Information Extraction

session in the course “Web Search” at the
École nationale supérieure des télécommunications
in Paris/France in spring 2011

by Fabian M. Suchanek

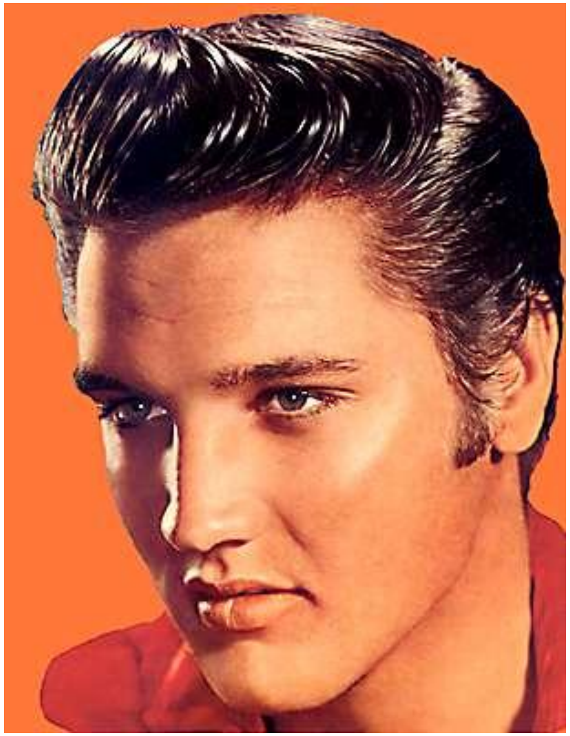
Organisation

- 3h class on Information extraction

- 3h lab session

- Web-sites:
  - http://suchanek.name/ → Teaching
  - http://pierre.senellart.com/enseignement/2010-2011/inf396/

Motivation



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

Elvis Presley
1935 - 1977

Will there ever be someone like him again?

Motivation

Google

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

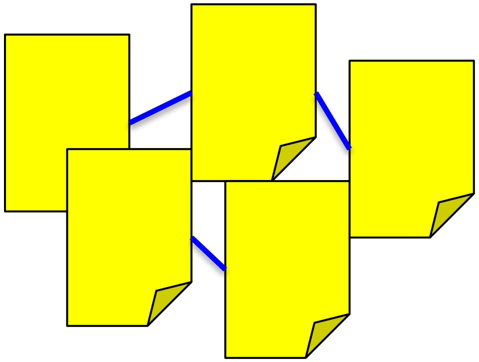
Another singer called Elvis, young

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



✗

Information Extraction

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

GName	FName	Occupation
Elvis	Presley	singer
Elvis	Hunter	painter
…	…	

1: Elvis Presley
2: Elvis …
3. Elvis …

Motivation: Definition

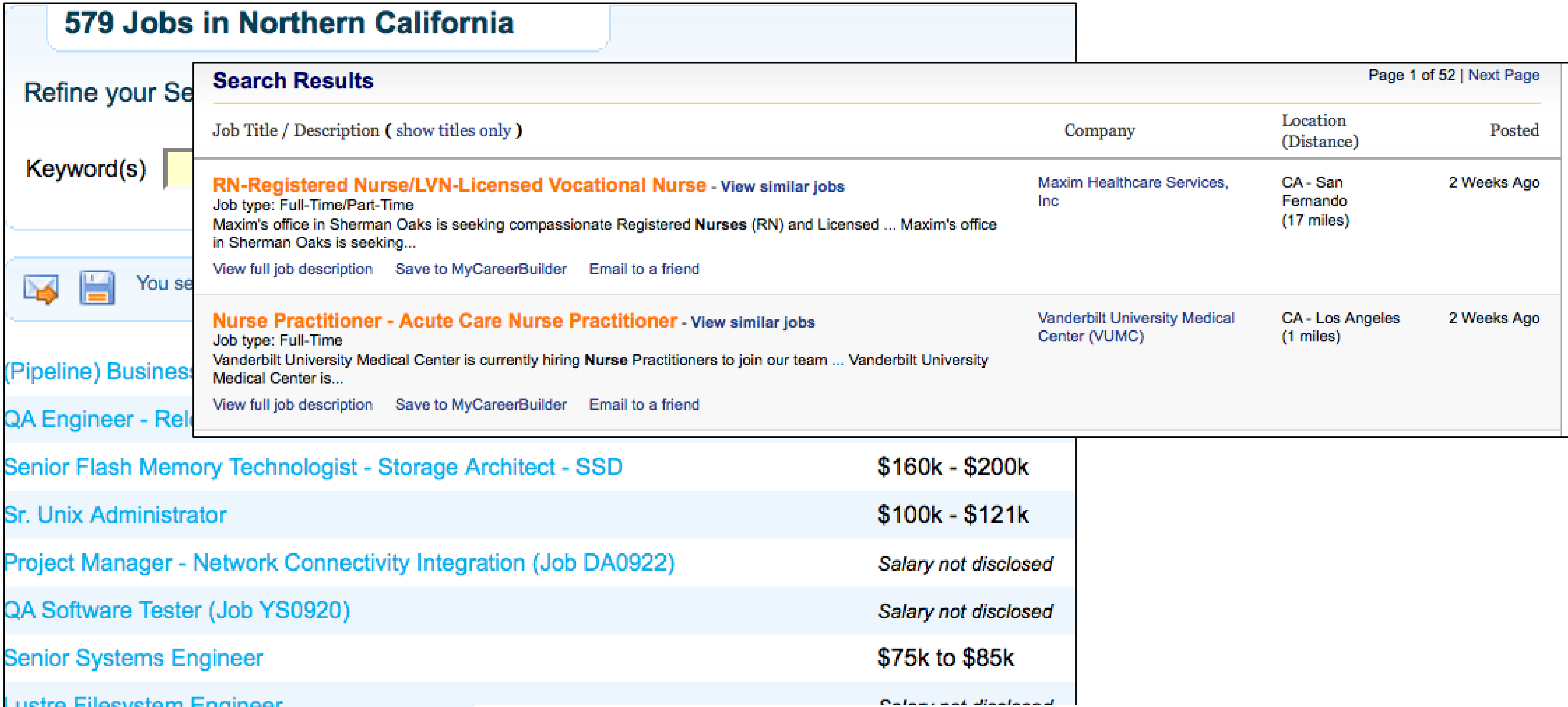
Information Extraction (IE) is the process
of extracting structured information (e.g., database tables)
from unstructured machine-readable documents (e.g., Web documents).

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
...	...	

Motivation: Examples



579 Jobs in Northern California

Refine your Se

Keyword(s)

You se

Search Results

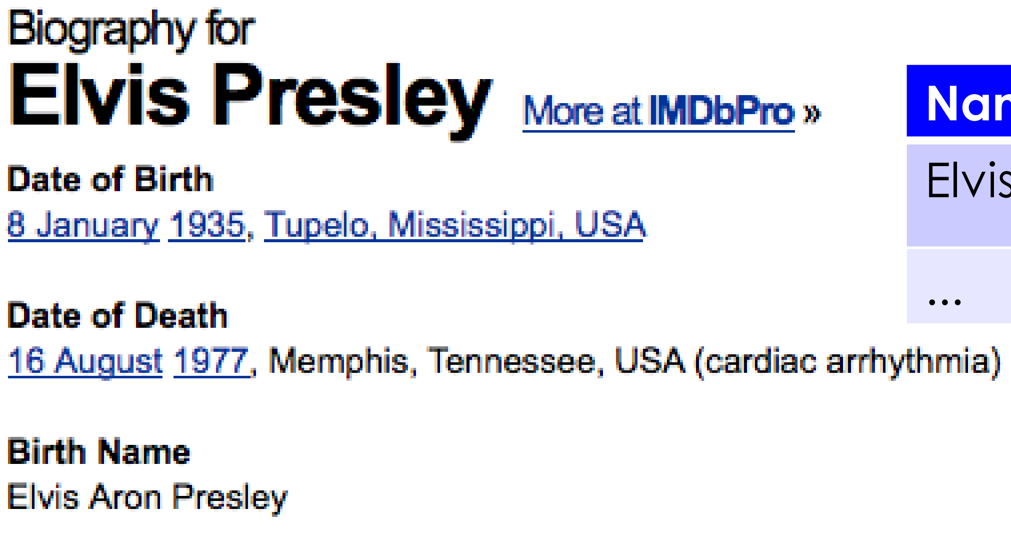
Page 1 of 52 | Next Page

Job Title / Description (show titles only)	Company	Location (Distance)	Posted
RN-Registered Nurse/LVN-Licensed Vocational Nurse - **View similar jobs** Job type: Full-Time/Part-Time Maxim's office in Sherman Oaks is seeking compassionate Registered **Nurses** (RN) and Licensed ... Maxim's office in Sherman Oaks is seeking... View full job description Save to MyCareerBuilder Email to a friend	Maxim Healthcare Services, Inc	CA - San Fernando (17 miles)	2 Weeks Ago
Nurse Practitioner - Acute Care Nurse Practitioner - **View similar jobs** Job type: Full-Time Vanderbilt University Medical Center is currently hiring **Nurse** Practitioners to join our team ... Vanderbilt University Medical Center is... View full job description Save to MyCareerBuilder Email to a friend	Vanderbilt University Medical Center (VUMC)	CA - Los Angeles (1 miles)	2 Weeks Ago

(Pipeline) Business	
QA Engineer - Rel	
Senior Flash Memory Technologist - Storage Architect - SSD	$160k - $200k
Sr. Unix Administrator	$100k - $121k
Project Manager - Network Connectivity Integration (Job DA0922)	*Salary not disclosed*
QA Software Tester (Job YS0920)	*Salary not disclosed*
Senior Systems Engineer	$75k to $85k
Lustre Filesystem Engineer	*Salary not disclosed*

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

Motivation: Examples



Biography for

Elvis Presley More at IMDbPro »

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

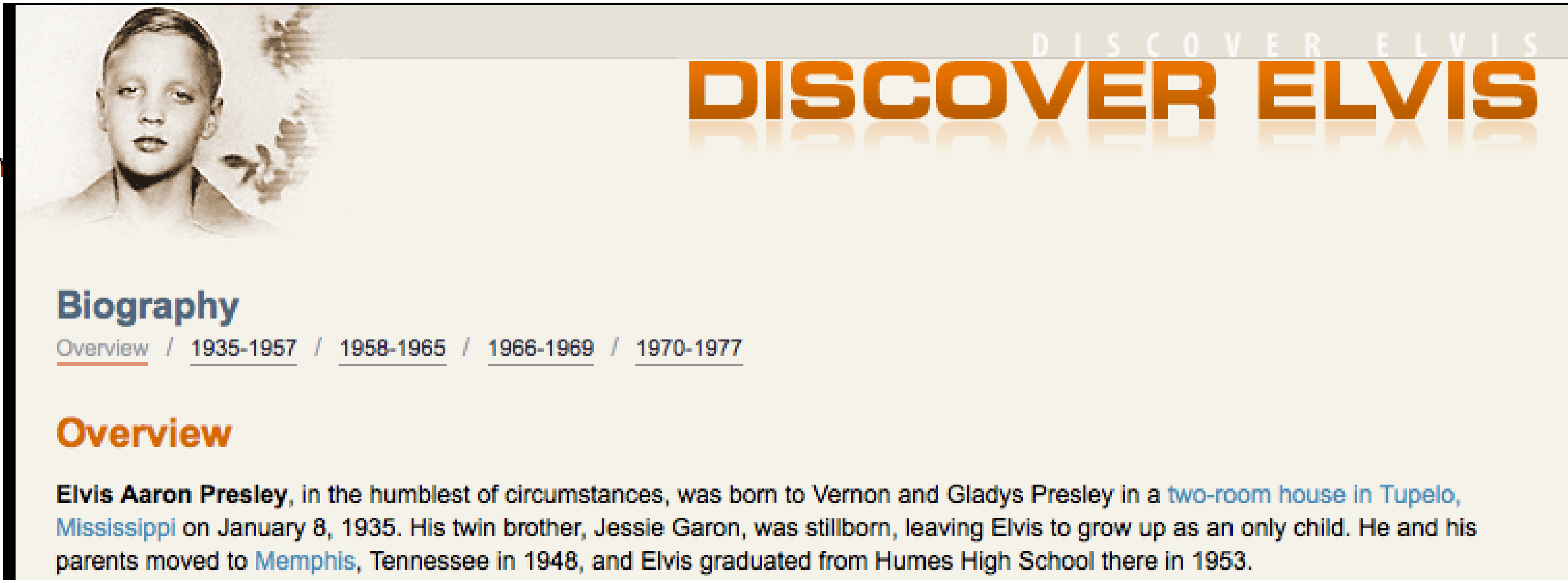
Height

6' (1.83 m)

Mini Biography

Elvis Aaron Presley

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



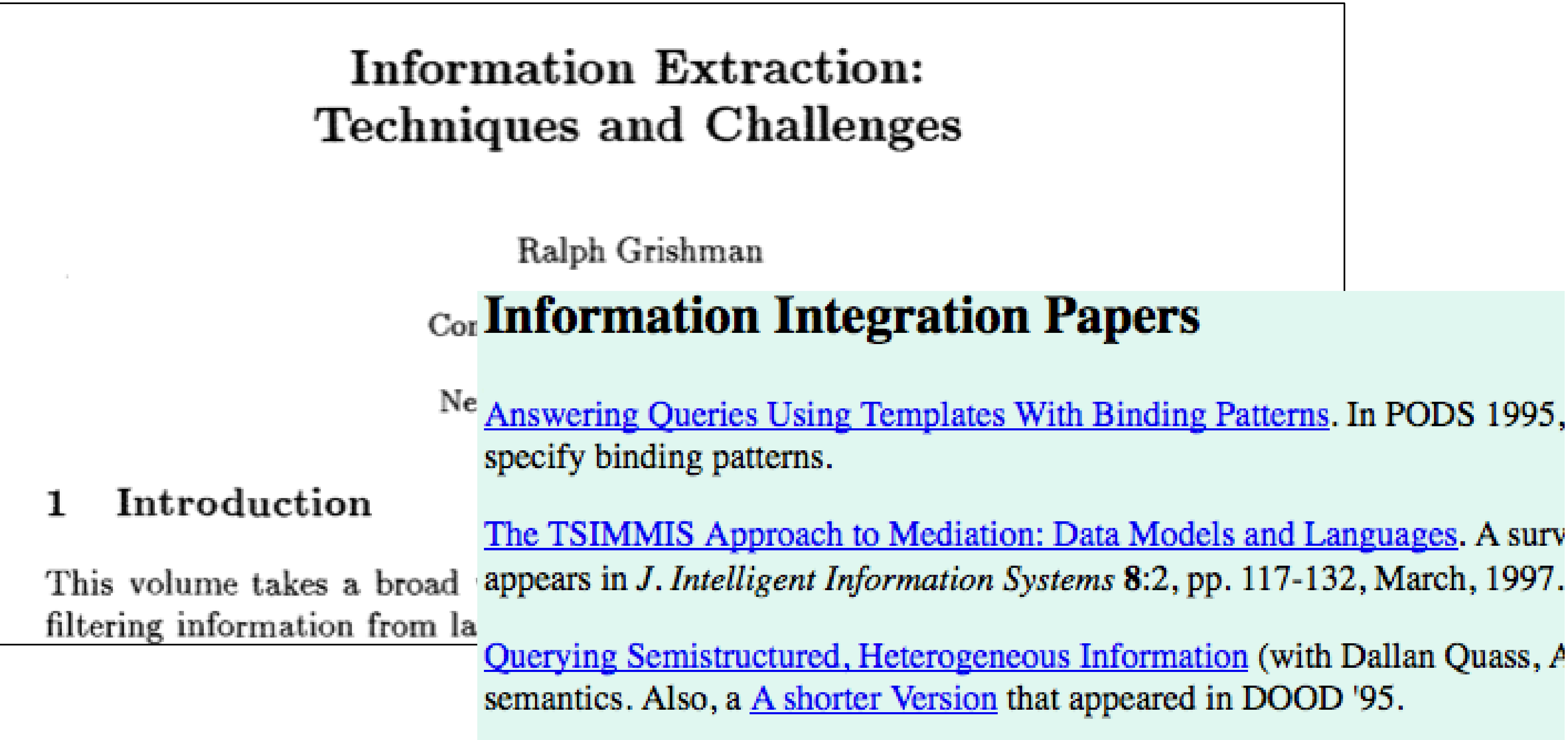
Biography

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.

Motivation: Examples



Information Extraction:
Techniques and Challenges

Ralph Grishman

Co

Ne

1 Introduction

This volume takes a broad
filtering information from la

Information Integration Papers

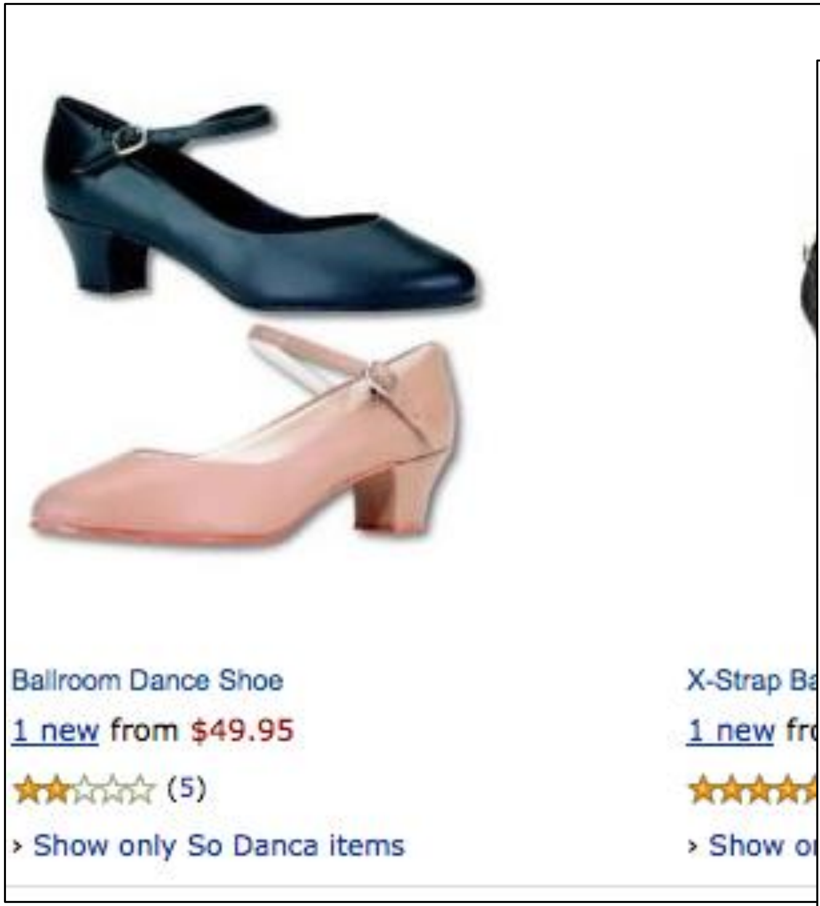
Answering Queries Using Templates With Binding Patterns. In PODS 1995,
specify binding patterns.

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

Querying Semistructured, Heterogeneous Information (with Dallan Quass, A
semantics. Also, a A shorter Version that appeared in DOOD '95.

