

Title: Location and Language Use in Social Media

http://www.mpi-sws.org/~cristian/LACSS_2014.html

<http://acl2014.org/>

Title: Location and Language Use in Social Media

Abstract:

We now know that social interactions are critical in many knowledge and information processes. In this talk, I plan to illustrate a model-driven approach to understanding social behavior around user location and different languages in social media.

First, in 2010, we performed the first in-depth study of user location field in Twitter user profiles. We found that 34% of users did not provide real location information, frequently incorporating fake locations or sarcastic comments that can fool traditional geographic information tools. We then performed a simple machine learning experiment to determine whether we can identify a user's location by only looking at contents of a user's tweets. We found that a user's country and state can in fact be determined easily with decent accuracy, indicating that users implicitly reveal location information, with or without realizing it.

Second, despite the widespread adoption of Twitter in different locales, little research has investigated the differences among users of different languages. In prior research, the natural tendency has been to assume that the behaviors of English users generalize to other language users. We studied 62 million tweets collected over a four-week period. We discovered cross-language differences in adoption of features such as URLs, hashtags, mentions, replies, and retweets. We also found interesting patterns of how multi-lingual Twitter users broker information across these language boundaries. We discuss our work's implications for research on large-scale social systems and design of cross-cultural communication tools.

Bio: Ed H. Chi is a Staff Research Scientist at Google, focusing on social interaction research relating to social search, recommendation, annotations, and analytics. Previous to Google, he was the Area Manager and a Principal Scientist at Palo Alto Research Center's Augmented Social Cognition Group, where he led the group in understanding how Web2.0 and Social Computing systems help groups of people to remember, think and reason. Ed completed his three degrees (B.S., M.S., and Ph.D.) in 6.5 years from University of Minnesota, and has been doing research on user interface software systems since 1993. He has been featured and quoted in the press, including the Economist, Time Magazine, LA Times, and the Associated Press.

With over 20 patents and over 90 research articles, he is known for research in Web and online social sites, and the effects of social signals on user behavior. For example, he led a group of researchers at PARC to understand the underlying mechanisms in Wikipedia. He has also worked on information visualization, computational molecular biology, ubicomp, and recommendation/search engines, and has won awards for both teaching and research. In his spare time, Ed is an avid photographer and snowboarder.



Location and Language Use in Social Media

Ed H. Chi (edchi@)
Staff Research Scientist, Google Research



Talk in 2 Parts

1. User Behavior in Location Disclosure in Twitter

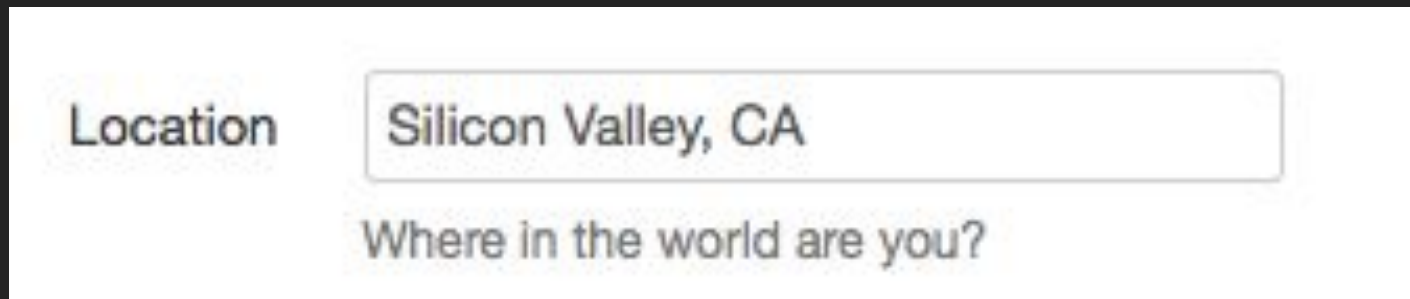
[Hecht et al., CHI2011]

2. Break Language Barriers in Social Media

[Hong et al., ICWSM2011]

“Locations” Signals in Social Media

Hecht, Hong, Suh & Chi, CHI 2011



A screenshot of a social media interface showing a location input field. The field is labeled "Location" and contains the text "Silicon Valley, CA". Below the input field, the text "Where in the world are you?" is displayed.



Your Occupation:

Your Hometown:

City you live in now:

Country:



facebook.

Brent Hecht » Edit Profile

Current City:  Evanston, Illinois

Hometown:

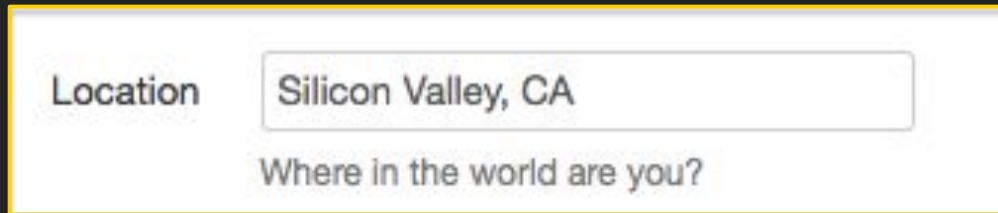
I Am: Male

Birthday: Apr 29 1983



Varying Forms of Location Field

Twitter



A screenshot of the Twitter location field. It features a label "Location" followed by a text input box containing "Silicon Valley, CA". Below the input box is the text "Where in the world are you?".

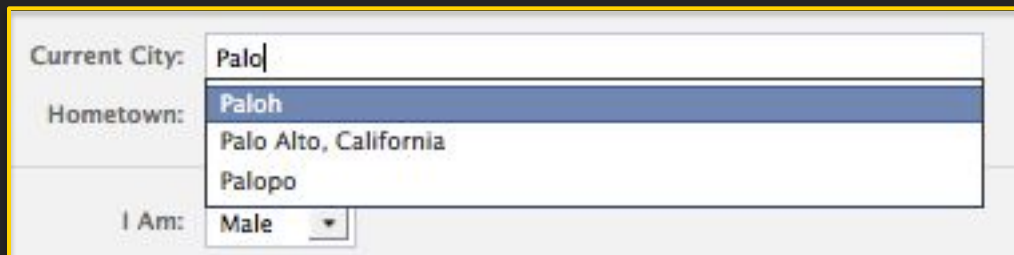
- Free-form
- Current location

Google

- Free-form
- Multiple locations



A screenshot of the Google "Places I've lived" interface. It shows a list of locations with location pins: "Silicon Valley, CA", "Fujian, China", and "type a city name". Each entry has a close button (X) to its right. At the bottom are "Save" and "Cancel" buttons.



A screenshot of the Facebook location field. It shows a "Current City:" label and a text input box with "Palo". Below it is a "Hometown:" label and a dropdown menu with "Palo" selected. Other options in the dropdown are "Palo Alto, California" and "Palopo". At the bottom is an "I Am:" label and a dropdown menu with "Male" selected.

