Information Extraction from the World Wide Web

Information Extraction

Definition:

The automatic extraction of *structured* information from *unstructured* documents.

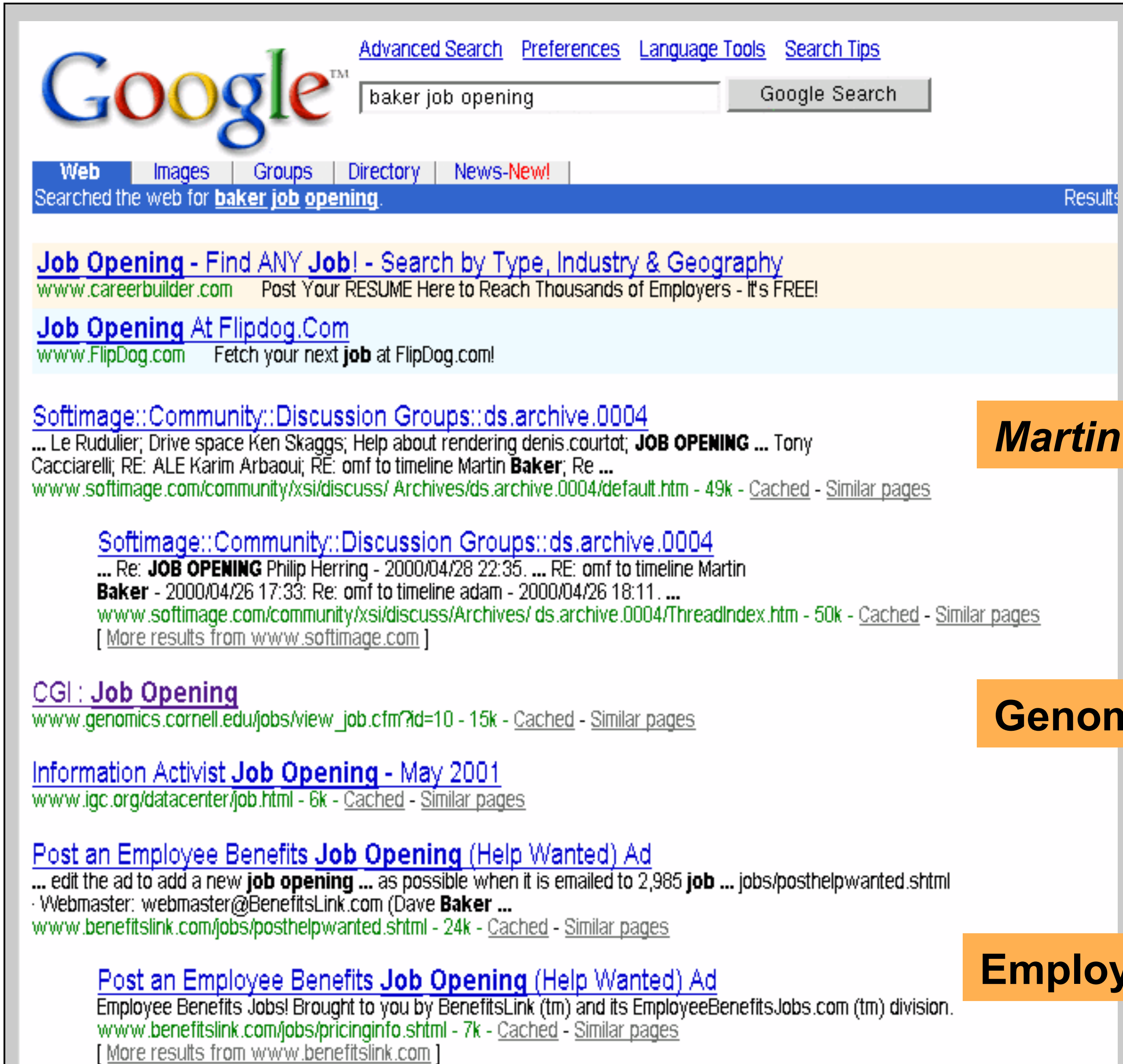
Different from Information Retrieval (search engines): users are presented with retrieved documents, ranked by relevance. User must open web pages and extract information.

Overall Goals:

- Making information more accessible *to people*
- Making information more *machine-processable*

Practical Goal: Build large knowledge bases

Example: The Problem (seeking jobs)



***Martin Baker*, a person**

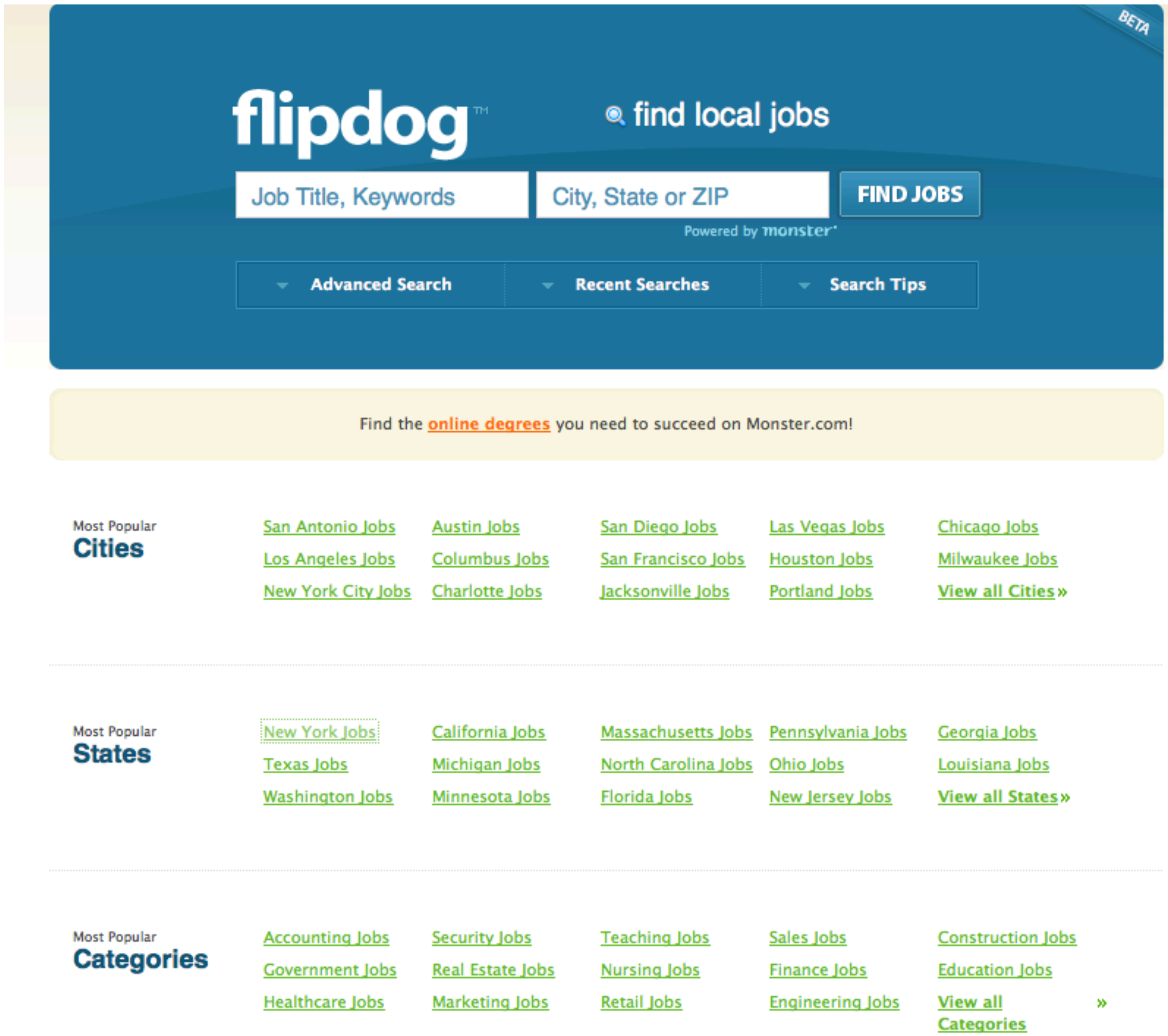
Genomics job

Employers job posting form

Example: A Solution

Extracting Job Openings from the Web

Flipdog job search engine



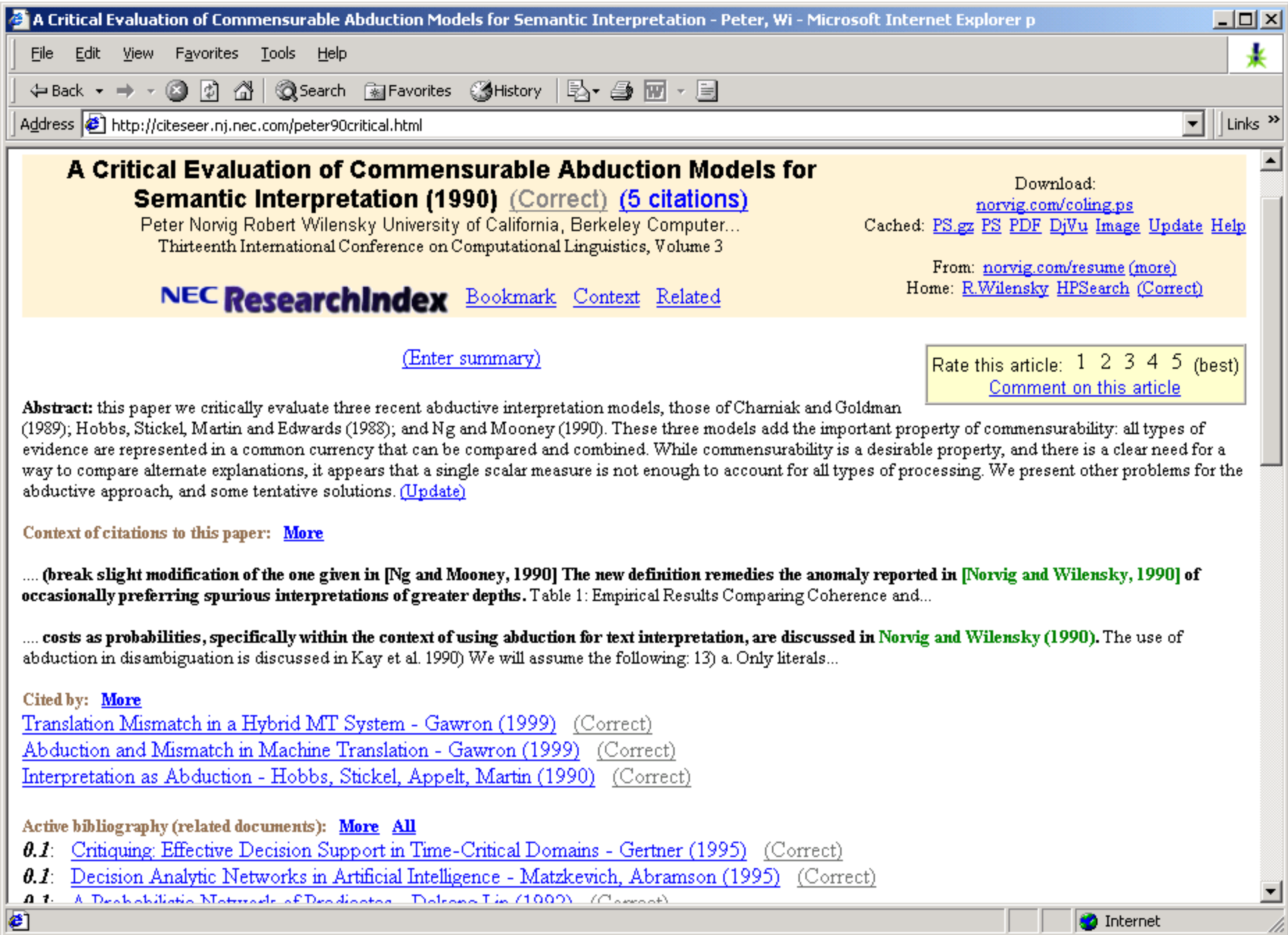
Create a database with the Extracted Job Information

Job Openings:
Category = ***Food Services***
Keyword = ***Baker***
Location = ***Continental U.S.***

Food Pantry Workers at Lutheran Social Services	October 11, 2002	Archbold, OH
Cooks at Lutheran Social Services	October 11, 2002	Archbold, OH
Bakers Assistants at Fine Catering by Russell Morin	October 11, 2002	Attleboro, MA
Baker's Helper at Bird-in-Hand	October 11, 2002	United States
Assistant Baker at Gourmet To Go	October 11, 2002	Maryland Heights, MO
Host/Hostess at Sharis Restaurants	October 10, 2002	Beaverton, OR
Cooks at Alta's Rustler Lodge	October 10, 2002	Alta, UT
Line Attendant at Sun Valley Coporation	October 10, 2002	Huntsville, UT
Food Service Worker II at Garden Grove Unified School District	October 10, 2002	Garden Grove, CA
Night Cook / Baker at SONOCO	October 10, 2002	Houma, LA
Cooks/Prep Cooks at GrandView Lodge	October 10, 2002	Nisswa, MN
Line Cook at Lone Mountain Ranch	October 10, 2002	Big Sky, MT
Production Baker at Whole Foods Market	October 08, 2002	Willowbrook, IL
Cake Decorator/Baker at Mandalay Bay Hotel and Casino	October 08, 2002	Las Vegas, NV
Shift Supervisors at Brueggers Bagels	October 08, 2002	Minneapolis, MN

Data Mining the Extracted Job Information

Example 2: IE from Research Papers



A Critical Evaluation of Commensurable Abduction Models for Semantic Interpretation - Peter, Wi - Microsoft Internet Explorer p

File Edit View Favorites Tools Help

Back Search Favorites History

Address http://citeseer.nj.nec.com/peter90critical.html Links

A Critical Evaluation of Commensurable Abduction Models for Semantic Interpretation (1990) (Correct) (5 citations)

Peter Norvig Robert Wilensky University of California, Berkeley Computer...

Thirteenth International Conference on Computational Linguistics, Volume 3

NEC ResearchIndex Bookmark Context Related

Download: norvig.com/coling.ps

Cached: PS.gz PS PDF DjVu Image Update Help

From: norvig.com/resume (more)

Home: R.Wilensky HPSearch (Correct)

(Enter summary)

Rate this article: 1 2 3 4 5 (best)

Comment on this article

Abstract: this paper we critically evaluate three recent abductive interpretation models, those of Charniak and Goldman (1989); Hobbs, Stickel, Martin and Edwards (1988); and Ng and Mooney (1990). These three models add the important property of commensurability: all types of evidence are represented in a common currency that can be compared and combined. While commensurability is a desirable property, and there is a clear need for a way to compare alternate explanations, it appears that a single scalar measure is not enough to account for all types of processing. We present other problems for the abductive approach, and some tentative solutions. (Update)

Context of citations to this paper: **More**

.... **(break slight modification of the one given in [Ng and Mooney, 1990] The new definition remedies the anomaly reported in [Norvig and Wilensky, 1990] of occasionally preferring spurious interpretations of greater depths.** Table 1: Empirical Results Comparing Coherence and...

.... **costs as probabilities, specifically within the context of using abduction for text interpretation, are discussed in Norvig and Wilensky (1990).** The use of abduction in disambiguation is discussed in Kay et al. 1990) We will assume the following: 13) a. Only literals...

Cited by: **More**

Translation Mismatch in a Hybrid MT System - Gawron (1999) (Correct)

Abduction and Mismatch in Machine Translation - Gawron (1999) (Correct)

Interpretation as Abduction - Hobbs, Stickel, Appelt, Martin (1990) (Correct)

Active bibliography (related documents): **More** **All**

0.1: Critiquing: Effective Decision Support in Time-Critical Domains - Gertner (1995) (Correct)

0.1: Decision Analytic Networks in Artificial Intelligence - Matzkevich, Abramson (1995) (Correct)

Internet

IE from financial statements

This filing covers the period from December 1996 to September 1997.

ENRON GLOBAL POWER & PIPELINES L.L.C.
CONSOLIDATED BALANCE SHEETS
(IN THOUSANDS, EXCEPT SHARE AMOUNTS)

	SEPTEMBER 30, 1997 (UNAUDITED)	DECEMBER 31, 1996
ASSETS		
Current Assets		
Cash and cash equivalents	$ 54,262	$ 24,582
Accounts receivable	8,473	6,301
Current portion of notes receivable	1,470	1,394
Other current assets	336	404
Total Current Assets	71,730	32,681
Investments in to Unconsolidated Subsidiaries	286,340	298,530
Notes Receivable	16,059	12,111
Total Assets	$374,408	$343,843
LIABILITIES AND SHAREHOLDERS' EQUITY		
Current Liabilities		
Accounts payable	$ 13,461	$ 11,277
Accrued taxes	1,910	1,488
Total Current Liabilities	15,371	49,348
Deferred Income Taxes	525	4,301

The U.S. energy markets in 1997 were subject to significant fluctuation

Data mine these reports for
- **suspicious behavior,**
- **to better understand what is normal.**

What is "Information Extraction"

As a task: Filling slots in a database from sub-segments of text.

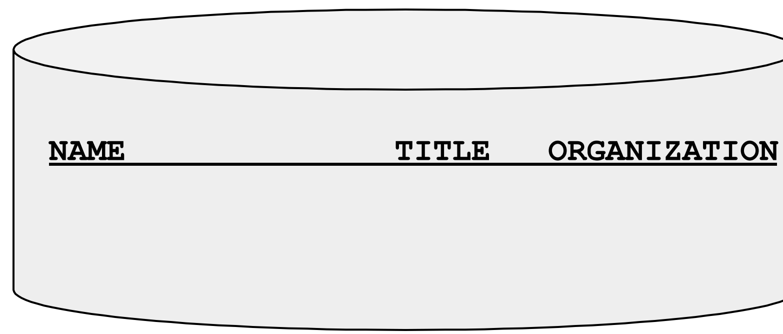
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…



What is "Information Extraction"

As a task: **Filling slots in a database from sub-segments of text.**

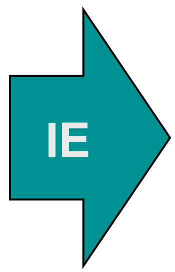
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…



NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

What is “Information Extraction”

As a sequence of techniques:

Information Extraction =
segmentation + classification + clustering + slot filling

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

Identify named entities

What is "Information Extraction"

As a family of techniques:

Information Extraction =
segmentation + classification + association + slot filling

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

classify entities according to categories, e.g. person, role, company

What is "Information Extraction"

As a family of techniques:

Information Extraction =
segmentation + classification + association + slot filling

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

associate related entities

Microsoft Corporation
CEO
Bill Gates

Microsoft
Gates

Microsoft

Bill Veghte
Microsoft
VP

Richard Stallman
founder
Free Software Foundation

What is "Information Extraction"

As a family of techniques:

Information Extraction =
segmentation + classification + association + slot filling

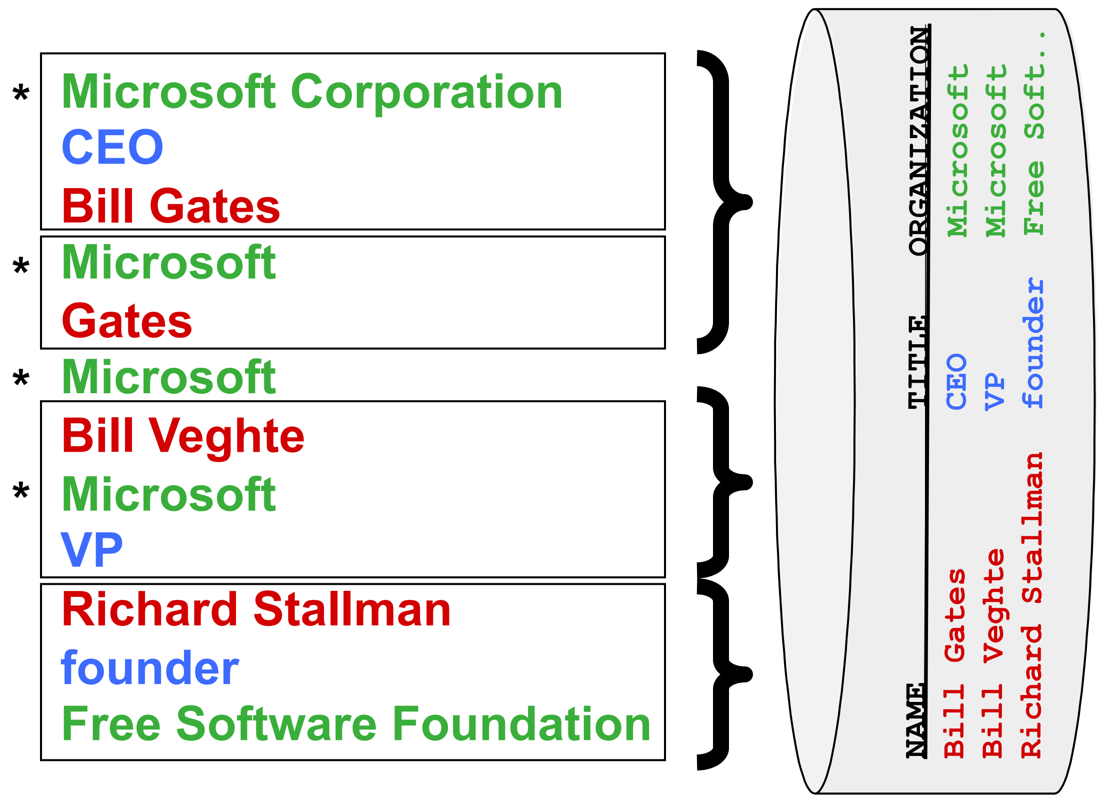
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…



Why IE from the Web?

