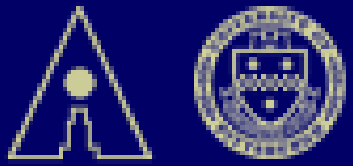


Intelligent Systems Program
University of Pittsburgh

February 3, 2006

Acquiring Ontological Relations From Biomedical Resources

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

General framework

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

Types of relations

- Lexical relations
  - Synonymy
- Ontological relations
  - Intra-ontological
    - Subsumption (*isa*)
    - Meronymy (*part of*)
    - [Instantiation]
  - Trans-ontological
    - Dependence relations
    - Contingent relations

Types of methods

- 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

Types of objectives

- Validate ontologies
  - Compare with asserted knowledge
- Extend ontologies
  - With terms extracted from a corpus
- Enrich ontologies
  - Maintenance
  - Alignment
- Link ontologies to other ontologies

Biomedical resources available

- 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

Ontologies *vs.* 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 degree of manual curation

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

Overview

- Validate
- Extend
- Enrich
- Link

	Lexico-syntactic	Statistical
Intra-ontological	Adjectival modif. LS patterms Reified part of Prep. attachment	Clustering
Trans-ontological	Semantic interpretation	Vector space model Co-occurrence anal. Assoc. rule mining

- Enrich
- Link

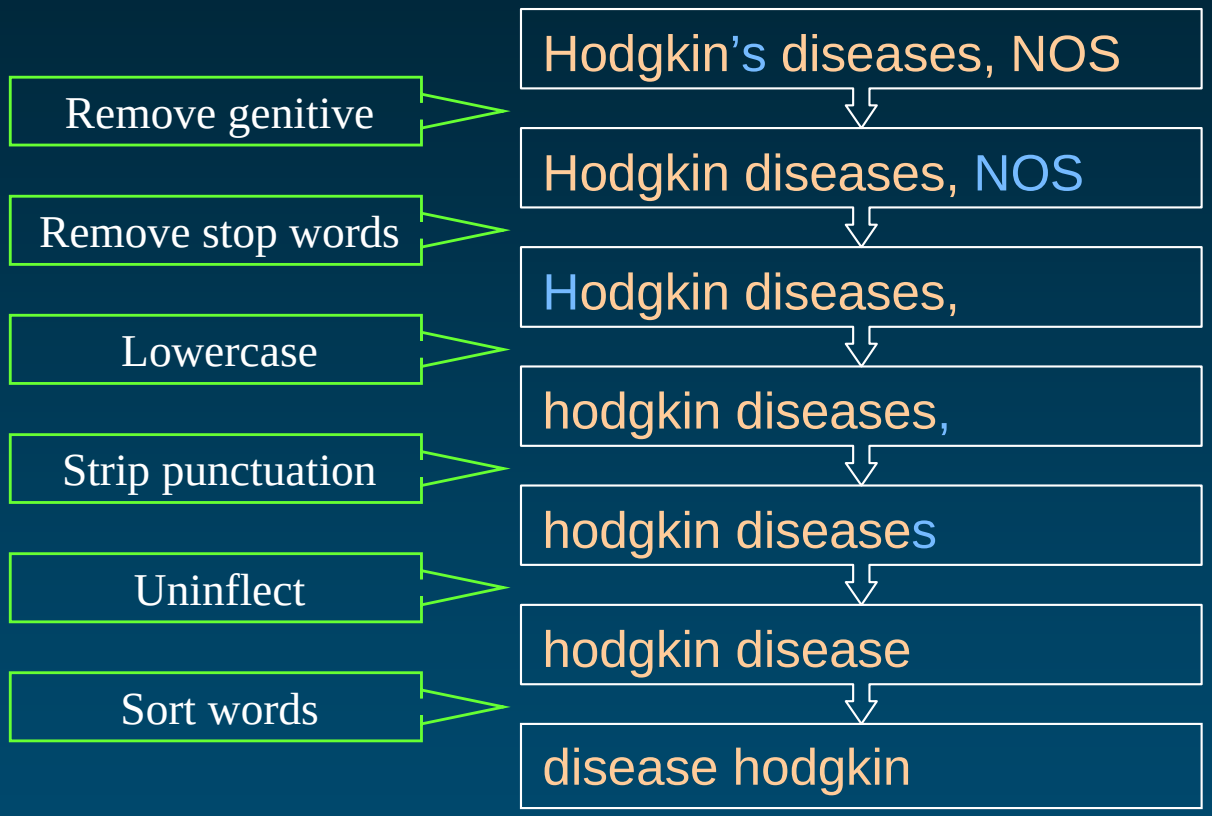
Lexico-syntactic methods

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

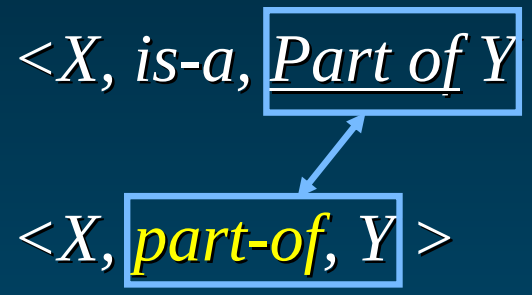
- Source: text corpus / terminology
- Example of modifiers
  - Adjective
    - *Tuberculous Addison's disease*
    - *Acute hepatitis*
  - Noun (noun-noun compounds)
    - *Prostate cancer*
    - *Carbon monoxide poisoning*

