Chapter 1. Introduction

Jingbo Xia
College of Informatics, HZAU

Outline

- Brief intro of BioNLP
- What make BioNLP unique, if compared with general text mining
- Main research issues
- Timetable for this term

Outline

- Brief intro of BioNLP
- What make BioNLP unique, if compared with general text mining
- Main research issues
- Timetable for this term

Definitions:

- Text mining -> Natural language process /Computational linguistics

- Biomedical Language Process (BioNLP), also known as Biomedical text mining

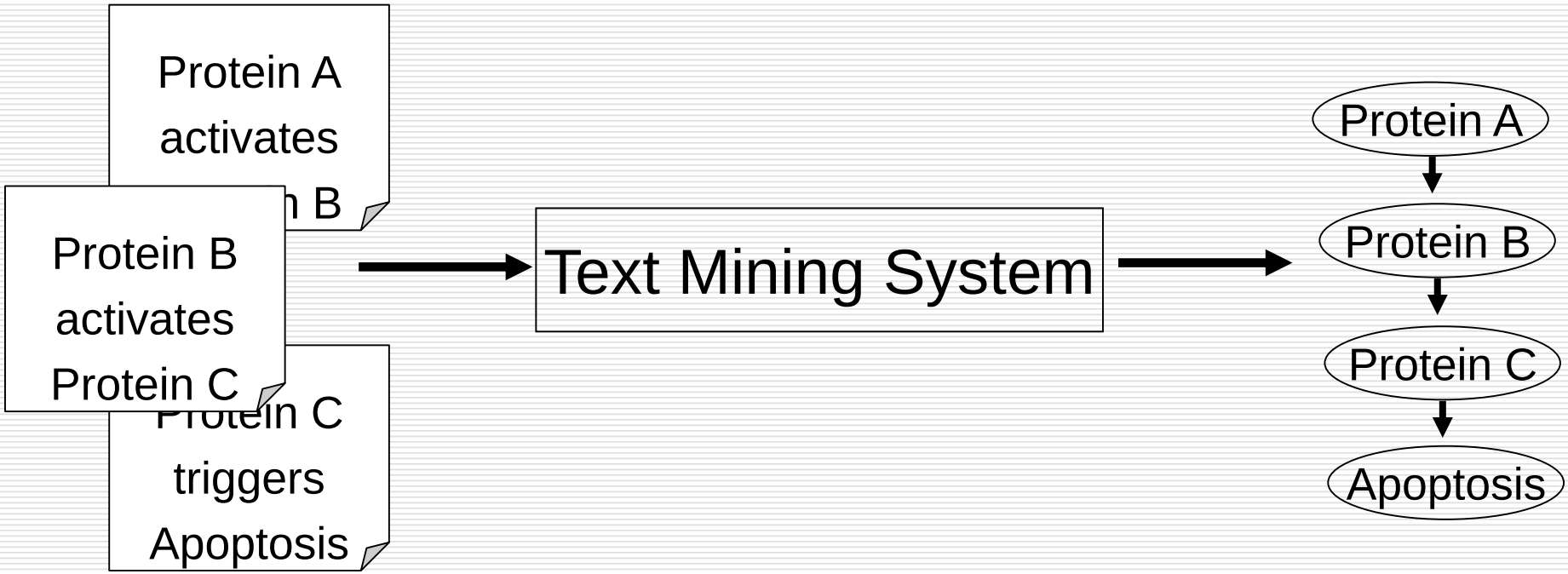
Text mining

- **Text mining** is
  - the process of *automatically extracting knowledge* from *large text collections*
  - data mining applied to text documents / knowledge discovery from text
  - a modular process similar to *reading*, where facts from different articles / books are combined for novel inference (de Bruijn 2002)

BioNLP

- **BioNLP** refers to text mining applied to texts and literature of the biomedical and molecular biology domain. It is a rather recent research field on the edge of natural language processing, bioinformatics,medical informatics and computational linguistics.

Examples in BioNLP



"BioNLP, as a newly developed cross-disciplinary research method, belongs to the scope of systematic biology, and it aims to supply systematical knowledge discovery upon unique bio-medical issues."

Outline

- Brief intro of BioNLP
- **What make BioNLP unique, if compared with general text mining**
- Main research issues
- Timetable for this term

What make BioNLP unique

- Biomedicine researches have and will being produce unique and enormous text data
- Cross-disciplinary field demands various integrative knowledge including:

  Data mining,

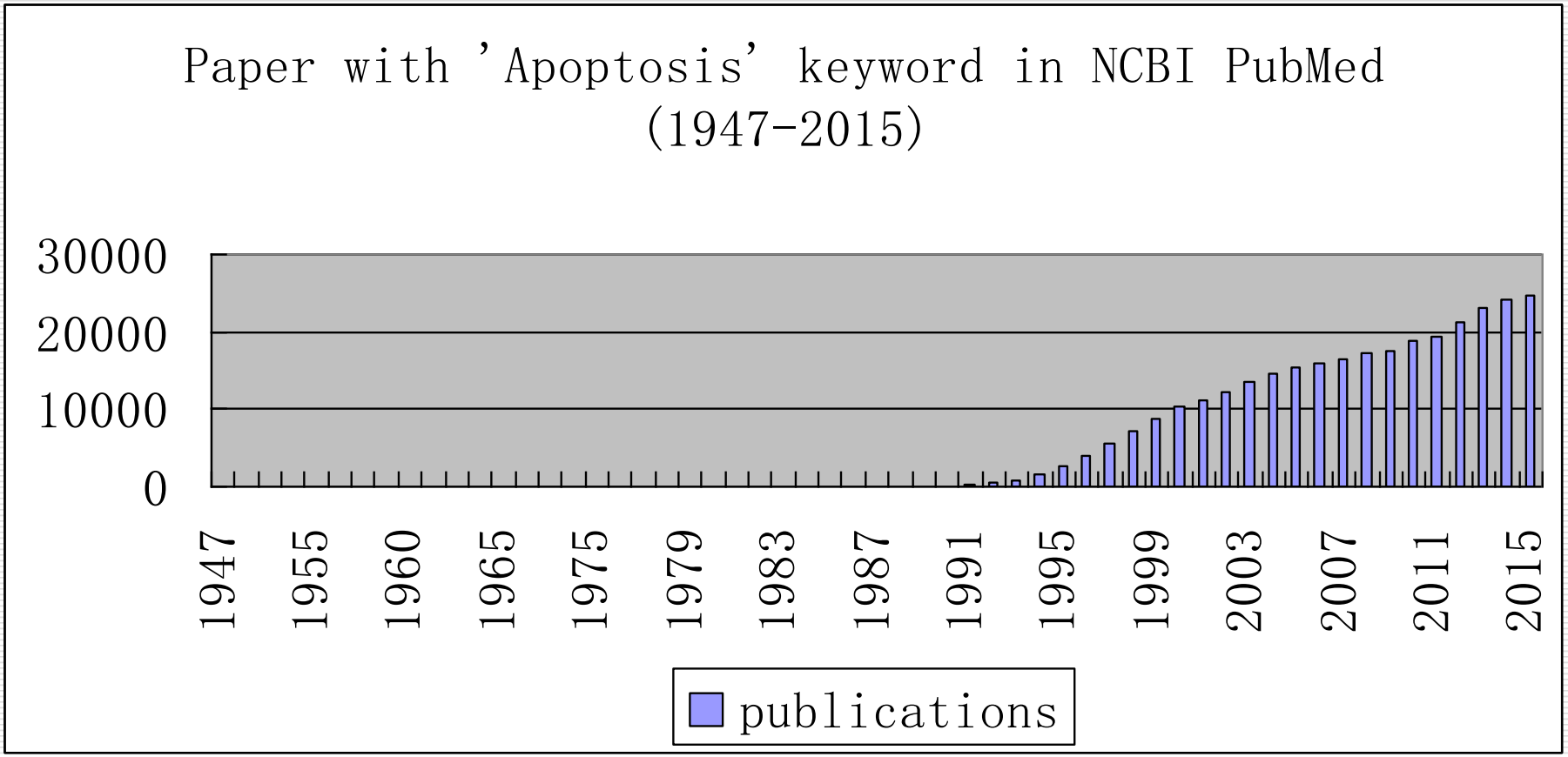
  Bioinformatics,

  Math and Sta,

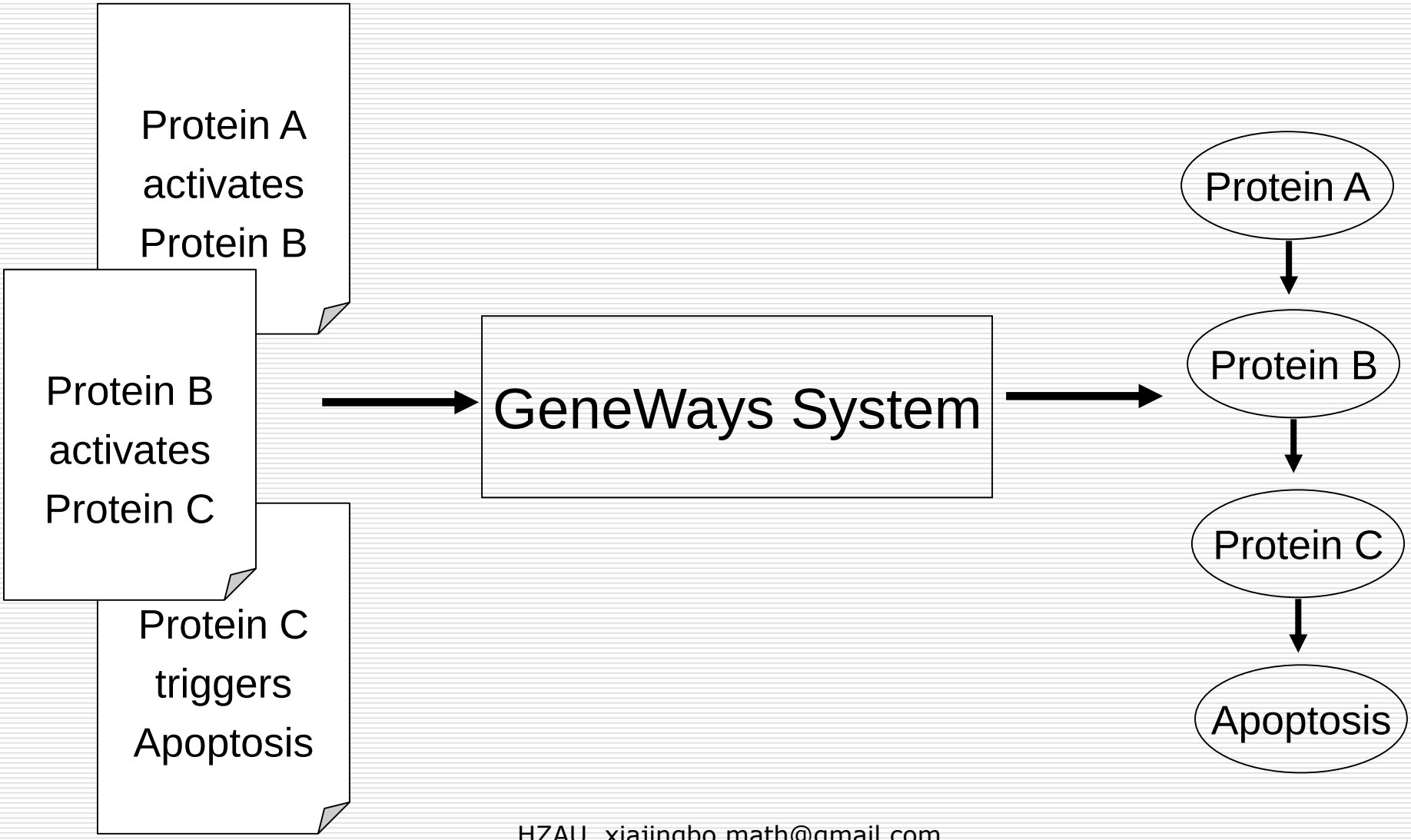
  Linguistics,

  Domain knowledge.

Information Explosion



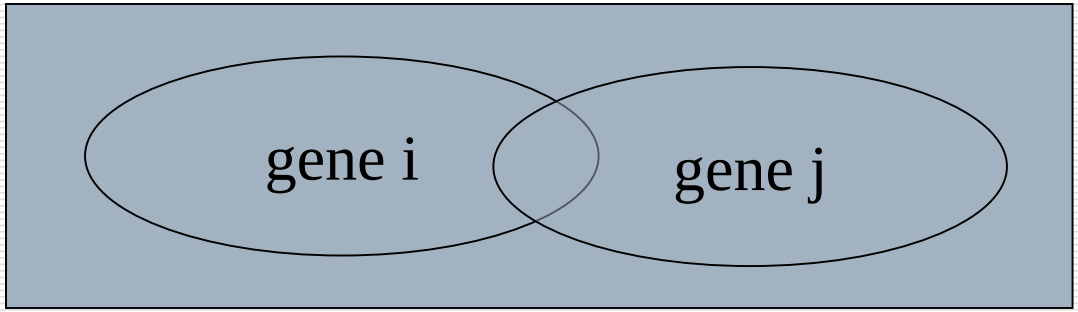
Mining Molecular Interactions



Mining Molecular Interactions (Cont.)

Statistical Methods

- Stapley (2000): Measuring gene associations
- Venn diagram of a set of Medline documents showing the Intersection of documents containing both genes *i* and *j*.
- Bio-Bibliometric distance: $d_{ij}=(|i|+|j|) / (|ij|)$



Stapley, B. J. and G. Benoit (2000). "Biobibliometrics: information retrieval and visualization from co- occurrences of gene names in Medline abstracts." Pac Symp Biocomput: 529-40.

Mining Molecular Interactions (Cont.)

Pattern Matching

- Pattern matching (~regexp) to extract protein-protein interactions
- <gene> <*interact with*> <gene>

Blaschke, C., M. A. Andrade, et al. (1999). "Automatic extraction of biological information from scientific text: protein-protein interactions." Proc Int Conf Intell Syst Mol Biol: 60-7.

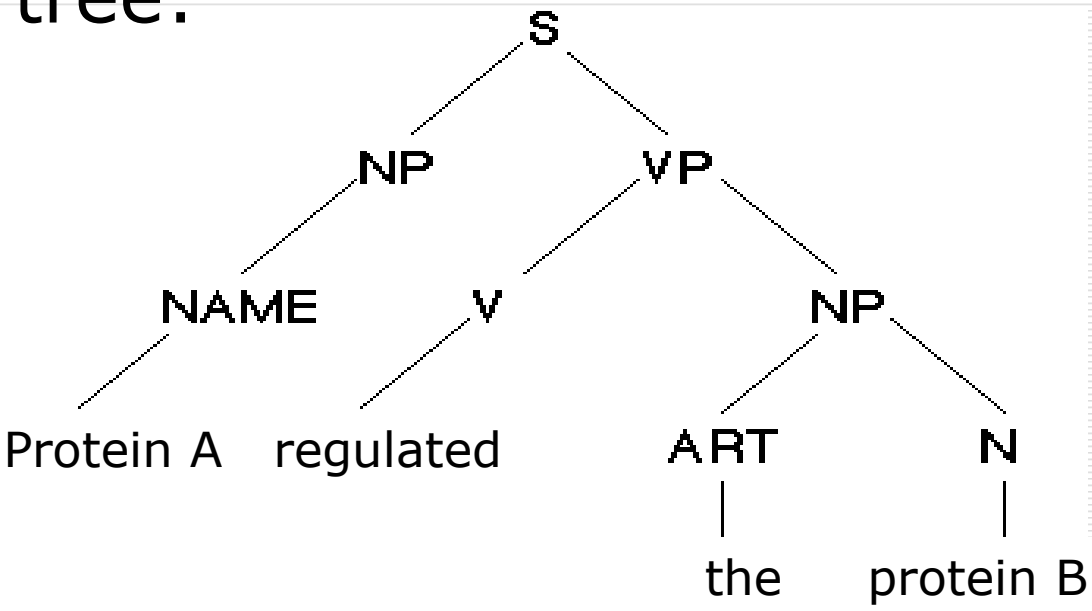
Ng, S. K. and M. Wong (1999). "Toward Routine Automatic Pathway Discovery from On-line Scientific Text Abstracts." Genome Inform Ser Workshop Genome Inform 10: 104-112.

Ono, T., H. Hishigaki, et al. (2001). "Automated extraction of information on protein-protein interactions from the biological literature." Bioinformatics 17(2): 155-61.

Mining Molecular Interactions (Cont.)

Full Parsing

- Parsing: Detect sequence of grammar rules that describe internal structure of sentence
- Grammar rule: S -> NP VP
- [The protein]$_{NP}$ [was degenerated]$_{VP.}$
- Syntax parse tree:



Mining Molecular Interactions (Cont.)

Full Parsing

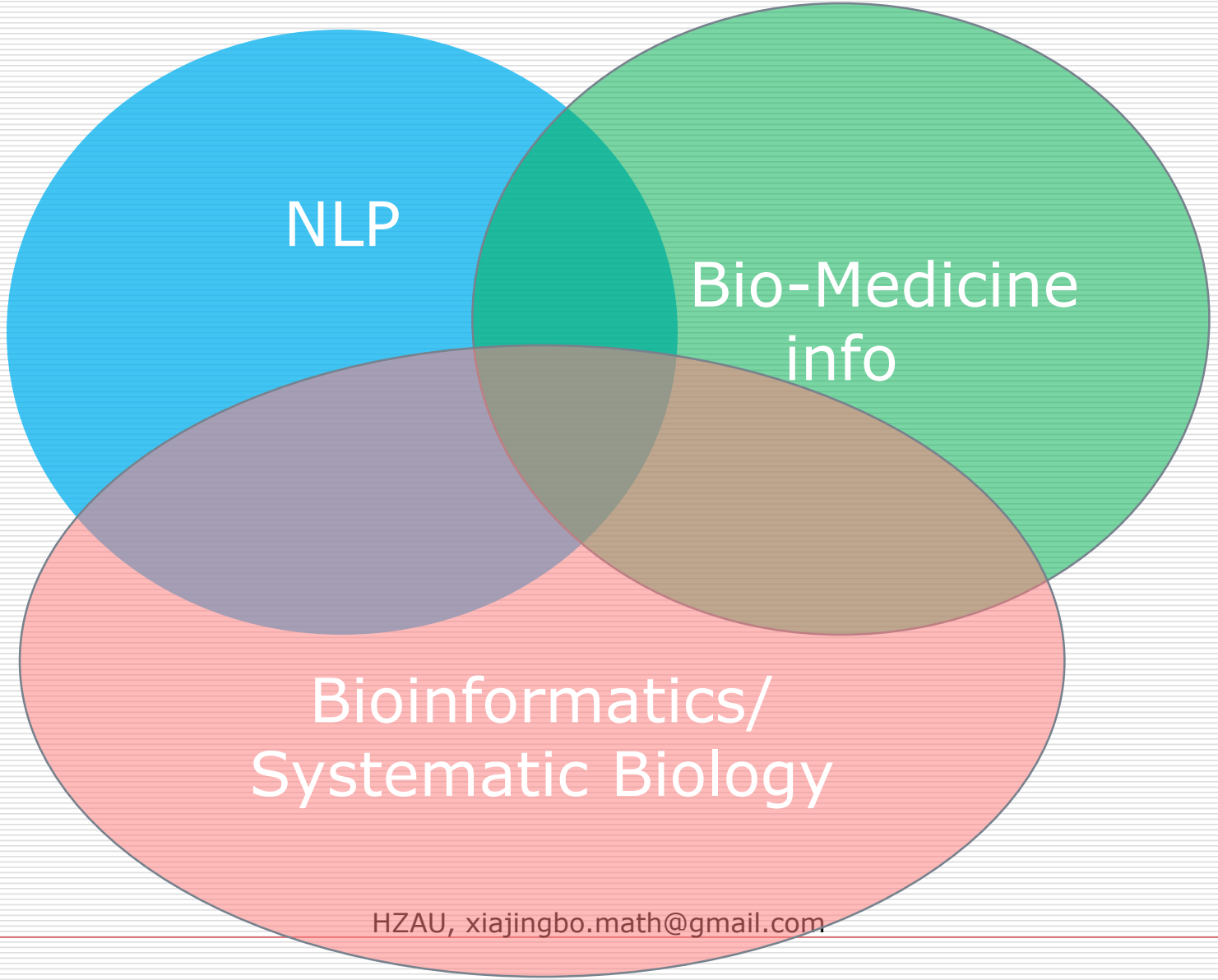
- GENIES: parser for molecular domain. Extracts molecular interactions.
- Frame representation: Each frame is a list beginning with the elements **type, value,** possibly followed by additional frames:

[protein, Il-2, [state, active]]

- For example, the parse of *Raf-1 activates Mek-1* is

[action, activate, [protein, Raf-1], [protein, Mek-1]]

BioNLP in our focus as in HZAU



Outline

- Brief intro of BioNLP
- What make BioNLP unique, if compared with general text mining
- **Main research issues**
- Timetable for this term

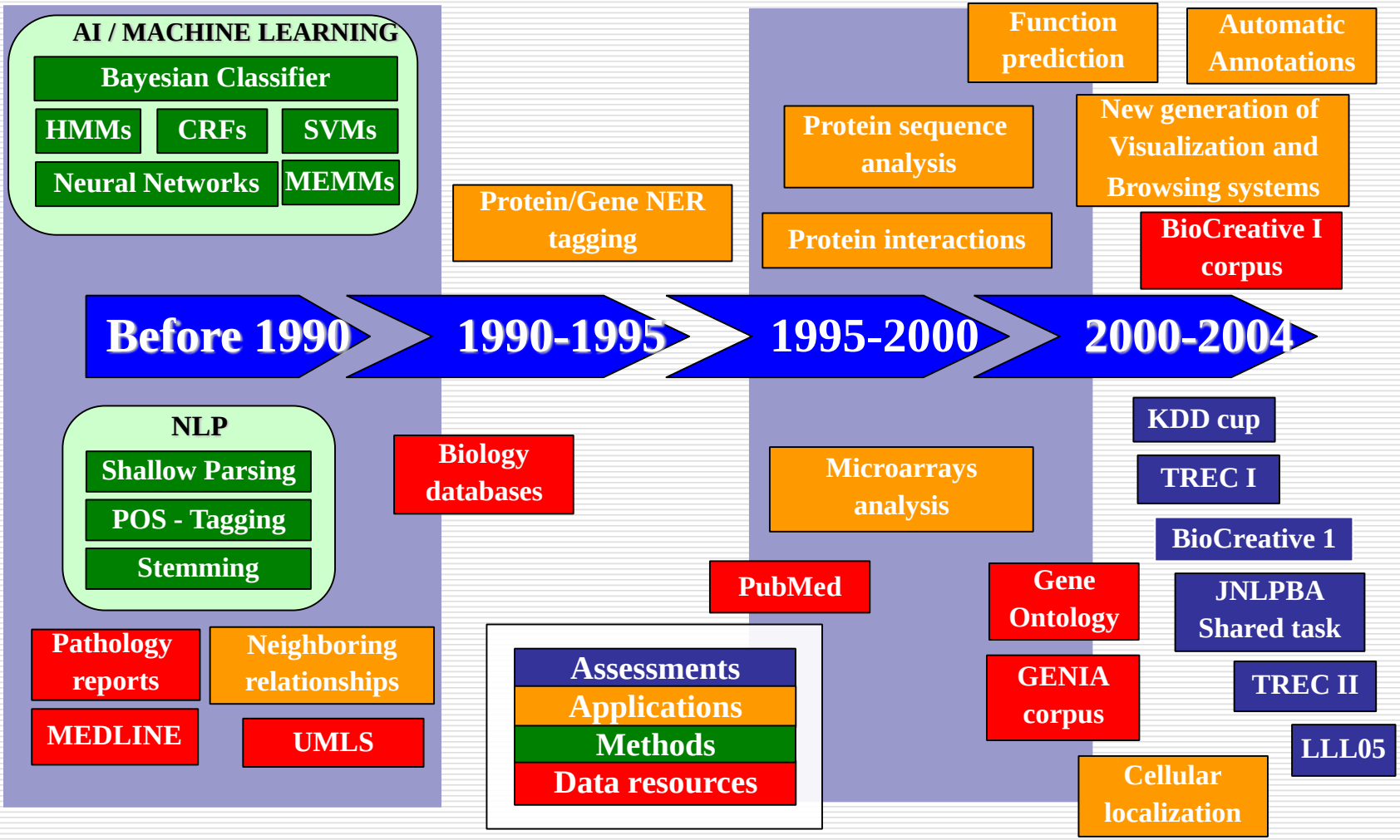
Outline

- Brief intro of BioNLP
- What make BioNLP unique, if compared with general text mining
- **Main research issues**
  - Issues in the early days
  - NLP challenge in BioNLP– timeline (2002-2014)
  - New trend in recent year
- Timetable for this term

Outline

- Brief intro of BioNLP
- What make BioNLP unique, if compared with general text mining
- **Main research issues**
  - **Issues in the early days**
  - NLP challenge in BioNLP– timeline (2002-2014)
  - New trend in recent year
- Timetable for this term

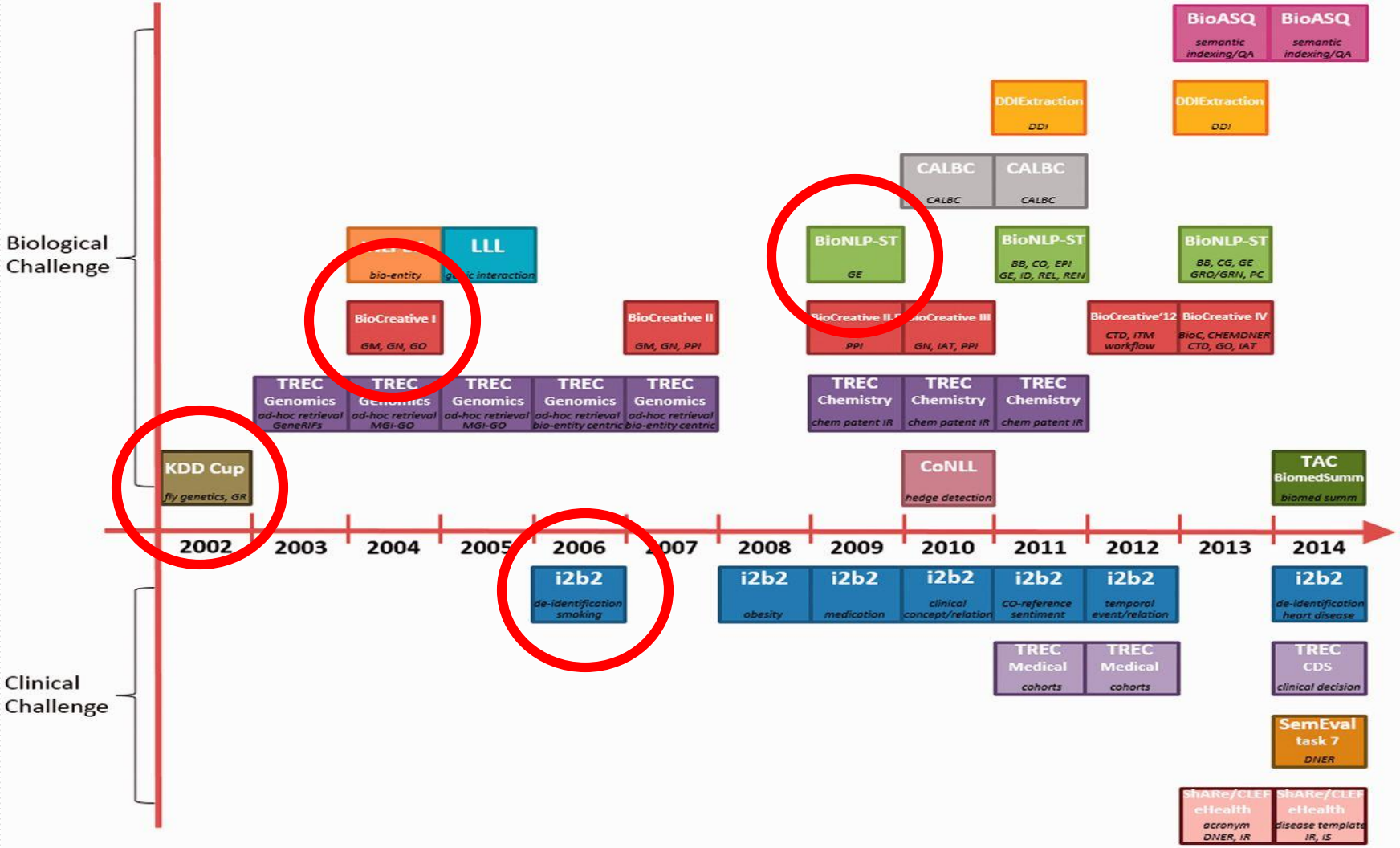
NLP in Molecular Biology – timeline (1990-2004)



Outline

- Brief intro of BioNLP
- What make BioNLP unique, if compared with general text mining
- **Main research issues**
  - Issues in the early days
  - **NLP challenge in BioNLP– timeline (2002-2014)**
  - New trend in recent year
- Timetable for this term

NLP challenge in BioNLP– timeline (2002-2014)



Chung-Chi Huang, and Zhiyong Lu. Brief Bioinform 2016;17:132-144

Briefings in Bioinformatics

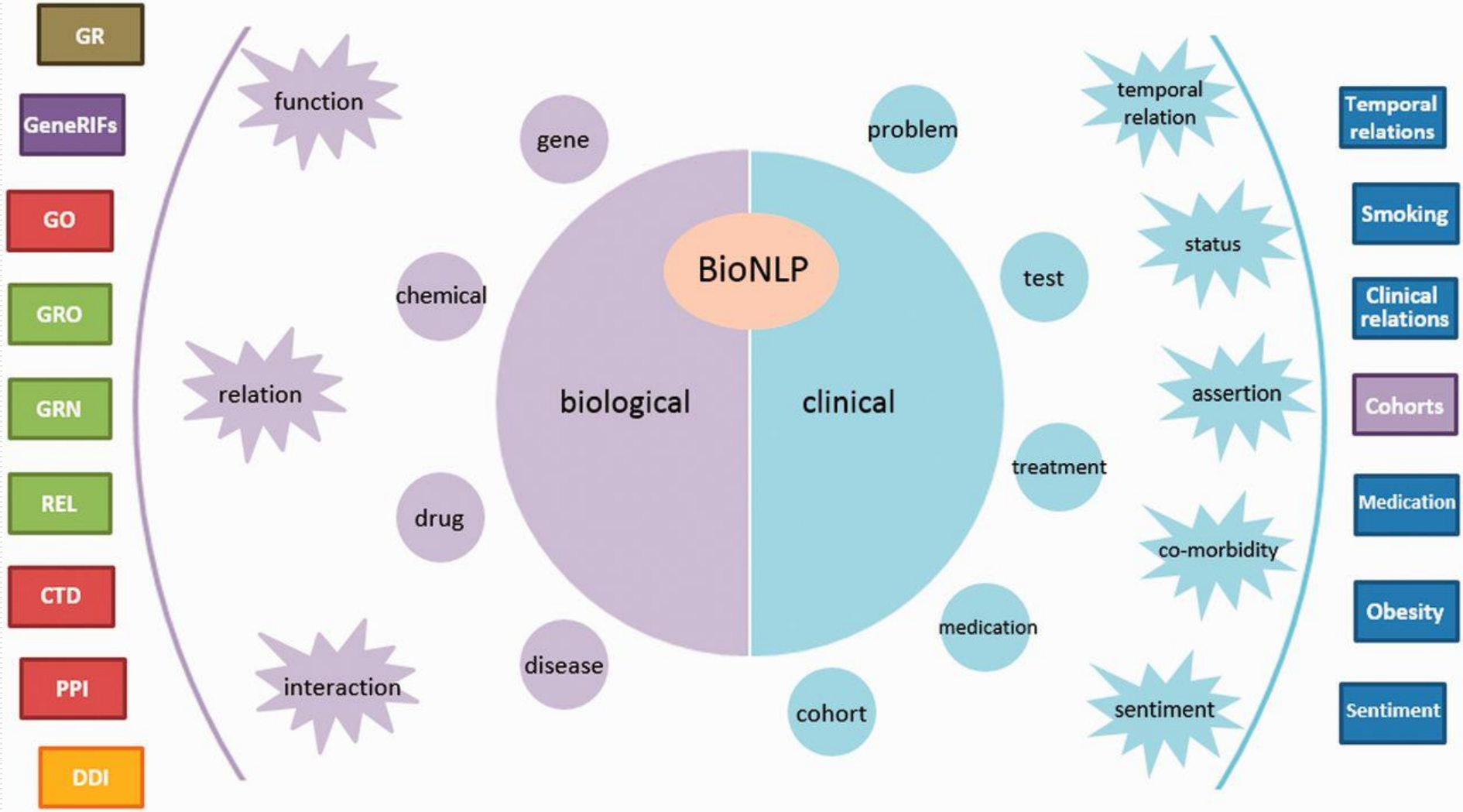
Representative Challenges:

1. KDD 2002
2. BioCreative (From 2004)
3. I2b2 (From 2006)
4. BioNLP Shared Task (From 2009)

Outline

- Brief intro of BioNLP
- What make BioNLP unique, if compared with general text mining
- **Main research issues**
  - Issues in the early days
  - NLP challenge in BioNLP– timeline (2002-2014)
  - **New trend in recent year**
- Timetable for this term

Different biological and clinical problems targeted by BioNLP challenges.



Chung-Chi Huang, and Zhiyong Lu. Brief Bioinform 2016;17:132-144

Briefings in Bioinformatics

Outline

- Brief intro of BioNLP
- What make BioNLP unique, if compared with general text mining
- Main research issues
- **Timetable for this term**

Schedule

- Week 1:

(19, Apr) Ch1. Introduction

(22, Apr) Ch2. Foundation of Mathematical Algorithm

- Week 2:

(26, Apr) Ch3. Foundation of Linguistics

(29, Apr) Ch4. Dataset and Text Retrieval

- Week 3:

(10, May) Ch5. Case Study I. Text Classification

(13, May) Ch6. Case Study II. NER

Schedule

- Week 4:

(17, May) Ch7. Case Study III. Entity and Relation

(20, May) Ch8. Case Study IV. Clinical Info

- Week 5:

(24, May) Ch9. Discussion I (Group Discussion)

(27, May) Ch10. Case Study V. Pheno-Genotype TM

- Week 6:

(31, May) Ch11. Case Study VI: Corpus-based Method

(3, Jun) Ch12. Discussion II (Conclusion Discussion)

Textbook

[1] Cohen, K. B., & Demner-Fushman, D. (2014). Biomedical natural language processing (Vol. 11). John Benjamins Publishing Company

Textbook

[2] Alex Chengyu Fang. English corpora and Automated Grammatical Analysis. The Commercial Press.

Textbook

[3] Daniel Jurafsky, James H. Martin. Speech and Langusge Processing. 人民邮电出版社（引进）.

Textbook

[4] 宗成庆. 统计自然语言处理. 清华大学出版社

Paper for Group Discussion (1):

[1] Langedijk, J., Mantel-Teeuwisse, A. K., Slijkerman, D. S., & Schutjens, M. H. D. (2015). Drug repositioning and repurposing: terminology and definitions in literature. Drug discovery today.

Paper for Group Discussion (2):

(Choose one from the two)

[2] Huang, C. C., & Lu, Z. (2015). Community challenges in biomedical text mining over 10 years: success, failure and the future. Briefings in bioinformatics, bbv024.

[3] Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. Science, 349(6245), 261-266.

Paper for Group Discussion (3):

(Choose one from the two)

[4] Zhang, W., Zou, H., Luo, L., Liu, Q., Wu, W., & Xiao, W. (2015). Predicting potential side effects of drugs by recommender methods and ensemble learning. Neurocomputing.

[5] Patki, A., Sarker, A., Pimpalkhute, P., Nikfarjam, A., Ginn, R., O'Connor, K., ... & Gonzalez, G. (2014). Mining adverse drug reaction signals from social media: going beyond extraction. Proceedings of BioLinkSig, 2014.

Paper for Group Discussion (4):

[6] Hao, T., Chen, X., & Huang, G. (2015). Discovering Commonly Shared Semantic Concepts of Eligibility Criteria for Learning Clinical Trial Design. InAdvances in Web-Based Learning–ICWL 2015 (pp. 3-13). Springer International Publishing.

Paper for Group Discussion (5):

[7] Wang, Z. Y., &Zhang, H. Y. (2013). Rational drug repositioning by medical genetics. Nature biotechnology, 31(12), 1080-1082.

Reference:

Michael Krauthammer. Text Mining in Biomedicine. Department of Pathology,Yale University School of Medicine