



NLP @ OEG

- Pablo Calleja
 - “Named Entity Recognition over unstructured documents”
- David Chavez-Fraga
 - “Ontology Learning (Evaluation) From Text”
- Carlos Badenes-Olmedo
 - “Text Mining on Large Datasets with Topic Models”
- Víctor Fernández-Rico
 - “Knowledge Graphs Embeddings”
- Mariano Rico
 - “NL-guided queries”



Named Entity Recognition over unstructured documents

Pablo Calleja Ibáñez

Unstructured documents



Structured Information



Tokenization

Sentence Segmentation

Named Entity Recognition

Co-reference Resolution

Relation Extraction

...

Named Entity Recognition



Named entity: real-world concept denoted with a referent term or proper name. Main classifications:

- Organizations
- Persons
- Places
- Temporal units
- Numerical units

Biomedicine:

- Diseases
- Proteins
- Genes
- Substances

Named entities in a natural language document

Linguistic models

Rules

"Mr." + { Mayus Noun }

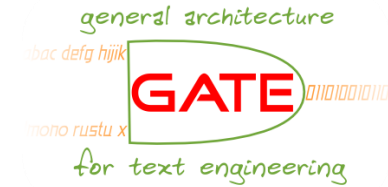


Person

{Noun _itis}



Disease



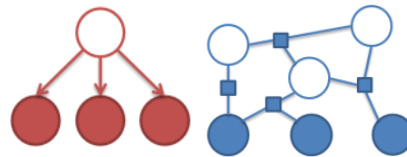
Gazetteers



Probabilistic models



+



Training Corpus

Algorithms:

- Hidden Markov Models
- Conditional random fields
- Bi-LSTM



Summary of Product characteristics

1- Drug Name

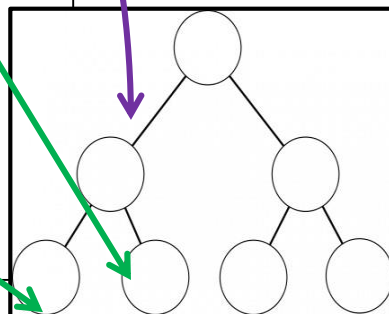
Asenapina 300 mg

4.1 Therapeutical information

For patients with schizophrenia.

4.7 Adverse reactions

May produce headaches.



SNOMED-CT

High noisy and unstructured texts:
- Leaks, emails, phone tapping...



- Person
- Company
- Phone Number
- Direction
- ...



Calleja, P., García-Castro, R., Aguado-de-Cea, L., Gómez-Pérez, A. (2017). **Expanding SNOMED-CT through Spanish Drug Summaries of Product Characteristics**. In Proceedings of the 9th International Conference on Knowledge Capture (K-CAP).



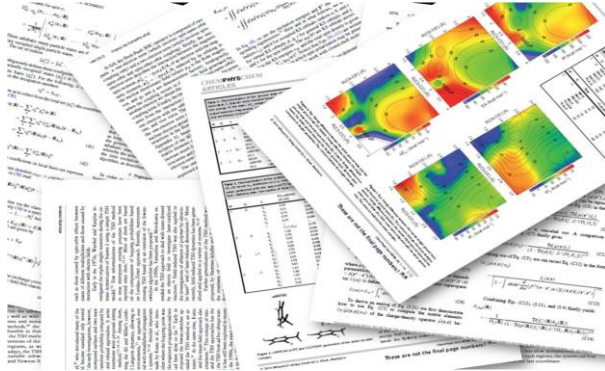
Calleja, P., García-Castro, R., Aguado-de-Cea, L., Gómez-Pérez, A. (2017). **Role-based model for Named Entity Recognition**. In Proceedings of the 11th International Conference Recent Advances in Natural Language Processing (RANLP).



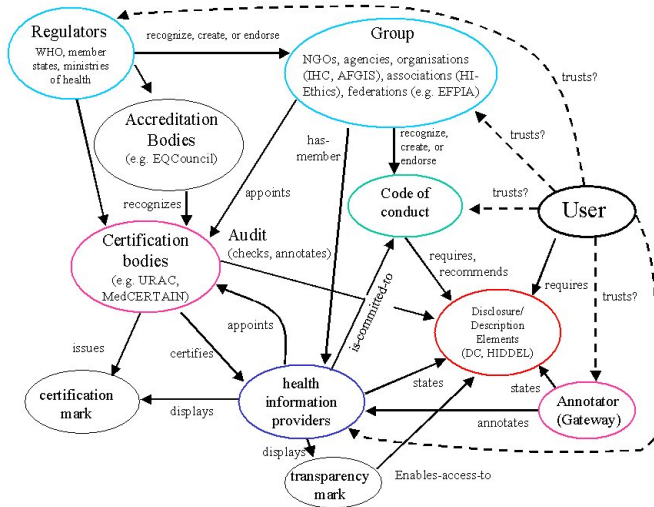
Ontology Learning (Evaluation) From Text

