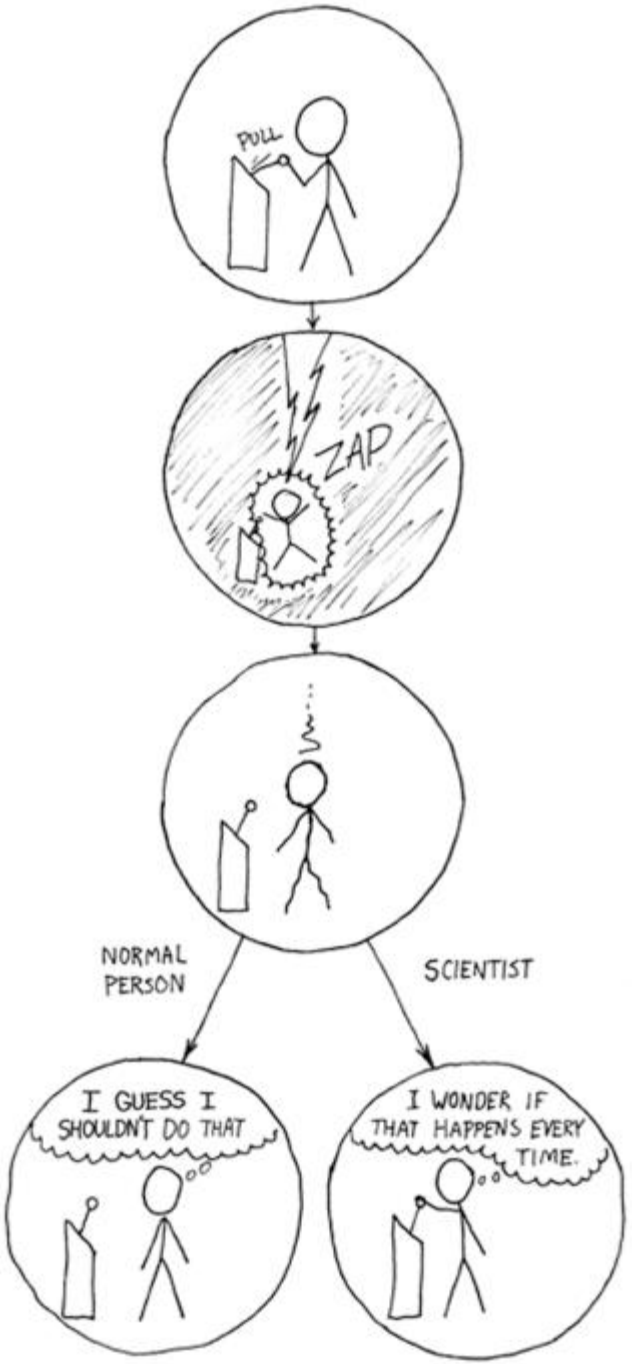
Data Science For The Win!

Michel Dumontier, Ph.D.

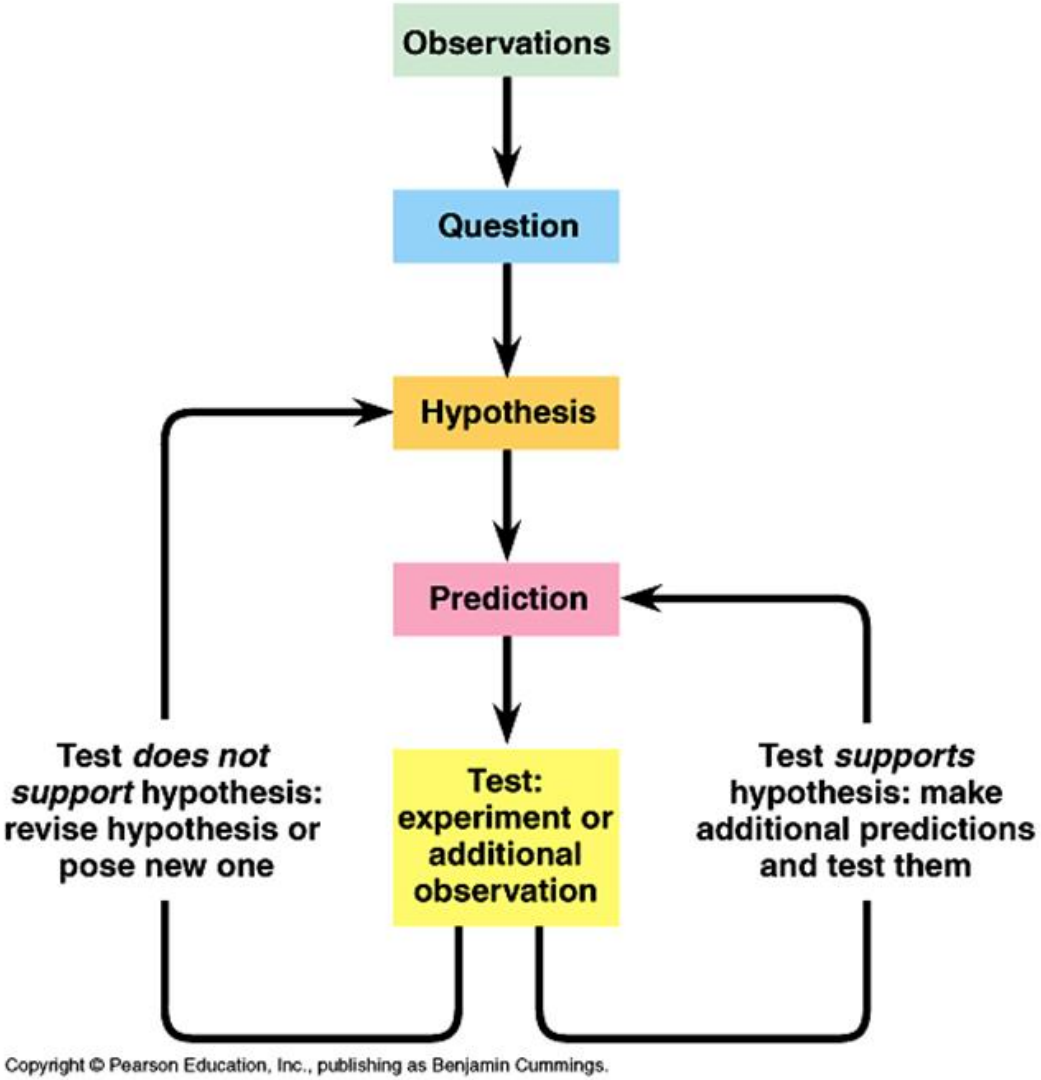
Distinguished Professor of Data Science

Maastricht University



HTTP://XKCD.COM/242/

Science!


Observations
Question
Hypothesis
Prediction
Test: experiment or additional observation
Test *does not support* hypothesis: revise hypothesis or pose new one
Test *supports* hypothesis: make additional predictions and test them
Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

Most published research findings are false.

- John Ioannidis, Stanford University

Non-reproducibility of **65–89%** in pharmacological studies and **64%** in psychological studies.

PLoS Med 2005;2(8): e124.

Science is hard.

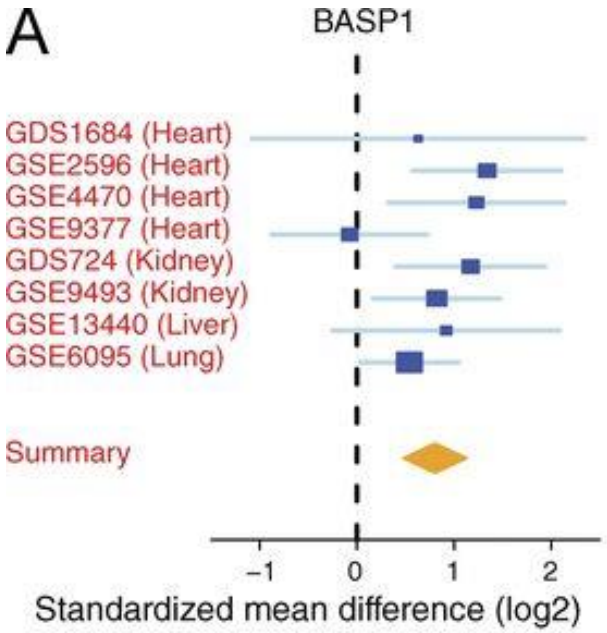
Statistics aren't sufficient.

