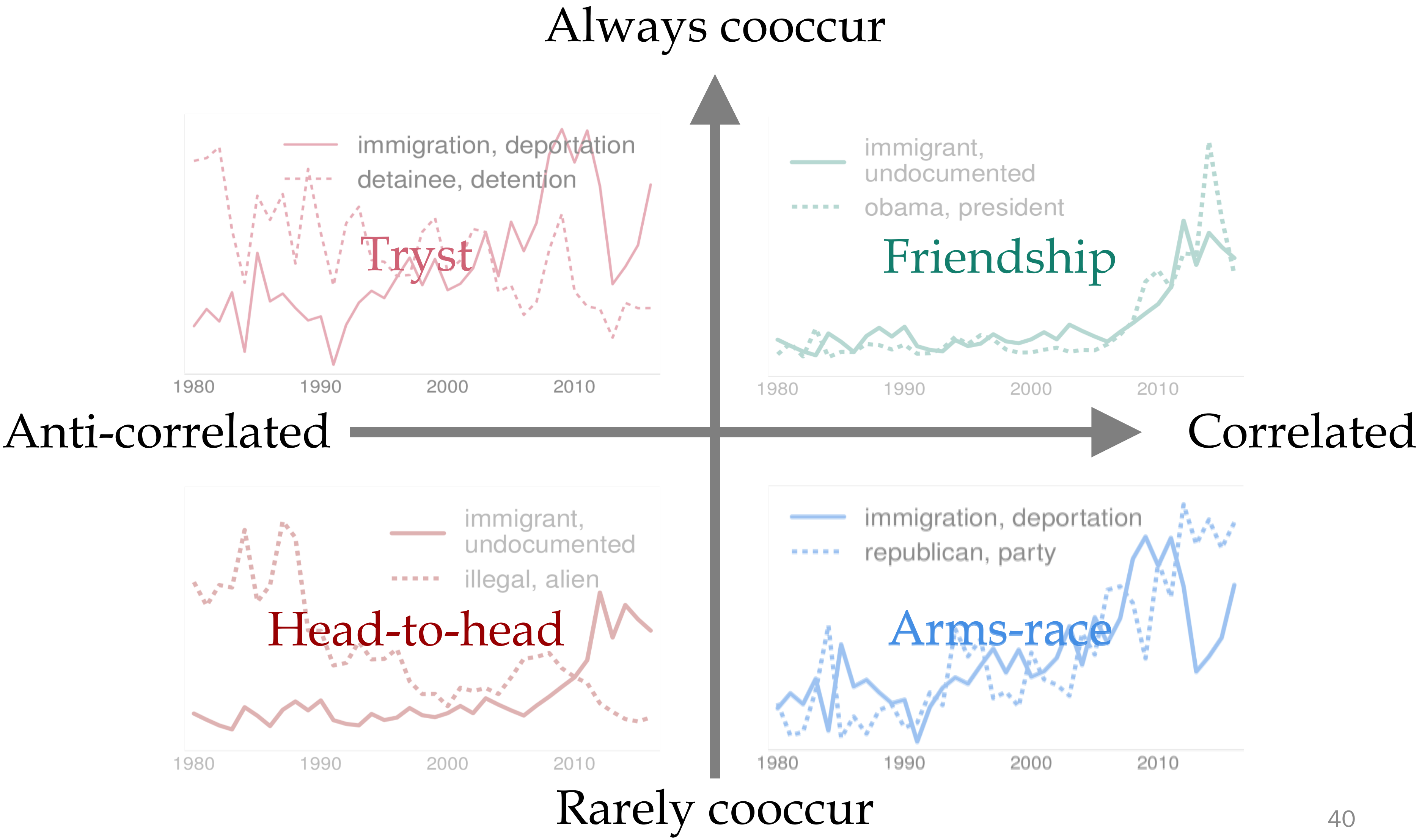
Correlated

Head-to-head

Arms-race

Rarely cooccur

RELATIONS BETWEEN IDEAS



Always cooccur

