April 25, 2022
First International Workshop on **Semantics-enabled Biomedical Literature Analytics** (SeBiLAn2022)

Powering Semantic Analysis with Bio-ontologies

Olivier Bodenreider

Acting Director, Lister Hill National Center for Biomedical Communications

National Institutes of Health, National Library of Medicine, Bethesda, Maryland, USA

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.

Outline

- Introduction to the Lister Hill National Center for Biomedical Communications
- Bio-ontologies as a source of semantic information
- UMLS and semantic interoperability
- Role of bio-ontologies in biomedical text processing

Introduction to the Lister Hill National Center for Biomedical Communications

National Library of Medicine

- Largest biomedical library in the world
- Started in 1836 as a small collection of medical books and journals in the office of the United States Army Surgeon General
- Part of the National Institutes of Health since 1962
- Flagship products and services (among many others)
  - PubMed/MEDLINE
  - ClinicalTrials.gov
  - Unified Medical Language System (UMLS)

National Library of Medicine

- Curating data, not just (dusty) books
  - Biomedical literature
  - Gene sequences
  - Clinical trials
  - Information for lay public
  - […]
- Research, not just products and services
  - NLM Intramural Research program
    - Computational Biology – National Center for Biotechnology Information (NCBI)
    - Computational Health – Lister Hill National Center for Biomedical Communications
  - NLM Intramural Training program
- Extramural programs (grants)

Lister Hill National Center for Biomedical Communications (LHC)

- With NCBI, one of the two research & development centers of NLM
- Established in 1968
- Initially focused on biomedical communications
- Later reorganized around health informatics
  - Clinical data science
  - Interoperable data
  - Development of scalable methods for clinical text and images
  - Translation of research insights into operations

LHC activities

- Natural Language Processing
  - **Identifying biomedical concepts and relations in clinical text / literature**
  - Clinical question answering
- Image processing
  - Application of machine learning/deep learning techniques to imaging datasets to support diagnostics
- Health information standards
  - **Terminology standards (UMLS, SNOMED CT, MeSH, RxNorm, LOINC, …)**
  - Information model standards (common data models, FHIR – Fast Healthcare Interoperability Resource)
- Health data-powered discovery
  - Getting insights from large observational databases (EHR and claims data)

Bio-ontologies as a source of semantic information

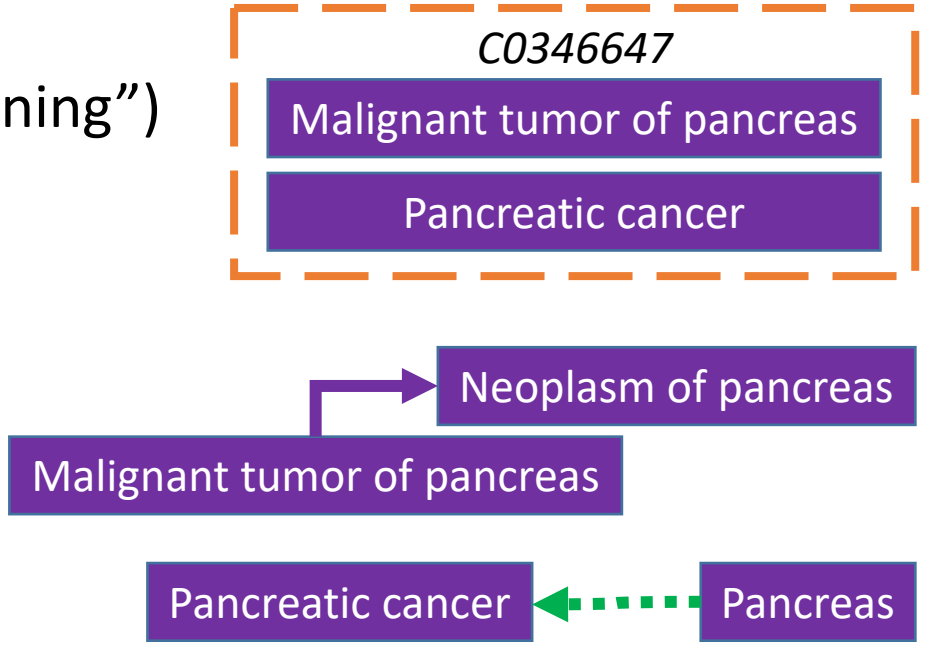
Bio-ontologies

- Ontologies are a technique or technology used to represent and share knowledge about a domain by modeling **the things in that domain** and **the relationships between those things**
  - The labels used for the things provide a language for a community to talk about their domain
  - By agreeing on a particular ontological representation, a common vocabulary can be used to describe and ultimately analyze data
- Ontologies vs. terminologies
  - Many ontologies have terminological features
  - Many terminologies have ontological features

Brief Bioinform. 2006 Sep; 7(3): 256–274.

Bio-ontologies and semantics

- Bio-ontologies contain more than the vocabulary of the biomedical domain
  - Grouping of terms into concepts (“unit of meaning”)
    - Synonymy relations
  - Concept categorization
    - Diseases, proteins, …
  - Relations among concepts
    - Hierarchical relations: subclass (isa) relations
    - Associative (“transversal”) relations
  - High-level conceptualization of the domain
    - Drugs *treat* Diseases



Examples from 3 bio-ontologies

- Medical Subject Headings (MeSH)
  - Developed by the National Library of Medicine
  - Used for indexing and retrieval of the biomedical literature
- Gene Ontology (GO)
  - Developed by the GO Consortium
  - Used for annotating gene products and reasoning with biological information
- SNOMED CT
  - Developed by SNOMED International
  - Used for coding and exchanging clinical information and for clinical analytics

MeSH Words…

https://meshb.nlm.nih.gov/

Alzheimer Disease MeSH Descriptor Data 2022

Details | Qualifiers | MeSH Tree Structures | Concepts

Alzheimer Disease *Preferred*

Expand All

Concept UI M0000842

Scope Note A degenerative disease of the BRAIN characterized by the insidious onset of DEMENTIA. Impairment of MEMORY, judgment, attention span, and problem solving skills are followed by severe APRAXIAS and a global loss of cognitive abilities. The condition primarily occurs after age 60, and is marked pathologically by severe cortical atrophy and the triad of SENILE PLAQUES; NEUROFIBRILLARY TANGLES; and NEUROPIL THREADS. (From Adams et al., Principles of Neurology, 6th ed, pp1049-57)

Terms
- **Alzheimer Disease** *Preferred Term*
- **Alzheimer Dementia**
- **Alzheimer's Disease**
- **Dementia, Senile**
- **Dementia, Alzheimer Type**
- **Alzheimer-Type Dementia (ATD)**
- **Alzheimer Type Senile Dementia**
- **Primary Senile Degenerative Dementia**
- **Dementia, Primary Senile Degenerative**
- **Alzheimer Sclerosis**
- **Alzheimer Syndrome**
- **Alzheimer's Diseases**
- **Senile Dementia, Alzheimer Type**

Acute Confusional Senile Dementia *Narrower*
Dementia, Presenile *Narrower*
Alzheimer Disease, Late Onset *Narrower*
Alzheimer's Disease, Focal Onset *Narrower*
Familial Alzheimer Disease (FAD) *Narrower*
Alzheimer Disease, Early Onset *Narrower*

page delivered in 0.163s

MeSH … and more than words

Nervous System Diseases [C10]
- Central Nervous System Diseases [C10.228]
  - Brain Diseases [C10.228.140]
    - Dementia [C10.228.140.380]
      - AIDS Dementia Complex [C10.228.140.380.070]
      - **Alzheimer Disease [C10.228.140.380.100]**
      - Aphasia, Primary Progressive [C10.228.140.380.132] ⊕
      - Creutzfeldt-Jakob Syndrome [C10.228.140.380.165]
      - Dementia, Vascular [C10.228.140.380.230] ⊕
      - Diffuse Neurofibrillary Tangles with Calcification [C10.228.140.380.254]
      - Frontotemporal Lobar Degeneration [C10.228.140.380.266] ⊕
      - Huntington Disease [C10.228.140.380.278]
      - Kluver-Bucy Syndrome [C10.228.140.380.326]
      - Lewy Body Disease [C10.228.140.380.422]

Nervous System Diseases [C10]
- Neurodegenerative Diseases [C10.574]
  - Tauopathies [C10.574.945]
    - **Alzheimer Disease [C10.574.945.249]**
    - Corticobasal Degeneration [C10.574.945.312]
    - Diffuse Neurofibrillary Tangles with Calcification [C10.574.945.374]
    - Supranuclear Palsy, Progressive [C10.574.945.500]

Mental Disorders [F03]
- Neurocognitive Disorders [F03.615]
  - Dementia [F03.615.400]
    - AIDS Dementia Complex [F03.615.400.050]
    - **Alzheimer Disease [F03.615.400.100]**
    - Aphasia, Primary Progressive [F03.615.400.125] ⊕
    - Creutzfeldt-Jakob Syndrome [F03.615.400.300]
    - Dementia, Vascular [F03.615.400.350] ⊕
    - Diffuse Neurofibrillary Tangles with Calcification [F03.615.400.370]
    - Frontotemporal Lobar Degeneration [F03.615.400.380] ⊕
    - Huntington Disease [F03.615.400.390]
    - Kluver-Bucy Syndrome [F03.615.400.431]
    - Lewy Body Disease [F03.615.400.512]

Gene Ontology Words…

https://www.ebi.ac.uk/QuickGO/

GO:0045597 P JSON

positive regulation of cell differentiation

Biological Process

Definition (GO:0045597 GONUTS page)

Any process that activates or increases the frequency, rate or extent of cell differentiation.

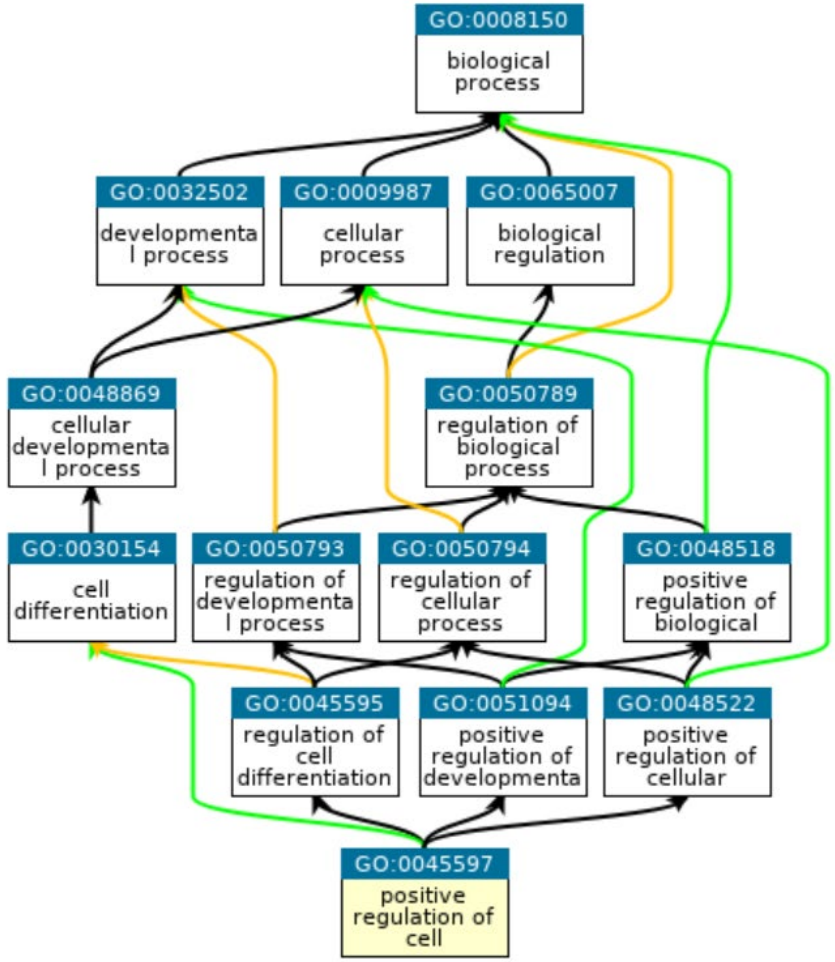
68,041 annotations

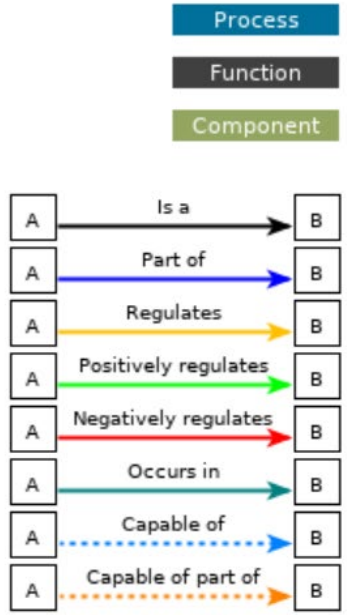
Synonyms

Synonyms are alternative words or phrases closely related in meaning to the term name, with indication of the relationship between the name and synonym given by the synonym scope.

Synonym	Type
upregulation of cell differentiation	exact
up-regulation of cell differentiation	exact
activation of cell differentiation	narrow
up regulation of cell differentiation	exact
stimulation of cell differentiation	narrow

Gene Ontology … and more than words



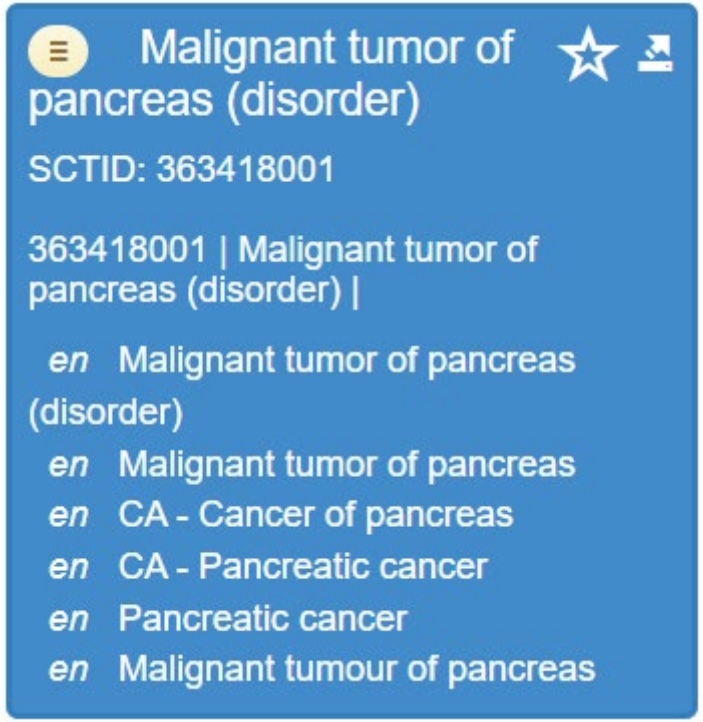


Gene Product	Symbol	Qualifier	GO Term	Evidence
UniProtKB:A0A015KU87	BG52_00880	involved_in	GO:0045881 P positive regulation of sporulation resulting in formation of a cellular spore	ECO:0000256 IEA
UniProtKB:A0A023CT81	H839_00355	involved_in	GO:0045881 P positive regulation of sporulation resulting in formation of a cellular spore	ECO:0000256 IEA

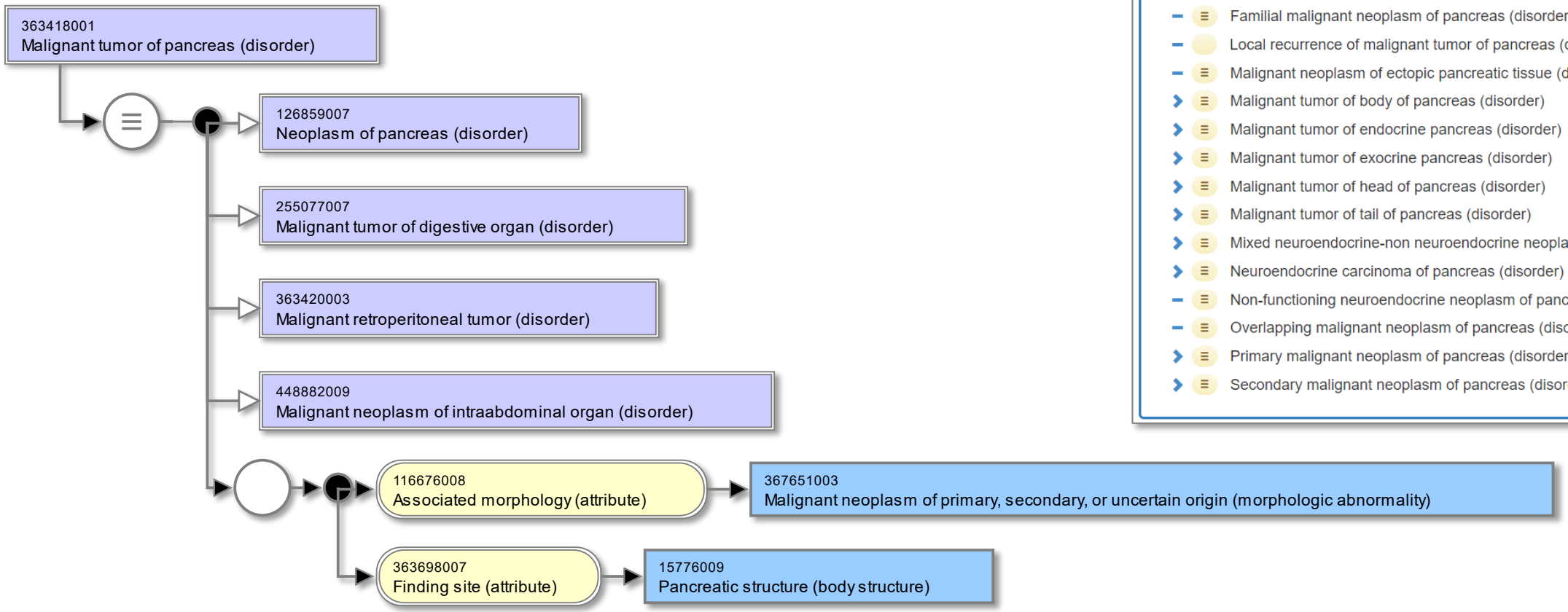
SNOMED CT Words…

SNOMED International Leading healthcare terminology, worldwide

https://browser.ihtsdotools.org/



SNOMED CT … and more than words



UMLS and semantic interoperability

Many bio-ontologies

- Different purposes
  - Clinical documentation – fine grained
  - Morbidity and mortality statistics – classification (avoid double-counting)
  - Indexing/retrieval – abstraction
  - Text mining – lexical variation
- Developed independently
  - Standard Development Organizations
  - No standard for developing standards
  - Different funding mechanisms
  - Different legacy products

Degrees of semantic interoperability

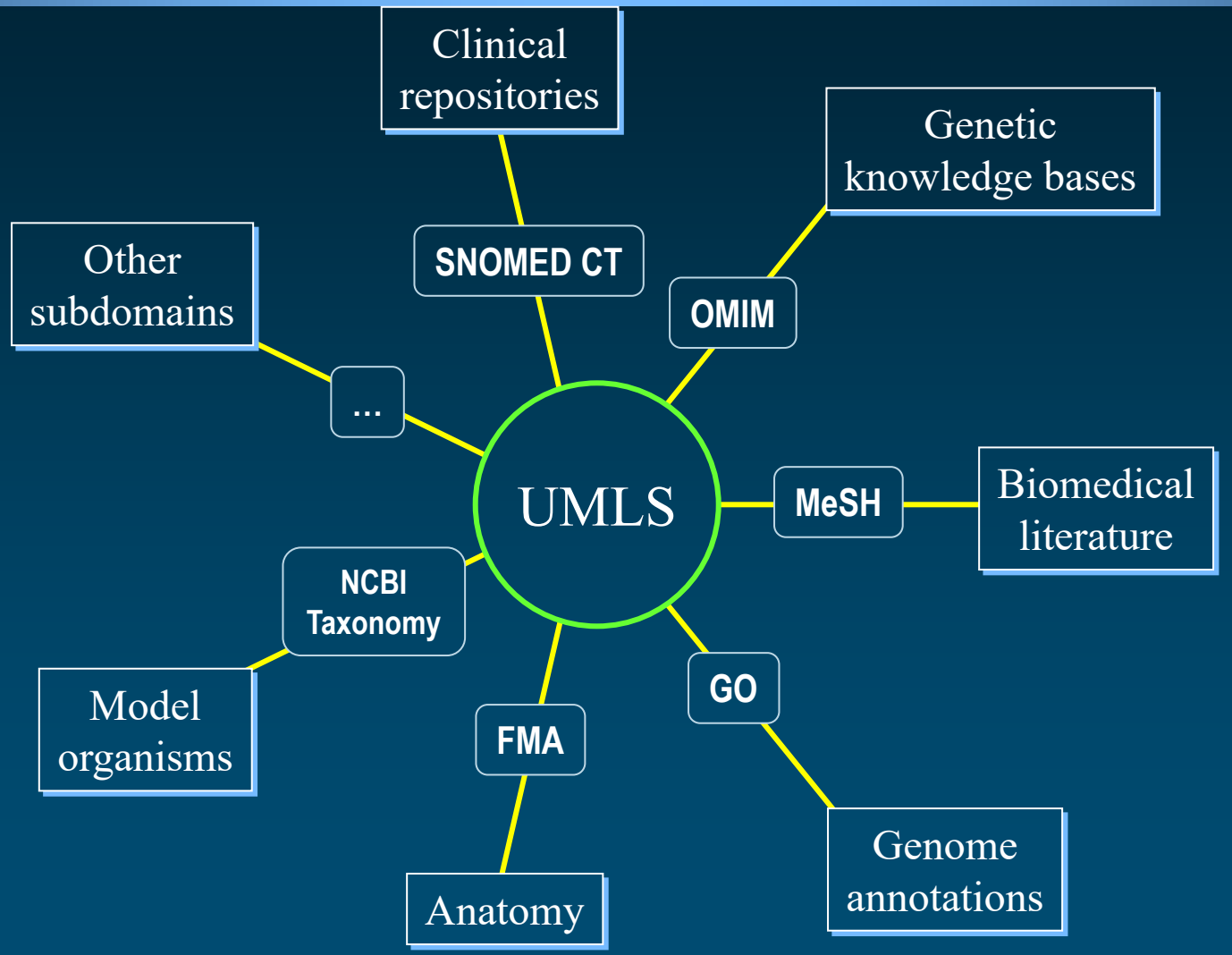
- Synonymy
  - Equivalence between terms (or concepts)
  - Myocardial infarction ↔ Heart attack
- Mapping
  - Closest term for the source term in the target terminology
  - Lipitor → Atorvastatin
- Closest ancestor
  - Closest term in the target terminology among the ancestors in the source
  - Pancreatic cancer → Pancreatic neoplasm
- Post-coordination
  - One term equivalent to the combination of several terms in the target terminology
  - Diabetic nephropathy → Nephropathy + Diabetes mellitus

UMLS Metathesaurus

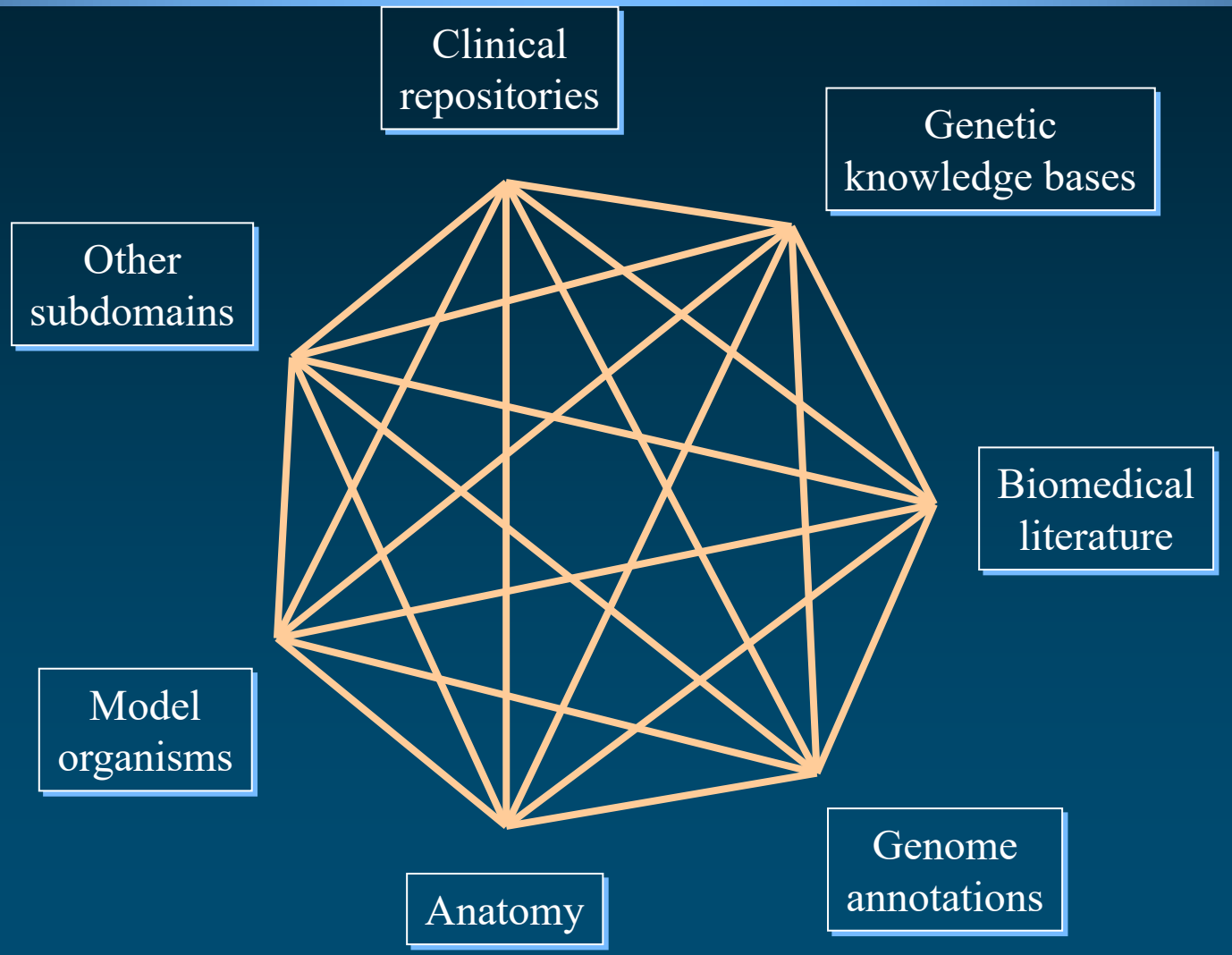
(2021AB)

- Unified Medical Language System
- 158 families of source vocabularies
  - Not counting 58 translations
- 25 languages
- Broad coverage of biomedicine
  - 12.8M names (normalized)
  - ~4.5M concepts
  - >10M relations
- Common presentation

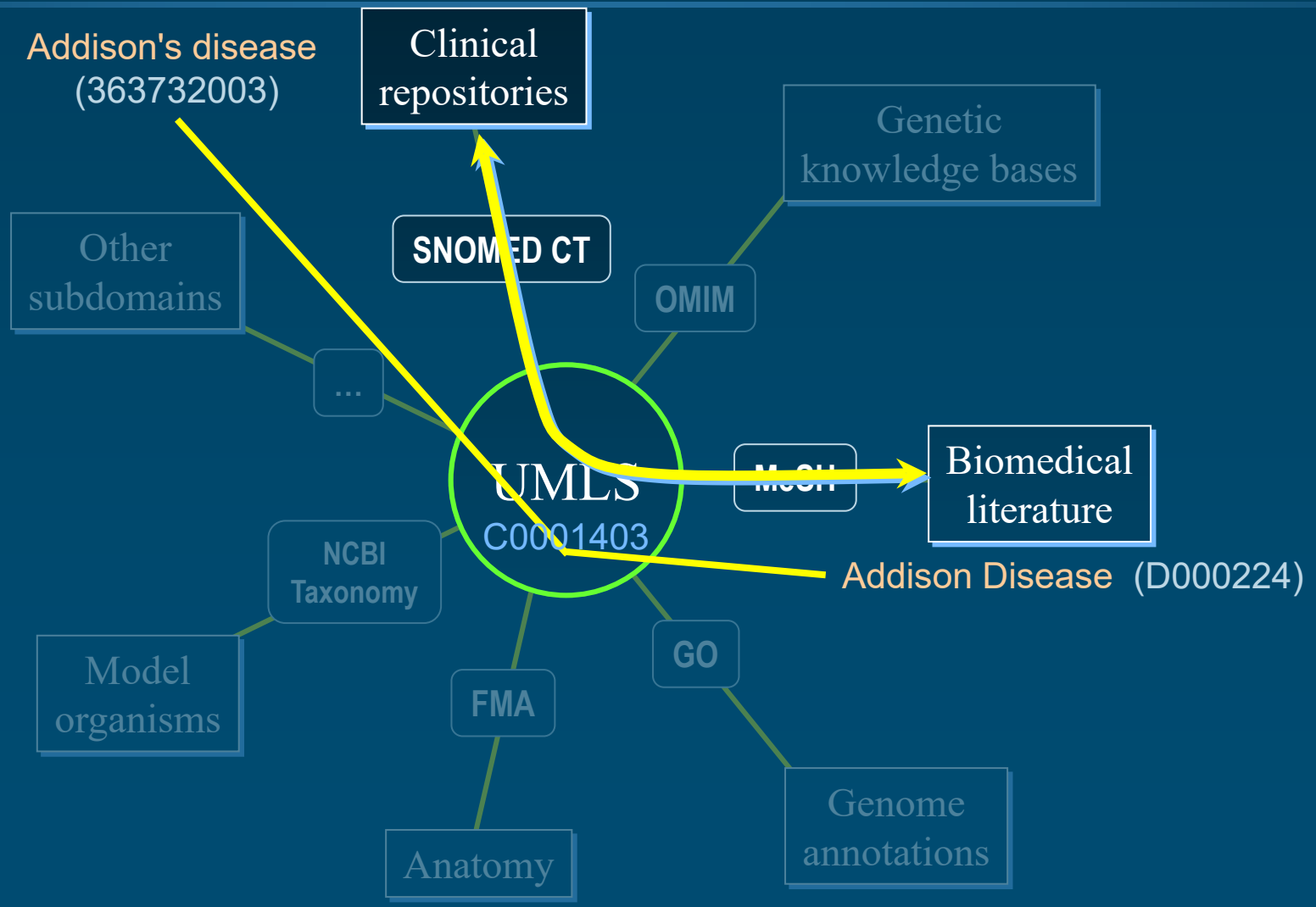
Integrating datasets through terminology integration



Point-to-point mappings are impractical



Integration through a reference (e.g., UMLS)



Semantic interoperability through UMLS

- Synonymy
  - *Synonymous terms clustered into the same ULMS concept*
  - Myocardial infarction ↔ Heart attack
- Mapping
  - *Existing mapping tables integrated into UMLS (e.g., ICD10 to SNOMED CT)*
  - Lipitor → Atorvastatin
- Closest ancestor
  - *Hierarchical relations are recorded in UMLS and can be navigated*
  - Pancreatic cancer → Pancreatic neoplasm
- Post-coordination
  - *Logical definitions for concepts recorded in UMLS (whenever available)*
  - Diabetic nephropathy → Nephropathy + Diabetes mellitus

Role of bio-ontologies in biomedical text processing

illustrated through some flagship NLM resources

Some NLM resources for text processing

- Applications
  - Named entity recognition (MetaMap)
  - Relation extraction (SemRep)
- Supporting resources
  - UMLS Metathesaurus
  - UMLS Semantic Network
- Resulting datasets
  - SemMedDB

https://lhncbc.nlm.nih.gov/ii/tools/MetaMap.html

https://lhncbc.nlm.nih.gov/ii/tools/SemRep_SemMedDB_SKR/SemRep.html

https://www.nlm.nih.gov/research/umls/index.html

https://lhncbc.nlm.nih.gov/semanticnetwork/

https://lhncbc.nlm.nih.gov/ii/tools/SemRep_SemMedDB_SKR/SemRep.html

Neurofibromatosis 2

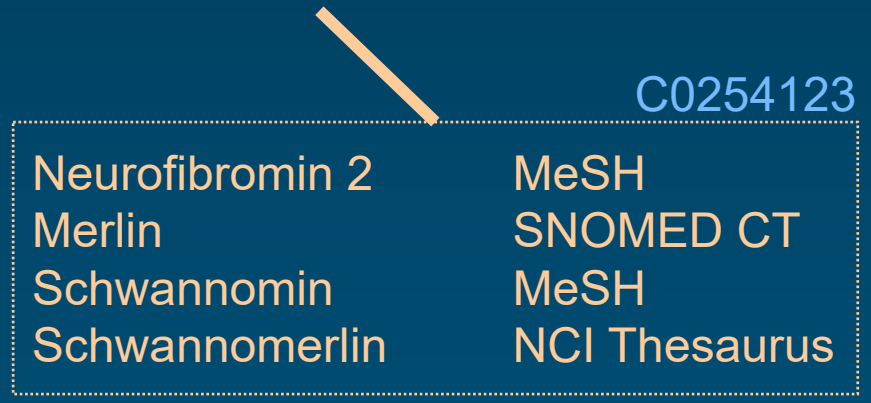
Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.

[Uppal, S., and A. P. Coatesworth. "Neurofibromatosis Type 2." *Int J Clin Pract*, 57, no. 8, 2003, pp. 698-703.]

Entity recognition and normalization (linking)

Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.

C0027832 C0027832 C0027831 C0027832 C0027859 C0027832 C0026882 C0254123 C0008665



Lexical/terminological resources

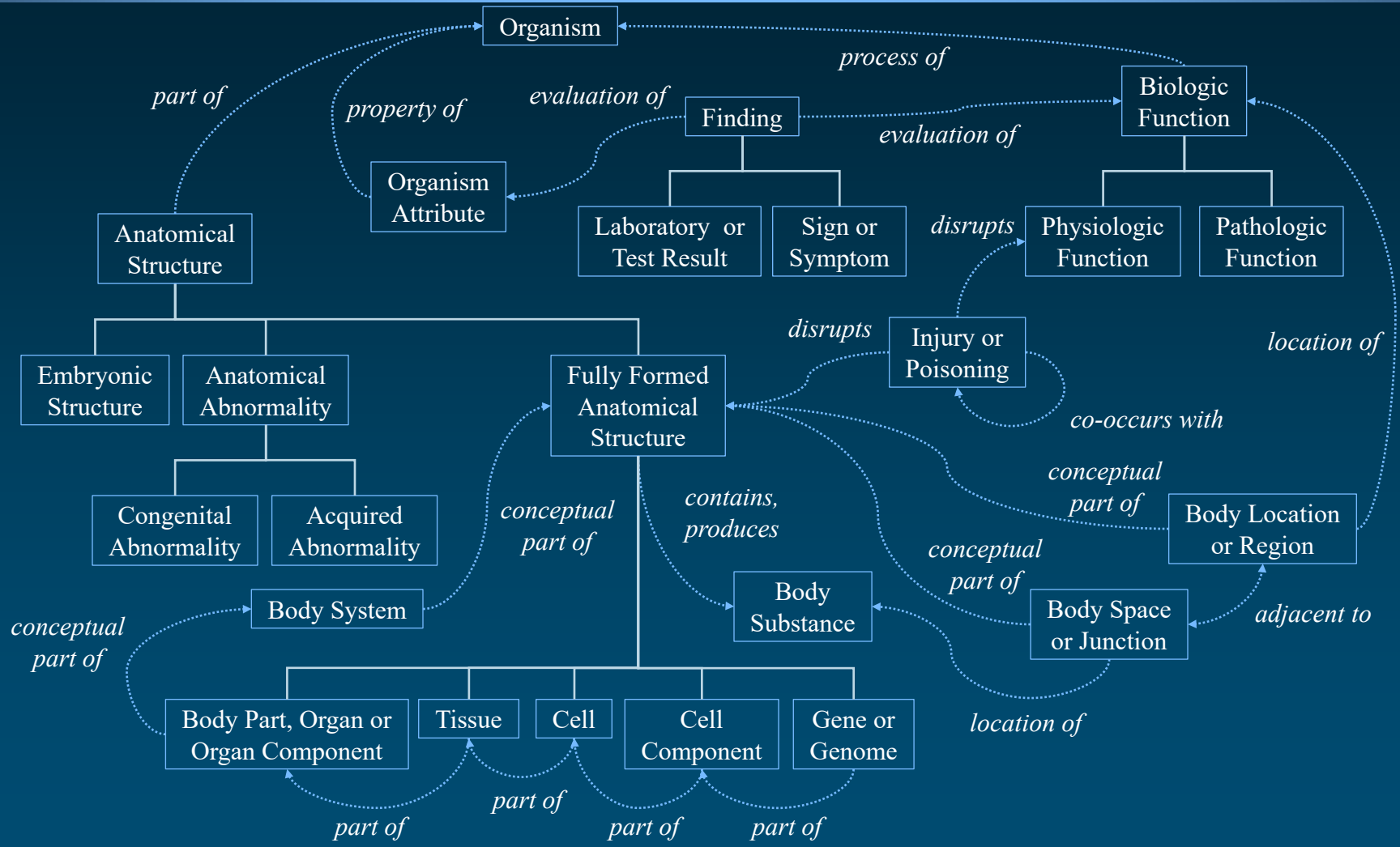
Relation extraction

Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.

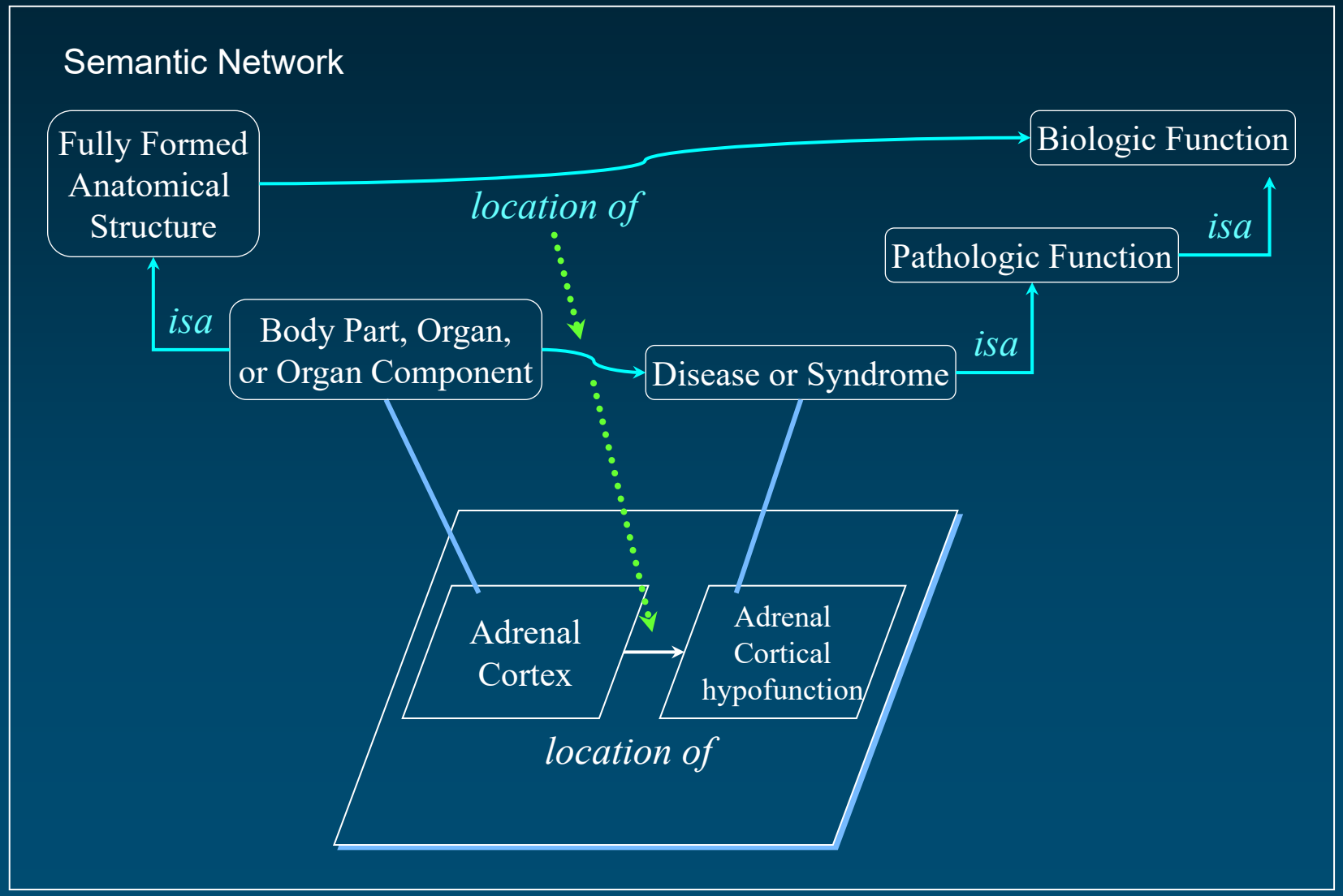
- vestibular schwannomas *manifestation of* neurofibromatosis 2
- neurofibromatosis 2 *associated with* mutation of NF2 gene
- NF2 gene *located on* chromosome 22

Ontologies

UMLS Semantic Network – 127 semantic types linked by isa and associative relationships



A semantic scaffolding for relation extraction



Summary

Semantic continuum

- Lexical resources
  - Collections of lexical items
  - Additional information
    - Part of speech
    - Spelling variants
  - Useful for entity recognition
  - UMLS SPECIALIST Lexicon

- Ontological resources
  - Collections of
    - kinds of entities (substances, qualities, processes)
    - relations among them
  - Useful for relation extraction
  - UMLS Semantic Network

- Terminological resources
  - Collections lexical items + identifiers
  - Useful for entity resolution
  - UMLS Metathesaurus

Parting thoughts – Bio-ontologies and AI

Do we still need bio-ontologies when semantics can be learned from vast amounts of textual data through neural networks?

Can we use neural networks to build better bio-ontologies?

olivier.bodenreider@nih.gov