Author	Publication	Year
Grishman	Information Extraction…	2006
…	…	…

Motivation: Examples



Ballroom Dance Shoe

1 new from $49.95

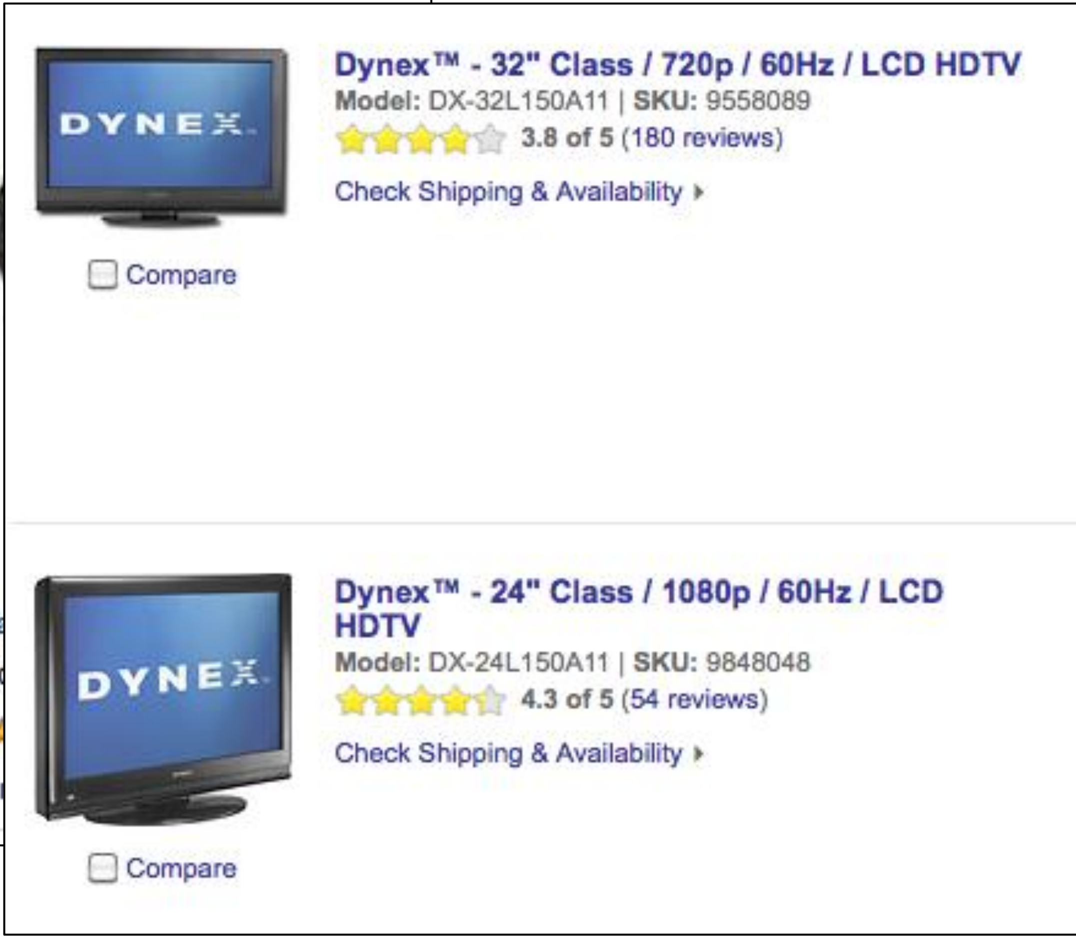
★★☆☆☆ (5)

› Show only So Danca items

X-Strap Ba

1 new fr

› Show o



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

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

Named Entity Recognition

...married Elvis
on 1967-05-01

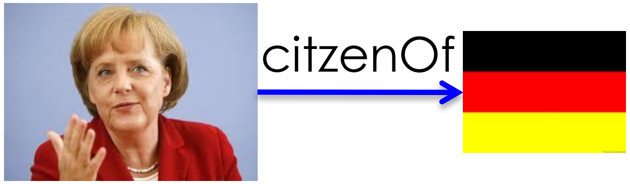
Instance Extraction

Elvis Presley	singer
Angela Merkel	politician

Fact Extraction

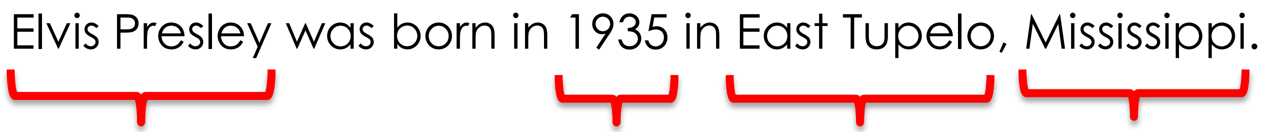
Person	**Nationality**
Angela Merkel	German

Ontological Information Extraction



Named Entity Recognition

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


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

We can extract different types of entities:

- Entities for which we have an exhaustive dictionary (**closed set extraction**)

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

States of the USA

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

Countries of the World (?)

May not always be trivial...

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

Named Entity Recognition

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

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

We can extract different types of entities:

- Entities for which we have an exhaustive dictionary (**closed set extraction**)
- Proper names (**open set extraction**)

... together with the software engineer Bob "the coder" Miller...

People

... The region of Northern Urzykistan has been at war with Southern Urzykistan ever since 1208, when...

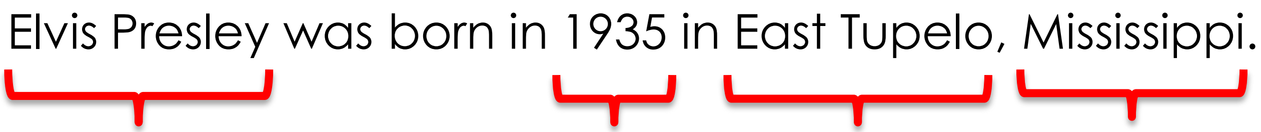
Locations

... BrightFridge Inc. presented their new product, the self-reloading fridge, at this year's exposition in Paris...

Organizations

Named Entity Recognition

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



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

We can extract different types of entities:

- Entities for which we have an exhaustive dictionary (**closed set extraction**)
- Proper names (**open set extraction**)
- Entities that follow a certain pattern

... 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

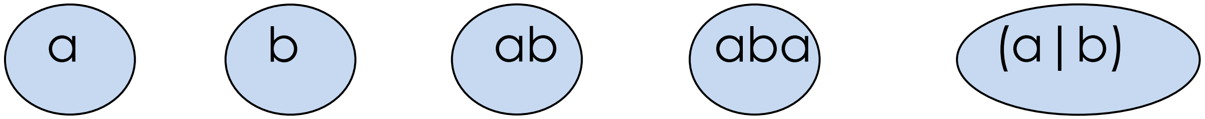
Phone numbers
(groups of digits)

NER: 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 Σ={a,b}, the following strings are regular expressions:



a

b

ab

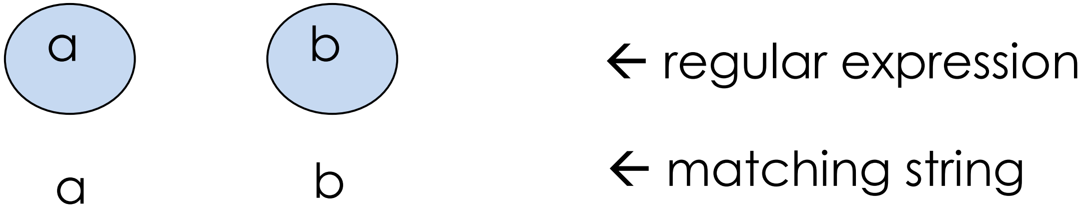
aba

(a|b)

NER: Regular Expressions

Matching

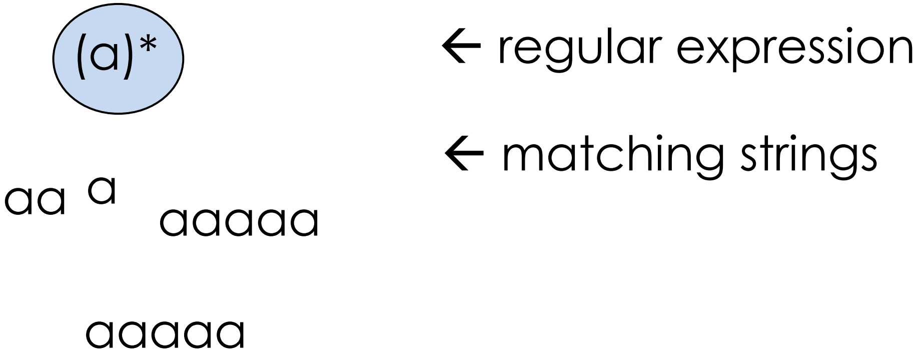
- a string **matches** a regex of a single character
  if the string consists of just that character



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

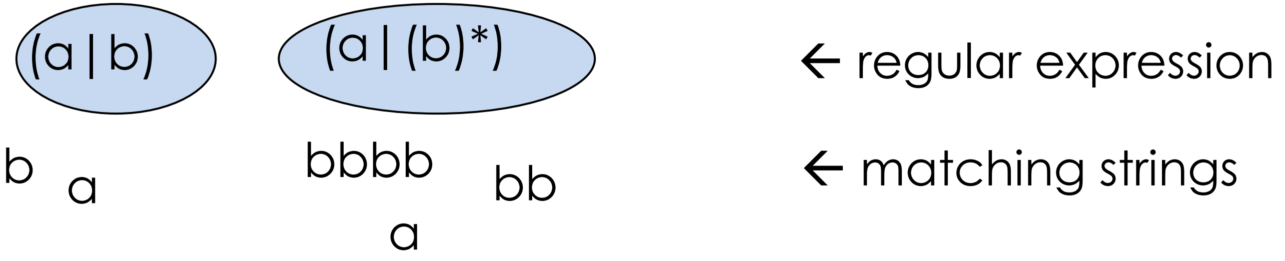
aa a aaaaa ← matching strings

aaaaa

NER: Regular Expressions

Matching

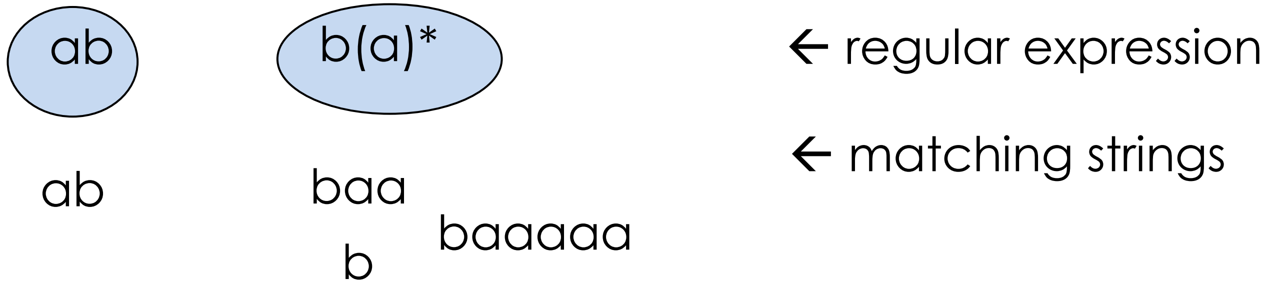
- a string matches a regex of the form (A|B)
  if it matches either A or B



(a|b) (a|(b)*) ← regular expression

b a bbbb bb a ← matching strings

- 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



ab b(a)* ← regular expression

ab baa baaaaa b ← matching strings

NER: Regular Expressions

Given an ordered set of symbols Σ, we define

- [x-y] for two symbols x and y, x<y, to be the alternation
  x | ... | 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...A | A...A | A...A | ... | A...A   (meaning: x to y A’s)

f{4,6} = ffff | fffff | ffffff

- A? for a regex A to be
  (| A)   (meaning: an optional A)

ab? = a(| b)

- . to be an arbitrary symbol from Σ

NER: Regular Expressions

A | B	Either A or B
A*	Zero or more occurrences of A
A+	One or more 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

Example

A digit

A digit or a letter

A sequence of 8 digits

5 pairs of digits, separated by space

5 pairs of digits, separated by a space or a hyphen

Numbers in scientific format

HTML attributes

Dates

NER: Regular Expressions

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

NER: Regular Expressions

A | B	Either A or B
A*	Zero or more occurrences of A
A+	One or more 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

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 syntactic pattern

NER: Normalization

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

  May be too strong: “President Bush” == “president bush”

- **Removing accents** and **umlauts**

  résumé → resume, Universität → Universitaet

- **Normalizing abbreviations**

  U.S.A. → USA, US → USA

NER: Normalization

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

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

Named Entity Recognition

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)
- Extraction with Regular Expressions (if the entities follow a pattern)

We often need normalization in addition.

Information Extraction

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

✓

Named Entity Recognition

...married Elvis on 1967-05-01

Instance Extraction

Elvis Presley	singer
Angela Merkel	politician

Fact Extraction

Person	**Nationality**
Angela Merkel	German

Ontological Information Extraction

Instance Extraction

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

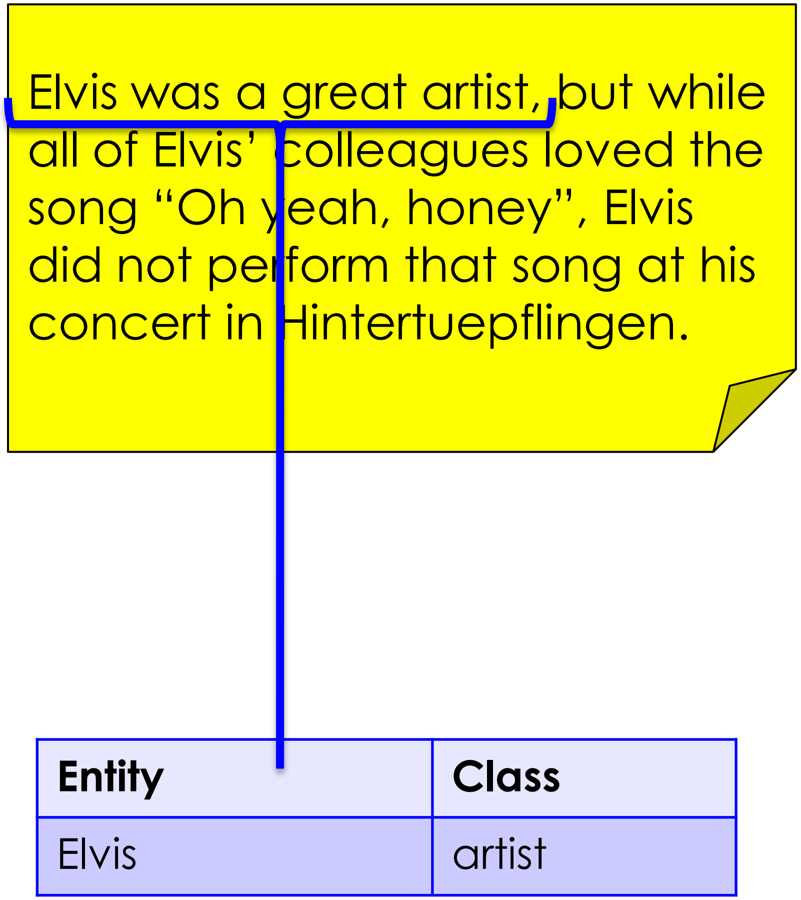
Elvis was a great artist, but while all of Elvis' colleagues loved the song "Oh yeah, honey", Elvis did not perform that song at his concert in Hintertuepflingen.

Entity	**Class**
Elvis	artist
Oh yeah, honey	song
Hintertuepflingen	location

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

Instance Extraction: 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 Germany.

Idea (by Hearst):

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

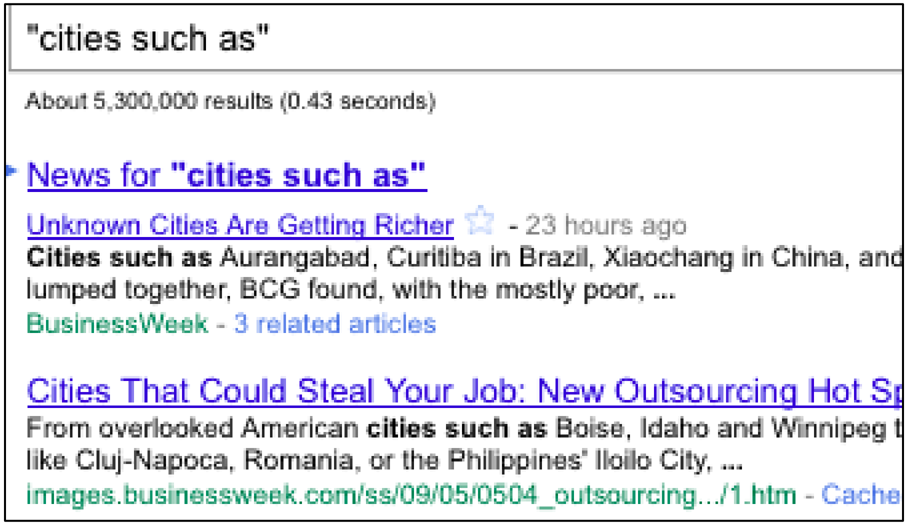
Hearst patterns:

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

Try this

Instance Extraction: Hearst Patterns

Hearst Patterns on Google



"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 Sp
From overlooked American **cities such as** Boise, Idaho and Winnipeg t
like Cluj-Napoca, Romania, or the Philippines' Iloilo City, ...
images.businessweek.com/ss/09/05/0504_outsourcing.../1.htm - Cache

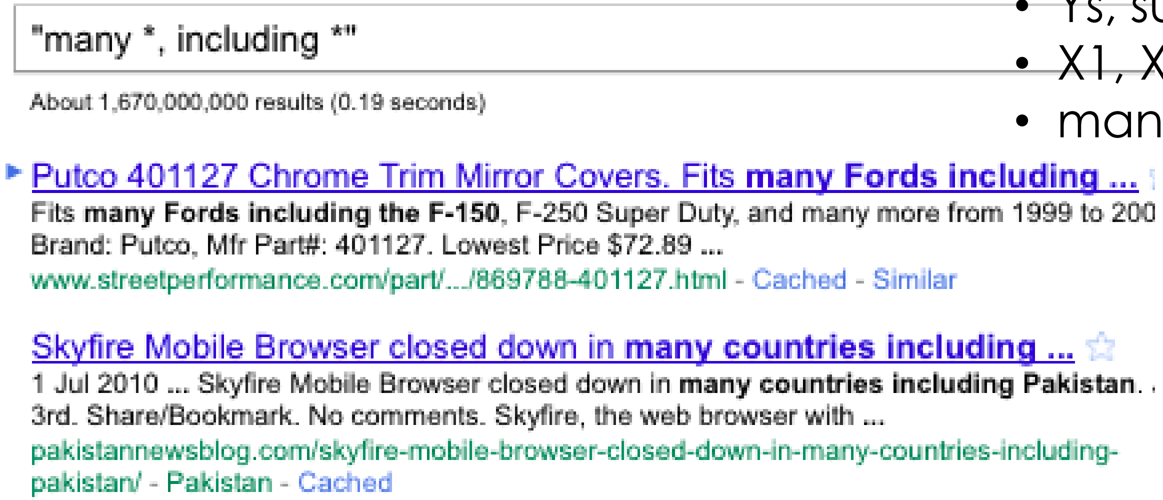
Idea (by Hearst):

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

Hearst patterns:

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

Wildcards on Google



"many *, including *"

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

Putco 401127 Chrome Trim Mirror Covers. Fits **many Fords including ...**
Fits **many Fords including the F-150**, F-250 Super Duty, and many more from 1999 to 200
Brand: Putco, Mfr Part#: 401127. Lowest Price $72.89 ...
www.streetperformance.com/part/.../869788-401127.html - Cached - Similar

Skyfire Mobile Browser closed down in **many countries including ...**
1 Jul 2010 ... Skyfire Mobile Browser closed down in **many countries including Pakistan**. .
3rd. Share/Bookmark. No comments. Skyfire, the web browser with ...
pakistannewsblog.com/skyfire-mobile-browser-closed-down-in-many-countries-including-pakistan/ - Pakistan - Cached

Instance Extraction: Hearst Patterns

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

Instance Extraction: POS

noun

Elvis is a great rock star who is adored by everybody.

X is a Y

Elvis is a great

Elvis is a great rock

➡ Elvis is a great rock star

Elvis is a great rock star who

?

Instance Extraction: POS

noun

Elvis is a great rock star who is adored by everybody.

The **Part-of-Speech** (POS) of a word in a sentence is the grammatical role that this word takes.

Open Part-of-Speech classes:

- Proper nouns: Alice, Fabian, Elvis, ...
- Nouns: computer, weekend, ...
- Adjectives: self-reloading fridge, ...
- Verbs: download, ...

Closes Part-of-Speech classes:

- Pronouns: he, she, it, this, ... (≈ what can replace a noun)
- Determiners: the, a, these, your, my, ... (≈ what goes before a noun)
- Prepositions: in, with, on, ... (≈ what goes before determiner + noun)
- Subordinators: who, whose, that, which, because, ...
  (≈ what introduces a sub-ordinate sentence)

Instance Extraction: POS

noun

Elvis is a great rock star who is adored by everybody.

Elvis/ProperNoun is/Verb a/Det great/Adj rock/Noun star/noun who/Sub is/verb adored/Verb …

POS tagging is the process of, given a sentence, determining the part of speech of each word.

Instance Extraction: POS

POS tagging is the process of, given a sentence, determining the part of speech of each word.

