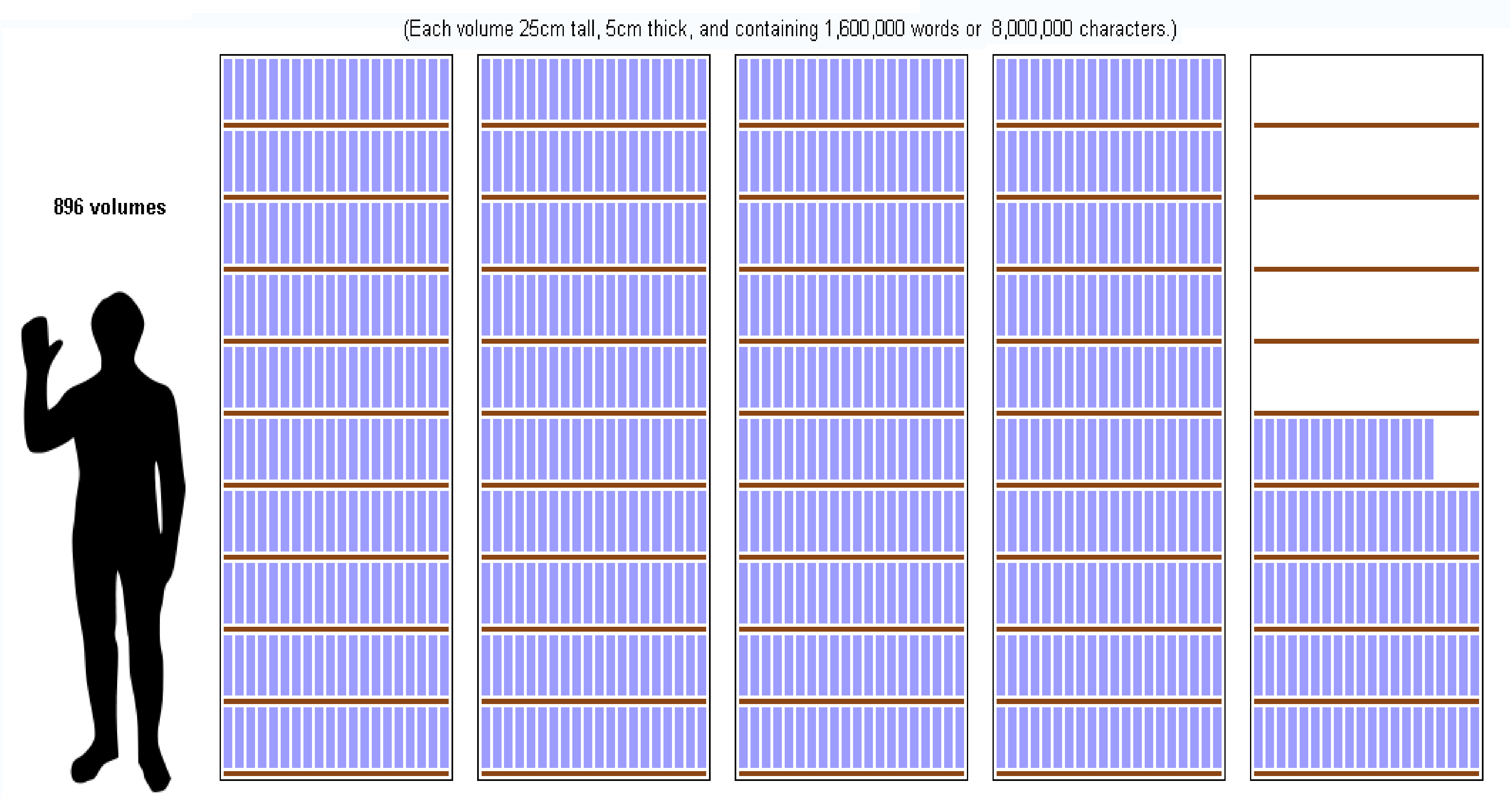
Wikipedia

- Wikipedia is available in dozens of languages,

- Its English version is the largest of all with 400+ million words in over one million articles
  - compared to 44 million words in 65,000 articles in Encyclopaedia Britannica.

- Interestingly, the open editing approach yields remarkable quality
  - a recent study [Giles,2005] found Wikipedia accuracy to rival that of Britannica.

The size of Wikipedia


(Each volume 25cm tall, 5cm thick, and containing 1,600,000 words or 8,000,000 characters.)
896 volumes

Concepts based on Wikipedia

- Explicit Semantic Analysis : an approach to representing semantics of natural language texts using natural concepts.

- A uniform way for computing relatedness of both individual words and arbitrarily long text fragments.

- The results of using ESA for computing semantic relatedness of texts are superior to the existing state of the art.

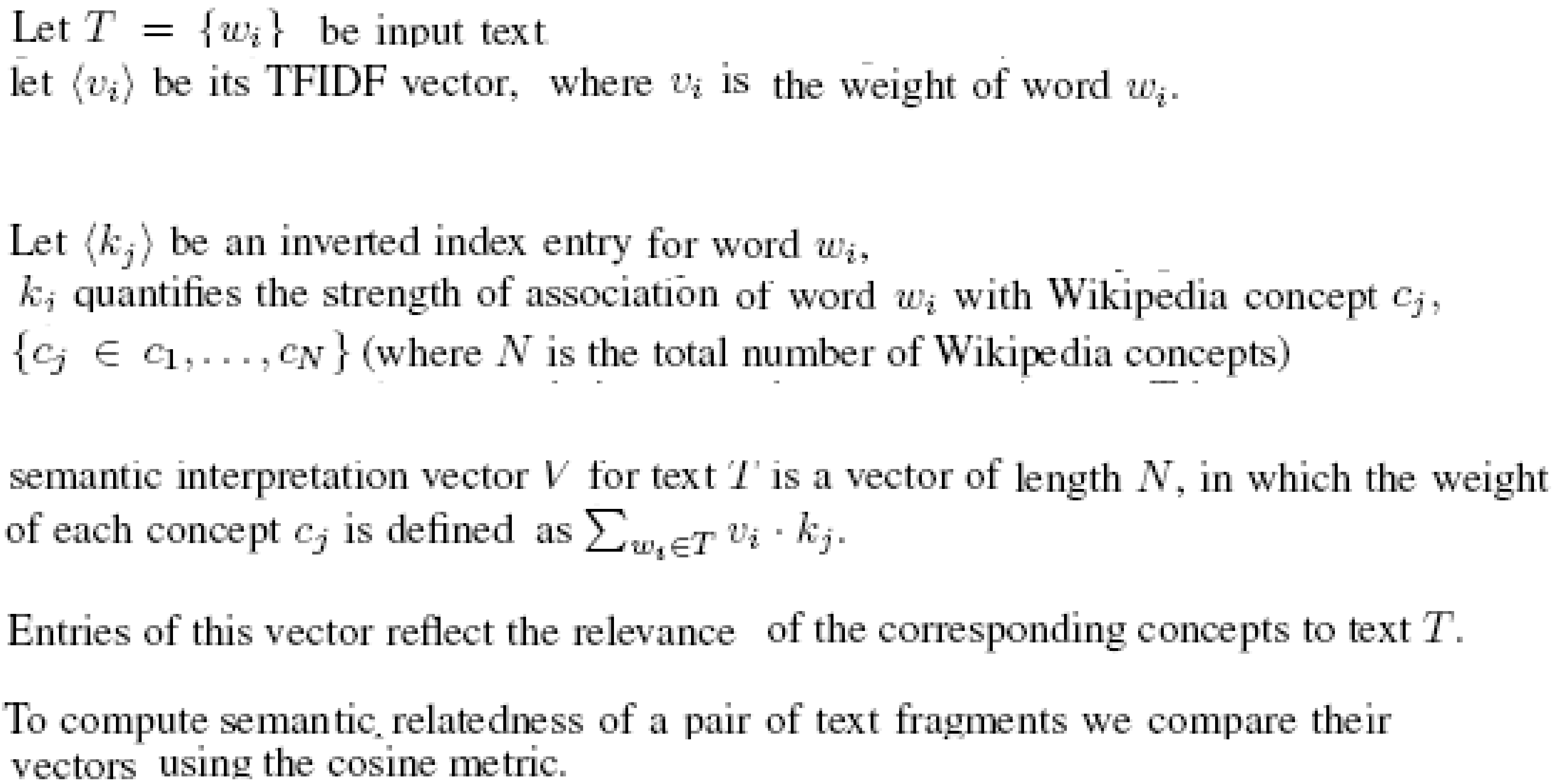
Overview

- Each Wikipedia article defines a concept.
- Examples : Computer Science, India
- Texts are represented as weighetd vectors of concepts called interpretation vectors
- These vectors can be compared using the cosine measure

