The Internet for Social Machines

The end of data sharing as we know it

FAIR > The main issue(s)

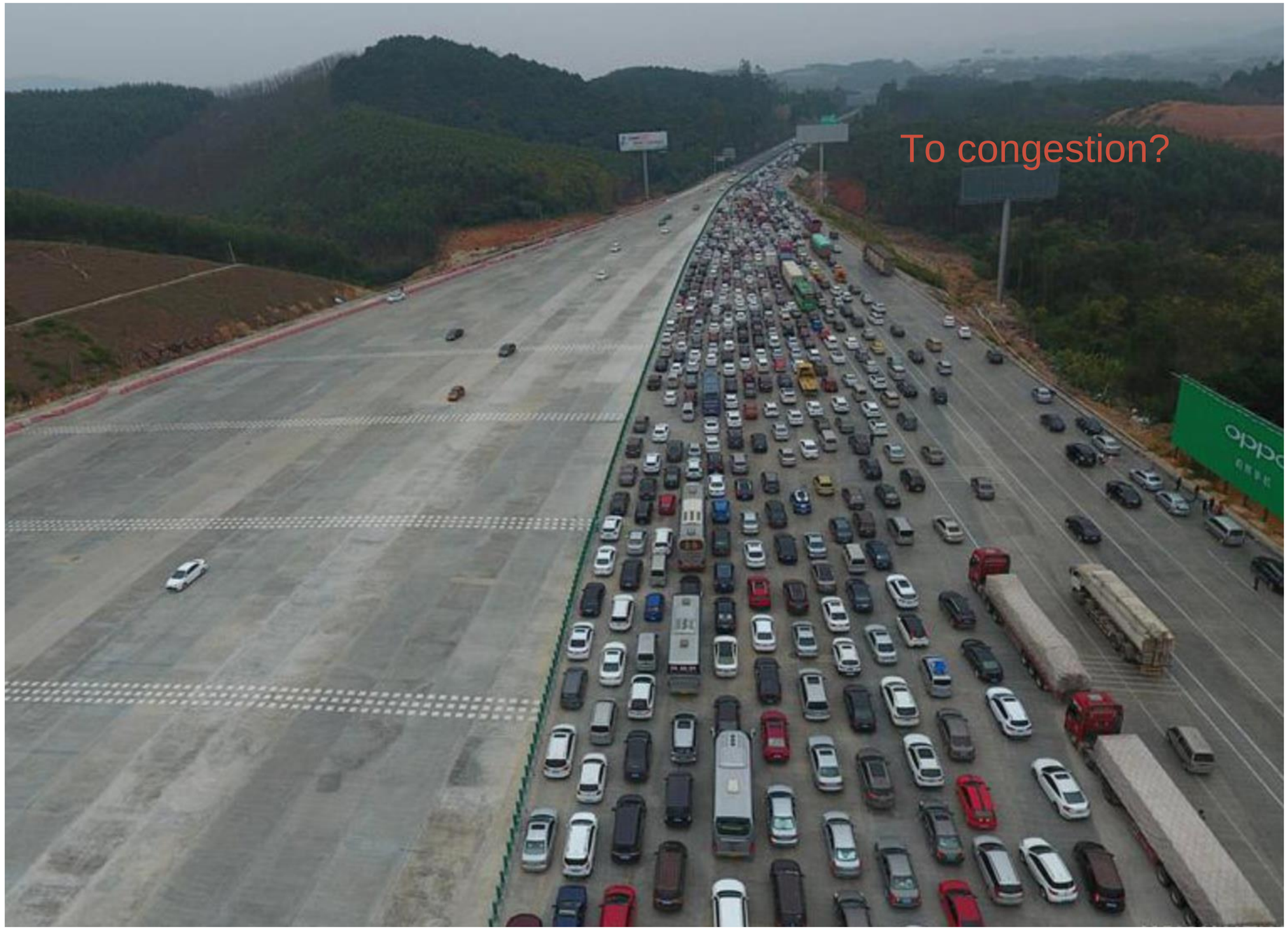
Barend Mons

April 24, 2019

‘The Machine knows what I mean’

The Road to FAIRness

The Road to FAIRness



CODATA

independent

supra-disciplinary

no competition for research funding

Global Scope

mandated to be impartial

not funding agencies

diplomacy

policy

large initiatives

endorsement

MoU

MoU

Convene expert groups

Community building

define and develop common practices

represent communities

coordinate communities

advocacy

plenaries

working groups

Interest groups

Grassroots

facilitates cross-IN issues

makes it easy to comply

PPP

implementation Networks

MoU

RDA RESEARCH DATA ALLIANCE

GO FAIR

serving the international data community

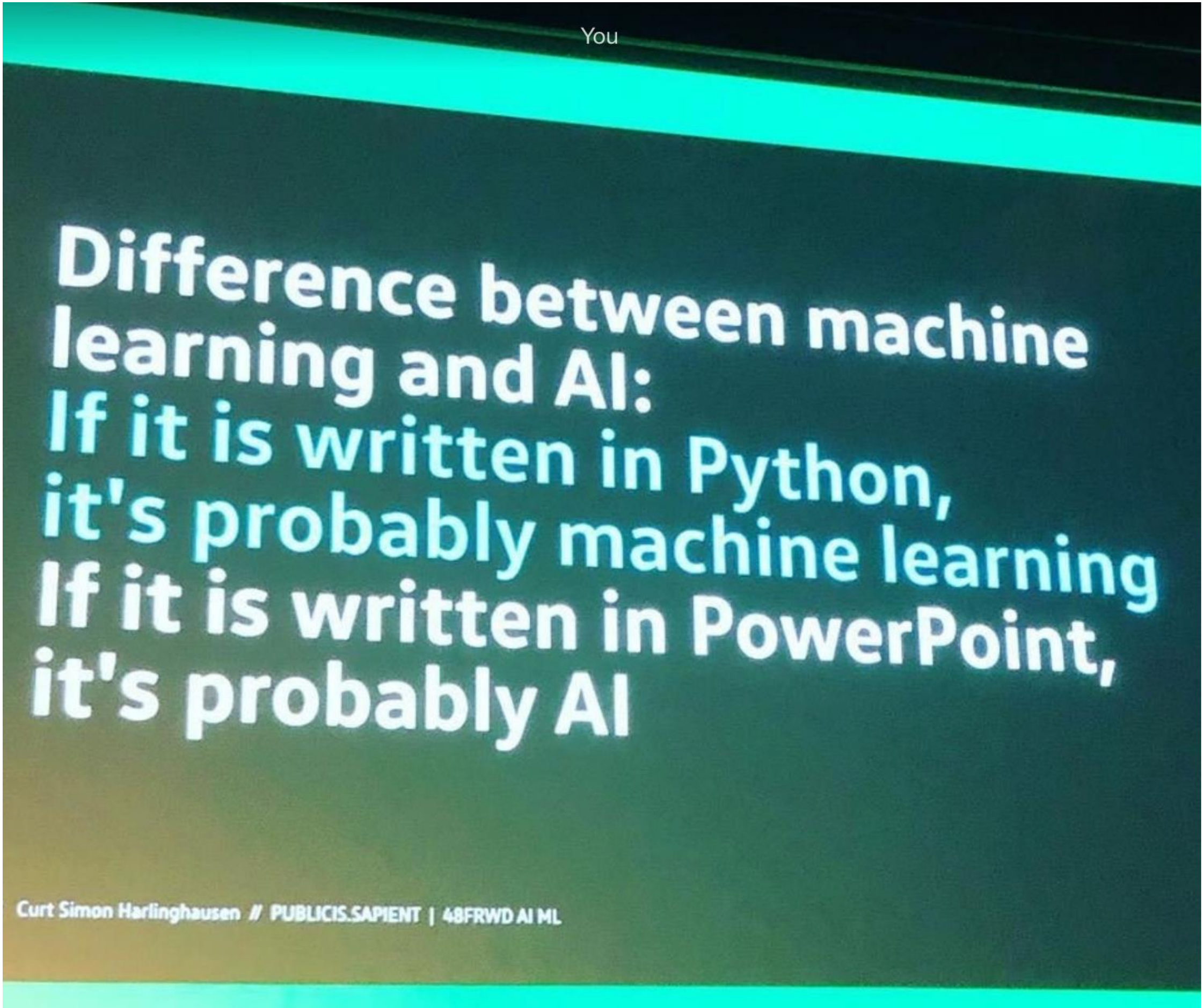
The Internet….

The end of data sharing as we know it

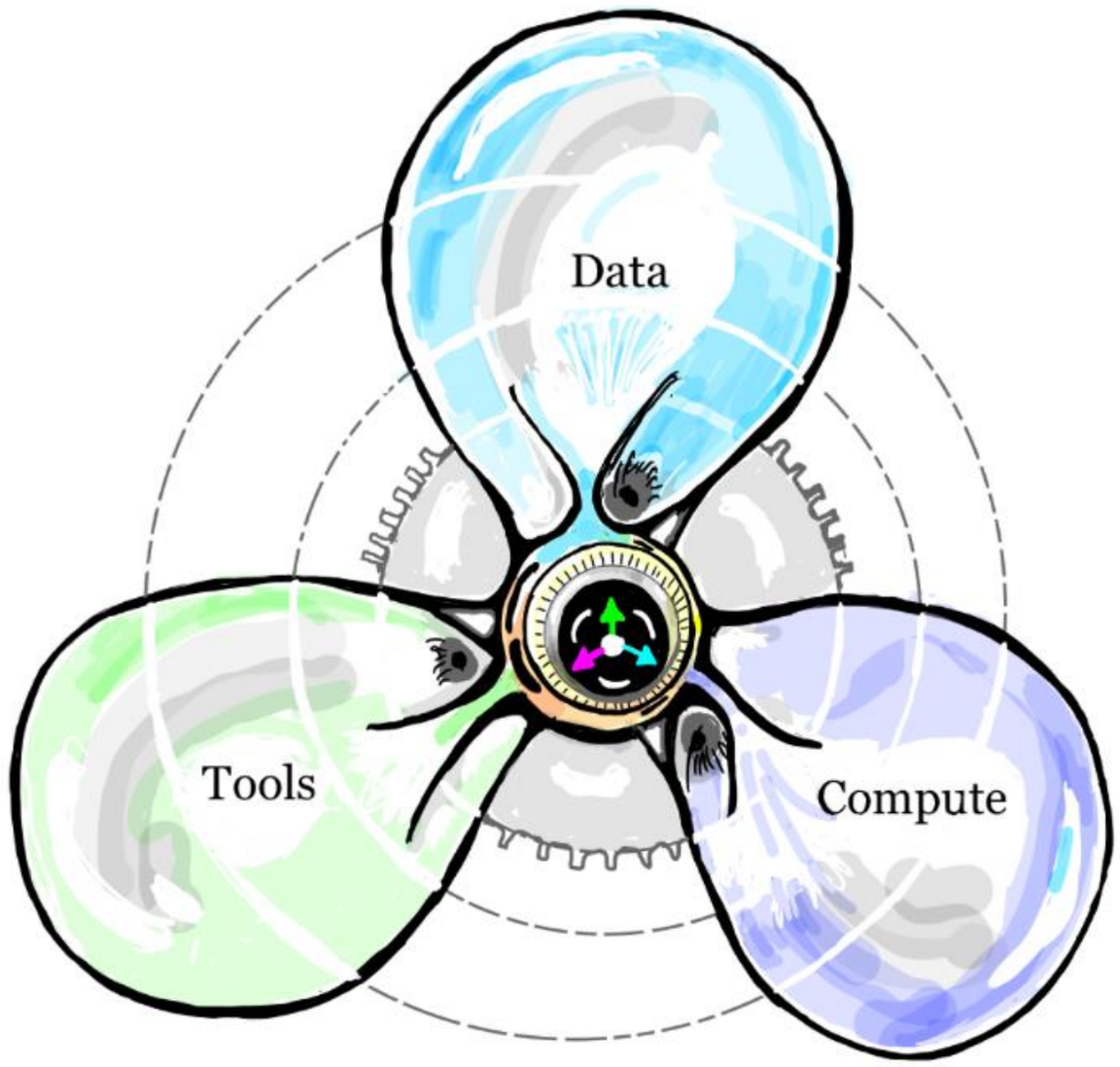
from data sharing to data visiting in

The Internet for Social Machines

You


Difference between machine learning and AI:
If it is written in Python, it's probably machine learning
If it is written in PowerPoint, it's probably AI
Curt Simon Harlinghausen // PUBLICIS.SAPIENT | 48FRWD AI ML

The Internet of FAIR data and Services

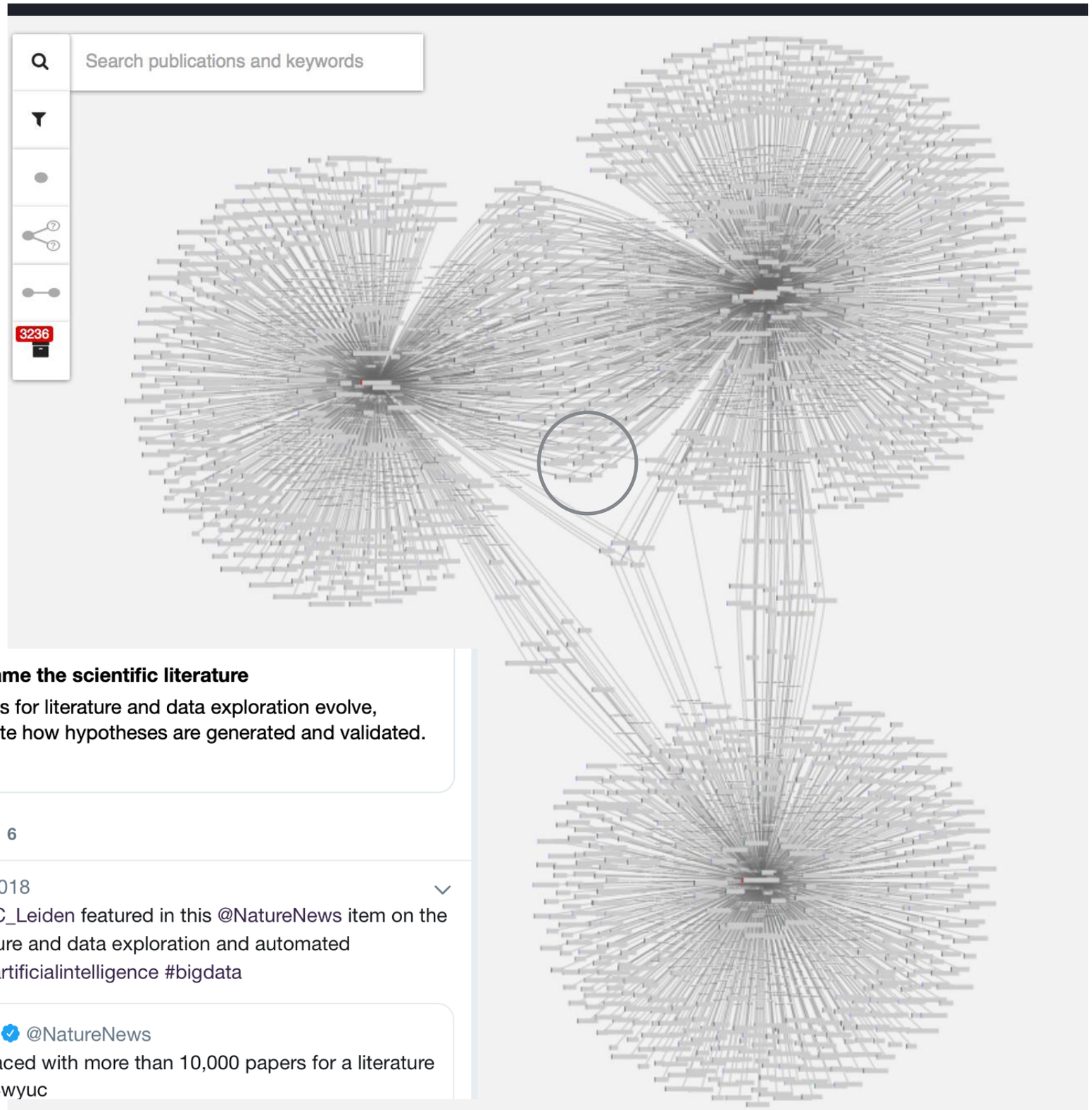


The Internet for Social Machines

The end of data sharing as we know it

complexity is beyond human comprehension, not only in life sciences!

5 objects are shared between all three knowlets
(in this case: metabolic syndrome, diabetes, and e.o Alzheimer)



How AI technology can tame the scientific literature

As artificially intelligent tools for literature and data exploration evolve, developers seek to automate how hypotheses are generated and validated.

nature.com

1 6

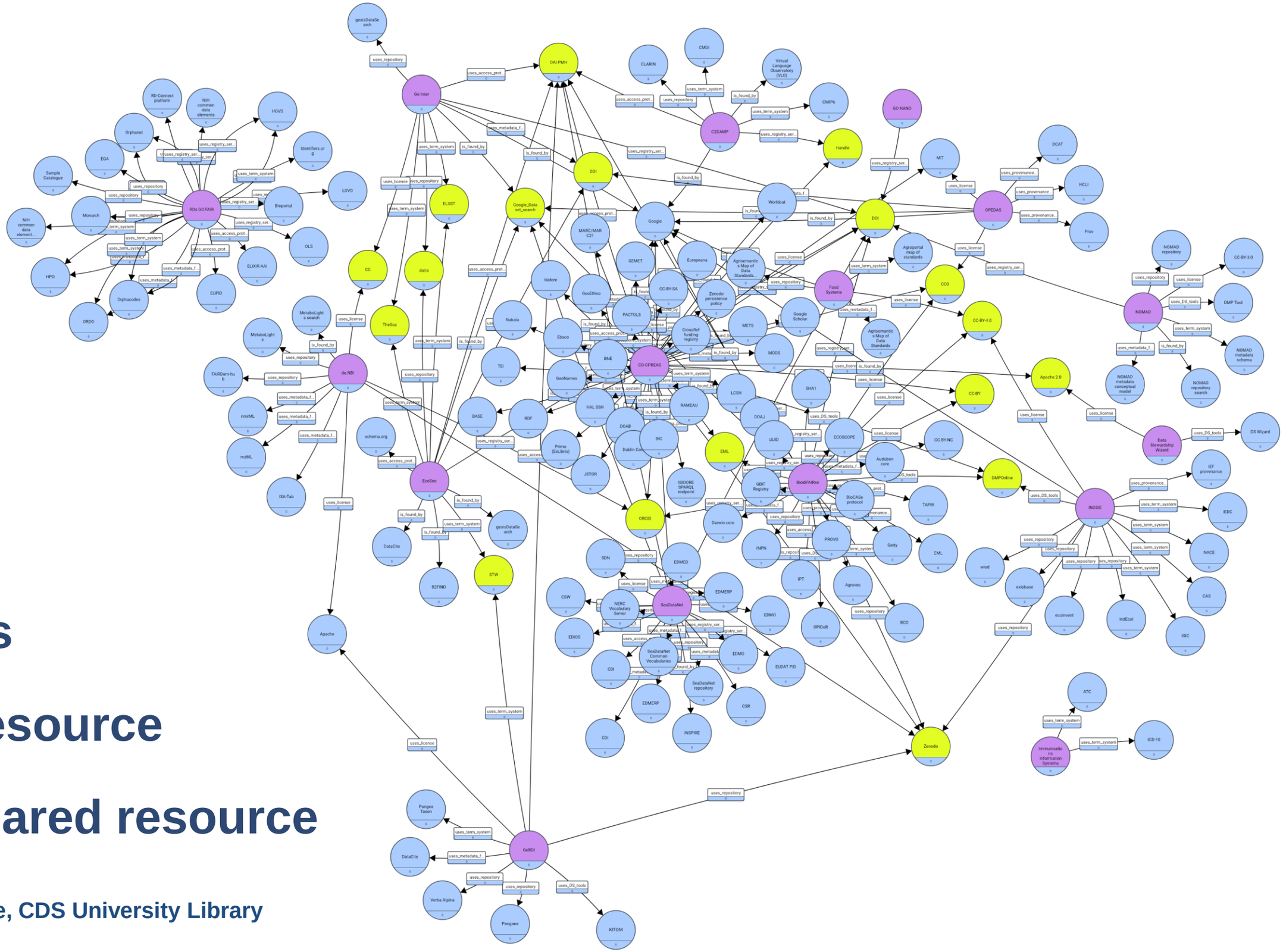
Euretos @Euretos · 10 Sep 2018

@euretos and partner @LUMC_Leiden featured in this @NatureNews item on the use of AI technology in literature and data exploration and automated hypotheses generation #AI #artificialintelligence #bigdata

Nature News & Comment @NatureNews

What would you do when faced with more than 10,000 papers for a literature review? go.nature.com/2N4wyuc

Community Implementation Choices & Challenges



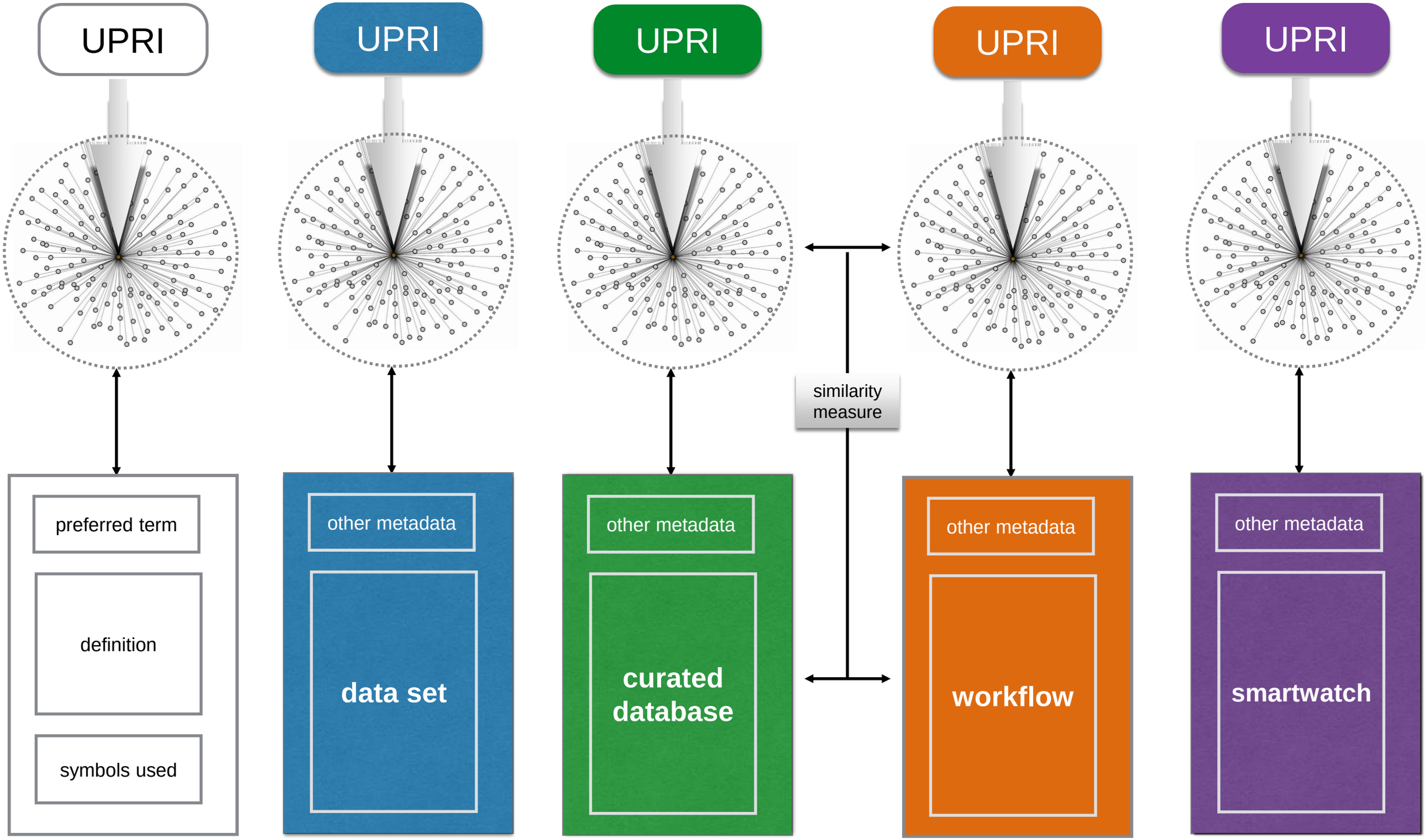
INs

Resource

Shared resource

Kristina Hettne, CDS University Library

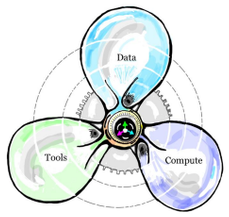
This will also bring (the right) people together!



This community does not make the classical mistakes

first, some term-bashing

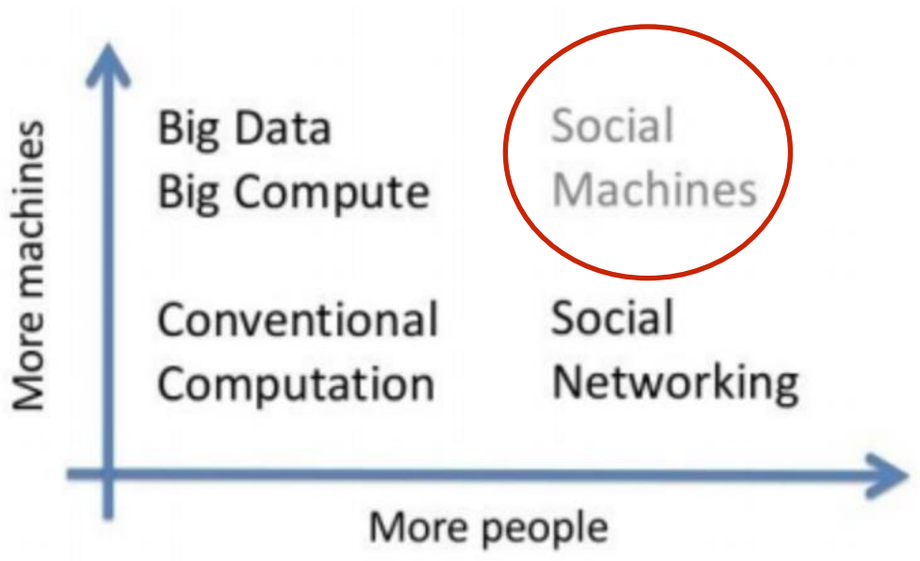
- un-FAIR <> Re-useless
- Standard <> **Guiding principle**
- Open <> **Accessible under well defined conditions**
- AI <> **Machine learning**
- Management <> **Stewardship**
- Sharing <> **Visiting**



Distributed learning by VM's

Which Gene Did You Mean?
why bury it first and then mine it again?

