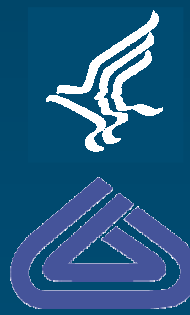
Ontology and Biomedical Informatics
Rome, Italy – May 1, 2005

Lexical and Statistical Approaches to Acquiring Ontological Relations

Formal Methods for Casual Ontology?



Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA

Introduction

- Biomedical ontologies
  - Precisely defined (e.g., formal ontology)
  - Limited size
  - Built manually
- Large amounts of knowledge
  - Not represented explicitly by symbolic relations
  - But expressed implicitly
    - By lexico-syntactic relations (i.e., embedded in terms)
    - By statistical relations (e.g., co-occurrence)
  - Can be extracted automatically

Formal vs. *casual* ontology

- Formal ontology
  - Provides a framework for building sound ontologies
  - Too labor-intensive for building large ontologies
- Casual ontology
  - Usually unsuitable for reasoning
  - Tools for automatic acquisition available

General framework

- Ontology learning
  - [Maedche & Staab, Velardi]
  - ECAI, IJCAI
- Term variation [Jacquemin]
- Terminology / Knowledge TKE, TIA
- Knowledge acquisition/capture K-CAP
- Information extraction

Sources of knowledge for casual ontology

- Long tradition of terminology building
  - Over 100 terminologies available in electronic format
- Large corpora available (e.g., MEDLINE)
  - Entity recognition tools available
    - E.g., MetaMap (UMLS-based)
    - Several for gene/protein names
  - Information extraction methods
- Large annotation databases available
  - MEDLINE citations indexed with MeSH
  - Model organism databases annotated with GO

Formal methods for casual ontology

- Lexico-syntactic methods
  - Lexico-syntactic patterns
  - Nominal modification
  - Prepositional phrases
  - Reified relations
  - Semantic interpretation
- Statistical methods
  - Clustering
  - Statistical analysis of co-occurrence data
  - Association rule mining

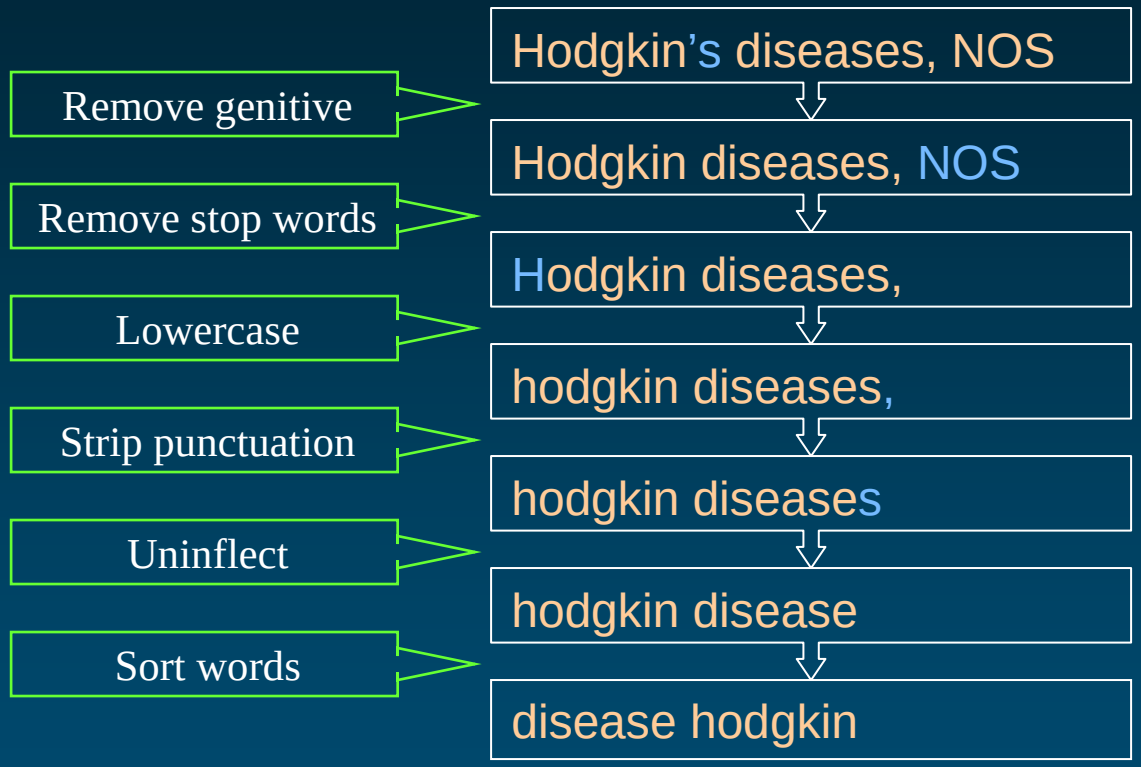
Lexico-syntactic methods

Synonymy

- Source: terminology
- Lexical similarity
  - Lexical variant generation program (UMLS)
  - *norm*
- Limitations
  - Clinical synonymy vs. Synonymy
  - Molecular biology

[McCray & al., SCAMC, 1994]

Normalization



Normalization Example

Hodgkin Disease
HODGKINS DISEASE
Hodgkin's Disease
Disease, Hodgkin's
Hodgkin's, disease
HODGKIN'S DISEASE
Hodgkin's disease
Hodgkins Disease
Hodgkin's disease NOS
Hodgkin's disease, NOS
Disease, Hodgkins
Diseases, Hodgkins
Hodgkins Diseases
Hodgkins disease
hodgkin's disease
Disease, Hodgkin

Taxonomic relations Lexico-syntactic patterns

- Source: text corpus
- Example of patterns
  - *Lamivudin is a nucleoside analogue with potent antiviral properties.*
  - *The treatment of schizophrenia with old typical antipsychotic drugs such as haloperidol can be problematic.*

[Hearst, COLING, 1992]
[Fiszman & al., AMIA, 2003]

Taxonomic relations Nominal modification

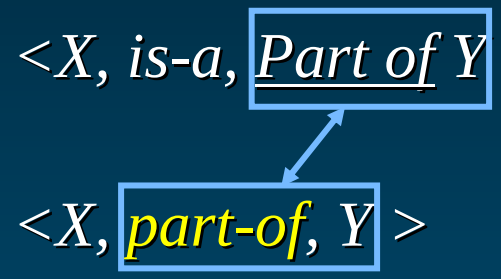
- Source: text corpus / terminology
- Example of modifiers
  - Adjective
    - *Tuberculous Addison's disease*
    - *Acute hepatitis*
  - Noun (noun-noun compounds)
    - *Prostate cancer*
    - *Carbon monoxide poisoning*

Terminology: constrained environment (increased reliability)

[Jacquemin, ACL, 1999]
[Bodenreider & al., TIA, 2001]

Reified relations

- Source: terminology
- Example: reification of part of



- Augmented relations from reified *part-of* relations
  - Reified: <*Cardiac chamber*, *is-a*, *Subdivision of heart*>
  - Augmented: <*Cardiac chamber*, *part-of*, *Heart*>

[Zhang & al., ISWC/Sem. Int., 2003]

Prepositional attachment

- Source: text corpus / terminology
- Example: *of*
  - *Lobe of lung* → part of *Lung*
  - *Bone of femur* → part of *Femur*
- Restrictions
  - Validity of preposition-to-relation correspondence may be limited to a subdomain (e.g., anatomy)
  - Not applicable to complex terms
    - *Groove for arch of aorta* → NOT part of *Aorta*

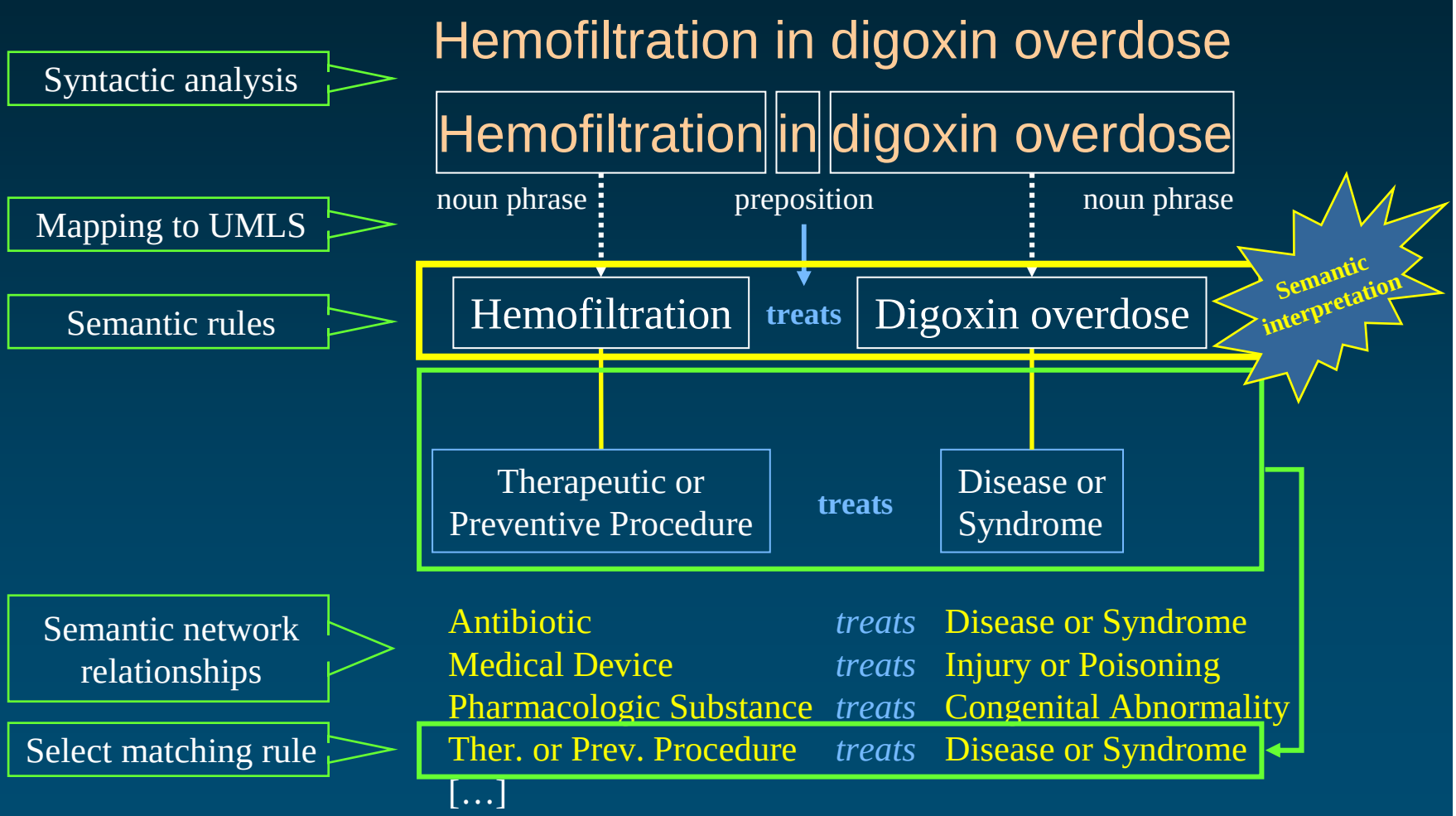
[Zhang & al., ISWC/Sem. Int., 2003]

Semantic interpretation

- Source: text corpus / terminology
- Correspondence between
  - Linguistic phenomena
  - Semantic relations
- Semantic constraints provided by ontologies

[Navigli & al., TKE, 2002]
[Romacker, AIME, 2001]
[Rindflesch & al., JBI, 2003]

Semantic interpretation



Compositional features of terms

- Lexical items [Baud & al., AMIA, 1998]
- Terms within a vocabulary
  - Clinical vocabularies [McDonald & al., AMIA, 1999]
  - Gene Ontology [Ogren & al., PSB, 2004] [Mungall, CFG, 2004]
- Terms across vocabularies
  - SNOMED / LOINC [Dolin, JAMIA, 1998]
  - GO / ChEBI [Burgun, SMBM, 2005]
- Lexicon / Terms
  - Semantic lexicon [Johnson, JAMIA, 1999] [Verspoor, CFG, 2005]

Statistical methods

Taxonomic relations Clustering

- Source: text corpus
- Principle: similarity between words reflected in their contexts
  - Co-occurring words (+ frequencies)
  - Hierarchical clustering algorithms
    - Similarity measure (cosine, Kullback Leibler)
- Can be refined using classification techniques (e.g., k nearest neighbors)

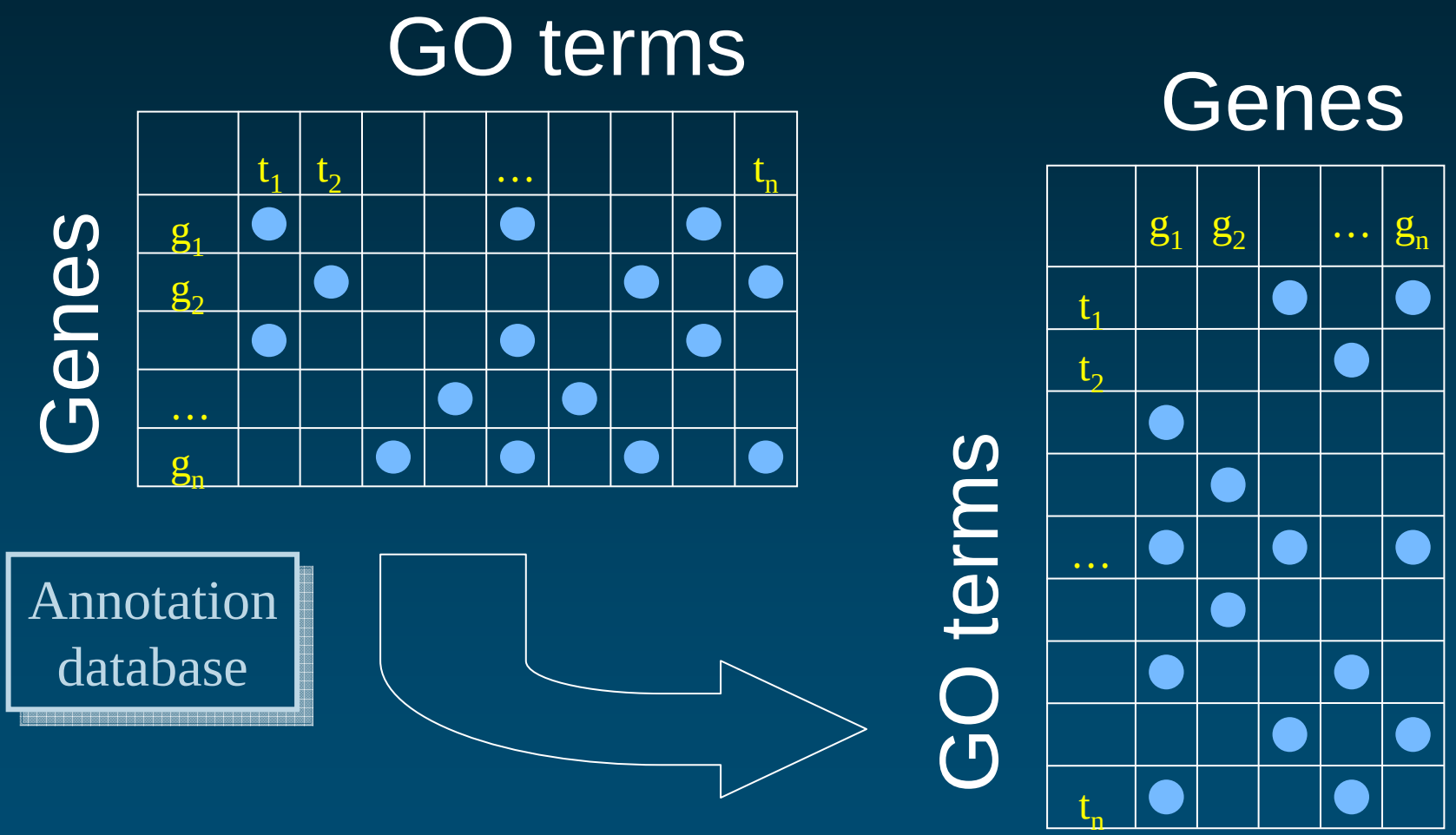
[Faure & al., LREC, 1998]
[Maedche & al., HoO, 2004]

Associative relations

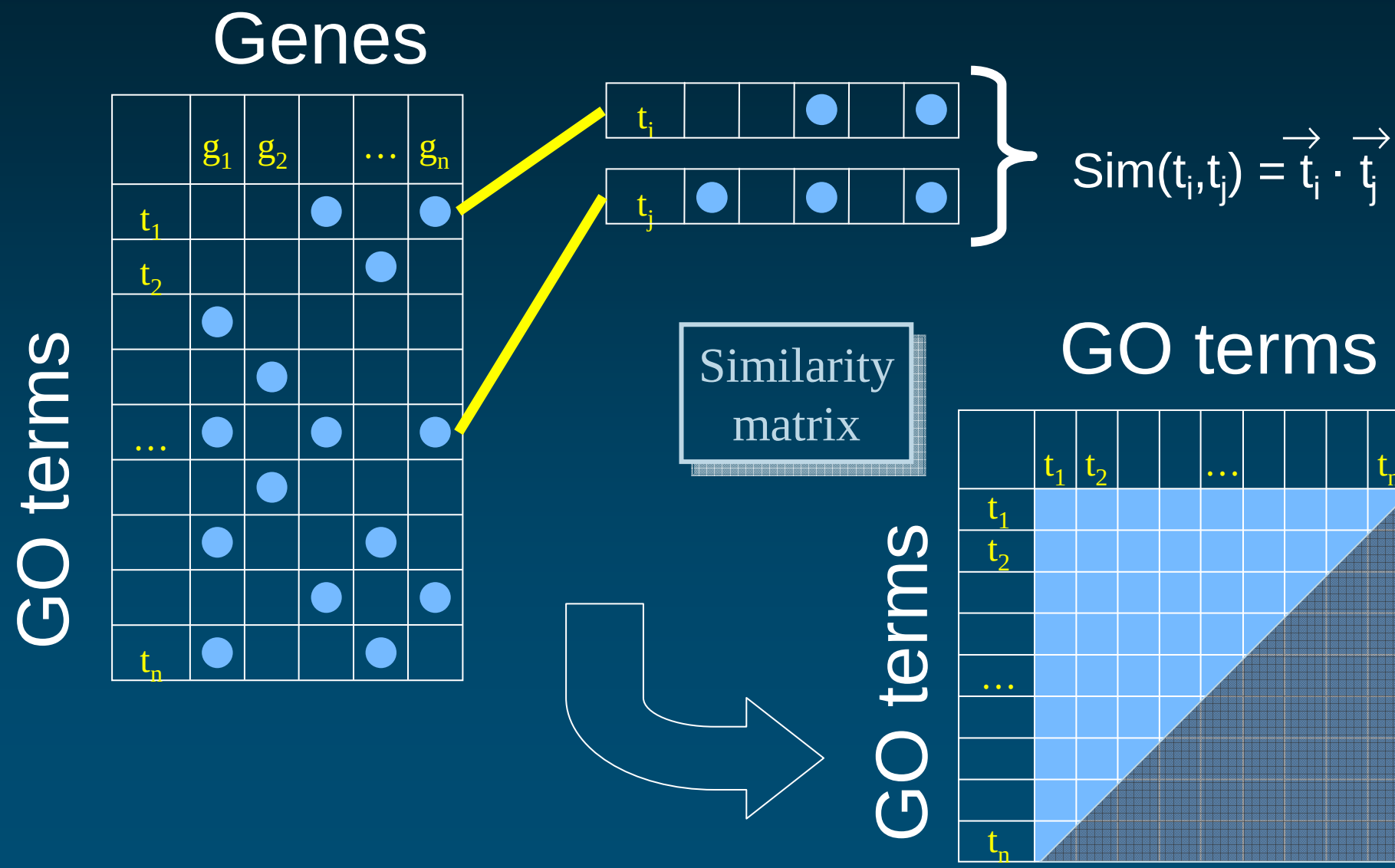
- Source: text corpus / annotation databases
- Principle: dependence relations
  - Associations between terms
- Several methods
  - Vector space model
  - Co-occuring terms
  - Association rule mining
- Limitations: no semantics

[Bodenreider & al., PSB, 2005]

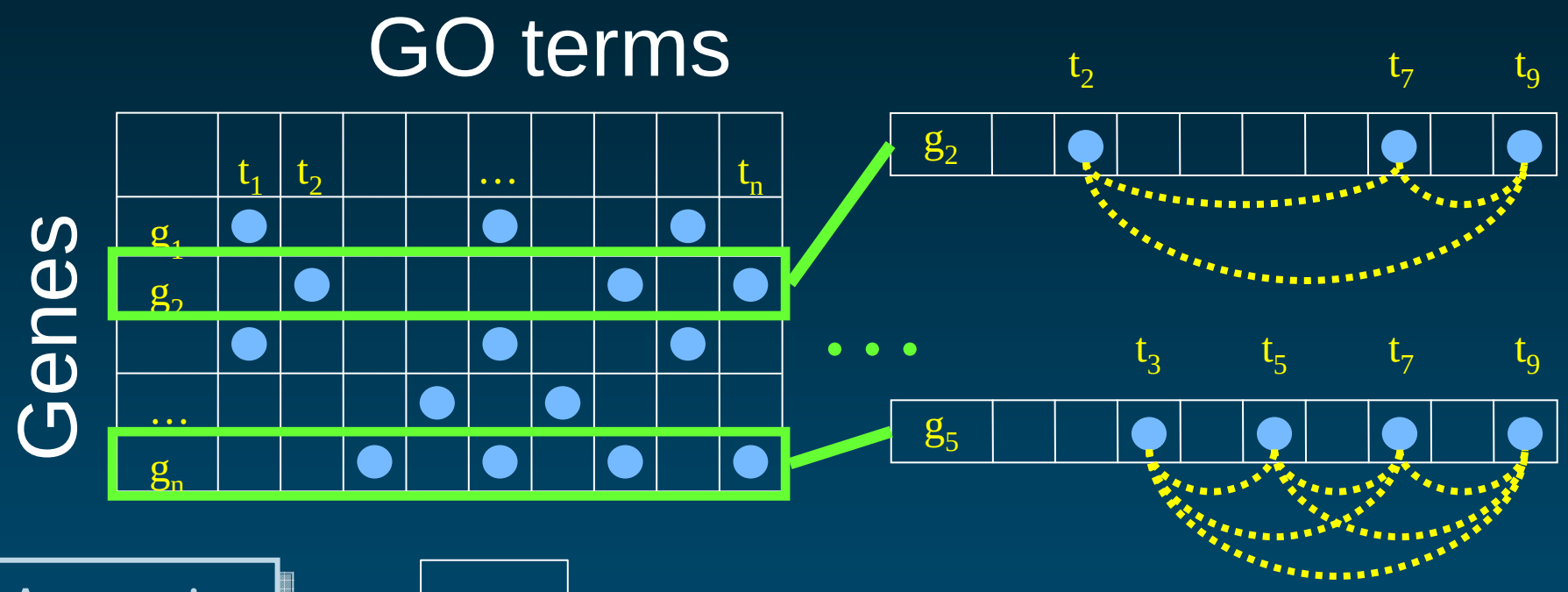
❶ Similarity in the vector space model



❶ Similarity in the vector space model



❷ Analysis of co-occurring GO terms



Annotation database

t_2-t_7	1
t_2-t_9	1
t_7-t_9	2
…	

t_5	1
t_7	2
t_9	2
…	

❷ Analysis of co-occurring GO terms

- Statistical analysis: test independence
  - Likelihood ratio test (G^2)
  - Chi-square test (Pearson's χ^2)
- Example from GOA (22,720 annotations)
  - C0006955 [BP] Freq. = 588
  - C0008009 [MF] Freq. = 53

} Co-oc. = 46

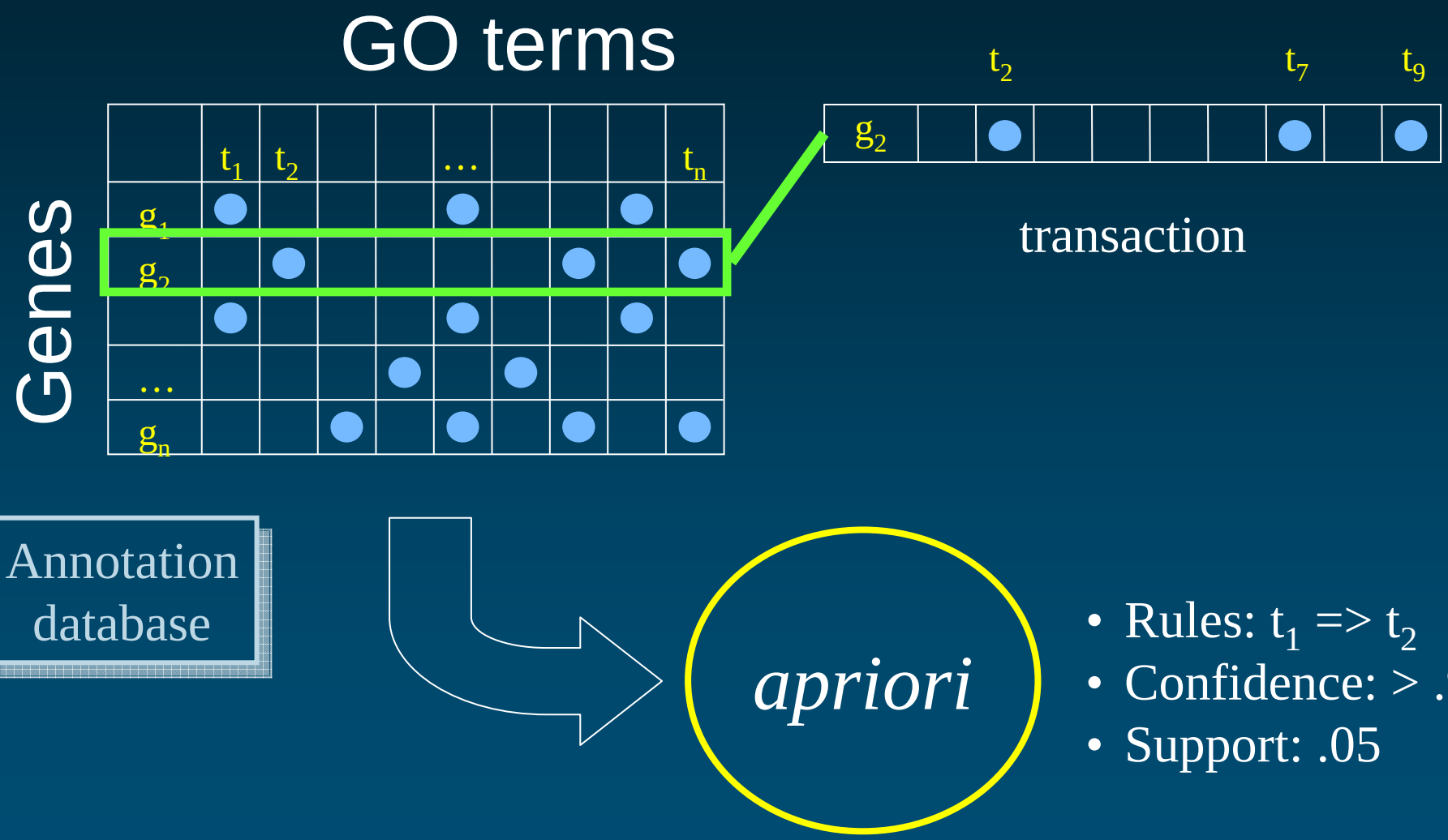
GO:0008009 *immune response*

GO:0006955 *chemokine activity*	present	absent	Total
present	46	542	588
absent	7	21,583	22,132
total	53	22,125	22,720

$G^2 = 298.7$

$p < 0.000$

❸ Association rule mining



Example of associations (GO)

- Vector space model
  - MF: *ice binding*
  - BP: *response to freezing*
- Co-occurring terms
  - MF: *chromatin binding*
  - CC: *nuclear chromatin*
- Association rule mining
  - MF: *carboxypeptidase A activity*
  - BP: *peptolysis and peptidolysis*

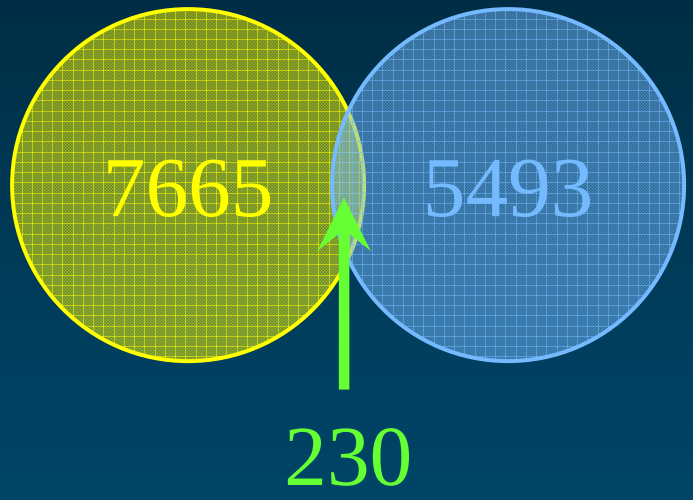
Discussion and Conclusions

Combine methods

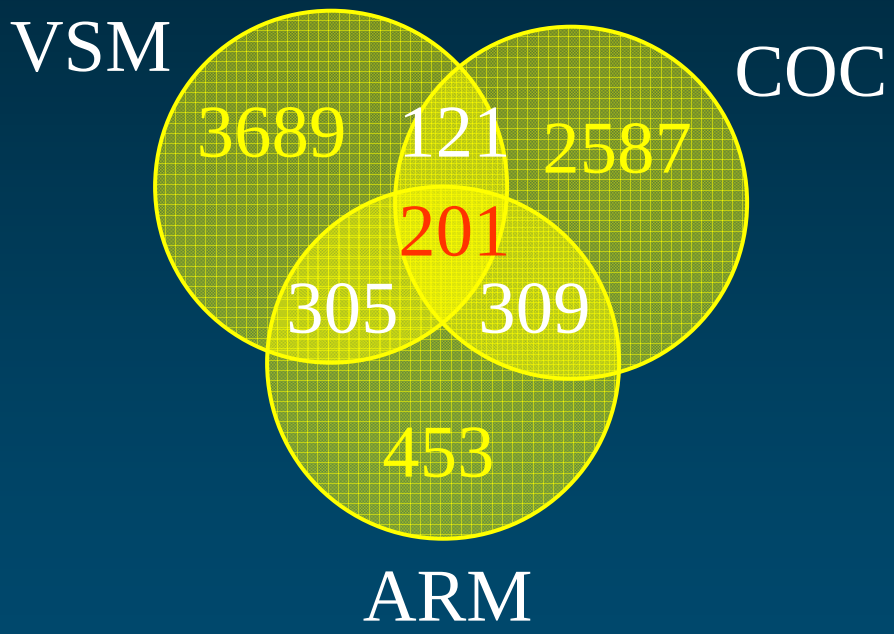
- Affordable relations
  - Computer-intensive, not labor-intensive
- Methods must be combined
  - Cross-validation
  - Redundancy as a surrogate for reliability
  - Relations identified specifically by one approach
    - False positives
    - Specific strength of a particular method
- Requires (some) manual curation
  - Biologists must be involved

Limited overlap among approaches

- Lexical vs. non-lexical



- Among non-lexical



[Bodenreider & al., PSB, 2005]

Reusing thesauri

- First approximation for taxonomic relations
  - No need for creating taxonomies from scratch in biomedicine
- Beware of purpose-dependent relations
  - *Addison's disease isa Autoimmune disorder*
- Relations used to create hierarchies vs. hierarchical relations
- Requires (some) manual curation

[Wroe & al., PSB, 2003]
[Hahn & al., PSB, 2003]

Formal vs. Casual

- Formal ontology
  - Provides a framework for building sound ontologies
  - Too labor-intensive for building large ontologies
- Casual ontology
  - Usually unsuitable for reasoning
  - Tools for automatic acquisition available

What is *not* useful

- Formal ontology = righteous
- Casual ontology = sloppy

Formal and Casual

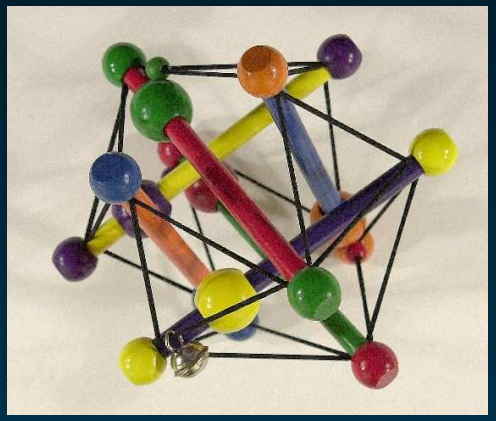
- Formal ontology
  - Provides a framework which can be used as a reference
  - Help us think clearly (?) about
    - Concepts
    - Relations (e.g., isa: is a kind of / is an instance of)
- Casual ontology
  - Supported by “cheap” (but formal) methods
  - Extracted from large amounts of data
  - Helps populating the framework from formal ontology

Combining *formal* and *casual*

- Formal ontology
  - Provides a framework for building sound ontologies
  - Too labor-intensive for building large ontologies
  - Can benefit from loosely defined ontologies
- Casual ontology
  - Usually unsuitable for reasoning
  - Tools for automatic acquisition available
  - Can benefit from formal ontology
    - Organization
    - Validation

Casual ontology as a bridge

- Casual ontology
  - Speaks the language of biologists
    - Extracted from text or terminologies
  - Passes (part of) the rigorous framework of formal ontology on to biologists
- Casual ontologist
  - Not a sloppy ontologist
    - Uses the formal methods of casual ontology
  - Mediator between formal ontology and biology



Medical Ontology Research

Contact: olivier@nlm.nih.gov

Web: mor.nlm.nih.gov

Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA