W3C Semantic Web for Healthcare and Life Sciences

PRISM

Lambertville, NJ

Eric Neumann

October 12, 2005

Semantic Web

"--The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation." - *Tim Berners-Lee*

What is the Semantic Web?

- Focuses on Data Semantics rather than Syntax (Semantic Interoperability)
- Open-World Graph model of all information on the Web (or in your intranet)
- RDF - Web-based descriptive model of all information
- OWL - Web Ontology Language; 3 levels of expressivity
- New Semantic Browsers with logic-driven style-sheets
- Rules and Inference engines work great with OWL and RDF

Semantic Web Vision

- Early goals
  - Using RDF-OWL technologies
  - Bottom-Up development
  - Interlinking of current ontologies
- Long-term goals
  - Intelligent connectivity of all of the Web
  - New level of services and transactions
  - Knowledge and Trust Networks

What can the Semantic Web do NOW?

- Data Integration via RDF triples and views
- Tagging of mined-Literature
- Data Federation (SPARQL)
- Knowledge Management of ideas (IP, hypotheses, decisions)
- Web service management (OWL-S)
- Business Workflow and Policy management using rules (SWRL)

Semantic Web for Life Sciences Participants MIT, Oct 27, 28 2004

Jackson Laboratories	**Berlex Biosciences**	**Novartis**	**Sanofi-Aventis**	Woods Hole Oceanographic Institute	Fred Hutchinson Cancer Research Center
Infinity Pharmaceuticals	**AstraZeneca R&D**	Elsevier	**Millenium Pharmaceuticals**	Nature Publishing Group	Pacific Northwest National Laboratory
Stanford Medical Informatics	Harvard Partners	Affymetrix	Mayo Clinic	American Chemical Society	European Bioinformatics Institute
National Science Foundation	Hewlett-Packard	**Pfizer**	**Genentech**	MacArthur Foundation	National Center for Genome Resources
Oracle	BioGrid	SemantxLS	PRISM Forum	Swiss Institute of Bioinformatics	National Cancer Institue (Center for Bioinformatics)
Children's Hospital	IBM	INRIA	University of Michigan	University of Massachusetts Boston	Harvard Medical School
AGFA Healthcare	MIT / CSBi	KEVRIC	Chevron Texaco	University of Cambridge (UK)	Fujitsu Laboratories of America
Broad Institute / MIT	MITRE	Genstruct	Network Inference	Alzheimer's Research Forum	German Cancer Research Center
Stanford Medical Informatics	Annotea	BioPAX	HydroJoule	University of Manchester	VTT Finland
Matsushita / W3C	SkyPrise	Djinnisys	Siderean	Yale Center for Medical Informatics	MIND (University of Maryland)
DSTC Pty Ltd	Technion – Israel Institute of Technology	Columbia University	Intelligent Solutions	Panther Informatics	Image Bioinformatics Lab, University of Oxford
University of Colorado	Northeastern University	Tucana Technologies		University of Georgia	Japan Biological Information Consortium
University of Zurich	University of Michigan	Life Sciences Insights	Object Management Group	**De Novo Pharmaceuticals**	European Network of Excellence REWERSE

W3C HCLSIG -
Semantic Web for HealthCare and Life Sciences Interest Group

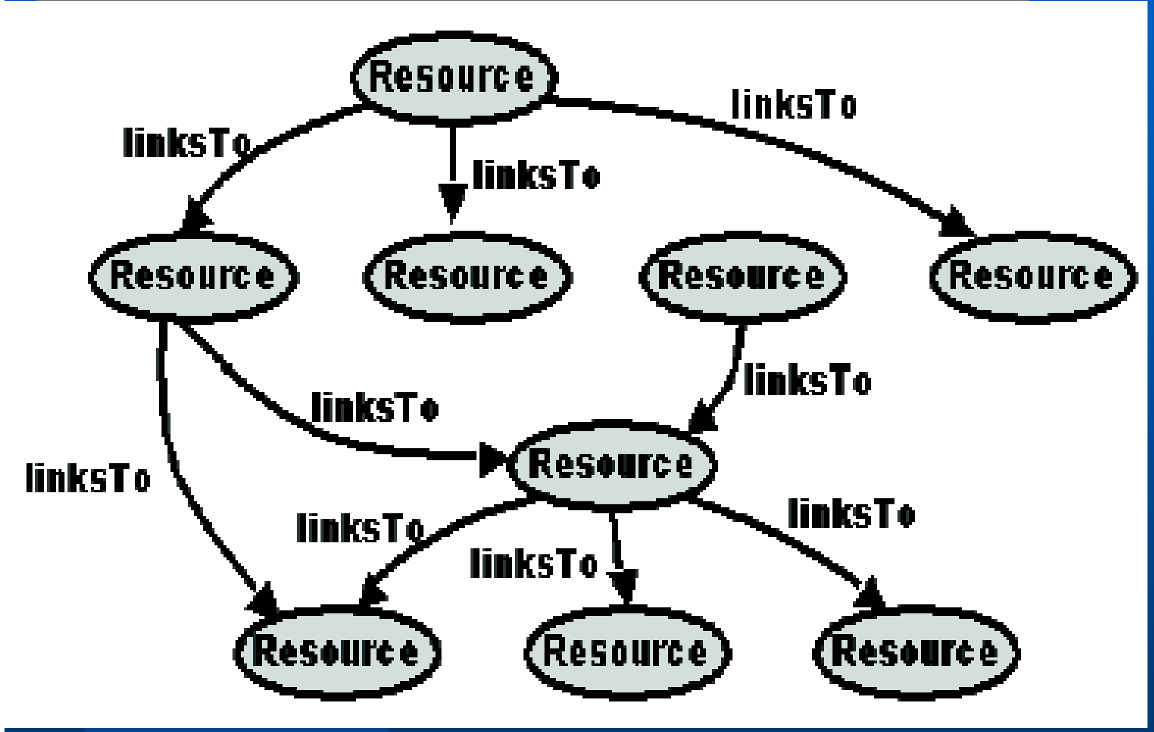
- **An Open Scientific Forum for Discussing, Capturing, and Showcasing Best Practices**
- **Draft Charter open for comment: http://www.w3.org/2005/05/swlsig-charter**
- **To be formalized in the next few weeks**
- **First formal meeting planned for Dec 13-14**
- **SW Supporting Vendors: Oracle, IBM, HP**
- **Recent life science members: Pfizer, Merck, caBIG/NCI, TeraNode, Partners HealthCare**
- **Please join us! http://www.w3.org/2005/04/swls/**

HCLSIG First Set of Activities

- **Core and Bridge Ontology work**
- **LSID refinement for SW**
- **Collect and Showcase Best Practices**
- **Construct powerful demonstrations in key areas**
- **Help create Work Groups in identified areas**

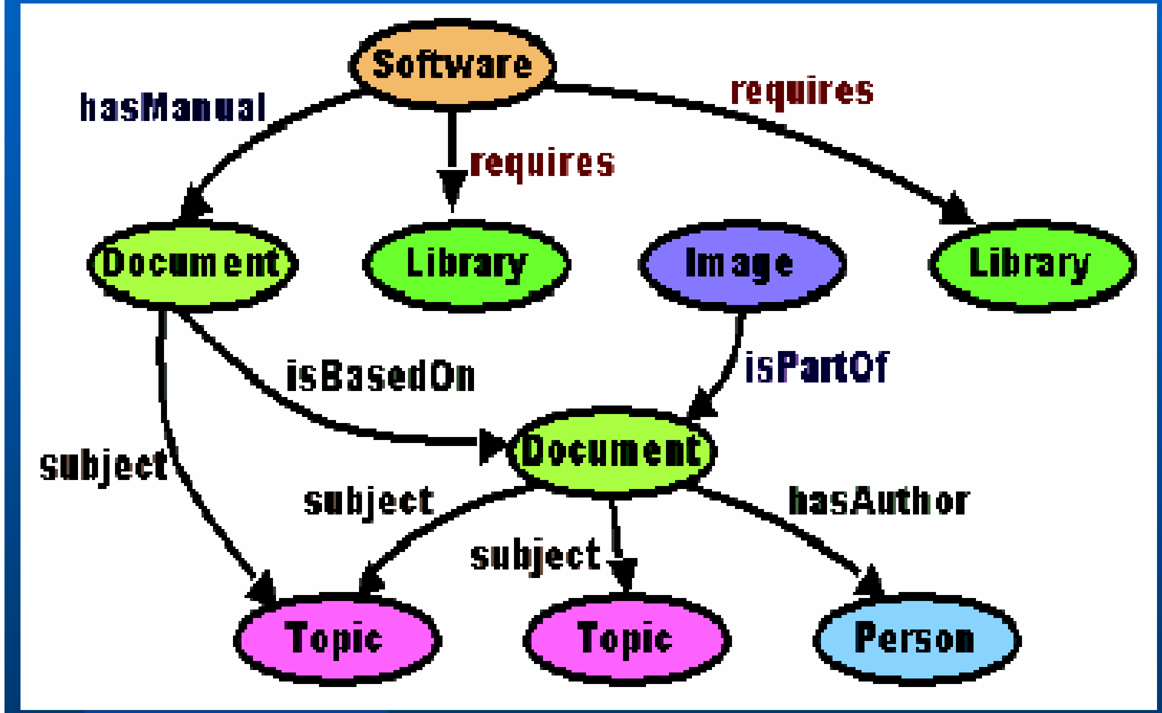
Semantic Web Model

The Current Web



- What the computer sees: “Dumb” links
- No semantics - <a href> treated just like <bold>
- Minimal machine-processable information

The Semantic Web



- Machine-processable semantic information
- Semantic context published – making the data more informative to both humans and machines

The Technologies: RDF

- **Think: "Relational Data Format"**
- **W3C standard for making statements of fact or belief**
- **Descriptive statements are expressed as triples: (Subject, Verb, Object)**
  - **We call verb a “predicate” or a “property”**

RDF Example (Gene)

Subject

Verb

Object

RDF

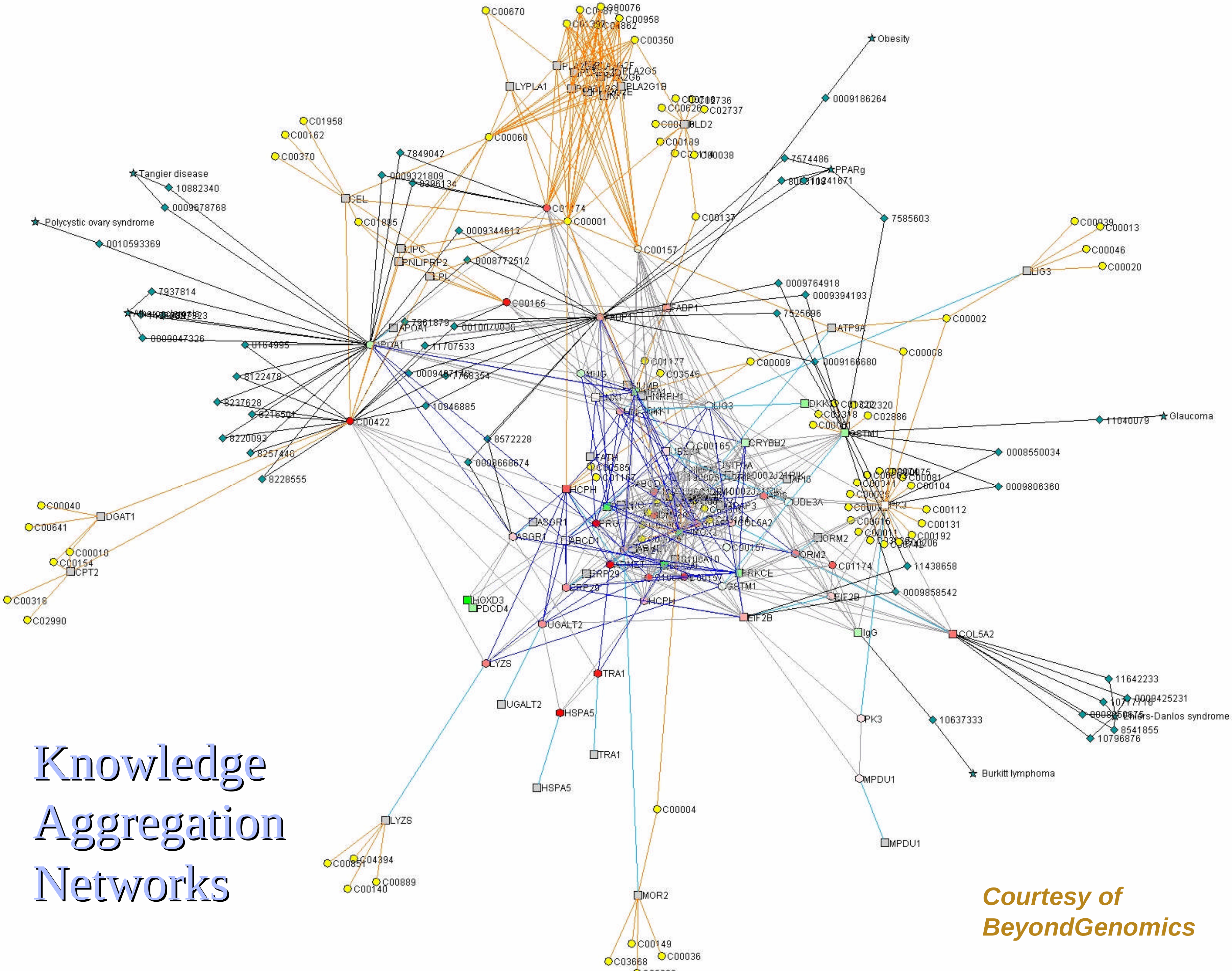
```
<Gene
    rdf:about="http://www.w3.org/2005/04/swls/gsk3b/gsk
    #GSK3b">
    <rdfs:label>Glycogen Synthase Kinase 3
    beta</rdfs:label>
  <translatedAs>
    rdf:resource="http://www.w3.org/2005/04/swls/gsk3b/g
    sk#GSK3betaProt"/>
    <dc:source>http://www.ncbi.nlm.nih.gov/entrez/query.
    fcgi?db=gene&#38;cmd=Retrieve&#38;dopt=Graphics
    &#38;list_uids=2932</dc:source>
  <exonSet rdf:parseType="Resource">
      <rdf:first
    rdf:resource="http://www.w3.org/2005/04/swls/gsk3b/g
    sk#ex1"/>
  </exonSet>
  <genomeLoc>14q3.2</genomeLoc>
  <hasAnnotation>Single locus </hasAnnotation>
  <hasTranscriptVariant
    rdf:resource="http://www.w3.org/2005/04/swls/gsk3b/g
    sk#GSK3betaSV"/>
  <isImplicatedIn
    rdf:resource="http://www.w3.org/2005/04/swls/gsk3b/g
    sk#Alzheimers"/>
  <isImplicatedIn
    rdf:resource="http://www.w3.org/2005/04/swls/gsk3b/g
    sk#DiabetesType2"/>

</Gene>
```

=

N3

```
gsk:GSK3b
    rdf:type  :Gene ;
    rdfs:label “Glycogen Synthase Kinase 3 beta” ;
    :translatedAs gsk:GSK3betaProt ;
    dc:source
    “http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=ge
    ne&#38;cmd=Retrieve&#38;dopt=Graphics&#38;list_ui
    ds=2932” ;
   :exonSet @( gsk:ex1 ) ;
   :genomeLoc “14q3.2” ;
    :hasAnnotation “Single locus”
    :hasTranscriptVariant gsk:GSK3betaSV ;
    :isImplicatedIn gsk:Alzheimers ;
    :isImplicatedIn gsk:DiabetesType2 .
```

N3 form is isomorphic with RDF, but more readable



Knowledge Aggregation Networks

Courtesy of BeyondGenomics

Statements Structure

ApoA1 …

… is produced by the Liver

… is expressed less in Atherosclerotic Liver

… is correlated with DKK1

… is cited regarding Tangier's disease

… has Tx Reg elements like HNFR1

Subject → Verb → Object

RDF-XML

RDF vs. XML example

Wang et al., Nature Biotechnology, Sept 2005

AGML

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

HUPML

RDF Stripe Mode

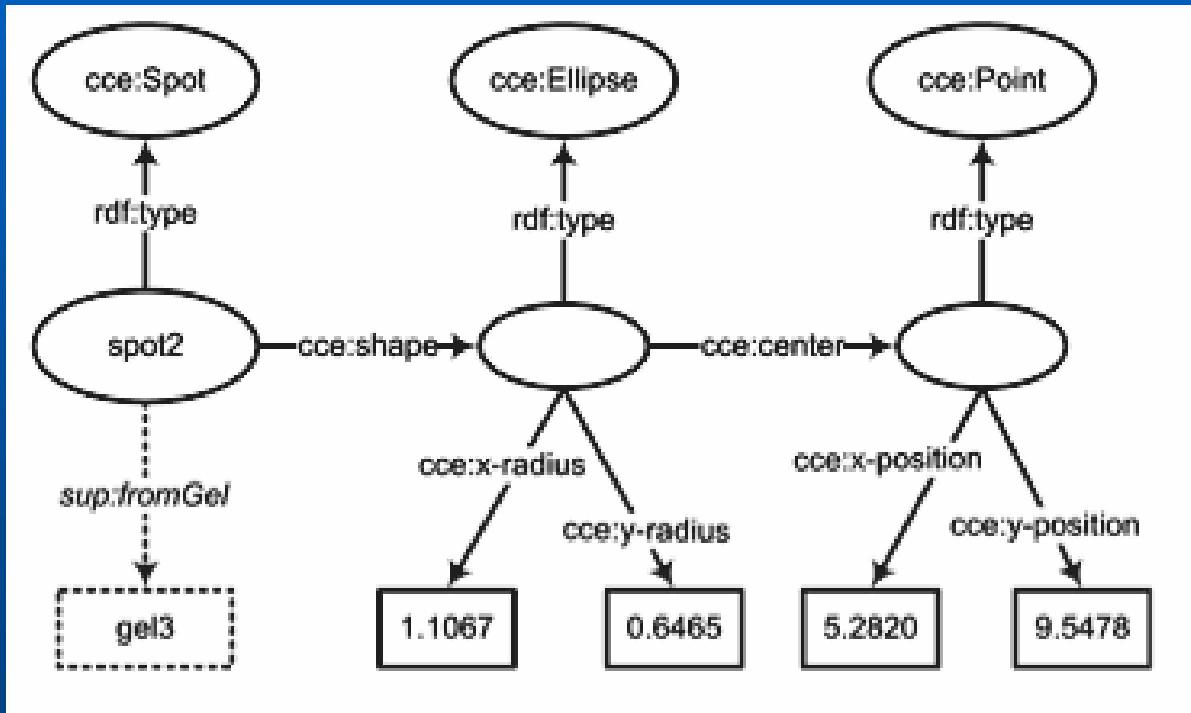
Document 1:

```
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:cce="http://www.charlestoncore.org/ontology/example#">
[redacted]
     <cce:shape>
[redacted]
         <cce:x-radius>1.1067</cce:x-radius>
         <cce:y-radius>0.6465</cce:y-radius>
         <cce:center>
[redacted]
             <cce:x-position>5.2820</cce:x-position>
             <cce:y-position>9.5478</cce:y-position>
           </cce:Point>
         </cce:center>
       </cce:Ellipse>
      </cce:shape>
   </cce:Spot>
</rdf:RDF>
```

