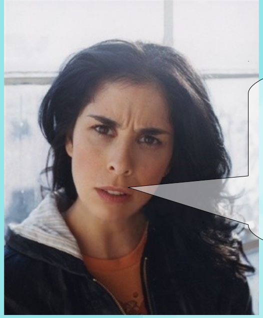
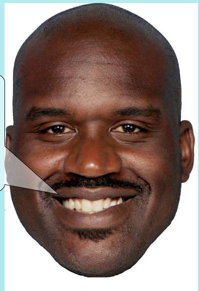
BAD LANGUAGE!

...on the INTERNET!!

Jacob EISENSTEIN
GEORGIA Institute of TECHnology



Boom! Ya ur website suxx bro



...dats why pluto is pluto it can neva be a star

michelle obama great. job. and. whit all my. respect she. look. great. congrats. to. her.

I now h v an iphone

What can we do about it?

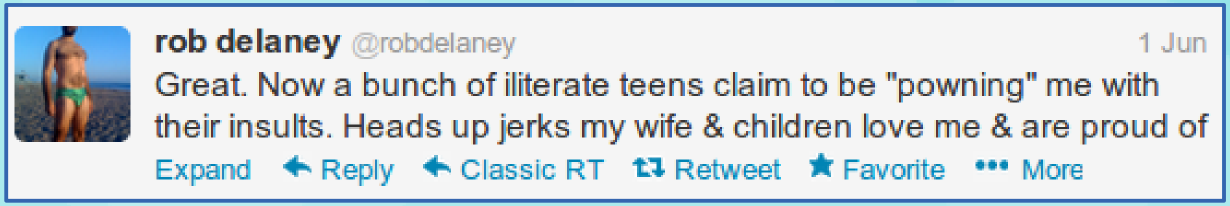
Why don't they just write ***NORMALLY??***

Can our software ever ***ADAPT???***

How does language go bad?

Illiteracy? No.

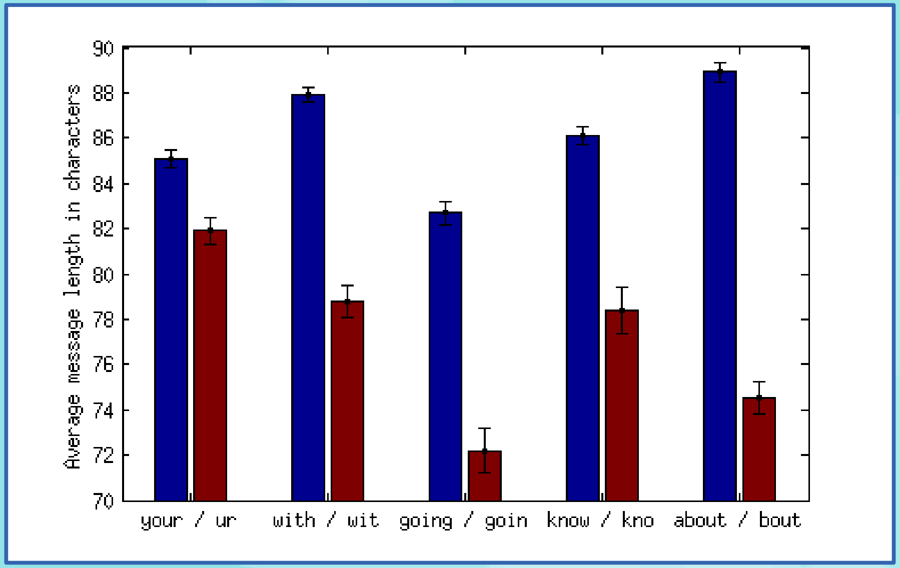
(Tagliamonte and Denis 2008; Drouin and Davis 2009)



rob delaney @robdelaney 1 Jun

Great. Now a bunch of illiterate teens claim to be "powning" me with their insults. Heads up jerks my wife & children love me & are proud of

Length limits? (probably not)

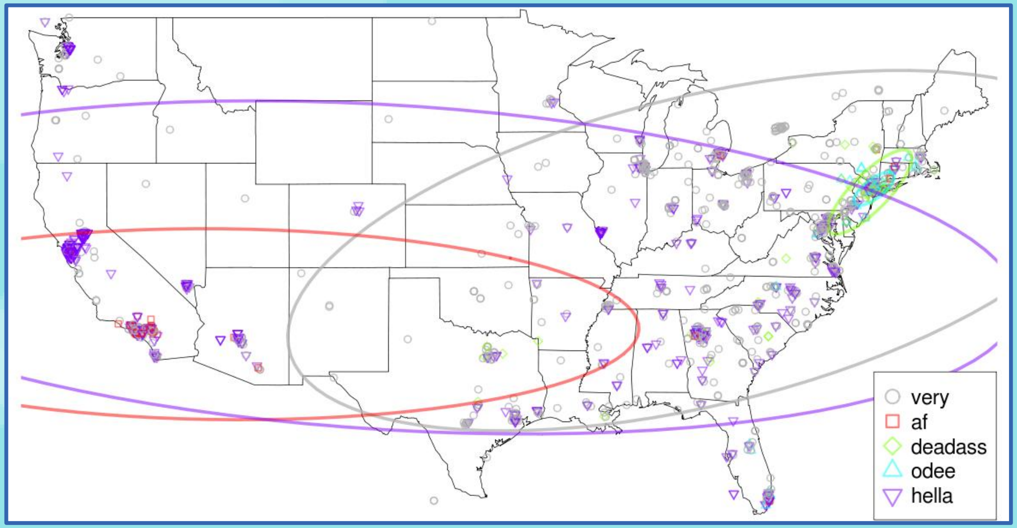


Hardware input constraints?

(Gouws et al 2011)

Social variables

- Non-standard language does *identity work*, signaling authenticity, solidarity, etc.
- Social variation is usually inhibited in written language, but social media is less regulated than other written genres.

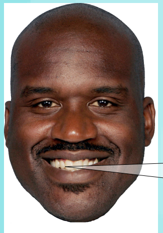


Abnormal

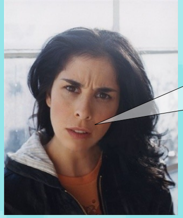
Normal

normalization →

imma → I'm gonna → I am going to

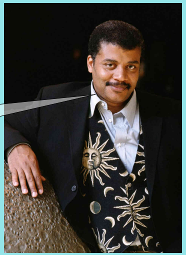


dats why pluto is pluto



Boom! Ya ur
website suxx bro

That's why pluto is pluto



Boom! Yes, your
website sucks, brother.

Irredeemably abnormal?

hella

wtf

jawn

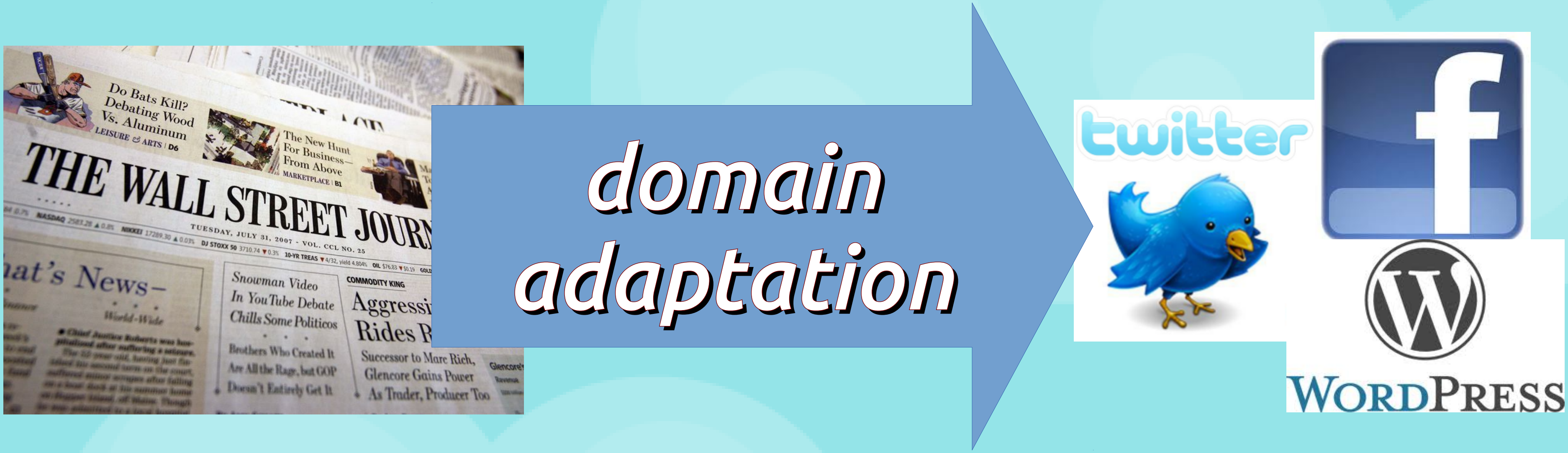
Whose norm?

