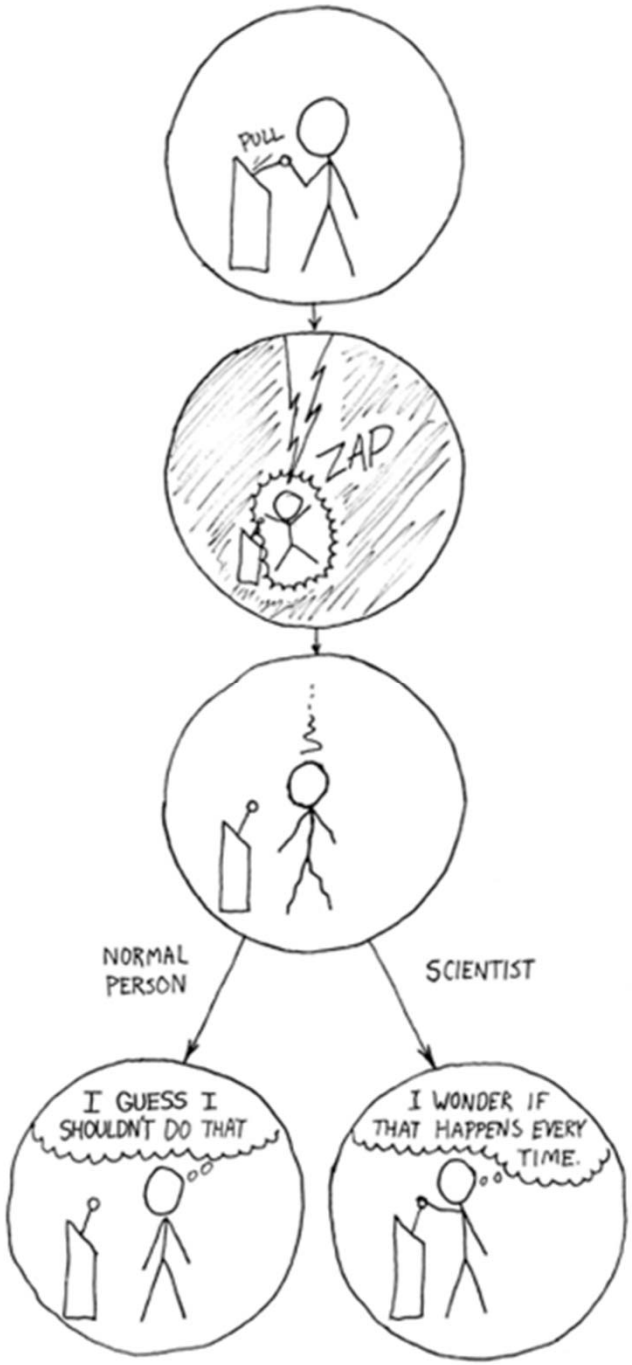
Semantic Approaches for Biomedical Knowledge Discovery

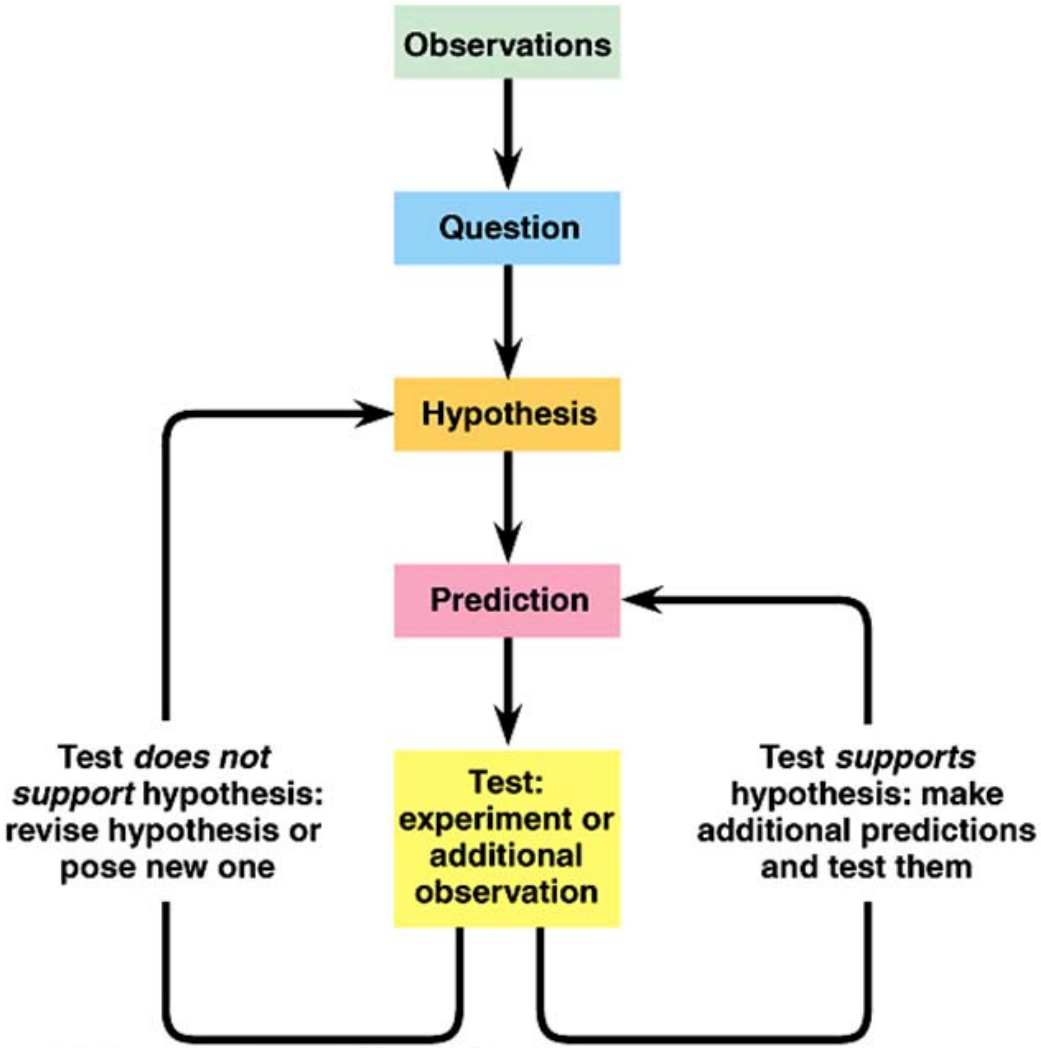
Michel Dumontier, Ph.D.

Associate Professor of Medicine (Biomedical Informatics)
Stanford University

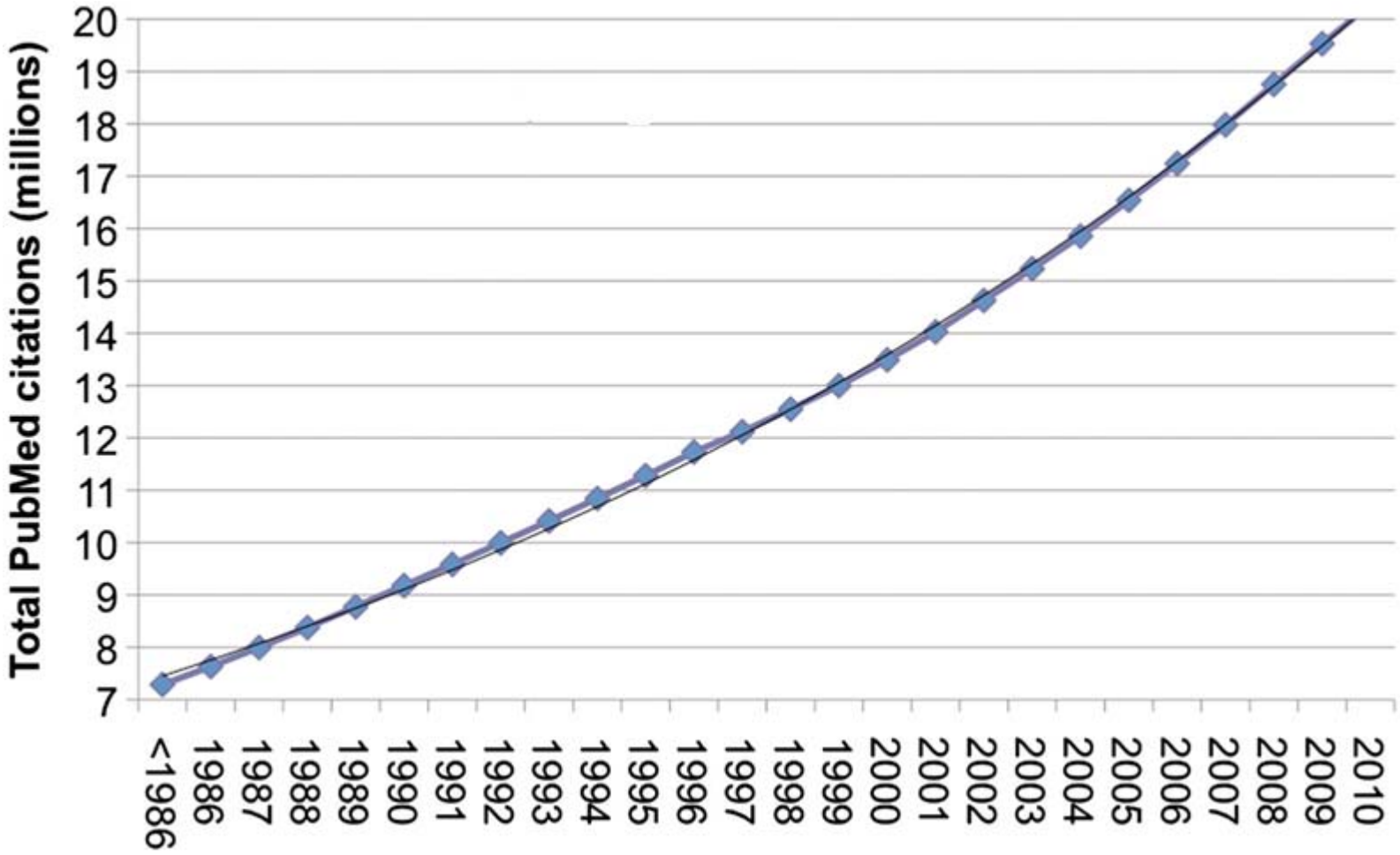


HTTP://XKCD.COM/242/

Science



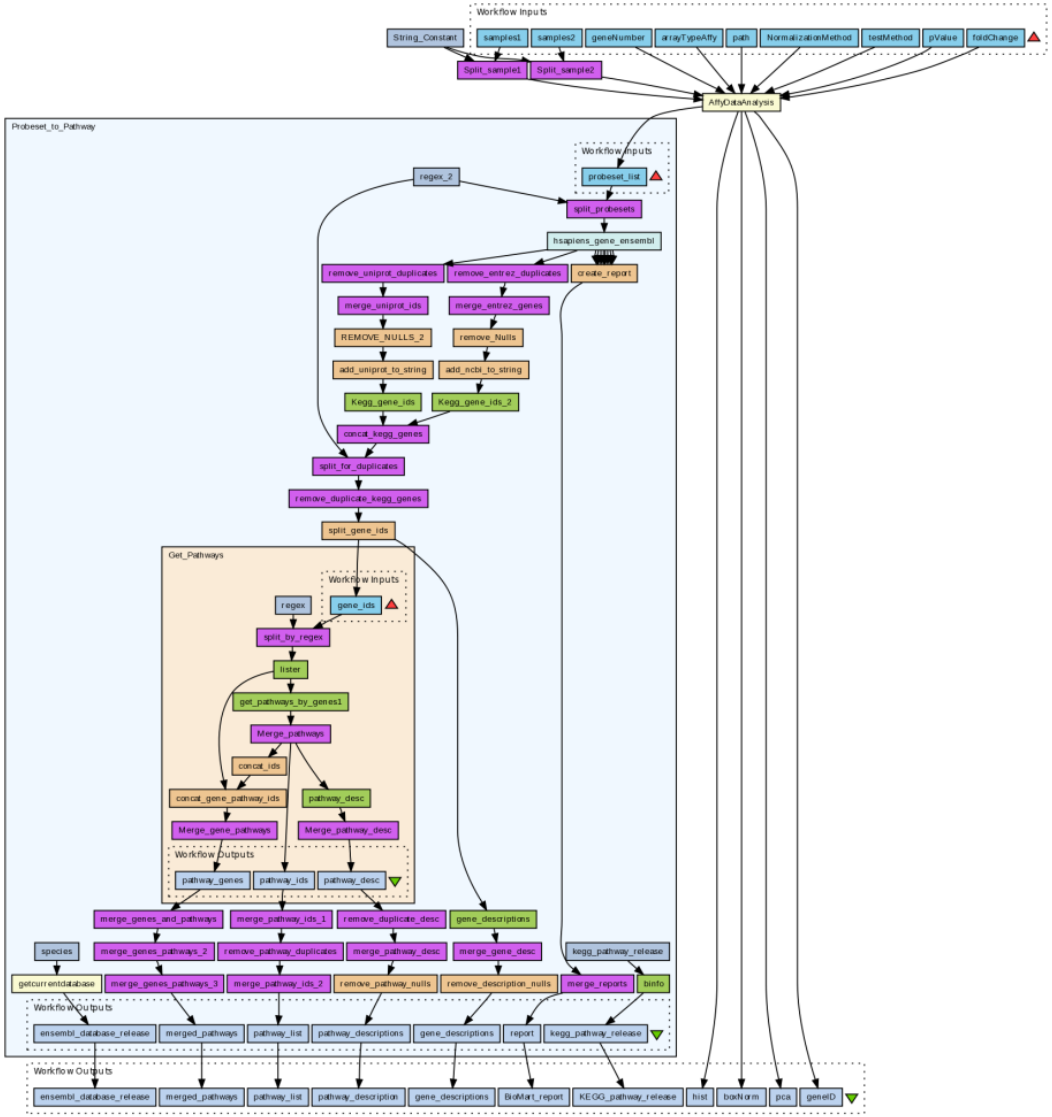
The unbelievable growth of scientific knowledge



Thousands of databases curate the literature into consumable facts (problems: access, format, identifiers & linking)

Software is needed to analyze, predict and evaluate (problems: OS, versioning, input/output formats)

Ultimately, we develop fairly sophisticated programs/workflows to test our hypotheses



Wouldn't it be great if we could just find the evidence required to support or dispute a scientific hypothesis using the most *up-to-date* and *relevant* data, tools and scientific knowledge?

