

Data Science For The Win!

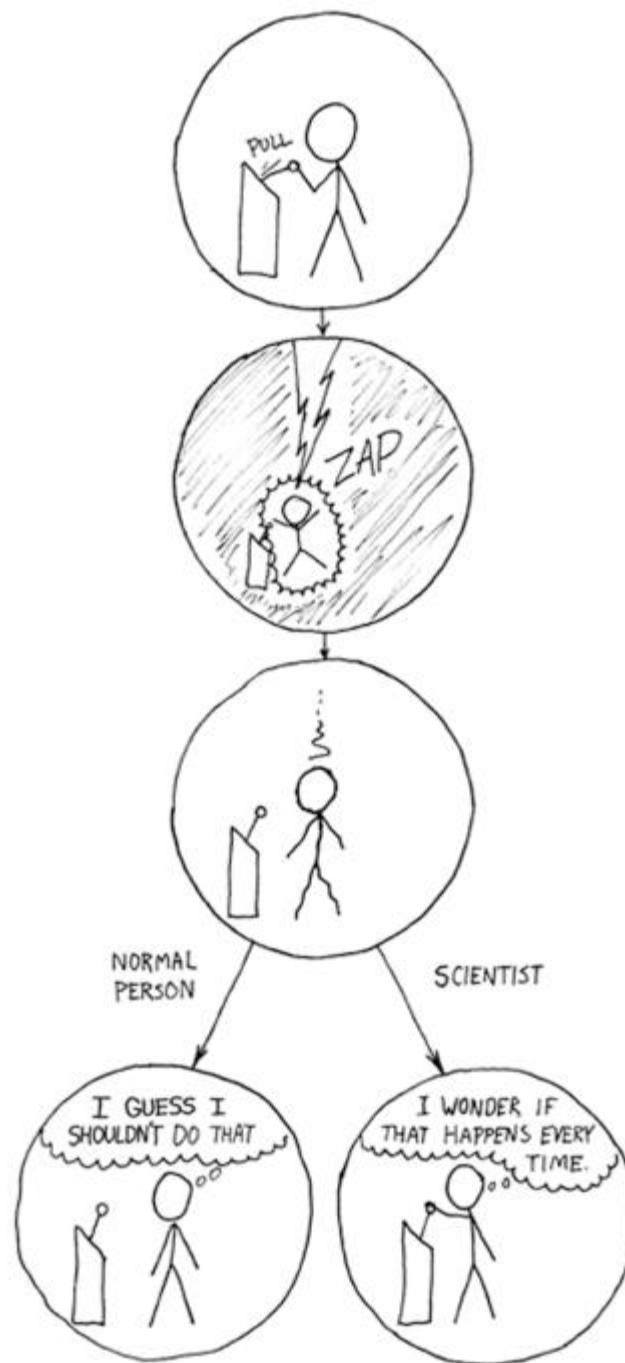


Michel Dumontier, Ph.D.

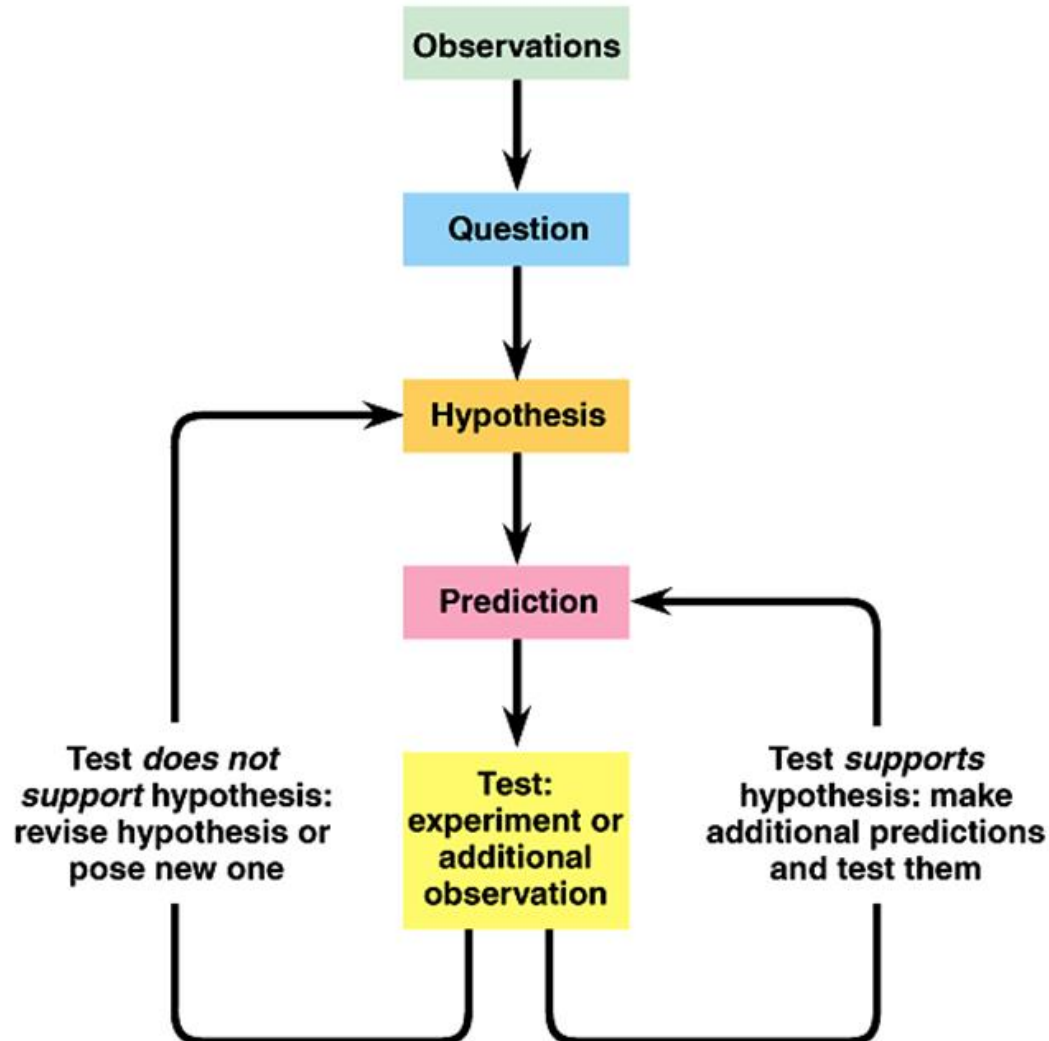
Distinguished Professor of Data Science



Maastricht University



Science!



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.

Science is hard.