Mapping Words to Concepts

- Each Wikipedia concept is represented as an attribute vector of words that occur in the corresponding article.
- Entries of these vectors are assigned weights using TFIDF scheme.
- These weights quantify the strength of association between words and concepts.
- To speed up semantic interpretation, an *inverted index* is used, which maps each word into a list of concepts in which it appears.

Comparing texts using Concepts



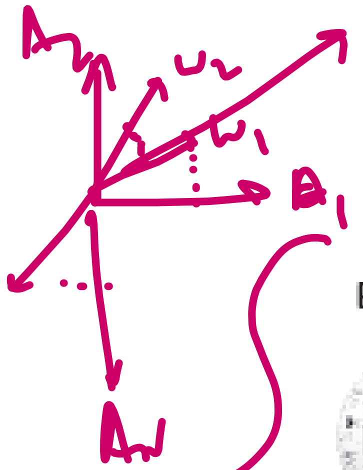
Let $T = \{w_i\}$ be input text
let $\langle v_i \rangle$ be its TFIDF vector, where v_i is the weight of word w_i.

Let $\langle k_j \rangle$ be an inverted index entry for word w_i,
k_j quantifies the strength of association of word w_i with Wikipedia concept c_j,
$\{c_j \in c_1, \ldots, c_N\}$ (where N is the total number of Wikipedia concepts)

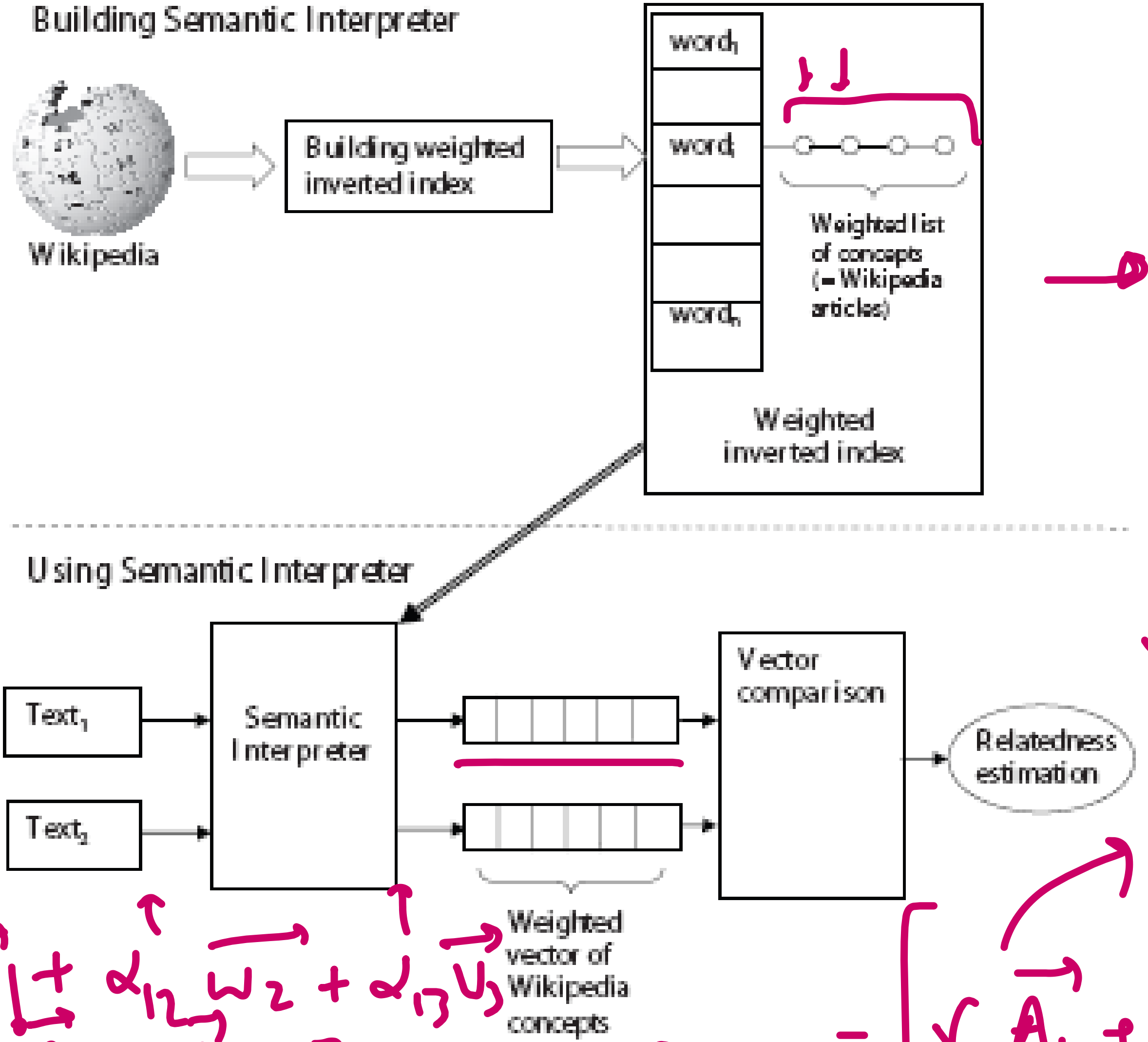
semantic interpretation vector V for text T is a vector of length N, in which the weight of each concept c_j is defined as $\sum_{w_i \in T} v_i \cdot k_j$.

Entries of this vector reflect the relevance of the corresponding concepts to text T.

To compute semantic relatedness of a pair of text fragments we compare their vectors using the cosine metric.



$\{$ Article → Concept
Synset → Concept $\}$


Building Semantic Interpreter

Wikipedia

Building weighted inverted index

word$_1$

word$_i$

word$_n$

Weighted list of concepts (= Wikipedia articles)