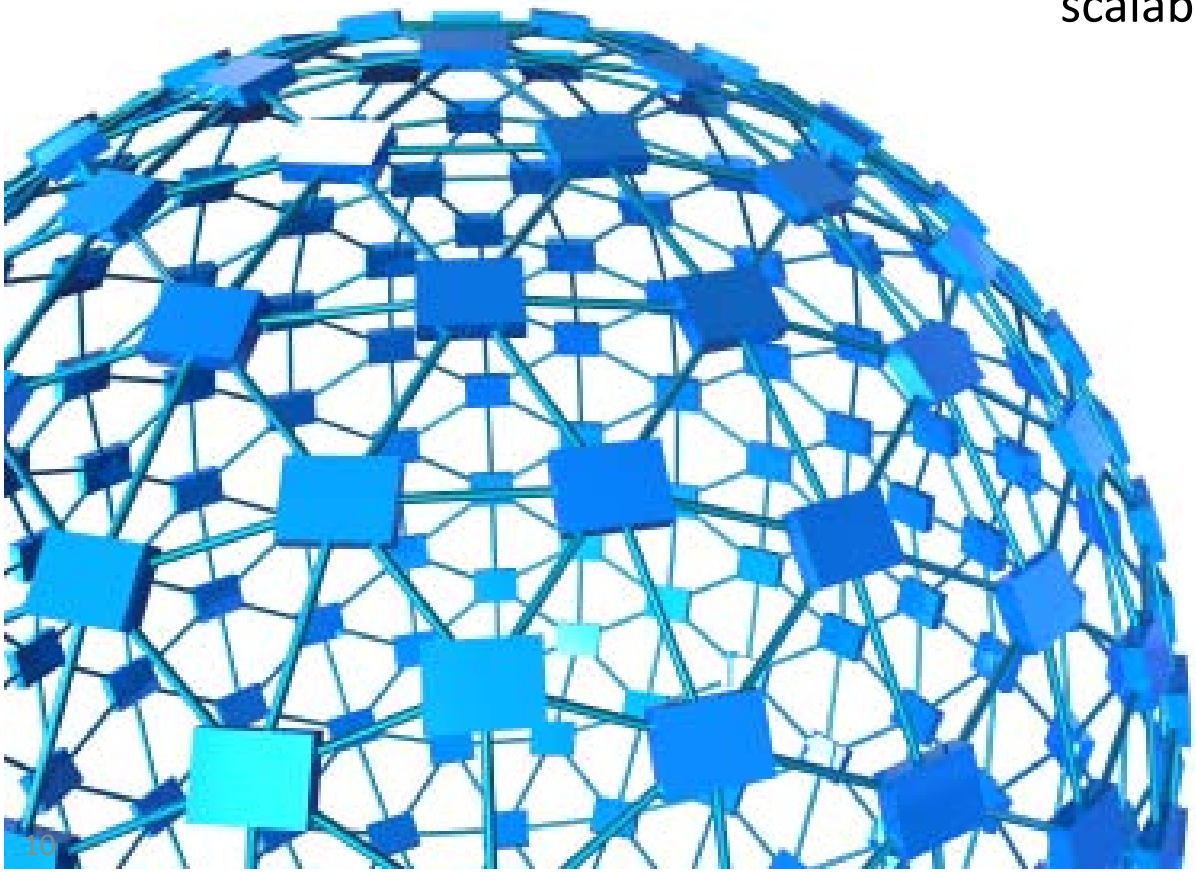
So what do we need to achieve this?

1. Standards to construct and interrogate a massive, decentralized network of *interconnected* data and software
2. Methods and Tools
   - To prepare, interlink, and query data
   - To mine and discover associations
   - To identify novel, supported associations
3. Incentives and penalties
   - Funding agencies, journals, institutions, societies, conferences, workshops

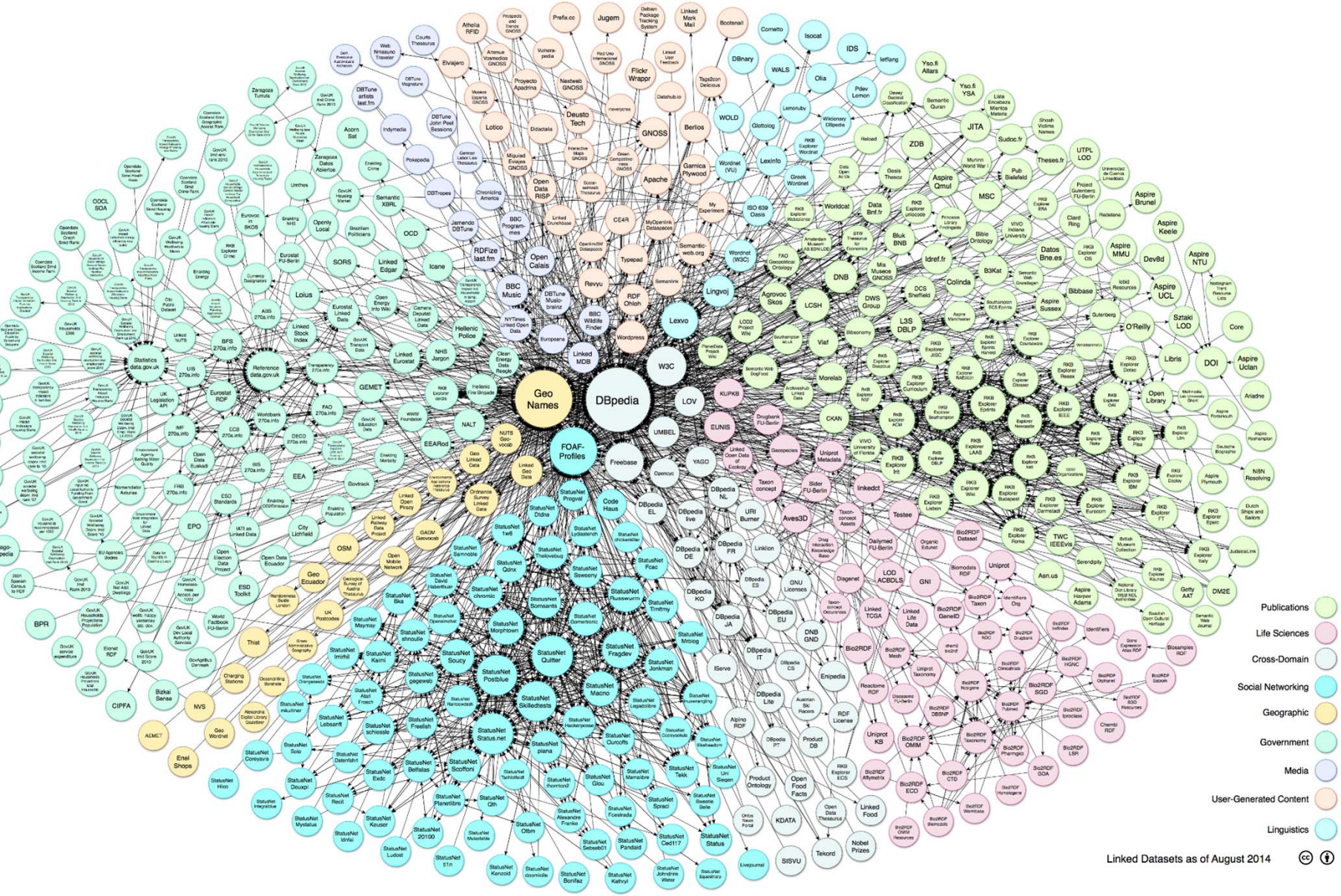
The Semantic Web
is the new global **web of knowledge**

standards for publishing, sharing and querying
facts, expert knowledge and services

scalable approach for the discovery
of *independently formulated*
and *distributed* knowledge



we're building a massive network of linked data

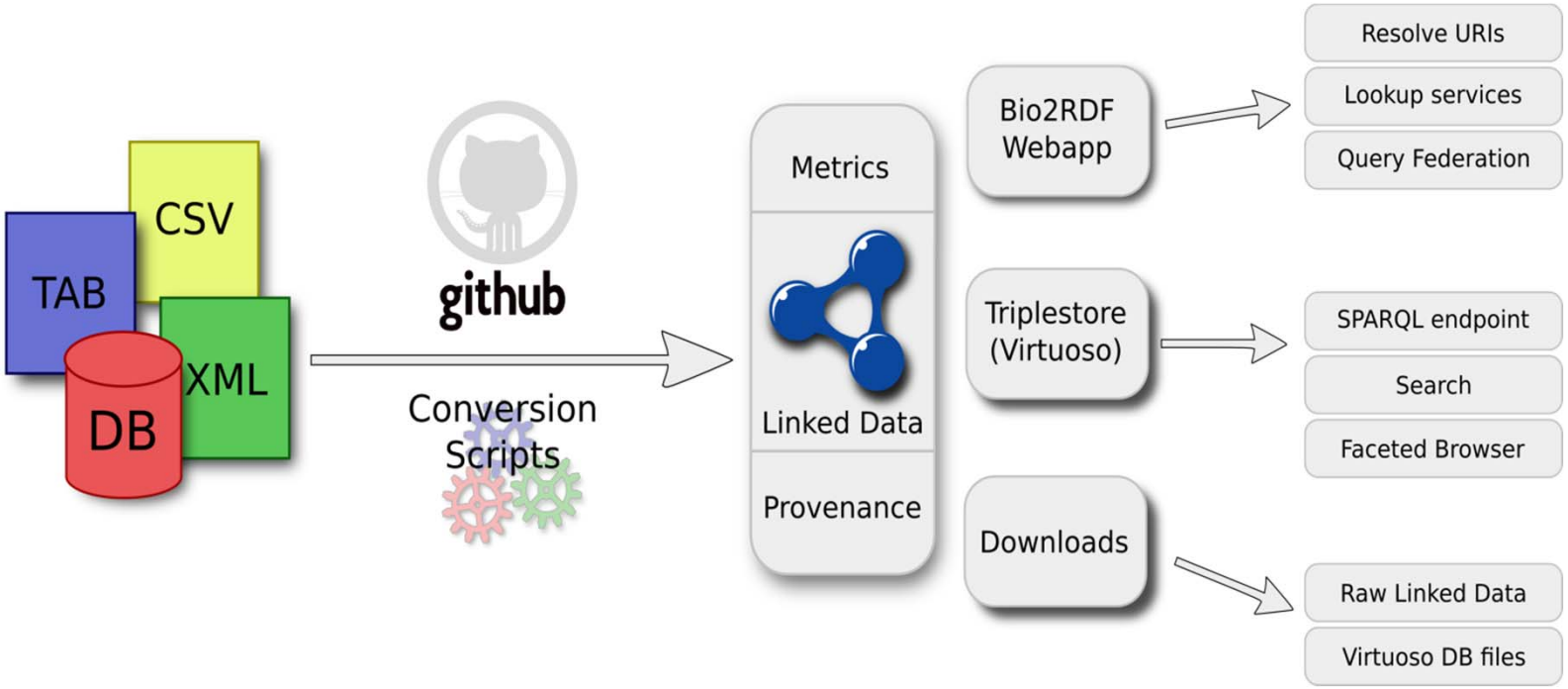


Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/"

BIO2RDF

Linked Data for the Life Sciences

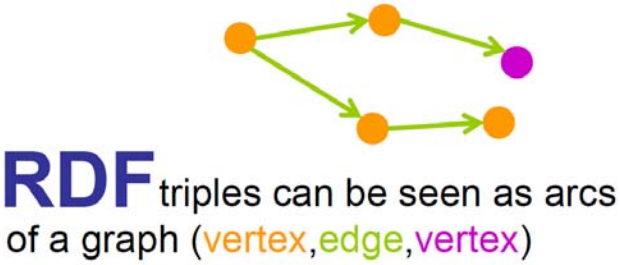
Bio2RDF is an open source project to unify the representation and interlinking of biological data using RDF.



