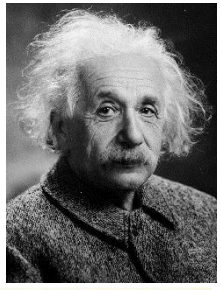
Chapter 15: Information Extraction and Knowledge Harvesting

The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning.

-- Sir Tim Berners-Lee

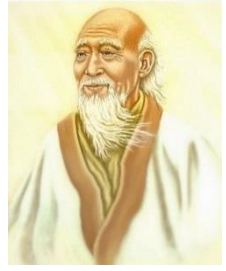


The only source of knowledge is experience.

-- Albert Einstein

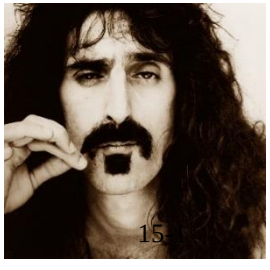
To attain knowledge, add things everyday.
To attain wisdom, remove things every day

-- Lao Tse



Information is not knowledge.
Knowledge is not wisdom.
Wisdom is not truth.
Truth is not beauty.
Beauty is not love.
Love is not music.
Music is the best.

-- Frank Zappa



Outline

15.1 Motivation and Overview 

15.2 Information Extraction Methods  

15.3 Knowledge Harvesting at Large Scale 

8.1 Motivation and Overview

What?

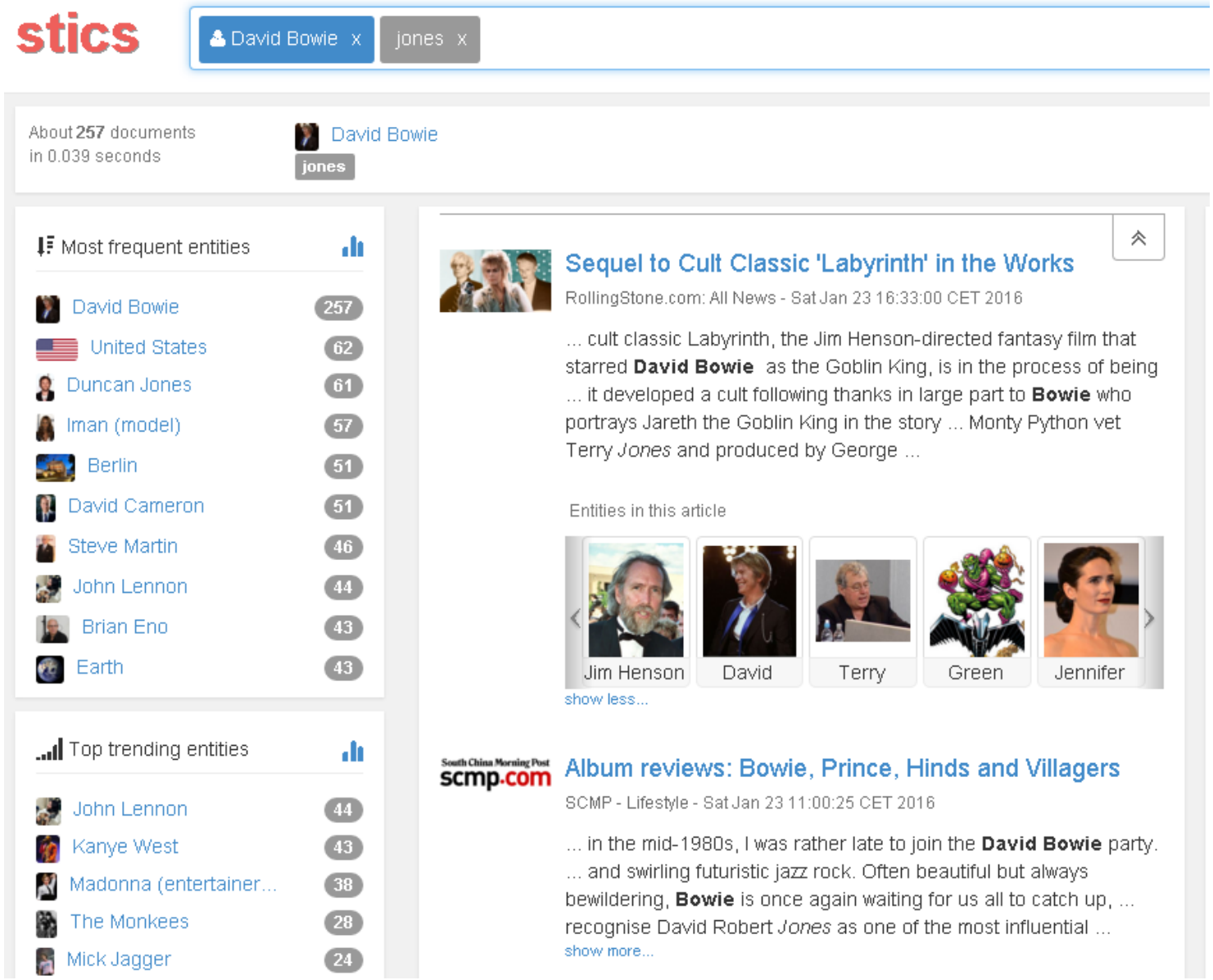
- **extract** entities and attributes from (Deep) **Web sites**
- **mark-up** entities and attributes in **text & Web pages**
- **harvest** relational facts from **the Web** to populate knowledge base

Overall: lift Web and text to level of “crisp“ structured data

Why?

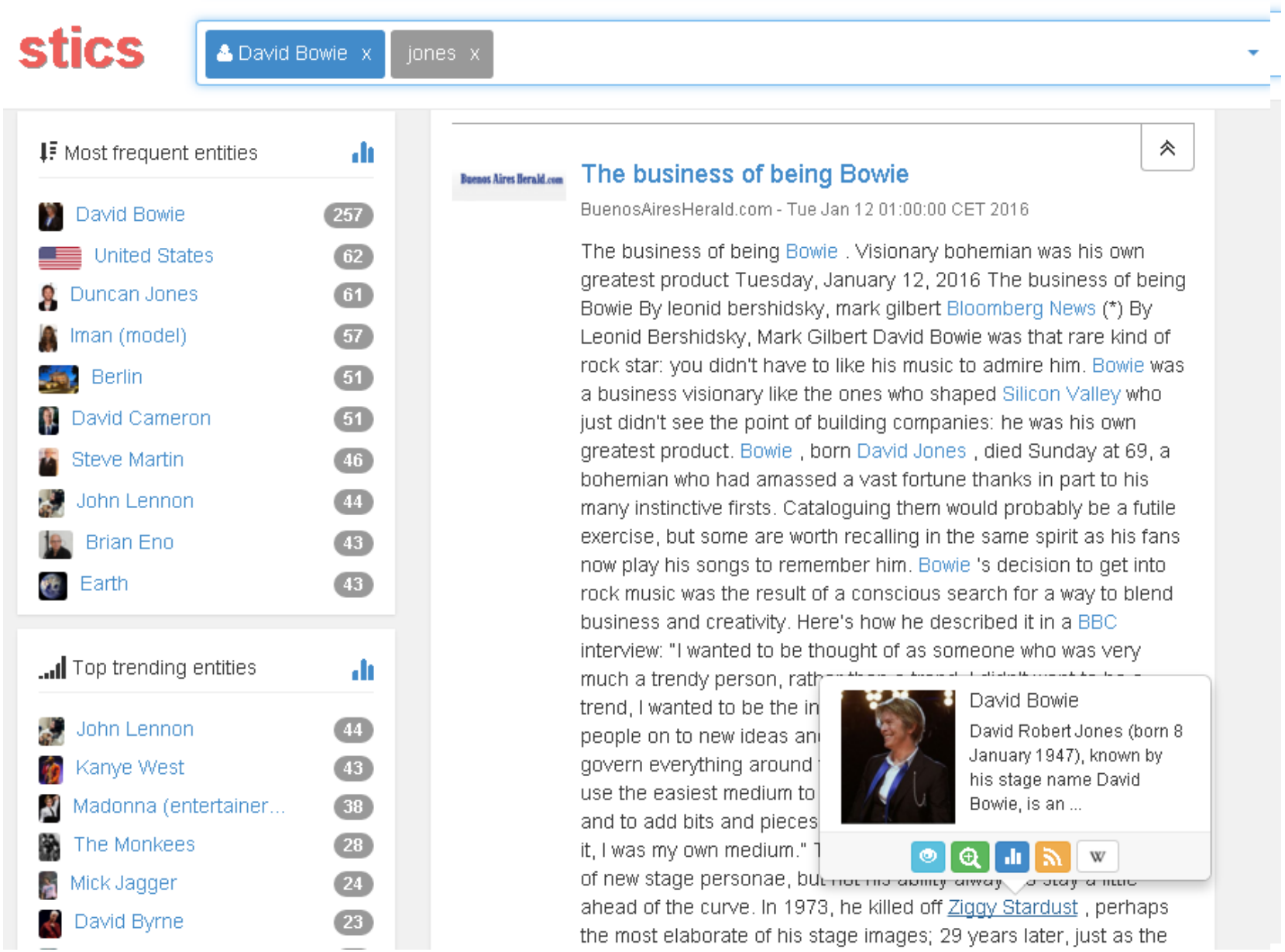
- compare values (e.g. prices) across sites
- extract essential info fields (e.g. job skills & experience from CV)
- more precise queries:
  - semantic search with/for “things, not strings“
  - question answering and fact checking
- constructing comprehensive knowledge bases
- sentiment mining (e.g. about products or political debates)
- context-aware recommendations
- business analytics

Use-Case Example: News Search



http:/stics.mpi-inf.mpg.de

Use-Case Example: News Search



http:/stics.mpi-inf.mpg.de

Use-Case Example: Biomedical Search

MEDIE — See *what causes cancer?*

subject	verb	object	search	clear	stop
	cause	colon cancer	advanced search		

Results **1-10** for **cause colon cancer**
» show query

Sort by Rank Date Sort

sentence | *article* | *table* show 10 results subject verb object **gene** **disease**

show next »

1. **DIXDC1** targets **p21** and **cyclin D1** via PI3K pathway activation to promote **colon cancer** cell proliferation. »XML
   Lei Wang, Xi-Xi Cao, Qi Chen, Teng-Fang Zhu, Hong-Guang Zhu, Li Zheng, pp. 1801-8, Volume 100, Issue 10, Cancer science, 2009 [PMID:19572978]

   Both siRNA knockdown of **DIXDC1** and blocking the PI3K pathway using a specific inhibitor caused G1/S phase arrest, as well as down-regulation of **cyclin D1** and up-regulation of **p21** in **DIXDC1** over-expression colon cancer cells. »XML

2. Induction and down-regulation of **Sox17** and its possible roles during the course of gastrointestinal tumorigenesis. »XML
   Yu-Chen Du, Hiroko Oshima, Keisuke Oguma, Takanori Kitamura, Hiraku Itadani, Takashi Fujimura, Ying-Shi Piao, Tanihiro Yoshimoto, Toshinari Minamoto, Hidehito Kotani, Makoto M Taketo, Masanobu Oshima, pp. 1346-57, Volume 137, Issue 4, Gastroenterology, 2009 [PMID:19549530]

   BACKGROUND & AIMS: The activation of **Wnt/beta-catenin** signaling causes the development of gastric and colon cancers. »XML

3. A **colorectal cancer** expression profile that includes **transforming growth factor beta inhibitor BAMBI** predicts metastatic potential. »XML
   Johannes Fritzmann, Markus Morkel, Daniel Besser, Jan Budczies, Frauke Kosel, Felix H Brembeck, Ulrike Stein, Iduna Fichtner, Peter M Schlag, Walter Birchmeier, pp. 165-75, Volume 137, Issue 1, Gastroenterology, 2009 [PMID:19328798]

