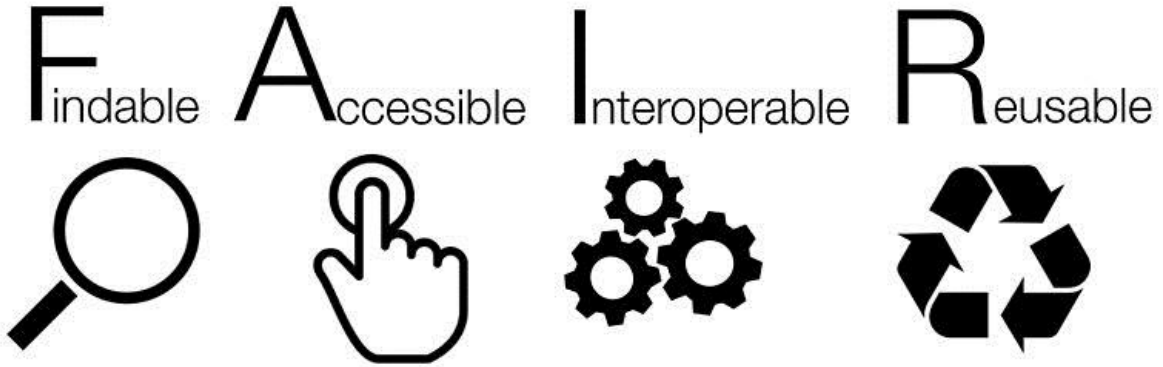
Accelerating biomedical discovery with an Internet of FAIR data and services



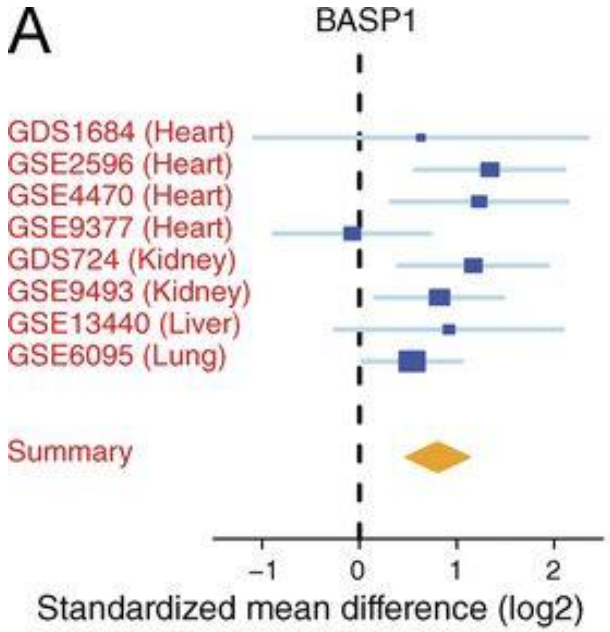
Michel Dumontier, Ph.D.
Distinguished Professor of Data Science
Director, Institute of Data Science

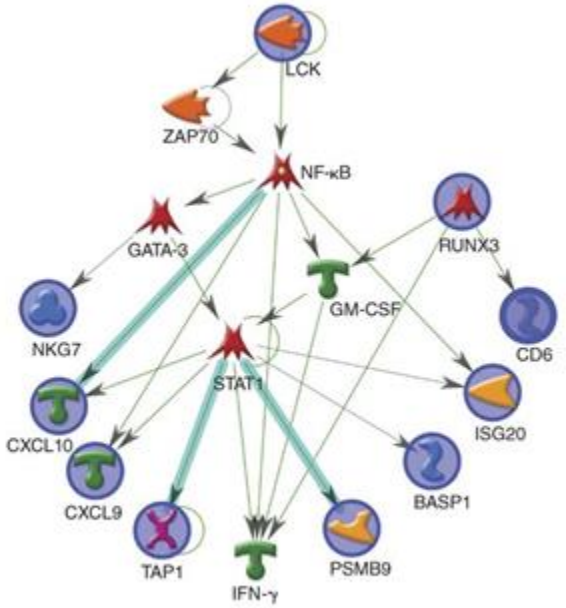
Maastricht University

An increasing number of discoveries are made using *already* available data

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





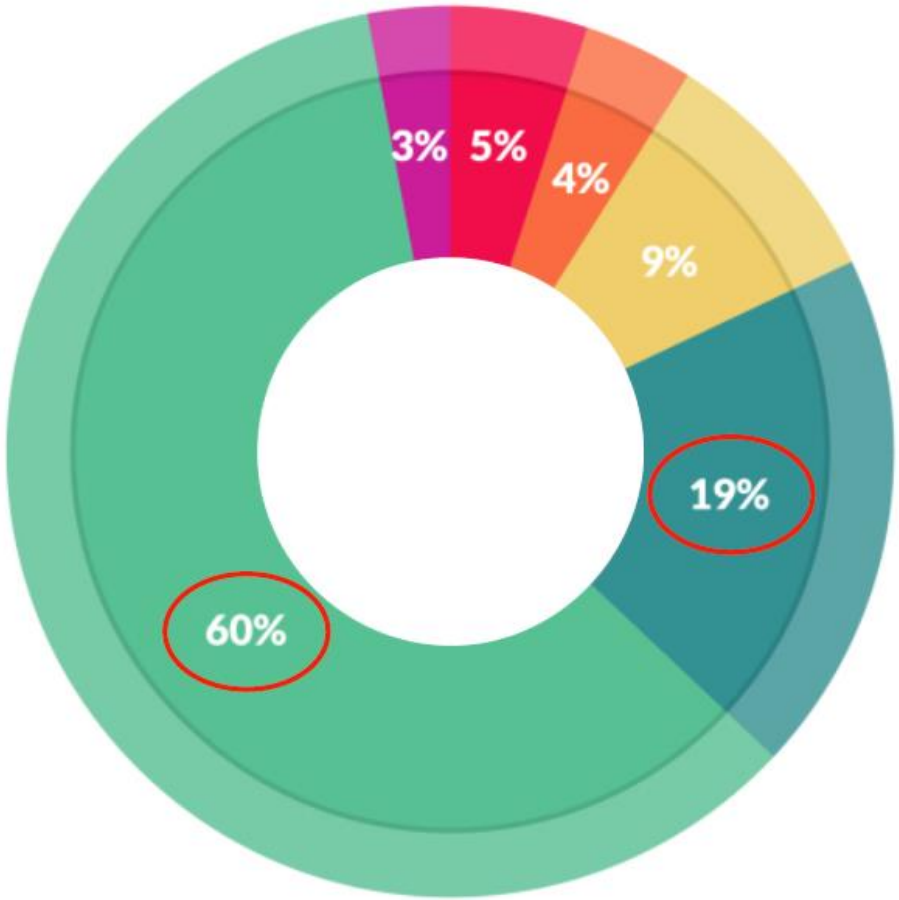
Main Findings:

1. CRM genes **predicted future injury** to a graft
2. Mice treated with **drugs against the CRM genes extended graft survival**
3. Retrospective **EHR analysis supports treatment prediction**

Key Observations:

1. **Meta-analysis** offers a **more reliable estimate** of the magnitude of the effect
2. Data can be used to **generate and support/dispute new hypotheses**

However, *significant effort* is still needed to find the right dataset(s), make sense of them, and use for a new purpose



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

Our ability to reproduce landmark studies is surprisingly low:

39% (39/100) in psychology[1]
21% (14/67) in pharmacology[2]
11% (6/53) in cancer[3]
unsatisfactory in machine learning[4]

[1]doi:10.1038/nature.2015.17433 [2]doi:10.1038/nrd3439-c1 [3]doi:10.1038/483531a [4]https://openreview.net/pdf?id=By4l2PbQ-

Most published research findings are false.

- John Ioannidis, Stanford University

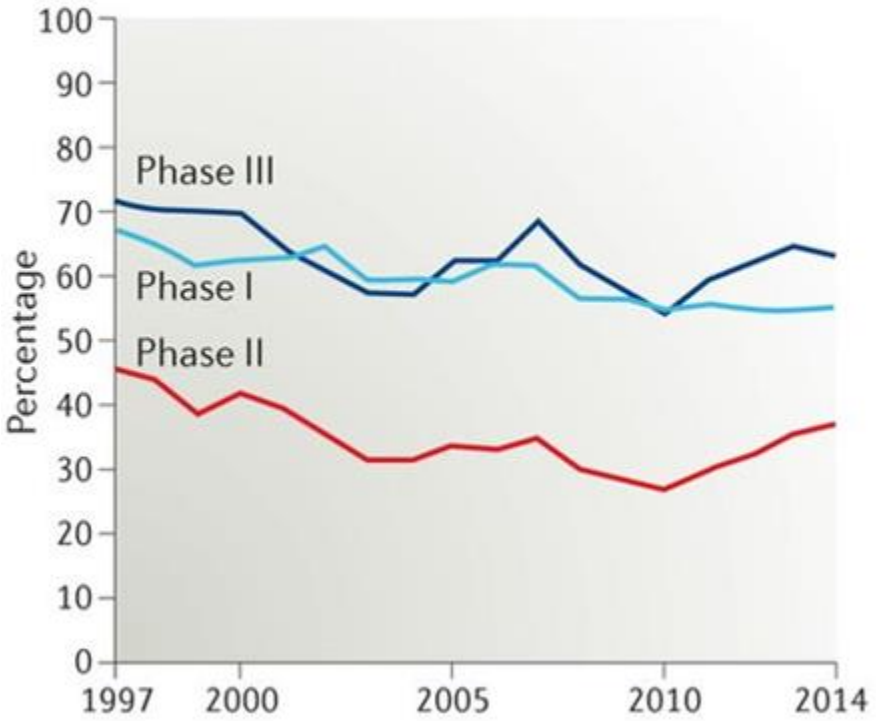
PLoS Med 2005;2(8): e124.