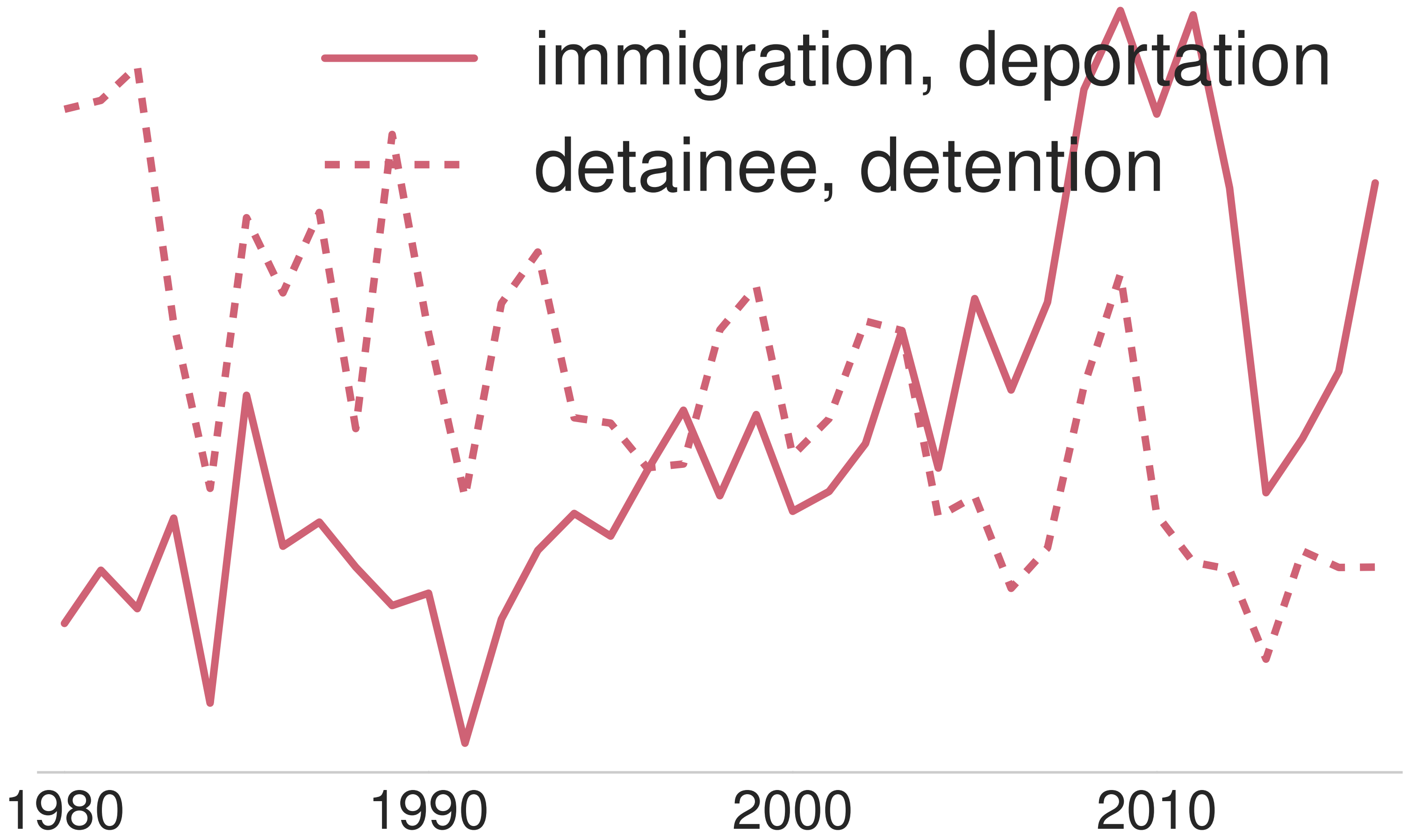
Anti-correlated

Correlated

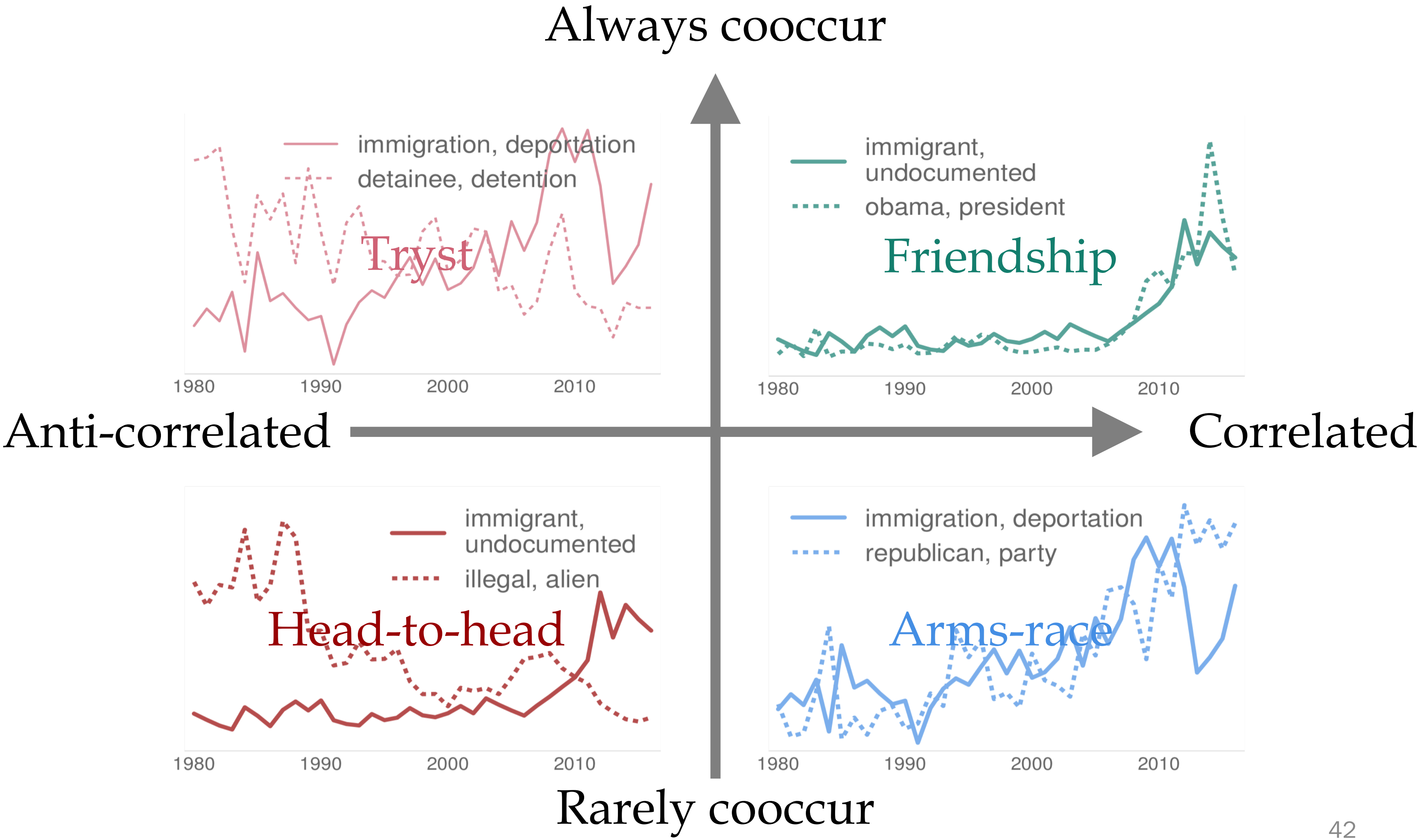
Rarely cooccur

TRYST

(ANTI-CORRELATED, LIKELY TO COOCCUR)



