

Information Extraction

3 sessions in the Module INF347
at the École nationale supérieure des
Télécommunications
in Paris/France in Summer 2011

by [Fabian M. Suchanek](#)

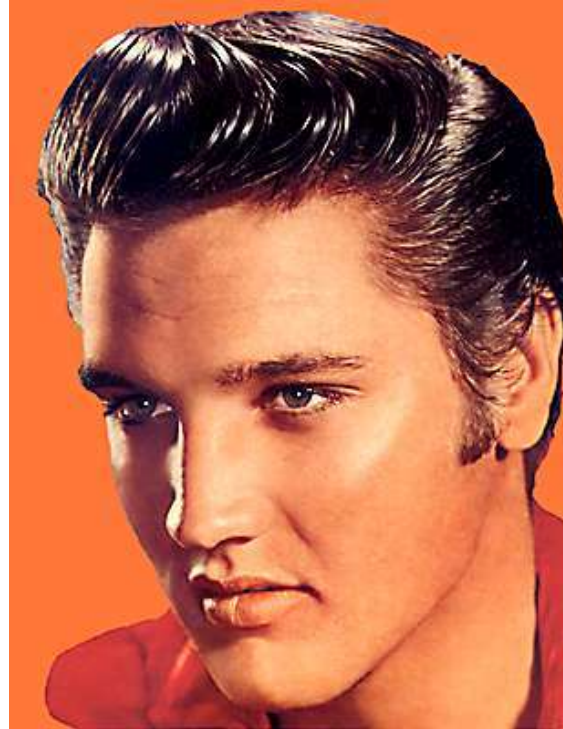


This document is available under a
[Creative Commons Attribution Non-Commercial License](#)

Organisation

- 3 sessions (each 1.5h) on Information extraction
- 1 lab session 1.5h
- Web-sites:
<http://www.infres.enst.fr/~danzart/INF347/>
<http://suchanek.name/> → Teaching

Motivation



*Elvis, when I
need you, I
can hear you!*

Elvis Presley
1935 - 1977

Will there ever be someone like him again?

Motivation



Another Elvis

Elvis Presley: The Early Years

Elvis spent more weeks at the top of the charts than any **other** artist.

www.fiftiesweb.com/elvis.htm

Motivation



Another singer called Elvis, young

[Personal relationships of Elvis Presley – Wikipedia](#)

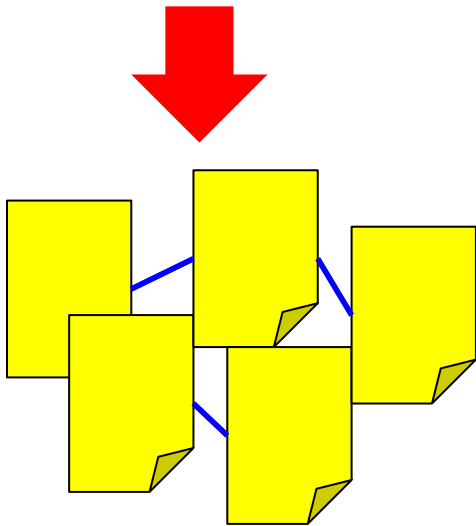
...when Elvis was a **young** teen.... **another** girl whom the **singer's** mother hoped Presley would The writer **called Elvis** "a hillbilly cat"

en.wikipedia.org/.../Personal_relationships_of_Elvis_Presley

Motivation

Google™

Another Elvis



X

```
SELECT * FROM person
WHERE gName='Elvis'
AND occupation='singer'
```

**Information
Extraction**



GName	FName	Occupation
Elvis	Presley	singer
Elvis	Hunter	painter
...	...	

- 1: Elvis Presley
- 2: Elvis ...
3. Elvis ...

Definition of IE

Information Extraction (IE) is the process of extracting structured information (e.g., database tables) from unstructured machine-readable documents (e.g., Web documents).

Elvis Presley was a famous rock singer.
...
Mary once remarked that the only attractive thing about the painter Elvis Hunter was his first name.

**Information
Extraction**



GName	FName	Occupation
Elvis	Presley	singer
Elvis	Hunter	painter
...	...	

“Seeing the Web as a table”

Motivating Examples

579 Jobs in Northern California

Refine your search

Keyword(s)

Search Results

Job Title / Description (show titles only)	Company
<p>RN-Registered Nurse/LVN-Licensed Vocational Nurse - View similar jobs</p> <p>Job type: Full-Time/Part-Time</p> <p>Maxim's office in Sherman Oaks is seeking compassionate Registered Nurses (RN) and Licensed ... Maxim's office in Sherman Oaks is seeking...</p> <p>View full job description Save to MyCareerBuilder Email to a friend</p>	Maxim Healthcare Services, Inc
<p>Nurse Practitioner - Acute Care Nurse Practitioner - View similar jobs</p> <p>Job type: Full-Time</p> <p>Vanderbilt University Medical Center is currently hiring Nurse Practitioners to join our team ... Vanderbilt University Medical Center is...</p> <p>View full job description Save to MyCareerBuilder Email to a friend</p>	Vanderbilt University Medical Center (VUMC)
QA Engineer - Release Engineer - Quality Assurance	\$50k - \$90k
Senior Flash Memory Technologist - Storage Architect - SSD	\$160k - \$200k

Title	Type	Location
Business strategy Associate	Part time	Palo Alto, CA
Registered Nurse	Full time	Los Angeles
...	...	

Motivating Examples

Biography for

Elvis Presley [More at IMDb](#)

Date of Birth

[8 January 1935](#), [Tupelo, Mississippi, USA](#)

Date of Death

[16 August 1977](#), [Memphis, Tennessee, USA](#) (cardiac arrhythmia)

Birth Name

Elvis Aron Presley

Nickname

The Pelvis
The King
The King Of Rock 'n'

Height

6' (1.83 m)

Mini Biography

Elvis Aaron Presley

Name	Birthplace	Birthdate
Elvis Presley	Tupelo, MI	1935-01-08
...	...	



Biography

[Overview](#) / [1935-1957](#) / [1958-1965](#) / [1966-1969](#) / [1970-1977](#)

Overview

Elvis Aaron Presley, in the humblest of circumstances, was born to Vernon and Gladys Presley in a [two-room house in Tupelo, Mississippi](#) on January 8, 1935. His twin brother, Jessie Garon, was stillborn, leaving Elvis to grow up as an only child. He and his parents moved to [Memphis, Tennessee](#) in 1948, and Elvis graduated from Humes High School there in 1953.

Motivating Examples

Information Extraction: Techniques and Challenges

Ralph Grishman


Information Integration Papers

[Answering Queries Using Templates With Binding Patterns](#). In PODS 1995. specify binding patterns.

[The TSIMMIS Approach to Mediation: Data Models and Languages](#). A survey appears in *J. Intelligent Information Systems* 8:2, pp. 117-132, March, 1997.

Author	Publication	Year
Grishman	Information Extraction...	2006
...

Motivating Examples



Ballroom Dance Shoe
1 new from **\$49.95**
 ★★☆☆☆ (5)
[Show only So Danca items](#)

X-Strap Ballroom Dance Shoe
1 new from **\$49.95**
 ★★★★★ (5)
[Show only So Danca items](#)



Dynex™ - 32" Class / 720p / 60Hz / LCD HDTV
 Model: DX-32L150A11 | SKU: 9558089
 ★★★★★ 3.8 of 5 (180 reviews)
[Check Shipping & Availability](#) ▶
[Compare](#)

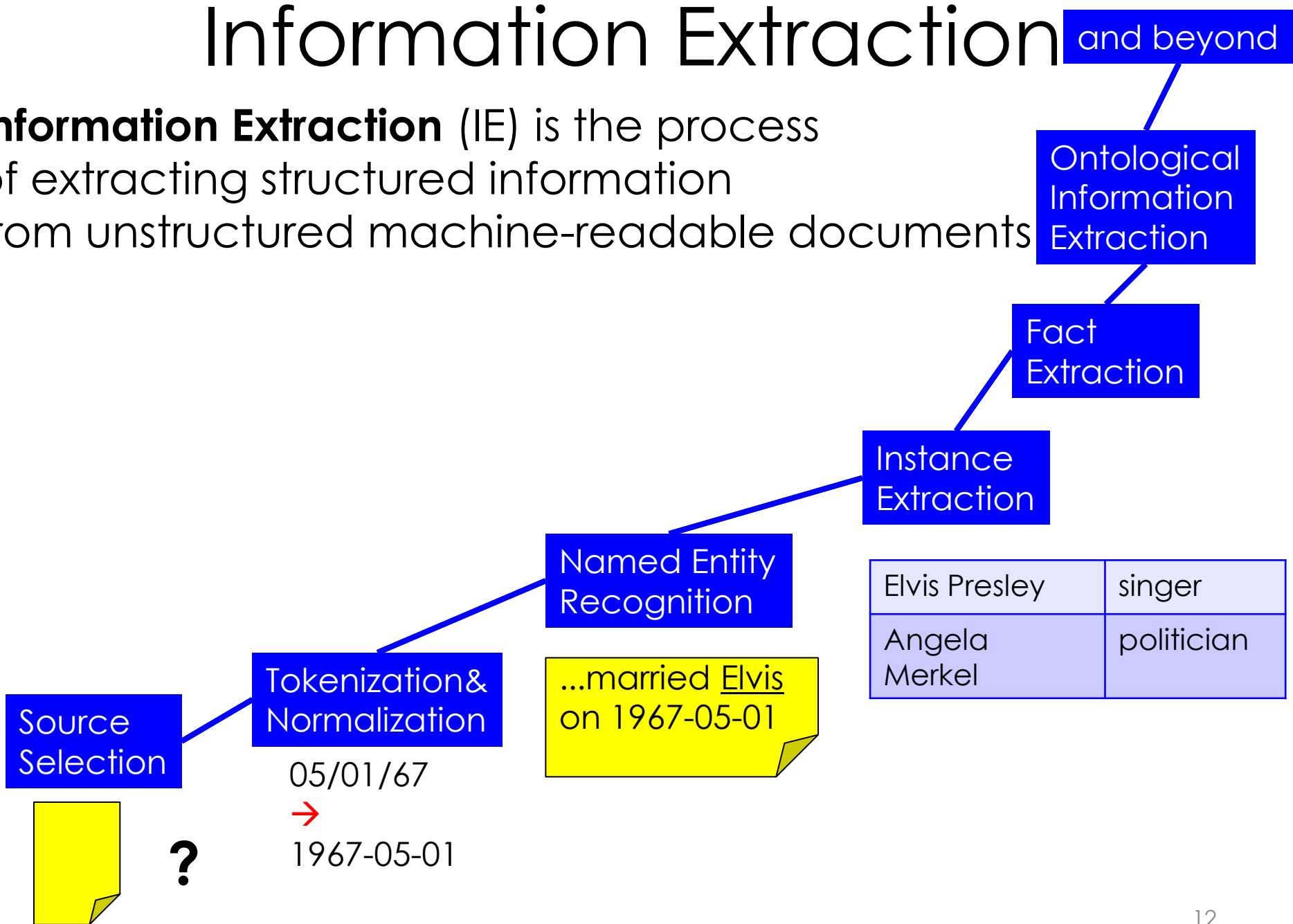


Dynex™ - 24" Class / 1080p / 60Hz / LCD HDTV
 Model: DX-24L150A11 | SKU: 9848048
 ★★★★★ 4.3 of 5 (54 reviews)
[Check Shipping & Availability](#) ▶
[Compare](#)

Product	Type	Price
Dynex 32"	LCD TV	\$1000
...	...	

Information Extraction and beyond

Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents



The Web



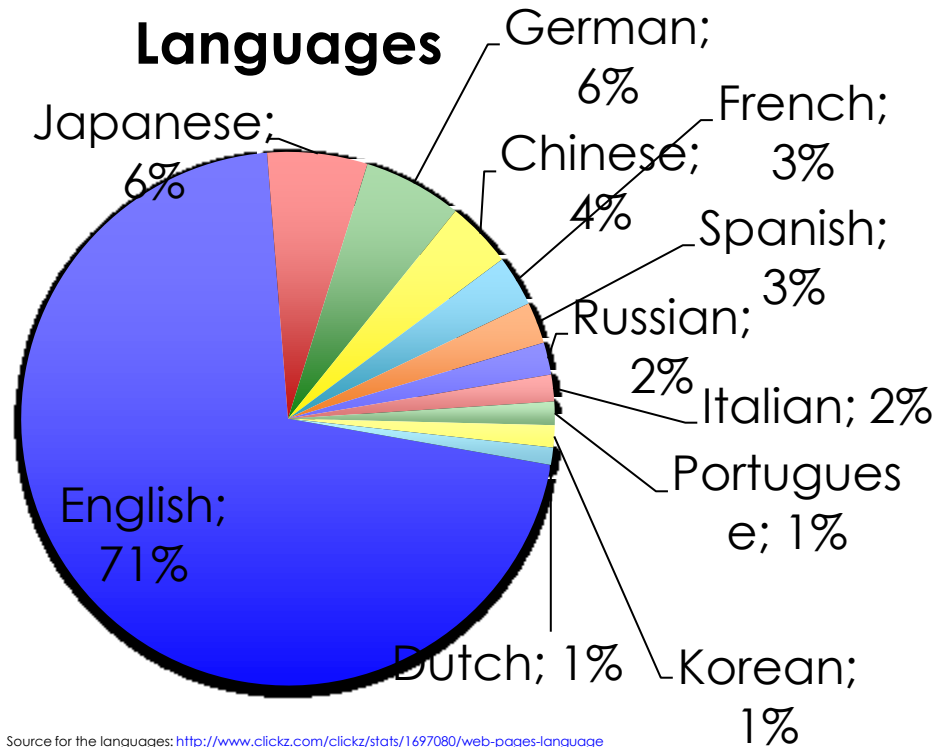
Relational Transducers for Electronic Commerce

Serge Abiteboul*
I.N.R.I.A.-Rocquencourt
Serge.Abiteboul@inria.fr

Victor Vianu*
U.C. San Diego
vianu@cs.ucsd.edu

Brad Fordham
Oracle Corporation
bfordham@us.oracle.com

(1 trillion Web sites)



Source for the languages: <http://www.clickz.com/clickz/stats/1697080/web-pages-language>
Need not be correct

IE Restricted to Domains

Restricted to one Internet Domain (e.g., Amazon.com)

Restricted to one Thematic Domain (e.g., biographies)

Restricted to one Language (e.g., English)

amazon.com. VIEW CART

WELCOME YOUR STORE BOOKS ELECTRONICS DVD TOYS & GAMES

SEARCH BROWSE SUBJECTS

Get \$5 off

Machine Learning by Tom M. Mitchell

LOOK INSIDE!

Learning in Graphical Models by Michael Irwin Jordan (Editor)

List Price: \$60.00
Price: \$60.00

This item ships for FREE with Super Saver Shipping

Availability: Usually ships within 2 to 3 days

Used & new from \$20.00

Edition: Paperback | All Editions

See more product details

Great Buy

Buy this book with *Probabilistic Reasoning in Intelligent Systems*

Buy Together Today: \$128.95

Buy both now!

Jason D. M. Rennie

Massachusetts Institute of Technology
MIT AI Lab NE43-733
200 Technology Sq.
Cambridge, MA 02139

Research Interests

My main interests lie in the automated analysis of data for the purposes of classification, estimation and the acquiring of new knowledge. I have both interests in applying such techniques to real-world problems and in the analysis of general, abstract, and theoretical problems.

L. Douglas Baker

Home Address: available upon request

Office Address: Wean Hall, 8102 School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213

Office Phone: (412) 683-6036

Home Page: http://www.cs.cmu.edu/~ldbpp

Objective: A position in a dynamic, highly-skilled applied research and development team using statistical machine learning to solve large-scale, real-world tasks such as Information Retrieval and Text Classification.

Education: Carnegie Mellon University (Pittsburgh, PA) - Ph.D., Computer Science, in progress; M.S., Computer Science, 1999; Technical University of Berlin (Berlin, Germany) - Exchange Fellow, 1992-1993; University of Michigan (Ann Arbor, MI) - M.S.E., Computer Science and Engineering, 1984 B.S.E., Computer Engineering, Summa Cum Laude, 1982

Research Experience: Carnegie Mellon University (1984-present)

I am currently pursuing my dissertation research: a hierarchical probabilistic model for novelty detection in text. This work is being done as part of the Topic Detection and Tracking project at CMU under the direction of Victor Hancock. The

8:30 - 9:30 AM	Invited Talk: Plausibility Measures: A General Approach <i>Joseph Y. Halpern, Cornell University</i>		
9:30 - 10:00 AM	Coffee Break		
10:00 - 11:30 AM	Technical Paper Sessions:		
Cognitive Robotics	Logic Programming	Natural Language Generation	Complexity Analysis
739: A Logical Account of Causal and Topological Maps <i>Emilio Remolina and Benjamin Kuipers</i>	116: A-System: Solving through Abduction <i>Marc Denecker, Antonis Kakas, and Bert Van</i>	758: Title Generation for Machine-Translated Documents <i>Rong Jin and Alexander G. Hauptmann</i>	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories <i>Marco Cadoli,</i>

Dr. Steven Minton - Founder/CTO

Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC's Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

Frank Huybrechts - COO
Mr. Huybrechts has over 20 years of

- Press Contact
- General information
- Directions maps

Finding the Sources



How can we find the documents to extract information from?

- The document collection can be given a priori (**Closed** Information Extraction)
e.g., a specific document, all files on my computer, ...
- We can aim to extract information from the entire Web (**Open** Information Extraction)
For this, we need to crawl the Web (see previous class)
- The system can find by itself the source documents
e.g., by using an Internet search engine such as Google

Scripts

Elvis Presley was a rock star.

(Latin script)

猫王是摇滚明星

(Chinese script,
“simplified”)

אלביס היה כוכב רוק

(Hebrew)

وكان ألفيس بريسلي نجم الروك

(Arabic)

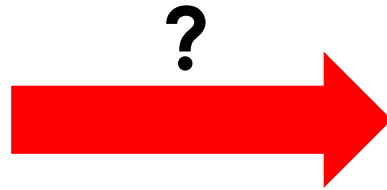
(Korean script)

Elvis Presley ถูกดาวร็อก

(Thai script)

Char Encoding: ASCII

100,000 different characters from 90 scripts



One byte with 8 bits per character (can store numbers 0-255)

How can we encode so many characters in 8 bits?

