# CS 626 Large Scale Data Science

## Lecture 6 – MapReduce Examples

Licong Cui
licong.cui@uky.edu

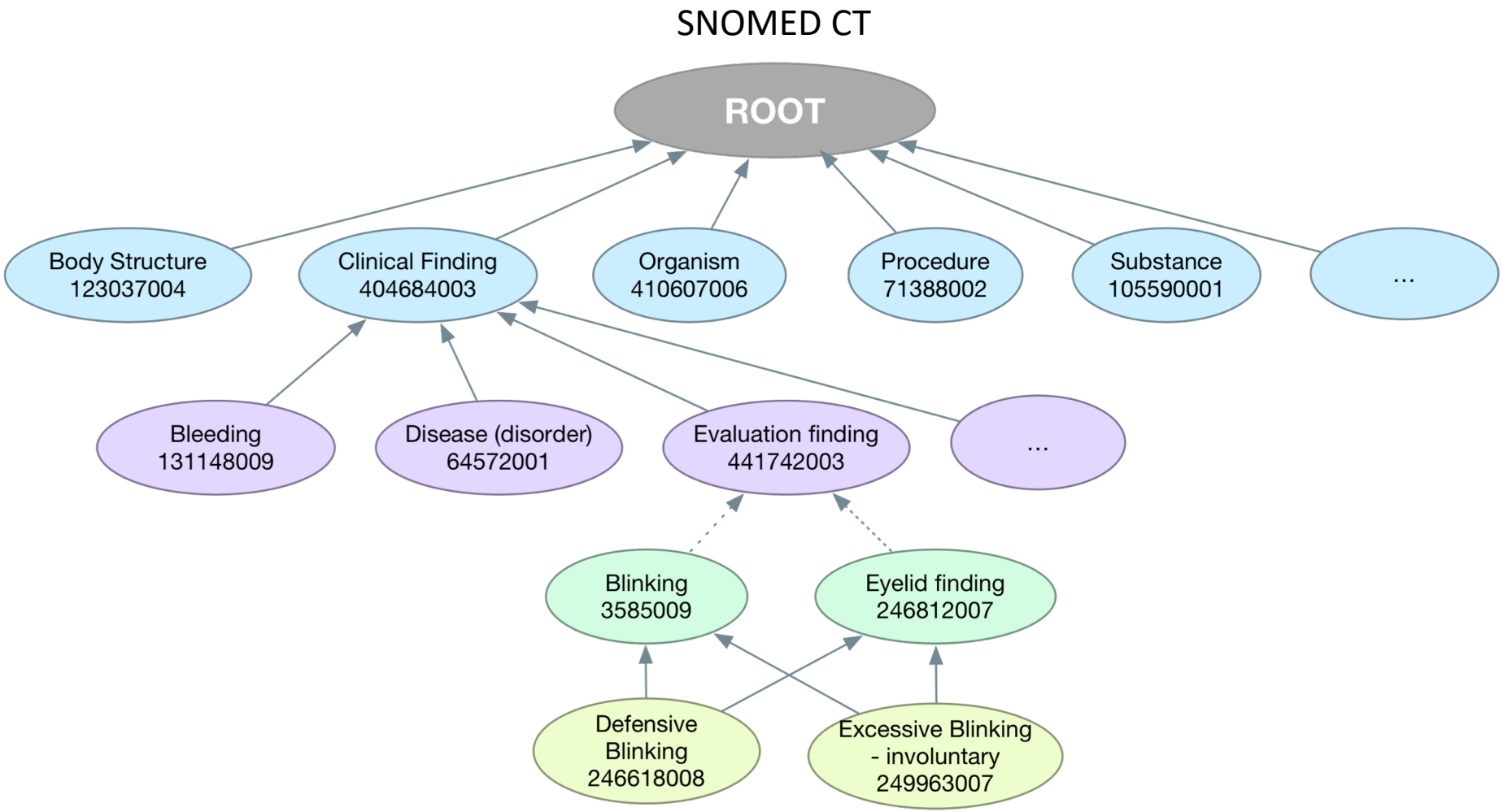
# Biomedical Ontology Quality Assurance (OQA)

# Outline

- ✓ Biomedical Ontology Quality Assurance
  - ➢ Cross-Ontology Hierarchical Relationship Examination (COHeRE)[1]
  - ➢ Non-lattice-based Auditing[2,3]

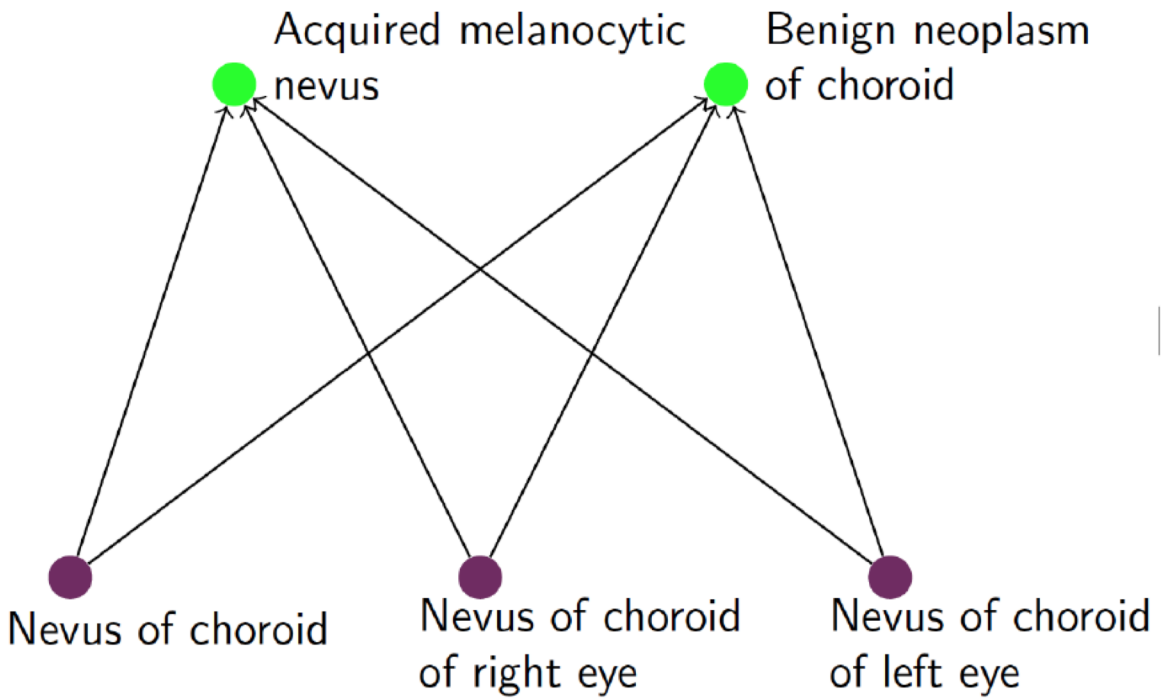

[1] **Cui L**. COHeRE: Cross-Ontology Hierarchical Relation Examination for Ontology Quality Assurance. ***AMIA Annual Symp Proc*** 2015, pp. 456-465.
[2] **Cui L**, Tao S, Zhang GQ. Biomedical Ontology Quality Assurance Using a Big Data Approach. ***ACM Transactions on Knowledge Discovery from Data***, 2016;10(4):41.
[3] **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. ***Journal of the American Medical Informatics Association***, 2017;24(4): 788-798.