Facebook

- Limited options,
no "Bay Area, CA"
- 2 locations

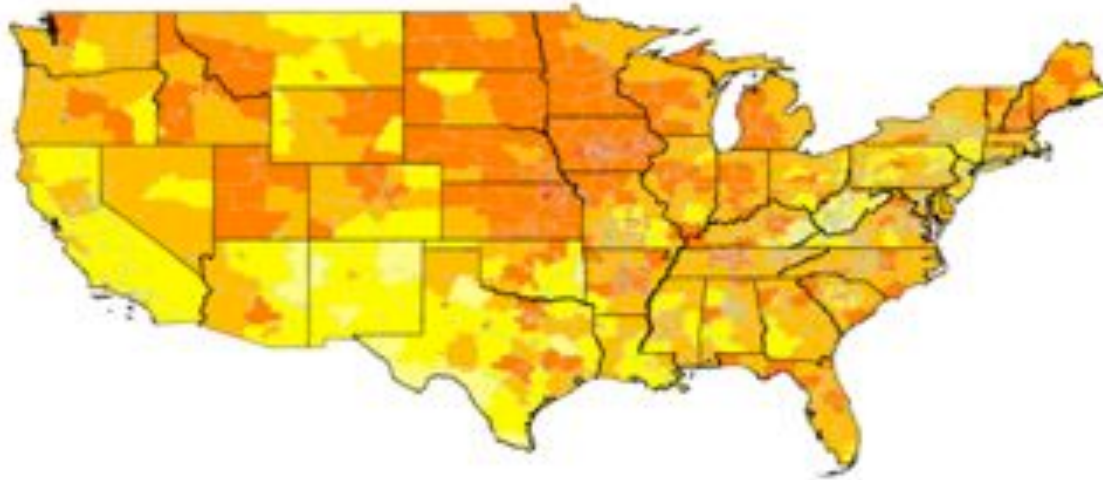
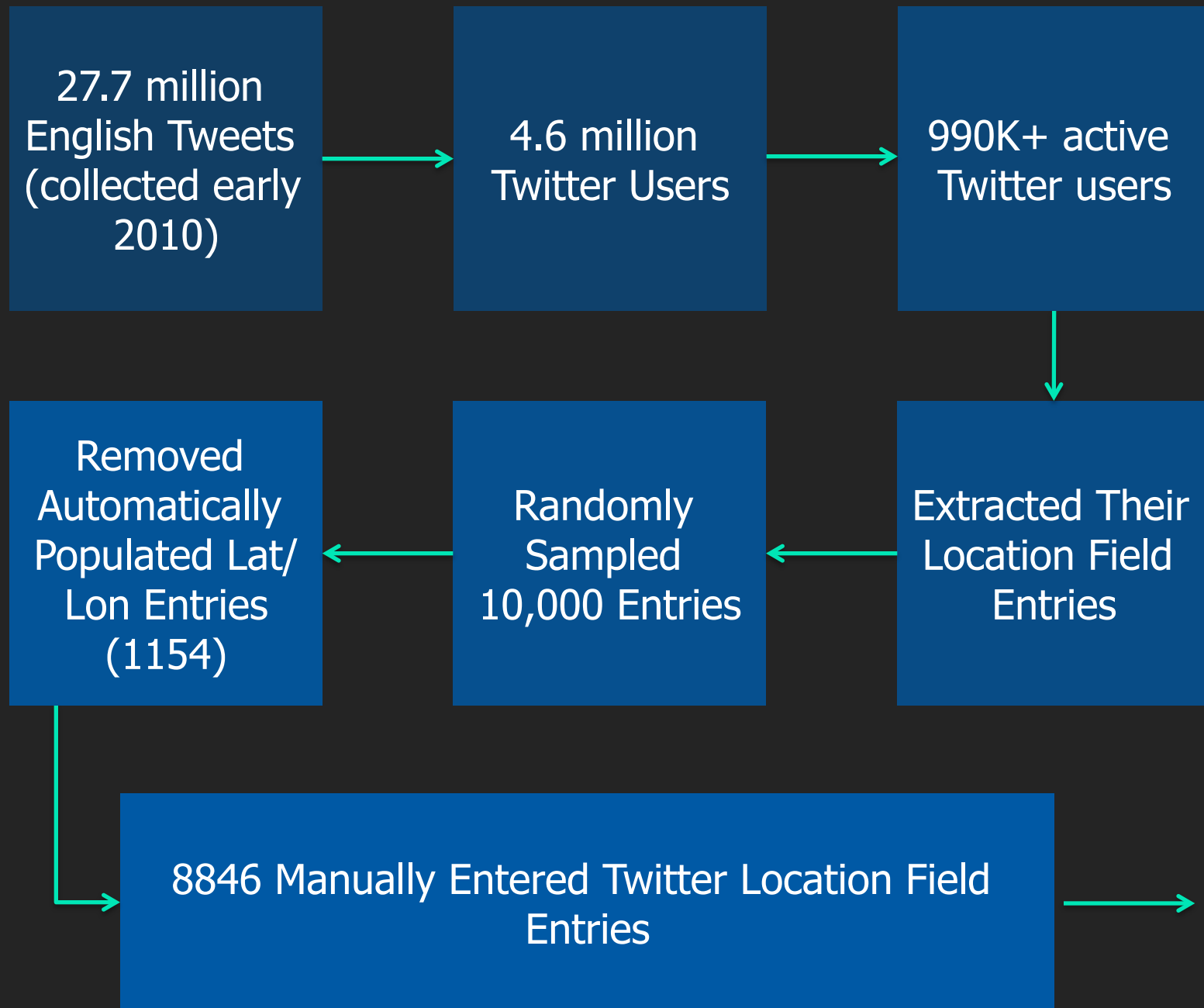


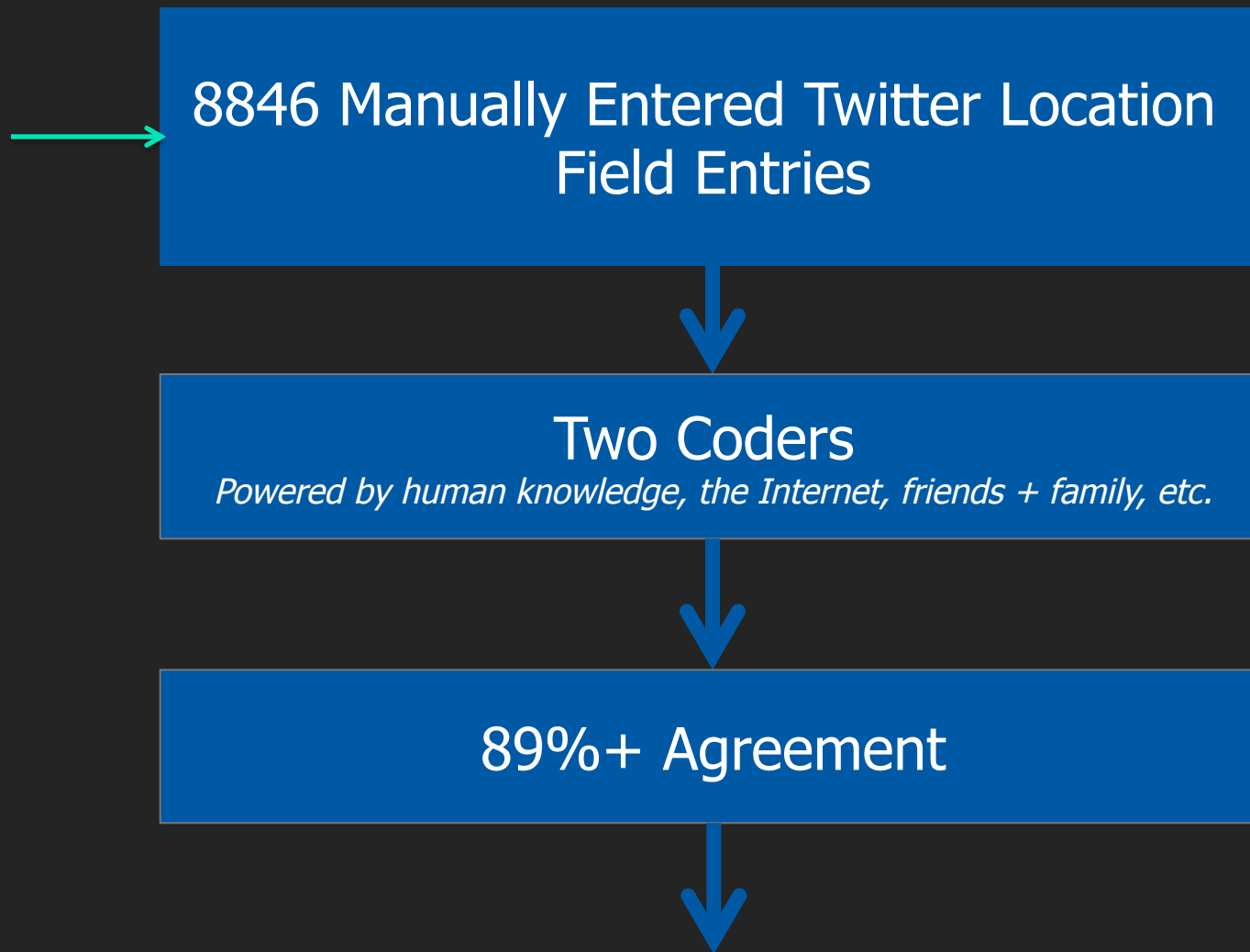
Figure 3: *Facebook penetration using user-provided addresses.* As a proportion of population, users in the midwest share more addresses on Facebook. However, this corresponds closely to overall Facebook penetration, shown in the next figure.

Backstrom et al. 2010

Assumptions about the Location Field

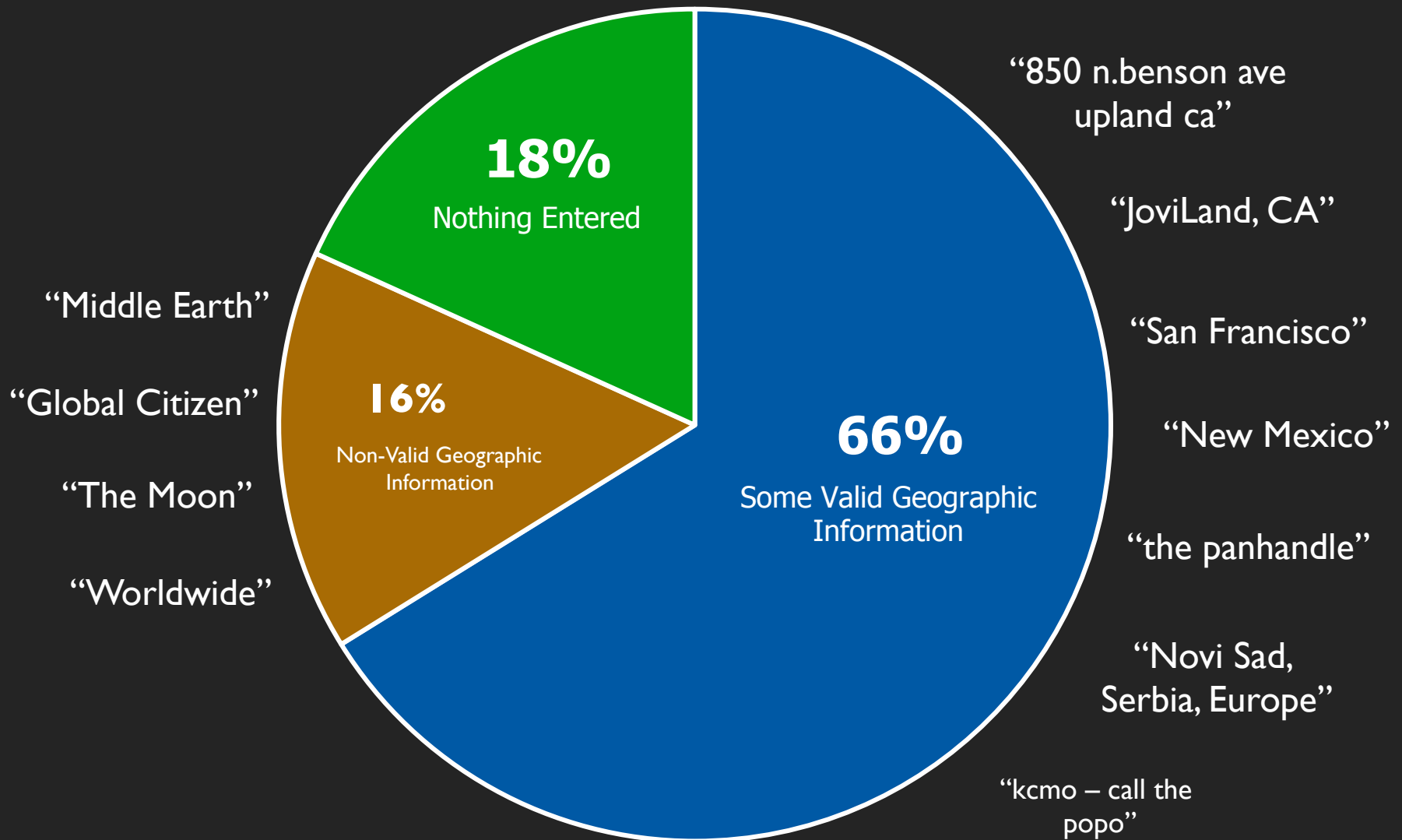
1. Strongly-typed geo information
2. Little noise
3. Good precision





