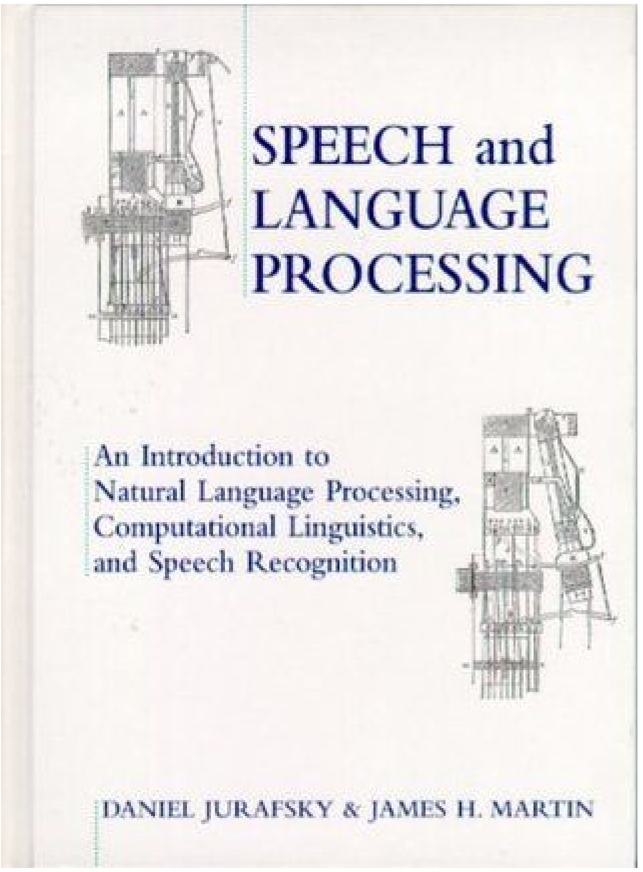
Introducción a NLP

Diego Fernández Slezak

Ciencia de Datos - 2do. cuat. 2017

La biblia...

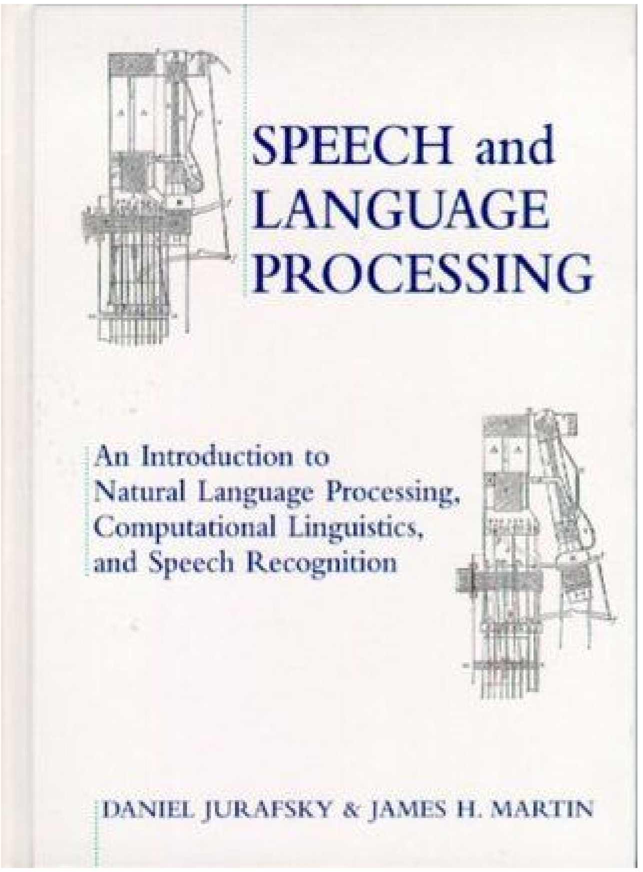
https://web.stanford.edu/~jurafsky/slp3/



- Introduction
- Regular Expressions, Text Normalization, and Edit Distance
- Finite State Transducers
- Language Modeling with N-Grams
- Spelling Correction and the Noisy Channel
- Naive Bayes Classification and Sentiment
- Logistic Regression
- Neural Nets and Neural Language Models
- Hidden Markov Models
- Neural Sequence Modeling: RNNs and LSTMs
- Part-of-Speech Tagging
- Formal Grammars of English
- Syntactic Parsing
- Statistical Parsing
- Dependency Parsing
- Vector Semantics
- Semantics with Dense Vectors
- Computing with Word Senses: WSD and WordNet
- Lexicons for Sentiment and Affect Extraction
- The Representation of Sentence Meaning
- Computational Semantics
- Information Extraction
- Semantic Role Labeling and Argument Structure
- Coreference Resolution and Entity Linking
- Discourse Coherence
- Seq2seq Models and Machine Translation
- Summarization
- Question Answering
- Dialog Systems and Chatbots
- Advanced Dialog Systems
- Speech Recognition
- Speech Synthesis

La biblia...

https://web.stanford.edu/~jurafsky/slp3/

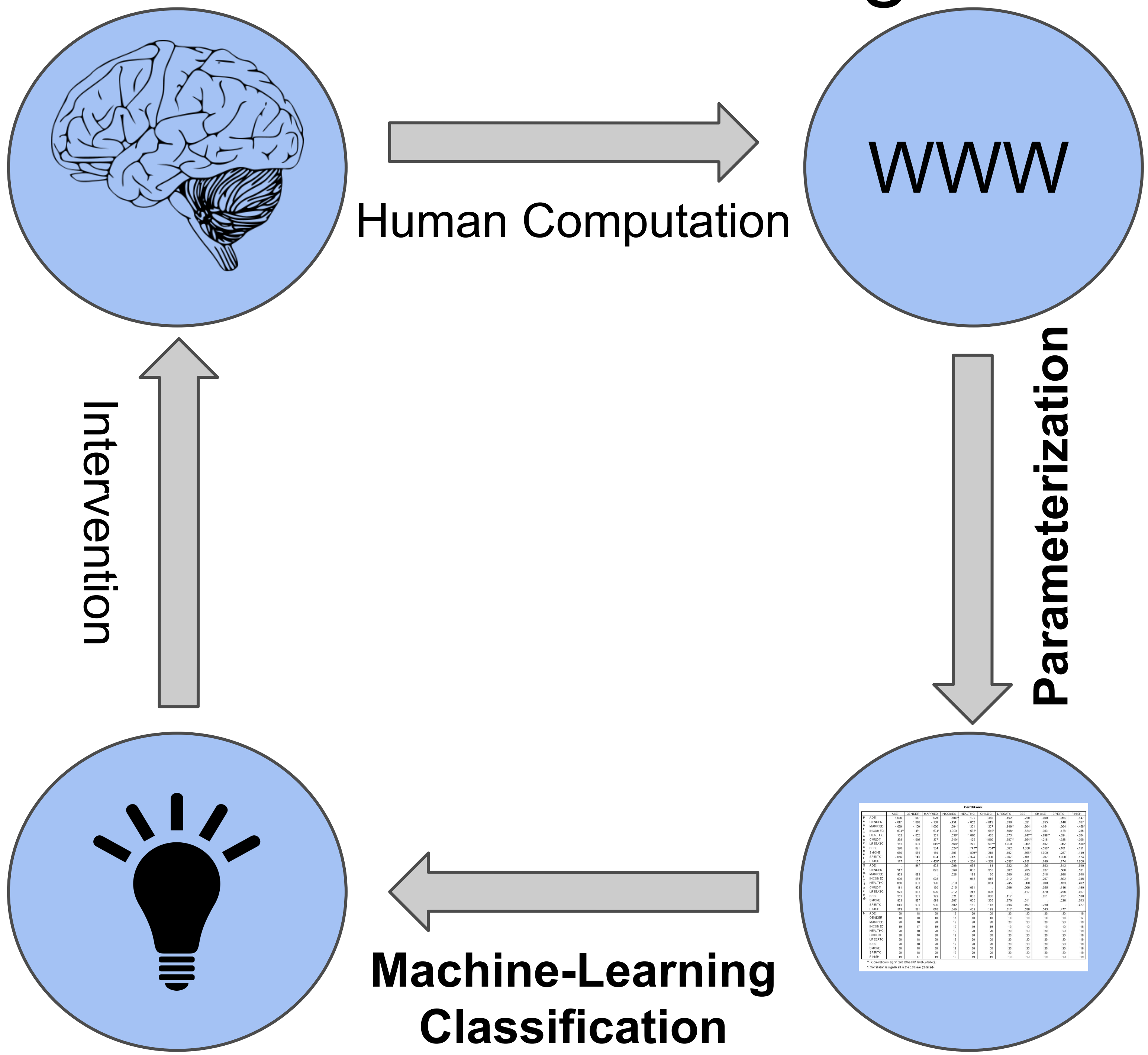


Mi motivación: ¡aplicaciones!

- ¿De qué están hablando las sociedades?
  - Text mining en millones de libros, blogs, páginas web, etc.

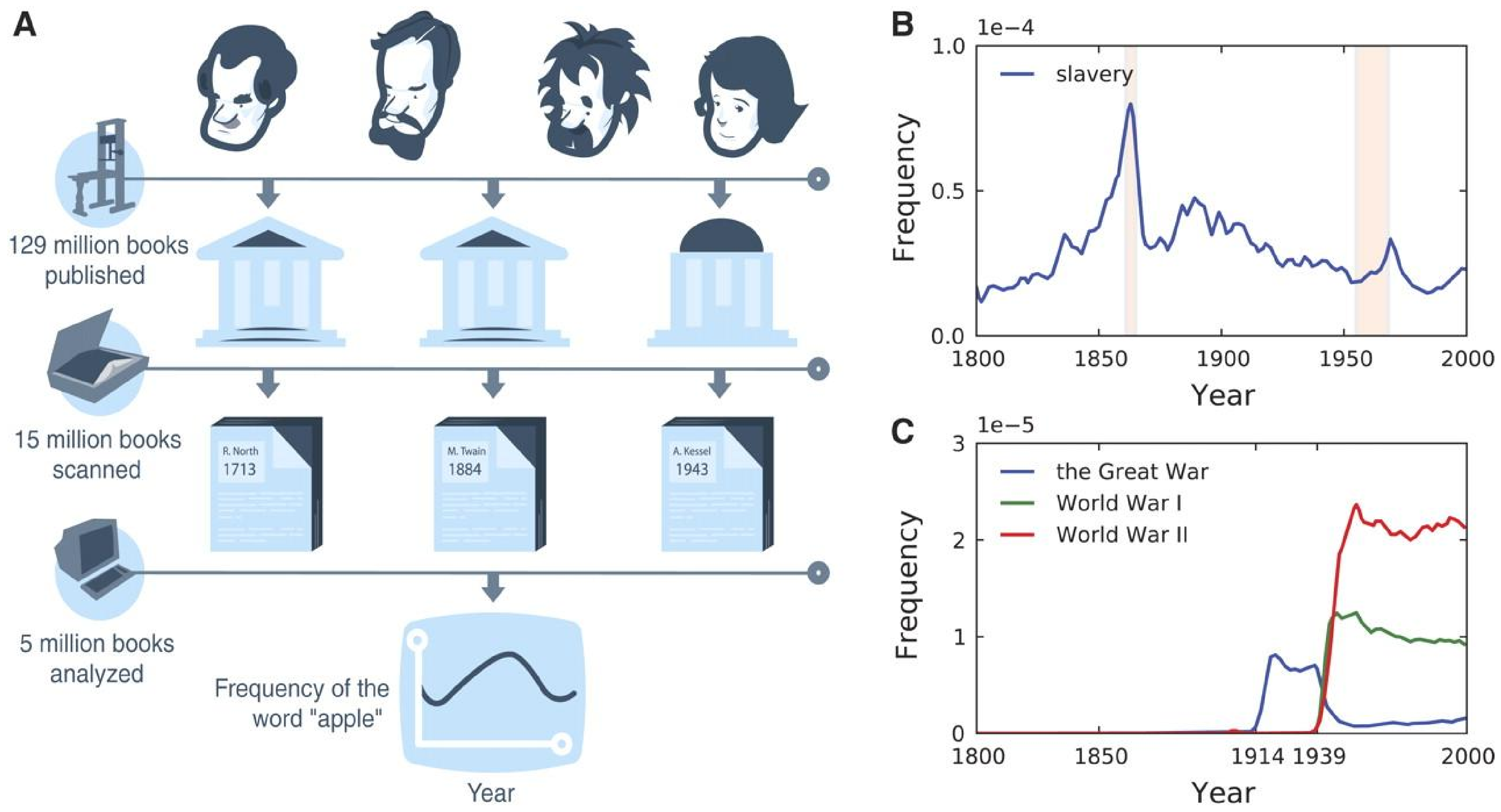
- Caracterización de estados mentales
  - Psiquiatría Computacional

Characterization of human behavior with the advent of Big Data



What people and societies are talking about...

- Michel, J.B. et. Al, *Quantitative analysis of culture using millions of digitized books,* **Science**, 2011.



Looking for abstract concepts

- Word Association
  - K. Church, P. Hanks, *Word association norms, mutual information and lexicography*, **J. of Computational Linguistics**, Volume 16 Issue 1, March 1990, pp 22-29, MIT Press.
- Latent Semantic Analysis
  - Deerwester, S. et. al, *Indexing by latent semantic analysis*, **J. of the American society for information science**, vol 416, pp 391-407, 1990.
  - Hofmann, T.*Probabilistic latent semantic indexing*, **Proc. of the 22nd annual international ACM SIGIR**, ACM, 1999.
- Latent Dirichlet Allocation
  - Blei, D., Ng, A. and Jordan, M., *Latent dirichlet allocation*, **Journal of machine Learning research,** 3: 993-1022, 2003.
- Deep Learning in NLP
  - Socher, R, Lin, C., Manning, C., Ng, A., *Parsing natural scenes and natural language with recursive neural networks*, **Proc of the 28th International Conference on Machine Learning (ICML-11)**. 2011.

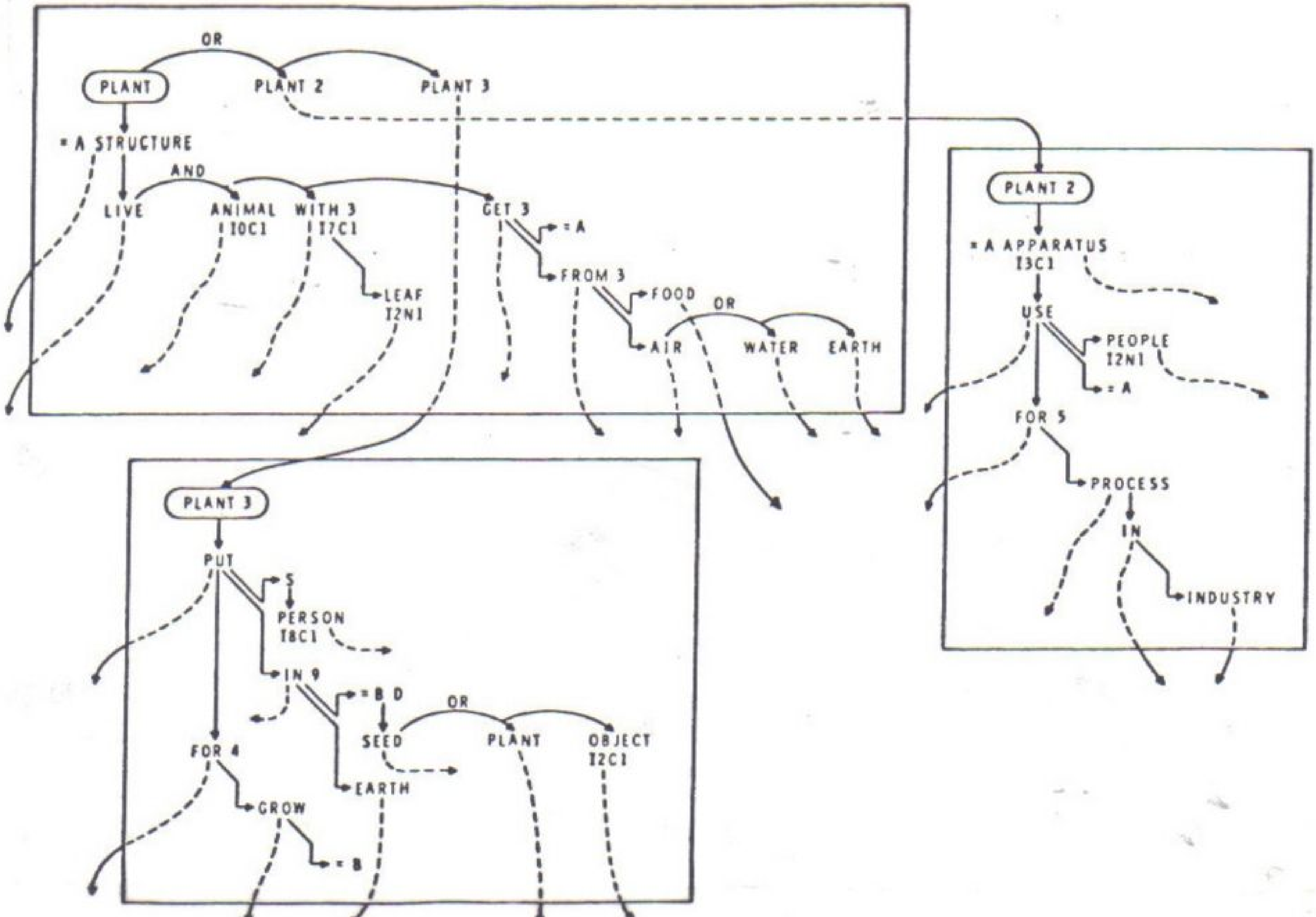
Word association

"Generally speaking, subjects respond quicker than normal to the word nurse if it follows a highly associated word such as doctor."

