

# An Introduction to Text Mining Research Papers

Petr Knoth  
Phil Gooch

Mendeley  
22 September 2015

# What is text mining?

‘the process of deriving high-quality information from text’ (*Wikipedia*)

‘the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking ... of the extracted information ... to form new facts or new hypotheses to be explored further’ (*Hearst, 2003*)

‘a burgeoning new field that attempts to glean meaningful information from natural language text ... the process of analyzing text to extract information that is useful for particular purposes’  
(*Witten, 2004*)

# Why text mine research papers?

*“Research papers are the most complete representation of human knowledge.”*

# Why is it so much talked about now?

The idea is not new, but up until recently access to large amounts of research papers was controlled by a handful of companies having bespoke arrangements with publishers.

The Open Access movement has recently largely contributed to decreasing the barriers to text-mining of research papers.

The availability of tools, developments in machine learning, and reduction in the costs of computing power and storage, has removed some of the technical and financial barriers.

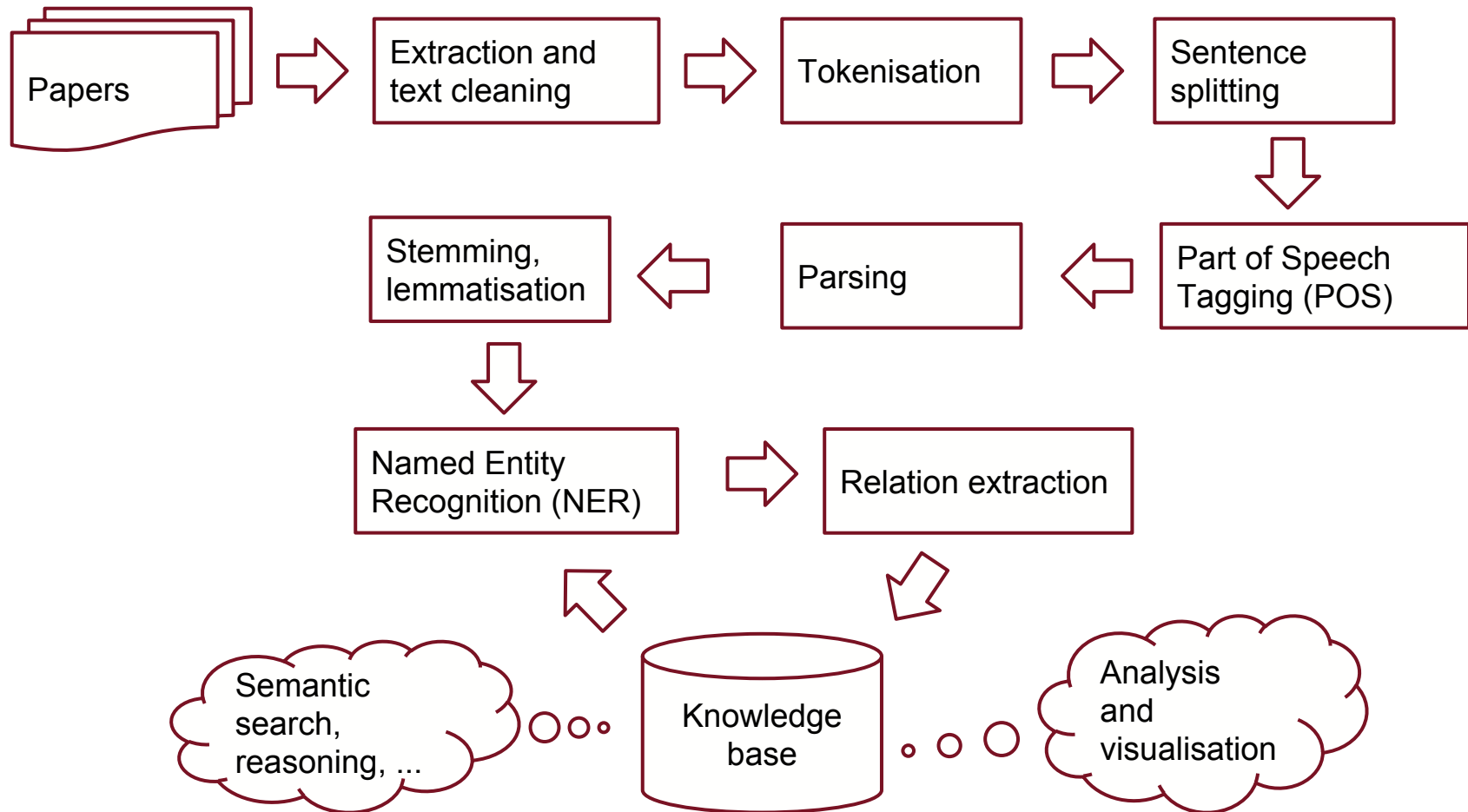
# What are the opportunities of text mining research papers?

- **Literature Based Discovery (LBD)**
  - Undiscovered Public Knowledge ([Swanson, 1986](#)).
  - Mining relationships for which there is “hidden” evidence in the research literature, yet they are not explicitly stated, such as *magnesium deficiency* and *migraine*, *fish oil* and *Raynaud's disease*.
  - Swanson's discoveries simulated by automated techniques ([Weeber et al., 2001](#)).

# Other use cases

- Supporting exploratory access to research literature
  - staying up-to-date with research
  - analysing, comparing and contrasting research findings
- Summarisation of research findings
- Systematic literature review automation
  - ‘snowballing’
- Question answering and semantic search from papers
- Understanding the research impact of articles, individuals, institutions, countries, ...
- Monitoring research trends
- Understanding how to direct research funding ...
- Evidence of reuse and plagiarism detection ...