Study I: “Geographicness”

data quality of the location field



Study 1: Non-Geo Information

types of non-geographic information entered into the location field

Information Type	# of Users
Popular Culture Reference	195 (12.9%)
Privacy-Oriented	18 (1.2%)
Insulting or Threatening to Reader	69 (4.6%)
Non-Earth Location	75 (5.0%)
Negative Emotion Towards Current Location	48 (3.2%)
Sexual in Nature	49 (3.2%)

Study 1: Non-Geo Information

types of non-geographic information entered into the location field

Information Type	# of Users
Popular Culture Reference	195 (12.9%)
Privacy-Oriented	18 (1.2%)
Insulting or Threatening to Reader	69 (4.6%)
Non-Earth Location	75 (5.0%)
Negative Emotion Towards Current Location	48 (3.2%)
Sexual in Nature	49 (3.2%)

Study 1: Popular Culture References

Non-geographic information in the location field in user's profiles

“BieberTown”

“My World”

“belieber wonderland”

“JaeJoongs heart”

“Next to Waldo :D”

“somewhere in Glambertville”

“Los Angeles, 2019 (GET IT?)”

“Schrute Farms”

Study 1: Privacy References

Non-geographic information in the location field in user's profiles

“Stalker City”

“Stalking me here isnt enough?”

“MindingMyOwn”

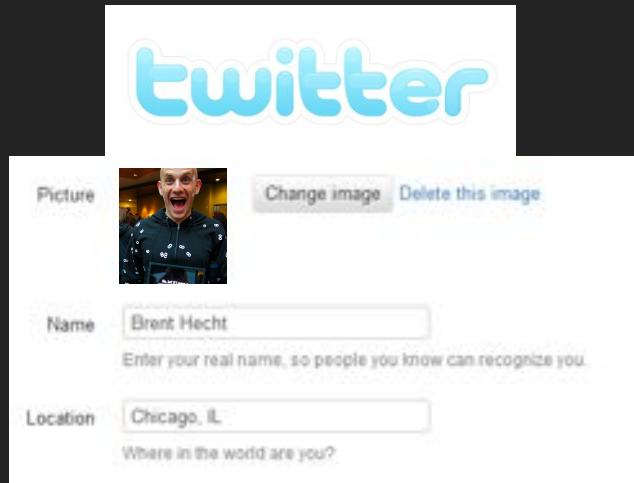
“For me to know n u to find out”

“NONE YA BISNESS”

“UM...STALKER!!”

“kgb answers”

Study 1: Implications



A screenshot of the Twitter profile creation form. At the top is the Twitter logo. Below it is a profile picture placeholder with a 'Change image' button and a 'Delete this image' link. The 'Name' field contains 'Brent Hecht' with a subtext 'Enter your real name, so people you know can recognize you.' The 'Location' field contains 'Chicago, IL' with a subtext 'Where in the world are you?'.

**STRONGLY-TYPED
GEOGRAPHIC
INFORMATION
REQUIRED**



Geocoder



Latitude and
Longitude
Coordinates

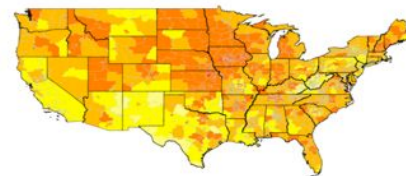
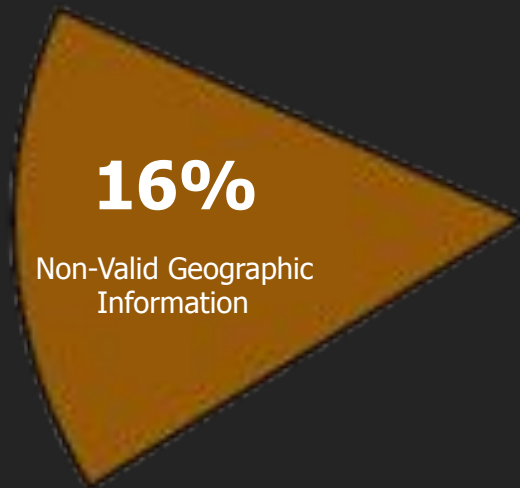


Figure 3: Facebook penetration using user-provided addresses. As a proportion of population, users in the midwest share more addresses on Facebook. However, this corresponds closely to overall Facebook penetration, shown in the next figure.

Study 1: Quality Implications

**STRONGLY-TYPED
GEOGRAPHIC
INFORMATION
REQUIRED**



Geocoder

Latitude and
Longitude
Coordinates

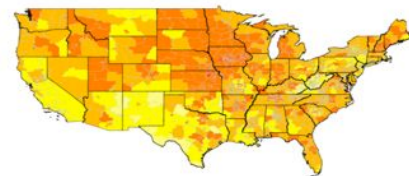
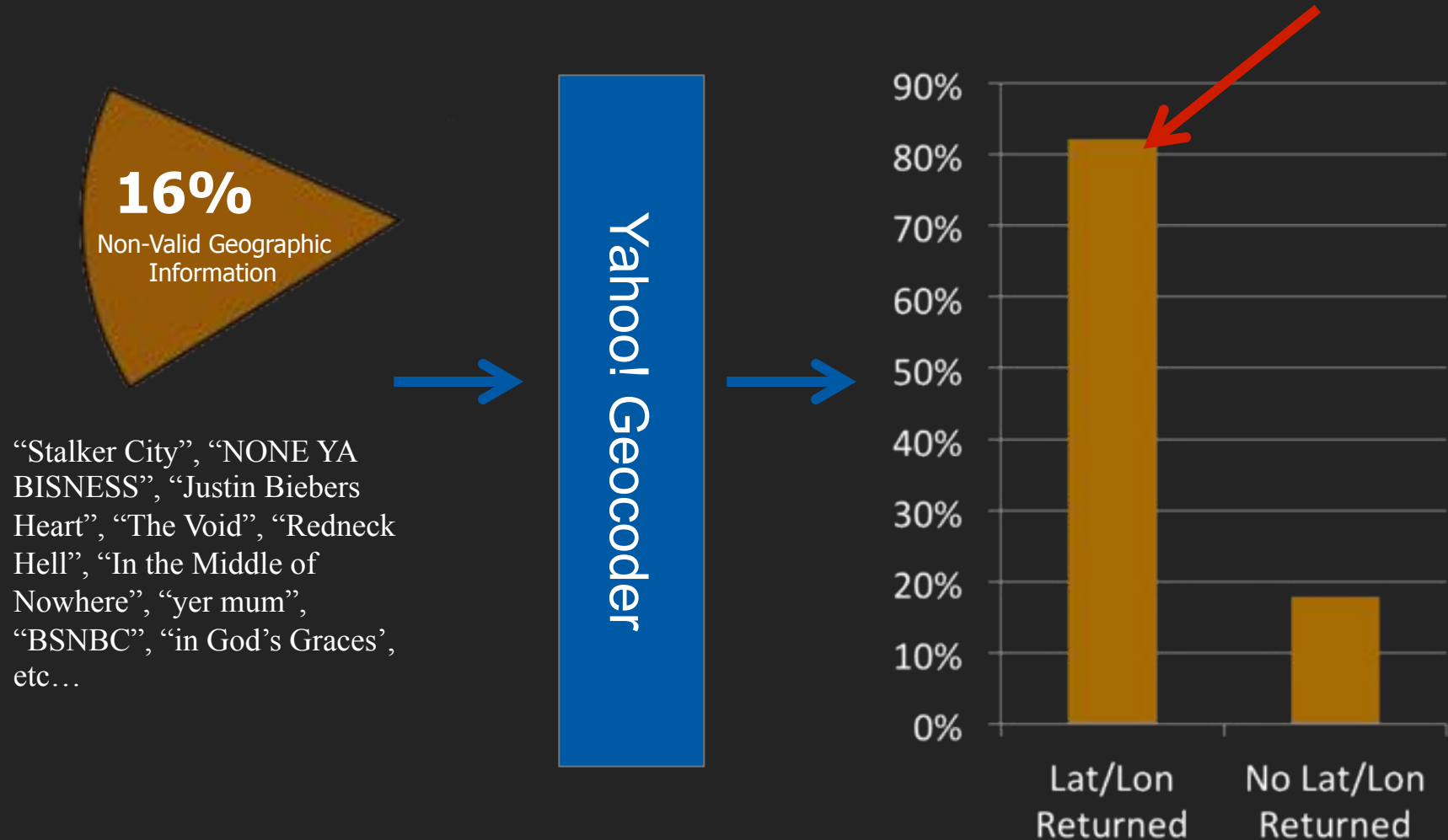