Weighted inverted index

Using Semantic Interpreter

Text$_1$

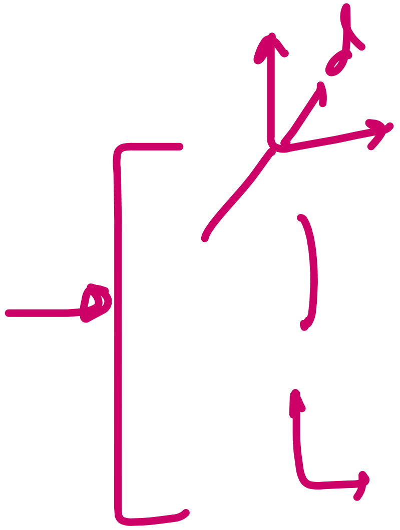
Text$_2$

Semantic Interpreter

Vector comparison

Relatedness estimation

Weighted vector of Wikipedia concepts

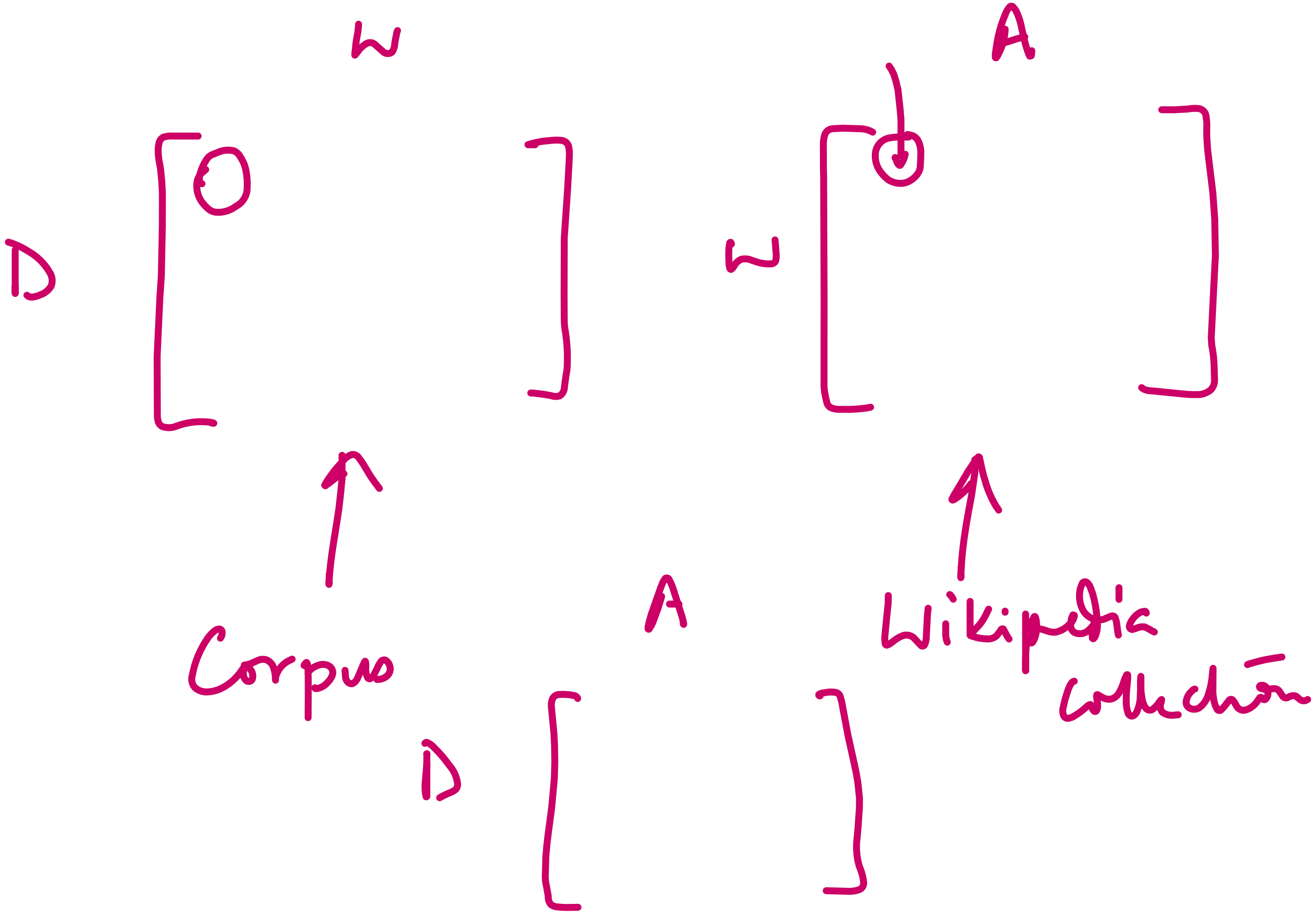


$r_1 = \alpha_{11}\beta_{11} + \alpha_{12}\beta_{12}$

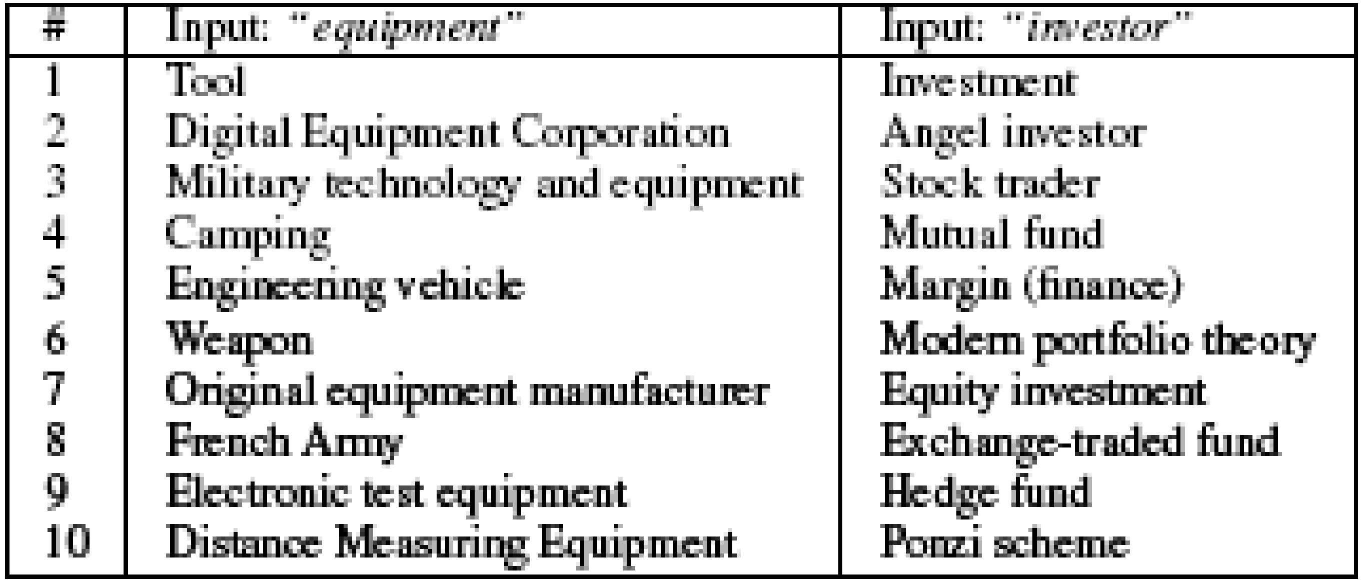
$D_1 = \alpha_{11}\vec{w_1} + \alpha_{12}\vec{w_2} + \alpha_{13}\vec{w_3}$

$= \alpha_{11}[\beta_{11}A_1 + \beta_{12}A_2 + \cdots \beta_N \vec{A_N}] + \cdots [$

