



Data Integration in the Life Sciences
Evry, France
June 26, 2008

Ontologies and data integration in biomedicine

Success stories and challenging issues



Olivier Bodenreider

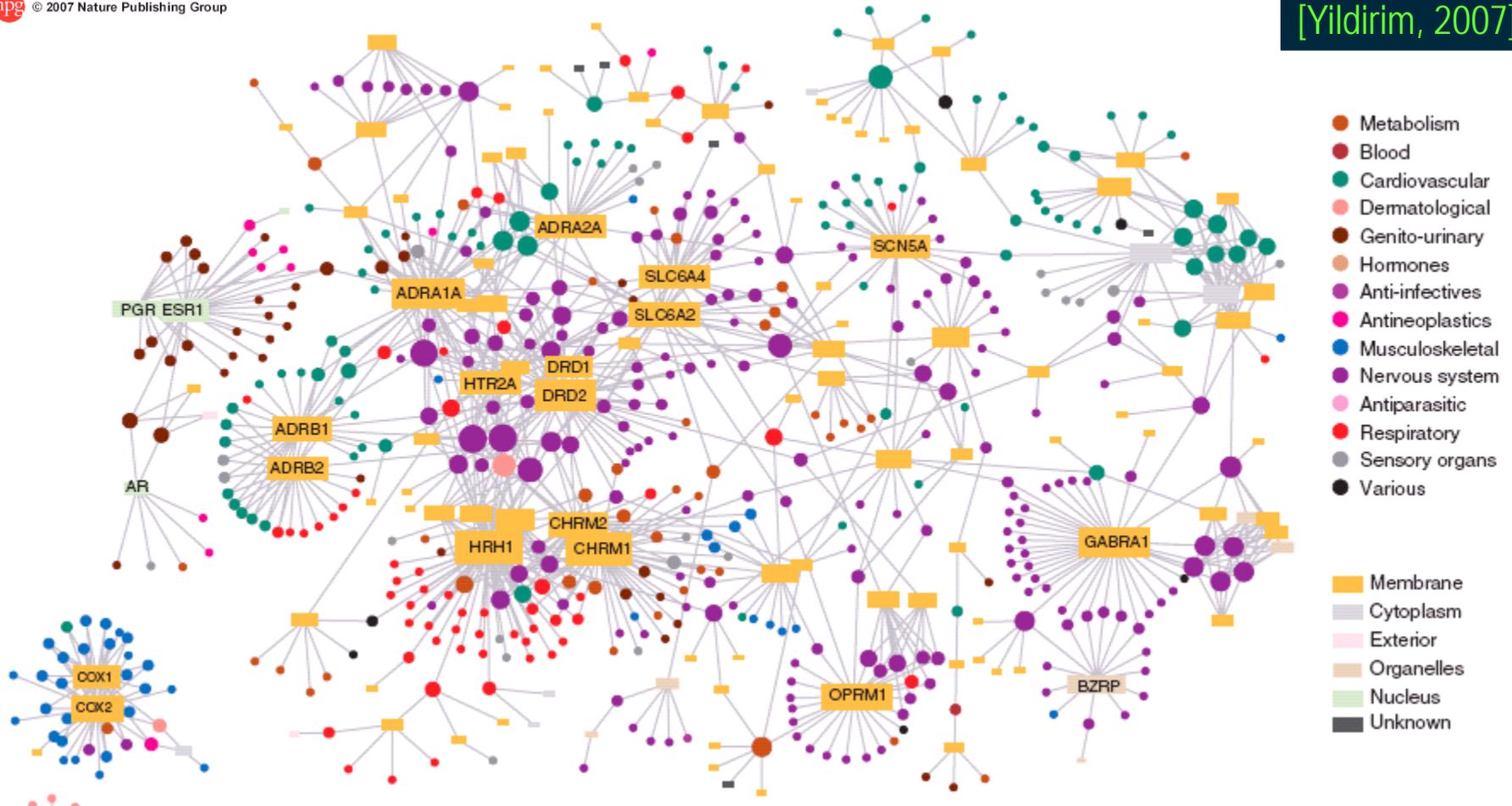
Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA

Why integrate data?

Integration yields nice pictures!

mpg © 2007 Nature Publishing Group

[Yildirim, 2007]



Motivation Translational research

- ◆ “Bench to Bedside”
- ◆ Integration of clinical and research activities and results
- ◆ Supported by research programs
 - NIH Roadmap
 - Clinical and Translational Science Awards (CTSA)
- ◆ Requires the effective integration and exchange and of information between
 - Basic research
 - Clinical research



Translational research NIH Roadmap



NIH Roadmap FOR MEDICAL RESEARCH



Re-engineering the Clinical Research Enterprise

- ▶ [Overview](#)
- ▶ [Implementation Group Members](#)
- ▶ [Funding Opportunities](#)
- ▶ [Funded Research](#)
- ▶ [Meetings](#)
- ▶ [Mid-course Reviews](#)

▶ [CTSAweb.org](#) [EXIT Disclaimer](#)

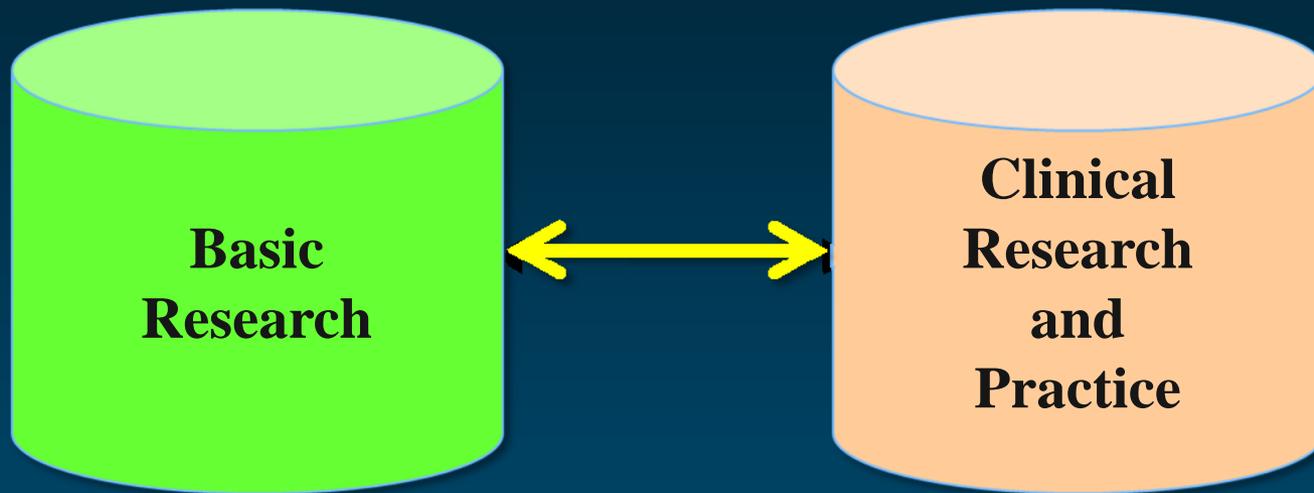
TRANSLATIONAL RESEARCH

OVERVIEW

To improve human health, scientific discoveries must be translated into practical applications. Such discoveries typically begin at "the bench" with basic research — in which scientists study disease at a molecular or cellular level — then progress to the clinical level, or the patient's "bedside."

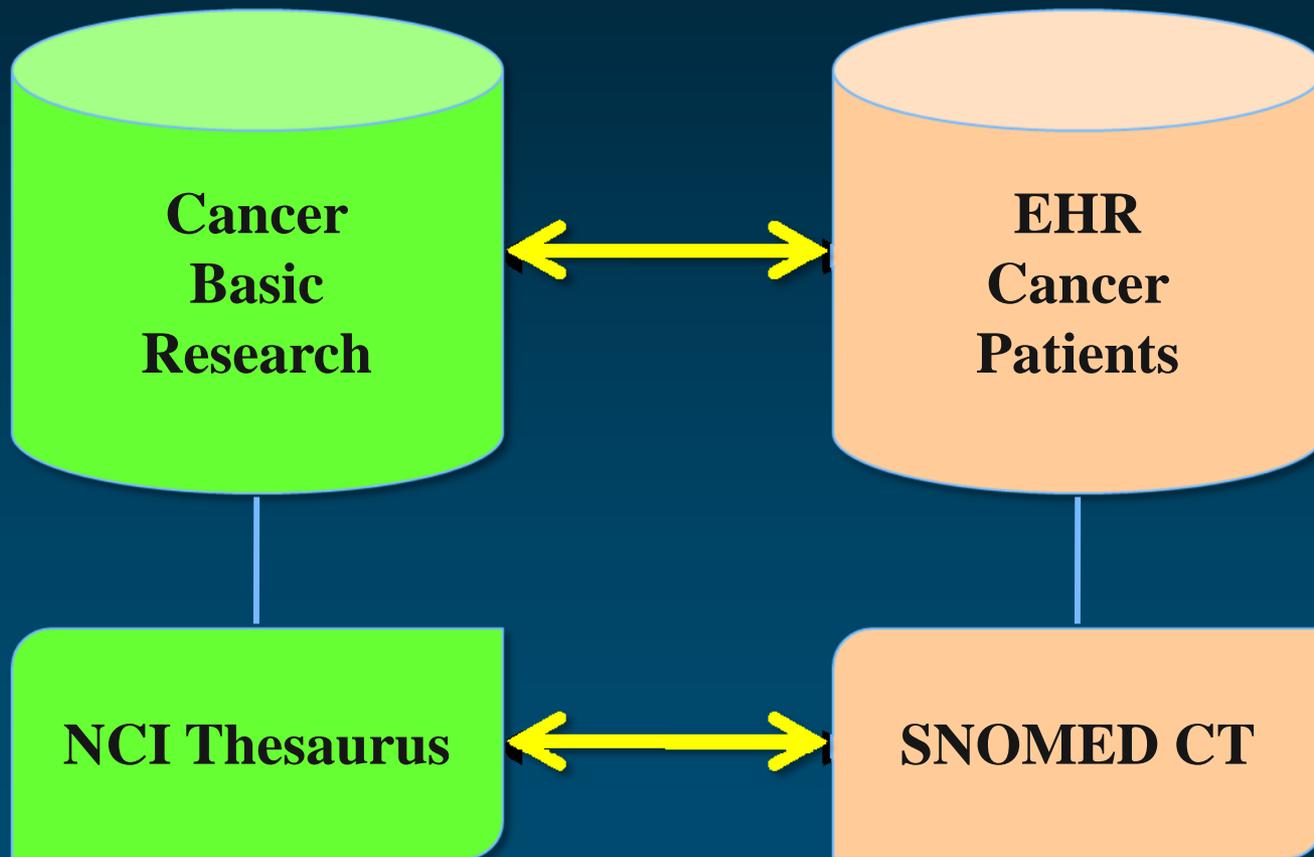
Scientists are increasingly aware that this bench-to-bedside approach to translational research is really a two-way street. Basic scientists provide clinicians with new tools for use in patients and for assessment of their impact, and clinical researchers make novel observations about the nature and progression of disease that often stimulate basic investigations.

