Big Data Analytics

FOSCA GIANNOTTI AND LUCA PAPPALARDO

HTTP://DIDAWIKI.DI.UNIPI.IT/DOKU.PHP/BIGDATAANALYTICS/BDA/

DIPARTIMENTO DI INFORMATICA - Università di Pisa
anno accademico 2018/2019

Suggested Bibliography

GoogleFluTrend

Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents, Peter Sheridan Dodds Æ Christopher M. Danforth, J Happiness Stud (2010)

The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place Lewis Mitchell1 , Morgan R. Frank, Kameron Decker Harris1,2, Peter Sheridan Dodds, Christopher M. Danforth1

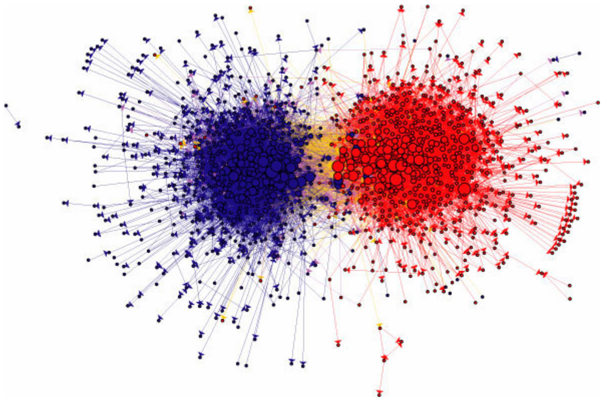
SUMMARY

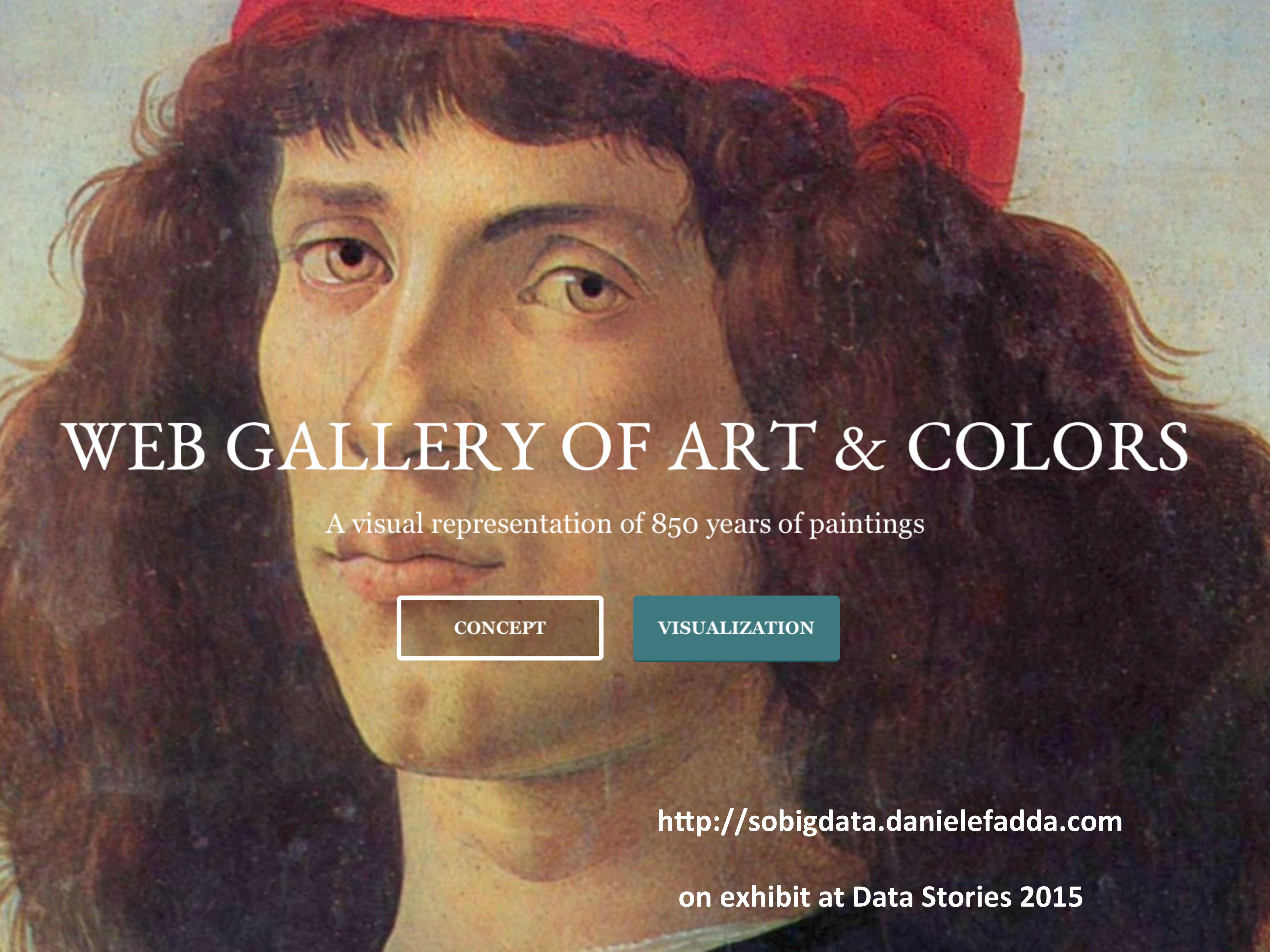
- Social media data and analytics (examples)
- Nowcasting Flu trend with search data –
- Predicting Happyness with lyrics ..and twitter – with predefined lexicon and mood weight
- Superdiversity
- Quantifying opinions with ISA algorithm – Forecasting Peruvian Elections
- Twitter data (UGC) for Migration Studies

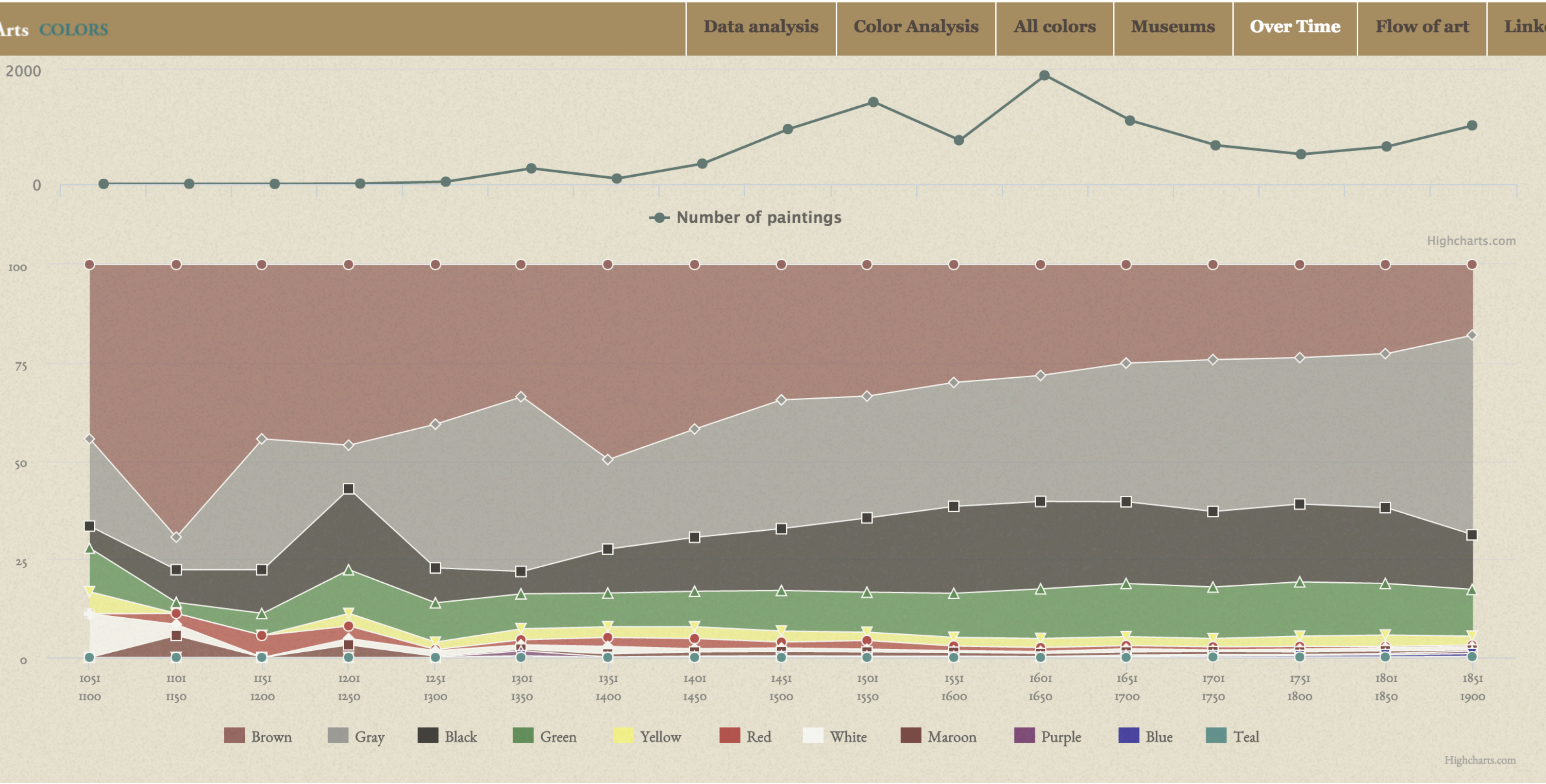
Social Media Data

Raw data +

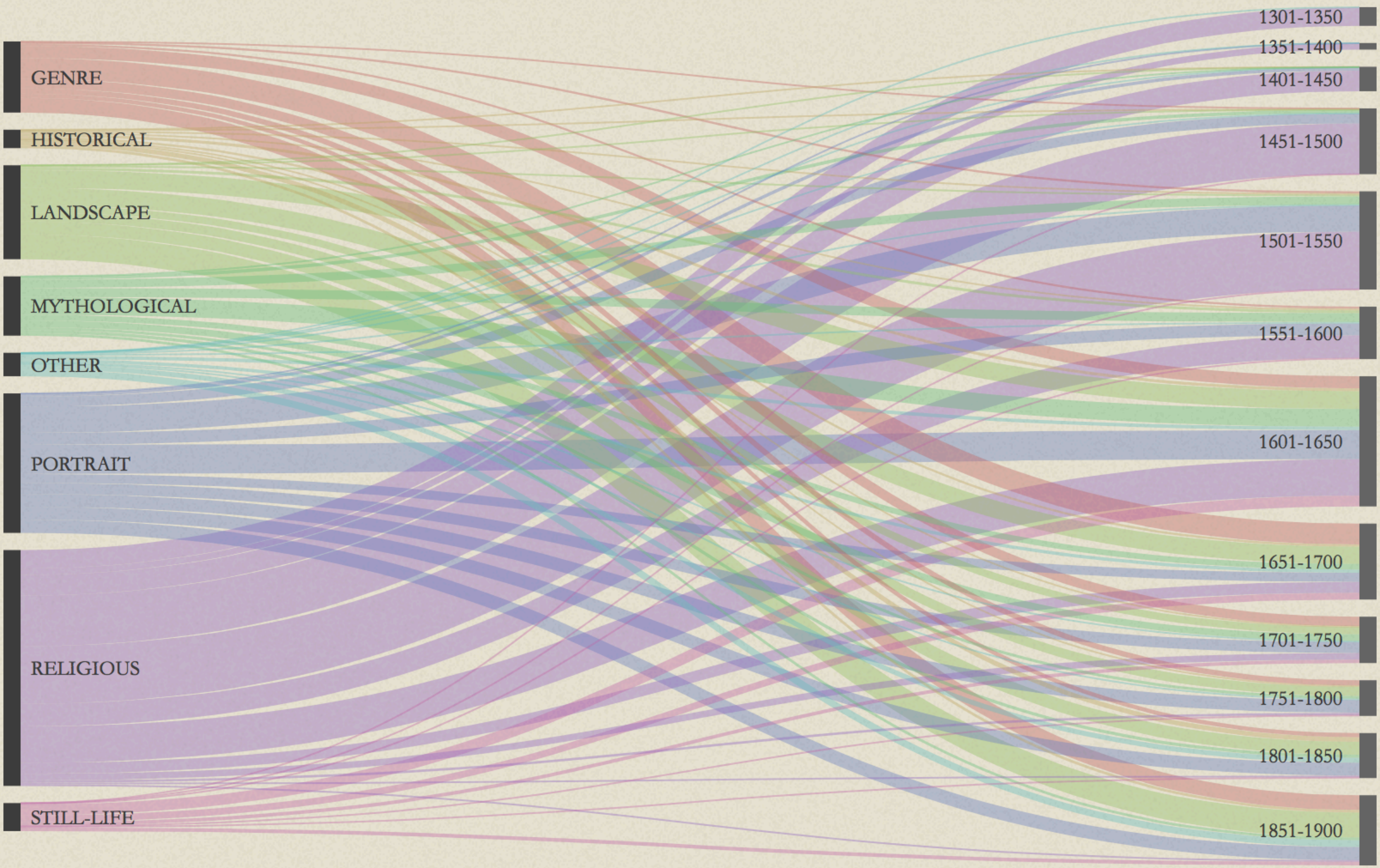
Shares, Likes, Mentions, Impressions, Hashtag usage, URL clicks, Keyword analysis, New followers, Comments




WEB GALLERY OF ART & COLORS
A visual representation of 850 years of paintings
CONCEPT
VISUALIZATION
http://sobigdata.danielefadda.com
on exhibit at Data Stories 2015



http://sobigdata.danielefadda.com

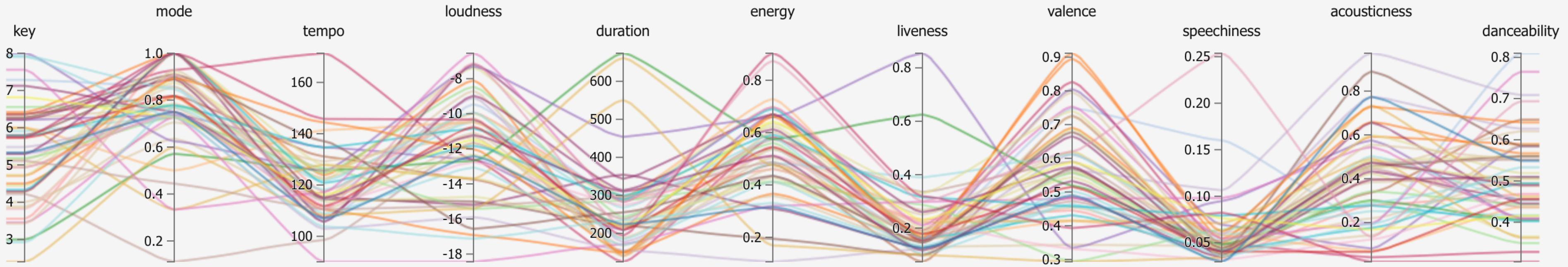


GENRE
HISTORICAL
LANDSCAPE
MYTHOLOGICAL
OTHER
PORTRAIT
RELIGIOUS
STILL-LIFE
1301-1350
1351-1400
1401-1450
1451-1500
1501-1550
1551-1600
1601-1650
1651-1700
1701-1750
1751-1800
1801-1850
1851-1900

Rolling Stone 500 Albums — Top 50 Analysis

Keep Exclude Export

50/50 Lines at **61.8%** opacity. Hide Ticks Dark



The secret of success?

What you see is a multidimensional explorer of musical features from the top 50 Albums of Rolling Stone 500 Greatest Albums.

We used Echo Nest API and Python to get the musical features for every song in an album, then we cleaned, filtered and merged the files to get one JSON per album with the average features.

You can clearly see a common pattern between the albums; it seems that the "best" albums mostly share the same musical features.

Controls

Brush: Drag vertically along an axis.
Remove Brush: Tap the axis background.
Reorder Axes: Drag a label horizontally.
Invert Axis: Tap an axis label.
Remove Axis: Drag axis label to the left edge.

Credits

Visualization adapted from the Nutrient Database Explorer by Kai Chang @ 2012, all credits to him

Project Done by Samuele Borgheresi, Tommaso Ferrari Aggradi and Marco Da Campo

Genres

- 1 Alternative Rock
- 2 Blues
- 1 Blues Rock
- 1 Country Rock
- 2 Folk
- 4 Folk Rock
- 1 Glam Rock
- 1 Grunge
- 2 Hard Rock
- 2 Jazz
- 1 Pop
- 6 Pop Rock
- 2 Progressive Rock
- 4 Psychedelic Rock
- 2 Punk
- 2 Punk Rock
- 1 Rap
- 1 Reggae
- 2 Rhythm and blues
- 5 Rock
- 3 Rock & Roll
- 1 Soft Rock
- 2 Soul
- 1 Southern Rock

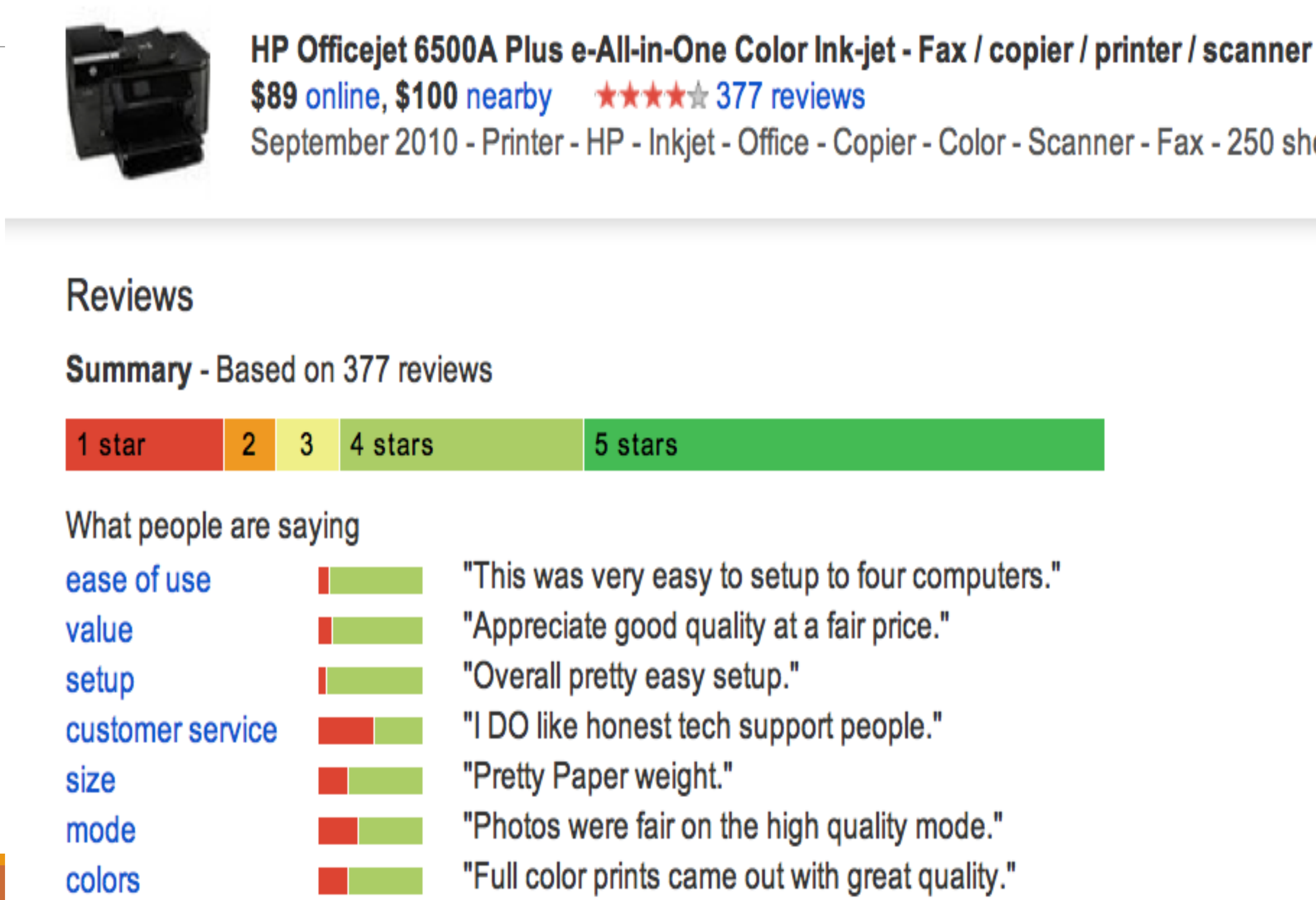
Sample of 25 Albums

Search...