$= [r_1\vec{A_1} + r_2\vec{A_2} + \cdots + r_N\vec{A_N}]$

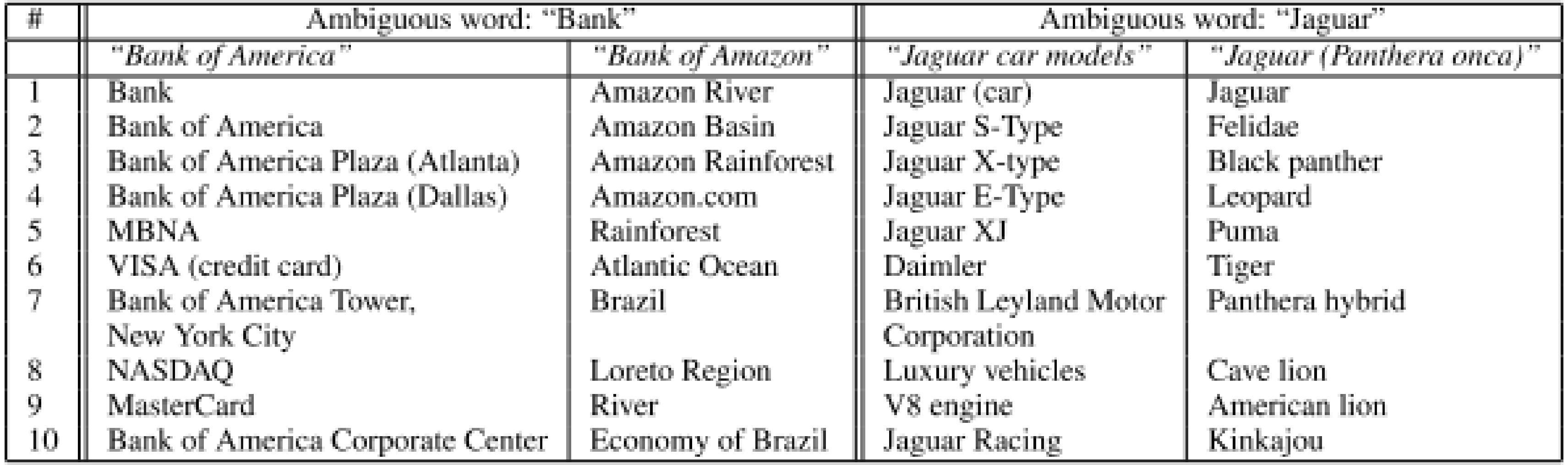

W
D
A
W
A
D
Corpus
Wikipedia collection

First 10 concepts in sample interpretation vectors



#	Input: *"equipment"*	Input: *"investor"*
1	Tool	Investment
2	Digital Equipment Corporation	Angel investor
3	Military technology and equipment	Stock trader
4	Camping	Mutual fund
5	Engineering vehicle	Margin (finance)
6	Weapon	Modern portfolio theory
7	Original equipment manufacturer	Equity investment
8	French Army	Exchange-traded fund
9	Electronic test equipment	Hedge fund
10	Distance Measuring Equipment	Ponzi scheme