THE CLINICAL-TRIAL CLIFF

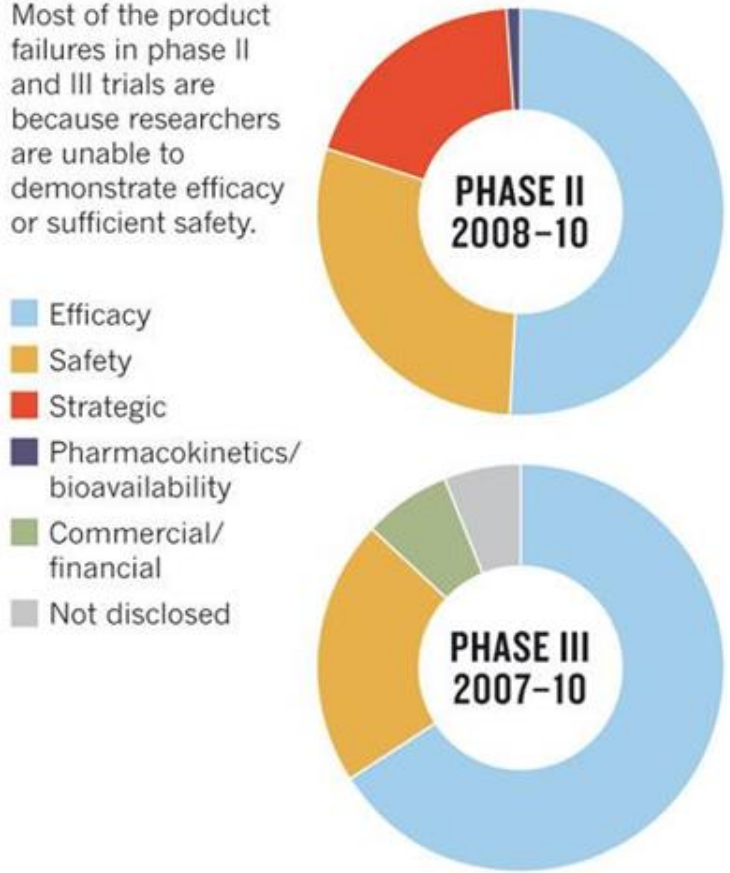
Drug companies are removing more compounds from the pipeline at all levels of testing than ever before.

Success rates by phase

Percentage likelihood of moving to next phase, 3-year rolling average*

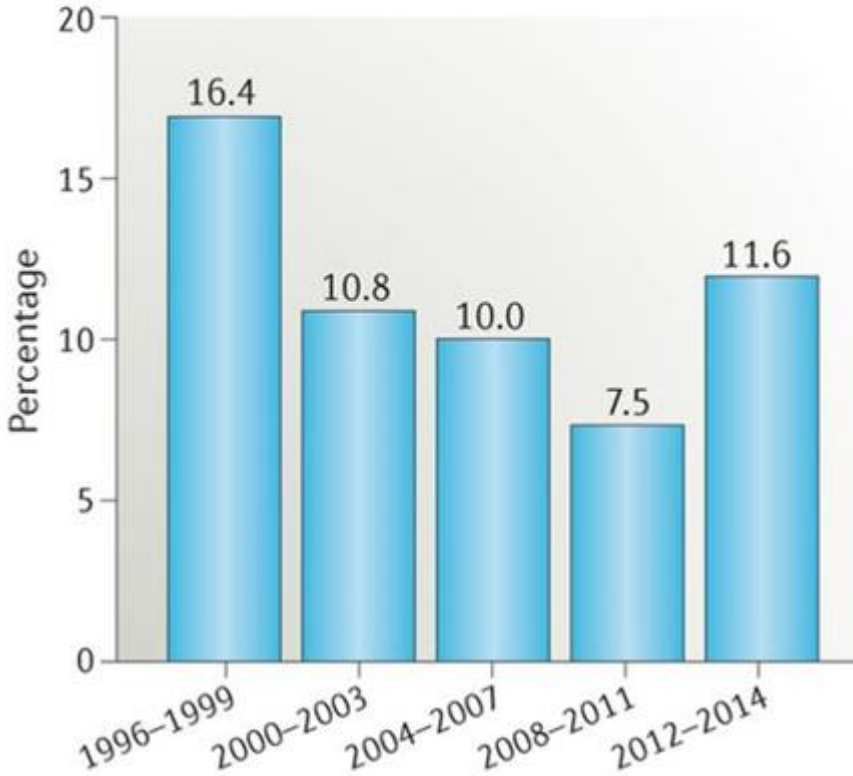


Most of the product failures in phase II and III trials are because researchers are unable to demonstrate efficacy or sufficient safety.

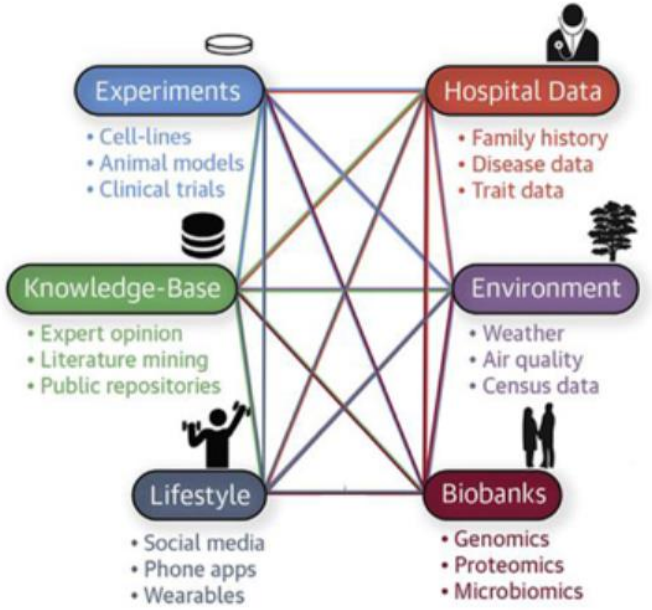


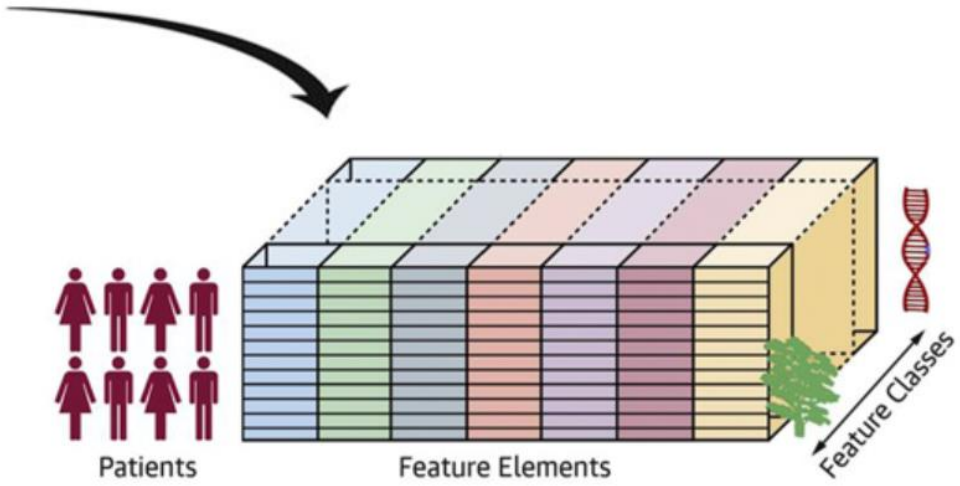
Cumulative success rate Phase I to launch

Percentage likelihood of moving from Phase I to launch



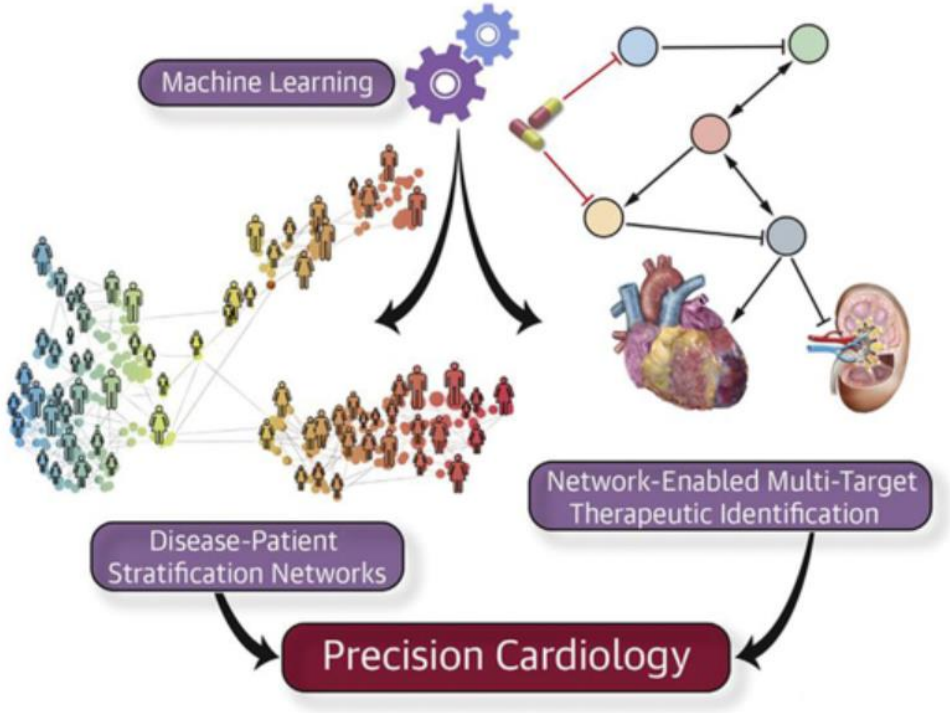
Nature Reviews | Drug Discovery





How will we ever get to

Precision Medicine?



