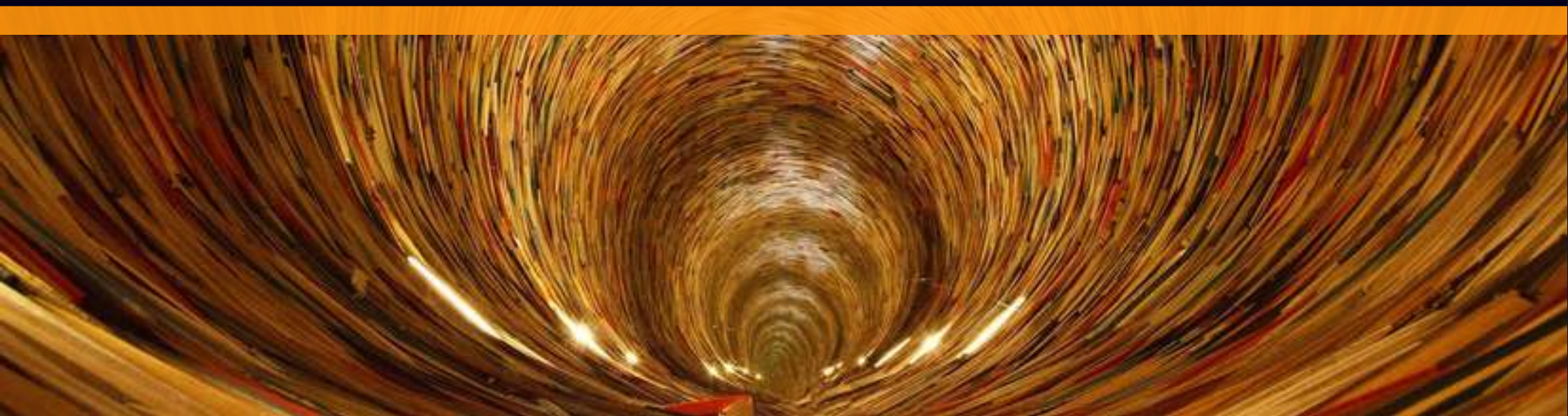


Neural Vector Representations beyond Words: Sentence and Document Embeddings

Gerard de Melo

<http://gerard.demelo.org>

Rutgers University



Outline

- **Word Representations**
- **Phrase Representations**
- **Sentence Representations**
- **Document Representations**
- **Applications and Outlook**

Phrases/Multiword Expressions

Kind	Example
Compound Noun	dog park
Adjective Noun	fresh food
Verb Object	win money
Verb Particle Phrase	walk up the stairs
Named Entities	East London
...	...

Compositional Models



Models for Composition

Composition Function

$$\mathbf{v}_p = f(\mathbf{v}_{w_1}, \dots, \mathbf{v}_{w_n})$$

Vector for phrase,
e.g. “mobile phone”

Vectors for parts of phrase
e.g. “mobile” and “phone”

Models for Composition: Elementwise Composition

$$f(\mathbf{v}_{w_1}, \dots, \mathbf{v}_{w_n}) = \mathbf{v}_{w_1} + \dots + \mathbf{v}_{w_n}$$

Vector 1		Vector 2		Result Vector
0.0		0.0		0.0
0.3		0.1		0.4
0.4	+	0.2	=	0.6
0.0		0.1		0.1
0.3		0.0		0.3
0.2		0.0		0.2

Models for Composition: Elementwise Composition

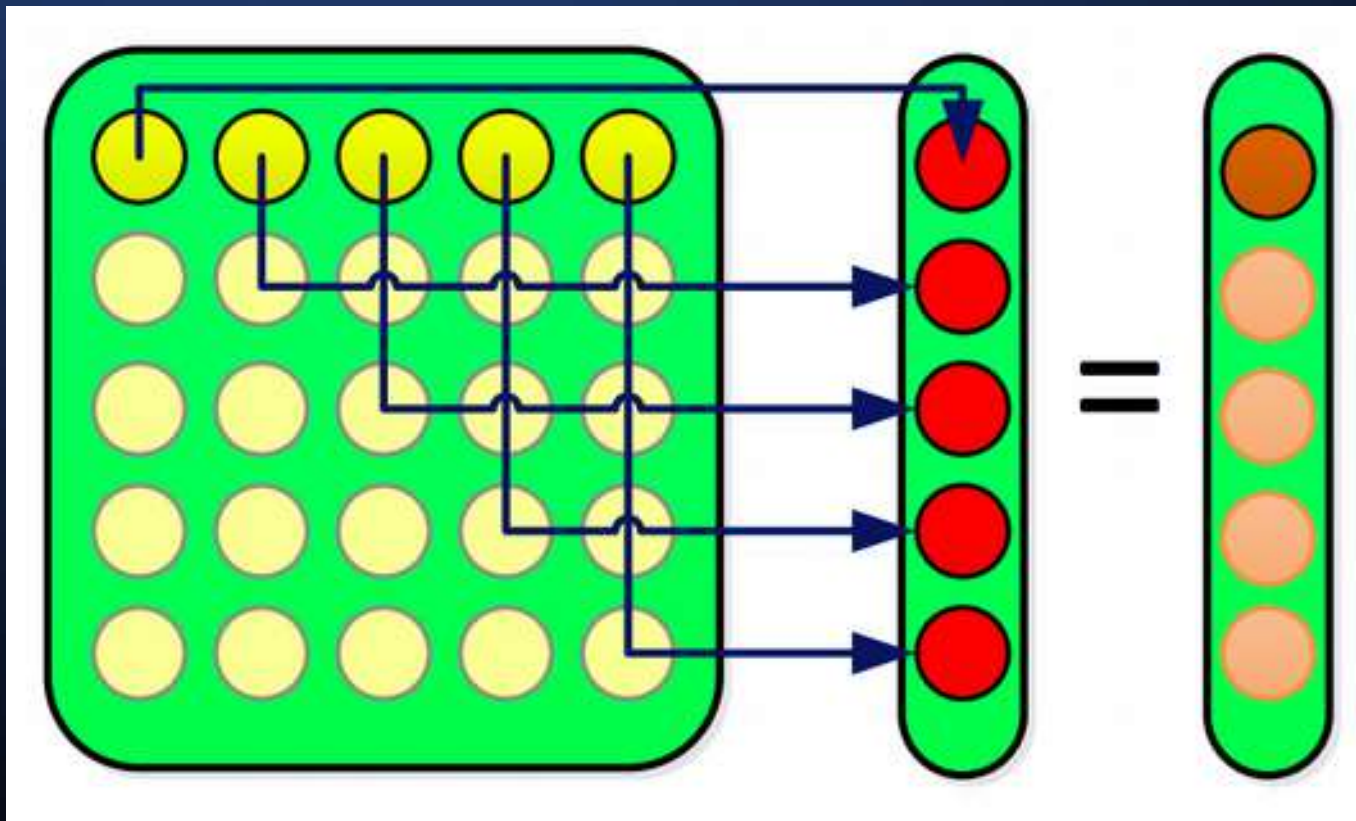
$$f(\mathbf{v}_{w_1}, \dots, \mathbf{v}_{w_n}) = \mathbf{v}_{w_1} \odot \dots \odot \mathbf{v}_{w_n}$$

Elementwise multiplication

Vector 1		Vector 2		Result Vector
0.0		0.0		0.00
0.3		0.1		0.03
0.4		0.2		0.08
0.0		0.1		0.00
0.3		0.0		0.00
0.2		0.0		0.00
	+		=	

Models for Composition: Tensor Models

Tensors that can be applied to one or more vectors for arguments



Phrases/Multiword Expressions

Kind	Example
Compound Noun	dog park
Adjective Noun	fresh food
Verb Object	win money
Verb Particle Phrase	walk up the stairs
Named Entities	East London
...	...

Phrases/Multiword Expressions

Kind	Compositional Example	Non-Compositional Example
Compound Noun	dog park	zebra crossing
Adjective Noun	fresh food	hot dog
Verb Object	win money	kick the bucket
Verb Particle Phrase	walk up the stairs	wrap up the session
Named Entities	East London	Los Angeles
...

Learning from Text



word2vec Implementation

1. PMI-like scores to find frequent MWEs

$$\text{score}(w_i, w_j) = \frac{\text{count}(w_i w_j) - \delta}{\text{count}(w_i) \times \text{count}(w_j)}$$

Bigram Count

Unigram Counts

The diagram illustrates the components of the PMI formula. A green arrow points from the label 'Bigram Count' to the term $\text{count}(w_i w_j)$ in the numerator. Two green arrows point from the label 'Unigram Counts' to the terms $\text{count}(w_i)$ and $\text{count}(w_j)$ in the denominator.

2. Select MWEs above threshold, then repeat to find even longer MWEs

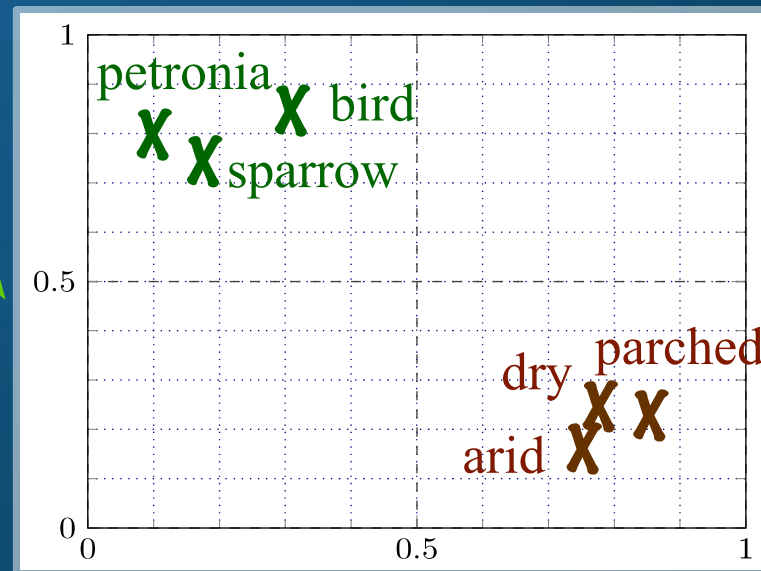
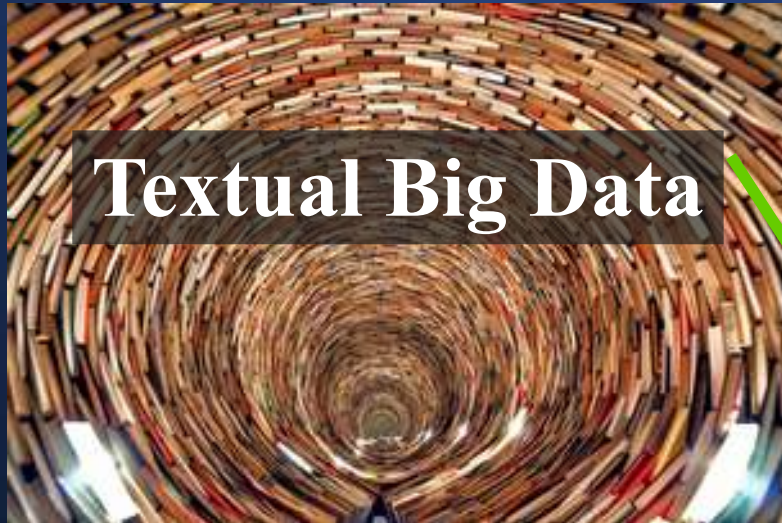
3. Finally, simply treat MWEs as unigrams

