



Text Mining in der Biomedizin

Forschung am Lehrstuhl für Computerlinguistik in Jena

Ekaterina Buyko

Friedrich-Schiller-Universität Jena

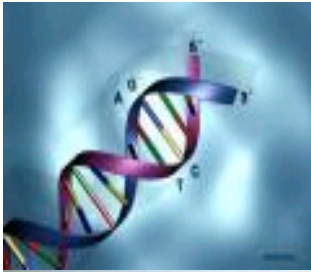




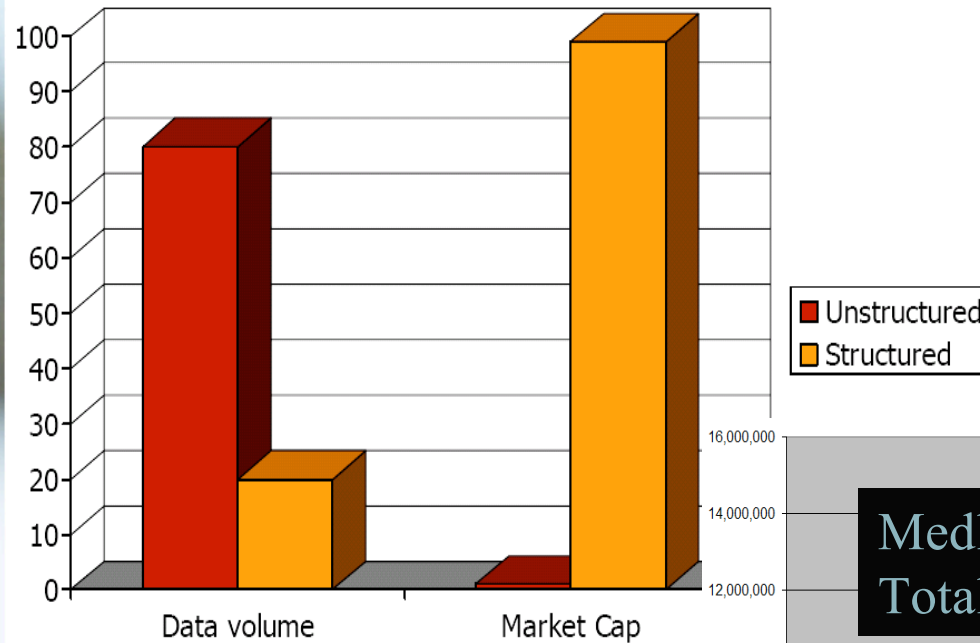
Warum NLP für Life Sciences?

- Der Informationszugang wird erschwert durch:
 - Biomedizinisches Wissen liegt zumeist in unstrukturierter Form in Freitexten
 - Riesige und schnell wachsende Textmengen
 - Geringe Vernetzung von Texten mit strukturierten Wissensquellen (DBs)
- Der Wettbewerbsvorteil bedeutet:
 - Redundante Forschung vermeiden
 - Kosten reduzieren
 - Schneller neue Informationen in die Forschung integrieren

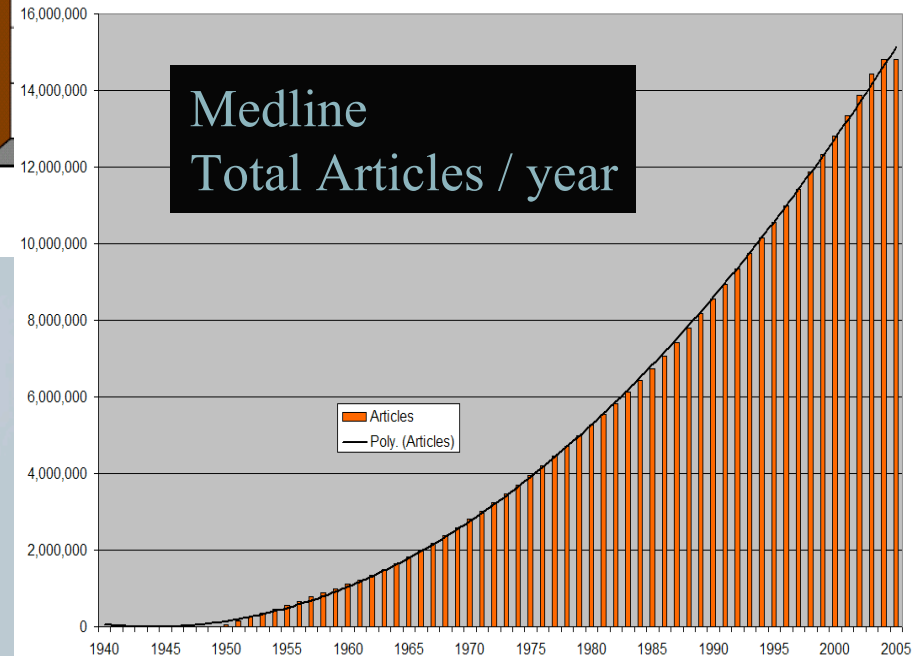




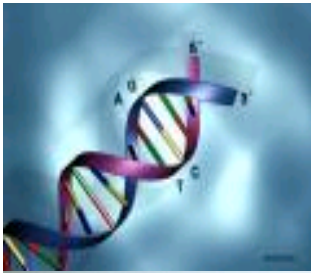
Warum NLP für Life Sciences?



Medline
Total Articles / year

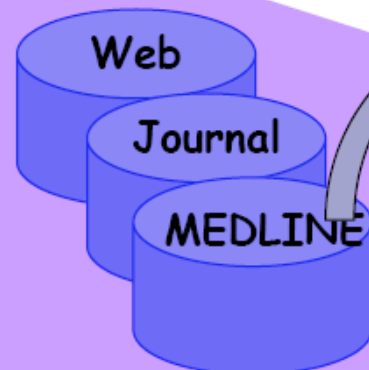


Quelle: Prabhakar, Raghavan,
Verity (2002)

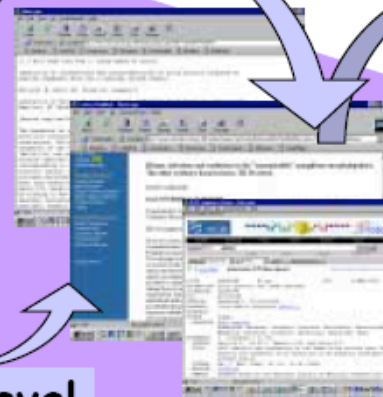


Von IR zu Text Mining

**Collections:
Gigabytes**



**Documents:
Megabytes**



**Question Answering:
question to answer**

**Lists, Tables:
Kilobytes**

**Phrases:
Bytes**

**Protease-resistant
prion protein
interacts with...**

**Information Retrieval
and Classification:
key words to
document classes**

**Information Extraction:
documents to entities, relations**



MITRE



Ausschnitt aus PubMed Abstract

... Electrophoretic mobility shift assays indicate that MS-2beta and MS-2gamma bind to nuclear factors that are induced during U937 differentiation. ...





Entity Recognition

... Electrophoretic mobility shift assays indicate that **MS-2beta** and **MS-2gamma** bind to **nuclear factors** that are induced during U937 differentiation. ...





Relation Recognition

... Electrophoretic mobility shift assays indicate that **MS-2beta** and **MS-2gamma** **bind to nuclear factors** that are induced during U937 differentiation. ...

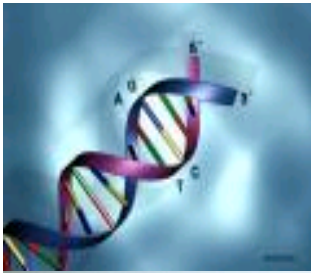




Relation Extraction

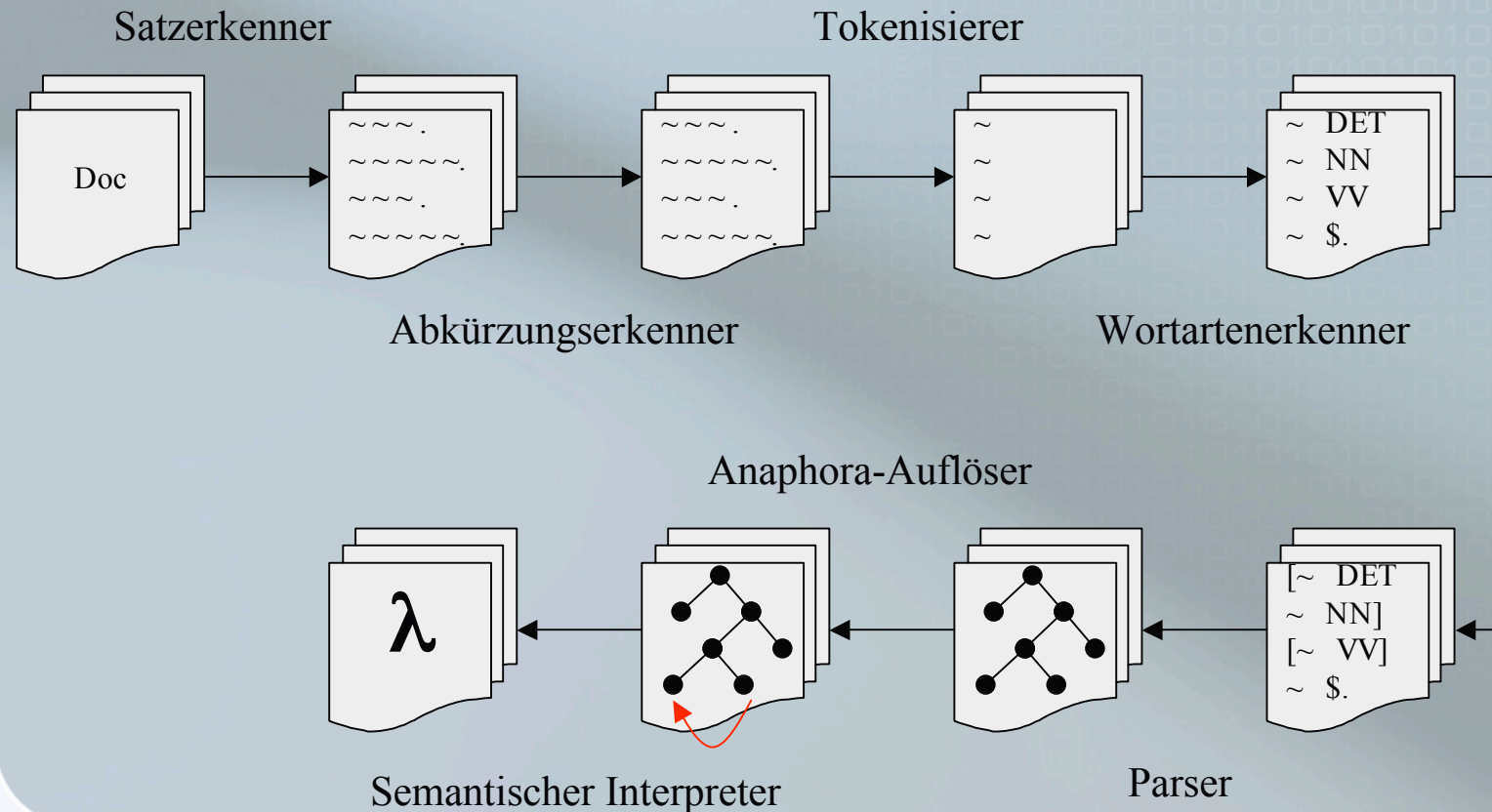
... Electrophoretic mobility shift assays indicate that **MS-2beta** and **MS-2gamma** **bind to nuclear factors** that are induced during U937 differentiation. ...

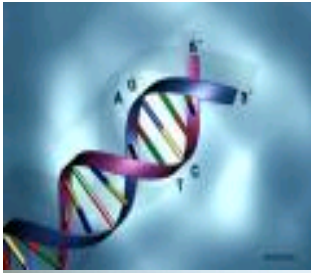
[**bind to** , **MS-2beta** & **MS-2gamma** , **nuclear factors**]



NLP Pipeline für Text Mining

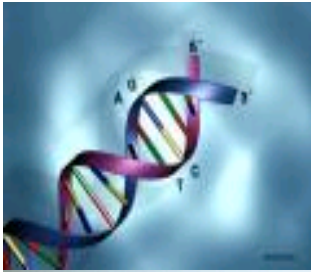
Ablauf einer Pipeline





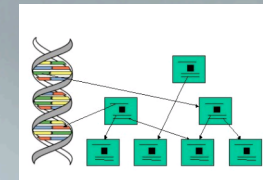
Aktueller Forschungsstand

	Zeitung (man)	Zeitung (auto)	Biomed (man)	Biomed (auto)
Wortarten	99	98-99	97-98	97
Entity	95-97	90-95	84-89	75-85
Anaphora	95	89	90-95	86
IE	97	75-80	70-90	55-75



Evaluation

- Wettbewerbe in Text Mining in der Biologie
 - BioCreative (seit 2003)
 - 2003-2004 (27 Gruppen aus 10 Ländern)
 - Genomics TREC (seit 2003)
 - 2005 (32 Gruppen)
 - LLLChallenge (seit 2005)

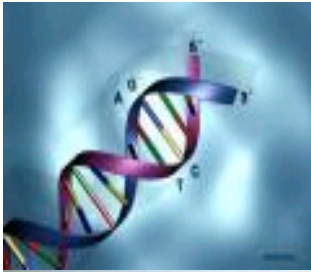




Organisationen

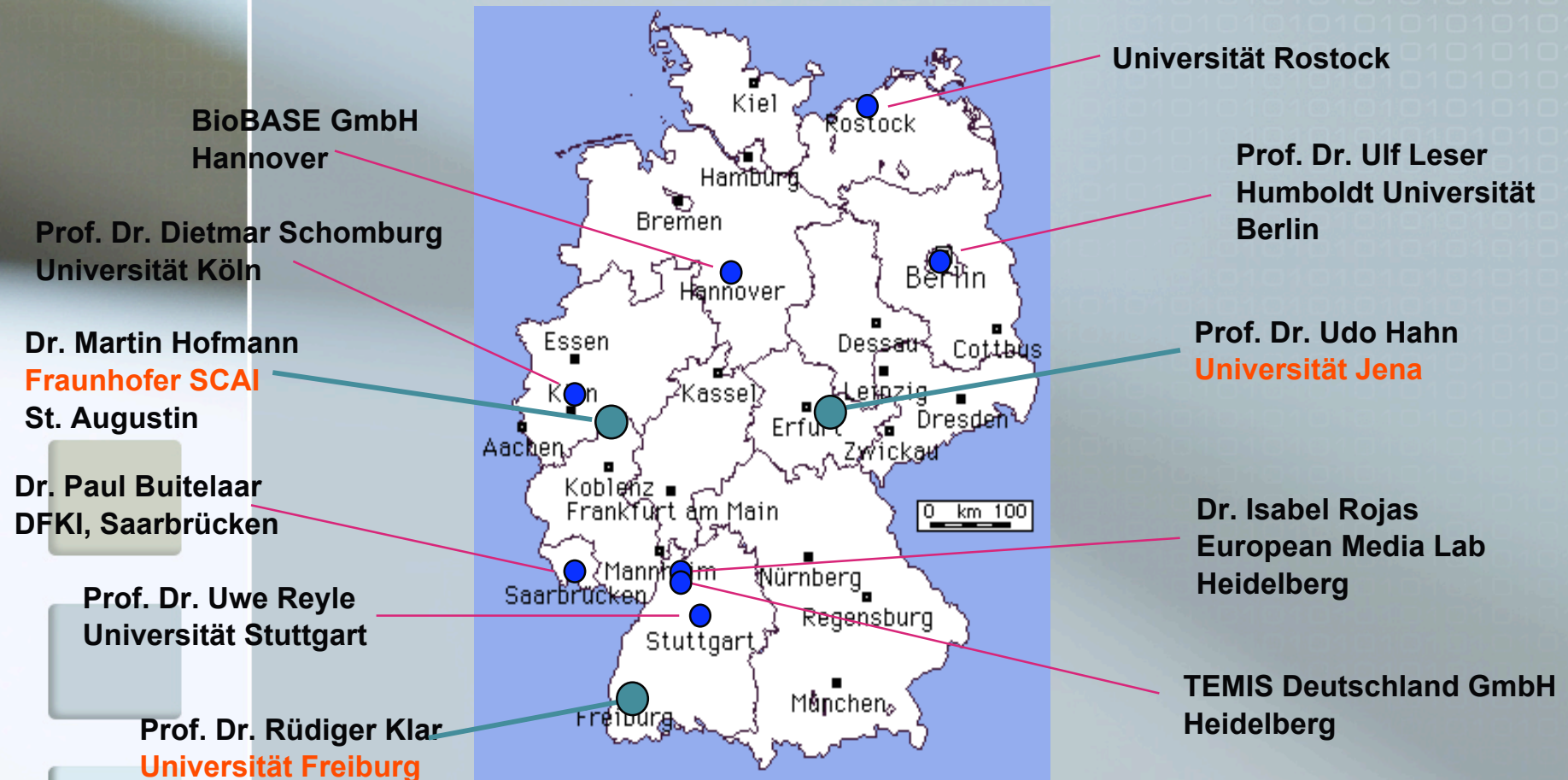
- National Library of Medicine (NLM USA)
- Tsujii Laboratory (Japan)
- National Centre for Text Mining (NaCTeM Großbritannien)
- BioTem (Deutschland)





BioTem-Initiative

<http://www.biotem.de>

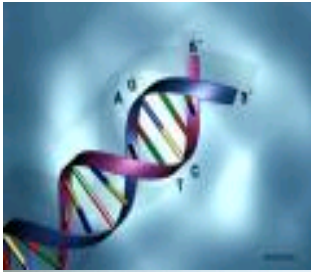




Universität Jena, Julie Lab

- Inhaber Prof. Dr. Udo Hahn (seit 2004)
- 7 Mitarbeiter, 10 Hiwis
- In der Lehre Kooperation mit Bioinformatik
- Forschung in Text Mining in der Biomedizin



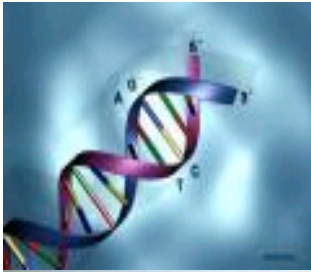


BootStrep Projekt (EU)



- Ziele
 - Akquisition von dynamischen Ressourcen für Text Mining in der Biologie
- Beteiligte Gruppen
 - Friedrich-Schiller Universität Jena, Computerlinguistik (Deutschland, Koordinator)
 - NacTeM (Großbritannien)
 - EMBL-EBI (Großbritannien)
 - Institute for Infocomm Research (Singapur)
 - Université de Rennes (Frankreich)
 - Istituto di Linguistica Computazionale (Italien)
 - Universitätsklinikum Freiburg (Deutschland)
- <http://www.bootstrep.eu>





StemNet Projekt (BMBF)

■ Konkreter Anwendungsfall

- Text-Mining: Verbesserung der Verträglichkeitsprüfung zwischen Empfänger und Spender bei Transplantation von Blut-Stammzellen (z.B. bei Leukämie-Patienten)
- Integration von Text-Mining- und Data-Mining-Methoden

■ Beteiligte Gruppen

- Clarity AG (Bad Homburg)
- Institut für Transfusionsmedizin (Medizinhochschule Hannover)





Offene Stellen

■ Profil

- Kern-NLP, Statistikkenntnisse
- Implementation (Java)
- Erfahrung im Maschinellen Lernen

■ Aufgaben

- Information Extraktion
- Text Mining





Kontakt

- Prof. Dr. Udo Hahn

hahn@coling.uni-jena.de

<http://www.coling.uni-jena.de>

Friedrich-Schiller-Universität Jena

Institut für Germanistische
Sprachwissenschaft

Fürstengraben 30

07743 Jena





Text Mining in der Biomedizin

Forschung am Lehrstuhl für Computerlinguistik in Jena

Ekaterina Buyko

buyko@coling-uni-jena.de

