



Academy of Mathematics and Systems Science
Chinese Academy of Sciences

Beijing, China
August 29, 2017

Using lexical and structural features for quality assurance of biomedical ontologies

Application to SNOMED CT



Olivier Bodenreider

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



U.S. National Library of Medicine







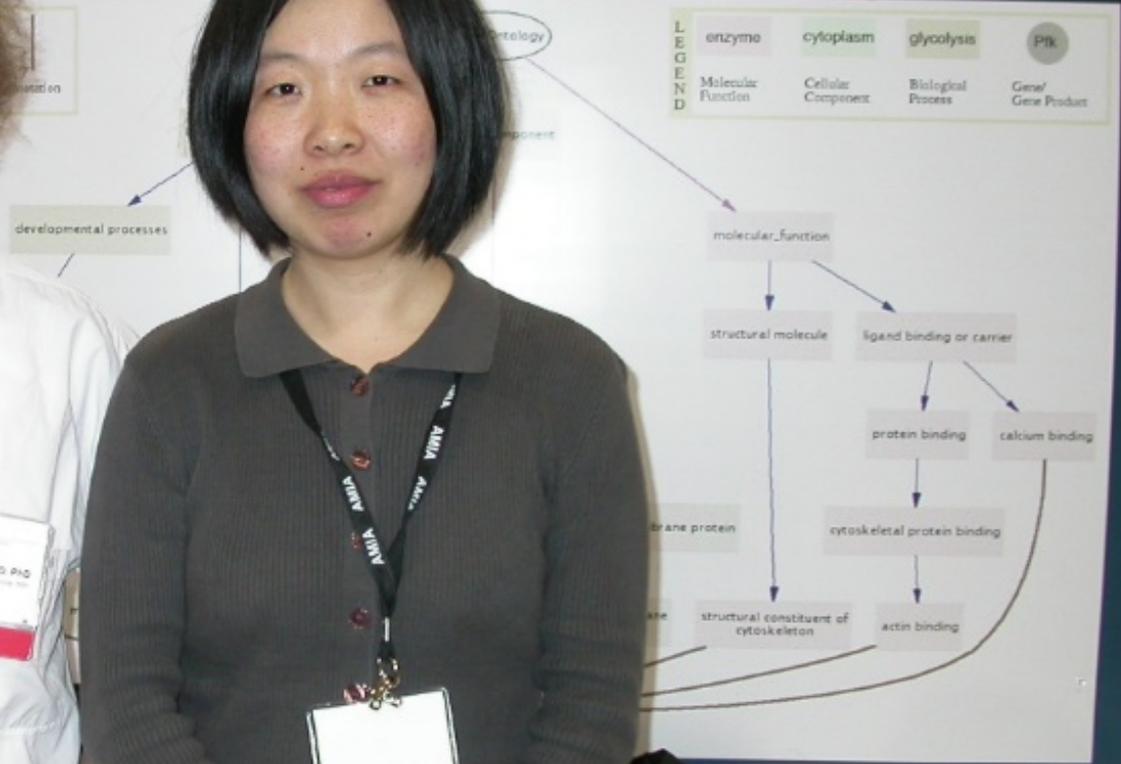
GenNav: Visualizing Gene Ontology as a graph

Olivier Bodenreider, MD, PhD, National Library of Medicine, Bethesda, MD

2002

Most browsers available for visualizing and navigating the Gene Ontology™ as a list of indented components

Gene Ontology
#0000001: Gene Ontology (GO)
#0000002: Cellular Component
#0000003: Molecular Function
#0000004: Biological Process



from
data
inter
visual
pack
the la

http://

ww

w

2007



Towards an Integrated View on Drug Information

Kelly Zeng, Olivier Bodenreider, John T. Kilbourne, Stuart J. Nelson
National Library of Medicine, Bethesda, Maryland, USA

P386

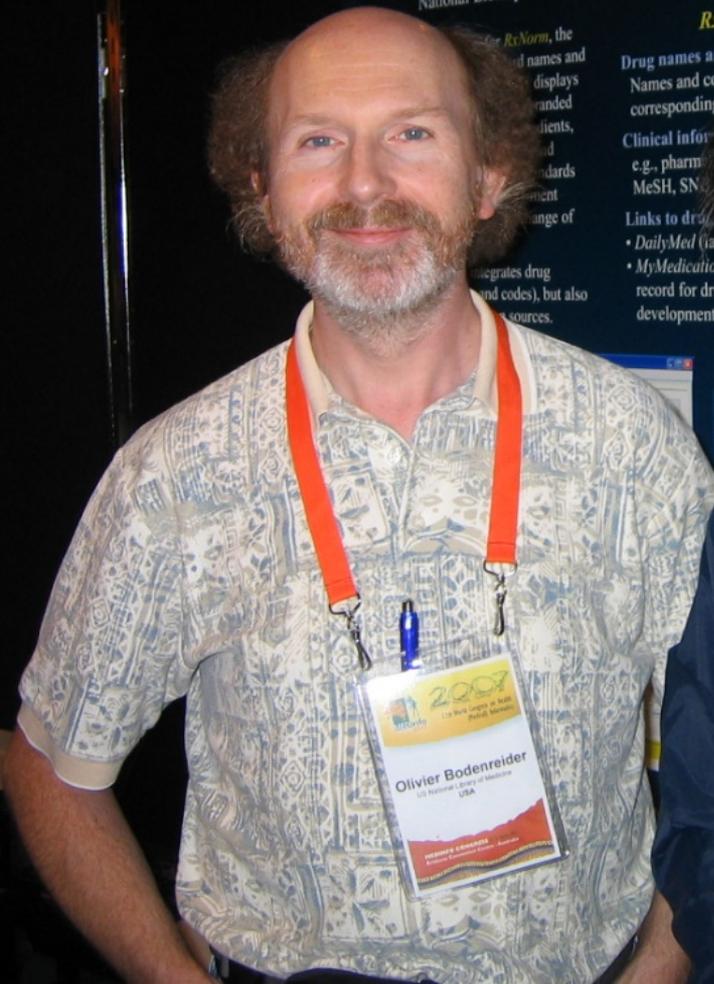
For RxNorm, the
drug names and
displays
branded
patients,
and
standards
ment
ange of
integrates drug
(and codes), but also
sources.

RxNav views

Drug names and co
Names and code
corresponding
Clinical info
e.g., pharm
MeSH, SN
Links to dr
• DailyMed (tab
• MyMedicationLi
record for drugs, c
development)

Codes for

- RXNORM
- Multum
- Nat'l



Disclaimer

The views and opinions expressed do not necessarily state or reflect those of the U.S. Government, and they may not be used for advertising or product endorsement purposes.