(2005)



(2018)

What does FAIR eventually entail?
The Machine knows what I mean

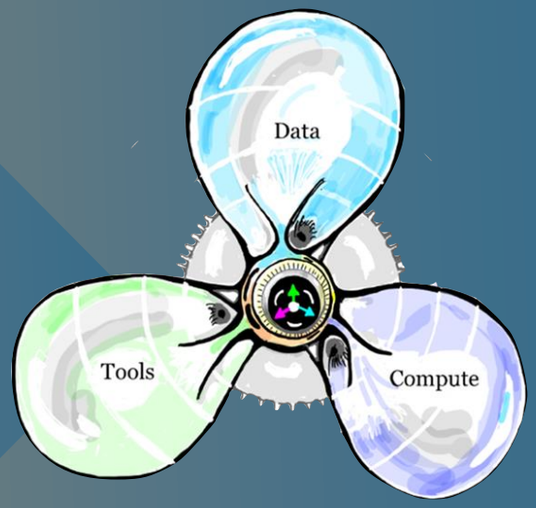
FAIR and GO FAIR

Lorentz

EOSC

FAIRdICT

IFDS



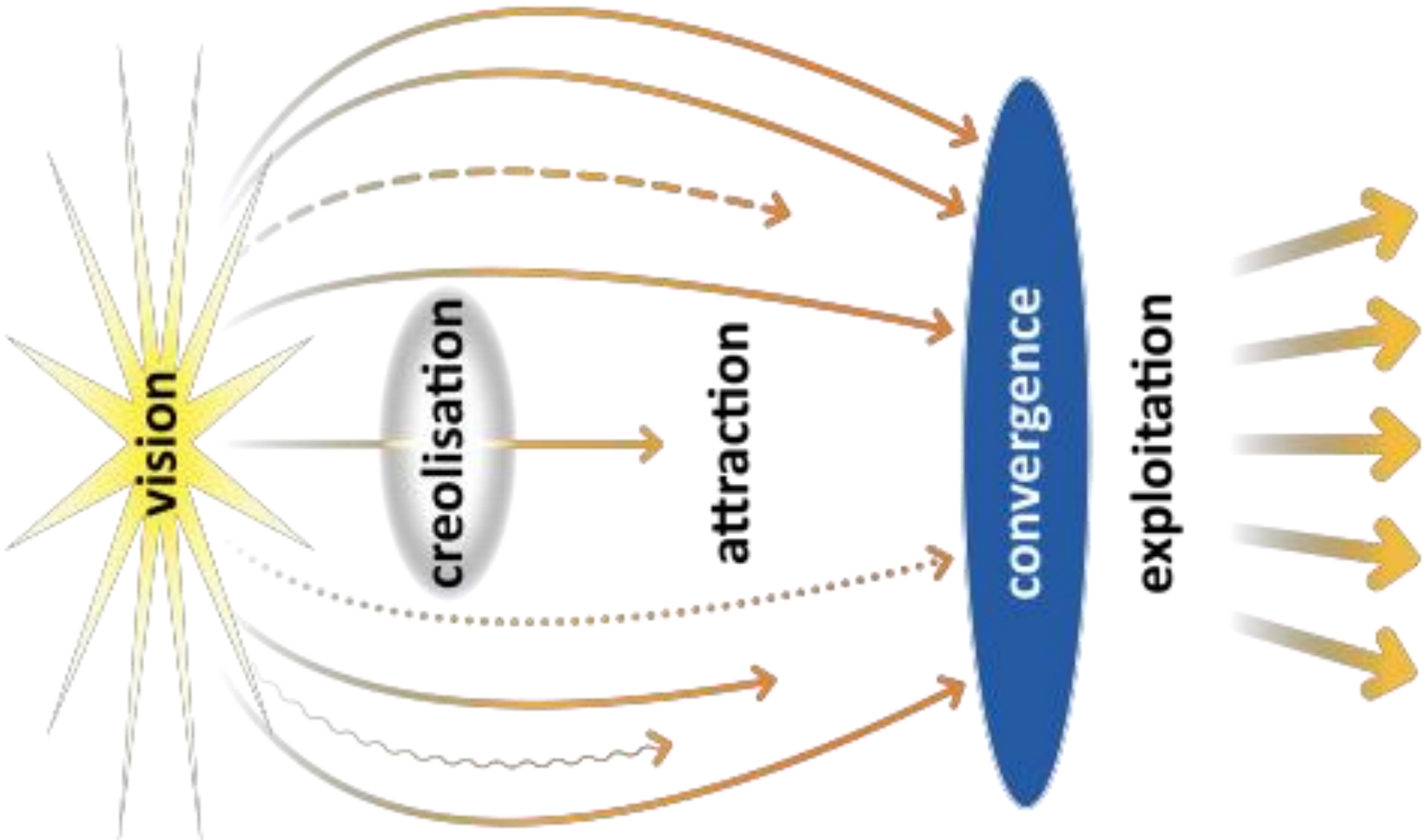
Birth	Infancy		Adolescence		Maturity
2014	2015	2016	2017	2018…	

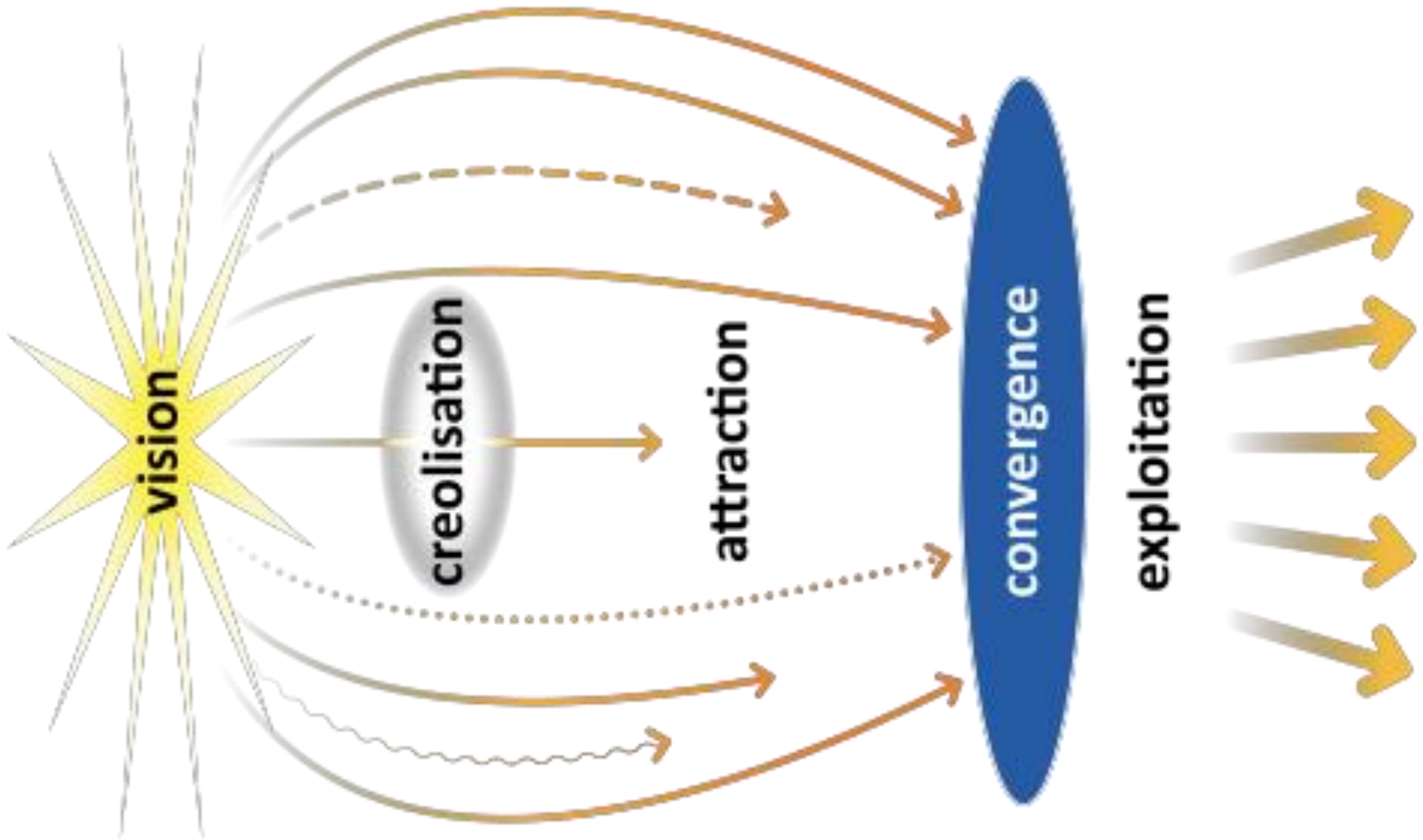
How is the Internet for Social Machines likely to develop?

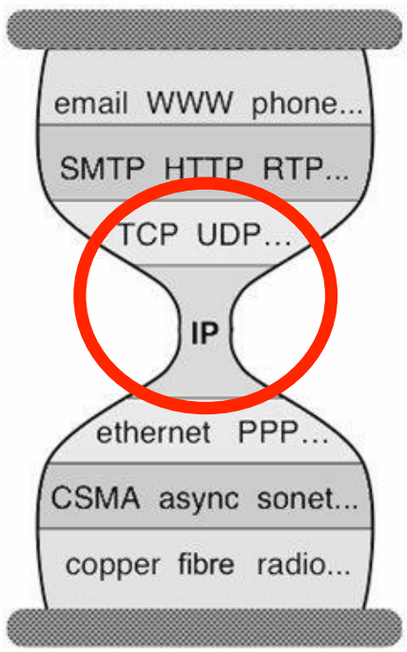
From looming congestion to exploitation !

Common Patterns in Revolutionary Infrastructures and Data

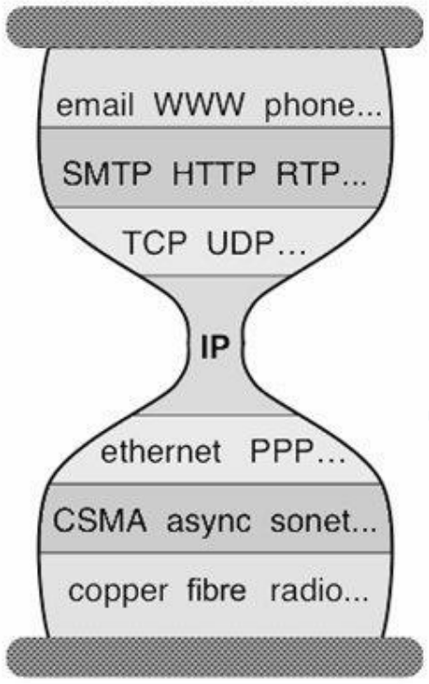
Peter Wittenburg, Max Planck Computing and Data Facility, George Strawn, US National Academy of Sciences, February 2018
https://www.rd-alliance.org/sites/default/files/Common_Patterns_in_Revolutionising_Infrastructures-final.pdf







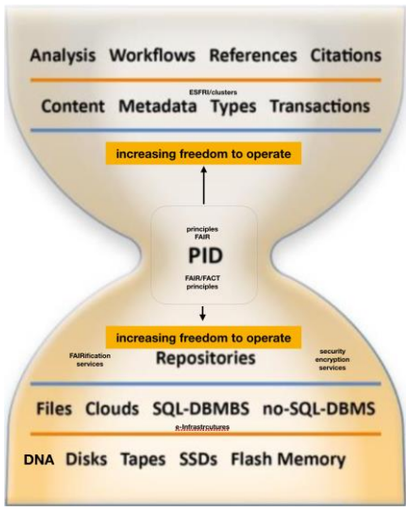
- **Minimal standards**
- **Voluntary participation**
- **Critical mass**