Document 2:

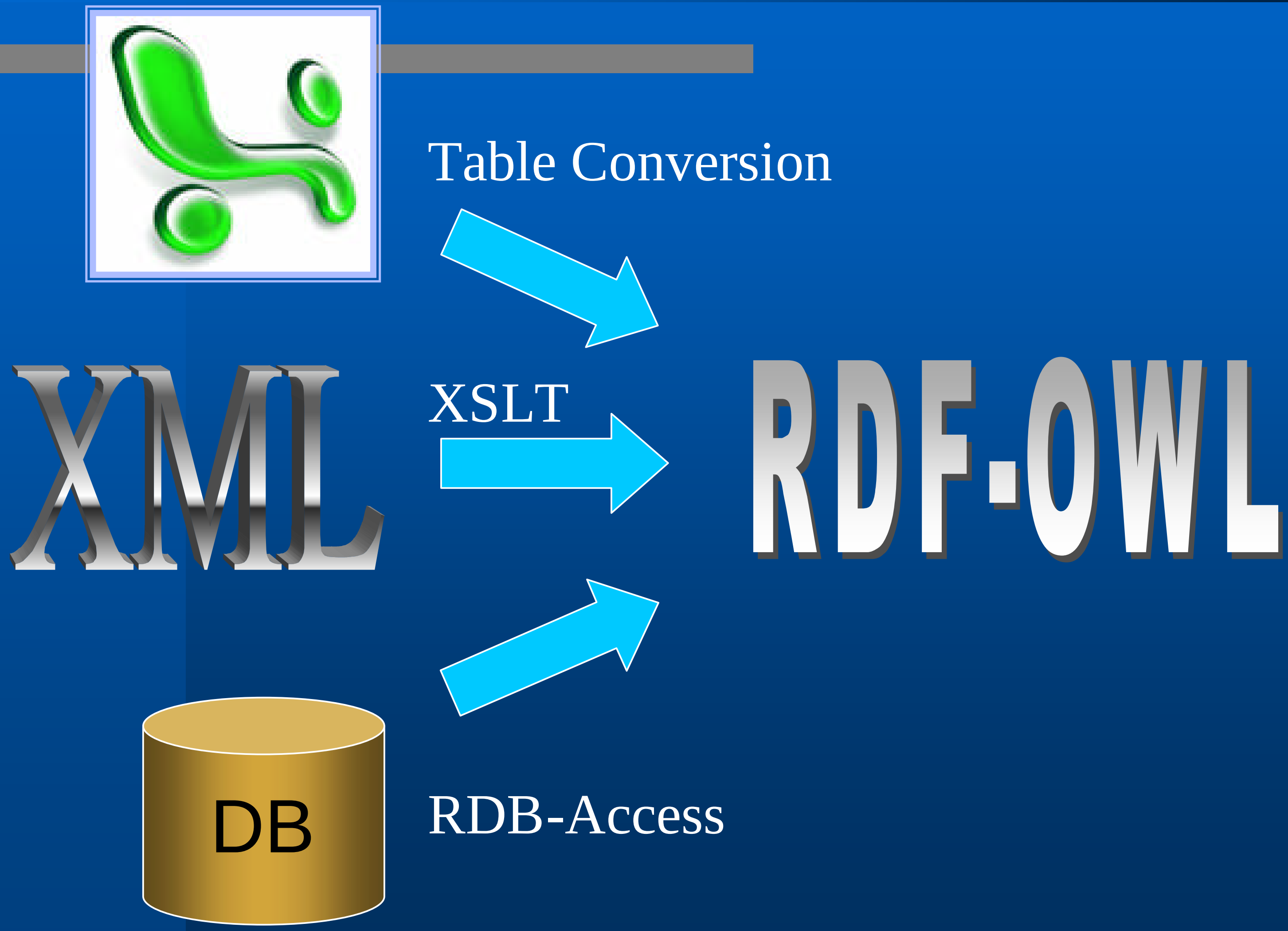
```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:cce="http://www.charlestoncore.org/ontology/example#"
  xmlns:sup="http://www.charlestoncore.org/ontology/supplement#">
[redacted]
     <sup:virtualGel
[redacted]
   </cce:Spot>
</rdf:RDF>
```

RDF Graph

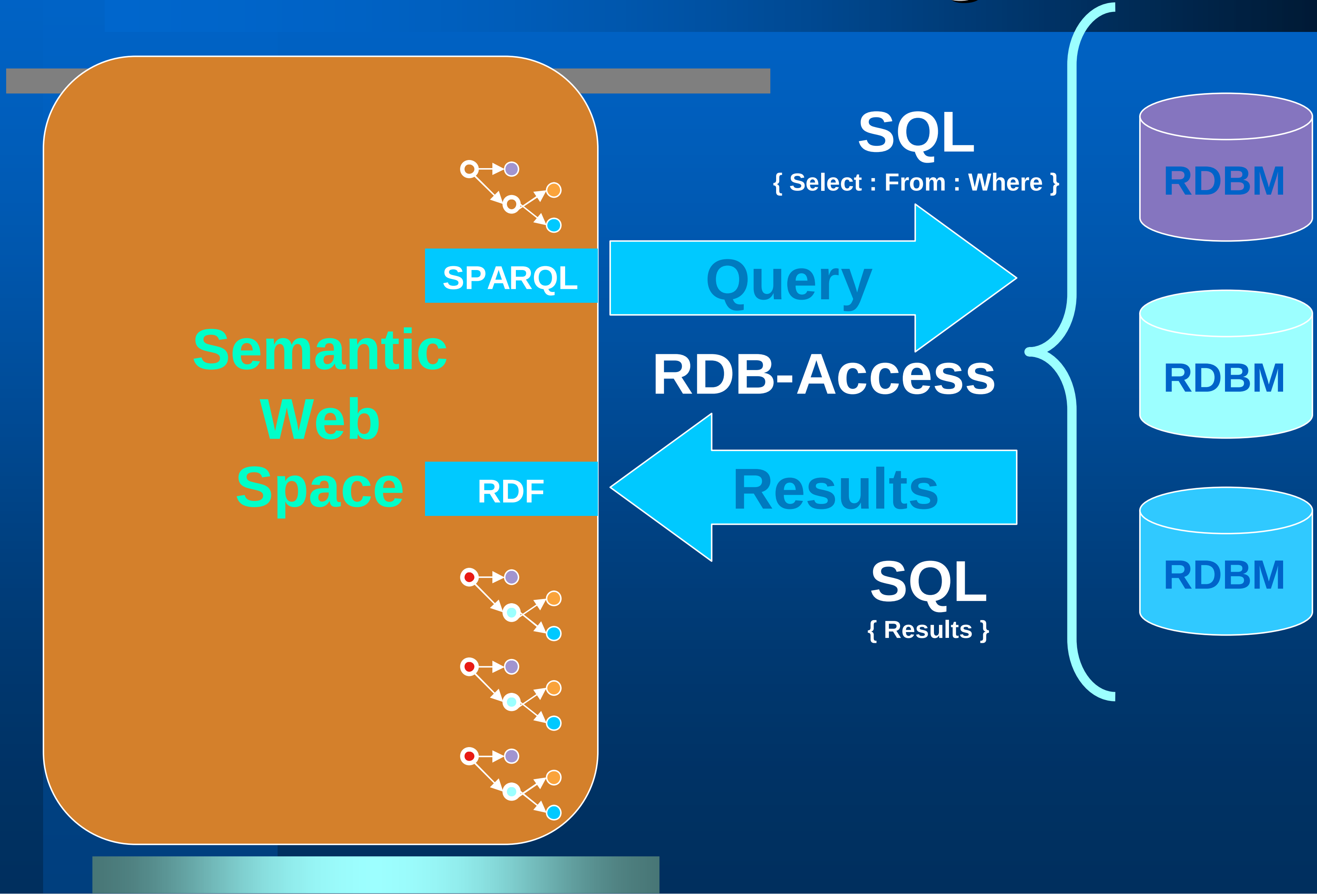


cce:Spot
cce:Ellipse
cce:Point
rdf:type
rdf:type
rdf:type
spot2
cce:shape
cce:center
cce:x-radius
cce:y-radius
cce:x-position
cce:y-position
sup:fromGel
gel3
1.1067
0.6465
5.2820
9.5478

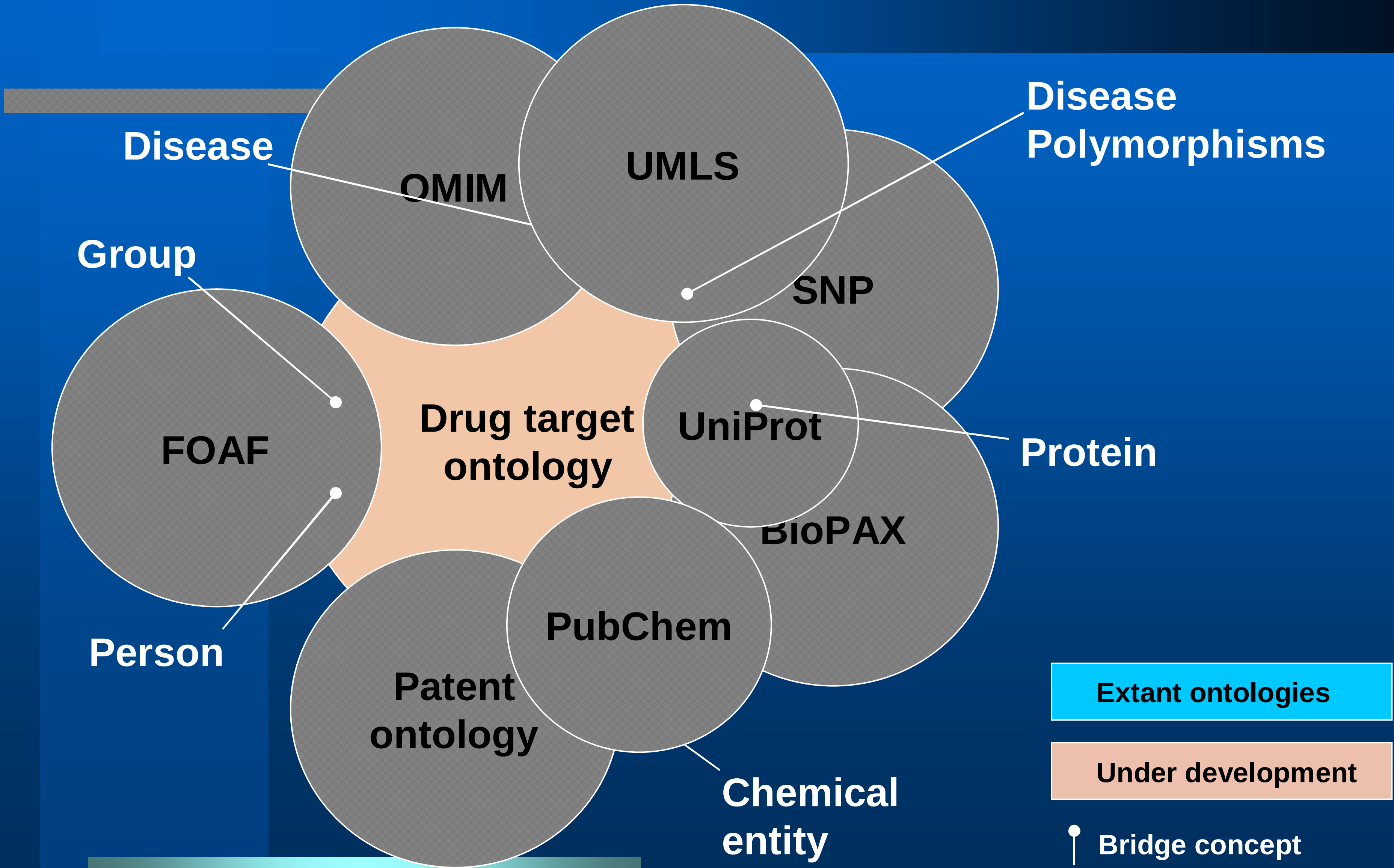
Mapping from Current Formats



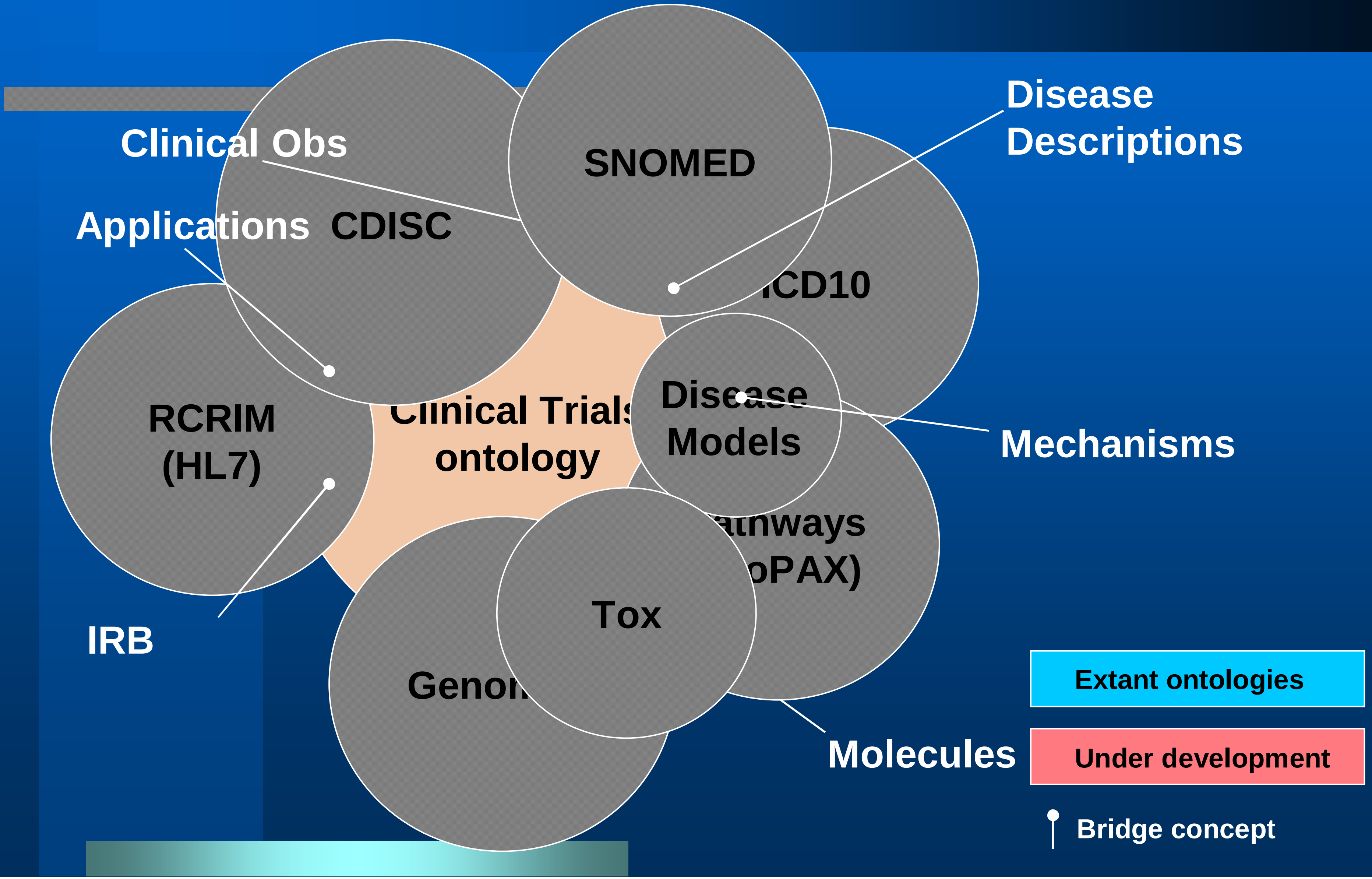
Semantic Federation using SPARQL



Multiple Ontologies Used Together



Potential Linked Clinical Ontologies



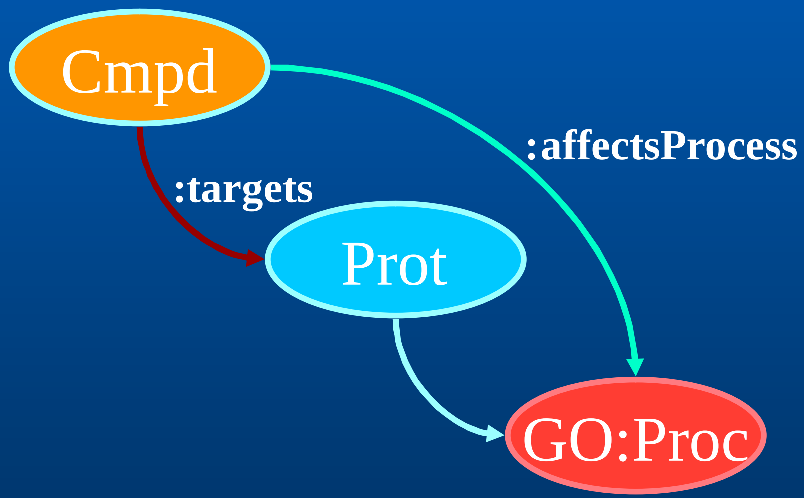
Applications of Ontologies

- Controlling vocabulary (ala GO)
- Controlling data types (concepts)
- Integrating data (instance serialization)
- Tagging of text to associate meta-data (publishing and search)
- Reasoning over aggregated information
- Web-service categorization

Bridging Domains via Rules

Projecting protein properties onto compounds

{:cmpd :targets ?prot .
?prot :bio-process ?proc}
➔
{:cmpd :affectsProcess ?proc}.