David Chaves-Fraga

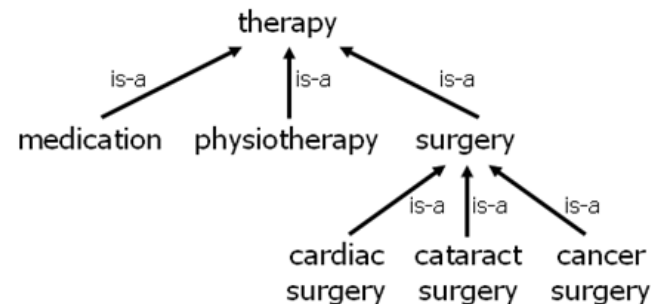
Big corpus of scientific papers from a domain



Bag of terms

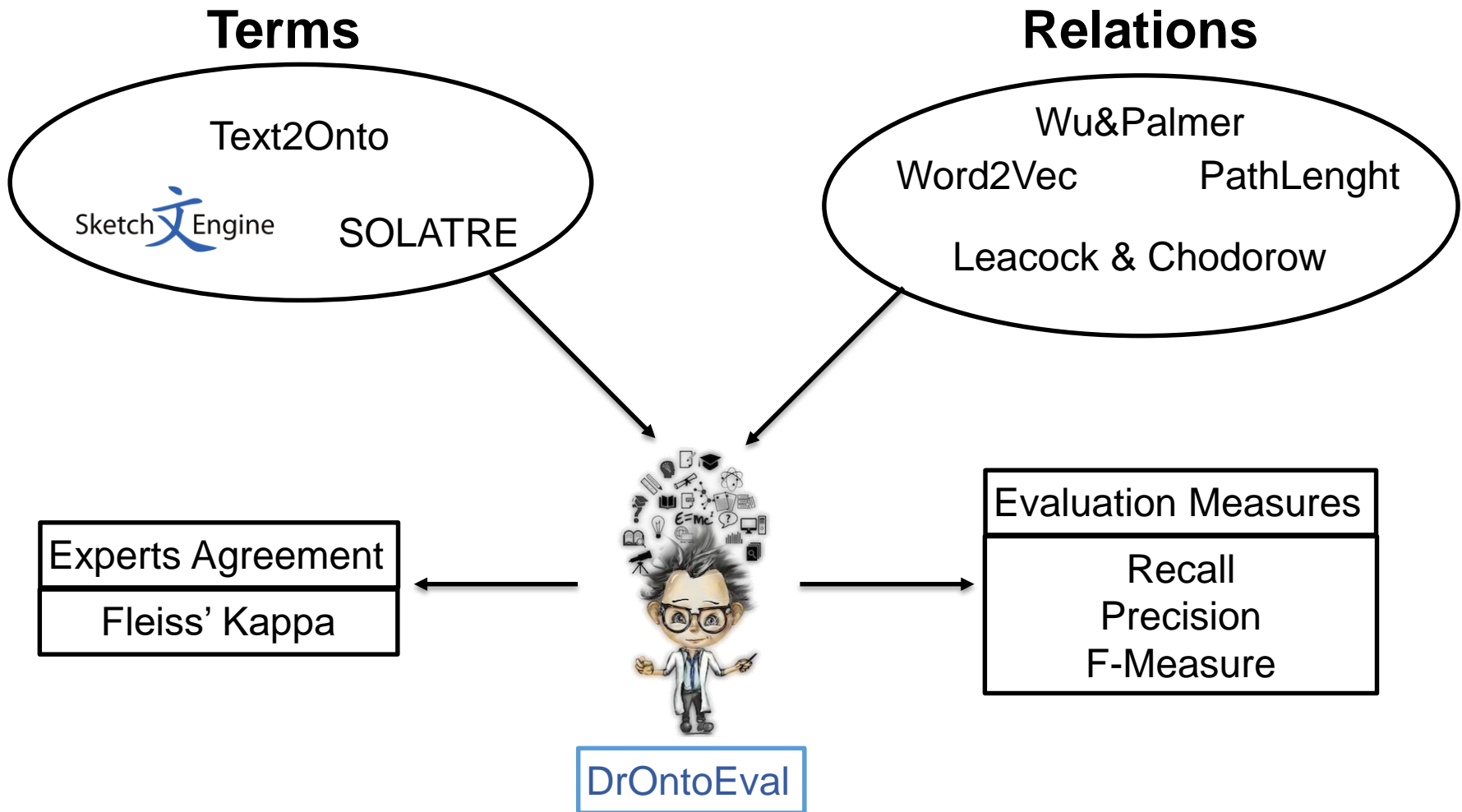


Ontology



Taxonomy

- OEG's ontology learning approach: SOLATRE
 - Only extracts terms (not very well)
 - Using by Zaporozhye National University
- Problems:
 - Developed by an ex-OEG member
 - Performance
 - External Word2Vec approach to identify relations
- Future work:
 - Implementation of new approach (part of Librairy)
 - From taxonomy to terms
 - Analysis of new systems like LexNET





Text Mining on Large Datasets with Topic Models

Carlos Badenes-Olmedo

Probabilistic Topic Models

Topics

Documents

Topic proportions
and assignments

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

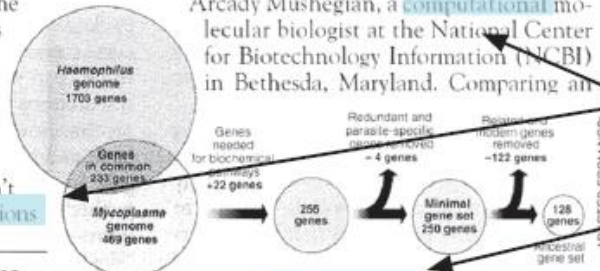
data 0.02
number 0.02
computer 0.01
...

