POLITÉCNICA
"Ingeniamos el futuro"

CAMPUS
DE EXCELENCIA
INTERNACIONAL

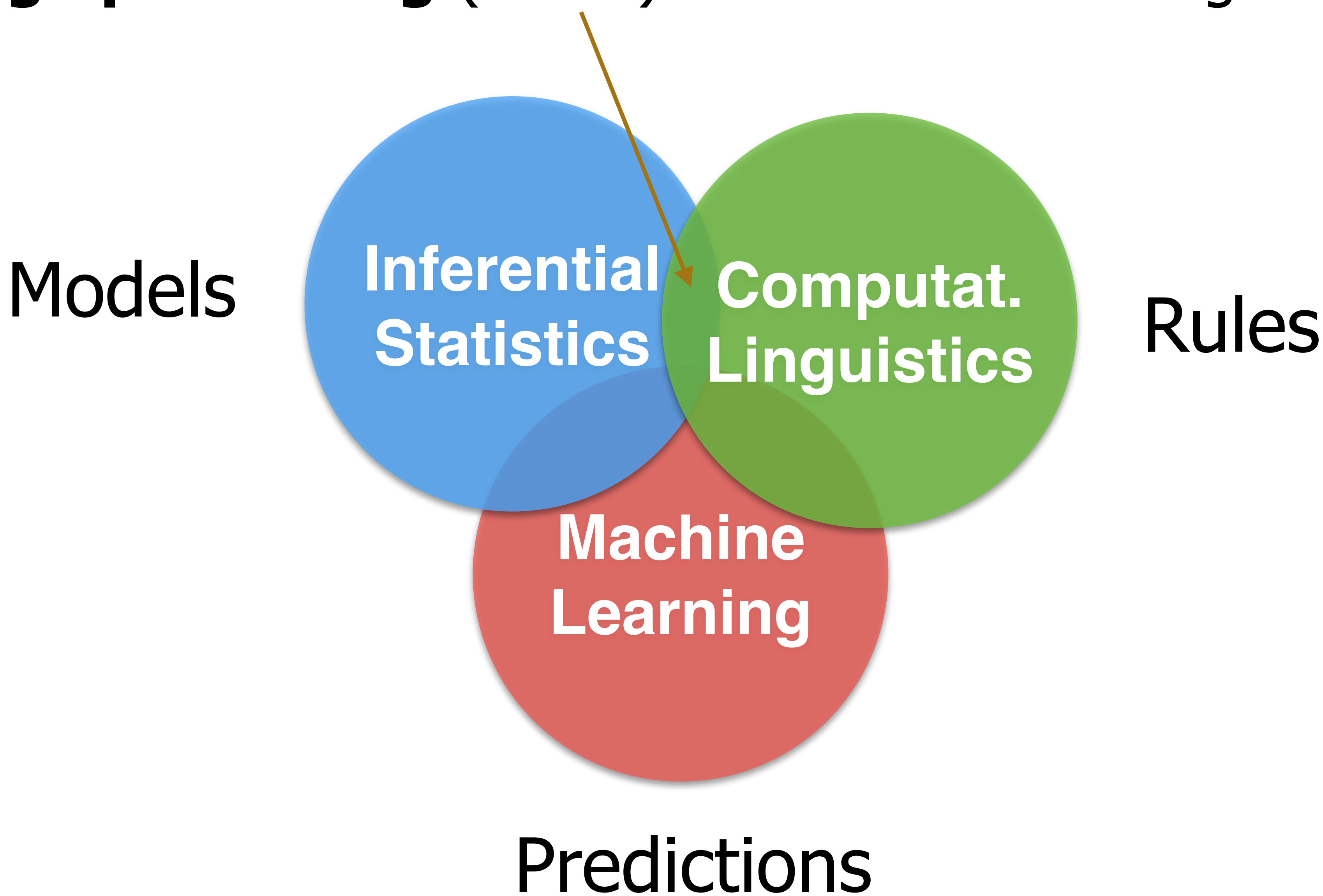
Text Mining 1
Introduction

9th Madrid Summer School (2014) on
Advanced Statistics and Data Mining

Florian Leitner
florian.leitner@upm.es

"Text Mining" or "Text Analytics"

The discovery of {new or existing} facts by applying **natural language processing** ("NLP") & statistical learning techniques.



Language Understanding = Artificial Intelligence ?

"Her" Movie, 2013

"Watson" & "CRUSH" IBM's future bet: Mainframes & AI

"The Singularity" Ray Kurzweil
(Google's director of engineering)

"predict crimes before they happen"

cognitive computing:
"processing information more like a human than a machine"

Examples of Language Processing Applications

Text Mining

Spam filtering

Document Classification

Date/time event detection

Information Extraction

(Web) Search engines

Information Retrieval

Watson in Jeopardy! (IBM)

Question Answering

Twitter brand monitoring

Sentiment Analysis (Stat. NLP)

Language Processing

Siri (Apple) and Google Now

Language Understanding

Spelling Correction

Statistical Language Modeling

Website translation (Google)

Machine Translation

"Clippy" Assistant (Microsoft)

Dialog System

Finding similar items (Amazon)

Recommender System

Current Topics in Text Mining

Course requirements…

Basic Linear Algebra and Probability Theory; Computer Savvy

You will learn about…

- Language Modeling
- String Processing
- Text Classification
- Information Extraction

Other topics…

- Information Retrieval
- Question Answering
- Dialogue Systems
- Text Summarization
- Machine Translation
- Language Understanding

Words, Tokens, Shingles, and N-Grams

Character-based,
Regular Expressions,
Probabilistic, ...

Text with words: This is a sentence.

"tokenization"

Tokens: This | is | a | sentence | NB: .

Token **N-Grams**

2-Shingles: This is | is a | a sentence | sentence .

a.k.a. **k-Shingling**

3-Shingles: This is a | is a sentence | a sentence .

Character **N-Grams**: all **trigrams** of "sentence":
[sen, ent, nte, ten, enc, nce]

Beware: the terms "k-shingle" and "n-gram" are not used consistently...