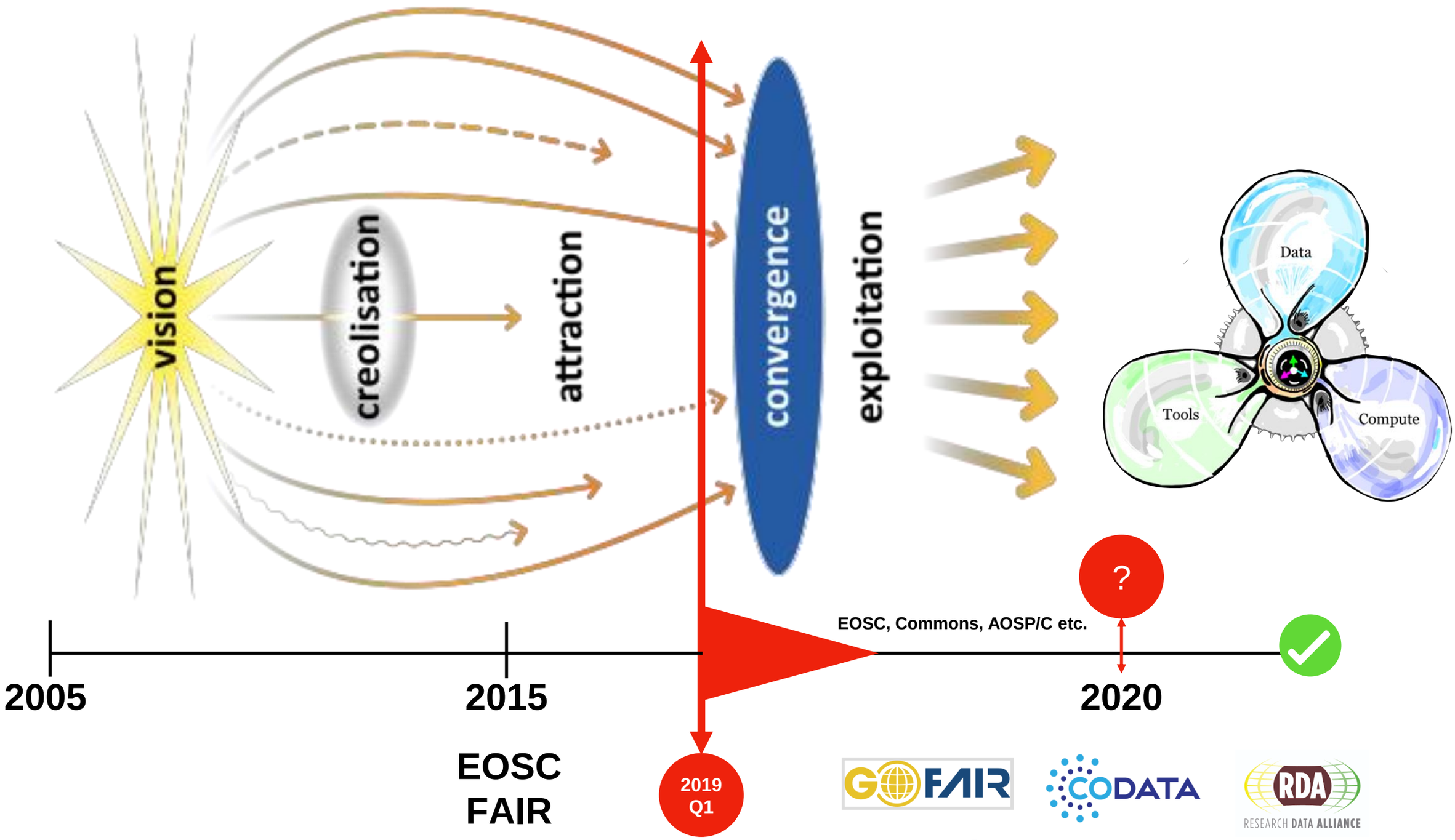
Lessons from the Internet for People:

1. Minimal standards only
2. Rough consensus/Running code
3. Don't tell anyone else what to do
4. Critical mass of lead-players

Now, for the Internet for Machines



Its happening RIGHT NOW!



- Minimal standards
- Voluntary participation
- Critical mass
- Rough consensus and running code

From attraction to convergence !!

IN Profile Matrix

Erik Schultes, PhD
International Science Coordinator
GO FAIR International Support and Coordination Office
erik.schultes@go-fair.org
go-fair.org

February 27, 2019

GO FAIR IN Profile Matrix

January 15-16, Leiden

Survey https://docs.google.com/forms/d/1Oug6GowuG1jNZNsjklXOeEvPbUrhyuS_F-d185SOy6A/edit
Matrix https://docs.google.com/spreadsheets/d/1MUZn7uh4x5YLPjqxi-V8XubsSEEonQWvx2jBlcyyNdU/edit#gid=0

IN Profile Matrix

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

100% $ % .0 .00 123 Helvetica ... 10

FAIR Implementation Matrix

On the OSF https://osf.io/n7uwp/

Red indicates waist of hourglass

*Blue is an Implementation **Choice***

*Orange is Implementation **Challenge***

Green highlight indicates a service provided by the IN or spin-off

Blank cell is not relevant for IN

FAIR Principle	Services	Component	Most used	C2CAMP	OPEDAS	PHT	Rare-Diseases	GERI
	central to all	DOIP	DOIP	DOIP	DOIP	DOIP	DOIP	
	central to all	Metadata format	RDF		RDF	RDF	RDF	
	central to all	Metadata access protocol			LDP/FDP	LDP/FDP	LDP/FDP	
	central to all	Metadata core elements	TBD on M4M		TBD on M4M	TBD on M4M	TBD on M4M	
	Technology	Data Format			RDF for interop.	RDF for interop.	RDF for interop.	
	Technology	Data Access Protocols (MR/A)			LDP/FDP	PHT-standard	PHT-standard	
	Technology	Computer-actionable license description language			RDF	RDF	RDF	
	Tooling	Repository (Data/Metadata)		DONA	IFDS Data Station	IFDS Data Station	ERN?	GERI
	Tooling(Repository)	https://www.dataone.org						
	Tooling	Registry Service		DONA	IFDS Station Registry	IFDS Station Registry	ERN?	
	tooling	Metadata forms/creators			CEDAR/CASTOR			
	Tooling	Search capability		DOIP	IFDS Station Registry	IFDS Station Registry	IFDS Station Registry	
	Policy	Persistence Policy			TBD	TBD	TBD	
	Technology	Computer-actionable policy description language			RDF	RDF	RDF	
	Tooling	License protocols			TBD	TBD	TBD	
	Tooling	Training Materials			Training-IN	Training-IN	EJP	

GO FAIR IN Profile Matrix

January 15-16, Leiden

Survey https://docs.google.com/forms/d/1Oug6GowuG1jNZNsjklXOeEvPbUrhyuS_F-d185SOy6A/edit
Matrix https://docs.google.com/spreadsheets/d/1MUZn7uh4x5YLPjqxi-V8XubsSEEonQWvx2jBlcyyNdU/edit#gid=0

IN Profile Matrix

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

100% $ % .0 .00 123 Helvetica ... 10

Column 1

	A	B	C	D	E	F	G	H
1	FAIR Implementation Matrix							
2	*On the OSF*	https://osf.io/n7uwp/						
3	*Red indicates waist of hourglass*							
4	*Blue is an Implementation* ***Choice***							
5	*Orange is Implementation* ***Challenge***							
6	*Green highlight indicates a service provided by the IN or spin-off*							
7	*Blank cell is not relevant for IN*							
8	**FAIR Principle**	**Services**	**Component**	**Most used**	**C2CAMP**	**OPEDAS**	**PHT**	**Rare-Diseases**
9		**central to all**	DOIP	DOIP	DOIP	DOIP	DOIP	DOIP
10		**central to all**	Metadata format	RDF		RDF	RDF	RDF
11		**central to all**	Metadata access protocol			LDP/FDP	LDP/FDP	LDP/FDP
12		**central to all**	Metadata core elements	TBD on M4M		TBD on M4M	TBD on M4M	TBD on M4M
13		**Technology**	Data Format			RDF for interop.	RDF for interop.	RDF for interop.
14		**Technology**	Data Access Protocols (MR/A)			LDP/FDP	PHT-standard	PHT-standard
15		**Technology**	Computer-actionable license description language			RDF	RDF	RDF
16		**Tooling**	Repository (Data/Metadata)		DONA	IFDS Data Station	IFDS Data Station	ERN?
17		**Tooling(Repository)**	https://www.dataone.org					
18		**Tooling**	Registry Service		DONA	IFDS Station Registry	IFDS Station Registry	ERN?
19		**tooling**	Metadata forms/creators			CEDAR/CASTOR		
20		**Tooling**	Search capability		DOIP	IFDS Station Registry	IFDS Station Registry	IFDS Station Registry
21		**Policy**	Persistence Policy			TBD	TBD	TBD
22		**Technology**	Computer-actionable policy description language			RDF	RDF	RDF
23		**Tooling**	License protocols			TBD	TBD	TBD
24		**Tooling**	Training Materials			Training-IN	Training-IN	EJP

Community Implementation Choices & Challenges

FAIR Guiding Principles

- *EOSC*
- *NIH Data Commons*
- *Preclinical Trials*
- *Funders*
- *American Geophysical Union*
- *Bayer*
- *Journalists*
- *Financial industry*

Self-Identified
Community
aiming to become more FAIR

Implementation
Choice

Community chooses to re-use existing metadata elements as needed to implement FAIR

Inspect IN Profile Matrix

Implementation
Challenge

Community accepts challenge to create new metadata elements as needed to implement FAIR

GO FAIR

Maximizing Reuse
- **Tools & registries**
- **Sustainability plans**
- **Funded infrastructure**
- **PPP services providers**

Community Implementation Choices & Challenges

FAIR Guiding Principles

Self-Identified
Community
aiming to become more FAIR

- *EOSC*
- *NIH Data Commons*
- *Preclinical Trials*
- *Funders*
- *American Geophysical Union*
- *Bayer*
- *Journalists*
- *Financial industry*

Implementation
Choice

Community chooses to re-use existing metadata elements as needed to implement FAIR

Inspect IN Profile Matrix

Implementation
Challenge

Community accepts challenge to create new metadata elements as needed to implement FAIR

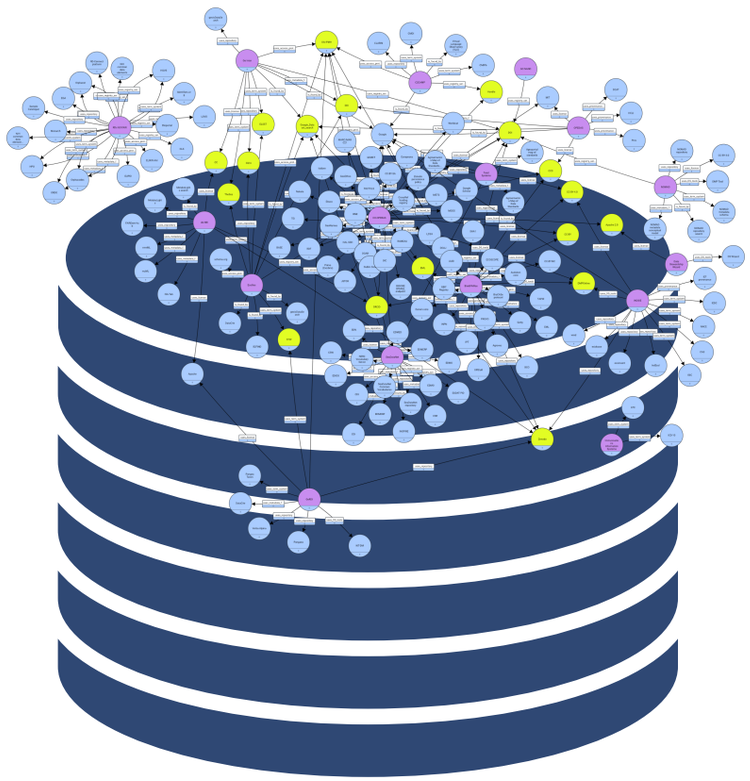
GO FAIR

Maximizing Reuse
- **Tools & registries**
- **Sustainability plans**
- **Funded infrastructure**
- **PPP services providers**

Building to Column 1

SUBJECT	PREDICATE	OBJECT	FAIR Principles
name of IN (UPRI)	has-coordinator	ORCID	
name of IN (UPRI)	has-participant	ORCID	
name of IN (UPRI)	has-member-organisation	VIVO / CrossRef	
name of IN (UPRI)	uses-repository	CTS?	
name of IN (UPRI)	uses-registry-service	PW ?	F1
name of IN (UPRI)	provides-registry-service		F1
name of IN (UPRI)	uses-data-format	format-PID	F2
name of IN (UPRI)	provides-data-format	format-PID	F2
name of IN (UPRI)	provides-access-protocol	format-PID	A1
name of IN (UPRI)	uses-access-protocol	protocol-PID	A1
name of IN (UPRI)	has-persistence-policy	policy	F1 / A2
name of IN (UPRI)	is found by	Search engine	F4
name of IN (UPRI)	uses-term-system	Term System-PID	I
name of IN (UPRI)	provides-term-system	Term System-PID	I
name of IN (UPRI)	uses-license	MR-license ID	R1.1
name of IN (UPRI)	uses-metadata-format	format-PID	R1.2
name of IN (UPRI)	provides-meta-data-format	Format-PID	R1.2
name of IN (UPRI)	provides-training-material	Resource-ID	
name of IN (UPRI)	uses-uses-training-material	Resource-ID	
name of IN (UPRI)	provides-DS-tools	Resource-ID	
name of IN (UPRI)	uses-DS-tools	Resource-ID	
name of IN (UPRI)	uses-workspace-tool	Resource-ID	
name of IN (UPRI)	Provides-workspace-tool	Resource-ID	

Community Implementation Choices & Challenges



Open, FAIR Data Point
Hosted by trusted party
(e.g. CDS, NIST…)