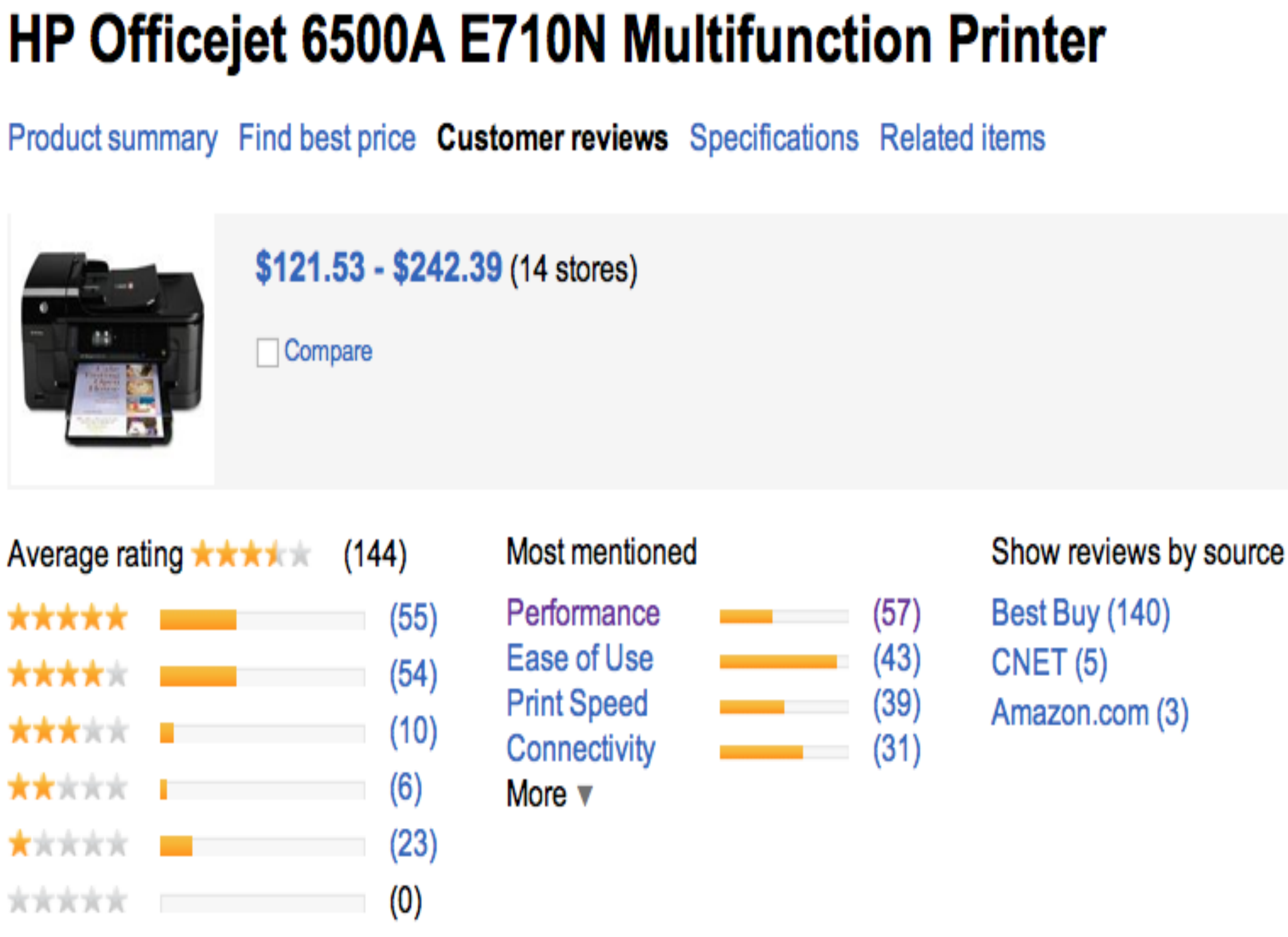
- Bob Dylan - Blonde on Blonde (1966)
- Bob Dylan - Blood on the Tracks (1975)
- Bob Dylan - Bringing It All Back Home (1965)
- Bob Dylan - Highway 61 Revisited (1965)
- Carole King - Tapestry (1971)
- Chuck Berry - The Great Twenty-Eight (1982)
- David Bowie - The Rise and Fall of Ziggy Stardust and the Spiders from
- James Brown - Live at the Apollo (1963)
- John Coltrane - A Love Supreme (1965)
- Led Zeppelin - Led Zeppelin (1969)
- Love - Forever Changes (1967)
- Michael Jackson - Thriller (1982)
- Miles Davis - Kind of Blue (1959)
- Pink Floyd - Dark Side of the Moon (1973)
- Public Enemy - It Takes a Nation of Millions to Hold Us Back (1965)
- Sex Pistols - Never Mind the Bollocks, Here's the Sex Pistols (1977)
- Stevie Wonder - Innervisions (1973)
- The Band - Music from Big Pink (1968)
- The Beach Boys - Pet Sounds (1966)
- The Beatles - Please Please Me (1963)
- The Beatles - Sgt. Pepper's Lonely Hearts Club Band (1967)
- The Jimi Hendrix Experience - Are You Experienced? (1967)
- The Rolling Stones - Exile on Main St. (1972)
- The Who - Who's Next (1971)
- U2 - The Joshua Tree (1987)

http://sobigdata.borgheresi.it/music_in_numbers/

Google Product Search

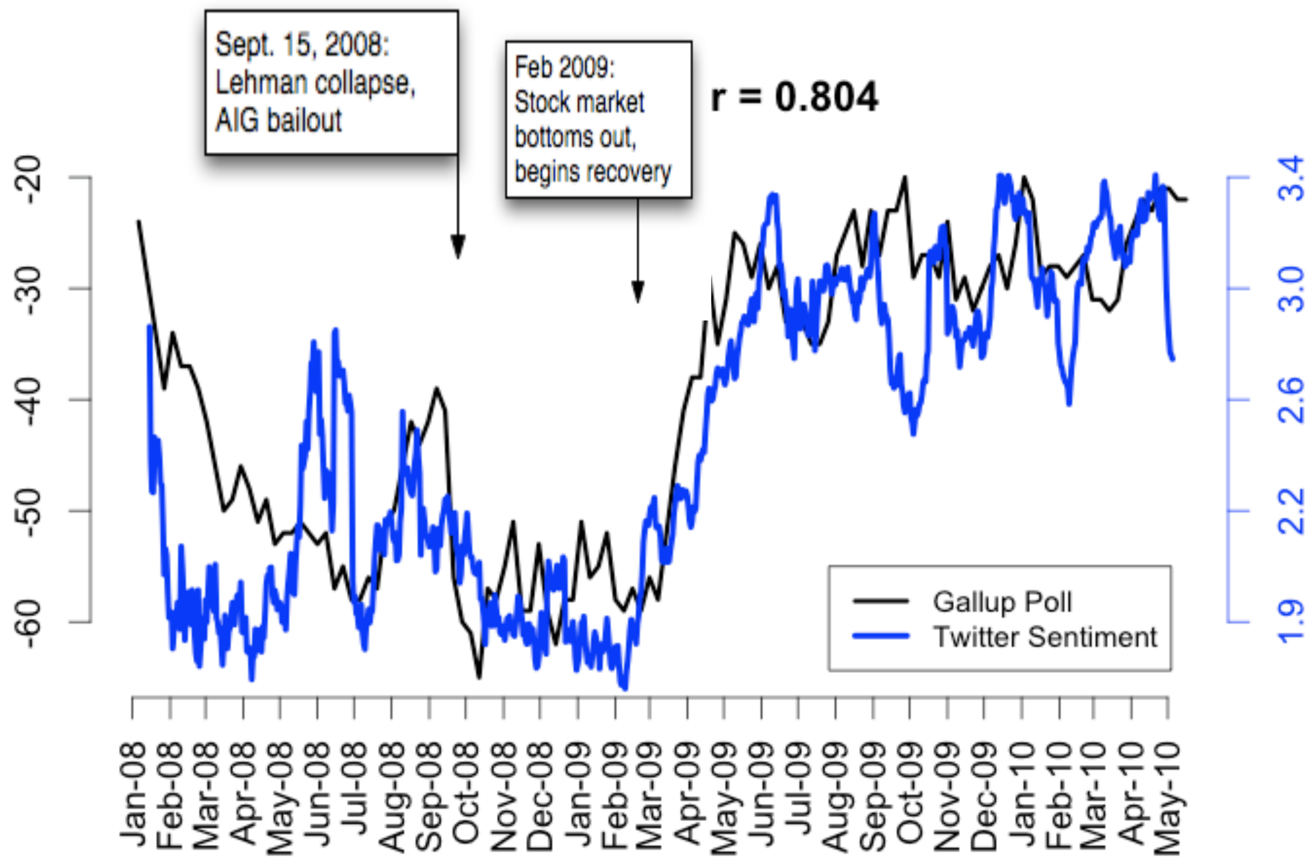


Bing Shopping



Twitter sentiment versus Gallup Poll of Consumer Confidence

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In ICWSM-2010

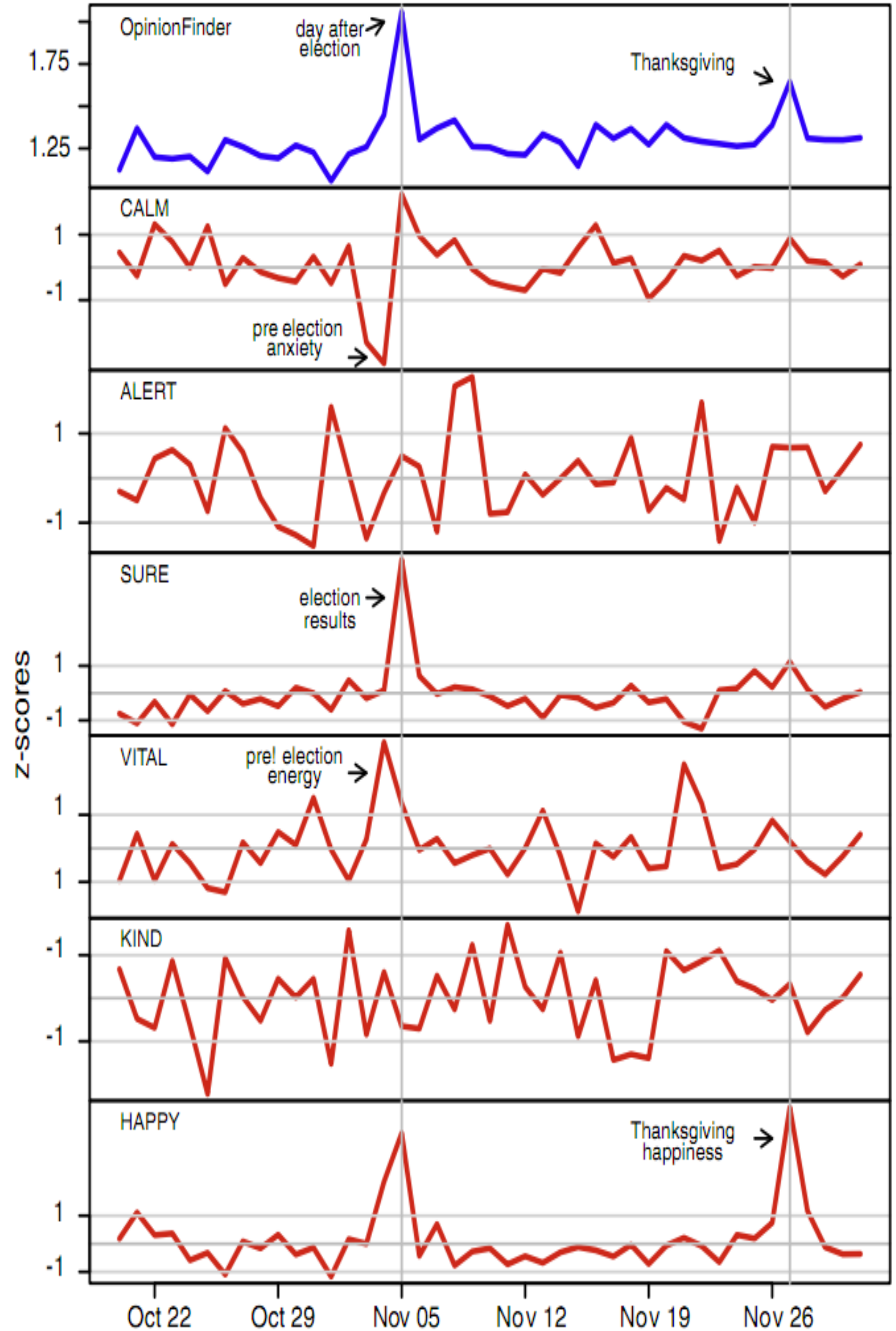


Twitter sentir

Johan Bollen, Huina Mao, Xiaojun Zeng. 2011. [Twitter mood predicts the stock market,]()

Journal of Computational Science 2:1, 1-8. 10.1016/j.jocs.2010.12.007.

Opinion Finder ed GPOMOS, 7 Time series



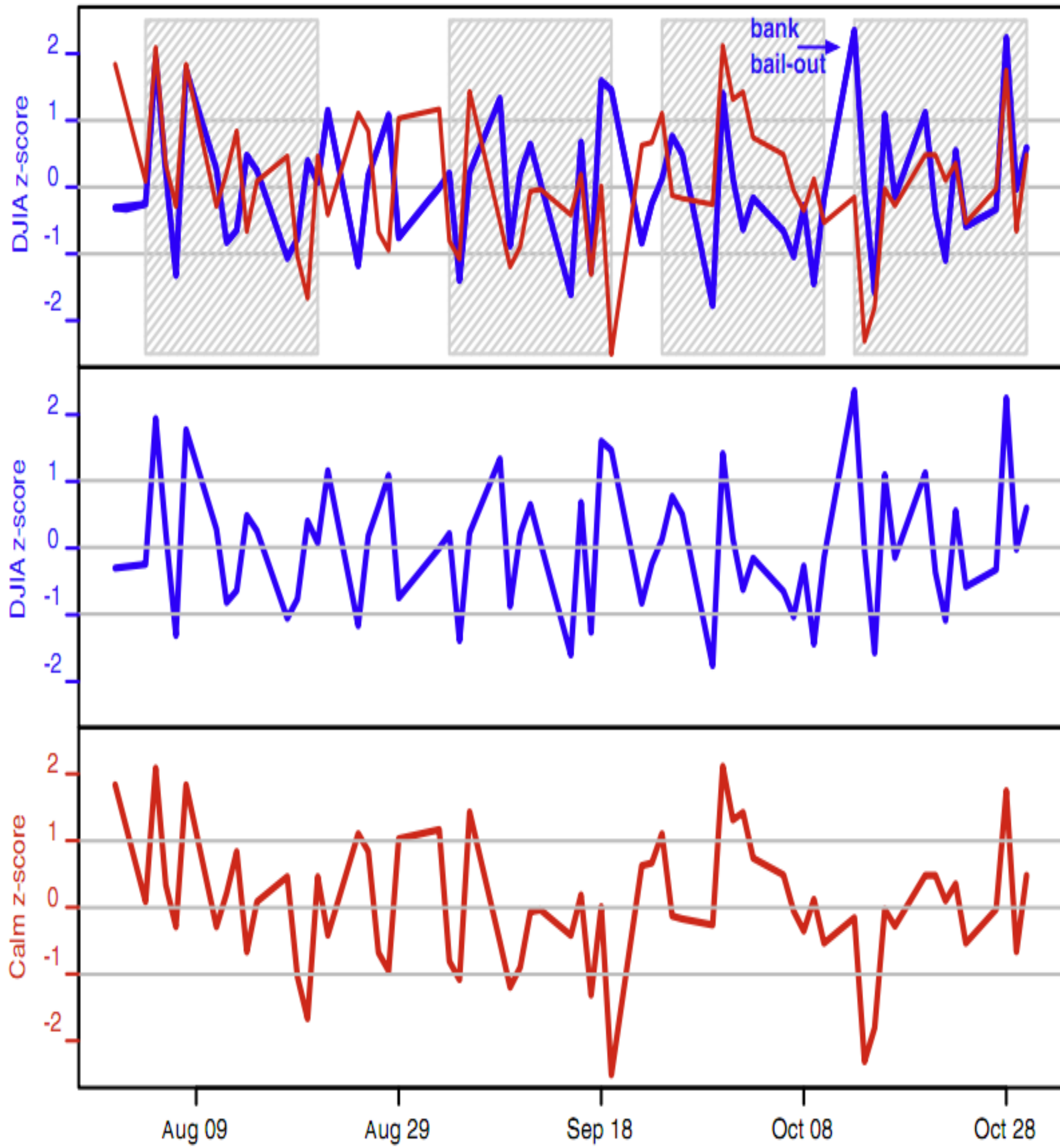
Bollen et al. (2011)

CALM predicts DJIA 3 days later

At least one current hedge fund uses this algorithm

Dow Jones

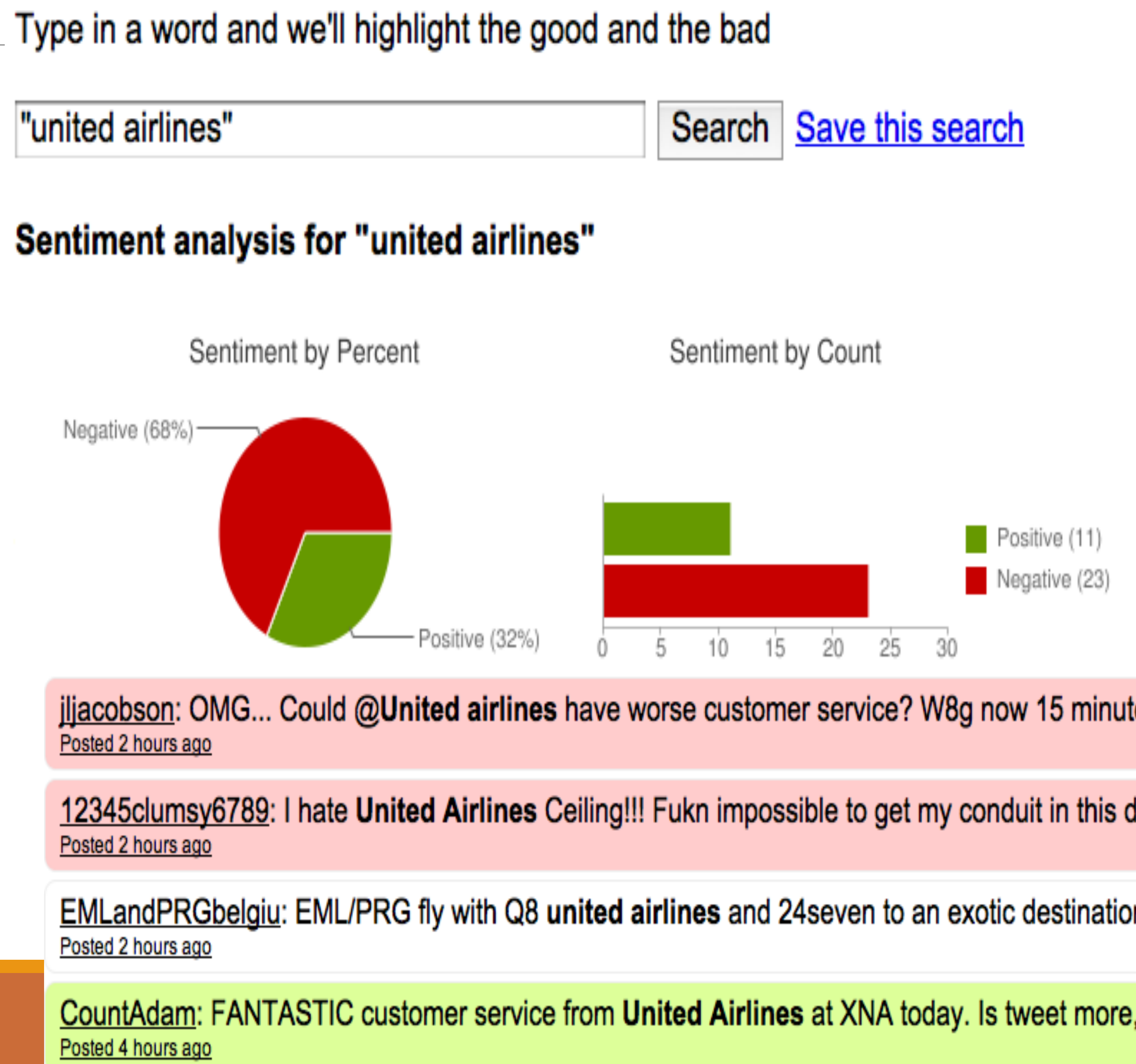
CALM



Target Sentiment on Twitter

[Twitter Sentiment App]

Alec Go, Richa Bhayani, Lei Huang. 2009. Twitter Sentiment Classification using Distant Supervision



Nowcasting: Predicting the present

The traps of big data

Google Flu Trends : search data can help predict the incidence of influenza-like diseases

Close relationship between number of people searching for flu-related topics and how many people have symptoms

Prediction models compared to real-world cases of flu

Hyunyoung Choi and Hal Varian. *Predicting the present with google trends*. Technical Report, 2009.

How it works

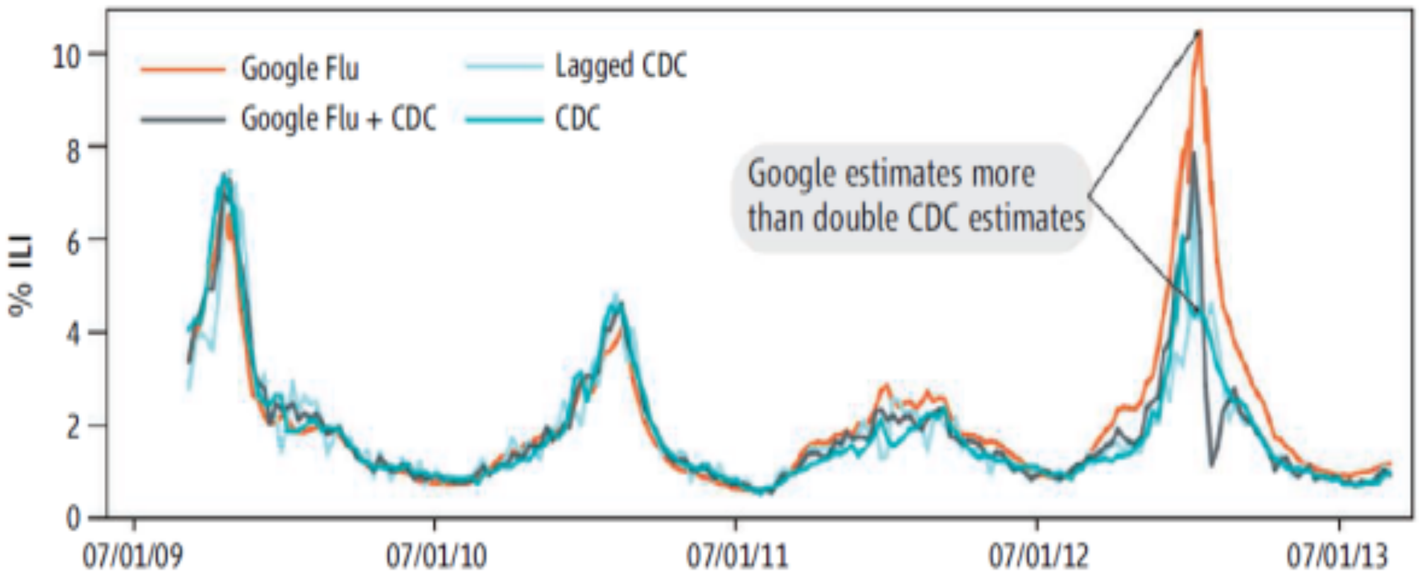
- a time series is computed for about 50 million common queries entered weekly from 2003 to 2008.
- georeferenced by identifying the IP address associated with each search, the state in which this query was entered can be determined.
- linear model is used to compute the log-odds of Influenza-like illness (ILI) physician visit (official data) and the log-odds of ILI-related search query:

$$\mathrm{logit}(P) = \beta_0 + \beta_1 \times \mathrm{logit}(Q) + \epsilon$$

How it works

- Each of the 50 million queries is tested as Q to see if the result computed from a single query could match the actual history ILI data obtained from the U.S. Centers for Disease Control and Prevention (CDC).
- This process produces a list of top queries which gives the most accurate predictions of CDC ILI data when using the linear model. Then the top 45 queries are chosen because, when aggregated together, these queries fit the history data the most accurately.
- Finally, the trained model is used to predict flu outbreak across all regions in the United States.

But..



- In 2009, completely missed the non-seasonal influenza A-H1N1
- In 2013, double doctor visits than CDC

David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. *The parable of google flu: Traps in big data analysis*. Science Magazine, 343(6176):1203–1205, 2014.

Happiness as Subjective Well-Being Indicator

MEASURING HAPPYNESS

Subjective Well-Being

- Perceptions and evaluations affect the way people face life and take advantage of opportunities in different ways.
- Subjective indicators are useful complement to strictly objective indicators, because they allow evaluating the possible differences between what people report and what it is captured by objective indicators.

