

Advancing Natural Language Understanding with Collaboratively Generated Content

Evgeniy Gabrilovich

Yahoo! Research

gabr@yahoo-inc.com

Making computers understand human language (at least to some extent 😊)

2

- The importance of **world knowledge** for natural language understanding
- Paradigm shift: from small-scale resources to **collaboratively generated content (CGC)**
- Distilling world knowledge from CGC repositories
- Future research directions
 - ▣ Obtaining semantic clues from behavioral information (e.g., the authoring process)



So what differentiates us from computers ?

dinner

3



d i n n e r



Motivating example

ODP (dmoz.org)



www.teckcominco.com

Excerpt from document
#15264 (Reuters-21578)

“... Cominco and Teck’s
22 pct-owned Lornex
agreed in January 1986
to form the joint venture,
merging their Highland
Valley operations ...”

Copper

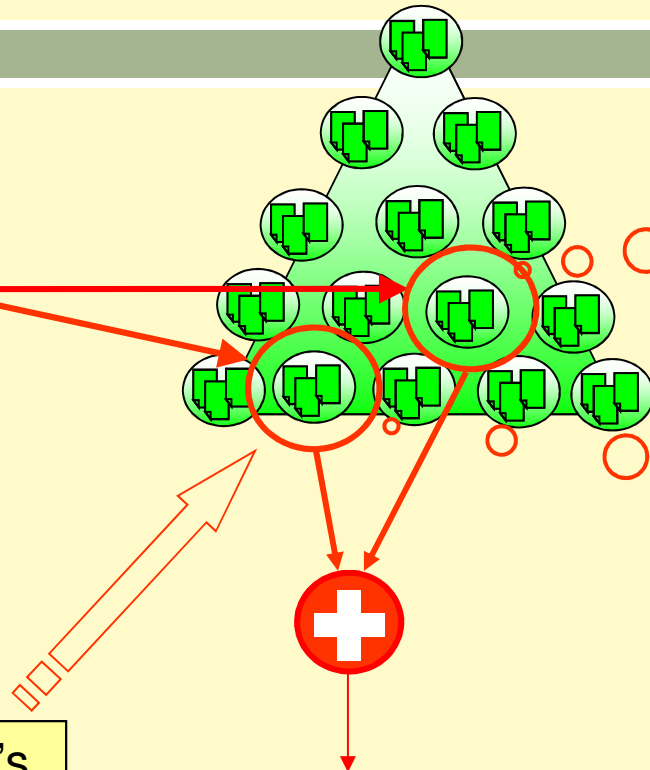
Bag of words

Classifier

Copper

Science/
Technology/
Mining

Business/
Mining and
Drilling



Sources of knowledge

5

Collaboratively Generated Content

Before



articles: **65,000**
Since 1768

WordNet

entries: **150,000**
Since 1985

CYC

assertions: **4,600,000**
Since 1984

After

3,600,000 articles
Since 2001



WIKIPEDIA
The Free Encyclopedia

2,500,000 entries
Since 2002



WIKTIONARY
the free dictionary

1,000,000,000 Q&A
Since 2005



4,000,000,000 images
Since 2004



Example problem:

Semantic relatedness of words and texts

6

□ **How related are**

cat $\Leftrightarrow$ mouse



Augean stables $\Leftrightarrow$ golden apples of the Hesperides

慕田峪 (Mutianyu) $\Leftrightarrow$ 司马台 (Simatai)

□ **Used in many applications**

▣ Information retrieval

▣ Word-sense disambiguation

▣ Error correction (“Web **site**” or “Web **sight**”)

Semantic relatedness using Wikipedia

7

□ Structure (links)



Strube & Ponzetto '06 (**WikiRelate**)



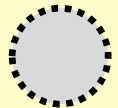
Milne & Witten '08 (**WLM**)

□ Content (text)

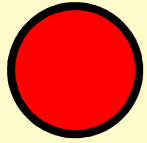


Gabrilovich & Markovitch '09 (**ESA**)

□ Structure + content



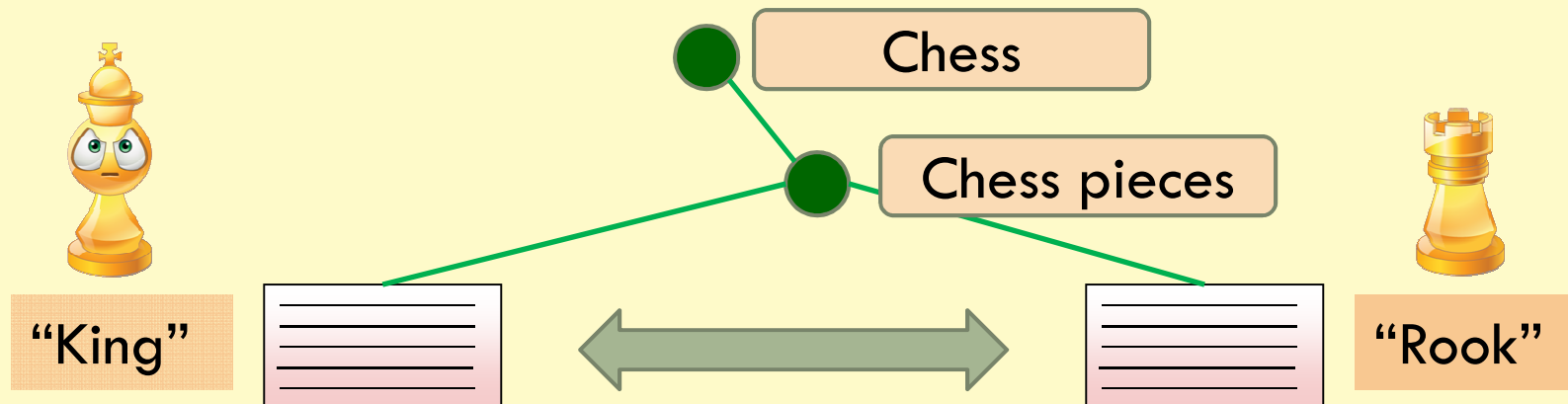
Yeh et al. '09 (**WikiWalk**)

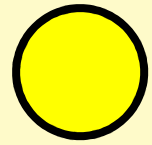


WikiRelate [Strube & Ponzetto '06]

8

- Find Wikipedia articles titled with the given words
 - ▣ Attempt joint disambiguation, if possible
 - King → Monarch, King (chess), King (magazine), ...
 - Rook → Rook (bird), Rook (chess), Rook (card game), ...
 - ▣ Otherwise, choose the first sense
- Compute relatedness
 - ▣ Compare full article texts or use the **category hierarchy**

