

W3C Semantic Web  
Health Care and Life Sciences Interest Group  
BioRDF Teleconference  
June 5, 2006

From terminology integration  
to information integration

*Unified Medical Language System (UMLS)*



*Olivier Bodenreider*

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

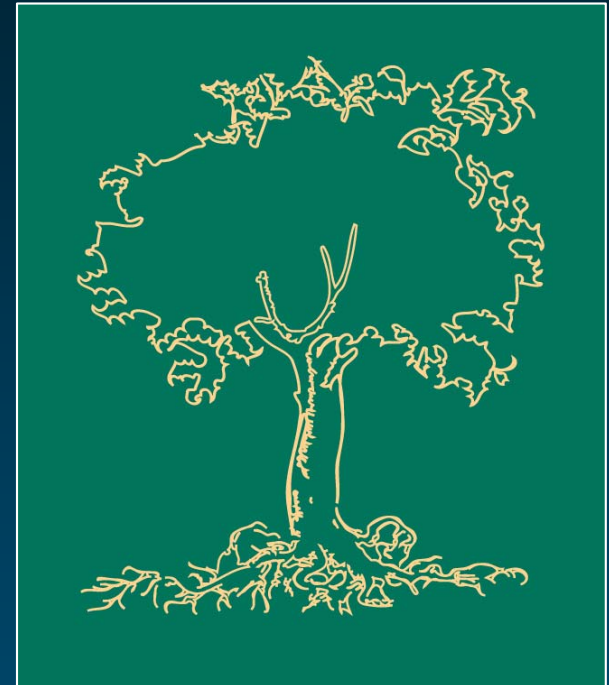
# Outline

- ◆ Overview through an example
- ◆ UMLS components
  - Lexical resources
  - Metathesaurus
  - Semantic Network
- ◆ UMLS and information integration



# What does UMLS stand for?

- ◆ Unified
- ◆ Medical
- ◆ Language
- ◆ System



UMLS<sup>®</sup>  
Unified Medical Language System<sup>®</sup>  
UMLS Metathesaurus<sup>®</sup>



# Motivation

- ◆ Started in 1986
- ◆ National Library of Medicine
- ◆ “Long-term R&D project”

«[...] the UMLS project is an effort to overcome two significant barriers to effective retrieval of machine-readable information.

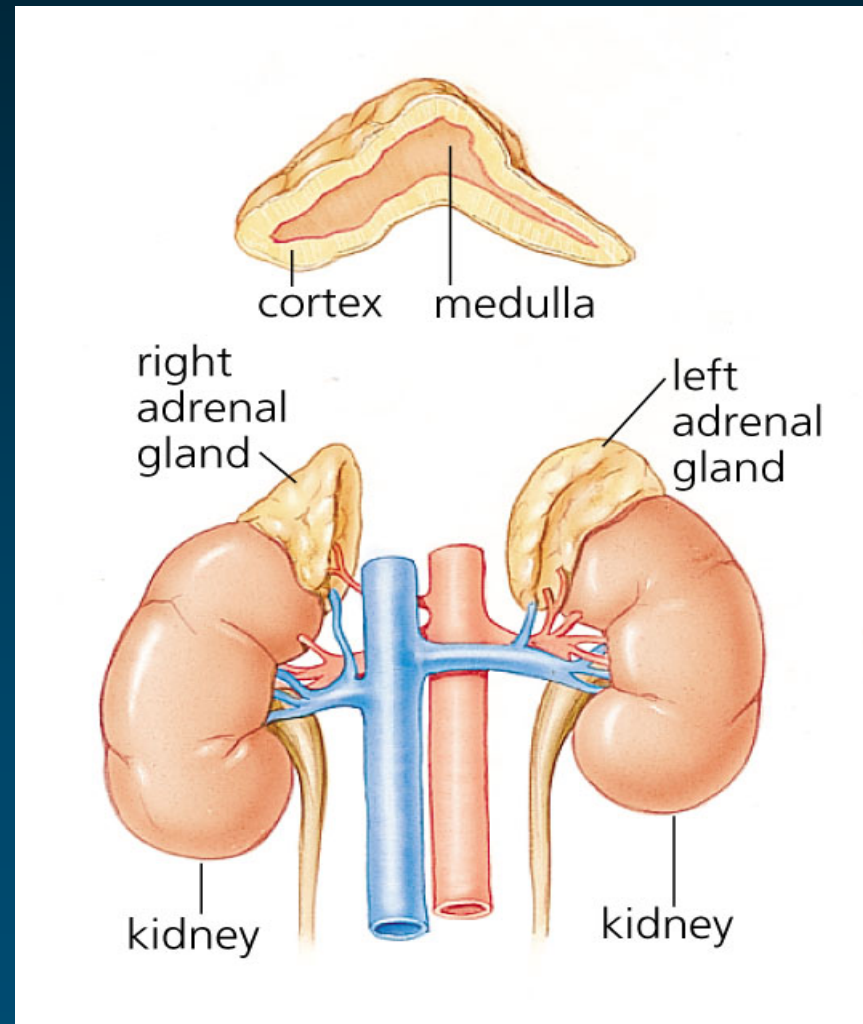
- The first is the variety of ways the same concepts are expressed in different machine-readable sources and by different people.
- The second is the distribution of useful information among many disparate databases and systems.»



Overview through an example

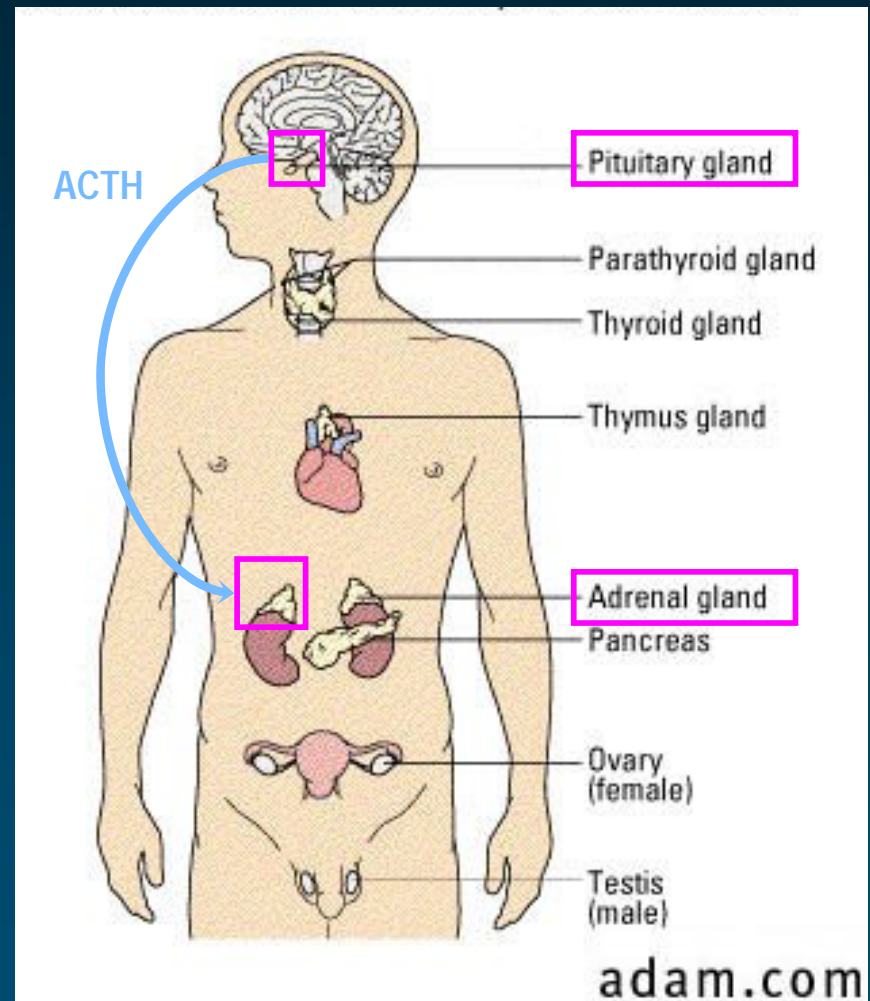
# Addison's disease

- ◆ Addison's disease is a rare endocrine disorder
- ◆ Addison's disease occurs when the adrenal glands do not produce enough of the hormone cortisol
- ◆ For this reason, the disease is sometimes called chronic adrenal insufficiency, or hypocortisolism



# Adrenal insufficiency Clinical variants

- ◆ Primary / Secondary
  - Primary: lesion of the adrenal glands themselves
  - Secondary: inadequate secretion of ACTH by the pituitary gland
- ◆ Acute / Chronic
- ◆ Isolated / Polyendocrine deficiency syndrome



# Addison's disease: Symptoms

- ◆ Fatigue
- ◆ Weakness
- ◆ Low blood pressure
- ◆ Pigmentation of the skin (exposed and non-exposed parts of the body)
- ◆ ...



# AD in medical vocabularies

## ◆ Synonyms: different terms

- Addisonian syndrome
  - Bronzed disease
  - Addison melanoderma
  - Asthenia pigmentosa
  - Primary adrenal deficiency
  - Primary adrenal insufficiency
  - Primary adrenocortical insufficiency
  - Chronic adrenocortical insufficiency
- } eponym
- } symptoms
- } clinical variants

## ◆ Contexts: different hierarchies



# Organize terms

- ◆ Synonymous terms clustered into a concept
- ◆ Preferred term
- ◆ Unique identifier (CUI)

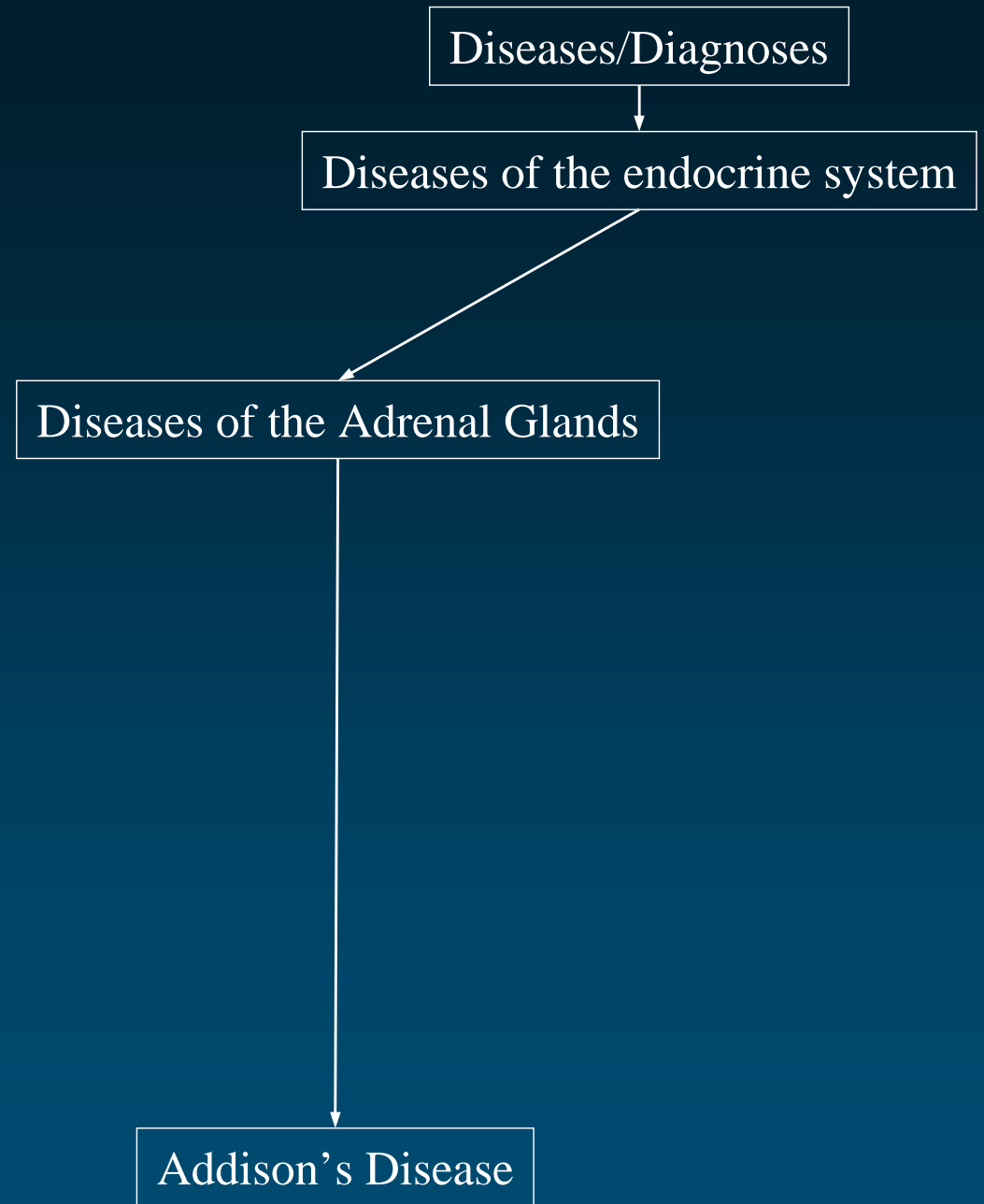
Addison Disease	MeSH	D000224
Primary hypoadrenalism	MedDRA	10036696
Primary adrenocortical insufficiency	ICD-10	E27.1
Addison's disease (disorder)	SNOMED CT	363732003