Happiness as Subjective Well-Being Indicator

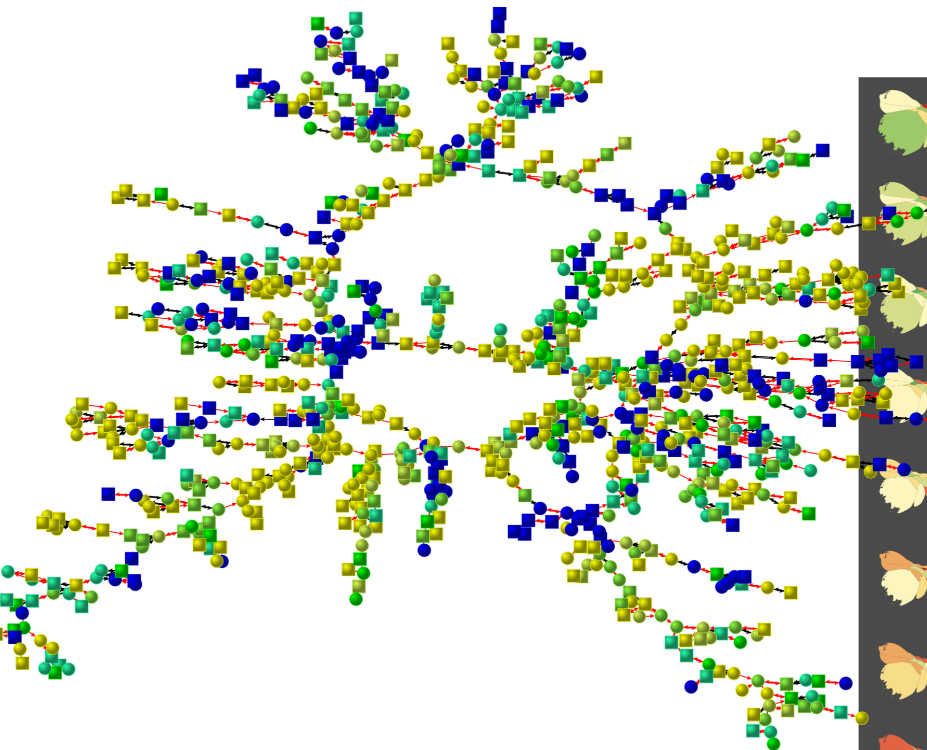
All the subjective well-being indicators observed in the literature come from surveys, and are a bag of ratings of questions like: life satisfaction, leisure time satisfaction, utilities closeness, etc.

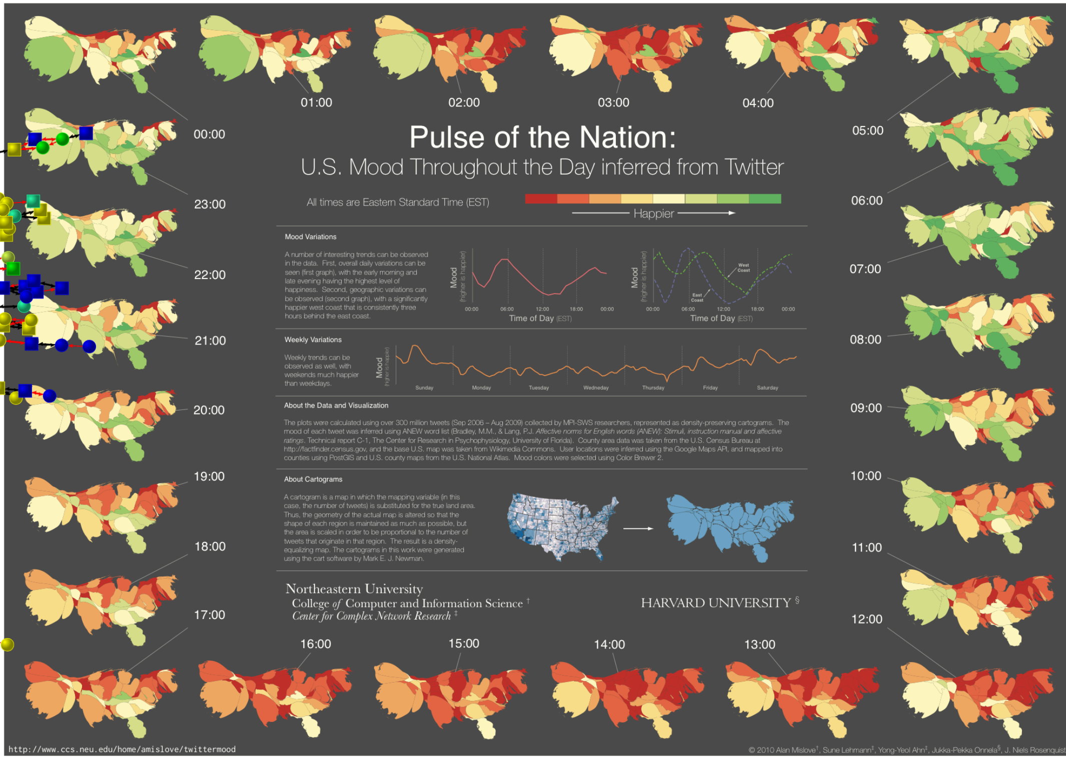
From a data-driven point of view, the best way to estimate the level of subjective well-being is to estimate the level of happiness.

Happiness is intrinsically correlated with, and it is a consequence of, the percepted level of wellbeing driven by social relationships, health, work condition, etc.

Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents.Peter Sheridan Dodds et al. Journal of Happiness Studies. 2010.

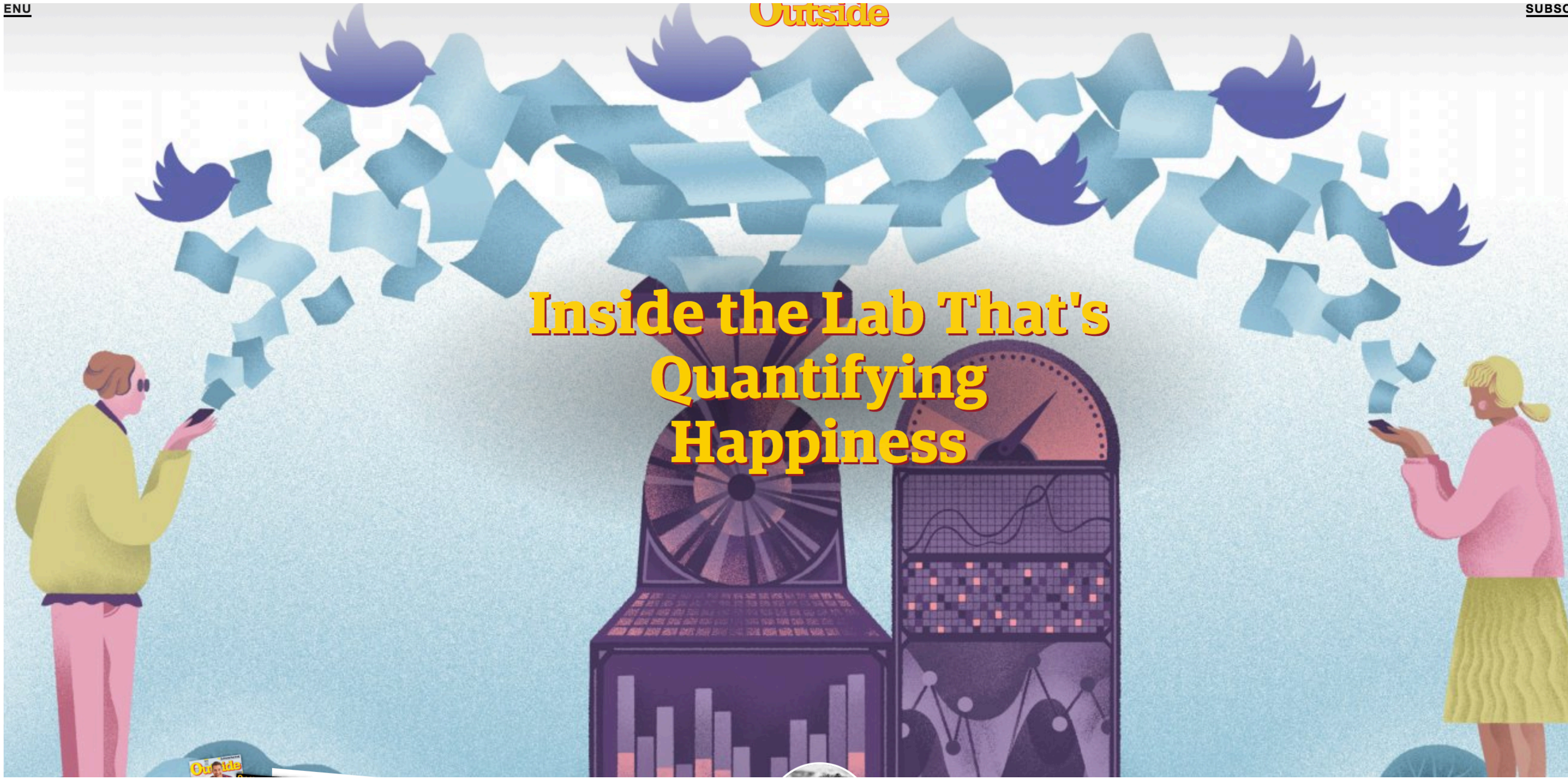
Measuring Happyness





Text Data Sources

- Every written text (books, songs, blogs, posts, etc.) reveals a positive or negative sentiment with respect to their content. Online Social Networks like Twitter and Facebook are the expression of our personal moods and our opinions.
- They are a very powerful data source to reveal the level of happiness in a certain region and in a certain ***period***.
- Indeed one of the biggest issue of traditional indicators based on surveys is that they "measure" a certain variable in the instant the user answer the survey.
- The mood of a user and consequently the answer to a certain question could be different just a week after because of family or work problems for example.



Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents, Peter Sheridan Dodds Æ Christopher M. Danforth, J Happiness Stud (2010)

The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place Lewis Mitchell1 , Morgan R. Frank, Kameron Decker Harris1,2, Pete
Sheridan Dodds, Christopher M. Danforth1

https://www.outsideonline.com/2230891/inside-lab-thats-quantifying-happiness

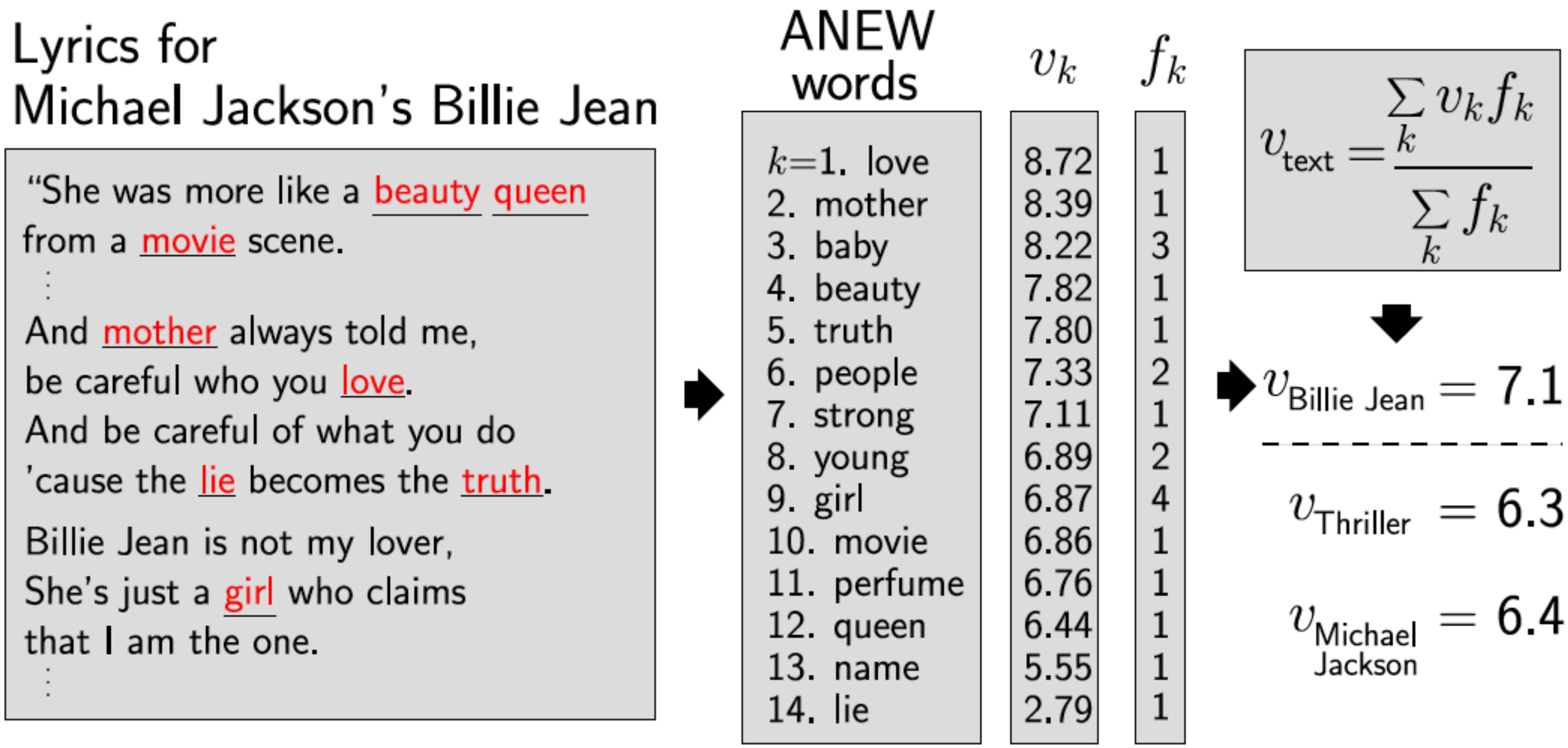
A data-driven Happiness Indicator

ANEW: Affective Norms for English Words are 1034 words rated between 1-9 (good-bad) in a sort of happiness scale.

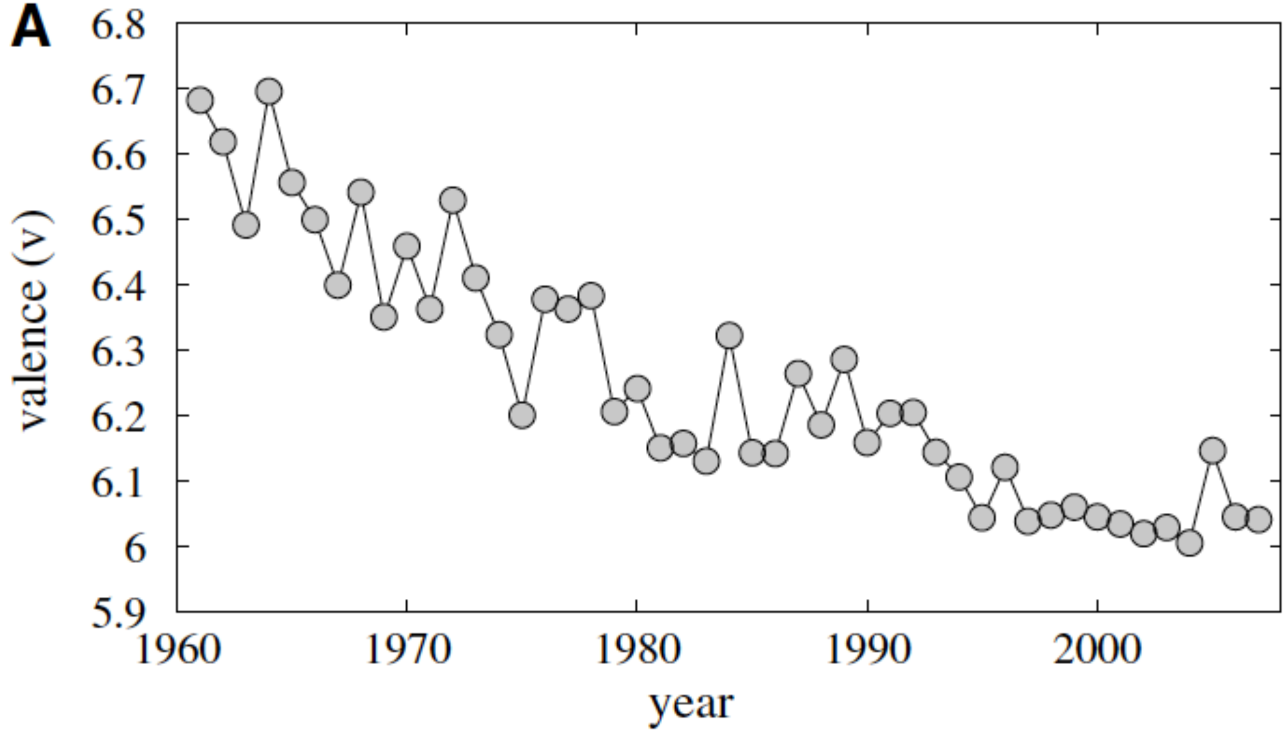
These words can be used to evaluate the level of happiness of a text called **valence**:

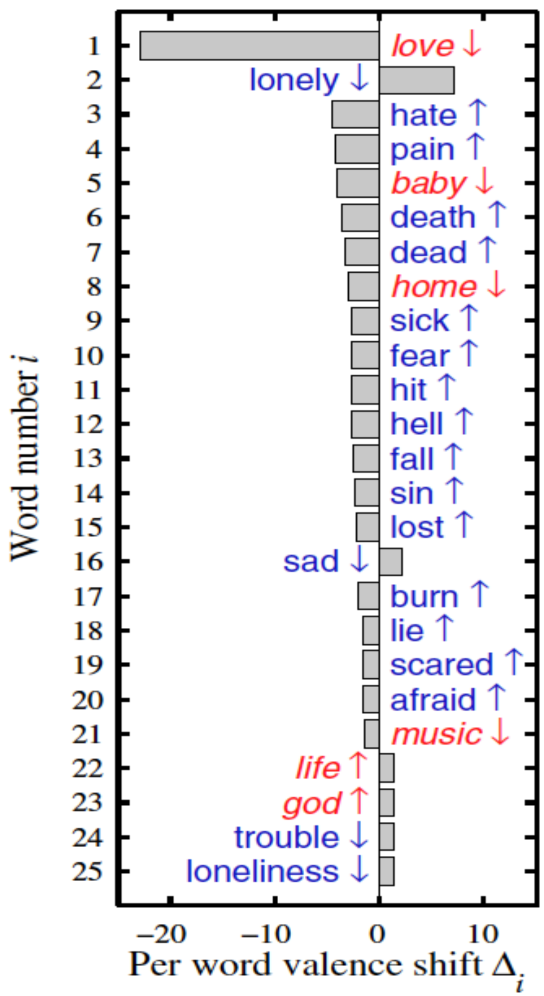
$$v_{\text{text}} = \frac{\sum_{i=1}^{n} v_i f_i}{\sum_{i=1}^{n} f_i}$$

Valence of lyrics - example

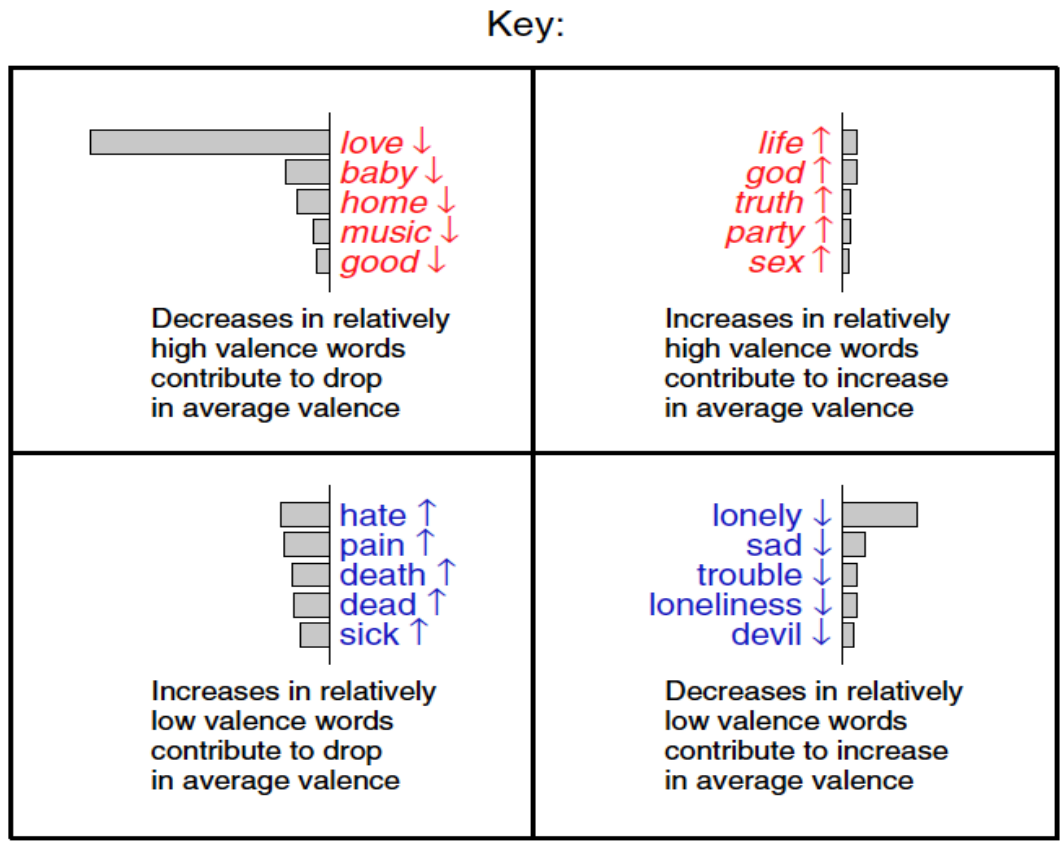


Song lyrics valence across decades



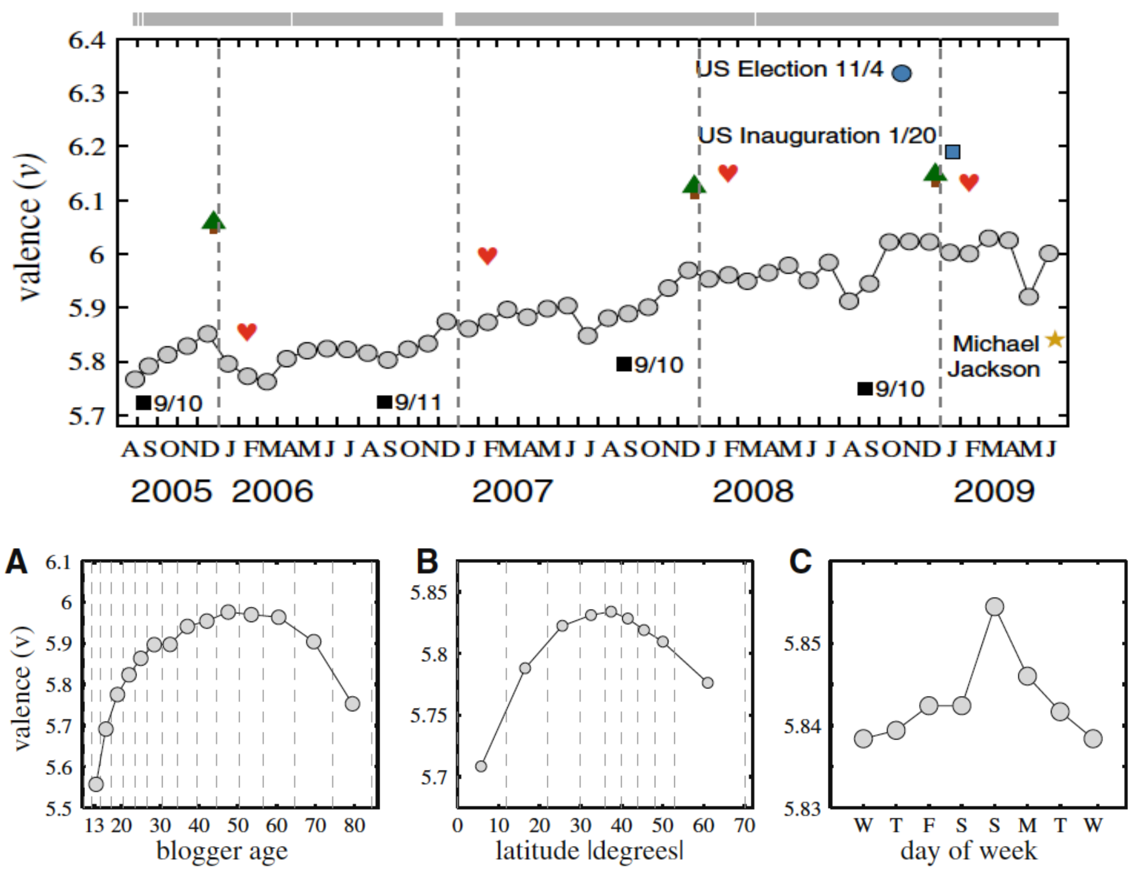


Word number i
Per word valence shift Δ_i
–20
–10
0
10
1 love ↓
2 lonely ↓
3 hate ↑
4 pain ↑
5 baby ↓
6 death ↑
7 dead ↑
8 home ↓
9 sick ↑
10 fear ↑
11 hit ↑
12 hell ↑
13 fall ↑
14 sin ↑
15 lost ↑
16 sad ↓
17 burn ↑
18 lie ↑
19 scared ↑
20 afraid ↑
21 music ↓
22 life ↑
23 god ↑
24 trouble ↓
25 loneliness ↓