POS tagging is not trivial, because the same word can appear with different POS:

- Some words belong to two word classes ("run" as a verb or noun)
- Some word forms may be ambiguous:

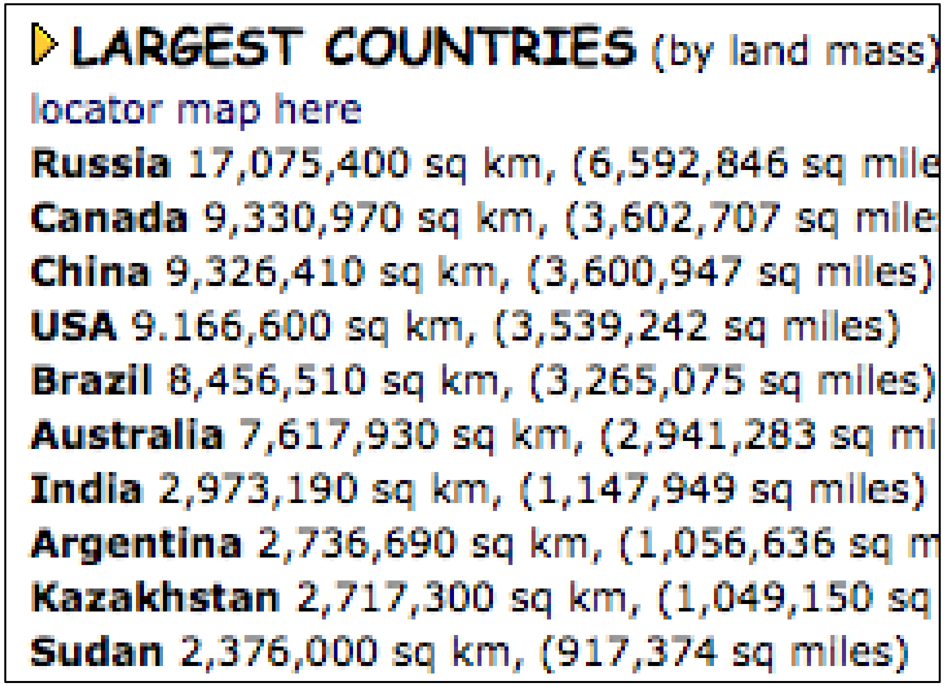
  Sound sounds sound sound.

Common techniques:

- rule-based
- statistical
- using dynamic programming

Instance Extraction: 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 mile

Canada 9,330,970 sq km, (3,602,707 sq mile

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 mi

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

Argentina 2,736,690 sq km, (1,056,636 sq m

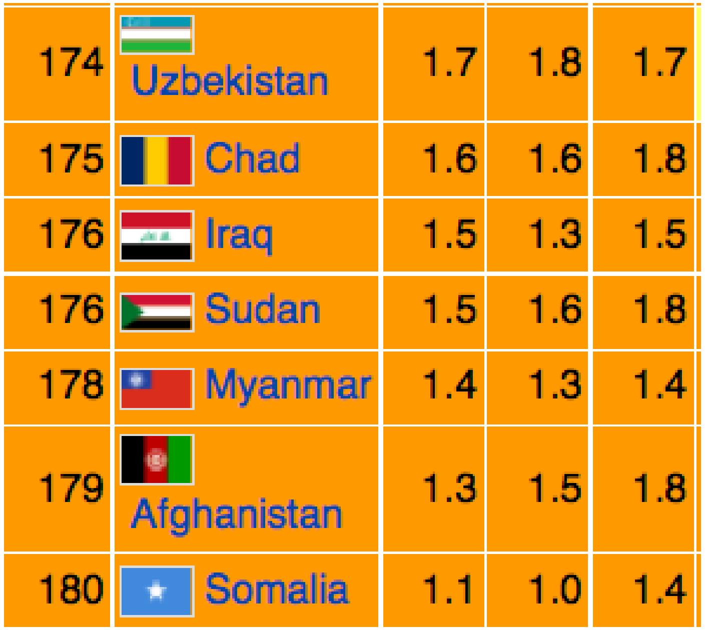
Kazakhstan 2,717,300 sq km, (1,049,150 sq

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

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

Instance Extraction: Set Expansion

Most corrupt countries



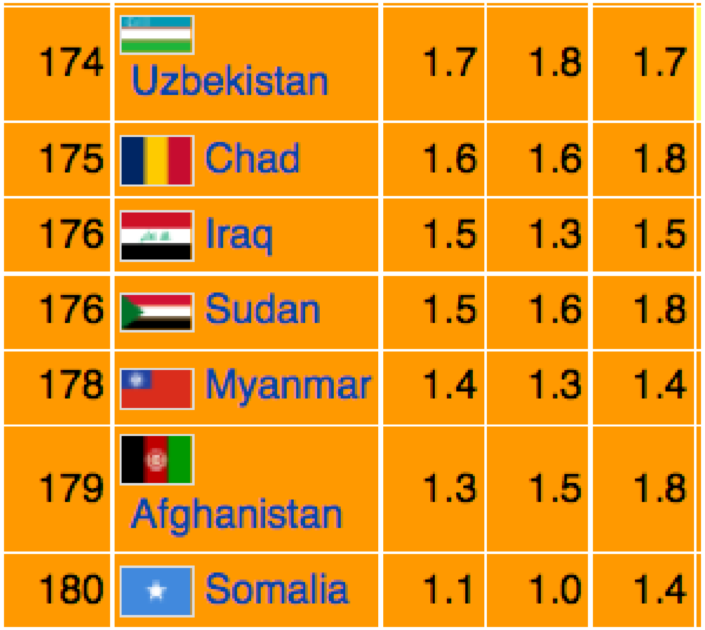
174	Uzbekistan	1.7	1.8	1.7
175	Chad	1.6	1.6	1.8
176	Iraq	1.5	1.3	1.5
176	Sudan	1.5	1.6	1.8
178	Myanmar	1.4	1.3	1.4
179	Afghanistan	1.3	1.5	1.8
180	Somalia	1.1	1.0	1.4

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

Instance Extraction: Set Expansion

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

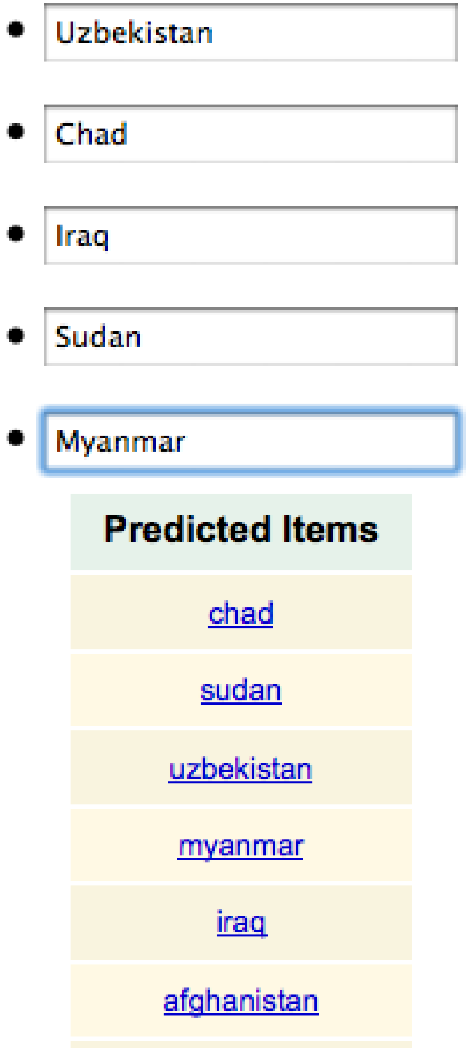
Most corrupt countries



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

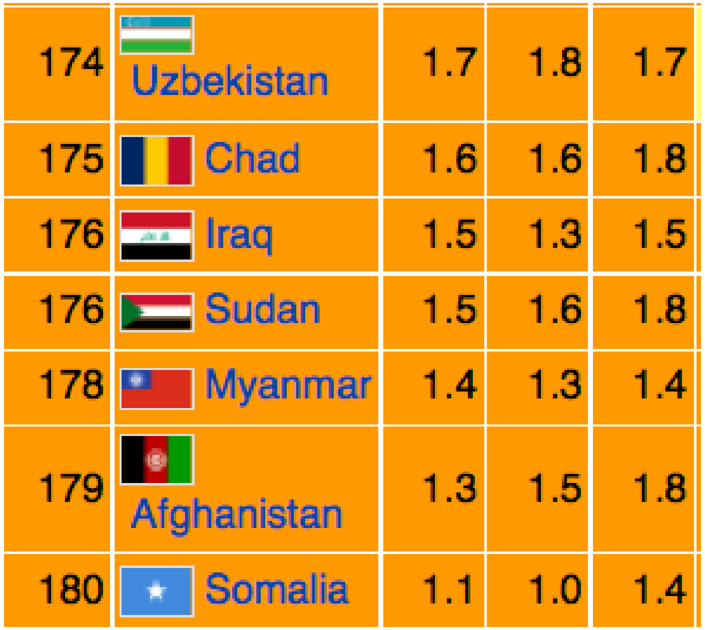
uzbekistan

myanmar

iraq

afghanistan

Instance Extraction: Set Expansion



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

Set Expansion can extract instances from tables or lists.

Input:
- seed pairs

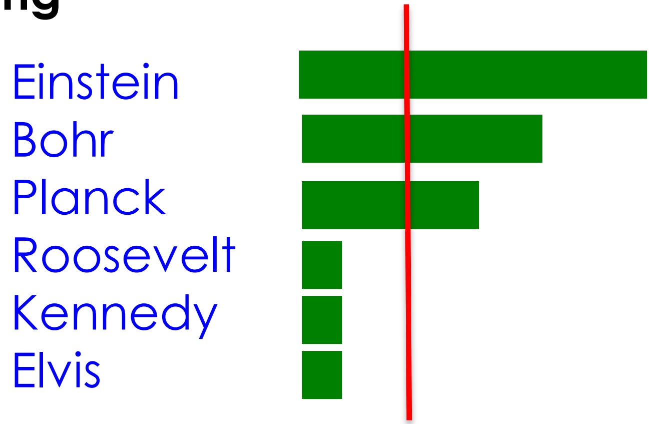
Condition:
- a corpus full of tables

Instance Extraction: Cleaning

Information Extraction nearly always produces **noise** (minor false outputs)

Approaches:

- **Thresholding**



(number of times extracted)

- **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), ...

Instance Extraction: 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 = {Einstein, Bohr, Planck, Clinton, Obama}

✓ ✓ ✓ ✗ ✗

Gold standard:
G = {Einstein, Bohr, Planck, Heisenberg}

✓ ✓ ✓ ✗

Precision:
What proportion of the output is correct?

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

Recall:
What proportion of the gold standard did we get?

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

Instance Extraction: Evaluation

Explorative algorithms extract everything they find.

(very low threshold)

Algorithm output:
O = {Einstein, Bohr, Planck, Clinton, Obama, Elvis, Heisenberg, ...}

Gold standard:
G = {Einstein, Bohr, Planck, Heisenberg}

Precision:
What proportion of the output is correct?

BAD

Recall:
What proportion of the gold standard did we get?

GREAT

Instance Extraction: Evaluation

Conservative algorithms extract only things about which they are very certain

(very high threshold)

Algorithm output:
O = {Einstein}

Gold standard:
G = {Einstein, Bohr, Planck, Heisenberg}

Precision:
What proportion of the output is correct?

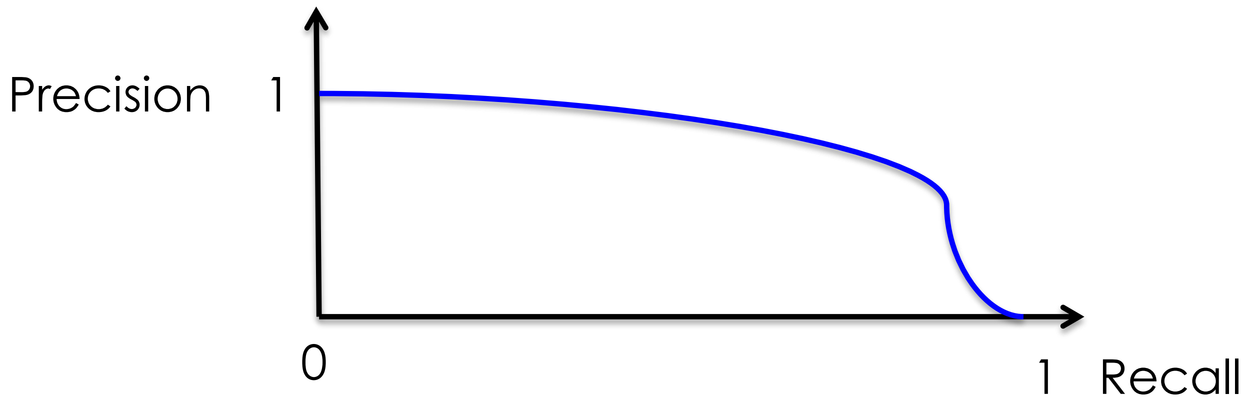
GREAT

Recall:
What proportion of the gold standard did we get?

BAD

Instance Extraction: Evaluation

You can't get it all...



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

F1 = 2 * precision * recall / (precision + recall)

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)
- Set Expansion (for tables and lists)
- ...many others...

On top of that:

- Iteration
- Cleaning
- POS-tagging

And finally:

- Evaluation

Information Extraction

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

✓

Named Entity Recognition

...married Elvis
on 1967-05-01

✓

Instance Extraction

Elvis Presley	singer
Angela Merkel	politician

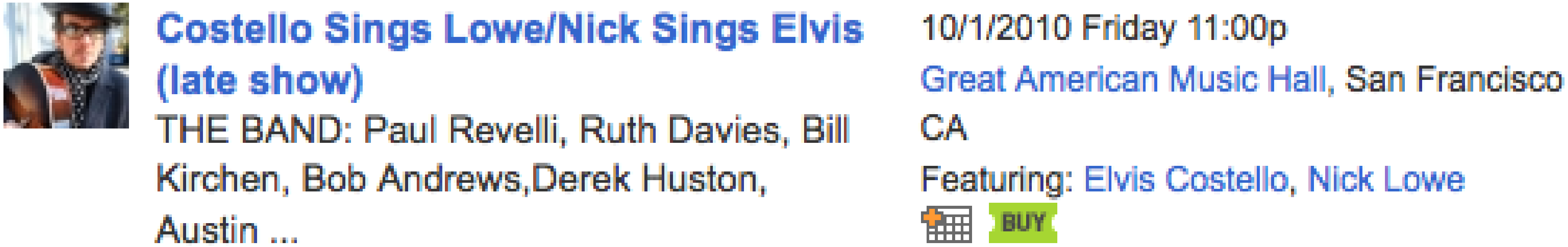
Fact Extraction

Person	Nationality
Angela Merkel	German

Ontological Information Extraction

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

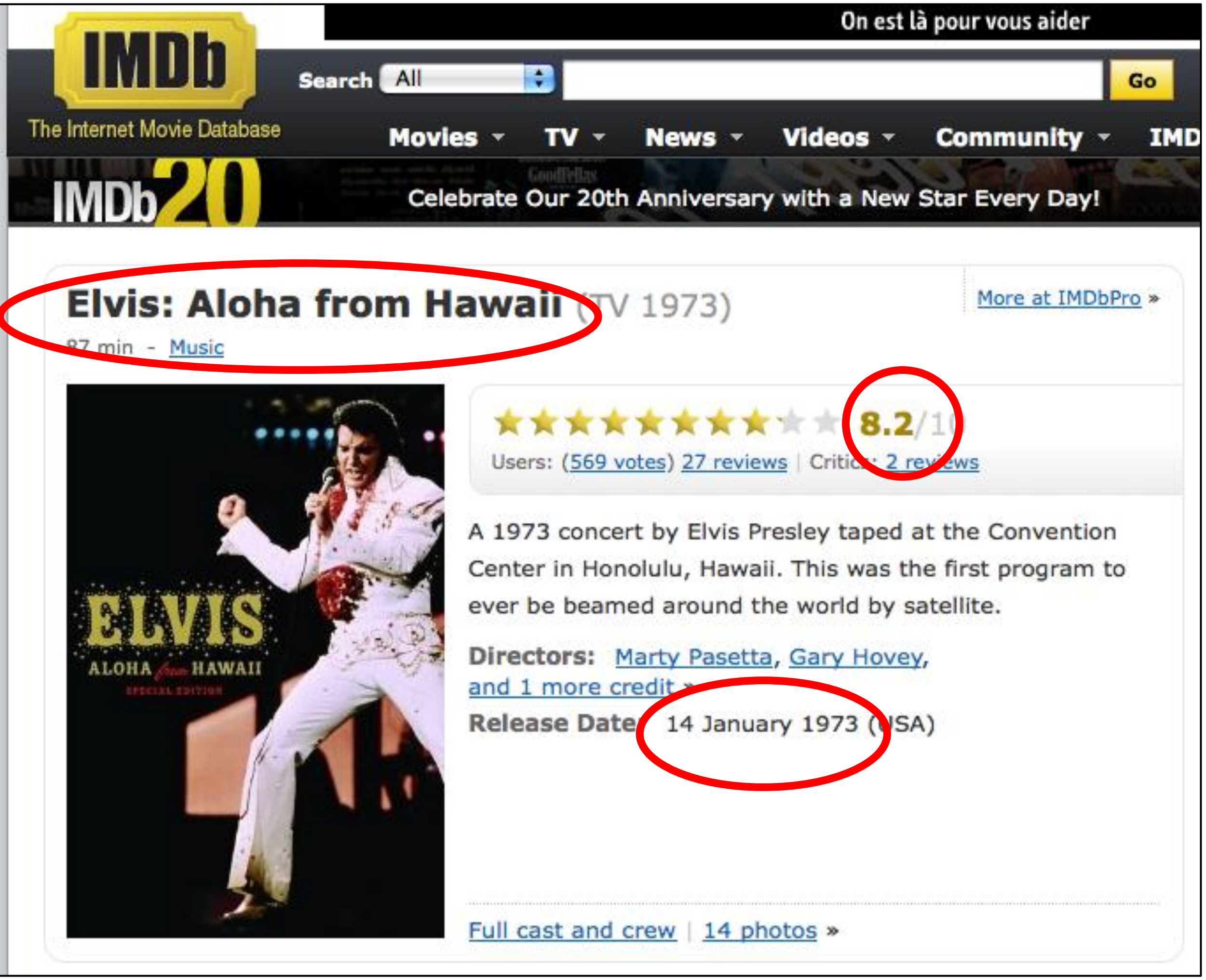
Great American Music Hall, San Francisco CA

Featuring: Elvis Costello, Nick Lowe

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

Fact Extraction: 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
IMD
IMDb 20
Celebrate Our 20th Anniversary with a New Star Every Day!
More at IMDbPro »
ELVIS
ALOHA from HAWAII
SPECIAL EDITION

Elvis: Aloha from Hawaii (TV 1973)

87 min - Music

8.2/10

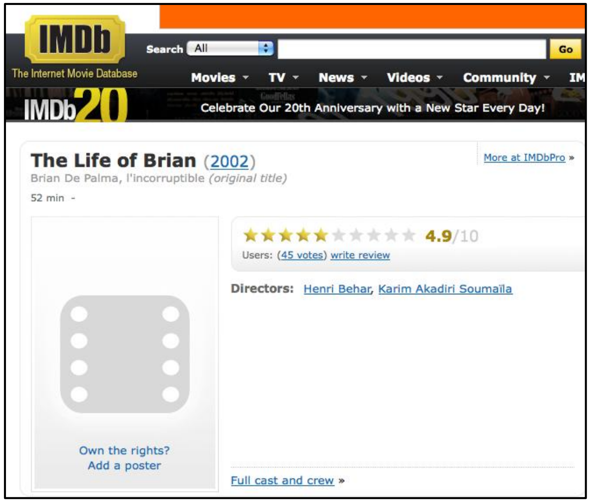
Users: (569 votes) 27 reviews | Critics: 2 reviews

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 »



The Life of Brian (2002)

Brian De Palma, l'incorruptible *(original title)*

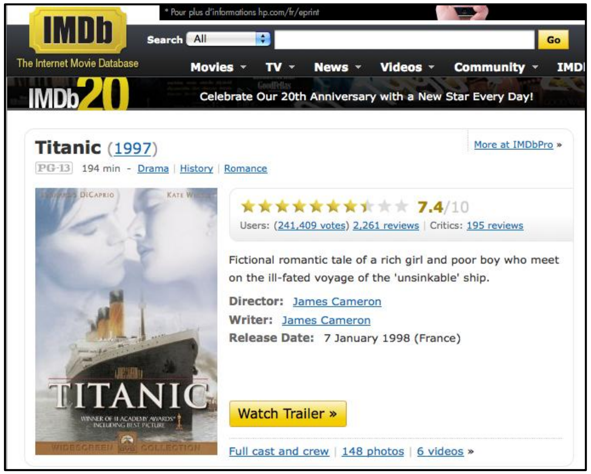
52 min -

4.9/10

Users: (45 votes) write review

Directors: Henri Behar, Karim Akadiri Soumaïla

Full cast and crew »


* Pour plus d'informations hp.com/fr/eprint
IMDb
The Internet Movie Database
Search
All
Go
Movies
TV
News
Videos
Community
IMDI
IMDb 20
Celebrate Our 20th Anniversary with a New Star Every Day!
More at IMDbPro »
DICAPRIO
KATE W
TITANIC
WINNER OF 11 ACADEMY AWARDS INCLUDING BEST PICTURE
WIDESCREEN COLLECTION
Watch Trailer »

Titanic (1997)

PG-13 194 min - Drama | History | Romance

7.4/10

Users: (241,409 votes) 2,261 reviews | Critics: 195 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

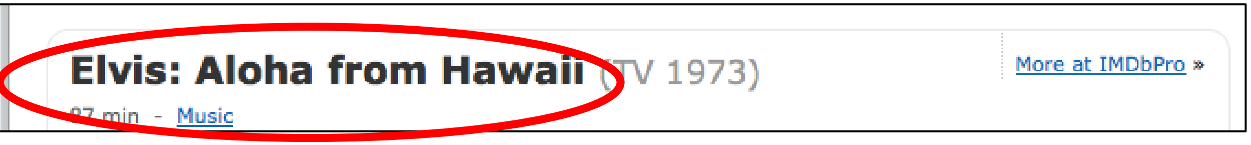
Release Date: 7 January 1998 (France)

Full cast and crew | 148 photos | 6 videos »

Fact Extraction: 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 or multiple spots on a page is called a **wrapper**.



