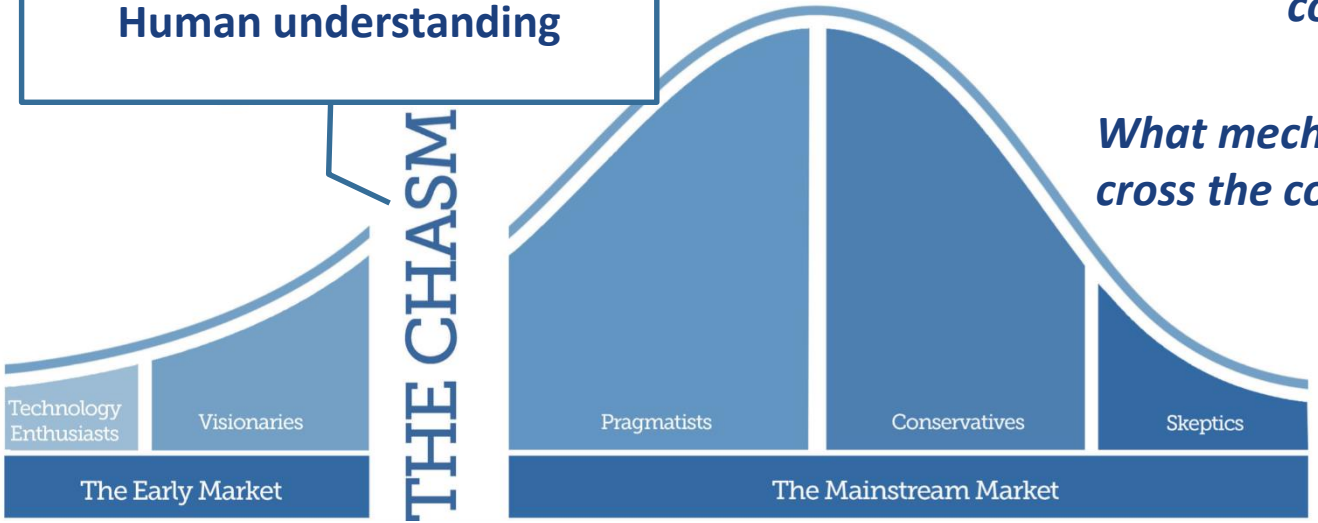
Towards a Vecsigrafo
Portable Semantics in Knowledge-based Text Analytics

Ronald Denaux & José Manuel Gómez Pérez

HSSUES– Oct. 21st, 2017

The Cognitive Chasm

Machine understanding vs. Human understanding



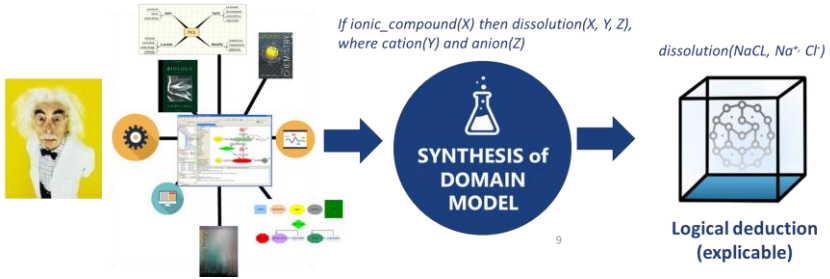
How can humans and AI interact with and understand each other?

Is this possible or are they cognitively disconnected?

What mechanisms are needed to cross the cognitive chasm?

How can knowledge representation be both flexible, scalable, deep and logical?

Pros and cons of structured knowledge



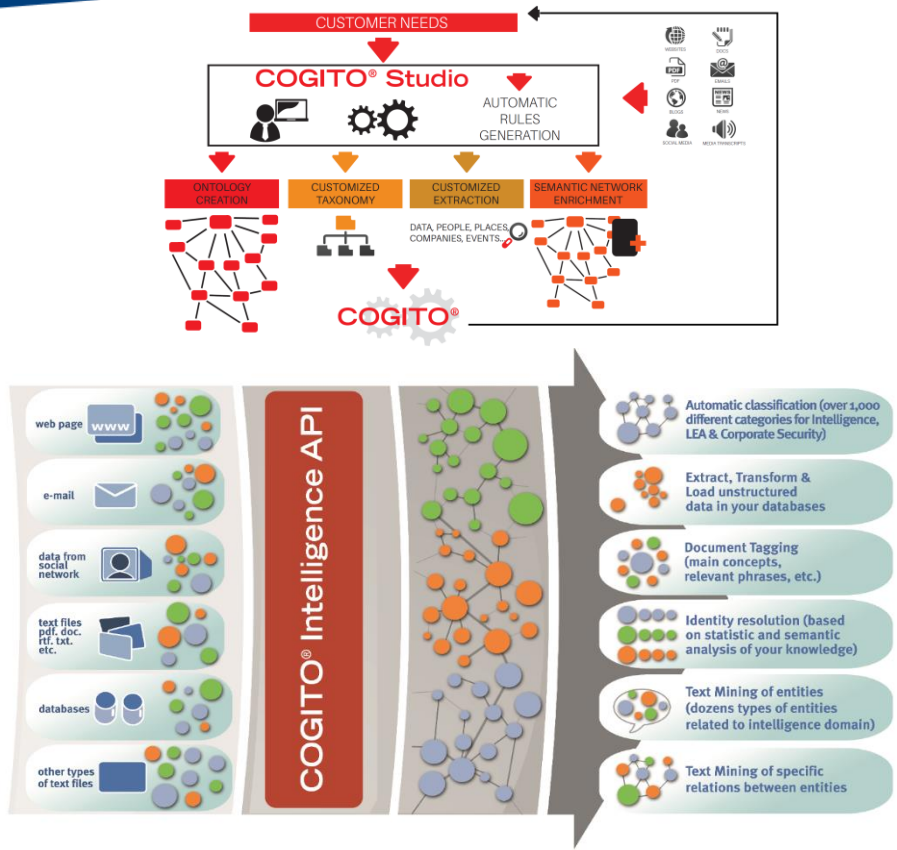
PROS

- Humans have a rich understanding of the domain, resulting in detailed, expressive models
- Underlying formalisms support logical explanations
- Reasonable response times
- Tooling can optimize cost, enabling user-entered knowledge

CONS

- Requires a considerable amount of well trained, centralized labor to manually encode knowledge
- Lacks scalability with large corpora and still costly due to humans in the loop
- Possible bias, hard to generalize
- Brittleness

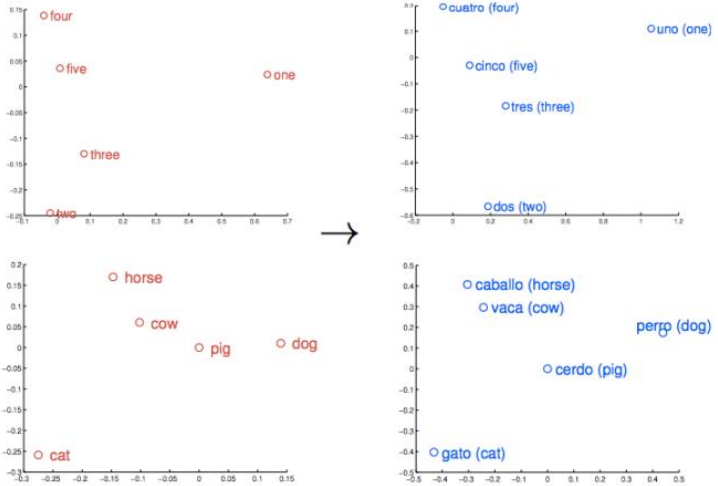
Structured knowledge (Sensigrafo)

