Diachronic Analysis of Language exploiting Google Ngram

Dr Annalina Caputo
ADAPT Centre

Diachronic Linguistics

The scientific study of language change over time
also called **Historical Linguistics**

Synchronic vs.
Diachronic

Synchronic
It describes the language rules at a specific point in time without taking its history into account.

Diachronic
It considers the evolution of a language over time.

Diachronic Linguistics

Why?

- Observe changes in particular languages
- Reconstruct the pre-history of languages
- Develop general theories about how and why language changes
- Describe the history of speech communities
- Etymology

https://en.wikipedia.org/wiki/Historical_linguistics

Google Book Ngram

5,195,769 books

4% all published books

500 billion words

1500-2012 time span

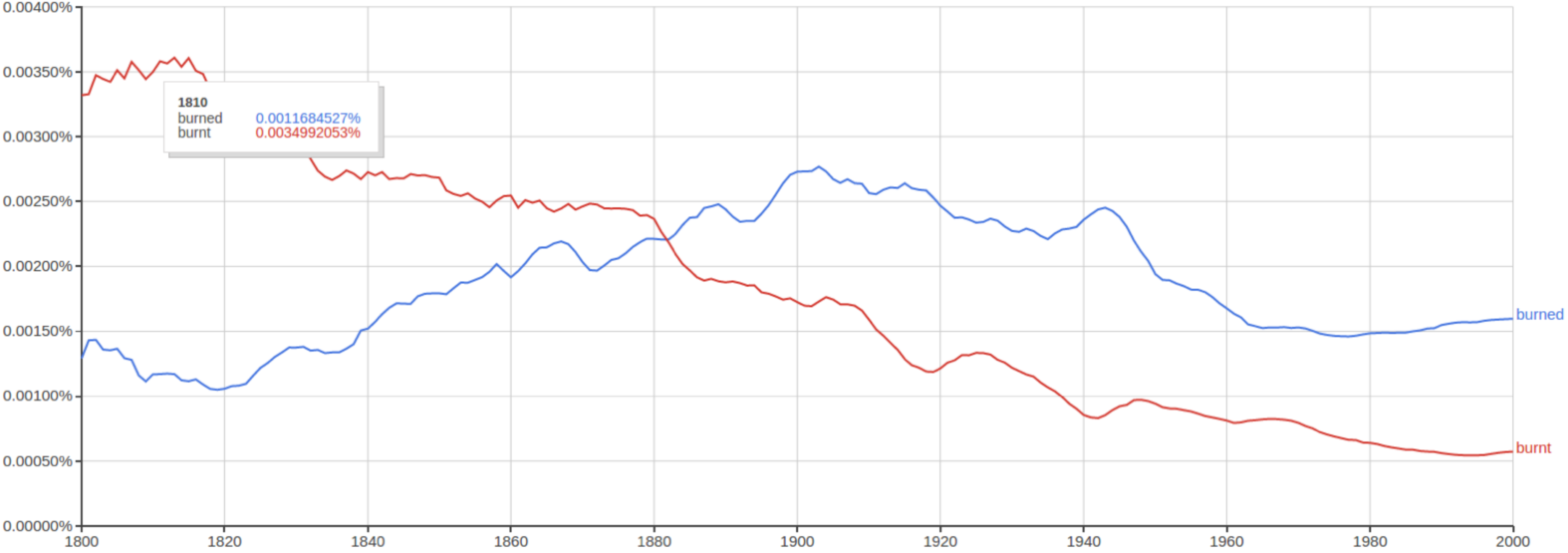
CULTUROMICS

A form of computational lexicology that studies **human behavior** and **cultural trends** through the **quantitative analysis** of digitized texts.

J.-B. Michel et al., *Quantitative Analysis of Culture Using Millions of Digitized Books,* Science, 2011

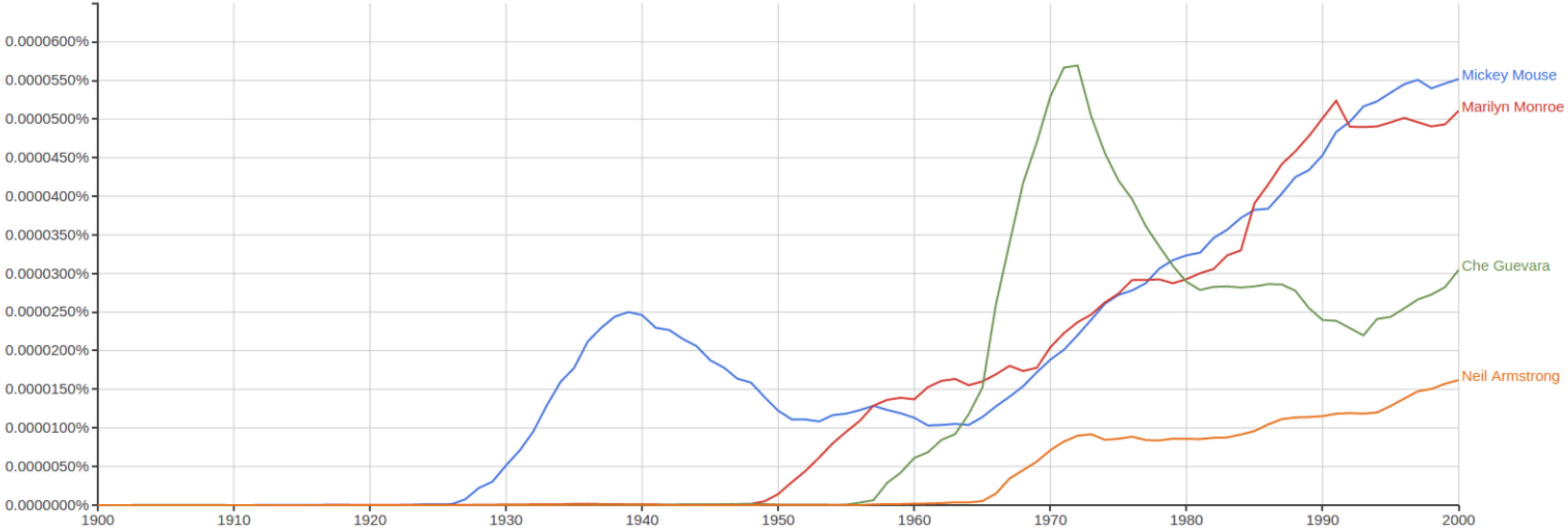
Culturomics

Grammar Evolution



J.-B. Michel et al., *Quantitative Analysis of Culture Using Millions of Digitized Books,* Science, 2011

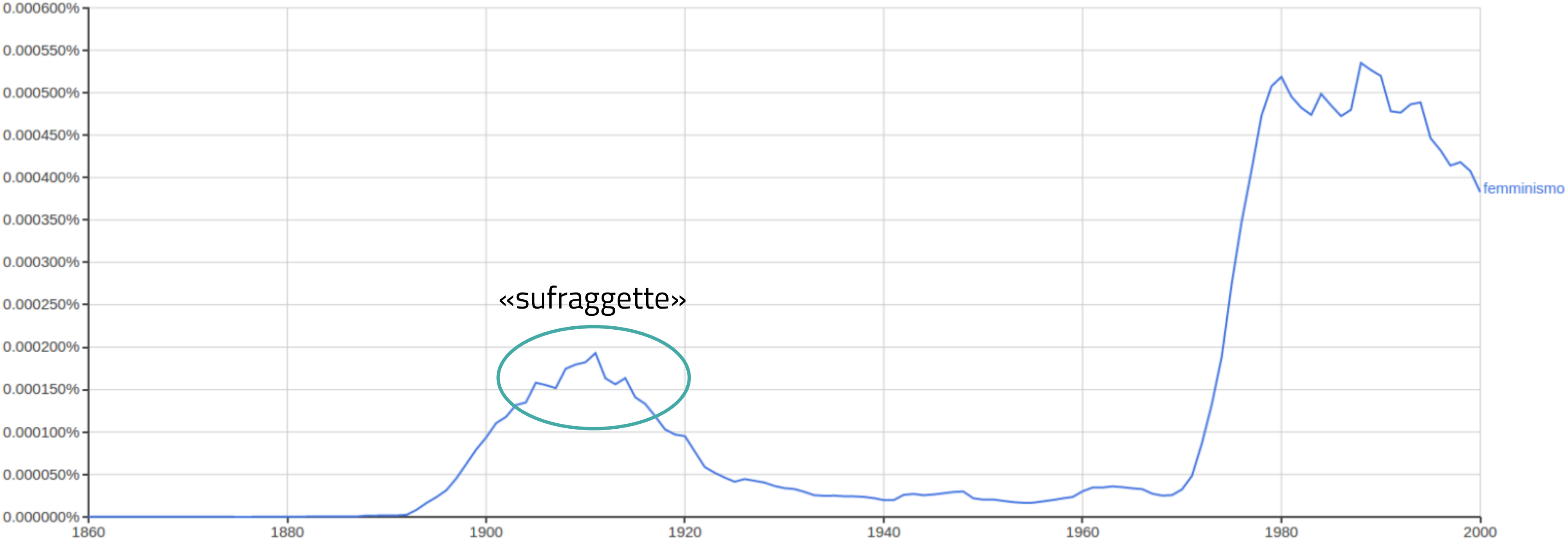
Culturomics
Popularity



J.-B. Michel et al., *Quantitative Analysis of Culture Using Millions of Digitized Books,* Science, 2011

Culturomics

feminism (Italian)

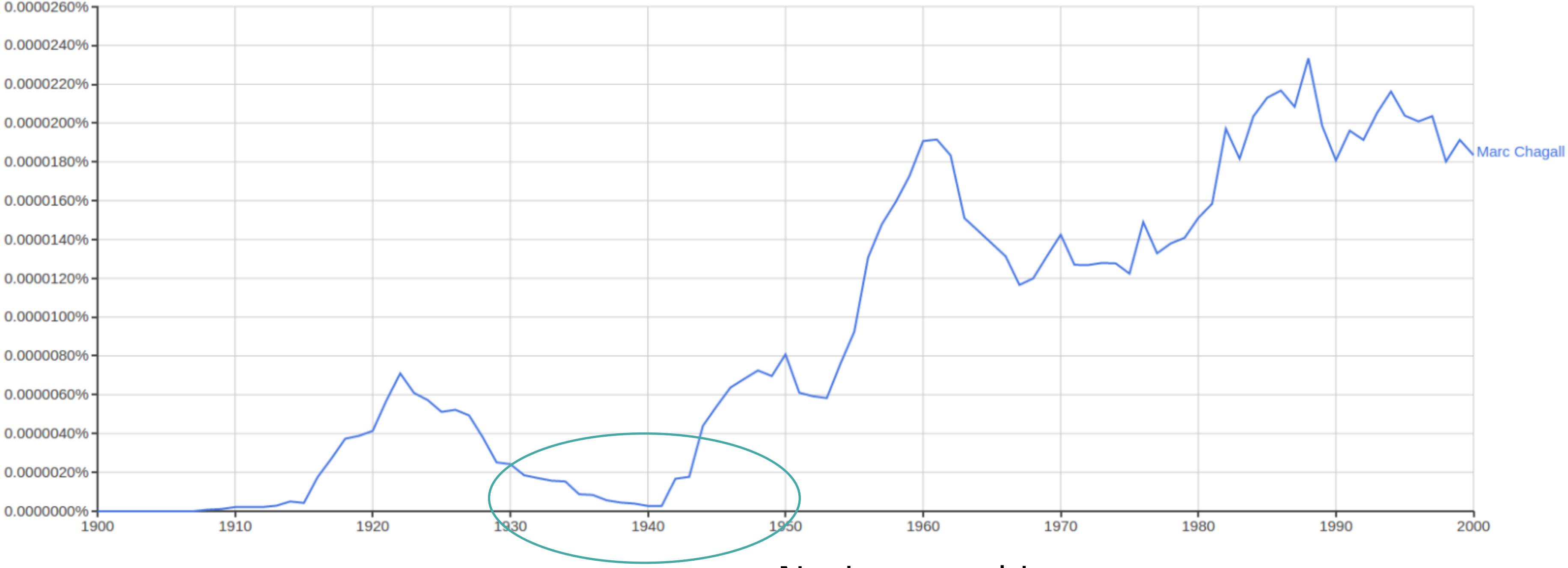


J.-B. Michel et al., *Quantitative Analysis of Culture Using Millions of Digitized Books,* Science, 2011

Culturomics

Censorship

Marc Chagall (German)

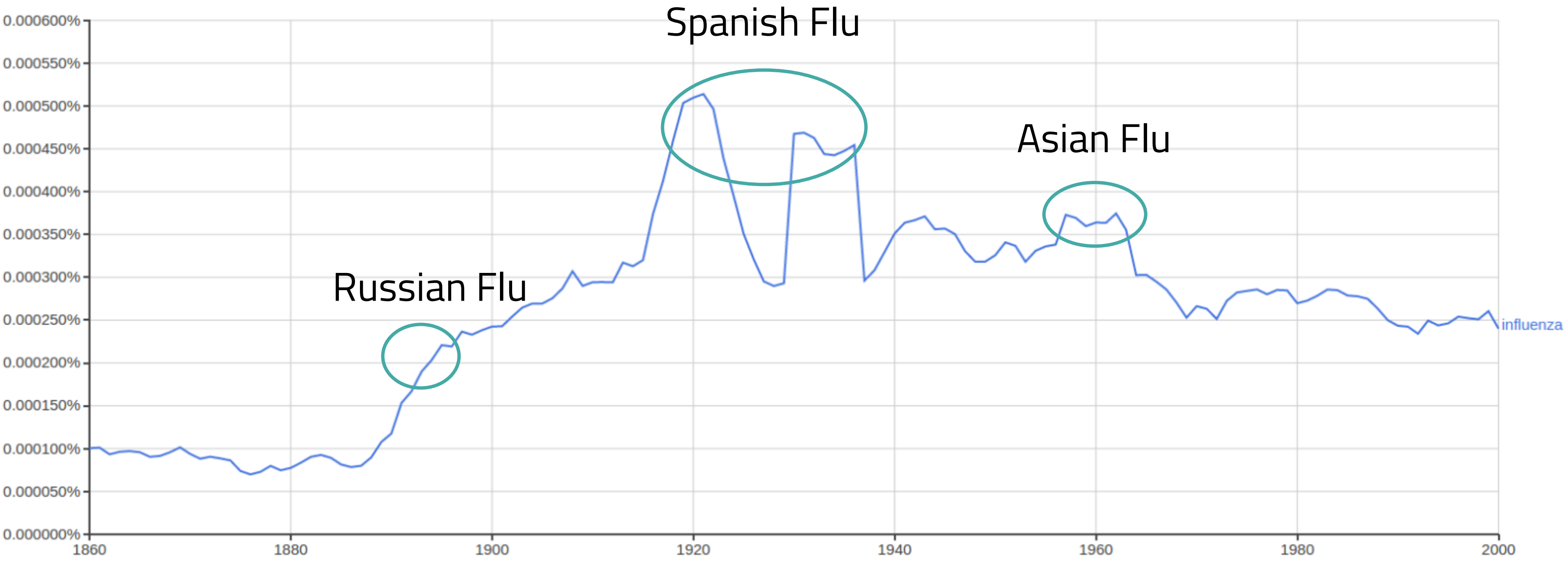


Nazi censorship

J.-B. Michel et al., *Quantitative Analysis of Culture Using Millions of Digitized Books,* Science, 2011

Culturomics

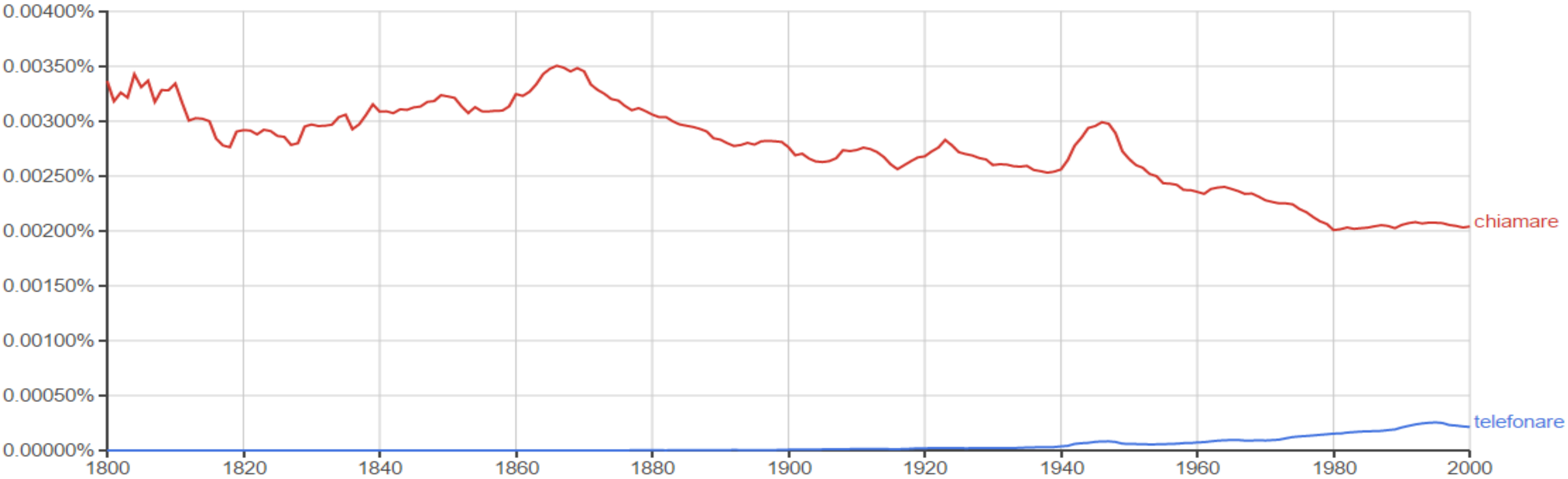
Events



J.-B. Michel et al., *Quantitative Analysis of Culture Using Millions of Digitized Books*, Science, 2011

Limit

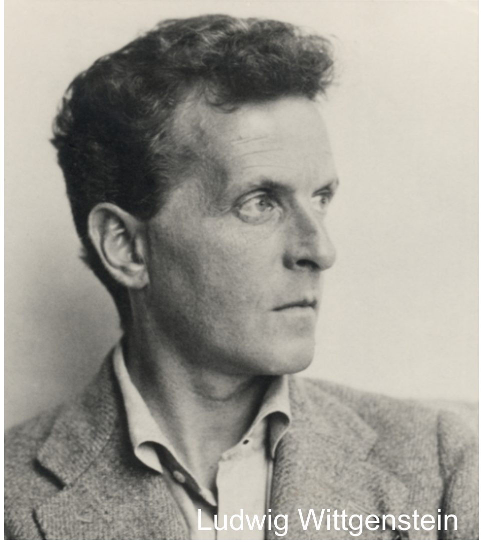
call (chiamare) vs. phone (telefonare)



Distributional semantic models

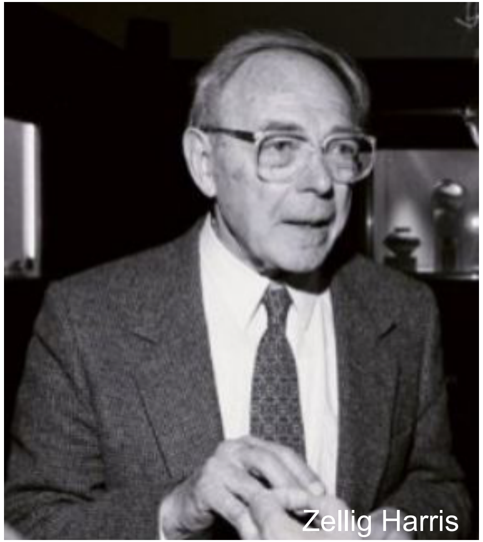
You shall know a word by the company it keeps!