Empirical estimates of word association

Palermo, D. and Jenkins, J. 1964 "Word Association Norms." University of Minnesota Press, Minneapolis, MN.

Redes semánticas de Quillian



Page 8 / 124

La evidencia I: Collins & Quillian

- Buscaban encontrar evidencia de la estructura jerárquica de la memoria semántica y del modelo de propagación de la activación (Spreading of activation).
- Tarea: con una tecla comunicar si la oración es verdadera o falsa.
- Oraciones de propiedades (del individuo o de la clase) o de pertenencia de clase.
- P0: Un roble tiene bellotas. P1: … tiene hojas.
- S0: Un perro es un perro. S1: …. es un mamífero.
- Usaban oraciones falsas además.

Priming y estructura jerárquica.

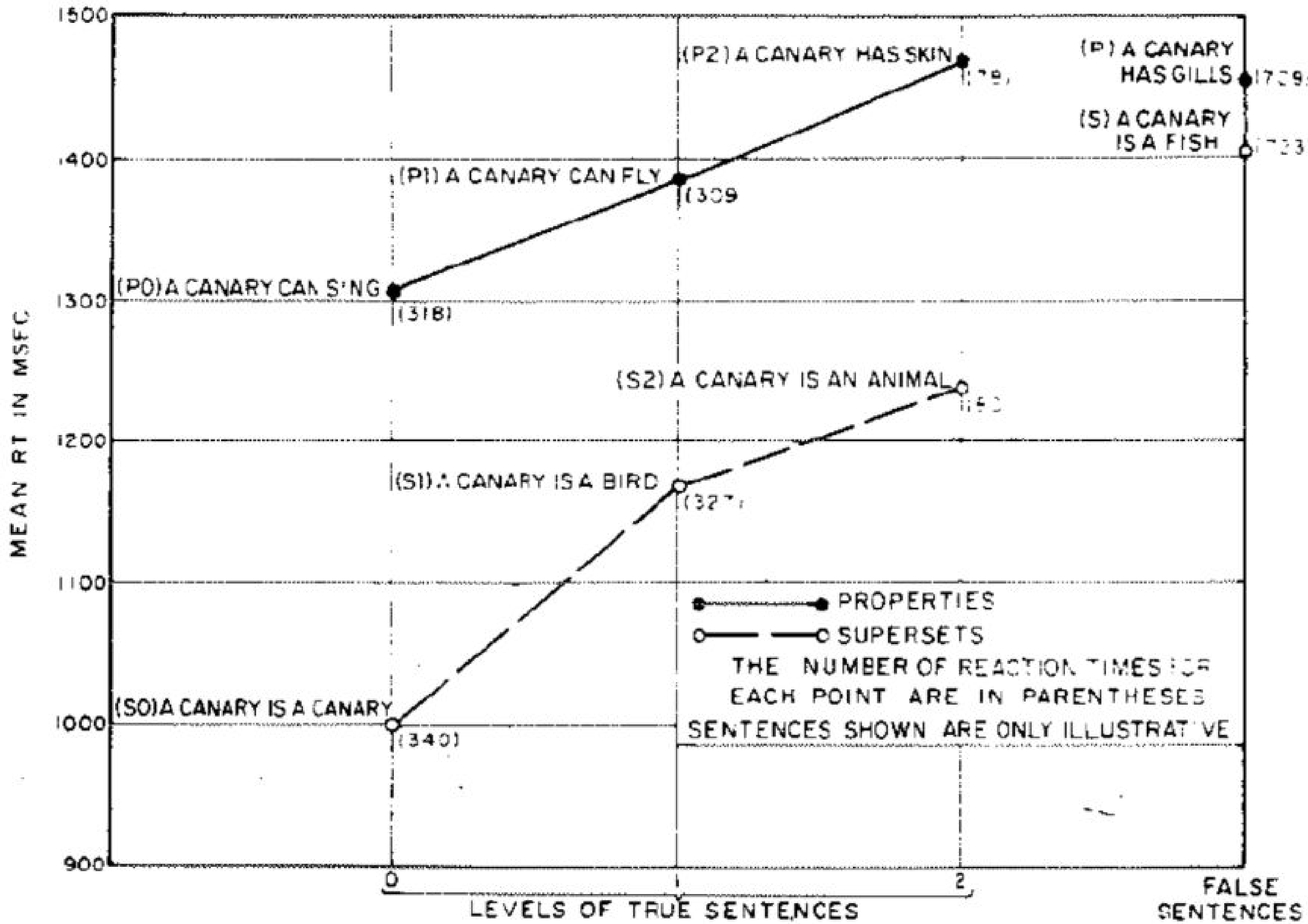
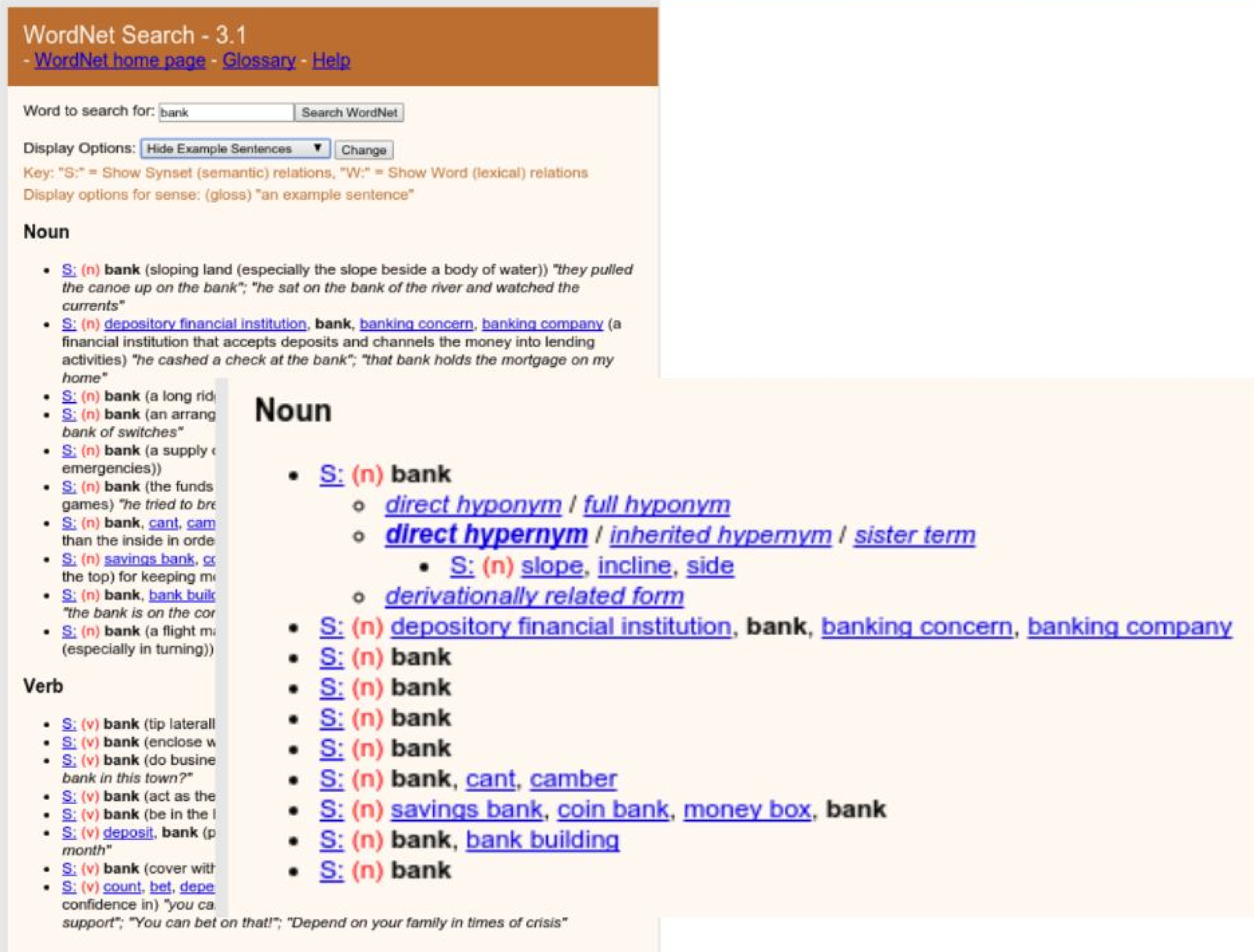


FIG. 2. Average reaction times for different types of sentences in three experiments.

Palabras y conceptos

- Visión, 1 palabra ↔ 1 concepto
  - Perro ↔ PERRO
  - Banco ↔ FINANCIERO? DE PLAZA? DE PECES?
  - .... ↔ Usar un clip para destapar la bombilla del mate
  - La ↔ Concepto u operador sobre concepto?
- Varias palabras ↔ un concepto y conceptos sin palabras.
- Palabras con diferentes sentidos (disco).
- Partes de conceptos (pestaña, etc).

Un ejemplo: wordnet



WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: bank Search WordNet

Display Options: Hide Example Sentences Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **bank** (sloping land (especially the slope beside a body of water)) *"they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"*
- S: (n) depository financial institution, **bank**, banking concern, banking company (a financial institution that accepts deposits and channels the money into lending activities) *"he cashed a check at the bank"; "that bank holds the mortgage on my home"*
- S: (n) **bank** (a long rid
- S: (n) **bank** (an arrang *bank of switches"*
- S: (n) **bank** (a supply emergencies))
- S: (n) **bank** (the funds games) *"he tried to bre*
- S: (n) **bank**, cant, cam than the inside in orde
- S: (n) savings bank, co the top) for keeping m
- S: (n) **bank**, bank buil *"the bank is on the cor*
- S: (n) **bank** (a flight m (especially in turning))

Verb

- S: (v) **bank** (tip laterall
- S: (v) **bank** (enclose w
- S: (v) **bank** (do busine *bank in this town?"*
- S: (v) **bank** (act as the
- S: (v) **bank** (be in the
- S: (v) deposit, **bank** (p *month"*
- S: (v) **bank** (cover wit
- S: (v) count, bet, depe confidence in) *"you ca support"; "You can bet on that!"; "Depend on your family in times of crisis"*

Noun

- S: (n) **bank**
  - *direct hyponym* / *full hyponym*
  - ***direct hypernym*** / *inherited hypernym* / *sister term*
    - S: (n) slope, incline, side
  - *derivationally related form*
- S: (n) depository financial institution, **bank**, banking concern, banking company
- S: (n) **bank**
- S: (n) **bank**
- S: (n) **bank**
- S: (n) **bank**
- S: (n) **bank**, cant, camber
- S: (n) savings bank, coin bank, money box, **bank**
- S: (n) **bank**, bank building
- S: (n) **bank**

An information theoretic measure

Association ratio, based on the information theoretic concept of **mutual information**.

$$I(x, y) \equiv \log_2 \frac{P(x, y)}{P(x)P(y)}$$

- If there is a **genuine association** between *x* and *y*, then *P(x,y)* will be much larger than *P(x)* . *P(y)*, and ***I(x,y) >> 0***.

- If there is **no interesting relationship** between *x* and *y*, then *P(x,y) = P(x) . P(y)*, and thus, ***I(x,y) ~ 0***.

- If x and y are in **complementary distribution**, then P(x,y) will be much less than P(x) P(y), forcing ***I(x,y) <<0***.

Probability estimation

P(x) and P(y) are estimated by counting the number of observations of *x* and *y* in a corpus, *f(x)* and *f(y)*, and normalizing by N, the size of the corpus.

P(x,y), are estimated by counting the number of times that *x* is followed by *y* in a window of *w* words, $f_w(x,y)$, and normalizing by N.

(...the window size, *w*, will be set to five words as a compromise)

Since the association ratio becomes unstable when the counts are very small, we will not discuss word pairs with $f(x,y) < 6$.

Measure properties

- I (x, y) = I(y, x). But, f(x, y) ~= f(y, x)
- Expected: f (x, y) <= f(x) and f(x, y) <= f(y)
  - "Library workers were prohibited from saving books from this heap of ruins,"
  - f(prohibited) = 1 and f(prohibited, from) = 2
  - This problem can be fixed by dividing f(x, y) by (w-1)

Corpus

- This paper: Associated Press
- Other corpus: TASA, Pagina12, La Nación, Google

Implementation

Python

- NLTK package for Natural Language Processing

  http://www.nltk.org/

Example

```
>>> import nltk.data
>>> text = '''
... Punkt knows that the periods in Mr. Smith and Johann S. Bach
... do not mark sentence boundaries.  And sometimes sentences
... can start with non-capitalized words.  i is a good variable
... name.
... '''
>>> sent_detector =
nltk.data.load('tokenizers/punkt/english.pickle')
>>> sents = sent_detector.tokenize(text.strip()))
```

Implementation

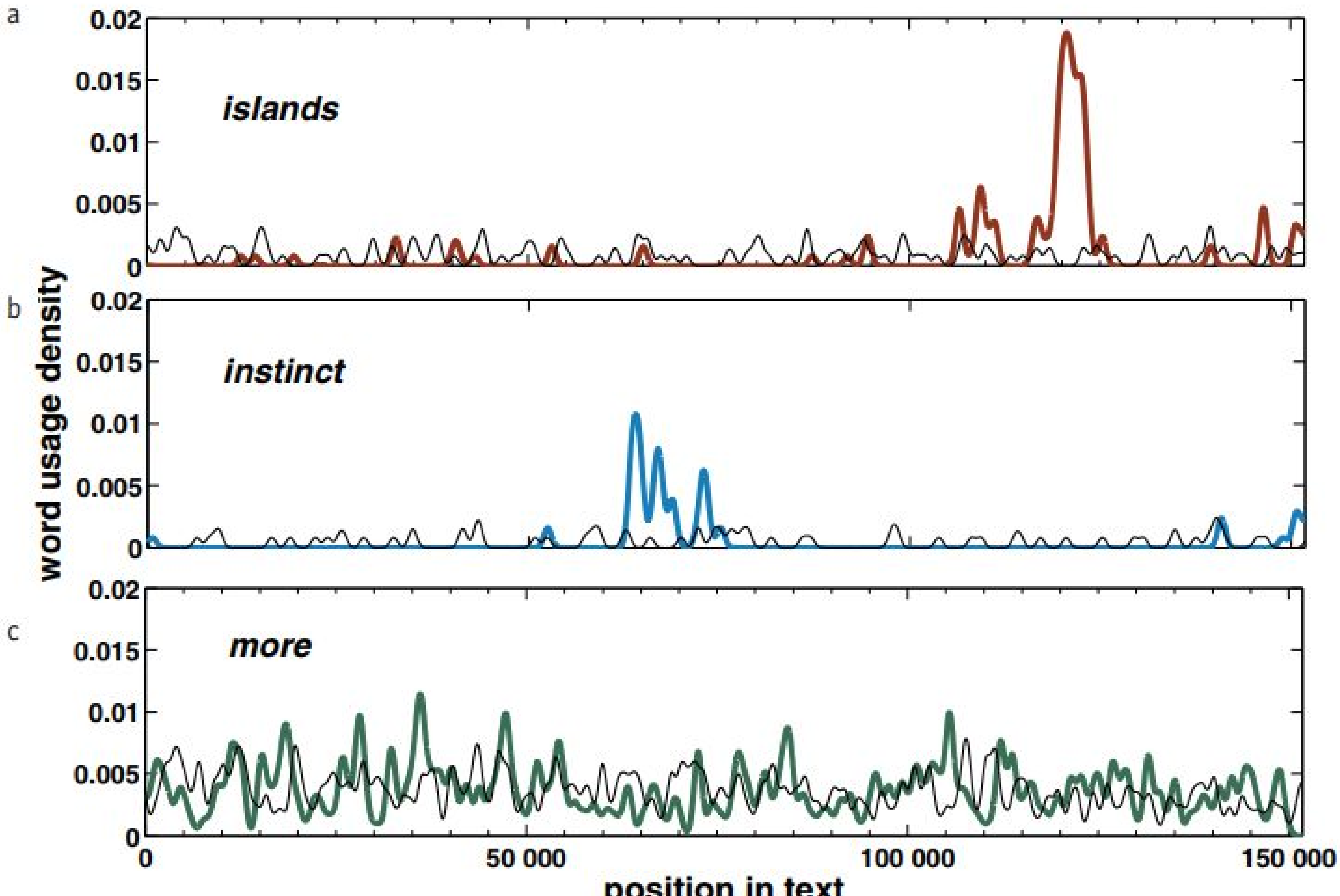
Word Association: Given the corpus *X* and w_i each different word in *X*