- Ignore all non-English characters (**ASCII standard**)

26 letters + 26 lowercase letters + punctuation \approx 100 chars

Encode them as follows:

A=65,

B=66,

C=67,

...

Disadvantage: Works only for English

Char Encoding: Code Pages

- For each script, develop a different mapping (a **code-page**)

Hebrew code page:, 226=כ,...

Western code page:, 226=à,...

Greek code page:, 226=α, ...

(most code pages map characters 0-127 like ASCII)

([Example](#))

Disadvantages:

- We need to know the right code page
- We cannot mix scripts

Char Encoding: HTML

- Invent special sequences for special characters (e.g., **HTML entities**)

è = è, ...

([Example](#), [List](#))

Disadvantage: Very clumsy for non-English documents

Char Encoding: Unicode

- Use 4 bytes per character (**Unicode**)

...65=A, 66=B, ..., 1001=α, ..., 2001=

([Example](#), [Example2](#))

Disadvantage: Takes 4 times as much space as ASCII

Char Encoding: UTF-8

- Compress 4 bytes Unicode into 1-4 bytes (**UTF-8**)

Characters **0 to 0x7F** in Unicode:

Latin alphabet, punctuation and numbers

Encode them as follows:

0xxxxxxx

(i.e., put them into a byte, fill up the 7 least significant bits)

A = 0x41 = 1000001



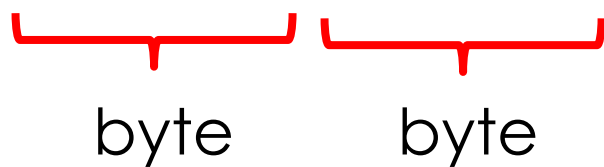
01000001

Advantage: An UTF-8 byte that represents such a character is equal to the ASCII byte that represents this character.

Char Encoding: UTF-8

Characters $0x80-0x7FF$ in Unicode (11 bits):
Greek, Arabic, Hebrew, etc.

Encode as follows:

$110xxxxx$ $10xxxxxx$

byte byte

$\zeta = 0xE7 = 00011100111$



11000011 10100111

f a ζ a d e

0x66 0x61 0xE7 0x61

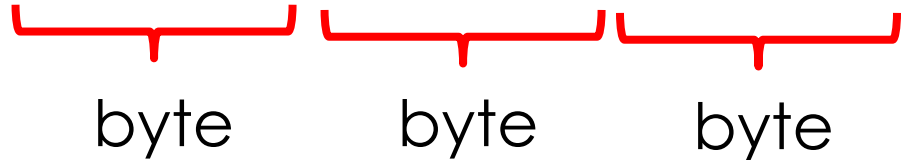
01100110 01100001 11000011 10100111 01100001

[Example](#)

Char Encoding: UTF-8

Characters `0x800-0xFFFF` in Unicode (16 bits):
mainly Chinese

Encode as follows:

1110xxxx 10xxxxxx 10xxxxxx

byte byte byte

Char Encoding: UTF-8

Decoding (mapping a sequence of bytes to characters):

- If the byte starts with `0xxxxxxx`
=> it's a "normal" character 00-0x7F
- If the byte starts with `110xxxxx`
=> it's an "extended" character 0x80 - 0x77F
one byte will follow
- If the byte starts with `1110xxxx`
=> it's a "Chinese" character, two bytes follow
- If the byte starts with `10xxxxxx`
=> it's a follower byte, you messed it up, dude!

01100110 01100001 11000011 10100111 01100001
f a ç a ... 24

Char Encoding: UTF-8

UTF-8 is a way to encode all Unicode characters into a variable sequence of 1-4 bytes

Advantages:

- common Western characters require only 1 byte (😊)
- backwards compatibility with ASCII
- stream readability (follower bytes cannot be confused with marker bytes)
- sorting compliance

In the following, we will assume that the document is a sequence of characters, without worrying about encoding

Language detection

How can we find out the language of a document?

Elvis Presley ist einer der
größten Rockstars aller Zeiten.

Different techniques:

- Watch for certain characters or scripts (umlauts, Chinese characters etc.)
But: These are not always specific, Italian similar to Spanish
- Use the meta-information associated with a Web page
But: This is usually not very reliable
- Use a dictionary
But: It is costly to maintain and scan a dictionary for thousands of languages

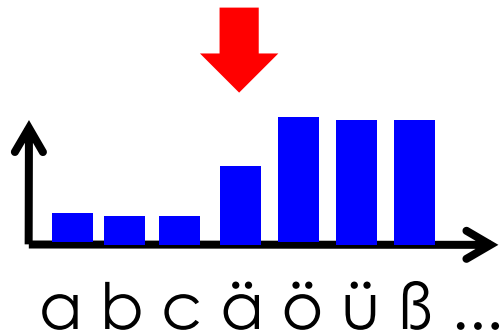
Language detection

Histogram technique for language detection:

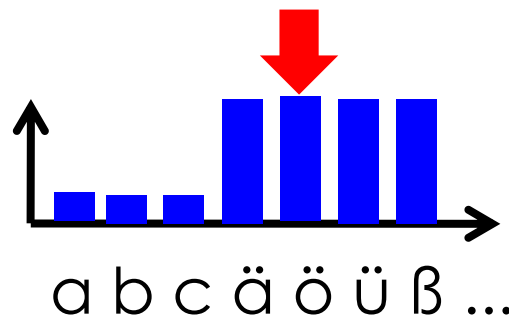
Count how often each character appears in the text.

Document:

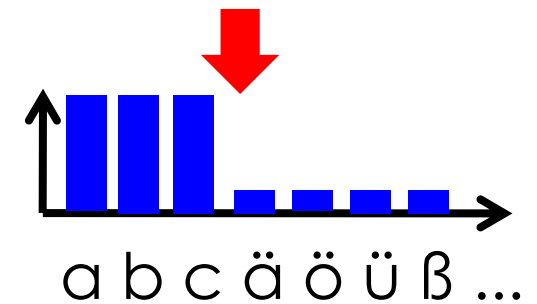
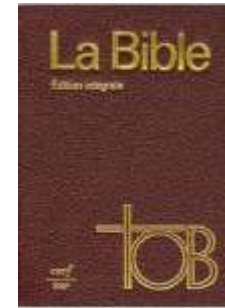
Elvis Presley ist
...



German corpus:



French corpus:



similar *not very similar*

Then compare to the counts on standard corpora.

Sources: Structured

Name	Number
D. Johnson	30714
J. Smith	20934
S. Shenker	20259
Y. Wang	19471
J. Lee	18969
A. Gupta	18884
R. Rivest	18038

**Information
Extraction**



Name	Citations
D. Johnson	30714
J. Smith	20937
...	...

File formats:

- TSV file (values separated by tabulator)
- CSV (values separated by comma)

Sources: Semi-Structured

```
<catalog>
```

```
  <cd>
```

```
    <title>
```

```
      Empire Burlesque
```

```
    </title>
```

```
    <artist>
```

```
      <firstName>
```

```
        Bob
```

```
      </firstName>
```

```
      <lastName>
```

```
        Dylan
```

```
      </lastName>
```

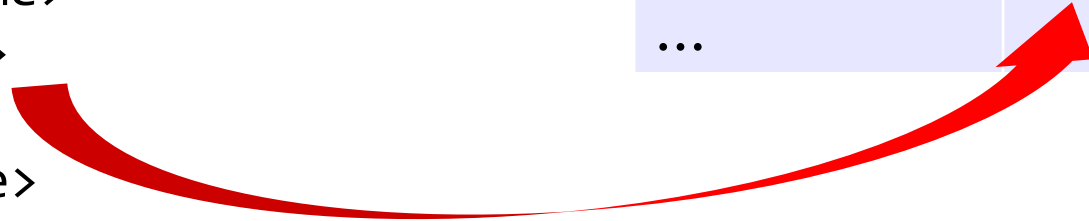
```
    </artist>
```

```
  </cd>
```

**Information
Extraction**



Title	Artist
Empire Burlesque	Bob Dylan
...	



File formats:

- XML file (Extensible Markup Language)
- YAML (Yaml Ain't a Markup Language)

Sources: Semi-Structured

Release Date (Year/M./Day)	Title	Num. Tracks
2008-11-24	Miles Away	7
2008-01-01	Hard Candy Cover	15
2008-01-01	Hard Candy Cover	12
2008-01-01	4 Minutes (4-Track Maxi-Single)	4
2008-01-01	4 Minutes (Single) Cover	1

<table>

<tr>

<td> 2008-11-24

<td> Miles away

<td> 7

<tr>

...

Information
Extraction



Title	Date
Miles away	2008-11-24
...	...

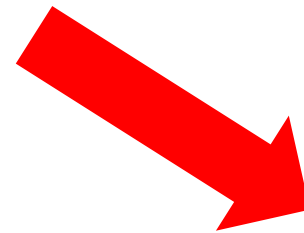
File formats:

- HTML file with table (Hypertext Markup Lang.)
- Wiki file with table (later in this class)

Sources: “Unstructured”

Founded in 1215 as a colony of Genoa, Monaco has been ruled by the House of Grimaldi since 1297, except when under French control from 1789 to 1814.

Designated as a protectorate of Sardinia from 1815 until 1860 by the Treaty of Vienna, Monaco's sovereignty ...







**Information
Extraction**

Event	Date
Foundation	1215
...	...

File formats:

- HTML file
- text file
- word processing document

Sources: Mixed

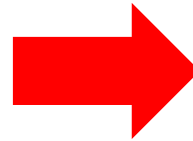
Barto, Andrew G.	(413) 545-2109	barto@cs.umass.edu	CS276
Professor. Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.			 
Berger, Emery D.	(413) 577-4211	emery@cs.umass.edu	CS344
Assistant Professor.			 

`<table>`

`<tr>`

`<td>` Professor.
Computational
Neuroscience,
...

Information
Extraction



Name	Title
Barte	Professor
...	...

...

Different IE approaches work with different types of sources

Source Selection Summary

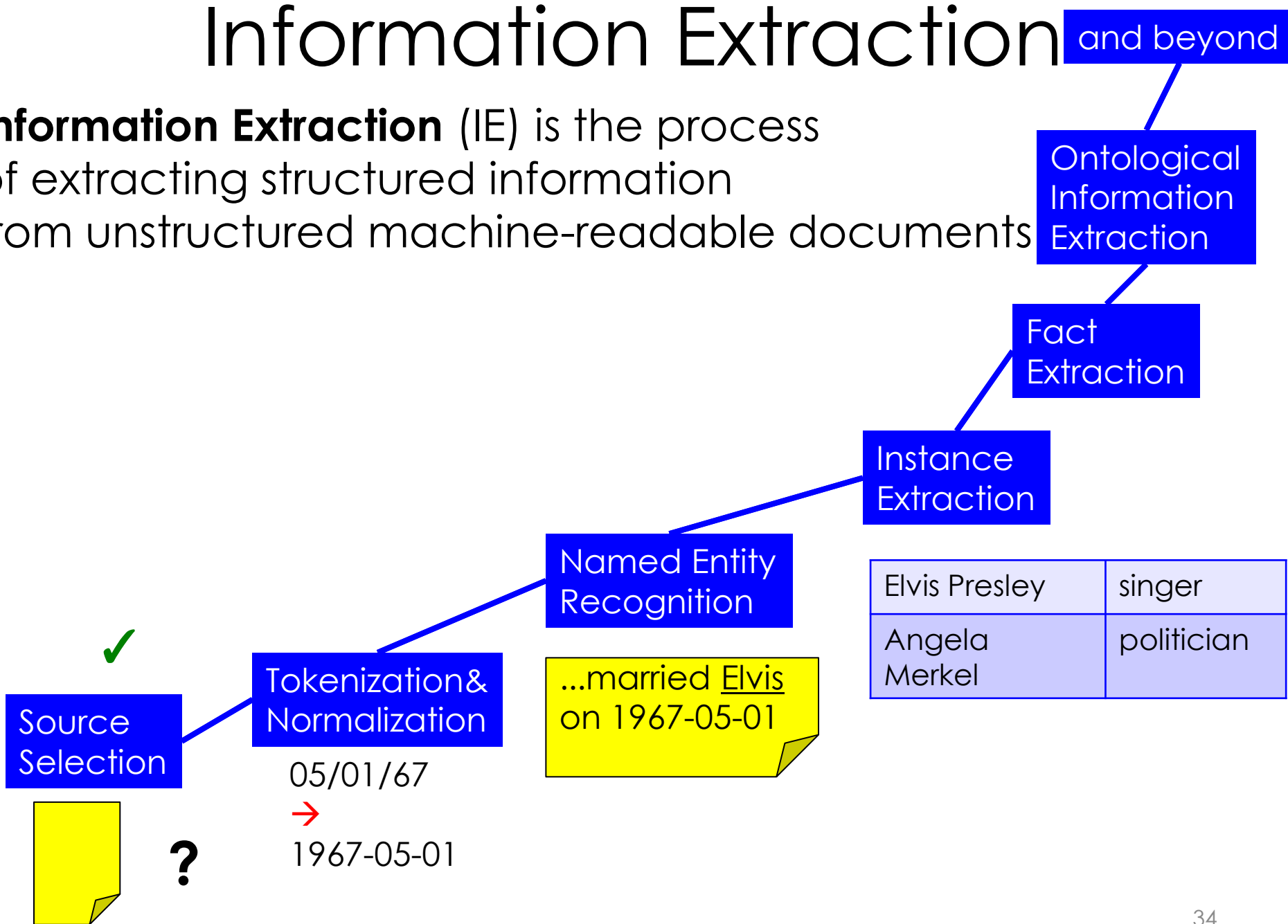
We can extract from the entire Web, or from certain Internet domains, thematic domains or files.

We have to deal with character encodings (ASCII, Code Pages, UTF-8,...) and detect the language

Our documents may be structured, semi-structured or unstructured.

Information Extraction and beyond

Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents



Tokenization

Tokenization is the process of splitting a text into tokens.

A **token** is

- a word
- a punctuation symbol
- a url
- a number
- a date
- or any other sequence of characters regarded as a unit

In|2011|,| President|Sarkozy|spoke|this|sample|sentence|. .

Tokenization Challenges

In |2011|, |President| |Sarkozy| spoke |this| sample |sentence|.

Challenges:

- In some languages (Chinese, Japanese), words are not separated by white spaces
- We have to deal consistently with URLs, acronyms, etc.
<http://example.com>, 2010-09-24, U.S.A.
- We have to deal consistently with compound words
[hostname](#), [host-name](#), [host name](#)

⇒ Solution depends on the language and the domain.

Naive solution: split by white spaces and punctuation 36

Normalization: Strings

Problem: We might extract strings that differ only slightly and mean the same thing.

Elvis Presley	singer
ELVIS PRESLEY	singer

Solution: **Normalize** strings, i.e., convert strings that mean the same to one common form:

- **Lowercasing**, i.e., converting all characters to lower case
- **Removing accents and umlauts**
résumé → resume, Universität → Universitaet
- **Normalizing abbreviations**
U.S.A. → USA, US → USA

Normalization: Literals

Problem: We might extract different **literals** (numbers, dates, etc.) that mean the same.

Elvis Presley	1935-01-08
Elvis Presley	08/01/35

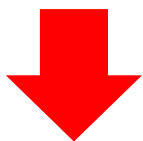
Solution: **Normalize** the literals, i.e., convert equivalent literals to one standard form:

08/01/35

01/08/35

8th Jan. 1935

January 8th, 1935



1935-01-08

1.67m

1.67 meters

167 cm

6 feet 5 inches

3 feet 2 toenails



1.67m

Normalization

Conceptually, normalization groups tokens into equivalence classes and chooses one representative for each class.

resume

résumé,
resume,
Resume

1935-01-08

8th Jan 1935,
01/08/1935

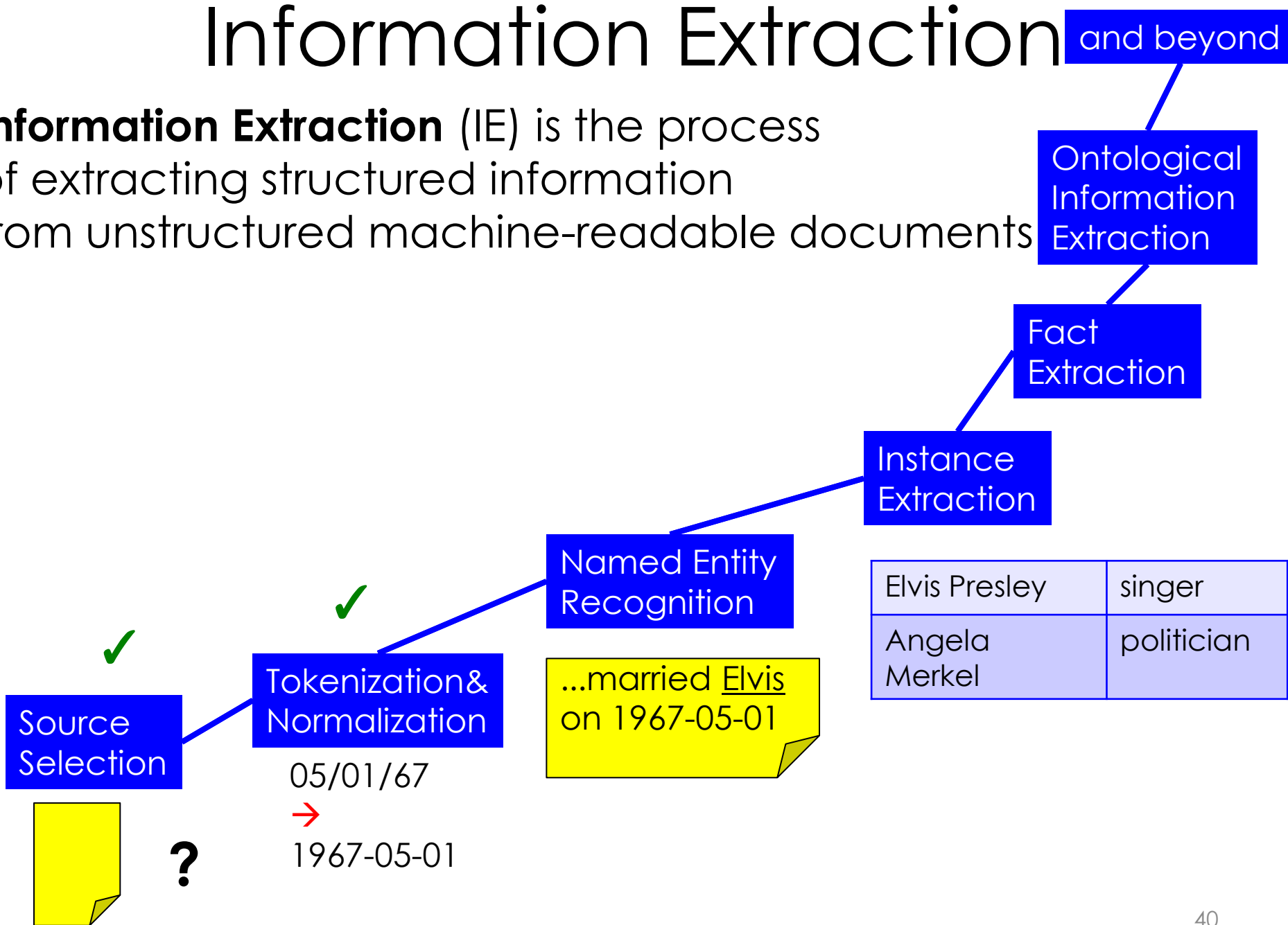
Take care not to normalize too aggressively:

bush



Information Extraction and beyond

Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents



Named Entity Recognition

Named Entity Recognition (NER) is the process of finding entities (people, cities, organizations, dates, ...) in a text.