https://goo.gl/nY4els



Meaning of a word is determined by its usage.

https://goo.gl/mD1oKn

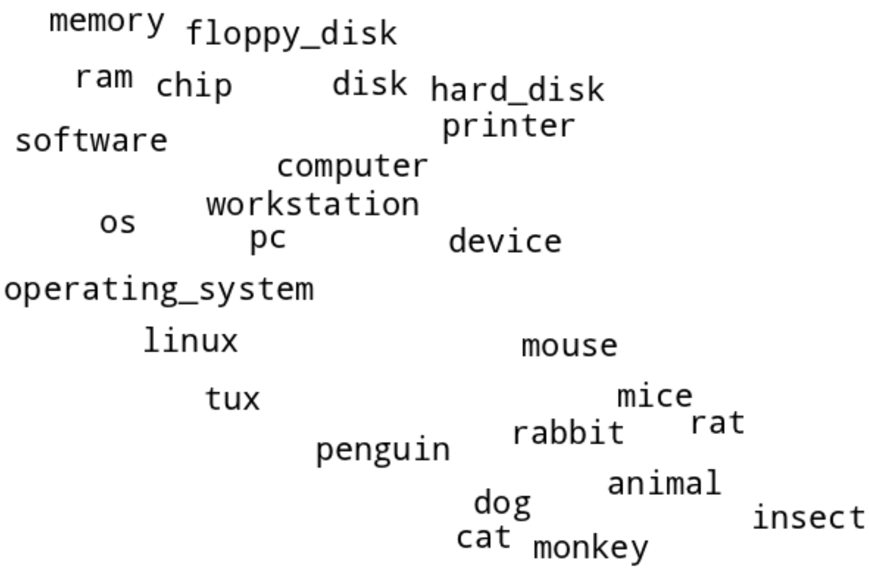


Distributional structure
Mathematical structures of language

https://goo.gl/b3sMtC

Distributional Semantic Models

- Analysis of word-usage statistics over huge corpora
- Geometric space of concepts
- Similar words are represented close in the space



> “ *A* ***WordSpace*** *is a snapshot of a specific corpus it does not take into account temporal information*

Random Indexing

Building the WordSpace

- Assign a random vector to each term in the corpus vocabulary
- Semantic vector for a term is the sum of the context vectors co-occurring with the term

Random Vector

…-1 0 1 0 0 0 0 0 0 0 0 0 -1 …

- Sparse
- high dimensional
- ternary {-1, 0, +1}
- small number of randomly distributed non-zero elements

https://github.com/semanticvectors/semanticvectors

Temporal Random Indexing

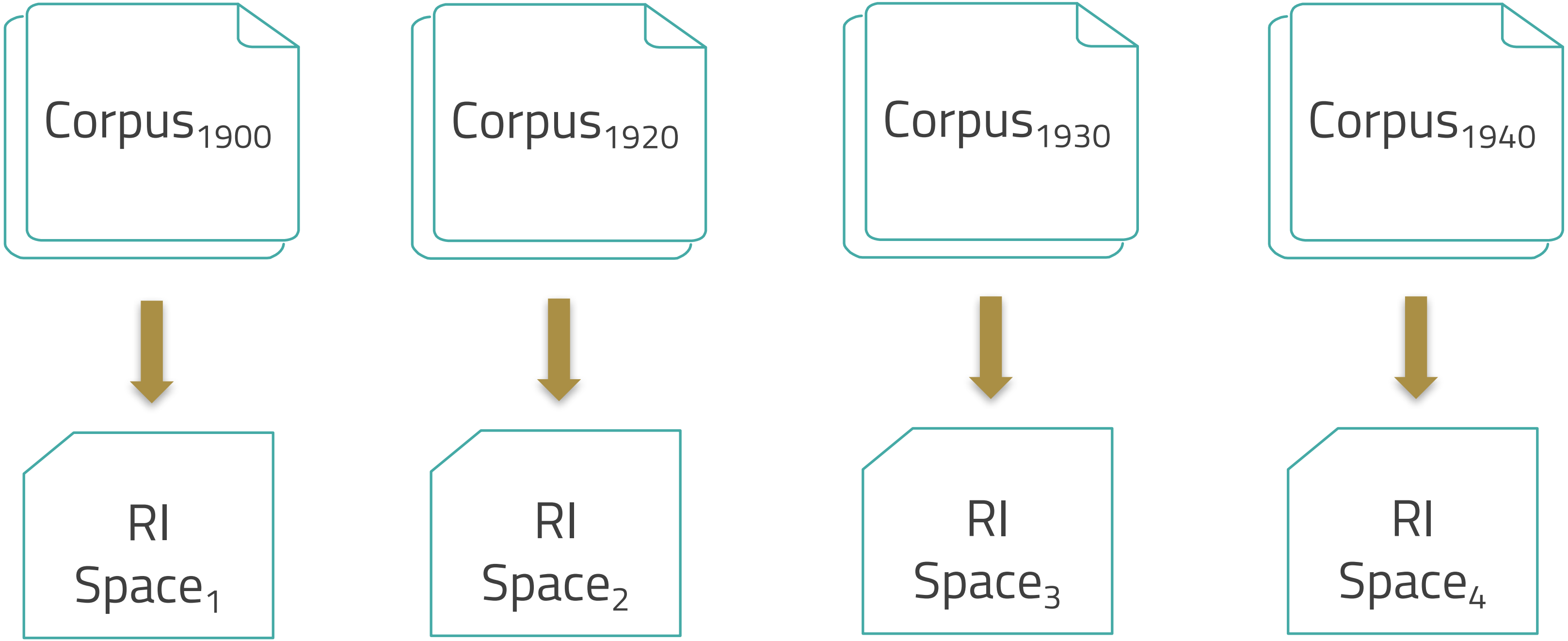
TRI

- Corpus with temporal information: split the corpus in several time periods
- Build a WordSpace for each time period
- Words in different WordSpaces are **comparable**!

