Automated Phenotypic Networks for the Integration of Heterogeneous Databases

Yves A. Lussier [1,2]

Xiaoyan Wang [1]

1 Dept of Biomedical Informatics

2 Dept of Medicine

Columbia University

Preview of Take Home Points

- Exponential growth of heterogeneous DBs
  - **difficult for human to review and recall**
- Complexity of Phenotypes
  - **Span scales of Biology, different granularity of description leading to compositional variants, ambiguity**
- Beyond Ontologies,
  Computational Networks of Phenotypes
  - **map knowledge of genomic databases in reusable representations**

Outline

- **Challenge**
- **Introduction:**
  - **Data representation vs Schema**
  - **Curation vs Automation**
  - **Direct Maps vs Phenotypic Networks (PN)**
- **Hypotheses**
- **Methods**
- **Results**
- **Conclusions**

Challenges

- **Heterogeneously data representation**
  - **Structural differences**
  - **Naming conventions & standards differences across fields**
  - **Semantic differences**
  - **Context differences**
- **Variable Database Schema**

Examples of Interoperability

- **Based on Schema**

  Requires compatible indexes,

  supports unrelated schema

  - **Mork P, Halevy A, Tarczay-Hornoch P. A model for data integration systems of biomedical data applied to online genetic databases. Proc AMIA Symp 2001:473-7.**

- **Based on Data Representation**

  can map unrelated data dictionaries

  Requires compatible schema

Interoperability

Based on Data Representation

- **Manual Curation**

  e.g.: UMLS, NCI Metathesaurus

  - **rate-limiting for data sets using current terminologies**
    - **delayed and incomplete synchronization**
  - **High throughput unattainable for uncoordinated data sets**

- **Computational Curation / Automation**

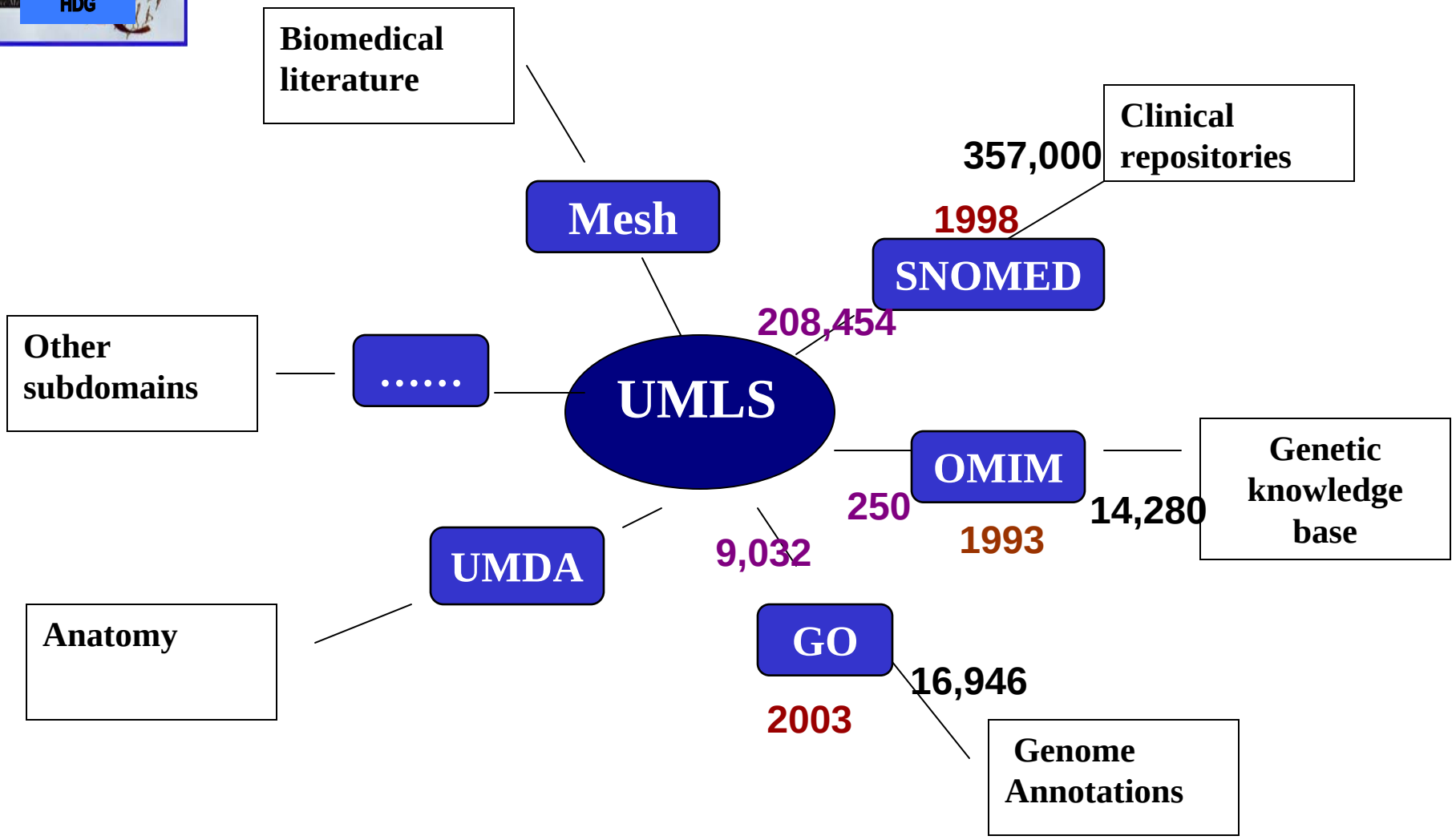
  E.g. automated indexing

Introduction

Interoperability based on **Manual Curation**

- rate-limiting for data sets using current terminologies
  - delayed and incomplete synchronization
- High throughput unattainable for uncoordinated data sets

Manual Indexing / Curation



Introduction: Automated Indexing

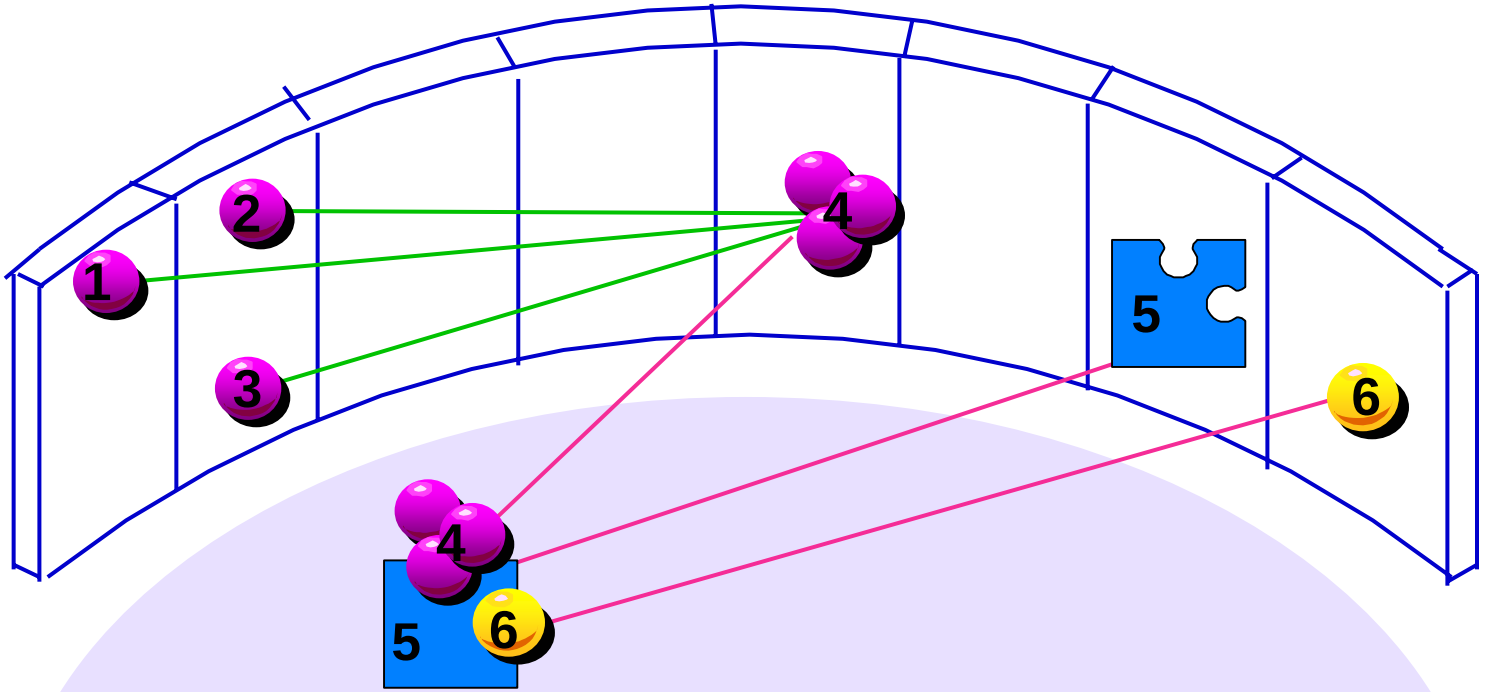
- **Automated Indexing**
  - **Direct maps between two unrelated data dictionaries**
  - **No use of networks of relationships**
  - **Rare studies in clinical genetics and molecular biology;**
  - **Lexical matching**
    - *Sperzel WD et al. Biomedical database interconnectivity: An experiment linking MIM, GENBANK, and META-1 via MEDLINE. Proc Annu Symp Comput Appl Med Care 1991:190-193.*
  - **Lexical and semantics**
    - *Bodenreider O. Pac Symp Biocomputing 2004*
    - *Sarkar IN, Lussier YA et al.. Linking biomedical information and knowledge resources: GO and UMLS. Pac Symp Biocomputing 2003;8:427-50.*