# Ontology


SNOMED CT
ROOT
Body Structure
123037004
Clinical Finding
404684003
Organism
410607006
Procedure
71388002
Substance
105590001
...
Bleeding
131148009
Disease (disorder)
64572001
Evaluation finding
441742003
...
Blinking
3585009
Eyelid finding
246812007
Defensive
Blinking
246618008
Excessive Blinking
- involuntary
249963007

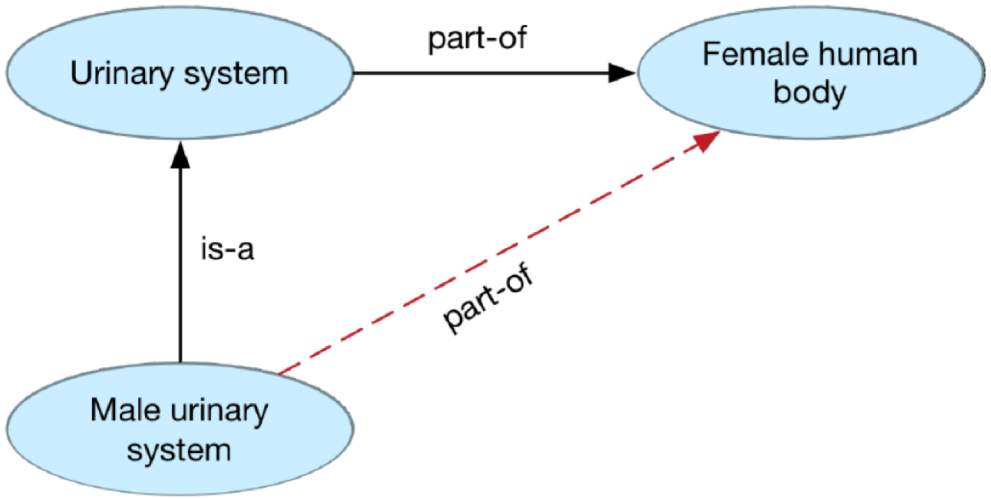
# Ontology Quality Assurance Needs

- A *Non-lattice Subgraph* Example in the SNOMED CT



# Ontology Quality Assurance Needs (cont.)

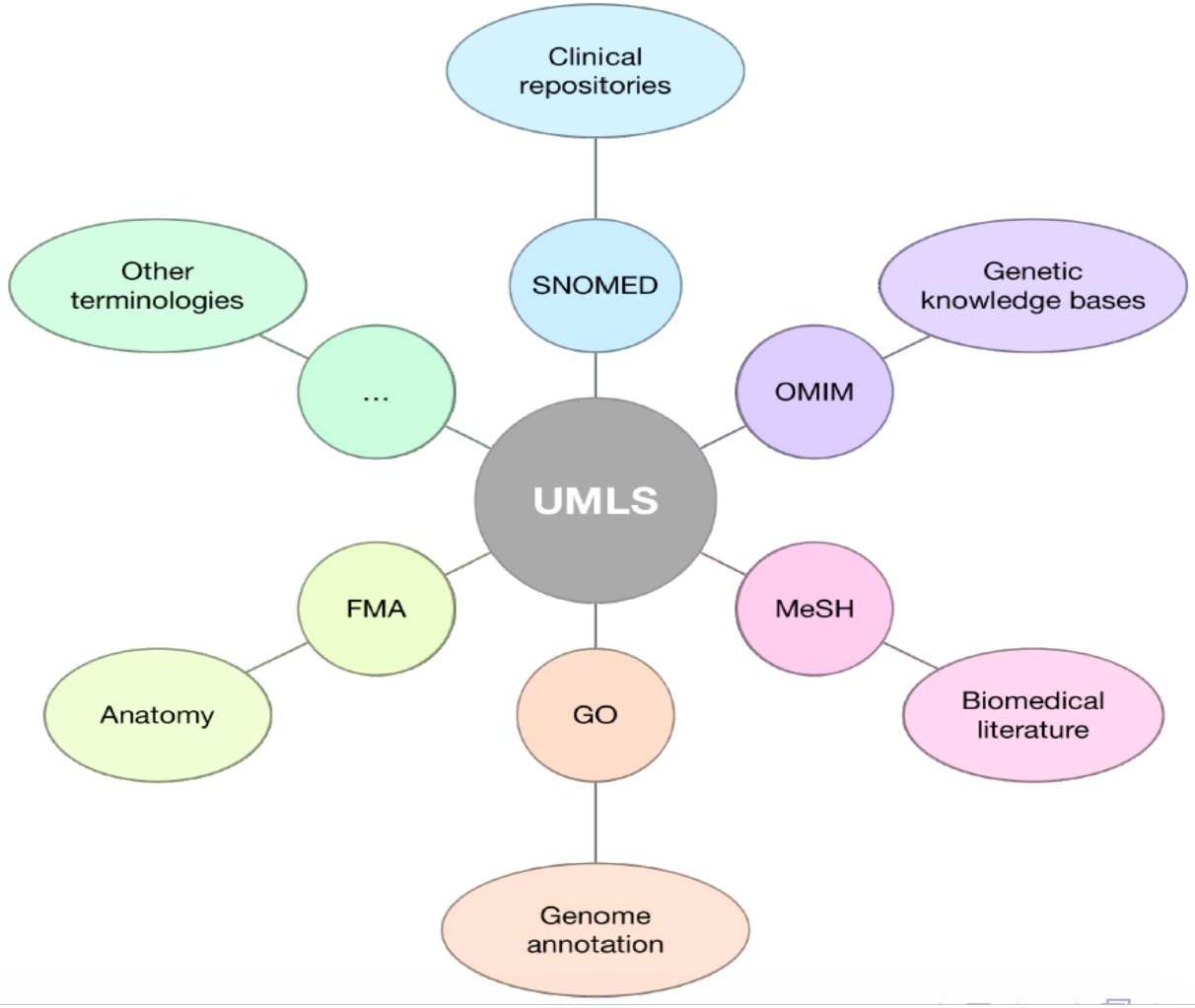
- Example in the Foundational Model of Anatomy (FMA)



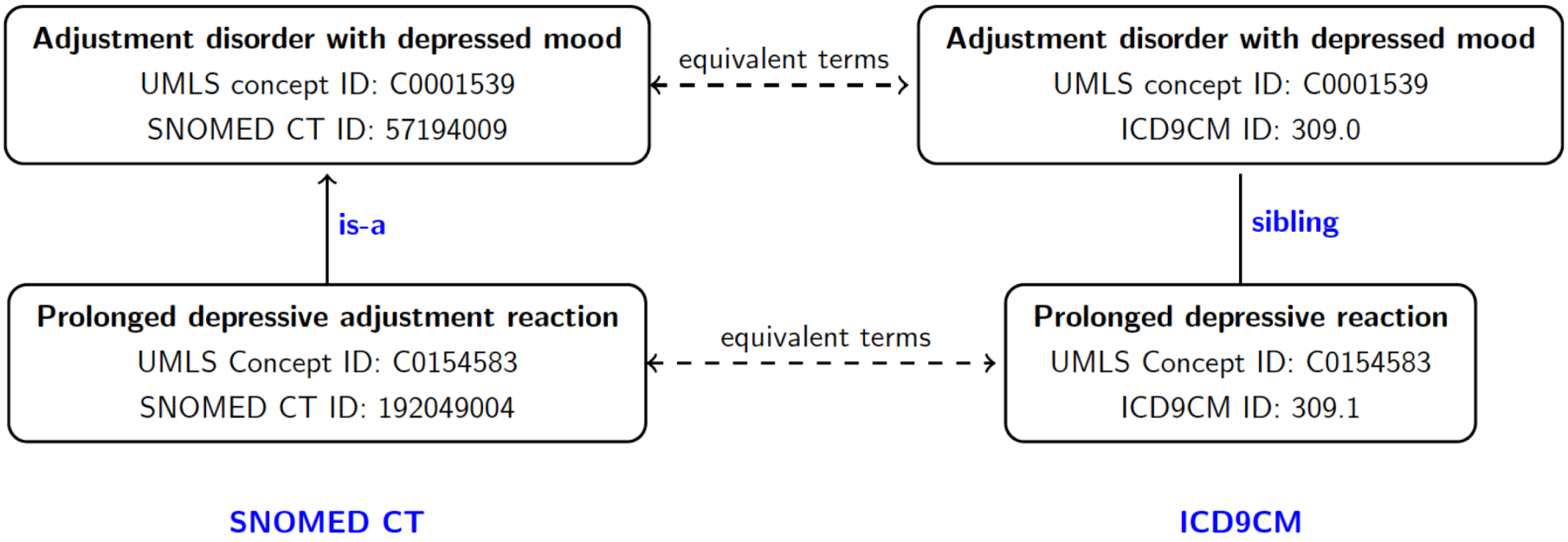
# COHeRE: Cross-Ontology Hierarchical Relationship Examination

