

# **Chapter VI:**

# **Information Extraction**

Information Retrieval & Data Mining  
Universität des Saarlandes, Saarbrücken  
Wintersemester 2013/14

# Chapter VI: Information Extraction

## **VI.1 Motivation & Applications**

Knowledge Queries, Entities & Relations, RDF(S), SPARQL

## **VI.2 Natural Language Processing Basics**

Part-of-Speech Tagging, Dependency Parsing, Word Sense Tagging

## **VI.3 Rule-Based Information Extraction**

Wrapper Induction

## **VI.4 Learning-Based Information Extraction**

Hidden Markov Models, Conditional Random Fields

## **VI.5 Named Entity Reconciliation**

Fellegi-Sunter Model

## **VI.6 Knowledge Base Construction & Open IE**

SNOWBALL, YAGO, TextRunner, NELL

# VI.1 Motivation & Applications

- Beyond keywords as queries and documents as retrieval units
  - **extract entities** and annotate text documents or web pages (e.g., named entity recognition)
  - **find instances** of semantic classes (e.g., not yet known in WordNet)
  - **extract facts** (relations among entities) from text documents or web pages (e.g., Wikipedia) to automatically populate ontology/knowledge base
  - **answer questions** by analyzing natural language and translating it into machine-processable format
- Technologies:
  - Lexicon lookups (name dictionaries, geo gazetteers, etc.)
  - NLP (PoS tagging, chunking/parsing, semantic role labeling, etc.)
  - Pattern matching & rule learning (regular expressions, FSAs)
  - Statistical learning (HMMs, CRFs, etc.)
  - Text mining

# Google Knowledge Graph

[Web](#) [Images](#) [Maps](#) [Shopping](#) [News](#) [More ▾](#) [Search tools](#)

About 127,000,000 results (0.44 seconds)

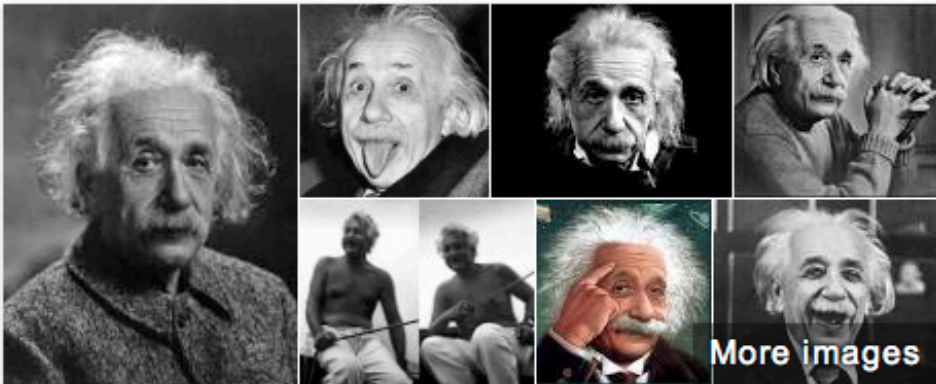
[Albert Einstein - Wikipedia, the free encyclopedia](#)  
[en.wikipedia.org/wiki/Albert\\_Einstein](http://en.wikipedia.org/wiki/Albert_Einstein) ▾  
**Albert Einstein** (/ˈælbɜrt ˈaɪnstʌɪn/; German: [ˈalbɐt ˈaɪnʃtaɪn] (listen); 14 March 1879 – 18 April 1955) was a German-born theoretical physicist who ...  
[Hans Albert Einstein](#) - [Eduard Einstein](#) - [Mileva Marić](#) - [Elsa Einstein](#)

[Albert Einstein – Wikipedia](#)  
[de.wikipedia.org/wiki/Albert\\_Einstein](http://de.wikipedia.org/wiki/Albert_Einstein) ▾ [Translate this page](#)  
**Albert Einstein** (\* 14. März 1879 in Ulm; † 18. April 1955 in Princeton, New Jersey ) war ein theoretischer Physiker. Seine Forschungen zur Struktur von Materie, ...  
[Relativitätstheorie](#) - [Ulm](#) - [Zionismus](#) - [Thomas Harvey](#)

[Albert Einstein - Biographical - Nobelprize.org](#)  
[www.nobelprize.org/nobel\\_prizes/physics/laureates/.../einstein-bio.html](http://www.nobelprize.org/nobel_prizes/physics/laureates/.../einstein-bio.html) ▾  
**Albert Einstein** - Biographical. **Albert Einstein** was born at Ulm, in Württemberg, Germany, on March 14, 1879. Six weeks later the family moved to Munich, where ...

[News for albert einstein](#)  
[Providence digital-content company chosen for online publishing aspects of Albert Einstein collection](#)  
[The Providence Journal](#) - 17 hours ago  
PROVIDENCE – As Princeton University Press works to publish "The Collected Papers of **Albert Einstein**," it has selected a Providence ...  
[Maybe he's a relative? Caterpillar bears incredible resemblance to Albert Ein...](#)  
[Daily Mail](#) - 3 days ago  
[Congress proves Albert Einstein's definition of insanity](#)  
[San Jose Mercury News](#) - 1 day ago

[Einstein Archives Online](#)  
[www.alberteinstein.info/](http://www.alberteinstein.info/) ▾  
The homepage of the repository of the personal papers of the great scientist, humanist and Jew, **Albert Einstein**.




**Albert Einstein**  
Theoretical Physicist

Albert Einstein was a German-born theoretical physicist who developed the general theory of relativity, one of the two pillars of modern physics.  
Wikipedia

**Born:** March 14, 1879, Ulm  
**Died:** April 18, 1955, Princeton, New Jersey, United States  
**Children:** [Eduard Einstein](#), [Hans Albert Einstein](#), [Lieserl Einstein](#)  
**Education:** [University of Zurich](#) (1905), [ETH Zurich](#) (1901), Aargau Cantonal School (1895–1896), Luitpold Gymnasium  
**Spouse:** [Elsa Einstein](#) (m. 1919–1936), [Mileva Marić](#) (m. 1903–1919)  
**Awards:** Nobel Prize in Physics, Copley Medal, Franklin Medal, [More](#)


People also search for

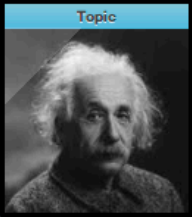


<http://www.google.com>




# Freebase

 Find... Browse Query Help



Topic

## Albert Einstein <sup>en</sup>

mid: /m/0jcx notable type: /education/academic on the web:  wikipedia.org

Albert Einstein was a German-born theoretical physicist who developed the general theory of relativity, one of the two pillars of modern physics. While best known for his mass–energy equivalence formula  $E = mc^2$ , he received the 1921 Nobel Prize in Physics "for his service to physics and his discovery of the law of the photoelectric effect". The latter was pivotal in establishing quantum theory. Near the beginning of his career, Einstein thought that Newtonian mechanics was no longer enough to reconcile the laws of classical mechanics with the laws of the electromagnetic field. This led to the development of his special theory of relativity. He realized, however, that the principle of relativity could also be extended to gravitational fields, and with his subsequent theory of gravitation in 1916, he published a paper on the general theory of relativity. He continued to deal with problems of statistical mechanics and quantum theory, which led to his explanations of particle theory and the motion of molecules. He also investigated the thermal properties of light which laid the foundation of the photon theory of light. In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. Wikipedia [-]

Properties

118n

Keys

Links

View and edit specific domains, types, or properties:

Filter options: ☐ Show all domains and properties

Common /common

Freebase Commons

Topic /common/topic

X

Also known as /common/topic/alias

Einstein







5 values total +

Description /common/topic/description

Albert Einstein was a German-born theoretical physicist who developed the general theory of relativity, one of the two pillars of modern physics. While best known for his mass–energy equivalence formula  $E = mc^2$ , he received the 1921 Nobel Prize in Physics "for his services to theoretical physics, and especially for his discovery of the law of the photoelectric effect". The latter was pivotal in establishing quantum theory. Near the beginning of his career, Einstein thought that Newtonian mechanics was no longer enough to reconcile the laws of classical mechanics with the laws of the electromagnetic field. This led to the development of his special theory of relativity. He realized, however, that the principle of relativity could also be extended to gravitational fields, and with his subsequent theory of gravitation in 1916, he published a paper on the general theory of relativity. He continued to deal with problems of statistical mechanics and quantum theory, which led to his explanations of particle theory and the motion of molecules. He also investigated the thermal properties of light which laid the foundation of the photon theory of light. In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. Wikipedia