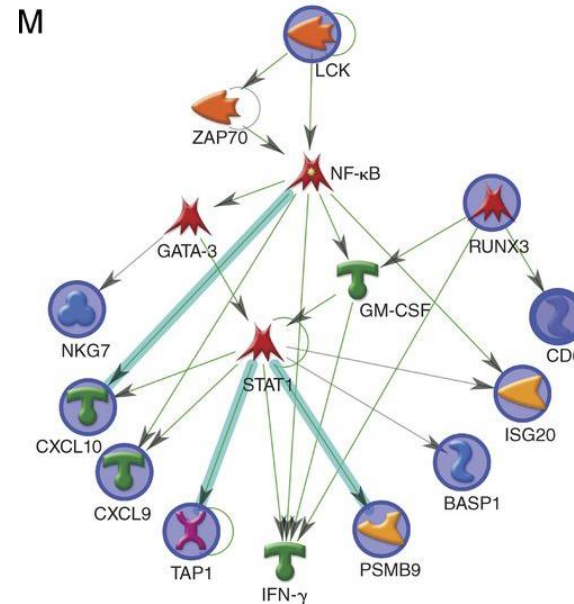
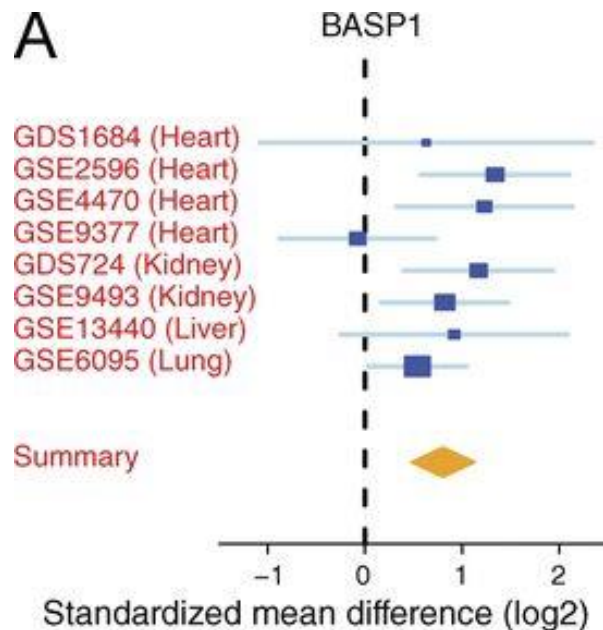
Statistics aren't sufficient.

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

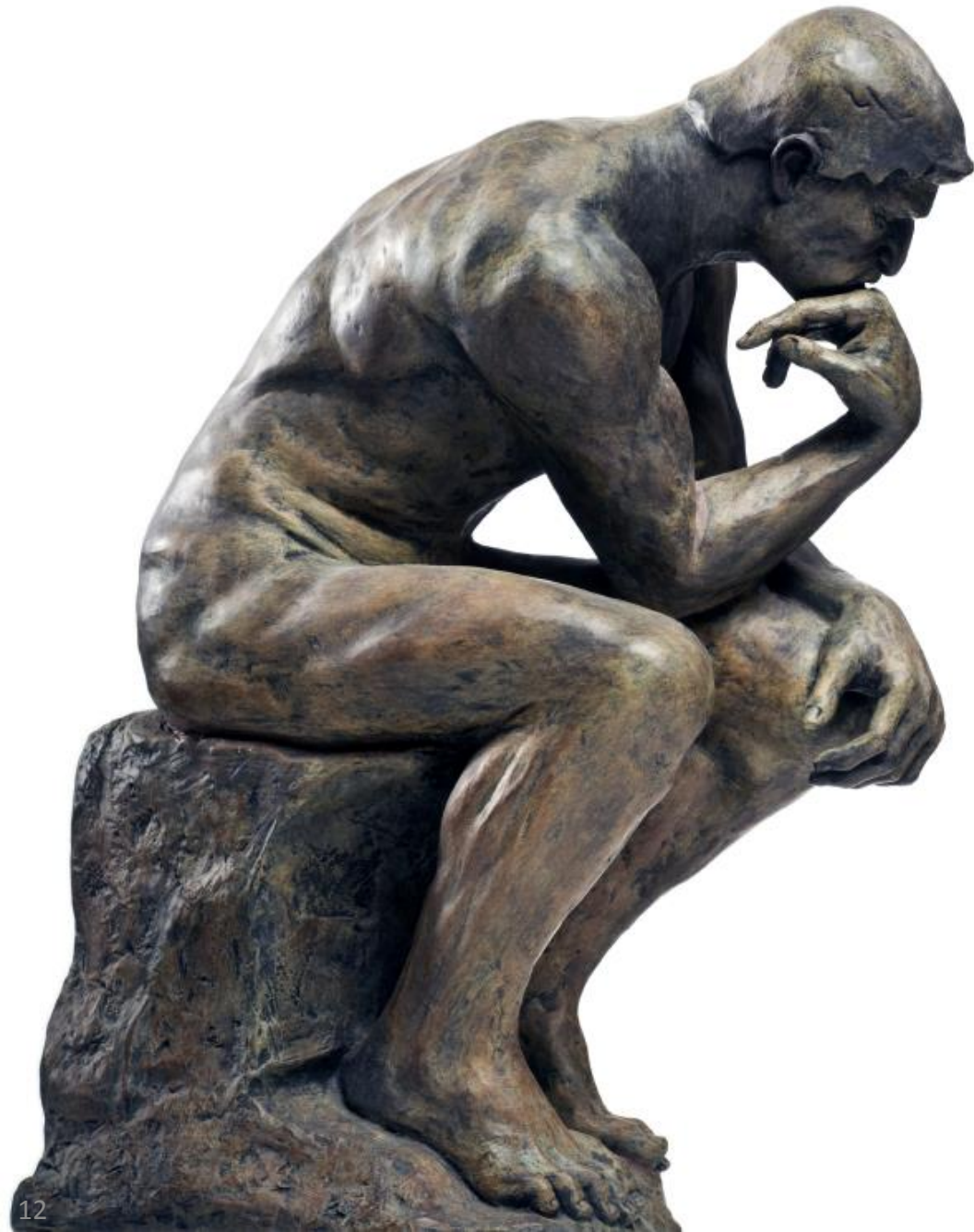
Interact with known genes

...

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*	fkf-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	unc-54	✓						✓	✓	✓



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

About *Scientific Data*

Scientific Data is an open-access, peer-reviewed journal for descriptions of scientifically valuable datasets. Our primary article-type, the **Data Descriptor**, is designed to make your data more discoverable, interpretable and reusable.