word2vec Implementation

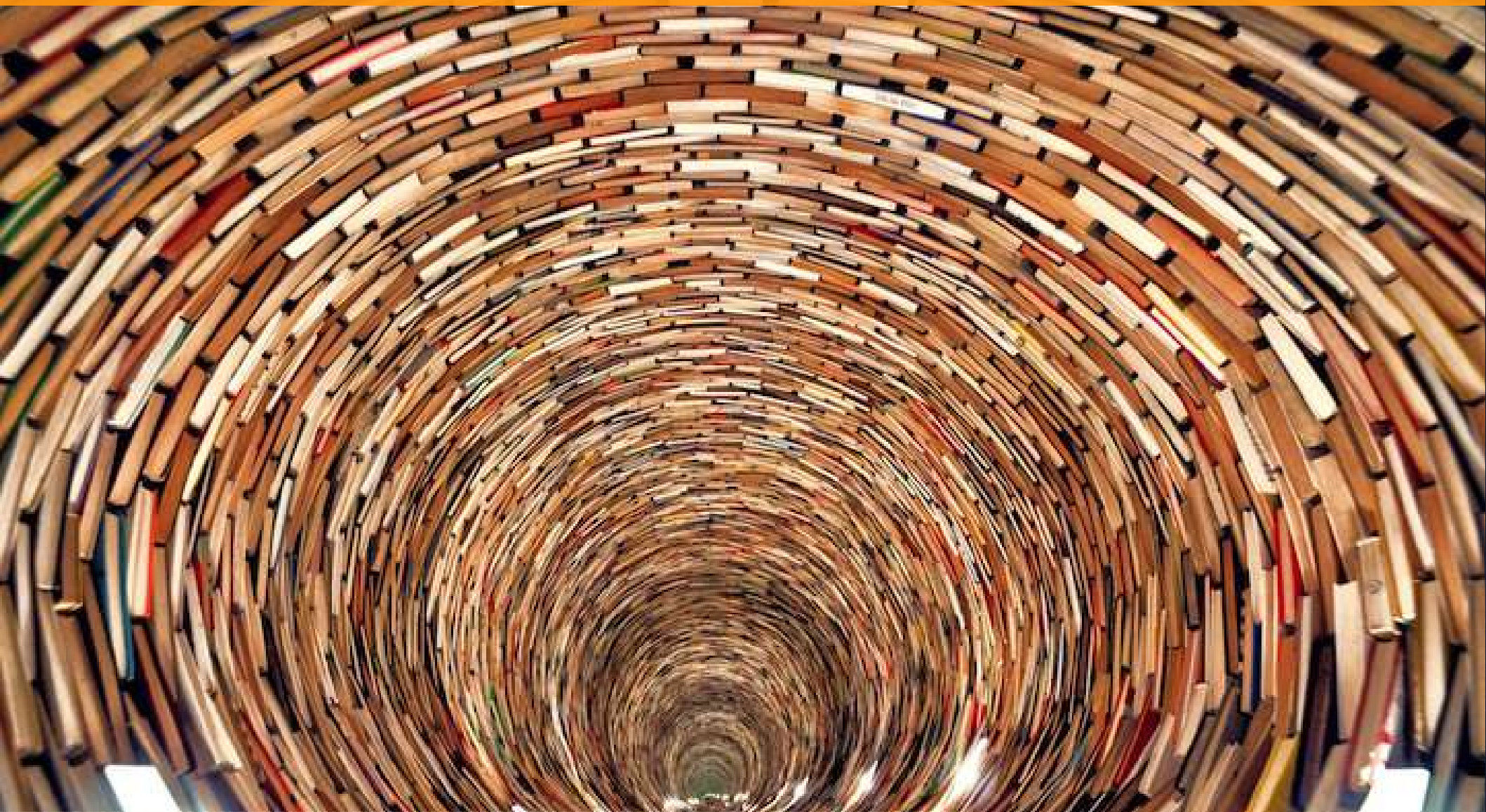
Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

Nearest neighbours of Component-wise sums
are meaningful

Learning from Heterogeneous Data



Textual Big Data



Matej Kren: Idiom. Prague Municipal Library

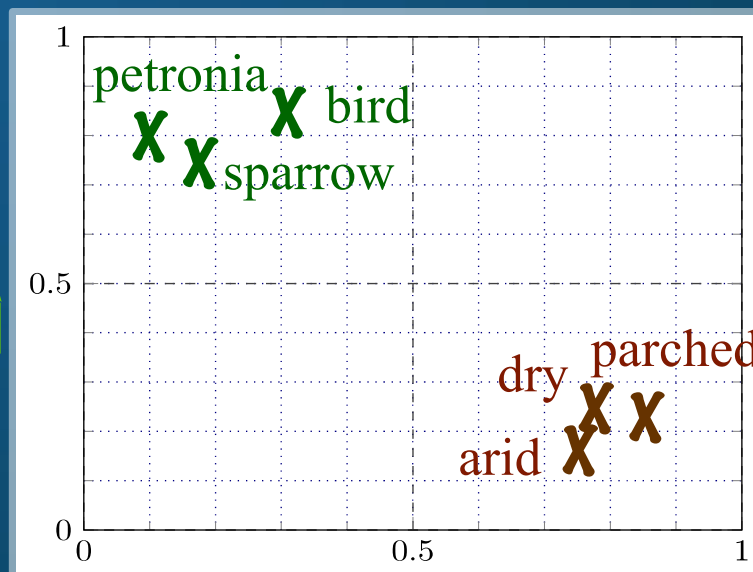
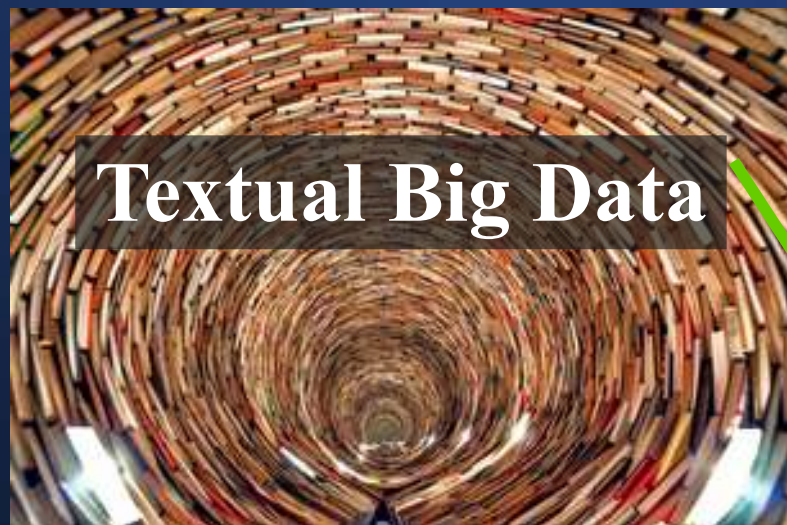
<https://www.flickr.com/photos/ill-padrino/6437837857/>

High-Quality Knowledge

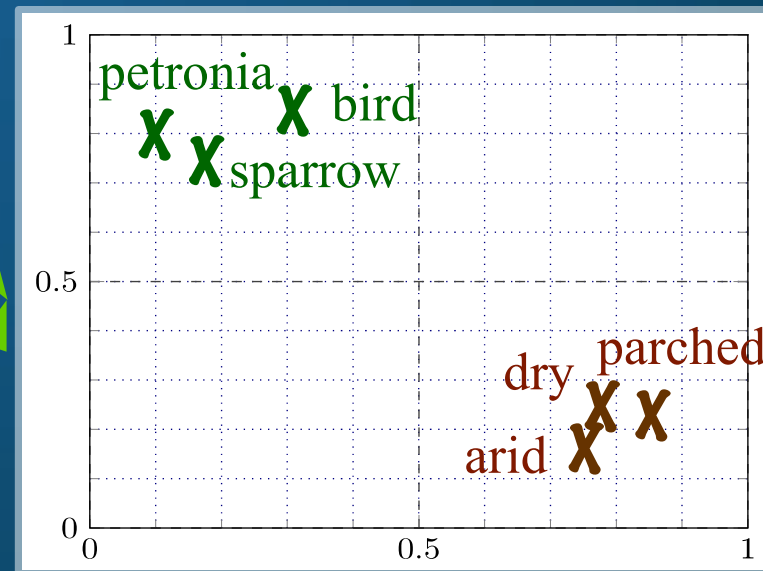
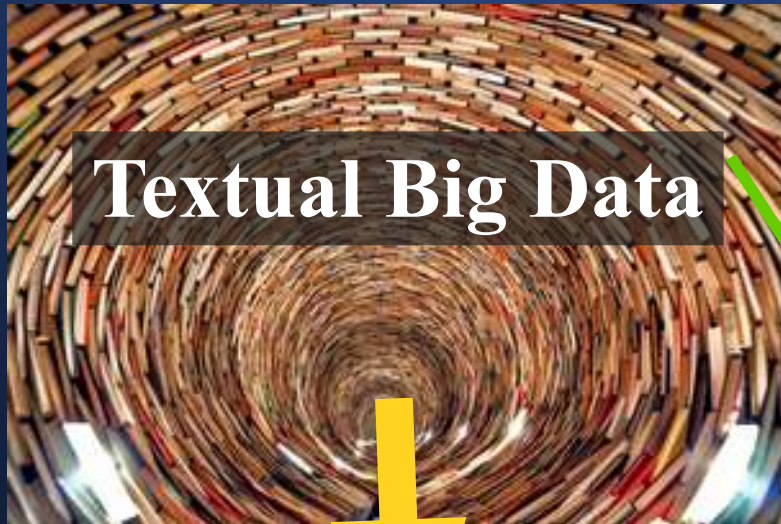


Theological Hall, Strahov Monastery Library, Prague

Learning from Heterogeneous Data



Information Extraction Approach



Information Extraction Approach



semantic!

...Greek and Roman mythology...

**look for semantically
salient contexts in text!**

Information Extraction Approach: Joint Training

Given an extracted pair of semantically related words, the intuition is that the embeddings for the two words should be pulled together

$$\max \frac{1}{T} \sum_{t=1}^T \sum_{w_r} \log p(w_r | w_t)$$

w_r : related words of w_t

Best Paper Award
at NAACL
Vector Space
Modeling Workshop

Information Extraction Approach: Joint Training

$$\max \frac{1}{T} \sum_{t=1}^T \sum_{w_r} \log p(w_r | w_t)$$

softmax function $p(w_r | w_t) = \frac{\exp(V'_{w_r} V_{w_t})}{\sum_{w=1}^W \exp(V'_w V_{w_t})}$

To compute its gradient is impractical. It is proportional to T , which is often large ($10^5 - 10^7$ terms)

Best Paper Award
at NAACL
Vector Space
Modeling Workshop

Case Study: Information Extraction

List Extraction

- Look for **repeated** occurrences of **commas**
- Short units of roughly equal length
- noun phrases, adjectives

Best Paper Award
at NAACL
Vector Space
Modeling Workshop

Case Study: Information Extraction

List Extraction

- Look for **repeated** occurrences of **commas**
- Short units of roughly equal length
- noun phrases, adjectives
- Also: Hearst patterns, e.g.
“cities **such as** New York, London, ...”

Best Paper Award
at NAACL
Vector Space
Modeling Workshop

Case Study: Information Extraction

Extracted Lists

player captain manager director vice-chairman

group race culture religion organisation person person

Italian Mexican Chinese Creole French

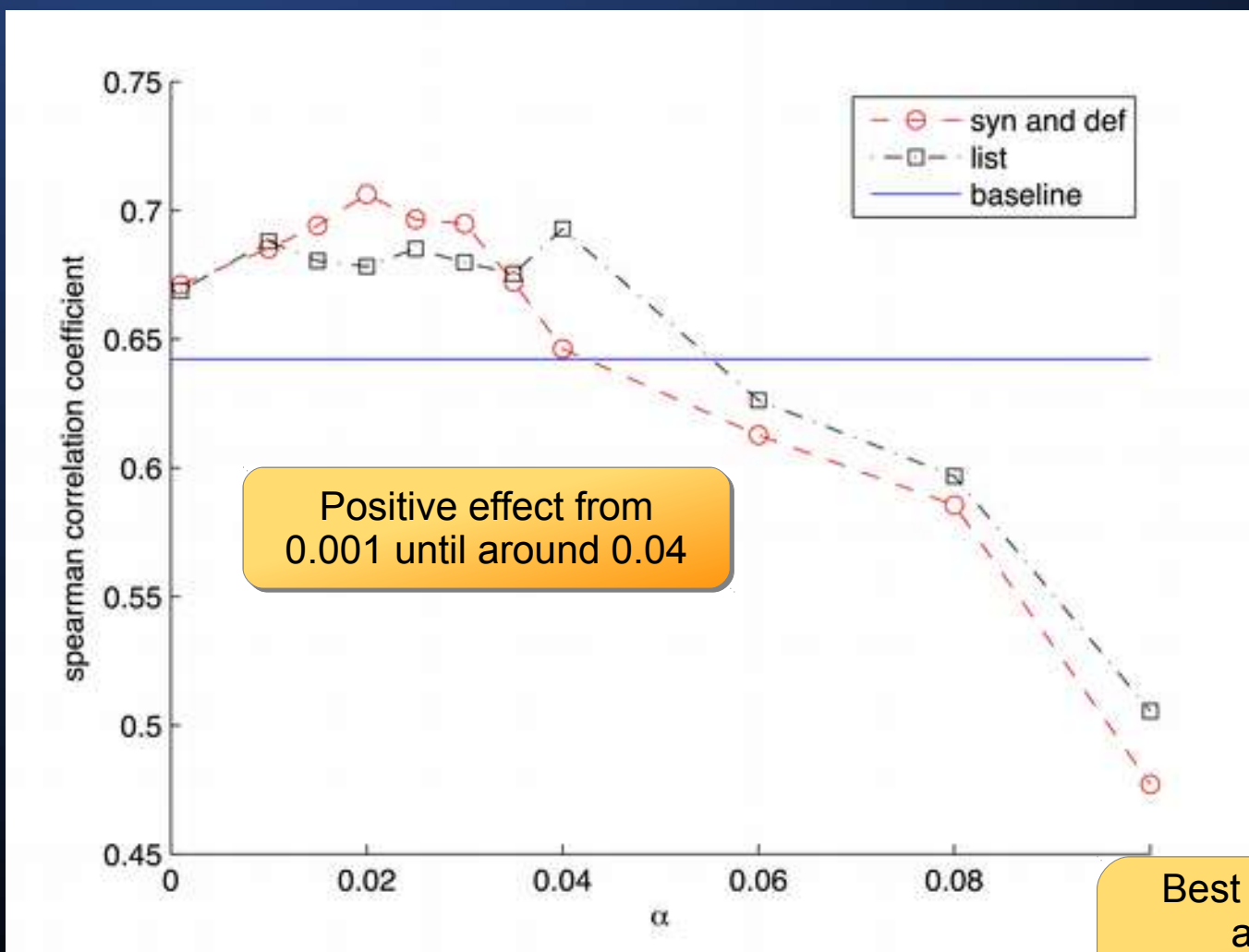
*Self-Portraits Portraits iris Still-Lives with Sunflowers view
from the Asylum Works after Millet Vineyards*

