

Introduction to Natural Language Processing

Bernardo Magnini
FBK, Trento, Italy
magnini@fbk.eu

Outline

- **What is Natural Language Processing (NLP)**
- **Challenges in NLP**
 - Ambiguity, redundancy
 - Lack of knowledge, need of inferences
 - Probabilistic judgments
- **Natural Language Processing: where we are**
 - Applications
 - Current limitations
- **Several approaches**
 - Frame semantics
 - Distributional semantics
 - Probabilistic models

What is Computational Linguistics

Computational Linguistics (CL) is the scientific study of language from a computational perspective. [www.aclweb.org/]

The long term goal is to realize *machines that understand natural languages* (e.g. English, German, Italian) both spoken and written

Other terminology:

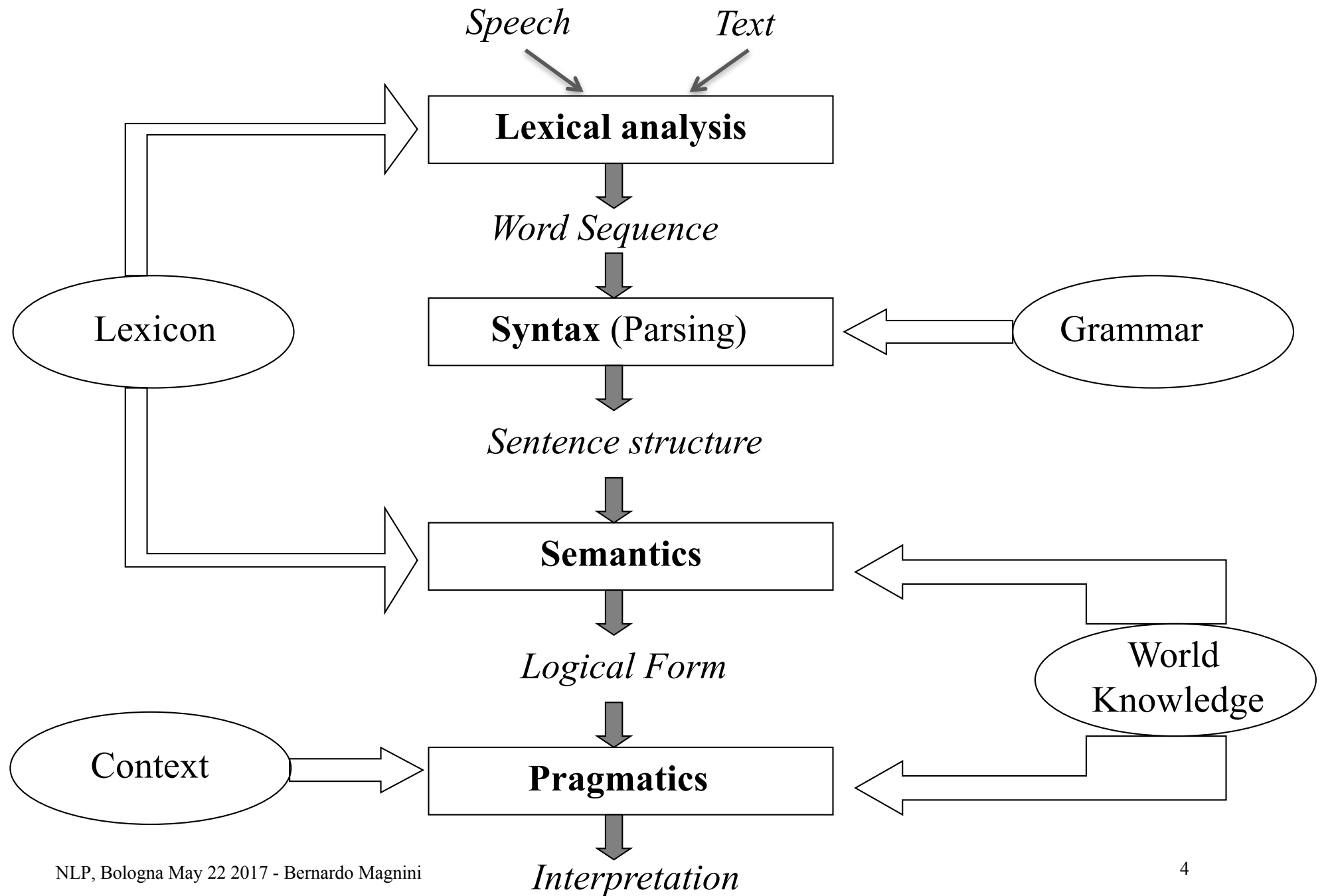
- Natural Language Processing (NLP)
- Human Language Technology (HLT)

A *Computational Linguist* is a scientist in CL: background in linguistics or computer science

The *Association for Computational Linguistics* (ACL) is the referent scientific society for CL

- Journals: Computational Linguistics, Transactions of the ACL, JNLE, etc.
- Conferences: ACL, EACL, EMNLP, COLING, LREC, etc.

Natural Language Interpretation



Lexical Analysis

Word level

- Tokenization (the role of punctuation)
- Morphological analysis
 - Lemma, part of speech, morphological features

World War One veteran becomes world's oldest man.

1. World (WORLD NOUN COMMON M SING)
2. War (WAR NOUN COMMON F SING)
3. One (ONE ADV NEG)
4. veteran (VETERAN NOUN COMMON SING)
5. becomes (BECOME VERB PRES 3 SING)
6. world (WORLD NOUN COMMON M SING)
7. 's (`S POS)
8. oldest (OLD ADJECTIVE M SING)
9. man (MAN NOUN COMMON M SING)
- 10.. (.)

Syntactic Analysis

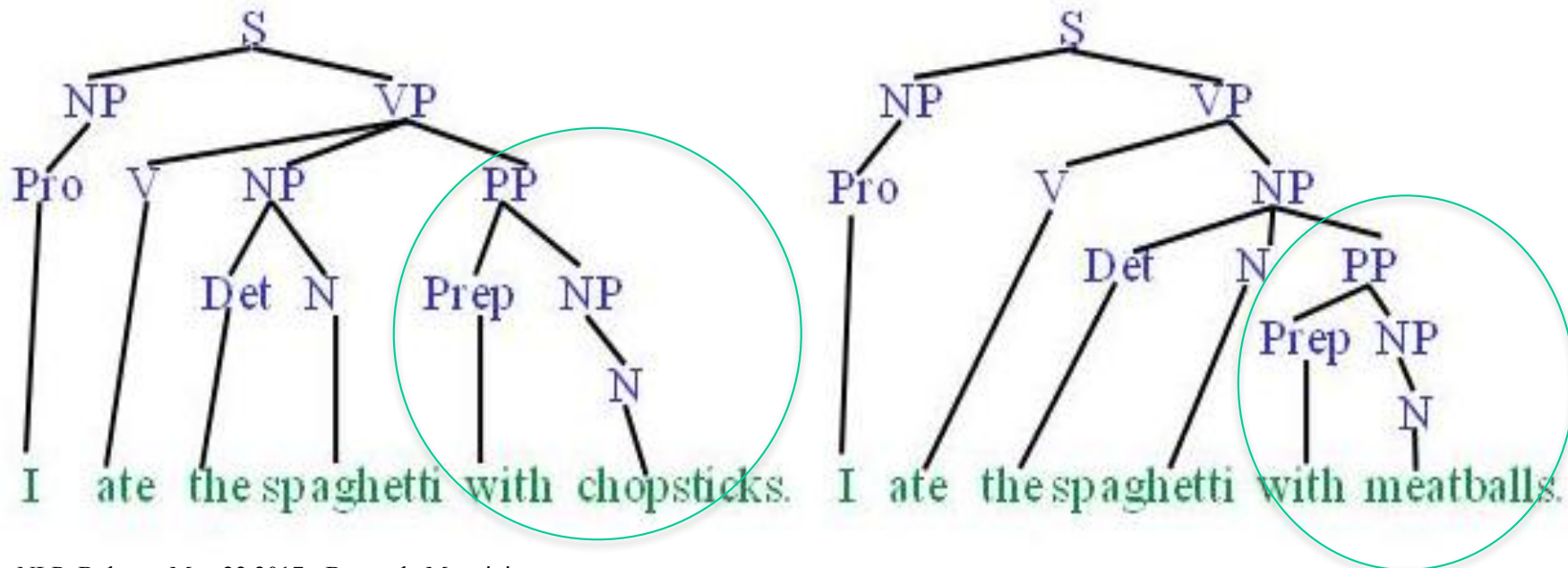
- **Sentence level**

- **Shallow parsing:** chunking

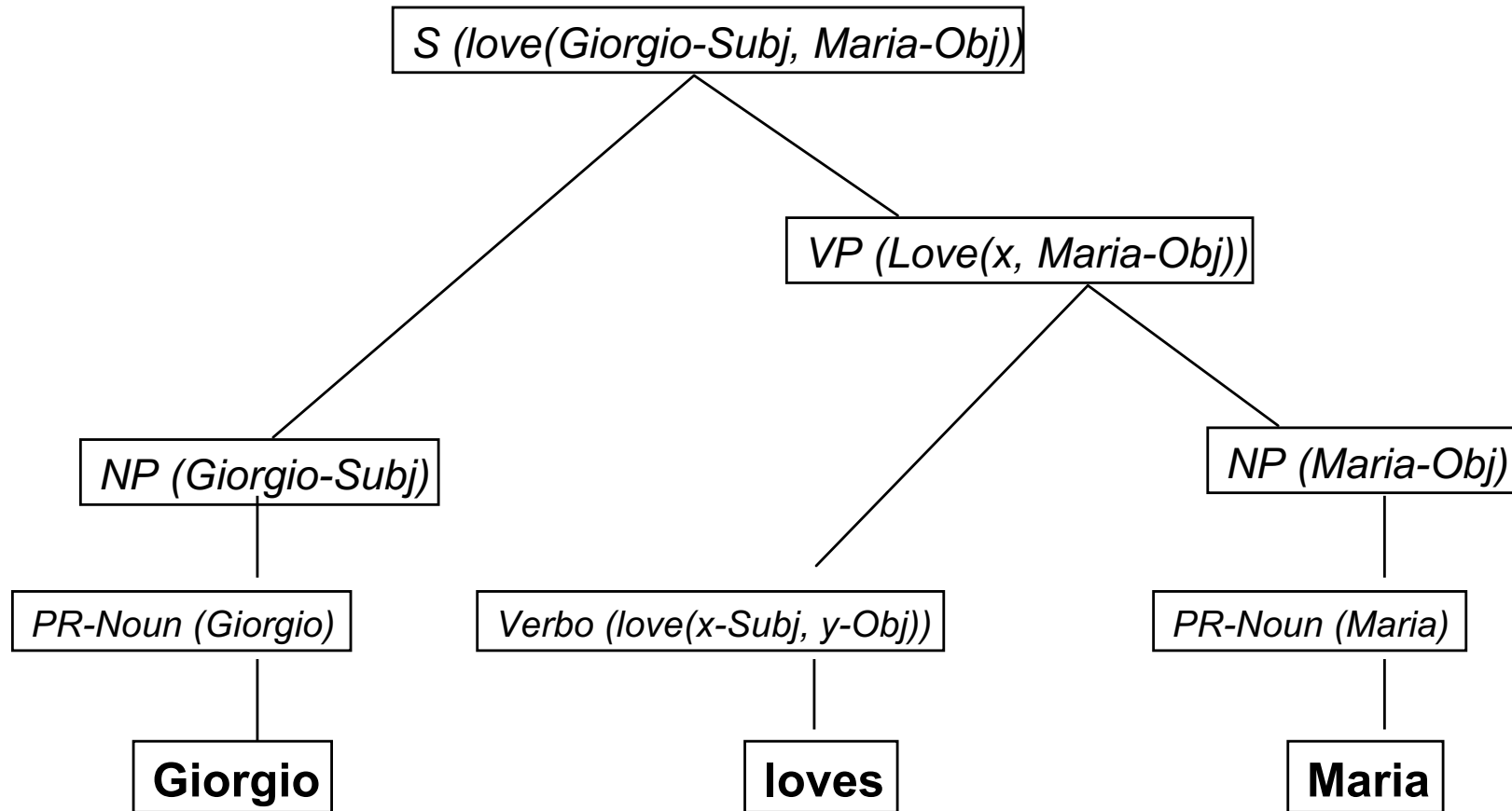
[NP I] [VP ate] [NP the spaghetti] [PP with] [NP chopsticks].

[NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs].

- **Deep parsing:** the syntactic structure of a sentence is recognized
- Syntactic constituents, syntactic ambiguities



Semantic Intepretation



- Compositional view of meaning
 - Word sense disambiguation
 - The meaning of a sentence is built on the meaning of words

Discourse Analysis and Pragmatics

- Sentences are interpreted in the communicative context in which they are uttered
 - Non linguistic context (e.g. time, place)
 - Anaphora and ellipsis resolution

I put pasta in the dish and then I ate it

- World knowledge is required (e.g. dishes are not food)
- Discourse relations (e.g. temporal connectives)
- User model (e.g. profiling)

NLP: A Difficult Challenge



mozart salzburg



[Salzburg - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/Salzburg](#) - Traduci questa pagina

The city is noted for its Alpine setting. **Salzburg** was the birthplace of 18th-century composer Wolfgang Amadeus **Mozart**. His mother was born at St Gilgen on the ...

[Mozart: Visit Salzburg & Wolfgang Amadeus Mozart](#)

[www.visit-salzburg.net](#) › ... › [art & culture](#) - Traduci questa pagina

Short article on Wolfgang Amadeus **Mozart** and his native city of **Salzburg**, written for prospective visitors of **Salzburg**. »»

[Mozart's House, Salzburg - Things to Do - VirtualTourist](#)

[www.virtualtourist.com](#) › ... › [Things to Do](#) - Traduci questa pagina

The house where **Mozart** was **born** on January 27, 1756 is one of the main attractions in Salzburg. I have to say that I find the pricing a bit much - it is a nice place ...

[Wolfgang Amadeus Mozart - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/.../Wolfgang_Amadeus_Moz...](#) - Traduci questa pagina

Wolfgang Amadeus **Mozart** was **born** to Leopold **Mozart** (1719–1787) and Anna Maria, née Pertl (1720–1778), at 9 Getreidegasse in Salzburg, capital of the ...