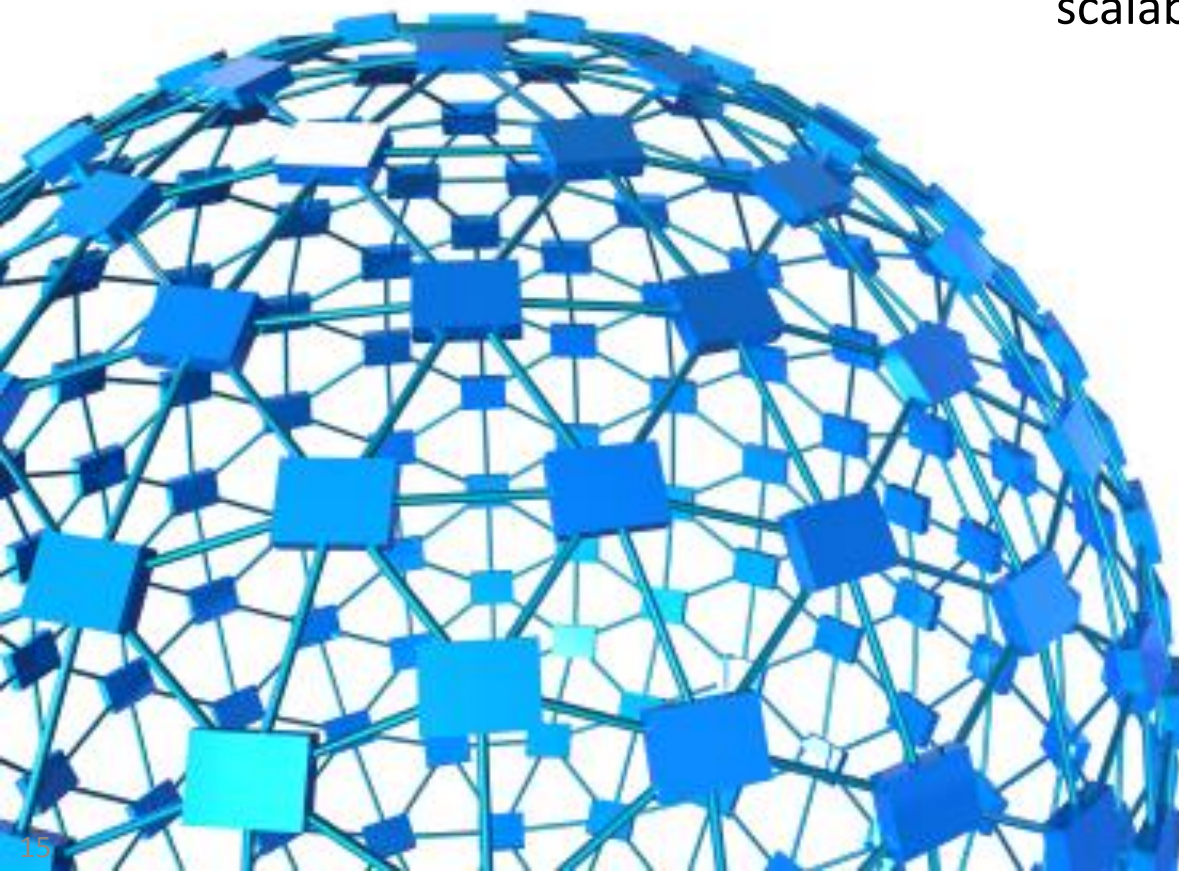
FAIR: Findable, Accessible, Interoperable, Re-usable

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

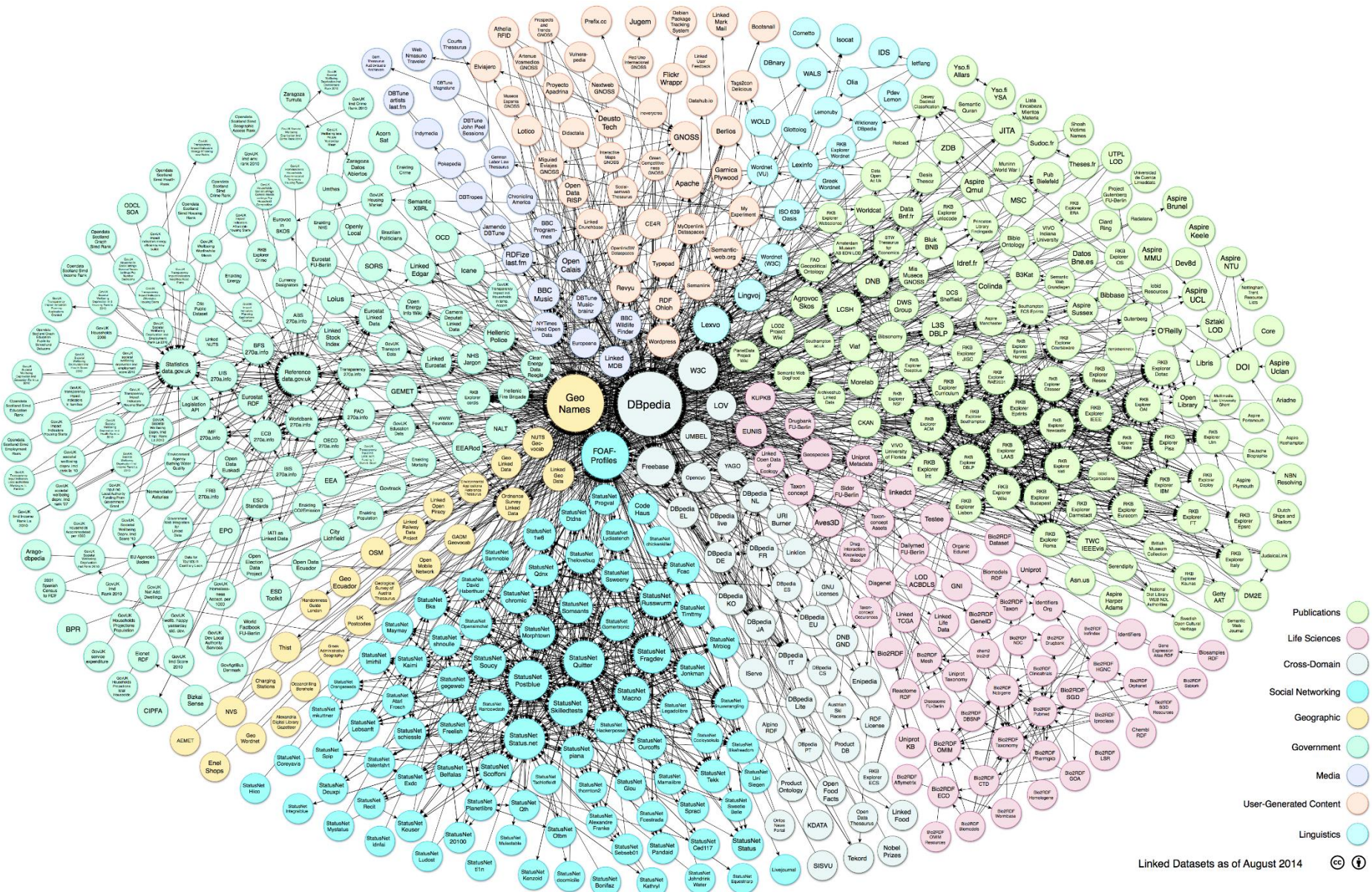
The Semantic Web is the new global **web of knowledge**

