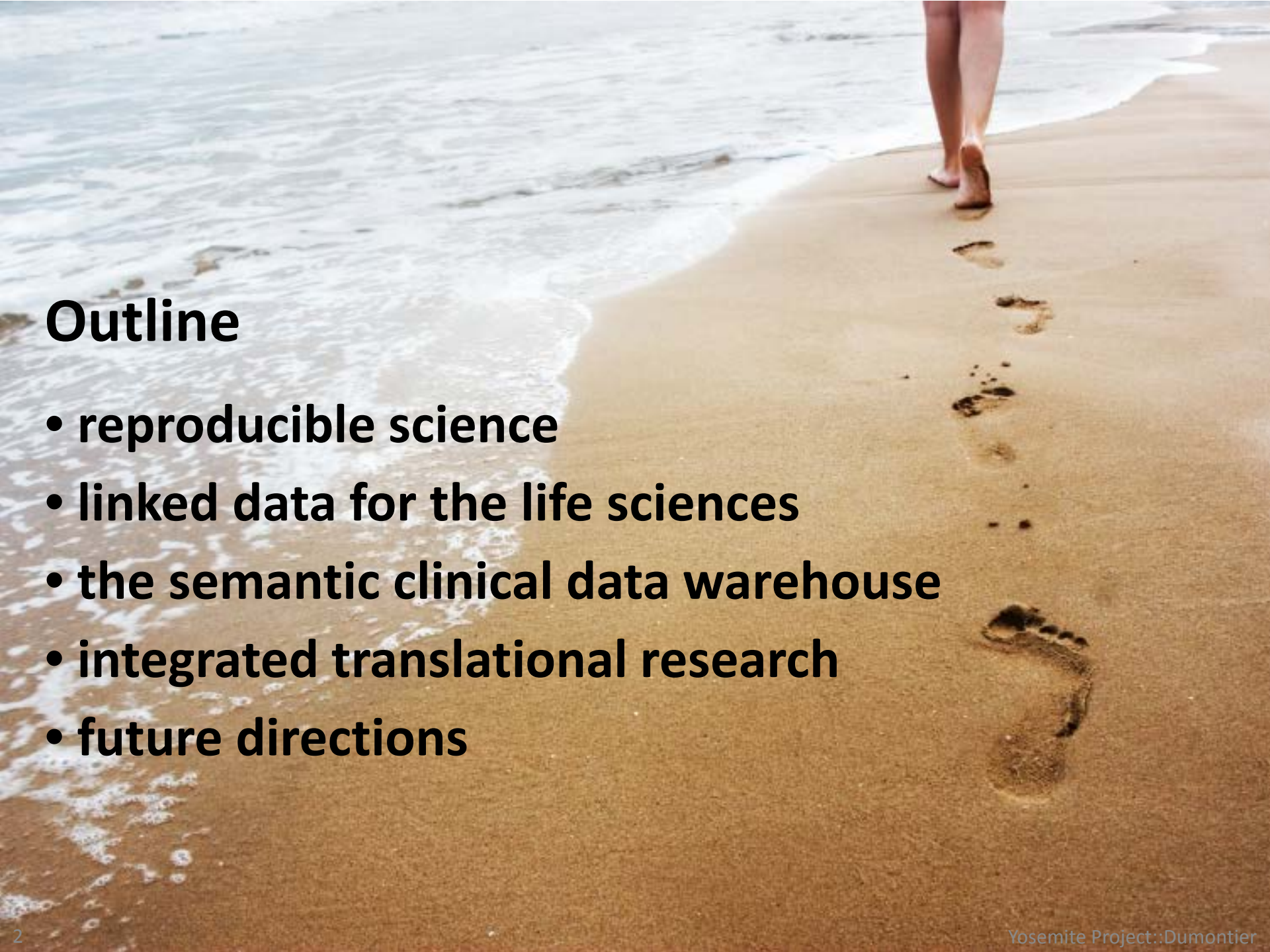
Data-Driven Biomedical Research With Semantic Web Technologies