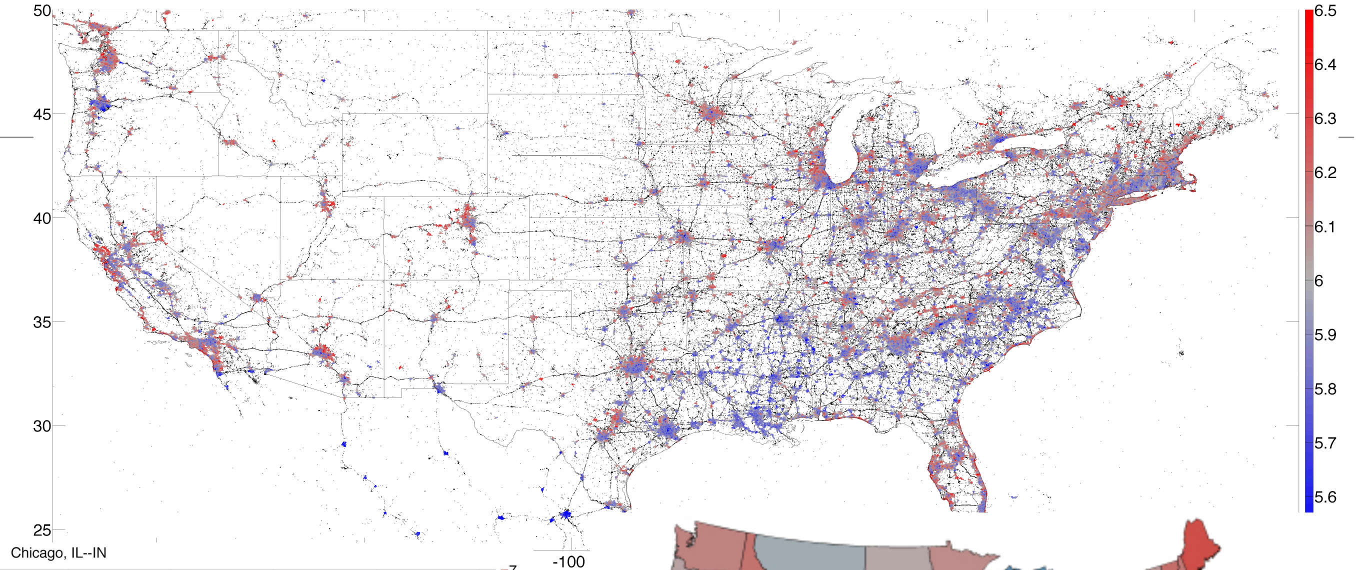


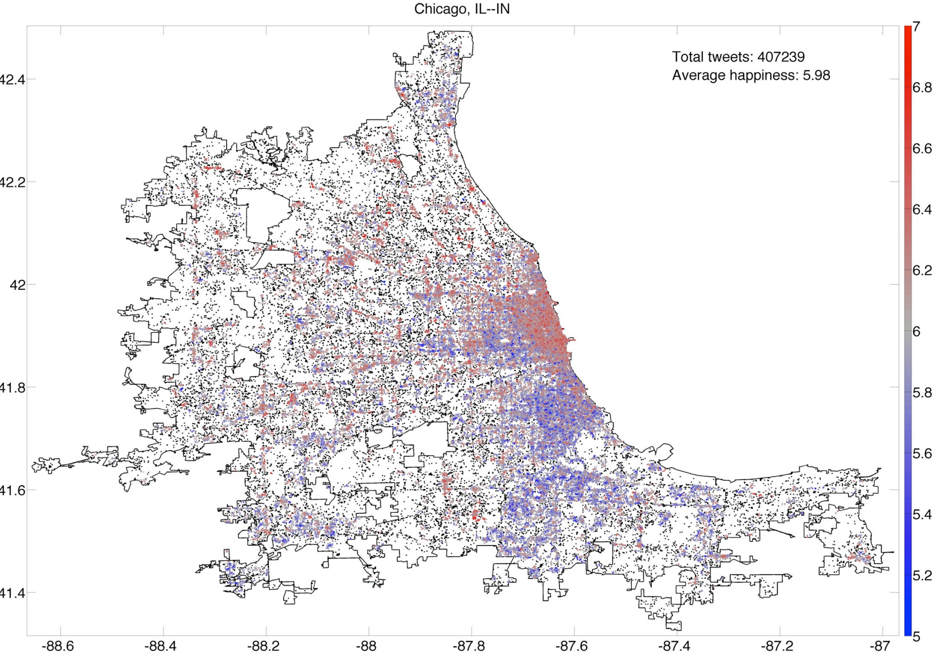
Key:
love ↓
baby ↓
home ↓
music ↓
good ↓
Decreases in relatively high valence words contribute to drop in average valence
life ↑
god ↑
truth ↑
party ↑
sex ↑
Increases in relatively high valence words contribute to increase in average valence
hate ↑
pain ↑
death ↑
dead ↑
sick ↑
Increases in relatively low valence words contribute to drop in average valence
lonely ↓
sad ↓
trouble ↓
loneliness ↓
devil ↓
Decreases in relatively low valence words contribute to increase in average valence

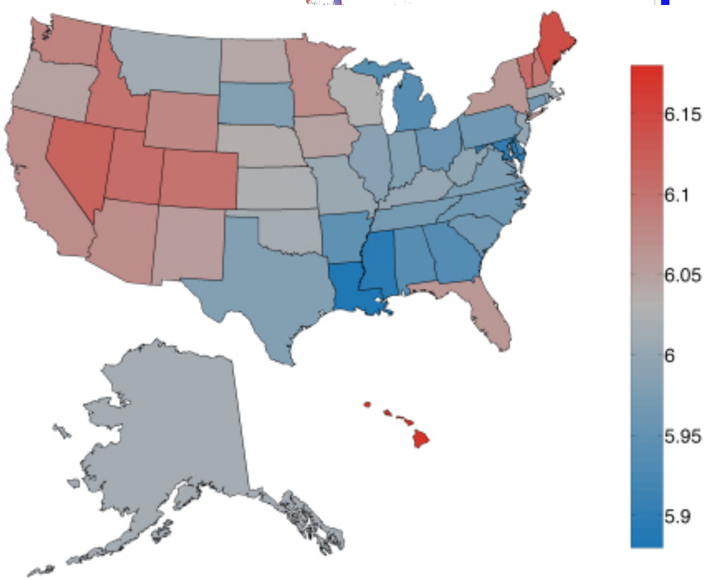
Blog valence



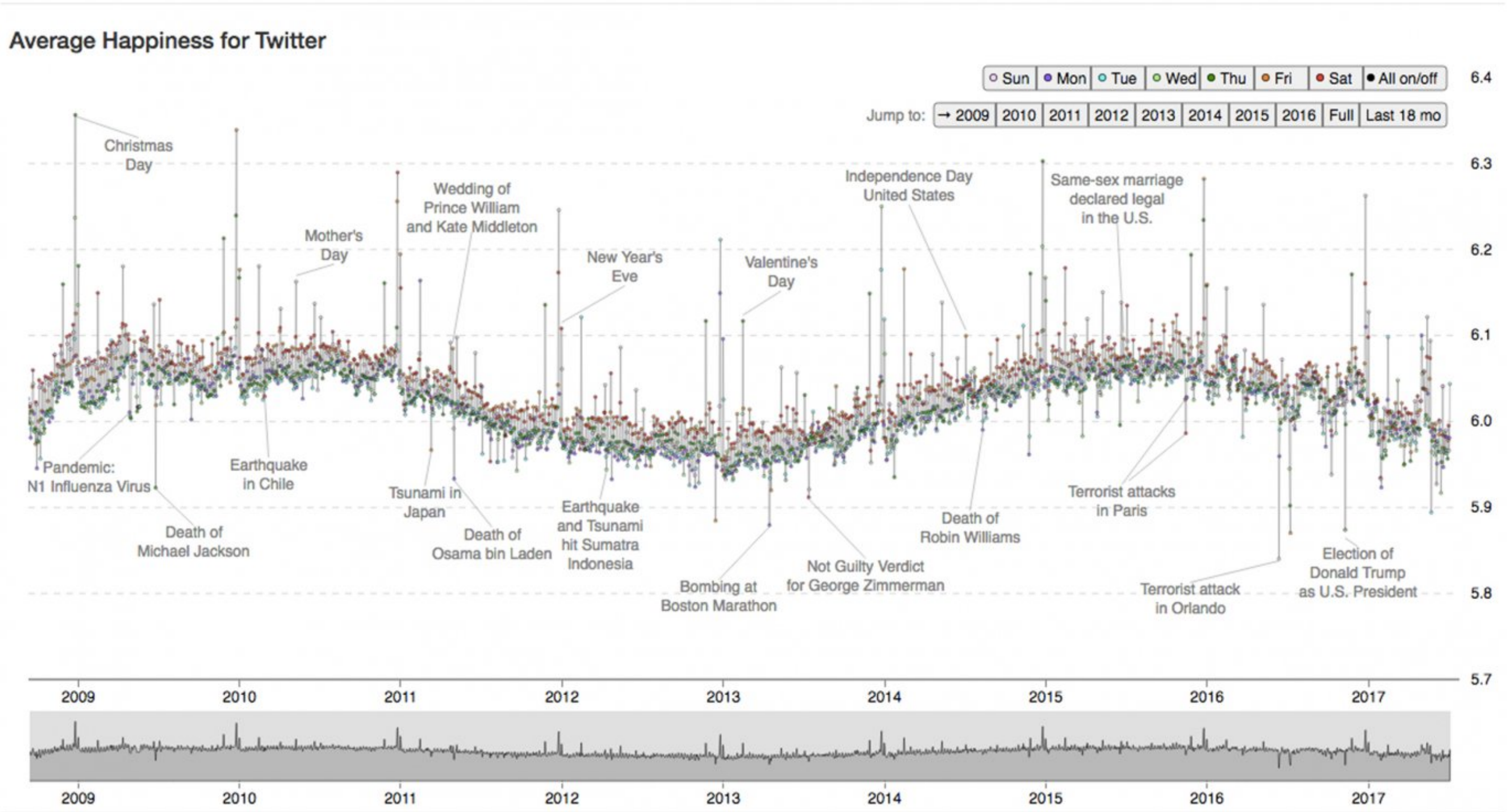
Twitter hedonometer







Average Happiness for Twitter



PARTIAMO DALLE PAROLE:
DOODS

Rapimento	2.76
Dolore	2.46
Angosciato	2.12
Sanguinoso	2.90

Spiaggia	8.03
Amore	8.72
Bacio	8.26
Casa	7.91

COME CALCOLIAMO IL MOOD?

- Frase: "io amo le spiagge"
- Lemmatizzazione:

  io → io

  amo → amare

  le → la

  spiagge → spiaggia
- Applicazione della formula di Dodds

LEMMATIZZAZIONE:

RIPORTARE LE FORME FLESSE A LEMMI

Io amo le spiagge → io amare la spiaggia

Spiaggia 8.03
Amore 8.72

Io amare la spiaggia

8.72 8.03

DOODS:

CALCOLIAMO LA SOMMATORIA DEGLI SCORE TROVATI MOLTIPLICATI PER LA LORO FREQUENZA DIVISO LA SOMMATORIA DELLE FREQUENZE DI TUTTI I LEMMI TROVATI

$$\frac{\sum_{i=1}^{n} v_i f_i}{\sum_{i=1}^{n} f_i} = \frac{8.72 * 1 + 8.03 * 1}{1 + 1} = 8.3$$

Measuring the “salad bowl” -Superdiversity on Twitter-

ALINA SÎRBU

WITH

LAURA POLLACCI, FOSCA GIANNOTTI, DINO PEDRESCHI

UNIVERSITY OF PISA

ISTI CNR

Outline

Superdiversity - a novel index based on sentiment
Validation against immigration data
Comparison with other indices
Discussion: is nowcasting possible?

Superdiversity

Superdiversity - a new level of cultural diversity due to immigration and cultural differences among immigrants themselves (Vertovec, 2007)

Measuring superdiversity - difficult task

- Cultural diversity - number of languages spoken in a region
- Immigration rates - official statistics - generally low time and space resolution
- Use social big data
  - Here: geolocalised tweets

Superdiversity Index (SI)

Main idea:

- different cultures assign **different emotional valence** to different words.
- compute **SI** as a **distance** between the '**standard**' emotional valence of a set of words and the '**used**' valence in the population of a region
  - standard - manually tagged lexicon - ANEW (Bradley and Lang 1999)
  - used - estimate from Twitter data

Estimating sentiment valences of words on Twitter

Using the algorithm from *Pollaci et al. 2017*.

Extend a sentiment-tagged lexicon using a Twitter corpus

- built to enhance lexicon-based sentiment analysis on Twitter
- starts from a small seed lexicon with sentiment valences
- builds a co-occurrence network of words from the Twitter corpus
- sentiment valences diffuse from the seed to the other words

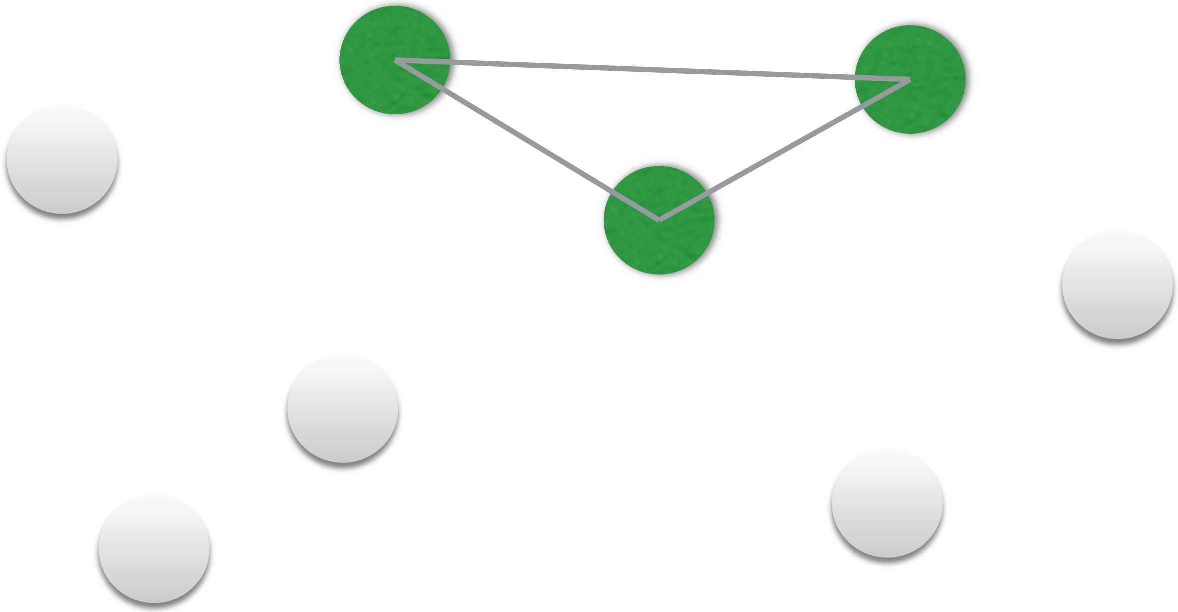
Laura Pollacci, Alina Sîrbu, Fosca Giannotti, Dino Pedreschi, Claudio Lucchese,and Cristina Ioana Muntean. 2017. Sentiment Spreading: An Epidemic Model for Lexicon-Based Sentiment Analysis on Twitter. In Conference of the Italian Association for Artificial Intelligence. Springer, 114–127.

Sentiment Spreading: An Epidemic Model for Lexicon-Based Sentiment Analysis

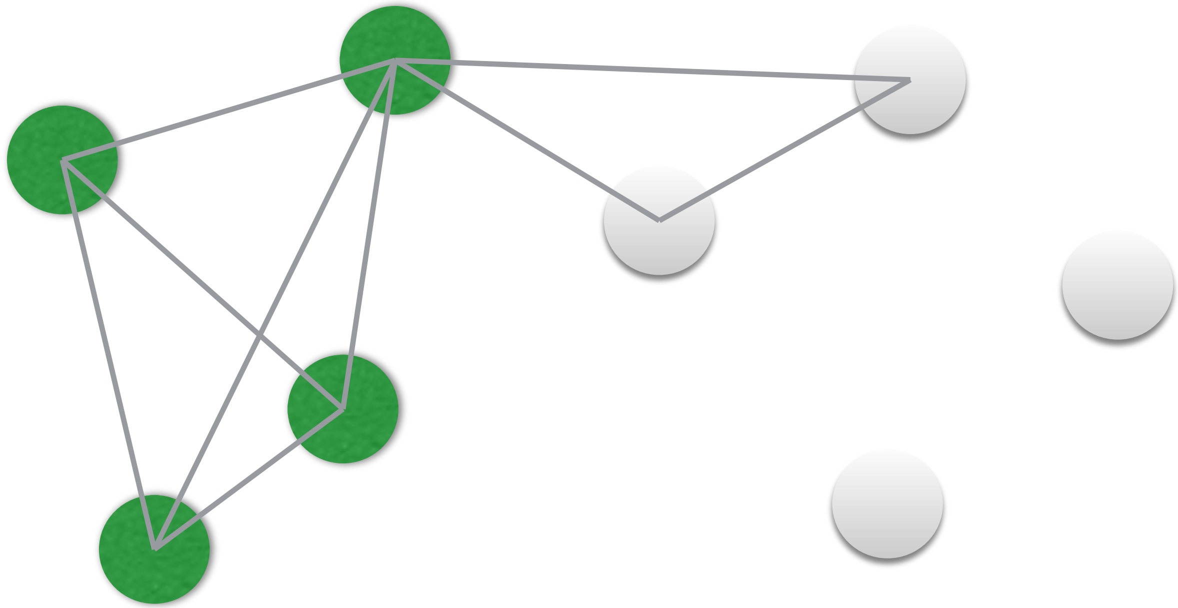
- Extend the dictionary and assign valence to tweet using a epidemic based approach (opinion dynamics like)
- Network of words - words that co-occur in a tweet are connected
- Some words have a valence, the other words take the valence of the neighbours (mean)
- After many iterations the system converges to a stable valence
- Method helps enlarge the dictionary, classify tweets but also characterise the group where the tweets came from

Estimating sentiment valences of words on Twitter

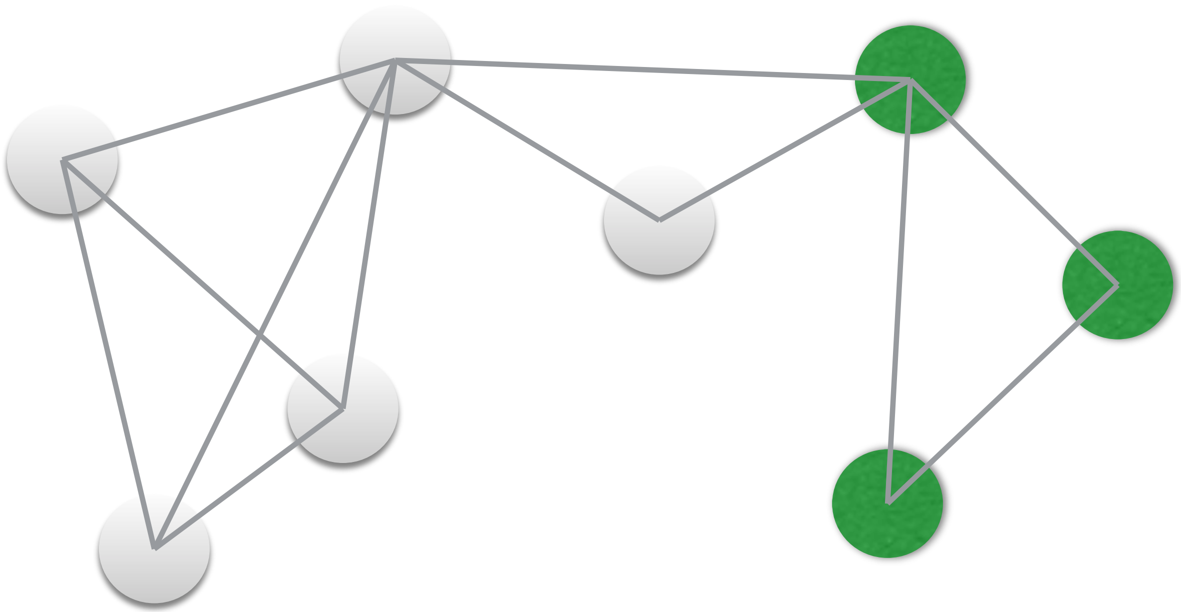
Estimating sentiment valences of words on Twitter



