Is a minority dialect "noisy text"?: Social media NLP, analysis, and variation

Brendan O'Connor (http://brenocon.com/)
College of Information and Computer Sciences
University of Massachusetts Amherst

Workshop on Noisy User-Generated Text, July 31, 2015
https://noisy-text.github.io/

- Why analyze noisy user-generated text?
  It's where the data is

<u>To analyze:</u>

Social phenomena in social media datasets

1

- Political speech under Chinese censorship
- Sentiment and topics by social group
- Social determinants of language evolution

<u>How to analyze:</u>

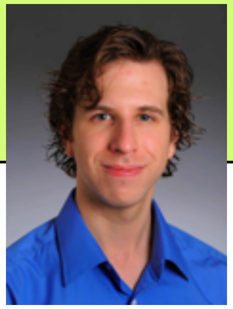
NLP capabilities we need to do these better

- Word segmentation
- Part of speech tagging
- Entities
- Syntactic, semantic parsing

2

1

Censorship and Deletion Practices in Chinese Social Media.
David Bamman, Brendan O'Connor, Noah Smith.
First Monday, 2012.



Chinese Internet Censorship

- Blocking information access
  - IP/DNS blocking (Facebook, Twitter, YouTube etc.)
  - Network filtering
  - Search engine results filtering

- Blocking content creation
  - This work
  - King, Pan and Roberts, 2013

  - “may be the most extensive effort to selectively censor human expression ever implemented”

sina新浪网
sina.com.cn

~~[message]~~

create

delete

Users

issues directives

Human censors
(~content moderators)

Propaganda Ministry

Our inference goal:
Characterize what types of content are targeted for deletion

Message Deletion

Download 56,951,585 realtime posted messages from Sina Weibo, over the period 2011/06/27 – 2011/09/30

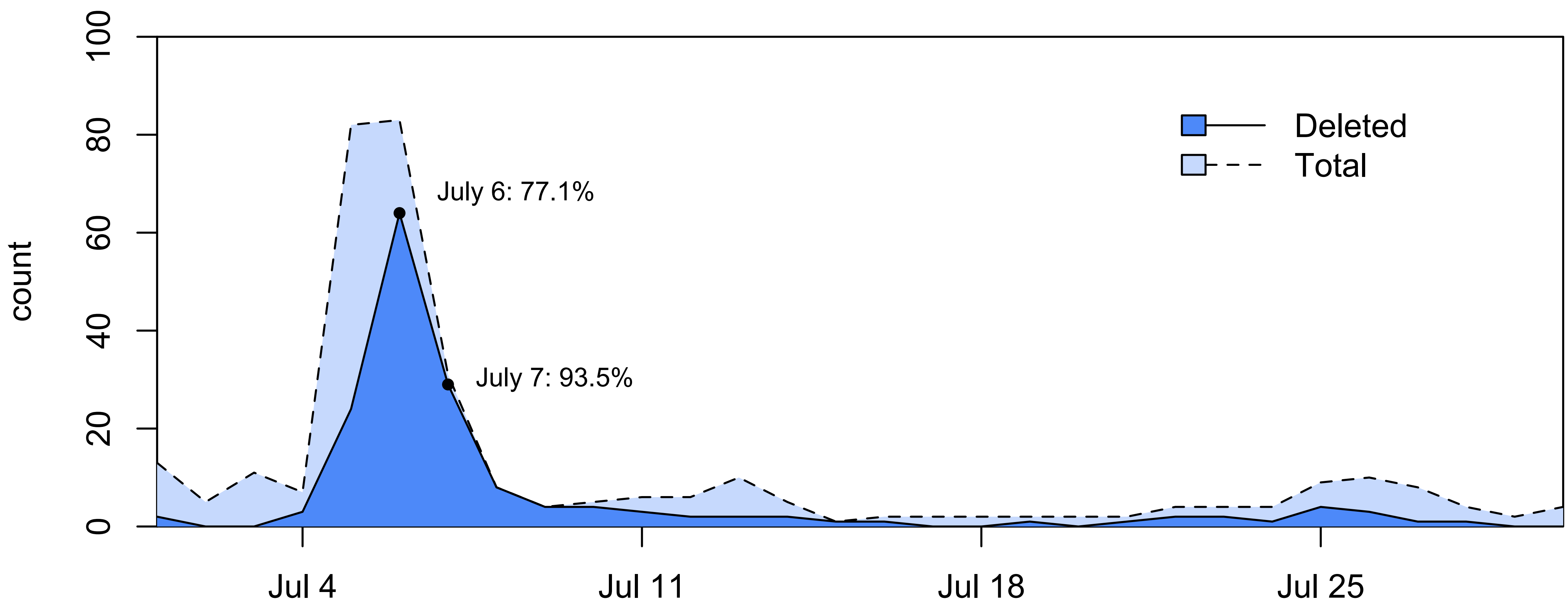
3 months after posting, check if deleted.

`“target weibo does not exist.”`

<u>Our inference goal:</u>
Characterize what types of content are targeted for deletion

Message Deletion

Messages containing "Jiang Zemin" (江泽民)