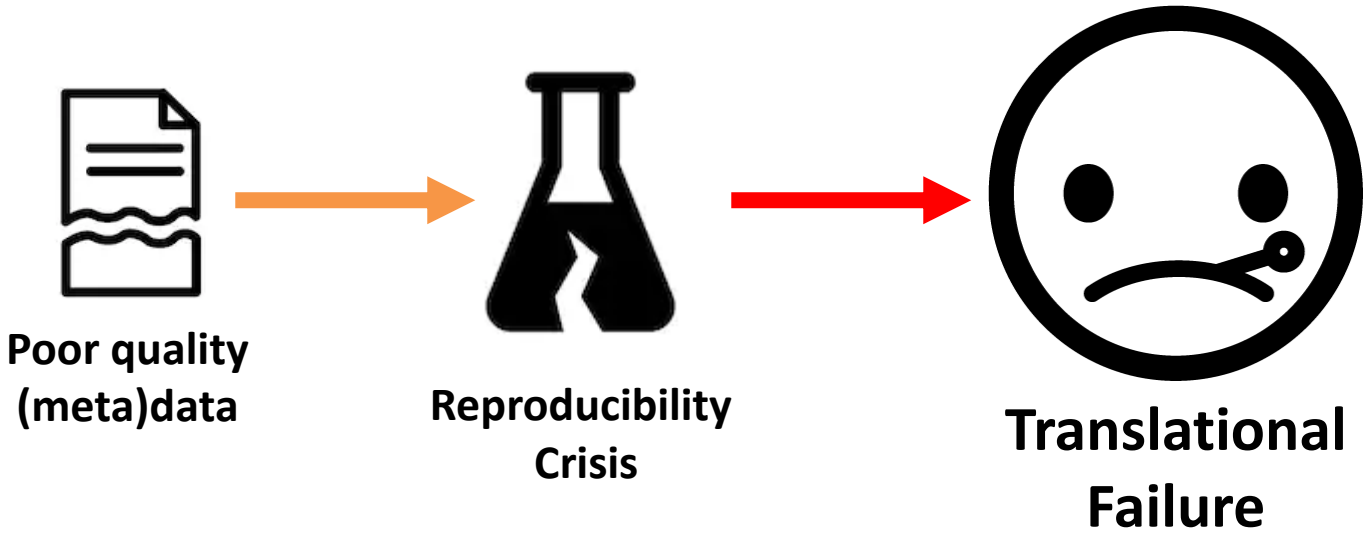
Maastricht University

Broken windows theory

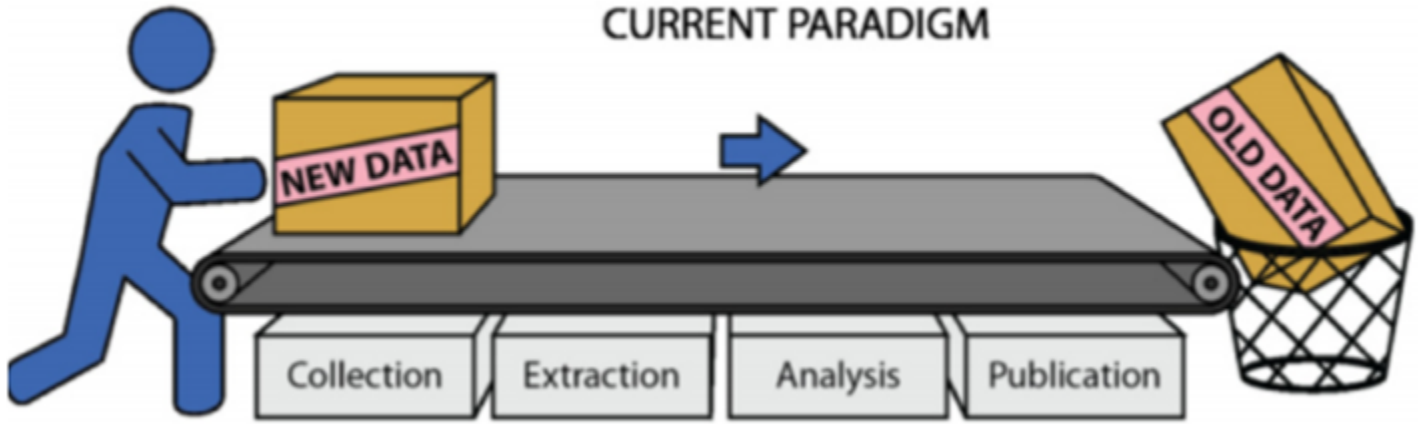
visible signs of crime, anti-social behavior, and civil disorder create an urban environment that encourages further crime and disorder, including serious crimes

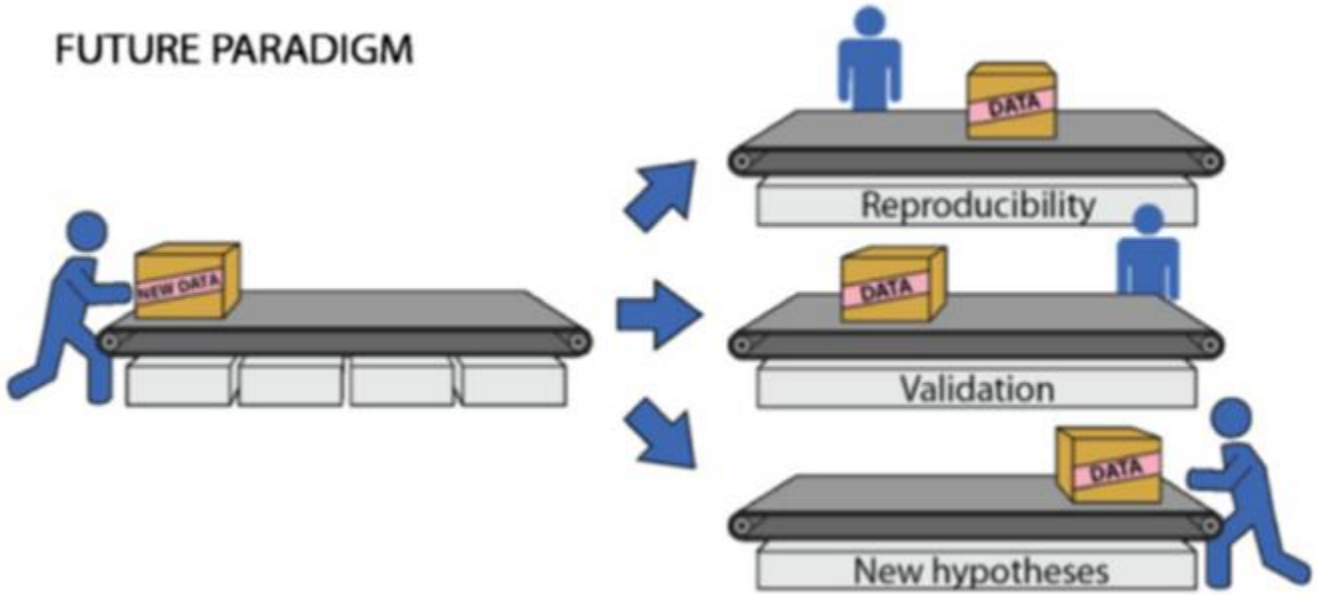
Inadequate reusability theory

Poor quality metadata and the inaccessibility of original research results make it less likely to reproduce original work, resulting in an ineffective translation of research into useful applications



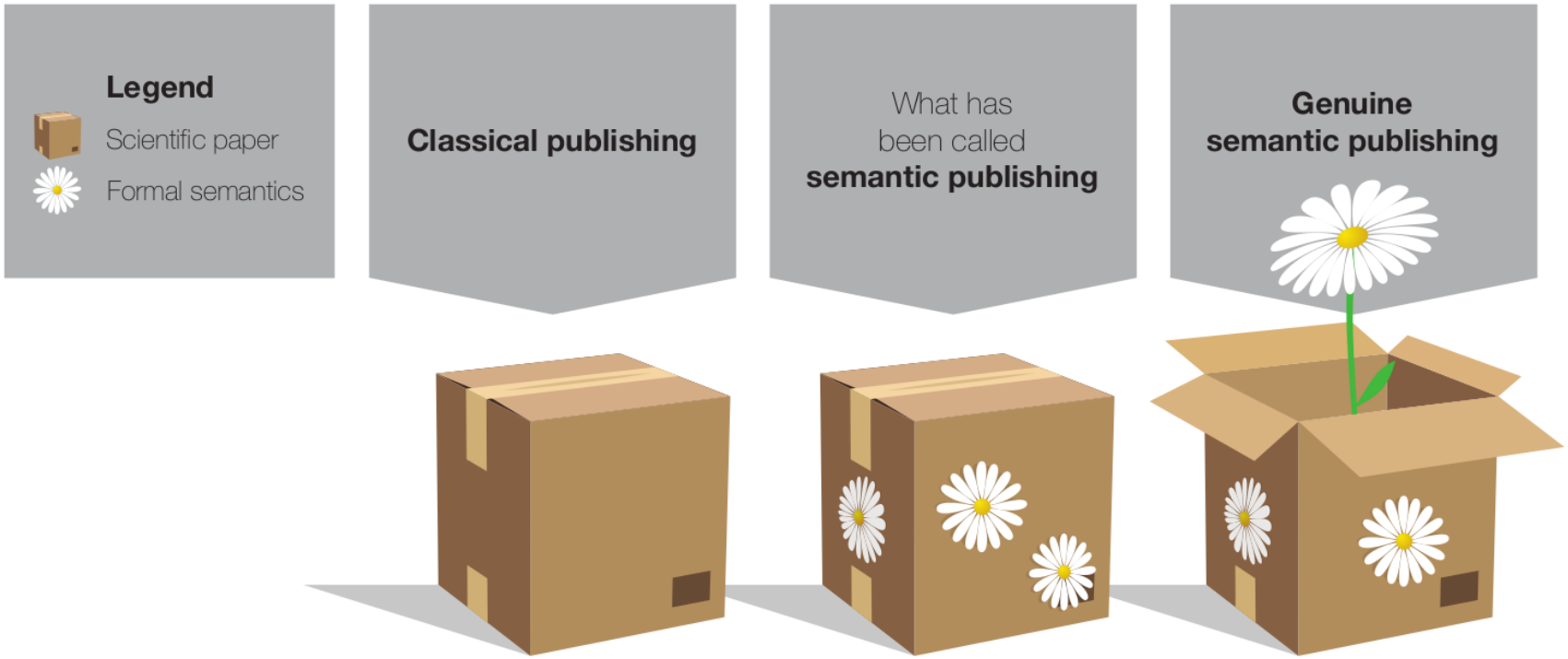
It’s time to completely rethink how we perform research (and how we report it)





Lambin et al. Radiother Oncol. 2013. 109(1):159-64. doi: 10.1016/j.radonc.2013.07.007

Rethinking Publishing Scientific Research



Genuine Semantic Publishing

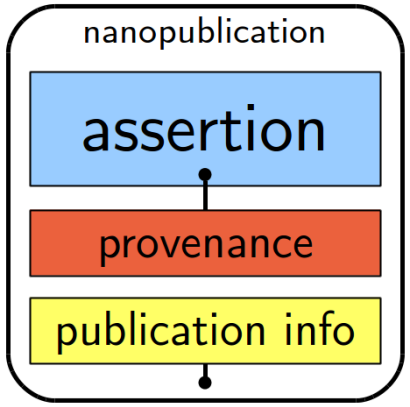
by Tobias Kuhn and Michel Dumontier

Content:

- as PDF
- as HTML/Dokieli
- as HTML/RASH
- as RDF/Turtle
- as RDF/TriG

Data Science. 2017 1(1-2):139-154. DOI: 10.3233/DS-170010
http://www.tkuhn.org/pub/sempub/

A growing network of nanopublications



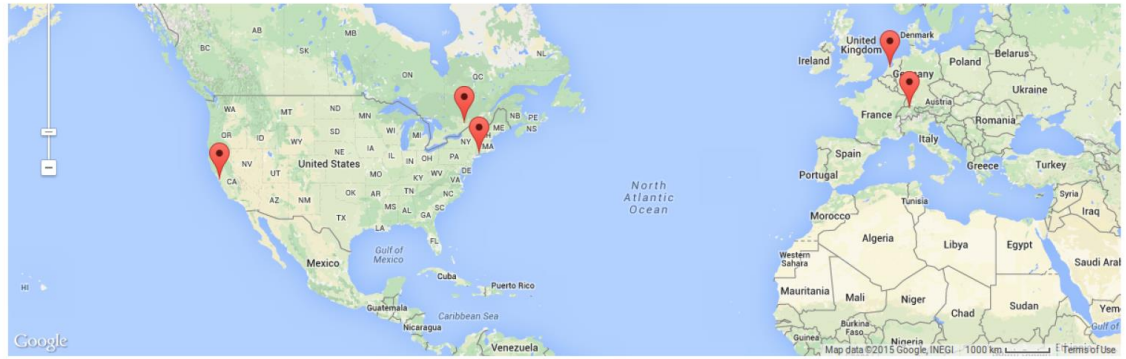
Trusty URI

```
@prefix this: <http://purl.org/np/RAzquSkwsTAZm61nReG6MOjXEXUx8fNVfdWnAzyn6sOhU> .
...

sub:Assertion {
  sub:Interaction occurs-in: obo:ENVO_01000240 ;
    has-participant: sub:Organism_1 , sub:Organism_2 ;
    a obo:GO_0044419 ;
    prov:atTime "1962-12-01T00:00:00Z"^^xsd:dateTime .
  sub:Organism_1 eats: sub:Organism_2 ;
    rdfs:label "Picoides villosus" .
  sub:Organism_2 a itis:114936 ;
    rdfs:label "Ips" .
}

sub:Provenance {
  sub:Assertion prov:wasDerivedFrom sub:Study .
  sub:Study dcterms:bibliographicCitation "Otvos, I. S. and R. W. Stark. 1985. Arthropod food of
some forest-inhabiting birds. Canadian Entomologist 117:971-990." .
}

sub:Pubinfo {
  this: dcterms:license <https://creativecommons.org/licenses/by/4.0/> ;
    pav:createdBy <https://doi.org/10.5281/zenodo.1212599> ;
    prov:wasDerivedFrom <https://github.com/hurlbertlab/dietdatabase> .
  <https://github.com/hurlbertlab/dietdatabase> dcterms:bibliographicCitation "Allen Hurlbert.
2017. Avian Diet Database." .
}
```

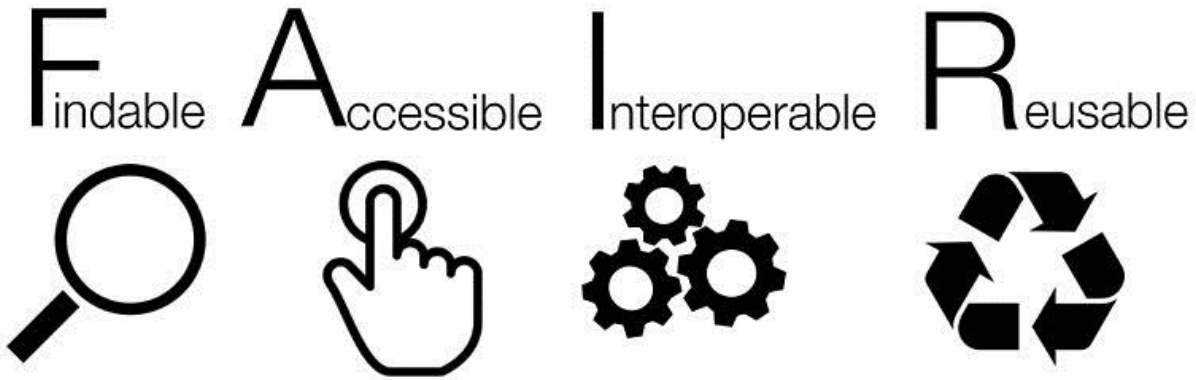


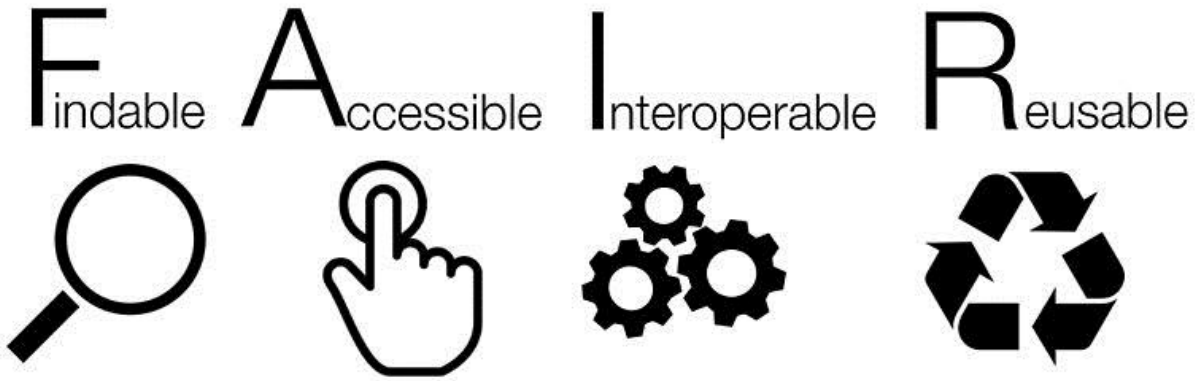
URL	Status	Success Ratio	Avg Response Time	Distance	Last Seen OK	Nanopub Count	Serve
http://nanopubs.semanticscience.org/	OK	99.993126%	257 ms	6129 km	April 9, 2015 11:18:38 AM CEST	5252183	Ottawa, Ca
http://ristretto.med.yale.edu:8080/nanopub-server/	OK	99.87604%	233 ms	6212 km	April 9, 2015 11:18:40 AM CEST	5252183	New Haver
http://np.inn.ac/	OK	99.993126%	4 ms	0 km	April 9, 2015 11:18:40 AM CEST	5252183	Zurich, Sw
http://nanopub-server.ops.labs.vu.nl/	OK	96.30011%	62 ms	615 km	April 9, 2015 11:18:40 AM CEST	5252183	Amsterdan
http://nanopubs.stanford.edu/nanopub-server/	OK	100.0%	456 ms	9393 km	April 9, 2015 11:18:42 AM CEST	5252183	Stanford,

Kuhn T., Chichester C., Krauthammer M., Dumontier M. (2015) **Publishing Without Publishers: A Decentralized Approach to Dissemination, Retrieval, and Archiving of Data**. In: Arenas M. et al. (eds) The Semantic Web - ISWC 2015. ISWC 2015. Lecture Notes in Computer Science, vol 9366. Springer, Cham

We need a new *social contract*, supported by *legal* and *technological* infrastructure to make digital resources available in a responsible manner

Human Machine collaboration will be crucial to our future success


Findable
Accessible
Interoperable
Reusable



An international, bottom-up paradigm for the discovery and reuse of digital content

for the machines that people use

SCIENTIFIC DATA

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

http://www.nature.com/articles/sdata201618

Last updated: Mon, 9 Sep 2019 14:18:07 GMT

Total citations

66k	718	957
Article Accesses	Web of Science	CrossRef

Online attention

Altmetric score (what's this?)

- Tweeted by **1230**
- Blogged by **77**
- On **18** Facebook pages
- Mentioned in **6** Google+ posts
- Picked up by **82** news outlets

Show more

This Altmetric score means that the article is:

- in the 99th percentile (ranked 76th) of the 264,573 tracked articles of a similar age in all journals
- in the 1st percentile (ranked 1st) of the 1 tracked articles of a similar age in *Scientific Data*

FAIR: Impact

EUROPEAN COMMISSION

Press Release Database

European Commission > Press releases database > Press Release details

European Commission - Statement

G20 Leaders' Communique Hangzhou Summit

Hangzhou, 5 September 2016

1. We, the Leaders of the G20, met in Hangzhou, China on 4-5 September 2016.

12. To achieve innovation-driven growth and the creation of innovative ecosystems, we support dialogue and cooperation on innovation, which covers a wide range of domains with science and technology innovation at its core. We deliver the G20 2016 Innovation Action Plan. We commit to pursue pro-innovation strategies and policies, support investment in science, technology and innovation (STI), and support skills training for STI - including support for the entry of more women into these fields - and mobility of STI human resources. We support effort to promote voluntary knowledge diffusion and technology transfer on mutually agreed terms and conditions. Consistent with this approach, we support appropriate efforts to promote open science and facilitate appropriate access to publicly funded research results on findable, accessible, interoperable and reusable (FAIR) principles. In furtherance of the above, we emphasize the importance of open trade and investment regimes to facilitate innovation through intellectual property rights (IPR) protection, and improving public communication in science and technology. We are committed to foster exchange of knowledge and experience by supporting an online G20 Community of Practice within the existing Innovation Policy Platform and the release of the 2016 G20 Innovation Report.

NIH

Big Data to Knowledge(BD2K)

DCPPC

DATA COMMONS PILOT PHASE CONSORTIUM

FAIR in a nutshell

FAIR aims to create **social** and **economic impact** by facilitating the discovery and reuse of **digital resources** through a set of requirements:

- **unique identifiers** to retrieve all forms of digital content and knowledge
- **high quality meta(data)** to enhance discovery of digital resources
- **use of common vocabularies and ontologies** to share terms and facilitate query
- **establishment of community standards** for more facile knowledge utilisation
- **detailed provenance** to provide context and reproducibility
- **registered in appropriate repositories** with high quality metadata for future content seekers
- **social and technological commitments** to realize reliable access
- **simpler terms of use** to clarify expectations and intensify innovation

G8 science ministers statement: London, 12 June 2013

G8 science ministers written statement from their UK meeting on international issues that need global cooperation.

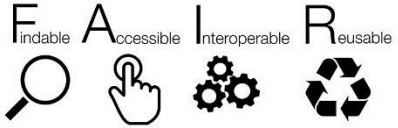
Published 13 June 2013

i. To the greatest extent and with the fewest constraints possible publicly funded scientific research data should be open, while at the same time respecting concerns in relation to privacy, safety, security and commercial interests, whilst acknowledging the legitimate concerns of private partners.

ii. Open scientific research data should be easily discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards.

FAIR != Open

Open as possible
closed as is necessary



COMMENT · 04 JUNE 2019 · CORRECTION 05 JUNE 2019

Make scientific data FAIR

All disciplines should follow the geosciences and demand best practice for publishing and sharing data, argue Shelley Stall and colleagues.

Shelley Stall ✉, Lynn Yarmey, Joel Cutcher-Gershenfeld, Brooks Hanson, Kerstin Lehnert, Brian Nosek, Mark Parsons, Erin Robinson & Lesley Wyborn

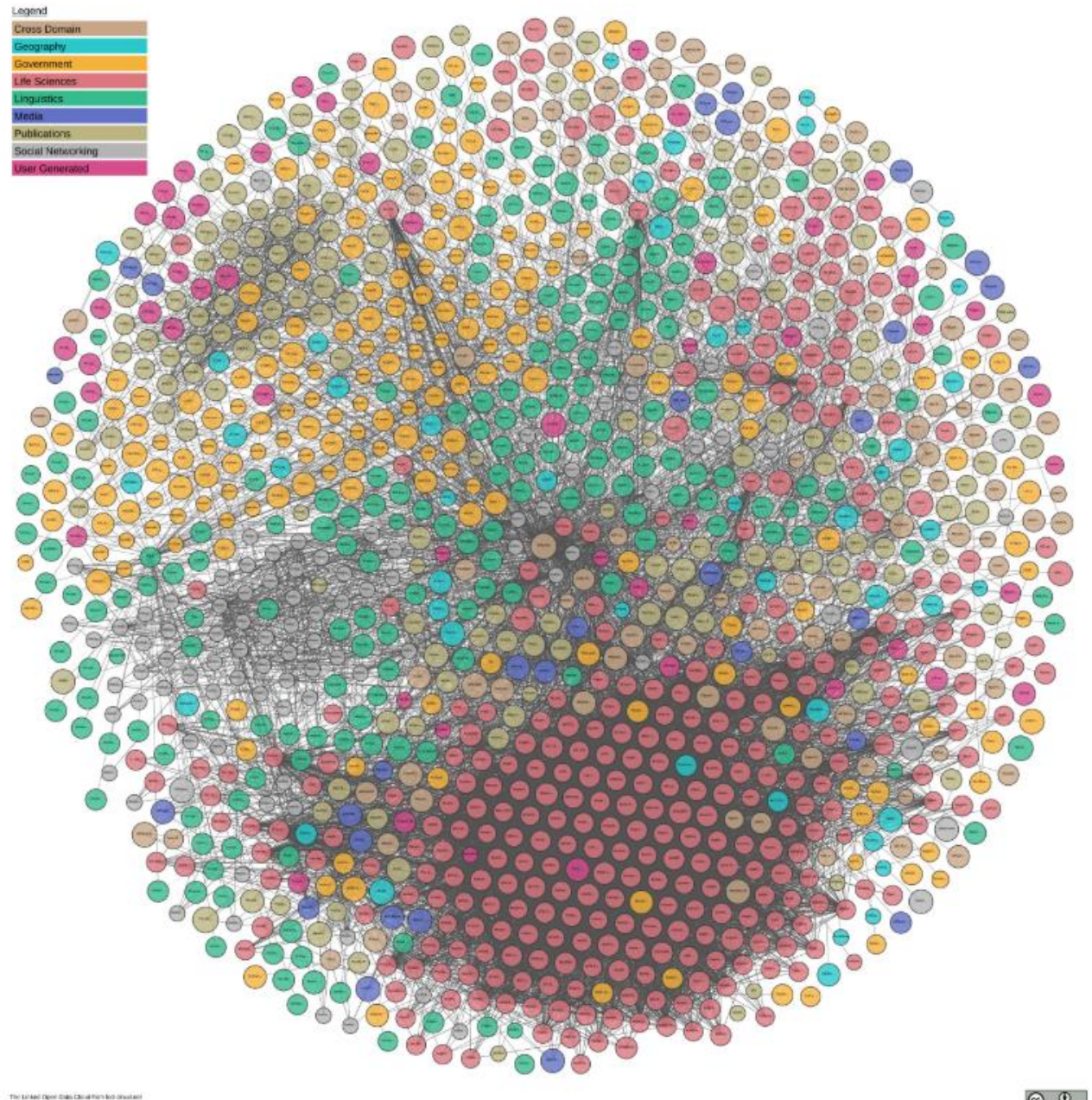
That's why more than 100 repositories, communities, societies, institutions, infrastructures, individuals and publishers (including the Springer Nature journals *Nature* and *Scientific Data*) have signed up since last November to the Enabling FAIR Data Project's Commitment Statement in the Earth, Space, and Environmental Sciences for depositing and sharing data (see go.nature.com/2wv2jxd). The principles state that research data should be 'findable, accessible, interoperable and reusable' (FAIR)[2]. The idea is not new, but aligning this broad community around common data guidelines is a radical step.

Why Should *I* Go FAIR?