- Science
  - Grand old dream of AI: Build large KB* and reason with it. IE from the Web enables the creation of this KB.
  - IE from the Web is a complex problem that inspires new advances in machine learning.
- Profit
  - Many companies interested in leveraging data currently "locked in unstructured text on the Web".
  - Not yet a monopolistic winner in this space.

* KB = "Knowledge Base"

Information Extraction in Applications

- Structured Search
- Opinion Mining/Sentiment Extraction
- Data Mining over Extracted Relationships

What makes IE from the Web Different?

Partly avoid deep NLP analysis, exploit formatting & linking

Newswire

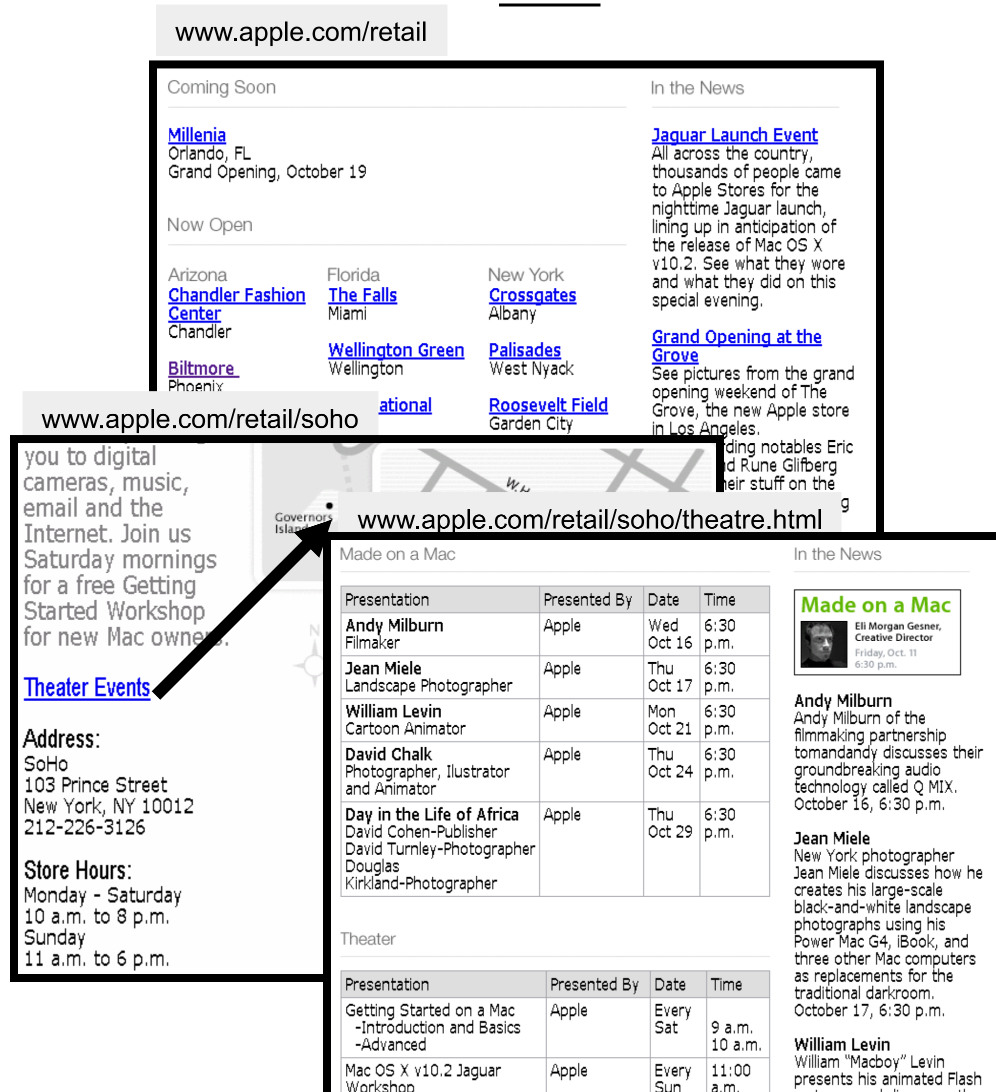
Apple to Open Its First Retail Store in New York City

MACWORLD EXPO, NEW YORK--July 17, 2002--Apple's first retail store in New York City will open in Manhattan's SoHo district on Thursday, July 18 at 8:00 a.m. EDT. The SoHo store will be Apple's largest retail store to date and is a stunning example of Apple's commitment to offering customers the world's best computer shopping experience.

"Fourteen months after opening our first retail store, our 31 stores are attracting over 100,000 visitors each week," said Steve Jobs, Apple's CEO. "We hope our SoHo store will surprise and delight both Mac and PC users who want to see everything the Mac can do to enhance their digital lifestyles."

The directory structure, link structure, formatting & layout of the Web is its own new grammar.

Web



Landscape of IE Tasks (1/4): Pattern Feature Domain

Text paragraphs without formatting

Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.

Grammatical sentences and some formatting & links

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

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

- Press

Contact

- General information
- Directions maps

Non-grammatical snippets, rich formatting & links

Barto, Andrew G.	(413) 545-2109	barto@cs.umass.edu	CS276
Professor. Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.			
Berger, Emery D.	(413) 577-4211	emery@cs.umass.edu	CS344
Assistant Professor.			
Brock, Oliver	(413) 577-0334	oli@cs.umass.edu	CS246
Assistant Professor.			
Clarke, Lori A.	(413) 545-1328	clarke@cs.umass.edu	CS304
Professor. Software verification, testing, and analysis; software architecture and design.			
Cohen, Paul R.	(413) 545-3638	cohen@cs.umass.edu	CS278
Professor. Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.			

Tables

8:30 - 9:30 AM	**Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty** *Joseph Y. Halpern, Cornell University*				
9:30 - 10:00 AM	Coffee Break				
10:00 - 11:30 AM	Technical Paper Sessions:				
Cognitive Robotics	**Logic Programming**	**Natural Language Generation**	**Complexity Analysis**	**Neural Networks**	**Games**
739: A Logical Account of Causal and Topological Maps *Emilio Remolina and Benjamin Kuipers*	116: A-System: Problem Solving through Abduction *Marc Denecker, Antonis Kakas, and Bert Van Nuffelen*	758: Title Generation for Machine-Translated Documents *Rong Jin and Alexander G. Hauptmann*	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories *Marco Cadoli, Thomas Eiter, and Georg Gottlob*	179: Knowledge Extraction and Comparison from Local Function Networks *Kenneth McGarry, Stefan Wermter, and John MacIntyre*	71: Iterative Widening *Tristan Cazenave*
549: Online-Execution of ccGolog Plans *Henrik Grosskreutz*	131: A Comparative Study of Logic Programs with	246: Dealing with Dependencies between Content Planning and	470: A Perspective on Knowledge Compilation	258: Violation-Guided Learning for Constrained	353: Temporal Difference Learning Applied to a

Landscape of IE Tasks (2/4):
Wrapper induction

Wrapper extracts content of a particular information source and translates it into a relational form.

Web site specific

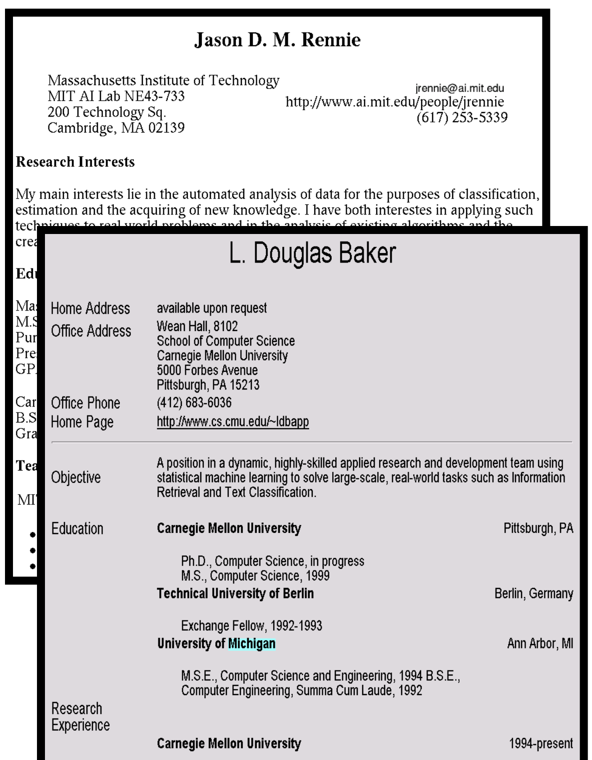
Formatting

E.g.: Amazon.com Book Pages

Genre specific

Layout

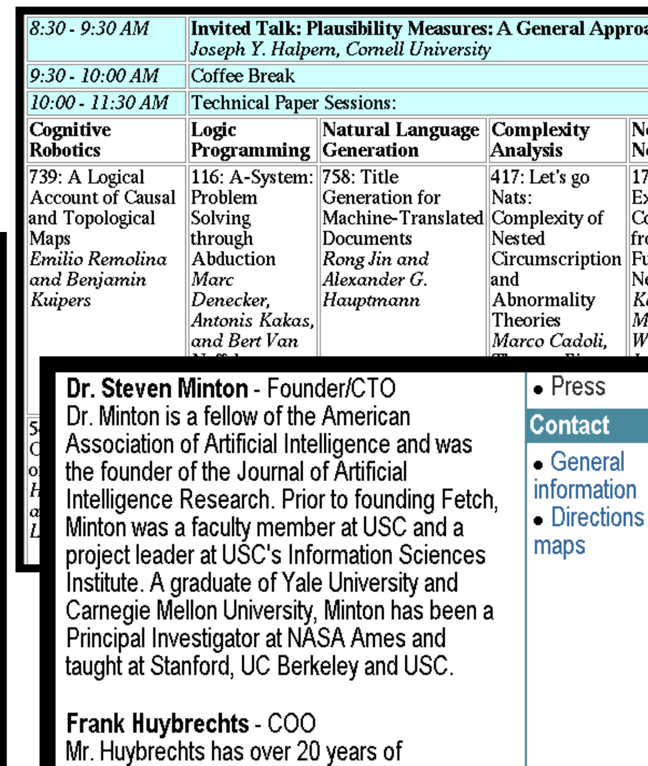
E.g. Resumes



Wide, non-specific

Language

E.g. University Names



Landscape of IE Tasks (3/4): Pattern Complexity

E.g. word patterns:

Closed set

U.S. states

He was born in Alabama…

The big Wyoming sky…

Regular set

U.S. phone numbers

Phone: (413) 545-1323

The CALD main office can be reached at 412-268-1299

Complex pattern

U.S. postal addresses

University of Arkansas
P.O. Box 140
Hope, AR 71802

Headquarters:
1128 Main Street, 4th Floor
Cincinnati, Ohio 45210

Ambiguous patterns, needing context and many sources of evidence

Person names

…was among the six houses sold by Hope Feldman that year.

Pawel Lake, Software Engineer at WhizBang Labs.

Landscape of IE Tasks (4/4): Pattern Combinations

Jack Welch will retire as CEO of General Electric tomorrow. The top role at the Connecticut company will be filled by Jeffrey Immelt.

Single entity

Person: **Jack Welch**

Person: **Jeffrey Immelt**

Location: **Connecticut**

Binary relationship

Relation: **Person-Title**
Person: **Jack Welch**
Title: **CEO**

Relation: **Company-Location**
Company: **General Electric**
Location: **Connecticut**

N-ary record

Relation: **Succession**
Company: **General Electric**
Title: **CEO**
Out: **Jack Welsh**
In: **Jeffrey Immelt**

"Named entity" extraction

Evaluation of Single Entity Extraction

TRUTH:

Michael Kearns and Sebastian Seung will start Monday's tutorial, followed by Richard M. Karpe and Martin Cooke.

PREDICTED:

Michael Kearns and Sebastian Seung will start Monday's tutorial, followed by Richard M. Karpe and Martin Cooke.

$$\text{Precision} = \frac{\text{\# correctly predicted segments}}{\text{\# predicted segments}} = \frac{2}{6}$$

$$\text{Recall} = \frac{\text{\# correctly predicted segments}}{\text{\# true segments}} = \frac{2}{4}$$

$$\text{F1} = \text{Harmonic mean of Precision \& Recall} = \frac{1}{((1/P) + (1/R)) / 2}$$

Parade of IE tasks (and some technique)

Information Extraction and beyond

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

- Ontological Information Extraction
- Fact Extraction
- Instance Extraction

Elvis Presley	singer
Angela Merkel	politician

- Named Entity Recognition

...married <u>Elvis</u> on 1967-05-01

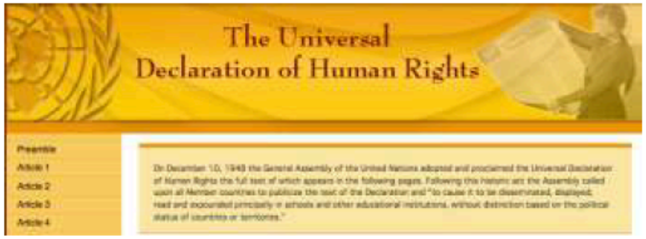
- Tokenization& Normalization

05/01/67 → 1967-05-01

- Source Selection

?

Source extraction: the web

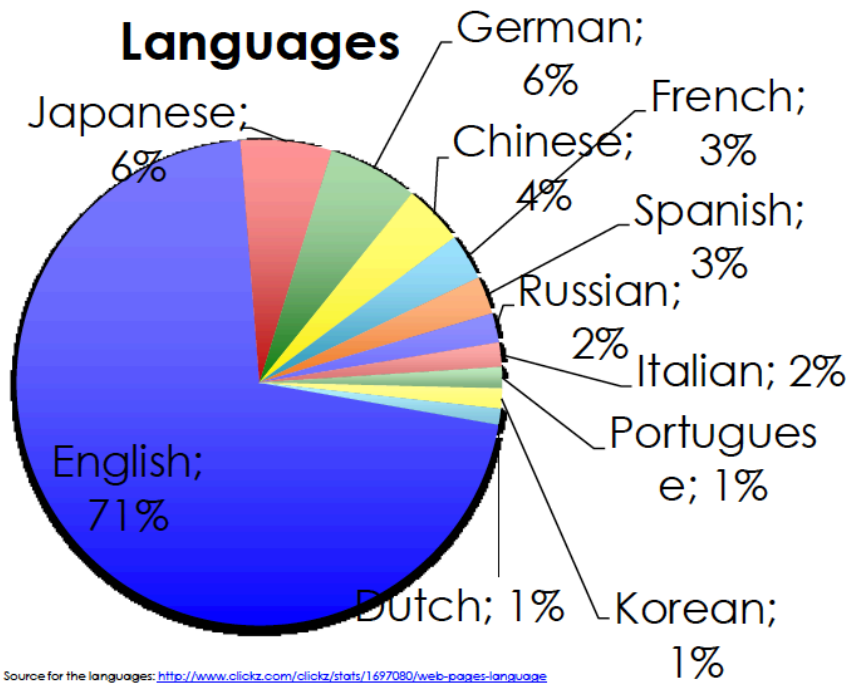


Relational Transducers for Electronic Commerce

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

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

Brad Fordham
Oracle Corporation
bfordham@us.oracle.com



(1 trillion Web sites)

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

Finding the Sources

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

Sources: structured

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

Information Extraction

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

File formats:
- TSV file (values separated by tabulator)
- CSV (values separated by comma)

Sources: semi-structured

```
<catalog>
  <cd>
    <title>
      Empire Burlesque
    </title>
    <artist>
      <firstName>
        Bob
      </firstName>
      <lastName>
        Dylan
      </lastName>
     <artist>
  </cd>
...
```

Information Extraction

Title	Artist
Empire Burlesque	Bob Dylan
…	

File formats:

- XML file (Extensible Markup Language)
- YAML (Yaml Ain‟t a Markup Language)

Sources: unstructured

Founded in 1215 as a colony of Genoa, Monaco has been ruled by the House of Grimaldi since 1297, except when under French control from 1789 to 1814. Designated as a protectorate of Sardinia from 1815 until 1860 by the Treaty of Vienna, Monaco's sovereignty ...

Information Extraction

Event	Date
Foundation	1215
...	...

File formats:
- HTML file
- text file
- word processing document

Sources: mixed

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

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

Information Extraction

Name	Title
Barte	Professor
...	...

Summary of Sources

- We can extract from the entire Web, or from certain Internet domains, thematic domains or files.
- We have to deal with character encodings (ASCII, Code Pages, UTF-8,…) and detect the language.
- Our documents may be structured, semi-structured or unstructured.

Information Extraction and beyond

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

Source Selection

?

✓

Tokenization& Normalization

05/01/67
→
1967-05-01

Named Entity Recognition

...married Elvis on 1967-05-01

Instance Extraction

Elvis Presley	singer
Angela Merkel	politician

Fact Extraction

Ontological Information Extraction

Tokenization

- Tokenization is the process of splitting a text into tokens.
- A token is
  - a word
  - a punctuation symbol
  - a url
  - a number
  - a date
  - or any other sequence of characters regarded as a unit

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

Normalization

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

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

- Solution: Normalize the literals, i.e., convertequivalent literals to one standard form:

08/01/35
01/08/35
8^{th} Jan. 1935
January 8^{th}, 1935

1935-01-08

1.67m
1.67 meters
167 cm
6 feet 5 inches
3 feet 2 toenails

1.67m

Normalization

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

resume

résumé, resume, Resume

1935-01-08

8th Jan 1935, 01/08/1935

Take care not to normalize too aggressively:

bush

Bush

Information Extraction and beyond

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

Ontological Information Extraction

Fact Extraction

Instance Extraction

Elvis Presley	singer
Angela Merkel	politician

Named Entity Recognition

...married Elvis on 1967-05-01

✓

Tokenization& Normalization

05/01/67 → 1967-05-01

✓

Source Selection

?