RELATIONS BETWEEN IDEAS



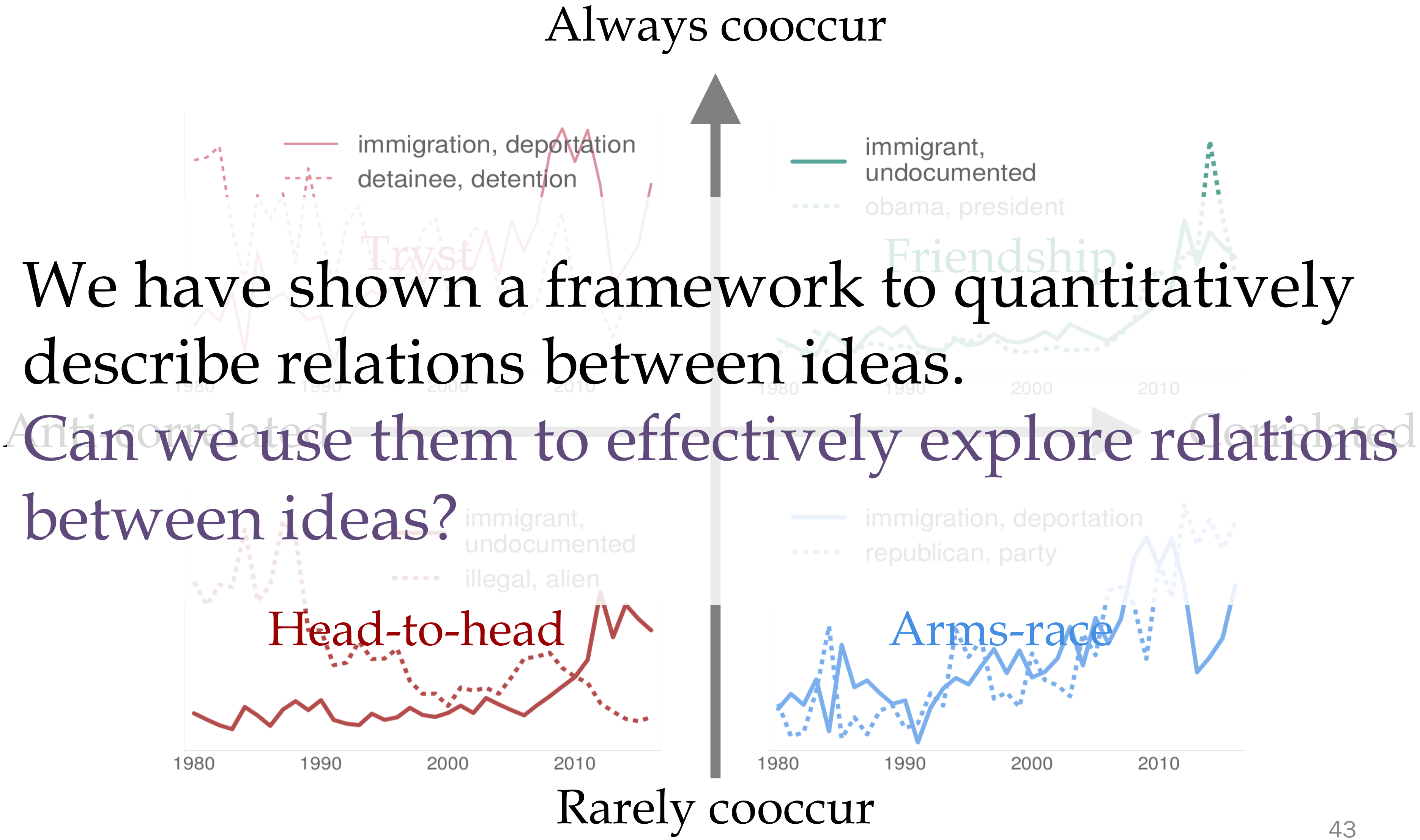
Always cooccur

Anti-correlated

Correlated

Rarely cooccur

RELATIONS BETWEEN IDEAS



We have shown a framework to quantitatively describe relations between ideas.

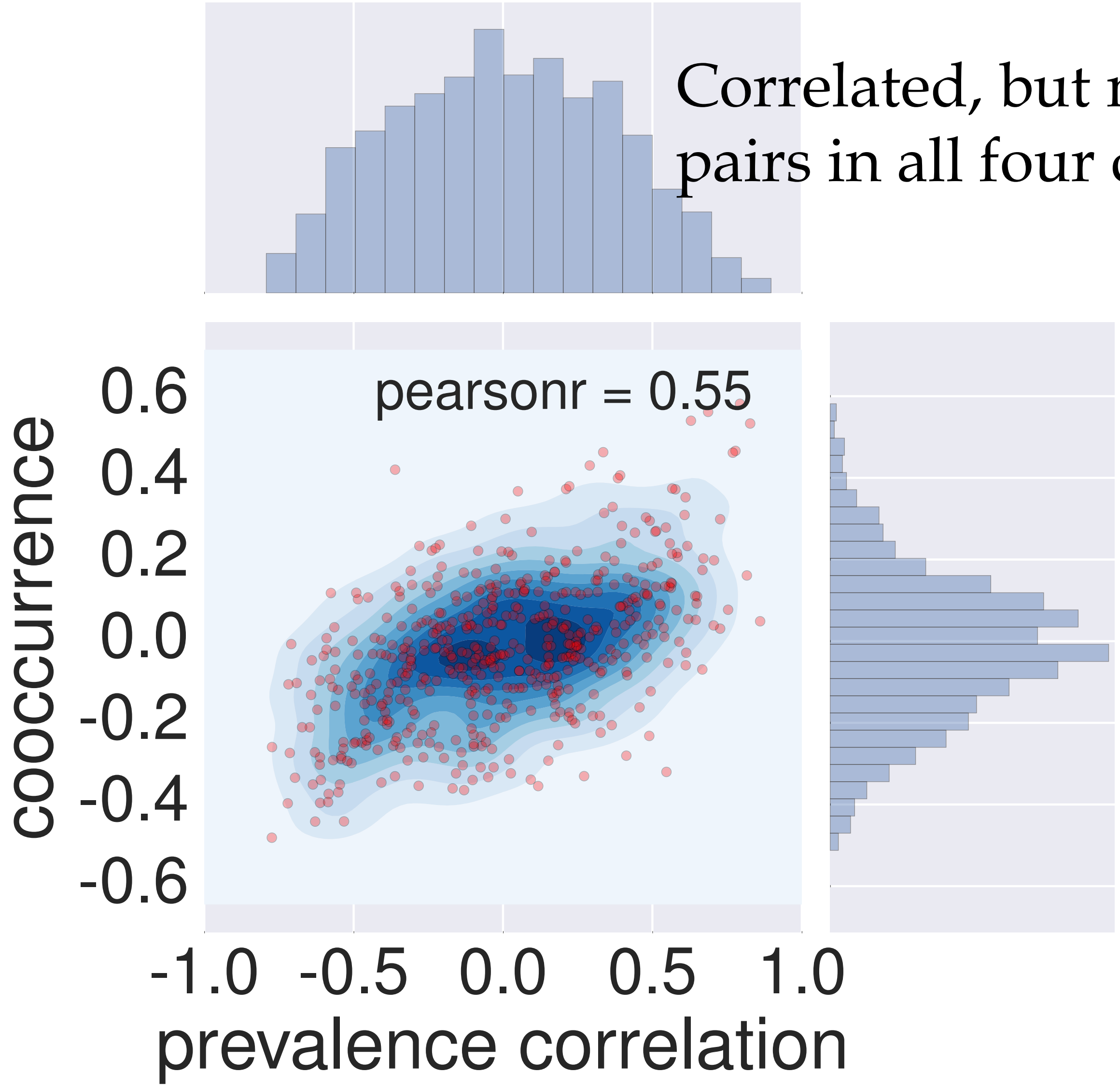
Can we use them to effectively explore relations between ideas?