*ballscrews leadscrews worm gear screwjacks linear
actuator*

*Cleveland Essex Lincolnshire Northamptonshire
Nottinghamshire Thames Valley South Wales*

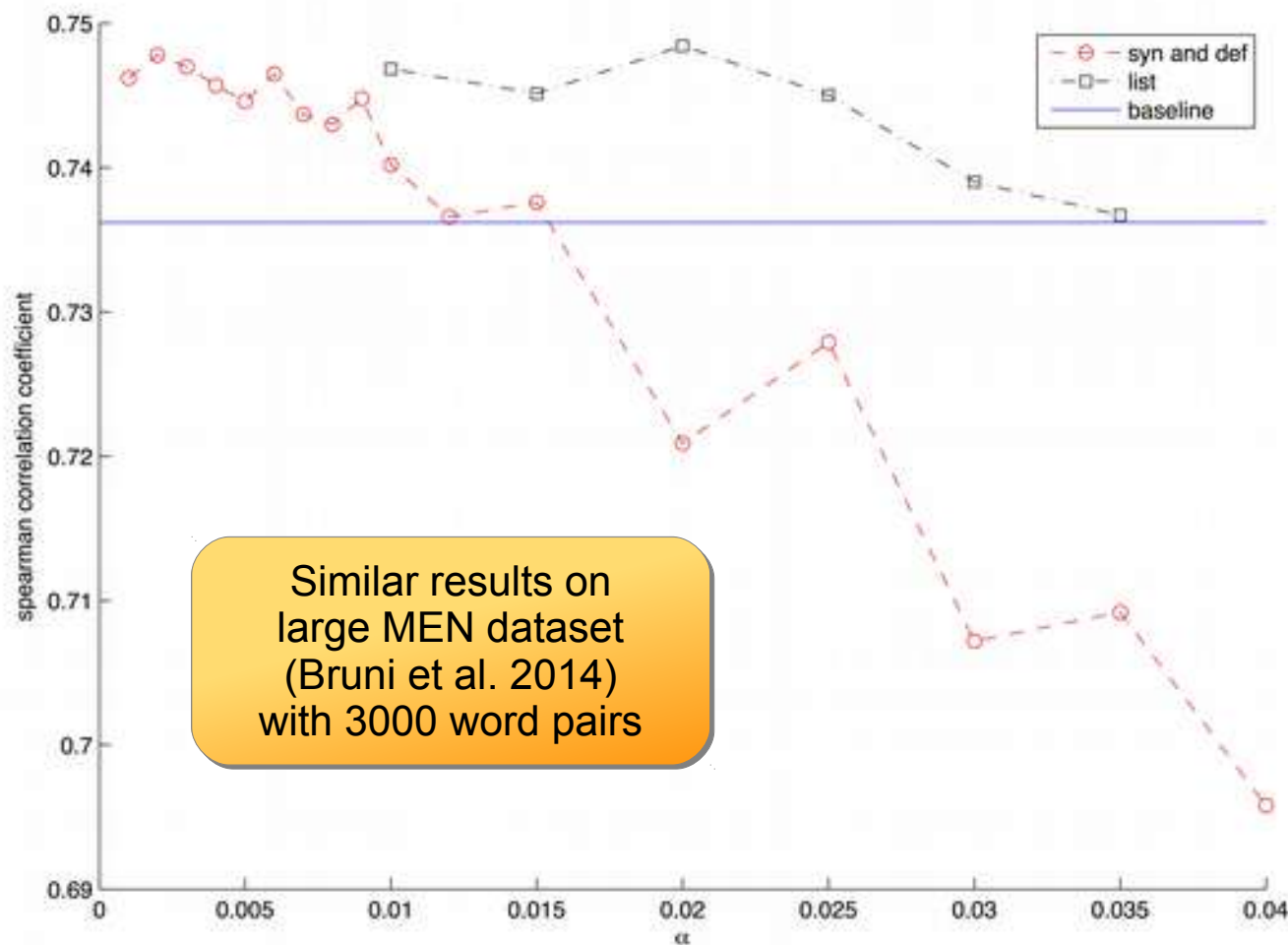
*ant.py dimdriver.py dimdriverdatafile.py
dimdriverdatasetdef.py dimexception.py dimmaker.py
dimoperators.py dimparser.py dimrex.py dimension.py*