Named Entity Recognition (NER)

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

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

NER: closed set

- If we have an exhaustive set of the entities we want to extract, we can use closed set extraction:
- Comparing every string in the text to every string in the set.

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

States of the USA
{ Texas, Mississippi,… }

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

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

May not always be trivial...

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

NER: patterns

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

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

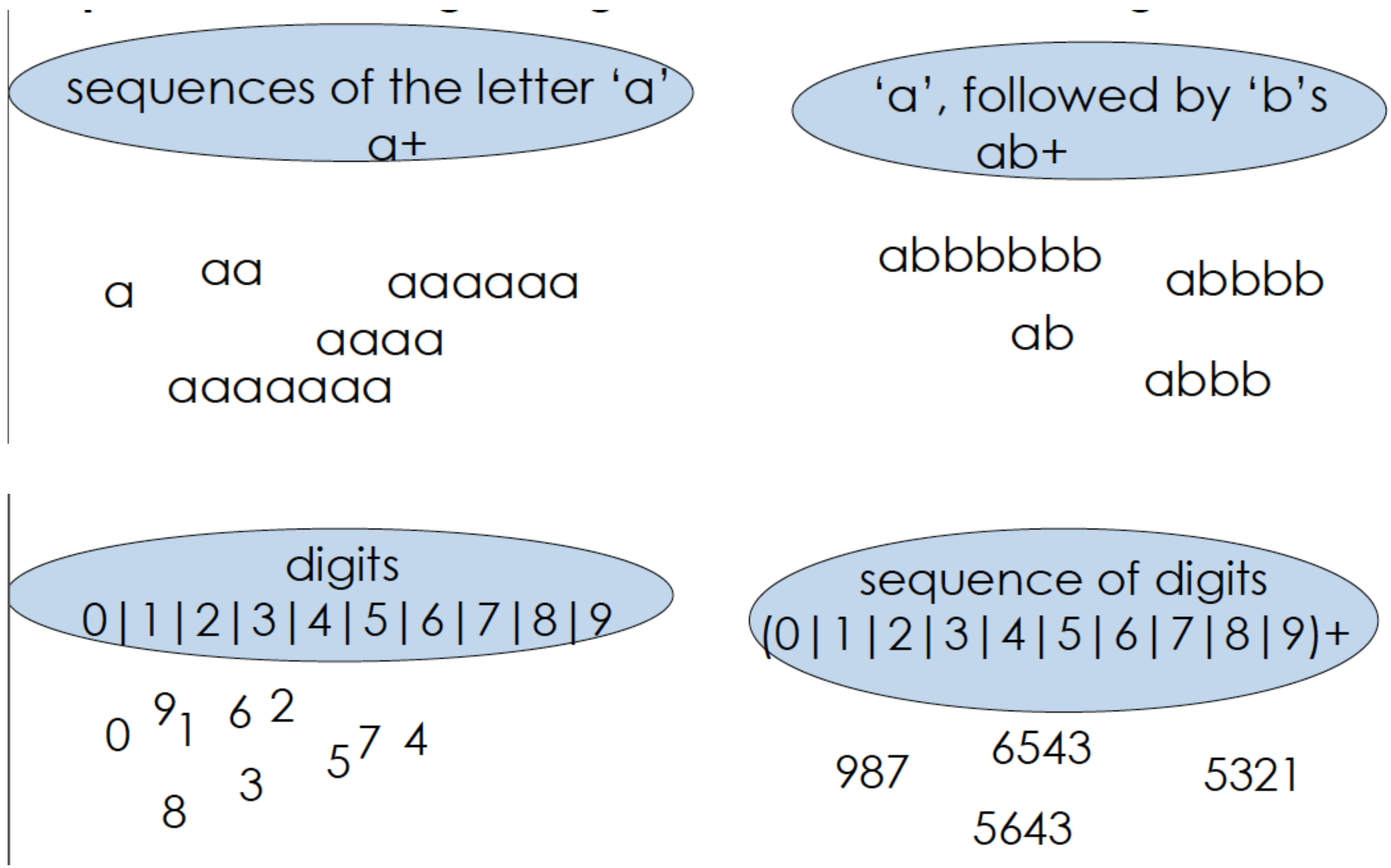
Years
(4 digit numbers)

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

Phone numbers
(groups of digits)

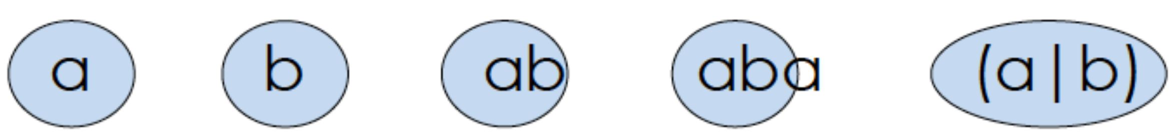
Patterns

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



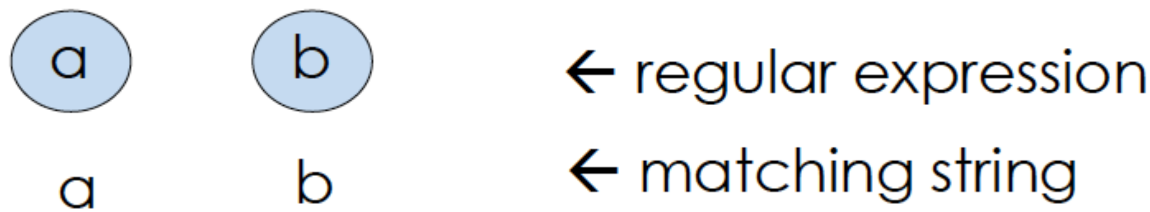
Regular Expressions (RegEx)

- A regular expression (regex) over a set of symbols Σ is:
  1. the empty string
  2. or the string consisting of an element of Σ (a single character)
  3. or the string AB where A and B are regular expressions (concatenation)
  4. or a string of the form (A|B), where A and B are regular expressions (alternation)
  5. or a string of the form (A)*, where A is a regular expression (Kleene star, like (A)+ without the empty string)
- For example, with **Σ**={a,b}, the following strings are regular expressions:

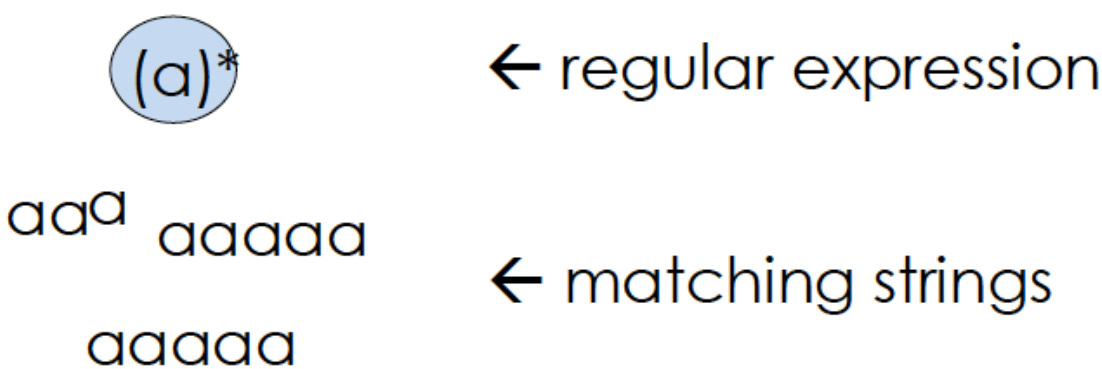


RegEx (2)

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

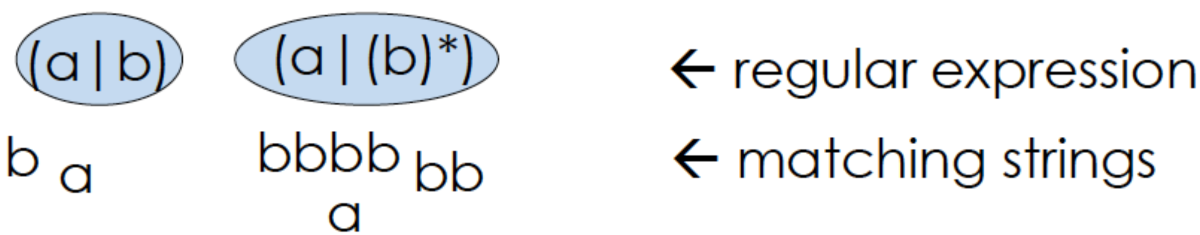


- a string matches a regular expression of the form (A)* if it consists of zero or more parts that match A

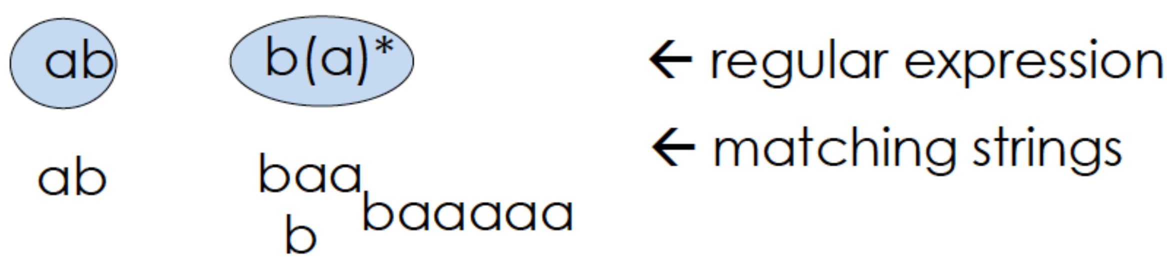


RegEx (3)

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



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



RegEx (4)

- Given an ordered set of symbols **Σ**, we define [x-y] for two symbols x and y, x<y, to be the alternation x|...|y (meaning: any of the symbols in the range)

  [0-9] = 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9

- A+ for a regex A to be A(A)* (meaning: one or more A‟s)

  [0-9]+ = [0-9][0-9]*

- A{x,y} for a regex A and integers x<y to be A...A|A...A|A...A|...|A...A (meaning: x to y A's) f{4,6} = ffff | fffff | ffffff
- A? for a regex A to be (|A) (meaning: an optional A) ab? = a(|b)
- **.** to be an arbitrary symbol from **Σ** **(wild char)**

Names & Groups in RegEx

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

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

- Parts of a regular expression can be singled out by bracketed groups (brakets; “( )” or “/ /”):

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

A Simple Exercise

- Write a regular expression to find all instances of the determiner “the”:

  /the/

  /[tT]he/

  /\b[tT]he\b/

  /(^|[^a-zA-Z][tT]he[^a-zA-Z]/

The recent attempt by the police to retain their current rates of pay has not gathered much favor with the southern factions.

A Simple Exercise

- Write a regular expression to find all instances of the determiner "the":

/the/

/[tT]he/

/\b[tT]he\b/

/(^|[^a-zA-Z][tT]he[^a-zA-Z]/

The recent attempt by the police to retain their current rates of pay has not gathered much favor with southern factions.

A Simple Exercise

- Write a regular expression to find all instances of the determiner “the”:

  /the/

  /[t|T]he/ (also capital T)

  /\b[tT]he\b/

  /(^|[^a-zA-Z][tT]he[^a-zA-Z]/

The recent attempt by the police to retain ***the****ir current rates of pay has not ga****the****red much favor with the sou****the****rn factions.*

A Simple Exercise

- Write a regular expression to find all instances of the determiner “the”:

  /the/

  /[tT]he/

  /\b[t|T]he\b/ (begin end string)

  /(^|[^a-zA-Z][tT]he[^a-zA-Z]/

The recent attempt by <u>the</u> police to retain their current rates of pay has not gathered much favor with <u>the</u> southern factions.

A Simple Exercise

- Write a regular expression to find all instances of the determiner "the":

  /the/

  /[tT]he/

  /\b[tT]he\b/

  /(^|[^a-zA-Z])[t|T]he[^a-zA-Z]/

  (nothing or no charachters before or after)

<u>The</u> recent attempt by <u>the</u> police to retain their current rates of pay has not gathered much favor with <u>the</u> southern factions.

More high-level examples

- **Create rules to extract locations**
  - Capitalized word + {city, center, river} indicates location

    Ex. *New York city*

    Hudson river
  - Capitalized word + {street, boulevard, avenue} indicates location

    Ex. *Fifth avenue*

Perl regex: http://www.cs.tut.fi/~jkorpela/perl/regexp.html

Metacharacters

char	meaning
^	beginning of string
$	end of string
.	any character except newline
*****	match 0 or more times
+	match 1 or more times
?	match 0 or 1 times; *or*: shortest match
|	alternative
( )	grouping; "storing"
[ ]	set of characters
{ }	repetition modifier
****	quote or special

Repetition

*a**	zero or more *a*'s
a+	one or more *a*'s
a?	zero or one *a*'s (i.e., optional *a*)
a{*m*}	exactly *m a*'s
a{*m*,}	at least *m a*'s
a{*m*,*n*}	at least *m* but at most *n a*'s
repetition?	same as *repetition* but the *shortest* match is taken

Perl regex

Special notations with \

Single characters

\t	tab
\n	newline
\r	return (CR)
\x*hh*	character with hex. code *hh*

"Zero-width assertions"

\b	"word" boundary
\B	not a "word" boundary

Matching

\w	matches any *single* character classified as a "word" character (alphanumeric or "_")
\W	matches any non-"word" character
\s	matches any whitespace character (space, tab, newline)
\S	matches any non-whitespace character
\d	matches any digit character, equiv. to [0-9]
\D	matches any non-digit character

Perl regex

Character sets: specialities inside [...]

Different meanings apply inside a character set ("character class") denoted by [...] so that, **instead of** the normal rules given here, the following apply:

[*characters*]	matches any of the characters in the sequence
[*x*-*y*]	matches any of the characters from *x* to *y* (inclusively) in the ASCII code
[\-]	matches the hyphen character "-"
[\n]	matches the newline; other single character denotations with \ apply normally, too
[^*something*]	matches any character *except* those that [*something*] denotes; that is, immediately after the leading "[", the circumflex "^" means "not" applied to all of the rest

Examples

<table>
<tr><th>expression</th><th>matches...</th></tr>
<tr><td>abc</td><td>abc (that exact character sequence, but anywhere in the string)</td></tr>
<tr><td>^abc</td><td>abc at the beginning of the string</td></tr>
<tr><td>abc$</td><td>abc at the end of the string</td></tr>
<tr><td>a|b</td><td>either of a and b</td></tr>
<tr><td>^abc|abc$</td><td>the string abc at the beginning or at the end of the string</td></tr>
<tr><td>ab{2,4}c</td><td>an a followed by two, three or four b's followed by a c</td></tr>
<tr><td>ab{2,}c</td><td>an a followed by at least two b's followed by a c</td></tr>
<tr><td>ab*c</td><td>an a followed by any number (zero or more) of b's followed by a c</td></tr>
<tr><td>ab+c</td><td>an a followed by one or more b's followed by a c</td></tr>
<tr><td>ab?c</td><td>an a followed by an optional b followed by a c; that is, either abc or ac</td></tr>
<tr><td>a.c</td><td>an a followed by any single character (not newline) followed by a c</td></tr>
<tr><td>a\.c</td><td>a.c exactly</td></tr>
<tr><td>[abc]</td><td>any one of a, b and c</td></tr>
<tr><td>[Aa]bc</td><td>either of Abc and abc</td></tr>
<tr><td>[abc]+</td><td>any (nonempty) string of a's, b's and c's (such as a, abba, acbabcacaa)</td></tr>
<tr><td>[^abc]+</td><td>any (nonempty) string which does not contain any of a, b and c (such as defg)</td></tr>
<tr><td>\d\d</td><td>any two decimal digits, such as 42; same as \d{2}</td></tr>
<tr><td>\w+</td><td>a "word": a nonempty sequence of alphanumeric characters and low lines (underscores), such as foo and 12bar8 and foo_1</td></tr>
<tr><td>100\s*mk</td><td>the strings 100 and mk optionally separated by any amount of white space (spaces, tabs, newlines)</td></tr>
<tr><td>abc\b</td><td>abc when followed by a word boundary (e.g. in abc! but not in abcd)</td></tr>
</table>

NER RegEx examples (in Pearl)

- Software name extraction: "*one or more capitalized words followed by a version number*" *(Mac OS X v.10.6.8)*

([A-Z]\w*\s*)+[Vv]?(\d+\.?)+.

one or more capitalized words followed by space

followed by (0 or 1) instances of V or v

followed by one or more digits, one or zero "." followed by anything else

NER RegEx examples (in Pearl)

Create regular expressions to extract:

Telephone number

blocks of digits separated by hyphens

RegEx = (\d+\-)+\d+

- **matches valid phone numbers like 900-865-1125 and 725-1234**
- **incorrectly extracts social security numbers 123-45-6789**
- **fails to identify numbers like 800.865.1125 and (800)865-CARE**

Improved RegEx = (\d{3}[-.\ ()]){1,2}[\dA-Z]{4}

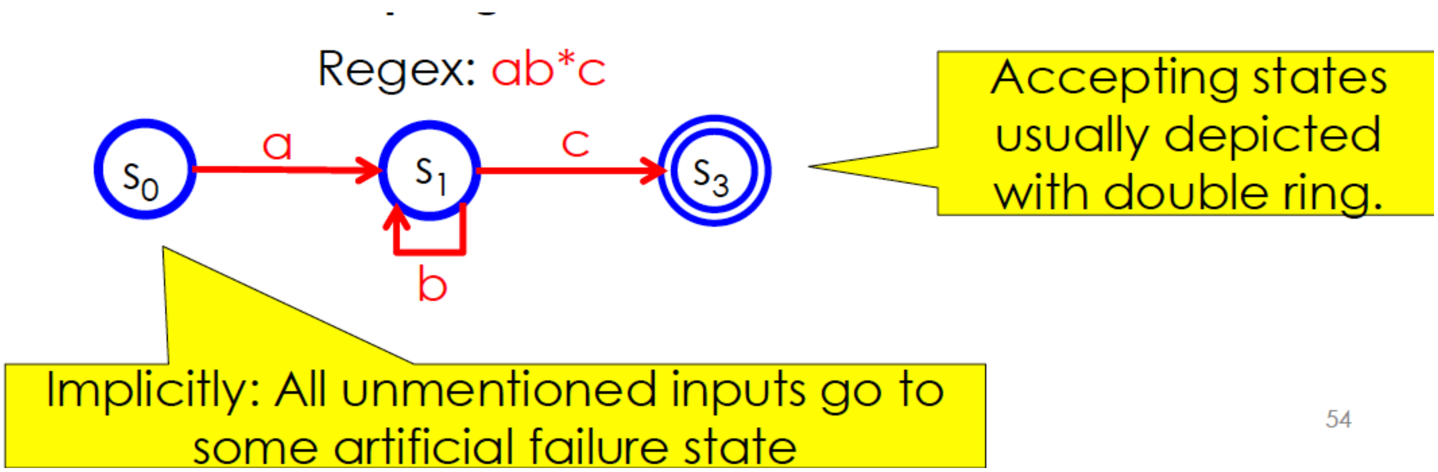
Another example

John James, Jr. Smith
John James Smith, Jr.
John James Smith Jr.
John, Jr. Smith
John Smith, Jr.
John Smith Jr.

