Sociolinguistics from data, on the back of an envelope

Leon Richard Anthony Strømberg-Derczynski
Leon Derczynski for short

– sociolinguistics time!

ITU Copenhagen

28 November 2019

IT UNIVERSITY OF COPENHAGEN

About me

- Faculty in Natural Language Processing at ITU CPH

- Computer scientist w. prior careers
- PhD from U. Sheffield (England’s biggest AI group)
- Prior positions at UCSD, Innopolis U, AU

- Research topics:
  - high-variation language, e.g. soc med, clinical
  - temporal and spatial semantics
  - “information extraction”
  - tools for Danish
  - veracity

Data Science

- Turning data into knowledge
- Telling stories creatively
- Doing the right thing based on facts

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY SAVE SHARE 12 COMMENT TEXT SIZE PRINT $8.95 BUY COPIES

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place

Language as data

- Language becomes our scientific observation
- Don’t adjust the input!

- Canonicality is not well-defined for language
- In fact, we’re more interested in the “noise”

Language as data

- Computer science
  - Natural science:
  - Real world phenomena
  - Hypothesise, experiment, analyse
  - .. and theorise. There’s already a field for that!

Where can we get the best data?

Where can we get the best data?

Raw
Unadjusted
Spontaneous
Colloquial
Varied social settings

One candidate: user-generated text

e.g.: facebook, instagram, VK, linkedin, wechat, hestenettet, twitter, trustpilot, nationen, ..

No editor

Comes direct from end user

NYSE
NYSE
NYSE
NYSE
NYSE
NYSE
NYSE
CYPRESS

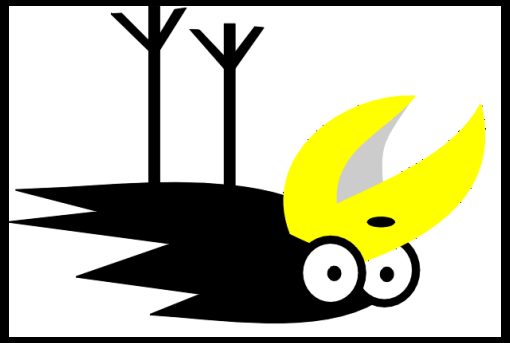
NYSE
NYSE
NYSE
NYSE
NYSE
NYSE
NYSE
CYPRESS

NYSE
CYPRESS

NYSE
NYSE
NYSE
NYSE
NYSE
NYSE
NYSE
CYPRESS

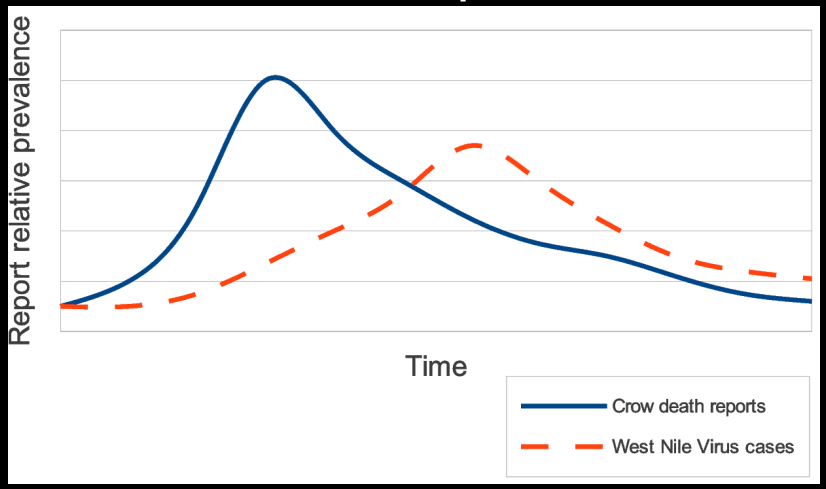
Utility in trivia

Retrospective analyses into cause and effect



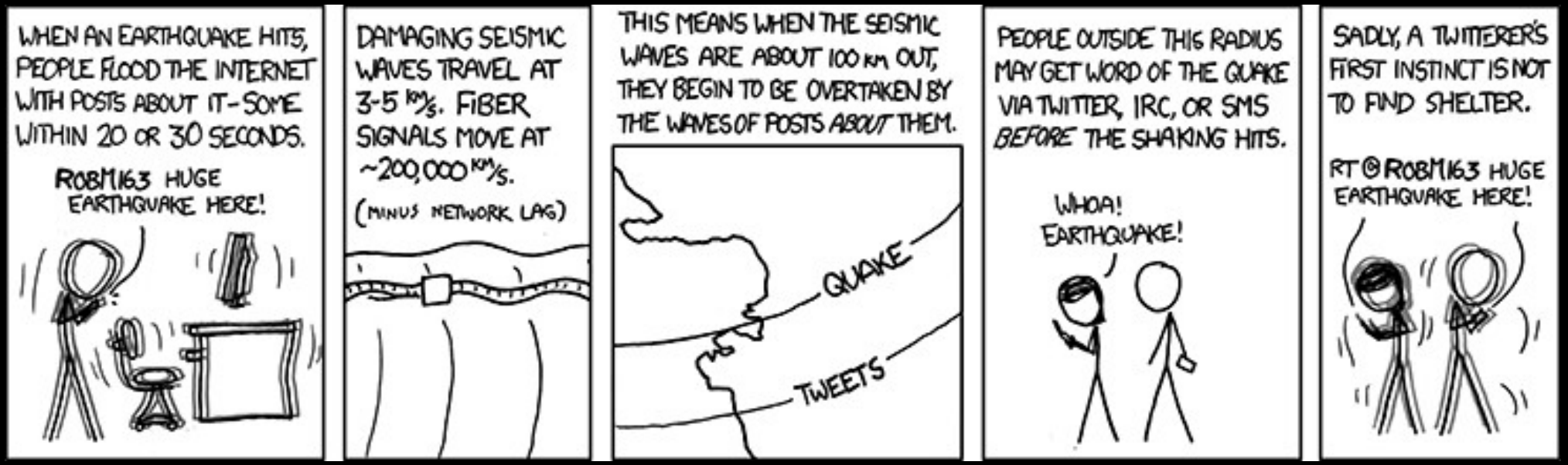
"There's a dead crow in my garden"

Social media mentions of dead crows predict WNV in humans [1]



1. Sugumaran & Voss 2012: "Real-time spatio-temporal analysis of West Nile Virus using Twitter Data", *Proc. Int'l conference on Computing for Geospatial Research and Applications*

Kind of a silly idea



Light travels faster than earthquakes

Detect earthquake images and discussion

Build a prediction

Send the warning

1.5 – 2 seconds extra time

Now used by USA, Japan authorities to send alarms, SMS alerts

But this is really difficult to process!