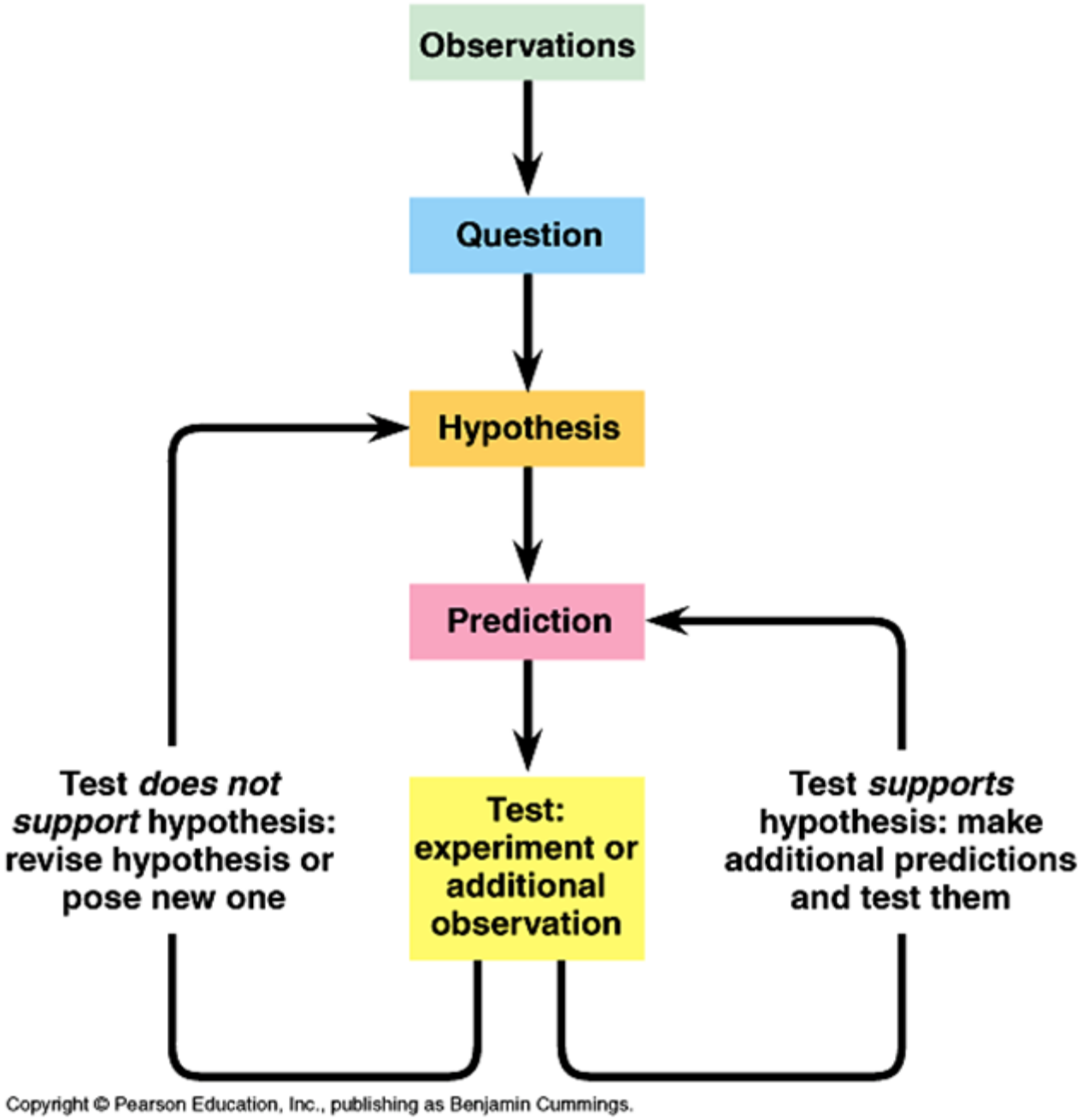
Michel Dumontier, Ph.D.

Associate Professor of Medicine (Biomedical Informatics)
Stanford University



Outline

- reproducible science
- linked data for the life sciences
- the semantic clinical data warehouse
- integrated translational research
- future directions