(<first name regexp>)\s(<optional middle regexp>),\s(<optional Jr.|Sr.|II|III|IV>)\s(<last name regexp>),\s(<optional Jr.|Sr.|II|III|IV>)

Matching RegEx

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



RegEx summary

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

Input:
- Manual design of the regex

Condition:
- Entities follow a pattern

Sliding Windows

- What if we do not want to specify regexes

by hand? Use sliding windows:

- Sliding windows method is based on ML learning algorithms and annotated datasets (sentences annotated with named entities of selected types)

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

Sliding Windows

$$\text{Information } \underbrace{\text{Extraction:}}_{\text{Prefix window}}\ \underbrace{\text{Tuesday}}_{\text{Content window}}\ \underbrace{\text{10:00 am}}_{\text{Postfix window}}\text{, Rm 407b}$$

- Choose certain features (properties) of windows that could be important:
  - window contains colon, comma, or digits
  - window contains week day, or certain other words
  - window starts with lowercase letter
  - window contains only lowercase letters
  - ...

Feature Vectors

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

Prefix window | Content window | Postfix window

Features	Feature Vector
Prefix colon	1
Prefix comma	0
...	...
Content colon	1
Content comma	0
...	...
Postfix colon	0
Postfix comma	1

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

Sliding Windows Corpus

time location

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

From this corpus, compute the feature vectors with labels:

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

Nothing Nothing Time Nothing Location

Machine Learning

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

Use the labeled feature vectors as training data for Machine Learning

$$\begin{bmatrix}1\\0\\0\\0\\1\\1\\1\end{bmatrix} \quad \begin{bmatrix}1\\1\\0\\0\\1\\0\end{bmatrix} \quad \rightarrow$$

Nothing Location

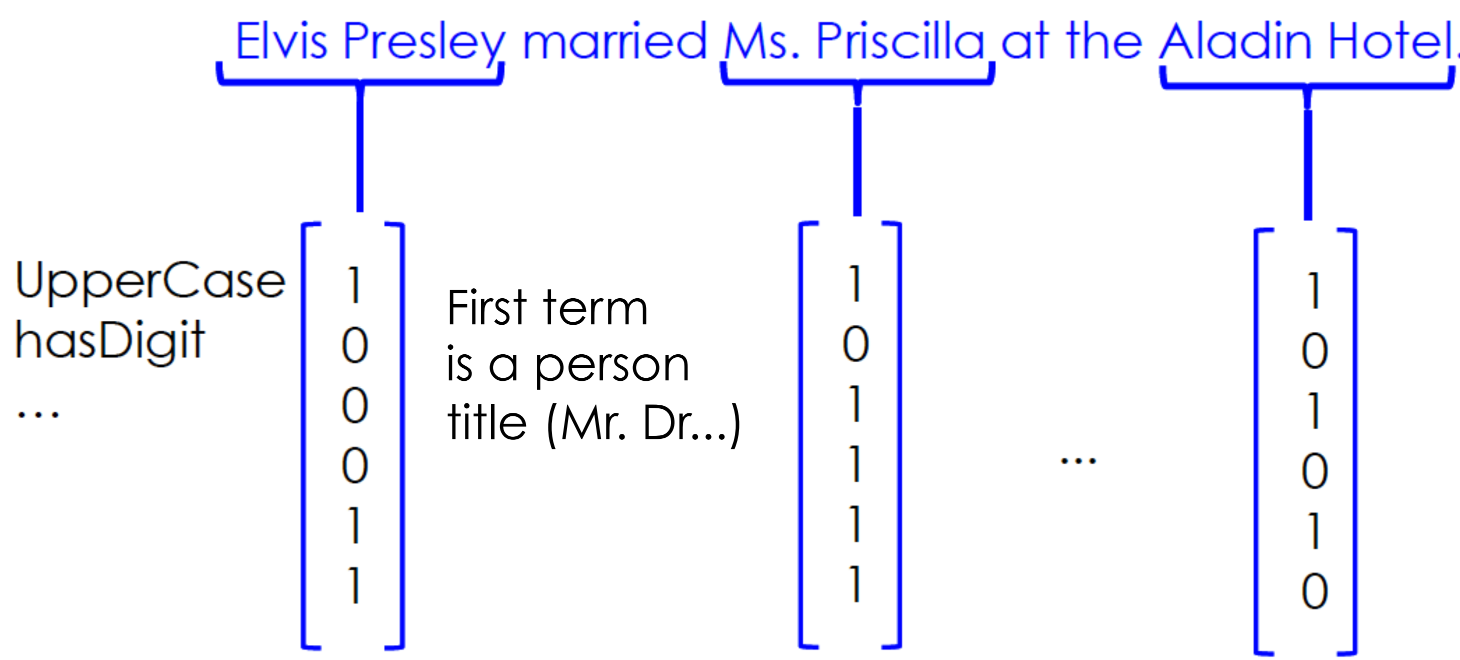
Machine Learning

classify

$$\begin{bmatrix}1\\0\\1\\0\\1\\0\end{bmatrix} \xrightarrow{\text{Result}} \text{Time}$$

Sliding Windows Exercise

- What features would you use to recognize person names?

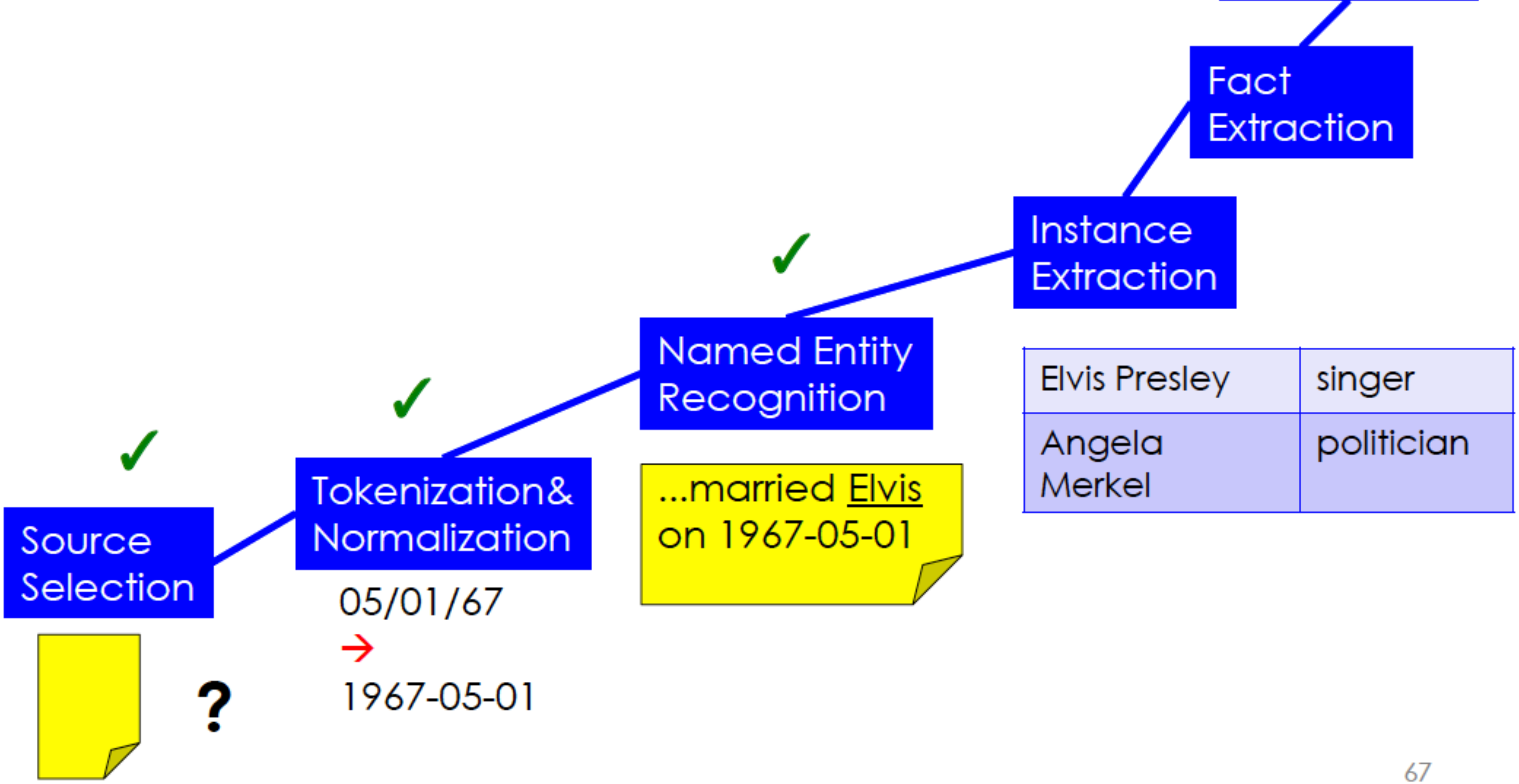


NER summary (but we learned general techniques..)

- Named Entity Recognition **(NER)** is the process of finding entities (people, cities, organizations, ...) in a text.
  - We have seen different techniques
  - Closed-set extraction (if the set of entities is known)
  - Extraction with Regular Expressions (if the entities follow a pattern). Can be done efficiently with Finite State Automata
  - Extraction with sliding windows / Machine Learning (if the entities share some syntactic features)

Information Extraction

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



Instance/relation Extraction

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

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

Entity	Class
Elvis	artist
Oh yeah, honey	song
Hintertuepflingen	location

Instance/relation extraction with patterns

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

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

Entity	Class
Elvis	artist

Pattern: X was a great Y

Instance/relation extraction using patterns

Elvis was a great artist

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

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

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

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

Can write RegEx for each pattern

Instance/relation extraction using patterns

- Manually writing patterns is difficult
- Lexical patterns do not generalize, e.g.
  - Elvis was a great pianist
  - Elvis was a pianist
  - Elvis, the pianist..
- Learn patterns with ML techniques
- Generalise patterns using HMM or lattices

Learning hypernym relations from definitions

Information Extraction and beyond

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

Ontological Information Extraction

Fact Extraction

Instance Extraction

Elvis Presley	singer
Angela Merkel	politician

Named Entity Recognition

...married <u>Elvis</u> on 1967-05-01

Tokenization& Normalization

05/01/67 → 1967-05-01

Source Selection

?

Fact extraction

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

Costello Sings Lowe/Nick Sings Elvis (late show)
THE BAND: Paul Revelli, Ruth Davies, Bill Kirchen, Bob Andrews,Derek Huston, Austin ...

10/1/2010 Friday 11:00p
Great American Music Hall, San Francisco CA
Featuring: Elvis Costello, Nick Lowe
BUY

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

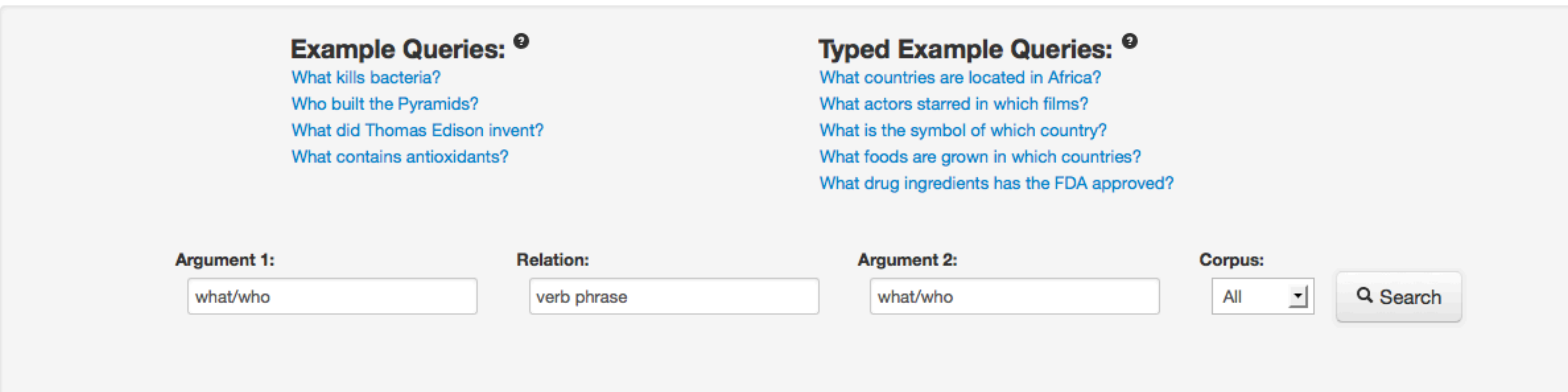
Fact Extraction

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

ReVerb (http://openie.cs.washington.edu/)

Pattern Based approach: read “Identifying Relations for Open Information Extraction” EMNLP 2011

Freely downloadable and available on-line



ReVerb

Argument 1: hammer Relation: used Argument 2: All Search

93 answers from **138 sentences**

all | location (8) | newspaper circulation area (5) | invention (3) | misc.

- a weapon (14)
- *html files* (4)
- nails (4)
- conjunction (3)
- excitation (3)
- *religious amulets* (3)
- hard materials (3)
- jewelry including gold (3)
- the operation of firearms (2)
- *specific locations* (2)
- the world (2)
- carpenters (2)
- **Portland cement** (2)
- protective talismans (2)
- some coat of arms (2)

nails

Extracted Synonyms:

a nail

Extracted from these sentences:

is used for Just as **a hammer** is used for **nails** , and a screwdriver is used for screws , so each type of geometric structure requires its own geometry . (via Google)

Round inches can generally be substituted for square inches in geometry when calculating the area of circles. Just as circular mils are used to measure the area of wire , so round inches can be used to measure the areas of circles .Of course , just as there is no exact way to measure circular areas in terms of straight lines , there is no way to exactly measure the area of squares and rectangles using circular geometry . Just as **a hammer** is used for **nails** , and a screwdriver is used for screws , so each type of geometric structure requires its own geometry . (via ClueWeb09)

The bottom line ; just as **a hammer** is used for **nails** and a screw drive for screws , assembly and compiled code each have a place in embedded applications . (via Google)

It s just choosing the right tool for the job , and mf and dslrs are like a hammer and screwdriver **a hammer** is used for **a nail** and a screwdriver for a screw . (via ClueWeb09)

Information Extraction

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

and beyond

Ontological Information Extraction

Fact Extraction ✓

Instance Extraction ✓

Named Entity Recognition ✓

Tokenization& Normalization ✓

Source Selection ✓

Person	Nationality
Angela Merkel	German

Ontology Extraction

- An ontology is consistent knowledge basewithout redundancy
- Every entity appears only with exactly the same name
- There are no semantic contradictions

Person	Nationality
Angela Merkel	German
Merkel	Germany
A. Merkel	French

Entity	Relation	Entity
Angela Merkel	citizenOf	Germany

Ontology Extraction

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

Entity	Relation	Entity
Angela Merkel	citizenOf	Germany

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

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

Person	Nationality
Angela Merkel	German
Merkel	Germany
A. Merkel	French

120

Ontological IE Challenges

- Challenge 1

Map names to names that are already known

Entity	Relation	Entity
Angela Merkel	citizenOf	Germany

Merkel

Angie

A. Merkel

Ontological IE Challenges

- Challenge 2

Be sure to map the names to the right known names

Entity	Relation	Entity
Angela Merkel	citizenOf	Germany
Una Merkel	citizenOf	USA

?

Merkel is great!

Ontological IE Challenges

- Challenge 3

Map to known relationships

Entity	Relation	Entity
Angela Merkel	citizenOf	Germany

… has nationality …
… has citizenship …
… is citizen of …

Ontological IE Challenges

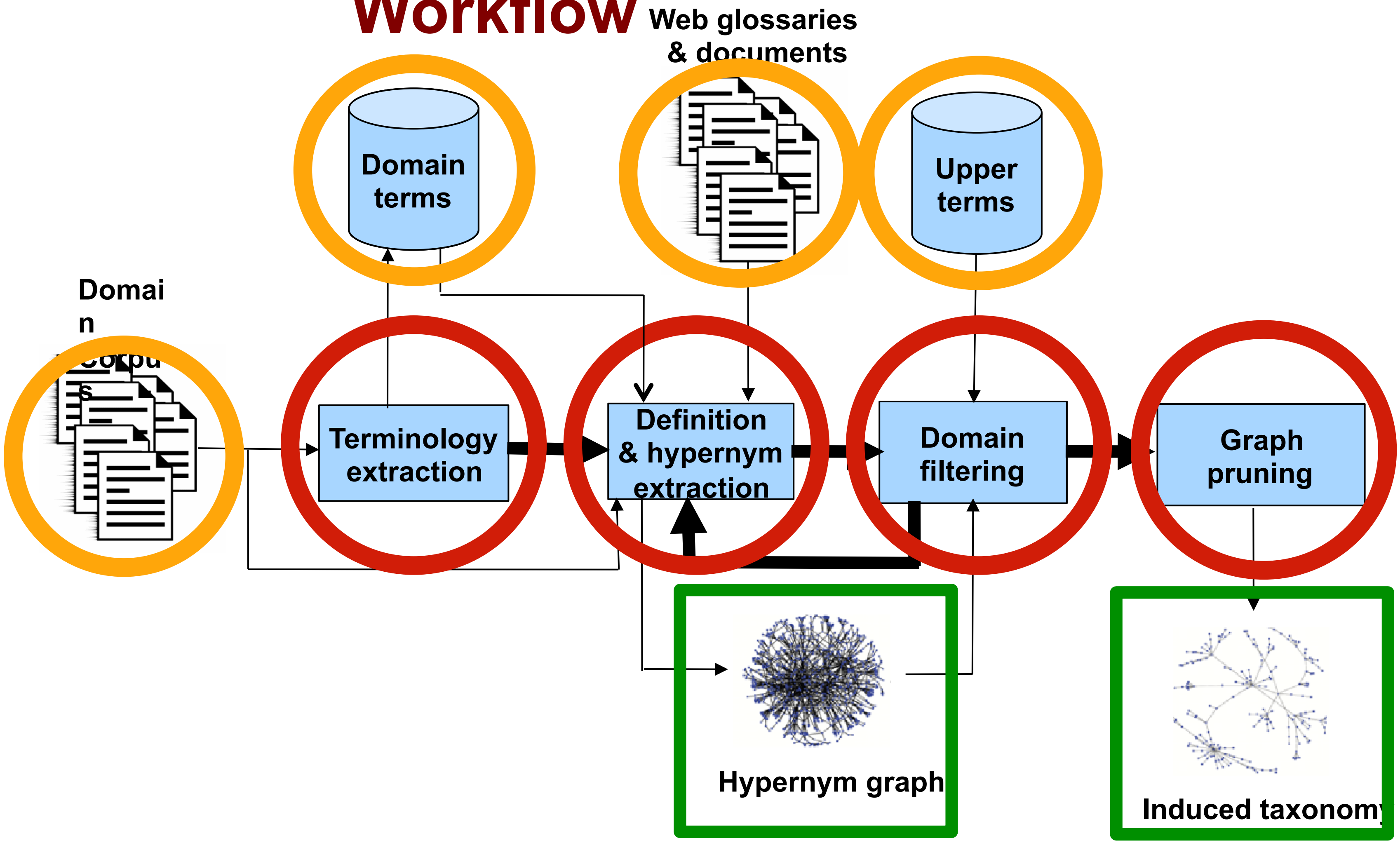
- Challenge 4
- Find hypernymy relations