# Generic text mining workflow



# Example 1: Literature Based Discovery

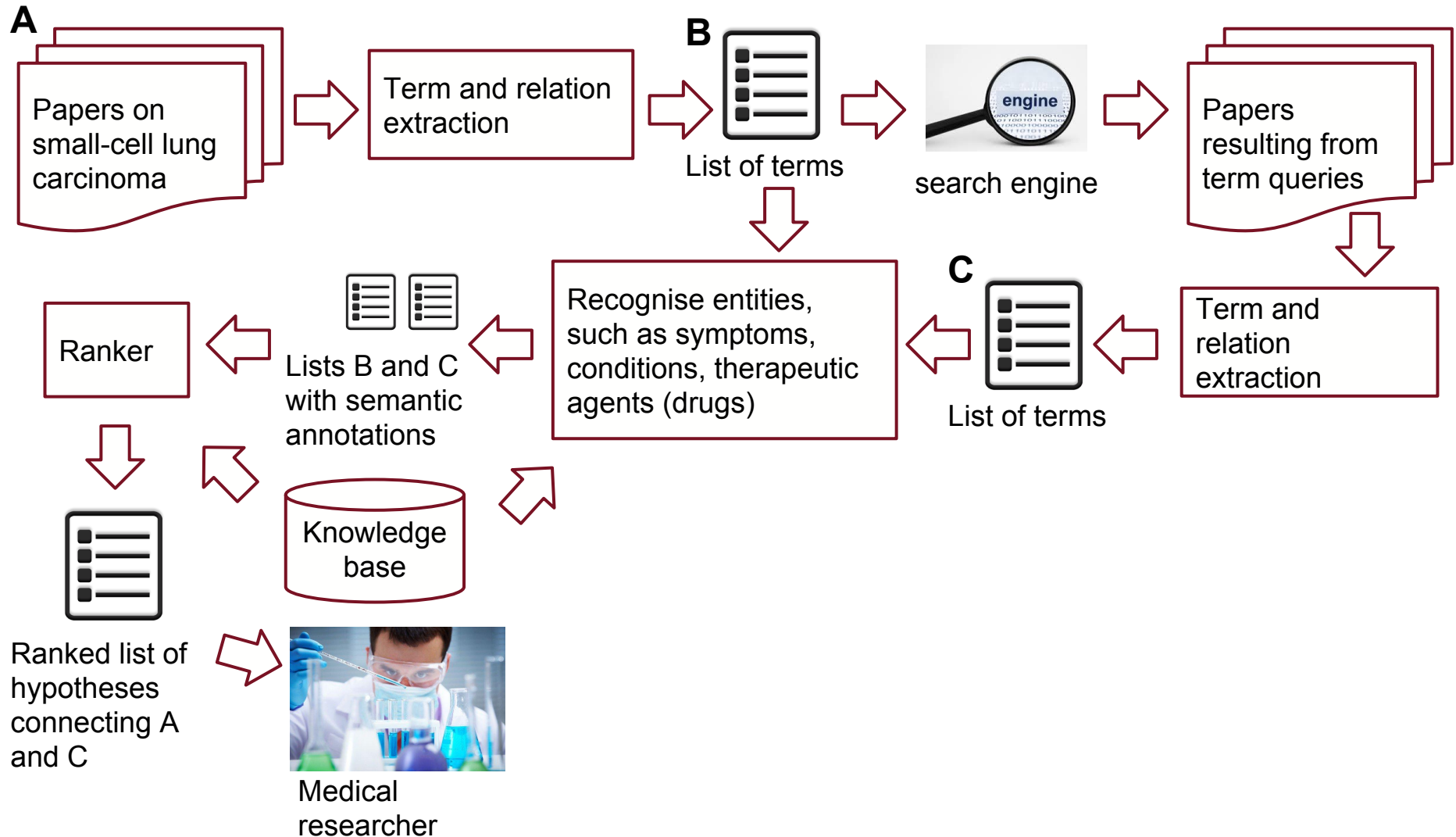
- A range of techniques ([Smalheiser, 2012](#))
- A typical approach: ABC method (e.g. [Hristovski et. al. 2008](#)):
  - **A** affects/binds/regulates/interacts with **B**
  - **B** affects/binds/regulates/interacts with **C**
  - **A** and **C** are not explicitly linked in any article