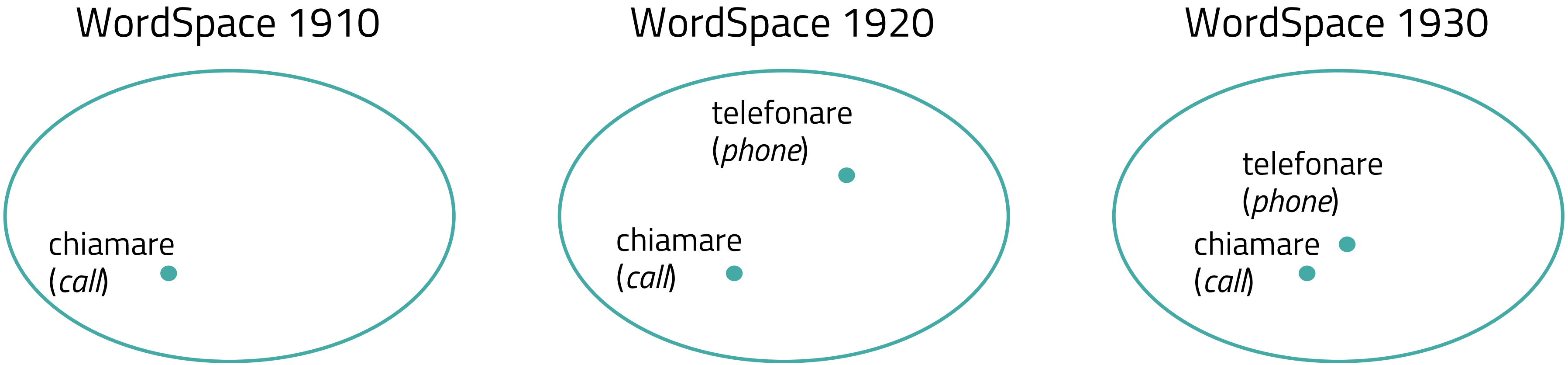
P. Basile, A. Caputo, G. Semeraro. *Temporal random indexing: A system for analysing word meaning over time*. IJCoL vol. 1
https://github.com/pippokill/tri

Temporal Random Indexing

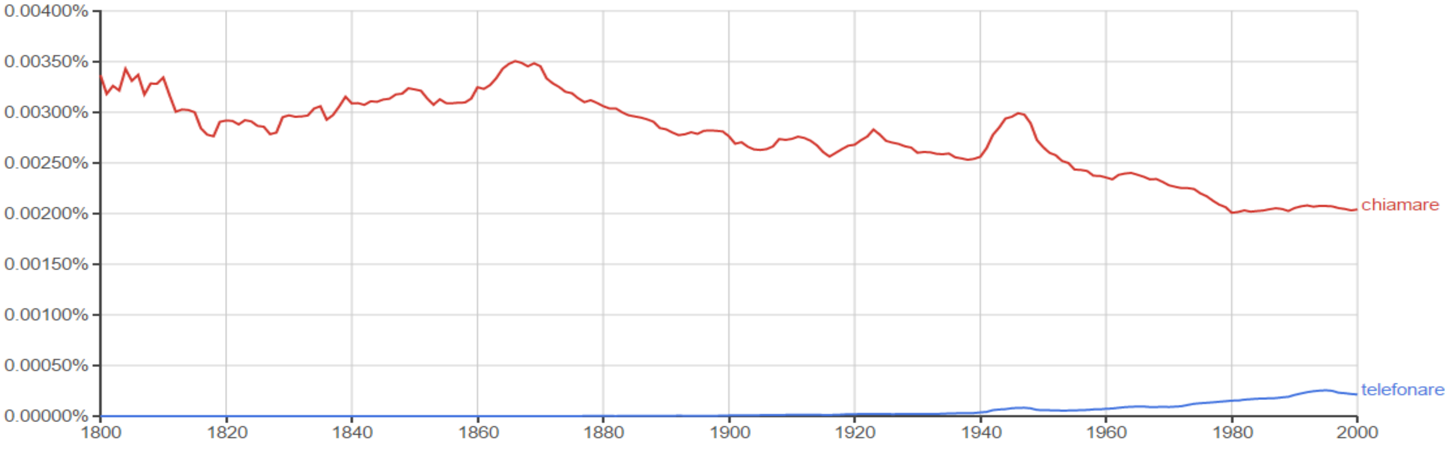
TRI



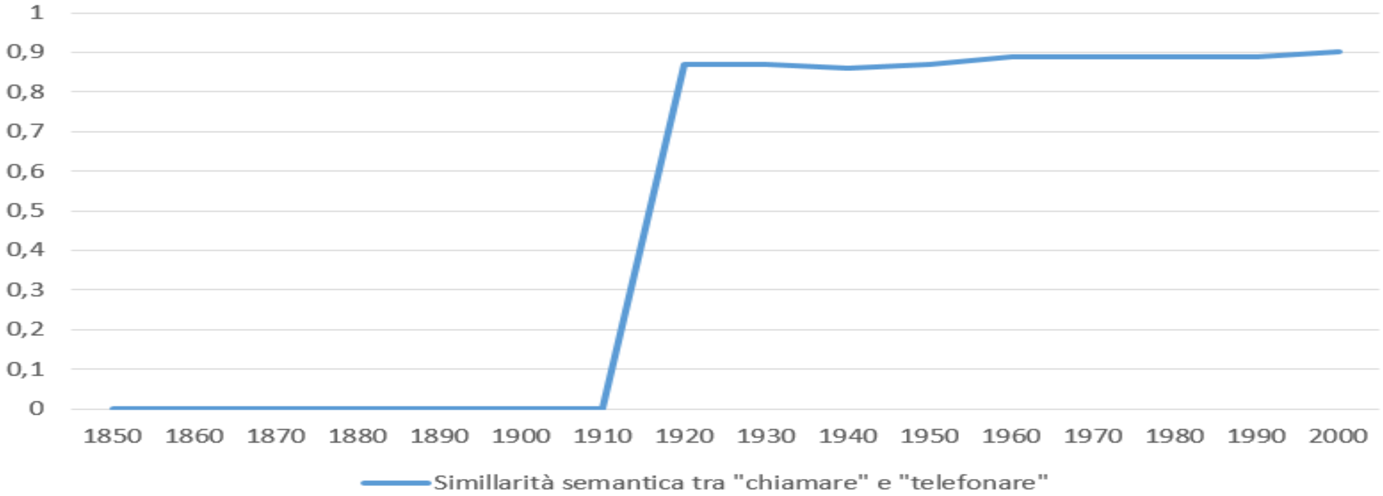
Similarity between words can change over time



Google Ngram



TRI



Methodology

TRI

Run TRI on Google Ngram: a WordSpace for each time period is built (10 years)

Time Series

Provide a time series for each word

Change Point Detection

Detect significant changes in the time series

Time Series

Several time series **Γ** at the time interval ***k***

log frequency	$\Gamma_k(t_i) = \frac{\#t_i^k}{C_k}$	Word frequency in each time period k
point-wise	$\Gamma_k(t_i) = cos_{sim}(sv_i^{k-1} \cdot sv_i^k)$	Cosine similarity between word vectors across two time periods
cumulative	$\Gamma_k(t_i) = cos_{sim}(\sum_{j=0}^{k-1} sv_i^j \cdot sv_i^k)$	Considers a cumulative vector of the previous *k-1* time periods

Change point detection

Mean shift model

- Mean shift of Γ pivoted at time period *j*

$$K(\Gamma) = \frac{1}{l-j}\sum_{k=j+1}^{l}\Gamma_k - \frac{1}{j}\sum_{k=1}^{j}\Gamma_k$$

- Search statistical significant mean shift
- Bootstrapping approach under the null hypothesis that there is no change in the meaning

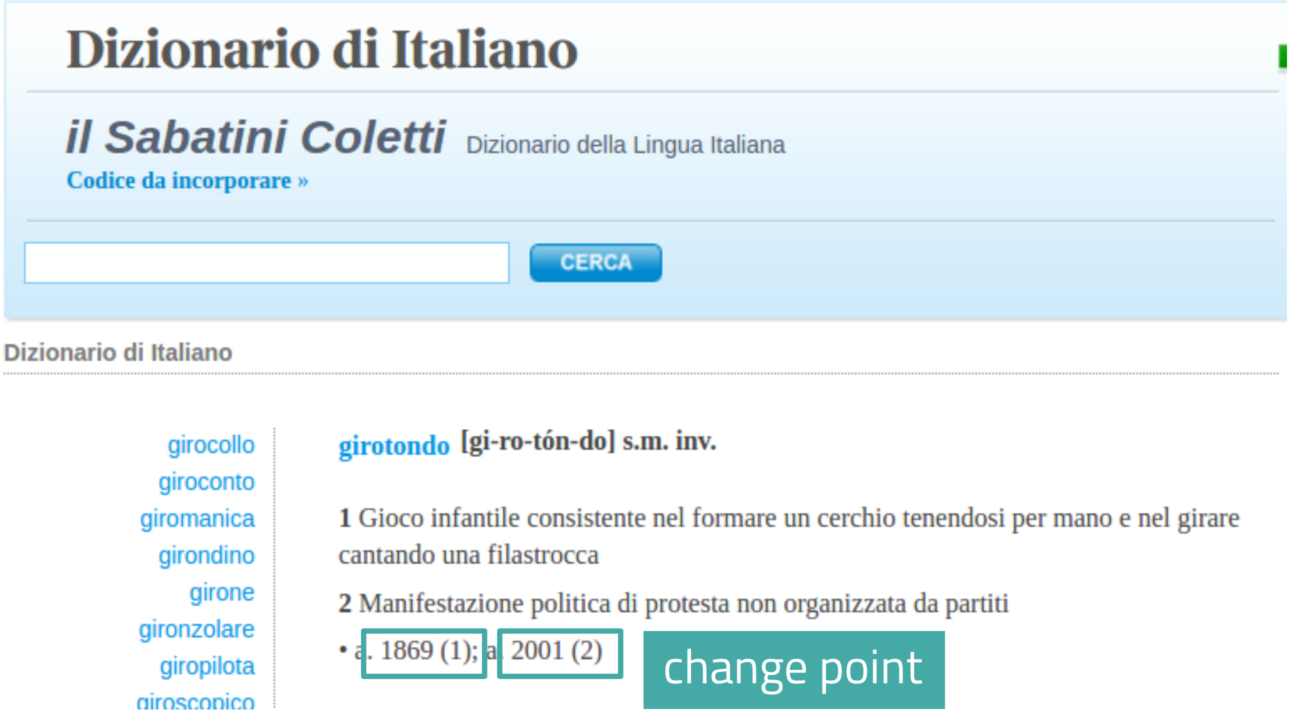
V. Kulkarni, et al. Statistically significant detection of linguistic change. WWW 2015.

Evaluation

- Build TRI by relying on the **Italian Google Ngram** corpus
- Build a standard benchmarking for meaning shift detection for the Italian language
  - "Dizionario Sabatino Coletti"
  - "Dizionario Etimologico Zanichelli"
- Evaluate the performance of TRI
  - compare the system output with **manual annotations** provided by **experts**

P. Basile, A. Caputo, G. Semeraro. *Diachronic Analysis of the Italian Language exploiting Google Ngram*. CLIC-it 2016

Build a gold standard for the evaluation



Dizionario di Italiano

il Sabatini Coletti Dizionario della Lingua Italiana

Codice da incorporare »

Dizionario di Italiano

girocollo
giroconto
giromanica
girondino
girone
gironzolare
giropilota
giroscopico

girotondo **[gi-ro-tón-do] s.m. inv.**

1 Gioco infantile consistente nel formare un cerchio tenendosi per mano e nel girare cantando una filastrocca

2 Manifestazione politica di protesta non organizzata da partiti

• a. 1869 (1); a. 2001 (2)

change point

http://dizionari.corriere.it/dizionario_italiano/

Evaluation

Results

Accuracy: the year predicted by the system must be equal or greater than one of the years reported in the gold standard

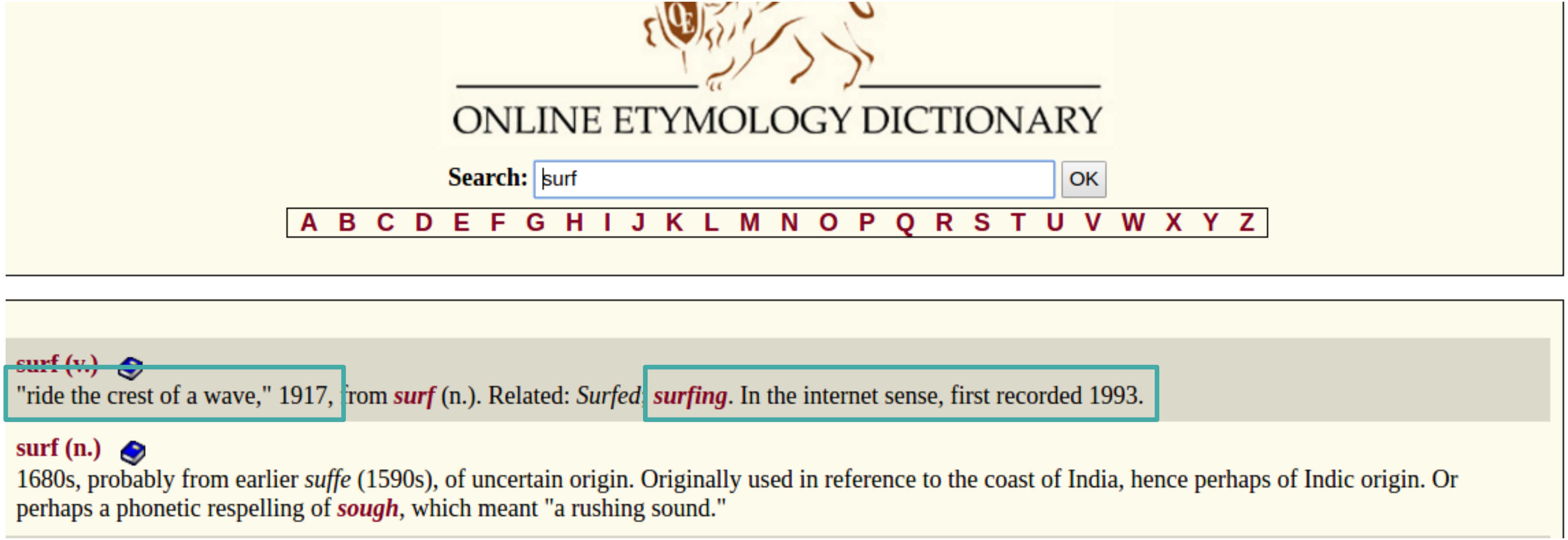
TRR1 and TRR2 are variants of TRI based on *Reflective Random Indexing*