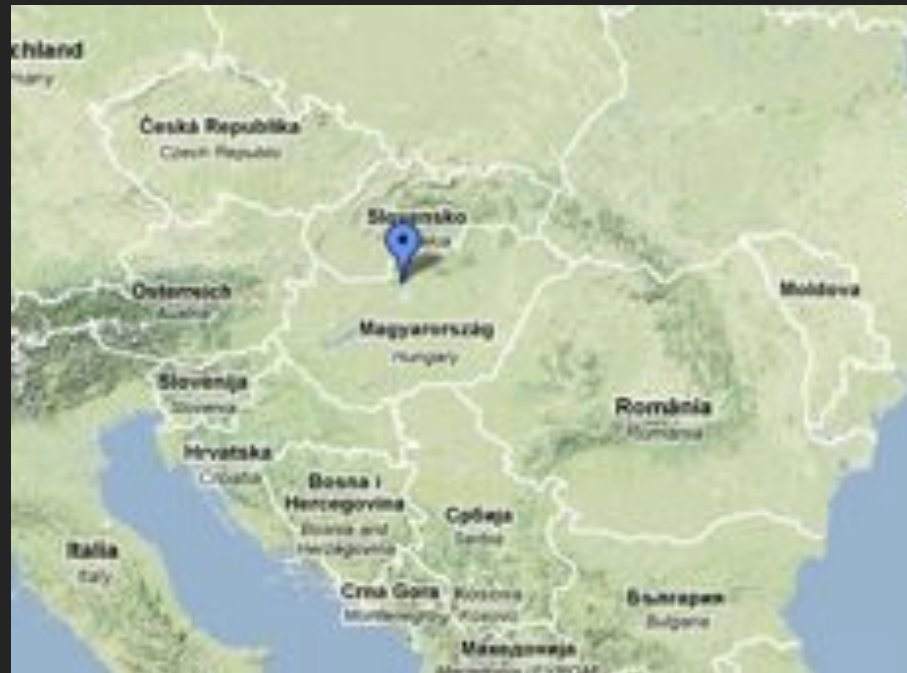
Figure 3: Facebook penetration using user-provided addresses. As a proportion of population, users in the midwest share more addresses on Facebook. However, this corresponds closely to overall Facebook penetration, shown in the next figure.

Study 1: Quality Implications





“Loserville :)”
(-71.397524, 42.28904)

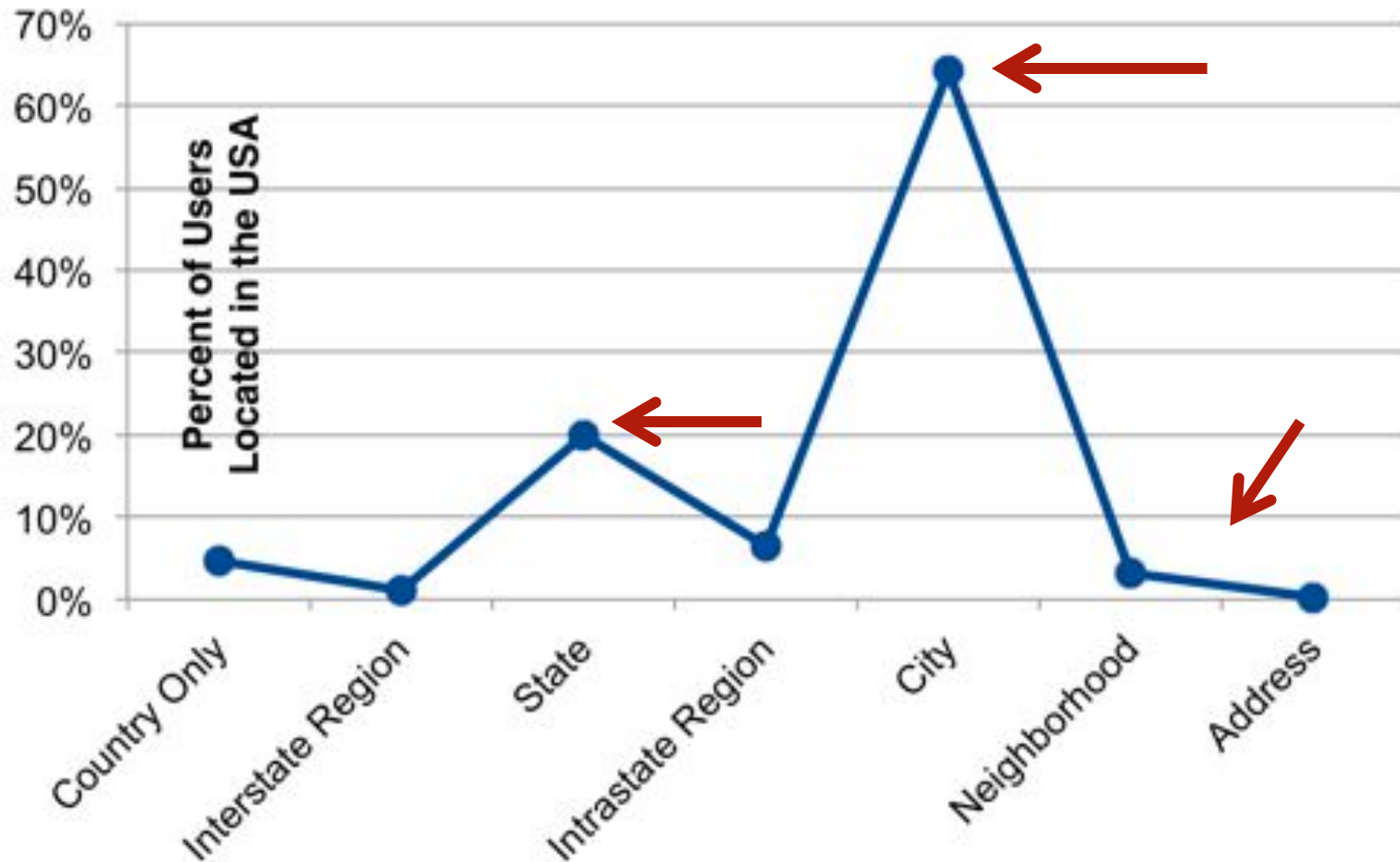


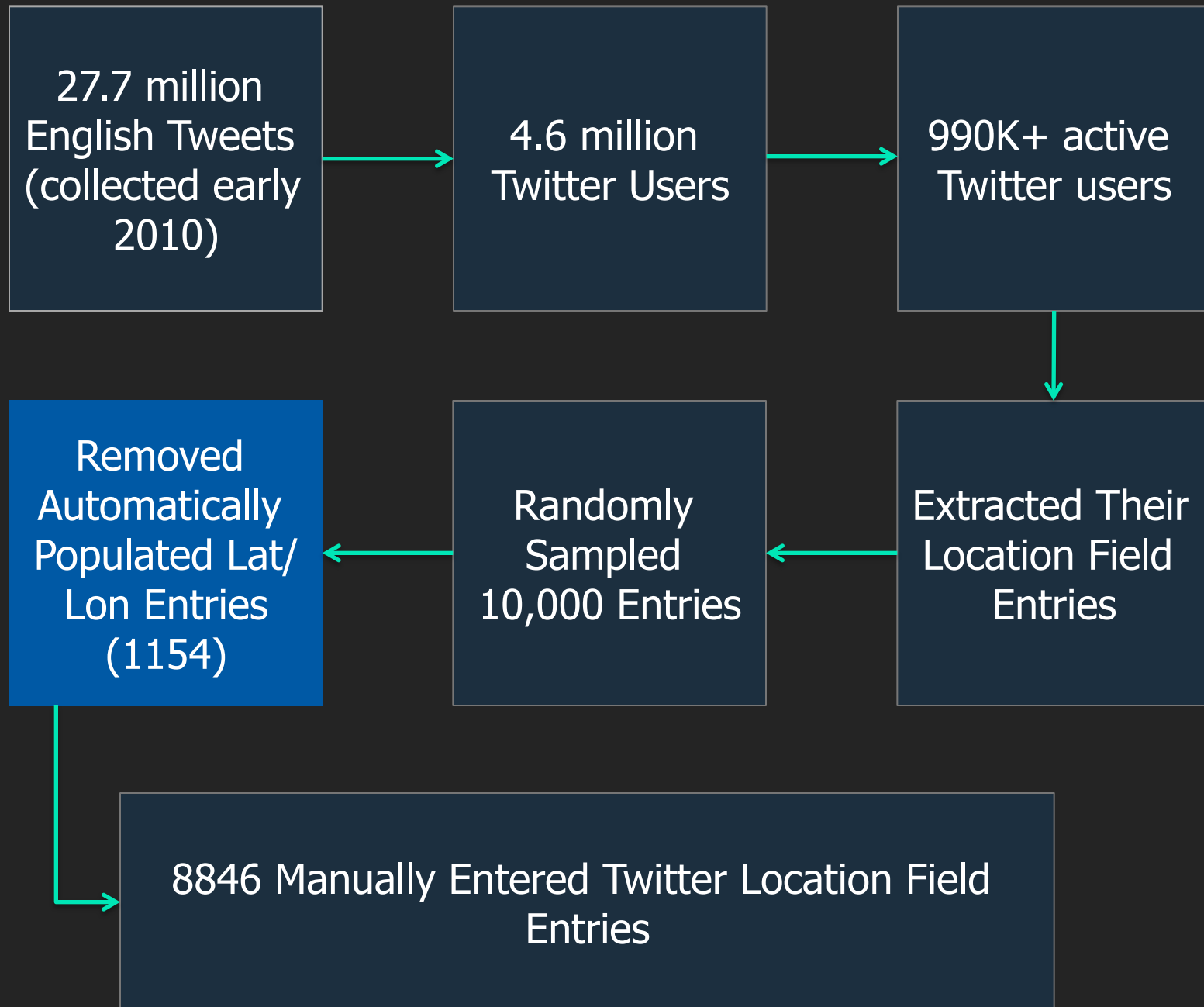
"With God"
(19.13683,47.705132)