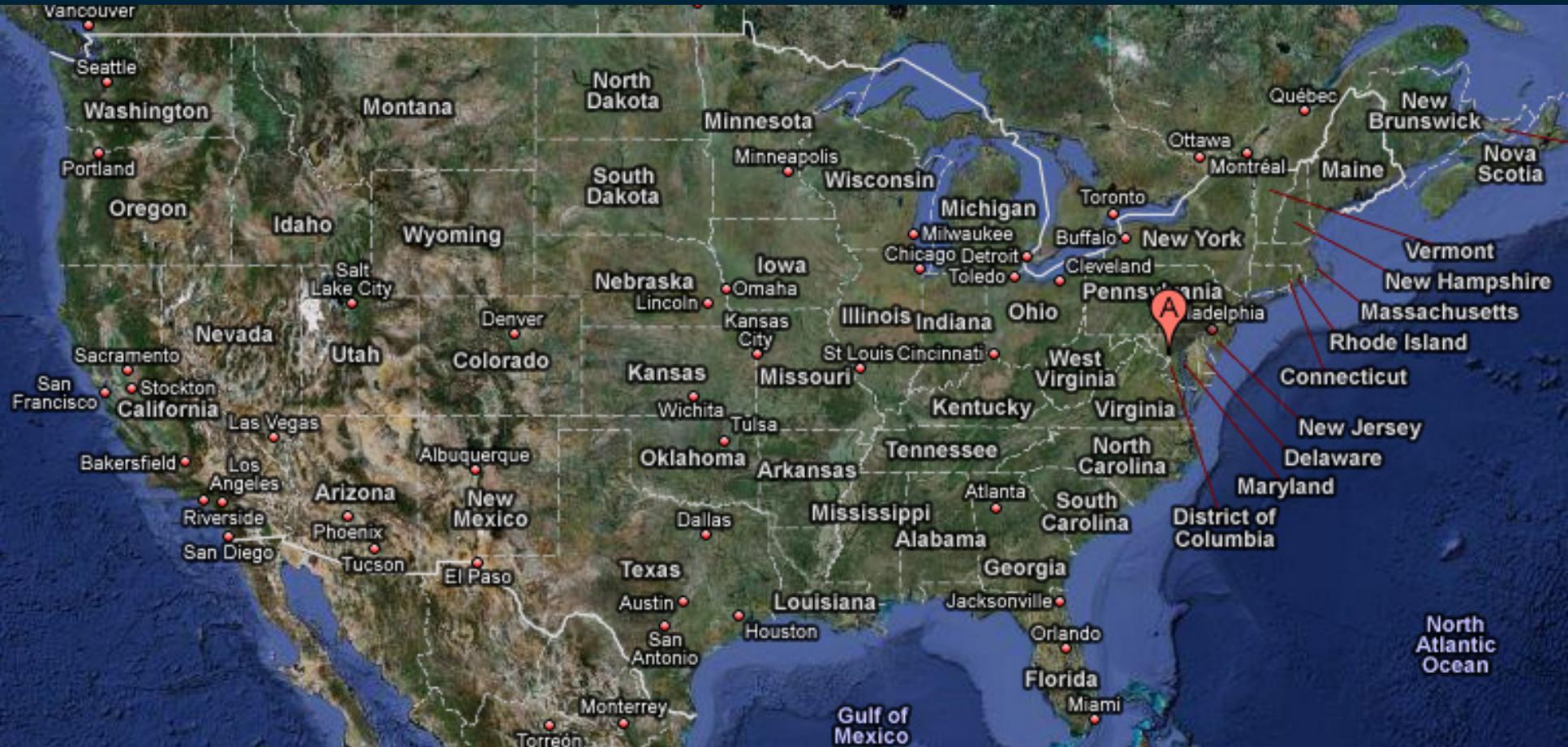


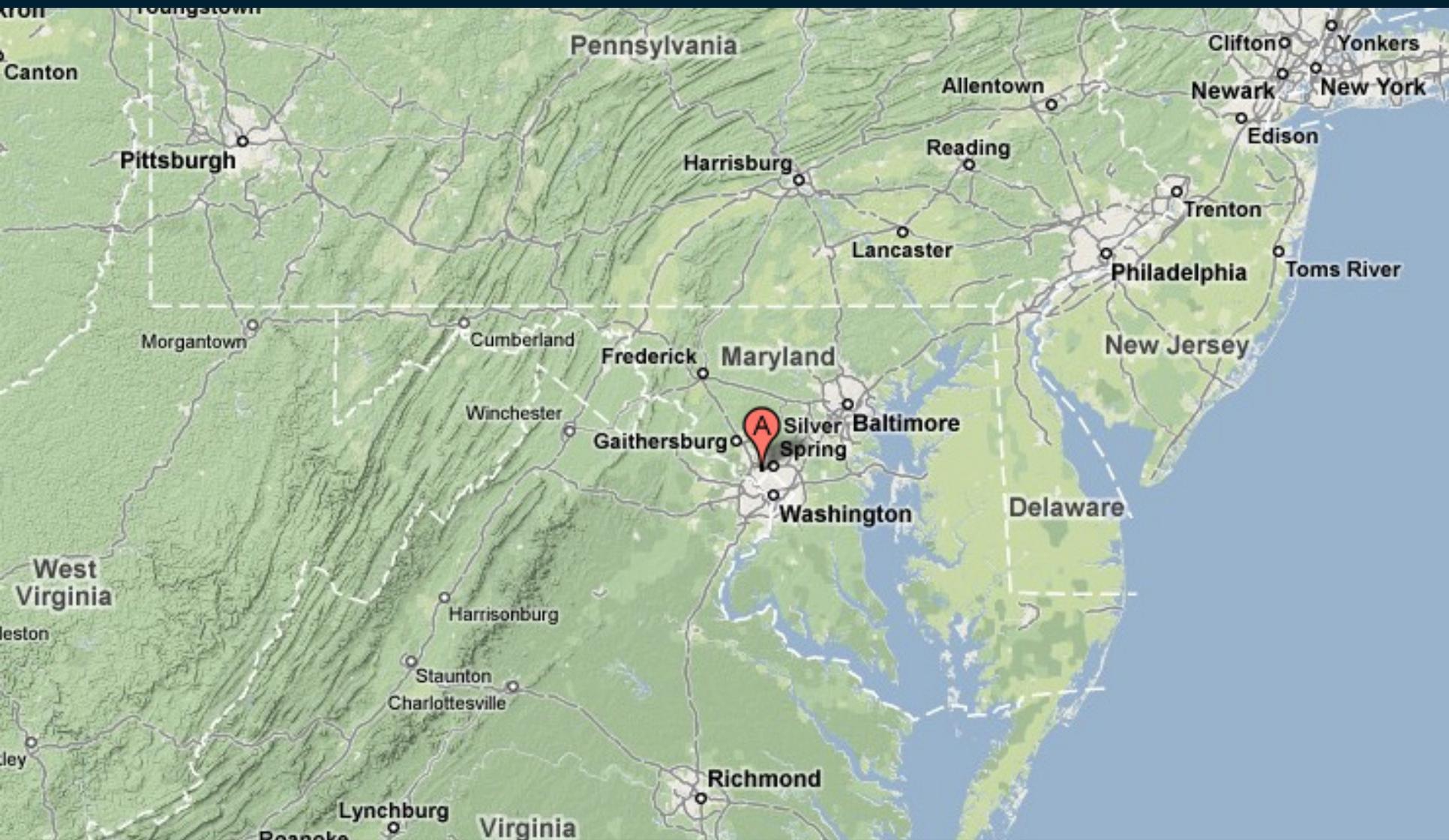
National Institutes of Health

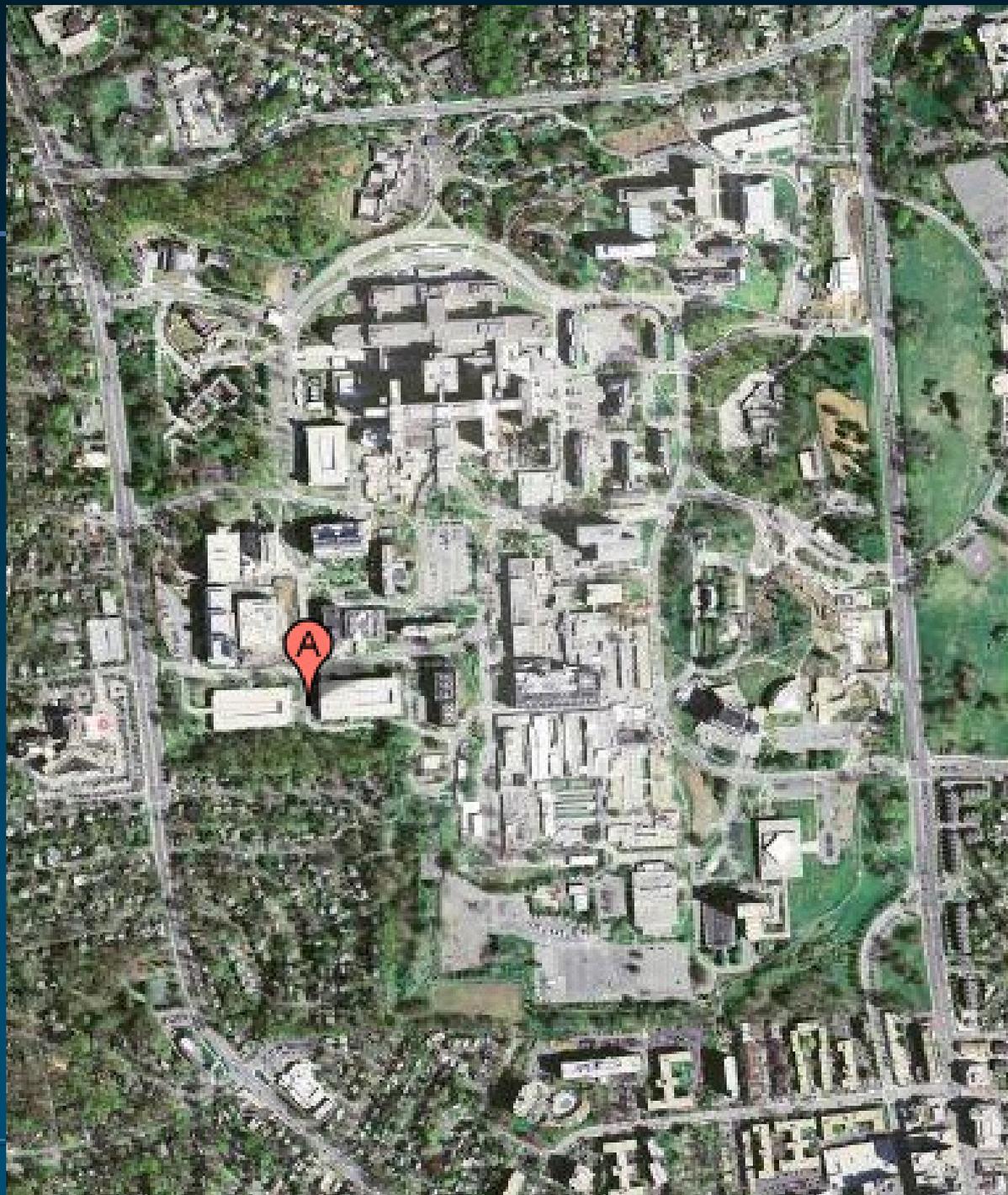


National Institutes of Health

Turning Discovery Into Health



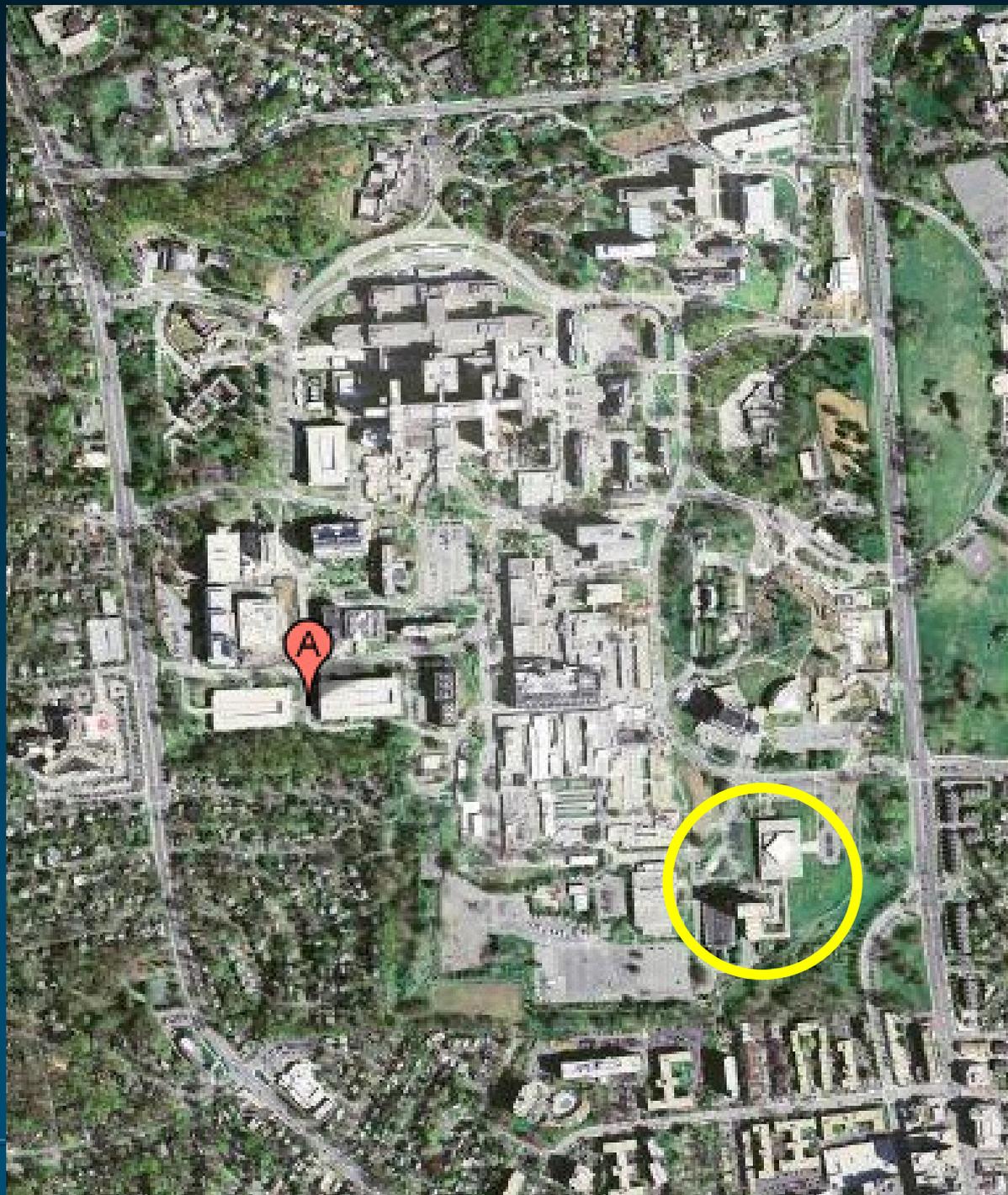




National Institutes of Health

- ◆ NIH seeks to enhance health, lengthen life, and reduce illness and disability
- ◆ Annual budget: ~ \$32 billion
- ◆ 27 Institutes and Centers
 - Institutes (Cancer; Heart, Lung & Blood)
 - Centers (Translational Research; Scientific Review)
- ◆ Intramural – 6,000 researchers at NIH Institutes
- ◆ Extramural – grants for research, research resources, training







National Library of Medicine



U.S. National Library of Medicine

NLM Customer Support



Databases

- PubMed/MEDLINE
- MeSH
- UMLS
- ClinicalTrials.gov
- MedlinePlus
- TOXNET
- Images from the History of Medicine
- Digital Collections
- LocatorPlus
- All NLM Databases & APIs

NIH MedlinePlus
MAGAZINE
Summer 2017

New NIH MedlinePlus Magazine Now Available
"Access Hollywood" host Liz Hernandez shares family experience with Alzheimer's.

1 2 3 4

Find, Read, Learn

- Search biomedical literature
- Find medical terminologies
- Search NLM collections
- Read about diseases
- Learn about drugs
- Explore history
- Find a clinical trial
- Use a medical dictionary
- Find free full-text articles

Explore NLM

- About NLM
- Health Information
- Library Catalog & Services
- History of Medicine
- Online Exhibitions & Digital Projects
- Information for Publishers
- Visit the Library

Research at NLM

- Human Genome Resources
- Biomedical Research & Informatics
- Environmental Health & Toxicology
- Health Services Research & Public Health
- Health Information Technology

NLM for You

- Grants & Funding
- Meaningful Use Tools
- Training & Outreach
- National Network of Medical Libraries
- Regional Activities
- Careers @ NLM
- Mobile Gallery

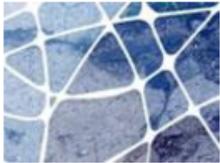
News, Events, Videos

- Funders Reflect on Lessons Learned in Funding International Open Science Prize (08/01/17)
- Jerry Sheehan Appointed Deputy Director of the National Library of Medicine (07/31/17)