# Unified Medical Language System (UMLS)

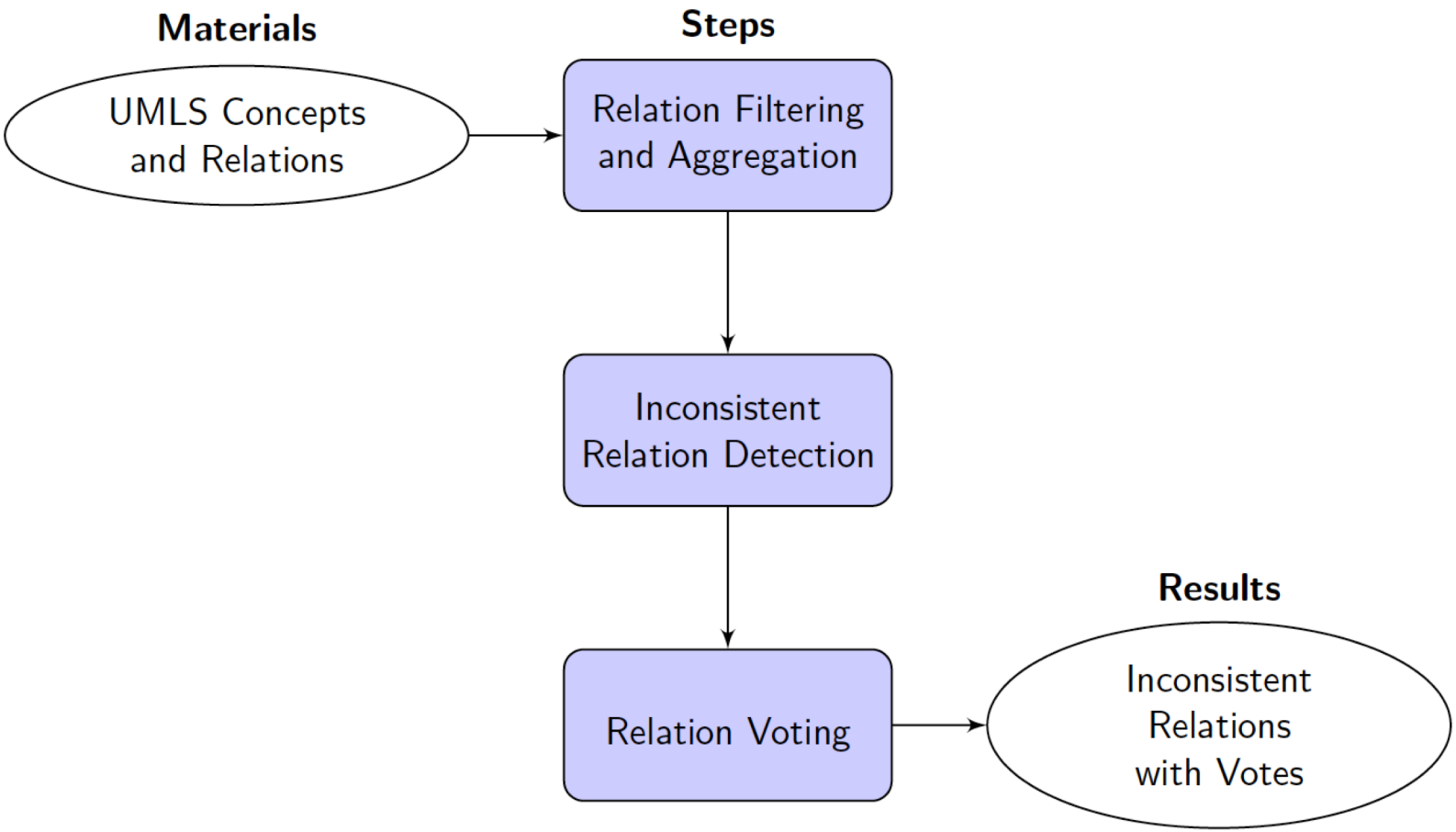
- Over 150 source vocabularies
- Over 2.8 million concepts and 60 million relations



# Example of Inconsistency



# Overview of the Proposed Method COHeRE



# Materials

- UMLS 2014AB release
  - A subset of source vocabularies
- Concepts
  - Concept Unique Identifier (CUI) e.g., C0027051
    - Heart attack
    - Infarction of heart
    - Myocardial infarction
    - Cardiovascular Stroke

| Acronym | Vocabulary Name |
|---|---|
| ATC | Anatomical Therapeutic Chemical Classification System |
| CPT | Current Procedural Terminology |
| FMA | Foundational Model of Anatomy Ontology |
| GO | Gene Ontology |
| ICD10CM | International Classification of Diseases, 10th Edition |
| ICD9CM | International Classification of Diseases, Ninth Revision |
| MDR | Medical Dictionary for Regulatory Activities Terminology |
| MEDLINEPLUS | MedlinePlus Health Topics |
| MSH | Medical Subject Headings |
| MTH | UMLS Metathesaurus |
| NCBI | NCBI Taxonomy |
| NCI | NCI Thesaurus |
| NDFRT | National Drug File - Reference Terminology |
| OMIM | Online Mendelian Inheritance in Man |
| RXNORM | RxNorm Vocabulary |
| SNOMEDCT_US | US Edition of SNOMED CT |
| VANDF | Veterans Health Administration National Drug File |

# Materials

- Relations
  - Relationships

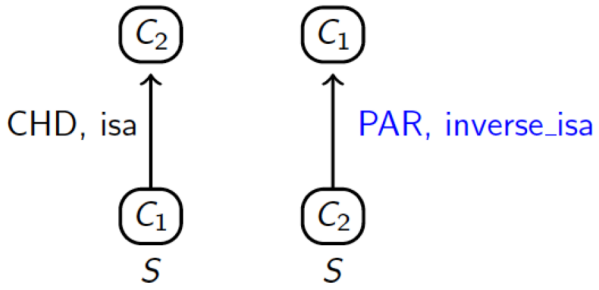
| | Relationship (R) | Relationship attribute (RA) Examples |
|---|---|---|
| SY | synonymy | same as, alias of |
| CHD | has child relationship | is a, part of |
| SIB | has sibling relationship | sibling in is a, sibling in part of |
| RN | has a narrower relationship | tradename of, form of |
| AQ | allowed qualifier | actual outcome of, modifies |
| RO | has other relationship | measured by, measures |

  - $(C_1, C_2, R, RA, S)$
    e.g., (C3853287, C3853286, CHD, is a, SNOMEDCT_US)