Elvis Presley was born in 1935 in East Tupelo, Mississippi.



Closed Set Extraction

If we have an exhaustive set of the entities we want to extract, we can use **closed set extraction**:

Comparing every string in the text to every string in the set.

... in Tupelo, Mississippi, but ...

States of the USA
{ Texas, Mississippi,... }

... while Germany and France were opposed to a 3rd World War, ...

Countries of the World (?)
{ France, Germany, USA,... }

May not always be trivial...

... was a great fan of France Gall, whose songs...

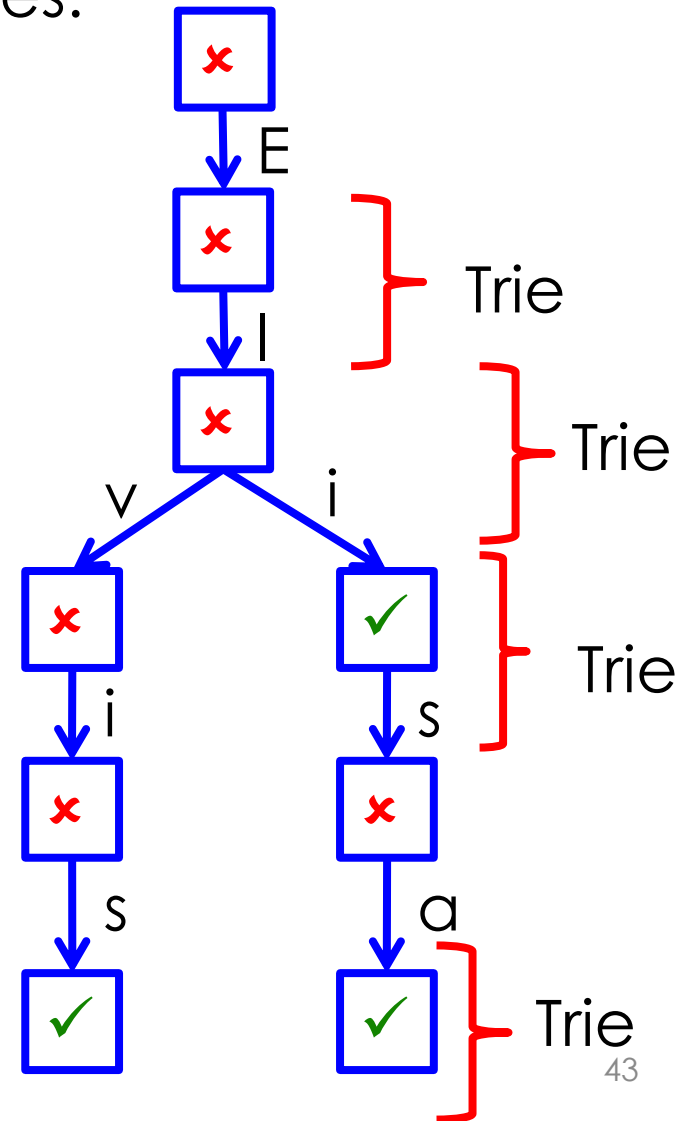
How can we do that efficiently?

Tries

A **trie** is pair of a boolean truth value, and a function from characters to tries.

Example: A trie containing “Elvis”, “Elisa” and “Eli”

A trie contains a string, if the string denotes a path from the root to a node marked with TRUE (✓)



Adding Values to Tries

Example: Adding "Elis"

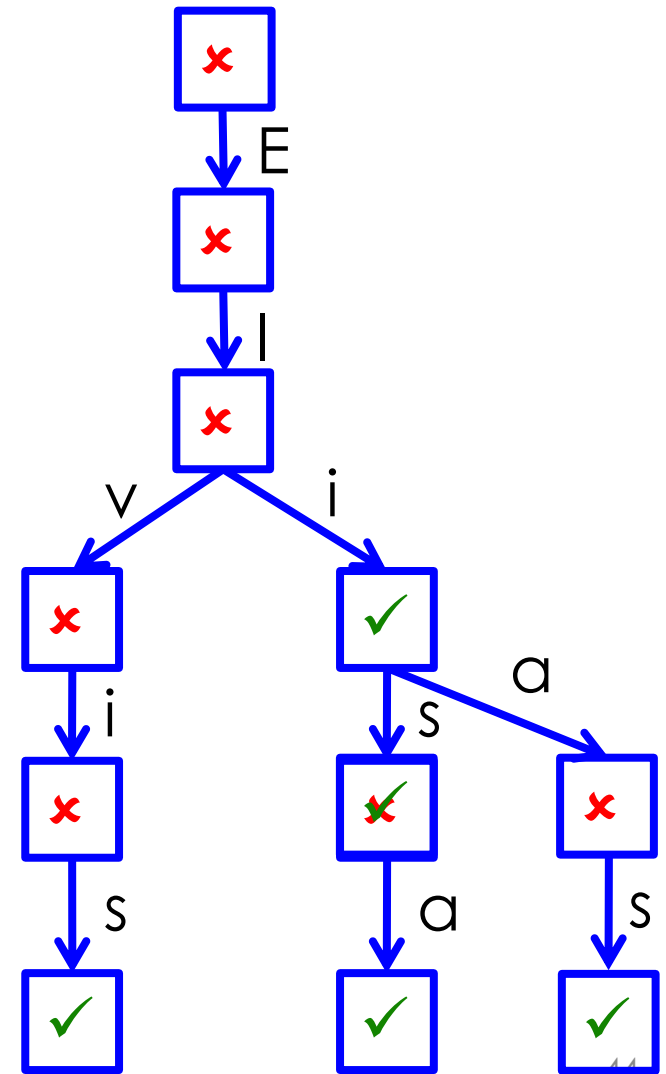
Switch the sub-trie to TRUE (✓)

Example: Adding "Elias"

Add the corresponding sub-trie

Start with an empty trie

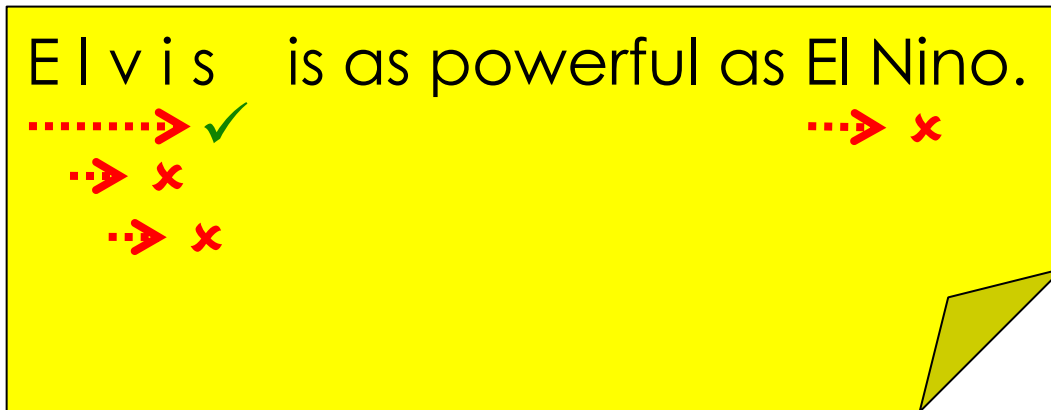
- *Add baby*
- *Add banana*



Parsing with Tries

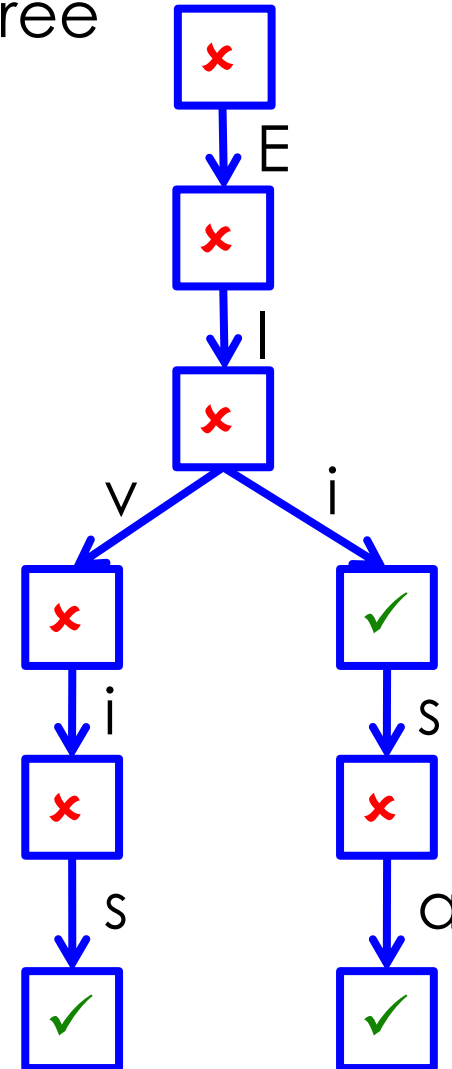
For every character in the text,

- advance as far as possible in the tree
- report match if you meet a node marked with TRUE (✓)



=> found Elvis

Time: $O(\text{textLength} * \text{longestEntity})$



NER: Patterns

If the entities follow a certain pattern, we can use **patterns**

... was born in 1935. His mother...
... started playing guitar in 1937, when...
... had his first concert in 1939, although...

Years
(4 digit numbers)

Office: 01 23 45 67 89
Mobile: 06 19 35 01 08
Home: 09 77 12 94 65

Phone numbers
(groups of digits)

Patterns

A **pattern** is a string that generalizes a set of strings.

sequences of the letter 'a'

a^+

a aa aaaaaa
 aaaa
aaaaaaa

'a', followed by 'b's

ab^+

abbbbbbb abbbb
 ab abbb
 abbb

digits

0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9

0 9 1 6 2
8 3 5 7 4

sequence of digits

$(0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9)^+$

987 6543 5321
 5643

=> Let's find a systematic way of expressing patterns

Regular Expressions

A **regular expression** (regex) over a set of symbols Σ is:

1. the empty string
2. or the string consisting of an element of Σ
(a single character)
3. or the string AB where A and B are regular expressions
(concatenation)
4. or a string of the form $(A | B)$,
where A and B are regular expressions **(alternation)**
5. or a string of the form $(A)^*$,
where A is a regular expression **(Kleene star)**

For example, with $\Sigma=\{a,b\}$, the following strings are regular expressions:

a

b

ab

aba

(a | b)

Regular Expression Matching

Matching

- a string **matches** a regex of a single character if the string consists of just that character

a

b

← regular expression

a

b

← matching string

- a string matches a regular expression of the form $(A)^*$ if it consists of zero or more parts that match A

(a)*

← regular expression

aaa aaaaa

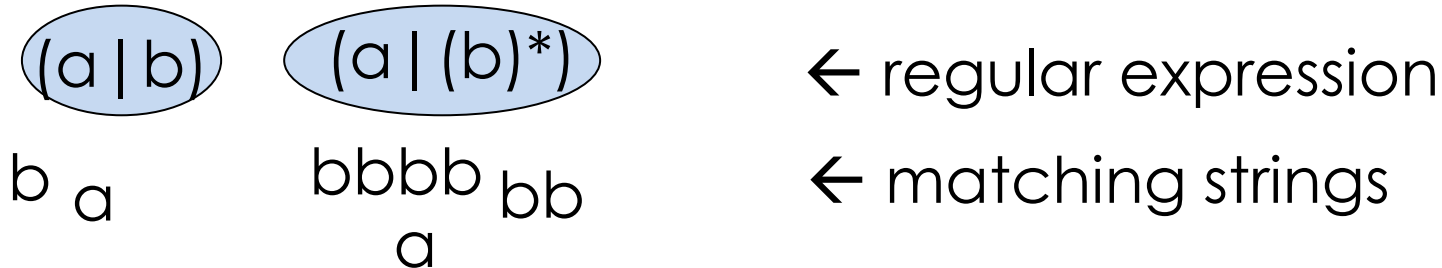
← matching strings

aaaaa

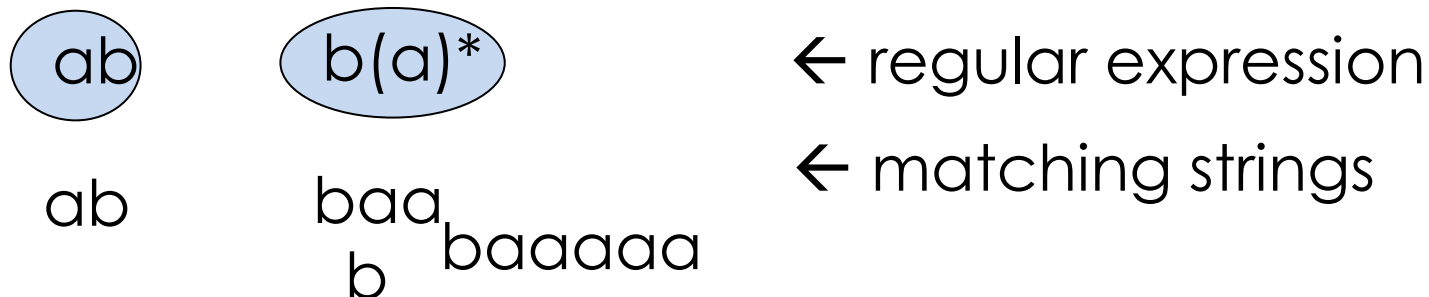
Regular Expression Matching

Matching

- a string matches a regex of the form $(A | B)$ if it matches either A or B



- a string matches a regular expression of the form AB if it consists of two parts, where the first part matches A and the second part matches B



Additional Regexes

Given an ordered set of symbols Σ , we define

- $[x-y]$ for two symbols x and y , $x < y$, to be the alternation $x | \dots | y$ (meaning: any of the symbols in the range)
 $[0-9] = 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9$
- A^+ for a regex A to be $A(A)^*$ (meaning: one or more A 's)
 $[0-9]^+ = [0-9][0-9]^*$
- $A\{x,y\}$ for a regex A and integers $x < y$ to be $A \dots A | A \dots A | A \dots A | \dots | A \dots A$ (meaning: x to y A 's)
 $f\{4,6\} = ffff | fffff | fffffff$
- $A?$ for a regex A to be $(| A)$ (meaning: an optional A)
 $ab? = a(| b)$
- $.$ to be an arbitrary symbol from Σ

Regular Expression Exercise

A B	Either A or B	(Use a backslash for the character itself, e.g., \+ for a plus)
A*	Zero+ occurrences of A	
A+	One+ occurrences of A	
A{x,y}	x to y occurrences of A	
A?	an optional A	
[a-z]	One of the characters in the range	
.	An arbitrary symbol	

A digit

A digit or a letter

A sequence of 8 digits

5 pairs of digits, separated by space

HTML tags

Person names:

Dr. Elvis Presley

Prof. Dr. Elvis Presley

[Example](#)

Names & Groups in Regexes

When using regular expressions in a program, it is common to **name** them:

```
String digits="[0-9]+";  
String separator="( |-)";  
String pattern=digits+separator+digits;
```

Parts of a regular expression can be singled out by bracketed **groups**:

```
String input="The cat caught the mouse."
```

```
String pattern="The ([a-z]+) caught the ([a-z]+)\\.\\."
```

 first group: "cat"
second group: "mouse"

[Try this](#)

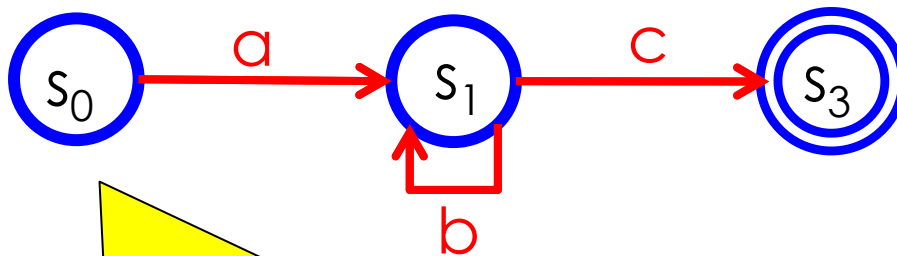
Finite State Machines

A regex can be matched efficiently by a Finite State Machine (Finite State Automaton, FSA, FSM)

A **FSM** is a quintuple of

- A set Σ of symbols (the **alphabet**)
- A set S of **states**
- An **initial state**, $s_0 \in S$
- A **state transition function** $\delta: S \times \Sigma \rightarrow S$
- A **set of accepting states** $F \subset S$

Regex: ab^*c



Accepting states usually depicted with double ring.

Implicitly: All unmentioned inputs go to some artificial failure state

Finite State Machines

A FSM **accepts** an input string, if there exists a sequence of states, such that

- it starts with the start state
- it ends with an accepting state
- the i -th state, s_i , is followed by the state $\delta(s_i, \text{input.charAt}(i))$

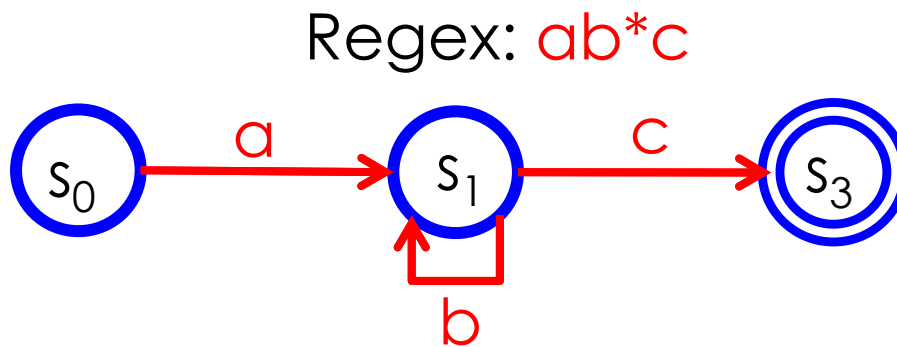
Sample inputs:

abbbc

ac

aabbbc

elvis

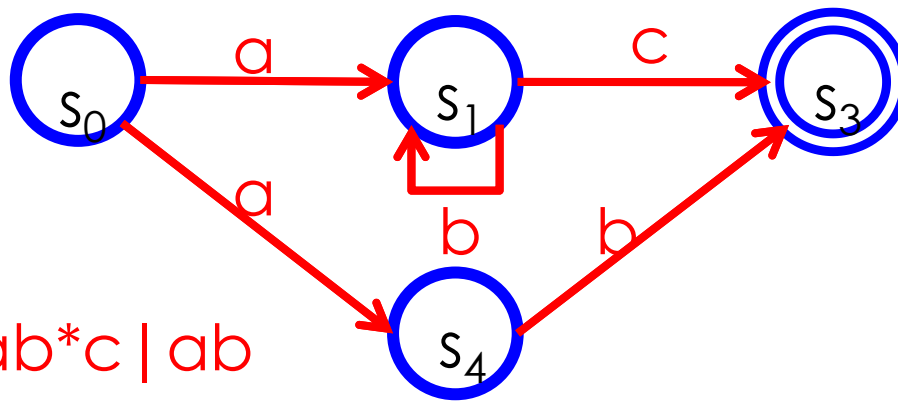


Non-Deterministic FSM

A **non-deterministic FSM** has a transition function that maps to a **set of states**.

A FSM **accepts** an input string, if there exists a sequence of states, such that

- it starts with the start state
- it ends with an accepting state
- the i -th state, s_i , is followed by a state in the set $\delta(s_i, \text{input.charAt}(i))$



Regex: $ab^*c \mid ab$

Sample inputs:

abbbc

ab

abc

elvis

Regular Expressions Summary

Regular expressions

- can express a wide range of patterns
- can be matched efficiently
- are employed in a wide variety of applications (e.g., in text editors, NER systems, normalization, UNIX grep tool etc.)

Input:

- Manual design of the regex


Condition:

- Entities follow a pattern

Sliding Windows

Alright, what if we do not want to specify regexes by hand? Use sliding windows:

Information Extraction: Tuesday 10:00 am, Rm 407b



For each position, ask: Is the current window a named entity?

Window size = 1

Sliding Windows

Alright, what if we do not want to specify regexes by hand? Use sliding windows:

Information Extraction: Tuesday 10:00 am, Rm 407b



For each position, ask: Is the current window a named entity?

Window size = 2

Features

Information Extraction: Tuesday 10:00 am, Rm 407b

Prefix
window

Content
window

Postfix
window

Choose certain **features** (properties) of windows that could be important:

- window contains colon, comma, or digits
- window contains week day, or certain other words
- window starts with lowercase letter
- window contains only lowercase letters
- ...

Feature Vectors

Information Extraction: Tuesday 10:00 am, Rm 407b

Prefix
window

Content
window

Postfix
window

Prefix colon
Prefix comma
...
Content colon
Content comma
...
Postfix colon
Postfix comma

$$\begin{bmatrix} 1 \\ 0 \\ \dots \\ 1 \\ 0 \\ \dots \\ 0 \\ 1 \end{bmatrix}$$

The **feature vector** represents the presence or absence of features of one content window (and its prefix window and postfix window)

Features

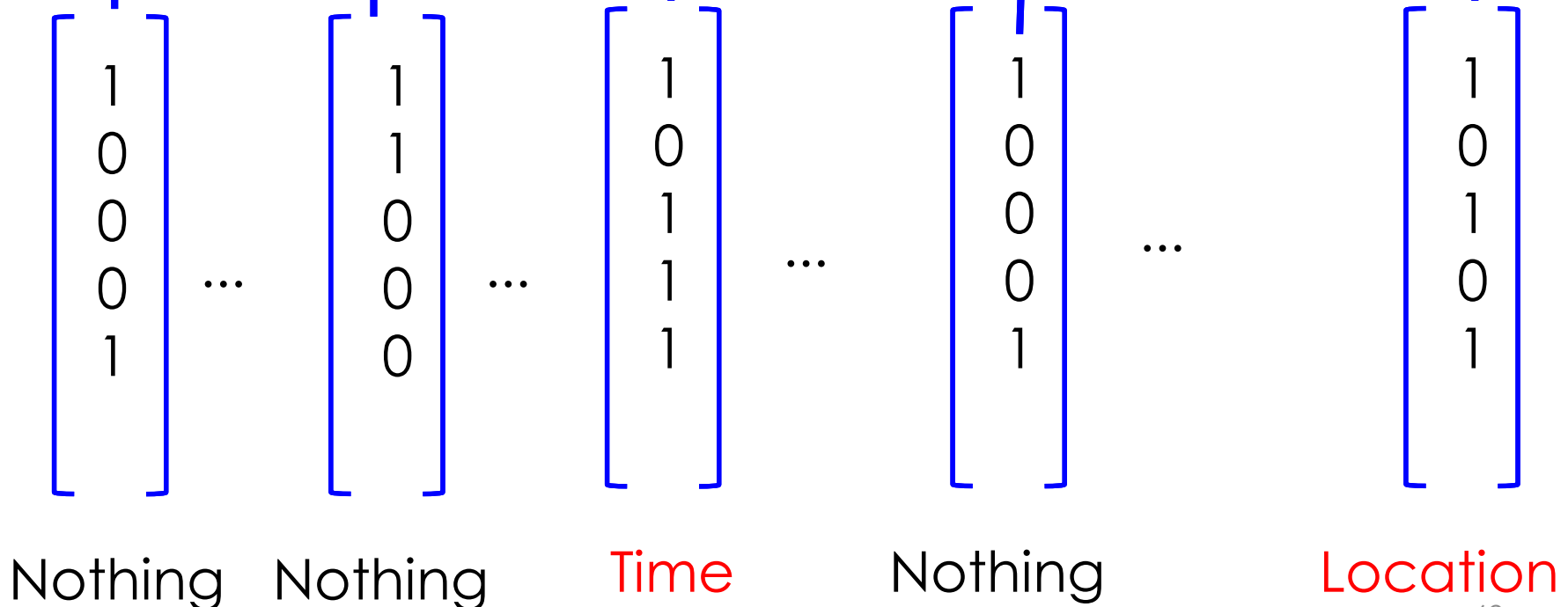
Feature Vector

Sliding Windows Corpus

Now, we need a **corpus** (set of documents) in which the entities of interest have been manually labeled.

NLP class: Wednesday, 7:30am and Thursday all day, rm 667

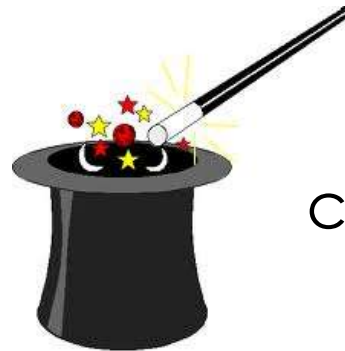
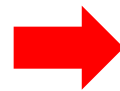
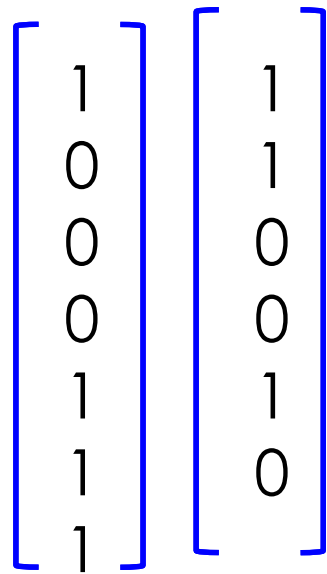
From this corpus compute the feature vectors with labels:



Machine Learning

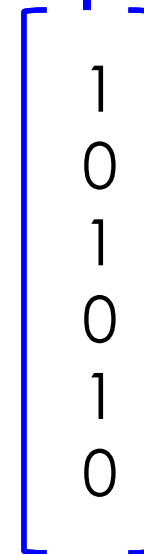
Information Extraction: Tuesday 10:00 am, Rm 407b

Use the labeled feature vectors as training data for Machine Learning



Machine Learning

classify



Time

Nothing Location

Sliding Windows Exercise

What features would you use to recognize person names?

Elvis Presley married Ms. Priscilla at the Aladin Hotel.

UpperCase
hasDigit
...

[1
0
0
0
1
1]

[1
0
1
1
1
1]

...

[1
0
1
0
1
0]

Sliding Windows Summary

The Sliding Windows Technique can be used for Named Entity Recognition for nearly arbitrary entities

Input:

- a labeled corpus
- a set of features

The features can be arbitrarily complex and the result depends a lot on this choice



Condition:

- The entities share some syntactic similarities

The technique can be refined by using better features, taking into account more of the context (not just prefix and postfix) and using advanced Machine Learning.

NER Summary

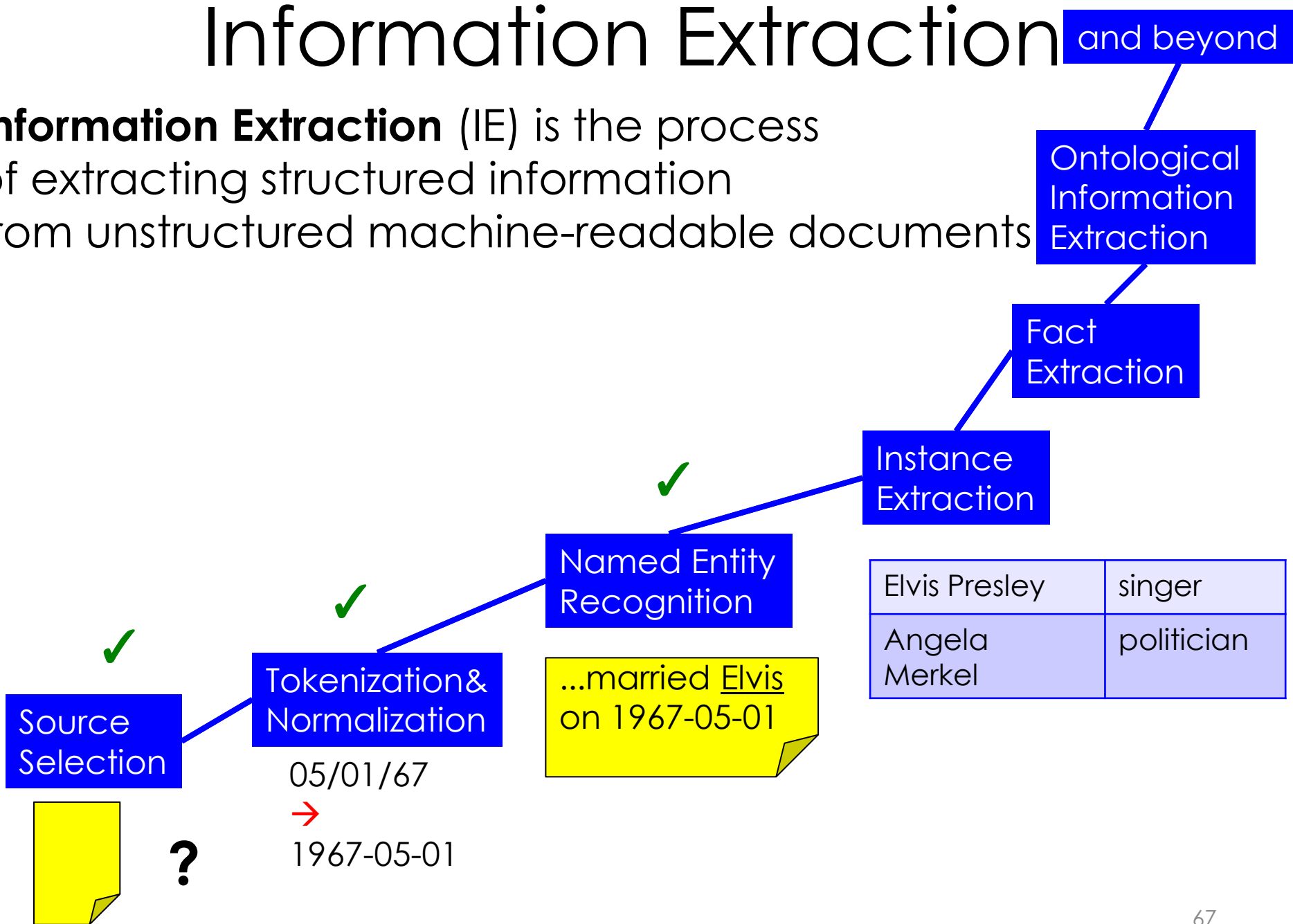
Named Entity Recognition (NER) is the process of finding entities (people, cities, organizations, ...) in a text.

We have seen different techniques

- Closed-set extraction (if the set of entities is known)
Can be done efficiently with a trie
- Extraction with Regular Expressions
(if the entities follow a pattern)
Can be done efficiently with Finite State Automata
- Extraction with sliding windows / Machine Learning
(if the entities share some syntactic features)

Information Extraction and beyond

Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents



Instance Extraction

Instance Extraction is the process of extracting entities with their **class** (i.e., concept, set of similar entities)

Elvis was a great artist, but while all of Elvis' colleagues loved the song "Oh yeah, honey", Elvis did not perform that song at his concert in Hintertuepflingen.

Entity	Class
Elvis	artist
Oh yeah, honey	song
Hintertuepflingen	location

...some of the class assignment might already be done by the Named Entity Recognition.

Hearst Patterns

Instance Extraction is the process of extracting entities with their **class** (i.e., concept, set of similar entities)

Elvis was a great artist,
but while all of Elvis' colleagues loved the song "Oh yeah, honey", Elvis did not perform that song at his concert in Hintertuepflingen.

Entity	Class
Elvis	artist

Idea (by Hearst):

Sentences express class membership in very predictable patterns. Use these patterns for instance extraction.

Hearst patterns:

- X was a great Y

Instance Extraction: Hearst Patterns

Elvis was a great artist

Many scientists, including Einstein, started to believe that matter and energy could be equated.

He adored Madonna, Celine Dion and other singers, but never got an autograph from any of them.

Many US citizens have never heard of countries such as Guinea, Belize or France.

Idea (by Hearst):

Sentences express class membership in very predictable patterns. Use these patterns for instance extraction.

Hearst patterns:

- X was a great Y
- Ys, such as X1, X2, ...
- X1, X2, ... and other Y
- many Ys, including X

Hearst Patterns on Google

Hearst Patterns on Google

[Try it out](#)

"cities such as"

About 5,300,000 results (0.43 seconds)

▶ [News for "cities such as"](#)

[Unknown Cities Are Getting Richer](#) ☆ - 23 hours ago
Cities such as Aurangabad, Curitiba in Brazil, Xiaochang in China, and lumped together, BCG found, with the mostly poor, ...
[BusinessWeek](#) - 3 related articles

[Cities That Could Steal Your Job: New Outsourcing Hot Spots](#)
From overlooked American cities such as Boise, Idaho and Winnipeg to cities like Cluj-Napoca, Romania, or the Philippines' Iloilo City, ...
[images.businessweek.com/ss/09/05/0504_outsourcing.../1.htm](#) - Cache

Idea (by Hearst):

Sentences express class membership in very predictable patterns. Use these patterns for instance extraction.

Wildcards on Google

"many ", including **"

About 1,670,000,000 results (0.19 seconds)

▶ [Putco 401127 Chrome Trim Mirror Covers. Fits many Fords in](#)
Fits many Fords including the F-150, F-250 Super Duty, and many more from Brand: Putco, Mfr Part#: 401127. Lowest Price \$72.89 ...
[www.streetperformance.com/part/.../869788-401127.html](#) - Cached - Similar

Hearst patterns:

- X was a great Y
- Ys, such as X1, X2, ...
- X1, X2, ... and other Y
- many Ys, including X

Hearst Patterns Summary

Hearst Patterns can extract instances from natural language documents

Input:

- Hearst patterns for the language (easily available for English)

Condition:

- Text documents contain class + entity explicitly in defining phrases

Idea (by Hearst):

Sentences express class membership in very predictable patterns. Use these patterns for instance extraction.

Hearst patterns:

- X was a great Y
- Ys, such as X1, X2, ...
- X1, X2, ... and other Y
- many Ys, including X

Instance Classification

Suppose we have $\text{scientists}=\{\text{Einstein, Bohr}\}$
 $\text{musician}=\{\text{Elvis, Madonna}\}$

When Einstein
discovered the U86
plutonium
hypercarbonate...

In 1940, Bohr
discovered the
 $\text{CO}_2\text{H}_3\text{X}$.

Elvis played the
guitar, the piano,
the flute, the
harpsichord,...

Rengstorff made
multiple important
discoveries, among
others the theory of
recursive
subjunction.

Stemmed context of the entity without stop words:

{discover
U86
plutonium}

{1940,
discover,
 $\text{CO}_2\text{H}_3\text{X}$ }

{play,
guitar,
piano}

{make,
important,
discover}

Scientist

Scientist

Musician

What is
Rengstorff?