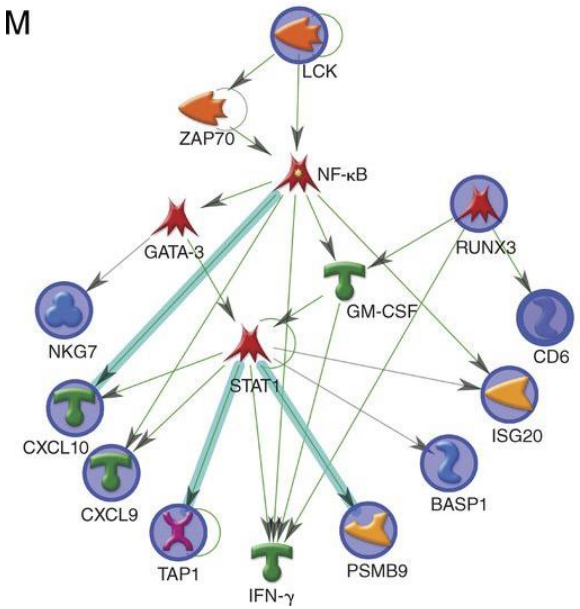
Biology is unruly.

Medicine is complicated.

we need new ways to think
about **discovery science**
using more **knowledge**
while paying attention to
reproducibility

Our confidence in a *finding* is strengthened when found in multiple independent datasets of the same type





A common rejection module (CRM) for acute rejection across multiple organs identifies novel therapeutics for organ transplantation

Khatri et al. JEM. 210 (11): 2205

DOI: 10.1084/jem.20122709

Our confidence in a *finding* is *strengthened* when we can uncover evidence at multiple levels

A study to examine what support exists for the role of genes in aging in the worm.

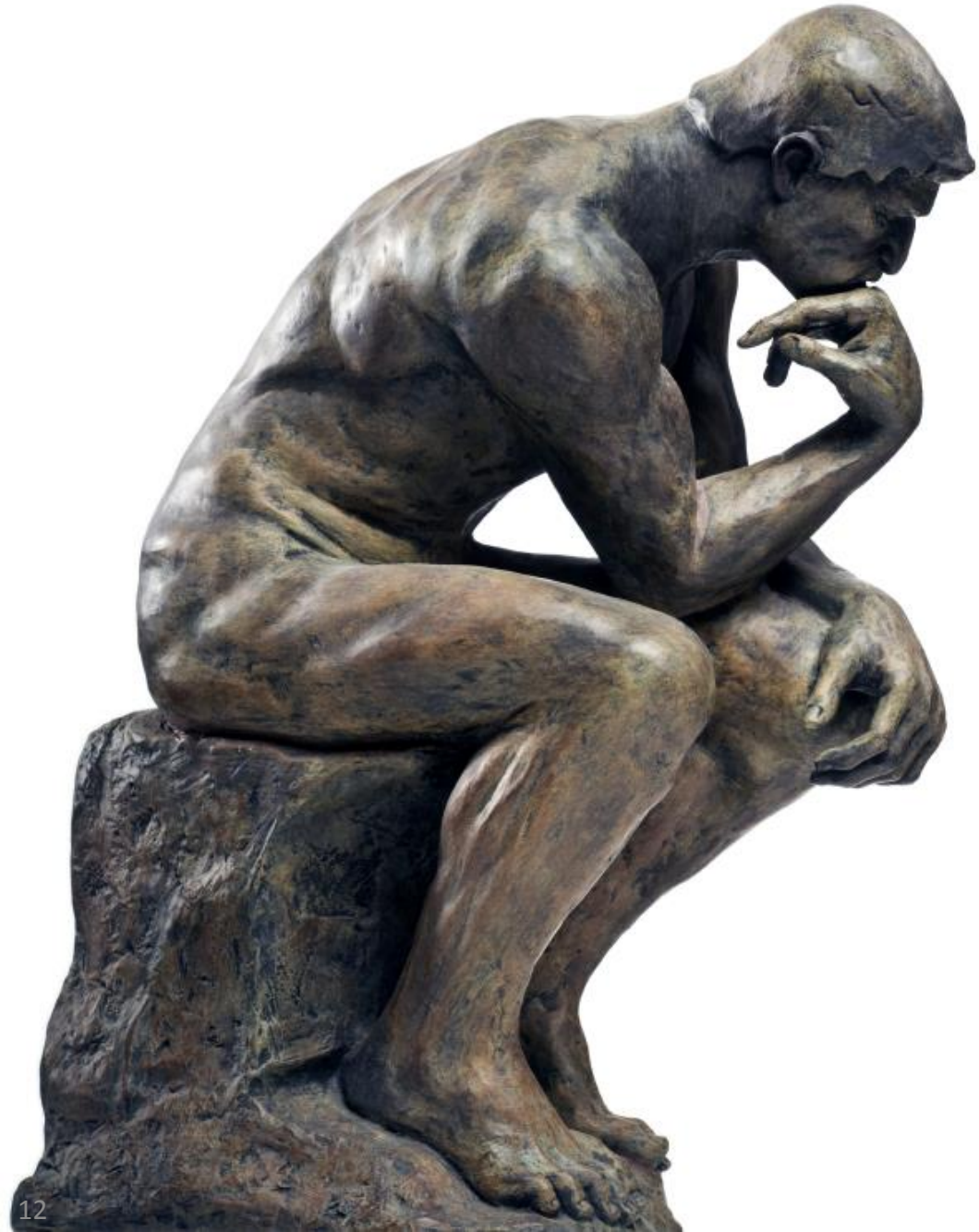
Table 3 8 *C. elegans* genes that received the highest HyQue evaluations for their role in aging, the PubMed identifiers of papers describing their roles in regulating longevity, and the data evaluation functions that contributed to their scores

WormBase identifier	Symbol	Score	PMID	Satisfied data evaluation function								
				1	2	3	4	5	6	7	8	9
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	✓	✓			✓	✓	✓	✓	✓

Known to be associated through perturbation expts.
Mentioned in scientific text together
Interact with known genes
…
Use statistical methods to find significant patterns and predictive models.

Table 4 31 highest scoring *C. elegans* genes that received HyQue evaluation scores for their role in aging without existing aging-related annotations, and the data evaluation functions that contributed to their scores

WormBase identifier	Symbol	Satisfied data evaluation function								
		1	2	3	4	5	6	7	8	9
WBGene00000252	bli-2		✓					✓	✓	✓
WBGene00000255	bli-5		✓				✓		✓	✓
WBGene00000262	bra-1		✓					✓	✓	✓
WBGene00000479	cgh-1						✓	✓	✓	✓
WBGene00000915*	daf-21						✓	✓	✓	✓
WBGene00001165	efn-4		✓					✓	✓	✓
WBGene00001428*	fkb-3		✓				✓		✓	✓
WBGene00001543*	gcy-18		✓				✓		✓	✓
WBGene00001578	ges-1		✓				✓		✓	✓
WBGene00001746	gsk-3		✓					✓	✓	✓
WBGene00001824	hbl-1		✓				✓		✓	✓
WBGene00001974	hmg-4		✓					✓	✓	✓
WBGene00001979	hmp-2		✓					✓	✓	✓
WBGene00002005*	hsp-1				✓		✓		✓	✓
WBGene00002013*	hsp-12.6		✓		✓				✓	✓
WBGene00002069*	ikb-1		✓		✓				✓	✓
WBGene00002881	let-756		✓					✓	✓	✓
WBGene00003029	lin-44		✓					✓	✓	✓
WBGene00003058	lov-1		✓					✓	✓	✓
WBGene00003210	mel-28						✓	✓	✓	✓
WBGene00003473	mtl-1		✓				✓		✓	✓
WBGene00003497	mup-4		✓					✓	✓	✓
WBGene00003977*	pes-2.1		✓				✓		✓	✓
WBGene00004392	rnr-2						✓	✓	✓	✓
WBGene00004765	sel-8		✓					✓	✓	✓
WBGene00006789	unc-54		✓					✓	✓	✓
WBGene00007036	sod-1		✓				✓	✓	✓	✓



How can we automatically find the evidence that support or dispute a scientific hypothesis using the totality of available data, tools and scientific knowledge?