Lemmatization, Part-of-Speech (PoS) Tagging, and Named Entity Recognition (NER)

PoS Tagset:
Penn Treebank

Token	Lemma	PoS	NER
Constitutive	constitutive	JJ	O
binding	**binding**	**NN**	O
to	to	TO	O
the	the	DT	O
peri-κ	**peri-kappa**	NN	**B-DNA**
B	B	NN	**I-DNA**
site	site	NN	**I-DNA**
is	**be**	VBZ	O
seen	**see**	VBN	O
in	in	IN	O
monocytes	**monocyte**	NNS	**B-cell**
.	.	.	O

B-I-O
NER Tagging

Information Retrieval (IR)

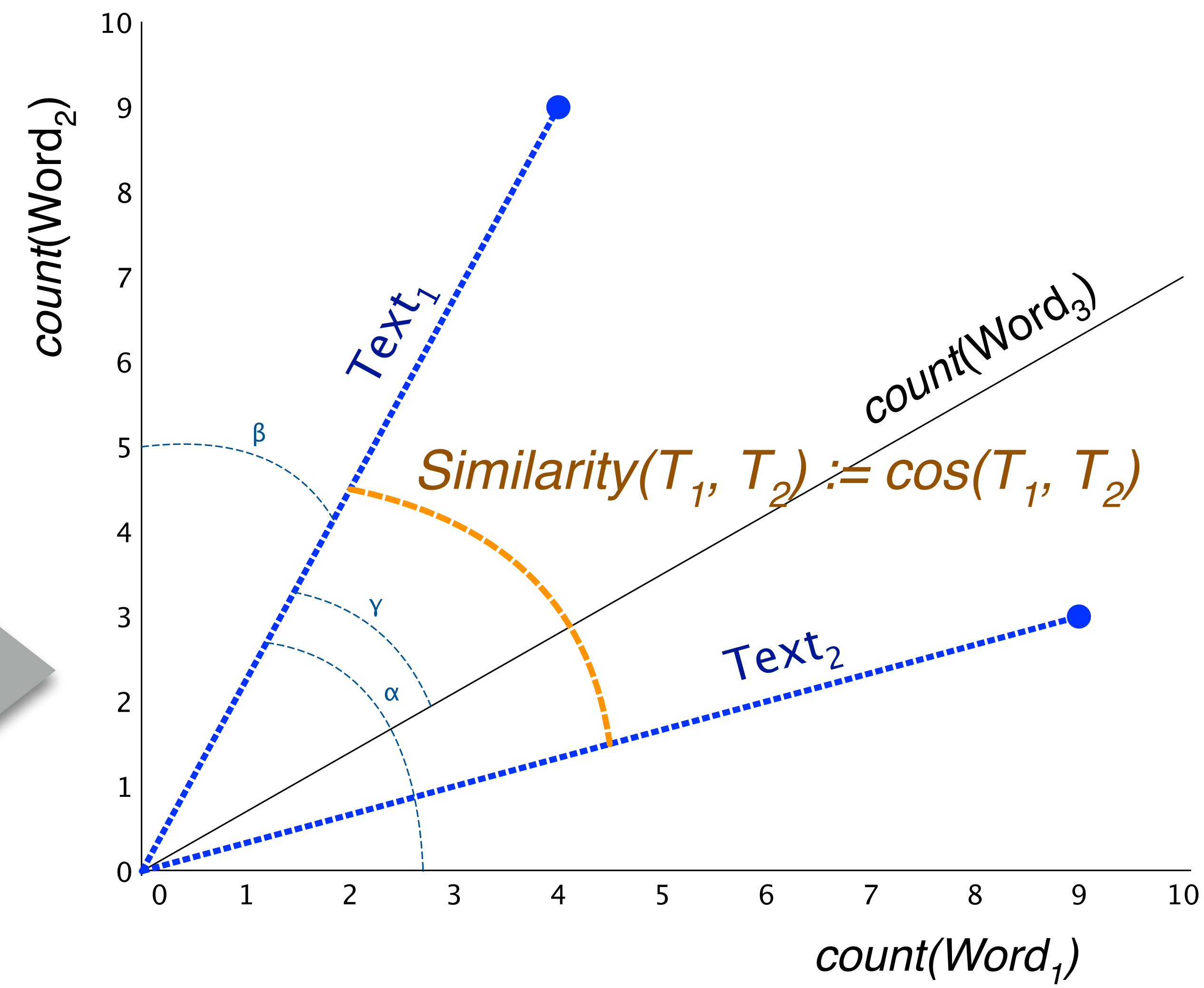
Text Vectorization: Inverted Index

Text 1: He that not wills to the end neither wills to the means.

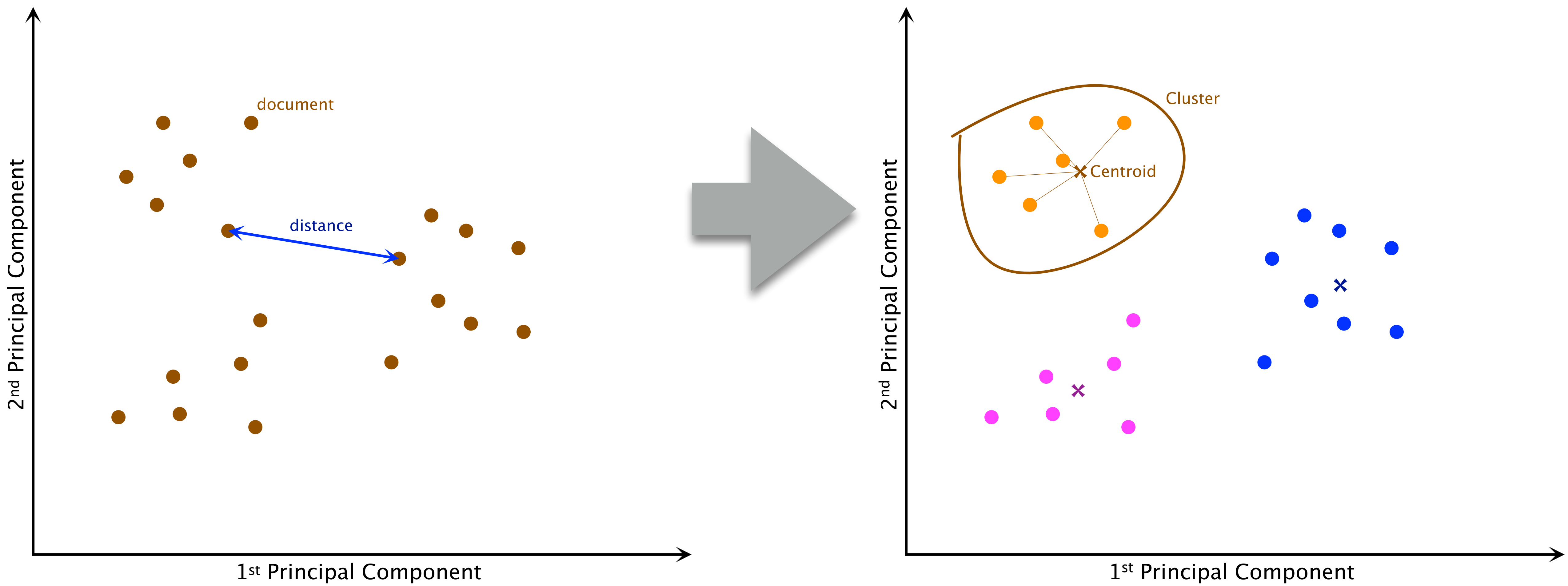
Text 2: If the mountain will not go to Moses, then Moses must go to the mountain.