- Let *k* be the window size
- For each w_i calculate $f(w_i)$
- For each w_i and w_j calculate $f_k(w_i, w_j)$

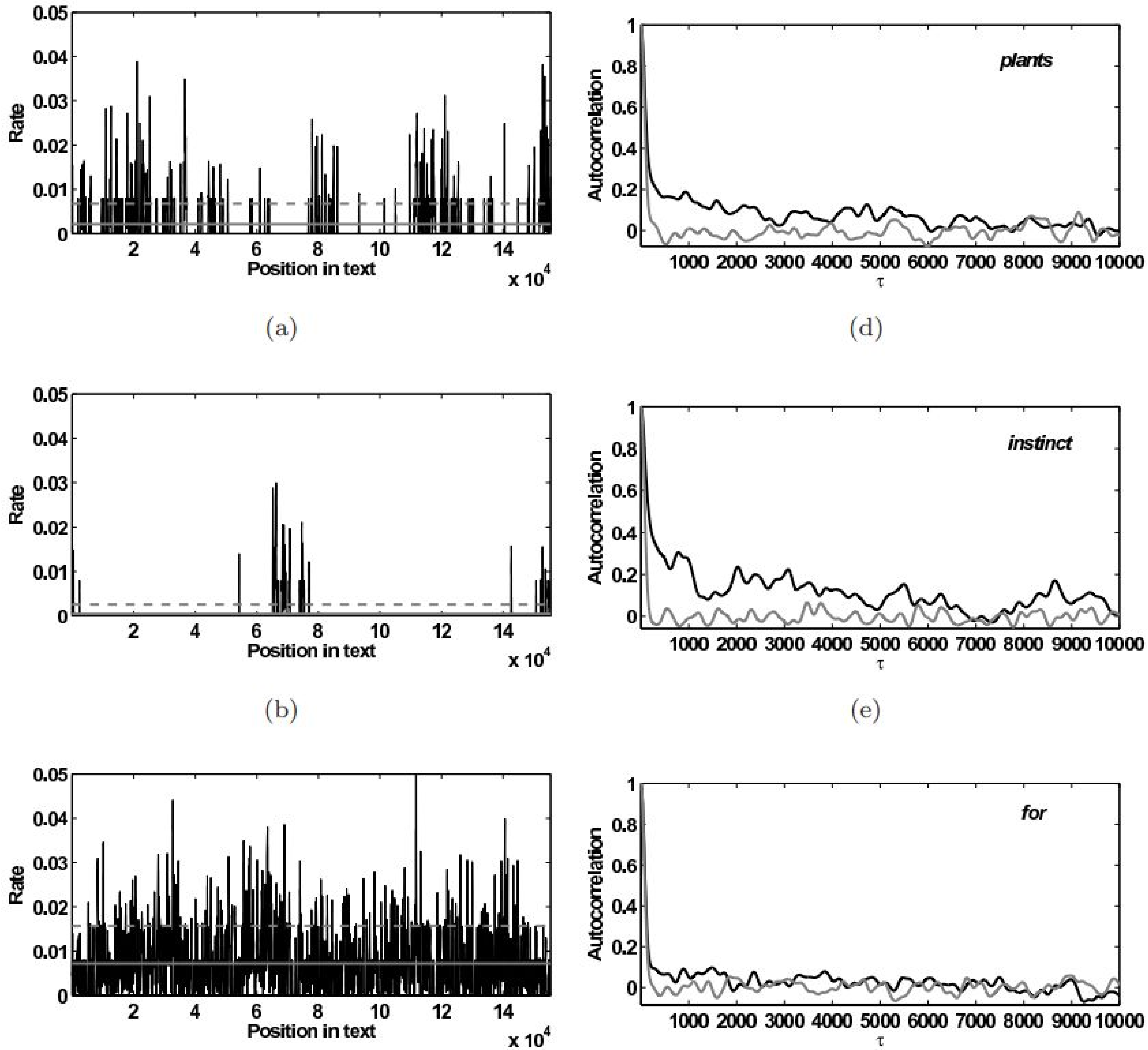
Pseudocode

- Let FM be the co-occurrence matrix
- For each position m,
  - For each position p between m+1 and m+k
    - $FM(X_m,X_p)$++

Información léxica



Escala de cada palabra



Pertenencia de palabras a secciones

$$H(J|w) = -\sum_{j=1}^{P} \frac{n_j}{n} \log_2 \frac{n_j}{n},$$

Suma sobre todas las palabras del texto

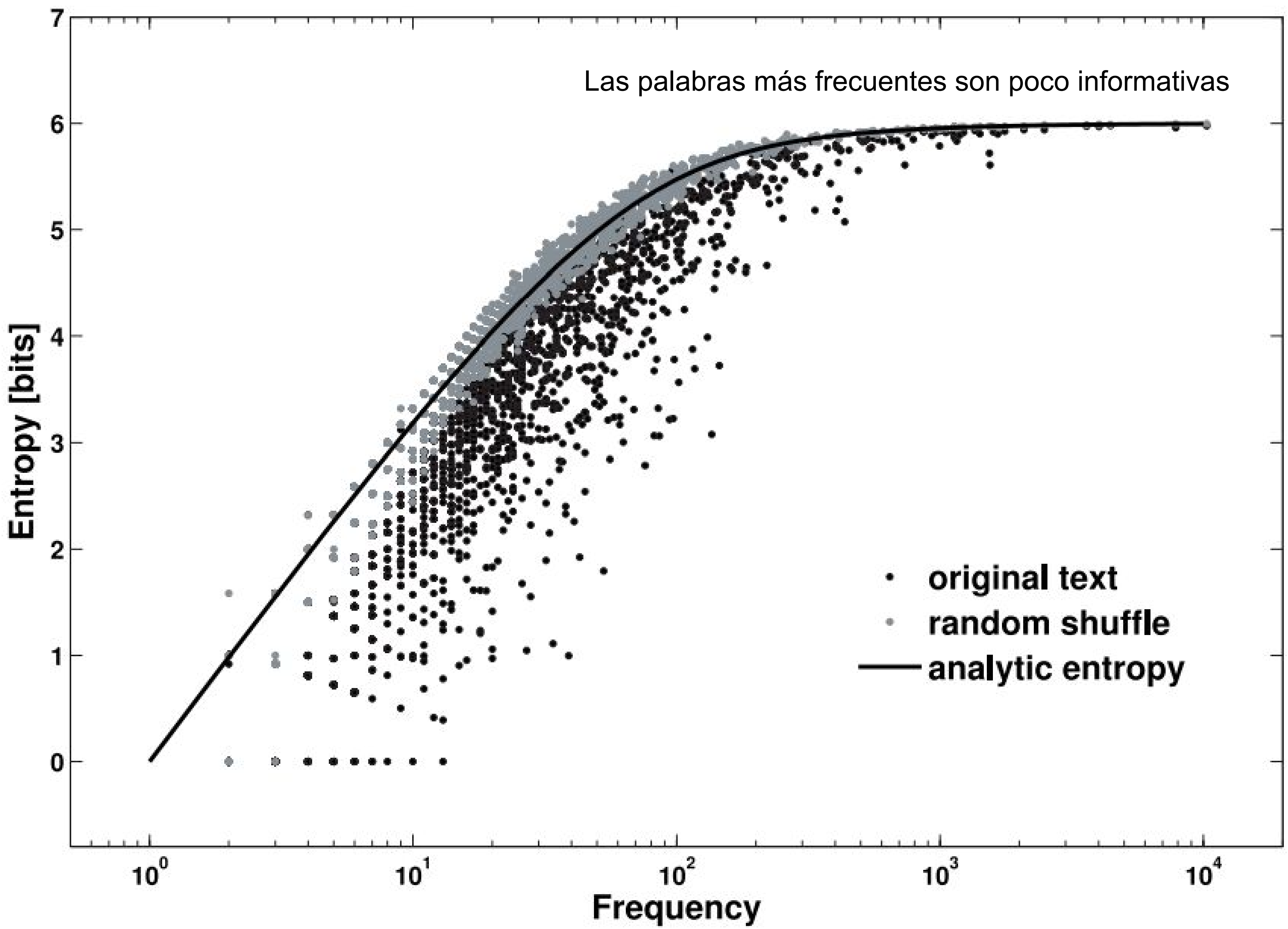
$$\Delta I(s) = \sum_{w=1}^{K} p(w)[\langle \hat{H}(J|w) \rangle - H(J|w)],$$

La entropia de la palabra w en las partes del texto. Es maxima si esta distribuida homogeneamente. Es minima si esta concentrada en una sola parte. Es una informacion mutua porque nos dice cuanto informa la palabra respecto de la parte. Es una funcion parametrica de la escala de la particion.

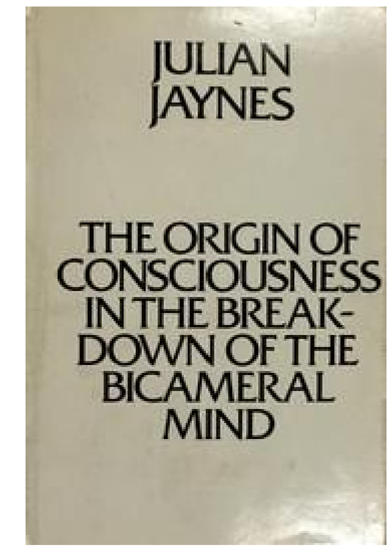
La frecuencia de la palabra w

La entropia permutada (el maximo que puede tomar)

Entropía de las palabras



Looking for abstract concepts. Introspection: A case study



J. Jaynes, *The origins of consciousness in the breakdown of the bicameral mind*, Mariner Books, 2000.

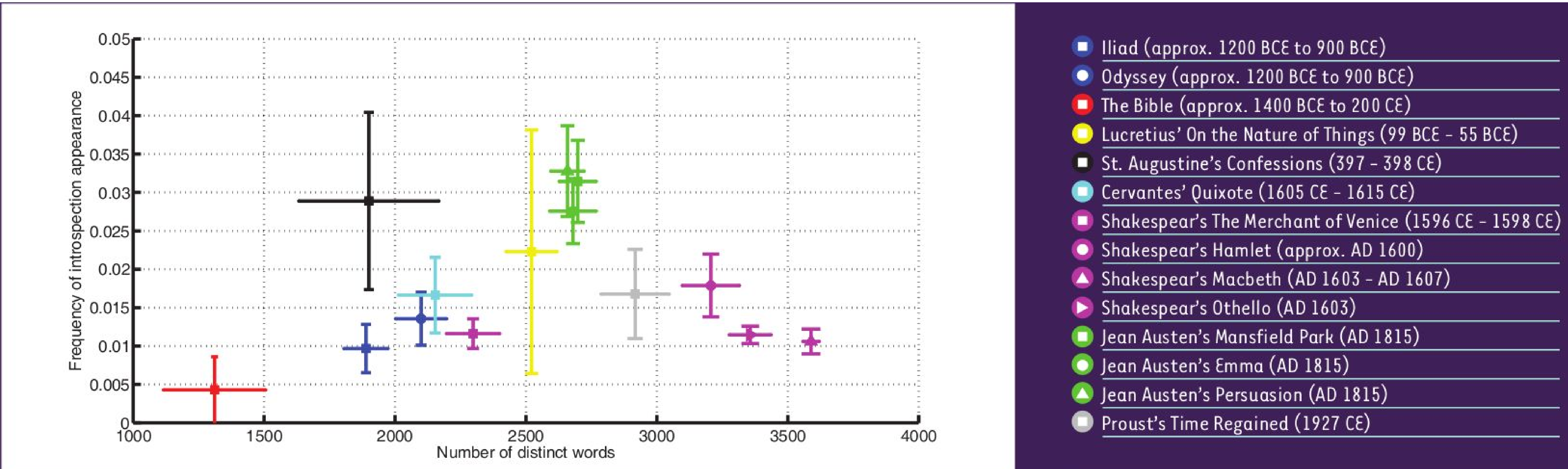
- **Psychological hypothesis**: human mind with cognitive functions divided in:
  - one part of the brain which appears to be "speaking", *acoustical hallucinations*
  - and a second part which listens and obeys
- Starting in about 10,000 BC and extending through about 1000 BC: agrarian empires and city-states, i.e. Egypt, the Levant, Mesopotamia, Greece, etc
- Introspection, Self-awareness, or Consciousness, was the cultural evolution
  - Development of language?
  - Cultural disruptions?

Diuk, Carlos G., D. Fernandez Slezak, Iván Raskovsky, Mariano Sigman, and Guillermo A. Cecchi. "A quantitative philology of introspection." *Frontiers in integrative neuroscience* 6 (2012).

Using regular expressions

Words associated with introspection:

- think+ thought myself mind+ feel+ felt



Frequency of words related to introspection versus the amount of different words. Each author is identified by a unique color. Error bars shows the standard deviation in both axis.

Selected text corpus representative of different ages in literature from the MIT classic text archive based on references in Jaynes' book.

Latent semantic analysis (LSA)

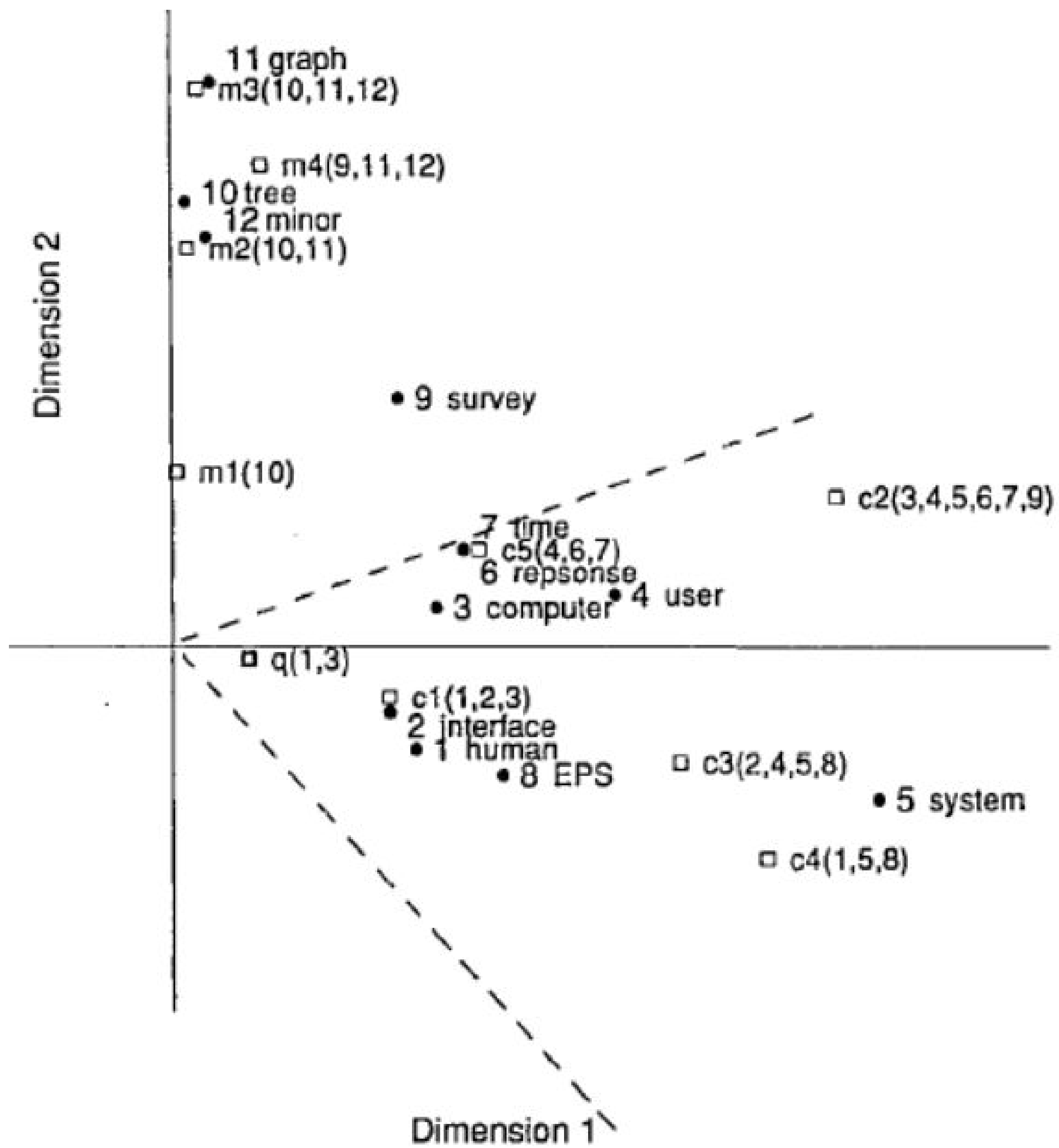
Similarity between **concepts** based on **documents** and **frequencies** of terms