# Relation Filtering and Aggregation

- Relation filtering
  - Remove redundancy

    $(C_2, C_1, \text{PAR}, \text{inverse\_isa}, S)$

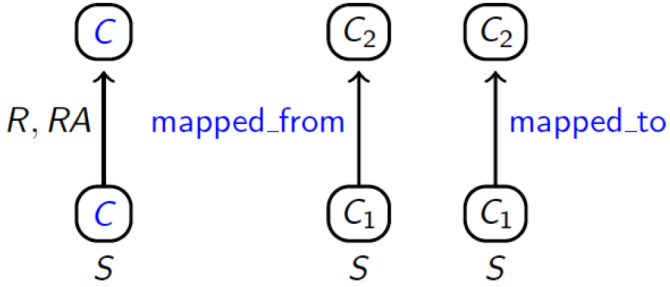



  - Filter relations

    $(C, C, R, RA, S)$

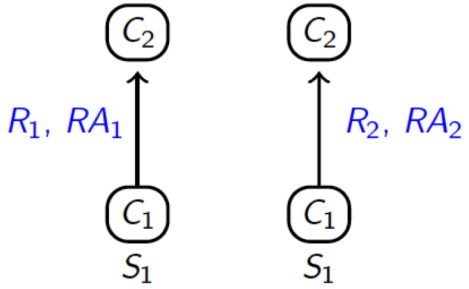
    $(C_1, C_2, R, \text{mapped\_from}, S)$

    $(C_1, C_2, R, \text{mapped\_to}, S)$



# Relation Filtering and Aggregation

- Relation aggregation
  - $(C_1, C_2) \rightarrow (R_1, RA_1, S_1) \mid (R_2, RA_2, S_2) \mid \ldots \mid (R_n, RA_n, S_n)$
  - Filter out multiply related concept pairs in a single source vocabulary



# Relation Filtering and Aggregation Using MapReduce

- MapReduce
  - Map:

$$(C_1, C_2, R, RA, S) \xRightarrow{\text{filtering}} (C_1, C_2), (R, RA, S)$$

  - Reduce:

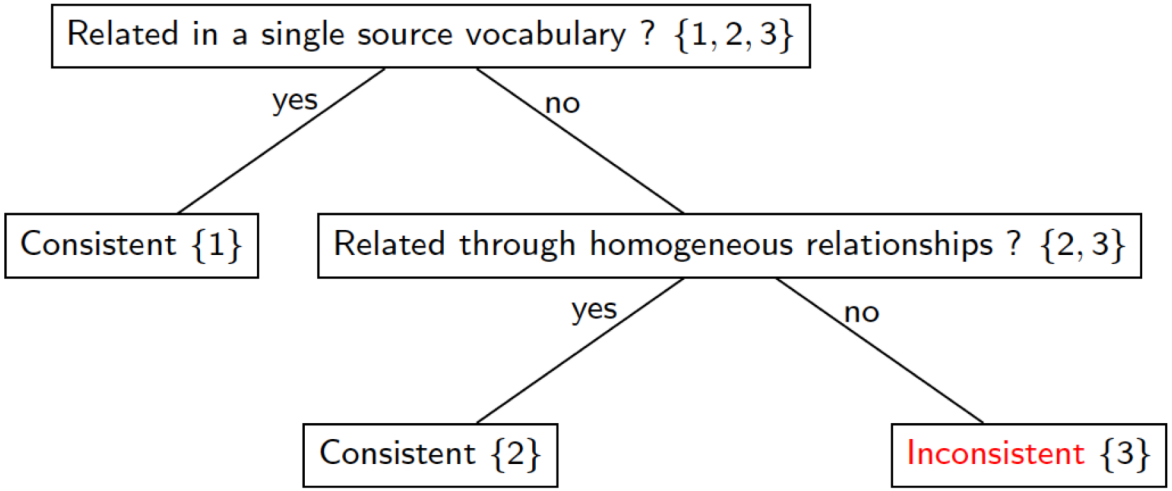
$$\{(C_1, C_2), (R_i, RA_i, S_i)\}$$

$$\Downarrow \text{ aggregation}$$

$$(C_1, C_2) \rightarrow (R_1, RA_1, S_1) \mid (R_2, RA_2, S_2) \mid \ldots \mid (R_n, RA_n, S_n)$$

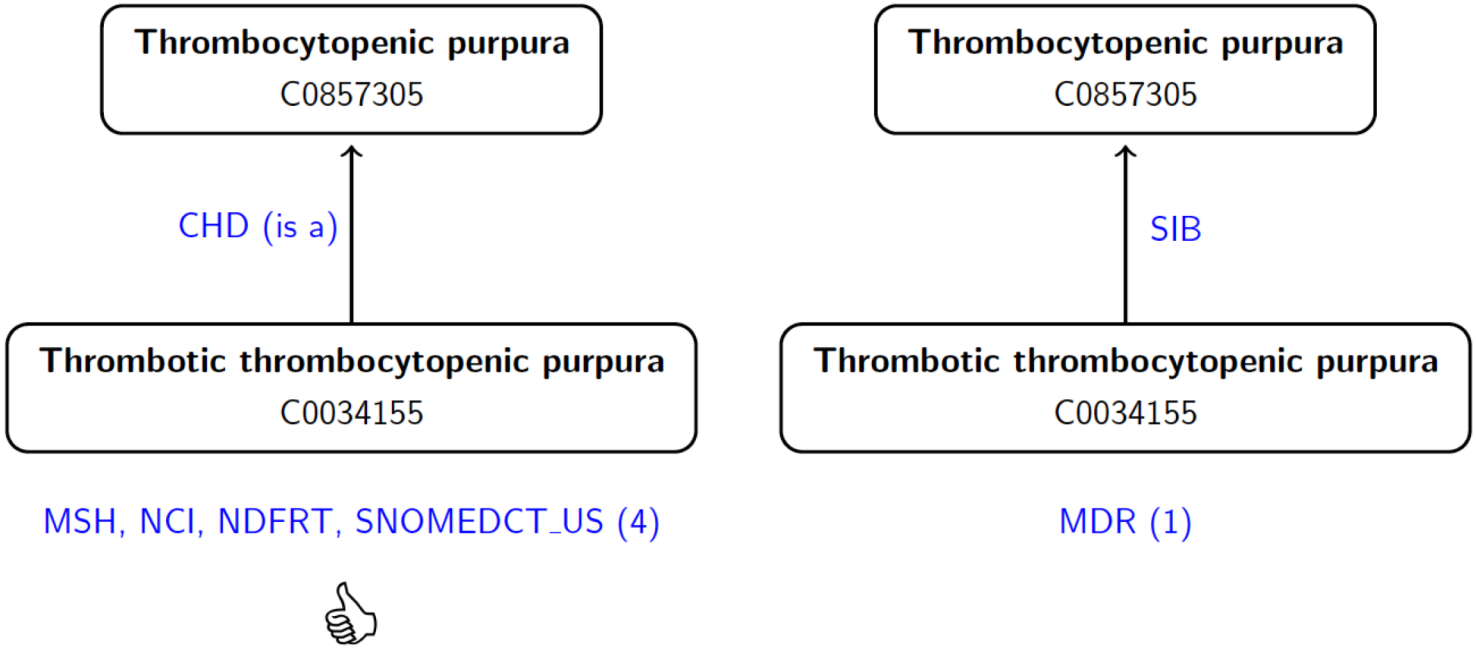
# Inconsistent Relation Detection

| No. | Concept pair $(C_1, C_2)$ | Relationships and source vocabularies $\{(R, RA, S)\}$ |
|---|---|---|
| 1 | *Arginine supplement (C3853287), Arginine and glutamine supplement (C3853286)* | *CHD, is a, SNOMEDCT_US* |
| 2 | *Hexosamines (C0019478) Fructosamine (C0060765)* | *CHD, null, MSH \| CHD, null, NDFRT* |
| 3 | *17-Oxosteroid (C0000167), Androstenedione (C0002860)* | *SIB, null, CPM \| CHD, null, MSH \| CHD, null, NDFRT* |

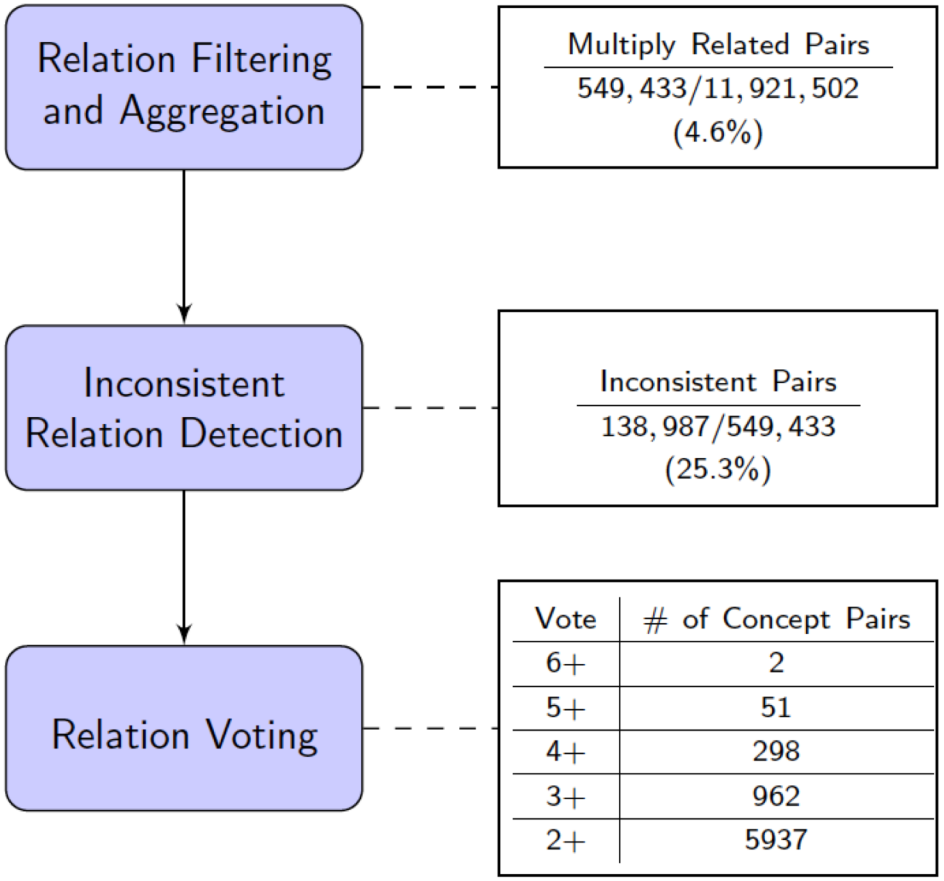


# Relation Voting

- Rank by votes
- Suggest top-ranked relationship with at least 2 votes



# Results

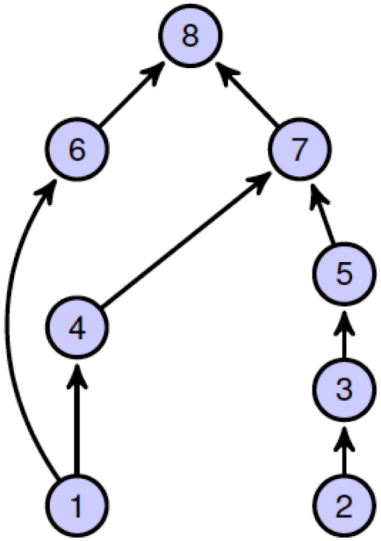



| Vote | # of Concept Pairs |
|---|---|
| 6+ | 2 |
| 5+ | 51 |
| 4+ | 298 |
| 3+ | 962 |
| 2+ | 5937 |

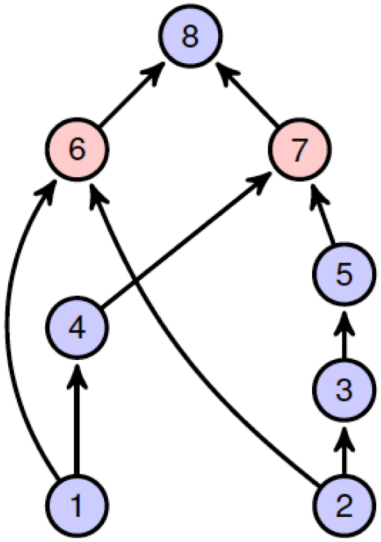
# Non-lattice-based Auditing of SNOMED CT

# Semilattice, Non-lattice

- An **upper semilattice** is a *partially ordered set* in which any two elements have a *unique minimal upper bound*.

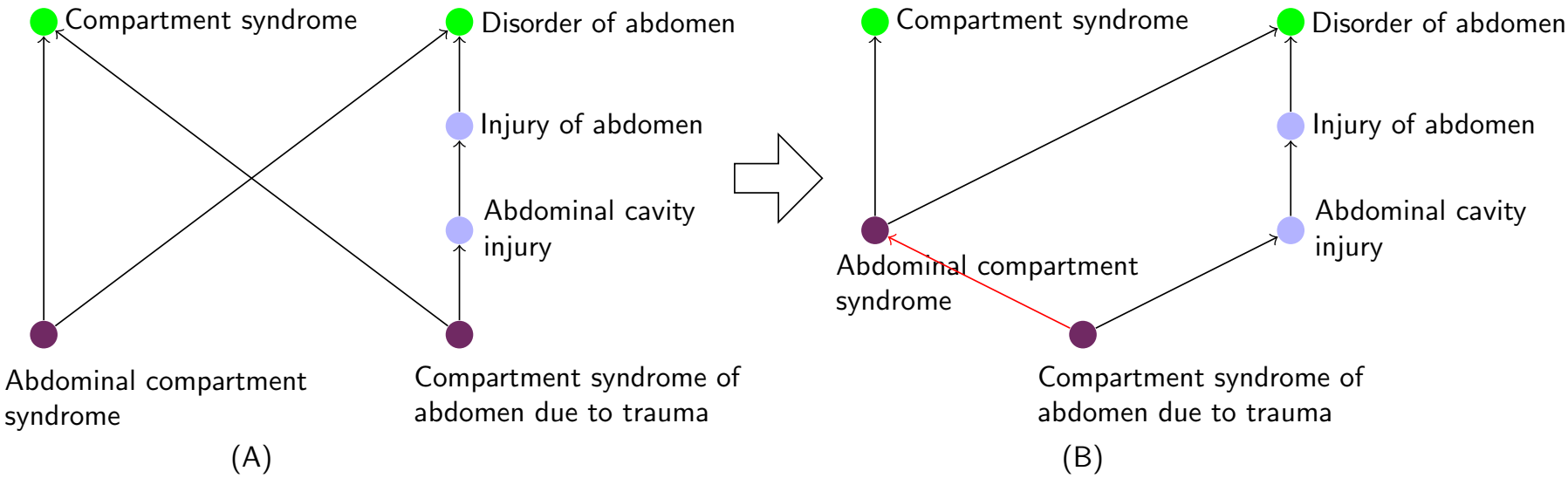



Upper semilattice




Non-lattice

# Example in SNOMED CT



# Non-lattice Detection (I)

Given a poset $(L, \leq)$ and $X \subseteq L$

- $\Uparrow X$ denotes the set of all common ancestors of $X$

$$\{a \mid \forall x \in X, x < a\}$$

- $\uparrow X$ represents its strict upper closure

$$\{a \mid \exists x \in X, x < a\}$$

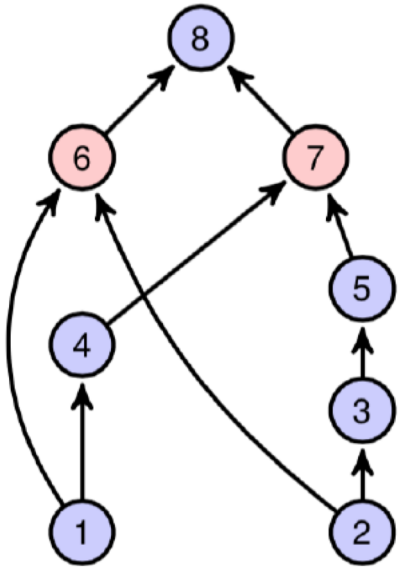
# Non-lattice Detection (II)

**Theorem:** Let $X$ be a set of pairwise incomparable elements of a poset $L$, with $|X| \geq 2$. We have

$$\mathrm{mub}(X) = {}^{\Uparrow}X - {}^{\uparrow}({}^{\Uparrow}X).$$

**Special Case:** Given a pair of incomparable elements $x$ and $y$,

$$\mathrm{mub}(\{x, y\}) = ({}^{\uparrow}x \cap {}^{\uparrow}y) - \bigcup_{t \in {}^{\uparrow}x \cap {}^{\uparrow}y} {}^{\uparrow}t$$

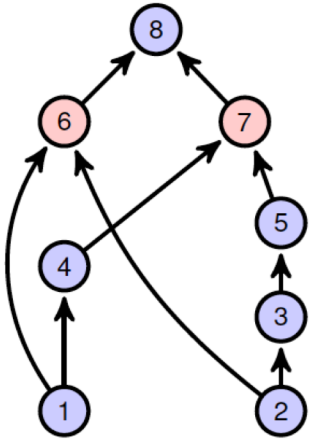



${}^{\uparrow}1 = \{4, 6, 7, 8\}$
${}^{\uparrow}2 = \{3, 5, 6, 7, 8\}$

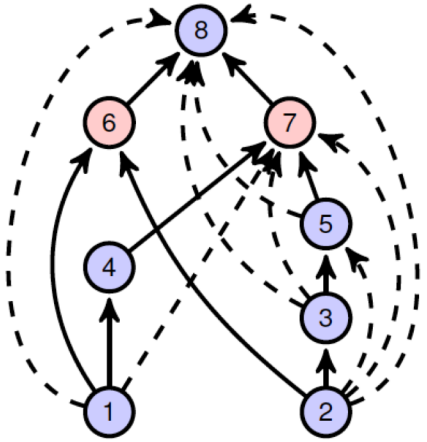
${}^{\uparrow}1 \cap {}^{\uparrow}2 = \{6, 7, 8\}$
${}^{\uparrow}6 \cup {}^{\uparrow}7 \cup {}^{\uparrow}8 = \{8\}$

$\mathrm{mub}(\{1, 2\}) = \{6, 7\}$

# Non-lattice Detection (III): Computing Transitive Closure using MapReduce



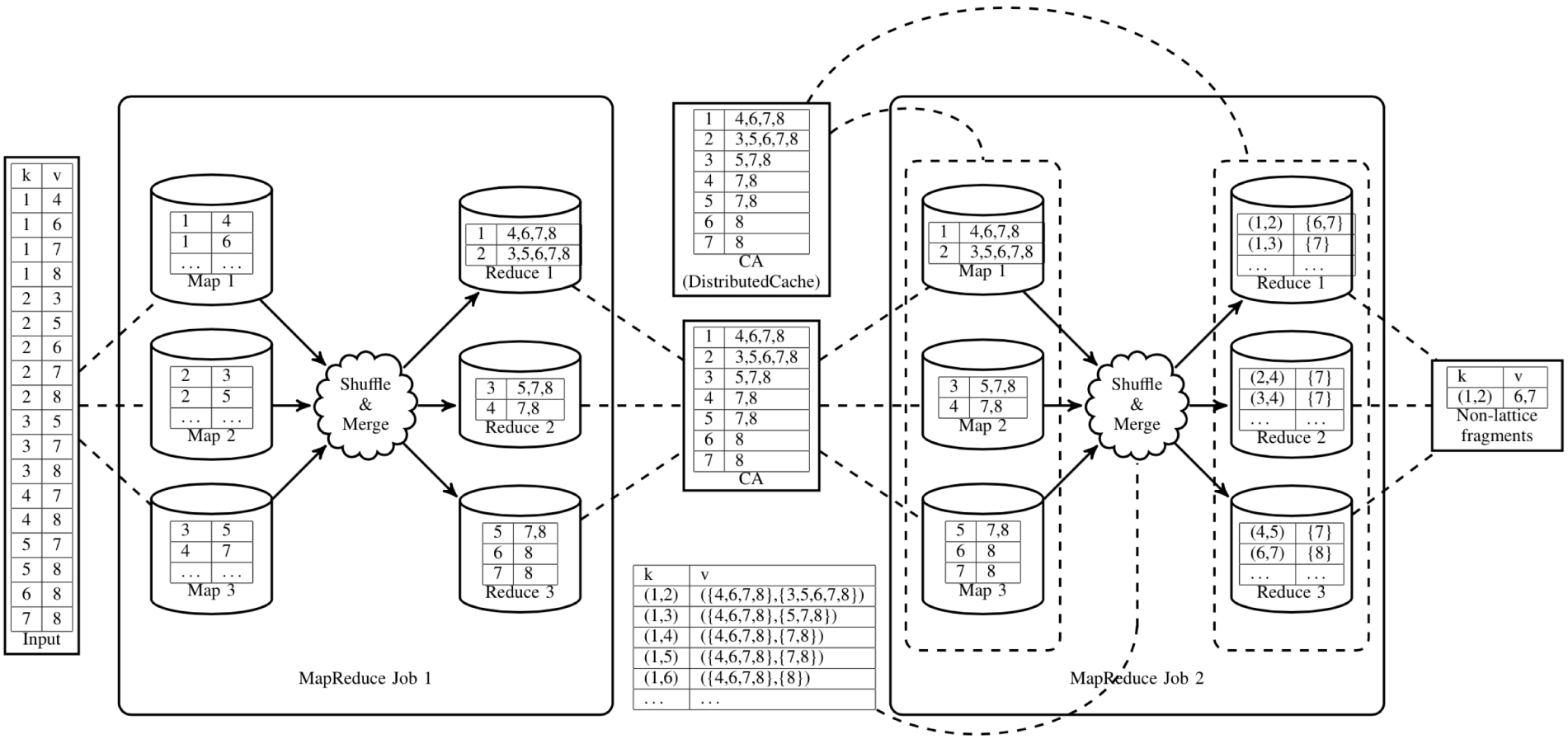

(1,4), (1, 6), (2, 3), (2, 6), (3, 5),
(4, 7), (5, 7),(6, 8), (7, 8)




(1,4), (1, 6), (1, 7), (1, 8), (2, 3), (2, 5),
(2, 7), (2, 8), (3, 5), (3, 7), (3, 8), (4, 7),
(4, 8), (5, 7), (5, 8), (6, 8), (7, 8)

- Map: $(c, a) \rightarrow (c, \{a_1, a_2, \ldots, a_n\})$
- Reduce: $(c, \{a_1, a_2, \ldots, a_n\}) \rightarrow (c, {}^{\uparrow}c)$

# MapReduce Pipeline for Detecting Non-lattice Pairs




| k | v |
|---|---|
| (1,2) | ({4,6,7,8},{3,5,6,7,8}) |
| (1,3) | ({4,6,7,8},{5,7,8}) |
| (1,4) | ({4,6,7,8},{7,8}) |
| (1,5) | ({4,6,7,8},{7,8}) |
| (1,6) | ({4,6,7,8},{8}) |
| ... | ... |

$\uparrow t$

$$\mathrm{mub}(\{x, y\}) = (\uparrow x \cap \uparrow y) - \bigcup_{t \in \uparrow x \cap \uparrow y} \uparrow t$$

# Results for Non-lattice Pairs

- 30-node Hadoop local cloud
- 8 versions from 07/2009 to 03/2014
- Average total computing time: $<$ 3 hours (~~3 month~~)

| | 07/2009 | 01/2010 | 01/2011 | 01/2012 | 07/2012 | 01/2013 | 07/2013 | 03/2014 |
|---|---|---|---|---|---|---|---|---|
| Concepts | 306,627 | 290,078 | 292,405 | 294,797 | 295,311 | 296,876 | 297,695 | 299,286 |
| IS-A Relations | 445,549 | 430,489 | 435,294 | 438,711 | 438,927 | 441,589 | 443,796 | 445,357 |
| Non-lattice Pairs | 559,182 | 566,239 | 576,688 | 568,535 | 581,754 | 574,204 | 584,934 | 586,771 |
| Compute Time (s) | 10,168 | 74,169 | 6,726 | 7,822 | 7,883 | 8,176 | 8,430 | 11,166 |

# Mining Non-lattice Subgraphs using Lexical Patterns

- Consider the fully specified name of a concept as a set (bag) of words in lower case:
  - Containment

$$U_i \subset U_j \text{ or } L_i \subset L_j$$

  - Intersection

$$L_i \cap L_j = U_k$$

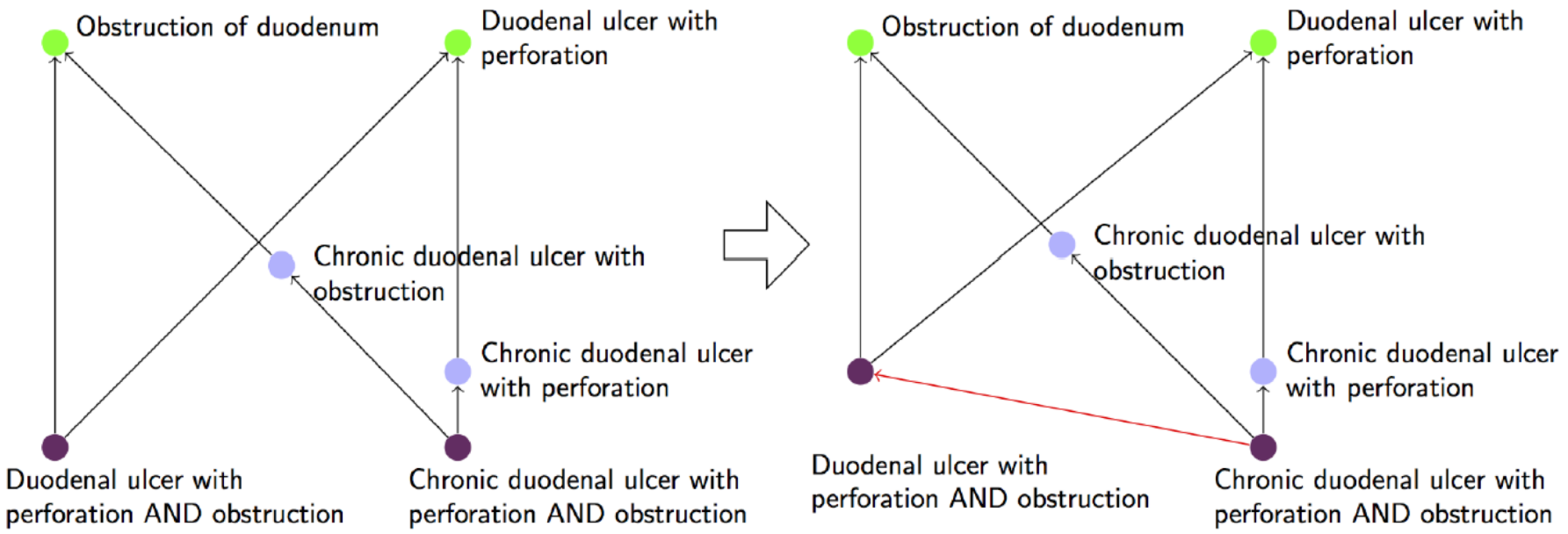
  - Union

$$U_i \cup U_j = L_k$$

  - Union-Intersection

$$U_i \cup U_j = L_s \cap L_t$$

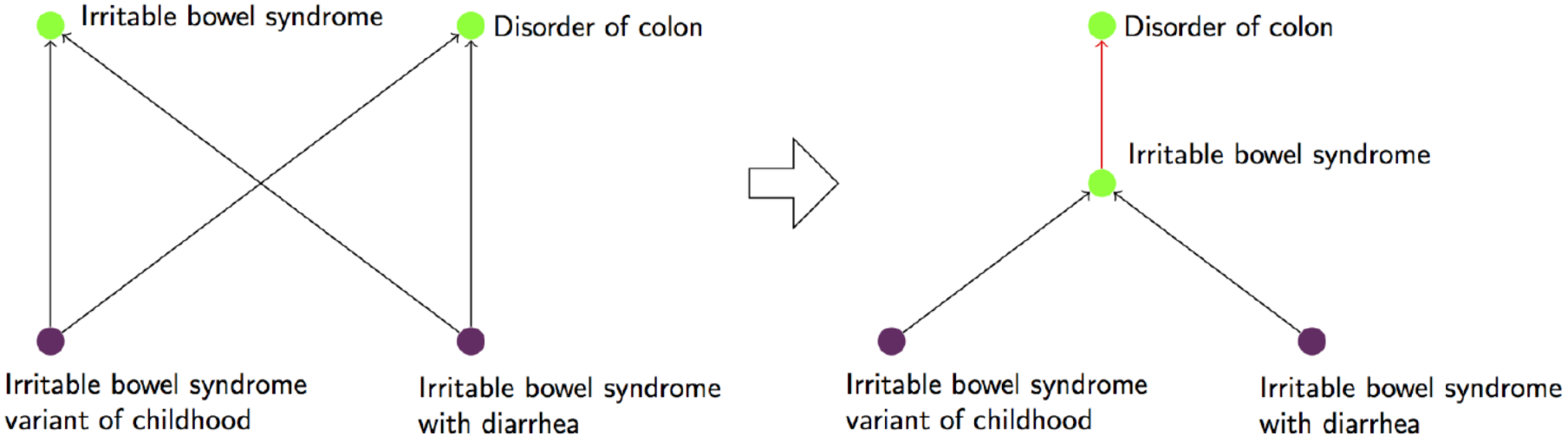
# Containment ($L_i \subset L_j$)




$\{duodenal, ulcer, perforation, obstruction\} \subset \{chronic, duodenal, ulcer, perforation, obstruction\}$

- This situation generally suggests a missing hierarchical relation between concepts in the upper boundary or the lower boundary.

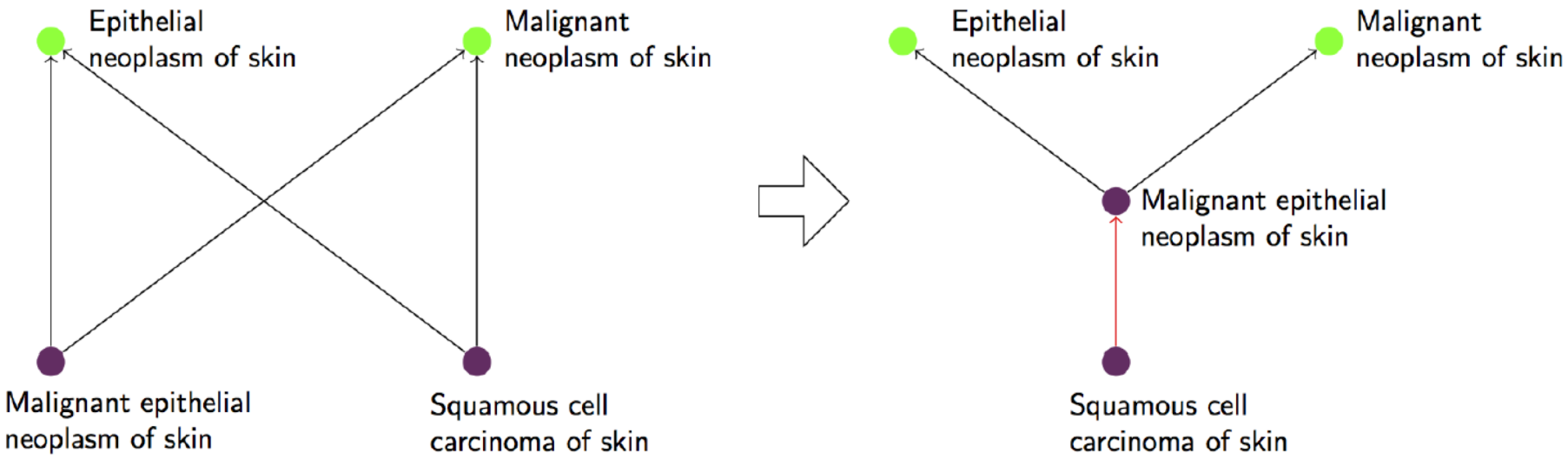
# Intersection ($L_i \cap L_j = U_k$)




$$\{irritable, bowel, syndrome, variant, childhood\} \cap \{irritable, bowel, syndrome, diarrhea\}$$
$$= \{irritable, bowel, syndrome\}$$

- This situation generally suggests a missing hierarchical relation between concepts in the upper boundary.

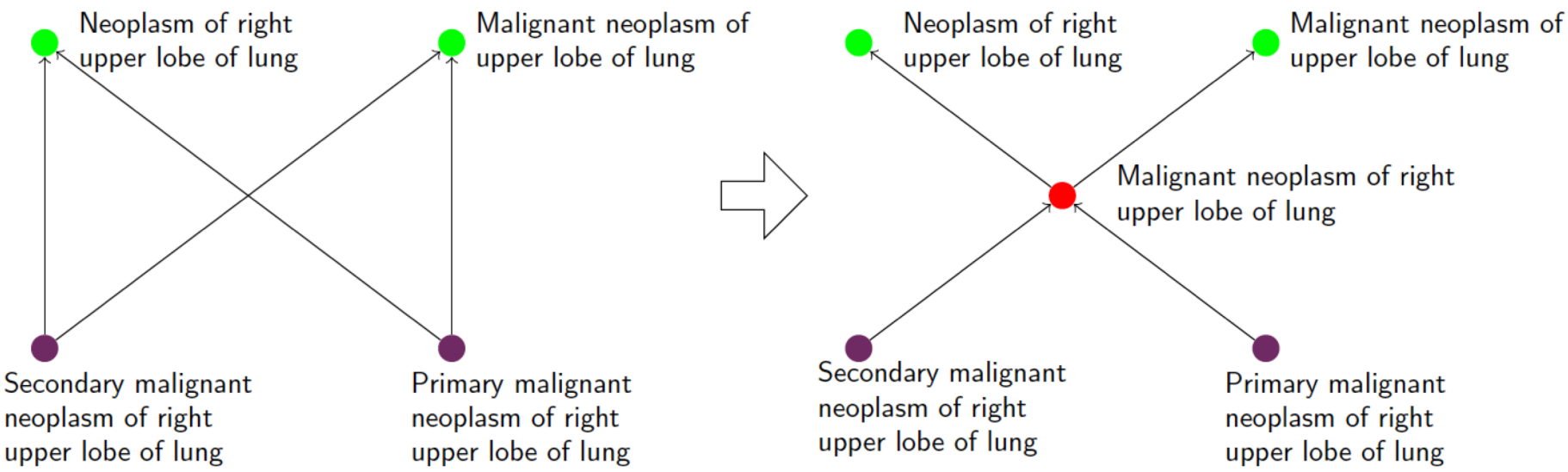
# Union ($U_i \cup U_j = L_k$)




$$\{epithelial, neoplasm, skin\} \cup \{malignant, neoplasm, skin\}$$
$$= \{epithelial, neoplasm, skin, malignant\}$$

- This situation generally suggests a missing hierarchical relation between concepts in the lower boundary.

# Union-intersection (UI) ($U_i \cup U_j = L_s \cap L_t$)




$$\{neoplasm, right, upper, lobe, lung\} \cup \{malignant, neoplasm, upper, lobe, lung\} =$$
$$\{secondary, malignant, neoplasm, right, upper, lobe, lung\} \cap \{primary, malignant, neoplasm, right, upper, lobe, lung\}$$

- This situation generally suggests a missing intermediary concept between the upper boundary and the lower boundary.

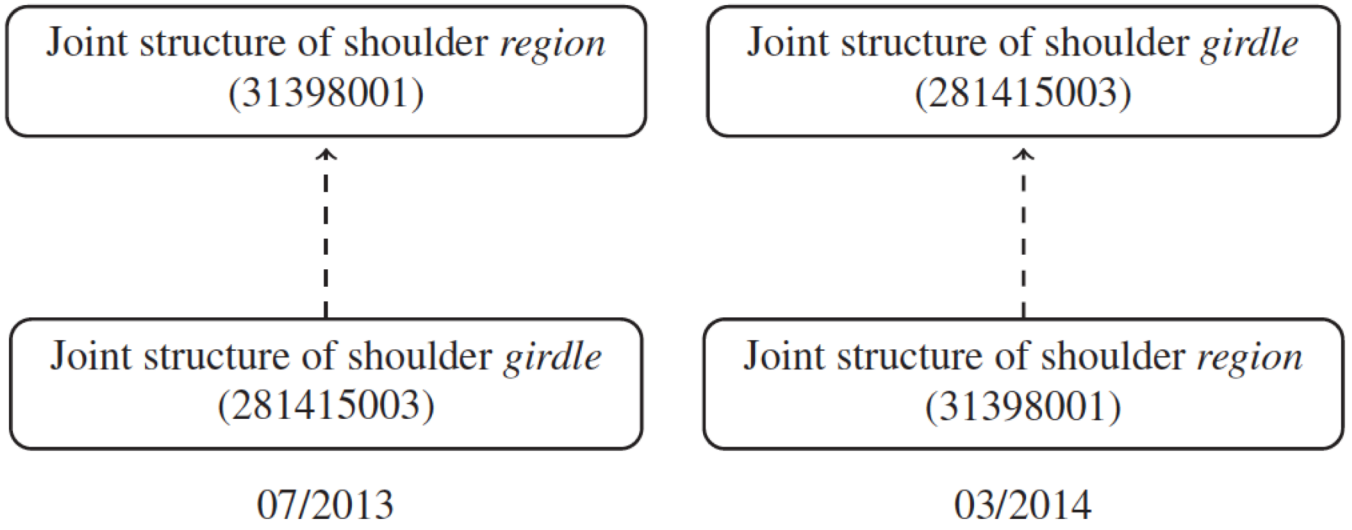
# Results for Ming Non-lattice Subgraphs

Material: September 2015 version of SNOMED CT (U.S. edition)

- 631,006 non-lattice pairs were detected.
- 171,011 non-lattice subgraphs were extracted.
- Sizes of non-lattice subgraphs ranged from 4 to 5,137.
- 90% of the non-lattice subgraphs had sizes 4 to 100.

# More Examples on OQA

- ✓ Mining Relation Reversals in the Evolution of SNOMED CT Using MapReduce*




*Tao S, **Cui L**, Zhu W, Sun M, Bodenreider O, Zhang GQ. Mining Relation Reversals in the Evolution of SNOMED CT Using MapReduce. ***AMIA Joint Summits on Translational Science*** 2015, pp. 46-50.