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

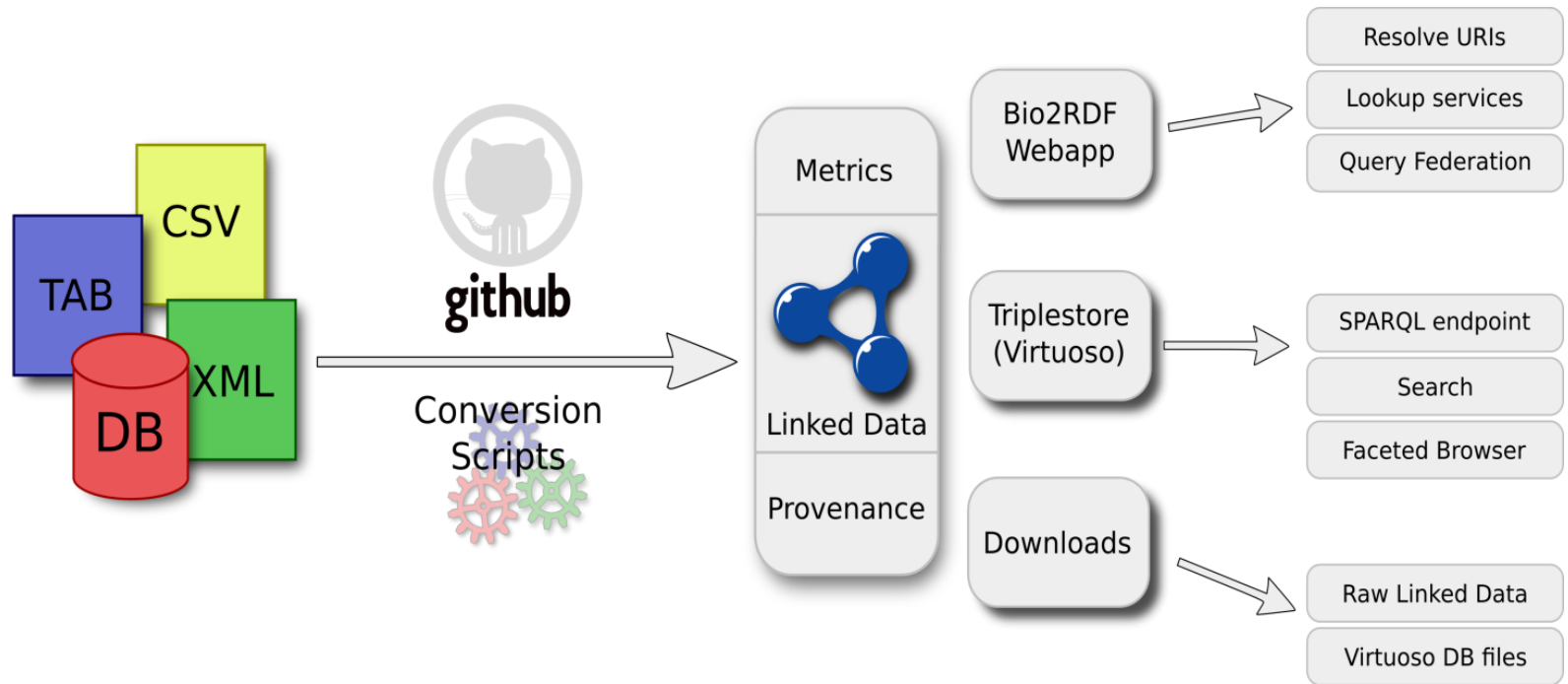
scalable approach for the discovery
of *independently formulated*
and *distributed* knowledge