41 values total +

Image /common/topic/image



<http://www.freebase.com>

## Browse YAGO2

Search:  eng

<Albert\_Einstein>



<p>&lt;Mileva_Marić&gt; &lt;Elsa_Einstein&gt;</p> <p>&lt;isMarriedTo&gt;</p> <p>&lt;id_zirval_1cp_h1oldc&gt; &lt;id_tiwmcu_16x_11ovtp&gt; &lt;id_vw40n3_1co_1yicwlj&gt; &lt;id_1gjh4l_1co_1m6tzij&gt; &lt;id_1m87rp4_1cp_unqotn&gt; &lt;id_zirval_1co_167wix7&gt; &lt;id_tiwmcu_16x_1ad43sb&gt; &lt;id_tiwmcu_ab2_1t9dinx&gt; &lt;id_tiwmcu_ab2_ec4aec&gt; &lt;id_1m87rp4_1co_h1oldc&gt; &lt;id_tiwmcu_ab2_9o3zmc&gt; &lt;id_lb2dll_1co_1yicwlj&gt;</p> <p>&lt;extractionSource&gt;</p>		<p>"Albert Einstein"@br "Алберт Айншайн"@bg "Albert Einstein"@bs "Альберт Эйнштэйн"@be "Albert Einstein"@bcl "Albert Einstein"@en → &lt;extractionSource&gt; &lt;http://en.wikipedia.org/wiki/Albert_Einstein&gt; "আলবার্ট আইনস্টাইন"@bn "Albert Einstein"@map-bms "Albert Eynsteyn"@az "Albert Einstein"@bm "Albert Einstein"@als "Albert Einstein"@af "The Development of Our Views on the Composition and Essence of Radiation"@s "Albert Einstein"@ay "AE"@en "Albert Einstein"@an "ألبرت أينشتاين"@ar "Albert Einstein"@ang "አልበርት አይንስታይን"@am "Albert Einstein"@ast "এলবার্ট আইনস্টাইন"@as</p> <p>rdfs:label</p>
<p>&lt;Einstein_family&gt;</p> <p>&lt;hasChild&gt;</p> <p>&lt;David_Hume&gt; &lt;Crookes_radiometer&gt; &lt;Camille_Pissarro&gt; &lt;COINTELPRO&gt; &lt;Carl_Sagan&gt; &lt;Carl_Friedrich_Gauss&gt; &lt;Corcovado&gt; &lt;California_Institute_of_Technology&gt; &lt;Copenhagen_interpretation&gt; &lt;Alfred_Lawson&gt; &lt;David_Hilbert&gt; &lt;Alan_Turing&gt; &lt;Albert_Schweitzer&gt; &lt;Claude_Shannon&gt; &lt;Arthur_Schopenhauer&gt; &lt;Baruch_Spinoza&gt; &lt;Albert_Brooks&gt; &lt;Aarau&gt; &lt;Arthur_Eddington&gt; &lt;Czech_Republic&gt; &lt;Bertrand_Russell&gt;</p> <p>&lt;linksTo&gt;</p>		<p>&lt;hasAcademicAdvisor&gt;</p> <p>&lt;Alfred_Kleiner&gt; → &lt;extractionSource&gt; &lt;http://en.wikipedia.org/wiki/Albert_Einstein&gt; &lt;extractionSource&gt;</p>
<p>&lt;Arthur_Schopenhauer&gt; &lt;Thomas_Young_(scientist)&gt; &lt;Baruch_Spinoza&gt; &lt;David_Hume&gt; &lt;Moritz_Schlick&gt; &lt;Ernst_Mach&gt;</p> <p>&lt;influences&gt;</p>		<p>&lt;Translational_symmetry&gt; &lt;Hans_Albert_Einstein&gt; &lt;Princeton_New_Jersey&gt; &lt;California_Institute_of_Technology&gt; &lt;Max_Talmey&gt; &lt;Henri_Poincaré&gt; &lt;John_Francis_Hylan&gt; &lt;Matura&gt; &lt;University_of_Bern&gt; &lt;ETH_Zurich&gt; &lt;Matteucci_Medal&gt; &lt;Einstein_field_equations&gt; &lt;Corbis&gt; &lt;BBC_News&gt; &lt;BBC&gt; &lt;Judaism&gt; &lt;Nobel_Foundation&gt; &lt;German_nuclear_energy_project&gt; &lt;Manhattan_Project&gt; &lt;Linus_Pauling&gt; &lt;Equivalence_principle&gt;</p> <p>&lt;linksTo&gt;</p>

<http://www.yago-knowledge.org>

## About: Dave Grohl

An Entity of Type : [musical artist](#), from Named Graph : <http://live.dbpedia.org>, within Data Space : [live.dbpedia.org](http://live.dbpedia.org)



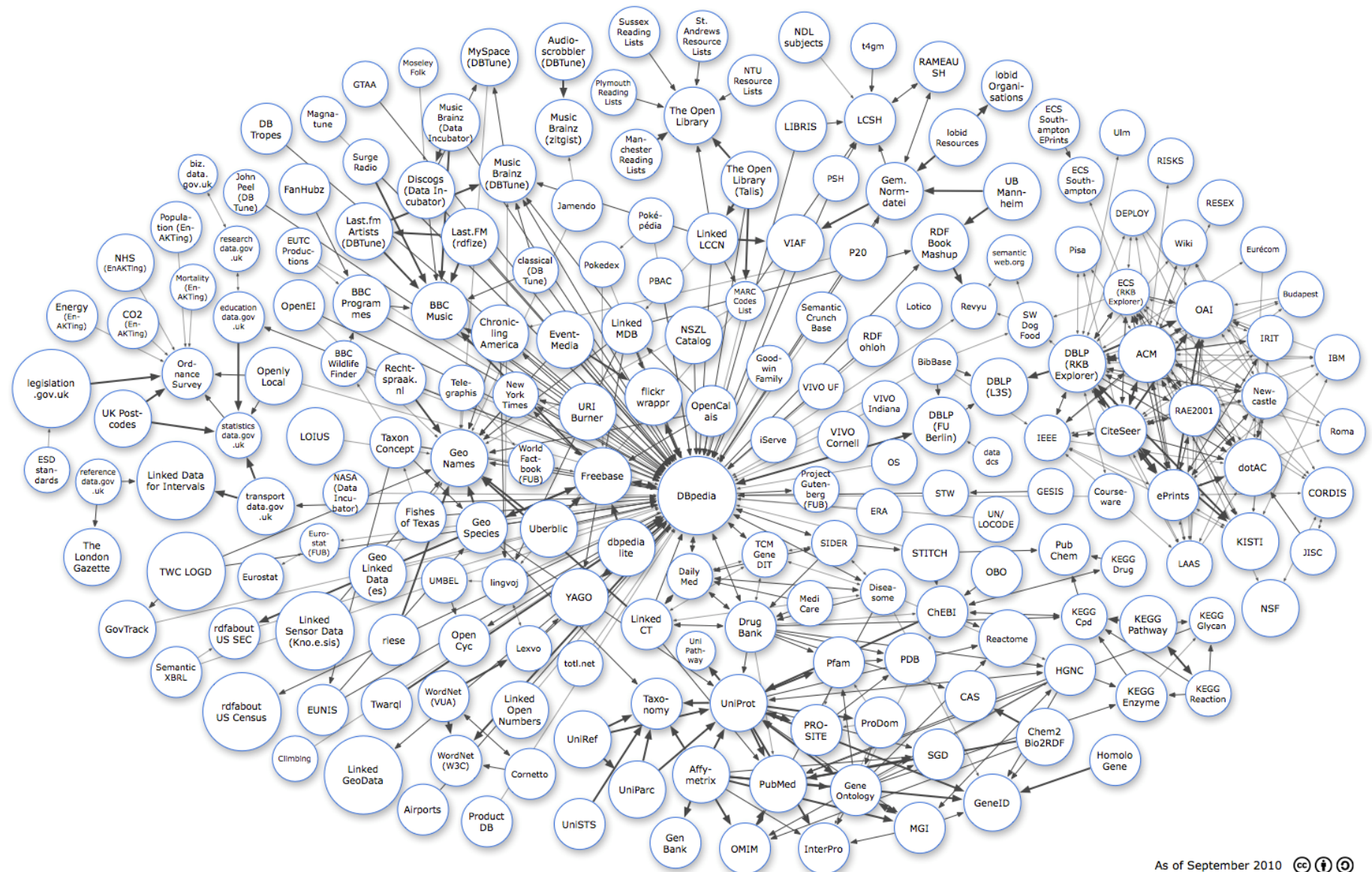
David Eric "Dave" Grohl (born January 14, 1969) is an American rock musician, multi-instrumentalist, singer-songwriter and film director, who is the lead vocalist, guitarist, primary or main songwriter and founder of the band Foo Fighters. Prior to Foo Fighters, Grohl was the drummer for the grunge band Nirvana. He is also the drummer and co-founder of the rock supergroup Them Crooked Vultures.

Property	Value
<a href="#">dbpedia-owl:abstract</a>	<ul style="list-style-type: none"> <li>David Eric "Dave" Grohl (born January 14, 1969) is an American rock musician, multi-instrumentalist, singer-songwriter and film director, who is the lead vocalist, guitarist, primary or main songwriter and founder of the band Foo Fighters. Prior to Foo Fighters, Grohl was the drummer for the grunge band Nirvana. He is also the drummer and co-founder of the rock supergroup Them Crooked Vultures. Grohl has additionally written all the music and performed all the instruments for his short-lived side projects Late! and Probot, as well as being involved with Queens of the Stone Age numerous times throughout the past decade. He has performed session work (as a drummer) for a variety of musicians, including Garbage, Killing Joke, Nine Inch Nails, David Bowie, Paul McCartney, The Prodigy, Slash, Iggy Pop, Juliette Lewis, Tenacious D, Tom Petty and the Heartbreakers, Lemmy and Stevie Nicks.</li> </ul>
<a href="#">dbpedia-owl:activeYearsStartYear</a>	<ul style="list-style-type: none"> <li>1984-01-01 00:00:00 (xsd:date)</li> <li>1984-01-01 00:00:00 (xsd:date)</li> </ul>
<a href="#">dbpedia-owl:alias</a>	<ul style="list-style-type: none"> <li>Davy Grolton, Dale Nixon, Late! (pseudonym for his solo album Pocketwatch), and Dr. G (as Tenacious D's drummer) .</li> </ul>
<a href="#">dbpedia-owl:associatedBand</a>	<ul style="list-style-type: none"> <li><a href="#">dbpedia:Paul_McCartney</a></li> <li><a href="#">dbpedia:Stevie_Nicks</a></li> <li><a href="#">dbpedia:The_Prodigy</a></li> <li><a href="#">dbpedia:Trent_Reznor</a></li> <li><a href="#">dbpedia:Tom_Petty_and_the_Heartbreakers</a></li> <li><a href="#">dbpedia:Rick_Springfield</a></li> <li><a href="#">dbpedia:Killing_Joke</a></li> <li><a href="#">dbpedia:Probot</a></li> <li><a href="#">dbpedia:Mondo_Generator</a></li> <li><a href="#">dbpedia:Tenacious_D</a></li> <li><a href="#">dbpedia:Foo_Fighters</a></li> <li><a href="#">dbpedia:Queens_of_the_Stone_Age</a></li> <li><a href="#">dbpedia:Them_Crooked_Vultures</a></li> <li><a href="#">dbpedia:Dain_Bramage</a></li> <li><a href="#">dbpedia:Nirvana_(band)</a></li> <li><a href="#">dbpedia:Slash_(musician)</a></li> <li><a href="#">dbpedia:Scream_(band)</a></li> </ul>
<a href="#">dbpedia-owl:associatedMusicalArtist</a>	<ul style="list-style-type: none"> <li><a href="#">dbpedia:Paul_McCartney</a></li> <li><a href="#">dbpedia:Stevie_Nicks</a></li> <li><a href="#">dbpedia:The_Prodigy</a></li> <li><a href="#">dbpedia:Trent_Reznor</a></li> <li><a href="#">dbpedia:Tom_Petty_and_the_Heartbreakers</a></li> <li><a href="#">dbpedia:Rick_Springfield</a></li> <li><a href="#">dbpedia:Killing_Joke</a></li> <li><a href="#">dbpedia:Probot</a></li> <li><a href="#">dbpedia:Mondo_Generator</a></li> <li><a href="#">dbpedia:Tenacious_D</a></li> <li><a href="#">dbpedia:Foo_Fighters</a></li> <li><a href="#">dbpedia:Queens_of_the_Stone_Age</a></li> <li><a href="#">dbpedia:Them_Crooked_Vultures</a></li> <li><a href="#">dbpedia:Dain_Bramage</a></li> <li><a href="#">dbpedia:Nirvana_(band)</a></li> <li><a href="#">dbpedia:Slash_(musician)</a></li> <li><a href="#">dbpedia:Scream_(band)</a></li> </ul>
<a href="#">dbpedia-owl:background</a>	<ul style="list-style-type: none"> <li>solo_singer</li> </ul>
<a href="#">dbpedia-owl:birthDate</a>	<ul style="list-style-type: none"> <li>1969-01-14 (xsd:date)</li> <li>1969-01-14 (xsd:date)</li> </ul>
<a href="#">dbpedia-owl:birthPlace</a>	<ul style="list-style-type: none"> <li><a href="#">dbpedia:United_States</a></li> <li><a href="#">dbpedia:Ohio</a></li> <li><a href="#">dbpedia:Warren,_Ohio</a></li> <li><a href="#">dbpedia:Norrköping,_Sweden</a></li> </ul>
<a href="#">dbpedia-owl:genre</a>	<ul style="list-style-type: none"> <li><a href="#">dbpedia:Hardcore_punk</a></li> <li><a href="#">dbpedia:Alternative_rock</a></li> <li><a href="#">dbpedia:Hard_rock</a></li> <li><a href="#">dbpedia:Heavy_metal_music</a></li> <li><a href="#">dbpedia:Post-grunge</a></li> <li><a href="#">dbpedia:Grunge</a></li> </ul>