- NLM Biomedical Informatics Course to Be Restructured to Help Accelerate Research (07/27/17)

Lister Hill National Center for Biomedical Communications

LHNCBC Research Areas

Selected Projects



Clinical Data Standards & Electronic Medical Records

Through R&D in standard clinical terminologies and interoperability across clinical information systems and NLM resources, LHNCBC advances user-tailored information retrieval.

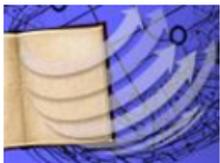
Clinical Vocabulary Standards
InfoBot
Medical Ontology Research



Collaboration Technologies & Mobile Health Applications

This LHNCBC R&D enables remote collaboration, education, training, and access to NLM information resources and disaster aids anytime, anywhere, and from devices like smart phones.

People Locator for Disasters
Remote Virtual Dialog System
Virtual Microscope



Document Processing

LHNCBC conducts R&D in text and data mining, machine learning, electronic preservation and on-line access for multi-media, print-only, and centuries-old biomedical documents.

Medical Article Record System
Interactive Publications
Turning The Pages



Health Information Resources

R&D staff at LHNCBC are developing and enhancing large, complex information systems to meet new needs in health information, biomedical research, and historical preservation.

Consumer Health Question Answering
Genetics Home Reference
Open-i



Health Information Resources

R&D staff at LHNCBC are developing and enhancing large, complex information systems to meet new needs in health information, biomedical research, and historical preservation.

Consumer Health Question Answering
Genetics Home Reference
Open-i



Image Processing & Visualization

For use in biomedical education and the diagnosis and treatment of diseases, LHNCBC conducts R&D in the analysis, presentation, and retrieval of images and the creation of visualizations.

Computer-aided TB Screening on Chest X-rays
Imaging Tools for Cancer Research
Visible Human Project



Natural Language Processing

LHNCBC's NLP R&D improves search and retrieval and facilitates discovery through advances in analyzing biomedical texts, graphical presentation of results, and multi-language search.

Lexical Systems & Tools
Automated Indexing Research
Semantic Knowledge Representation

**Medical Informatics
Training Program**

References

- ◆ Cui L, Zhu W, Tao S, Case JT, Bodenreider O, Zhang GQ. Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in SNOMED CT. J Am Med Inform Assoc. 2017;24(4):788-798. PMID: 28339775

Journal of the American Medical Informatics Association, 24(4), 2017, 788–798

doi: 10.1093/jamia/ocw175

Advance Access Publication Date: 19 February 2017

Research and Applications

AMIA
INFORMATICS PROFESSIONALS LEADING THE WAY

OXFORD

Research and Applications

Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in SNOMED CT

Licong Cui,^{1,2} Wei Zhu,² Shiqiang Tao,^{2,3} James T Case,⁴ Olivier Bodenreider,⁴ and Guo-Qiang Zhang^{2,3}

¹Department of Computer Science, University of Kentucky, Lexington, KY, USA, ²Institute for Biomedical Informatics, University of Kentucky,

³Division of Biomedical Informatics, College of Medicine, University of Kentucky and ⁴National Library of Medicine, Bethesda, MD, USA



SNOMED Clinical Terms

SNOMED
International

Leading healthcare
terminology, worldwide

SNOMED CT Characteristics

- ◆ Developed by SNOMED International
 - Consortium of 30 member countries
- ◆ Largest clinical terminology in the world
 - ~320,000 active concepts
 - ~ 1 million terms (“descriptions”)
- ◆ Major organizing principles
 - Logical definitions (incomplete: many primitives)
 - Built using description logics (EL++)



SNOMED CT Top level

- ▼ ● SNOMED CT Concept
 - ▶ ● Body structure (body structure)
 - ▶ ● Clinical finding (finding)
 - ▶ ● Environment or geographical location (environment / location)
 - ▶ ● Event (event)
 - ▶ ● Observable entity (observable entity)
 - ▶ ● Organism (organism)
 - ▶ ● Pharmaceutical / biologic product (product)
 - ▶ ● Physical force (physical force)
 - ▶ ● Physical object (physical object)
 - ▶ ● Procedure (procedure)
 - ▶ ● Qualifier value (qualifier value)
 - ▶ ● Record artifact (record artifact)
 - ▶ ● Situation with explicit context (situation)
 - ▶ ● SNOMED CT Model Component (metadata)
 - ▶ ● Social context (social concept)
 - ▶ ● Special concept (special concept)
 - ▶ ● Specimen (specimen)
 - ▶ ● Staging and scales (staging scale)
 - ▶ ● Substance (substance)

SNOMED CT Example

Parents

- ▶ ☰ Operation on appendix (procedure)
- ▶ ☰ Partial excision of large intestine (procedure)

☰ Appendectomy (procedure) ☆ ↗

SCTID: 80146002

80146002 | Appendectomy (procedure) |

Appendectomy
Excision of appendix
Appendicectomy
Appendectomy (procedure)

Procedure site - Direct → Appendix structure
Method → Excision - action

Children (8)

- ☰ Appendectomy with drainage (procedure)
- ▶ ☰ Emergency appendectomy (procedure)
- ● Excision of appendiceal stump (procedure)
- ● Excision of ruptured appendix by open approach (procedure)
- ● Incidental appendectomy (procedure)
- ● Interval appendectomy (procedure)
- ▶ ☰ Laparoscopic appendectomy (procedure)
- ☰ Non-emergency appendectomy (procedure)

SNOMED CT Challenges

◆ Legacy

- Many primitive concepts
- Not amenable to automatic DL classification

◆ Maintenance

- Developed by many human editors
- Error prone

◆ Quality assurance

- Difficult due to its size
- Ontology design patterns (“concept model”)
 - Difficult to apply retrospectively



Quality assurance approaches

Quality assurance approaches

- ◆ Three types of QA approaches applied to SNOMED CT by researchers
 - Lexical
 - Structural
 - Semantic

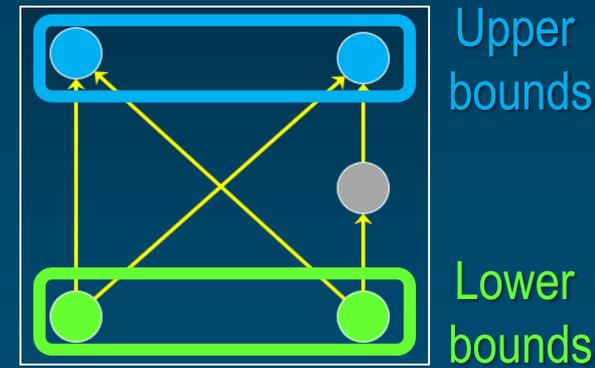
Lattice-based structural QA

◆ Lattice

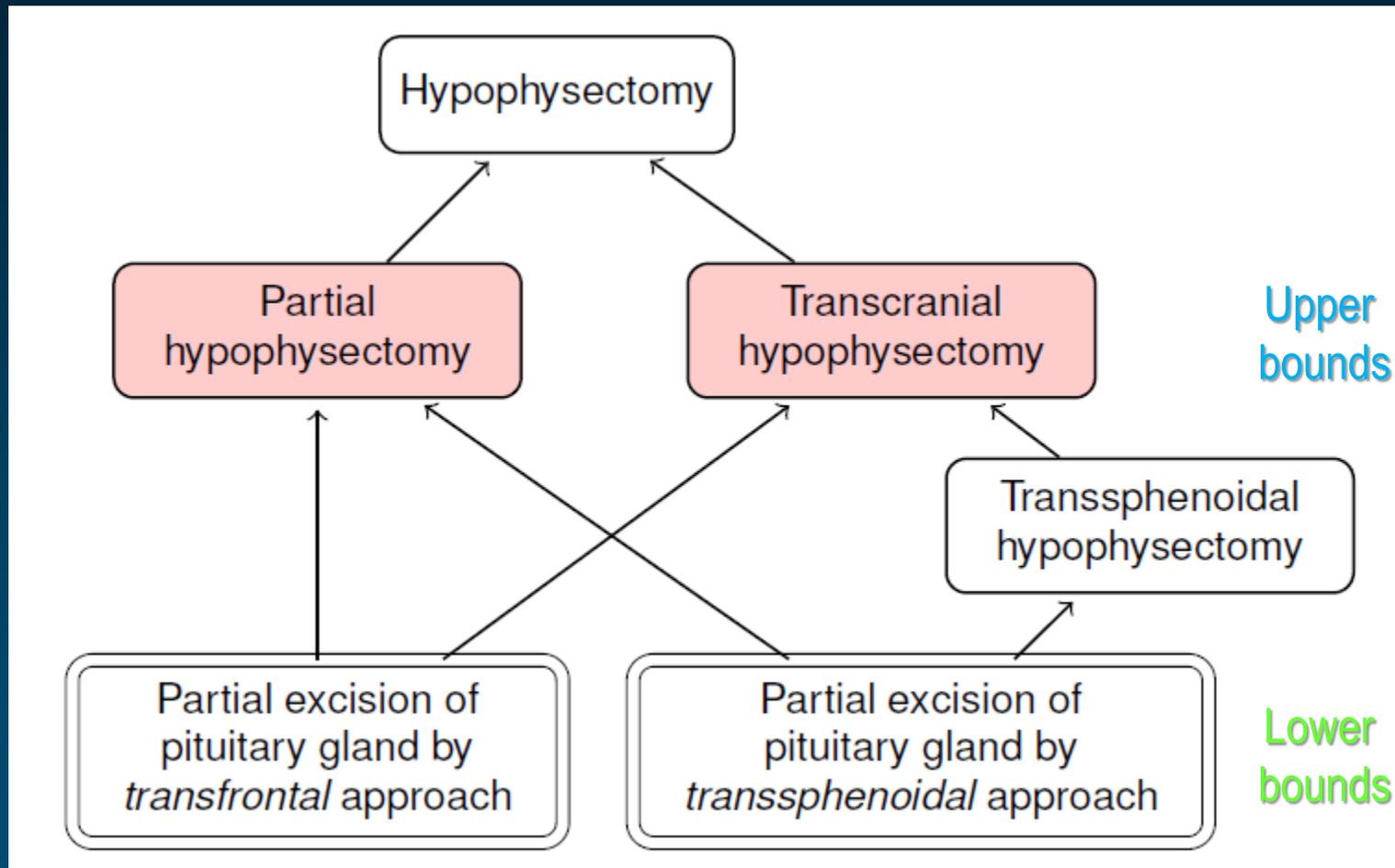
- Specific type of directed acyclic graph (DAG)
- Any two nodes have a unique maximal common descendant, as well as a unique minimal common ancestor

◆ Non-lattice fragments are often indicative of a problem in ontology construction

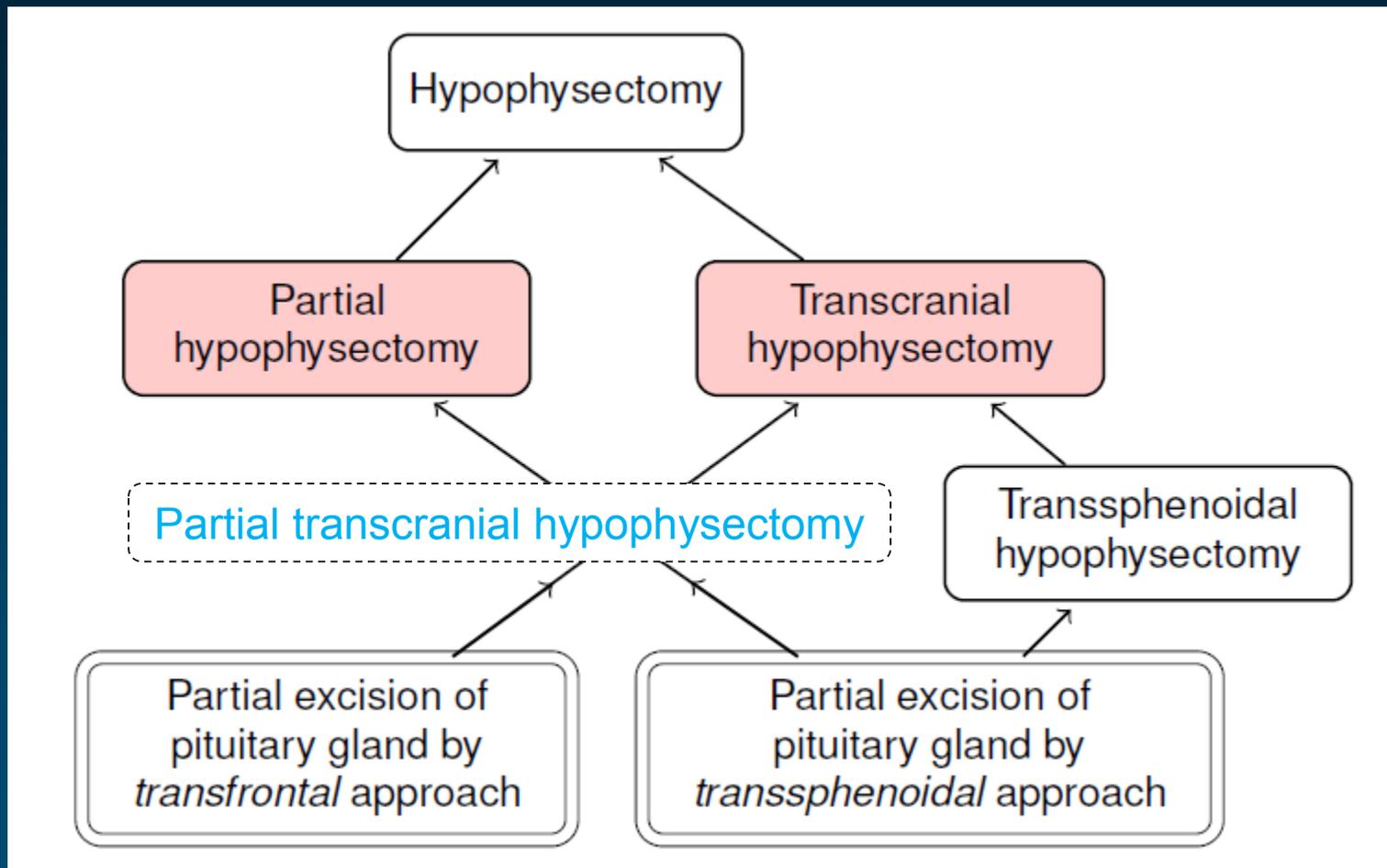
- Missing hierarchical relation
- Missing intermediary concept



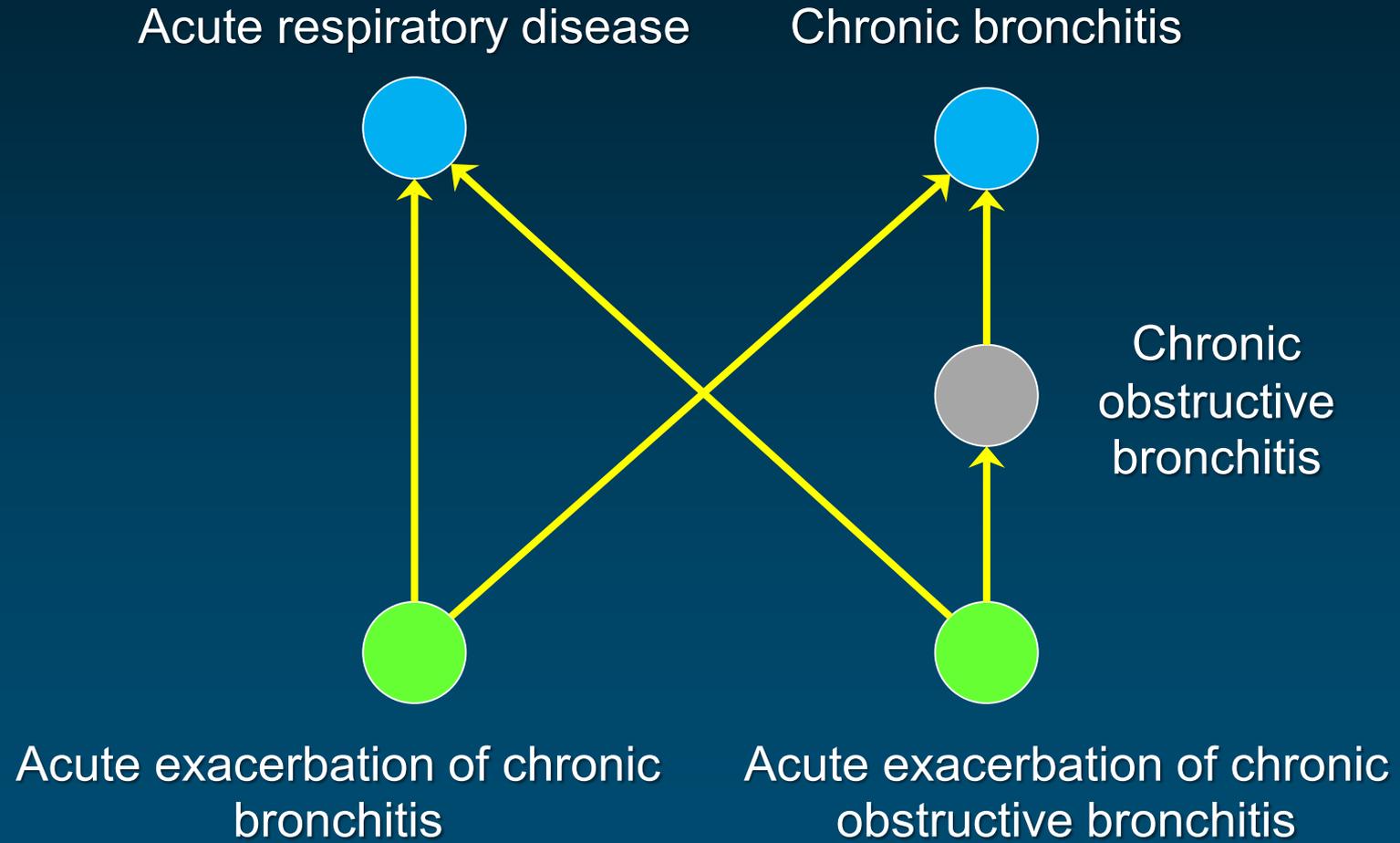
Example of non-lattice fragment



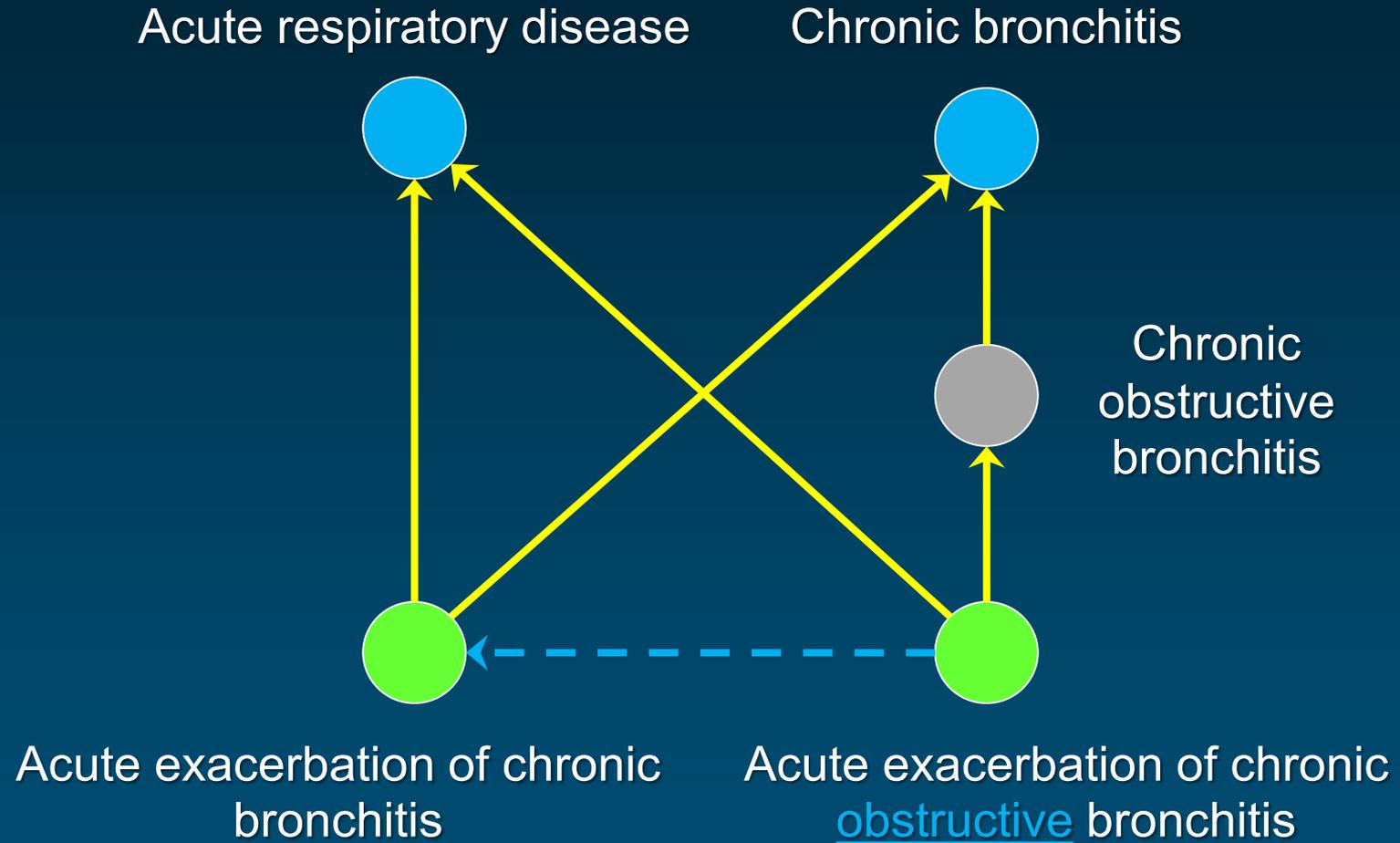
Missing intermediary concept



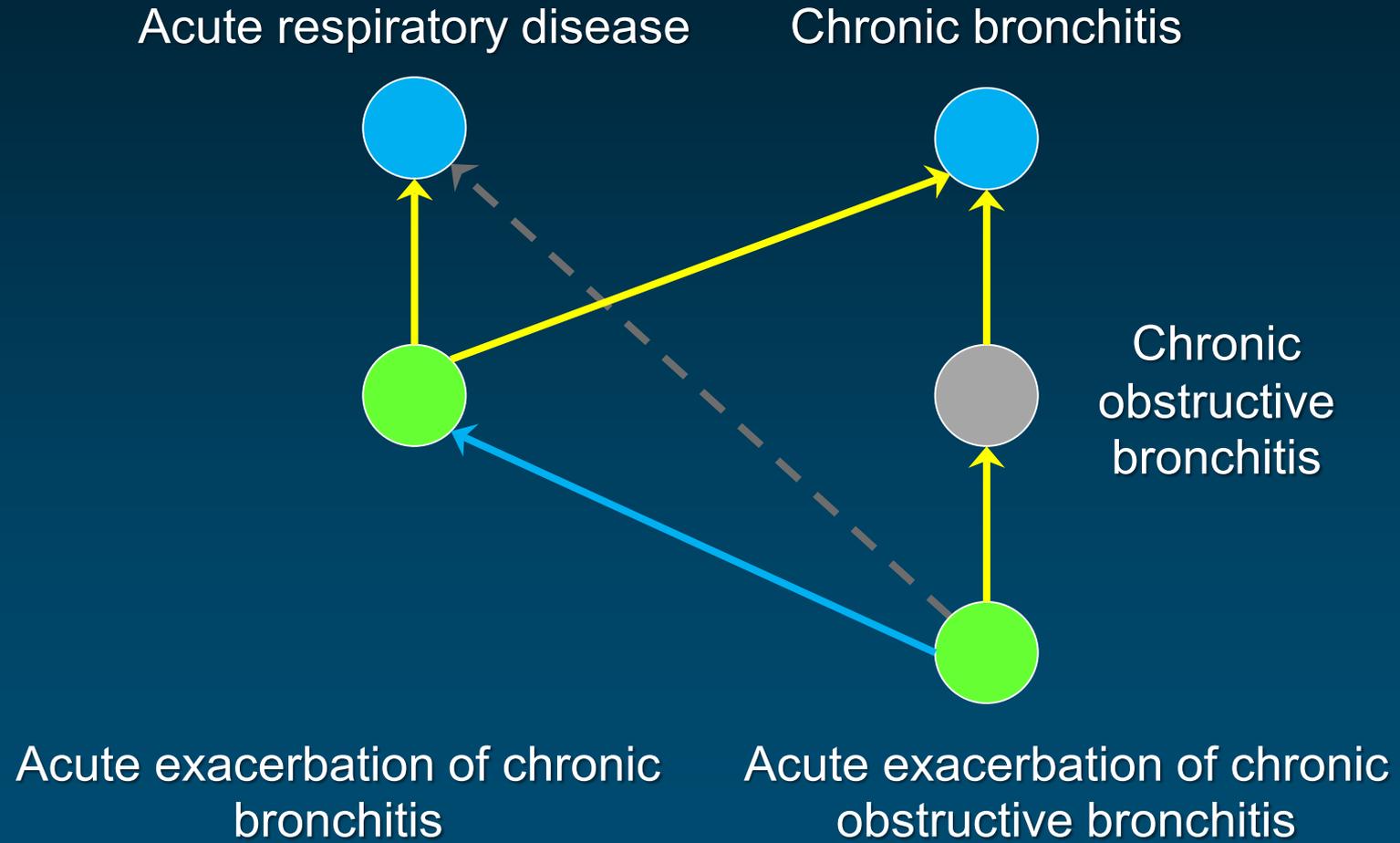
Non-lattice fragment in SNOMED CT



Missing hierarchical relation

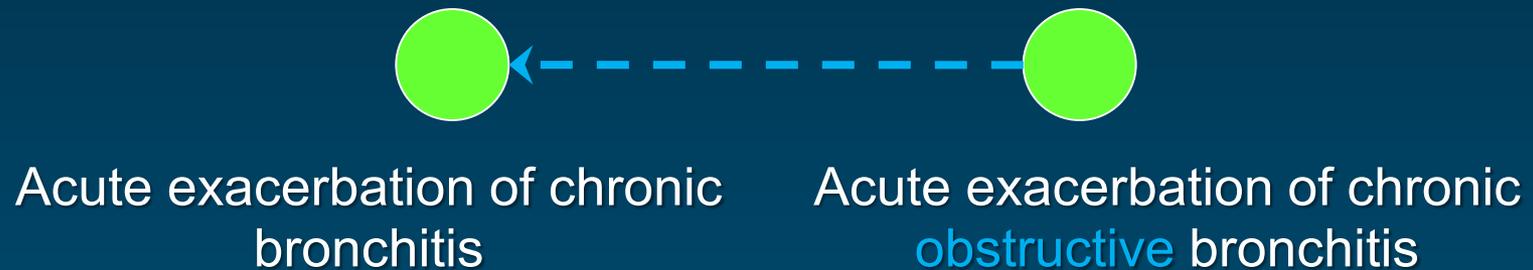


~~Non-lattice~~ fragment in SNOMED CT



QA based on lexical patterns

- ◆ Lexical differences among terms are often indicative of semantic relations among them



Suggested missing hierarchical relations

Child name	Parent name
Alveolar bone graft <u>to mandible</u>	Alveolar bone graft
<u>Basal cell carcinoma</u> <u>of skin</u> of lip	Carcinoma of lip
Carcinoma <u>in situ of</u> palate	Palate carcinoma
Chronic <u>bacterial</u> otitis externa	Chronic otitis externa
<u>Congenital</u> vascular anomaly of eyelid	Vascular anomaly of eyelid
<u>Electrocoagulation</u> of retina <u>for</u> repair <u>of</u> tear	Repair of retina

Objectives

- ◆ To combine lexical and structural QA approaches to automatically and precisely identifying missing hierarchical relations and missing concepts in SNOMED CT
- ◆ To suggest remediation for such inconsistencies
- ◆ Materials: September 2015 version of SNOMED CT (U.S. edition)

Methods & Results

Overview

- ◆ Identifying non-lattice pairs and subgraphs
- ◆ Identifying lexical patterns indicative of missing concepts and relations
- ◆ Analyzing non-lattice subgraphs with lexical patterns
- ◆ Evaluation

Identifying non-lattice pairs and subgraphs

- ◆ Hadoop-based technique
- ◆ 30 hours to analyze all pairs of SNOMED CT concepts
- ◆ Aggregation of non-lattice pairs with the same shared ancestors into non-lattice subgraphs
 - Smaller subgraphs contained in larger subgraphs
- ◆ 631,006 non-lattice pairs
- ◆ 171,011 non-lattice subgraphs
 - Focus on small subgraphs



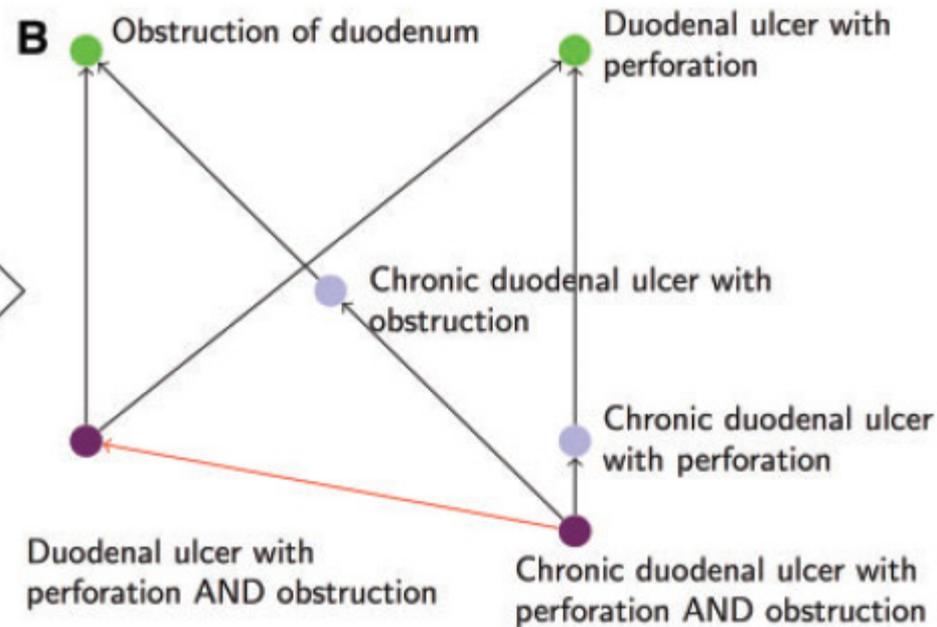
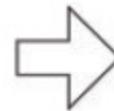
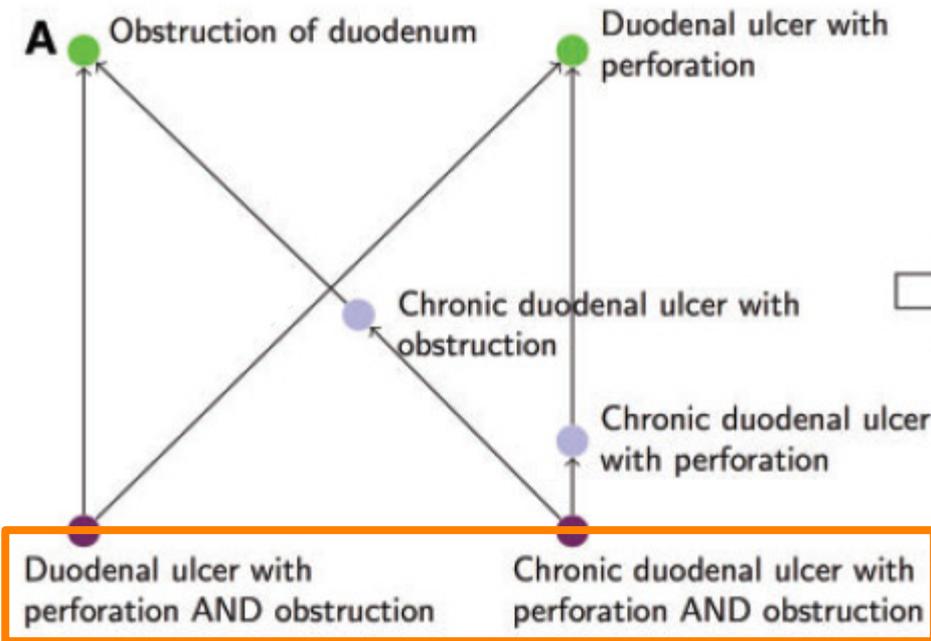
Lexical patterns (1) Containment

- ◆ The set of words for one concept in the upper (resp. lower) bounds is contained in the set of words for another concept in the upper (resp. lower) bounds
- ◆ Suggests a *missing hierarchical relation* between concepts in the upper (resp. lower) bounds
- ◆ 736 small non-lattice subgraphs with this pattern

Lexical patterns (1) Containment

Non-lattice subgraph

Suggested remediation



Duodenal ulcer with perforation AND obstruction \supset Chronic duodenal ulcer with perforation AND obstruction

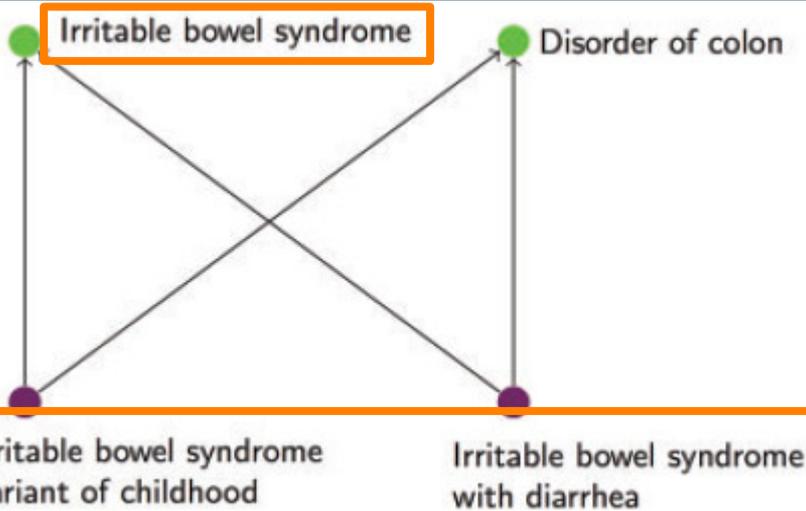
Lexical patterns (2) Intersection

- ◆ The intersection of sets of words for concepts in the lower bounds is equal to the set of words for some concept in the upper bounds
- ◆ Suggests a *missing hierarchical relation* between concepts in the upper bounds
- ◆ 1085 small non-lattice subgraphs with this pattern

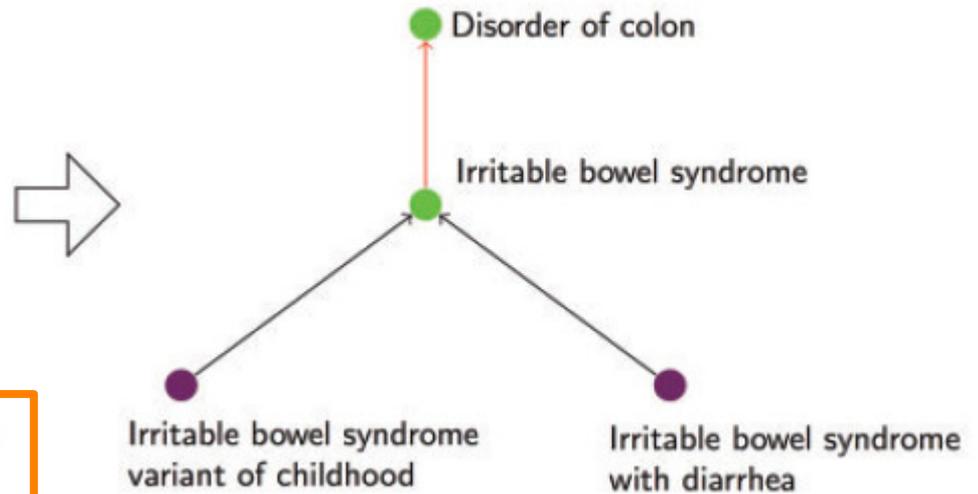


Lexical patterns (2) Intersection

Non-lattice subgraph



Suggested remediation



Irritable bowel syndrome

Irritable bowel syndrome
variant of childhood



Irritable bowel syndrome
with diarrhea

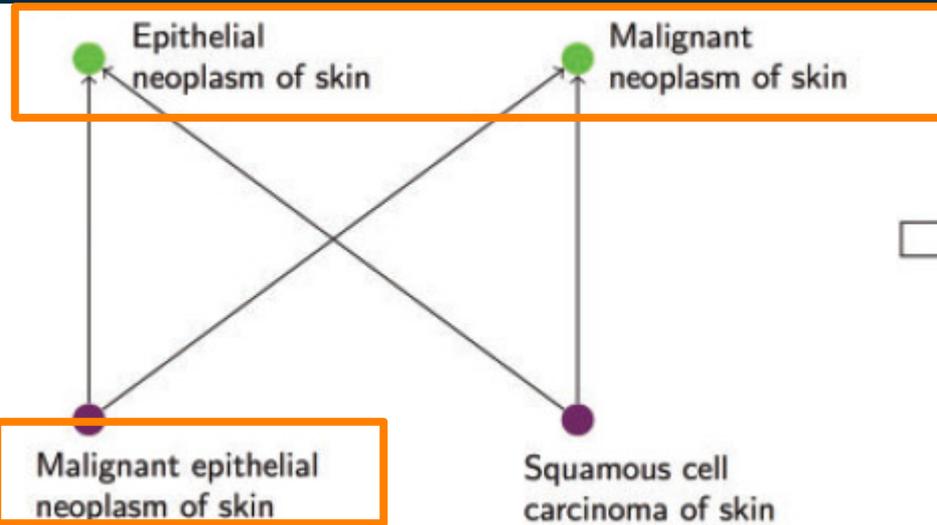
Lexical patterns (3) Union

- ◆ The union of the sets of words for concepts in the upper bounds is equal to the set of words for some concept in the lower bounds
- ◆ Suggests a *missing hierarchical relation* between concepts in the lower bounds
- ◆ 164 small non-lattice subgraphs with this pattern

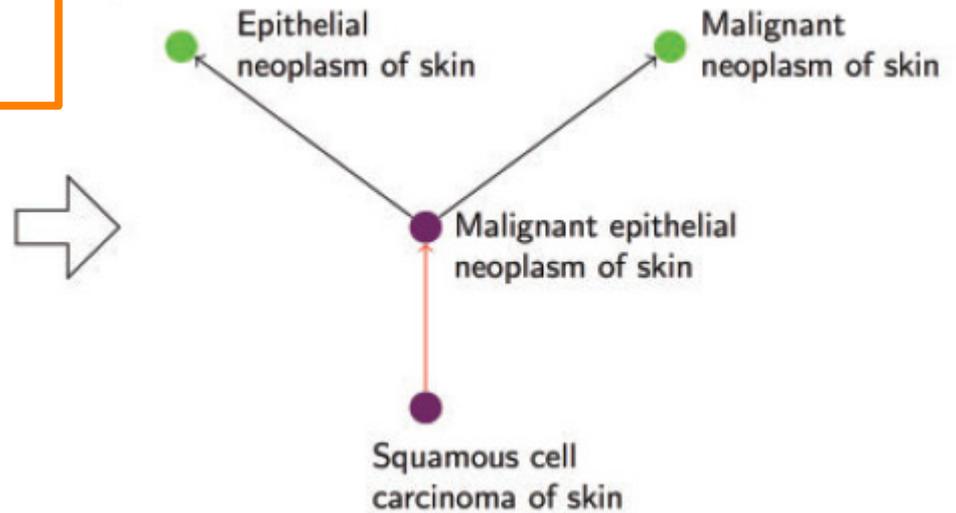


Lexical patterns (3) Union

Non-lattice subgraph



Suggested remediation



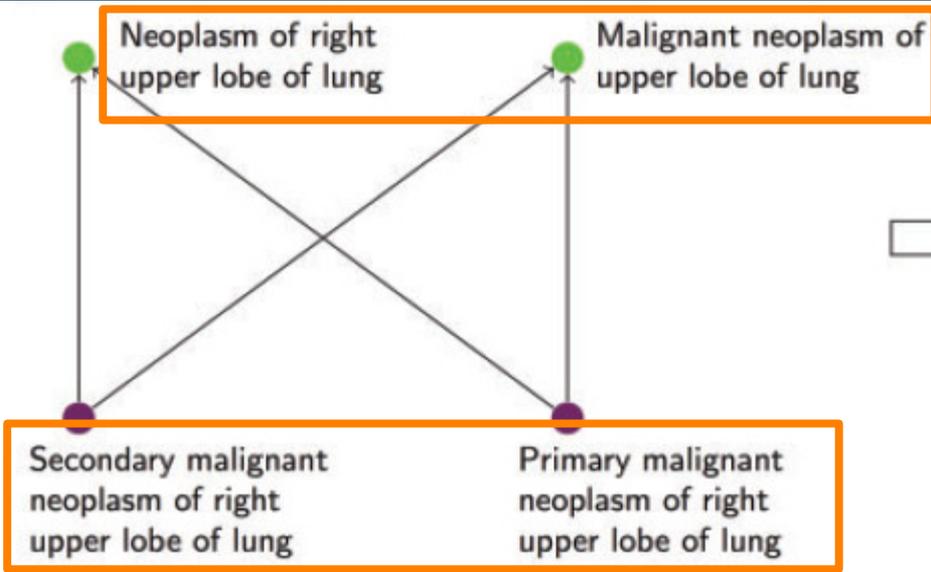
Epithelial neoplasm of skin \cup *Malignant* neoplasm of skin
Malignant epithelial
neoplasm of skin

Lexical patterns (4) Union-Intersection

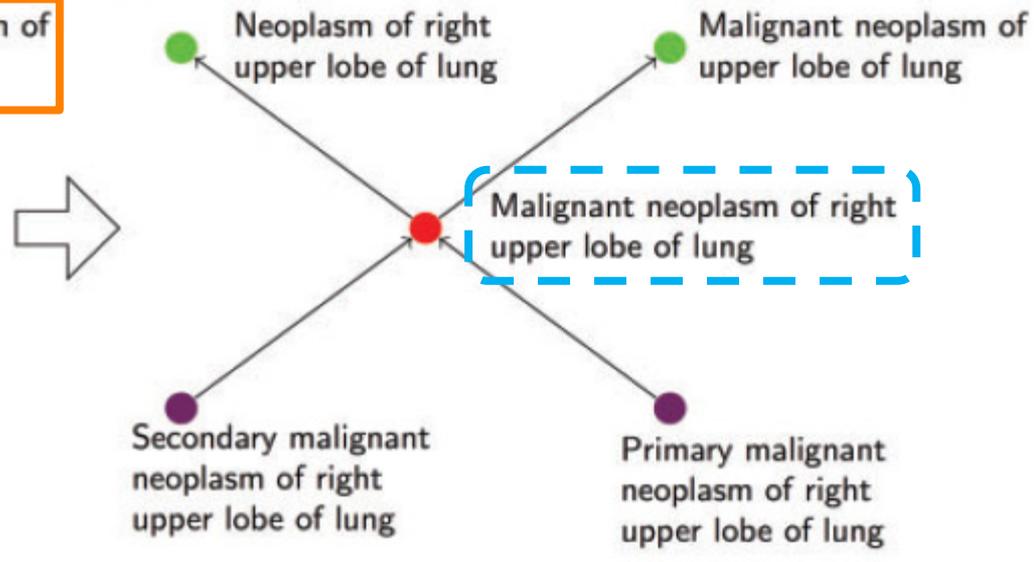
- ◆ The union of the sets of words for concepts in the upper bounds is equal to the intersection of sets of words for concepts in the lower bounds
- ◆ Suggests a *missing intermediary concept* between the upper bounds and the lower bounds
- ◆ 61 small non-lattice subgraphs with this pattern

Lexical patterns (4) Union-Intersection

Non-lattice subgraph



Suggested remediation



Neoplasm of *right* upper lobe of lung



Malignant neoplasm of upper lobe of lung



Secondary malignant neoplasm of right upper lobe of lung



Primary malignant neoplasm of right upper lobe of lung

Evaluation

- ◆ 59 subgraphs independently reviewed by 2 experts after triaging
 - Differences resolved by discussion
- ◆ All contained errors – 61 errors
 - Missing hierarchical relation: 59
 - Missing intermediary concept: 2
- ◆ Lexical patterns
 - Containment: 34; Intersection: 14; Union: 8; U/I: 3
- ◆ Suggested remediation
 - Accepted for 53 subgraphs
 - Rejected for 6 subgraphs (deeper modeling issues)

Discussion

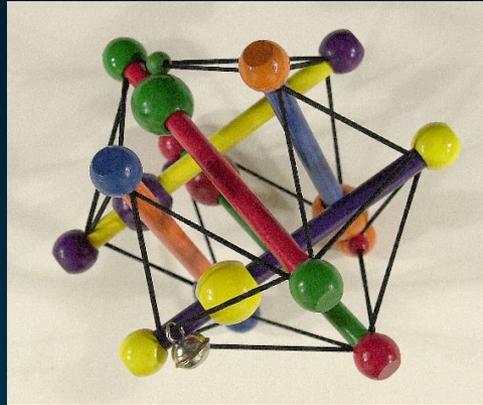
Significance

- ◆ Most terminology QA techniques merely identify potential errors
- ◆ Our approach
 - Identified unreported errors
 - Confirmed by experts
 - Suggested appropriate remediation in many cases
- ◆ Should greatly facilitate error correction by the developers of SNOMED CT
- ◆ Scalable and applicable to other terminologies

Limitations and future work

- ◆ Suggested remediation (e.g., to add missing hierarchical relations) is based on the inferred concept hierarchy of SNOMED CT
 - Does not address the root cause (e.g., incomplete/inaccurate logical definition)
 - Root cause needs to be addressed by the SNOMED CT editors
- ◆ Only 4 lexical patterns considered
 - Could be refined with additional patterns
- ◆ Only used the preferred terms
 - Could also use synonyms





Medical Ontology Research

Contact: olivier@nlm.nih.gov

Web: <http://mor.nlm.nih.gov>



Olivier Bodenreider

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