   In mice, overexpression of **BAMBI** caused colon cancer cells to form tumors that metastasized more frequently to liver and lymph nodes than control cancer cells. »XML

4. Activation of aminoimidazole carcinogens by nitrosation: mutagenicity and nucleotide adducts. »XML
   Terry V Zenser, Vijaya M Lakshmi, Herman A J Schut, Hui-jia Zhou, P David Josephy, pp. 109-15, Volume 673, Issue 2, Mutation research, 2009 [PMID:19449459]

   2-Amino-3-methylimidazo[4,5-f]quinoline (IQ) and 2-amino-3,8-dimethylimidazo[4,5-f]quinoxaline(MeIQx) are heterocyclic amines (HCAs) derived from high temperature cooking of meat and thought to cause colon cancer in humans. »XML

5. Inhibition of **human sirtuins** by in situ generation of an acetylated lysine-ADP-ribose conjugate. »XML
   Tomomi Asaba, Takayoshi Suzuki, Rie Ueda, Hiroki Tsumoto, Hidehiko Nakagawa, Naoki Miyata, pp. 6989-96, Volume 131, Issue 20, Journal of the American Chemical Society, 2009 [PMID:19413317]

   Compound 2k also caused a dose-dependent increase of **p53** acetylation in human colon cancer HCT116 cells, indicating inhibition of **SIRT1** in the cells. »XML

6. Mismatch repair polymorphisms and risk of colon cancer, tumour microsatellite instability and interactions with lifestyle factors. »XML
   P T Campbell, K Curtin, C M Ulrich, W S Samowitz, J Bigler, C M Velicer, B Caan, J D Potter, M L Slattery, pp. 661-7, Volume 58, Issue 5, Gut, 2009 [PMID:18523027]

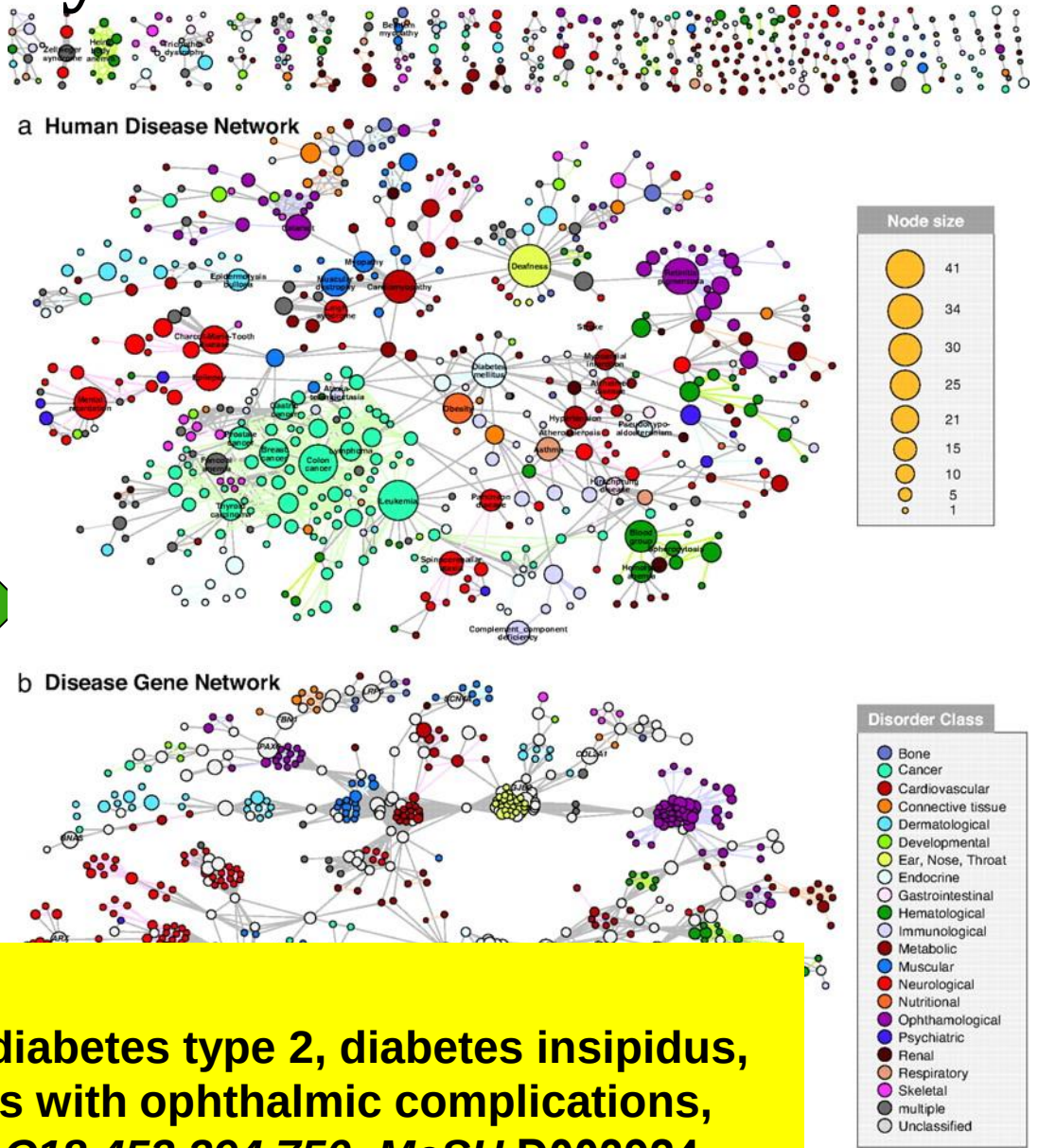
   BACKGROUND: Germline mutations in DNA mismatch repair (MMR) genes cause lynch syndrome colon cancers. »XML

7. Preclinical development of the **nicotinamide phosphoribosyl transferase** inhibitor prodrug GMX1777. »XML
   Pierre Beauparlant, Dominique Bédard, Cynthia Bernier, Helen Chan, Karine Gilbert, Daniel Goulet, Michel-Olivier Gratton, Manon Lavoie, Anne Roulston, Emilie Turcotte, Mark Watson, pp. 346-54, Volume 20, Issue 5, Anti-cancer drugs, 2009 [PMID:19369827]

   Consistent with the requirement for a prolonged exposure for cytotoxicity in vitro, a dose of 75 mg/kg of GMX1777 administered as two bolus intravenous injections in 1 day were not effective at reducing the growth of multiple myeloma (IM-9) tumors, over a 24 h intravenous infusion caused tumor regression in the IM-9 model, a small-cell lung cancer (**SHP-77**) model, and a colon carcinoma (HCT-116) model. »XML

http://www.nactem.ac.uk/medie/search.cgi

Use-Case Text Analytics: Disease Networks



But not so easy with:

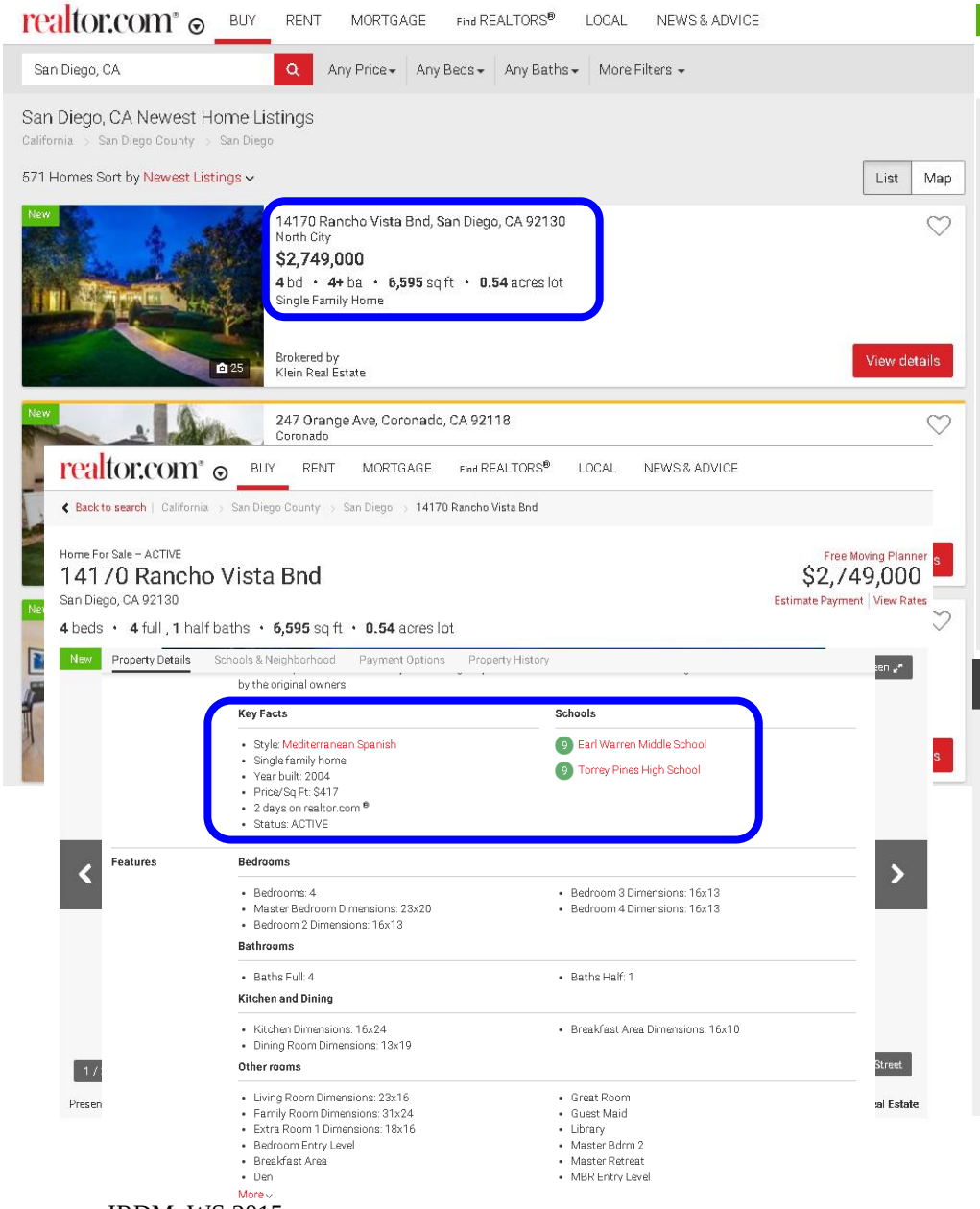
diabetes mellitus, diabetis type 1, diabetes type 2, diabetes insipidus, insulin-dependent diabetes mellitus with ophthalmic complications, ICD-10 E23.2, OMIM 304800, MeSH *C18.452.394.750, MeSH* D003924, …