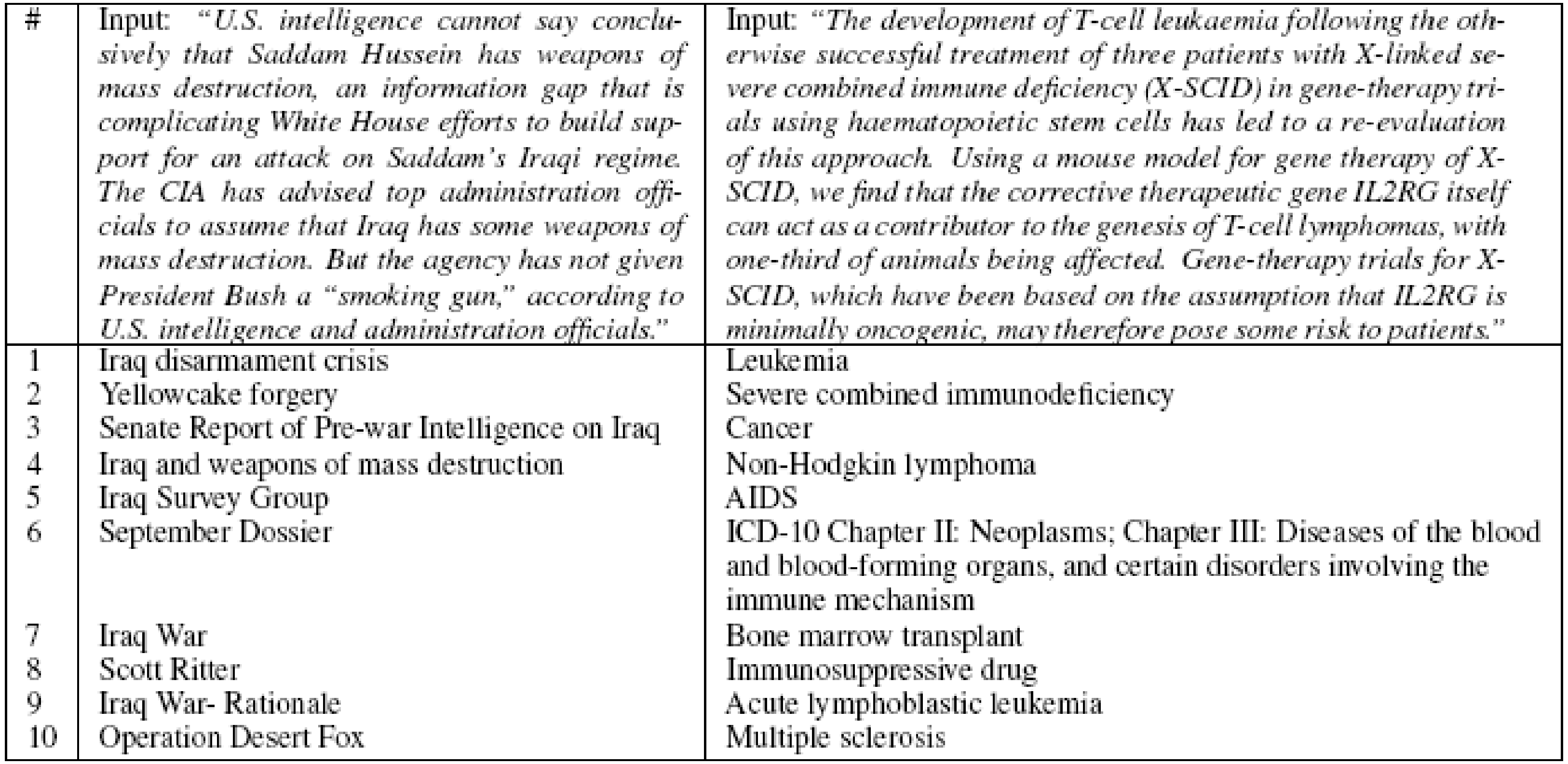
Disambiguation



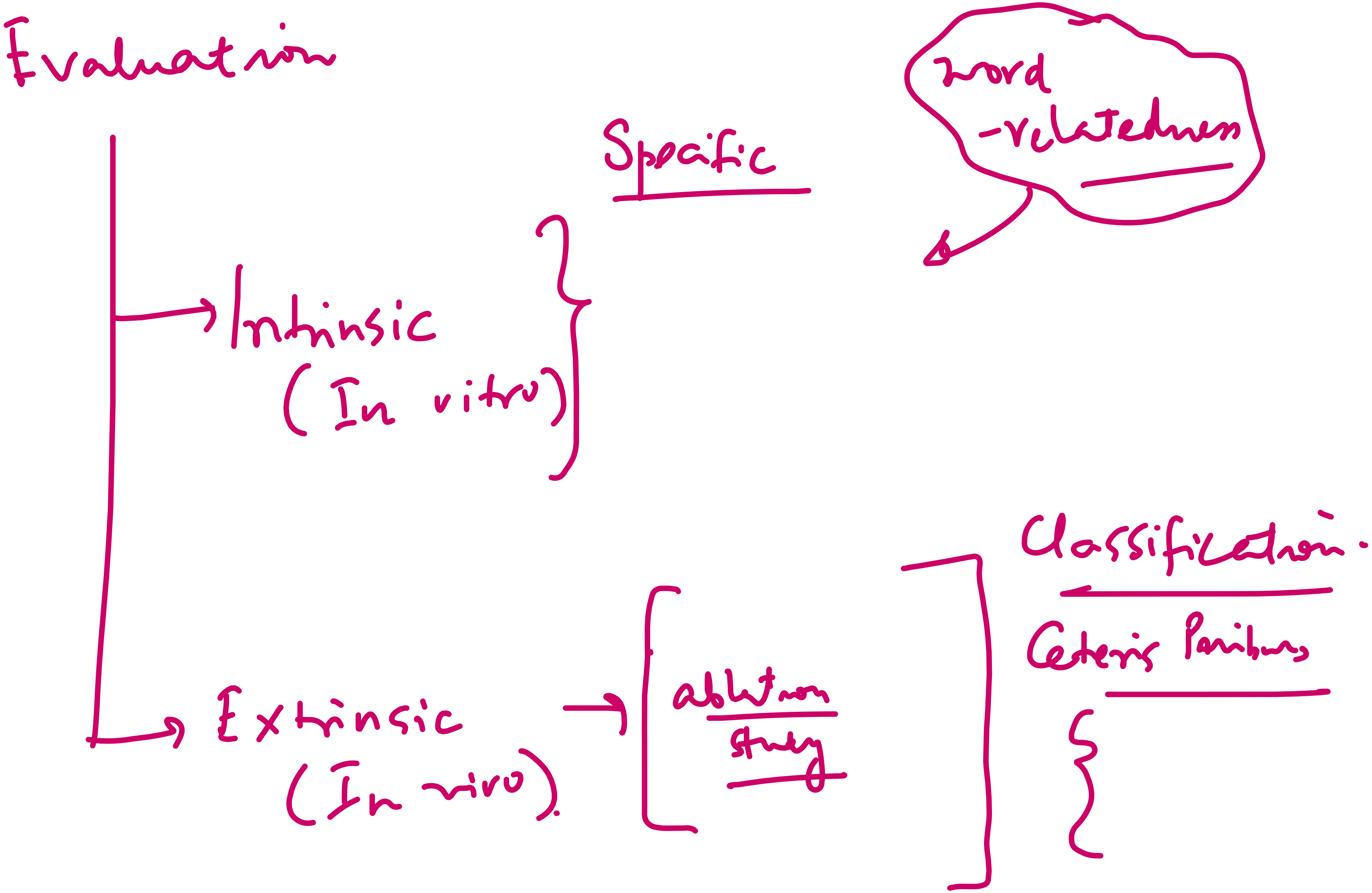
#	Ambiguous word: “Bank”		Ambiguous word: “Jaguar”	
	“Bank of America”	*“Bank of Amazon”*	*“Jaguar car models”*	*“Jaguar (Panthera onca)”*
1	Bank	Amazon River	Jaguar (car)	Jaguar
2	Bank of America	Amazon Basin	Jaguar S-Type	Felidae
3	Bank of America Plaza (Atlanta)	Amazon Rainforest	Jaguar X-type	Black panther
4	Bank of America Plaza (Dallas)	Amazon.com	Jaguar E-Type	Leopard
5	MBNA	Rainforest	Jaguar XJ	Puma
6	VISA (credit card)	Atlantic Ocean	Daimler	Tiger
7	Bank of America Tower, New York City	Brazil	British Leyland Motor Corporation	Panthera hybrid
8	NASDAQ	Loreto Region	Luxury vehicles	Cave lion
9	MasterCard	River	V8 engine	American lion
10	Bank of America Corporate Center	Economy of Brazil	Jaguar Racing	Kinkajou

First ten concepts of the interpretation vectors for texts with ambiguous words.

First ten concepts of the interpretation vectors for sample text fragments



#	Input: *"U.S. intelligence cannot say conclusively that Saddam Hussein has weapons of mass destruction, an information gap that is complicating White House efforts to build support for an attack on Saddam's Iraqi regime. The CIA has advised top administration officials to assume that Iraq has some weapons of mass destruction. But the agency has not given President Bush a "smoking gun," according to U.S. intelligence and administration officials."*	Input: *"The development of T-cell leukaemia following the otherwise successful treatment of three patients with X-linked severe combined immune deficiency (X-SCID) in gene-therapy trials using haematopoietic stem cells has led to a re-evaluation of this approach. Using a mouse model for gene therapy of X-SCID, we find that the corrective therapeutic gene IL2RG itself can act as a contributor to the genesis of T-cell lymphomas, with one-third of animals being affected. Gene-therapy trials for X-SCID, which have been based on the assumption that IL2RG is minimally oncogenic, may therefore pose some risk to patients."*
1	Iraq disarmament crisis	Leukemia
2	Yellowcake forgery	Severe combined immunodeficiency
3	Senate Report of Pre-war Intelligence on Iraq	Cancer
4	Iraq and weapons of mass destruction	Non-Hodgkin lymphoma
5	Iraq Survey Group	AIDS
6	September Dossier	ICD-10 Chapter II: Neoplasms; Chapter III: Diseases of the blood and blood-forming organs, and certain disorders involving the immune mechanism
7	Iraq War	Bone marrow transplant
8	Scott Ritter	Immunosuppressive drug
9	Iraq War- Rationale	Acute lymphoblastic leukemia
10	Operation Desert Fox	Multiple sclerosis



Evaluation

Intrinsic
(In vitro)

Specific

word
-relatedness

Extrinsic
(In vivo)

ablation
study

Classification.

Ceteris Paribus

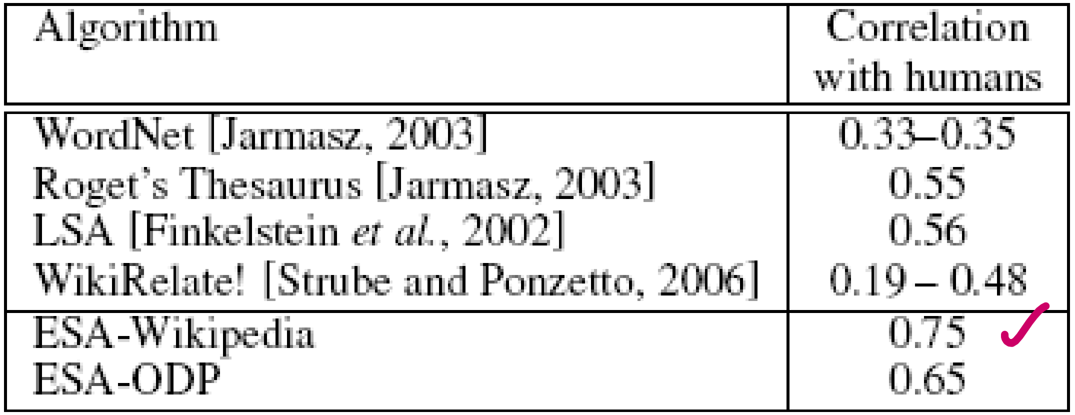
Evaluation

WordSim 353

353 word pairs.

13-16

A₁ > A₂ on T on E on D ass. 5

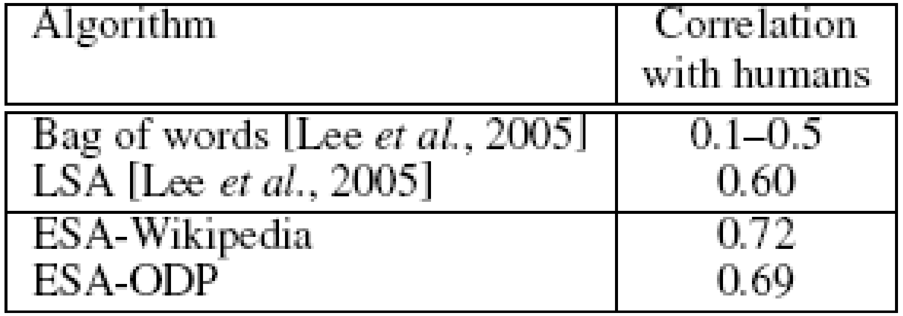


Algorithm	Correlation with humans
WordNet [Jarmasz, 2003]	0.33–0.35
Roget's Thesaurus [Jarmasz, 2003]	0.55
LSA [Finkelstein *et al.*, 2002]	0.56
WikiRelate! [Strube and Ponzetto, 2006]	0.19 – 0.48
ESA-Wikipedia	0.75 ✓
ESA-ODP	0.65