Elvis: Aloha from Hawaii (TV 1973)

97 min - Music

More at IMDbPro »

```
<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]

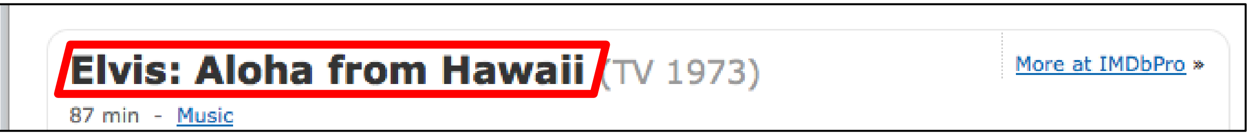
It can also be a search text/regex

>.*</b>(TV

Fact Extraction: Wrapper Induction

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

title



Elvis: Aloha from Hawaii (TV 1973)

87 min - Music

More at IMDbPro »

Try it out

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

Title:
div[1] → div[2]

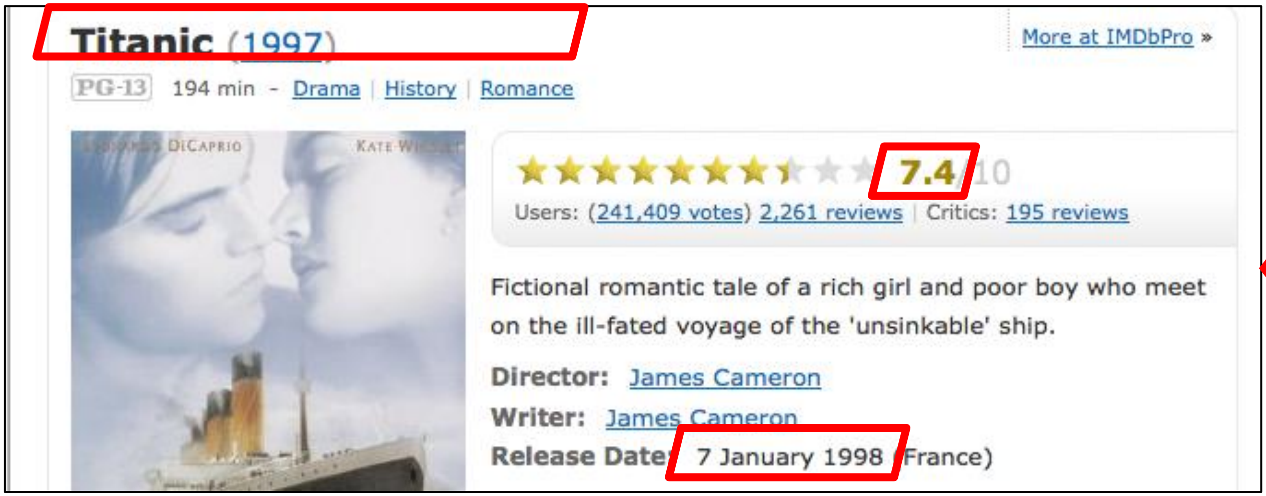
Rating:
div[7] → span[2] → b[1]

ReleaseDate:
div[10] → i[1]

Fact Extraction: Wrapper Induction

We manually label the fields to be extracted, and produce the corresponding wrappers.

Then we **apply** the wrappers to all pages in the domain (i.e., we determine the spots of the pages that the wrappers point to).



Titanic (1997)

More at IMDbPro »

PG-13 194 min - Drama | History | Romance

7.4/10

Users: (241,409 votes) 2,261 reviews | Critics: 195 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)

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

Fact Extraction: Wrapper Induction

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

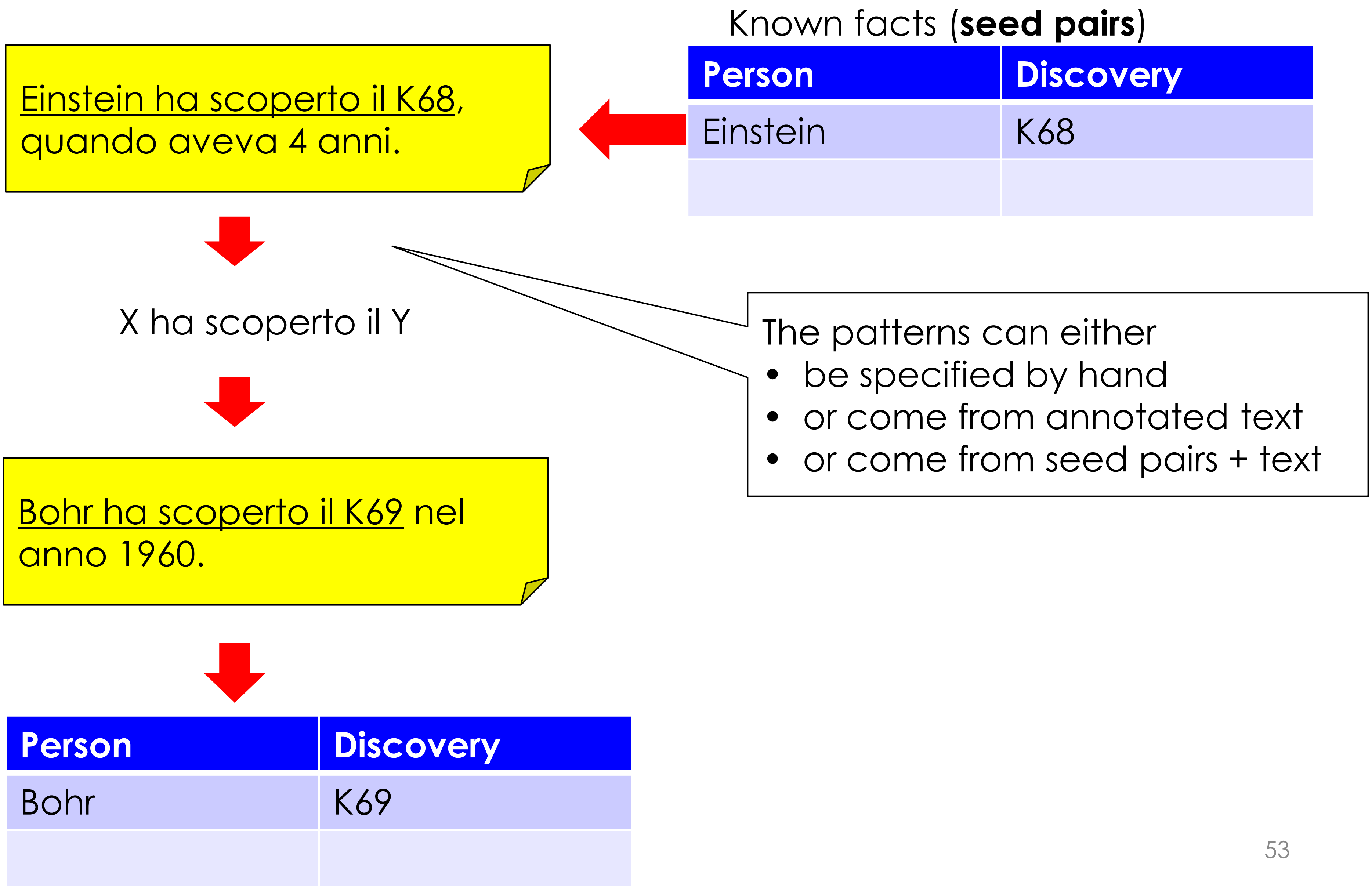
Condition:

- 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?

Fact Extraction: Pattern Matching



Known facts (**seed pairs**)

Person	Discovery
Einstein	K68

Einstein ha scoperto il K68, quando aveva 4 anni.

X ha scoperto il Y

The patterns can either

- be specified by hand
- or come from annotated text
- or come from seed pairs + text

Bohr ha scoperto il K69 nel anno 1960.

Person	Discovery
Bohr	K69

Fact Extraction: Pattern Matching

Einstein ha scoperto il K68, quando aveva 4 anni.

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

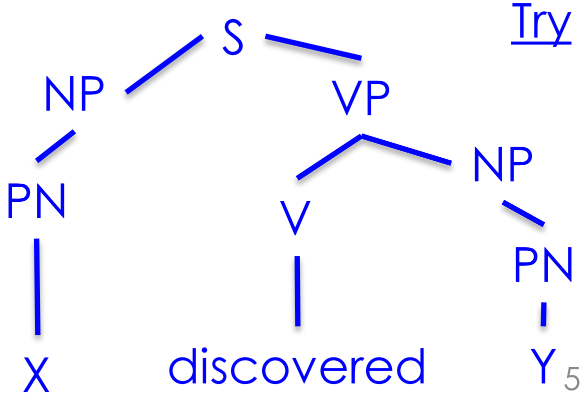
  X discovered the .{0,20} Y

- POS patterns

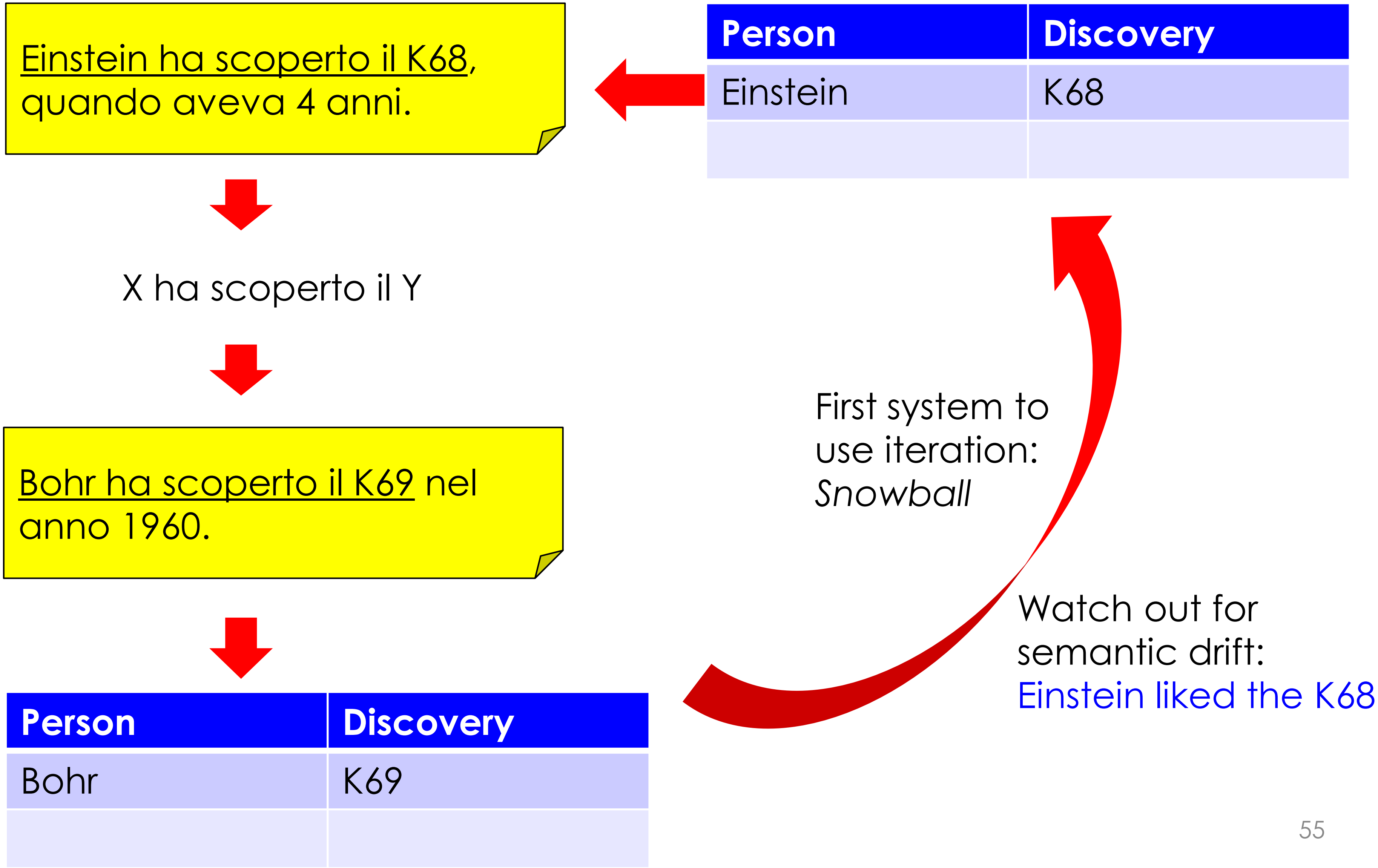
  X discovered the ADJ? Y

- Parse trees

Try



Fact Extraction: Pattern Matching



Einstein ha scoperto il K68, quando aveva 4 anni.

Person	Discovery
Einstein	K68

X ha scoperto il Y

Bohr ha scoperto il K69 nel anno 1960.

Person	Discovery
Bohr	K69

First system to use iteration: *Snowball*

Watch out for semantic drift: Einstein liked the K68

Fact Extraction: Pattern Matching

Einstein ha scoperto il K68, quando aveva 4 anni.

Pattern matching can extract facts from natural language text corpora.

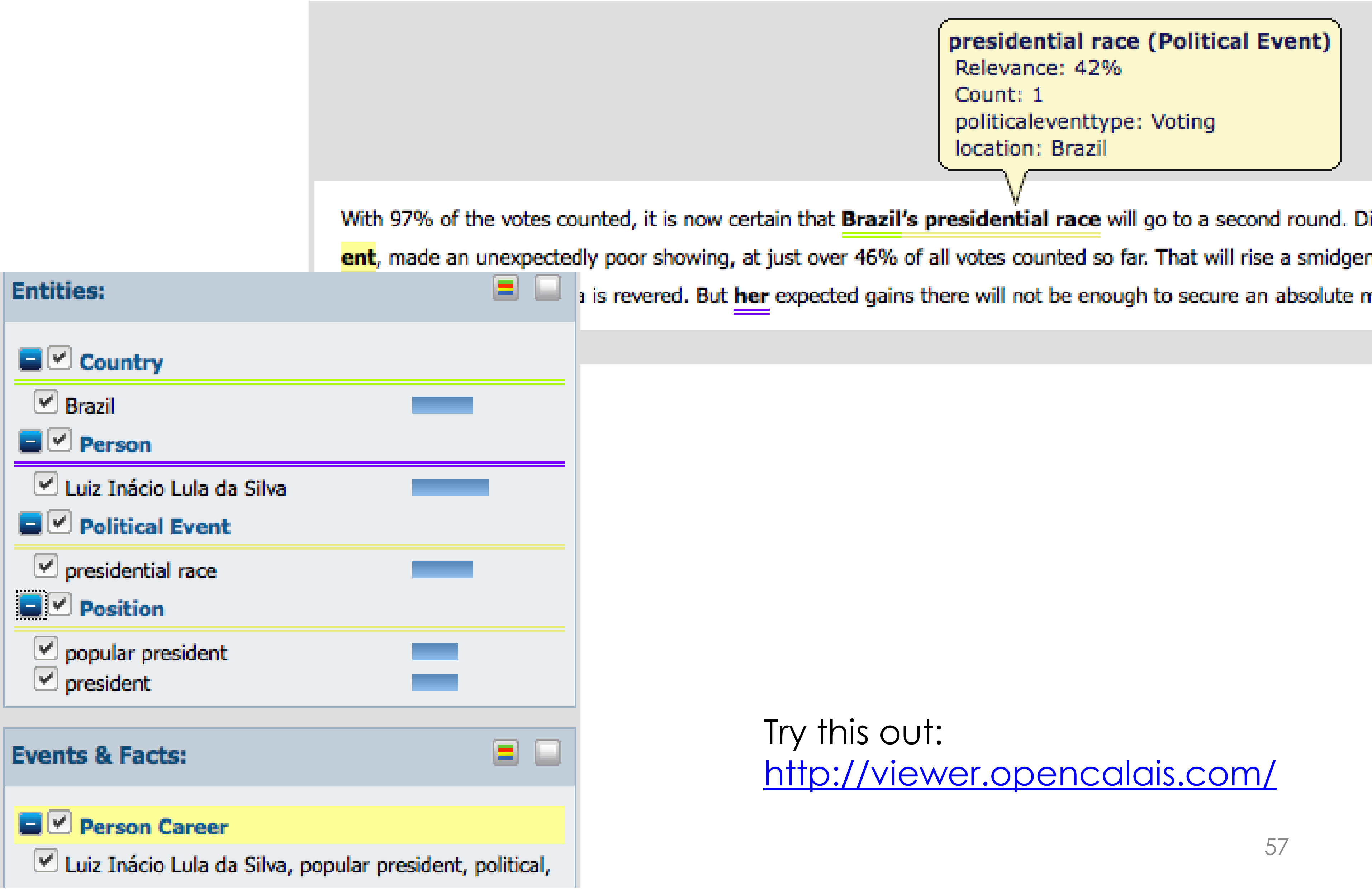
Input:

- 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

Fact Extraction: Pattern Matching



presidential race (Political Event)
Relevance: 42%
Count: 1
politicaleventtype: Voting
location: Brazil

With 97% of the votes counted, it is now certain that **Brazil's presidential race** will go to a second round. Di
ent, made an unexpectedly poor showing, at just over 46% of all votes counted so far. That will rise a smidger
a is revered. But **her** 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/

Fact Extraction: Cleaning

Fact Extraction commonly produces huge amounts of garbage.

Web page contains bogus information

Web page contains misleading items (advertisements, error messages)

Deviation in iteration

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

Formatting problems (bad HTML, character encoding mess)

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

Something has changed over time (facts or page formatting)

⇒ 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:

- 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 (e.g., database tables)
from unstructured machine-readable documents
(e.g., Web documents).

Ontological Information Extraction

✓

Fact Extraction

Person	Nationality
Angela Merkel	German

✓

Instance Extraction

Elvis Presley	singer
Angela Merkel	politician

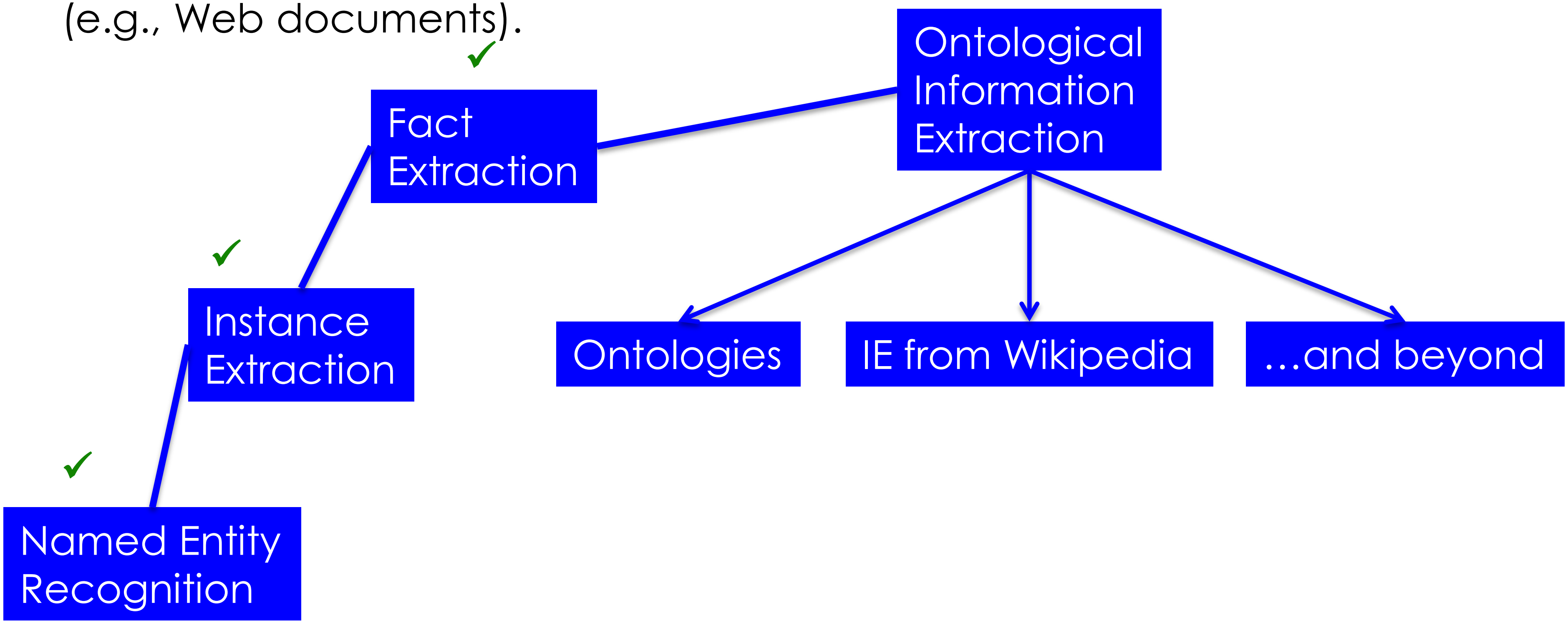
✓

Named Entity Recognition

...married Elvis
on 1967-05-01

Information Extraction

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



✓ Fact Extraction

✓ Instance Extraction

✓ Named Entity Recognition

Ontological Information Extraction

Ontologies

IE from Wikipedia

…and beyond