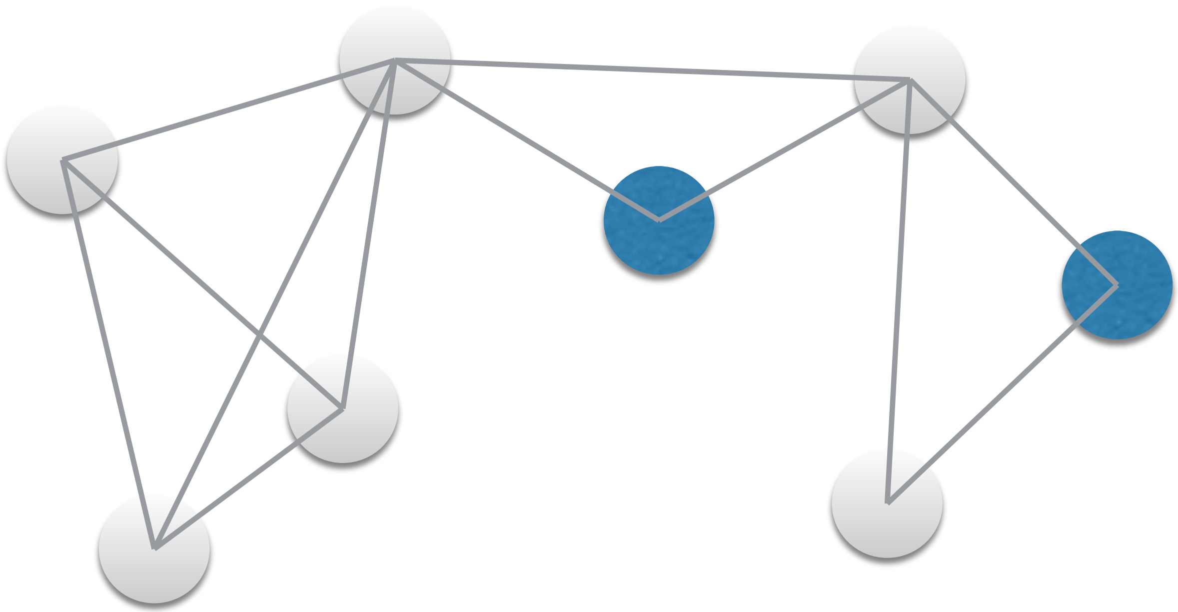
Estimating sentiment valences of words on Twitter



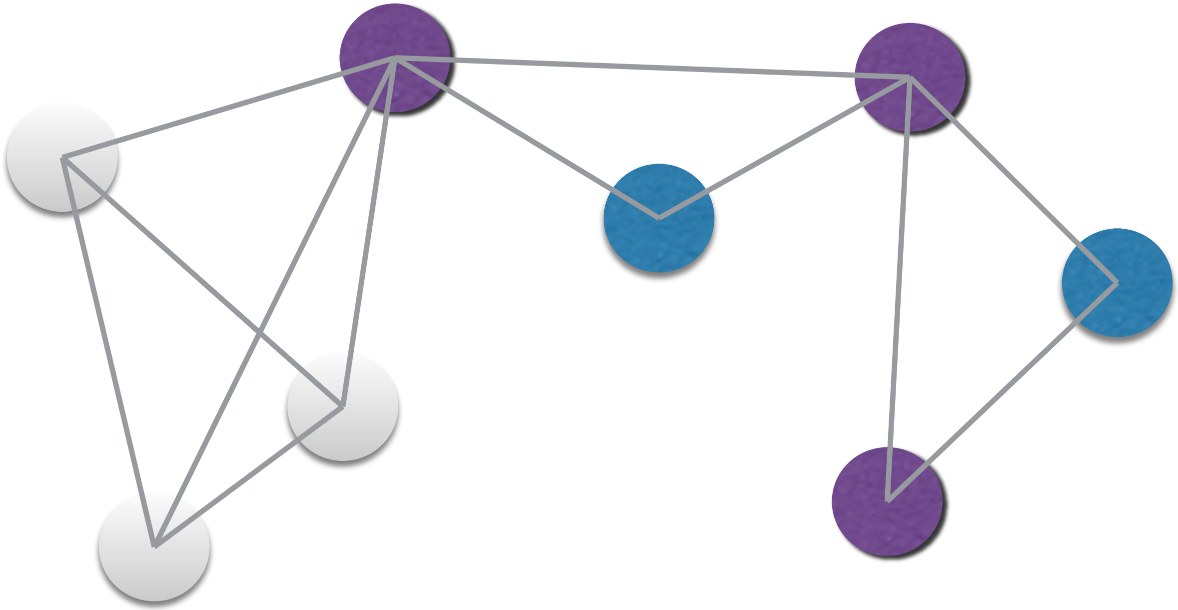
Estimating sentiment valences of words on Twitter



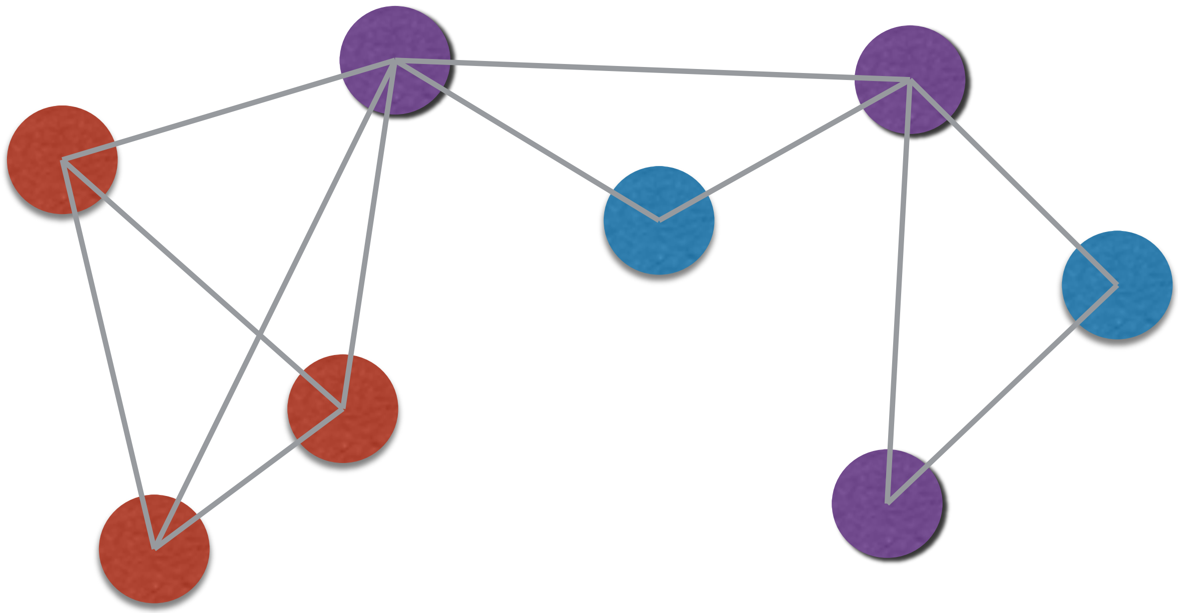
Estimating sentiment valences of words on Twitter



Estimating sentiment valences of words on Twitter



Estimating sentiment valences of words on Twitter



Estimating sentiment valences of words on Twitter

The resulting dictionary

- is larger - enhanced sentiment analysis on Twitter
- depends on the way language is used - the new valences are population dependent
  - we use the new valences as estimates for the real emotional content of the words in the population - **compute the distance to a manually tagged lexicon**

Compute SI for different geographical regions

- 10% of Geolocalised tweets for 3 months (August-October 2015)

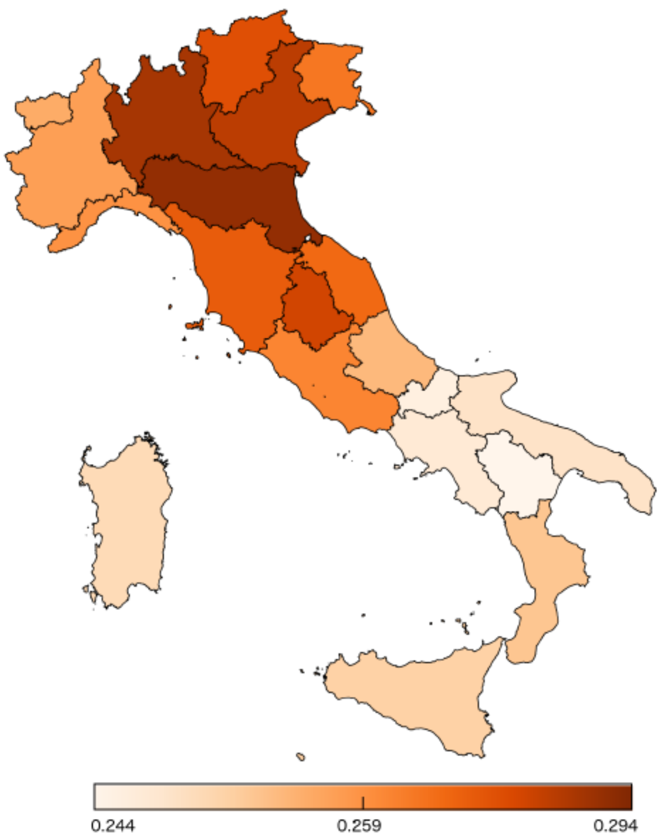
Compare SI with immigration rates

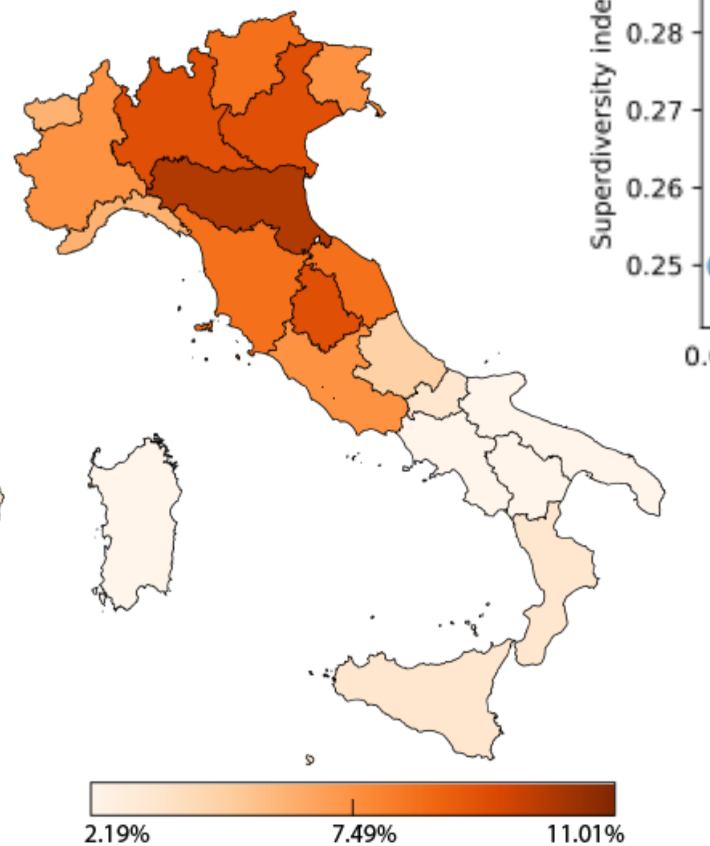
- JRC D4I dataset - immigration rates at various NUTS* levels (https://bluehub.jrc.ec.europa.eu/datachallenge/)
- We analyse NUTS1, NUTS2, NUTS3 for Italy and UK

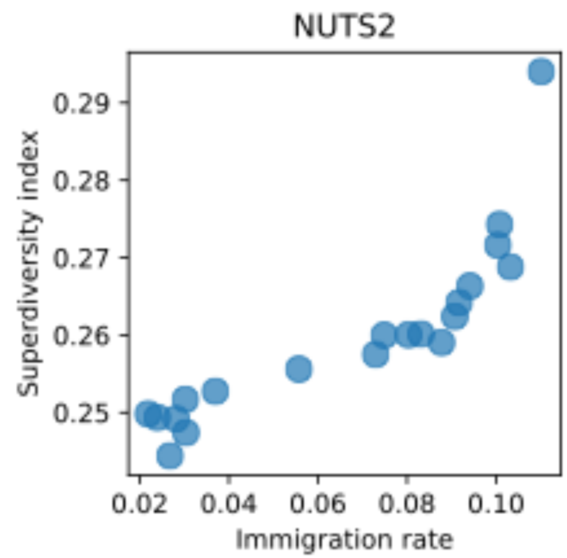
European Commission

**Nomenclature of Territorial Units for Statistics*

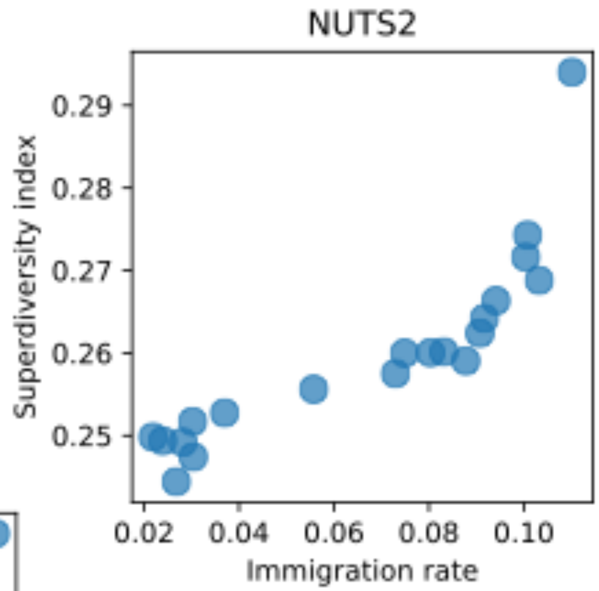
SI Evaluation

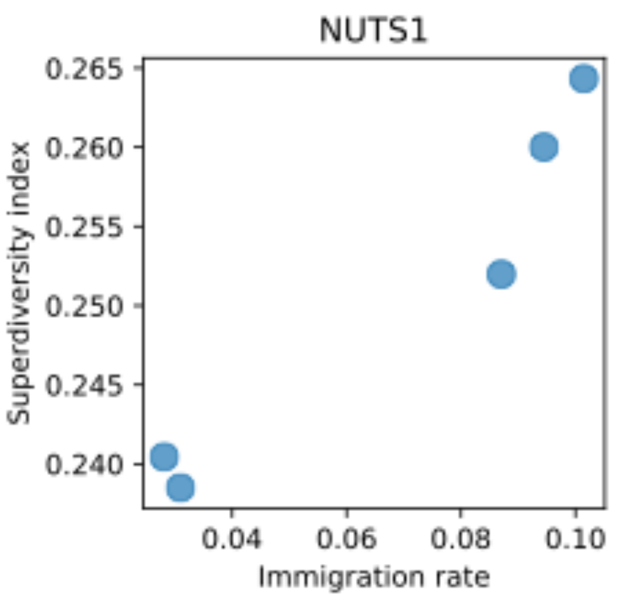




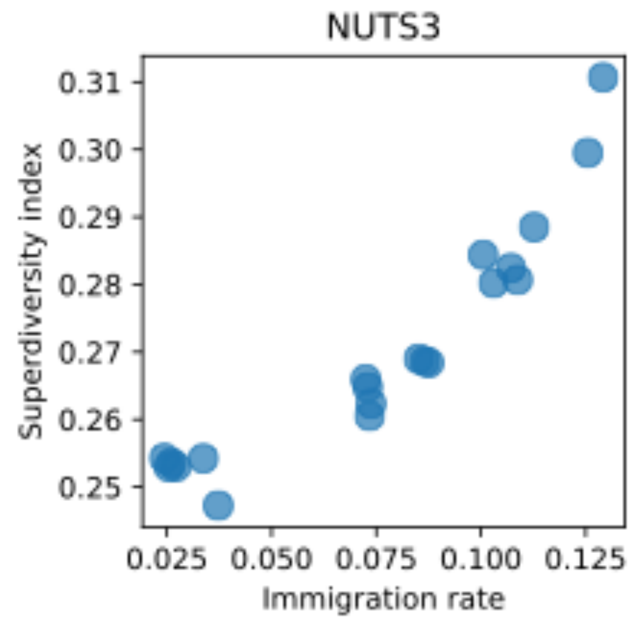


SI Evaluation

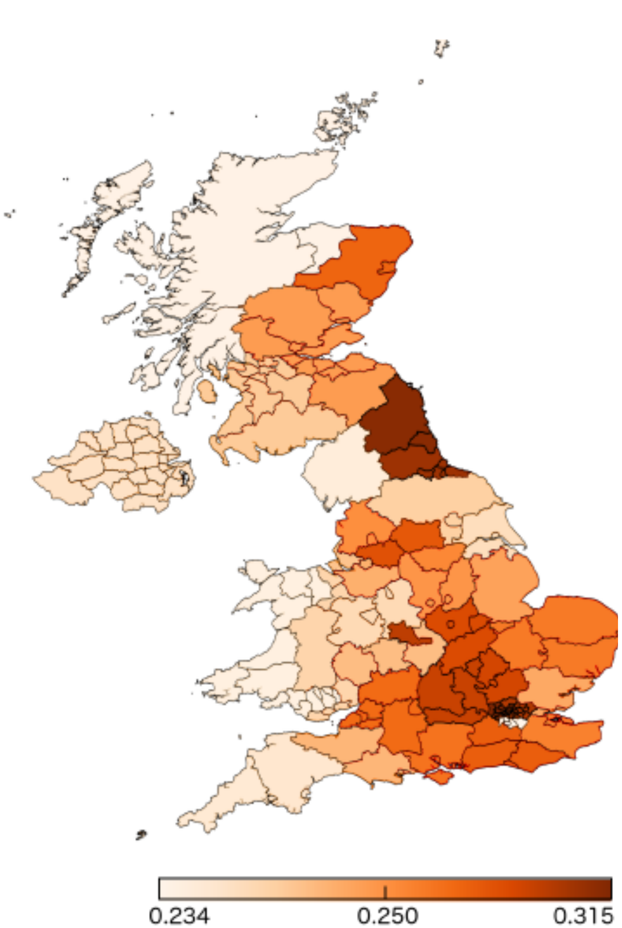


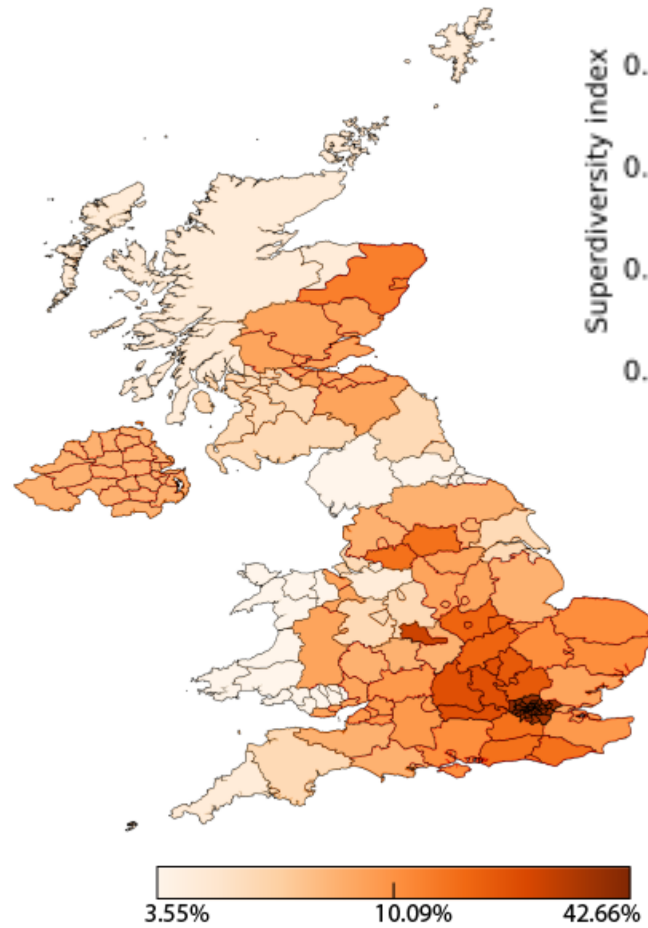


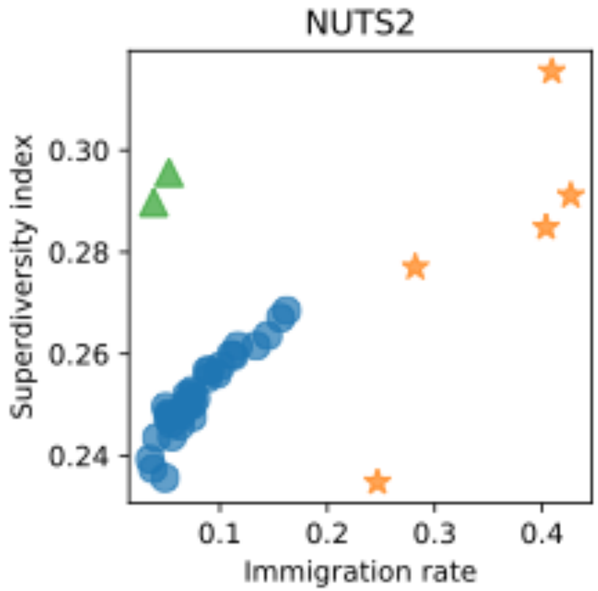
Italy

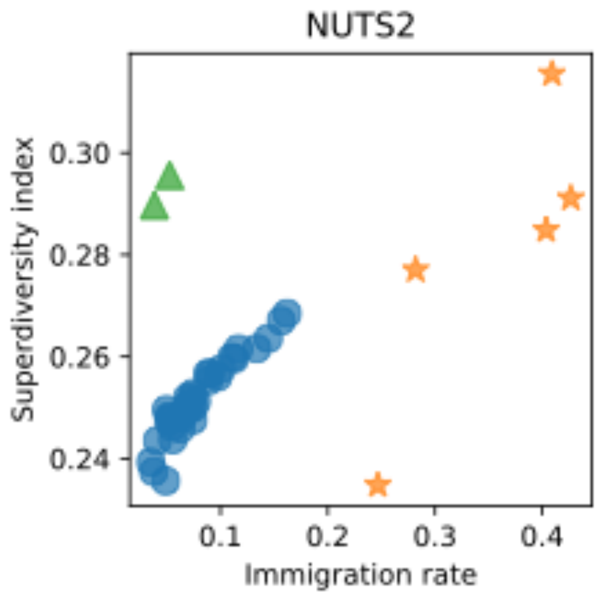


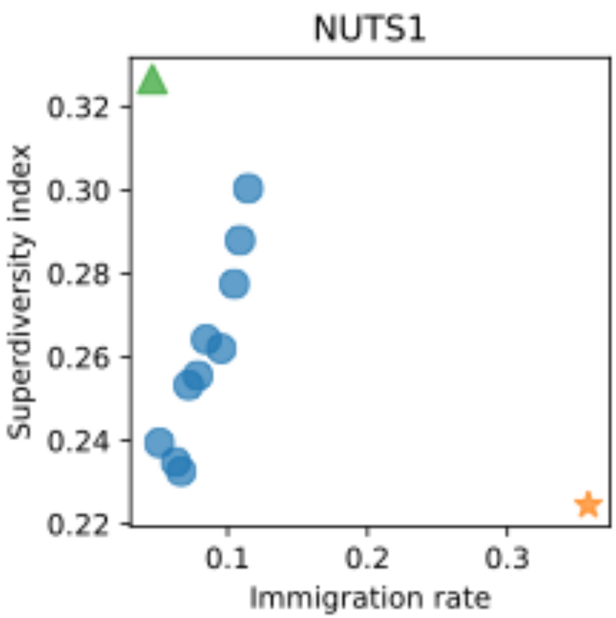
SI Evaluation



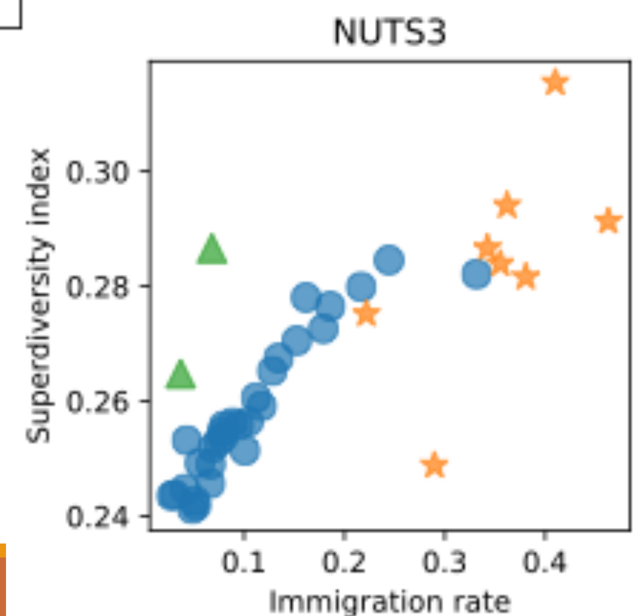





NUTS2
Superdiversity index
0.30
0.28
0.26
0.24
0.1
0.2
0.3
0.4
Immigration rate


NUTS1
Superdiversity index
0.32
0.30
0.28
0.26
0.24
0.22
0.1
0.2
0.3
Immigration rate