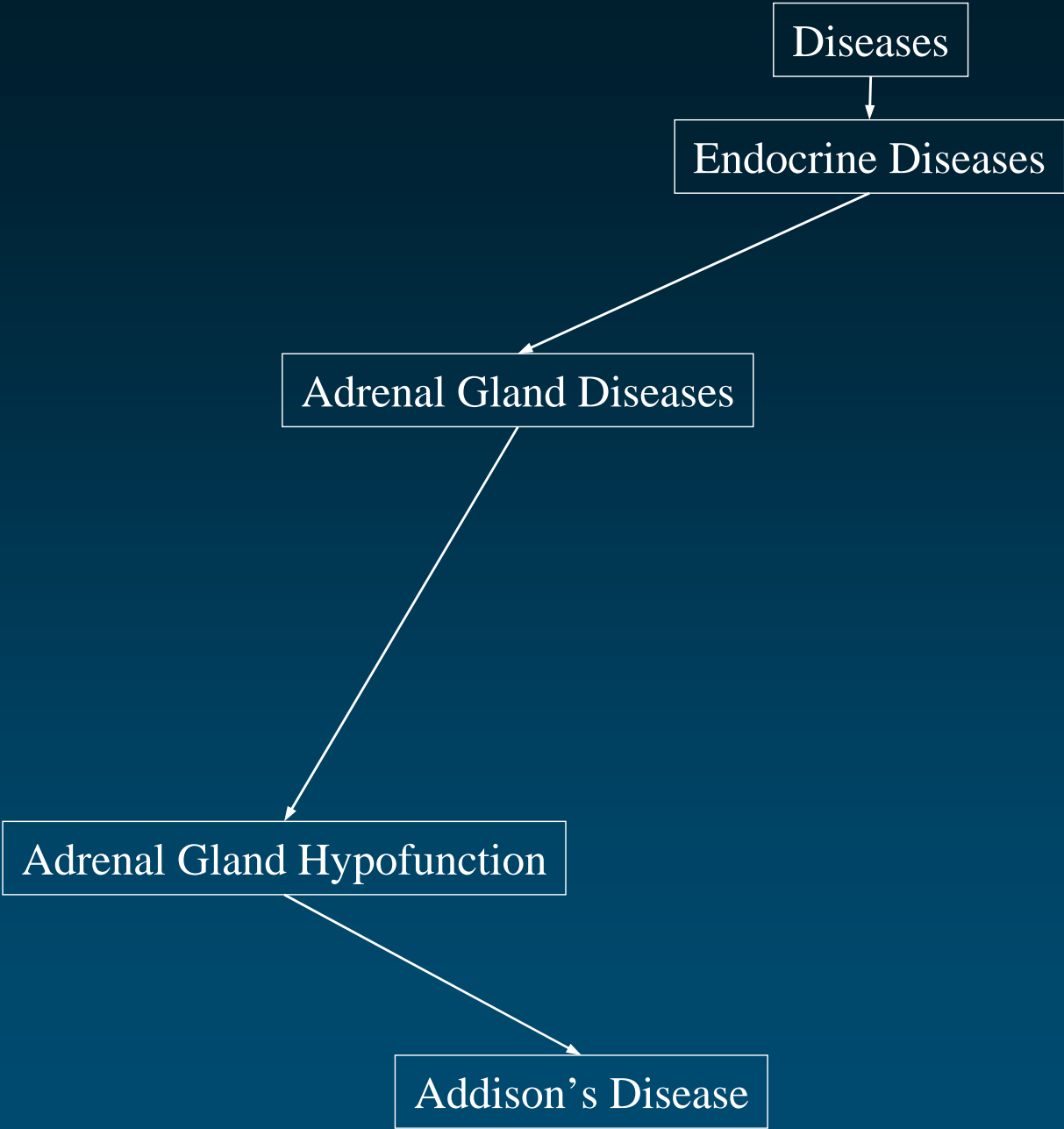
C0001403

Addison's disease

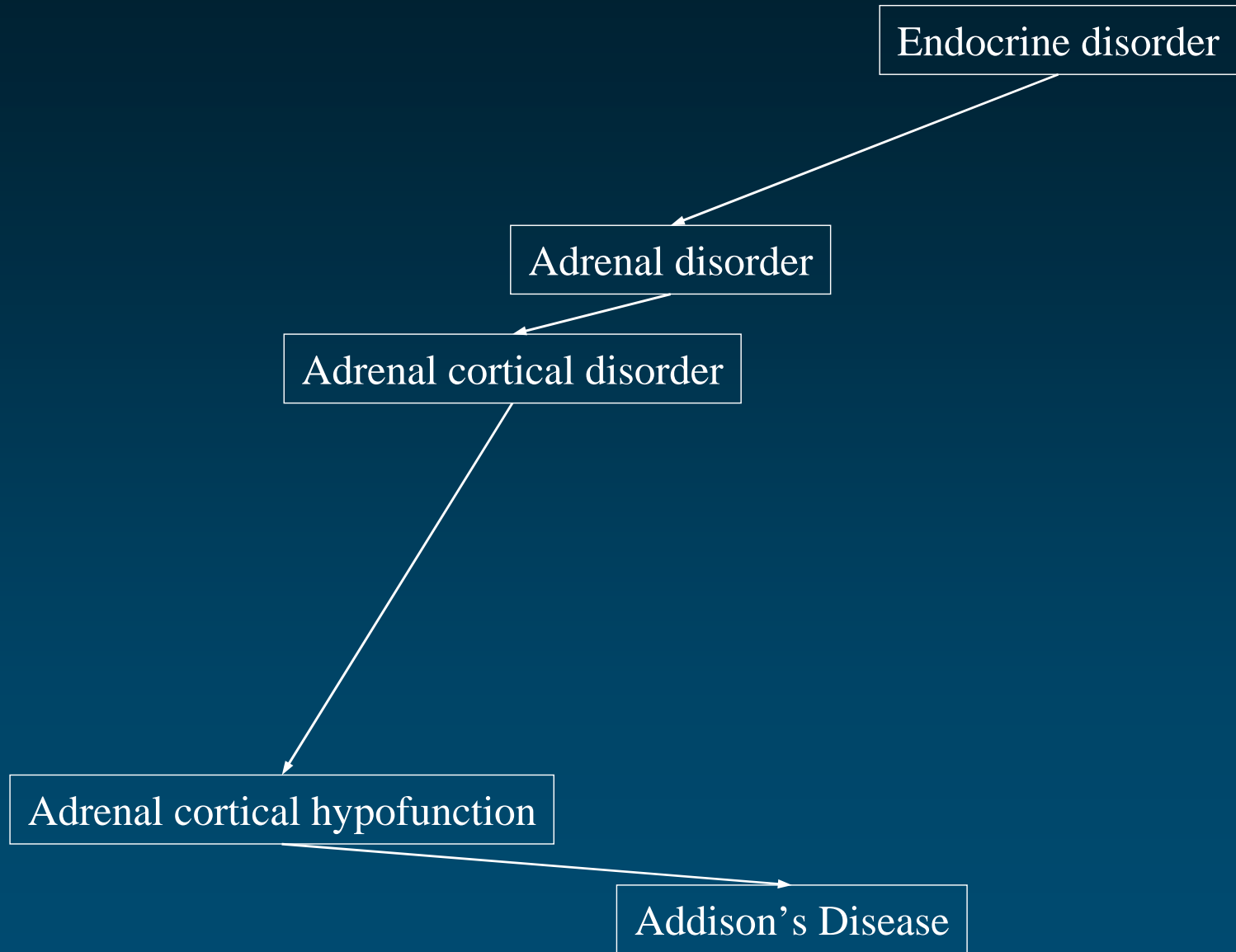


# SNOMED International

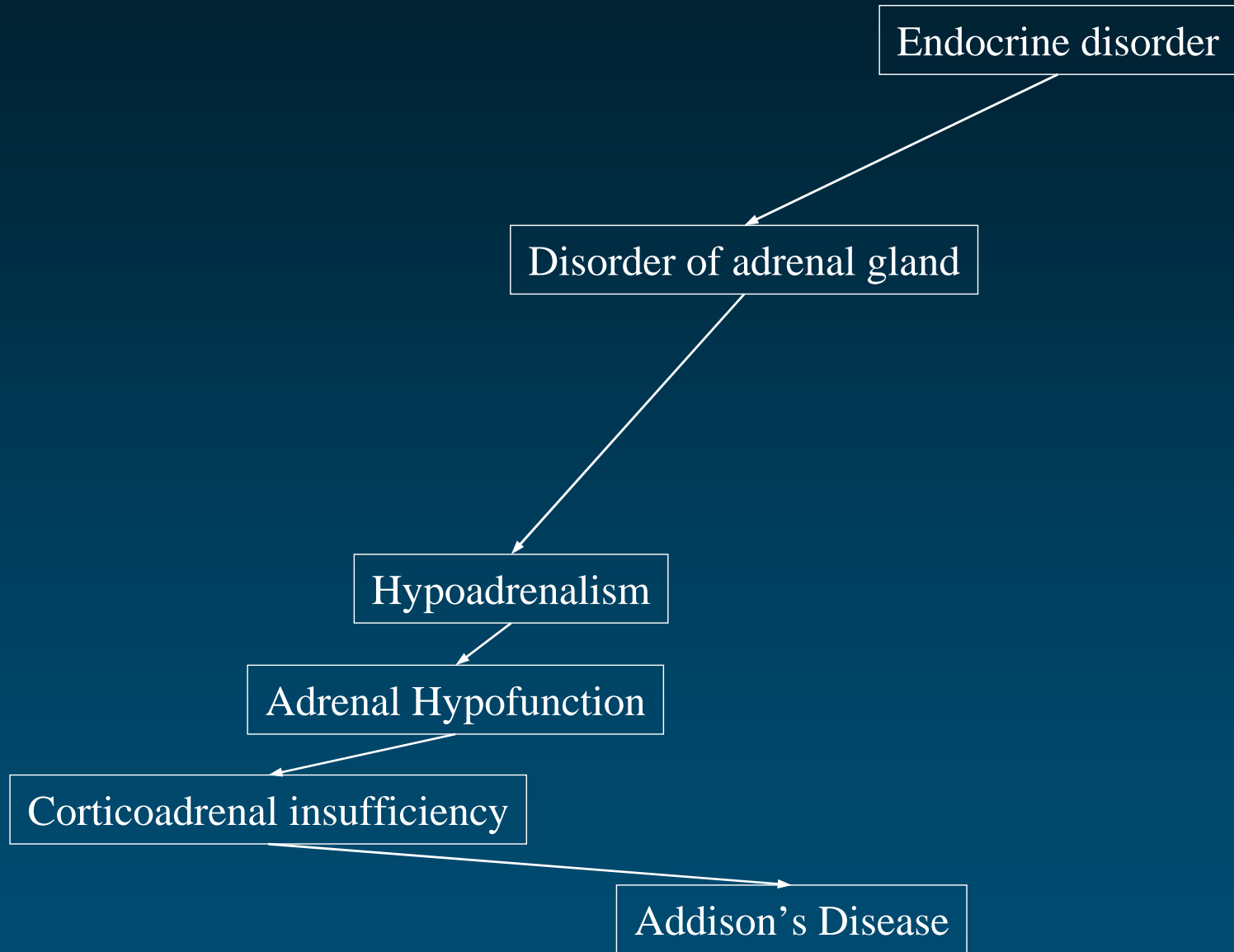




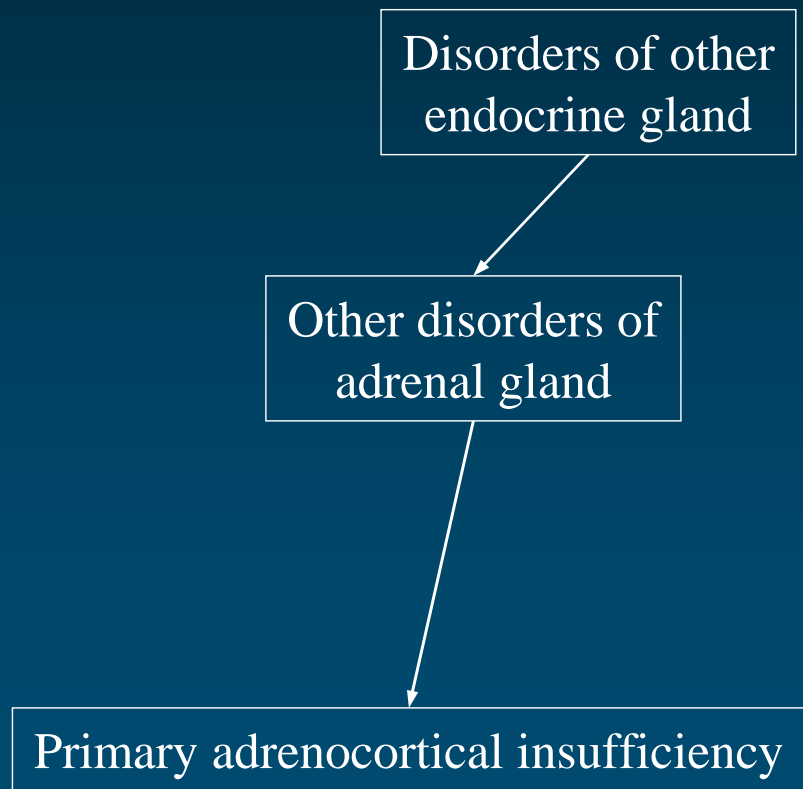
# AOD



## Read Codes

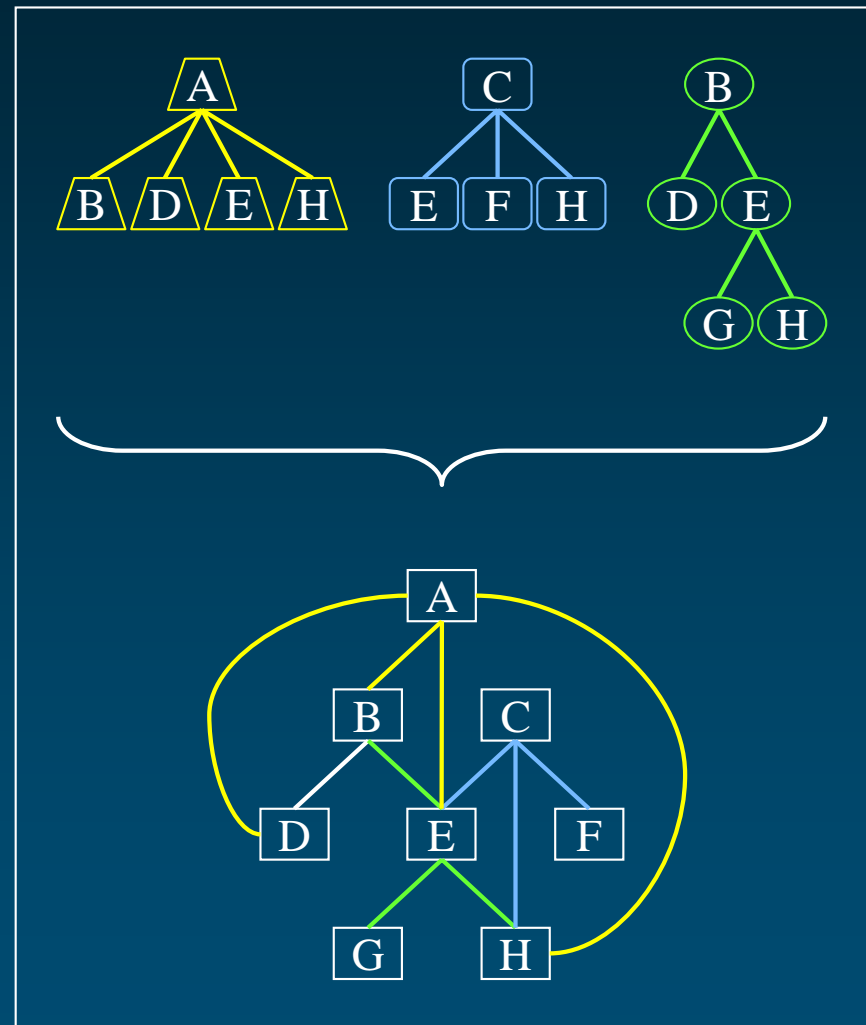


# ICD-10



# Organize concepts

- ◆ Inter-concept relationships: hierarchies from the source vocabularies
- ◆ Redundancy: multiple paths
- ◆ One graph instead of multiple trees (multiple inheritance)





*organize concepts*

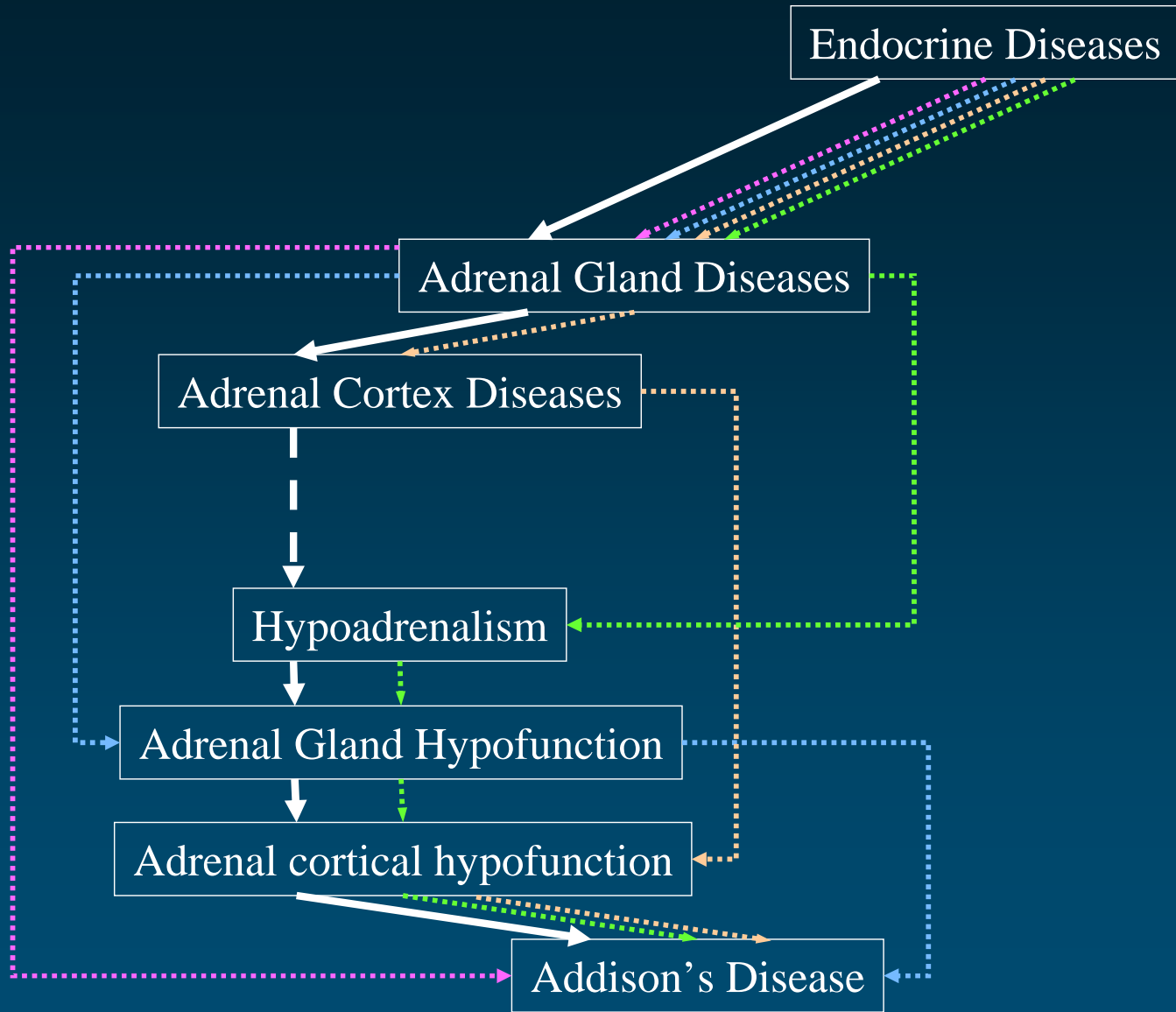