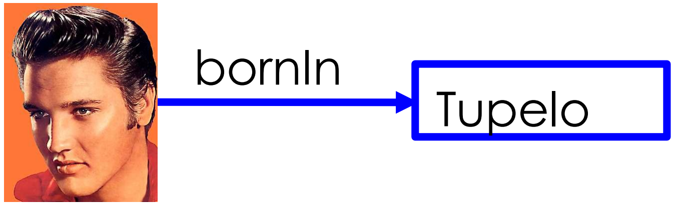
Ontologies: RDF

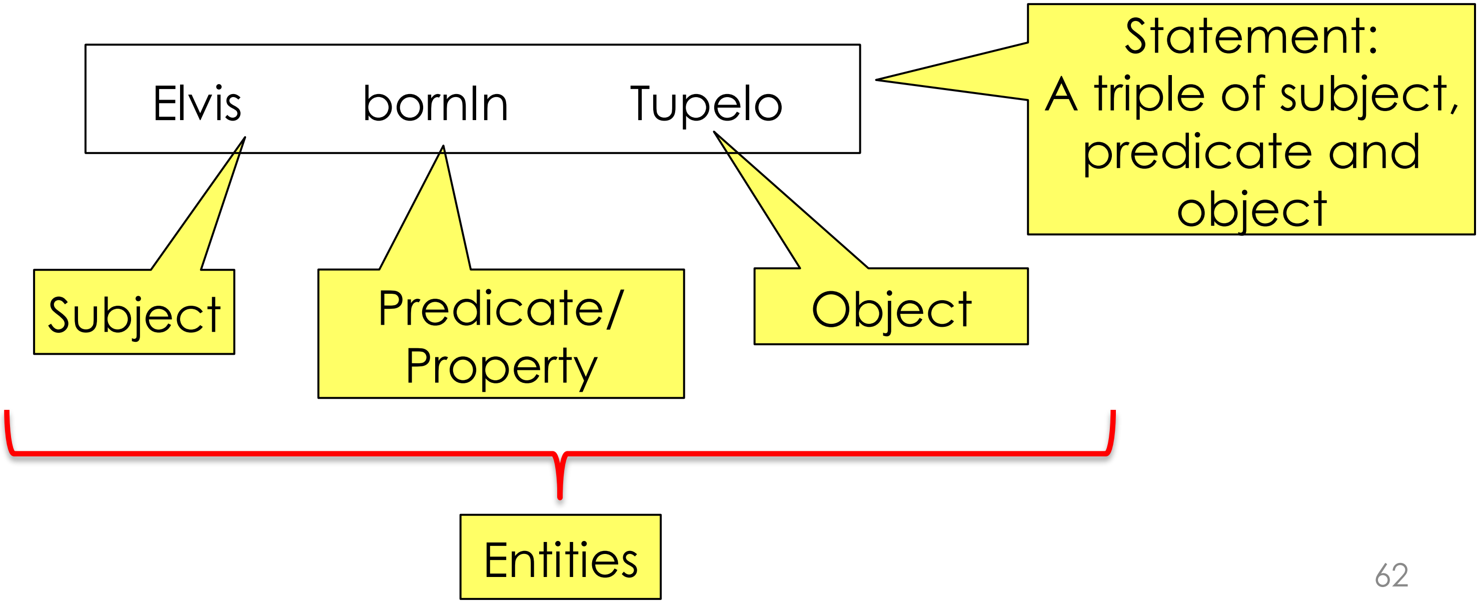
An **ontology** is a structured collection of world knowledge.

(Here, we are concerned mainly with RDF ontologies. RDF is a W3C endorsed standard)

RDF(the Resource Description format) is a format of knowledge representation that is similar to the Entity-Relationship-Model.

“Elvis was born in Tupelo”





Elvis bornIn Tupelo

Statement:
A triple of subject, predicate and object

Subject

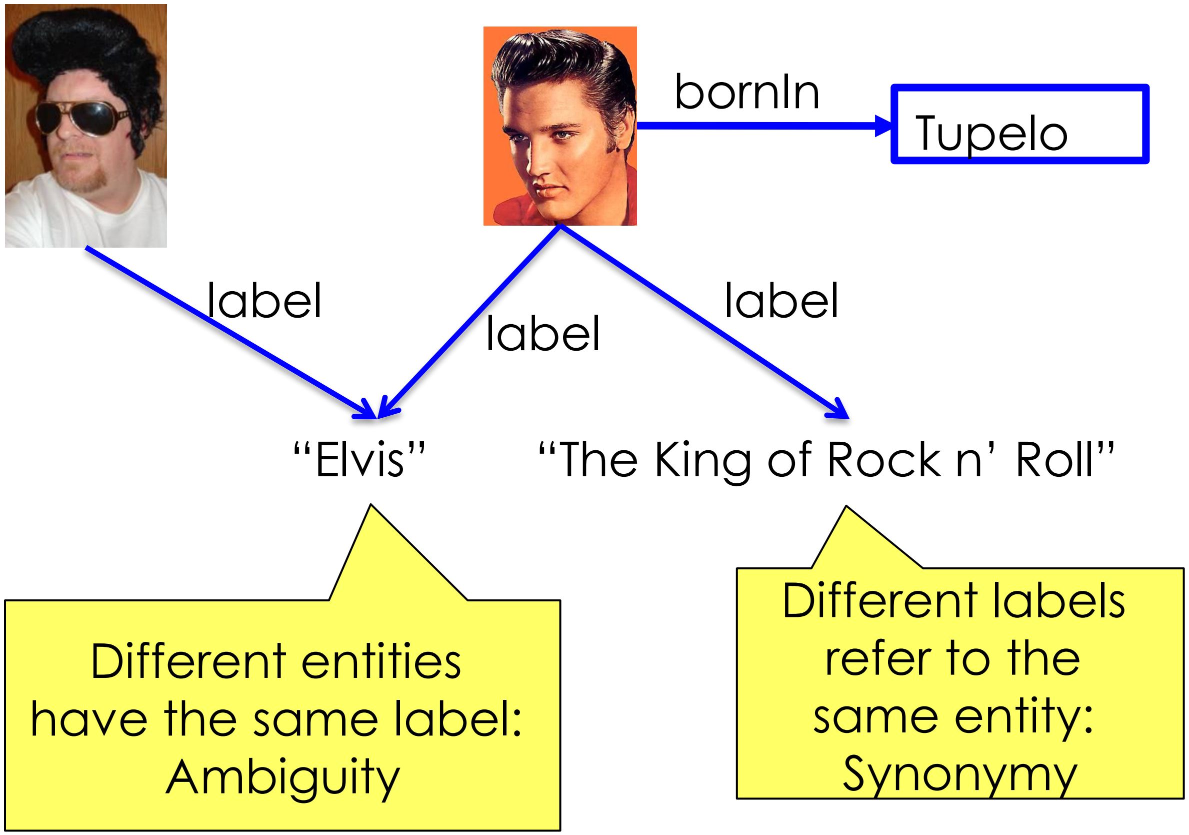
Predicate/ Property

Object

Entities

Ontologies: Labels

An ontology distinguishes between the **entity** and its **label**.



bornIn

Tupelo

label

label

label

“Elvis”

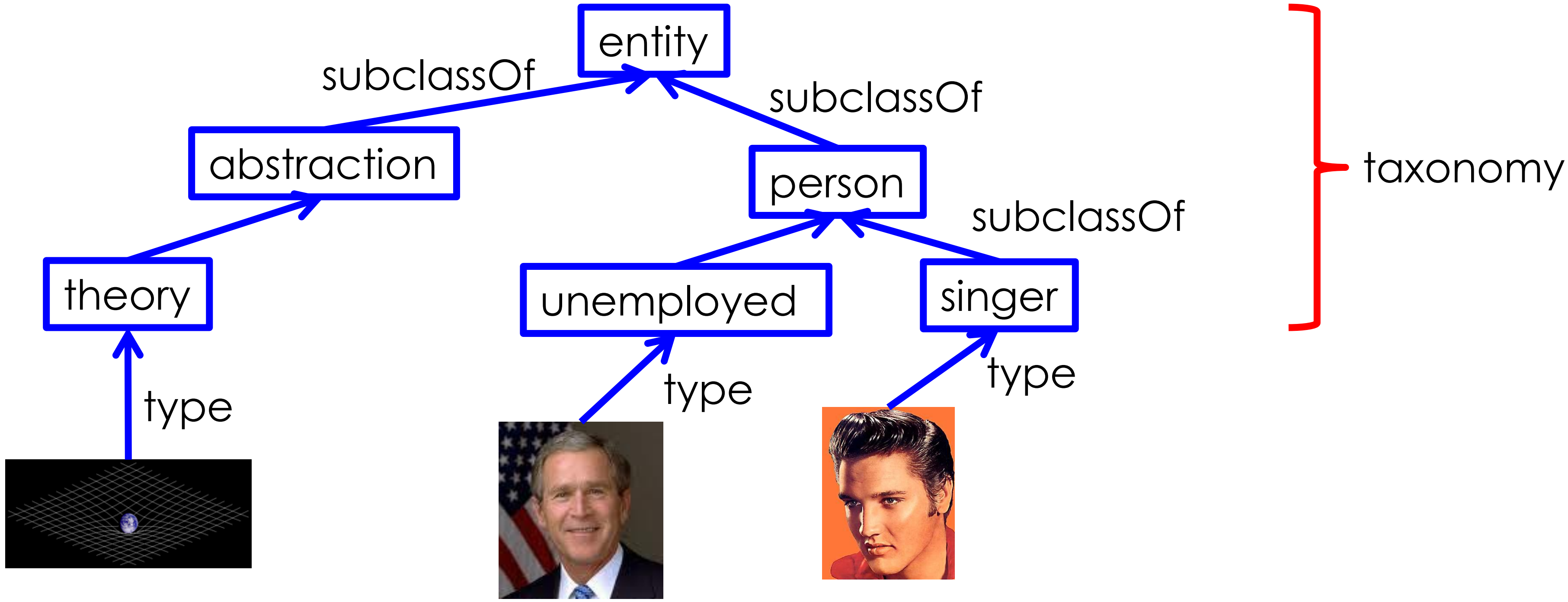
“The King of Rock n’ Roll”

Different entities have the same label: Ambiguity

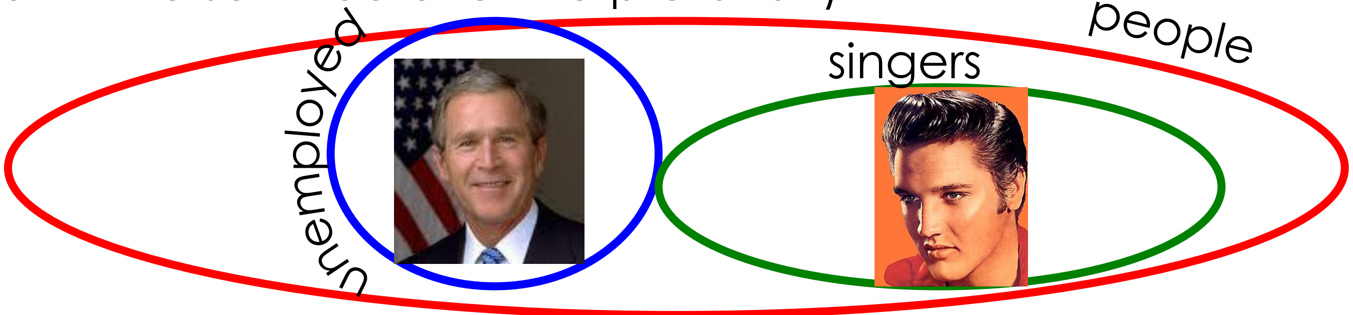
Different labels refer to the same entity: Synonymy

Ontologies: Classes

A **class** (also called concept) can be understood as a set of similar entities.

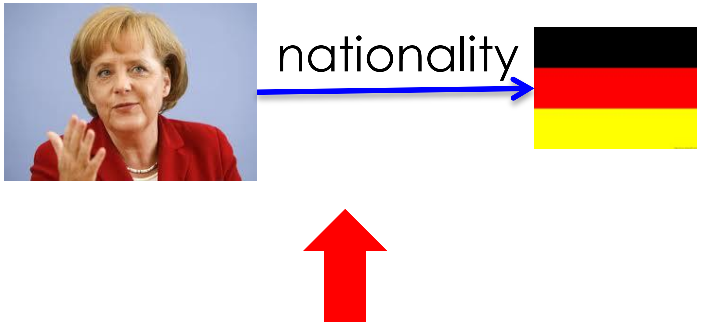


A **super-class** of a class is a class that is more general than the first class (a super-set in the set-theoretic interpretation)



Ontologies: IE

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



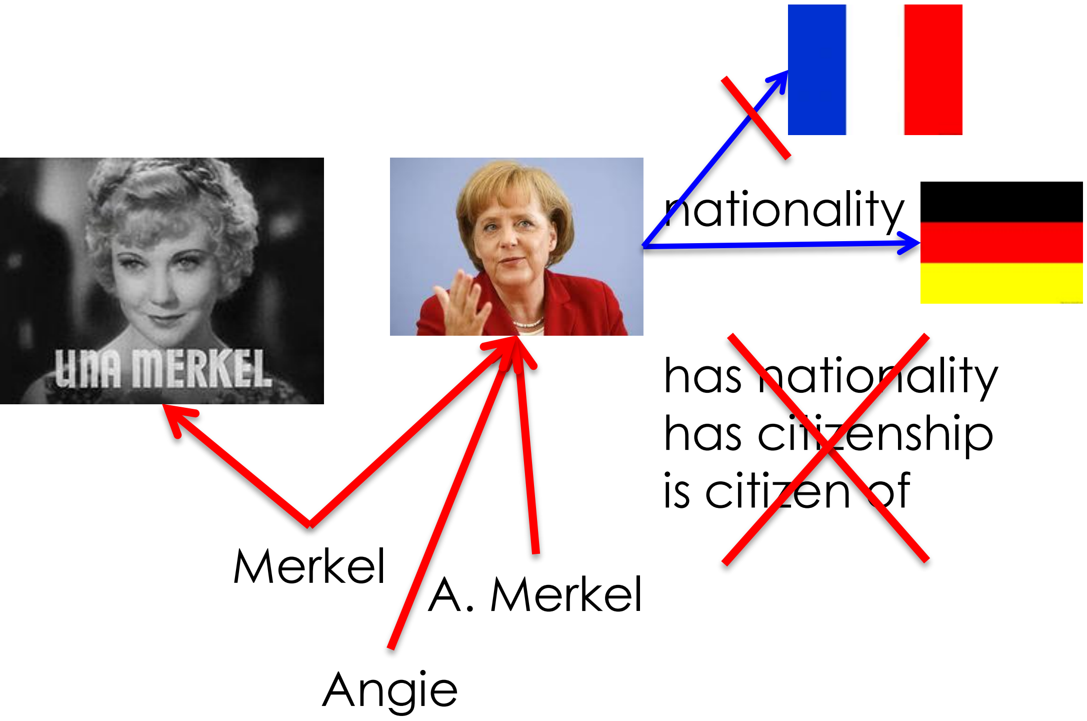
Angela Merkel is the German chancellor....
...Merkel was born in Germany...

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

Person	Nationality
Angela Merkel	German
Merkel	Germany
A. Merkel	French

Ontologies: IE

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



nationality

~~has nationality~~
~~has citizenship~~
~~is citizen of~~

Merkel

A. Merkel

Angie

Challenges:

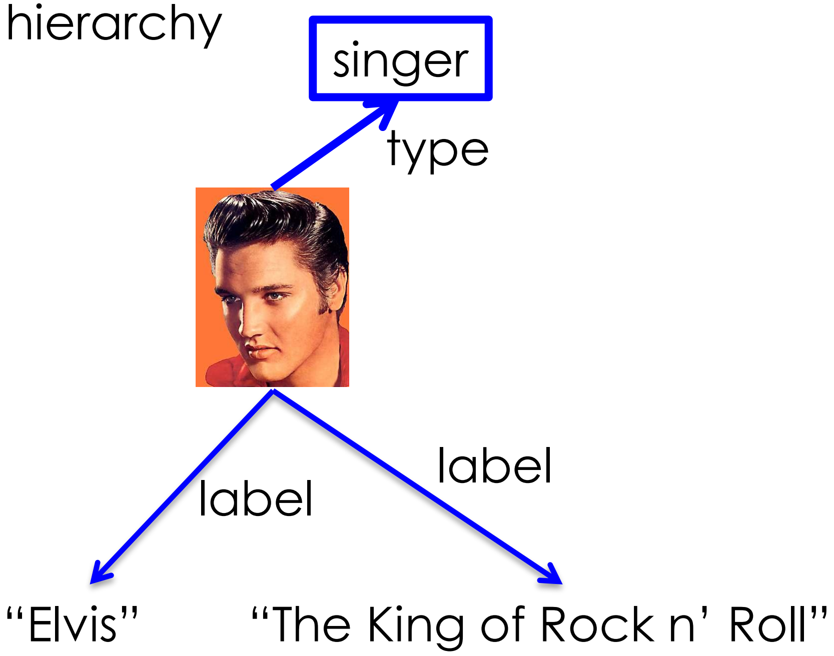
1. Map entity names to ontological entities
2. Disambiguate entity names
3. Use the relationships from the ontology
4. Make the ontology consistent

Ontologies

An **ontology** is a structured collection of world knowledge.

In an RDF ontology

- entities are unique
- entities can have different labels
- facts are represented as triples
- entities are arranged in a class hierarchy



singer

type

label

label

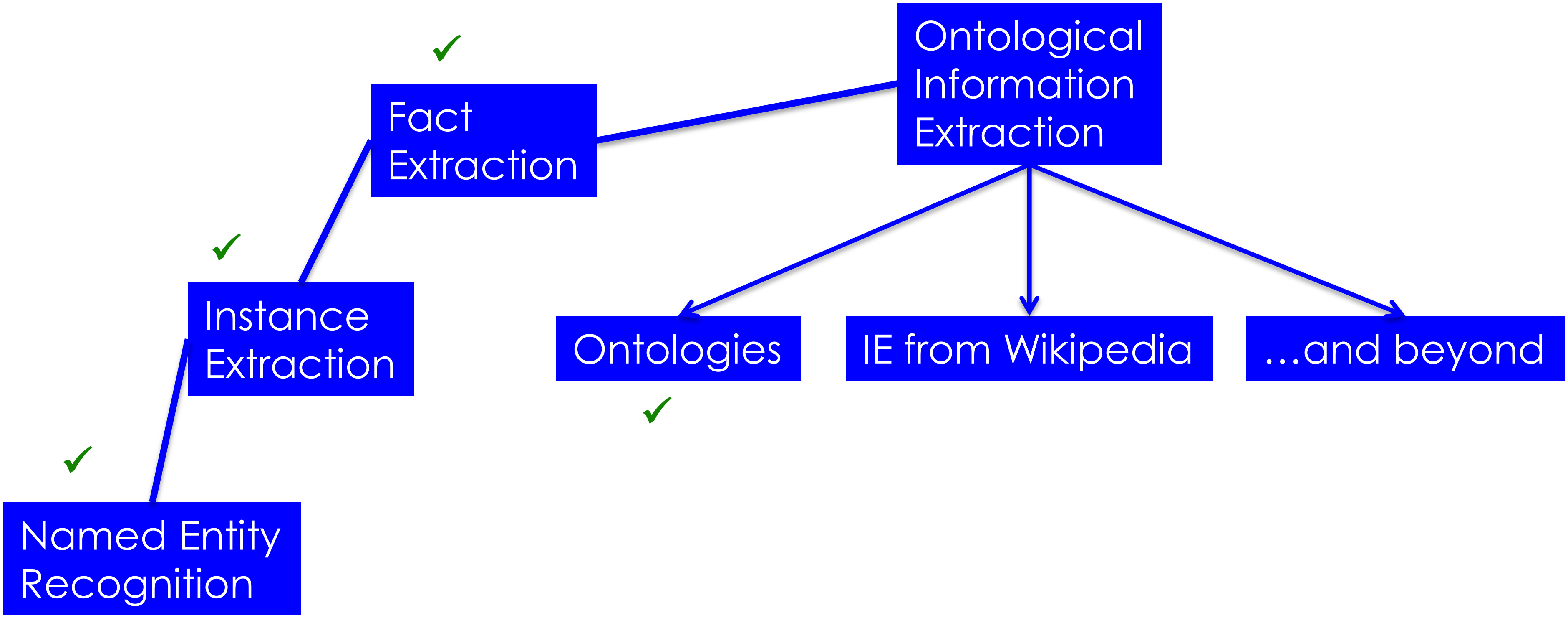
"Elvis"

"The King of Rock n' Roll"

…hence:
IE for ontologies is difficult.

Information Extraction

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



✓ Fact Extraction

Ontological Information Extraction

✓ Instance Extraction

Ontologies ✓

IE from Wikipedia

...and beyond

✓ Named Entity Recognition

IE from 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 and covers many different aspects (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,...)

IE from Wikipedia

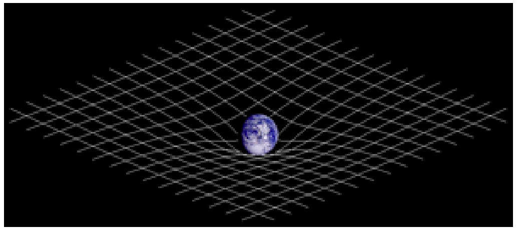
Wikipedia is a free online encyclopedia
- 3.4 million articles in English
- 16 million articles in dozens of languages

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

Angela_Merkel

Una_Merkel

Germany

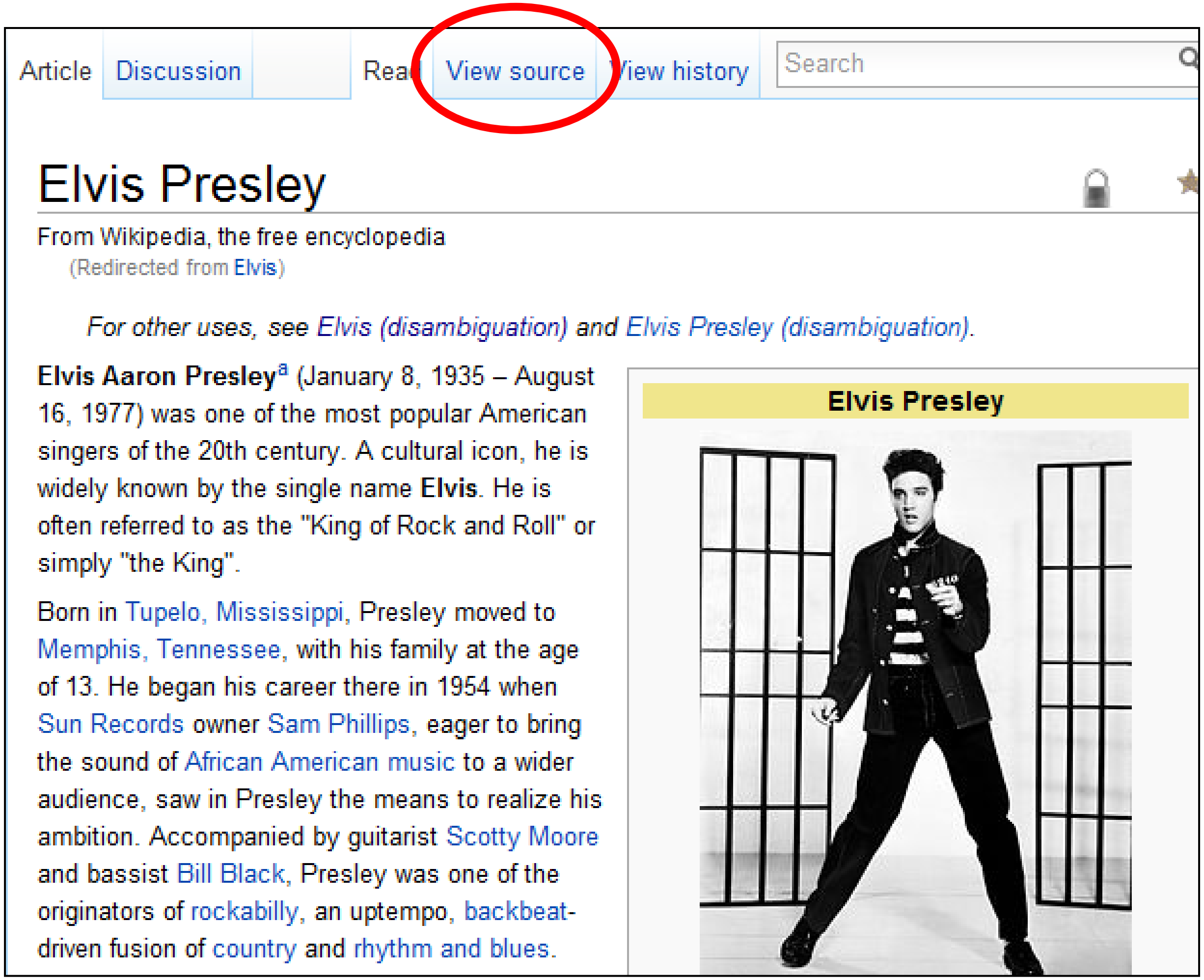


Theory_of_Relativity

IE from Wikipedia: Markup

Wikipedia uses the Wiki markup language

[Try this]



Elvis Presley

From Wikipedia, the free encyclopedia
(Redirected from Elvis)

For other uses, see Elvis (disambiguation) and Elvis Presley (disambiguation).

Elvis Aaron Presley[a] (January 8, 1935 – August 16, 1977) was one of the most popular American singers of the 20th century. A cultural icon, he is widely known by the single name **Elvis**. He is often referred to as the "King of Rock and Roll" or simply "the King".

Born in Tupelo, Mississippi, Presley moved to Memphis, Tennessee, with his family at the age of 13. He began his career there in 1954 when Sun Records owner Sam Phillips, eager to bring the sound of African American music to a wider audience, saw in Presley the means to realize his ambition. Accompanied by guitarist Scotty Moore and bassist Bill Black, Presley was one of the originators of rockabilly, an uptempo, backbeat-driven fusion of country and rhythm and blues.

Elvis Presley

IE from Wikipedia: Markup

Special formatting

Hyperlinks to other pages

Hyperlinks with alternative text