A WIDE RANGE OF DATASETS

- Newspapers and research articles as datasets
  - Immigration
  - Terrorism
  - Same-sex marriage
  - Abortion
  - Tobacco

  - ACL
  - NIPS

JOINT DISTRIBUTIONS

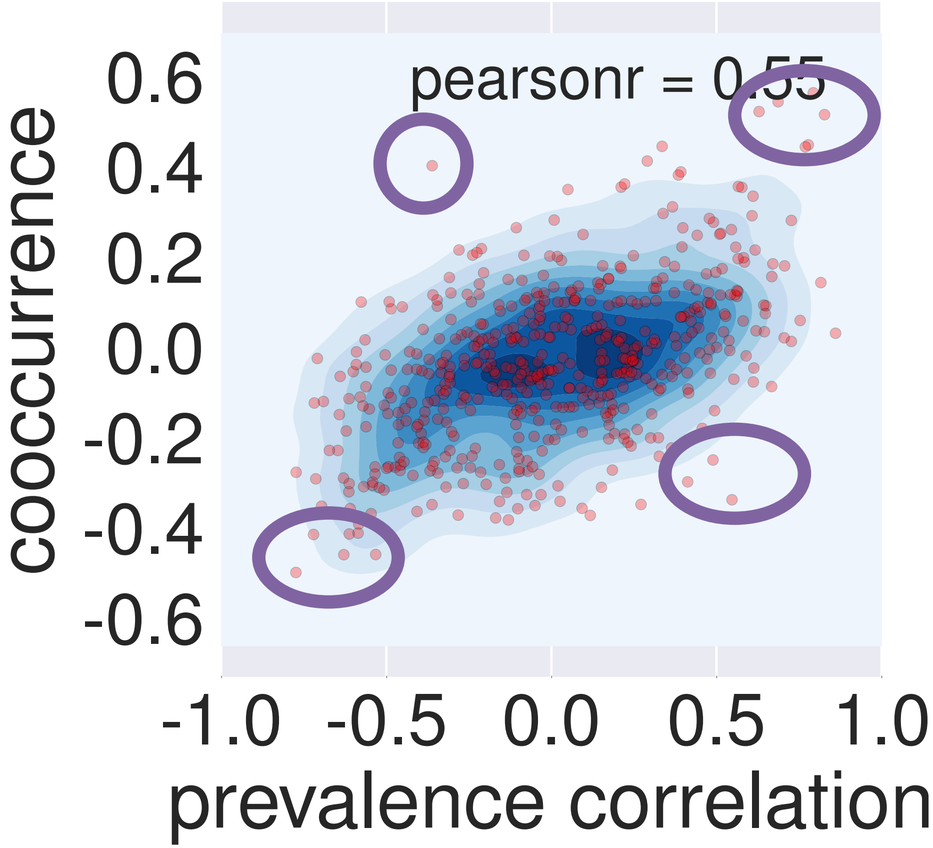


Correlated, but many pairs in all four quadrants!

THE STRENGTH OF RELATIONS

Strength = |PMI| × |correlation|

Extreme pairs are the interesting ones!



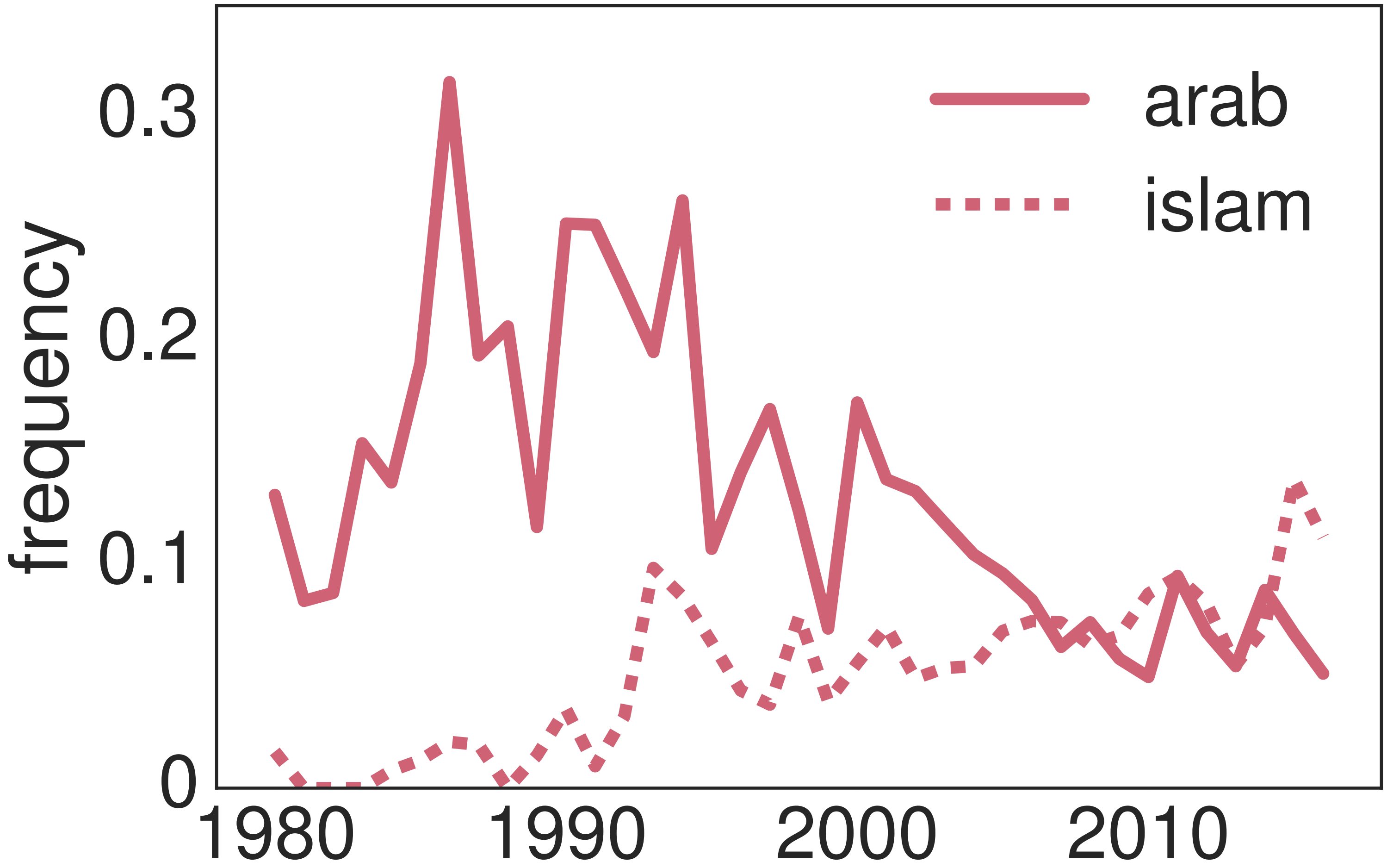
EFFECTIVELY EXPLORE RELATIONS BETWEEN IDEAS