Motivation Translational research



Why ontologies?

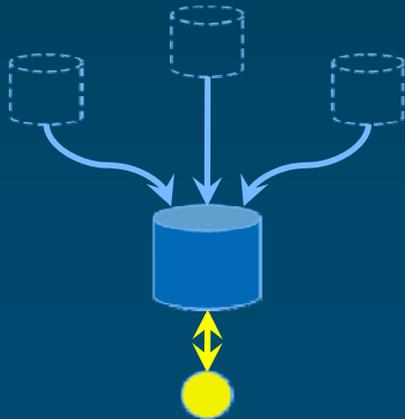
Terminology and translational research



Approaches to data integration

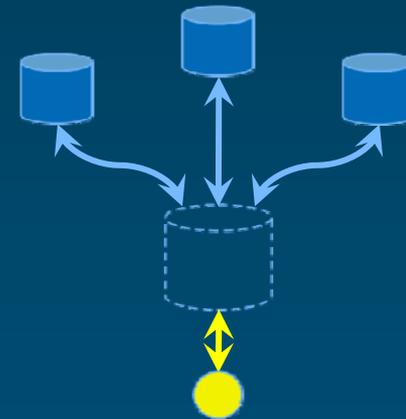
◆ Warehousing

- Sources to be integrated are transformed into a common format and converted to a common vocabulary



◆ Mediation

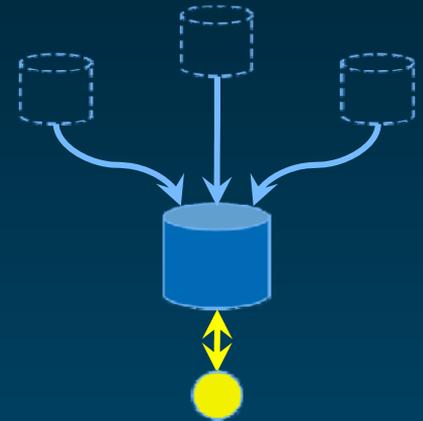
- Local schema (of the sources)
- Global schema (in reference to which the queries are made)



Ontologies and warehousing

◆ Role

- Provide a conceptualization of the domain
 - Help define the schema
 - Information model vs. ontology
- Provide value sets for data elements
- Enable standardization and sharing of data



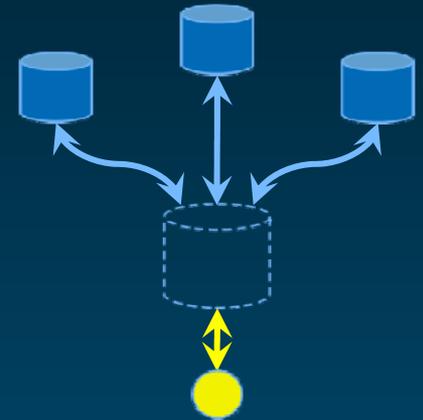
◆ Examples

- Annotations to the Gene Ontology
- Repositories for translational research (CTSA)
- Clinical information systems

Ontologies and mediation

◆ Role

- Reference for defining the global schema
- Map between local and global schemas



◆ Examples

- TAMBIS
- BioMediator
- OntoFusion

Success stories

Gene Ontology

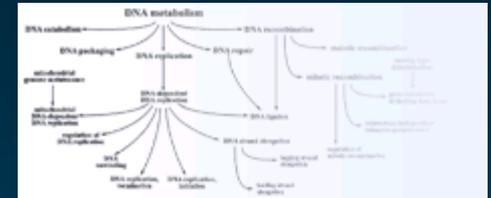
<http://www.geneontology.org/>



Annotating data

◆ Gene Ontology

- Functional annotation of gene products in several dozen model organisms



◆ Various communities use the same controlled vocabularies

◆ Enabling comparisons across model organisms

◆ Annotations

- Assigned manually by curators
- Inferred automatically (e.g., from sequence similarity)

GO Annotations for Aldh2 (mouse)

GO Annotations in Tabular Form (Text View) (GO Graph



Category	Classification Term	Evidence
Molecular Function	aldehyde dehydrogenase (NAD) activity	IEA
Molecular Function	oxidoreductase activity	IEA
Molecular Function	oxidoreductase activity	IEA
Cellular Component	mitochondrion	IDA
Biological Process	metabolic process	IEA
Biological Process	oxidation reduction	IEA

[http:// www.informatics.jax.org/](http://www.informatics.jax.org/)

GO ALD4 in Yeast

GO Annotations

Molecular Function

Manually curated

Biological Process

Manually curated

Cellular Component

Manually curated

High-throughput

All **ALD4** GO evidence and references

*View Computational GO annotations for **ALD4***

- aldehyde dehydrogenase (NAD) activity (IDA, IMP, ISS)
- aldehyde dehydrogenase [NAD(P)+] activity (IDA)

- ethanol metabolic process (IMP)

- mitochondrial nucleoid (IDA)
- mitochondrion (IMP, ISS)
- mitochondrion (IDA)



<http://db.yeastgenome.org/>



GO Annotations for ALDH2 (Human)



Function						
GO:0016491	oxidoreductase activity	interpro	IEA	IPR015590	UniProt	9606
GO:0016491	oxidoreductase activity	interpro	IEA	IPR016160	UniProt	9606
GO:0016491	oxidoreductase activity	interpro	IEA	IPR016162	UniProt	9606
GO:0016491	oxidoreductase activity	interpro	IEA	IPR016161	UniProt	9606
GO:0016491	oxidoreductase activity	spkw	IEA	KW-0560	UniProt	9606
GO:0004029	aldehyde dehydrogenase (NAD) activity	1306115	TAS		PINC	9606
GO:0004030	aldehyde dehydrogenase [NAD(P)+] activity	8903321	TAS		PINC	9606
GO:0009055	electron carrier activity	8903321	TAS		UniProt	9606
GO:0004029	aldehyde dehydrogenase (NAD) activity	enzyme	IEA	1.2.1.3	UniProt	9606

<http://www.ebi.ac.uk/GOA/>

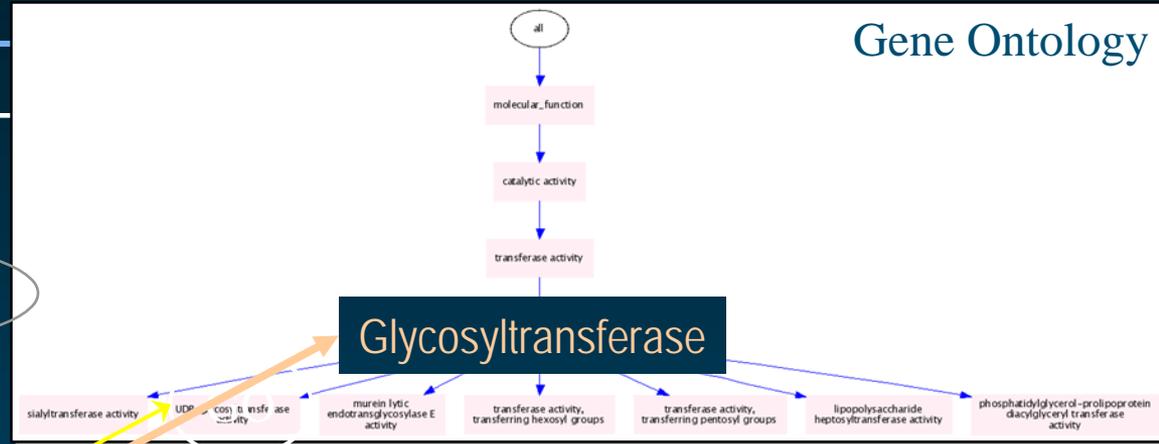
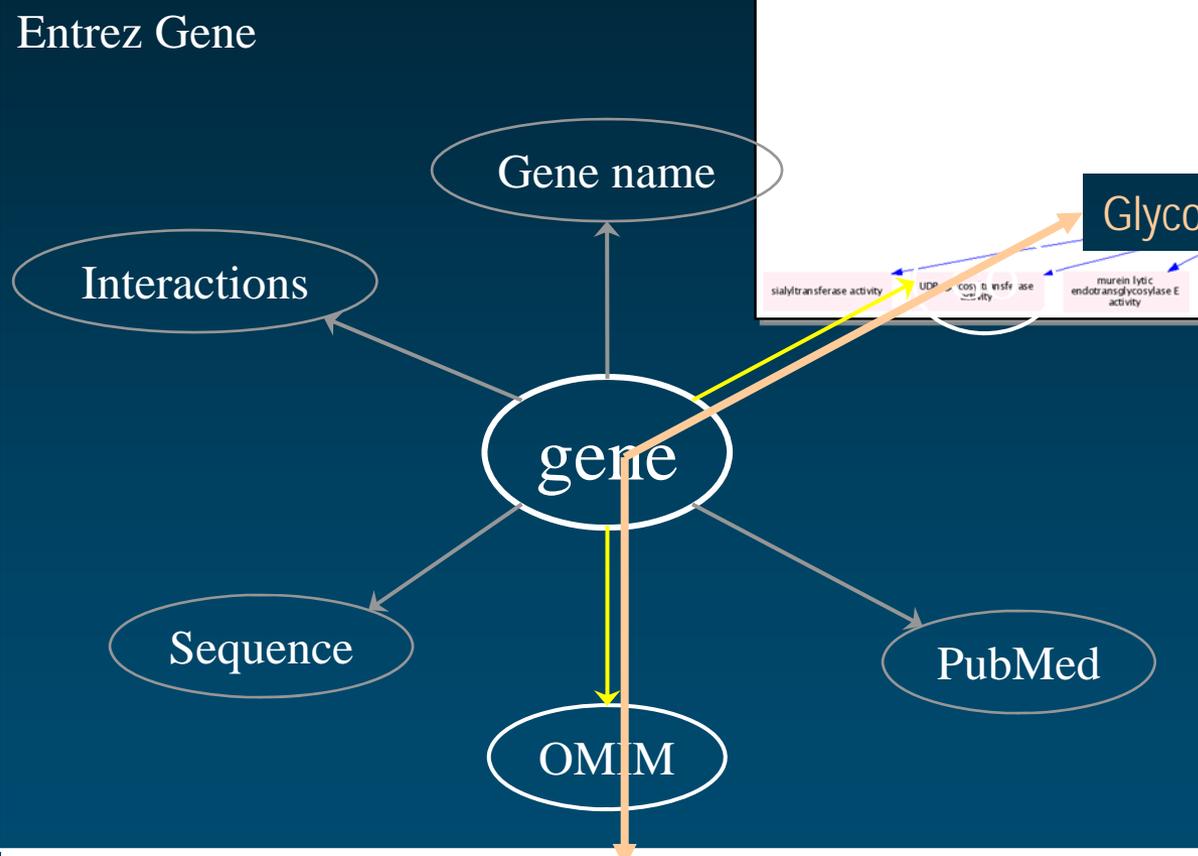


Integration applications

- ◆ Based on shared annotations
 - Enrichment analysis (within/across species)
 - Clustering (co-clustering with gene expression data)
- ◆ Based on the structure of GO
 - Closely related annotations
 - Semantic similarity [Lord, PSB 2003]
- ◆ Based on associations between gene products and annotations [Bodenreider, PSB 2005]
- ◆ Leveraging reasoning [Sahoo, Medinfo 2007]



Integration Entrez Gene + GO

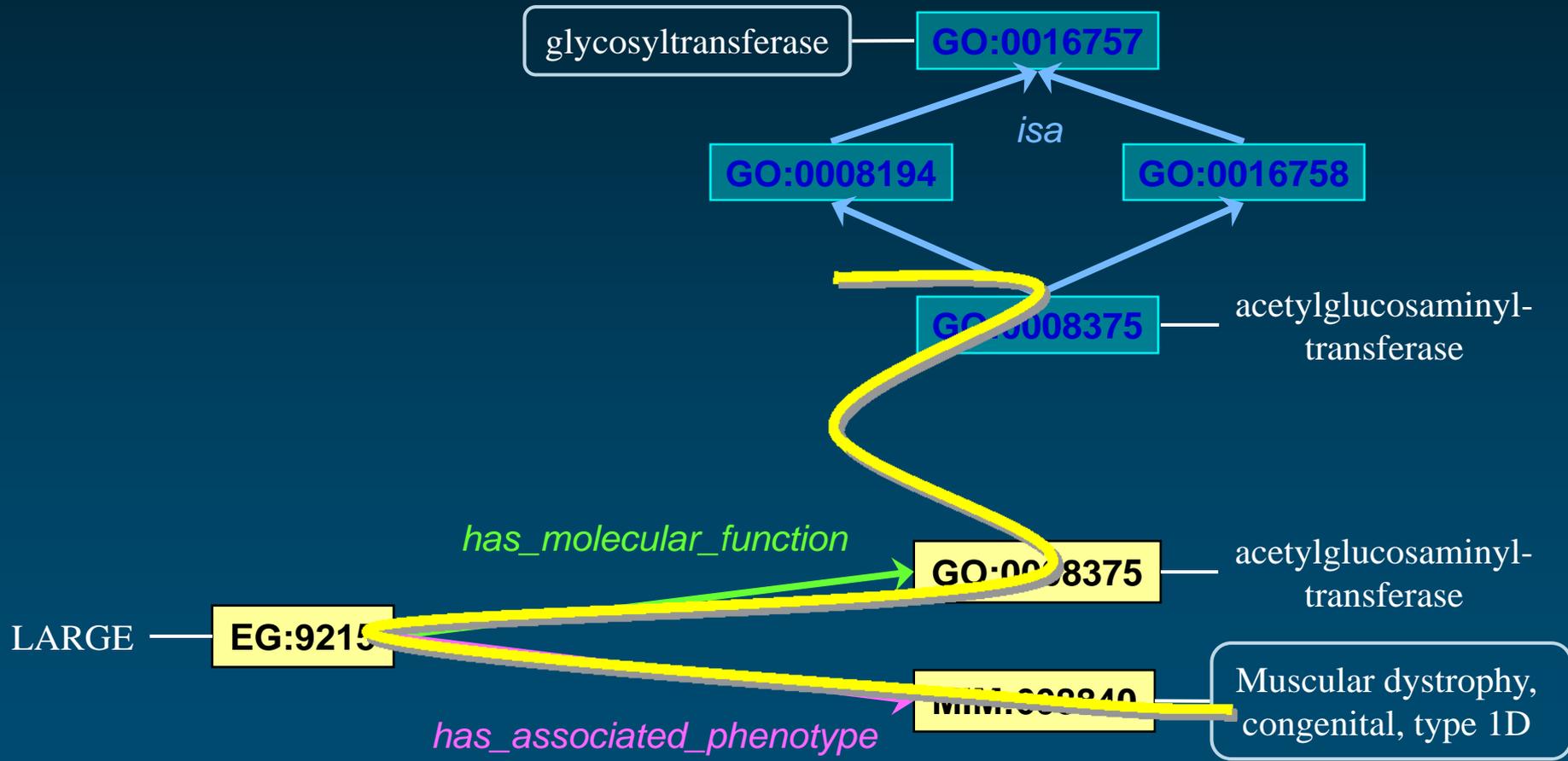


[Sahoo, Medinfo 2007]

Congenital muscular dystrophy



From *glycosyltransferase* to congenital muscular dystrophy



Success stories

caBIG

<http://cabig.nci.nih.gov/>



National Cancer Institute



caBIG[™]

Cancer Biomedical
Informatics Grid[™]

Cancer Biomedical Informatics Grid

- ◆ US National Cancer Institute
- ◆ Common infrastructure used to share data and applications across institutions to support cancer research efforts in a grid environment
- ◆ Data and application services available on the grid
- ◆ Supported by ontological resources



caBIG services

- ◆ caArray
 - Microarray data repository
- ◆ caTissue
 - Biospecimen repository
- ◆ caFE (Cancer Function Express)
 - Annotations on microarray data
- ◆ ...

- ◆ caTRIP
 - Cancer Translational Research Informatics Platform
 - Integrates data services



Ontological resources

◆ NCI Thesaurus

- Reference terminology for the cancer domain
- ~ 60,000 concepts
- OWL Lite

◆ Cancer Data Standards Repository (caDSR)

- Metadata repository
- Used to bridge across UML models through Common Data Elements
- Links to concepts in ontologies



Success stories

*Semantic Web
for Health Care and Life Sciences*

<http://www.w3.org/2001/sw/hcls/>



W3C Health Care and Life Sciences IG



Semantic Web Health Care and Life Sciences (HCLS) Interest Group

Introduction

The **mission** of the Semantic Web Health Care and Life Sciences Interest Group, part of the [Semantic Web Activity](#), is to develop, advocate for, and support the use of Semantic Web technologies for biological science, translational medicine and health care. These domains stand to gain tremendous benefit by adoption of Semantic Web technologies, as they depend on the interoperability of information from many domains and processes for efficient decision support.

The group will:

- ◆ Document use cases to aid individuals in understanding the business and technical benefits of using Semantic Web technologies.
- ◆ Document guidelines to accelerate the adoption of the technology.
- ◆ Implement a selection of the use cases as proof-of-concept demonstrations.
- ◆ Explore the possibility of developing high level vocabularies.
- ◆ Disseminate information about the group's work at government, industry, and academic events.

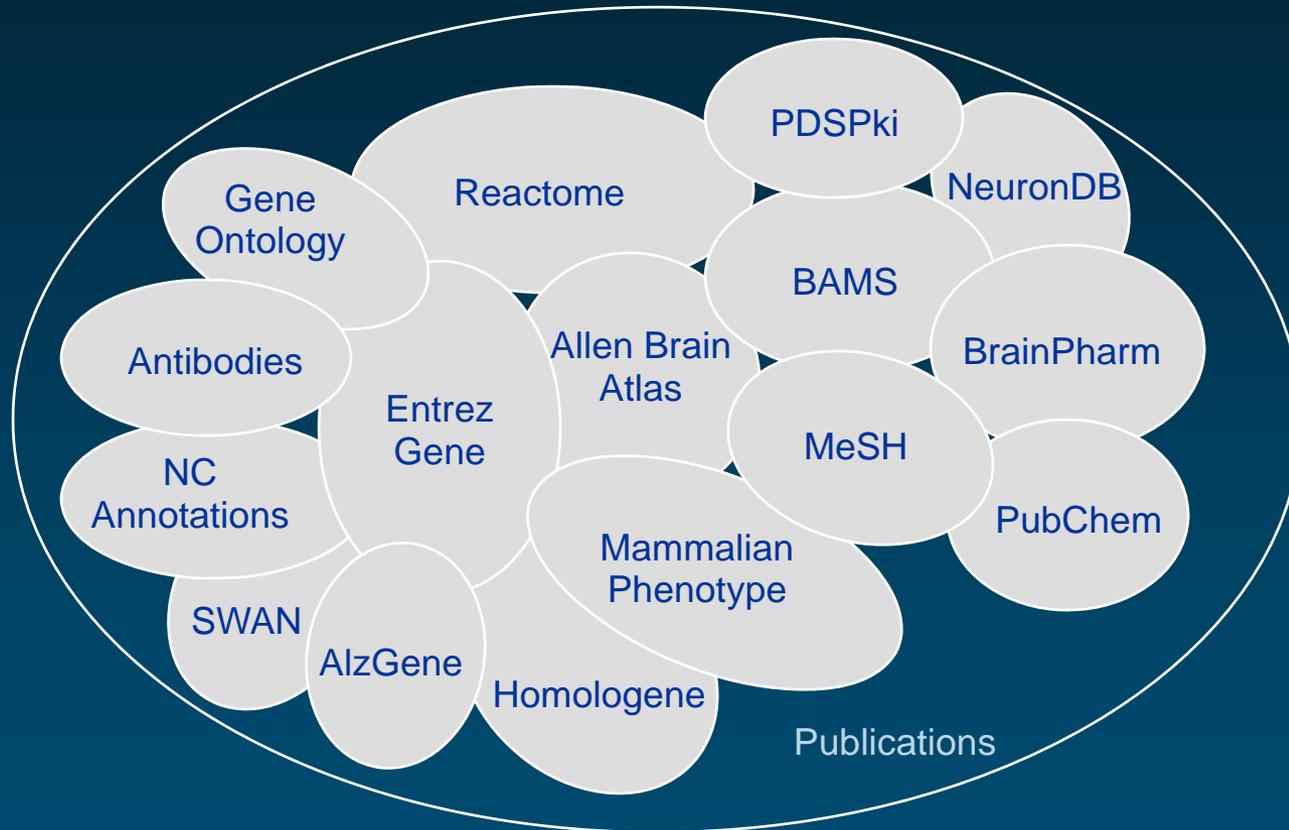
Biomedical Semantic Web

- ◆ Integration
 - Data/Information
 - E.g., translational research
- ◆ Hypothesis generation
- ◆ Knowledge discovery

[Ruttenberg, 2007]



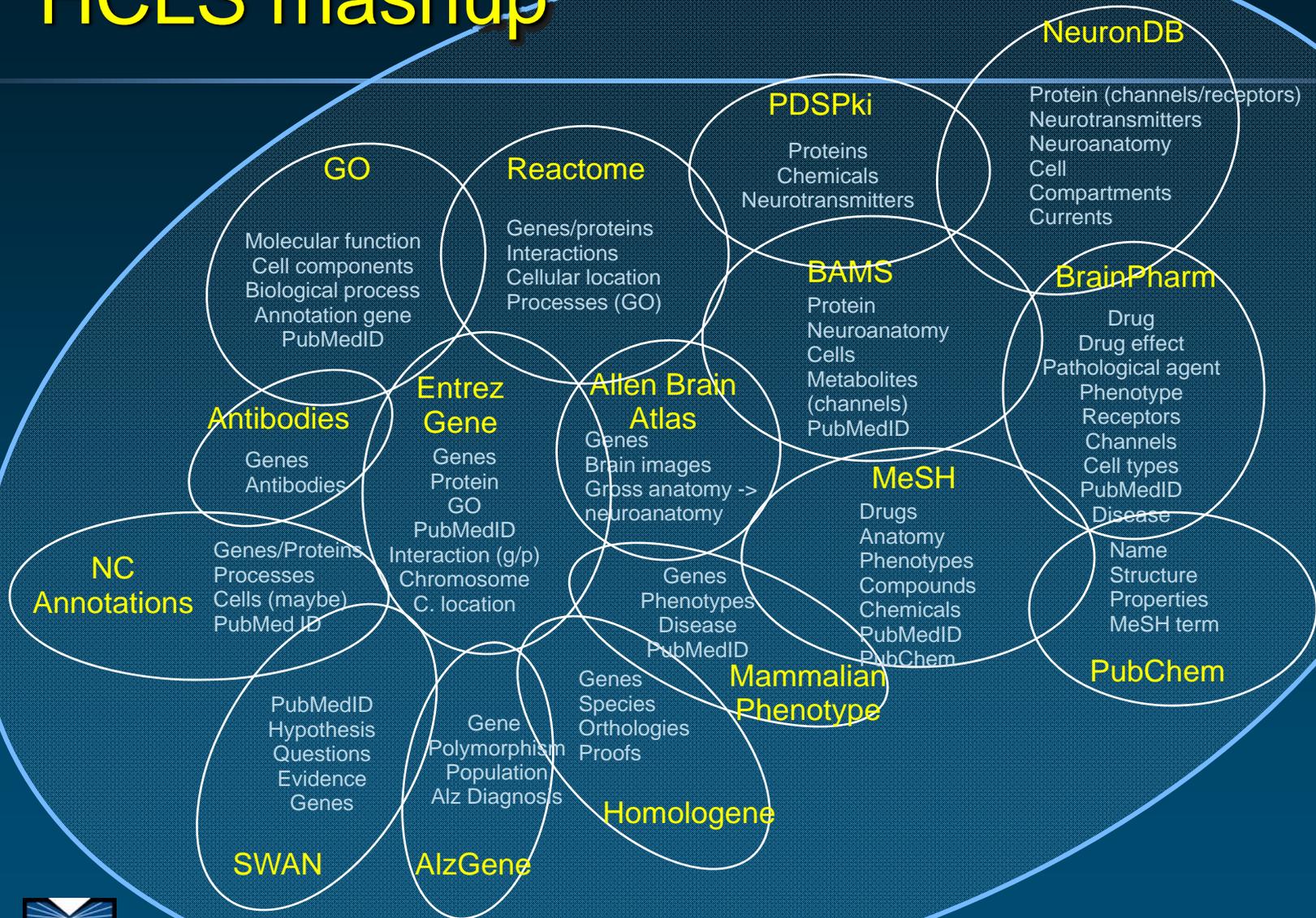
HCLS mashup of biomedical sources



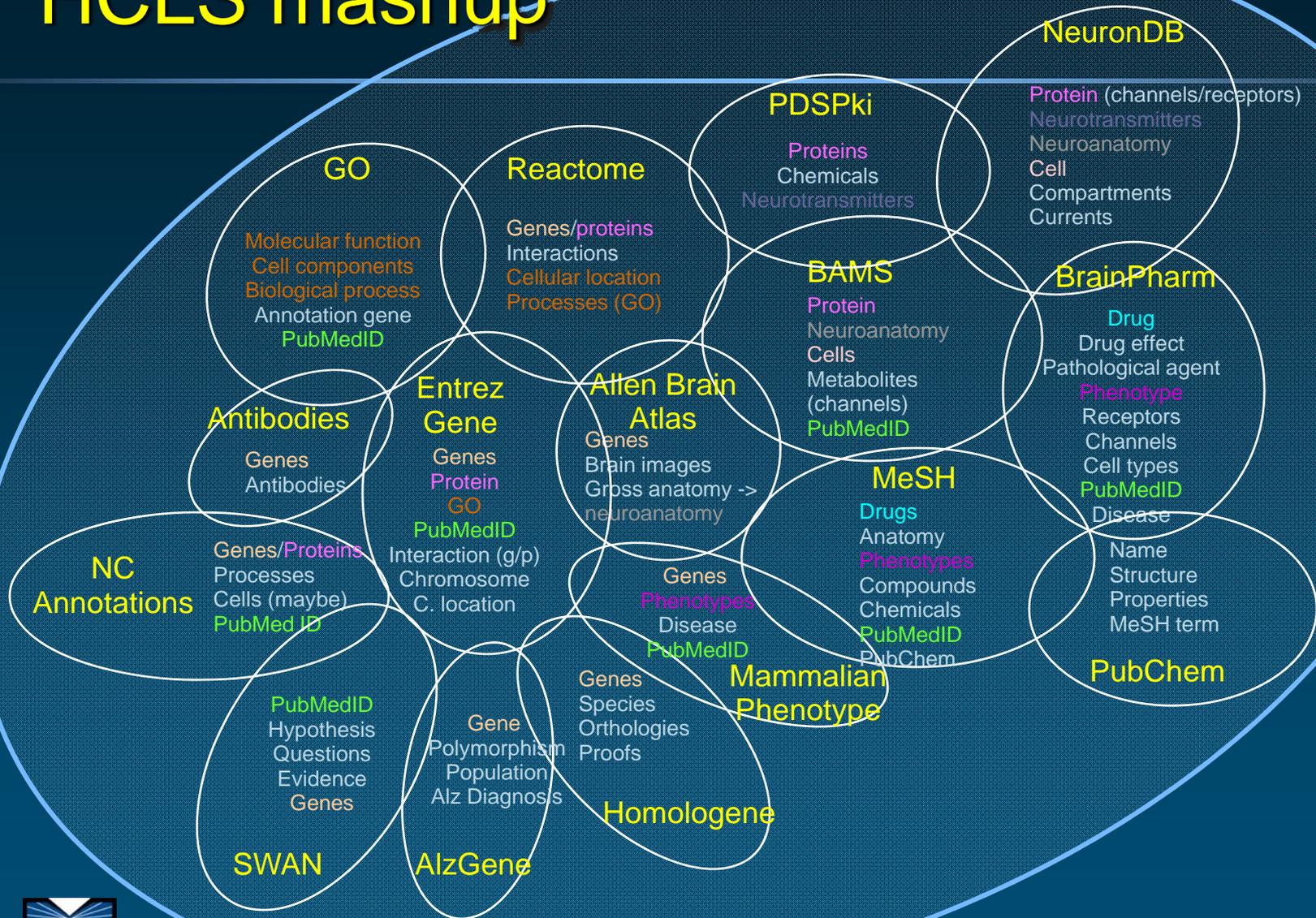
http://esw.w3.org/topic/HCLS/HCLSIG_DemoHomePage_HCLSIG_Demo



HCLS mashup



HCLS mashup



HCLS mashups

- ◆ Based on RDF/OWL
- ◆ Based on shared identifiers
 - “Recombinant data” (E. Neumann)
- ◆ Ontologies used in some cases
- ◆ Support applications (SWAN, SenseLab, etc.)

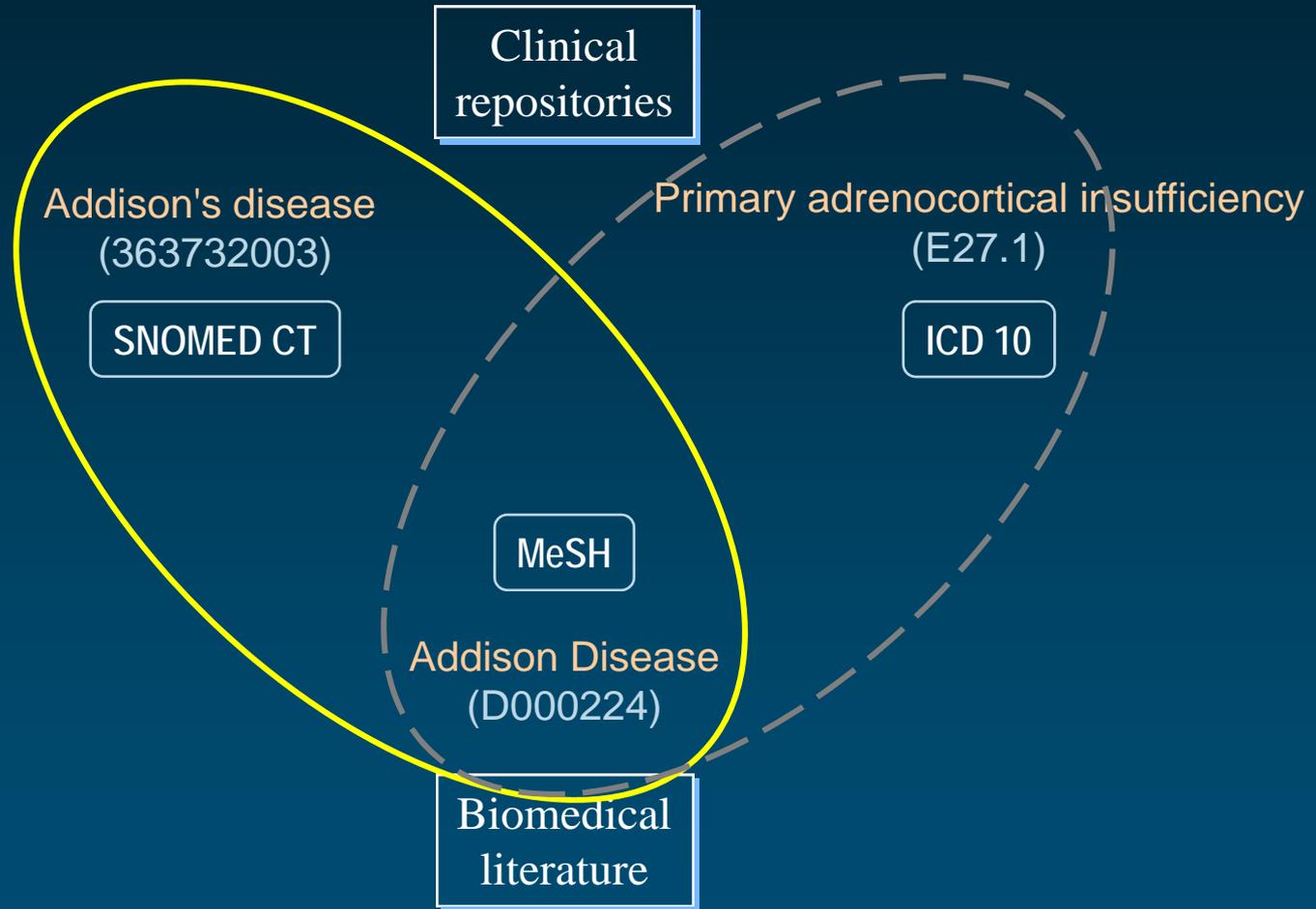
- ◆ Journal of Biomedical Informatics
special issue on Semantic Bio-mashups
(forthcoming)



Challenging issues

Bridges across ontologies

Trans-namespace integration



(Integrated) concept repositories

- ◆ Unified Medical Language System

<http://umlsks.nlm.nih.gov>

- ◆ NCBO's BioPortal

<http://www.bioontology.org/tools/portal/bioportal.html>

- ◆ caDSR

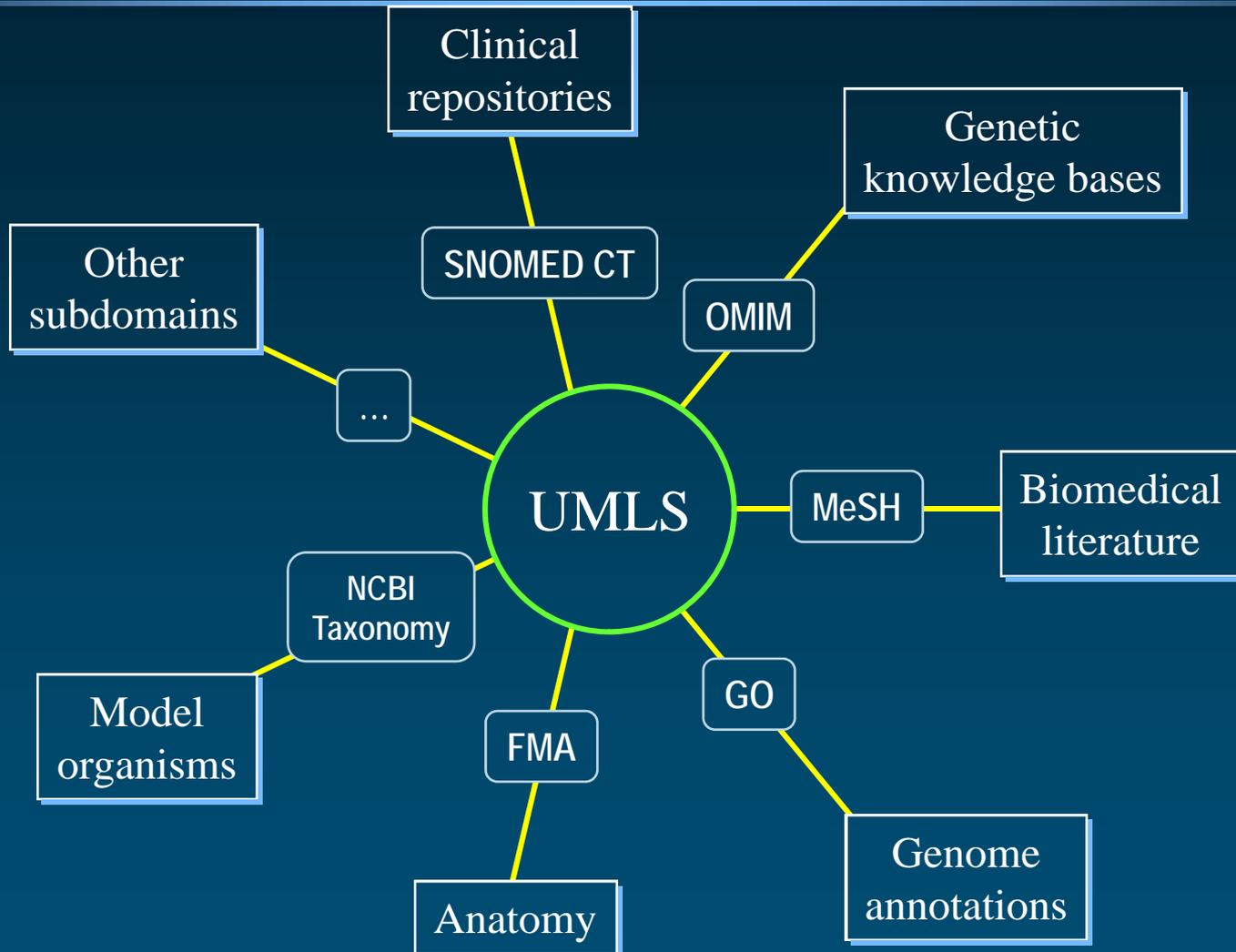
http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/cadsr

- ◆ Open Biomedical Ontologies (OBO)

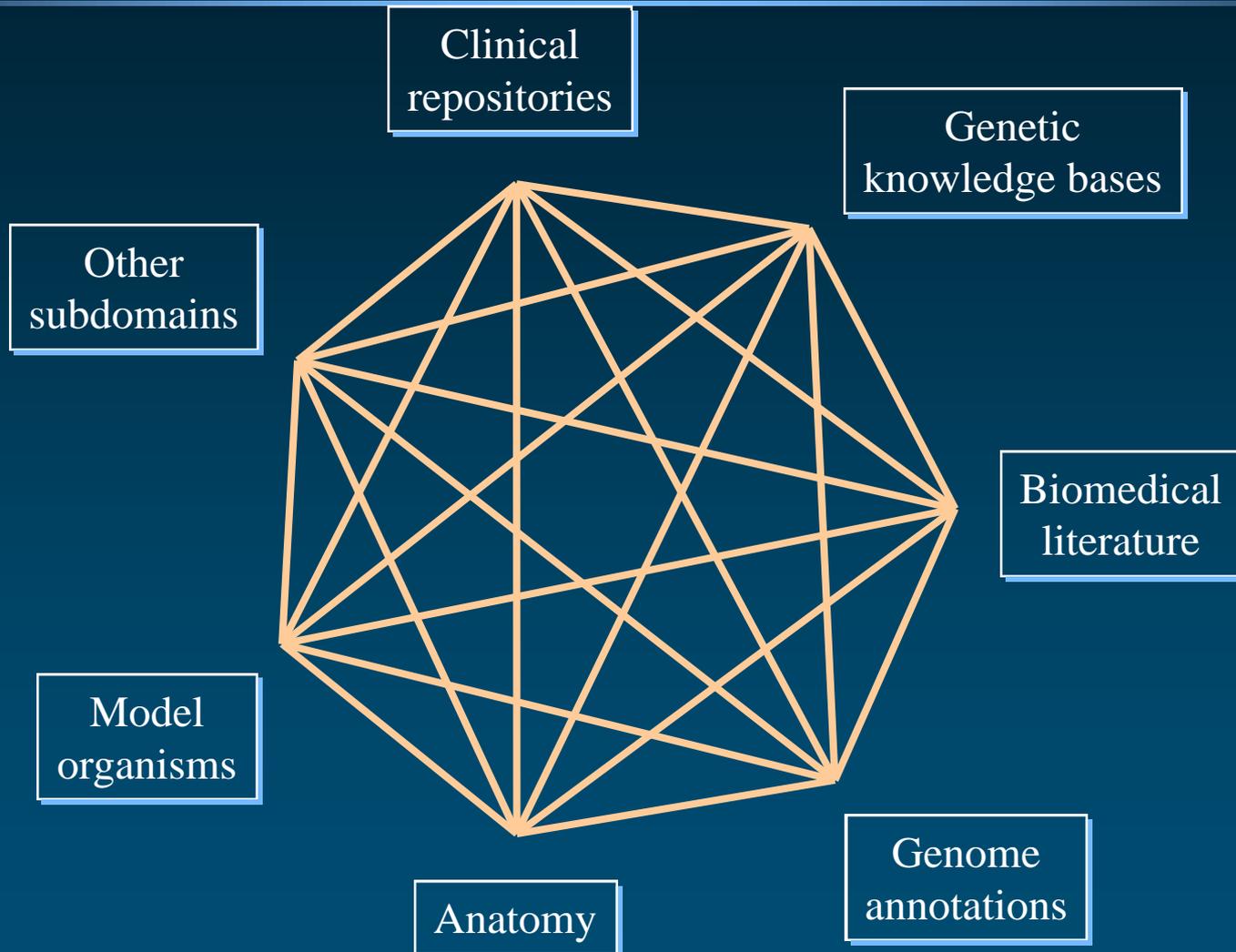
<http://obofoundry.org/>



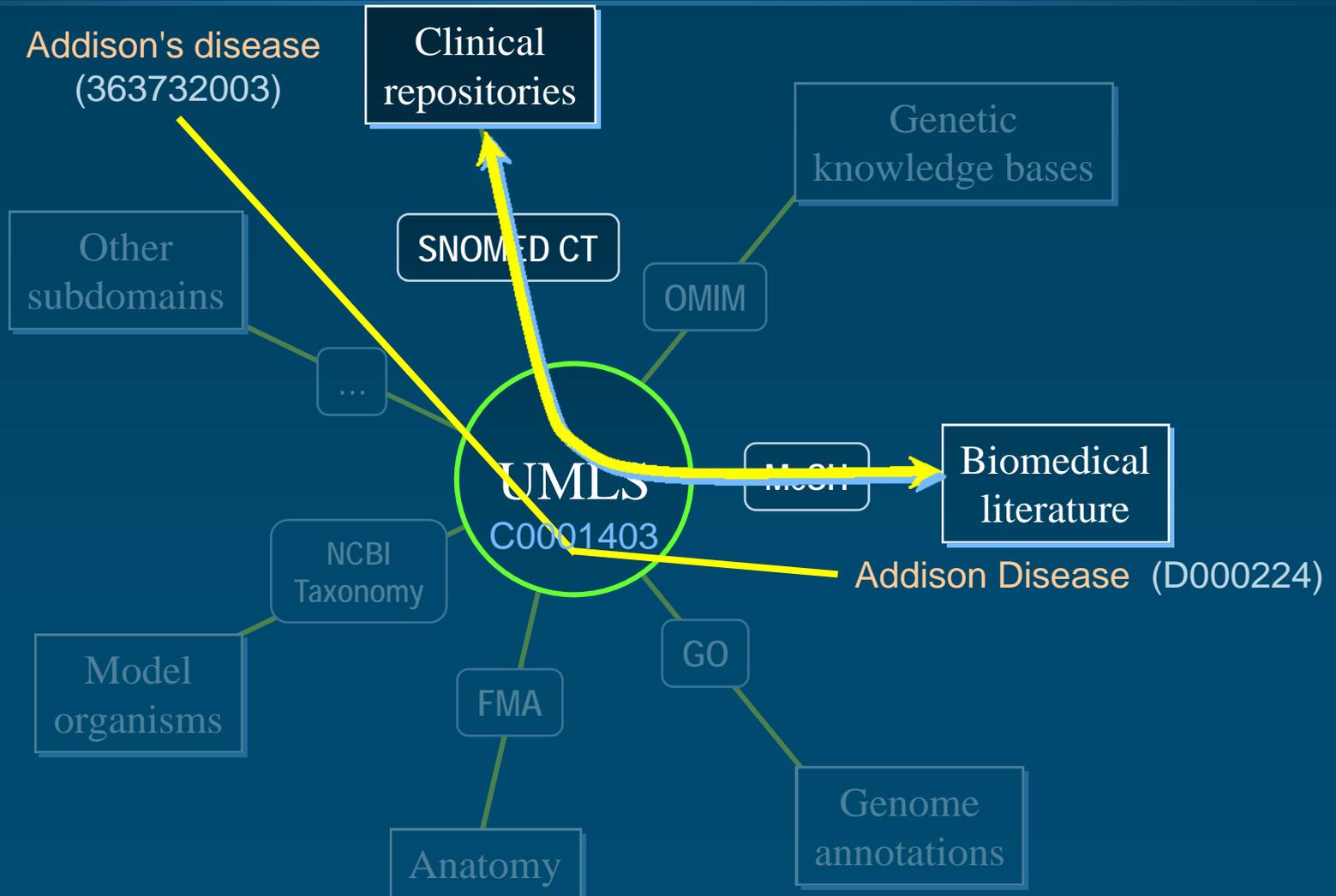
Integrating subdomains



Integrating subdomains



Trans-namespace integration



Mappings

- ◆ Created manually (e.g., UMLS)
 - Purpose
 - Directionality
- ◆ Created automatically (e.g., BioPortal)
 - Lexically: ambiguity, normalization
 - Semantically: lack of / incomplete formal definitions
- ◆ Key to enabling semantic interoperability
- ◆ Enabling resource for the Semantic Web

Challenging issues

Permanent identifiers for biomedical entities

Identifying biomedical entities

- ◆ Multiple identifiers for the same entity in different ontologies
- ◆ Barrier to data integration in general
 - Data annotated to different ontologies cannot “recombine”
 - Need for mappings across ontologies
- ◆ Barrier to data integration in the Semantic Web
 - Multiple possible identifiers for the same entity
 - Depending on the underlying representational scheme (URI vs. LSID)
 - Depending on who creates the URI

Possible solutions

- ◆ PURL <http://purl.org>
 - One level of indirection between developers and users
 - Independence from local constraints at the developer's end
- ◆ The institution creating a resource is also responsible for minting URIs
 - E.g., URI for genes in Entrez Gene
- ◆ Guidelines: “URI note”
 - W3C Health Care and Life Sciences Interest Group



Challenging issues

Other issues

Availability

- ◆ Many ontologies are freely available
- ◆ The UMLS is freely available for research purposes
 - Cost-free license required
- ◆ Licensing issues can be tricky
 - SNOMED CT is freely available in member countries of the IHTSDO
- ◆ Being freely available
 - Is a requirement for the Open Biomedical Ontologies (OBO)
 - Is a *de facto* prerequisite for Semantic Web applications



Discoverability

- ◆ Ontology repositories
 - UMLS: 143 source vocabularies
(biased towards healthcare applications)
 - NCBO BioPortal: ~100 ontologies
(biased towards biological applications)
 - Limited overlap between the two repositories
- ◆ Need for discovery services

Formalism

◆ Several major formalism

- Web Ontology Language (OWL) – NCI Thesaurus
- OBO format – most OBO ontologies
- UMLS Rich Release Format (RRF) – UMLS, RxNorm

◆ Conversion mechanisms

- OBO to OWL
- LexGrid (import/export to LexGrid internal format)



Ontology integration

- ◆ *Post hoc* integration , form the bottom up
 - UMLS approach
 - Integrates ontologies “as is”, including legacy ontologies
 - Facilitates the integration of the corresponding datasets
- ◆ Coordinated development of ontologies
 - OBO Foundry approach
 - Ensures consistency *ab initio*
 - Excludes legacy ontologies

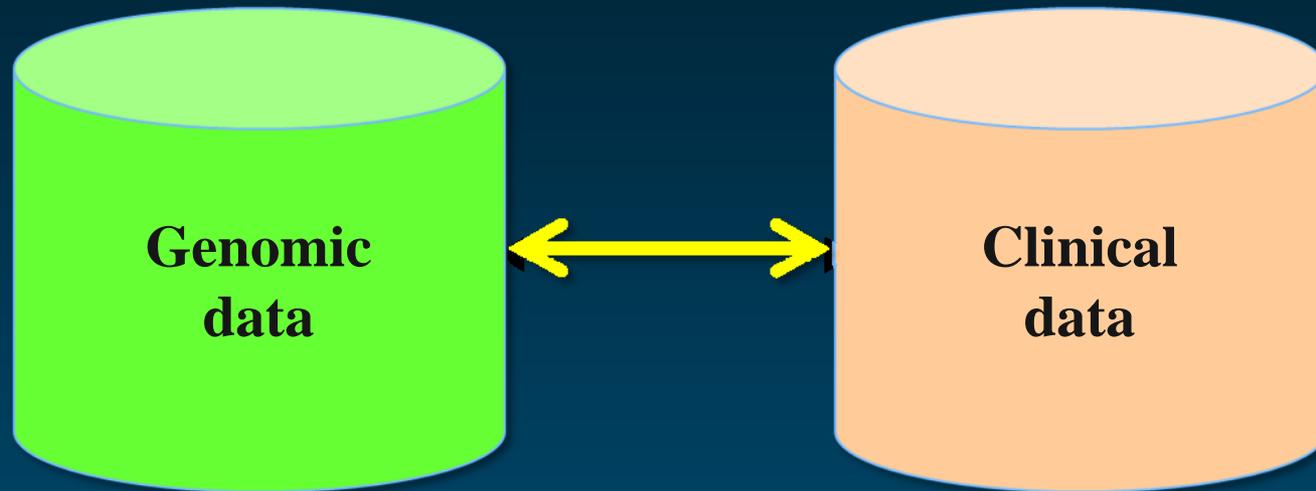
Quality

- ◆ Quality assurance in ontologies is still imperfectly defined
 - Difficult to define outside a use case or application
- ◆ Several approaches to evaluating quality
 - Collaboratively, by users (Web 2.0 approach)
 - Marginal notes enabled by BioPortal
 - Centrally, by experts
 - OBO Foundry approach
- ◆ Important factors besides quality
 - Governance
 - Installed base / Community of practice

Thinking outside the integration box

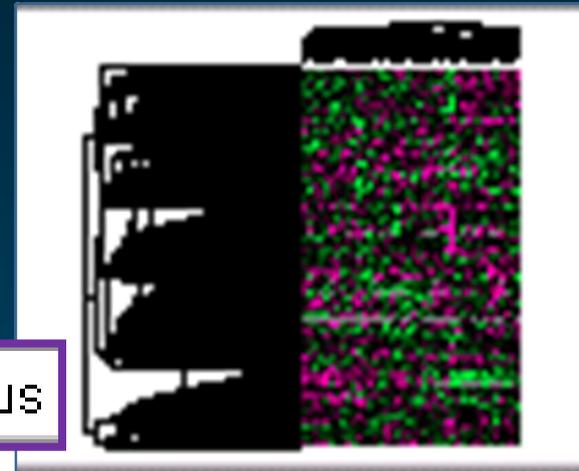
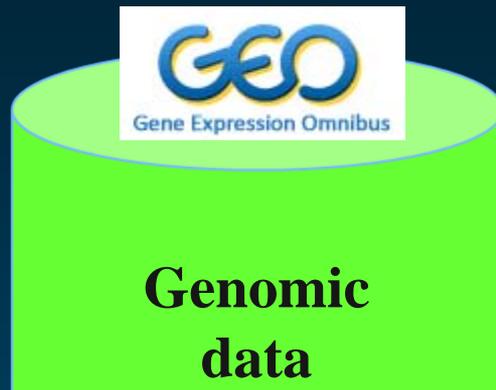
The Butte approach

Integrating genomic and clinical data



- ◆ No genomic data available for most patients
- ◆ No precise clinical data available associated with most genomic data (GWAS excepted)

Integrating genomic and clinical data



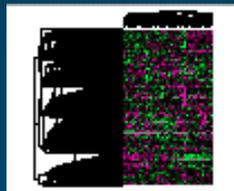
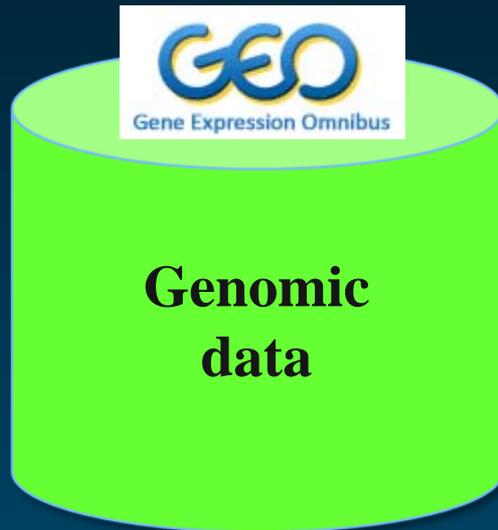
Aging and cognitive impairment: hippocampus

Accession:	GDS2639 View Expression (GEO profiles)	
Title:	Aging and cognitive impairment: hippocampus	
DataSet type:	gene expression array-based (RNA / in situ oligonucleotide)	
Summary:	Analysis of hippocampi from aged learning-impaired animals on the last day of training in the Morris water maze (MWZ) or 21 days post-training. The MWZ task is a dorsal hippocampal-dependent task. Results provide insight into the molecular basis of aging-related cognitive impairment.	
Platform:	GPL341: Affymetrix GeneChip Rat Expression Set 230 Array RAE230A	
Citations:	Rowe WB, Blalock EM, Chen KC, Kadish I et al. Hippocampal expression analyses reveal selective association of immediate-early, neuroenergetic, and myelinogenic p rats. <i>J Neurosci</i> 2007 Mar 21;27(12):3098-110. PMID: 17376971	
Sample organism:	Rattus norvegicus	Platform organism:
Feature count:	15923	Value type:
Series:	GSE5666	Series published:
Last GDS update:	04/27/2007	

MeSH Terms:

- ◆ [Age Factors](#)
- ◆ [Animals](#)
- ◆ [Cognition Disorders/genetics](#)

Integrating genomic and clinical data



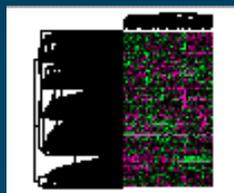
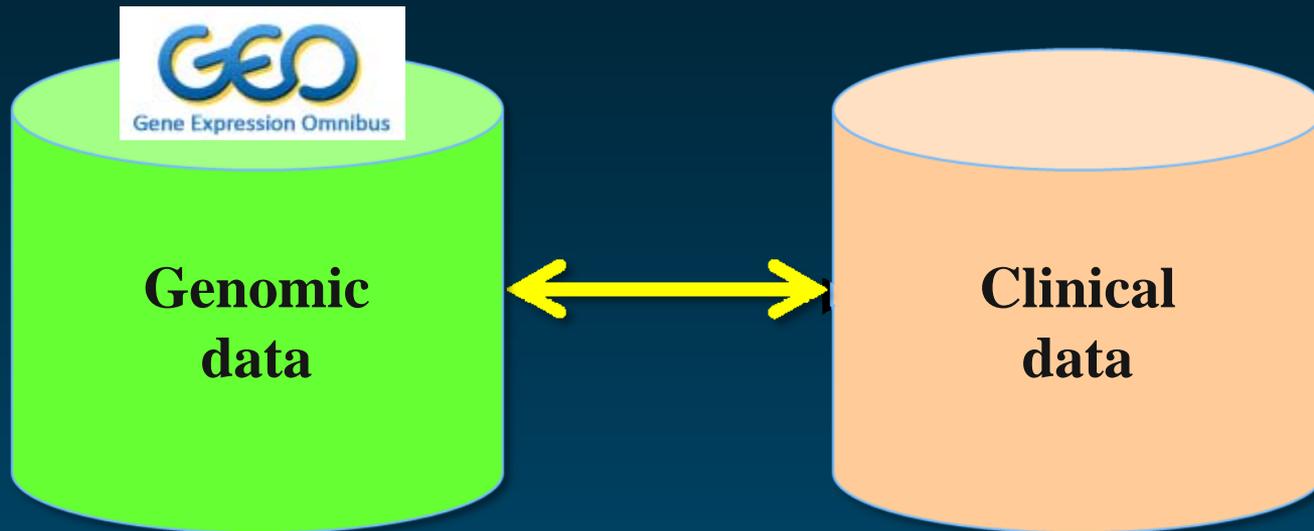
Upregulated genes

GEO Summary			
Accession:	GSE20000 View Expression (GEO profiles)		
Title:	Aging and cognitive impairment: hippocampus		
DataSet type:	gene expression array-based (RNA in situ oligonucleotide)		
Summary:	Analysis of hippocampus from aged learning impaired animals on the last day of training in the Morris water maze (MWM) or 21 days post-training. The MWM task is a dorsal hippocampal-dependent task. Results provide insight into the molecular basis of aging-related cognitive impairment.		
Platform:	GPL341 Affymetrix GeneChip Rat Expression Set 230 Array RA230A		
Citation:	Rowe WB, Blalock EM, Chen KC, Kadish I et al. Hippocampal expression analyses reveal selective association of immediate early, neuroenergetic, and neurotrophic pathways with cognitive impairment in aged rats. <i>J Neurosci</i> 2007 Mar 21;27(12):3096-110. PMID: 17376873		
Sample organism:	Rattus norvegicus	Platform organism:	Rattus norvegicus
Feature count:	16923	Value type:	count
Series:	GSE20000	Series published:	03/15/2007
Last GD5 update:	04/27/2007		

Diseases
(extracted from text
+ MeSH terms)



Integrating genomic and clinical data



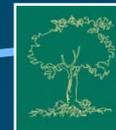
GEO Summary	
Accession:	G000000 View Expression (GEO profiles)
Title:	Aging and cognitive impairment: hippocampus
DataSet type:	gene expression array-based (RNA in situ oligonucleotide)
Summary:	Analysis of hippocampus from aged learning impaired animals on the last day of training in the Morris water maze (MWM) or 21 days post-training. The MWM task is a dorsal hippocampal-dependent task. Results provide insight into the molecular basis of aging-related cognitive impairment.
Platform:	GPL341 Affymetrix GeneChip Rat Expression Set 230 Array RAE230A
Chronic:	Rhee WB, Blalock EM, Chen JC, Kadish I et al. Hippocampal expression analyses reveal selective association of immediate early, neuroenergetic, and neurotrophic pathways with cognitive impairment in aged rats. <i>J Neurosci</i> 2007 Mar 21;27(12):3096-110. PMID: 17376573
Sample organism:	Rattus norvegicus
Platform organism:	Rattus norvegicus
Feature count:	16023
Value type:	count
Series:	G000000
Series published:	03/15/2007
Last GD5 update:	04/27/2007

Coded discharge summaries

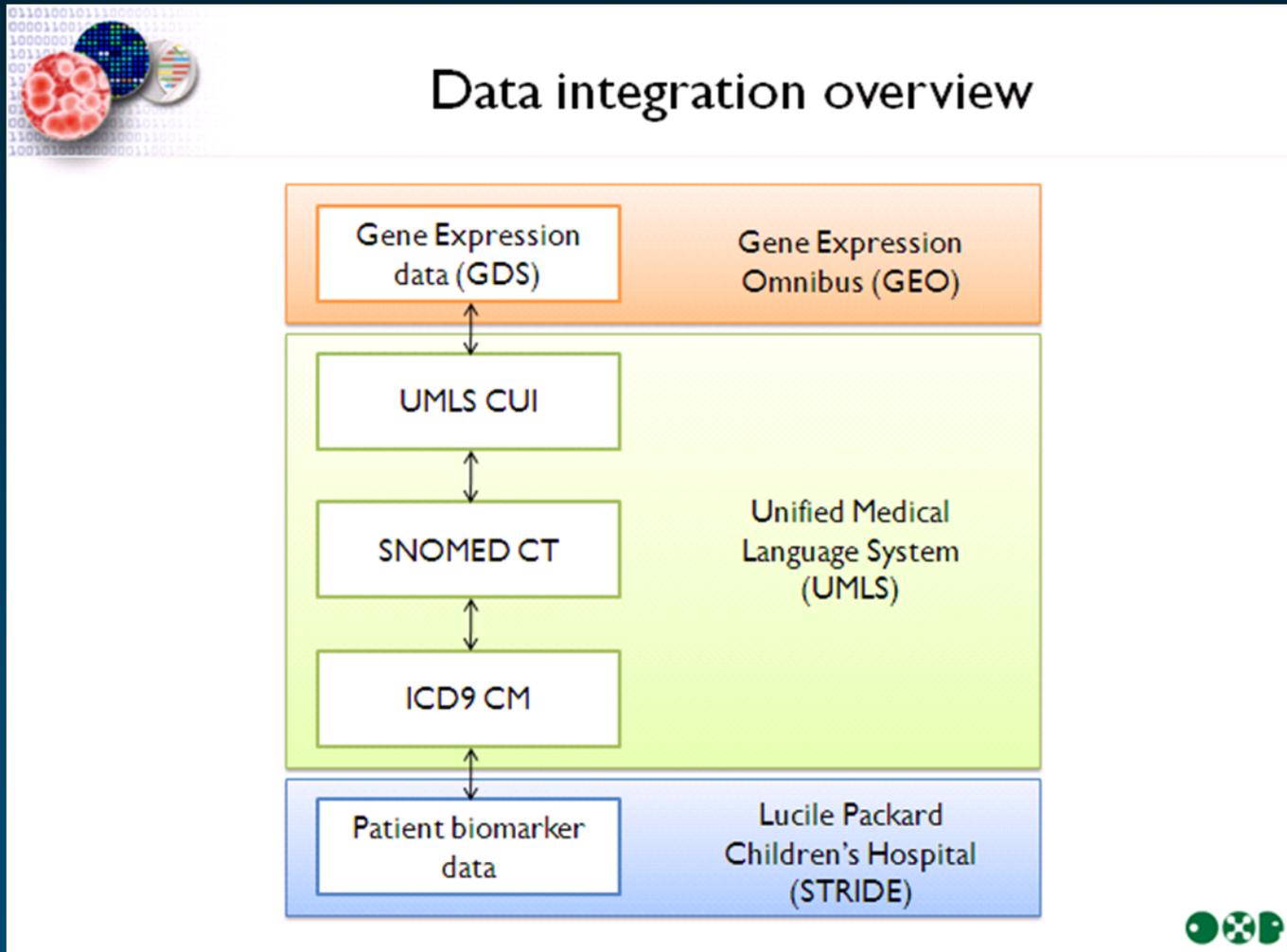
Laboratory data

Upregulated genes

Diseases (extracted from text + MeSH terms)



The Butte approach Methods



The Butte approach Results



Results

- 737 GEO Data Sets that were related to human disease
- 238 disease concepts were associated with GDS subsets
- 29,541 microarray samples were coded with SNOMED CT identifiers
- Note, we only included GDS that compared disease state to normal state
- 13,452 patients (of 49,414) mapped to 211 (of 238) of the disease concepts



The Butte approach

- ◆ Extremely rough methods
 - No pairing between genomic and clinical data
 - Text mining
 - Mapping between SNOMED CT and ICD 9-CM through UMLS
 - Reuse of ICD 9-CM codes assigned for billing purposes
- ◆ Extremely preliminary results
 - Rediscovery more than discovery
- ◆ Extremely promising nonetheless



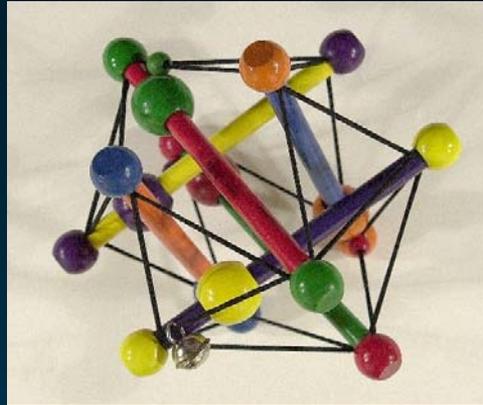
The Butte approach References

- ◆ Dudley J, Butte AJ "Enabling integrative genomic analysis of high-impact human diseases through text mining." *Pac Symp Biocomput* 2008; 580-91
- ◆ Chen DP, Weber SC, Constantinou PS, Ferris TA, Lowe HJ, Butte AJ "Novel integration of hospital electronic medical records and gene expression measurements to identify genetic markers of maturation." *Pac Symp Biocomput* 2008; 243-54
- ◆ Butte AJ, "Medicine. The ultimate model organism." *Science* 2008; 320: 5874: 325-7



Conclusions

- ◆ Ontologies are enabling resources for data integration
- ◆ Standardization works
 - Grass roots effort (GO)
 - Regulatory context (ICD 9-CM)
- ◆ Bridging across resources is crucial
 - Ontology integration resources / strategies (UMLS, BioPortal / OBO Foundry)
- ◆ Massive amounts of imperfect data integrated with rough methods might still be useful



Medical Ontology Research

Contact: olivier@nlm.nih.gov

Web: mor.nlm.nih.gov



Olivier Bodenreider

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