Case Study: Results on WS353

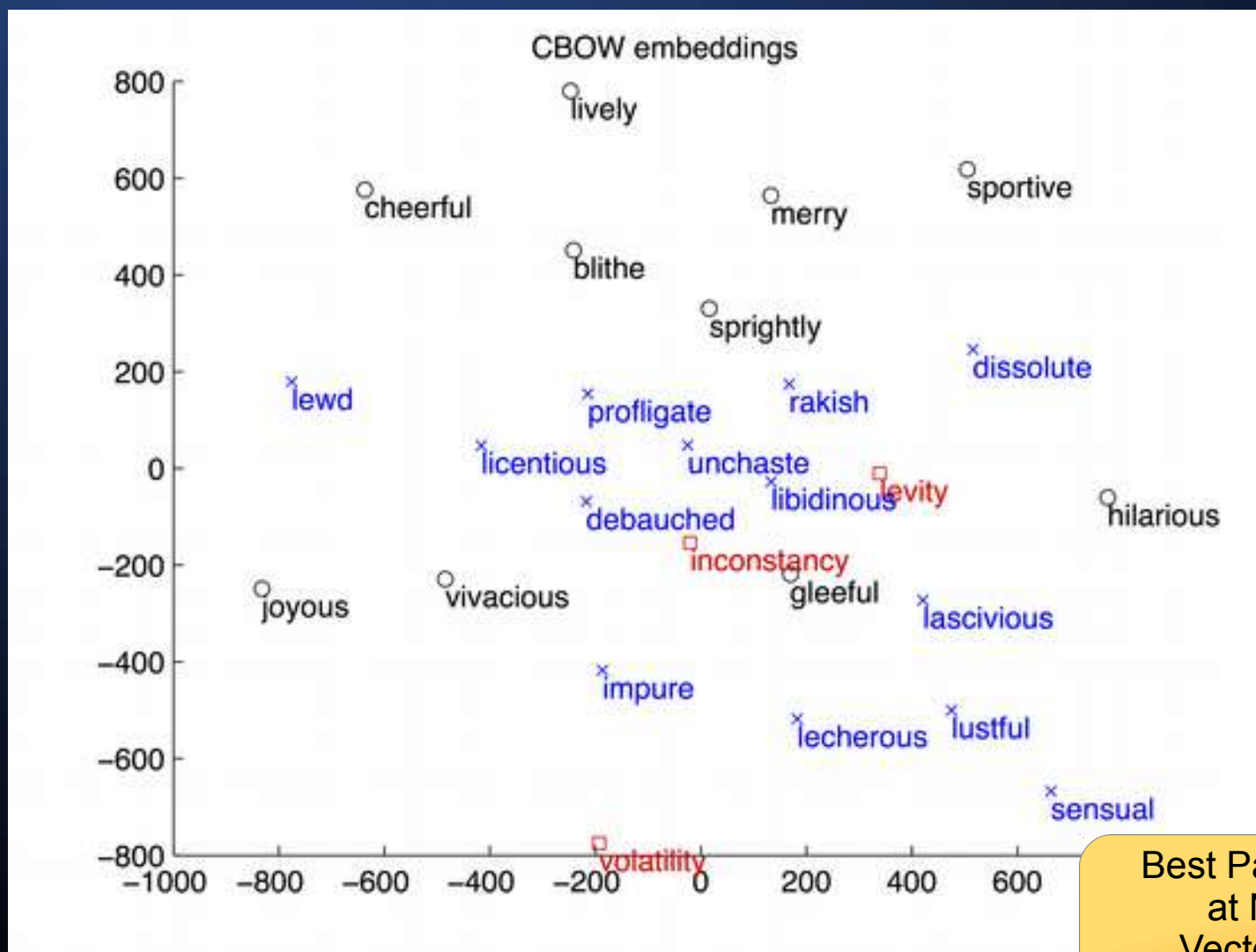


Best Paper Award
at NAACL
Vector Space
Modeling Workshop

Case Study: Results on MEN

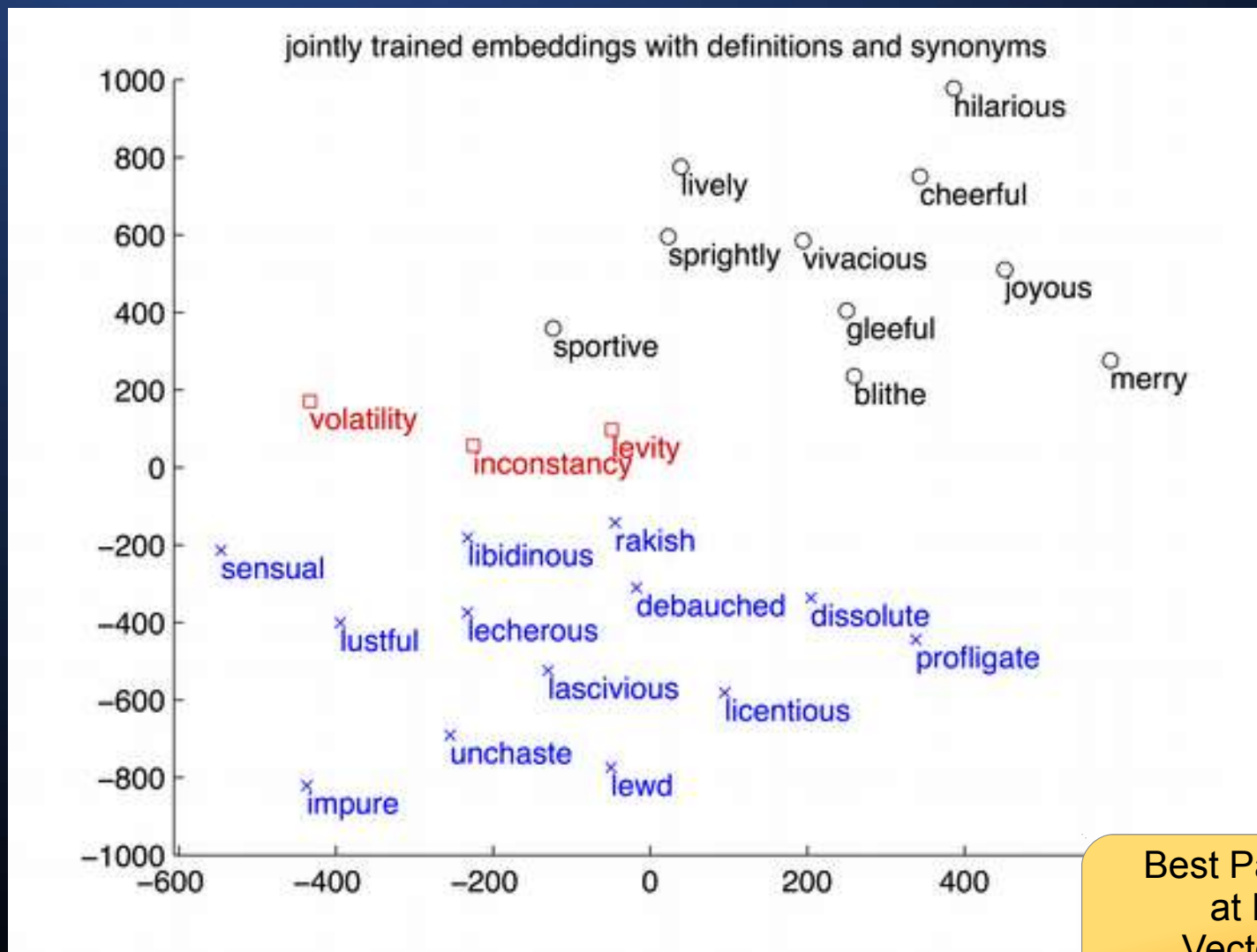


Case Study: Example



Best Paper Award
at NAACL
Vector Space
Modeling Workshop

Case Study: Example

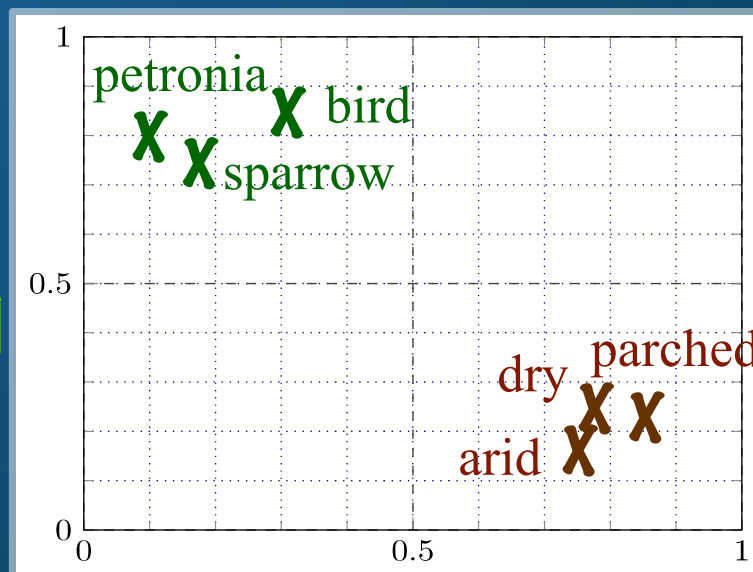
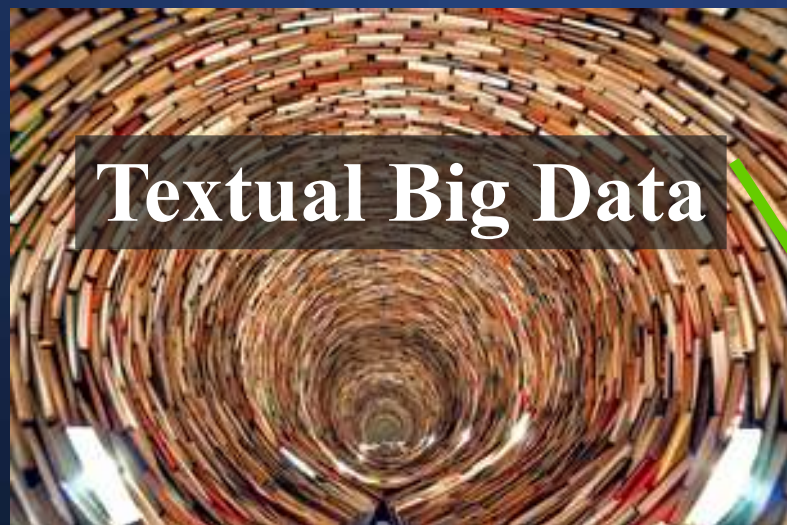


Best Paper Award
at NAACL
Vector Space
Modeling Workshop

Learning from Text and Structured Data



Learning from Heterogeneous Data



Lexical Knowledge



Portuguese-Chinese Dictionary by Ruggieri et al. (1580s)
The first European-Chinese dictionary

Lexical Knowledge: Wiktionary

Wiktionary

English
The free dictionary
5 884 000+ entries

Français
Le dictionnaire libre
3 383 000+ articles

Русский
Свободный словарь
997 000+ статей

Deutsch
Das freie Wörterbuch
734 000+ Einträge

Ελληνικά
Το ελεύθερο λεξικό
462 000+ λήμματα

Polski
Wolny słownik
644 000+ stron

Español
El diccionario libre
885 000+ entradas

日本語
フリー多機能辞典
214 000+ 項目

Italiano
Il dizionario libero
433 000+ lemmi

Português
O dicionário livre
249 000+ entradas

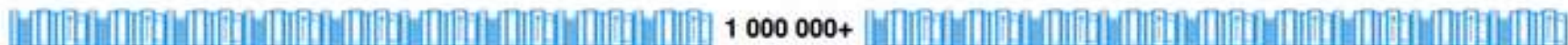


English

Q

Find a language: →

Provides translations,
antonyms, etc.



Lexical Knowledge: Wiktionary

a multilingual free encyclopedia

Wiktionary

['wɪkʃənɹɪ] *n.*,
a wiki-based Open
Content dictionary

Wilek ['wɪl kən]

[Main Page](#)
[Community portal](#)
[Preferences](#)
[Requested entries](#)
[Recent changes](#)
[Random entry](#)
[Help](#)
[Donations](#)
[Contact us](#)

Tools
[What links here](#)
[Related changes](#)

Entry [Discussion](#) [Citations](#)

déjeûna

French [\[edit\]](#)

Pronunciation [\[edit\]](#)

- Homophones: [déjeûnas](#), [déjeûnât](#)

Verb [\[edit\]](#)

déjeûna

- third-person singular past historic of [déjeuner](#)*

Categories: [French verb forms](#) | [French non-lemma forms](#)

Lexical Knowledge: Wiktionary

Etymology 1 [edit]

From Latin *pulsus* ("beat"), from *pellere* ("to drive"), from Proto-Indo-European **pel* ("to drive, strike, thrust").

For spelling, the -e (on -ise) is so the end is pronounced /ɪs/, rather than /ɪz/ as in *pulls*, and does not change the vowel ('u'). Compare *else*, *false*, *convulse*.

Pronunciation [edit]

• IPA^(key): /pʌls/

• Audio (US)  0:00 MENU

Noun [edit]

pulse (*plural* **pulses**)

1. (*physiology*) A normally regular beat felt when arteries are depressed, caused by the pumping action of the heart.
2. A beat or throb. (*quotations ▼*)
3. (*music*) The beat or *tactus* of a piece of music.
4. An *autosoliton*.

Related terms [edit]

- impulse
- repulse

Translations [edit]

regular beat caused by the heart

Select targeted languages

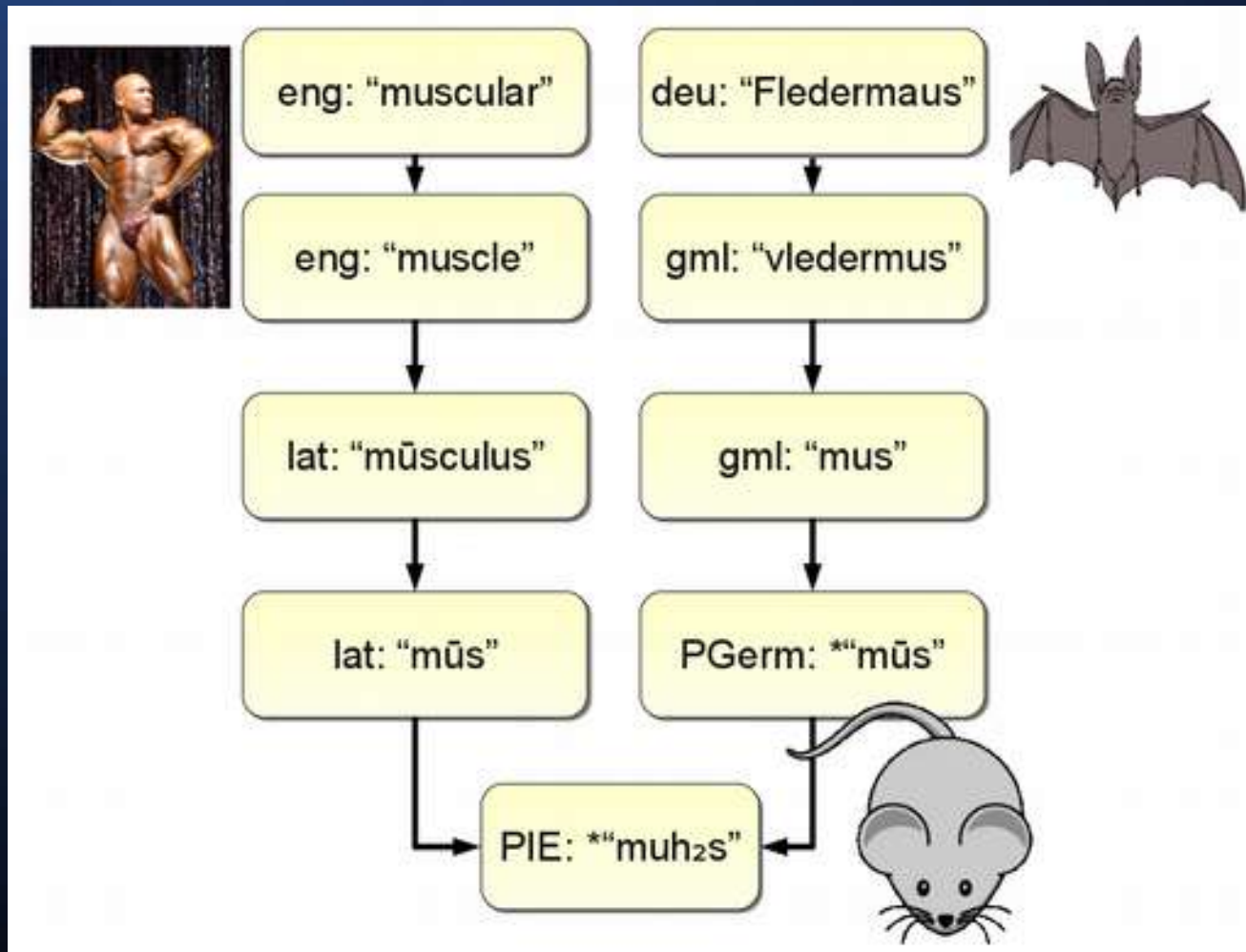
- | | |
|---|---|
| • Arabic: نبضة <i>f</i> (nábDa) | • Irish: cuisle <i>f</i> |
| • Chinese: | • Italian: polso <i>m</i> , battito ^(it) <i>m</i> |
| Mandarin: 脈搏, 脉搏 ^(zh) (máibó), 脈 ^(zh) , 脉 ^(zh) (mài) | • Japanese: 脈搏 (みやくはく, myakuhaku), 脈 ^(ja) (みやく, myaku) |
| • Czech: tep ^(cs) <i>m</i> , puls <i>m</i> | • Norman: pouls <i>m</i> |
| • Dutch: pols ^(nl) <i>m</i> | • Korean: 맥박 ^(ko) (maekbak) (脈搏) |
| • Esperanto: pulso ^(eo) | • Malay: nadi |
| • Faroese: æðraslættur <i>m</i> | • Norwegian: puls |
| • Finnish: pulssi | • Persian: نبض ^(fa) (nabz) |
| • French: pouls ^(fr) <i>m</i> | • Portuguese: pulso ^(pt) <i>m</i> |
| Old French: pouls <i>m</i> | • Russian: пульс ^(ru) <i>m</i> (pul's) |
| Middle French: pouls <i>m</i> | • Slovak: pulz <i>m</i> |

Etymological Wordnet

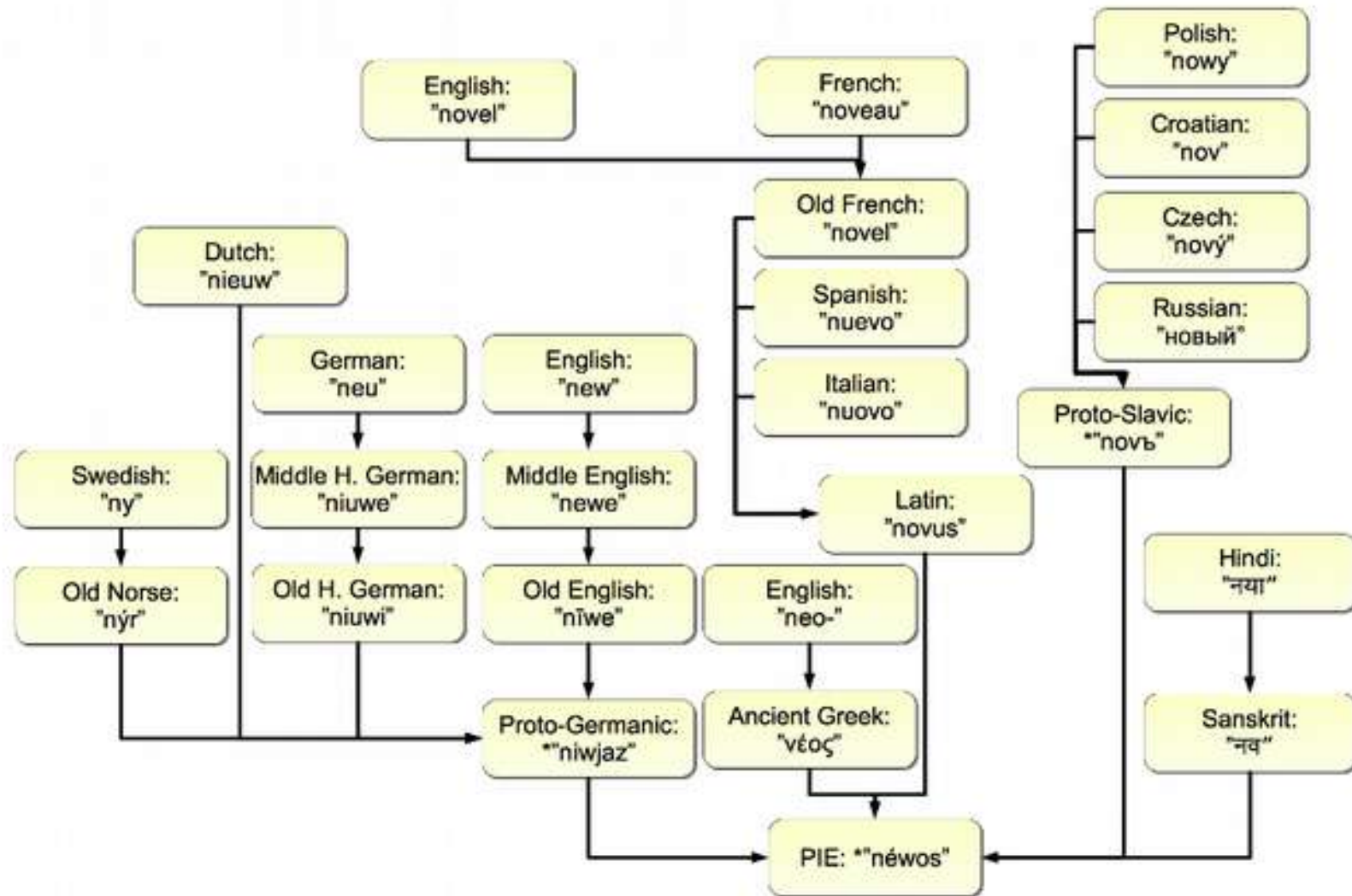


e.g. “salary” < Lat. “salarium” < Lat. “sal” (salt)

Etymological Wordnet



Etymological Wordnet



Lexical Knowledge: Wiktionary

Etymology 1 [\[edit\]](#)

From Latin *pulsus* ("beat"), from *pellere* ("to drive"), from Proto-Indo-European **pel* ("to drive, strike, thrust").