- Terrorism
  - Keywords
  - Topics

EFFECTIVELY EXPLORE RELATIONS BETWEEN IDEAS

- Terrorism
  - Keywords
  - Topics

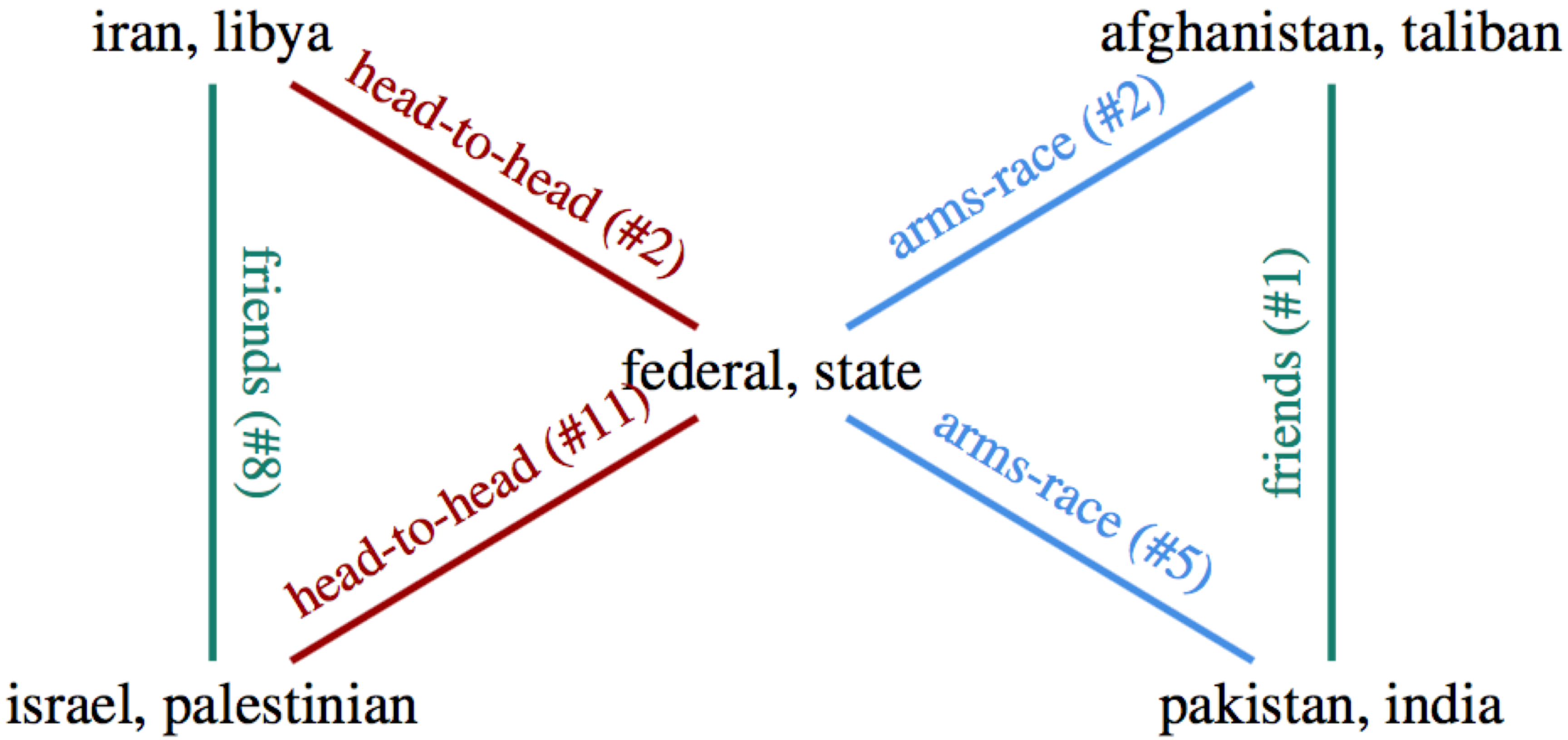
#2 IN TRYSTS



EFFECTIVELY EXPLORE RELATIONS BETWEEN IDEAS

- Terrorism
  - Keywords
  - Topics

TOP RELATIONS BETWEEN IDEAS



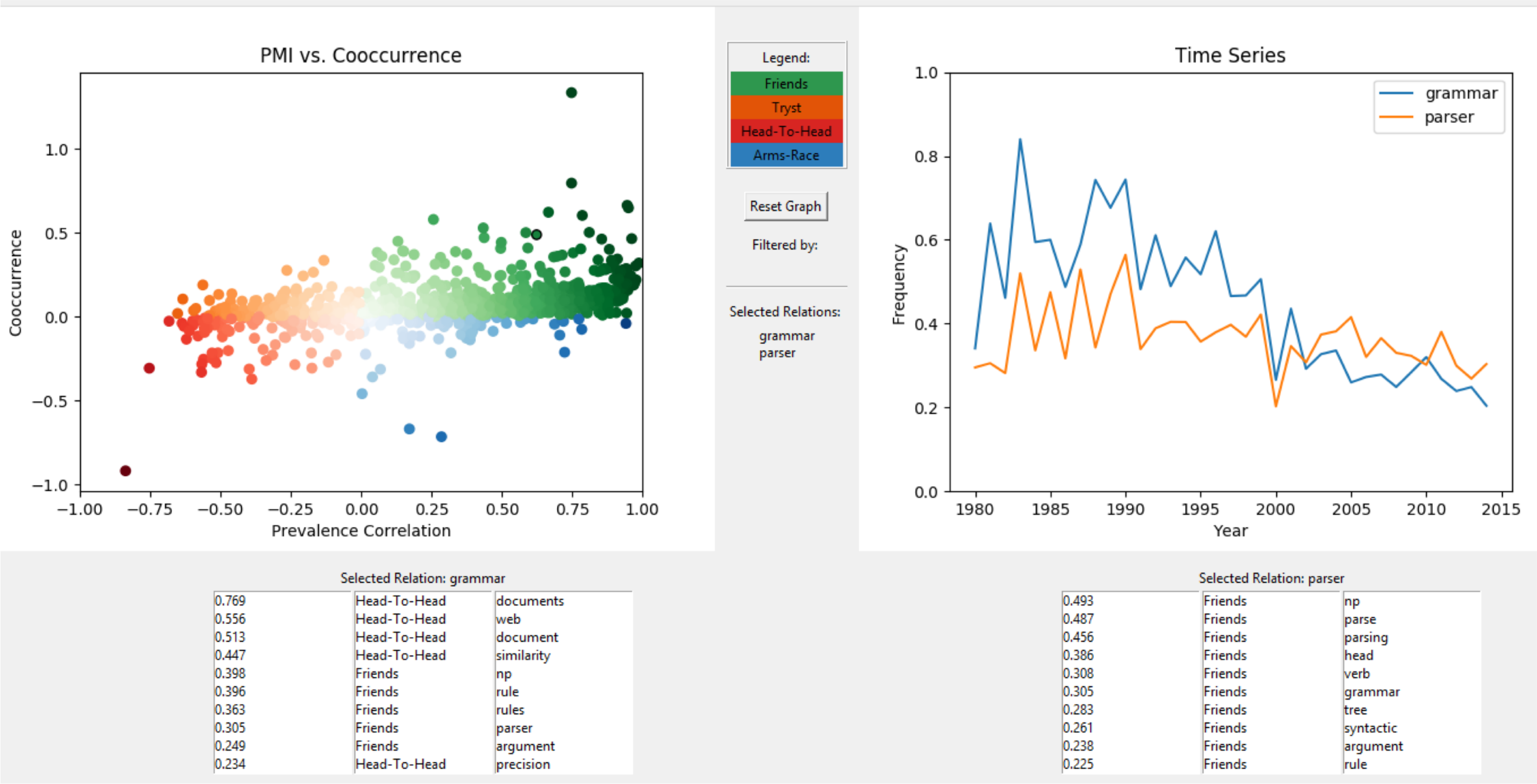
The relations between these topics are consistent with structural balance theory: the enemy of an enemy is a friend [Cartwright and Harary, 1956; Heider, 1946]

EFFECTIVE EXPLORATIONS

			Rank among all relations		
			PMI	**Correlation**	**Joint**
Keywords	arab	islam	106	1,494	2
Topics	federal, state	afghanistan, taliban	43	99	2
	federal, state	iran, lybia	36	56	2

The “interesting” pair is ranked much higher according to our framework.

VISUALIZATION TOOL



Selected Relations:
grammar
parser

Selected Relation: grammar

0.769	Head-To-Head	documents
0.556	Head-To-Head	web
0.513	Head-To-Head	document
0.447	Head-To-Head	similarity
0.398	Friends	np
0.396	Friends	rule
0.363	Friends	rules
0.305	Friends	parser
0.249	Friends	argument
0.234	Head-To-Head	precision

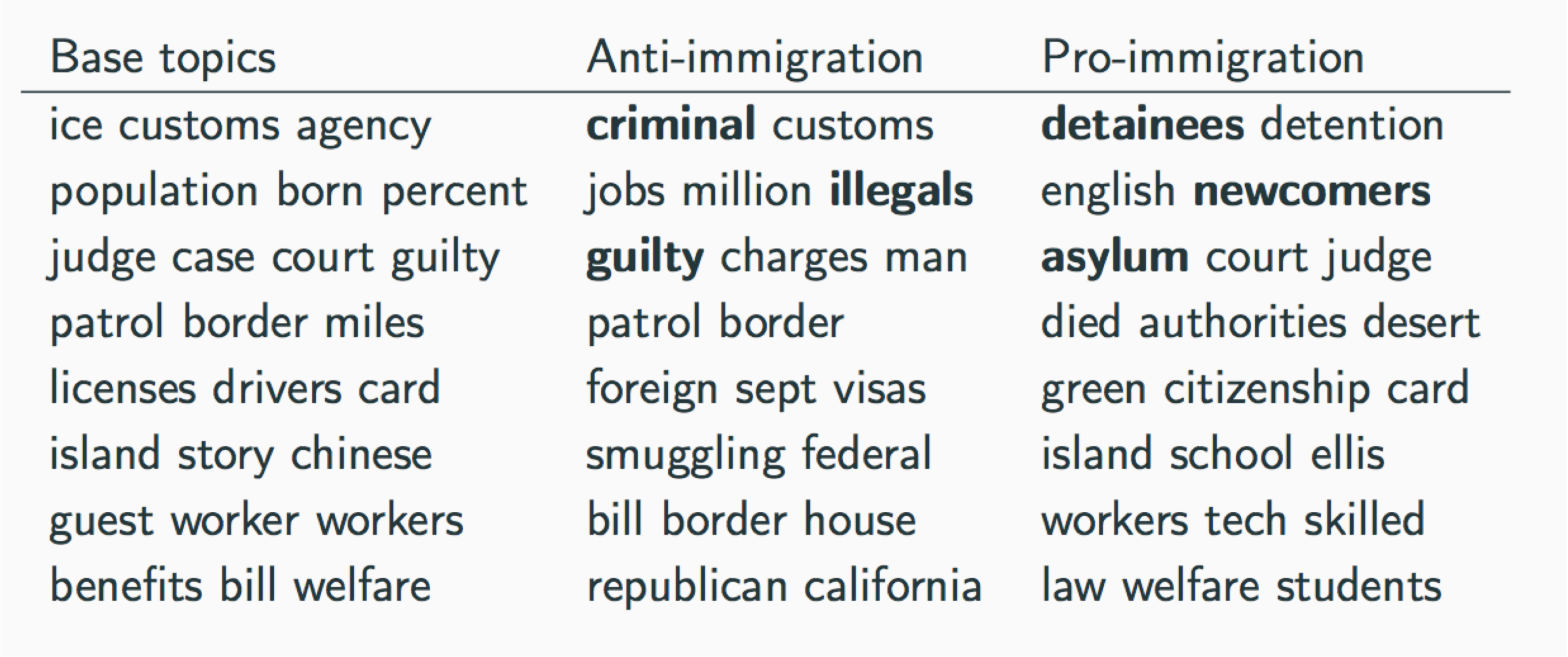
Selected Relation: parser

0.493	Friends	np
0.487	Friends	parse
0.456	Friends	parsing
0.386	Friends	head
0.308	Friends	verb
0.305	Friends	grammar
0.283	Friends	tree
0.261	Friends	syntactic
0.238	Friends	argument
0.225	Friends	rule

https://github.com/Noahs-ARK/idea_relations

https://github.com/nwrush/Visualization

NEURAL MODELS FOR DOCUMENTS WITH METADATA



Base topics	Anti-immigration	Pro-immigration
ice customs agency	**criminal** customs	**detainees** detention
population born percent	jobs million **illegals**	english **newcomers**
judge case court guilty	**guilty** charges man	**asylum** court judge
patrol border miles	patrol border	died authorities desert
licenses drivers card	foreign sept visas	green citizenship card
island story chinese	smuggling federal	island school ellis
guest worker workers	bill border house	workers tech skilled
benefits bill welfare	republican california	law welfare students

https://github.com/dallascard/scholar

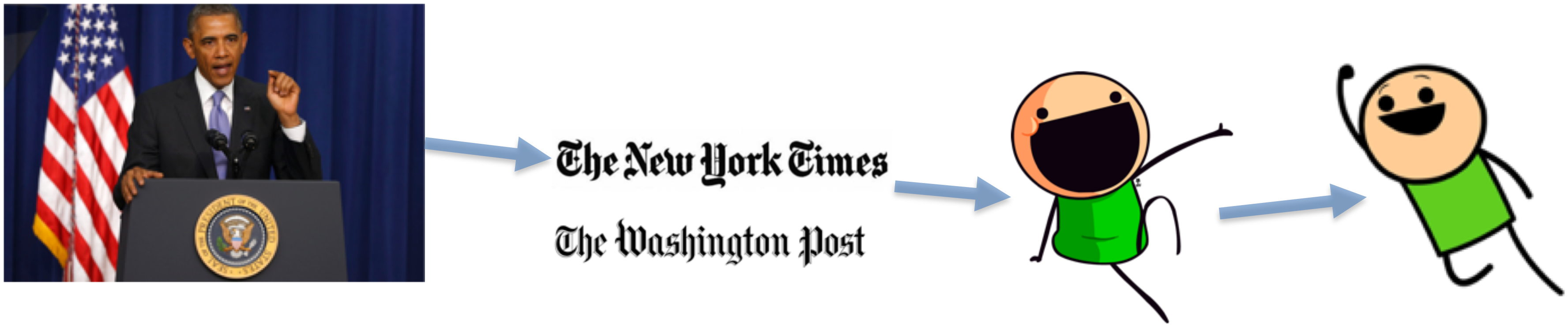
THE ECOSYSTEM OF IDEAS

- Competition and collaboration
  - – Natural selection [Dawkins 1976]
  - – Marketplace of ideas [Milton 1644; Mill 1859]

    Tan, Card, Smith, ACL'17
- Evolution of ideas
  - – Telephone game

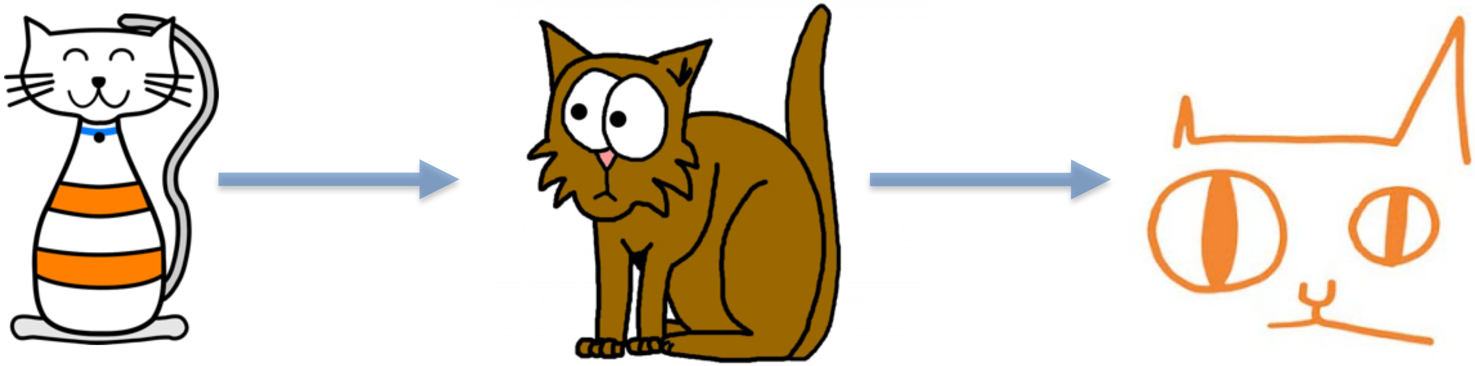
    Tan, Friggeri, Adamic, ICWSM'16

IDEAS REACH US VIA DIFFERENT WAYS

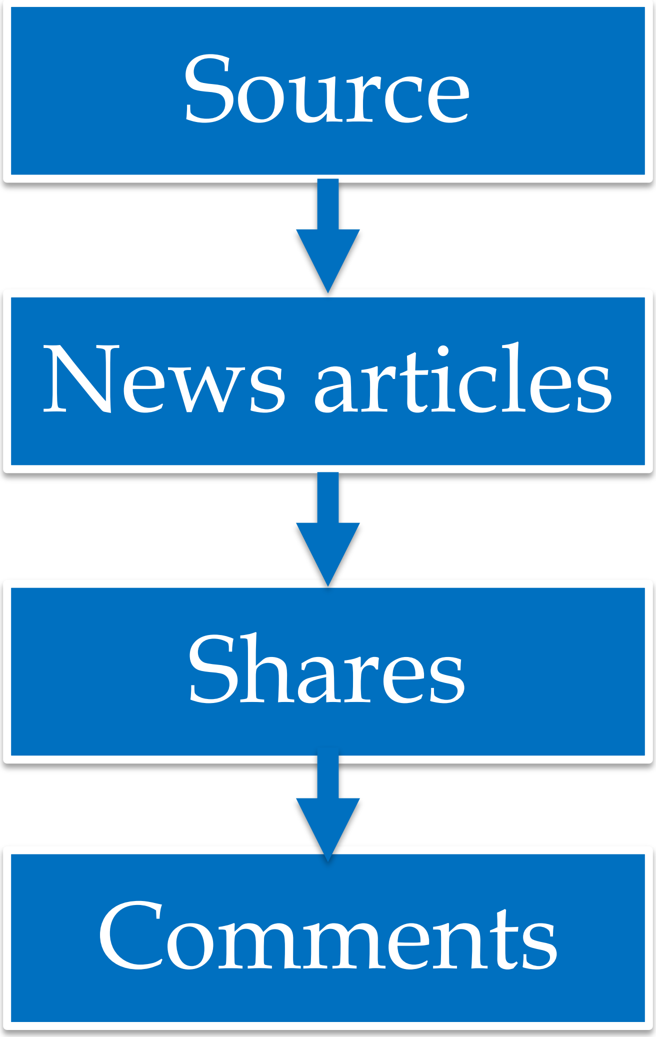


The New York Times

The Washington Post



EVOLUTION OF IDEAS



Source

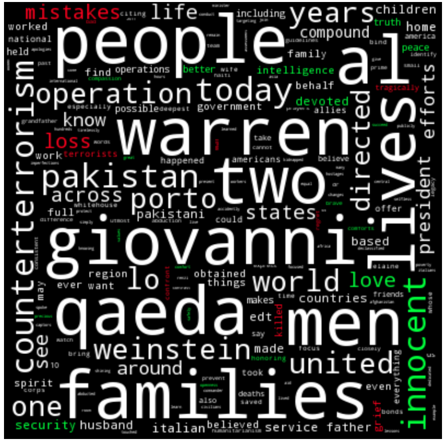
News articles

Shares

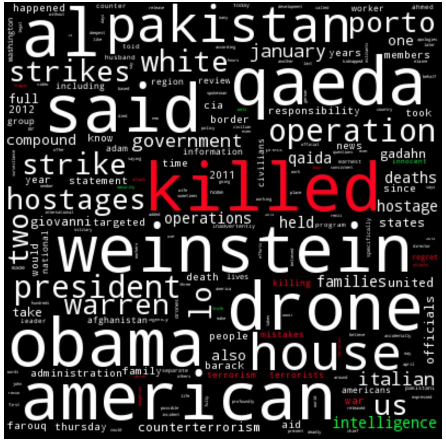
Comments

OBAMA'S SPEECH ON DEATHS OF WARREN AND GIOVANNI

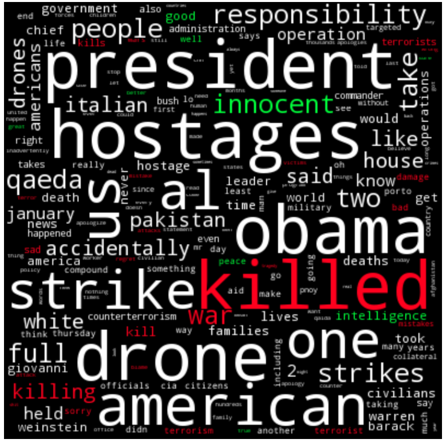
information source



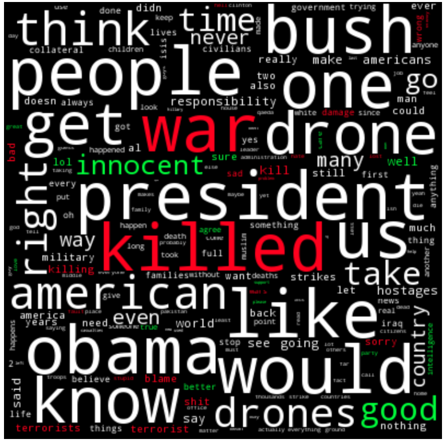
news articles



shares

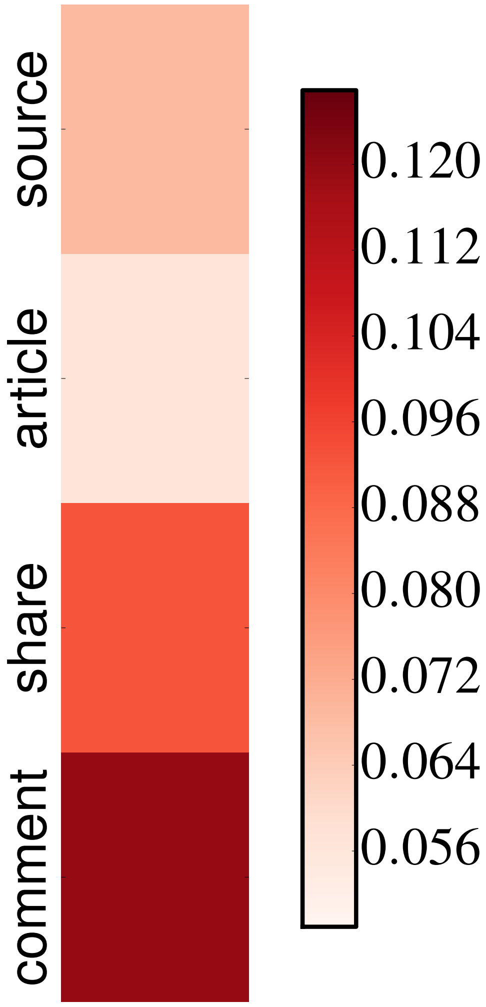


comments



SENTIMENT EVOLUTION

- Subjectivity defined as the fraction of emotional words
- News articles are less subjective than sources, but shares and comments get more and more subjective

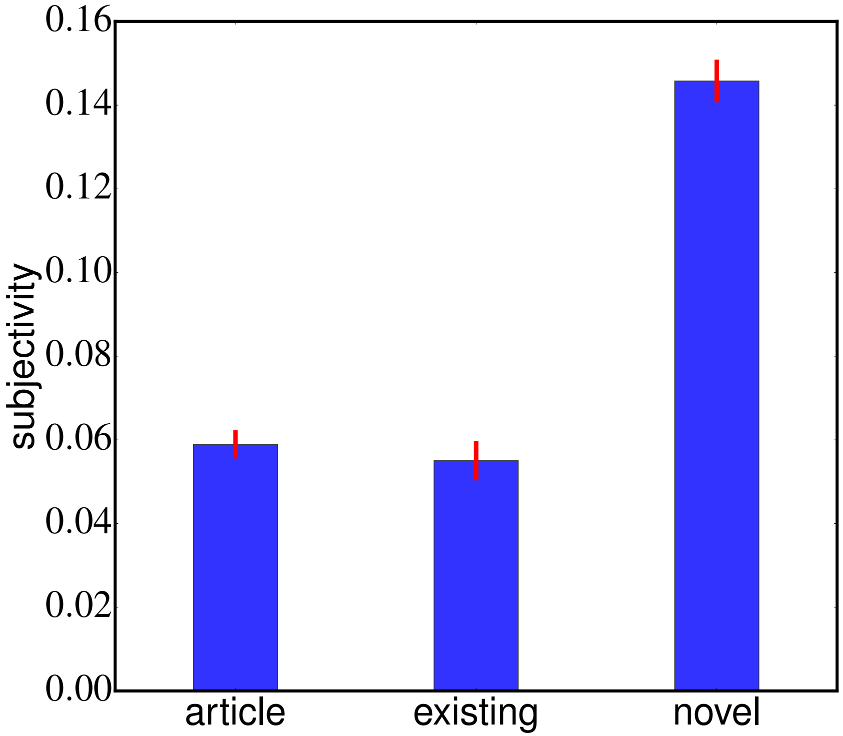


Red: more subjective

WHY DOES SUBJECTIVITY INCREASE?

News articles → Shares

- words already in news articles *(magnifier)*
- novel words added by individuals *(creator)*



CHARACTERIZING THE ECOSYSTEM OF IDEAS

better understand
existing ideas in the data

better generate
ideas for the future

The beginning of an exciting journey!

Chenhao Tan
chenhao@chenhaot.com
https://chenhaot.com
@ChenhaoTan

http://www.xiangke.com/html/news/100284/61936252.html