<http://www.dbpedia.org>



# The Linked Data Project



As of September 2010 © ⓘ ⓘ

- As of 2011:
  - 295 sources
  - 32 billion RDF triples
  - 504 million links

<http://www.linkeddata.org>







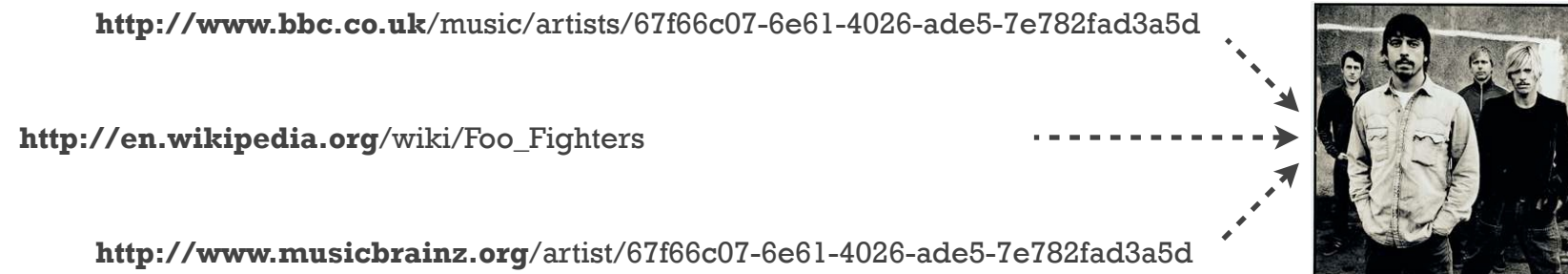
# Semantic Web

- **Semantic Web** [Berners-Lee '01] is an extension of the World Wide Web to make its contents **interpretable for machines**
- **World Wide Web Consortium** (W3C) Semantic Web standards
  - **Unified Resource Identifier** (URI)  
to uniquely identify abstract or physical resources
  - **Resource Description Framework** (RDF)  
to describe properties of abstract or physical resources
  - **Resource Description Framework Schema** (RDF/S)  
to describe schemata
  - **Web Ontology Language** (OWL)  
to describe ontologies
  - **SPARQL Protocol and Query Language** (SPARQL)  
to formulate queries



# Unified Resource Identifier

- **Unified Resource Identifier** (URI) is a string of characters that uniquely identifies an **abstract or physical resource**



- `http://www.host.org/pub/bands?query=FF#albums`
  - **scheme** (e.g., http, ftp, urn) determines interpretation of URI
  - **authority** indicates who is responsible for the resource (e.g., a host)
  - **path** provides hierarchical information for identifying the resource
  - **query** provides non-hierarchical information for identifying the resource
  - **fragment** refers to a specific part of the resource

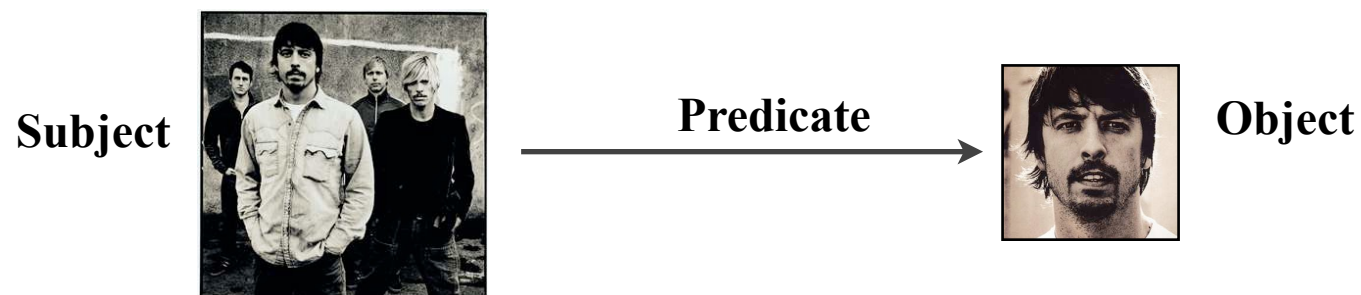
# RDF

- **Resource Description Framework** (RDF) provides a data model to describe properties of resources (identified by their URI)
- **RDF statements** are (S, P, O) triples consisting of a **subject** (URI), a **predicate** (a URI), and an **object** (a URI or literal)
- Example: Dave Grohl is a member of Foo Fighters

<http://dbtune.org/musicbrainz/page/artist/67f66c07-6e61-4026-ade5-7e782fad3a5d> (**S**)

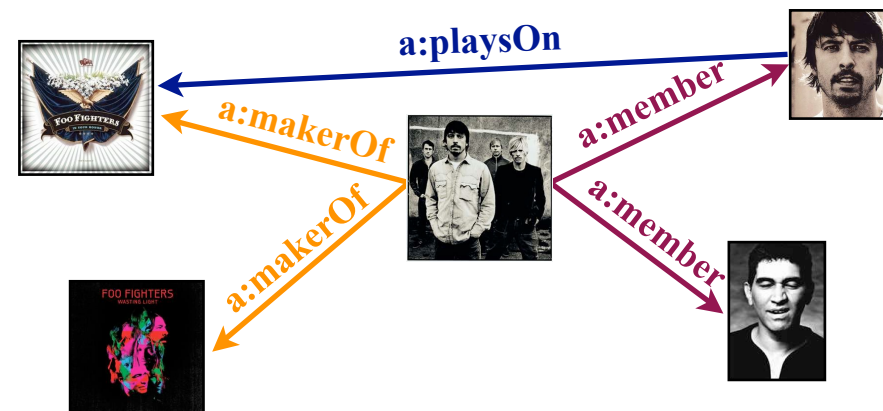
<http://xmlns.com/foaf/spec/20100809.html#member> (**P**)

<http://dbtune.org/musicbrainz/page/artist/4d5f891d-9bce-45ae-ad86-912dd27252fa> (**O**)



# RDF (cont'd)

- RDF triples form a **RDF graph**



- **Namespaces** represent **common URI prefixes** and allow for a more compact representation of RDF data

@prefix a: <http://allaboutmusic.org/>

- **RDF/N3** as one possible text representation of RDF data

```
@ prefix    a:  http://allaboutmusic.org

a:Foo_Fighters  a:member  a:Dave_Grohl
a:Foo_Fighters  a:member  a:Pat_Smear
```

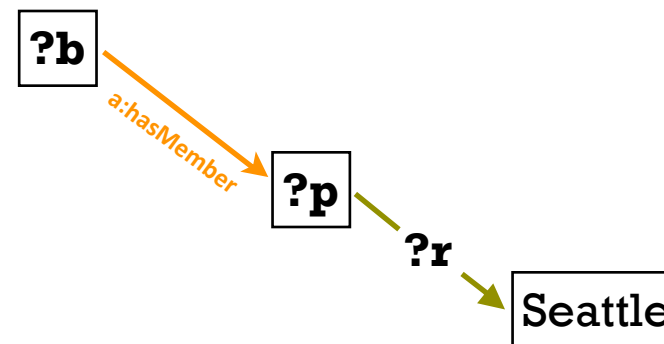
# SPARQL

- **SPARQL Protocol and Query Language** (SPARQL) is a **query language** for the Semantic Web standardized by the W3C
- SPARQL has a **SQL-inspired syntax** to define **graph patterns** and retrieves all matching subgraphs as query results
- Example:

Query

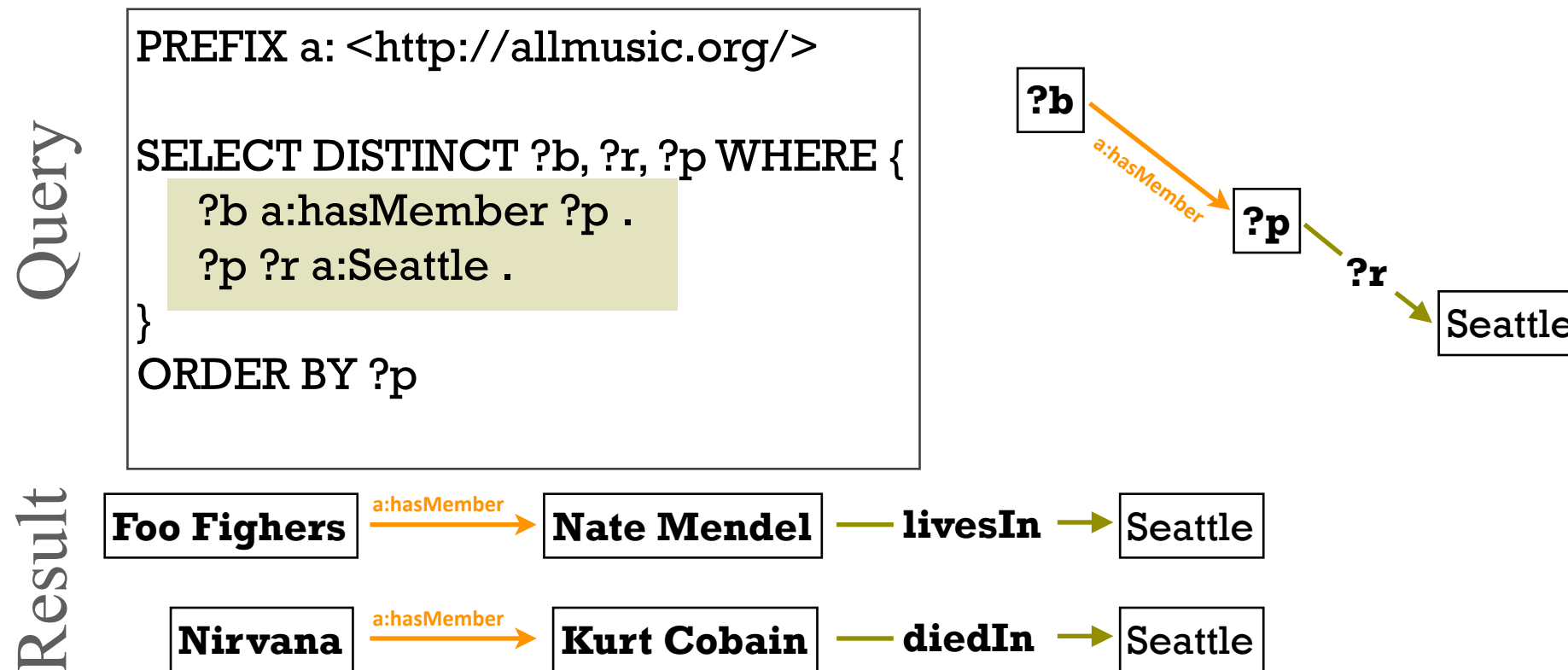
```
PREFIX a: <http://allmusic.org/>

SELECT DISTINCT ?b, ?r, ?p WHERE {
  ?b a:hasMember ?p .
  ?p ?r a:Seattle .
}
ORDER BY ?p
```




# SPARQL

- **SPARQL Protocol and Query Language** (SPARQL) is a **query language** for the Semantic Web standardized by the W3C
- SPARQL has a **SQL-inspired syntax** to define **graph patterns** and retrieves all matching subgraphs as query results
- Example:



# Wolfram Alpha



Who was German chancellor in 1992?

Examples Random

Input interpretation:

Germany Chancellor 03. December 1992

Result:

Helmut Kohl

Basic information:

official position	Chancellor
country	Germany
political affiliation	Christian Democratic Union
start date	01. October 1982 (31 years 2 months 2 days ago)
end date	27. October 1998 (15 years 1 month 7 days ago)
duration of leadership	16 years 26 days

Sequence:

Tuesday, November 22, 2005 to Tuesday, December 3, 2013	Angela Merkel (Christian Democratic Union)
Tuesday, October 27, 1998 to Tuesday, November 22, 2005 (7 years)	Gerhard Schröder (Social Democratic Party of Germany)
Friday, October 1, 1982 to Tuesday, October 27, 1998 (16 years)	Helmut Kohl (Christian Democratic Union)
Thursday, May 16, 1974 to Friday, October 1, 1982 (8 years 5 months)	Helmut Schmidt (Social Democratic Party of Germany)
Tuesday, May 7, 1974 to Thursday, May 16, 1974 (9 days)	Walter Scheel (acting) (Free Democratic Party)

Personal information:

full name	Helmut Michael Kohl
date of birth	03. April 1930 (age: 83 years)
place of birth	Ludwigshafen, Rhineland-Palatinate, Germany

Timeline:  
Helmut Kohl



How was the weather last year?

Examples Random

Input interpretation:

weather last year

Recorded weather for Saarbrücken, Germany:

Show non-metric More

time range	2012
temperature	(-21 to 37) °C (average low: 4 °C   average high: 15 °C)
relative humidity	average: 75%
wind speed	average: 3 m/s (maximum: 15 m/s)

Units »

Give us your feedback:

Send

About | Pro | Products | Mobile Apps | Business Solutions | For Developers | Resources & Tools

Blog | Forum | Participate | Contact | Connect

© 2013 Wolfram Alpha LLC—A Wolfram Research Company | Terms | Privacy Policy

<http://www.wolframalpha.com>



# Information Extraction (IE): Text to Relations



Max Planck

The Nobel Prize in Physics 1918

*Max Karl Ernst Ludwig Planck was born in Kiel, Germany, on April 23, 1858, the son of Julius Wilhelm and Emma (née Patzig) Planck. Planck studied at the Universities of Munich and Berlin, where his teachers included Kirchhoff and Helmholtz, and received his doctorate of philosophy at Munich in 1879. He was Privatdozent in Munich from 1880 to 1885, then Associate Professor of Theoretical Physics at Kiel until 1889, in which year he succeeded Kirchhoff as Professor at Berlin University, where he remained until his retirement in 1926. Afterwards he became President of the Kaiser Wilhelm Society for the Promotion of Science, a post he held until 1937. He was also a gifted pianist and is said to have at one time considered music as a career. Planck was twice married. Upon his appointment, in 1885, to Associate Professor in his native town Kiel he married a friend of his childhood, Marie Merck, who died in 1909. He remarried her cousin Marga von Hösslin. Three of his children died young, leaving him with two sons.*

# Information Extraction (IE): Text to Relations



Max Planck

The Nobel Prize in Physics 1918

bornOn(Max Planck, 23 April 1858)

bornIn(Max Planck, Kiel)

*Max Karl Ernst Ludwig Planck was born in Kiel, Germany, on April 23, 1858, the son of Julius Wilhelm and Emma (née Patzig) Planck. Planck studied at the Universities of Munich and Berlin, where his teachers included Kirchhoff and Helmholtz, and received his doctorate of philosophy at Munich in 1879. He was Privatdozent in Munich from 1880 to 1885, then Associate Professor of Theoretical Physics at Kiel until 1889, in which year he succeeded Kirchhoff as Professor at Berlin University, where he remained until his retirement in 1926. Afterwards he became President of the Kaiser Wilhelm Society for the Promotion of Science, a post he held until 1937. He was also a gifted pianist and is said to have at one time considered music as a career. Planck was twice married. Upon his appointment, in 1885, to Associate Professor in his native town Kiel he married a friend of his childhood, Marie Merck, who died in 1909. He remarried her cousin Marga von Hösslin. Three of his children died young, leaving him with two sons.*

# Information Extraction (IE): Text to Relations



Max Planck

The Nobel Prize in Physics 1918

bornOn(Max Planck, 23 April 1858)

bornIn(Max Planck, Kiel)

type(Max Planck, physicist)

*Max Karl Ernst Ludwig Planck was born in Kiel, Germany, on April 23, 1858, the son of Julius Wilhelm and Emma (née Patzig) Planck. Planck studied at the Universities of Munich and Berlin, where his teachers included Kirchhoff and Helmholtz, and received his doctorate of philosophy at Munich in 1879. He was Privatdozent in Munich from 1880 to 1885, then Associate Professor of Theoretical Physics at Kiel until 1889, in which year he succeeded Kirchhoff as Professor at Berlin University, where he remained until his retirement in 1926. Afterwards he became President of the Kaiser Wilhelm Society for the Promotion of Science, a post he held until 1937. He was also a gifted pianist and is said to have at one time considered music as a career. Planck was twice married. Upon his appointment, in 1885, to Associate Professor in his native town Kiel he married a friend of his childhood, Marie Merck, who died in 1909. He remarried her cousin Marga von Hösslin. Three of his children died young, leaving him with two sons.*

# Information Extraction (IE): Text to Relations



Max Planck

The Nobel Prize in Physics 1918

bornOn(Max Planck, 23 April 1858)

bornIn(Max Planck, Kiel)

type(Max Planck, physicist)

plays(Max Planck, piano)

*Max Karl Ernst Ludwig Planck was born in Kiel, Germany, on April 23, 1858, the son of Julius Wilhelm and Emma (née Patzig) Planck. Planck studied at the Universities of Munich and Berlin, where his teachers included Kirchhoff and Helmholtz, and received his doctorate of philosophy at Munich in 1879. He was Privatdozent in Munich from 1880 to 1885, then Associate Professor of Theoretical Physics at Kiel until 1889, in which year he succeeded Kirchhoff as Professor at Berlin University, where he remained until his retirement in 1926. Afterwards he became President of the Kaiser Wilhelm Society for the Promotion of Science, a post he held until 1937. He was also a gifted pianist and is said to have at one time considered music as a career. Planck was twice married. Upon his appointment, in 1885, to Associate Professor in his native town Kiel he married a friend of his childhood, Marie Merck, who died in 1909. He remarried her cousin Marga von Hösslin. Three of his children died young, leaving him with two sons.*



# Information Extraction (IE): Text to Relations



Max Planck

The Nobel Prize in Physics 1918

bornOn(Max Planck, 23 April 1858)

bornIn(Max Planck, Kiel)

type(Max Planck, physicist)

plays(Max Planck, piano)

spouse(Max Planck, Marie Merck)

spouse(Max Planck, Marga Hösslin)