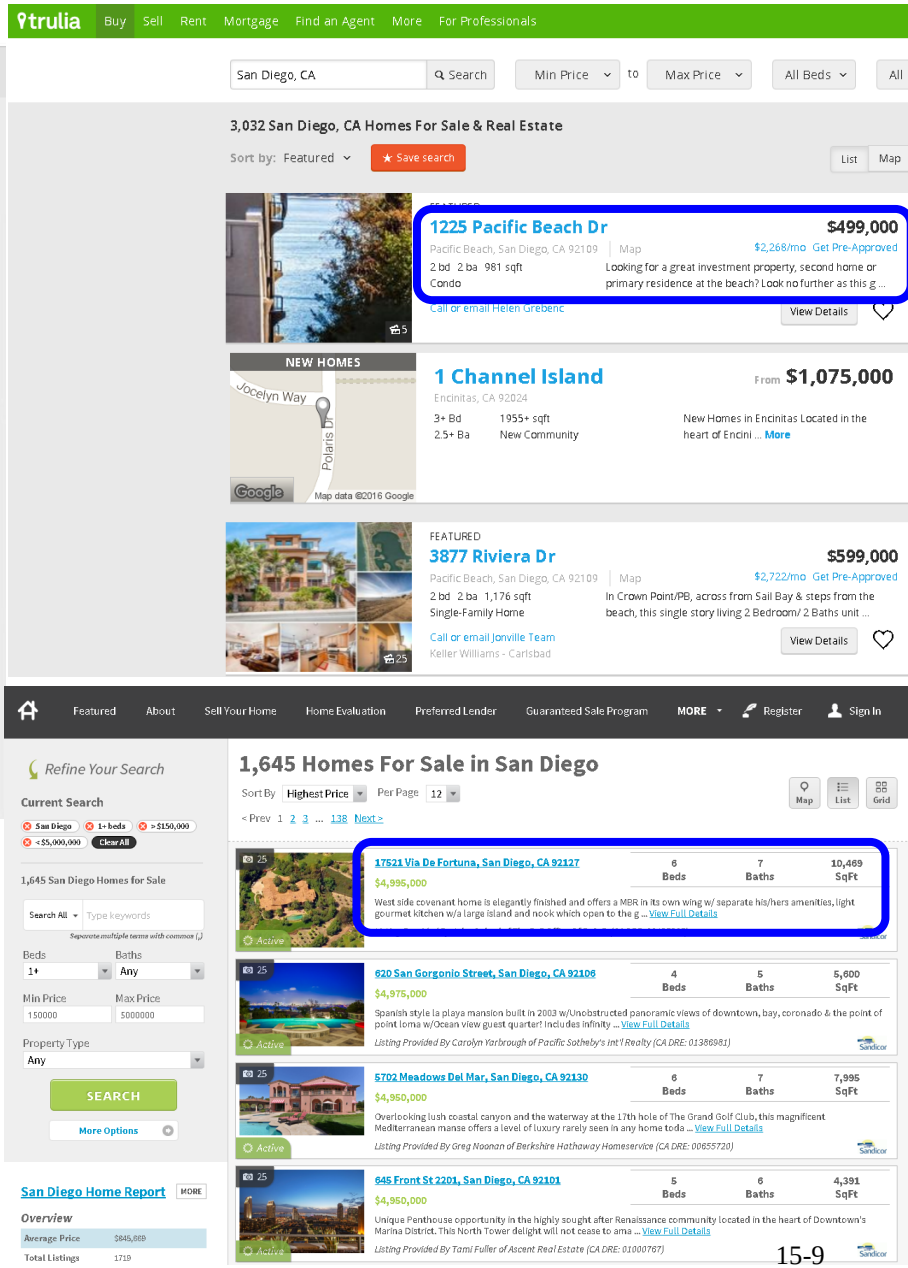
K.Goh,M.Kusick,D.Valle,B.Childs,M.Vidal,A.Barabasi: The Human Disease Network, PNAS, May 2007

Methodologies for IE

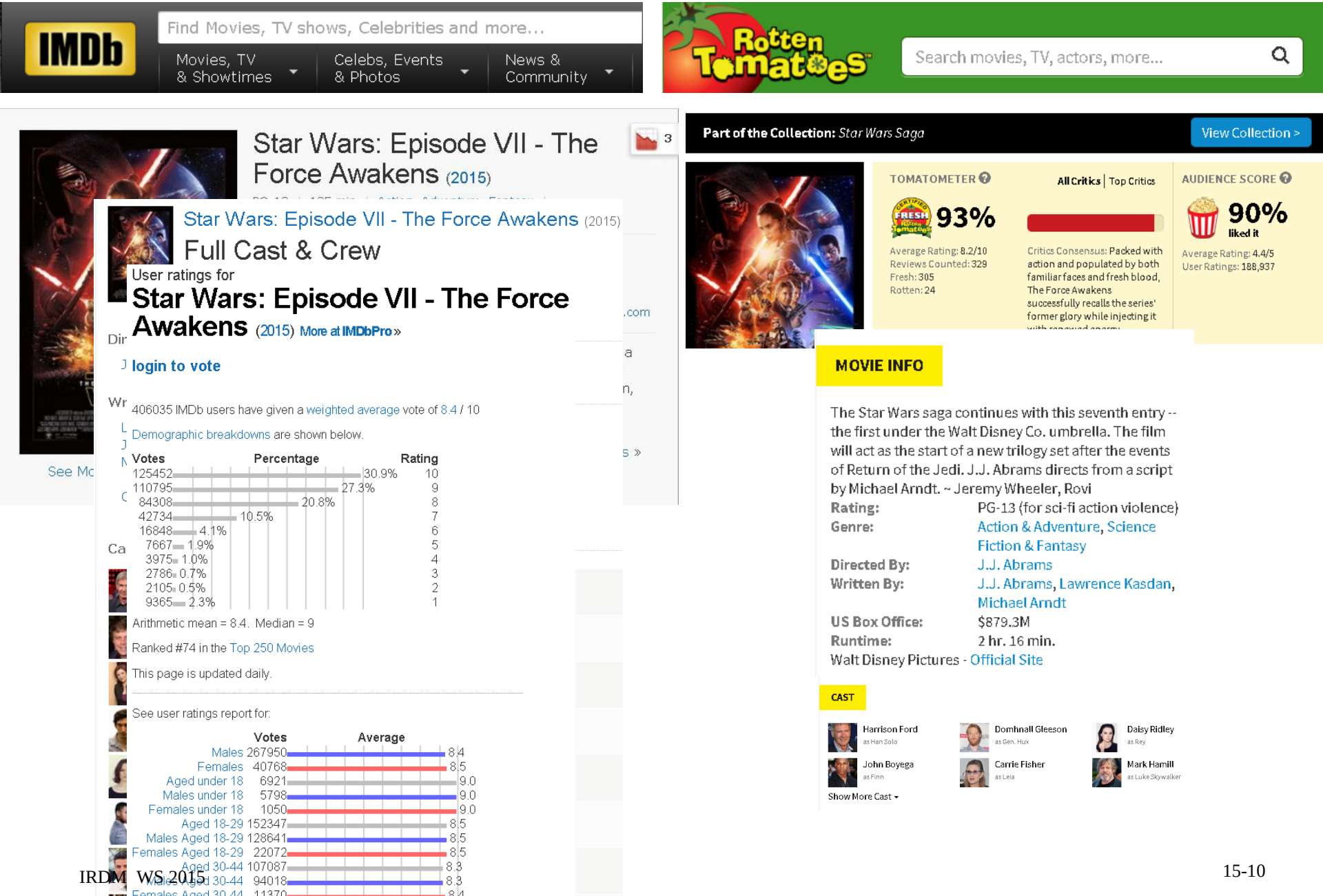
- **Rules & patterns**, especially **regular expressions**
- Pattern matching & pattern learning
- **Distant supervision** by dictionaries, taxonomies, ontologies etc.
- **Statistical machine learning**: classifiers, HMMs, CRFs etc.
- **Natural Language Processing (NLP)**: POS tagging, parsing, etc.
- Text mining algorithms in general

IE Example: Web Pages to Entity Attributes





IE Example: Web Pages to Entity Attributes



IE Example: Text to Opinions on Entities

IMDb

Find Movies, TV shows, Celebrities and more...

Movies, TV & Showtimes | Celebs, Events & Photos | News & Community

Reviews & Ratings for

Star Wars: Episode VII - The Force Awakens

More at IMDbPro »

Filter: Loved It | Hide Spoilers:

Interleaved...

Reviews from users who voted this title more than 8.4.

Reviews from users who voted this title more than 8.4.

Page 4 of 98: [1] [2] [3] **[4]** [5] [6] [7] [8] [9] [10] [11] [12] [13] [14]

Index 973 matching reviews (3317 reviews in total)

7 out of 26 people found the following review useful:

New ... But still Star Wars

Author: chris246 from United Kingdom

23 December 2015

I feel like this is brand new, with new actors and ideas being thrown in the mix. Yet it still feels like Star Wars. This burst of new life into Star Wars is defiantly worth a watch despite complaints of similarities to episode IV, whether a Star Wars fan or not.

I love how they have brought back the older characters (e.g Han Solo, Leia, ...) To pass on the mantle of Star Wars to some new characters. As much as I love this film it would be nice to have some more light sabre fight scenes but I feel as though this can be overlooked due to the fact that this is the first instalment of the series and brings you up to pace with the current affairs of a galaxy far far away a long time ago.

Was the above review useful to you? Yes No

6 out of 25 people found the following review useful:

A-Mazing!!!

Author: Alex Rivett from Peterborough, England

22 December 2015

****** This review may contain spoilers ******

Being a massive Star Wars fan, I went along to the midnight launch in 3D (UK) and I was completely blown away from start to finish! I didn't get to sleep until 4am and I was due in work later that morning... I've been back to see it again, it was that good!

But I have to say this movie had all of the ingredients you need for a major blockbuster and I salute you Mr Abrams for putting together a stunning masterpiece for all of us fans. Not only has it re- energized the Star Wars universe/franchise but it has introduced us to some exciting new characters in Rey, Finn, Kylo Ren and BB-8 - to name a few as well as reuniting us with the faces from the past! Can I just mention that Daisy Ridley is truly fantastic in her role as Rey, she has made me fall in love with her character and I now have a new Star Wars character fave!

Rotten Tomatoes

Search movies, TV, actors, more...

AUDIENCE REVIEWS FOR *STAR WARS: EPISODE VII - THE FORCE AWAKENS*

★★★★

Thought the scriptwriting was brilliant, but I didn't enjoy the actual sets and scenes so much.

Letitia Lew
★ Super Reviewer

★★★

Rey, a young smuggler, is thrust into a battle between the First Order and the resistance when she teams up with a storm trooper who suffered a crisis of conscience.
The new entry into the Star Wars universe is profoundly derivative, essentially an updated retelling of A New Hope, and while ignoring the backstory about the First Order largely mutes

Read More

Jim Hunter
★ Super Reviewer

★★★★★

Extraordinarily faithful to the tone and style of the originals, The Force Awakens brings back the Old Trilogy's heart, humor, mystery, and fun. Since it is only the first piece in a new three-part journey it can't help but feel incomplete. But everything that's already there, from the stunning visuals, to the thrilling action sequences, to the charismatic new characters,

Read More

Matthew Samuel Mirliani
★ Super Reviewer

★★★½

JJ Abrams is very good and knowing what his audience wants and giving just that to them. He is not great, however, because he rarely shows us something we didn't know we wanted. This film derives a lot from the first Star Wars, and just goes along as you might expect, yet it is still very enjoyable because it's Star Wars. The old faces were cool to see, and the new ones do

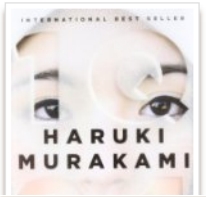
Read More

Sanjay Rema
★ Super Reviewer

View All Audience Reviews (13449) ›

IE Example: Web Pages to Facts & Opinions

LibraryThing Home Groups Talk Zeitgeist



1Q84

by Haruki Murakami

Other authors: See the other authors section.
Series: 1Q84 (1-3)

Common Knowledge

You must log in to edit Common Knowledge data.
For more help see the Common Knowledge help page.