flavor

flavour

Ain't

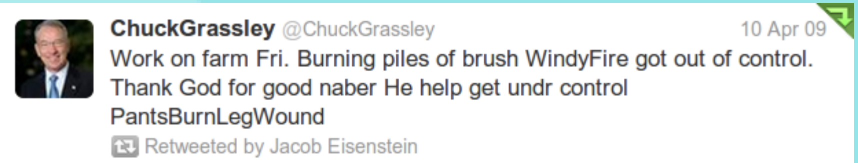
Source   Target(s)



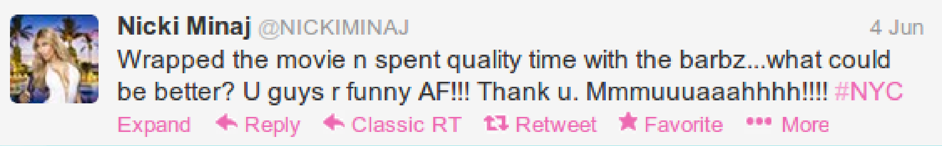
Lots of work on "X-for-Twitter" using domain adaptation.

- POS: Gimpel et al 2011
- NER: Finin et al 2010, Ritter et al 2011
- Parsing: Foster et al 2011

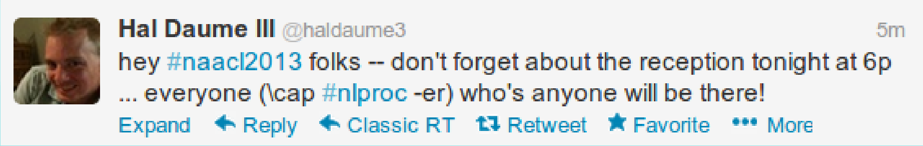
Is social media a domain?
Is Twitter?



ChuckGrassley @ChuckGrassley 10 Apr 09
Work on farm Fri. Burning piles of brush WindyFire got out of control. Thank God for good naber He help get undr control PantsBurnLegWound
Retweeted by Jacob Eisenstein