- Makes it easier for **me to use my own data for a new purpose**
- Makes it easier for **other people to find, use and *cite* my data**, and for them to understand what I expect in return
- Makes it easier/possible for people to **verify my work**
- Ensure that the **data are available in the future**, especially as I may not want the responsibility
- **Satisfy the expectations** around data management from institution, funding agency, journal, my peers

Let’s build the Internet of FAIR data and services

The Linked Open Data Cloud

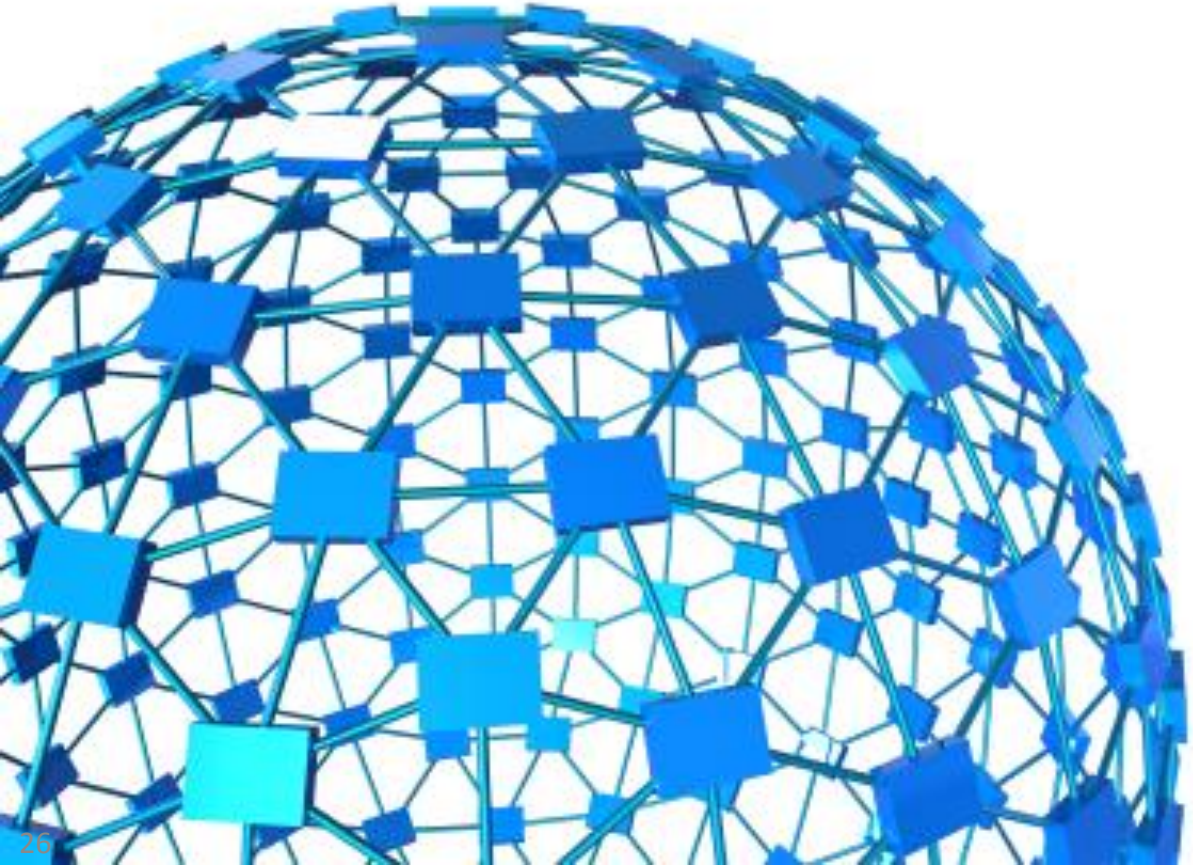


https://lod-cloud.net/

The Semantic Web
is a portal to the **web of knowledge**

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

scalable approach for the discovery
of *independently constructed,*
collaboratively described,
distributed knowledge
(in principle)



W3C Semantic Web

W3C Recommendation

RDF 1.1 Concepts and Abstract Syntax

W3C Recommendation 25 February 2014

This version:
http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/
Latest published version:
http://www.w3.org/TR/rdf11-concepts/
Previous version:
http://www.w3.org/TR/2014/PR-rdf11-concepts-20140109/
Previous Recommendation:
http://www.w3.org/TR/rdf-concepts
Editors:
Richard Cyganiak, DERI, NUI Galway
David Wood, 3 Round Stones
Markus Lanthaler, Graz University of Technology

W3C Recommendation

SPARQL 1.1 Overview

W3C Recommendation 21 March 2013

This version:
http://www.w3.org/TR/2013/REC-sparql11-overview-20130321/
Latest version:
http://www.w3.org/TR/sparql11-overview/
Previous version:
http://www.w3.org/TR/2012/PR-sparql11-overview-20121108/
Editor:
The W3C SPARQL Working Group, see Acknowledgements <public-rdf-dawg-comments@w3.org>

W3C Recommendation

OWL 2 Web Ontology Language Document Overview (Second Edition)

W3C Recommendation 11 December 2012

This version:
http://www.w3.org/TR/2012/REC-owl2-overview-20121211/
Latest version (series 2):
http://www.w3.org/TR/owl2-overview/
Latest Recommendation:
http://www.w3.org/TR/owl-overview
Previous version:
http://www.w3.org/TR/2012/PER-owl2-overview-20121018/
Editors:
W3C OWL Working Group (see Acknowledgements)

Data on the Web Best Practices

W3C Recommendation 31 January 2017

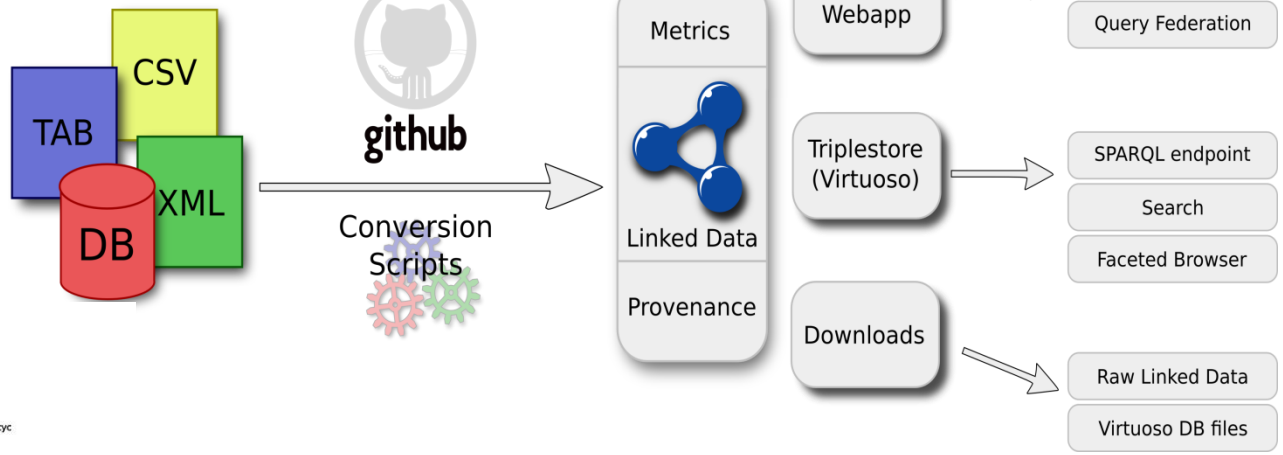
This version:
https://www.w3.org/TR/2017/REC-dwbp-20170131/
Latest published version:
https://www.w3.org/TR/dwbp/
Latest editor's draft:
http://w3c.github.io/dwbp/bp.html
Implementation report:
http://w3c.github.io/dwbp/dwbp-implementation-report.html
Previous version:
https://www.w3.org/TR/2016/PR-dwbp-20161215/
Editors:
Bernadette Farias Lóscio, CIn - UFPE, Brazil
Caroline Burle, NIC.br, Brazil
Newton Calegari, NIC.br, Brazil

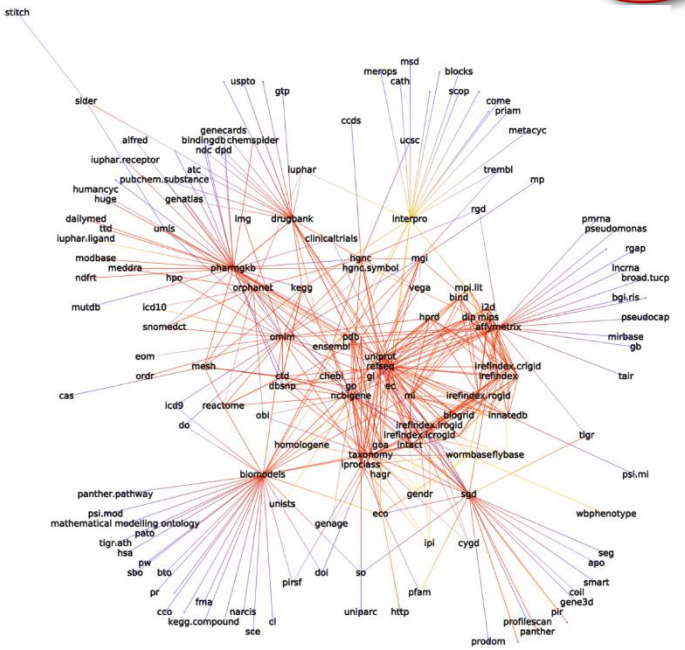
BIO2RDF

Linked Data for the Life Sciences

Bio2RDF is an open source project that uses semantic web technologies to make it easier to reuse biomedical data

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





- **30+** biomedical data sources
- **10B+** interlinked statements
- EBI, SIB, NCBI, DBCLS, NCBO, and many others produce this content

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

Federated query over the biological web of data

Phenotypes of knock-out mouse models for the targets of a selected drug (Imatinib)

Endpoint : http://drugbank.bio2rdf.org/sparql

Output : Table

Configure request

```
PREFIX dct: <http://purl.org/dc/terms/>
SELECT DISTINCT ?phenotype_label
WHERE {
  SERVICE <http://drugbank.bio2rdf.org/sparql> {
        ?drug <http://bio2rdf.org/drugbank_vocabulary:target> ?target .
         FILTER(?drug = <http://bio2rdf.org/drugbank:DB00619>)
        ?target <http://bio2rdf.org/drugbank_vocabulary:x-hgnc> ?hgnc .
  }
  SERVICE <http://hgnc.bio2rdf.org/sparql> {
    ?hgnc <http://bio2rdf.org/hgnc_vocabulary:x-mgi> ?marker .
  }
  SERVICE <http://mgi.bio2rdf.org/sparql> {
        ?model <http://bio2rdf.org/mgi_vocabulary:marker> ?marker .
        ?model <http://bio2rdf.org/mgi_vocabulary:allele> ?all .
        ?all <http://bio2rdf.org/mgi_vocabulary:allele-attribute> ?allele_type .
        ?model <http://bio2rdf.org/mgi_vocabulary:phenotype> ?phenotypes .
        FILTER (str(?allele_type) = "Null/knockout")
  }
  SERVICE <http://bioportal.bio2rdf.org/sparql> {
        ?phenotypes rdfs:label ?phenotype_label .
  }
}
```

	phenotype_label
1	"hemorrhage [mp:0001914]"@en
2	"intracranial hemorrhage [mp:0001915]"@en
3	"perinatal lethality [mp:0002081]"@en

Reproduce original research

Mol Syst Biol. 2011; 7: 496.
Published online 2011 Jun 7. doi: 10.1038/msb.2011.26
PMCID: PMC3159979

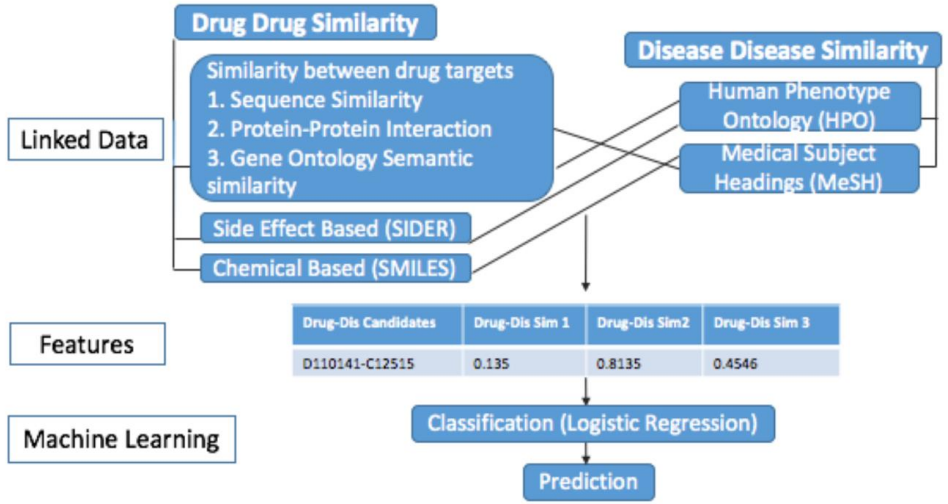
PREDICT: a method for inferring novel drug indications with application to personalized medicine

Assaf Gottlieb,[1] Gideon Y Stein,[2,3] Eytan Ruppin,[1,2] and Roded Sharan[a,1]

AUC 0.91 across all therapeutic indications

Scripts not available. Feature tables available.

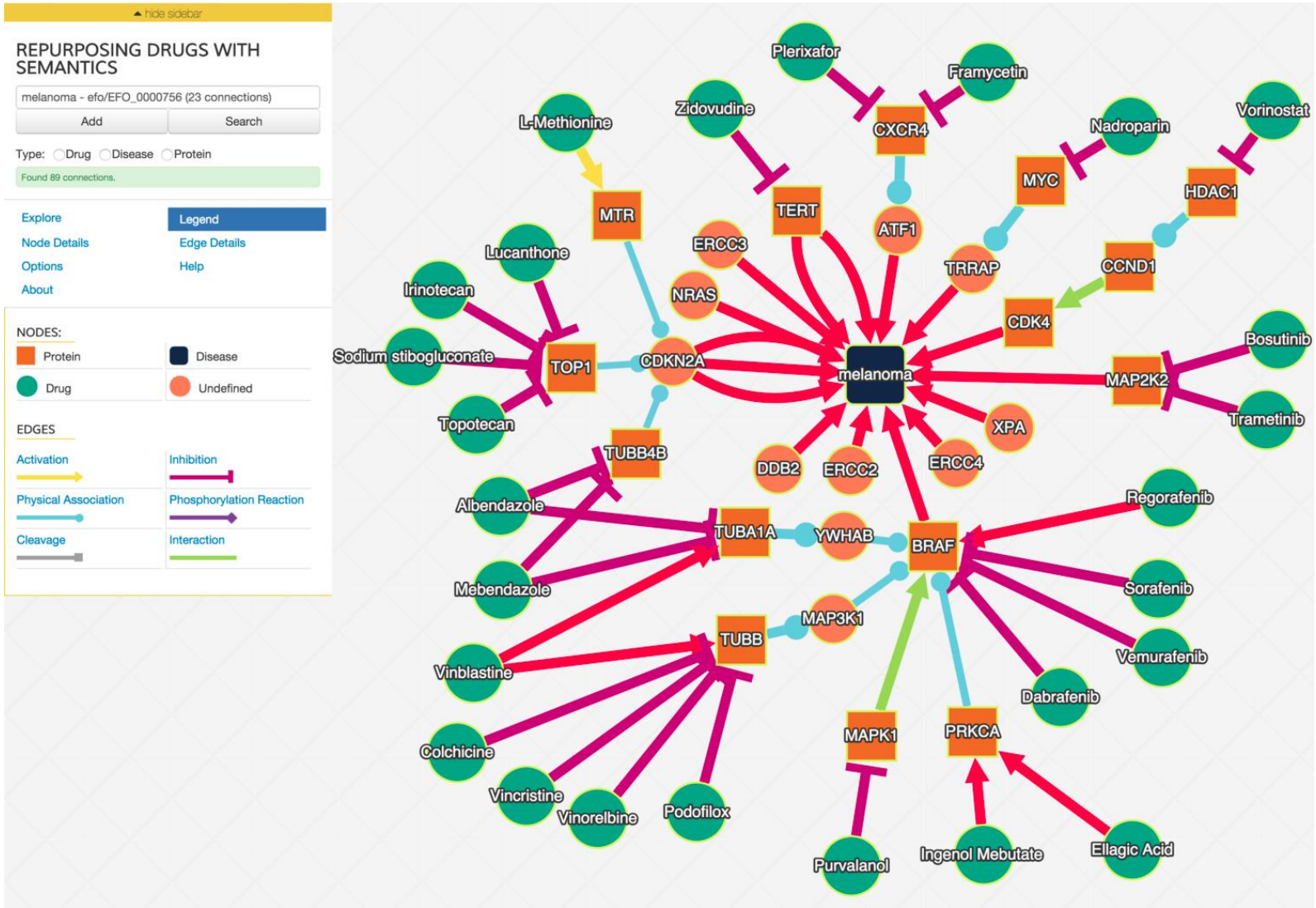
BIO2RDF



Result of reproducibility study: ROCAUC 0.83

Efficiently explore the web of data

by exploring a probabilistic semantic knowledge graph



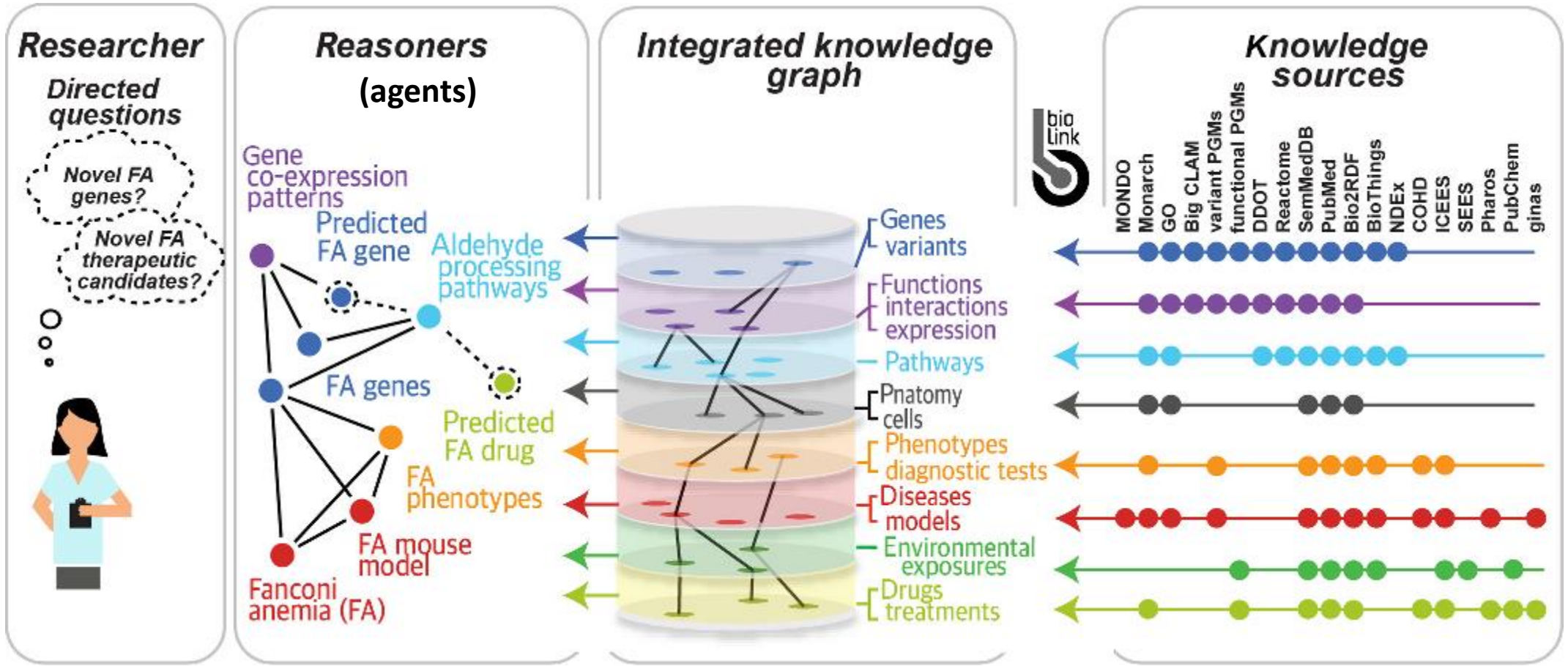
And validate them against pipelines for drug discovery