Series (with order)	1Q84 (3.5|1-3)
Canonical title	1Q84
Original title	1Q84 (ichi-kyû-hachi-yon)
Alternative titles	
Original publication date	2009 (vol 1-2) 2010 (vol 3) 2011-10-25 (1-3 · English)
People/Characters	Masami Aomame Tengo Kawana Toshiharu Ushikawa Tamaru Eriko Fukada (show all 9 items)
Important places	Tokyo, Japan
Important events	
Related movies	
Awards and honors	Man Asian Literary Prize Longlist (2011) New York Times bestseller (Fiction, 2011)

Member: g33kgrrl

Collections	Your library (1,609), Read but unowned (19), Chicago (13), Cookbooks (3), Crafty (7), Textbooks (33), Currently reading (9), To read (67), New To Read (189), Favorites (176), On Loan (3), Nursing (40), Sold/Given Away (135), Ebook (57), Library Books (6), Finish unread books (207), Will's books (2), Children's Books (15), Zines (1), All collections (1,761)
Reviews	156 reviews
Tags	favorites (184), comic (173), non-fiction (114), read2008 (97), gift (97), read2007 (95), read2006 (88), read2013 (86), read2009 (83), read2010 (71) — see all tags
Media	Not set (12), Book (1,744), Paper Book (1,642), Audiobook (2), Ebook (45), Other (5), Software (1)
Clouds	tag cloud, author cloud, tag mirror
About me	Once when I was very young, I built a chair out of my Baby Sitter's Club books. I then sat in it to read more books. I've since given up on making furniture out of books, but my dream is to someday have a library with in-wall shelves.
Groups	All Things Discworldian - The Guild of Pratchett Fans, Chicagoans
Venues	Favorites | Visited
Favorite bookstores	Book Loft, Brookline Booksmith, Chicago Comics, City Lit Books, Myopic Books, Uncharted Books
Favorite libraries	Billy Ireland Cartoon Library & Museum (The Ohio State University), Chicago Public Library - Logan Square Branch
Also on	Flickr, Last.fm, Ravelry, Twitter
Membership	LibraryThing Early Reviewers/Member Giveaway
Location	Chicago IL
Favorite authors	Not set
Account type	public, lifetime
URLs	/profile/g33kgrrl (profile) /catalog/g33kgrrl (library)
Member since	Nov 18, 2005
Currently reading	House of Leaves: The Remastered Full-Color Edition by Mark Z. Danielewski The Complete Calvin and Hobbes (Calvin & Hobbes) by Bill Watterson Just an Ordinary Day: The Uncollected Stories Of Shirley Jackson by Shirley Jackson American Lion: Andrew Jackson in the White House by Jon Meacham The Nibelungenlied: Prose Translation (Penguin Classics) by Anonymous show all (9)

Member revie

English (195) Sp

Showing 1-5 of 195

I love Murakar

Jen.ODriscoll.Le

I love Murakar

Jen.ODriscoll.Le

I liked the sto
a lot of unnec
audio so I had
★★★★)

eadieburke | Ja

5 stars and a
as dealing wit
about Aomam
keep me think

sashinka | Jan 1

I actually thou
character for
descriptions o
an editor in ge

g33kgrrl | Jan 6, 2016 |

Showing 1-5 of 195 (next | show all)

Published reviews

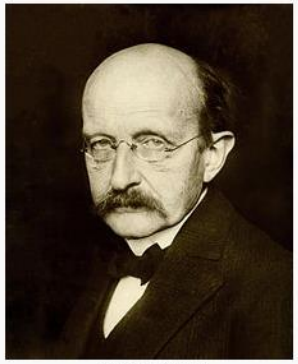
13 reviews add a review

Murakami name-drops George Orwell's laugh-riot 1984 several times. Both books deal with the concept of manipulated realities. And while Murakami's book
as long, it's also more fun to read.

Weekly Alibi, John Bear (Jan 26, 2012)

IE Example: Web Pages to Facts on Entities

Max Planck



Planck in 1933

Born	Max Karl Ernst Ludwig Planck 23 April 1858 Kiel, Duchy of Holstein
Died	4 October 1947 (aged 89) Göttingen, Lower Saxony, Germany
Nationality	German
Fields	Physics

```
{{Redirect|Planck}}

{{Infobox scientist
| name  = Max Planck
| image  = Max Planck 1933.jpg
|caption        = Planck in 1933
| image_size = 225px
|birth_name= Max Karl Ernst Ludwig Planck
| birth_date = {{birth date|df=yes|1858|4|23}}
| birth_place = [[Kiel]], [[Duchy of Holstein]]
| death_date = {{death date and age|df=yes|1947|10|4|1858|4|23}}
| death_place = [[Göttingen]], [[Lower Saxony]], [[Allied-occupied Germany|Germany]]
| spouse      = Marie Merck (1887-1909)<br />{{nowrap|Marga von Hösslin (1911-1947)}}
| nationality = [[Germany|German]]
| religion        = [[Lutheranism]]<ref>[http://books.google.com/books?id=d5zKH2Bx2AwC&pg=PA183
of an upright man: Max Planck and the fortunes of German science'']. J. L. Heilbron. [[Harvard
| field  = [[Physics]]
| alma_mater = [[Ludwig Maximilian University of Munich]]
| work_institutions = {{Plainlist|
* {{nowrap|[[University of Kiel]]}}
* {{nowrap|[[University of Göttingen]]}}
* {{nowrap|[[Kaiser Wilhelm Society]]}}
}}
| doctoral_advisor  = [[Alexander von Brill]]<br>[[Gustav Kirchhoff]]<ref>Physics Tree profile
[[Hermann von Helmholtz]]
```

	Max Planck Medal (1929) Goethe Prize (1945)
Spouse	Marie Merck (1887–1909) Marga von Hösslin (1911–1947)

Stephen Hawking

Hawking at NASA, 1980s

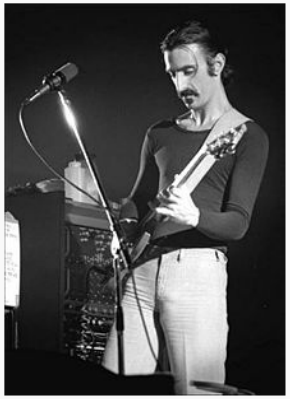
Born	Stephen William Hawking 8 January 1942 (age 74) Oxford, England
Residence	United Kingdom
Known for	Hawking radiation Penrose–Hawking theorems *A Brief History of Time* (1988)

David Bowie

Bowie during the Heathen Tour in Chicago, August 2002

Born	David Robert Jones 8 January 1947 Brixton, London, England
Died	10 January 2016 (aged 69) Manhattan, New York, United
	John Lennon · Queen · Nine Inch Nails · Mick Jagger · Tina Turner · Mott the Hoople
Website	davidbowie.com

Frank Zappa



Zappa performing in Ekeberghallen, Oslo, on January 16, 1977

Born	Frank Vincent Zappa December 21, 1940 Baltimore, Maryland, U.S.
Died	December 4, 1993 (aged 52) Los Angeles, California, U.S.
Instruments	Vocals · guitar · bass guitar · keyboards · drums · percussion
Labels	Verve · Bizarre · Straight · DiscReet · Zappa · Barking

```
{{redirect|Zappa}}
{{pp-move-indef|small=yes}}
{{Use mdy dates|date=September 2015}}
{{Infobox person
| name = Frank Zappa
| image = Zappa 16011977 01 300.jpg
| caption = Zappa performing in [[Ekeberghallen]], Oslo, on January 16, 1977
| birth_name = Frank Vincent Zappa
| birth_date = {{birth date|1940|12|21|mf=y}}
| birth_place = [[Baltimore, Maryland]], U.S.
| death_date = {{death date and age|1993|12|4|1940|12|21|mf=y}}
| death_place = [[Los Angeles]], California, U.S.
| death_cause = [[Prostate cancer]]
| monuments =
| education = [[Chaffey College]]
| occupation = {{flatlist|
*Musician
*songwriter
*composer
*conductor
*record producer
*activist
*filmmaker
*actor
*author
```

Zhang Yimou

Zhang at the 15th Busan International Film Festival

Chinese name	張藝謀 (traditional)
Chinese name	张艺谋 (simplified)
Pinyin	Zhāng Yìmóu (Mandarin)
Origin	China
Born	14 November 1951 (age 64) Xi'an, Shaanxi, China
Occupation	Film director, producer, cinematographer and actor
Spouse(s)	Xiao Hua (肖华) (m. 1978–88) (m. 2011)

ilm Festival

	, China
	965 (age 50) Shenyang, Liaoning, China
Occupation	Actress
Years active	1987–present
Spouse(s)	Ooi Hoe Soeng (1996–2010)

IE Example: 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 1 He was Privatdozent in Munich from 1880 to 1885, the 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 retiren Afterwards he became President of the Kaiser Wilhel for the Promotion of Science, a post he held until 193

He was also a gifted pianist and is said to have at one considered music as a career.
Planck was twice married. Upon his appointment, in to Associate Professor in his native town Kiel he married a friend of his childhood, Marie Merck, w in 1909. He remarried her cousin Marga von Hösslin. Three of his children died young, leaving him with tw

bornOn (Max Planck, 23 April 1858)
bornIn (Max Planck, Kiel)
type (Max Planck, physicist)
advisor (Max Planck, Kirchhoff)
advisor (Max Planck, Helmholtz)
AlmaMater (Max Planck, TU Munich)
plays (Max Planck, piano)
spouse (Max Planck, Marie Merck)
spouse (Max Planck, Marga Hösslin)

Person	BirthDate	BirthPlace	...
Max Planck	4/23, 1858	Kiel	
Albert Einstein	3/14, 1879	Ulm	
Mahatma Gandhi	10/2, 1869	Porbandar	

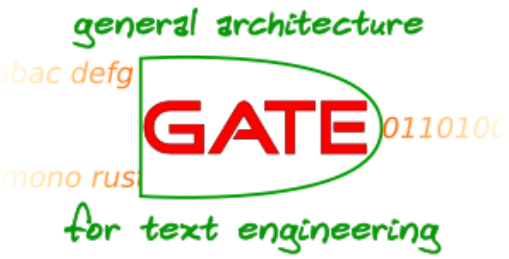
Person	Award
Max Planck	Nobel Prize in Physics
Marie Curie	Nobel Prize in Physics
Marie Curie	Nobel Prize in Chemistry

IE Example: Text to Annotations

Named Entity Recognition with ANNIE

GATE is an open source infrastructure for developing and deploying software components that process human language. GATE excels at text analysis of all shapes and sizes. From large corporations to small startups, from multi-million research consortia to undergraduate projects. More than €5 million has been invested in GATE development and our objective is to make sure that this continues to be money well spent for all GATE's users.

GATE is distributed with an example Information Extraction system, known as ANNIE, which has formed the basis of many commercial and research systems. While ANNIE is capable of recognising a number of different entity types this simple demo focuses on the annotion of people, locations, and organizations.