```
'''Elvis Aaron Presley'''([[January 8]], [[1935]] --
[[August 16]], [[1977]]), middle name sometimes written
'''Aron''', was an [[United States|American]] [[singer]],
[[musician]] and [[actor]]. …

{{Infobox musical artist
|Name              = Elvis Presley
|Img               = Elvis Presley 1970.jpg
|Born              = {{birth date|1935|1|8|}}
|Occupation        = [[singer]], [[actor]]
…
}}

[[Category:1935 births]] [[Category:1977 deaths]]…
```

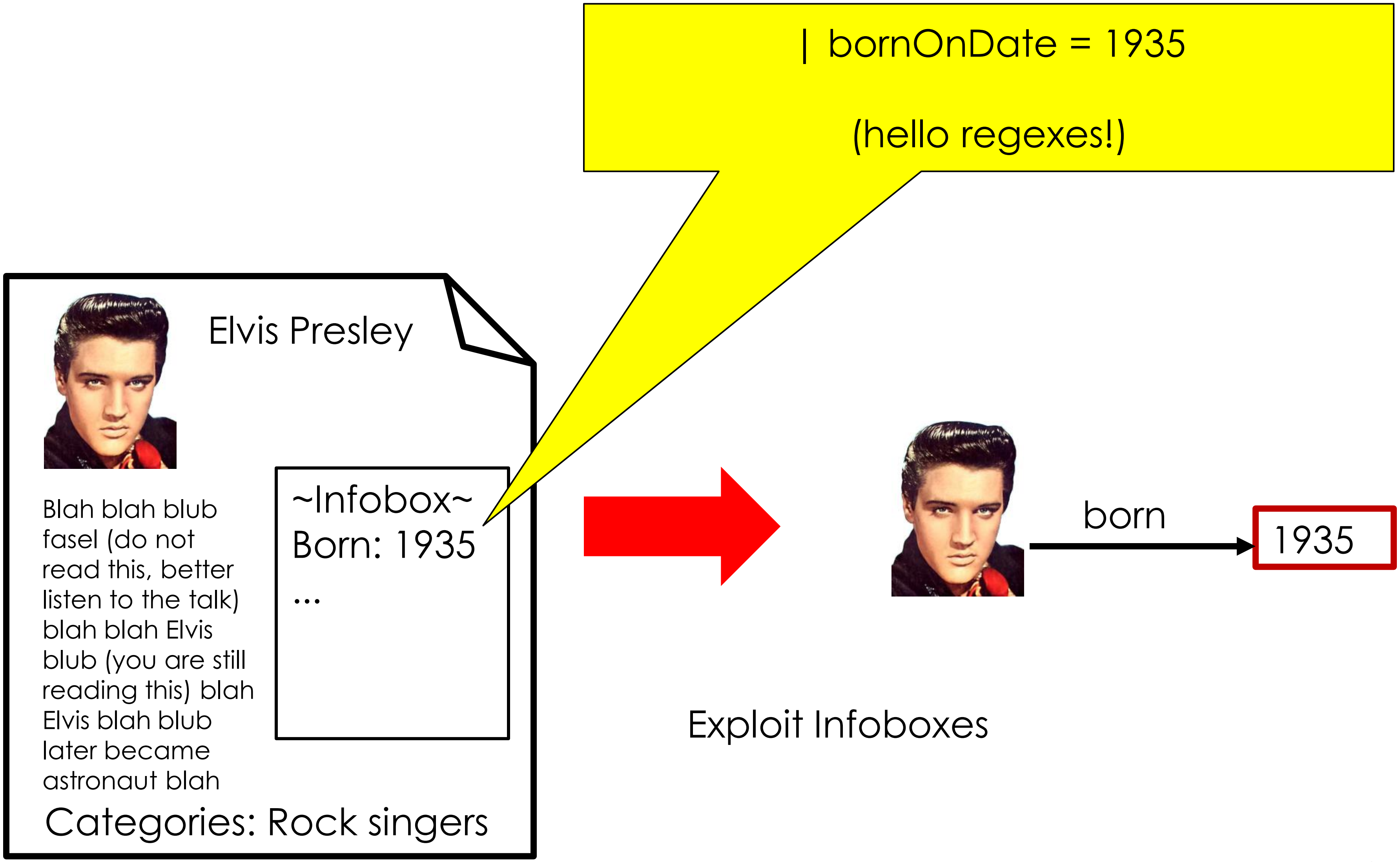
Infoboxes with type

Micro formats

Attribute value pairs

Categories

IE from Wikipedia: YAGO



| bornOnDate = 1935

(hello regexes!)

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

~Infobox~
Born: 1935
…

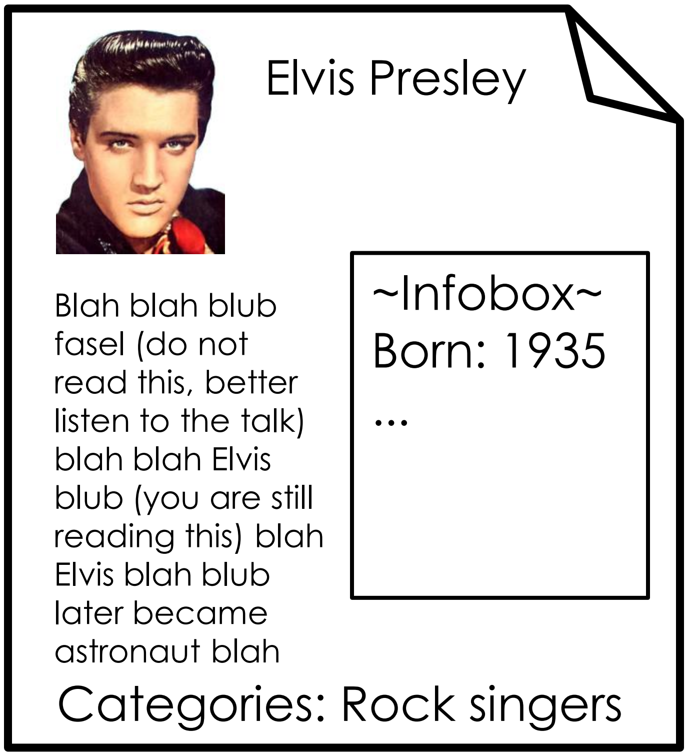
Categories: Rock singers

born

1935

Exploit Infoboxes

IE from Wikipedia: YAGO

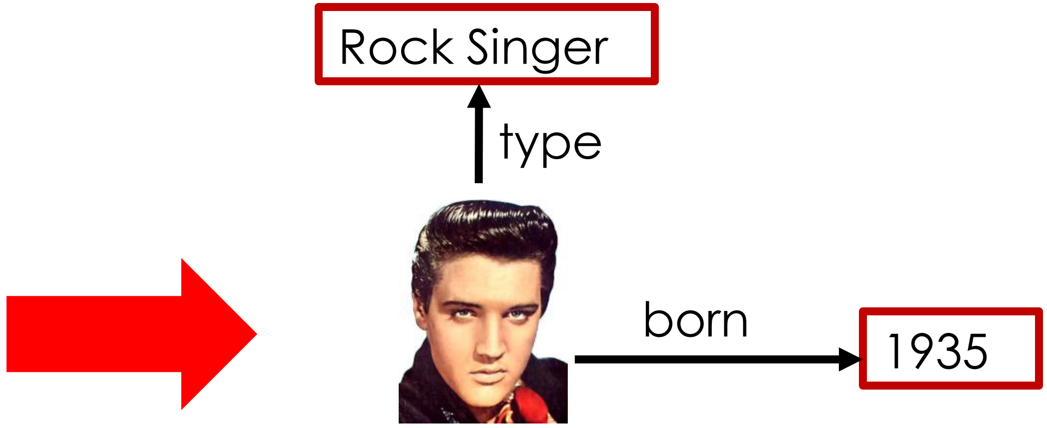


Elvis Presley

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

~Infobox~
Born: 1935
…

Categories: Rock singers



Rock Singer

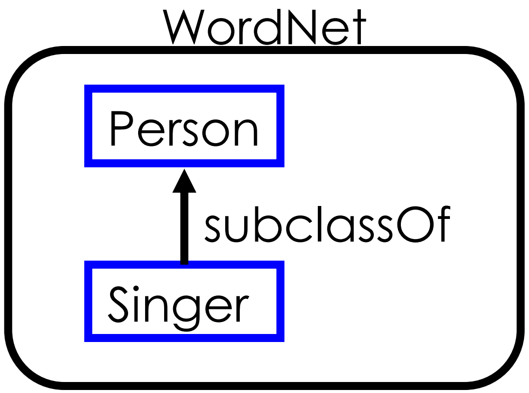
type

born

1935

Exploit Infoboxes
Exploit conceptual categories

IE from Wikipedia: YAGO

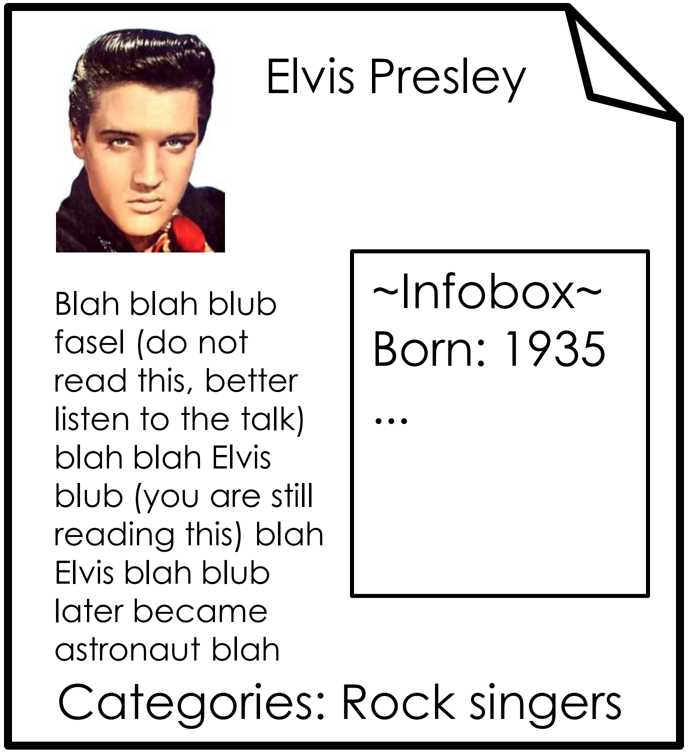


WordNet

Person

subclassOf

Singer

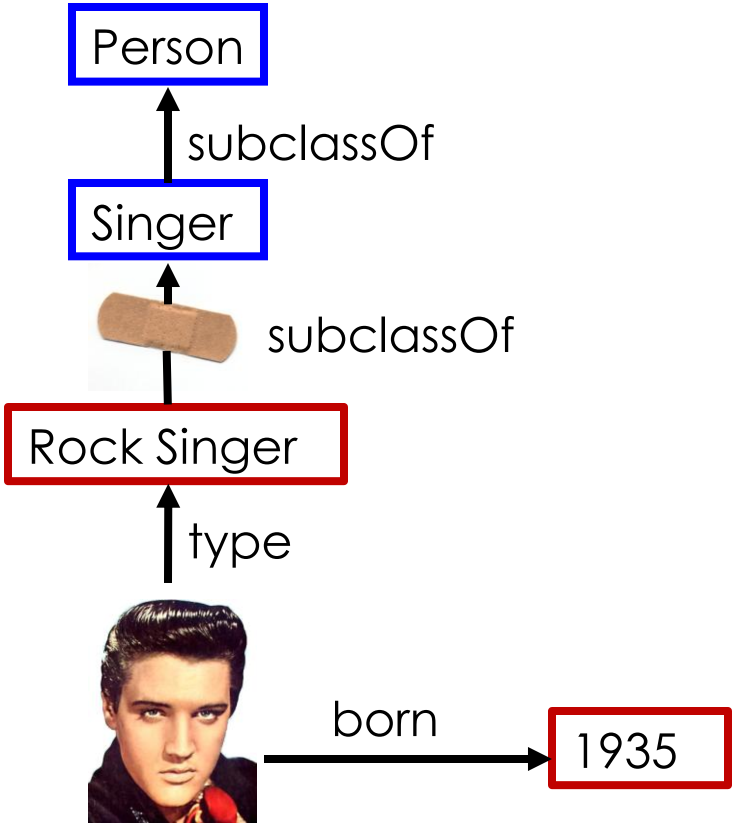


Elvis Presley

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

~Infobox~
Born: 1935
…

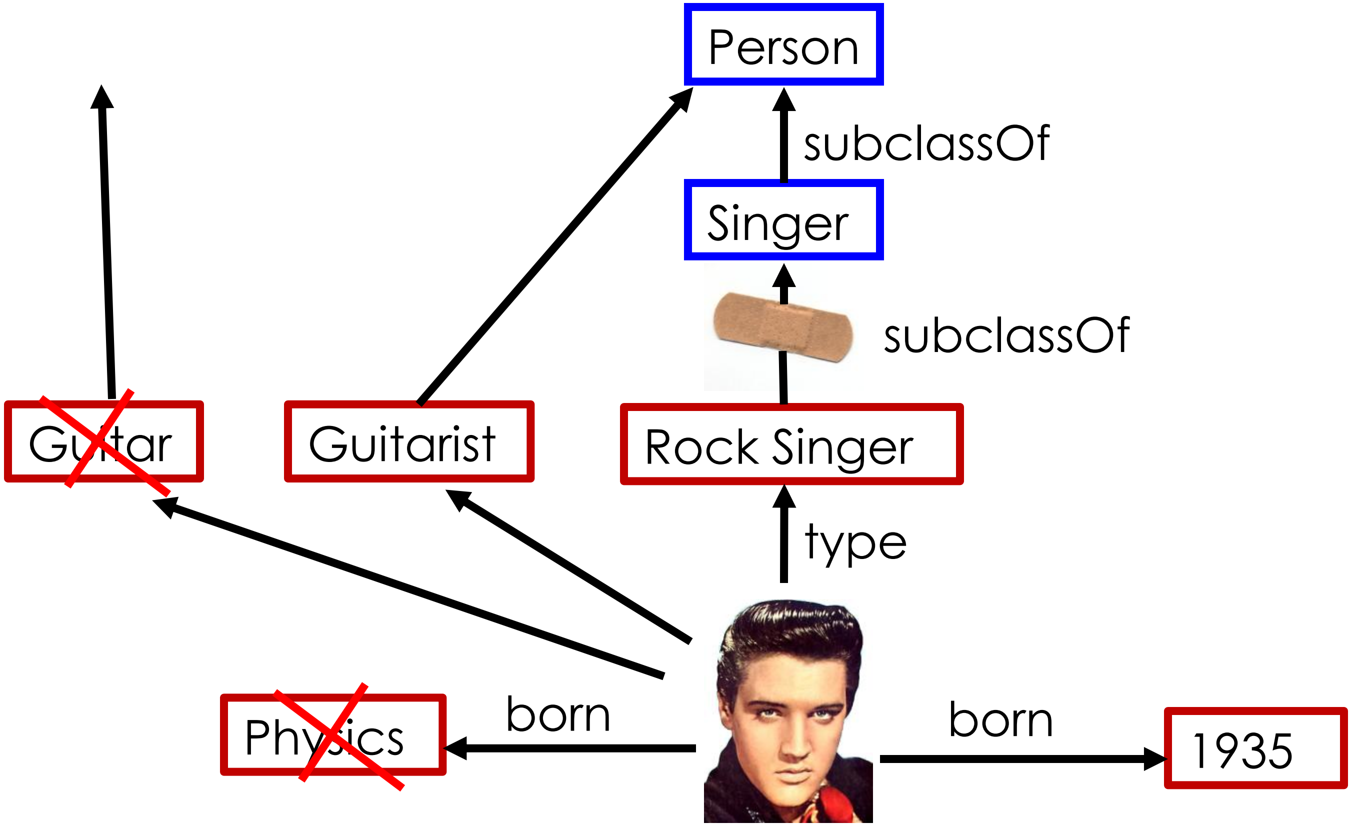
Categories: Rock singers



Exploit Infoboxes

Exploit conceptual categories

IE from Wikipedia: YAGO



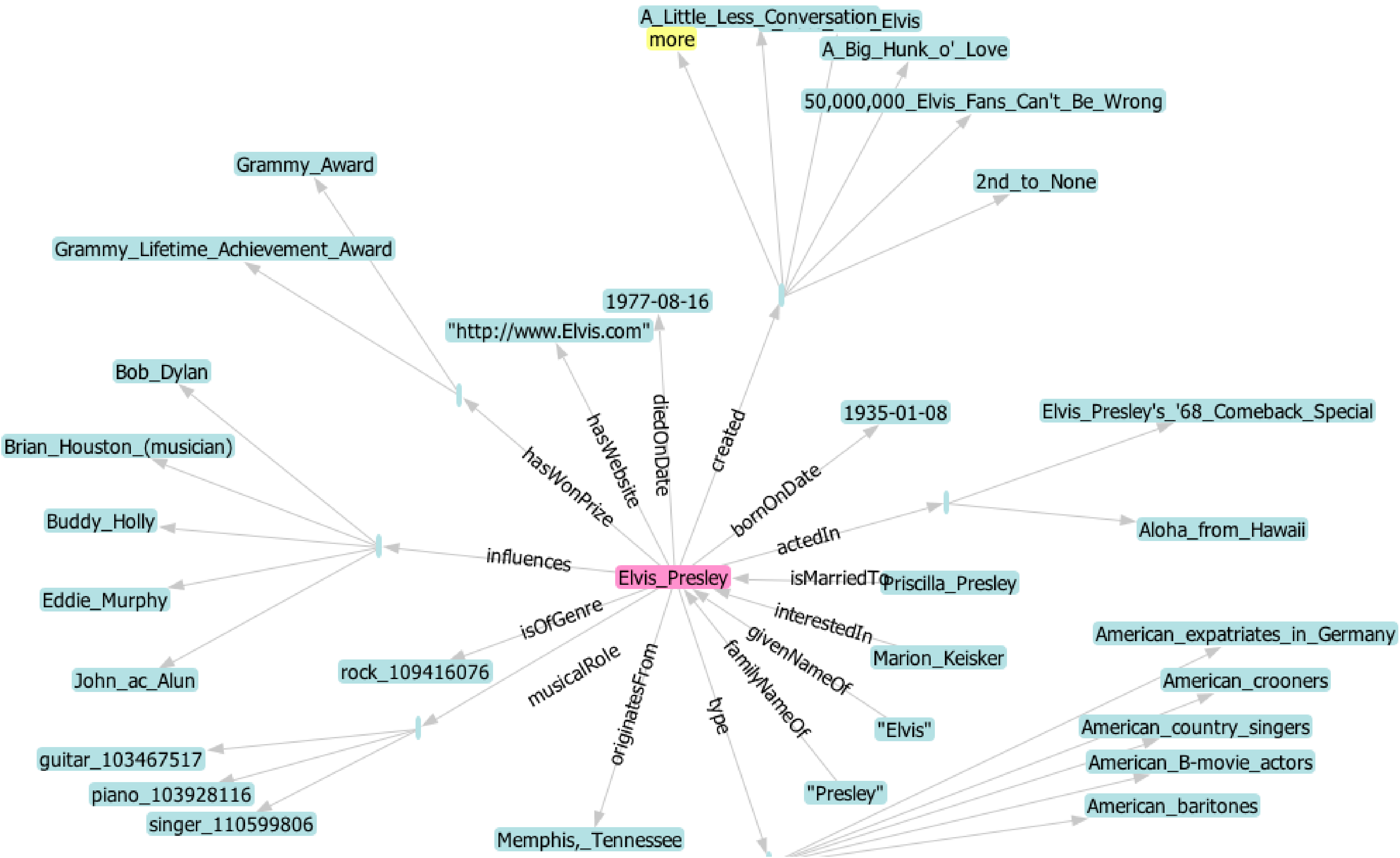
Check uniqueness of entities and functional arguments

Check domains and ranges of relations

Check type coherence

IE from Wikipedia: YAGO

Example: Elvis in YAGO



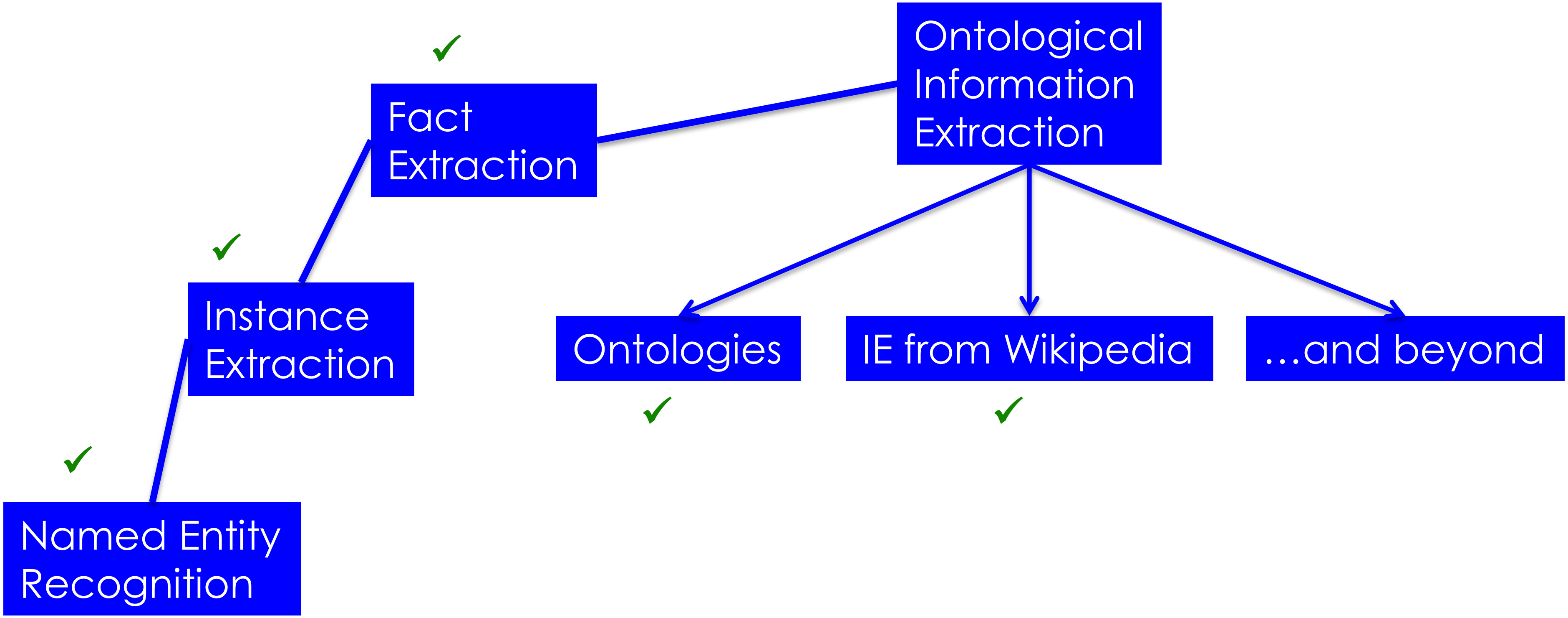
IE from Wikipedia: Summary

Wikipedia is very well suited for ontological IE

Numerous ontology projects make use of Wikipedia:

Information Extraction

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



Ontological IE: Open systems

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

Open Information Extraction/Machine Reading/Macro Reading
aims at information extraction from the entire Web.

Vision of Open Information Extraction:

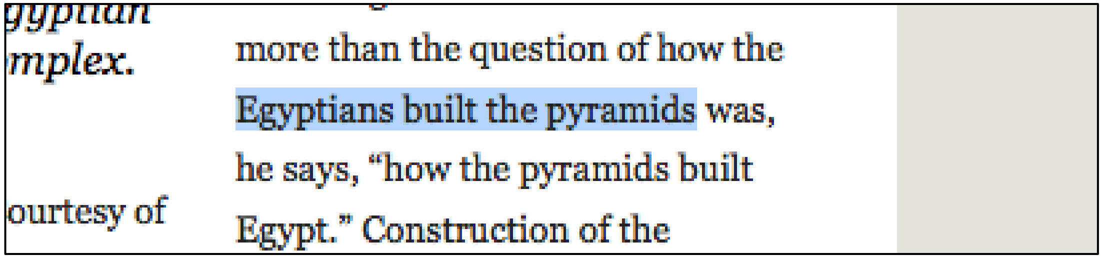