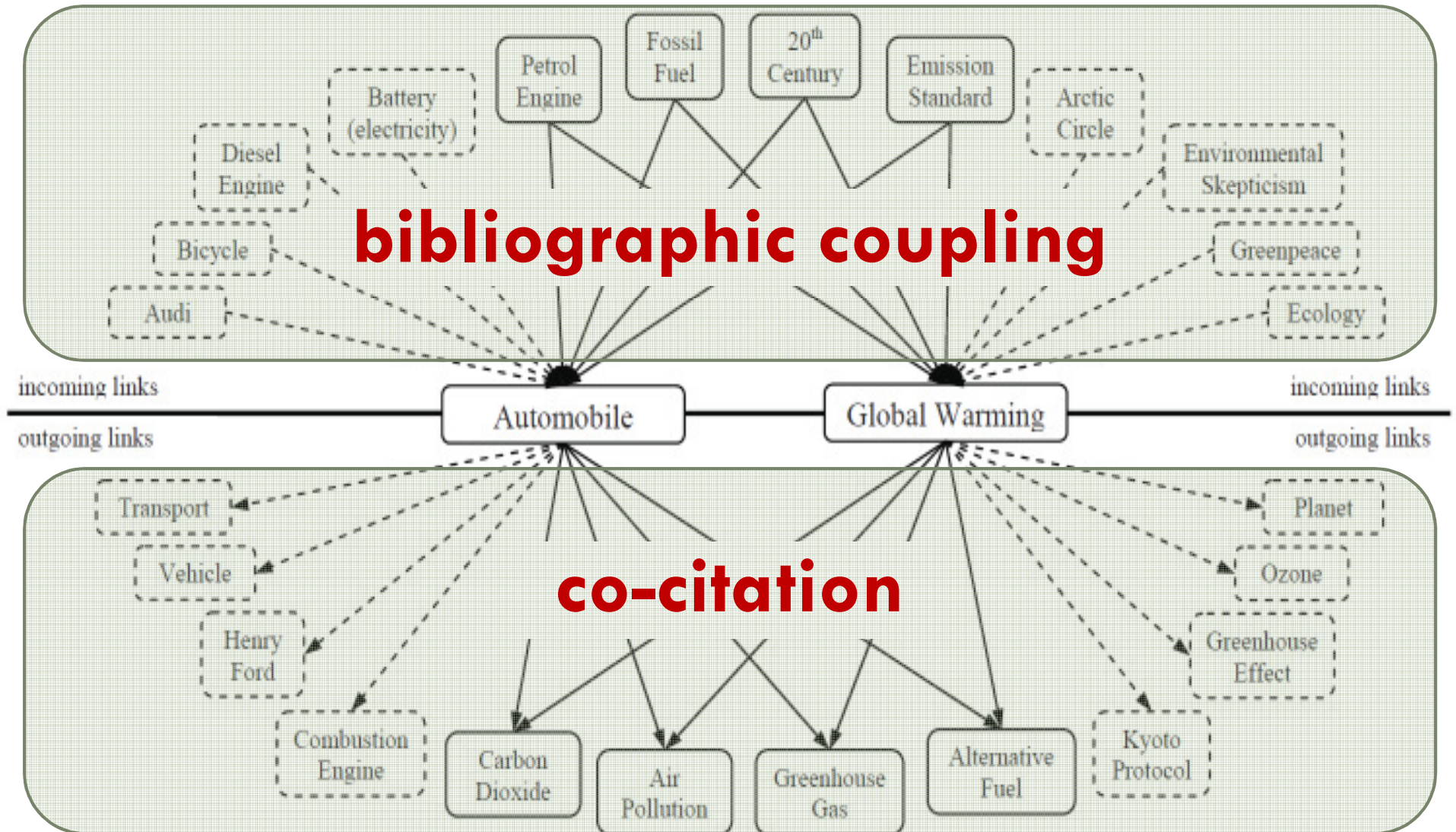


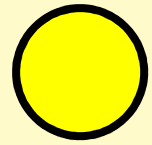


Wikipedia Link-based Measure (WLM)

Using **all** Wikipedia links [Milne & Witten '08]

9



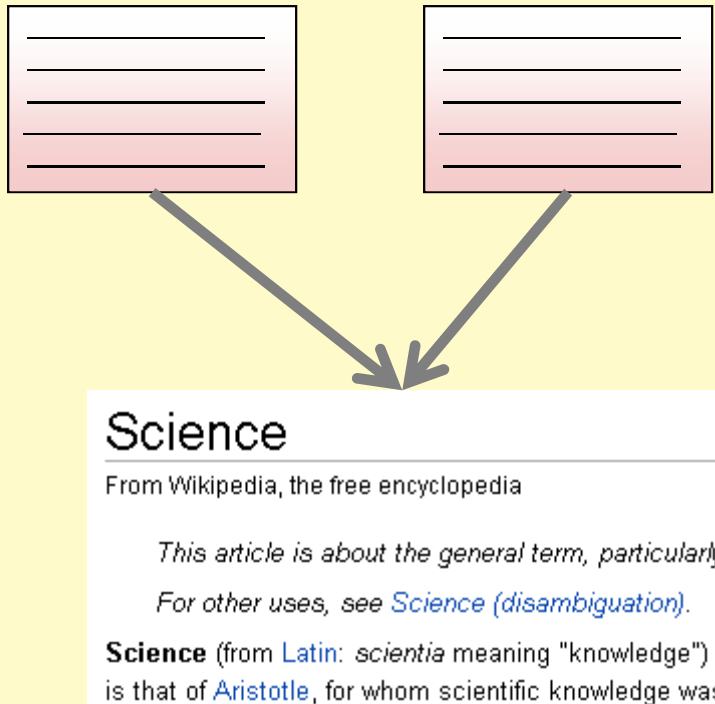


Wikipedia Link-based Measure (continued)

Not all links are equal !