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic numbers** game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing all



* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. **Computer analysis** yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

[0.1, 0.2, 0.4, 0.1]

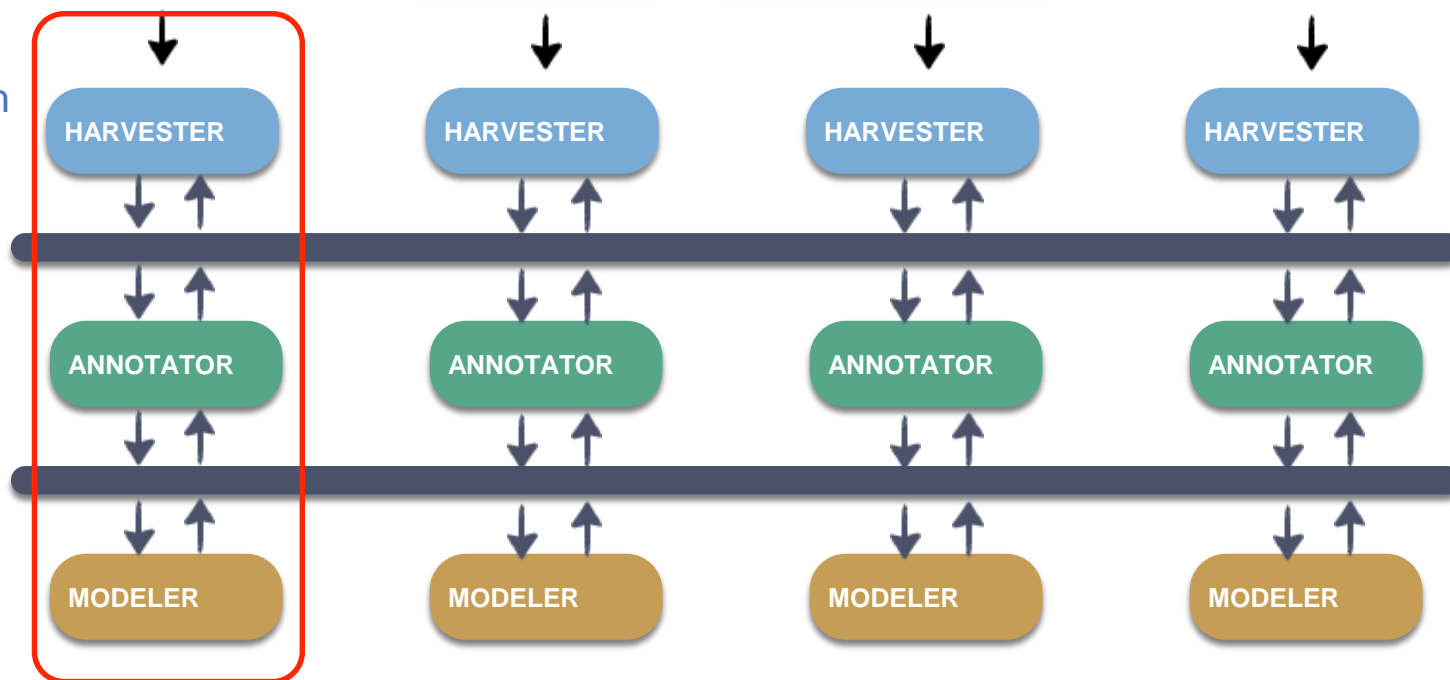
[0.3, 0.1, 0.2, 0.4]

[...]