Computing word relatedness

interpretability

effectiveness

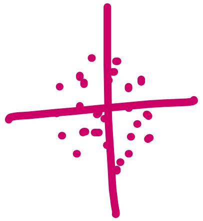


Algorithm	Correlation with humans
Bag of words [Lee *et al.*, 2005]	0.1–0.5
LSA [Lee *et al.*, 2005]	0.60
ESA-Wikipedia	0.72
ESA-ODP	0.69

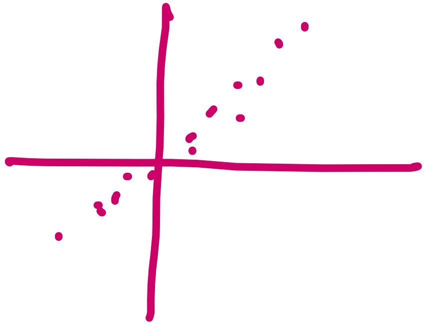
Computing text relatedness

Spearman Correlation Coeffn.

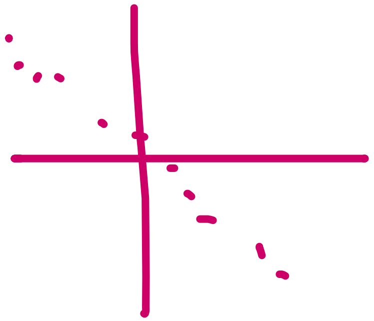
$$\left\{ \text{Pearson Corr. Coeff} \right\} \times \left[\frac{Cov(x,y)}{\sigma_x \sigma_y} \right]$$



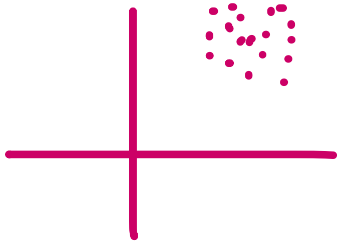
$\boxed{0}$



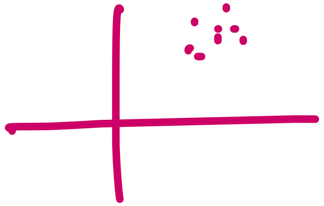
→ +1



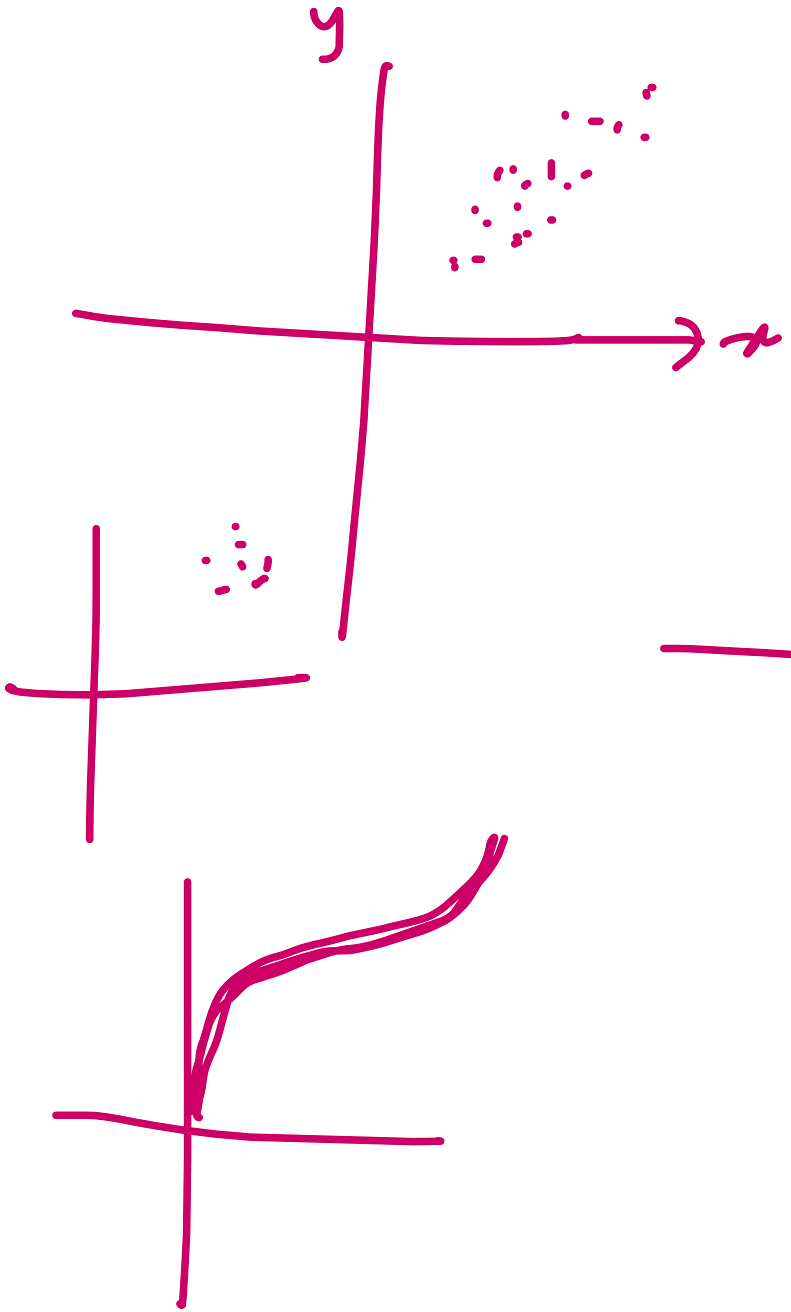
→ -1

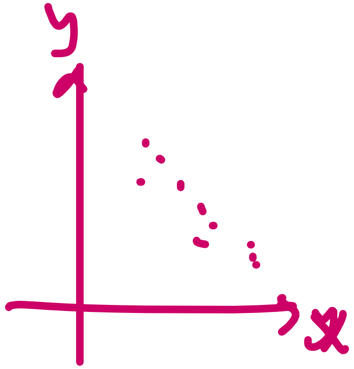


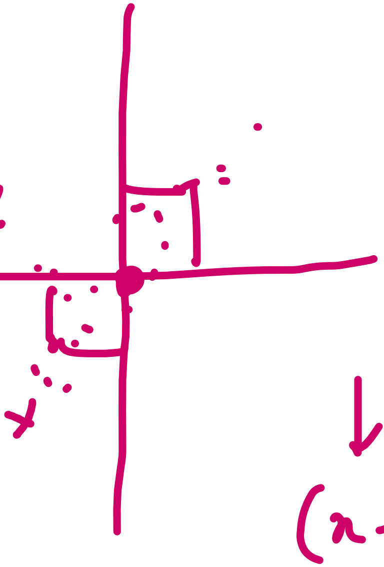
$$\sum (x_i - \bar{x})(y_i - \bar{y})$$



[-1, 1]

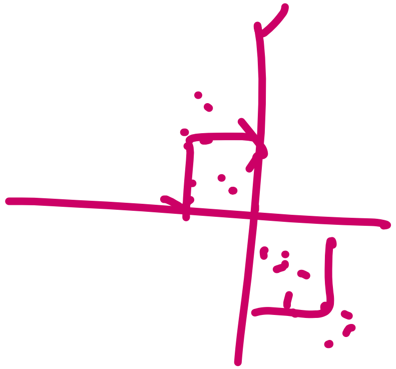


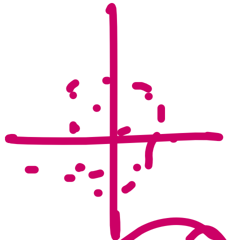




large positive value

$$(x - \bar{x})(y - \bar{y})$$





0

$$\underbrace{-ve \qquad 0 \qquad +ve}_{[-1,\ +1]}$$

1, 2, 3. 353.