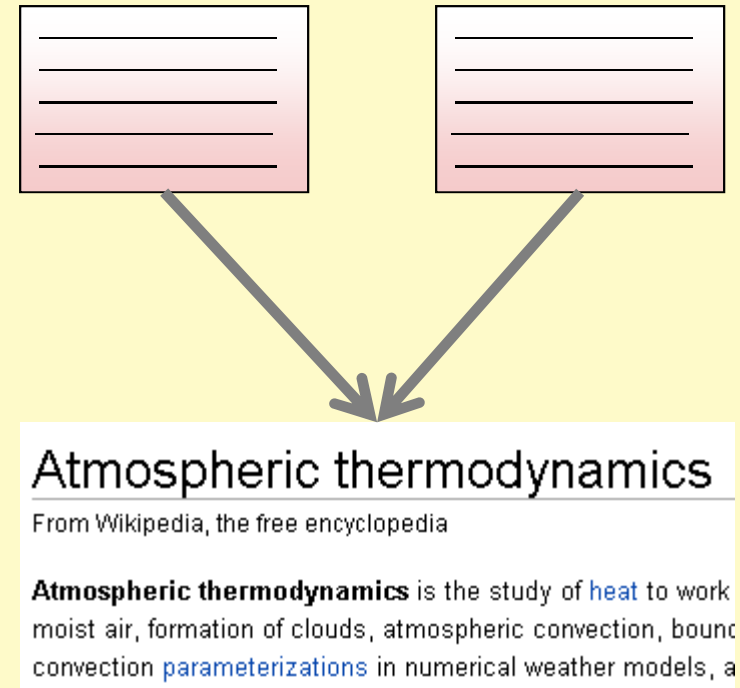
10

Loosely related

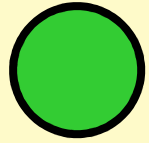


Very general concept

Tightly related

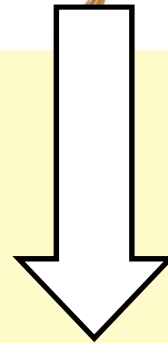
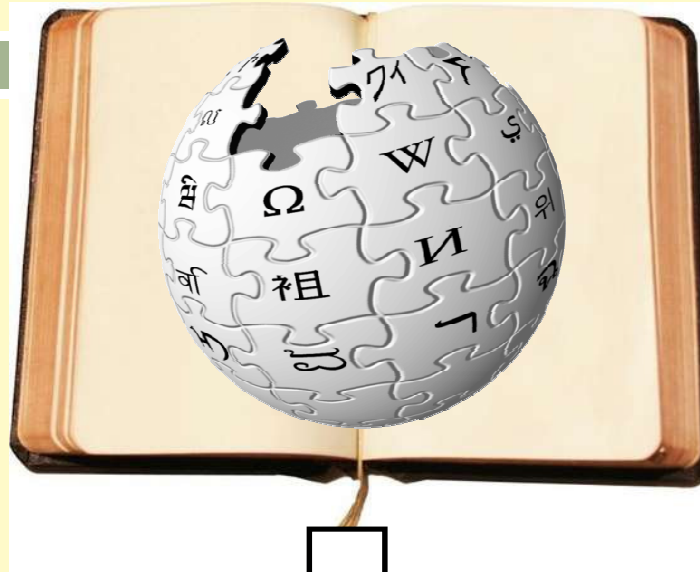