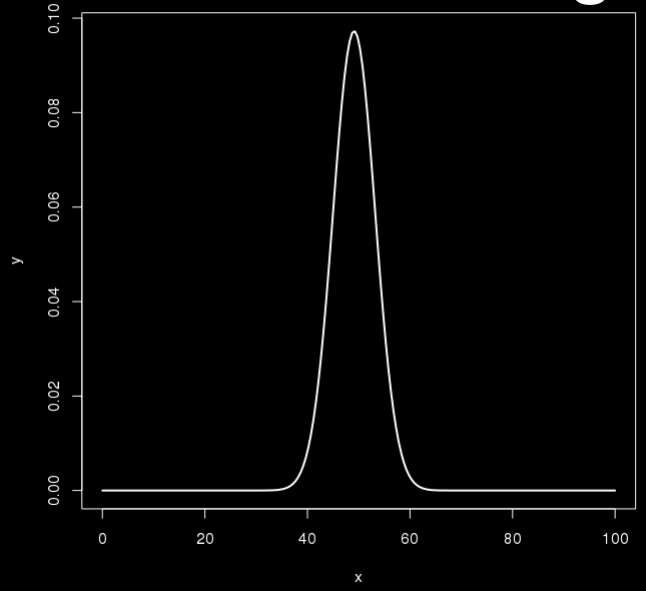
Why?

Why this particular social media source?

Twitter as D. Melanogaster of social media (Tufekci ‘14)

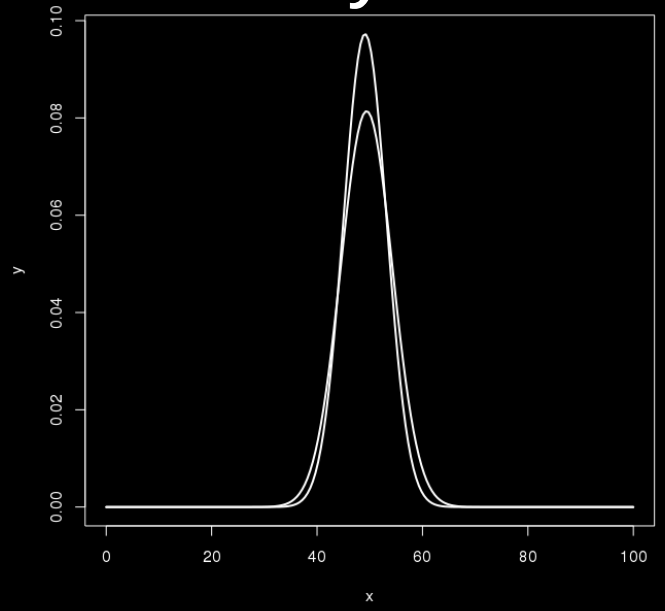
Huge variety, hard to process

Users intelligent, literate – individual styles



Spelling ability in net slang users

Overlaid with slang users



social media text is **surprisingly** formal

they see me rollin

- a typo?

they see me rollin
they hatin

- perhaps not. G-dropping mapped from speech

they see me rollin
they hatin
patrollin

they see me rollin
they hatin
patrollin
tryna catch me ridin dirty

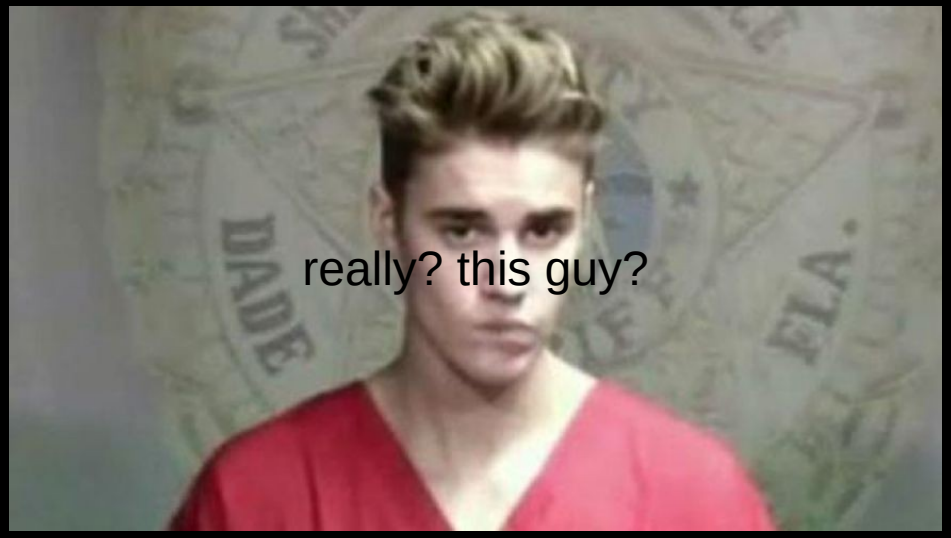
- flawless; not a single mistake

omb **x**

- surely they mean “omg”*?*

omb ✔

- the keys are like, right next to each other



Age?
Gender?
Languages spoken at home?

What kind of variance do we see?

Temporal:

- concept drift over time
- daily cycles (work, family, socialising)
- weekly cycles (weekend, hump day)
- time of year (seasonal behaviours)

Spatial

- many different anglophone regions (Botswana, Singapore,..)
- different surface forms in each (jam jelly jell-o)
- different signifiers (LLC – Ltd. - DAC)

Social

- WSJ readers' and writers' interests
- net celebrities (pewdiepie, LeVar Burton)
- tv characters (Dænærys, McNulty)

Tweets are *dirty*

- You all know what Twitter is, so let's just look at some difficult tweets
- Orthography: Kk its 22:48 friday nyt :D really tired so imma go to sleep :) good nyt x god bles xxxxx
- Fragments: Bonfire tonite. All are welcome, joe included
- Capitalisation: Don't Have Time To Stop In??? Then, Check Out Our Quick Full Service Drive Thru Window :)
- Nonverbal acts: RT @Huddy85: @Mz_Twilightxxx *kisses your ass**sneezes after* Lol

Word order still matters.. just

- Hard for tweets: exclamations and fragments
- Whole sentences a bit rare
- @FeeninforPretty making something to eat, aint ate all day
- Peace green tea time‼ Happyzone‼‼ :)))))

- Sentence structure cues (e.g. capitals) often:
  - missing
  - too many

Qualitative description

- Great diversity in social media users, but they're not illiterate
  - People want to represent their own dialects and accents (Jones 2010)
  - They pick and choose from the entire stylistic repertoire of language (Tagliamonte 2008)
  - Same literacy scores in standard and non-standard vocabulary users (Drouin 2009)
- Emoticons have more than just an expressive function
  - Pragmatic function, e.g. demonstrating a less stressed stance (Dresner 2010)
  - Not just pictograms: phrasal abbreviations are also included – smh, lol
  - Lexical items are made nonstandard through lengthening – cooolll (Brody 2011)