Human	123	150	1	-	-	-	-
	↕	↕					↓
ESA	120	90	5				

$\boxed{d_i}$

Spearman Correlation Coeff^n

	1	2	3	3	5
H	5	4	3	2	1
E	1	2	3	4	5

[ ]

$$\left[1 - \frac{6\sum d_i^2}{n(n^2-1)}\right].$$

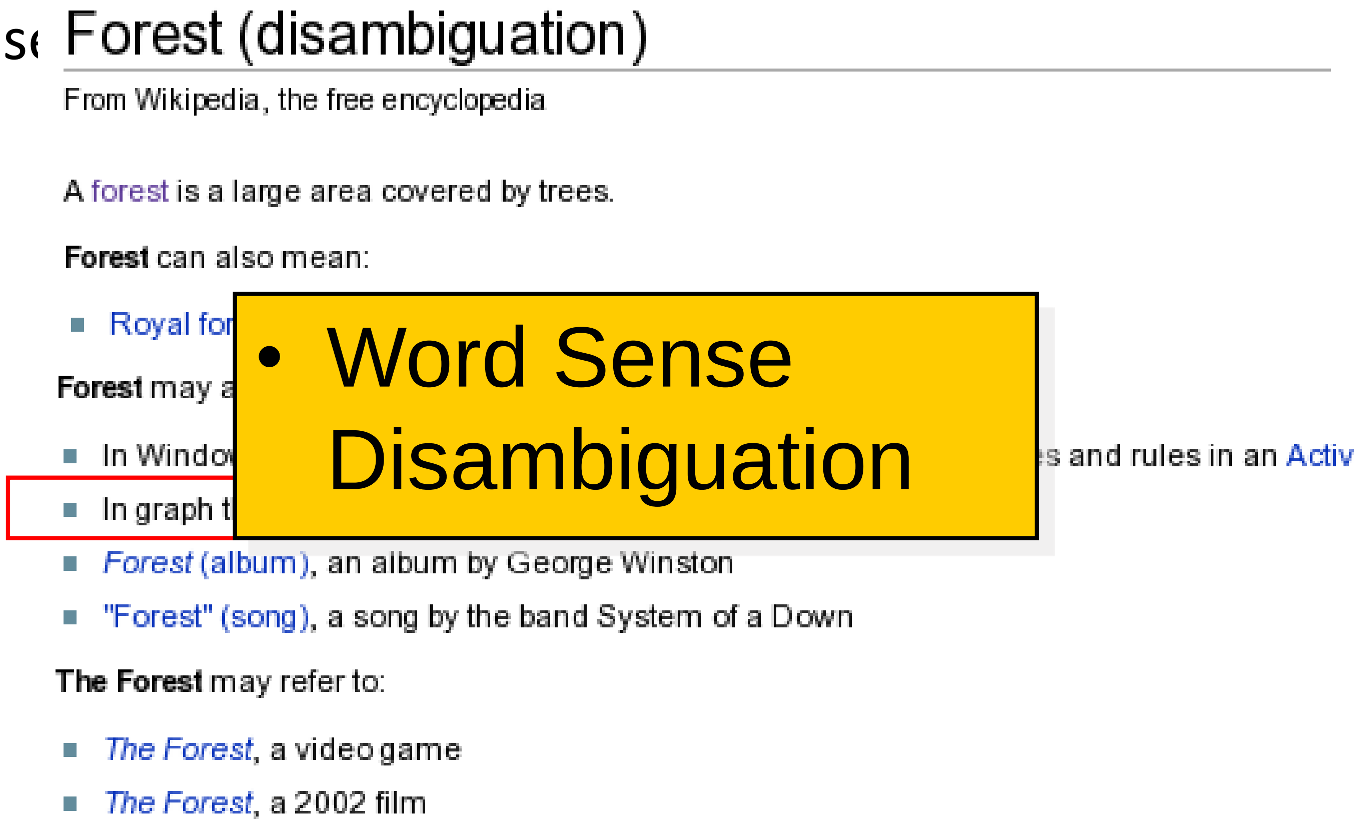
↓

Open Questions

- Concepts beyond words?
    - Child knowledge acquisition
- Can we *grow* concept descriptions instead of *build*ing them?

Wikipedia – Disambiguation pages

- Sense inventory
  - Domain specific se



Forest (disambiguation)

From Wikipedia, the free encyclopedia

A forest is a large area covered by trees.

Forest can also mean:

- Royal for

Forest may a

- In Windo s and rules in an Activ
- In graph t
- *Forest* (album), an album by George Winston
- "Forest" (song), a song by the band System of a Down

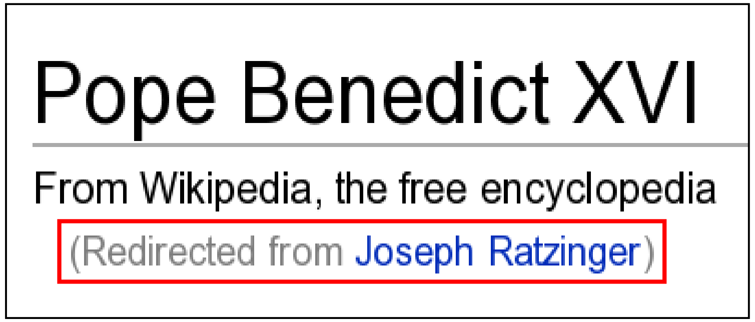
The Forest may refer to:

- *The Forest*, a video game
- *The Forest*, a 2002 film

- Word Sense Disambiguation

Wikipedia – Redirect pages

- Synonyms
  - *Pope Benedict XVI*
  - *Joseph Ratzinger*
  - *Joseph Cardinal Ratzinger*
- Spelling variations
  - *Benedict the Sixteenth*
  - *Benedict the 16th*
  - *Benedict 1*
  - *Benedict 1*
  - *Benedict X*
  - *Benedict x*
- Misspellings
  - *Josef Ratzi*
- Abbreviations
  - *PB16*



Pope Benedict XVI

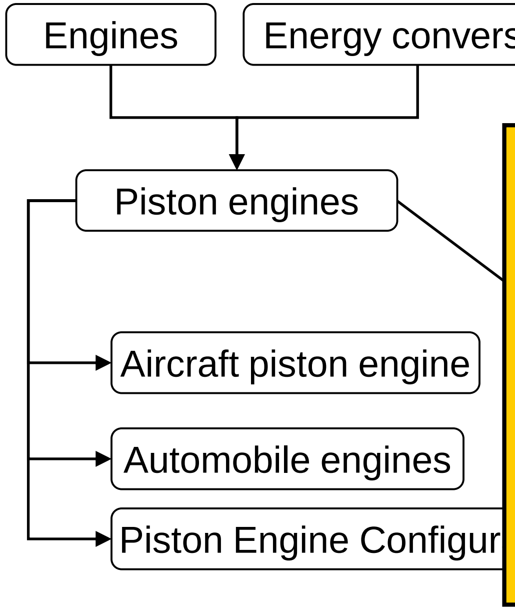
From Wikipedia, the free encyclopedia

(Redirected from Joseph Ratzinger)