Technical Memo Example

Titles

- c1: *Human* machine *interface* for Lab ABC *computer* applications
- c2: A *survey of user* opinion of *computer system response time*
- c3: The *EPS user interface* management *system*
- c4: *System* and *human system* engineering testing of *EPS*
- c5: Relation of *user*-perceived *response time* to error measurement
- m1: The generation of random, binary, unordered *trees*
- m2: The intersection *graph* of paths in *trees*
- m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
- m4: *Graph minors: A survey*

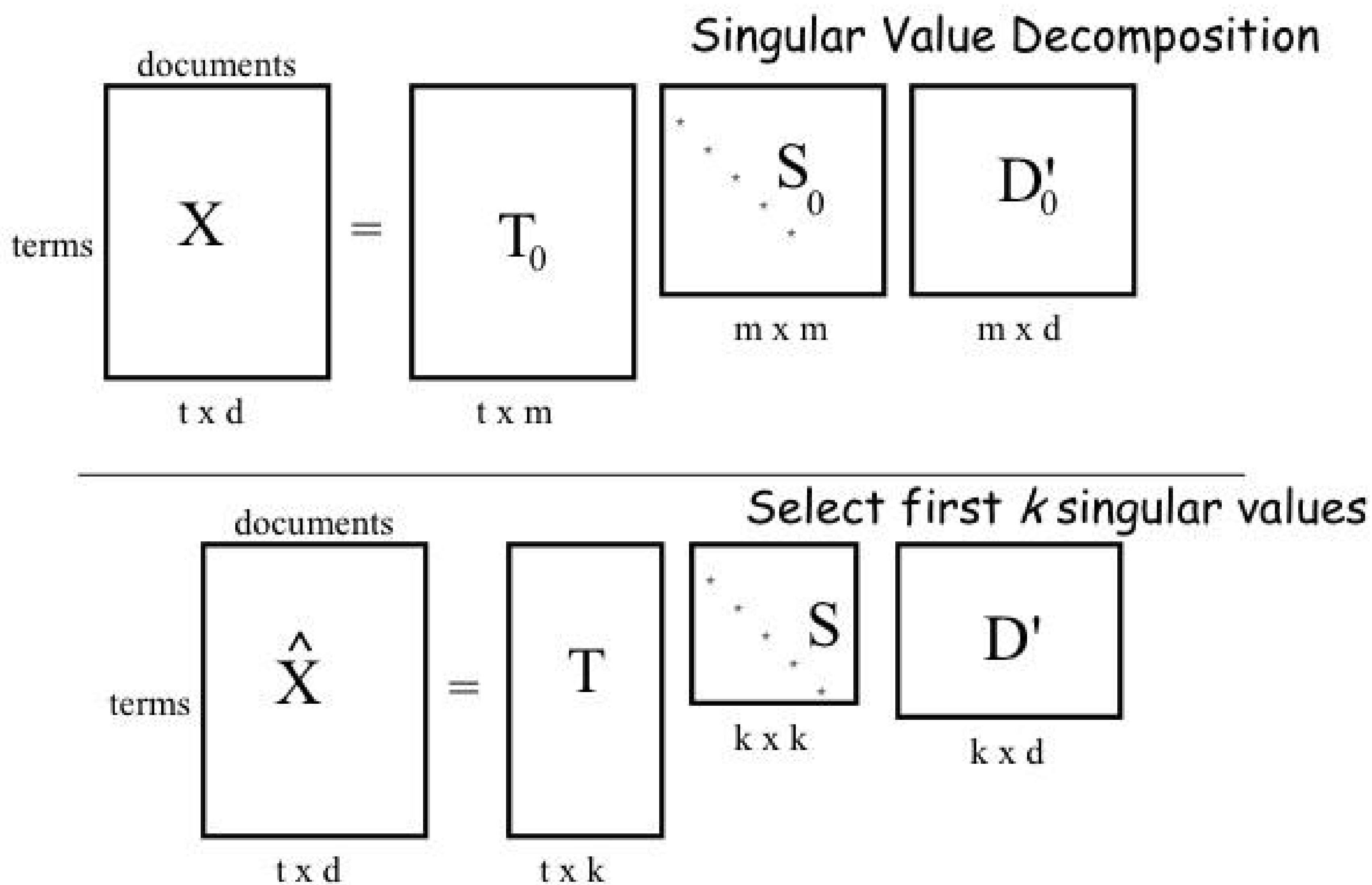
Terms	Documents								
	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1



Deerwester, S. et. al,, 1990.

LSA in one slide

Given a text database, let *X* be the occurrence matrix of words in each text (#docs × #terms).



LSA trained with TASA corpus

- TASA corpus: Touchstone Applied Science Associates, Inc.
  - texts, novels, newspaper articles, and other information used in elementary, secondary school and college.
  - TASA corpus used to develop The Educator's Word Frequency Guide.

- LSA @ Colorado: http://lsa.colorado.edu/

- We run LSA to TASA corpus with 300 components → we can measure distance between all words present in TASA.
  - Landauer, T.K., Dumais, S.T., *A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge*, **Psychological Review**, 104, 211-240, 1997.

Implementation

Python

- Gensim
  - http://radimrehurek.com/gensim/

```python
from gensim import corpora, models, similarities
dictionary = corpora.Dictionary.load('/tmp/deerwester.dict')
corpus = corpora.MmCorpus('/tmp/deerwester.mm')
lsi = models.LsiModel(corpus, id2word=dictionary, num_topics=2)
```

```python
doc = "Human computer interaction"
vec_bow = dictionary.doc2bow(doc.lower().split())
vec_lsi = lsi[vec_bow] # convert the query to LSI space
print(vec_lsi)
[(0, -0.461821), (1, 0.070028)]
```

Using LSA...

Selected word: *introspection*

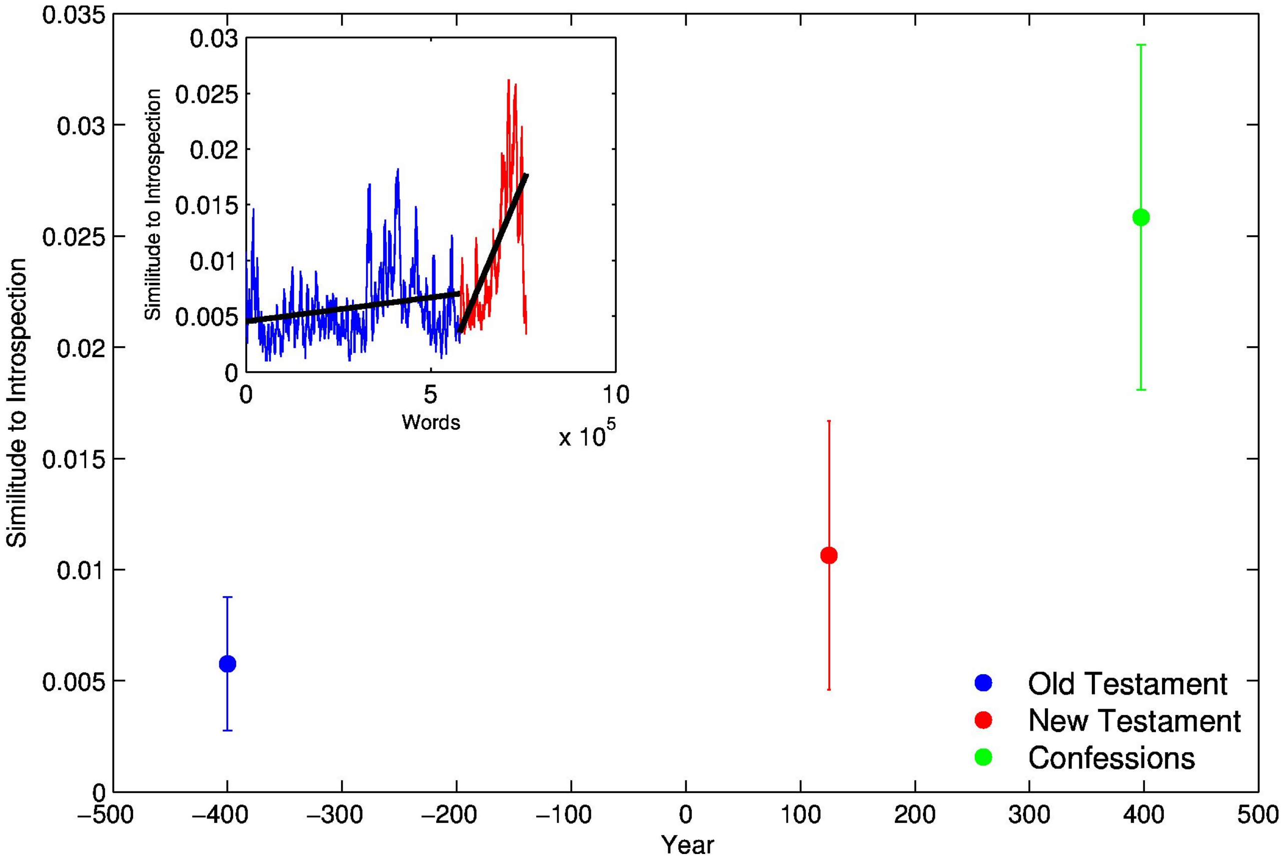
- For each document *d*:
  - For each word *w* in *d*:
    - Calculate LSA distance to introspection: d_{LSA}(w,'introspection')

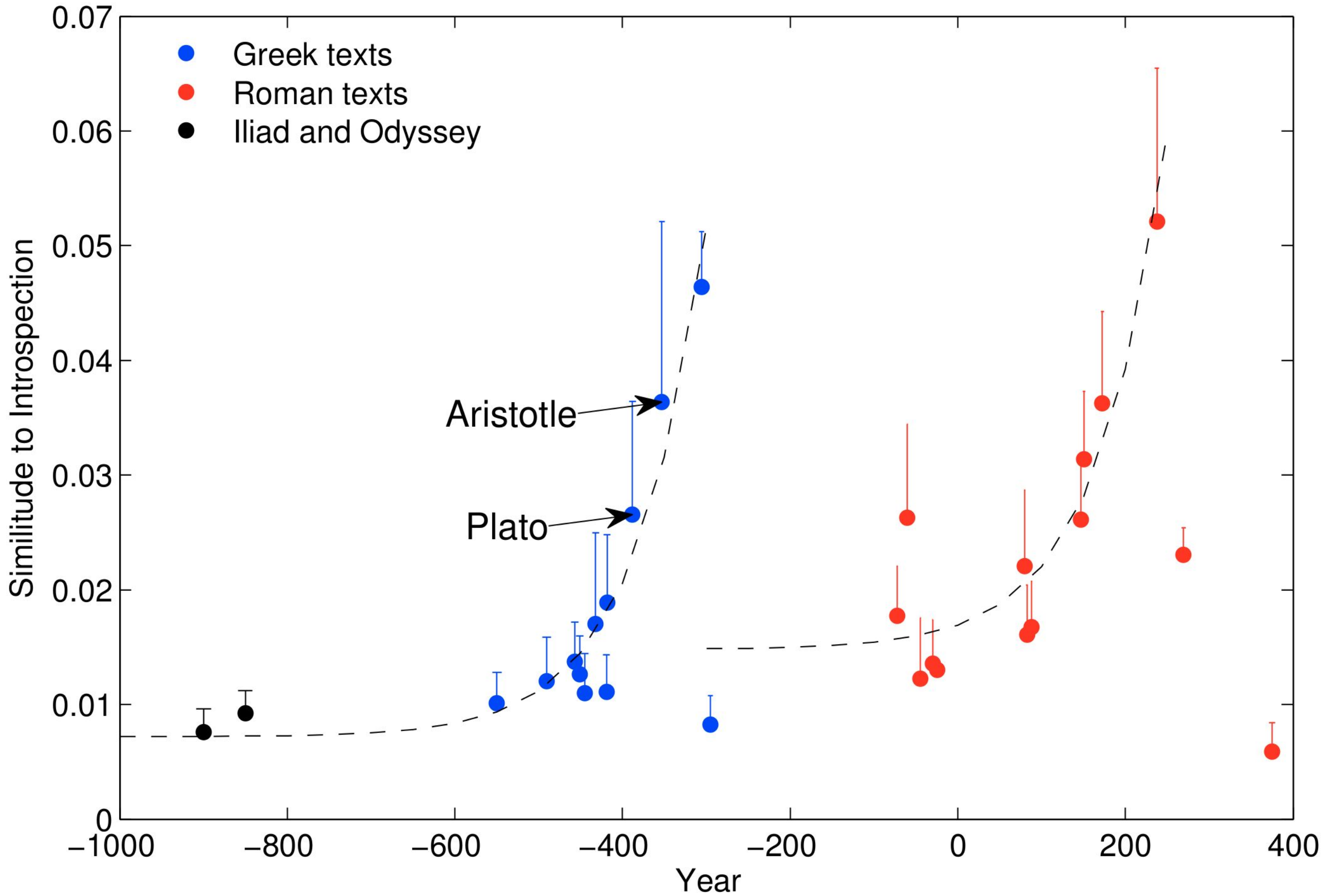
- Calculate mean, std, median, etc, on d_{LSA}

- Compare with control words: 200 concepts deemed "fundamental" across 87 Indo-European
  - Pagel, M., Atkinson, Q. D., & Meade, A., *Frequency of word-use predicts rates of lexical evolution throughout Indo-European history*. **Nature**, 449 (7163), 717–720, 2007.

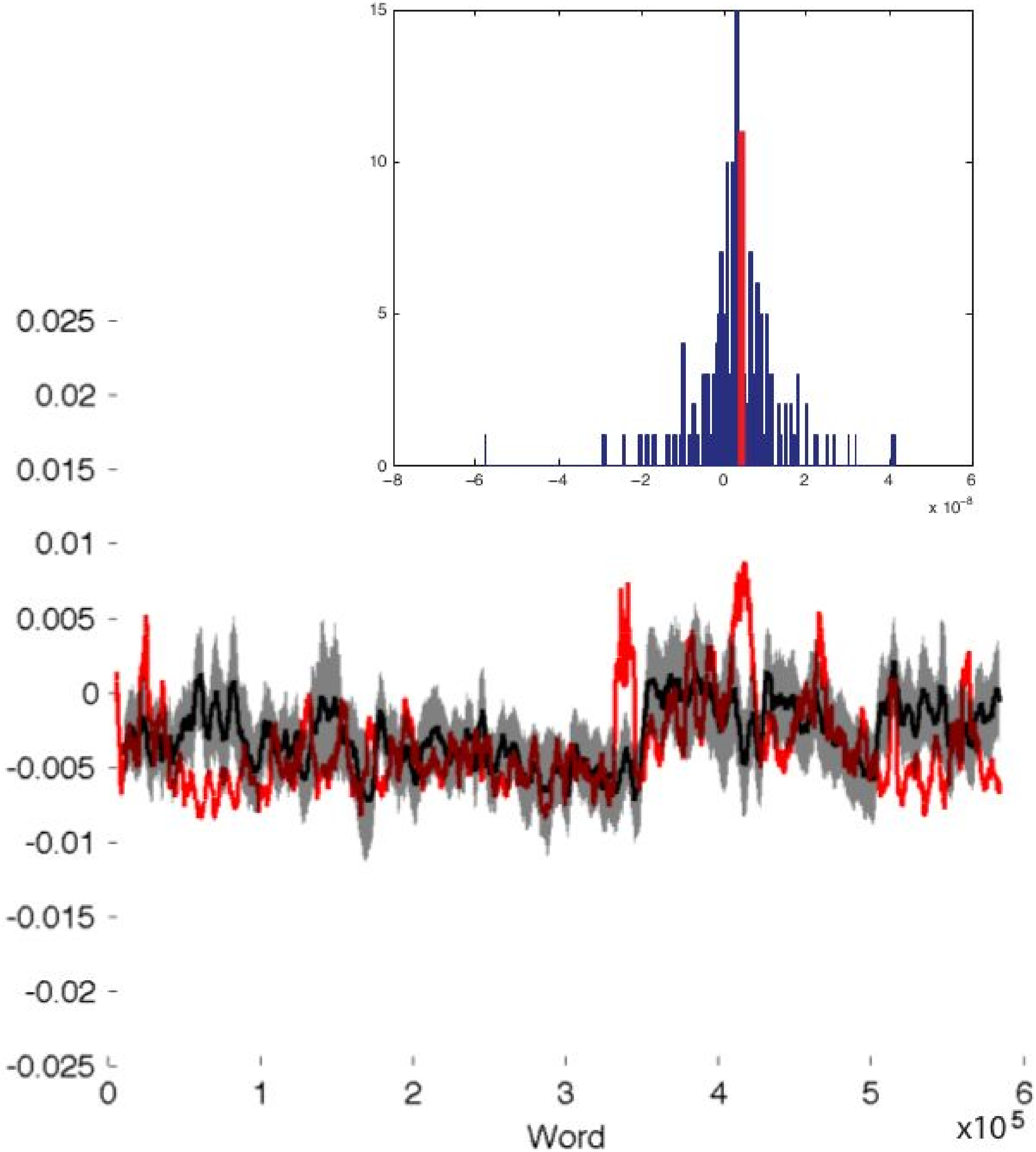
Confirmation of transition hypothesis

- Transition from the Bible to modern literature: Confessions, San Augustine (398 AC). First Autobiography, considered the first introspective text.