Very specific concept



The AI Dream

11



Problem



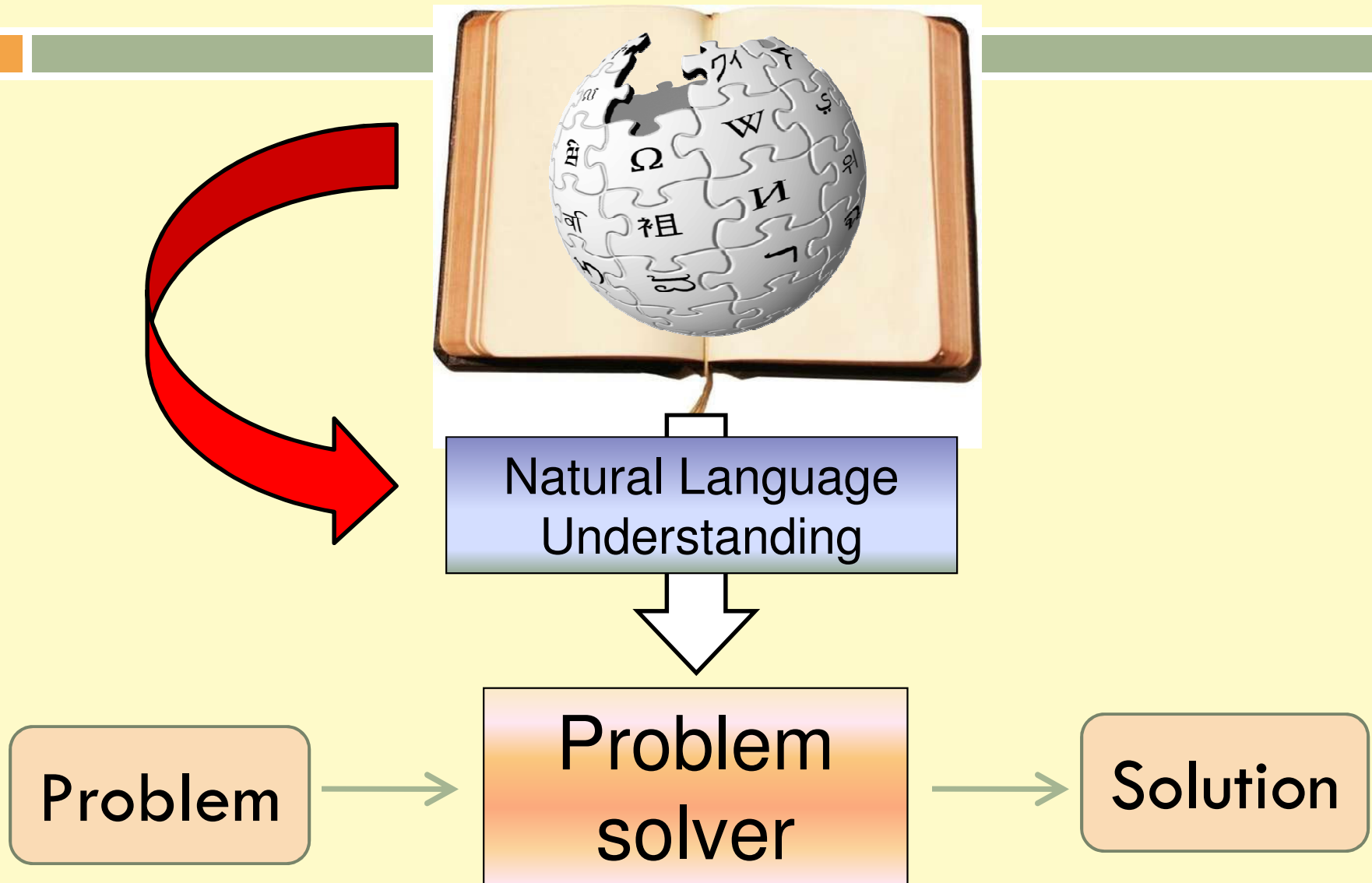
Problem
solver

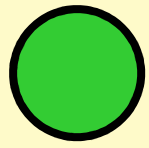


Solution

The circular dependency problem

12

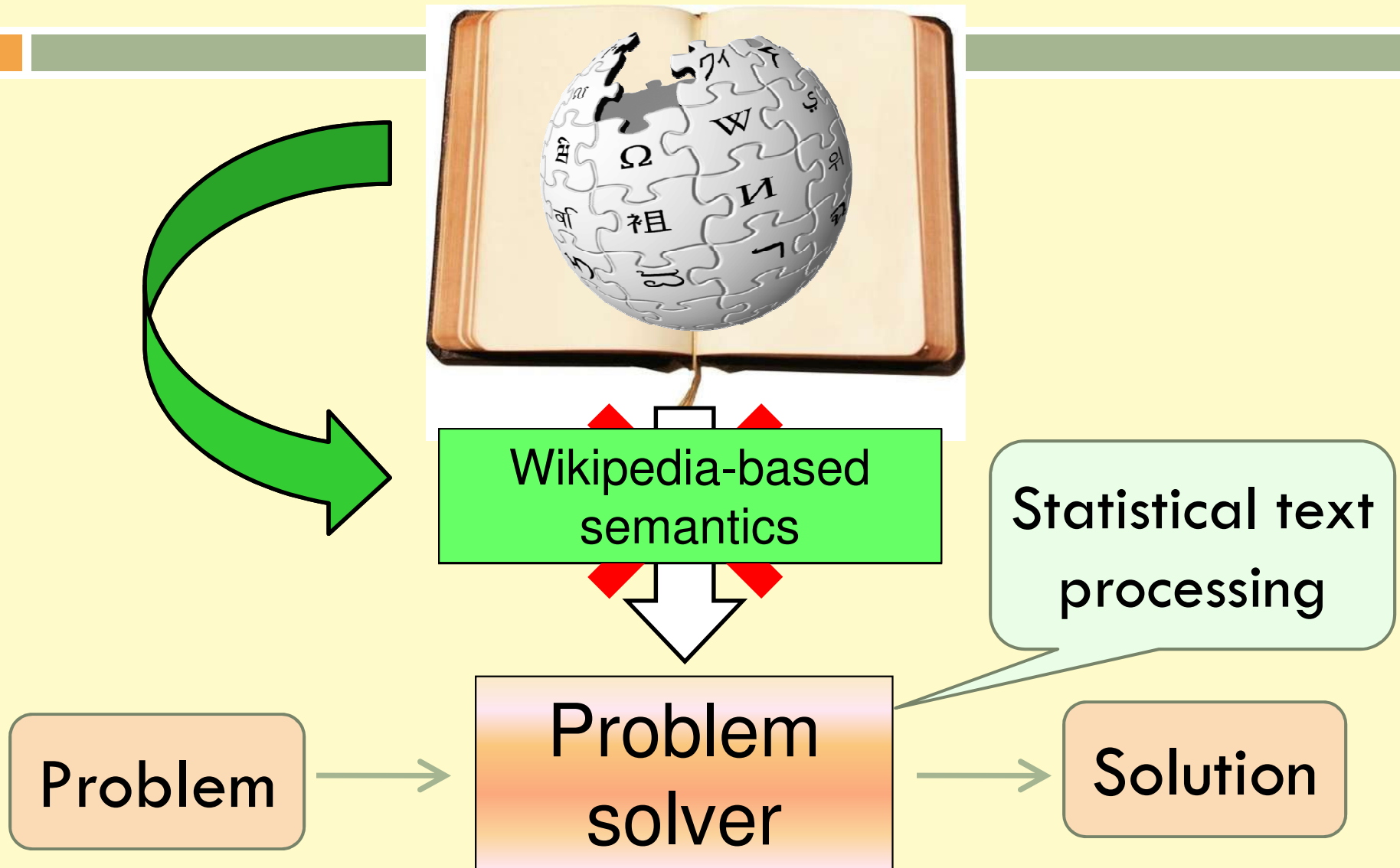




Circumventing the need for deep Natural Language Understanding

Using Wikipedia content

13



Every Wikipedia article represents a **concept**

14



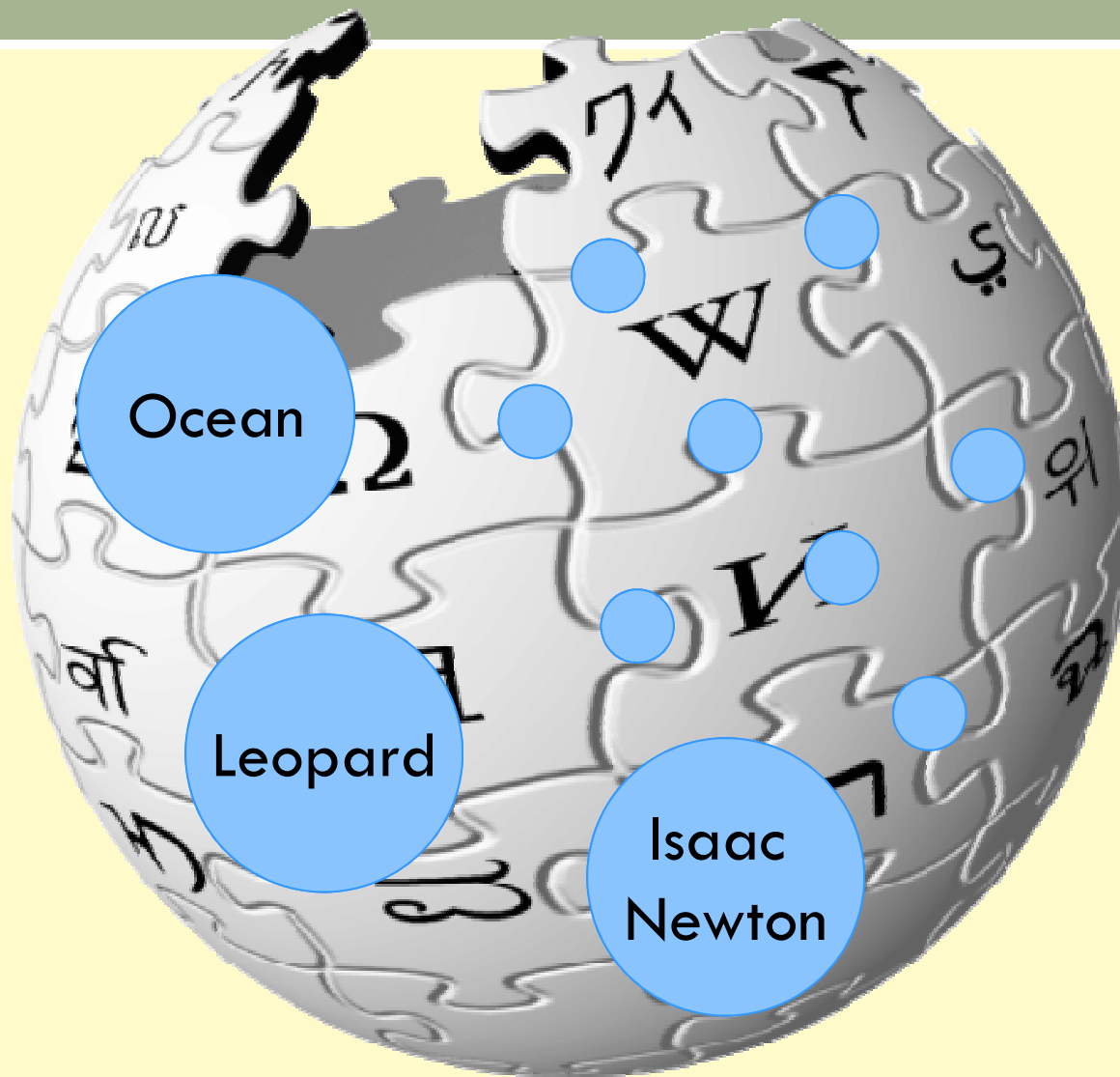
The screenshot shows the Wikipedia article for "Leopard". At the top, there's a navigation bar with links for "article", "discussion", "edit this page", and "history". Below this, the title "Leopard" is displayed, followed by the text "From Wikipedia, the free encyclopedia". A note states: "Leopards" redirects here. For other uses, see *Leopards (disambiguation)*. The main text describes the leopard (*Panthera pardus*) as a member of the Felidae family, the smallest of the four "big cats" in the genus *Panthera*, along with the tiger, lion, and jaguar. It mentions its distribution across southern Asia and Africa, from Korea to South Africa, and notes that its range has decreased due to hunting and habitat loss. It also mentions that the greatest concentration of leopards now occurs in sub-Saharan Africa, with fragmented populations in Pakistan, India, Indochina, Malaysia, and China. Due to the loss of range and declines in population, it is graded as a "Near Threatened" species. Its numbers are greater than other *Panthera* species, all of which face more acute conservation concerns. A small image of a leopard standing on a tree branch is shown on the right. Below the image, the text "Leopard^[1]" is followed by "Fossil range: Late Pliocene or Early Pleistocene to Recent".