Be FAIR by creating Linked Data



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

← → ↻ <http://bio2rdf.org/drugbank:DB00586>

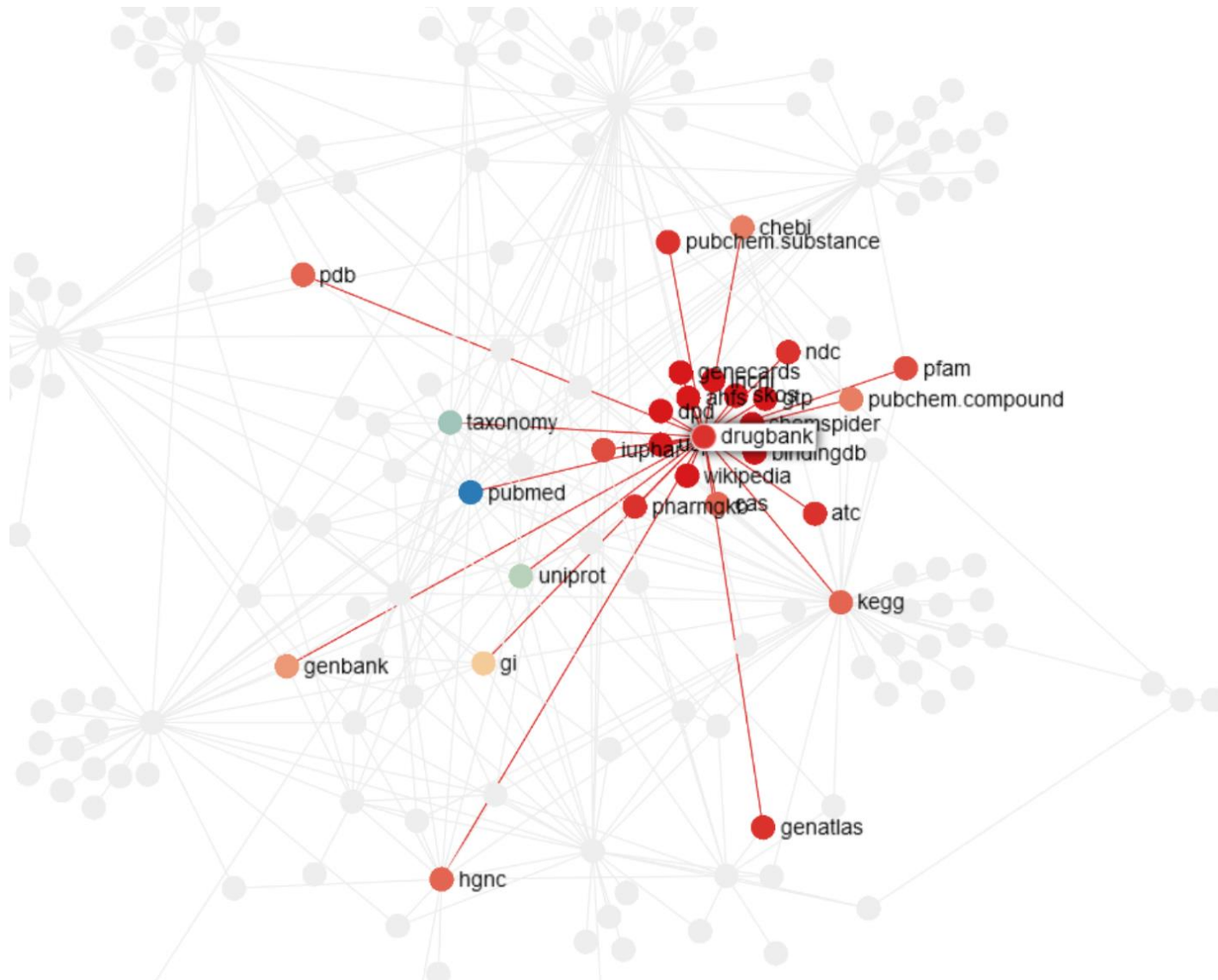
 OPENLINK SOFTWARE

Facets (new session)

About: <http://bio2rdf.org/drugbank:DB00586> [Sponge](#) [Permalink](#)
An Entity of Type : http://bio2rdf.org/drugbank_vocabulary:Small-molecule, within Data Space : drugbank.bio2rdf.org associated with source [dataset\(s\)](#)
Type: http://bio2rdf.org/drugbank_vocabulary:Small-molecule ▼ [New Facets Session with This Class](#)

Attributes	Values
rdf:type	http://bio2rdf.org/drugbank_vocabulary:Drug http://bio2rdf.org/drugbank_vocabulary:Resource http://bio2rdf.org/drugbank_vocabulary:Small-molecule
rdfs:label	Diclofenac [drugbank:DB00586]
rdfs:seeAlso	http://www.drugbank.ca/drugs/DB00586 http://www.drugs.com/cdi/diclofenac-drops.html http://www.rxlist.com/cgi/generic/diclofen.htm
owl:sameAs	http://identifiers.org/drugbank/DB00586
dcterms:title	Diclofenac
dcterms:description	A non-steroidal anti-inflammatory agent (NSAID) with antipyretic and analgesic actions. It is primarily available as the sodium salt. [PubChem]
dcterms:identifier	drugbank:DB00586
void:inDataset	http://bio2rdf.org/drugbank_resource:bio2rdf.dataset.drugbank.R3
http://bio2rdf.org...bulary:identifier	DB00586
http://bio2rdf.org...abulary:namespace	drugbank
http://bio2rdf.org...df_vocabulary:uri	http://bio2rdf.org/drugbank:DB00586
http://bio2rdf.org...x-identifiers.org	http://identifiers.org/drugbank/DB00586
http://bio2rdf.org...bulary:absorption	http://bio2rdf.org/drugbank_resource:af3a8b347e732d3c3b48a5428a6160e0
http://bio2rdf.org...affected-organism	http://bio2rdf.org/drugbank_vocabulary:Humans-and-other-mammals

Discover connections across datasets



Efficiently find and explore data

The screenshot displays the OntoForce web application interface. At the top, a navigation bar shows a search path: **asthma** (Disease, 336) → **Drug/chemi...** (1.1k, 54) → **Links** (845k). The main content area is divided into several panels:

- Data Sources (12)**: A list of data sources with counts. MeSH is highlighted with 54 results.
- Substance Type (4)**: A list of substance types. Small_molecule is highlighted with 54 results.
- Group (9)**: A list of groups. 'approved' is highlighted with 54 results.
- Manufacturer (280)**: A list of manufacturers. 'Sandoz inc' is highlighted with 12 results.
- Drug/chemical (54)**: A list of specific chemical entities. The first three are Indacaterol, Lodosyn, and aspirin, each with its molecular formula and source information.

The right sidebar shows an **Overview** tab and a **Clear filters** button.

Examine the facts and their provenance

← Go back to all Drug/chemical results 1/54

Data Sources

ChEMBL DrugBank ChEBI IUPHAR family

UNII MeSH PubChem-substance

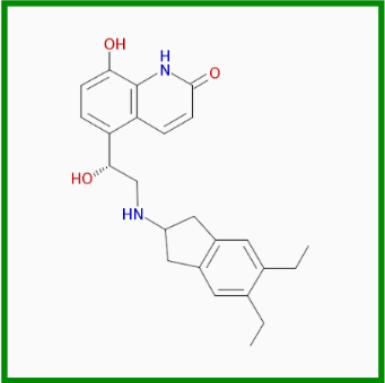
SureChEMBL UMLS

☐ Indacaterol

Linked Data

Publication (223)	Clinical study (90)	Drug/chemi... (66)
Patent (61)	Relation (46)	Uncategoriz... (21)
Protein (10)	Medicine (9)	Disease (4)

Structure



Molecular Formula

C₂₄H₂₈N₂O₃

Synonym

5-(2-{5,6-Diethylindan-2-ylamino}-1-hydroxyethyl)-8-hydroxy-1H-quinolin-2-one
5-(2-{5,6-Diethylindan-2-ylamino}-1-hydroxyethyl)-8-hydroxy-1H-quinolin-2-one
5-[(1R)-2-[(5,6-diethyl-2,3-dihydro-1H-inden-2-yl)amino]-1-hydroxyethyl]-8-hydroxyquinolin-2(1H)-one
8-Hydroxy-5-[(R)-1-hydroxy-2-{5,6-diethylindan-2-ylamino}-ethyl]-1H-quinolin-2-one
Arcanta Neohaler Cream

Description

A monohydroxyquinoline that consists of 5-[(1<stereo>R</stereo>)-2-amino-1-hydroxyethyl]-8-hydroxyquinolin-2-one having a 5,6-diethylindan-2-yl group attached to the amino function. Used as the maleate salt for treatment of chronic obstructive pulmonary disease.
Indacaterol is a novel, ultra-long-acting, rapid onset β(2)-adrenoceptor agonist developed for Novartis for the once-daily management of asthma and chronic obstructive pulmonary disease. It was

CAS Registry Num...