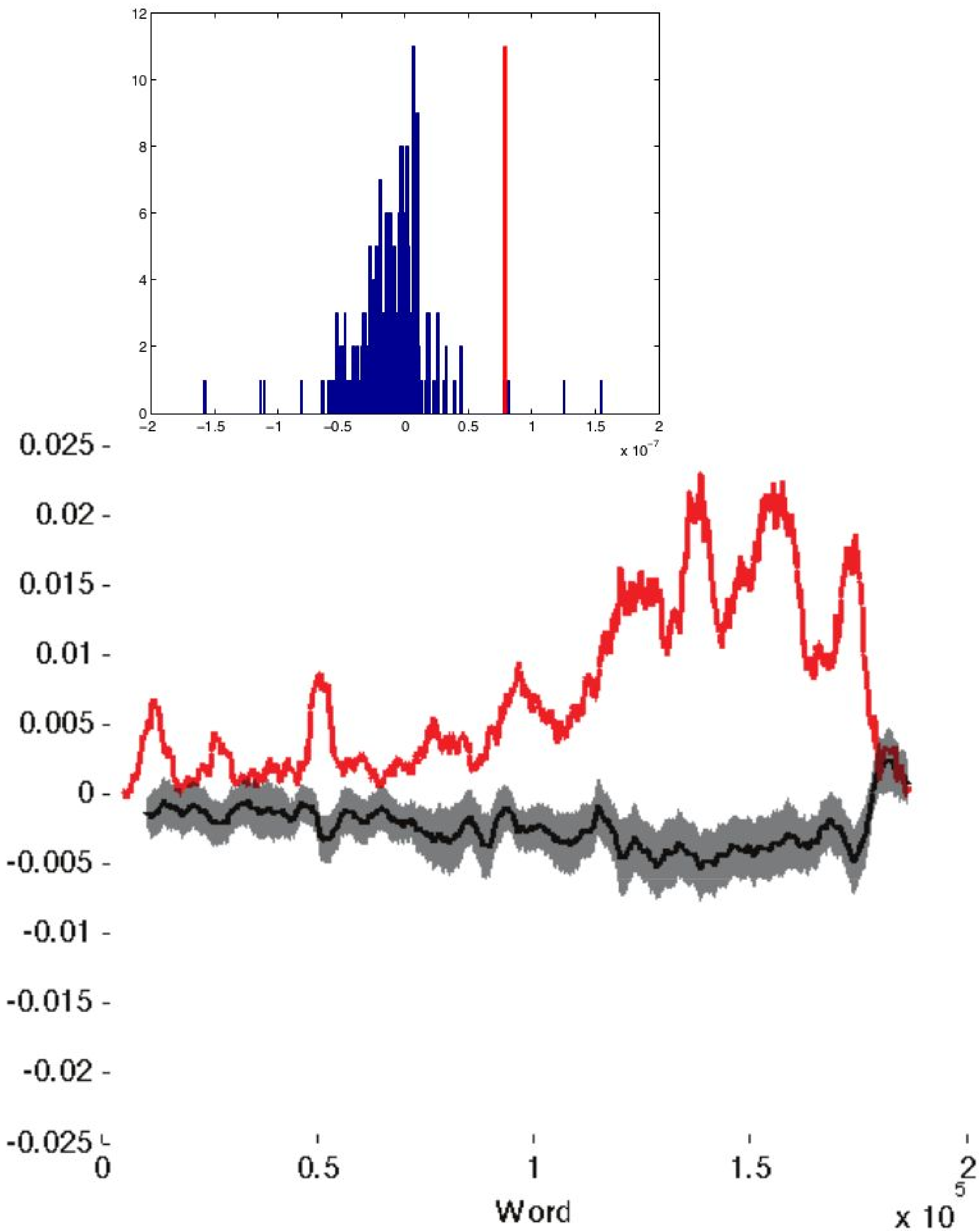




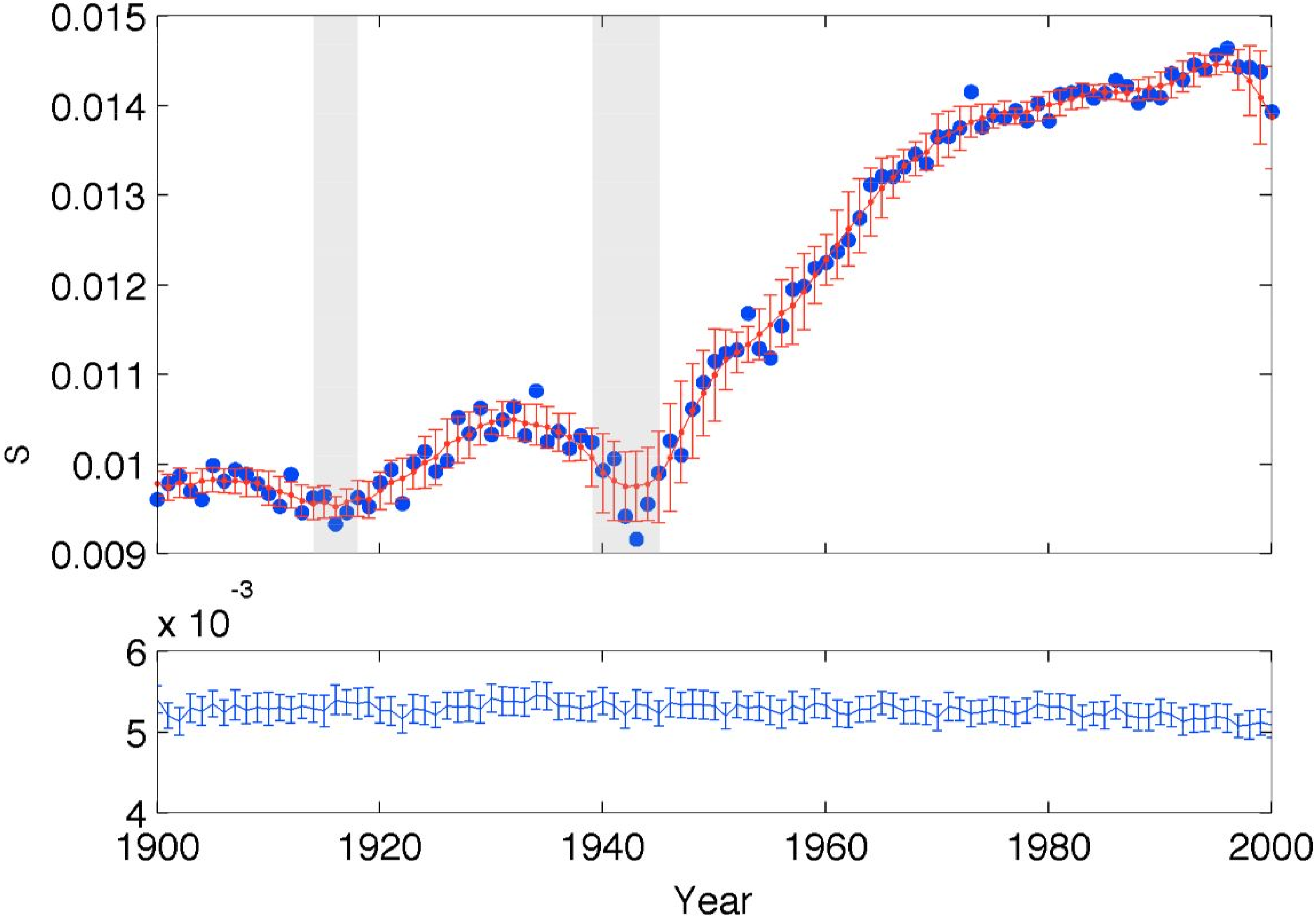
Introspection vs. control words @ Old testament



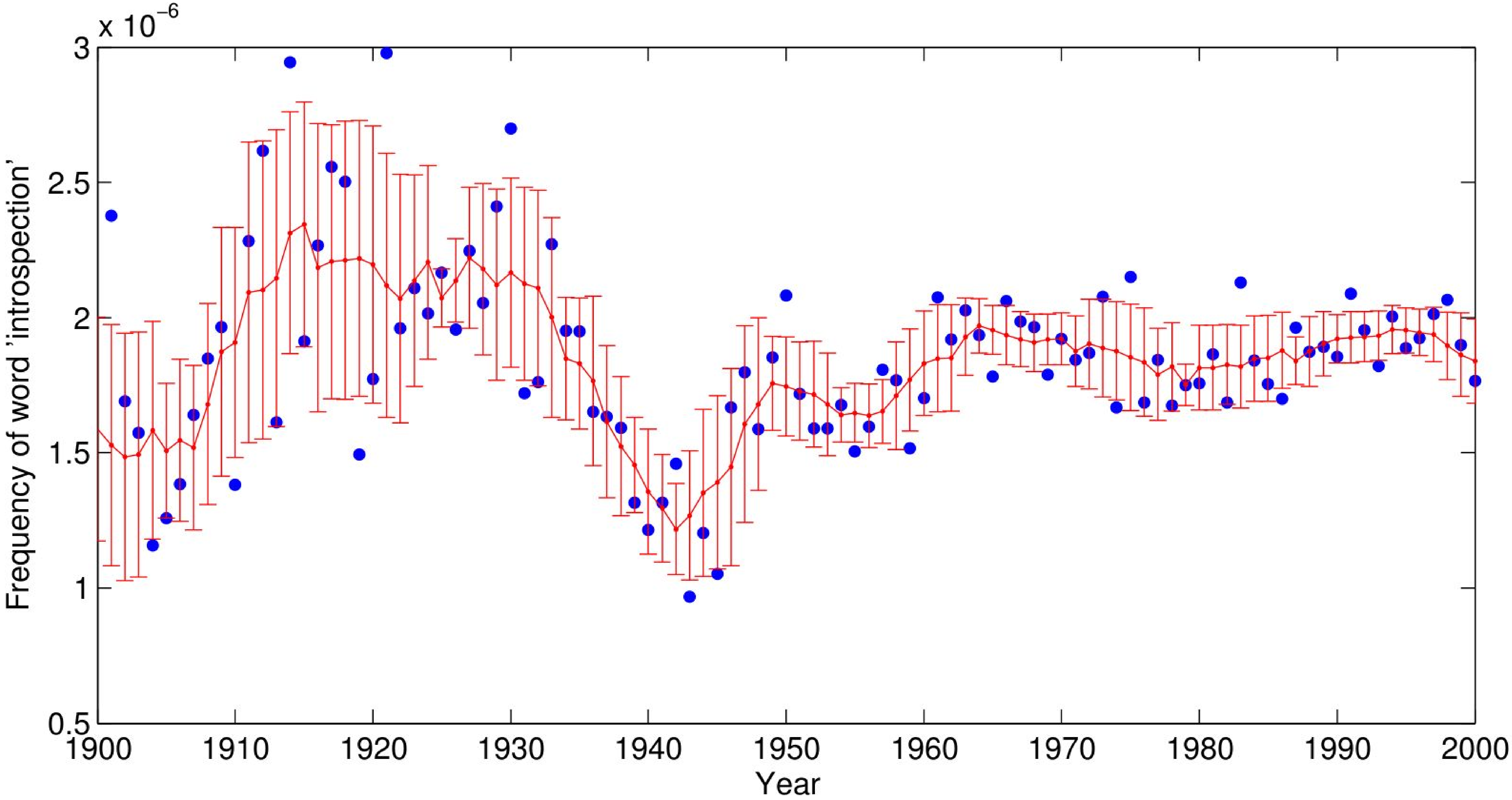
Introspection vs. control words @ New testament



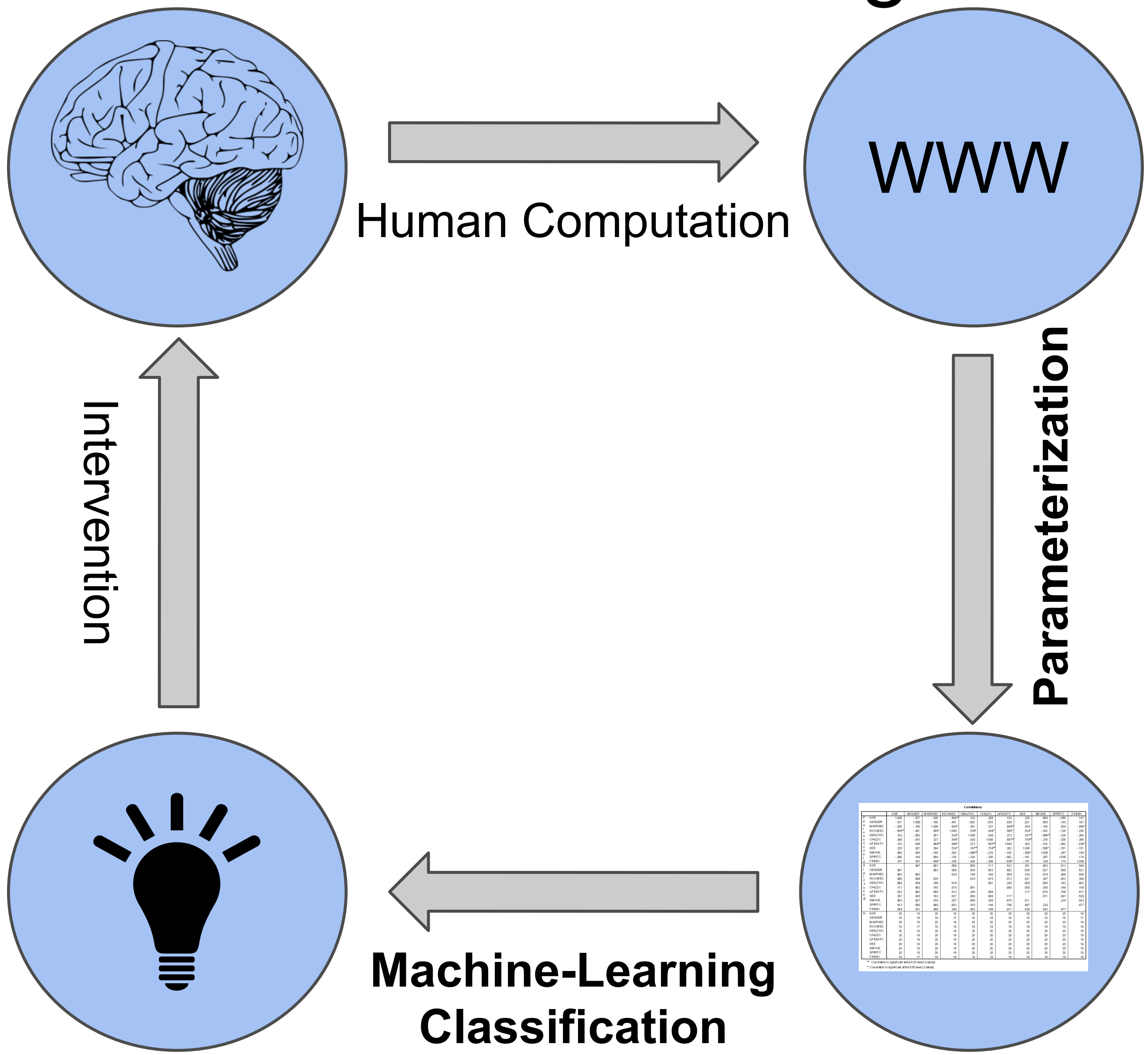
Evolution of Introspection in modern cultural records



Introspection @ Google Ngrams



Characterization of human behavior with the advent of Big Data



Can we analyse text from subjects to identify mental states?

Computational Psychiatry

- Characterize mental states by studying behavior through problem solving
    - Montague, P., et al. *Computational psychiatry*, **Trends in cognitive sciences** 16.1 (2012): 72-80.