[List of compositions](#) - [Death](#) - [Antonio Salieri](#) - [Salzburg](#)

NLP: A Difficult Challenge

Google

mozart salzburg

Redundancy: different expressions (birthplace, native city) for the same meaning

[Salzburg - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Salzburg - Traduci questa pagina

The city is noted for its Alpine setting. **Salzburg** was the birthplace of 18th-century composer Wolfgang Amadeus **Mozart**. His mother was born at St. Gilgen on the ...

[Mozart: Visit Salzburg & Wolfgang Amadeus Mozart](#)

www.visit-salzburg.net > ... > art & culture - Traduci questa pagina

Short article on Wolfgang Amadeus **Mozart** and his native city of **Salzburg**, written for prospective visitors of **Salzburg**.

[Mozart's House, Salzburg - Things to Do - VirtualTourist](#)

www.virtualtourist.com > ... > Things to Do - Traduci questa pagina

The house where **Mozart** was **born** on January 27, 1756 is one of the main attractions in Salzburg. I have to say that I find the pricing a bit much - it is a nice place ...

[Wolfgang Amadeus Mozart - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/.../Wolfgang_Amadeus_Moz... - Traduci questa pagina

Wolfgang Amadeus **Mozart** was **born** to Leopold **Mozart** (1719–1787) and Anna Maria, née Pertl (1720–1778), at 9 Getreidegasse in Salzburg, capital of the ...

[List of compositions](#) - [Death](#) - [Antonio Salieri](#) - [Salzburg](#)

NLP: A Difficult Challenge



mozart salzburg

Ambiguity: (which “Mozart”)
same expression has different
meanings

[Salzburg - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/Salzburg](#) - Traduci questa pagina

The city is noted for its Alpine setting. **Salzburg** was the birthplace of 18th-century composer Wolfgang Amadeus **Mozart**. His mother was born at St Gilgen on the ...

[Mozart: Visit Salzburg & Wolfgang Amadeus Mozart](#)

[www.visit-salzburg.net](#) › ... › art & culture - Traduci questa pagina

Short article on Wolfgang Amadeus **Mozart** and his native city of **Salzburg**, written for prospective visitors of **Salzburg**. >>

[Mozart's House, Salzburg - Things to Do - VirtualTourist](#)

[www.virtualtourist.com](#) › ... › Things to Do - Traduci questa pagina

The house where **Mozart** was **born** on January 27, 1756 is one of the main attractions in Salzburg. I have to say that I find the pricing a bit much - it is a nice place ...

[Wolfgang Amadeus Mozart - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/.../Wolfgang_Amadeus_Moz...](#) - Traduci questa pagina

Wolfgang Amadeus **Mozart** was **born** to Leopold **Mozart** (1719–1787) and Anna Maria, née Pertl (1720–1778), at 9 Getreidegasse in Salzburg, capital of the ...

[List of compositions](#) - [Death](#) - [Antonio Salieri](#) - [Salzburg](#)

CL: A Difficult Challenge

Google

mozart salzburg

Incompleteness: inferences are needed (house - located_in - Salzburg)

[Salzburg - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Salzburg - Traduci questa pagina

The city is noted for its Alpine setting. **Salzburg** was the birthplace of 18th-century composer Wolfgang Amadeus **Mozart**. His mother was born at St Gilgen on the ...

[Mozart: Visit Salzburg & Wolfgang Amadeus Mozart](#)

www.visit-salzburg.net > ... > art & culture - Traduci questa pagina

Short article on Wolfgang Amadeus **Mozart** and his native city of **Salzburg**, written for prospective visitors of **Salzburg**. >>

[Mozart's House, Salzburg - Things to Do - VirtualTourist](#)

www.virtualltourist.com > ... > Things to Do - Traduci questa pagina

The house where **Mozart** was **born** on January 27, 1756 is one of the main attractions in Salzburg. I have to say that I find the pricing a bit much - it is a nice place ...

[Wolfgang Amadeus Mozart - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/.../Wolfgang_Amadeus_Moz... - Traduci questa pagina

Wolfgang Amadeus **Mozart** was **born** to Leopold **Mozart** (1719–1787) and Anna Maria, née Pertl (1720–1778), at 9 Getreidegasse in Salzburg, capital of the ...

[List of compositions](#) - [Death](#) - [Antonio Salieri](#) - [Salzburg](#)

CL: A Difficult Challenge



mozart salzburg

Non literal meaning: “saw the light”

[Salzburg - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Salzburg - Traduci questa pagina

The city is noted for its Alpine setting. **Salzburg** was the birthplace of 18th-century composer Wolfgang Amadeus **Mozart**. His mother was born at St Gilgen on the ...

[Mozart: Visit Salzburg & Wolfgang Amadeus Mozart](#)

www.visit-salzburg.net > ... > art & culture - Traduci questa pagina

Short article on Wolfgang Amadeus **Mozart** and his native city of **Salzburg**, written for prospective visitors of **Salzburg**. >>

[Salzburg and Surroundings - Tourismusverband Hallein / Bad ...](#)

www.hallein.com/.../salzburg-umgebung.php?navi... - Traduci questa pagina

Salzburg, the capital of the beautiful federal province of **Salzburg**, can be reached ...

Salzburg where Wolfgang Amadeus **Mozart** first **saw the light** of day in 1756.

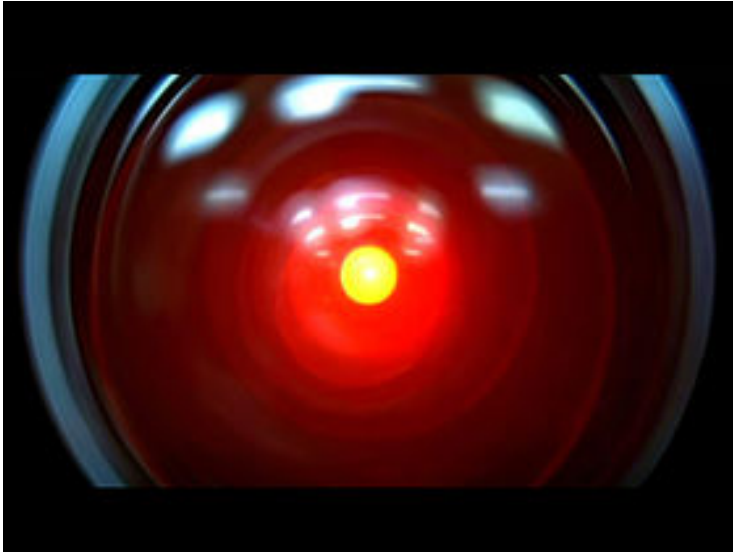
[Wolfgang Amadeus Mozart - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/.../Wolfgang_Amadeus_Moz... - Traduci questa pagina

Wolfgang Amadeus **Mozart** was **born** to Leopold **Mozart** (1719–1787) and Anna Maria, née Pertl (1720–1778), at 9 Getreidegasse in Salzburg, capital of the ...

[List of compositions](#) - [Death](#) - [Antonio Salieri](#) - [Salzburg](#)

CL: Where we are



Hal 9000 Space Odyssey – 1968

- ✓ Rule-based systems: 70'-80'
- ✓ Data-driven approaches: 90'- now
- ✓ Enabling technology: search engines, translation, voice commands

- ✓ Natural Language Understanding
- ✓ Artificial Intelligence
- ✓ An interdisciplinary field: computer science, statistics, linguistics, psychology, philosophy of language, ...



IBM Watson at Jeopardy Challenge - 2011

Where we are: Personal Assistant

Voice commands

- Speech recognition
- Interpretation of simple questions

Context aware

- Know where you are
“the closest restaurant...”

Personalized

- Know your social network
“send a message to my wife...”



Where we are: Personal Assistant

Voice commands

- Speech recognition
- Interpretation of simple questions

Context aware

- Know where you are
“the closest restaurant...”

Personalized

- Know your social network
“send a message to my wife...”

Still, poor dialogue, restricted domains, noisy environments, ...



Where we are: Subtitling and Translation



- Real-time transcription
- ✓ Quasi real-time translation
- ✓ Applications:
 - ✓ media content (BBC)
 - ✓ Skype translator
 - ✓ education (lectures)
- ✓ Language acquisition

Where we are: Subtitling and Translation



Real-time transcription

✓ Quasi real-time translation

✓ Applications:

✓ media content (BBC)

✓ Skype translator

✓ education (lectures)

✓ Language acquisition

Still, poor quality of translation (compared to professional level), ...

Where we are: Semantic Tagging

The semantic web (3.0)

- Using metadata to tag “post”
- Crucial for semantic search
- Multimedia tagging
- Sentiment

The web of data

- Linking text to structured data
(Open Linked Data, Wikipedia)



Where we are: Semantic Tagging

The semantic web (3.0)

- Using metadata to tag “post”
- Crucial for semantic search
- Multimedia tagging
- Sentiment

The web of data

- Linking text to structured data
(Open Linked Data, Wikipedia)

Still, tagging “big data” is computationally expensive, portability (domains, languages) is very poor



Where we are: Entity-Based Search

From key-words to objects

Information extraction from large-scale archives: entities, persons, locations, institutions, relations.

da vinci

Web Immagini Maps Shopping Notizie Altro Strumenti di ricerca

Circa 48.400.000 risultati (0,32 secondi)

I cookie ci aiutano a fornire i nostri servizi. Utilizzando tali servizi, accetti il nostro utilizzo dei cookie.
OK Ulteriori informazioni

[Leonardo da Vinci - Wikipedia](#)
it.wikipedia.org/wiki/Leonardo_da_Vinci
Leonardo di ser Piero da Vinci (Vinci, 15 aprile 1452 – Amboise, 2 maggio 1519) è stato un pittore, ingegnere e scienziato italiano. Uomo d'ingegno e talento ...
Itifallico - Genio - Francesco Melzi - Gioconda

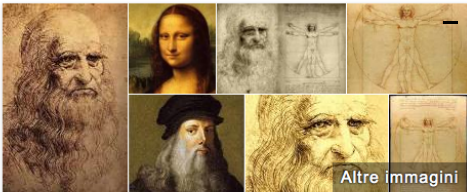
[Liceo Da Vinci - Trento](#)
www.liceodavincitn.it/
Mercoledì 13 novembre 2013, alle 9.30, presso l'aula magna del liceo "Leonardo da Vinci" di Trento, presentazione del secondo numero della rivista "Presenti" ...
Indirizzi - Valutazione - Info Genitori - Udienze

[UNITN | Lifelong Learning Programme - Leonardo da Vinci - Trento](#)
www.unitn.it/outgoing/.../lifelong-learning-programme-leonardo-da-vin...
Nell'ambito del programma LLP-Leonardo da Vinci, parte del più ampio Programma per l'Apprendimento Permanente 2007–2013 lanciato dalla Commissione ...

°°° [CENTRO SERVIZI VOLONTARIATO DELLA PROVINCIA DI...](#)
aziende.virgilio.it > ... > Associazioni ed enti > Associazioni di volon...
CENTRO SERVIZI VOLONTARIATO DELLA PROVINCIA DI TRENTO - Piazza Leonardo Da Vinci - Trento - ASSOCIAZIONI DI VOLONTARIATO PREVENZIONE ...

[Ministero degli Affari Esteri - D.A.V.I.N.C.I.](#)
www.esteri.it/davinci/
Benvenuti nella banca dati DAVINCI. COS'E' DAVINCI? DAVINCI è una banca dati accessibile via Internet, predisposta dal Ministero degli Affari Esteri Italiano ...

da Vinci Surgery - Minimally Invasive Robotic Surgery with the da








Altre immagini

Leonardo da Vinci

Leonardo di ser Piero da Vinci è stato un pittore, ingegnere e scienziato italiano. Uomo d'ingegno e talento universale del Rinascimento, incarnò in pieno lo spirito della sua epoca, portandolo alle ... Wikipedia

Data di nascita: 15 aprile 1452, Vinci
Data di morte: 2 maggio 1519, Amboise, Francia
Altezza: 1,94 m
Genitori: Piero Fruosino di Antonio da Vinci, Caterina da Vinci
Fratelli: Bartolomeo da Vinci, Bernedetto Ser Piero, Altro

Opera d'arte

 Gioconda 1517	 Ultima Cena 1498	 Uomo vitruviano 1490	 Dama con l'ermellino 1490	 Vergine delle Rocce 1486
---	--	--	--	--



Ultima Cena

Leonardo

L'Ultima Cena è un dipinto parietale a tempera grassa su intonaco di Leonardo da Vinci, databile al 1494-1498 e conservato nell'ex-refettorio del convento adiacente al santuario di Santa Maria delle Grazie a Milano.
Wikipedia

Artista: Leonardo da Vinci

Luogo: Chiesa di Santa Maria delle Grazie

Data creazione: 1495–1498

Soggetto: Gesù

Dimensioni: 4,6 m x 8,8 m

Supporti: Gesso, Pistacia lentiscus, Pittura a tempera, Pece

Where we are: Entity-Based Search

From key-words to objects

Information extraction from large-scale archives: entities, persons, locations, institutions, relations.

The screenshot shows a search engine interface with the query 'da vinci' in the search bar. Below the search bar, there are navigation tabs for 'Web', 'Immagini', 'Maps', 'Shopping', 'Notizie', 'Altro', and 'Strumenti di ricerca'. The search results are displayed in a grid format. The first result is a Wikipedia entry for 'Leonardo da Vinci', which includes a small image of Leonardo's face and a link to the full article. Below this, there are several other search results, including a page from 'Liceo Da Vinci - Trento' and a page from 'UNITN | Lifelong Learning Programme - Leonardo da Vinci - Trento'. Each result includes a small thumbnail image and a brief description of the content.



