

# computational social media

## lecture 3: tweeting

### part 1

daniel gatica-perez

# this lecture



a human-centric view of twitter

1. introduction

2. twitter users & uses

3. understanding large-scale human behavior

4. inferring real-world events & trends

5. spreading information in the real world

# twitter basic official statistics

<https://about.twitter.com/company>, accessed 2014, 2016-2020

## mission

“to give everyone the power to create and share ideas and information instantly, without barriers”

	2014	2016	2018 & later
monthly active users	241M	320M	N/A
tweets sent per day	500M	N/A	N/A
active users on mobile	76%	80%	N/A
accounts outside the US	77%	79%	N/A
supported languages	35+	35+	N/A
employees	2700	3900	N/A



"an echo chamber of random chatter"

"140 characters: between an SMS  
(with larger audience) and  
a blog (but less cumbersome)"

280 characters (Fall 2017)



Jack Dorsey 

@jack

 Follow

just setting up my twttr

 Reply  Retweet  Favorite  More

RETWEETS

21,609

FAVORITES

19,815



12:50 PM - 21 Mar 2006

<https://twitter.com/jack/status/20>

<https://about.twitter.com/milestones>

# before twitter...

古池や蛙飛び込む水の音  
ふるいけやかわずとびこむみずのおと

old pond . . .  
a frog leaps in  
water's sound

Bashō (17<sup>th</sup> century)

<https://en.wikipedia.org/wiki/Haiku>

The Dinosaur

On waking, the dinosaur was still there.

Augusto Monterroso (20<sup>th</sup> century)

<https://es.wikipedia.org/wiki/Microrrelato>

Form No. 105, M. D.

# THE WESTERN UNION TELEGRAPH COMPANY. C

24,000 OFFICES IN AMERICA. INCORPORATED CABLE SERVICE TO ALL THE WORLD.

This Company TRANSMITS and DELIVERS messages only on conditions limiting its liability, which have been assented to by the sender of the following message. Errors can be guarded against only by repeating a message back to the sending station for comparison, and the Company will not hold itself liable for errors or delays in transmission or delivery of Unrepeated Messages, beyond the amount of tolls paid thereon, nor in any case where the claim is not presented in writing within sixty days after the message is filed with the Company for transmission.

This is an UNREPEATED MESSAGE, and is delivered by request of the sender, under the conditions named above.  
ROBERT C. CLOWRY, President and General Manager.

RECEIVED at 126 Ch H9-B:- 37 Paid Govt.

Wh:- The White House, Washington, D. C. Apr. 18-06.

Hon. George C. Pardee,

Governor of California

Sacramento, Cal.

Hear rumors of great disaster through an earthquake in San Francisco but know nothing or the real facts call upon me for any assistance I can render.

Theodore Roosevelt.

12:51 P.M.

MONEY TRANSFERRED BY TELEGRAPH.

CABLE OFFICE.

Roosevelt Telegram Great SF Quake, California State Capitol Museum  
Photo by Chris Boyer on Unsplash: <https://unsplash.com/photos/PEd6MBB1kZQ>







CHAINS REQUIRED  
20 MILES AHEAD  
EXPECT DELAYS

CAM



# this lecture

a human-centric view of twitter

1. introduction

2. twitter users & uses

3. understanding large-scale human behavior

4. inferring real-world events & trends

5. spreading information in the real world

# what is twitter made of?



**follow** (2006)  
users subscribe to other users' tweets



**hashtags #** (2007, official 2009)  
words articulating a topic or event  
allow for search and clustering



**retweet RT** (2007, official 2009)  
repost tweets towards one's followers  
enables trends by retweeting

<https://about.twitter.com/press/brand-assets>

<https://about.twitter.com/milestones>

J. van Dijck The culture of connectivity, Oxford University Press, 2013



# hashtags

link **strangers** into  
larger conversations

facilitate **impromptu**  
interactions

not directed  
communication  
but a **stream**

enable the **emergence**  
of trending topics



**Swiss Embassy**  @SwissEmbassyUSA · 7h

Today we celebrate [#WorldWaterDay](#) as a reminder of this precious & essential resource on [#earth](#). [#Switzerland](#) is committed to sustaining high [#environmental](#) standards. Strict laws & regulations ensure access to fresh [#water](#) and 1/3 can be drunk without treatment.





**Wengen Switzerland** @WengenSwiss · Mar 19

One of the best skiing days in this winter season. [#wengen](#) [#Switzerland](#) [#inlovewithswitzerland](#) [#skiing](#)



**Switzerland Tourism**  @MySwitzerland\_e · Mar 20

Today is the [#FirstDayOfSpring](#)! The first signs of spring can also be seen in various places around [#Switzerland](#).   Are you looking forward to the new season?

<https://twitter.com/SwissEmbassyUSA/status/1374086527328907270>

<https://twitter.com/WengenSwiss/status/1372877615472738308>

[https://twitter.com/MySwitzerland\\_e/status/1373205370017964035](https://twitter.com/MySwitzerland_e/status/1373205370017964035)





**Chris Messina™**  
@chrismessina



Follow

how do you feel about using # (pound) for groups. As in [#barcamp](#) [msg]?

Reply Retweet Favorite More

RETWEETS  
**346**

FAVORITES  
**614**



12:25 PM - 23 Aug 2007

<https://twitter.com/chrismessina/status/223115412>



**Eric Rice**  
@ericrice

Follow

ReTweet: [jmalthus @spin](#) Yes! Web2.0 is about social media, and guess what people like to be social about? Themselves. Social Narcissism

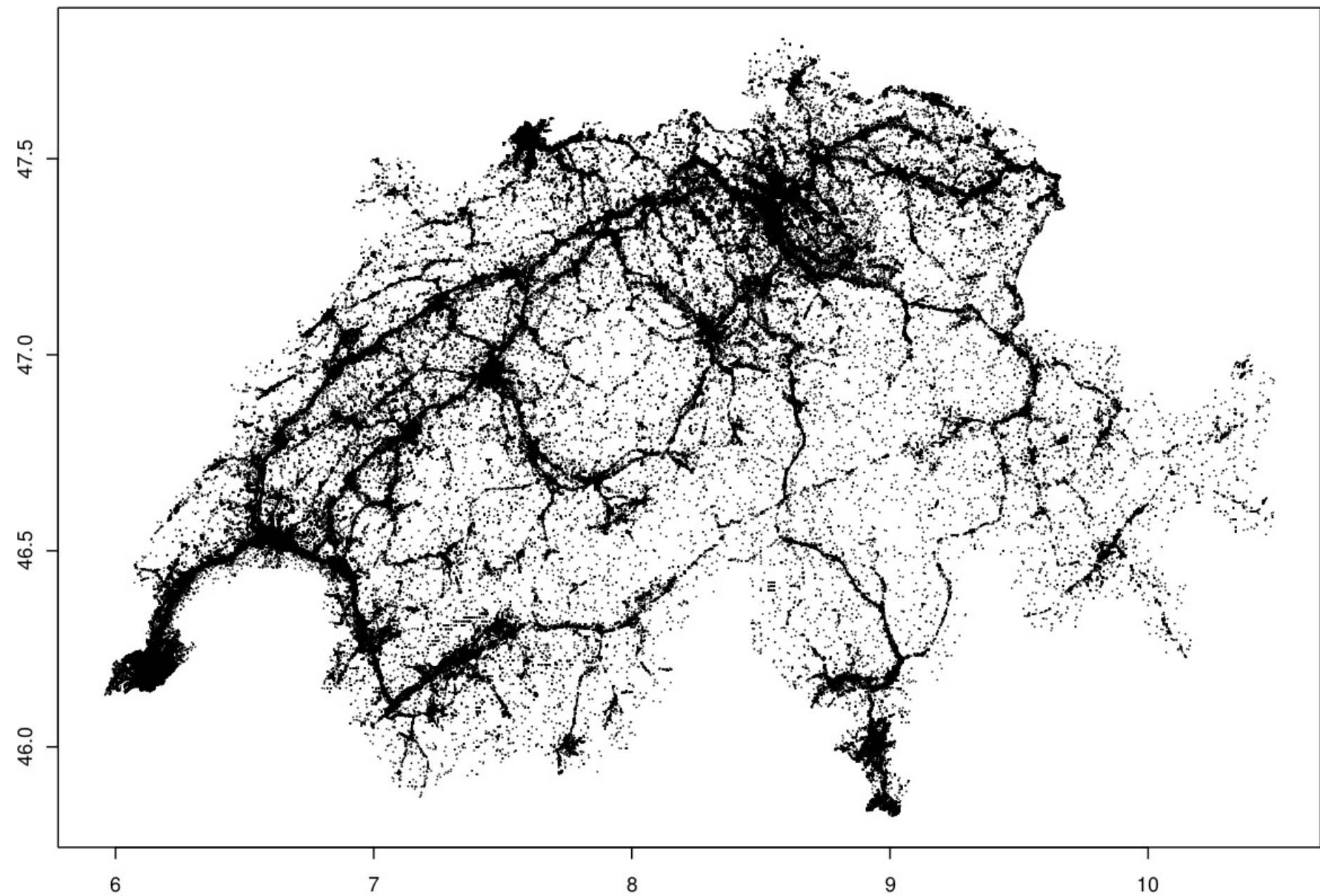
5:33 AM - 18 Apr 2007

19 RETWEETS 34 FAVORITES



<https://twitter.com/ericrice>

# geolocation



# users and usage

- 2006: older professional users in business and news
- 2009: shift to younger adults, then mainstream