312753-06-3

InChIKey

QZZUEBNBZAPZLX-QFIPXVFZSA-N
QZZUEBNBZAPZLX-QFIPXVFZSA-N
QZZUEBNBZAPZLX-QFIPXVFZSA-N
QZZUEBNBZAPZLX-QFIPXVFZSA-N

Smiles Representa...

CCC1=C(CC)C=C2CC(C2=C1)NC[C@H](O)C1=C2C=CC(=O)NC2=C(O)C=C1
CCC1=CC2=C(CC(C2)NC[C@H](O)C2=C3C=CC(=O)NC3=C(O)C=C2)C=C1CC(Cc1cc2CC(Cc2cc1CC)NC[C@H](O)c1ccc(c2c1ccc(O)[nH]2)O)O
CCc1cc2CC(Cc2cc1CC)NC[C@H](O)c1ccc(O)c2[nH]c(O)ccc12
CCc1cc2c(cc1CC)CC(C2)NC[C@H](O)c1ccc(c4c3ccc(O)[nH]4)O)O

Substance Type

Small_molecule
Small_molecule
Synthetic organic

Group

approved
approved

Highest Developm...

4

Indication

Asthma
Chronic obstructive airway disease
For the long term, once-daily-dosing maintenance of airflow obstruction in patients with chronic obstructive pulmonary disease (COPD), including chronic bronchitis and/or emphysema.
Heart failure
Lung diseases, obstructive

External Link

<http://identifiers.org/chebi/CHEBI:68575>
<http://www.drugbank.ca/drugs/DB05039>
<http://www.drugs.com/international/indacaterol.html>

Develop timely knowledge portals

Bio2RDF Ebola Virus Knowledgebase

About Bio2RDF

Download Ebola Virus Knowledgebase

SPARQL Endpoint

Contact

Ebola Virus Genomic Information

Genome Wheel

Ebola Virus (EBOV) Genes and Protein Domains

(683 Publications (P)) (1783 Ligands (L))

- [NP_066243.1] NP Nucleoprotein (P: 120) (L: 141)
 - [CDD:147601] Ebola nucleoprotein (P: 20) (L: 2)
- [NP_066244.1] VP35 Polymerase Complex Protein (P: 45) (L: 589)
 - [CDD:145320] Filoviridae VP35 (P: 20) (L: 0)
- [NP_066245.1] VP40 Matrix Protein (P: 29) (L: 420)
 - [CDD:116068] Matrix protein VP40 (P: 29) (L: 11)
- [NP_066246.1] GP Spike Glycoprotein (P: 40) (L: 0)
 - [CDD:110602] Filovirus glycoprotein (P: 20) (L: 2)
 - [CDD:197367] heptad repeat 1-heptad repeat 2 region of the transmembrane subunit of Filoviridae viruses, Ebola virus and Marburg virus, and related domains (P: 104) (L: 112)
- [NP_066247.1] GP Small Secreted Glycoprotein (P: 33) (L: 23)
 - [CDD:110602] Filovirus glycoprotein (P: 20) (L: 2)
- [NP_066248.1] GP Second Secreted Glycoprotein (P: 20) (L: 16)
 - [CDD:110602] Filovirus glycoprotein (P: 20) (L: 2)

Annotations

All Annotations

Interpro Annotations

[IPR014459] Minor nucleoprotein VP30, Filoviridae type	🔗
[IPR002953] Filoviridae VP35 protein	🔗
[IPR002561] Filoviruses glycoprotein, extracellular domain	🔗
[IPR014023] Mononegavirales RNA-directed RNA polymerase catalytic domain	🔗
[IPR014625] Filoviruses glycoprotein	🔗
[IPR017235] RNA-directed RNA polymerase L, filovirus	🔗
[IPR025786] RNA-directed RNA polymerase L	🔗
[IPR008609] Ebola nucleoprotein	🔗
[IPR026890] Mononegavirales mRNA-capping domain V	🔗
[IPR008986] EV matrix protein	🔗
[IPR009433] Filovirus membrane-associated VP24	🔗

Biological Process Annotations

[0016032] viral process	🔗
[0006139] nucleobase-containing compound metabolic process	🔗

3D Molecular Structure View

3D Structure View of 4LDD 🔗



Publications

All Publications

[1] Formate assay in body fluids: application in methanol poisoning. 🔗

[10023770] Crystal structure of an MHC class I presented glycopeptide that generates carbohydrate-specific CTL. 🔗

[10023771] Crystal structures of two H-2Db/glycopeptide complexes suggest a molecular basis for CTL cross-reactivity. 🔗

[10049310] Structural dynamics of ligand diffusion in the protein matrix: A study on a new myoglobin mutant Y(B10) Q(E7) R(E10). 🔗

[10074941] Crystal structure of a scavenger receptor cysteine-rich domain sheds light on an ancient superfamily. 🔗

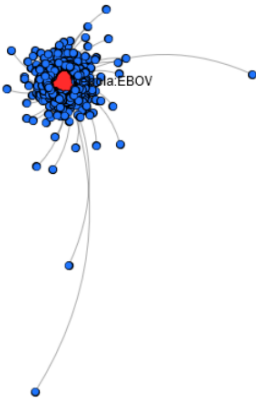
[10077567] Core structure of the envelope glycoprotein GP2 from Ebola virus at 1.9-Å resolution. 🔗

[10080897] Crystallographic evaluation of internal motion of human alpha-lactalbumin refined by full-matrix least-squares method. 🔗

[10089503] Structure determination of echovirus 1. 🔗

[10097078] Crystal structure of human p32, a doughnut-shaped acidic 🔗

MeSH Terms View



Ligands

All Ligands

[017] (3R,3AS,6AR)-HEXAHYDROFURO[2,3-B]FURAN-3-YL(1S,2R)-3-[[[4-AMINOPHENYL]SULFONYL]([SOBUTYL]AMINO)-1-BENZYL-2-HYDROXYPROPYLCA 🔗