Leonardo da Vinci

Leonardo di ser Piero da Vinci è stato un pittore, ingegnere e scienziato italiano. Uomo d'ingegno e talento universale del Rinascimento, incarnò in pieno lo spirito della sua epoca, portandolo alle ... [Wikipedia](#)

Data di nascita: 15 aprile 1452, Vinci

Data di morte: 2 maggio 1519, Amboise, Francia

Altezza: 1,94 m

Genitori: Piero Fruosino di Antonio da Vinci, Caterina da Vinci

Fratelli: Bartolomeo da Vinci, Bernedetto Ser Piero, Altro

Opera d'arte



Ultima Cena

Leonardo

L'Ultima Cena è un dipinto parietale a tempera grassa su intonaco di Leonardo da Vinci, databile al 1494-1498 e conservato nell'ex-refettorio del convento adiacente al santuario di Santa Maria delle Grazie a Milano. [Wikipedia](#)

Artista: [Leonardo da Vinci](#)

Luogo: [Chiesa di Santa Maria delle Grazie](#)

Data creazione: 1495–1498

Soggetto: Gesù

Dimensioni: 4,6 m x 8,8 m

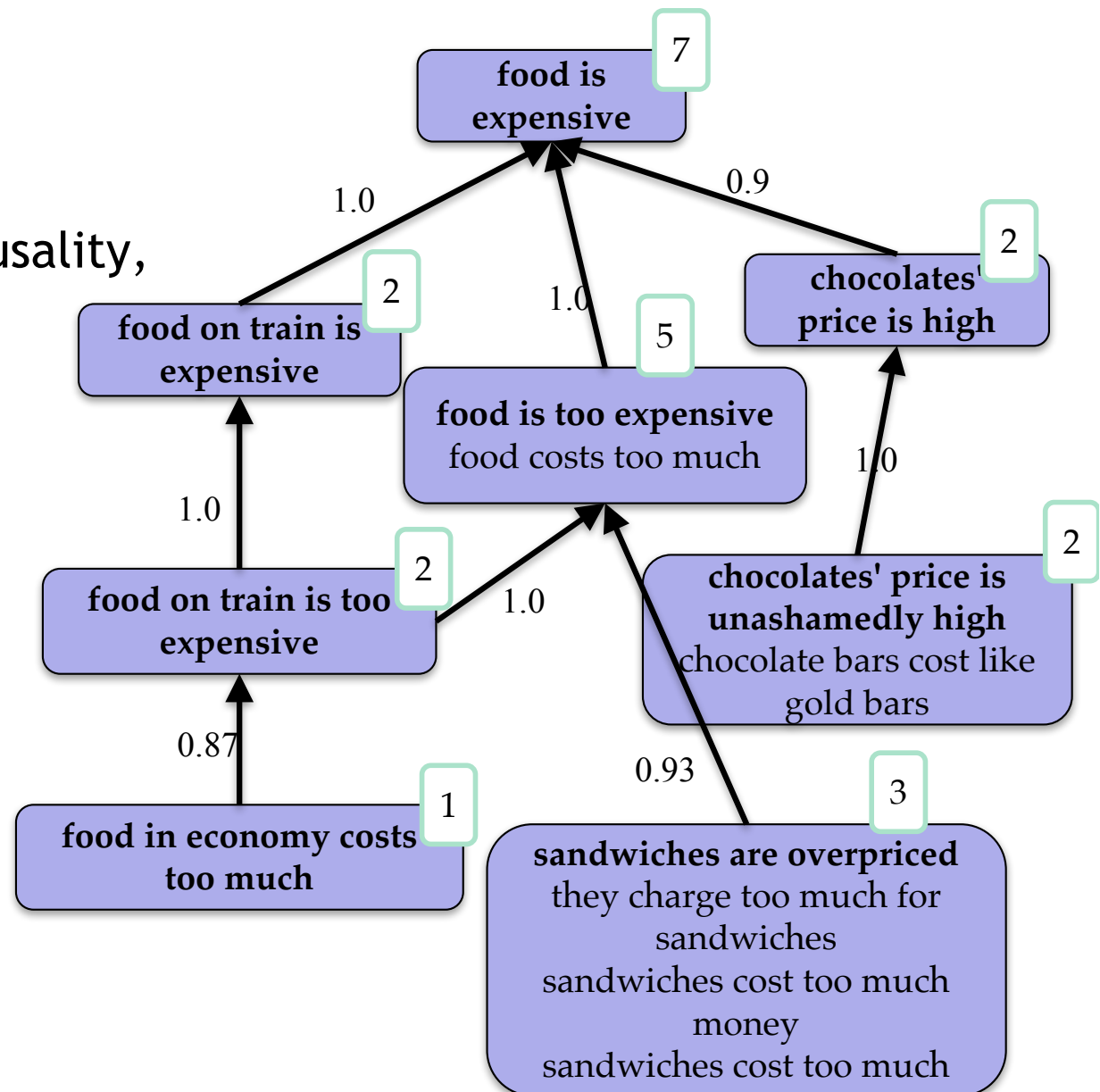
Supporti: Gesso, [Pistacia lentiscus](#), Pittura a tempera, Pece

Still, only simple entities, no events, coreference problematic for low frequent entities, ...

Where we are: Deep Understanding

Semantic inferences

- Entailment, similarity, causality, temporal relations
- Probabilistic judgements

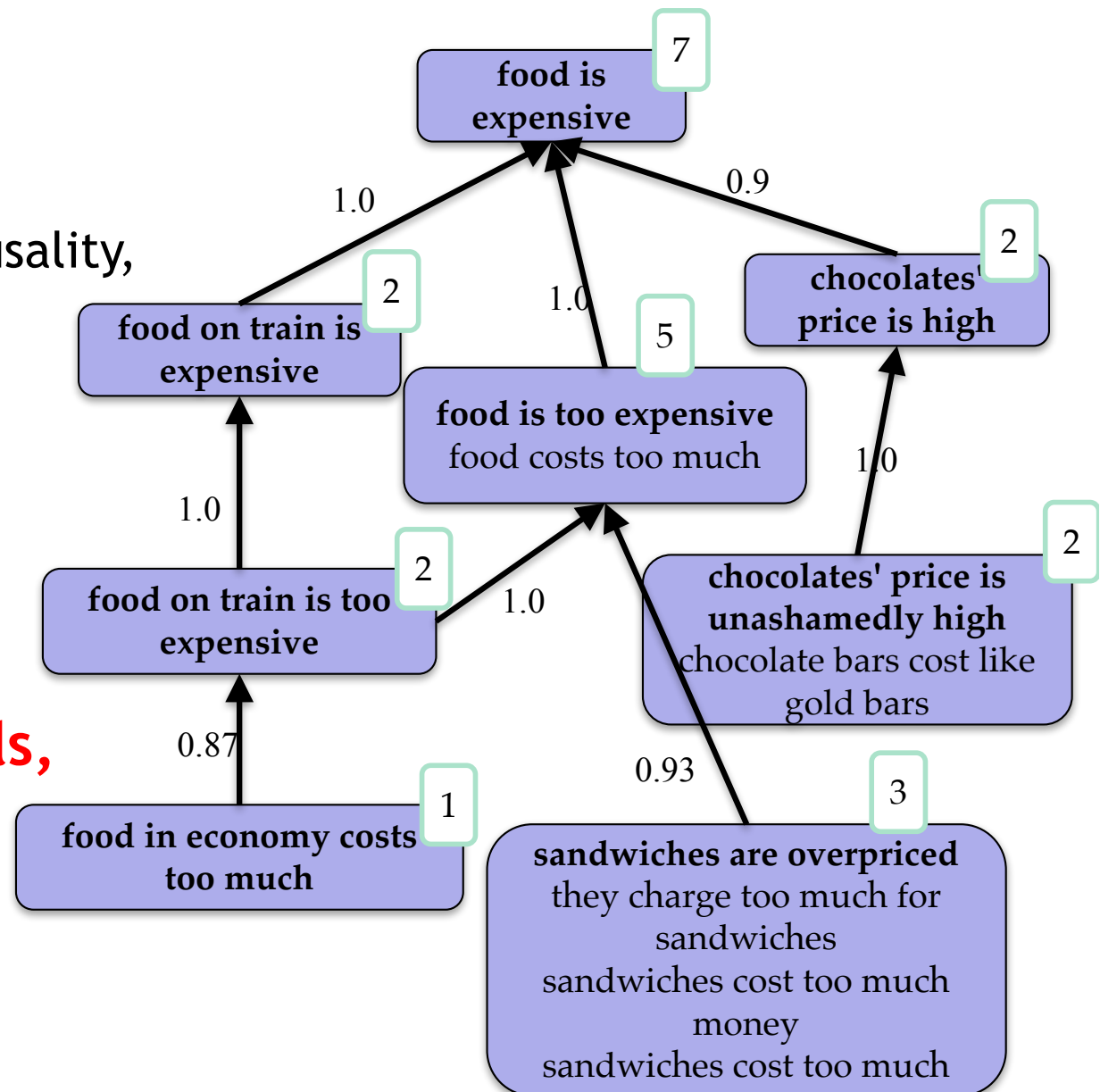


Where we are: Deep Understanding

Semantic inferences

- Entailment, similarity, causality, temporal relations
- Probabilistic judgements

Still, performance are poor, lack of clear models, available datasets are very small, ...





1. Frame Semantics

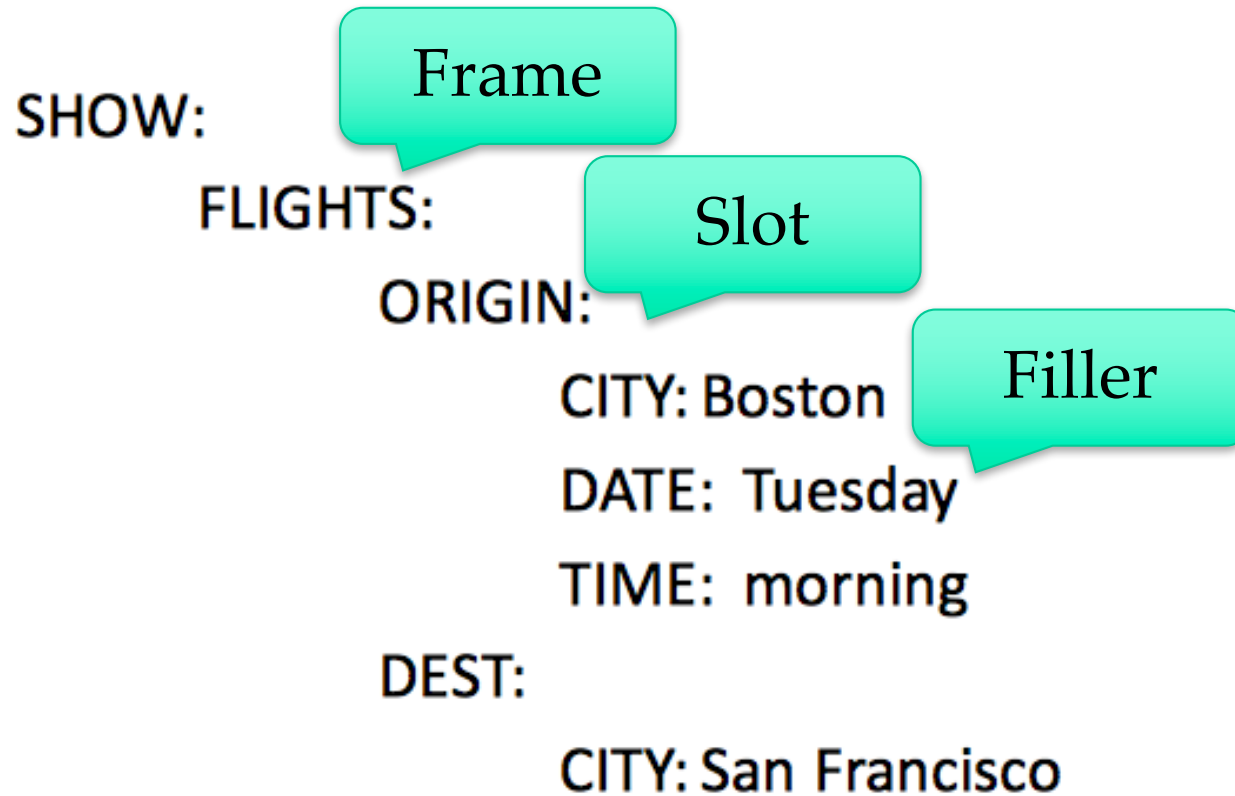
Apple Siri

- <http://www.apple.com/it/ios/siri/>
- Personal voice assistant on all Apple devices
- Since 2011, several languages
- Vocal commands: email, appointments, news, maps, points of interest, whether forecast, booking, search, ...
- Integrated into third party app: WhatsApp, LinkedIn, Pinterest, ...



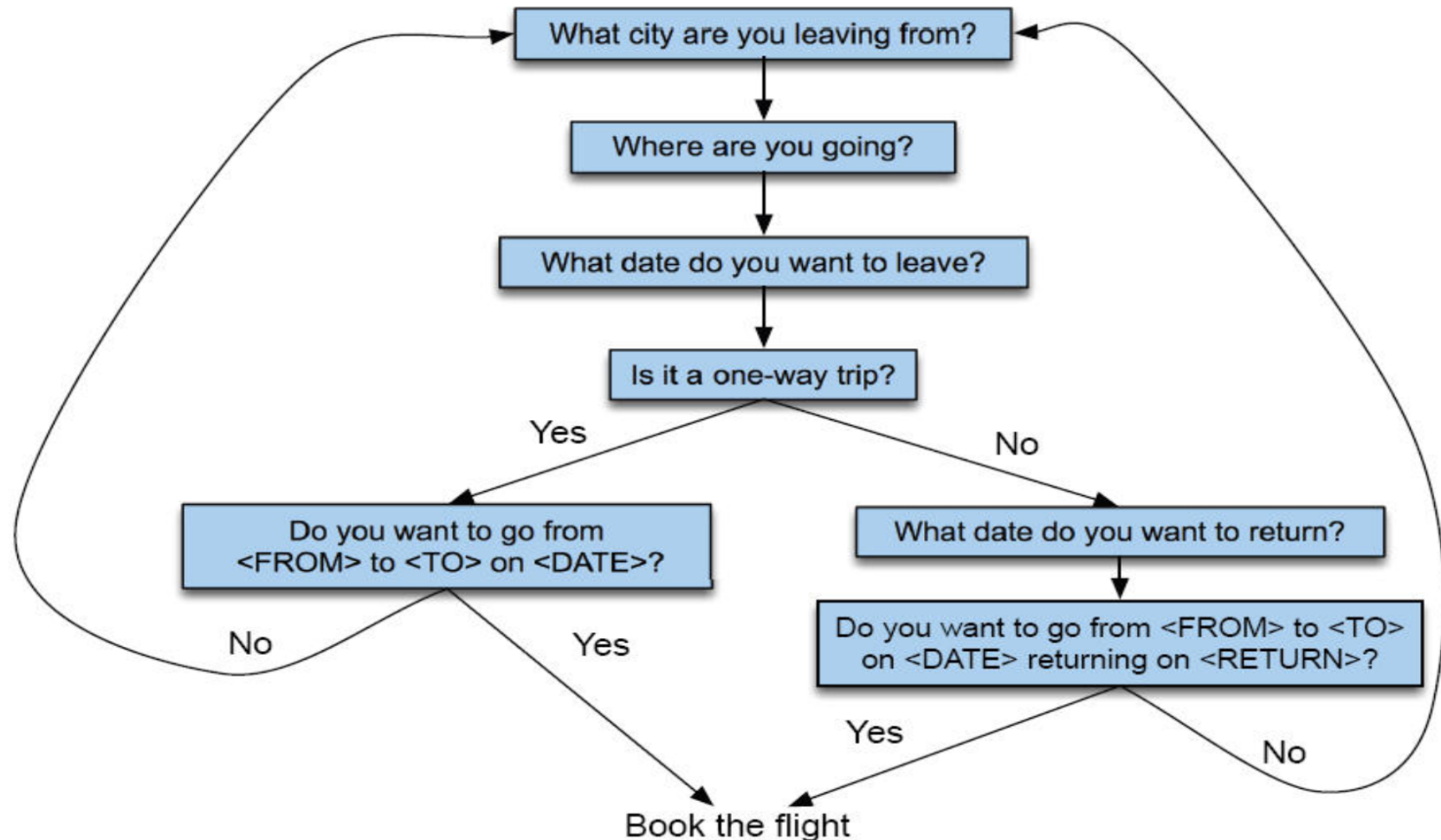
“Frame and Slots” Semantics

Show me morning flights from Boston to SF on Tuesday.