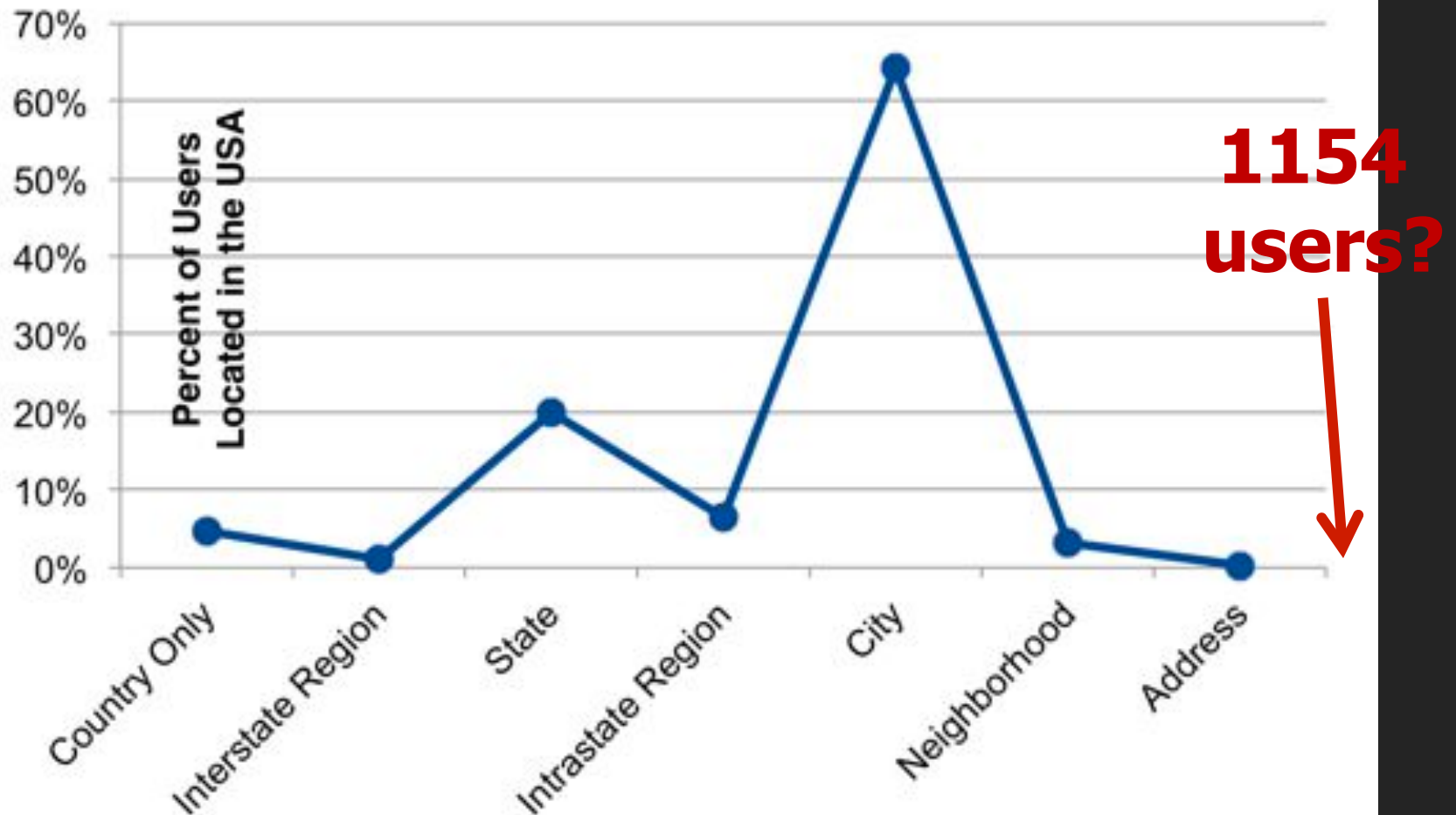
“Justin Biebers heart!”
(-91.700189, 36.328785)

Geographic Scale (In Order of Increasing Localness from Left to Right)





33.687456,-84.244945
Seriously?





edchi edchi

wonderful afternoon at the 20th anni. of HCI at [StanfordU](#). Terry Winograd toasted by many, inc. my ex and current boss: S.Card and L.Page.

21 Feb



edchi edchi

Disagree. RT @businessinsider: ex Sun CEO: [Silicon Valley](#) No Longer Best Place To Start Biz. Gov Crushing Job Growth
<http://read.bi/hWoEff>

13 Feb



edchi edchi

NYTimes: [U.S.](#) Urged to Raise Teachers' Status <http://nyti.ms/dDXnPv>

16 Mar



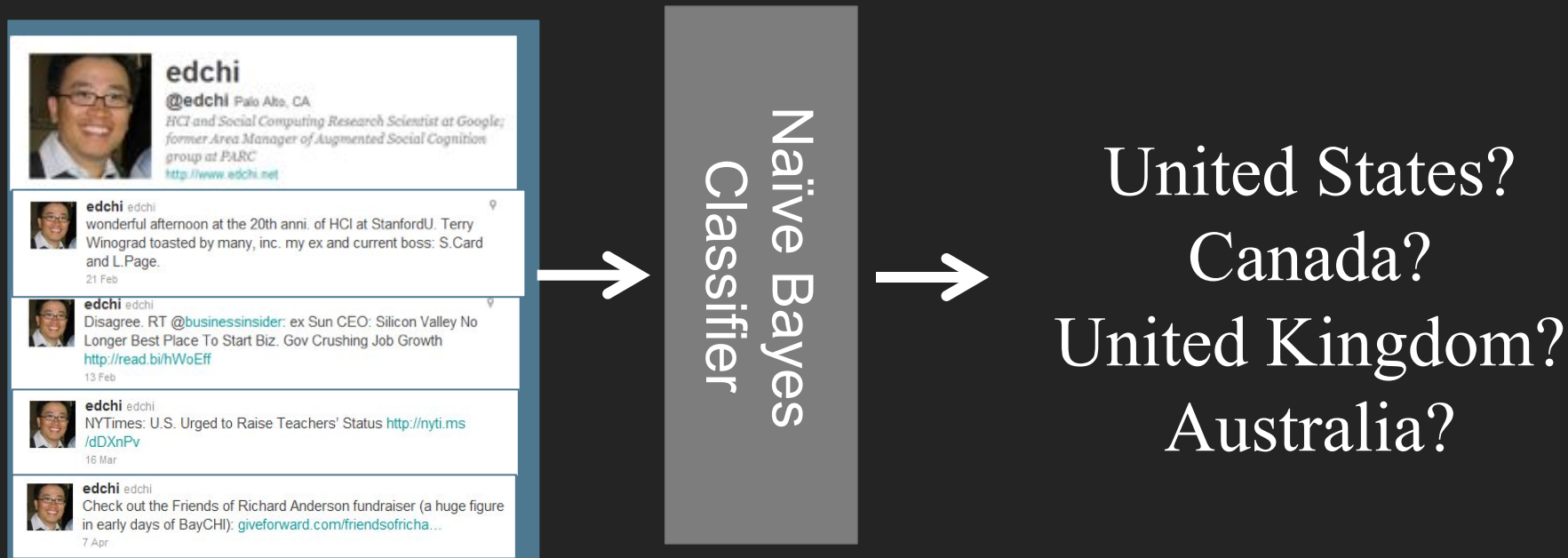
edchi edchi

Check out the Friends of Richard Anderson fundraiser (a huge figure in early days of [BayCHI](#)): giveforward.com/friendsofricha...

7 Apr

Study 2: Country Experiments

Uniform Sampling



72.10% Accuracy

2.91x better than random

Study 2: State Experiments

Uniform Sampling



Naïve Bayes
Classifier



California?
Arkansas?
New York?
Washington?
Texas?

...