**SNOMED**

**MeSH**

**AOD**

**Read Codes**

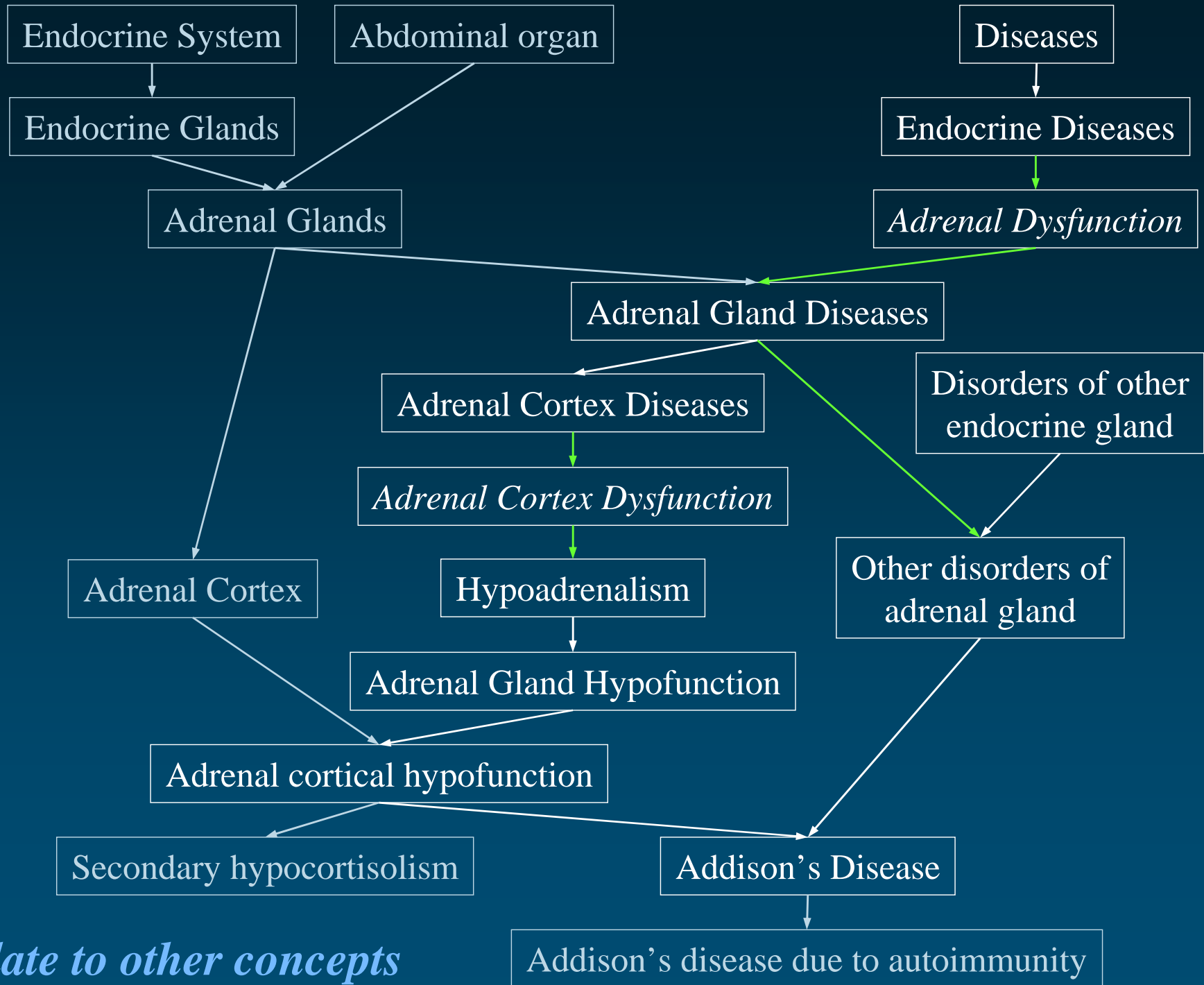
**UMLS**



# Relate to other concepts

- ◆ Additional hierarchical relationships
  - link to other trees
  - make relationships explicit
- ◆ Non-hierarchical relationships
- ◆ Co-occurring concepts
- ◆ Mapping relationships

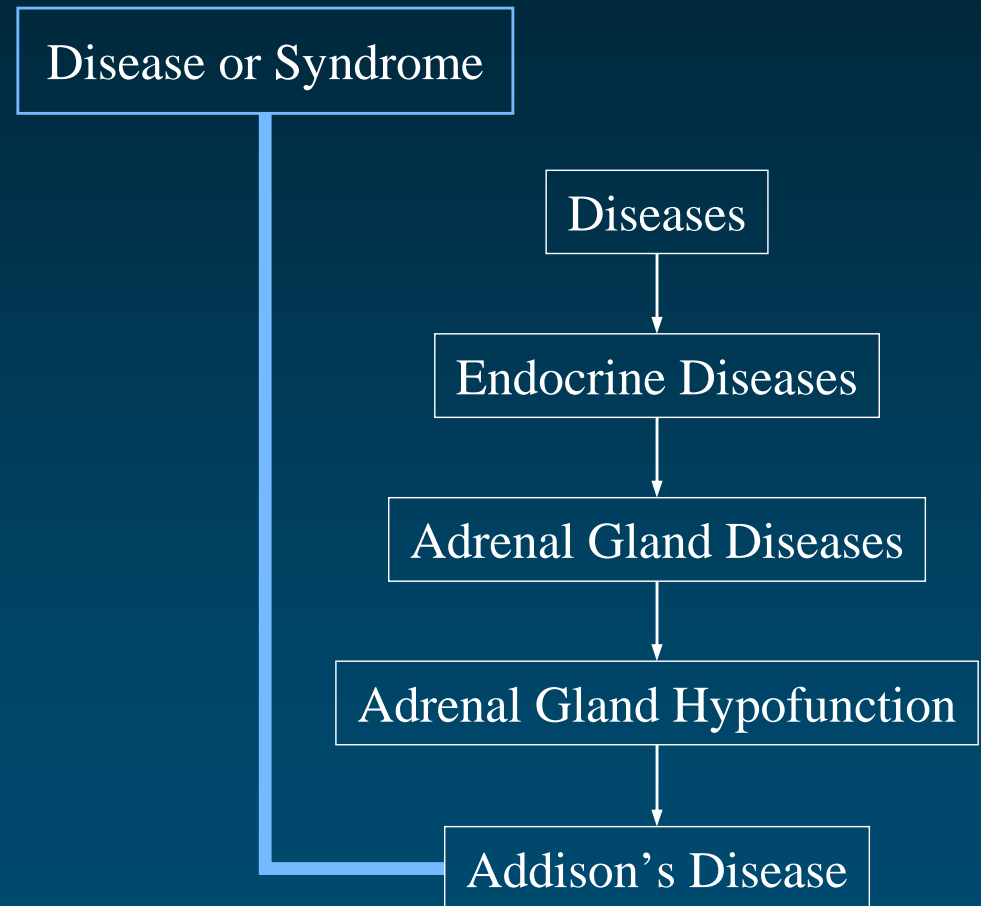




*relate to other concepts*

# Categorize concepts

- ◆ High-level categories (semantic types)
- ◆ Assigned by the Metathesaurus editors
- ◆ Independently of the hierarchies in which these concepts are located