Large Datasets

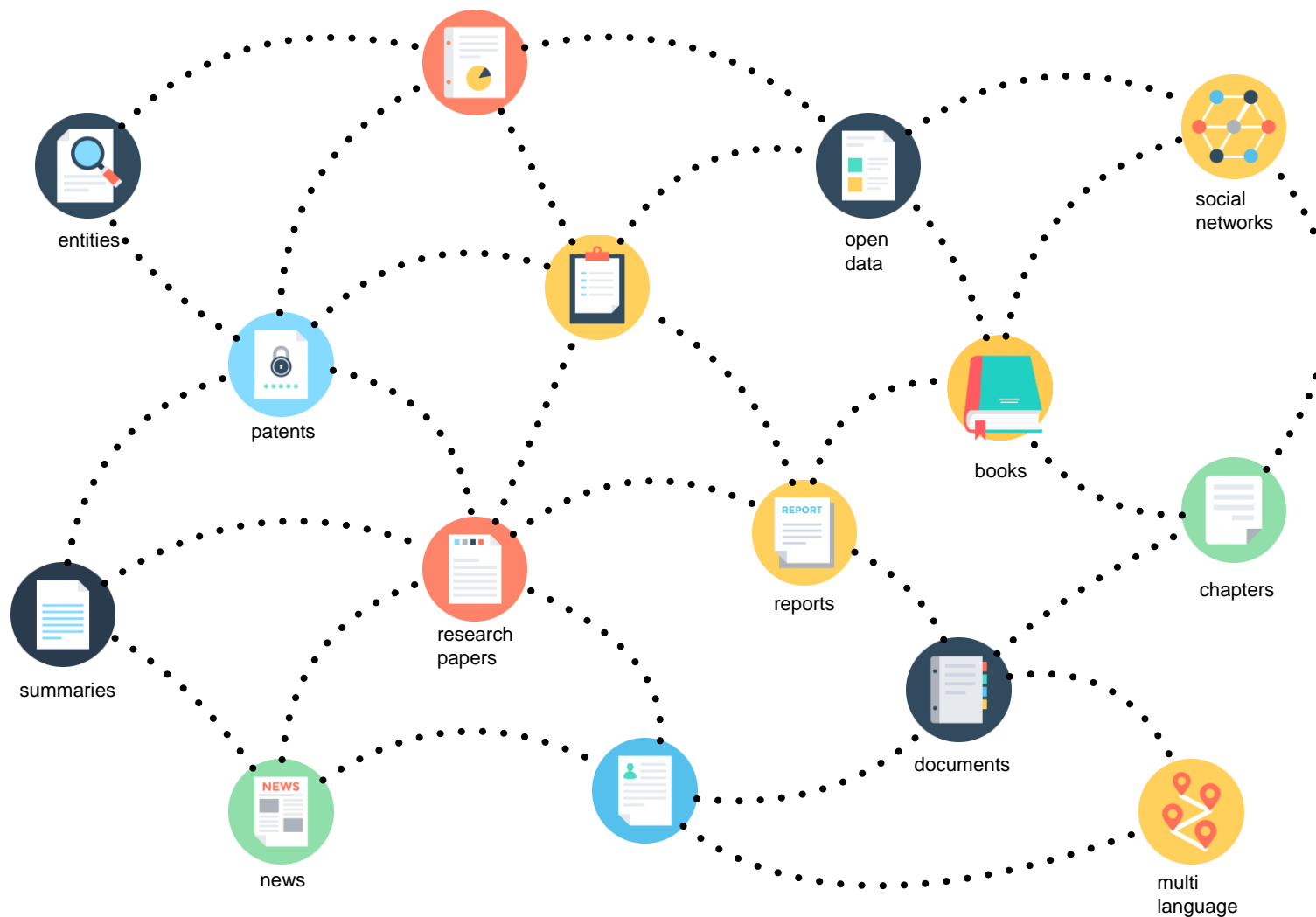


classical
approach



Badenes-Olmedo, C., Redondo-Garcia, J. L., & Corcho, O. (2017). **Distributing Text Mining tasks with librAlry.**

In Proceedings of the 17th ACM Symposium on Document Engineering (DocEng). <http://doi.org/https://doi.org/10.1145/3103010.3121040>



Badenes-Olmedo, C., Redondo-Garcia, J. L., & Corcho, O. (2017). **Efficient Clustering from Distributions over Topics**. In Proceedings of the 9th International Conference on Knowledge Capture (K-CAP).



Badenes-Olmedo, C., Redondo-Garcia, J. L., & Corcho, O. (2017). **An Initial Analysis of Topic-based Similarity among Scientific Documents based on their rhetorical discourse parts**. In Proceedings of the 1st SEMSCI workshop co-located with ISWC.



Knowledge Graphs Embeddings

Víctor Fernández-Rico

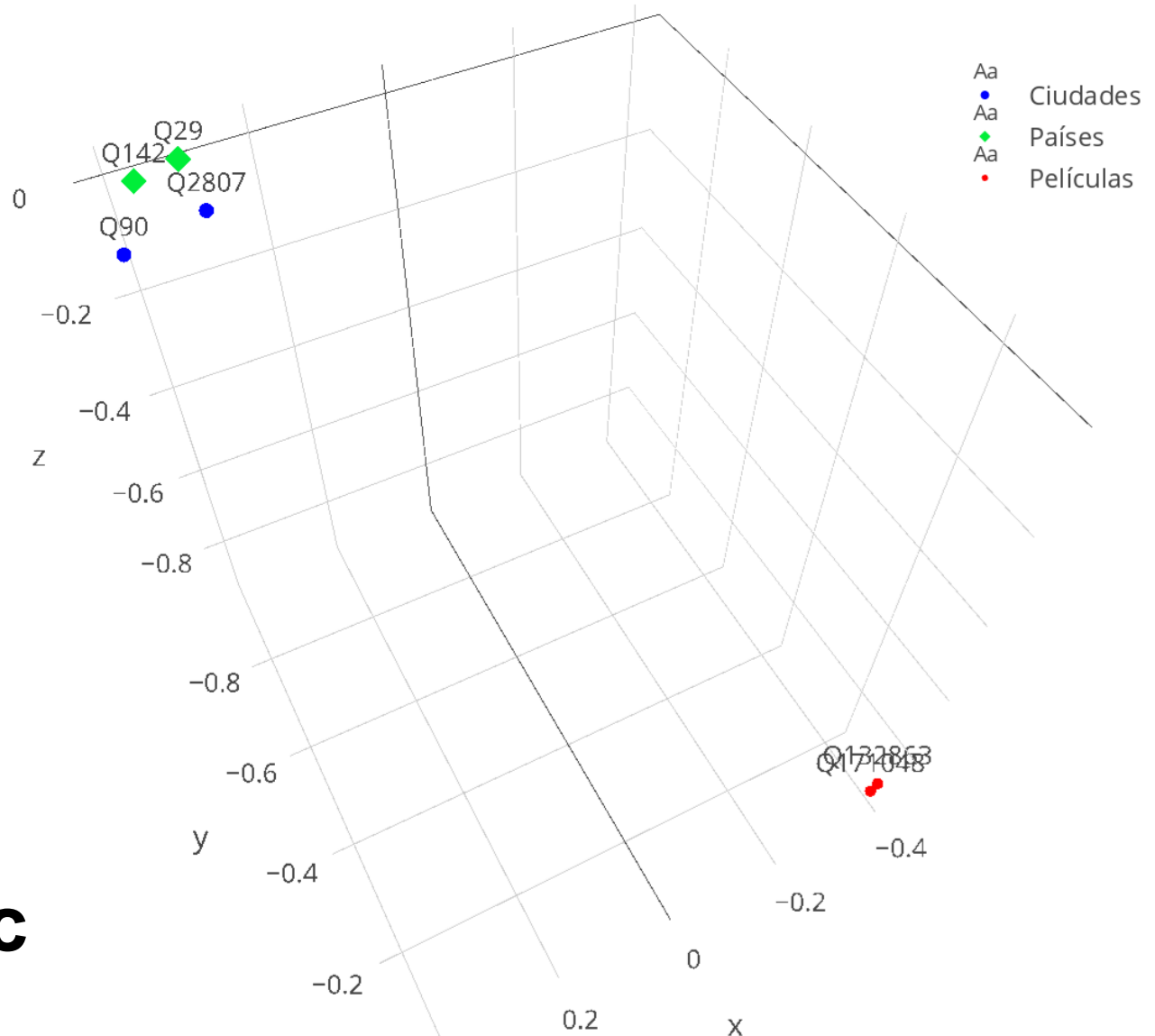
Madrid

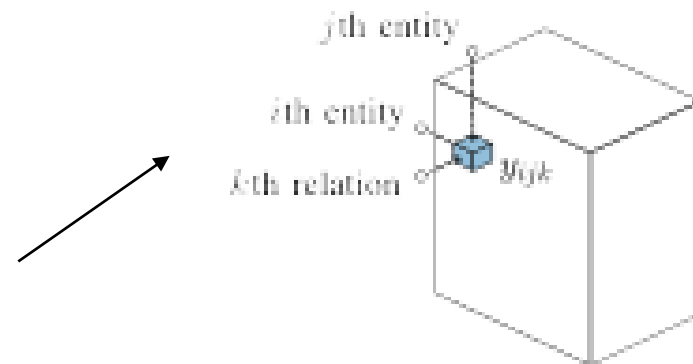
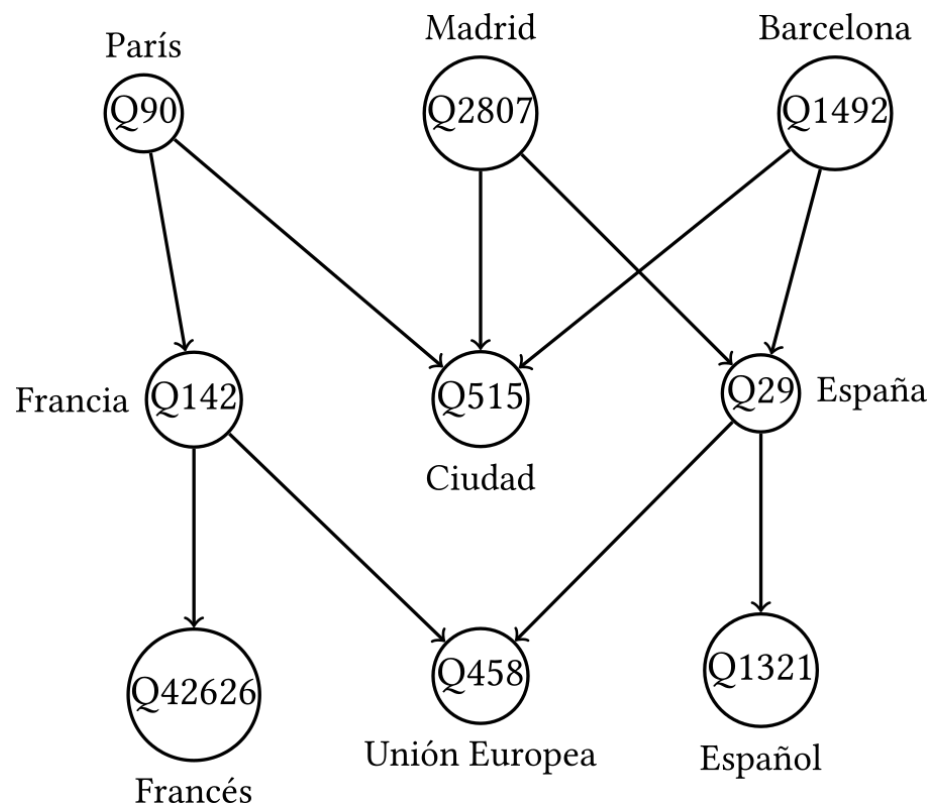


$M = [0.73 \quad \dots \quad 0.28]$

$M \in \mathbb{R}^n$

Word2vec



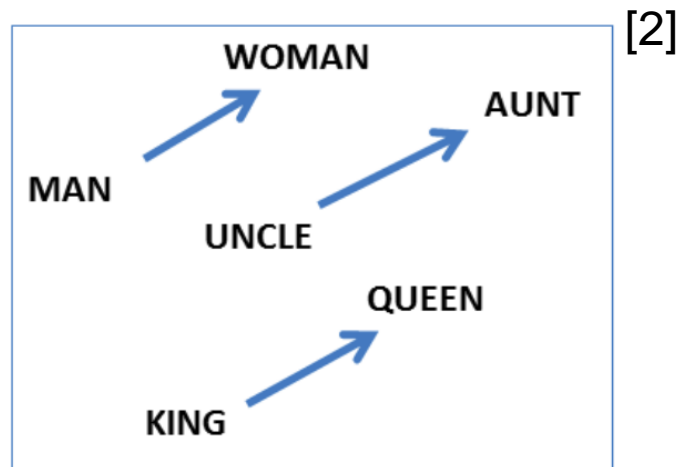


TransE [1]

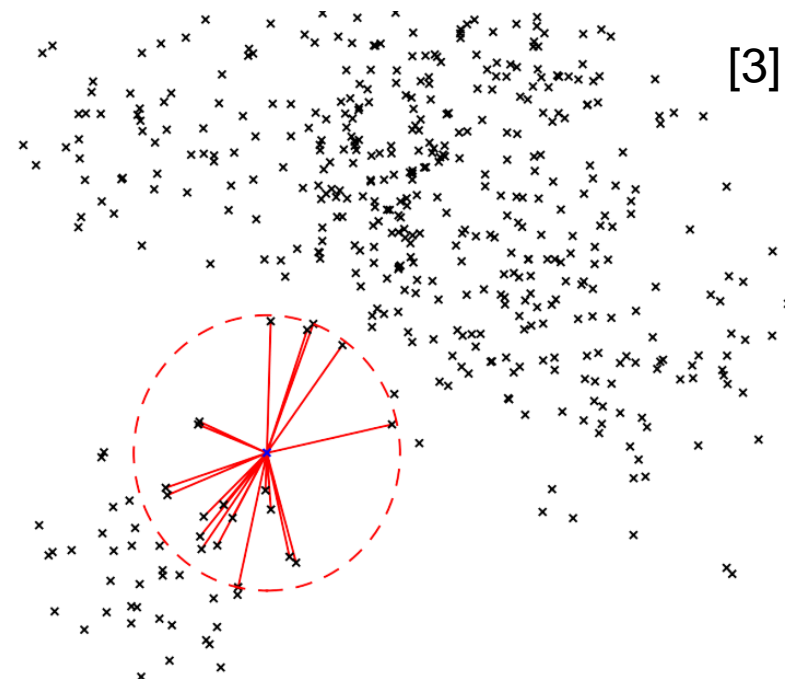
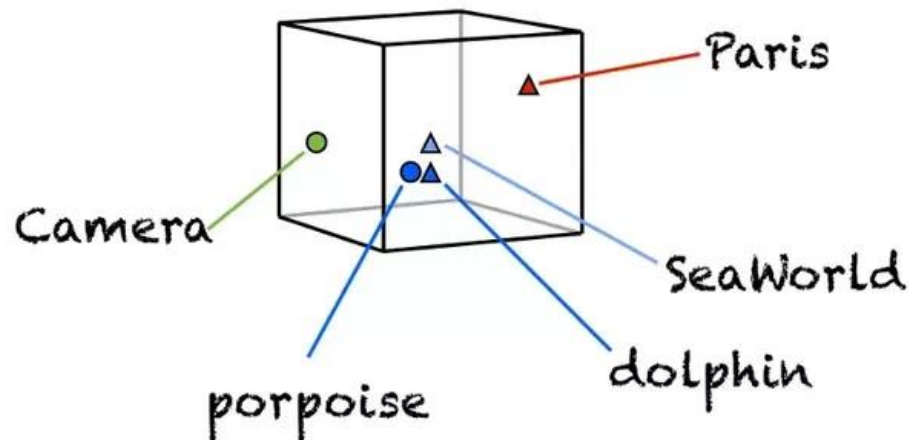
```
[ 0.73 ... 0.28 ... 0.65 ;
  0.48 ... 0.92 ... 0.42 ;
  ... ]
```

[1] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston y O. Yakhnenko, «Translating embeddings for modeling multi-relational data», en Advances in Neural Information Processing Systems 26.

Properties and applications of *Embeddings*



$$\text{Madrid} \approx E_{\text{Paris}} - E_{\text{France}} + E_{\text{Spain}}$$



[2] Linguistic Regularities in Continuous Space Word Representations – Mikolov et al. 2013

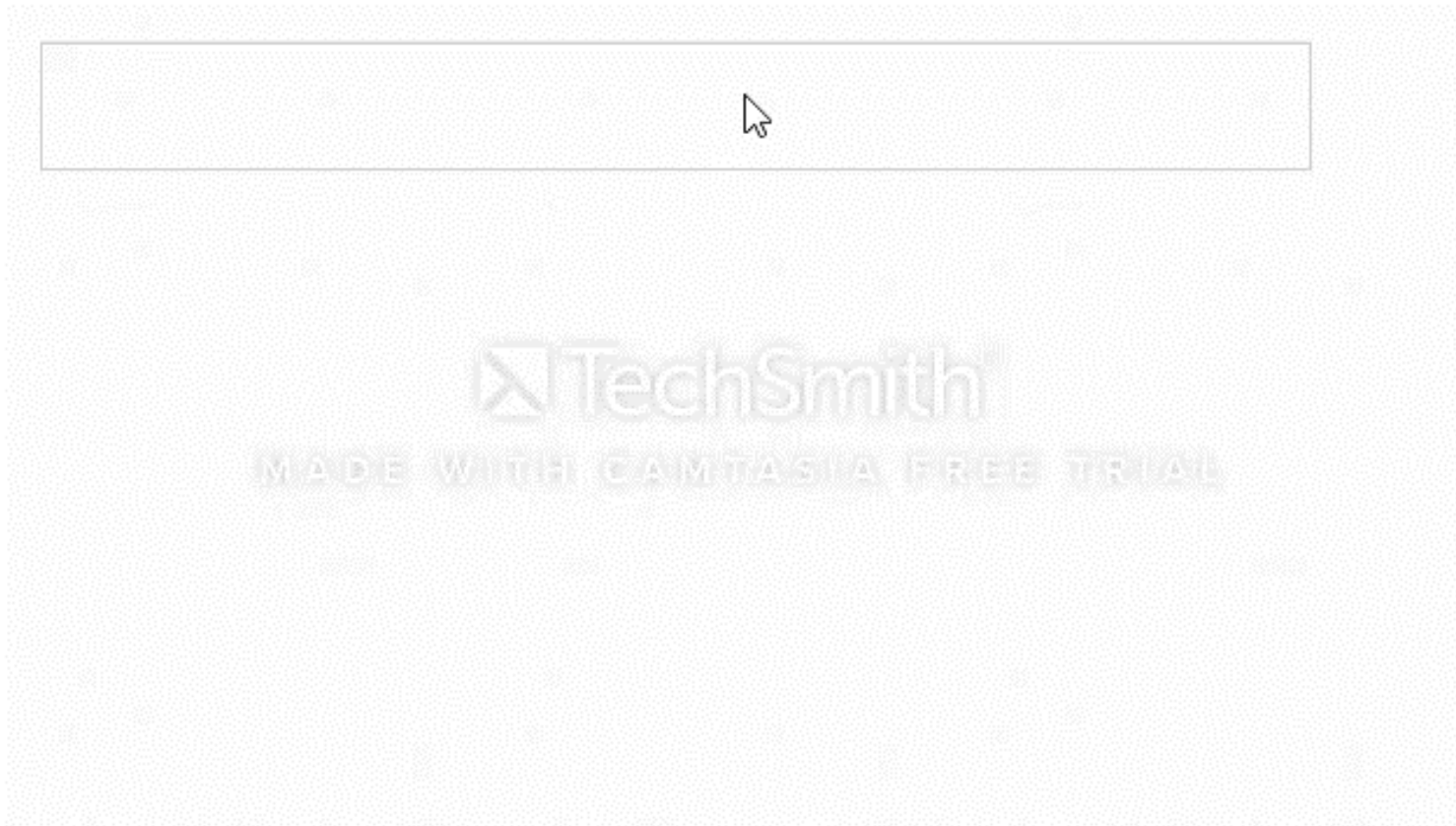
[3] Nearest neighbor methods and vector models - Erik Bernhardsson – Annoy (Spotify)