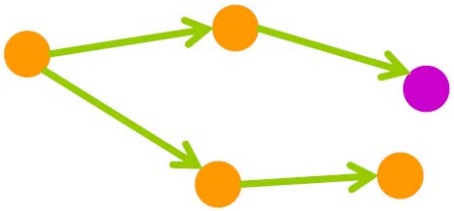
chemicals/drugs/formulations,
genomes/genes/proteins, domains
Interactions, complexes & pathways
animal models and phenotypes
Disease, genetic markers, treatments
Terminologies & publications

- Release 3 (June 2014): 11B+ interlinked statements from 35 biomedical datasets
- dataset description, provenance & statistics
- Partnerships with EBI, NCBI, DBCLS, NCBO, OpenPHACTS, and commercial tool providers

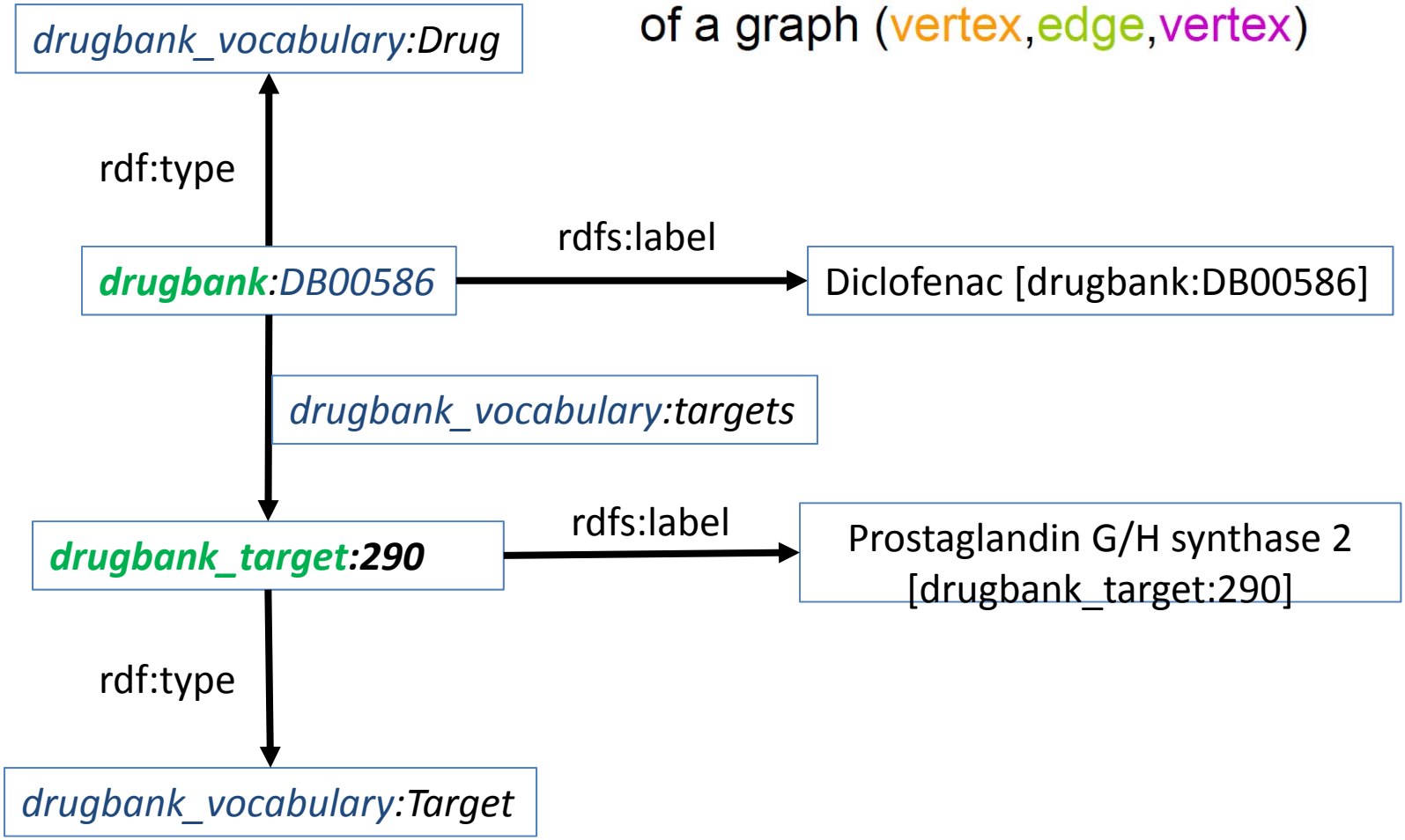
Resource Description Framework



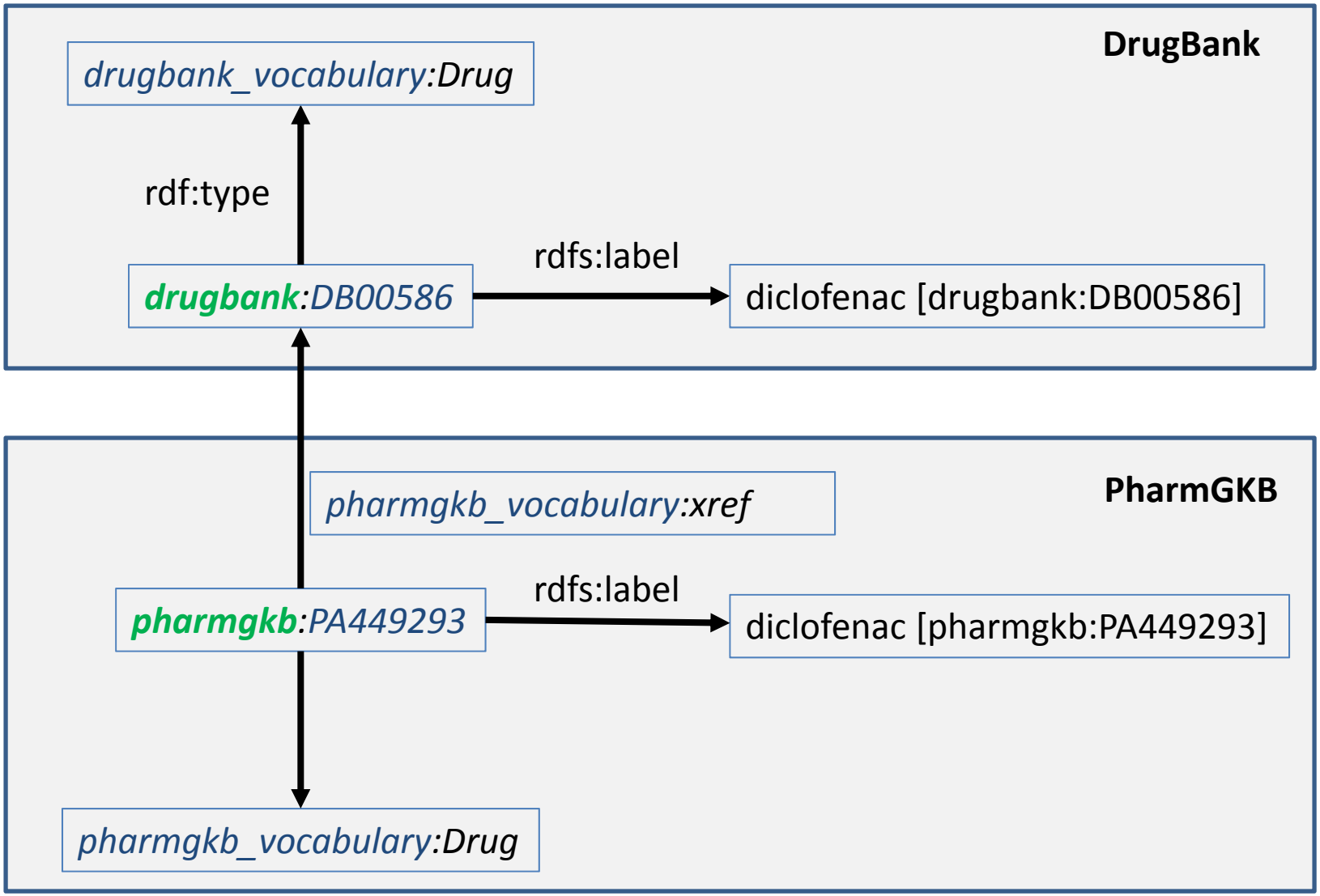
- It's a language to represent knowledge
  - Logic-based formalism -> automated reasoning
  - graph-like properties -> data analysis
- Good for
  - Describing in terms of type, attributes, relations
  - Integrating data from different sources
  - Sharing the data (W3C standard)
  - Reusing what is available, developing what you need, and contributing back to the web of data.



RDF triples can be seen as arcs of a graph (vertex,edge,vertex)



The linked data network expands with every reference



BIO2RDF

Linked Data for the Life Sciences

–Release 3–

[website] [datasets] [documentation]

DrugBank

The DrugBank database is a bioinformatics and chemoinformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information.

Keywords: drug,protein

Homepage: http://www.drugbank.ca/

Organization: DrugBank

License: license

Identifier Regex Pattern: ^DB\d{5}$

SPARQL Endpoint URL: http://cu.drugbank.bio2rdf.org/sparql

Faceted Browser URL: http://cu.drugbank.bio2rdf.org/fct

Conversion Script URL: http://github.com/bio2rdf/bio2rdf-scripts/tree/master/drugbank

Download URL: http://download.bio2rdf.org/release/3/drugbank

Contents

- Basic metrics
- Types
- Object Properties
- Datatype Properties
- Subject Type and Property
- Property and Object Type
- Type-Property-Type List
- Dataset-Property-Dataset List

Basic metrics

Search:

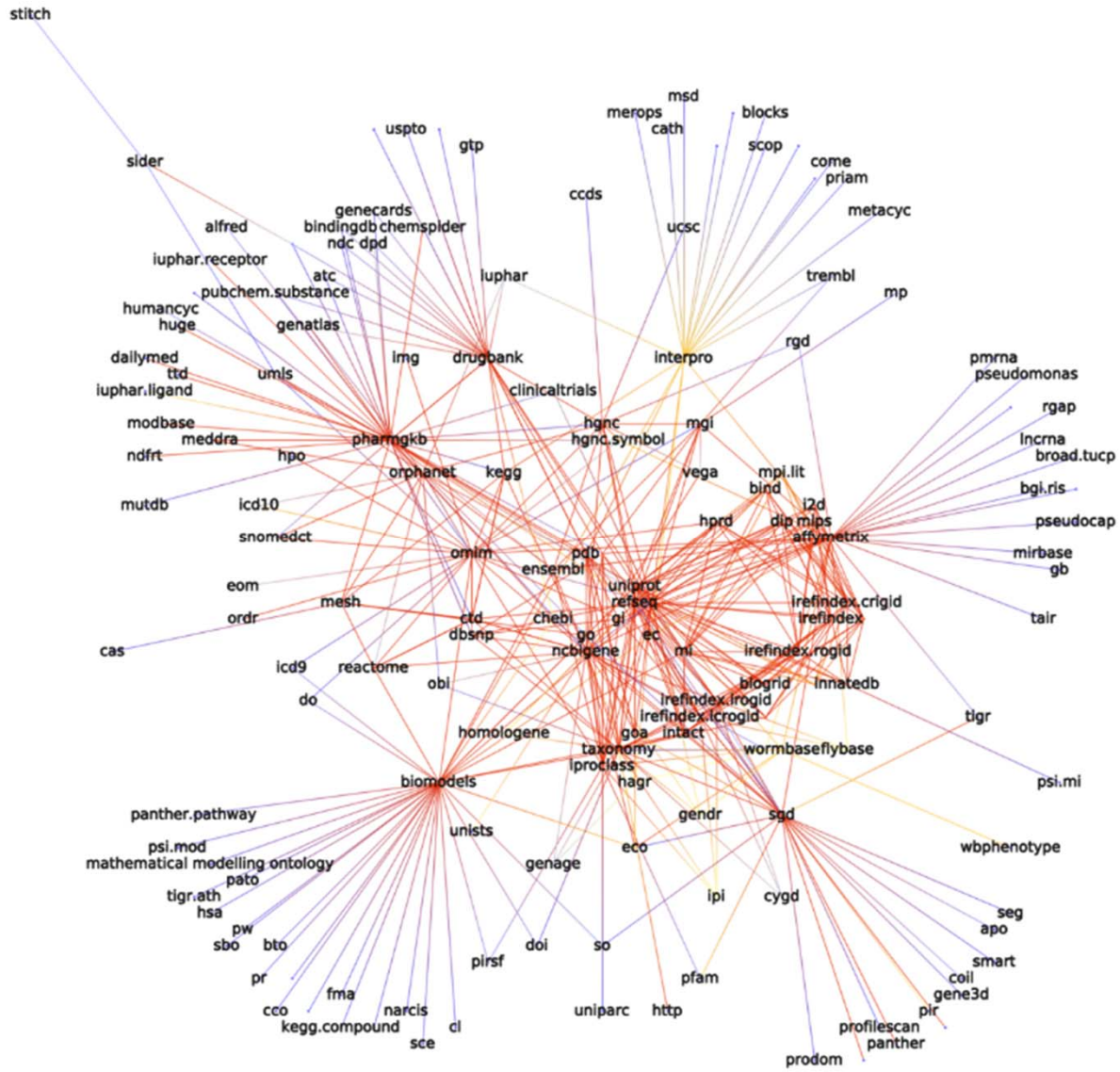
Triples	3672531
Distinct Literals	1459302
Distinct Objects	370346
Distinct Entities	316950
Distinct Subjects	316950
Distinct Properties	105
Distinct Types	91

Dataset-Property-Dataset List

Search:

Dataset	Property	Dataset
drugbank	x wikipedia [drugbank_vocabulary:x-wikipedia] http://bio2rdf.org/drugbank_vocabulary:x-wikipedia	wikipedia
drugbank	x uniprot [drugbank_vocabulary:x-uniprot] http://bio2rdf.org/drugbank_vocabulary:x-uniprot	uniprot
drugbank	x pubchemsubstance [drugbank_vocabulary:x-pubchemsubstance] http://bio2rdf.org/drugbank_vocabulary:x-pubchemsubstance	pubchem.substance
drugbank	x pubchemcompound [drugbank_vocabulary:x-pubchemcompound] http://bio2rdf.org/drugbank_vocabulary:x-pubchemcompound	pubchem.compound
drugbank	x pharmgkb [drugbank_vocabulary:x-pharmgkb] http://bio2rdf.org/drugbank_vocabulary:x-pharmgkb	pharmgkb
drugbank	x pdb [drugbank_vocabulary:x-pdb] http://bio2rdf.org/drugbank_vocabulary:x-pdb	pdb
drugbank	x ndc [drugbank_vocabulary:x-ndc] http://bio2rdf.org/drugbank_vocabulary:x-ndc	ndc
drugbank	x kegg [drugbank_vocabulary:x-kegg] http://bio2rdf.org/drugbank_vocabulary:x-kegg	kegg
drugbank	x iuphar [drugbank_vocabulary:x-iuphar] http://bio2rdf.org/drugbank_vocabulary:x-iuphar	iuphar
drugbank	x hgnc [drugbank_vocabulary:x-hgnc] http://bio2rdf.org/drugbank_vocabulary:x-hgnc	hgnc
drugbank	x gtp [drugbank_vocabulary:x-gtp] http://bio2rdf.org/drugbank_vocabulary:x-gtp	gtp
drugbank	x gi [drugbank_vocabulary:x-gi] http://bio2rdf.org/drugbank_vocabulary:x-gi	gi
drugbank	x genecards [drugbank_vocabulary:x-genecards] http://bio2rdf.org/drugbank_vocabulary:x-genecards	genecards
drugbank	x genbank [drugbank_vocabulary:x-genbank] http://bio2rdf.org/drugbank_vocabulary:x-genbank	genbank
drugbank	x genatlas [drugbank_vocabulary:x-genatlas] http://bio2rdf.org/drugbank_vocabulary:x-genatlas	genatlas
drugbank	x dpd [drugbank_vocabulary:x-dpd] http://bio2rdf.org/drugbank_vocabulary:x-dpd	dpd
drugbank	x chemspider [drugbank_vocabulary:x-chemspider] http://bio2rdf.org/drugbank_vocabulary:x-chemspider	chemspider

Bio2RDF offers a highly connected network of data



Graph summarization for query formulation

Type-Property-Type List

Search: drug-drug-interaction

Total Subjects	Distinct Subjects	Subject Type	Property	Object Type	Distinct Objects	Total Objects
24095	1176	Drug [drugbank_vocabulary:Drug] http://bio2rdf.org/drugbank_vocabulary:Drug	ddi interactor in [drugbank_vocabulary:ddi-interactor-in] http://bio2rdf.org/drugbank_vocabulary:ddi-interactor-in	drug-drug interaction [drugbank_vocabulary:Drug-Drug-Interaction] http://bio2rdf.org/drugbank_vocabulary:Drug-Drug-Interaction	12104	24095
12104	12104	drug-drug interaction [drugbank_vocabulary:Drug-Drug-Interaction] http://bio2rdf.org/drugbank_vocabulary:Drug-Drug-Interaction	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	drugbank resource [drugbank_vocabulary:Resource] http://bio2rdf.org/drugbank_vocabulary:Resource	1	12104
12104	12104	drug-drug interaction [drugbank_vocabulary:Drug-Drug-Interaction] http://bio2rdf.org/drugbank_vocabulary:Drug-Drug-Interaction	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Class	1	12104
12104	12104	drug-drug interaction [drugbank_vocabulary:Drug-Drug-Interaction] http://bio2rdf.org/drugbank_vocabulary:Drug-Drug-Interaction	http://rdfs.org/ns/void#inDataset	http://purl.org/dc/terms/Dataset	1	12104
12104	12104	drug-drug interaction [drugbank_vocabulary:Drug-Drug-Interaction] http://bio2rdf.org/drugbank_vocabulary:Drug-Drug-Interaction	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2000/01/rdf-schema#Resource	1	12104
24208	1199	drugbank resource [drugbank_vocabulary:Resource] http://bio2rdf.org/drugbank_vocabulary:Resource	ddi interactor in [drugbank_vocabulary:ddi-interactor-in] http://bio2rdf.org/drugbank_vocabulary:ddi-interactor-in	drug-drug interaction [drugbank_vocabulary:Drug-Drug-Interaction] http://bio2rdf.org/drugbank_vocabulary:Drug-Drug-Interaction	12104	24208

```
PREFIX drugbank_vocabulary: <http://bio2rdf.org/drugbank_vocabulary:>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?ddi ?d1name ?d2name
WHERE {
        ?ddi a drugbank_vocabulary:Drug-Drug-Interaction .
        ?d1 drugbank_vocabulary:ddi-interactor-in ?ddi .
        ?d1 rdfs:label ?d1name .
        ?d2 drugbank_vocabulary:ddi-interactor-in ?ddi .
        ?d2 rdfs:label ?d2name.
    FILTER (?d1 != ?d2)
}
```

You can use query assistants

http://sindicetech.com/sindice-suite/sparqled/

```
1  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2
3  SELECT ?drugname ?genename WHERE {
4     ?s a <http://bio2rdf.org/pharmgkb_vocabulary:Drug-Gene-Association> .
5     ?s <
6  }
7  LIMIT 1
8
```

<http://bio2rdf.org/pharmgkb_vocabulary:association_type>
<http://www.w3.org/2000/01/rdf-schema#label>
<http://bio2rdf.org/pharmgkb_vocabulary:drug>
<http://bio2rdf.org/pharmgkb_vocabulary:gene>
<http://rdfs.org/ns/void#inDataset>
<http://bio2rdf.org/pharmgkb_vocabulary:article>
<http://bio2rdf.org/pharmgkb_vocabulary:pd_relationship>
<http://bio2rdf.org/pharmgkb_vocabulary:pk_relationship>

```
1  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2
3  SELECT ?drugname ?genename WHERE {
4     ?s a <http://bio2rdf.org/pharmgkb_vocabulary:Drug-Gene-Association> .
5     ?s <http://bio2rdf.org/pharmgkb_vocabulary:drug> ?drug .
6     ?s <http://bio2rdf.org/pharmgkb_vocabulary:gene> ?gene .
7     ?drug <http://www.w3.org/2000/01/rdf-schema#label> ?drugname .
8     ?gene <http://www.w3.org/2000/01/rdf-schema#label> ?genename .
9  }
10 LIMIT 10
```

graph: http://sindicetech.com/analytics

Federated Queries over Independent SPARQL EndPoints

Get all protein catabolic processes (and more specific) in biomodels

```
SELECT ?go ?label count(distinct ?x)
WHERE {
 service <http://bioportal.bio2rdf.org/sparql> {
   ?go rdfs:label ?label .
   ?go rdfs:subClassOf ?tgo
   ?tgo rdfs:label ?tlabel .
   FILTER regex(?tlabel, "^protein catabolic process")
   }
   service <http://biomodels.bio2rdf.org/sparql> {
    ?x <http://bio2rdf.org/biopax_vocabulary:identical-to> ?go .
    ?x a <http://www.biopax.org/release/biopax-level3.owl#BiochemicalReaction> .
   }
}
```

- protein catabolic process
  - catabolism by organism of protein in other organism
  - cellular protein catabolic process
  - glycoprotein catabolic process
  - hemoglobin catabolic process
  - insulin catabolic process
  - lipoprotein catabolic process
  - multicellular organismal protein catabolic process
  - proteasomal protein catabolic process

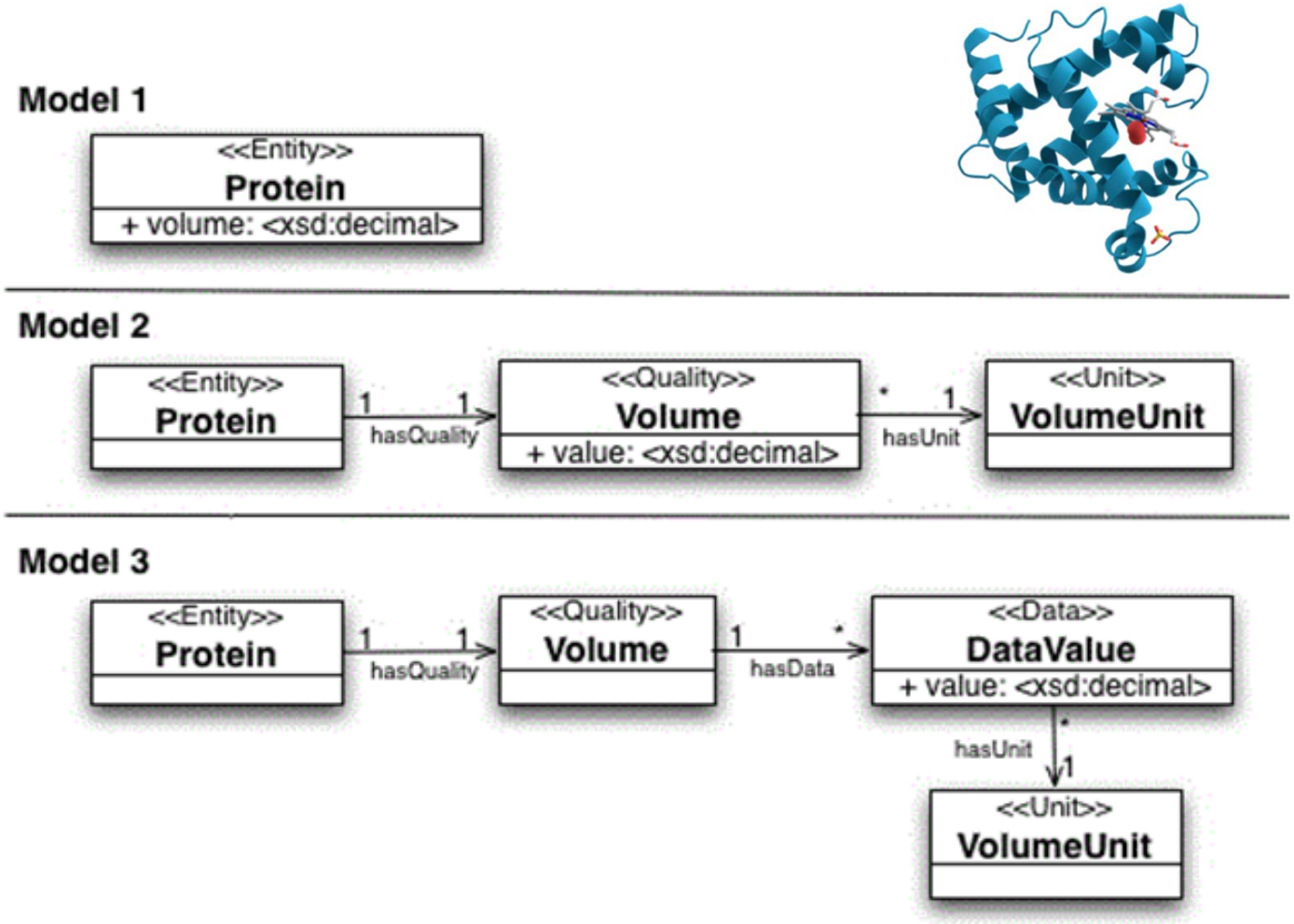
go	label	callret-2
http://bio2rdf.org/go:0044257	"cellular protein catabolic process [go:0044257]"@en	26

Bio2RDF: 2M+ SPARQL queries per month

Despite all the data, it's still hard to find answers to questions

Because there are many ways to represent the same data and each dataset represents it differently

multiple formalizations of the same kind of data do emerge, each with their own merit



Three ways to model the relationship between a protein and the volume it occupies.

Massive Proliferation of Ontologies / Vocabularies

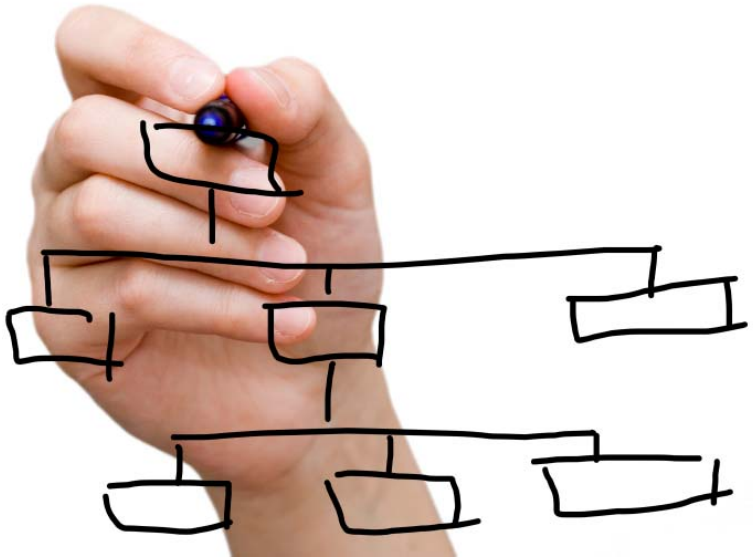
BioPortal

Statistics

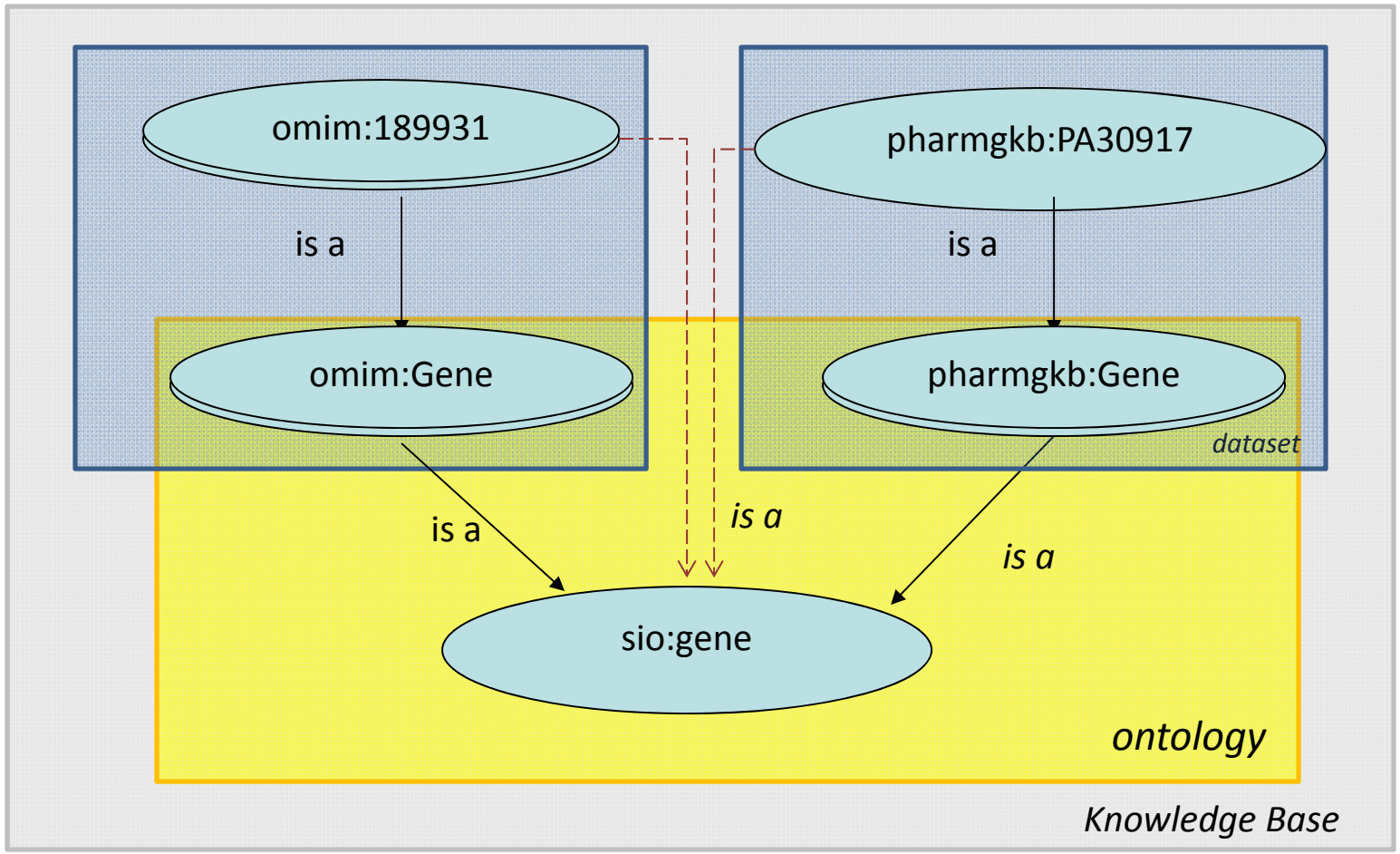
Ontologies	378
Classes	5,964,932
Resources Indexed	39
Indexed Records	5,126,145
Direct Annotations	1,883,854,337
Direct Plus Expanded Annotations	24,828,631,205

Linked Open Vocabularies (LOV)

Multi-Stakeholder Efforts to Standardize Representations are Reasonable, *Long Term Strategies* for Data Integration



Semantic data integration, consistency checking and query answering over Bio2RDF with the Semanticscience Integrated Ontology (SIO)



Querying Bio2RDF Linked Open Data with a Global Schema. Alison Callahan, José Cruz-Toledo and Michel Dumontier. Bio-ontologies 2012.

- entity
  - attribute
  - object
    - 'information content entity'
    - 'material entity'
      - 'chemical entity'
        - atom
        - 'chemical substance'
        - 'covalently connected entity'
          - molecule
            - antigen
            - catalyst
            - drug
            - isomer
            - ligand
            - 'molecular regulator'
            - 'organic molecule'
            - 'pharmaceutical component'
            - polymer
            - primer
            - product
            - 'signal transducer'
            - substrate
            - target
        - 'specialized chemical entity'
        - 'submolecular entity'
      - 'heterogeneous substance'
      - wave
    - 'spatial region'
  - process

- topObjectProperty
  - 'is related to'
    - 'has attribute'
    - 'is attribute of'
    - 'is comparable to'
      - 'is numerically comparable to'
      - 'is variant of'
        - 'is dissimilar to'
        - 'is homologous to'
        - 'is similar to'
          - 'is broader than (t)'
            - 'is broader than'
          - 'is match to'
          - 'is narrower than (t)'
        - 'is version of'
    - 'is generically related with'
    - 'is mutually related to'
    - 'is referred to by'
    - 'is spatiotemporally related to'
    - 'refers to'

SRIQ(D)
10700+ axioms
1300+ classes
201 object properties (inc. inverses)
1 datatype property

Annotations

description
"An isomer is a molecule that is compositionally identical to another molecule as a result of a different atomic connectivity."@en

label
"isomer"@en

Description: isomer

Equivalent classes

molecule
and ('is variant of' some molecule)

Superclasses

Inherited anonymous classes

'has part' some
(atom
and ('is covalently connected to' some atom))

'has component part' some 'covalent chemical bond'

'physical entity'
or 'abstract entity'

'has proper part' only 'material entity'

'has quality' some mass

'has quality' only 'physical quality'

'spatiotemporal region'
or ('is located in' some 'spatiotemporal region')

'has proper part' only 'physical entity'

'processual entity'
or 'material entity'
or region

'has part' some atom

Bio2RDF and SIO powered SPARQL 1.1 federated query: Find chemicals (from CTD) and proteins (from SGD) that participate in the same process (from GOA)

```
SELECT ?chem, ?prot, ?proc
FROM <http://bio2rdf.org/ctd>
WHERE {
    SERVICE <http://ctd.bio2rdf.org/sparql> {
      ?chemical a sio:chemical-entity.
      ?chemical rdfs:label ?chem.
      ?chemical sio:is-participant-in ?process.
      ?process rdfs:label ?proc.
    FILTER regex (?process, "http://bio2rdf.org/go:")
     }
      SERVICE <http://sgd.bio2rdf.org/sparql> {
                ?protein a sio:protein .
                ?protein sio:is-participant-in ?process.
                ?protein rdfs:label ?prot .
      }
}
```

tactical formalization

Take what you need
and represent it in a way that directly serves your objective

USER DRIVEN REPRESENTATION

identifying aberrant and pharmacological pathways

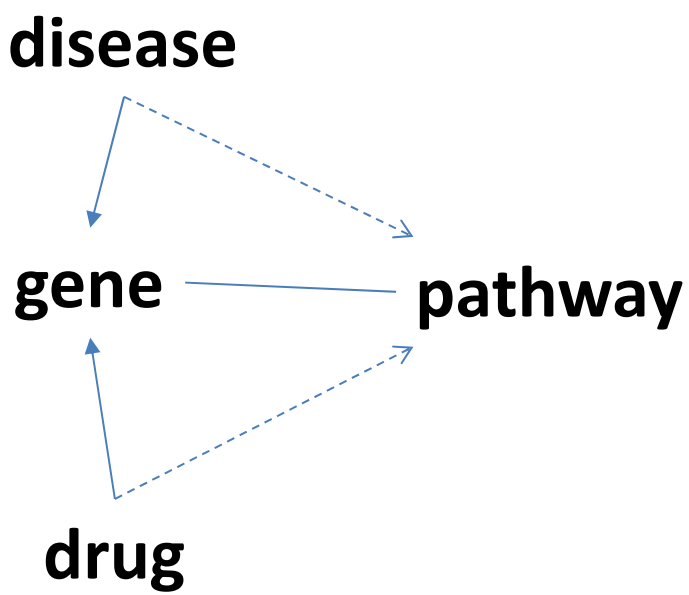
predicting drug targets using organism phenotypes

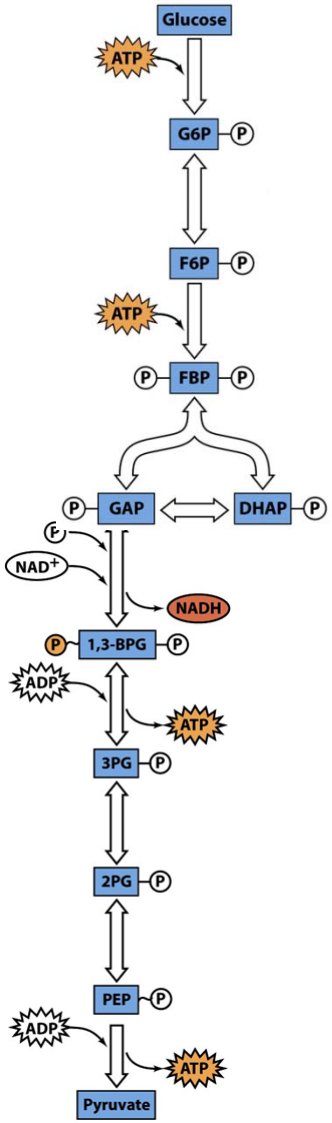
STANDARDS

Biopax-pathway exploration

FALDO-powered genome navigation

aberrant and pharmacological pathways





Q1. Can we identify pathways that are associated with a particular disease or class of diseases?

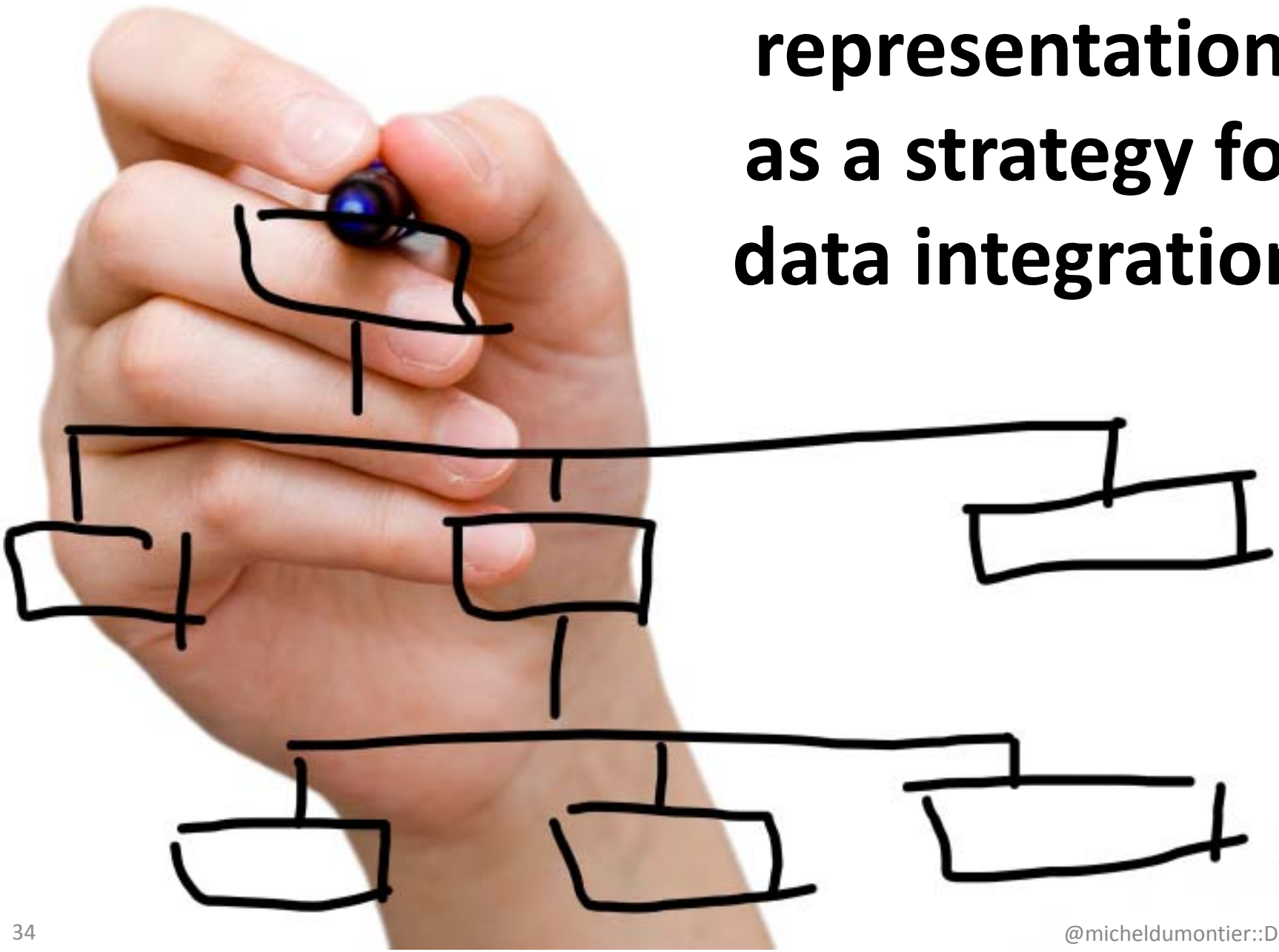
Q2. Can we identify pathways are associated with a particular drug or class of drugs?

Identification of drug and disease enriched pathways

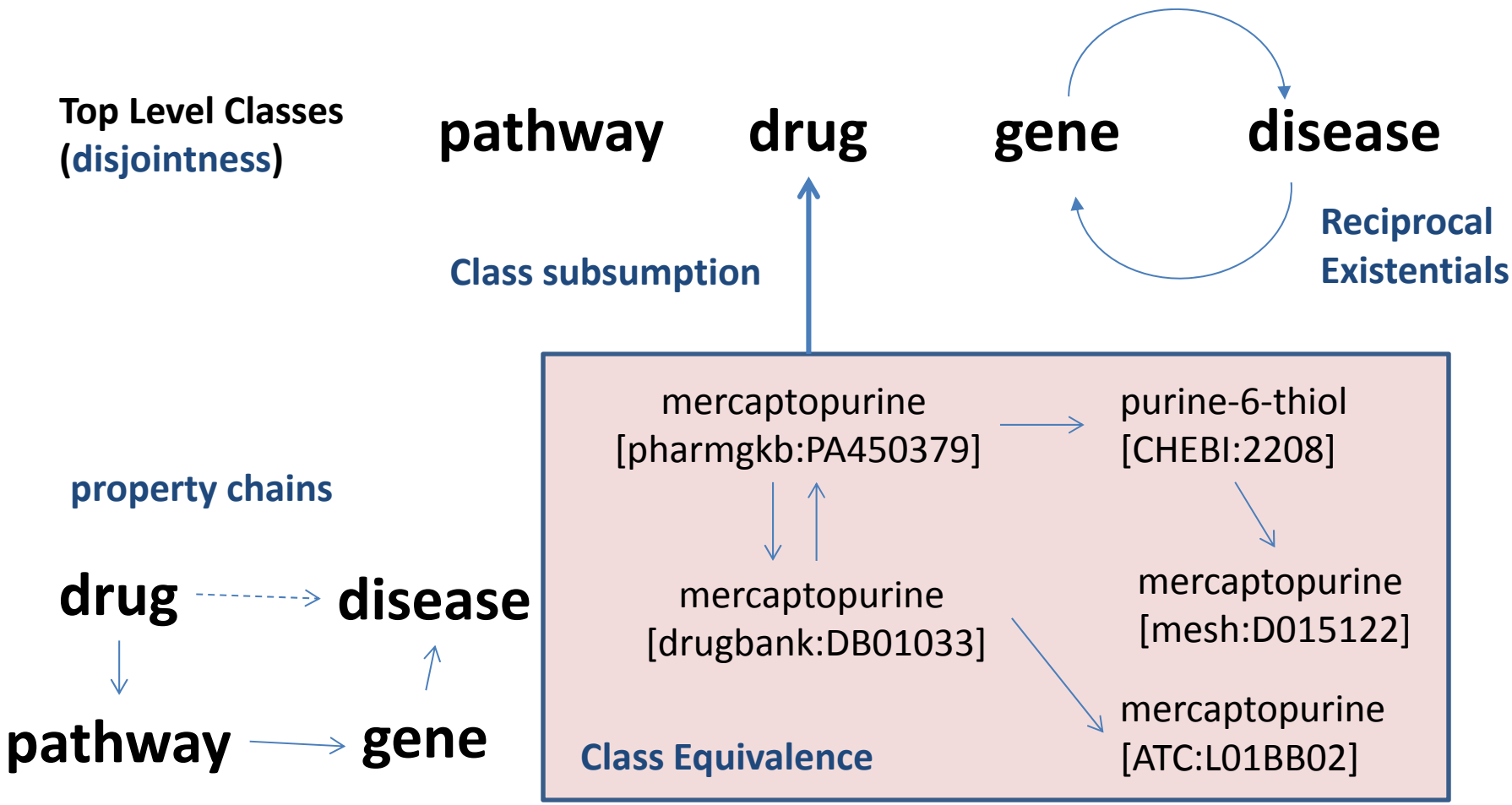
- Approach
  - Integrate 3 datasets
    - **DrugBank, PharmGKB and CTD**
  - Integrate 7 terminologies
    - **MeSH, ATC, ChEBI, UMLS, SNOMED, ICD, DO**
  - Formalize data of interest
  - Identify significant associations using enrichment analysis over the fully inferred knowledge base

Identifying aberrant pathways through integrated analysis of knowledge in pharmacogenomics. Bioinformatics. 2012.

Formal knowledge representation as a strategy for data integration



Have you heard of OWL?



Formalized as an OWL-EL ontology

650,000+ classes, 3.2M subClassOf axioms, 75,000 equivalentClass axioms

Benefits: Enhanced Query Capability

- Use <u>any</u> mapped terminology to query a target resource.
- Use knowledge in target ontologies to formulate <u>more precise questions</u>
  - ask for drugs that are associated with diseases of the joint: 'Chikungunya' (do:0050012) is defined as a viral infectious disease located in the 'joint' (fma:7490) and caused by a 'Chikungunya virus' (taxon:37124).
- Learn relationships that are <u>inferred by automated reasoning</u>.
  - alcohol (ChEBI:30879) is associated with alcoholism (PA443309) since alcoholism is directly associated with ethanol (CHEBI:16236)
  - 'parasitic infectious disease' (do:0001398) associated with 129 drugs, 15 more than are directly linked.

Knowledge Discovery through Data Integration *and* Enrichment Analysis

- **OntoFunc**: Tool to discover significant associations between sets of objects and ontology categories. enrichment of attribute among a selected set of input items as compared to a reference set. hypergeometric or the binomial distribution, Fisher's exact test, or a chi-square test.

- We found 22,653 disease-pathway associations, where for each pathway we find genes that are linked to disease.
  - **Mood disorder** (do:3324) associated with **Zidovudine Pathway** (pharmgkb:PA165859361). Zidovudine is for treating HIV/AIDS. Side effects include fatigue, headache, myalgia, malaise and anorexia

- We found 13,826 pathway-chemical associations
  - **Clopidogrel** (chebi:37941) associated with **Endothelin signaling pathway** (pharmgkb:PA164728163). Endothelins are proteins that constrict blood vessels and raise blood pressure. Clopidogrel inhibits platelet aggregation and prolongs bleeding time.

Tactical Formalization + Automated Reasoning Offers Compelling Value *for Certain Problems*

We need to be smart about the goal, and how best to achieve it. Tactical formalization is another tool in the toolbox.

We've formalized data as OWL ontologies to verify, fix and exploit Linked Data through expressive OWL reasoning

- To identify mistakes in human curated knowledge
- To identify conflicting meaning in terms
- To identify mistakes in the representation of RDF data
  - incorrect use of relations
  - incorrect assertion of identity (owl:sameAs)

Many other applications can be envisioned.

PhenomeDrug

A computational approach to predict drug targets, drug effects, and drug indications using phenotypes

Mouse model phenotypes provide information about human drug targets.
Hoehndorf R, Hiebert T, Hardy NW, Schofield PN, Gkoutos GV, **Dumontier M**. Bioinformatics. 2013.

animal models provide insight for *on target* effects

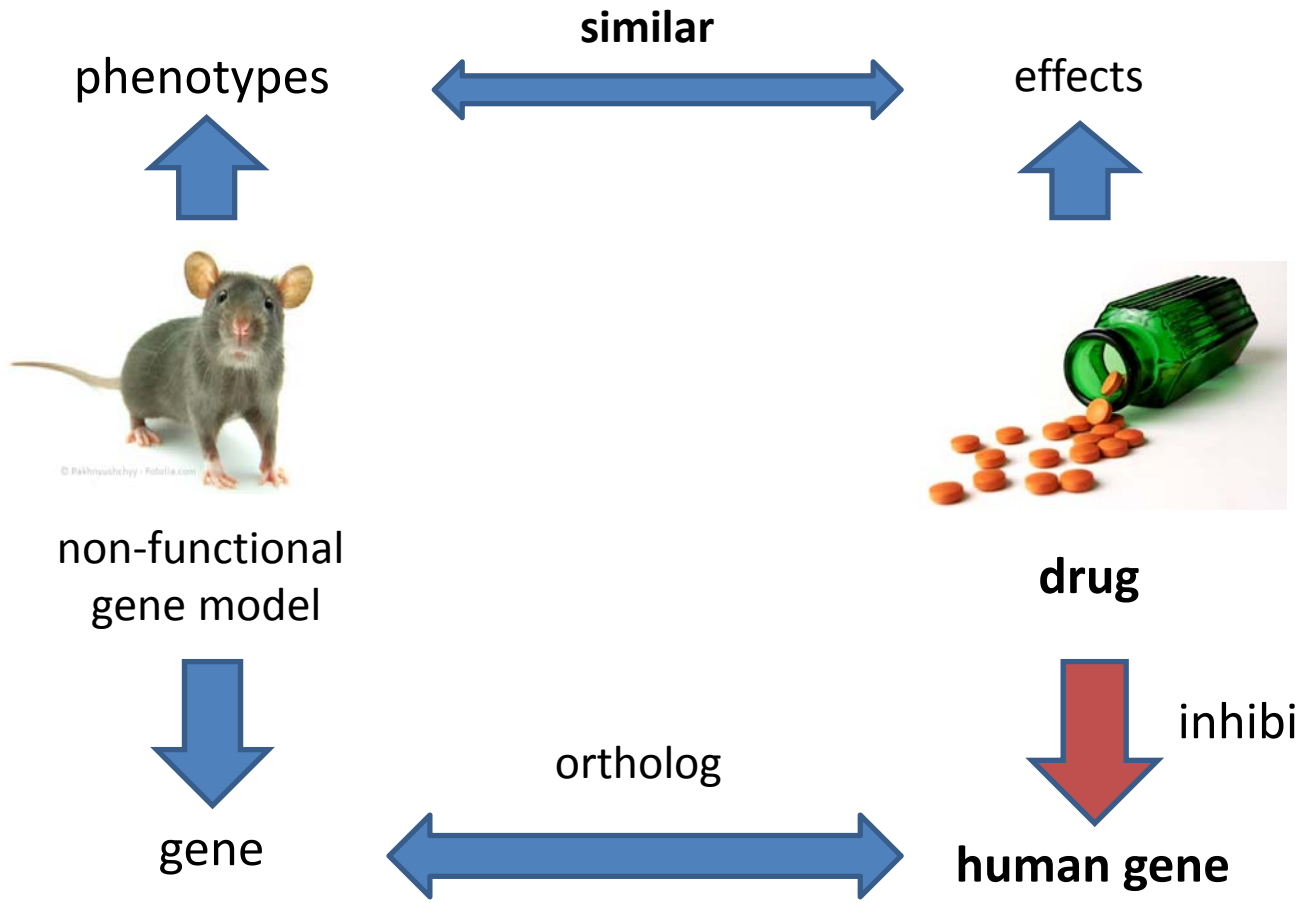
- In the majority of 100 best selling drugs ($148B in US alone), there is a direct correlation between knockout phenotype and drug effect

- **Immunological Indications**
  - Anti-histamines (Claritin, Allegra, Zyrtec)
  - KO of histamine H_1 receptor leads to decreased responsiveness of immune system
  - Predicts *on target* effects : drowsiness, reduced anxiety

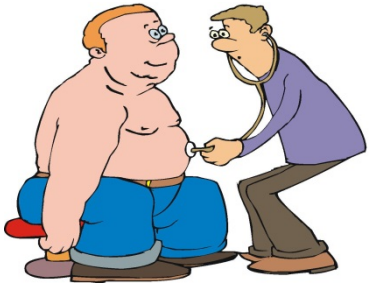
Zambrowicz and Sands. Nat Rev Drug Disc. 2003.

Identifying drug targets from mouse knock-out phenotypes

Main idea: if a drug's phenotypes matches the phenotypes of a null model, this suggests that the drug is an inhibitor of the gene



GENOTYPE

kcnj11$^{c14/c14}$; insr$^{t143/+}$(AB)	B6.Cg-Alms1$^{foz/fox}$/J	ALSM1(NM_015120.4) [c.10775delC] + [-] ?
		
increased weight, adipose tissue volume, glucose homeostasis altered	increased food intake, hyperglycemia, insulin resistance	obesity, diabetes mellitus, insulin resistance

PHENOTYPE

Terminological Interoperability
(we must compare apples with apples)

Mouse Phenotypes

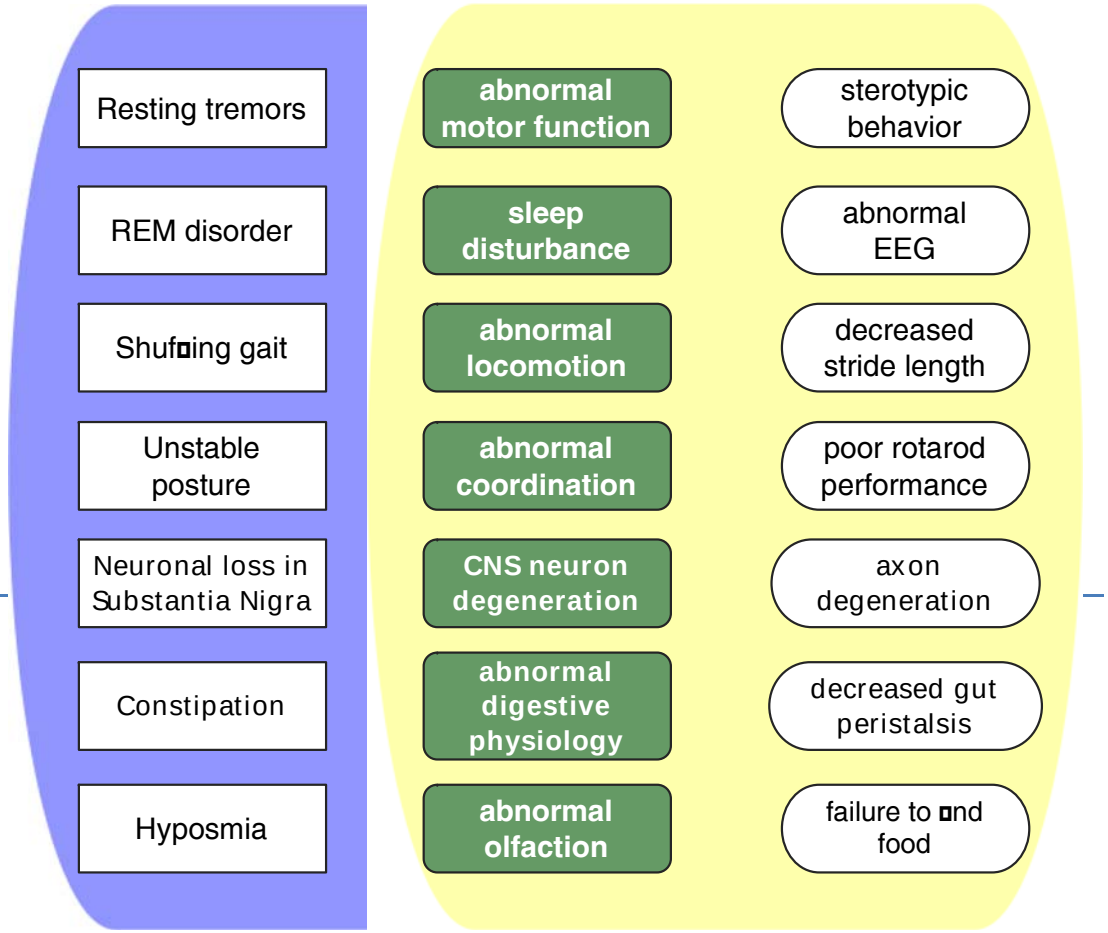
Mammalian Phenotype Ontology

PhenomeNet

PhenomeDrug

Drug effects

(mappings from UMLS to DO, NBO, MP)



Semantic Similarity

Given a drug effect profile D and a mouse model M, we compute the semantic similarity as an information weighted Jaccard metric.

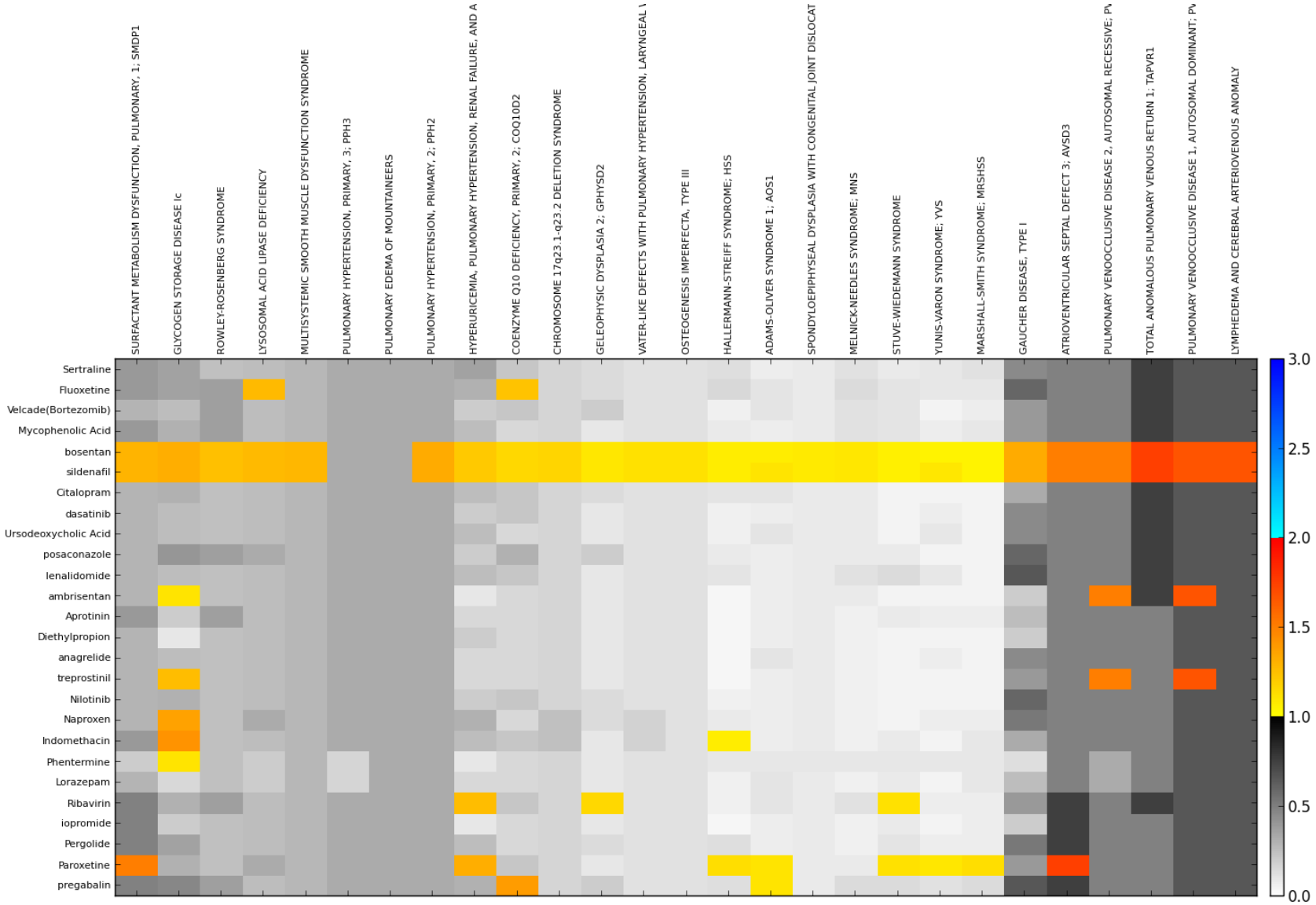
$$sim(D, M) = \frac{\sum_{x \in Cl(D) \cap Cl(M)} IC(x)}{\sum_{y \in Cl(D)} IC(y)}$$

The similarity measure used is non-symmetrical and determines the amount of information about a drug effect profile D that is covered by a set of mouse model phenotypes M.

Loss of function models predict targets of inhibitor drugs

- 14,682 drugs; 7,255 mouse genotypes
- Validation against known and predicted inhibitor-target pairs
  - 0.76 ROC AUC for human targets (DrugBank)
  - 0.81 ROC AUC for mouse targets (STITCH)
- diclofenac (STITCH:000003032)
  - NSAID used to treat pain, osteoarthritis and rheumatoid arthritis
  - Drug effects include liver inflammation (hepatitis), swelling of liver (hepatomegaly), redness of skin (erythema)
  - 49% explained by PPARg knockout
    - peroxisome proliferator activated receptor gamma (PPARg) regulates metabolism, proliferation, inflammation and differentiation,
    - Diclofenac is a known inhibitor
  - 46% explained by COX-2 knockout
    - Diclofenac is a known inhibitor

Phenotype-Based Drug Repurposing



Using the Semantic Web to Gather Evidence for Scientific Hypotheses

What evidence supports or disputes that TKIs are cardiotoxic?

FDA Use Case: TKI non-QT Cardiotoxicity

- **Tyrosine Kinase Inhibitors (TKI)**
  - Imatinib, Sorafenib, Sunitinib, Dasatinib, Nilotinib, Lapatinib
  - Used to treat cancer
  - Linked to cardiotoxicity.
- **FDA** launched drug safety program to detect toxicity
  - Need to integrate data and ontologies (Abernethy, CPT 2011)
  - Abernethy (2013) suggest using public data in genetics, pharmacology, toxicology, systems biology, to predict/validate adverse events
- What evidence could we gather to give credence that TKI's causes non-QT cardiotoxicity?

Database or knowledge base	URL
SIDER (computer-readable side effect resource)	http://sideeffects.embl.de
DrugBank	http://www.drugbank.ca
Chemical Effects in Biological Systems (CEBS)	http://cebs.niehs.nih.gov/
NCBI Database of Genotypes and Phenotypes (dbGaP)	http://www.ncbi.nlm.nih.gov/gap/
Comparative Toxicogenomics Database	http://ctd.mdibl.org/
Genetic Association Database (archive of human genetic association studies of complex diseases and disorders)	http://geneticassociationdb.nih.gov
Kyoto Encyclopedia of Genes and Genomes (KEGG) (bioinformatics resource for linking genomics to life)	http://www.genome.jp/kegg
The Pharmacogenomics Knowledgebase (PharmGKB) (resource describing how variation in human genetics leads to variation in response to drugs)	http://www.pharmgkb.org
Gene Expression Omnibus (GEO) (database repository of high-throughput gene expression data and hybridization arrays, chips, and microarrays)	http://www.ncbi.nlm.nih.gov/geo
Connectivity Map (detailed map that links gene patterns associated with disease to corresponding patterns produced by drug candidates and a variety of genetic manipulations)	http://www.broadinstitute.org/genome_bio/connectivitymap.html
The Gene Ontology (GO) (standardized representation of gene and gene product attributes across species and databases)	http://www.geneontology.org
Tox21 (Computational Toxicology Research program)	http://epa.gov/ncct/Tox21
International HapMap Project (database of genes associated with human disease and response to pharmaceuticals)	http://hapmap.ncbi.nlm.nih.gov
Human Interactome Database (database of human binary protein-protein interaction networks)	http://interactome.dfci.harvard.edu/H_sapiens
European Bioinformatics Institute (EBI) ArrayExpress Archive	http://www.ebi.ac.uk/microarray-as/ae/
NCI-60 DTP Human Tumor Cell Line Screen	http://dtp.nci.nih.gov/branches/btb/ivclsp.html
Library of Integrated Network-Based Cellular Signatures (LINCS)	http://commonfund.nih.gov/lincs/
Reactome	http://www.reactome.org/ReactomeGWT/entrypoint.html
Online Mendelian Inheritance in Man®	http://www.ncbi.nlm.nih.gov/omim

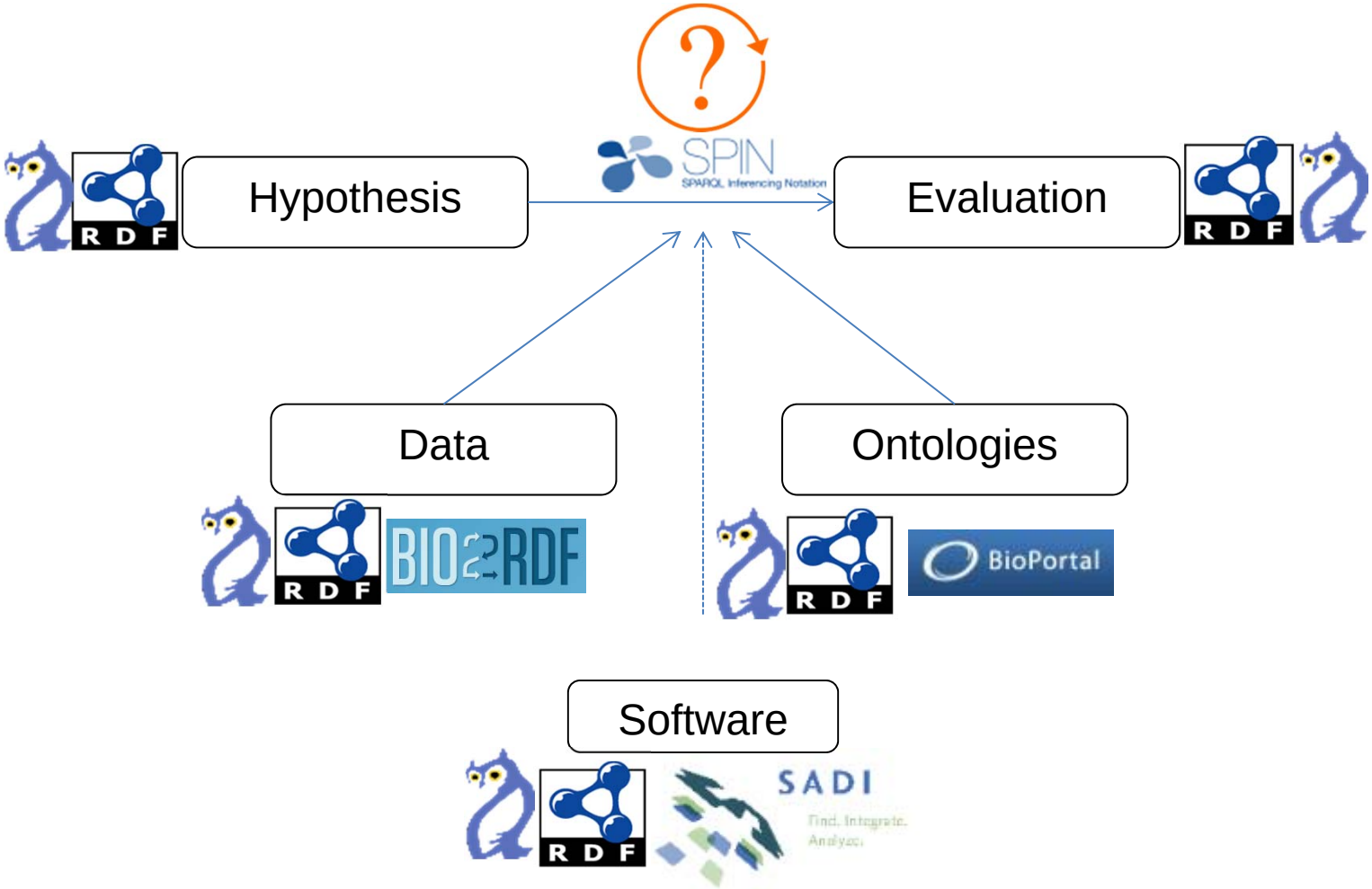
Jane P.F. Bai and Darrell R. Abernethy. Systems Pharmacology to Predict Drug Toxicity: Integration Across Levels of Biological Organization. Annu. Rev. Pharmacol. Toxicol. 2013.53:451-473

HyQue

- The goal of HyQue is retrieve and evaluate evidence that supports/disputes a hypothesis
  - hypotheses are described as a set of events
    - e.g. binding, inhibition, phenotypic effect
  - events are associated with types of **evidence**
    - a query is written to retrieve **data**
    - a weight is assigned to provide **significance**
- Hypotheses are written by people who seek answers
- data retrieval rules are written by people who know the data and how it should be interpreted ☺

1. HyQue: Evaluating hypotheses using Semantic Web technologies. J Biomed Semantics. 2011 May 17;2 Suppl 2:S3.
2. Evaluating scientific hypotheses using the SPARQL Inferencing Notation. Extended Semantic Web Conference (ESWC 2012). Heraklion, Crete. May 27-31, 2012.

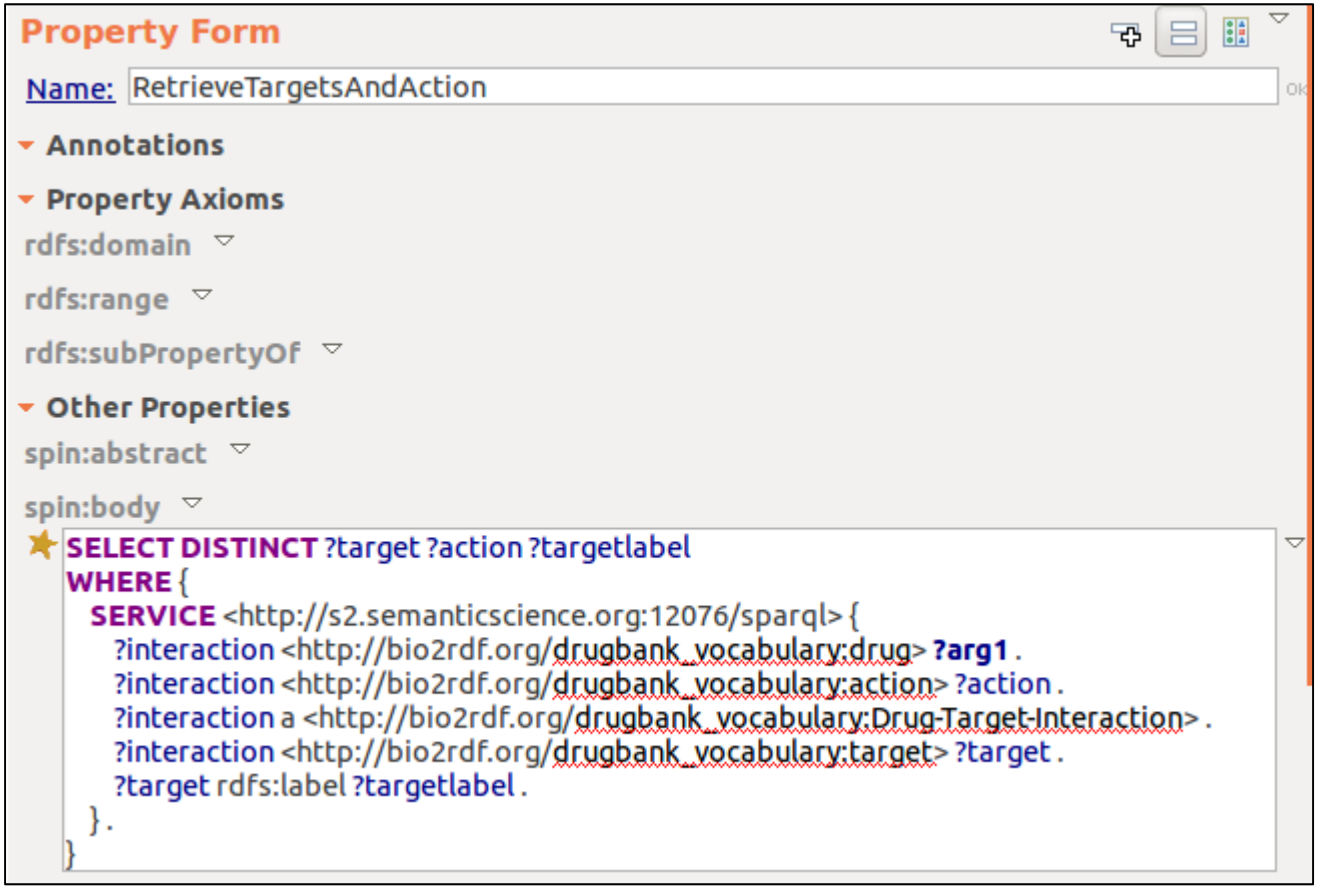
HyQue: A Semantic Web Application



What evidence might we gather?

- clinical: Are there cardiotoxic effects associated with the drug?
  - Literature (*studies*) [curated db]
  - Product labels (*studies*) [**r3:sider**]
  - Clinical trials (*studies*) [**r3:clinicaltrials**]
  - Adverse event reports [**r2:pharmgkb/onesides**]
  - Electronic health records (*observations*)
- pre-clinical associations:
  - genotype-phenotype (null/disease models) [**r2:mgi, r2:sgd; r3:wormbase**]
  - *in vitro* assays (IC50) [**r3:chembl**]
  - drug targets [**r2:drugbank; r2:ctd; r3:stitch**]
  - drug-gene expression [**r3:gxa**]
  - pathways [**r2:kegg**; **r3:reactome**]
  - Drug-pathway, disease-pathway enrichments [aberrant pathways]
  - Chemical properties [**r2:pubchem; r2.drugbank**]
  - Toxicology [**r1.toxkb/cebs**]

Data retrieval is done with SPARQL



Data Evaluation is done with SPIN rules

Class Form

Name: IsRAF1orPDGFRsorVEGFRorAMPKTargetAndIsActionInhibit

Annotations

rdfs:comment

Targets IDs in drugnbank: RAF1=40, PDGFR alpha=950, PDGFR beta=228, VEGFR=32

Class Axioms

rdfs:subClassOf

spin:Functions

Other Properties

spin:abstract

spin:body

```
ASK WHERE {
  ( ?adrug ) :RetrieveTargetsAndAction ( ?target ?action ) .
  FILTER ((((?target = <http://bio2rdf.org/drugbank.target:40>) || (?target = <http://
bio2rdf.org/drugbank.target:950>)) || (?target = <http://bio2rdf.org/
drugbank.target:228>)) || ((?target = <http://bio2rdf.org/drugbank.target:32>) &&
(xsd:string(?action) = "inhibitor"))) .
}
```

spin:constraint

Argument **arg:adrug** :

Describe your hypothesis

Hypothesis title

Sunitinib has cardiotoxic effects

e.g. TKI hypothesis 1

Hypothesis description

e.g. drug A causes cardiotoxicity

Author

A Callahan

e.g. John Doe

Source of hypothesis [if applicable]

20376335

e.g. PMID:11340206

Event label

Sunitinib is an agent in a cardiotoxicity event

e.g. "Drug A is an agent in a cardiotoxicity event"

Event type *

Cardiotoxicity

Is your hypothesis that the event does NOT occur? For example, that drug A does NOT cause cardiotoxicity? *

- ○ Yes
- ◉ No

Agent *

http://bio2rdf.org/drugbank:DB01268

Search by entity name, e.g. "imatinib" or using namespace:identifier, e.g. "drugbank:DB01268"

Target

e.g. gene X, or chebi:28260

Perturbation context

e.g. gal3p-ovexp

Reset

Add another event | Next >>

http://bio2rdf.org/drugbank:DB01268

Overall hypothesis evaluation: HYPOTHESIS SUPPORTED

Evidence summary for hypothesis_20131115091904_e1

Evidence type	Evaluation
Known drug side effects	SUPPORTS HYPOTHESIS
TUNEL assay results	NEUTRAL
Literature-sourced drug side effects	SUPPORTS HYPOTHESIS
hERG inhibition	NEUTRAL
Literature-sourced drug targets	SUPPORTS HYPOTHESIS
Known cardiotoxicity assays	NEUTRAL
Known gene targets and associated mouse model phenotypes	NEUTRAL
Known drug targets and effects	SUPPORTS HYPOTHESIS

Literature-sourced drug side effects

Side effect	Source article
arterial thrombosis [umls:C0151942]	PUBMED:20351323
congestive heart failure [umls:C0018802]	PUBMED:17457301
congestive heart failure [umls:C0018802]	PUBMED:19734999
congestive heart failure [umls:C0018802]	PUBMED:21283106
ejection fraction decreased [umls:C0743400]	PUBMED:19734999
hypertension [umls:C0020538]	PUBMED:19734999
myocardial infarction [umls:C0027051]	PUBMED:19734999

Literature-sourced side effects retrieval query

```
SELECT DISINCT ?effect ?effect_article ?effectlabel
WHERE {
SERVICE <http://s2.semanticscience.org:12084/sparql> {
?drug a <http://bio2rdf.org/cardiotox_vocabulary:Drug> .
?drug <http://bio2rdf.org/cardiotox_vocabulary:hasCardiotoxicEffect> ?effect .
SERVICE <http://s3.semanticscience.org:12074/sparql> {
?effects rdfs:label ?effectlabel .
} .
?effects <http://bio2rdf.org/cardiotox_vocabulary:hasArticle> ?effect_article .
} .
}
```

BioAssay data (source: CHEMBL)

No hERG inhibition IC50 results found.

IC50 assay data retrieval query

```
SELECT DISTINCT ?stdValue ?assay ?target ?targetlabel
WHERE {
SERVICE <http://s2.semanticscience.org:12076/sparql> {
?drug rdfs:seeAlso ?dbid .
SERVICE <http://www.ebi.ac.uk/rdf/services/chembl/sparql> { .
?activity a <http://rdf.ebi.ac.uk/terms/chembl#Activity> .
?activity <http://rdf.ebi.ac.uk/terms/chembl#hasMolecule> ?chemblmolecule .
?chemblmolecule <http://rdf.ebi.ac.uk/terms/chembl#moleculeXref> ?dbid .
?activity <http://rdf.ebi.ac.uk/terms/chembl#hasAssay> ?assay .
?activity <http://rdf.ebi.ac.uk/terms/chembl#standardValue> ?stdValue .
?activity <http://rdf.ebi.ac.uk/terms/chembl#standardType> "IC50" .
?assay <http://rdf.ebi.ac.uk/terms/chembl#hasTarget> ?target .
?target rdfs:label ?targetlabel .
FILTER (?target = <http://rdf.ebi.ac.uk/resource/chembl/target/CHEMBL240>) .
} .
} .
}
```

No TUNEL assay results found.

TUNEL assay data retrieval query

```
SELECT DISTINCT ?value ?assay
WHERE {
SERVICE <http://s2.semanticscience.org:12076/sparql> {
?drug rdfs:seeAlso ?dbid .
SERVICE <http://www.ebi.ac.uk/rdf/services/chembl/sparql> {
?activity a <http://rdf.ebi.ac.uk/terms/chembl#Activity> .
?activity <http://rdf.ebi.ac.uk/terms/chembl#hasMolecule> ?chemblmolecule .
?chemblmolecule <http://rdf.ebi.ac.uk/terms/chembl#moleculeXref> ?dbid .
?activity <http://rdf.ebi.ac.uk/terms/chembl#hasAssay> ?assay .
?assay <http://purl.org/dc/terms/description> ?description .
?activity <http://rdf.ebi.ac.uk/terms/chembl#publishedValue> ?value .
?activity <http://www.bioassayontology.org/bao#BAO_0000208> <http://www.bioassayontology.org/bao#BAO_0001103> .
```

In Summary

- This talk was about making sense and using the structured data we already have
- RDF-based Linked Open Data acts as a substrate for query answering and task-based formalization in OWL
- Discovery through the generation of testable hypotheses in the target domain.
- Using Linked Data to evaluate scientific hypotheses

Looking to the Future

- Community **guidelines** for RDF-based data and dataset descriptions (e.g. CEDAR)
- Alignment and **consolidation** of OWL ontologies (e.g. UMLS)
- Identifying and filling **gaps** in our knowledge (e.g. Adam the Robot scientist)
- Improving our **coverage** of available evidence (e.g. HyQue)
- More sophisticated data **mining** (e.g. you!)

Acknowledgements

Bio2RDF Release 2:

Allison Callahan, Jose Cruz-Toledo, Peter Ansell

Aberrant Pathways: Robert Hoehndorf, Georgios Gkoutos

PhenomeDrug: Tanya Hiebert, Robert Hoehndorf, Georgios Gkoutos, Paul Schofield

TKI Cardiotoxicity: Alison Callahan, Tania Hiebert, Beatriz Lujan, Sira Sarntivijai (FDA)

dumontierlab.com

michel.dumontier@stanford.edu

Website: http://dumontierlab.com

Presentations: http://slideshare.com/micheldumontier