- the system runs perpetually, constantly gathering new information
- the system creates meaning on its own from the gathered data
- the system learns and becomes more intelligent,
  i.e. better at gathering information

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

Ontological IE: KnowItAll &Co

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



gyptian
mplex.

ourtesy of

more than the question of how the Egyptians built the pyramids was, he says, "how the pyramids built Egypt." Construction of the

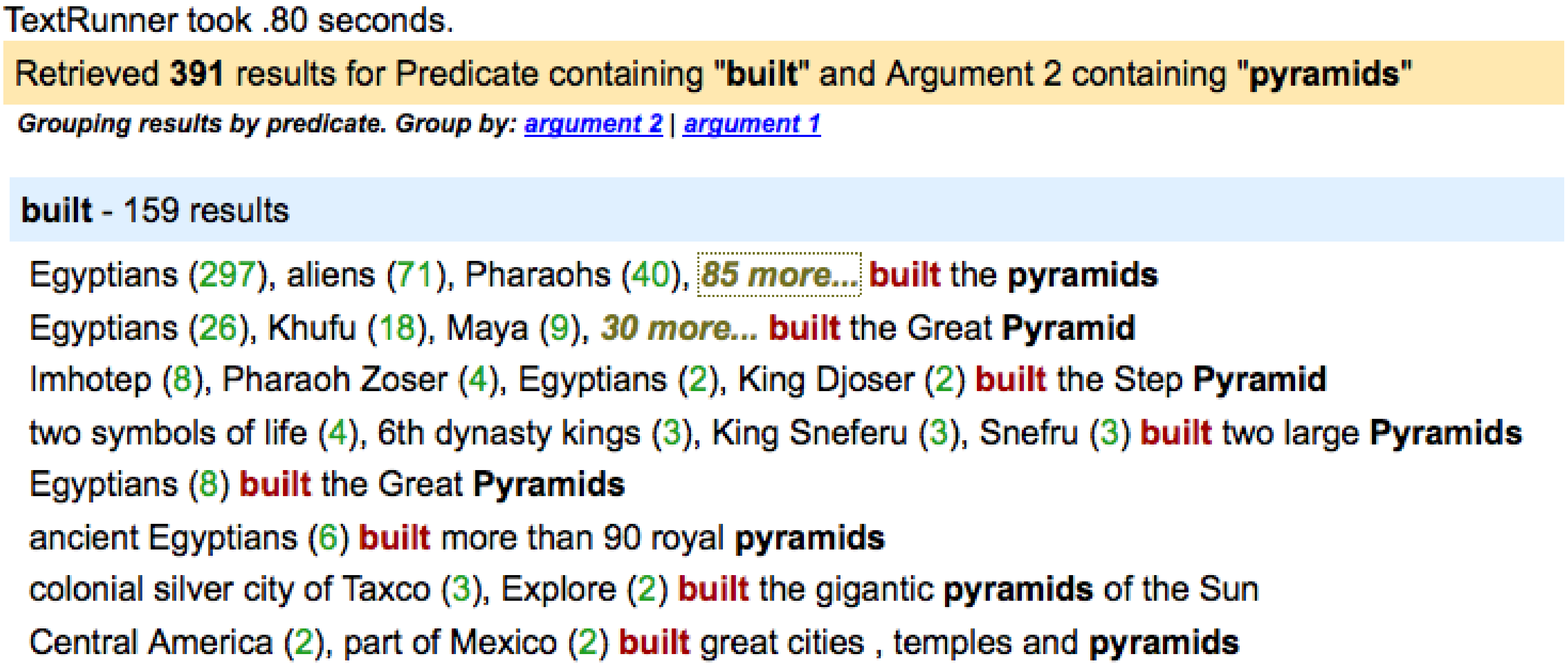
Subject	Verb	Object	Count
Egyptians	built	pyramids	400
Americans	built	pyramids	20
…	…	…	…

Valuable common sense knowledge (if filtered)

http://www.cs.washington.edu/research/textrunner/

Ontological IE: KnowItAll &Co

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



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**

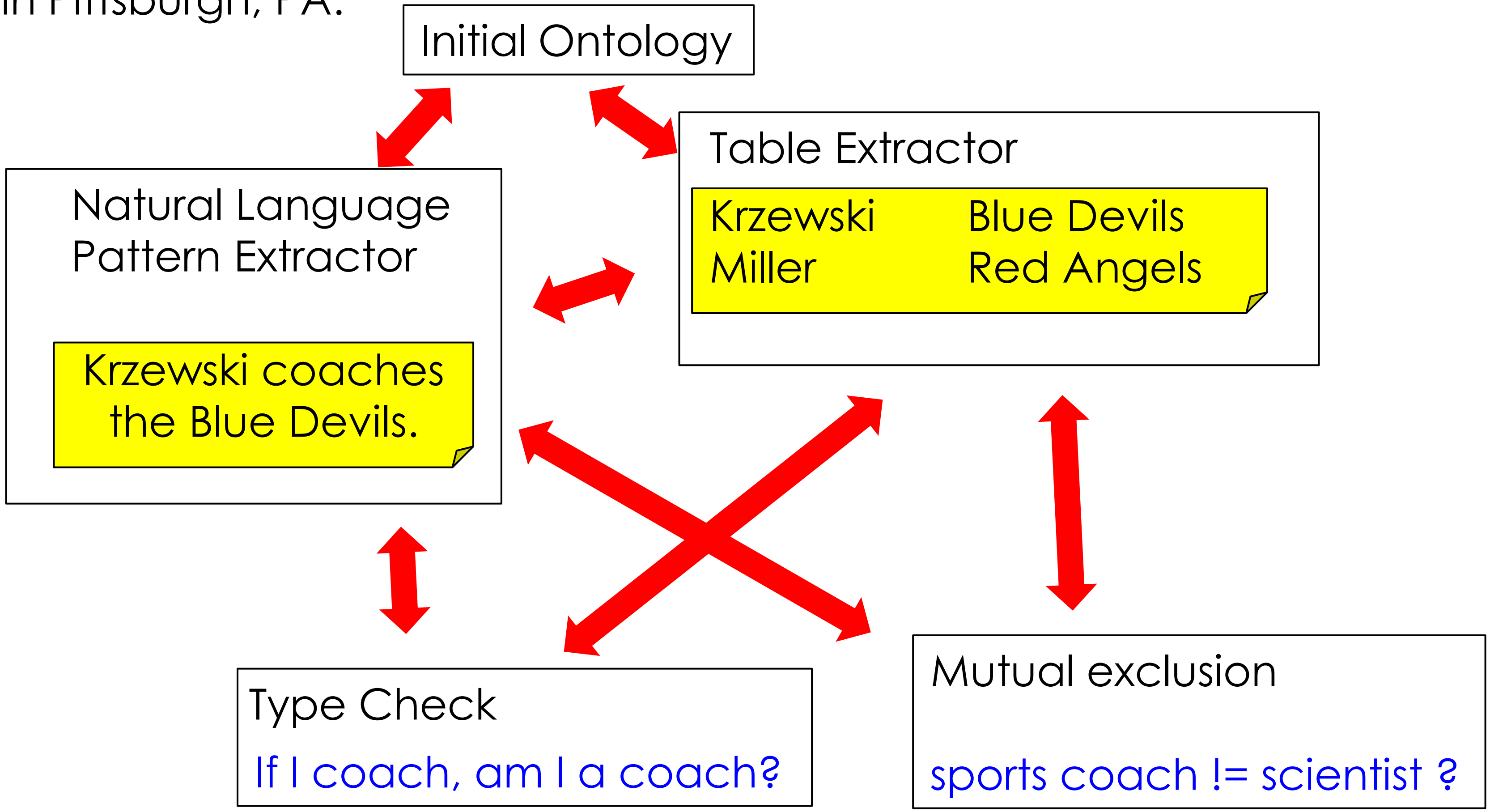
colonial silver city of Taxco (3), Explore (2) **built** the gigantic **pyramids** of the Sun

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

http://www.cs.washington.edu/research/textrunner/

Ontological IE: Read the Web

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



Initial Ontology

Table Extractor

Krzewski	Blue Devils
Miller	Red Angels

Natural Language Pattern Extractor

Krzewski coaches the Blue Devils.

Type Check

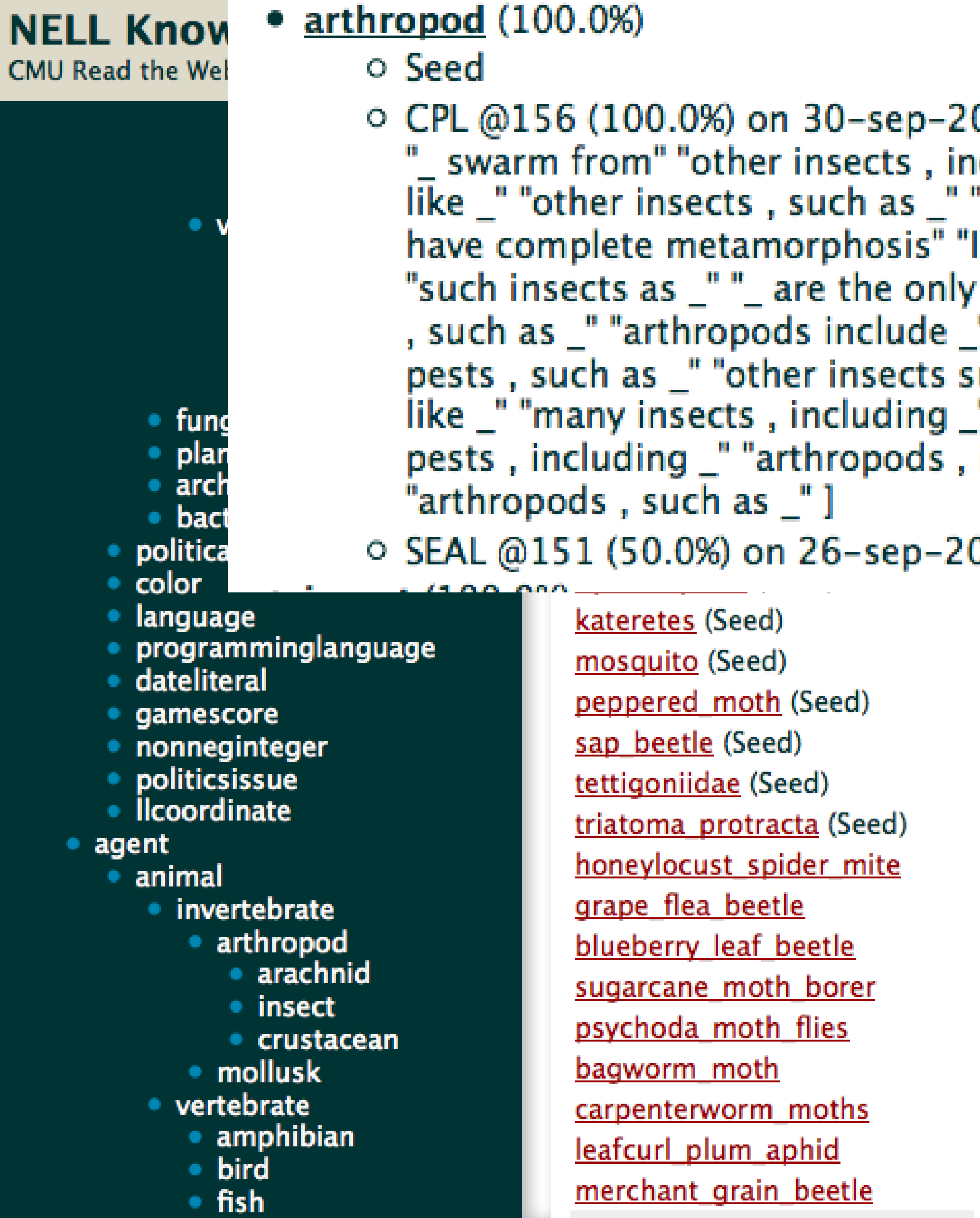
If I coach, am I a coach?

Mutual exclusion

sports coach != scientist ?

http://rtw.ml.cmu.edu/rtw/

Ontological IE: Read the Web



NELL Know
CMU Read the Wel

- 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 si
    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 ir
    , 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]

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

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

http://rtw.ml.cmu.edu/rtw/

Ontological IE: Summary

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

Main hot projects

- TextRunner
- Read the Web

Input:

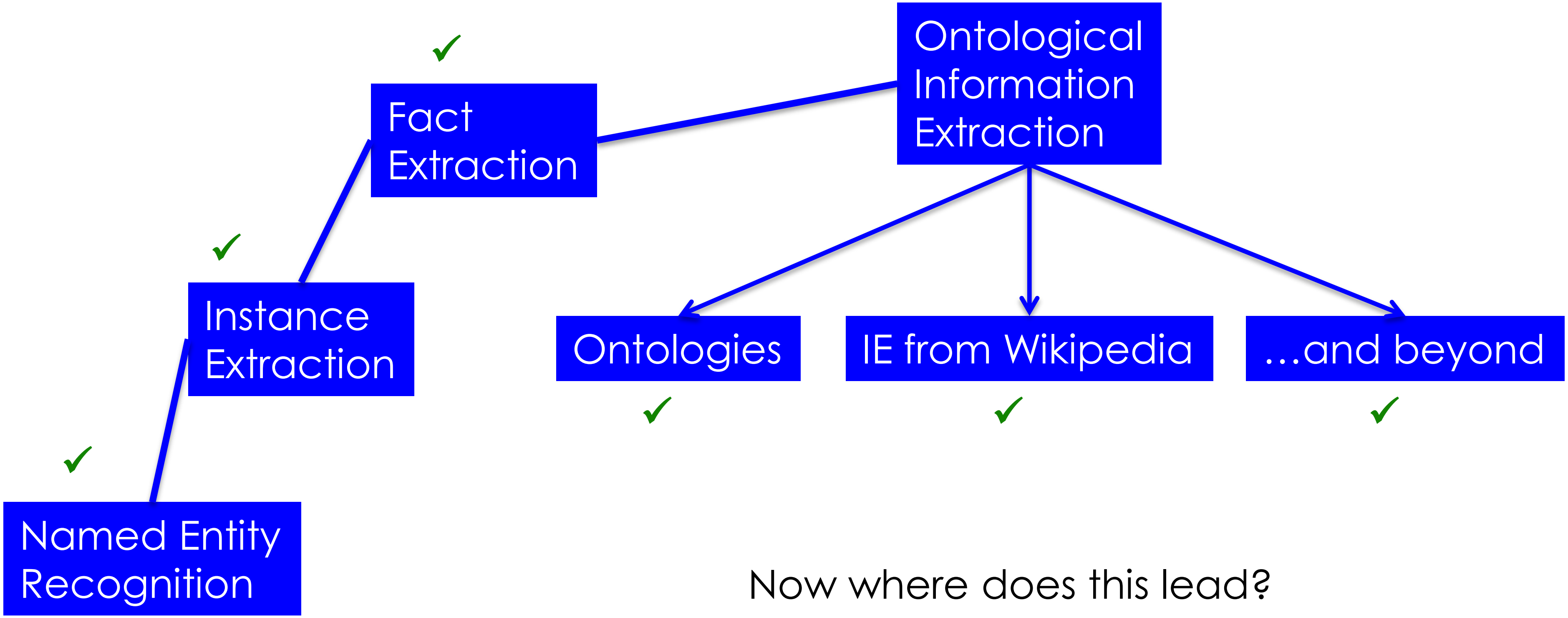
- The Web
- Read the Web: Manual rules
- Read the Web: initial ontology

Conditions

- none

Information Extraction