For spelling, the -e (on -ise) is so the end is pronounced /ɪz/, rather than /ɪz/ as in *pulls*, and does not change the vowel ('u'). Compare *else*, *false*, *convulse*.

Pronunciation [\[edit\]](#)

· IPA^(key): /pʌls/

· Audio (US)  0:00 [MENU](#)

Noun [\[edit\]](#)

pulse (plural **pulses**)

1. *(physiology)* A normally regular beat felt when arteries are depressed, caused by the pumping action of the heart.
2. A beat or throb. [\[quotations ▼\]](#)
3. *(music)* The beat or *tactus* of a piece of music.
4. An *autosoliton*.

Related terms [\[edit\]](#)

- impulse
- repulse

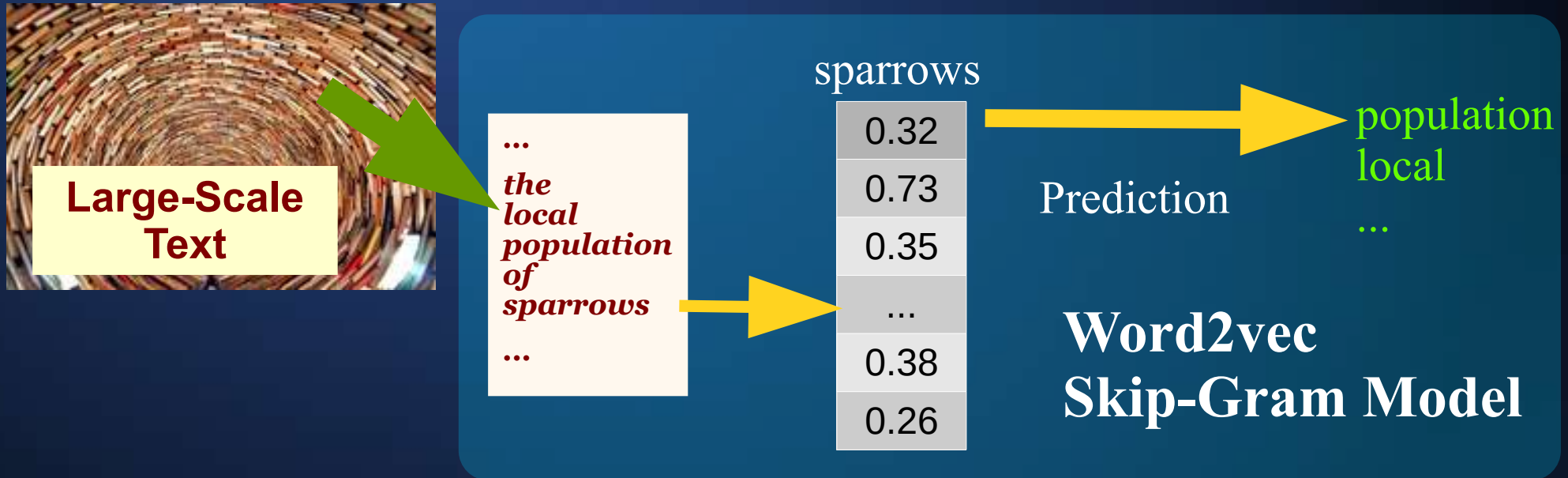
Translations [\[edit\]](#)

regular beat caused by the heart

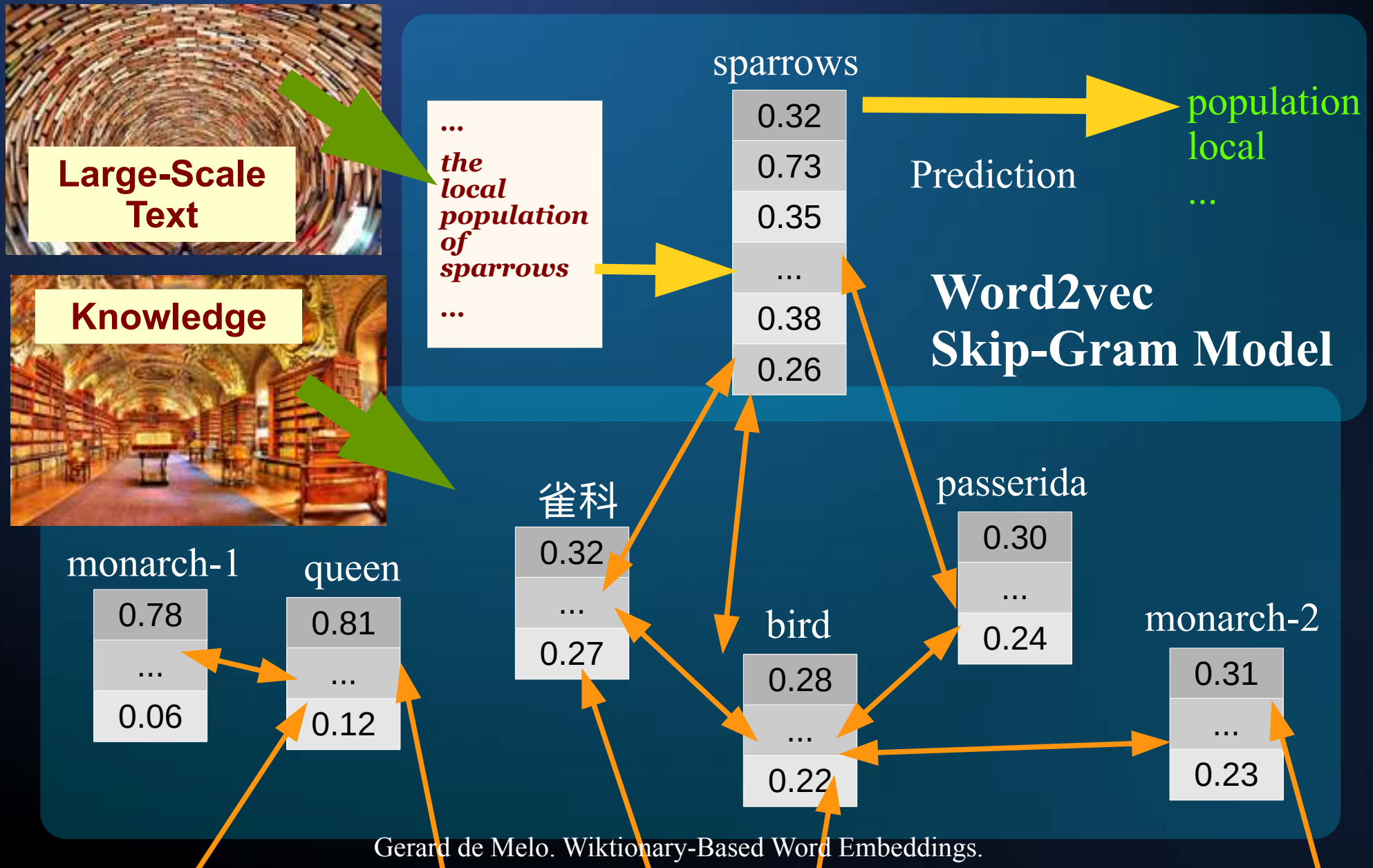
Select targeted languages

- | | |
|--|---|
| · Arabic: نبضة <i>f</i> (nábDa) | · Irish: <i>cuisle</i> <i>f</i> |
| · Chinese: | · Italian: <i>polso</i> <i>m</i> , <i>battito</i> ^(it) <i>m</i> |
| Mandarin: 脈搏 , 脉搏 ^(zh) (máibó), 脈 ^(zh) , 脉 ^(zh) (mài) | · Japanese: 脈搏 (みやくはく, myakuhaku), 脈 ^(ja) (みやく, myaku) |
| · Czech: tep ^(cs) <i>m</i> , puls <i>m</i> | · Norman: <i>pouls</i> <i>m</i> |
| · Dutch: pols ^(nl) <i>m</i> | · Korean: 맥박 ^(ko) (maekbak) (脈搏) |
| · Esperanto: pulso ^(eo) | · Malay: <i>nadi</i> |
| · Faroese: æðraslættur <i>m</i> | · Norwegian: <i>puls</i> |
| · Finnish: pulssi | · Persian: نبض ^(fa) (nabz) |
| · French: pouls ^(fr) <i>m</i> | · Portuguese: pulso ^(pt) <i>m</i> |
| Old French: poultz <i>m</i> | · Russian: пульс ^(ru) <i>m</i> (pul's) |
| Middle French: pouls <i>m</i> | · Slovak: pultz <i>m</i> |