Instance Classification

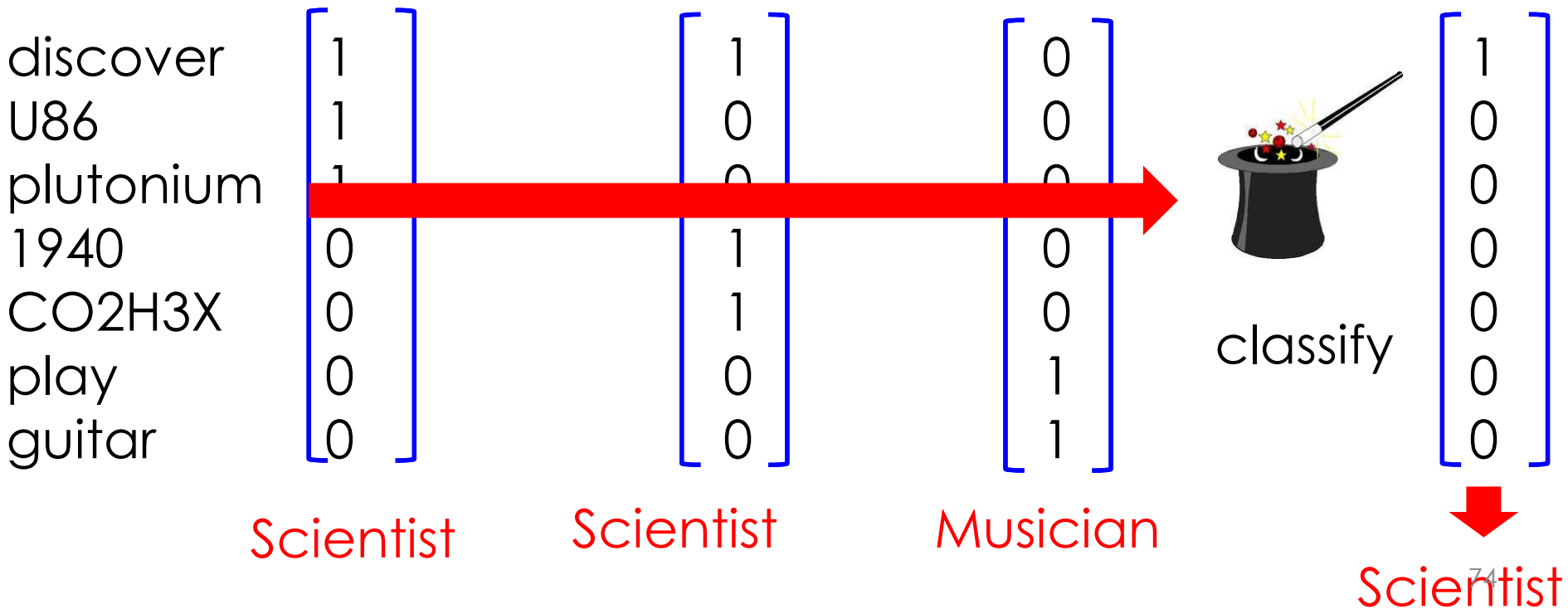
Suppose we have $\text{scientists}=\{\text{Einstein, Bohr}\}$
 $\text{musician}=\{\text{Elvis, Madonna}\}$

When Einstein discovered the U86 plutonium hypercarbonate...

In 1940, Bohr discovered the $\text{CO}_2\text{H}_3\text{X}$.

Elvis played the guitar, the piano, the flute, the harpsichord,...

Rengstorff made multiple important discoveries, among others the theory of recursive subjunction.



Instance Classification

Instance Classification can extract instances from text corpora without defining phrases.

Condition:

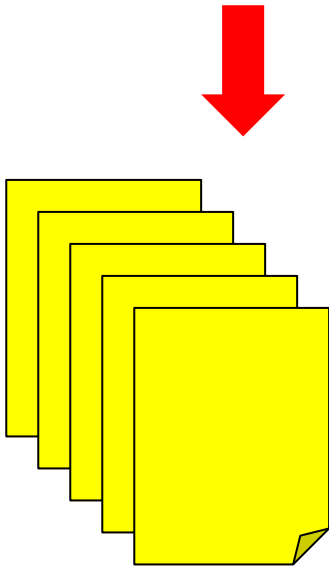
- The texts have to be homogenous

Input:

- Known classes
- seed sets

Instance Extraction Iteration

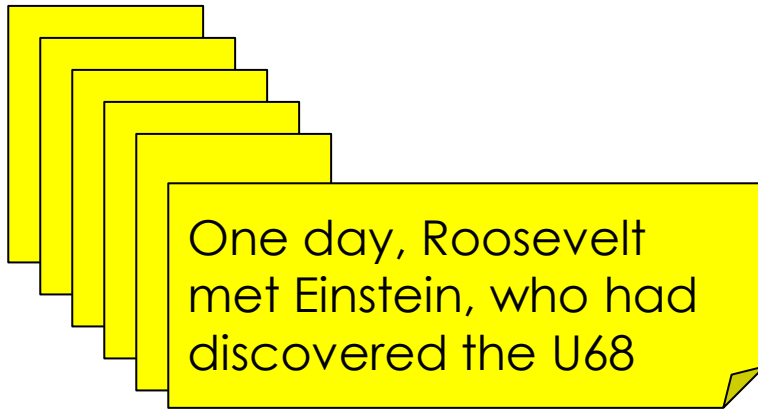
Seed set: {Einstein, Bohr}



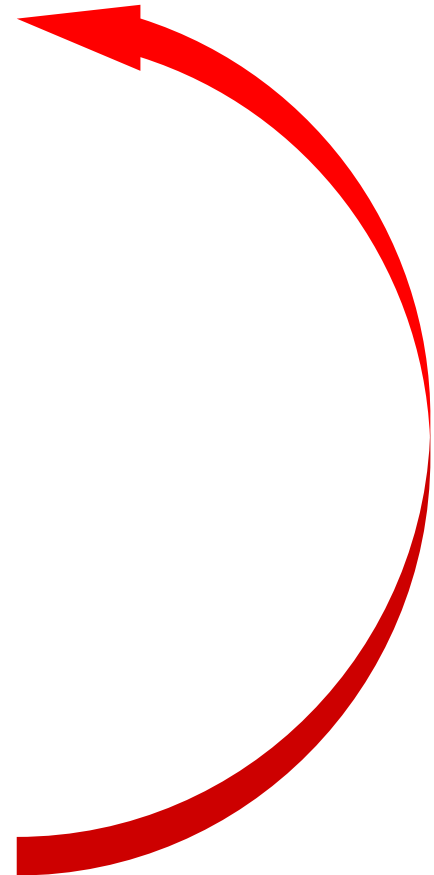
Result set: {Einstein, Bohr, Planck}

Instance Extraction Iteration

Seed set: {Einstein, Bohr, Planck}

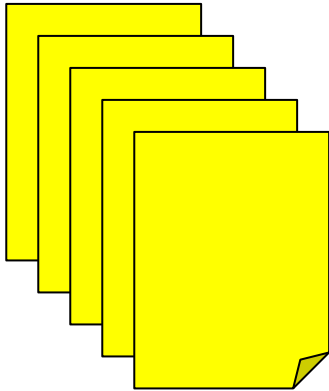


Result set: {Einstein, Bohr, Planck, Roosevelt}



Instance Extraction Iteration

Seed set: {Einstein, Bohr, Planck, Roosevelt}



Semantic Drift is a problem that can appear in any system that reuses its output

Result set: {Einstein, Bohr, Planck, Roosevelt, Kennedy, Bush, Obama, Clinton}

Set Expansion

Seed set: {Russia, USA, Australia}



▶ **LARGEST COUNTRIES** (by land mass)

[locator map here](#)

Russia 17,075,400 sq km, (6,592,846 sq miles)

Canada 9,330,970 sq km, (3,602,707 sq miles)

China 9,326,410 sq km, (3,600,947 sq miles)

USA 9,166,600 sq km, (3,539,242 sq miles)

Brazil 8,456,510 sq km, (3,265,075 sq miles)

Australia 7,617,930 sq km, (2,941,283 sq miles)

India 2,973,190 sq km, (1,147,949 sq miles)

Argentina 2,736,690 sq km, (1,056,636 sq miles)

Kazakhstan 2,717,300 sq km, (1,049,150 sq miles)

Sudan 2,376,000 sq km, (917,374 sq miles)



Result set: {Russia, Canada, China, USA, Brazil, Australia, India, Argentina, Kazakhstan, Sudan}

Set Expansion

Most corrupt countries

174	 Uzbekistan	1.7	1.8	1.7
175	 Chad	1.6	1.6	1.8
176	 Iraq	1.5	1.3	1.5
176	 Sudan	1.5	1.6	1.8
178	 Myanmar	1.4	1.3	1.4
179	 Afghanistan	1.3	1.5	1.8
180	 Somalia	1.1	1.0	1.4

Result set: {Russia, Canada, China, USA, Brazil, Australia, India, Argentina, Kazakhstan, Sudan}

Set Expansion

Seed set: {Russia, Canada, ...}



Most corrupt countries

174	 Uzbekistan	1.7	1.8	1.7
175	 Chad	1.6	1.6	1.8
176	 Iraq	1.5	1.3	1.5
176	 Sudan	1.5	1.6	1.8
178	 Myanmar	1.4	1.3	1.4
179	 Afghanistan	1.3	1.5	1.8
180	 Somalia	1.1	1.0	1.4



Result set: {Uzbekistan, Chad, Iraq, ...}

Try, e.g., Google sets:

<http://labs.google.com/sets>

- Uzbekistan
- Chad
- Iraq
- Sudan
- Myanmar

Predicted Items

[chad](#)

[sudan](#)

[uzbekistan](#)

[myanmar](#)

[iraq](#)

[afghanistan](#)

Set Expansion

Set Expansion can extract instances from tables or lists.

174	 Uzbekistan	1.7	1.8	1.7
175	 Chad	1.6	1.6	1.8
176	 Iraq	1.5	1.3	1.5
176	 Sudan	1.5	1.6	1.8
178	 Myanmar	1.4	1.3	1.4
179	 Afghanistan	1.3	1.5	1.8
180	 Somalia	1.1	1.0	1.4

Input:

- seed pairs

Condition:

- a corpus full of tables

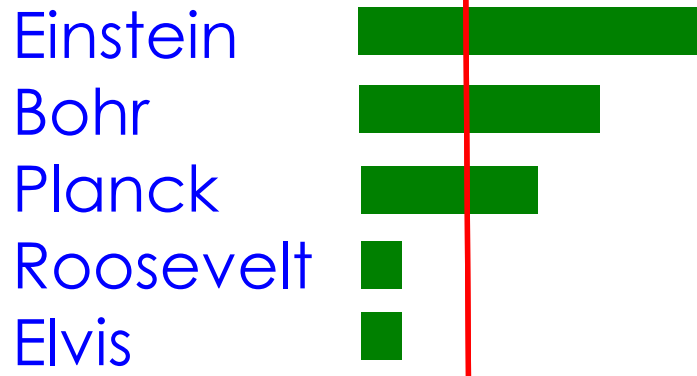
Cleaning

IE nearly always produces **noise** (minor false outputs)

Solutions:

- **Thresholding**

(Cutting away instances that were extracted few times)



- **Heuristics**

(rules without scientific foundations that work well)

Accept an output only if it appears on different pages,
merge entities that look similar (Einstein, EINSTEIN), ...

Evaluation

In science, every system, algorithm or theory should be **evaluated**, i.e. its output should be compared to the **gold standard** (i.e. the ideal output).

Algorithm output:

$O = \{\text{Einstein}, \text{Bohr}, \text{Planck}, \text{Clinton}, \text{Obama}\}$

Gold standard:

$G = \{\text{Einstein}, \text{Bohr}, \text{Planck}, \text{Heisenberg}\}$

Precision:

What proportion of the output is correct?

$$\frac{|O \cap G|}{|O|}$$

Recall:

What proportion of the gold standard did we get?

$$\frac{|O \cap G|}{|G|}$$

Explorative Algorithms

Explorative algorithms extract everything they find.
(very low threshold)

Algorithm output:

$O = \{\text{Einstein, Bohr, Planck, Clinton, Obama, Elvis, ...}\}$

Gold standard:

$G = \{\text{Einstein, Bohr, Planck, Heisenberg}\}$

Precision:

What proportion of the output is correct?

BAD

Recall:

What proportion of the gold standard did we get?

GREAT

Conservative Algorithms

Conservative algorithms extract only things about which they are very certain

(very high threshold)

Algorithm output:

$O = \{\text{Einstein}\}$

Gold standard:

$G = \{\text{Einstein, Bohr, Planck, Heisenberg}\}$

Precision:

What proportion of the output is correct?

GREAT

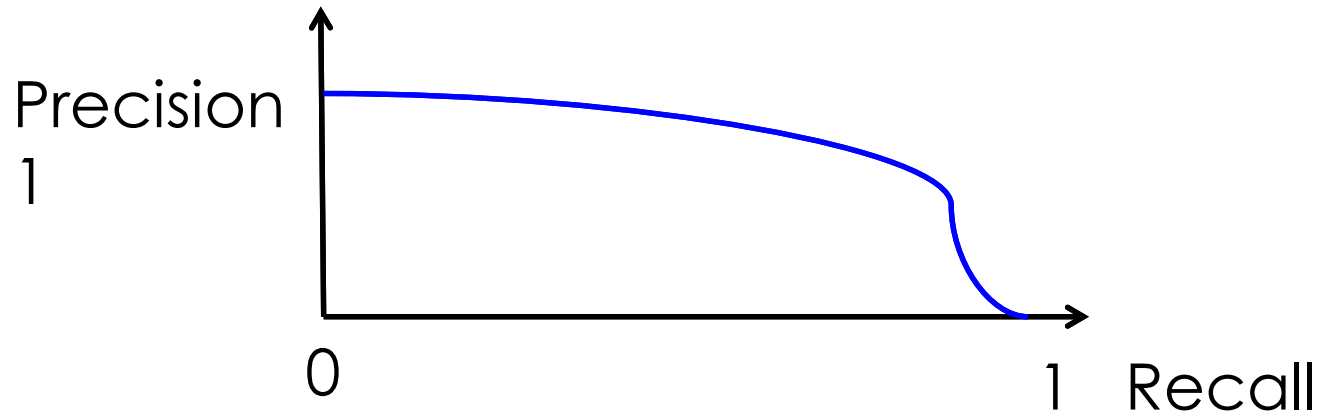
Recall:

What proportion of the gold standard did we get?

BAD

F1 - Measure

You can't get it all...



The F1-measure combines precision and recall as the harmonic mean:

$$F1 = 2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$$

Precision & Recall Exercise

What is the algorithm output, the gold standard, the precision and the recall in the following cases?

1. Nostradamus predicts a world war for every century from the 15th to the 20th incl.
2. In an exercise, you do 3 out of 21 questions right.
3. On Elvis Radio TM, 90% of the songs are by Elvis. An algorithm learns to detect Elvis songs. Out of 100 songs on Elvis Radio, the algorithm says that 20 are by Elvis (and 5 were not).

output={e1,...,e15, x1,...,x5}

gold={e1,...,e90}

prec=15/20=75 %, rec=15/90=16%

4. How can you improve the algorithm?

Instance Extraction

Instance Extraction is the process of extracting entities with their **class** (i.e., concept, set of similar entities)

Approaches:

- Hearst Patterns
(work on natural language corpora)
- Classification
(if the entities appear in homogeneous contexts)
- Set Expansion
(for tables and lists)
- ...many others...

On top of that:

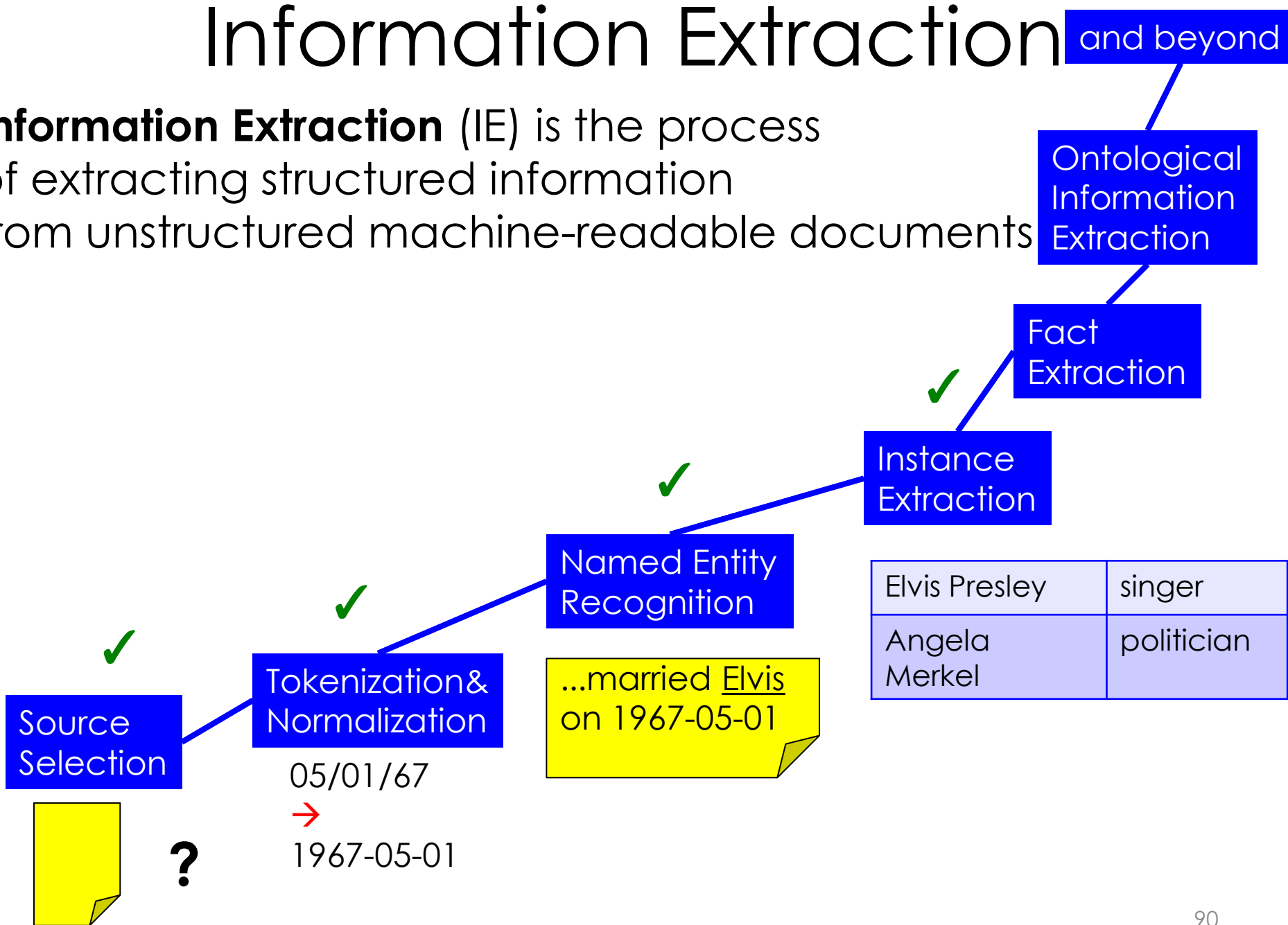
- Iteration
- Cleaning

And finally:

- Evaluation

Information Extraction and beyond

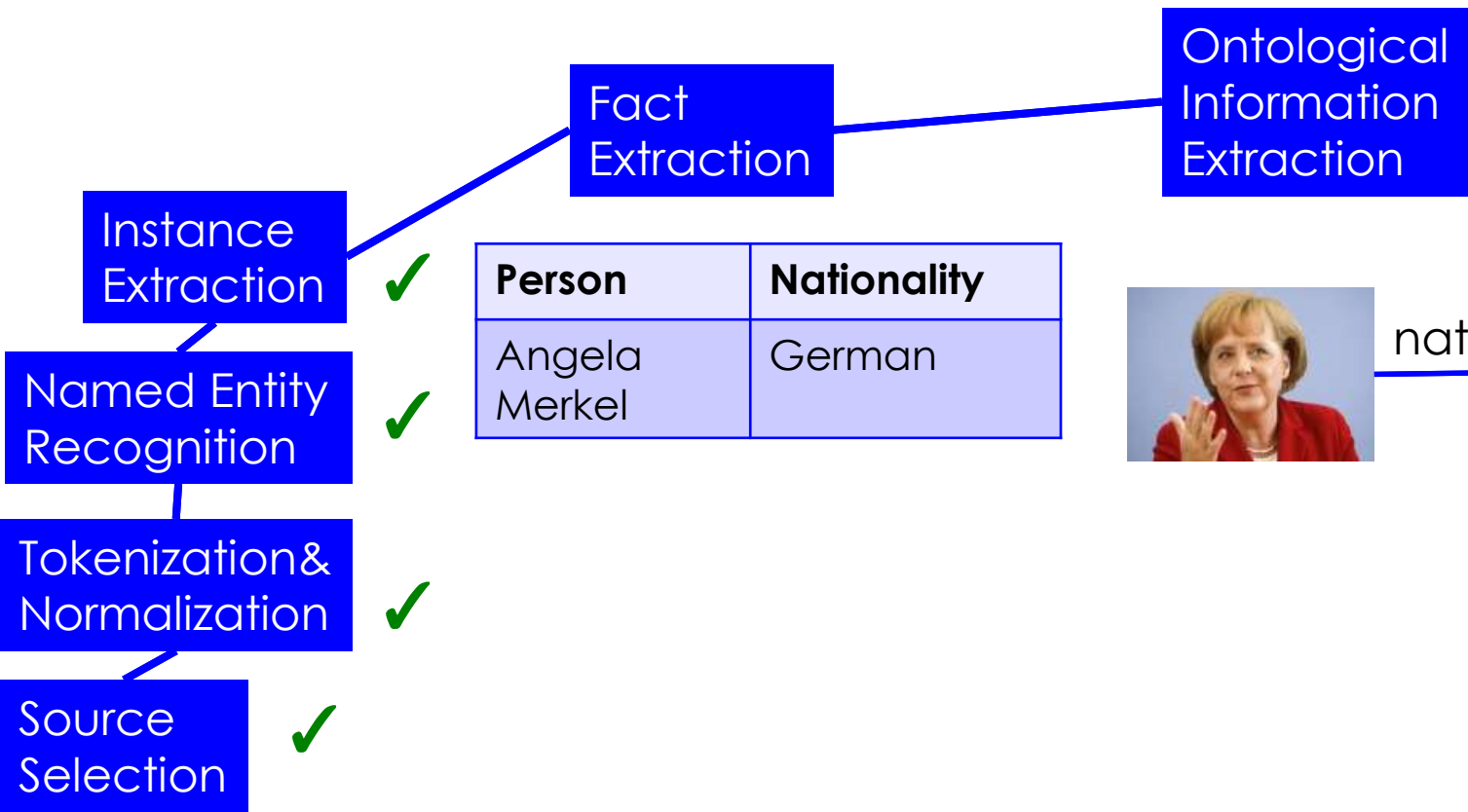
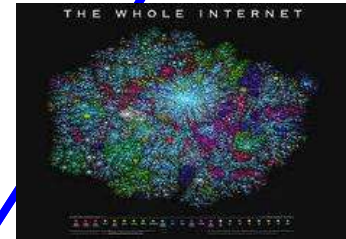
Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents



Information Extraction

Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents

and beyond



Fact Extraction

Fact Extraction is the process of extracting pairs (triples,...) of entities together with the relationship of the entities.



Costello Sings Lowe/Nick Sings Elvis (late show)

THE BAND: Paul Revelli, Ruth Davies, Bill Kirchen, Bob Andrews, Derek Huston, Austin ...

10/1/2010 Friday 11:00p

Great American Music Hall, San Francisco CA

Featuring: [Elvis Costello](#), [Nick Lowe](#)



BUY



Event	Time	Location
Costello sings...	2010-10-01, 23:00	Great American...

Wrapper Induction

Observation: On Web pages of a certain domain, the information is often in the same spot.

On est là pour vous aider

IMDb The Internet Movie Database

Search All Go

Movies TV News Videos Community IMDb

IMDb20 Celebrate Our 20th Anniversary with a New Star Every Day!

Elvis: Aloha from Hawaii (TV 1973) [More at IMDbPro](#) »

87 min - [Music](#)

★★★★★ 8.2/10
Users: (569 votes) [27 reviews](#) | Critics: [2 reviews](#)

A 1973 concert by Elvis Presley taped at the Convention Center in Honolulu, Hawaii. This was the first program to ever be beamed around the world by satellite.

Directors: [Marty Pasetta](#), [Gary Hovey](#), and [1 more credit](#) »

Release Date: **14 January 1973 (USA)**

[Full cast and crew](#) | [14 photos](#) »

IMDb The Internet Movie Database

Search All Go

Movies TV News Videos Community IMDb

IMDb20 Celebrate Our 20th Anniversary with a New Star Every Day!

The Life of Brian (2002) [More at IMDbPro](#) »

Brian: On Palmes, Tikourounglilhe (original title)

52 min -

★★★★★ 4.9/10
Users: [145 votes](#) with [1 review](#)

Directors: [Héctor Belloc](#), [Néstor Akaziri-Soumeia](#)

[Own the rights? Add a poster](#)

[Full cast and crew](#) »

IMDb The Internet Movie Database

Search All Go

Movies TV News Videos Community IMDb

IMDb20 Celebrate Our 20th Anniversary with a New Star Every Day!

Titanic (1997) [More at IMDbPro](#) »

(PG-13) 194 min - [Drama](#) | [History](#) | [Romance](#)

★★★★★ 7.4/10
Users: [1261,832 votes](#) | [2,281 reviews](#) - Other: [188 reviews](#)

Fictional romantic tale of a rich girl and poor boy who meet on the ill-fated voyage of the 'unsinkable' ship.

Director: [James Cameron](#)
Writer: [James Cameron](#)
Release Date: 7 January 1998 (France)

[Watch Trailer](#) »

[Full cast and crew](#) | [148 photos](#) | [9 videos](#) »

Wrapper Induction

Observation: On Web pages of a certain domain, the information is often in the same spot.

Idea: Describe this spot in a general manner.

A description of one spot on a page is called a **wrapper**.



```
<html>
<body>
<div>
  ...
  <div>
  ...
  <div>
  ...
  <b>Elvis: Aloha from Hawaii</b> (TV...
```

A wrapper can be similar to an XPath expression:

`html → div[1] → div[2] → b[1]`

It can also be a search text or regex

`>.*(TV`

Wrapper Induction

We manually label the fields to be extracted, and produce the corresponding wrappers (usually with a GUI tool).

title



[Try it out](#)

```
<html>
<body>
<div>
  ...
  <div>
  ...
  <div>
  ...
  <b>Elvis: Aloha from Hawaii</b>
```



Title:

div[1] → div[2]

Rating:

div[7] → span[2] → b[1]

ReleaseDate:

div[10] → i[1]

Wrapper Induction

We manually label the fields to be extracted, and produce the corresponding wrappers (usually with a GUI tool).

Then we **apply** the wrappers to all pages in the domain.



Title:
div[1] → div[2]

Rating:
div[7] → span[2] → b[1]

ReleaseDate:
div[10] → i[1]

Title	Rating	ReleaseDate
Titanic	7.4	1998-01-07

Xpath

Xpath:

basic syntax: /label/sublabel/...

n-th child: .../label[n]/...


attributes: .../label[@attribute=value]/...

```
<html>
  <body>
    <div>News *** News *** News</div>
    <div id="content">
      Elvis caught with chamber maid in New York hotel
    </div>
  </body>
</html>
```





















```
<html>
  <body>
    <div> News *** News *** News </div>
    <div>Buy Elvis CDs now!!</div>
    <div id="content">
      Carla Bruni works as chamber maid in New York.
    </div>
  </body>
</html>
```

Wrapper Induction

Wrappers can also work inside one page, if the content is repetitive.

DAYPACK OVERVIEW 

ORDER BY Products found 11 Pages 1 2 >

<p>J-PACK DE LUXE »</p>  <p>An exceptionally roomy daypack with a removable notebook section</p> <p>from 89.95 EUR*</p>  	<p>J-PACK XT »</p>  <p>A roomy daypack with a laptop/notebook compartment</p> <p>from 79.95 EUR*</p>  	<p>BRAIN STORM »</p>  <p>Office pack with laptop case and several useful compartments</p> <p>from 69.95 EUR*</p>  
<p>FAST TRACK 24 »</p>  <p>Sporty daypack with excellent ventilation</p> <p>from 69.95 EUR*</p>  	<p>SACRAMENTO »</p>  <p>The big-format book pack</p> <p>from 59.95 EUR*</p>   	<p>SEATTLE »</p>  <p>A versatile book pack with a timeless design</p> <p>from 49.95 EUR*</p>   

Wrapper Induction on 1 Page

Wrappers can also work inside one page, if the content is repetitive.

<p>SACRAMENTO »</p>  <p>The big-format book pack</p> <p>from 59.95 EUR*</p> <p>■ ■ ■</p> <p>in stock</p>	<p>SEATTLE »</p>  <p>A versatile book pack with a timeless design</p> <p>from 49.95 EUR*</p> <p>■ ■ ■</p>
--	--

Problem:

some parts of the repetitive items may be optional or again repetitive

⇒ learn a stable wrapper

Road Runner

```
627.626:<br />
628.Plus .0.628
  0.And
    0.* [Ange Pitou is the third book ...,One of the best-loved Jane Au... ]
    1.934:<img /[[height, vspace, border, width, src]]>
    2.Hook
      0.And
        0.935:<br />
        1.936:<img /[[border, width, alt, src, align]]>
        2.* [The Metamorphosis and Other S...,Lord Jim ]
        3.938:<br />
        4.* [by John Dos Passos,by H. Rider Haggard ]
        5.940:<br />
        6.941:<img /[[height, border, width, src]]>
        7.942:<br />
        8.* [ Our price: FREE, Our Price: FREE ]
        9.944:<br />
        10.945:<br />
        11.946:<a[[href]]>
        12.947:<img /[[height, border, width, alt, src]]>
        13.948:</a[[href]]>
        14.949:<br /[[clear]]>
      3.950:<br />
    629.* [After a dark, mysterious pref...,Burney's Cecilia is an heires... ]
```

Sample system: RoadRunner

<http://www.dia.uniroma3.it/db/roadRunner/>

Wrapper Induction Summary

Wrapper induction can extract entities and relations from a set of similarly structured pages.

Input:

- Choice of the domain
- (Human) labeling of some pages
- Wrapper design choices

Condition:

- All pages are of the same structure

Can the wrapper say things like

“The last child element of this element”

“The second element, if the first element contains XYZ”

?

If so, how do we generalize the wrapper?

Pattern Matching

Known facts (**seed pairs**)

Person	Discovery
Einstein	K68

Einstein ha scoperto il K68, quando aveva 4 anni.

X ha scoperto il Y

Bohr ha scoperto il K69 nel anno 1960.

Person	Discovery
Bohr	K69

- The patterns can either
- be specified by hand
 - or come from annotated text
 - or come from seed pairs + text

Pattern Matching

Einstein ha scoperto il K68, quando aveva 4 anni.

Known facts (**seed pairs**)

Person	Discovery
--------	-----------

Einstein	K68
----------	-----

X ha scoperto il Y

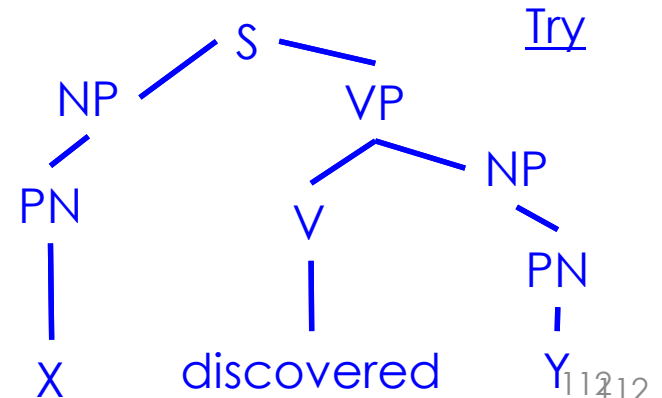
Bohr ha scoperto il K69 nel anno 1960.

Person	Discovery
--------	-----------

Bohr	K69
------	-----

The patterns can be more complex, e.g.

- regular expressions
X found $\{0,20\}$ Y
- parse trees



Pattern Matching

Einstein ha scoperto il K68, quando aveva 4 anni.

Known facts (**seed pairs**)

Person	Discovery
Einstein	K68

X ha scoperto il Y

Bohr ha scoperto il K69 nel anno 1960.

First system to use iteration:
Snowball

Person	Discovery
Bohr	K69

Watch out for semantic drift:
Einstein liked the K68

Pattern Matching

Pattern matching can extract facts from natural language text corpora.

Input:

- a known relation
- seed pairs or labeled documents or patterns

Condition:

- The texts are homogenous
(express facts in a similar way)
- Entities that stand in the relation do not stand in another relation as well

Open Calais

presidential race (Political Event)

Relevance: 42%

Count: 1

politicaleventtype: Voting

location: Brazil

that Brazil's presidential race will go to a second round. Did it over 46% of all votes counted so far. That will rise a smidger expected gains there will not be enough to secure an absolute m

Entities:

Country

Brazil

Person

Luiz Inácio Lula da Silva

Political Event

presidential race

Position

popular president

president

Events & Facts:

Person Career

Luiz Inácio Lula da Silva, popular president, political,

Try this out:

<http://viewer.opencalais.com/>

Cleaning

Fact Extraction commonly produces huge amounts of garbage.

Web page contains misleading items (advertisements, error messages)

Web page contains bogus information

Deviation in iteration

Formatting problems (bad HTML, character encoding mess)

Regularity in the training set that does not appear in the real world

Something has changed over time (facts or page formatting)

Different thematic domains or Internet domains behave in a completely different way

⇒ Cleaning is usually necessary, e.g., through thresholding or heuristics

Fact Extraction Summary

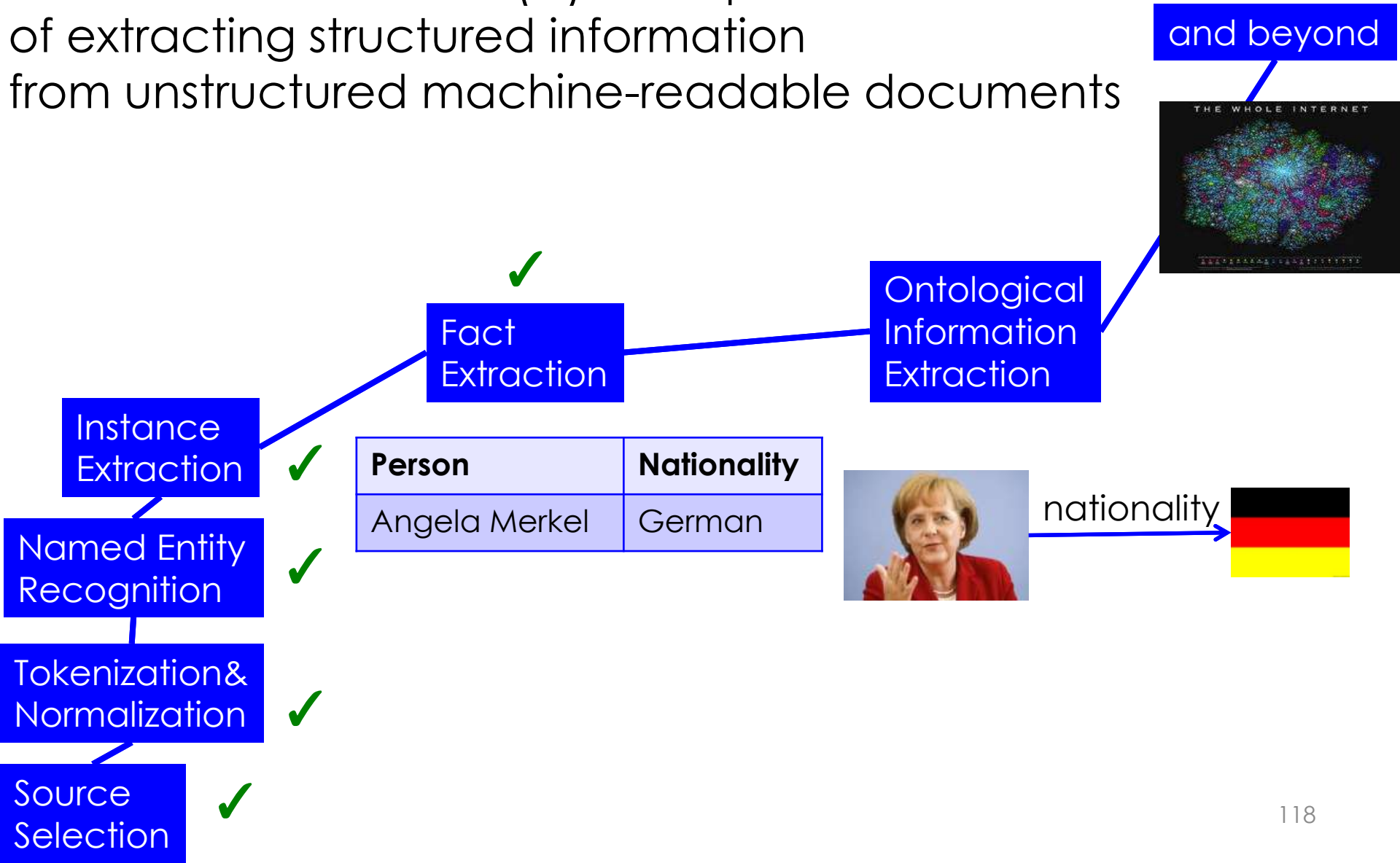
Fact Extraction is the process of extracting pairs (triples,...) of entities together with the relationship of the entities.

Approaches:

- Fact extraction from tables
(if the corpus contains lots of tables)
- Wrapper induction
(for extraction from one Internet domain)
- Pattern matching
(for extraction from natural language documents)
- ... and many others...

Information Extraction

Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents



Ontologies

An **ontology** is consistent knowledge base without redundancy

Person	Nationality
Angela Merkel	German
Merkel	Germany
A. Merkel	French



Entity	Relation	Entity
Angela Merkel	citizenOf	Germany



- Every entity appears only with exactly the same name
- There are no semantic contradictions

Ontological IE

Ontological Information Extraction (IE) aims to create or extend an ontology.

Entity	Relation	Entity
Angela Merkel	citizenOf	Germany



Angela Merkel is the German chancellor....
...Merkel was born in Germany...

...A. Merkel has French nationality...

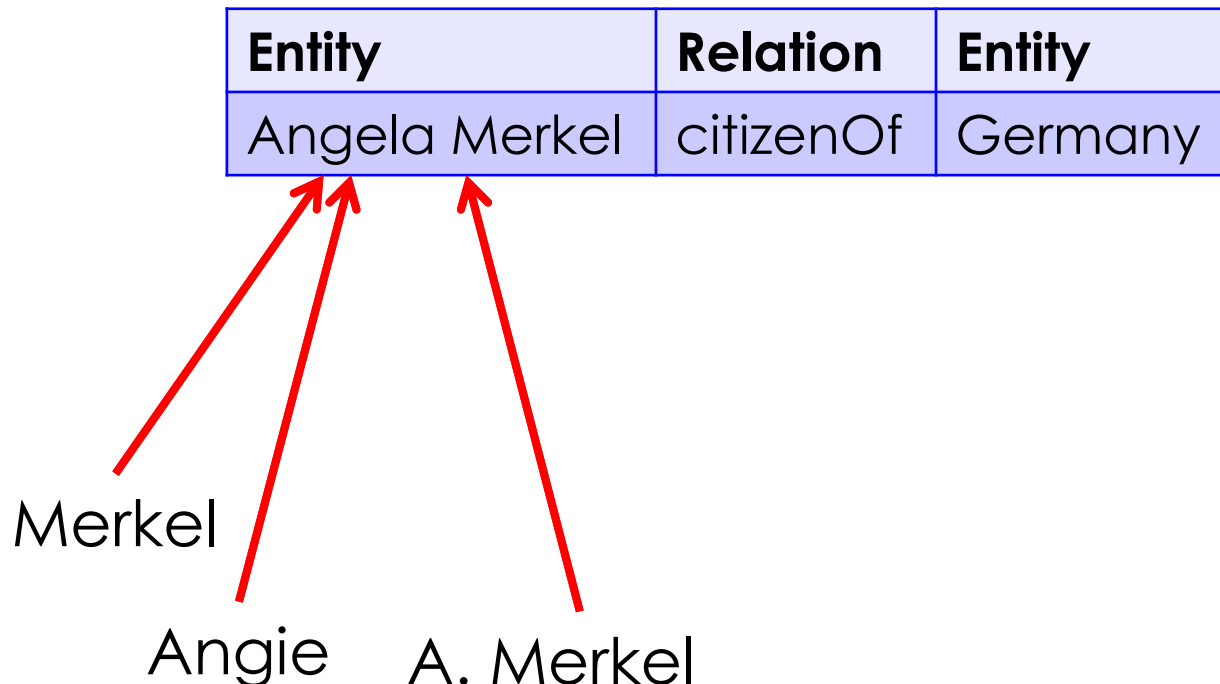


Person	Nationality
Angela Merkel	German
Merkel	Germany
A. Merkel	French

Ontological IE Challenges

Challenge 1:

Map names to names that are already known



Ontological IE Challenges

Challenge 2:

Be sure to map the names to the right known names

Entity	Relation	Entity
Angela Merkel	citizenOf	Germany
Una Merkel	citizenOf	USA



?

Merkel is great!

Ontological IE Challenges

Challenge 3:

Map to known relationships

Entity	Relation	Entity
Angela Merkel	citizenOf	Germany



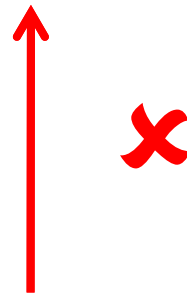
... has nationality ...
... has citizenship ...
... is citizen of ...

Ontological IE Challenges

Challenge 4:

Take care of consistency

Entity	Relation	Entity
Angela Merkel	citizenOf	Germany



Angela Merkel is
French...

Triples

A **triple** (in the sense of ontologies) is a tuple of an entity, a relation name and another entity:

Entity	Relation	Entity
Angela Merkel	citizenOf	Germany

=



=

<Angela Merkel, citizenOf, Germany>

Triples

A **triple** (in the sense of ontologies) is a tuple of an entity, a relation name and another entity:

Entity	Relation	Entity
Angela Merkel	citizenOf	Germany

Most ontological IE approaches produce triples as output. This decreases the variance in schema.

Person	Country
Angela	Germany

Citizen	Nationality
Angela	Germany

Person	Birthdate	Country
Angela	1980	Germany

Wikipedia



- Wikipedia is a free online encyclopedia
- 3.4 million articles in English
 - 16 million articles in dozens of languages

Why is Wikipedia good for information extraction?

- It is a huge, but homogenous resource
(more homogenous than the Web)
- It is considered authoritative
(more authoritative than a random Web page)
- It is well-structured with infoboxes and categories
- It provides a wealth of meta information
(inter article links, inter language links, user discussion,...)

Ontological IE from Wikipedia



- Wikipedia is a free online encyclopedia
- 3.4 million articles in English
 - 16 million articles in dozens of languages

Every article is (should be) unique
=> We get a set of unique entities
that cover numerous areas of interest



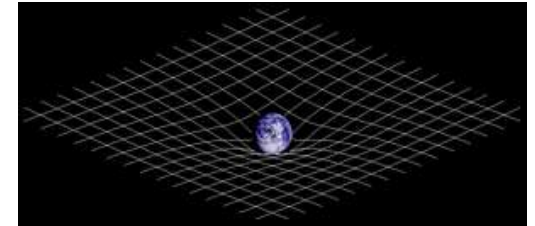
Angela_Merkel



Una_Merkel

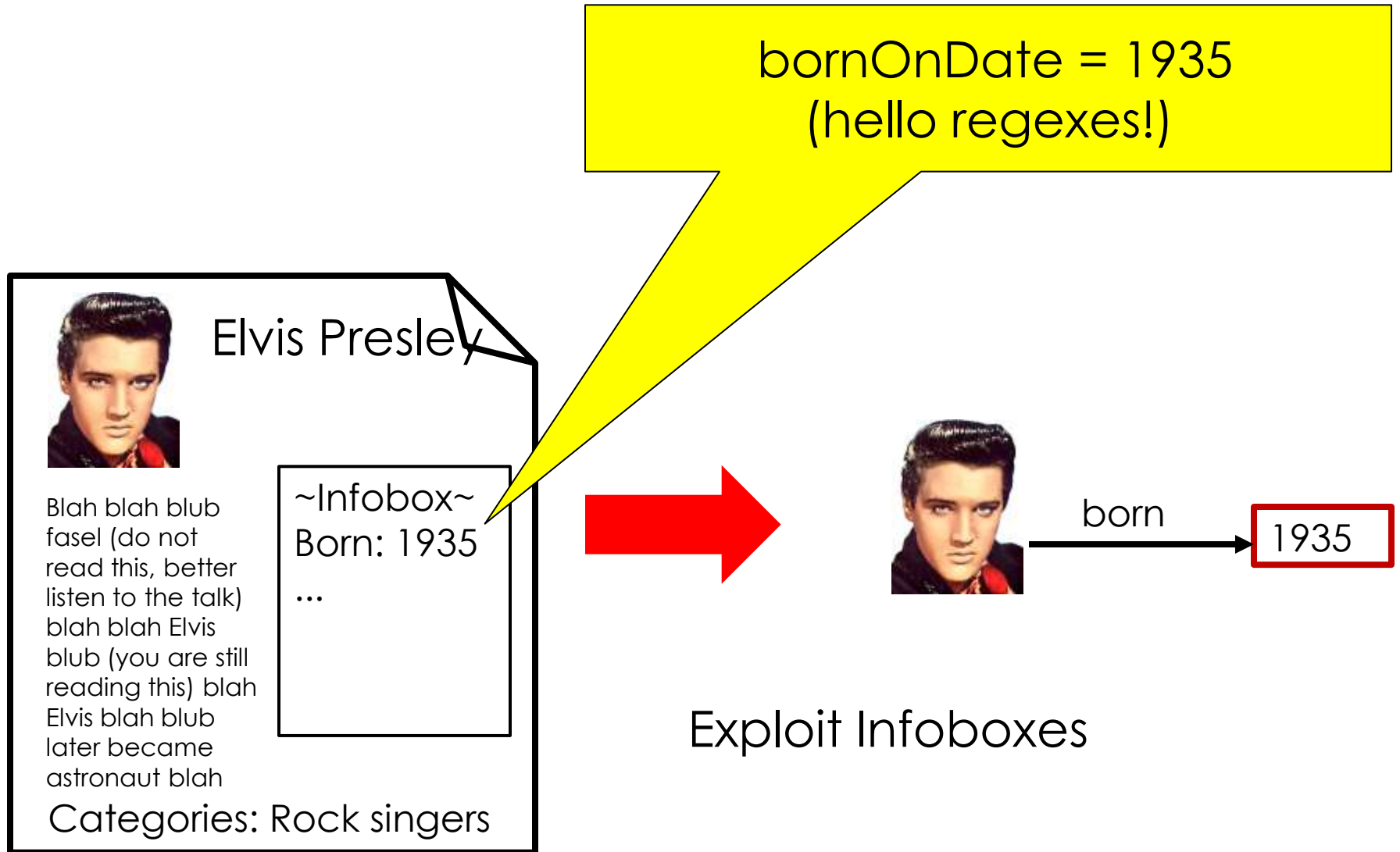


Germany




Theory_of_Relativity

IE from Wikipedia



IE from Wikipedia

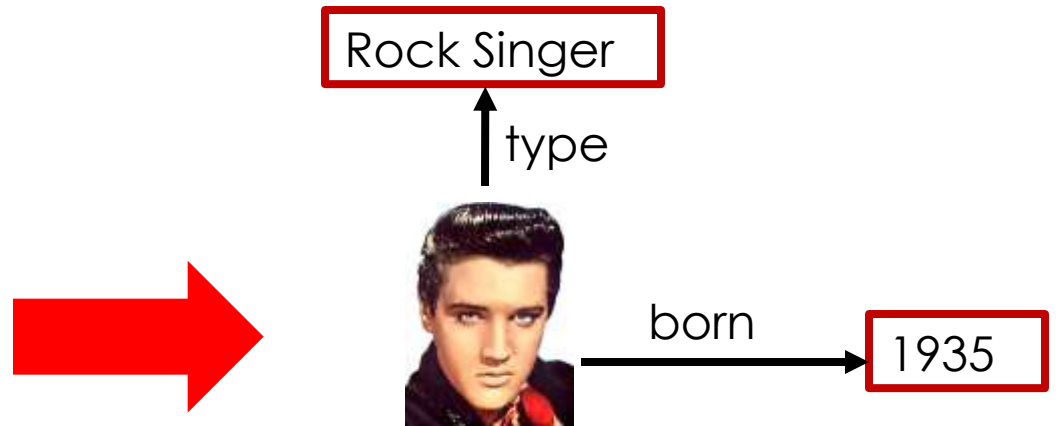


Elvis Presley

Blah blah blub
fasel (do not
read this, better
listen to the talk)
blah blah Elvis
blub (you are still
reading this) blah
Elvis blah blub
later became
astronaut blah

Categories: Rock singers

~Infobox~
Born: 1935
...


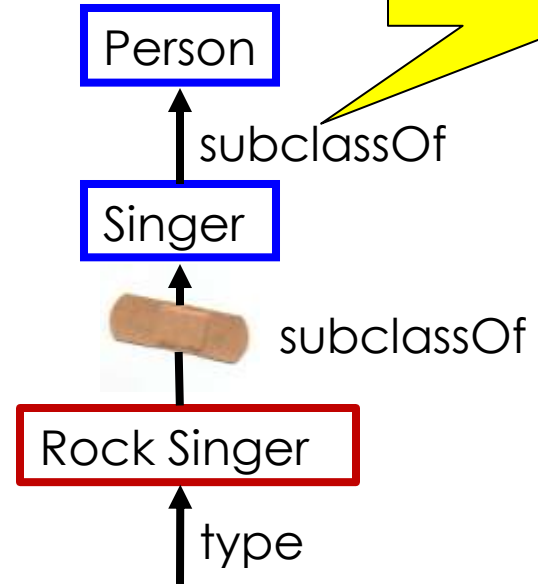
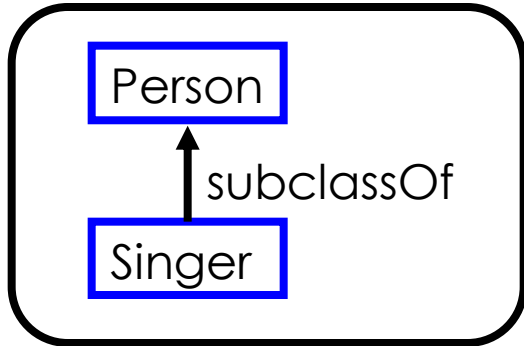


Exploit Infoboxes
Exploit conceptual categories

IE from Wikipedia

Every singer is a person

WordNet

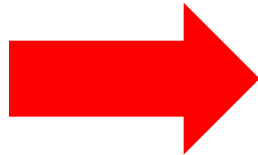


Elvis Presley

Blah blah blub
fasel (do not read this, better listen to the talk)
blah blah Elvis blub (you are still reading this) blah
Elvis blah blub later became astronaut blah

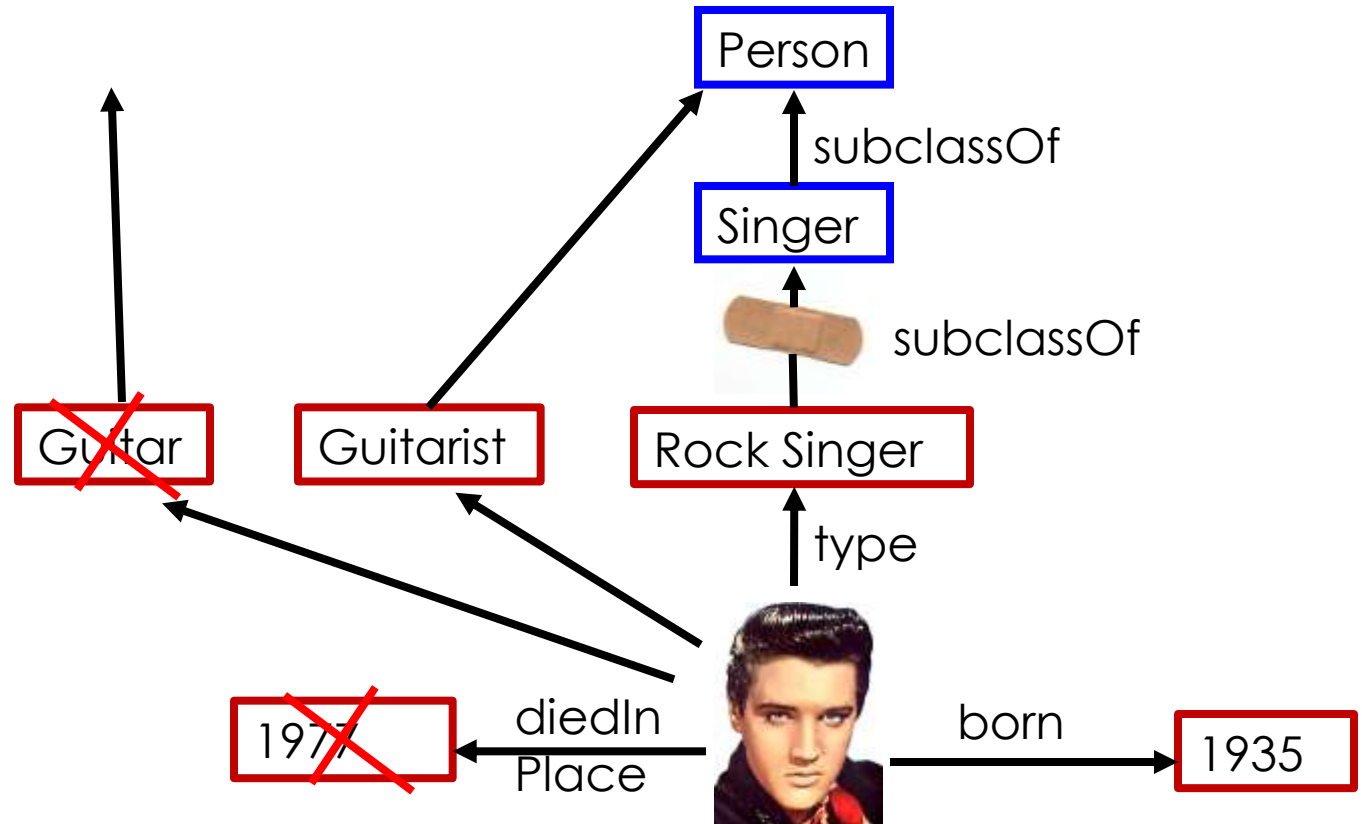
Categories: Rock singers

~Infobox~
Born: 1935
...