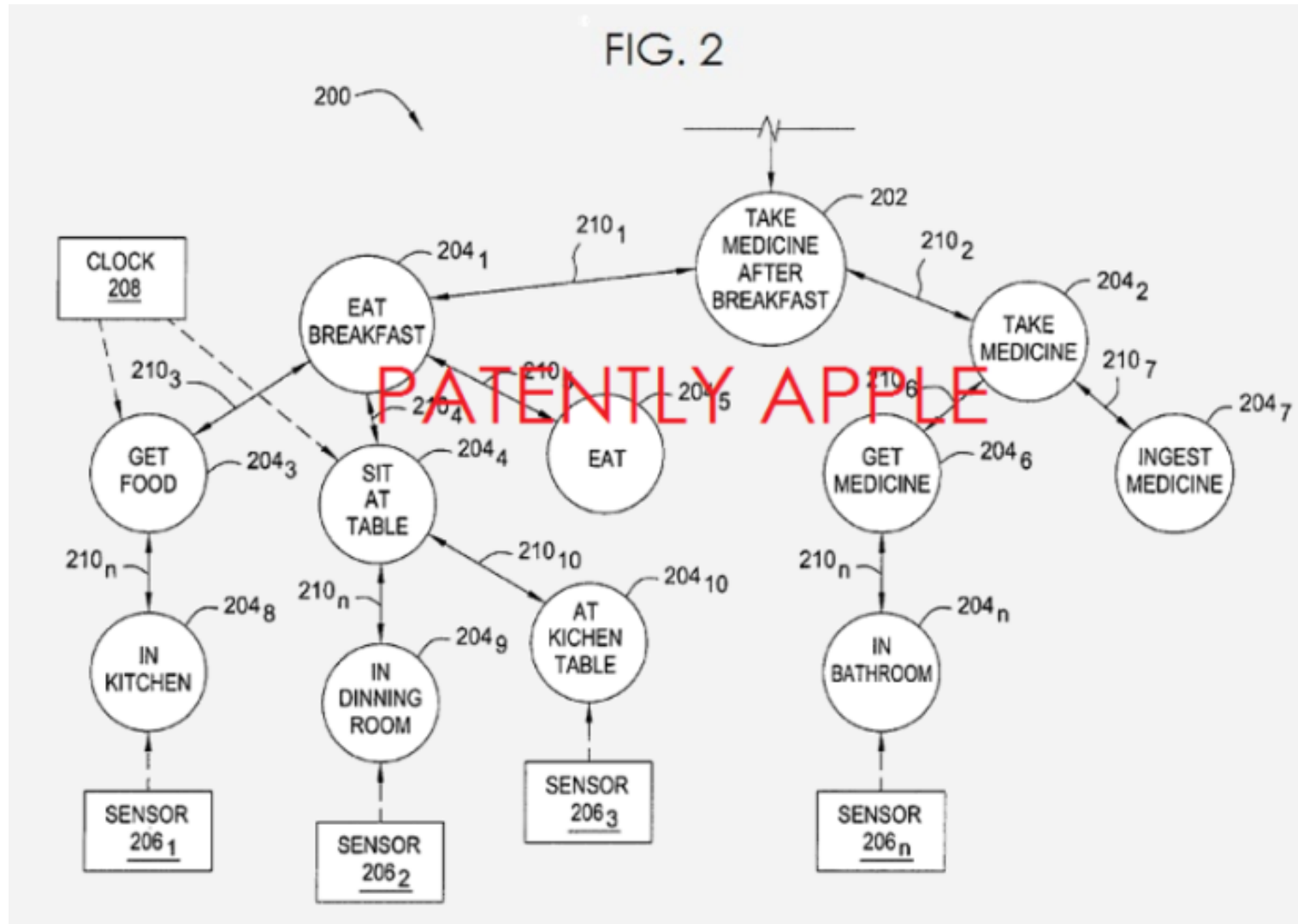
Add Statistical classifiers to map words to semantic frame- fillers

Managing conversations...



- A rule-based dialogue manager. At each dialogue state the system tries to fill a slot with user information.

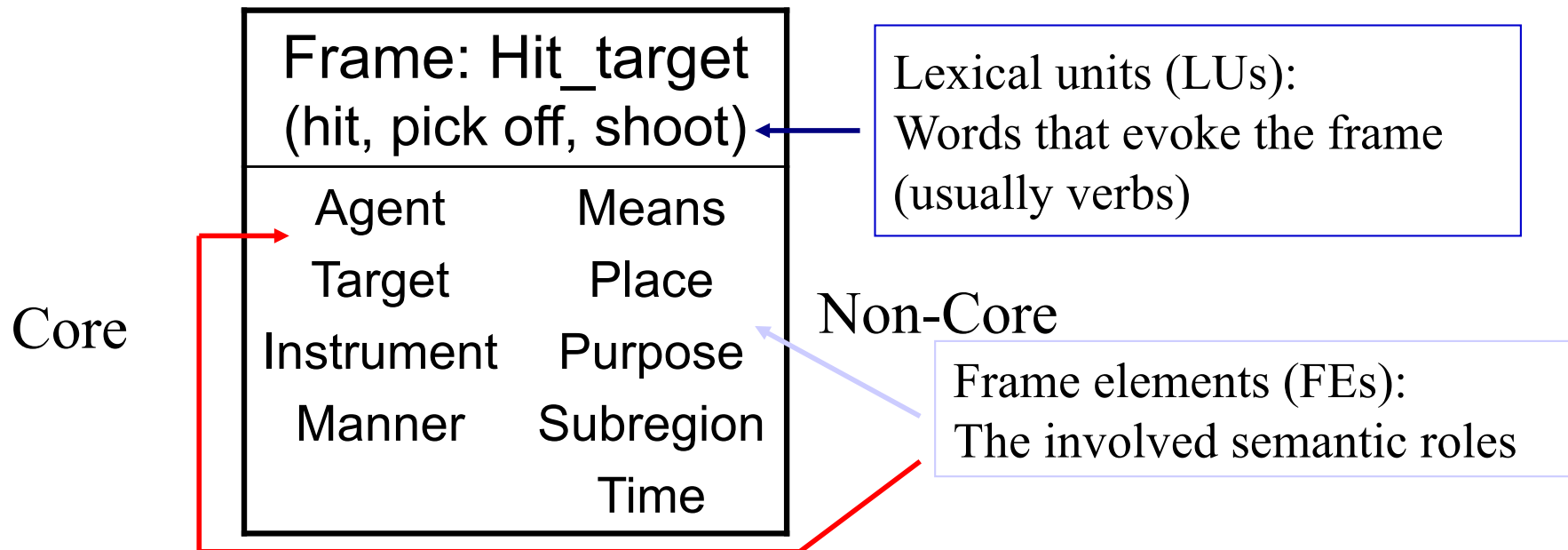
“Active Ontology” (Apple Siri)



- **Active Ontology:** relational network of concepts
- **rule sets** that perform actions on concepts

Frames and Linguistics

FrameNet [Fillmore et al. 01]



[Agent *Kristina*] **hit** [Target *Scott*] [Instrument *with a baseball*] [Time *yesterday*].

Abstract Meaning Representation

(s / **see-01**
:ARG0 (i / i)
:ARG1 (p / picture
:mod (m / magnificent)
:location (b2 / book :wiki -
:name (n / name :op1 "True" :op2 "Stories" :op3 "from" :op4 "Nature")
:topic (f / forest
:mod (p2 / primeval))))
:mod (o / once)
:time (a / age-01
:ARG1 i
:ARG2 (t / temporal-quantity :quant 6
:unit (y / year))))

Semantic roles as slots

*Once when I was six years old I saw a magnificent picture in a book ,
called True Stories from Nature , about the primeval forest .*

Semantic Role Labeling (SRL)

- SRL can be treated as an sequence labeling problem.
- For each verb, try to extract a value for each of the possible semantic roles for that verb.
- Employ any of the standard sequence labeling methods (e.g. machine learning algorithms)

SRL with Parse Trees

- Assume that a syntactic parse is available.
- For each predicate (verb), label each node in the parse tree as either not-a-role or one of the possible semantic roles.

Color Code:

not-a-role

agent

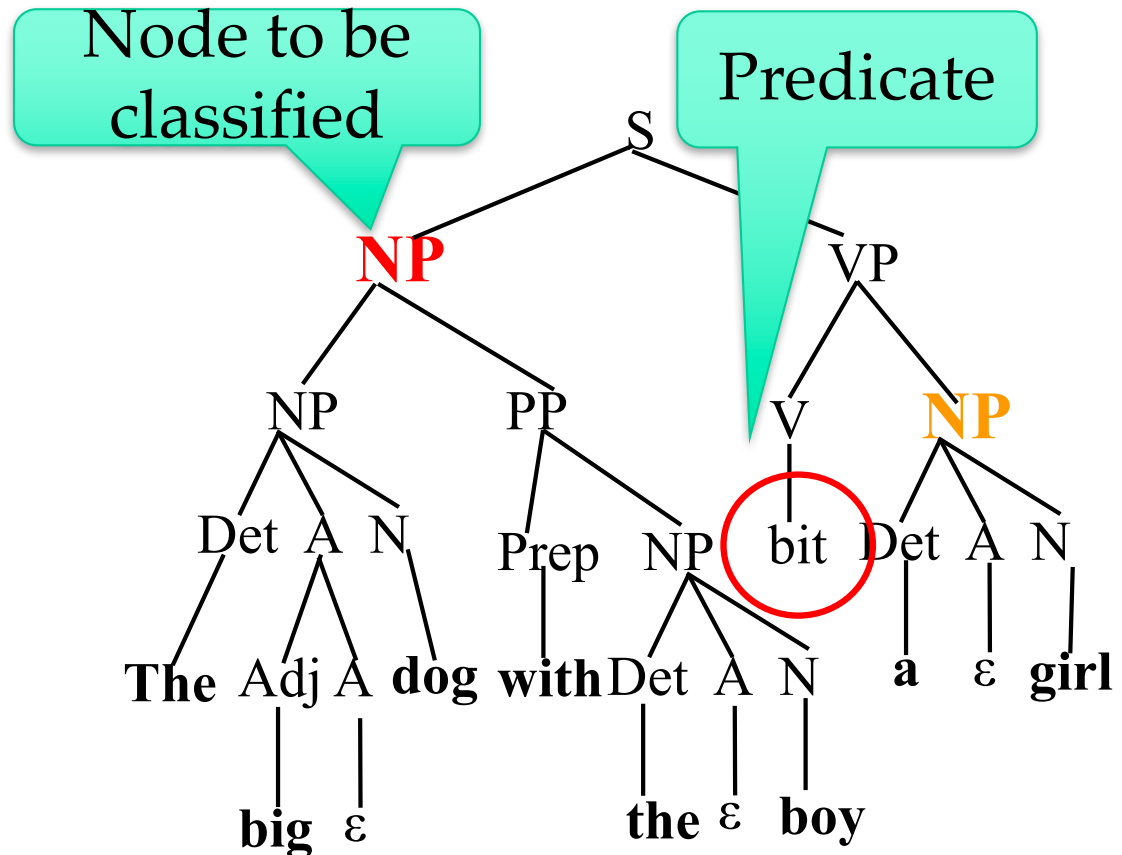
patient

source

destination

instrument

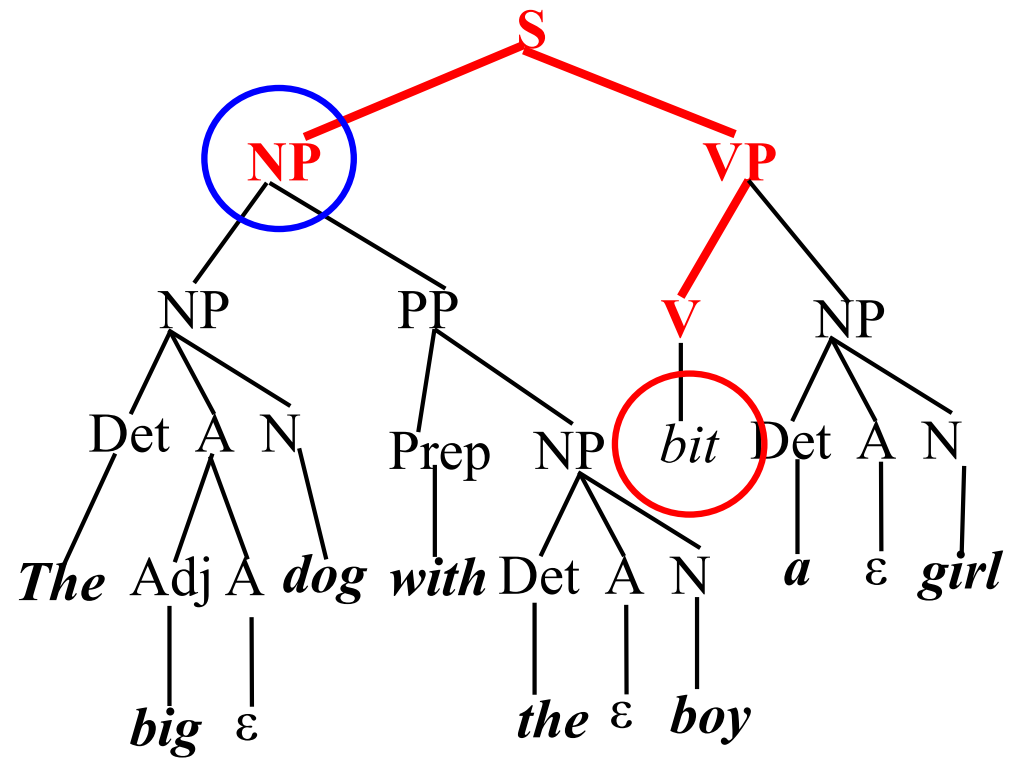
beneficiary



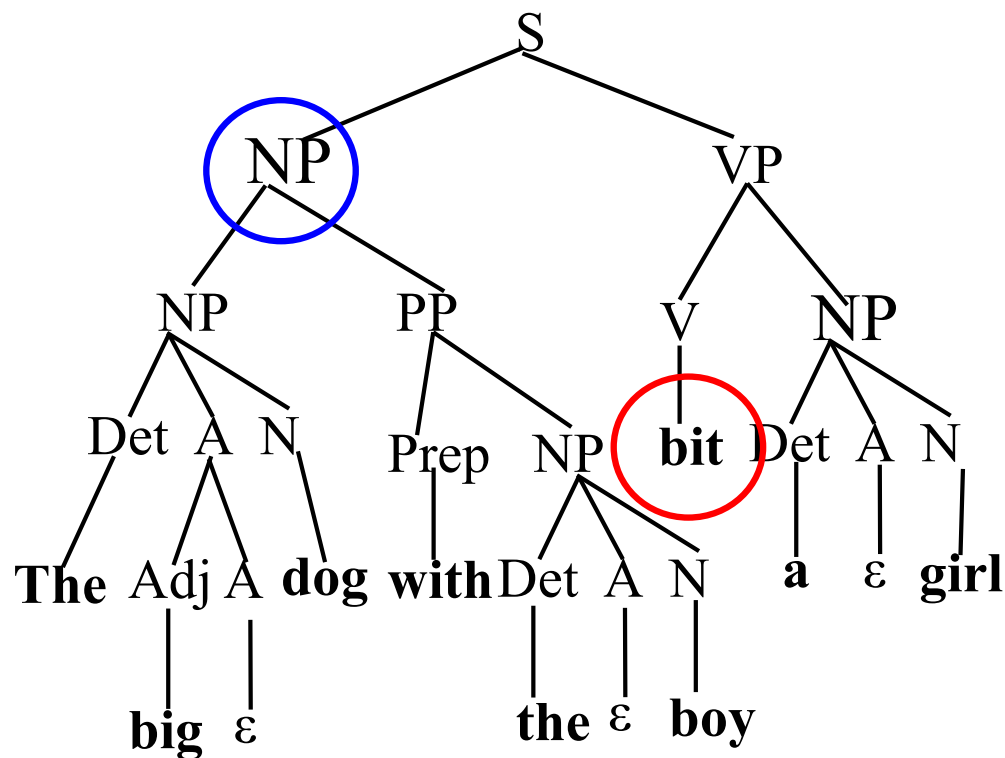
Parse Tree Path Feature

Path Feature Value:

V ↑ **VP** ↑ **S** ↓ **NP**



SRL Features

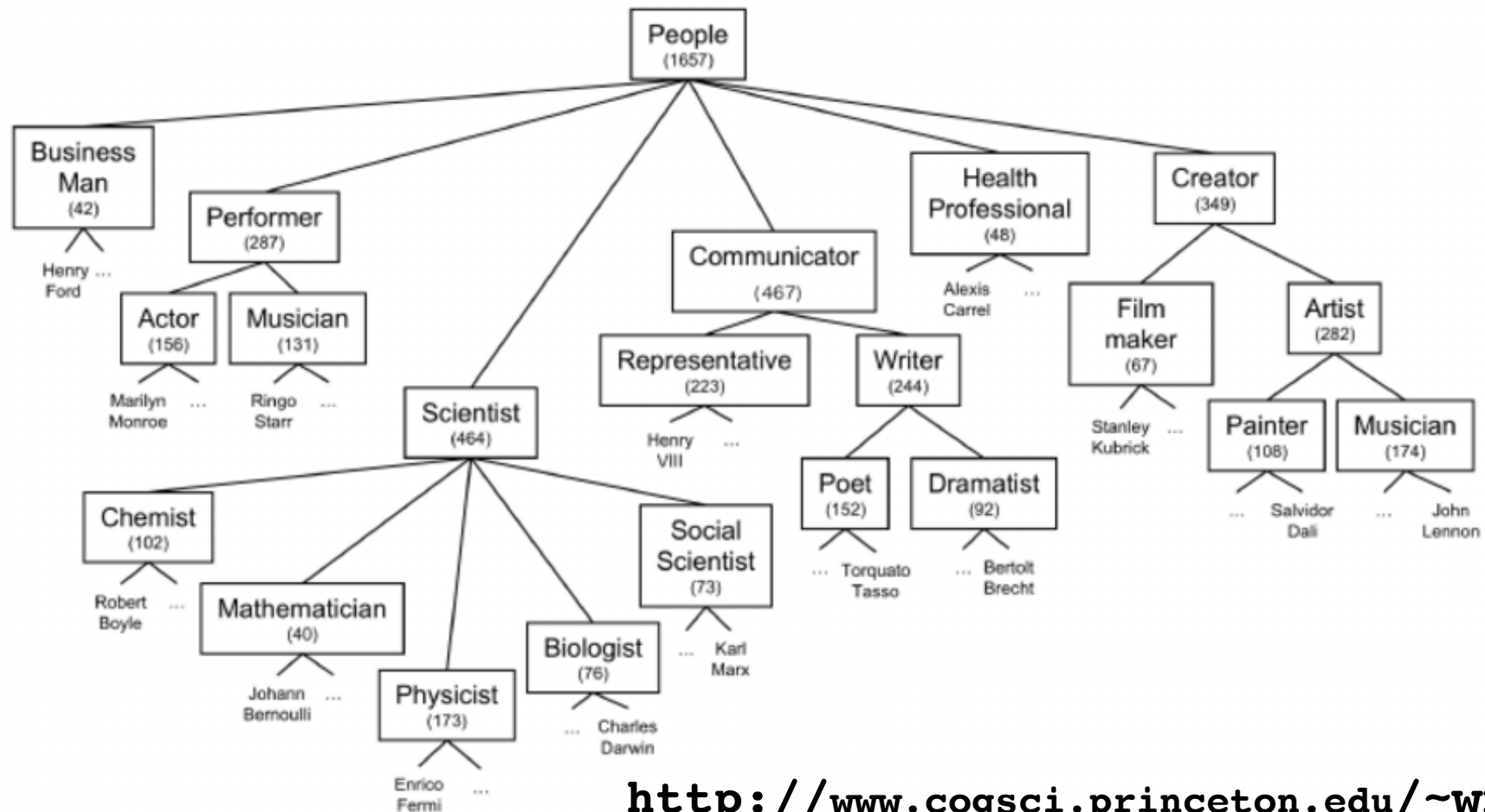


Phrase type	Parse Path	Position	Voice	Head word
NP	V↑VP↑S↓NP	precede	active	dog

Lexical Meaning

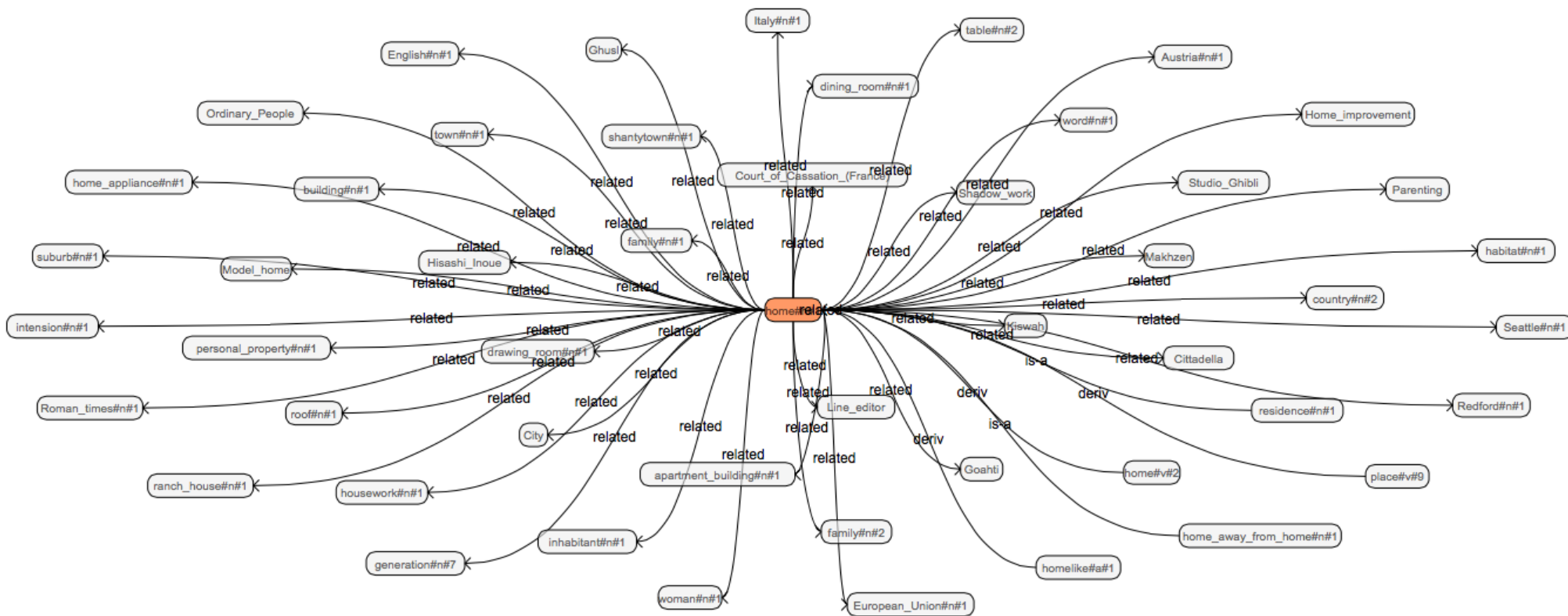
Word Meanings may help slot filling (e.g. a **food** is expected as patient of **eat**)

We need **lexical dictionaries**: A snapshot of the **WordNet** hierarchy



<http://www.cogsci.princeton.edu/~wn/>

BabelNet: A Rich and Multilingual Semantic Network



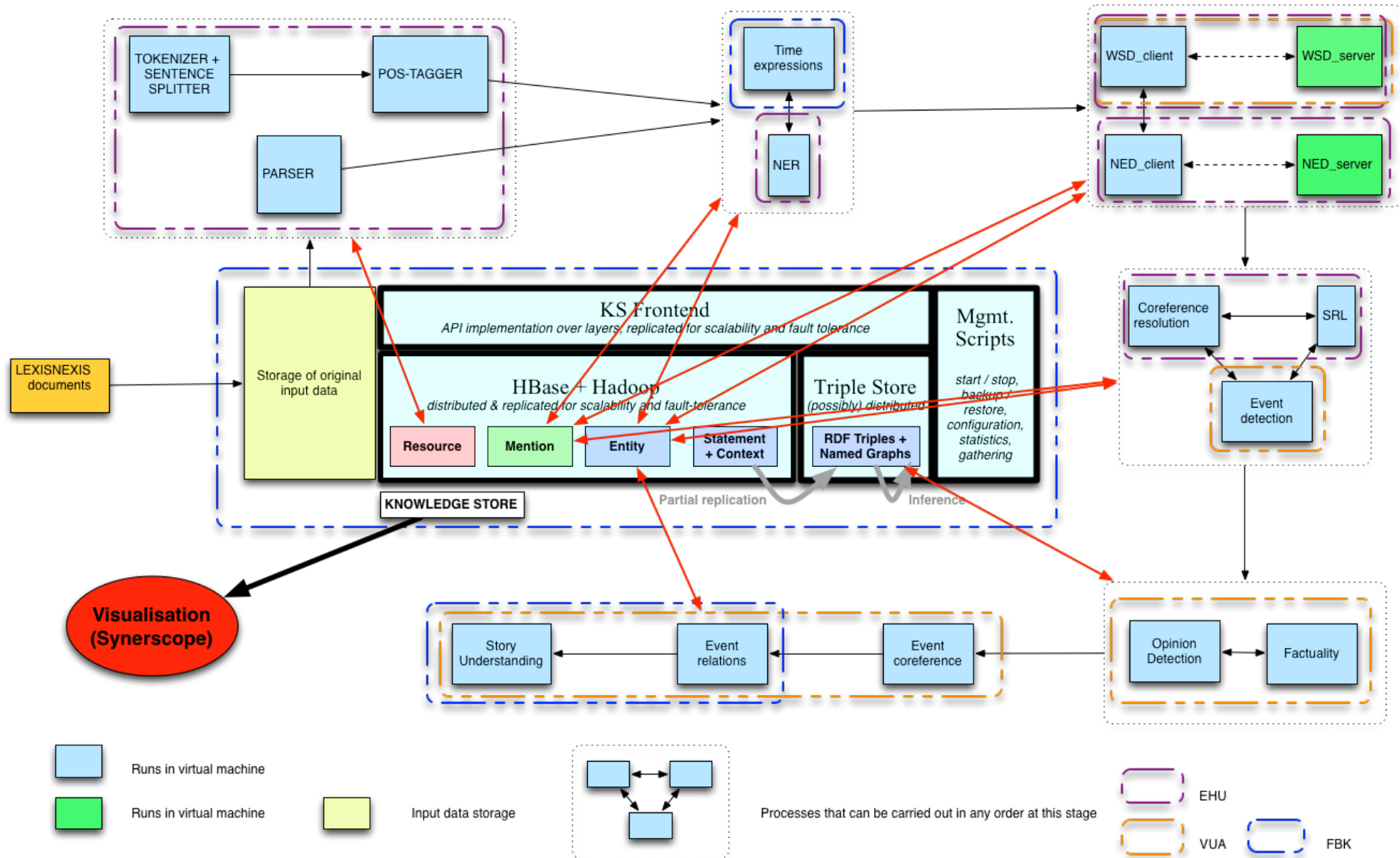
<http://babelnet.org>

Ontology Population and Linking

- Detecting and collecting as much as possible information from text (**unstructured**) and store it in a knowledge-base (**structured**). The goal is then to retrieve and use information from the KB rather than from texts.
- **Linking** against existing repositories
 - Wikipedia info-boxes
- **Cross-document coreference**: are we talking about the same entity / event?
 - Need of world knowledge to fill textual gaps
- Building a **KnowledgeGraph** (Google-like)

Fiat 500	
	
1970 Fiat 500 L	
Overview	
Manufacturer	Fiat
Production	1957–1975 3,893,294 units ^[1]
Assembly	Turin, Italy Desio, Italy ^[1] Termini Imerese (PA), Italy ^[1]
Designer	Dante Giacosa
Body and chassis	
Class	City car (A)
Body style	2-door saloon 3-door estate 3-door Panel van
Layout	RR layout

KnowledgeStore



A Multimodal KnowledgeStore

Interpreting / extracting / aligning knowledge from different media (e.g., video, commentary, images, text, ...)