Chancellor of Germany

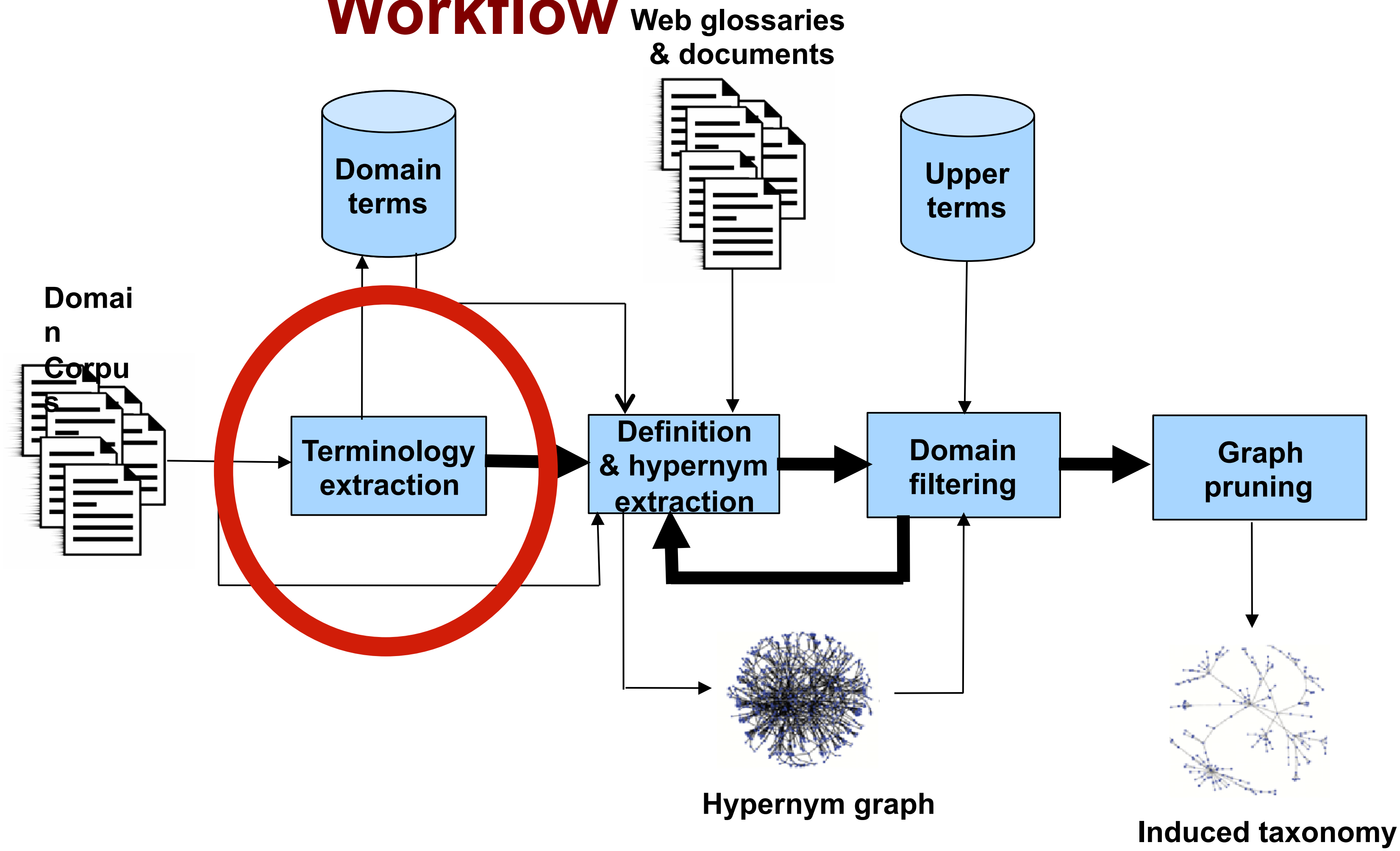
Ontolearn (IJCAI 2011, CL 2013)

Inducing lexical taxonomies from scratch

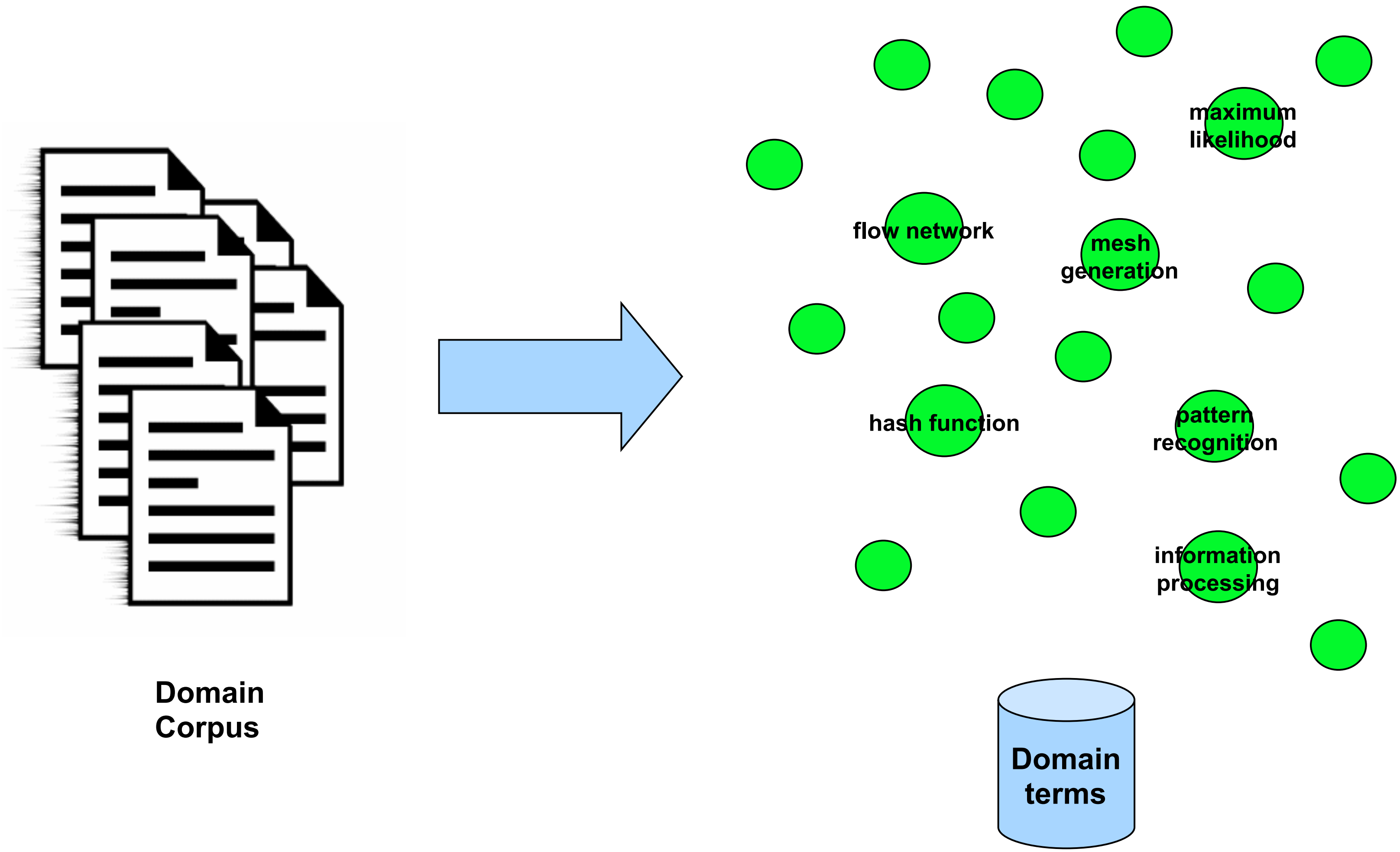
Taxonomy Learning Workflow



Taxonomy Learning Workflow

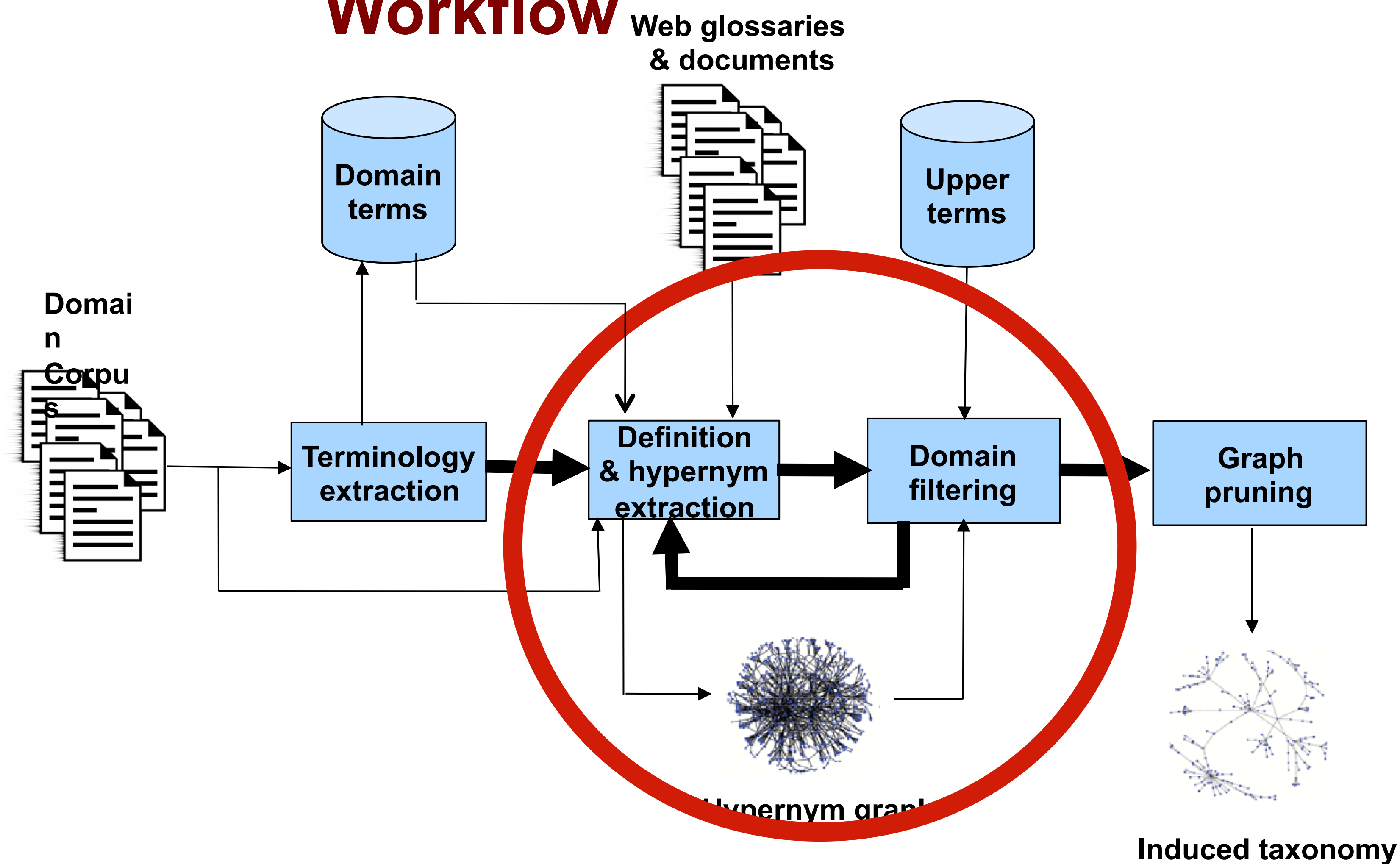


Terminology Extraction

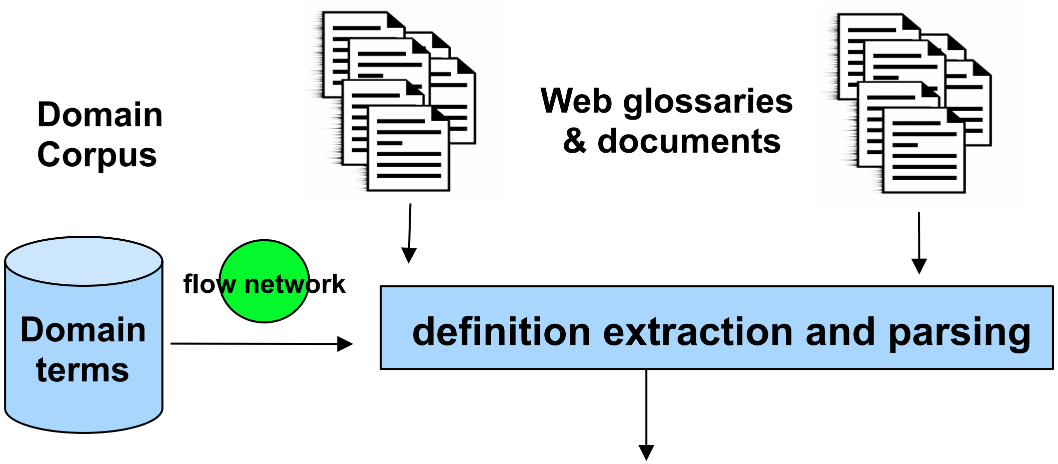


several on- line terminology extractors (e.g. TermExtractor http://hal.di.uniroma1.it/termextractor/public/demo.faces)

Taxonomy Learning Workflow

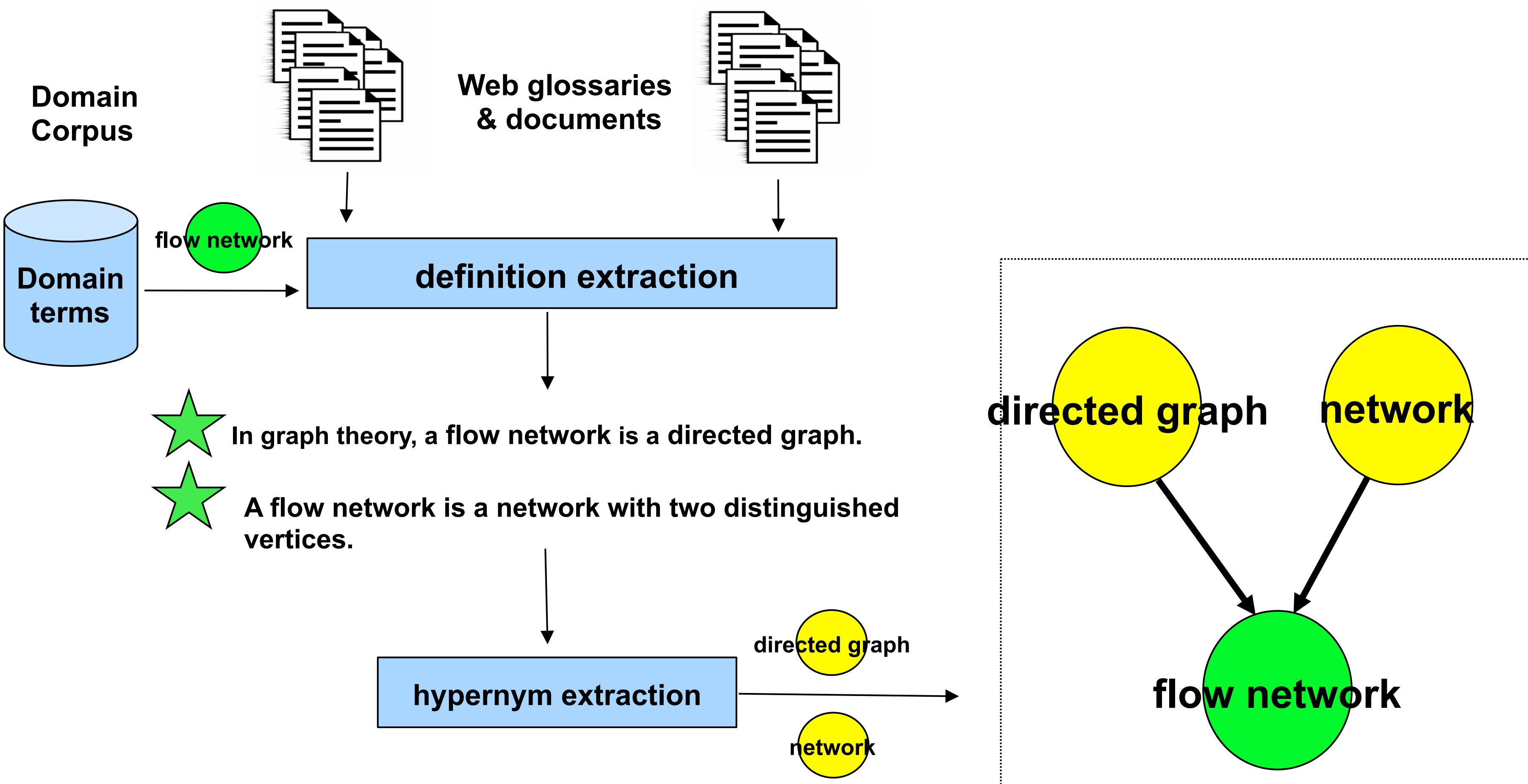


Definition & Hypernym Extraction + Domain Filtering

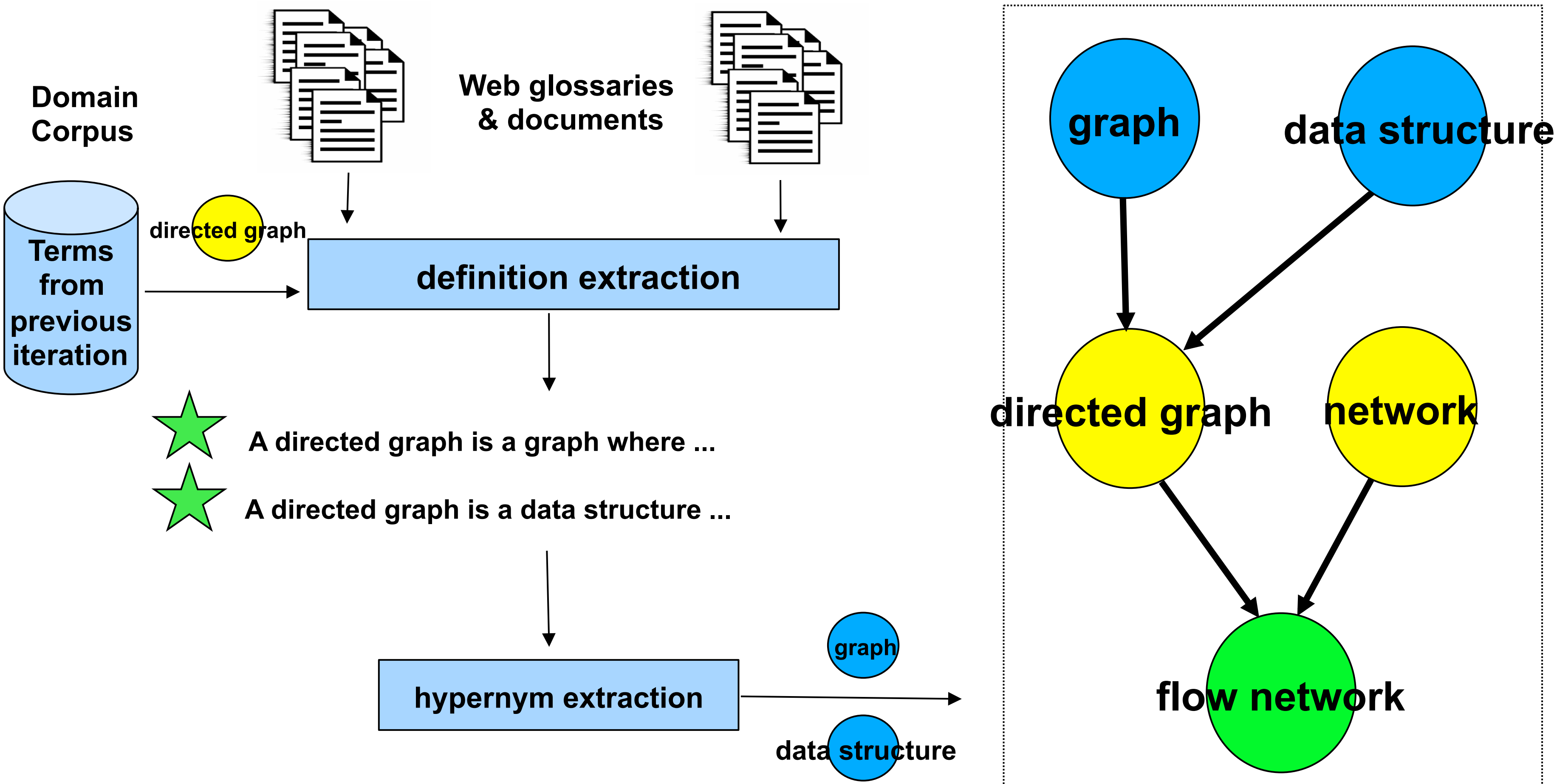


	domain	non domain
In graph theory, a flow network is a directed graph.	★	
Global Cash Flow Network is a business opportunity to make money online.		★
A flow network is a network with two distinguished vertices.	★	

Definition & Hypernym Extraction + Domain Filtering



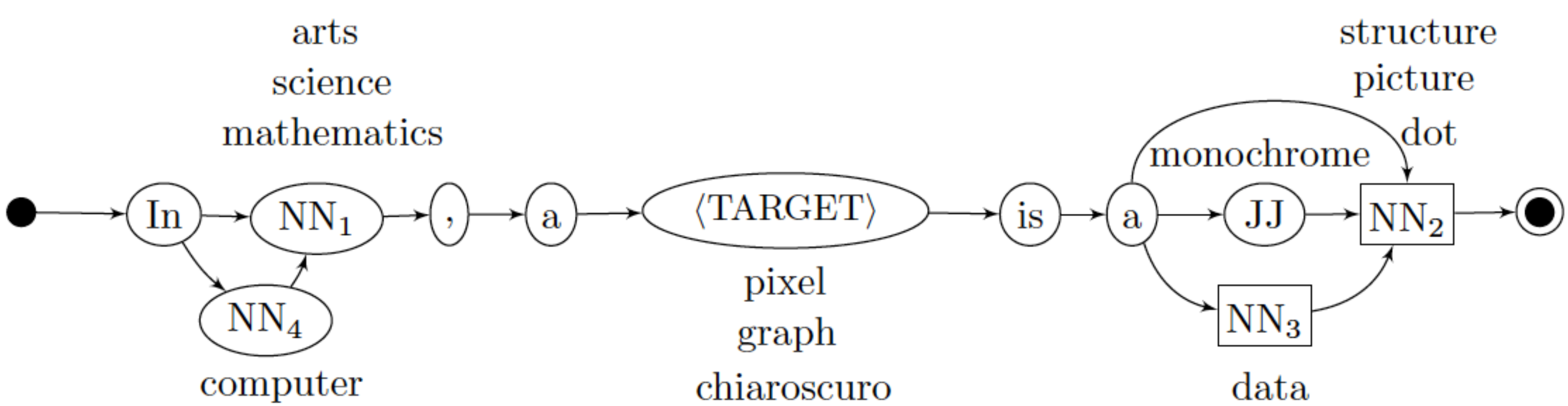
Definition & Hypernym Extraction + Domain Filtering



Hypernym Extraction Algorithm

Based on Word-Class Lattices, i.e. lattice-based definition models learned by means of a greedy definition alignment algorithm

- Determine whether a sentence is definitional
- If so, returns the hypernym(s) of the defined term



Performance in Definition Extraction

Algorithm	P	R	F_1	A
WCL-1	**99.88**	42.09	59.22	76.06
WCL-3	**98.81**	60.74	**75.23**	**83.48**
Star patterns	86.74	66.14	**75.05**	81.84
Bigrams	66.70	**82.70**	73.84	75.80
Random BL	50.00	50.00	50.00	50.00

Wikipedia

Algorithm	P	R†
WCL-1	**98.33**	39.39
WCL-3	**94.87**	56.57
Star patterns	44.01	**63.63**
Bigrams	46.60	45.45
Random BL	50.00	50.00

UKWac corpus

Outperforms existing methods for definition extraction

Precision in Hypernym Extraction

Algorithm	Full	Substring
WCL-1	**42.75**	77.00
WCL-3	40.73	**78.58**

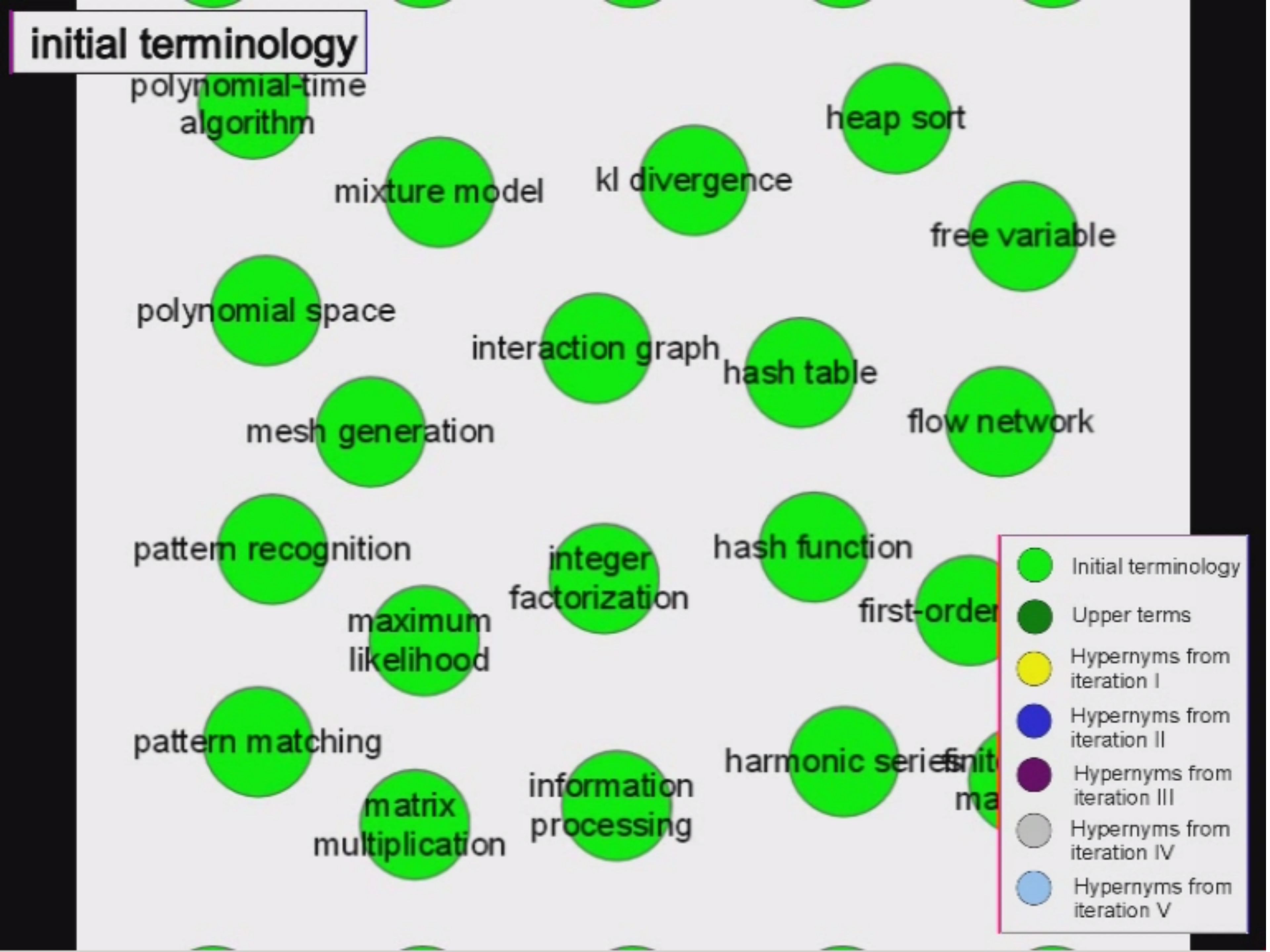
Wikipedia

Algorithm	Full		Substring	
WCL-1	86.19	(206)	**96.23**	(230)
WCL-3	**89.27**	(383)	**96.27**	(413)
Hearst	65.26	(62)	88.42	(84)

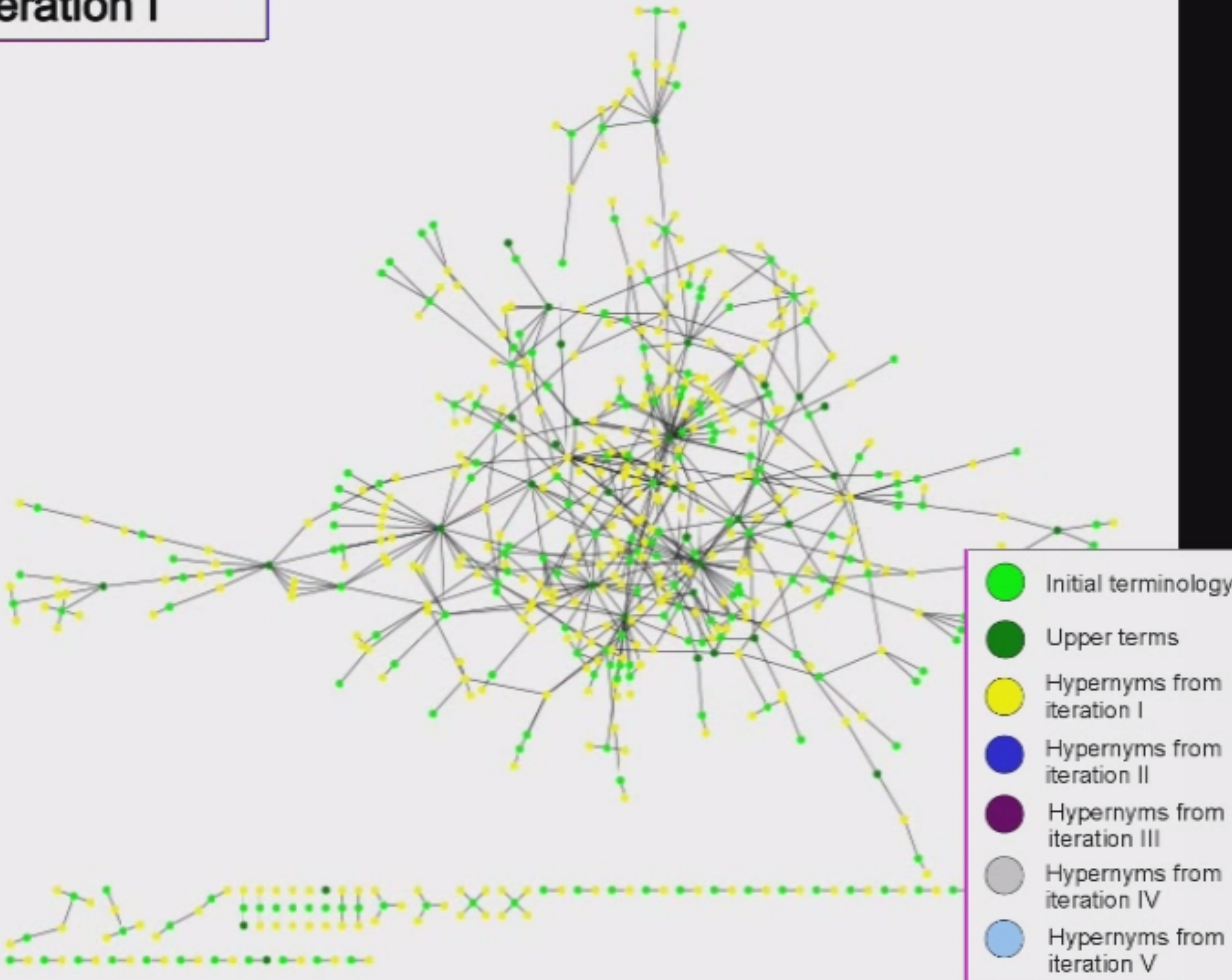
UKWac

Pattern-based methods achieve much lower recall: 62 vs. 383 hypernyms extracted from UKWac

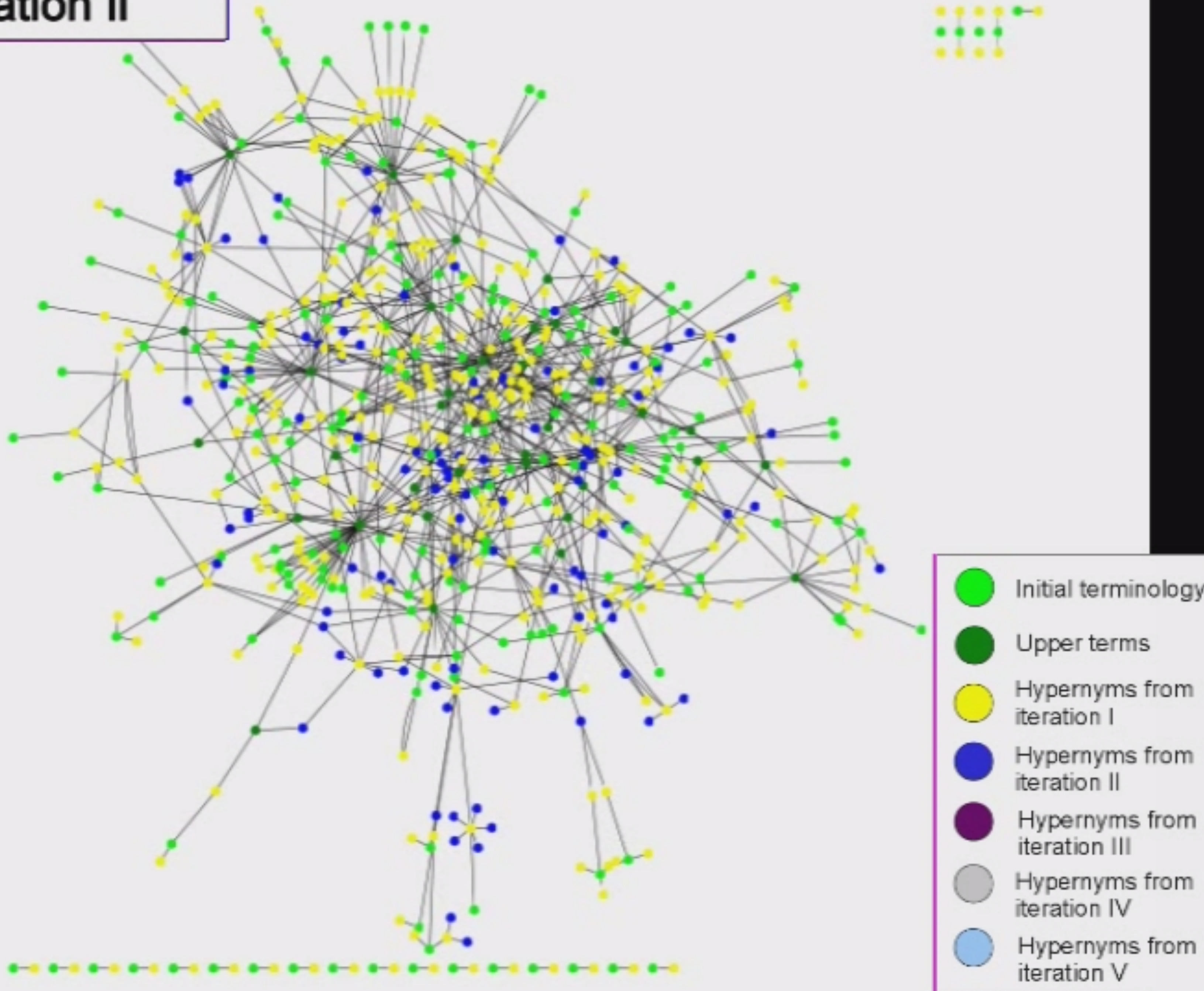
The iterative growth of the hypernym graph


initial terminology
polynomial-time algorithm
heap sort
mixture model
kl divergence
free variable
polynomial space
interaction graph
hash table
mesh generation
flow network
pattern recognition
integer factorization
hash function
first-order
maximum likelihood
pattern matching
harmonic series
information processing
matrix multiplication
Initial terminology
Upper terms
Hypernyms from iteration I
Hypernyms from iteration II
Hypernyms from iteration III
Hypernyms from iteration IV
Hypernyms from iteration V

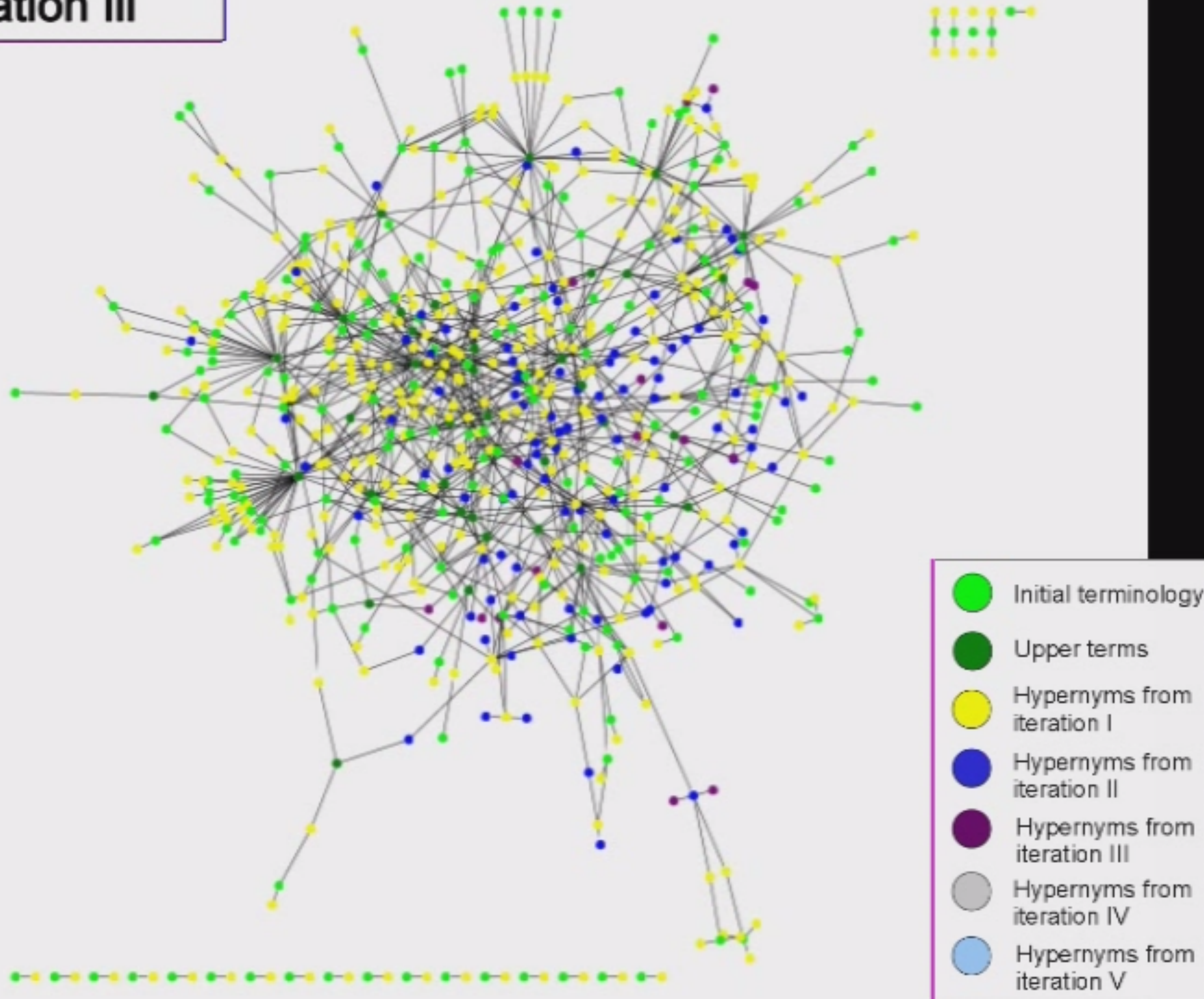
Iteration I


Initial terminology
Upper terms
Hypernyms from iteration I
Hypernyms from iteration II
Hypernyms from iteration III
Hypernyms from iteration IV
Hypernyms from iteration V