Qualitative description

- Social variables associated with certain transformations
  - Slang is less inhibited in informal settings (Labov 1972)
  - G-dropping mapped from speech to writing (Eisenstein 2010)
  - Lexemes can have a spatial association within a language (Eisenstein 2011)
- This socio-linguistic variation in social media highlights bias in existing resources
  - Most corpora text was curated predominantly by working-age white men (Eisenstein 2013)
  - Social media is not curated, so has different biases
  - We have little data that is free from this demographic bias

Quantitative description

- General style
  - Twitter is more conservative and formal, less conversational than SMS and online chat;
  - Brief, with careful word choice, dense in lexical words (Halliday 2004);
  - Tweets used to share news or broadcast personal status
- Individual style:
  - Individualistic style, or address large audience? (Yates 1996)
  - Users develop linguistically unique styles
  - Both 1st and 3rd person pronouns are common
  - Intensifiers of a young audience - “really” vs. “very” (Ito 2003).

Usage of names is confusing for systems,
but tells us about social relations

derek

x *is a person*

miles

x *might be a person*

Marie Claire

x *should not be a person*

Exodus Porter

x *probably an OK person, but actually a beer*

Spicy Pickle Jr.

x *apparently actually a person*

Spicy Pickle Jr.

x *apparently actually a person*

???

Understanding the text

Annotator recall

Pretty serious problem

Individuals have limited knowledge, limited world experience

Expert annotators actually not good – we're desperately narrow

Don't believe me? Who can explain this real document?

KKTNY in 45 min!!!!!

KOURTNEY
& KIM
TAKE
NEW YORK

Annotation challenges

Annotator recall

Pretty serious problem

People have limited knowledge, limited world experience

Expert annotators actually not good – we're desperately overfit

Don't believe me? Who can explain this real document?

KKTNY in 45 min!!!!!

Solution:

Don't try to make people agree

Just accept everything you get: it's the observation

Rare knowledge ≠ Wrong knowledge

Variation in individual knowledge

Group	Recall over final annotations	Agreement (F1)
Expert	0.309	0.835
Crowd	0.837	0.350

Experts are self-consistent, but don’t get far

Crowd is varied and inconsistent, but gets superior recall performance

What does the variation tell us?

Do you feel luccy, punk?

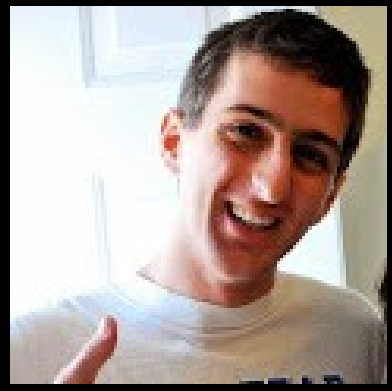
Do you feel luccy, punk?

Male?
US?
12-30?
Low education level?
Lower income family? - not reliable!
Ethnicity? - even less reliable

“Now we stronger than ever”

- Let's look at AAVE
  - “African American vernacular English”
- Hyp 1:
  - Ethnicity represented in syntax
- Hyp 2:
  - Women use more style in their language

- We already have a dataset…



Ian Stewart, EACL 2014

“Now we stronger than ever”

- AAVE already affects our toolchain
- Twitter word splitter (CMU)
  - Slang is hard to split, e.g. “ima” - “I am going to”
- Part-of-speech tagging
  - Nonstandard tense-aspect-mood not in tagsets
    - “finna”
  - Need to detect 3rd person sing. “s” drop
    - “oh man she runs”
    - “oh man she run”

“Now we stronger than ever”

- Map language with zip code
- AAVE more common in cities
  - Standard American English outside cities
- Language predicts race well
  - You need to look at many variations
- Women use slightly more stylised language
  - Though not with everything

AAVE in Canonical context

- False positives:
  - Copula deletion in Singapore English
  - Quoted text, e.g. rap lyrics – is this a persona?
- “finna” presents all kinds of problems
  - Should be VP, not NN
  - Systems to handle non-standard morphemes don't exist

Using author information

- Demographics help!
  - Hovy 'ACL 14
  - Group word over many documents, then also split things by author
  - Split groups are better guessing attitude, finding mistakes, etc

- Hmm, interesting..

- Python notebook time!

User reviews for sociolinguistics

- Start with reviews from Trustpilot
  - Danish, Dutch, English, Finnish, French, German, Italian, Norwegian, Polish, Portuguese, Russian, Spanish, Swedish
- Users are labelled with location, age, gender

- Variation in language then
  - Data mined (χ^2 test)
  - Used to predict each of these three categories

Hovy et al. WWW 2015

User reviews for sociolinguistics

- Gender

	Denmark	France	Germany	UK	US
female	nemt	**satisfaite**	gerne	flowers	customer_service
	bestille	**contente**	qualität	really	**love**
	varer	suis	tolle	**pleased**	we
	tilfreds	produits	wir	thank_you	**pleased**
	rigtig	chaussures	ware	definitely	really
	handle	étiquettes	bestellen	lovely	petflow
	dejligt	sont	sind	**love**	free_shipping
	nem	ma	karten	quickly	thank_you
	hjemmeside	été	bestellung	received	our
	mine	spartoo	vielen_dank	impressed	food
male	lovet	**problème**	keys	wife	these_guys
	deres	rien	günstig	parts	parts_geek
	problemer	bonne	service	deal	fast
	alt	matériel	mmoga	product	partsgeek
	aftalt	**conforme**	support	00	prices
	virker	ldlc	weiter_so	an	best
	tiden	service	top	first_class	shipping
	ok	sans	key	tyres	customer
	pris	bon	guter	could_find	part
	00	satisfait	gerne_wieder	best	parts

User reviews for sociolinguistics

- Location

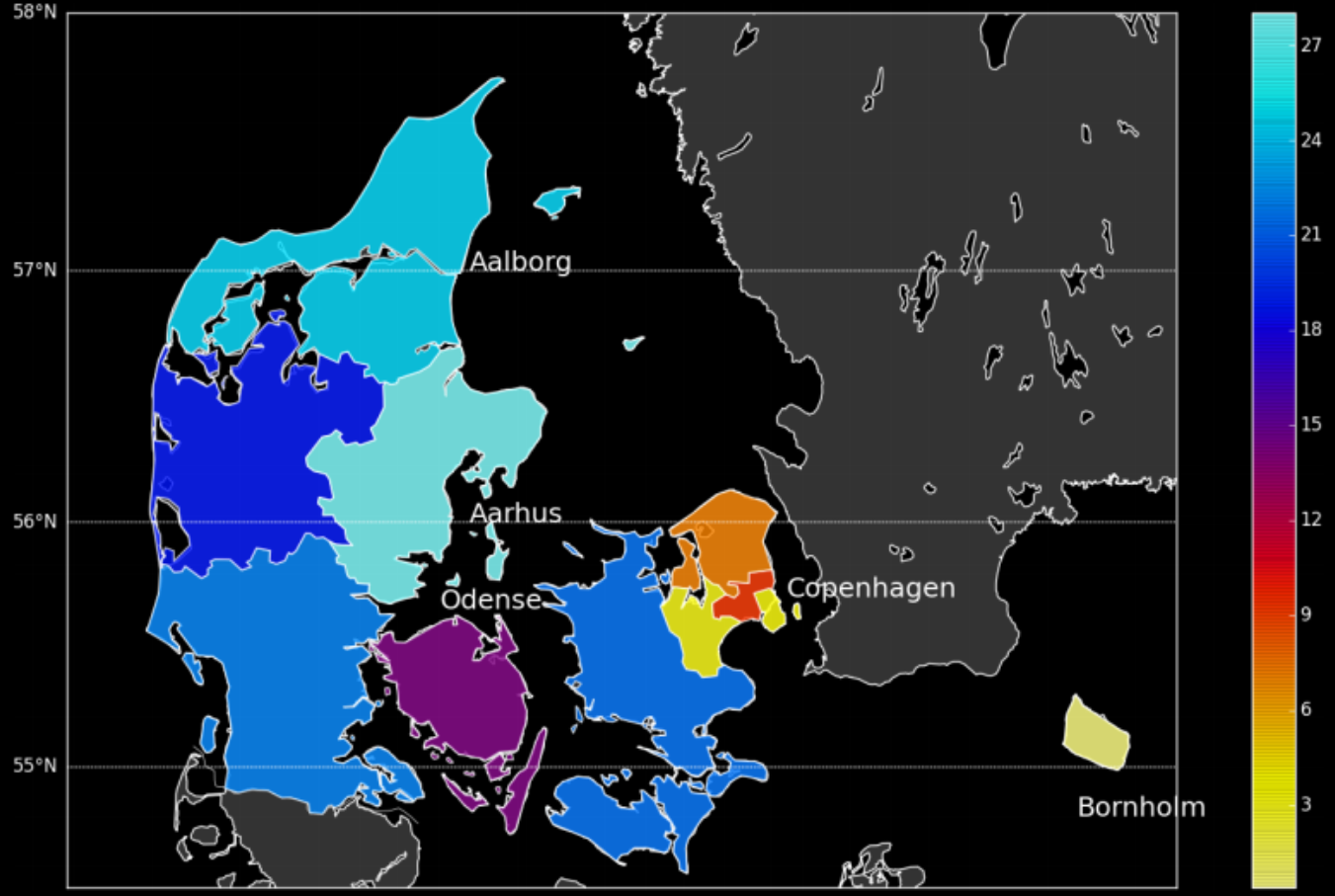


Figure 5: Regional distribution of "sin/sit" (ratio of all pronouns per NUTS-2 region)

User reviews for sociolinguistics

- Location

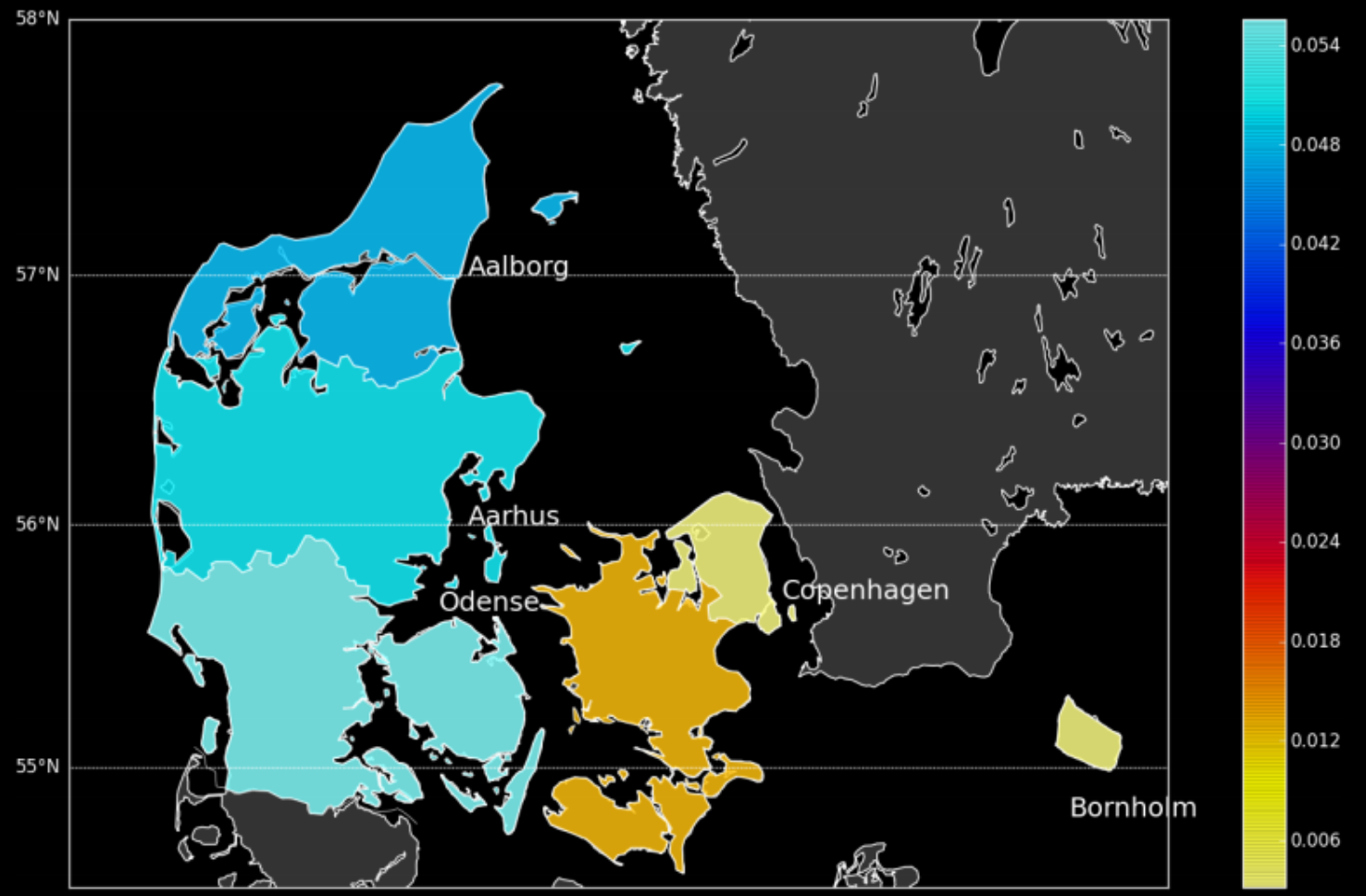


Figure 6: Regional distribution of the word *træls* (ratio of all words per NUTS-2 region).

User reviews for sociolinguistics

- Gender

- Females use emoticons twice as often
- Top product categories:
  - F: Pets, clothes/fashion, hotels, fashion accessories
  - M: Computers, car lights, hotels, fashion accessories
- Females *slightly* prefer noseless smileys
- Same spread of review scores

User reviews for sociolinguistics

- Age - Biggest indicator: NOSES

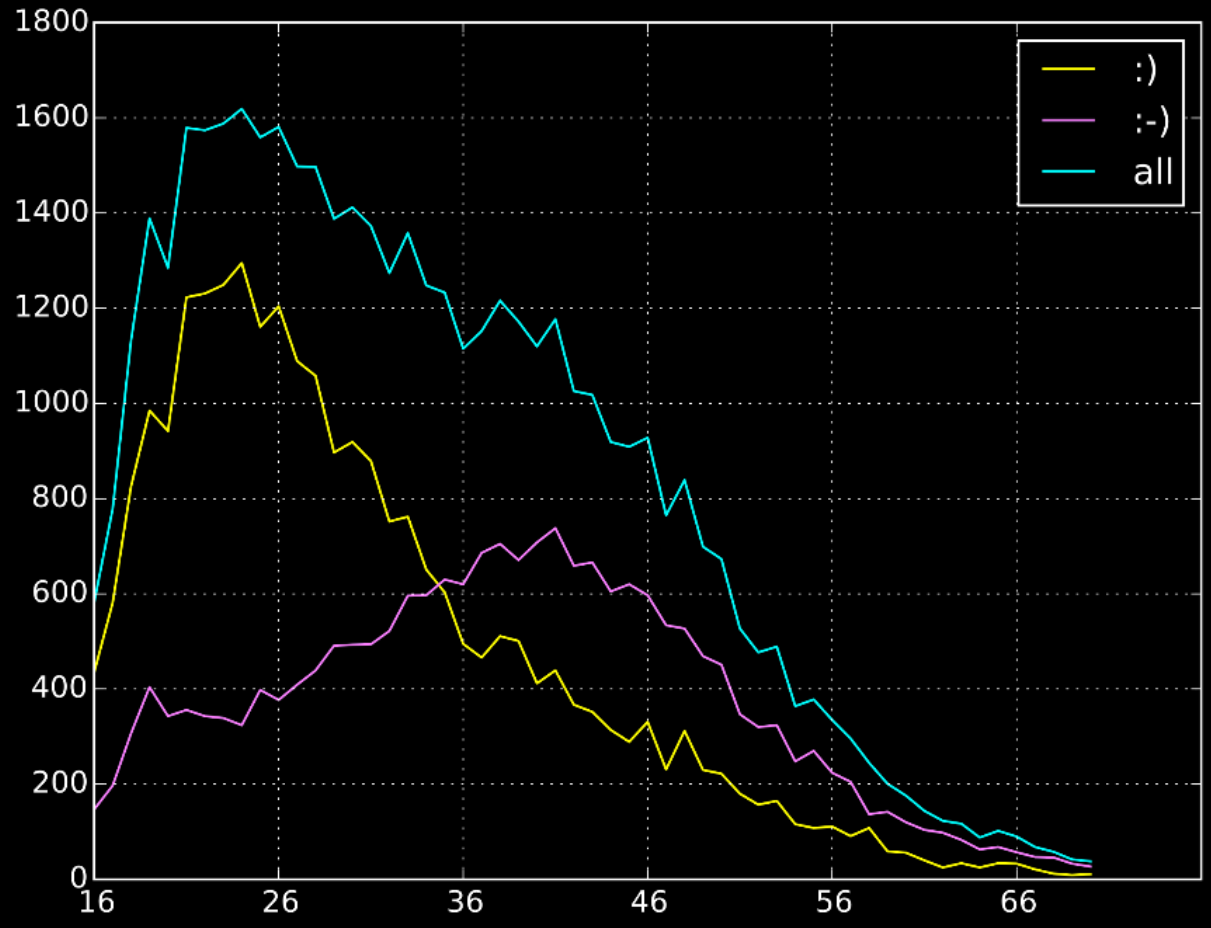


Figure 3: Usage of emoticons with and without nose by age group, aggregated over all countries

What can we say about this?

User reviews for sociolinguistics

- Age - Biggest indicator: NOSES

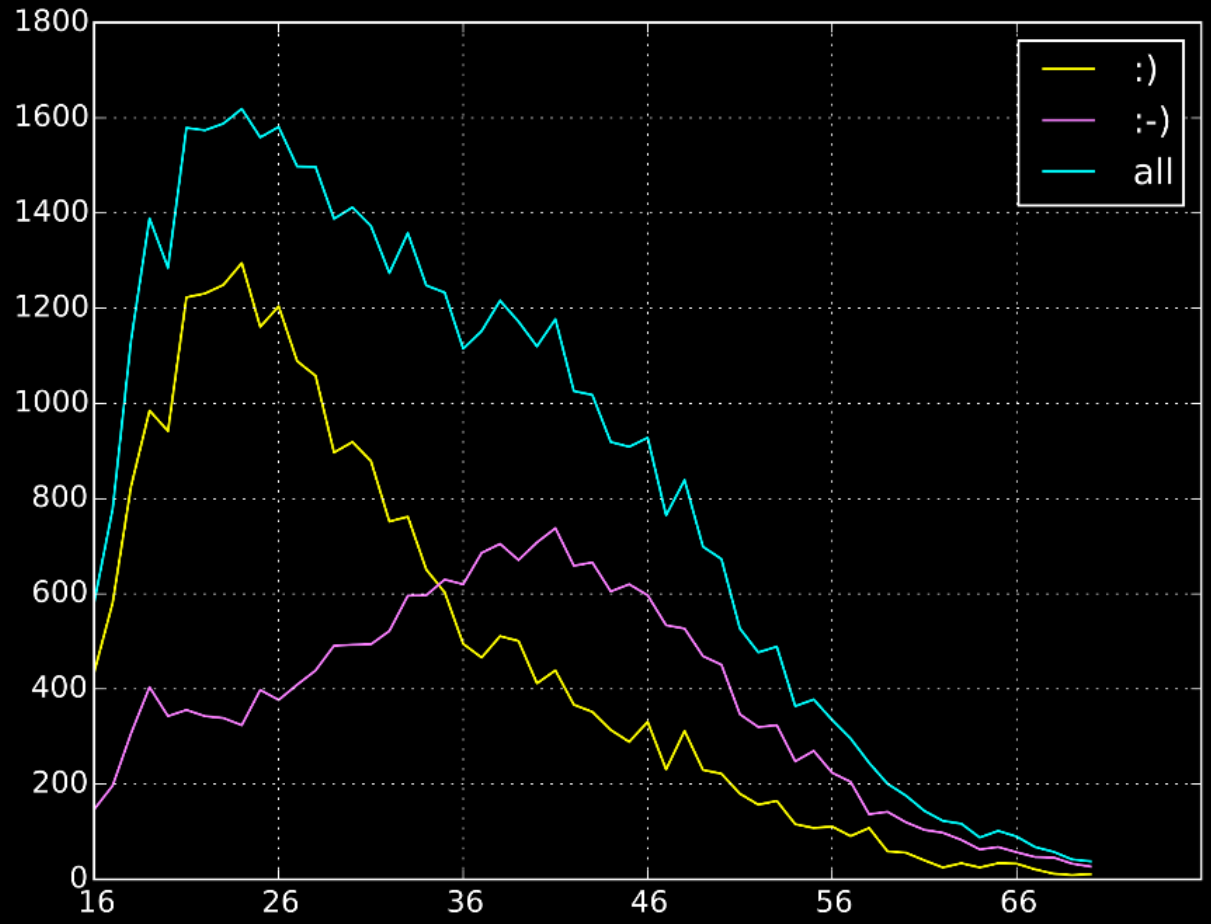


Figure 3: Usage of emoticons with and without nose by age group, aggregated over all countries

Use a nose?
Probably 35+

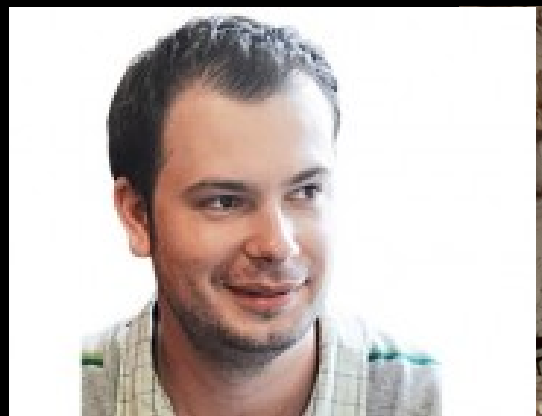
Nose in emojis = born before 1981

we’re losing this signal

– by @EmojiMashupFace

Social media predicts income

- Job *(Preotiuc, Lampos, Aletras ACL'15)*
- 9-class system: Management, professional, skilled trades, sales etc
- Data from Twitter:
  - Followers & friends
  - Hashtag, RT, link usage
  - Tweet rate
  - % in English