To try the demo please enter either a URL:

or some free text to process:

```
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.
```

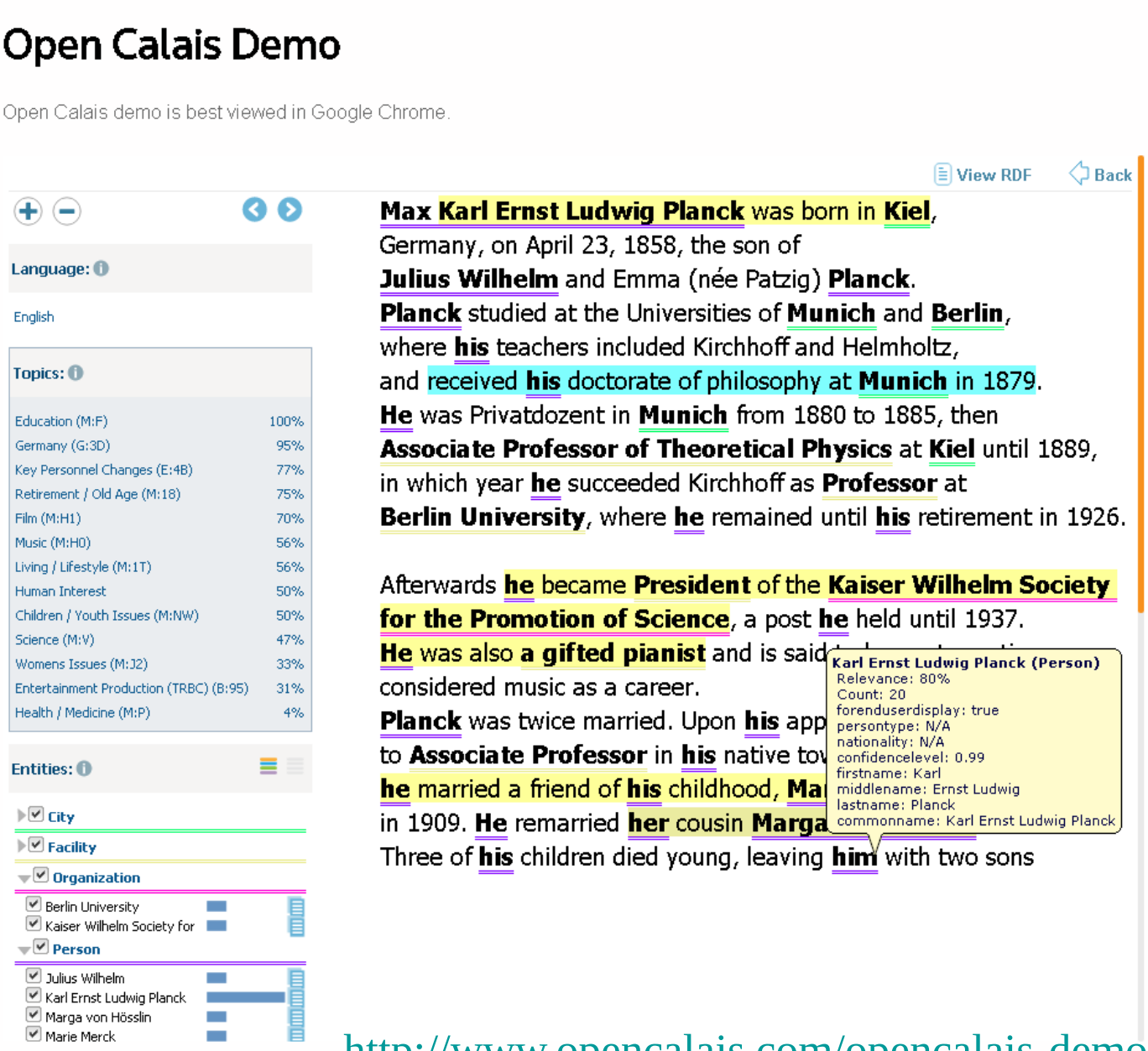
↓ Process Text ↓

Max Karl Ernst Ludwig Planck was born in Kiel, Germany, on April 23, 1858, the son of Julius Wilhelm and Emma (ne 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

---

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

IE Example: Text to Annotations



http://www.opencalais.com/opencalais-demo/

Info Extraction vs. Knowledge Harvesting

Surajit obtained his PhD in CS from Stanford University under the supervision of Prof. Jeff Ullman. He later joined HP and worked closely with Umesh Dayal …

source-centric IE

1) recall !
2) precision

instanceOf (Surajit, scientist)
inField (Surajit, computer science)
hasAdvisor (Surajit, Jeff Ullman)
almaMater (Surajit, Stanford U)
workedFor (Surajit, HP)
friendOf (Surajit, Umesh Dayal)
…

one source

- targeted: hasAdvisor, almaMater
- open: worked for, affiliation, employed by, romance with, affair with, …

many sources

yield-centric harvesting

1) precision !
2) recall

hasAdvisor

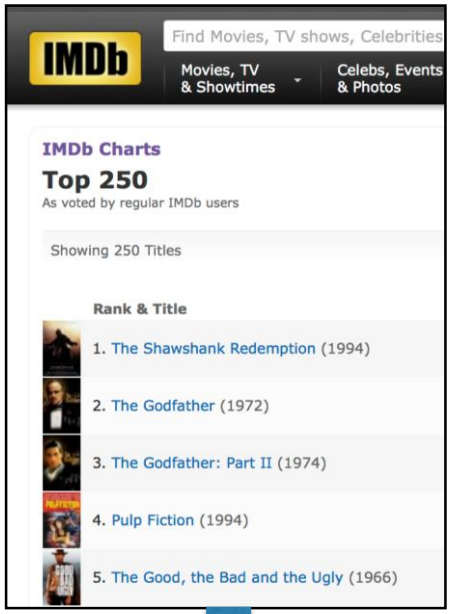
Student	Advisor
Surajit Chaudhuri	Jeffrey Ullman
Alon Halevy	Jeffrey Ullman
Jim Gray	Mike Harrison
…	…

almaMater

Student	University
Surajit Chaudhuri	Stanford U
Alon Halevy	Stanford U
Jim Gray	UC Berkeley

15.2.1 IE with Rules on Patterns (aka. Web Page Wrappers)

Goal: Identify and extract entities and attributes in regularly structured HTML page, to generate database records



Rule-driven regular expression matching

- regex over alphabet Σ of tokens:
  ε, $\sigma \in \Sigma$, (expr1|expr2), (expr)*
- Interpret pages from same source (e.g. Web site to be wrapped) as regular language (FSA, Chomsky-3 grammar)
- Specify rules by regex's for detecting and extracting attribute values and relational tuples

Title	Year
The Shawshank Redemption	1994
The Godfather	1972
The Godfather - Part II	1974
Pulp Fiction	1994
The Good, the Bad, and the Ugly	1966

LR Rules: Left and Right Tokens

L token (left neighbor) ***pre-filler pattern*** fact token ***filler pattern*** **R token** (right neighbor) ***post-filler pattern***

Example:

L = <B>, **R** = </B>

→ ***MovieTitle***

L = <I>, **R** = </I>

→ ***Year***

```
<HTML>
    <TITLE>Top-250 Movies</TITLE>
<BODY>
    <B>Godfather 1</B><I>1972</I><BR>
    <B>Interstellar</B><I>2014</I><BR>
    <B>Titanic</B><I>1997</I><BR>
</BODY>
</HTML>
```

produces relation with tuples: <Godfather 1, 1972>, <Interstellar, 2014>, <Titanic, 1997>

Rules can be combined and generalized

→ RAPIER [Califf and Mooney '03]

Advanced Rules: HLRT, OCLR, NHLRT, etc

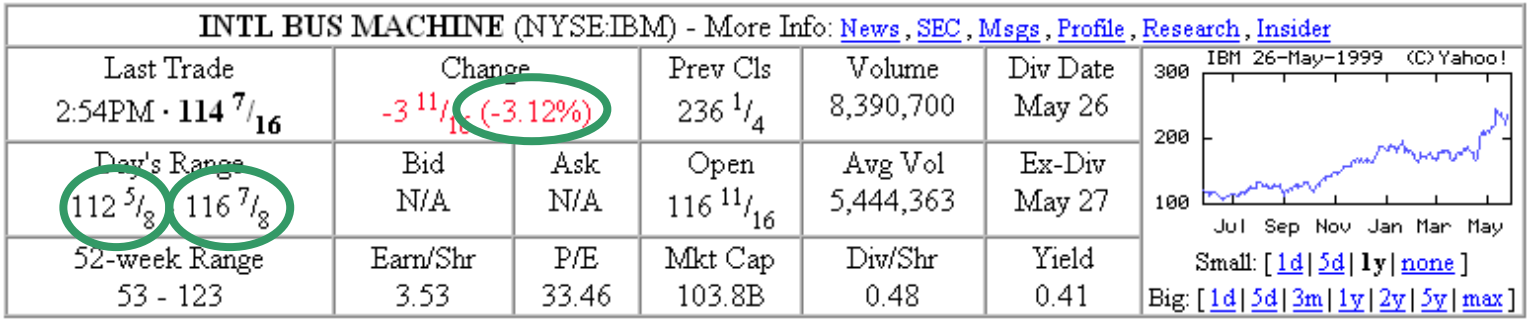
Idea: Limit application of LR rules to proper context (e.g., to skip over HTML table header)

```
<TABLE>
    <TR><TH><B>Country</B></TH><TH><I>Code</I></TH></TR>
    <TR><TD><B>Godfather 1</B></TD><TD><I>1972</I></TD></TR>
    <TR><TD><B>Interstellar</B></TD><TD><I>2014</I></TD></TR>
    <TR><TD><B>Titanic</B></TD><TD><I>1997</I></TD></TR>
</TABLE>
```

- **HLRT rules** (head left token right tail)
  apply LR rule only if inside HT (e.g., H = <TD> T = </TD>)
- **OCLR rules** (open (left token right)* close):
  O and C identify tuple, LR repeated for individual elements
- **NHLRT** (nested HLRT):
  apply rule at current nesting level,
  open additional levels, or return to higher level

Rules for HTML DOM Trees

- Use HTML tag paths from root to target element
- Use more powerful operators for matching, splitting, extracting



INTL BUS MACHINE (NYSE:IBM) - More Info: News , SEC , Msgs , Profile , Research , Insider

Last Trade	Change		Prev Cls	Volume	Div Date	IBM 26-May-1999 (C) Yahoo!
2:54PM · **114 7/16**	-3 11/16 (-3.12%)		236 1/4	8,390,700	May 26	300 / 200 / 100 — Jul Sep Nov Jan Mar May
Day's Range	Bid	Ask	Open	Avg Vol	Ex-Div	Small: [1d | 5d | **1y** | none]
112 5/8 — 116 7/8	N/A	N/A	116 11/16	5,444,363	May 27	Big: [1d | 5d | 3m | 1y | 2y | 5y | max]
52-week Range	Earn/Shr	P/E	Mkt Cap	Div/Shr	Yield	
53 - 123	3.53	33.46	103.8B	0.48	0.41	

Source: A. Sahuguet, F. Azavant: Looking at the Web through <XML> glasses, http://db.cis.upenn.edu/research/w4f.html

Example:
extract the volume
table.tr[1].td[*].txt, **match /Volume/**
extract the % change
table.tr[1].td[1].txt, **match /[(](.*?)[)]/**
extract the day's range for the stock:
table.tr[2].td[0].txt, **match/Day's Range (.*)/, split /-/**

match /.../, split /…/ return lists of strings

Learning Regular Expressions (aka. Wrapper Induction)

Input: Hand-tagged examples of a regular language

Output: (Restricted) regular expression for the language of a finite-state transducer that reads sentences of the language and outputs token of interest

Example:

*This apartment has **3** bedrooms.
 The monthly rent is $ **995**.*
*This apartment has **4** bedrooms.
 The monthly rent is $ **980**.*
*The number of bedrooms is **2**.
 The rent is $ **650** per month.*