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 Computer Science. 2017. 3:e106 https://doi.org/10.7717/peerj-cs.106

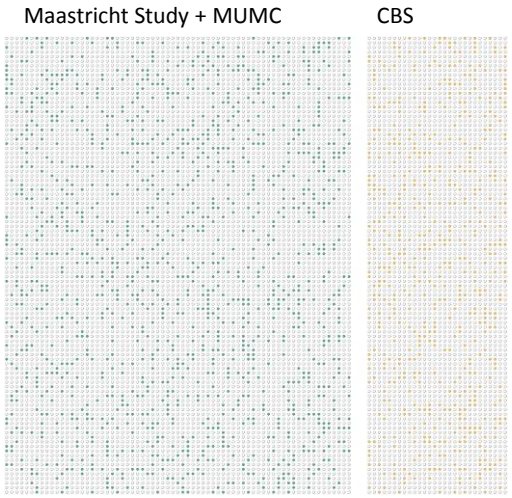
National Center for Advancing Translational Sciences

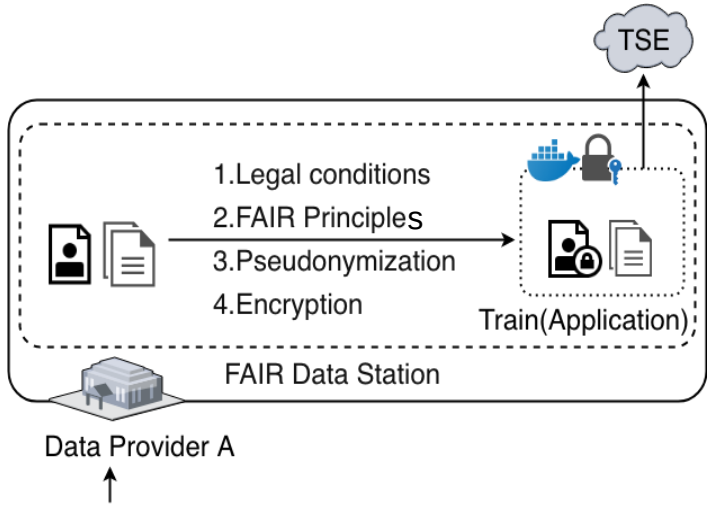
Biomedical Data Translator

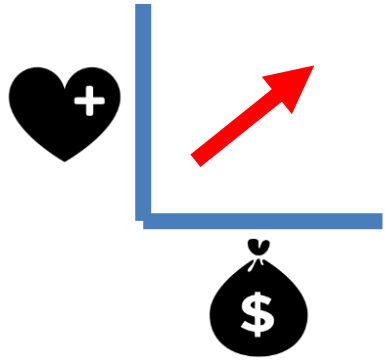


A community building a shared infrastructure…

Mine distributed, access restricted FAIR datasets in a privacy preserving manner



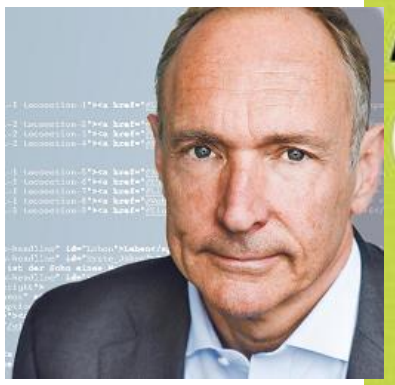




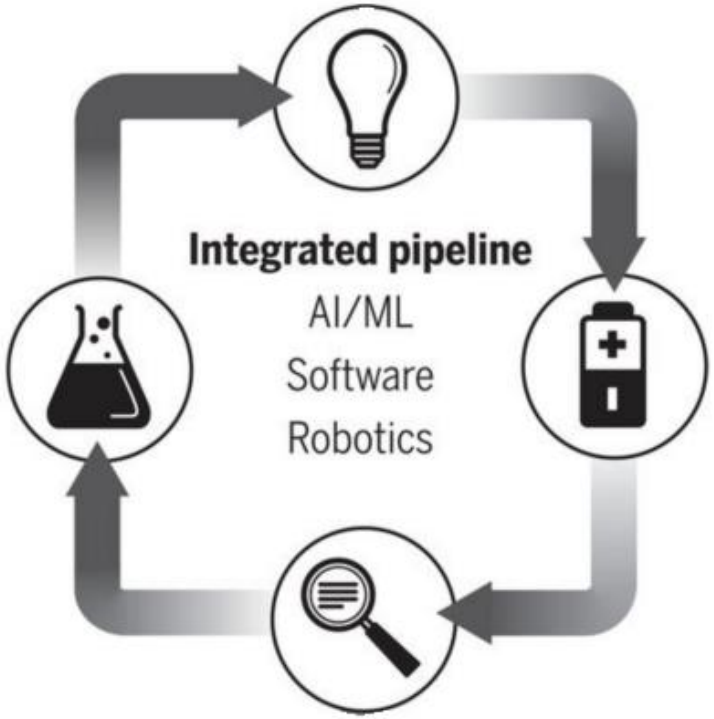
Goal is to learn **high confidence** determinants of health in a **privacy preserving** manner over ***vertically partitioned data*** from the Maastricht Study and Statistics Netherlands. The data are made available through **FAIR data stations** that provide access to *allowable* subsets of data to *authorized* users of *approved* algorithms.

Establish **a new social, legal, ethical and technological infrastructure** for discovery science in and across health and non-health settings, including scalable **governance** and flexible **consent** to underpin the responsible use of Big Data.

FAIR is a part of the solution that will enable arbitrary machines to work with each other



Tim Berners-Lee



Ross King

Semantic Web ⟷ Robot Science

Large Scale, Autonomous Scientific Discovery

Summary

FAIR represents a global initiative to enhance the discovery and reuse of all kinds of digital resources. ***It is a work in progress!***

It demands a **new social, legal, ethical, scientific and technological infrastructure** that currently doesn't exist *in whole*, but has to be built for and adopted by digital savvy communities! It must answer the questions:

- Can we build and use shared terminologies and representations to reduce the effort needed to answer questions across data collections?
- How can we share data and perform analyses in a responsible manner?
- What incentives, rewards and penalties are needed to maximize trust, participation, legality, and utility?

Semantics, coupled with AI technologies, may enable **humans, aided by intelligent machine agents, to exploit the Internet of FAIR data and services**, and hence to accelerate discovery in biomedicine and in other disciplines.

Acknowledgements

Dumontier Lab (Maastricht University, Stanford University, Carleton University)

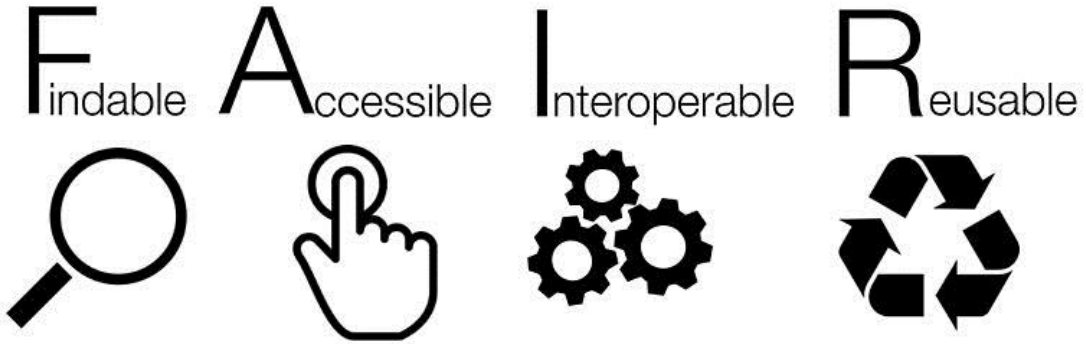
MU: Seun Adekunle, Remzi Celebi, Dorina Claessens, Ricardo De Miranda Azevedo, Pedro Hernandez Serrano, Massimiliano Grassi, Andine Havelange, Lianne Ippel, Alexander Malic, Kody Moodley, Stuti Nayak, Nadine Rouleaux, Claudia van open, Chang Sun, Amrapali Zaveri
SU: Sandeep Ayyar, Remzi Celebi, Shima Dastgheib, Maulik Kamdar, David Odgers, Maryam Panahiazar, Amrapali Zaveri
CU: Alison Callahan, Jose Toledo-Cruz, Natalia Villaneuva-Rosales

FAIR

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons

FAIR metrics

Mark D Wilkinson, Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos, Michel Dumontier



The mission of the **Institute of Data Science at Maastricht University** is to foster a collaborative environment for multi-disciplinary data science research, interdisciplinary training, and data-driven innovation .

We tackle key **scientific, technical, social, legal, ethical issues** that advance our understanding across a variety of disciplines and strengthen our communities in the face of these developments.

michel.dumontier@maastrichtuniversity.nl

Website: http://maastrichtuniversity.nl/ids

Maastricht University