Method	Accuracy
$\mathbf{TRI}_{\mathbf{point}}$	**0.3086**
$\mathrm{TRI}_{\mathrm{cum}}$	0.2963
$\mathrm{TRR1}_{\mathrm{point}}$	0.2716
log freq	*0.2346*
$\mathrm{TRR2}_{\mathrm{point}}$	0.1728
$\mathrm{TRR1}_{\mathrm{cum}}$	0.1605
$\mathrm{TRR2}_{\mathrm{cum}}$	0.1235

On going work...

English Google Ngram

- Build a gold standard for the English language



ONLINE ETYMOLOGY DICTIONARY

Search: surf

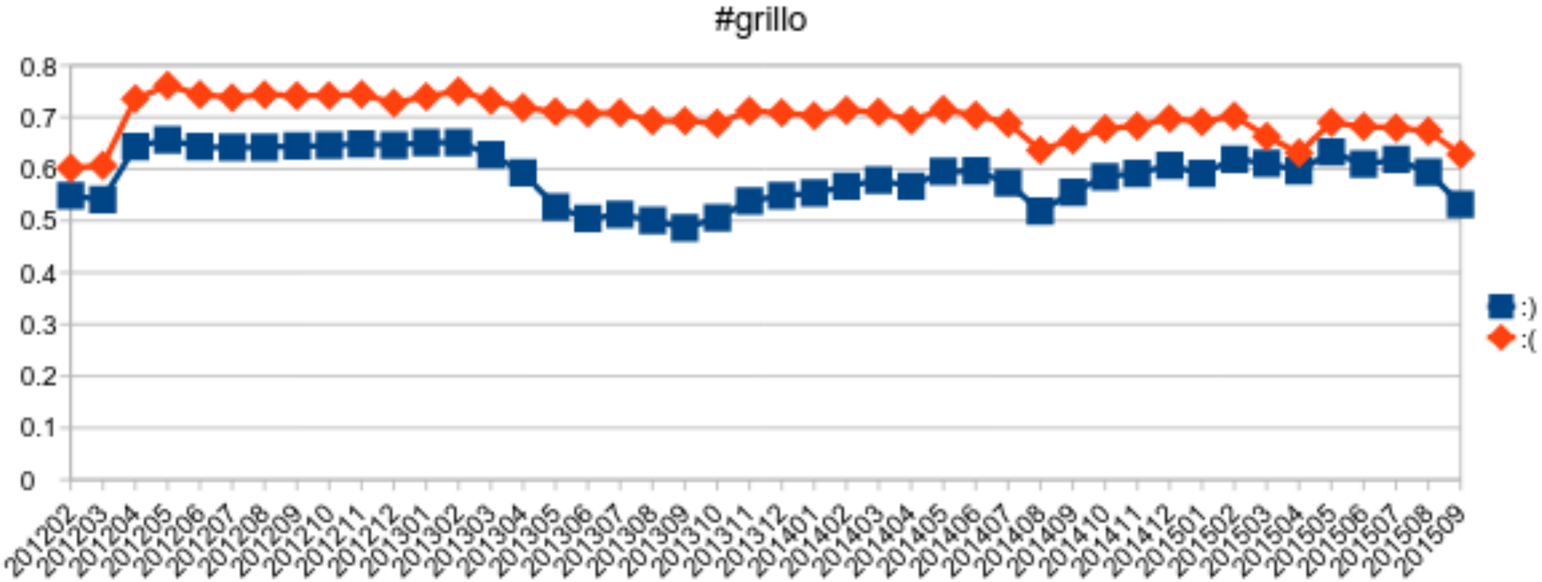
surf (v.)
"ride the crest of a wave," 1917, from ***surf*** (n.). Related: *Surfed*; ***surfing***. In the internet sense, first recorded 1993.

surf (n.)
1680s, probably from earlier *suffe* (1590s), of uncertain origin. Originally used in reference to the coast of India, hence perhaps of Indic origin. Or perhaps a phonetic respelling of ***sough***, which meant "a rushing sound."