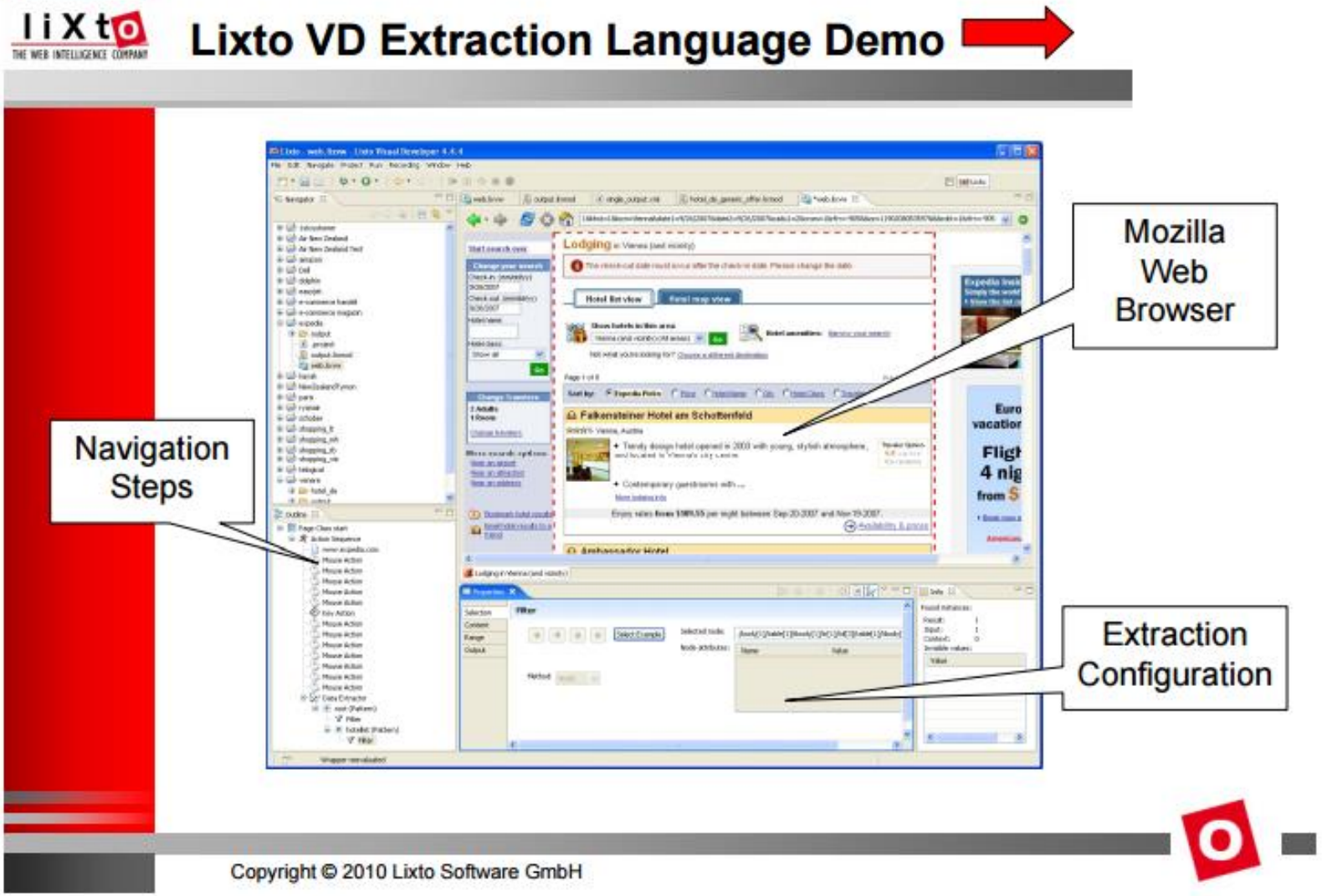
yields | * **<digit>** * "
" * "$" **<digit>+** * | as learned pattern

Problem: Grammar inference for general regular languages is hard.

→ restricted class of regular languages

(e.g. WHISK [Soderland 1999], LIXTO [Baumgartner 2001])

Example of Markup Tool for Supervised Wrapper Induction



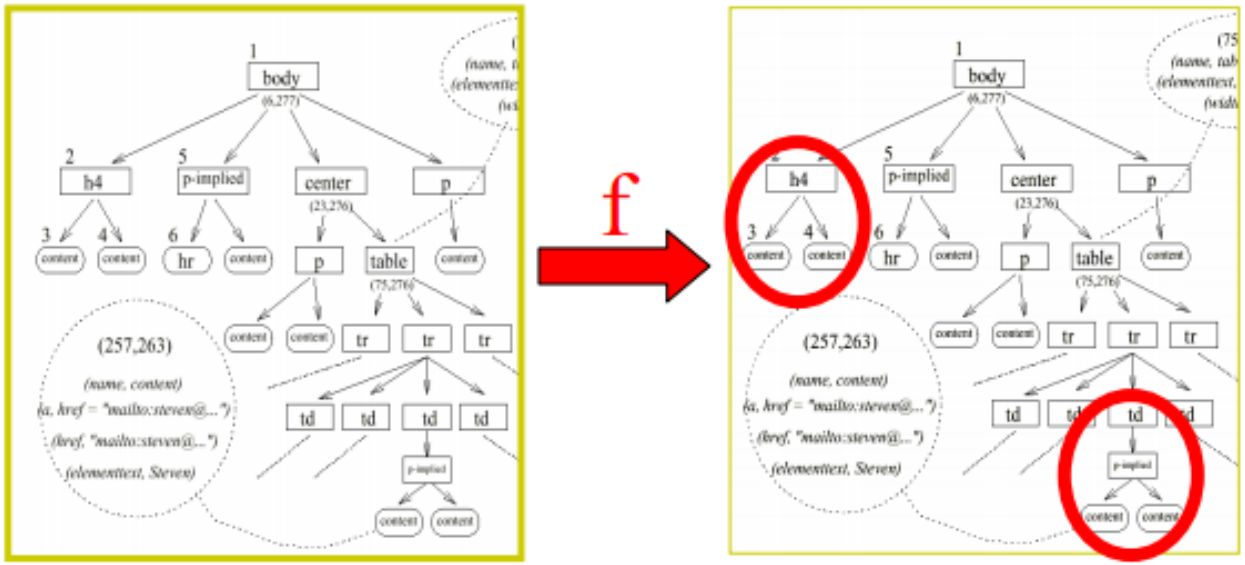
Source: R. Baumgartner, Datalog-related Aspects in Lixto Visual Developer, 2010, http://datalog20.org/slides/baumgartner.pdf

Example of Markup Tool for Supervised Wrapper Induction

Extraction from Tree Structure

Wrapper as Function f: HTML DOM Tree ➲ Subtrees

Leaves of subtrees are among leaves of original tree



⇨„Visually“ define such functions

Source: R. Baumgartner, Datalog-related Aspects in Lixto Visual Developer, 2010, http://datalog20.org/slides/baumgartner.pdf

Limitations and Extensions of Rule-Based IE

- Powerful for wrapping **regularly structured web pages**
  (e.g., template-based from same Deep Web site / CMS)
- Many **complications** with real-life HTML
  (e.g., misuse of tables for layout)
- Extend flat view of input to **trees**:
  - **hierarchical document structure** (DOM tree, XHTML)
  - extraction patterns for restricted regular languages on trees
    (e.g. fragements and variations of XPath)
- **Regularities with exceptions** are difficult to capture
  - Identify positive and negative cases and use statistical models

15.2.2 IE with Statistical Learning

For heterogeneous Web sources and for natural-language text

- **NLP techniques** (PoS tagging, parsing) for tokenization
- **Identify patterns** (regular expressions) as **features**
- **Train statistical learners** for segmentation and labeling (e.g., HMM, CRF, SVM, etc.) augmented with lexicons
- Use learned model to **automatically tag** new input sequences

- Example for labeled training data:

  The WWW conference in 2007 takes place in Banff in Canada.

  Today's keynote speaker is Dr. Berners-Lee from W3C.

  with tags of the following kinds:

  event, person, location, organization, date

IE as Boundary Classification

Idea: Learn **classifiers** to **recognize start token** and **end token** for the facts under consideration. Combine multiple classifiers (ensemble learning) for more robust output.

Example:

There will be a talk by Alan Turing at the University at 4 PM.

Prof. Dr. James Watson will speak on DNA at MPI at 6 PM.

The lecture by Francis Crick will be in the IIF at 3:15 today.

Trained classifiers test each token
(with PoS tag, LR neighbor tokens, etc. as features)
for two classes: begin-fact, end-fact

IE as Text Segmentation and Labeling

Idea: Observed text is concatenation of structured record with limited reordering and some missing fields

Examples: Addresses and bibliographic records

House number	Building	Road	City	State	Zip
4089	Whispering Pines	Nobel Drive	San Diego	CA	92122

Author	Year	Title	Journal	Volume	Page

P.P.Wangikar, T.P. Graycar, D.A. Estell, D.S. Clark, J.S. Dordick (1993) Protein and Solvent Engineering of Subtilising BPN' in Nearly Anhydrous Organic Media J.Amer. Chem. Soc. 115, 12231-12237.

Source: S. Sarawagi: Information Extraction, 2008

→ **Hidden Markov Model (HMM)** !