- Text as a vehicle

- http://computationalpsychiatry.org/

Paradigm Pipeline

Interview

Manual
transcription

Hack

Machine Learning

Psychiatric Interview

input:
- Interviewer
- Subjects
- Interview

output:
- interview audio

problems:
- interviewer bias

Interview

Manual transcription

Hack

Machine Learning

Manual Transcription

input:

interviews audio

output:

transcriptions texts + tags

problems:

audio quality
transcriber interpretation
punctuation, indeterminism

Interview

Manual transcription

Hack

Machine Learning

Hack:

input:

interview texts + tags

hypothesis, clues, etc

output:

features

problems:

computational performance, availability, NLP

Interview

Manual transcription

Hack

Machine Learning

Machine Learning:

input:

features & classes

output:

classifier

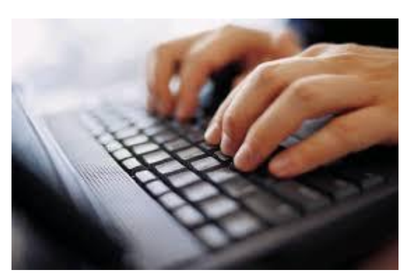
performance

problems:

overfitting (data < features, feature selection)

validation strategies

Entrevista psiquiátrica

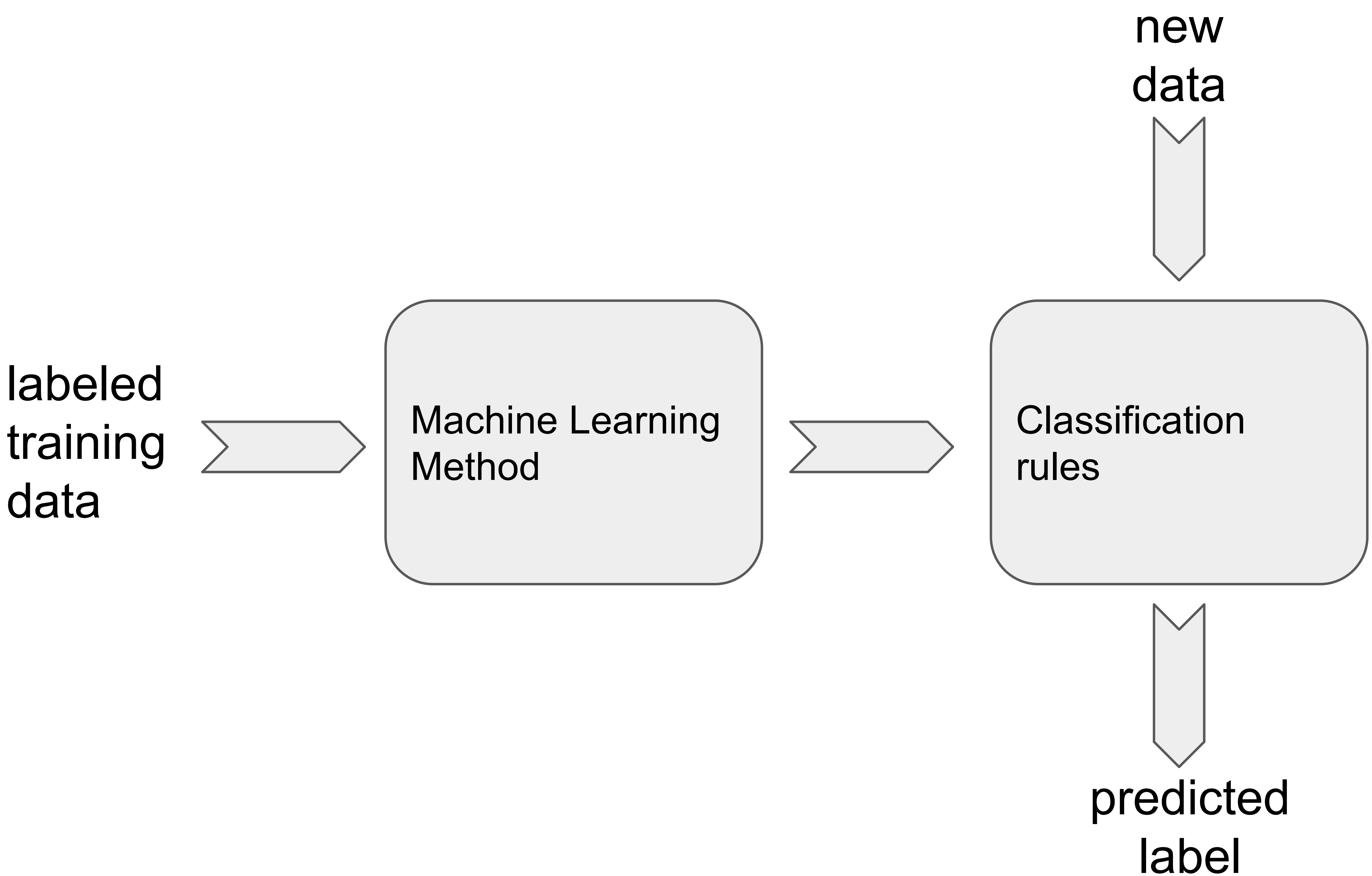


Transcripción manual

Hack

Machine Learning

Classification

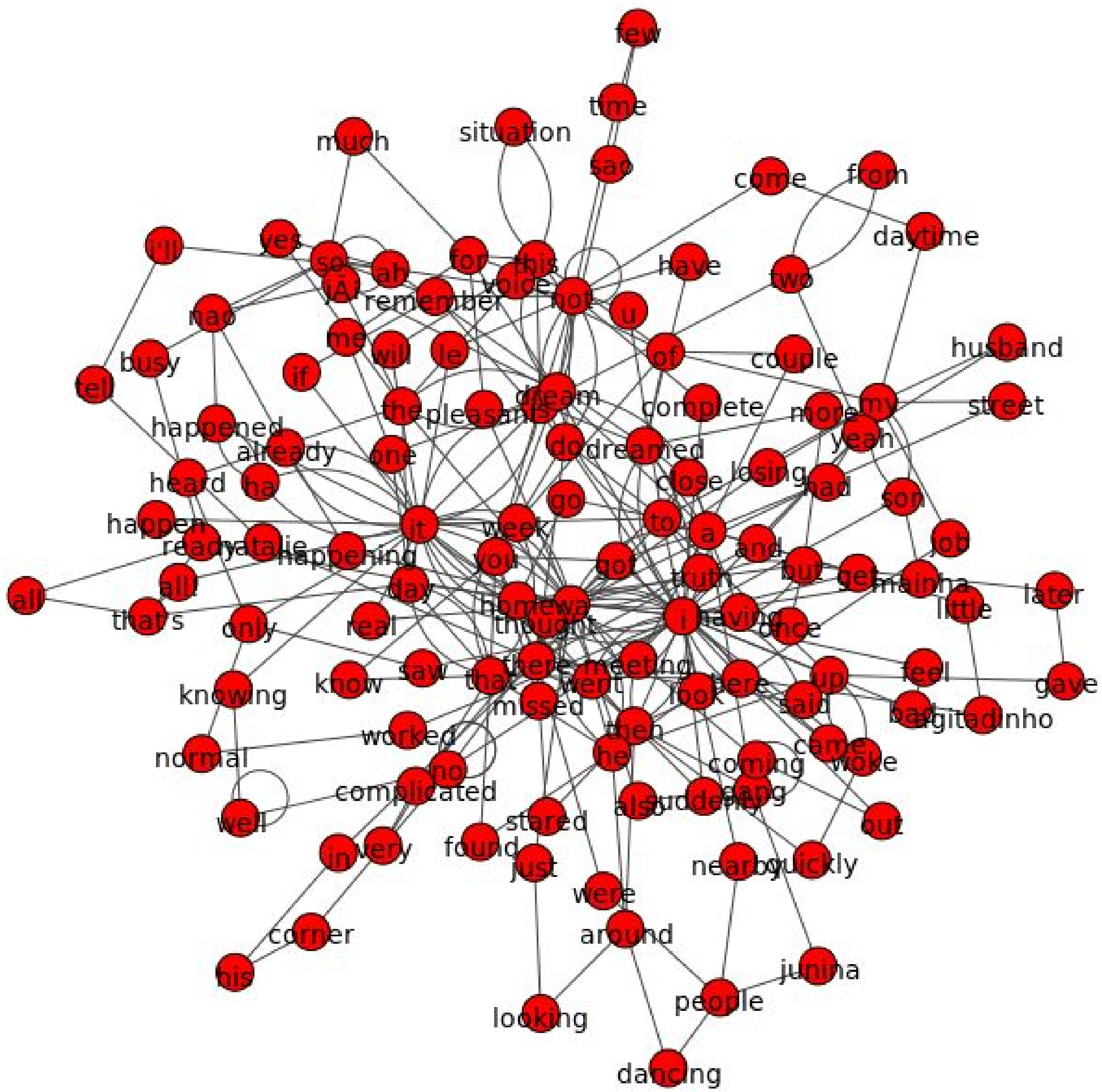


Applications

Characterizing mental states

- 63 millions of psychiatric interviews per year in USA
- Classification of speech as a marker of thought disorder in psychosis
- Classification of interviews of subjects under the effect of Ecstasy
- Classification of interviews of prodrome for schizophrenia

Psychographs



word ⟶ **node**

Classic metrics for graphs:

- # nodes, # edges
- parallel edges
- diameter
- shortest path mean

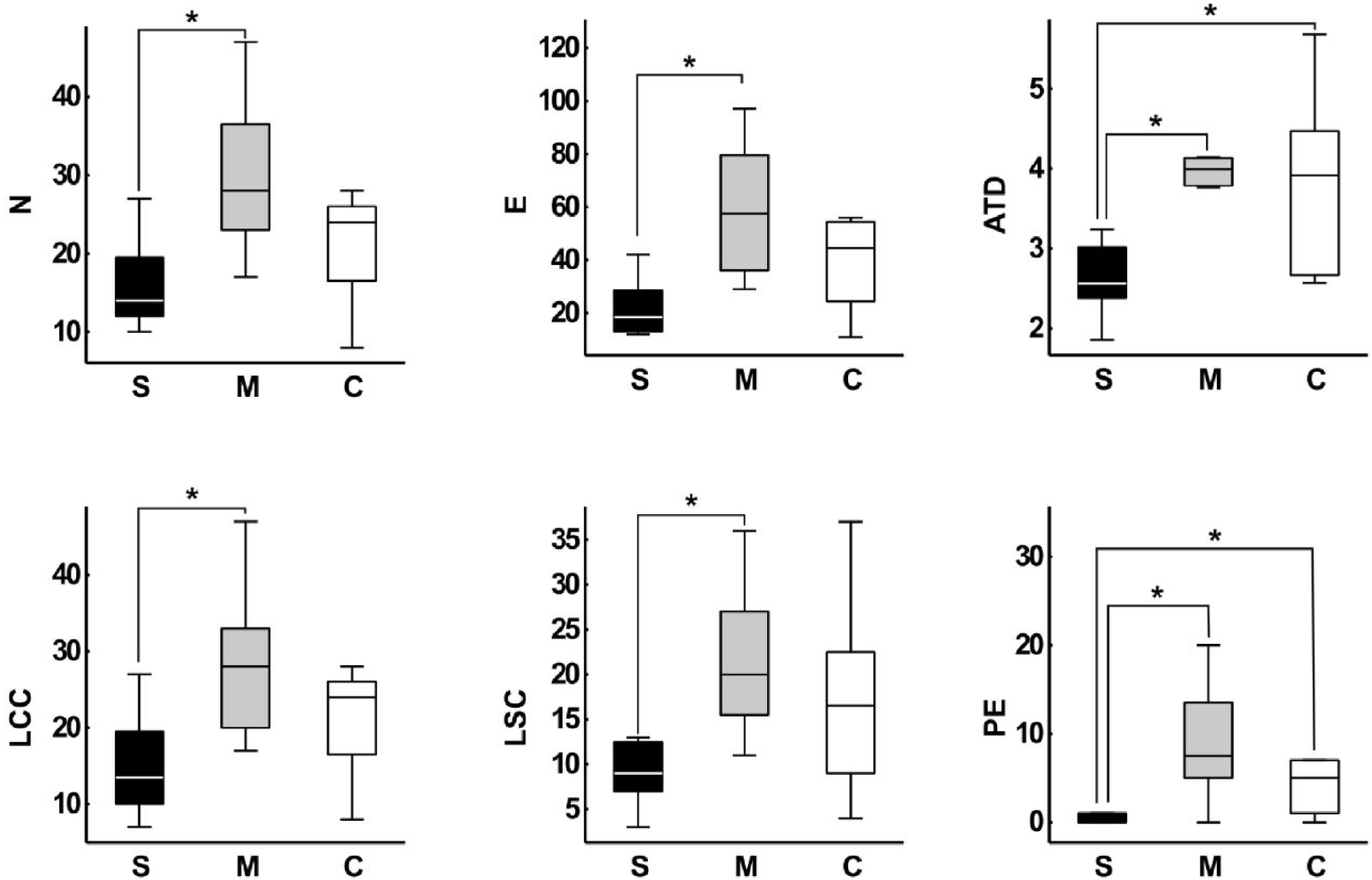
Manual process with ad hoc transformations

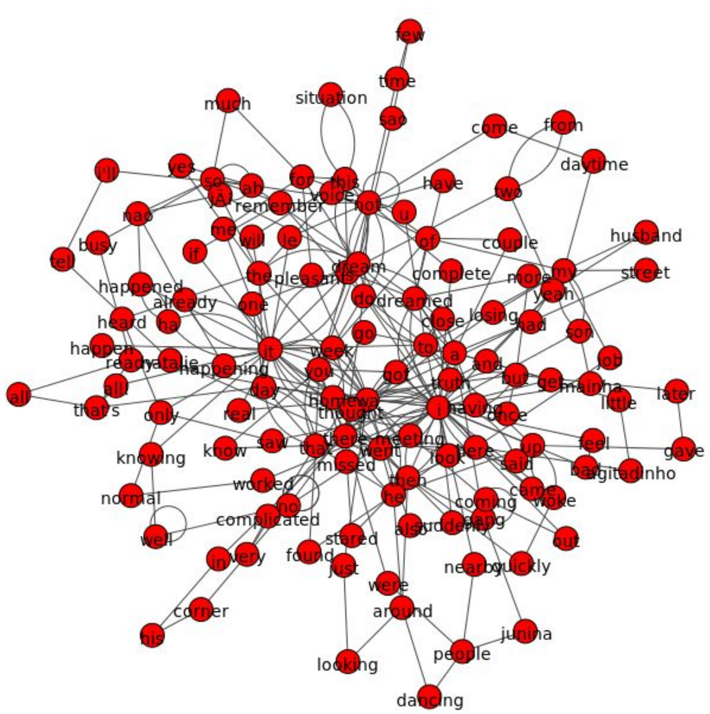
[1] : N.B. Mota, N.A.P. Vasconcelos, N. Lemos, A.C. Pieretti, O. Kinouchi, G.A. Cecchi, M. Copelli, and S. Ribeiro. Speech graphs provide a quantitative measure of thought disorder in psychosis. PloS one, 7(4):e34928, 2012.

Single metrics classification

Subjects

- 8 **S**chizophrenic
- 8 **M**aniac
- 8 **C**ontrol





[1] : N.B. Mota, N.A.P. Vasconcelos, N. Lemos, A.C. Pieretti, O. Kinouchi, G.A. Cecchi, M. Copelli, and S. Ribeiro. Speech graphs provide a quantitative measure of thought disorder in psychosis. PloS one, 7(4):e34928, 2012.

Graphs (automated)

- Automate process

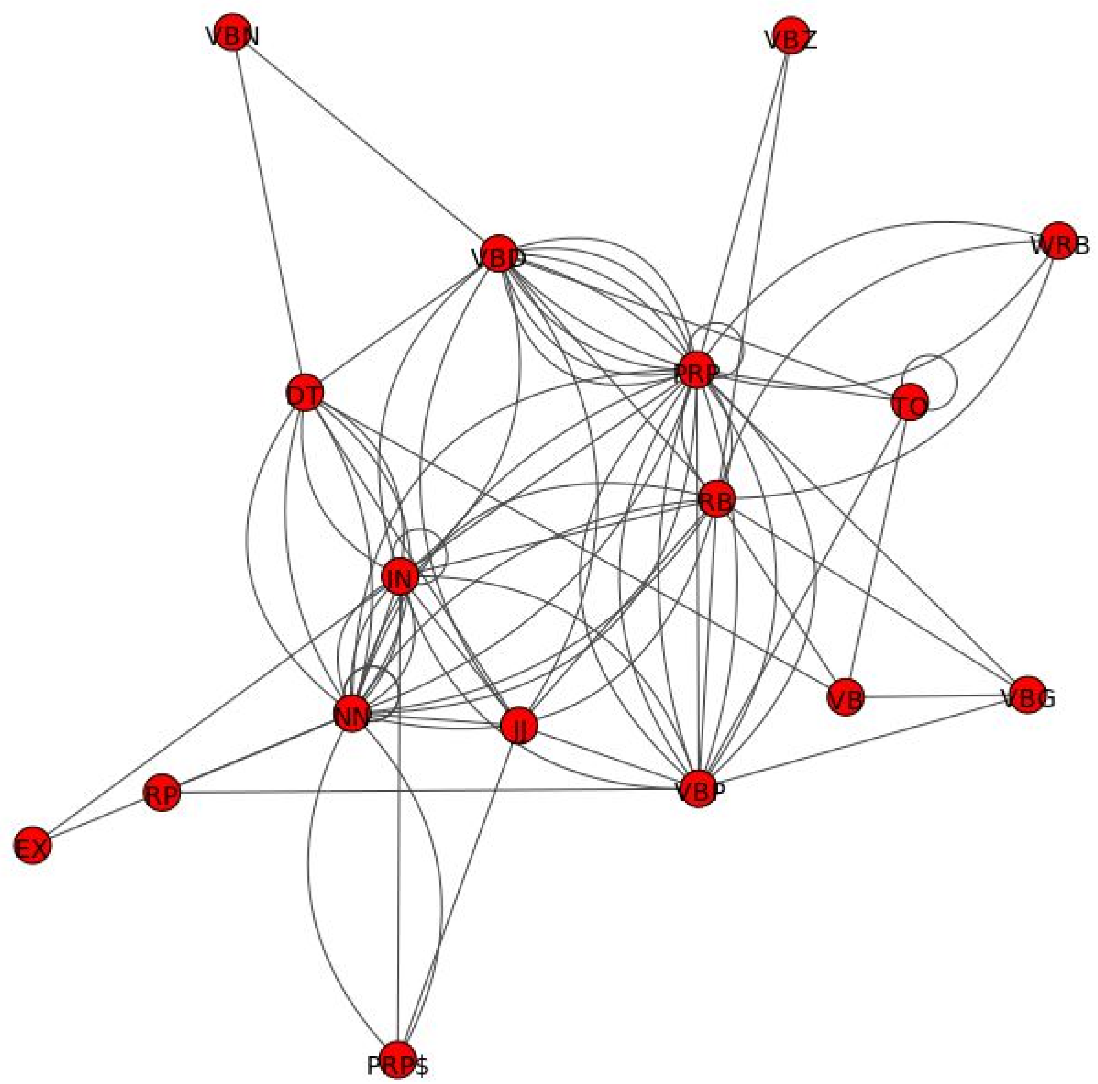
- **More** features!