# How do they do that?

- ◆ Lexical knowledge
- ◆ Semantic pre-processing
- ◆ UMLS editors



# Lexical knowledge

Adrenal gland diseases

Adrenal disorder

Disorder of adrenal gland

Diseases of the adrenal glands

C0001621

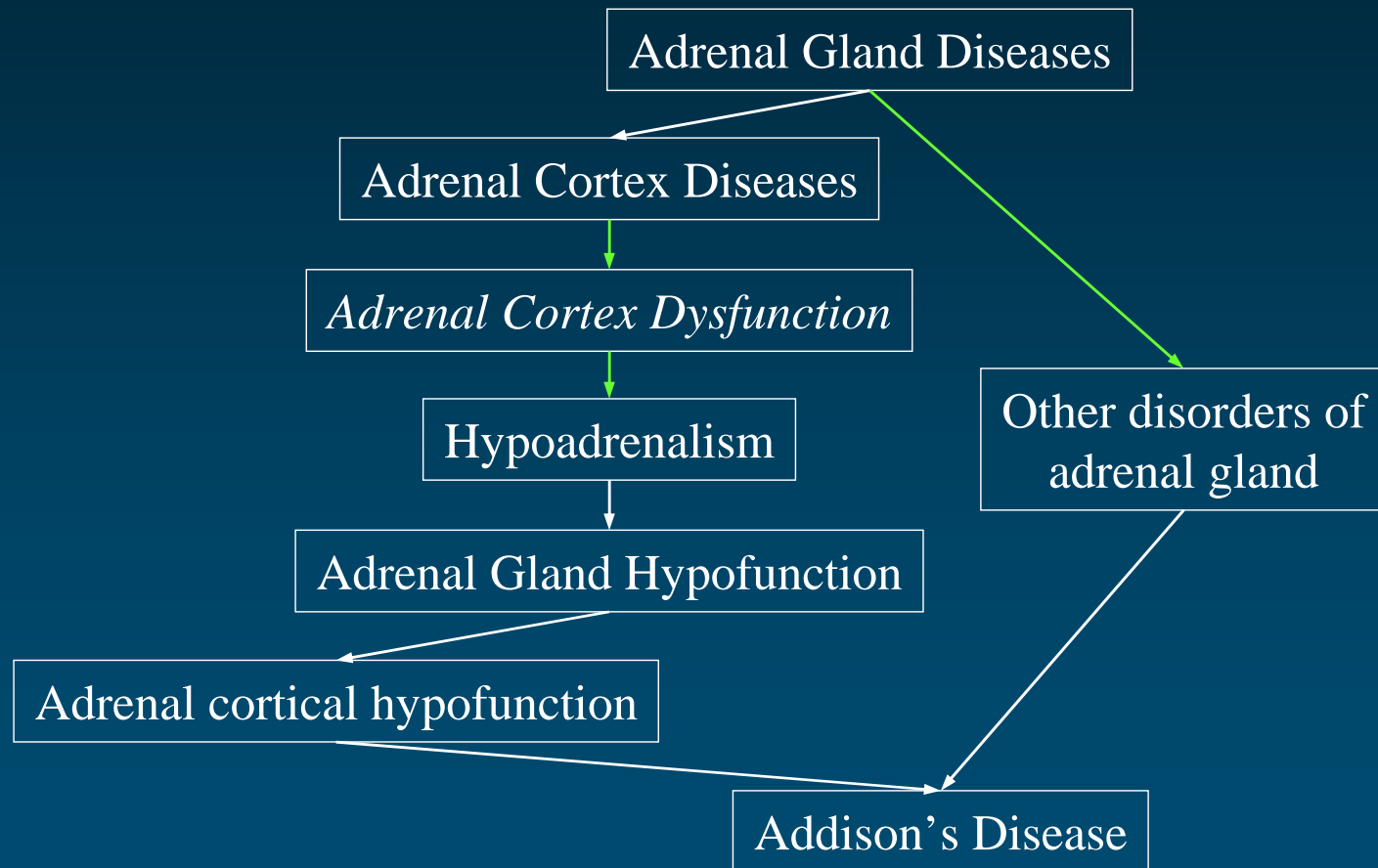


# Semantic pre-processing

- ◆ Metadata in the source vocabularies
- ◆ Tentative categorization
- ◆ Positive (or negative) evidence for tentative synonymy relations based on lexical features



# Additional knowledge: UMLS editors





# UMLS: 3 components



## ◆ SPECIALIST Lexicon

- 200,000 lexical items
- Part of speech and variant information

Lexical  
resources

## ◆ Metathesaurus

- 5M names from over 100 terminologies
- 1M concepts
- 16M relations

Terminological  
resources

## ◆ Semantic Network

- 135 high-level categories
- 7000 relations among them

Ontological  
resources



# UMLS Metathesaurus



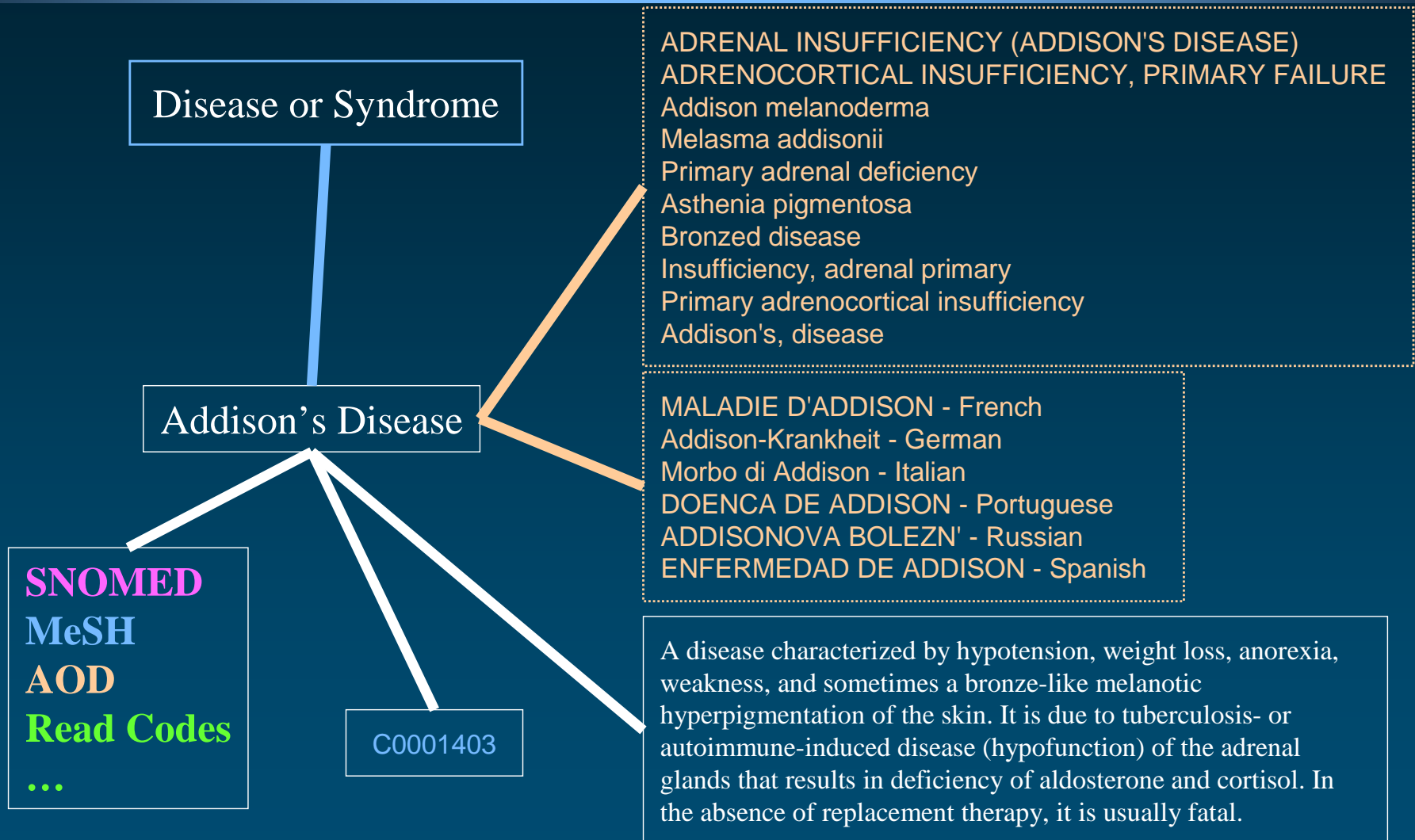
# Source Vocabularies

(2006AB)

- ◆ 139 source vocabularies
  - 17 languages
- ◆ Broad coverage of biomedicine
  - 5.1M names
  - 1.3M concepts
  - 16M relations
- ◆ Common presentation



# Addison's Disease: Concept



# Metathesaurus Concepts (2006AB)

- ◆ Concept (> 1.3M) CUI
  - Set of synonymous concept names
- ◆ Term (> 4.6M) LUI
  - Set of normalized names
- ◆ String (> 5.1M) SUI
  - Distinct concept name
- ◆ Atom (> 6.2M) AUI
  - Concept name in a given source

A0000001 headache (source 1)  
A0000002 headache (source 2)  
**S0000001**

A0000003 Headache (source 1)  
A0000004 Headache (source 2)  
**S0000002**

**L0000001**

A0000005 Cephalgia (source 1)  
**S0000003**

**L0000002**

**C0000001**



# Cluster of synonymous terms

Concept  
C0001621

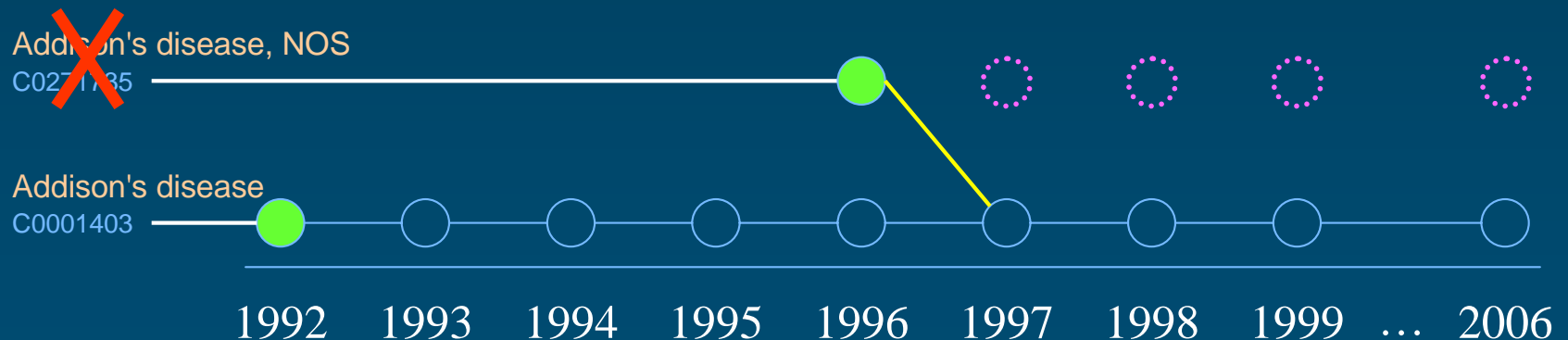
Term L0001621	<p>S0011232 <i>Adrenal Gland Diseases</i></p> <p>S0011231 Adrenal Gland Disease</p> <p>S0000441 Disease of adrenal gland</p> <p>S0481705 Disease of adrenal gland, NOS</p> <p>S0220090 Disease, adrenal gland</p> <p>S0044801 Gland Disease, Adrenal</p>	[...]
Term L0041793	<p>S0860744 <i>Disorder of adrenal gland, unspecified</i></p> <p>S0217833 Unspecified disorder of adrenal glands</p>	
Term L0161347	<p>S0225481 <i>ADRENAL DISORDER</i></p> <p>S0627685 DISORDER ADRENAL (NOS)</p>	[...]
Term L0181041	<p>S0632950 <i>Disorder of adrenal gland</i></p> <p>S0354509 Adrenal Gland Disorders</p>	[...]
Term L0368399	<p>S0586222 <i>Adrenal disease</i></p> <p>S0466921 ADRENAL DISEASE, NOS</p>	[...]
Term L1279026	<p>S1520972 <i>Nebennierenkrankheiten</i></p>	GER
Term L0162317	<p>S0226798 <i>SURRENALE, MALADIES</i></p>	FRE

[...]



