Text mining & Information Extraction

UA.DETI
José Luis Oliveira

Outline

- Motivation & Background
  - What's text mining?
  - Fields of application
    - Biomedical text mining
    - Social network mining
- Text mining tasks
  - Named entity recognition
  - Named entity normalization
  - Relation extraction
    - Example: PPI networks
  - Event Extraction
    - Example: Measuring flu incidence rates from Twitter data
- Some examples

(Some) References

- S. Sarawagi. Information Extraction. Foundations and Trends in Databases, 1(3):261–377, 2008.
  - https://www.cse.iitb.ac.in/~sunita/papers/ieSurvey.pdf
- C.D. Manning, H Schütze, Foundations of Statistical Natural Language Processing, MIT Press, 1999.
  - http://nlp.stanford.edu/fsnlp/
- C. D. Manning, P. Raghavan and H. Schuetze, Introduction to Information Retrieval, Cambridge University Press. 2008.
  - http://nlp.stanford.edu/IR-book/)
- S. Sakurai (Ed.), Theory and Applications for Advanced Text Mining, 2012, DOI: 10.5772/3115.
  - http://www.intechopen.com/books/theory-and-applications-for-advanced-text-mining

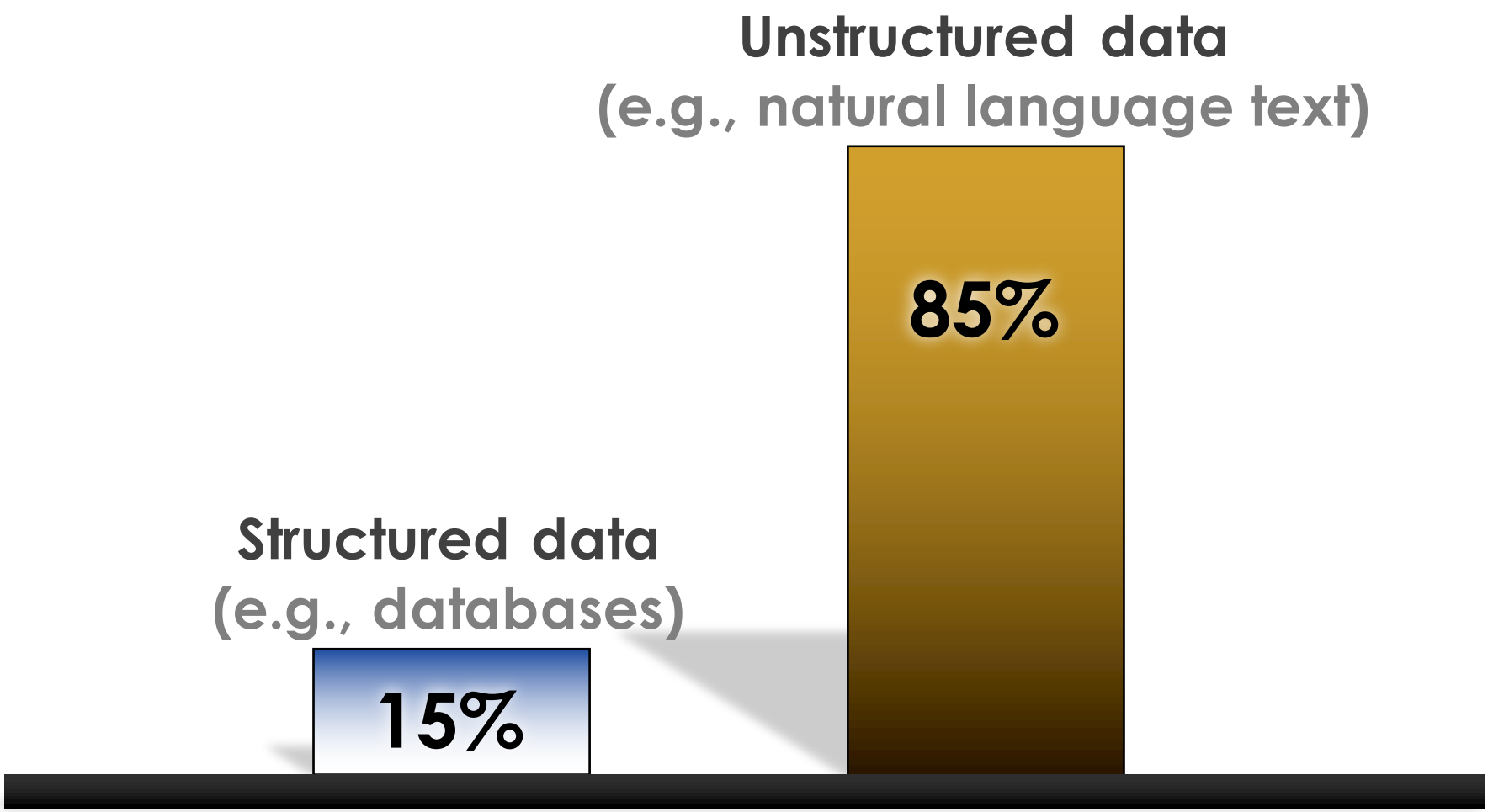
Group research

- D. Campos, S. Matos, and J. L. Oliveira, "A document processing pipeline for annotating chemical entities in scientific documents", **Journal of Cheminformatics**, 7(Suppl 1):S7, January 2015. pdf
- D. Campos, J. Lourenco, S. Matos, and J. L. Oliveira, "Egas: a collaborative and interactive document curation platform", **Database**, June 2014. pdf
- V. M. Prieto, S. Matos, M. Álvarez, F. Cacheda, and J. L. Oliveira, "Twitter: A Good Place to Detect Health Conditions", **PLoS ONE**, January 2014. pdf
- D. Campos, Q. C. Bui, S. Matos, and J. L. Oliveira, "TrigNER: automatically optimized biomedical event trigger recognition on scientific documents", **Source Code for Biology and Medicine**, vol. 9:1, January 2014. pdf
- D. Campos, S. Matos, and J. L. Oliveira, "A modular framework for biomedical concept recognition", **BMC Bioinformatics**, 14:281, 2013. pdf
- T. Nunes, D. Campos, S. Matos, and J. L. Oliveira, "BeCAS: biomedical concept recognition services and visualization", **Bioinformatics**, Vol. 29(15):1915-6, August 2013. pdf
- D. Campos, S. Matos, and J. L. Oliveira, "Gimli: open source and high-performance biomedical name recognition", **BMC Bioinformatics**, 14:54, February 2013. pdf

How are you sharing knowledge?

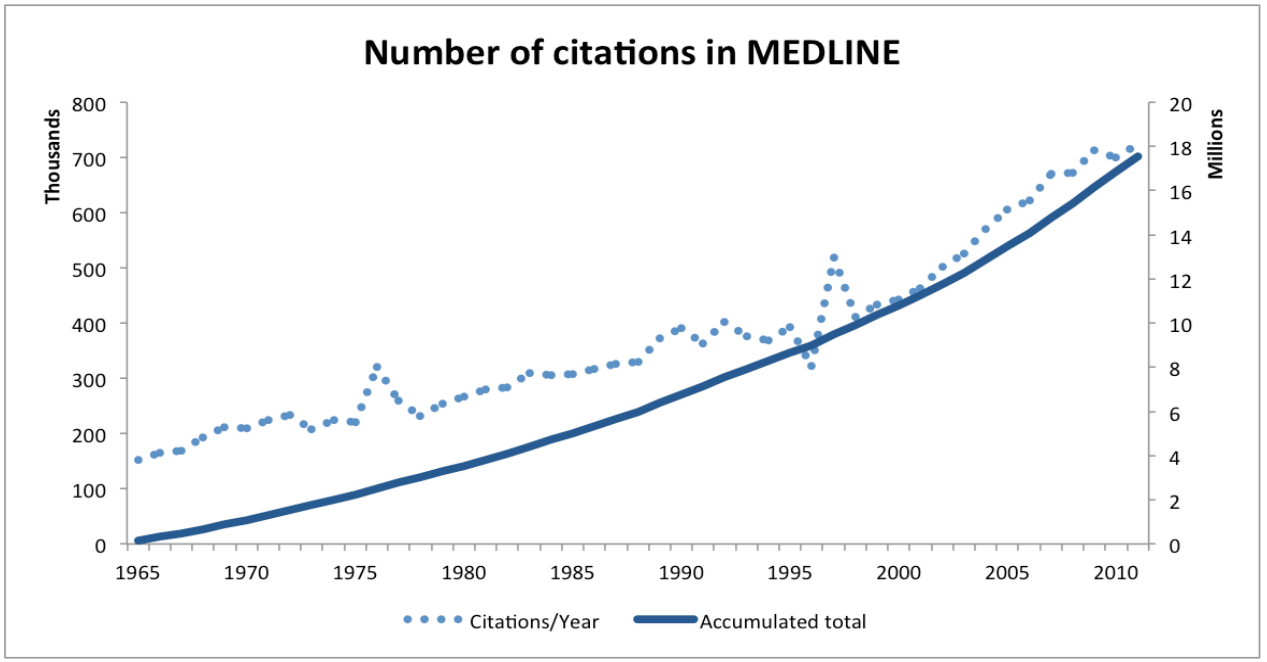
How are you sharing knowledge?

How are we capturing knowledge?



Motivation

- Huge amounts of natural language text
  - User-generated content (blogs, Twitter, comments, …)
    - valuable information, with many possible applications
  - Scientific literature
    - the most complete and up-to-date source of information



2000 new citations added every day

Motivating Examples

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

Slide from Suchanek

Motivating Examples

Biography for
Elvis Presley More at IMDbPro »

Date of Birth
8 January 1935, Tupelo, Mississippi,

Date of Death
16 August 1977, Memphis, Tennesse

Birth Name
Elvis Aron Presley

Nickna
The Pe
The Kir
The Kir

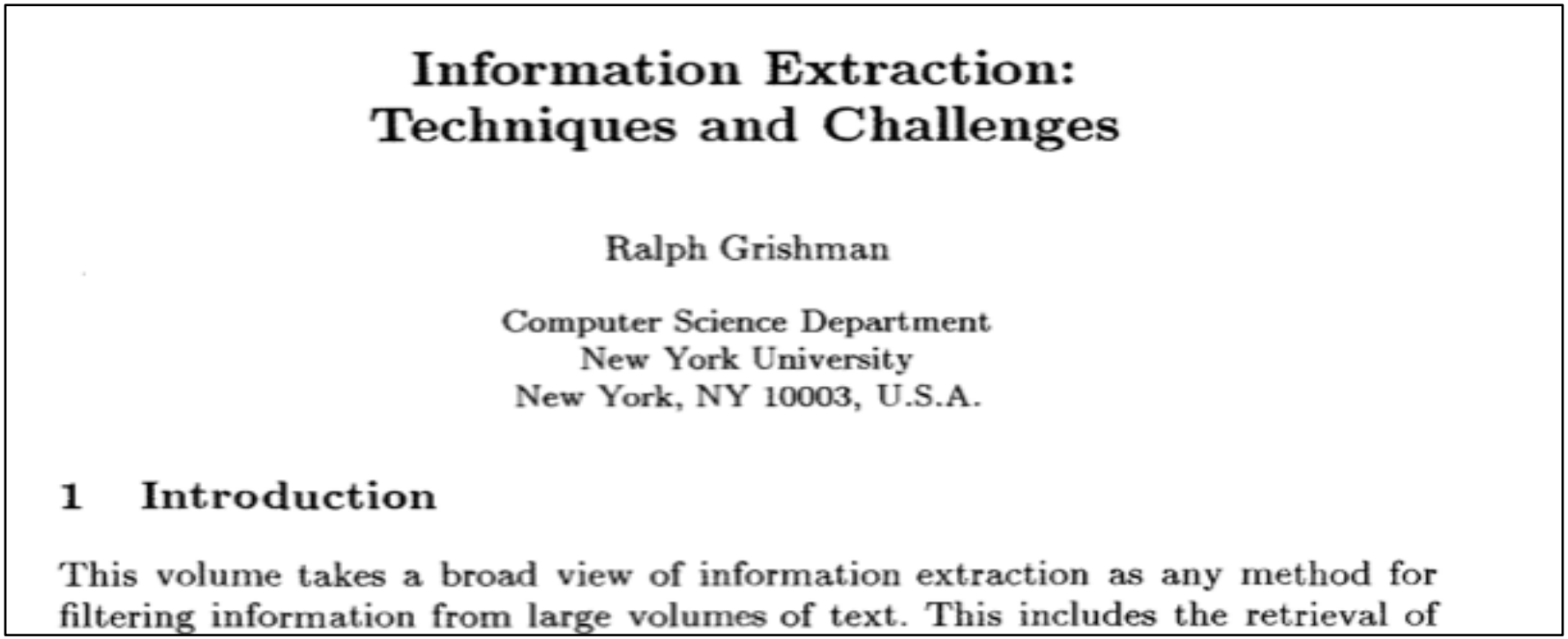
Height
6' (1.83

Mini Bi
Elvis Aa

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

Slide from Suchanek

Motivating Examples



Information Extraction:
Techniques and Challenges

Ralph Grishman

Computer Science Department
New York University
New York, NY 10003, U.S.A.

1 Introduction

This volume takes a broad view of information extraction as any method for filtering information from large volumes of text. This includes the retrieval of

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

Slide from Suchanek

What is Text Mining (TM)?

- "Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources.", M. Hearst,
  - http://people.ischool.berkeley.edu/~hearst/text-mining.html

- **Text mining** is a broader area as compared to information retrieval or information extraction.

TM, IR & IE

- Typical **text mining** tasks include document classification, document clustering, building ontology, sentiment analysis, document summarization, Information extraction etc.
- **Information retrieval** typically deals with crawling, parsing and indexing document, retrieving documents.
- **Information extraction** is the task of automatically extracting structured information from unstructured or semi-structured machine-readable documents.

What is Information Extraction (IE)?

- First note this semantic slippage:
  - Information Retrieval doesn't retrieve information
  - You have an information need, but what you get back isn't information but documents, which you hope have the information
- Information extraction is one approach to going further for a special case:
  - There's some relation you're interested in
  - Your query is for elements of that relation
  - A limited form of natural language understanding

What is Information Extraction (IE)?

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

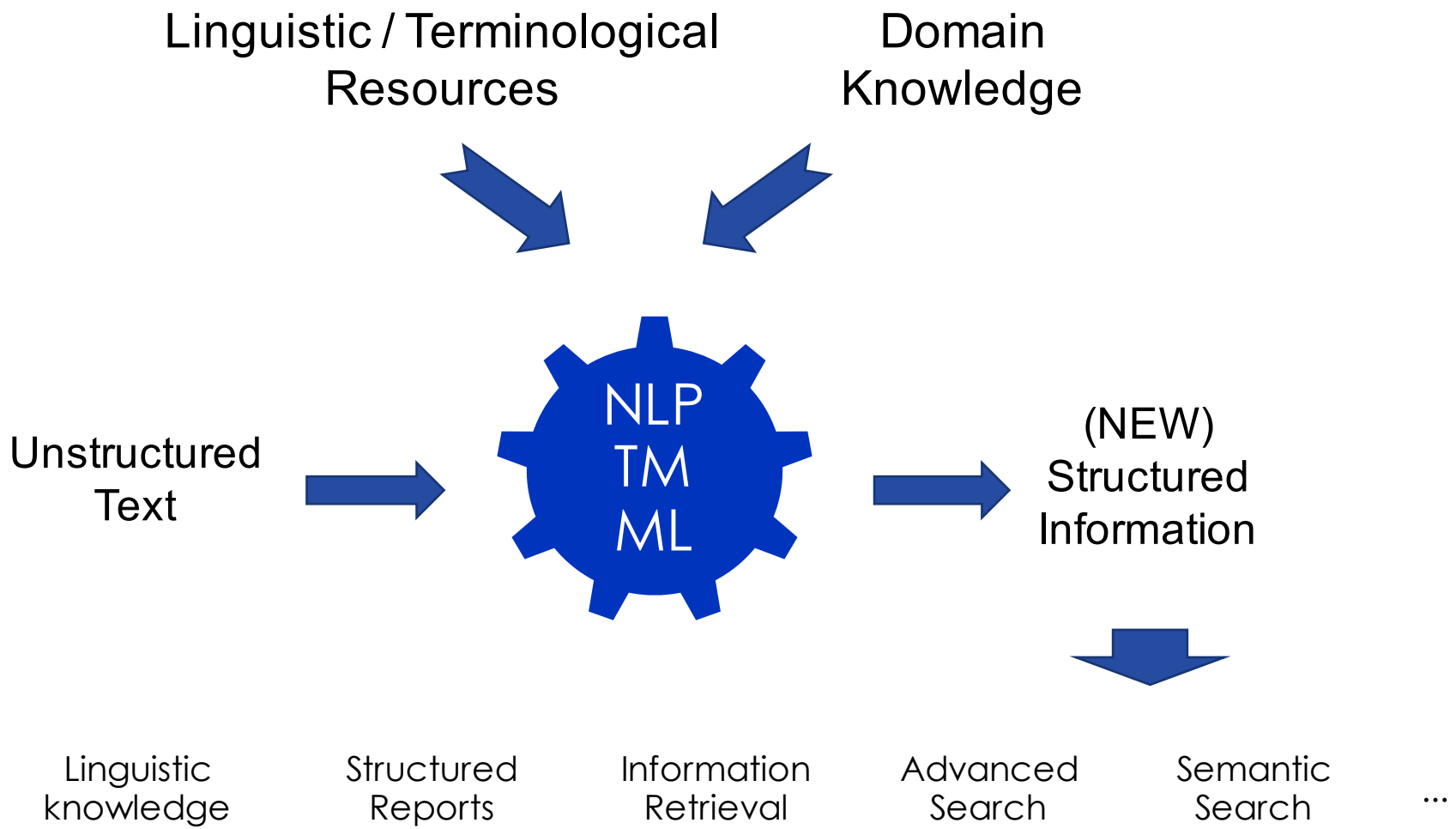
Elvis Presley was a famous rock singer.
...

Mary once remarked that the only attractive thing about the painter Elvis Hunter was his first name.

Information Extraction

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

Typical Pipeline



Application areas

- Military
- News
- Finance
- Law
- Clinical
- Scientific research
- Social research
- Social networks
- Advertisement / Recommendation systems
- Machine translation
- (…)

Sources

- News
- Enterprise reports, patents
- Court rulings, compensation claims
- Patient records, medical reports
- Drug information leaflets
- Scientific publications
- User Generated Content
  - Emails
  - Chat rooms
  - Blogs
  - Social web posts
  - User comments

Some popular services / tools

- Knowledge engines, Q&A
  - Ask.com, WolframAlpha, IBM Watson
- Recommendation systems
  - Amazon, Last.fm, Facebook
- Open source tools
  - GATE, UIMA, tm (R), RapidMiner, NEJI, ..
- Enterprise tools
  - SAS TextMiner, Xerox Fact Spotter, Digital Reasoning

- And much more going on "behind the scenes":
  - Microsoft Research, Google Research, Yahoo! Labs, IBM Natural Language Processing, Xerox Research Centre Europe

Text mining tasks

Retrieval

- Find (and order) the most relevant documents for a given search

Categorization

- Organize documents according to the major topic(s)

Entity Recognition

- Identify specific entity names in text

Disambiguation

- Associate entity mentions with univocal identifiers (e.g. Uniprot acession)

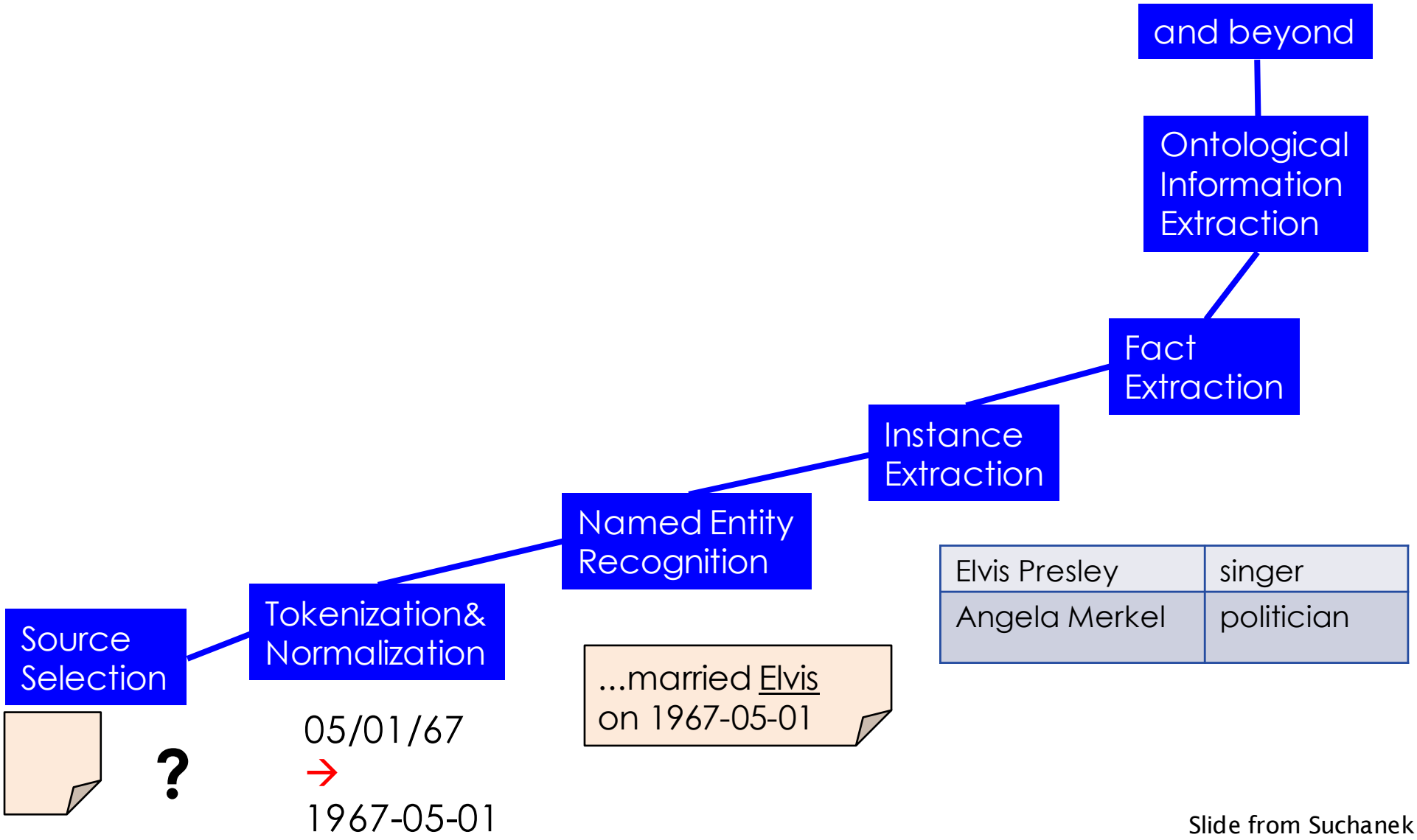
Relation / Event Extraction

- Extract relations or events related to the entities

Summarization

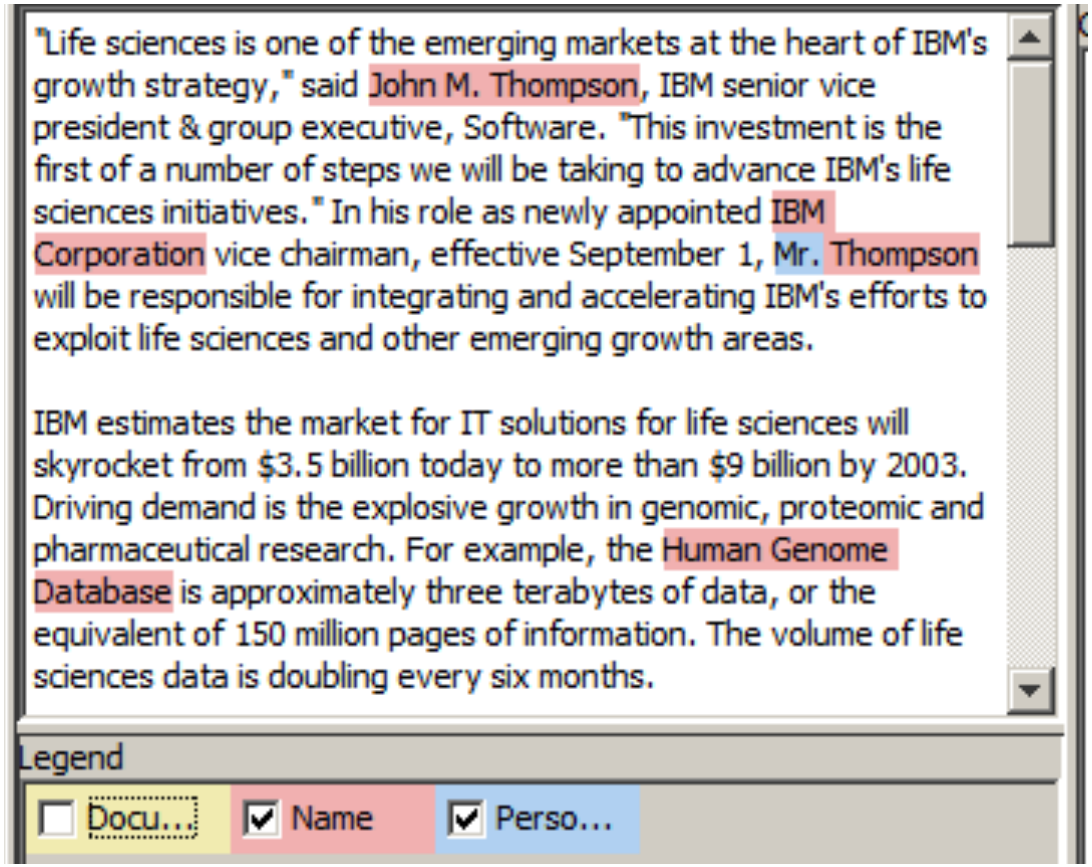
- Extract and organize the main ideas from a (set of) text(s)