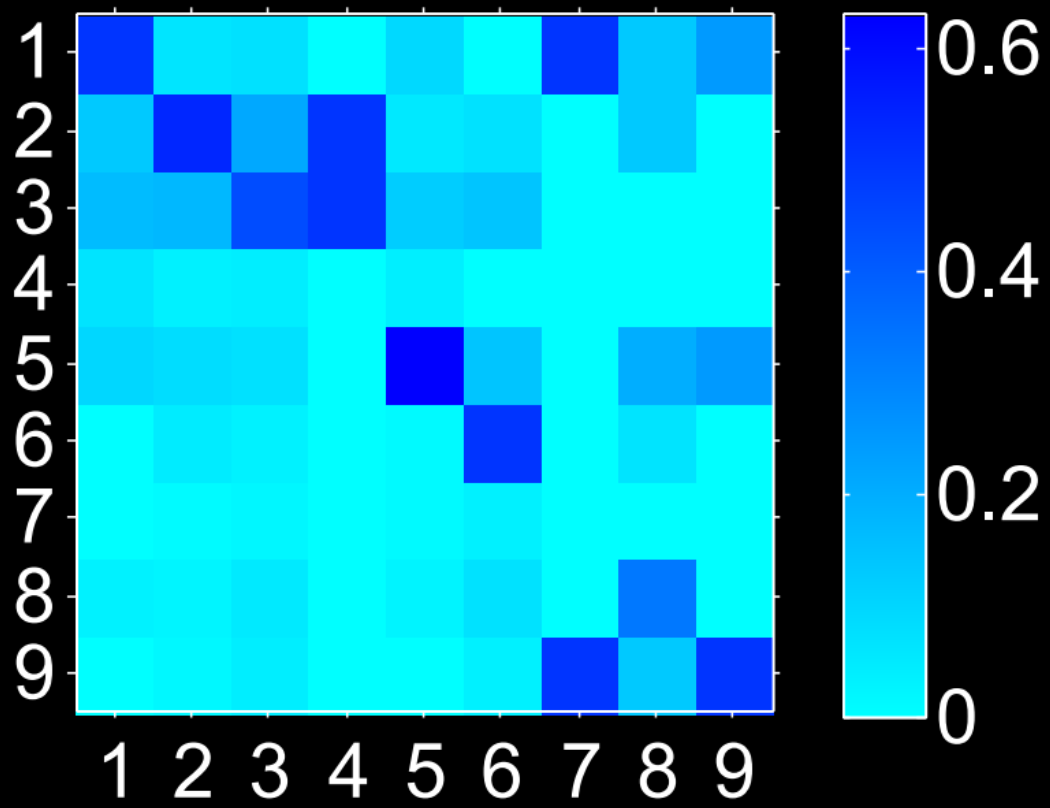


Social media predicts income

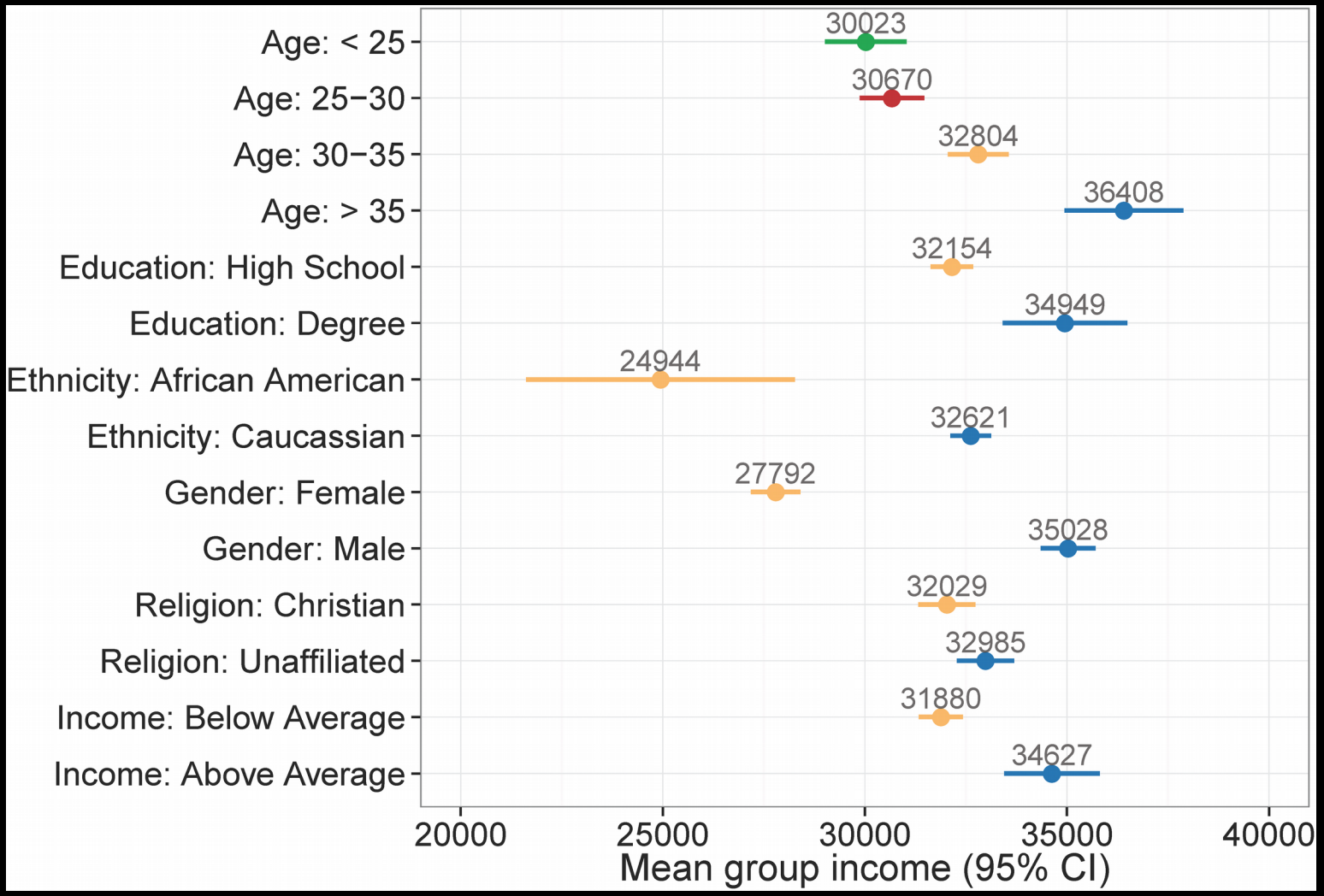
- W2V and SVD feature extraction

- Dark on the diagonal = we guessed correct

- Reasonable accuracy
- Enables mining of:
  - Related topics
  - Behaviour

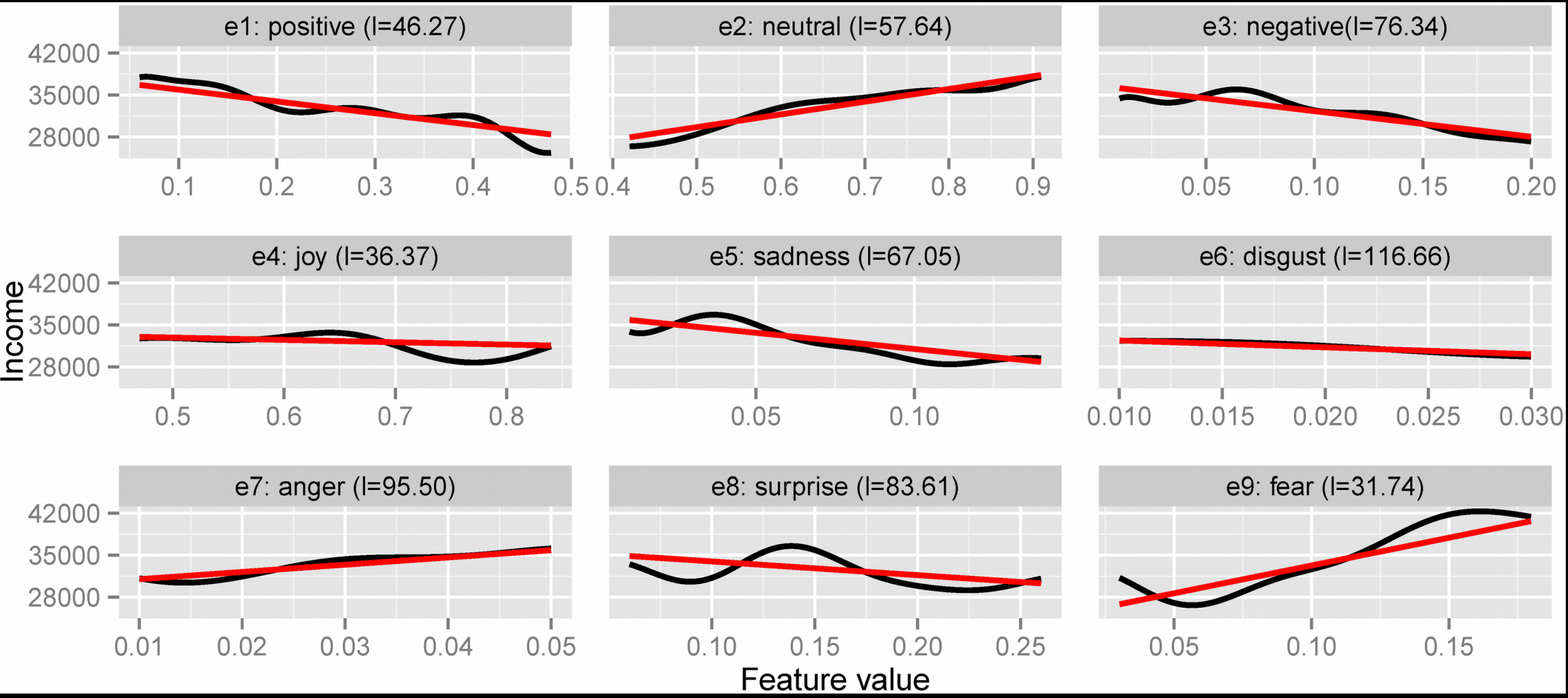


Social media predicts income



- Model interpretability: neural language models

Social media predicts income



- Expressions in language predict income

Social media predicts income

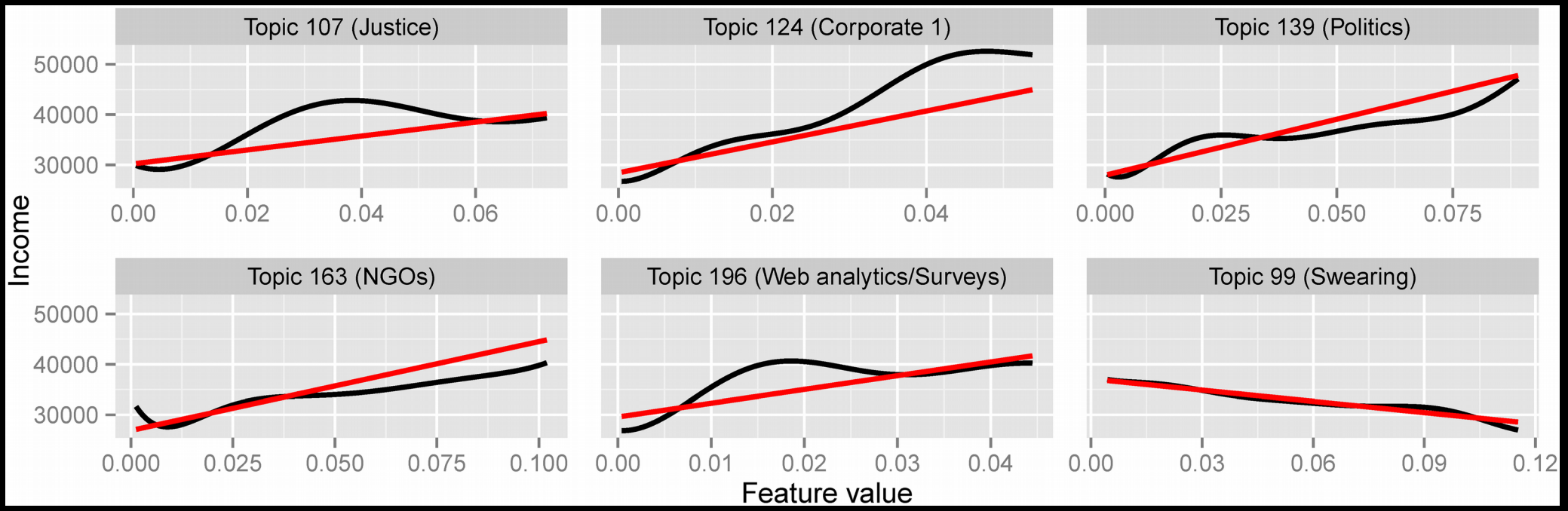
- What are people talking about? (most reliable first)

Rank	Topic #	Label	Topic	*I*
1	139	Politics	republican democratic gop congressional judiciary hearings abolishing oppose legislation governors congress constitutional lobbyists democrat republicans	3.10
2	163	NGOs	advocacy organization organizations advocates disadvantaged communities organisations participation outreach associations non-profit nonprofit orgs educators initiative	3.44
3	196	Web analytics / Surveys	#measure analytics #mrx #crowdsourcing crowdsourcing #socialmedia #analytics whitepaper #li metrics #roi startup #social #smm segmentation	3.68
5	124	Corporate 1	consortium institutional firm's acquisition enterprises subsidiary corp telecommunications infrastructure partnership compan aims telecom strategic mining	6.48
6	91	Corporate 2	considerations provides comprehensive cost-effective enhance advantages selecting utilizing resource essential additionally specialized benefits provide enhancing	7.44
7	107	Justice	allegations prosecution indictment alleged convicted allegation alleges accused charges extortion defendant investigated prosecutor sentencing unlawful	7.84
8	92	Link words	otherwise unless wouldn't whatever either maybe pretend anyone's assume eventually assuming or bother couldn't however	8.39
9	173	Beauty	hair comb bleached combed slicked hairs eyebrows ponytail trimmed curlers dye dyed curls waxed bangs	9.75
10	40	Sport shows	first-ever roundup sport's round-up rundown poised previewing spotlight thursday's com's long-running joins concludes prepares observer	10.57
11	99	Swearing	messed f'd picking effed cracking f*cked hooking tearing catching lighten picked cracks ganging warmed fudged	11.09

doi:10.1371/journal.pone.0138717.t004

Social media predicts income

- Topic vs. income



- One takeaway: avoid rich people if you want to capture creative uses of language?
- These can be valuable: “De kan rende og hønse”

Risks

- The more you can be profiled, the more you can be predicted
- The more you can be predicted, the more you can manipulated
  - E.g. Facebook's emotional manipulation experiments
- Huge and zero-consent loss of privacy
- People talk differently and that's OK
  - Should linguistic micro-profiling be legal?

Side effects of ignoring diversity of language

Getting lol right

- Lexicalisation of real laughter

  THAT'S A GOOD JOKE, CHUCK. LOL.

- As discourse marker

  oh man I was drunk last night

  lol

- As irony marker

  this job is so...idk just kill me lol

  climate change is killing us all and that's fine lol

Getting lol right

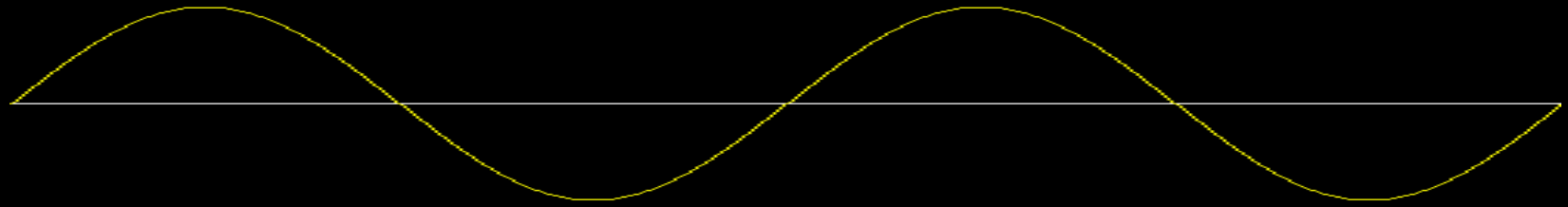
- Lexicalisation of real laughter
- As discourse marker
- As irony marker

- Consider:
  - You look good in red lol
  - I love you lol

Critical points in learning from data

- Systems are very bad at generalising beyond their data
  - It’s hard enough to generalise within it
- If a behaviour isn’t included in the training data, it will be rejected by systems
- Crucial to include broad variation
- Luckily, we know ML systems learn better and faster from diverse data → CO2 (and also money) savings
- But what if we <u>don’t</u> use diverse data?

news



social media

most of our language AI was trained on news text

the bias is:

- middle class
- white
-working age
- educated
- male
- 1980s/1990s
- from the US
- journalist
- following AP guidelines

your phone rewards you if you talk and write like

(ok.. sort of)

your phone rewards you if you talk and write like

(ok.. sort of)

.. and punishes you when you don't.

(not cool!)

Thank you for your time!

There is hope:

Jersey Shore is overrated. studying and history homework then a fat night of sleep!

Do you have any questions?