NLP for computational social science

Dong Nguyen, Alan Turing Institute

The Alan Turing Institute

Traditional data sources in the social sciences

Surveys

Observation

Interviews

Traditional data sources in the social sciences

Surveys

Observation

Interviews

Time consuming ☹

Traditional data sources in the social sciences

Surveys

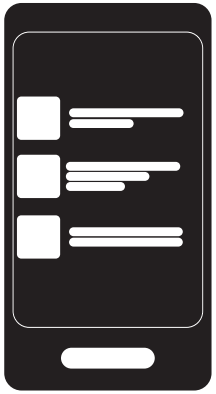
Observation

Interviews

Observer's paradox ☹

Labov (1972): "the aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain these data by systematic observation".

Big social and cultural data



Big social and cultural data

- Informal
- Large amounts of data
- Interaction patterns
- Over time
- Multimodal

Today: Focus on methodological challenges

- NLP for theory building and explanation
- Data, data, data...
  - Biases in data
  - Small vs. big data
- Ethical challenges

NLP for theory building and explanation

Exploration vs prediction

Natural Language Processing

Focus on tasks: *accuracy*, *f-score*, *precision*, *recall*

"[...] there has been an over-focus on numbers, on beating the state of the art."
Manning, 2016

Social sciences & humanities

Interpretation (why?)
(*theory*, *causality*, *interpretability*)

Opinions..

But faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete. [...]
There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

(Anderson, 2008, https://www.wired.com/2008/06/pb-theory)

The two goals in analyzing data which Leo calls prediction and information I prefer to describe as "management" and "science." Management seeks profit, practical answers (predictions) useful for decision making in the short run. Science seeks truth, fundamental knowledge about nature which provides understanding and control in the long run.

(Parzen, comment on Statistical Modeling: The Two Cultures by Leo Breiman, 2001)

Explanation vs. prediction

Explanatory modeling

- minimize bias
- model validation: R^2, significance coefficients, etc.
- risks: type I and type II
- causal relationships
- variables: small number of variables, interpretable

Predictive modeling

- minimize bias + variance
- model validation: external test set
- risks: overfitting
- associations
- variables: Many variables, black box?

Predictive analytics in information systems research, Shmueli and Koppius, MIS Quarterly Vol. 35 No. 3 pp. 553-572, 2011

NLP for theory building and explanation

- 'Traditional' hypothesis testing but use NLP to operationalize variables
- Theory discovery using unsupervised methods
- Large-scale testing of existing theories using prediction models
- Theory discovery using black box(?) prediction models

NLP for theory building and explanation

- 'Traditional' hypothesis testing but use NLP to operationalize variables
- Theory discovery using unsupervised methods
- Large-scale testing of existing theories using prediction models
- Theory discovery using black box(?) prediction models

Cultural fit in organisations

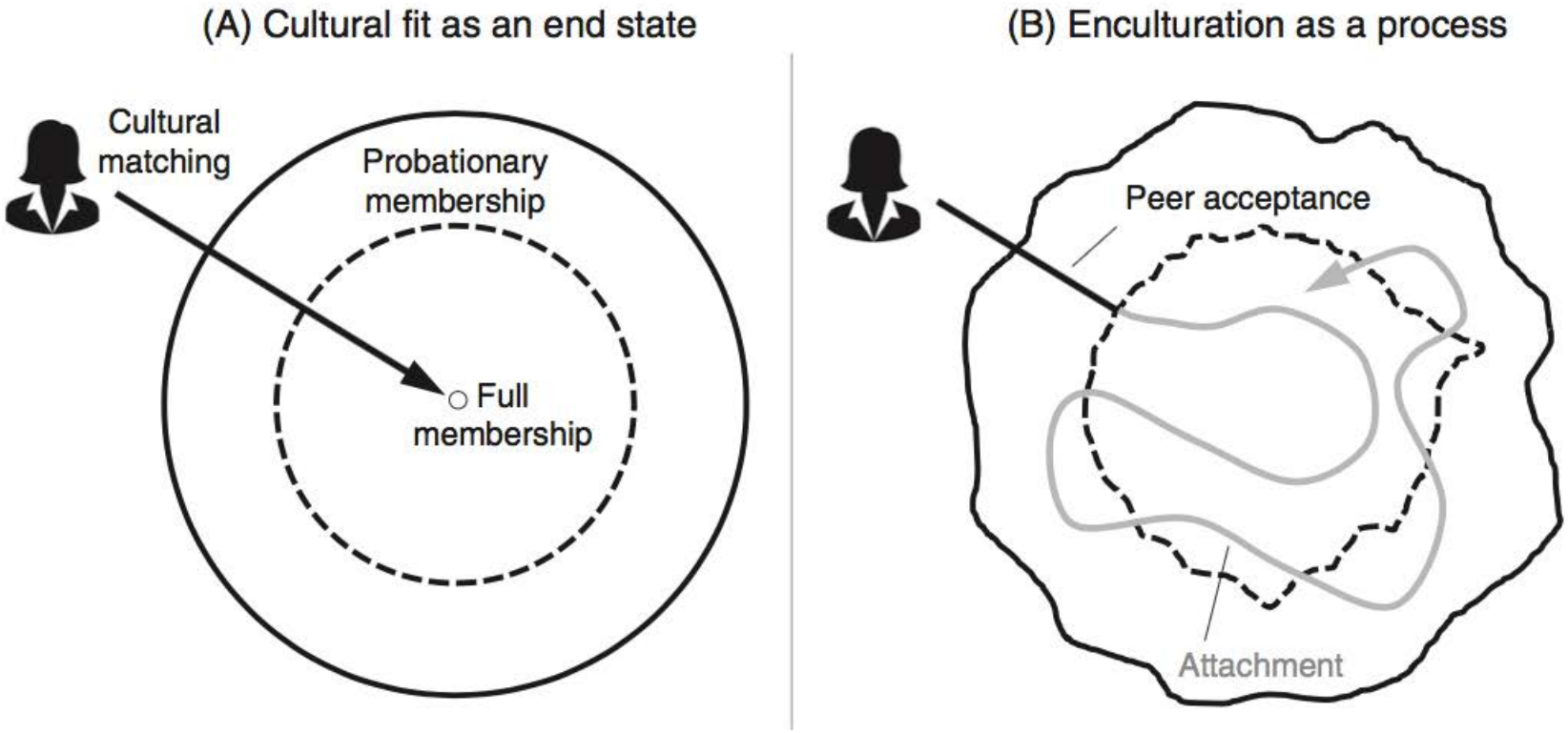
- Data: 10.24 million emails over five years. 601 employees of a midsized U.S. for-profit technology firm.

- How do people adapt in organisations? How does this affect career outcomes?
  - Previously: self reports
    - prone to bias
    - coarse categories
    - difficult to measure temporal variations
    - difficult to scale
  - Now: Measure based on language use

Enculturation Trajectories: Language, Cultural Adaptation, and Individual Outcomes in Organizations, Srivastava et al., 2017

Cultural fit in organisations II

enculturation trajectory: an individual's temporal pattern of cultural fit



Enculturation Trajectories: Language, Cultural Adaptation, and Individual Outcomes in Organizations, Srivastava et al., 2017

Cultural fit in organisations III

Cultural fit: linguistic alignment between an individual and her interaction partners in the organization.

- Measure alignment between incoming and outcoming messages
- Time windows: months

LIWC (Linguistic Inquiry and Word Count):

- Counts words in predefined categories (e.g., swear words, pronouns, insight, anxiety)

$$JS(O\|I) = \tfrac{1}{2}KL(O\|M) + \tfrac{1}{2}KL(I\|M),$$
$$M = \tfrac{1}{2}(O + I)$$

JS: Jensen–Shannon divergence
KL: Kullback-Leibler divergence

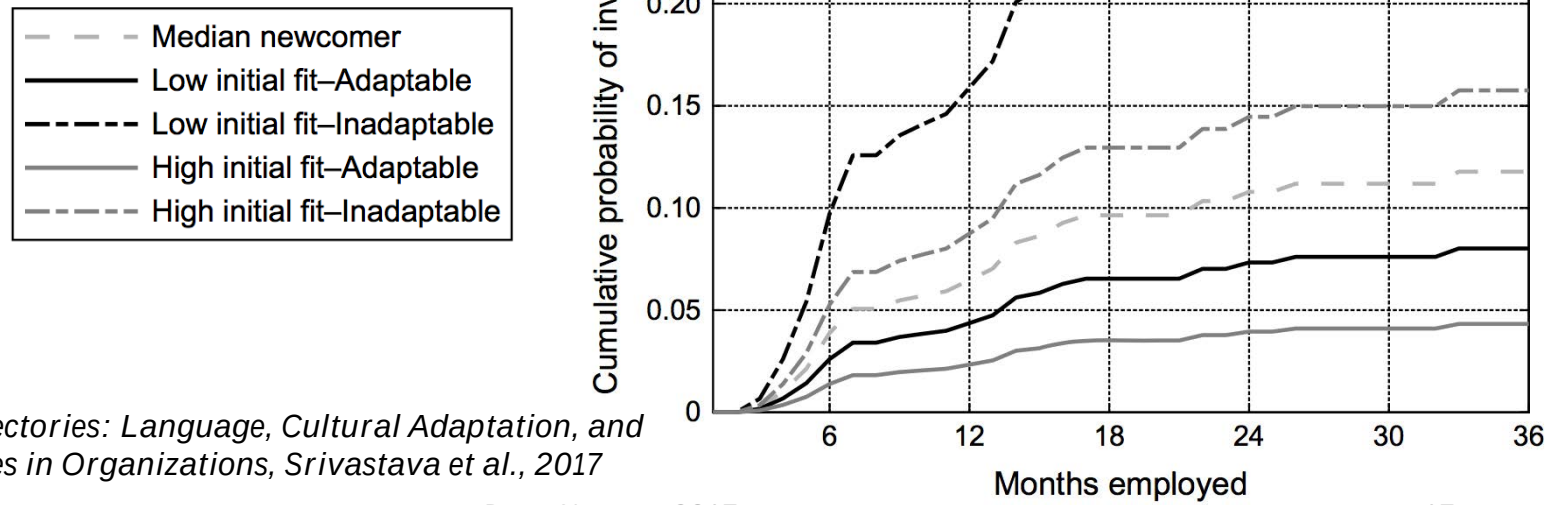
O: distribution over LIWC categories in outcoming messages in period T
I: distribution over LIWC categories in incoming messages in period T

Enculturation Trajectories: Language, Cultural Adaptation, and Individual Outcomes in Organizations, Srivastava et al., 2017

Cultural fit in organisations IV

Findings

- Slow enculturation rates in early stages: more likely to exit involuntarily
- Cultural fit can decline later in careers: sign of voluntary exit



Enculturation Trajectories: Language, Cultural Adaptation, and Individual Outcomes in Organizations, Srivastava et al., 2017

NLP for theory building and explanation

- 'Traditional' hypothesis testing but use NLP to operationalize variables
- Theory discovery using unsupervised methods
- Large-scale testing of existing theories using prediction models
- Theory discovery using black box(?) prediction models

Topic modeling

- Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003)

- Assumptions:
  - A document is a mixture over topics
  - A topic: distribution over a vocabulary
  - Number of topics need to be specified beforehand

- Limitations:
  - Context? (bag of words), interpretative ability.
  - Sensitive to preprocessing steps.
  - Not all topics are meaningful.

LDA example

money	emails	us
tax	evidence	war
government	information	russia
pay	fbi	our
taxes	case	military
business	email	world
our	had	russian
us	wikileaks	foreign
million	investigation	american
work	proof	government

Each topic is a distribution over words
Each document is a mixture of topics

Topic modeling examples

- Themes and author gender in 19th-century literature (Jockers and Mimno, Poetics 2013)
- Framing and agenda setting in four years of public statements issued by members of the U.S. Congress (Tsur et al., ACL-IJCNLP 2015)
- Trends in academic fields based on dissertation abstracts (McFarland et al., Poetics 2013)
- Trends in literary studies (Goldstone and Underwood, 2014)

Grounded theory

- Glaser & Strauss, 1967

- Inductive methodology
- Emergence of conceptual categories
- Grounded in data
- Iterative process (often also repeated data collection)

- Drawbacks: time-consuming, biases of the researcher

Topic modeling vs grounded theory I

- Topic modeling and grounded theory on the same data (survey data with free-text responses).
- Data: Social media user leaves a site and becomes a non-user. 5,245 participants (opt-in to share with researchers)
- Question such as “How did your friends react [to you leaving Facebook]?”

Comparing Grounded Theory and Topic Modeling: Extreme Divergence or Unlikely Convergence?, Baumer et al., JASIST 2017

Topic modeling vs grounded theory II

- Similarities:
  - Iterative process
  - Grounded in data
  - Identify thematic patterns

- Grounded theory: Two researchers. Iterative process: categories were created/combined/removed/changed. Later on initial categories grouped into broader themes.

- LDA: 10 topics. Some pre-processing (lowercase, stop word removal, etc.)

Comparing Grounded Theory and Topic Modeling: Extreme Divergence or Unlikely Convergence?, Baumer et al., JASIST 2017

Topic modeling vs grounded theory III

No simple one-to-one correspondence between topics and themes:

- Topics captured components of a theme
- Most themes associated with at least one, usually two or three, topics
- Topics tend to have a lower level of abstraction

"The grounded theory analysis took two researchers several hours of work per week over roughly 2.5 months. In contrast, a single researcher conducted and wrote up the topic modeling results within a few hours over 2 days."

"these methods involve surprisingly similar processes and produce surprisingly similar results."

Comparing Grounded Theory and Topic Modeling: Extreme Divergence or Unlikely Convergence?, Baumer et al., JASIST 2017

NLP for theory building and explanation

- 'Traditional' hypothesis testing but use NLP to operationalize variables
- Theory discovery using unsupervised methods
- Large-scale testing of existing theories using prediction models
- Theory discovery using black box(?) prediction models


MOVEMBER

Scaling up the Social Identity Model of Collective Action (van Zomeren et al., 2008)

- Injustice: A shared emotion that includes both affective (e.g., anger) and cognitive perceptions (ideology) of an unfair situation

"I had testicular cancer"

"my dad"

"because men's health is important to me"

#SupportTheCause: Identifying Motivations to Participate in Online Health Campaigns, Nguyen et al., 2015

Scaling up the Social Identity Model of Collective Action (van Zomeren et al., 2008)

- Injustice: A shared emotion that includes both affective (e.g., anger) and cognitive perceptions (ideology) of an unfair situation
- Social identity: A sense of belonging together that emerges out of common attributes, experiences and external labels

"my friends asked me again to join them"

"a great excuse to grow a stache"

#SupportTheCause: Identifying Motivations to Participate in Online Health Campaigns, Nguyen et al., 2015

Scaling up the Social Identity Model of Collective Action (van Zomeren et al., 2008)

- Injustice: A shared emotion that includes both affective (e.g., anger) and cognitive perceptions (ideology) of an unfair situation
- Social identity: A sense of belonging together that emerges out of common attributes, experiences and external labels
- Collective efficacy: The shared belief that ones group is capable of resolving its grievances through a campaign

"this campaign can make a difference!"

#SupportTheCause: Identifying Motivations to Participate in Online Health Campaigns, Nguyen et al., 2015

Philip Bloom

Like

My motivation

For 9 years I have been taking part in Movember. It's a cause that is very important to me having lost my grandfather to prostate cancer. Because of this my Uncle got checked and found he also had it but thankfully is OK as he caught it early. Awareness is hugely important as is raising money to beat it! Thanks for helping my campaign! Philip

My fundraising link

http://mobro.co/bloom Copy to clipboard

Donations

£27,605

Target: £25,000

Donate to Philip

Philip has raised £195,444 since 2008

Automatic classification of November profiles

	F1
Injustice	0.816
Social Identity	0.788
Collective efficacy	0.627

Final system

Logistic Regression
unigrams, bigrams, topics, text length, country

Injustice	Social Identity	Collective Efficacy
LDA topic[a]	fun	LDA topic[b]
cancer	team	beat
friend	moustache	and family
lost	mo	change
father	grow	yourself
had	mustache	all of
survivor	LDA topic[c]	awareness
prostrate	fuzz	for movember
for my	movement	awareness of
my	look	last

Table 4: Top-weighted features for free-text motivation experiments.

[a] topic about family/friends who had cancer
[b] topic about raising funds for research
[c] topic about the Movember campaign

#SupportTheCause: Identifying Motivations to Participate in Online Health Campaigns, Nguyen et al., 2015

Findings

- Campaign participants with an injustice motivation raise significantly ($p < 0.001$) more money
- Participants that are part of a team raise significantly more money ($p < 0.001$)
- Participants with a social identity motivation are more often part of a team

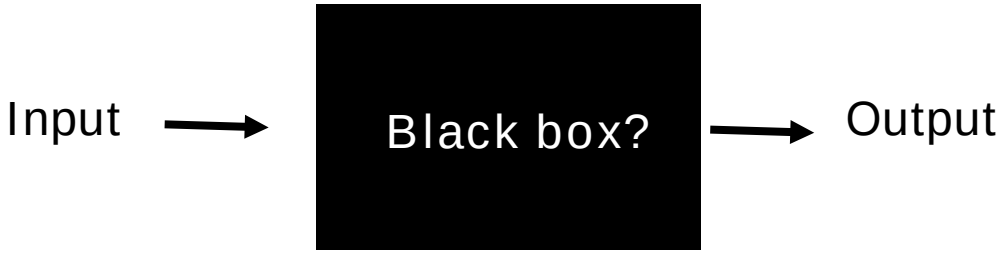
	Injustice	Identity	Efficacy
UK ($)	203.74	128.36	123.39
US ($)	234.47	156.07	169.03

n=90,484

#SupportTheCause: Identifying Motivations to Participate in Online Health Campaigns, Nguyen et al., 2015

Alternative

- Predict amount of donations directly
- Interpret the underlying model?



NLP for theory building and explanation

- 'Traditional' hypothesis testing but use NLP to operationalize variables
- Theory discovery using unsupervised methods
- Large-scale testing of existing theories using prediction models
- Theory discovery using black box(?) prediction models

Interpretable models

- Support theory building and explanation
- Reveal incompleteness in the problem formalization (Doshi-Velez and Kim, 2017)
- Support error analyses and feature discovery (Aubakirova and Bansal 2016)
- Reveal biases in the data

Making the model more interpretable

- Using a simpler model (e.g., logistic regression) instead of a less interpretable model (e.g., deep neural network)
- Regularization (e.g., Lasso/L1)
- Adding monotonicity constraints
  - E.g., probability of having cancer increases monotonically with age (Freitas, 2013)

Extract an interpretable model

- Extract a proxy (a more interpretable model, e.g., decision tree) from a neural network
- Fidelity: Fraction of cases that the proxy agrees with the complex model
- Craven and Shavlik (NIPS 1995) extracted decision trees from neural network
  - Build a decision tree using an oracle
  - Oracle: determines class (as predicted by the neural network) for submitted queries
  - Queries can be instances or specific constraints on the values that the feature can take

Extracting tree-structured representations of trained networks (Craven and Shavlik, NIPS 1995)

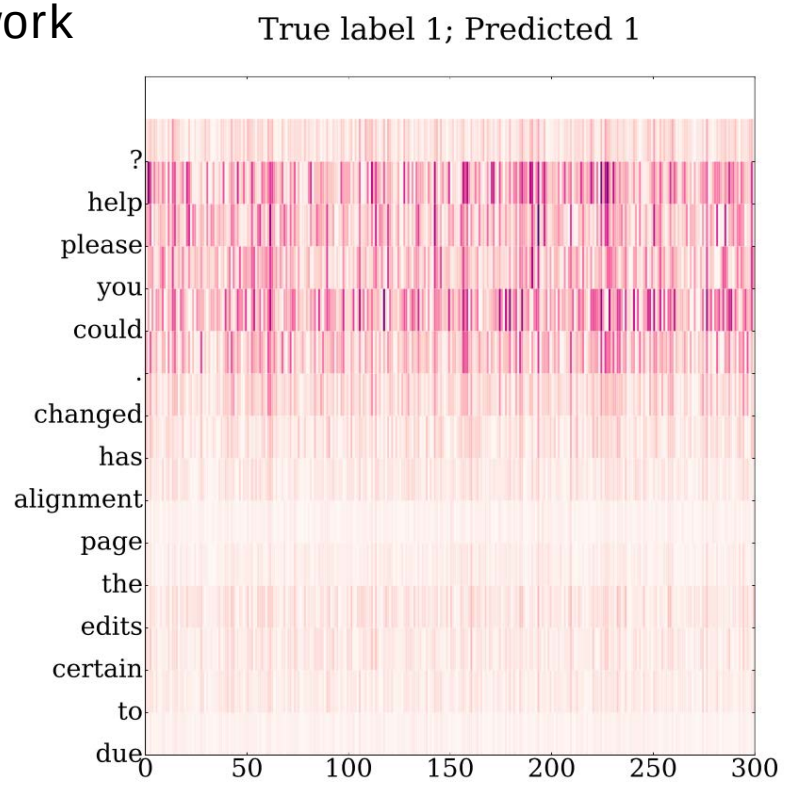
Global vs. local explanation

- Global explanation:
  - Explain the workings of the whole model
  - Could hurt performance (e.g., when adding constraints)
  - Sometimes too complex to explain as a whole

- Local explanation:
  - Explain a specific prediction
  - Can be misleading if local fidelity is low (e.g., doesn't match the black box model)

Interpreting neural networks for politeness prediction I

- Task: distinguish between polite and impolite requests
- Classifier: Convolutional Neural Network
- Methods:
  - Gradients
    - Magnitude of first derivative wrt features



Interpreting Neural Networks to Improve Politeness Comprehension, Aubakirova and Bansal, EMNLP 2016

Interpreting neural networks for politeness prediction I

- Task: distinguish between polite and impolite requests
- Classifier: Convolutional Neural Network
- Methods:
  - Gradients
  - Activation clusters: analyse top-scoring instances for individual units in the CNN
    - Suggested new features/strategies:
      - Indefinite pronouns (*am i missing something here?)*
      - Punctuation (*original article????*)
    - Adding these features improved SVM with linguistic features

Interpreting Neural Networks to Improve Politeness Comprehension, Aubakirova and Bansal, EMNLP 2016

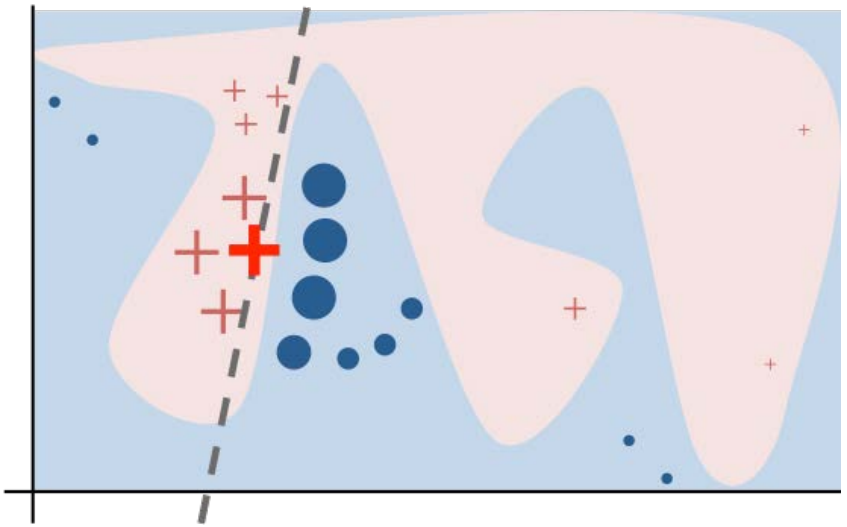
Local explanation: LIME I

Desired characteristics:

- local fidelity: the proxy must behave like the model in the neighborhood of the point of interest
- 'interpretable': e.g., decision trees, linear model
- preferably also: model agnostic

Steps:

- sample around the point of interest by perturbing the data
- fit an interpretable model



"Why Should I Trust You?" Explaining the Predictions of Any Classifier, Ribeiro et. al 2016

https://homes.cs.washington.edu/~marcotcr/blog/lime/

https://github.com/marcotcr/lime

Local explanation: LIME II

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

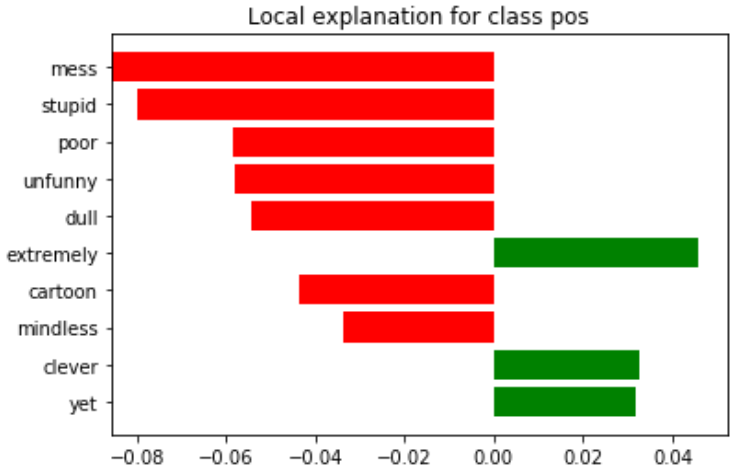
unfaithfulness *g* in approximating *f*.
π_x measures proximity

Complexity of *g*. E.g.,
number of non-zero weights (linear model),
depth of tree (decision tree)

g: interpretable model
f: black box model

"Why Should I Trust You?" Explaining the Predictions of Any Classifier, Ribeiro et. al, KDD 2016

Local explanation: LIME III



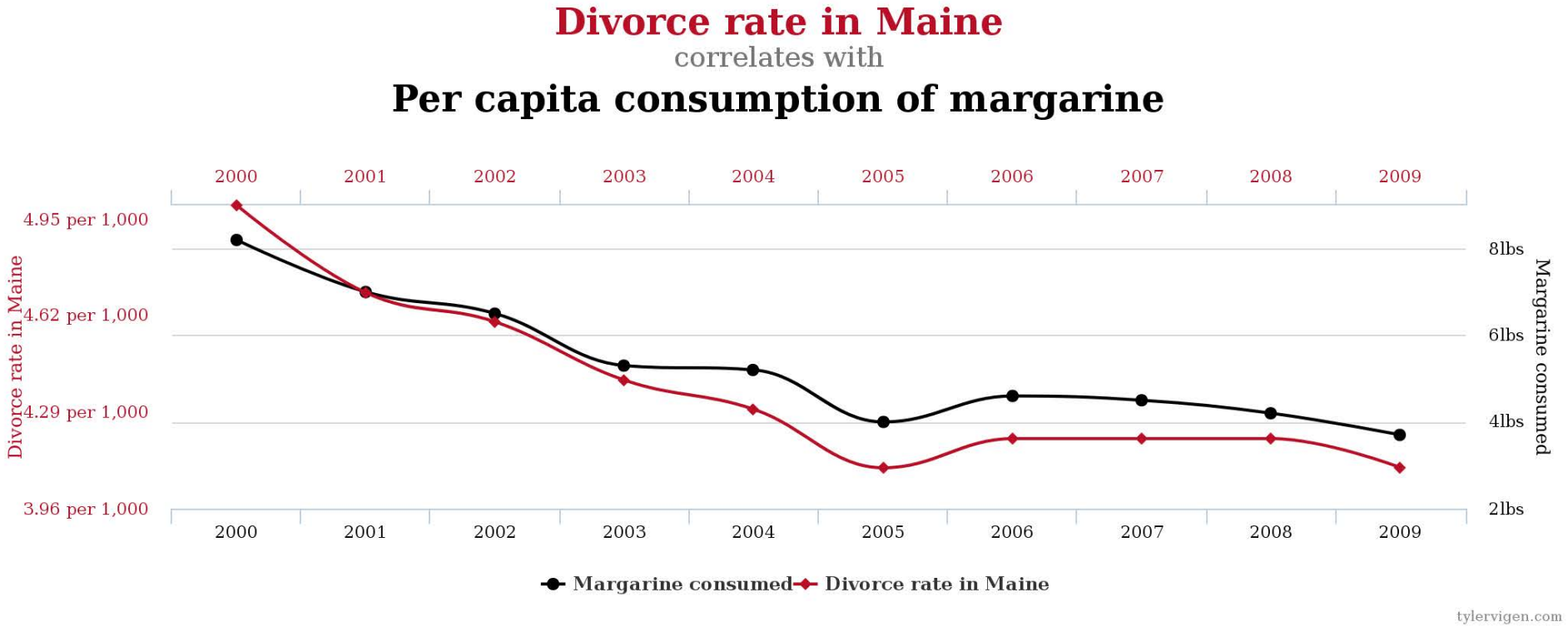
its a stupid little movie that trys to be clever and sophisticated, yet trys a bit too hard. with the voices of woody allen, [..] journey out into the world to find a meaning for life. about 15 minutes into the picture, i began to wonder what the point of the film was. halfway through, i still didn't have an answer. by the end credits, i just gave up and ran out. antz is a mindless mess of poor writing and even poorer voice-overs. allen is nonchalant , while i would have guessed, if i hadn't seen her in the mighty and basic instinct, stone can't act , even in a cartoon. this film is one for the bugs: unfunny and extremely dull. hey, a bug's life may have a good time doing antz in.

"Why Should I Trust You?" Explaining the Predictions of Any Classifier, Ribeiro et. al, KDD 2016

Interpretable models?

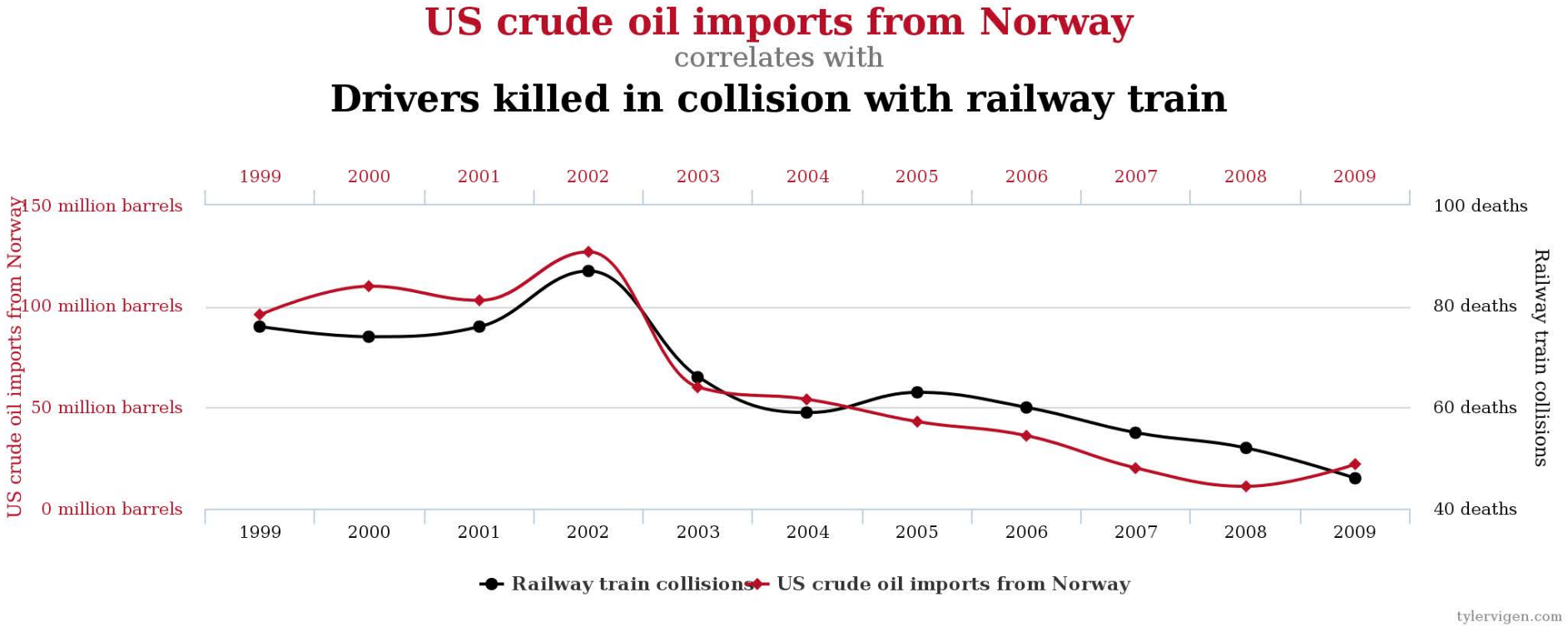
- Might be a way for predictive modeling to support theory building and explanation
- But... interpretability is not well defined (Lipton 2016)
- Many challenges in evaluation

Causality I



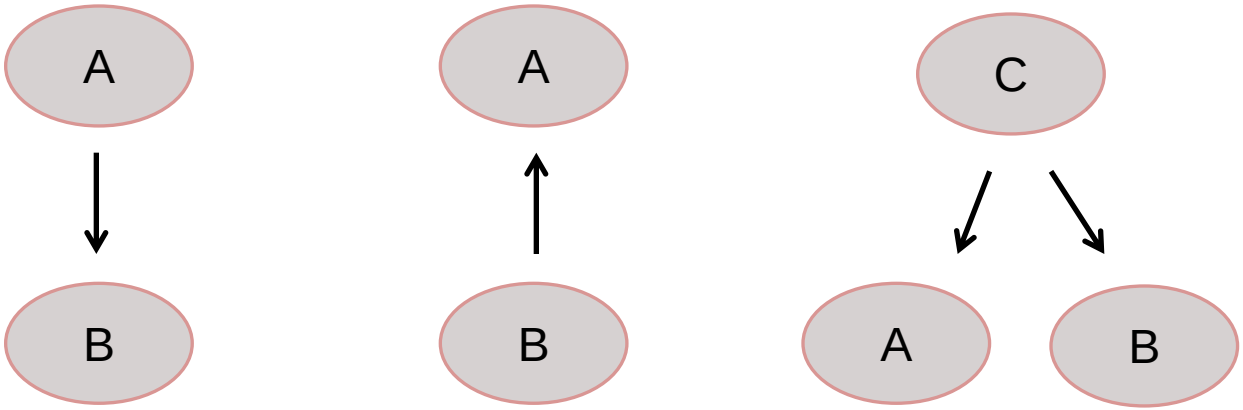
http://tylervigen.com/spurious-correlations

Causality II



http://tylervigen.com/spurious-correlations

Causality III



Summary

- Many different ways for NLP to contribute to theory building and explanation in the social sciences!

- Challenges:
  - Explanation vs. prediction. Interpretability of models.
  - Correlation vs causation.

References

- Comprehensible classification models: a position paper, Freitas, ACM SIGKDD Explorations Newsletter 2013
- Predictive analytics in information systems research, Shmueli and Koppius, MIS Quarterly Vol. 35 No. 3 pp. 553-572, 2011
- “Why Should I Trust You?” Explaining the Predictions of Any Classifier, Ribeiro et al., KDD 2016
- Comparing Grounded Theory and Topic Modeling: Extreme Divergence or Unlikely Convergence?, Baumer et al., JASIST 2017
- #SupportTheCause: Identifying Motivations to Participate in Online Health Campaigns, Nguyen et al. EMNLP 2015.
- Extracting tree-structured representations of trained networks, Craven and Shavlik, NIPS 1995
- Interpreting Neural Networks to Improve Politeness Comprehension, Aubakirova and Bansal, EMNLP 2016
- The Mythos of Model Interpretability, Lipton 2016

Data bias

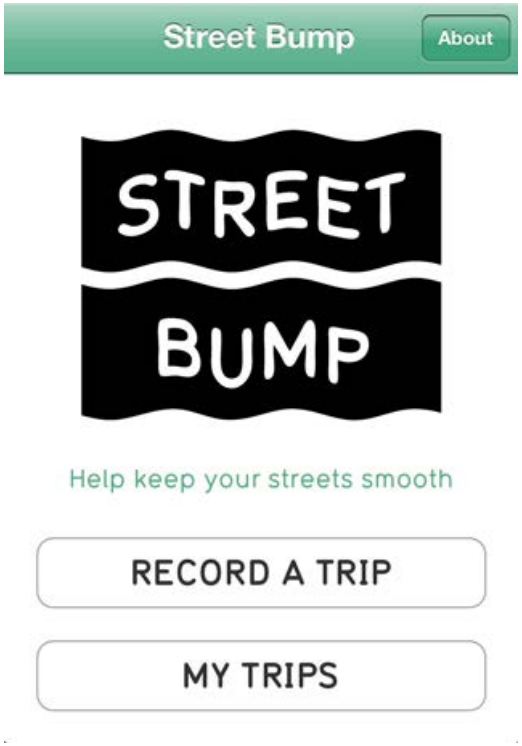
Representativeness & bias

- Representativeness
  - Offline population
  - Relevant content
  - Behavior

Biases may be introduced during the complete research pipeline. Here: focus on biases due to data source selection and data collection

Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. Olteanu et al.
http://www.aolteanu.com/SocialDataLimitsTutorial/

Data source selection I



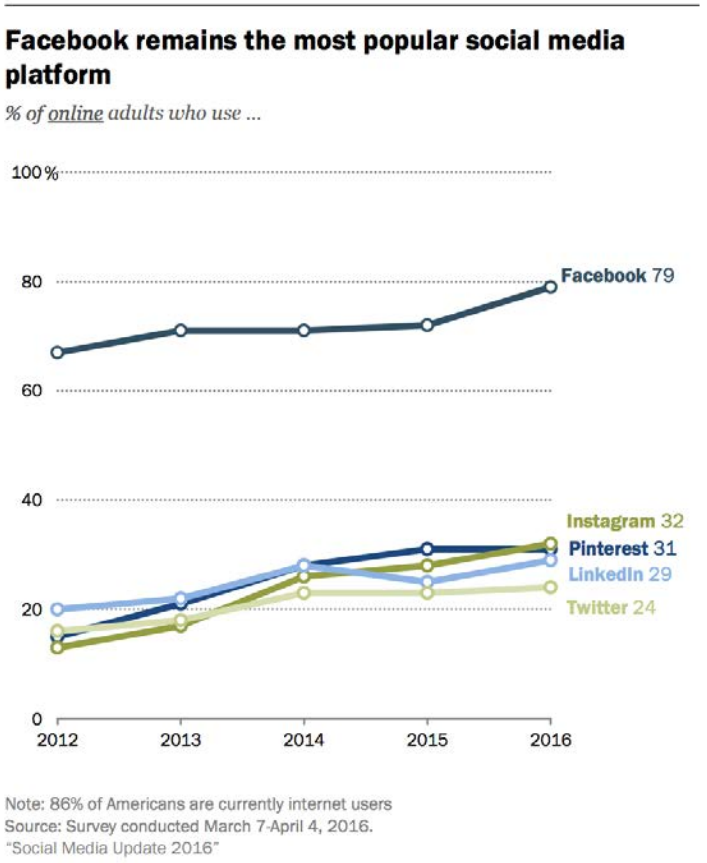
Lower income groups and older residents are less likely have smartphones

Twitter and Foursquare data to study hurricane Sandy. Data gives the illusion that Manhattan was the center of disaster.

https://hbr.org/2013/04/the-hidden-biases-in-big-data
(Crawford, 2013)

Data source selection II

Twitter is the 'model organism' for social media studies (Tufekci, 2014)



Pew Research social media update (Nov 2016)
http://www.pewinternet.org/2016/11/11/social-media-update-2016/pi_2016-11-11_social-media-update_0-01/

Data source selection III

Twitter is the 'model organism' for social media studies (Tufekci, 2014)

But different platforms have different:

- mechanisms that shape user behaviour
- norms
- demographics
- Etc...

Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. Olteanu et al.

Bias in Twitter: demographics

	20-	20-40	40+
M	796	488	265
F	1078	316	157

Table 2: Age and gender

Nguyen et al., ICWSM 2013

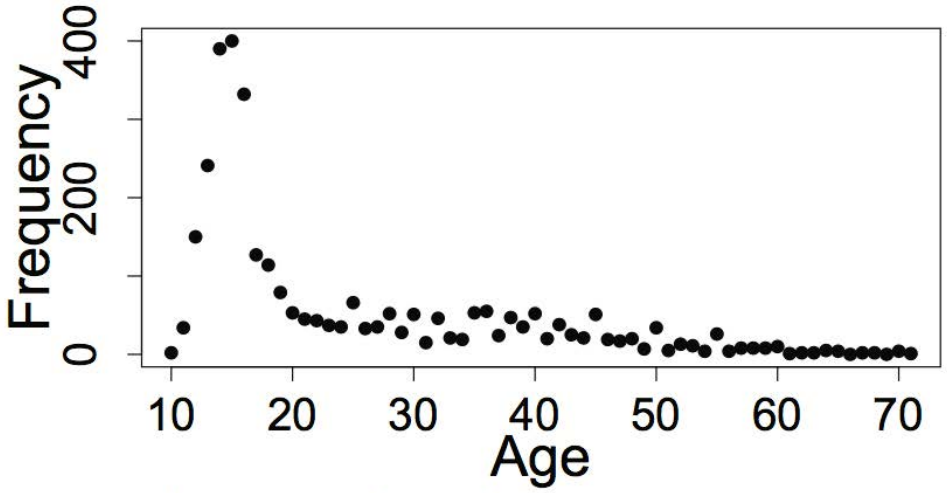


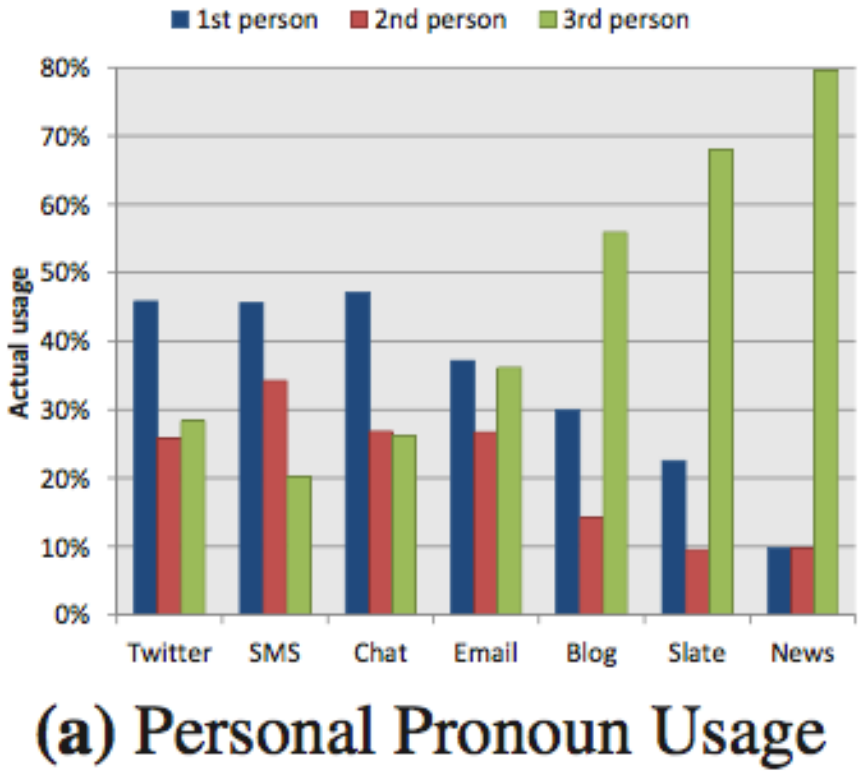
Figure 1: Plot of frequencies per age

Controlling for only one demographic variable is not enough!

Also observed in blogs (Schler et al. , 2006)

Bias in Twitter: language I

Dude, srsly?: The Surprisingly Formal Nature of Twitter's Language, Hu et al., ICWSM 2013



(a) Personal Pronoun Usage

Bias in Twitter: language II

Corpus	Documents	Average words per document
TWITTER-1	1 000 000	11.8 ± 8.3
TWITTER-2	1 000 000	11.6 ± 8.1
COMMENTS	874 772	15.8 ± 18.6
FORUMS	1 000 000	23.2 ± 29.3
BLOGS	1 000 000	147.7 ± 339.3
WIKIPEDIA	200 000	281.2 ± 363.8
BNC	3141	$31\,609.0 \pm 30\,424.3$

Corpus	Parseable				Unparseable
	strict		informal		
	full	frag	full	frag	
TWITTER-1	13.8	23.9	22.2	2.5	37.4
TWITTER-2	13.9	23.8	22.8	1.7	37.6
COMMENTS	18.0	22.2	26.4	1.4	31.9
FORUMS	23.9	14.1	24.7	1.5	35.6
BLOGS	25.6	17.5	18.8	2.7	35.3
WIKIPEDIA	48.7	4.5	18.9	1.5	26.2
BNC	38.4	12.0	24.0	2.2	23.2

How Noisy Social Media Text, How Diffrnt Social Media Sources?, Baldwin et al., IJCNLP 2013

Bias Twitter: sampling

Political orientation
(republicans vs democrats):

Sampled users according to a different degree of political engagement on Twitter: 'normal' users are much harder to classify!

Classifying Political Orientation on Twitter: It's Not Easy!, Cohen and Ruths, ICWSM 2013

GPS-tagged tweets are written more often by young people and by women.

Confounds and consequences in geotagged twitter data, Pavalanathan and Eisenstein, EMNLP 2015

Twitter API
Twitter Streaming API (1%) vs full access (Firehose). Identified issues with estimating top hashtags based on streaming API.

Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose, Morstatter et al., ICWSM 2013

Google books I

- July 2012 (Version 2) and July 2009 (Version 1)
- Over 8 million books
- English, Chinese (simplified), French, German, Hebrew, Russian, Spanish, Italian

```
circumvallate 1978 335 91
circumvallate 1979 261 91
```

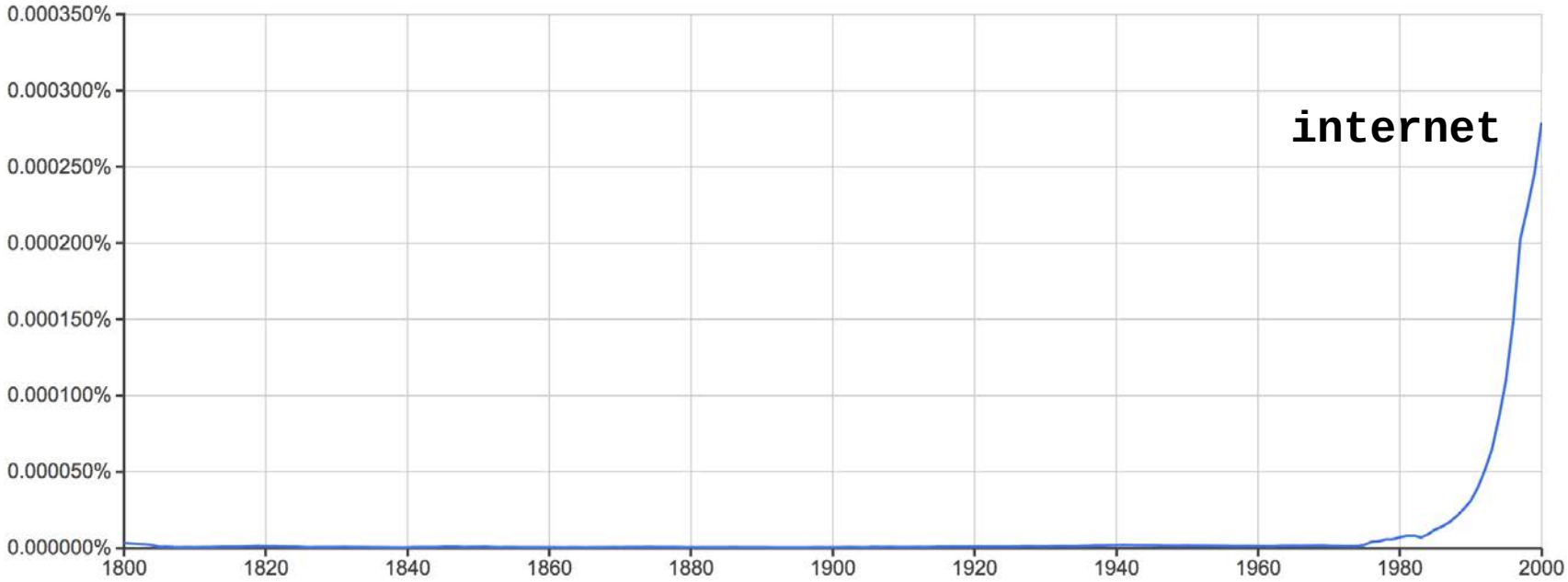
Syntactic Annotations for the Google Books Ngram Corpus, Lin et al., ACL 2012

Michel et al., Quantitative analysis of culture using millions of digitized books. Science., 2011

Web interface: https://books.google.com/ngrams

Download data:
http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

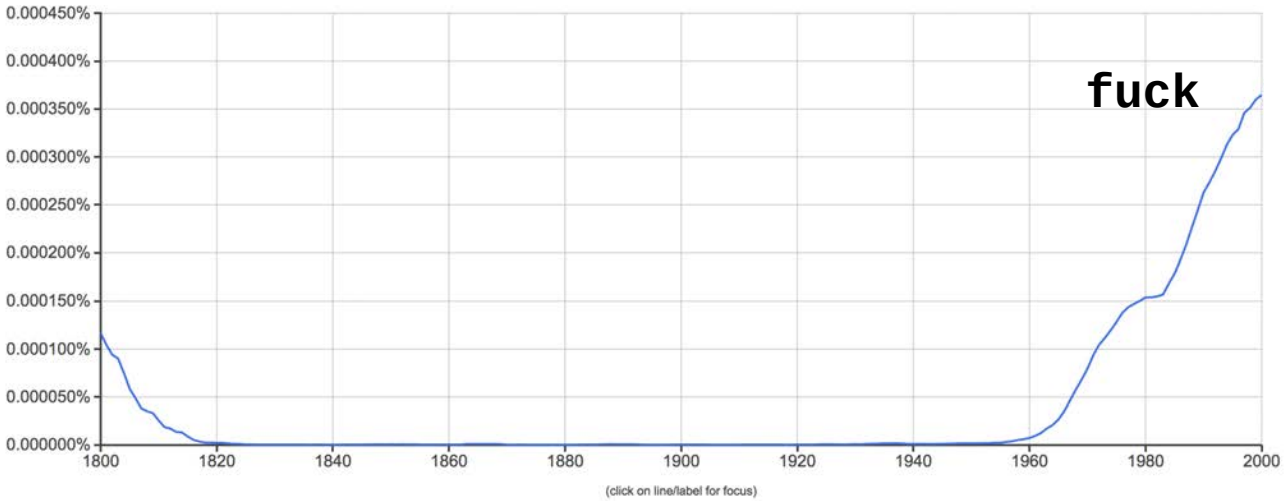
Google books I



Google books II

OCR errors... ☹

lowercase long *s* confused with *f*



(Wikipedia)

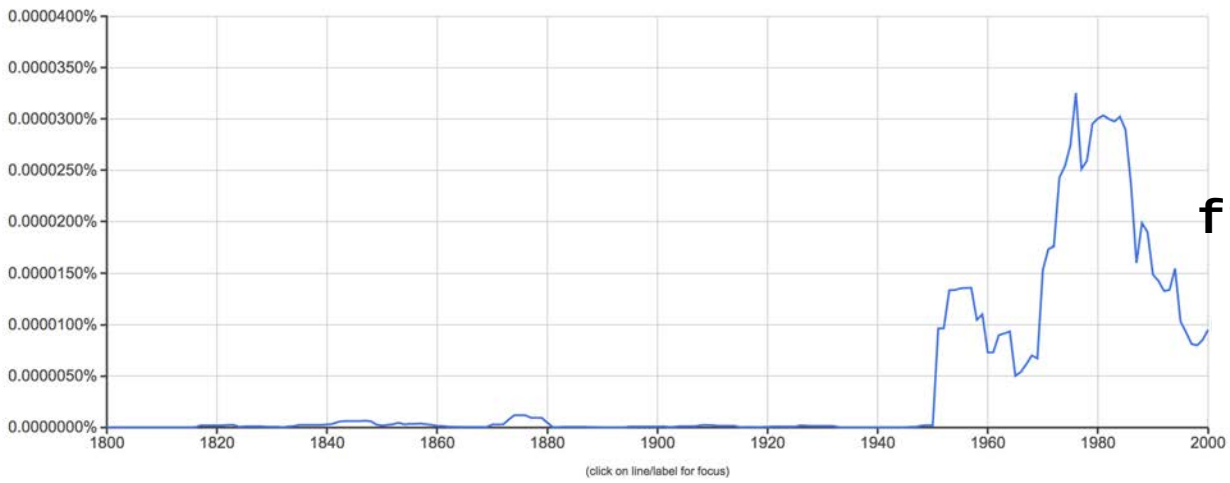
https://www.wired.com/2015/10/pitfalls-of-studying-language-with-google-ngram

Discussions: http://languagelog.ldc.upenn.edu/nll/?p=2847

Google books III

" [...] is a reflection of a library in which only one of each book is available"

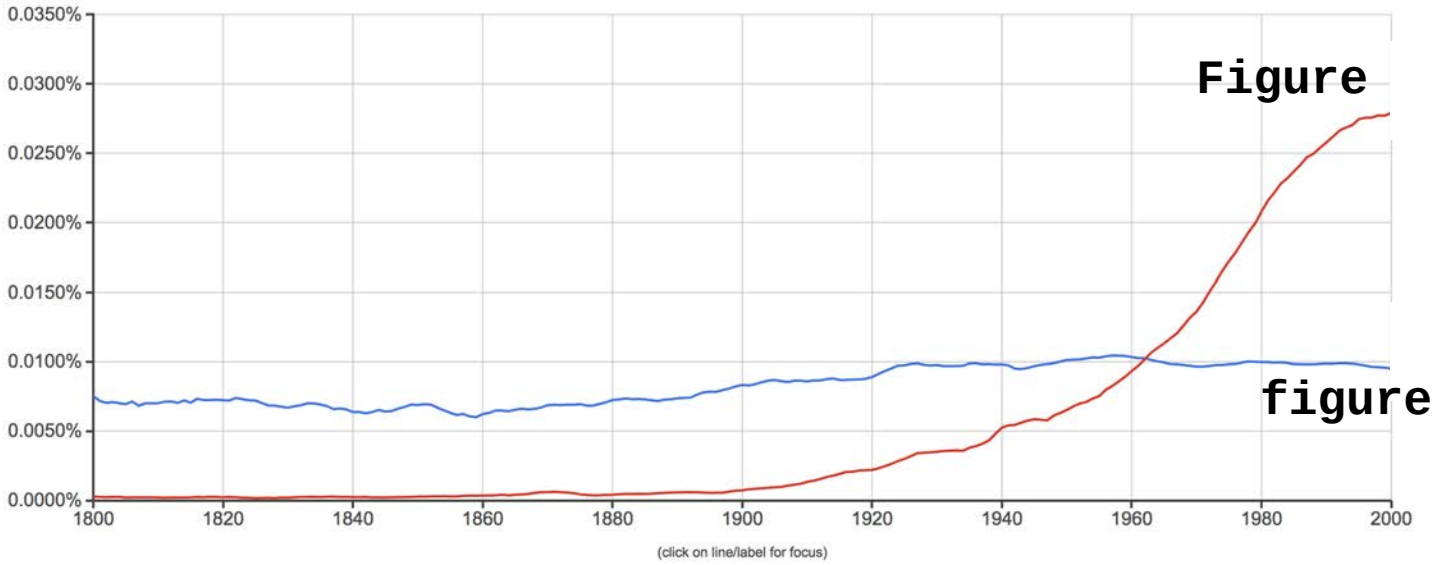
"[...] conveys an illusion of large-scale cultural popularity."



Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution, Pechenick et al. PLOS ONE 2015

Google books IV

Impact of inclusion of scientific texts



Pechenick et al. recommend: second version of the English Fiction data set

Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution, Pechenick et al. PLOS ONE 2015

Summary

- Possible biases are introduced in the complete research pipeline
- Impact of bias depends on your research questions
- Be aware and report (potential) biases in your data!
- Even better: compare across datasets, correct for demographic bias (e.g., Wang et al. 2015, Zagheni and Weber 2015)

References

- Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. Olteanu et al. http://www.aolteanu.com/SocialDataLimitsTutorial/
- Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution, Pechenick et al. PLOS ONE 2015
- Classifying Political Orientation on Twitter: It's Not Easy!, Cohen and Ruths, ICWSM 2013
- Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose, Morstatter et al., ICWSM 2013
- Confounds and consequences in geotagged twitter data, Pavalanathan and Jacob Eisenstein, EMNLP 2015
- How Noisy Social Media Text, How Diffrnt Social Media Sources?, Baldwin et al., IJCNLP 2013
- Dude, srsly?: The Surprisingly Formal Nature of Twitter's Language, Hu et al., ICWSM 2013
- Big questions for social media big data: Representativeness, validity and other methodological pitfalls, Tufekci, ICWSM 2014
- "How Old Do You Think I Am?": A Study of Language and Age in Twitter, Nguyen et al., ICWSM 2013
- Effects of Age and Gender on Blogging, Schler et al., Proceedings of AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs 2006
- Demographic research with non-representative internet data, Zagheni and Weber, International Journal of Manpower, 36(1):13–25, 2015.
- Forecasting elections with nonrepresentative polls, Wang et al., International Journal of Forecasting, 31(3):980 – 991, 2015

Small vs. big data

Making big data small again

"Big Data allows us to produce summaries of human behavior at a scale never before possible. But in the push to produce these summaries, we risk losing sight of a secondary but equally important advantage of Big Data—the plentiful representation of minorities. ***Women, minorities and statistical outliers have historically been omitted from the scientific record, with problematic consequences. Big Data affords the opportunity to remedy those omissions."***

On minorities and outliers: The case for making Big Data small, Foucault Welles, Big Data & Society, 2014

NLP for small data

- Semi-supervised learning
- Domain adaptation
- Incorporate background knowledge
- ...

ICML 2016 Workshop on Data-Efficient Machine Learning, https://sites.google.com/site/dataefficientml/

Supporting small data analysis

- Retrieval systems to support social scientists
- Computational approaches to identify 'interesting cases' for closer reading and coding

```
oetverkocht ('sold out')

('oet': Limburgish, 'verkocht': Dutch)
```

Nguyen & Cornips, 2016

Code-switching within words: <<0.1%

References

- Small data in the era of big data, Kitchin and Lauriault, GeoJournal (2015)
- On minorities and outliers: The case for making Big Data small, Foucault Welles, Big Data & Society 2014

Ethical challenges

Privacy I

The editorial policy followed in citing CMC data in this volume makes a distinction between restricted- and open-access electronic fora, the former of which are considered private, while the latter are public. (Herring, 1996)

we are confronted with media texts that combine private and public aspects on various levels. They may be public in the sense that they are within the public space and can be read by a large and anonymous audience, while at the same time discussing topics which we think of as 'private' and using language which is associated with informal and private conversations. (Landert and Jucker, 2011)

Privacy II

- Removal of posts by user
- Discussing individual users and including their posts (text/images/location?) in research articles
- Sharing data

Data representativeness

Wall Street Journal articles from 1989 are a big part of the Penn Treebank.

Audience: *older*, *richer*, *men*, *well-educated*?

The social impact of natural language processing, Hovy and Spruit, ACL 2016

POS taggers: age groups

- Hovy and Søgaard (2015) compared the performance of two POS taggers on user reviews with known gender, age and location.
- The taggers were trained on the Wall Street Journal portion from the Penn Treebank.
- Significant performance difference: the taggers perform better on reviews written by older authors (>45 years vs <35 years).

Tagging Performance Correlates with Author Age, Hovy and Søgaard, 2015

POS taggers: AAVE

POS taggers perform significantly worse on African American Vernacular English (AAVE) tweets

	STANFORD	GATE	ARK
AAVE	61.4	**79.1**	77.5
non-AAVE	74.5	**83.3**	77.9
Δ(+,-)	13.1	4.2	0.4

Table 5: POS tagging accuracies (%)

Challenges of studying and processing dialects in social media, Jørgensen et al., 2015)

Language identification: AAE

	AAE	White-Aligned
langid.py	13.2%	7.6%
Twitter-1	8.4%	5.9%
Twitter-2	24.4%	17.6%

Proportion of tweets in AA- and white-aligned corpora classified as non-English by different classifiers

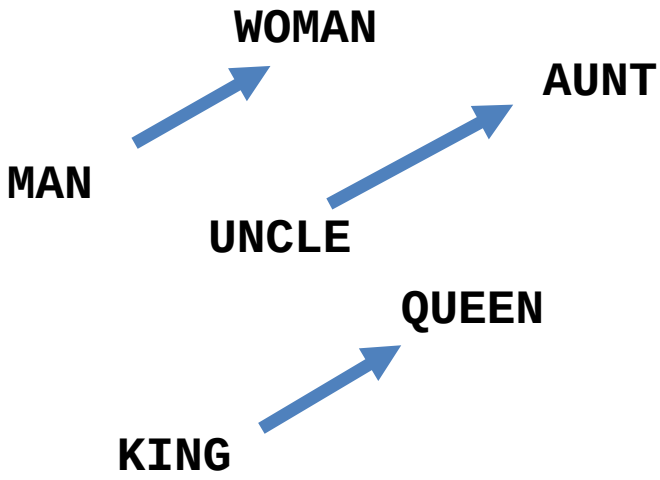
Demographic Dialectal Variation in Social Media: A Case Study of African-American English, Blodgett et al. EMNLP 2016

Perpetuation of bias: word embeddings I

Words are mapped onto a continuous vector space (word2vec, GloVe, etc.)

king – man + woman = queen

Word embeddings also capture gender relations



Linguistic Regularities in Continuous Space Word Representations, Mikolov et al. 2013

Perpetuation of bias: word embeddings II

Occupations closest to *she* and *he*

Extreme *she*	Extreme *he*
homemaker	maestro
nurse	skipper
receptionist	protege
librarian	philosopher
socialite	captain
hairdresser	architect
nanny	financier

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (Bolukbasi, et al. NIPS 2016)

Finder gender stereotype analogies

$$S_{(a,b)}(x,y) = \cos(\vec{a} - \vec{b}, \vec{x} - \vec{y}) \quad \text{if } \|\vec{x} - \vec{y}\| \leq \delta, \quad 0 \text{ else}$$

(a,b)=(*she*,*he*)

Gender stereotype *she-he* analogies
nurse-surgeon
sassy-snappy
cupcakes-pizzas
lovely-brilliant
vocalist-guitarist

Gender appropriate *she-he* analogies
queen-king
sister-brother
ovarian cancer-prostate cancer
mother-father
convent-monastery

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (Bolukbasi, et al. NIPS 2016)

Detecting bias: Word-Embedding Association Test

- The Implicit Association Test (IAT) based on response times and has been widely used.

- Word-Embedding Association Test (WEAT) by Caliskan et al: similarity between a pair of vectors (cosine similarity score) as analogous to reaction time in the IAT

https://en.wikipedia.org/wiki/Implicit-association_test

Semantics derived automatically from language corpora contain human-like biases, Caliskan, Bryson, Narayanan, Science 2017

Detecting bias: Word-Embedding Association Test

- The Implicit Association Test (IAT) based on response times and has been widely used.

Were able to replicate well-known IAT findings!

- Word-Embedding Association Test (WEAT) by Caliskan et al: similarity between a pair of vectors (cosine similarity score) as analogous to reaction time in the IAT

Semantics derived automatically from language corpora contain human-like biases, Caliskan, Bryson, Narayanan, Science 2017

Perpetuation of bias in sentiment analysis

"I had tried building an algorithm for sentiment analysis based on word embeddings [..]When I applied it to restaurant reviews, I found it was ranking Mexican restaurants lower. The reason was not reflected in the star ratings or actual text of the reviews. It's not that people don't like Mexican food. ***The reason was that the system had learned the word "Mexican" from reading the Web."***

https://blog.conceptnet.io/2017/04/24/conceptnet-numberbatch-17-04-better-less-stereotyped-word-vectors/

Reinforcing stereotypes? gender in NLP I

- Various datasets: Twitter (Rao et al., 2010; Bamman et al., 2014; Fink et al., 2012; Bergsma and Van Durme, 2013; Burger et al., 2011), blogs (Mukherjee and Liu, 2010; Schler et al., 2005), telephone conversations (Garera and Yarowsky, 2009), YouTube (Filippova, 2012), etc.

- Features
  - Females: more pronouns, emoticons, emotion words.
  - Males: more numbers, technology words, and links.

- Accuracy on Twitter users: Bergsma and Van Durme (2013) report an accuracy of 87%, Bamman et al. (2014) 88%.

Reinforcing stereotypes? gender in NLP II

Current supervised machine learning models learn stereotypical models

predictive words for females in (Dutch) tweets:

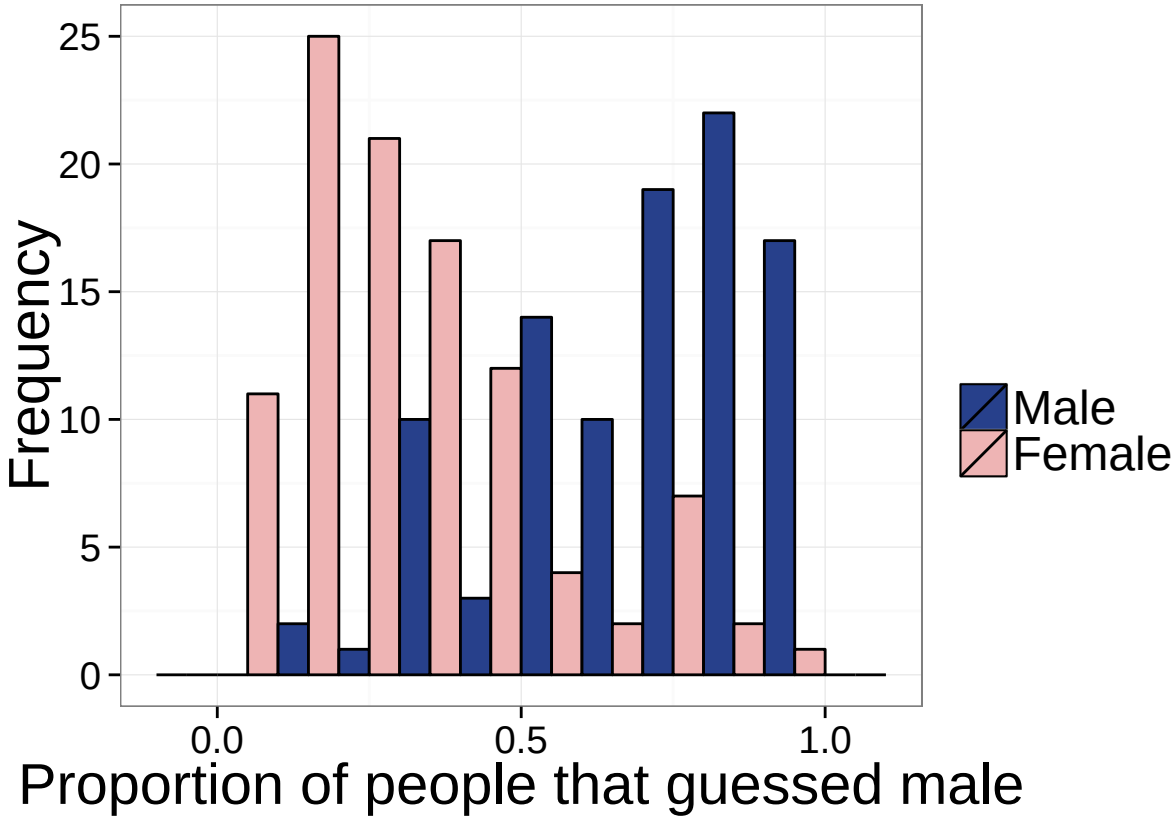
- *my man*
- *bye*
- *omg*
- *mom*
- *sweet*
- *girlfriends*
- *xx*
- *nails*

predictive words for males in (Dutch) tweets:

- *man*
- *bro*
- *[name of soccer team]*
- *fifa*
- *beer*
- *nice*
- *my woman*
- *game*

Reinforcing stereotypes? gender in NLP III

For example,
25 female users
where
10 - 20% of
the players guessed
they were male.



Based on 42K guesses, Nguyen et al. 2014

Reinforcing stereotypes? gender in NLP IV

- Automatic gender predictions on YouTube data correlated more strongly with the dominant gender in a user's network than the user-reported gender (Filippova 2012).
- Incorrectly labeled Twitter users had fewer same-gender connections in experiments by Bamman et al. 2014
- Clusters of Twitters users who used linguistic markers that conflicted with population-level findings (Bamman et al. 2014)

Reinforcing stereotypes? gender in NLP V

- Current supervised machine learning models learn stereotypical models.
- Not effective for users who don't fit gender stereotypical behavior
- Need to be careful with reporting findings.. could reinforce stereotypes

References: critical look on gender in NLP

- Gender as a variable in natural-language processing: ethical considerations. Larson, Proceedings of the First Workshop on Ethics in Natural Language Processing, 2017.
- Gender identity and lexical variation in social media. Bamman et al., Journal of Sociolinguistics, 2014.
- These are not the stereotypes you are looking for: bias and fairness in authorial gender attribution. Koolen and van Cranenburgh, Proceedings of the First Workshop on Ethics in Natural Language Processing, 2017.
- User demographics and language in an implicit social network. Filippova, EMNLP-CoNLL 2012.
- Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. Nguyen et al., COLING 2014.

Conclusion

- Ethical issues arise in the complete research pipeline!
  - Data collection
  - Data processing
  - Analysis
  - Applications
  - Reporting

Conclusion

Conclusion

- New datasets and methods enable studying language and social behavior in a variety of situations on a very large scale.

- But... still many challenges!
  - Technical
  - Ethical
  - Methodological: bridging the gap between disciplines

Questions?

Dong Nguyen
@dongng
www.dongnguyen.nl