## from social network to information network

### tool for communication

- people's everyday small talk
- journalism
- political grassroots activism
- emergencies and disasters
- community participation
- misinformation

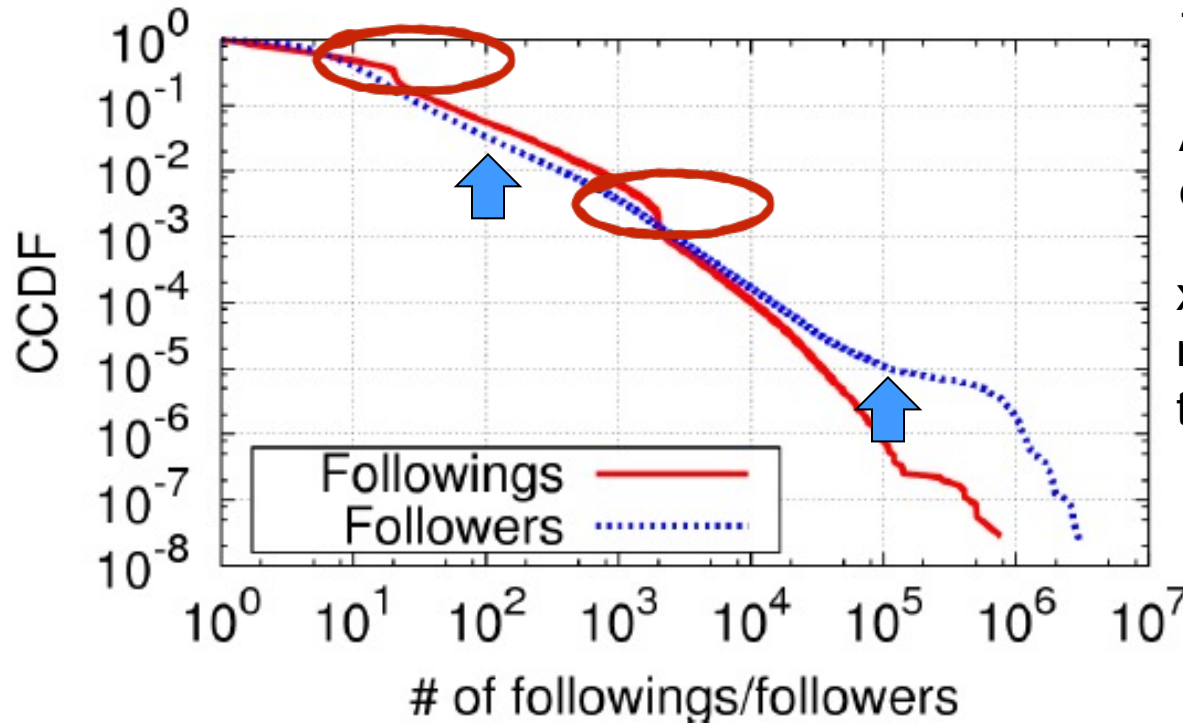
### tool for self-promotion

- celebrities, stars
- politicians

“the impulse to make life a publicly annotated experience has blurred the distinction between advertising and self-expression, marketing and identity”  
(Hagan 2011)

# twitter basic descriptive statistics (2009)

## 41.7M user profiles



Fewer than 10% of users have 100 or more followers

A tiny fraction of users have over 100,000 followers

$x = 20$  : twitter recommended new users to follow 20 users to start with

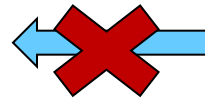
$x=2000$ : before 2009 there was limit on the number of people one could follow (rule removed later)

CCDF: Complementary Cumulative Distribution Function

$$\text{CCDF}(x) = 1 - F(x) = P(X > x)$$



# following vs. friending



**“connection with very low expectation”**: weak ties (Murthy, 2013)

**low reciprocity, highly asymmetric links** (Kwak, 2010)

“77.9% of user pairs with any link between them are connected one-way.”

