The Internet for Social Machines

The end of data sharing as we know it

FAIR > The Machine Knows what I Mean

Barend Mons

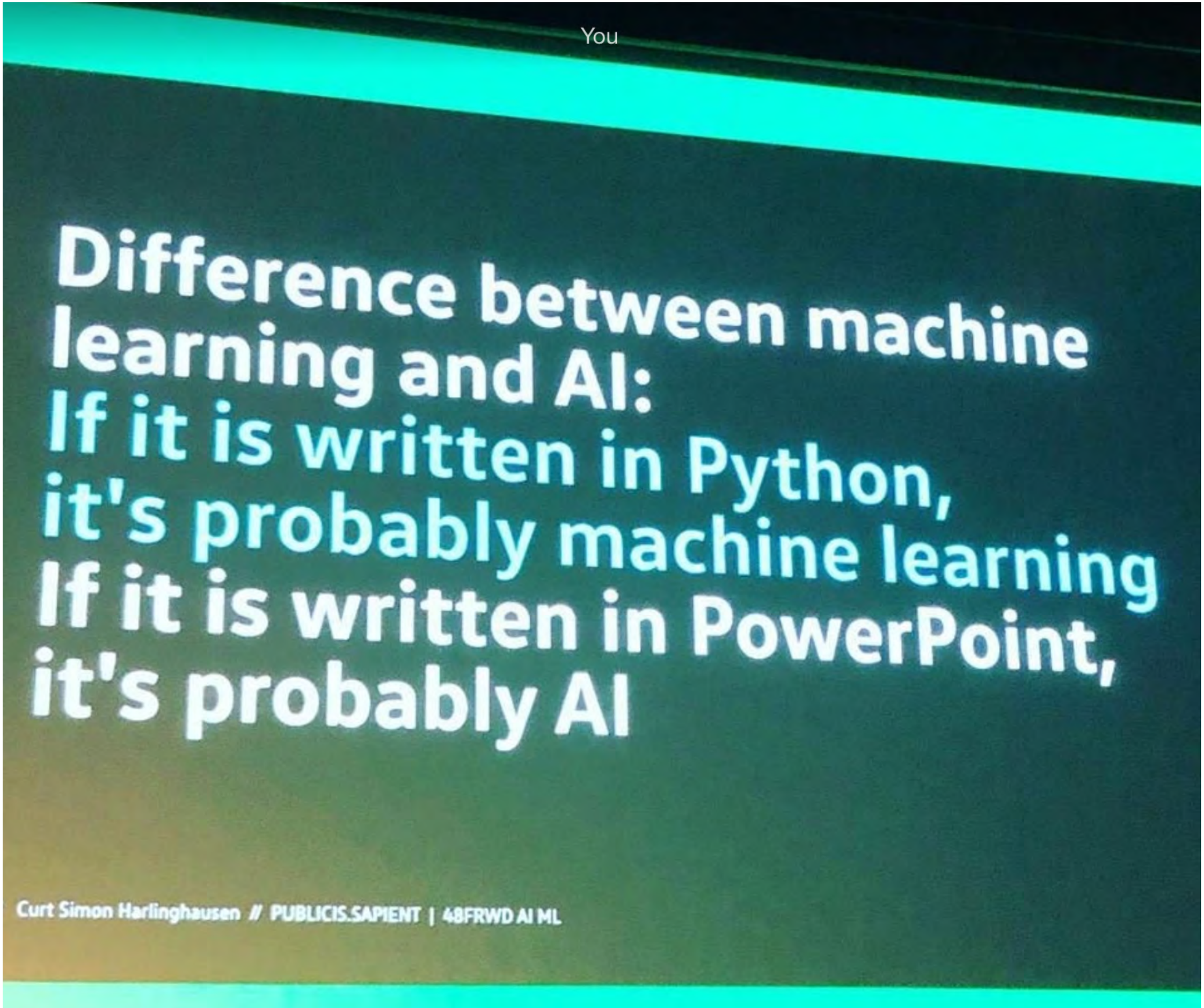
July 8th 2019, Madrid

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**

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

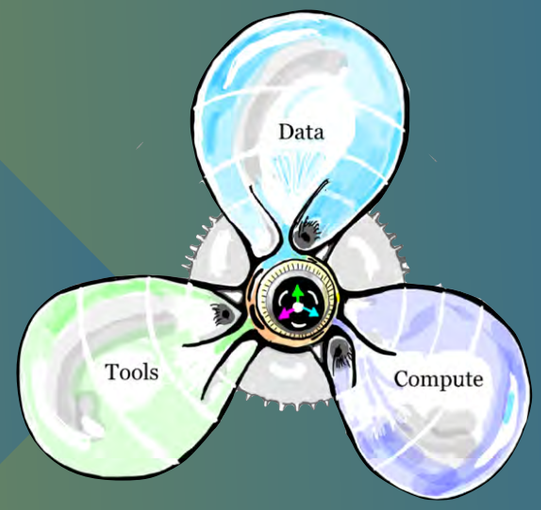
FAIR and GO FAIR

Lorentz

EOSC

FAIRdICT

IFDS



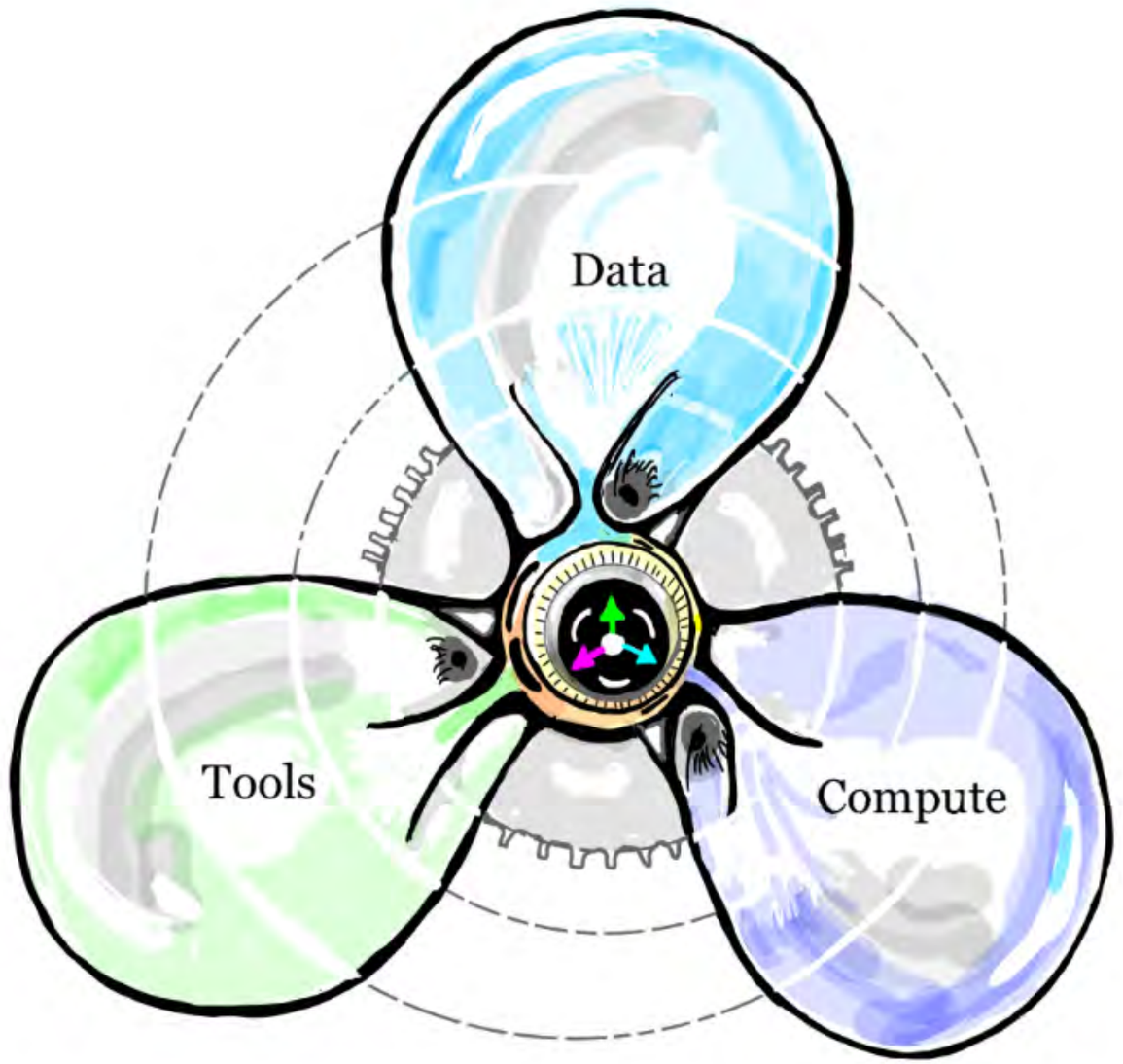
EOSC

Birth	Infancy		Adolescence		Maturity
2014	2015	2016	2017	2018…	

The Road to FAIRness

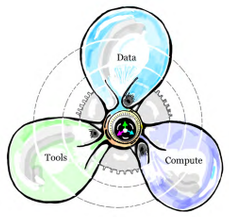
The Road to FAIRness

The Internet of FAIR data and Services



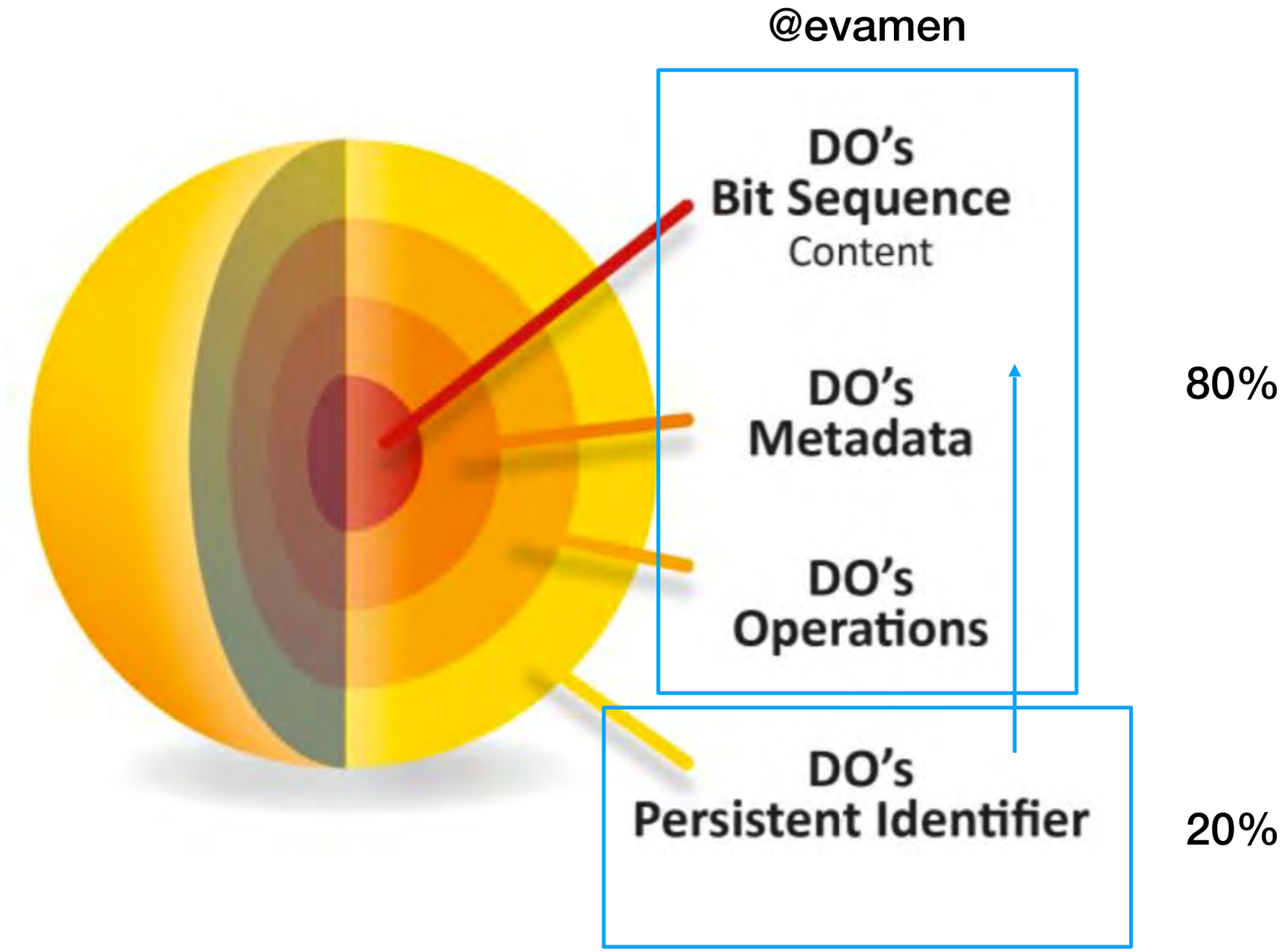
The Internet for Social Machines

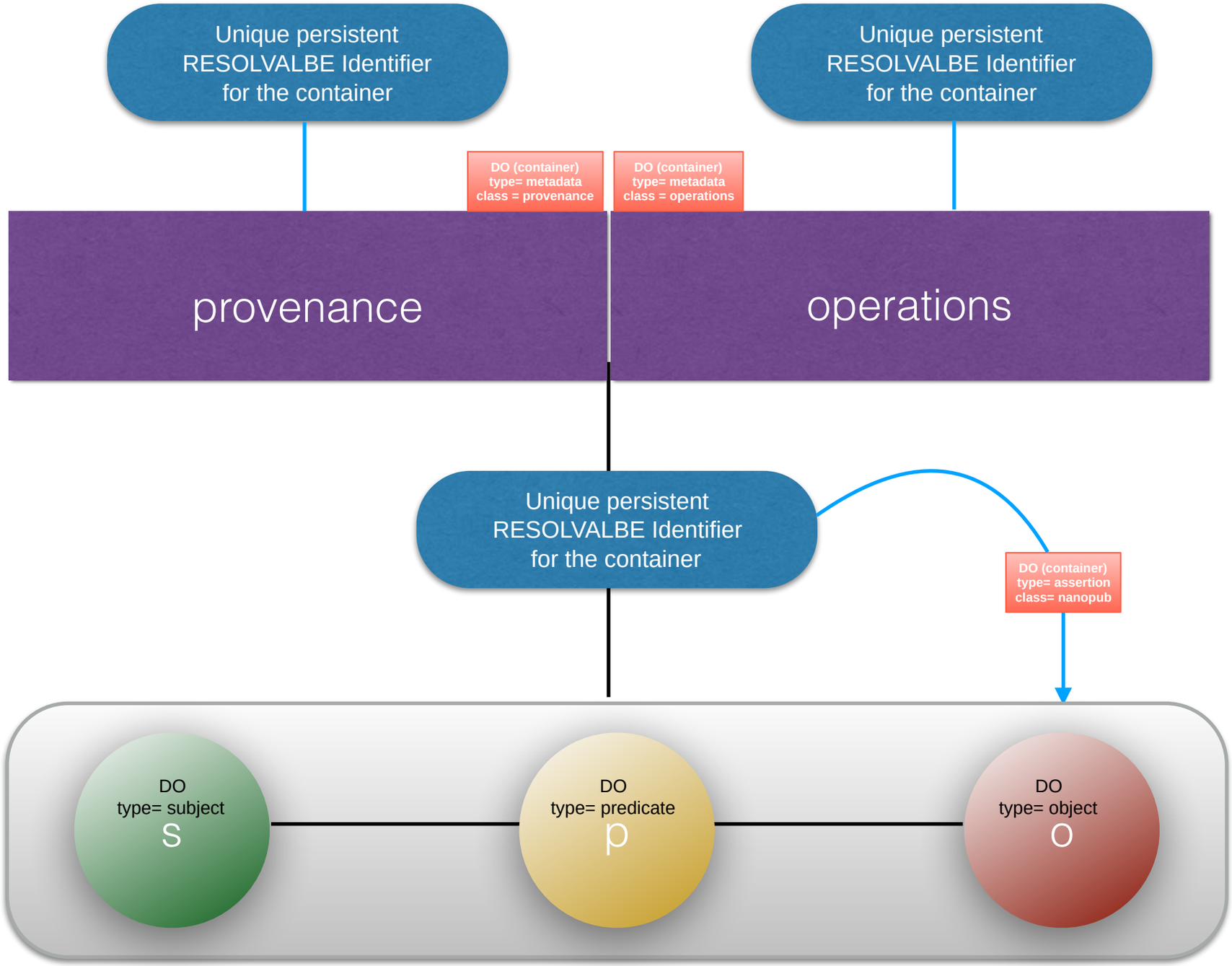
The end of data sharing as we know it

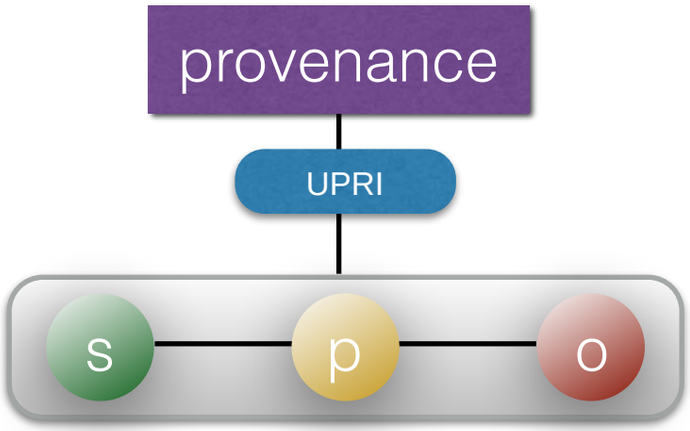


Distributed learning by VM's

Minimally: treat everything as a digital, transferable FAIR object

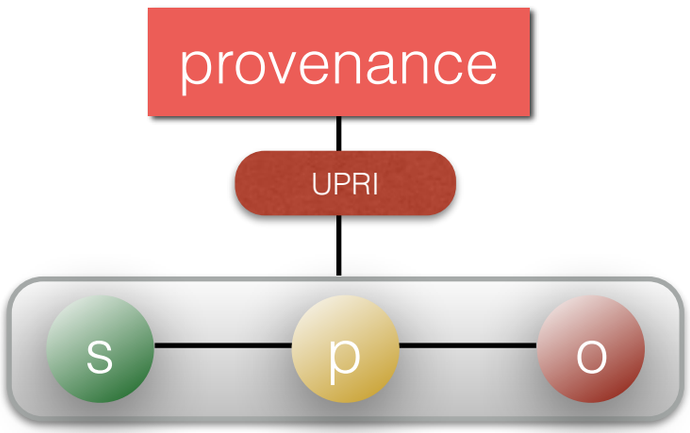



Unique persistent
RESOLVALBE Identifier
for the container
Unique persistent
RESOLVALBE Identifier
for the container
DO (container)
type= metadata
class = provenance
DO (container)
type= metadata
class = operations
provenance
operations
Unique persistent
RESOLVALBE Identifier
for the container
DO (container)
type= assertion
class= nanopub
DO
type= subject
S
DO
type= predicate
p
DO
type= object
O