HMM Example: Postal Address

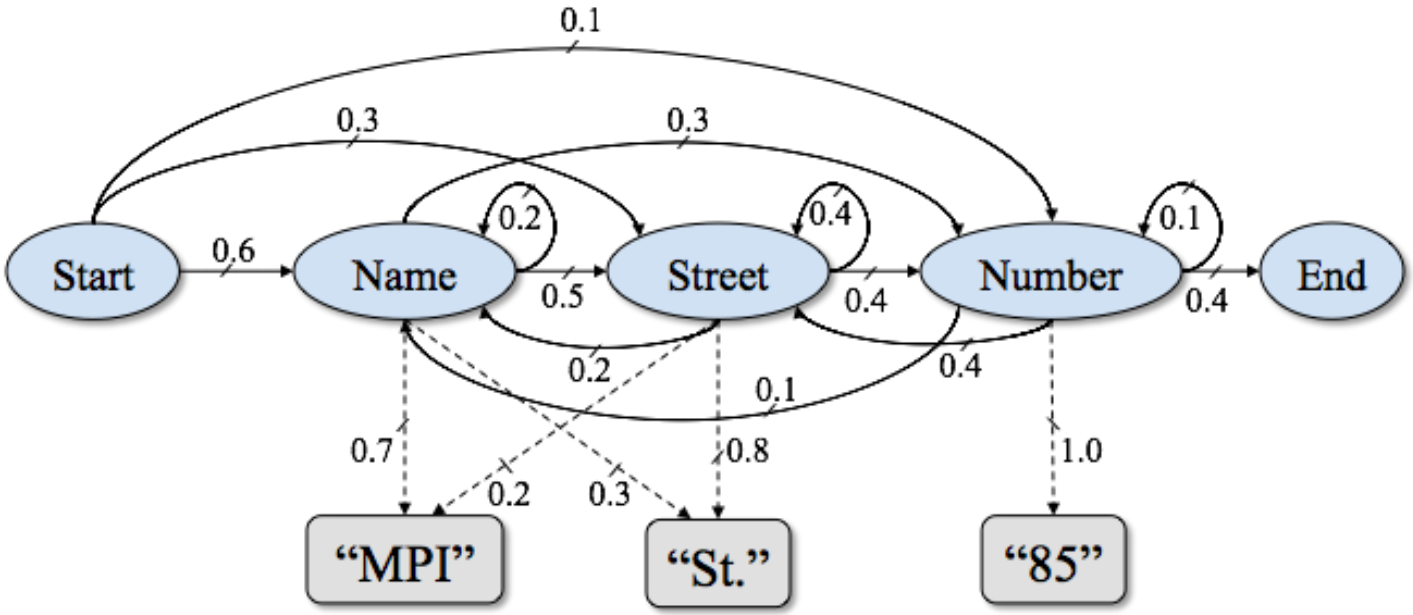
Goal: Label the tokens in sequences

Max-Planck-Institute, Stuhlsatzenhausweg 85

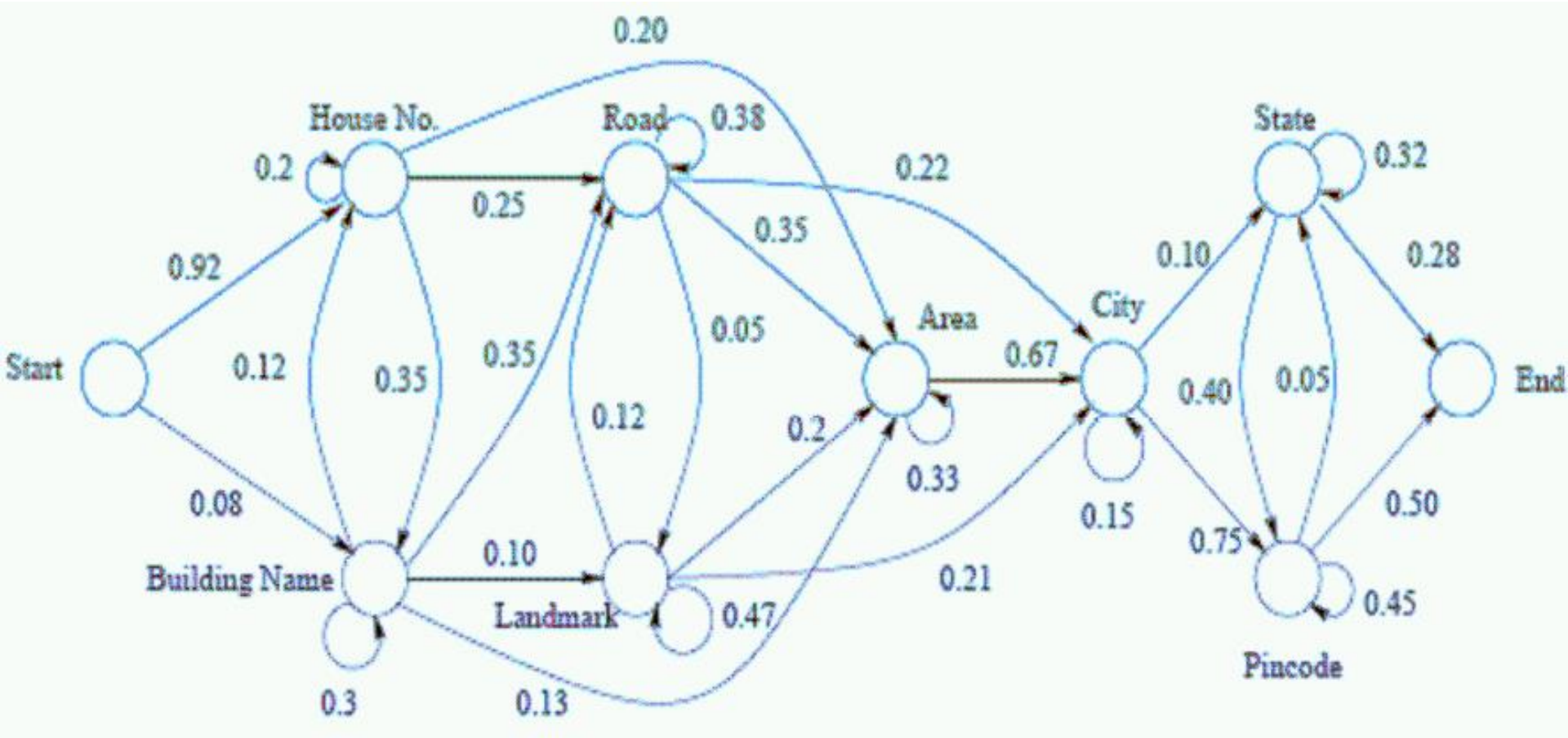
with the labels **Name**, **Street**, **Number**

Σ = {"MPI", "St.", "85"} // output alphabet

S = {Name, Street, Number} // (hidden) states



HMM Example: Postal Addresses



Source: Eugene Agichtein and Sunita Sarawagi, Tutorial at KDD 2006

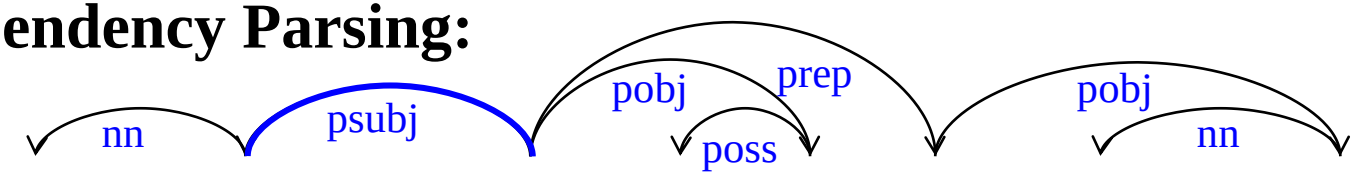
Basics from NLP for IE (in a Nutshell)

Surajit Chaudhuri obtained his PhD from Stanford University
under the supervision of Prof. Jeff Ullman

Part-of-Speech (POS) Tagging:

Surajit Chaudhuri obtained his PhD from Stanford University
NNP NNP VBD PRP NN IN NNP NNP
under the supervision of Prof. Jeff Ullman
IN DT NN IN NNP NNP NNP

Dependency Parsing:



Surajit Chaudhuri obtained his PhD from Stanford University

under the supervision of Prof. Jeff Ullman

NLP: Part-of-Speech (POS) Tagging

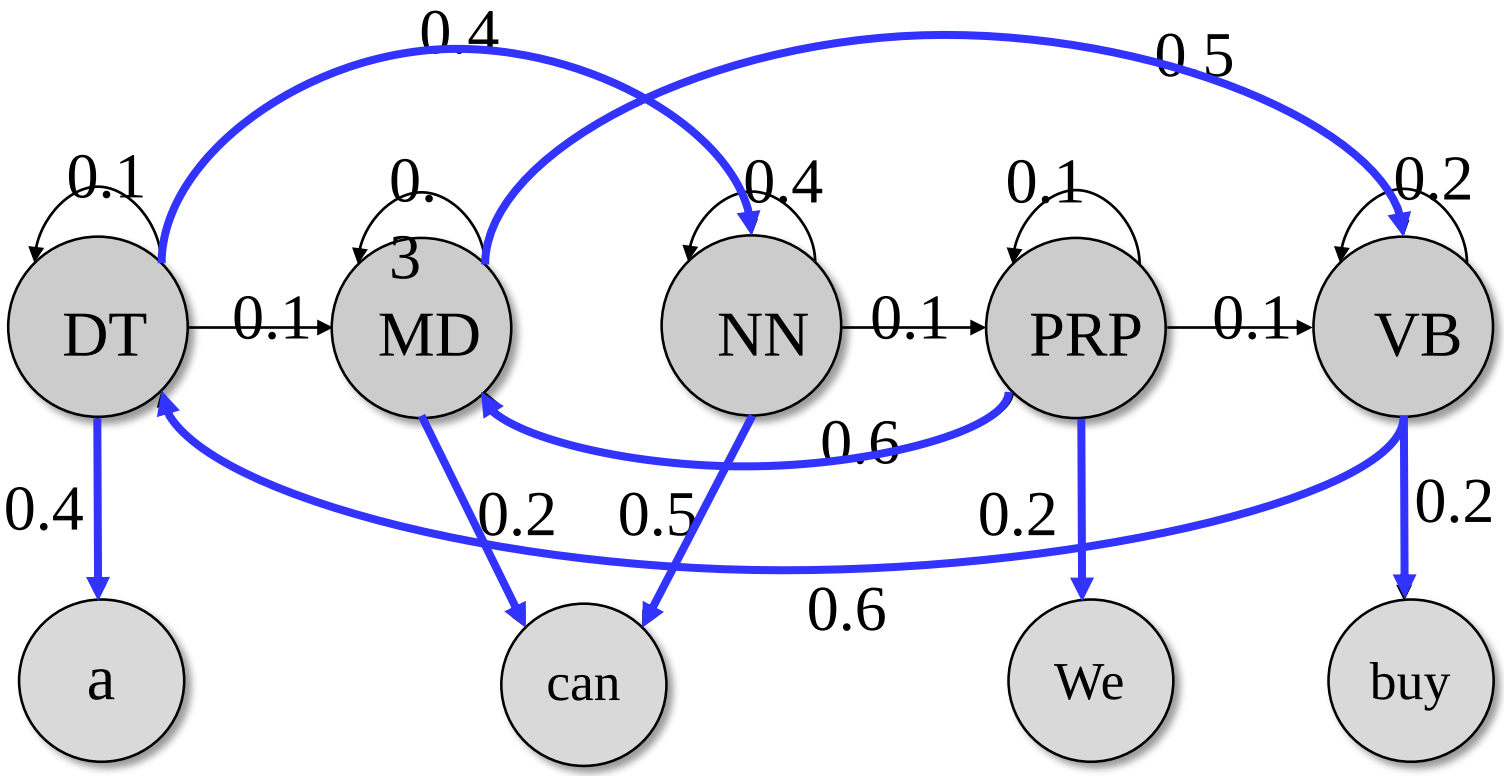
Tag each word with its grammatical role (noun, verb, etc.)
Use HMM or CRF trained over large corpora

POS Tags (Penn Treebank):

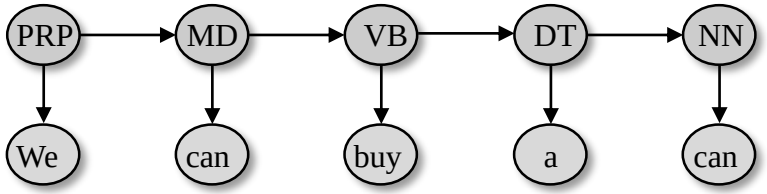
CC coordinating conjunction
CD cardinal number
DT determiner
EX existential *there*
FW foreign word
IN preposition or subordinating conjunction
JJ adjective
JJR adjective, comparative
JJS adjective, superlative
LS list item marker
MD modal
NN noun
NNS noun, plural
NNP proper noun
NNPS proper noun, plural
PDT predeterminer
POS possessive ending
PRP personal pronoun
PRP$ possessive pronoun
RB adverb
RBR adverb, comparative
RBS adverb, superlative
RP particle
SYM symbol
TO *to*
UH interjection
VB verb, base form
VBD verb, past tense
VBG verb, gerund or present participle
VBN verb, past participle
VBP verb, non-3rd person singular present
VBZ verb, 3rd person singular present
WDT wh-determiner (which …)
WP wh-pronoun (what, who, whom, …)
WP$ possessive wh-pronoun
WRB wh-adverb

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

HMM for Part-of-Speech Tagging



How to find the best sequence of POS tags for sentence "**We can buy a can**"?



(Linear-Chain) Conditional Random Fields (CRFs)

Extend HMMs in several ways:

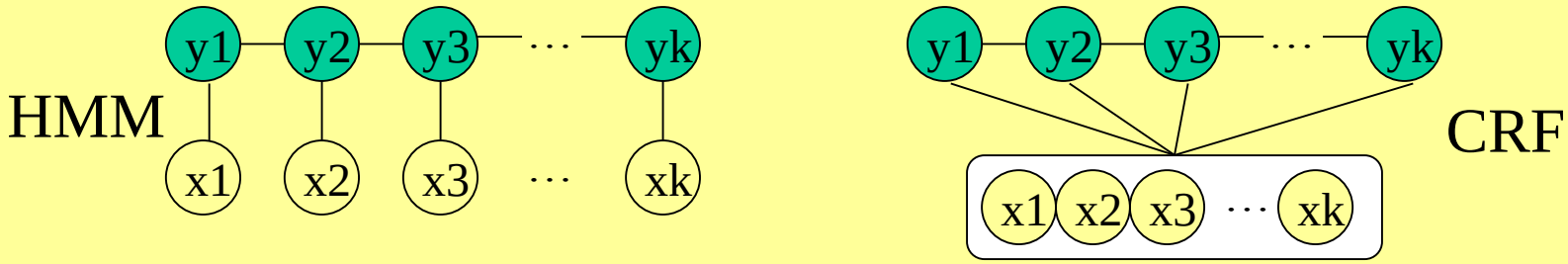
- exploit **complete input sequence** for predicting state transition, not just last token
- use **features** of input tokens (e.g. hasCap, isAllCap, hasDigit, isDDDD, firstDigit, isGeoname, hasType, afterDDDD, directlyPrecedesGeoname, etc.)