*Max Karl Ernst Ludwig Planck was born in Kiel, Germany, on April 23, 1858, the son of Julius Wilhelm and Emma (née Patzig) Planck. Planck studied at the Universities of Munich and Berlin, where his teachers included Kirchhoff and Helmholtz, and received his doctorate of philosophy at Munich in 1879. He was Privatdozent in Munich from 1880 to 1885, then Associate Professor of Theoretical Physics at Kiel until 1889, in which year he succeeded Kirchhoff as Professor at Berlin University, where he remained until his retirement in 1926. Afterwards he became President of the Kaiser Wilhelm Society for the Promotion of Science, a post he held until 1937. He was also a gifted pianist and is said to have at one time considered music as a career. Planck was twice married. Upon his appointment, in 1885, to Associate Professor in his native town Kiel he married a friend of his childhood, Marie Merck, who died in 1909. He remarried her cousin Marga von Hösslin. Three of his children died young, leaving him with two sons.*

# IE for Knowledge Base Construction

```
{{Infobox_Scientist
| name = Max Planck
| birth_date = [[April 23]], [[1858]]
| birth_place = [[Kiel]], [[Germany]]
| death_date = [[October 4]], [[1947]]
| death_place = [[Göttingen]], [[Germany]]
| residence = [[Germany]]
| nationality = [[Germany|German]]
| field = [[Physicist]]
| work_institution = [[University of Kiel]]</br>
[[Humboldt-Universität zu Berlin]]</br>
[[Georg-August-Universität Göttingen]]
| alma_mater = [[Ludwig-Maximilians-Universität München]]
| doctoral_advisor = [[Philipp von Jolly]]
| doctoral_students =
[[Gustav Ludwig Hertz]]</br>
...
| known_for = [[Planck's constant]],
[[Quantum mechanics|quantum theory]]
| prizes = [[Nobel Prize in Physics]] (1918)
...
}}
```

Categories: 1858 births | 1947 deaths | German Nobel laureates | German physicists | Members of the Pontifical Academy of Sciences | Members of the Prussian Academy of Sciences | Nobel laureates in Physics | Recipients of the Copley Medal | People from Kiel | People from the Province of Schleswig-Holstein | Quantum physicists | Recipients of the Pour le Mérite (civil class) | Theoretical physicists | Thermodynamicists | University of Munich alumni | University of Munich faculty | Humboldt University of Berlin alumni | Humboldt University of Berlin faculty | University of Kiel faculty | Columbia University faculty | German Christians | Religion and science | Fellows of the Leopoldina

Max Planck



Born	April 23, 1858 Kiel, Holstein
Died	October 4, 1947 (aged 89) Göttingen, West Germany
Nationality	German
Fields	Physics
Institutions	University of Kiel University of Berlin University of Göttingen Kaiser-Wilhelm-Gesellschaft
Alma mater	Ludwig-Maximilians-Universität München
Doctoral advisor	Alexander von Brill
Doctoral students	Gustav Ludwig Hertz Erich Kretschmann Walther Meißner Walter Schottky Max von Laue Max Abraham Moritz Schlick Walther Bothe Julius Edgar Lilienfeld
Known for	Planck's constant Planck postulate Planck's law of black body radiation
Notable awards	Nobel Prize in Physics (1918)
Religious stance	Protestant <sup>[1]</sup>
Notes	He is the father of Erwin Planck who was hanged in 1945 by the Gestapo for his part in the July 20 plot.



# NLP-Based IE on the Web

ANNIE Output for [http://en.wikipedia.org/wiki/Che\\_Guevarra](http://en.wikipedia.org/wiki/Che_Guevarra)

Annotation Key:

**Person** **Location** **Organization** **Date** **Address** **Money** **Percent**

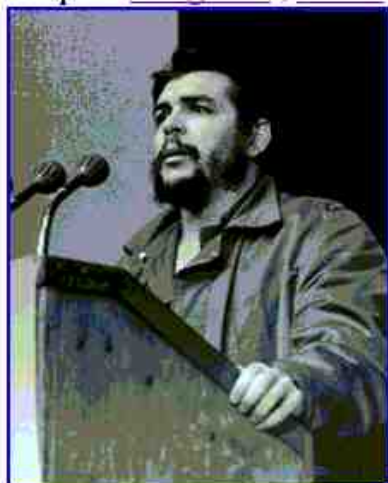
>> /\*\*/ > /\*\*/

## Che Guevara

From Wikipedia, the free encyclopedia.

(Redirected from **Che Guevarra** )

Jump to: [navigation](#) , [search](#)




**Che Guevara**

**Ernesto Rafael Guevara de la Serna** ( **June 14 , 1928** <sup>[1]</sup> ? **October 9 , 1967** ), commonly known as **Che Guevara** or **el Che**, was an **Argentine** -born **Marxist revolutionary** and **Cuban guerrilla leader**. **Guevara** was a member of **Fidel Castro** 's " **26th of July Movement** " that seized power in **Cuba** in **1959** . After serving in various important posts in the new government, **Guevara** left **Cuba** in **1965** with the hope of fomenting revolutions in other countries, first in the Congo-**Kinshasa** (currently the **Democratic Republic of the Congo** ) and later in **Bolivia** , where he was captured in a **CIA** -organized military operation. It is believed by some that the **CIA** wished to keep **Guevara** alive for **interrogation** but, after his capture in the Yuro ravine, he died at the hands of the **Bolivian Army** in **La Higuera** near **Vallegrande** on **October 9 , 1967** . Testimony by various individuals who were participants in, or

<http://services.gate.ac.uk/annie/>

# NLP-Based IE on the Web



[Show RDF](#)[Entry Page](#)

+

-

Topics:

War Conflict94%

Social Tags:

War Conflict☆☆☆

Cuban Revolution☆☆☆

Argentine people☆☆☆

Che Guevara☆☆☆

Granma☆☆☆

26th of July Movement☆☆☆

Guerrilla Warfare☆☆☆

Che☆☆☆

Guerrillero Heroico☆☆☆

Fidel Castro☆☆☆

Marxist theorists☆☆☆

Politics☆☆☆

Argentina☆☆☆

Entities:

+

City

+

Company

+

Continent

+

Country

+

Facility

+

Industry Term

+

Natural Feature

+

Organization

+

Person

+

Position

+

Region

**Ernesto "Che" Guevara** (Spanish pronunciation: [ˈtʃe ye ˈβaɾa];[7] June 14,[1] 1928 – October 9, 1967), commonly known as el **Che** or simply **Che**, was an Argentine Marxist revolutionary, **physician, author, guerrilla leader, diplomat, and military theorist**. A major figure of the Cuban Revolution, **his** stylized visage has become a ubiquitous countercultural symbol of rebellion and global insignia within popular culture.[8]

As a young medical student, **Guevara** traveled throughout **South America** and was radicalized by the poverty, hunger, and disease **he** witnessed.

[9] **His** burgeoning desire to help overturn what **he** saw as the capitalist exploitation of **Latin America** by the **United States** prompted **his** involvement in **Guatemala's** social reforms under **President Jacobo Árbenz**, whose eventual **CIA**-assisted overthrow at the behest of the **United Fruit Company** solidified **Guevara's** political ideology.

[9] Later, in **Mexico City**, **he** met **Raúl** and **Fidel Castro**, joined their 26th of July Movement, and sailed to **Cuba** aboard the yacht, **Granma**, with the intention of overthrowing US-backed Cuban dictator **Fulgencio Batista**. [10] **Guevara** soon rose to prominence among the insurgents, was promoted to second-in-command, and played a pivotal role in the victorious two-year guerrilla campaign that deposed the Batista regime. [11]

Following the Cuban Revolution, **Guevara** performed a number of key roles in the new government. These included reviewing the appeals and firing squads for those convicted as war criminals during the revolutionary tribunals,

[12] instituting agrarian land reform as **minister** of industries, helping spearhead a successful nationwide literacy campaign, serving as both national **bank president and instructional director** for **Cuba's** armed forces, and traversing the globe as **a diplomat** on behalf of Cuban socialism. Such positions also allowed **him** to play a central role in training the militia forces who repelled **the Bay of Pigs** Invasion [13] and bringing **the Soviet nuclear-armed ballistic missiles** to **Cuba** which precipitated the 1962 Cuban Missile Crisis.

[14] Additionally, **he** was **a prolific writer** and diarist, composing a seminal manual on guerrilla warfare, along with a best-selling memoir about **his** youthful continental motorcycle journey. **His** experiences and studying of Marxism–Leninism led **him** to posit that the Third World's underdevelopment and dependence was an intrinsic result of imperialism, neocolonialism, and monopoly capitalism, with the only remedy being proletarian internationalism and world revolution. [15] [16] **Guevara** left **Cuba** in 1965 to foment revolution abroad, first unsuccessfully in **Congo-Kinshasa** and later in **Bolivia**, where **he** was captured by **CIA**-assisted **Bolivian forces** and summarily executed. [17]

**Guevara** remains both a revered and reviled historical figure, polarized in the collective imagination in a multitude of biographies, memoirs, essays, documentaries, songs, and films. As a result of **his** perceived martyrdom, poetic invocations for class struggle, and desire to create the consciousness of a "new man" driven by moral rather than material incentives, **he** has evolved into a quintessential icon of various leftist-inspired movements. **Time** magazine named **him** one of the 100 most influential people of the 20th century, [18] while an Alberto Korda photograph of **him** entitled Guerrillero Heroico (shown), was cited by the **Maryland Institute College of Art** as "the most famous photograph in the world". [19]

<http://www.opencalais.com>

# Extracting Structured Records from the Deep Web

amazon

Try Prime

Your Amazon.com

Today's Deals

Gift Cards

Sell

Help

Shop by Department

Search

All

mining the web

Go

Help

Sign in

Your Account

Try Prime

Cart

Wish List

CYBER MONDAY DEALS WEEK

See the deals

Presented by Amazon.com Rewards Visa Card

Books

Advanced Search

New Releases

Best Sellers

The New York Times® Best Sellers

Children's Books


Textbooks

Sell Your Books


Best Books of the Month

Deals in Books

Click to LOOK INSIDE!