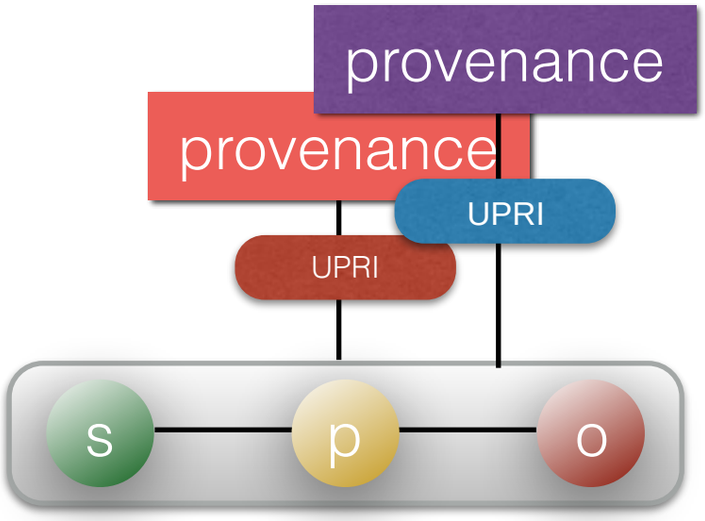
A

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



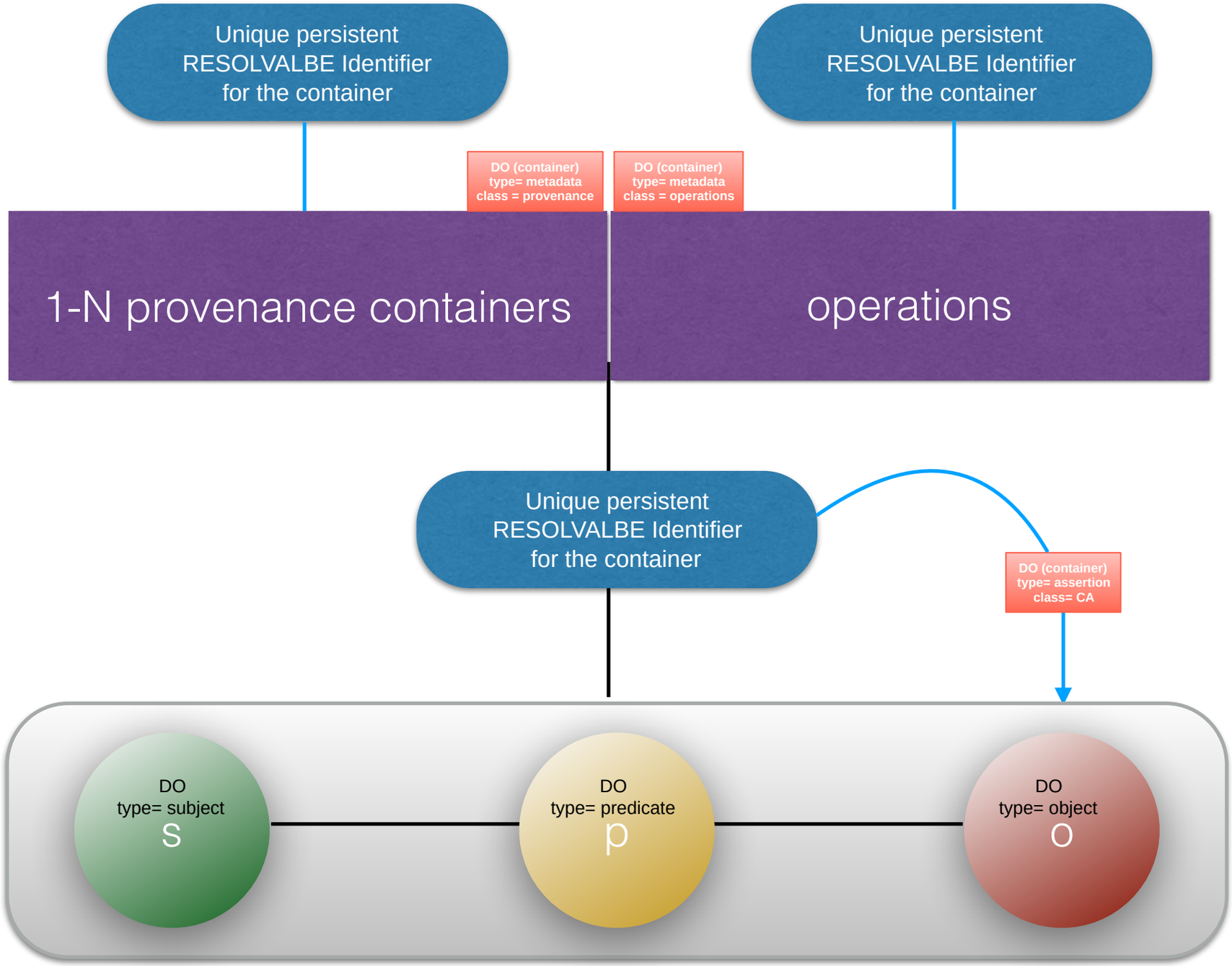
B

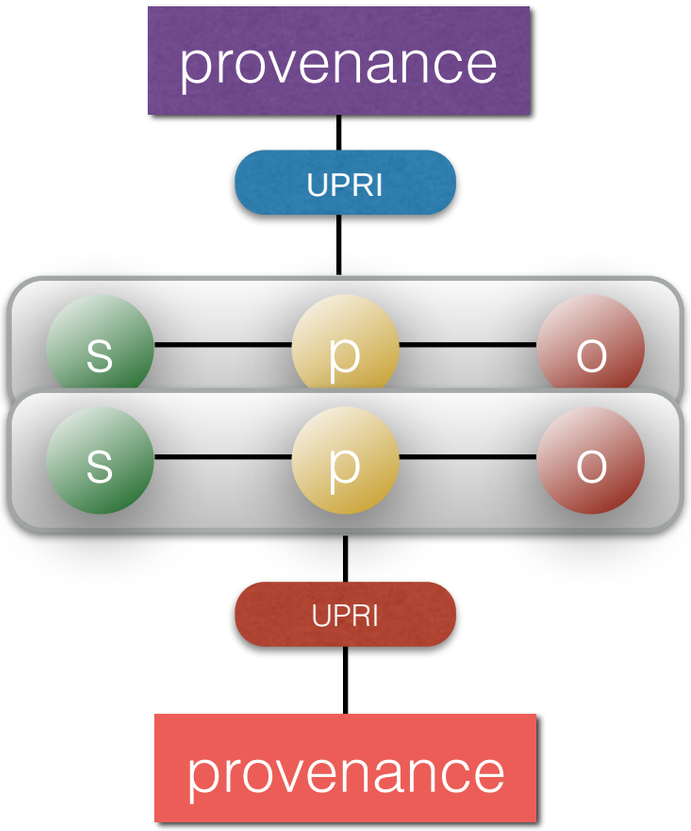
Two nanopublications representing the same meaningful assertion, i.e. the Subject-Predicate-Object triples are identical
may have **different provenance** (they come from different sources)
They each have their Persistent and resolvable Identifier.
and different provenance graphs



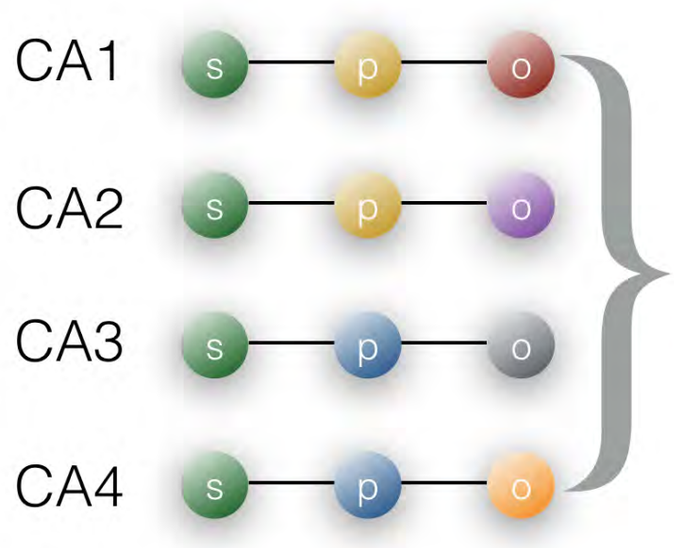
C

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

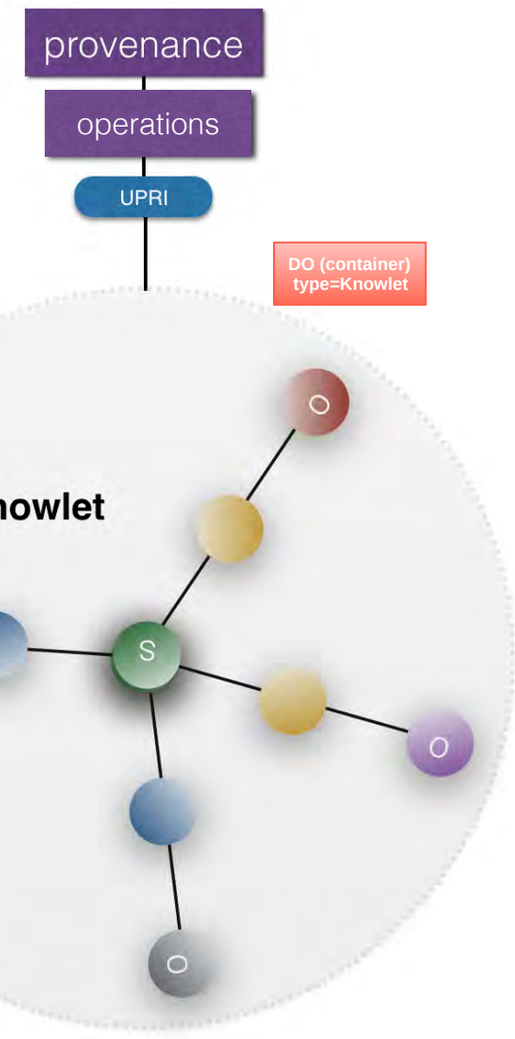

Unique persistent
RESOLVALBE Identifier
for the container
Unique persistent
RESOLVALBE Identifier
for the container
DO (container)
type= metadata
class = provenance
DO (container)
type= metadata
class = operations
1-N provenance containers
operations
Unique persistent
RESOLVALBE Identifier
for the container
DO (container)
type= assertion
class= CA
DO
type= subject
S
DO
type= predicate
p
DO
type= object
O



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



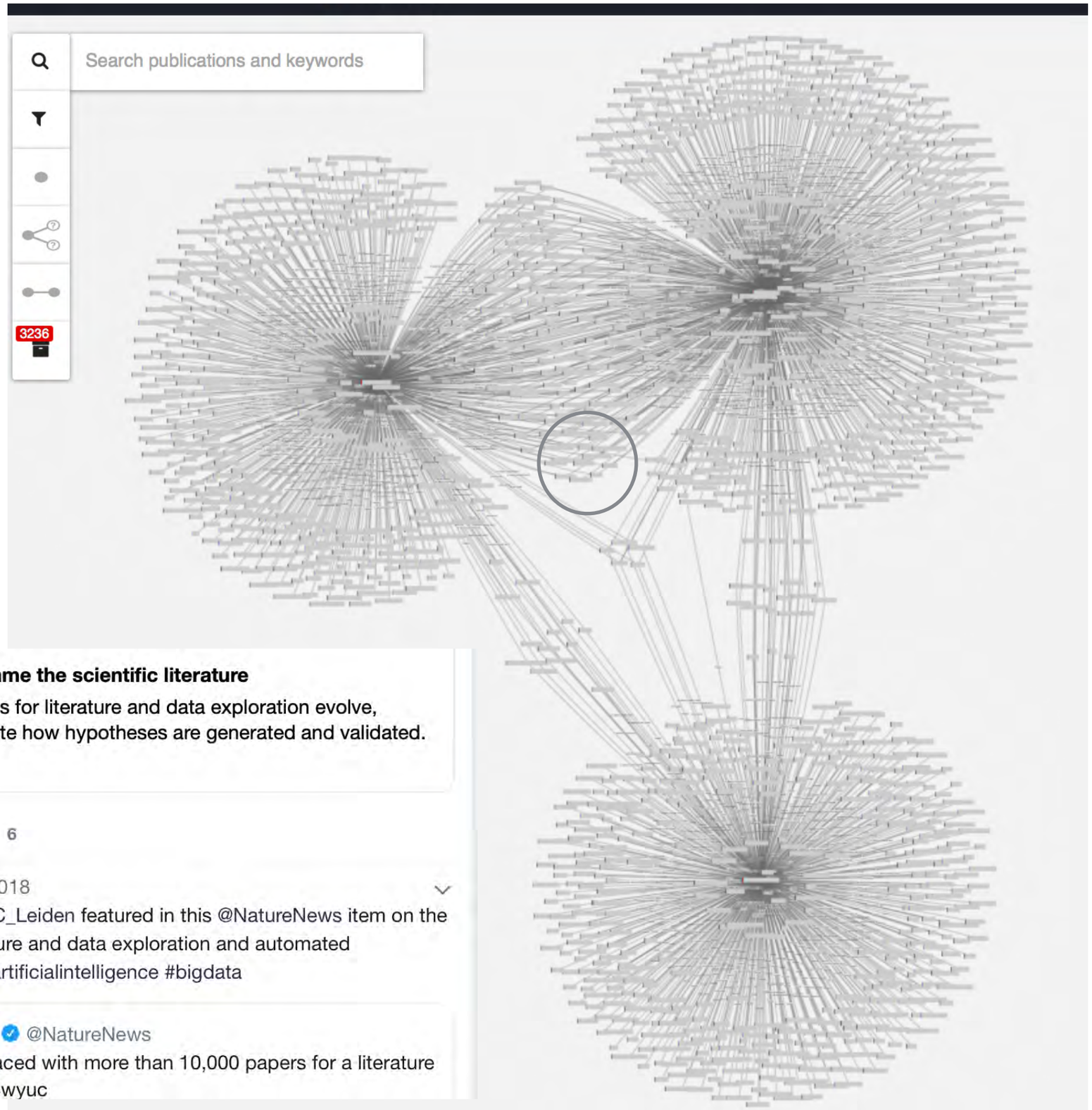
form a → **Knowlet**



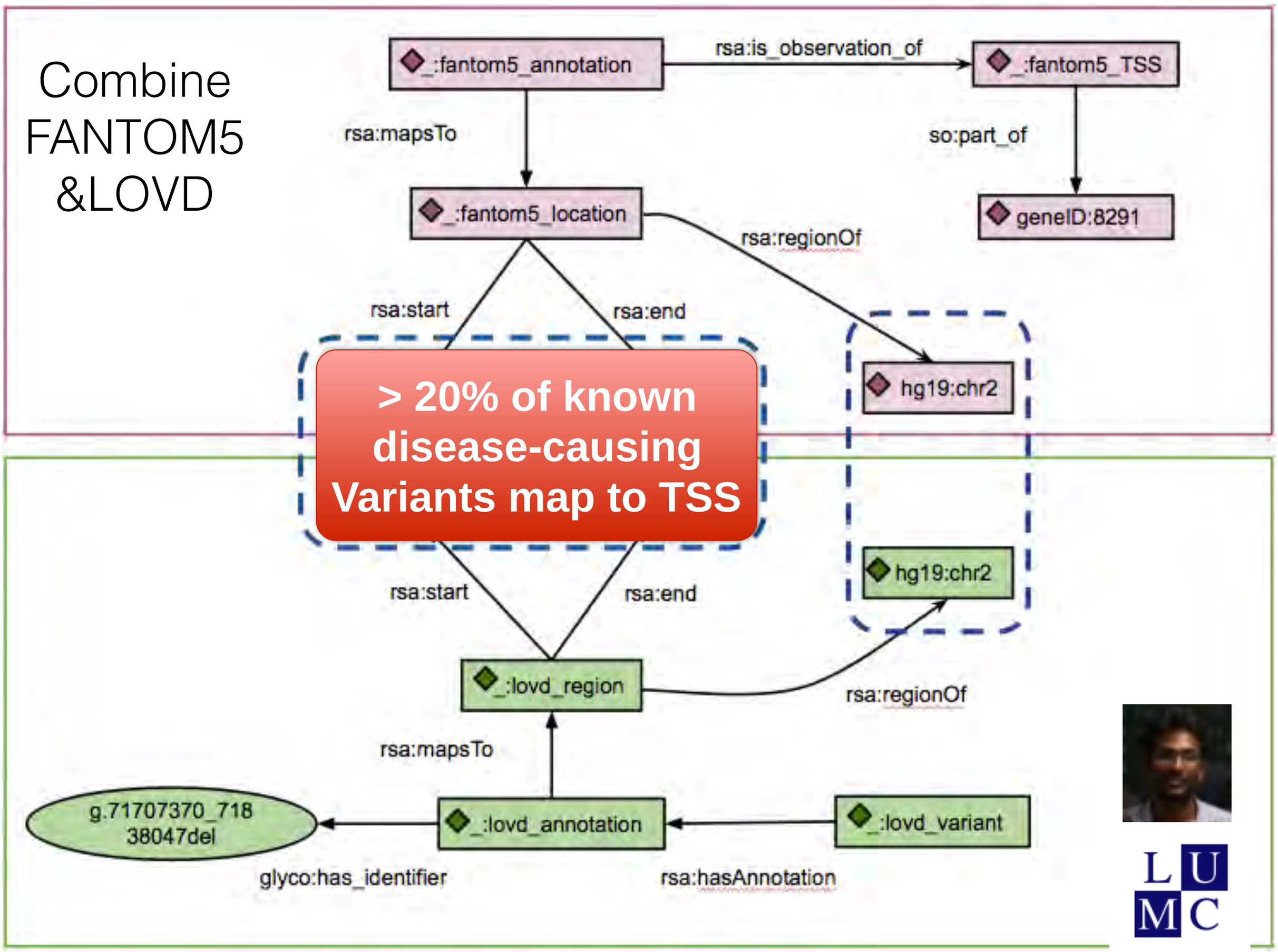
* UPRI's and Provenance not depicted for simplicity reasons

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)

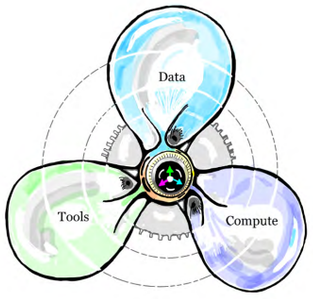


Combine FANTOM5 &LOVD



LUMC

Internet for (social) machines



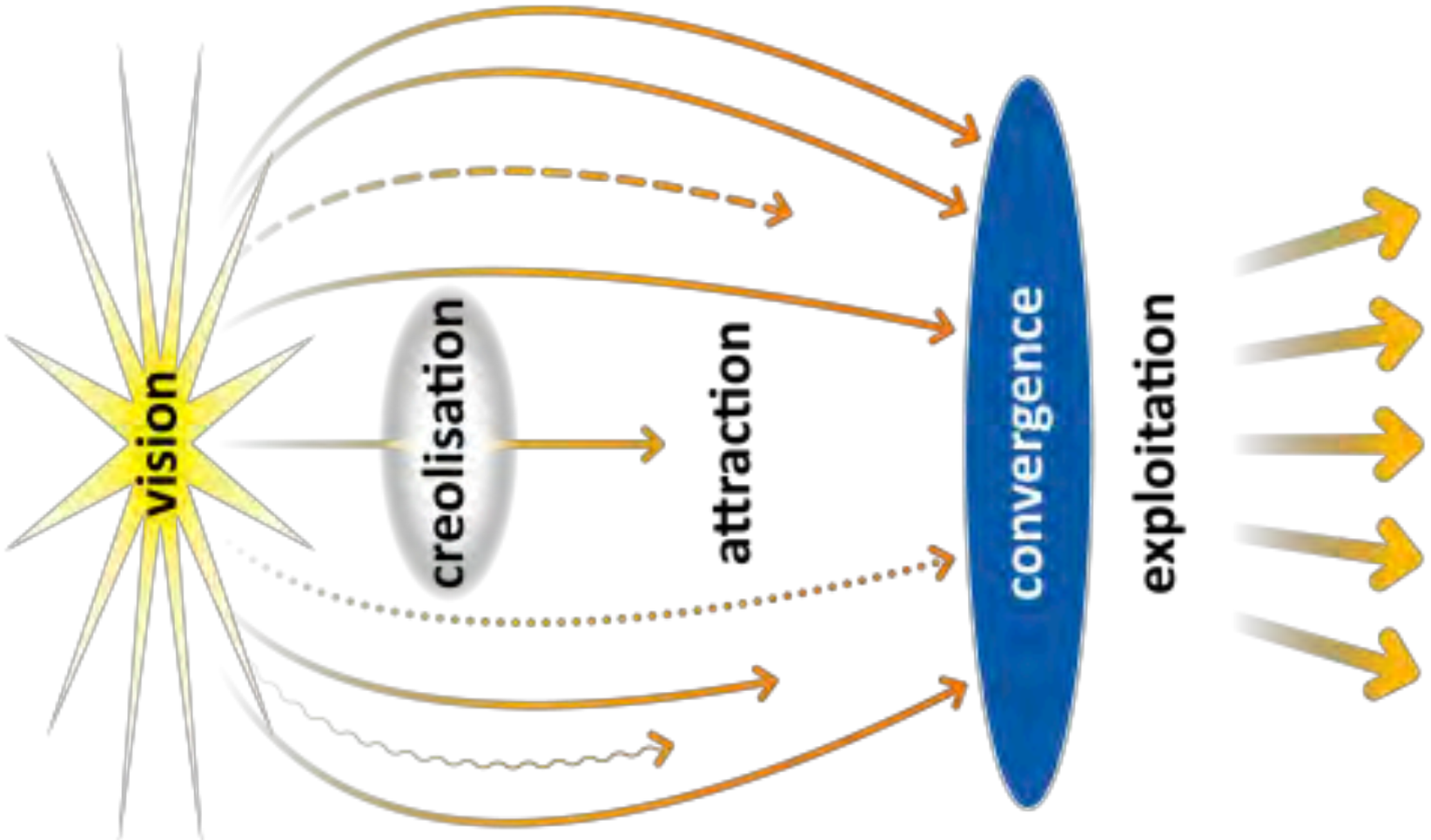
- The Machine Knows what I mean
- As open as possible, as closed as necessary
- As distributed as possible, as central as necessary
- Global: (FAIR in Europe, **F**ully **AI**-**R**eady in USA) :)

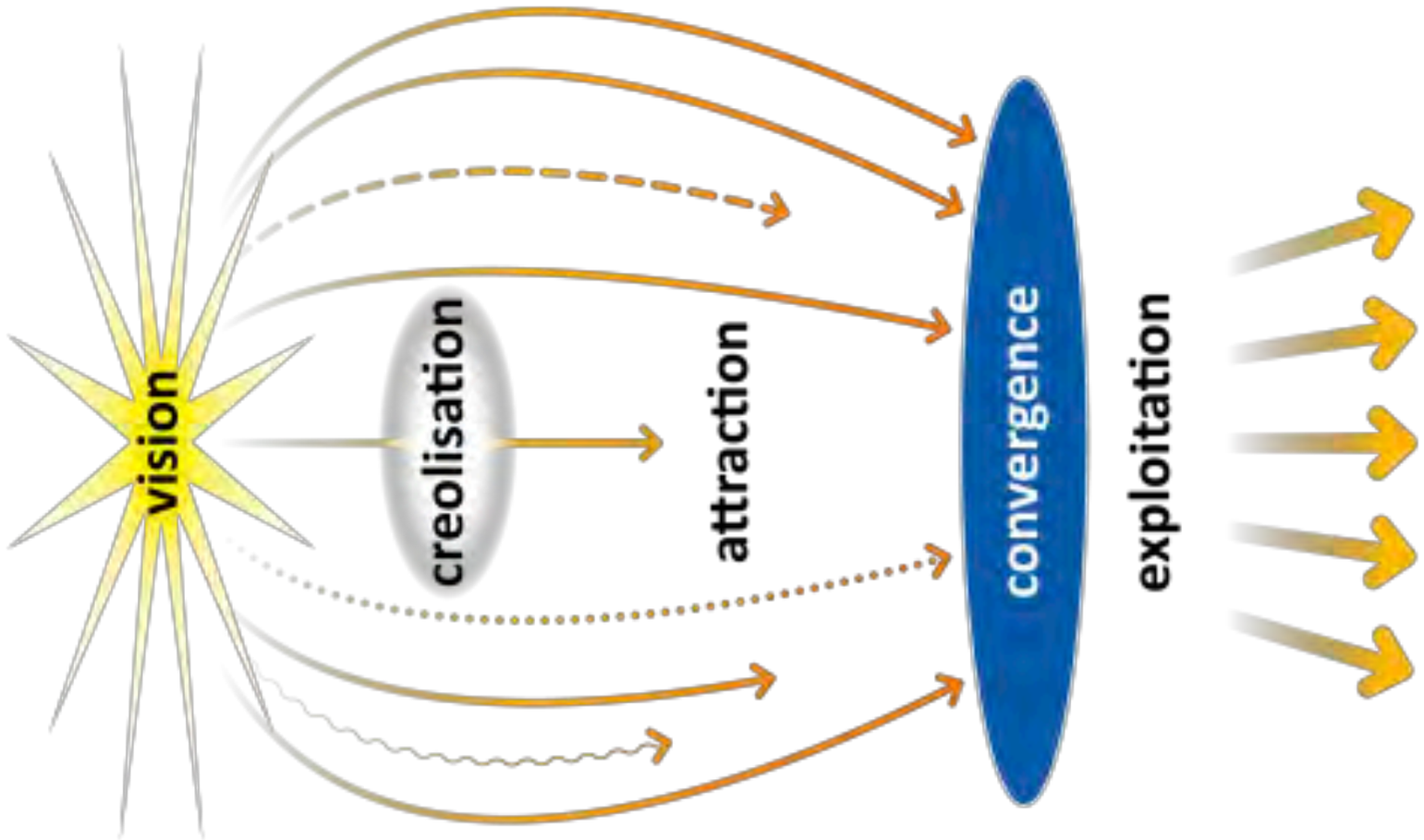
How is the Internet for Social Machines likely to develop?

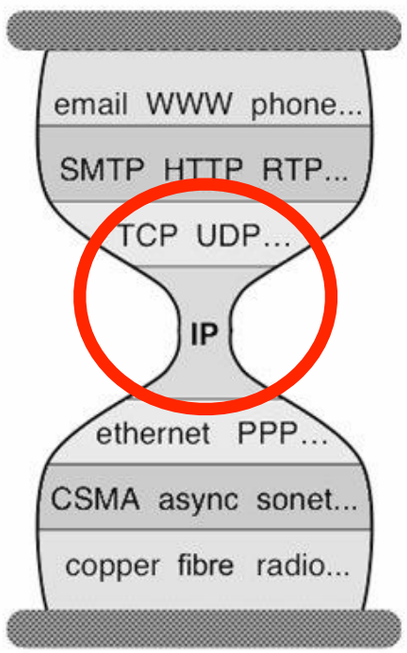
From looming congestion to exploitation !

Common Patterns in Revolutionary Infrastructures and Data

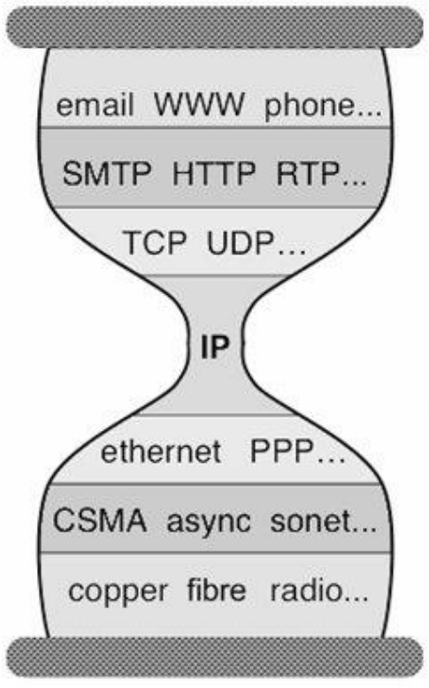
Peter Wittenburg, Max Planck Computing and Data Facility, George Strawn, US National Academy of Sciences, February 2018
https://www.rd-alliance.org/sites/default/files/Common_Patterns_in_Revolutionising_Infrastructures-final.pdf







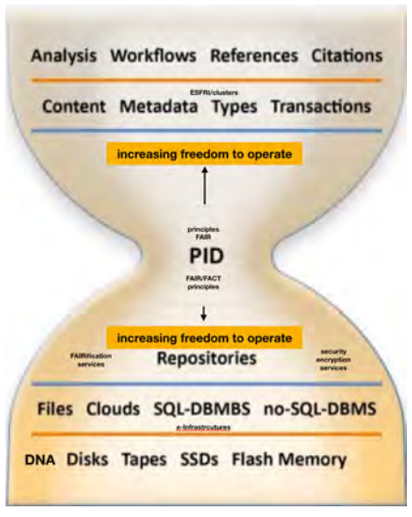
- **Minimal standards**
- **Voluntary participation**
- **Critical mass**



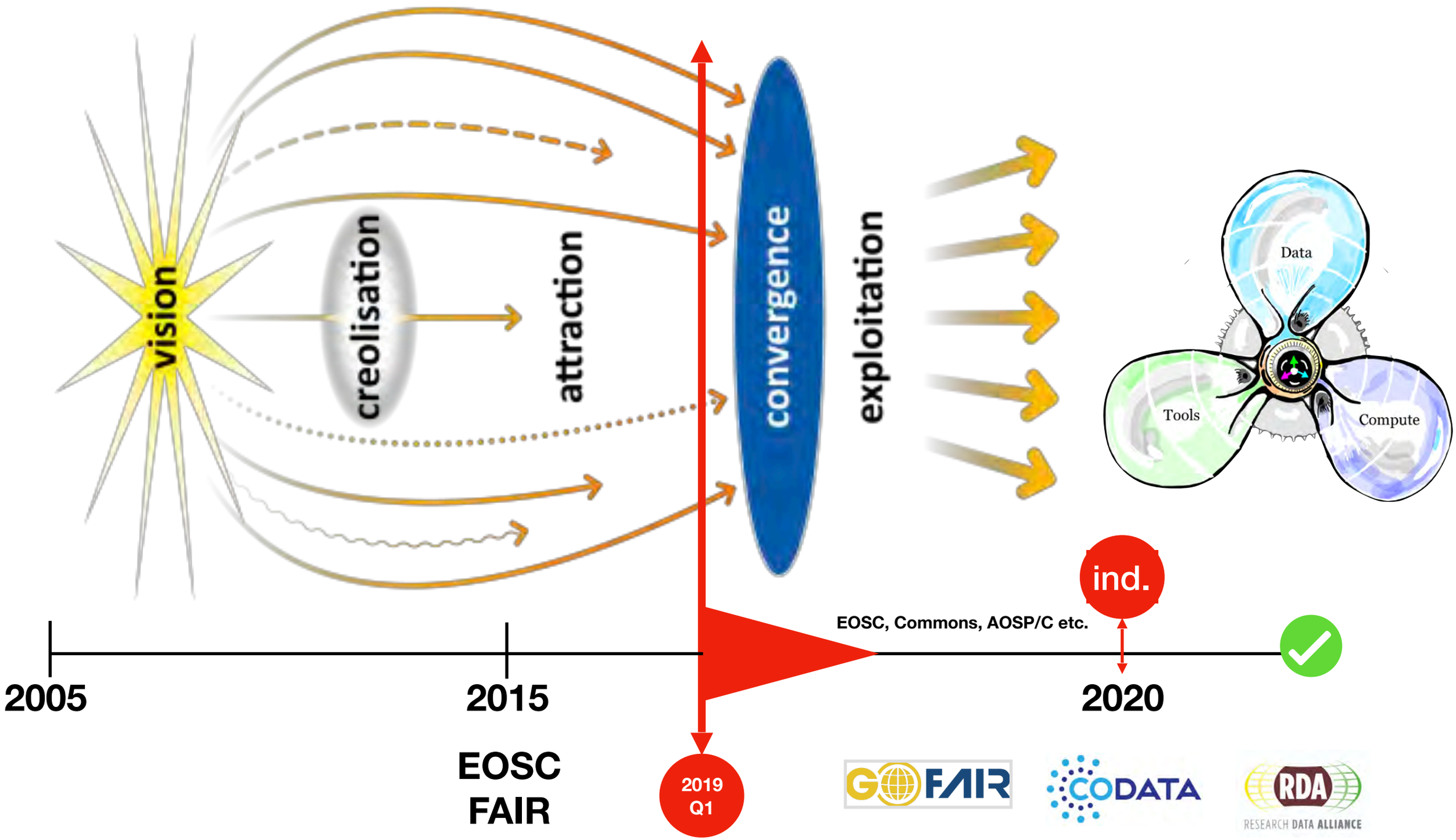
Lessons from the Internet for People:

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 GO FAIR



Its happening RIGHT NOW!



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

CODATA

independent

Global Scope

supra-disciplinary

mandated to be impartial

no competition for research funding

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

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

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
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

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

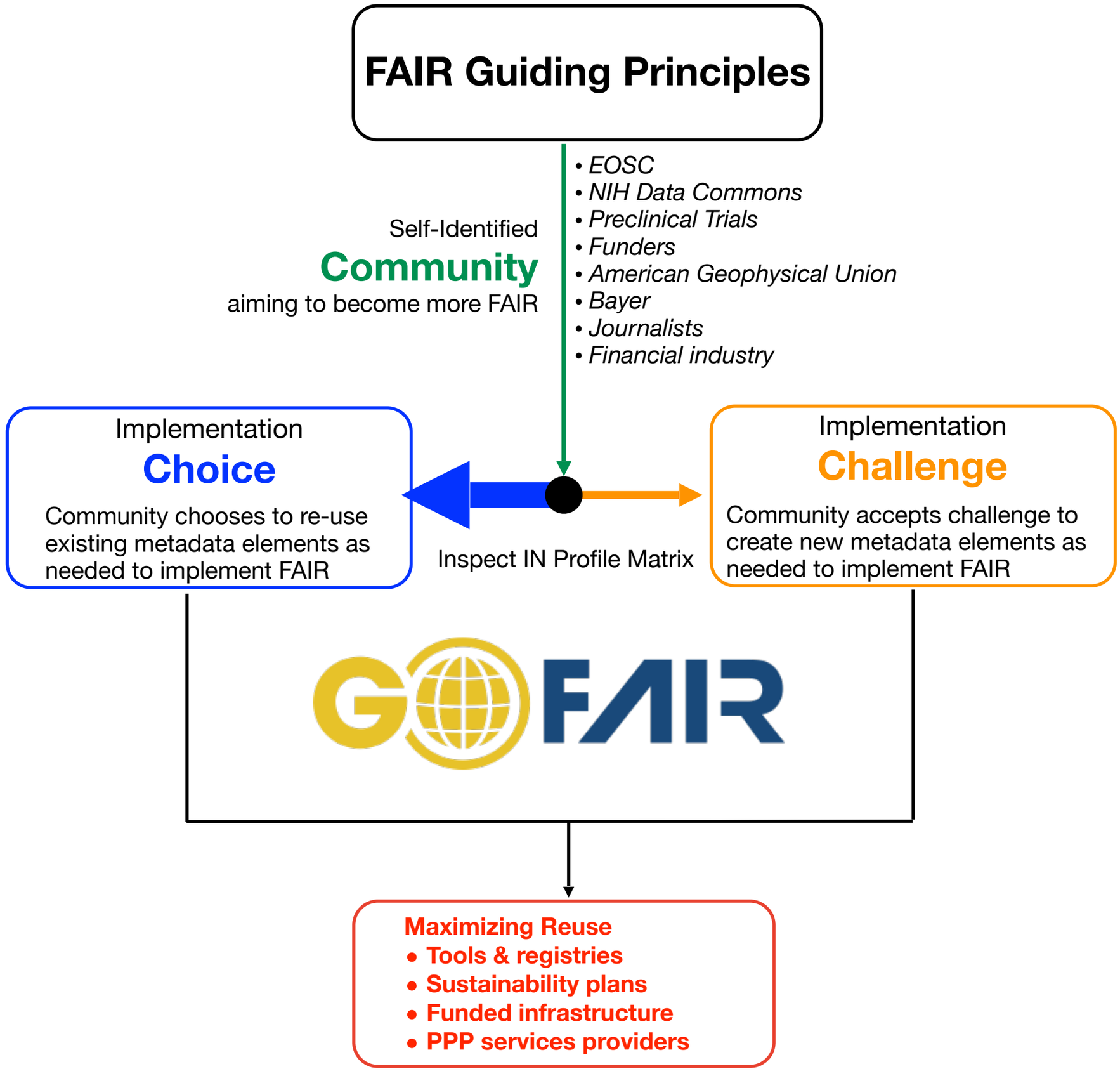
IN Profile Matrix

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

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

	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
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

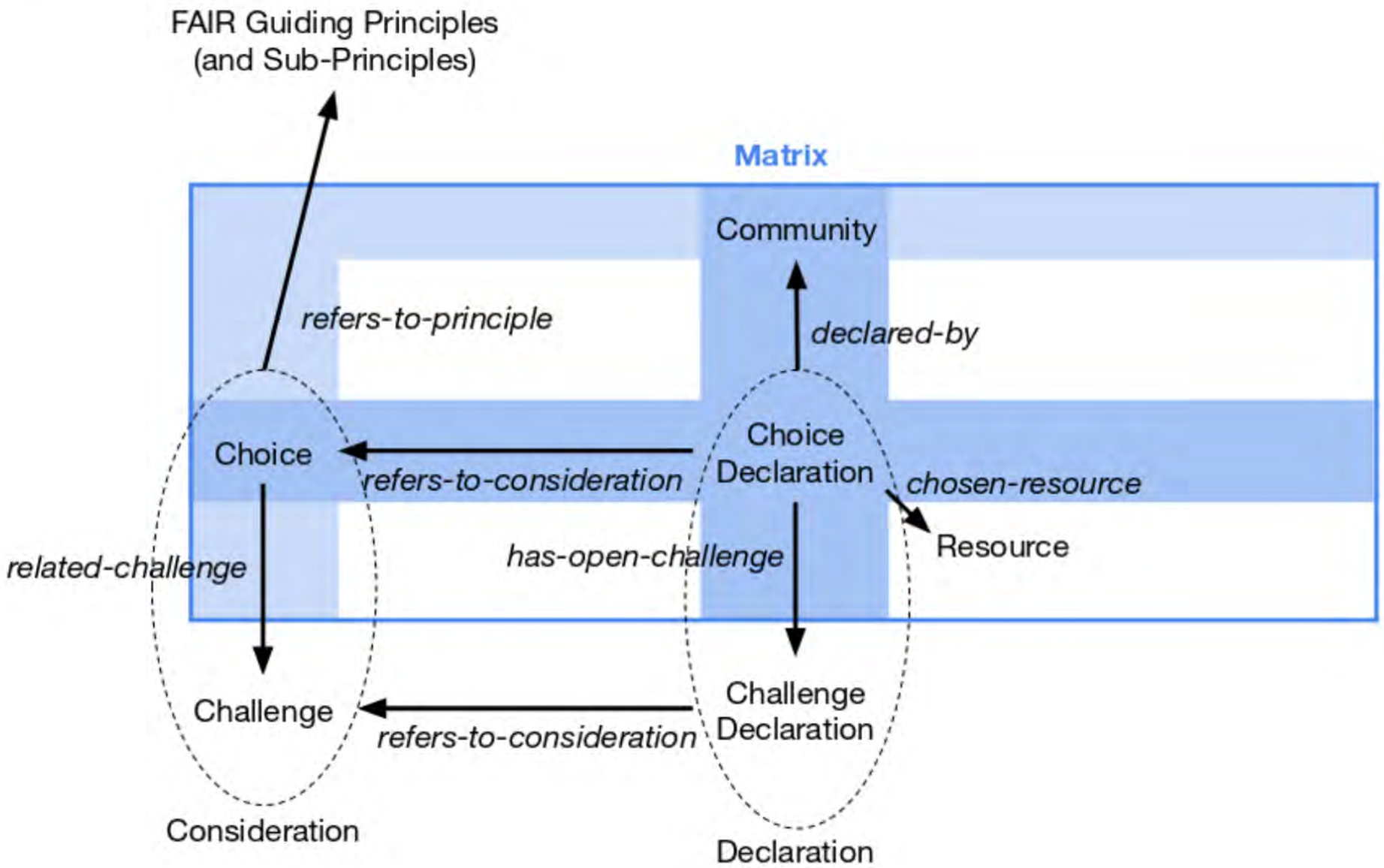
Community Implementation Choices & Challenges