For token sequence $x=x_1 \ldots x_k$ and state sequence $y=y_1..y_k$

HMM models joint distr. $P[x,y] = \Pi_{i=1..k} P[y_i|y_{i-1}] * P[x_i|y_i]$

CRF models conditional distr. $P[y|x]$

with conditional independence of non-adjacent y_i's given x



CRF Training and Inference

graph structure of conditional-independence assumptions leads to:

$$P[y \mid x] = \frac{1}{Z(x)} exp\left(\sum_{j=1}^{m} \lambda_j \sum_{t=1}^{T} f_j(y_{t-1}, y_t, x) \right)$$

where j ranges over feature functions and
Z(x) is a normalization constant

parameter estimation with n training sequences:
MLE with regularization

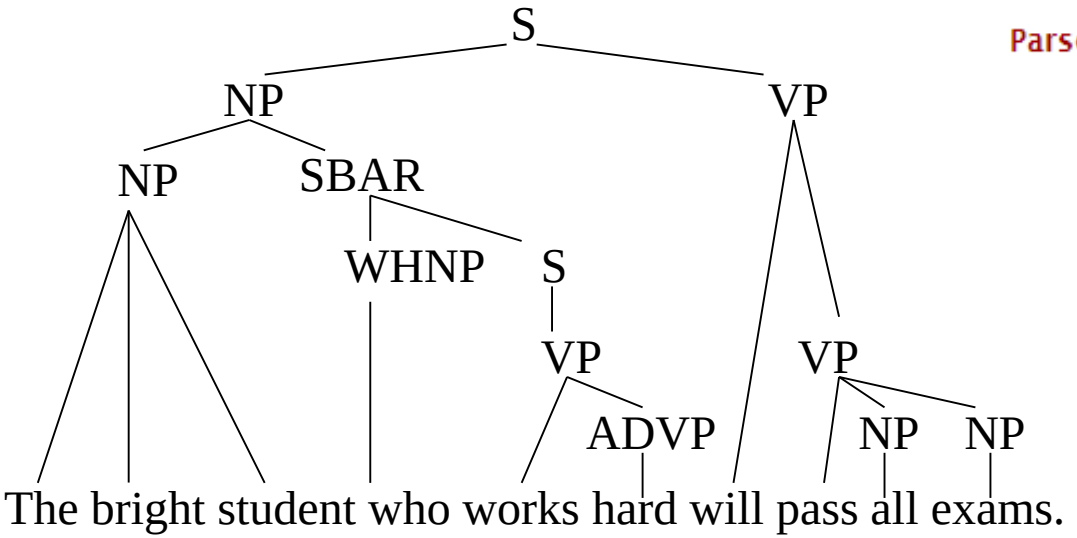
$$log\, L(\theta) = \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{j=1}^{m} \lambda_j f_j(y_{t-1}^{(i)}, y_t^{(i)}, x_t^{(i)}) - \sum_{i=1}^{n} log\, Z(x^{(i)}) - \sum_{j=1}^{m} \frac{\lambda_j^2}{2\sigma^2}$$

inference of most likely (x,y) for given x:
dynamic programming (similar to Viterbi)

CRFs can be further generalized to undirected graphs
of coupled random variables (aka. MRF: Markov random field)

NLP: Deep Parsing for Constituent Trees

- Construct syntax-based tree of sentence constituents
- Use **non-deterministic context-free grammars** natural ambiguity
- Use **probabilistic grammar (PCFG):** likely vs. unlikely parse trees (trained on corpora)



Parse

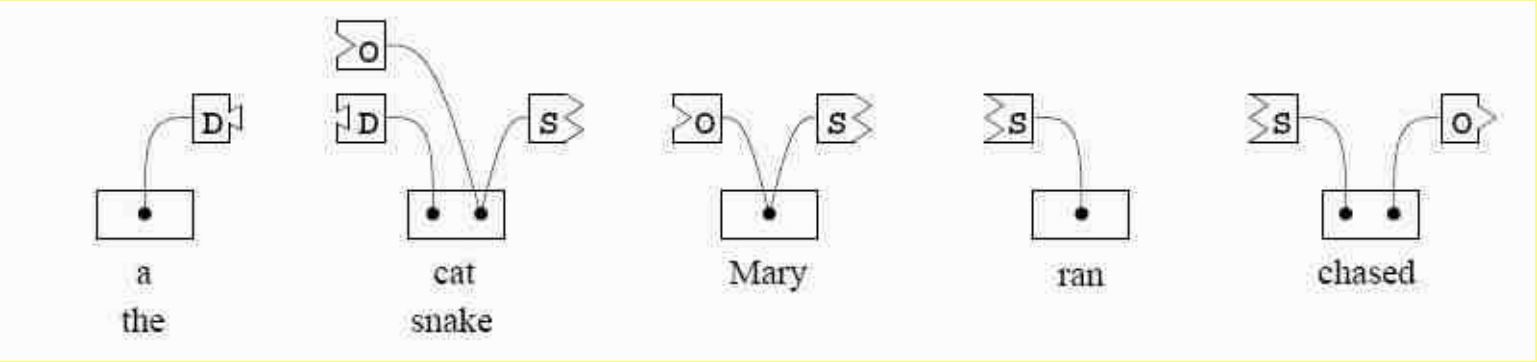
```
(ROOT
  (S
    (NP (JJ Colorless) (JJ green) (NNS ideas))
    (VP (VBP sleep)
      (ADVP (RB furiously)))
    (. .)))
```

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

NLP: Link-Grammar-Based Dependency Parsing

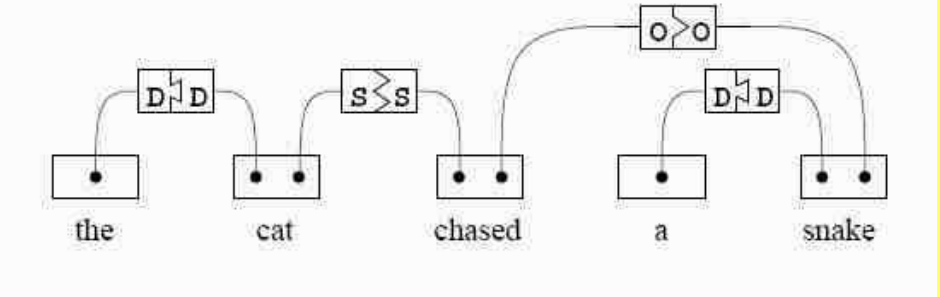
Dependency parser based on grammatical rules for left & right connector

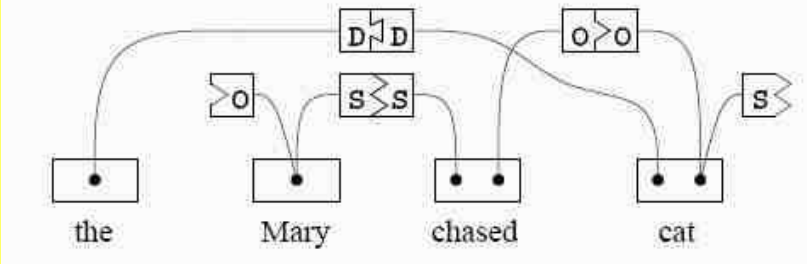


[Sleator/ Temperley 1991]

rules have form: **w1 → left: { A1 | A2 | …} right: { B1 | B2 | …}**
w2 → left: { C1 | B1 | …} right: {D1 | D2 | …}
w3 → left: { E1 | E2 | …} right: {F1 | C1 | …}

- **Parser finds all matchings that connect all words into planar graph (using dynamic programming for search-space traversal)**
- **Extended to probabilistic parsing and error-tolerant parsing**





$O(n^3)$ algorithm with many implementation tricks, and grammar size n is huge

Dependency Parsing Examples (1)

http://www.link.cs.cmu.edu/link/

```
++++Time                                          0.01 seconds (77.95 total)
Found 4 linkages (2 with no P.P. violations)
  Linkage 1, cost vector = (UNUSED=0 DIS=0 AND=0 LEN=18)

    +-----------------------------------Xp-----------------------------------+
    +----------Wd----------+-------------Ss------------+                     |
    |      +-------Ds------+-----Bs-----+               |      +----Opn---+   |
    |      |       +----A---+---R--+--RS-+--MVa--+      +---I--+--O-+     |   |
    |      |       |        |      |     |       |      |      |    |     |   |
LEFT-WALL the bright.a student.n who works.v hard.e will.v pass.v all exams.n
```

```
    +---------------------------Xp----------------------------+
    |                      +-------MVp------+                 |
    |                      |    +---Js--+   +-------Js-------+ |
    +---Wd--+------Ss-----+-MVp+  +--G-+    |  +--G-+-YS-+-D-+ |
    |       |             |    |  |    |    |  |    |    |   | |
LEFT-WALL Alice and Bob met.v in New York on New Year 's.p Eve .
```

Selected tags (CMU Link Parser), out of ca. 100 tags (plus variants):

MV connects verbs to modifying phrases like adverbs, time expressions, etc.

O connects transitive verbs to direct or indirect objects

J connects prepositions to objects

B connects nouns with relative clauses

Dependency Parsing Examples (2)

http://nlp.stanford.edu/software/lex-parser.shtml

Your sentence

The bright student who works hard will pass all exams.

Tagging

```
The/DT  bright/JJ  student/NN  who/WP  works/VBZ  hard/JJ  will/MD  pass/VB  all/DT  exams/NNS  ./.
```

Parse

```
(ROOT
  (S
    (NP
      (NP (DT The) (JJ bright) (NN student))
      (SBAR
        (WHNP (WP who))
        (S
          (VP (VBZ works)
            (ADJP (JJ 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(works-5, student-3)
nsubj(pass-8, student-3)
rel(works-5, who-4)
rcmod(student-3, works-5)
acomp(works-5, hard-6)
aux(pass-8, will-7)
det(exams-10, all-9)
dobj(pass-8, exams-10)
```

Selected tags (Stanford Parser), out of ca. 50 tags:

nsubj: nominal subject	amod; adjectival modifier
rel: relative	rcmod: relative clause modifier
dobj: direct object	acomp: adjectival complement
det: determiner	poss: possession modifier

…

Additional Literature for 15.2

- S. Sarawagi: Information Extraction, Foundations & Trends in Databases 1(3), 2008
- H. Cunningham: Information Extraction, Automatic.
  in: Encyclopedia of Language and Linguistics, 2005, http://www.gate.ac.uk/ie/
- M.E. Califf, R.J. Mooney: Bottom-Up Relational Learning of Pattern Matching Rules for Information Extraction, JMLR 2003
- S. Soderland: Learning Information Extraction Rules for Semi-Structured and Free Text, Machine Learning Journal 1999
- N. Kushmerick: Wrapper induction: Efficiency and expressiveness, Art. Intelligence 2000
- A. Sahuguet, F. Azavant: Building light-weight wrappers for legacy web data-sources using W4F, VLDB 1999
- R Baumgartner et al.: Visual Web Information Extraction with Lixto, VLDB 2001
- G. Gottlob et al.: The Lixto data extraction project, PODS 2004
- B. Liu: Web Data Mining, Chapter 9, Springer 2007
- C. Manning, H. Schütze: Foundations of Statistical Natural Language Processing, MIT Press 1999