Click to open expanded view



Share your own customer images

Search inside this book

Mining the Web: Discovering Knowledge from Hypertext Data (Hardcover)

Soumen Chakrabarti (Author)

★★★★★ (2 customer reviews)

List Price: \$81.95

Price: **\$62.82** & FREE Shipping. Details

Your Save: \$29.13 (32%)

Only 2 left in stock (more on the way).

Ships from and sold by Amazon.com. Gift-wrap available.

Want it tomorrow, Dec. 4? Order within 4 hrs 55 mins and choose One-Day Shipping at checkout. Details

Ordering for Christmas? To ensure delivery by December 24 choose FREE Shipping at checkout. Read more about holiday shipping.

32 new from \$16.00 36 used from \$17.00

FREE TWO-DAY SHIPPING FOR COLLEGE STUDENTS

MEMBERS

amazon student

Format	Amazon Price	New from	Used from
Kindle Edition	\$54.99	--	--
Hardcover	\$62.82	\$16.00	\$17.00

Buy New

\$62.82

Quantity: 1

Yes, I want FREE Two-Day Shipping with Amazon Prime

Add to Cart

or

Sign in to turn on 1-Click ordering

Add to Wish List

Sell Us Your Item

For a \$1.87 Gift Card

Trade In

Learn more

More Buying Choices

62 used & new from \$17.00

Have one to sell? Sell on Amazon

Share

Get up to 75% Back

Sell Your Books

Get up to 75% back when you sell your books on Amazon. Ship your books for free and get Amazon.com Gift Cards. Learn more.

Product Details

Hardcover: 344 pages

Publisher: Morgan Kaufmann; 1 edition (October 23, 2002)

Language: English

ISBN-10: 1558607544

ISBN-13: 978-1558607545

Product Dimensions: 1 x 7.4 x 9.4 inches

Shipping Weight: 1.4 pounds (View shipping rates and policies)

Average Customer Review: ★★★★★ (2 customer reviews)

Amazon Best Sellers Rank: #810,751 in Books (See Top 100 in Books)

Did we miss any relevant features for this product? Tell us what we missed.

Would you like to update product info, give feedback on images, or tell us about a lower price?




# Extracting Structured Records from the Deep Web

amazon [Try Prime](#) [Your Amazon.com](#) [Today's Deals](#) [Gift Cards](#) [Sell](#) [Help](#) **CYBER MONDAY DEALS WEEK** [See the deals](#) Presented by Amazon.com Rewards Visa Card

Shop by Department [All](#)   [Hello, Sign in Your Account](#) [Try Prime](#) [Cart](#) [Wish List](#)

Books [Advanced Search](#) [New](#)

Click to **LOOK INSIDE** [See a Guide](#)



Click to open expanded view

[Share your own opinion about this book](#) [Search inside this book](#)

Get up to **75% Back** [Sell Your Books](#) Get up to 75% back. [Learn more.](#)

**Product Details**  
Hardcover: 344 pages  
Publisher: Morgan Kaufmann  
Language: English  
ISBN-10: 1558607544  
ISBN-13: 978-1558607544  
Product Dimensions: 1 x  
Shipping Weight: 1.4 pounds  
Average Customer Review: [See all reviews](#)  
Amazon Best Sellers Rank: [See the Best Sellers rank](#)  
Did we miss any relevant information?  
Would you like to [update](#) this information?

```
<div class="buying"><b class="sans">Mining the Web: Analysis of Hypertext and  
Semi Structured Data (The Morgan Kaufmann Series in Data Management  
Systems)  
(Hardcover)</b><br />by  
<a href="/exec/obidos/search-handle-url/index=books&field-author-  
exact=Soumen%20Chakrabarti&rank=-relevance%2C%2Bavailability%2C-daterank/  
102-8395894-5490548">Soumen Chakrabarti</a>  
<div class="buying" id="priceBlock">  
<style type="text/css">  
  td.productLabel { font-weight: bold; text-align: right; white-space:  
nowrap; vertical-align: top; padding-right: 5px; padding-left: 0px; }  
  table.product { border: 0px; padding: 0px; border-collapse: collapse;  
}</style>  
<table class="product">  
  <tr>  
    <td class="productLabel">List Price:</td>  
    <td>$62.82</td>  
  </tr>  
  <tr>  
    <td class="productLabel">Price:</td>  
    <td><b class="price">$62.82</b>  
& this item ships for <b>FREE with Super Saver Shipping</b>.  
...
```


# Extracting Structured Records from the Deep Web

amazon [Try Prime](#) [Your Amazon.com](#) [Today's Deals](#) [Gift Cards](#) [Sell](#) [Help](#) **CYBER MONDAY DEALS WEEK** [See the deals](#) Presented by Amazon.com Rewards Visa Card

Shop by Department [All](#)   [Hello, Sign in Your Account](#) [Try Prime](#) [Cart](#) [Wish List](#)

Books [Advanced Search](#) [New](#)

Click to **LOOK INSIDE** [See a sample](#)



Click to open expanded view

[Share your own reviews](#) [Search inside this book](#)

Get up to **75% Back** [Sell Your Books](#) Get up to 75% back. Cards. [Learn more.](#)

**Product Details**  
Hardcover: 344 pages  
Publisher: Morgan Kaufmann  
Language: English  
ISBN-10: 1558607544  
ISBN-13: 978-1558607544  
Product Dimensions: 1 x  
Shipping Weight: 1.4 pounds  
Average Customer Review:  
Amazon Best Sellers Rank:  
Did we miss any relevant info?  
Would you like to [update](#)

```
<div class="buying"><b class="sans">Mining the Web: Analysis of Hypertext and  
Semi Structured Data (The Morgan Kaufmann Series in Data Management  
Systems)  
(Hardcover)</b><br />by  
<a href="/exec/obidos/search-handle-url/index=books&field-author-  
exact=Soumen%20Chakrabarti&rank=-relevance%2C%2Bavailability%2C-daterank/  
102-8395894-5490548">Soumen Chakrabarti</a>  
<div class="buying" id="priceBlock">  
<style type="text/css">  
td.productLabel { font-weight: bold; text-align: right; white-space:  
nowrap; vertical-align: top; padding-left: 0px; }  
table.product { border: 0px; padding: 0px; border-collapse: collapse;  
}</style>  
<table class="product">  
<tr>  
<td class="productLabel">List Price:</td>  
<td>$62.82</td>  
</tr>  
<tr>  
<td class="productLabel">Price:</td>  
<td><b class="price">$62.82</b>  
& this item ships for <b>FREE with Super Saver Shipping</b>.  
...
```

**Extracted Record:**  
**Title:** Mining the Web  
**Author:** Soumen Chakrabarti  
**Hardcover:** 344 pages  
**Publisher:** Morgan Kaufmann  
**Language:** English

# Jeopardy!

*A big U.S. city with two airports, one named after a World War II hero, and one named after a World War II battle field.*



# Jeopardy!

*A big U.S. city with two airports, one named after a World War II hero, and one named after a World War II battle field.*



# Jeopardy!

*A big U.S. city with two airports, one named after a World War II hero, and one named after a World War II battle field.*

O'Hare International Airport – Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/O%27Hare\_International\_Airport

## History [\[edit\]](#)

### World War II [\[edit\]](#)

*See also: [Illinois World War II Army Airfields](#)*

The airport was constructed in 1942–43 as a manufacturing plant for [Douglas C-54s](#) during [World War II](#).<sup>[12]</sup> The site was chosen for its proximity to the city and transportation.<sup>[12]</sup> The two million square foot (180,000 m²) factory needed easy access to the workforce of the nation's then-second-largest city, as well as its extensive railroad infrastructure. [Orchard Place](#) was a small nearby farming community.<sup>[12]</sup>

Douglas Company's contract ended in 1945 and though plans were proposed to build commercial aircraft, the company ultimately chose to concentrate production on the west coast. With the departure of Douglas, the airport took the name **Orchard Field Airport**, the source of its three-letter IATA code **ORD**.

In 1945, the facility was chosen by the city of Chicago as the site for a facility to meet future aviation demands. Matthew Laflin Rockwell (1915–1988) was the director of planning for the [U.S. Army Corps of Engineers](#) and responsible for the site selection and design of O'Hare International Airport. He was the great grandson of [Matthew Laflin](#), a founder and pioneer of Chicago.

In 1949, the airport was renamed "O'Hare International Airport" to honor [Edward O'Hare](#), the U.S. Navy's first flying ace and [Medal of Honor](#) recipient in [World War II](#). Its IATA code, "ORD", remained unchanged, however, resulting in the infrequent case of an airport's three-letter designation bearing no connection to the airport name or metropolitan area.



Model of "Butch" O'Hare's Grumman F4F-3 Wildcat on display in Terminal 2 of the airport



# Jeopardy!

*A big U.S. city with two airports, one named after a World War II hero, and one named after a World War II battle field.*

Chicago Midway International Airport - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Chicago\_Midway\_International\_Airport

Edit links

## History [\[edit\]](#)

### Early history (1923–1962) [\[edit\]](#)


Originally named **Chicago Air Park**,<sup>[6]</sup> Midway Airport was built on a 320-acre (1.3 km<sup>2</sup>) plot in 1923 with one **cinder** runway that primarily served **airmail** services. In 1926, the city leased the airport for commercial purposes. The airport was designated as Chicago Municipal Airport on December 12, 1927.<sup>[7]</sup> By 1928, the airport had twelve hangars and four runways, lit for night operations.<sup>[8]</sup>

In 1931, a new passenger terminal opened at 62nd St;<sup>[8]</sup> the following year the airport claimed to be the "**World's Busiest**" with over 100,846 passengers on 60,947 flights.<sup>[9]</sup> (The July 1932 Official Aviation Guide shows 206 scheduled airline departures a week.)

The March 1939 OAG shows 47 weekday departures: 13 on United, 13 American, 9 TWA, 4 Northwest, and two each on Eastern, Braniff, **Pennsylvania Central**, and **C&S**.<sup>[10]</sup> New York's airport (Newark, then LaGuardia by the end of 1939) was then the busiest airline airport in the United States, but Midway passed LaGuardia in 1948 and kept the title until 1960.<sup>[8]</sup>

More construction was funded in part by \$1 million from the **Works Progress Administration**; the airport expanded to fill the square mile in 1938–41 after a court ordered the **Chicago and Western Indiana Railroad** to reroute tracks that had crossed the square along the northern edge of the older field.