Semantic Information Model of SNOMED

Compositional, multiaxial, multi-hierarchic

T M F C D P G L

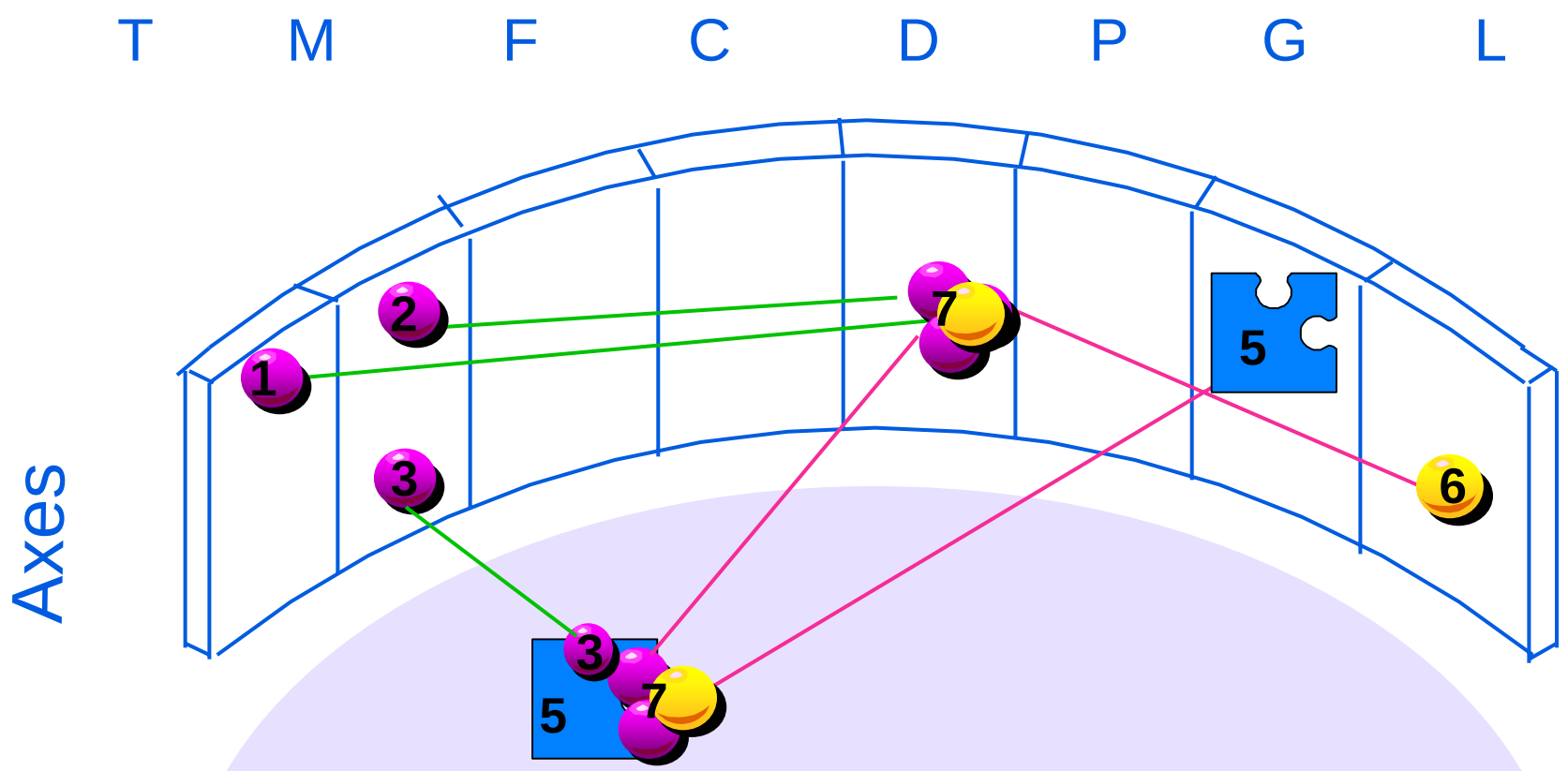
Axes



H. Pylori associated heamorrhagic Gastric Ulcer =
(4) D5-32220 Gastric (1) Ulcer (2) with haemorrhage (3)
G-C002 associated with (5) L-13551 H. pylori (6)

SNOMED Information Model:

Representational variant



H. pylori associated haemorrhagic Gastric Ulcer =
(7) DE-16016 H.pylori (6) associated Gastric (1) Ulcer (2) with (5) M-37000 haemorrhage (3)

Outline

- **Challenge**
- **Introduction: Phenotypic Networks (PN)**
- **Hypotheses**
- **Methods:**
  - **Curation vs Automated mappings**
  - **Direct maps vs network-based maps**
- **Results**
- **Conclusions**

Hypothesis

Proof-of-Concept Study:

Automated Networks of Phenotypes
can increase recall and precision
of queries across two heterogeneous databases
sharing no cross-indexes.

Method

- **Automated terminology networks**
  - **Databases**
  - **Computational network of phenotypes**
  - **Incremental Lexico-semantic techniques**
    - **Lexical method**
    - **Semantic constrains**
    - **Multi-strategy / Incremental exploitation of the network**
  - **Network's pathways**
  - **Accuracy measurements**
- **Evaluation**
  - **Gold standard**

Method: databases

Target databases

- **Human Disease Genes Database (HDG)**

  Jimenez-Sanchez G, Childs B, Valle D. Human disease genes. Nature 2001 409: 853-5

  – **Manually compiled database to classify disease genes & their products according to function**

  – **921 disease genes are documented in the database**

- **SNOMED-Clinical Term (clinical medicine)**

  – **Concept-based clinical terminology**

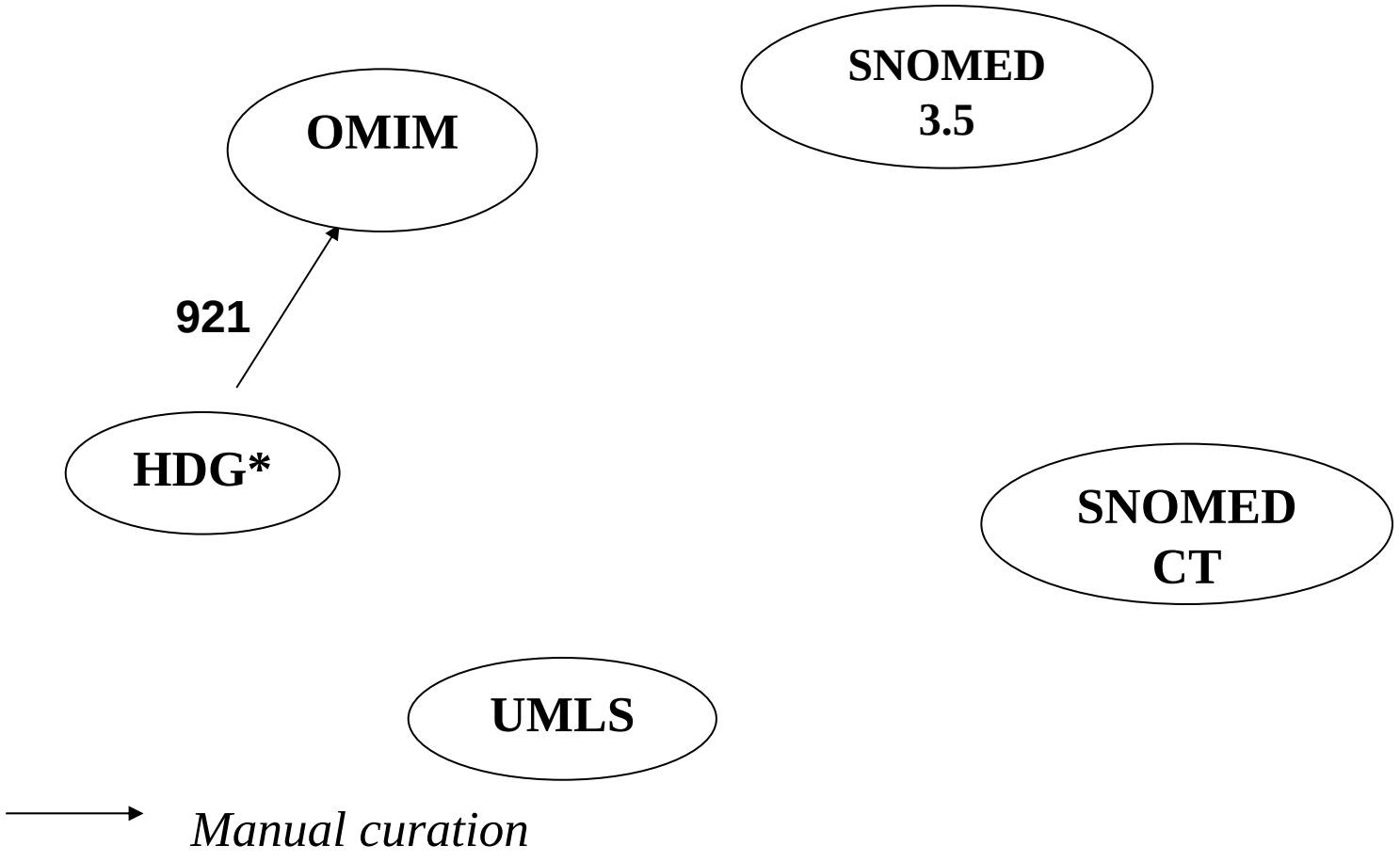
  – **Version used: July, 2002 ; 333,325 concepts.**

Method: databases

Intermediating databases/terminologies

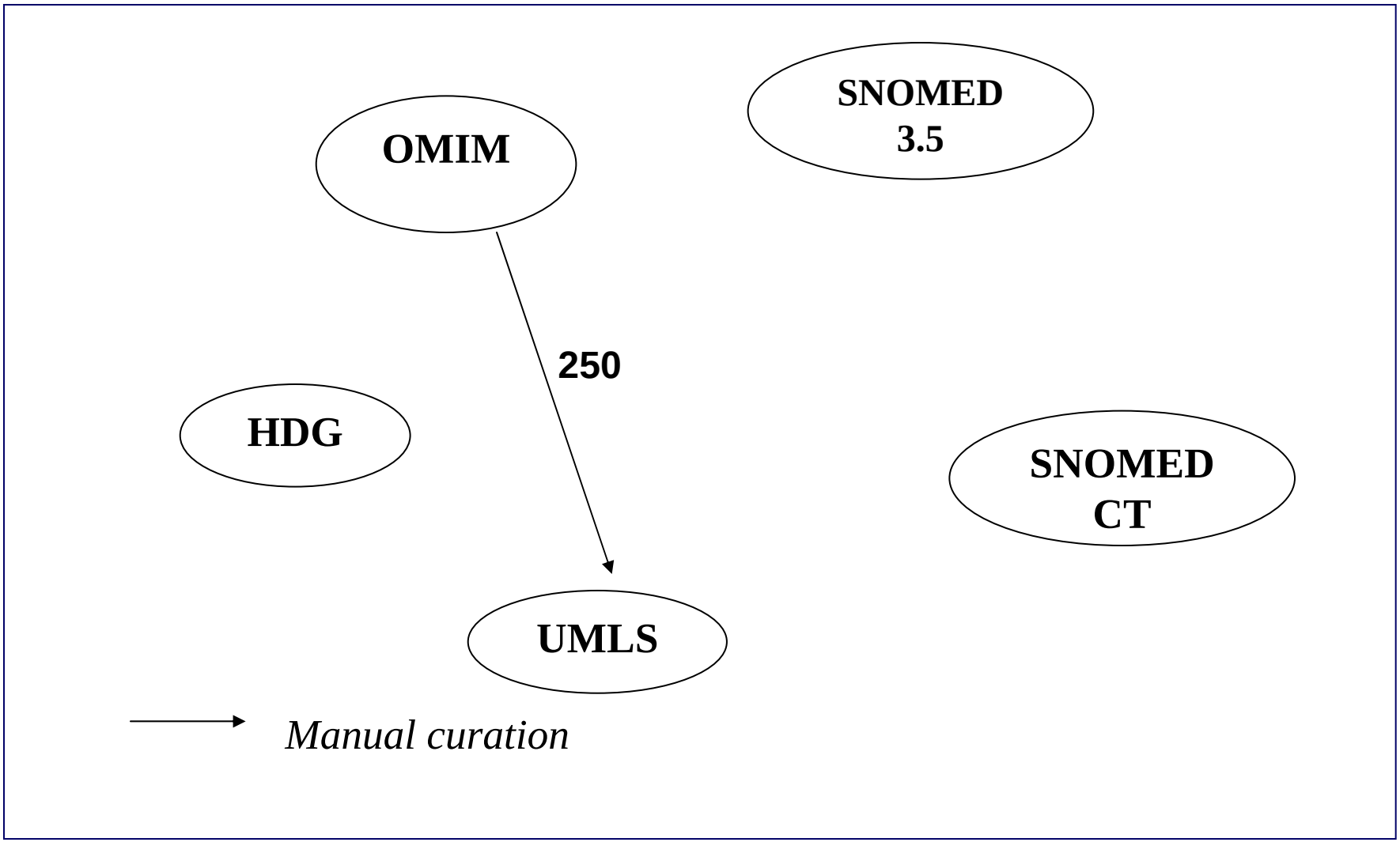
- **Online Mendelian Inheritance in Man (OMIM);**
  - **14,280 entries (Loci and diseases)**
- **Unified Medical Language System (UMLS);**
  - **871,584 concepts (version 2002AB)**
- **SNOMED 3.5**
  - **208,454 concepts (version SNOMED Intern., 3.5/ 1998)**

Method: Manual Curation

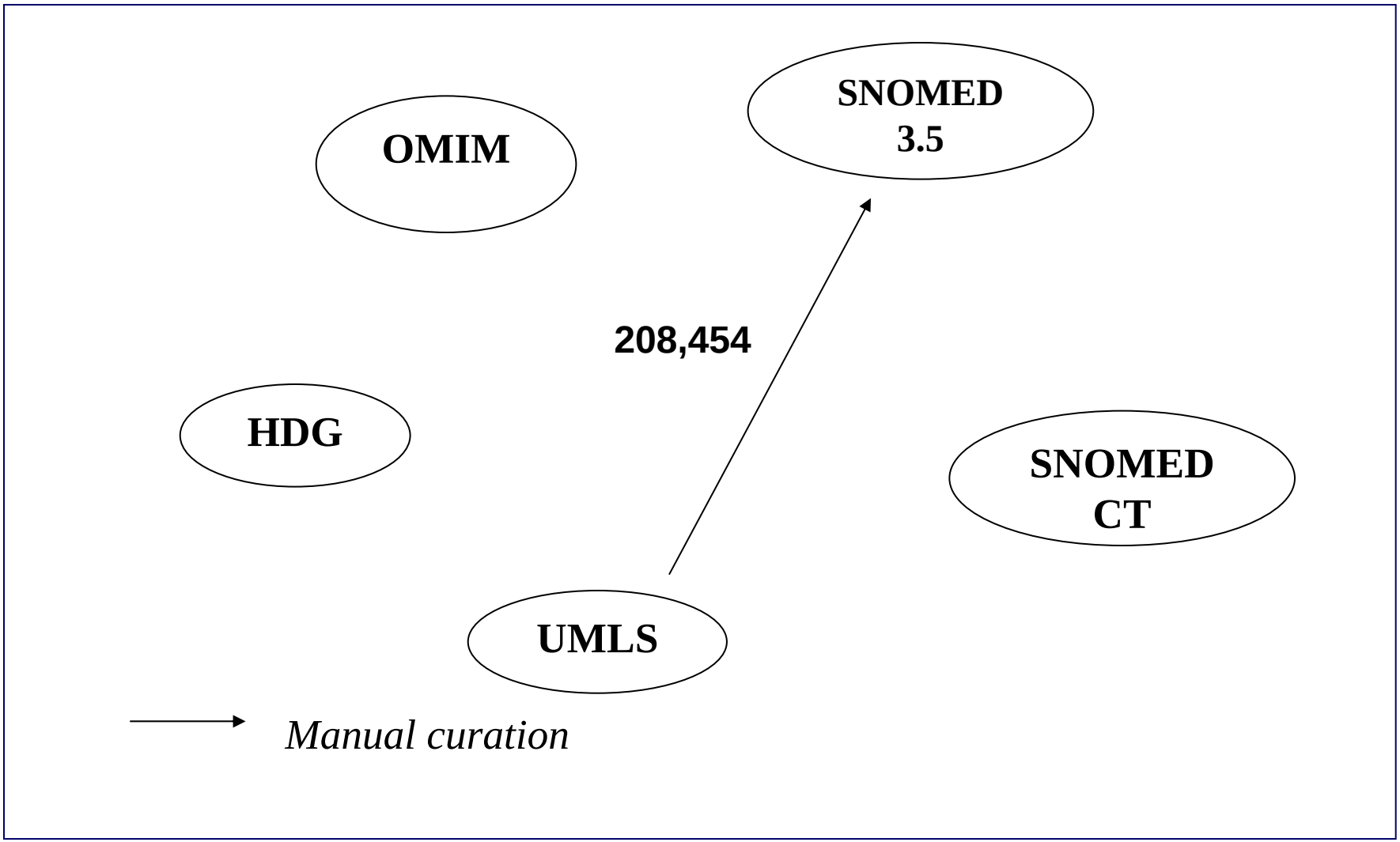


** Jimenez-Sanchez G, Childs B, Valle D. Human disease genes. Nature 2001 409: 853-5*

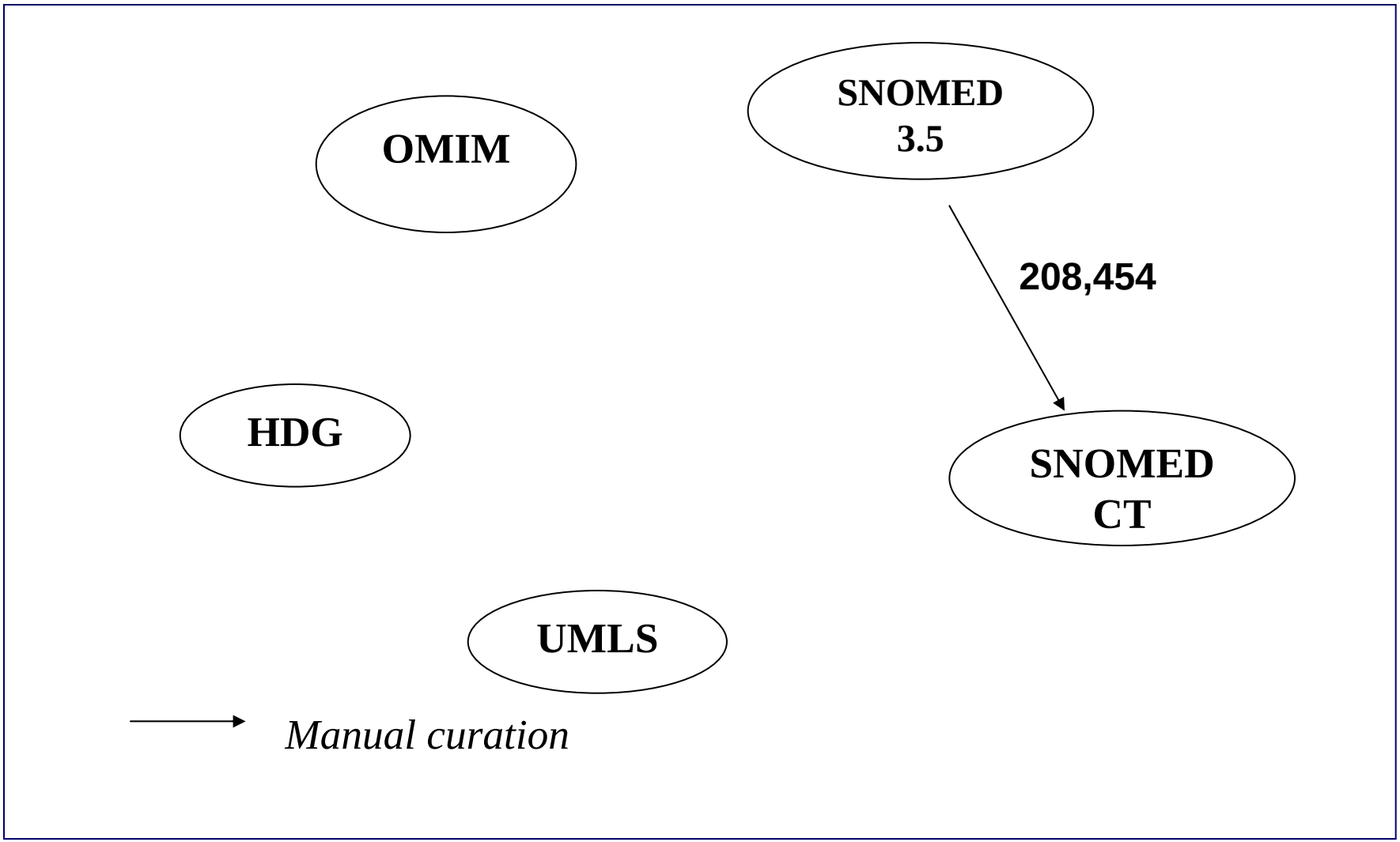
Method: Manual Curation



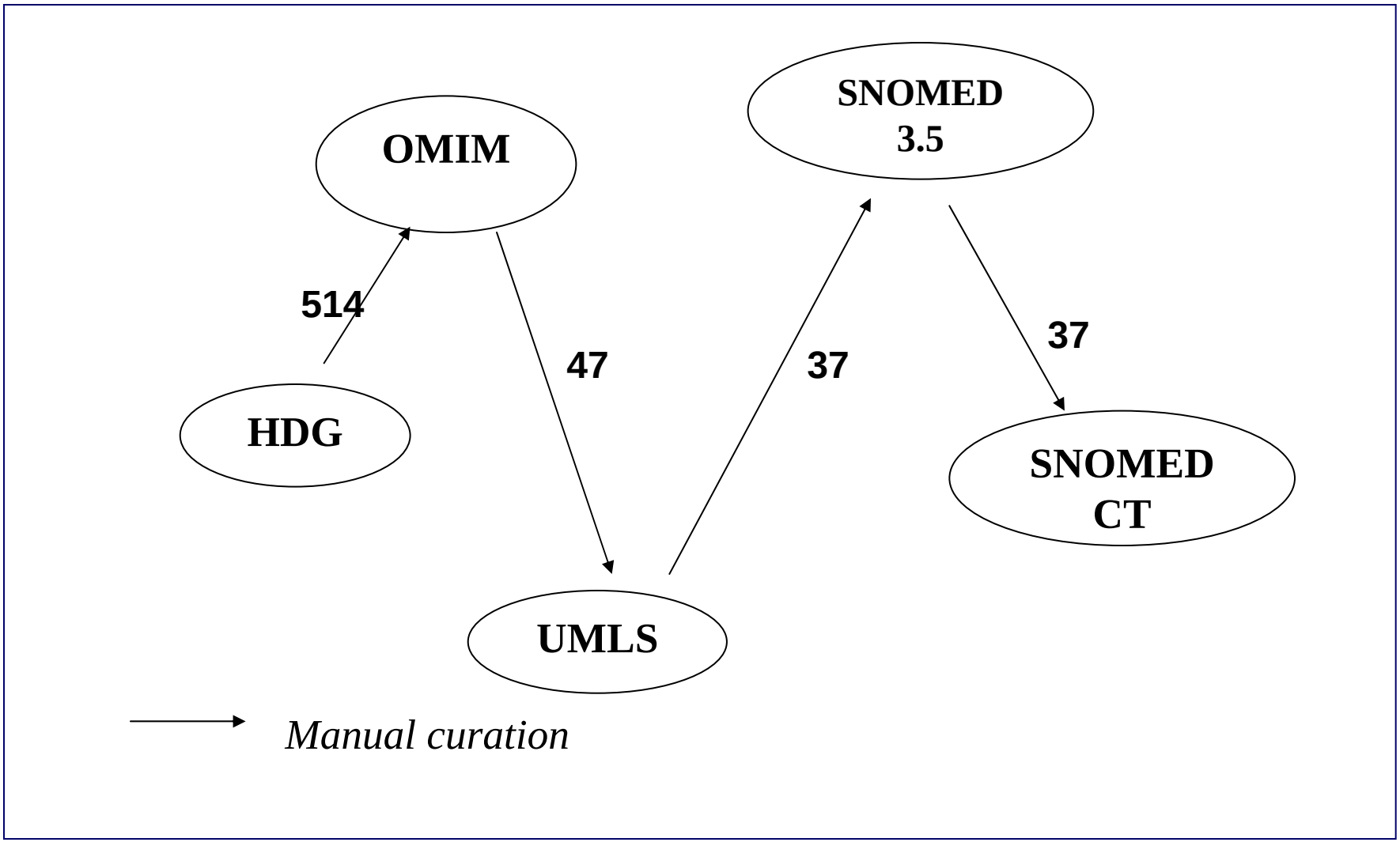
Method: Manual Curation



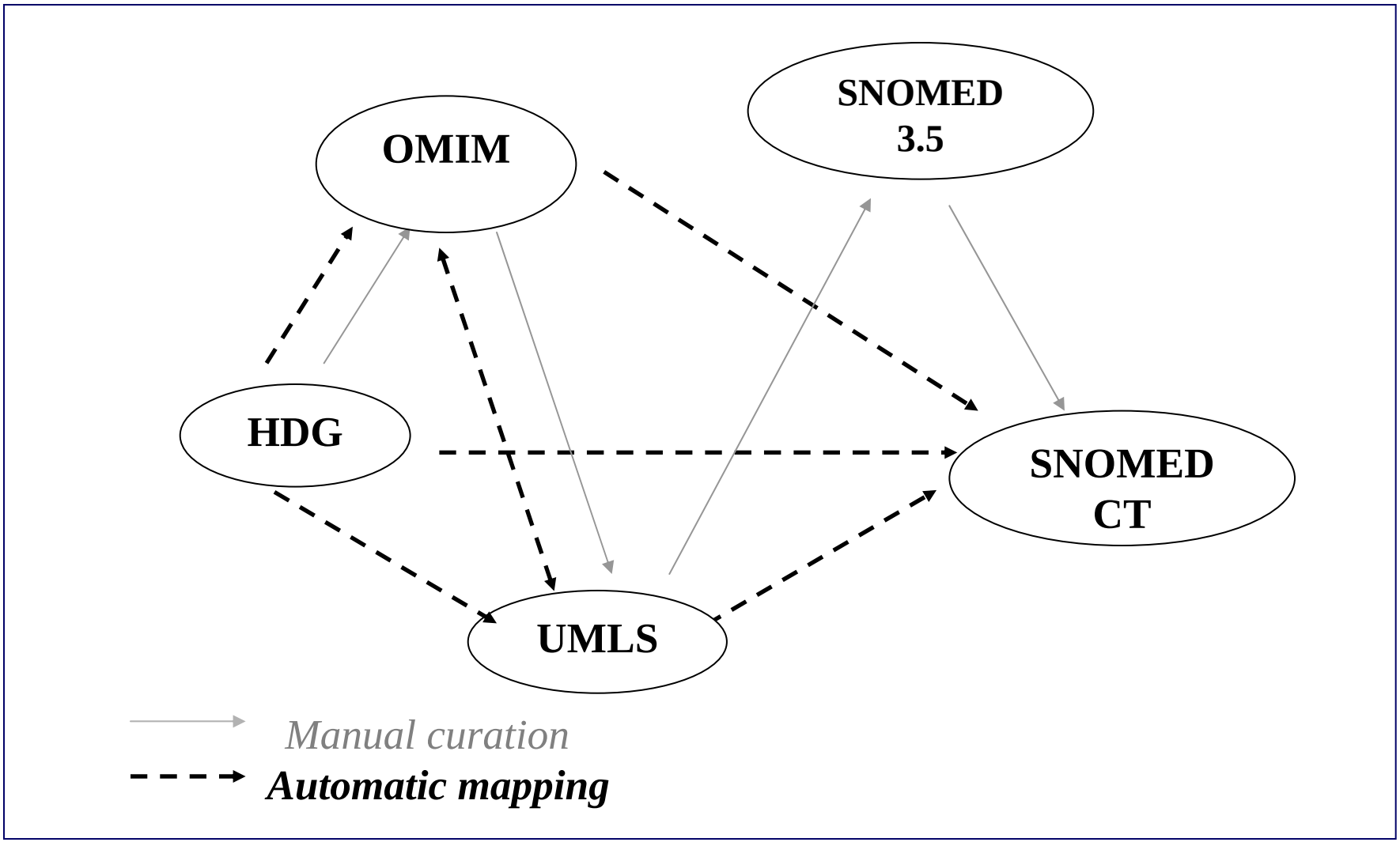
Method: Manual Curation



Method: Manual Curation



Method: Automated Terminology Network: ATN



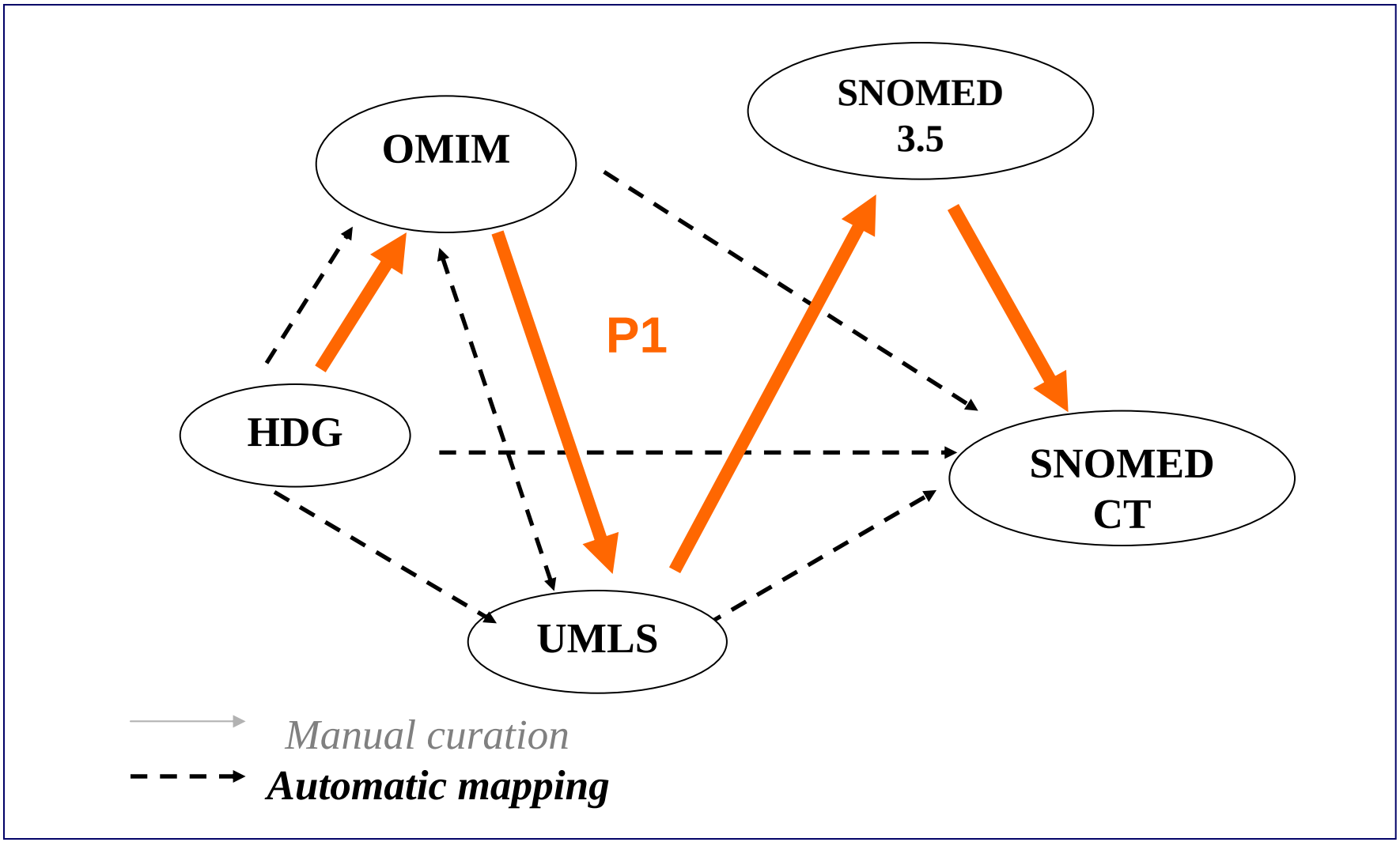
Method: Paths derived from the network

Path Name	Intermediating terminologies (#)	Complete Path
P1	3	HDG = OMIM = UMLS = SMOMED3-5=SNOMED-CT
P2	0	HDG → SNOMED-CT
P3	1	HDG → UMLS → SNOMED-CT
P4	1	HDG → OMIM (Disease) → SNOMED-CT
P5	1	HDG → OMIM (Title) → SNOMED-CT
P6	2	HDG → UMLS → OMIM → SNOMED-CT
P7	2	HDG → OMIM → UMLS → SNOMED-CT

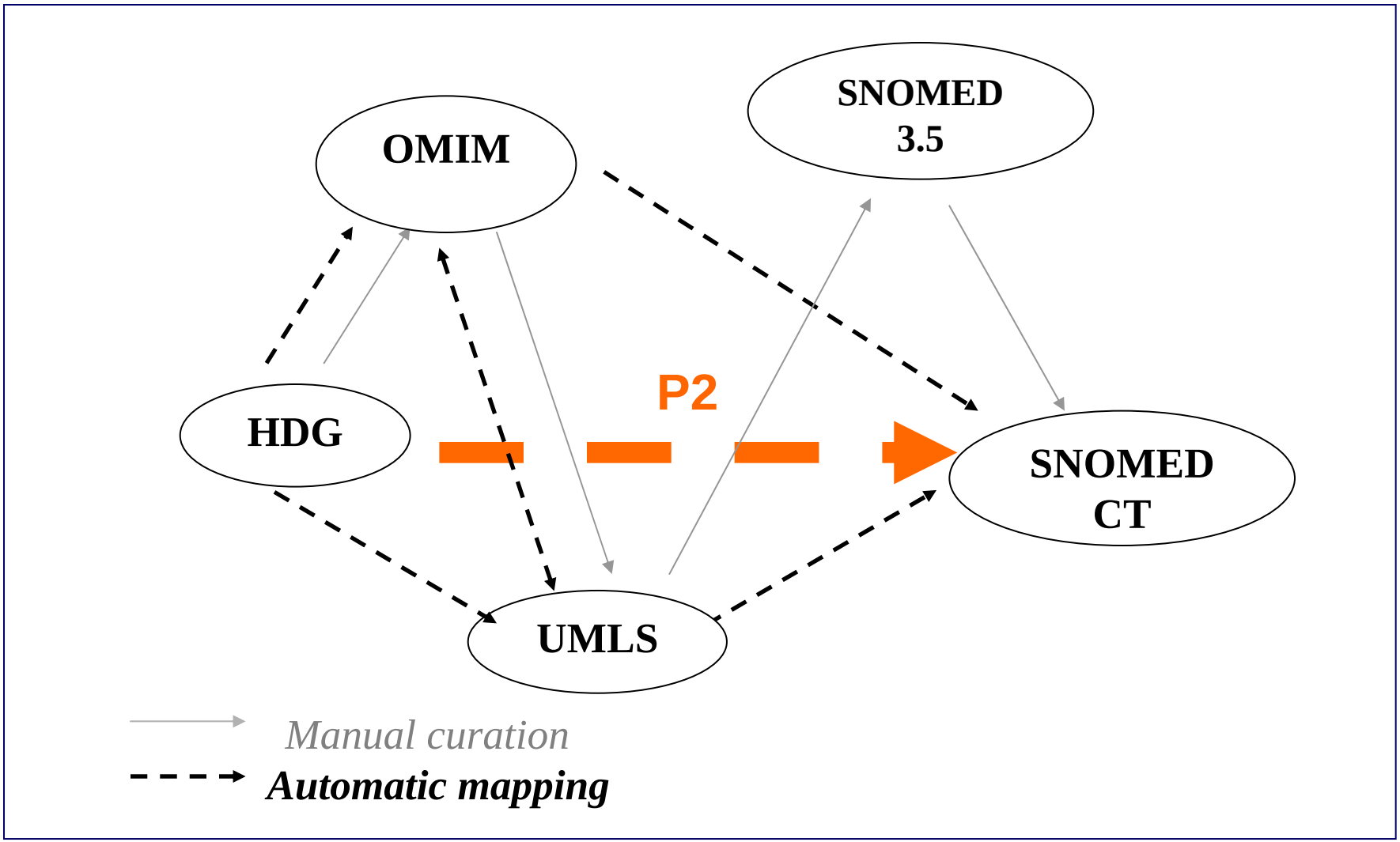
A = B Manual Curation / Mapping of terms via a common index between databases A and B.

A→B Automated Mapping / lexico-semantic mapping of terms between databases A and B.

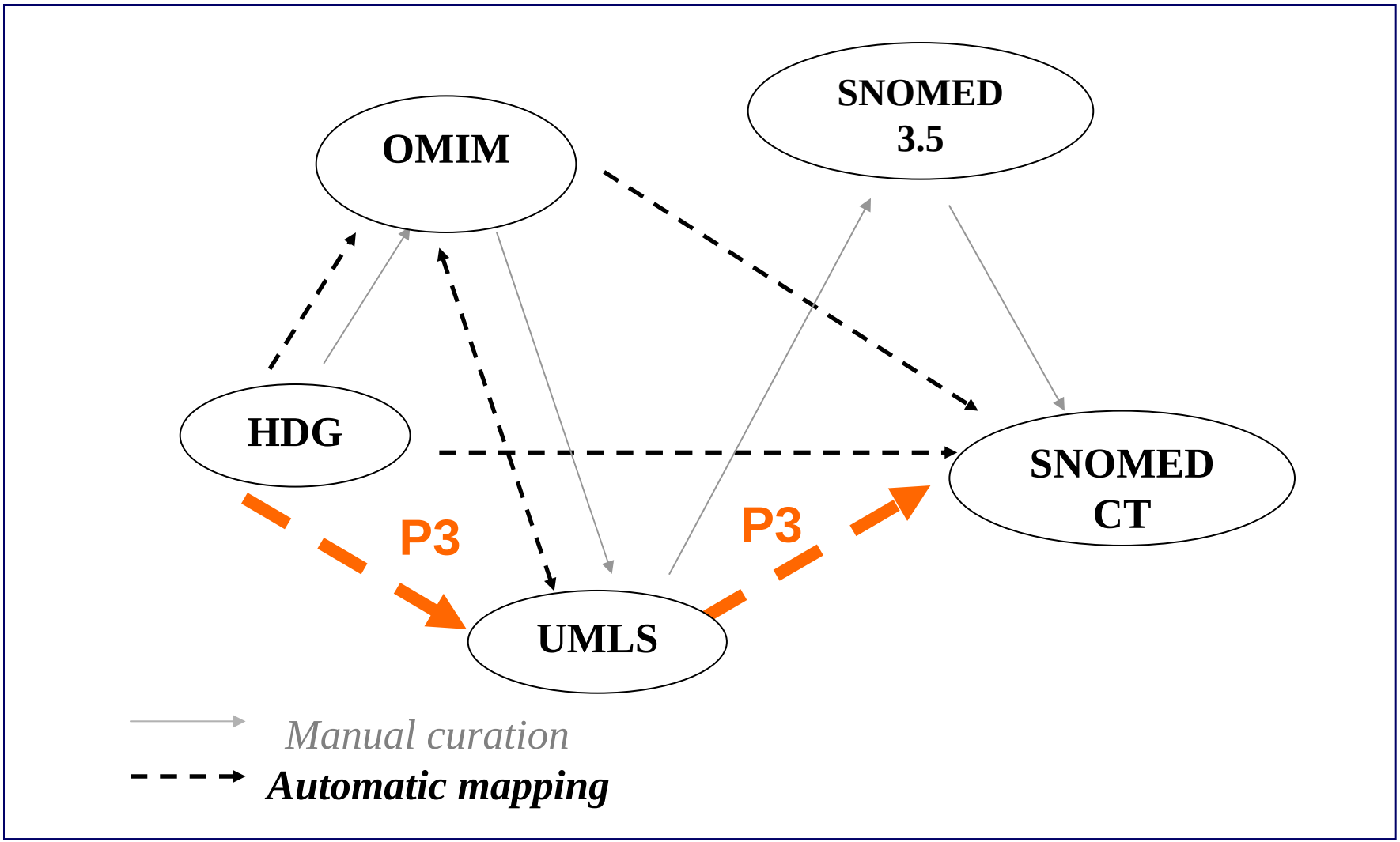
Method: Automated Terminology Network: ATN



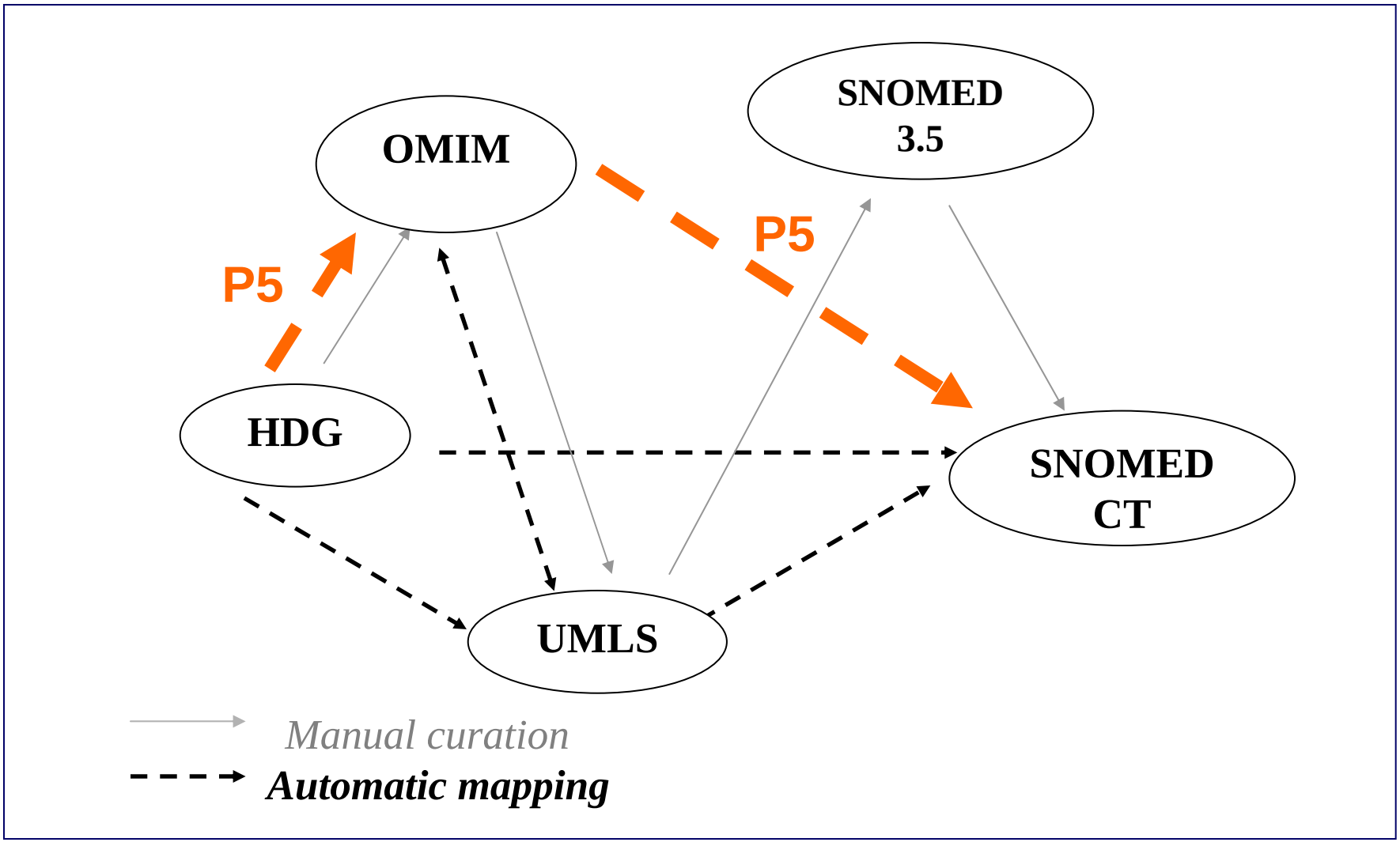
Method: Automated Terminology Network: ATN



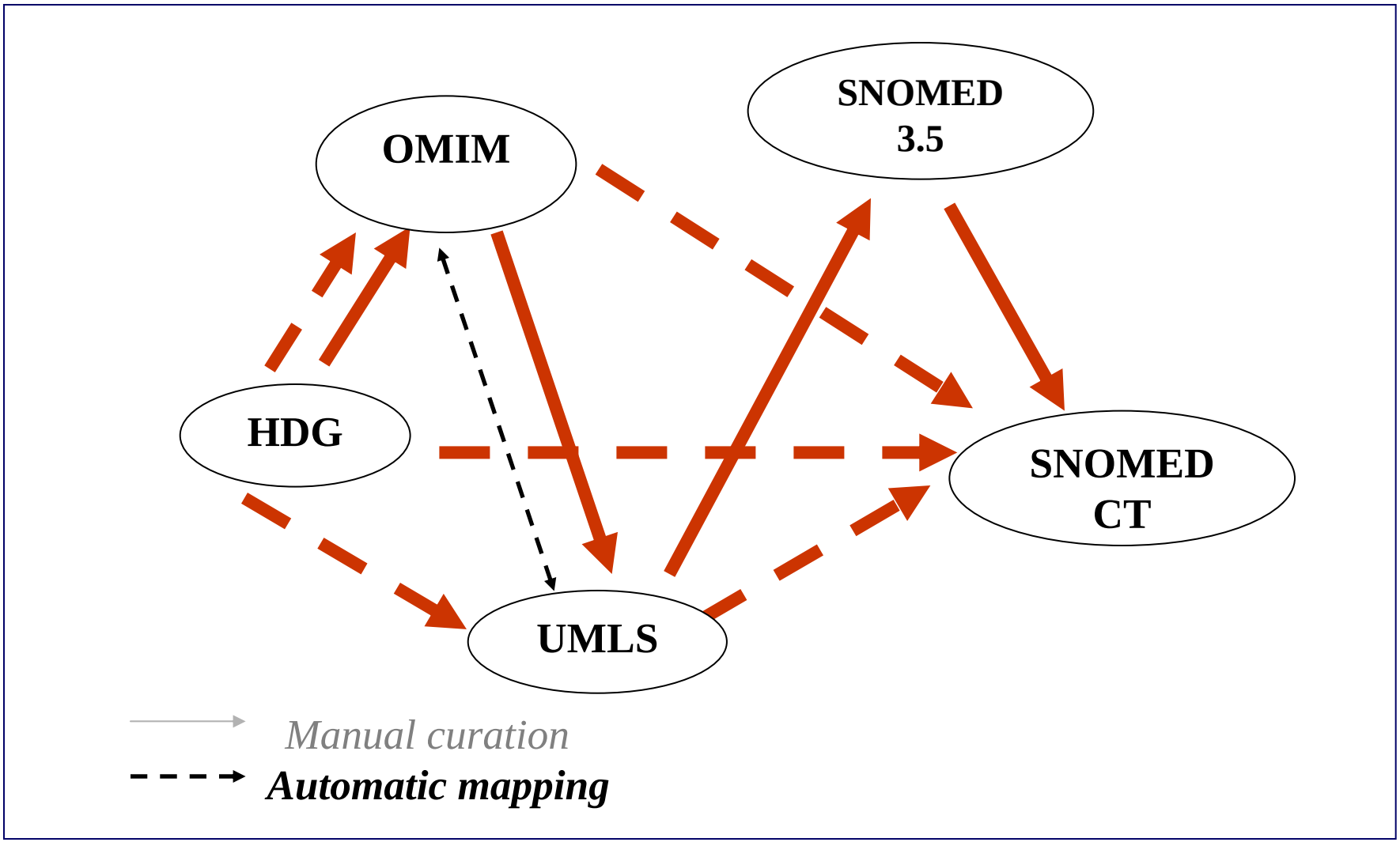
Method: Automated Terminology Network: ATN



Method: ATN


SNOMED
3.5
OMIM
P5
P5
HDG
SNOMED
CT
UMLS
Manual curation
Automatic mapping

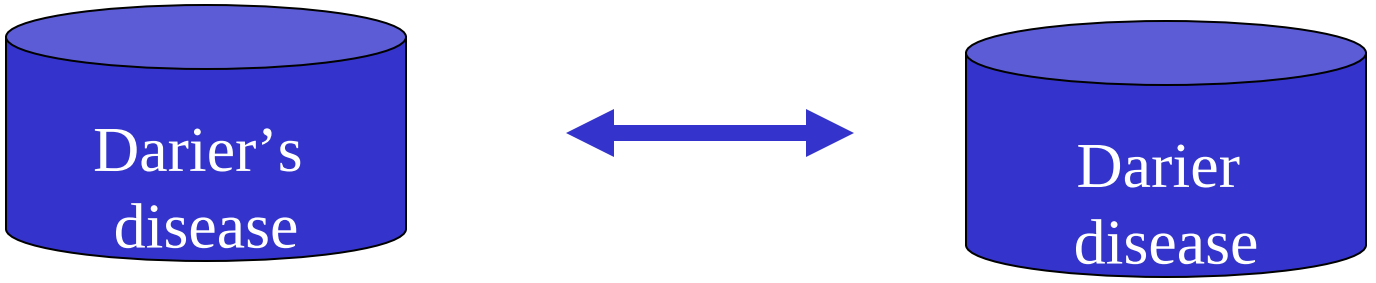
Method: Multistrategy ATN



Method: Lexico-Semantic techniques

Lexical Method: NORM

- Punctuations removed
- stop & duplicate words
- Conversion to base form
- Sort in alphabetical order



Method: how it works

HDG
306700
HEMOPHILIA A

OMIM
306700
HEMOPHILIA A
HEMOPHILIA, CLASSIC; HEMA
COAGULATION FACTOR VIIIC,
PROCOAGULANT COMPONENT,
COAGULATION FACTOR VIII, F8

UMLS
C0272322
AHG Deficiency
AHG deficiency disease
C0358603
Intermediate factor VIII|3|…

SNOMED-CT
16872008
Hereditary factor VIII
deficiency disease
319871002
Factor VIII fraction products

SNOMED3.5
16872008
AHG Deficiency
AHG deficiency disease
319871002
Factor VIII fraction products

Method: how it works

HDG
306700
HEMOPHILIA A

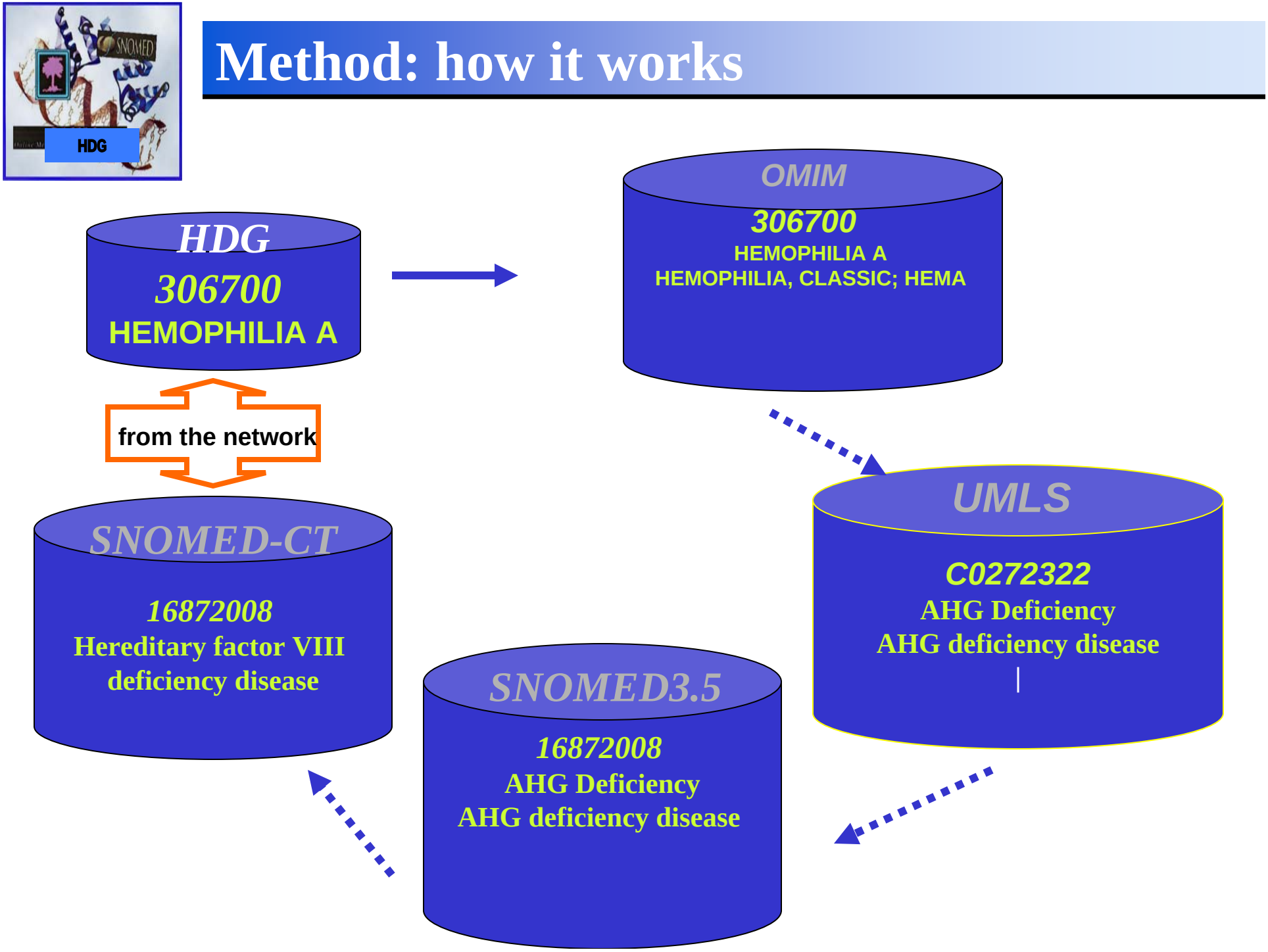
OMIM
306700
HEMOPHILIA A
HEMOPHILIA, CLASSIC; HEMA
COAGULATION FACTOR VIIC
PROCOAGULANT COMPONENT;
COAGULATION FACTOR VIII, P8

UMLS
C0272322
AHG Deficiency
AHG deficiency disease
C0259602
Intermediate factor VIII[2]

SNOMED-CT
16872008
Hereditary factor VIII deficiency disease
31587002
Factor VIII fraction products

SNOMED3.5
16872008
AHG Deficiency
AHG deficiency disease
31587002
Factor VIII fraction products

Method: how it works



Method: evaluation

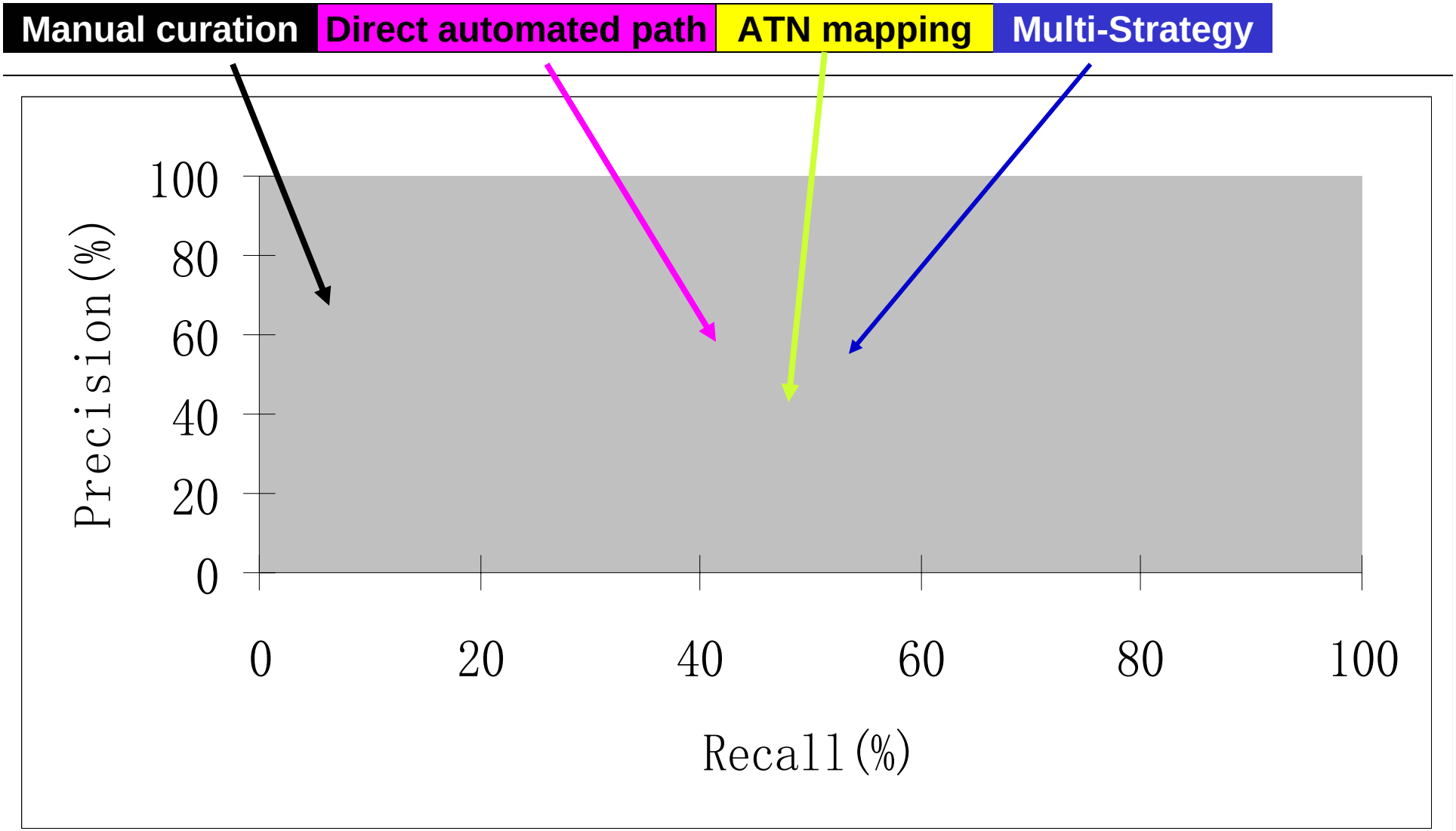
- **Gold Standard**
  - **3 independent curators**
  - **Agreement on 514 HDG-SNOMED maps**
- **Quantitative analysis**
  - **Recall: TP/ (TP+FN)**
  - **Precision: TP/(TP+FP)**

  TP = True positive, FN = false negative FP = False positive
- **Qualitative analysis**
  - **Ambiguity**
  - **Redundancy**

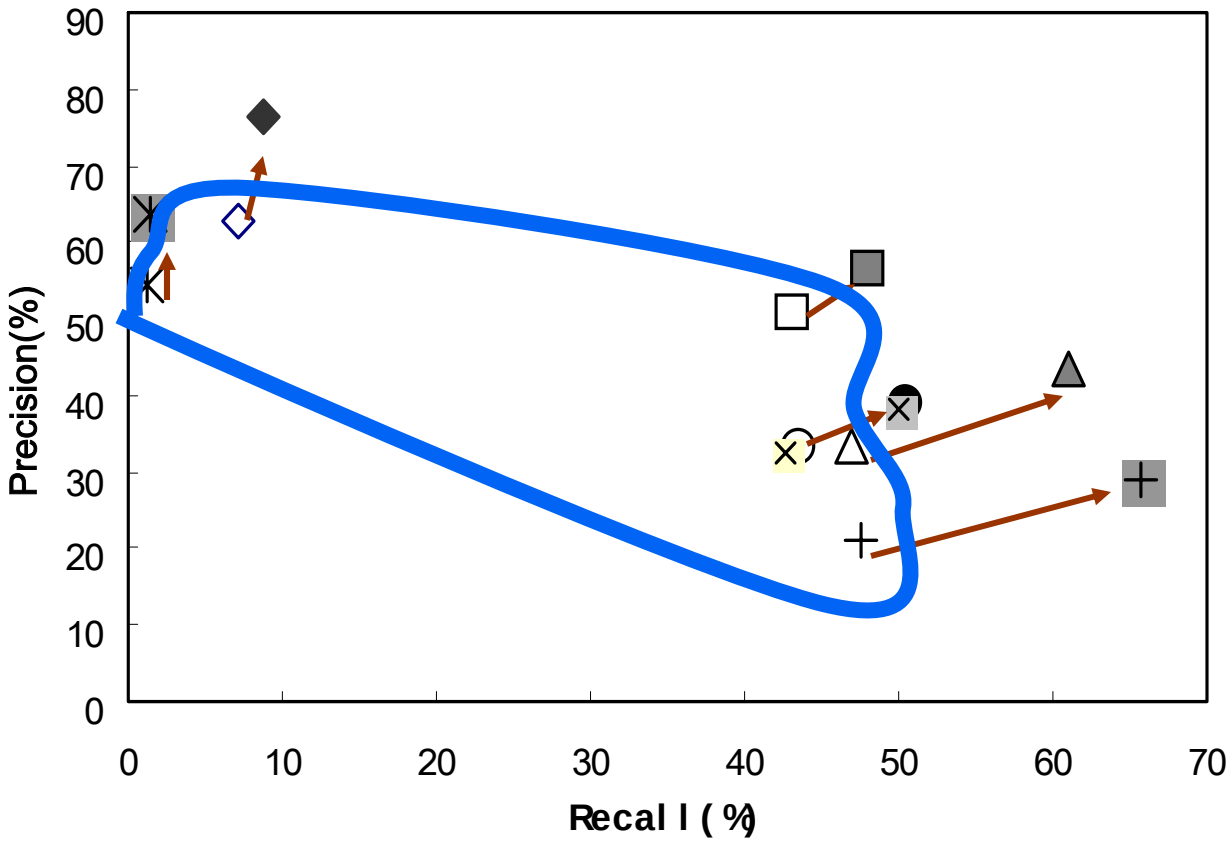
Outline

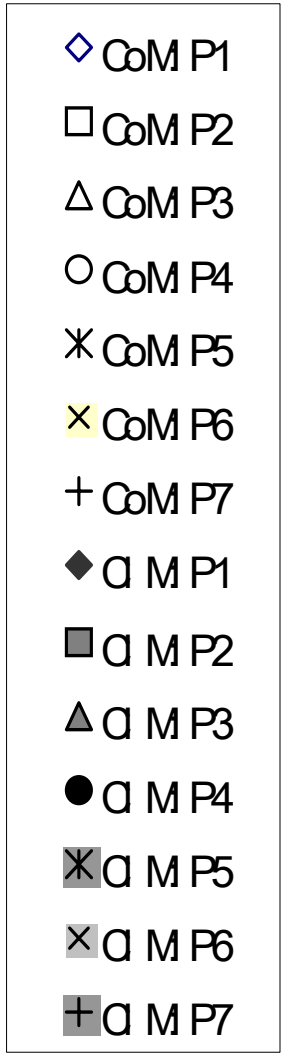
- **Challenge**
- **Introduction: Phenotypic Networks (PN)**
- **Hypotheses**
- **Methods**
- **Results**
  - **Accuracy of direct maps vs pathways**
  - **Accuracy of manual vs automated curation**
  - **Accuracy of multi-strategy method**
- **Conclusions**

Result: Quantitative analysis



Precision vs. recall of each of the linking paths in the ATN

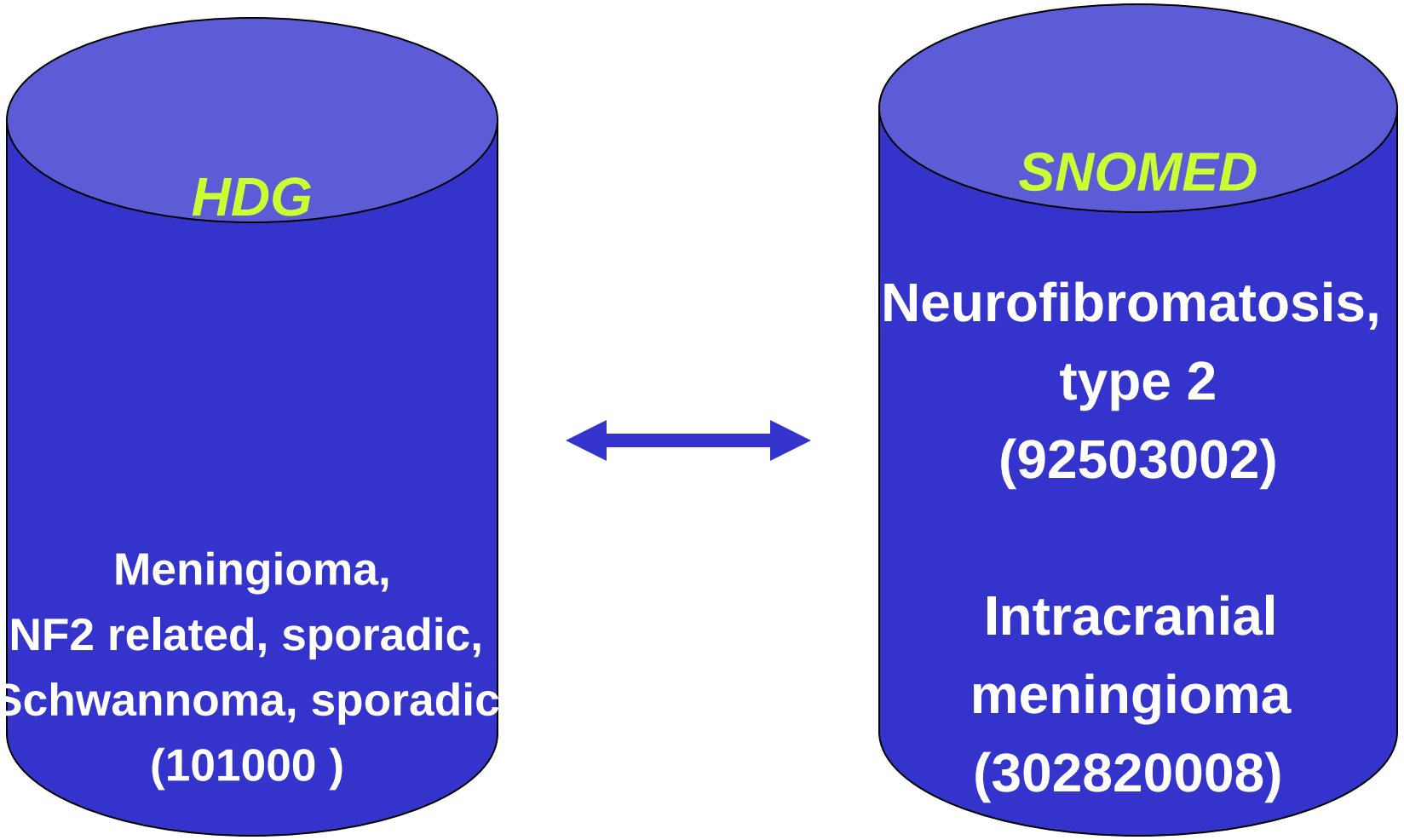




CoM: Concept-Based Mapping; ClM: Class-Based Mapping

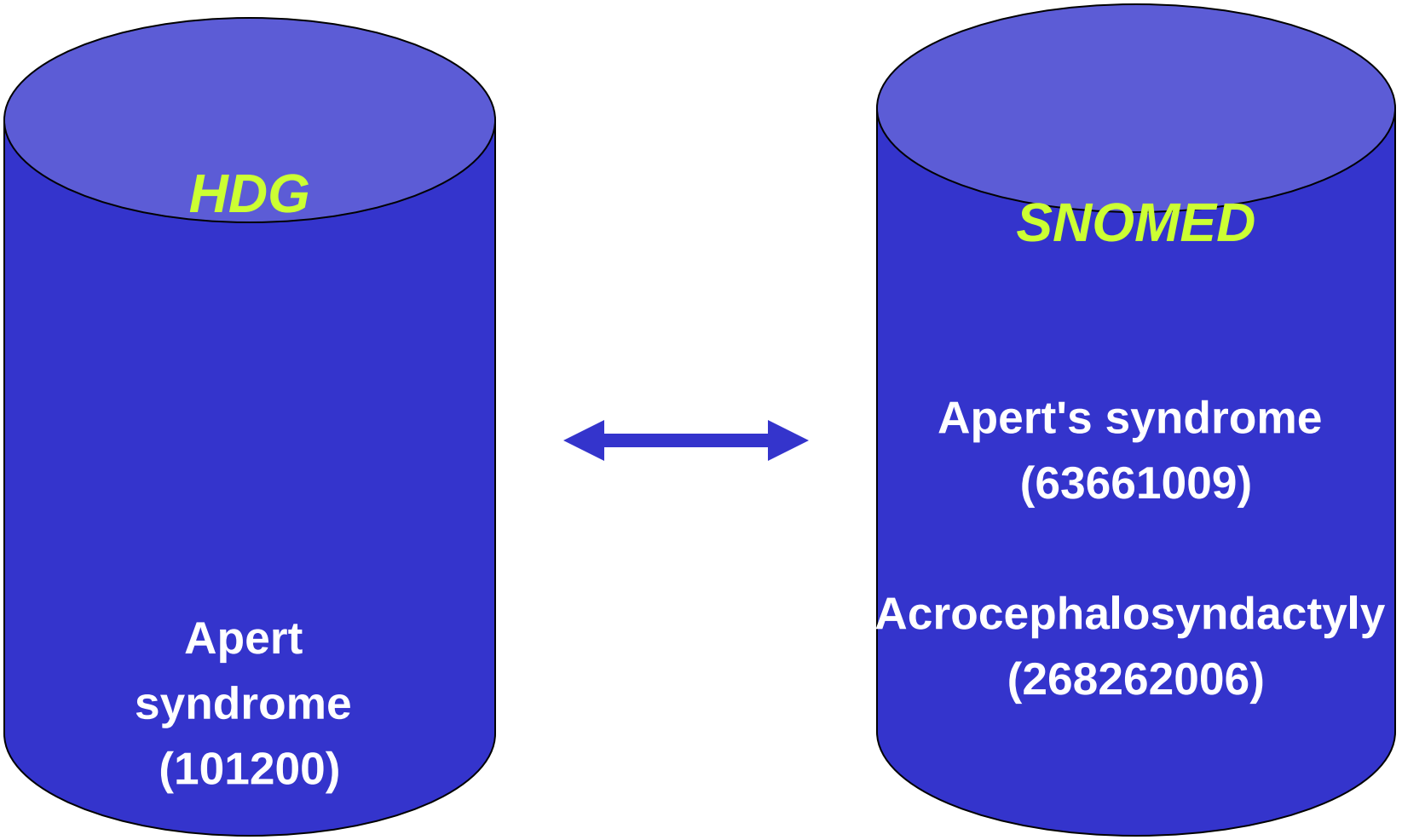
Result: Qualitative Analysis: P1

Ambiguity in HDG: 18%



Result: Qualitative Analysis: P1

Redundancy in SNOMED: 15%



Conclusions

- **Automated mapping traversing a network of terminologies can have significantly improved recall (six fold increase in this study) over that of manual indexes, with minimal impact on precision.**
- **Direct automated mapping (non-network) performed significantly worse than any other method.**
- **Incremental and class-based methods not investigated in this study, have shown in a previous study to increase precision.**
- **Automated terminology Networks may allow for high-throughput linkages between disparate biomedical databases**

Limitations

- **Small GS**
- **Compositional mapping has not been addressed with these methods**

Future Directions

- **Support compositional mapping**
- **Predict the accuracy of terminological pathways large-scale networks.**

Acknowledgments

- Trainees: Michael Cantor, Aylit Schultz, Hui Nar Quek
- Staff: Jianrong Li

- National Institute of Allergy and Infectious Diseases (NIAID),
- New York State Office of Science, Technology, and Academic Research (NYSTAR)-sponsored Center for Advanced Technology at Columbia University
- Office of Advanced Telemedicine (OAT) of the Health Resources and Services Administration (HRSA),
- Virginia Commonwealth University's Medical Informatics and Technology Applications Consortium, a National Aeronautics and Space Administration (NASA) Commercial Space Center.

Thank you!