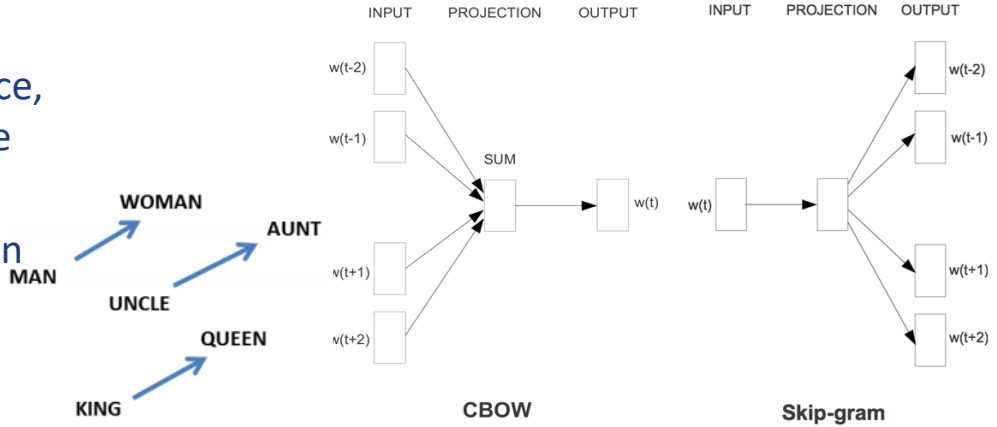


- **Sensigrafo, a knowledge graph** containing word definitions, related concepts and linguistic information
- Main entities include **syncons** (concepts), **lemmas** (canonical representation of a word) and **relations** (properties, taxonomical, polysemy, synonymy…)
  - 301,582 syncons
  - 401,028 lemmas
  - 80+ relation types that yield ~2.8 million links
- **Internal representation** that leverages external resources, both general and domain-specific
- **Word-sense disambiguation**, based on the context of a word in Sensigrafo
- **Categorization** and **extraction** supported through **Sensigrafo plus lexical-syntactic rules**

Building multiple language models

- **Word2vec** represents words in a vector space, making natural language computer-readable
- Neural word embeddings enable **word similarity, analogy and relatedness** based on vector arithmetic (cosine similarity)
- **Essential property: Semantic portability**



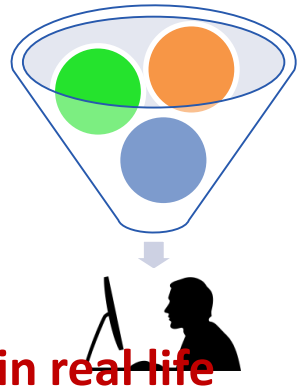


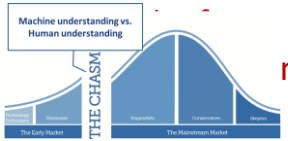
Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Towards Natural Learning at Expert System

- Knowledge embedded in document corpora
- Broad, flexible, scalable
- Good for POS tagging, parsing, semantic relatedness
- Statistic induction, not logical explanation

- Knowledge encoded in the mind of the expert
- Structured knowledge base
- Good for logical deduction and explanation
- Deep, but rigid and brittle
- Human is a bottleneck: hand-engineered features and powerful modeling tools needed





understanding of lexical semantics and pragmatics

Automatically learning how language is used in real life and materializing that in structured knowledge graphs

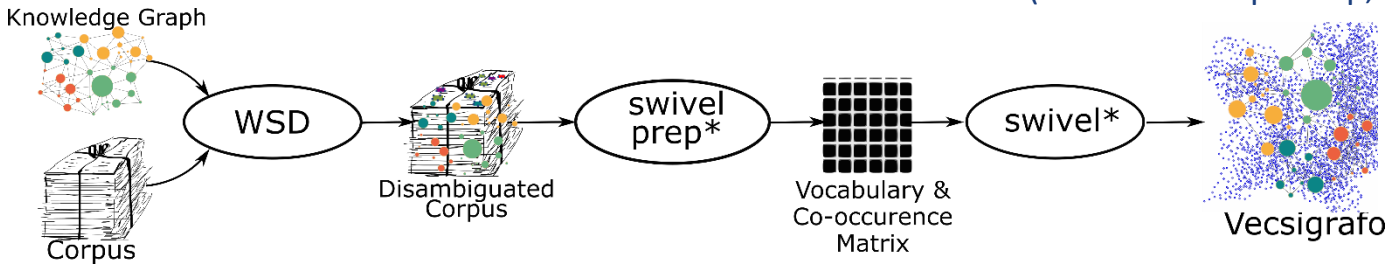
EXPERT SYSTEM
SEMANTIC INTELLIGENCE

Vecsigrafo – Putting it all together

Vocab elements	EN-grafo		ES-grafo	
	Sensi	Vecsi	Sensi	Vecsi
Lemmas	398	80	268	91
Concepts	300	67	226	52
Total	**698**	**147**	**474**	**143**

Corpus	Sentences	Spanish words	English words
Euparl	1,965,734	51,575,748	49,093,806
UN.en-es	**21,911,121**	**678,778,068**	**590,672,799**

- **Two parallel corpora**, focused on English and Spanish (**Europarl** and **UN**)
- **Meaning extracted from corpora and related to Sensigrafo** (**21%** and **30%** Sensigrafo covered, resp.)
- **Tokenized, lemmatized and disambiguated** with COGITO
- Learned **monolingual joint word-concept models** and a **(non-linear) transformation** between vector spaces for crosslinguality
- **Deeplearning4j** with Skip-gram, minFreq 10, vector dimensionality 400
- **TensorFlow** and **Swivel** for better vectorization time (~16x & ~20x speedup, 80 epochs)



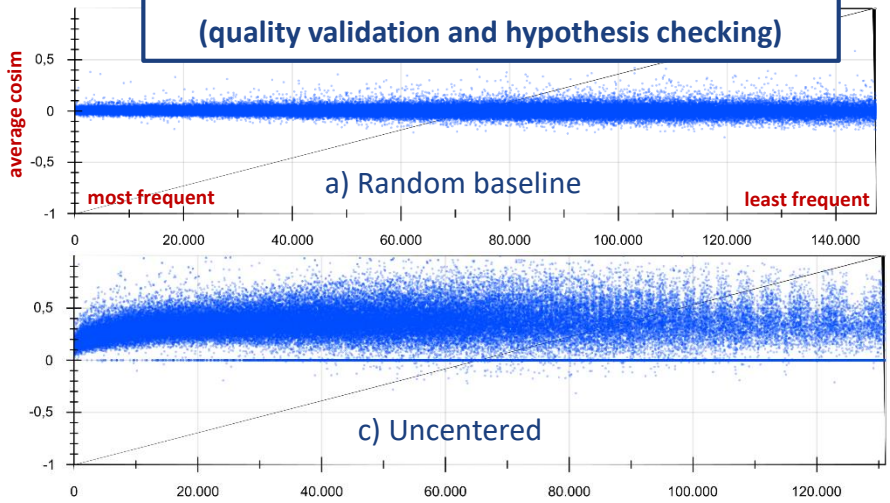
Vecsigrafo - Evaluation

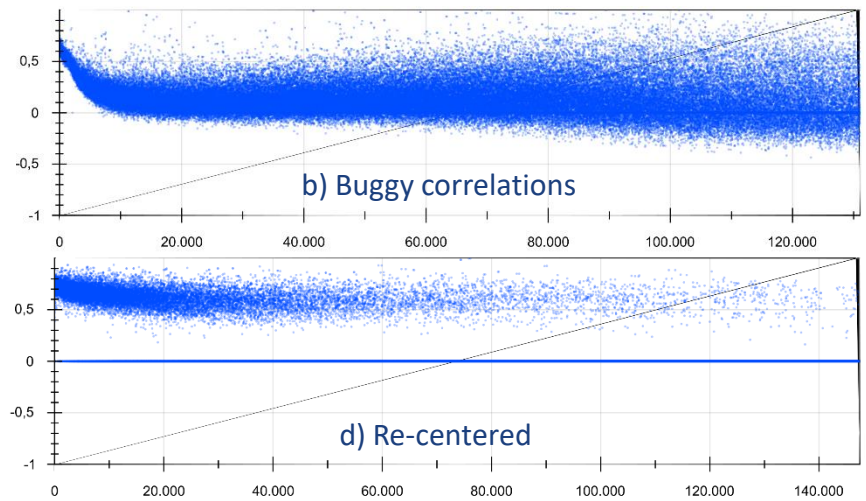
Model	WSim	WSrel	Simlex999	Rarewords	Simverb
SotA 2015	79.4	70.6	43.3	50.8	n/a
Swivel	74.8	61.6	40.3	48.3	62.8
Swivel$_{UN, en}$	**58.8**	**45.0**	**18.3**	**37.8**	**15.3**
~~**Vecisgrafo$_{UN,en}$**~~	~~**47.6**~~	~~**24.1**~~	~~**12.4**~~	~~**30.8**~~	~~**13.2**~~

- Corpus **size and distribution** matters
- ~~**Overall performance equivalent at lemma level** (Swivel, same corpus)~~
- ~~**Including concepts has a cost**~~
- **Visual inspection** (t-SNE, PCA) and **manual** (relatedness, analogy…)
- **Further insight needed**

Word Prediction Plots

(quality validation and hypothesis checking)





Vecsigrafo – Word Similarity Redux

Model	WSim	WSrel	Simlex999	Rarewords	Simverb
SotA 2015	79.4	70.6	43.3	50.8	62.8
Swivel	74.8	61.6	40.3	48.3	n/a
Swivel$_{UN, en}$	**58.8**	45.0	18.3	37.8	15.3
Swivel$_{UN, en}$ recentered	57.7	**47.2**	**21.3**	**39.2**	**17.0**
~~Vecisgrafo$_{UN,en}$~~	~~47.6~~	~~24.1~~	~~12.4~~	~~30.8~~	~~13.2~~
Vecisgrafo$_{UN,en}$	**69.9**	**51.6**	38.2	**50.3***	**30.6**
Vecisgrafo$_{UN,en}$ recentered	59.3	43.0	**42.4**	49.3	30.4
Vecisgrafo$_{UN,en}$ NN aligned to es	65.8	45.3	39.2	49.3	28.5

- Better than swivel for same corpus
- Effect of **recentering**
- Effect of **aligning to Spanish**
- **Further insight needed**
  - **How similar are two vecsigrafos?**
  - **Which relations are inferred?**
  - **How are relations encoded in the embedding space?**

Vecsigrafo – Application Roadmap

Map individual Vecsigrafos

Correlate and identify modeling gaps in Sensigrafos

Suggest crosslingual synonyms

Crosslinguality

Assisted Sensigrafo Learning

Fast internationalization at Expert System (EU, US, LATAM) and growing customer needs in 14 languages

Mapping and correlation

- **Mapping vector spaces in different languages:** Linear transformation suggested by (Mikolov, 2013) produced poor results. **Non-linear** transformation using NNs: **hit@5 = 0.78 and 90% semantic relatedness**
- Manual inspection showed **only 28% exact correspondence EN→ES**, due to volume (75K concepts less in Spanish Sensigrafo) and strategic modeling decisions
- **How to address the gap?**

Alignment performance

Method	Nodes	hit@5
TM	n/a	0.36
NN2	4K	0.61
NN2	5K	0.68
NN2	**10K**	**0.78**
NN3	5K	0.72

Manual inspection EN→ES

	in dict.	out dict.
#concepts	46	64
hit@5	**0.72**	**0.28**
no $concept_{ES}$	2	33

Examples

"Financing" (EN→ES)

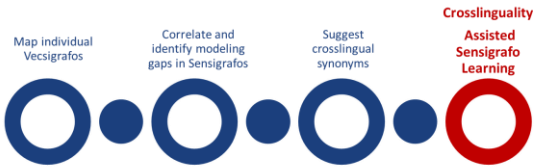
@	lemma/syncon	cosim	comment
1	financing	0.96	lemma
2	finance	0.85	lemma
3	funding	0.80	lemma
4	en#178501: adverb for financing	0.79	
5	en#75764: verb for fund, finance	0.76	
6	**es#126922: noun financiación,**	**0.75**	**synonym**

"Scrap value" (EN→ES)

@	lemma/syncon	cosim	comment
1	salvage value	0.92	lemma
2	scrap value	0.90	lemma
3	replacement cost	0.72	
4	en#57338: replacement cost	0.72	
30	en#195309: reduced price, sale	0.61	
??	precio de compra	0.48	
??	**es#20836: cambio, valor comercial**	**0.23**	

"PYME" (ES→EN)

@	lemma/syncon	cosim	comment
1	es#92662: pequeña y mediana empresa PYME	0.99	
2	**en#2739337:SME**	**0.81**	**synonym**
3	**sme**	**0.81**	**synonym**
4	mediana empresa	0.79	narrower
5	es#307734: mediana empresa	0.78	narrower

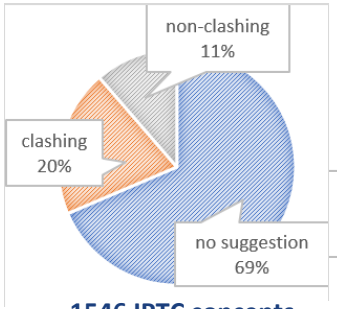


Crosslingual synonym suggester

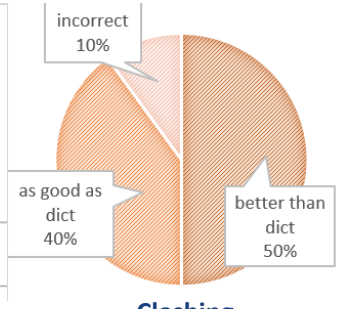
Combines features from bilingual vecsigrafo, the target and source Sensigrafos and a dictionary (PanLex)

1. For each concept in the source language, **find the *n* nearest concepts** in the target language that match grammar type (noun, verb, adjective, etc.)
2. For each candidate, **calculate hybrid features** (lemma translation, glossa similarity, cosine similarity, shared hypernyms and domains)
3. **Combine into a single score and rank**
4. Check if suggested synonym candidate is **already mapped** to a different concept and compare
5. Suggestion made if score is over a threshold

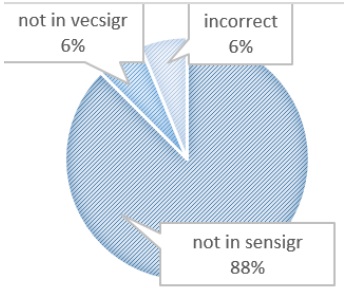
Manual inspection EN→ES (1546 concepts, IPTC)



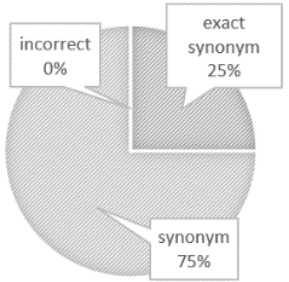
1546 IPTC concepts



Clashing

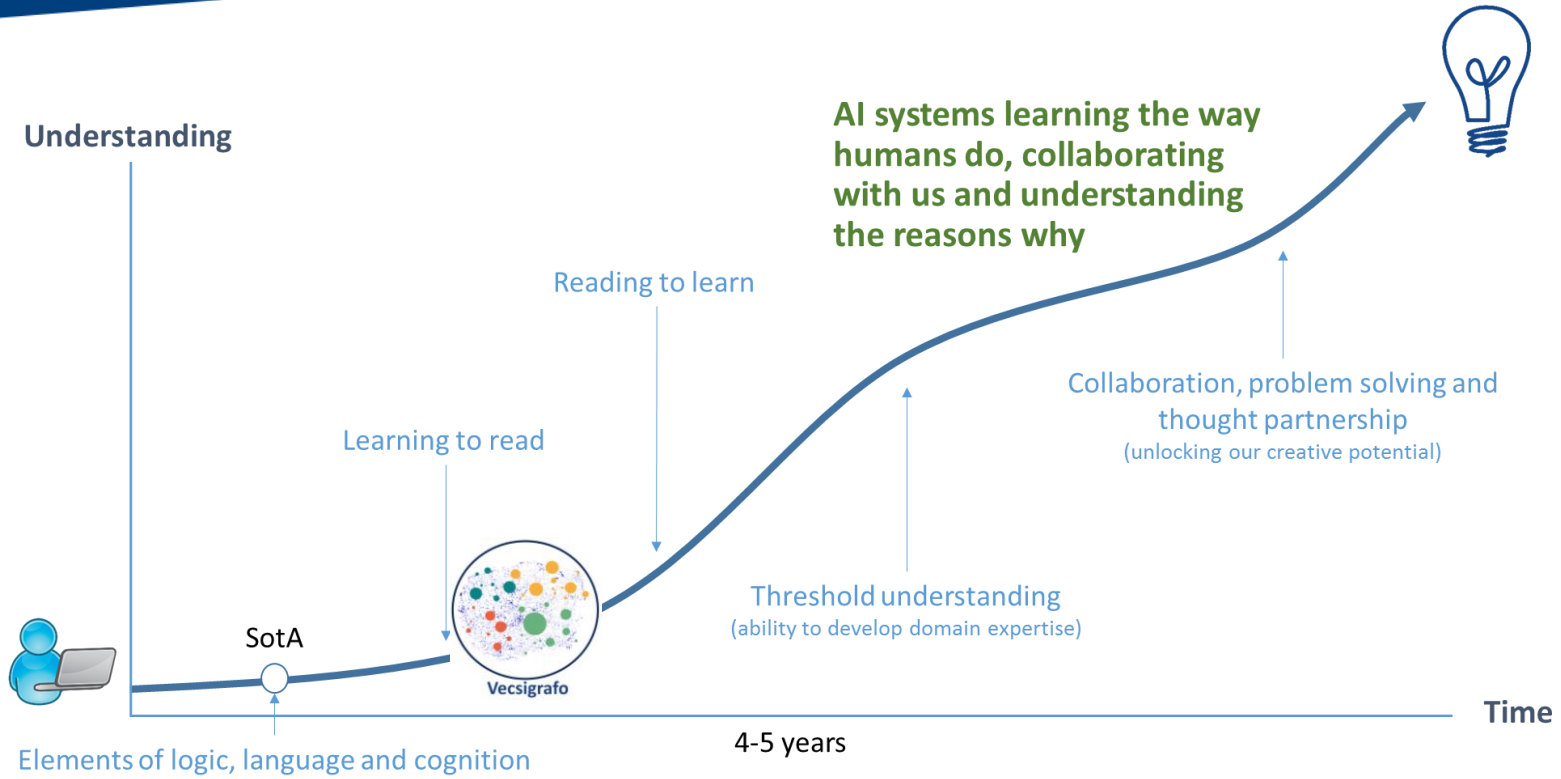


No suggestions



Non clashing

Wrapping up



From David Ferrucci, 2016

Ronald Denaux
Senior Researcher
rdenaux@expertsystem.com

Jose Manuel Gomez-Perez
Director R&D
jmgomez@expertsystem.com

Denaux R, Gomez-Perez JM. ***Towards a Vecsigrafo: Portable Semantics in Knowledge-based Text Analytics.*** *To appear in proceedings of the Intl. Workshop on Hybrid Statistical Semantic Understanding and Emerging Semantics (HSSUES), collocated with the 16th Intl. Semantic Web Conference (ISWC), Vienna, 2017.*


linkedin.com/company/expert-system

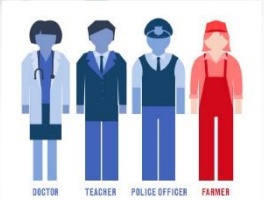

twitter.com/Expert_System


info@expertsystem.com


BANKING


Health
Life
Travel
Insurance
Business
House
Car

STOCK MARKETS


DOCTOR
TEACHER
POLICE OFFICER
FARMER

Correlation calculation

Develop an indicative list of advisory and conciliatory measures to encourage full compliance;

Tokenize & WSD

en#67083|develop en#89749|indicative en#113271|list en#88602|advisory en#85521|conciliatory en#33443|measure en#77189|encourage en#84127|full en#4941|compliance

Correlation for en_lem_list (window 2, harmonic weight)

token	Distance	weight
en#67083	2	½
develop	2	½
en#89749	1	1
indicative	1	1
en#113271	0	1

token	Distance	weight
list	0	1
en#88602	1	1
advisory	1	1
en#85521	2	½
conciliatory	2	½