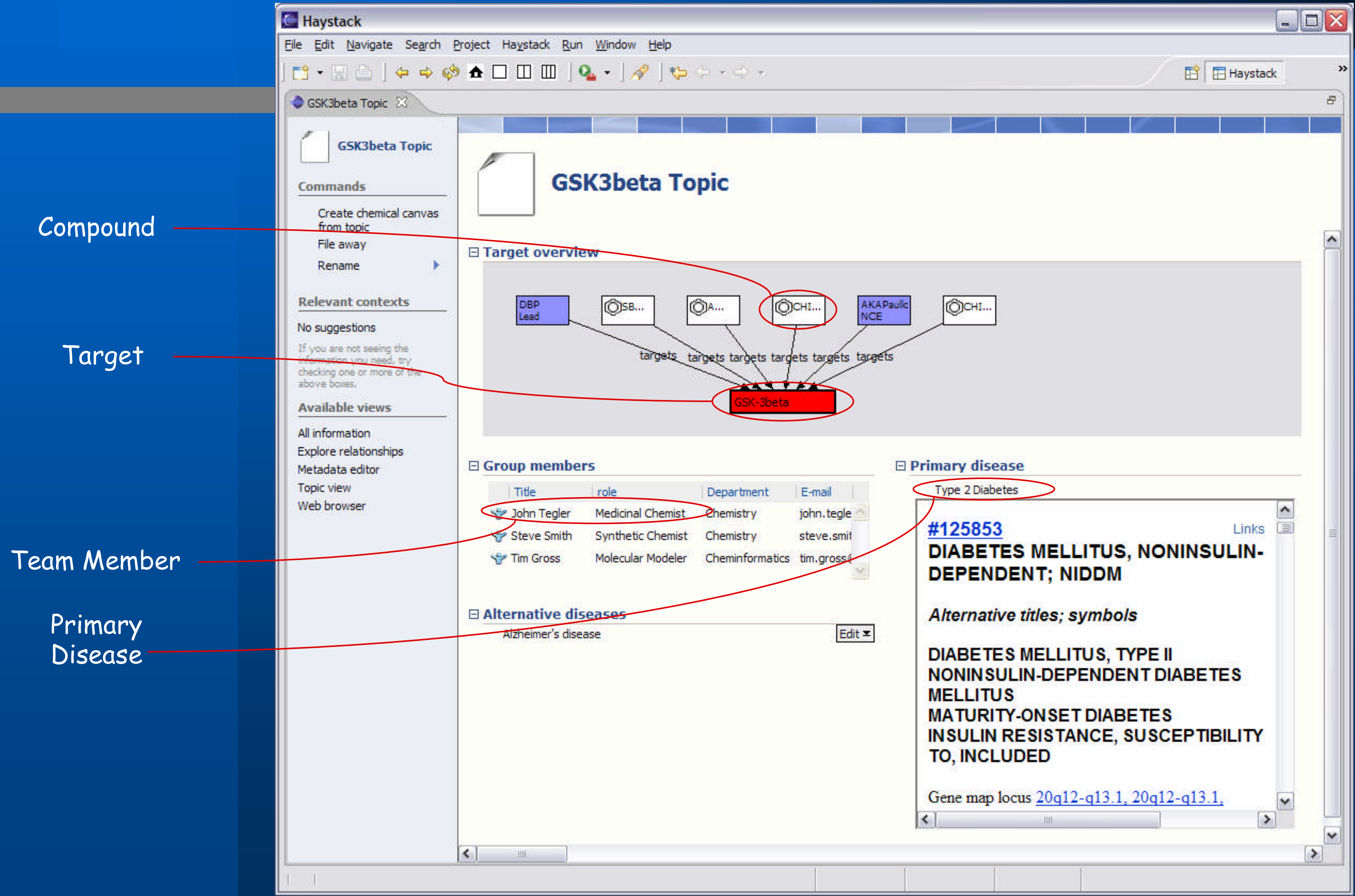


Semantic Browsers

- Next generation browsers extend viewing functionality to semantic information via “semantic lenses”
- Renders OWL-RDF, XML, and HTML documents
- Lenses act as info aggregators and style-sheets
- Most lenses require no Java, C, or perl programming; similar to how HTML can be rendered by any browser today
- New lenses can easily be created and added
- BioDash uses the Haystack browser from MIT

BioDASH Semantic Browser

http://www.w3.org/2005/04/swls/BioDash

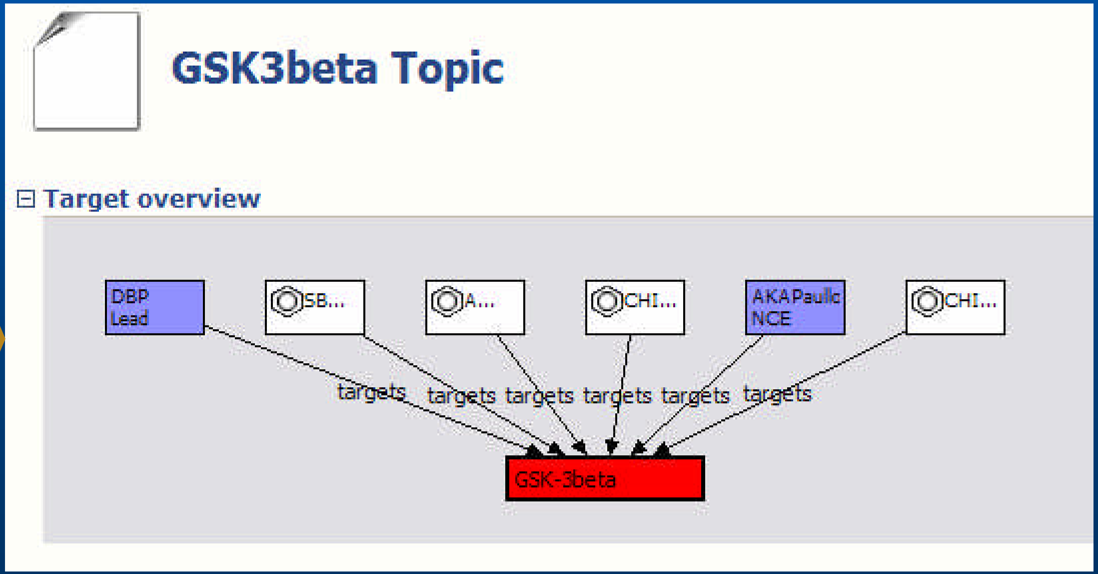


BioDASH Semantic Browser

All **viewed information is in RDF, which is then rendered according to "semantic lenses" to GUI components.**

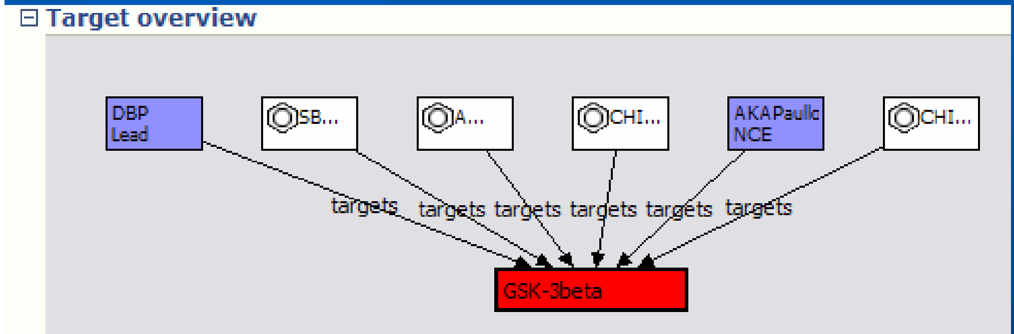
:Target gsk:GSK3beta
:rdfs:label "GSK-3beta" ;
:contextDisease gsk:DiabetesType2 ;
:targetFor gsk:AKAP ;
:targetFor gsk:ARA014418 ;
:targetFor gsk:CHIR98014 ;
:targetFor gsk:CHIR99021 ;
:targetFor gsk:DBP ;
:targetFor gsk:SB216763 ;
:inPathway gsk:Wnt ;
:references gsk:GSK3b ;
:xref
<urn:lsid:uniprot.org.lsid.biopathways.org:uniprot:P49841>

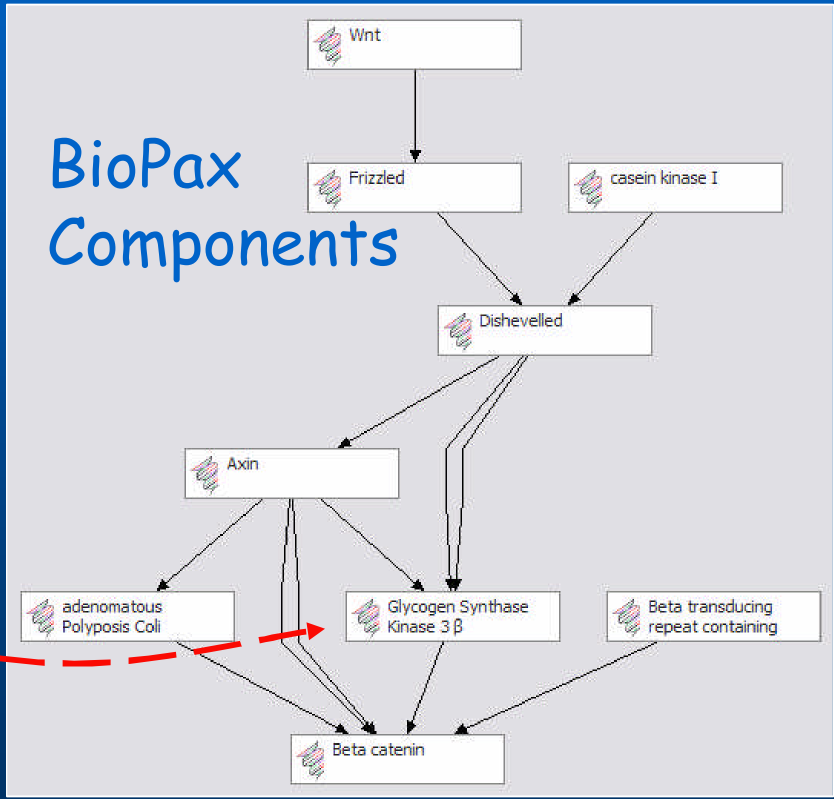
LENS



Bridging Chemistry and Molecular Biology

- Different Views have different semantics: Semantic Lenses
- When a correspondence between objects is determined, a semantic binding is made



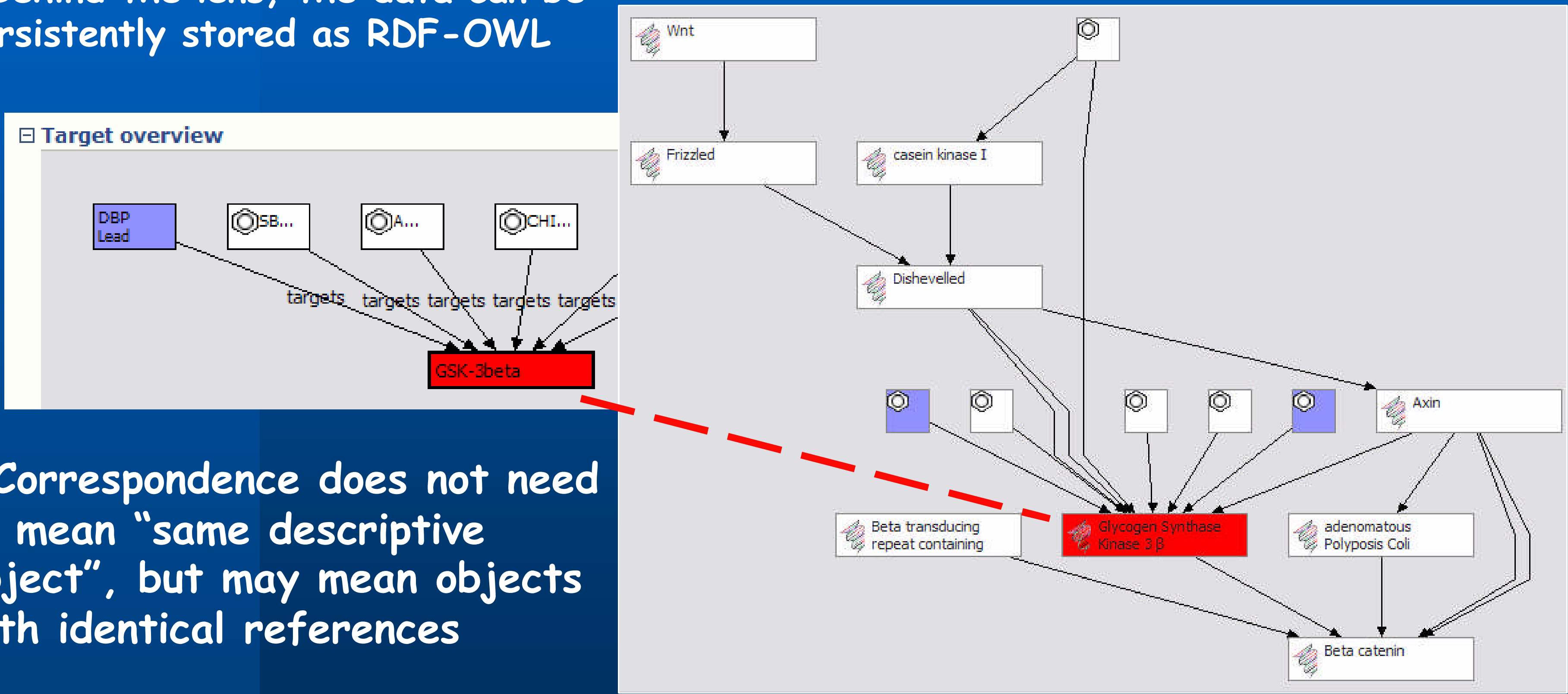


urn:lsid:uniprot.org:uniprot:P49841

Apply Correspondence Rule:
if ?target.xref.lsid == ?bpx:prot.xref.lsid
then ?target.correspondsTo.?bpx:prot

Bridging Chemistry and Molecular Biology

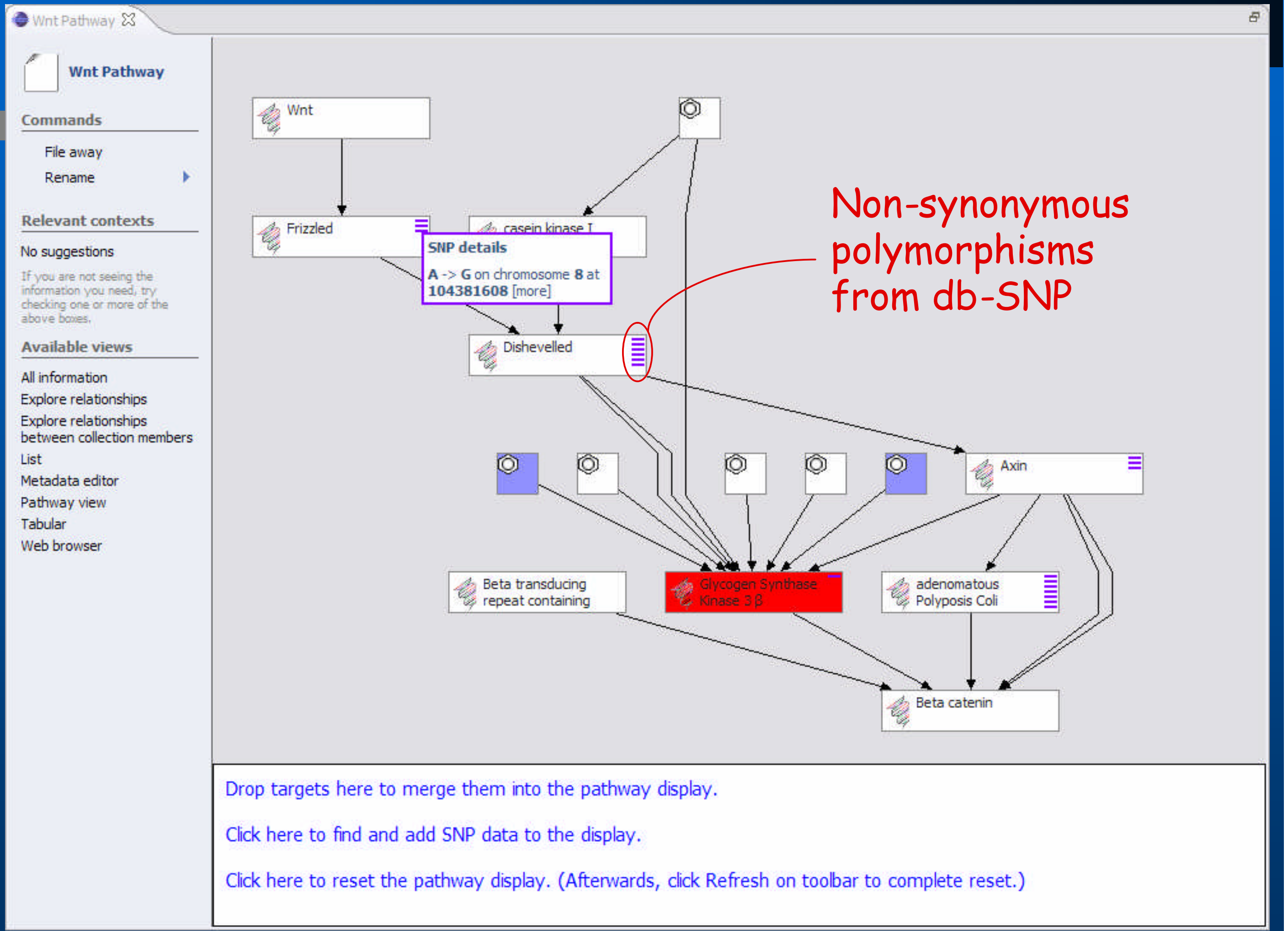
- Lenses can aggregate, accentuate, or even analyze new result sets
- Behind the lens, the data can be persistently stored as RDF-OWL



- Correspondence does not need to mean "same descriptive object", but may mean objects with identical references

Pathway Polymorphisms

- Merge directly onto pathway graph
- Identify targets with lowest chance of genetic variance
- Predict parts of pathways with highest functional variability
- Map genetic influence to potential pathway elements
- Select mechanisms of action that are minimally impacted by polymorphisms



Pathway Semantic Lens example

```
add { :predicateSet
        rdf:type graph:PredicateSet ;
        dc:title "BioPAX pathway arrows" ;
        hs:member    biopax:NEXT-STEP ;
        hs:member    :pointingTo ;
        hs:member ${
            rdf:type     vowl:RDFQueryLens ;
            vowl:sourceExistential    ?s ;
            vowl:targetExistential    ?t ;
            rdfs:label                "" ;
            vowl:existentials     @( ?s ?t ?type ) ;
            vowl:statement ${
                vowl:subject ?type ;
                vowl:predicate   biopax:LEFT ;
                vowl:object      ?s
            } ;
            vowl:statement ${
                vowl:subject ?type ;
                vowl:predicate   biopax:RIGHT ;
                vowl:object      ?t
            }
        }
}
```

S emantic lenses are defined using RDF (or N3) as well !

Microarray Data* ala Semantic Web

**or any other kind of tabular data*

```
:diabetes-hepatocyte-dataset
        rdf:typels:MicroArrayExpt ;

        dc:title"GSK3beta Expression Study" ;
        dc:date "4/15/04" ;
        ls:experimentalist "David Brucker" ;
        ls:targetSystem "mHep-R1 Hepatocytes" ;
        ls:design :GSK3b_mus_Protocol ;

ls:valueTypes @( ls:GE_Expected_Ratio ) ;

ls:indivCell ${ rdf:type ls:GE_Cell; ls:probeHub gsk:CaseinK;
        ls:conditionHub :GSK3b_RNAi_perturb; ls:GE_Expected_Ratio "0.909"; ls:GE_Variance "0.007" } ;

 ls:indivCell ${ rdf:type ls:GE_Cell; ls:probeHub gsk:DVL;
        ls:conditionHub :DBAP_perturb; ls:GE_Expected_Ratio "1.055"; ls:GE_Variance "0.007" } ;

 ls:indivCell ${ rdf:type ls:GE_Cell; ls:probeHub gsk:Axin;
        ls:conditionHub :SB216763_perturb; ls:GE_Expected_Ratio "0.881"; ls:GE_Variance "0.007" } ;
```

Any valid combination of qualia and quantities can be described for each Data Point

Gene Expression Studies

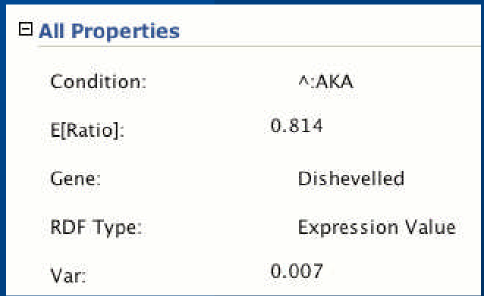
- Sets of Genes and Conditions linked to each data cell via RDF properties
- Hierarchical Clustering
- Semantic links to each object

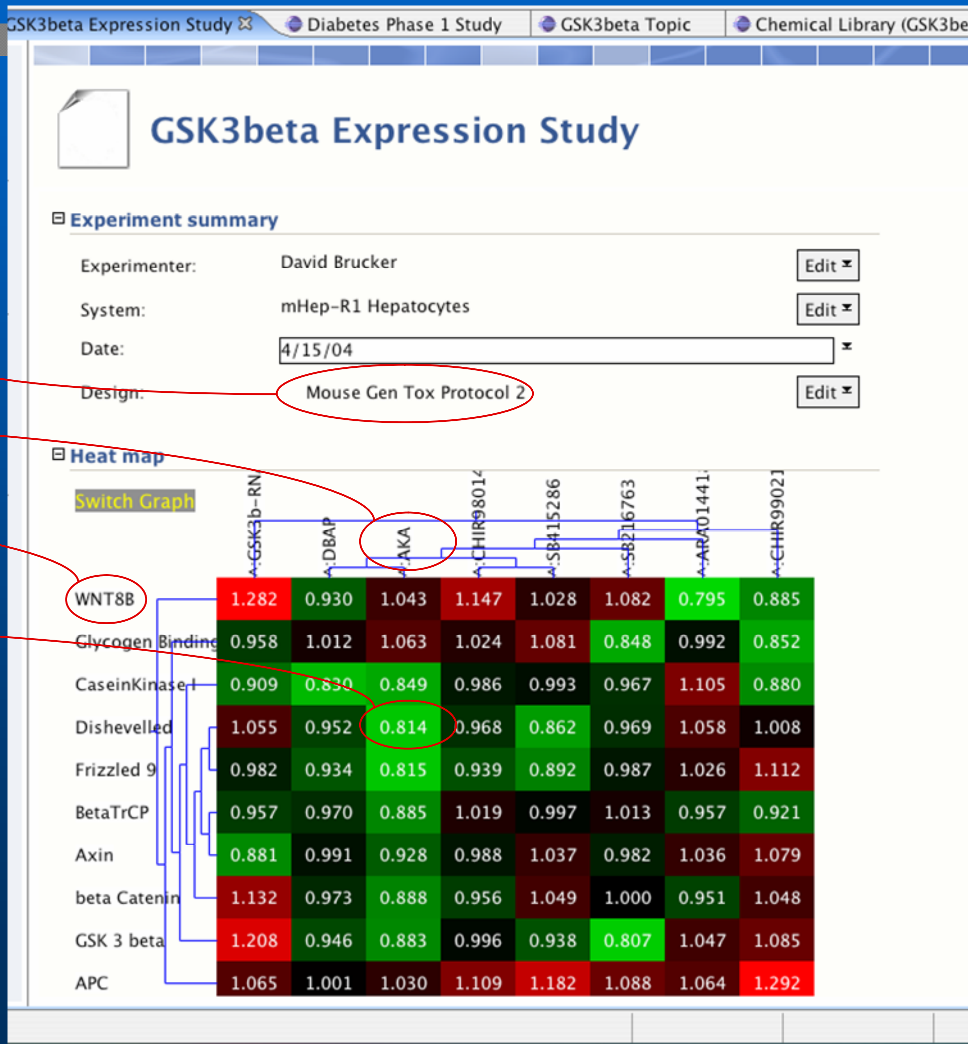
Protocol Object

Condition Hub

Gene (Probe) Hub

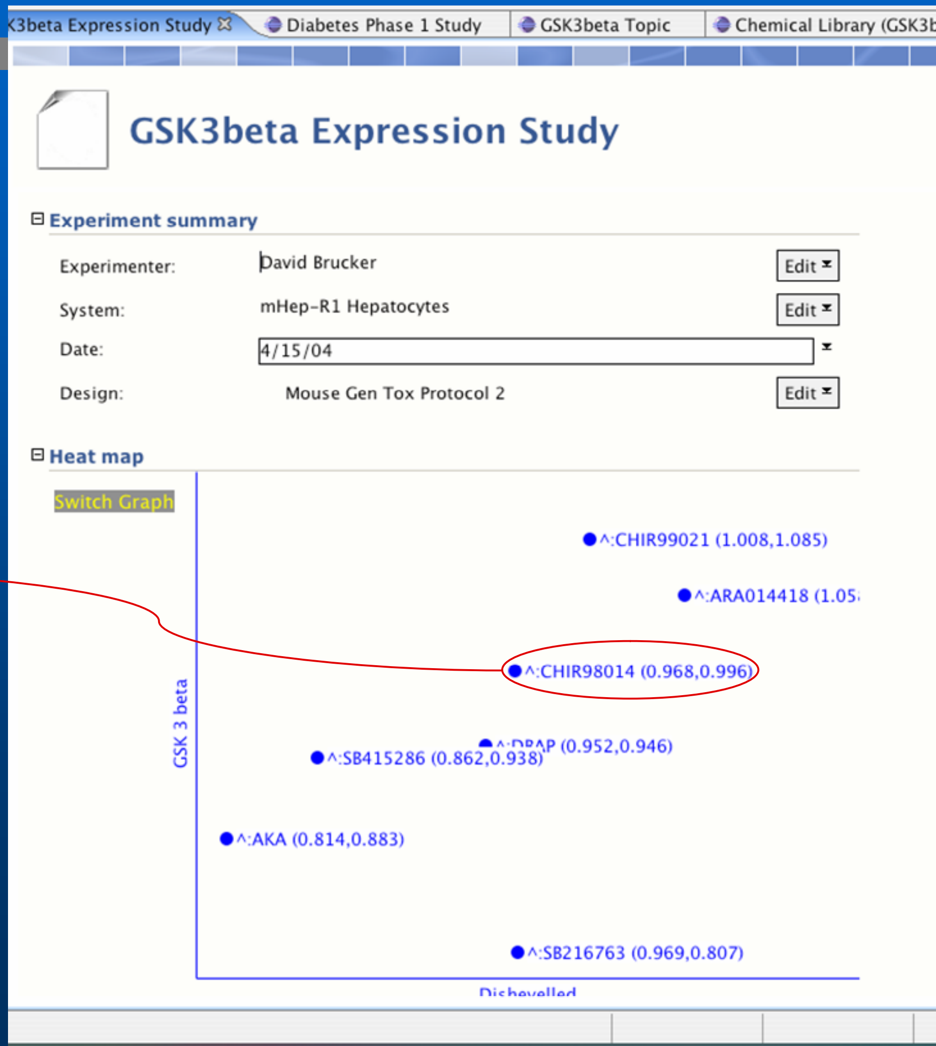
GE Node





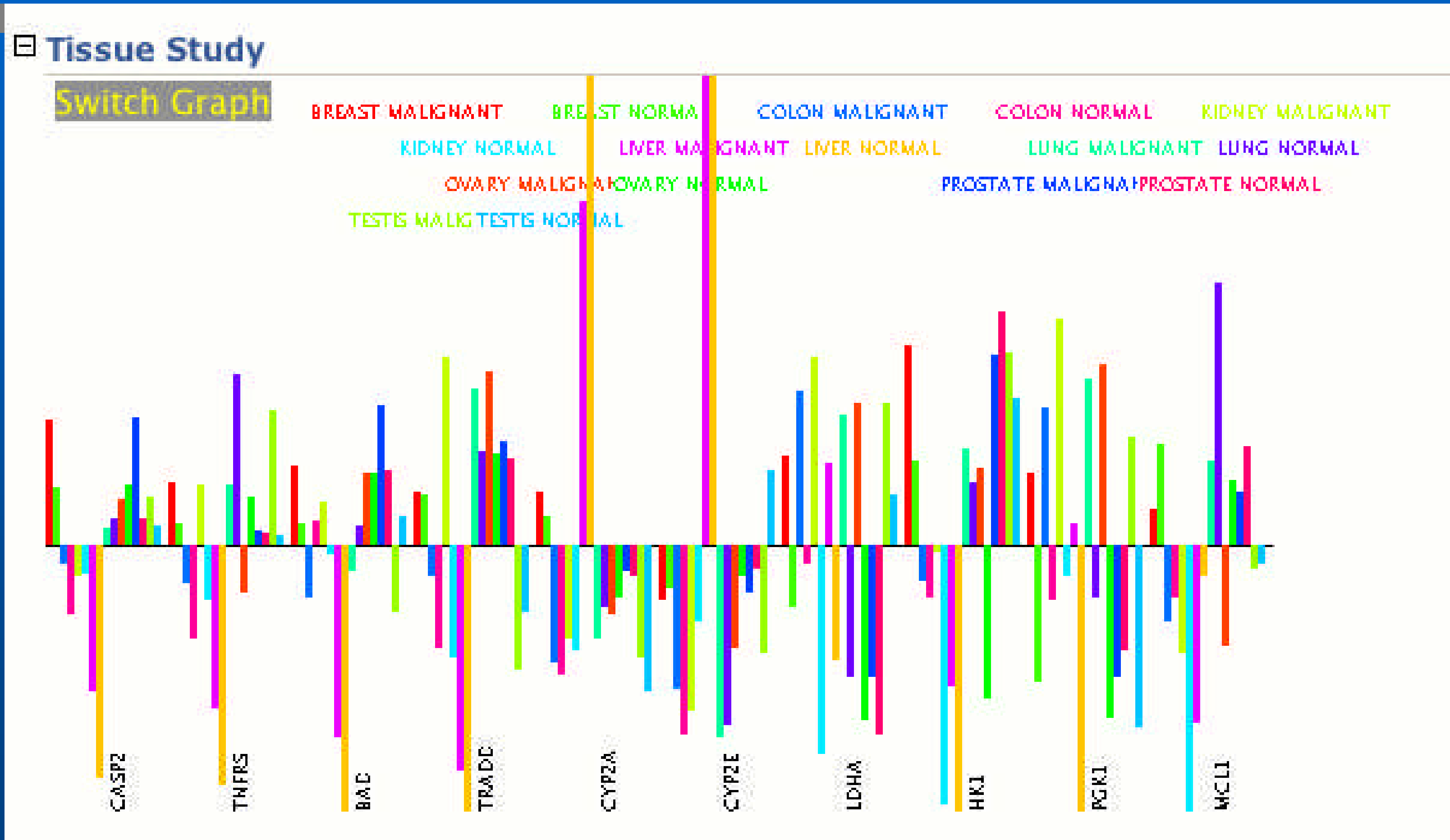
Semantic Graphs

- The full set of probes or conditions is captured
- Data points can contain a wealth of semantic information



Each Data Point is a Semantic Hub

Alternative Views



Preclinical and Clinical Applications
- *in support of the* ***Critical Path***

Critical Path

Support Innovation and Safety through-

- **Capturing all information more carefully (semantically)**
- **Biomarkers, PGx, Pathways, drug profiling, etc.**
- **More Knowledge re-use**
  - *Lessons Learned: Failures + Successes*
- **Better coordination between groups**
- **Using better tools for Decision Making**
- **Paper trails**
- **Semantic Lenses**

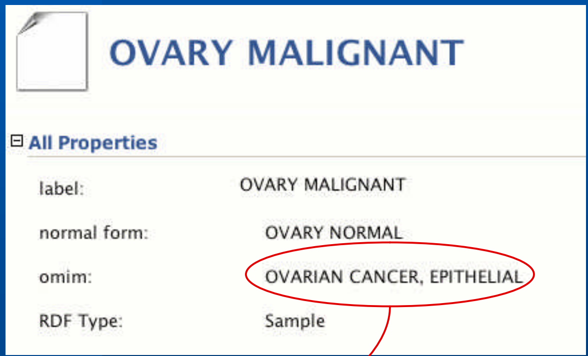
Supporting Drug Safety
'Safety Lenses'

- Semantic Lenses used to 'see' data in specific ways
- Can be "wrapped" around selected statistical tools
- Used by both reviewers and sponsors (different sets)
- View findings under different lenses depending on area of interest: hepatoxicity vs. genotoxicity
- Aggregate other sources and findings (*knowledge)* in context with a particular project
- Support special "Alert-channels" by regulators for each different toxicity issue
- Align animal studies with clinical results
- Easy access to any new published and validated mechanisms of actions

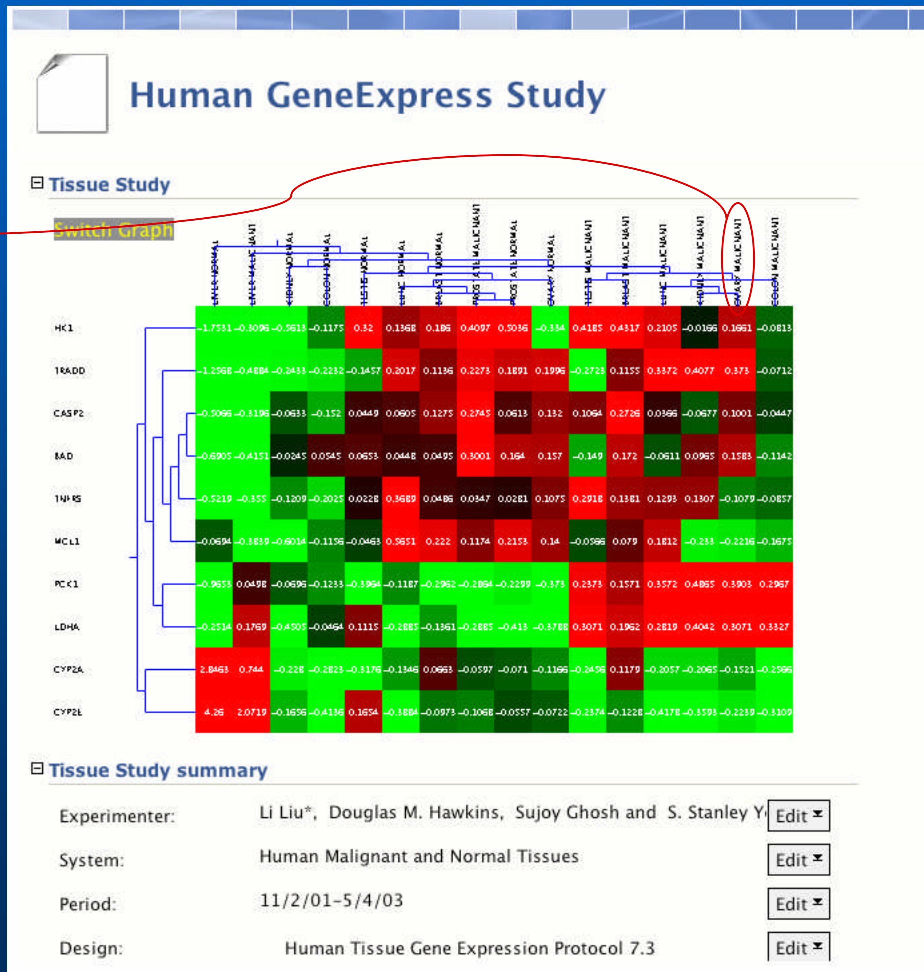
GeneLogic *GeneExpress* Data

- **Additional relations and aspects can be defined additionally**

Diseased Tissue

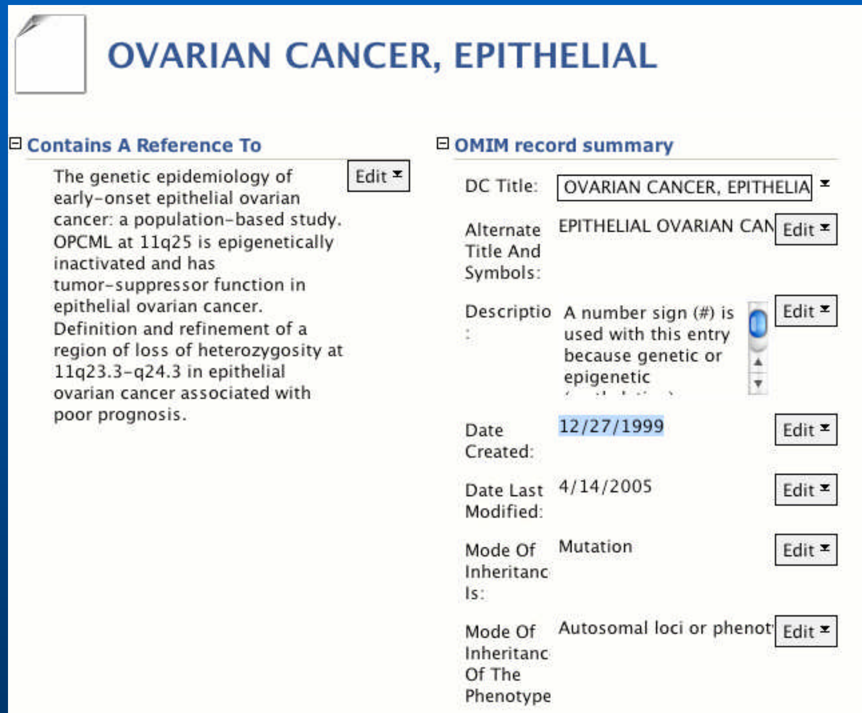


Links to OMIM (RDF)



Diseases Links

- All diseases linked dynamically to OMIM references
- OMIM is converted to RDF chunks, each queriable and annotatable.

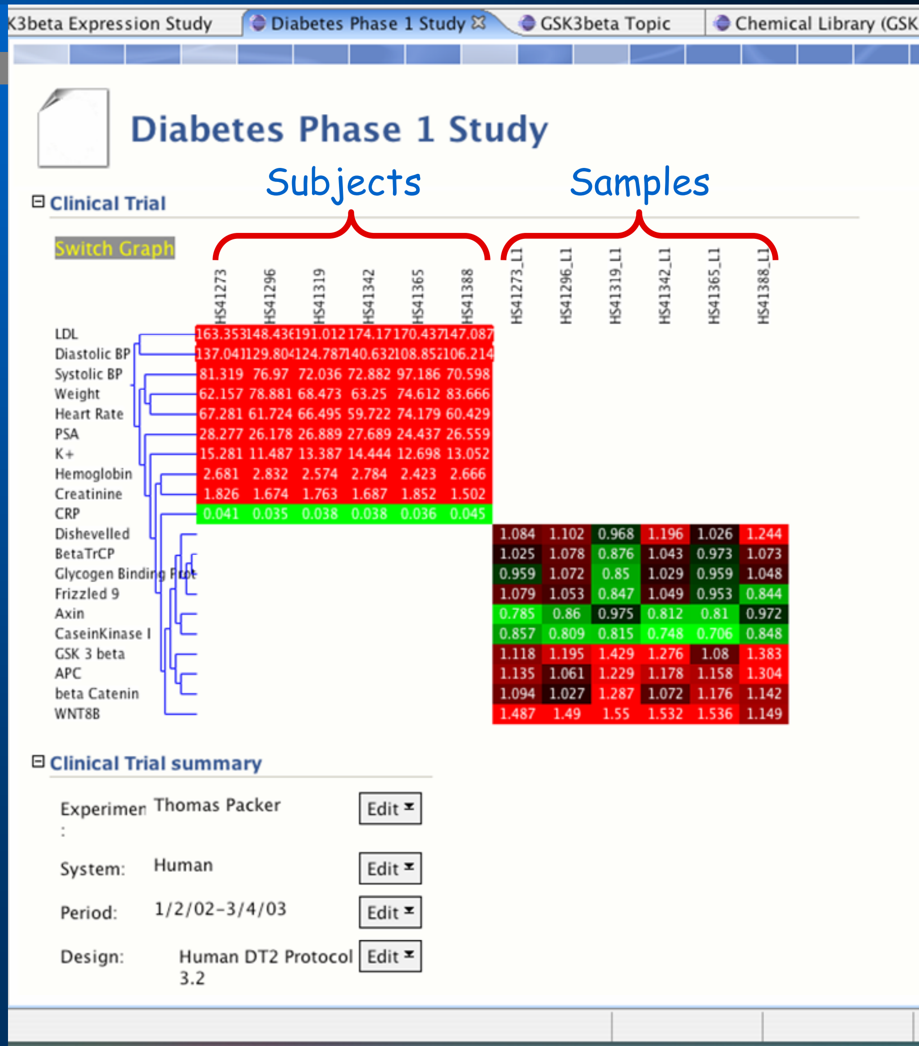


Clinical Trials

- Combine Clinical Observations with other measurements (e.g., gene expression)
- Semantic links between samples and subjects
- Semantically defined Clinical protocols

Clinical Obs

Expression Data



Clinical Trials

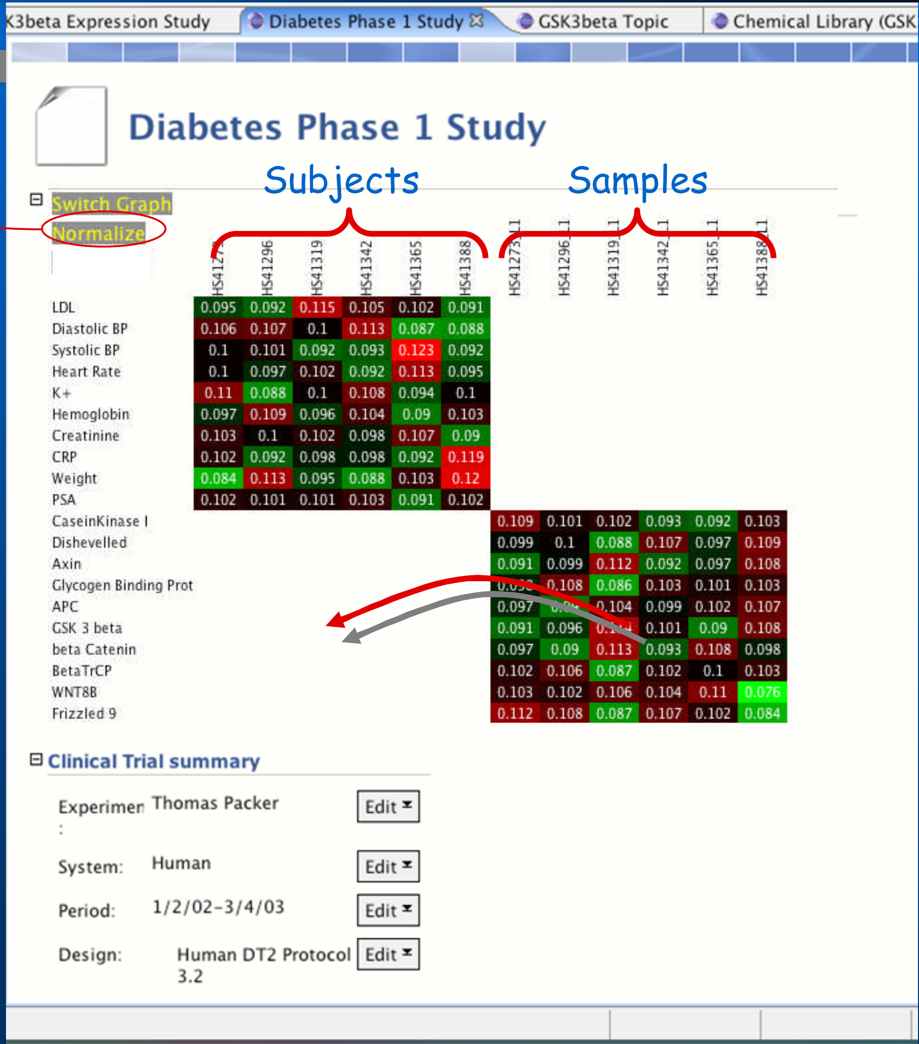
•Values can be normalized across all measurables (rows)

•Samples can be aligned to their subjects using RDF rules

Normalize

Clinical Obs

Expression Data



	HS4127	HS41296	HS41319	HS41342	HS41365	HS41388	HS41273_1	HS41296_1	HS41319_1	HS41342_1	HS41365_1	HS41388_1
LDL	0.095	0.092	0.115	0.105	0.102	0.091						
Diastolic BP	0.106	0.107	0.1	0.113	0.087	0.088						
Systolic BP	0.1	0.101	0.092	0.093	0.123	0.092						
Heart Rate	0.1	0.097	0.102	0.092	0.113	0.095						
K+	0.11	0.088	0.1	0.108	0.094	0.1						
Hemoglobin	0.097	0.109	0.096	0.104	0.09	0.103						
Creatinine	0.103	0.1	0.102	0.098	0.107	0.09						
CRP	0.102	0.092	0.098	0.098	0.092	0.119						
Weight	0.084	0.113	0.095	0.088	0.103	0.12						
PSA	0.102	0.101	0.101	0.103	0.091	0.102						
CaseinKinase I							0.109	0.101	0.102	0.093	0.092	0.103
Dishevelled							0.099	0.1	0.088	0.107	0.097	0.109
Axin							0.091	0.099	0.112	0.092	0.097	0.108
Glycogen Binding Prot							[illegible]	0.108	0.086	0.103	0.101	0.103
APC							0.097	[illegible]	0.104	0.099	0.102	0.107
GSK 3 beta							0.091	0.096	[illegible]	0.101	0.09	0.108
beta Catenin							0.097	0.09	0.113	0.093	0.108	0.098
BetaTrCP							0.102	0.106	0.087	0.102	0.1	0.103
WNT8B							0.103	0.102	0.106	0.104	0.11	0.076
Frizzled 9							0.112	0.108	0.087	0.107	0.102	0.084

Clinical Trials

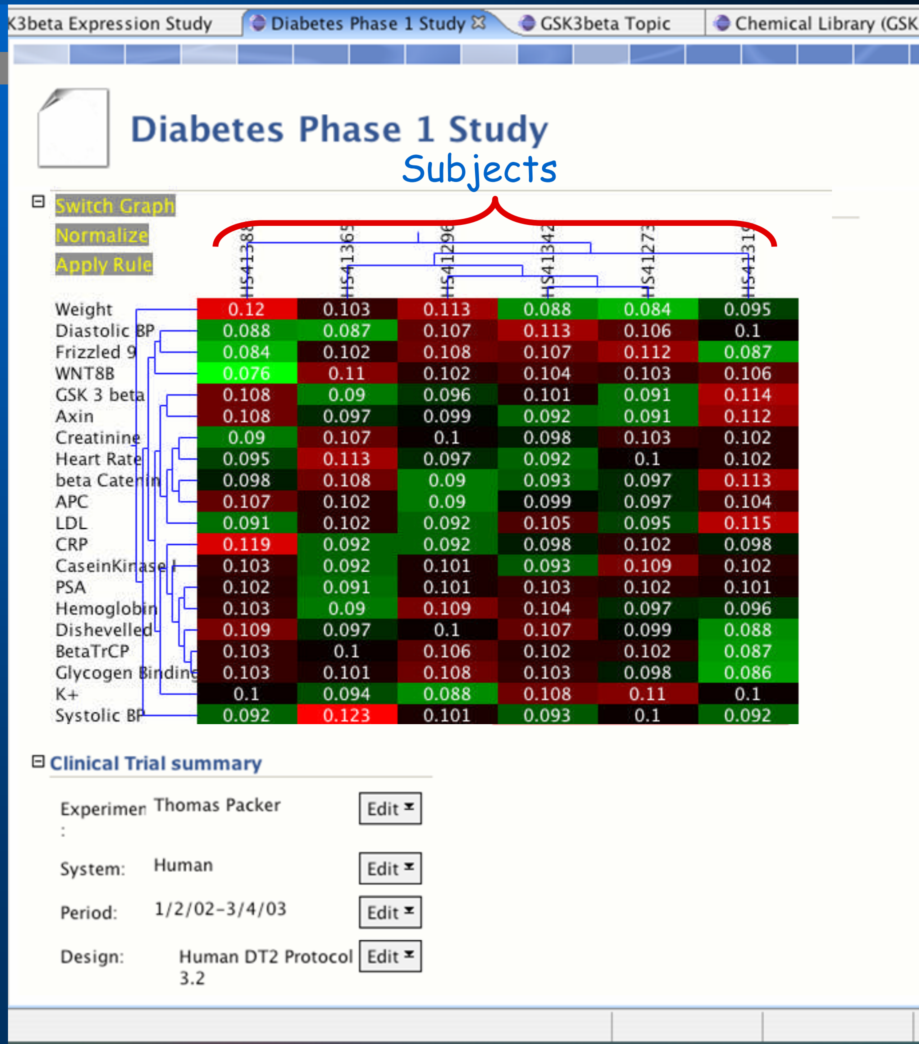
- Values can be normalized across all measurables (rows)

- Samples can be aligned to their subjects using RDF rules

- Clustering can now be done over all measureables (rows)

Clinical Obs

Expression Data



	HS41388	HS41365	HS41296	HS41342	HS41273	HS4131
Weight	0.12	0.103	0.113	0.088	0.084	0.095
Diastolic BP	0.088	0.087	0.107	0.113	0.106	0.1
Frizzled 9	0.084	0.102	0.108	0.107	0.112	0.087
WNT8B	0.076	0.11	0.102	0.104	0.103	0.106
GSK 3 beta	0.108	0.09	0.096	0.101	0.091	0.114
Axin	0.108	0.097	0.099	0.092	0.091	0.112
Creatinine	0.09	0.107	0.1	0.098	0.103	0.102
Heart Rate	0.095	0.113	0.097	0.092	0.1	0.102
beta Catenin	0.098	0.108	0.09	0.093	0.097	0.113
APC	0.107	0.102	0.09	0.099	0.097	0.104
LDL	0.091	0.102	0.092	0.105	0.095	0.115
CRP	0.119	0.092	0.092	0.098	0.102	0.098
CaseinKinase I	0.103	0.092	0.101	0.093	0.109	0.102
PSA	0.102	0.091	0.101	0.103	0.102	0.101
Hemoglobin	0.103	0.09	0.109	0.104	0.097	0.096
Dishevelled	0.109	0.097	0.1	0.107	0.099	0.088
BetaTrCP	0.103	0.1	0.106	0.102	0.102	0.087
Glycogen Binding	0.103	0.101	0.108	0.103	0.098	0.086
K+	0.1	0.094	0.088	0.108	0.11	0.1
Systolic BP	0.092	0.123	0.101	0.093	0.1	0.092

CDISC and the Semantic Web

- Reduce the need to write data parsers to any CDISC XML Schema
- Make use of ontologies and terminologies directly using RDF
- Easier inclusion of Genomic data
- Use Semantic Lenses for Reviewers
- Easier acceptance by industry with their current technologies

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

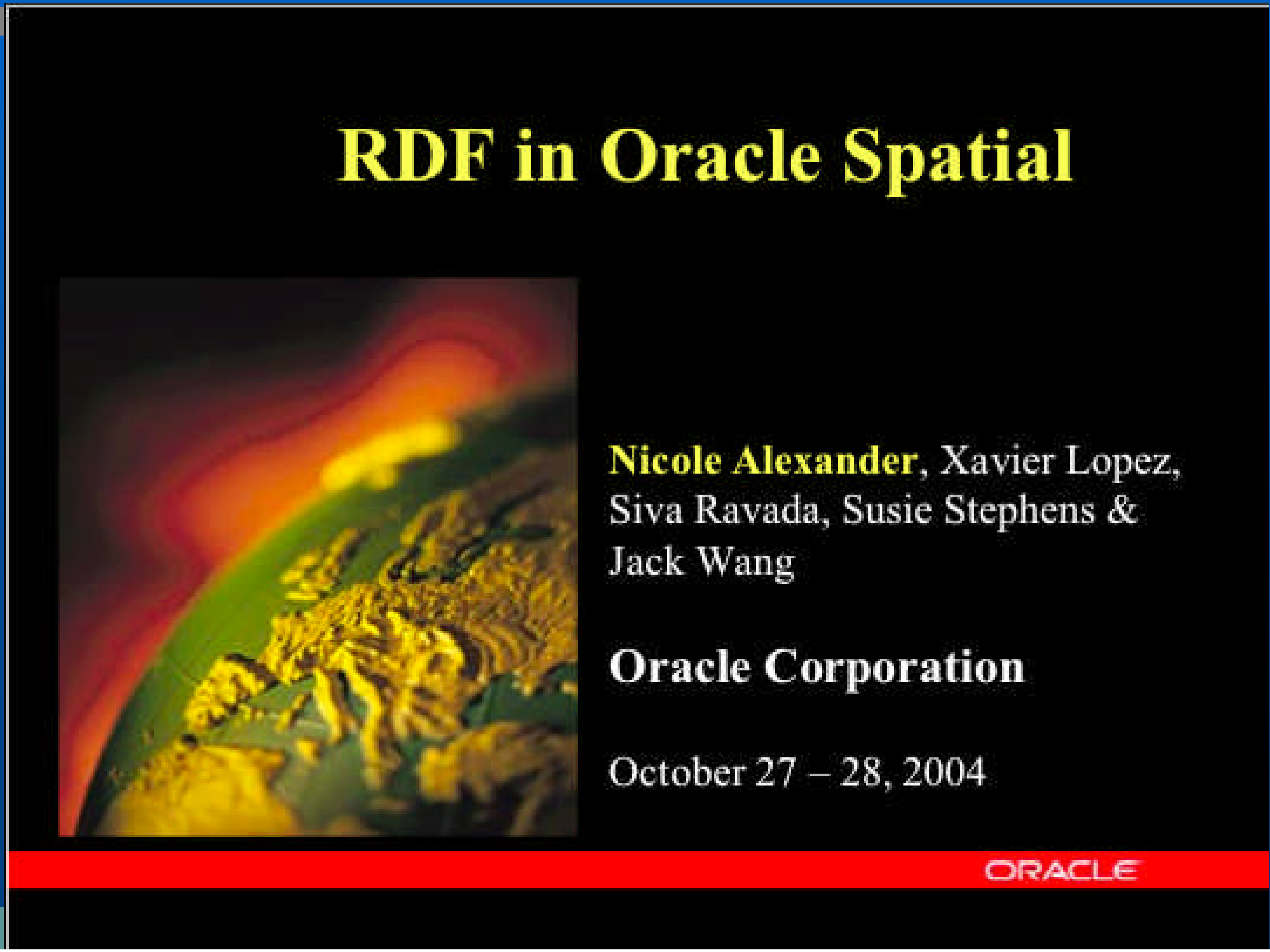
Merging Clinical Evidence with Interpretation and Models

- Semantic linking to the Narrative
  - Observations
  - Genomics
  - Biomarkers
  - PK/PD
  - Statistical Reasoning and Inference
- Justifications
  - Efficacy
  - Toxicity
  - Mechanism of Action
- Built into Application Submission
- Use of Semantic Lenses by reviewers to compare all evidence

How to get there?

- **Current Standards into SW format (OWL and RDF)**
  - **CDISC, RCRIM, MEDRA, UMLS, ICD10**
- **Use secure-web as infrastructure**
- **Semantic wrappers for Relational Databases (W3C, Oracle)**
- **Using SW to define and apply *policies* through rule-based semantics**

Vendor Adoption



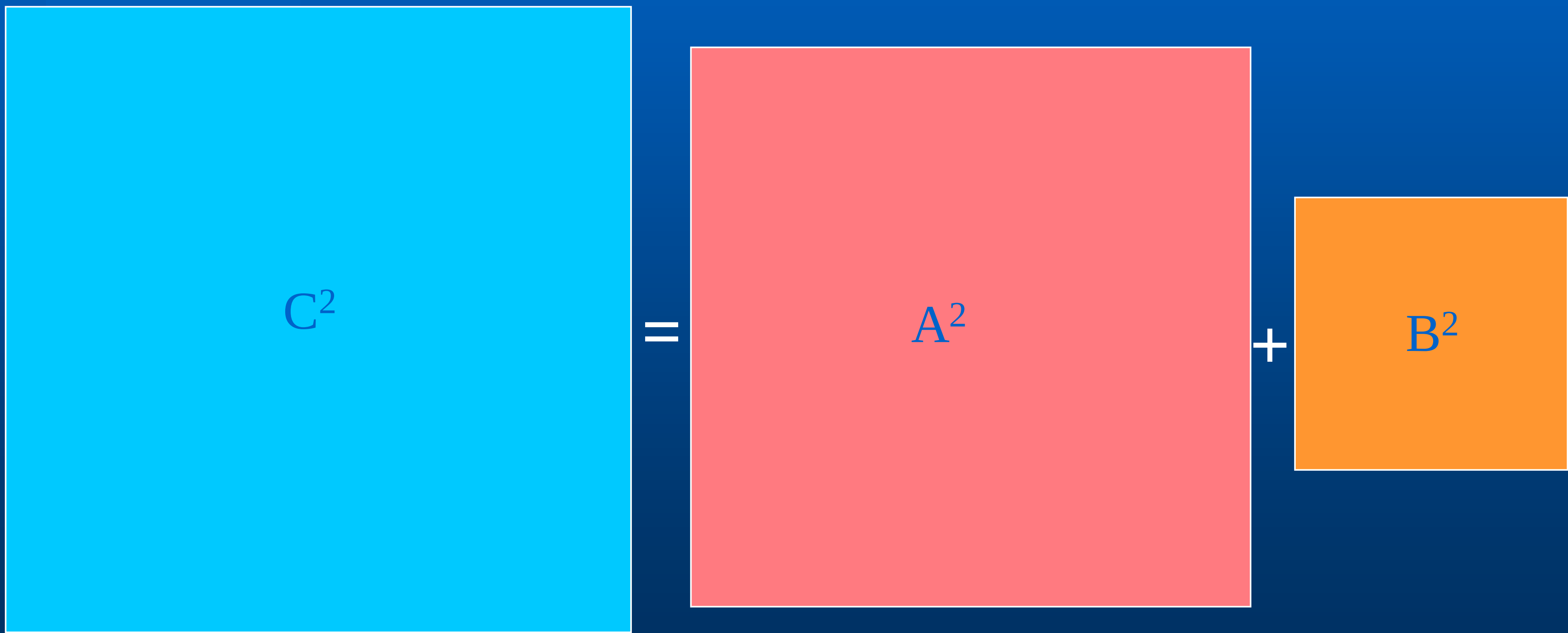
SW Public Domain Tools

- Haystack - http://haystack.lcs.mit.edu
- RDB-RDF-Access -
  - http://www.w3.org/2004/04/30-RDF-RDB-access/
  - http://www.w3.org/2004/10/04-pharmaFederate/
- XML-->RDF - http://www.w3.org/2005/02/13-KEGG/
- Protégé - http://protege.stanford.edu/
- SESAME - http://www.openrdf.org/
- JENA - http://www.hpl.hp.com/semweb/jena2.htm
- SWOOP - http://www.mindswap.org/2004/SWOOP/

Thank You

- Healthcare and Life Sciences Interest Group (HCLSIG)
  - –Tonya Hongsermeier, co-chair Partners Healthcare
  - – Eric Neumann, co-chair Clinical Semantics Group
- Eric Miller, Semantic Web Lead
- http://www.w3.org/2005/04/swls
- http://www.w3.org/sw

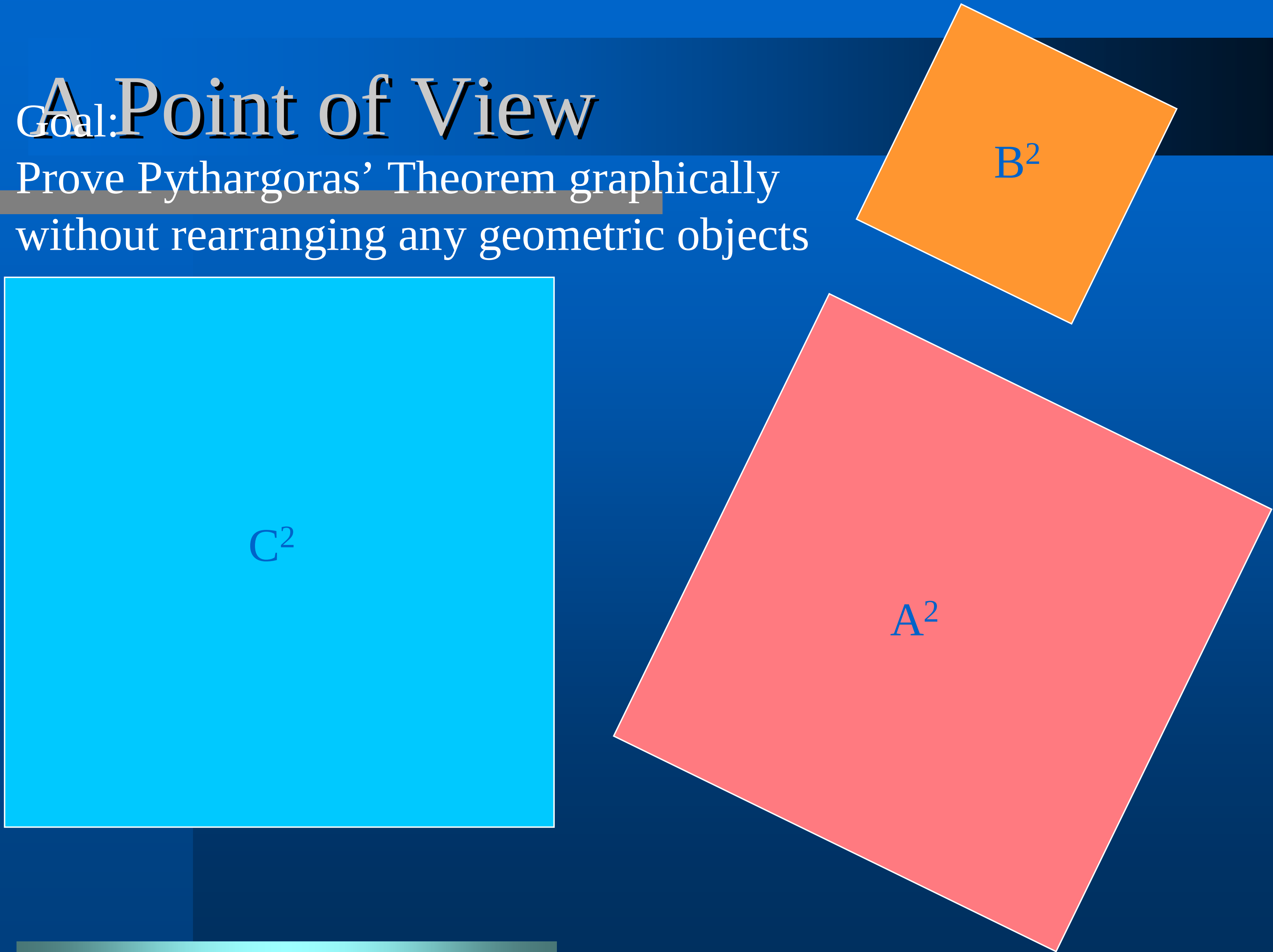
A Point of View



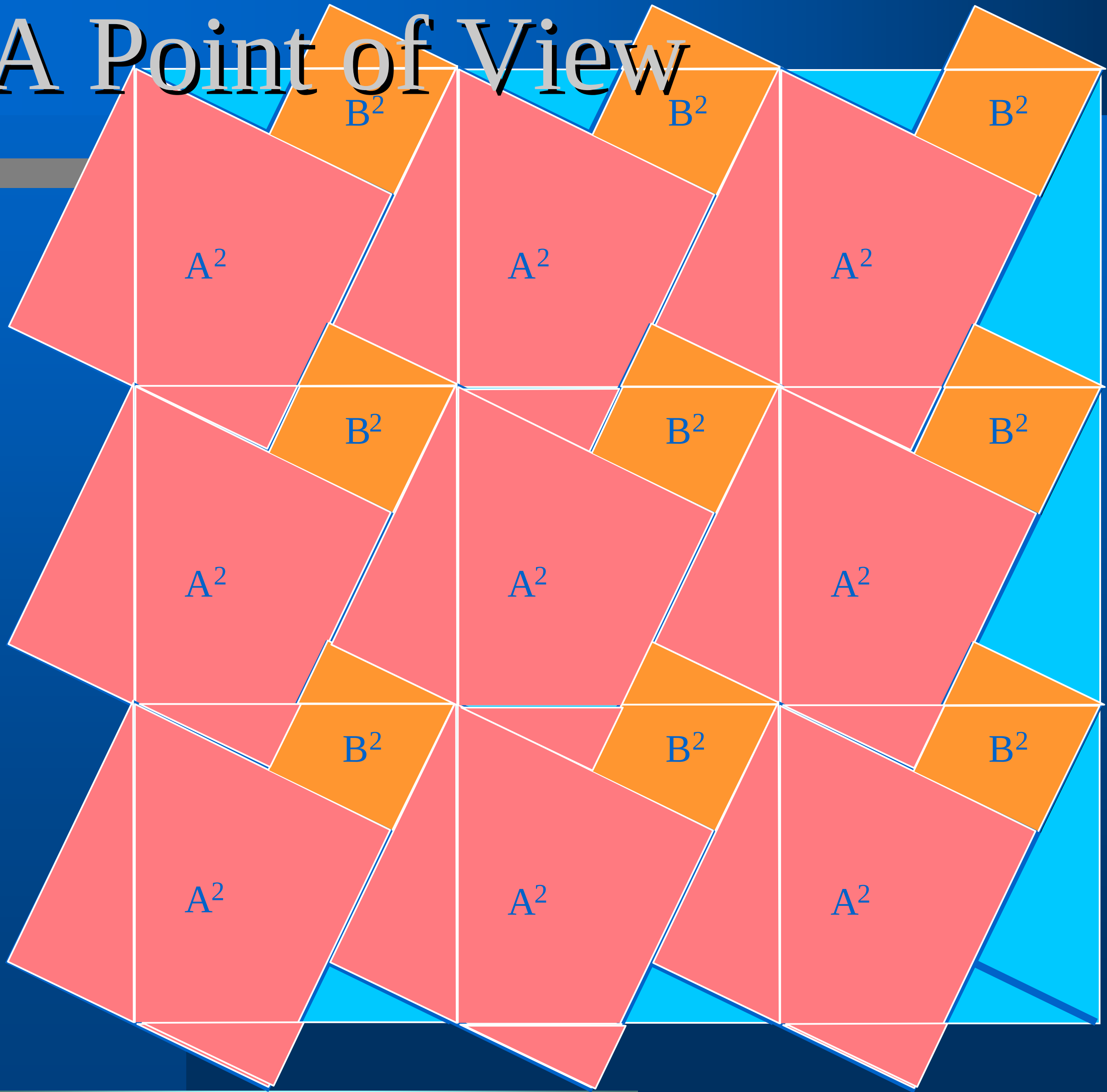
A Point of View

Goal:

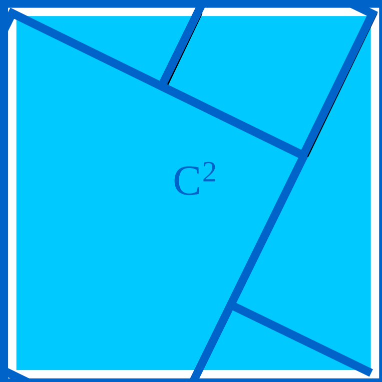
Prove Pythargoras' Theorem graphically without rearranging any geometric objects



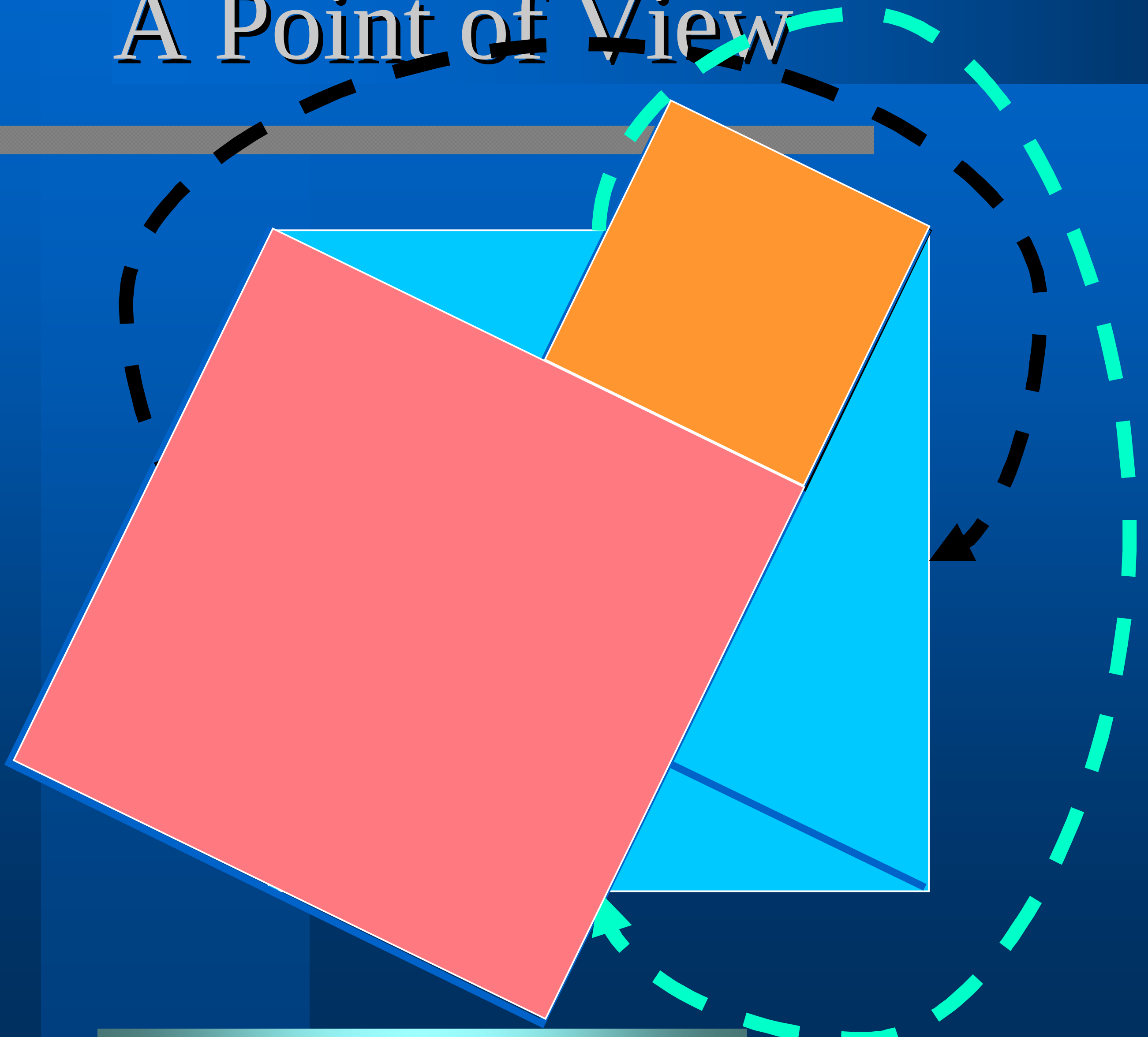
A Point of View



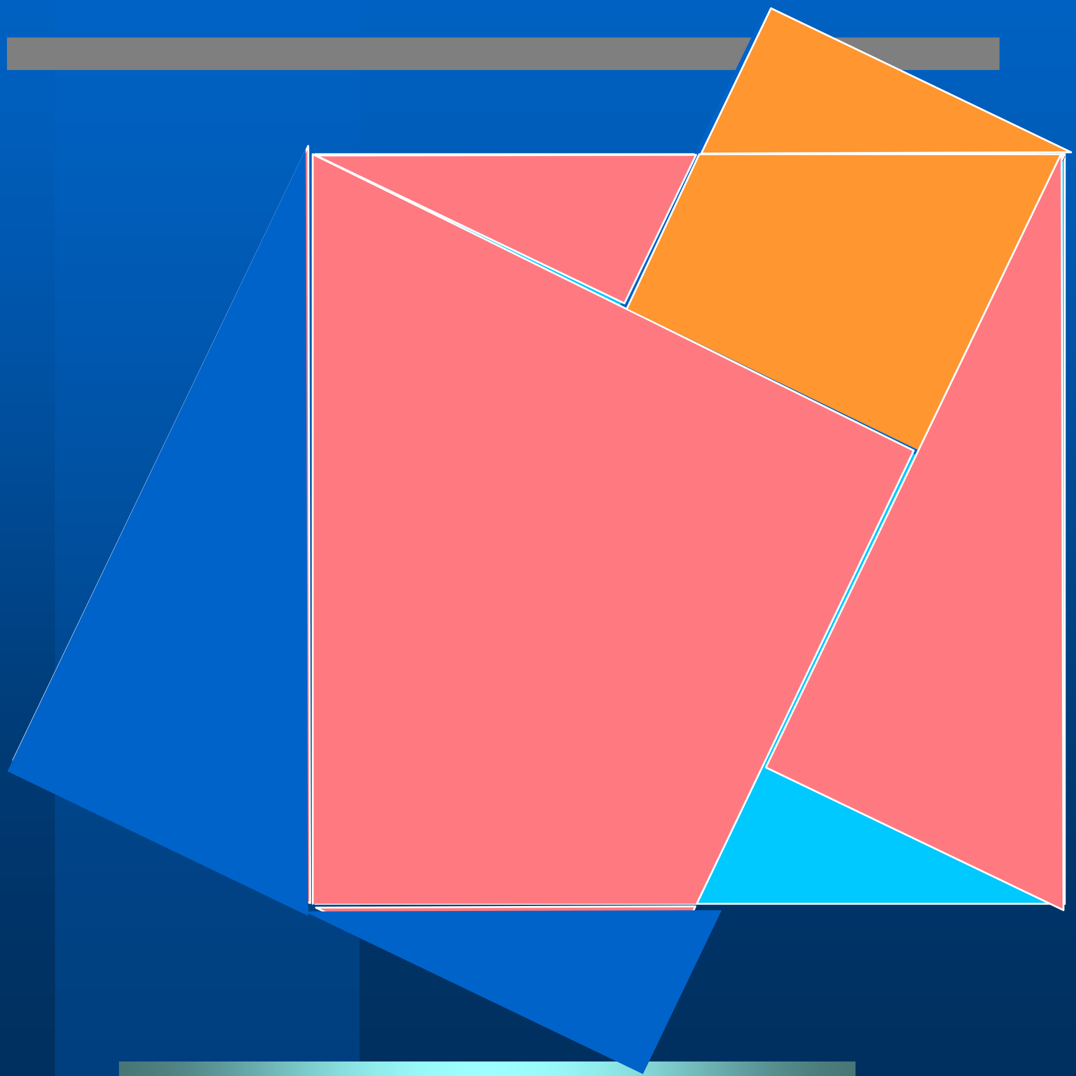
A Point of View



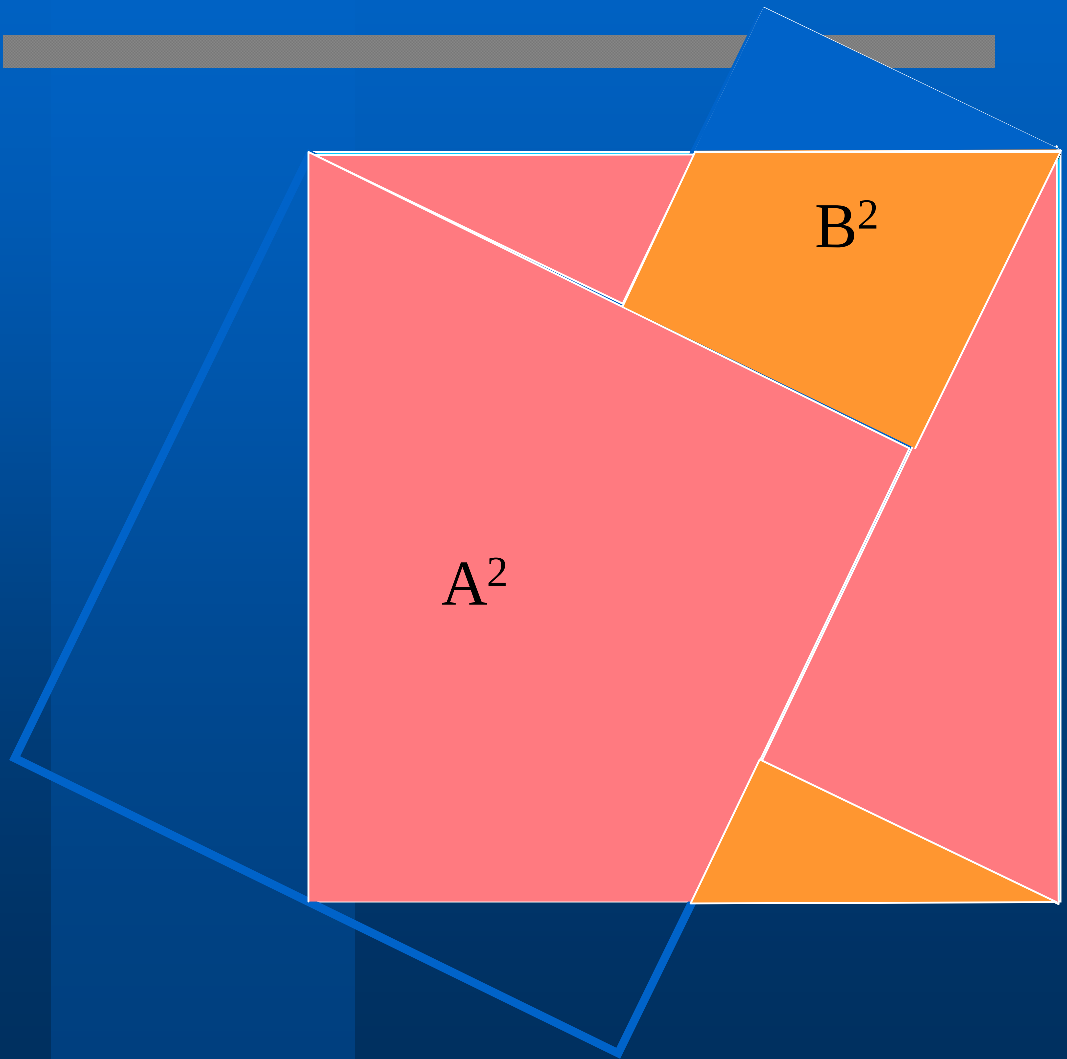
A Point of View



A Point of View



A Point of View



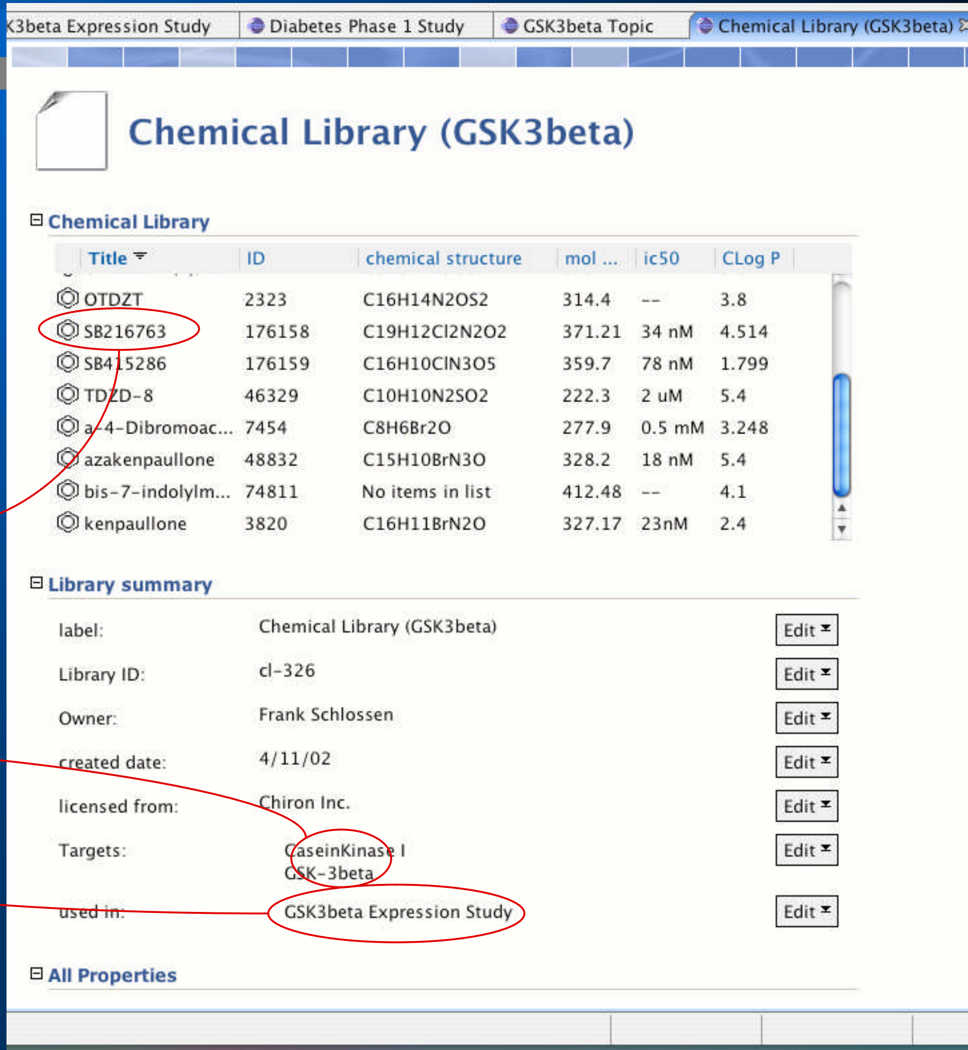
Chemical Libraries

- Lists of Compounds and progressed Chemical Entities
- All Chem properties semantically defined
- Target and assay information associated

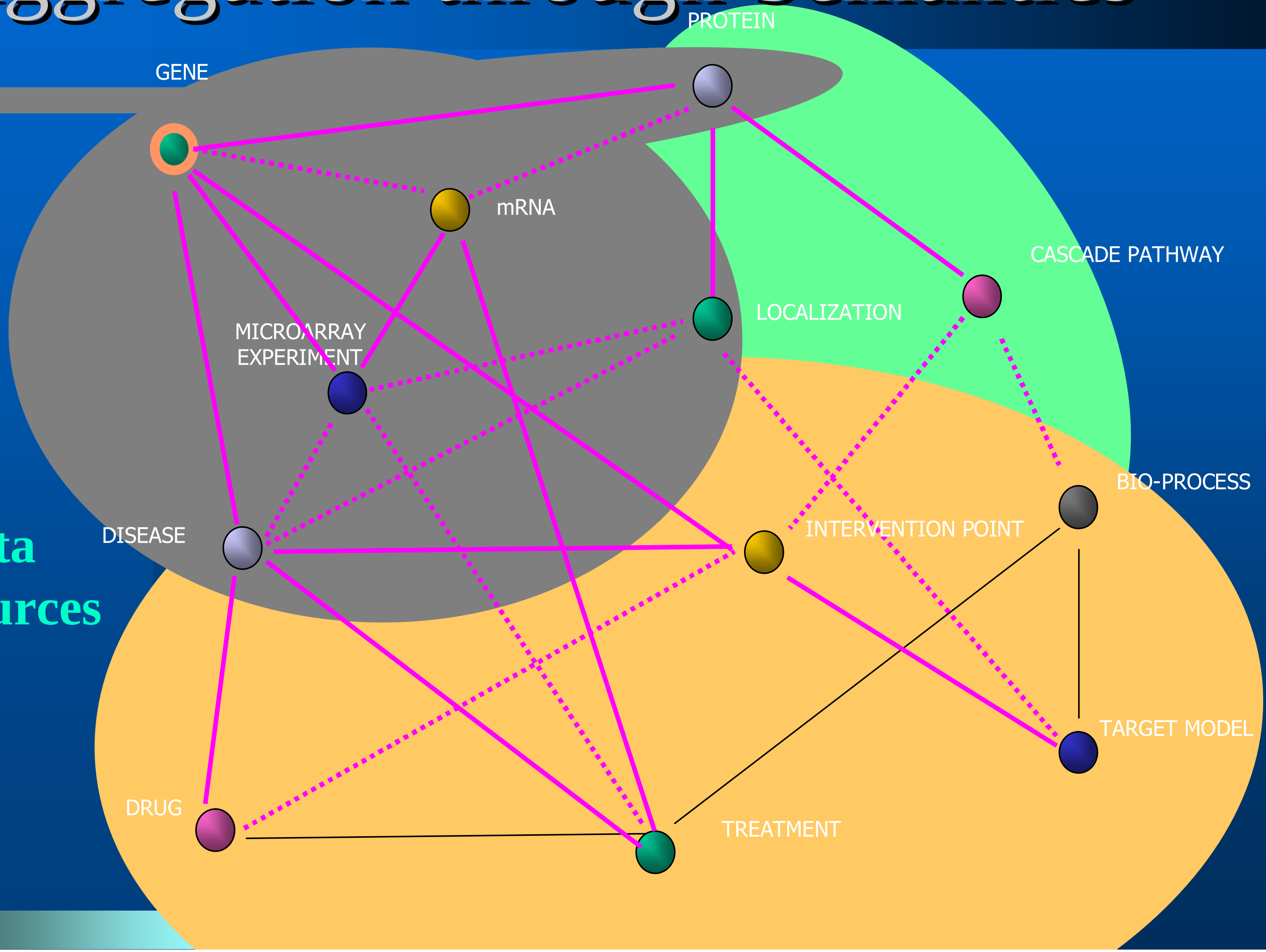
Compound Objects

Known Targets

Studies applied to



Aggregation through Semantics



Data Sources

Toxicogenomics

- Gene responses to administered drugs
- Drug Dosing can be associated with trends

2 Big Topics in Clinical Trials

- Safety
  - More Conclusive (making a better case)
  - Earlier Prediction (reduce costs)
- Efficacy
  - Clearly identify value to market
  - Better business planning

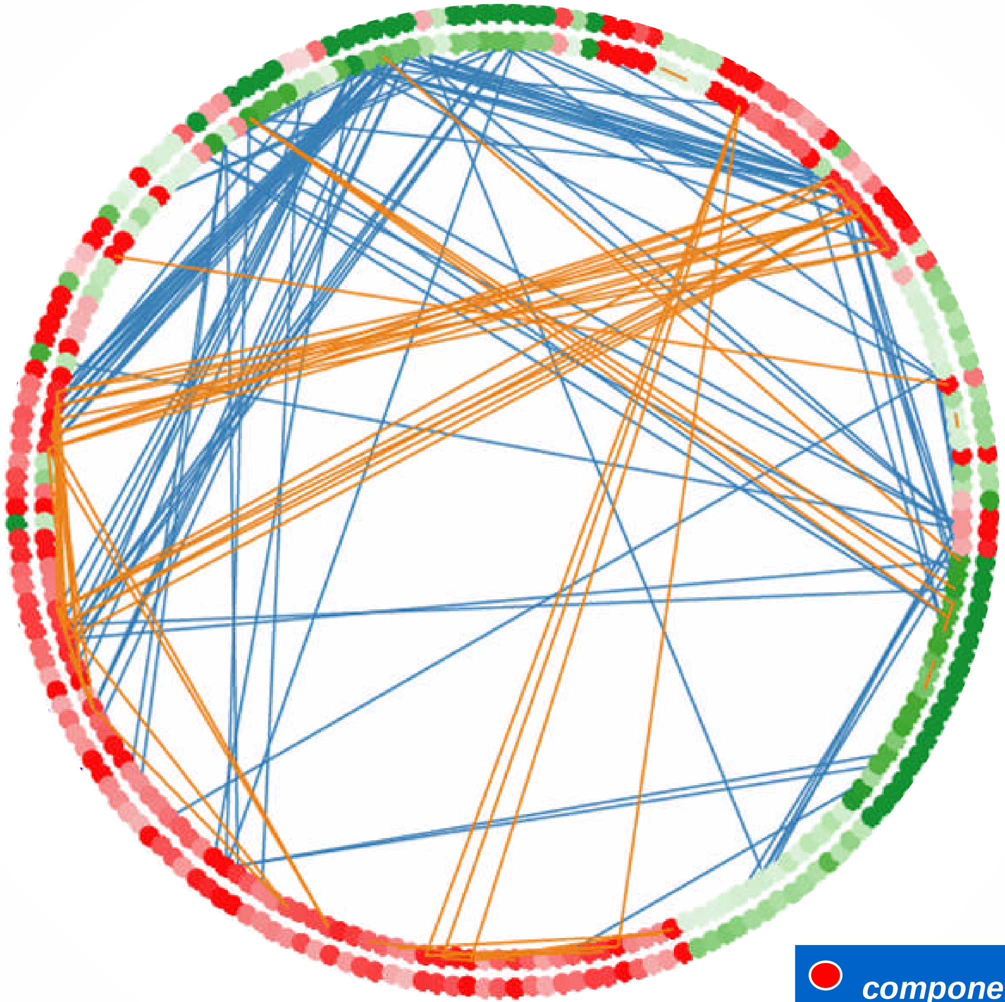
What's the Path to get there?

- Optimal Trial Designs
  - Using knowledge for recruitment
  - Judicious Choice of Biomarkers
  - Bayesian enablers

- Better information utilization
  - Animal-Human Cross studies
  - Compiled Tox-response knowledge
  - Just-in-Time use of external information

Correlation Network

- Cross Species Comparison: Human vs. Animal Metabolites

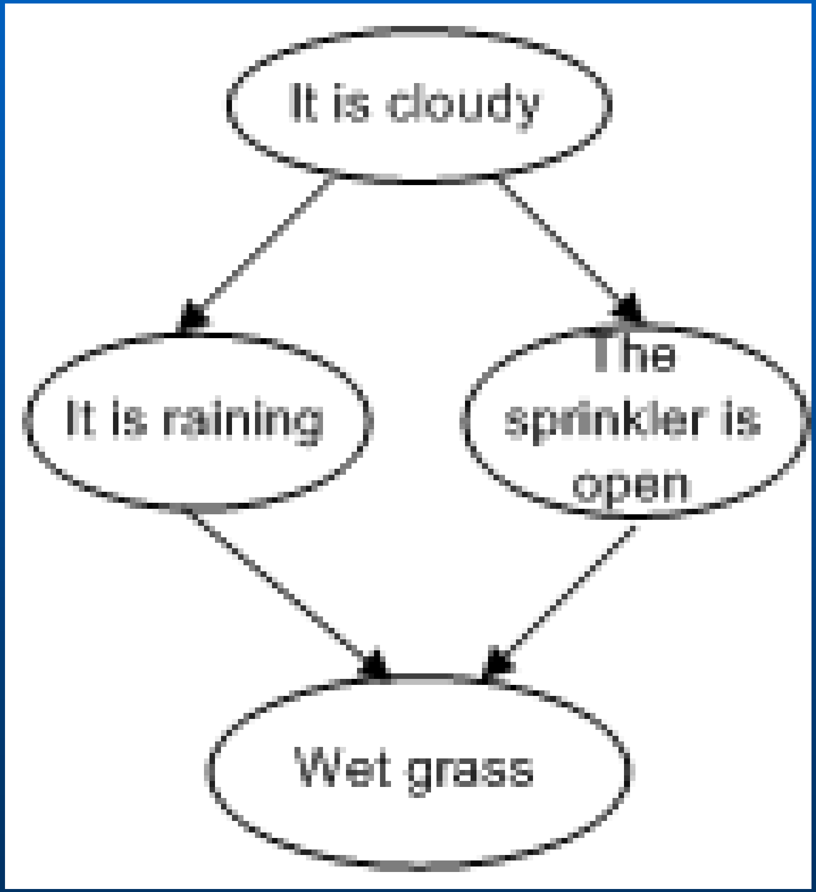


component higher in disease
component lower in disease

inner circle = affected **humans** vs. unaffected **humans**
outer circle = affected **animals** vs. unaffected **animals**

Bayesian Networks

- Captures Causality
- Usually a DAG
- Defines Conditional Independence
- Useful for Intervention Analysis



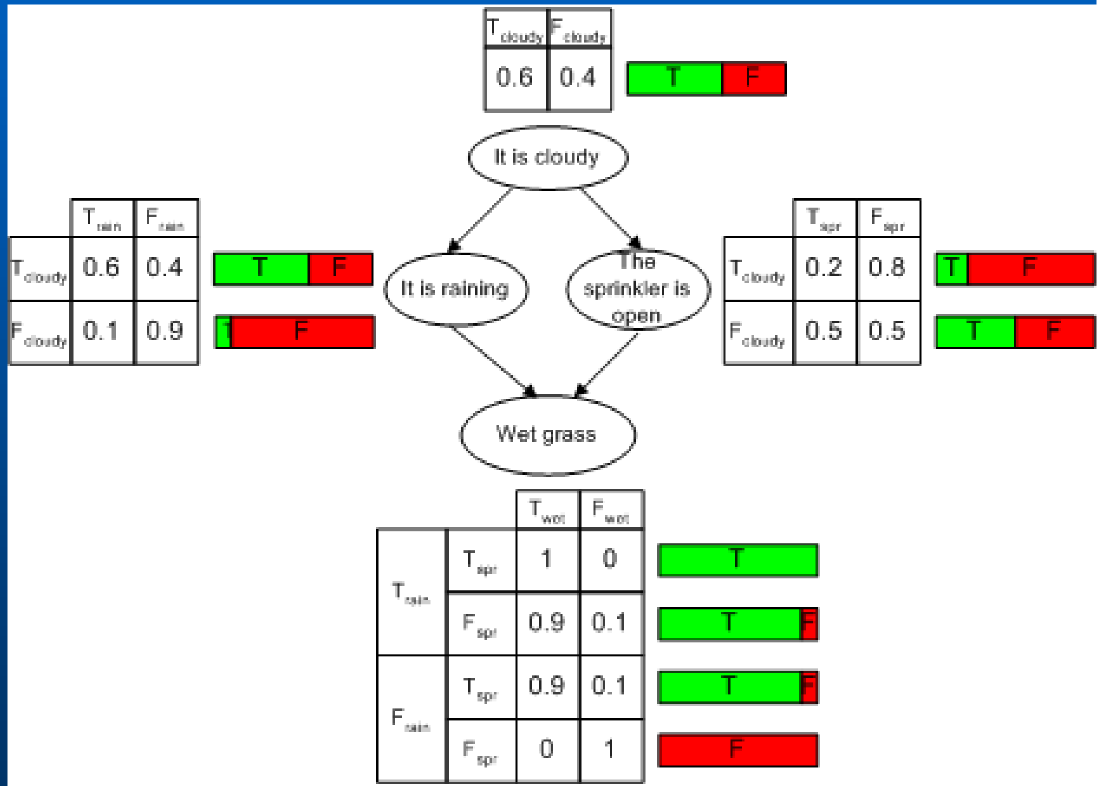
Bayesian Networks

Building Conditional Prob Tables

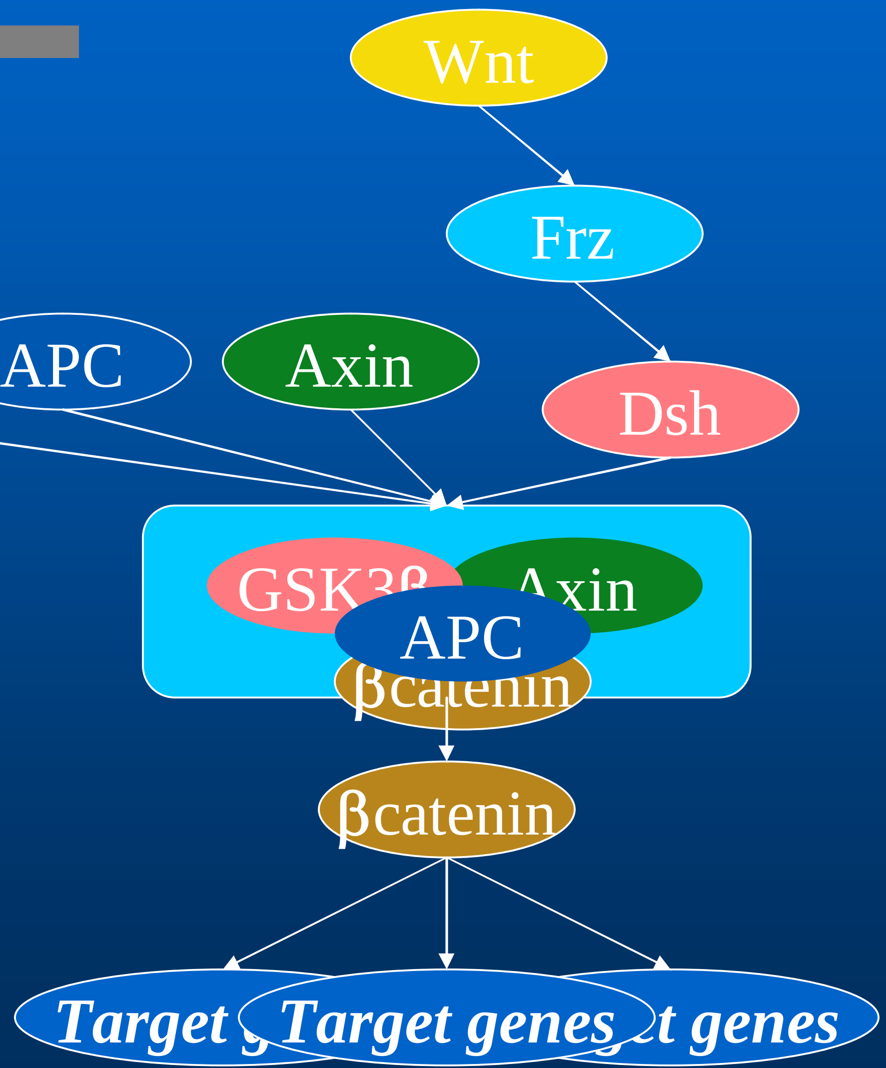
A path *p* between two variables *X* and *Y* is **d-separated** by a set of nodes *Z* iff

1. *p* contains a *chain i -> m -> j* or a *fork i <- m -> j* such that *m* is in *Z*, or
2. *p* does not contain a collider *i -> m <- j* such that *m* or any of its descendents are in Z.

A set *Z* is said to **d-separate** *X* from *Y* iff *Z* blocks every path between *X* and *Y*.



Causal Flow in Pathways



Each step is d-separated

⇒ Pathways can be inferred from analyses of d-separation

SB Impact on Informatics

- How to re-organize data and information around these
  - Experiments that generate multiple kinds of data (BG)
  - Heterogeneous info sources (DBs, literature)
  - Use of Pathway knowledge
  - Handling analyzed info for interpretation
- P roposing and encoding models and hypotheses
  - Data format (syntax) standards will not work
  - Need to code “Statements” - more like linguistics
- The need for a higher form of information service: ***Semantics***

Handling High-Throughput Data

Unsupervised

- **Cluster Analysis**
- **PCA**
- **Self-Organizing Maps**

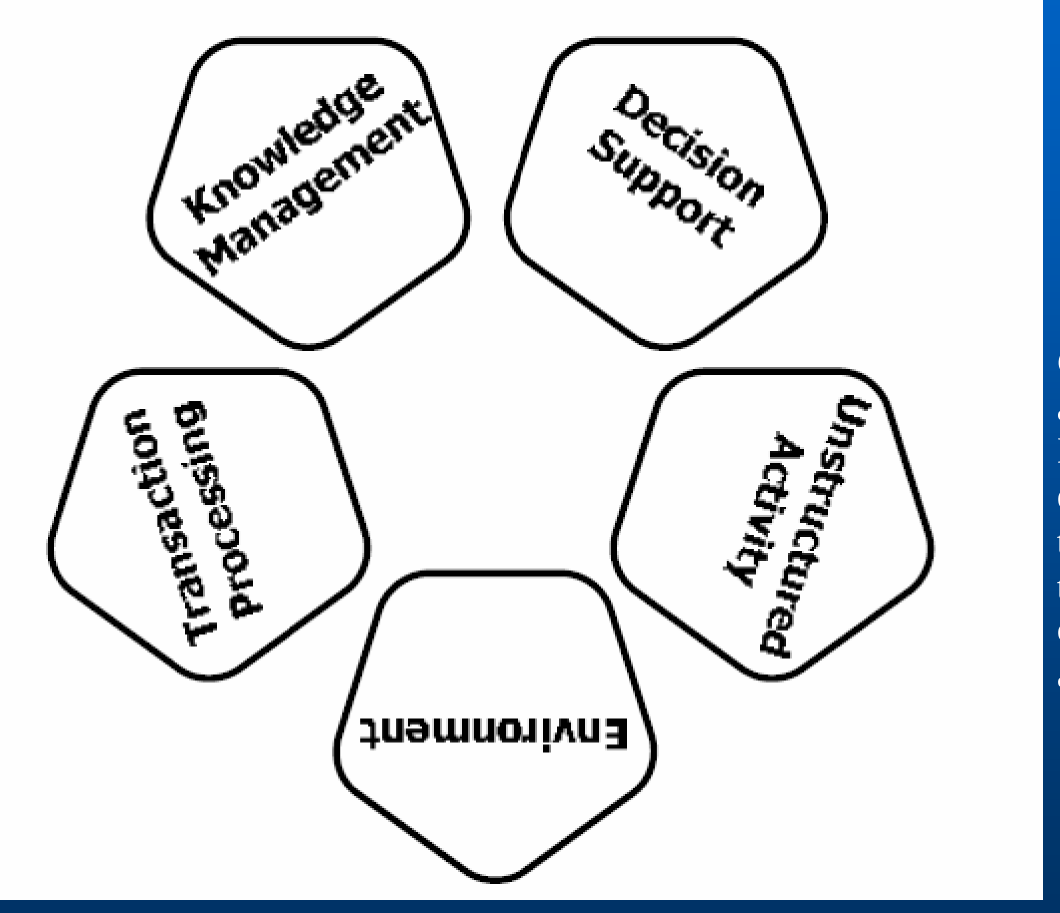
Supervised

- **Support Vector Machines (incl. NN)**
- **Bayesian Networks**

Five Areas of Semantic Applications

Ontologies and Semantic Consistency

EAI (Enterprise Application Integration), Service Oriented Architectures
Where 35% to 65% of budgets for integration and interoperation are spent because a large percentage of those systems have semantic mismatches as their basis. .



Information Consolidation and ETL QC

Content Management and Emails
Individual interaction of employees that affect tasks, resulting in transactions that are currently well captured and structured..

Enterprise Mapping and CRM
Supply chain and our demand chain, through which we are currently connected with e-commerce and B2B.