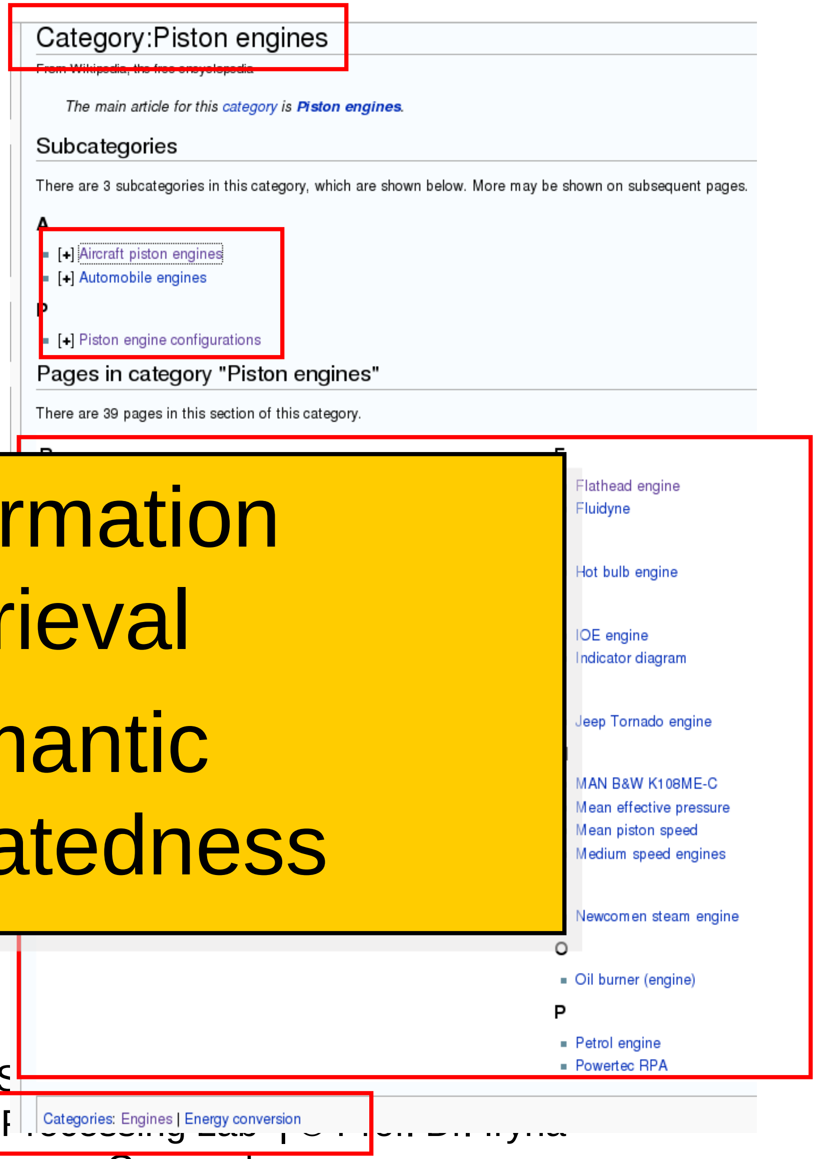
- Named Entity Recognition
- Co-reference Resolution

Wikipedia – Categories

- Articles
- Hierarchy



- Information Retrieval
- Semantic Relatedness



Category:Piston engines

From Wikipedia, the free encyclopedia

*The main article for this category is **Piston engines**.*

Subcategories

There are 3 subcategories in this category, which are shown below. More may be shown on subsequent pages.

A

- [+] Aircraft piston engines
- [+] Automobile engines

P

- [+] Piston engine configurations

Pages in category "Piston engines"

There are 39 pages in this section of this category.

- Flathead engine
- Fluidyne
- Hot bulb engine
- IOE engine
- Indicator diagram
- Jeep Tornado engine
- MAN B&W K108ME-C
- Mean effective pressure
- Mean piston speed
- Medium speed engines
- Newcomen steam engine

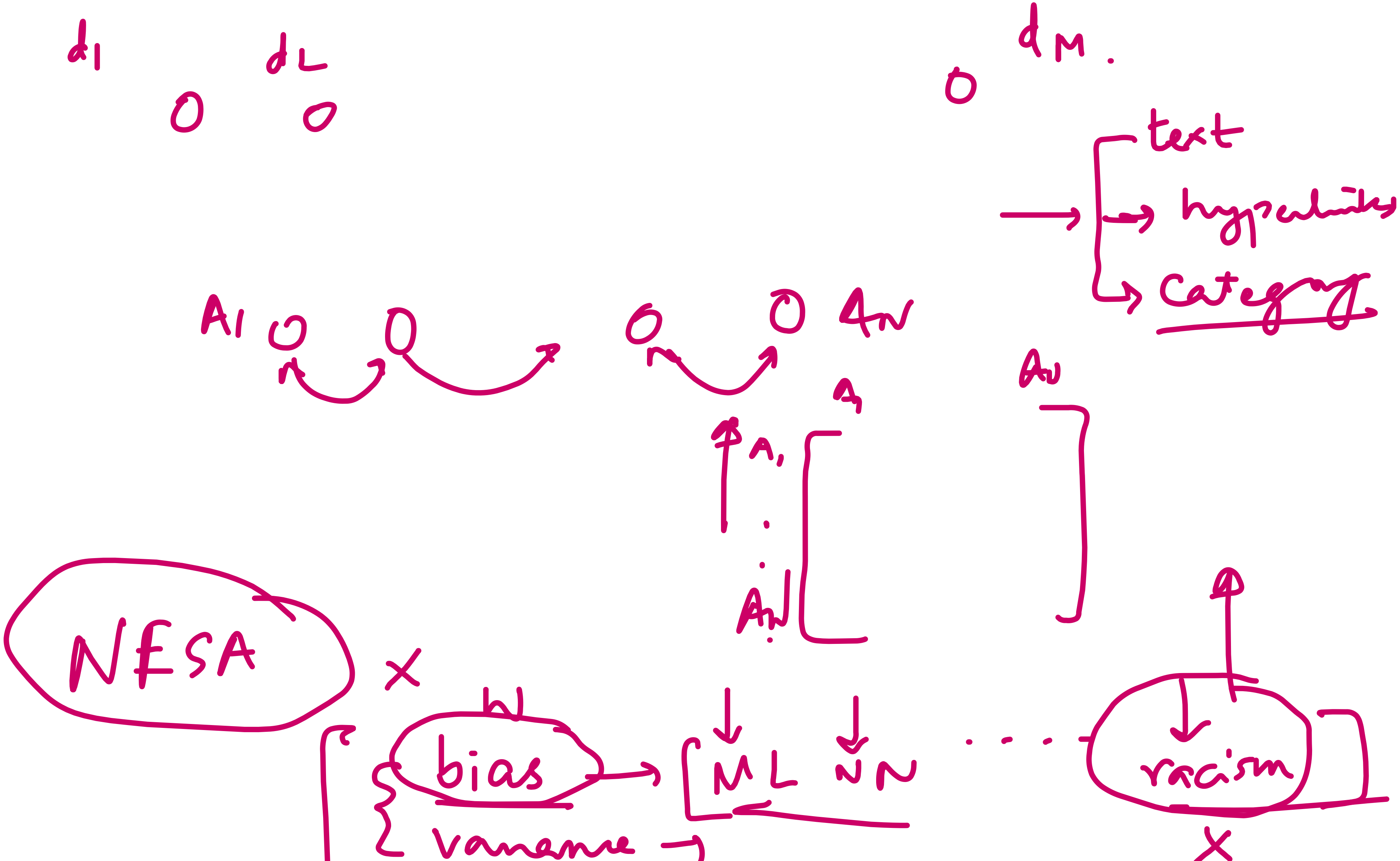
O

- Oil burner (engine)

P

- Petrol engine
- Powertec RPA

Categories: Engines | Energy conversion


d_1
d_L
d_M.
text
hyperlinks
category
A_1
A_N
A_N
A_1
A_1
A_N
NESA
X
W
bias
variance
ML
NN
racism
X