word → **node**

word → **Lemma** → **node**

word → **POS Tag** → **node**

- Generate **classifier**



Full automatic classification

Classification with only two parameters
(Naive Bayes - 10 folds cross validation)

	S vs M vs C	**S vs M**	**S vs C**	**M vs C**
Previous work	0.6250	0.9380	0.8750	0.6880
Naive	0.6250 (naive_ATD y naive_L2)	0.9375 (naive_Nodes y naive_PE)	0.8750 (naive_ATD y naive_L1)	0.6875 (naive_L1 y naive_L2)
Lemmatized	0.7500 (lem_ATD y lem_L1)	1.0000 (lem_ATD y naive_PE)	0.8750 (naive_ATD y lem_L1)	0.8125 (lem_L1 y naive_PE)

S: Schizophrenia
M: Mania
C: Control

New features… Application to Ecstasy

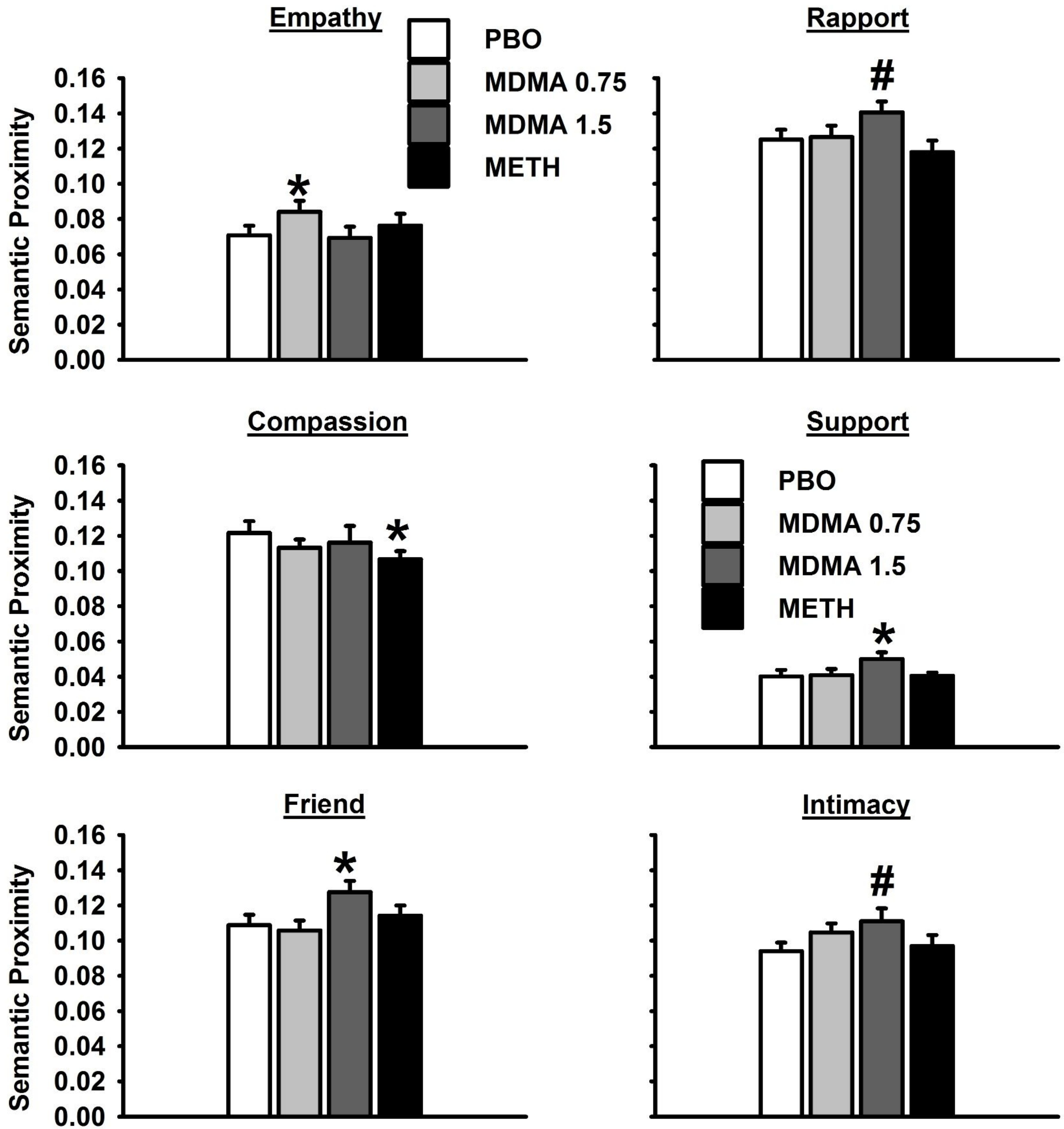
Experimental setup

- Regular Ecstasy Users (MDMA)

- Interviews in four conditions, blind to subject. (~2,000 words):
    - Placebo,
    - MDMA (high),
    - MDMA (low),
    - Methamphetamine.

- Measure semantic similarity to specific words: Empathy, Compassion, Forgiveness, etc

- Collapse interview to mean similarity to each concept word


Empathy
Rapport
Compassion
Support
Friend
Intimacy
PBO
MDMA 0.75
MDMA 1.5
METH
Semantic Proximity
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
*
#

A Window into the Intoxicated Mind

- Classifier: off-the-shelf SVM
- Leave-subject-out cross-validation

Condition I	Condition II	Accuracy		Baseline
MDMA1.5	PBO	88%	± 6%	50%
MDMA1.5	METH	84%	± 6%	50%
PBO	METH	69%	± 10%	50%
MDMA1.5	MDMA0.75	57%	± 11%	50%
METH	MDMA0.75	50%	± 11%	50%
PBO	MDMA0.75	46%	± 11%	50%
4-way		59%	± 6%	25%

Classification of interviews of prodrome for schizophrenia

Schizophrenia

Wikipedia

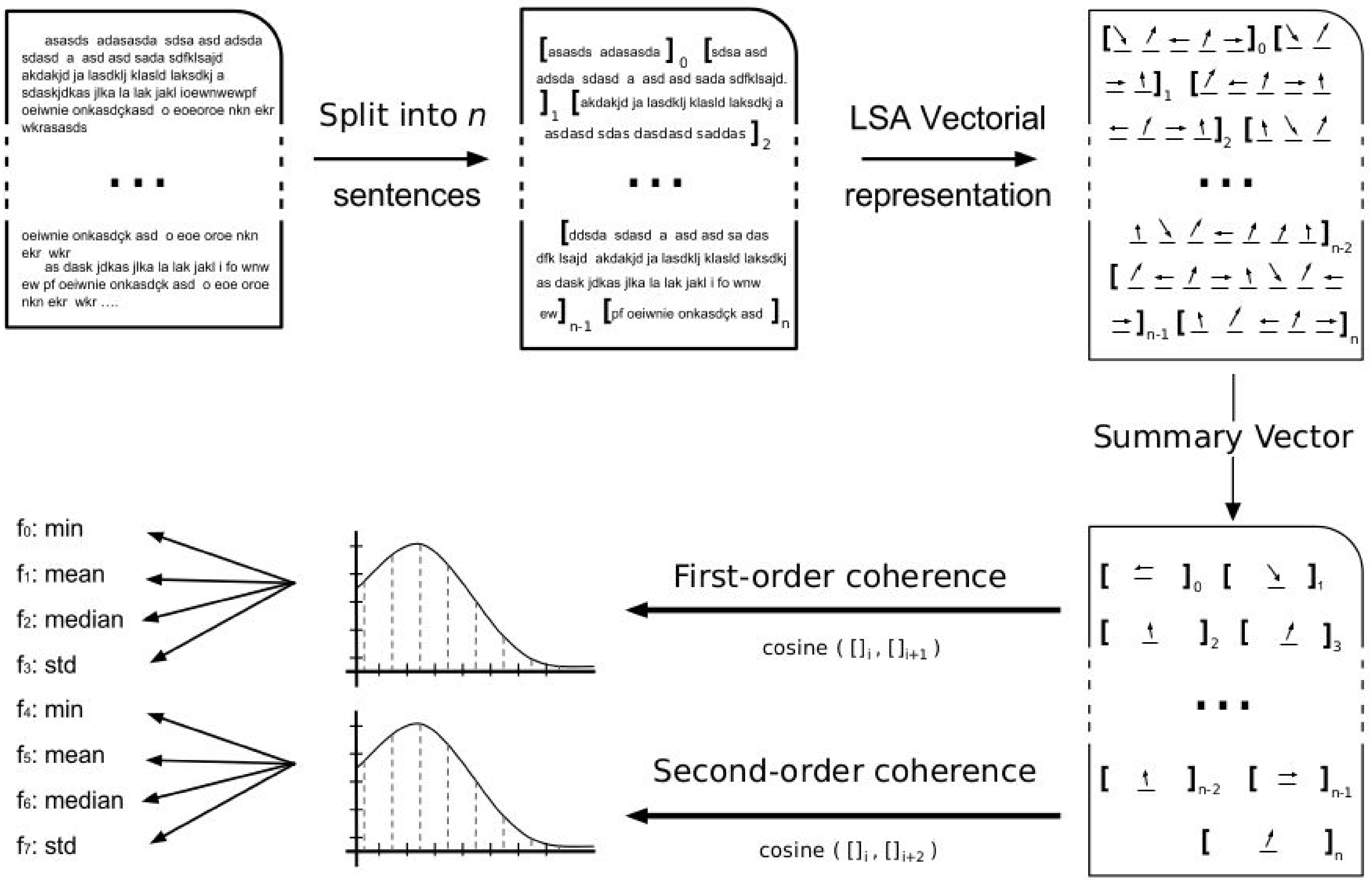
- **Mental disorder** often characterized by **abnormal social behavior** and **failure to recognize what is real**.

- **Diagnosis** is based on observed behavior and the **person's reported experiences**.

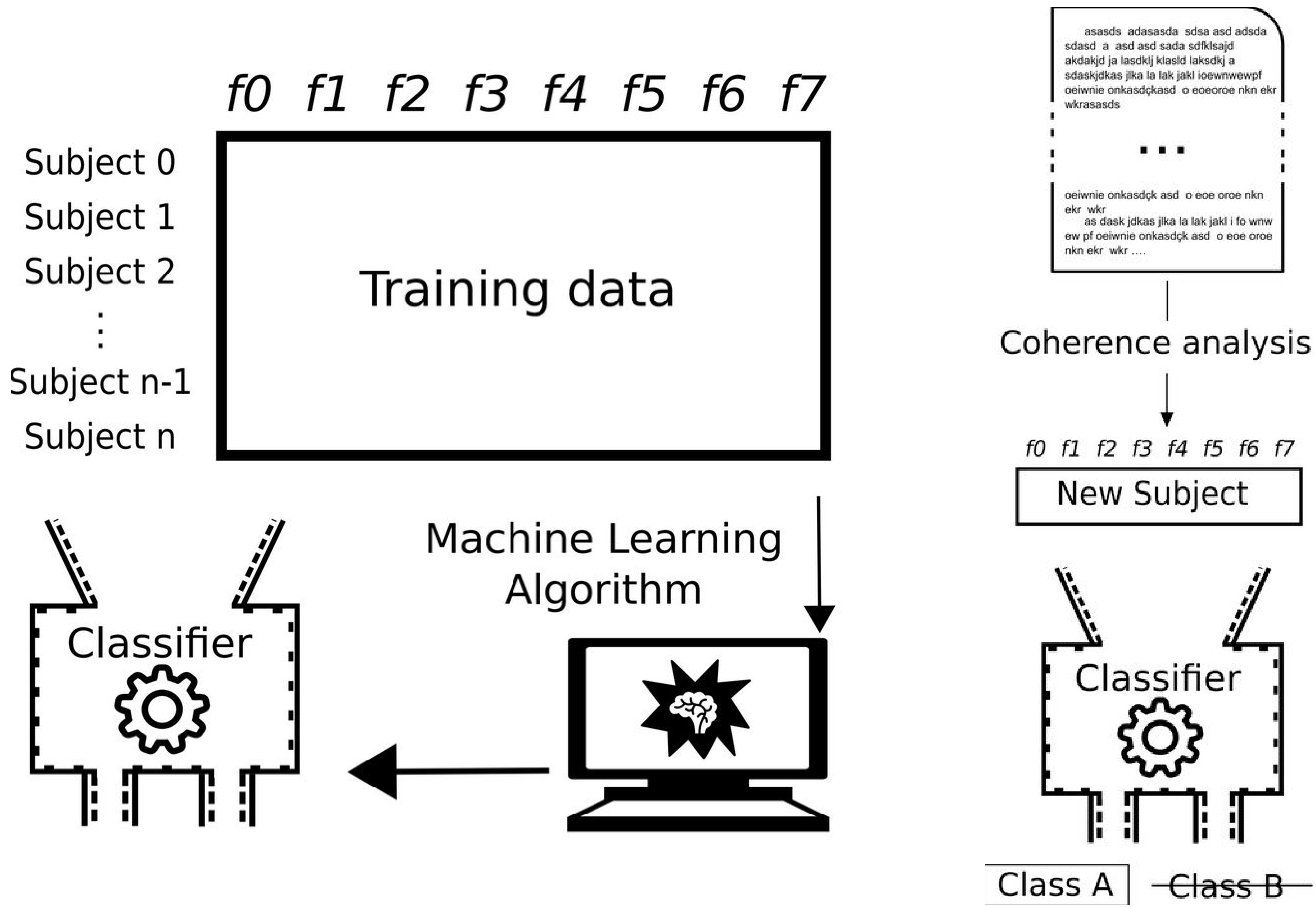
- Criteria
  - DSM-5: To be diagnosed with schizophrenia, **two diagnostic criteria** have to be met over much of the time of a period of at least one month, with a significant impact on social or occupational functioning for at least six months. The person had to be suffering from **delusions, hallucinations or disorganized speech**.