# Metathesaurus Evolution over time

- ◆ Concepts never die (in principle)
  - CUIs are permanent identifiers
- ◆ What happens when they do die (in reality)?
  - Concepts can merge or split
  - Resulting in new concepts and deletions



# Metathesaurus Relationships

- ◆ Symbolic relations: ~9 M pairs of concepts
  - ◆ Statistical relations : ~7 M pairs of concepts  
(co-occurring concepts)
  - ◆ Mapping relations: 100,000 pairs of concepts
- 

- ◆ Categorization: Relationships between concepts and semantic types from the Semantic Network





# Symbolic relations

## ◆ Relation

- Pair of “atom” identifiers
- Type
- Attribute (if any)
- List of sources (for type and attribute)

## ◆ Semantics of the relationship: defined by its *type* [and *attribute*]

Source transparency: the information  
is recorded at the “atom” level



# Symbolic relationships Type

## ◆ Hierarchical

- Parent / Child
- Broader / Narrower than

PAR/CHD

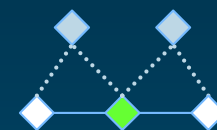
RB/RN



## ◆ Derived from hierarchies

- Siblings (children of parents)

SIB



## ◆ Associative

- Other

RO



## ◆ Various flavors of near-synonymy

- Similar
- Source asserted synonymy
- Possible synonymy

RL

SY

RQ



# Symbolic relationships Attribute

- ◆ Hierarchical
  - isa (is-a-kind-of)
  - part-of
- ◆ Associative
  - location-of
  - caused-by
  - treats
  - ...
- ◆ Cross-references (mapping)



Semantic Types

Anatomical Structure

Fully Formed Anatomical Structure

Embryonic Structure

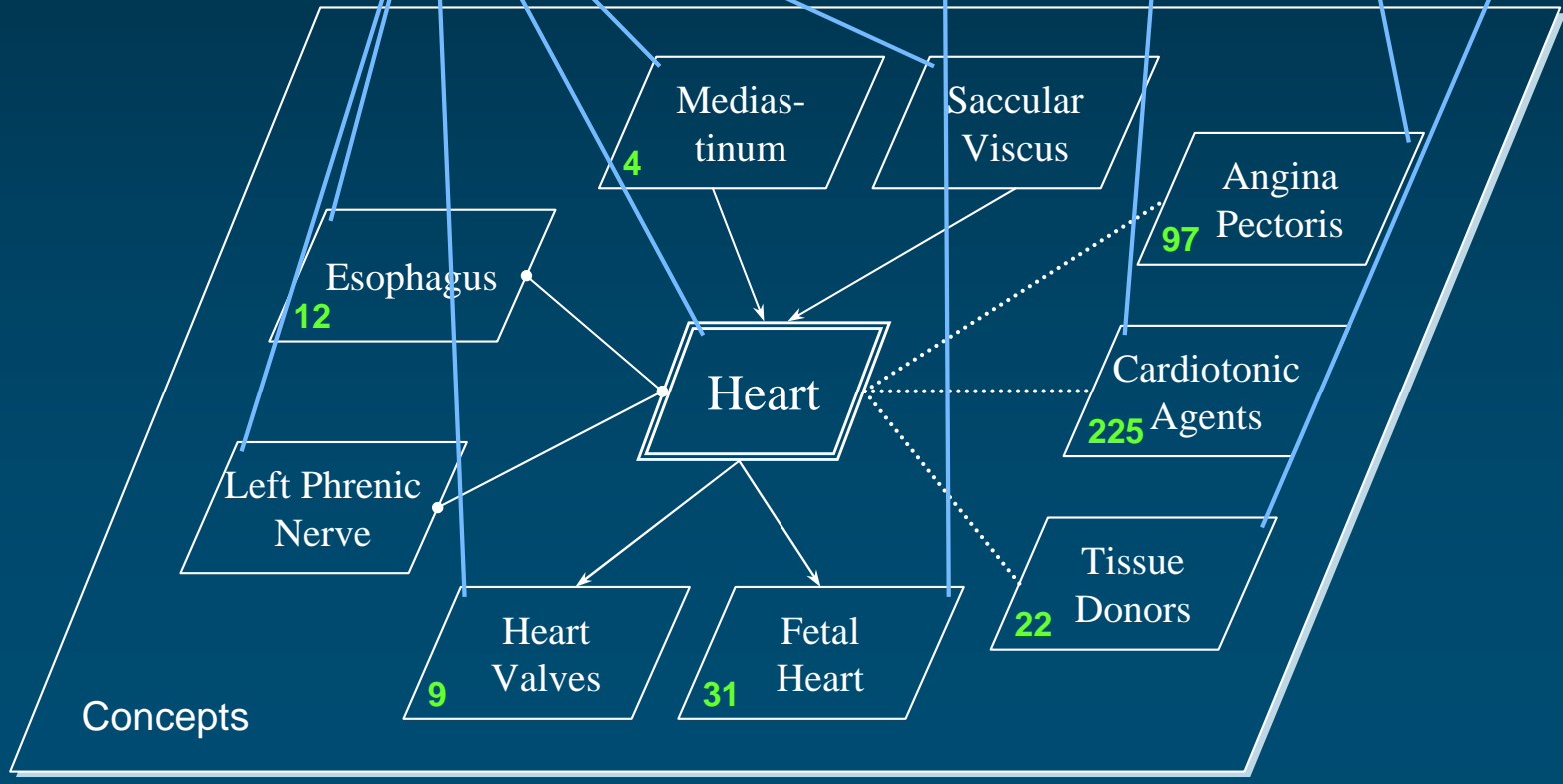
Disease or Syndrome

Body Part, Organ or Organ Component

Pharmacologic Substance

Population Group

Semantic Network



Metathesaurus

Concepts

# UMLS Semantic Network

# Semantic Network

## ◆ Semantic types (135)

- tree structure
- 2 major hierarchies
  - Entity
    - Physical Object
    - Conceptual Entity
  - Event
    - Activity
    - Phenomenon or Process

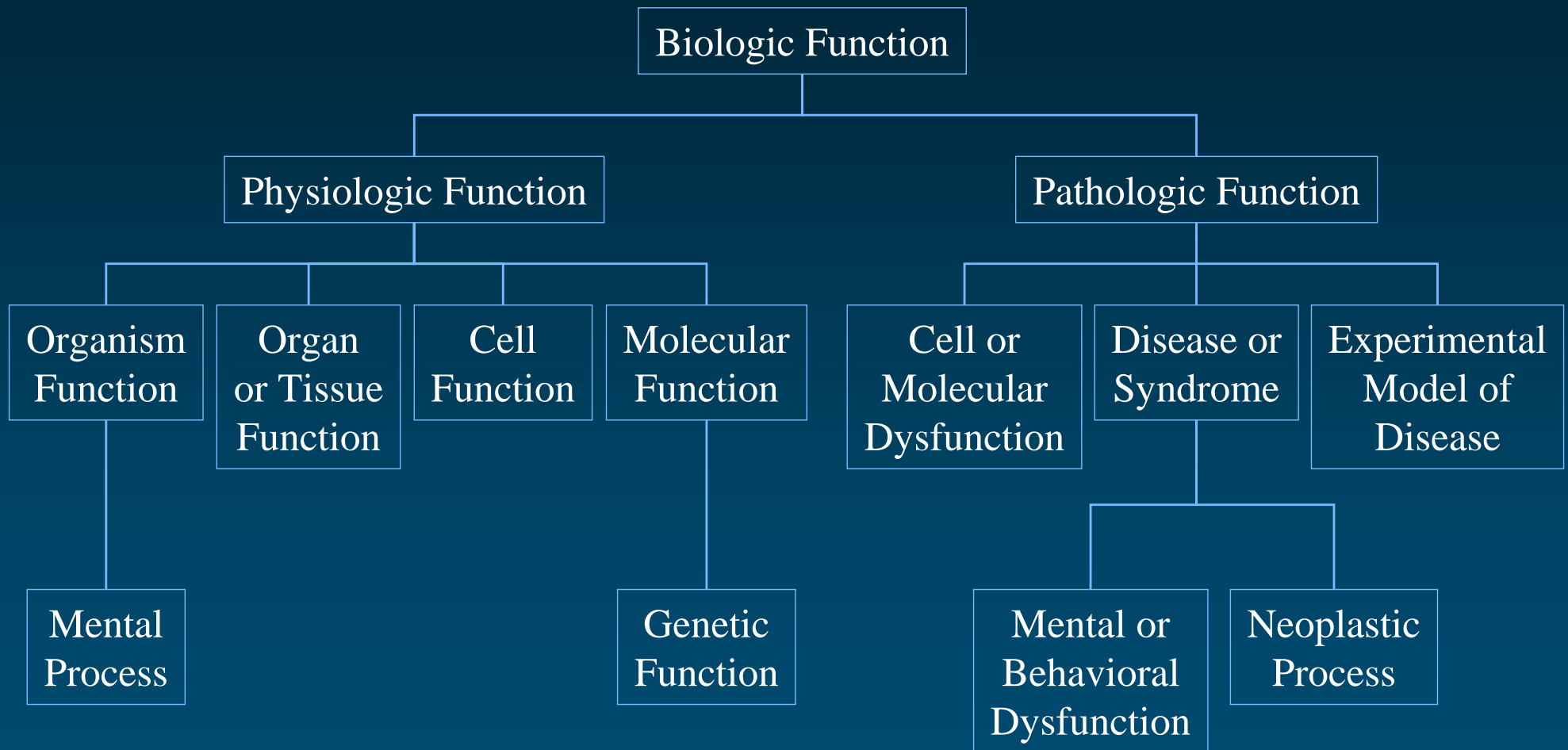


# Semantic Network

- ◆ Semantic network relationships (54)
  - hierarchical (isa = is a kind of)
    - among types
      - *Animal isa Organism*
      - *Enzyme isa Biologically Active Substance*
    - among relations
      - *treats isa affects*
  - non-hierarchical
    - *Sign or Symptom diagnoses Pathologic Function*
    - *Pharmacologic Substance treats Pathologic Function*