30.28% Accuracy

5.45x better than random

Study 2: Predictive Words

Word	Geography	Predictiveness
calgary	Canada	419.42
brisbane	Australia	137.29
coolcanuck	Canada	78.28
afl	Australia	56.24
clegg	UK	35.49
cbc	Canada	29.40
yelp	USA	19.80

Word	Geography	Predictiveness
elk	Colorado	90.74
redsox	Massachusetts	41.18
biggbi	Michigan	24.26
gamecock	South Carolina	16.00
crawfish	Louisiana	14.87

1.81x better than random

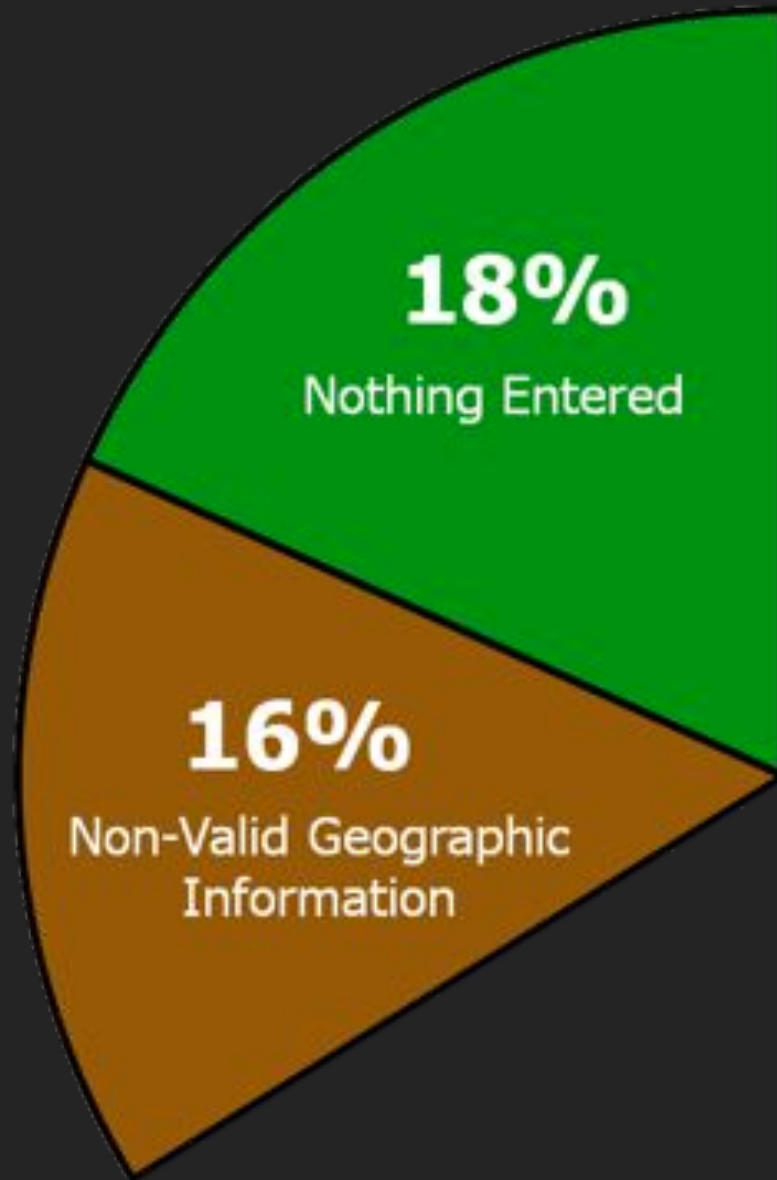
5.45x better than random

1.08x better than random

2.91x better than random



Tweets Have Implicit Location Information



This needs to be considered in the context of **implicit** location disclosure



Information Transmission across Language Barriers

Joint work with
Lichan Hong, Gregorio Convertino
[ICWSM July 2011]

* Work done at Xerox PARC

THIS IS ME....









SOCIAL MEDIA – BRINGING WORLD TOGETHER?!

- I speak 3 different languages fluently.
 - English
 - Mandarin
 - Cantonese
- And 3 other languages badly:
 - Taiwanese
 - German
 - Japanese



SOCIAL MEDIA – BRINGING WORLD TOGETHER?!



艾未未 Ai Weiwei @aiww

47m

回避事实RT @yancaiwm: 虽然觉得她的回复很无聊，但还是转给艾总。||猪油渣：你告诉他，如果有些人受到控制，必定还有一些人在控制，有些人是把别人当作达成个人目的的手段。我根本看不起一个向恶法低头的人，更不想成为他表演正义的道具。一个公众人物，与其千呼万唤，不如街头



Alan @GammaCounter

2h

Wasabi fire alarm. All of the deaf people tested, exposed to the odor of wasabi, woke up within 2.5 minutes. dvr.it/1L6324 @gohsuket

↻ Retweeted by aquigley



Alireza Sahami @alirezasahami

49m

ایران نوروزت ، و نوروزت ای ایران جاودانه باد . / Happy Nowruz



Search Update @SearchUpdate

49m

DATA COLLECTION & PROCESSING

Twitter stream

04/18/10-05/16/10 (4 weeks)



62M tweets



104 languages



Top 10
languages



Google Language API & LingPipe

TOP 10 LANGUAGES IN TWITTER

Language	Tweets	%	Users
English	31,952,964	51.1	5,282,657
Japanese	11,975,429	19.1	1,335,074
Portuguese	5,993,584	9.6	993,083
Indonesian	3,483,842	5.6	338,116
Spanish	2,931,025	4.7	706,522
Dutch	883,942	1.4	247,529
Korean	754,189	1.2	116,506
French	603,706	1.0	261,481
German	588,409	1.0	192,477
Malay	559,381	0.9	180,147

[Hong, Convertino, Chi. ICWSM July 2011]

ACCURACY OF LANGUAGE DETECTION

- Two Types of Errors
 - *Got ur dirct msg.i'm lukng 4 wrd 2 twt wit u too.so,wat doing ha...*(detected as Afrikaans)
 - High error rate for tweets of 1~2 words


COMMON TWITTER CONVENTIONS

 **sharoda** sharoda
Recently accepted **#icwsm11** poster on question asking and answering on Twitter (pdf): <http://bit.ly/f1UCPZ> (with **@koozda**, **@edchi**)


hashtag

mention

URL

 **edchi** edchi
@msbernst Haha! Glad to hear I had an effect---Super Geeky Coolness!

reply (per-tweet metadata)

 **koozda** Lichan Hong
Very interesting and relevant to our work! RT **@TechCrunch**: Look Out Quora, InboxQ Takes Q&A Off-Site And C... (cont)
<http://deck.ly/~Au74a>

retweet

USE OF URLS IN 62M TWEETS

Language	URLs
<i>All</i>	21%
English	25%
Japanese	13%
Portuguese	13%
Indonesian	13%
Spanish	15%
Dutch	17%
Korean	17%
French	37%
German	39%
Malay	17%

Chi Square tests confirmed that differences by language are significant.

SIGNIFICANT CROSS-LANGUAGE DIFFERENCES

Language	URLs	Hashtags	Mentions	Replies	Retweets
<i>All</i>	21%	11%	49%	31%	13%
English	25%	14%	47%	29%	13%
Japanese	13%	5%	43%	33%	7%
Portuguese	13%	12%	50%	32%	12%
Indonesian	13%	5%	72%	20%	39%
Spanish	15%	11%	58%	39%	14%
Dutch	17%	13%	50%	35%	11%
Korean	17%	11%	73%	59%	11%
French	37%	12%	48%	36%	9%
German	39%	18%	36%	25%	8%
Malay	17%	5%	62%	23%	29%

Chi Square tests confirmed that differences by language are significant

Design IMPLICATIONS

Language	URLs	Hashtags	Mentions	Replies	Retweets
<i>All</i>	21%	11%	49%	31%	13%
Korean	17%	11%	73%	59%	11%
German	39%	18%	36%	25%	8%

- Use of Social Media for social networking vs. information sharing
 - Different in different languages
- Design of recommendation engines
 - Korean users: promote conversational tweets
 - German users: promote tweets with URLs

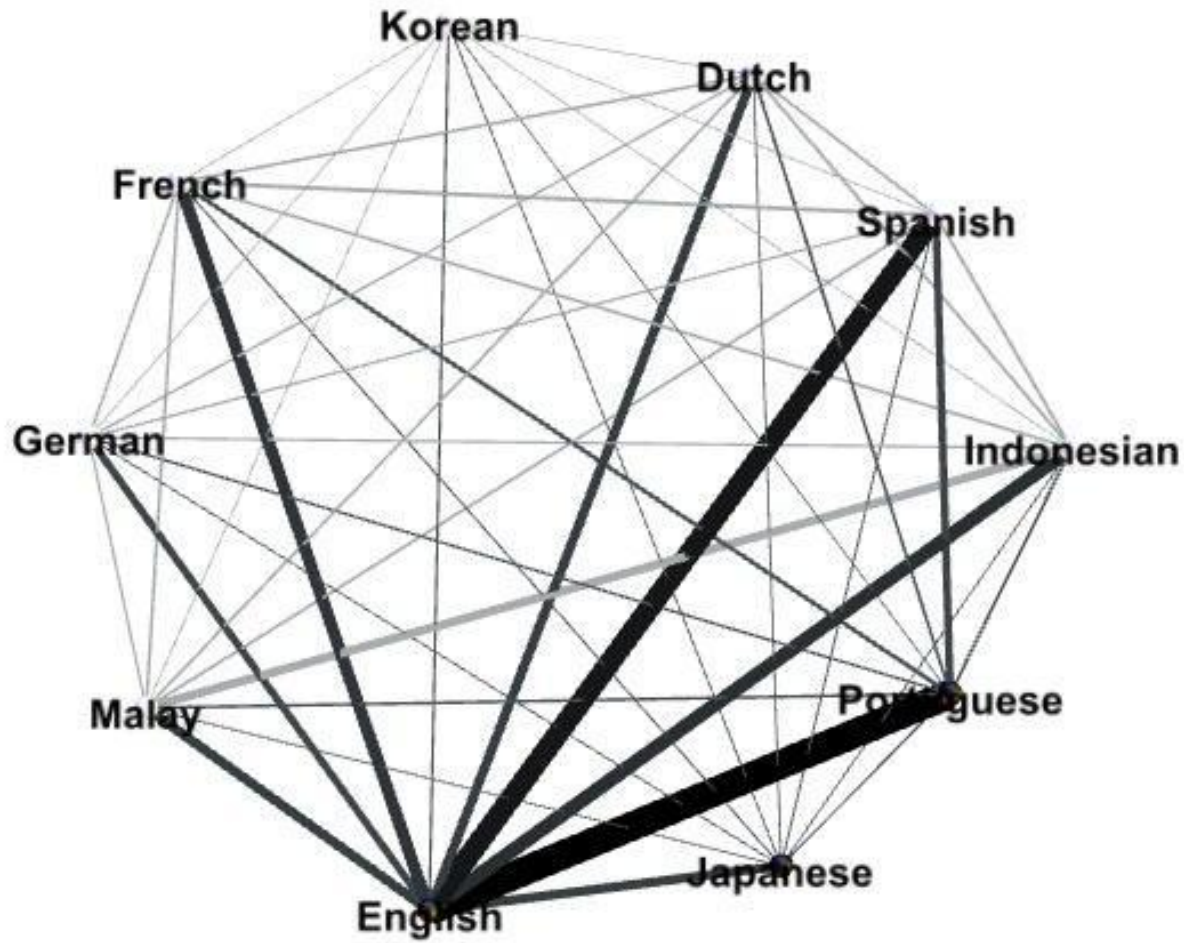
IDEAS BRIDGED BY BROKERS

- Importance of brokers. [Structural holes, Burt'92]
- Define bilingual brokers as Users who tweeted in a pair of languages

NUMBER OF BILINGUAL BROKERS

	E	J	P	I	S	D	K	F	G
J	140,730								
P	488,545	13,228							
I	230,023	4,825	29,405						
S	359,117	10,139	112,524	36,068					
D	150,041	6,383	30,855	34,906	30,916				
K	19,722	6,384	906	2,014	1,109	972			
F	194,931	10,463	53,607	34,586	49,445	33,568	1,244		
G	110,748	6,053	22,106	21,471	21,989	22,162	786	24,763	
M	148,365	4,208	31,184	135,427	31,967	29,331	1,518	30,257	18,301

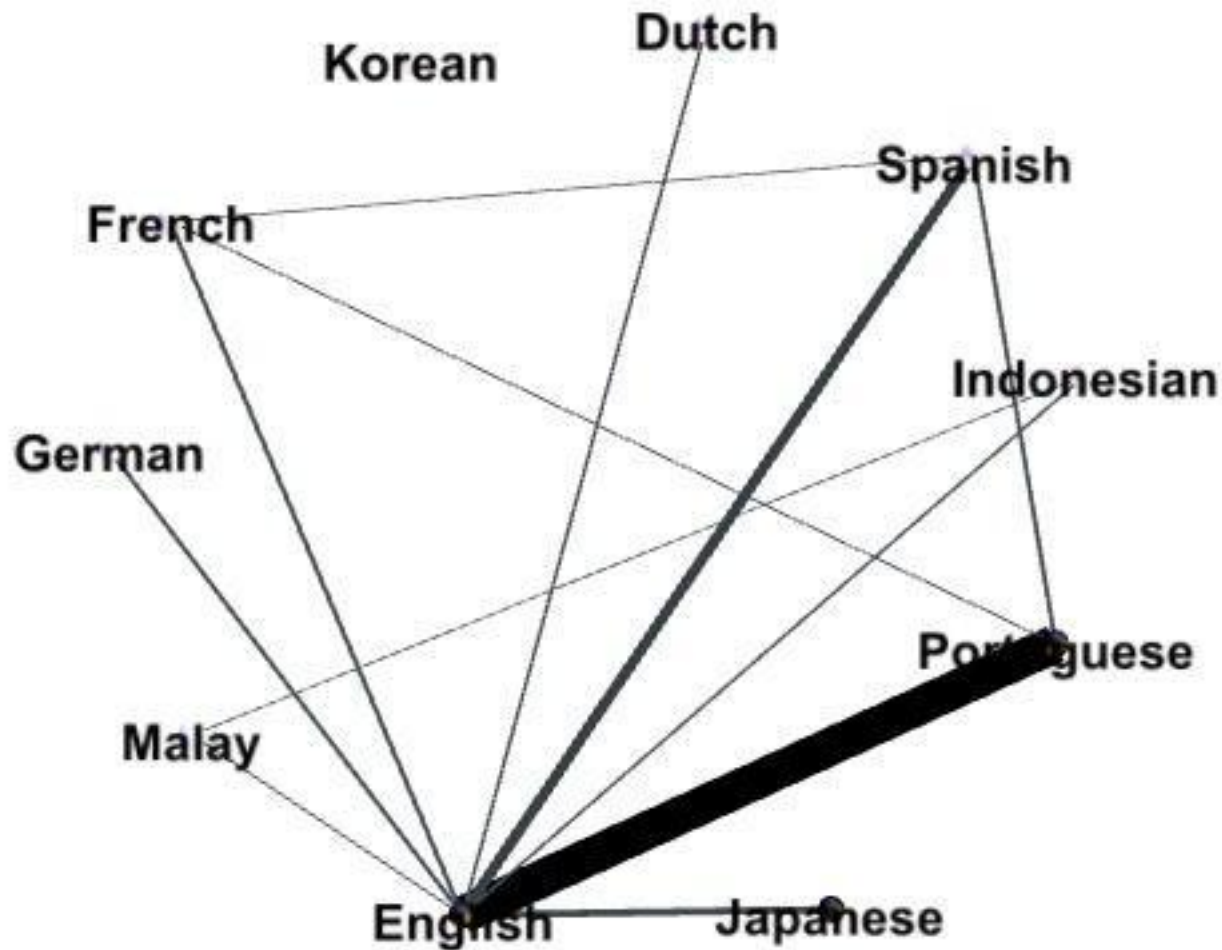
NUMBER OF BILINGUAL BROKERS



SHARING URLs ACROSS LANGUAGES

	E	J	P	I	S	D	K	F	G	M
E		3,013	18,399	985	4,986	1,144	212	1,791	1,647	540
J	3,013		77	37	58	29	43	59	46	18
P	18,399	77		74	1,644	198	2	453	168	123
I	985	37	74		67	64	1	53	38	279
S	4,986	58	1,644	67		139	0	286	139	53
D	1,144	29	198	64	139		2	112	126	48
K	212	43	2	1	0	2		3	3	1
F	1,791	59	453	53	286	112	3		157	53
G	1,647	46	168	38	139	126	3	157		40
M	540	18	123	279	53	48	1	53	40	

SHARING URLs ACROSS LANGUAGES

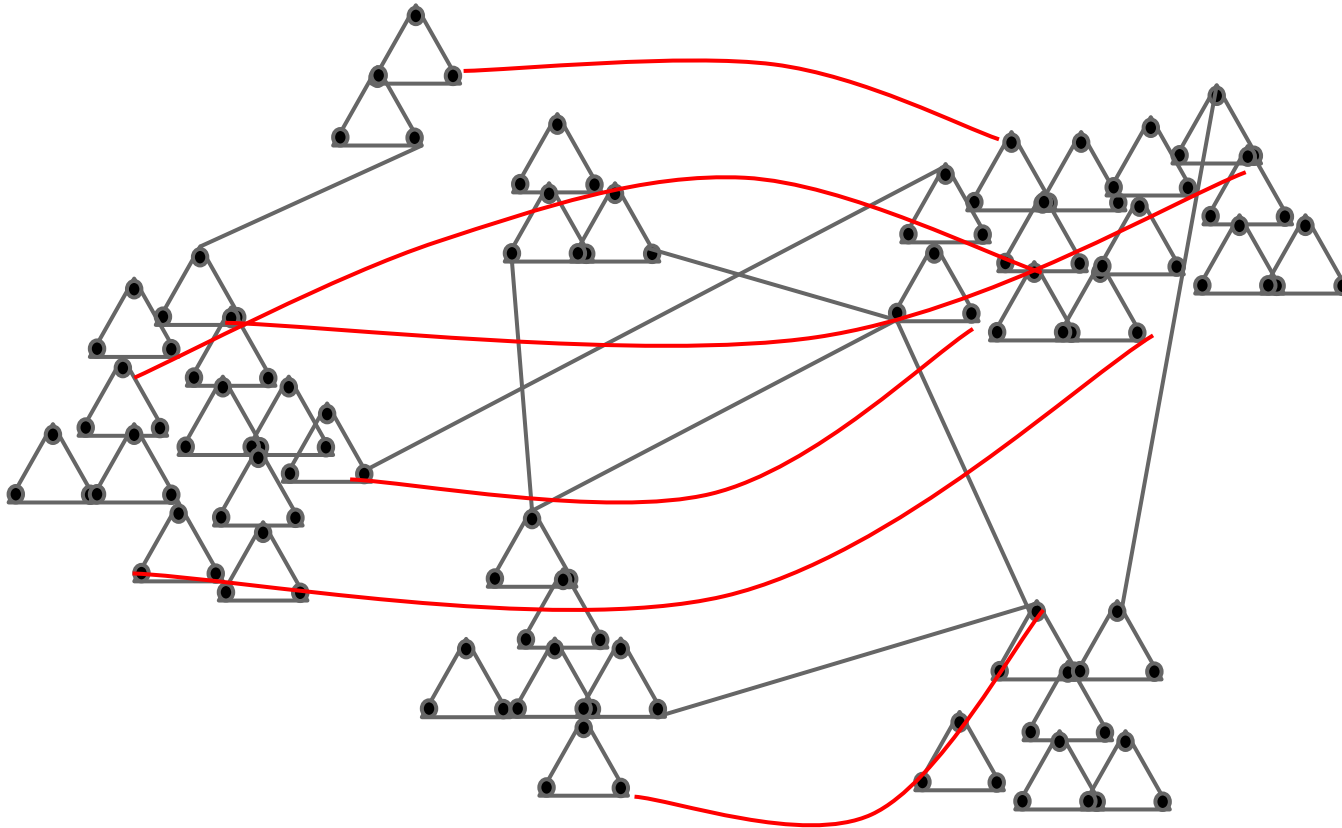


IMPLICATIONS

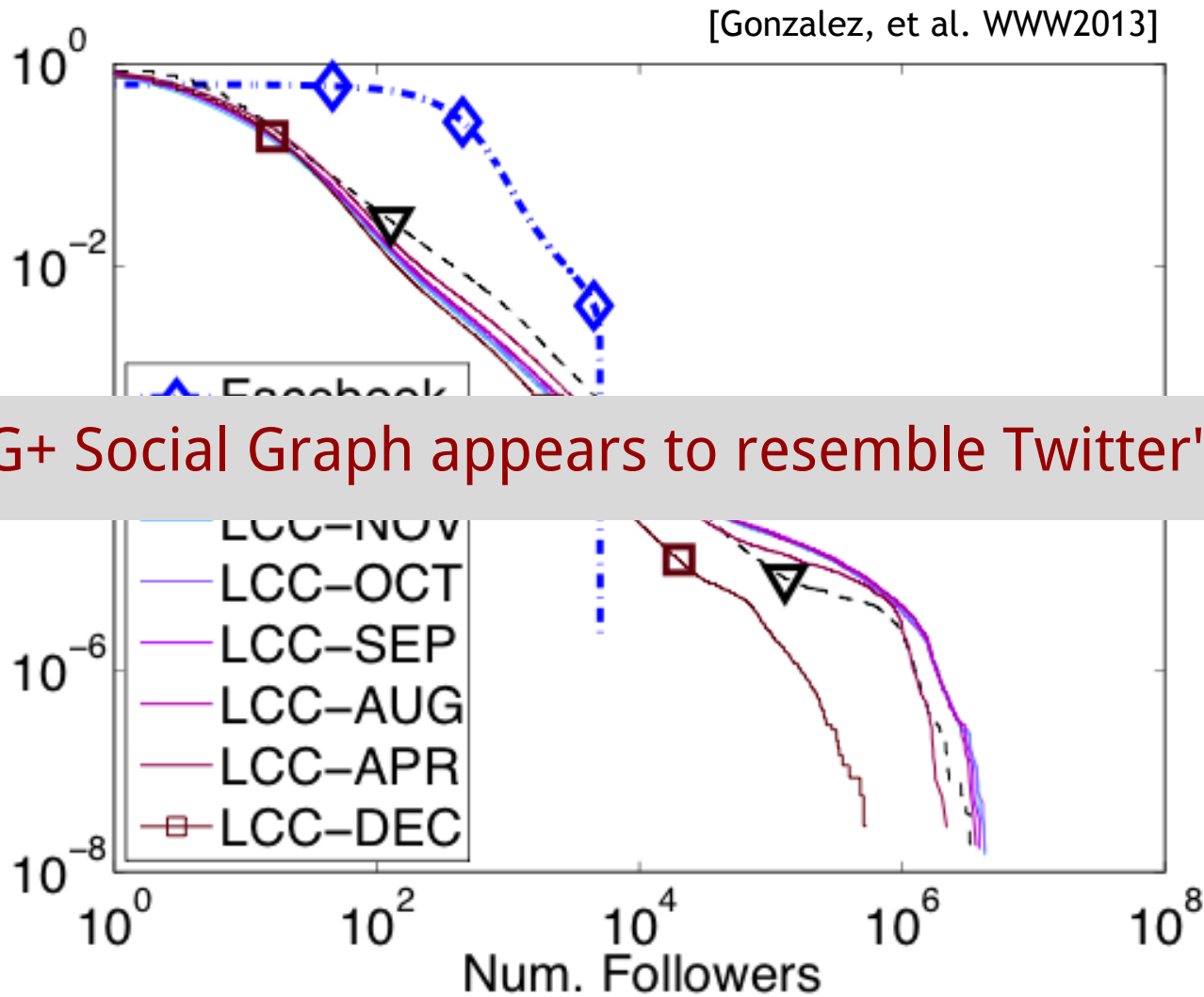
- Need to improve cross-language brokerage and communication



Reduce Isolation



Comparing to Twitter's graph structure



GIANT EXPERIMENT

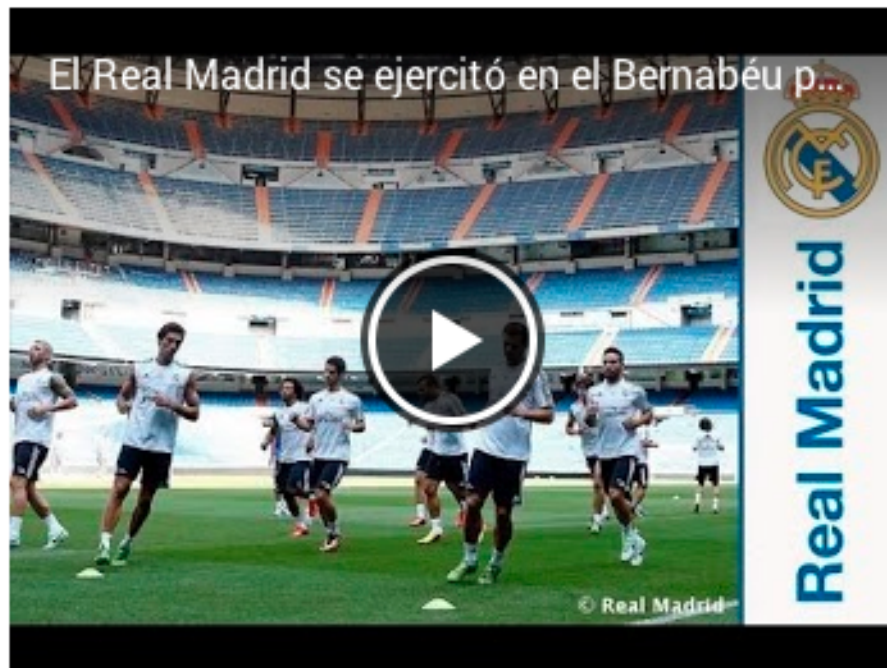


Real Madrid C.F.

Shared publicly - Aug 16, 2013

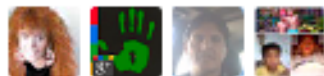
El Real Madrid se ejercitó en el Bernabéu por primera vez esta temporada

[Translate](#)



+231

8

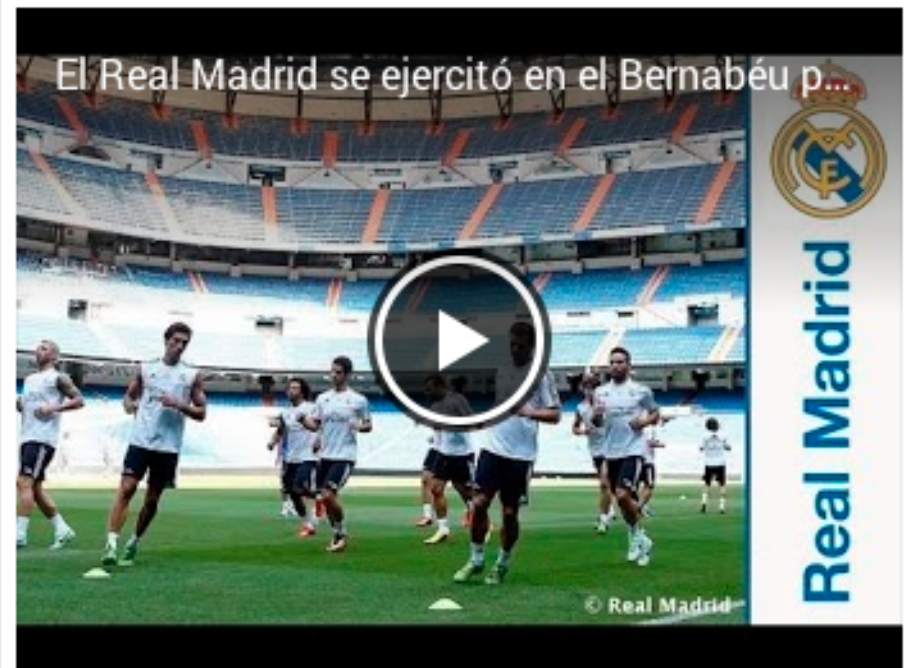


Real Madrid C.F.

Shared publicly - Aug 16, 2013

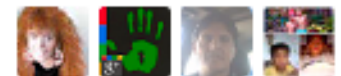
The Real Madrid trained at the Bernabeu for the first time this season

[Show original text](#)



+231

8





Talk in 2 Parts

1. User Behavior in Location Disclosure in Twitter

[Hecht et al., CHI2011]

2. Break Language Barriers in Social Media

[Hong et al., ICWSM2011]

User Behavior in Social Media

- Users adopt and adapt social media to suit their needs.
- Designing for the variety of behavior is critical to the success of social media.
- What we learned:
 - Location field can be quite expressive;
 - Language affects use cases;
 - Language can be a barrier to expression and information brokerage.



Thank you! Questions?

Social Interactions Research @ Google

Contact: edchi@google.com

Position Information

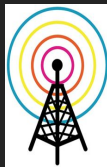
Sensor-based



Global Positioning System (GPS)



WiFi Access Point



Cell Phone Towers

Self-reported Implicitly Revealed!



Study 2: Country Experiments

Demographically Proportional Sampling



Naïve Bayes
Classifier



United States?
Canada?
United Kingdom?
Australia?

88.86% Accuracy

1.08x better than random

Study 2: State Experiments

Demographically Proportional Sampling



Naïve Bayes
Classifier



California?
Arkansas?
New York?
Washington?
Texas?

...

27.31% Accuracy

1.81x better than random