UK


NUTS3
Superdiversity index
0.30
0.28
0.26
0.24
0.1
0.2
0.3
0.4
Immigration rate

SI Evaluation

Compare with other possible indices and null model

- Null model obtained by shuffling tweets between regions, but maintaining the number of tweets
- Other indices:
  - Number of tweets (/capita), **number of languages**, entropy of languages, type-token ratio (TTR)

SI Evaluation

Italy

Geographical level	SI	null model SI	number of Italian tweets	number of Italian tweets per capita	number of languages	language entropy	TTR
NUTS1 (5 regions)	**0.963**	-0.437	0.735	0.696	0.183	-0.585	-0.727
NUTS2 (20 regions)	**0.859**	0.143	0.279	0.282	0.304	0.099	-0.243
NUTS3 (20 regions)	**0.924**	0.082	0.081	-0.148	0.216	0.021	0.091

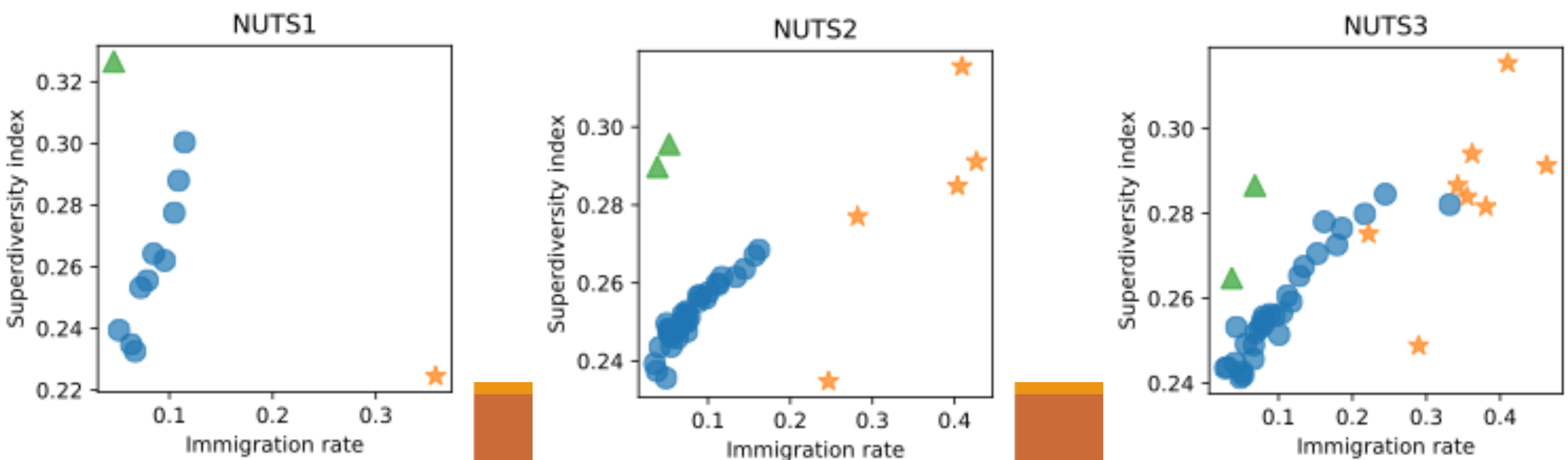
UK (except for London and Northeast England)

Geographical level	SI	null model SI	number of English tweets	number of English tweets per capita	number of languages	language entropy	TTR
NUTS1 (10 regions)	**0.943**	-0.236	0.328	-0.520	0.519	0.481	-0.005
NUTS2 (40 regions)	**0.941**	-0.137	0.332	0.007	0.362	0.288	-0.340
NUTS3 (40 regions)	**0.928**	-0.221	0.141	0.049	0.322	0.529	0.147

Discussion

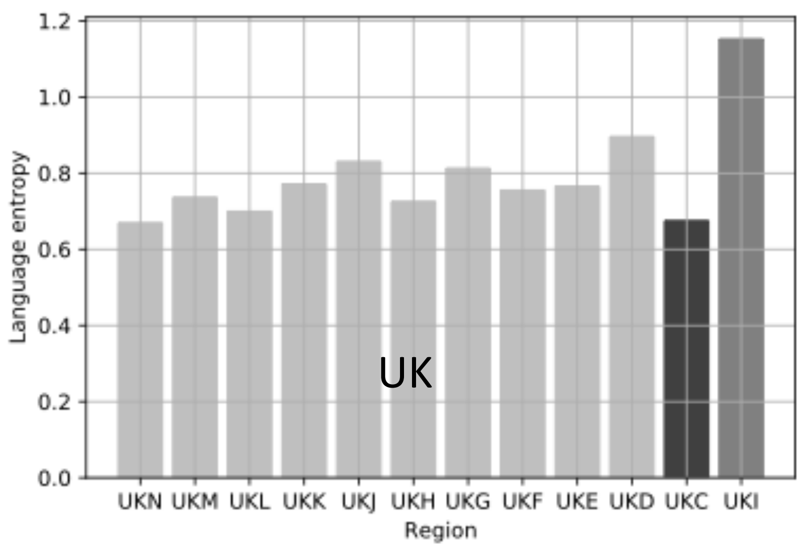
What about London and Northeast England?

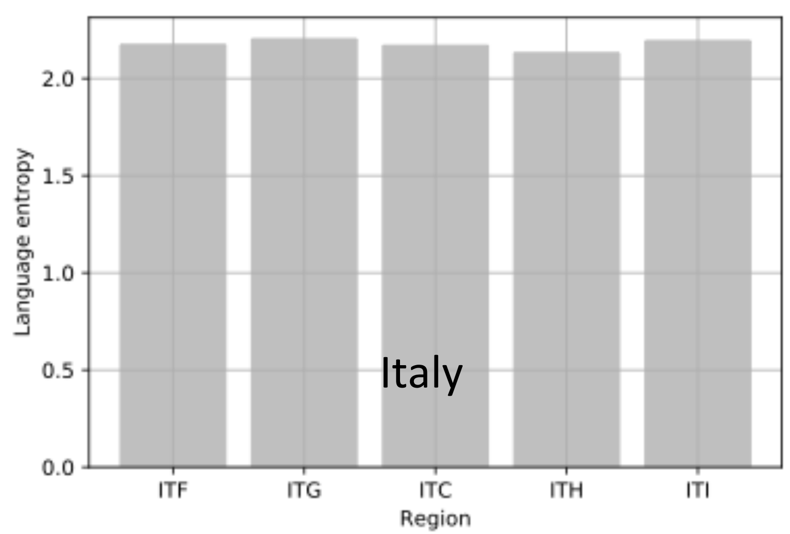
- SI appears to have different ranges in different region groups (and also spatial resolutions)
- Pre-clustering to identify these groups?



What about London and Northeast England?

- Pre-clustering to identify these groups?
  - Language entropy - some information
  - Need to identify further features - e.g. population density, local dialects





Can we use **superdiversity** to **nowcast immigration**?

- Correlations are, in general, very large.
- By adding pre-clustering and other correcting factors, within a ML model, it should be possible.
- We can estimate immigration levels at various resolutions where they are not available or up to date (clandestine immigration?)

Thank you!

Contact: alina.sirbu@unipi.it

References:

- Pollacci, Laura, Alina Sîrbu, Fosca Giannotti, Dino Pedreschi, Claudio Lucchese, and Cristina Ioana Muntean. *"Sentiment Spreading: An Epidemic Model for Lexicon-Based Sentiment Analysis on Twitter."* In AI*IA 2017.

- Pollacci, Laura, Alina Sîrbu, Fosca Giannotti, Dino Pedreschi, *"Measuring the 'Salad Bowl'- Superdiversity on Twitter."* Submitted.

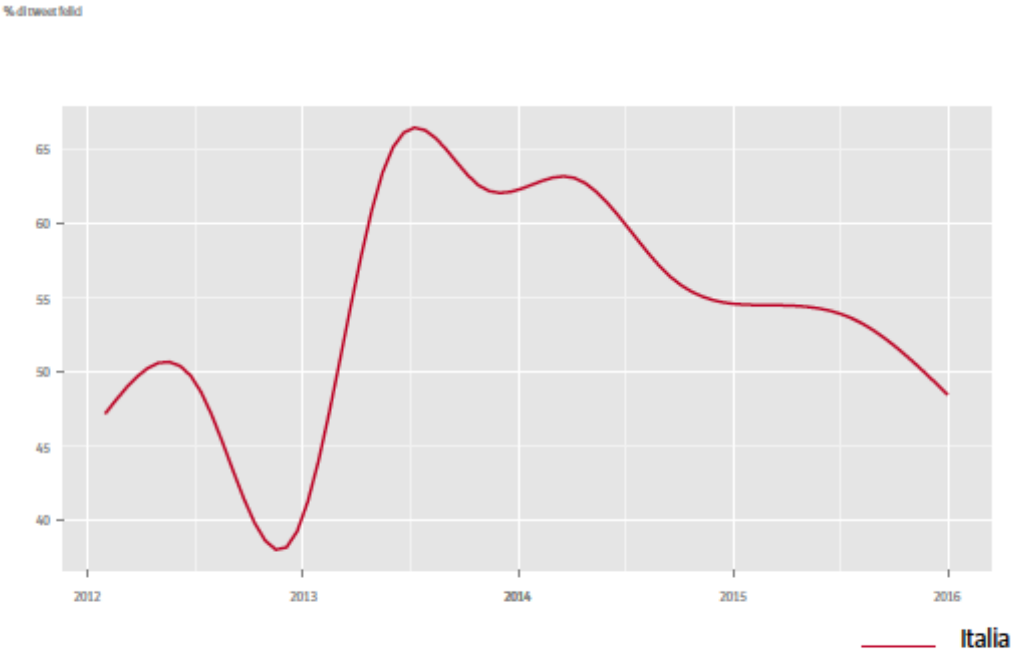
VOICES
from the Blogs

iHAPPY 2015

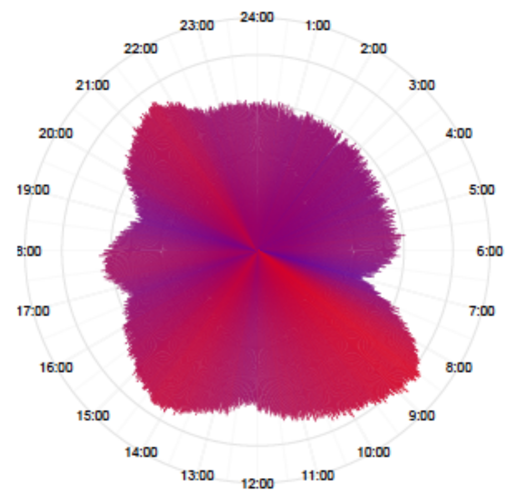
di
Andrea Ceron
Luigi Curini
Stefano M. Iacus

Politics and Big Data: Nowcasting and Forecasting Elections with Social Media. By Andrea Ceron, Luigi Curini, Stefano Maria

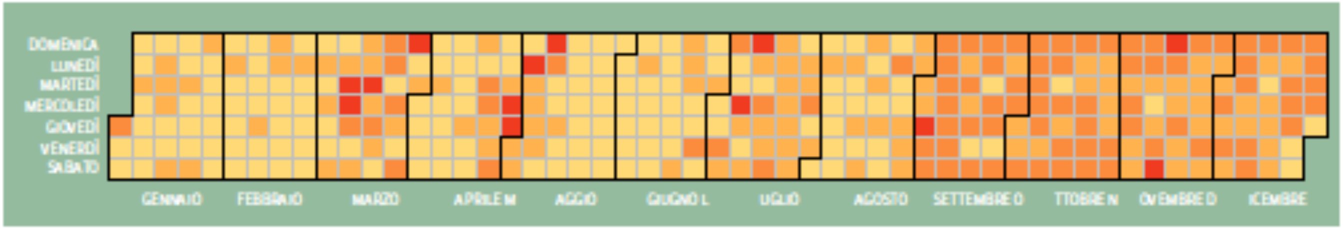
2012-15 ANDAMENTO DI *IHAPPY* IN ITALIA

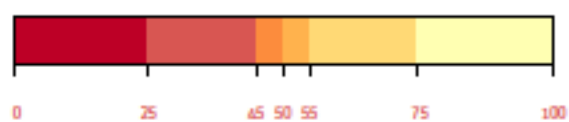


L'ORA ITALIANA PIÙ FELICE: *IHAPPY* MINUTO PER MINUTO

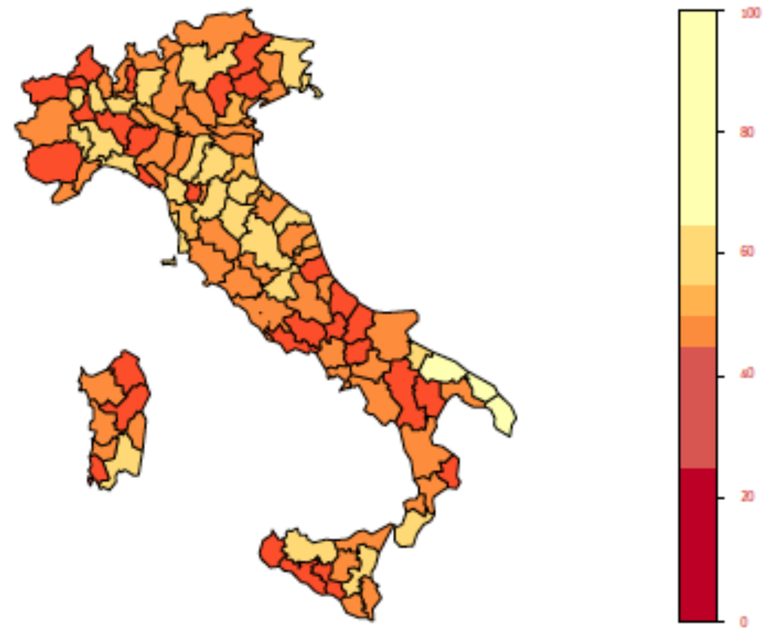


2015 CALENDARIO DELLA TWITTER-FELICITÀ





FELICITÀ MEDIA ANNUALE DELLE SINGOLE PROVINCE ITALIANE



LA CLASSIFICA REGIONALE

(La freccia indica se una regione è migliorata o peggiorata nella classifica rispetto al 2014)

Umbria 54,9% ↑	Puglia 54,4% ↓	Trentino Alto Adige 54% ↑	Toscana 53,6% ↑	Marche 53,3% ↓
Emilia Romagna 53,3% ↑	Piemonte 53,2% ↑	Liguria 53,1% ↓	Friuli V. G. 53% ↑	Sicilia 52,1% ↑
Lombardia 52% ↑	Sardegna 51,8% ↓	Calabria 51,7% ↓	Lazio 51,6% ↑	Veneto 51,6% ↑
Campania 51,5% ↑	Abruzzo 51,2% ↓	Basilicata 50,7% ↓	Valle d'Aosta 48,6% ↑	Molise 47,6% ↓

Text Analysis & Social Media

...dal rumore all'informazione

Come analizzare i social?

Re Tweet @re_assoluto · 30 apr
Doniamo una mamma di Baltimora a ognuno dei ragazzi di **#noexpo**

RETWEET 380 PREFERITI 387

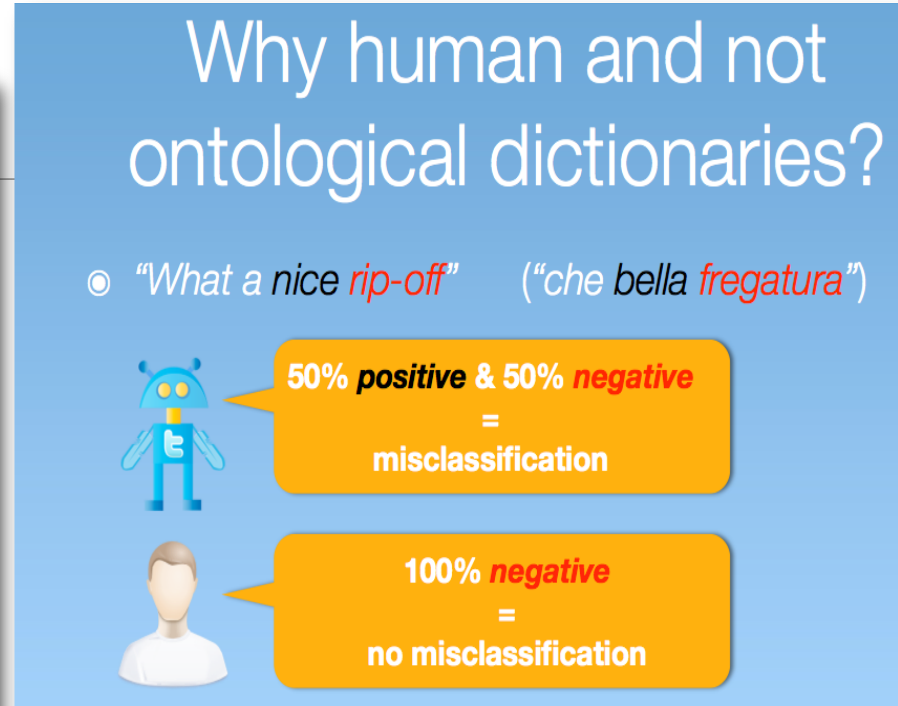
05:17 - 30 apr 2015 · Dettagli

[redacted] @[redacted] · 1 mag
I milanesi che, con spugna e sapone, ripuliscono i graffiti dei **#noexpo** sono il barlume di civiltà che ci serviva.

199 205 Foto

1) Non basta contare le mentions: Ma distinguere contenuto, ironia, giochi di parole, rumore, …

Come analizzare i social?



*"This movie has **good** premises. Looks like it has a **nice** plot, an **exceptional** cast, **first class** actors and Stallone gives his **best**. But it sucks"*

5 POSITIVE TERMS VS 1 NEGATIVE

2) Usare tecniche supervisionate no NLP

Come analizzare i social?

3) Stimare la distribuzione aggregata

Non cerchiamo l'ago nel pagliaio,

ma vogliamo sapere che aspetto ha l'intero pagliaio

Come analizzare i social?

4) Non solo "sentiment": analizziamo le opinioni

Tassonomia delle tecniche di analisi testuale

Principi della Text Analysis

Every quantitative linguistic model is wrong, but some can be useful

Principi della Text Analysis

Every quantitative linguistic model is wrong, but some can be useful

Quantitative methods help, but cannot replace human

Principi della Text Analysis

Every quantitative linguistic model is wrong, but some can be useful

Quantitative methods help, but cannot replace human

There exists not BEST or IDEAL technique of text analysis

Principi della Text Analysis

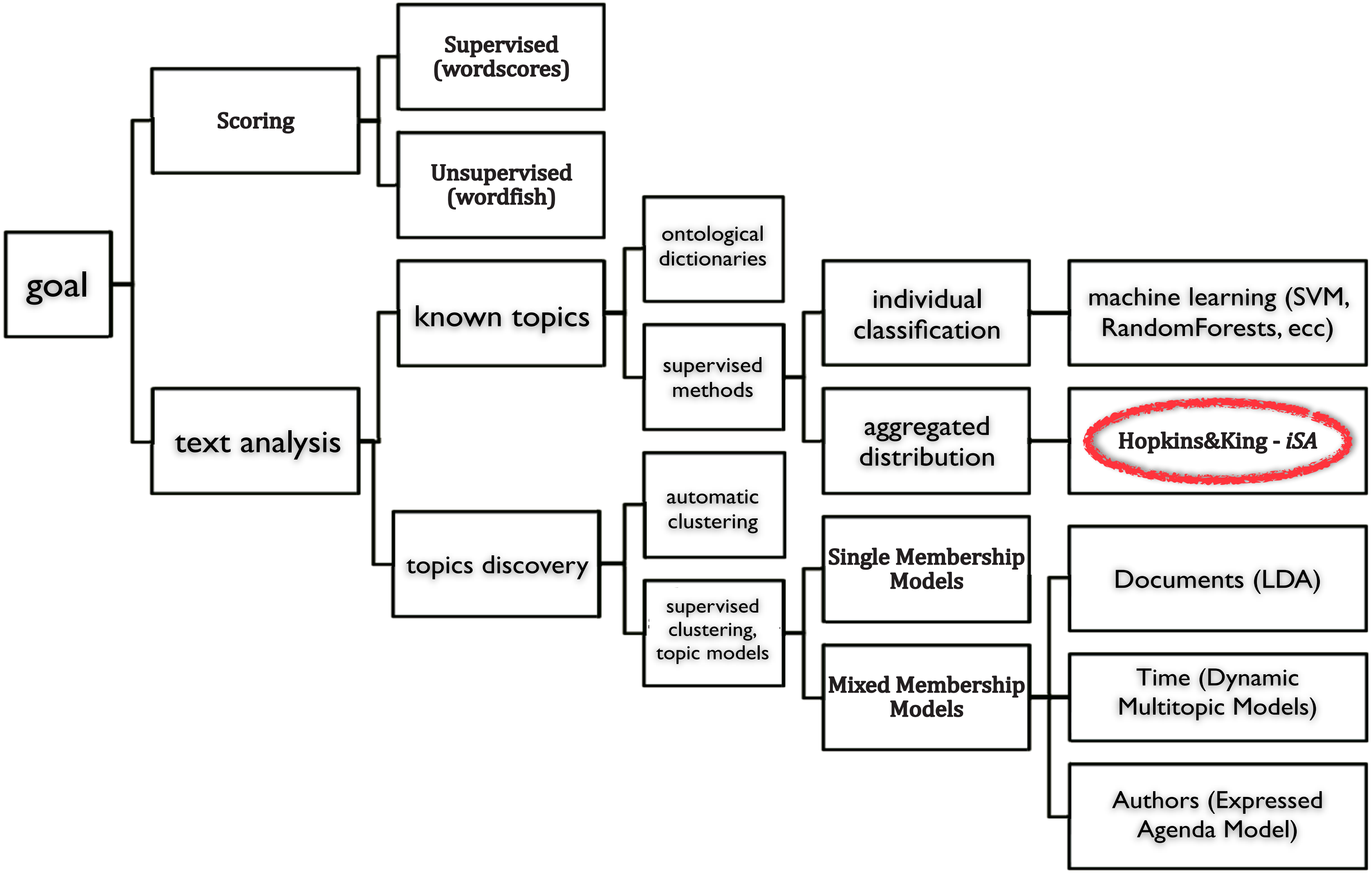
Every quantitative linguistic model is wrong, but some can be useful

Quantitative methods help, but cannot replace human

There exists not BEST or IDEAL technique of text analysis

Validate your analysis

Tassonomia delle tecniche



VOICES
from the Blogs

L'innovazione *i*SA ®

- La tecnologia iSA® (*integrated Sentiment Analysis*) sviluppata da **VOICES** rende possibile studiare i Big Data con la profondità di una analisi **qualitativa**

- iSA® è il **migliore algoritmo esistente al mondo** per efficacia, velocità di analisi e robustezza nello svolgere analisi sulle opinioni espresse sui Big Data

L’innovazione *i*SA®

Tool prima generazione	iSA®
Volumi di conversazione	**Significato** delle conversazioni
Sentiment positivo / negativo	**Motivazioni** dietro le opinioni
Conteggio parole	**Analisi** statistica
Dipendenti dalla lingua	**Indipendente** dalla lingua
Accuratezza medio-bassa (~85%)	**Accuratezza** alta (> 97%)

Come funziona in pratica?