So what do we need to achieve this?

1. **Data Science**

   Infrastructure to identify, represent, store, transport, retrieve, aggregate, query, mine, analyze *data* and execute *services* on demand in a reproducible manner.

   Methods to *discover* plausible, supported, prioritized, and experimentally verifiable associations.

2. **Community**

   to build a massive, decentralized network of *interconnected* and *interoperable* data and services

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, **Michel Dumontier** [...] **Barend Mons**

Affiliations | **Contributions** | **Corresponding author**

Scientific Data **3**, Article number: 160018 (2016) | doi:10.1038/sdata.2016.18

Received 10 December 2015 | Accepted 12 February 2016 | Published online 15 March 2016

FAIR: Findable, Accessible, Interoperable, Re-usable

Applies to *all* digital *resources* and their *metadata*

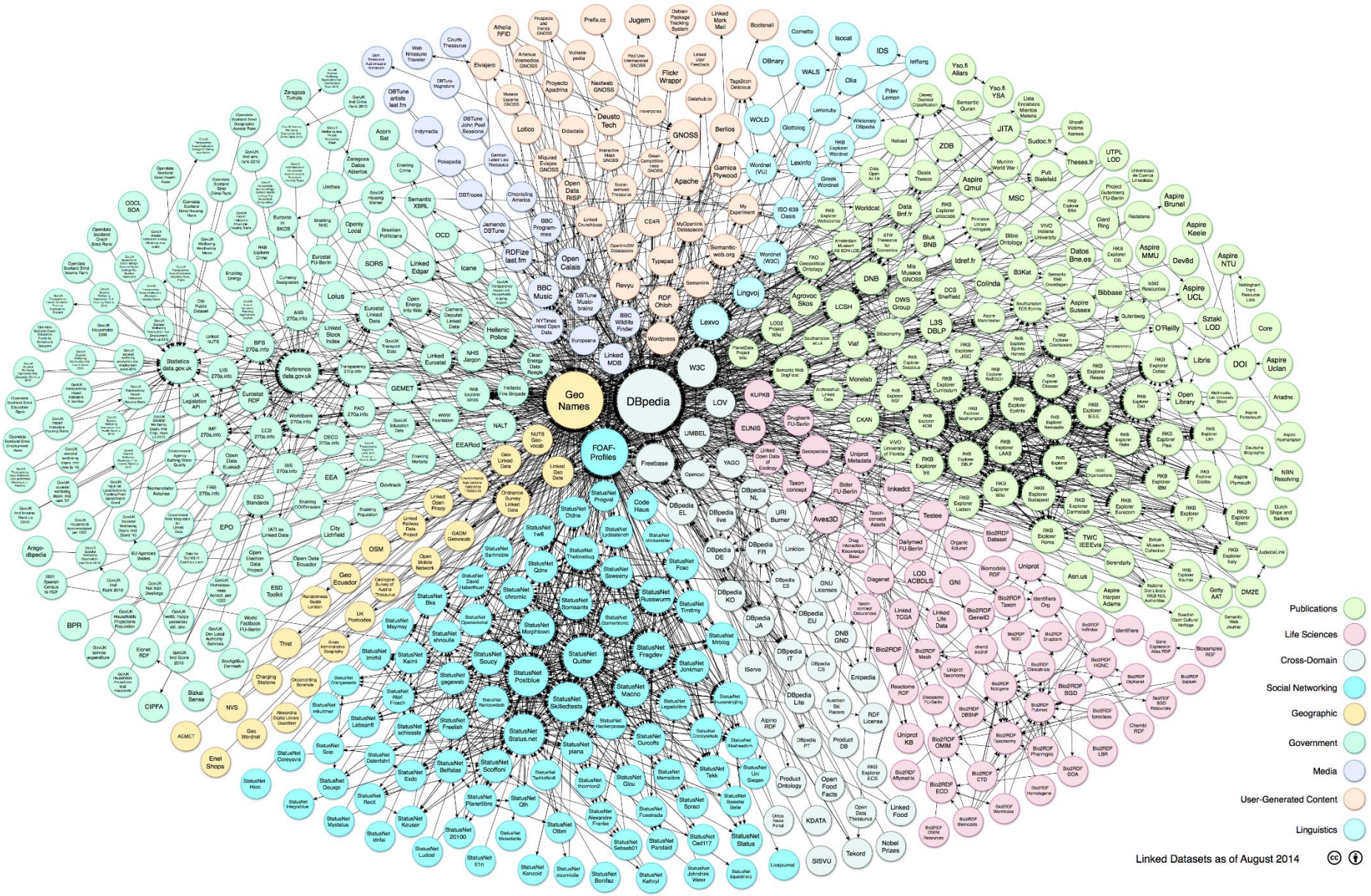
The Semantic Web
is the new global **web of knowledge**

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

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

W3C Semantic Web

Be FAIR by creating Linked Data

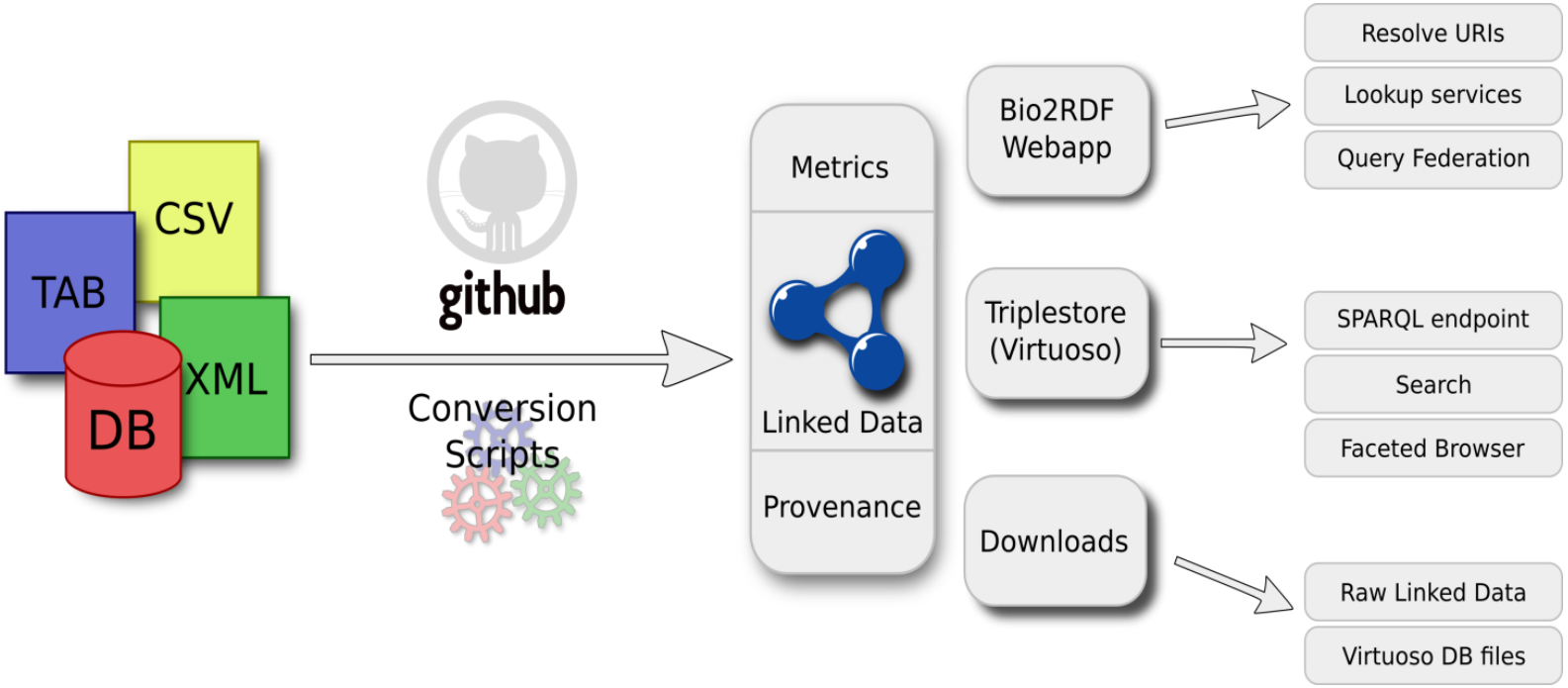


Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/"

BIO2RDF

Linked Data for the Life Sciences

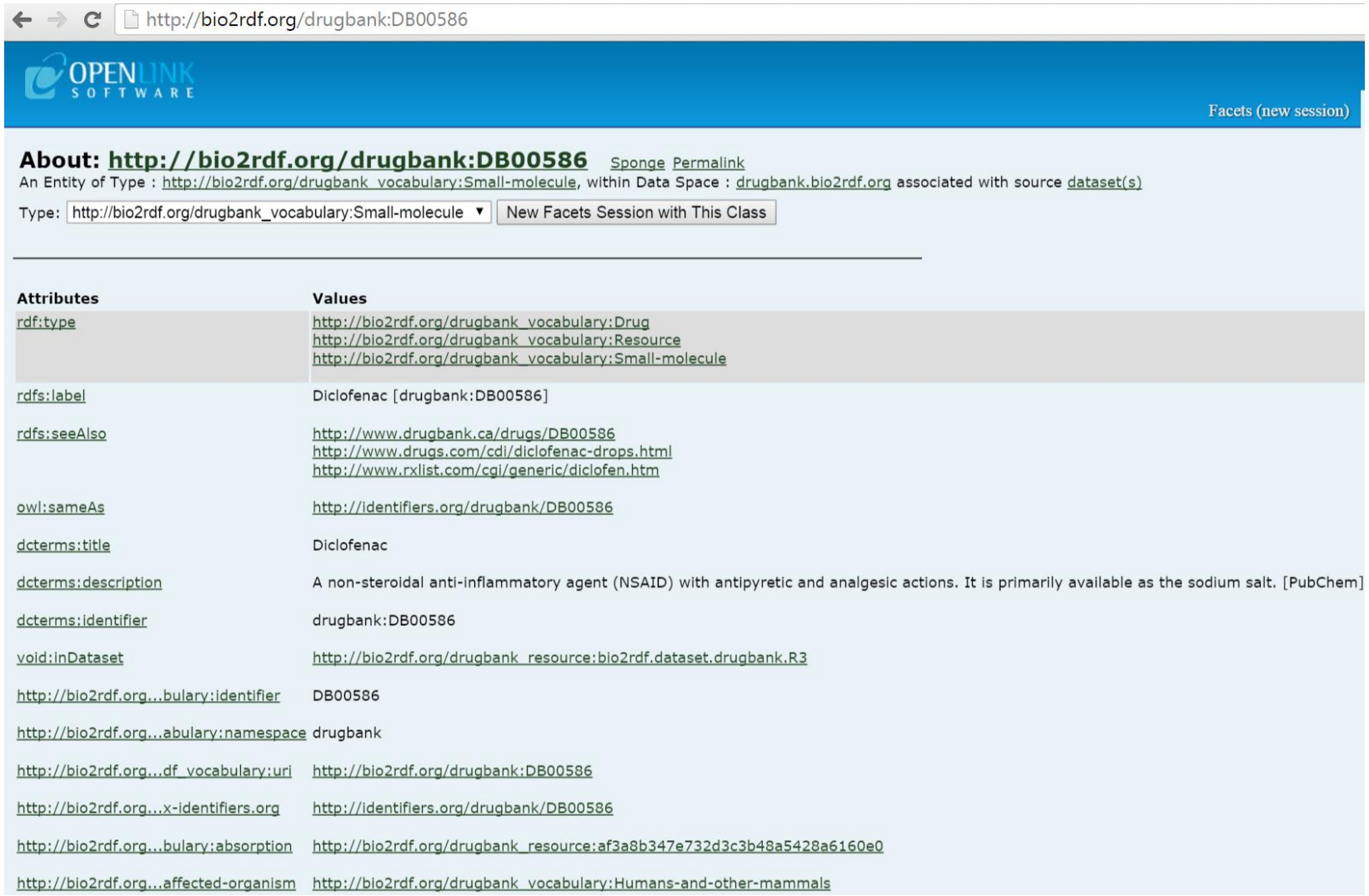
Bio2RDF is an open source project that uses semantic web technologies to make FAIR biomedical data



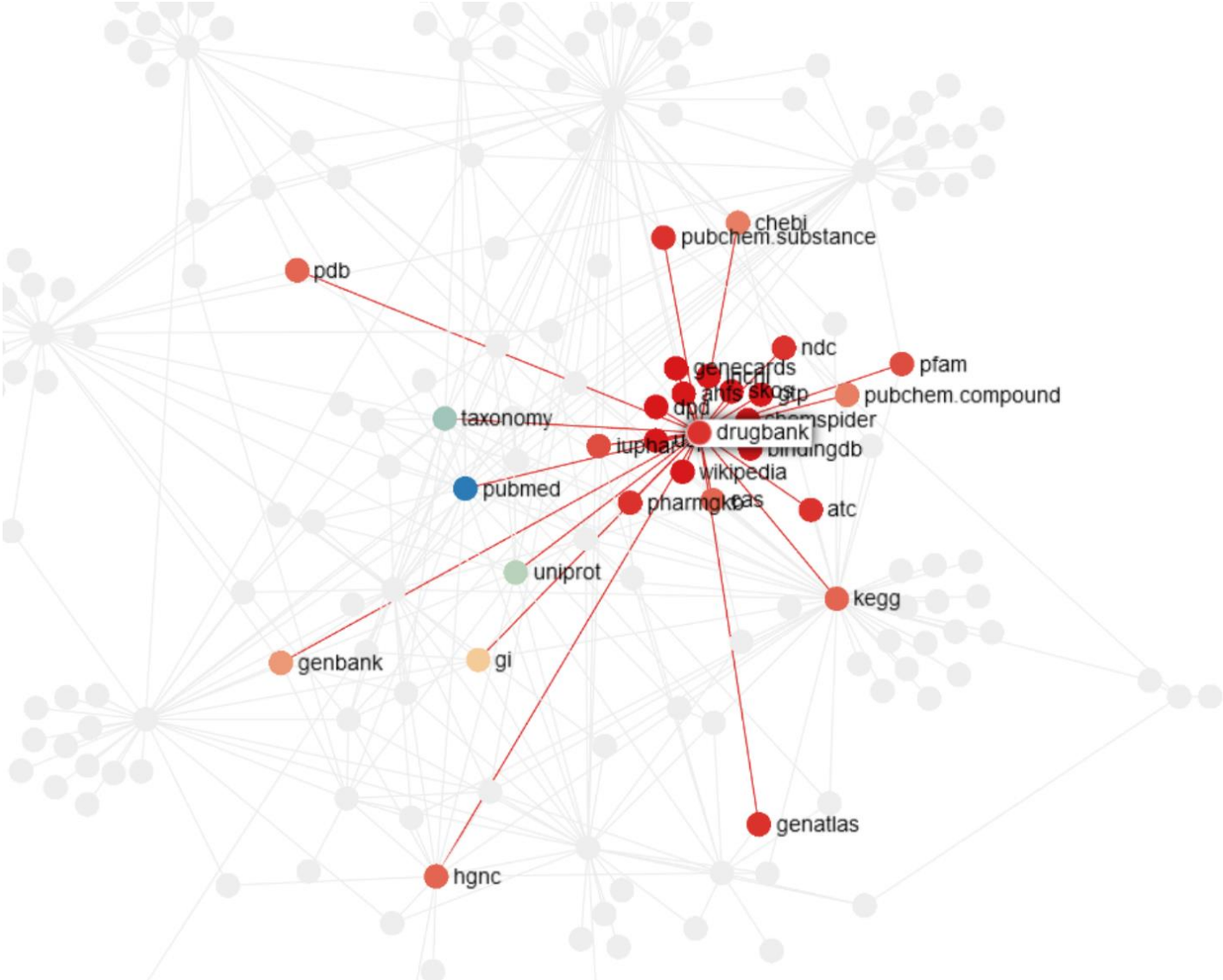
chemicals/drugs/formulations,
genomes/genes/proteins, domains
Interactions, complexes & pathways
animal models and phenotypes
Disease, genetic markers, treatments
Terminologies & publications

- Billions of facts from 35 biomedical datasets
- Provenance & statistics
- **A growing interoperable ecosystem with global partners: EBI, NCBI, DBCLS, NCBO, OpenPHACTS, and commercial tool providers**

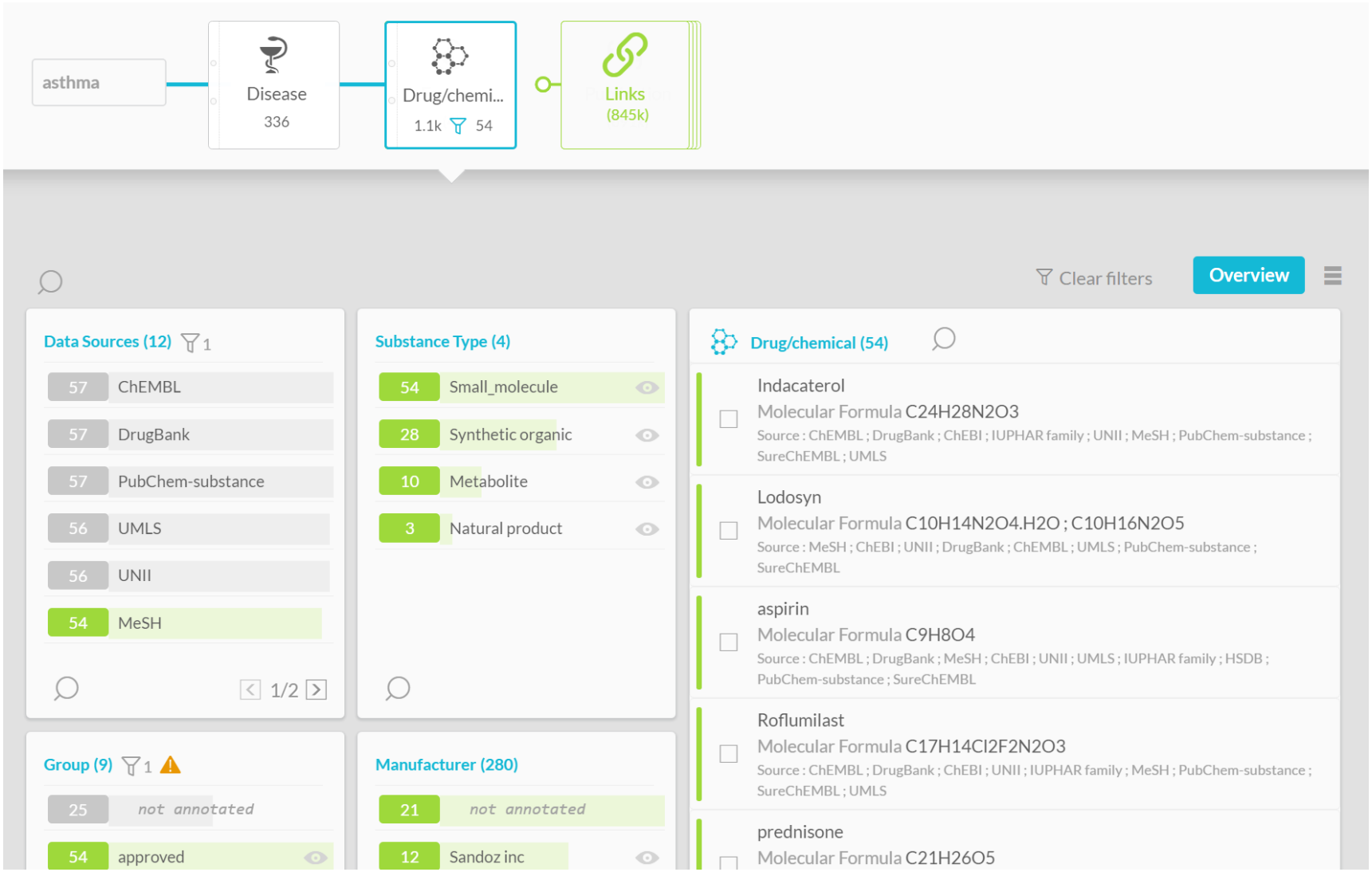
Get data in standardized representations using web technology



Discover connections across datasets

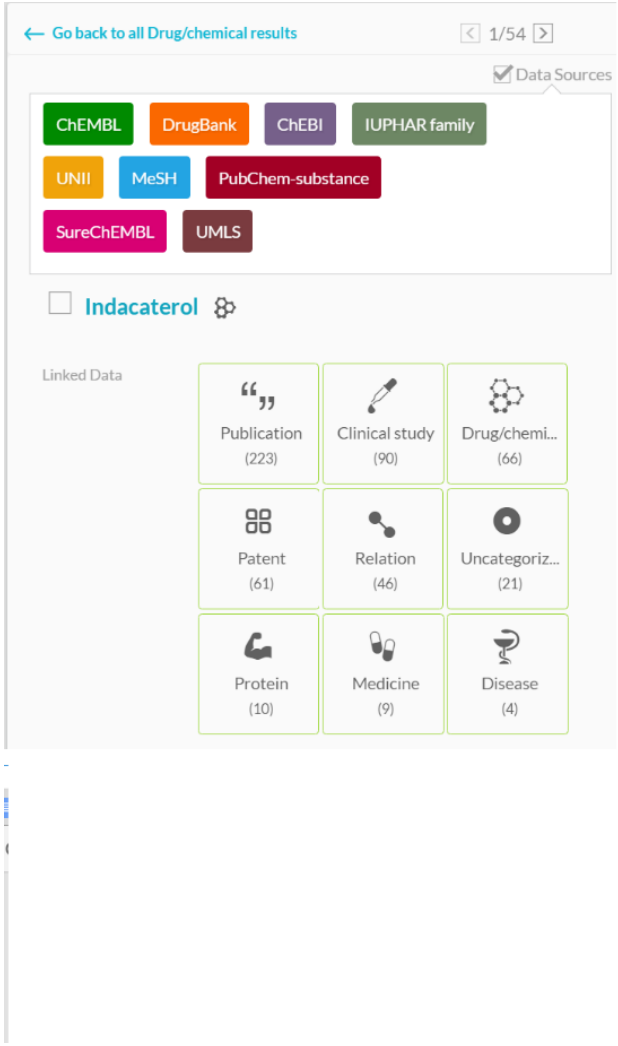


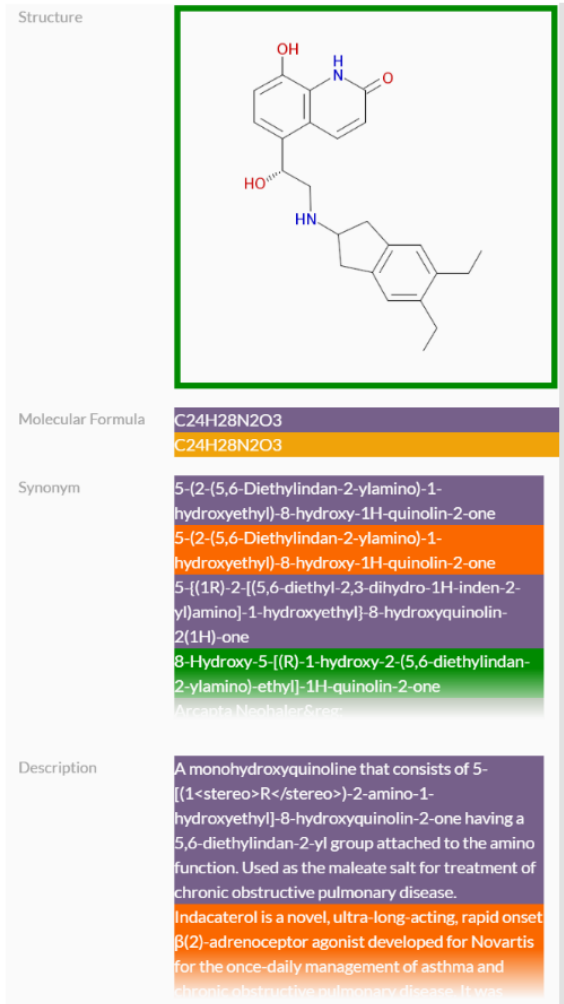
Efficiently find and explore data

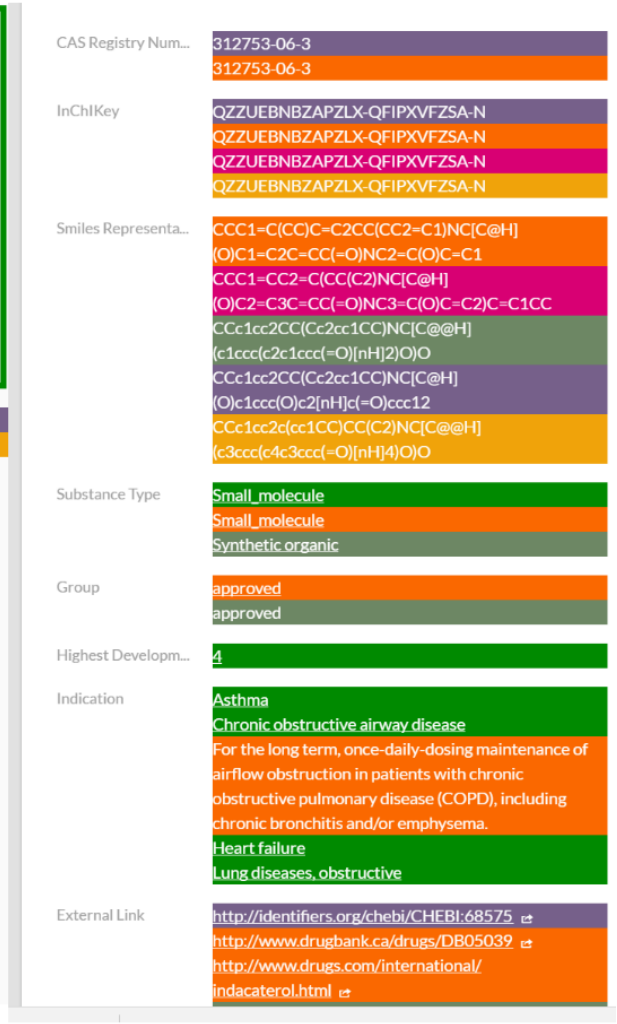


ONTOFORCE

Examine the facts and their provenance

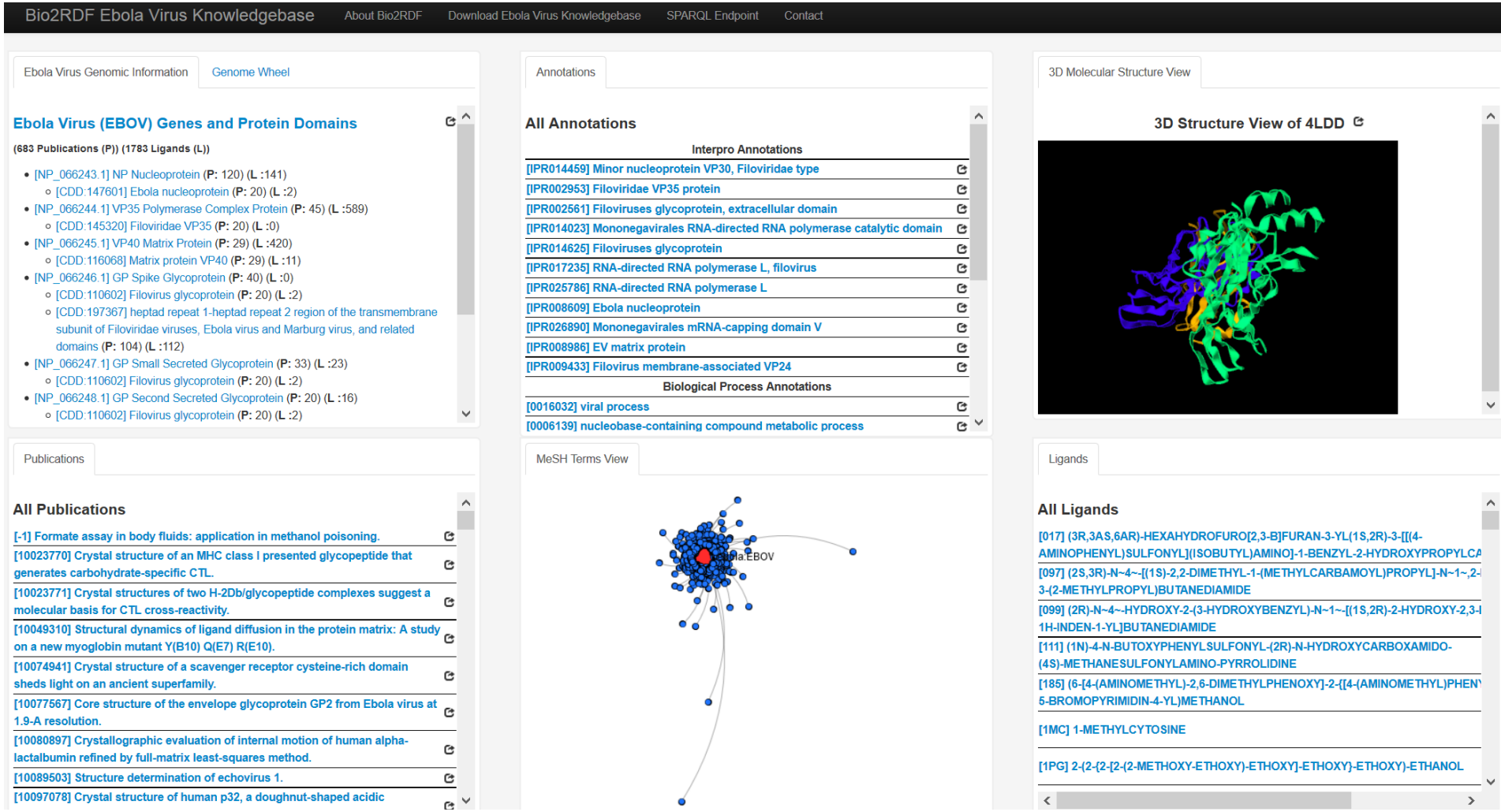






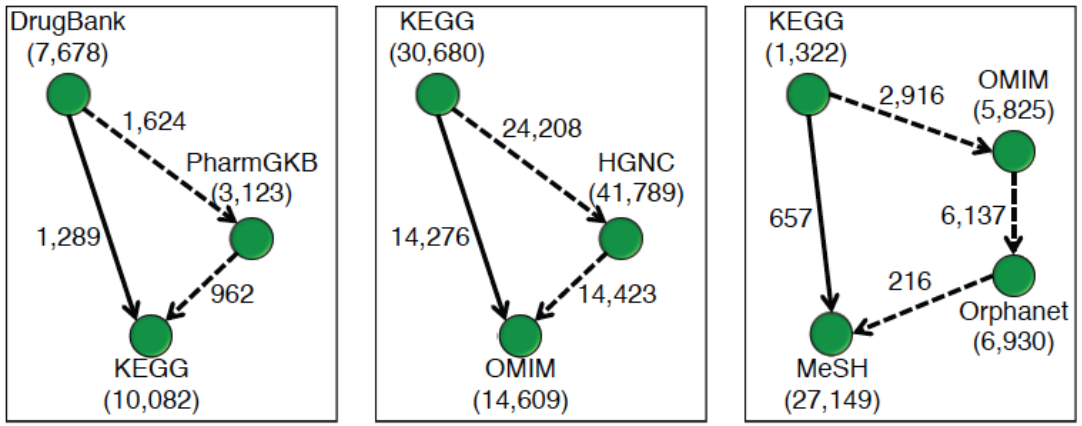
ONTOFORCE

Develop timely knowledge portals



Kamdar, Dumontier. An Ebola virus-centered knowledge base. Database. 2015 Jun 8;2015. doi: 10.1093/database/bav049.

Apply graph methods to assess find bad links and fill in the gaps



(a) Drugs (b) Human genes (c) Diseases

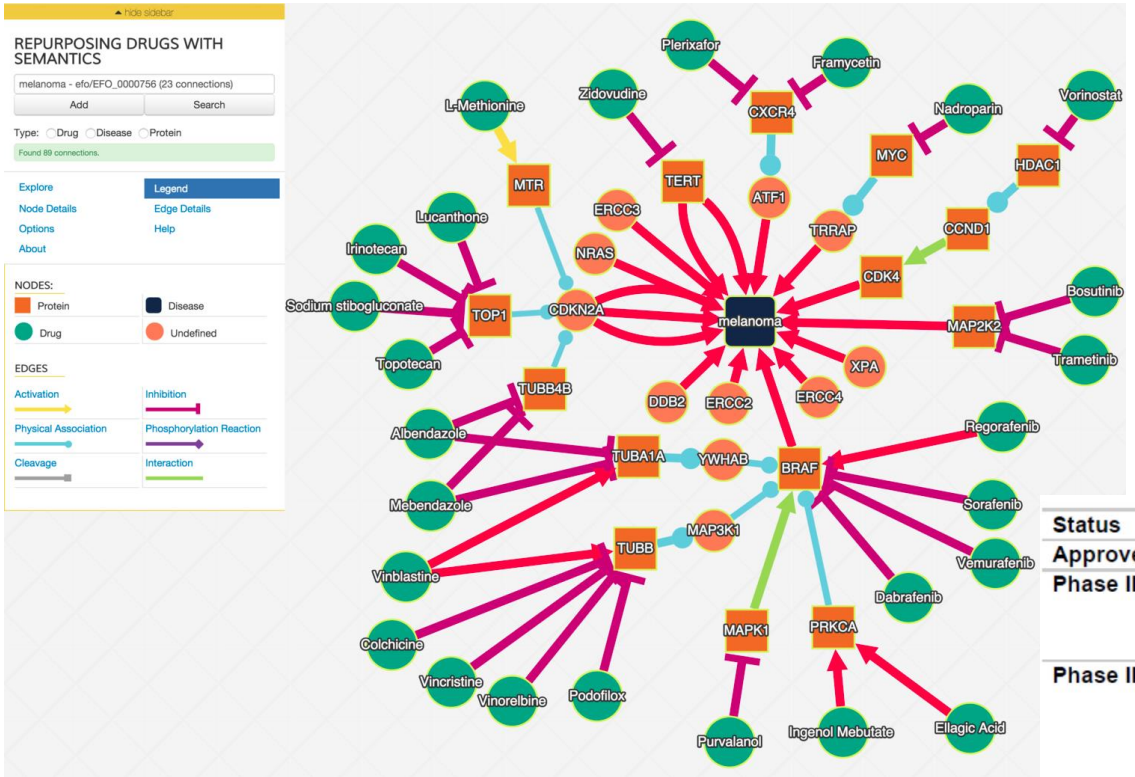
	Direct links	Transitive paths	Identical ending entities	Different entities	Missing direct	Missing transitive	Total
Drugs	1,289	954	946	6	2	343	1,297
Human genes	14,276	14,250	14,236	5	9	40	14,290
Diseases	657	33	8	18	7	649	682

Fig. 3. Transitivity analysis of entity links: (i) the value in each parenthesis denotes the number of entities given a specified topic; and (ii) the solid arcs represent direct links between entities while the dashed arcs form transitive paths. The value on each arc denotes the number of entity links from one dataset to the other.

W Hu, H Qiu, **M Dumontier**. Link Analysis of Life Science Linked Data. International Semantic Web Conference (2) 2015: 446-462.

Find new uses for existing drugs

using data mining



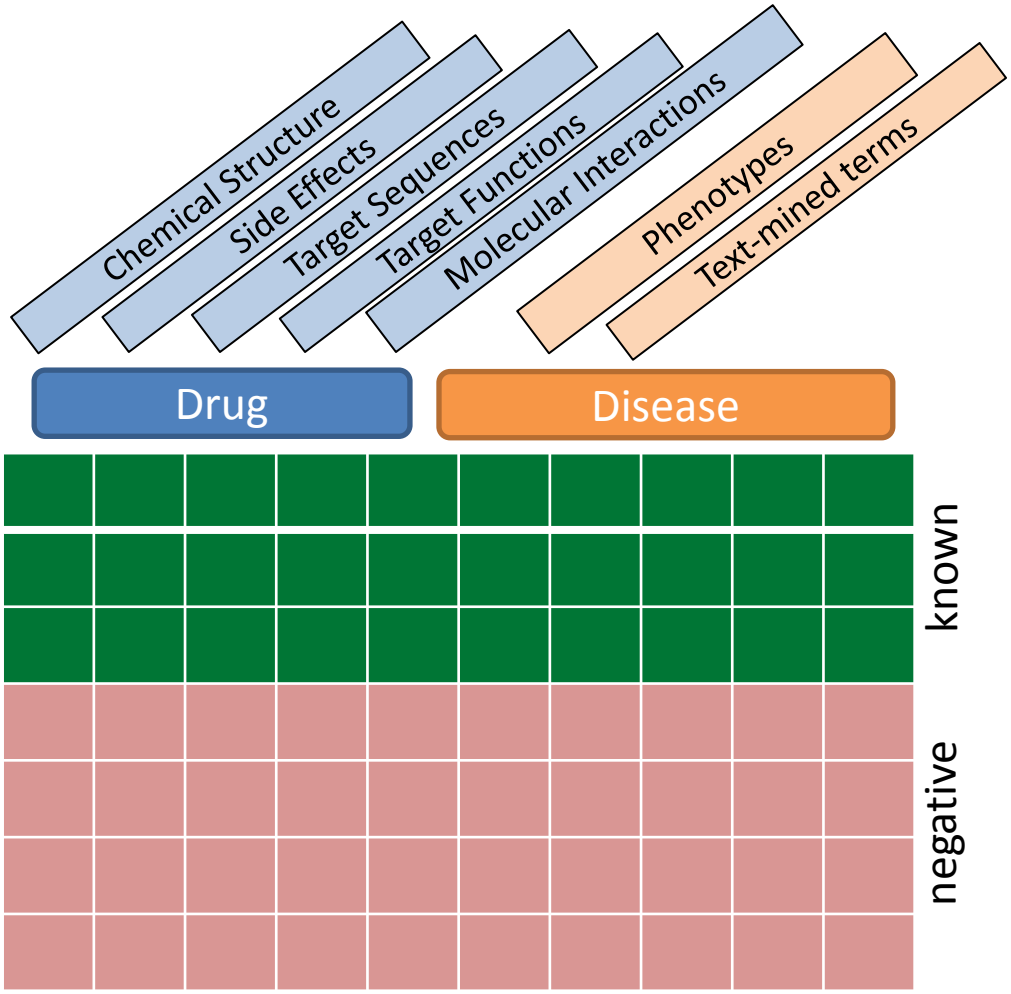
And validate them against pipelines for drug discovery

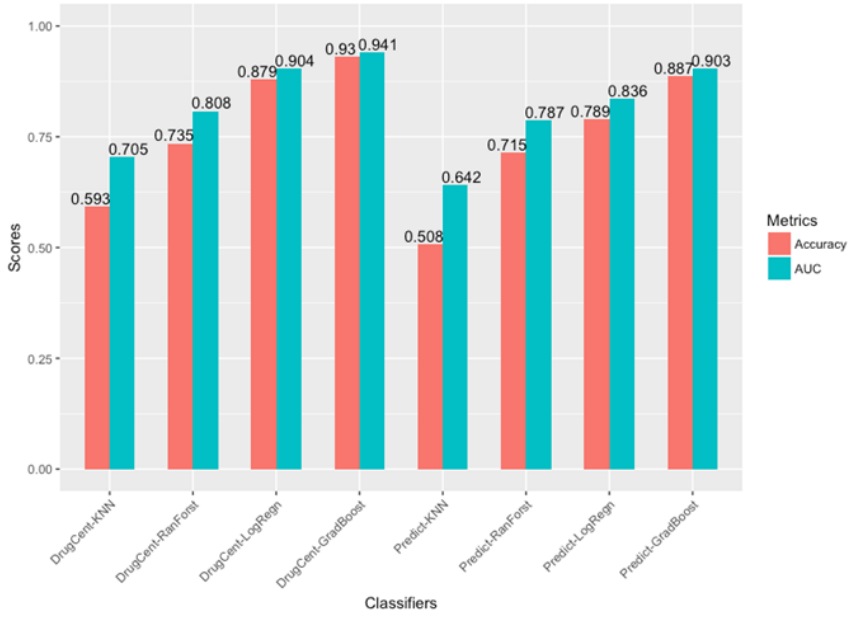
Status	Drug	Pathway	Steps	Joint p
Approved	Vemurafenib[2]	BRAF	2	0.98
Phase III	Dabrafenib[13]	BRAF	2	0.98
	Sorafenib[14]	BRAF	2	0.98
	Vinblastine[18]	MAP kinase	3	0.93
Phase II	Zidovudine[29]	TERT	2	0.98
	Trametinib[19]	MAP kinase	2	0.98
	Regorafenib[15]	BRAF	2	0.98
	Nadroparin[30]	MYC	3	0.97
	Vinorelbine[20]	MAP kinase	3	0.93
	Irinotecan[43]	CDKN2A	3	0.93
	Topotecan[44]	CDKN2A	3	0.93
Phase I	Sodium stibogluconate[45]	CDKN2A	3	0.93
Case Study	Ingenol Mebutate[46]	PRKCA/BRAF	3	0.95
In Vitro	Bosutinib[17]	MAP kinase	2	0.98
	Purvalanol[21]	MAP kinase/TP53	3	0.97
	Ellagic Acid[47]	PRKCA/BRAF	3	0.95
	Albendazole[48]	CDKN2A	3	0.93
	Colchicine[22]	MAP kinase	3	0.93
In Vivo	Plerixafor[27]	CXCR4	3	0.97
	Vincristine[23]	MAP kinase	3	0.93
	L-Methionine[49]	CDKN2A	3	0.93
	Mebendazole[50]	CDKN2A	3	0.93

Finding melanoma drugs through a probabilistic knowledge graph. *PeerJ Preprints*

Find new uses for existing drugs

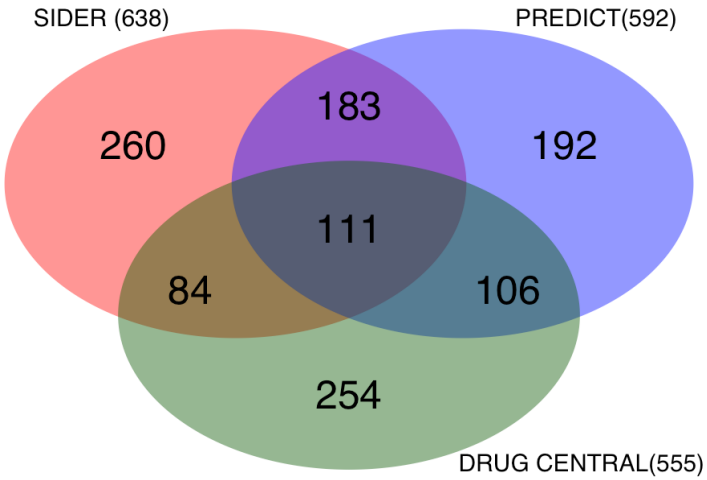
using machine learning



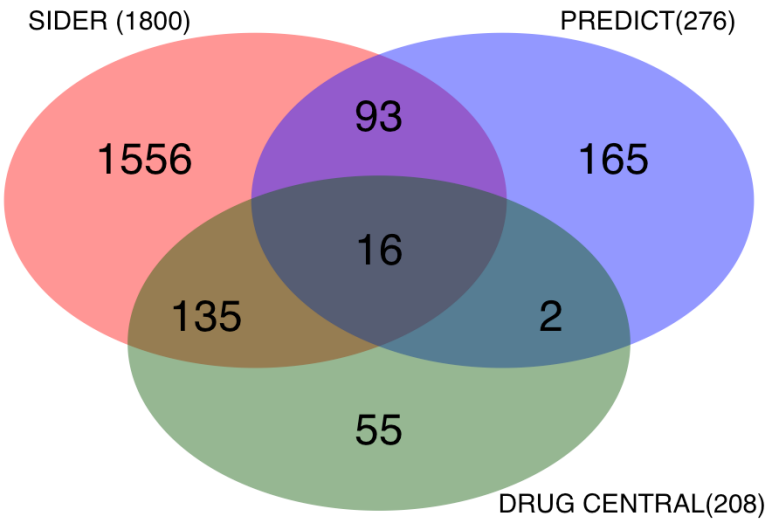


Examine overlap in “gold standards”

drugs



diseases



Uncover evidence in a transparent manner

HyQue

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	
hERG inhibition	
Literature-sourced drug targets	
Known cardiotoxicity assays	
Known gene targets and associated mouse model phenotypes	
Known drug targets and effects	

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 .
} .
}
```

Key Research Challenges to accelerate discovery science

- Scalable, shared, fault-tolerant, and readily re-deployable frameworks for **archiving** and **providing versioned and maximally FAIR biomedical (meta)data**
- Scalable methods for the *prospective* and *retrospective* **authoring, assessment, and repair of metadata**.
- Scalable frameworks for *open*, *transparent*, *reproducible* and *recurrent* **analysis** and **meta-analysis** of FAIR research data.
- Methods to identify **reporting *biases*** and **knowledge *gaps***
- Scalable and reliable methods **for the evaluation of scientific hypotheses using evidence gathered across scales and sources**
- Scalable methods for **validation** of research findings.

Institute of Data Science

The **mission** of the **Institute of Data Science** is to **accelerate scientific discovery**, **improve clinical care and well being**, and to **strengthen communities.** We will foster *a collaborative environment for inter- and multi-disciplinary research and training* founded on accurate, reproducible, multi-scale, distributed, and efficient computation.

Maastricht University

Institute of Data Science

Tackle impactful research problems through multidisciplinary teams involving students, researchers, partners, and stakeholders inside and out of the University.

- **Establish a core team of researchers** to broadly fulfill the mission of the Institute.
- **Highlight and engage the incredible data scientists** already at UM
- **Work with communities** to deploy and evaluate the impact of the application of our methods.
- **Train the next generation of data scientists** to beeven more collaborative and interdisciplinary.

michel.dumontier@maastrichtuniversity.nl

Website: http://dumontierlab.com

Presentations: http://slideshare.com/micheldumontier

Maastricht University