tokens	Text 1	Text 2
end	1	0
go	0	2
he	1	0
if	0	1
means	1	0
Moses	0	2
mountain	0	2
must	0	1
not	1	1
that	1	0
the	2	2
then	0	1
to	2	2
will	2	1

Comparing Word Vectors: Cosine Similarity



Document Classification



Supervised (“Learning to Classify”, e.g., spam filtering)

vs.

Unsupervised (“Exploratory Grouping”, e.g., topic modeling)

Inverted (I-) Indices

factors, normalization (len[text]), probabilities, and n-grams

Text 1: He that not wills to the end neither wills to the means.

Text 2: If the mountain will not go to Moses, then Moses must go to the mountain.

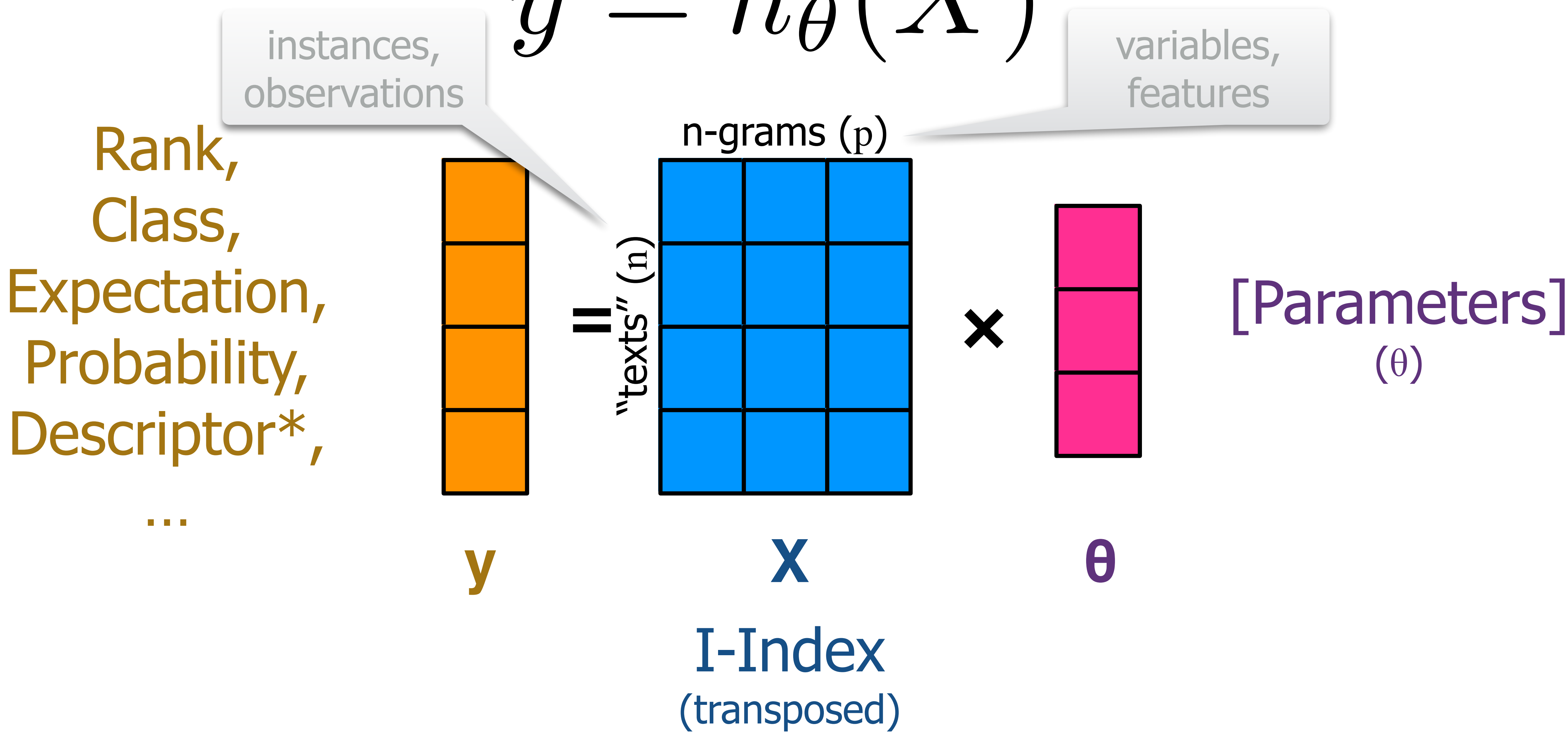
tokens	Text 1	Text 2
end	1	0
go	0	2
he	1	0
if	0	1
means	1	0
Moses	0	2
mountain	0	2
must	0	1
not	1	1
that	1	0
the	2	2
then	0	1
to	2	2
will	2	1

unigrams	T1	T2	p(T1)	p(T2)
end	1	**0**	0.09	0.00
go	**0**	2	0.00	0.13
he	1	**0**	0.09	0.00
if	**0**	1	0.00	0.07
means	1	**0**	0.09	0.00
Moses	**0**	2	0.00	0.13
mountain	**0**	2	0.00	0.13
must	**0**	1	0.00	0.07
not	1	1	0.09	0.07
that	1	**0**	0.09	0.00
the	2	2	0.18	0.13
then	**0**	1	0.00	0.07
to	2	2	0.18	0.13
will	2	1	0.18	0.07
SUM	11	15	1.00	1.00

bigrams	Text 1	Text 2
end, neither	1	**0**
go, to	**0**	2
he, that	1	**0**
if, the	**0**	1
Moses, must	**0**	1
Moses, then	**0**	1
mountain, will	**0**	1
must, go	**0**	1
not, go	**0**	1
not, will	1	**0**
that, not	1	**0**
the, means	1	**0**
the, mountain	**0**	2
then, Moses	**0**	1
to, Moses	**0**	1
to, the	2	1
will, not	**0**	1
will, to	2	**0**

I-Indices and the Central Dogma Machine Learning

$$y = h_\theta(X)$$



The Curse of Dimensionality

(RE Bellman, 1961) [inventor of dynamic programming]

- $\mathbf{p} \gg \mathbf{n}$ (more tokens/n-grams/features than texts/documents)
- Inverted indices (X) are very **sparse** matrices.
- Even with millions of training examples, **unseen** tokens will keep coming up in the "test set" or in production.
- In a high-dimensional hypercube, most instances are closer to the **face of the cube** ("nothing") than their nearest neighbor.