“67.6% of users are not followed by any of their followings. For these users Twitter is rather a source of information than a social networking site”

# real names in twitter



twitter does not require real names

pseudonyms are valuable in information networks: “not real names but persistent identity with reputation attached”

it works as identity service for individuals & entities whose long-time presence depends on being identified

downside: fake user accounts

# this lecture

a human-centric view of twitter

1. introduction

2. twitter users & uses

3. understanding large-scale human behavior

4. inferring real-world events & trends

5. spreading information in the real world

## **case study: twitter and human mood**

S. A. Golder and M. W. Macy, Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures, *Science*, 30 September 2011, Vol. 333 no. 6051 pp. 1878-1881



## mood

“a conscious state  
of mind or  
predominant emotion”  
(Merriam-Webster  
dictionary)

“a temporary state of  
mind or feeling”  
(Oxford dictionary)



# understanding mood expressed on Twitter

## positive affect (PA):

enthusiasm, delight, activeness, alertness

## negative affect (NA):

distress, fear, anger, guilt

## PA and NA are independent dimensions

low PA: absence of positive feelings, not presence of negative ones

**goal:** study variations in PA & NA over time of day, day of week, and world region using longitudinal twitter data

- \* 2.4 million twitter users worldwide
- \* 509 million tweets
- \* up to 400 public messages per user
- \* all users had at least 25 messages
- \* average: 212 tweets/user
- \* period: 02.2008 and 01.2010
- \* only english speakers

# extraction of positive affect (PA) and negative affect (NA)



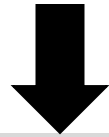
**Vice President Kamala Harris**  @VP · Mar 20

...

United States government official

Sending best wishes to [@SuluhuSamia](#) following her swearing in as Tanzania's new President - the first woman to hold the office. The United States stands ready to work with you to strengthen relations between our countries.

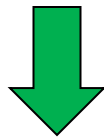
<https://twitter.com/VP/status/1373341947184693252>



Linguistic Inquiry & Word Count (LIWC)

Word categories related to psychological constructs and personal concerns

Word count per category



**PA**



**NA**



# LIWC: Linguistic Inquiry & Word Count (2022, 2015, 2007, 2001, 1999)



HOME

TRY IT NOW

▼ HELP

DICTIONARIES

CONTACT US

BUY NOW

## INTRODUCING LIWC-22

A NEW SET OF TEXT ANALYSIS TOOLS AT YOUR FINGERTIPS

People reveal themselves by the words they use. Using LIWC-22 to analyze others' language can help you understand their thoughts, feelings, personality, and the ways they connect with others. It can give you insights you've never had before into the people and world around you.

[BUY LIWC NOW](#)

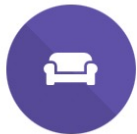
[CONTACT US](#)

<https://www.liwc.app>



# DISCOVER THE WORLD OF WORDS

Linguistic Inquiry and Word Count (LIWC) is the gold standard in software for analyzing word use. It can be used to study a single individual, groups of people over time, or all of social media.



## EASY TO USE

LIWC-22 requires no advanced linguistics or computer science skills. With it, you can analyze single or multiple text files, words in spreadsheets, or simply copy and paste text into the program. Your results can be displayed in word clouds, graphs, or data spreadsheets.



## THE HIGHEST SCIENTIFIC STANDARDS

LIWC-22 analyzes over 100 dimensions of text, all of which have been validated by respected labs around the world. Over [20,000 scientific articles](#) have already been published using LIWC.



## FLEXIBILITY FOR MORE ADVANCED USERS

LIWC-22 works as an all-inclusive desktop application which now integrates with other programming languages like Python and R while still taking advantage of the LIWC processing engine.

**Table 2. LIWC-22 Language Dimensions and Reliability**

Category	Abbrev.	Description/Most frequently used exemplars	Words/ Entries in category*	Internal Consistency: Cronbach's $\alpha$	Internal Consistency: KR-20
<b>Summary Variables</b>					
Word count	WC	Total word count			
Analytical thinking	Analytic	Metric of logical, formal thinking			
Clout	Clout	Language of leadership, status			
Authentic	Authentic	Perceived honesty, genuineness			
Emotional tone	Tone	Degree or positive (negative) tone			
Words per sentence	WPS	Average words per sentence			
Big words	BigWords	Percent words 7 letters or longer			
Dictionary words	Dic	Percent words captured by LIWC			
<b>Linguistic Dimensions</b>	Linguistic				
Total function words	function	the, to, and, I			
Total pronouns	pronoun	I, you, that, it			
Personal pronouns	ppron	I, you, my, me			
1st person singular	i	I, me, my, myself			
1st person plural	we	we, our, us, lets			
2nd person	you	you, your, u, yourself			
3rd person singular	shehe	he, she, her, his			
3rd person plural	they	they, their, them, themsel*			
Impersonal pronouns	ipron	that, it, this, what			
Determiners	det	the, at, that, my			
Articles	article	a, an, the, alot			
Numbers	number	one, two, first, once			
Prepositions	prep	to, of, in, for			