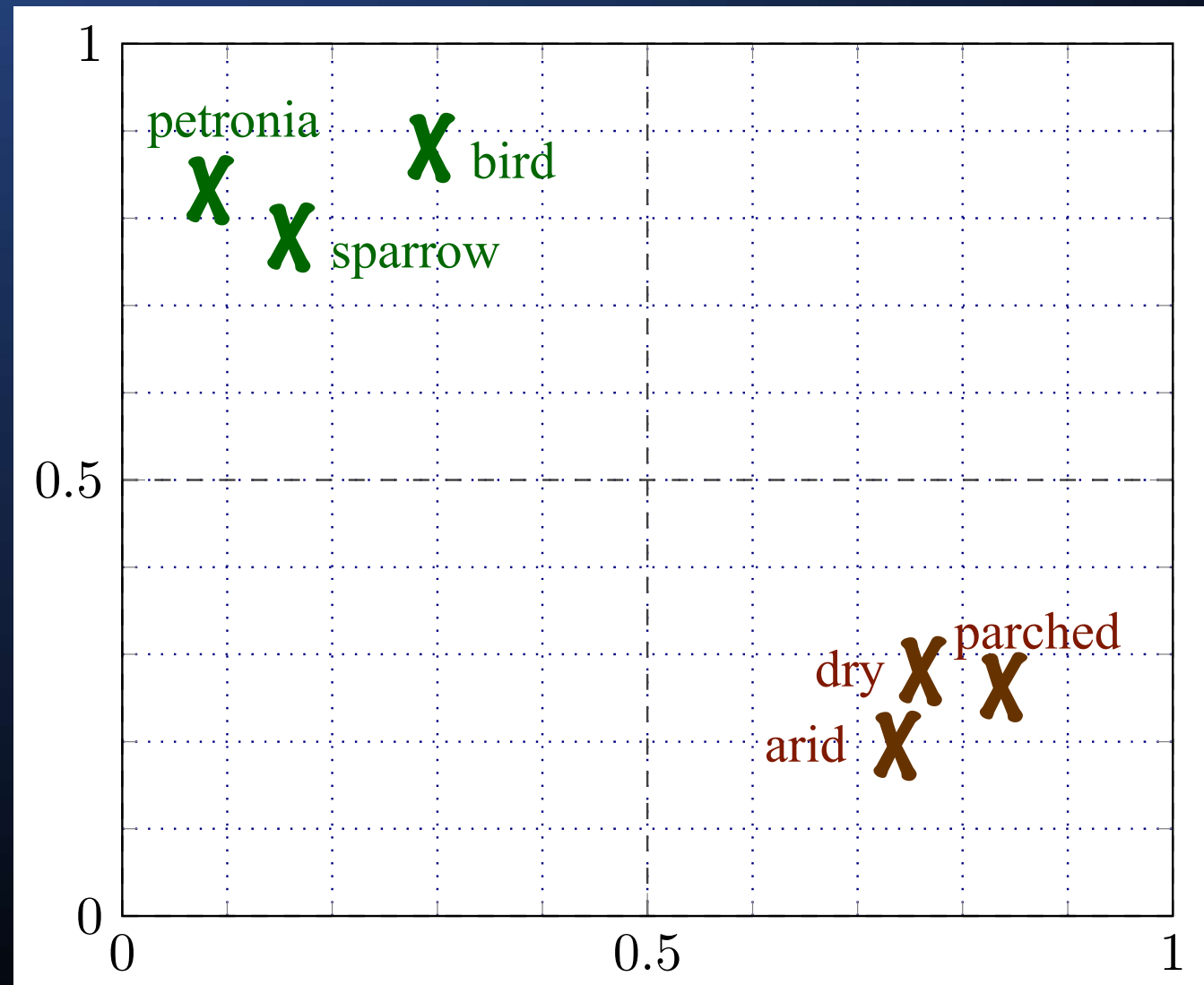
Word Vector Representations: word2vec



Word Vector Representations: word2vec



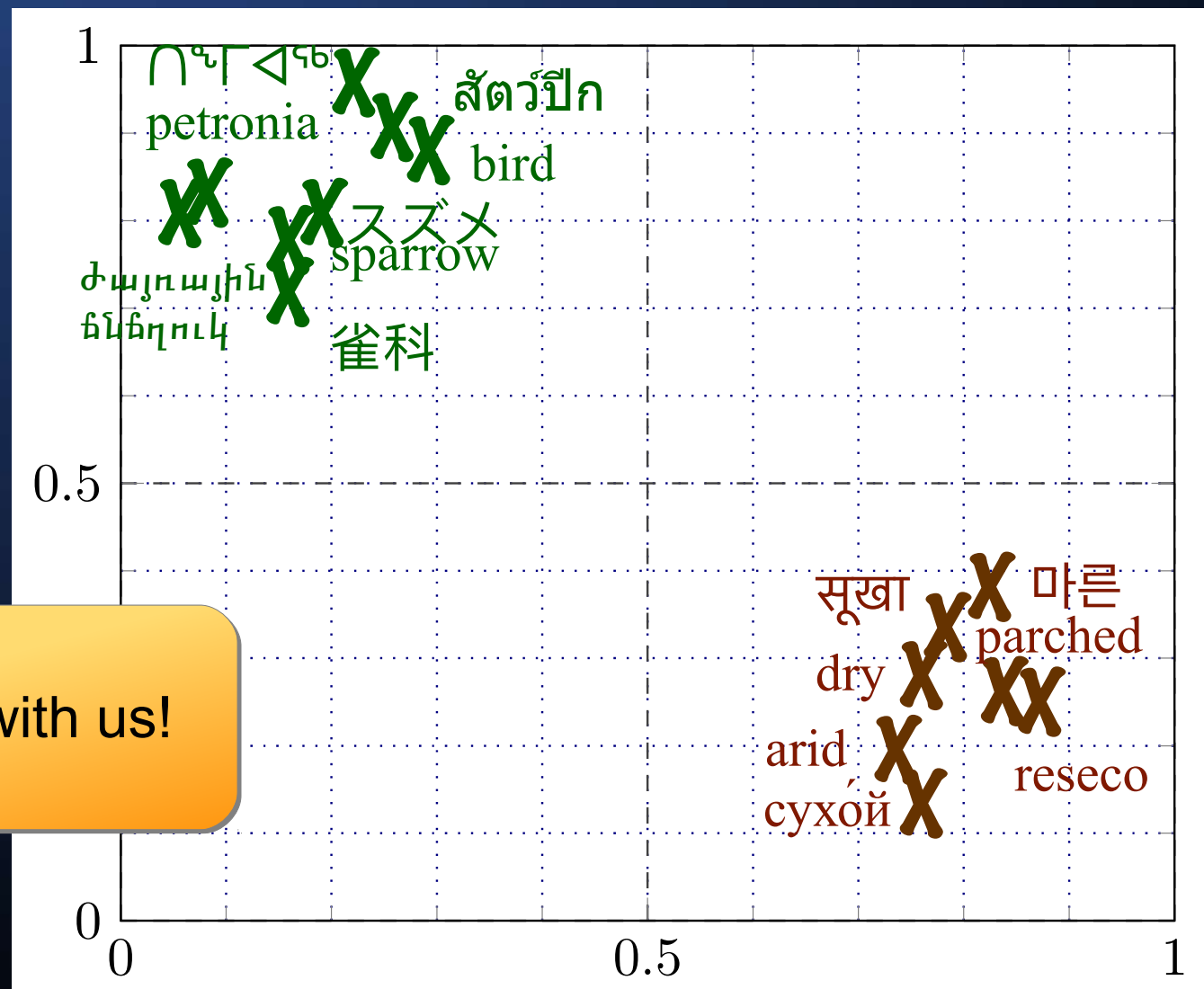
Multilingual Word Vectors



Multilingual Word Vectors

**5 million
words,
over 300
languages**

Get in touch with us!



Multilingual Word Vectors

Combine word2vec/GloVe Vectors
with multilingual lexical data from Wiktionary

Example:
German
language
datasets

RG65	Chandar A P et al. (2014) En-Es Vectors	0.629 @ 55.4%
	Ours (word2vec)	0.834 @ 100.0%
	Ours (Glove)	0.809 @ 100.0%
MC30	Chandar A P et al. (2014) En-Es Vectors	0.430 @ 60.0%
	Faruqui et al. (2015)	0.591 @ N/A
	Ours (word2vec)	0.811 @ 76.7%
	Ours (Glove)	0.848 @ 76.7%
WS353	Chandar A P et al. (2014) En-Es Vectors	0.256 @ 65.1%
	Ours (word2vec)	0.548 @ 65.6%
	Ours (Glove)	0.591 @ 65.6%

Multilingual Word Vectors

English–German RG65	Chandar A P et al. (2014) En-De Vectors	0.441 @ 38.4%
	Ours (word2vec)	0.812 @ 97.6%
	Ours (GloVe)	0.828 @ 97.6%
English–Spanish RG65	Chandar A P et al. (2014) En-Es Vectors	0.588 @ 59.5%
	Ours (word2vec)	0.869 @ 100.0%
	Ours (GloVe)	0.863 @ 100.0%
English–French RG65	Chandar A P et al. (2014) En-Fr Vectors	0.598 @ 52.0%
	Ours (word2vec)	0.864 @ 100.0%
	Ours (GloVe)	0.855 @ 100.0%
English–Spanish MC30	Chandar A P et al. (2014) En-Es Vectors	0.351 @ 70.0%
	Ours (word2vec)	0.745 @ 90.0%
	Ours (GloVe)	0.797 @ 90.0%
Spanish–English MC30	Chandar A P et al. (2014) En-Es Vectors	0.645 @ 56.7%
	Ours (word2vec)	0.713 @ 83.3%
	Ours (GloVe)	0.721 @ 83.3%
English–Spanish WS353	Chandar A P et al. (2014) En-Es Vectors	0.303 @ 75.9%
	Ours (word2vec)	0.582 @ 79.8%
	Ours (GloVe)	0.641 @ 79.8%
Spanish–English WS353	Chandar A P et al. (2014) En-Es Vectors	0.299 @ 73.3%
	Ours (word2vec)	0.550 @ 78.7%
	Ours (GloVe)	0.612 @ 78.7%

Multilingual Word Choice Quizzes

gourmet

- a) enjoys cooking
- b) has indigestion
- c) has an expert appreciation of food
- d) is hungry

dale

- a) plain
- b) retreat
- c) shelter
- d) valley

brace

- a) to scream
- b) prepare for danger
- c) hold your breath
- d) close your eyes

Word Meanings

Monarch



Word Meanings

Monarch



Image from:
Elizabeth: The Golden Age (2007)

Word Meanings

Kiwi



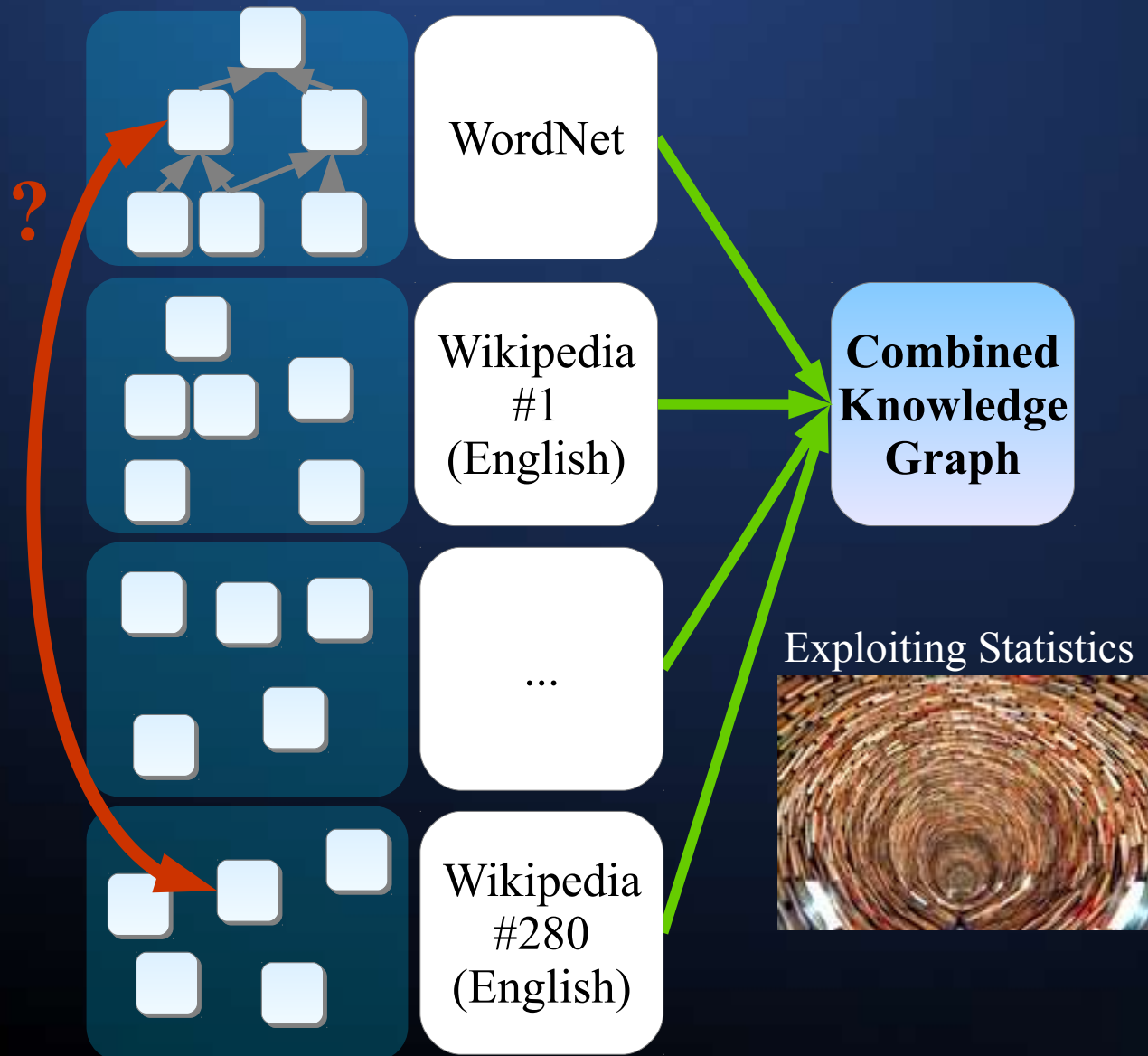
<https://commons.wikimedia.org/wiki/File:Tokoeka.jpg>

Word Meanings

Kiwi

<https://commons.wikimedia.org/wiki/File:Tokoeka.jpg>

Connecting Multiple Sources



Extraction from Wikipedia

Visit the main page



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes

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

Not logged in - Talk Contributions Create account Log in

Article Talk

Read Edit View history Search

Rutgers University

From Wikipedia, the free encyclopedia
(Redirected from Rutgers)

Coordinates:  40°30′6″N 74°26′53″W

"Rutgers" redirects here. For other uses, see Rutgers (disambiguation).

Rutgers, The State University of New Jersey (/rʌtɡərz/), commonly referred to as **Rutgers University**, **Rutgers**, or **RU**, is an American public research university and the largest institution for higher education in New Jersey.

Originally chartered as **Queen's College** on November 10, 1766, Rutgers is the eighth-oldest college in the United States and one of the nine colonial colleges chartered before the American Revolution.^{[9][10]} The college was renamed **Rutgers College** in 1825^[11] in honor of Colonel **Henry Rutgers** (1745–1830), a New York City landowner, philanthropist and former military officer, whose \$5000 bond donation to the school allowed it to reopen after years of financial difficulty.^[12] For most of its existence, Rutgers was a private liberal arts college affiliated with the Dutch Reformed Church. The college expanded its

Rutgers

The State University of New Jersey



Latin: *Universitas Rutgersensis Civitatis Novae Caesareae*^[1]

Former names Queen's College (1766–1825)
Rutgers College

Bertrand Russell



Born 18 May 1872
Trellech, Monmouthshire, UK

Died 2 February 1970 (aged 97)
Penryndreath, Wales, UK

Residence United Kingdom

Nationality British

Era 20th century philosophy

Region Western philosophy

School Analytic philosophy

Main interests Metaphysics, epistemology, logic, mathematics, philosophy of language, philosophy of mathematics, philosophy of science, ethics, philosophy of religion, history of philosophy

Notable ideas Analytic philosophy - logical atomism - theory of descriptions - knowledge by acquaintance and knowledge by description - Russell's paradox - Russell's teapot

Influenced by [\[show\]](#)

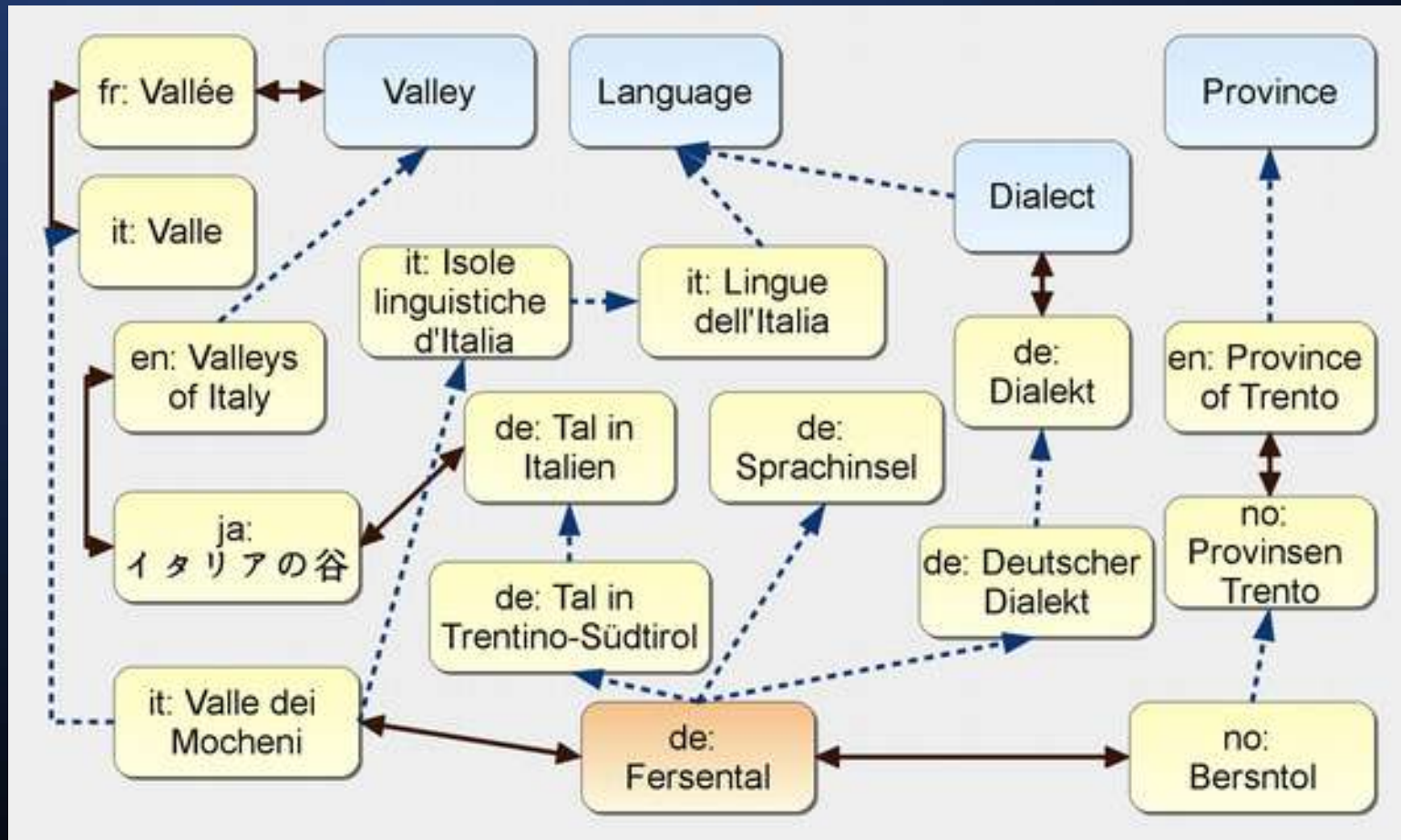
Influenced [\[show\]](#)

Awards Nobel Prize in Literature (1950)

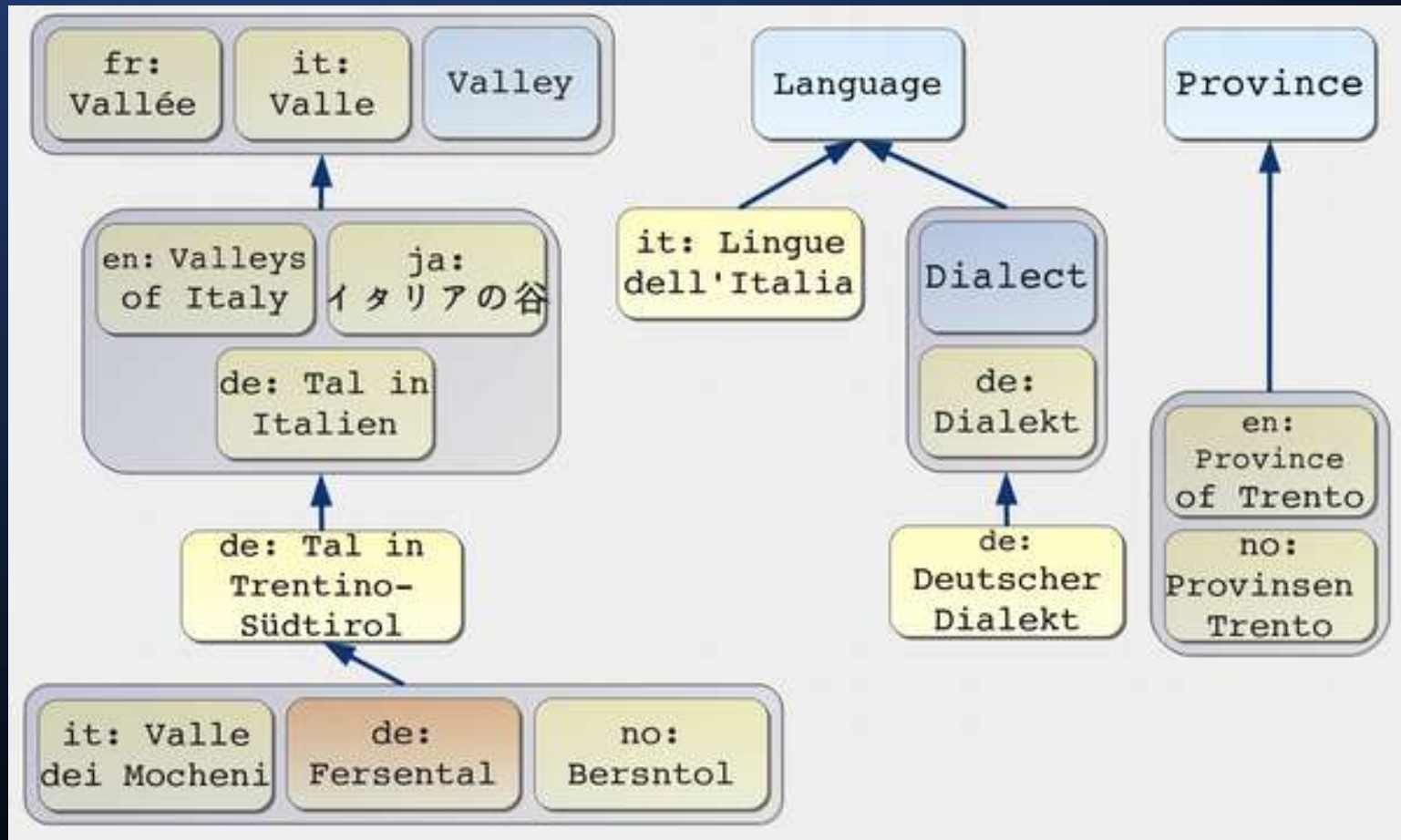
Signature 

MENTA

Integration Approach

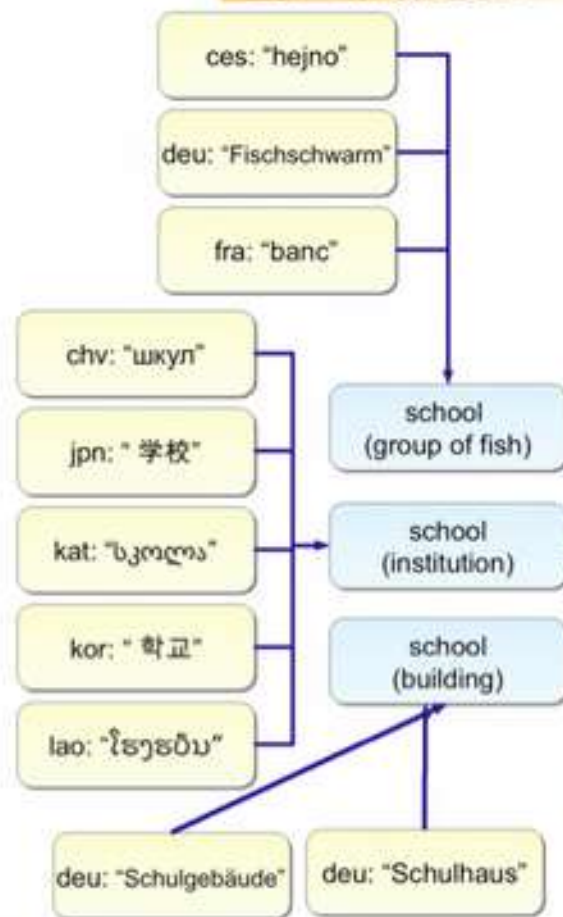


MENTA Integration Approach

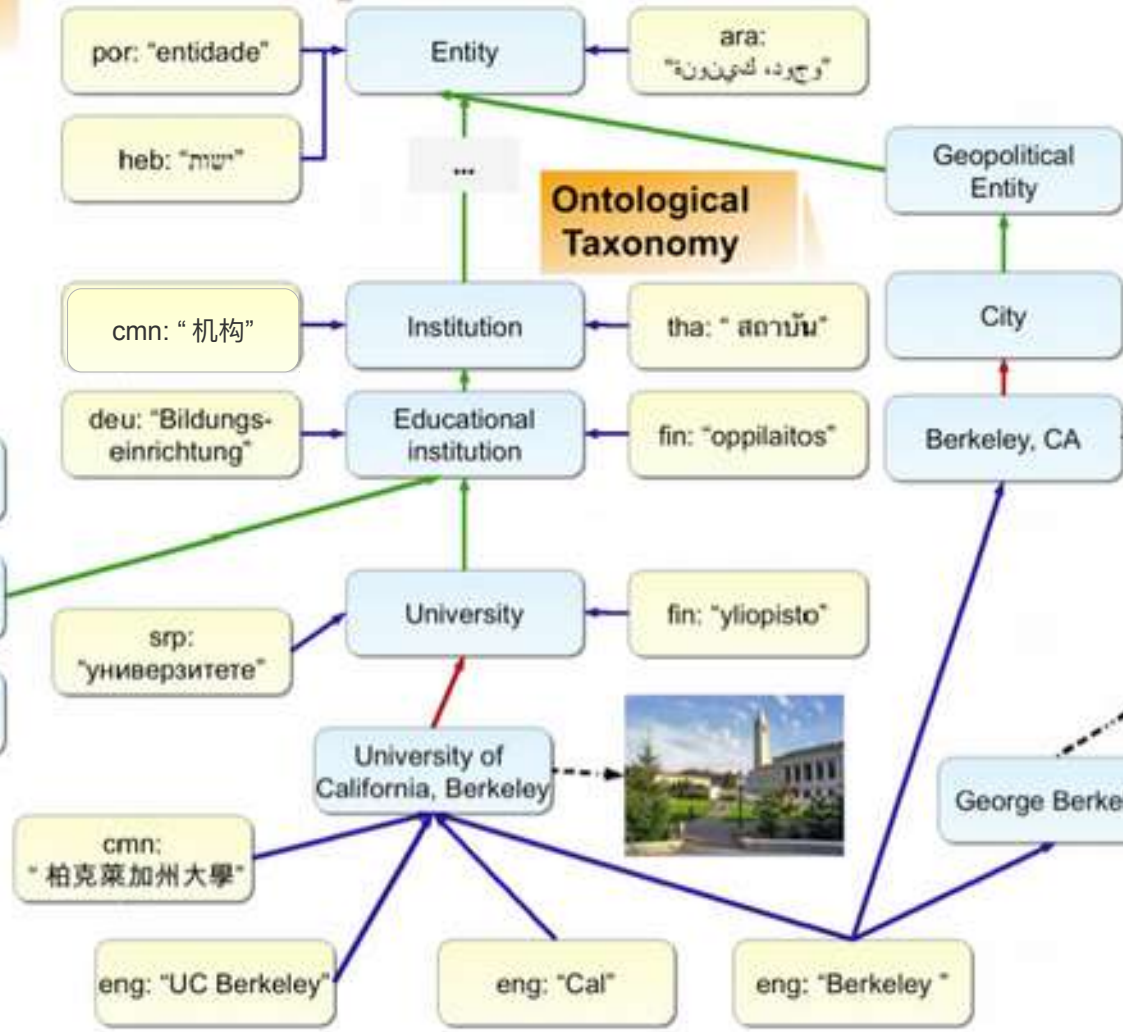


Universal Wordnet

200+ languages



UWN: Meaning Distinctions



Ontological Taxonomy

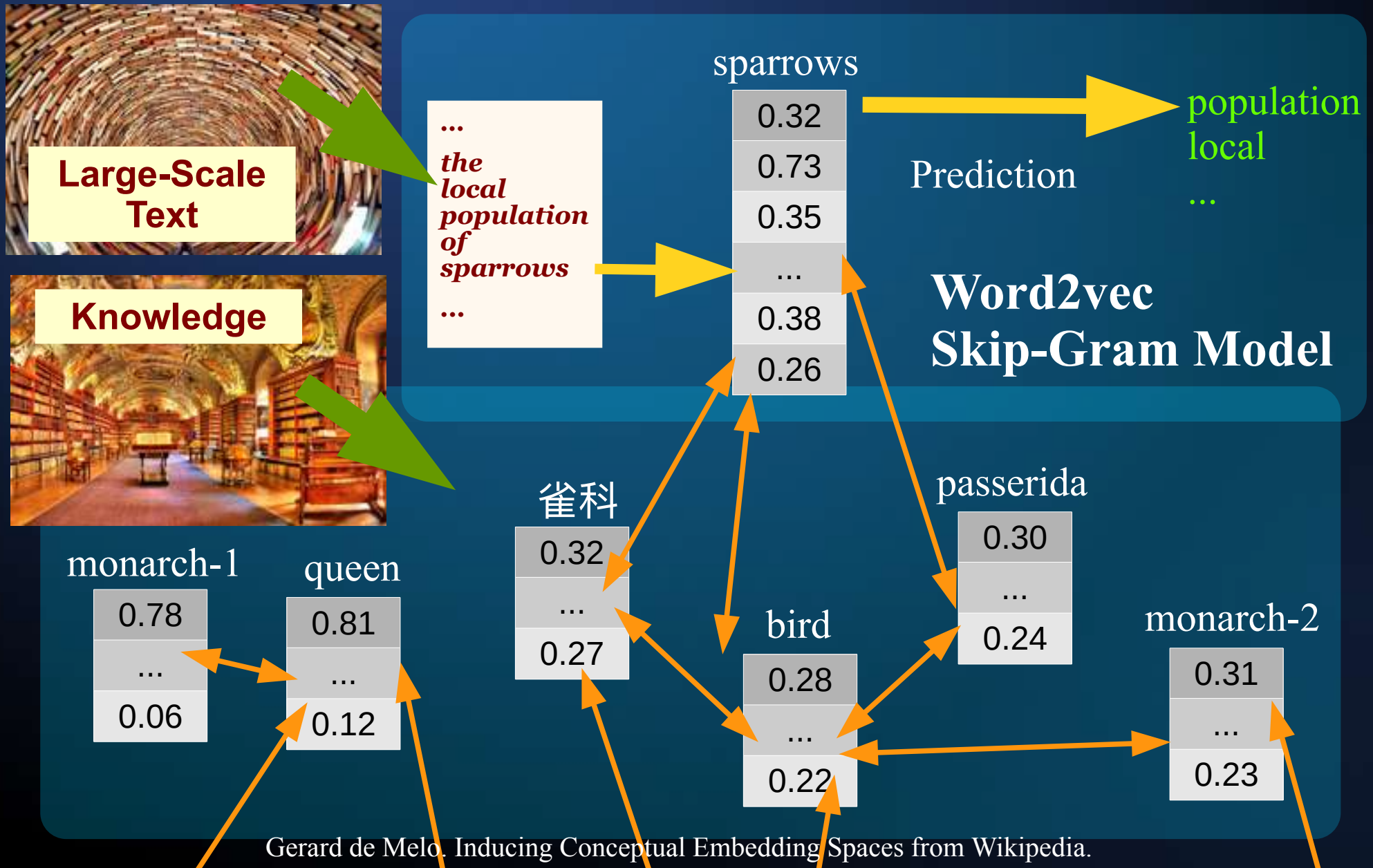
Encyclopedic Knowledge, Pictures, Video, Sounds, Maps



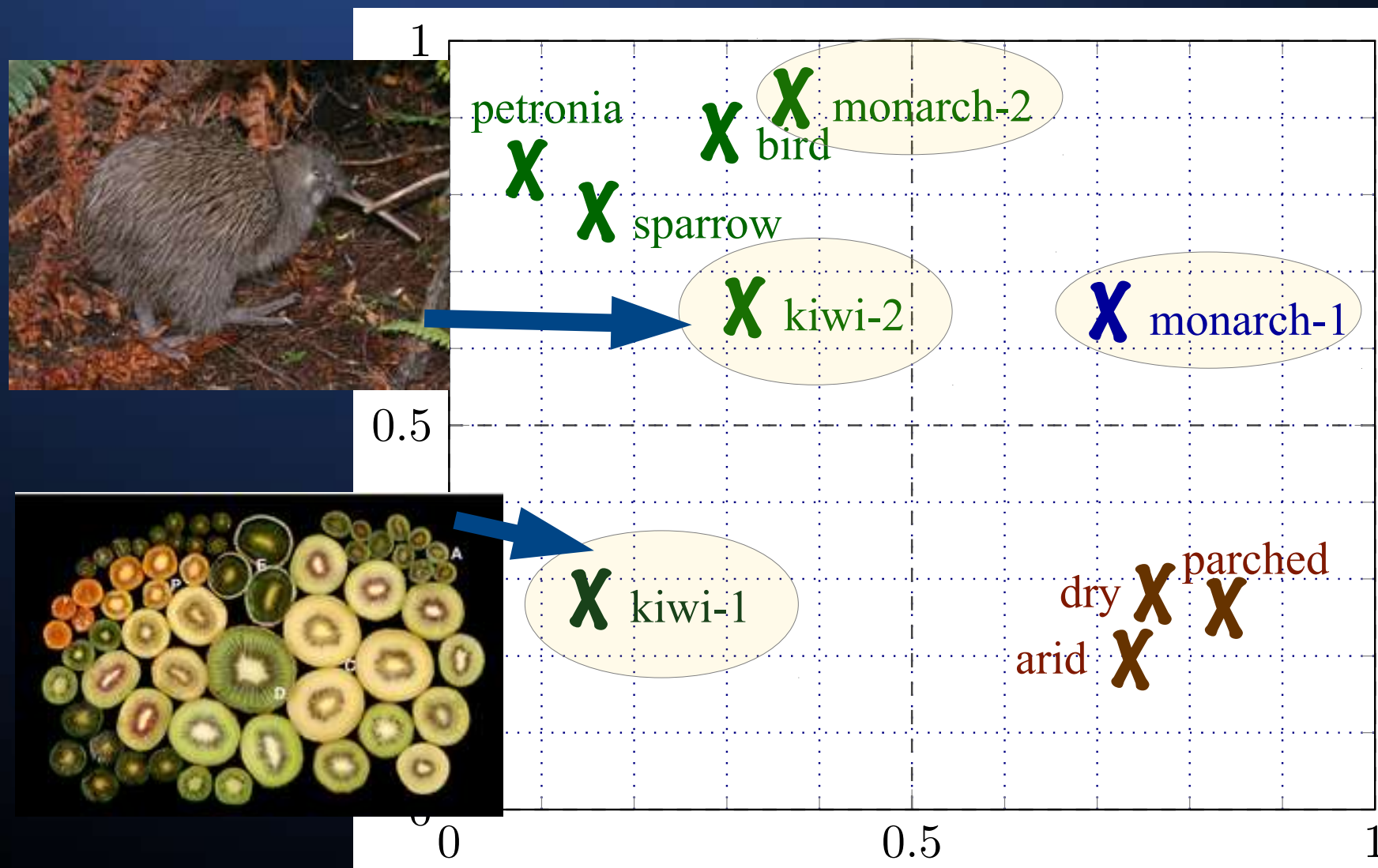
Millions of Named Entities (People, Places, Proteins, Asteroids, Companies, etc.)

<http://lexvo.org/uwn/>

Word Vector Representations: word2vec



Sense Representations



Application: Verbal Questions in IQ Tests

Identify two words (one from each set of brackets) that form a connection (analogy) when paired with the words in capitals.

*CHEMISTRY (laboratory, reaction, substances)
FAUNA (plants, animals, countryside)*

Answer: substances, animals

Multilingual Analogical Reasoning



Coffee is to Starbucks
as ... is to Lipton?

System: Tea

Source vectors: GLOVE

Gerard de Melo. Inducing Conceptual Embedding Spaces from Wikipedia.

Ongoing Research: Multilingual Analogical Reasoning



Apple is to Macbook
as ... is to Thinkpad?

System: 1. IBM 2. Compaq 3. Lenovo

Source vectors: GLOVE

Gerard de Melo. Inducing Conceptual Embedding Spaces from Wikipedia.

Ongoing Research: Multilingual Analogical Reasoning



m is to meters
as ... is to kilograms?

System: 1. kilos 2. kg 3. kilogrammes

Source vectors: GLOVE

Gerard de Melo. Inducing Conceptual Embedding Spaces from Wikipedia.

Ongoing Research: Multilingual Analogical Reasoning



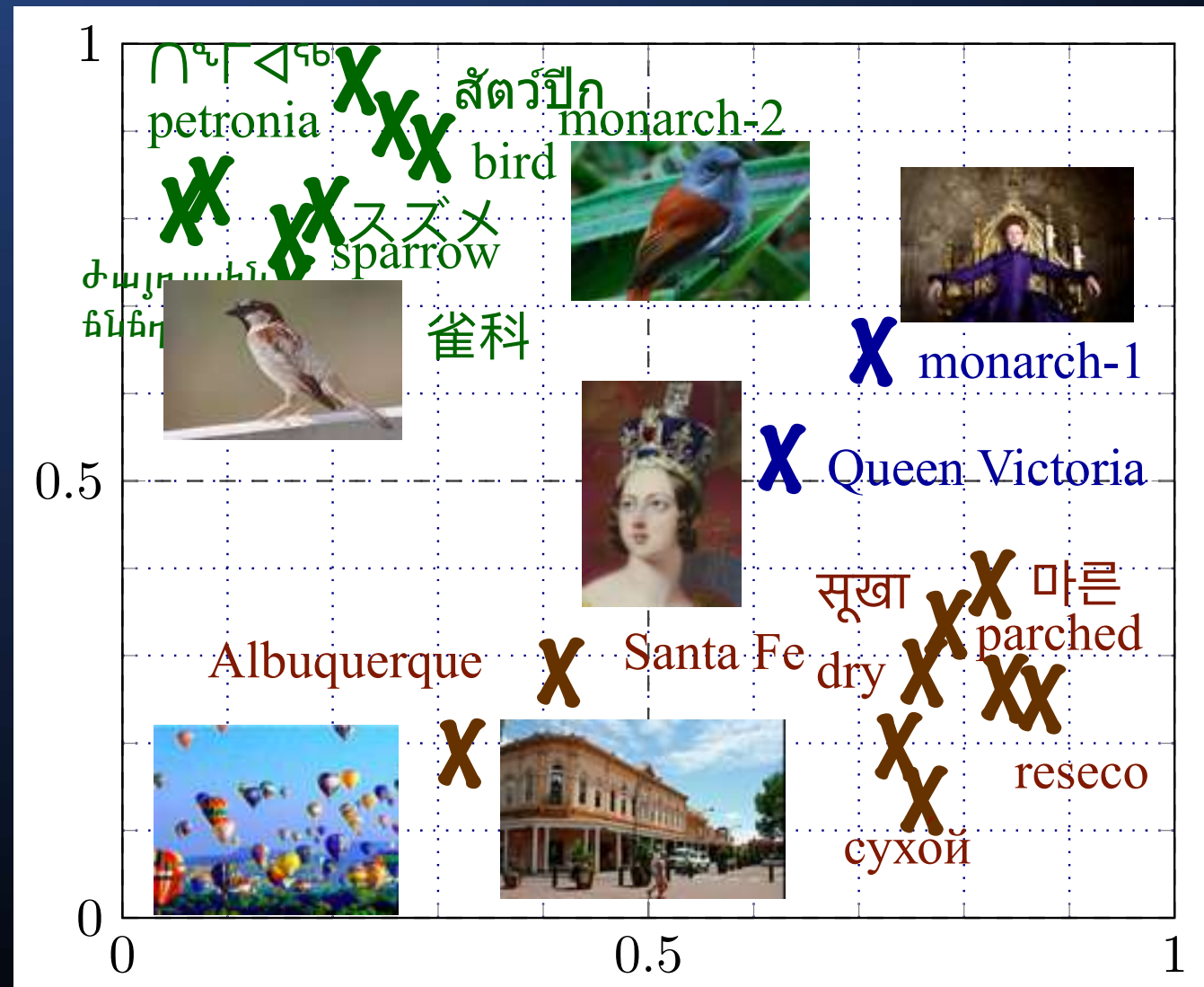
Germany is to Merkel
as ... is to Putin?

System: 1. Russia 2. Moscow 3. Europe

Source vectors: GLOVE

Gerard de Melo. Inducing Conceptual Embedding Spaces from Wikipedia.

Multilingual and Multimodal Word Vectors



Gerard de Melo (2017).
Inducing Conceptual
Embedding Spaces
from Wikipedia

Questions?



Image: <https://www.flickr.com/photos/opensourceway/5556249000>