Leopard

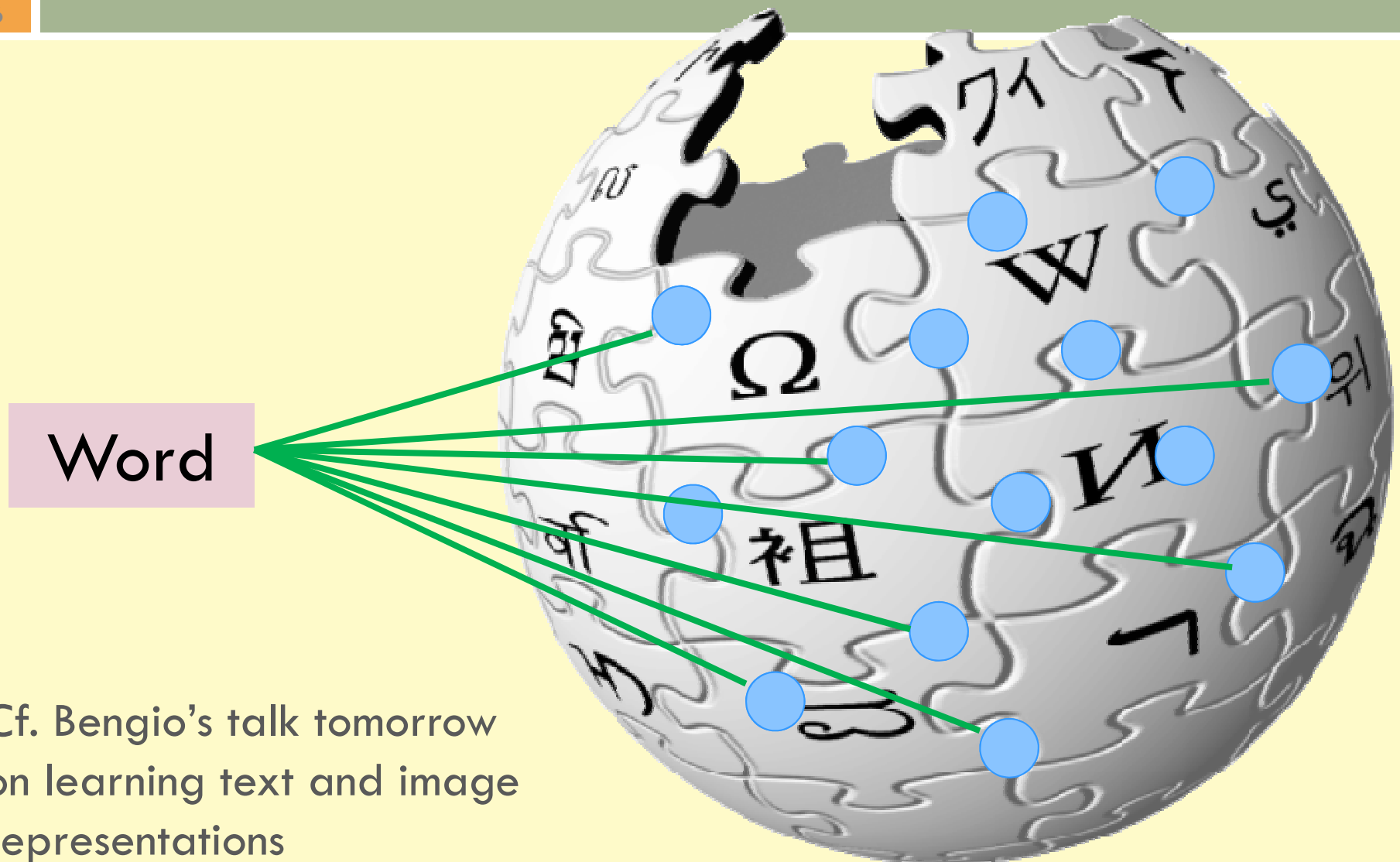
Wikipedia can be viewed as an ontology –
a collection of **concepts**

15



The **semantics** of a word is a vector of its **associations** with Wikipedia concepts

16



Cf. Bengio's talk tomorrow
on learning text and image
representations

Concept and word representation in Explicit Semantic Analysis

[Gabrilovich & Markovitch '09]

17

Panthera

Cat [0.92]

Leopard [0.84]

Roar [0.77]

Panthera

From Wikipedia, the free encyclopedia

Not to be confused with [Pantera](#).

For other uses, see [Panthera \(disambiguation\)](#).

Panthera is a [genus](#) of the [family Felidae](#) ([cats](#)) which contains four well-known living [species](#): the [tiger](#), the [lion](#), the [jaguar](#), and the [leopard](#). The genus comprises about half of the [Pantherinae](#) subfamily, the [big cats](#). The word **panther**, while technically referring to all members of the genus, is commonly used to specifically designate the [black panther](#).

Only the four *Panthera* cat species have the anatomical structure that enables them to [roar](#). The primary reason for this was formerly assumed to be the incomplete [ossification](#) of the [hyoid bone](#). However, new studies show that the ability to roar is due to other [morphological](#) features, especially of the [larynx](#). The [snow leopard](#), *Uncia uncia*, which is sometimes included within *Panthera*, does not roar. Although it has an incomplete ossification of the hyoid bone, it lacks the special morphology of the larynx.^[2]

[Contents](#) [\[show\]](#)

Name

[\[edit\]](#)

See also: [Panther \(legendary creature\)](#)

According to the *American Heritage Dictionary*, the origin of the word is unknown. A folk etymology derives the word from the [Greek](#) πάν *pan-* ("all") and θήρ *thēr* ("beast of prey") because they can hunt and kill almost everything. The Greek word πάνθηρ *pánthēr*, referred to all spotted *Felidae* generically. Although it came into English through the classical

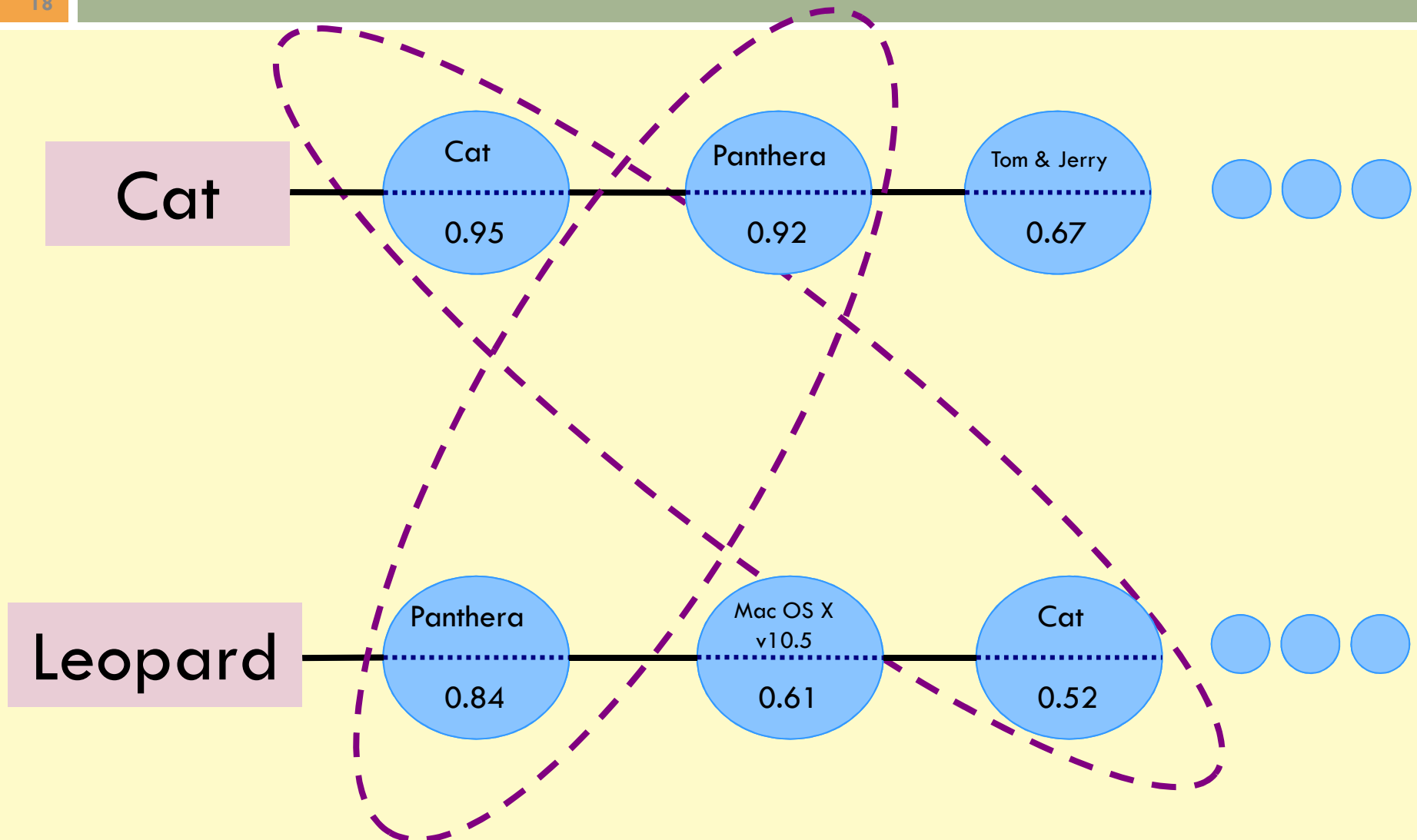
Panthera^[1]

Temporal range: Early [Pliocene](#) to
Recent



Computing semantic relatedness as **cosine** between two vectors

18



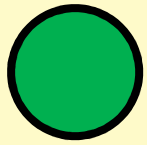
Experimental results (individual words)

19

Limitations of lexical resources

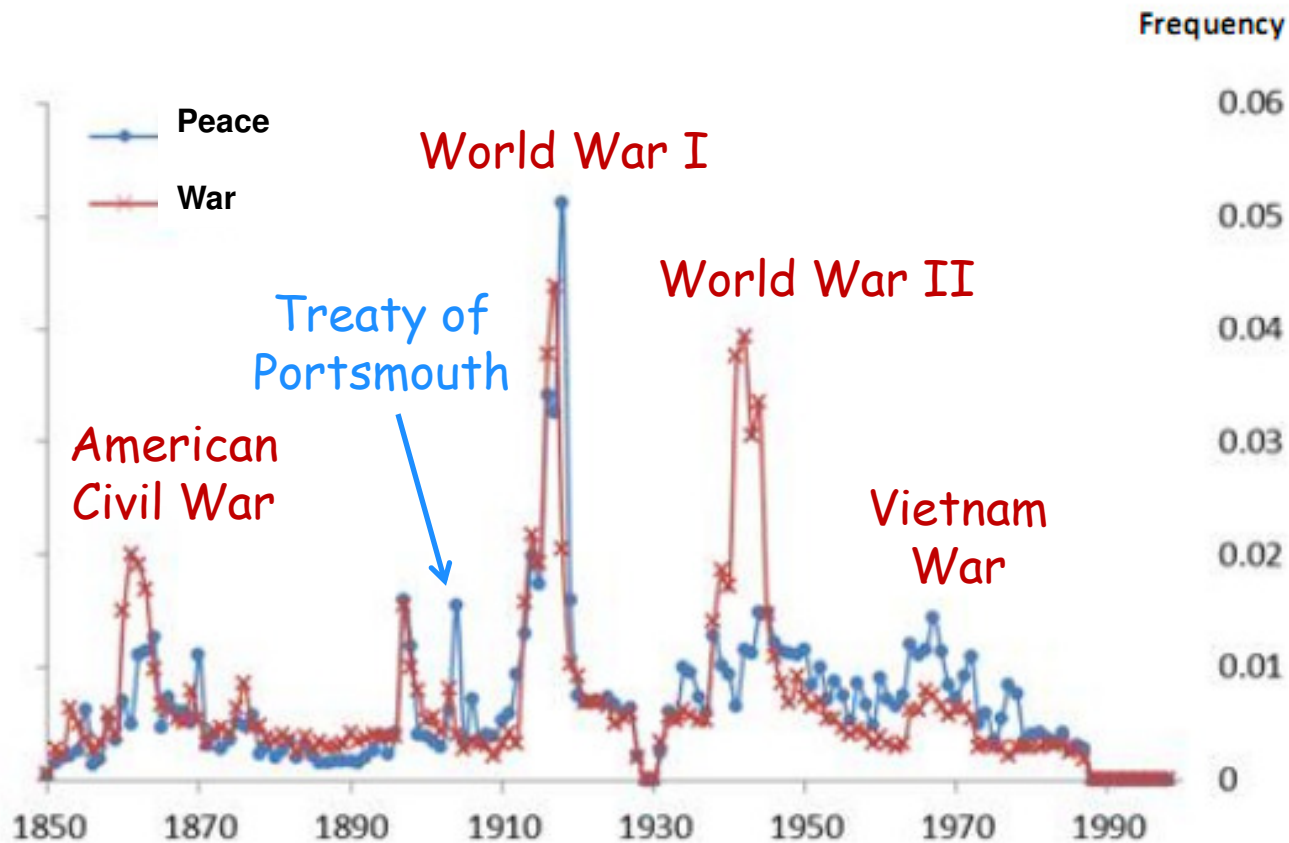
- Small coverage (especially proper names)
- Descriptions of senses are short and scarce
- Overly fine-grained sense distinctions
“go (v)” has 30 senses, “make (v)” – 48 (sic!)

Method		Correlation with human judgments
WordNet		0.33–0.35
Roget's Thesaurus		0.55
WikiRelate (categories)	●	0.49
WLM (all links)	●	0.69
ESA (Explicit Semantic Analysis)	●	0.75
TSA (Temporal Semantic Analysis)	●	0.8

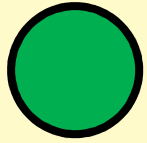


Augmenting Explicit Semantic Analysis with the **time dimension**

20



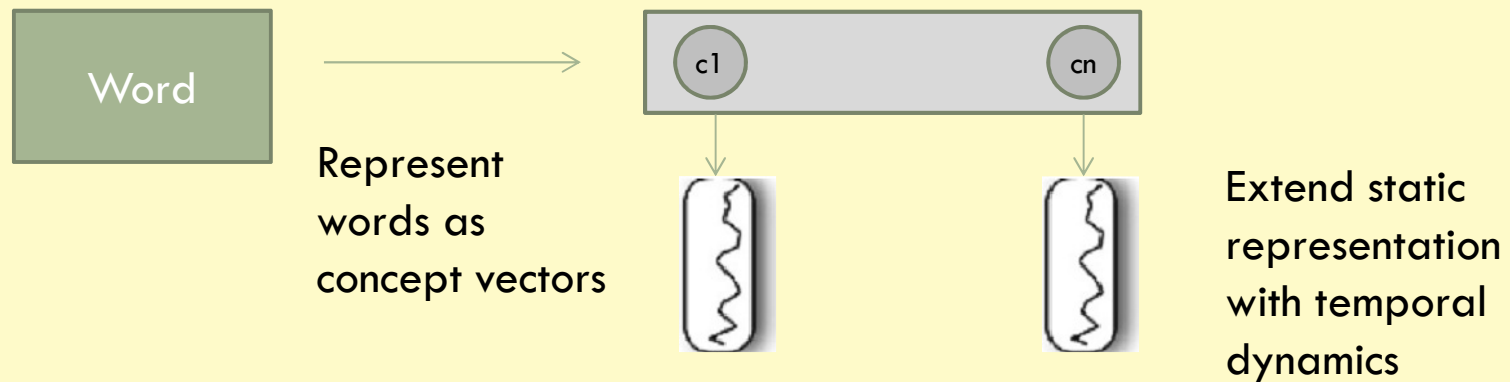
New York Times archives 1850–1990



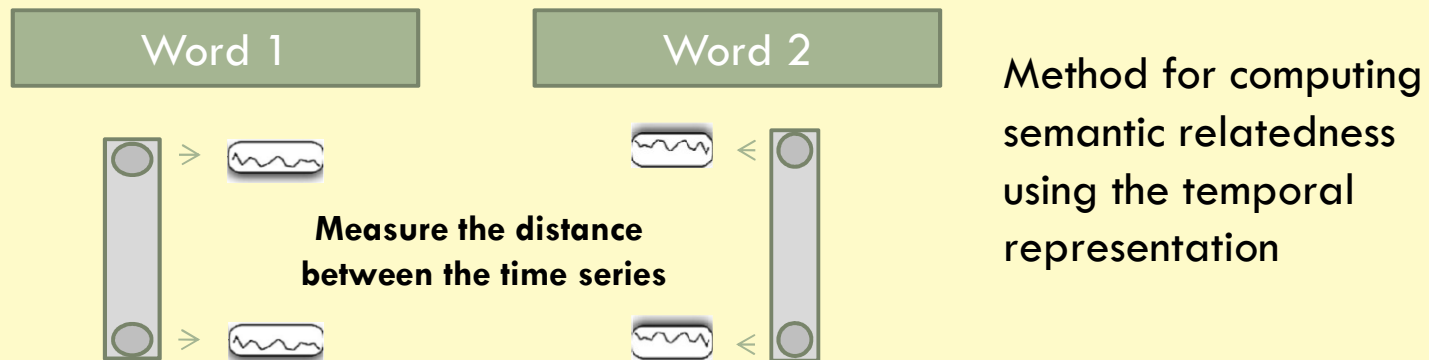
Temporal Semantic Analysis [Radinsky et al. '11]