Frame	Commentary	Knowledge
	“Sanchez, Sanchez,.. . goal. Sanchez equalizes for Chile”	dbpedia:Alexis_Sanchez scorestAt 32min
	“Yellow card for the Chilean defender”	dbpedia:Mauricio_Pinilla yellowCardAt 102min
	“Now is Marcelo turn, to kick the fourth penalty” “Marcelo. . . Goal”	dbpedia:Marcelo_Vieira kicks SuppPenalty4 SuppPenalty4 leadsTo goal



2. Representing Word Meaning with Vectors

What is the meaning of “bardiwac”?

- He handed her **glass** of **bardiwac**.
- Beef dishes are made to complement the **bardiwacs**.
- Nigel staggered to his feet, face flushed from **too much bardiwac**.
- Malbec, one of the lesser-known **bardiwac grapes**, responds well to Australia’s sunshine.
- I dined off bread and cheese and this excellent **bardiwac**.
- The **drinks** were delicious: **blood-red bardiwac** as well as light, sweet Rhenish.

⇒ **Bardiwac ???**

What is the meaning of “bardiwac”?

- He handed her **glass** of **bardiwac**.
- Beef dishes are made to complement the **bardiwacs**.
- Nigel staggered to his feet, face flushed from **too much bardiwac**.
- Malbec, one of the lesser-known **bardiwac grapes**, responds well to Australia’s sunshine.
- I dined off bread and cheese and this excellent **bardiwac**.
- The **drinks** were delicious: **blood-red bardiwac** as well as light, sweet Rhenish.

⇒ **bardiwac** is a heavy red alcoholic beverage made from grapes

Geometric interpretation of Word Meaning

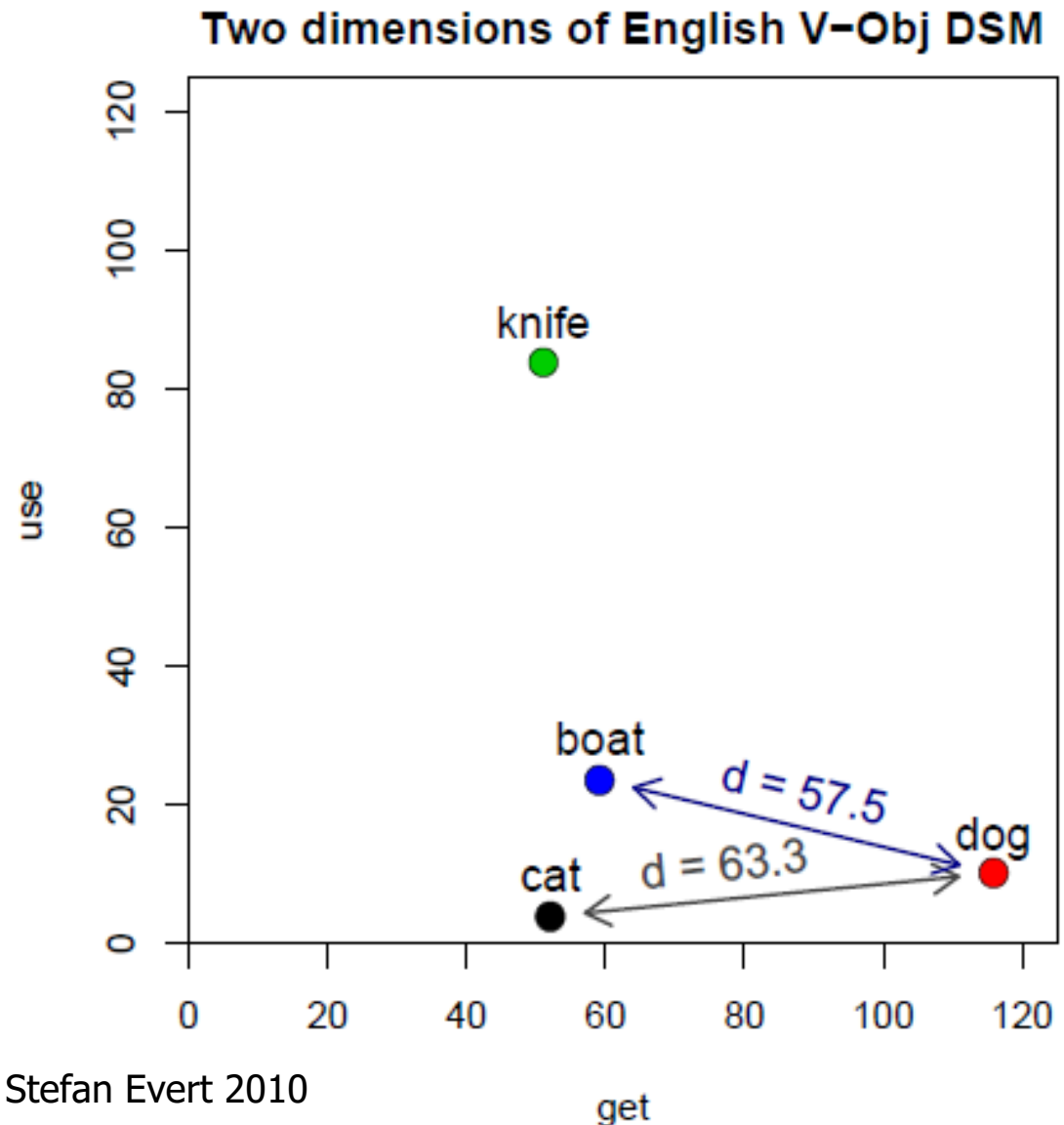
- row vector \mathbf{x}_{dog} describes frequency of word *dog* in the corpus
- can be seen as coordinates of point in n -dimensional Euclidean space \mathbb{R}^n

	get	see	use	hear	eat	kill
knife	51	20	84	0	3	0
cat	52	58	4	4	6	26
dog	115	83	10	42	33	17
boat	59	39	23	4	0	0
cup	98	14	6	2	1	0
pig	12	17	3	2	9	27
banana	11	2	2	0	18	0

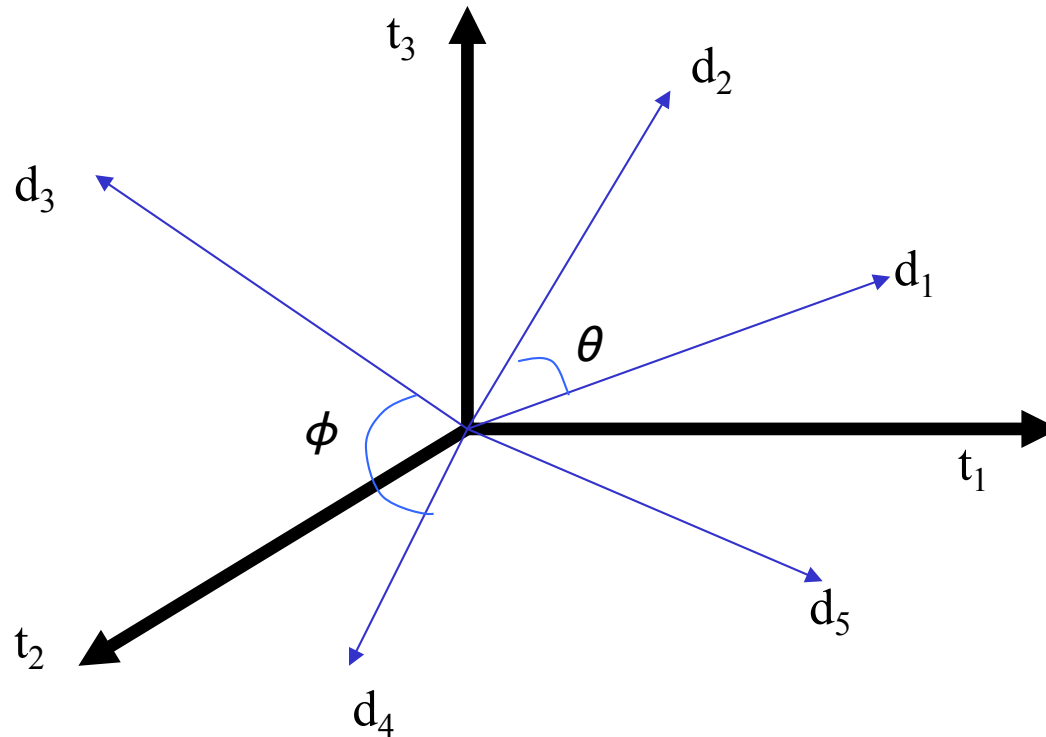
co-occurrence matrix M

Distance and similarity

- illustrated for two dimensions: *get* and *use*: $\mathbf{x}_{\text{dog}} = (115, 10)$
- similarity = spatial proximity (**Euclidean distance**)
- location depends on frequency of noun ($f_{\text{dog}} \approx 2.7 \cdot f_{\text{cat}}$)



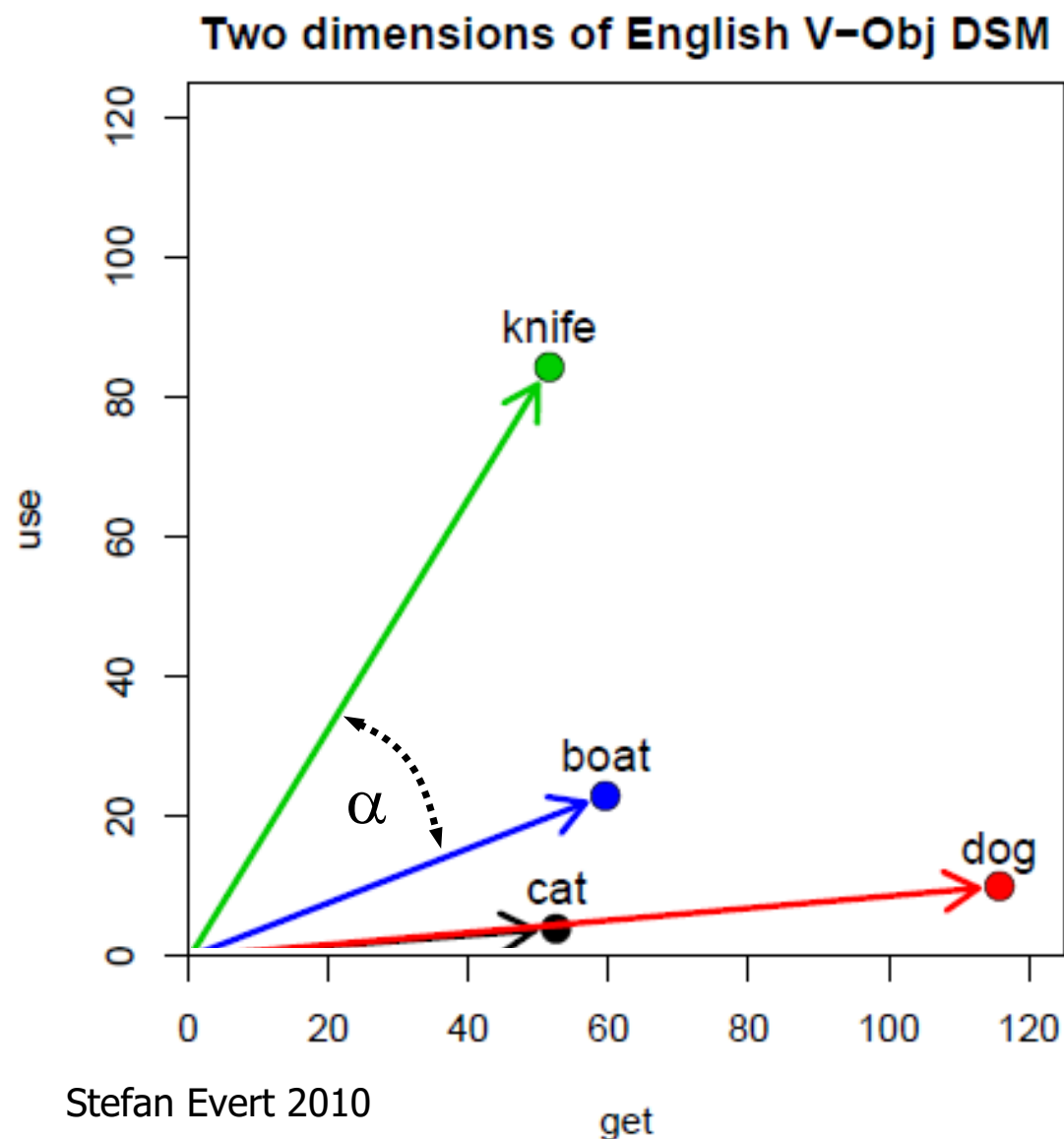
Vector Space



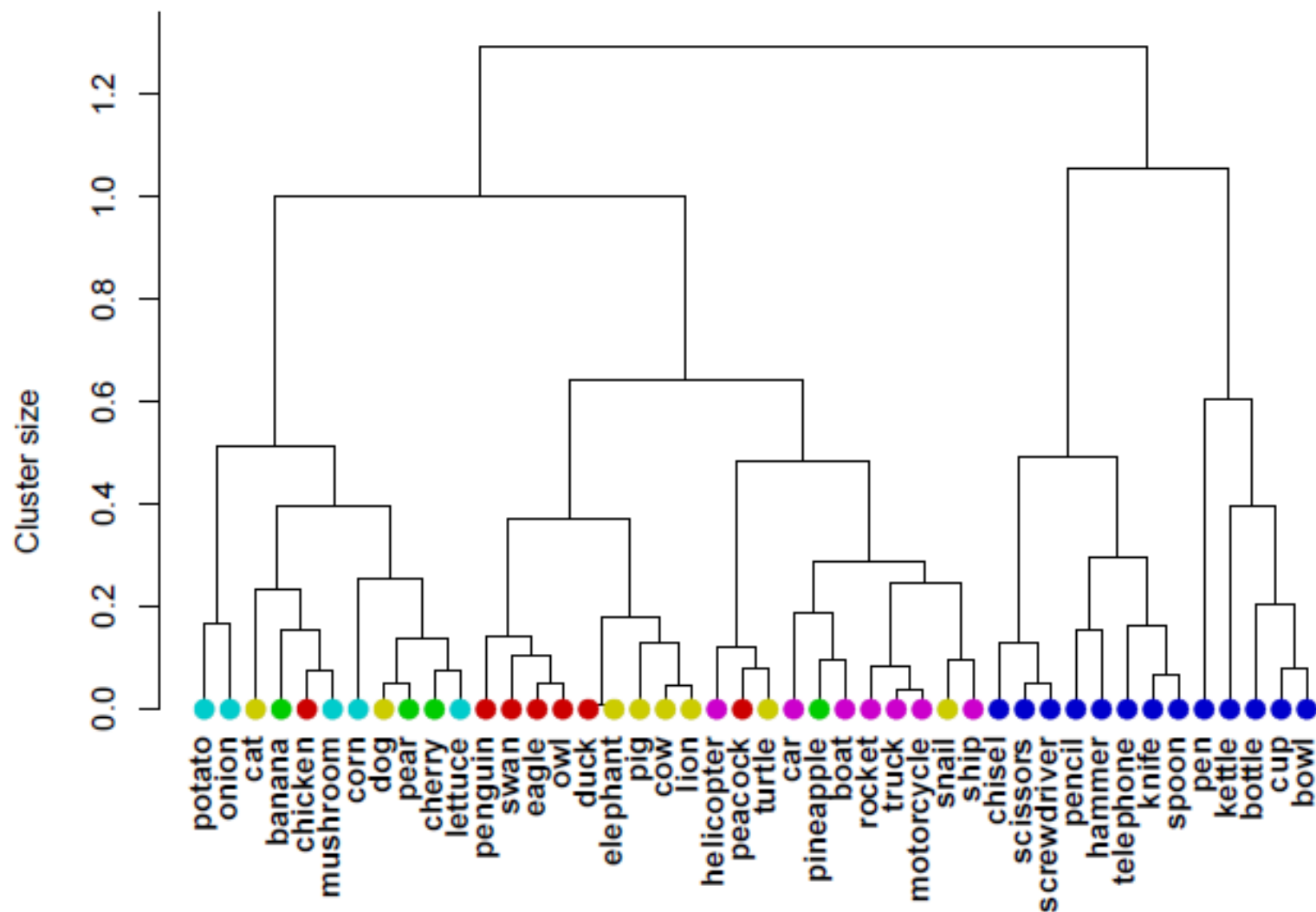
Postulate: Words that are “close together” in the vector space have similar meaning.
There is **one dimension for each term in the document collection.**

Angle and similarity

- **direction** more important than location
- normalise “length” $||\mathbf{x}_{\text{dog}}||$ of vector
- or use angle α as distance measure



Application: Clustering



Embeddings: Dimensionality Reduction

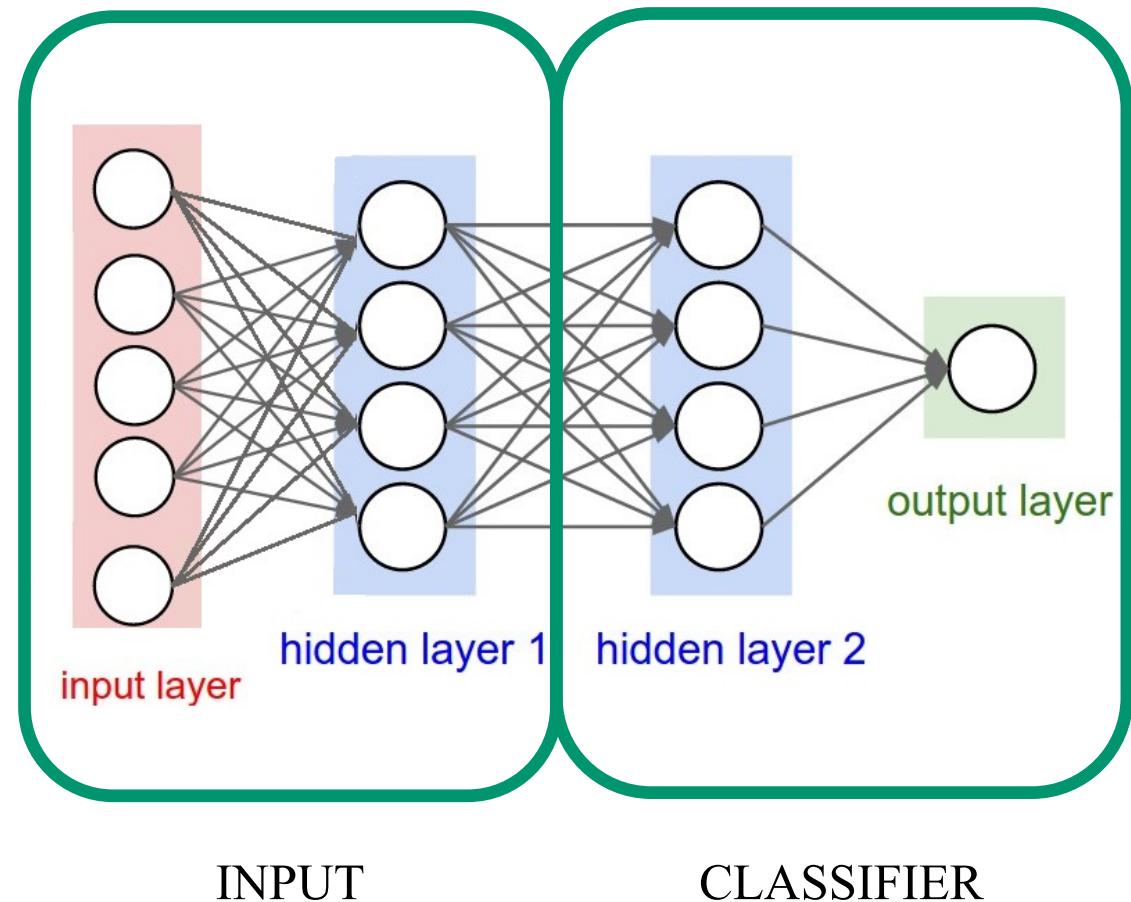
- **One-Hot vector** (a vector with only 1)
 - *Social*
= [0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]
 - *Private*
= [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0]
 - Easy to obtain, but computationally expensive (vector size = vocabulary size)
 - Do not generalize
 - Risk of overfitting (due to size)
- **Dense vectors** have better performances and reduce the risk of overfitting:
 - *Social* = [0.5, 0.3, 1.0]
 - *Private* = [0.2, 1.0, 0.4]

Using Neural Networks to Build Embeddings

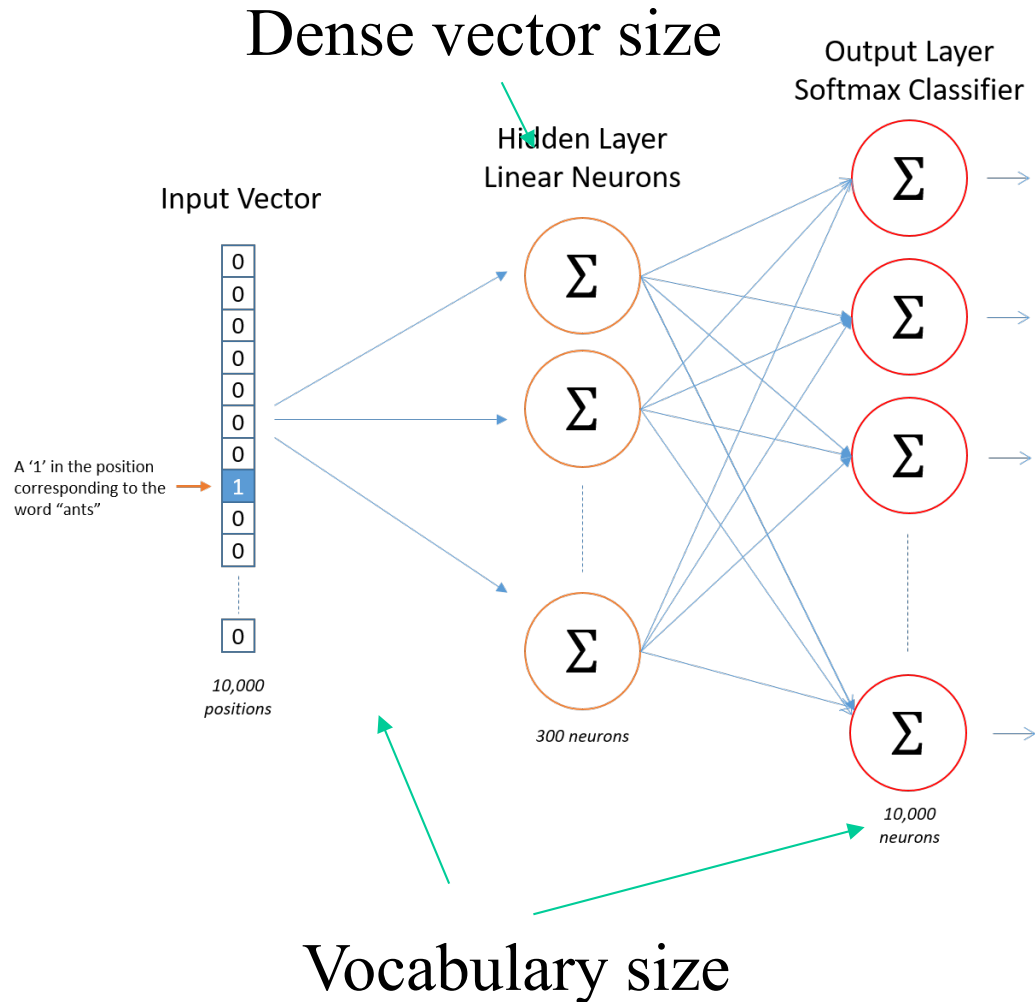
Transform the original sparse vector

INTO

A new dense and optimized vector



Neural Word Embedding

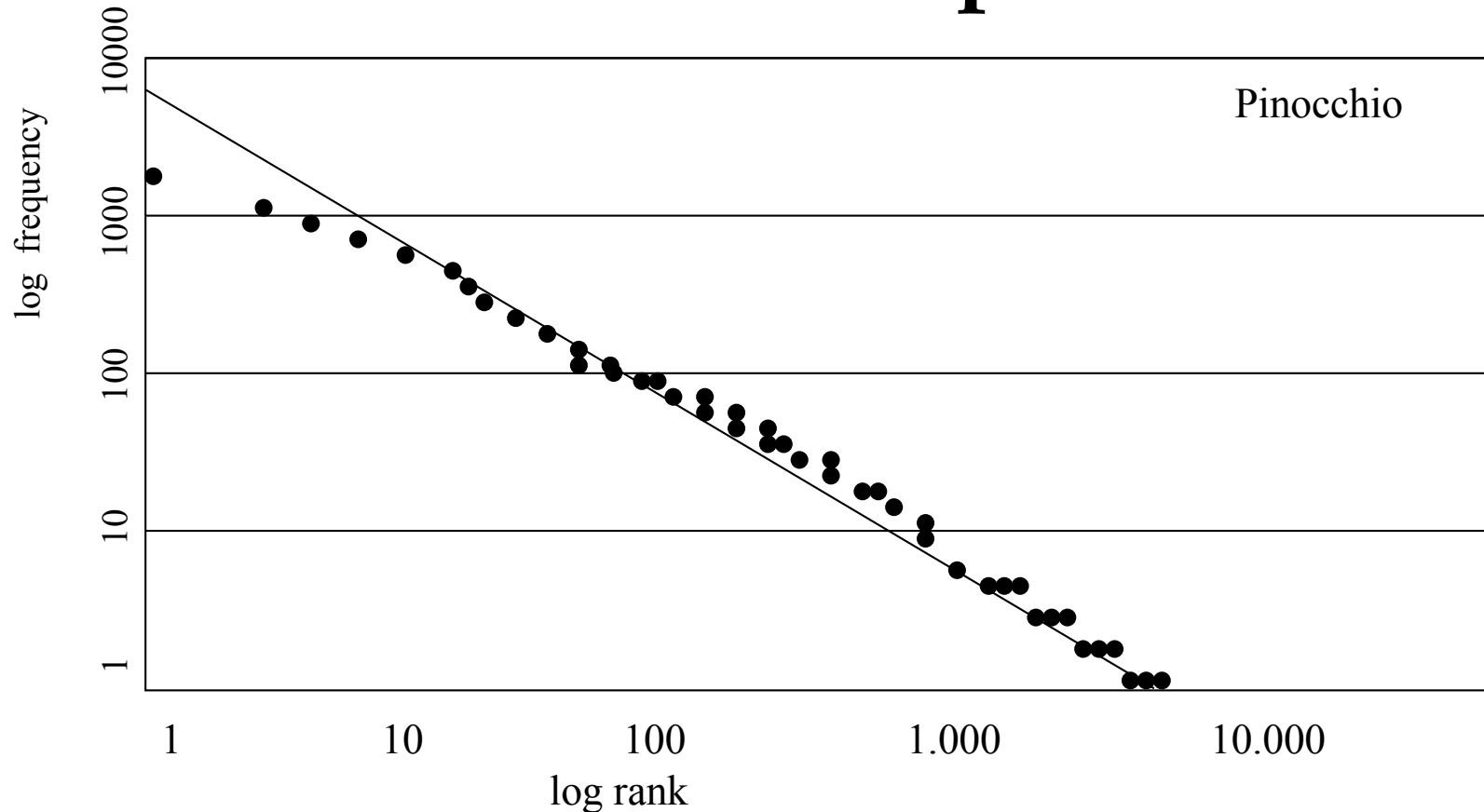


In this example we suppose 10,000 words in the vocabulary. The network extracts dense vectors of 300 elements for each word.



3. Representing Meaning with Probabilistic Models

Linguistic events are not uniformly distributed (Zipf Law)



Logarithmic scale: $\log f(z) = \log K - a \log z$

With $K = 6185$, i.e. $f \times z = 6185$

The slope of the curve is defined by the coefficient a (1)

Linguistic Events

- **Examples of linguistic events**
 - The probability that the word “dog” occurs after the word “the”
 - The probability that the word “race” is a NOUN given that the word before is an ARTICLE
 - The probability that the translation of “chair” is “sedia” given that the word before is translated as “tavolo”
 - The probability that “dog” is a SUBJECT of the verb “bark”
 - The probability that “John Smith” is a PERSON
 - The probability of the sentence “My dog barks”
 - The probability of a text like “Pinocchio” to be generated
- **Estimate posterior probabilities for linguistic events:**
 - use corpora for empirical observations

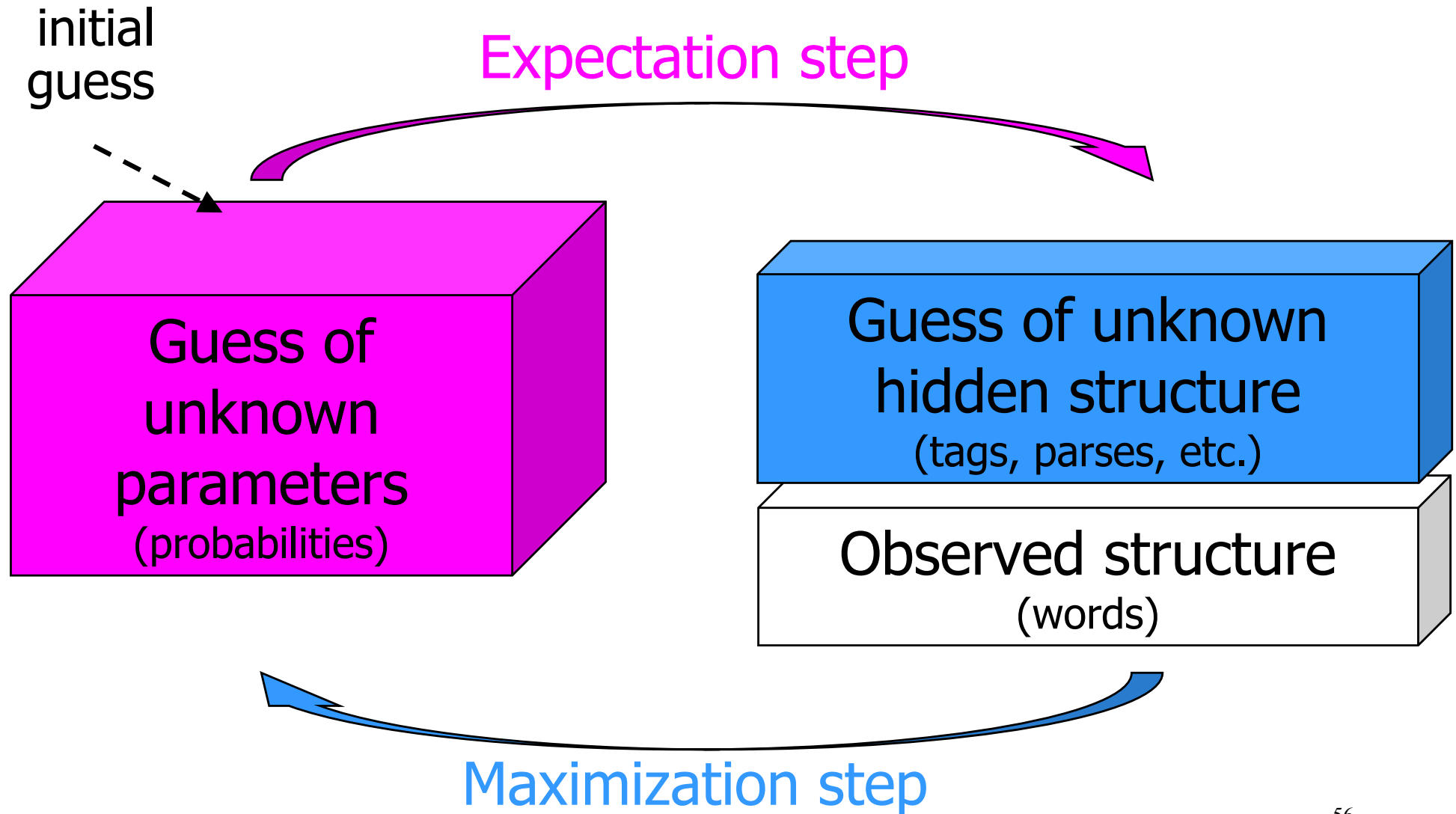
Language Model

- Define a statistical (stochastic) model of language
 1. Observe linguistic phenomena on a training corpus
 2. Estimate probabilities of linguistic events (i.e. the language model)
 3. Apply the model to a new (not yet observed) set of data
- A text T is seen as a sequence of simple events $e_1, e_2, \dots, e_{|T|}$, (not necessarily independent), each of them representing the occurrence of a word in T .
- Example:

*The₁ dog₂ sleeps₃ here₄ the₅ dog₆ sleeps₇ there₈ the₉ dog₁₀ eats₁₁
here₁₂ the₁₃ cat₁₄ eats₁₅ there₁₆ a₁₇ cat₁₈ sleeps₁₉*

- $e_1 = \text{The}, e_2 = \text{dog}, \dots$
 - $p(e_i = w)$ is the probability that the word type w occurs as the i event in T

Estimate probabilities for linguistic events: EM Approach



For Hidden Markov Models

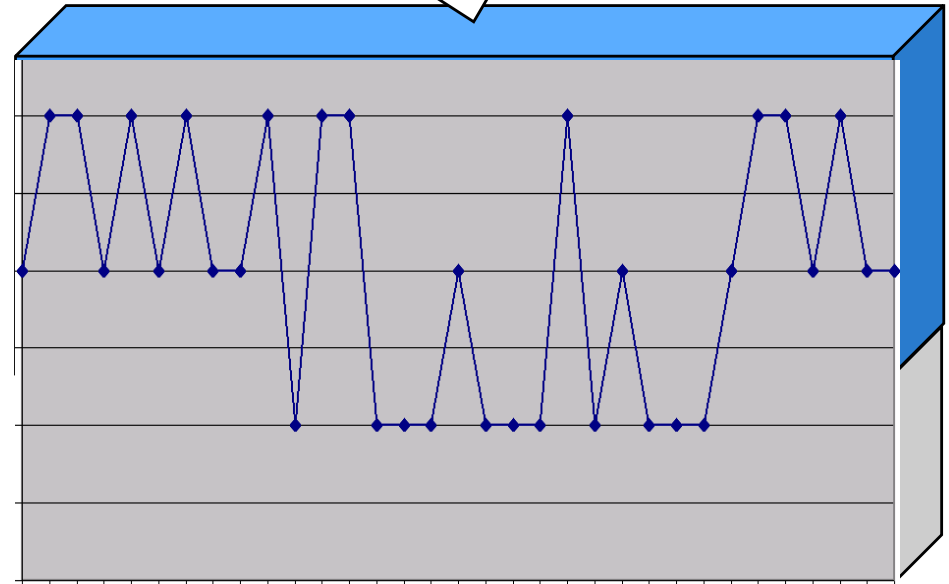
initial
guess

E step

Guess of

	$p(\dots C)$	$p(\dots H)$	$p(\dots START)$
$p(1 \dots)$	0.7	0.1	
$p(2 \dots)$	0.2	0.2	
$p(3 \dots)$	0.1	0.7	
$p(C \dots)$	0.8	0.1	0.5
$p(H \dots)$	0.1	0.8	0.5
$p(STOP \dots)$	0.1	0.1	0

(probabilities)



M step

For Hidden Markov Models

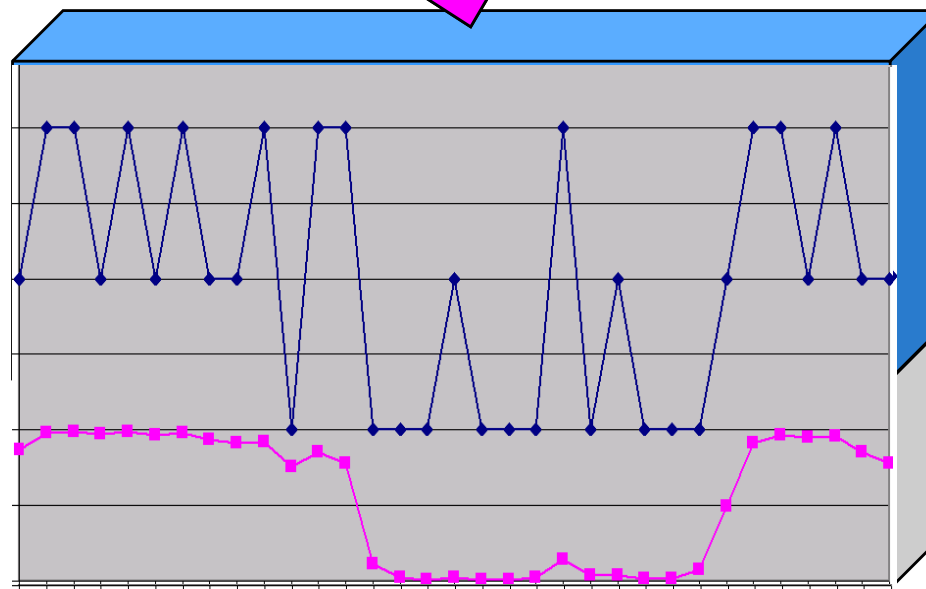
initial
guess

E step

Guess of

	$p(\dots C)$	$p(\dots H)$	$p(\dots START)$
$p(1 \dots)$	0.677	0.058	
$p(2 \dots)$	0.219	0.425	
$p(3 \dots)$	0.105	0.517	
$p(C \dots)$	0.876	0.093	0.129
$p(H \dots)$	0.109	0.865	0.871
$p(STOP \dots)$	0.015	0.042	0

(probabilities)



M step

For Hidden Markov Models

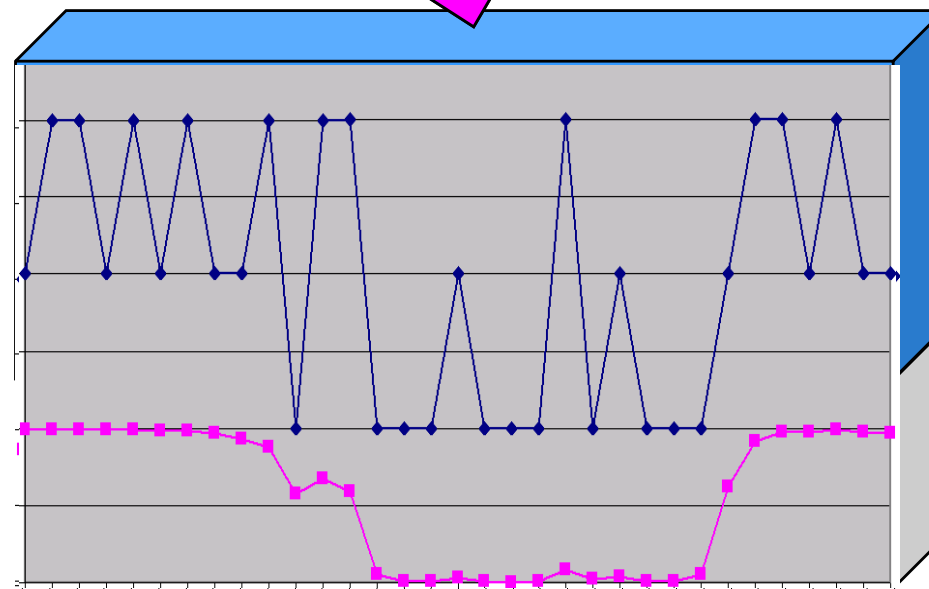
initial
guess

E step

Guess of

	p(... C)	p(... H)	p(... START)
p(1 ...)	0.697	0.04	
p(2 ...)	0.171	0.464	
p(3 ...)	0.132	0.496	
p(C ...)	0.904	0.077	0.012
p(H ...)	0.094	0.87	0.988
p(STOP ...)	0.002	0.053	0

(probabilities)



M step

NLP Research at FBK-irst (Trento)

- **Human Language Technologies (HLT)** research unit
 - Natural Language processing
 - Machine Translation
 - Speech Recognition
- **Language Technologies**
 - TextPro platform: basic text processing in Italian and English
 - Speech Recognition system
- **High level education**
 - PhD students
 - Internship, thesis
 - LCT seminars
- **Spin-off:** Pervoice, Spazio Dati, Cross Library Service, Semantic Valley

Textbooks

1. [JM] Jurafsky-Martin, *Speech and Language Processing*, 2009.
(third edition in preparation)
2. [MRS] Manning, Raghavan, Schutze, *An Introduction to Information Retrieval*, CUP, 2008.
3. [CFL] Clark, Fox Lappin, *The handbook of Computational Linguistics and Natural Language Processing*, 2010.
4. [MS] Manning-Schutze, *Foundation of Statistical Natural Language Processing*, 1999.
5. [CL] Mitkof, *Handbook of Computational Linguistics*, 2003.
6. [JM] Jackson-Moulinier, *Natural Language Processing for Online Applications*, 2002..
7. [AE] Agirre-Edmond, *Word Sense Disambiguation*, 2006.
8. [WN] Fellbaum, *WordNet*, MIT Press, 1998.
9. [AL] Allen, *Natural Language Processing*, 1995.