[097] (2S,3R)-N-4-~[(1S)-2,2-DIMETHYL-1-(METHYLCARBAMOYL)PROPYL]-N-1~,2- 🔗

3-(2-METHYLPROPYL)BUTANEDIAMIDE 🔗

[099] (2R)-N-4~-HYDROXY-2-(3-HYDROXYBENZYL)-N-1-~[(1S,2R)-2-HYDROXY-2,3-1H-INDEN-1-YL]BUTANEDIAMIDE 🔗

[111] (1N)-4-N-BUTOXYPHENYL SULFONYL-(2R)-N-HYDROXYCARBOXAMIDO- 🔗

(4S)-METHANESULFONYLAMINO-PYRROLIDINE 🔗

[185] (6-[4(AMINOMETHYL)-2,6-DIMETHYLPHENOXY]-2-[[4-(AMINOMETHYL)PHEN' 🔗

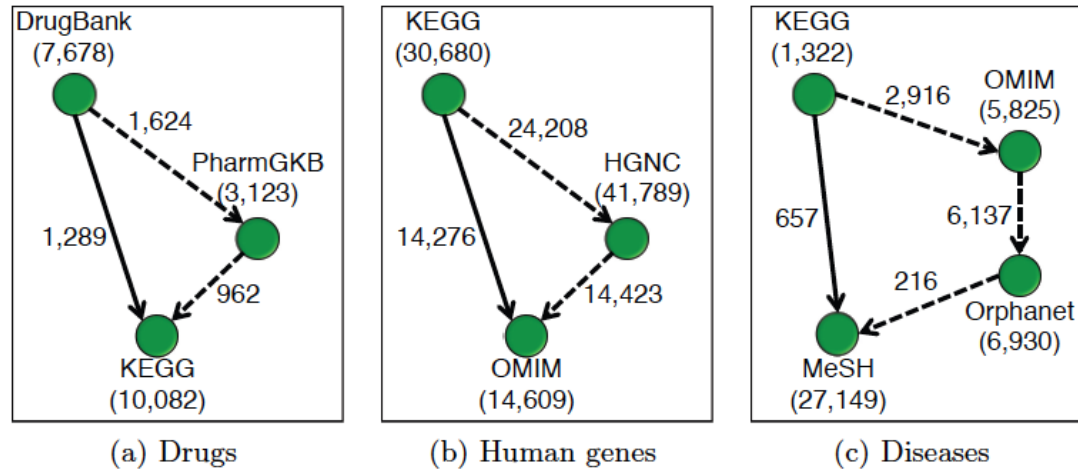
5-BROMOPYRIMIDIN-4-YL]METHANOL 🔗

[1MC] 1-METHYLCYTOSINE 🔗

[1PG] 2-(2-[2-(2-METHOXY-ETHOXY)-ETHOXY]-ETHOXY)-ETHOXY-ETHANOL 🔗

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



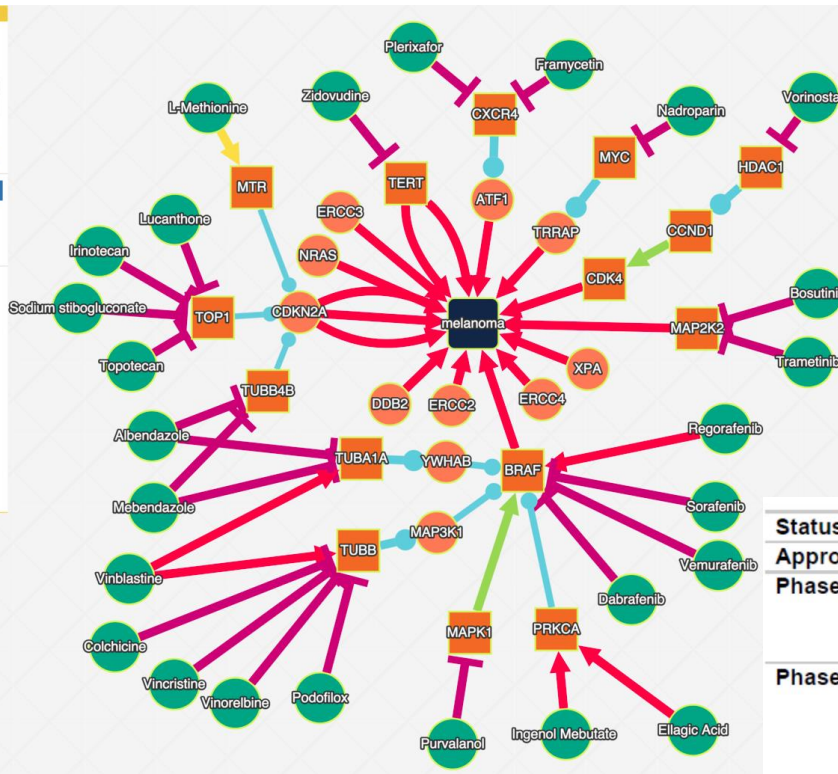
	Direct links	Transitive paths	Identical ending entities	Different entities	Missing direct	Missing transitive	Total
Drugs	1,289	954	946	6	2	343	1,297
Human genes	14,276	14,250	14,236	5	9	40	14,290
Diseases	657	33	8	18	7	649	682