Information Extraction tasks

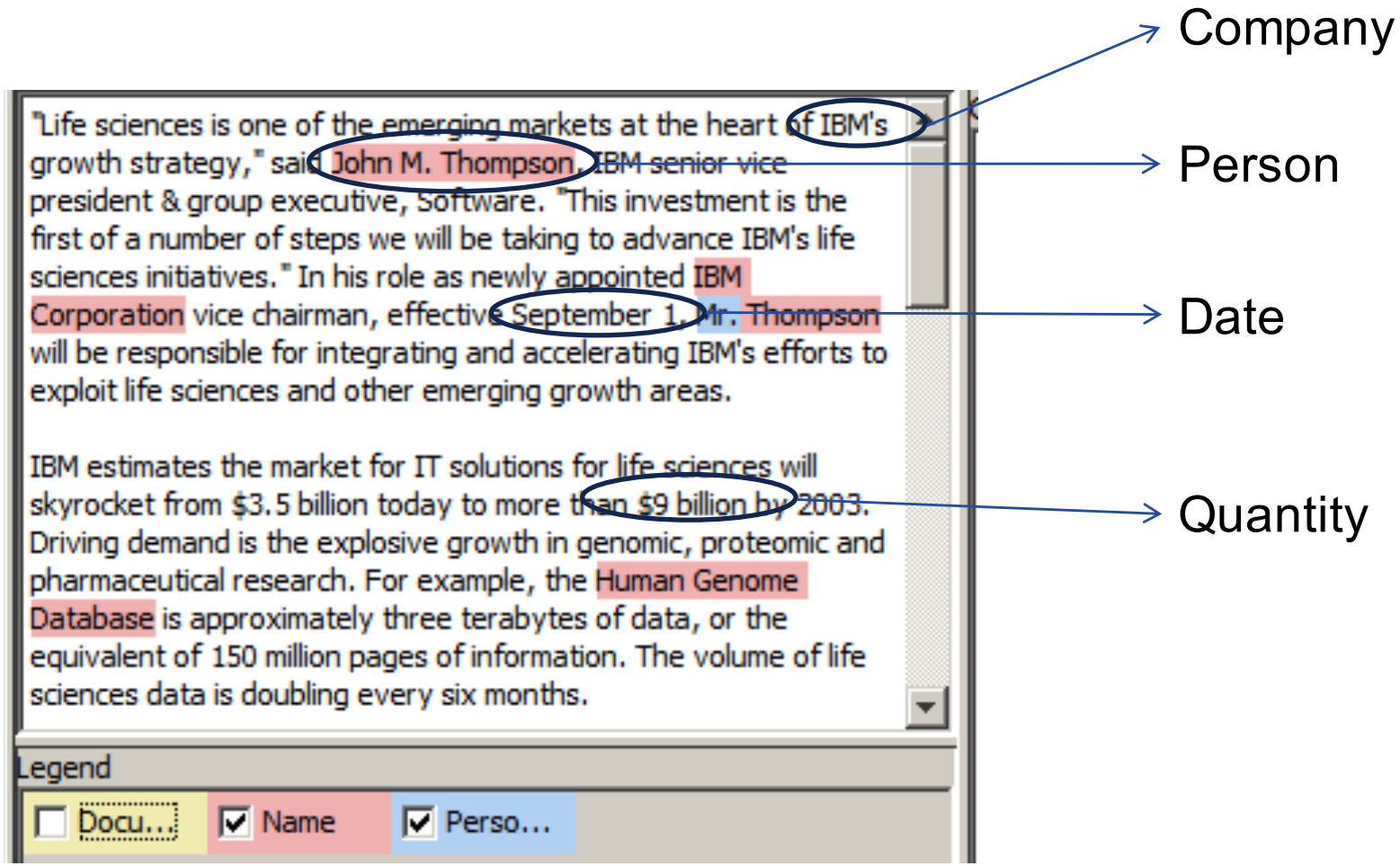


Slide from Suchanek

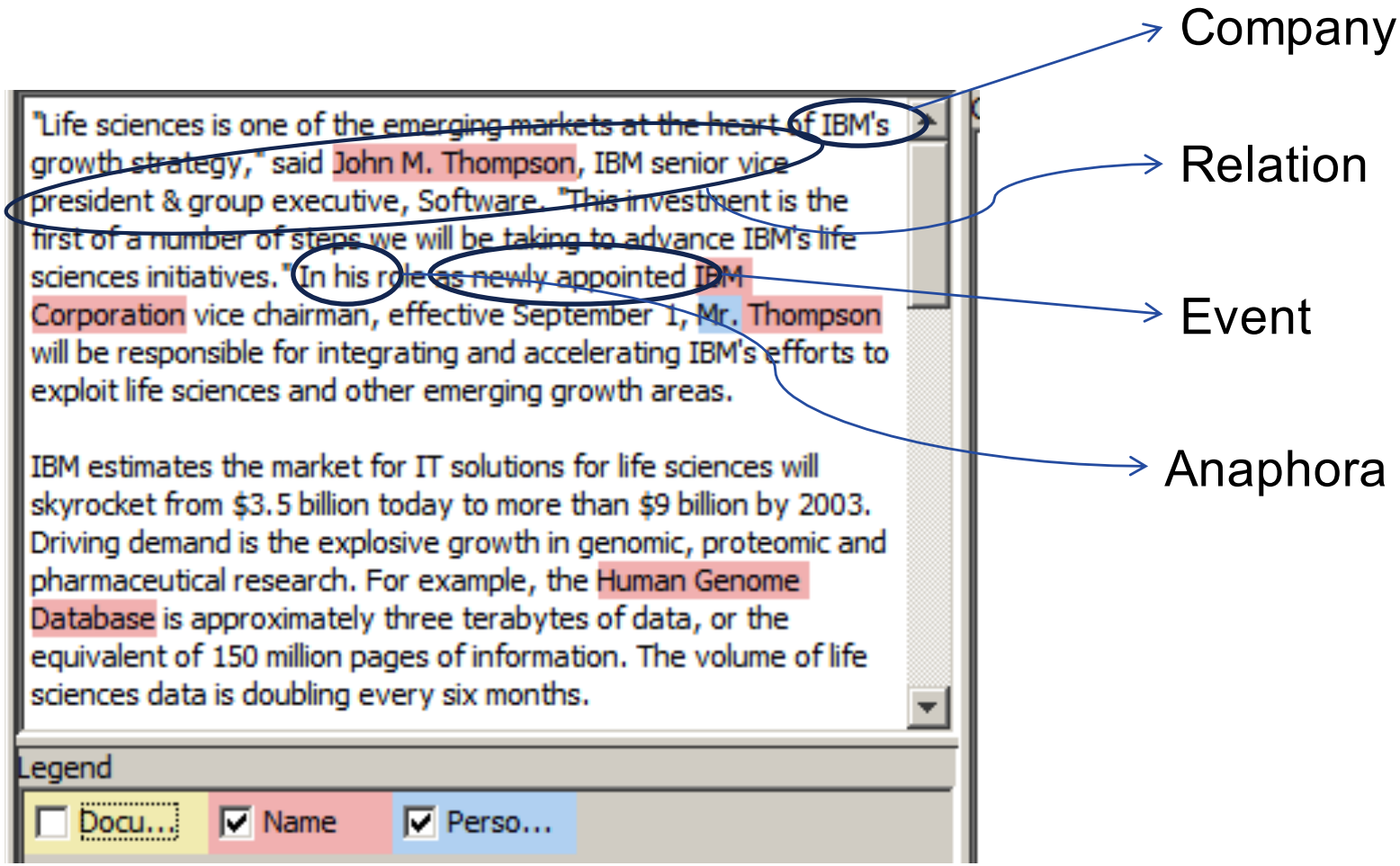
A simple example



A simple example



A simple example

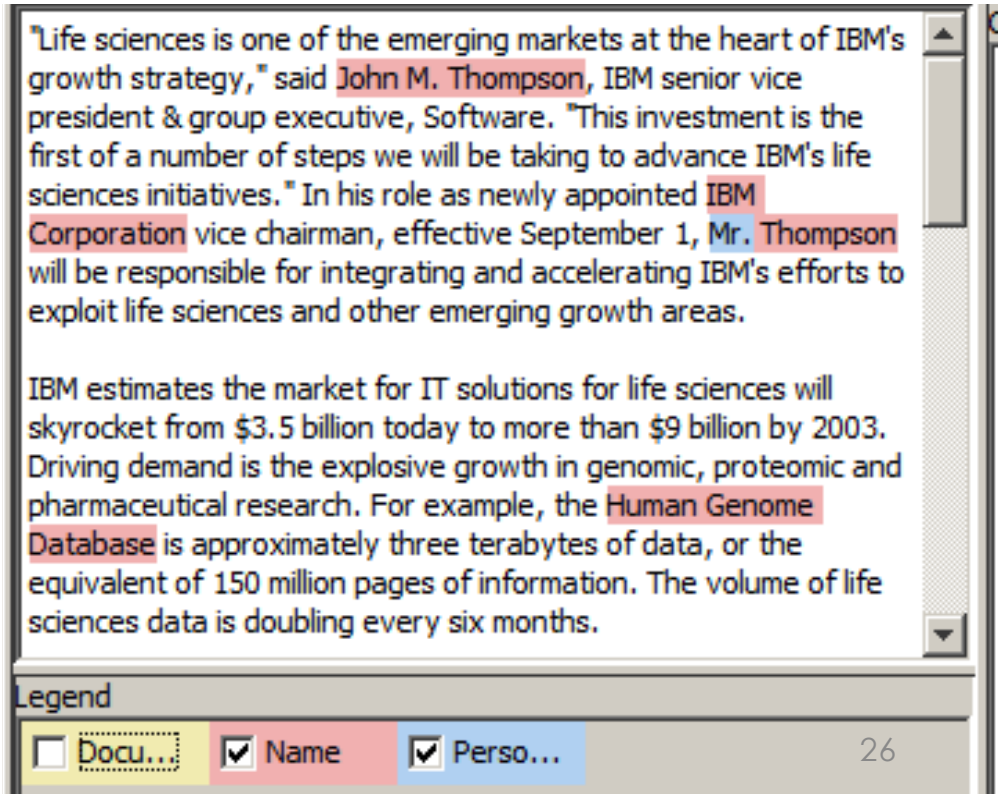


How is it done? Several approaches…

- Dictionaries
  - e.g. countries, cities, companies
- Templates / Rule-based
  - e.g. attributes, relations
  - example: "<PER> is the new CEO of <COMPANY>"
- Natural language processing (NLP)
  - e.g. syntactic parsing
- Statistics
- Machine Learning
- Hybrid

How is it done? Several steps

- tokenization
- sentence splitting
- part-of-speech (POS) tagging
- named-entity recognition
- linguistic parsing
- semantic interpretation
- discourse interpretation
- template filling
- Merging



How is it done? Example

- Sources: scientific publications, drug leaflets, patents, patient records
- Common steps:
  - Pre-processing — tokens, sentences, POS, …
  - NER — genes, proteins, drugs, diseases, …
  - Normalization — ATF2 [Entrez Gene=1386] Salbutamol [DrugBank DB01001]
  - Relations — gene/disease, drug-disease
  - Events — gene expression, regulation
  - …
  - Knowledge extraction — gene function, disease process

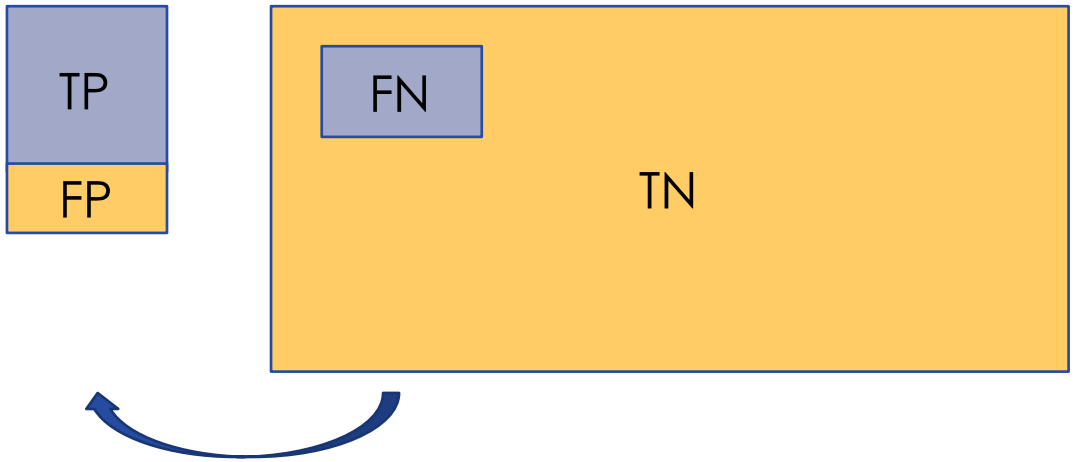
How is it done? Example

- Sources: scientific publications, drug leaflets, patents, patient records
- Common steps:
  - Pre-processing — tokens, sentences, POS, …
  - NER — genes, proteins, drugs, diseases, …
  - Normalization — ATF2 [Entrez Gene=1386]; Salbutamol [DrugBank DB01001]
  - Relations — gene/disease, drug-disease
  - Events — gene expression, regulation
  - …
  - Knowledge extraction — gene function, disease process

+ semantics
+ domain knowledge

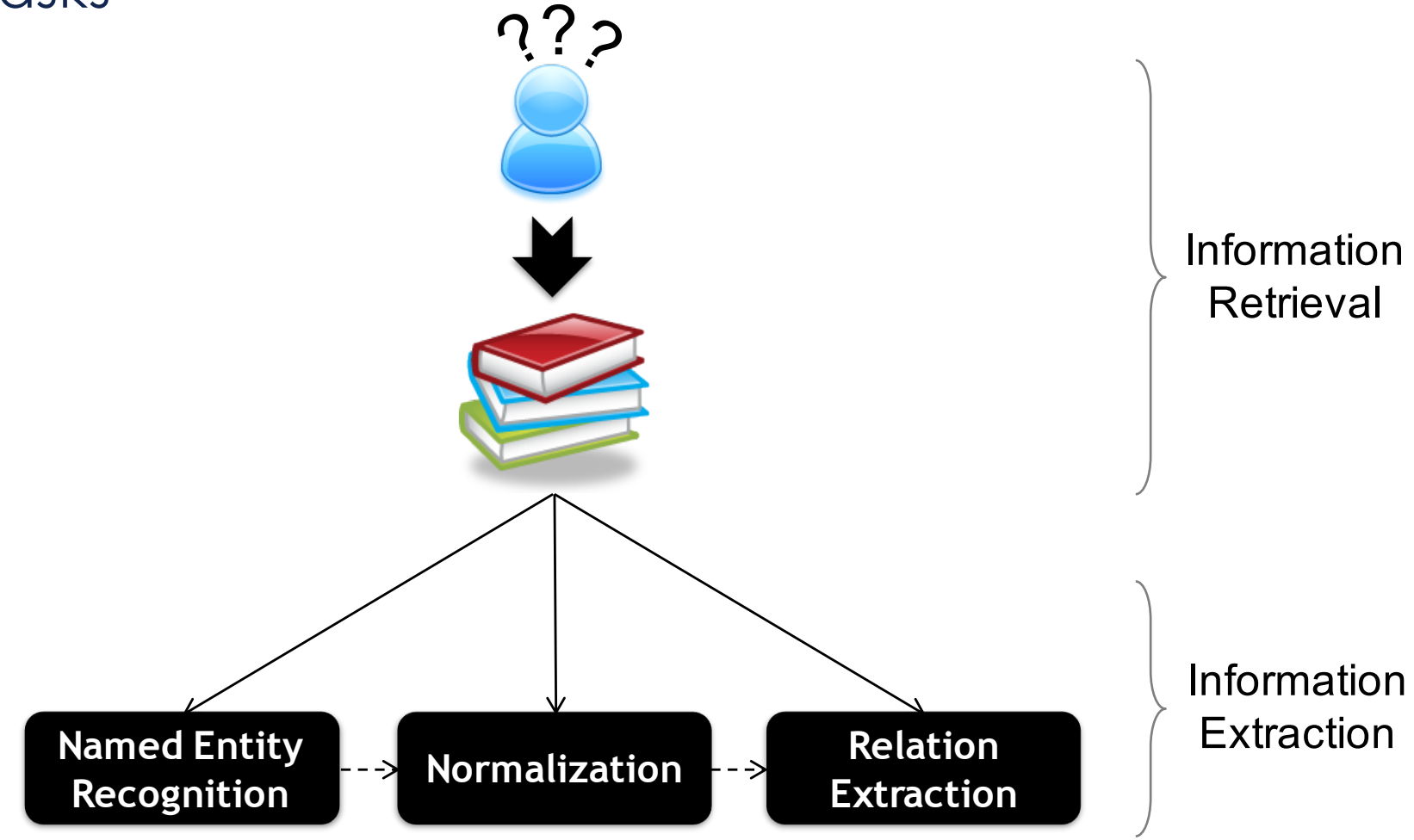
How is it done? Evaluation

- Given a set of manually annotated documents "Gold standard" (GS)
- Compare predicted results with GS
- Common measures
  - Precision: fraction of predictions which are correct
    - P = positive & predicted / total predicted
  - Recall: fraction of true results which were predicted
    - R = positive & predicted / total positive
  - F-measure
    - F = 2*P*R / (P+R)



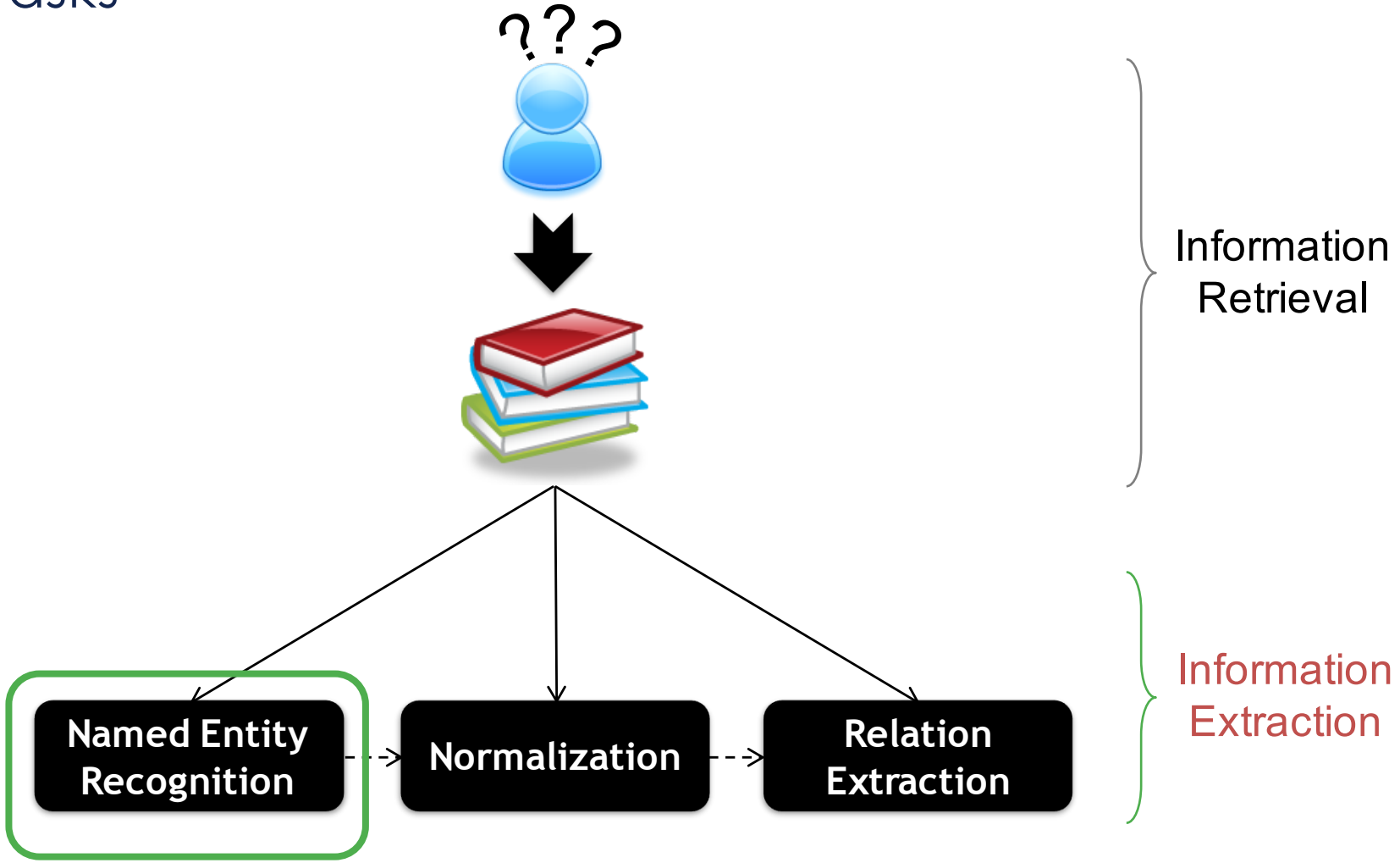
Text Mining

- Tasks



Text Mining

- Tasks



Named Entity Recognition

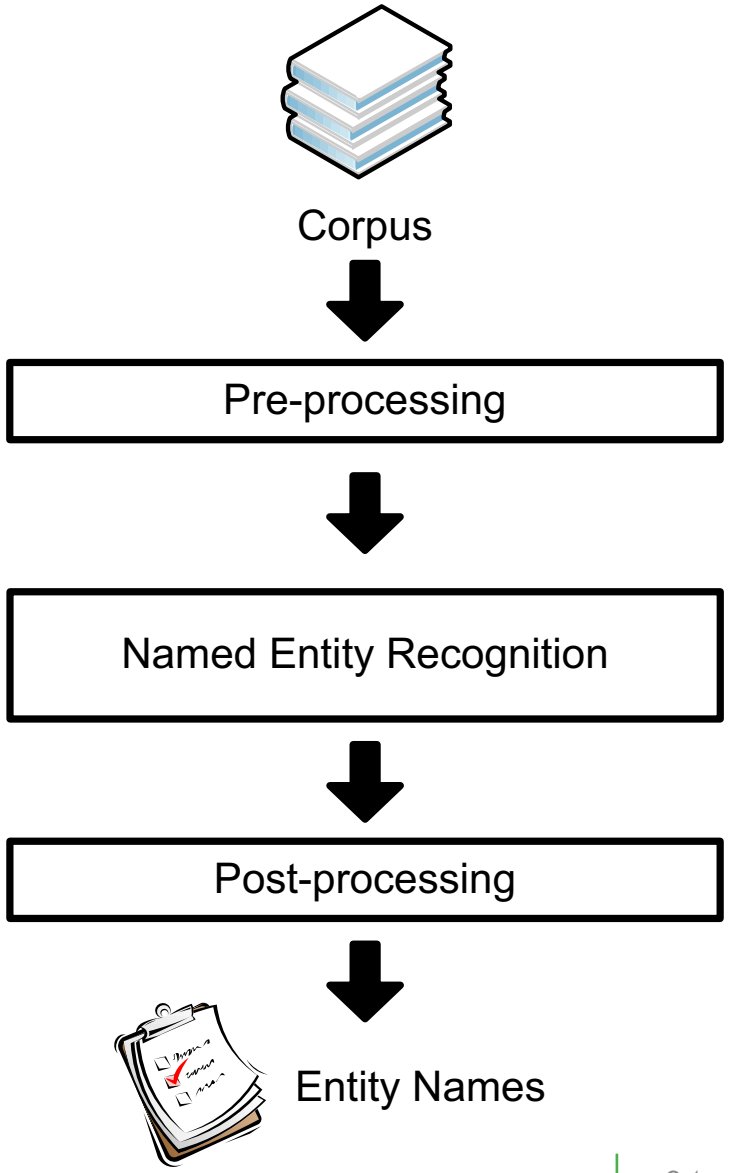
- Named Entity Recognition (NER) labels sequences of words in a text which are the names of things, such as person and company names, or gene and protein names http://nlp.stanford.edu/software/CRF-NER.shtml
- NER aims to "locate and classify **atomic elements** in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc."http://en.wikipedia.org/wiki/Named-entity_recognition
- Example:
  - **John J. Smith** lives in **Seattle**

Named Entity Recognition

- Named Entity Recognition (NER) labels sequences of words in a text which are the names of things, such as person and company names, or gene and protein names http://nlp.stanford.edu/software/CRF-NER.shtml
- NER aims to "locate and classify **atomic elements** in text into **predefined categories** such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc."http://en.wikipedia.org/wiki/Named-entity_recognition
- Example:
  - **John J. Smith** lives in **Seattle**
  - **<PER>**John J. Smith**</PER>** lives in **<LOC>**Seattle**</LOC>**

Named Entity Recognition

- Pre-processing
  - Simplify the recognition process:
    - Tokenization
    - Stopword removal
    - Word Normalization
- Post-processing
  - Refine the recognized names:
    - Abbreviation resolution
    - Error correction
- Approaches
  - Dictionary matching
  - Rule-based matching
  - Machine learning

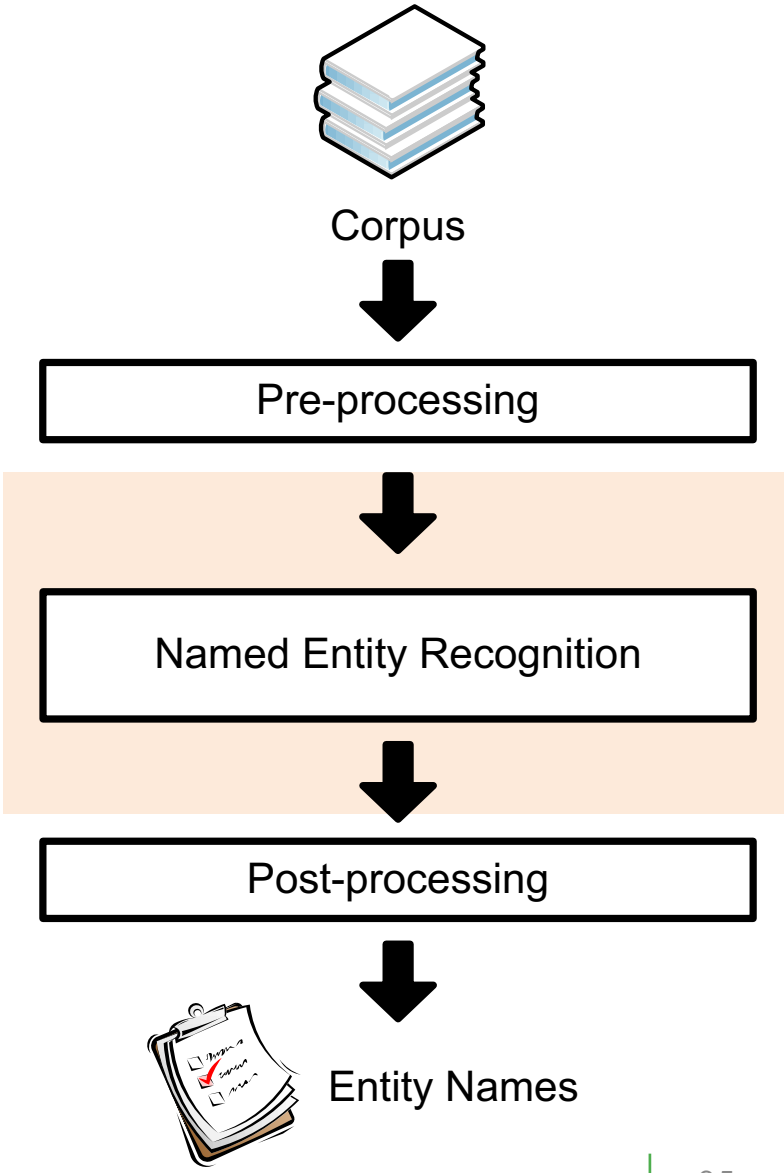


NER: Dictionary matching

Approaches

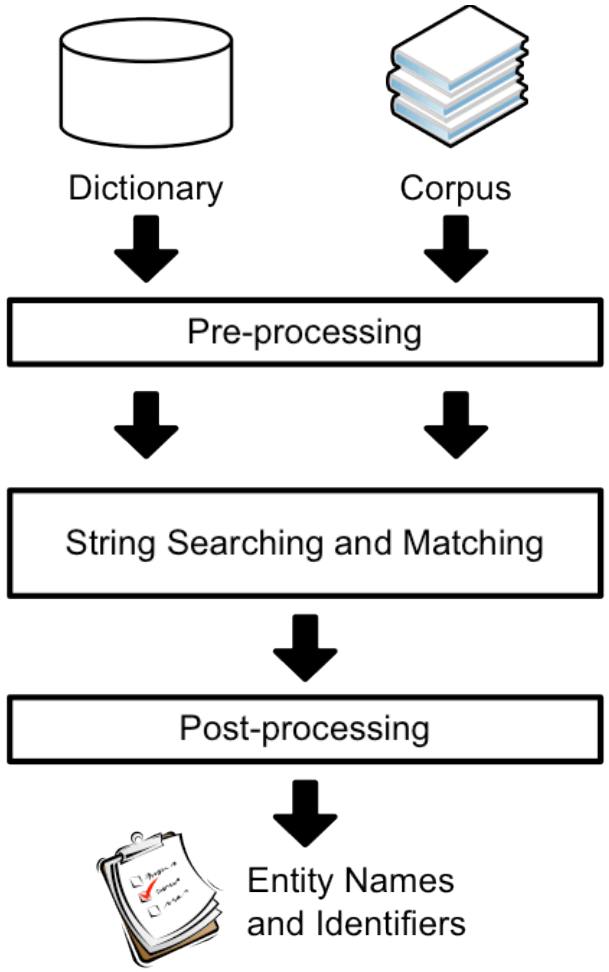
- **Dictionary matching**
- Rule-based matching
- Machine learning

Match between entity names in a **resource** and **text**



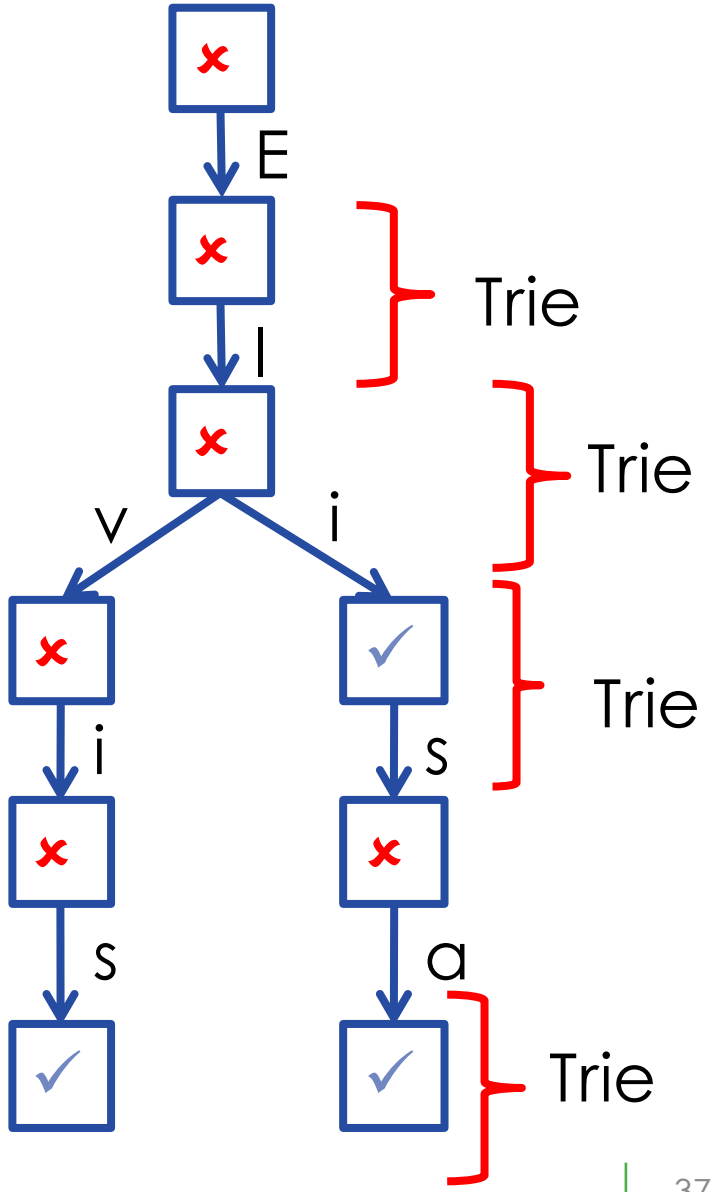
Dictionary matching

- Dictionary
  - Collect the maximum number of entity names
- String searching
  - Use a structure for the text to streamline the searching process
- String matching
  - Exact
  - Approximate



String search – Example with Tries

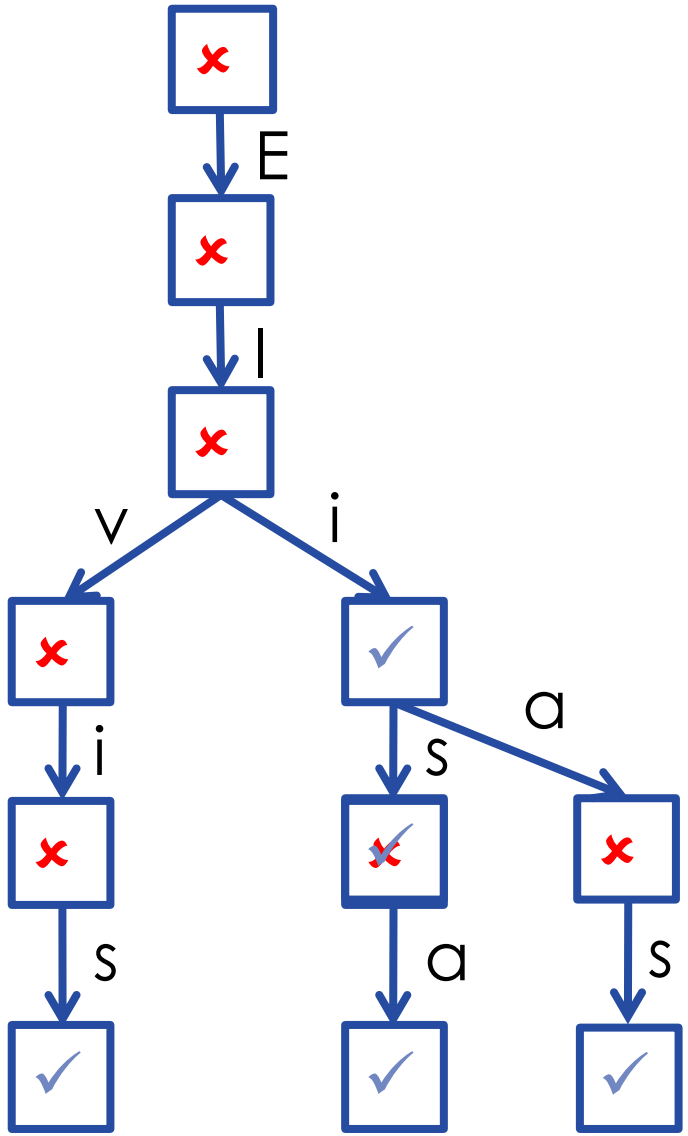
- A **trie** is pair of a boolean truth value, and a function from characters to tries.
- Example: A trie containing "Elvis", "Elisa" and "Eli"
- A trie contains a string, if the string denotes a path from the root to a node marked with TRUE (✓)



Slide from Suchanek

Adding Values to Tries

- Example: Adding "Elis"
  - Switch the sub-trie to TRUE (✓)
- Example: Adding "Elias"
  - Add the corresponding sub-trie



Slide from Suchanek

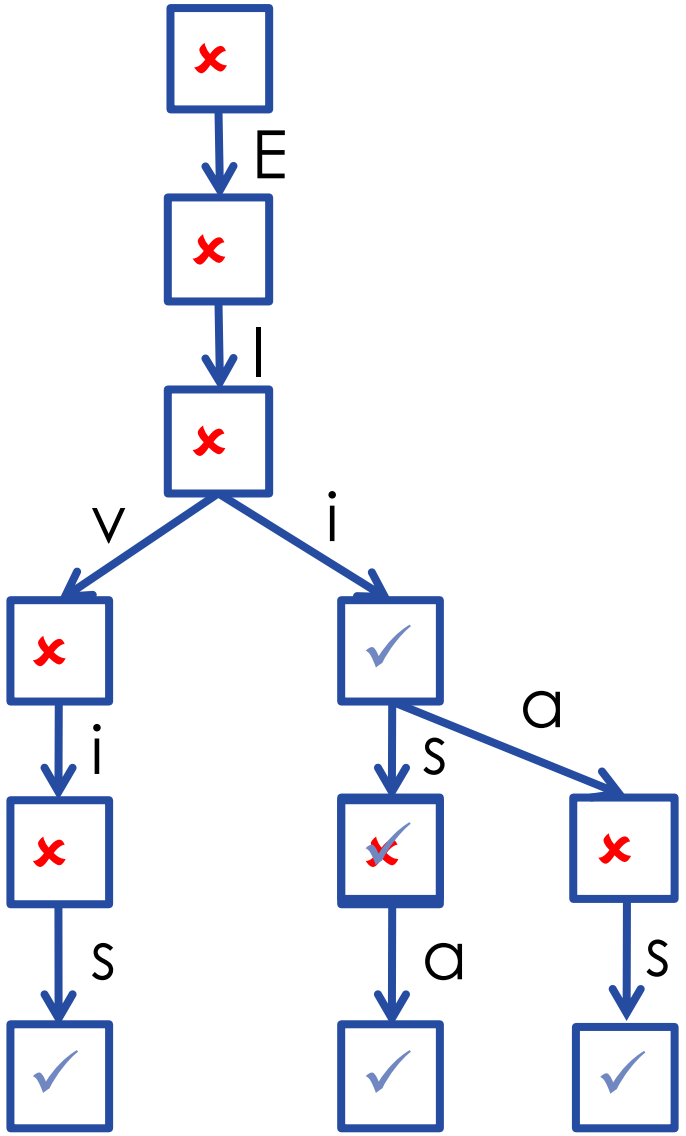
Parsing with Tries

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

Elvis is as powerful as El Nino.

=> found Elvis

Time: O(textLength * longestEntity)



Dictionary matching

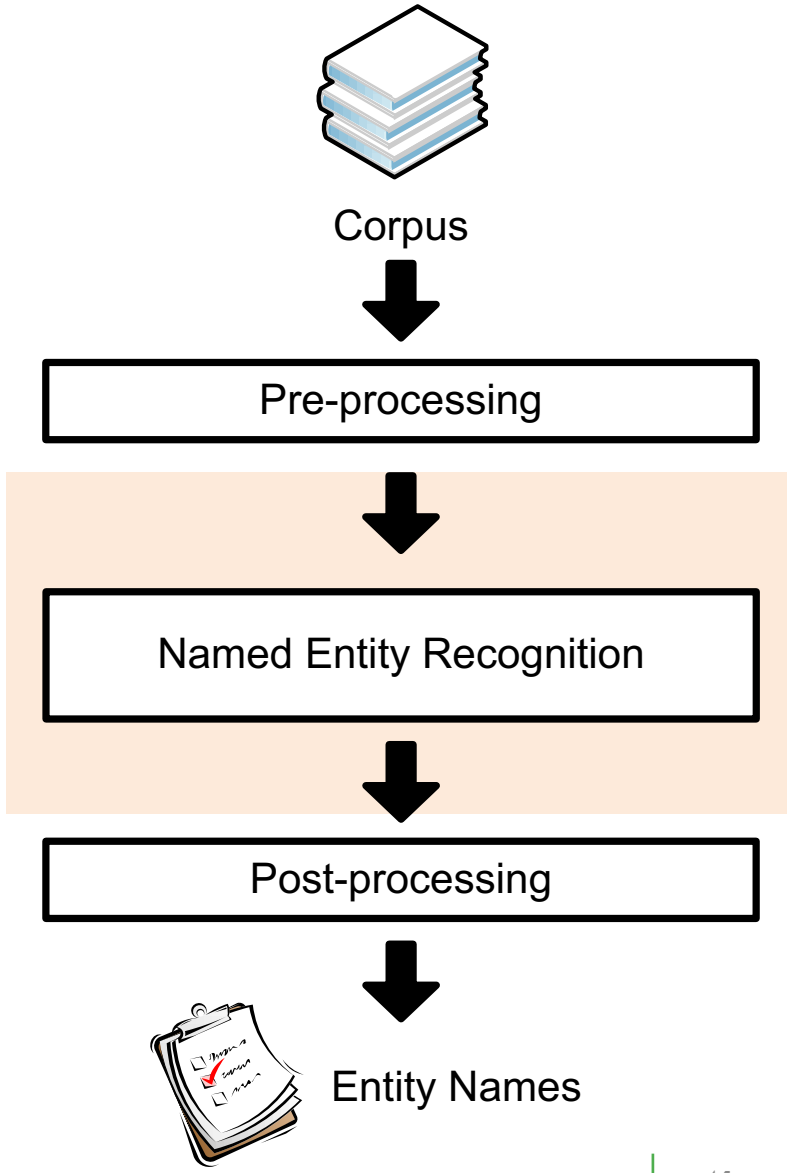
- Common applications
  - Places, companies, names, animal species
- Advantages / disadvantages:
  - Easy to match with text
  - Can be very specific (few FPs), e.g. companies
  - Usually low recall, e.g. people names
  - Dictionaries can become very large
  - Difficult to keep up-to-date
- Example (biomedical):
  - LINNAEUS [Gerner et al., 2010], http://linnaeus.sourceforge.net/

NER: Rule-based matching

Approaches

- Dictionary matching
- **Rule-based matching**
- Machine learning

Find patterns or **rules** in **text**



NER: Patterns examples

- 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)

Regular Expressions (regex)

- A **regular expression** over a set of symbols Σ is:
  - the empty string
  - or the string consisting of an element of Σ
    - (a single character)
  - or the string AB where A and B are regular expressions
    - (**concatenation**)
  - or a string of the form (A | B),
    - where A and B are regular expressions (**alternation**)
  - or a string of the form (A)*,
    - where A is a regular expression (**Kleene star**)
- For example, with Σ={a,b}, the following strings are regular expressions:

a b ab aba (a | b)

Things that are easy to express

- A | B — Either A or B
- A* — Zero+ occurrences of A
- A+ — One+ occurrences of A
- A{x,y} — x to y occurrences of A
- A? — an optional A
- [a-z] — One of the characters in the range
- . — An arbitrary symbol

Slide from Suchanek

Regex examples

- Email

  /^([a-z0-9_\.-]+)@([\da-z\.-]+)\.([a-z\.]{2,6})$/

- URL

  /^(https?:\/\/)?([\da-z\.-]+)\.([a-z\.]{2,6})([\/\w \.-]*)*\/?$/

- IP address

  /^(?:(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.){3}(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)$/

- HTML tag

  /^<([a-z]+)([^<]+)*(?:>(.*)<\/\1>|\s+\/>)$/

NER: Rule-based

- Common applications:
  - numbers, measures, time, addresses
- Encoded as regular expressions
  - [0-9]{4}-[0-9]{3}
  - (R.|Av.|Pr.) (de|da|dos)? [A-Z][a-z]+, [0-9]+
  - [0-9]+(\.[0-9]{2})? (€|£|$)
- Advantages / disadvantages:
  - Can be very specific (few FPs), e.g. urls, emails
  - May suffer from low recall, e.g. postal addresses
  - Time-consuming, difficult to maintain

NER: Hybrid solutions

- Using word lists with rules

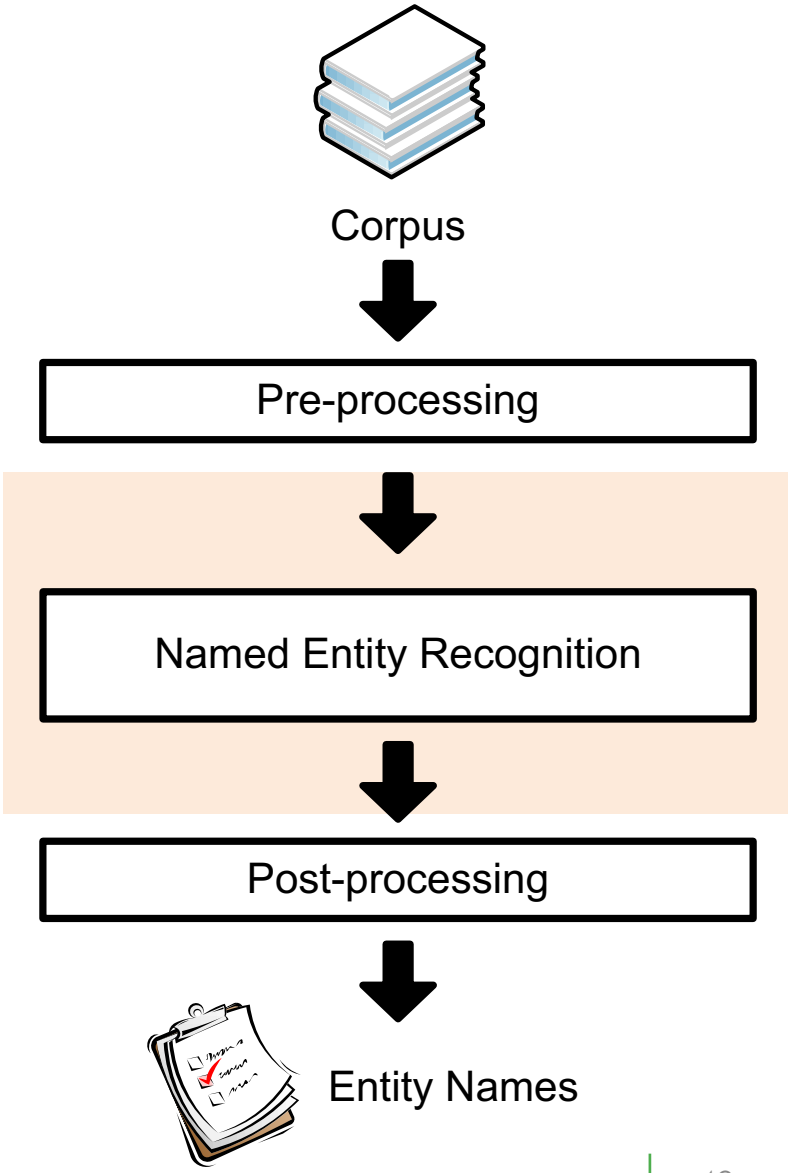
- Person names
  - Title + pattern
    - Mr. | Mrs. | Sir | … [A-Z][a-z]+ [A-Z][a-z]+
    - President | Prime-Minister | … [A-Z][a-z]+ [A-Z][a-z]+
  - Suffix + pattern
    - [A-Z][a-z]+ [A-Z][a-z]+ Jr. | Sr.
  - List of common name + pattern
    - Carl | John | Steve | … [A-Z][a-z]+
    - [A-Z][a-z]+ Silva | Smith | …
- Companies
  - ([A-Z][a-z]+) Corp. | Inc. | Lda.

NER: Machine learning

Approaches

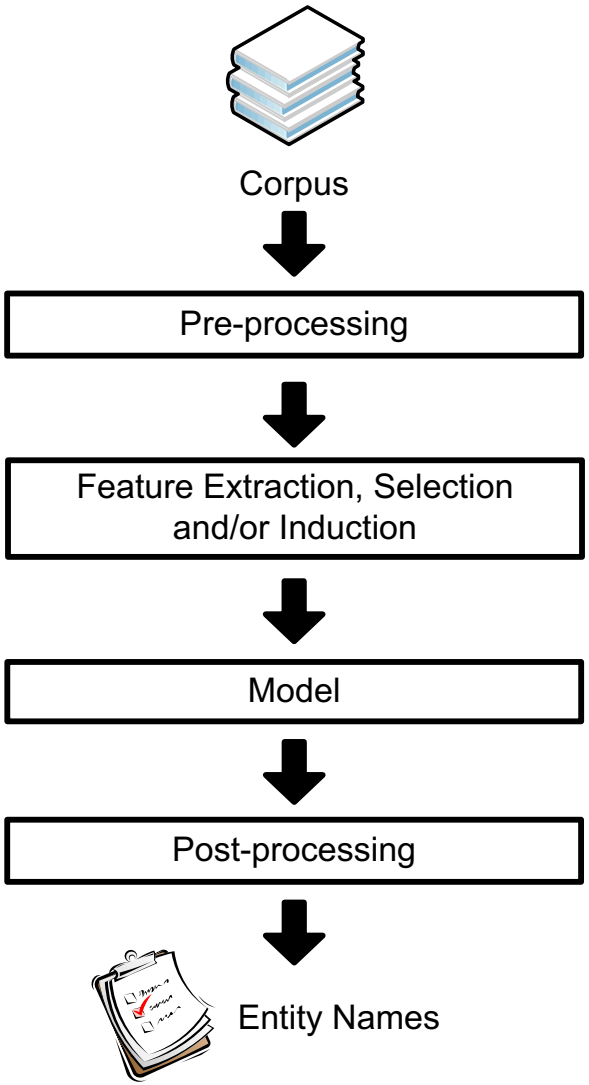
- Dictionary matching
- Rule-based matching
- **Machine learning**

Learn what is an entity name from (annotated) **examples**



Machine Learning

- Features
  - Extraction
    - Lowercase, all capitals, ...
  - Selection
    - Filter by higher contribution
  - Induction
    - Automatically extract and select
- Model
  - Supervised Learning
    - Use labelled data
  - Semi-Supervised Learning
    - Use labelled and unlabelled data



NER: Machine Learning

- John J. Smith lives in Seattle.
- (John) (J) (.) (Smith) (lives) (in) (Seattle) (.)
- NER as sequence labeling (or segmentation)
- / John J. Smith / lives in / Seattle / .

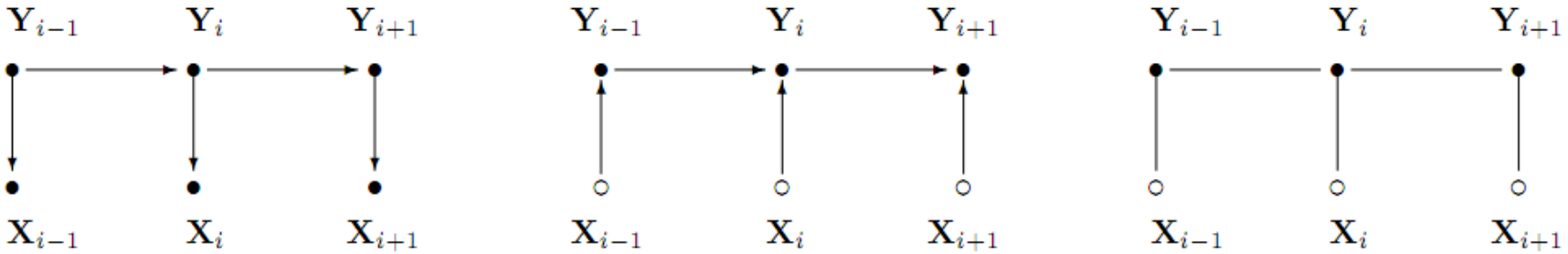
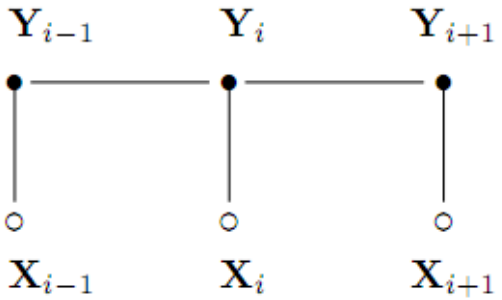


Image from: Lafferty et al., 2001

NER: Machine Learning

- NER as sequence labeling (or segmentation)



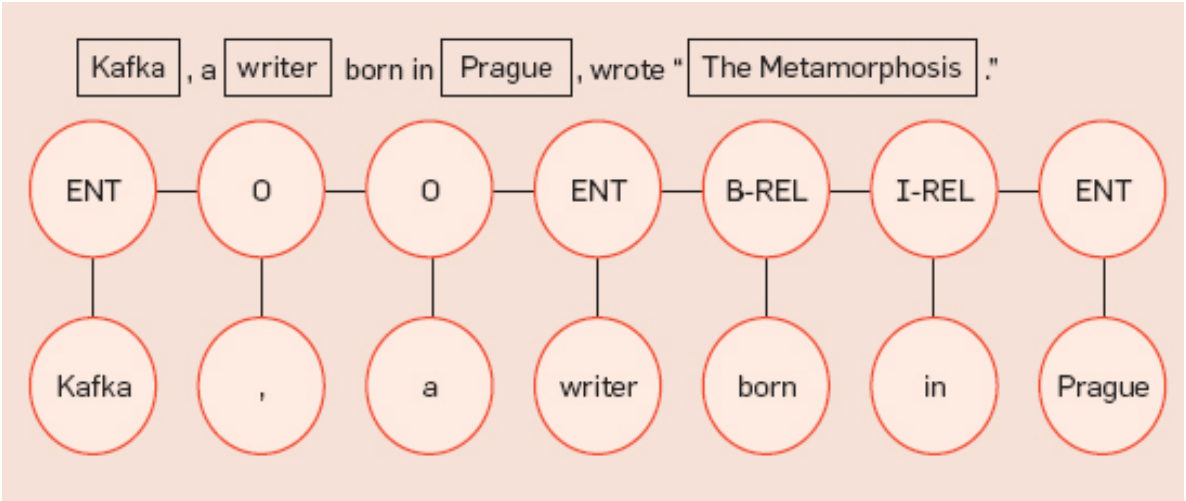
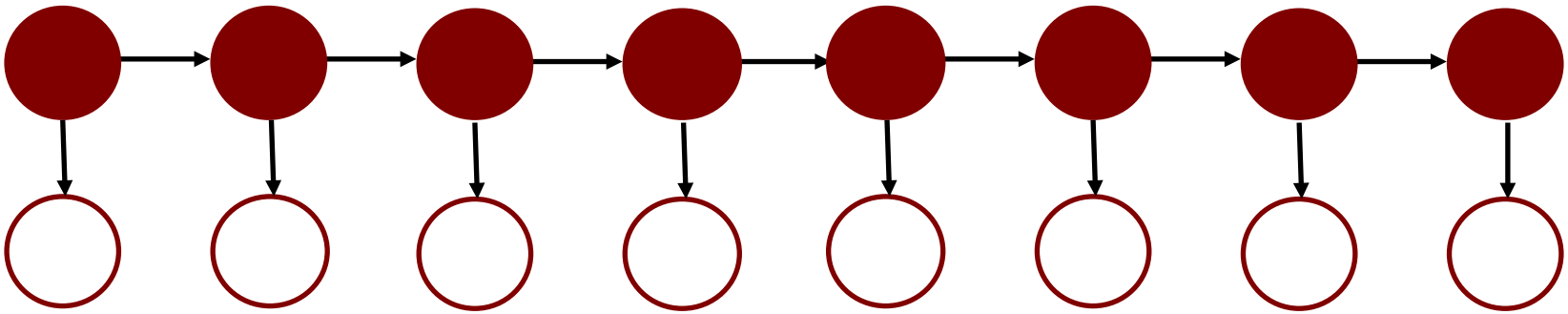


Image from: Etzioni et al., 2008

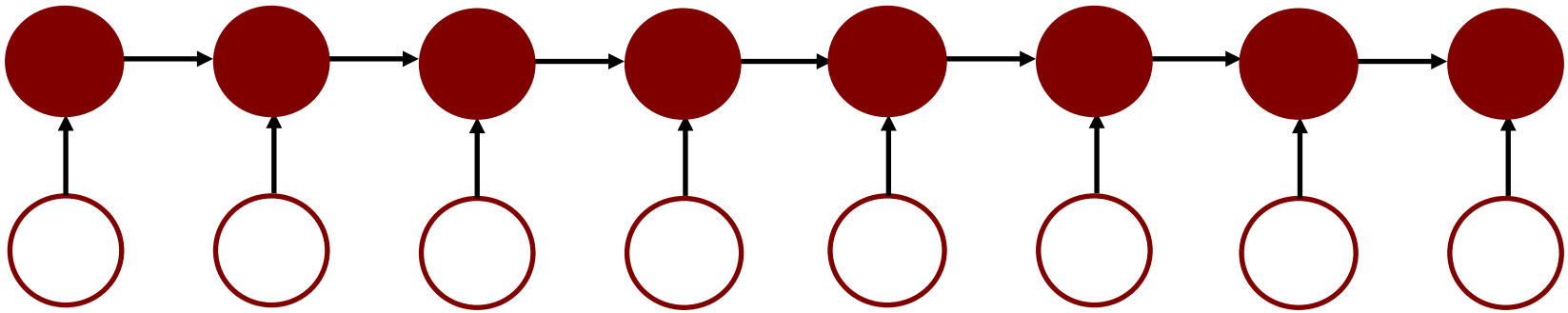
Hidden Markov Models (HMMs)

- Generative
  - Find parameters to maximize P(X,Y)
- Assumes features are independent
- When labeling Xi future observations are taken into account (forward-backward)



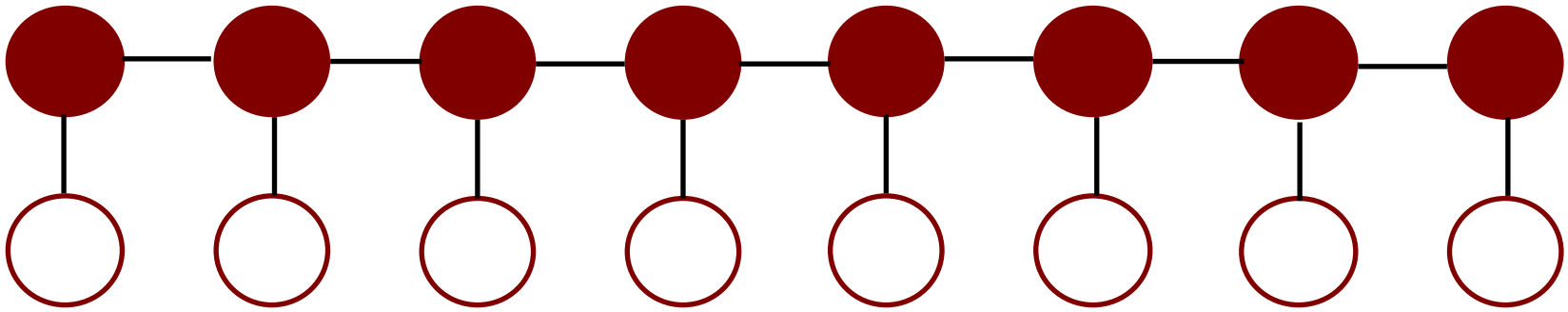
MaxEnt Markov Models (MEMMs)

- Discriminative
  - Find parameters to maximize P(Y | X)
- No longer assume that features are independent
- Do not take future observations into account (no forward-backward)



Conditional Random Fields (CRFs)

- Discriminative
- Doesn't assume that features are independent
- When labeling Yi future observations are taken into account
- ➔ The best of both worlds!



Model Trade-offs

	Speed	Discrim vs. Generative	Normalization
HMM	very fast	generative	local
MEMM	mid-range	discriminative	local
CRF	kinda slow	discriminative	global

NER: Machine Learning

- General architecture
  - Building the model

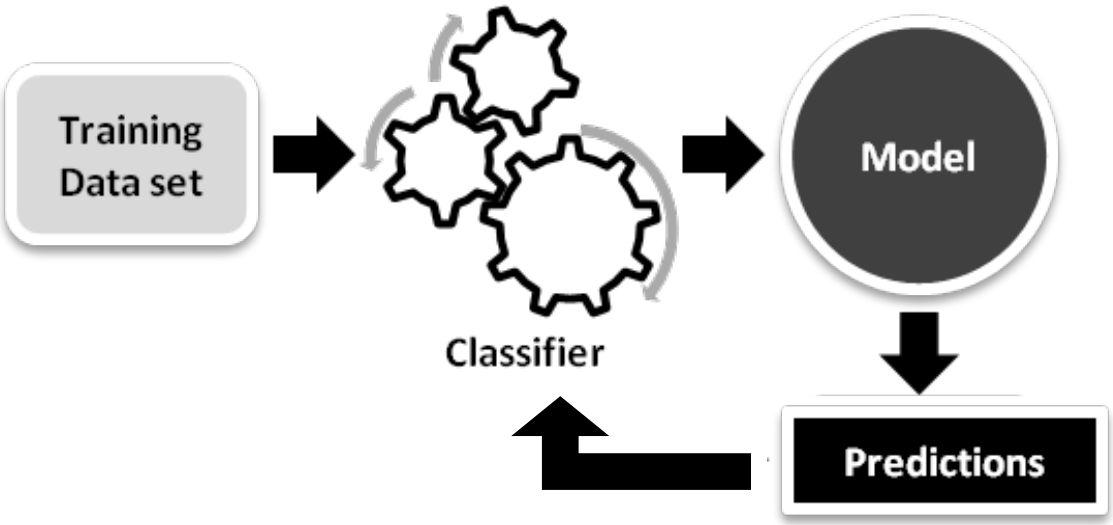


Image from David Campos

NER: Machine Learning

- General architecture
  - Applying the model

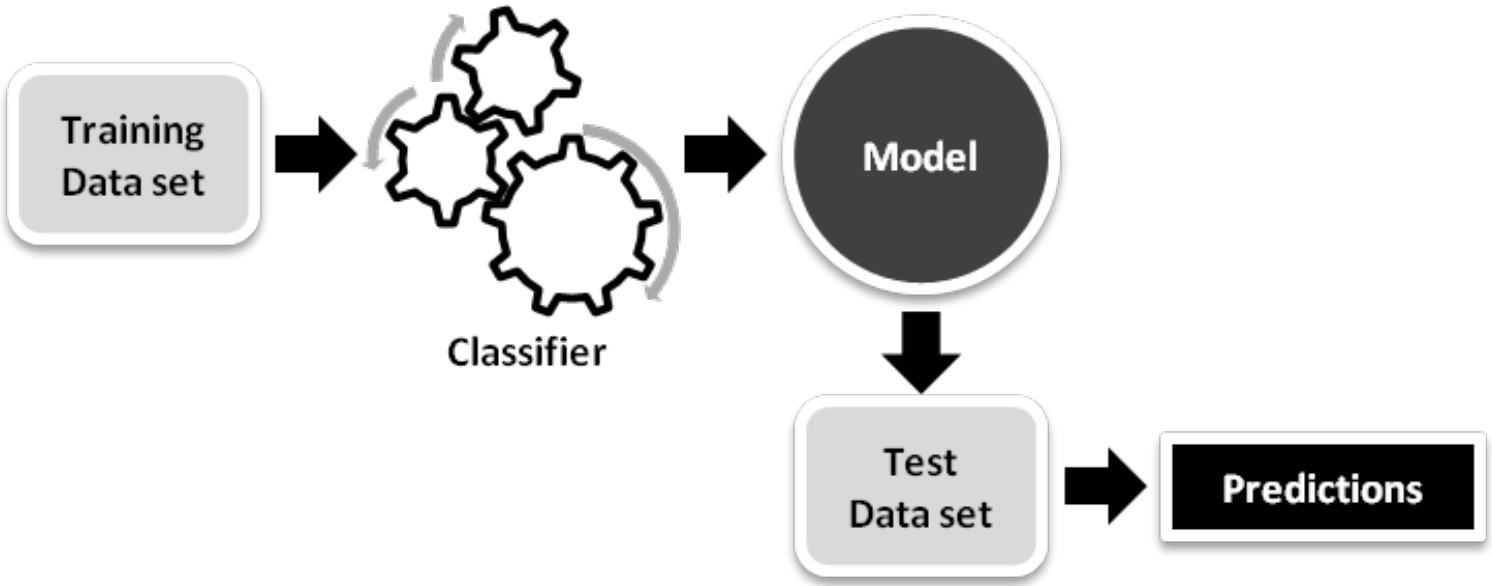


Image from David Campos

NER: Machine Learning

- General architecture
  - But how?