=> There might be an undiscovered relationship between **A** and **C**

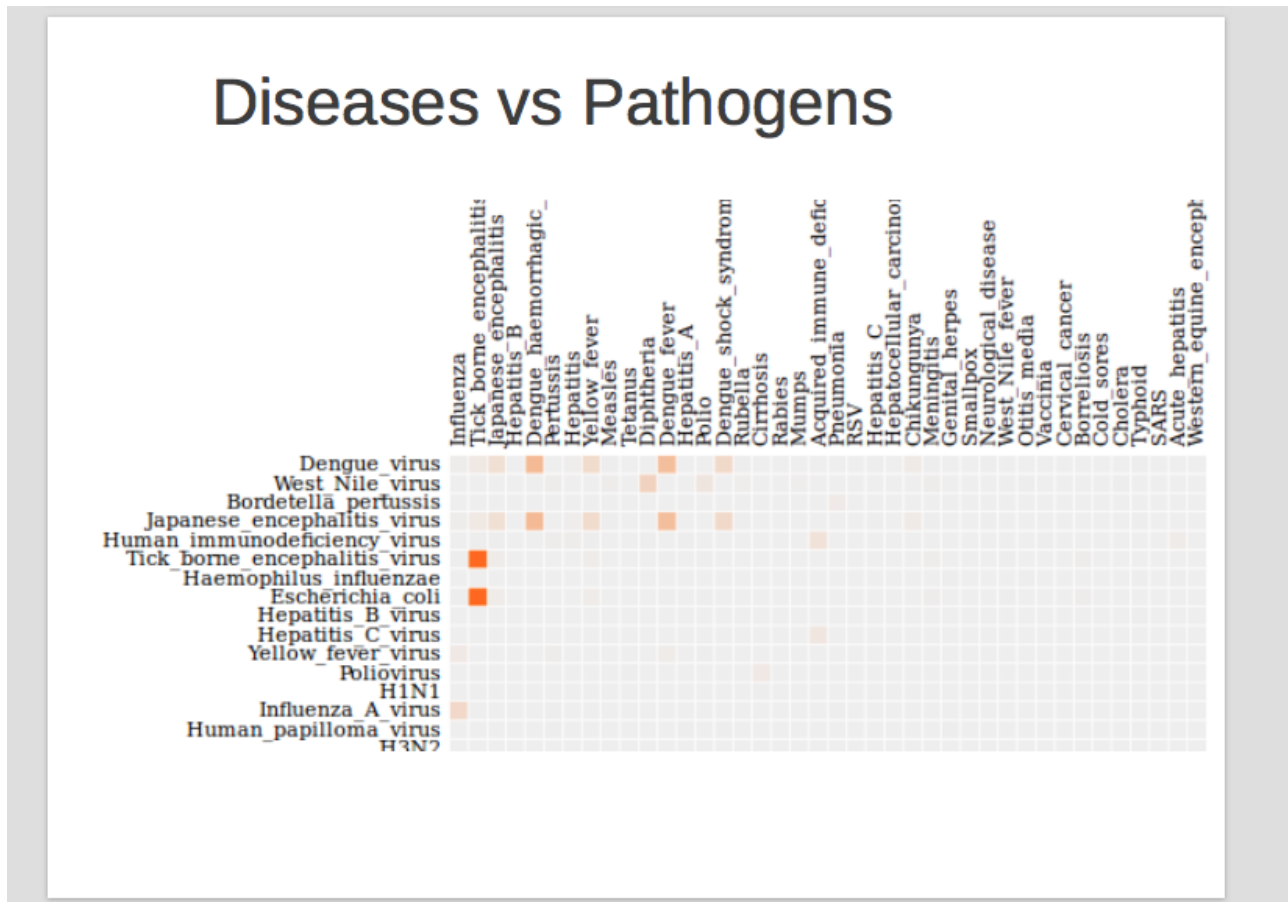
Various reasons why **A** and **C** might not be connected, e.g. B is a rare term.



# Example 1: Literature Based Discovery



# Example 1: Visualising the A-C hypotheses

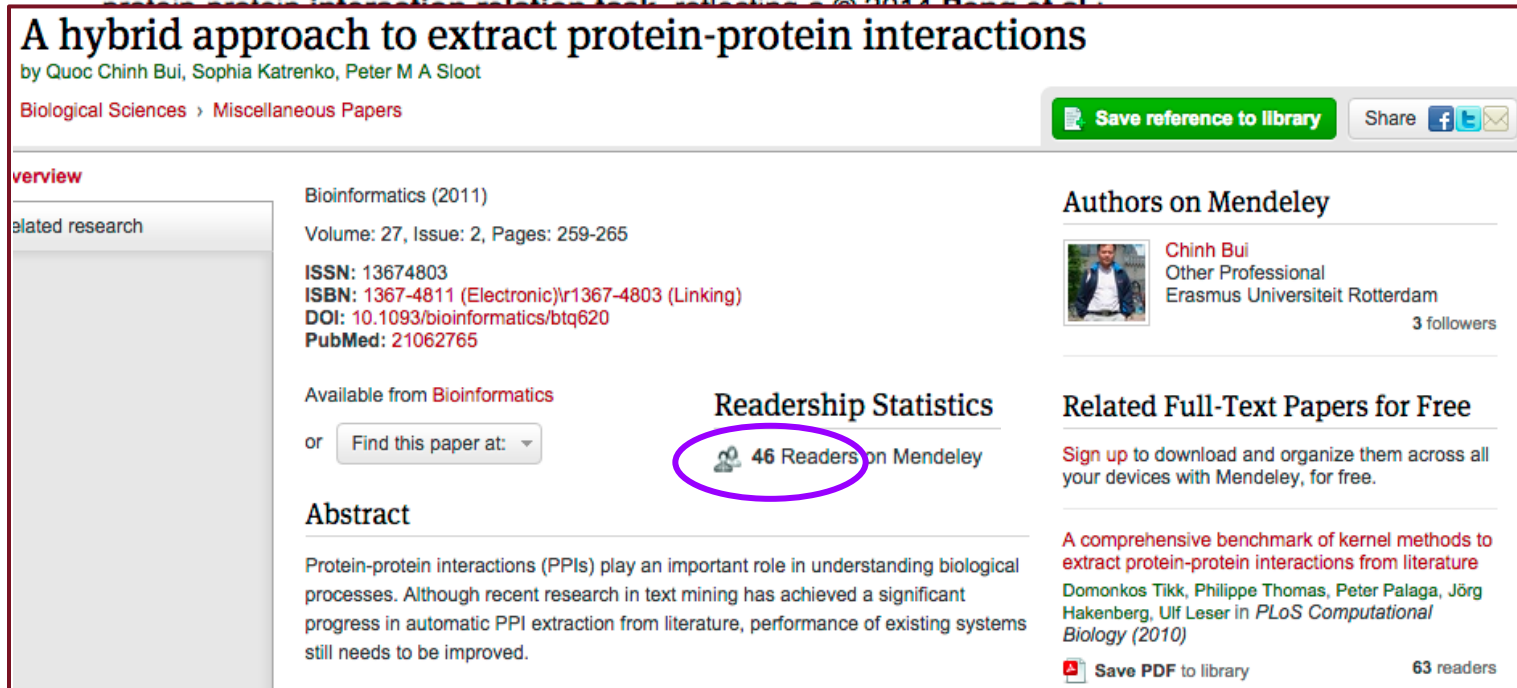


Reproduced with permission from [FastVac](#) and [Public Health England](#).

# Example 2: modeling research arguments using citation contexts

4. Bui QC, Katrenko S, Sloot PM: A hybrid approach to extract protein–protein interactions. *Bioinformatics* 2011, 27(2): 259–265. [[Mendeley](#)]

- **Section: Background** -- ... ities (e.g., proteins) reported in text. Approaches to the **relation extraction task** can be categorized into two major classes: (1) machine learningbased approaches and (2) pattern-based approaches. Machine learning-based approaches are data-driven and can derive models from a set of annotated data [\[\[1-7\]\]](#). The use of machine learning methods can be quite effective, but the performance of resulting systems depends on the quality and the amount of annotated data. For example, large annotated corpora become available for the



**A hybrid approach to extract protein-protein interactions**  
by Quoc Chinh Bui, Sophia Katrenko, Peter M A Sloot  
Biological Sciences > Miscellaneous Papers


[Save reference to library](#) [Share](#) [f](#) [t](#) [e](#)

**view**


related research

Bioinformatics (2011)  
Volume: 27, Issue: 2, Pages: 259-265  
ISSN: 13674803  
ISBN: 1367-4811 (Electronic) | 1367-4803 (Linking)  
DOI: 10.1093/bioinformatics/btq620  
PubMed: 21062765

Available from [Bioinformatics](#)  
or

**Readership Statistics**  
 **46 Readers** on Mendeley

**Abstract**  
Protein-protein interactions (PPIs) play an important role in understanding biological processes. Although recent research in text mining has achieved a significant progress in automatic PPI extraction from literature, performance of existing systems still needs to be improved.

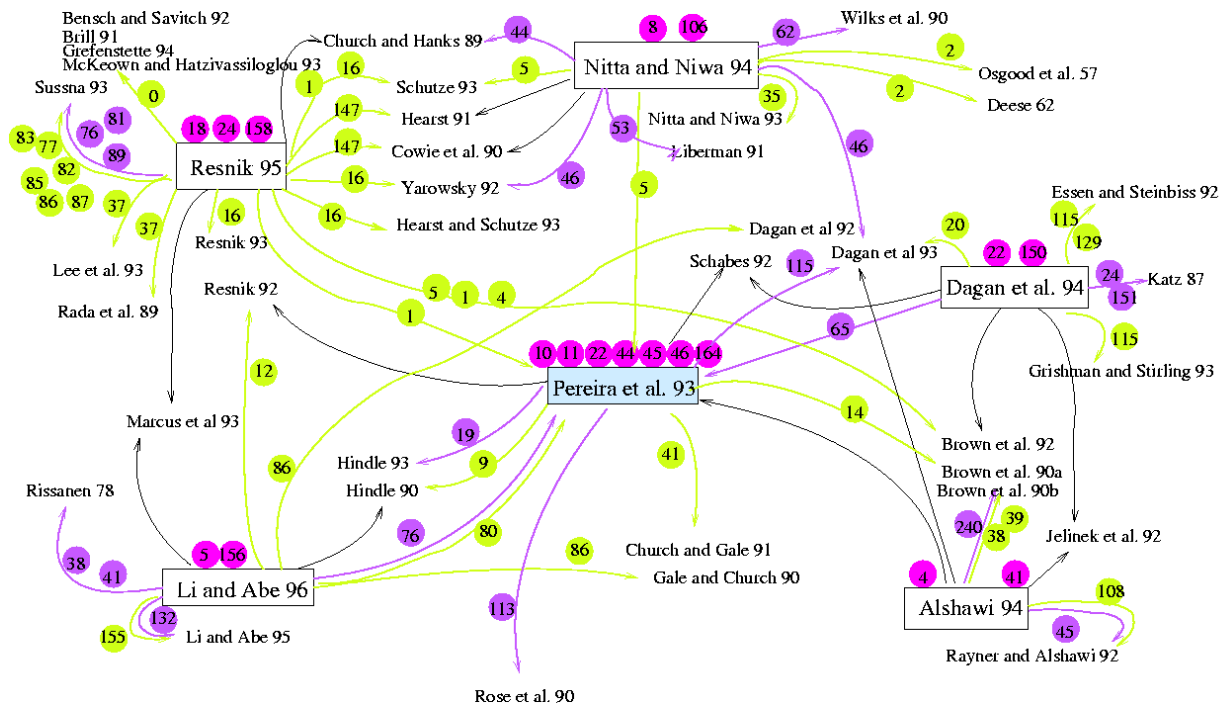
**Authors on Mendeley**  
 **Chinh Bui**  
Other Professional  
Erasmus Universiteit Rotterdam  
3 followers

**Related Full-Text Papers for Free**  
[Sign up](#) to download and organize them across all your devices with Mendeley, for free.

[A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature](#)  
Domonkos Tikk, Philippe Thomas, Peter Palaga, Jörg Hakenberg, Ulf Leser in *PLoS Computational Biology* (2010)  
[Save PDF to library](#) **63 readers**

# Example 2: modeling research arguments using citation contexts

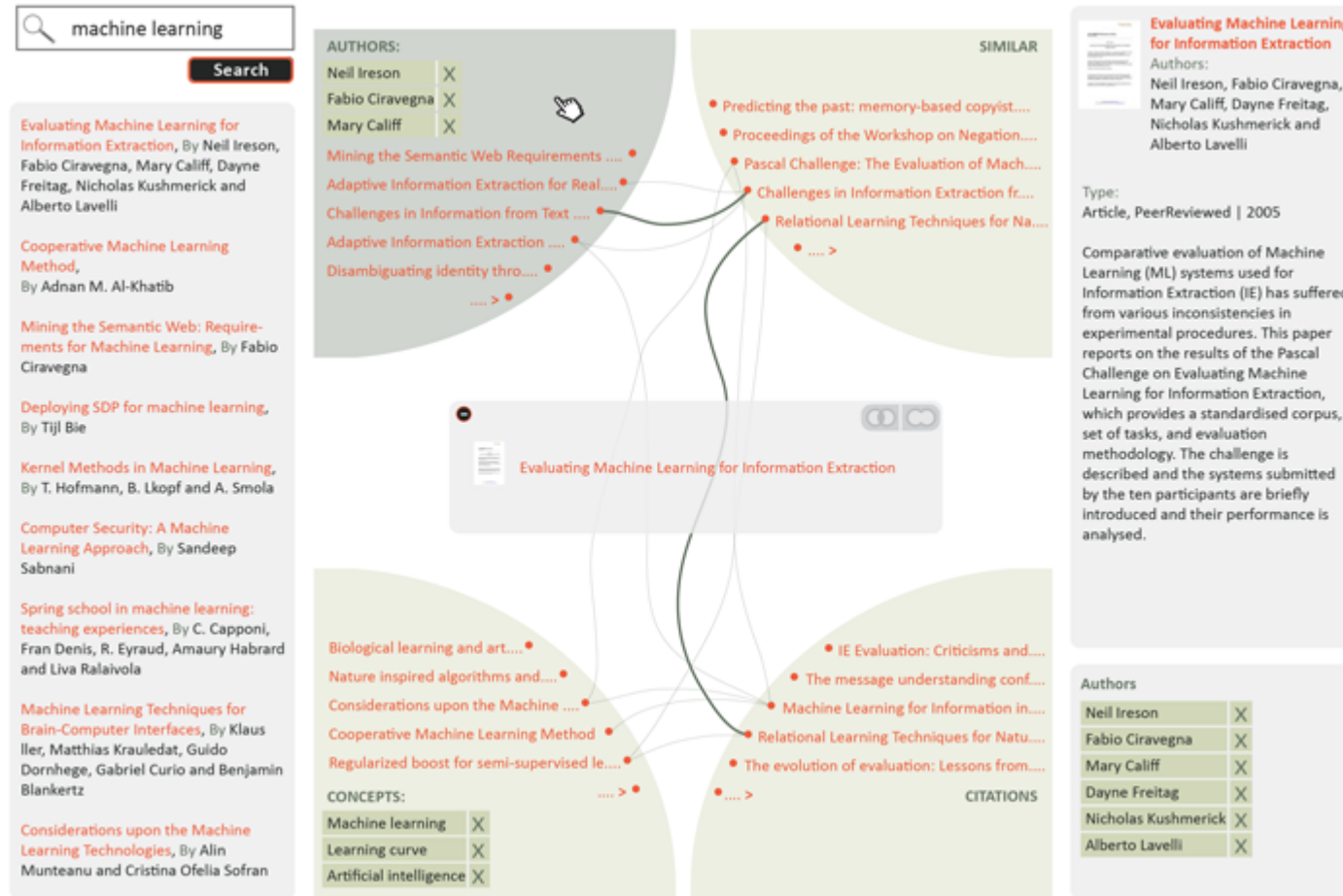
- Label the citation network according to the context of how they are cited
- Pioneering work by Simone Teufel in this area



Green: Contrastive or comparative statements  
 Pink: Statements about the aim of the paper

Source: Simone Teufel (1999) Argumentative Zoning: Information Extraction from Scientific Text, Phd Thesis, University of Edinburgh.

# Example 3: exploratory search over research papers



The screenshot shows a Mendeley search interface for the query "machine learning". The central area displays a network of related papers, with the selected paper "Evaluating Machine Learning for Information Extraction" highlighted in the center. The interface is divided into several panels:

- Search Bar:** Contains the query "machine learning" and a "Search" button.
- Left Panel (Search Results):** Lists several papers, including "Evaluating Machine Learning for Information Extraction", "Cooperative Machine Learning Method", "Mining the Semantic Web: Requirements for Machine Learning", "Deploying SDP for machine learning", "Kernel Methods in Machine Learning", "Computer Security: A Machine Learning Approach", "Spring school in machine learning: teaching experiences", and "Machine Learning Techniques for Brain-Computer Interfaces".
- Top Left Panel (AUTHORS):** Lists authors: Neil Ireson, Fabio Ciravegna, and Mary Califf, each with an 'X' mark.
- Top Right Panel (SIMILAR):** Lists similar papers: "Predicting the past: memory-based copyist...", "Proceedings of the Workshop on Negation...", "Pascal Challenge: The Evaluation of Mach...", "Challenges in Information Extraction fr...", and "Relational Learning Techniques for Na...".
- Bottom Left Panel (CONCEPTS):** Lists concepts: Machine learning, Learning curve, and Artificial intelligence, each with an 'X' mark.
- Bottom Right Panel (CITATIONS):** Lists citations: "IE Evaluation: Criticisms and...", "The message understanding conf...", "Machine Learning for Information in...", "Relational Learning Techniques for Natu...", and "The evolution of evaluation: Lessons from...".
- Right Panel (Paper Details):** Provides details for the selected paper: "Evaluating Machine Learning for Information Extraction", authors (Neil Ireson, Fabio Ciravegna, Mary Califf, Dayne Freitag, Nicholas Kushmerick and Alberto Lavelli), type (Article, PeerReviewed), and year (2005). It also includes a brief abstract.
- Bottom Right Panel (Authors):** Lists authors: Neil Ireson, Fabio Ciravegna, Mary Califf, Dayne Freitag, Nicholas Kushmerick, and Alberto Lavelli, each with an 'X' mark.

Source: [Herrmannova & Knoth \(2012\)](#) Visual Search for Supporting Content Exploration in Large Document Collections, DLib 18(8).

# How can I get started?

- Get the data
- Design/build your workflow
- Select a framework, tools, services to be used in implementing the workflow
- Understand how to evaluate the performance of each component

# Full-text Open Access article sources

- Subject repositories:

- [arXiv bulk data](#)

 arXiv.org

- [PubMed OA Subset](#)

 PubMed

- Institutional repositories:

- ~3k across the world, see [Directory of Open Access Repositories](#)

 OpenDOAR

- Open Access journals:

- >10k OA journals, see [Directory of Open Access Journals \(DOAJ\)](#)

 DOAJ  
DIRECTORY OF  
OPEN ACCESS  
JOURNALS

- Open Access subsets from publishers:

- [Elsevier OA STM Corpus](#)

 ELSEVIER

- [SpringerOpen](#)

 SpringerOpen

- Aggregators:

- [CORE \(API, Data dumps\)](#)

 CORE

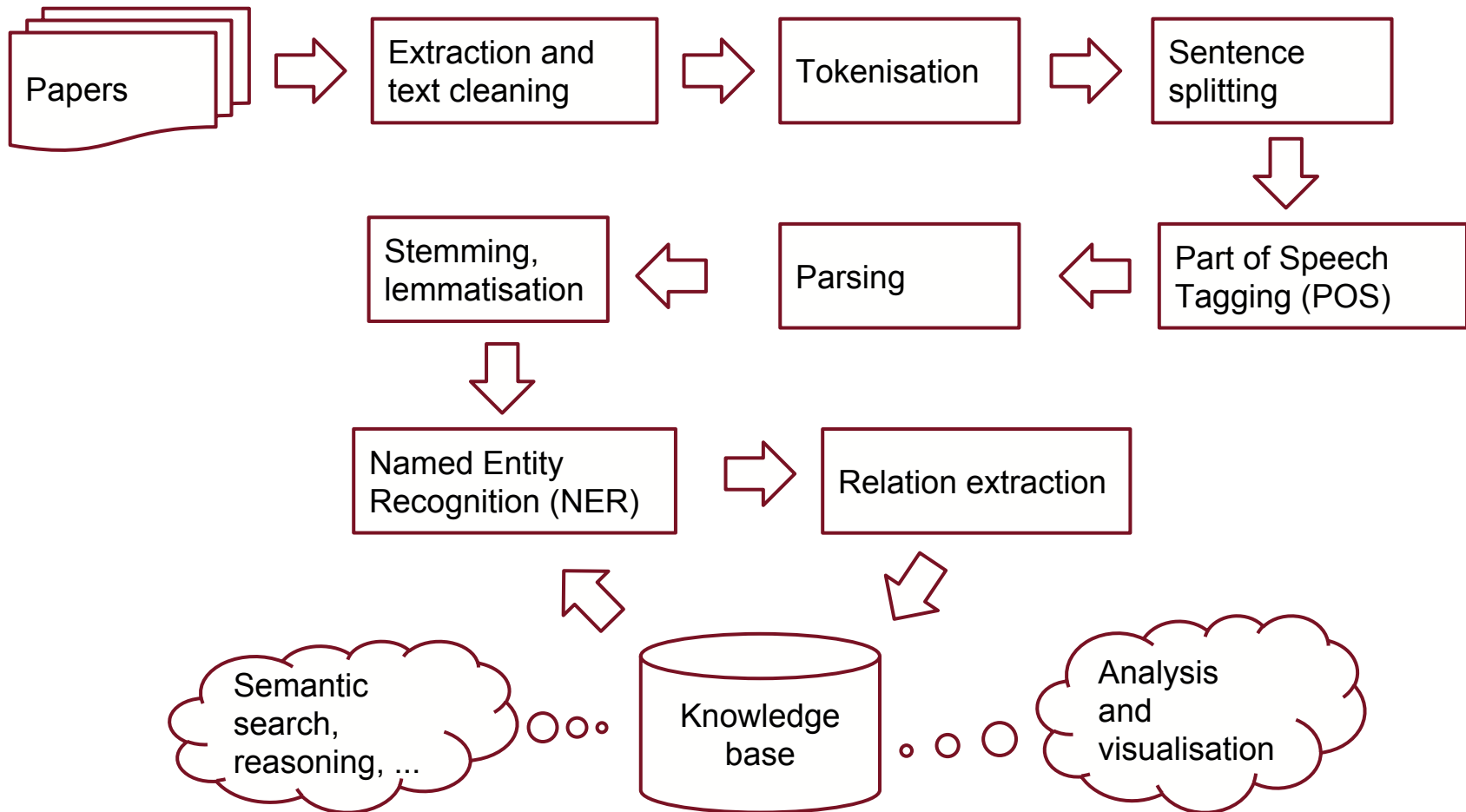
# Other full-text sources

Typically for non-commercial research only.

- Publisher full-text APIs:
  - [Elsevier's Text and Data Mining API](#):
  - (subscribed content on ScienceDirect)
- Domain specific corpuses:
  - [JSTOR Data For Research](#)
- Full-text content from multiple publishers:
  - [CrossRef Text and Data Mining API](#) (full-texts)
  - [Mendeley API](#) (abstracts)
- Other useful scholarly resources:
  - [Microsoft Academic Graph \(MAG\)](#)
- Clinical records
  - [i2b2 NLP Challenge Data Sets](#)



# Building a text mining application



# Example: Part of speech tagging

[Acute]<sub>JJ</sub> [lymphoblastic]<sub>JJ</sub> [leukemia]<sub>NN</sub> ([ALL]<sub>NN</sub>) [leads]<sub>VB</sub> [to]<sub>IN</sub> [an]<sub>DT</sub>  
[accumulation]<sub>NN</sub> [of]<sub>IN</sub> [immature]<sub>JJ</sub> [lymphoid]<sub>JJ</sub> [cells]<sub>NN</sub> [into]<sub>IN</sub> [the]<sub>DT</sub> [bone]<sub>NN</sub>  
[marrow]<sub>NN</sub>, [blood]<sub>NN</sub> [and]<sub>CC</sub> [other]<sub>JJ</sub> [organs]<sub>NN</sub>.

[A]<sub>DT</sub> [young]<sub>JJ</sub> [patient]<sub>NN</sub> [was]<sub>VB</sub> [treated]<sub>VB</sub> [unconventionally]<sub>RB</sub> [for]<sub>IN</sub>  
[Philadelphia]<sub>NN</sub> [positive]<sub>NN</sub> [ALL]<sub>NN</sub> [and]<sub>CC</sub> [Mucormycosis]<sub>NN</sub> [with]<sub>IN</sub>  
[Amphotericin]<sub>NN</sub> [B]<sub>NN</sub>.

[The]<sub>DT</sub> [patient]<sub>NN</sub> [achieved]<sub>VB</sub> [a]<sub>DT</sub> [disease-free]<sub>JJ</sub> [survival]<sub>NN</sub> [of]<sub>IN</sub> [12]<sub>CD</sub>  
[months]<sub>NN</sub> [with]<sub>IN</sub> [good]<sub>JJ</sub> [quality]<sub>NN</sub> [of]<sub>IN</sub> [life]<sub>NN</sub>.

# Example: Parsing

[Acute lymphoblastic leukemia]<sub>NP</sub> ([ALL]<sub>NP</sub>) [leads to]<sub>VP</sub> [an]<sub>DT</sub> [accumulation]<sub>NN</sub>  
[of]<sub>IN</sub> [immature lymphoid cells]<sub>NP</sub> [into]<sub>IN</sub> [the]<sub>DT</sub> [bone marrow]<sub>NP</sub>, [blood]<sub>NP</sub>  
[and]<sub>CC</sub> [other organs]<sub>NP</sub>.

[A]<sub>DT</sub> [young patient]<sub>NP</sub> [was treated unconventionally for]<sub>VP</sub> [Philadelphia  
positive ALL]<sub>NP</sub> [and]<sub>CC</sub> [Mucormycosis]<sub>NP</sub> [with]<sub>IN</sub> [Amphotericin B]<sub>NP</sub>.

[The]<sub>DT</sub> [patient]<sub>NP</sub> [achieved]<sub>VP</sub> [a]<sub>DT</sub> [disease-free survival]<sub>NP</sub> [of]<sub>IN</sub> [12 months]<sub>NP</sub>  
[with]<sub>IN</sub> [good]<sub>JJ</sub> [quality of life]<sub>PP</sub>.

# Example: Named entity recognition

[**Acute lymphoblastic leukemia**]<sub>Disease</sub> ([**ALL**]<sub>Disease</sub>) [**RESULT**]<sub>VP</sub> [accumulation]<sub>NN</sub> [of]<sub>IN</sub> [immature lymphoid cells]<sub>AnatomicalSite</sub> [into]<sub>IN</sub> [the]<sub>DT</sub> [bone marrow]<sub>AnatomicalSite</sub> [blood]<sub>AnatomicalSite</sub> [and]<sub>CC</sub> [other organs]<sub>AnatomicalSite</sub>

[A]<sub>DT</sub> [young patient]<sub>Person</sub> [**TREAT**]<sub>VP</sub> [**Philadelphia positive ALL**]<sub>Disease</sub> [and]<sub>CC</sub> [**Mucormycosis**]<sub>Disease</sub> [with]<sub>IN</sub> [**Amphotericin B**]<sub>Treatment</sub>

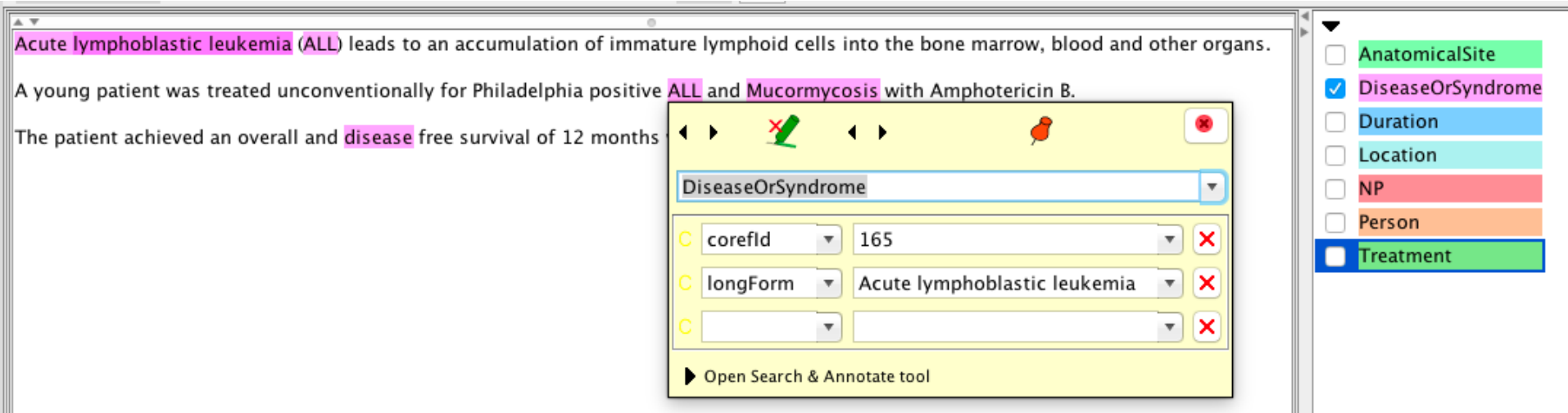
[The]<sub>DT</sub> [patient]<sub>Person</sub> [**RESULT**]<sub>VP</sub> [disease-free survival]<sub>Outcome</sub> [of]<sub>IN</sub> [12 months]<sub>Duration</sub> [with]<sub>IN</sub> [good quality of life]<sub>Outcome</sub>

# Example: Named entity recognition

Acute lymphoblastic leukemia (ALL) leads to an accumulation of immature lymphoid cells into the bone marrow, blood and other organs.

A young patient was treated unconventionally for Philadelphia positive ALL and Mucormycosis with Amphotericin B.

The patient achieved an overall and disease free survival of 12 months

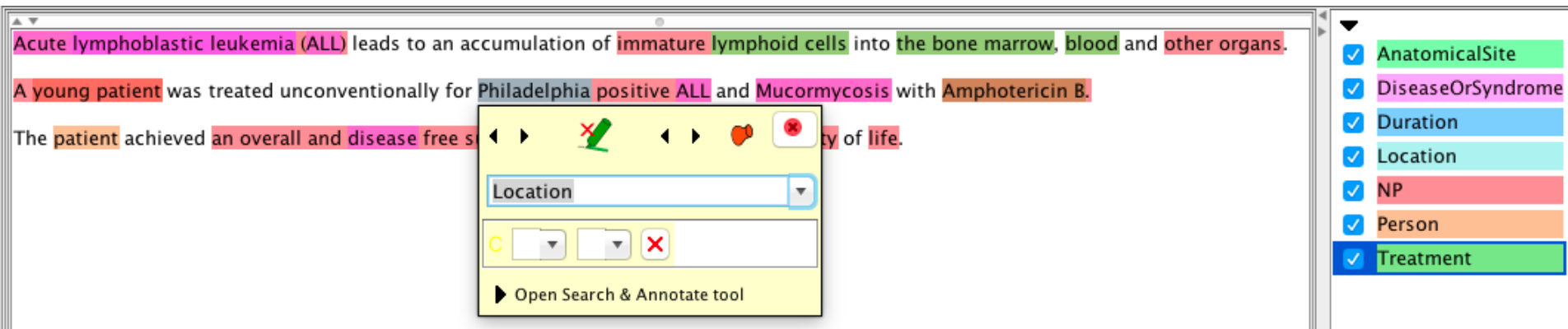


Entity types selected in the sidebar:

- AnatomicalSite
- DiseaseOrSyndrome
- Duration
- Location
- NP
- Person
- Treatment

- Here, *ALL* is automatically expanded to its full form: *acute lymphoblastic leukemia*
- *ALL* is automatically labelled as a DiseaseOrSyndrome. as the initial mention was also labelled as DiseaseOrSyndrome

# Example: Named entity recognition



Acute lymphoblastic leukemia (ALL) leads to an accumulation of immature lymphoid cells into the bone marrow, blood and other organs. A young patient was treated unconventionally for Philadelphia positive ALL and Mucormycosis with Amphotericin B. The patient achieved an overall and disease free survival rate of 50%.

Location

- AnatomicalSite
- DiseaseOrSyndrome
- Duration
- Location
- NP
- Person
- Treatment

- *Philadelphia* on its own can be labelled as a **Location**
- In this context it is part of a longer noun phrase 'Philadelphia positive ALL', of which *ALL* is the Disease *acute lymphoblastic leukemia*
- '*Philadelphia positive*' is also a term in a domain-specific dictionary
- Which label is chosen can be determined by a rule, either handwritten or learnt by a machine learning algorithm

# Frameworks

- General Architecture for Text Engineering (GATE)
- Apache UIMA
- Natural Language Toolkit (NLTK)
- OpenNLP

# Tools and Services

- [Open Calais](#): general purpose tagging of people, places, companies, facts, events and relationships
- [AlchemyAPI](#): similar to OpenCalais
- [National Centre for Text Mining](#): terms, entities, semantic search
  - [Cafetiere](#): text mine your own documents online
  - [EventMine](#): relations between terms: protein binding, synthesis inhibition
- [ContentMine](#): Fact extraction
- [ReVerb](#): open-domain relation extraction
- [XIP Parser](#): linguistic analysis
- [Linguamatics](#)
- Metadata and citation extraction:
  - [ParsCit](#)
  - [Grobid](#)
  - [Cermin](#)



# Open source demos

- Mimir
  - analysing, comparing and contrasting research findings
- EEXCESS
  - Recommend related resources
  - Narrative paths between resources

# Existing challenges

- Harmonising metadata formats across publishers, repositories, etc.
- Agreeing on standards/ontologies/formats used to share the outputs of research publications text-mining tools and all their components.
- Integration of text-mining tools with content providers' systems
- Building and maintaining text-mining web-services for research (building blocks)
- Promoting and adopting end-user tools utilising text-mining in researchers' daily workflows
- Building gold standards for various text-mining tasks and sharing them across researchers (issue of credit)

# Events and initiatives

- Events
  - [International Workshop on Mining Scientific Publications \(WOSP\)](#)
  - Joint Conference on Digital Libraries (JCDL)
  - ACL, COLING, IJC-NLP, CoNLL, LREC, etc.
- Projects
  - [OpenMinTeD](#) (EC funded)
  - [EEXCESS](#)
  - [OpenAIRE](#)



Thanks for listening ...

[datascience@mendeley.com](mailto:datascience@mendeley.com)