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



Now where does this lead?

Ontologies

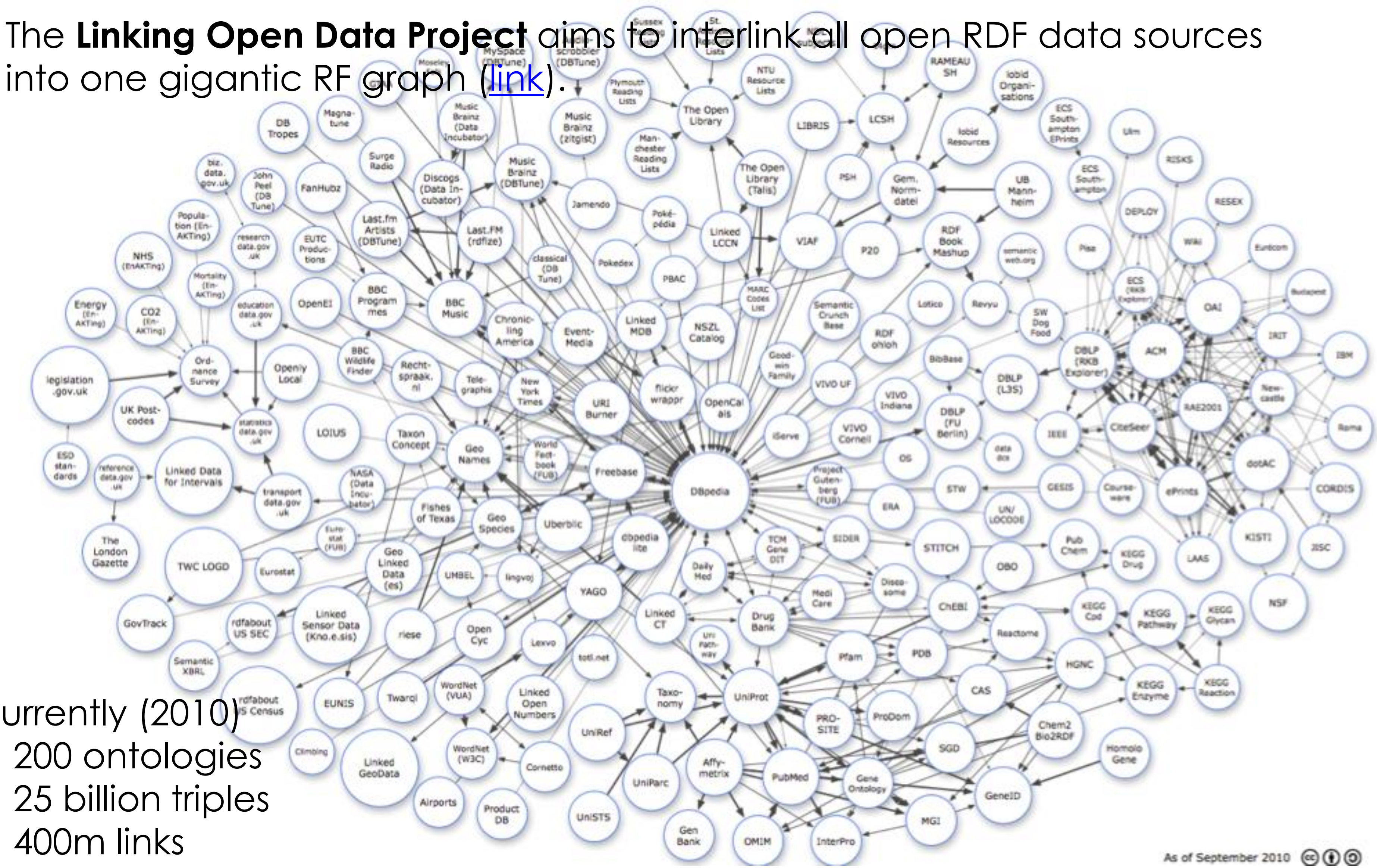
Hundreds of data sets are nowadays available in RDF
(http://www4.wiwiss.fu-berlin.de/lodcloud/)

- US census data
- BBC music database
- Gene ontologies
- DBpedia general knowledge (and hub vocabulary), + YAGO, + Cyc etc.
- UK government data
- geographical data in abundance
- national library catalogs (Hungary, USA, Germany etc.)
- publications (DBLP)
- commercial products
- all Pokemons
- ...and many more

(Only some of these ontologies come from IE.
Many of them are being used for IE)

The Linked Data Cloud

The **Linking Open Data Project** aims to interlink all open RDF data sources into one gigantic RF graph (link).

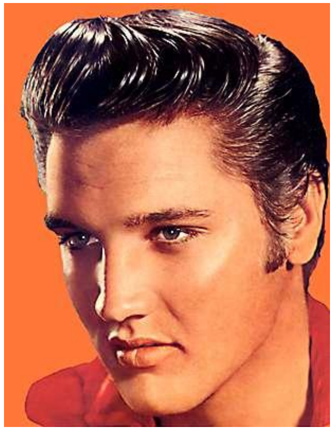


Currently (2010)

- 200 ontologies
- 25 billion triples
- 400m links

http://richard.cyganiak.de/2007/10/lod/imagemap.html

But back to the original question...

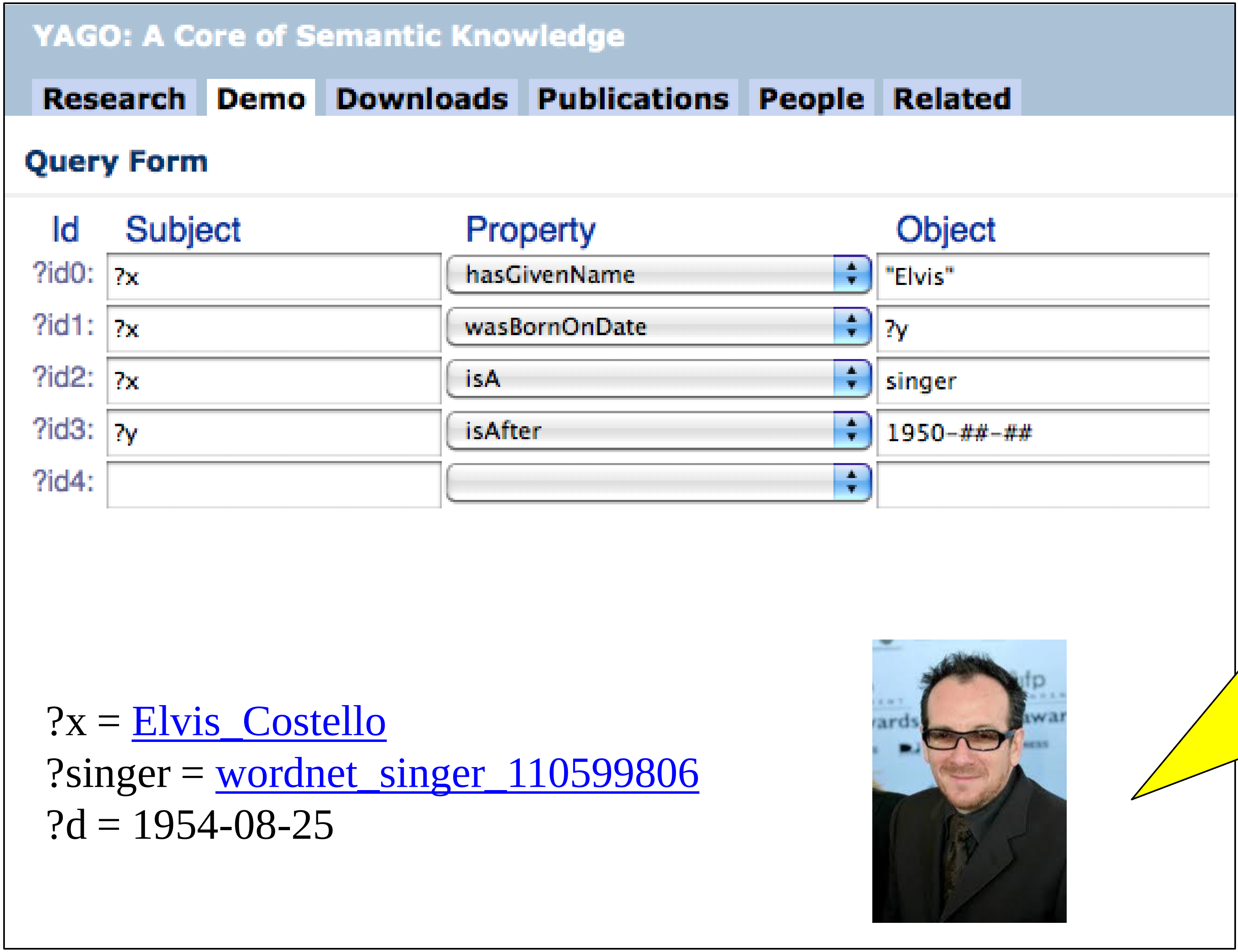


Will there ever be a famous singer called Elvis again?

=> Let’s go query an existing ontology!

But back to the original question…

http://mpii.de/yago



YAGO: A Core of Semantic Knowledge

Query Form

Id	Subject	Property	Object
?id0:	?x	hasGivenName	"Elvis"
?id1:	?x	wasBornOnDate	?y
?id2:	?x	isA	singer
?id3:	?y	isAfter	1950-##-##
?id4:			

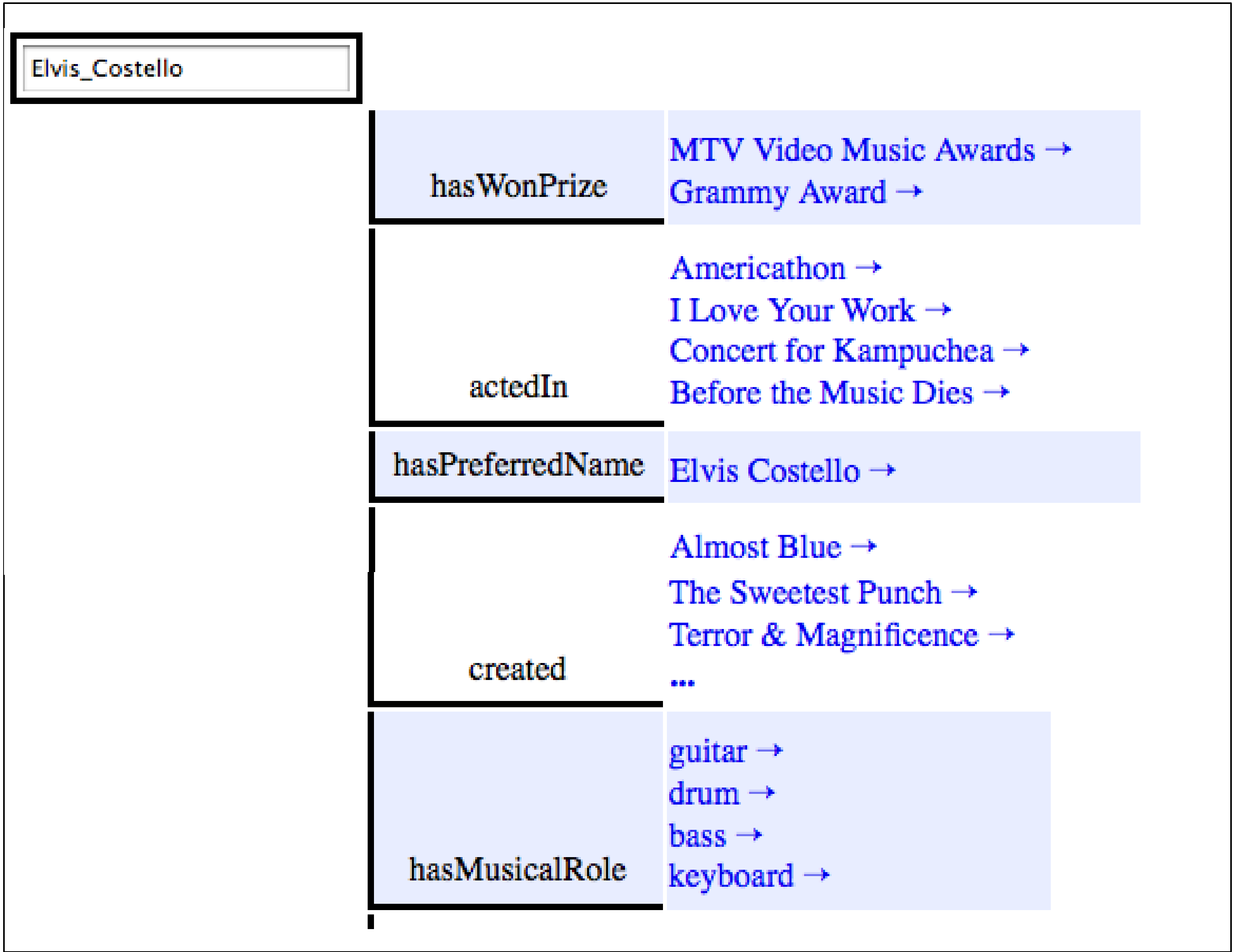
?x = Elvis_Costello
?singer = wordnet_singer_110599806
?d = 1954-08-25

We found him!

Can we find out more about this guy?

But back to the original question...

http://mpii.de/yago



Elvis_Costello

hasWonPrize	MTV Video Music Awards → Grammy Award →
actedIn	Americathon → I Love Your Work → Concert for Kampuchea → Before the Music Dies →
hasPreferredName	Elvis Costello →
created	Almost Blue → The Sweetest Punch → Terror & Magnificence → ...
hasMusicalRole	guitar → drum → bass → keyboard →

Summary

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

We have seen techniques for

- Named entity recognition
- Instance extraction
- Fact extraction

An **ontology** is a structured collection of world knowledge.

We have seen

- basic knowledge representation
- some techniques for ontological IE

And, yes, there is
hope for the
quality of music: