

Natural Language Processing (NLP)

for

Computational Social Science

Cristian Danescu-Niculescu-Mizil

and

Lillian Lee

<http://www.cs.cornell.edu/courses/cs6742/2015fa>

Datasets:

<http://www.cs.cornell.edu/home/llee/data/index.html>

[http://www.cs.cornell.edu/~cristian/Data Media Talks News.html](http://www.cs.cornell.edu/~cristian/Data_Media_Talks_News.html)

Natural Language Processing

Computational
Social Science

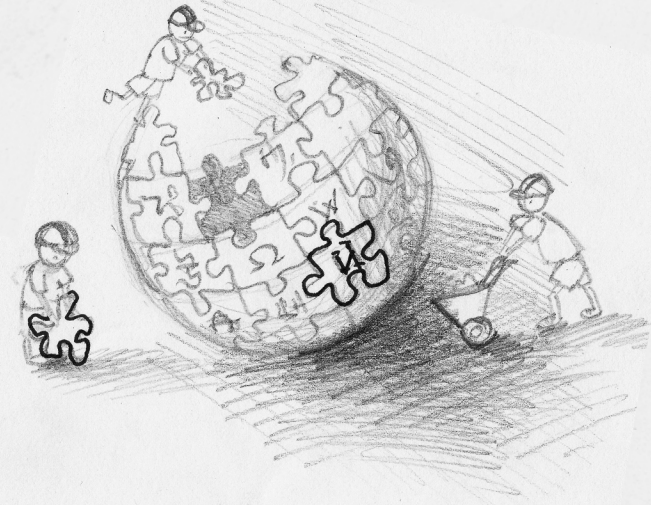
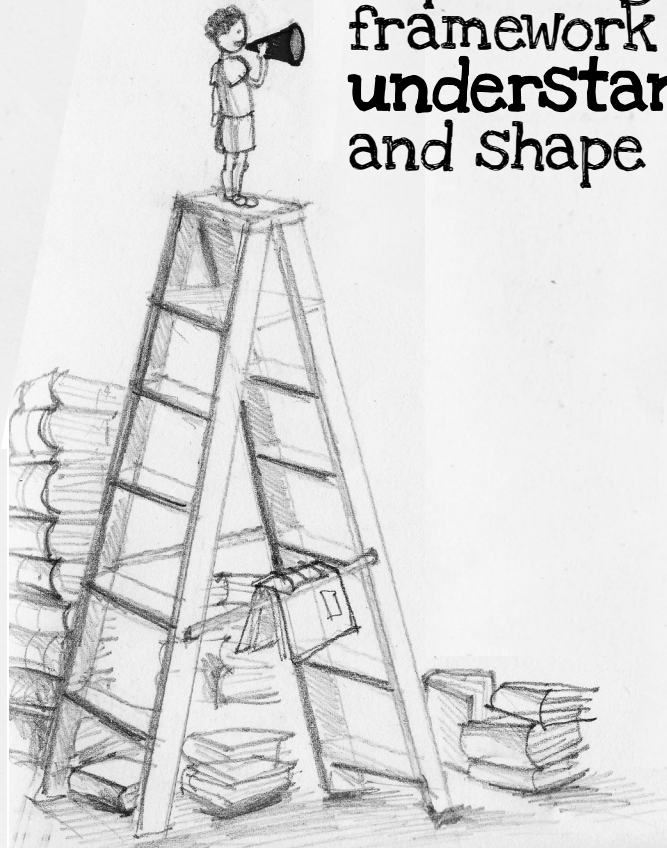
← Why do people do what they do
(when other people are involved)?

← NLP: a great way to find out!

Why NLP for CSS?

Much of online human activity leaves digital traces that are recorded in **natural-language format.**

Exploiting these resources under a computational framework can bring a **phase transition in our understanding of human social behavior** and shape the future of social-media systems.



TODAY:

... Research questions
persuasion, linguistic change, framing

... Techniques
language models, Bayesian feature analysis



... Research practices
controls, feasibility, data inspection

The social effects of linguistic subtleties

The social effects of linguistic subtleties

"Motivating voter turnout" (Bryan et al, 2011)

The social effects of linguistic subtleties

"Motivating voter turnout" (Bryan et al, 2011)

"How important it is to you to be a voter?"

"How important it is to you to vote?"

The social effects of linguistic subtleties

"Motivating voter turnout" (Bryan et al, 2011)

"How important it is to you to **be a voter**?" (identity)

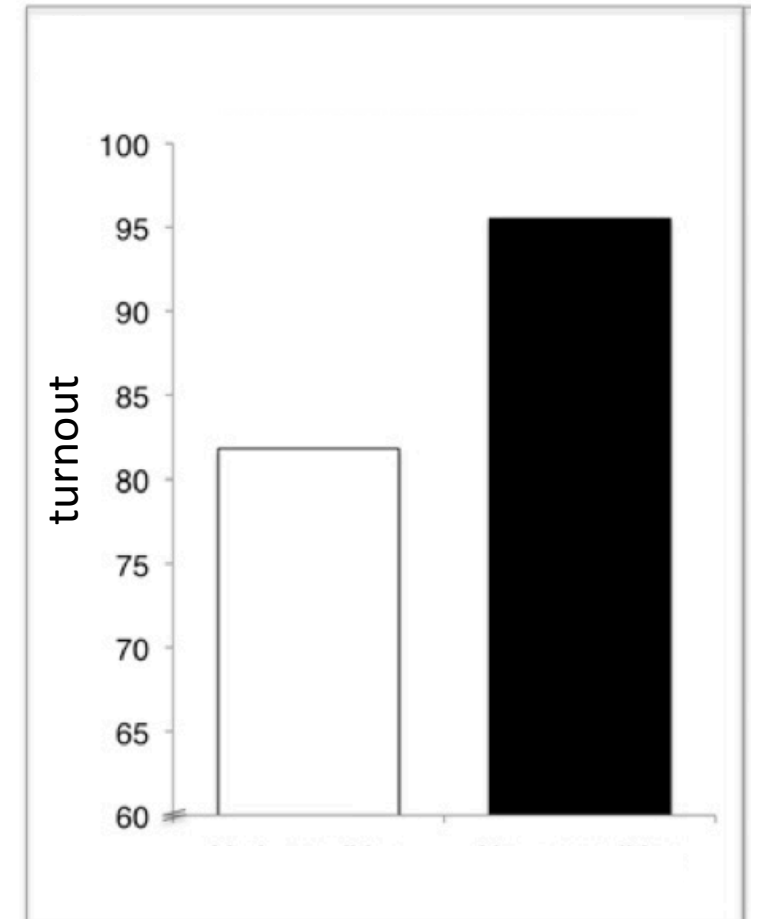
"How important it is to you to **vote**?" (action)

The social effects of linguistic subtleties

"Motivating voter turnout" (Bryan et al, 2011)

"How important it is to you to be a voter?" (identity)

"How important it is to you to vote?" (action)

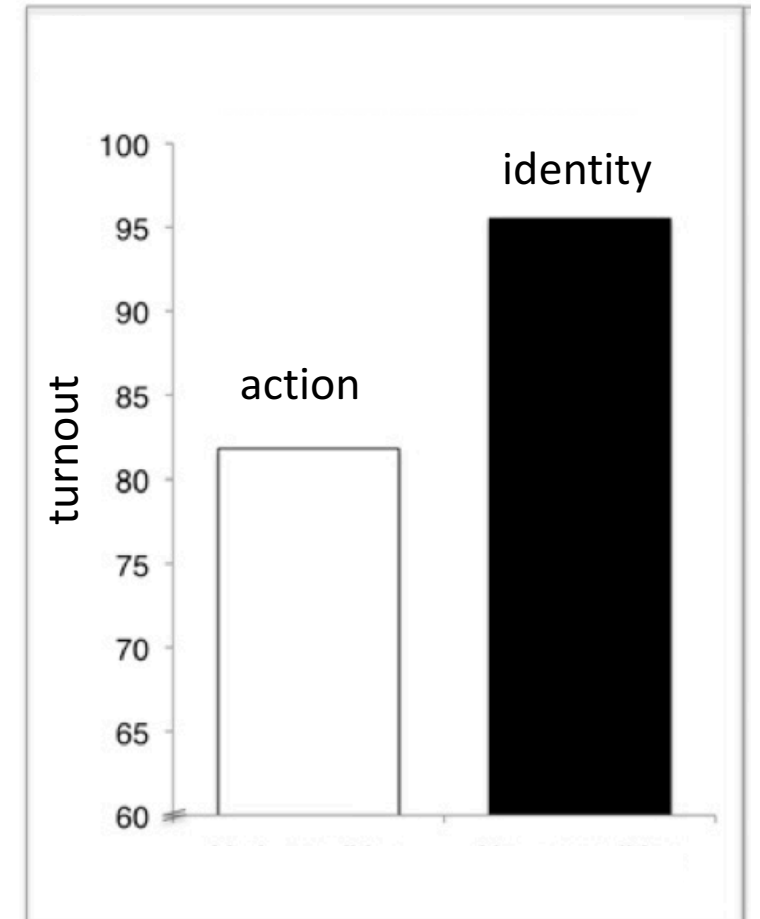


The social effects of linguistic subtleties

"Motivating voter turnout" (Bryan et al, 2011)

"How important it is to you to be a voter?" (identity)

"How important it is to you to vote?" (action)



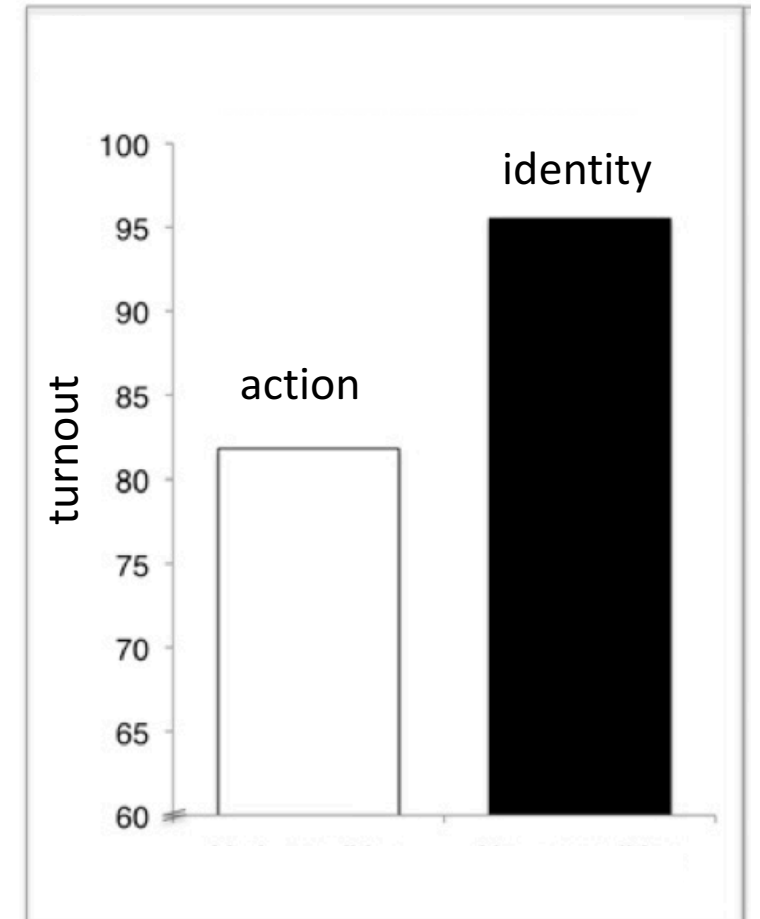
The social effects of linguistic subtleties

"Motivating voter turnout" (Bryan et al, 2011)

"How important it **is** to you to be a voter?" (identity)

"How important it **is** to you to vote?" (action)

How things are said
(vs what is said)



The social effects of linguistic subtleties

"The role of placebic information" (Langer et al, 1978)

"I have 5 pages. May I use the xerox machine?"

The social effects of linguistic subtleties

"The role of placebo information" (Langer et al, 1978)

"I have 5 pages. May I use the xerox machine?"

"I have 5 pages. May I use the xerox machine,
because I need to make copies?"

The social effects of linguistic subtleties

"The role of placebic information" (Langer et al, 1978)

"I have 5 pages. May I use the xerox machine?"