Exploit Infoboxes
Exploit conceptual categories

Consistency Checks



- Check uniqueness of functional arguments
- Check domains and ranges of relations
- Check type coherence

Wikipedia Source

Example: [Elvis on Wikipedia](#)

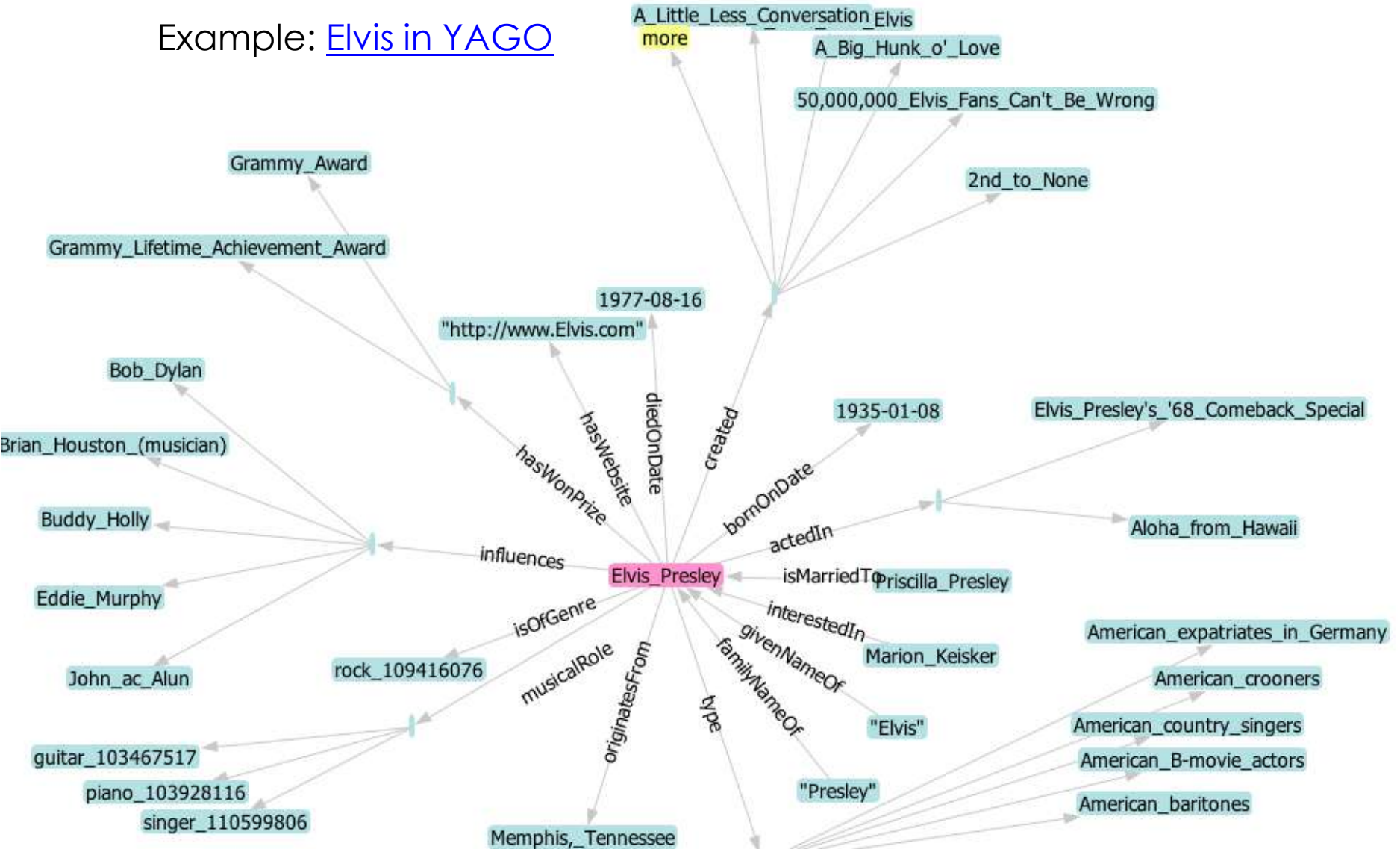
Background information

Birth name	Elvis Aaron Presley
Born	January 8, 1935 Tupelo, Mississippi , United States
Died	August 16, 1977 (aged 42) Memphis, Tennessee , United States
Genres	Rock and roll , pop , rockabilly , country , blues , gospel , R&B
Occupations	Musician, actor
Instruments	Vocals, guitar, piano
Years active	1954–77
Labels	Sun , RCA Victor
Associated acts	The Blue Moon Boys , The Jordanaires , The Imperials
Website	www.elvis.com 

```
| Birth_name = Elvis Aaron Presley  
| Born = {{Birth date | 1935 | 1 | 8}}<br />  
[[Tupelo, Mississippi | Tupelo]]
```

YAGO

Example: [Elvis in YAGO](#)



Ontological IE from Wikipedia



YAGO

- 3m entities, 28m facts
- focus on precision 95%
(automatic checking of facts)

<http://mpii.de/yago>



DBpedia

- 3.4m entities
- 1b facts (also from non-English Wikipedia)
- large community

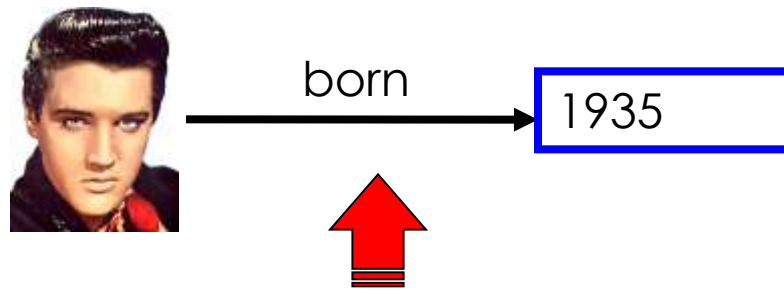
<http://dbpedia.org>



Community project on top of Wikipedia
(bought by Google, but still open)

<http://freebase.com>

Ontological IE by Reasoning



Elvis was born in 1935

Recap: The challenges:

- deliver canonic relations
- deliver canonic entities
- deliver consistent facts

died in, was killed in

Elvis, Elvis Presley, The King

born (Elvis, 1970)
born (Elvis, 1935)

Idea: These problems are interleaved,
solve all of them together.

Using Reasoning

Ontology



First Order Logic

```
type(Elvis_Presley,singer)
subclassof(singer,person)
...

appears("Elvis","was born in",
"1935")

...
means("Elvis",Elvis_Presley,0.8)
means("Elvis",Elvis_Costello,0.2)
...

born(X,Y) & died(X,Z) => Y<Z
appears(A,P,B) & R(A,B)
=> expresses(P,R)
appears(A,P,B) & expresses(P,R)
=> R(A,B)
...
```

Documents

Elvis was born in 1935

Consistency

Rules

birthdate<deathdate



born

1935

SOFIE
system

MAX SAT

A **Weighted Maximum Satisfiability Problem (WMAXSAT)**

is a set of propositional logic formulae with weights.

A	[10]
A => B	[5]
-B	[10]

A **solution** to a WMAXSAT is an assignment of the variables to truth values. Its weight is the sum of weights of satisfied formulas

Solution 1:

A=true

B=true

Weight: $10+5=15$

Solution 2:

A=true

B=false

Weight: $10+10=20$

MAX SAT

A **Weighted Maximum Satisfiability Problem (WMAXSAT)**

is a set of propositional logic formulae with weights.

The **optimal solution** is a solution is a solution that maximizes the sum of the weights of the satisfied formulae.

The optimal solution is NP hard to compute
=> use a (smart) approximation algorithm

Solution 1:

A=true

B=true

Weight: $10+5=15$

Solution 2:

A=true

B=false

Weight: $10+10=20$



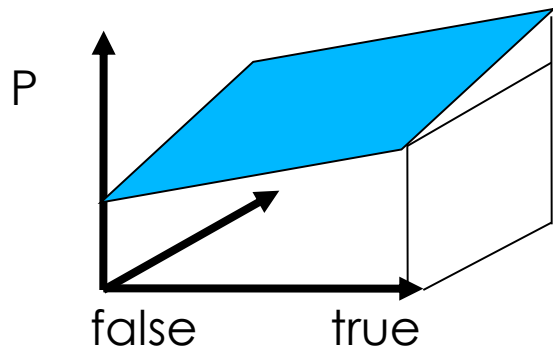
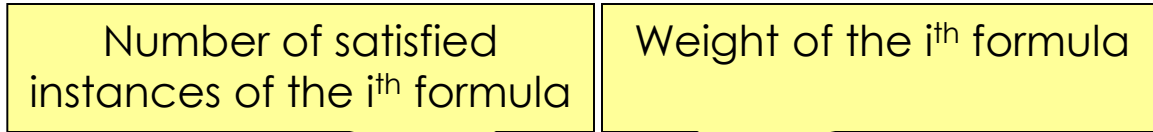
Markov Logic

A Markov Logic Program

is a set of propositional logic formulae with weights
(can be generalized to first order logic)

A	[10]
A => B	[5]
-B	[10]

... with a probabilistic interpretation:
Every solution (possible world) has
a certain probability



bornIn(Elvis, Tupelo)

$$P(X) \sim \prod e^{\text{sat}(i,X) w_i}$$

$$\max_x \prod e^{\text{sat}(i,X) w_i}$$

$$\max_x \log(\prod e^{\text{sat}(i,X) w_i})$$

$$\max_x \sum \text{sat}(i,X) w_i$$

Weighted MAX SAT problem

Ontological IE by Reasoning

Reasoning-based approaches use logical rules to extract knowledge from natural language documents.

Current approaches use either

- Weighted MAX SAT
- or Datalog
- or Markov Logic

Input:

- often an ontology
- manually designed rules

Condition:

- homogeneous corpus helps

Ontological IE Summary

Ontological Information Extraction (IE) tries to create or extend an ontology through information extraction.

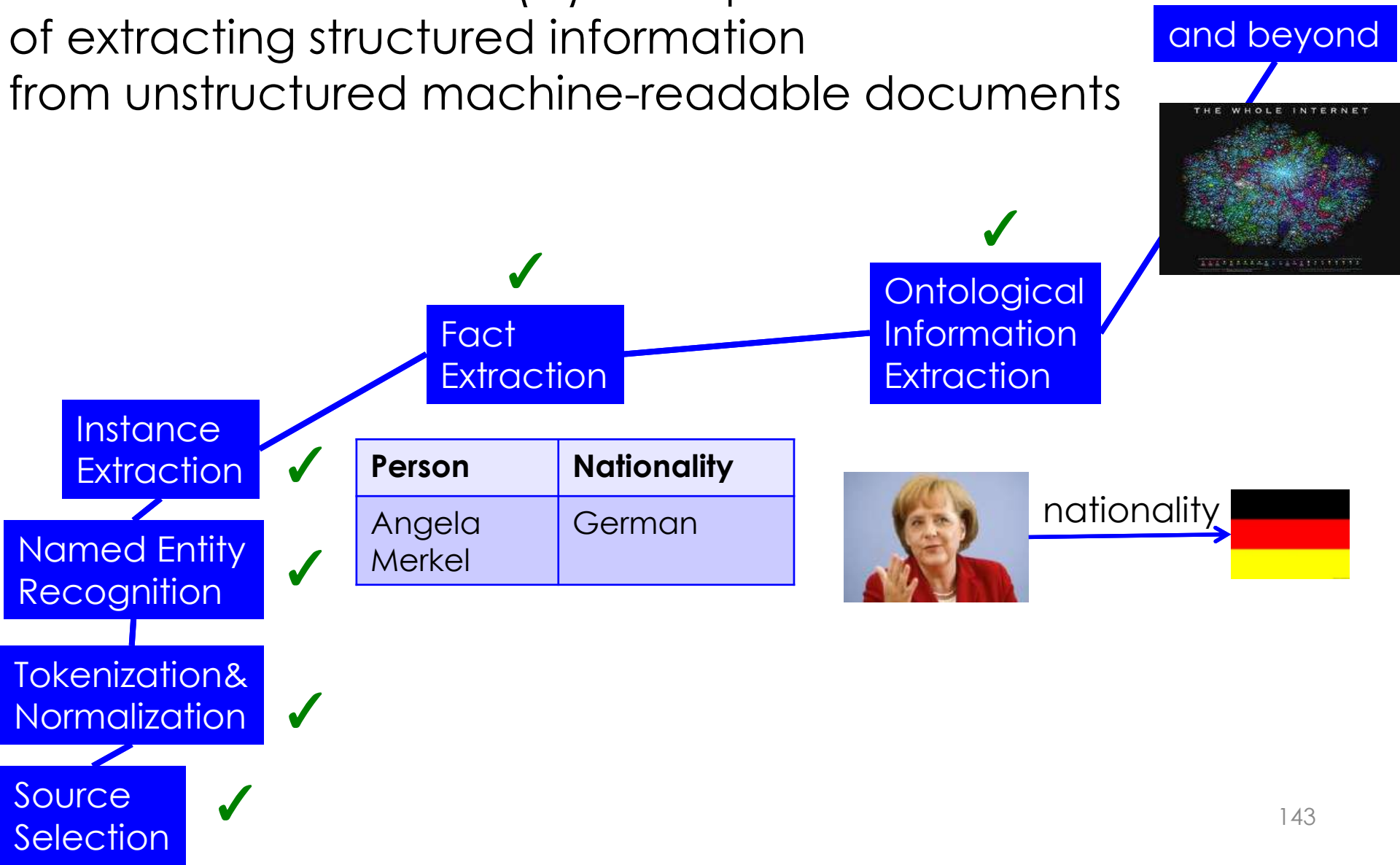


Current hot approaches:

- extraction from Wikipedia
- reasoning-based approaches

Information Extraction

Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents



Open Information Extraction

Open Information Extraction/Machine Reading

aims at information extraction from the entire Web.

Vision of Open Information Extraction:

- the system runs perpetually, constantly gathering new information
- the system creates meaning on its own from the gathered data
- the system learns and becomes more intelligent, i.e. better at gathering information

Open Information Extraction

Open Information Extraction/Machine Reading

aims at information extraction from the entire Web.

Rationale for Open Information Extraction:

- We do not need to care for every single sentence, but just for the ones we understand
- The size of the Web generates redundancy
- The size of the Web can generate synergies

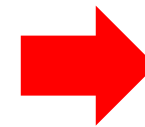
KnowItAll & Co

KnowItAll, KnowItNow and TextRunner are projects at the University of Washington (in Seattle, WA).

*Egyptian
complex.* more than the question of how the
Egyptians built the pyramids was,
he says, "how the pyramids built



Subject	Verb	Object	Count
Egyptians	built	pyramids	400
Americans	built	pyramids	20
...



Valuable
common sense
knowledge
(if filtered)

KnowItAll & Co

TextRunner took .80 seconds.

Retrieved **391** results for Predicate containing "**built**" and Argument 2 containing "**pyramids**"

Grouping results by predicate. Group by: [argument 2](#) | [argument 1](#)

built - 159 results

Egyptians (297), aliens (71), Pharaohs (40), [85 more...](#) **built** the **pyramids**

Egyptians (26), Khufu (18), Maya (9), [30 more...](#) **built** the Great **Pyramid**

Imhotep (8), Pharaoh Zoser (4), Egyptians (2), King Djoser (2) **built** the Step **Pyramid**

two symbols of life (4), 6th dynasty kings (3), King Sneferu (3), Snefru (3) **built** two large **Pyramids**

Egyptians (8) **built** the Great **Pyramids**

ancient Egyptians (6) **built** more than 90 royal **pyramids**

colonial silver city of Taxco (3), Explore (2) **built** the gigantic **pyramids** of the Sun

Central America (2), part of Mexico (2) **built** great cities , temples and **pyramids**

Read the Web

“Read the Web” is a project at the Carnegie Mellon University in Pittsburgh, PA.

Initial Ontology

Table Extractor

Krzewski	Blue Angels
Miller	Red Angels

Natural Language
Pattern Extractor

Krzewski coaches
the Blue Devils.

Mutual exclusion

sports coach != scientist

Type Check

If I coach, am I a coach?

Open IE: Read the Web

- arthropod (100.0%)

- Seed

- CPL @156 (100.0%) on 30-sep-2010 ["hind wings of _ "invertebrates , such as _ "
" _ swarm from" "other insects , including _ " _ marching home" "honeydew produce
like _ " "other insects , such as _ " _ do not eat wood" "many legs as _ " _ produce s
have complete metamorphosis" "I do n't see anymore _ " "ants , so _ " "insecticide fo
"such insects as _ " _ are the only insects" "red imported _ " "insects like _ " "social in
, such as _ " "arthropods include _ " "insect pests including _ " "meaty foods like _ " _
pests , such as _ " "other insects such as _ " "insects , in particular _ " _ release a ph
like _ " "many insects , including _ " _ are social insects" "insect pests such as _ " _ a
pests , including _ " "arthropods , including _ " _ are beneficial insects" _ are comm
"arthropods , such as _ "]

- SEAL @151 (50.0%) on 26-sep-2010 [1]

kateretes (Seed)

mosquito (Seed)

peppered moth (Seed)

sap beetle (Seed)

tettigoniidae (Seed)

triatoma protracta (Seed)

honeylocust spider mite

grape flea beetle

blueberry leaf beetle

sugarcane moth borer

psychoda moth flies

bagworm moth

carpenterworm moths

leafcurl plum aphid

merchant grain beetle

NELL Know

CMU Read the Web

- v
- fung
- plan
- arch
- bact
- politica
- color
- language
- programminglanguage
- dateliteral
- gamescore
- nonnegativeinteger
- politicsissue
- llcoordinate
- agent
 - animal
 - invertebrate
 - arthropod
 - arachnid
 - insect
 - crustacean
 - mollusk
 - vertebrate
 - amphibian
 - bird
 - fish

Open Information Extraction

Open Information Extraction/Machine Reading

aims at information extraction from the entire Web.

Main hot projects

- TextRunner
- Read the Web
- Prospera (from SOFIE)

Input:

- The Web
- Read the Web: Manual rules
- Read the Web: initial ontology

Conditions

- none

Information Extraction

Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents