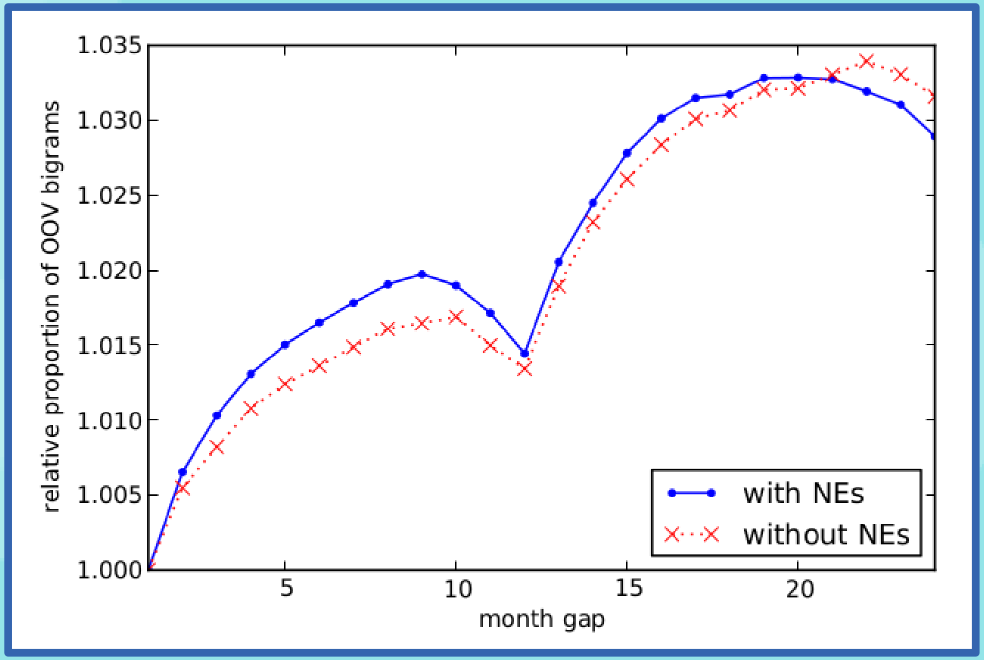
Nicki Minaj @NICKIMINAJ 4 Jun
Wrapped the movie n spent quality time with the barbz...what could be better? U guys r funny AF!!! Thank u. Mmmuuuaaahhhh!!!! #NYC

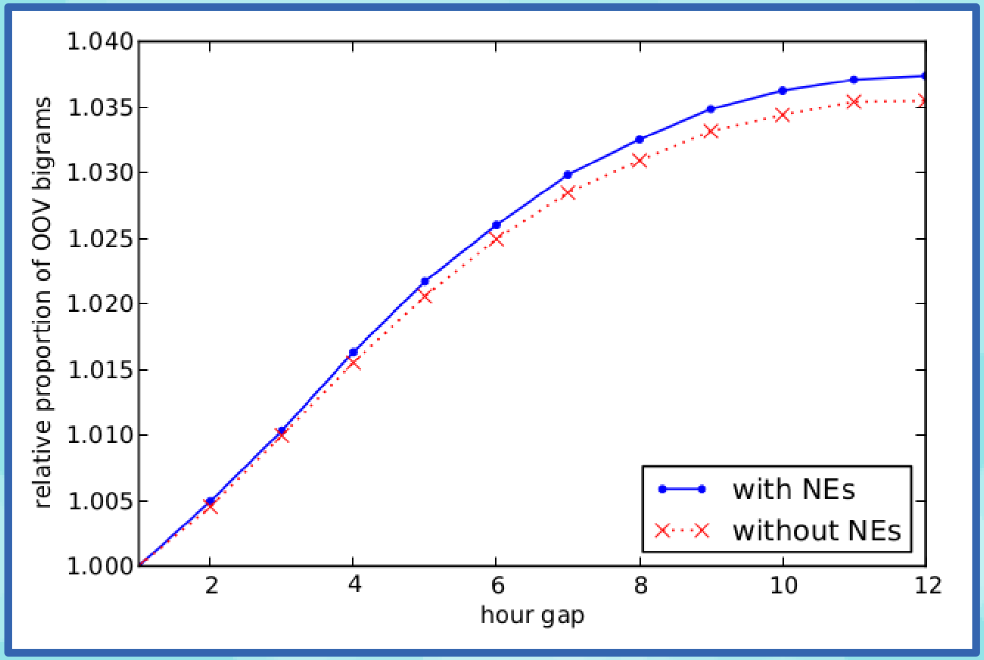


Hal Daume III @haldaume3 5m
hey #naacl2013 folks -- don't forget about the reception tonight at 6p ... everyone (\cap #nlproc -er) who's anyone will be there!

Coherence over time

- **Goal**: measure the linguistic coherence of Twitter
- **Data**: million-word samples at each month and hour
- **Measure**: relative proportion of OOV bigrams



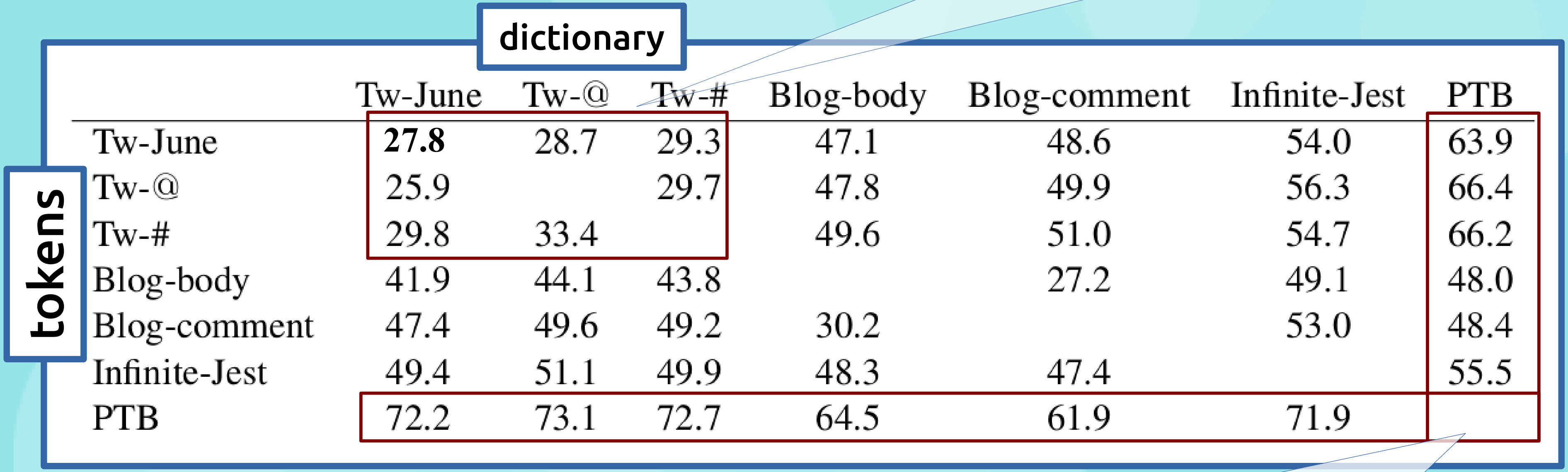


Social media language is changing continuously.

- We cannot annotate our way out of the Bad Language problem.
- Any annotated dataset rapidly becomes stale.

Coherence across media

Twitter is self-similar, but...
OOV rate increases significantly across usage scenarios

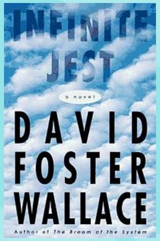


dictionary

tokens

	Tw-June	Tw-@	Tw-#	Blog-body	Blog-comment	Infinite-Jest	PTB
Tw-June	**27.8**	28.7	29.3	47.1	48.6	54.0	63.9
Tw-@	25.9		29.7	47.8	49.9	56.3	66.4
Tw-#	29.8	33.4		49.6	51.0	54.7	66.2
Blog-body	41.9	44.1	43.8		27.2	49.1	48.0
Blog-comment	47.4	49.6	49.2	30.2		53.0	48.4
Infinite-Jest	49.4	51.1	49.9	48.3	47.4		55.5
PTB	72.2	73.1	72.7	64.5	61.9	71.9	

PTB is the clear outlier
most OOV tokens in almost every comparison



WordPress

- Tw-June: randomly-selected messages from June 2011
- Tw-@: messages beginning with a username mention
- Tw-#: messages beginning with a hashtag
- Blog-body: posts from 2008 political blogs (Yano et al 2009)
- Blog-comment: from 2008 political blogs (Yano et al 2009)
- Infinite-Jest: the 1996 novel (Wallace 2012)
- PTB: section 2-21