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*

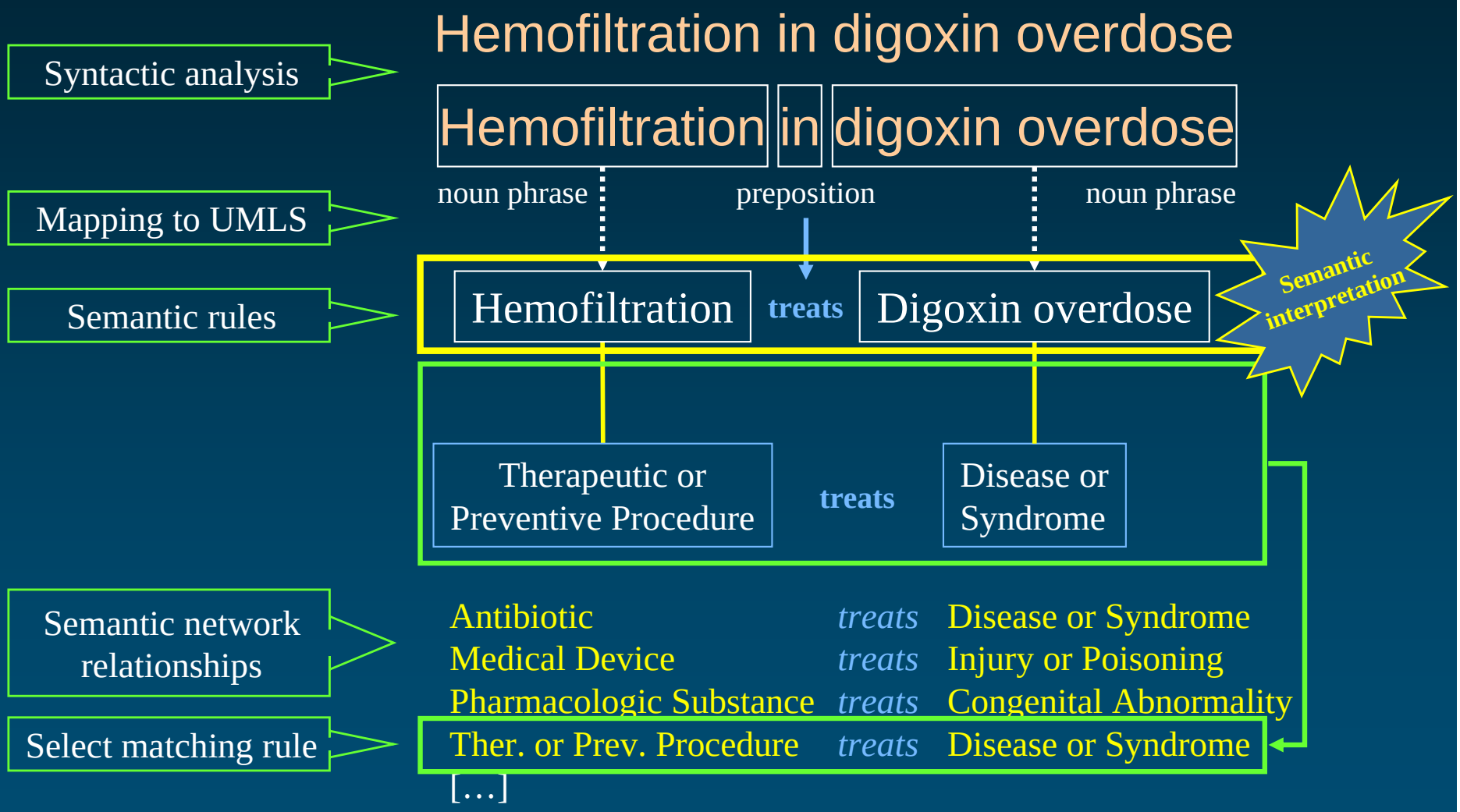
[Zhang & al., ISWC/Sem. Int., 2003]

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

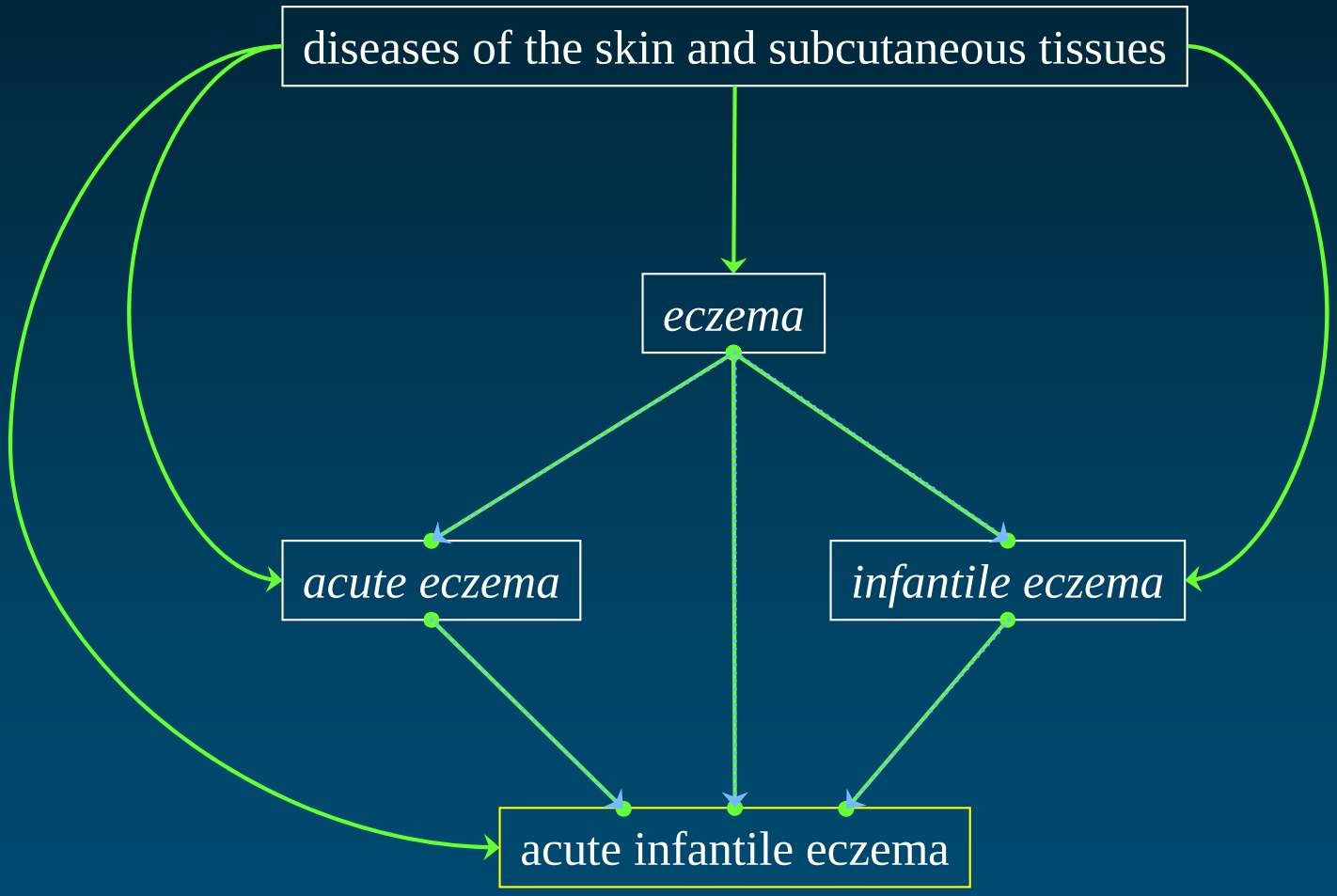
- Lexical items [Baud & al., AMIA, 1998]
- Terms within a vocabulary
  - Clinical vocabularies [McDonald & al., AMIA, 1999]
  - Gene Ontology [Ogren & al., PSB, 2004] [Mungall, CFG, 2004]
- Terms across vocabularies
  - SNOMED / LOINC [Dolin, JAMIA, 1998]
  - GO / ChEBI [Burgun, SMBM, 2005]
- Lexicon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

Applications Ontology validation

- 28,851 pairs of terms
  - Original SNOMED term
  - Transformed term (found in UMLS)
- Corresponding relationship in the Metathesaurus
  - Hierarchical in 50% of the cases
  - “Sibling” in 25% of the cases
  - Missing in 25% of the cases

[Bodenreider & al., TIA, 2001]

Lack of structure within a source

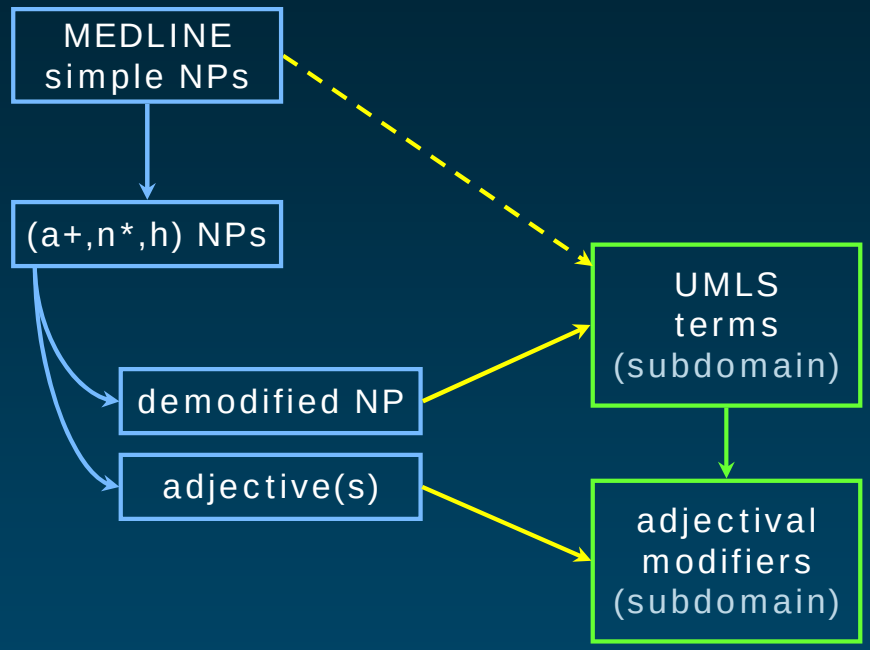


Plesionymy


posttransfusion hepatitis
posttransfusion viral hepatitis

Applications To extend ontologies

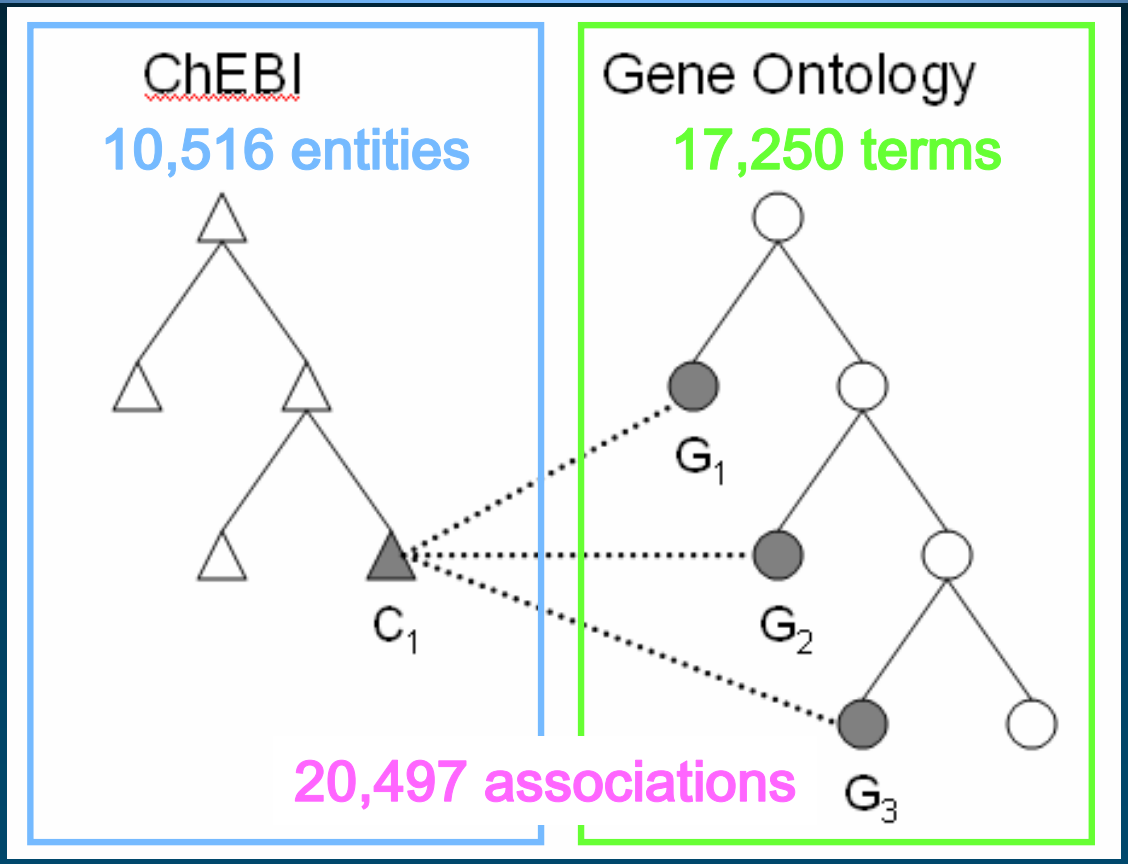
- 3 M “simple” MEDLINE NPs
- 21,000 already in the Metathesaurus (eliminated)
- 1.3 M (adj+, noun*, head) NPs
- 1.6 M demodified terms
- 125,464 candidate terms



- Manual review
- Relevance of the association: 83%

[Bodenreider & al., ACL/NLP-BioMed, 2002]

Applications To link ontologies



[Burgun & al., SMBM 2005]

- 2,700 ChEBI entities (27%) identified in some GO term
- 9,431 GO terms (55%) include some ChEBI entity in their names

Examples of links ChEBI-GO

- iron [CHEBI:18248]
  - **BP** iron ion transport [GO:0006826]
  - **MF** iron superoxide dismutase activity [GO:0008382]
  - **CC** vanadium-iron nitrogenase complex [GO:0016613]
- uronic acid [CHEBI:27252]
  - **BP** uronic acid metabolism [GO:0006063]
  - **MF** uronic acid transporter activity [GO:0015133]
- carbon [CHEBI:27594]
  - **BP** response to carbon dioxide [GO:0010037]
  - **MF** carbon-carbon lyase activity [GO:0016830]

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)

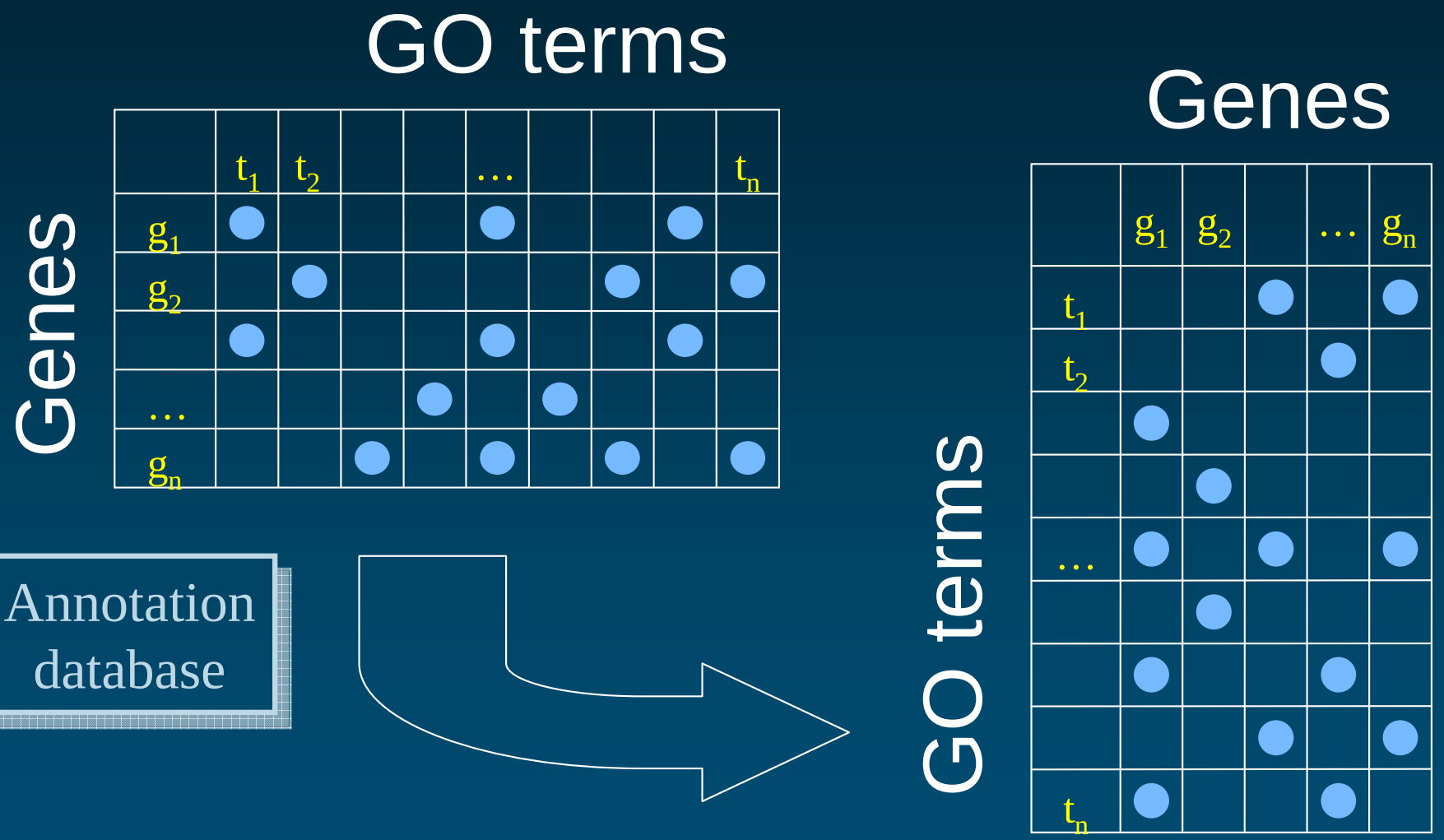
[Faure & al., LREC, 1998]
[Maedche & al., HoO, 2004]

Associative relations

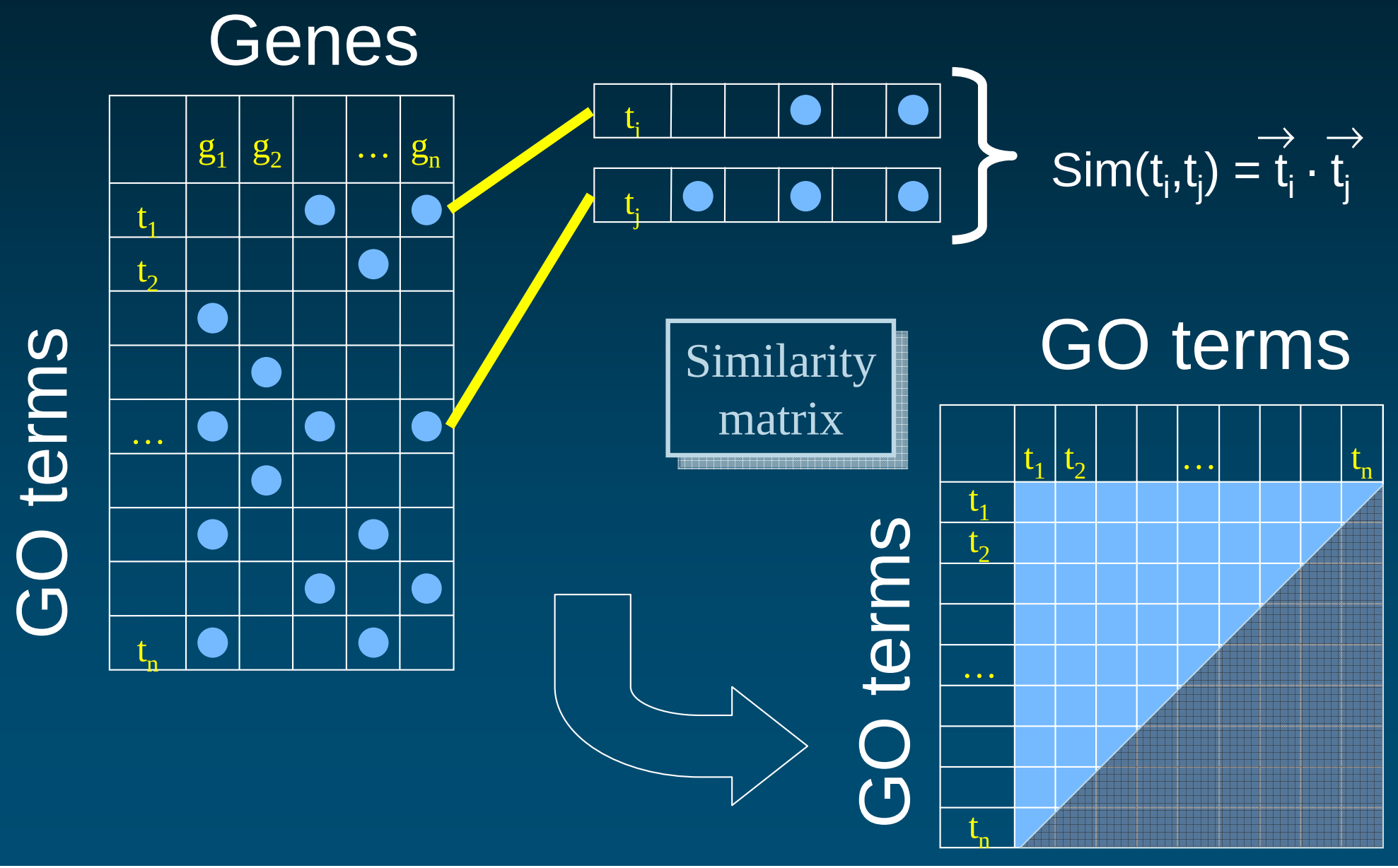
- Source: text corpus / annotation databases
- Principle: dependence relations
  - Associations between terms
- Several methods
  - Vector space model
  - Co-occuring terms
  - Association rule mining
- Limitations: no semantics

[Bodenreider & al., PSB, 2005]

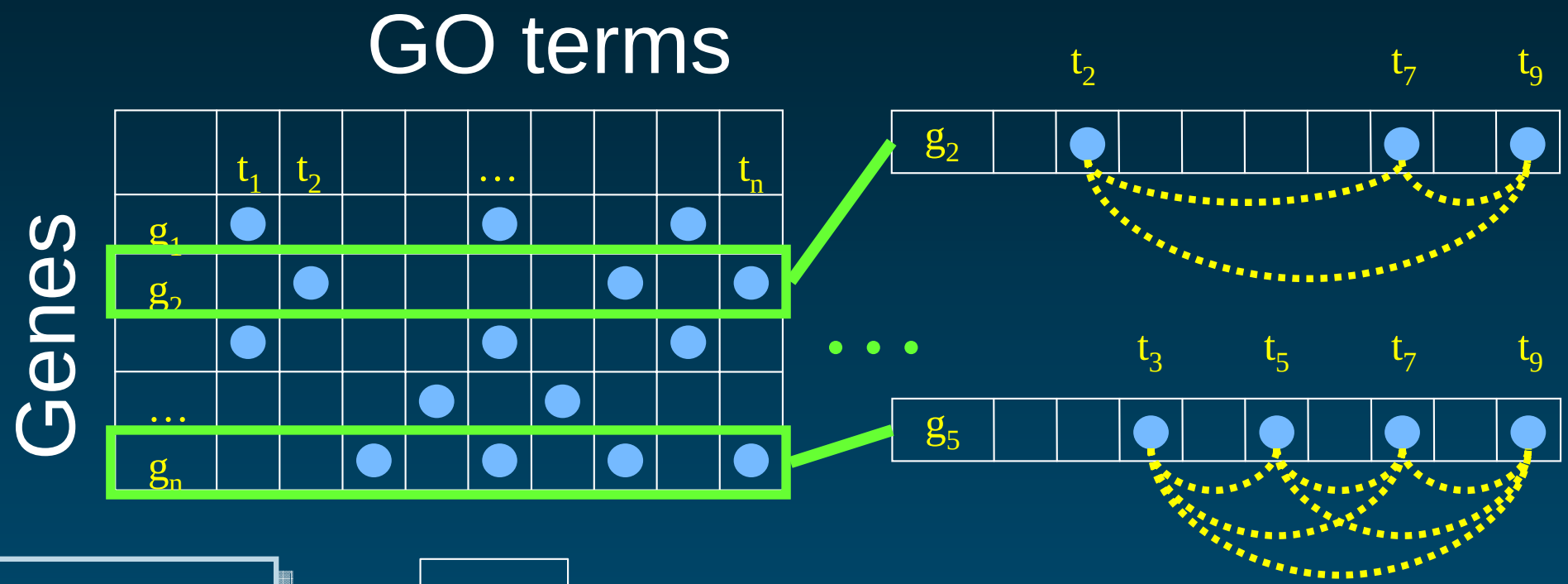
❶ Similarity in the vector space model



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

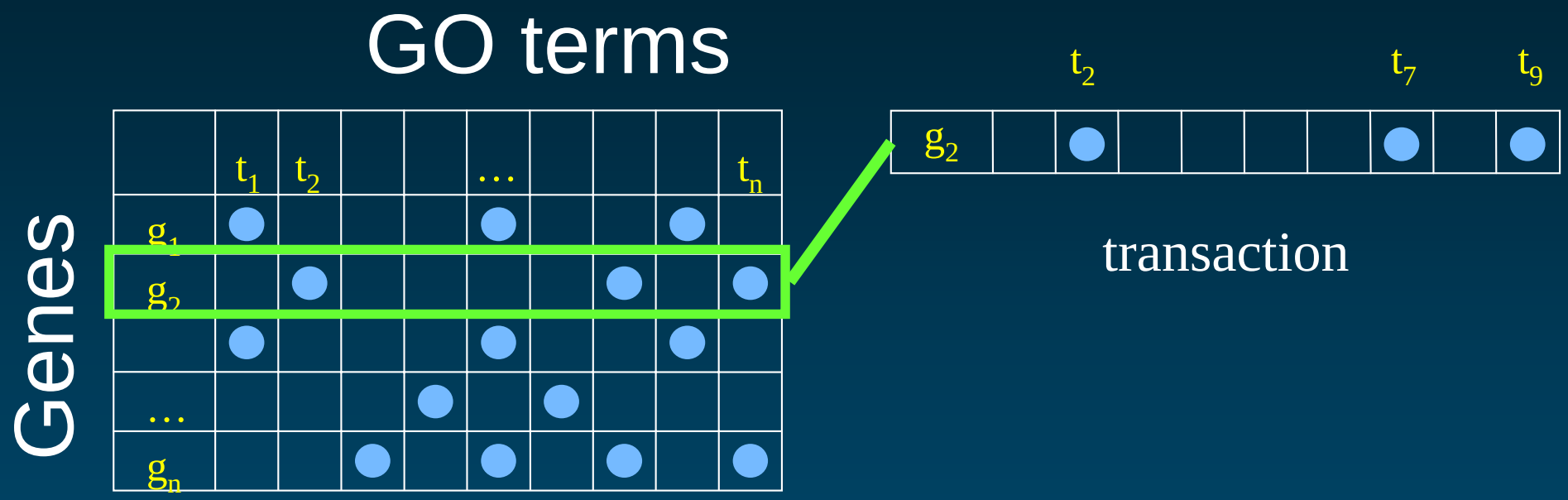
GO:0008009 *immune response*

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



Annotation database

apriori

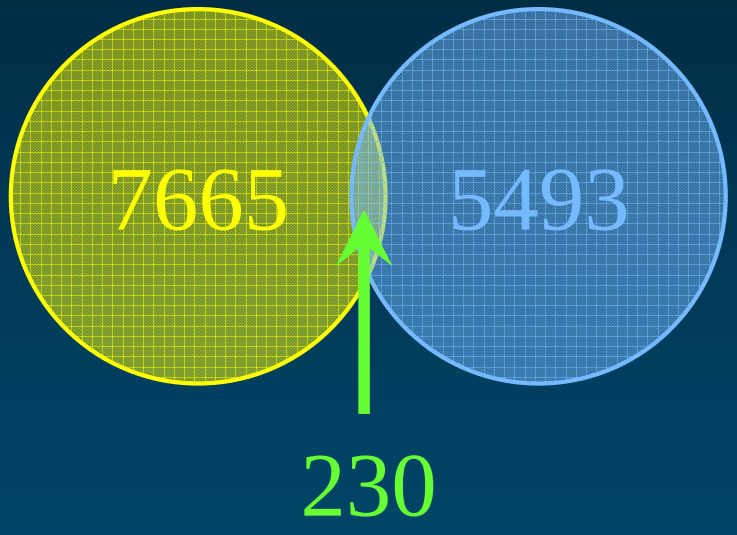
- Rules: $t_1 \Rightarrow t_2$
- Confidence: > .9
- Support: .05

Example of associations (GO)

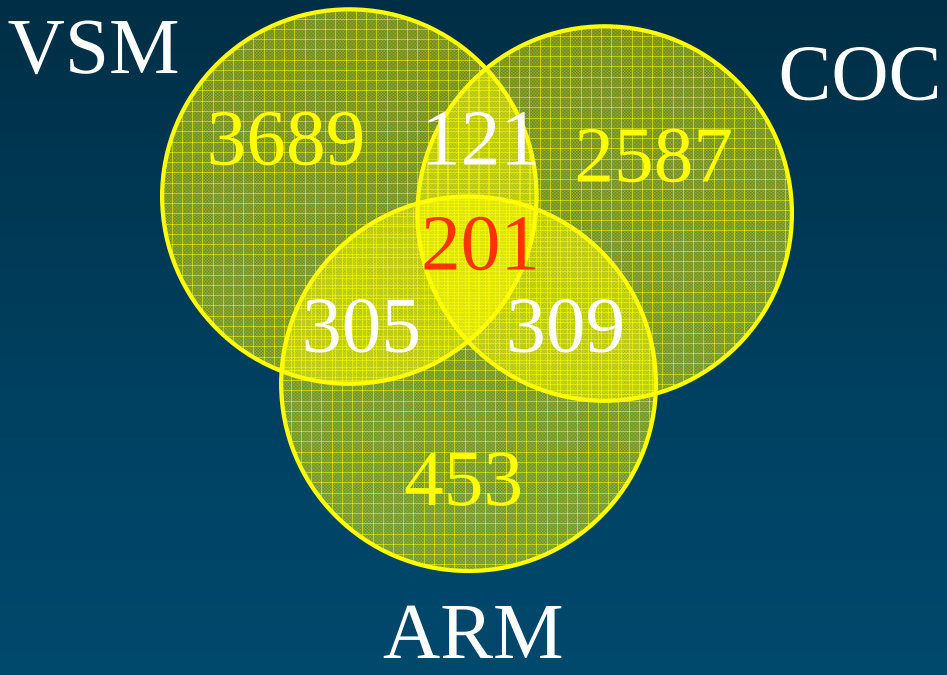
- Vector space model
  - MF: *ice binding*
  - BP: *response to freezing*
- Co-occurring terms
  - MF: *chromatin binding*
  - CC: *nuclear chromatin*
- Association rule mining
  - MF: *carboxypeptidase A activity*
  - BP: *peptolysis and peptidolysis*

Limited overlap among approaches

- Lexical vs. non-lexical



- Among non-lexical



Discussion

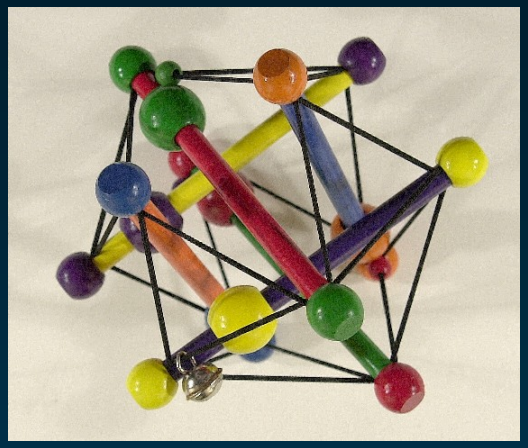
Lexico-syntactic *vs.* statistical

- Lexical
  - Based on terminologies/ontologies
  - Inferable semantics
- Statistical
  - Based on knowledge bases
  - No semantics

- Non-redundant, complementary techniques
- Both require some degree of manual curation (semi-automatic techniques)

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



Medical Ontology Research

Contact: olivier@nlm.nih.gov

Web: mor.nlm.nih.gov

NATIONAL LIBRARY OF MEDICINE

Olivier Bodenreider

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