✓ Remedy: the "blessing of non-uniformity" ➡ dimensionality reduction (a.k.a. [low-dimensional] **embedding**)

- feature extraction: PCA, LDA, factor analysis, unsupervised classification of tokens based on their surrounding tokens ("word embedding"), …
- feature "reduction": locality sensitivity hashing, random projections, …

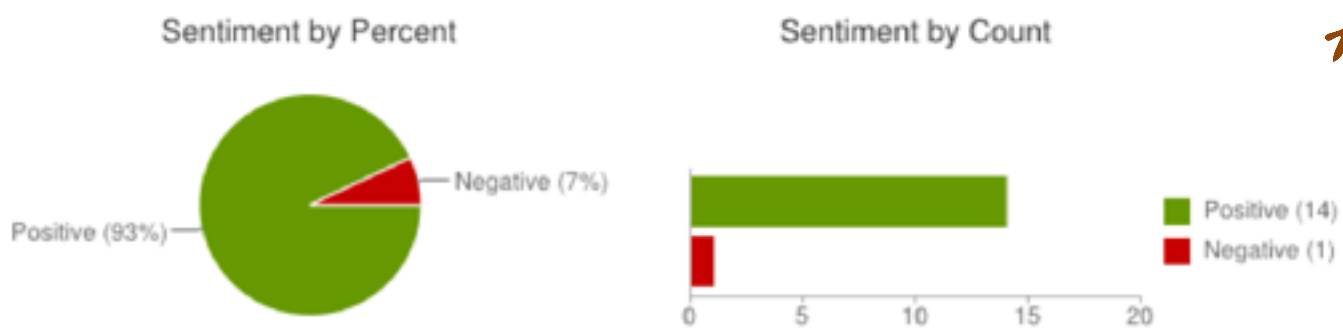
Sentiment Analysis

ElReyAbdica | Spanish | Search

http://www.sentiment140.com

feelings are complex and not black or white... (irony, negations)

Sentiment analysis for ElReyAbdica



andiuz: @cercodeartajona **#ElReyAbdica** tengo un fantasma que nadie conoce. A ver si lo encuentras... El rey nunca vino.
Posted: 1 minute ago

Elizabethtrip: Allá va la republicana Va pidiendo libertad Libre libre libertad Libre libre libertad Libertad y solidaridad! #ReferendumYA **#ElReyAbdica**
Posted: 1 minute ago

Nachette89: RT @horadelagente: Que dice ZP que reclamar #ReferendumYA es demagogia...y yo que pensaba que era DEMOCRACIA #IIIRepublica **#ElReyAbdica**
Posted: 1 minute ago

maideraretxe14: RT @YisucristEs: ya me estoy imaginando a Juan Carlos en el Burger King cn la corona recordando viejos tiempos. **#ElReyAbdica**
Posted: 1 minute ago

Ricitos11: RT @Malviviendo: RT si queréis al Kaki como nuevo Rey. **#elreyabdica** #melonesparatodos http://t.co/EnR0GPzYRx
Posted: 1 minute ago

marcelinelobita: RT @20m: **#ElReyAbdica** Más de 60 #manifestaciones en toda España reclamarán un #referéndum esta tarde http://t.co/xzRye3Dgi7 http://t.co/tus ?
Posted: 1 minute ago

tota_totita: RT @RPPNoticias: **#ElReyAbdica**: Así sería la línea de sucesión de la Corona española. (Vía .@el_pais) ? http://t.co/tO4dYkmO6O
Posted: 1 minute ago

Information Extraction (IE)

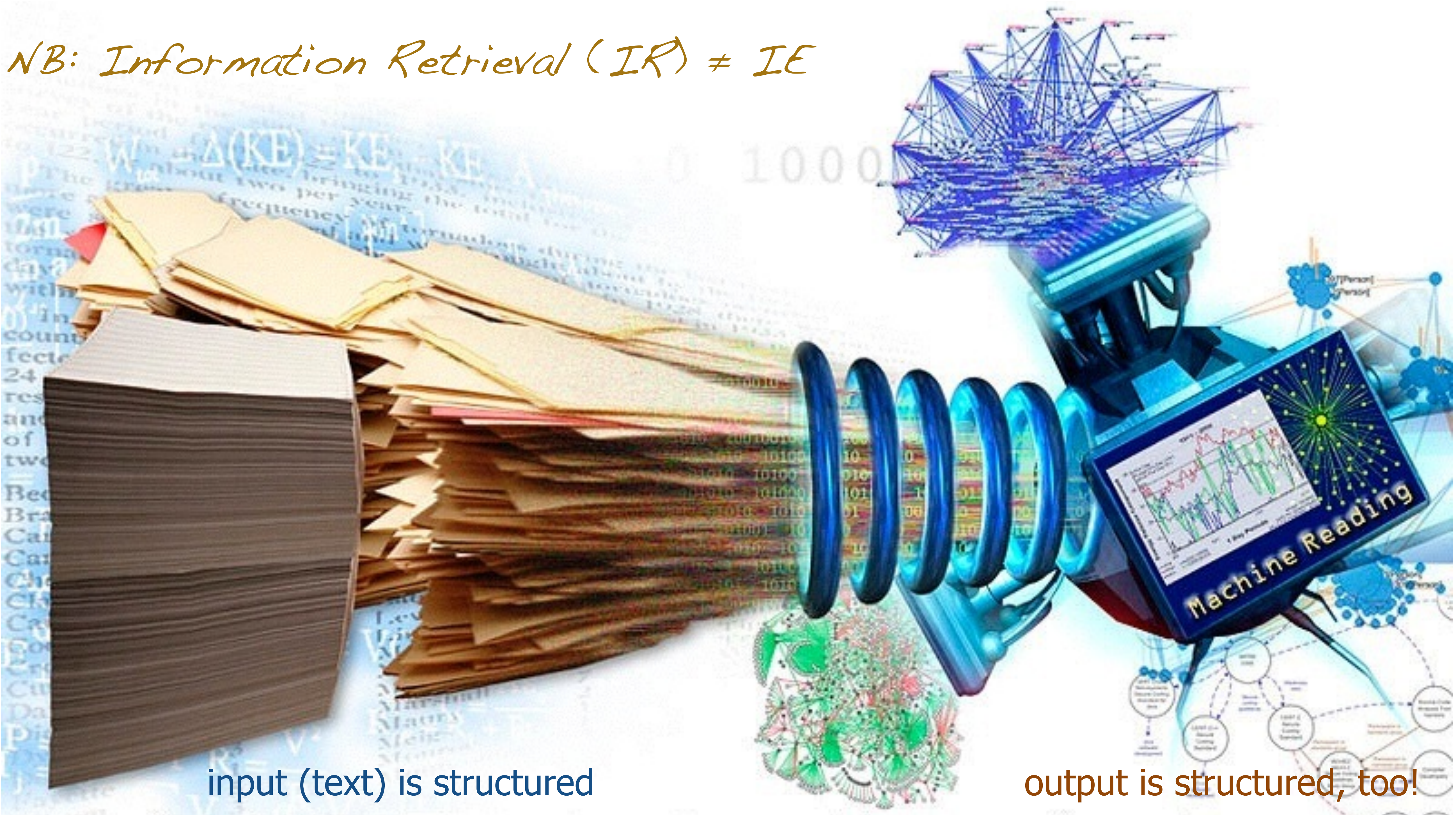


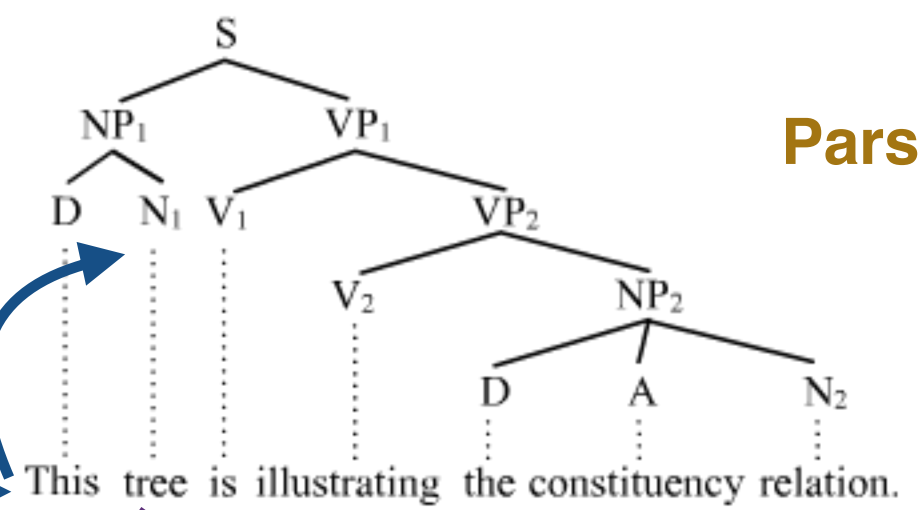
Image Source: www.DARPA.mil

"from **non-normalized** text to **connected** data"

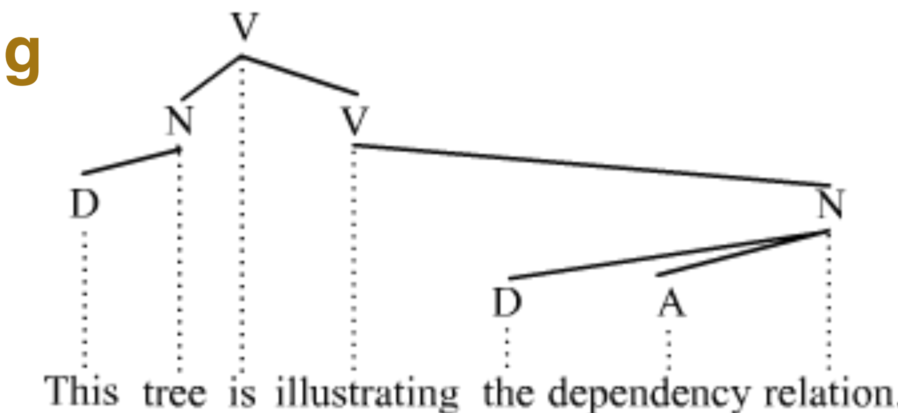
Language Understanding

Parsing

Coreference (Anaphora) Resolution

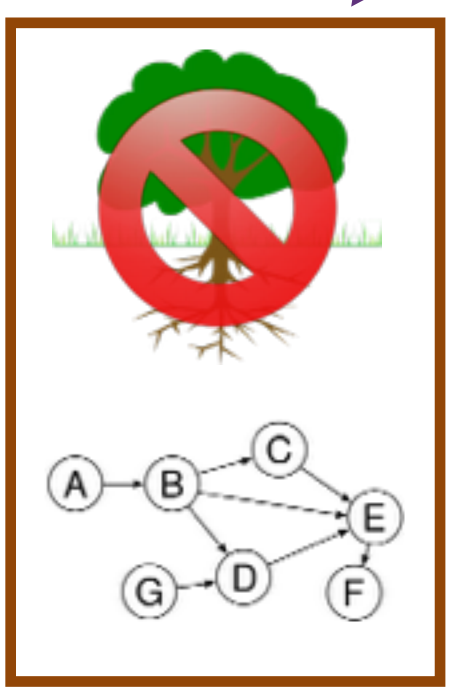


Constituency relation (PSG)

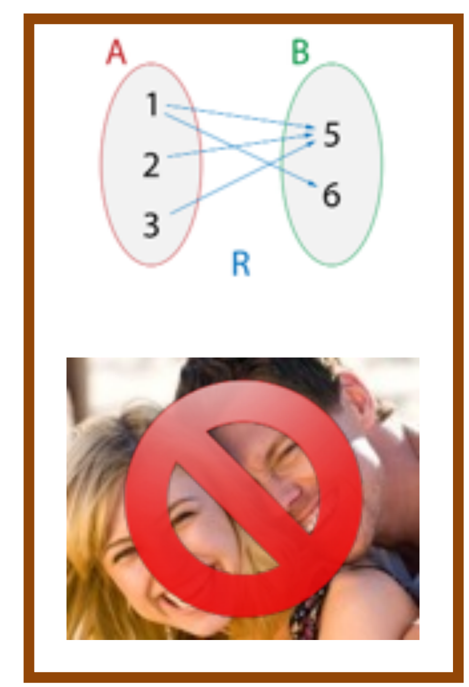


Dependency relation

Named Entity Recognition



Apple Siri



disambiguation!