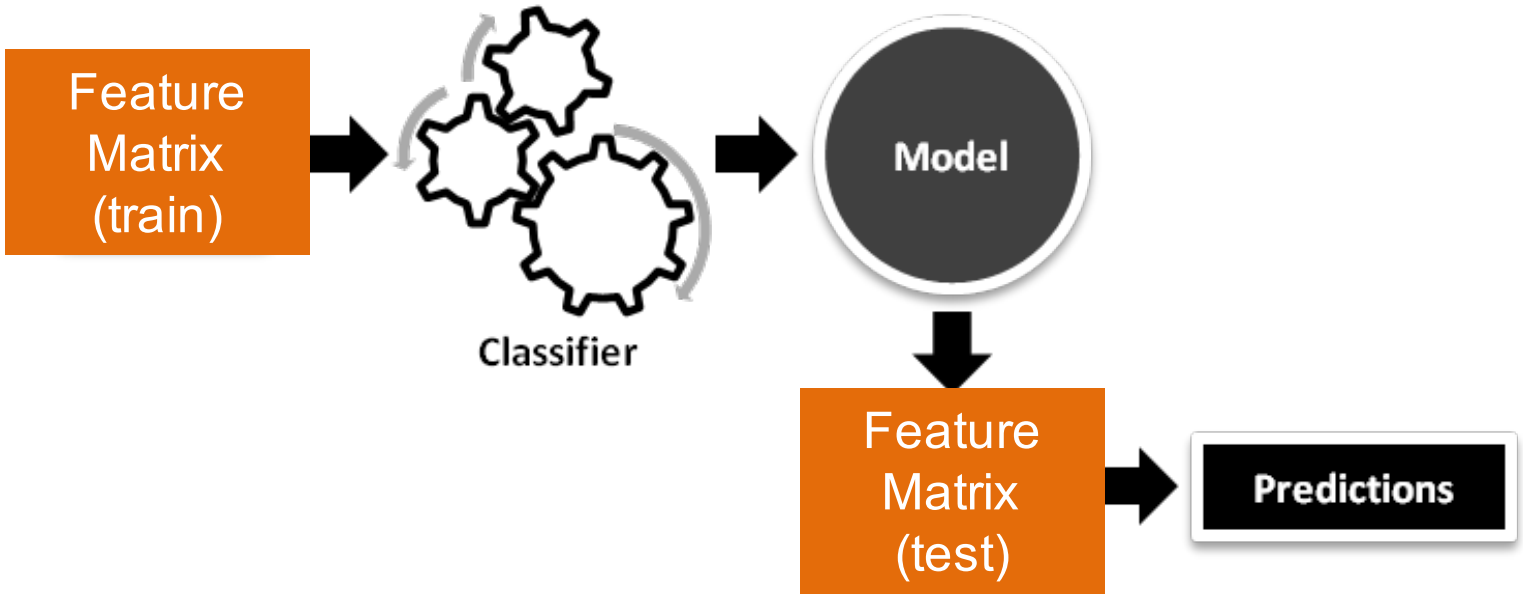


Image from David Campos

NER: Machine Learning features

- Ortographic
  - Capitalization, structure
  - E.g. initCap?, allCaps?, hasDigits?, hasSymbols?
- Morphological
  - Prefixes, sufixes, n-grams
  - e.g. ‘sea’, ‘eat’, ‘att’, ‘ttl’, ‘tle’
- Linguistic
  - POS, lemma, linguistic parsing
  - e.g. ‘lives’: V, lemma=‘live’
- Domain terms
  - e.g. biomedical: nucleotides, chemical compounds, ...
- (...)

Example of line features

- begins-with-number
- begins-with-ordinal
- begins-with-punctuation
- begins-with-question-word
- begins-with-subject
- blank
- contains-alphanum
- contains-bracketed-number
- contains-http
- contains-non-space
- contains-number
- contains-pipe
- contains-question-mark
- contains-question-word
- ends-with-question-mark
- first-alpha-is-capitalized
- indented
- indented-1-to-4
- indented-5-to-10
- more-than-one-third-space
- only-punctuation
- prev-is-blank
- prev-begins-with-ordinal
- shorter-than-30

A McCallum, *ICML 2000*

Example of features

Is Capitalized
Is Mixed Caps
Is All Caps
Initial Cap
Contains Digit
All lowercase
Is Initial
Punctuation
Period
Comma
Apostrophe
Dash
Preceded by HTML tag

In stopword list
(the, of, their, etc)
In honorific list
(Mr, Mrs, Dr, Sen, etc)
In person suffix list
(Jr, Sr, PhD, etc)
In name particle list
(de, la, van, der, etc)
In lastname list;
segmented by P(name)
In firstname list;
segmented by P(name)
In locations lists
(states, cities, countries)
In company name list
("J. C. Penny")
In list of company suffixes
(Inc, & Associates, Foundation)

Word Features
- lists of job titles,
- Lists of prefixes
- Lists of suffixes
- 350 informative phrases

HTML/Formatting Features
- {begin, end, in} x
- {<b>, <i>, <a>, <hN>} x
- {lengths 1, 2, 3, 4, or longer}
- {begin, end} of line

Slide from Kauchak

Sliding Windows

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

Prefix window

Content window

Postfix window

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

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

Slide from Suchanek

Feature Vectors

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

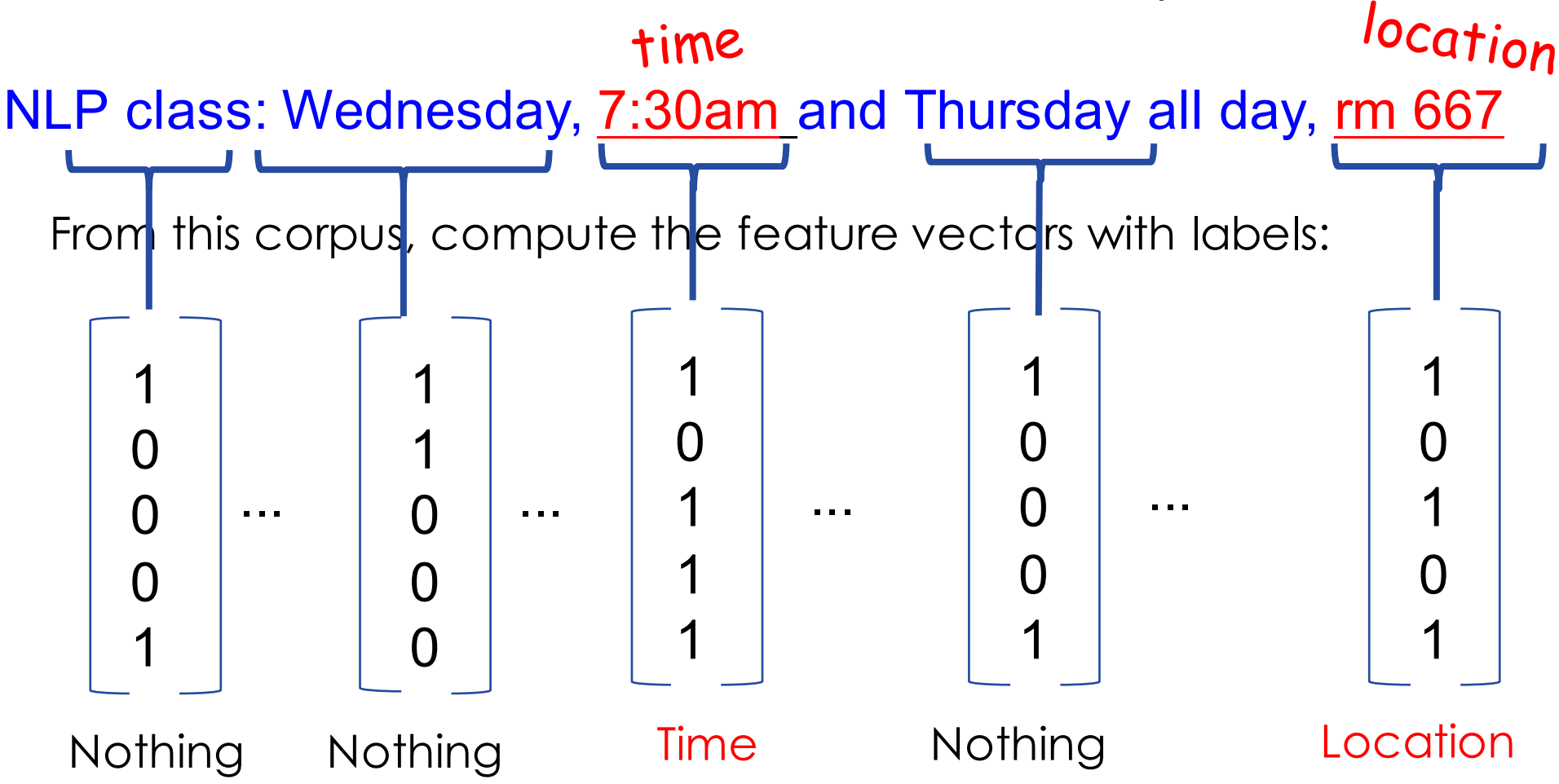
Prefix window | Content window | Postfix window

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

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

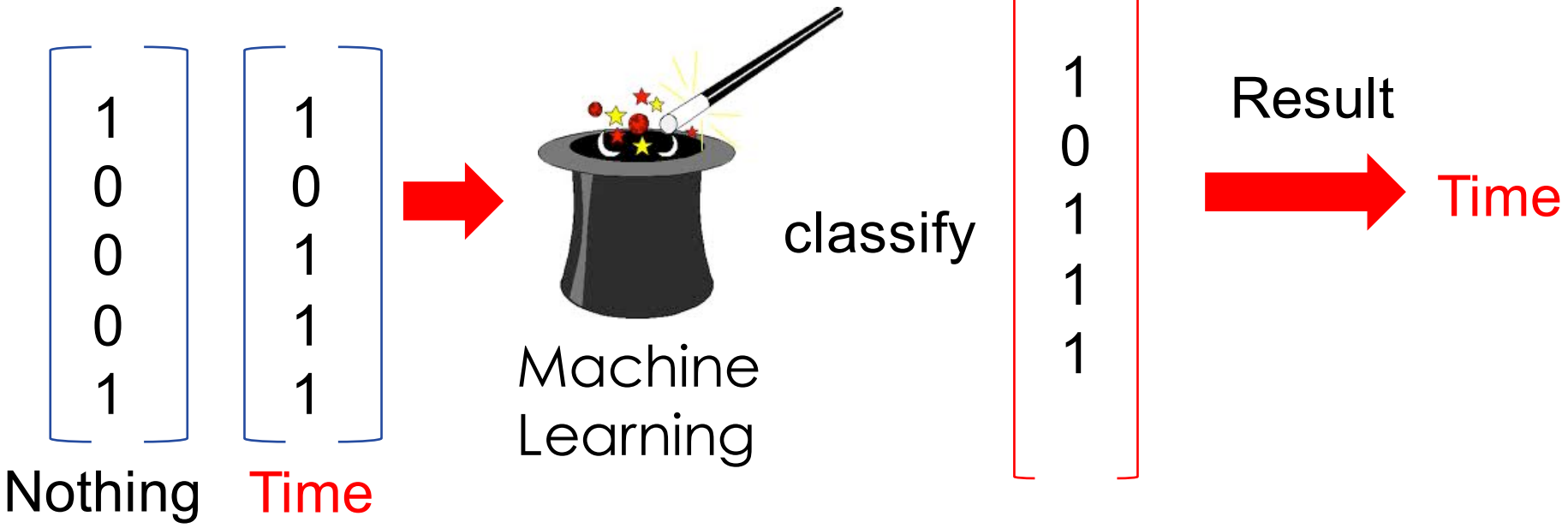


Slide from Suchanek

Machine Learning

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

Use the labeled feature vectors as training data for Machine Learning



NER: Machine Learning

- Feature encoding for ML

Token	POS	InitCap	AllCaps	Lowercase	Dot	Name	…	Class
John	N	1	0	0	0	1		PER
J	N	1	1	0	0	0		PER
.	DOT	0	0	0	1	0		PER
Smith	N	1	0	0	0	1		PER
lives	V	0	0	1	0	0		O
in	P	0	0	1	0	0		O
Seattle	N	1	0	0	0	0		LOC

NER: Machine Learning

- Feature encoding for ML

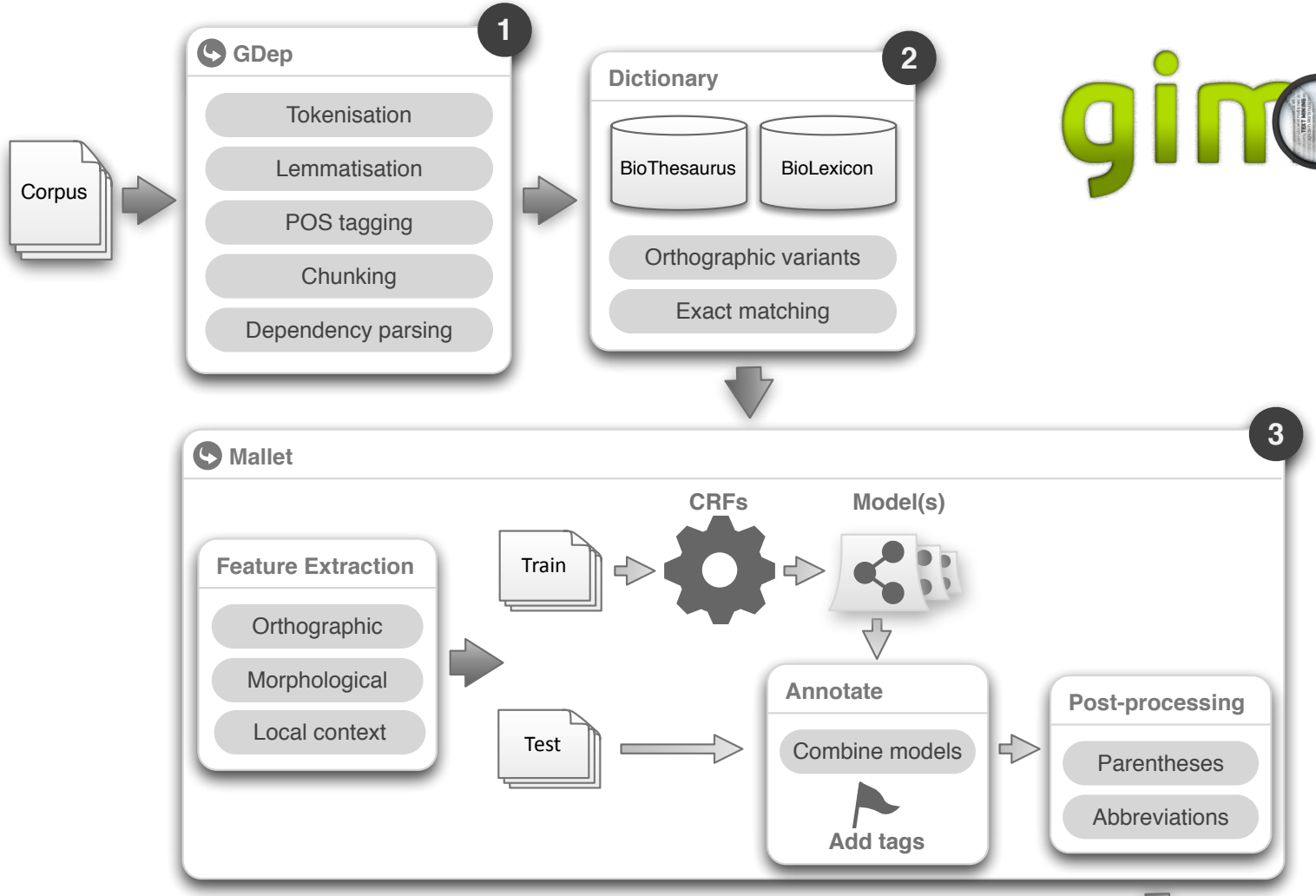
Token	POS	InitCap	AllCaps	Lowercase	Dot	Name	…	Class
John	N	1	0	0	0	1		PER_B
J	N	1	1	0	0	0		PER_I
.	DOT	0	0	0	1	0		PER_I
Smith	N	1	0	0	0	1		PER_I
lives	V	0	0	1	0	0		O
in	P	0	0	1	0	0		O
Seattle	N	1	0	0	0	0		LOC_B

BIO encoding

NER: Machine Learning

- Based on probabilistic models describing the surface characteristics of named entities of interest
- Model parameters estimated from training data
- Models
  - Maximum Entropy (ME), Hidden Markov Models (HMM), Maximum Entropy Markov Models (MEMM), Support Vector Machines (SVM), Conditional Random Fields (CRF)
- Advantages / disadvantages:
  - Increased recall
  - Requires training data
- Example (biomedical):
  - http://bioinformatics.ua.pt/software/gimli/
  - http://bioinformatics.ua.pt/neji/

GIMLI



Campos et al. BMC Bioinformatics 2013, 14:54
http://www.biomedcentral.com/1471-2105/14/54

BMC Bioinformatics

SOFTWARE Open Access

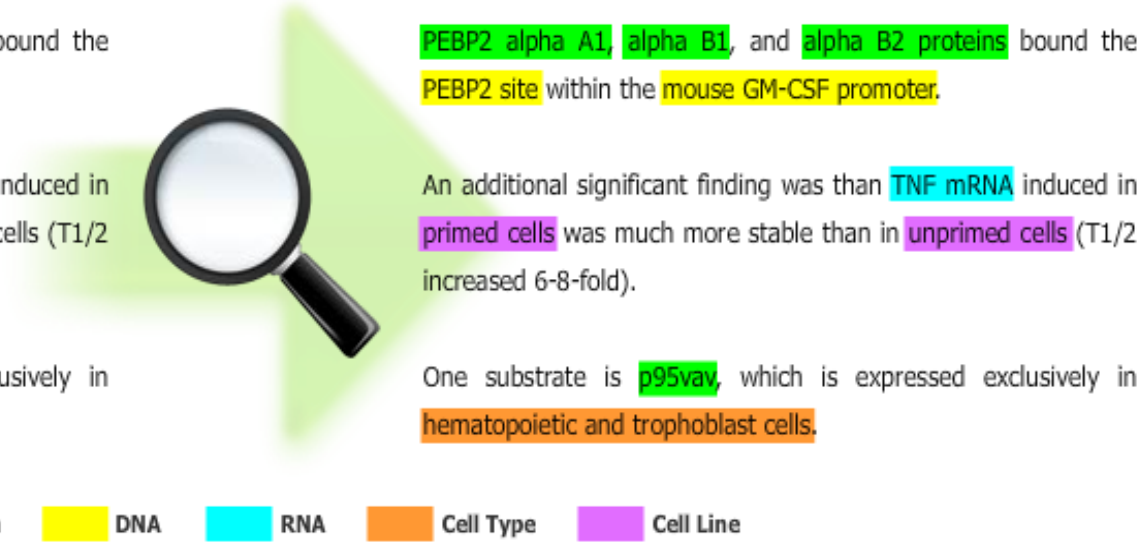
Gimli: open source and high-performance biomedical name recognition

GIMLI: some results

PEBP2 alpha A1, alpha B1, and alpha B2 proteins bound the PEBP2 site within the mouse GM-CSF promoter.

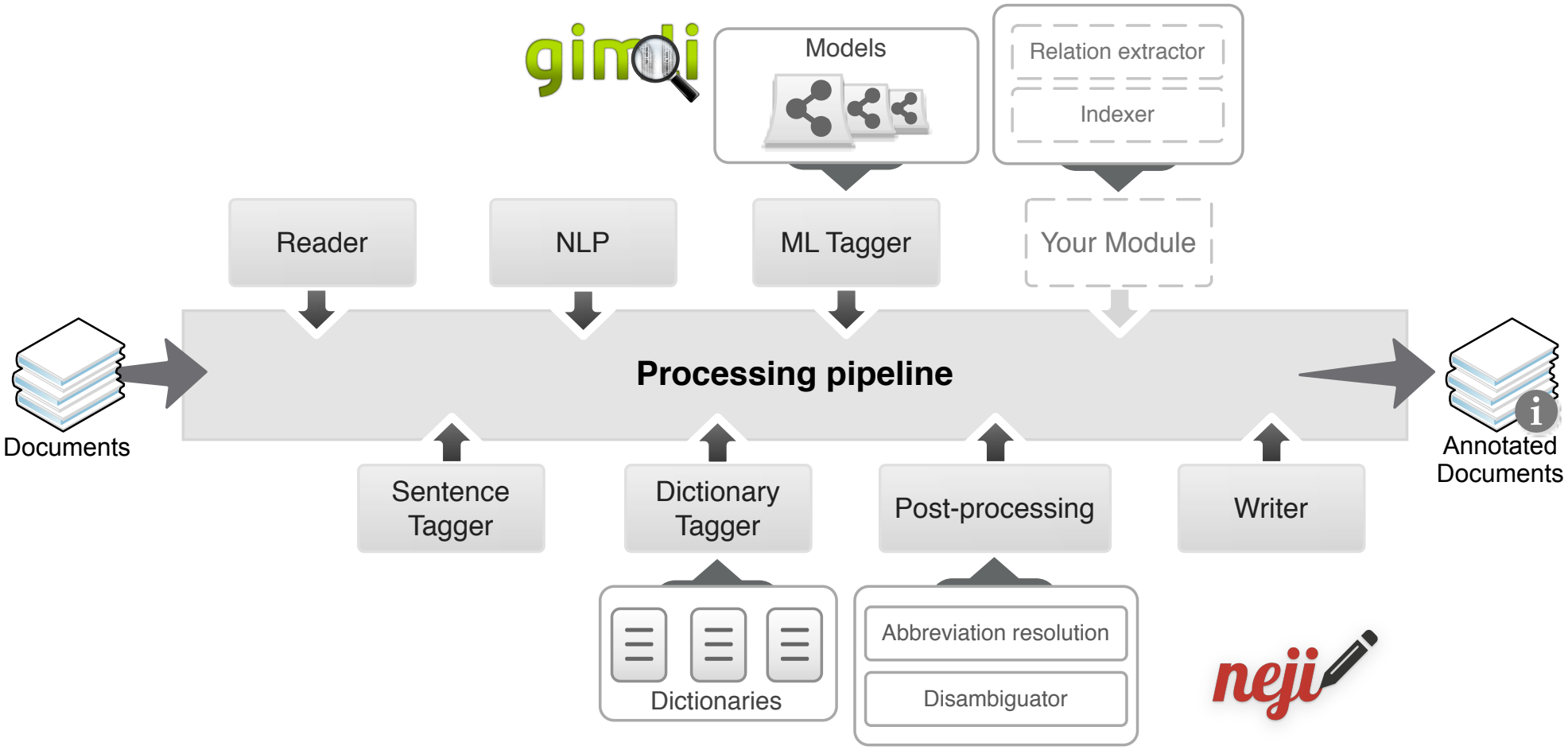
An additional significant finding was than TNF mRNA induced in primed cells was much more stable than in unprimed cells (T1/2 increased 6-8-fold).

One substrate is p95vav, which is expressed exclusively in hematopoietic and trophoblast cells.



- BioCreative: **87,54%**
- JNLPBA: **73,05%**

Neji: Processing pipeline



Campos et al. BMC Bioinformatics 2013, 14:54
http://www.biomedcentral.com/1471-2105/14/54

BMC Bioinformatics

SOFTWARE Open Access

Gimli: open source and high-performance biomedical name recognition

BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 9 2012, pages 1253–1261
doi:10.1093/bioinformatics/bts125

Data and text mining Advance Access publication March 13, 2012

Harmonization of gene/protein annotations: towards a gold standard MEDLINE

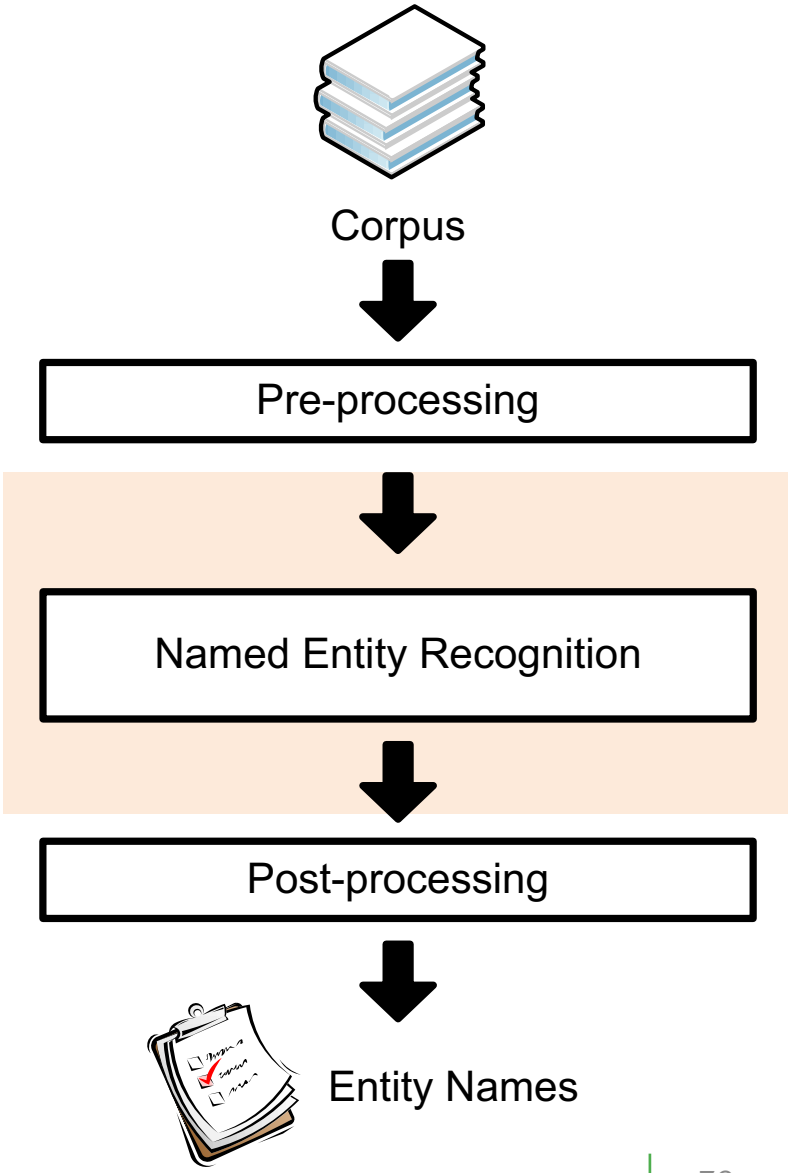
David Campos[1,*], Sérgio Matos[1], Ian Lewin[2], José Luís Oliveira[1] and Dietrich Rebholz-Schuhmann[2,*]

NER: Hybrid solutions

Approaches

- **Dictionary matching**
- **Rule-based matching**
- **Machine learning**

Exploit **strengths** of **each approach**



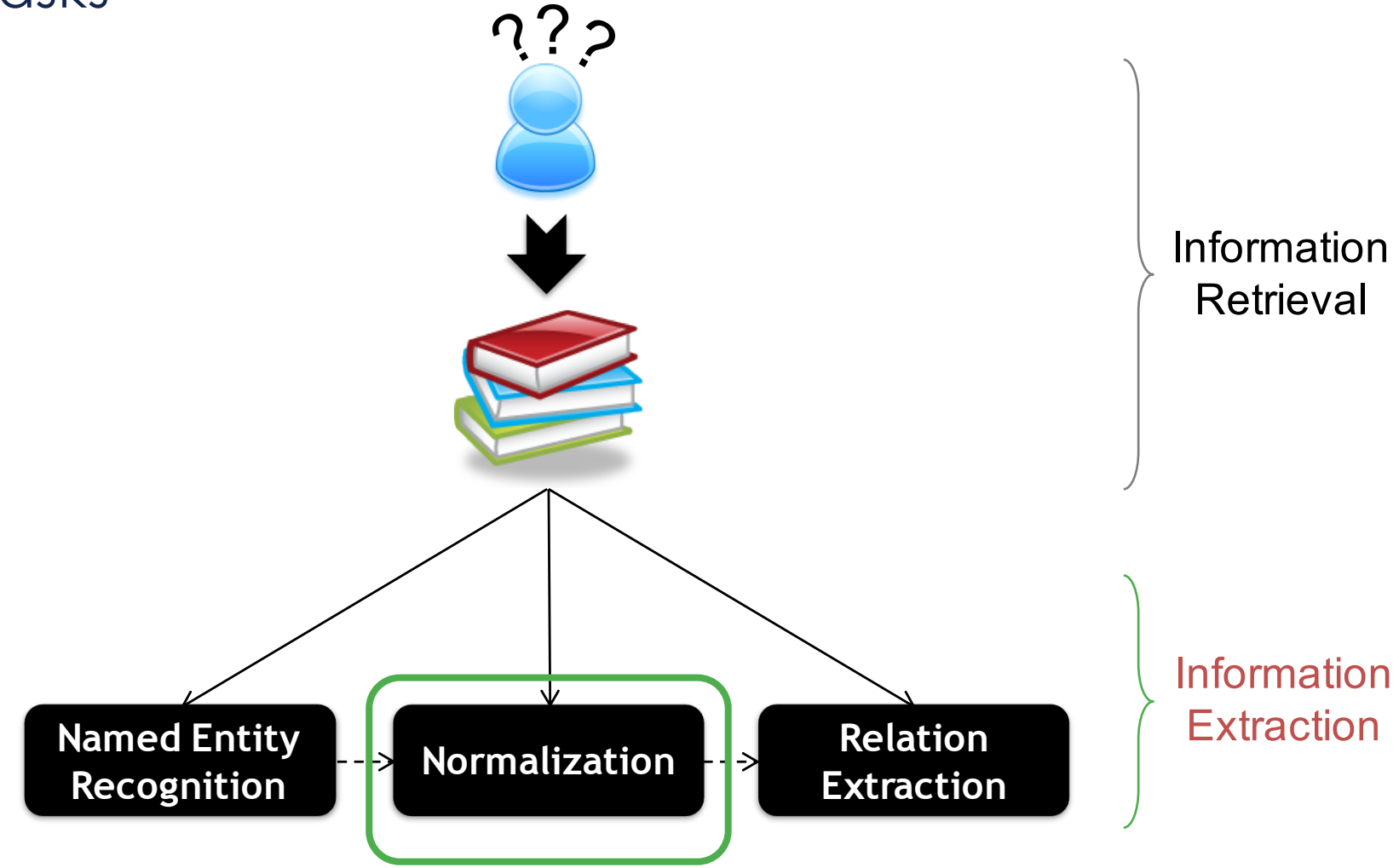
NER: Hybrid solutions

- ML + dictionaries
  - dictionaries used as features for ML and/or
  - to complete/correct the results of ML
  - ML used to correct results of dictionary matching (e.g. eliminate FPs)

- ML (+ dictionaries) + rules
  - rules used to correct results (e.g. eliminate FPs; add missed acronyms/long forms)
  - rules used to clean results (e.g. eliminate unpaired parentheses)

Text Mining

- Tasks



Normalization

- Goal
  - **Assign a unique identifier to each entity**

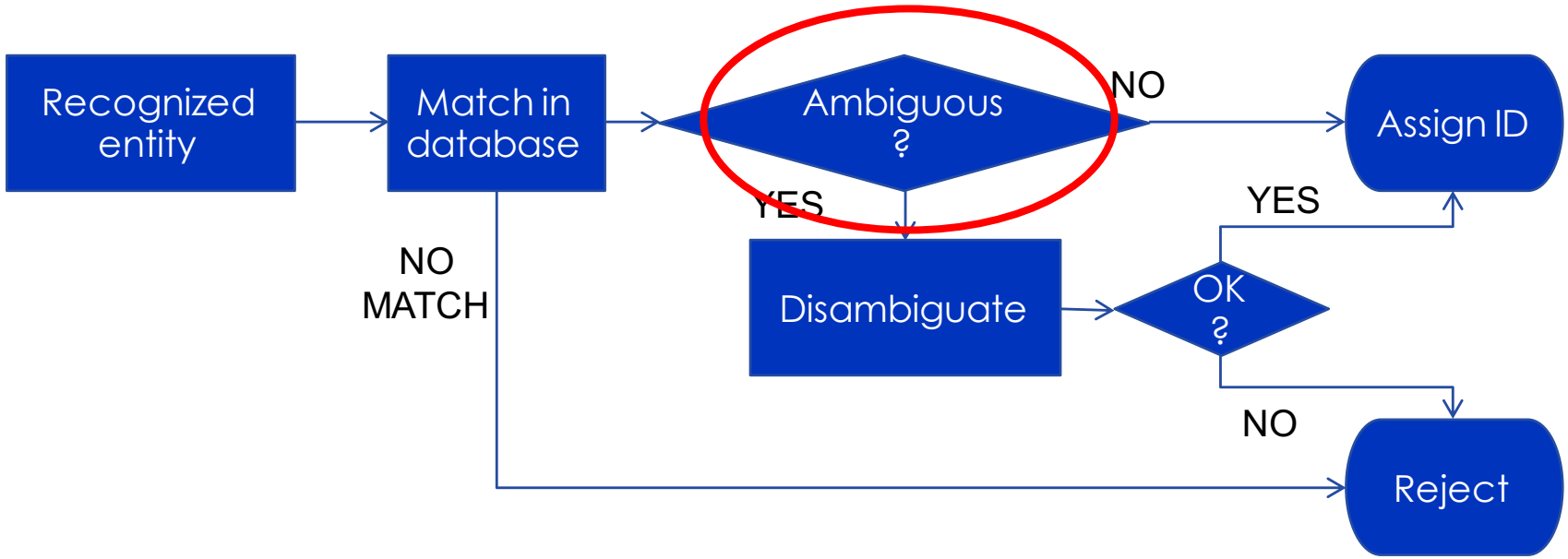
- Example
  - Text
    - "Folliculin encoded by the BHD gene …"
  - NER
    - "<protein>Folliculin</protein> encoded by the <gene>BHD</gene> gene …"
  - Normalization
    - "<protein:uniprot=Q8NFG4>Folliculin</protein> encoded by the <gene:Entrez=201163>BHD</gene> gene …"

Normalization

- Example
  - "<protein:uniprot=Q8NFG4>Folliculin</protein> encoded by the <gene:Entrez=201163>BHD</gene> gene …"
- Benefits
  - Link a textual mention of an entity to an unambiguous database entry (and link entities to documents mentioning those entities)
  - Link facts extracted from text to the correct entity
- Challenges
  - Synonymy
  - Acronyms, e.g. Birt-Hogg-Dube syndrome (BHD)
  - Ambiguity

Normalization

- General approach



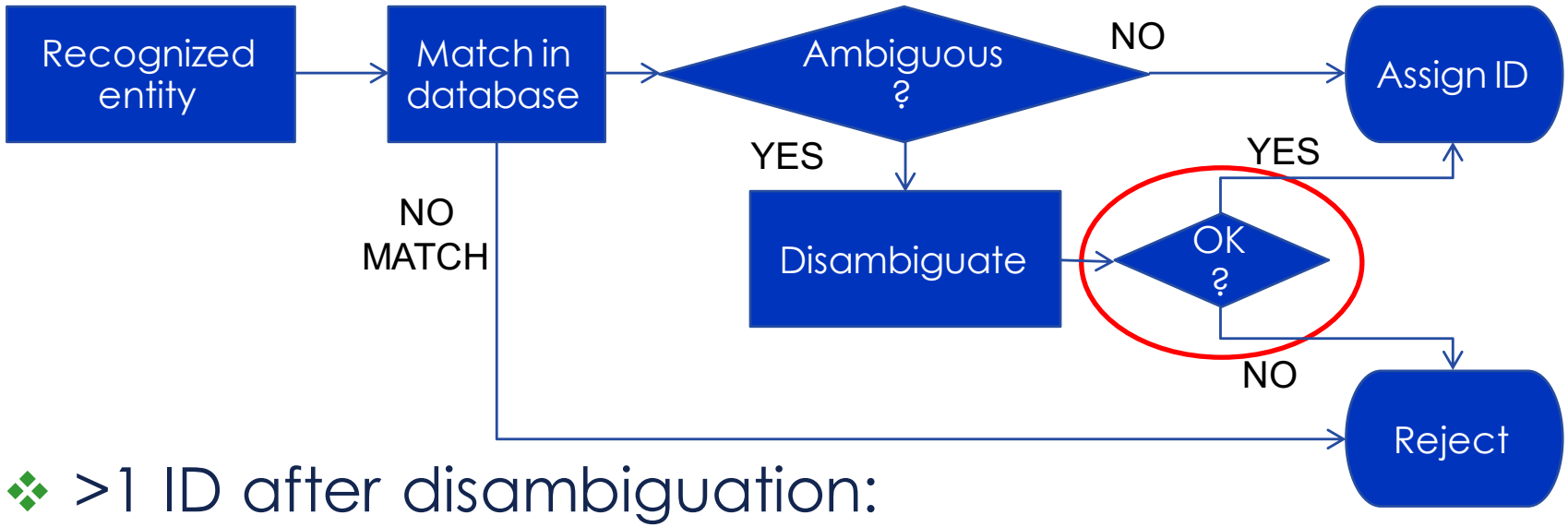
- Ambiguity problem
  - A match associates a name with multiple entity identifiers
  - Need to determine the type of an entity, e.g. Java, Paris

Disambiguation strategies

- Dictionary-based NER:
  - IDs can be included in the dictionary
- Linguistic
  - Syntactic patterns: "BHD, an inherited genetic condition..."
  - Lexical/domain clues: "Folliculin encoded by …"
- Statistical
  - Corpora-based profiles - Uses co-occurring terms in large corpora to define a disambiguating profile for each entity
- ML-based
  - Use training corpora to create a probabilistic model for disambiguation - Also based on co-occurring terms, used as features
- Use of domain or "universal" knowledge
  - Ontologies, folksonomies: e.g. Wikipedia

Normalization

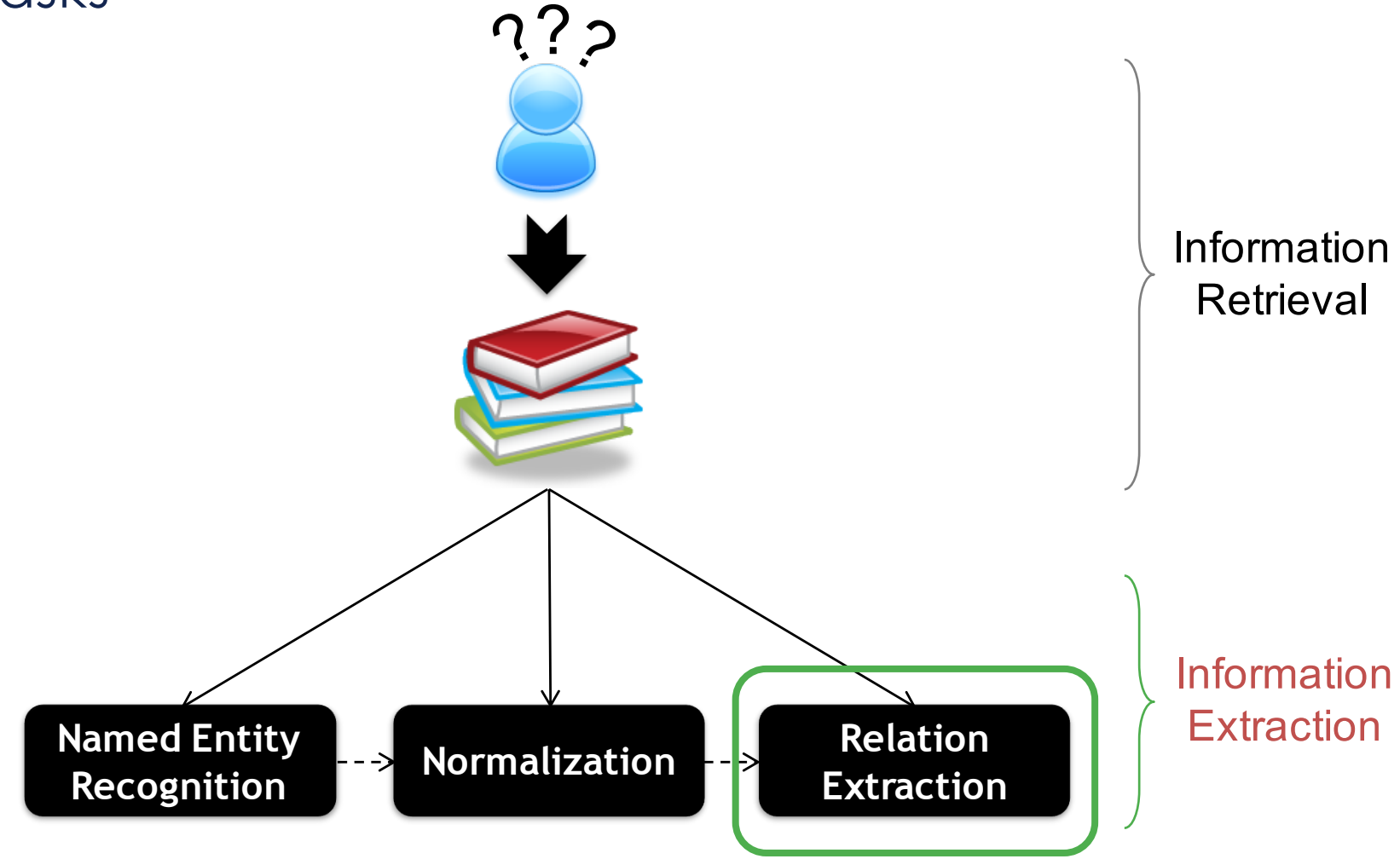
- General approach



- >1 ID after disambiguation:
  - Random ☺
  - Select top ranking option (w/ given probability)
  - Simply discard

Text Mining

- Tasks



Relation Extraction

- NLP Stanford course by Dan Jurafsky

IEETA IBT & Bioinformatics research

Research lines

- Genome analysis
  - mRNA mistranslation, Gene design, Microarrays
- Data and knowledge integration
  - Integration of clinical and genetic information
- **Text mining**
  - **Information extraction, NER, PPI, Normalization**
  - **Query expansion, semantics**
- Biomedical Informatics
  - PACS, P2P, Cloud,
  - Biomedical applications for health

Text mining in scientific literature

Metab Brain Dis. 2012 Dec;27(4):595-603. doi: 10.1007/s11011-012-9319-5. Epub 2012 May 27.

Effect of histidine administration to female rats during pregnancy and lactation on enzymes activity of phosphoryltransfer network in cerebral cortex and hippocampus of the offspring.

Rojas DB, de Andrade RB, Gemelli T, Oliveira LS, Campos AG, Dutra-Filho CS, Wannmacher CM.

Departamento de Bioquímica, Programa de Pós-Graduação em Ciências Biológicas: Bioquímica, Instituto de Ciências Básicas da Saúde, Universidade Federal do Rio Grande do Sul, Rua Ramiro Barcelos, 2600 anexo, 90035-003, Porto Alegre, RS, Brazil.

Abstract

Histidinemia is an inborn error of metabolism of amino acids caused by deficiency of histidase activity in liver and skin with consequent accumulation of histidine in plasma and tissues. Histidinemia is an autosomal recessive trait usually considered harmless to patients and their offspring, but some patients and children born from histidinemic mothers have mild neurologic alterations. Considering that histidinemia is one of the most frequently identified metabolic conditions, in the present study we investigated the effect of L-histidine load to female rats during pregnancy and lactation on some parameters of phosphoryltransfer network in cerebral cortex and hippocampus of the offspring. Pyruvate kinase, cytosolic and mitochondrial creatine kinase activities decreased in cerebral cortex and in hippocampus of rats at 21 days of age and this pattern remained in the cerebral cortex and in hippocampus at 60 days of age. Moreover, adenylate kinase activity was reduced in the cerebral cortex and in hippocampus of the offspring at 21 days of age, whereas the activity was increased in the two tissues at 60 days of age. These results suggest that administration of L-histidine to female rats in the course of pregnancy and lactation could impair energy homeostasis in the cerebral cortex and hippocampus of the offspring. Considering that histidinemia is usually a benign condition and little attention has been given to maternal histidinemia, it seems important to perform more studies in the children born from histidinemic mothers.

PMID: 22638695 [PubMed - indexed for MEDLINE]

Text mining in scientific literature

Metab Brain Dis. 2012 Dec;27(4):595-603. doi: 10.1007/s11011-012-9319-5. Epub 2012 May 27.

Effect of histidine administration to female rats during pregnancy and lactation on enzymes activity of phosphoryltransfer network in cerebral cortex and hippocampus of the offspring.

Rojas DB, de Andrade RB, Gemelli T, Oliveira LS, Campos AG, Dutra-Filho CS, Wannmacher CM.

Departamento de Bioquímica, Programa de Pós-Graduação em Ciências Biológicas: Bioquímica, Instituto de Ciências Básicas da Saúde, Universidade Federal do Rio Grande do Sul, Rua Ramiro Barcelos, 2600 anexo, 90035-003, Porto Alegre, RS, Brazil.

Abstract

Histidinemia is an inborn error of metabolism of amino acids caused by deficiency of histidase activity in liver and skin with consequent accumulation of histidine in plasma and tissues. Histidinemia is an autosomal recessive trait usually considered harmless to patients and their offspring, but some patients and children born from histidinemic mothers have mild neurologic alterations. Considering that histidinemia is one of the most frequently identified metabolic conditions, in the present study we investigated the effect of L-histidine load to female rats during pregnancy and lactation on some parameters of phosphoryltransfer network in cerebral cortex and hippocampus of the offspring. Pyruvate kinase, cytosolic and mitochondrial creatine kinase activities decreased in cerebral cortex and in hippocampus of rats at 21 days of age and this pattern remained in the cerebral cortex and in hippocampus at 60 days of age. Moreover, adenylate kinase activity was reduced in the cerebral cortex and in hippocampus of the offspring at 21 days of age, whereas the activity was increased in the two tissues at 60 days of age. These results suggest that administration of L-histidine to female rats in the course of pregnancy and lactation could impair energy homeostasis in the cerebral cortex and hippocampus of the offspring. Considering that histidinemia is usually a benign condition and little attention has been given to maternal histidinemia, it seems important to perform more studies in the children born from histidinemic mothers.

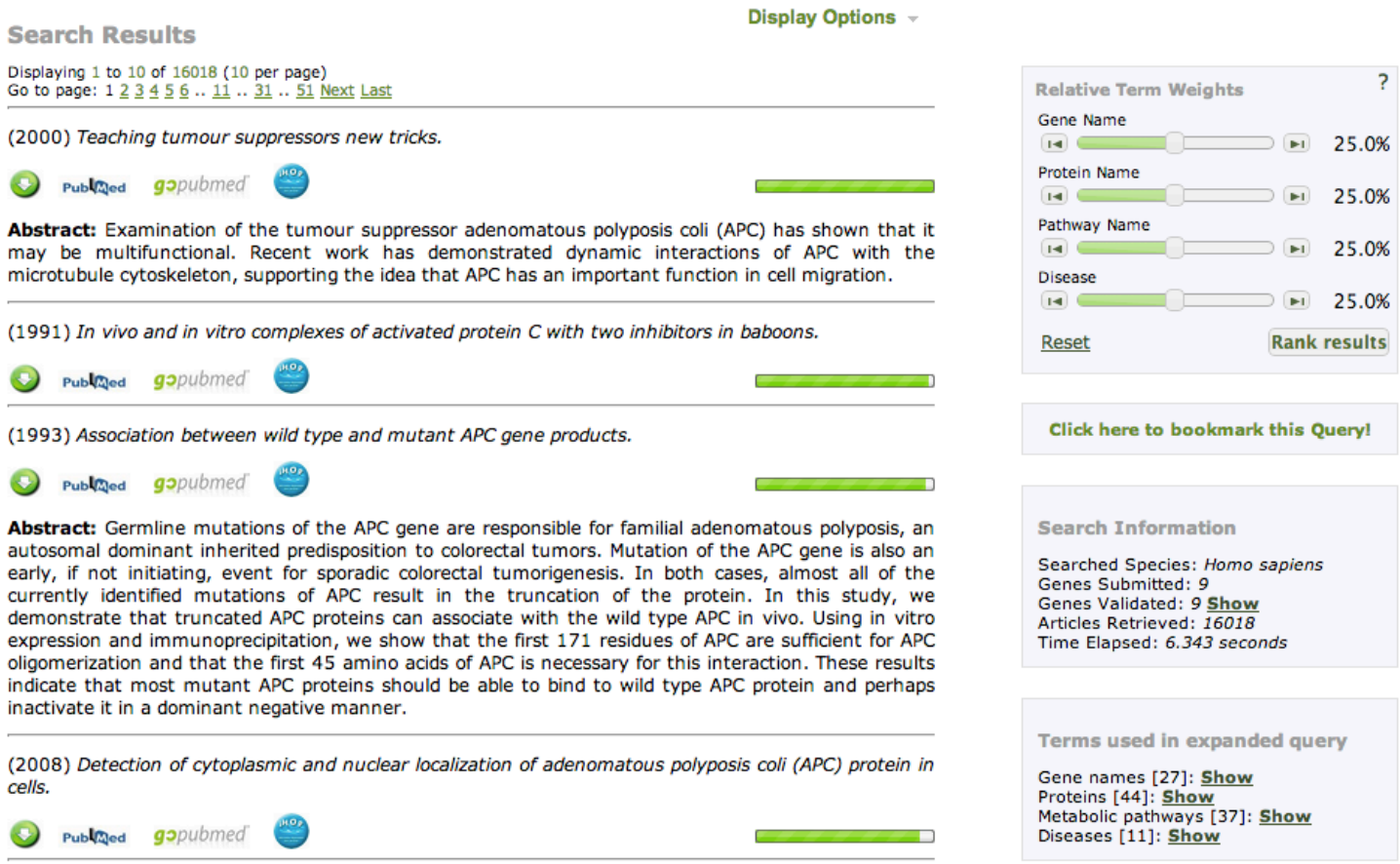
PMID: 22638695 [PubMed - indexed for MEDLINE]

Entities → Concepts **Relations, Events, Assertions**

Some research

- Query Expansion
  - Quext
- Named Entity Recognition
  - GIMLI
- Social Mining
- TM framework for biomedical domain
  - NEJI
- User interaction
  - BECAS, EGAS

Query expansion



- Data integration
- Focused indexing and search
- Query processing and expansion

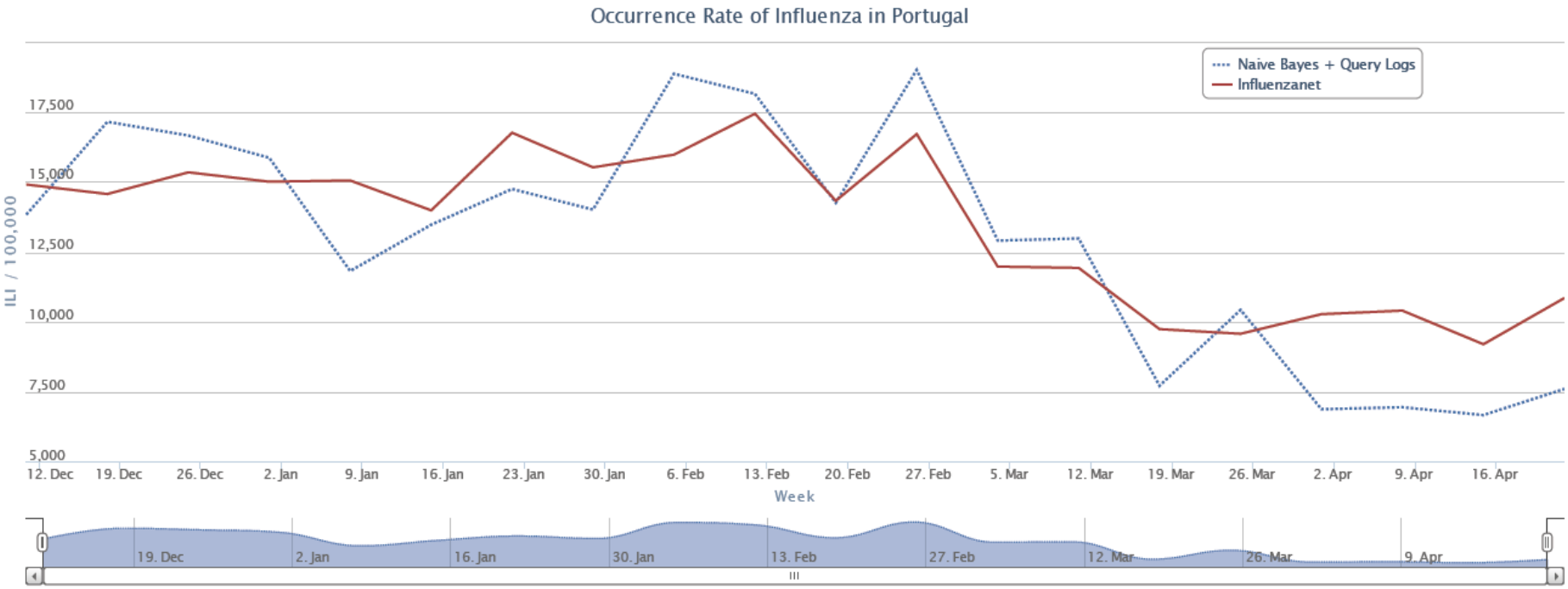
QuExT

Matos et al., *BMC Bioinformatics* 2010, **11**:212

Social mining

- Detection of health conditions

Social mining



THEORETICAL BIOLOGY AND MEDICAL MODELLING

Santos, J.C., Matos, S., Analysing Twitter and web queries for flu trend prediction

OPEN ACCESS Freely available online

PLOS ONE

Twitter: A Good Place to Detect Health Conditions

Víctor M. Prieto[1]*, Sérgio Matos[2], Manuel Álvarez[1], Fidel Cacheda[1], José Luís Oliveira[2]

Gimli

- **Machine learning-based tool for biomedical Named Entity Recognition (NER)**

http://bioinformatics.ua.pt/gimli

- Main features:
  - Train new machine learning models with custom feature sets and model parameters
  - Annotate documents with trained models

Campos *et al. BMC Bioinformatics* 2013, **14**:54
http://www.biomedcentral.com/1471-2105/14/54

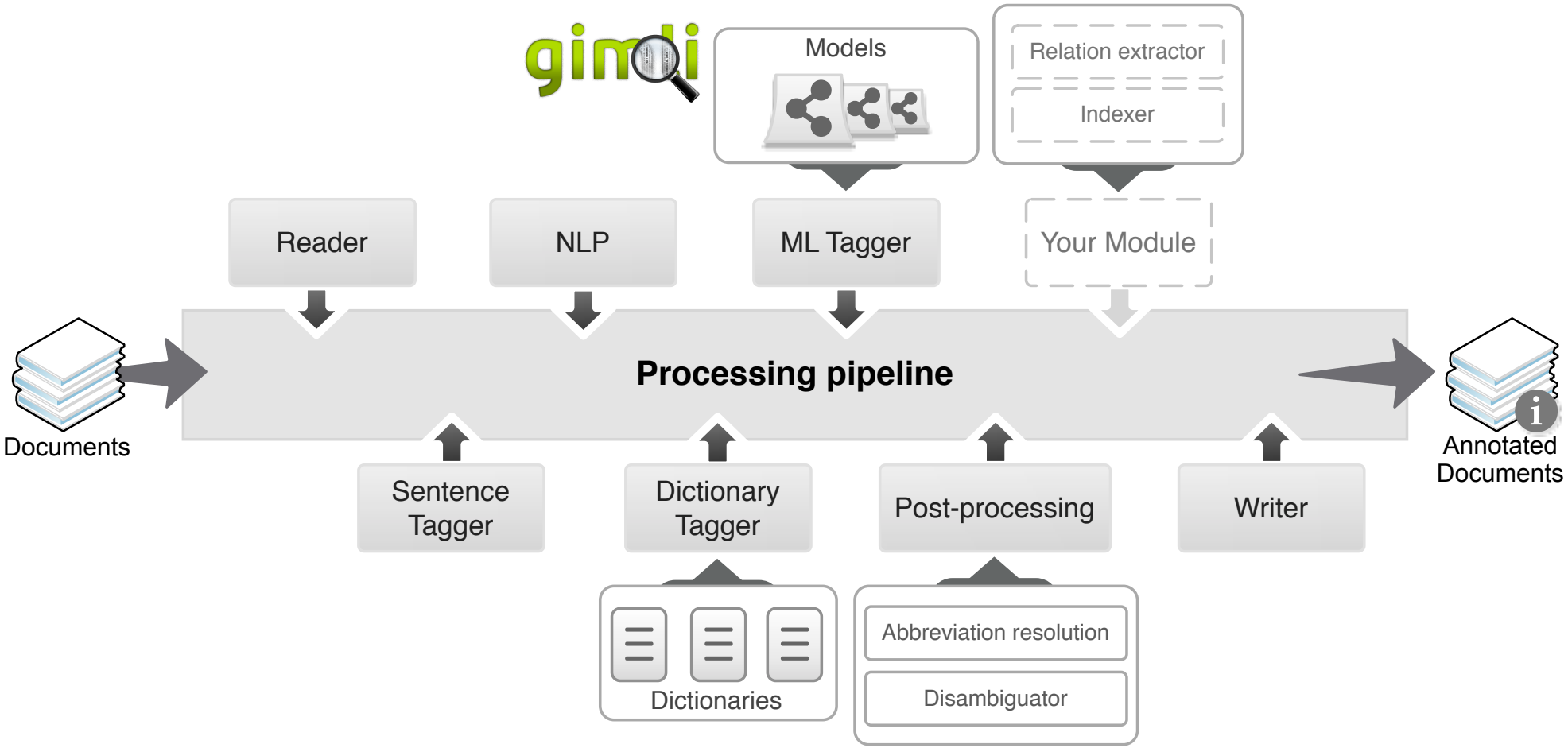
BMC Bioinformatics

SOFTWARE Open Access

Gimli: open source and high-performance biomedical name recognition

David Campos*, Sérgio Matos and José Luís Oliveira

Neji: Processing pipeline



BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 9 2012, pages 1253–1261
doi:10.1093/bioinformatics/bts125

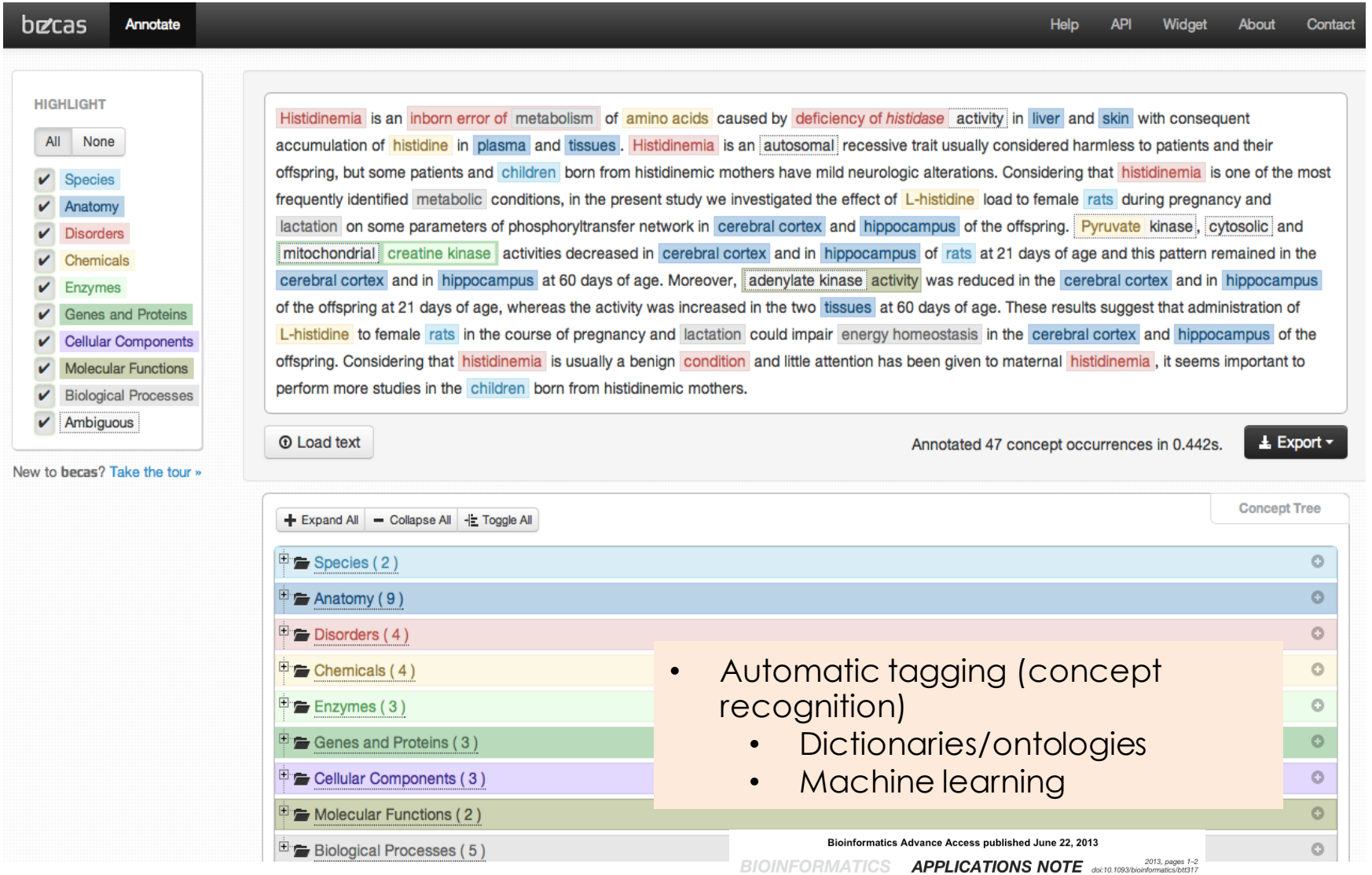
Data and text mining Advance Access publication March 13, 2012

Harmonization of gene/protein annotations: towards a gold standard MEDLINE

David Campos[1,*], Sérgio Matos[1], Ian Lewin[2], José Luís Oliveira[1] and Dietrich Rebholz-Schuhmann[2,*]

BeCAS - Concept recognition

becas



- Automatic tagging (concept recognition)
  - Dictionaries/ontologies
  - Machine learning


Bioinformatics Advance Access published June 22, 2013

BIOINFORMATICS APPLICATIONS NOTE 2013, pages 1–2 doi:10.1093/bioinformatics/btt317

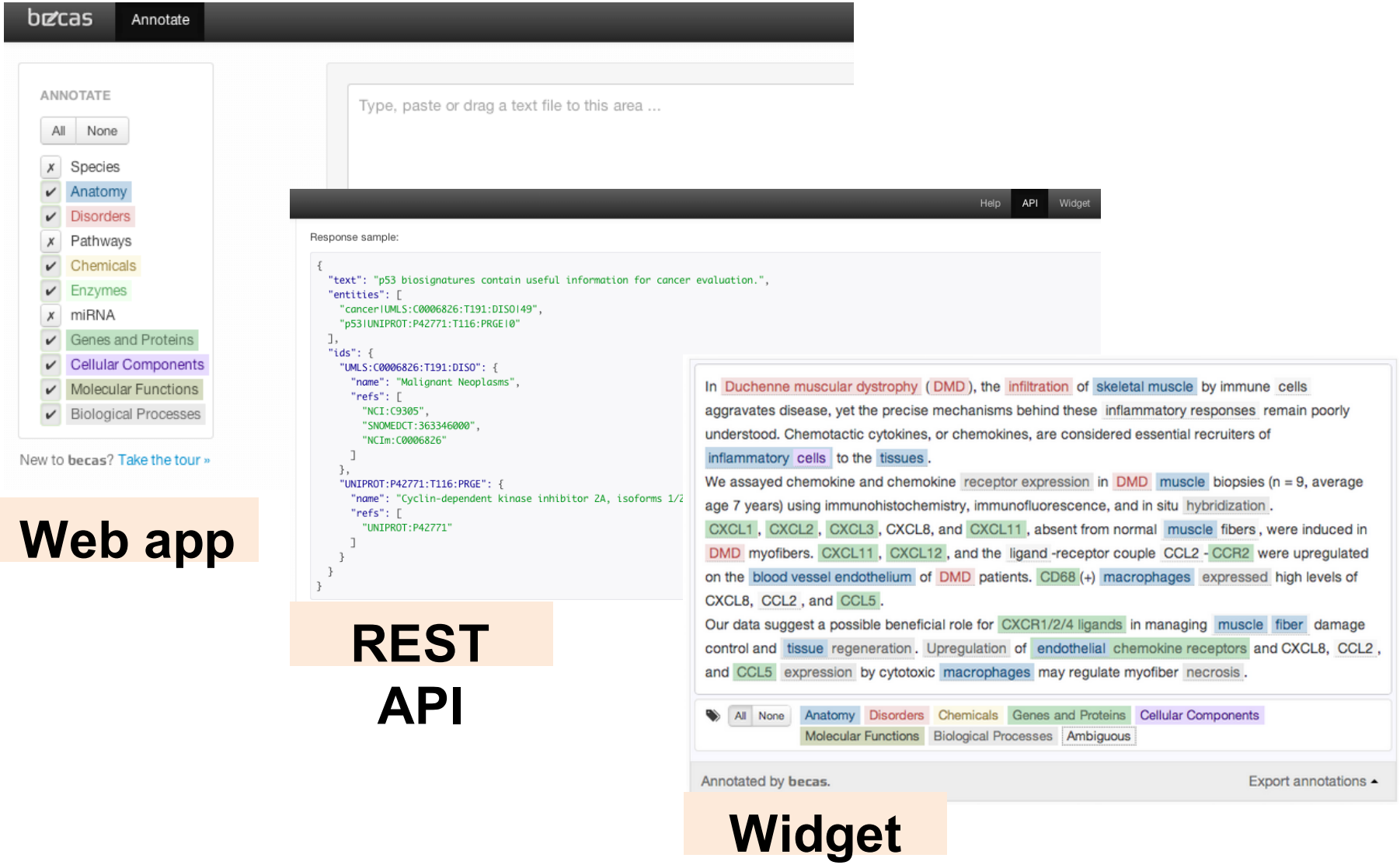
Data and text mining Advance Access publication June 4, 2013

BeCAS: biomedical concept recognition services and visualization

Tiago Nunes*, David Campos, Sérgio Matos and José Luís Oliveira*

BeCAS - Concept recognition

becas



Web app

REST API

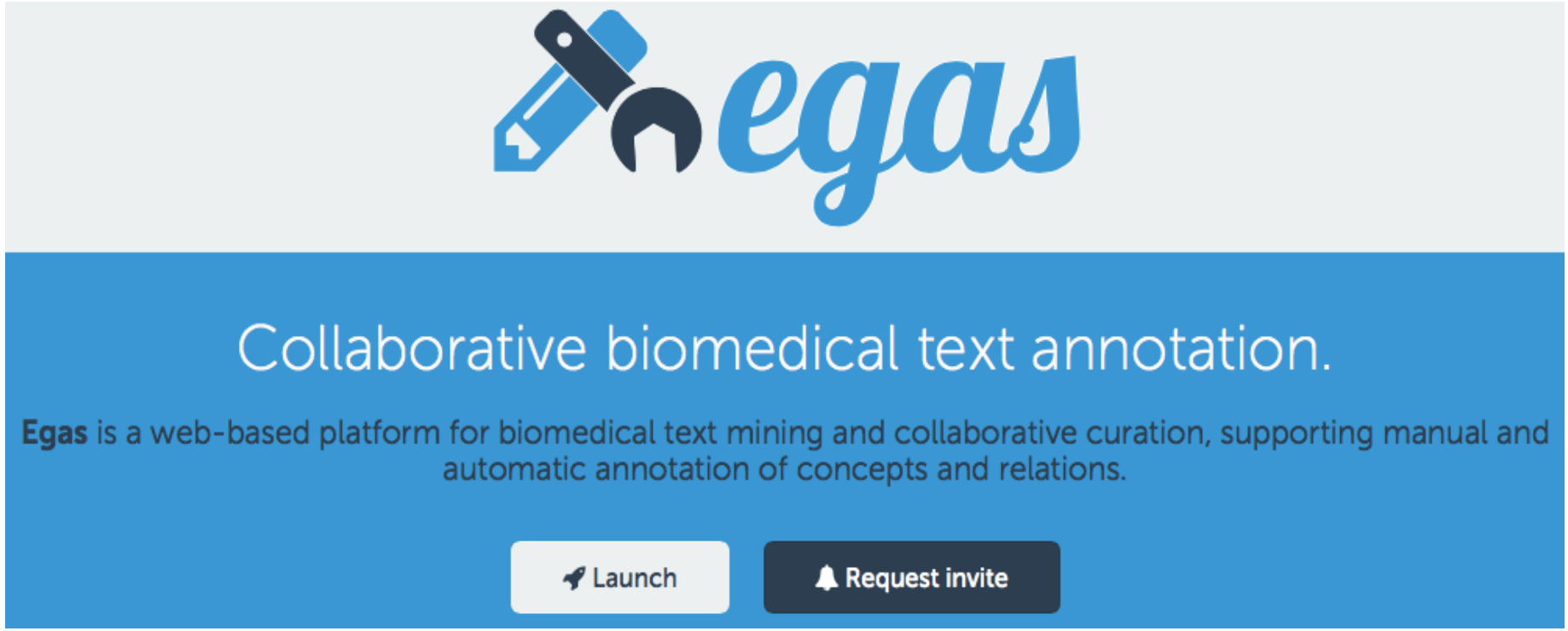
Widget

How far can we go?

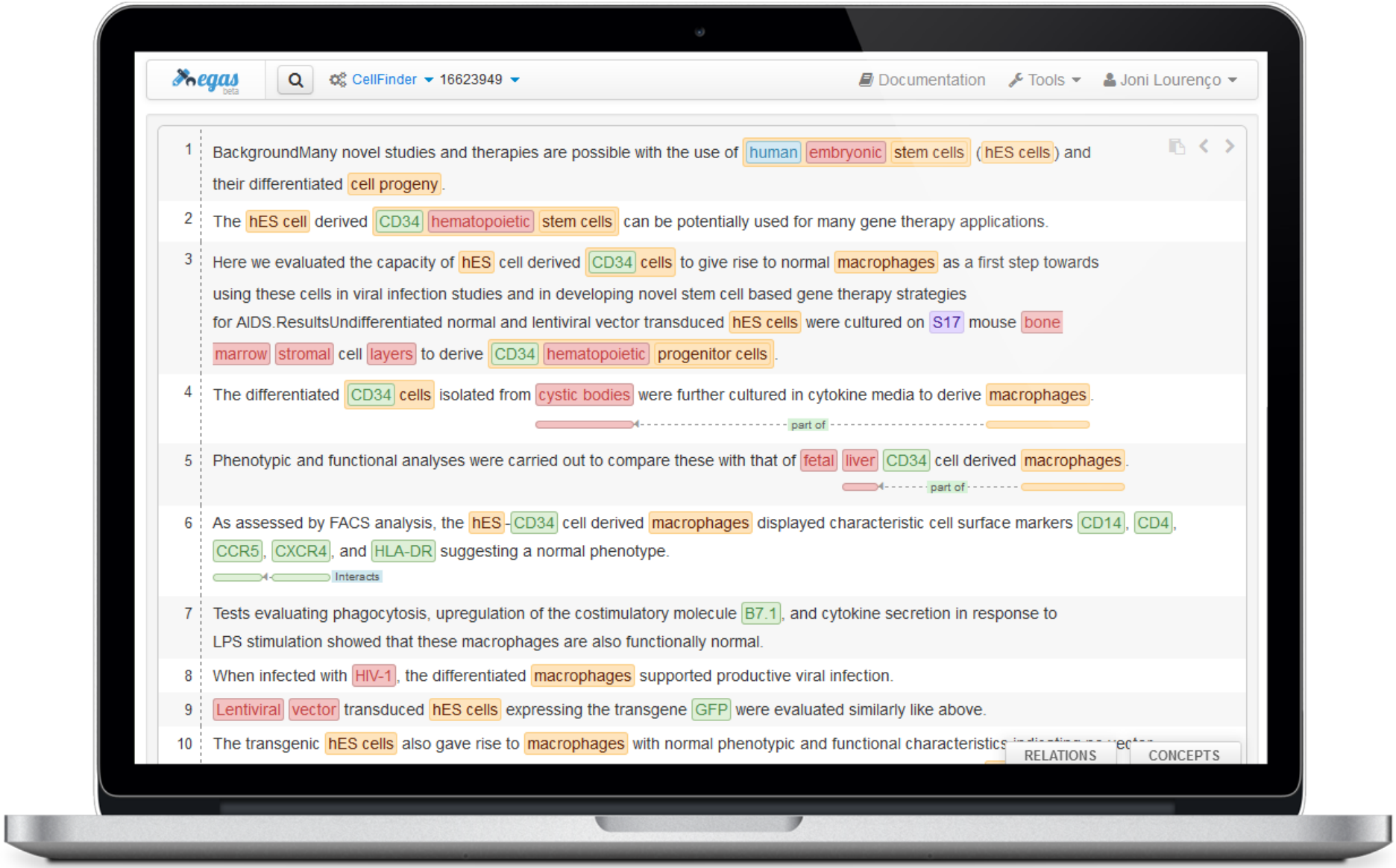
- Automatic text mining already reach F-measure ~80-90%
  - For common concepts (genes, proteins, diseases,..)

- What about the other 10-20% ?
  - Quality is a major issue
  - Curation is still needed

Collaborative curation



EGAS


egas beta
CellFinder 16623949
Documentation
Tools
Joni Lourenço
1 BackgroundMany novel studies and therapies are possible with the use of human embryonic stem cells (hES cells) and their differentiated cell progeny.
2 The hES cell derived CD34 hematopoietic stem cells can be potentially used for many gene therapy applications.
3 Here we evaluated the capacity of hES cell derived CD34 cells to give rise to normal macrophages as a first step towards using these cells in viral infection studies and in developing novel stem cell based gene therapy strategies for AIDS.ResultsUndifferentiated normal and lentiviral vector transduced hES cells were cultured on S17 mouse bone marrow stromal cell layers to derive CD34 hematopoietic progenitor cells.
4 The differentiated CD34 cells isolated from cystic bodies were further cultured in cytokine media to derive macrophages.
part of
5 Phenotypic and functional analyses were carried out to compare these with that of fetal liver CD34 cell derived macrophages.
part of
6 As assessed by FACS analysis, the hES-CD34 cell derived macrophages displayed characteristic cell surface markers CD14, CD4, CCR5, CXCR4, and HLA-DR suggesting a normal phenotype.
Interacts
7 Tests evaluating phagocytosis, upregulation of the costimulatory molecule B7.1, and cytokine secretion in response to LPS stimulation showed that these macrophages are also functionally normal.
8 When infected with HIV-1, the differentiated macrophages supported productive viral infection.
9 Lentiviral vector transduced hES cells expressing the transgene GFP were evaluated similarly like above.
10 The transgenic hES cells also gave rise to macrophages with normal phenotypic and functional characteristics
RELATIONS
CONCEPTS

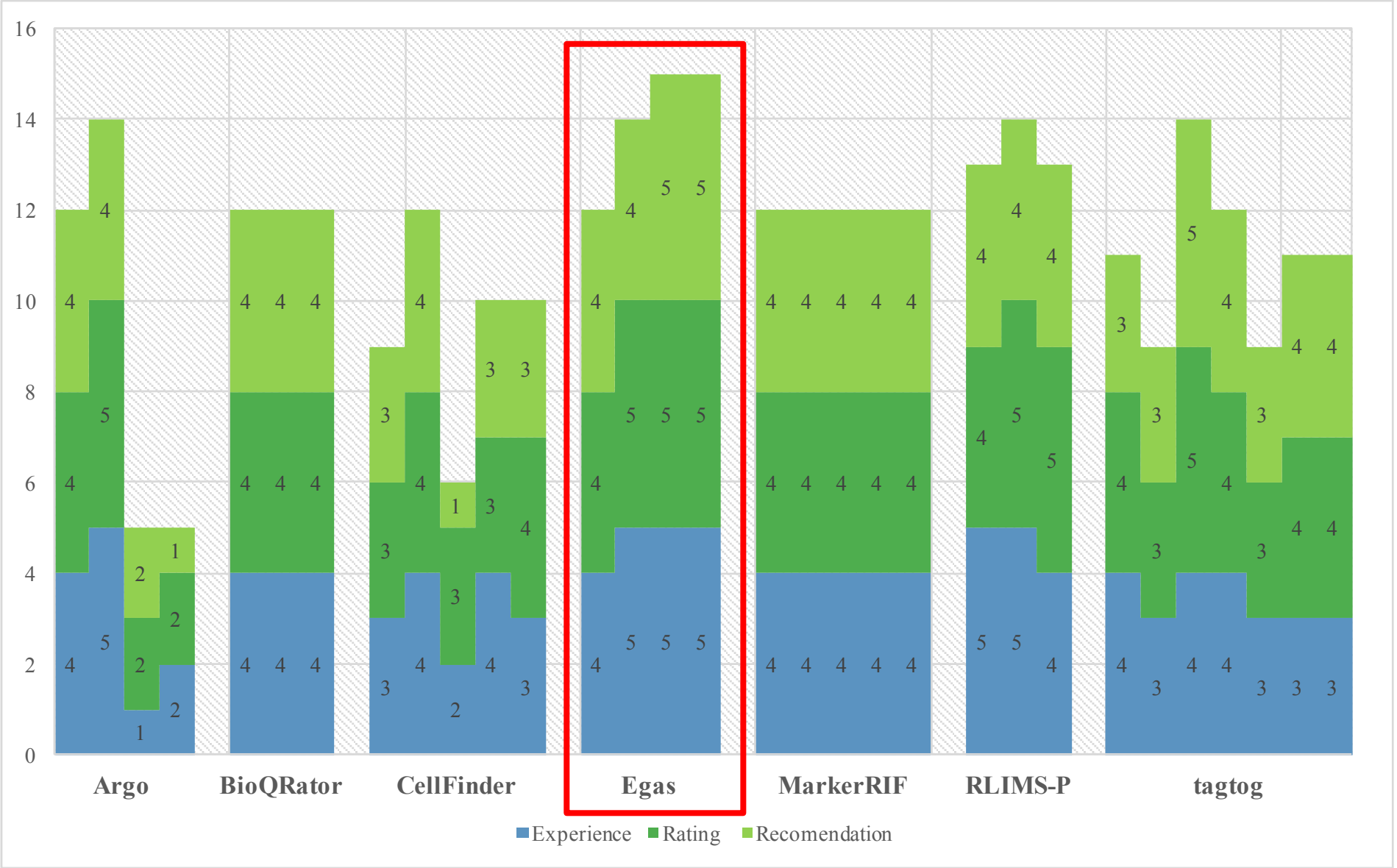
BioCreative IV

- Track 5- User Interactive Task (IAT)

- 4 annotators
- 50 documents x 4
  - 25 automatic annotation
  - 25 manual annotation

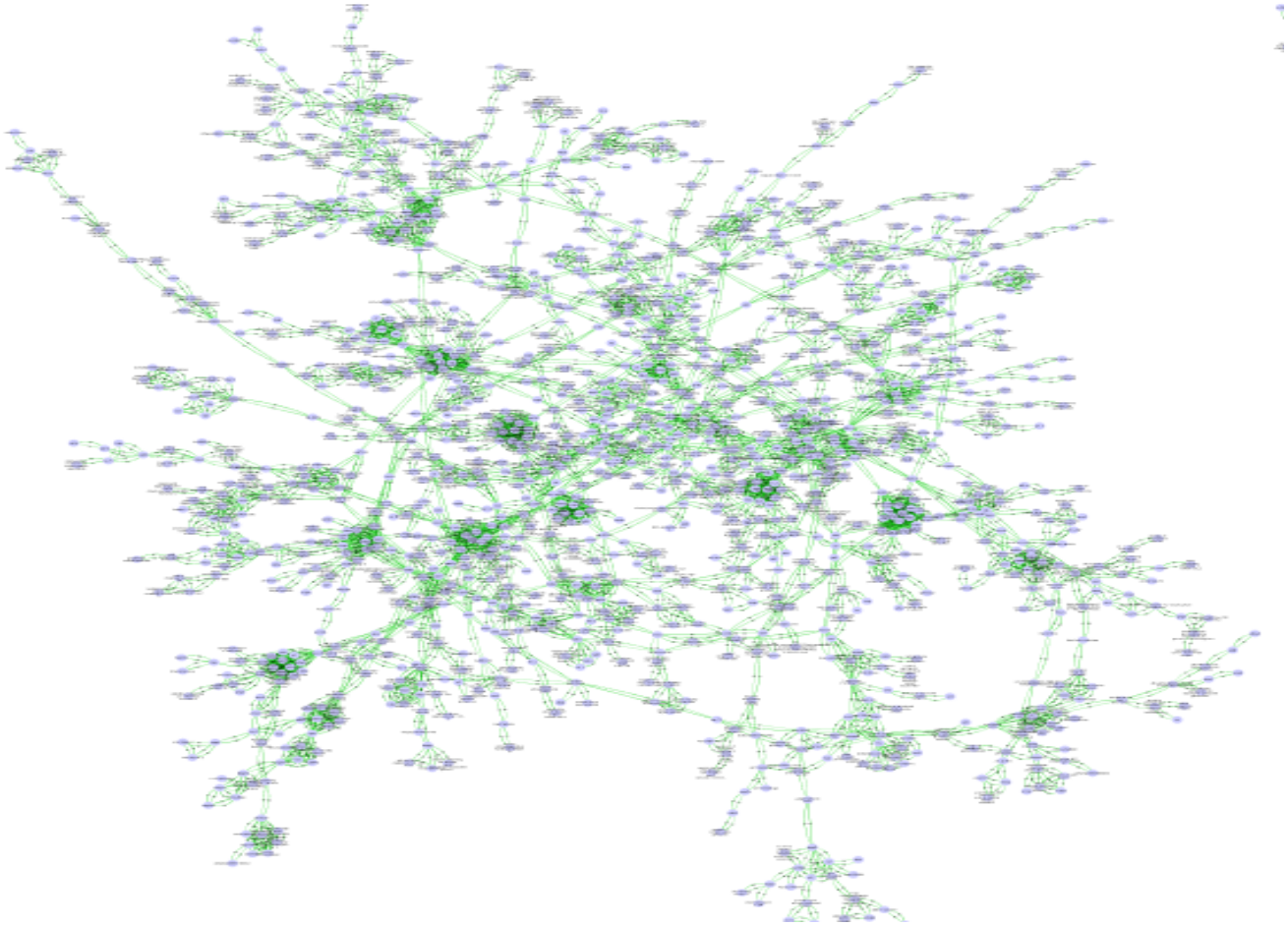
D. Campos, J. Lourenço, T. Nunes, R. Vitorino, P. Domingues, S. Matos, and J. L. Oliveira, ***Egas - Collaborative Biomedical Annotation as a Service***, in Fourth BioCreative Challenge Evaluation Workshop, Bethesda, Maryland, USA, Oct. 2013, p. 254–259;

BioCreative IV IAT



Relation Extraction

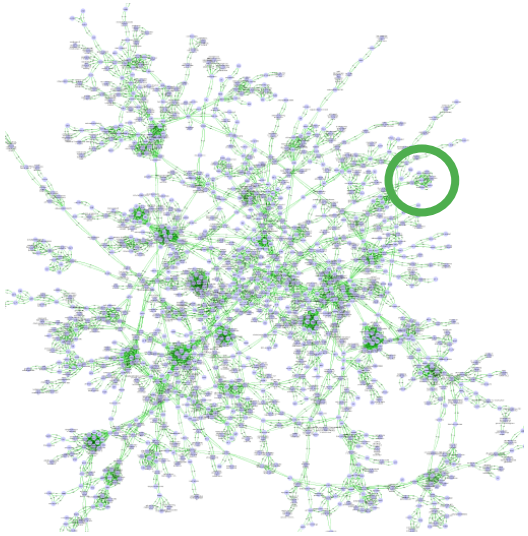
- Protein-protein interactions

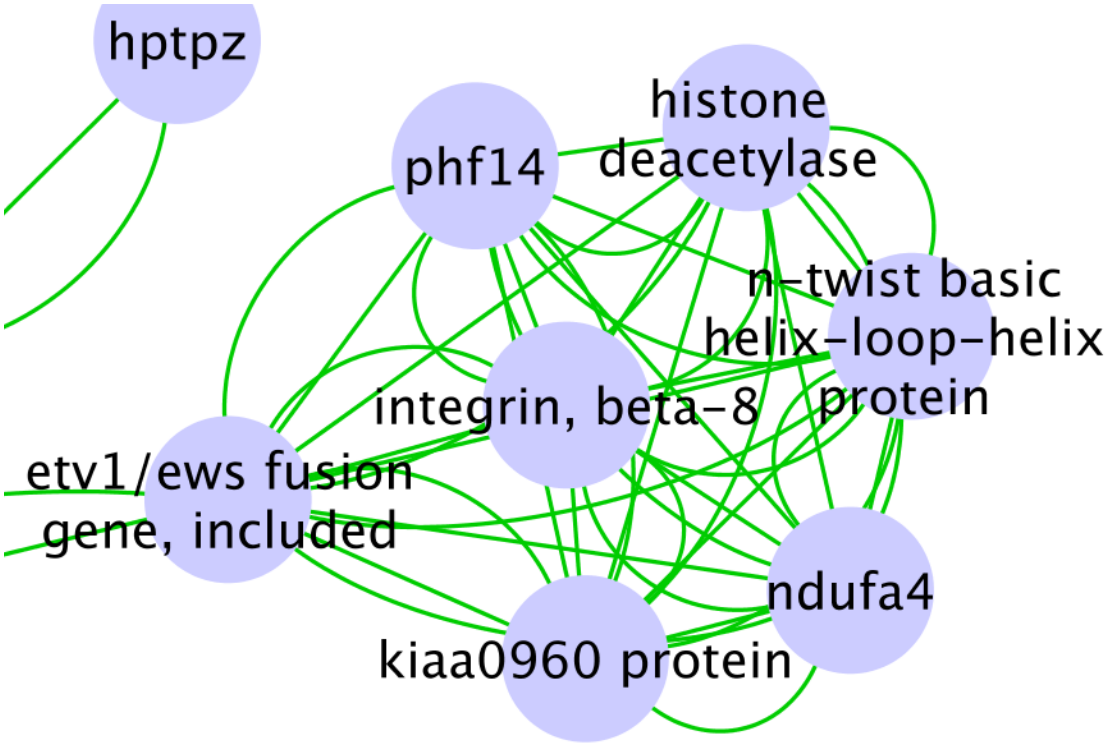


Graph created with Cytoscape [http://cytoscape.org/]

Relation Extraction

- Protein-protein interactions

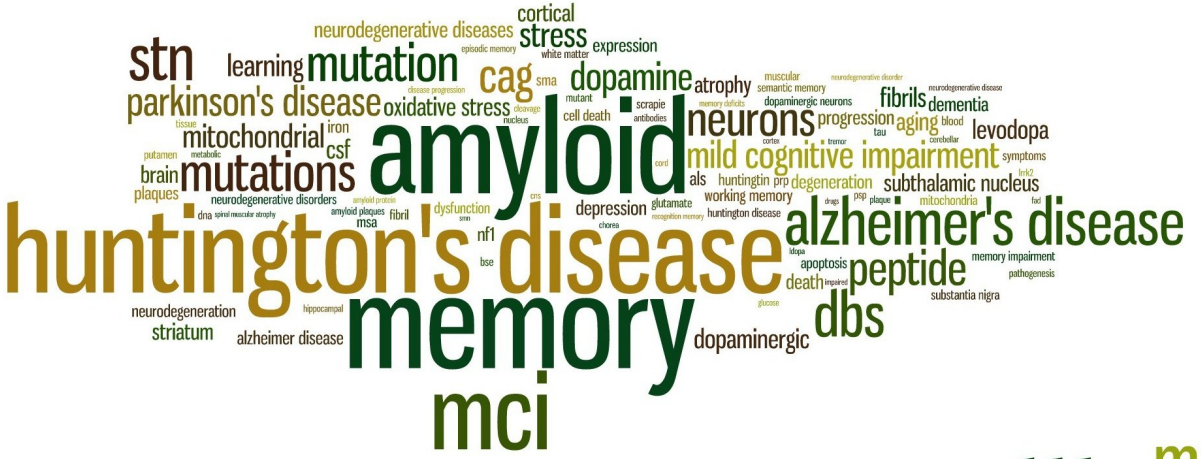


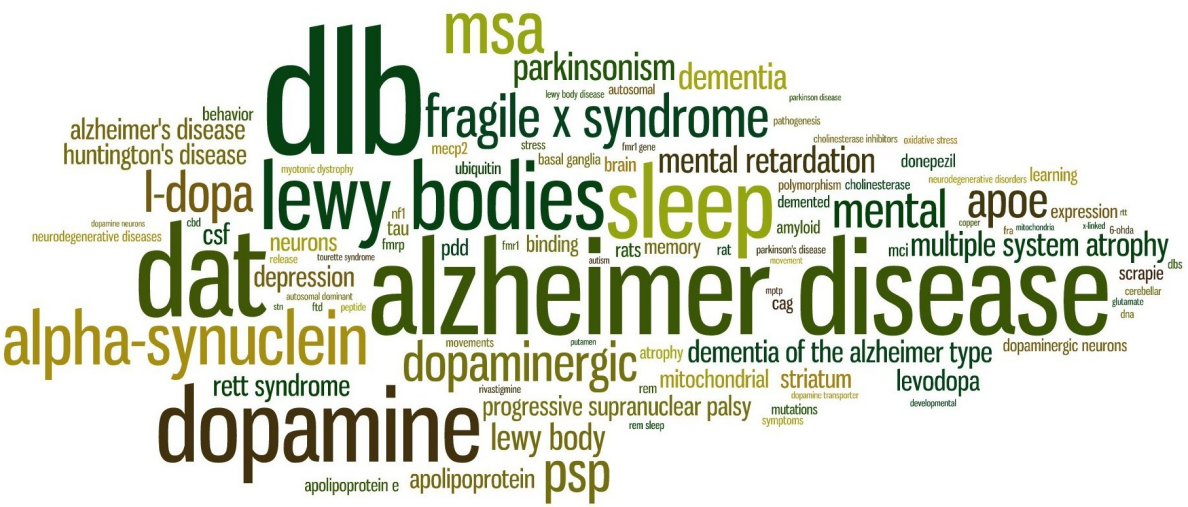


Graph created with Cytoscape [http://cytoscape.org/]

Latent Semantic Analysis

- Structuring literature search results





Competitions

- BioCreative III (2010)
  - Named entity recognition (genes/proteins)
  - PPI article classification and ranking
  - Chemical NER (F1 = 87.5%)*; interactive annotation*
- BioCreative III (2013)
  - IAT (Egas)
- SemEval 2014
  - Disease mentions in clinical text (P=81%, R=61%)
- BioCreative V IAT (2015)
  - BioC, ChemdNER and IAT tasks

Some current challenges

- Disambiguation / Normalization (concept recognition)
  - Gene vs. disease, CAT, …
- Relations mining
  - PPI, gene-drug relations, ..
- Event mining
  - gene events (e.g., expression, transcription, binding, regulation)
- Ranked retrieval
  - Besides quality (e.g. increases), rank quantity (e.g. how much?)
- EHR mining
  - Events, epidemiology
- Social media
  - Health monitoring, pharmacovigilance
- Multiple languages
  - EN, PT, DE, SP, ..
- Multimodal information
  - Images, figures, tables,..

Credits