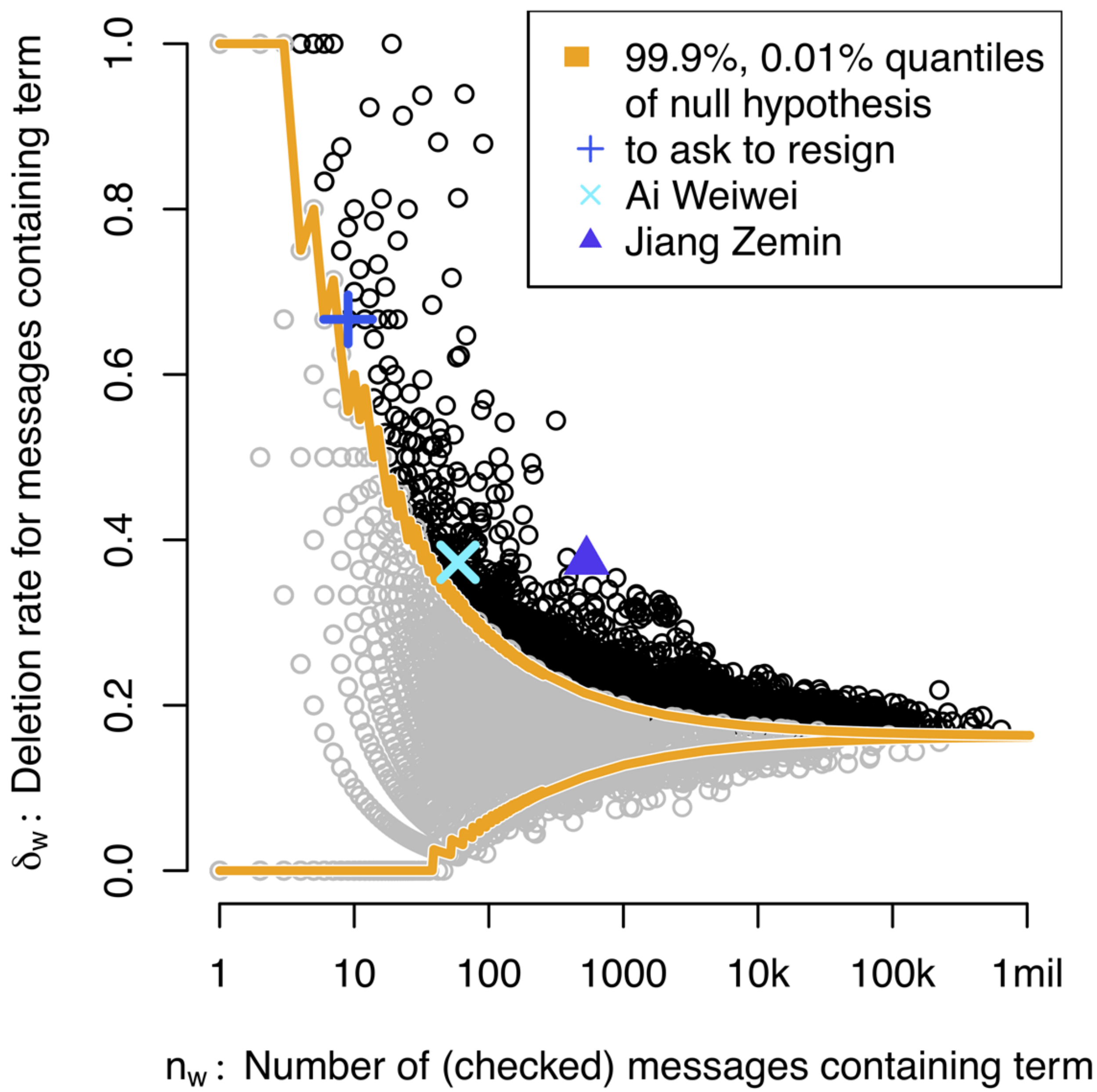


Message Deletion

- Message sample (1.6M) for deletion checks.
  Baseline deletion rate: 16.25%
- Social media word segmentation is hard NLP;
  instead use bilingual lexicon
  (CC-DICT + Wikipedia page titles)
- Find terms that are deleted with higher than expected rates.

Term Deletion

$$\delta_w \equiv P(\text{message becomes deleted} \mid \text{message contains term } w)$$



Multiple null hypothesis tests

False Discovery Rate
(Benjamini-Hochberg 1995, Efron 2010)

$$\mathrm{FDR}_{p_w<.001} < \frac{P_{null}(p_w < p)}{\hat{P}(p_w < p)} = \frac{0.001}{0.040} = 2.5\%$$

Highly deleted terms

- Spam (Movie titles etc.)
- Personal messages (Lantern festival, condolences)
- Known sensitive terms
  - 方滨兴 (Fang Binxing, architect of the GFW)
  - 法轮功 (Falun Gong, a banned spiritual group)
- Driven by current events
  - 请辞 (to ask someone to resign) -- Wenzhou train crash
  - 防核 (nuclear defense/protection), 碘盐 (iodized salt), and 放射性碘 (radioactive iodine) -- Fukushima
- Imperfect correspondence to GFW block status

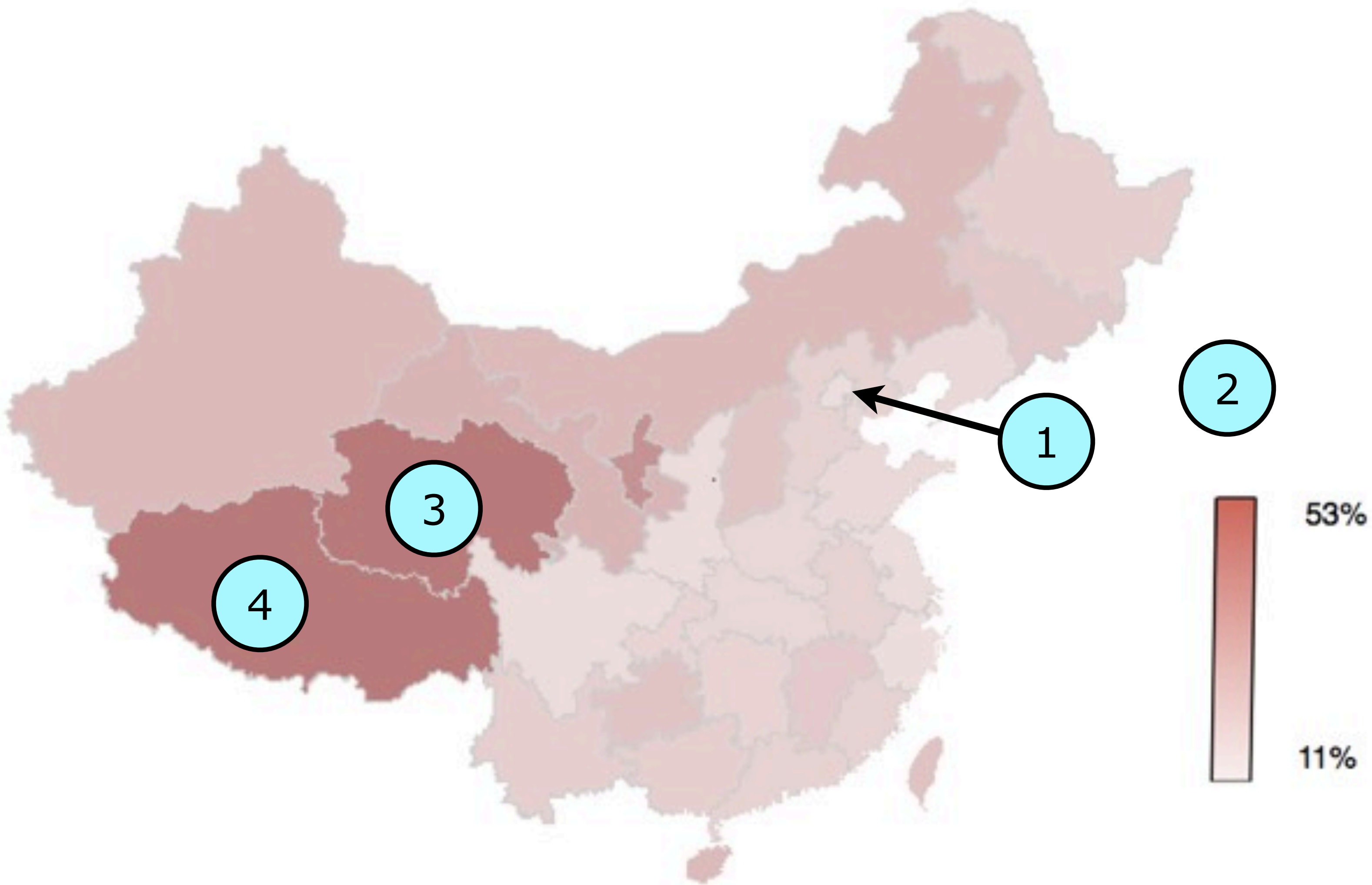
Geography

Deletion rate by region

Tibet	53.0	Hunan	16.4
Qinghai	52.1	Hubei	15.9
Ningxia	42.2	Outside China	15.5
Macau	32.1	Tianjin	15.2
Gansu	28.5	Henan	15.1
Xinjiang	27.0	Shandong	14.5
Hainan	26.5	Liaoning	14.1
Inner Mongolia	26.3	Jiangsu	13.9
Taiwan	23.9	Shaanxi	13.8
Guizhou	22.6	Sichuan	13.2
Shanxi	22.2	Zhejiang	12.9
Jilin	21.5	Beijing	12.0
Jiangxi	20.7	Shanghai	11.4
Other China	20.2		
Heilongjiang	18.3		
Guangxi	18.3		
Yunnan	18.2		
Hong Kong	17.8		
Hebei	17.3		
Guangdong	17.3		
Anhui	17.2		
Fujian	17.1		
Chongqing	16.8		

Geography

Words by region (PMI ranking)



1. Beijing: (1) 西直门 (Xizhimen neighborhood of Beijing); (2) 望京 (Wangjing neighborhood of Beijing); (3) 回京 (to return to the capital)
   ▹ (410) 钓鱼岛 (Senkaku/Diaoyu Islands)

2. Outside China: (1) 多伦多 (Toronto); (2) 墨尔本 (Melbourne); (3) 鬼佬 (foreigner [Cantonese])
   ▹ (632) 封锁 (to blockade/to seal off); (698) 人权 (human rights)

3. Qinghai: (1) 西宁 (Xining [capital of Qinghai]); (2) 专营 (special trade/monopoly); (3) 天谴 (divine retribution).
   ▹ (331) 独裁 (dictatorship); (803) 达赖喇嘛 (Dalai Lama)

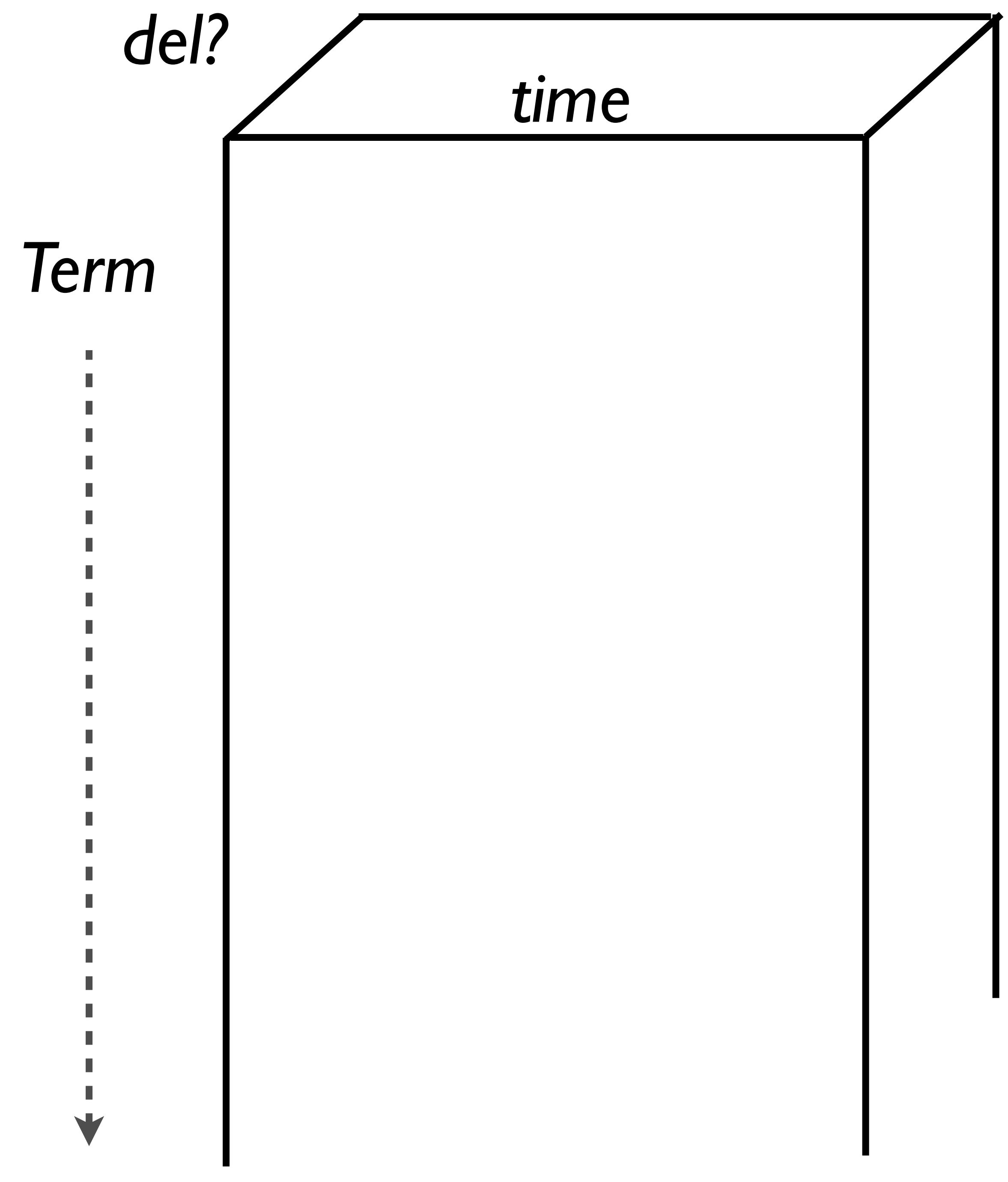
4. Tibet: (1) 拉萨 (Lhasa [capital of Tibet]); (2) 集中营 (concentration camp); (3) 贱格 (despicable)
   ▹ (50) 达赖喇嘛 (Dalai Lama); (108) 迫害 (to persecute)

- Analyzing text content sheds light on political science questions
- KPR 2013: government doesn't censor criticism, but rather collective action potential

NLP as social analysis tool

- All analyses on 4-dimensional message count table.
  - Term
  - Deleted?
  - Region
  - Time

NLP defines the content dimensions of analysis



- Ideally we'd want
  - Word segmentation
  - Topical clustering
  - Sentiment
  - Translation...

- Why analyze noisy user-generated text?
  It's where the data is

1

<u>To analyze:</u>

Social phenomena in social media datasets

- Political speech under Chinese censorship
- Sentiment and topics by social group
- Social determinants of language evolution

2

<u>How to analyze:</u>

NLP capabilities we need to do these better

- Word segmentation
- Part of speech tagging
- Entities
- Syntactic, semantic parsing

TweetMotif: Exploratory Search and Topic Summarization for Twitter.
Brendan O'Connor, Michel Krieger, and David Ahn.
ICWSM 2010.

Part-of-speech tagging for Twitter: Annotation, Features, and Experiments.
Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan and Noah A. Smith.
ACL 2011.

Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters.
Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider and Noah A. Smith.
NAACL 2013.

2

Tagger, tokenizer, clusters are available at
http://www.ark.cs.cmu.edu/TweetNLP/

NLP on social media's own terms

ikr	smh	he	asked	fir	yo	last
name	so	he	can	add	u	on
fb	lololol					

- Any NLP system, starting with POS tagging, needs different models/resources than traditional written English

Linguistic/speech act diversity on Twitter

Official announcements

Business advertising

Links to blog and web content

Celebrity self-promotion

Status messages

Group conversation

Personal conversation

[Slide credit: Jacob Eisenstein]

- Making a POS tagger
  - Tokenizer
  - Annotate small amount of POS data
  - Design features for supervised model
- Unsupervised word clusters for lexical generalization

- Analyzing the system reveals social confounds in social media NLP

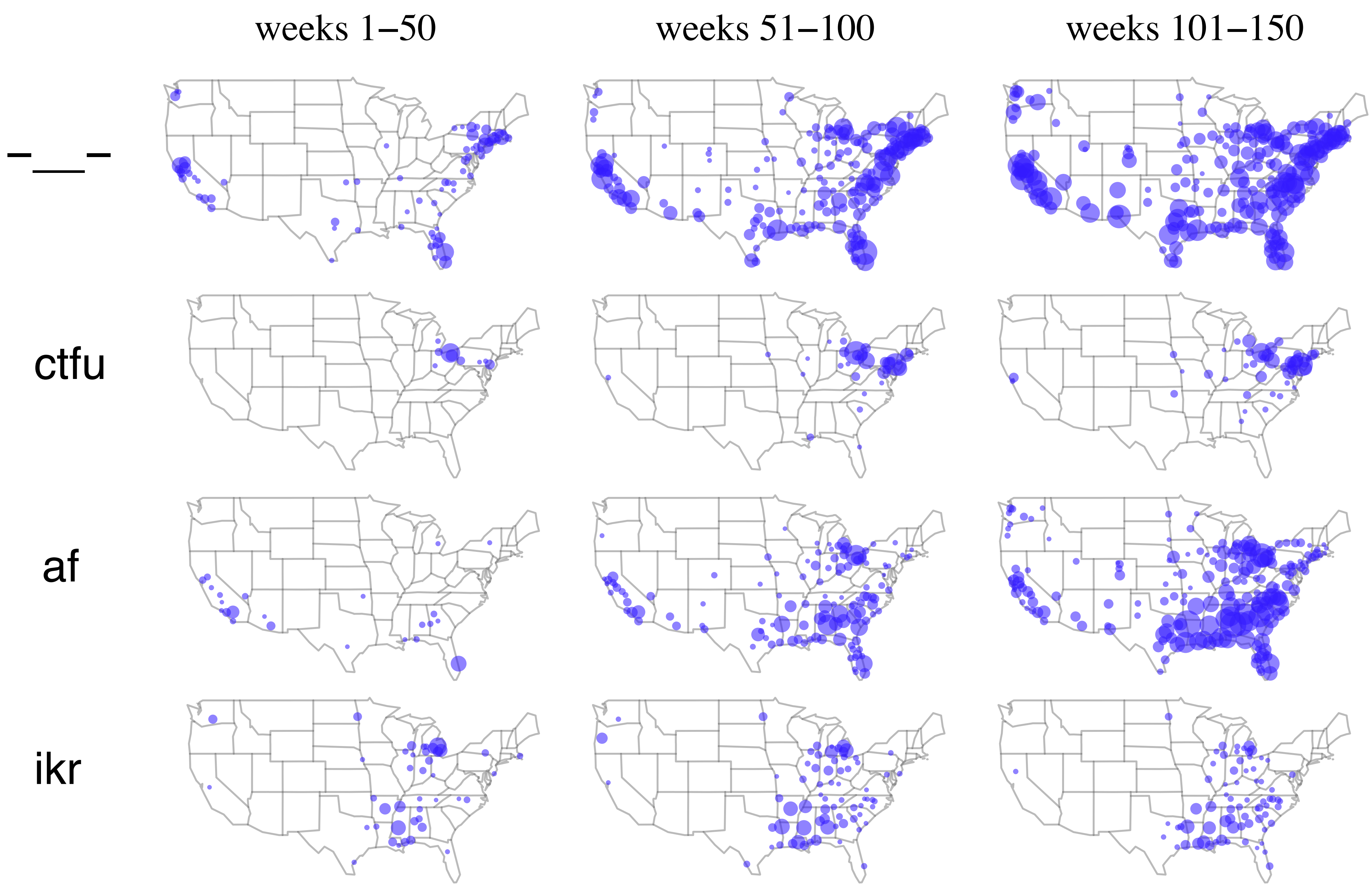
- POS taggers for English Twitter
  - This work: ARK TweetNLP *[Gimpel et al. 2011, Owoputi et al. 2013]*
  - See also GATE *[Derczynski et al. 2013]*

Tokenizer

- split [^a-z0-9] => "p" "d" are top-100 words
  [:-P :D]
- Strategy: recognize punctuation-heavy entities to protect from splitting (emoticon, URL regexes)
- Data-driven rule-based development: at each change, run on large unlabeled corpus, inspect diff
- *twokenize.py*, *Twokenize.java*
- Language change is already hurting the tokenizer
  - New emoticons, URL TLDs

Tokenizer matters for analysis

Geographic diffusion of novel terms, 2009–2012



Also note: geographic specificity of some terms

[Eisenstein, O'Connor, Smith, Xing, PLOS ONE 2014]

Just a little annotated data

	#Msg.	#Tok.	Tagset	Dates
OCT27	1,827	26,594	App. A	Oct 27-28, 2010
DAILY547	547	7,707	App. A	Jan 2011–Jun 2012
NPSCHAT	10,578	44,997	PTB-like	Oct–Nov 2006
(w/o sys. msg.)	7,935	37,081		
RITTERTW	789	15,185	PTB-like	unknown

- Focus: quality (well, consistency?) over quantity
  - Coarse tagset for ease of annotation
    - Twitter-specific: Emoticons, discourse markers, non-constituent hashtags
    - Compound tokens
  - Annotation process sharpened intuitions about the data
- Sustainability of small annotations approach to language diversity?

Features (MEMM tagger)

- Direct representations
  - Lexical identity, shape, prefix/suffix ngrams
- Regexes: Emoticons, hashtags, @-mentions
- Dictionary lookups
  - Traditional POS dictionary
  - Name lists
  - **Word clusters**

Word clustering

- Labeled data is small and sparse. Lexical generalization via induced word classes.
- Unsupervised HMM with hierarchical clustering [*Percy Liang (2005)*'s version of Brown clustering]
  - Word belongs to only one class (bad assumption, but better than alternative; *Blunsom et al. 2011*)
- Big Data vs. I Make My Own Data
  - Unlabeled: 56 M tweets, 847 M tokens
  - Labeled: 2374 tweets, 34k tokens
- 1000 clusters over 217k word types

http://www.ark.cs.cmu.edu/TweetNLP/cluster_viewer.html

What does it learn?

- Orthographic normalizations

so s0 -so so- $o /so //so

soo sooo soooo sooooo soooooo sooooooo soooooooo sooooooooo soooooooooo sooooooooooo soooooooooooo sooooooooooooo soso soooooooooooooo sooooooooooooooo soooooooooooooooo sososo superrr sooooooooooooooooo ssooo so0o superrrr so0 soooooooooooooooooo sosososo sooooooooooooooooooo ssoo sssooo soooooooooooooooooooo #too s0o ssoooo s00 sooooooooooooooooooooo so0o0o sososososo soooooooooooooooooooooo sssoooo ssooooo superrrrr very2 s000 sooooooooooooooooooooooo soooooooooooooooooooooooo sooooooooooooooooooooooooo _so_ soooooooooooooooooooooooooo /so/ sssooooo sosososososo

- suggests joint model against FSA -- *Wulff and Søgaard* here
- compare: word2vec learned embeddings, *Godin et al.* here

- Emoticons etc.
(Clusters/tagger useful for sentiment analysis: NRC-Canada SemEval 2013, 2014)

:d ^^ =d *-* :-d \o/ :dd \m/ 8d *--* *_* u.u :ddd ;;) *.* o/ ;3 =))) *---* \(´▽`)/ n_n b-) (^_^) ^o^ :dddd ;dd *__* :)))))) *----* d/ \o \: =dd n.n -q *___* :33 :ddddd :od -n *-----* xddddd <URL-crunchyroll.com> ^^v (x \= =:) *------* \0/ (˘_˘")

;) :p :-) xd ;-) ;d (; :3 ;p =p :-p =)) ;] xdd 😏 #gno xddd >:) ;-p >:d □ 8-) □ 😜 □ ;-d □ 😝 [; □ :^) =)))) ;-)) <URL-seesmic.com> :pp :~) x'd :op >:p ;^) >:] =))))) :>) <URL-hstl.co> ;)))) ;~) toort >:3 #eden ;pp

:) (: =) :)) :] ☺ :') =] ^_^ :))) ^.^ [: ;)) 😊 ((: ^__^ (= ^-^ :)))) 😁 👍 □ :-)) 😉 ✌ ^___^ (': :} :))))) □ 😃 👌 □ □ 😌 :") :]] □ =]] 😄 □ □ ü ;))) [= (-: ^____^ ;') :-))) (((:

:o o_o « o.o xp ;o ._. t.t t_t #wtf #lol o: x_x =o 0_o dx o_0 :-o ¬¬ --" 0_0 o__o »» u_u #help --' =3 (-_-) -) #confused □ #omg ¬_¬ t^t otl #igetthatalot 😱 xdddd o___o @@ cx t__t d8 □ :{ t___t ---------- #whodoesthat e_e :oo

:(:/ -_- -.- :-(:'(d: :| :s -__- =(=/ >.< -___- :-/ </3 :\ -____- ;(/: ☹ :((>_< =[:[#fml 😒 -_____- =\ >:(😔 -,- >.> >:o ;/ 😳 d; .-. -______- >_> :(((-_-" =s □ ;_; #ugh :-\ =.= □ -_______-

x xx xxx xxxx xxxxx qt xox xxxxxx xxxxxxx xxxxxxxx #pawpawty xxxxxxxxx xxxxxxxxxx #1dfamily #frys #1dqa xxxxxxxxxxx #askliam #dcth xxxxxxxxxxxx #askniall *rt #jbinpoland xxxxxxxxxxxxx #askharry x-x #wiimoms xxxxxxxxxxxxxx oxox #wlf #nipclub +) 1dhq xxxxxxxxxxxxxxx #20peopleilove <URL-paidmodels.com> yart #jedreply #elevensestime <URL-shrtn.us> #askzayn xxxxxxxxxxxxxxxx #wineparty +9 #amwritingparty #tweepletuesday #soumanodomano <URL-today.com> #twfanfriday 22h22

<3 ♥ xoxo <33 xo <333 ❤ ♡ #love s2 <URL-twitition.com> #neversaynever <3333 #swag x3 #believe #100factsaboutme ♥♥ 😘 <3<3 <33333 #blessed xoxoxo 😍 #muchlove #salute xoxox ♥♥♥ #excited ☀ □ #happy #leggo #cantwait <3<3<3 #loveit <333333 #please #dailytweet #thanks 🙏 (˘‿˘) 💜 #yay #thankyou #loveyou {} ε˘`) #nsn #iloveyou

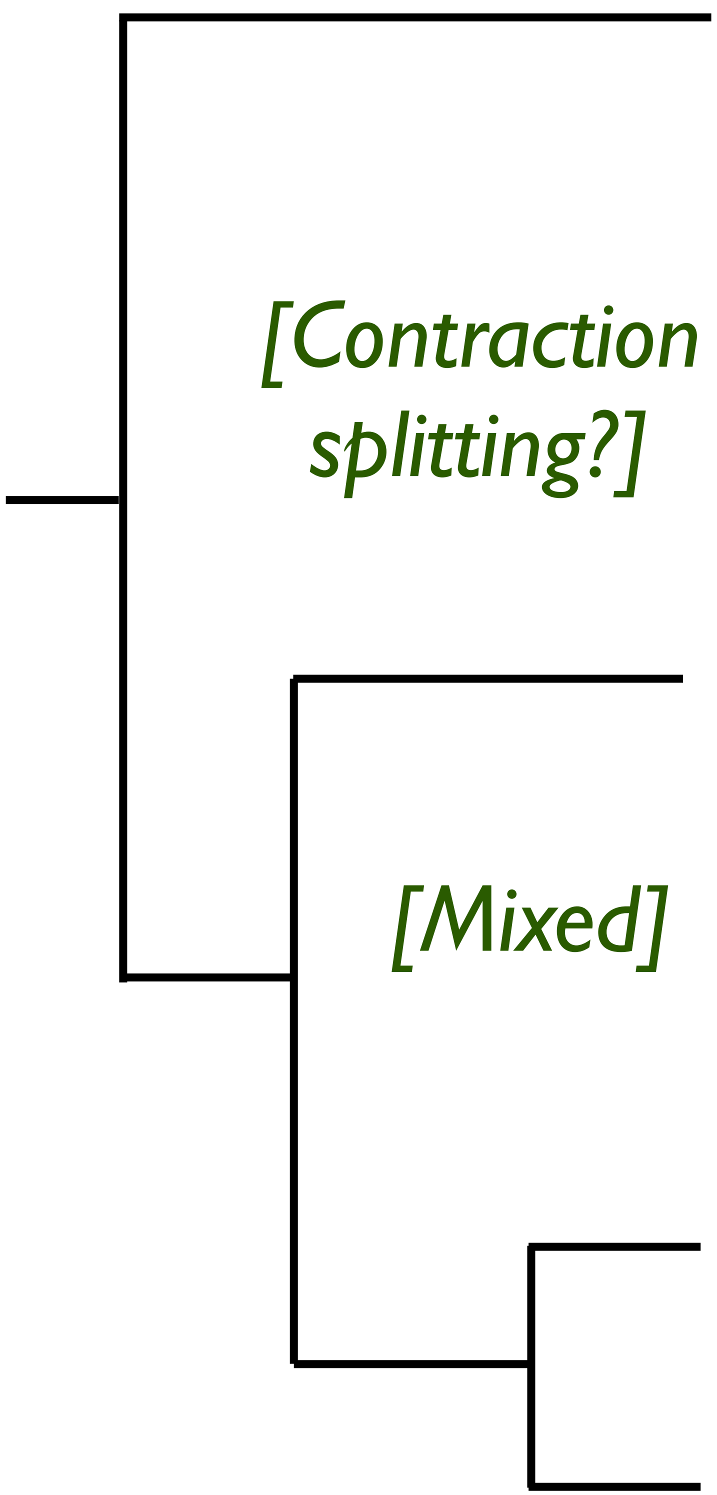
(Immediate?) future auxiliaries

gonna gunna gona gna guna gnna ganna qonna gonnna gana qunna gonne goona gonnaa g0nna goina gonnah goingto gunnah gonaa gonan gunnna going2 gonnnna gunnaa gonny gunaa quna goonna qona gonns goinna gonnae qnna gonnaaa gnaa

tryna gon finna bouta trynna boutta gne fina gonn tryina fenna qone trynaa qon boutaa funna finnah bouda boutah abouta fena bouttah boudda trinna qne finnaa fitna aboutta goin2 bout2 finnna trynah finaa ginna bouttaa fna try'na g0n trynn tyrna trna bouto finsta fnna tranna finta tryinna finnuh tryingto boutto

- finna ~ “fixing to”
- tryna ~ “trying to”
- bouta ~ “about to”

Subject-AuxVerb constructs



i'd you'd we'd he'd they'd she'd who'd i’d u'd youd you’d iwould theyd icould we’d i`d #whydopeople he’d i´d #iusedto they’d i'ld she’d #iwantsomeonewhowill i'de imust a:i'd you`d yu'd icud l'd

ill ima imma i'ma i'mma ican iwanna umma imaa #imthetypeto iwill amma #menshouldnever igotta #whywouldyou #iwishicould #sometimesyouhaveto #thoushallnot #ihatewhenpeople illl #thingspeopleshouldnotdo #howdareyou #thingsgirlswantboystodo im'a #womenshouldnever #thingsblackgirlsdo immma iima #ireallyhatewhenpeople ishould #thingspeopleshouldntdo #irefuseto itl #howtospoilahoodrat iwont imight #thingsweusedtodoaskids ineeda #thingswhitepeopledo we'l #whycantyoujust #whydogirls #everymanshouldknowhowto #ushouldnt #howtopissyourgirloff #amanshouldnot #uwannaimpressme #realfriendsdont immaa #ilovewhenyou

you'll we'll it'll he'll they'll she'll it'd that'll u'll that'd youll ull you’ll itll there'll we’ll itd there'd theyll this'll thatd thatll they’ll didja he’ll it’ll yu'll she’ll youl you`ll you'l you´ll yull u'l it'l we´ll we`ll didya that’ll it’d he'l shit'll they'l theyl she'l everything'll he`ll things'll u’ll this'd

i'll i’ll i'l i`ll i´ll i'lll l'll i\'ll i"ll -i'll /must @pretweeting she`ll

Syntactic slant

^**01**00110111*

called named considered spelled titled pronounced
finished completed finshed #crunchyroll #viggle finishd
started stopped began awoke stoped cbf startd
starts ends begins dies continues opens stops

^**01**11101110*

calls wins hits offers runs plays leaves leads
shows changes beats moves presents answers cuts
lives talks heads faces hearts minds bodies backs

- *called* / *calls* very far away in tree
- Weakness of Brown clustering (HMM favors local syntax; hard clustering doesn't do ambiguity), but is kinda OK for POS tagging

Word clusters as features

ikr	smh	he	asked	fir	yo	last
!	G	O	V	P	D	A
name	so	he	can	add	u	on
N	P	O	V	V	O	P
fb	lololol					
^	!					

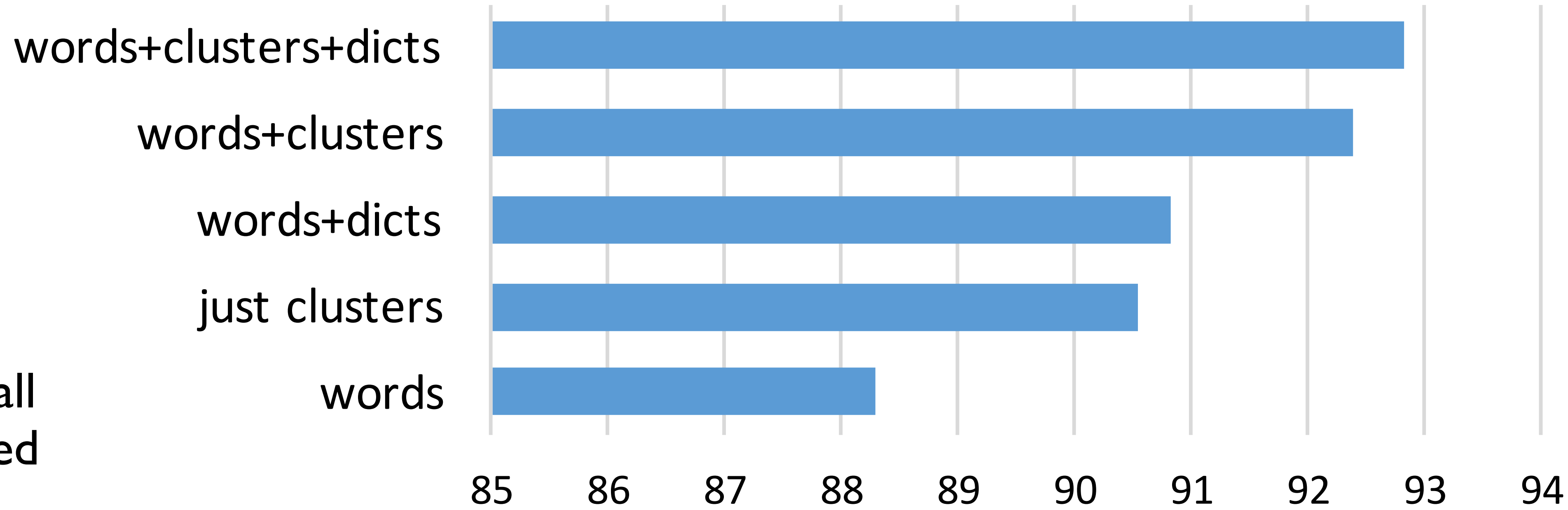
w fo fa fr fro ov fer **fir** whit abou aft serie fore fah fuh w/her w/that fron isn agains	“non-standard prepositions”
yeah yea nah naw yeahh nooo yeh noo noooo yeaa **ikr** nvm yeahhh nahh nooooo	“interjections”
facebook **fb** itunes myspace skype ebay tumblr bbm flickr aim msn netflix pandora	“online service names”
smh jk #fail #random #fact smfh #smh #winning #realtalk smdh #dead #justsaying	“hashtag-y interjections”??

Highest-weighted POS–treenode features

hierarchical structure generalizes nicely.

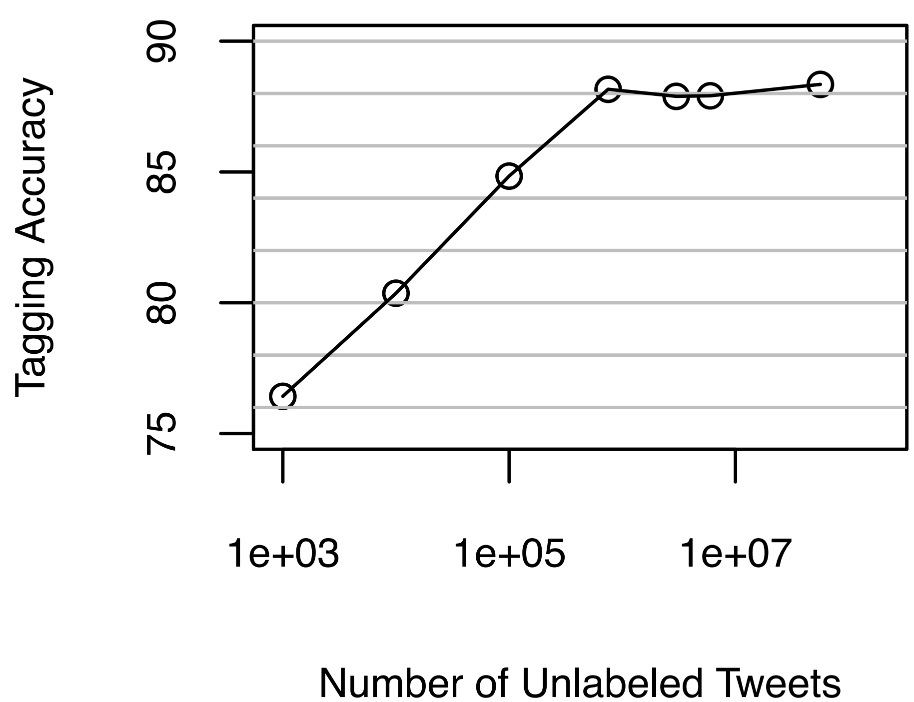
Cluster prefix	Tag	Types	Most common word in each cluster with prefix
11101010*	!	8160	lol lmao haha yes yea oh omg aww ah btw wow thanks sorry congrats welcome yay ha hey goodnight hi dear please huh wtf exactly idk bless whatever well ok
11000*	L	428	i'm im you're we're he's there's its it's
1110101100*	E	2798	x <3 :d :p :) :o :/
111110*	A	6510	young sexy hot slow dark low interesting easy important safe perfect special different random short quick bad crazy serious stupid weird lucky sad
1101*	D	378	the da my your ur our their his
01*	V	29267	do did kno know care mean hurts hurt say realize believe worry understand forget agree remember love miss hate think thought knew hope wish guess bet have
11101*	O	899	you yall u it mine everything nothing something anyone someone everyone nobody
100110*	&	103	or n & and

Clusters help POS tagging



Test set accuracy

"words": all handcrafted features



Dev set accuracy
using only clusters as features

Clusters help for nonstandard terms

Model	**In dict.**	**Out of dict.**
Full	93.4	85.0
No clusters	92.0 (-1.4)	79.3 (-5.7)
Total tokens	*4,808*	*1,394*

Table 3: DAILY547 accuracies (%) for tokens in and out of a traditional dictionary, for models reported in rows 1 and 3 of Table 2.

Many uses of word clusters

- Features for downstream tasks
- Exploratory analysis of lexicon
- Assist manual dictionary building
  - Name filter
  - Emotion keyword lists

- Where do nonstandard terms come from?
  - “Noise”: orthographic deviations from “true” form (accidental? intentional / creative?)
- Or...

imma

https://twitter.com/search?q=imma&src=typd&vertical=default&f=tweets

r.i.p spikeee @xEnvyme_alex · 2m
Feel like **imma** have it up way more my sophomore year than I did freshman year

Angelica @BrowncoatAnge · 2m
Imma start unfollowing all these celebs crying about a Lion and not even remotely interested in the struggle of the Black community.

Billi3 J3an @Misskooki3 · 2m
Niggas thought they wanted to pimp me before they noticed bitch **imma** boss.. I pimp the nigga who brought u here

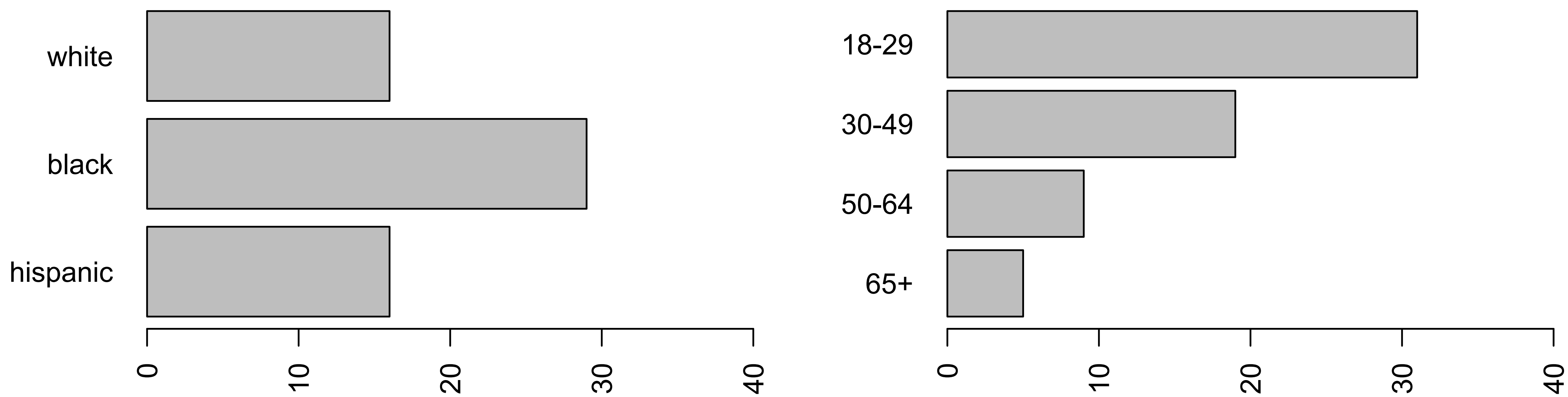
kye @hippys0ul · 2m
K imma just scroll my tweets to the day abel followed. 😬❤️

LifeΣtyleTr3 @LifeStyleTr3 · 2m
@Charismatic_Cee ay shoot me those pictures you calm you got of us n **imma** post em on the gram !

Ⓜarolex @Marolex_ · 2m
imma firin' mah razor

P(use twitter | demographics)

2013



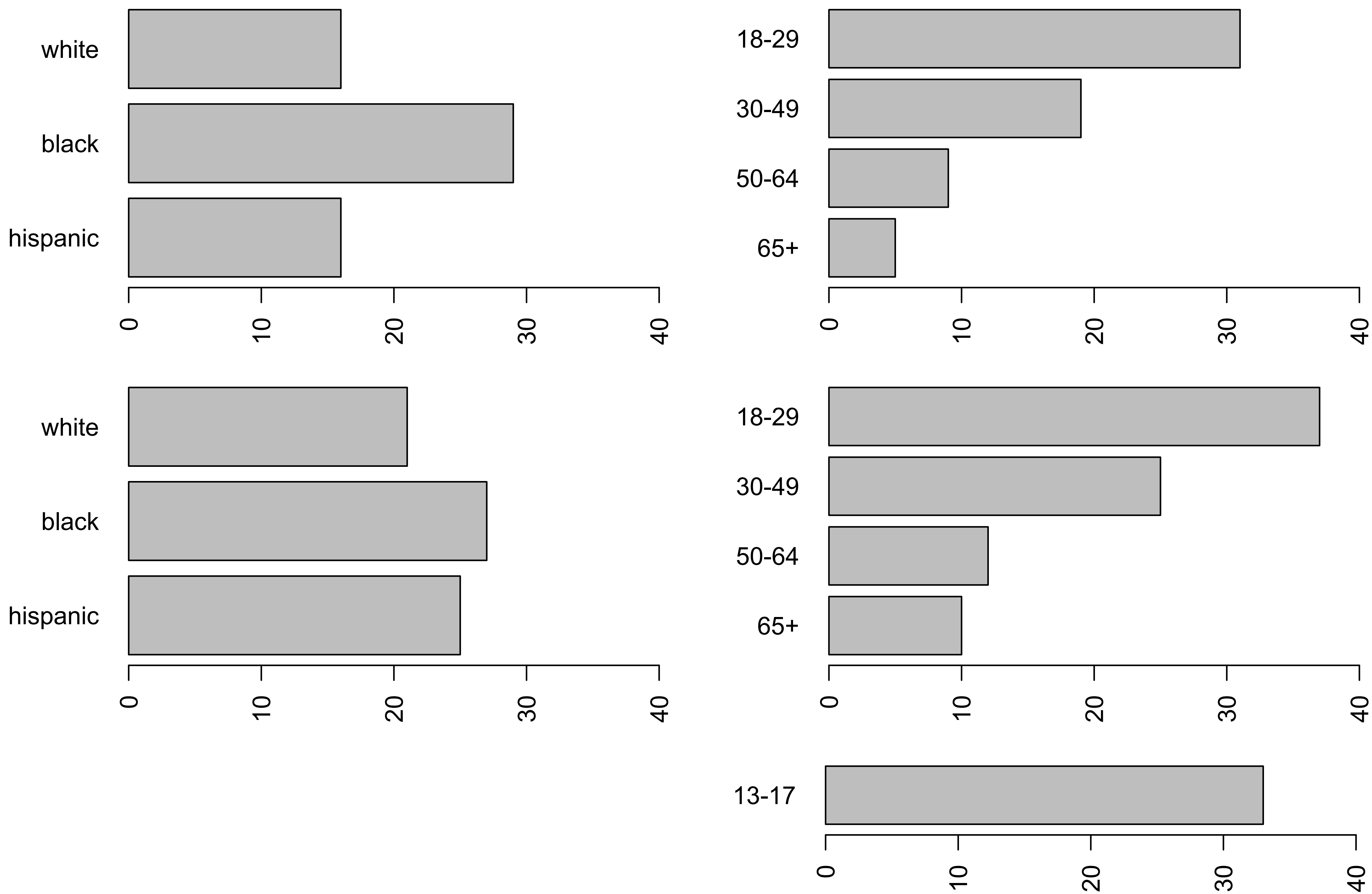
- Overrepresented: younger ages, urban areas, sometimes minorities
- U.S. data: Pew Research

P(use twitter | demographics)

2013

2014

2015

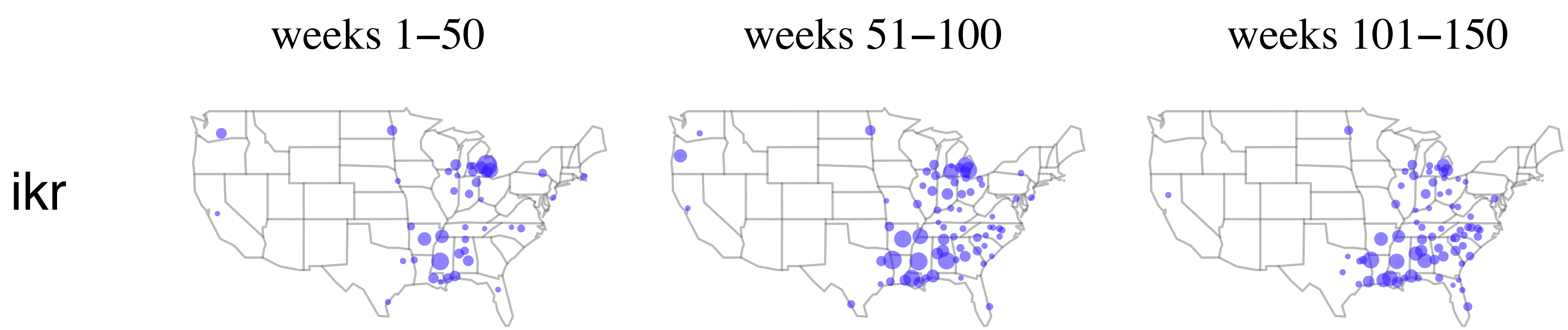


- Overrepresented: younger ages, urban areas, sometimes minorities
- U.S. data: Pew Research

Geographic and textual context give clues to meaning?

“ikr” =?= “I know, right?”

yeah yea nah naw yeahh nooo yeh noo noooo yeaa **ikr** nvm yeahhh nahh nooooo yh yeaaa yeaah yupp naa yeahhhh yeaaah iknow werd noes nahhh naww yeaaaa shucks yeaaaah yeahhhhh naaa naah nawl nawww yehh ino yeaaaaa yeeah yeeeah wordd yeaahh nahhhh naaah yeahhhhhh yeaaaaah naaaa yeeeeah nall yeaaaaaa



- Who are we building tools for?
  - Your noise is my dialect
  - Dominant vs minority language politics
    - Ebonics controversy; English as U.S. official language
    - Ukrainian/Russian
    - etc. etc. etc.
  - Compare: low-resource languages
- Usefulness of noise metaphor

- Are these forms of speech unique to the new medium, or is it novel digitized recording of long-standing dialectical variation (e.g. African-American English, or Egyptian Arabic...)?
  - *lol* vs. *imma* (?)
- Both "noise" and deeper variation exist. How to distinguish, and how much linguistic variation is due to each?
  - Phonological sources of social media spelling variation *[Eisenstein, J Socioling. 2015; Jørgensen et al., here]*
- Laboratory to analyze code switching, creoles, other non-formal language phenomena and corpus sociolinguistics more generally *[e.g. Hovy et al., WWW 2015]*
- Implications for NLP-driven social analysis?