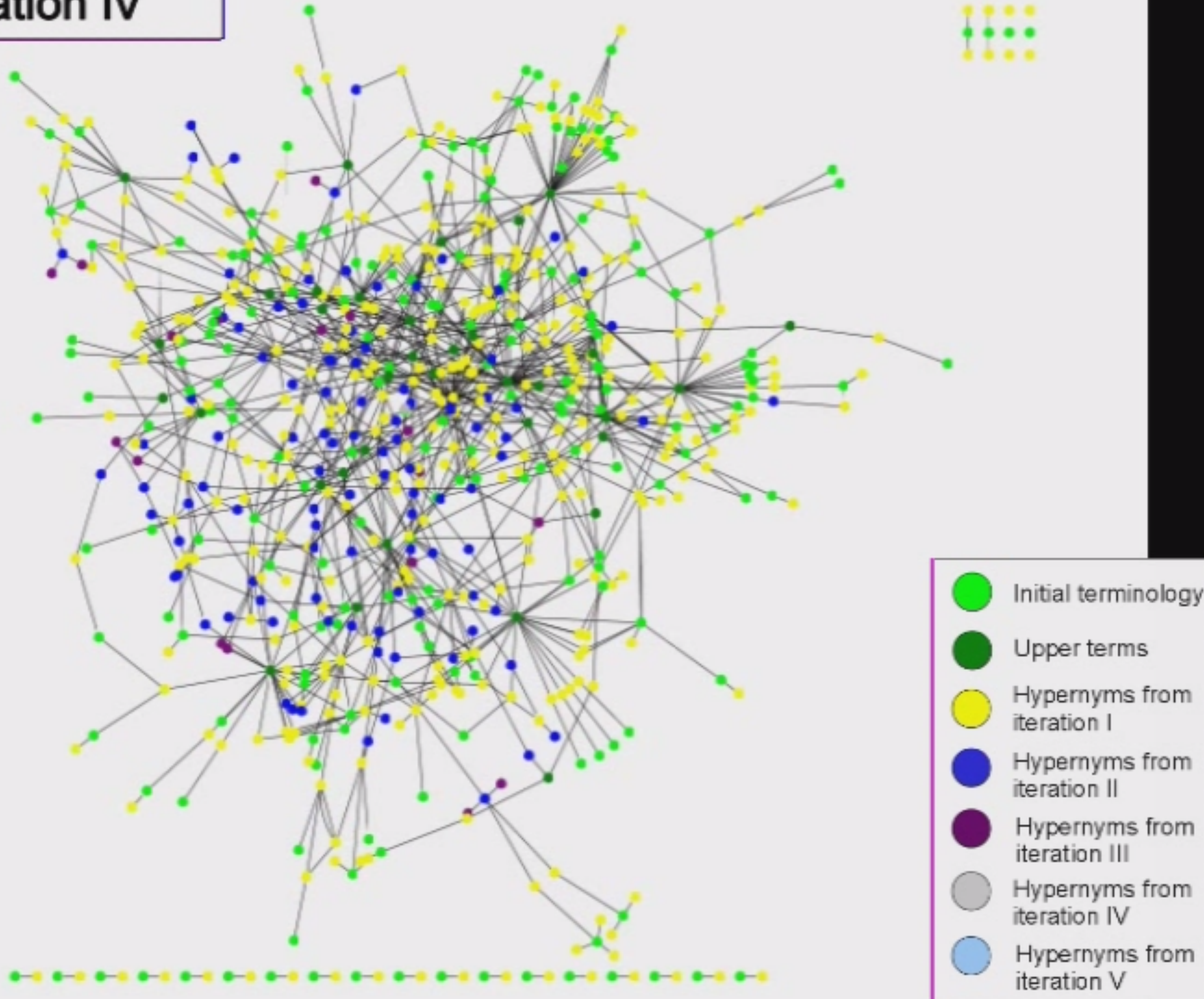
Iteration II



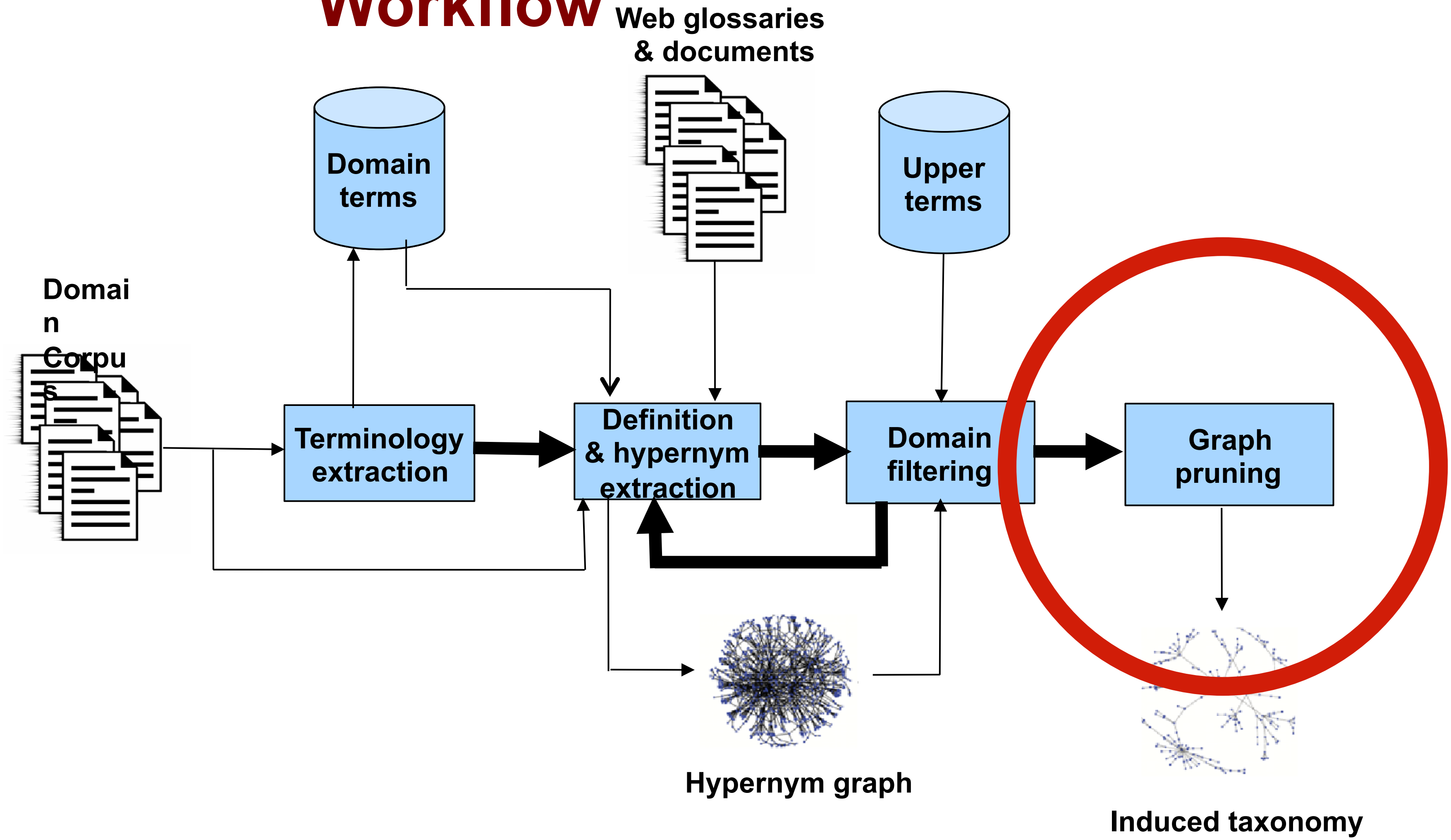
Iteration III



Iteration IV



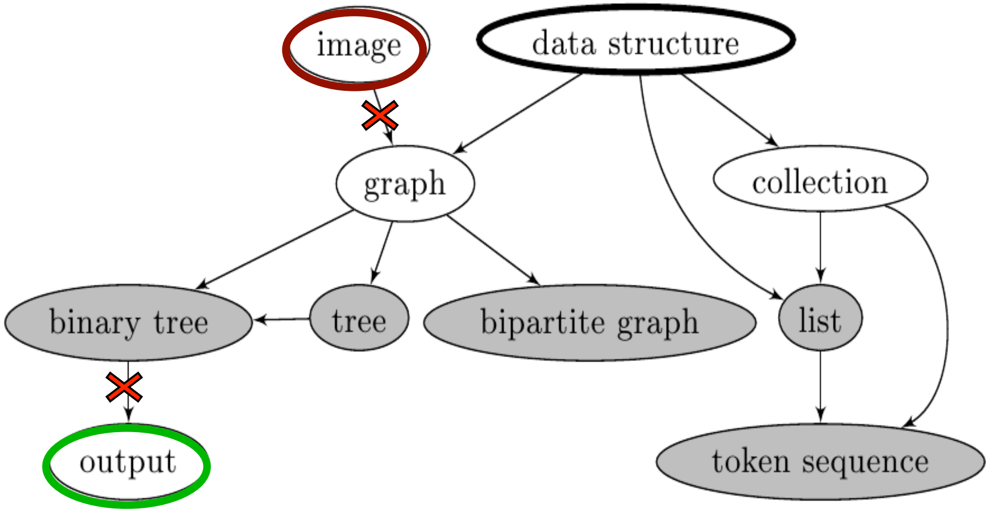
Taxonomy Learning Workflow



Graph Pruning

- Goal:
  - From noisy hypernym graph
  - To full-fledged taxonomy
- How:
  - A graph-based algorithm

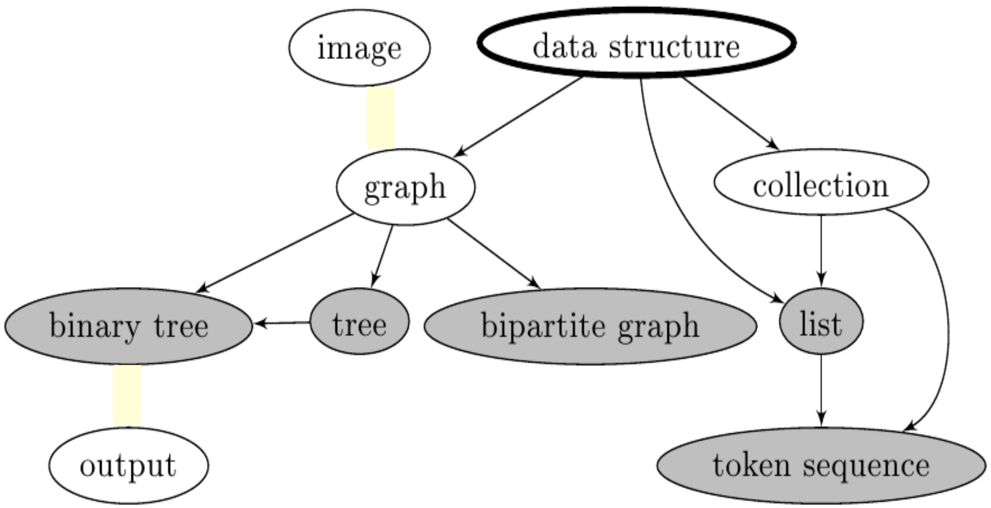
Graph Pruning: Preliminary step



Given the hypernym graph, we disconnect:

- false roots (root nodes not in the set of upper terms)
- false leaves (leaf nodes not in the initial terminology)

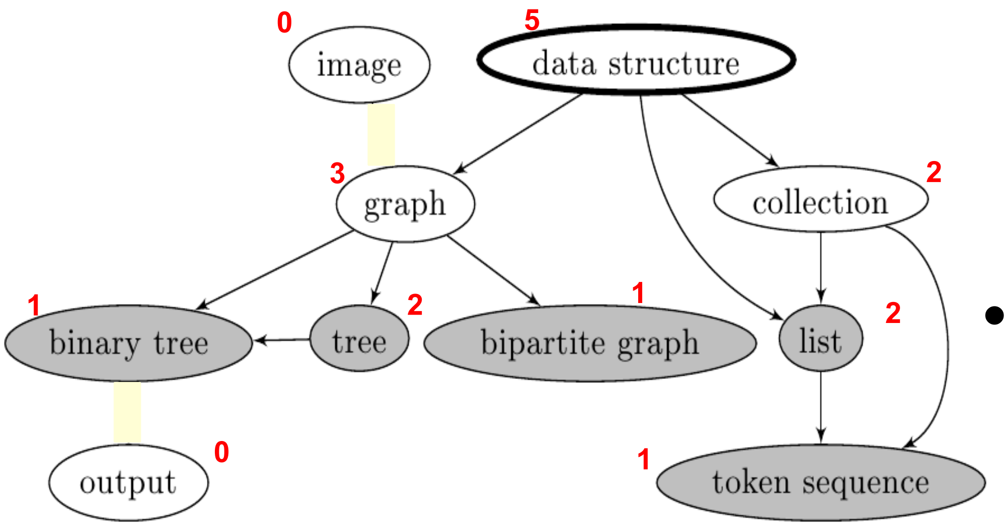
Graph Pruning: Preliminary step



Given the hypernym graph, we disconnect:

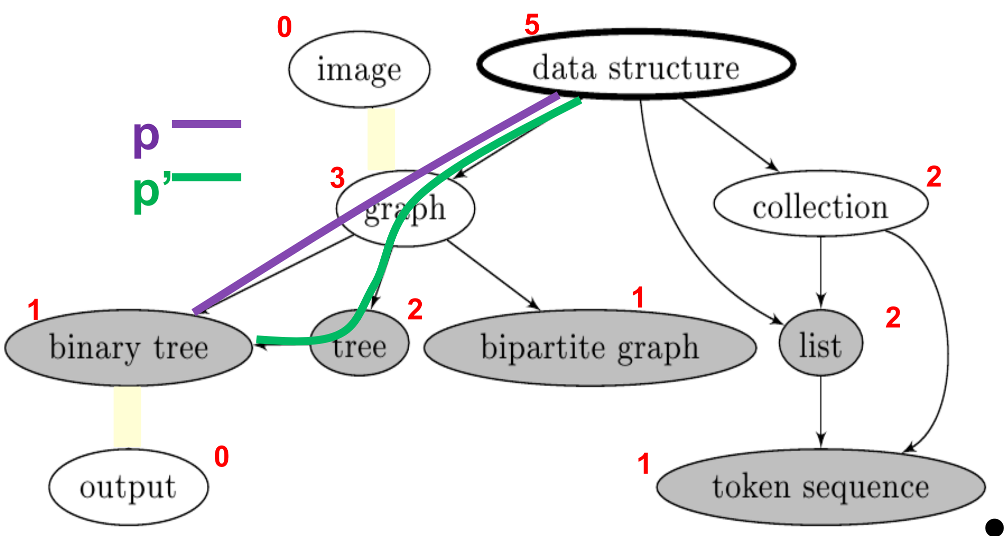
- false roots (root nodes not in the set of upper terms)
- false leaves (leaf nodes not in the initial terminology)

Graph Pruning (1)



- Weight each node *v* by the number of terminological nodes reachable from *v*
- E.g.:
  - w(collection) = 2
  - w(graph) = 3

Graph Pruning (2)

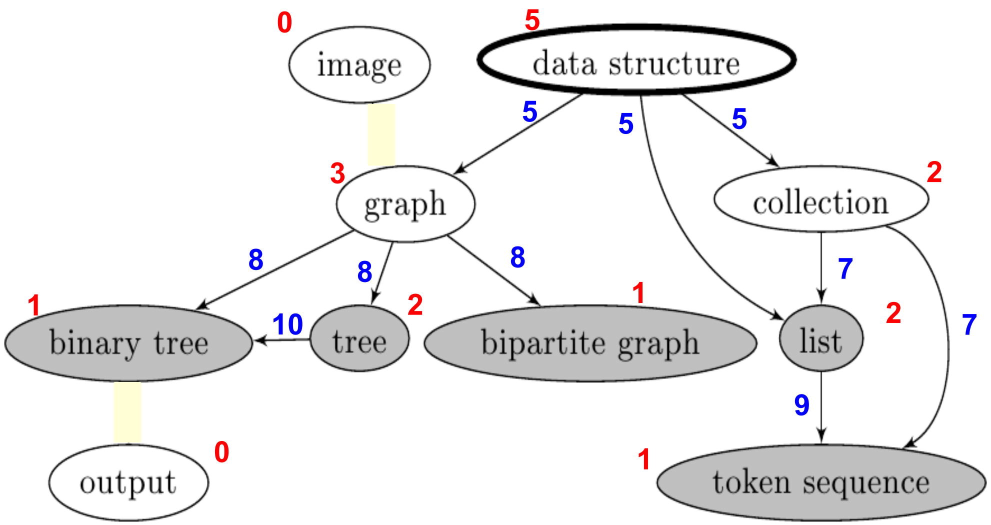


- For each path *p* from an upper term *r* to a node *v*, we calculate its cumulative weight:

$$\omega(p) = \sum_{v' \in p} w(v')$$

- E.g.:
  - $\omega(p) = 5+3 = 8$
  - $\omega(p') = 5+3+2 = 10$

Graph Pruning (3)