Community Implementation Choices & Challenges

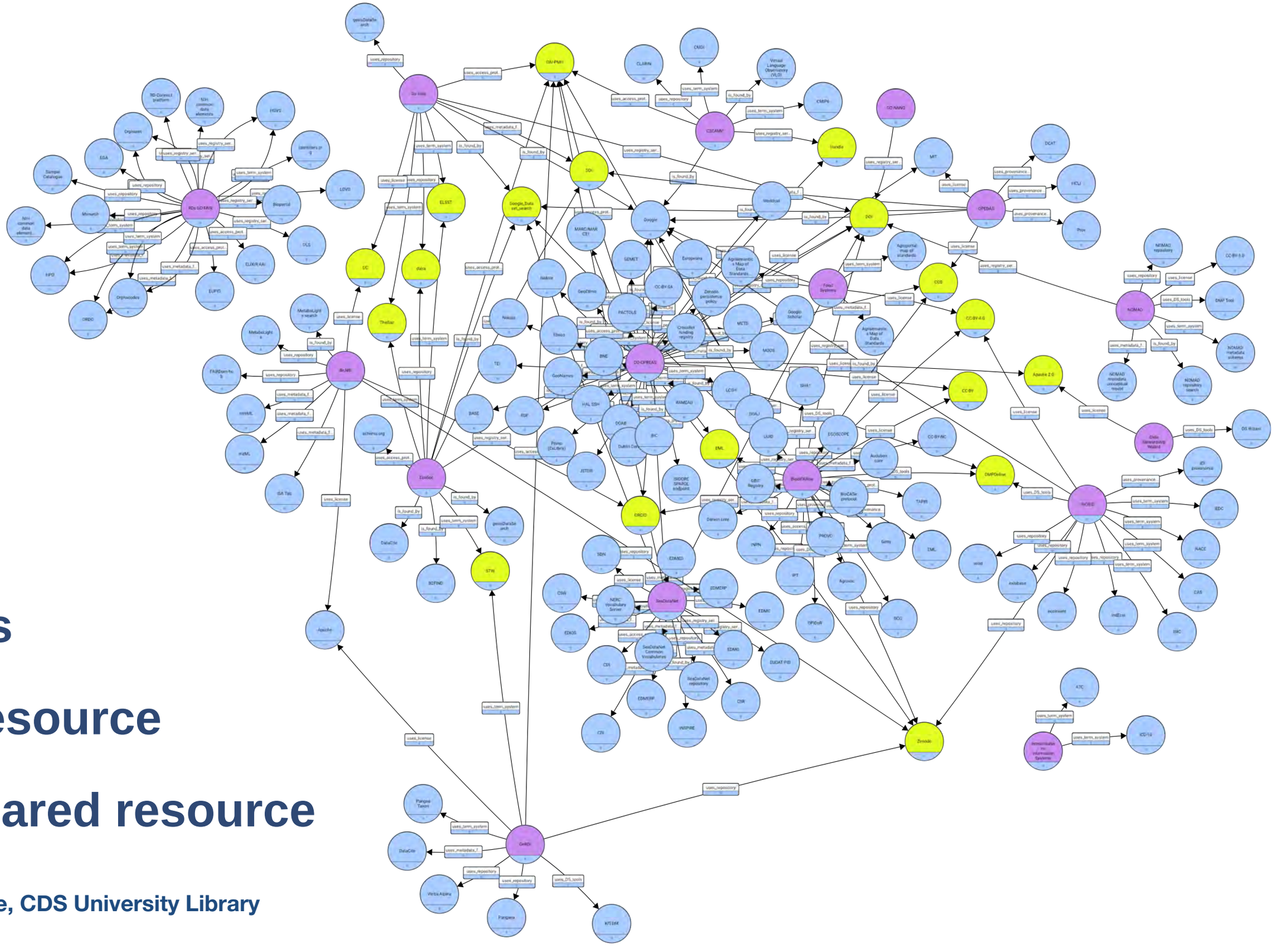
Overview of the FAIR Implementation Choices and Challenges Model (Draft)

Diagram



- PPP services providers

Community Implementation Choices & Challenges



INs

Resource

Shared resource

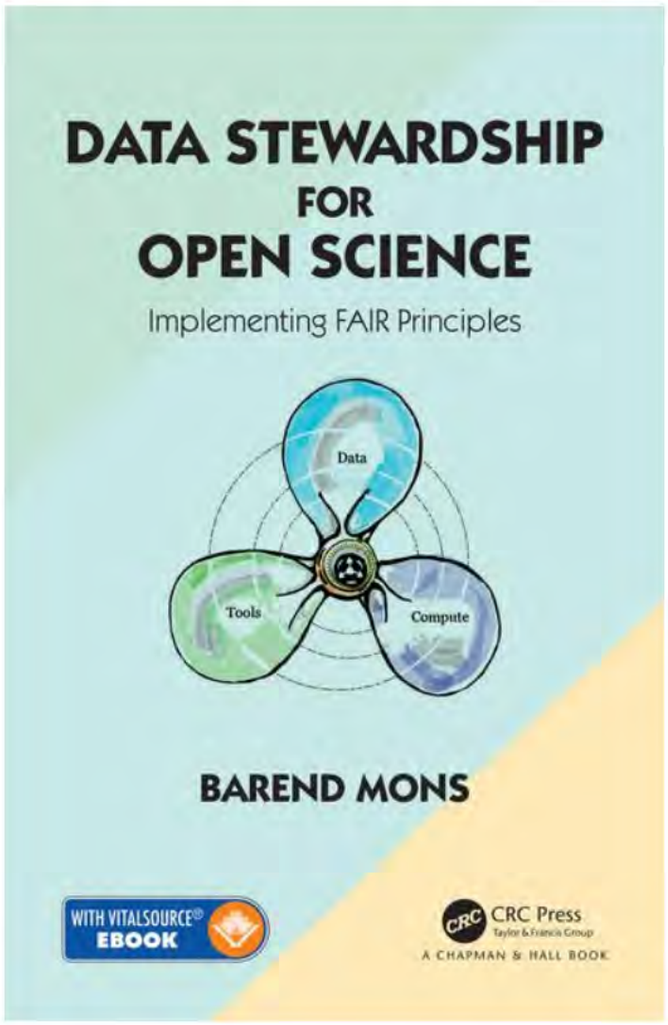
Kristina Hettne, CDS University Library

DATA Intelligence: First generation FAIR implementation choices and challenges

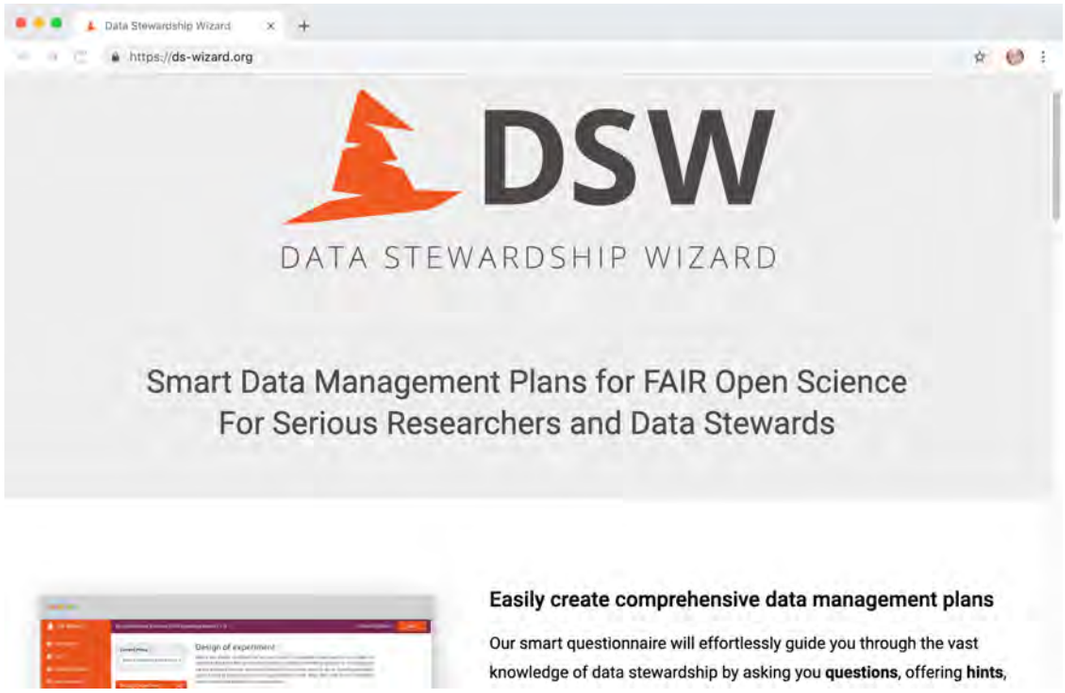
CRC Press
Taylor & Francis Group

About Us ▾ Resources ▾ Tex

Home / Computer Science & Engineering / Da



Preview this Book



Data Intelligence
special issue
30 + articles
September 2019

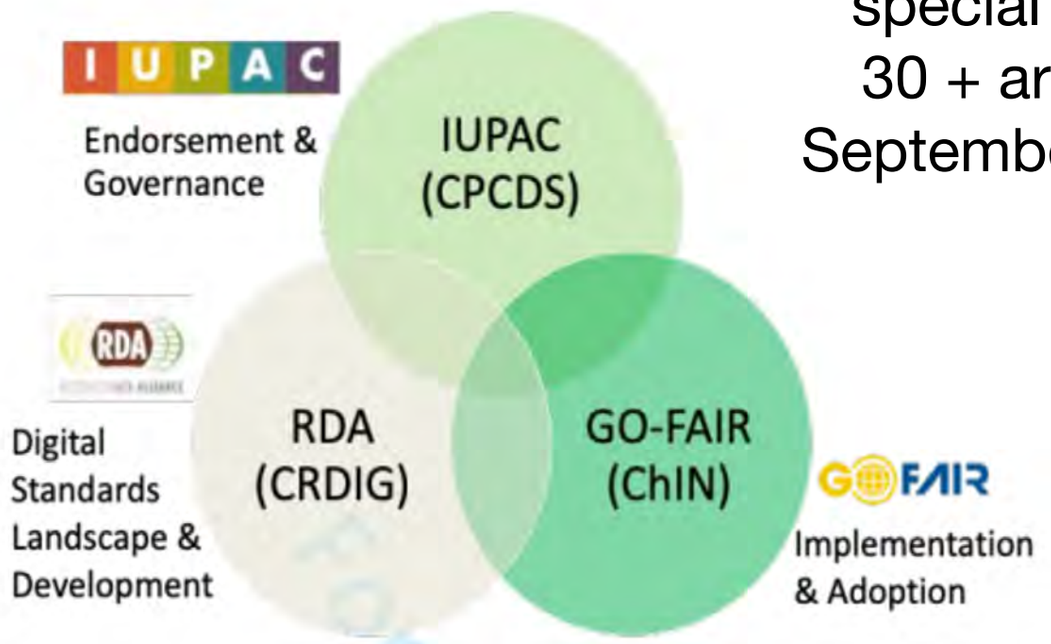


Figure 1: The interoperation between the ChIN and key community-leading organisations.

The seven capital sins of Open Science

1: Age factor….Reward only narrative.com

2: Ignore complexity and existing data

3: Disrespect other disciplines

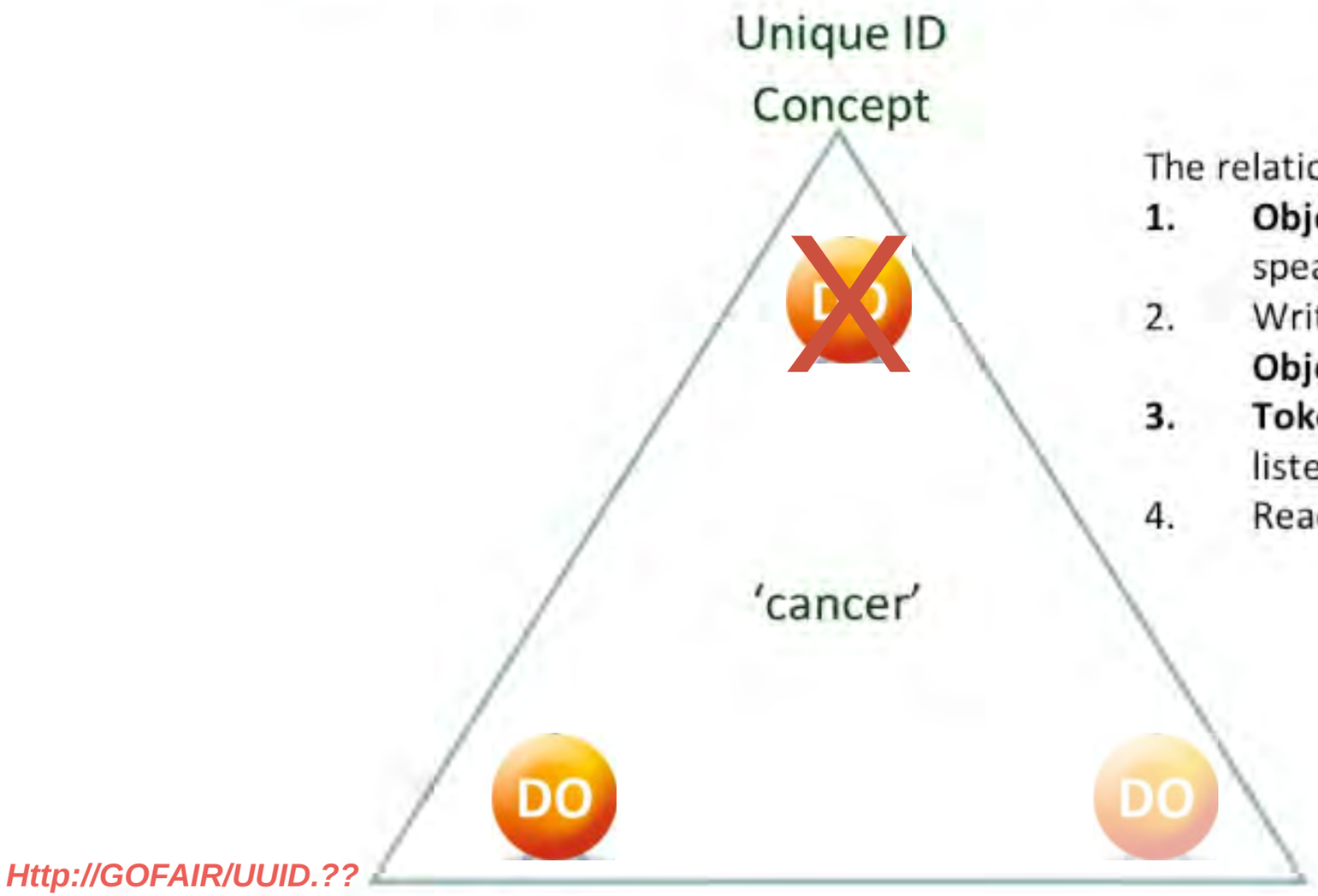
4: publish data without a supplementary paper

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:

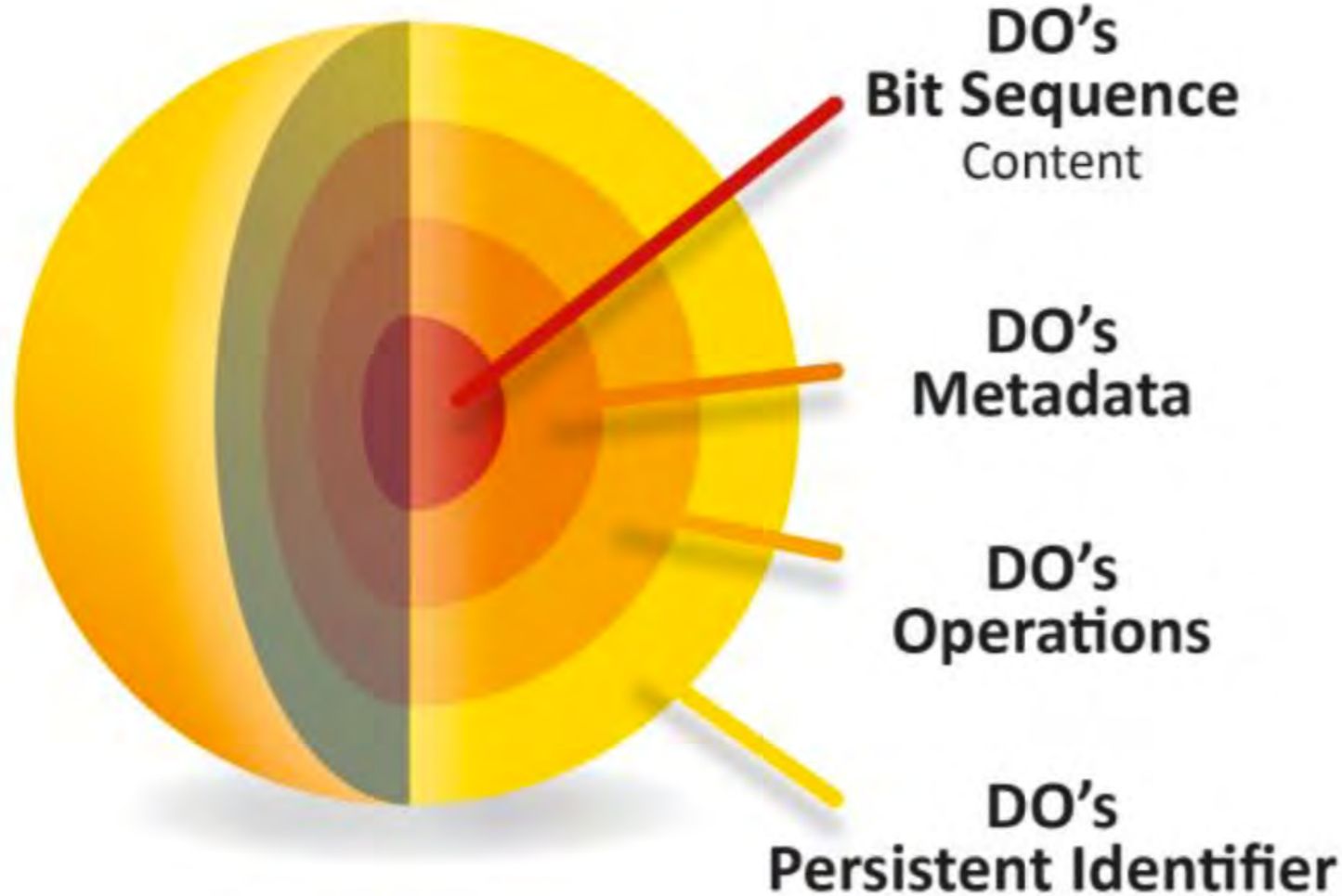
1. **Object** evokes **Concept** (in writer's or speaker's mind)
2. Writer/speaker uses **Token** to refer to **Object**
3. **Token** evokes **Concept** (in reader's or listener's mind)
4. Reader/listener refers **Token** back to **Object**

Token or word or icon:

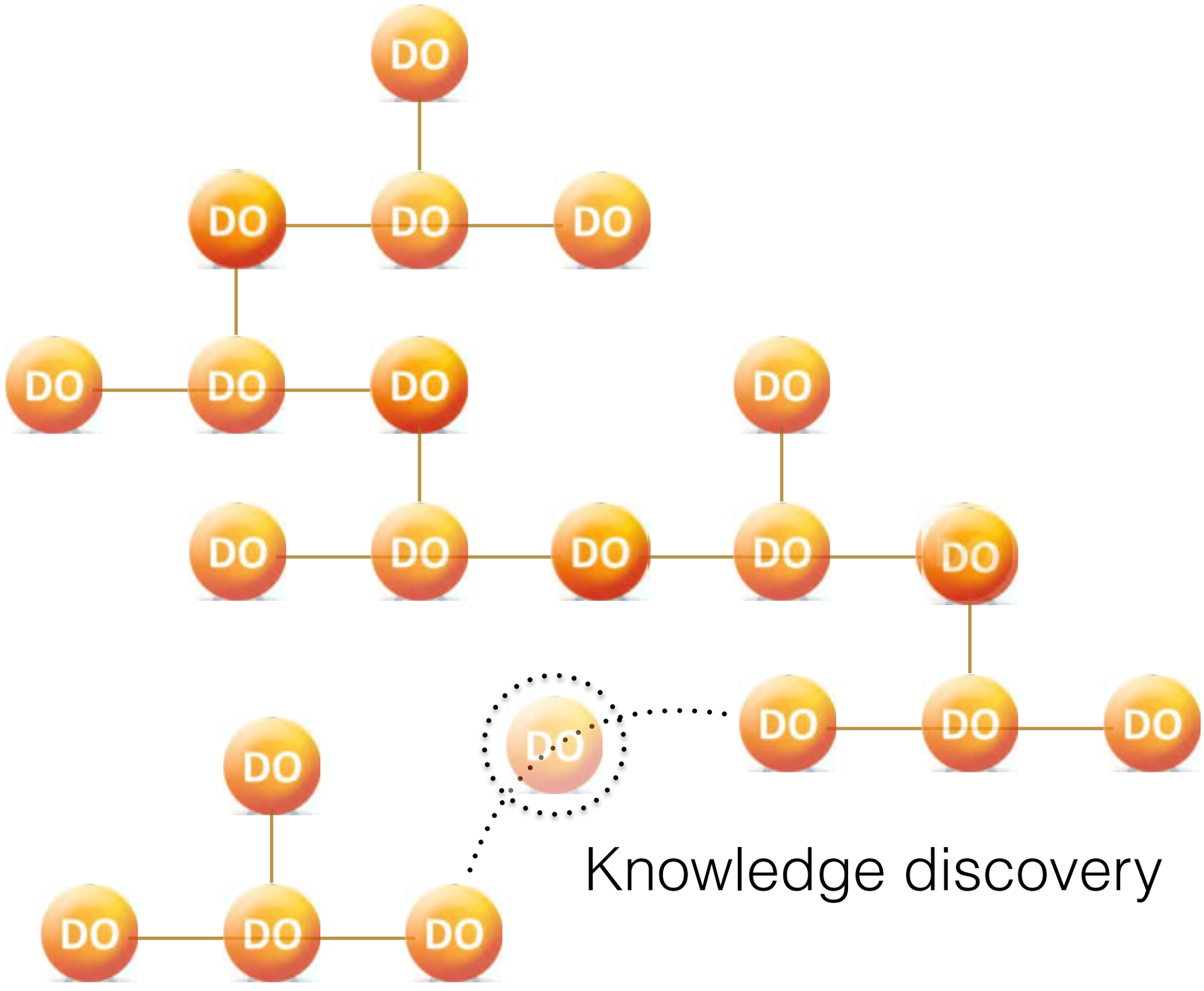
'cancer'
Malignant Neoplasms
Krebskrankheit
C0-265
Etc…

object, entity, defined meaning

could be a DO (f.i. Digital Twin of object)


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

SPO tripples as collections of connected DO’s



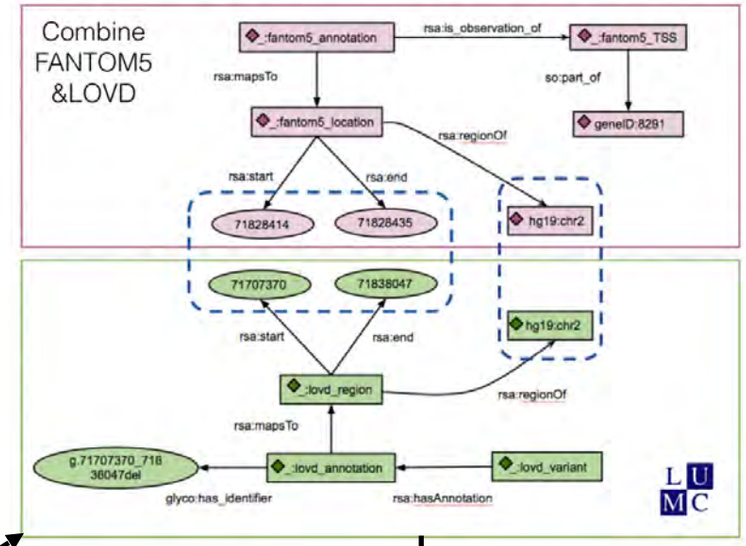
Search

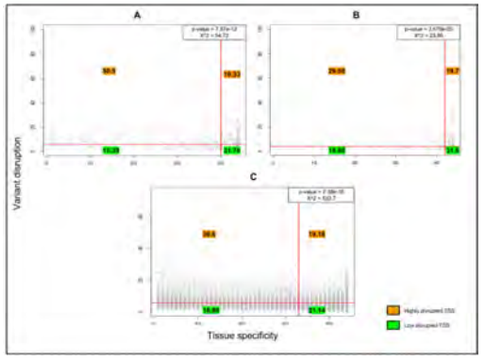
Variant-
TTS-
protein-
Pathology

get_variants_overlapping_with_tss.spar ql

LOVD
Leiden Open Variation Database

Combine
FANTOM5
&LOVD



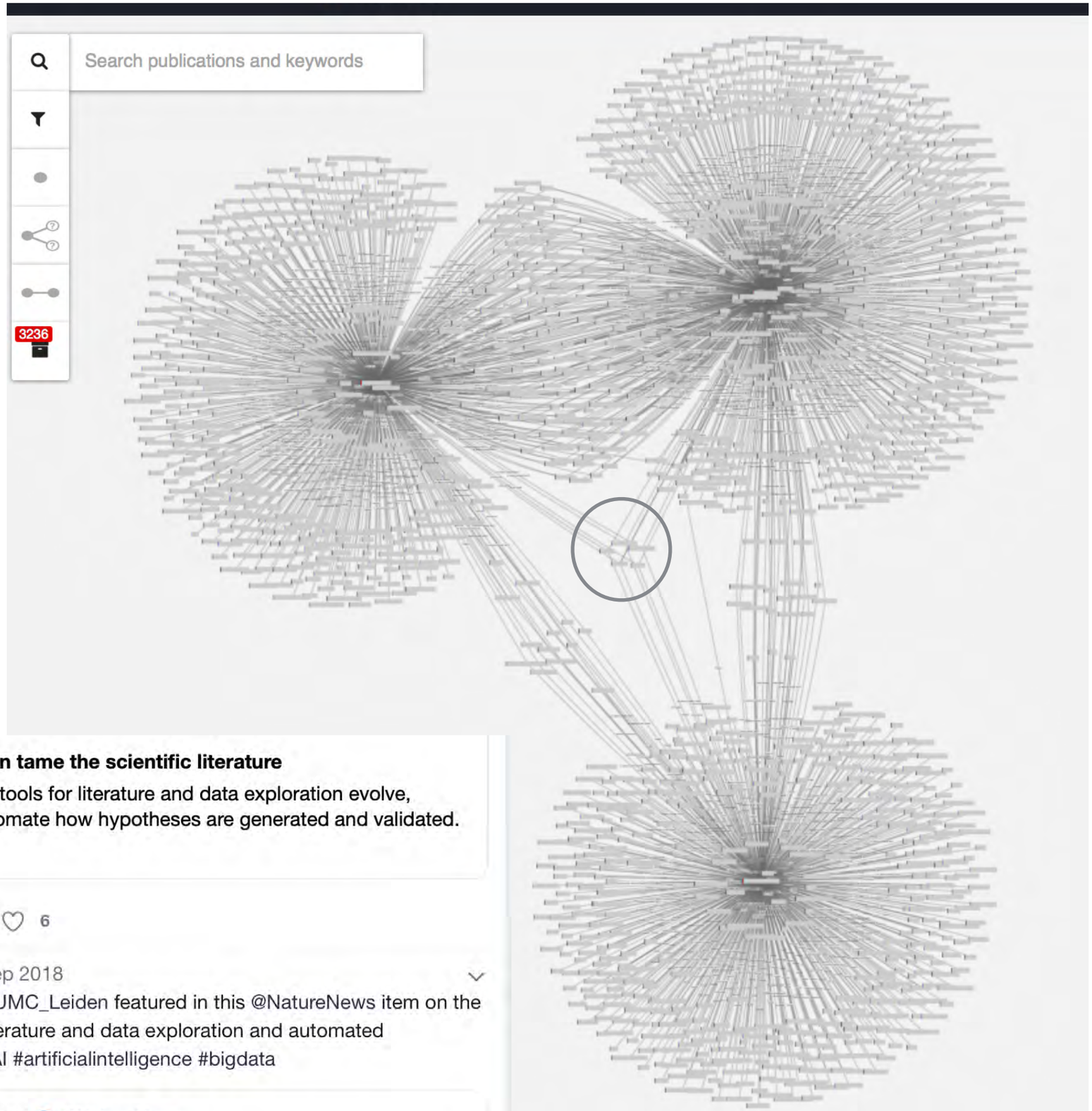


similarity
measure

FANTOM5
FUNCTIONAL ANNOTATION OF THE MAMMALIAN GENOME

Disease Causing Variants

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

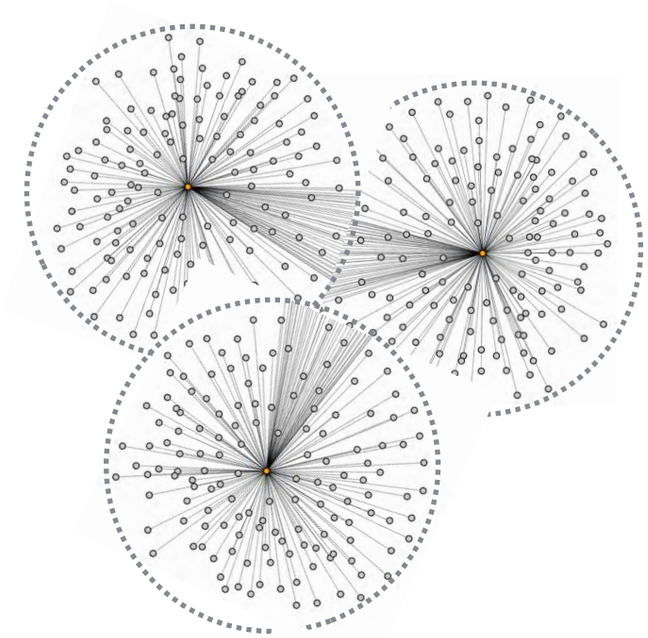
Nature News & Comment @NatureNews

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

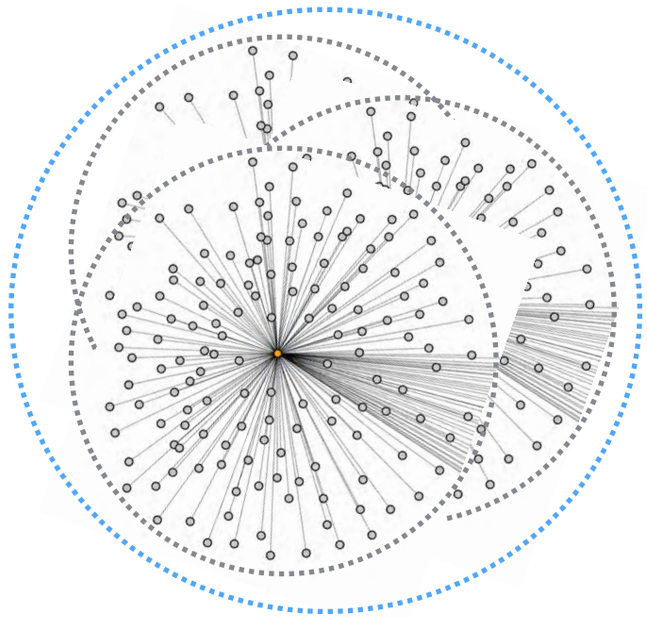
A

conceptual similarity
(hypothesis generation)



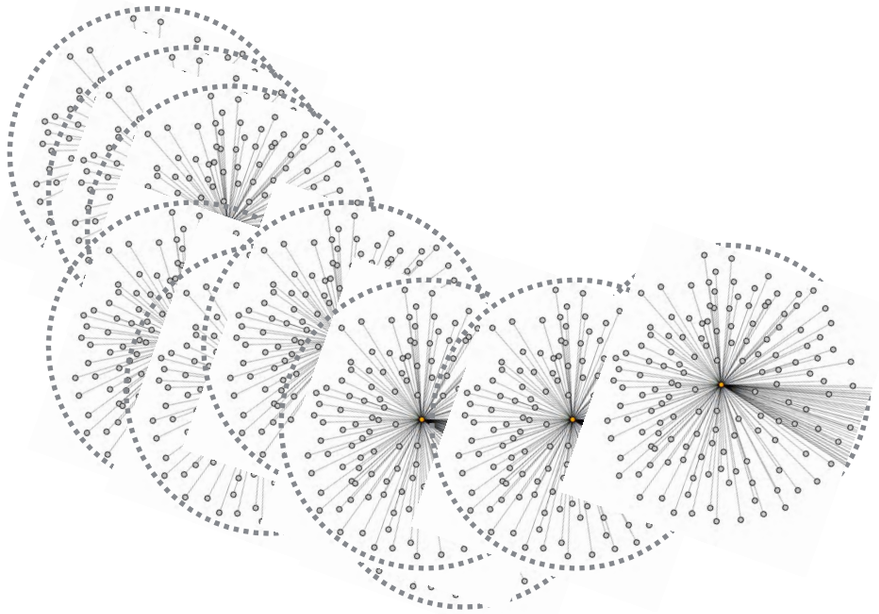
B

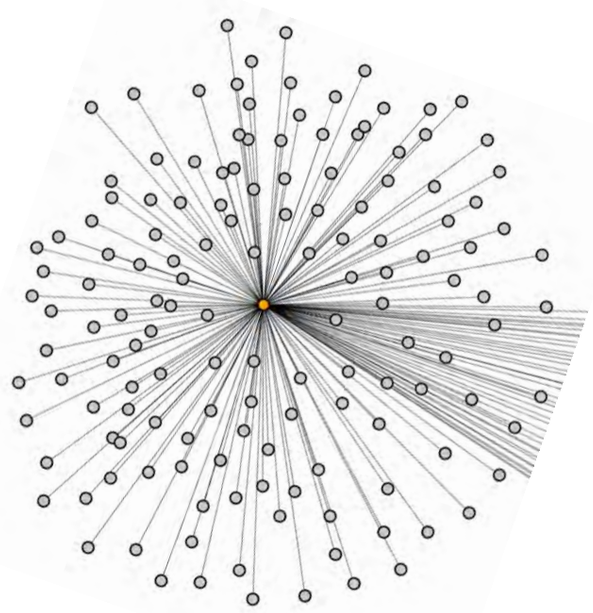
Near sameness
(semantic lenses)



C

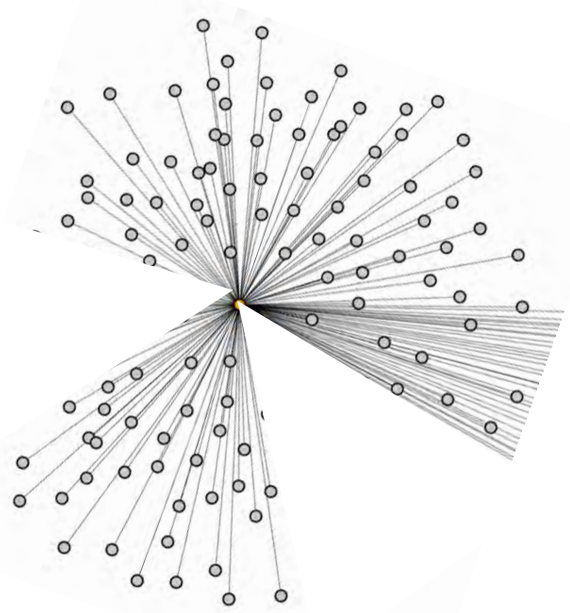
conceptual drift
(meta-data/blockchain)





D

QUA's
(semantic bias)



Thank you!

and

'see you at your data'