Text Summarization

...is hard because...

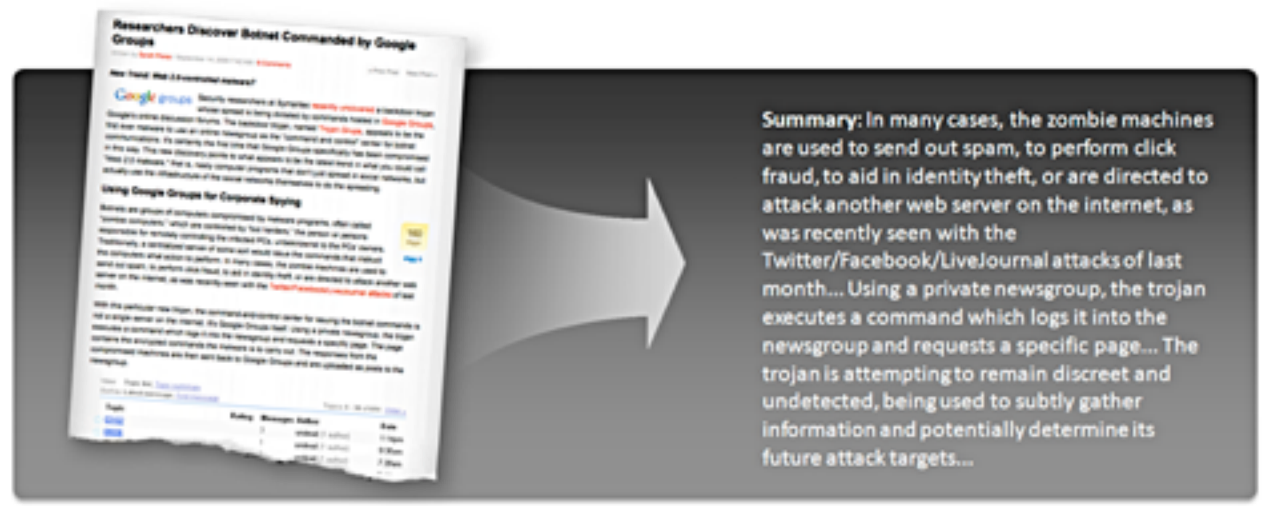
Variance/human agreement: When is a summary “correct”?

Coherence: providing **discourse structure** (text flow) to the summary.

Paraphrasing: important sentences are repeated, but with different wordings.

Implied messages: (the Dow Jones index rose 10 points → the economy is thriving)

Anaphora (coreference) **resolution**: very hard, but crucial.



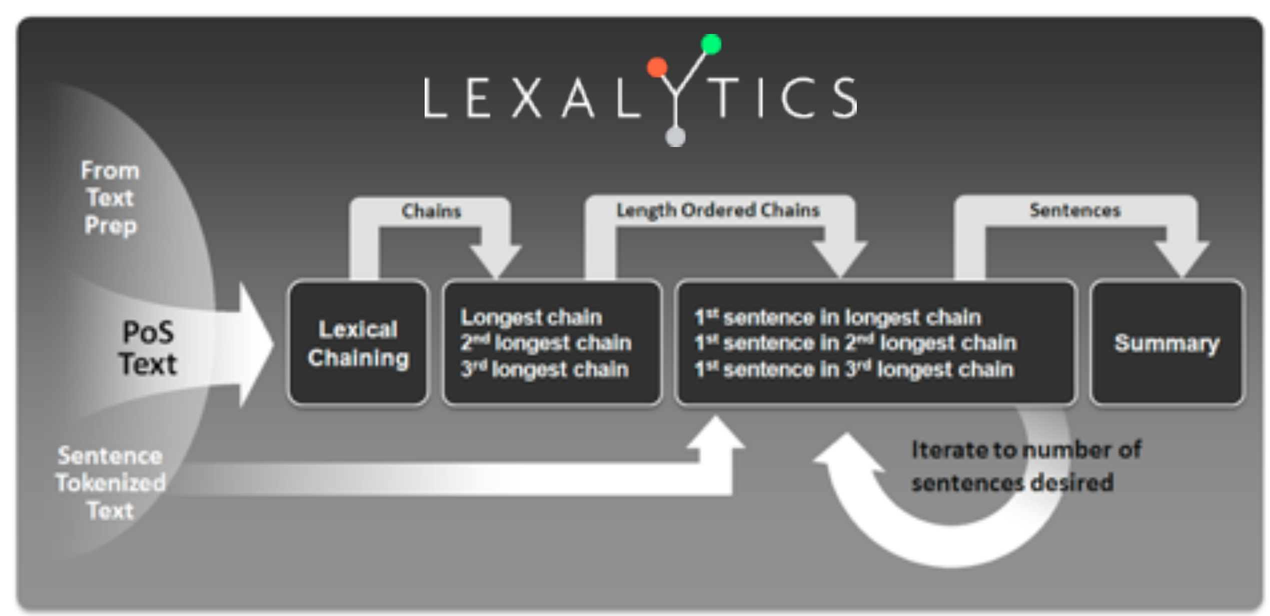
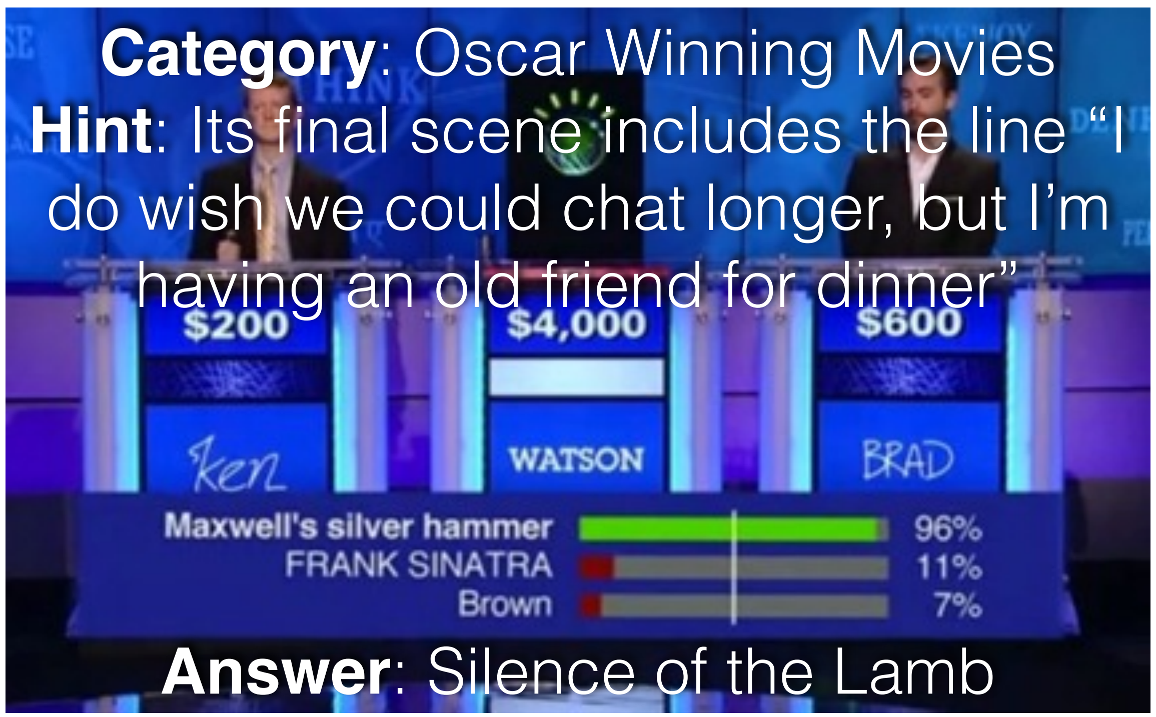


Image Source: www.lexalytics.com

Question Answering

Biggest issue: very domain specific



IBM

WolframAlpha

Machine Translation

Languages…

- have no gender (en: the) or use different genders (es/de: el/die 🌞; la/der 🌝; ??/das 👶)
- have different verb placements (es↔de).
- have a different concept of verbs (latin, arab, cjk).
- use different tenses (en↔de).
- have different word orders (latin, arab, cjk).

Ambiguity

It's all in the semantics! (Or is it?)

Part-of-Speech Tagging

The robot **wheels** out the iron.

Paraphrasing

Unemployment is on the rise.

vs

The economy is slumping.

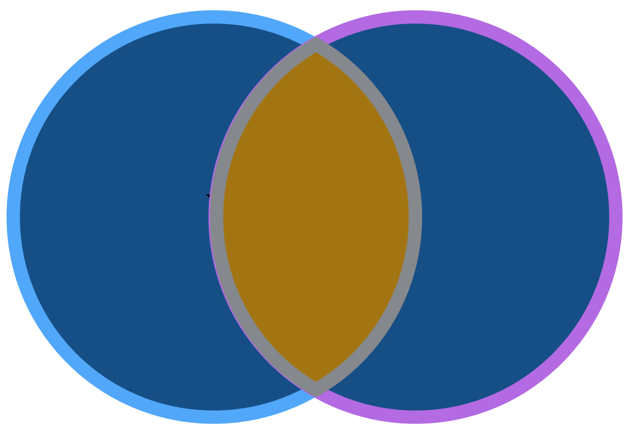
Anaphora Resolution

Carl and Bob were fighting:
"**You** should shut up,"
Carl told **him**.

Named Entity Recognition

Is **Princeton** really good for you?

The Conditional Probability for Dependent Events

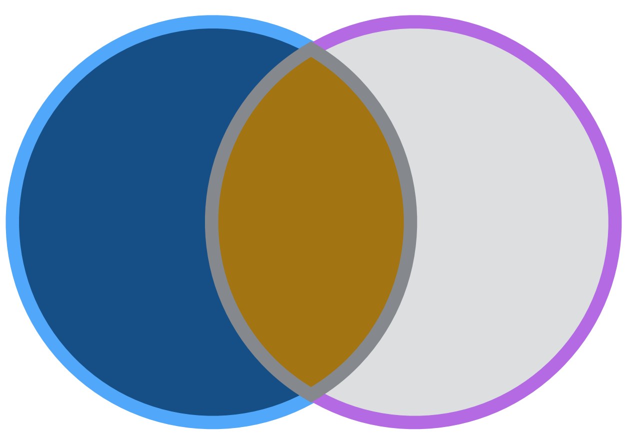


and **Joint Probability**

$P(X \cap Y) = P(X, Y) = P(X \times Y)$

The **multiplication principle** for dependent events*:

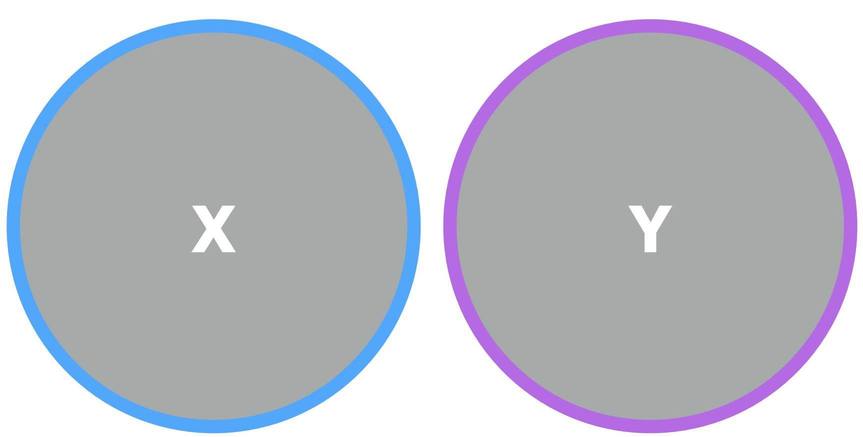
$P(X \cap Y) = P(Y) \times P(X \mid Y)$



therefore, by using a little algebra:

Conditional Probability

$P(X \mid Y) = P(X \cap Y) \div P(Y)$



***Independence**

$P(X \cap Y) = P(X) \times P(Y)$

$P(X \mid Y) = P(X)$

$P(Y \mid X) = P(Y)$

Marginal, Conditional and Joint Probabilities

variable/factor margin

contingency table

variable/factor

	X=x	X=x	M
Y=y	a/n = P(x	b/n = P(x	(a+b)/n = P(y
Y=y	c/n = P(x	d/n = P(x	(c+d)/n = P(y
M	(a+c)/n = P(x	(b+d)/n = P(x	

$\sum / n = 1 = P(X) = P(Y)$

Joint Probability*

$P(x_i, y_j) = P(x_i) \times P(y_j)$

Marginal Probability

$P(y_i)$

Conditional Probability

$P(x_i \mid y_j) = P(x_i, y_j) \div P(y_j)$

*for **independent** events

Bayes' Rule: Diachronic Interpretation

prior

likelihood

posterior $\longrightarrow$ $$P(H|D) = \frac{P(H) \times P(D|H)}{P(D)}$$

"normalizing constant"

(law of total probability)

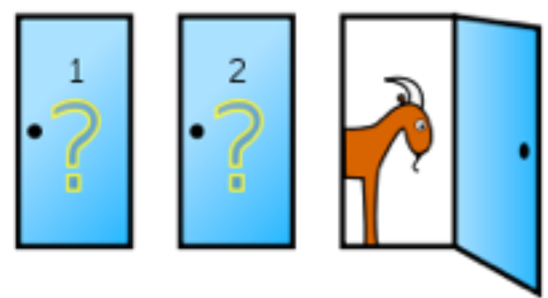
H - Hypothesis

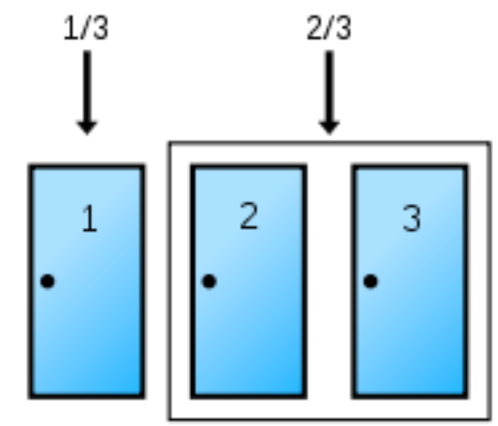
D - Data

Bayes' Rule: The Monty Hall Problem

Images Source: Wikipedia, Monty Hall Problem, Cepheus

$$P(H|D) = \frac{P(H) \times P(D|H)}{P(D)}$$





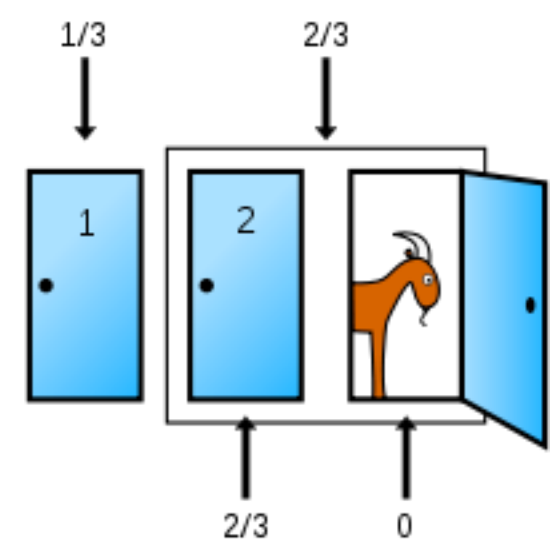
your pick

	1	2	3	
Prior p(H)	1/3	1/3	1/3	
Likelihood p(D|H)	1/2	1	0	**p(D) =Σ**
p(H)* p(D|H)	1/3×1/2 =1/6	1/3×1 =1/3	1/3×0 =0	1/6+1/3 =1/2
Posterior p(H|D)	1/6÷1/2 =1/3	1/3÷1/2 =2/3	0÷1/2 =0	

H=2 D=3

given the car is behind H=1, Monty Hall opens D=(1 or 2)

H=3 D=2



practical use: a trickster hides a stone with three cups...

An Overview of Open Source NLP Frameworks

- Natural Language ToolKit
  - NLTK, Python
- General Architecture for Text Engineering
  - GATE, Java
- Stanford NLP Framework
  - CoreNLP, Java
- Unstructured Information Management Architecture
  - UIMA, Java
  - Many framework-sized sub-projects, e.g., ClearNLP
- LingPipe Framework
  - LingPipe, Java (OpenSource, but only free for “non-commerical” use)
- FreeLing NLP Suite
  - FreeLing, C++
- The Lemur Toolkit
  - Lemur, C++ (IR + TextMining)
- The Bow Toolkit
  - Bow, C (Language Modeling)
- DeepDive Inference Engine
  - dp, Scala (+ SQL & Python)

Practicals :: Setup

- Install **Python**, **Numpy**, **SciPy**, **matplotlib**, **pandas**, and **IPython**
  - Via graphical installer: http://continuum.io/downloads
    - uses Continuum Analytics' "Anaconda Python 2.0.x", anaconda [for Py2.7, **recommended**] or anaconda3 [for Py3.4; if you are progressive & "know thy snake"]
  - Via command line: manual installation of above packages for Py2.7 or 3.4
    - http://fnl.es/installing-a-full-stack-python-data-analysis-environment-on-osx.html
      ...but you're on your own here!
- Install **NLTK 2.x**
  - Natural Language Toolkit http://www.nltk.org/install.html
    - Via Anaconda (Py2.7 only): **`conda install nltk`**
    - Default Python (Py2.7 only): **`pip install nltk`**
  - or download 3-alpha (for Py3.4):
    - http://www.nltk.org/nltk3-alpha
    - Run in directory: **`python setup.py install`**
- Install **SciKit-Learn 0.x**
  - http://scikit-learn.org/stable/install.html
    - Via Anaconda: **`conda install sklearn`**
    - Default Python: **`pip install sklearn`**
- [Install **gensim** (Py2.7 only)]
  - http://radimrehurek.com/gensim
    - Anaconda & Default Python: **`pip install gensim`**

Introduction to IPython , NLTK, NumPy, and SciPy

Ladies and Gentlemen, please start your engines!

Chatty Chatterbots

Create two chat bots with NLTK and let them talk to each other, printing each others answer on the screen.

http://www.nltk.org/api/nltk.chat.html

```
from nltk.chat import eliza; eliza.demo()

eliza??

from nltk.chat.util import \

   Chat, reflections

from nltk.chat.eliza import pairs as eliza_pairs

eliza = Chat(eliza_pairs, reflections)

eliza.respond?
```

“I do not fear computers. I fear the lack of them.”

Isaac Asimov, ~1980 (?)