21

1. Novel temporal representation of text



2. Novel temporal text-similarity measurement



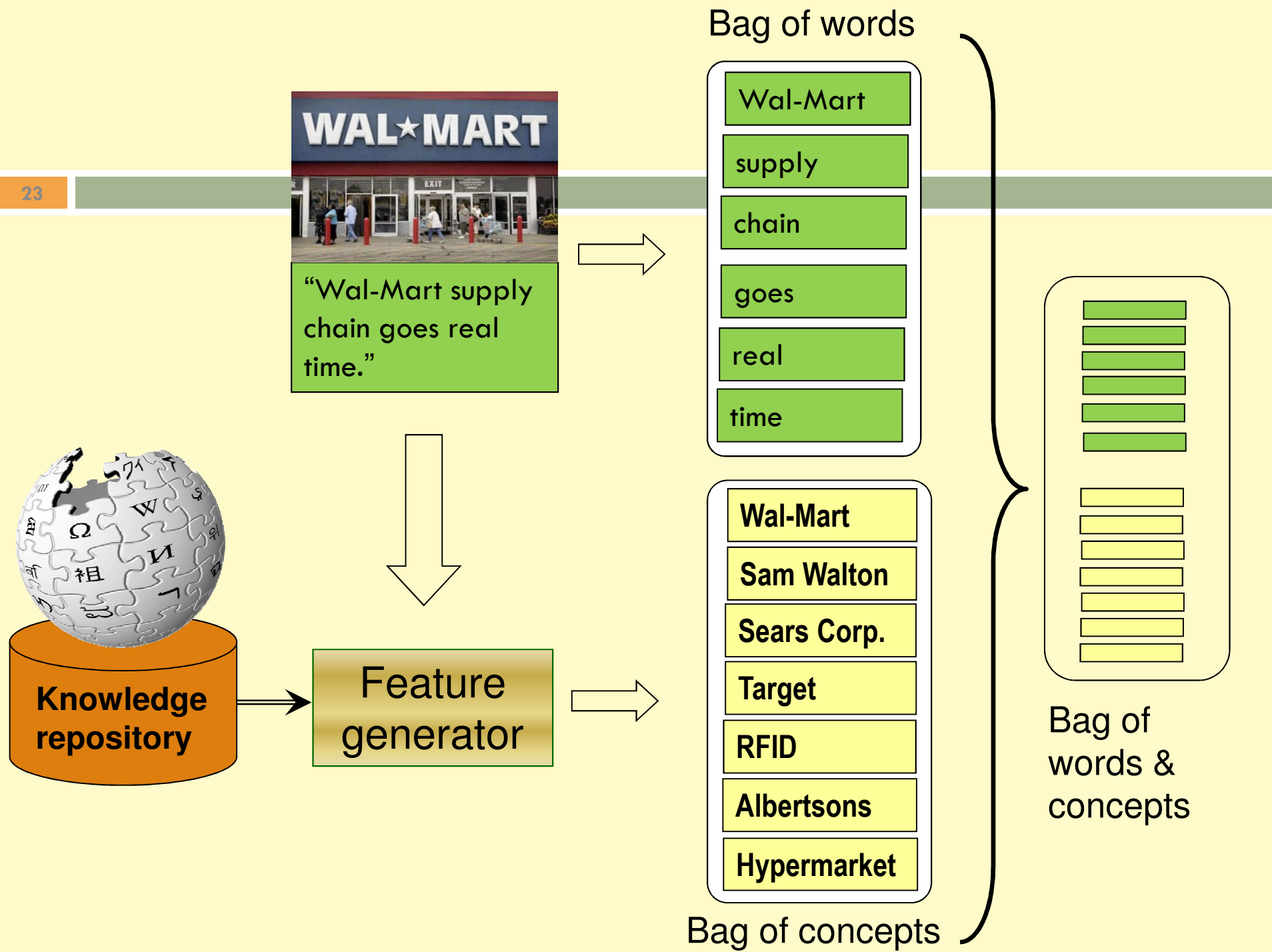
Concept-based information retrieval with Explicit Semantic Analysis

[Gabrilovich & Markovitch' 09; Egozi et al. '11]



22

- Comparing words / short texts → use concepts
- Comparing longer texts → use the **bag of words** & concepts
- **Text categorization:** assigning category labels to documents
 - ▣ Email filtering, news routing, classifying Web pages
- **Information retrieval:** find documents relevant to a query
 - ▣ Searching document collections (e.g., Web search)

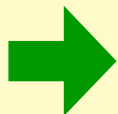


Cross-lingual Explicit Semantic Analysis (CL-ESA)

[Potthast et al., '08; Sorg & Cimiano '08]

24

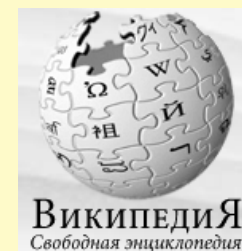
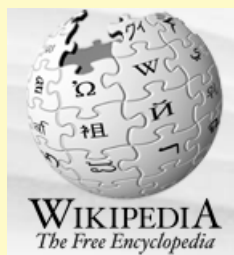
“Scientific
methods in
biology”



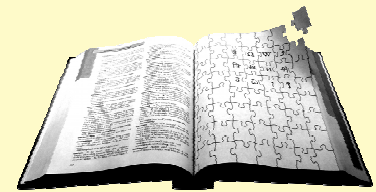
Biology
Scientific classification
Science
Chemical biology
Binomial nomenclature
Nature (journal)
Social sciences
Philosophy of biology
Scientist
History of biology



Биология
Биологическая систематика
Наука
No article or no link yet
Биномиальная номенклатура
Nature
Общественные науки
No article or no link yet
Учёный
История биологии



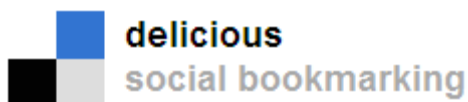
Summary



WIKTIONARY
the free dictionary

25

- **Exogenous knowledge** is crucial for language understanding
- **Collaboratively generated content** contains vast amounts of world knowledge
- **Concept-based representation** addresses two key problems: **synonymy** and **polysemy**
- Numerous other resources await to be tapped!
- Future work: Obtaining semantic clues from behavioral information (e.g., the authoring process)



Thank you!

gabr@yahoo-inc.com

<http://research.yahoo.com/~gabr>