"I have 5 pages. May I use the xerox machine,
because I need to make copies?"

The social effects of linguistic subtleties

"The role of placebic information" (Langer et al, 1978)

"I have 5 pages. May I use the xerox machine?" 60% agreed

"I have 5 pages. May I use the xerox machine,
because I need to make copies?" 93% agreed

The social effects of linguistic subtleties

"The role of placebo information" (Langer et al, 1978)

"I have 5 pages. May I use the xerox machine?" 60% agreed

"I have 5 pages. May I use the xerox machine,
because I need to make copies?" 93% agreed

"I have 5 pages. May I use the xerox machine,
because I am in a rush?" 94% agreed

The social effects of linguistic subtleties

Today's data → opportunity to discover and better understand social effects

The social effects of linguistic subtleties

Today's data → opportunity to discover and better understand social effects

A classification problem?

Example: (How) do male and female describe things differently?

The social effects of linguistic subtleties

Today's data → opportunity to discover and better understand social effects

A classification problem?

Example: (How) do male and female describe things differently?

Gender classification

The social effects of linguistic subtleties

Today's data → opportunity to discover and better understand social effects

A classification problem?

Example: (How) do male and female describe things differently?

Gender classification

Issue: Gender-Topic confound (Argamon et al 2003, Sarawgi et al 2011)

"Finance" trends male,

but what about females who talk about finance

The social effects of linguistic subtleties

Today's data → opportunity to discover and better understand social effects

Challenges:

- ★ maintaining the controlled, hypothesis-driven nature of traditional studies
 - › sense (and luck) to find **the right data**

The social effects of linguistic subtleties

"The role of placebic information" (Langer et al, 1978)

"I have 5 pages. May I use the xerox machine?" 60% agreed

"I have 5 pages. May I use the xerox machine,
because I need to make copies?" 93% agreed

"I have 5 pages. May I use the xerox machine,
because I am in a rush?" 94% agreed

The social effects of linguistic subtleties

"The role of placebo information" (Langer et al, 1978)

"I have 5 pages. May I use the xerox machine?" 60% agreed

"I have 20 pages. May I use the xerox machine,
because I need to make copies?" ~~93% agreed~~
24% agreed

"I have 5 pages. May I use the xerox machine,
because I am in a rush?" 94% agreed

The social effects of linguistic subtleties

Today's data → opportunity to discover and better understand social effects

Challenges:

- ★ maintaining the controlled, hypothesis-driven nature of traditional studies
 - › sense (and luck) to find **the right data**

The social effects of linguistic subtleties

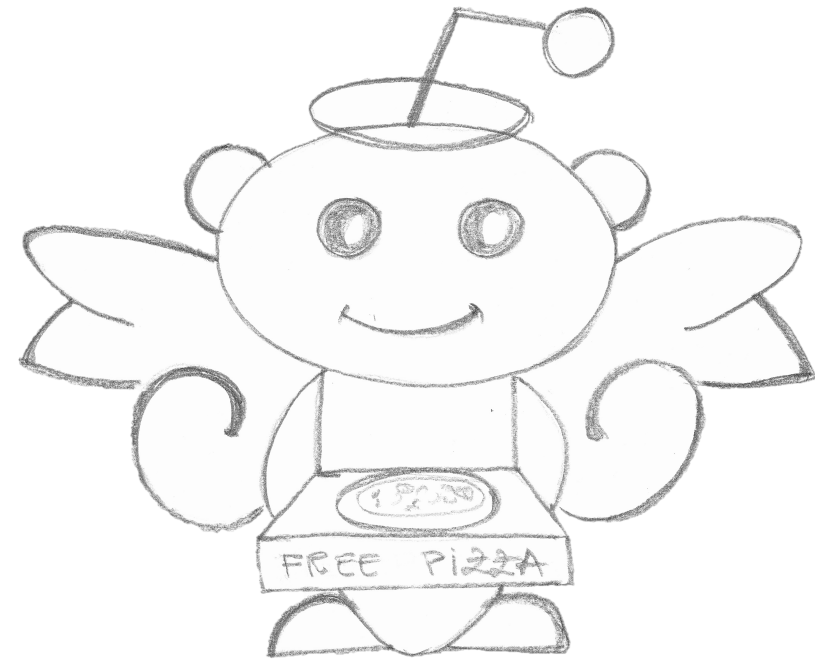
Today's data → opportunity to discover and better understand social effects



Challenges:

- ★ maintaining the controlled, hypothesis-driven nature of traditional studies
 - › sense (and luck) to find **the right data**

"How to ask for a favor" (Althoff et al, 2013)



20,000 requests for ... pizza





65  0  [Request] I have gotten pizza before from this subreddit, but it's Easter, and I'm stuck at school because finals for me start tomorrow, and I'm broke.


 submitted 3 days ago by [silentsly](#)

[comment](#) [share](#)



66  4  [Request] I've been working on my first computer for 6 hours, only to find my GPU was DOA. Can someone hit me up with some pizza please?


 submitted 3 days ago by [bigbootypanda](#)
[4 comments](#) [share](#)

68  1  [Request] Spooky podcasts go great with pizza! (California)



 submitted 3 days ago by [posolutelyabsotively](#)


[comment](#) [share](#)

69  23  [REQUEST] I know this is a long shot. But I've come to the end of college and have drained my funds for it 100% I am currently waiting on an email from said college that will basically determine my future. I have never been so stressed or scared. Pizza would be a comfort. Promise to pay it forward.



 submitted 3 days ago by [mrshansgruber](#)

[3 comments](#) [share](#)



65  0  [Request] I have gotten pizza before from this subreddit, but it's Easter, and I'm stuck at school because finals for me start tomorrow, and I'm broke.


 submitted 3 days ago by [silentsly](#)

[comment](#) [share](#)



66  4  [Request] I've been working on my first computer for 6 hours, only to find my GPU was DOA. Can someone [hit me up](#) with some pizza [please?](#)


 submitted 3 days ago by [bigbootypanda](#)
[4 comments](#) [share](#)

68  1  [Request] Spooky podcasts go great with pizza! (California)



 submitted 3 days ago by [posolutelyabsotively](#)



[comment](#) [share](#)

69  23  [REQUEST] I know this is a long shot. But I've come to the end of college and have drained my funds for it 100% I am currently waiting on an email from said college that will basically determine my future. I have never been so stressed or scared. Pizza [would](#) be a comfort. [Promise to pay it forward.](#)



 submitted 3 days ago by [mrshansgruber](#)

[3 comments](#) [share](#)



65  0  [Request] I have gotten pizza before from this subreddit, but it's Easter, and I'm stuck at school because finals for me start tomorrow, and I'm broke.



 submitted 3 days ago by [silentsly](#)  got pizza'd

[comment](#) [share](#)



66  4  [Request] I've been working on my first computer for 6 hours, only to find my GPU was DOA. Can someone hit me up with some pizza please?

 submitted 3 days ago by [bigbootypanda](#)
4 [comments](#) [share](#)

68  1  [Request] Spooky podcasts go great with pizza! (California)

 submitted 3 days ago by [posolutelyabsotively](#)  got pizza'd

[comment](#) [share](#)

69  23  [REQUEST] I know this is a long shot. But I've come to the end of college and have drained my funds for it 100% I am currently waiting on an email from said college that will basically determine my future. I have never been so stressed or scared. Pizza would be a comfort. Promise to pay it forward.

submitted 3 days ago by [mrshansgruber](#)
3 [comments](#) [share](#)

The social effects of linguistic subtleties

Today's data → opportunity to discover and better understand such effects

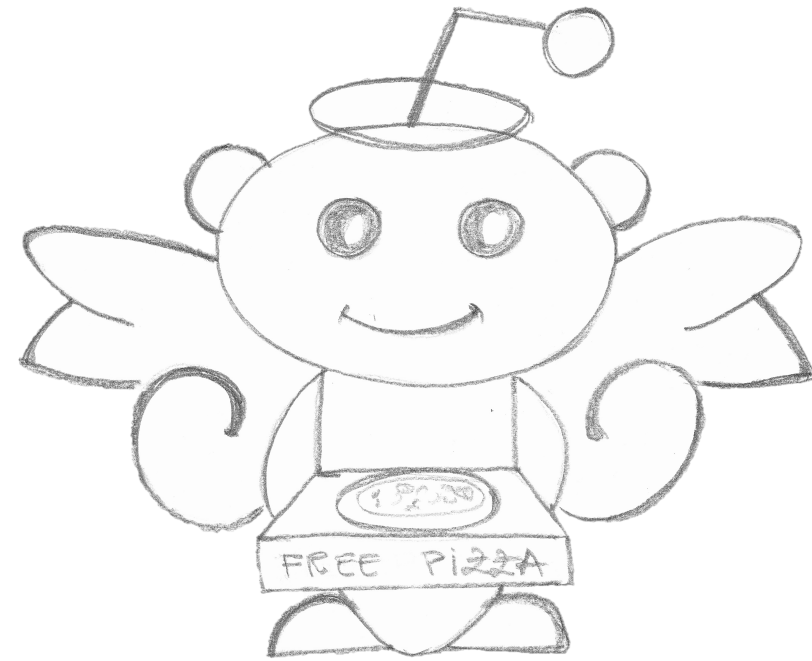
Challenges:

- ★ maintaining the controlled, hypothesis-driven nature of traditional studies
 - › sense (and luck) to find **the right data**

"How to ask for a favor" (Althoff et al, 2013)

20,000 requests for ... pizza

Language choices can increase success rate from 9% to 57%



The social effects of linguistic subtleties

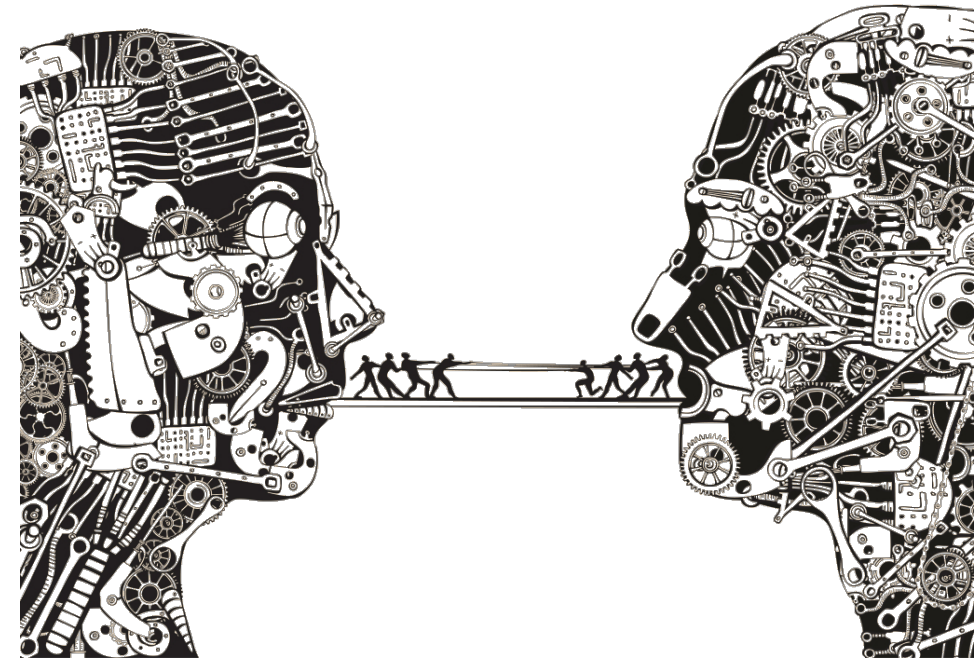
Today's data → opportunity to discover and better understand such effects

Challenges:

- ★ maintaining the controlled, hypothesis-driven nature of traditional studies
 - › sense (and luck) to find **the right data**

"Winning arguments" (Tan et al, 2016)

20,000 persuasion "contests"



Example persuasion attempts in ChangeMyView

Original post, by "blue" user

CMV: the Tontine should be legalized and made a common retirement strategy.
[Reference URL omitted] Basically, today we have a huge problem with retirement [...+73 words]
A tontine for retirement looks like [...+56 words] The yearly sum is divided evenly for all the surviving participants [...+25 words]. The key advantages as I see it are:
*We don't need actuaries [...+29 words...]
*Management fees can be quite low [...+22 words]
* [Another reason]
* [Another reason]
But CMV. Are there major risks I am not foreseeing? [+2 more questions]

A tontine is a pretty crappy retirement vehicle for most people. It pays out the least when you need the most, and the most when you need the least.
People's income needs in retirement generally fall as they age. [...+35 words]
[URL]



Very interesting. I'll give a Δ because I didn't have any idea that was true and changes my idea of how the tontine should work. That said, I don't think it's unsolvable [...+44 words]

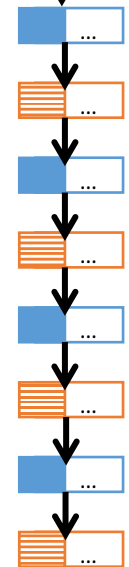
DeltaBot] Confirmed: 1 delta awarded to [red]

The Social Security system is basically one giant Tontine [...+13 words] So it's already legal.

I'd imagine the tontine as a secondary system to social security though, one that is optional for people to do, not mandatory like social security. [...+11 words]

There are some key differences though. First, Social Security is defined by the government [...+36 words]

And a tontine would be defined by your bank [...+79 words]



Then your back to needing actuaries, to predict [...+11 words]

Depends how exact you need to be [...+33 words]

The social effects of linguistic subtleties

Today's data → opportunity to discover and better understand such effects

Challenges:

- ★ maintaining the controlled, hypothesis-driven nature of traditional studies
 - › sense (and luck) to find the right data
 - › **taming wild data**: art to setting up the right comparisons

The social effects of linguistic subtleties

Today's data → opportunity to discover and better understand such effects

Challenges:

- ★ maintaining the controlled, hypothesis-driven nature of traditional studies
 - › sense (and luck) to find the right data
 - › taming wild data: art to setting up the right comparisons
- ★ need to develop/adapt computational tools

Case study: catchy language

(Some) people craft (some) political and ad slogans, news items, song lyrics, etc. to achieve cultural penetration.

A depressing possibility: does content actually matter, on average?

- **Maybe not:** Salganik, Dodds, Watts “MusicLab” paper, *Science* 2006



Case study: catchy language

(Some) people craft (some) political and ad slogans, news items, song lyrics, etc. to achieve cultural penetration.

A depressing possibility: does content actually matter, on average?

- **Maybe not:** Salganik, Dodds, Watts “MusicLab” paper, *Science* 2006



Movie quotes:
massively,
permanently viral



"FRANKLY, MY DEAR, I DON'T GIVE A DAMN" TOPS AFI'S LIST OF 100 GREATEST MOVIE QUOTES OF ALL TIME

OTHER WINNERS INCLUDE:

THE GODFATHER, **"I'M GOING TO MAKE HIM AN OFFER HE CAN'T REFUSE"**

THE WIZARD OF OZ, **"TOTO, I'VE GOT A FEELING WE'RE NOT IN KANSAS ANYMORE"**

AND CASABLANCA, **"HERE'S LOOKING AT YOU, KID"**



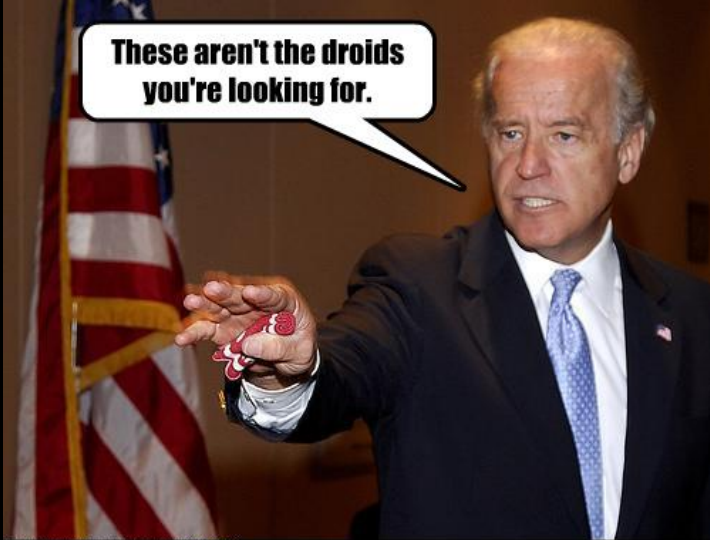
AFI'S 100 Years...100 Movie Quotes: America's Greatest Quips, Comebacks and Catchphrases

LOS ANGELES, June 22, 2005 — The American Film Institute revealed the top movie quotes of all time in **AFI's 100 Years...100 Movie Quotes**, a three-hour special television event on CBS hosted by actor and action star Pierce Brosnan with commentary from many of Hollywood's most celebrated actors and filmmakers. A jury of 1,500 film artists, critics and historians selected "Frankly, my dear, I don't give a damn," spoken by Clark Gable in the celebrated Civil War epic, GONE WITH THE WIND as the most memorable movie quote of all time.

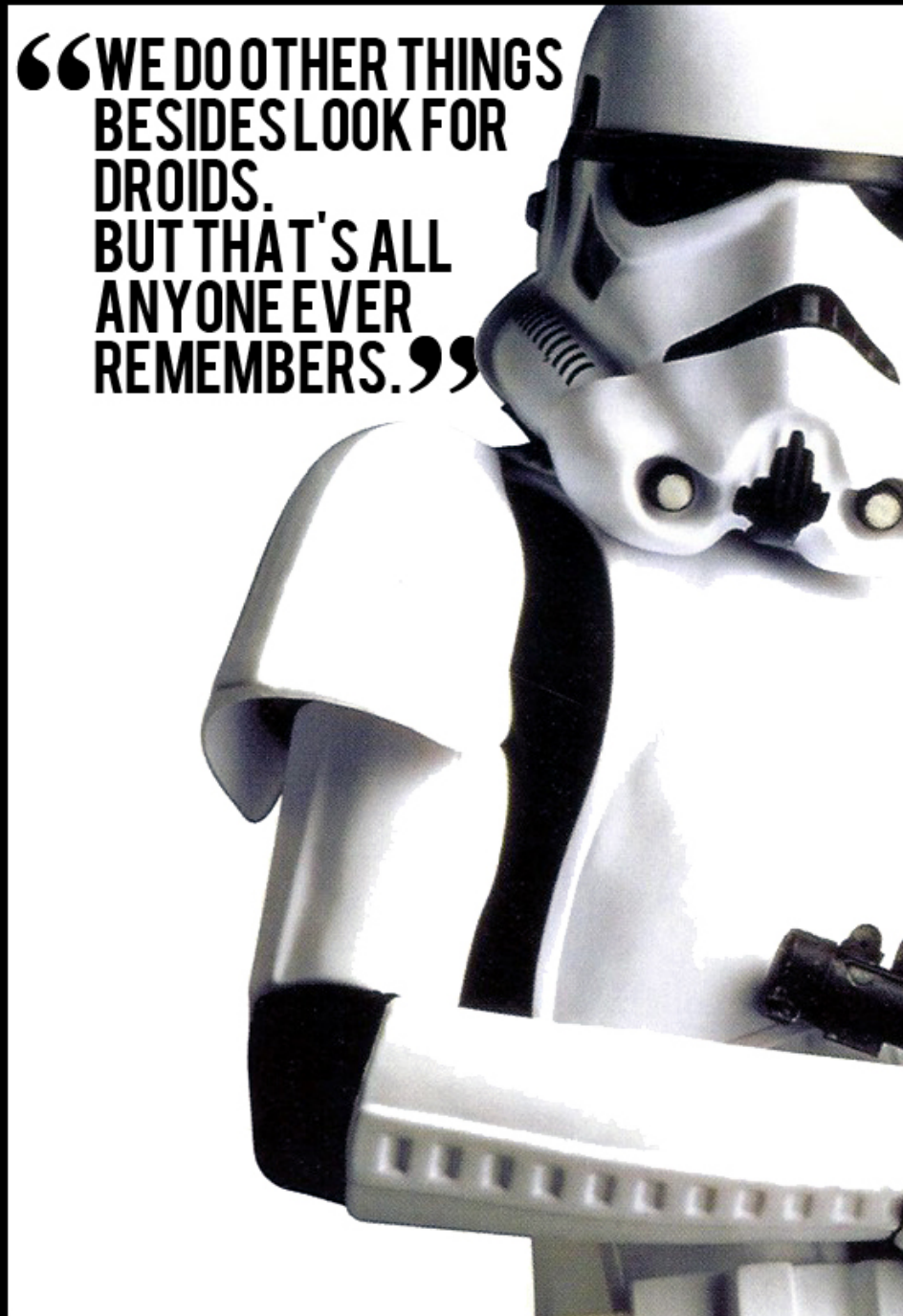
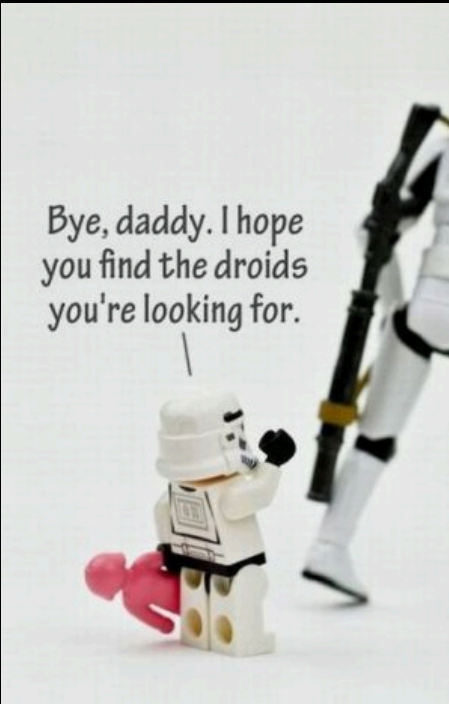
Research question: Does phrasing affect memorability?



Obi-Wan: You don't need to see his identification.
Stormtrooper: [ditto]
Obi-Wan: These aren't the droids you're looking for.
Stormtrooper: [ditto]
Obi-Wan: He can go about his business.
Stormtrooper: [ditto]
Obi-Wan: Move along.
Stormtrooper: [ditto]



<http://nikiedaniel.files.wordpress.com/2011/09/troopers.jpg>
http://ic.qlsc.livejournal.com/level_speak_47/11608882/245440/245440_original.jpg
<http://bloodyhell.com/wp-content/uploads/2009/11/droids-we-are-looking-for-1.jpg>



<http://mikedaniel.files.wordpress.com/2011/09/troopers.jpg>
http://ic.ics.livejournal.com/level_spoek_427/11608882/245440/245440_original.jpg
<http://bloodyot.com/wp-content/uploads/2009/11/droids-we-were-looking-for-1.jpg>

RQ: Does phrasing affect memorability?



**Data: Movie Scripts
with memorability labels (IMDB)**

Obi-Wan: You don't need to see his identification.
Stormtrooper: [ditto]
Obi-Wan: These aren't the droids you're looking for.
Stormtrooper: [ditto]
Obi-Wan: He can go about his business.
Stormtrooper: [ditto]
Obi-Wan: Move along.
Stormtrooper: [ditto]

<http://www.blu-ray.com/movies/screenshot.php?movieid=14903&position=6>

RQ: Does phrasing affect memorability?



Possible prediction setting:

memorable quotes vs. all the rest

Obi-Wan: You don't need to see his identification.

Stormtrooper: [ditto]

Obi-Wan: These aren't the droids you're looking for.

Stormtrooper: [ditto]

Obi-Wan: He can go about his business.

Stormtrooper: [ditto]

Obi-Wan: Move along.

Stormtrooper: [ditto]

RQ: Does **phrasing** affect memorability?



Possible prediction setting:

memorable quotes vs. all the rest

confounds:

memorable movies (e.g., Star Wars)

memorable characters (e.g., Obi-Wan)

memorable positions (e.g., last line of a movie)

length (shorter are easier to remember)

Obi-Wan: You don't need to see his identification.

Stormtrooper: [ditto]

Obi-Wan: These aren't the droids you're looking for.

Stormtrooper: [ditto]

Obi-Wan: He can go about his business.

Stormtrooper: [ditto]

Obi-Wan: Move along.

Stormtrooper: [ditto]

<http://www.blu-ray.com/movies/screenshot.php?movieid=14903&position=6>

RQ: Does **phrasing** affect memorability?



Controlled setting

Match each **memorable quote** with a **non-memorable quote**

Obi-Wan: You don't need to see his identification.

Stormtrooper: [ditto]

Obi-Wan: These aren't the droids you're looking for.

Stormtrooper: [ditto]

Obi-Wan: He can go about his business.

Stormtrooper: [ditto]

Obi-Wan: Move along.

Stormtrooper: [ditto]

RQ: Does **phrasing** affect memorability?



Controlled setting

Match each **memorable quote** with a **non-memorable quote**
from the same character
same place in the movie
same length

... to focus on the effect of phrasing

Obi-Wan: You don't need to see his identification.

Stormtrooper: [ditto]

Obi-Wan: These aren't the droids you're looking for.

Stormtrooper: [ditto]

Obi-Wan: He can go about his business.

Stormtrooper: [ditto]

Obi-Wan: Move along.

Stormtrooper: [ditto]

<http://www.blu-ray.com/movies/screenshot.php?movieid=14903&position=6>

Research question: Does phrasing affect memorability?



Obi-Wan: You don't need to see his identification.
Stormtrooper: [ditto]
Obi-Wan: These aren't the droids you're looking for.
Stormtrooper: [ditto]
Obi-Wan: He can go about his business.
Stormtrooper: [ditto]
Obi-Wan: Move along.
Stormtrooper: [ditto]

<http://www.blu-ray.com/movies/screenshot.php?movieid=14903&position=6>

Gain intuition: Look at the data

Research question: Does phrasing affect memorability?

First quote	Second quote
Half a million dollars will always be missed	I know the type, trust me on this.

Gain intuition: Look at the data

Research question: Does phrasing affect memorability?

First quote	Second quote
Half a million dollars will always be missed	I know the type, trust me on this.
I think it's time to try some unsafe velocities.	No cold feet, or any other parts of our anatomy.

Gain intuition: Look at the data

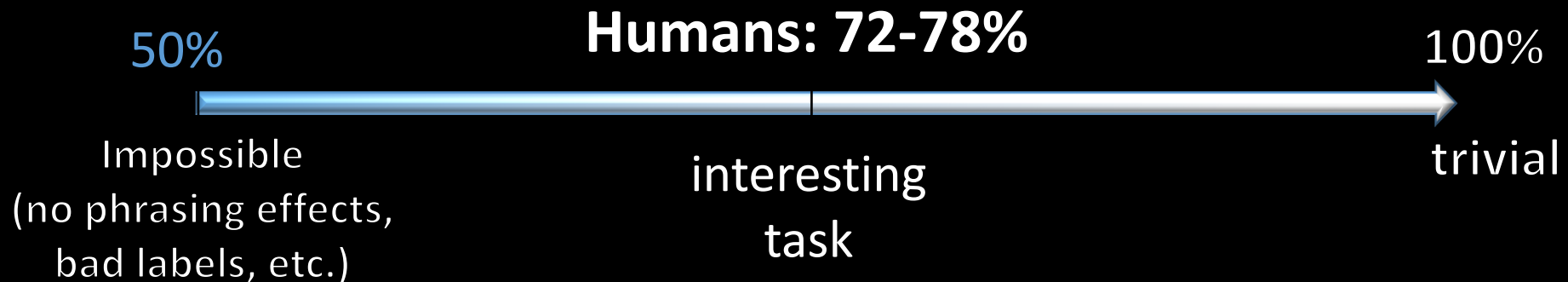
Research question: Does phrasing affect memorability?

First quote	Second quote
Half a million dollars will always be missed	I know the type, trust me on this.
I think it's time to try some unsafe velocities.	No cold feet, or any other parts of our anatomy.
A little advice about feelings kiddo; don't expect it always to tickle.	I mean there's someone besides your mother you've got to forgive.

Gain intuition: Look at the data

Research question: Does phrasing affect memorability?

First quote	Second quote
Half a million dollars will always be missed	I know the type, trust me on this.
I think it's time to try some unsafe velocities.	No cold feet, or any other parts of our anatomy.
A little advice about feelings kiddo; don't expect it always to tickle.	I mean there's someone besides your mother you've got to forgive.



Gain intuition: Look at the data

Hypothesis: Surprising combinations of words are memorable



Blu-ray.com

Hypothesis: Surprising language is memorable

Technique:

measure surprisingness using language models

Toolkits: KenLM, MIT LM Toolkit, SRILM

Creative part:

- A) Where to train the language model
i.e., "Surprising with respect to what?"
- B) How to represent the language?

Hypothesis: Surprising language is memorable

Technique:

measure surprisingness using language models

Toolkits: KenLM, MIT LM Toolkit, SRILM

Creative part:

- A) Where to train the language model
i.e., "Surprising with respect to what?"
- B) How to represent the language?

Here:

- A) Train on fiction that pre-dates the movies (to avoid contamination)

Hypothesis: Surprising language is memorable

Technique:

measure surprisingness using language models

Toolkits: KenLM, MIT LM Toolkit, SRILM

Creative part:

- A) Where to train the language model
i.e., "Surprising with respect to what?"
- B) How to represent the language?

Here:

- B) represent language as sequence of words
→ Surprising combinations of words are more memorable
e.g., "I see dead people."

Hypothesis: Surprising language is memorable

Technique:

measure surprisingness using language models

Toolkits: KenLM, MIT LM Toolkit, SRILM

Creative part:

- A) Where to train the language model
i.e., "Surprising with respect to what?"
- B) How to represent the language?

Here:

B) represent language as sequence of parts of speech

→ Common syntax is more memorable

e.g., "You're gonna need a bigger boat" vs. "You're gonna need a boat that is bigger"

Fitness and diffusion of cultural content (memes)

"Meme-tracking" Leskovec, Backstrom, Kleinberg. 2009

"Memes online" Simmons, Adamic, Adar. 2011

"What's in a name" Himabindu, McAuley, Leskovec. 2013

"QUOTUS" Zhang, Niculae, Suen, Danescu-Niculescu-Mizil, Leskovec. 2015

Another quick LM case study: gender bias in sports journalism
[Fu et al. 2016]

Inspired by covertheathlete.com

Hypothesis: questions to female players are less about the game

Technique:

measure surprisingness using language models

Hypothesis (rewritten in terms of surprise):

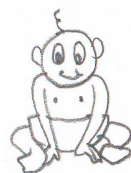
questions to female players are more surprising wrt a game-language

Creative part:

where to train the language model?

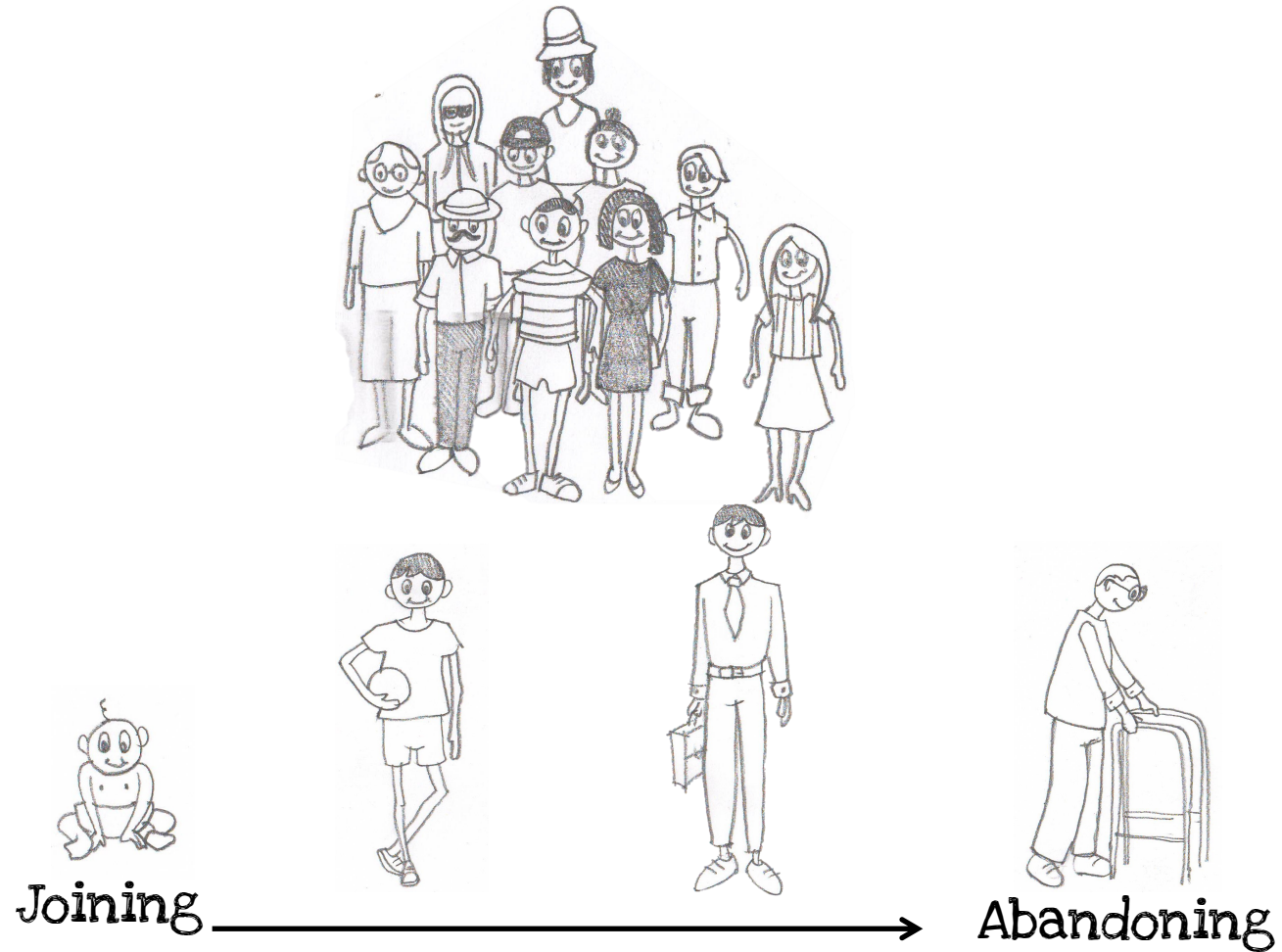
→ play-by-play game commentary

Language (models) capturing user-community dynamics



Joining

Language (models) capturing user-community dynamics



Main intuition: linguistic change

Language norms

- › build collective identity
- › foster individual expression

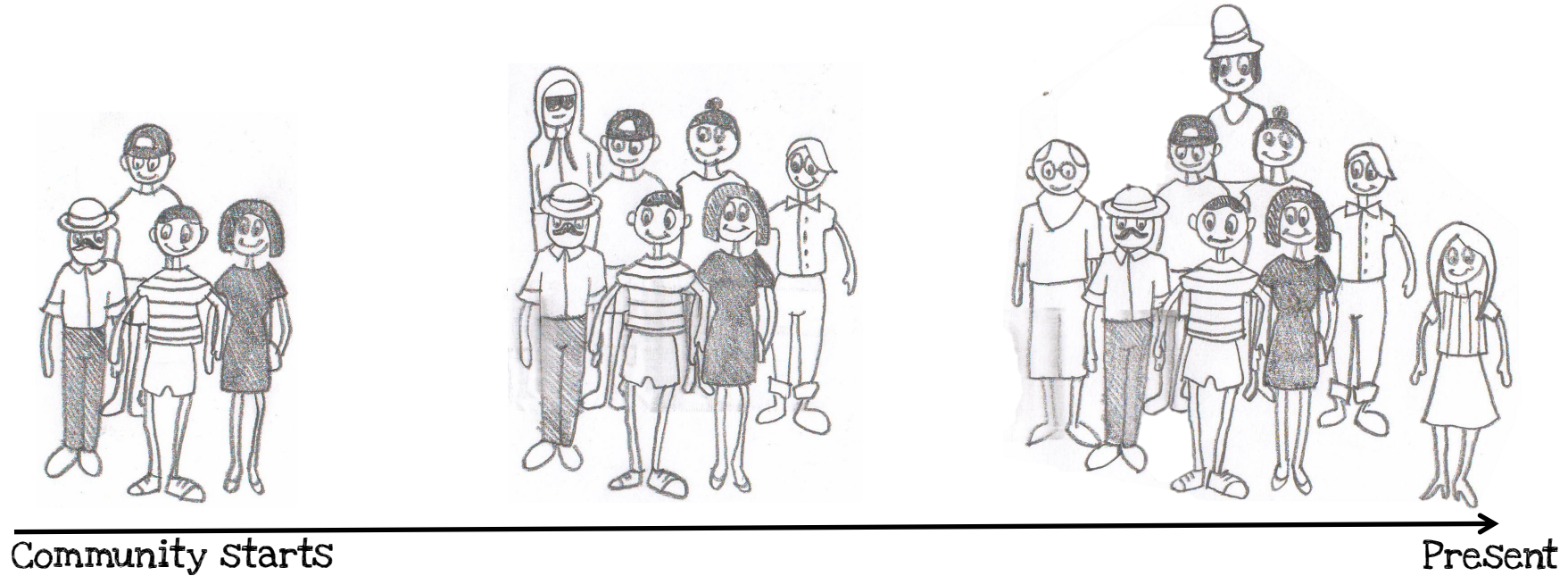
Linguistic change allows us to capture

- › relation between members and their community

“No country for old members” (DaneSCu-NiculeSCu-Mizil et al, 2013)

Longitudinal data

Complete linguistic record of **three** online communities:



Beeradvocate

rate**beer**



BREASTCANCER.ORG

Main intuition: linguistic change

Intuition check:

Norms form online: Language becomes less surprising over time

Main intuition: linguistic change

Intuition check:

Norms form online: Language becomes less **surprising** over time

Main intuition: linguistic change

Intuition check:

Norms form online: **Language** becomes less **surprising** over time

Main intuition: linguistic change

Intuition check:

NormS form online: **Language** becomes less **surprising** over time

Entropy:

$$H(\vec{\theta}) = \sum_i \theta_i \log \frac{1}{\theta_i}, \quad \theta_i = P(\text{string}_i)$$

Main intuition: linguistic change

Intuition check:

NormS form online: Language becomes less surprising over time

Entropy:

$$H(\vec{\theta}) = \sum_i \theta_i \log \frac{1}{\theta_i}, \quad \theta_i = P(\text{string}_i)$$

surprise to see string

Main intuition: linguistic change

Intuition check:

NormS form online: Language becomes less surprising over time

Entropy:

$$H(\vec{\theta}) = \sum_i \theta_i \log \frac{1}{\theta_i}, \quad \theta_i = P(\text{string}_i)$$

prob. [surprise] to see string

Main intuition: linguistic change

Intuition check:

NormS form online: Language becomes less surprising over time

Entropy: expected surprise in a language

$$H(\vec{\theta}) = \sum_i \theta_i \log \frac{1}{\theta_i}, \quad \theta_i = P(\text{string}_i)$$

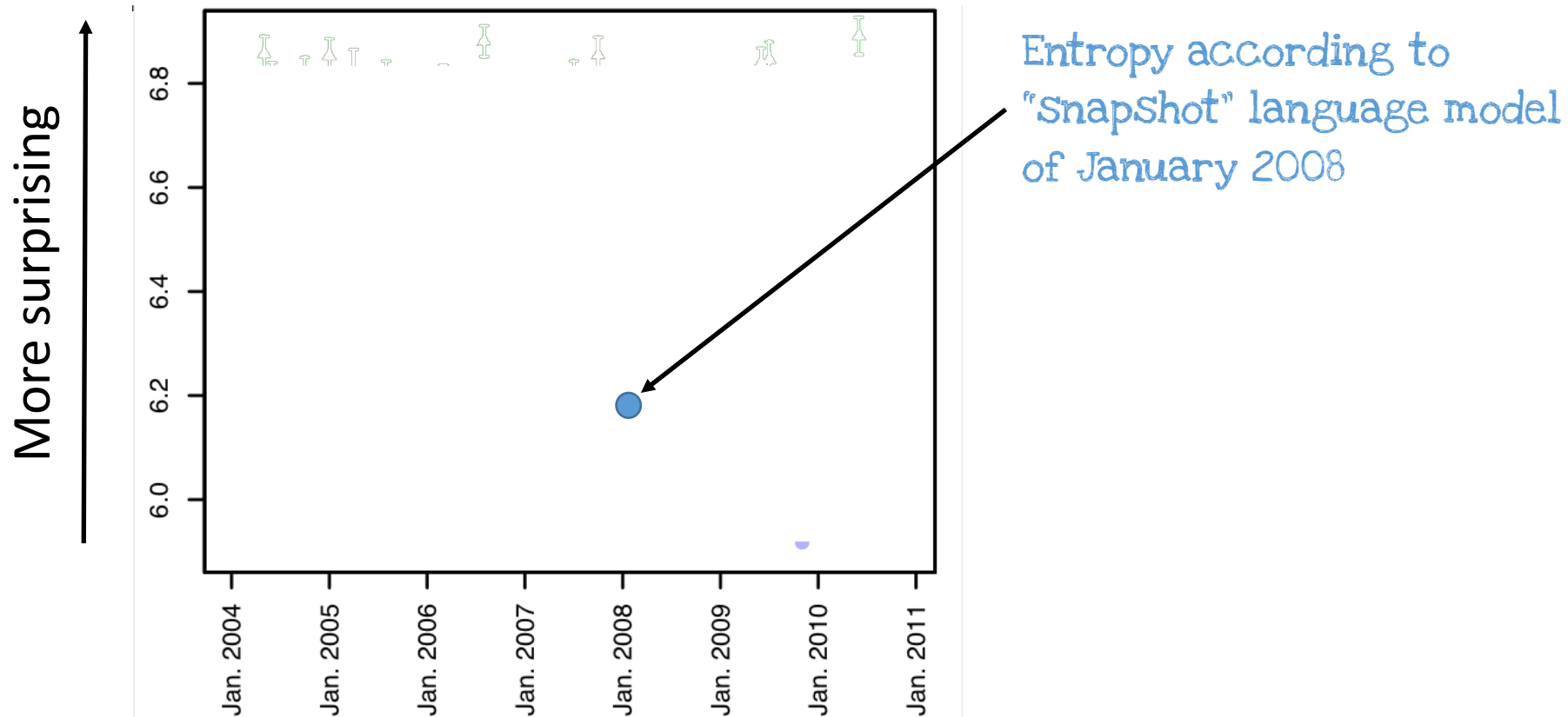
prob. [surprise] to see string

Main intuition: linguistic change

Intuition check:

NormS form online: **Language** becomes less **surprising** over time

Entropy: expected surprise in a language



Main intuition: linguistic change

Intuition check:

NormS form online: Language becomes less surprising over time

Entropy: expected surprise in a language

$$H(\vec{\theta}) = \sum_i \theta_i \log \frac{1}{\theta_i}, \quad \theta_i = P(\text{string}_i)$$

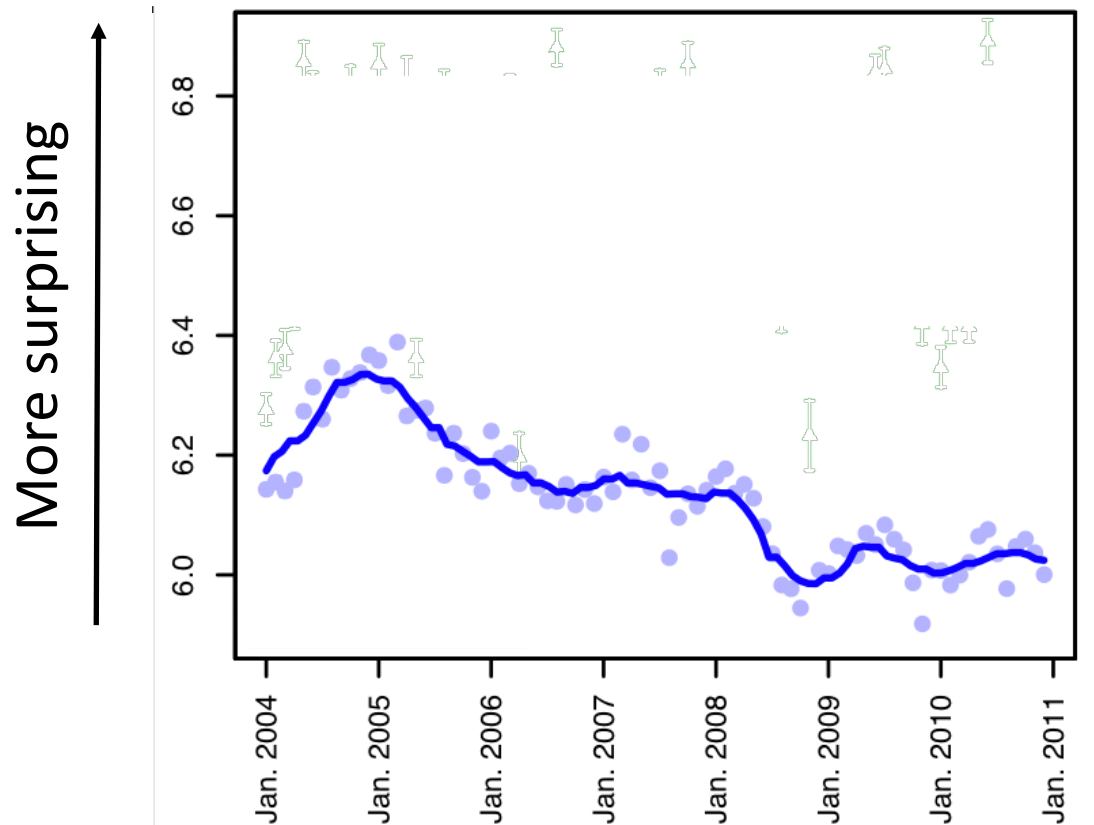
prob. [surprise] to see string

Main intuition: linguistic change

Intuition check:

NormS form online: **Language** becomes less **surprising** over time

Entropy: expected surprise in a language

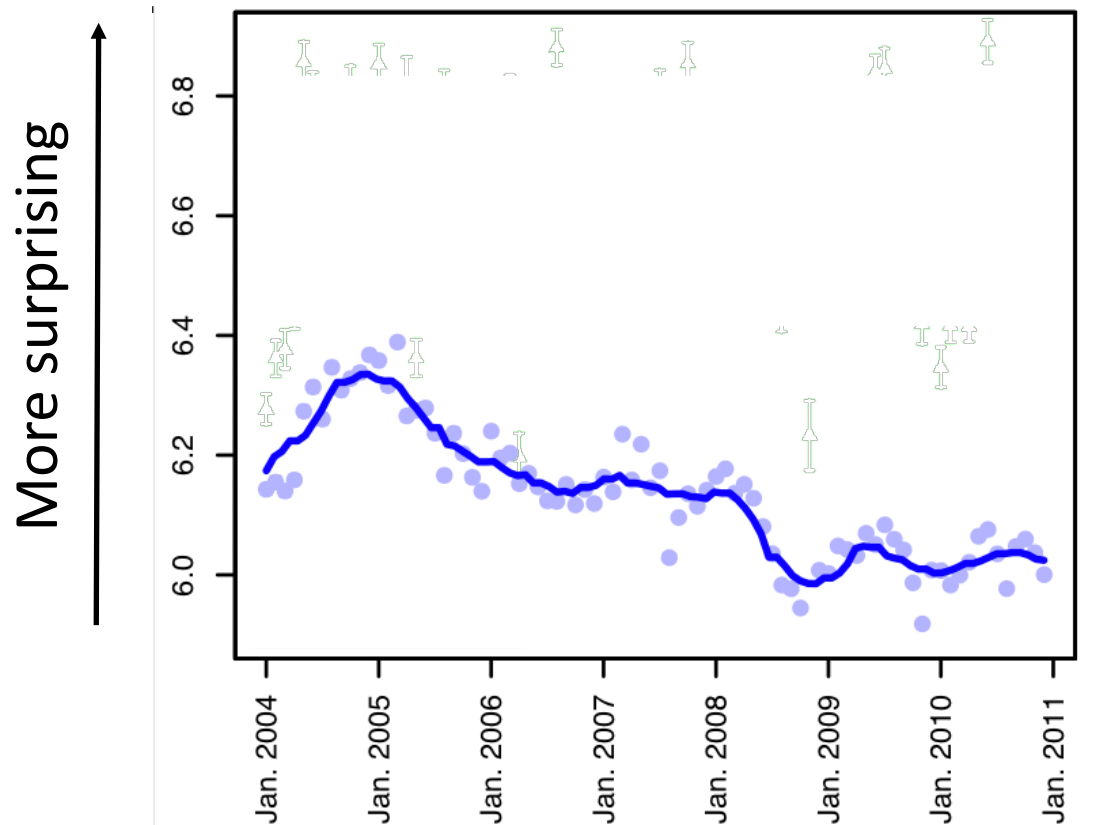


Main intuition: linguistic change

Intuition check:

NormS form online: **Language** becomes less **surprising** over time

Entropy: expected surprise in a language



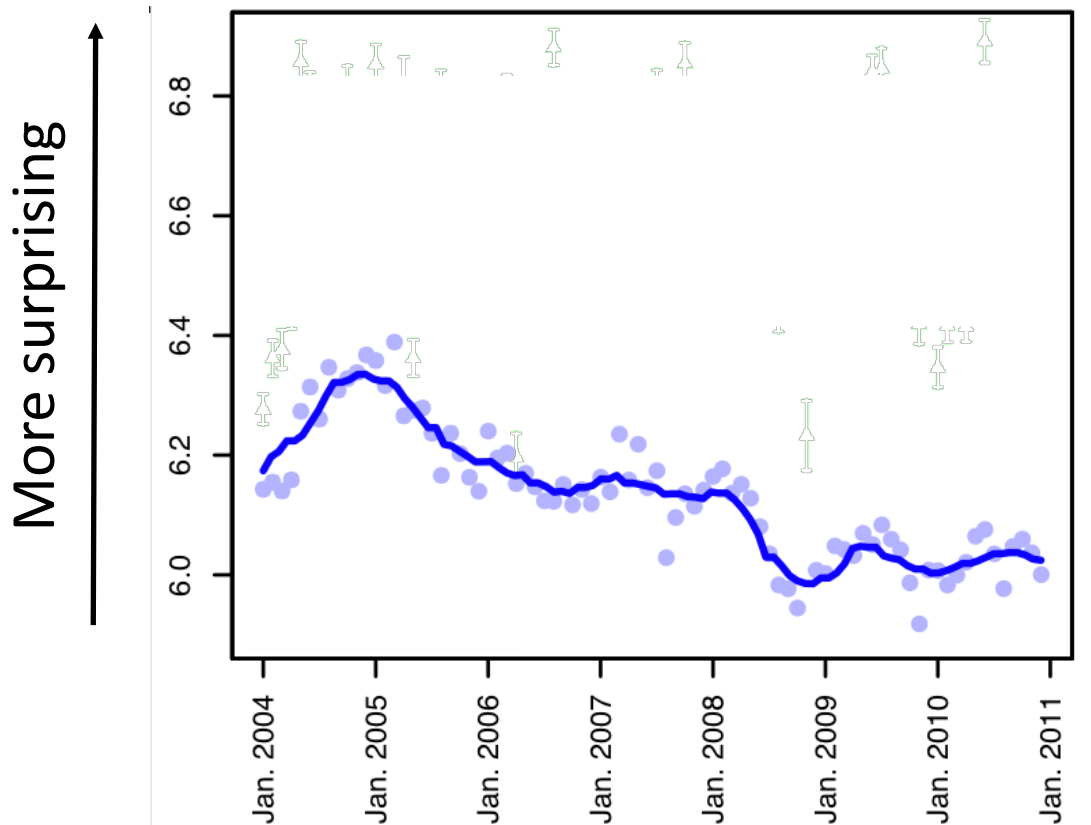
Alternative explanation

Main intuition: linguistic change

Intuition check:

NormS form online: **Language** becomes less **surprising** over time

Entropy: expected surprise in a language



Alternative explanation

Community size grows,
LM is more informed,
So harder to surprise

Main intuition: linguistic change

Intuition check:

Norms take time to learn: Newcomers start farther away

Main intuition: linguistic change

Intuition check:

Norms take time to learn: Newcomers start **farther** away

Main intuition: linguistic change

Intuition check:

Norms take time to learn: Newcomers start **farther** away

Cross-Entropy: expected surprise given a "known" language

Main intuition: linguistic change

Intuition check:

Norms take time to learn: Newcomers start **farther** away

Cross-Entropy: expected surprise given a "known" language

$$H(\vec{\theta}, \vec{\varphi}) = \sum_i \varphi_i \log \frac{1}{\theta_i},$$

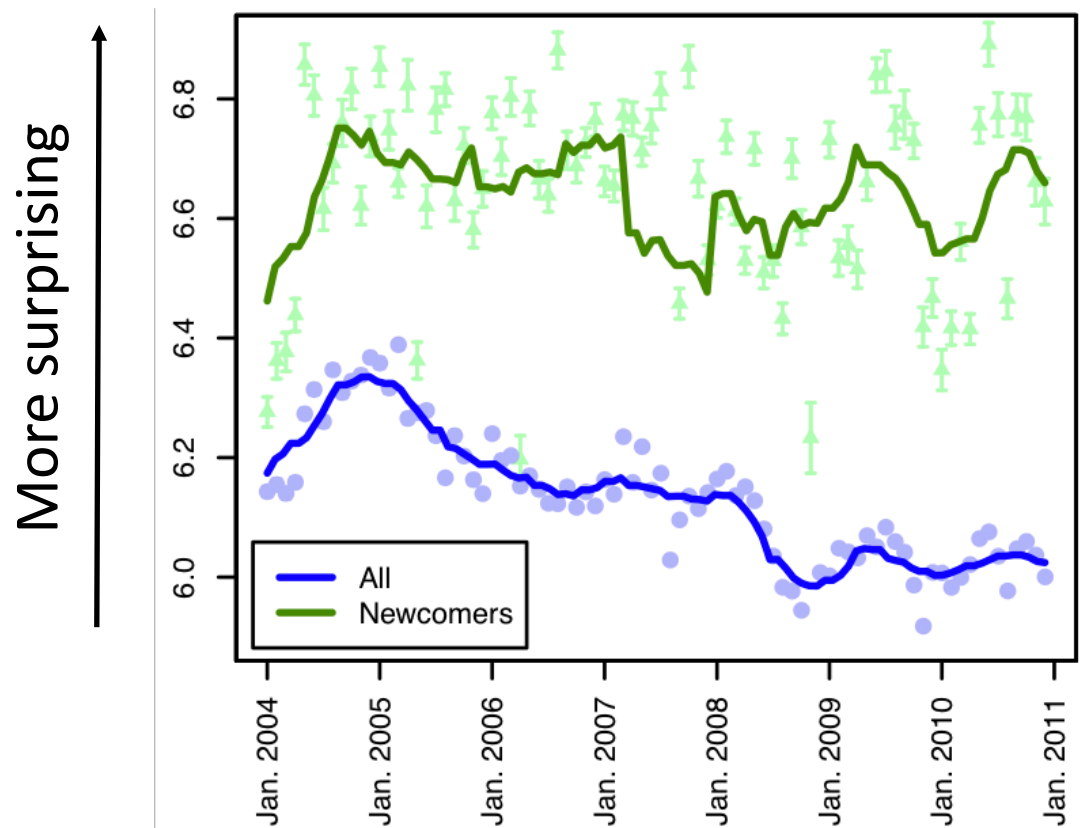
$\theta_i = P(\text{string}_i \text{ in "known" language})$
 $\varphi_i = P(\text{string}_i \text{ in "new" language})$

Main intuition: linguistic change

Intuition check:

Norms take time to learn: Newcomers start **farther** away

Cross-Entropy: expected surprise given a "known" language

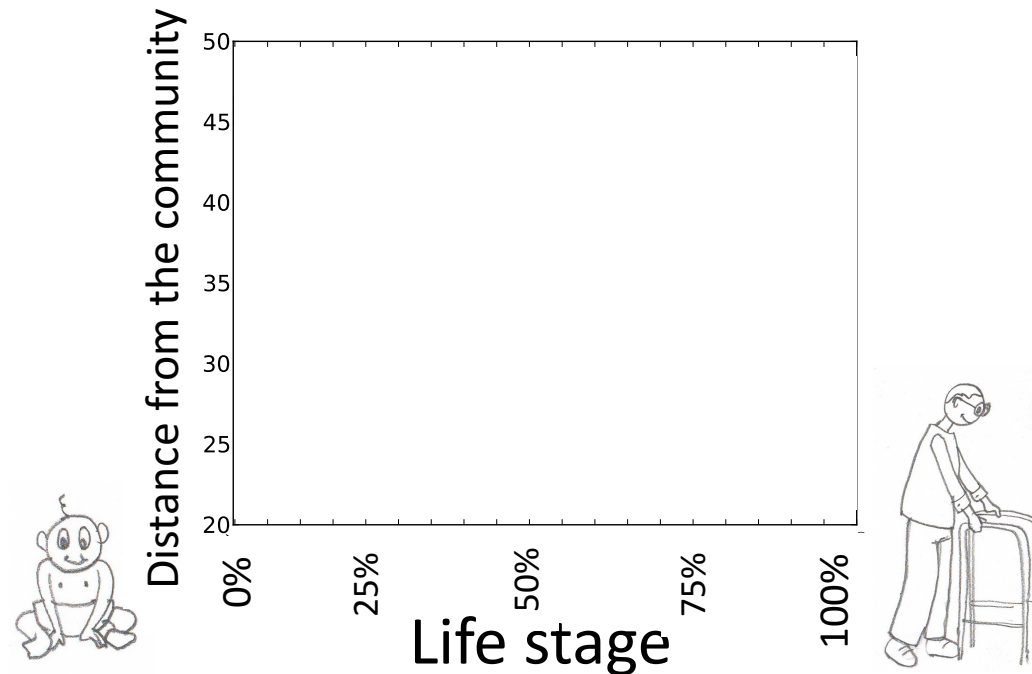


Main intuition: linguistic change

Main results: "No country for old members" (DaneSCu-NiculeSCu-Mizil et al, 2013)

Main intuition: linguistic change

Main results: "No country for old members" (DaneSCu-Niculescu-Mizil et al, 2013)

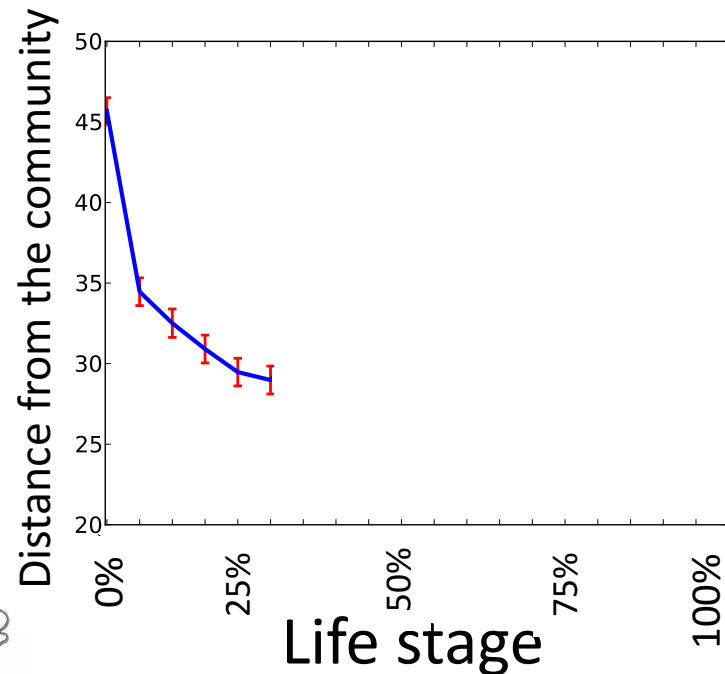


Main intuition: linguistic change

Main results: "No country for old members" (DaneSCu-Niculescu-Mizil et al, 2013)

Stage 1:

user **assimilates**
the language of
the community

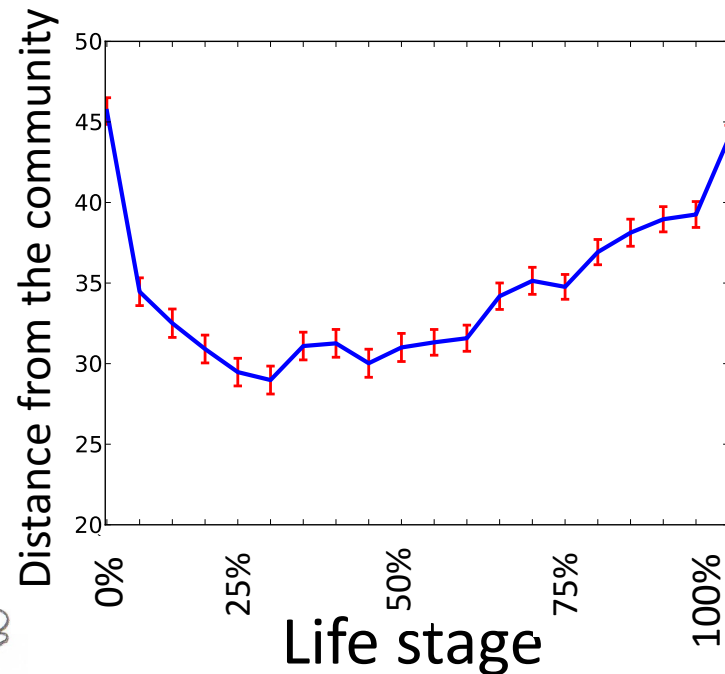
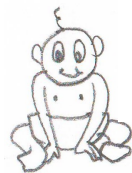


Main intuition: linguistic change

Main results: "No country for old members" (DaneSCu-Niculescu-Mizil et al, 2013)

Stage 1:

user **assimilates**
the language of
the community



Stage 2:

user's language
distances
itself from that
of the community



Language change and social dynamics

Other cool work (links & more on website):

"Social Dynamics of Language Change."
Goel, Soni, Goyal, Paparrizos, Wallach, Diaz, Eisenstein. 2016

"Regional dialects" Eisenstein. 2014

"Geographic variation" Kulkarni, Perozzi, Skiena, 2016

"Semantic Change" Hamilton, Jurafsky, Leskovec. 2016

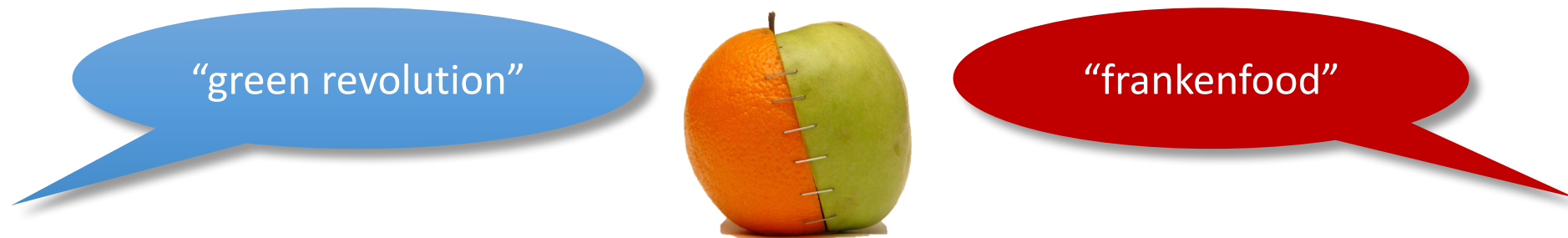
What makes two “languages” different?

Issues analyzed in Kleinberg (2004, *Data Stream Management* 2015)

Presentation/figures follow Monroe, Colaresi and Quinn, *Political Analysis* (2008)

Persuasion: *frame* competition

Example: public discussion of GMOs in food



The *framing* of an argument emphasizes certain principles or perspectives.

“One of the most important concepts in the study of public opinion”

James Druckman (2001)

Other “*CL” framing work includes: Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik (2013); Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, Geri Gay (2015); Oren Tsur, Dan Calacci, David Lazer (2015); Dallas Card, Justin Gross, Amber Boydston, and Noah Smith (2016).

Example: 106th U.S. Senate speeches on abortion

Frames we might expect from Democrats:

... women's rights ...
... privacy ...

Frames we might expect from Republicans:

... unborn children ...
... murder ...

Assume a joint vocabulary of terms v_i .

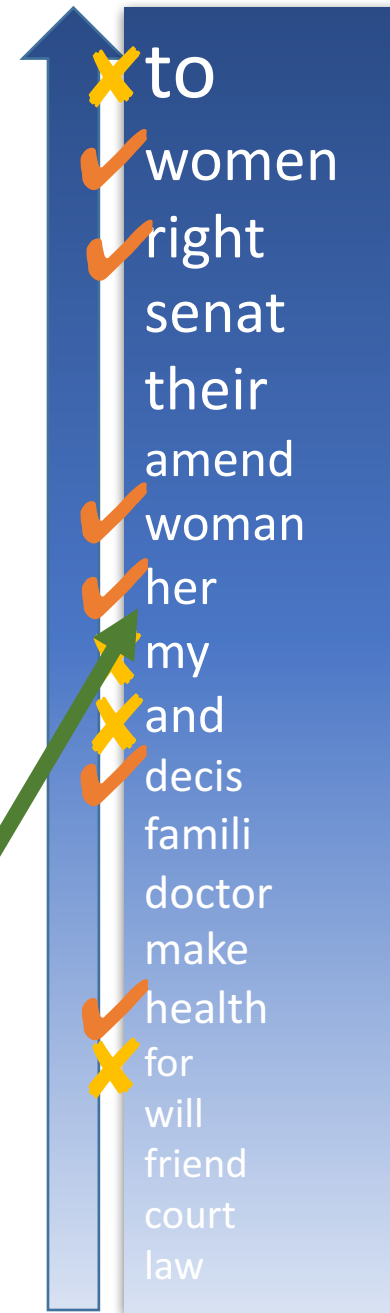
$p(v_i)$ and $p(v_i)$: relative frequency of v_i in the blue and red samples

Ranking using $P(x|class)$

Top and bottom 20 words according to

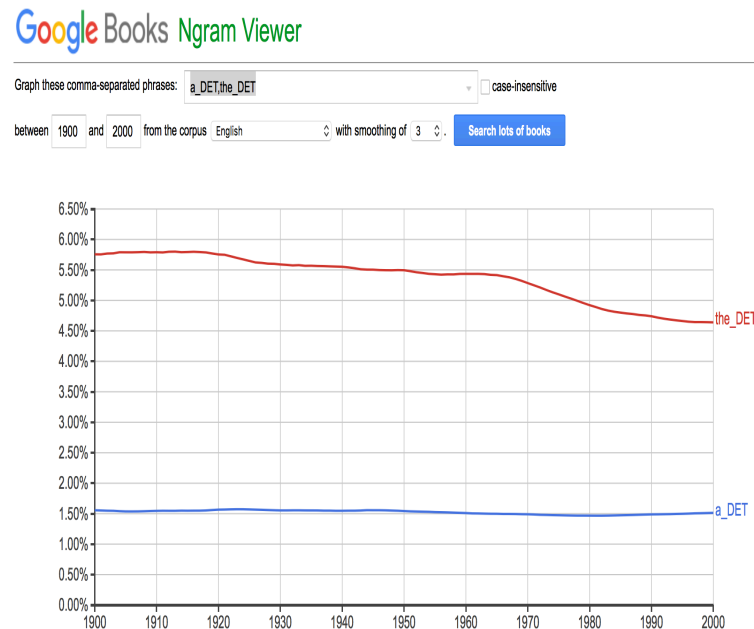
$$p(v_i) - p(v_i)$$

important, but would be lost with stopwords filtering



Aside: “stopword removal” not recommended

- Very-frequent terms have been proving “increasingly” useful, e.g., for stylistic or psychological cues
- “a” vs “the” is surprising

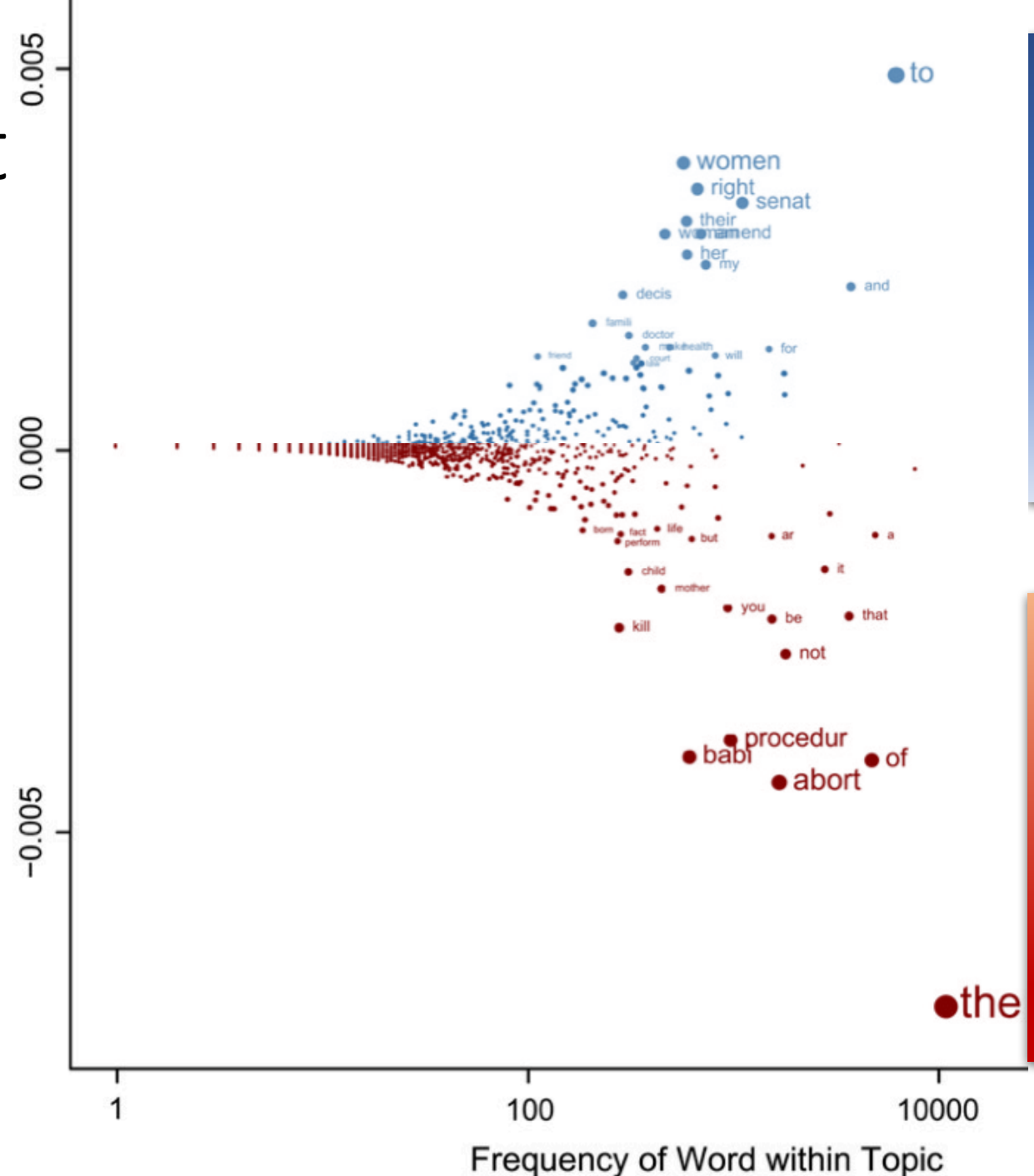


[for years LL assumed this was a bug, but see Language Log, Jan 3 2016]

$P(x|class)$ vs. count

$p(v_i)$ — $p(v_i)$ favors big counts, i.e., v_i towards the righthand side of this plot

(can't have a large difference between two small differences)



to
women
right
senat
their
amend
woman

kill
not
procedur
babi
of
abort
the

Ranking by log odds-ratio

$$\log \frac{p(v_i)/(1 - p(v_i))}{p(v_i)/(1 - p(v_i))}$$

bankruptc

snow

ratifi

confidenti

church

schumer

chosen

voter

wage

1974

attach

attornie

idaho

sadli

coverag

d

juri

mikulsi

tonight

necessarili

martin

peter

leg

harvest

frist

bright

anim

trade

taught

dayton

obvious

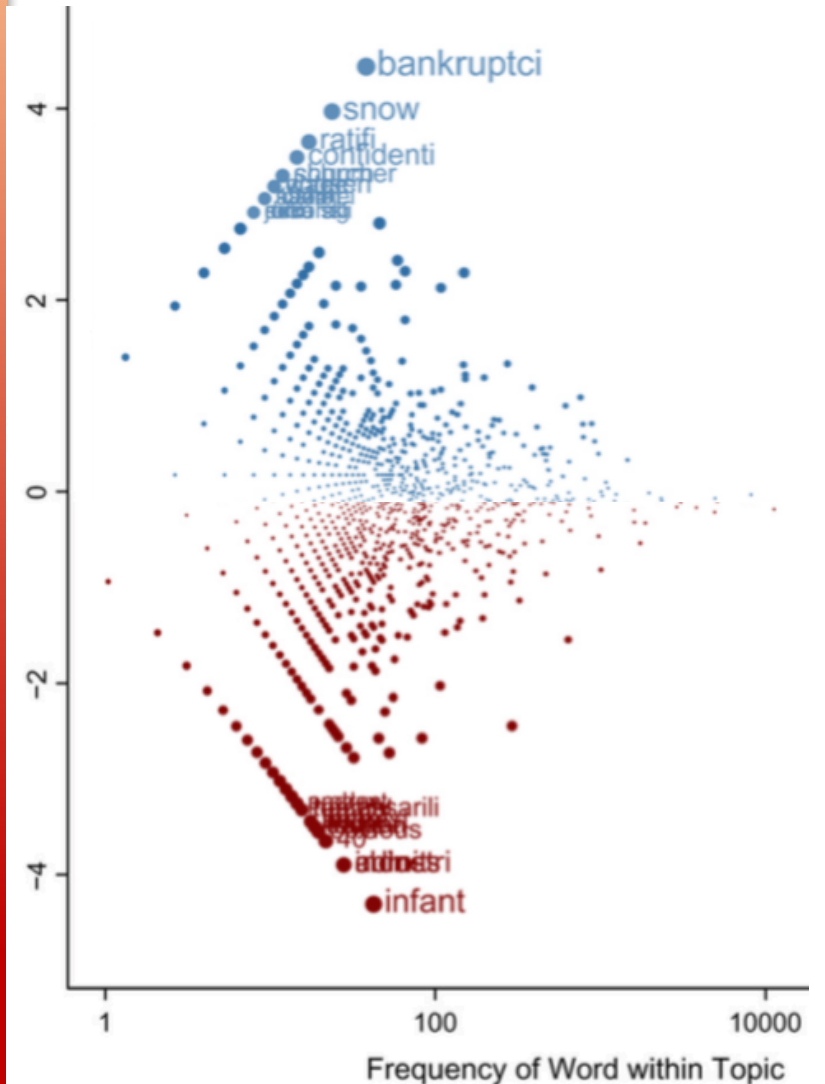
40

industri

chines

admit

infant



Ranking by z-score of log odds-ratio, with model of variance (uniform prior)

women

right

woman

their

decis

famili

amend

her

senat

friend

my

choos

doctor

durbin

serv

pennsylvania

santorum

of
dr

not

partial

fact

birth

head

you

perform

born

the

mother

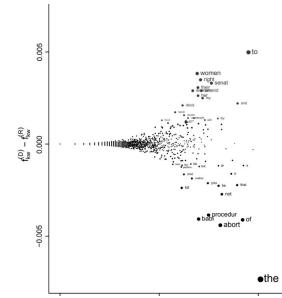
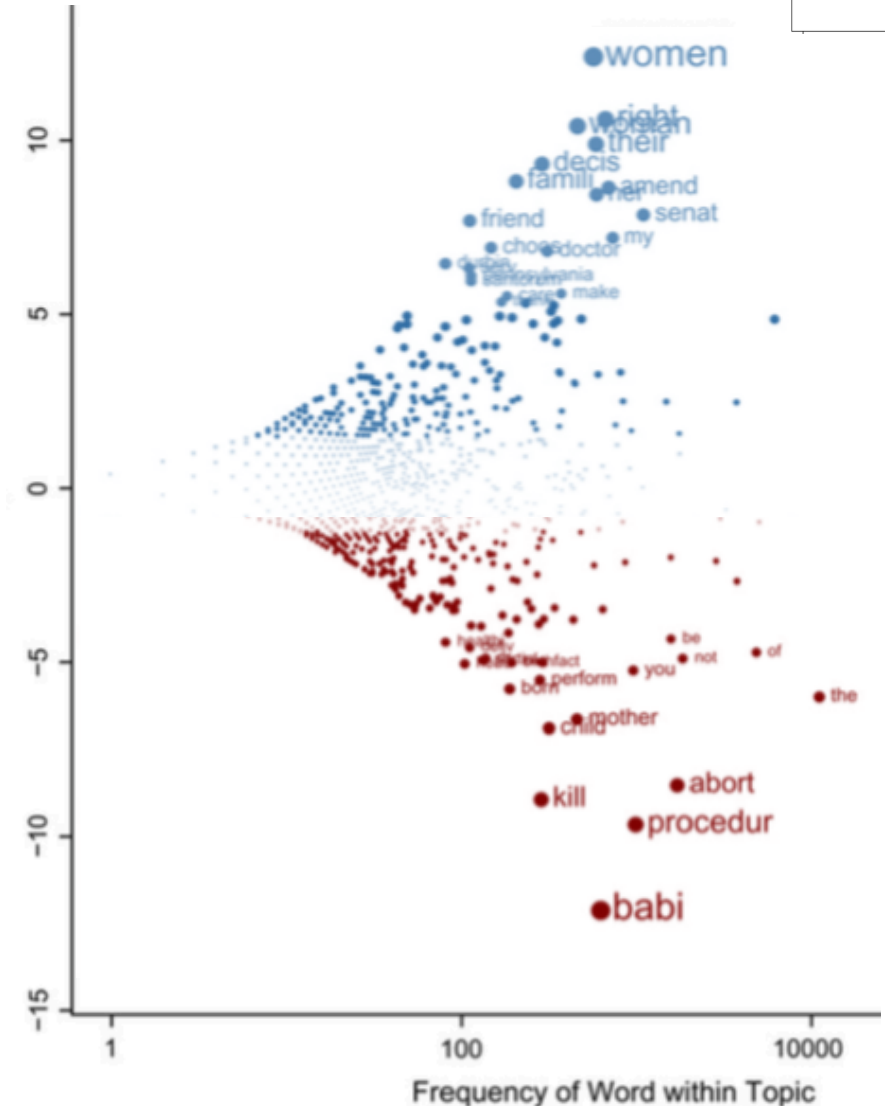
child

abort

kill

procedur

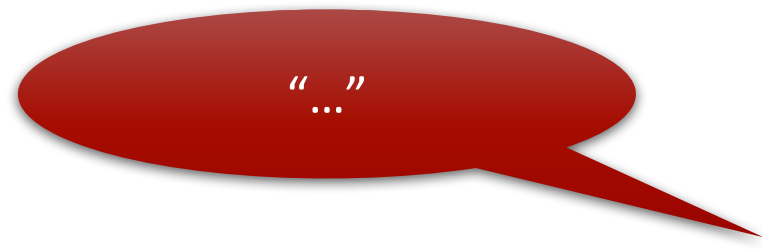
babi



Additional applications: Differentiating the language of

- **successful** vs. **unsuccessful** persuaders
- **low-status** vs. **high-status** people ...
- **males** vs **females**
- *your experimental condition A vs. your experimental condition B!!*

Also good for sanity-checking your data...



Drawing to a close

[The Duchess said,] 'You're thinking about something, my dear, and that makes you forget to talk. I can't tell you just now what the moral of that is, but I shall remember it in a bit.'

'Perhaps it hasn't one,' Alice ventured to remark.

'Tut, tut, child!' said the Duchess. 'Everything's got a moral, if only you can find it.'



Morals you *shouldn't* conclude (we only had two hours together...)

- ~~“More sophisticated NLP isn't used (or doesn't work) for computational social science.”~~
 - example: topic models for differentiating language samples (Blei, Ng, Jordan 2003)
 - example: syntactic correlates of gender differences (Sarawgi, Gajulapalli and Choi 2011)
 - example: discourse modeling of conversational flow
- ~~“We now know all the interesting problems and work there are in computational social science.”~~
 - not even close! (And that's not even counting ethics, fairness, and bias questions...)

Pointers to resources

This tutorial was based on our Cornell course

“Natural Language Processing and Social Interaction”.

For links to papers, conferences, datasets, toolkits, research ideas:

<http://www.cs.cornell.edu/courses/cs6742/> - most recent run (5 so far)

Add one of {2011fa,2013fa,2014fa,2015fa,2016fa} to URL to get that semester;

<http://www.cs.cornell.edu/courses/cs6742/2014fa> has scanned lecture notes

More datasets:

<http://www.cs.cornell.edu/home/llee/data/index.html>

http://www.cs.cornell.edu/~cristian/Data_Media_Talks_News.html

TODAY:

... Research questions
persuasion, linguistic change, framing

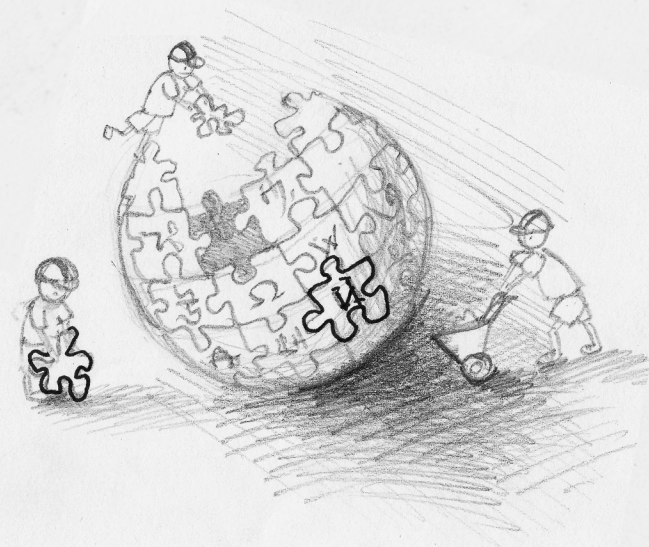
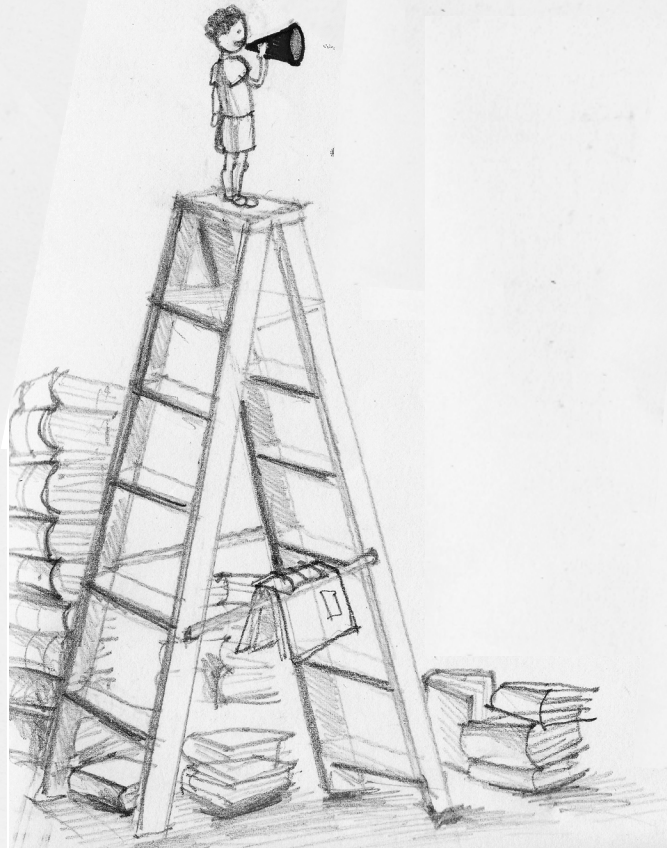
... Techniques
language models, Bayesian feature analysis



... Research practices
controls, feasibility, data inspection

LOOKING FORWARD:

Deeper interplay between
natural language processing
and
how people use and are affected by language
is a huge opportunity for all concerned.



I think this is the
beginning of a
beautiful friendship.

Thanks!