dictionary:

12,000 words (2022)

6,400 words i(2015)

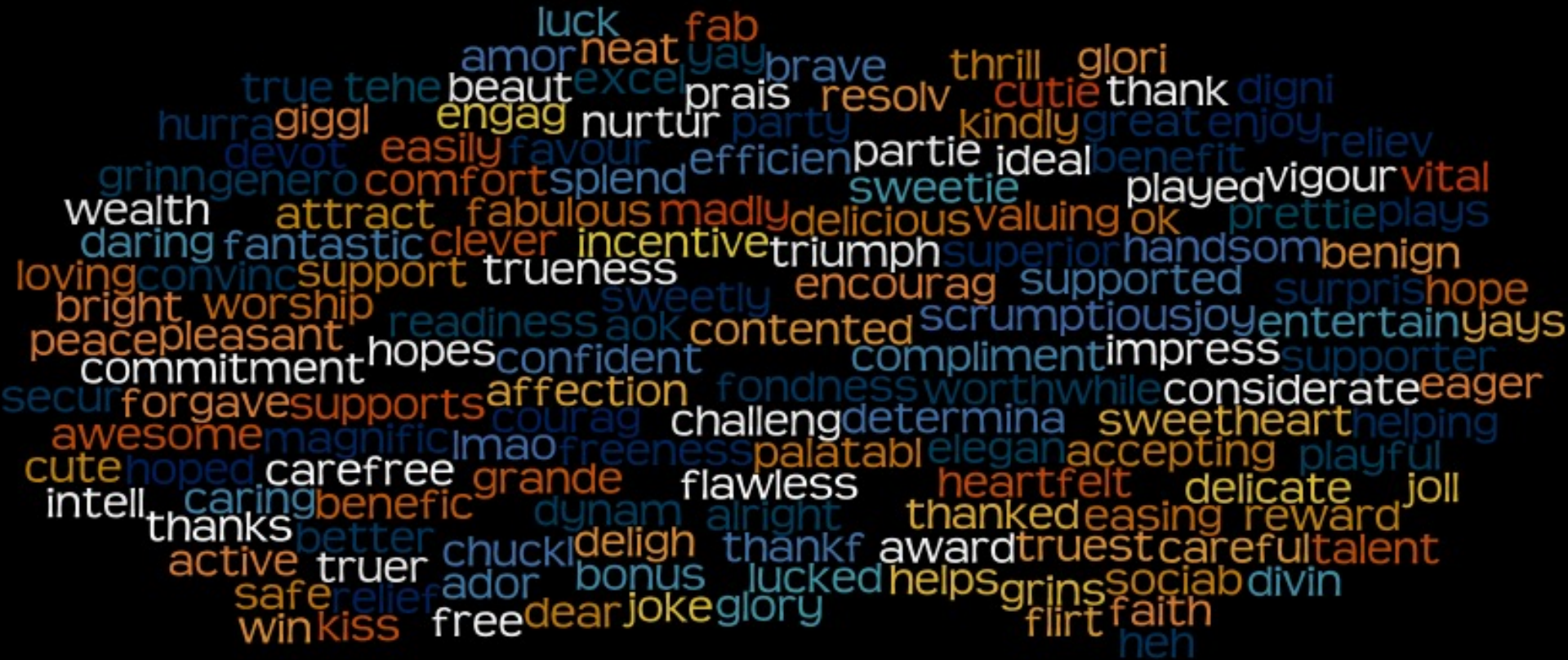
4,500 words (2007)

each word belongs to one or more categories

example: “agree” is part of: *affect*, *positive emotions* and *assent*

over 60 word categories (excluding punctuation)

# positive emotion category words









## measurements

$$PA_u(h) = \frac{\|PAWORDS_u(h)\|}{\|WORDS_u(h)\|} \quad (1)$$

where  $h \in H$  and  $H = \{0 \dots 167\}$ , or the 168 hours of the week (24 hours/day \* 7 days). The measure for NA was computed similarly, as were the measures taken over 24 hours.