http://www.etymonline.com/

On going work...
social media

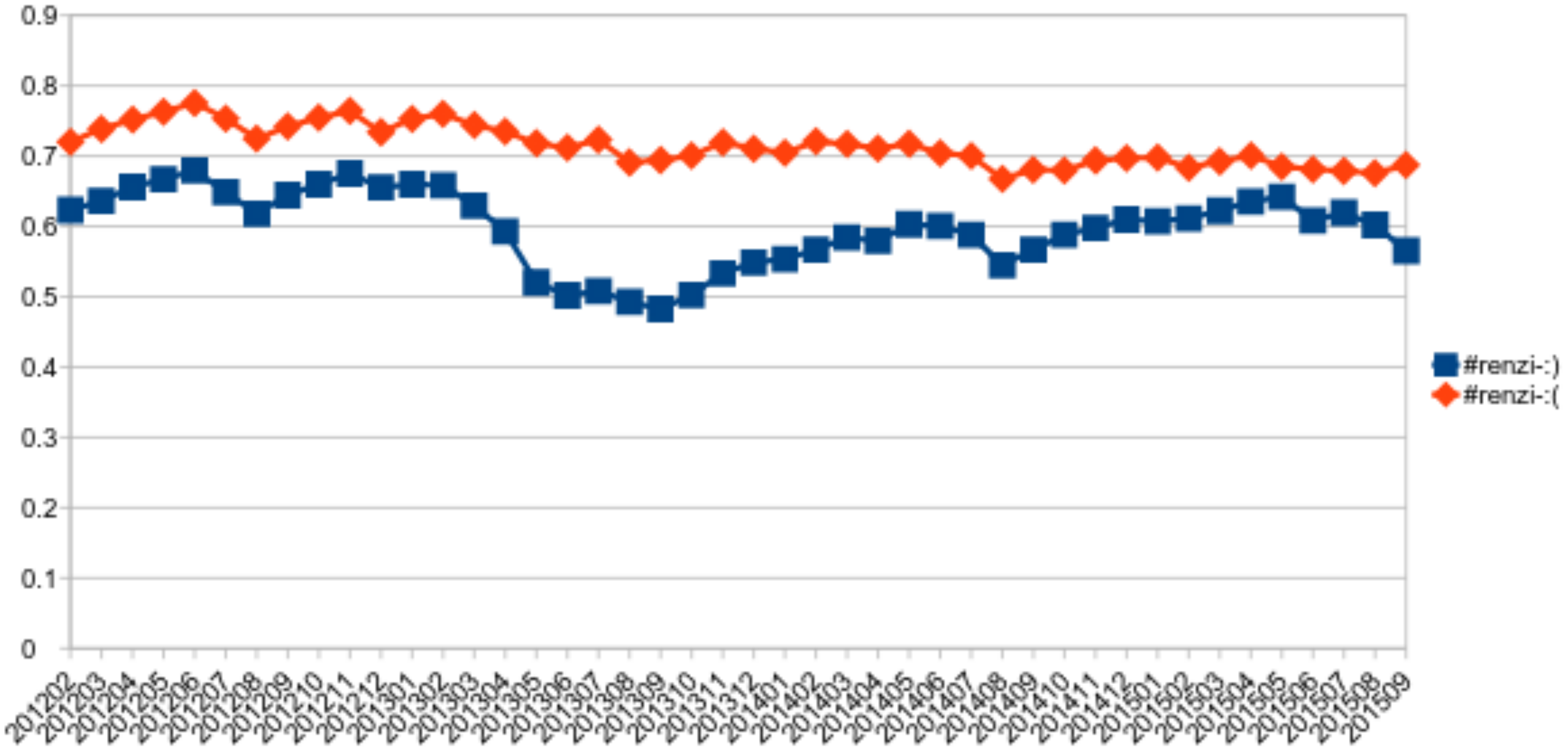
- Build TRI on **Twitter** (TWITA collection)
- About **500M tweets** (feb. 2012 – sep. 2015)



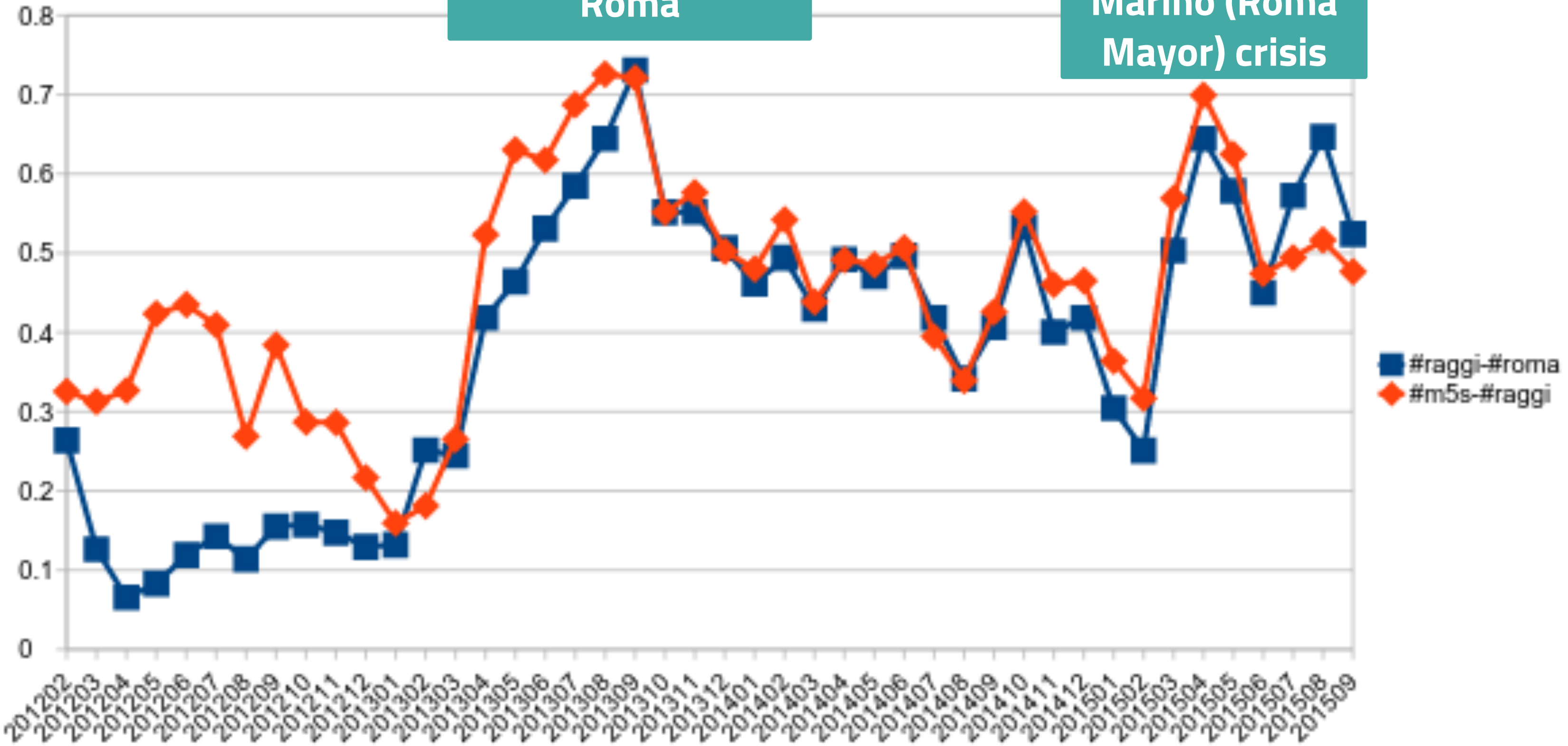
http://valeriobasile.github.io/twita/about.html

On going work...

social media



On going work...
social media



Local Election Roma

Marino (Roma Mayor) crisis

Workshop on

Temporal Dynamics in Digital Libraries @ TPDL2017

https://tddl2017.github.io/

Submission deadline: June 2, 2017

Thanks!

You can find me at @headlighty &
annalina.caputo@adaptcentre.ie &
annalina.github.io

Credits

- Thanks to Pierpaolo Basile for the material of this presentation
- The Google Ngram graphs are taken from J.-B. Michel et al., Quantitative Analysis of Culture Using Millions of Digitized Books. Science, 2011
- Presentation template by [SlidesCarnival]
- Photographs by [Unsplash]
- The source for every picture has been indicated below each of them. All copyrights belong to their respective owners.