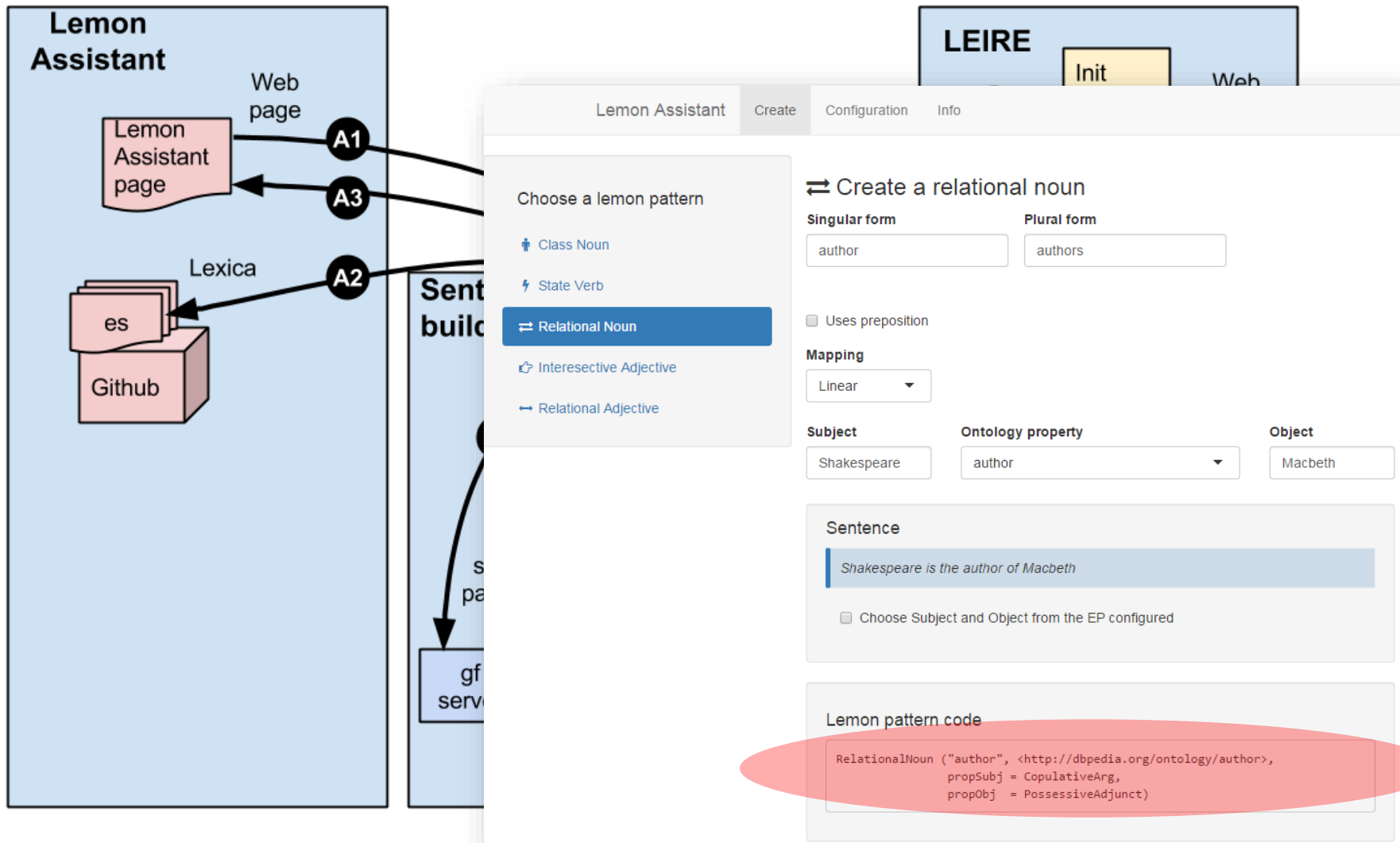
NL-guided queries

Mariano Rico

- An animation is worth 1K words



■ Lemonade: Lexicalizing ontologies



■ Pipeline

Customer

Knowledge
(vocabularies,
semantics of data)



Data
(SQL, non-SQL,
text files)



Structuration

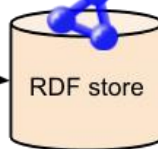
Ontology
specification



Ontology &
Vocabulary &
languages

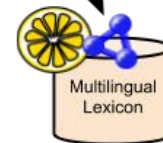
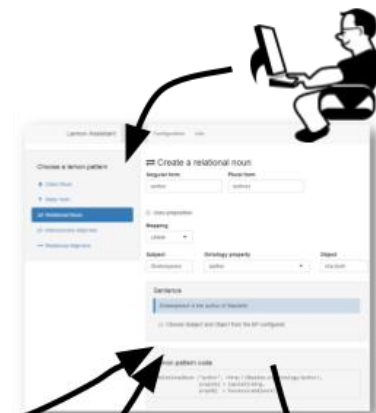


RDF data
generation



RDF store

Lexicalization



Multilingual
Lexicon

Search engine