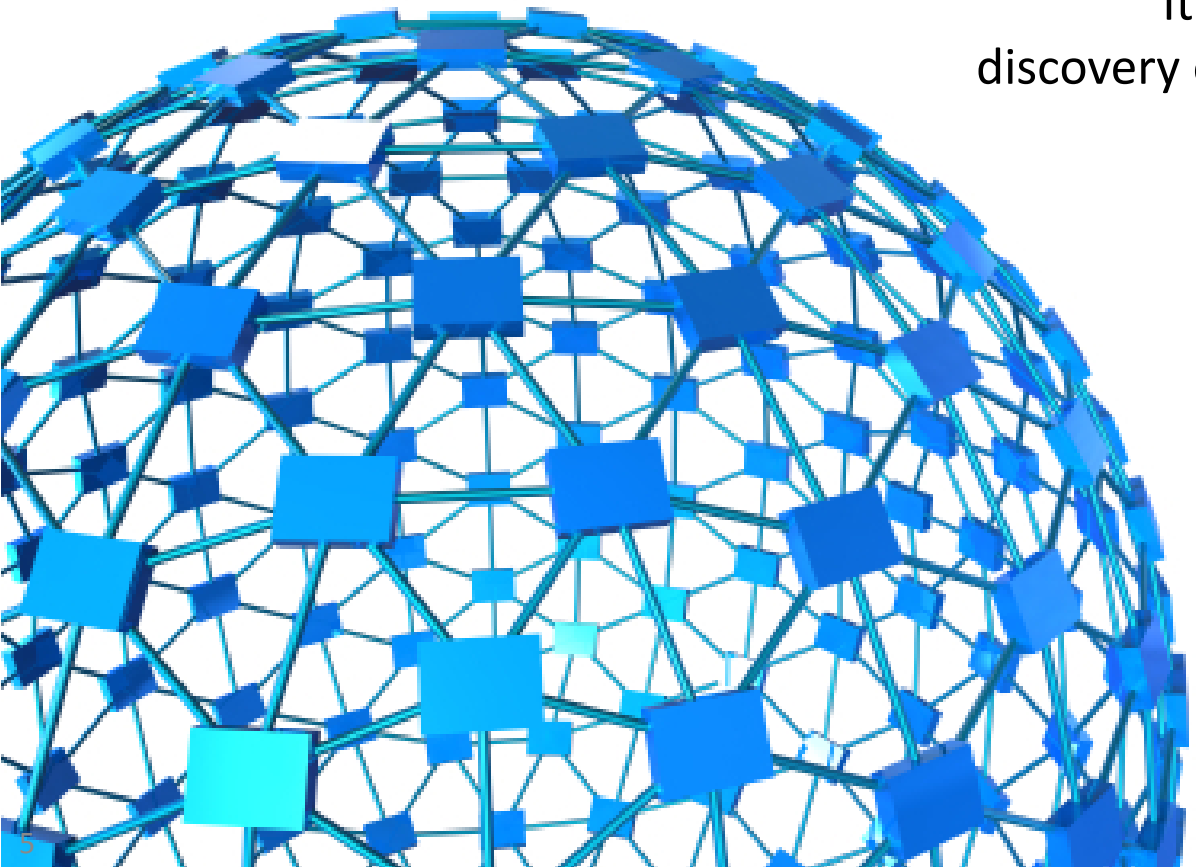
Scientists need to find evidence to support/refute a hypothesis which is, surprisingly, increasingly challenging with more data

need to know where to look,
understand the nature
and structure of data
and how to process it

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

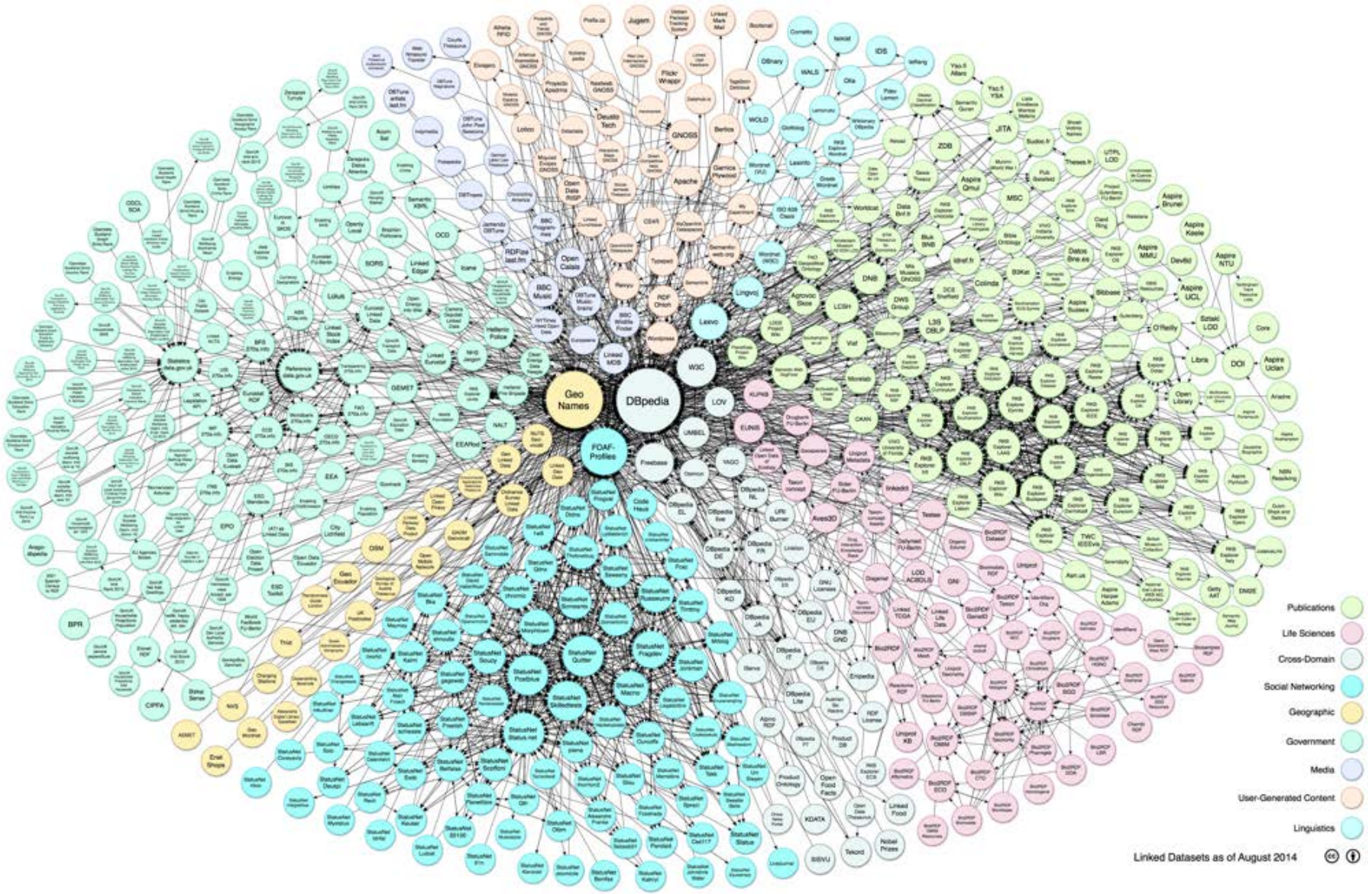
It involves standards for publishing, sharing and querying **facts, expert knowledge and services**