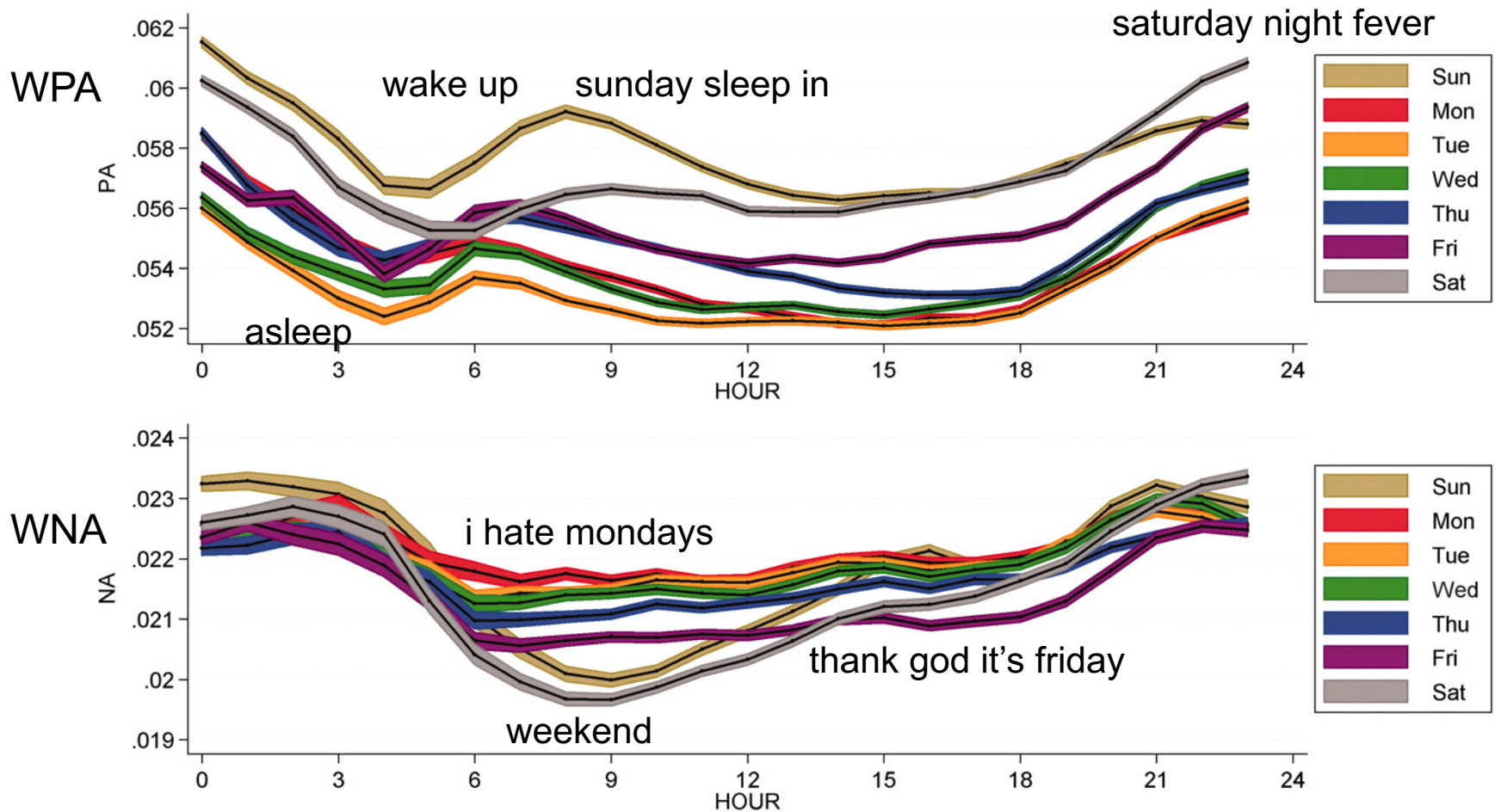
Between-individual variation captures how individuals differ from one another in their baseline affect regardless of the time of day or day of week. It is simply the individual's mean affect across all hours:

$$BPA_u = \overline{PA_u} = \frac{1}{\|H\|} \sum_{h \in H} PA_u(h) \quad (2)$$

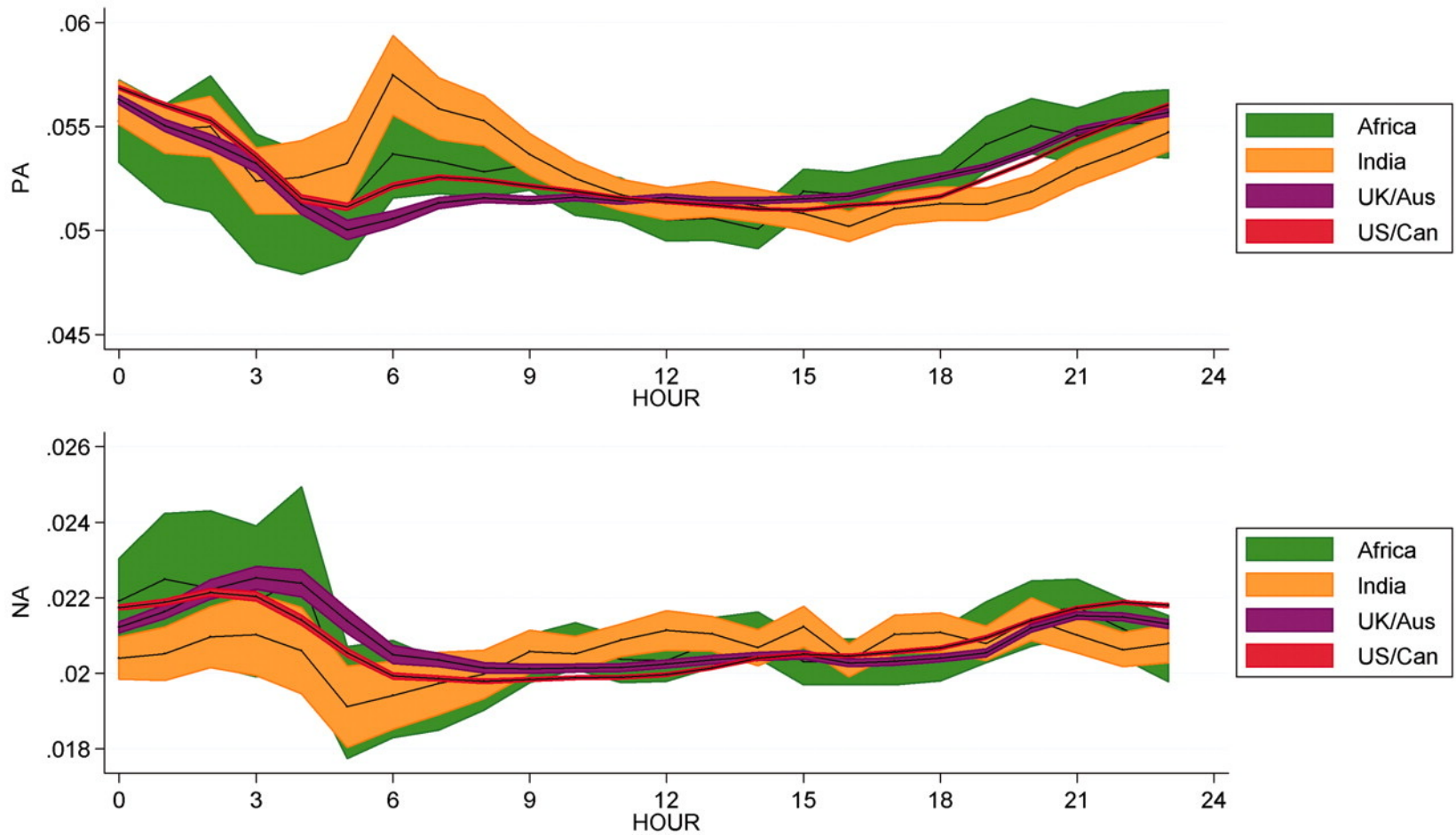
The within-individual PA score for a person-hour measures the signed difference between the person's score that hour and their baseline as defined in (2). Within-individual scores are comparable across people because individuals' baseline tendencies toward being upbeat or downbeat have been removed, leaving only the change over time that is within each individual:

$$WPA_u(h) = PA_u(h) - BPA_u + \frac{1}{\|UH\|} \sum_{(u,h) \in UH} PA_u(h) \quad (3)$$

where  $(u, h)$  pairs indicate user-hours and  $UH$  is the set of all such pairs in the dataset.<sup>1</sup>

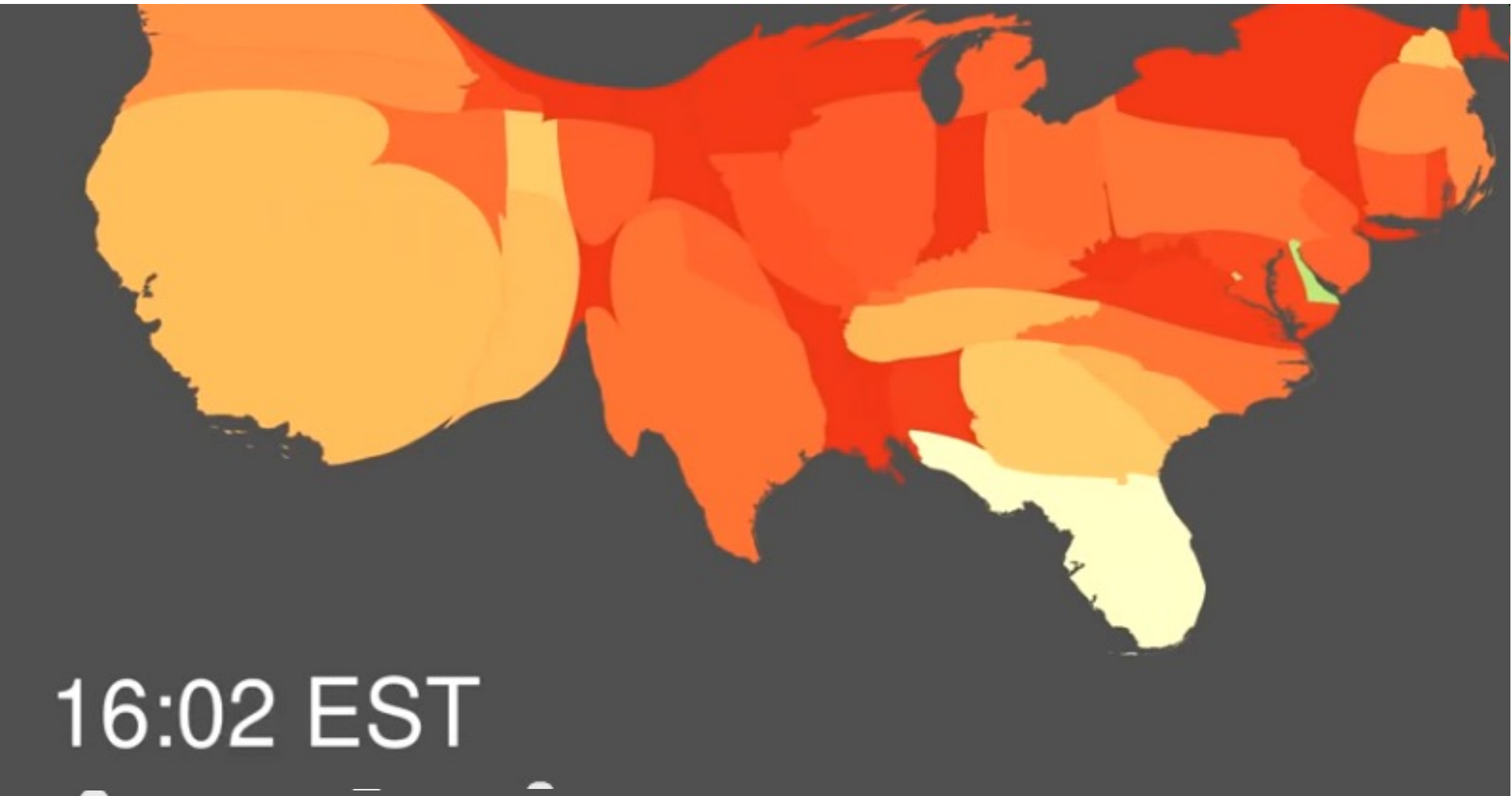


**Fig. 1 Hourly changes in individual affect broken down by day of the week (top, PA; bottom, NA). Each series shows mean affect (black lines) and 95% confidence interval (colored regions)**



**Fig. 2 Hourly changes in individual affect in four English-speaking regions.**  
 Each series shows mean affect (black lines) and 95% confidence interval (colored regions)

# visualizing twitter mood





# what to remember

## **twitter: an information network**

brevity has been a value across cultures & situations  
a network of weak links  
low reciprocity, highly asymmetric

## **large-scale human behavior**

language in tweets: short yet informative  
traces of human states like mood  
beware of biased data

**questions?**

daniel.gatica-perez@epfl.ch