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



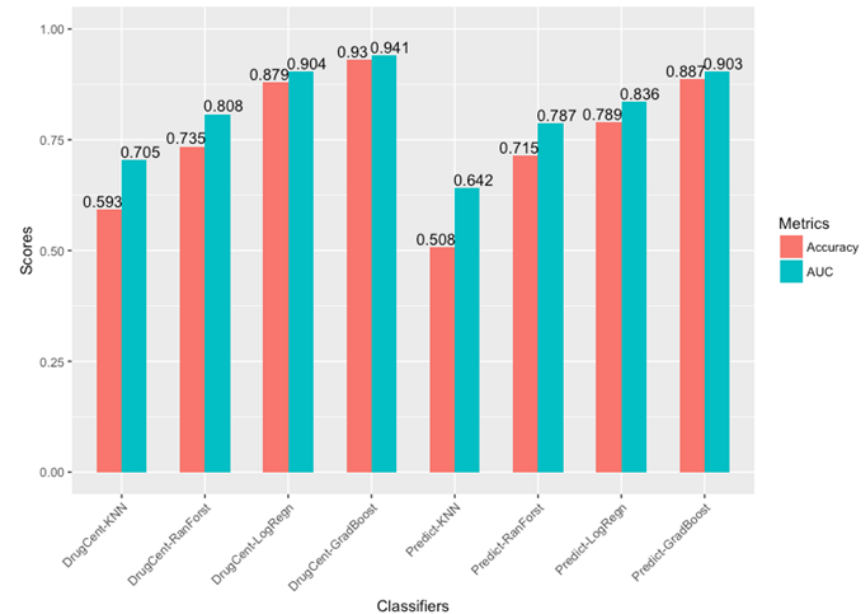
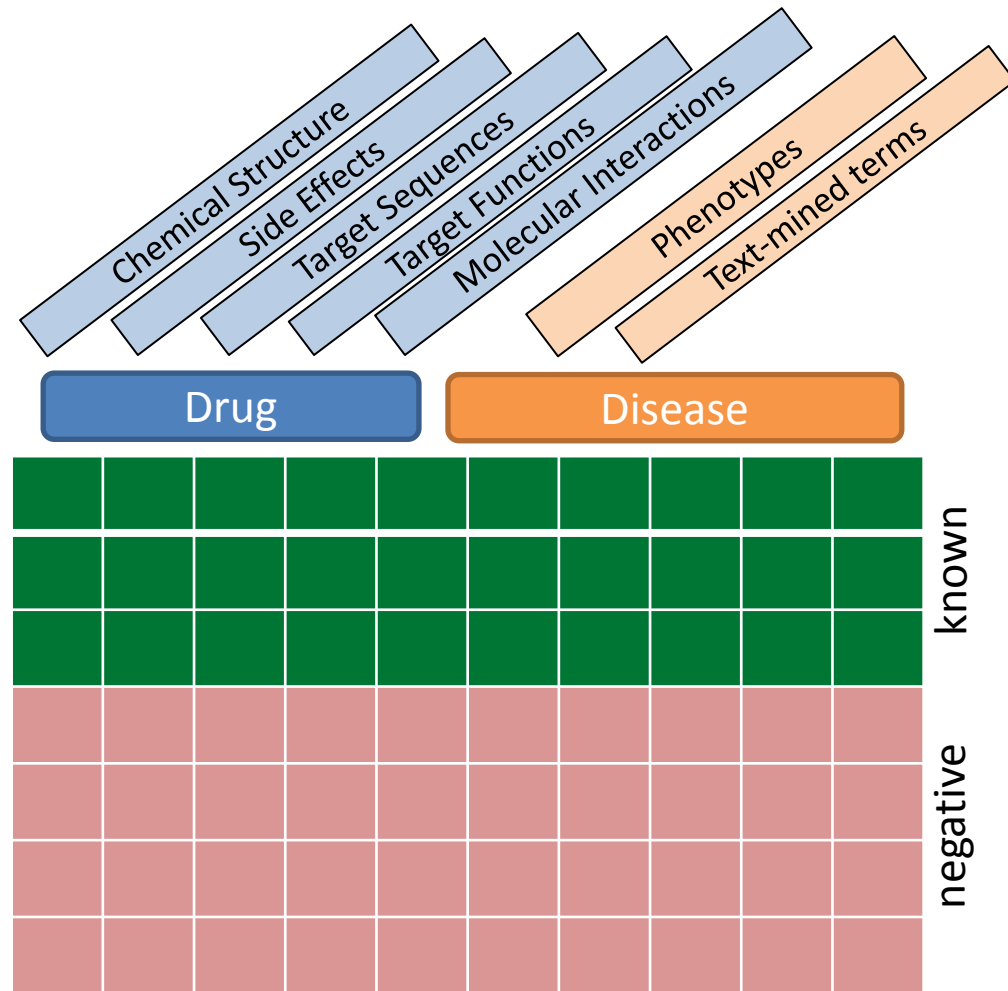
And validate them against pipelines for drug discovery

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
	Topotecan ⁴⁴	CDKN2A	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
	Sodium stibogluconate ⁴⁵	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
	Plerixafor ²⁷	CXCR4	3	0.97
In Vivo	Vincristine ²³	MAP kinase	3	0.93
	L-Methionine ⁴⁹	CDKN2A	3	0.93
	Mebendazole ⁵⁰	CDKN2A	3	0.93
	Regorafenib ¹⁵	BRAF	2	0.98

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

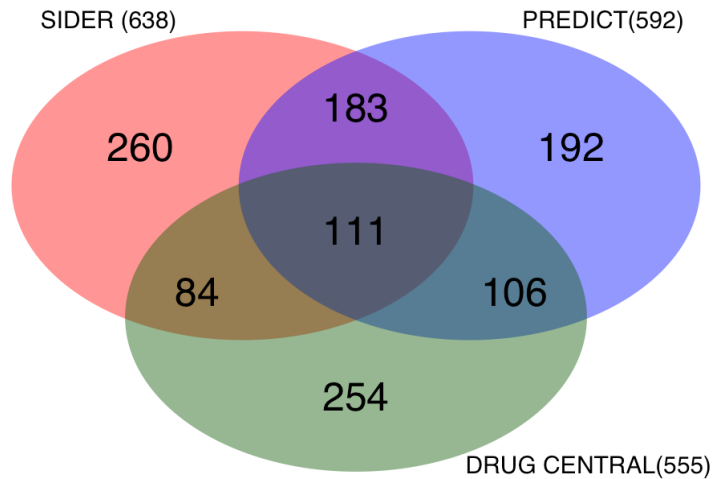
Find new uses for existing drugs

using machine learning

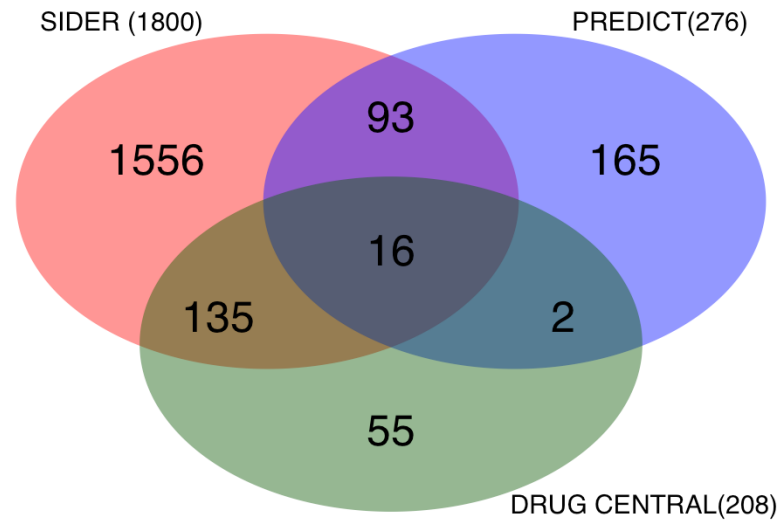


Examine overlap in “gold standards”

drugs



diseases



Uncover evidence in a transparent manner



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 DISTINCT ?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.

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 be even more collaborative and interdisciplinary.



michel.dumontier@maastrichtuniversity.nl

Website: <http://dumontierlab.com>

Presentations: <http://slideshare.com/micheldumontier>



Maastricht University