Come funziona in pratica?

Il workflow tipico

Crawling

Twitter, FB	Blog, Forum, Web	Company data	Other data feeds

Preliminary Screeing

Preview results	Setting the goal

Tagging Manuale

Tagging coder 1	(…)	Tagging coder N

Text PreProcessing

Cleaning	Stemming

Analysis

iSA	CrossTab	Scoring	Etc

Reporting

Visualization	Technical reports	Data Feed	Data linkage

Come funziona in pratica?

post#1: "il nucleare conviene perché è economico"
post#2: "il nucleare produce scorie"
post#3: "il nucleare mi fa paura per le radiazioni, le scorie e non riduce l'inquinamento"

Si dividono i post dei social network in due gruppi:

✴il **train set** : ovvero i testi che verranno letti da codificatori umani e con i quali si istruisce l'agoritmo affinché possa eseguire le stime

✴il **test set**: ovvero i testi che non verranno letti dai codificatori ma con i quali l'algoritmo stimerà la distribuzione aggregata delle opinioni

Ogni testo, sia del train che del test set, viene scomposto in stilemi/ parole detti "stem" dall'algoritmo

IsA at work

Post1: Nuclear energy is convenient as it is cheaper

Post2: Nuclear energy produces waste

Post3: Nuclear scarry me bacause of radiation

Training Set: produced by humans (annotators), used to train the model

Test Set: used to estimate the distribution of the opinion

post#1: “il nucleare conviene perché è economico”
post#2: “il nucleare produce scorie”
post#3: “il nucleare mi fa paura per le radiazioni, le scorie e non riduce l’inquinamento”

Codifica manuale — **Stemming**

	Post	Di	Word: nucleare	Word: paura	Word: radiazioni	Word: inquinamento	Word: scorie	Word: economico
train set	post#1	a favore	1	0	0	0	0	1
***test* set**	post#2	NA	1	0	0	0	1	0
train set	post#3	contro	1	1	1	1	1	0
train set	post#4	contro	1	1	1	1	1	0
train set	post#5	a favore	1	0	1	0	0	1
	...	...	...	...	...	...	...	...
test set	post#1000	NA	1	0	0	0	0	1

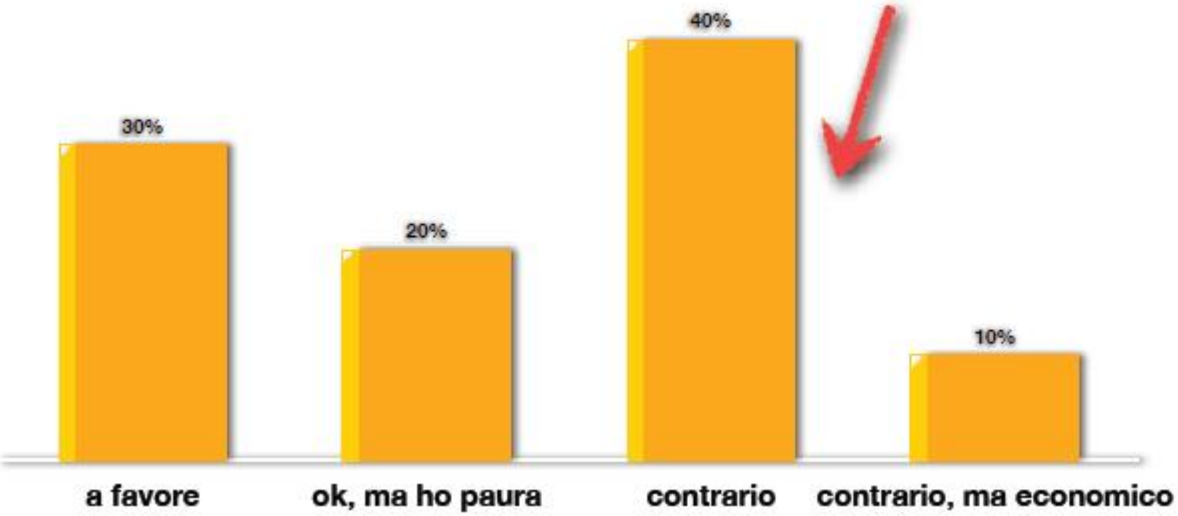
M

train set

Post	Di	Word: nuclearE	Word: paura	Word: radiazioni	Word: inquinamento	Word: scorie	Word: economico
post#1	a favore	1	0	0	0	0	1

Di = "a favore" Si = (1,0,0,0,0,1)

Goal: stima della distribuzione **P(D)**



	Post	Di	Word: nucleare	Word: paura	Word: radiazioni	Word: inquinamento	Word: scorie	Word: economico
train set	post#1	a favore	1	0	0	0	0	1
test set	post#2	NA	1	0	0	0	1	0
train set	post#3	contro	1	1	1	1	1	0
train set	post#4	contro	1	1	1	1	1	0
train set	post#5	a favore	1	0	1	0	0	1
	...	...	...	...	...	...	...	...
test set	post#1000	NA	1	0	0	0	0	1

Dizionario Treccani (italiano): 270k lemmi
Oxford Dictionary (English): 650k lemmi

In realtà, per ciascun argomento nel linguaggio comune si tende ad utilizzare al massimo M = 200 o 500 "stilemi" e questo rende possibile effettuare l'analisi statistica

L'analisi resta difficile da approcciare perché avremo potenzialmente 2^M righe diverse composte da 0 e 1 (ovvero tra 1,6*10^50 e 3,3*10^150)

Approcc

goal — train & test — train+test

$$P(D) = P(D|S) * P(S)$$

modello statistico classico produce missclassification

distribuzione degli stem

All machine learning methods affected, choose your own and add to the list below:

- Support Vector Machines
- Random Forests
- Neural Networks
- ecc

Problemi di tipo statistico: questo approccio cerca di rintracciare le opinioni considerando tutti i 2^M possibili "stilemi" → time-consuming e non garantisce risultati accurati perché diventa improbabile rintracciare "stilemi" che stiano davvero esprimendo un'opinione

Met

Approccio statistico innovativo *(King&Hopkins, 2010)*

train+test *train* *goal*

$$P(S) = P(S|D) * P(D)$$

$$P(S|D)^{-1} * P(S) = P(D)$$

Si guarda alla distribuzione degli Stem in ciascuna categoria di “opinioni” e non il contrario

“Semplice” quanto invertire una matrice!

Nessun problema a gestire la quantità di “Big” Data

Met

Approccio statistico innovativo *(King&Hopkins, 2010)*

train+test — *train* — *goal*

$$P(S) = P(S|D) * P(D)$$

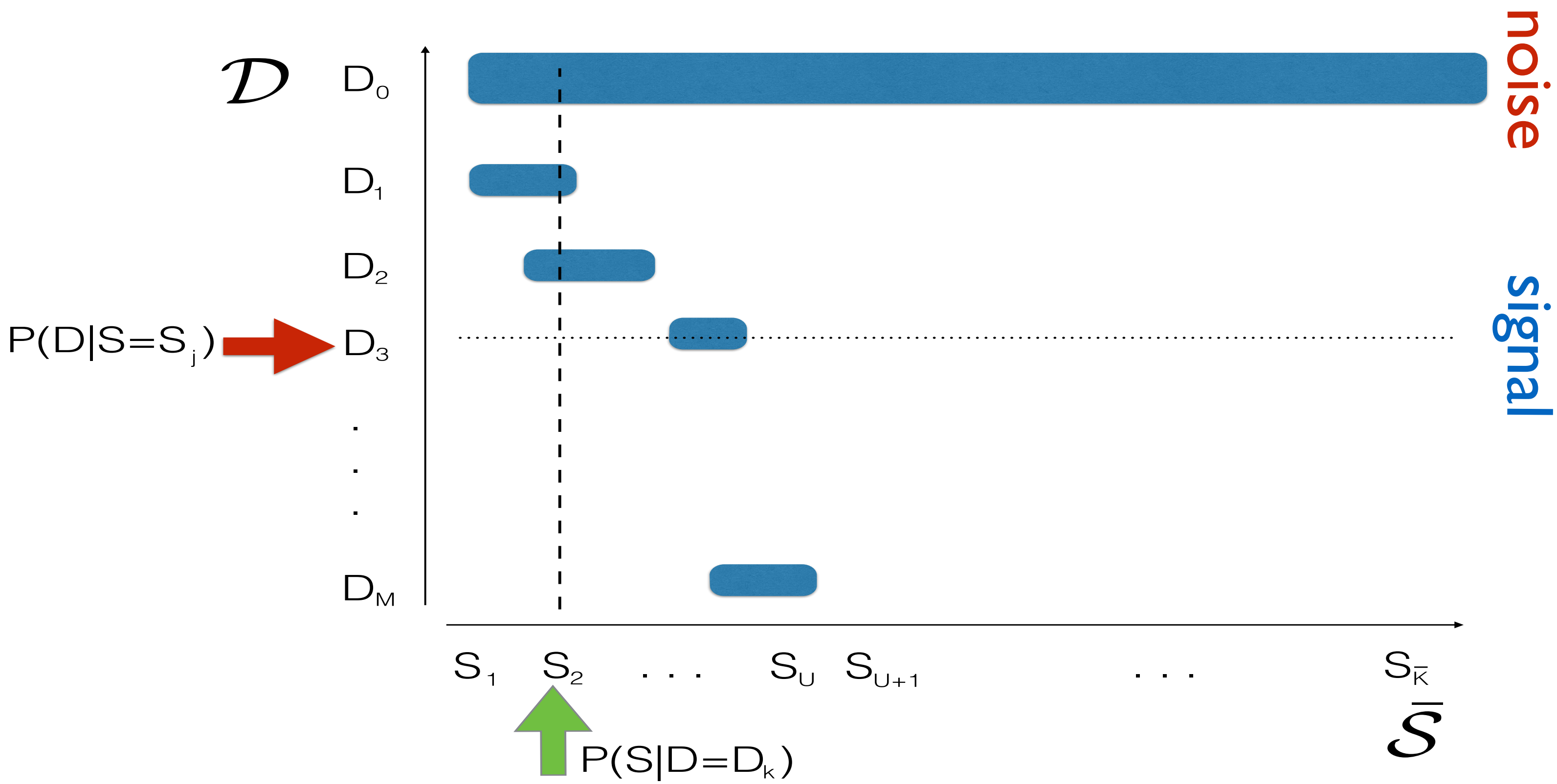
$$P(S|D)^{-1} * P(S) = P(D)$$

Si guarda alla distribuzione degli Stem in ciascuna categoria di “opinioni” e non il contrario

“Semplice” quanto invertire una matrice!

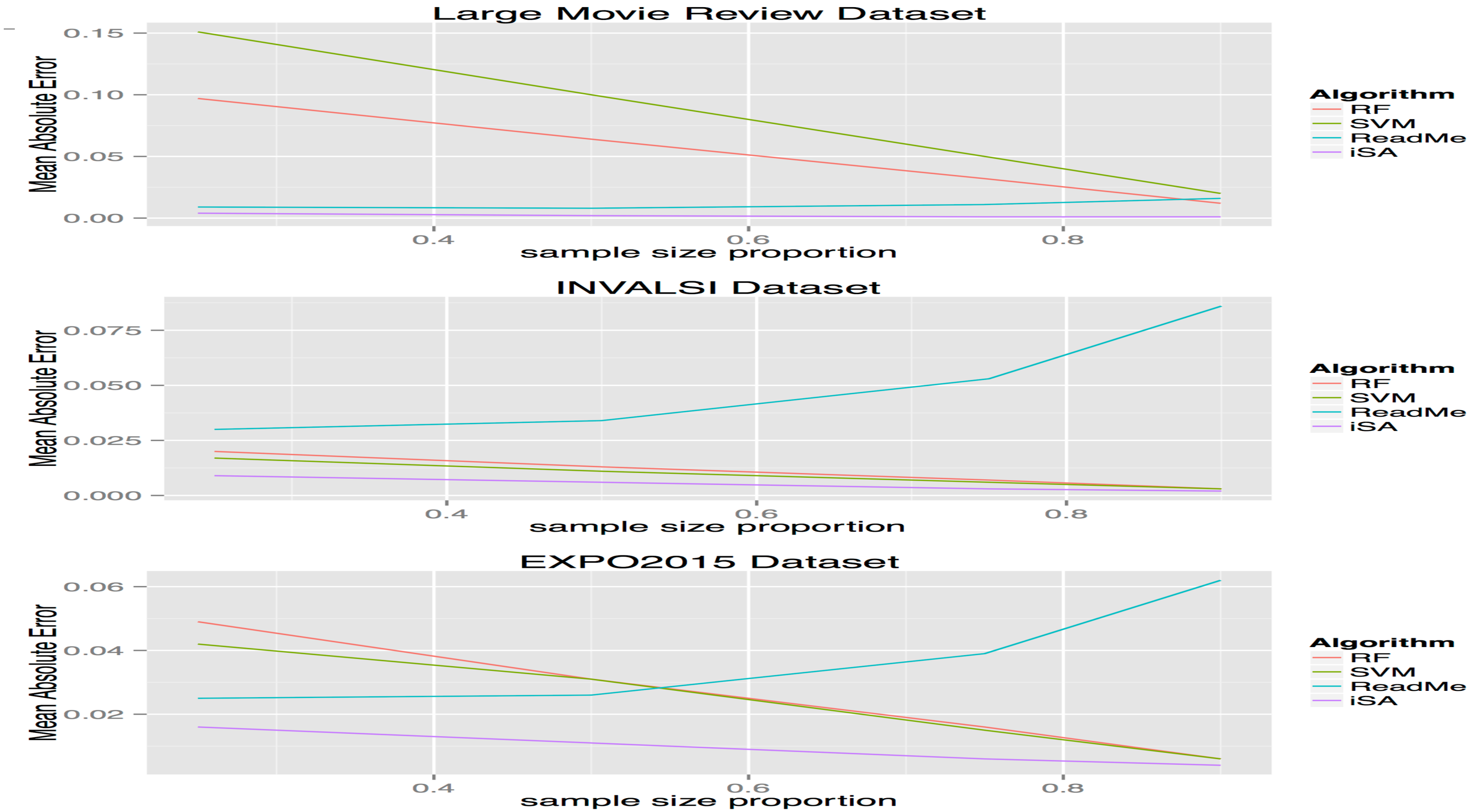
Nessun problema a gestire la quantità di “Big” Data

The space "Opinion x Stems" = $\mathcal{D} \times \bar{\mathcal{S}}$



Tipica performance statistica di iSA vs competitors

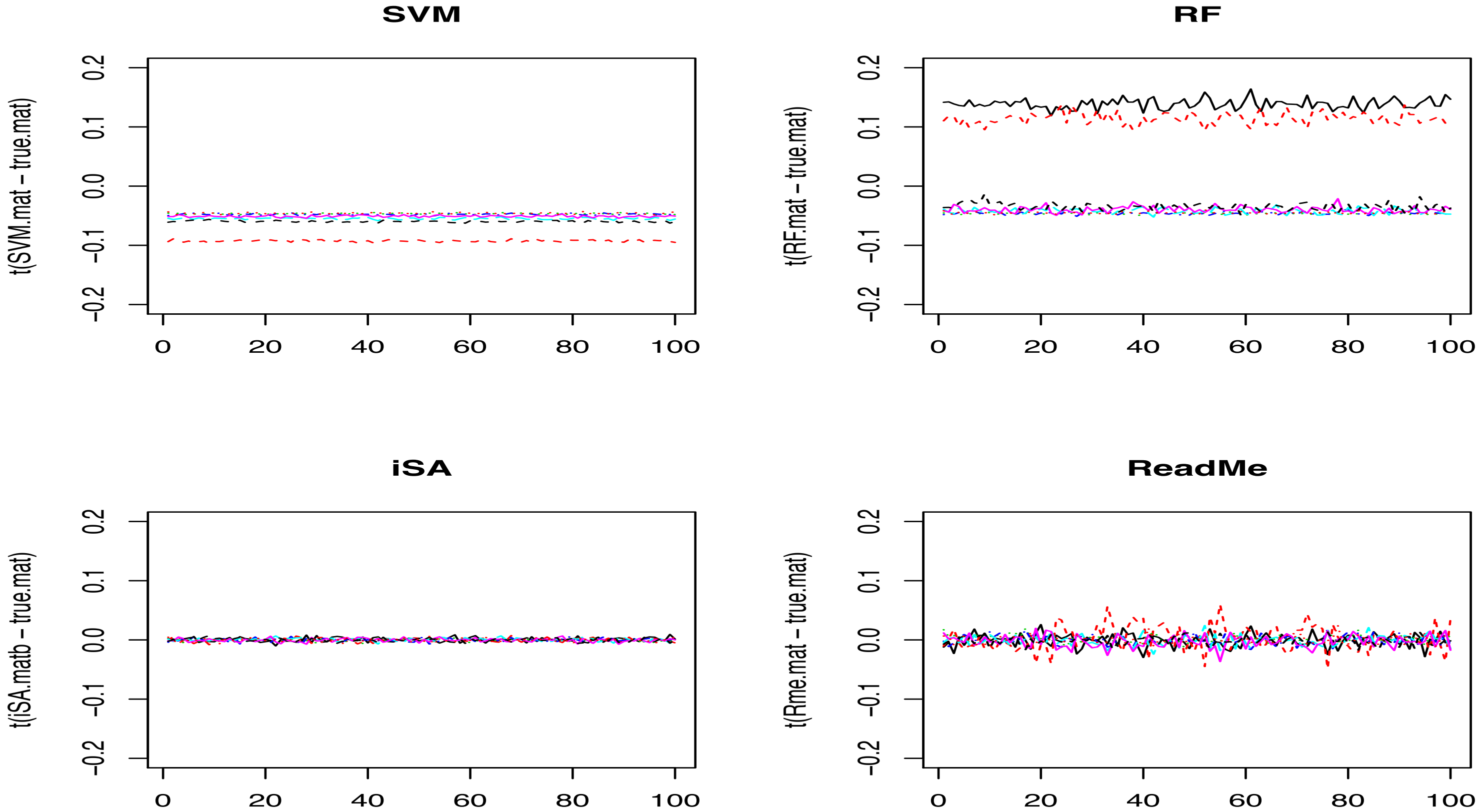
"Bias"



VOICES
from the Blogs

Tipica performance statistica di iSA vs competitors

"Precision" (Large Movie dataset)



Forecasting Peruvian Elections with twitter

Sentiment-enhanced Multidimensional Analysis of Online Social Networks:

Perception of the Mediterranean Refugees Crisis

Mauro Coletto ∗†, Andrea Esuli †, Claudio Lucchese †, **Cristina Ioana Muntean** †, Franco Maria Nardini †, Raffaele Perego †, Chiara Renso †

∗ IMT School for Advanced Studies Lucca - ITALY
† ISTI - CNR Pisa - ITALY

Refugees Crisis Perception Analysis

AQ1: What is the evolution of the discussions about refugees migration in Twitter?

AQ2: What is the sentiment of users across Europe in relation to the refugee crisis? What is the evolution of the perception in countries affected by the phenomenon?

AQ3: Are users more polarized in countries most impacted by the migration flow?

Analytical Framework

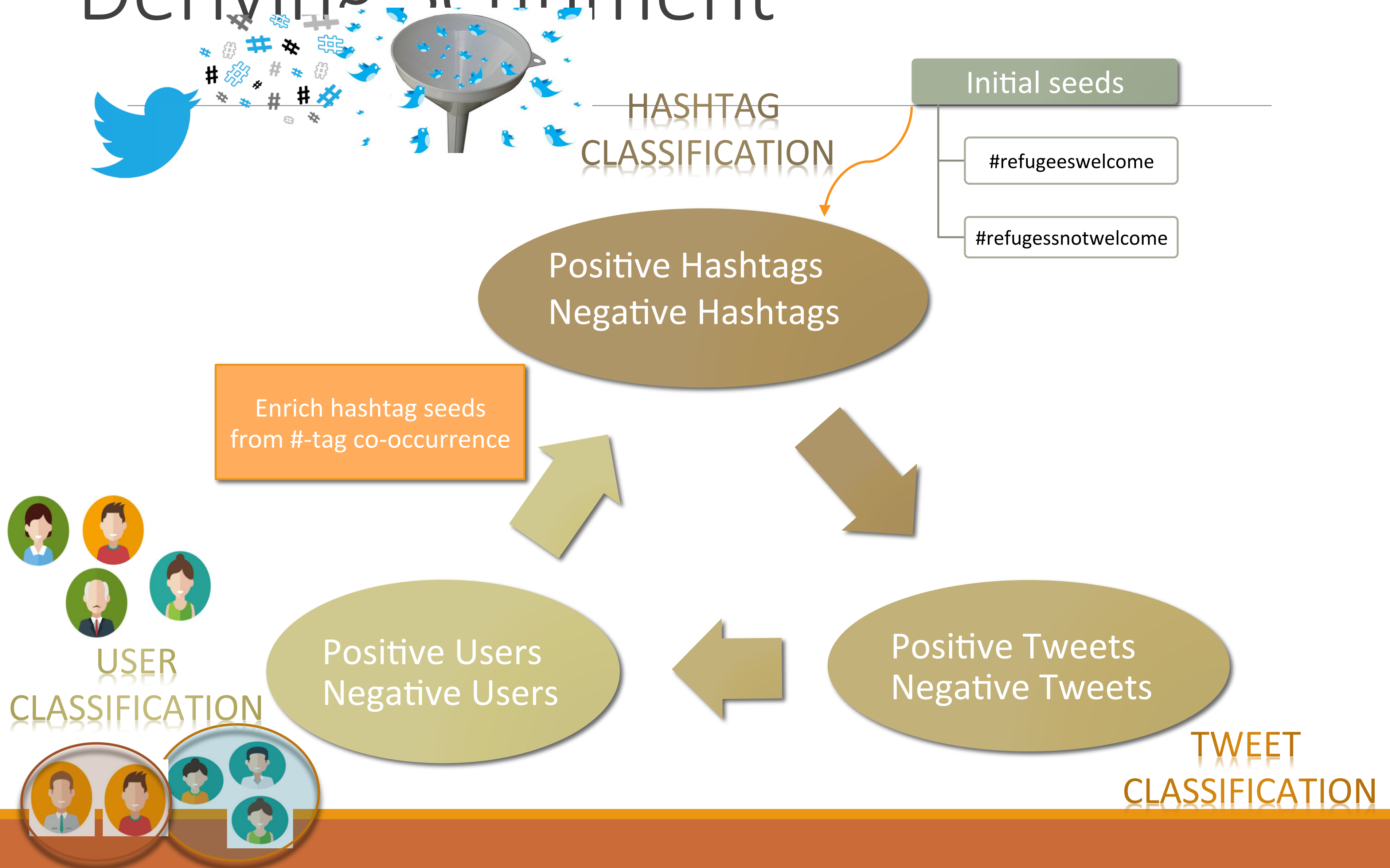
An analytical framework to interpret social trends from large tweet collections by extracting and crossing information about the following three dimensions:

- Time
- Space
  - User location
  - Location mentions
- Sentiment
  - Tweet sentiment
  - User sentiment

Symbol	Description	# Total
$\mathcal{G}$	Collected English tweets	97,693,321
$\mathcal{T}$	Tweets related to the refugee crisis	1,238,921
$\mathcal{T}_{c+}$	Positive sentiment tweets	459,544
$\mathcal{T}_{c-}$	Negative sentiment tweets	387,374
$\mathcal{T}_{ML}$	Tweets with mentioned location	421,512
$\mathcal{T}_{UL}$	Tweets with user location	101,765
$\mathcal{U}$	Users	480,660
$\mathcal{U}_{c+}$	Users with positive sentiment	213,920
$\mathcal{U}_{c-}$	Users with negative sentiment	104,126
$\mathcal{U}_{L}$	Users with country location	47,824

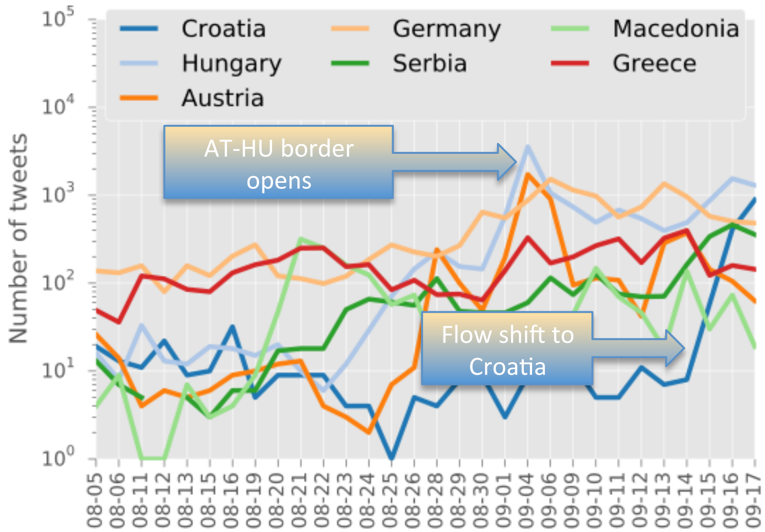
Perform multidimensional analyses considering content and locations in time

Deriving Sentiment

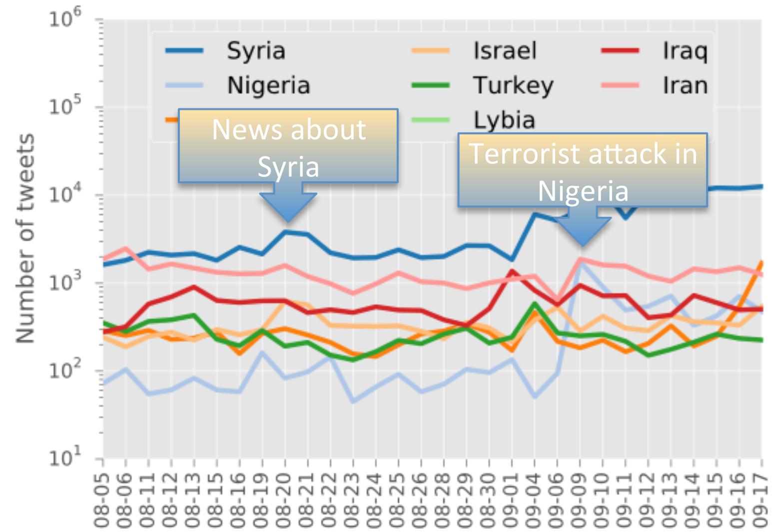


Sentiment on migration topics: Perception of the Mediterranean Refugee Crisis

European country mentions

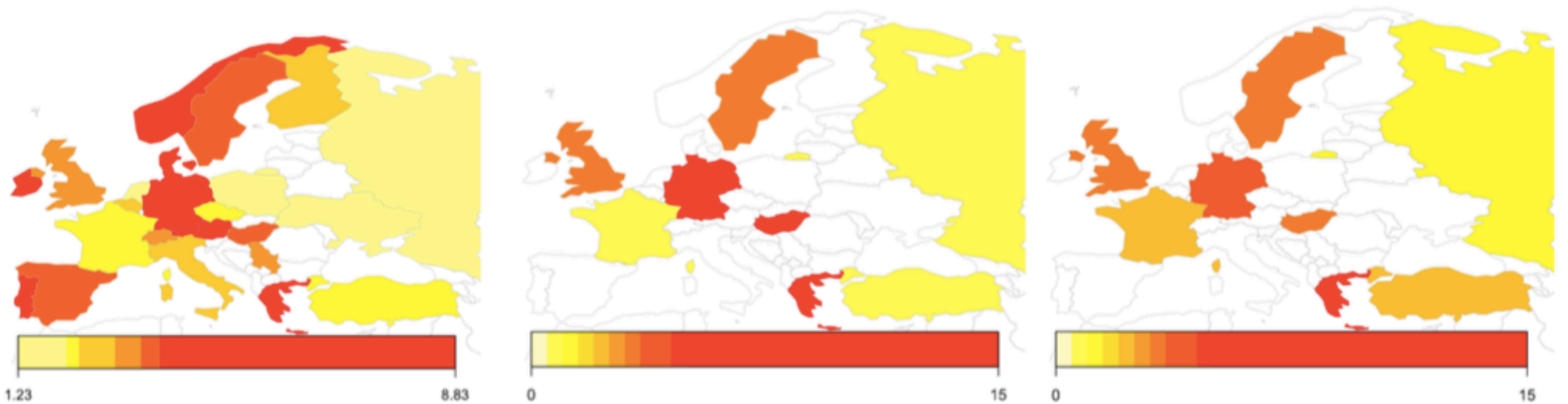


Africa & Middle East country mentions



Sentiment on migration topics: Perception of the Mediterranean Refugee Crisis

- Internal and external perception by country
  - Index **ρ** - the ratio between pro refugees users and against refugees users
  - Red means a higher predominance of positive sentiment, higher ρ
  - Yellow means a higher predominance of negative sentiment, lower ρ

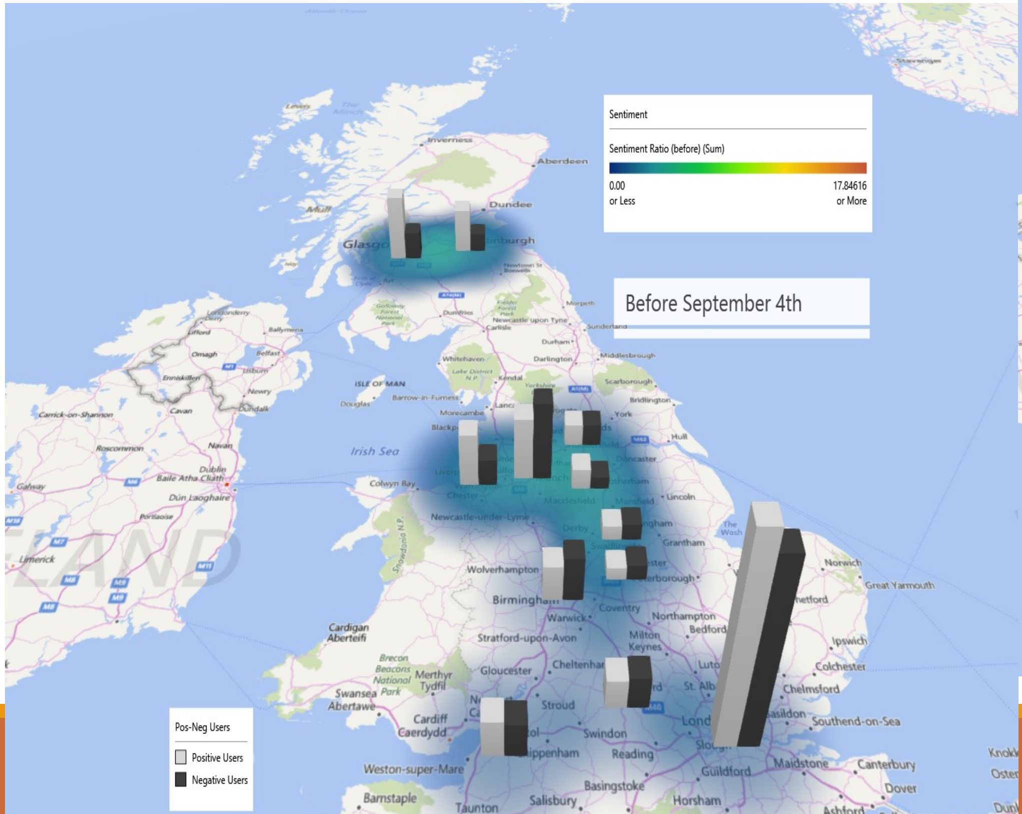


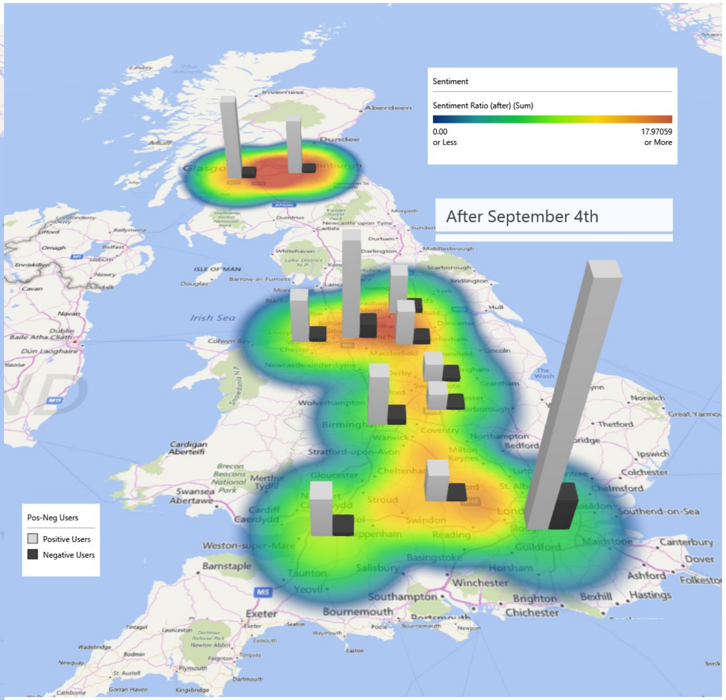
(a) Global perception (b) Internal perception (c) External perception

Sentiment Analysis in UK

Positive and negative users for different cities in UK before and after September 4 (death of Alan Kurdi, borders between AT HU, Germany welcomes refugees).

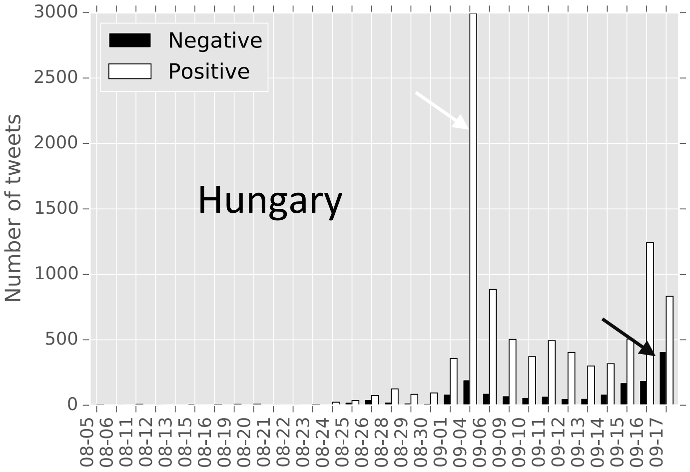
- bars show the number of polarized positive and negative users by city
- the heat map in background indicates the value of ρ

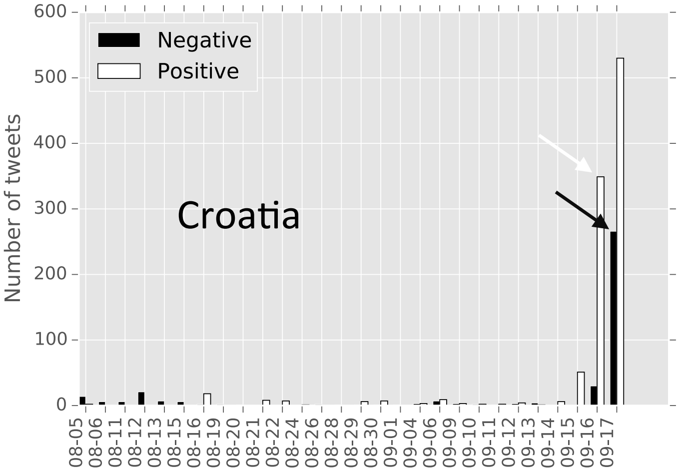


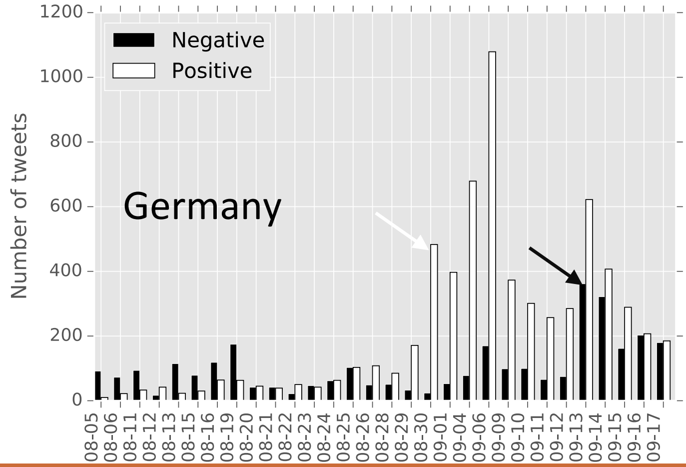


Sentiment Analysis

For mentioned locations analysis and tweet sentiment

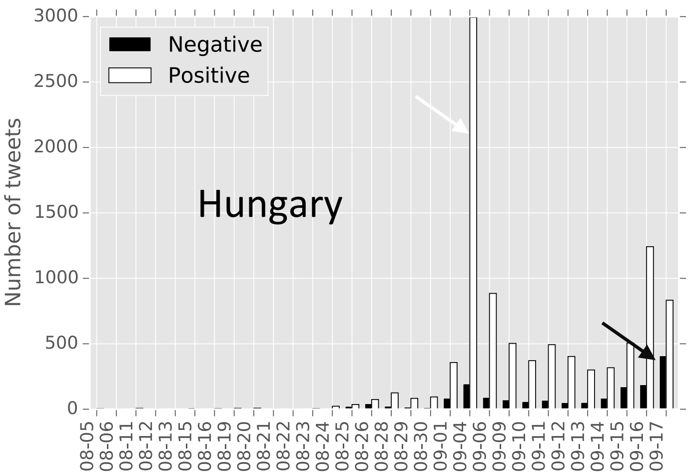


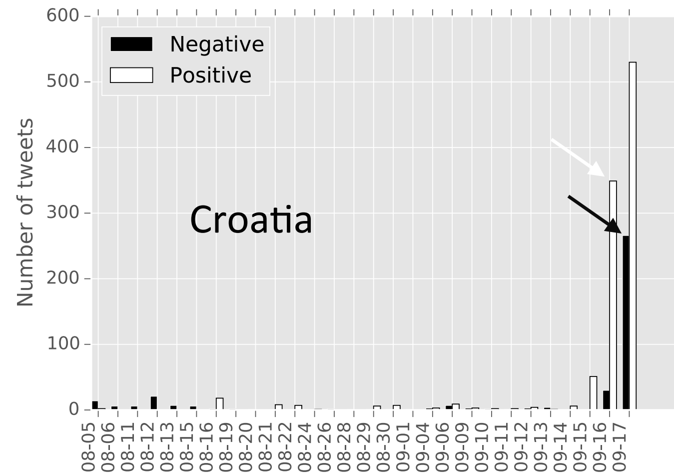


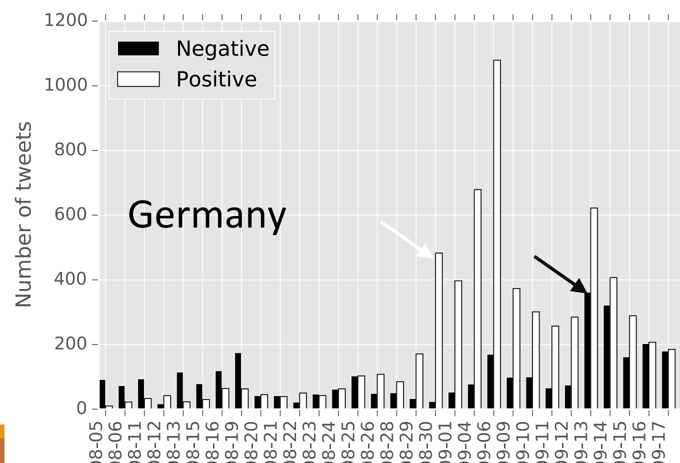


Sentiment Analysis

For mentioned locations analysis and tweet sentiment

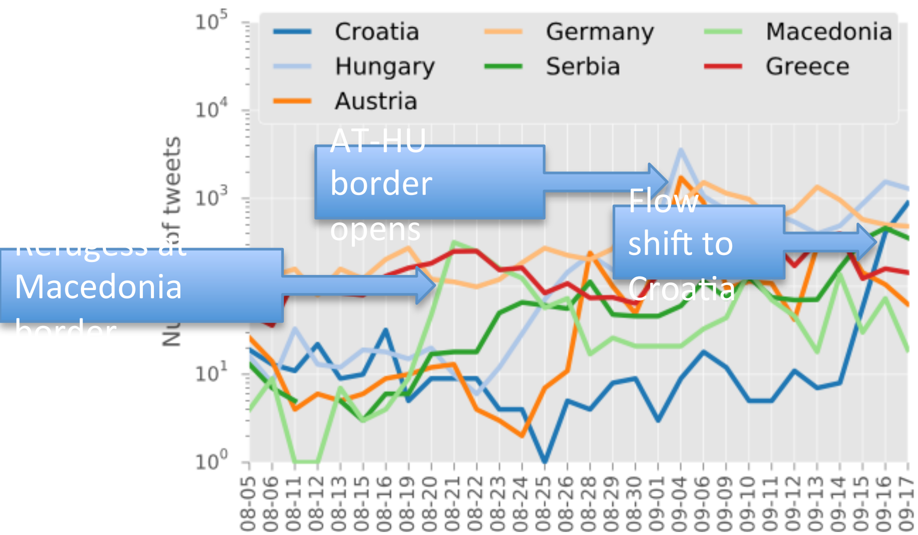




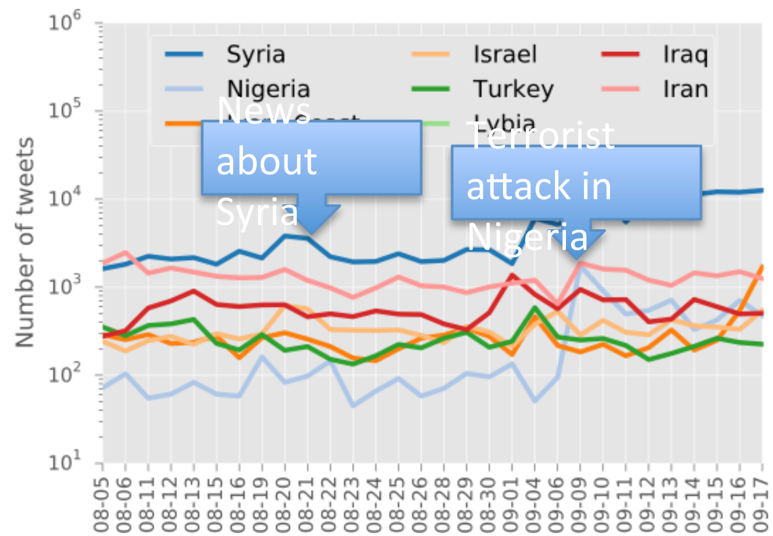


Space and Time analysis

European country mentions



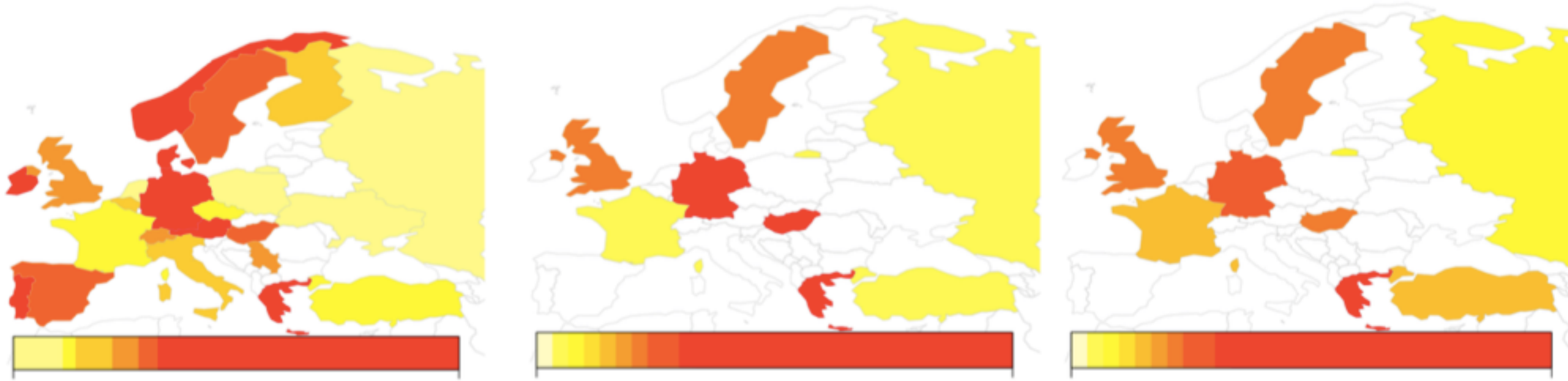
Africa & Middle East country mentions



Sentiment Analysis

Internal and external perception by country

- Index **ρ** - the ratio between pro refugees users and against refugees users
- Red means a higher predominance of positive sentiment, higher ρ



- +

(a) Overall.

- +

(b) Internal perception.

- +

(c) External perception.

THANK YOU !

Questions?