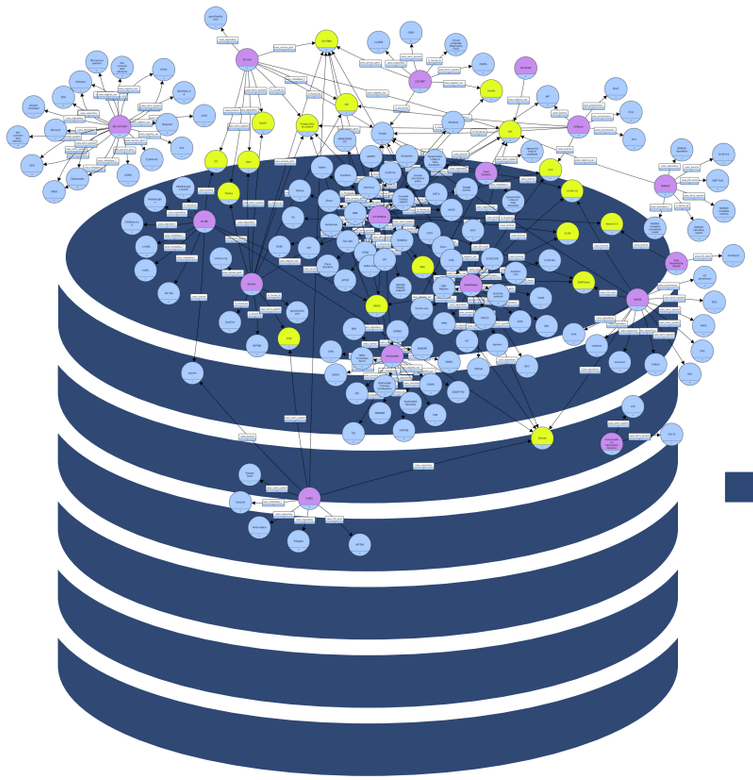
Community Implementation Choices & Challenges

Additional Community Profiles

AGU Enabling FAIR Data

MPS Hackathon

ENVRIFAIR



Open, FAIR Data Point

Column 1

Analytics
- **Reuse**
- **Overlaps**

Bottom-up Good Practices

FAIRsharing.org
standards, databases, policies

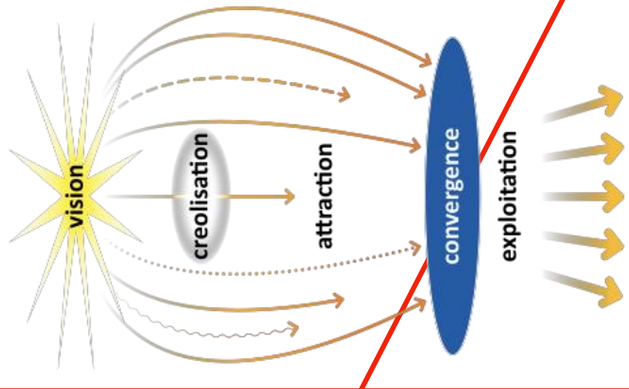
CODATA

RDA
RESEARCH DATA ALLIANCE

NSF

rfg

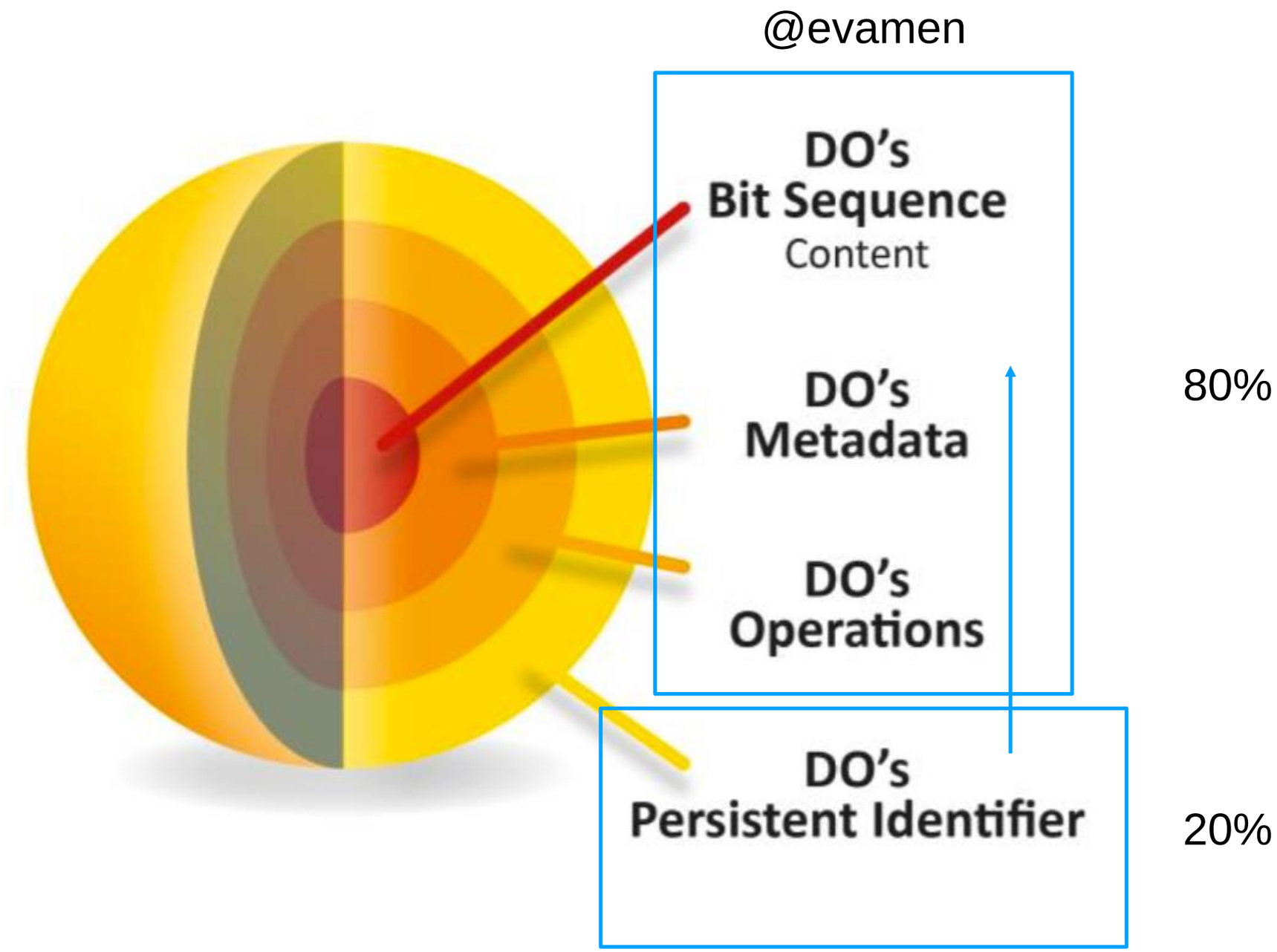
WILEY



Convergence


GO
FAIR

Minimally: treat everything as a digital, transferable FAIR object



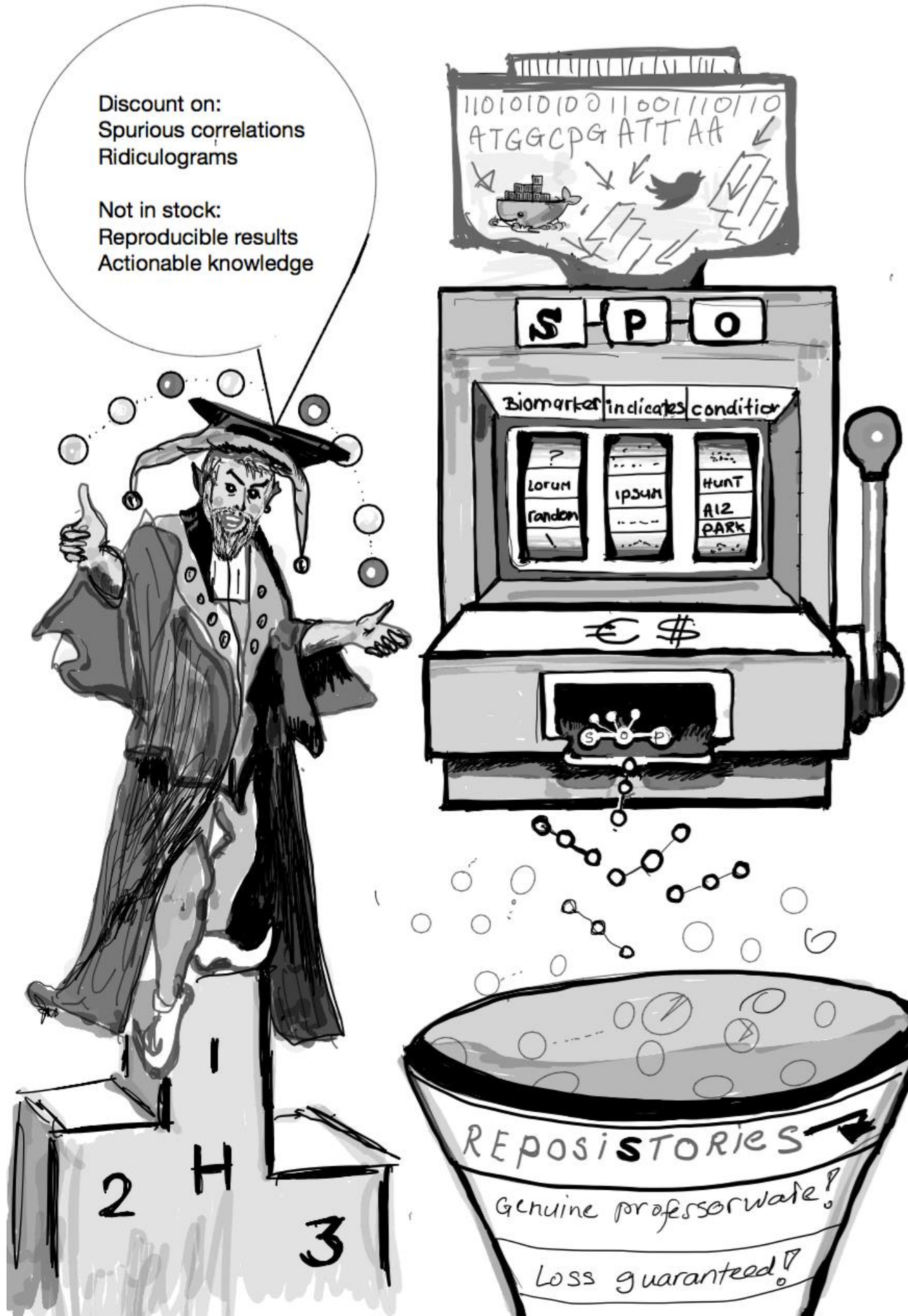
The seven capital sins of Open Science

1: Age factor….Reward only narrative.com

2: Ignore complexity and existing data

3: Disrespect other disciplines

4: publish data without a supplementary paper

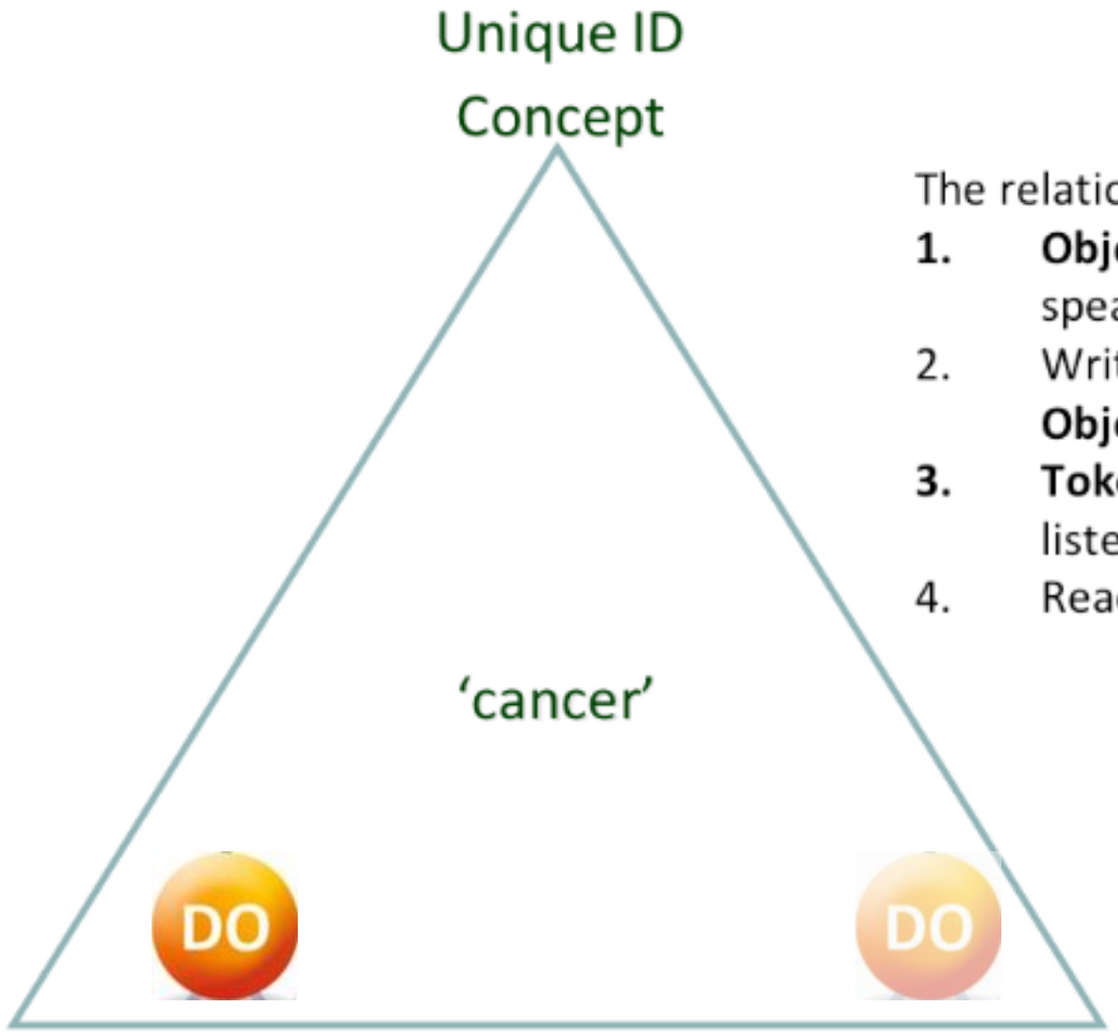


5: create a nightmare for machines

6: refuse to invest in research -infrastructure

7 : Create Data without a Data Stewardship plan

The Ogden Triangle – Concepts versus words



The relations between the corners:

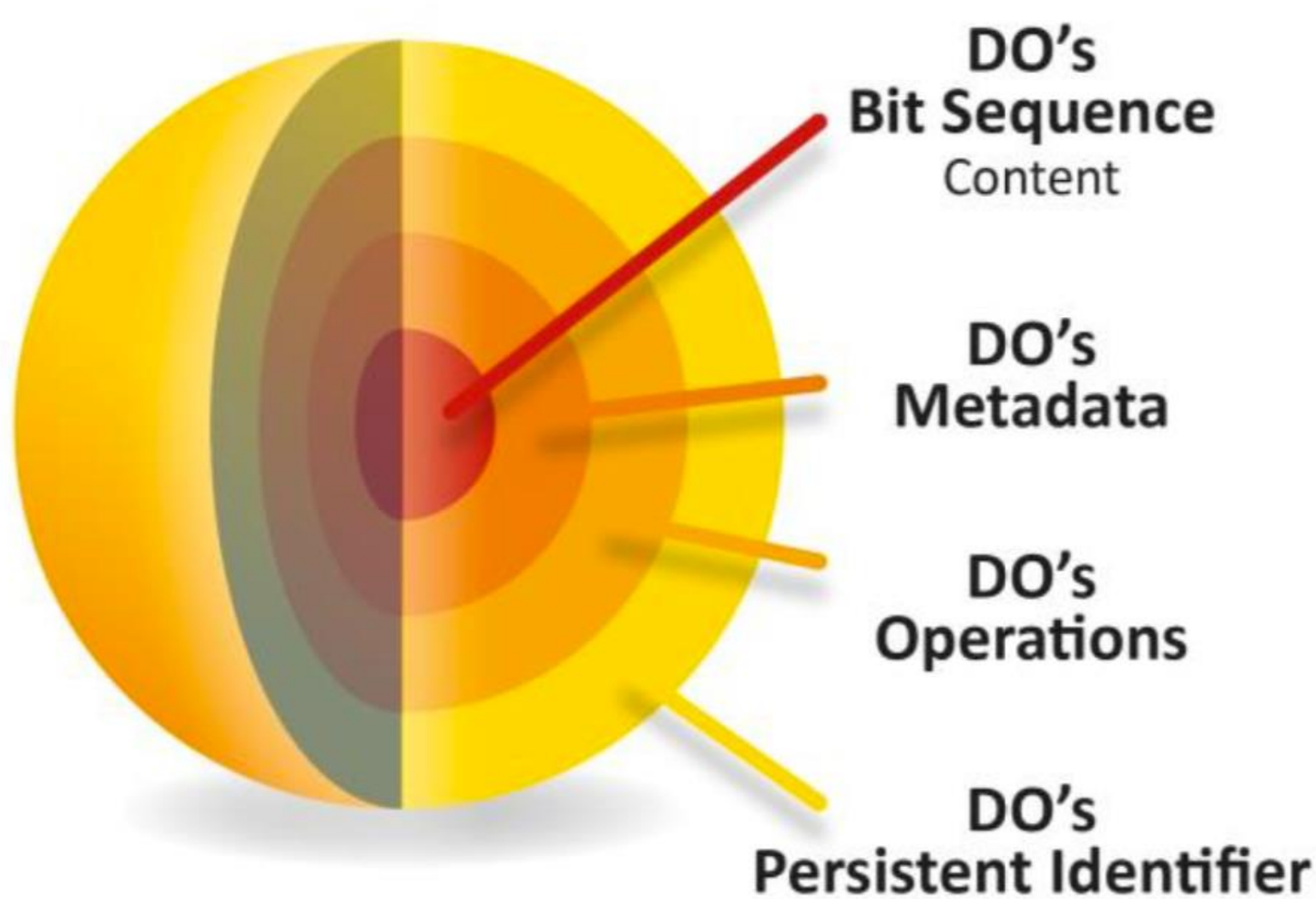
1. **Object** evokes **Concept** (in writer's or speaker's mind)
2. Writer/speaker uses **Token** to refer to **Object**
3. **Token** evokes **Concept** (in reader's or listener's mind)
4. Reader/listener refers **Token** back to **Object**

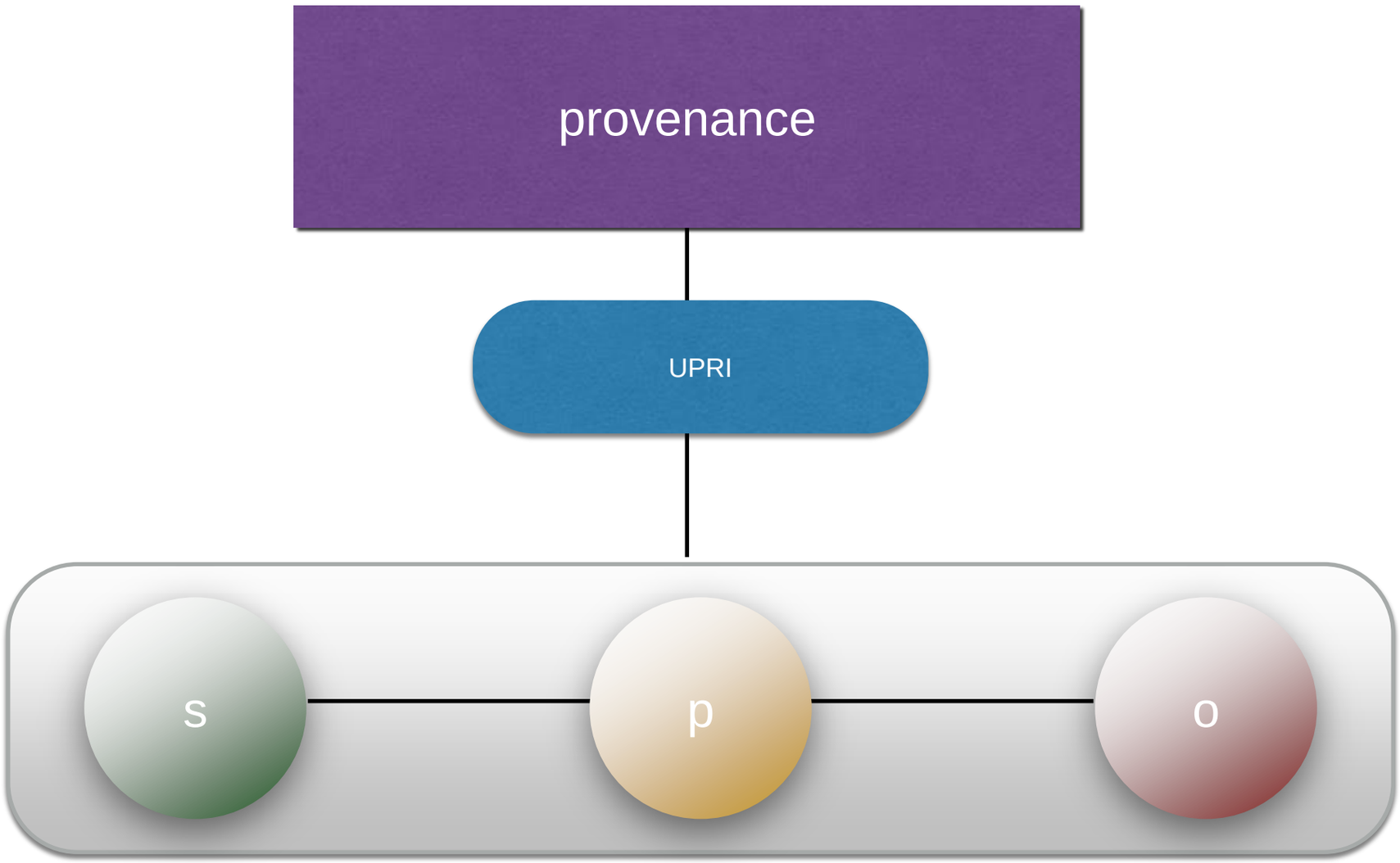
Http://GOFAIR/UUID.??

Token or word or icon:

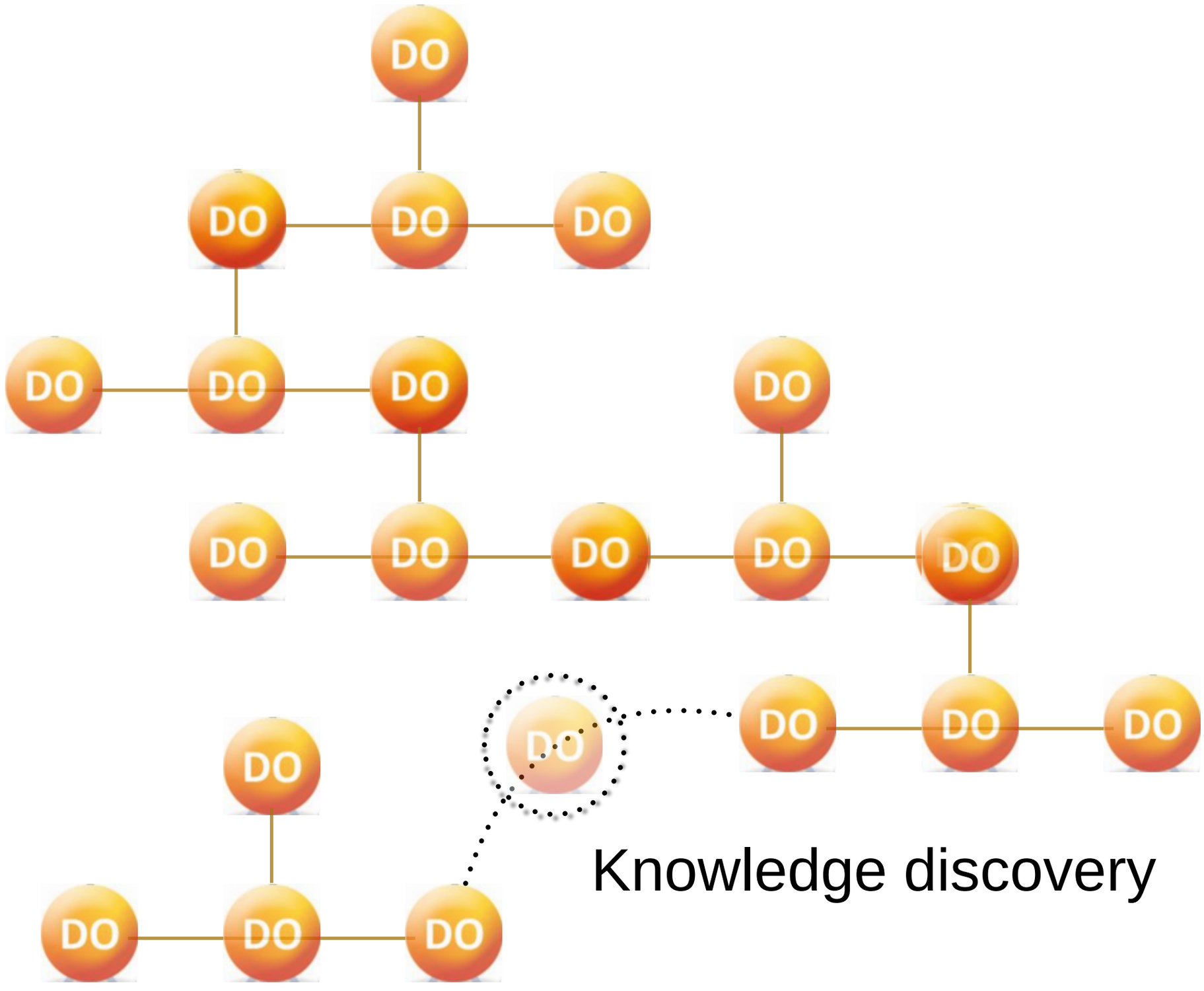
'cancer'
Malignant Neoplasms
Krebskrankheit
C0-265
Etc…

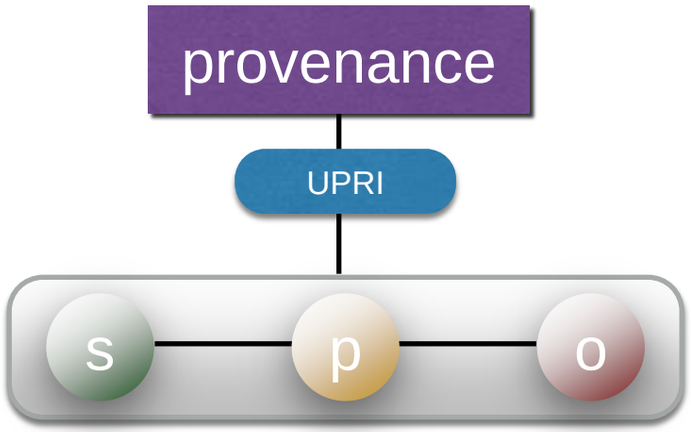
object, entity, defined meaning


DO's
Bit Sequence
Content
DO's
Metadata
DO's
Operations
DO's
Persistent Identifier


provenance
UPRI
s
p
o

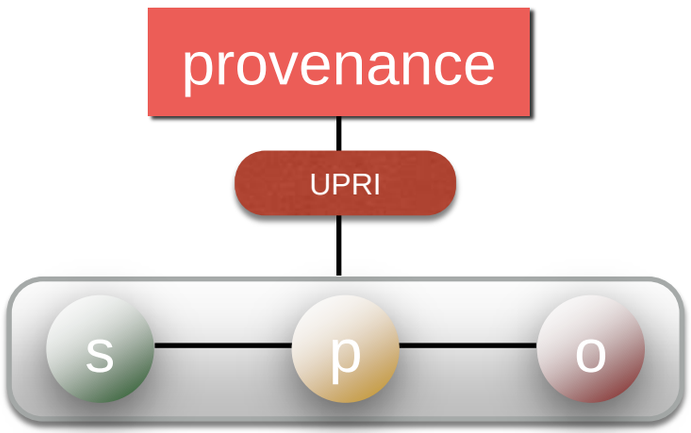
SPO tripples as collections of connected DO’s





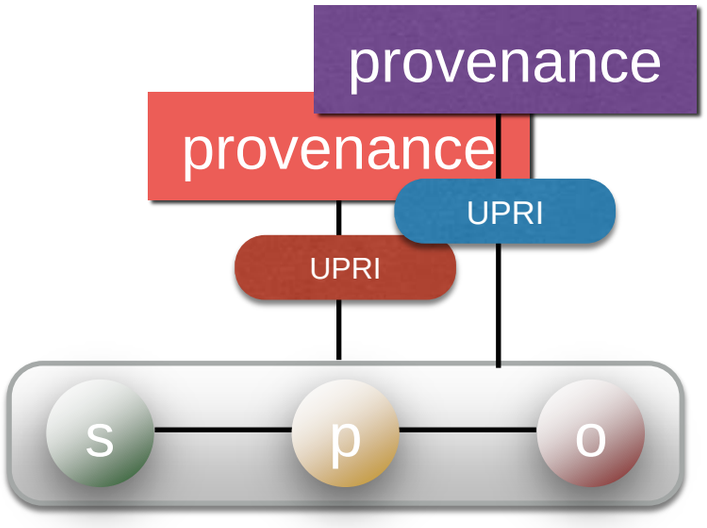
A

A **nanopublication** is the smallest meaningful assertion, minimally one Subject-Predicate-Object triple
S,P, & O are all concepts and thus all have Unique, Persistent and Resolvable Identifiers. Many nanopublications are small graphs with multiple triples forming the assertion



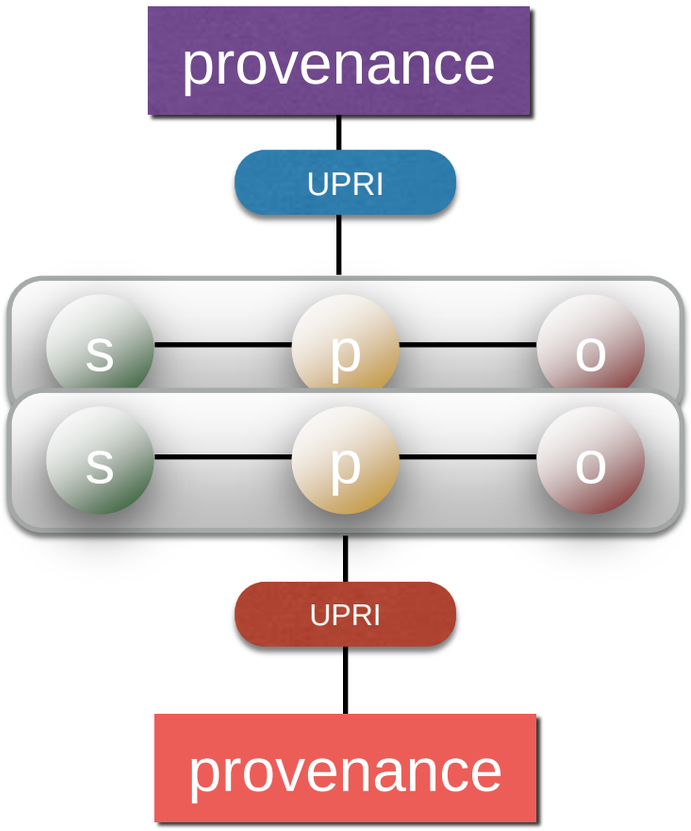
B

Two nanopublications representing the same meaningful assertion, i.e. the Subject-Predicate-Object triples are identical
may have **different provenance** (they come from different sources)
They each have their Persistent and resolvable Identifier.
and different provenance graphs

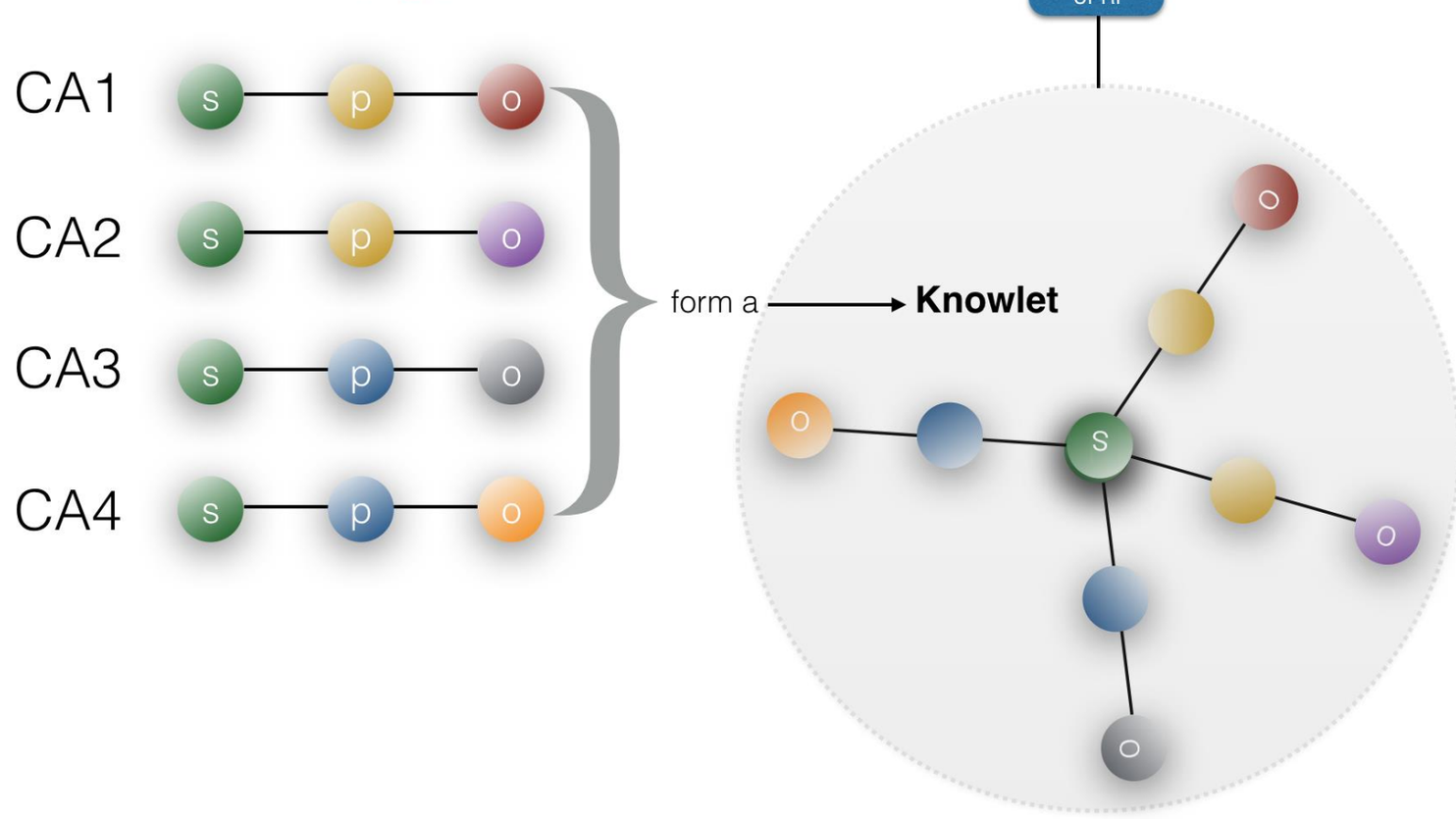


C

A **Cardinal Assertion** is one assertion that is linked to 1-n provenance graphs (up to thousands in some cases)



Multiple different cardinal Assertions*
with the same **subject**



* UPRI's and Provenance not depicted for simplicity reasons

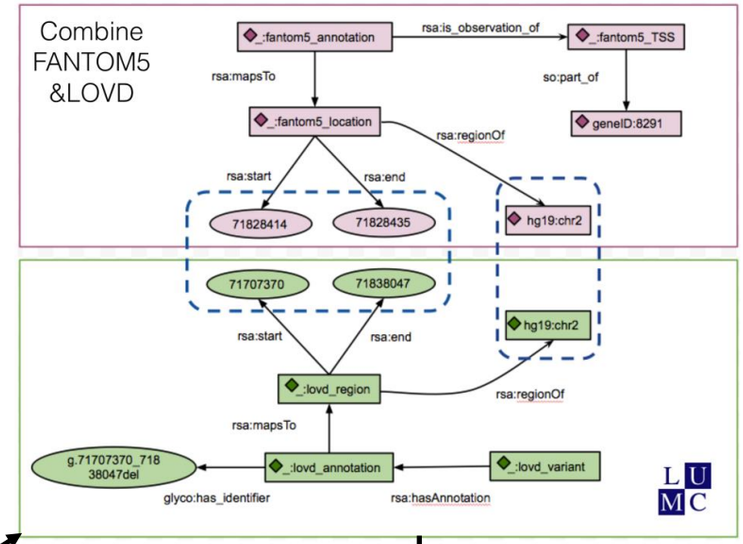
Search

Variant-TTS-protein-Pathology

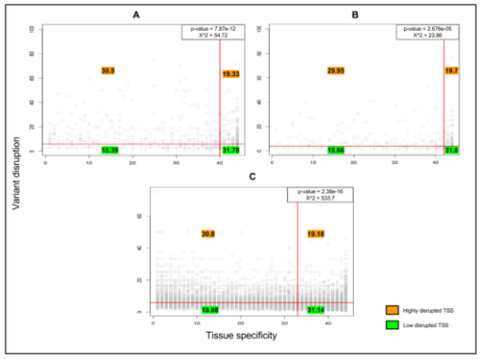
get_variants_overlapping_with_tss.spar ql

LOVD
Leiden Open Variation Database

Combine FANTOM5 &LOVD



similarity measure

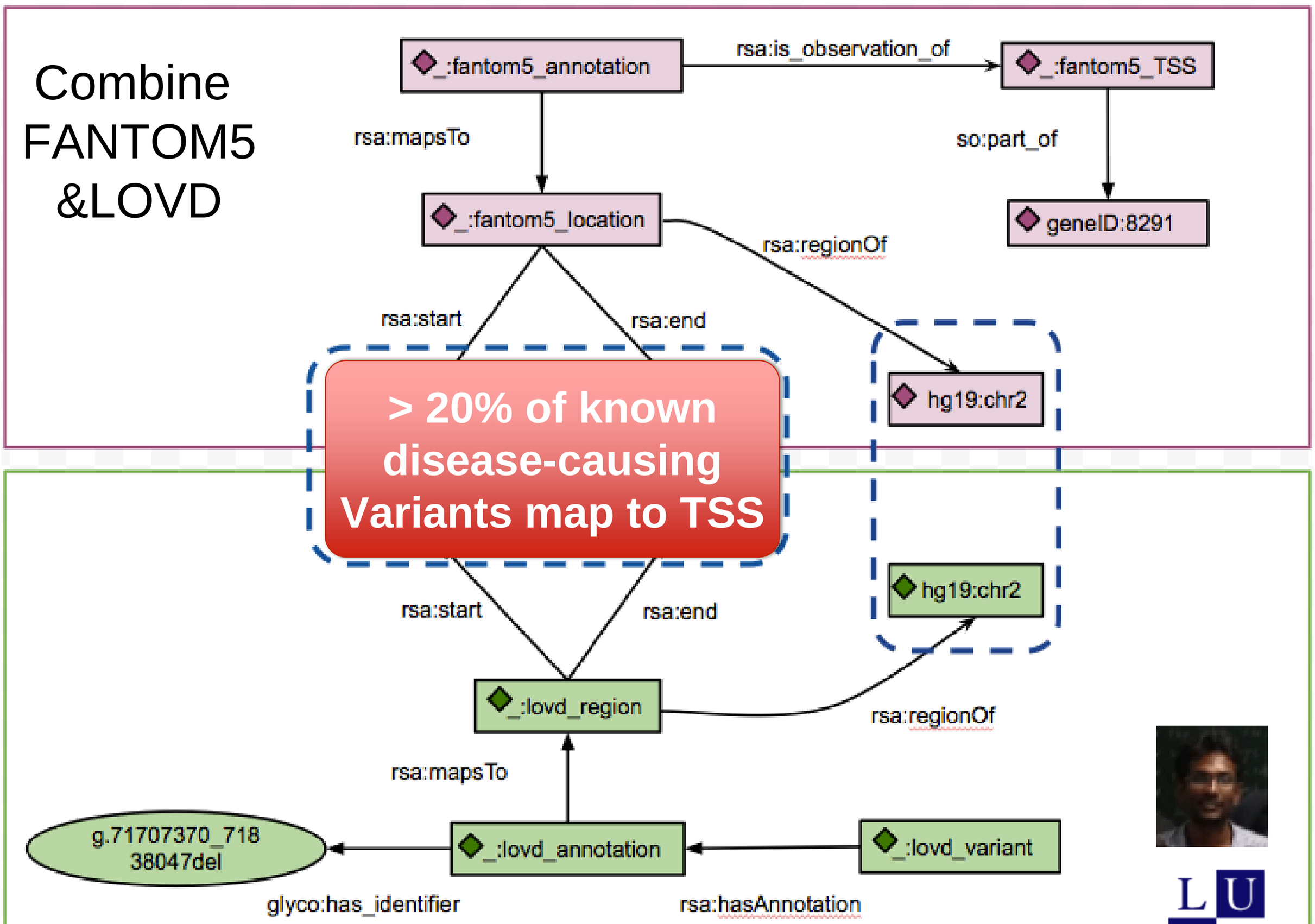


FANTOM5
FUNCTIONAL ANNOTATION OF THE MAMMALIAN GENOME

PATHOLOGY ATLAS

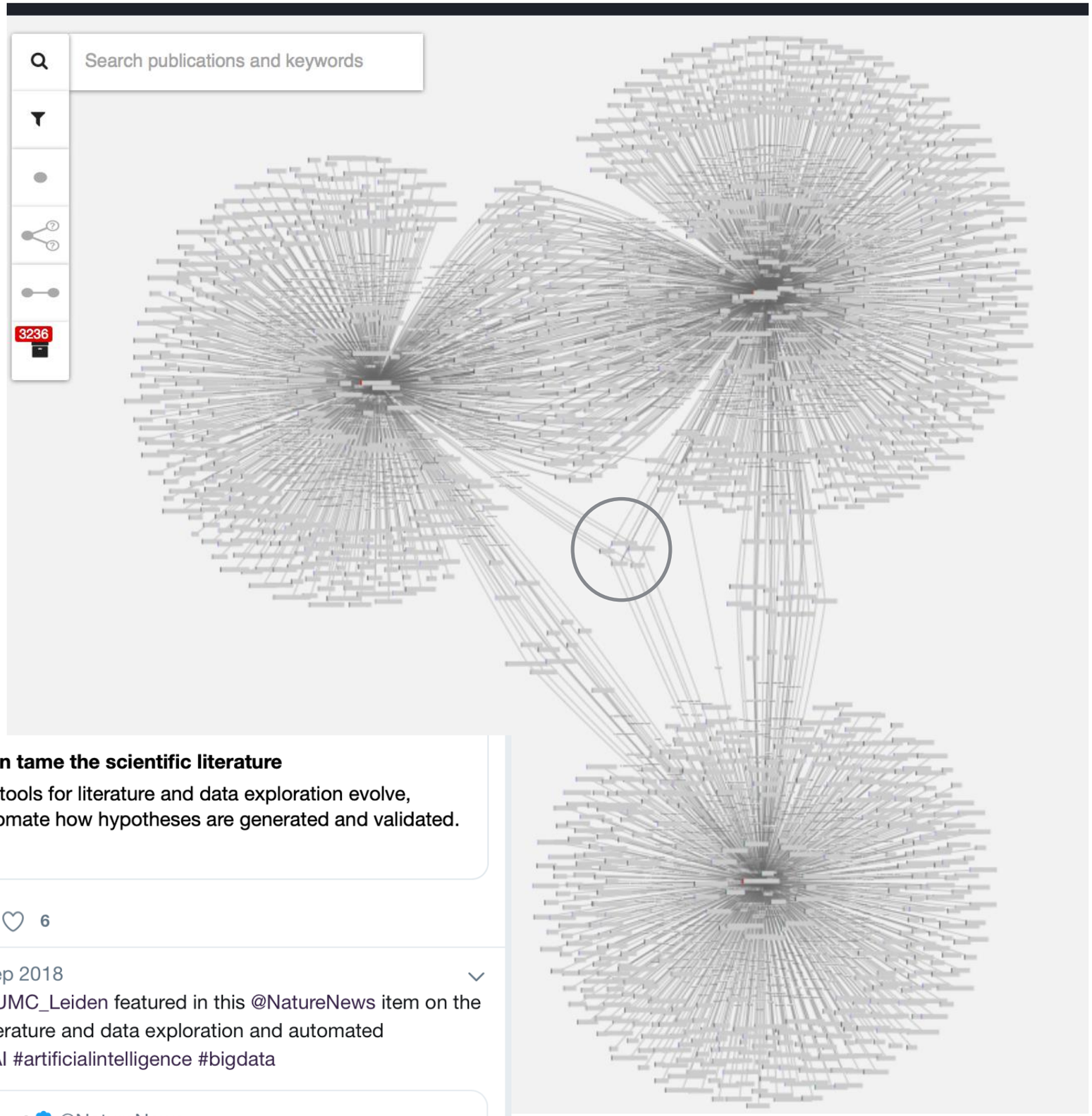
Disease Causing Variants

Combine FANTOM5 &LOVD



LUMC

5 objects are shared between all three knowlets
(in this case: metabolic syndrome, diabetes, and e.o Alzheimer)



How AI technology can tame the scientific literature

As artificially intelligent tools for literature and data exploration evolve, developers seek to automate how hypotheses are generated and validated.

nature.com

1 6

Euretos @Euretos · 10 Sep 2018

@euretos and partner @LUMC_Leiden featured in this @NatureNews item on the use of AI technology in literature and data exploration and automated hypotheses generation #AI #artificialintelligence #bigdata

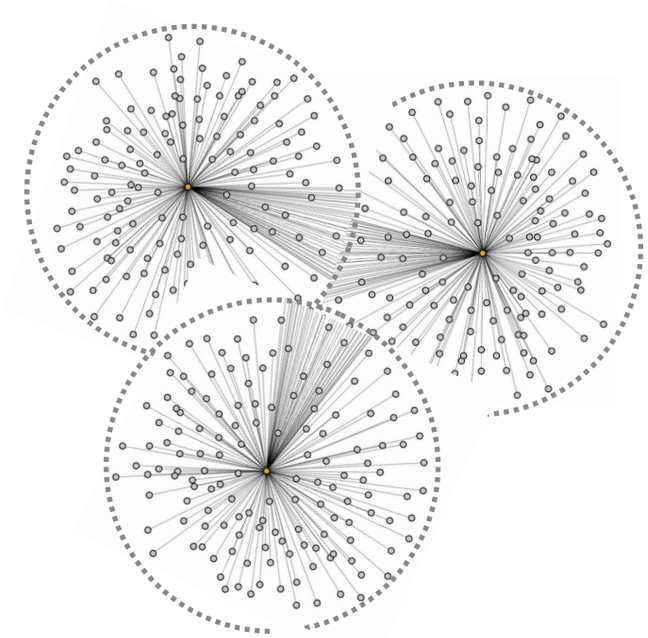
Nature News & Comment @NatureNews

What would you do when faced with more than 10,000 papers for a literature review? go.nature.com/2N4wyuc

The value of knowlets in dynamic ontological graphs

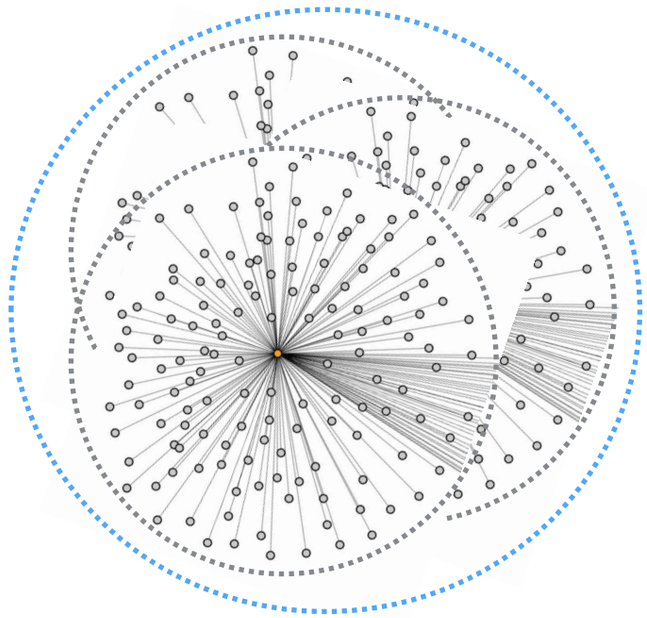
A

conceptual similarity
(hypothesis generation)



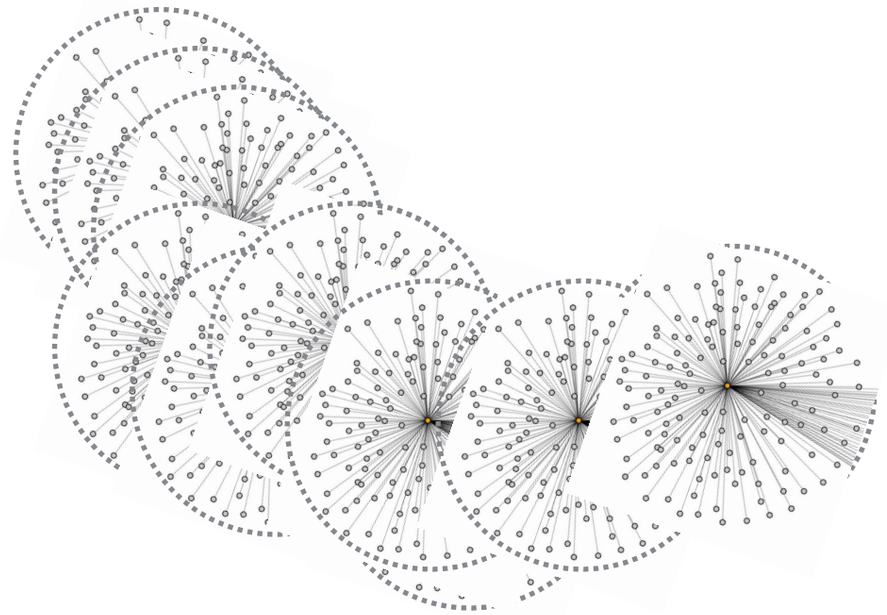
B

Near sameness
(semantic lenses)



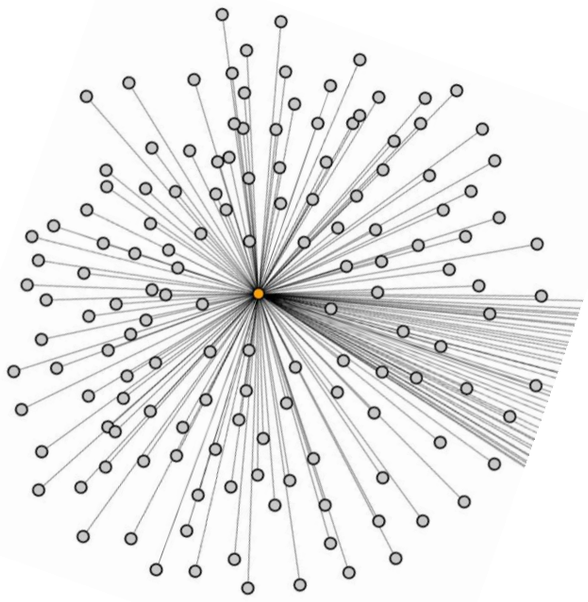
C

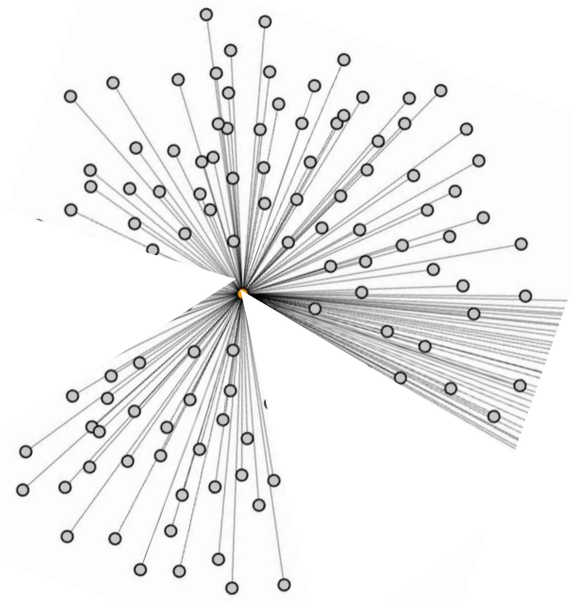
conceptual drift
(meta-data/blockchain)



D

QUA’s
(semantic bias)





'see you at your data'