Speech Coherence

- Coherence: semantic distance between sentences



Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia, B Elvevåg, et al., **Schizophrenia research**, 2007



f0 f1 f2 f3 f4 f5 f6 f7
Subject 0
Subject 1
Subject 2
Subject n-1
Subject n
Training data
Machine Learning Algorithm
Classifier
Coherence analysis
f0 f1 f2 f3 f4 f5 f6 f7
New Subject
Classifier
Class A
~~Class B~~

Speech coherence: application to prodrome

Subjects (35) **decide to go** to Columbia Hospital because of anxiety, depression and other symptoms

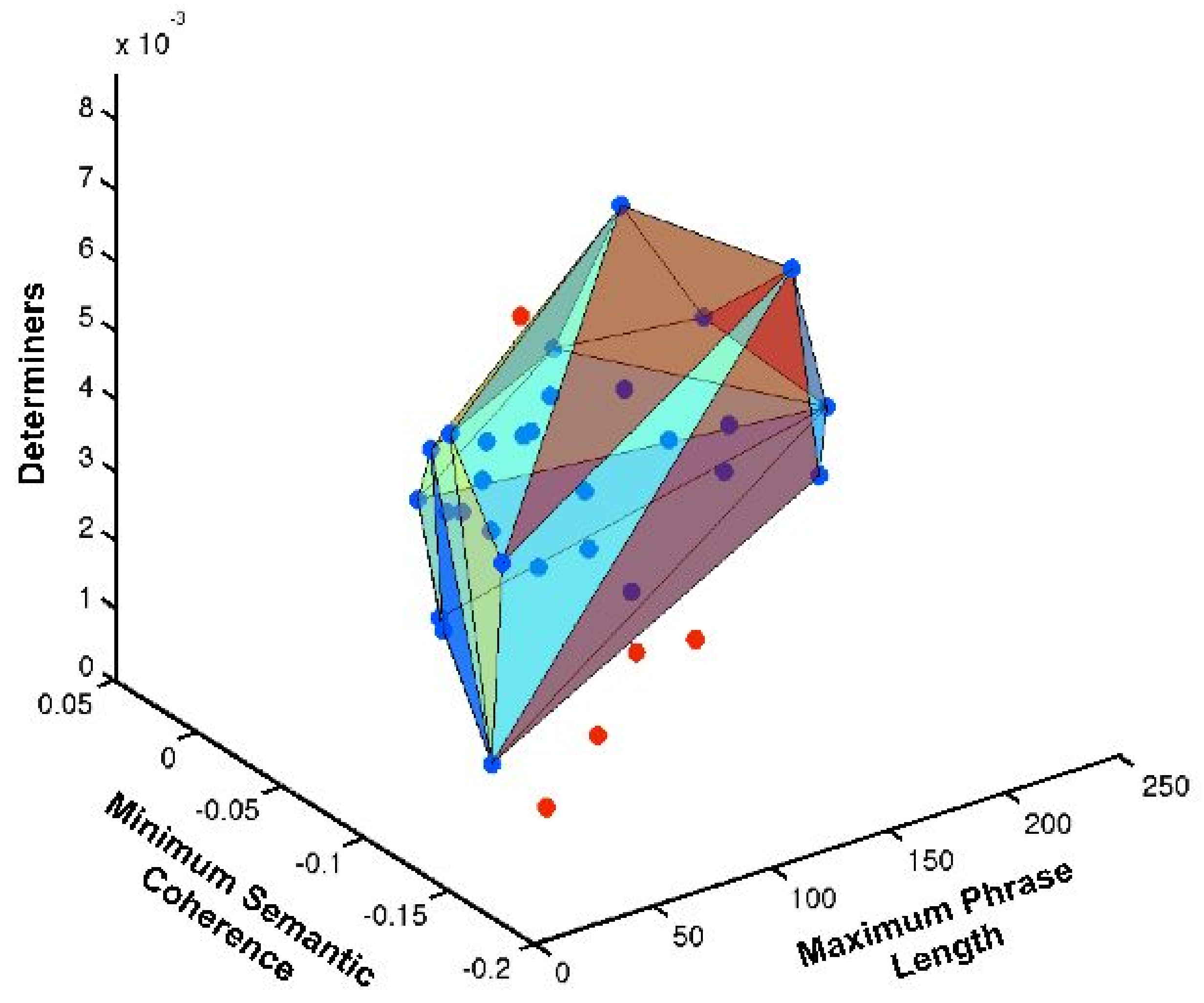
Guided interview **10'**: "Who are you? What are you feeling?"

No diagnostic before interview

One year later, **5 subjects** present psychotic episodes and declared **schizophrenic**

Classification Method

- Convex hull 3-features



Classification Performance

Method	PPV	NPV	Sens.	Spec	ROC
Convex Hull 3-features	100	100	100	100	1.0
‘Black box’ 22-features	67	96	80	93	0.8
SIPS/SOPS	33	89	40	86	0.47

PPV = Positive Predictive Value;

NPV = Negative Predictive Value;

Sens. = Sensitivity;

Spec. = Specificity;

ROC = Receiver Operating Characteristic Area Under the Curve;

SIPS/SOPS = Classification based on baseline scores on the Structured Interview for Prodromal Syndromes/Scale for Prodromal Symptoms;

Other semantic distances

Google Similarity Distance

- Cilibrasi, Rudi L., and Paul MB Vitanyi. "The google similarity distance", *IEEE Transactions on Knowledge and Data Engineering,* 19.3 (2007): 370-383.

$$\begin{aligned}\mathrm{NGD}(x,y) &= \frac{G(x,y)-\min(G(x),G(y))}{\max(G(x),G(y))} \qquad \text{(III.3)}\\ &= \frac{\max\{\log f(x),\log f(y)\}-\log f(x,y)}{\log N-\min\{\log f(x),\log f(y)\}},\end{aligned}$$

- Problems:
  - If you search for “Maradona”: About 8,080,000 results. **Very imprecise value**
  - IP blocking issues

Twitter similarity distance

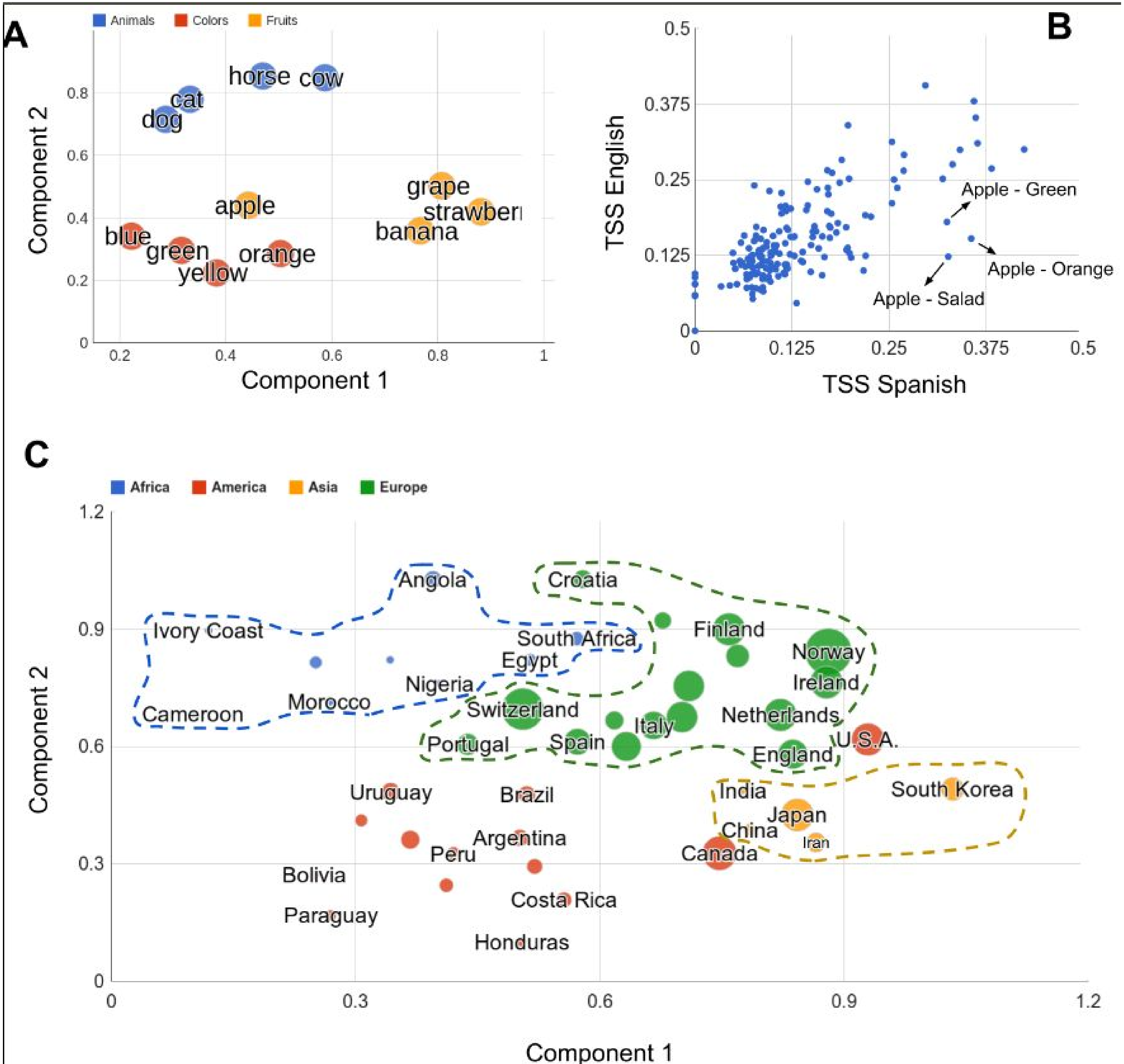
Twitter Similarity Distance

Considering the word w and the timestamp series T of tweets containing w:

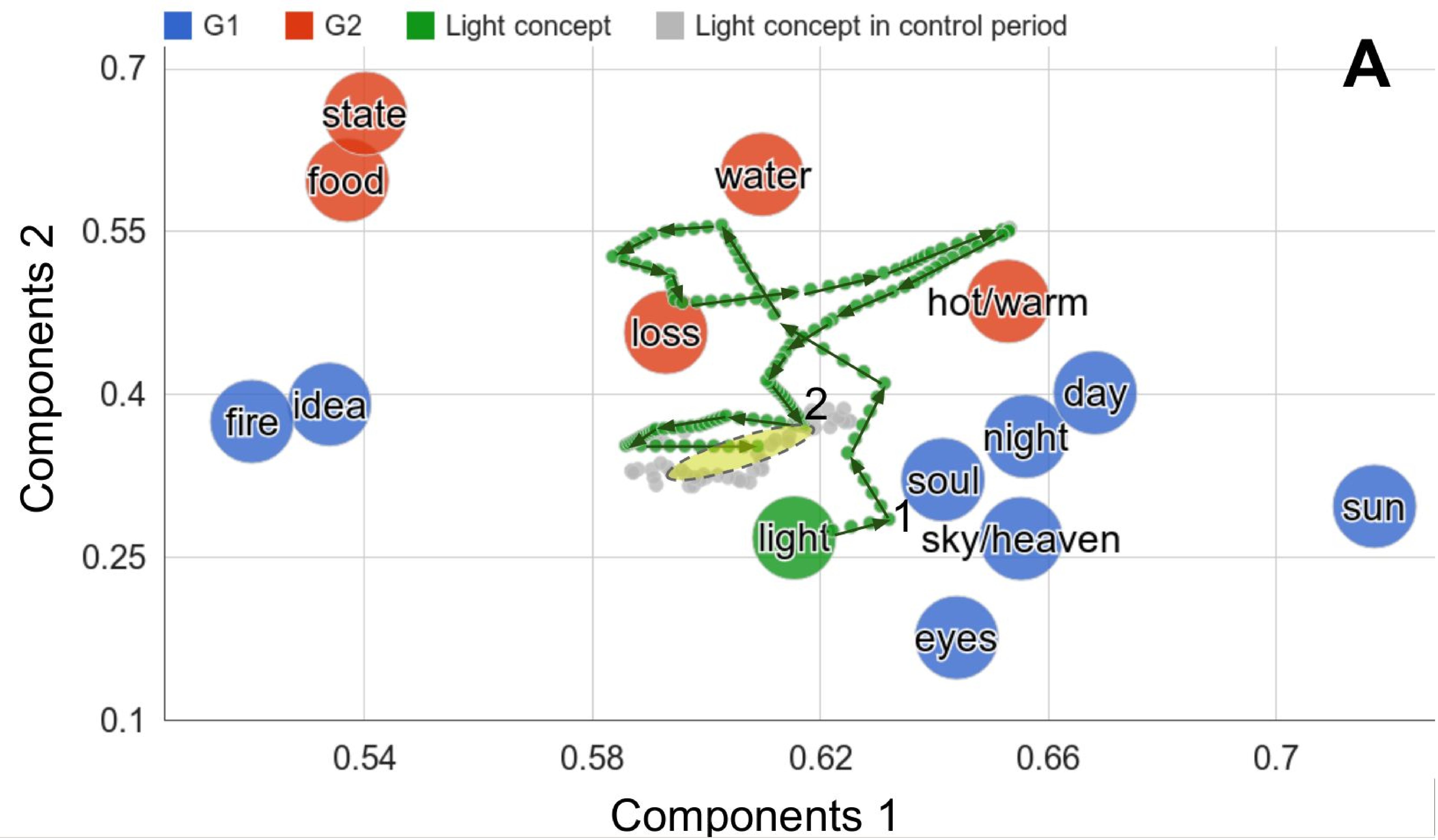
$$\Phi(w) = \left(\frac{\sum_{i=1}^{N-1} (\tau_{i+1}(w) - \tau_i(w))}{N-1} \right)^{-1}$$

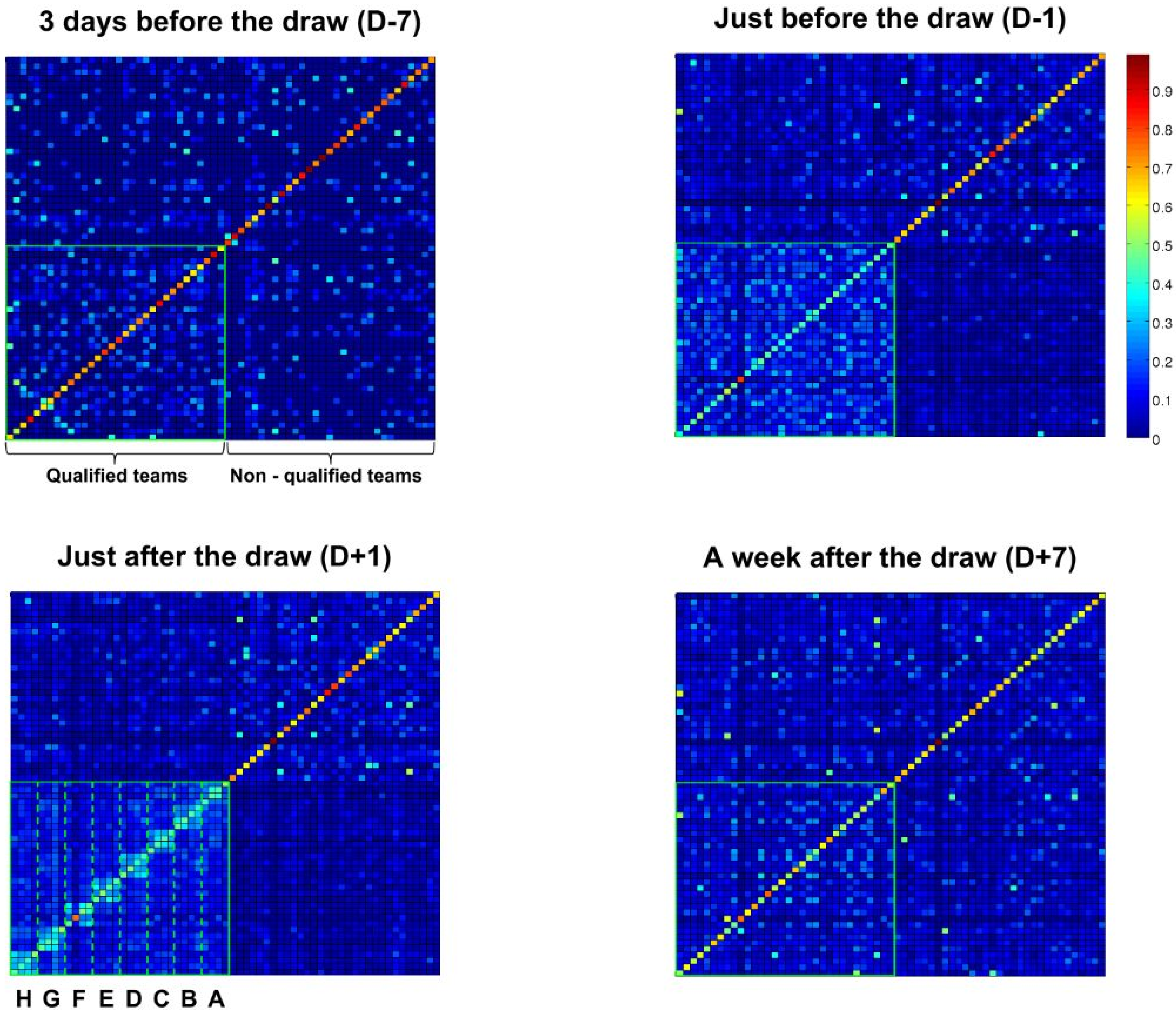
Then,

$$TSS(w_1, w_2) = \left(\frac{\Phi(w_1 \wedge w_2)}{\max(\Phi(w_1), \Phi(w_2))} \right)^{\alpha}$$



A
Animals
Colors
Fruits
horse
cow
cat
dog
grape
apple
strawberr
banana
blue
green
orange
yellow
Component 2
Component 1
B
TSS English
TSS Spanish
Apple - Green
Apple - Orange
Apple - Salad
C
Africa
America
Asia
Europe
Angola
Croatia
Ivory Coast
South Africa
Finland
Egypt
Norway
Nigeria
Ireland
Cameroon
Morocco
Switzerland
Netherlands
Italy
Portugal
Spain
U.S.A.
England
Uruguay
Brazil
India
South Korea
Japan
Argentina
China
Iran
Peru
Canada
Bolivia
Costa Rica
Paraguay
Honduras
Component 2
Component 1


G1
G2
Light concept
Light concept in control period
A
Components 2
Components 1
0.7
0.55
0.4
0.25
0.1
0.54
0.58
0.62
0.66
0.7
state
food
water
loss
hot/warm
fire
idea
day
night
soul
sun
light
sky/heaven
eyes
1
2



3 days before the draw (D-7)
Just before the draw (D-1)
Qualified teams
Non - qualified teams
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Just after the draw (D+1)
A week after the draw (D+7)
H G F E D C B A

Groups classification