In July 1949, the airport was renamed after the **Battle of Midway**.<sup>[9]</sup> That year Midway saw 3.2 million passengers; passengers peaked at 10 million in 1959.<sup>[11]</sup> The diagram on the January 1951 C&GS approach chart shows four parallel pairs of runways, all



are's Grumman  
lay in Terminal 2 of

corps of Engineers  
Matthew Laflin, a

ring ace and Medal  
ent case of an

# Structured Knowledge Queries

*A big **U.S. city** with two airports, one **named after a World War II hero**, and one **named after a World War II battle field**.*

```
SELECT DISTINCT ?c WHERE {  
    ?c type City . ?c locatedIn USA .  
    ?a1 type Airport . ?a2 type Airport .  
    ?a1 locatedIn ?c . ?a2 locatedIn ?c .  
    ?a1 namedAfter ?p . ?p type WarHero .  
    ?a2 namedAfter ?b . ?b type BattleField .  
}
```

- Use manually curated **templates** for mapping sentence patterns to structured queries
- Focus on **factoid** and **list questions**

# Deep QA

*William Wilkinson's "An Account of the Principalities of Wallachia and Moldavia" inspired this author's most famous novel*

*This town is known as "Sin City" & its downtown is "Glitter Gulch"*

*As of 2010, this is the only former Yugoslav republic in the EU*

*99 cents got me a 4-pack of Ytterlig coasters from this Swedish chain*



**Question classification & decomposition**



**Knowledge backends**



- Full details: [Ferrucci et al. '10] [Ferrucci et al. '12]



# More IE Applications

- **Comparison shopping & recommendation portals**  
(e.g., consumer electronics, used cars, real estate, pharmacy, etc.)
- **Business analytics** on customer dossiers, financial reports, etc.  
(e.g., how was company X performing in the last 5 years?)
- Market/customer, PR impact, and **media coverage analysis**  
(e.g., how are our products perceived by teenagers?)
- **Job brokering** (applications/resumes, job offers)  
(e.g., how well does the candidate match our desired profile?)
- **Knowledge management** in consulting companies  
(e.g., do we have experience on retail in Brasil?)
- **Knowledge extraction** from scientific literature  
(e.g., which HIV drugs have been found ineffective recently?)

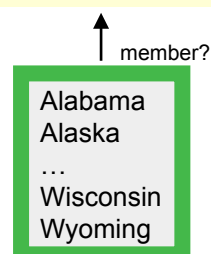
# IE Viewpoints and Approaches

- IE as learning (restricted) **wrappers/regular expressions** (wrapping pages with common structure from Deep Web)
- IE as learning **relations** (rules for identifying instances of  $n$ -ary relation)
- IE as learning **fact boundaries**
- IE as learning **text segmentation** (HMMs, etc.)
- IE as learning **contextual patterns**
- IE as **natural-language analysis** (NLP methods)
- IE as **large-scale text mining** for knowledge acquisition (combinations of tools including web queries)

# IE Viewpoints and Approaches

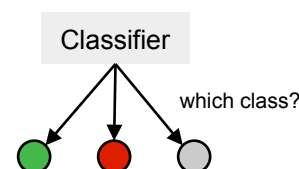
## Lexicons

Abraham Lincoln was born in Kentucky.



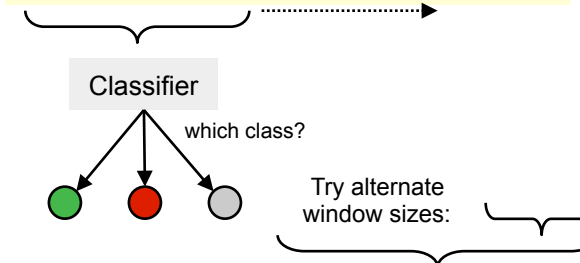
## Classify Pre-Segmented Candidates

Abraham Lincoln was born in Kentucky.



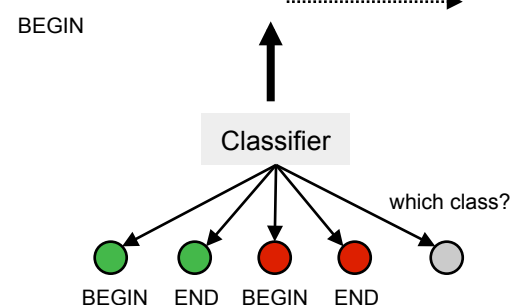
## Sliding Window (+ Classifier)

Abraham Lincoln was born in Kentucky.



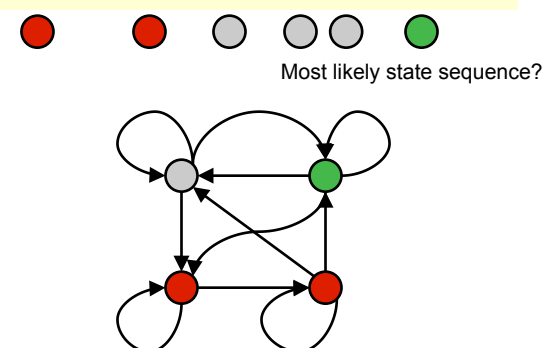
## Boundary Models

Abraham Lincoln was born in Kentucky.



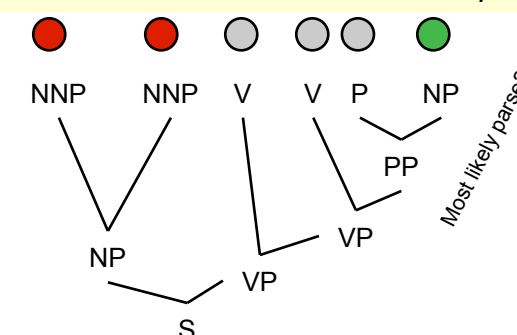
## Finite State Machines

Abraham Lincoln was born in Kentucky.



## Context Free Grammars

Abraham Lincoln was born in Kentucky.



- Source: [Cohen '03]

# IE Quality Assessment

- Fix IE task (e.g., extract all book records from bookseller website)
- Manually extract all correct records
- Use **standard IR effectiveness measures**
  - precision, (relative) recall, F1 measure, etc.
  - **statistical tests** w/ confidence intervals for precision, recall, etc. based on a **sample of manually inspected records**
- Benchmark settings:
  - MUC (Message Understanding Conference), discontinued
  - ACE (Automatic Content Extraction) (<http://www.nist.gov/speech/tests/ace/>)
  - TAC (Text Analysis Conference) (<http://www.nist.gov/tac/>)
  - ...



# Additional Literature for VI.1

- **E. Agichtein:** *Towards Web-Scale Information Extraction*, KDD Webcast 2007, <http://www.mathcs.emory.edu/~eugene/kdd-webinar/>
- **T. Berners-Lee, J. Hendler and O. Lassila:** *The Semantic Web*, Scientific American, 2001
- **H. Cunningham:** *Information Extraction*, Encyclopedia of Language and Linguistics, 2005
- **W. W. Cohen:** *Information Extraction and Integration: An Overview*, KDD 2002, <http://www.cs.cmu.edu/~wcohen/ie-survey.ppt>
- **D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, J. Murdock, E. Nyberg, J. Prager, N. Schlaefer, C. Welty:** *Building Watson: An Overview of the DeepQA Project*. AI Magazine 31(3):59-79, 2010
- **D. Ferrucci:** *Introduction to “This is Watson”*, IBM Journal of Research and Development, 56(3):1, 2012
- **S. Sarawagi:** *Information Extraction*, Foundations & Trends in Databases 1(3), 2008

## VI.2 Natural Language Processing Basics

- **Tokenization** of input documents into
  - **meaningful input units** (e.g., NL sentences, tables, lists, etc.)
  - **input tokens** (e.g., words, phrases, semantic sequences)
  - **token features** (e.g., position in document, capitalization, length, etc.)
- **Linguistic preprocessing** of input documents
  - **part-of-speech tagging** maps words to their grammatical role
  - **chunk parsing** maps a sentence to labeled segments
  - **dependency parsing** identifies logically connected segments
- Both are **important preprocessing steps** for many IE tasks

# Part-of-Speech Tagging

- **Part-of-Speech (PoS) tagging** maps each word (group) to its **grammatical role** (e.g., noun, verb, adjective, determiner, etc.)
- Often uses **Hidden Markov Models** trained on large corpora

- PoS Tags (Penn Treebank):

<b>CD</b>	cardinal number
<b>DT</b>	determiner
<b>EX</b>	existential <i>there</i>
<b>JJ</b>	adjective
<b>NN</b>	noun
<b>POS</b>	possessive ending
<b>PRP</b>	personal pronoun
<b>RB</b>	adverb
<b>VB</b>	verb, base form
<b>WDT</b>	<i>wh</i> -determiner ( <i>which</i> , ...)
<b>WP</b>	<i>wh</i> -pronoun ( <i>who</i> , <i>whom</i> , ...)

...

<http://www.lsi.upc.edu/~nlp/SVMTool/PennTreebank.html>

- Example: *The/DT bright/JJ student/NN who/WP works/VBZ hard/RB will/MD pass/VB all/DT exams/NNS*

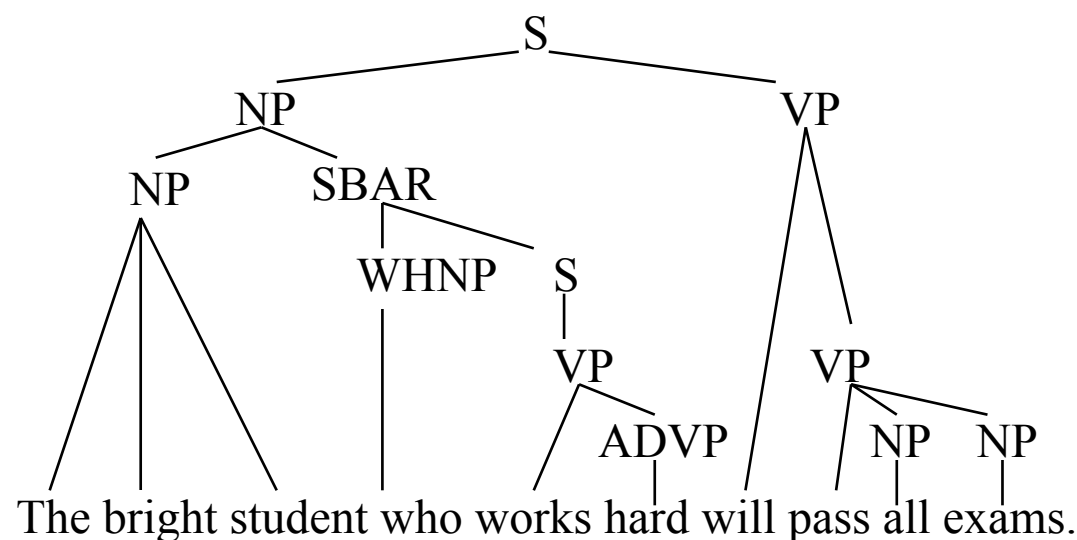
# Word Sense Tagging/Disambiguation

- Idea: Tag each **word** with its **word sense** (meaning, concept) by mapping to a thesaurus/ontology/lexicon such as WordNet
- Typical approach:
  - Form **context**  $con(w)$  **of word**  $w$  in sentence (or passage)
  - Form **context**  $con(s)$  **of candidate sense** (e.g., using the corresponding WordNet synset, gloss, neighboring concepts, etc.)
  - Assign  $w$  to  $s$  with **highest similarity** between  $con(s)$  and  $con(w)$  or **highest likelihood** of  $con(s)$  generating  $con(w)$
  - Incorporate **prior**, i.e., relative frequency of senses for same word
  - **Joint disambiguation**: map multiple words to their most likely meaning (taking into account semantic coherence, compactness)
- Benchmark initiative: <http://www.senseval.org>



# Deep Parsing for Constituent Trees

- Construct syntax-based parse tree of sentence constituents
  - **Non-deterministic context-free grammars** (natural ambiguity)
  - **Probabilistic context-free grammars** (likely vs. unlikely parse trees)



```
(ROOT
  (S
    (NP
      (NP (DT The) (JJ bright) (NN student))
      (SBAR
        (WHNP (WP who))
        (S
          (VP (VBZ works)
            (ADVP (RB hard))))))
      (VP (MD will)
        (VP (VB pass)
          (NP (DT all) (NNS exams))))))
    (VP (VBZ works)
      (ADVP (RB hard))))))
```

- Extensions and variations:
  - **lexical parser**: enhanced with lexical dependencies (e.g., only specific verbs can be followed by two noun phrases)
  - **chunk parser**: simplified to detect only phrase boundaries

# Dependency Parsing

- Reveal dependencies between **logically connected segments**

```
(ROOT
  (S
    (NP
      (NP (DT The) (JJ bright) (NN student))
      (SBAR
        (WHNP (WP who))
        (S
          (VP (VBZ works)
            (ADVP (RB hard))))))
    (VP (MD will)
      (VP (VB pass)
        (NP (DT all) (NNS exams))))))
```

## Typed dependencies:

```
det(student-3, The-1)
amod(student-3, bright-2)
nsubj(passes-7, student-3)
nsubj(works-5, who-4)
rcmod(student-3, works-5)
advmod(works-5, hard-6)
root(ROOT-0, passes-7)
det(exams-9, all-8)
dobj(passes-7, exams-9)
```

- Stanford Dependencies:

<b>nsubj</b>	nominal subject
<b>rel</b>	relative
<b>dobj</b>	direct object
<b>det</b>	determiner
<b>amod</b>	adjectival modifier
<b>rcmod</b>	relative clause modifier
<b>acompl</b>	adjectival complement
<b>advmod</b>	adverbial modifier

...

<http://nlp.stanford.edu/software/stanford-dependencies.shtml>

# Named Entity Recognition (NER)

- Identify mentions of **named entities**  
(e.g., **persons**, **locations**, **organizations**, **dates**, etc.)
- Runs text through part-of-speech tagging or probabilistic parsing
- Uses dictionaries to validate/falsify candidate entities
- Does not disambiguate candidate entities
- Example: *Bayern Munich* with their captain *Philipp Lahm* lost the final in *Munich* on *May 19 2012*

# Coreference Resolution (Anaphor Resolution)

- Connect **pronouns** etc. to **subject/object** of previous sentence.
- Example: *Diego Maradona* was soccer player of the year. *He* is

*also known as the hand of god.*





# Semantic Role Labeling (SRL)

- Identify **semantic types** of events or  $n$ -ary relations based on taxonomy (e.g., FrameNet, VerbNet, PropBank)
- Fill **components** of  $n$ -ary tuples (semantic roles, slots of frames)
- Example: *Thompson is understood to be accused of importing heroin into the United States*

```
<event>
  <type> drug-smuggling </type>
  <destination> <country>United States</country></destination>
  <source> unknown </source>
  <perpetrator> <person> Thompson </person> </perpetrator>
  <drug> heroin </drug>
</event>
```

# FrameNet Representation for SRL

## Smuggling

### Definition:

The words in this frame describe situations in which the **Perpetrator** secretly takes **Goods** into or out of a country or other area which are prohibited by law or on which one has not paid the required duty.

### FEs:

#### Core:

**Goal [Goal]** Goal is the location the Goods end up in.  
**Semantic Type**  
Goal

**Goods [Goods]** The FE Goods is anything (including labor, time, or legal rights) that can a country.

**Path [Path]** The path refers to (a part of the) ground the Goods travel over or to a

**Perpetrator [Perp]** This is the person (or other agent) that illegally takes the goods into or  
**Semantic Type**  
Sentient

**Source [Src]** The source is the location the goods occupy initially before change of lo  
**Semantic Type**  
Source

#### Non-Core:

**Duration [Dur]** The amount of time for which a state holds or a process is ongoing.  
**Semantic Type**  
Duration

**Event []** The unlawful movement of **Goods**.

**Frequency [Freq]** The number of times that a smuggling event occurs.  
Inmates **frequently** **SMUGGLE** marijuana into the prison

**Manner [Man]** A description of the **Event** not covered by more specific FEs, including secondary effects (*quietly, loudly*), and general descriptions comparing events (*the same way*). In most cases, it indicates salient characteristics of a **Perpetrator** that also affect the action (*presumptuously, coldly, deliberately, eagerly, carefully*).  
The rebels had **secretly** **SMUGGLED** in several tonnes of explosives.

**Means [Mns]** An act of the **Perpetrator** which allows them to smuggle the **Goods**.

**Place [Place]** Where the event takes place.  
**Semantic Type**  
Location

**Purpose [Purp]** The action that the **Perpetrator** is trying to accomplish by the act of smuggling.  
We **SMUGGLED** you in here **to try to help** but ...

**Reason [Reas]** The Reason for which an event occurs.

**Time [Time]** When the event occurs.  
**Semantic Type**  
Time

Inherits From: **Committing\_crime**  
Is Inherited By:  
Subframe of:  
Has Subframes:  
Precedes:

- Source: <http://framenet.icsi.berkeley.edu/>

# PropNet Representation for SRL

- Large collection of annotated newspaper articles; roles are simpler (more generic) than FrameNet

Arg0, Arg1, Arg2, ... and ArgM with modifiers

**LOC:** location

**ADV:** general purpose

**MOD:** modal verb

**TMP:** time

**MNR:** manner

**EXT:** extent

**NEG:** negation marker

**CAU:** cause

**PNC:** purpose

**DIR:** direction

- Example: *Revenue edged up 3.4% to \$904 million from \$874 million in last year's third quarter*

[Arg0: Revenue] *increased* [Arg2-EXT: by 3.4%] [Arg4: to \$904 million] [Arg3: from \$874 million] [ArgM-TMP: in last year's third quarter]

- Source: <http://verbs.colorado.edu/~mpalmer/projects/ace.html>

# Stanford CoreNLP

- **Stanford CoreNLP Tools**

- implemented in Java
- wrappers for Python, Ruby, Perl, etc.
- part-of-speech tagging
- dependency parsing
- coreference resolution
- named entity recognition
- sentiment analysis
- models for English, Arabic, Chinese, French, German

**Stanford CoreNLP**

Output format:

Please enter your text here:

Bayern Munich won the finals in Wembley in 2013.

**Part-of-Speech:**

1 Bayern Munich won the finals in Wembley in 2013.

**Named Entity Recognition:**

1 Bayern Munich won the finals in Wembley in 2013.

**Coreference:**

1 Bayern Munich won the finals in Wembley in 2013.

**Basic dependencies:**

1 Bayern Munich won the finals in Wembley in 2013.

<http://nlp.stanford.edu:8080/corenlp/>

- Link: <http://nlp.stanford.edu/downloads/corenlp.shtml>



# NLTK

- **Natural Language Toolkit**

- implemented in Python
- part-of-speech tagging
- dependency parsing
- named entity recognition
- sentiment analysis
- models for English, Chinese, and Spanish

**Tagging, Chunking & Named Entity Recognition with NLTK**

This is a demonstration of **NLTK part of speech taggers** and **NLTK chunkers** using **NLTK 2.0.4**. These taggers can assign part-of-speech tags to each word in your text. They can also identify certain phrases/chunks and named entities.

**Tag and Chunk Text**

Choose tagger/chunker  
Default Tagger & NE Chunker

Enter text  
San Francisco is very foggy.

Enter up to 50000 characters

Tag & Chunk

**Tagged Text**

San/NNP Francisco/NNP is/VBZ very/RB foggy/JJ ./.

**Phrases and Named Entities**

**GPE:**  
San/NNP

**PERSON:**  
Francisco/NNP

<http://text-processing.com/demo/tag/>

- Link: <http://nltk.org>

# Additional Literature for VI.2

- **C. Manning and H. Schütze:** *Foundations of Statistical Natural Language Processing*, MIT Press, 2000
- **D. Jurafsky and J. Martin:** *Speech and Language Processing*, Pearson Prentice Hall, 2008