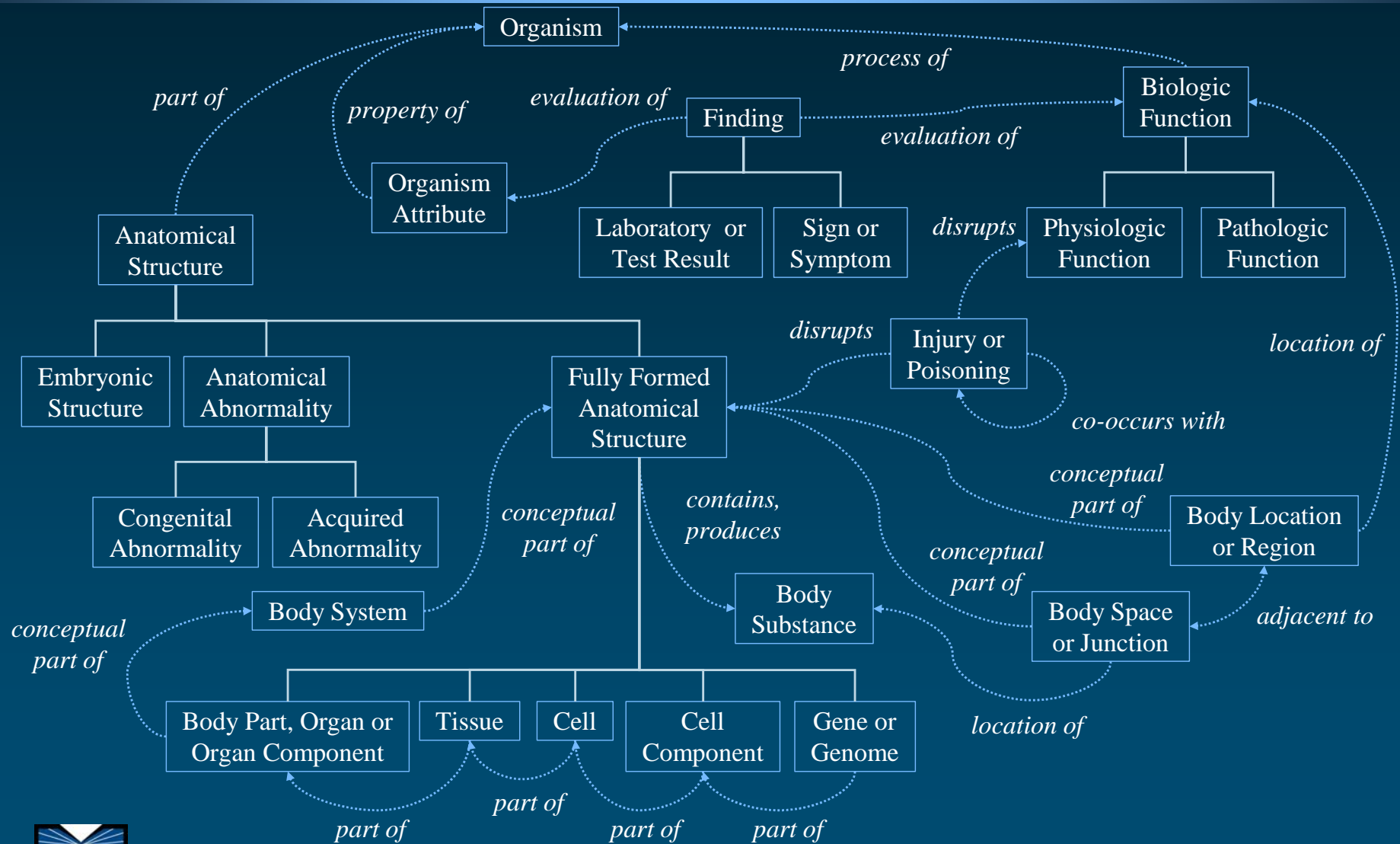


# “Biologic Function” hierarchy (isa)





# Associative (non-isa) relationships

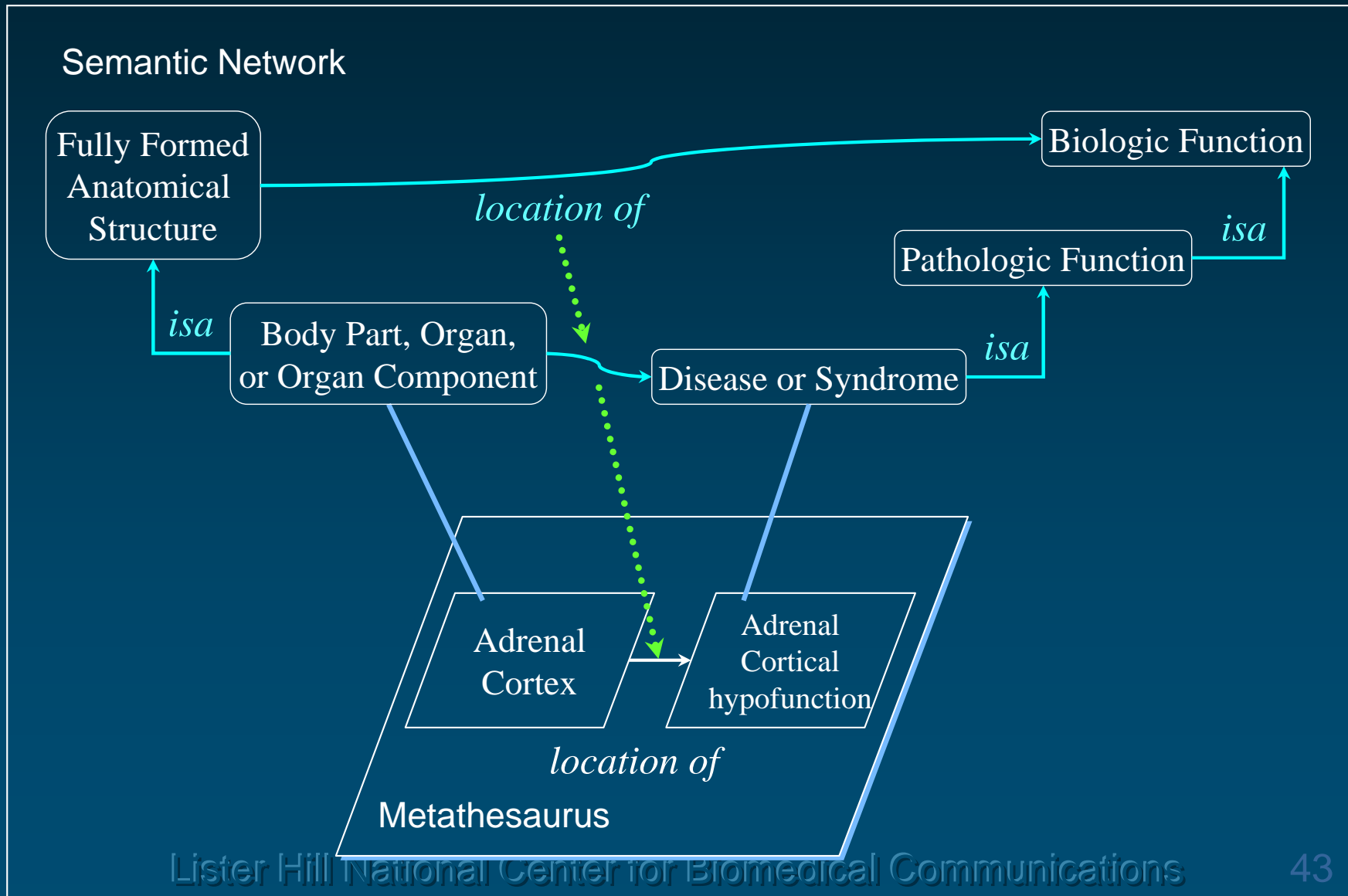


# Why a semantic network?

- ◆ Semantic Types serve as high level categories assigned to Metathesaurus concepts, *independently of their position in a hierarchy*
- ◆ A relationship between 2 Semantic Types (ST) is a possible link between 2 concepts that have been assigned to those STs
  - The relationship may or may not hold at the concept level
  - Other relationships may apply at the concept level



# Relationships can inherit semantics

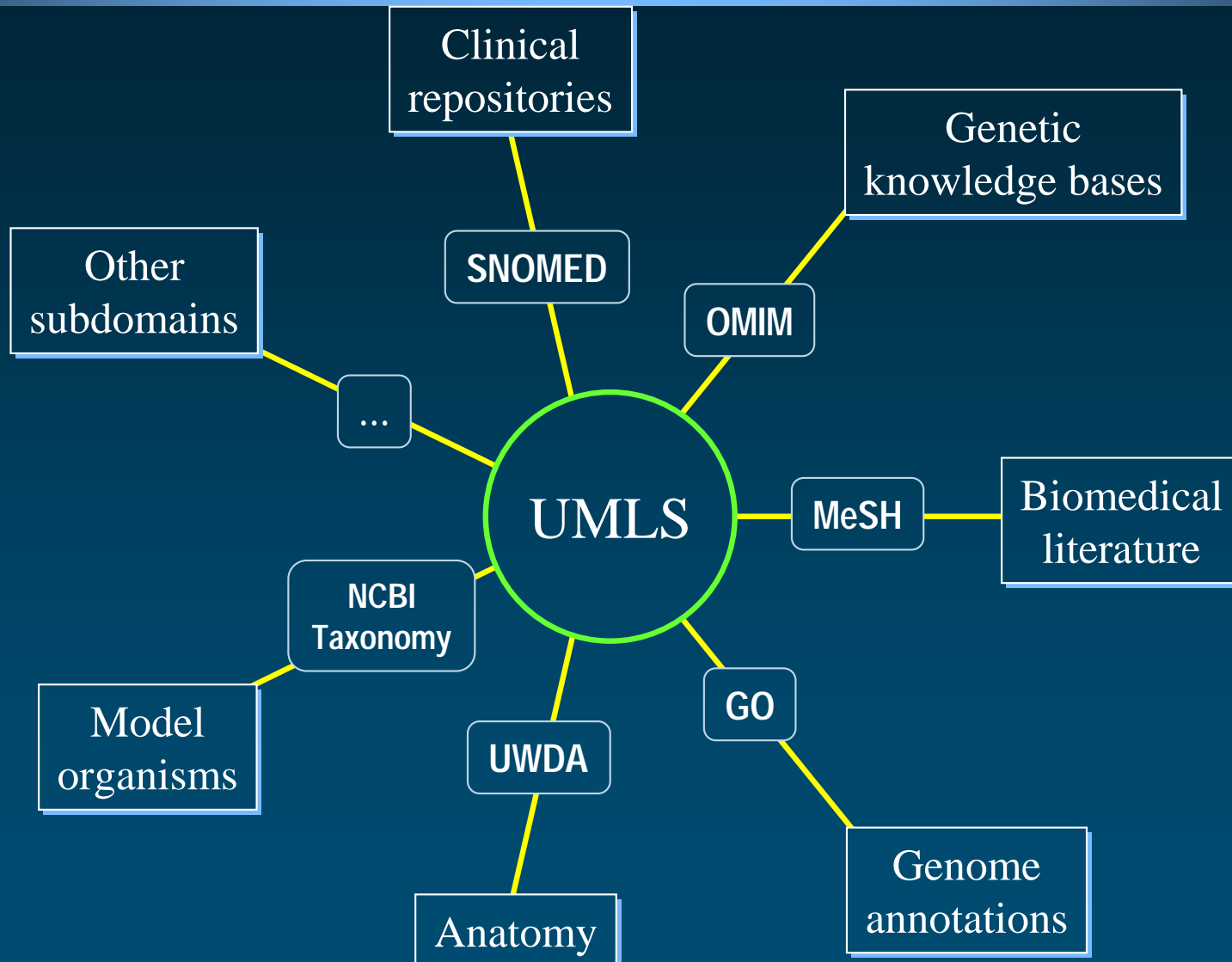


# UMLS Summary

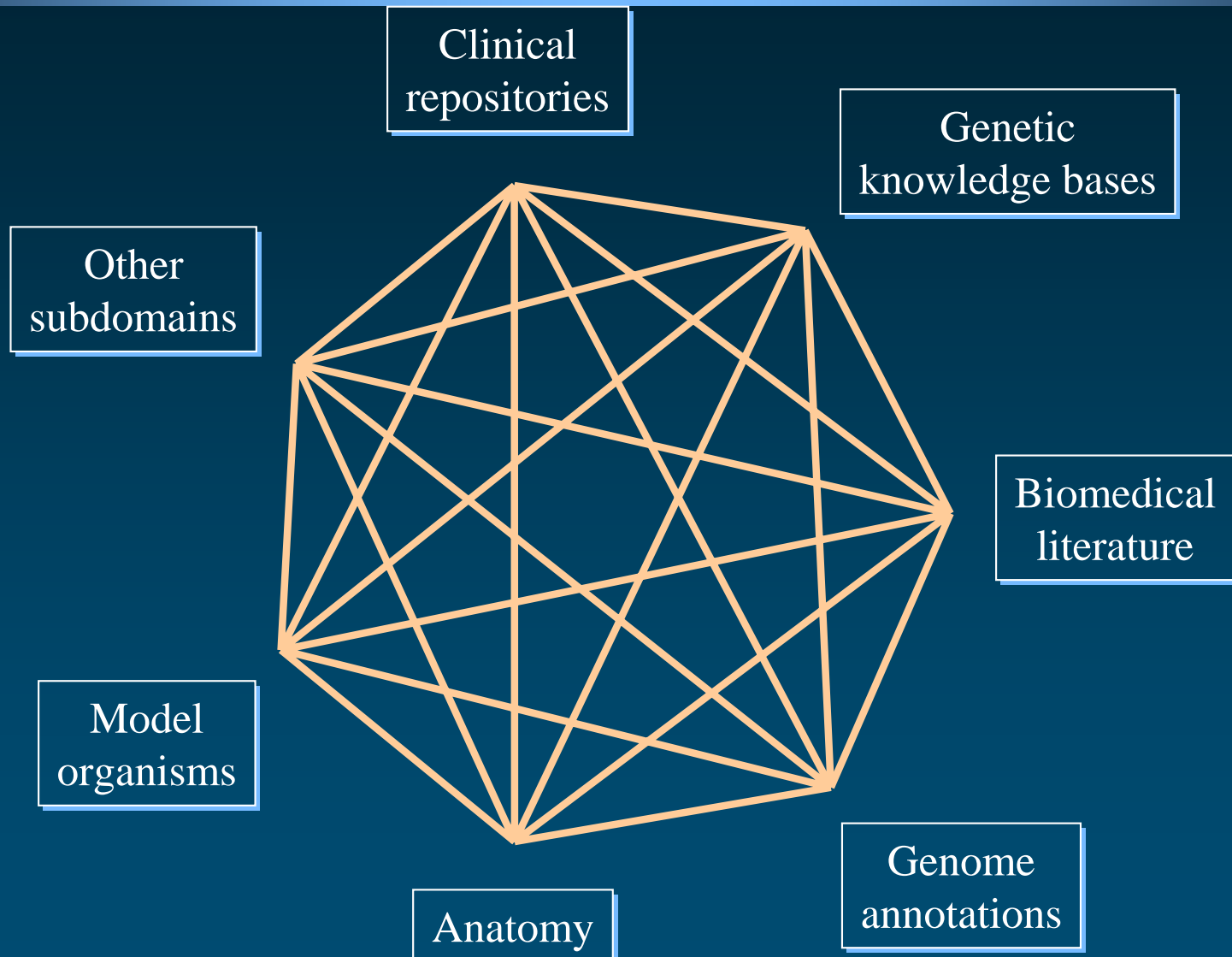
- ◆ Synonymous terms clustered into concepts
- ◆ Unique identifier
- ◆ Finer granularity
- ◆ Broader scope
- ◆ Additional hierarchical relationships
- ◆ Semantic categorization



# Integrating subdomains



# Integrating subdomains



# Information integration

*Genomics as an example*

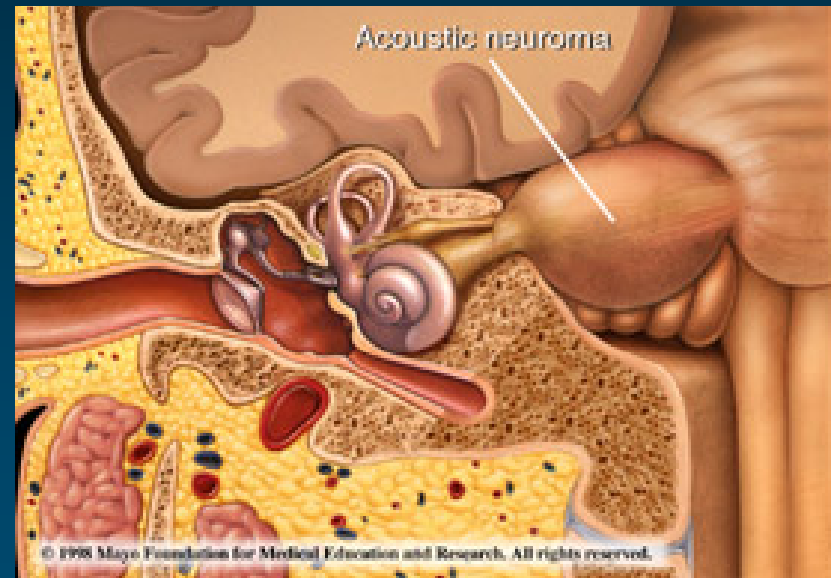
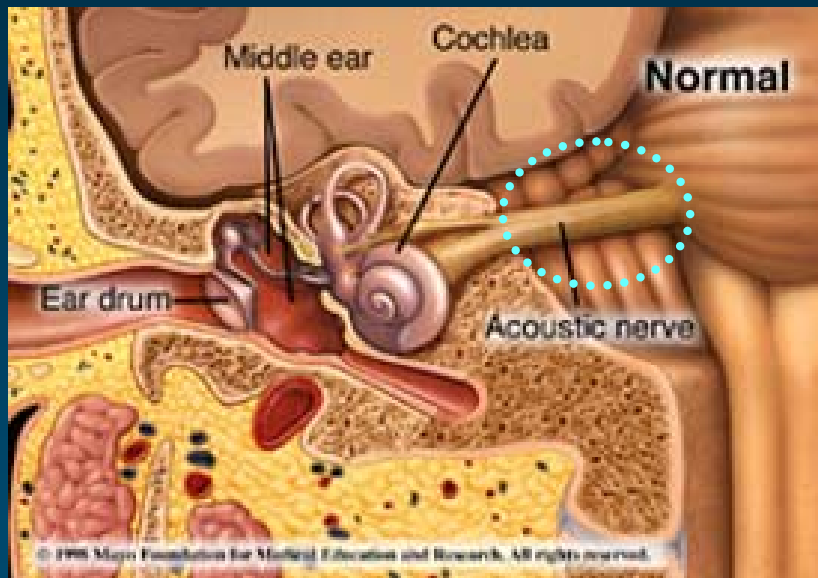
# NF2 Gene, protein, and disease

*Neurofibromatosis 2* is an autosomal dominant disease characterized by tumors called schwannomas involving the acoustic nerve, as well as other features. The disorder is caused by mutations of the *NF2 gene* resulting in absence or inactivation of the protein product. The protein product of NF2 is commonly called *merlin* (but also neurofibromin 2 and schwannomin) and functions as a tumor suppressor.





# Schwannoma (acoustic neuroma)



<http://www.mayoclinic.com>

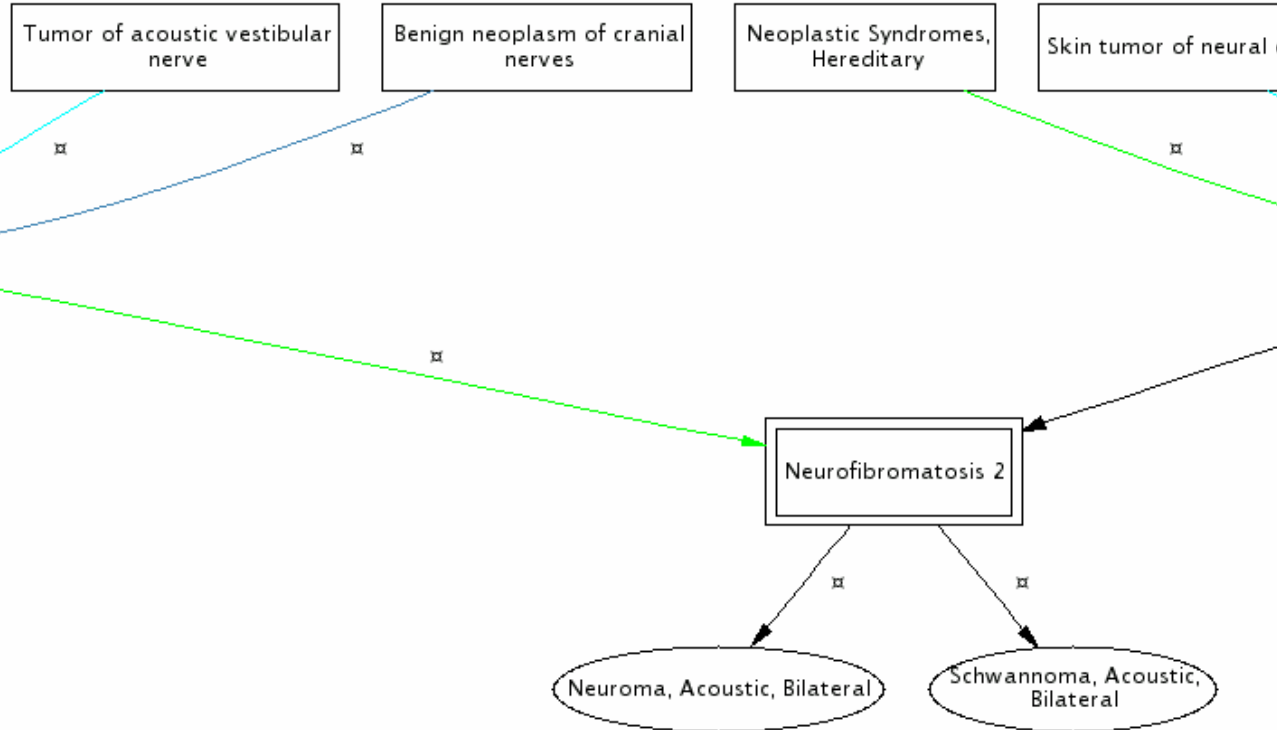
### Siblings

#### Disorders

- Cerebellopontine Angle Acoustic Neuroma
- Diffuse neurofibroma
- Melanocytic Vestibular Schwannoma
- Neurofibromatosis (nonmalignant)
- Neurofibromatosis 1
- neurofibromatosis 1 and 2 (NF1 and NF2)
- Neurofibromatosis 3
- Neurofibromatosis type 3
- NEUROFIBROMATOSIS TYPE IV, OF RICCARDI
- Neuroma, Acoustic, Unilateral
- Segmental neurofibromatosis

(11 siblings)

[direct children and narrower concepts of direct parents and broader concepts]



### Other Related Concepts

#### Anatomy

- Acoustic Nerve

#### Chemicals & Drugs

- Neurofibromin 2

#### Disorders

- Familial Acoustic Neuromas
- Neoplasm of uncertain behavior NOS
- Neurofibromatosis
- Neurofibromatosis

- Nerve Sheath Tumors [4]
- Nervous System Neoplasms [6]
- Neurilemmoma [35]
- Neurofibromatosis 1 [38]
- Neuroma, Acoustic [26]
- Peripheral Nervous System Diseases [3]
- Peripheral Nervous System Neoplasms [6]
- Postoperative Complications [9]
- Retinal Diseases [6]
- Skin Neoplasms [9]

**BCI** **Neurofibromatosis 2** **LEGEND \***

Start again Apply new parameters

Restrict to vocabulary: Show all

Highlight vocabulary: Nothing

UMLS data: UMLS\_2003

Type of hierarchical rel.:  All  Parent/Child only  Broader/Narrower only

**Similar Concepts**  
(none)

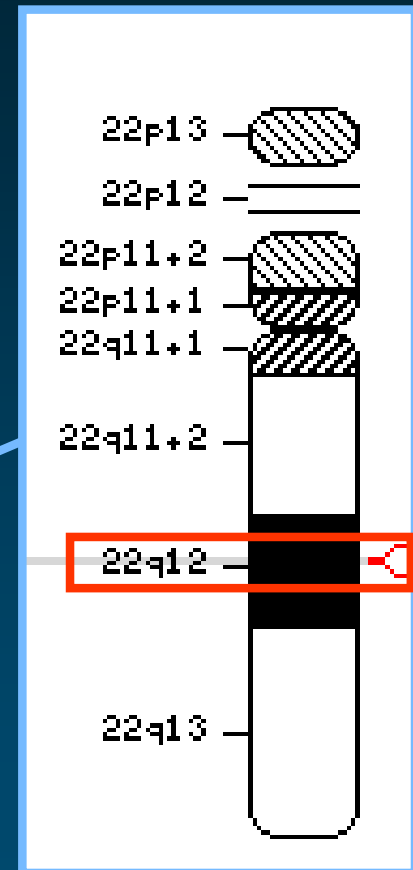
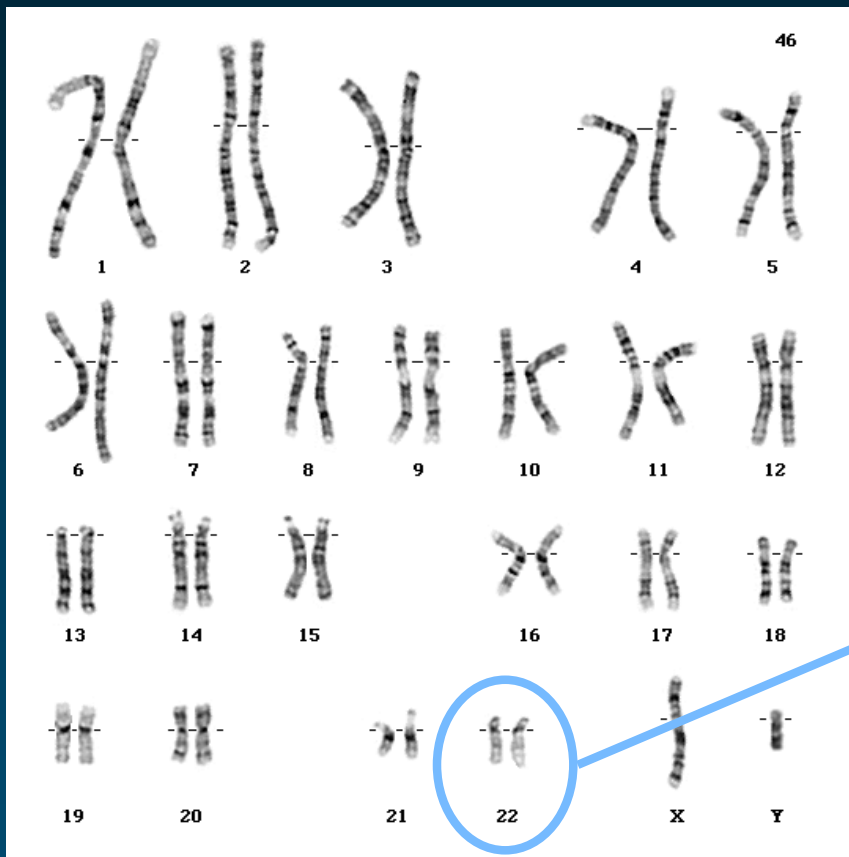
**Allegedly Synonyms**  
• Neurofibromatosis (nonmalignant)

**Closest MeSH Terms**

**Main Headings**  
• Neurofibromatosis 2

**Subheadings**

# NF2 gene



<http://staff.washington.edu/timk/cyto/human/>

<http://www.ncbi.nlm.nih.gov/mapview/>



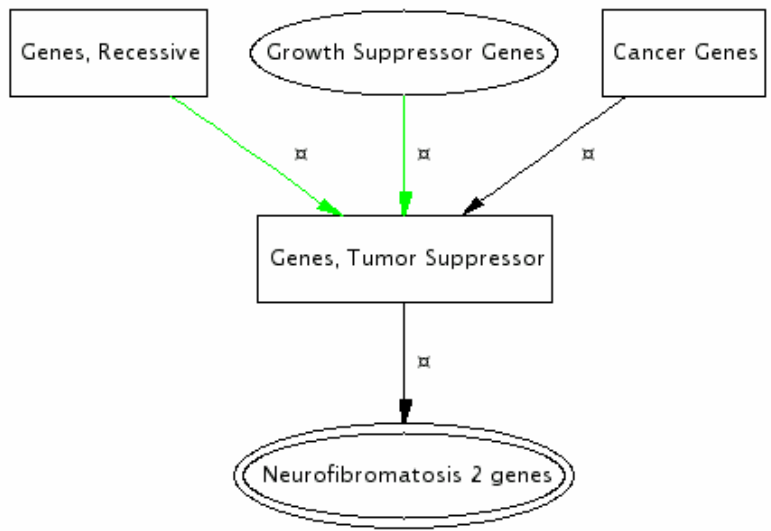
Siblings

Chemicals & Drugs

- ADAM11 protein, human
- DLG5 protein, human
- DPM3 protein, human
- HCCS-1protein, human
- hssh3bp1 protein, human
- HUGL protein, human
- LAPSER1 protein, human
- mitochondria proteolipid-like protein, human
- MRG protein, human
- p53 gene/protein
- PLAGL1 protein, human
- RARRES3 protein, human
- SEZ6L protein, human
- TES protein, human

Genes & Molecular Sequences

- APC Gene
- BAX Gene
- brca gene
- CDH1 gene
- CHES1 Gene
- cyclin-dependent kinase inhibitor 2A



Other Related Concepts

Chemicals & Drugs

- Neurofibromin 2

Disorders

- Neurofibromatosis 2

(2 other related concepts)

- Chromosome Deletion [7]
- Ependymoma [4]
- Glioma [4]
- Loss of Heterozygosity [7]
- Meningeal Neoplasms [25]
- Meningioma [30]
- mesothelioma
- <1> [4]
- Neoplasms [4]
- Neurilemmoma [20]
- Neurofibromatosis
- Neurofibromatosis 2 [64]
- Neuroroma, Acoustic [5]
- Spinal Cord Neoplasms [3]

**BCI** Neurofibromatosis 2 genes **LEGEND \***

Start again Apply new parameters

Restrict to vocabulary: Show all

Highlight vocabulary: Nothing

UMLS data: UMLS\_2003

Type of hierarchical rel.:  All  Parent/Child only  Broader/Narrower only

**Similar Concepts** (none)

**Allegedly Synonyms** (none)

**Closest MeSH Terms**

**Main Headings**

- Genes, Neurofibromatosis 2

**Subheadings**

# Merlin

## ◆ Synonyms

- Neurofibromin 2
- Schwannomin
- Schwannomerlin
- Neurofibromatosis-2

## ◆ 10 isoforms

## ◆ Annotations

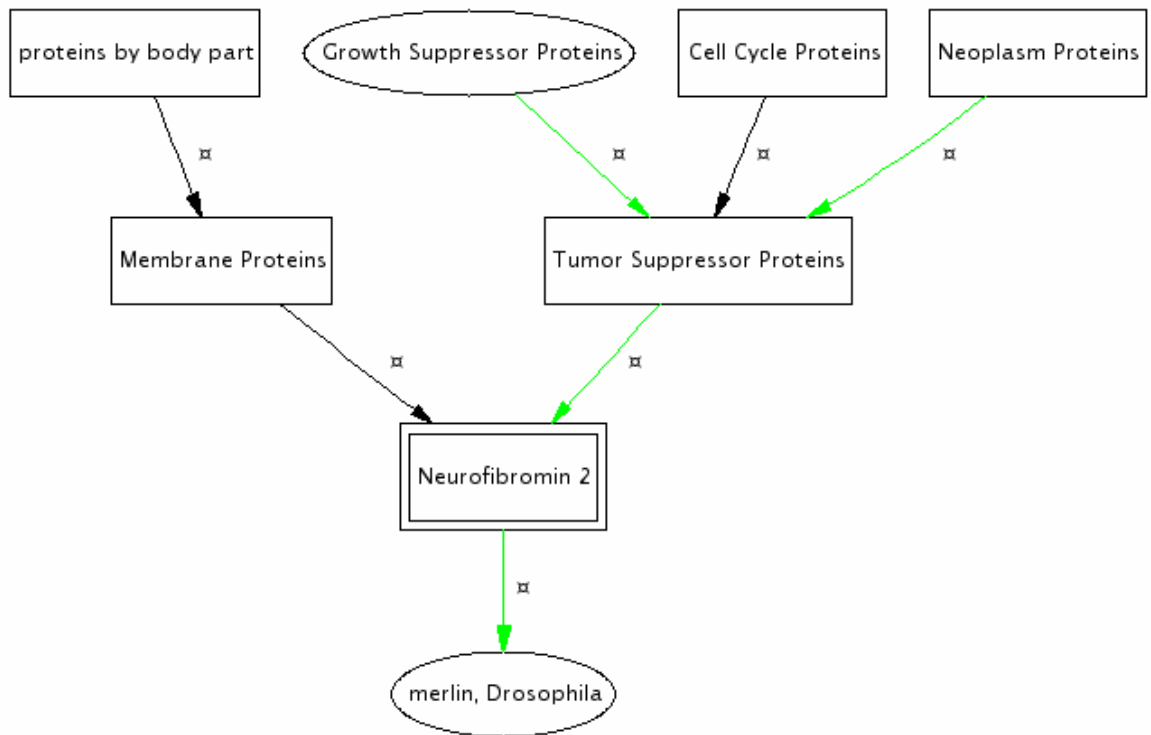
- Negative regulation of cell proliferation
- Cytoskeleton
- Plasma membrane



### Siblings

#### Chemicals & Drugs

- (LA)12 peptide
- (methyl)ammonium uptake carrier, Corynebacterium
- 120-kDa hemocyte-specific membrane protein, flesh fly
- 15a protein, Aedes aegypti
- 22.6-kDa antigen, Schistosoma japonicum
- 36-kDa vesicular integral membrane protein
- 38L protein
- 5-lipoxygenase-activating protein
- 59 kDa dystrophin-associated protein
- A-1 antigen
- A-kinase anchor protein 149
- A-kinase anchor protein 15
- A-kinase anchor protein 200
- A-kinase anchor protein KL
- A14.5L protein
- A15 protein
- ABC-me protein
- ABU-1 protein, C elegans
- AcfB protein
- ACR3 protein



### Other Related Concepts

#### Disorders

- Neurofibromatosis 2

#### Genes & Molecular Sequences

- Neurofibromatosis 2 genes

(2 other related concepts)

### Co-occurring Concepts

#### Anatomy

- Arachnoid [1]
- Cell Membrane [1]
- Cerebellum [1]
- Chromosomes, Human, Pair 22 [1]
- Cytoplasm [1]
- Cytoskeleton [2]
- Microfilaments [1]
- Purkinje Cells [1]
- Schwann Cells [1]
- Stem Cells [1]

BCI

### Neurofibromin 2

LEGEND \*

Start again Apply new parameters

Restrict to vocabulary: Show all

Highlight vocabulary: Nothing

UMLS data: UMLS\_2003

Type of hierarchical rel.: All Parent/Child only Broader/Narrower only

#### Similar Concepts

(none)

#### Allegedly Synonyms

(none)

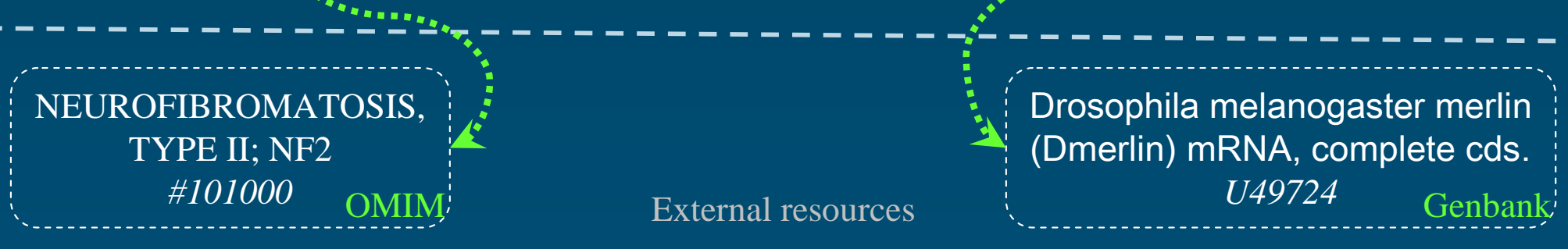
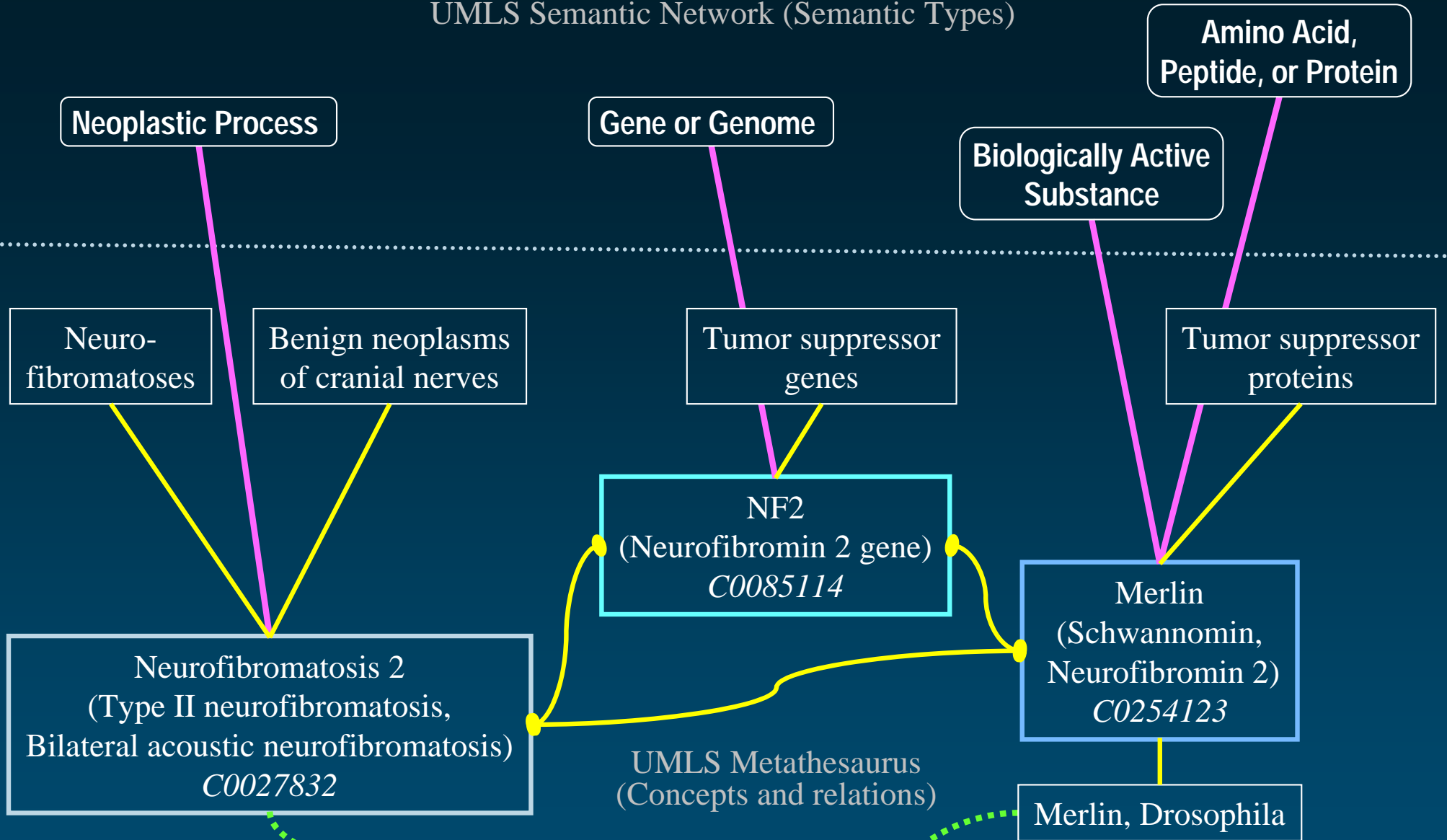
#### Closest MeSH Terms

#### Main Headings

- Neurofibromin 2

#### Subheadings

UMLS Semantic Network (Semantic Types)



# Limitations

- ◆ Genes not systematically represented
  - Most gene products and diseases are
- ◆ Gene/Gene product-Disease relations
  - Not systematically represented
  - Not explicitly represented (e.g., co-occurrence)
- ◆ Cross-references not systematically represented
- ◆ Naming conventions (genes)



# References

- ◆ UMLS

[umlsinfo.nlm.nih.gov](http://umlsinfo.nlm.nih.gov)

- ◆ UMLS browsers

(free, but UMLS license required)

- Knowledge Source Server: [umlsks.nlm.nih.gov](http://umlsks.nlm.nih.gov)

- Semantic Navigator:

<http://mor.nlm.nih.gov/perl/semnav.pl>

- RRF browser

(standalone application distributed with the UMLS)



# References

## ◆ Recent overviews

- Bodenreider O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*; D267-D270.
- Nelson, S. J., Powell, T. & Humphreys, B. L. (2002 ). The Unified Medical Language System (UMLS) Project. In: Kent, Allen; Hall, Carolyn M., editors. *Encyclopedia of Library and Information Science*. New York: Marcel Dekker. p.369-378.



# References

## ◆ UMLS as a research project

- Lindberg, D. A., Humphreys, B. L., & McCray, A. T. (1993). The Unified Medical Language System. *Methods Inf Med, 32*(4), 281-91.
- Humphreys, B. L., Lindberg, D. A., Schoolman, H. M., & Barnett, G. O. (1998). The Unified Medical Language System: an informatics research collaboration. *J Am Med Inform Assoc, 5*(1), 1-11.

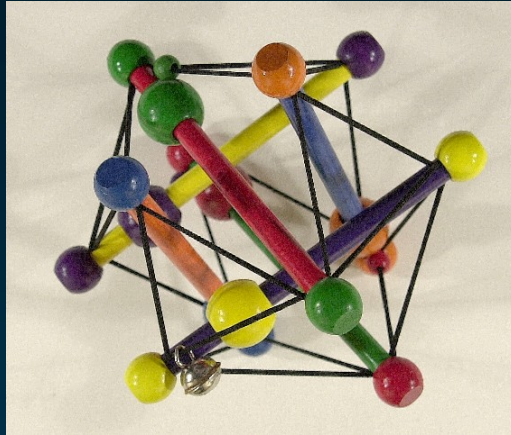


# References

## ◆ Technical papers

- McCray, A. T., & Nelson, S. J. (1995). The representation of meaning in the UMLS. *Methods Inf Med*, 34(1-2), 193-201.
- Bodenreider O. & McCray A. T. (2003). Exploring semantic groups through visual approaches. *Journal of Biomedical Informatics*, 36(6), 414-432.





# Medical Ontology Research

Contact: [olivier@nlm.nih.gov](mailto:olivier@nlm.nih.gov)

Web: [mor.nlm.nih.gov](http://mor.nlm.nih.gov)



*Olivier Bodenreider*

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