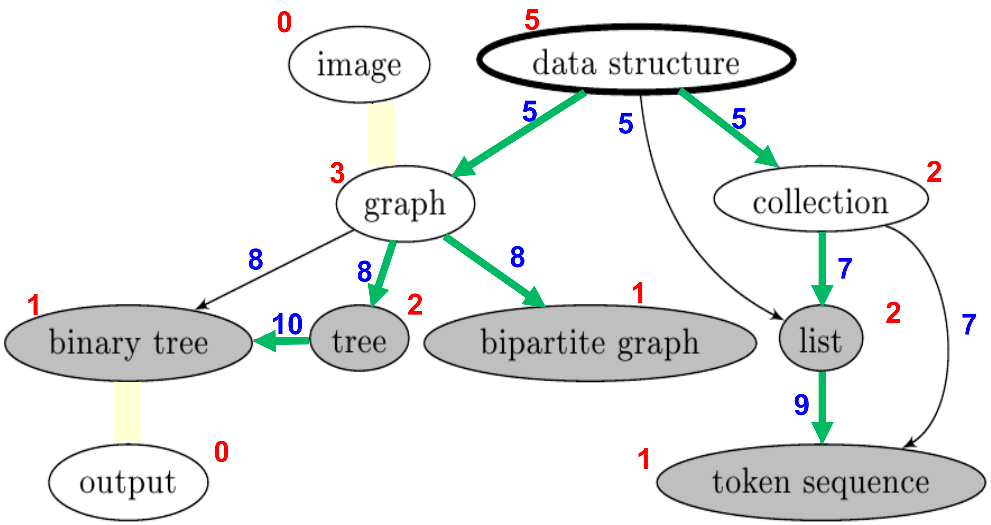


- Assign to each edge (h, v) the maximum cumulative weight among all the paths from any upper term to node v

$$w(h,v) = \max_{r \in U} \max_{p \in \Gamma(r,h)} \omega(p)$$

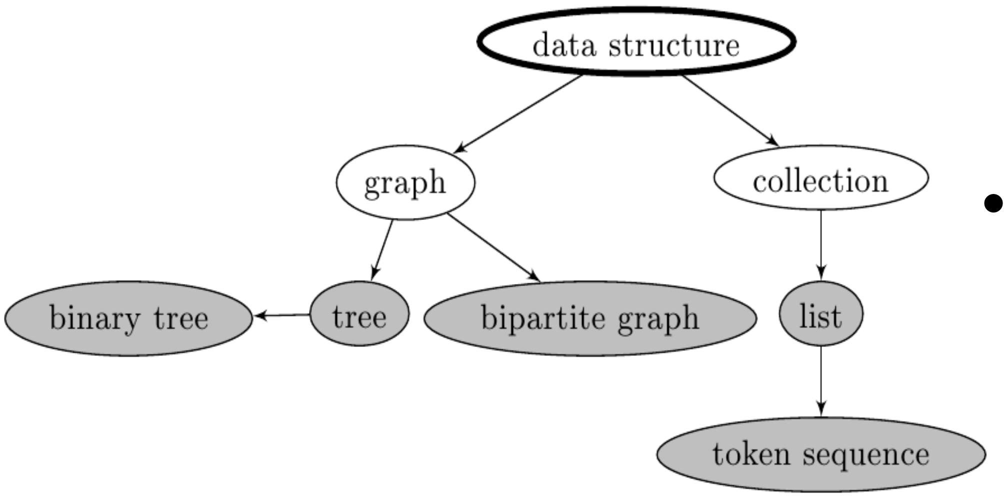
- E.g.:
  - w(graph, binary tree) = 8
  - w(tree, binary tree) = 10

Graph Pruning (4)



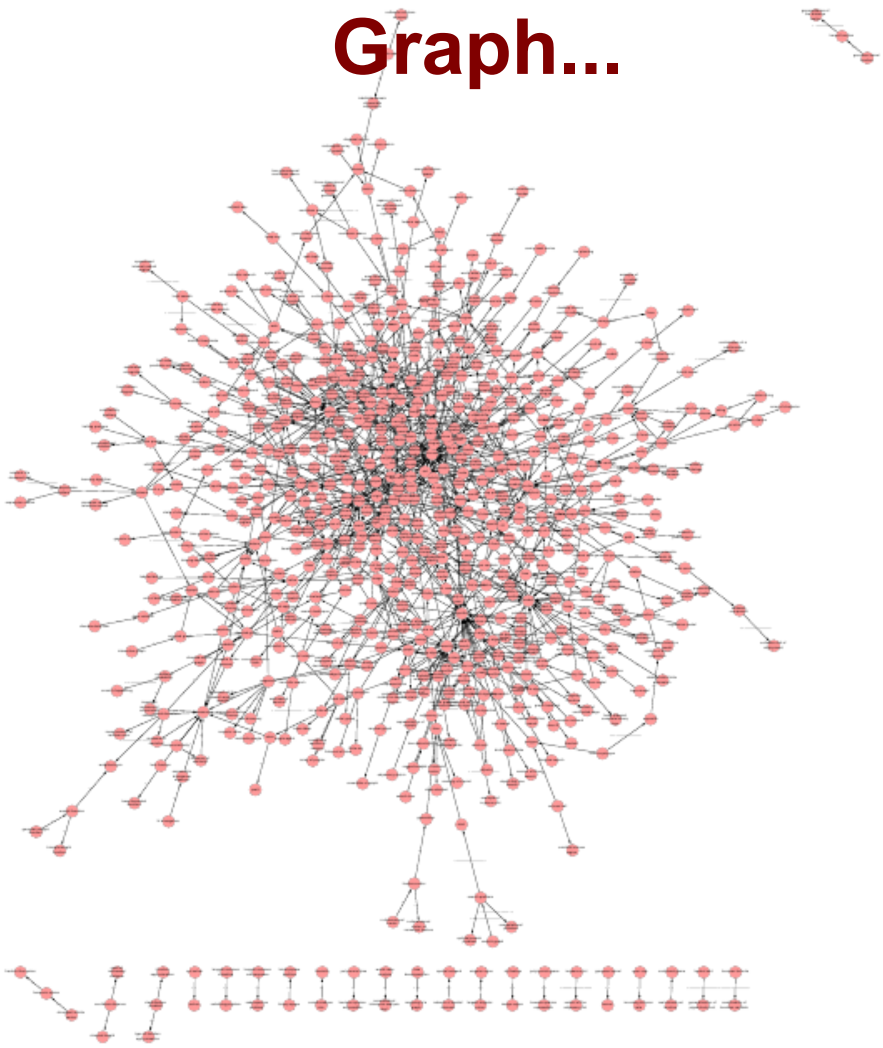
- Apply the Chu-Liu Edmonds [1967] algorithm

Graph Pruning (4)

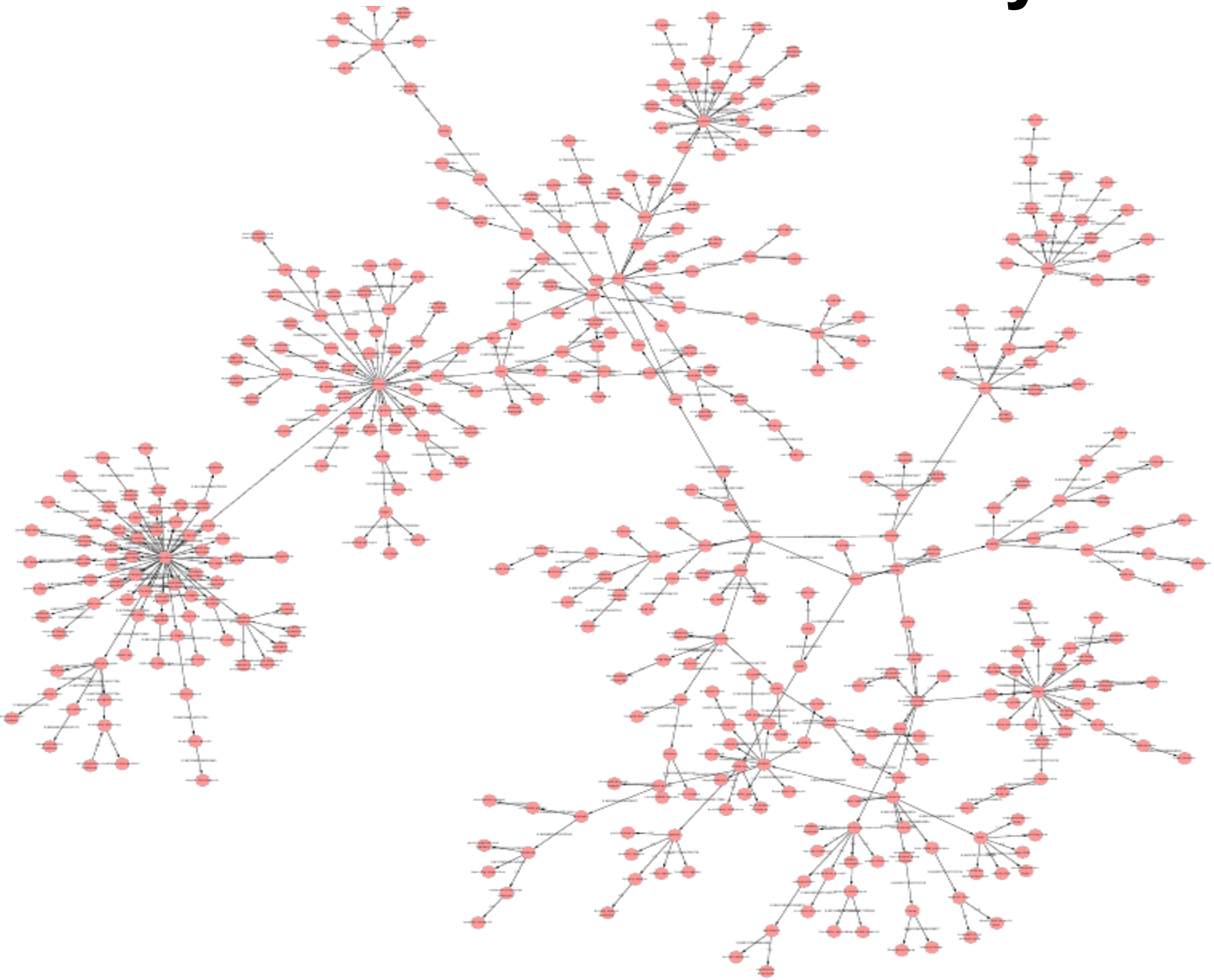


- Apply the Chu-Liu Edmonds [1967] algorithm
- As a result we obtain a tree-like taxonomy

From the Noisy Hypernym Graph...



...to a Tree-like Taxonomy



Pruning Recovery

- Many small connected components will be returned
- To recover from excessive pruning we apply a simple heuristic:
  - Let *r* be the root of such a component
  - Select the best-ranking edge (*v*, *r*) according to the domain score of the corresponding definition
  - If no edge exists, we use string inclusion
    - E.g., r=binary tree and v=tree

Evaluation

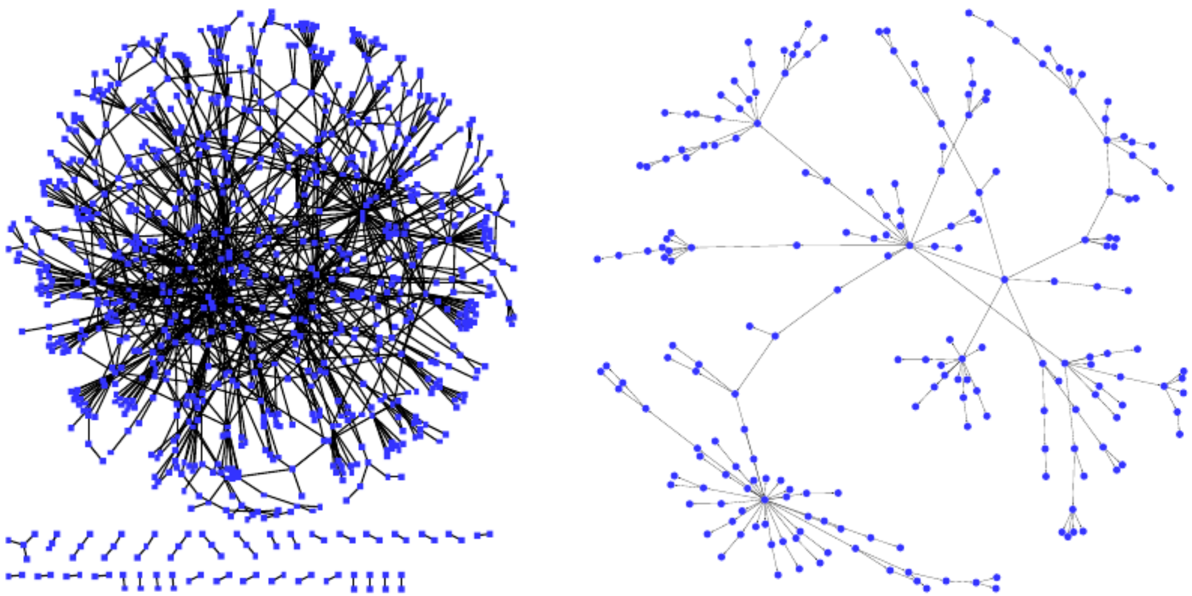
- Taxonomy evaluation is a hard task
- We performed two different experiments:
  - Inducing an Artificial Intelligence (AI) taxonomy
  - Reproducing existing sub-hierarchies in WordNet

Experiment 1: Inducing an AI Taxonomy

- Corpus: IJCAI 2009 proceedings (334 papers)
- Domain terminology: via term extraction
  - 374 initial domain terms
- Upper terms: we manually selected 13 terms (*process*, *abstraction*, *algorithm*)
  - Used as a stopping criterion for definition/hypernym extraction
- Definition and hypernym extraction: IJCAI corpus + Google define
  - 715 nodes and 1025 edges

Experiment 1: Inducing an AI Taxonomy

- Final graph: 427 nodes, 426 edges
- Compression ratio (against unpruned graph): 0.60 (nodes), 0.41 (edges)



Experiment 1: Inducing an AI Taxonomy

- Final graph: 427 nodes, 426 edges
- Compression ratio (against unpruned graph): 0.60 (nodes), 0.41 (edges)
- Manual evaluation of edge precision: 81.5% (347/426)
  - Note: many hypernyms would be equally valid (collaborative assessment?)
- The AI Taxonomy is available to the community:
  - http://lcl.uniroma1.it/taxolearn

Experiment 2: Evaluation against WordNet

- Same evaluation strategy as in Kozareva & Hovy (EMNLP 2010)
  - Domains: animals, plants, vehicles
- No terminology extraction: we use the terminology provided by K&H for each domain
- Upper terms: those in the synsets of animal#n#1, plant#n#2, vehicle#n#1
- Definition and hypernym extraction: no domain corpus, just Google define

Experiment 2: Evaluation against WordNet

- Statistics and manual evaluation:

	animals		plants		vehicles	
node compression ratio	0.48	978/2015	0.40	744/1840	0.41	144/353
edge compression ratio	0.49	977/1975	0.35	743/2138	0.35	143/413
coverage of initial terminology	0.71	484/684	0.79	438/554	0.73	85/117
precision by hand of edges not in WN (100 randomly chosen)	0.76	76/100	0.82	82/100	0.92	92/100
precision by hand of nodes not in WN (all)	0.70	218/312	0.77	158/206	0.69	9/13

Conclusions

- An algorithm to learn a lexical taxonomy truly from scratch
  - Based on the idea of exploiting the scholarly knowledge from definitions
  - A graph-based approach to learn a "clean" taxonomy
  - Can be applied to any domain of interest with little effort

References

- [Bikel et al 1997] Bikel, D.; Miller, S.; Schwartz, R.; and Weischedel, R. Nymble: a high-performance learning name-finder. In *Proceedings of ANLP'97*, p194-201.
- [Califf & Mooney 1999], Califf, M.E.; Mooney, R.: Relational Learning of Pattern-Match Rules for Information Extraction, in *Proceedings of the Sixteenth National Conference on Artificial Intelligence* (AAAI-99).
- [Cohen, Hurst, Jensen, 2002] Cohen, W.; Hurst, M.; Jensen, L.: A flexible learning system for wrapping tables and lists in HTML documents. *Proceedings of The Eleventh International World Wide Web Conference* (WWW-2002)
- [Cohen, Kautz, McAllester 2000] Cohen, W; Kautz, H.; McAllester, D.: Hardening soft information sources. *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining* (KDD-2000).
- [Cohen, 1998] Cohen, W.: Integration of Heterogeneous Databases Without Common Domains Using Queries Based on Textual Similarity, in *Proceedings of ACM SIGMOD-98*.
- [Cohen, 2000a] Cohen, W.: Data Integration using Similarity Joins and a Word-based Information Representation Language, *ACM Transactions on Information Systems*, 18(3).
- [Cohen, 2000b] Cohen, W. Automatically Extracting Features for Concept Learning from the Web, *Machine Learning: Proceedings of the Seventeeth International Conference (ML-2000).*
- [Collins & Singer 1999] Collins, M.; and Singer, Y. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.
- [De Jong 1982] De Jong, G. An Overview of the FRUMP System. In: Lehnert, W. & Ringle, M. H. (eds), *Strategies for Natural Language Processing*. Larence Erlbaum, 1982, 149-176.
- [Freitag 98] Freitag, D: Information extraction from HTML: application of a general machine learning approach, *Proceedings of the Fifteenth National Conference on Artificial Intelligence* (AAAI-98).
- [Freitag, 1999], Freitag, D. *Machine Learning for Information Extraction in Informal Domains*. Ph.D. dissertation, Carnegie Mellon University.
- [Freitag 2000], Freitag, D: Machine Learning for Information Extraction in Informal Domains, *Machine Learning* 39(2/3): 99-101 (2000).
- Freitag & Kushmerick, 1999] Freitag, D; Kushmerick, D.: Boosted Wrapper Induction. *Proceedings of the Sixteenth National Conference on Artificial Intelligence* (AAAI-99)
- [Freitag & McCallum 1999] Freitag, D. and McCallum, A. Information extraction using HMMs and shrinakge. In Proceedings AAAI-99 Workshop on Machine Learning for Information Extraction. AAAI Technical Report WS-99-11.
- [Kushmerick, 2000] Kushmerick, N: Wrapper Induction: efficiency and expressiveness, *Artificial Intelligence*, 118(pp 15-68).
- [Lafferty, McCallum & Pereira 2001] Lafferty, J.; McCallum, A.; and Pereira, F., Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *In Proceedings of ICML-2001*.
- [Leek 1997] Leek, T. R. *Information extraction using hidden Markov models*. Master's thesis. UC San Diego.
- [McCallum, Freitag & Pereira 2000] McCallum, A.; Freitag, D.; and Pereira. F., Maximum entropy Markov models for information extraction and segmentation, In *Proceedings of ICML-2000*
- [Miller et al 2000] Miller, S.; Fox, H.; Ramshaw, L.; Weischedel, R. A Novel Use of Statistical Parsing to Extract Information from Text. *Proceedings of the 1st Annual Meeting of the North American Chapter of the ACL (NAACL),* p. 226 - 233.

References

- [Muslea et al, 1999] Muslea, I.; Minton, S.; Knoblock, C. A.: *A Hierarchical Approach to Wrapper Induction*. Proceedings of Autonomous Agents-99.
- [Muslea et al, 2000] Musclea, I.; Minton, S.; and Knoblock, C. Hierarhical wrapper induction for semistructured information sources. *Journal of Autonomous Agents and Multi-Agent Systems*.
- [Nahm & Mooney, 2000] Nahm, Y.; and Mooney, R. A mutually beneficial integration of data mining and information extraction. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence*, pages 627--632, Austin, TX.
- [Punyakanok & Roth 2001] Punyakanok, V.; and Roth, D. The use of classifiers in sequential inference. *Advances in Neural Information Processing Systems 13*.
- [Ratnaparkhi 1996] Ratnaparkhi, A., A maximum entropy part-of-speech tagger, in *Proc. Empirical Methods in Natural Language Processing Conference*, p133-141.
- [Ray & Craven 2001] *Ray, S.; and Craven, Ml.* Representing Sentence Structure in Hidden Markov Models for Information Extraction. *Proceedings of the 17th International Joint Conference on Artificial Intelligence,* Seattle, WA. Morgan Kaufmann.
- [Soderland 1997]: Soderland, S.: Learning to Extract Text-Based Information from the World Wide Web. *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (KDD-97).
- [Soderland 1999] Soderland, S. Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34 (1/3):233-277.