It is a scalable approach to the discovery of *independently formulated* and *distributed* knowledge



W3C Semantic Web

We are building a massive network of linked open data

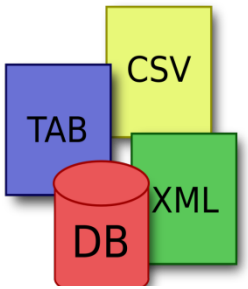


Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/"

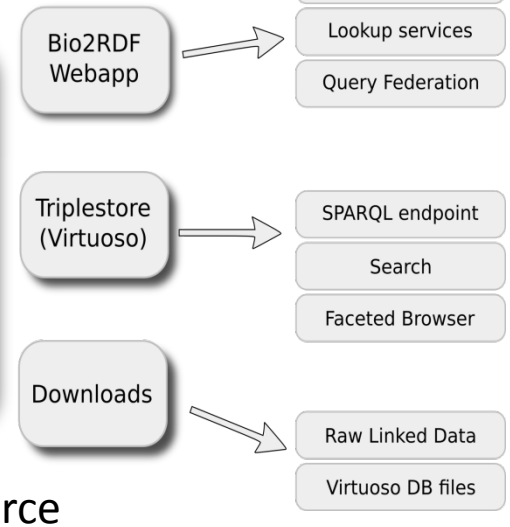
BIO2RDF

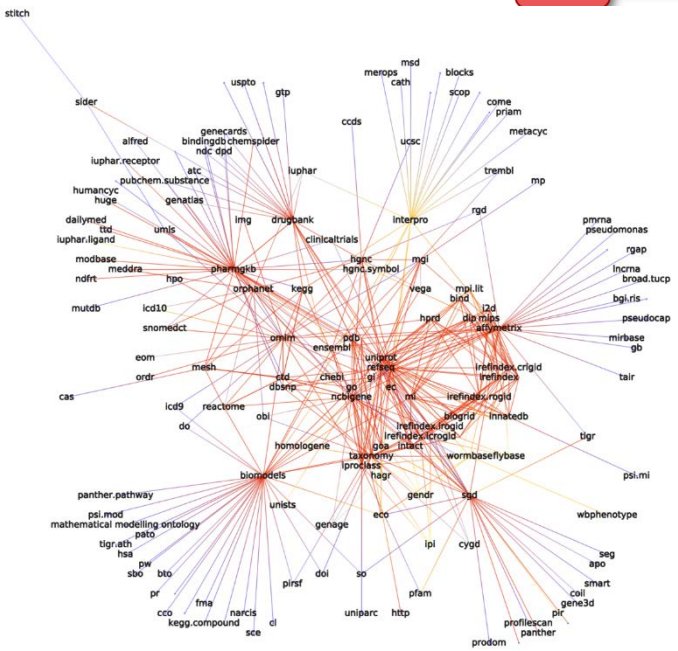
Linked Data for the Life Sciences

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



Conversion Scripts





- Free and open source
- Leverages Semantic Web standards
- **10B+** interlinked statements from **30+** conventional and high value datasets
- Partnerships with EBI, SIB, NCBI, DBCLS, NCBO, OpenPHACTS, and many others

Alison Callahan, Jose Cruz-Toledo, Peter Ansell, Michel Dumontier: Bio2RDF Release 2: Improved Coverage, Interoperability and Provenance of Life Science Linked Data. ESWC 2013: 200-212

HyQue

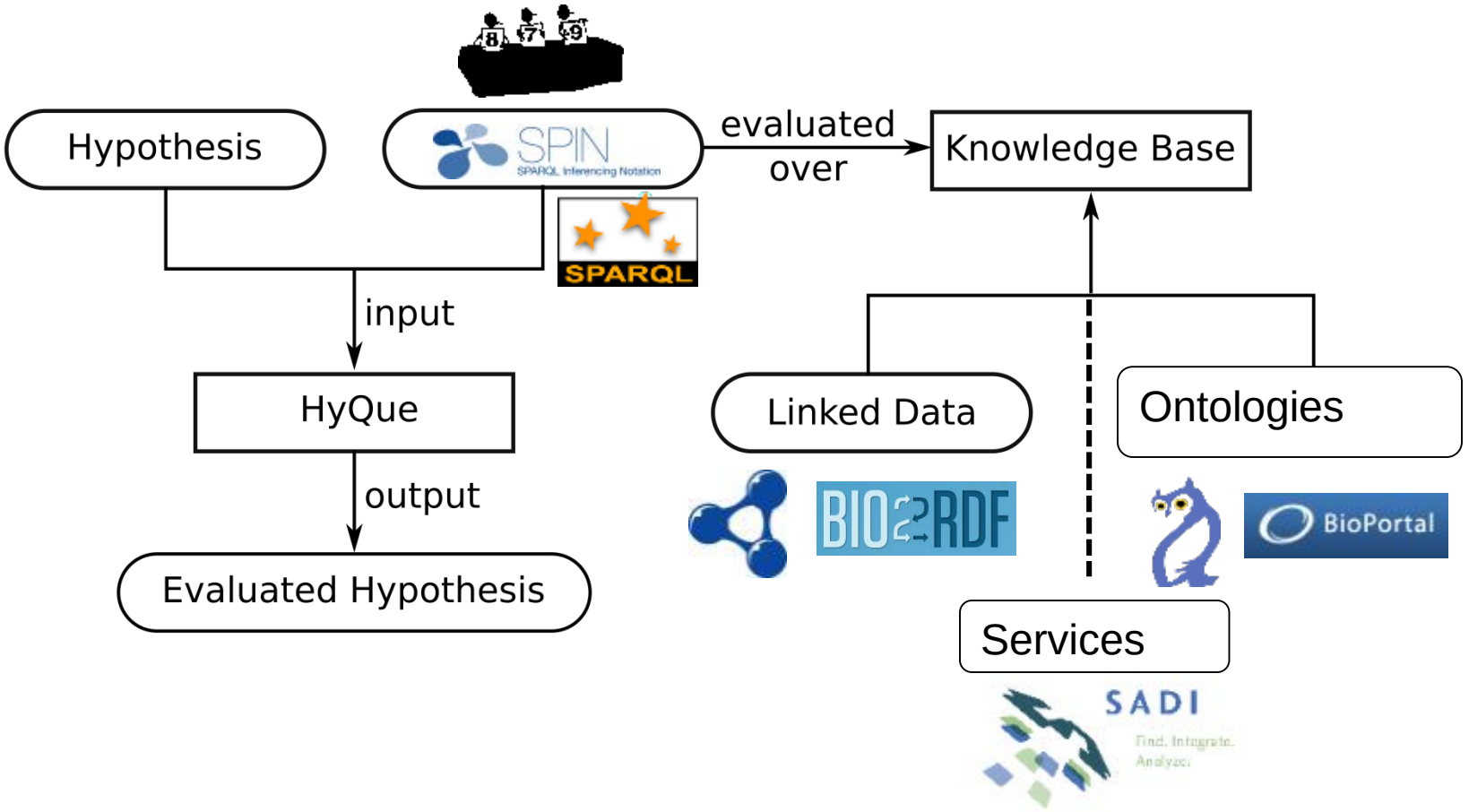
HyQue is the Hypothesis query and evaluation system

- A platform for knowledge discovery
- Facilitates hypothesis formulation and evaluation
- Leverages Semantic Web technologies to provide access to facts, expert knowledge and web services
- Pervasive Provenance
- Reproducible evaluation against positive and negative findings
- Transparent evidence weighting

HyQue: evaluating hypotheses using Semantic Web technologies. J Biomed Semantics. 2011 May 17;2 Suppl 2:S3.

Evaluating scientific hypotheses using the SPARQL Inferencing Notation. Extended Semantic Web Conference (ESWC 2012). Heraklion, Crete. May 27-31, 2012.

HyQue is a Semantic Web Application that uses RDF, OWL, SPARQL, SPIN, and SADI



FDA Use Case: TKI Cardiotoxicity

- FDA launched drug safety program to detect toxicity
  - Need to integrate data and ontologies (Abernethy, CPT 2011)
  - Development of organ-specific predictions (e.g. cardiotoxicity)
- Tyrosine Kinase Inhibitor
  - Imatinib, Sorafenib, Sunitinib, Dasatinib, Nilotinib, Lapatinib
  - Used to treat cancer
  - Recently linked to cardiotoxicity.
- Abernethy & Bai (2013) suggest using public data in genetics, pharmacology, toxicology, systems biology, to predict/validate adverse events

BIO2RDF

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

Gather the Evidence

- clinical: Are there cardiotoxic effects associated with the drug?
  - Past, current or planned Clinical trials (*studies*)
  - Product labels (*studies*)
  - Literature (*studies*)
  - Electronic health records (*observations*)
  - Adverse event reports (*reports*)
- pre-clinical:
  - *in vitro* assays
  - TUNEL assay (detects DNA fragmentation that results from apoptotic signaling cascades)
  - key targets: RAF1, PDGFR,VEGFR, AMPK or hERG?
  - Animal models of drug action, of disease
  - GWAS, Gene Expression data

Describe an event

Event label

Sunitinib is an agent in a cardiotoxicity event

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

Event ty

Cardic

Is your h

Yes

No

Agent *

http://b

Search b

Overall hypothesis evaluation: HYPOTHESIS SUPPORTED

Evidence summary

Evidence ty
Known drug
Known drug
hERG inhibiti
hERG Non-in
Known cardio
TUNEL assa

Drug action on known targets (source: DRUGBANK)

Target	Action
Mast/stem cell growth factor receptor [drugbank.target:504]	antagonist

Drug side effects (source: SIDER)

Side Effect
infection [umls:C0021311]
hypertension [umls:C0020538]
increased sgot [umls:C0151904]
increased sgpt [umls:C0151905]
bleeding [umls:C0019080]
flatulence [umls:C0016204]
dry mouth [umls:C0043352]
mouth ulcer [umls:C0149745]

Evidence-Based Approach: Cardiotoxicity

TKI	Score	Our classification of cardiotoxicity based on the score	Known cardiotoxicity based on Chen *et al*.	Confidence cardiotoxicity based on Chen *et al*.
dasatinib	0.50	Intermediate	Yes	low-moderate
erlotinib	0.22	Weak	No	N/A
gefitinib	0.22	Weak	No	N/A
imatinib	0.63	Strong	Yes	low
lapatinib	0.12	Weak	No	N/A
nilotinib	0.33	Intermediate	Yes	low
sorafenib	0.52	Intermediate	Yes	low
sunitinib	0.48	Intermediate	Yes	moderate

Evidence-Based Approach: Aging

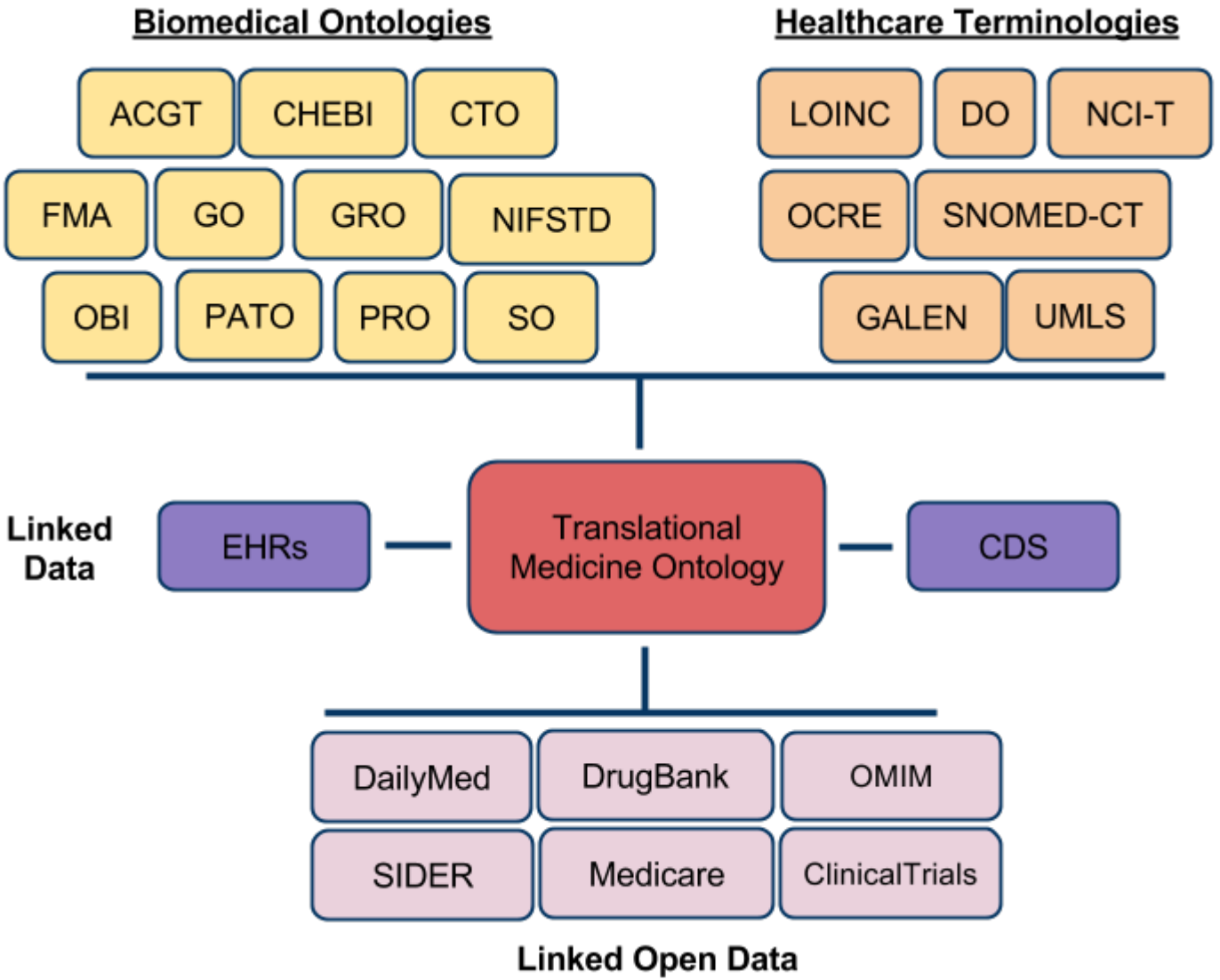
WormBase ID	Symbol	Score	PMID	Satisfied data evaluation function								
				DEF1	DEF2	DEF3	DEF4	DEF5	DEF6	DEF7	DEF8	DEF9
WBGene00008205	sams-1	0.89	16103914	✓	✓	✓	✓	✓	✓		✓	✓
WBGene00000371	cco-1	0.78	21215371	✓	✓			✓	✓	✓	✓	✓
WBGene00009741	drr-1	0.78	16103914	✓	✓		✓	✓	✓		✓	✓
WBGene00002178	jnk-1	0.78	15767565	✓	✓			✓	✓	✓	✓	✓
WBGene00004013	pha-4	0.78	19239417		✓		✓	✓	✓	✓	✓	✓
WBGene00004789	sgk-1	0.78	15068796	✓	✓			✓	✓	✓	✓	✓
WBGene00004800	sir-2.1	0.78	21938067	✓			✓	✓	✓	✓	✓	✓
WBGene00006796	unc-62	0.78	17411345	✓	✓			✓	✓	✓	✓	✓

Translational Research

Using a *Semantic* Clinical Data Warehouse

ontology as a strategy to formally represent and integrate knowledge

Semantic data integration through ontological mappings



J Biomed Semantics. 2011 May 17;2 Suppl 2:S1. doi: 10.1186/2041-1480-2-S2-S1.

Applications in biomedical and clinical research

Pharmaceutical Research

- Which existing marketed drugs might potentially be **re-purposed** for AD because they are known to modulate genes that are implicated in the disease?
  - *57 compounds or classes of compounds that are used to treat 45 diseases, including AD, hyper/hypotension, diabetes and obesity*

Clinical research

- Identify an AD clinical trial for a drug with a **different mechanism of action** (MOA) than the drug that the patient is currently taking
  - *Of the 438 drugs linked to AD trials, only 58 are in active trials and only 2 (Doxorubicin and IL-2) have a documented MOA. 78 AD-associated drugs have an established MOA.*

Health care

- Have any of my AD patients been treated for other neurological conditions as this might impact their diagnosis?
  - *Patient 2 is also being treated for depression.*

STRIDE-RDF

STRIDE [1] is a clinical data warehouse built from HL7 messages from the Stanford University Medical Center. Over 1.2 million pediatric and adult patients since 1995. Uses ICD9-CM, ICDO, CPT, RxNorm and SNOMED.

- We [2] converted patient demographics, diagnoses, laboratory tests, prescriptions, and text mined clinical notes into RDF.
- Demonstrate how federated SPARQL 1.1 queries can be used to answer the following questions:

	Question	Datasets used
1	Which co-morbidities are most often found in patients that suffer from Mucopolysaccharidosis?	STRIDE2RDF, ICD9
2	What disease genes are associated with Mucopolysaccharidosis co-morbidities?	STRIDE2RDF, ICD9, OMIM
3	Which adverse events experienced by Mucopolysaccharadosis patients taking Tromethamine are associated with this drug?	STRIDE2RDF, ICD9, RxNORM, SIDER

[1] Lowe et al . STRIDE. AMIA Annu Symp Proc. 2009; 2009: 391–395.
[2] Odgers & Dumontier. AMIA-TBI. 2015.

Translational Research: Identifying human drug targets with animal model phenotypes

semantic similarity

model phenotypes

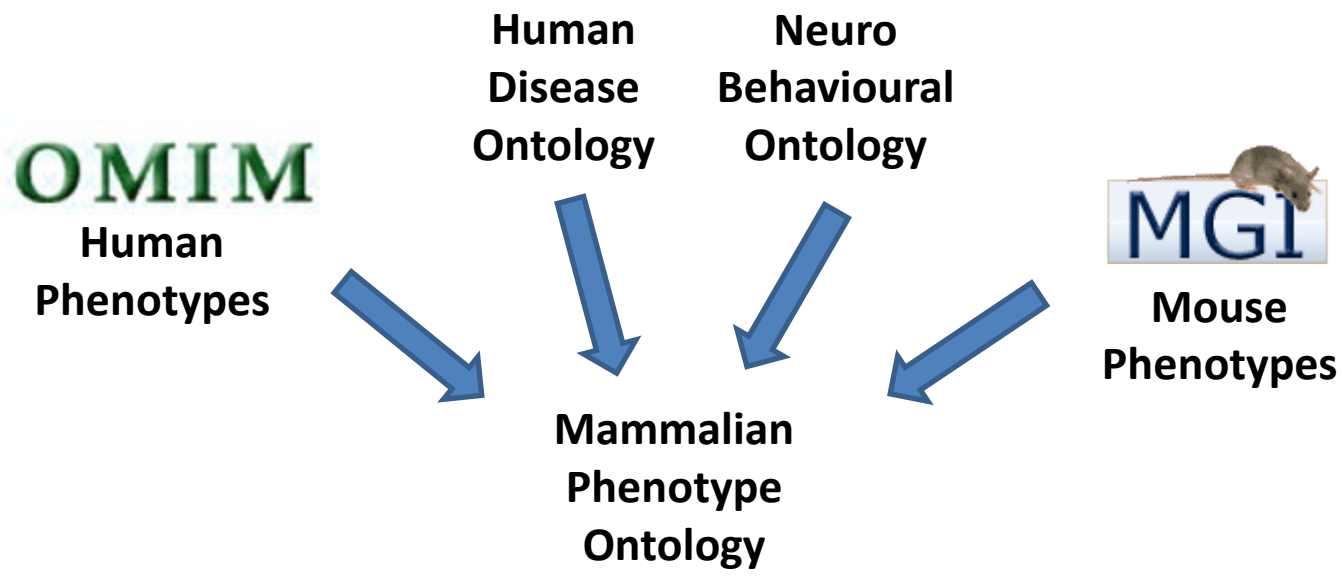
Human drug effects

knockout gene

inhibitor drug

Human gene

Terminological Interoperability



Terminological Interoperability *means* learning something new when you put them together.

human ‘overriding aorta [HP:0002623]’ EquivalentTo:

‘phenotype of’ some (‘has part’ some (‘aorta [FMA:3734]’ and ‘overlaps with’ some ‘membranous part of interventricular septum [FMA:7135]’)

mouse ‘overriding aorta [MP:0000273]’ EquivalentTo:

‘phenotype of’ some (‘has part’ some (‘aorta [MA:0000062]’ and ‘overlaps with’ some ‘membranous interventricular septum [MA:0002939]’

Uberon super-anatomy ontology provides inter-species mappings

‘aorta [FMA:3734]’ EquivalentTo: ‘aorta [MA:0002939]’

‘membranous part of interventricular septum [FMA:3734]’ EquivalentTo: ‘membranous interventricular septum [MA:0000062]

Thus, ‘overriding aorta [HP:0002623] EquivalentTo:‘overriding aorta[MP:0000273]’

Phenotypes of loss of function mutants largely predict inhibitor targets

- 14,682 drug formulations; 7,255 mouse genotypes
- Validate against known and predicted inhibitor-target pairs
  - 0.78 ROC AUC for human targets (DrugBank)
- diclofenac
  - 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

Research Aims and Directions

the **overall aim** of my research is

to understand how living systems respond to chemical agents

and developing small-molecule applications

My primary research interests are:

- Elucidating the mechanism of drug effects; polypharmacology
- Re-purposing drugs for rare, complex, and untreatable diseases
- Devising optimal drug combinations that maximize therapeutic value and minimize side effects
- Investigating the role of drug metabolic products in toxicology
- Empowering synthetic biology with small molecule chemistry

Let's get more out of the health data that we already have access to

- Access to de-identified patient data for research purposes
  - Use standardized, but evolving health terminologies
  - Text-mining to increase the amount of data available for analysis

- Interoperability between health and biomedical ontologies to enable translational research
  - Human Phenotype Ontology to be incorporated into the UMLS

- Use a growing suite of methods to access and integrate data.
  - RDF as a common platform for representing data
  - OWL ontologies as a means to formalize the meaning of terms so they become comparable
  - Methods to integrate, query, and semantically compare semantic data

- Envision new applications for testing, diagnosis, and treatment that makes the most out of the data we already have

Special Thanks

- Dumontier Lab
  - Jose Cruz-Toledo (IO Informatics)
  - Alison Callahan (Stanford)
  - Tanya Hiebert (recent grad)
  - Beatriz Lujan (recent grad)

New - post-docs wanted!

- Collaborators
  - Bio2RDF team
  - W3C HCLS Interest Group
  - Mark Wilkinson (UPM)
  - Robert Hoehndorf (KAUST)
  - George Gkoutos (Aberystwyth)
  - Nigam Shah (Stanford)

Yosemite Project
David Booth
Conor Dowling
Josh Mandel
Claude Nanjo
Rafael Richards

SemanticWeb.com
Eric Franzon

Let’s use semantic technologies to make it easier to do the work that needs to be done.

dumontierlab.com

michel.dumontier@stanford.edu

Website: http://dumontierlab.com
Presentations: http://slideshare.com/micheldumontier