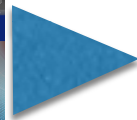
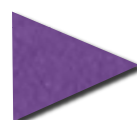




Barend Mons



Biosemanantics Group LUMC and EMC



LS integrator Netherlands eScience Center



Chair of DTL-data
Head of ELIXIR node NL



EC member of Open PHACTS



Chair of High Level Expert Group EOSC



The opinions expressed in this presentation are my personal opinions and do not necessarily reflect the draft report of the High Level Expert Group for the European Open Science Cloud.

Open Science as a Social Machine

Where (the.....) are the Data?

analysis, and dissemination of research data products. The 'map' below was used to start conversations and build more nuanced understandings with investigators.

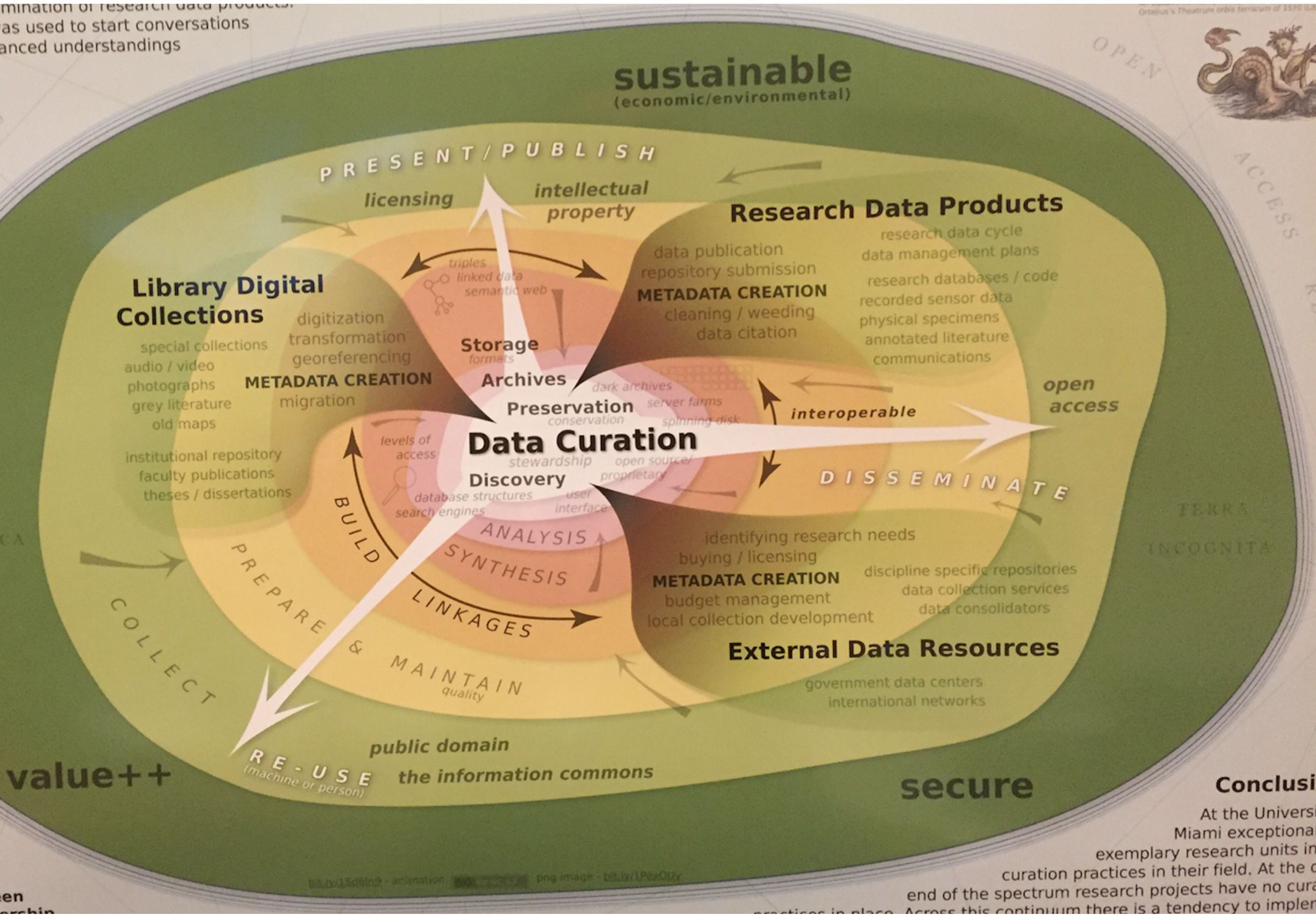
Compass card
Political Almanac



OF UNICORNS

TERRA
BIBLIOTHECA

Tensions between
Value and Ownership



bit (v/15d6in9 - animation) png image - bit.ly/1PexOUy

2 min.
of

Lamenting

: Data, the new currency >>>>>



Neelie Kroes ([@NeelieKroesEU](https://twitter.com/NeelieKroesEU))

16-03-12 14:25

'Data is the new oil': I urge [@ePSIplatform](https://twitter.com/ePSIplatform) conference to go out & make case for [#opendata](https://twitter.com/#!/opendata)
youtu.be/9Jq4Qy1UeAE

Neelie Kroes (then Vice-President of the European Commission, responsible for the “Digital Agenda” for the European union)
When she announced the EU’s Open Data Strategy she opened with “**Data is the New Gold**”. **We wish it were that simple.**

The value of data [OIL, GOLD, CURRENCY, bla,bla]

Nature Genetics, 43, 281–283 (2011)

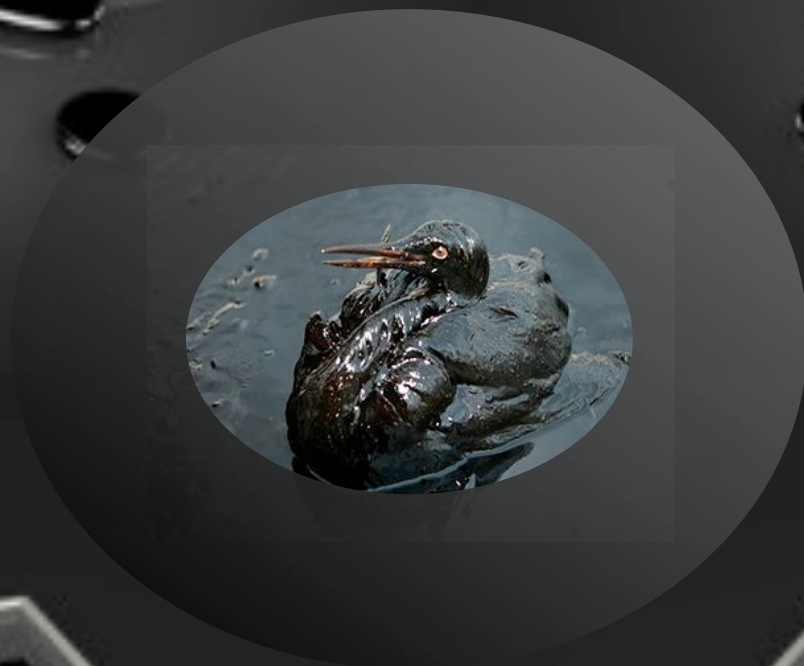
Barend Mons¹⁻⁴, Herman van Haagen¹, Christine Chichester^{2,4}, Peter-Bram ‘t Hoen^{1,4}, Johan T den Dunnen¹, Gertjan van Ommen^{1,4}, Erik van Mulligen^{3,4}, Bharat Singh^{2,3}, Rob Hooft^{2,4}, Marco Roos^{1,2,4}, Joel Hammond⁵, Bruce Kiesel⁵, Belinda Giardine⁶, Jan Velterop^{4,7}, Paul Groth^{4,8} & Erik Schultes^{1,4}

A large, dark silhouette of an oil pumpjack dominates the right side of the frame. The background is a hazy, light-colored sky, suggesting a sunrise or sunset. The pumpjack's long arm is extended upwards and to the right, with a counterweight visible. The overall mood is industrial and contemplative.

DATA

is the new oil

reality

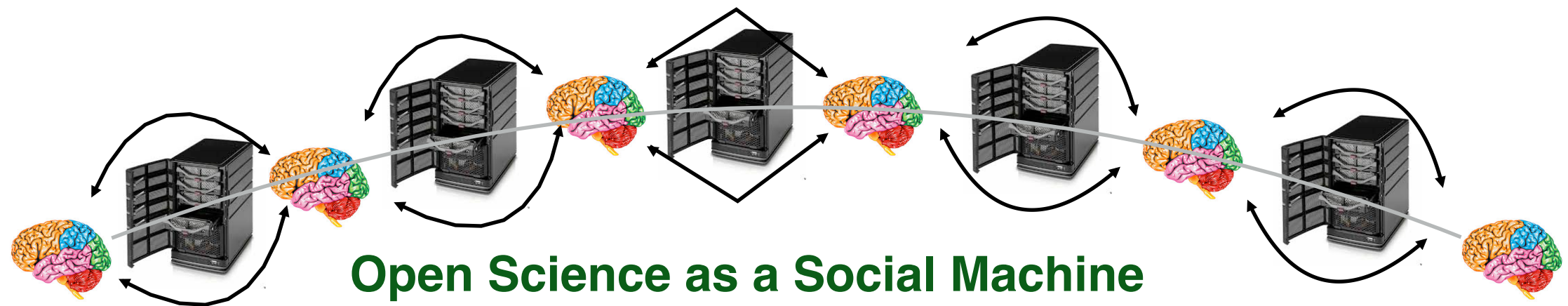


5 min.
of

Contemplation

From Individual Brilliant Minds to Social Machines





WE ARE HERE!...where are you !?

Data Repositories (OA/linked if lucky)

Christmas trees of Hyperlinks

HTML/XML

PDF

Print

The Knowledge Cone

human knowledge bubble:

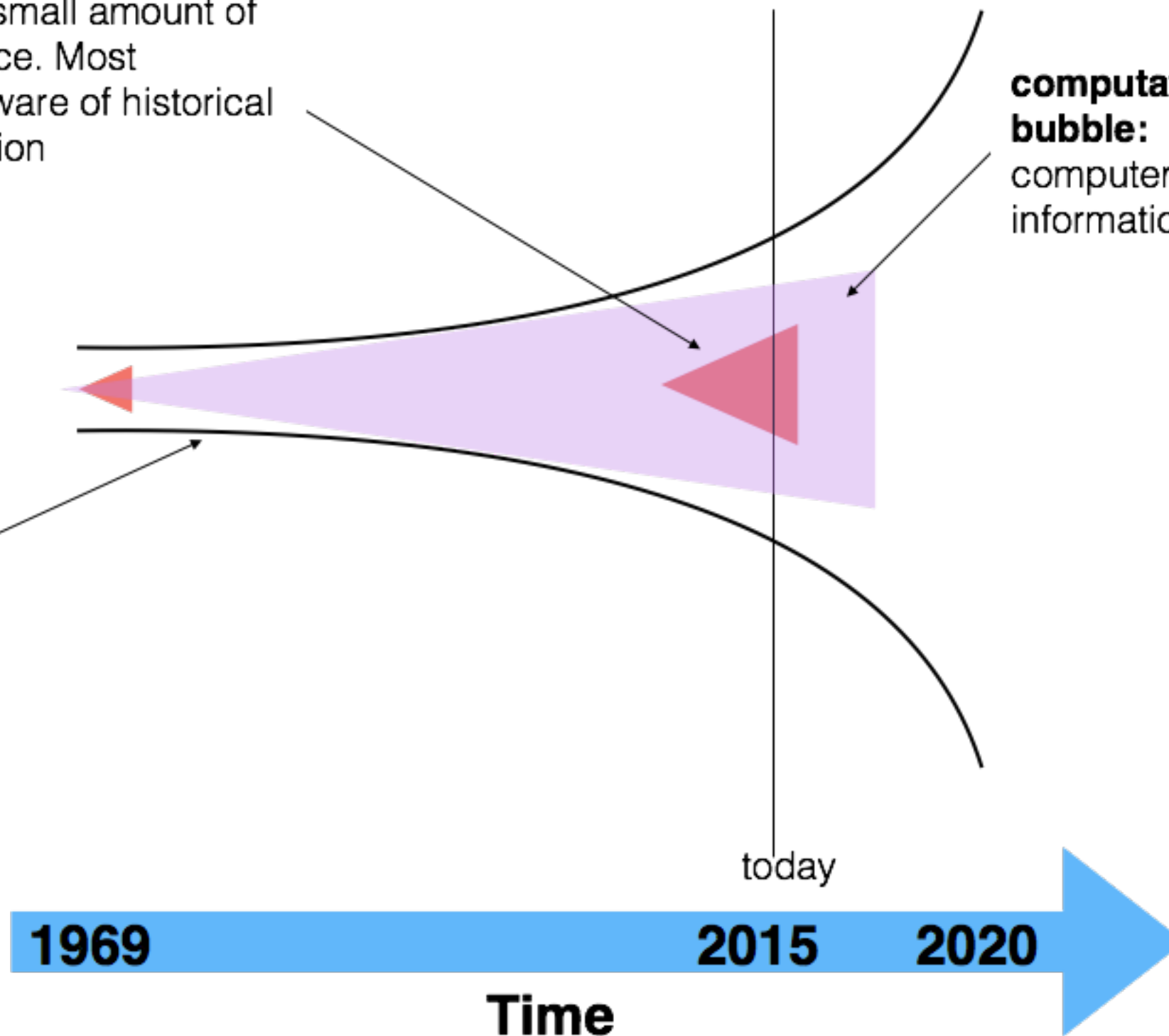
a wave of current trends in the field that factors in only small amount of the knowledge space. Most scientists are not aware of historical and lateral information

computational knowledge bubble:

computer can expand information awareness

knowledge space:

boundaries of biomedical knowledge (exponential growth)



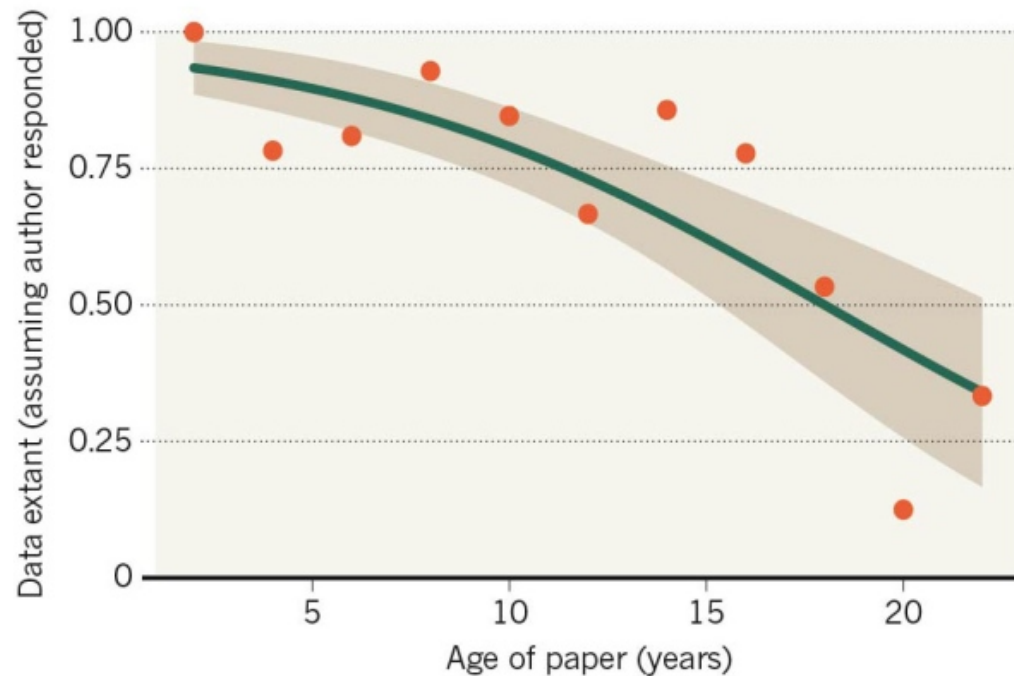
10 min.
of

Reality

Data loss is real and significant, while data growth is staggering

MISSING DATA

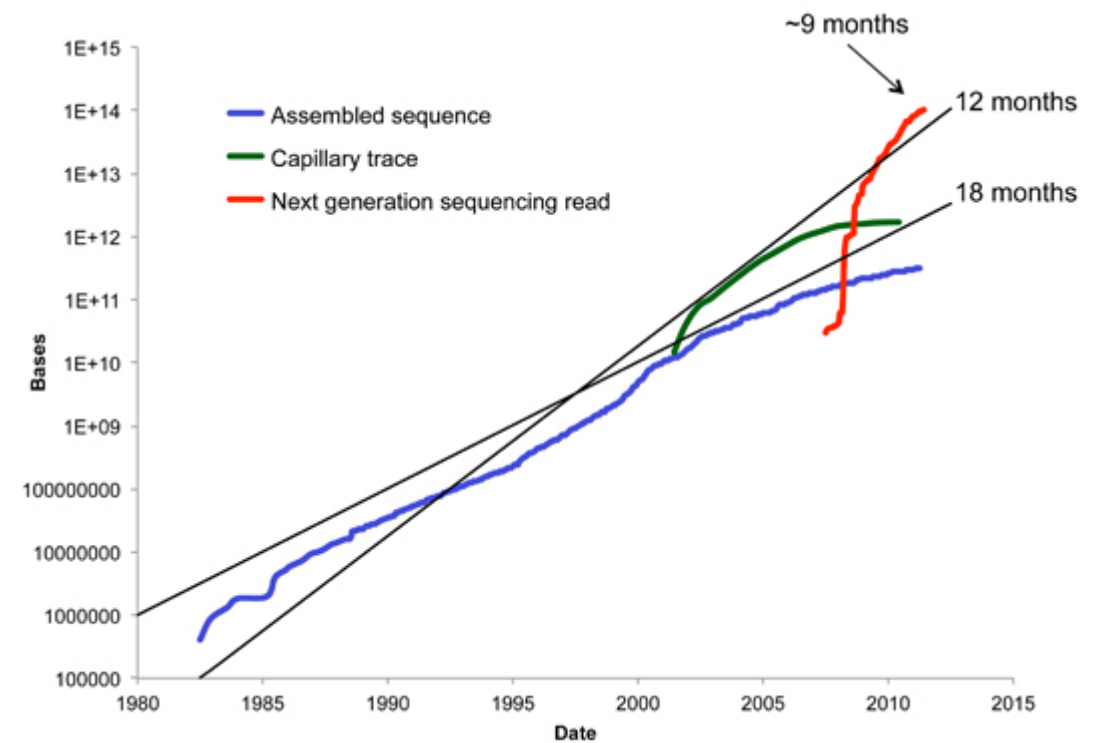
As research articles age, the odds of their raw data being extant drop dramatically.



Nature news, 19 December 2013



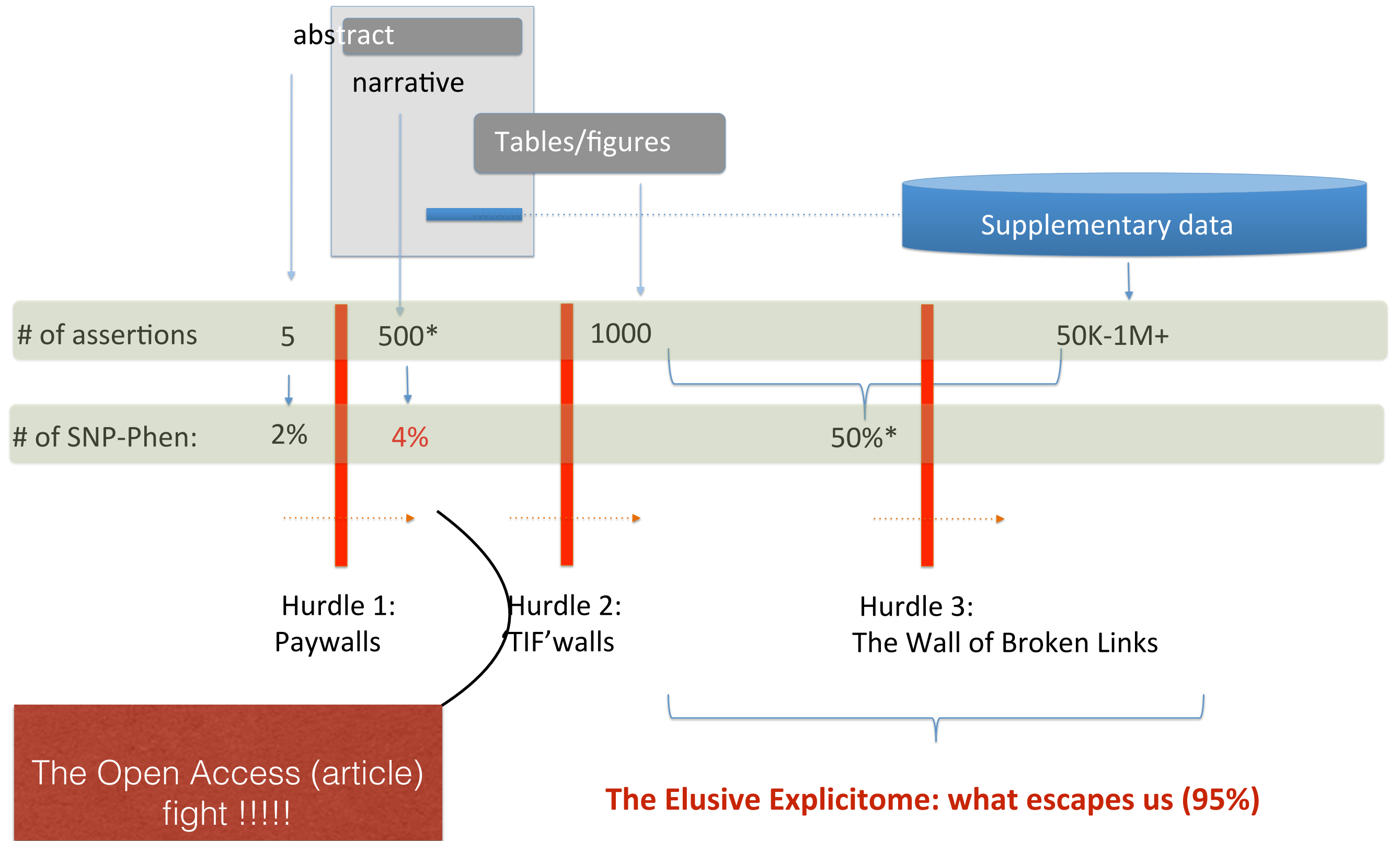
'Oops, that link was the laptop of my PhD student'



- Computer speed and storage capacity is **doubling every 18 months** and this rate is steady
- DNA sequence data is **doubling every 6-8 months** over the last 3 years and looks to continue for this

Current scholarly publishing and the Elusive Explicitome Phenomenon

example from: & Verspoor 2013



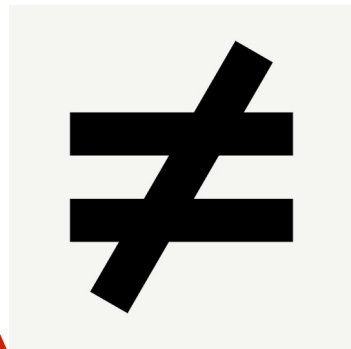
The **EXPLICITOME**

Everything we have 'claimed' in science

Estimate today (LS) : **10^{14} associations.....**

The **EXPLICITOME**

Open Science

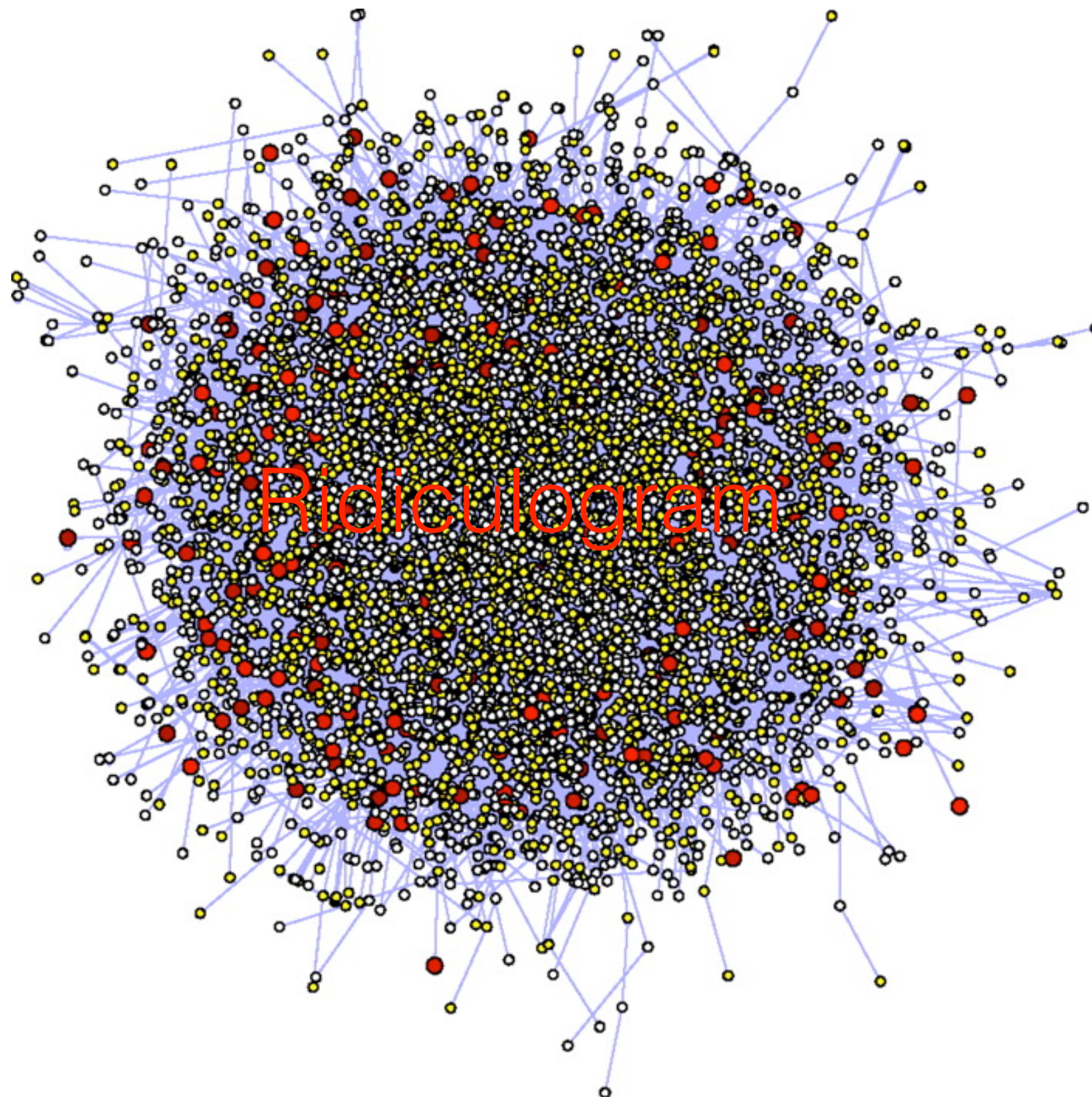


OA articles

5 min.
of

Open Science

Simplified e-Science



FAIR for computers

FAIR for people



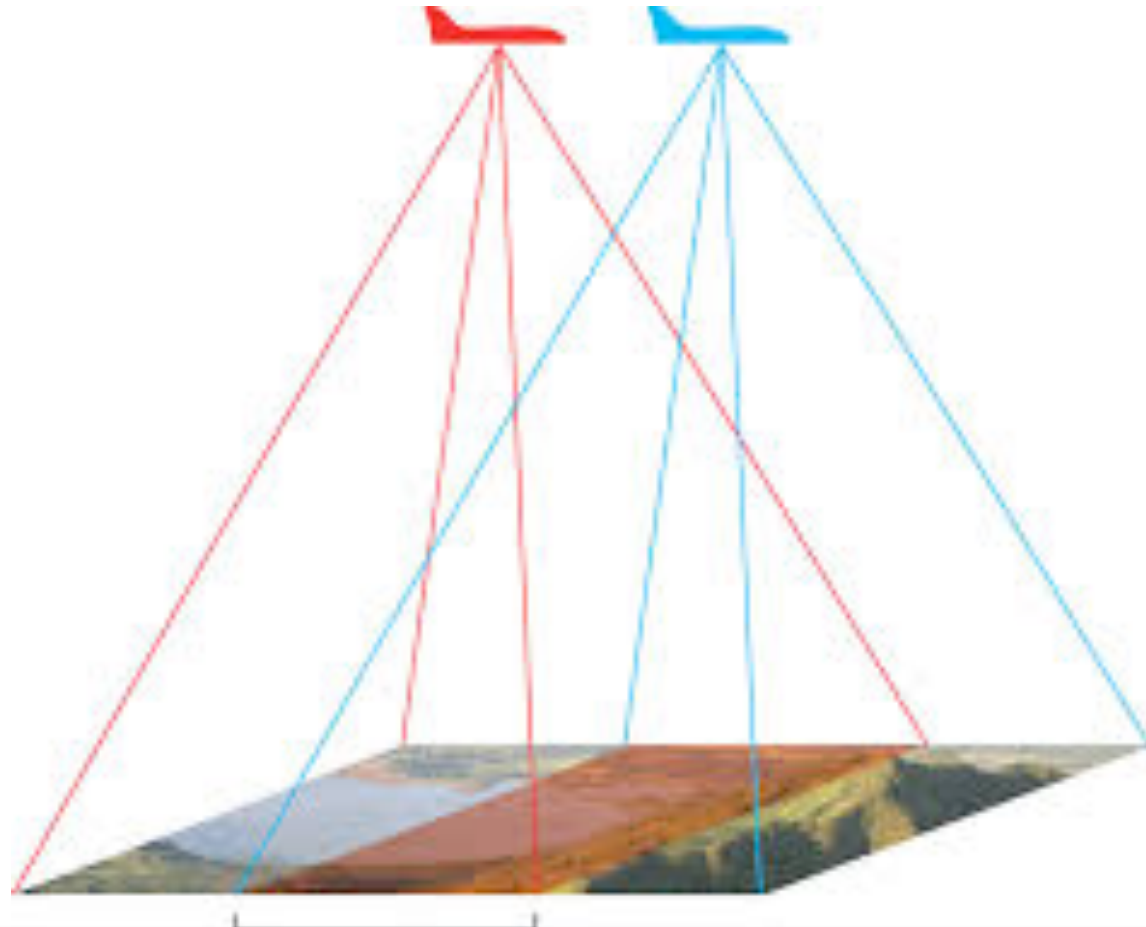
AERIAL SURVEY
pattern recognition in
Ridiculograms



HUMAN EXCAVATION
rationalisation and
'confirmational reading'

'Why would I believe this association'???

The Explicitome is spread over **Thousands** of databases

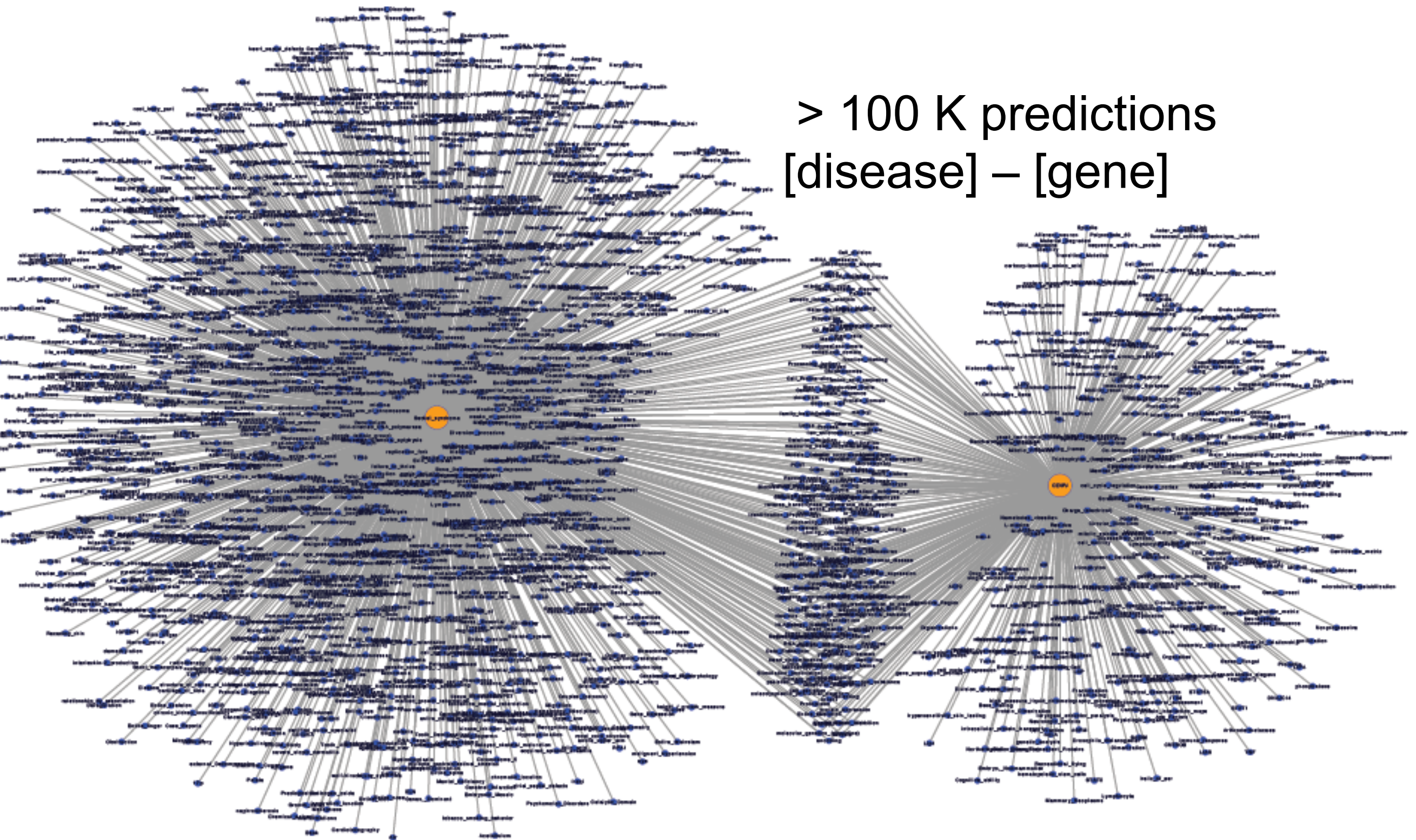


The Explicitome is estimated at 10^{14} assertions





The 'Cardinal Explicitome' is estimated to be 'only' 10^{11} assertions

We publish about less than a million LSConcepts !

> 100 K predictions
[disease] – [gene]

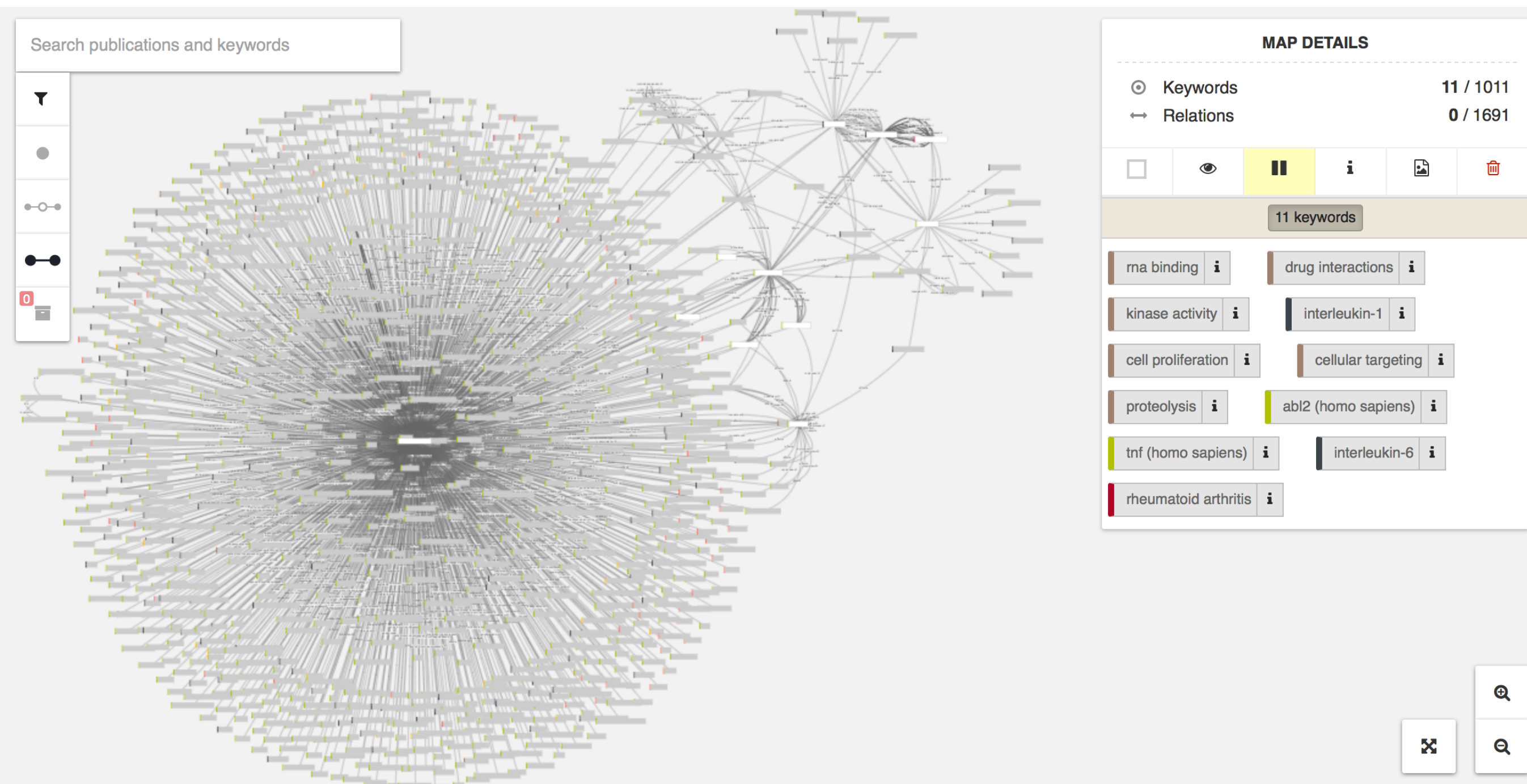




-  **Link over 70 databases**
-  **Discovery of indirect relationships**
-  **Rationalisation (workflows)**
-  **Dig into evidence via provenance**



The Knowlet is mostly connected
to the IL-6/IL-1/TNF/AR via RNA binding, drug interactions
Cell proliferation, cellular targeting, proteolysis and kinase activity





Keywords

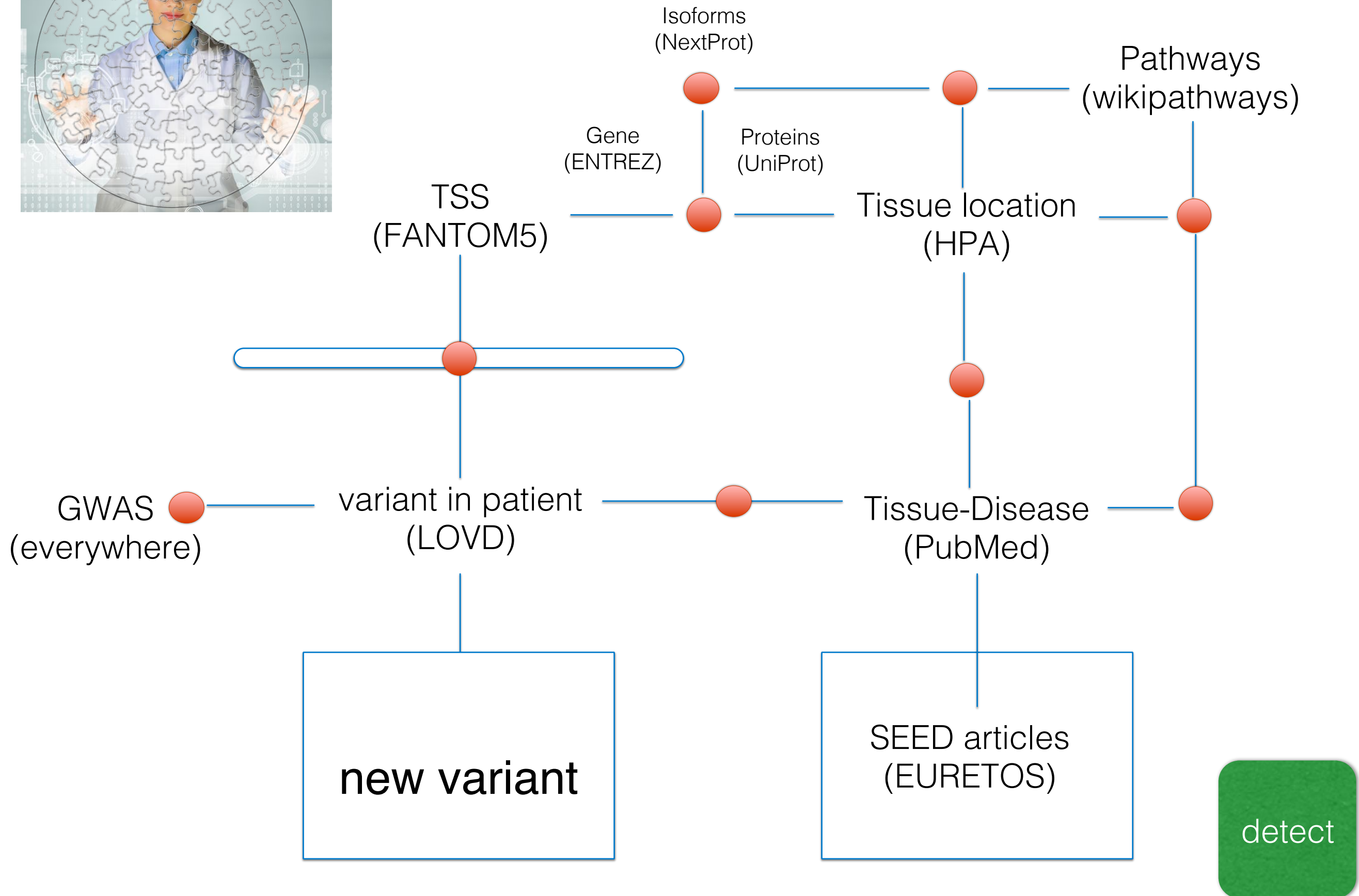
Relations



1 keyword

rheumatoid arthritis







User Decides

encrypted
Query to API

EURETOS

- EURETOS Surfaceome Publication Search 🔒 >
- EURETOS Systematic Literature Review 🔒 🔒 >
- EURETOS Genomic Biomarker Candidates 🔒 🔒 >
- EURETOS Gene Evaluation Workflow 🔒 🔒 >
- EURETOS Variant - Disease Analysis 🔒 🔒 >
- EURETOS Biomarker Workflow 🔒 🔒 >
- Expert Finder 🔒 >
- Utopia Utopia Annotator and Hypothesis 🔒 🔒 >

secure (polymorphic encrypted)
Query



User Decides



EURETOS
Scholar

I ❤️
open
access

annotation

I ❤️
open
access

nanopublication

assertion

provenance

publication info

I ❤️
open
access

EURETOS

CONTENTS:

Summary
Disorder details
Gene details
Gene products
Gene expression
Gene inhibiting molecules
Gene stimulating molecules
Interactions via biological phenomena
Disorders
Anatomy
Macromolecules
Small molecules
Physiology
Interactions via gene affected molecules
References

seed article

Summary

Included gene products and sub-disorders	2
Included sub-disorders	5
Direct Interactions	0
Direct gene disease interactions	0
Strength of biological association (100 is maximum)	87.11
Interactions via biological phenomena	32
Disorders (disease-disease interactions)	41
Anatomy (gene expression)	0
Genes (gene-gene interactions)	64
Macromolecules (including protein-protein interactions)	35
Small molecules (chemical interactions)	39
Interactions via gene affected molecules	7116
Molecular and physiological interactions	7116
Gene inhibiting and stimulating molecules	5
Gene inhibiting molecules	5
Gene stimulating molecules	0

all references in PDF > Lazarus EURETOS-Graph...)

User has subscription to (all) publishers

New S-(p?)-O's



5 min.
of




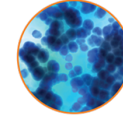
Vision

ELIXIR: An international distributed infrastructure for biological data

Technical platforms

-  Data
-  Standards
-  Tools
-  Compute
-  Training

User communities

-  Marine metagenomics
-  Crop and forest plants
-  Human data
-  Rare diseases



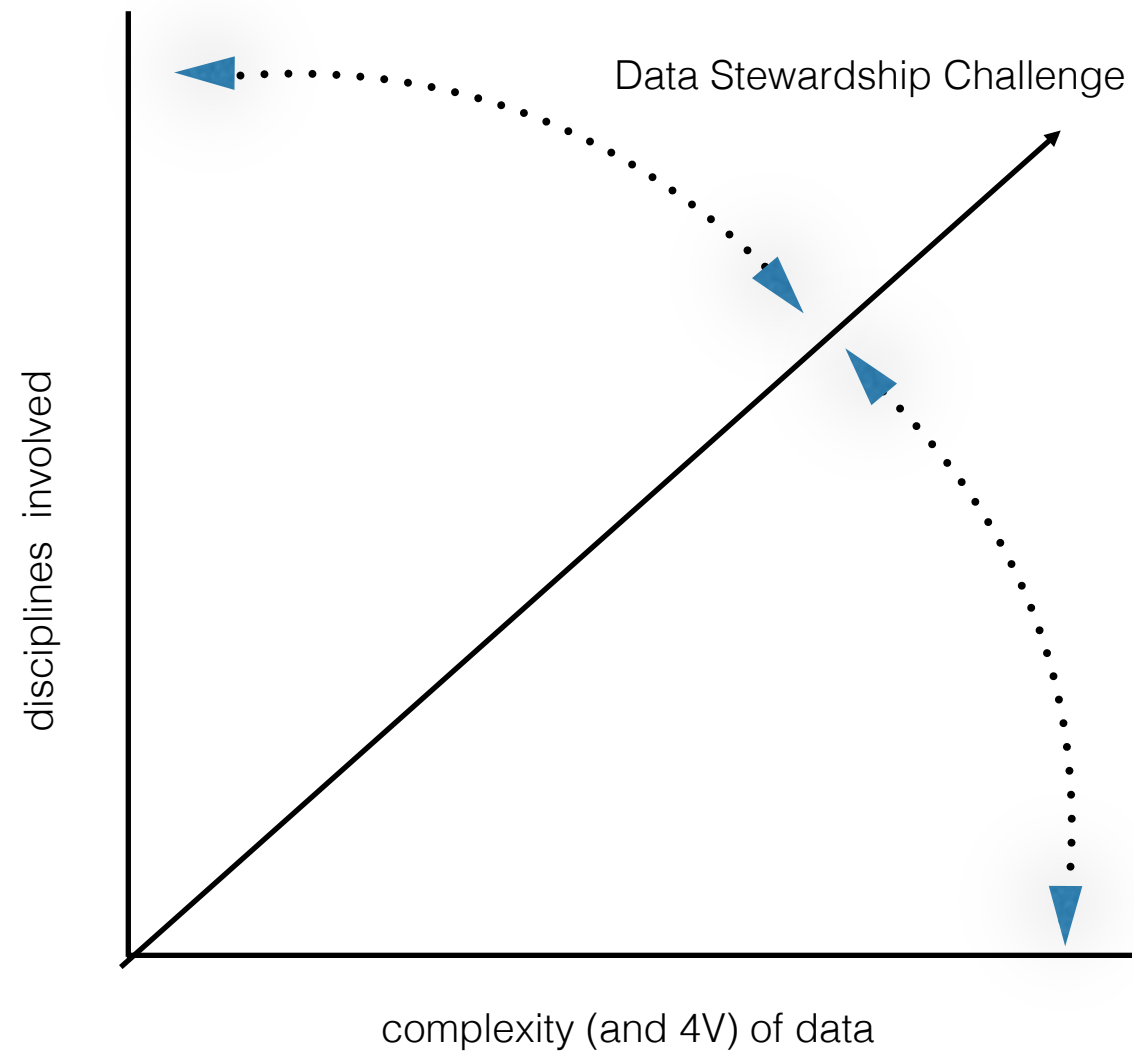
- European
 - Open
 - Science
 - Cloud

- European
- Open Science
 - Cloud

The EOSC

Open Science

Open ≠ (articles)



EOSC: **Framing**

- **Trusted access** to **services** & **systems**
- **Re-use** of shared **data**
- **Across** disciplinary, social and geographical **borders**
- **Federated** environment, across Member States

EOSC: ‘Internet approach’

- **Minimal** international guidance and governance
- **Maximum** freedom to implement.
- **Globally** interoperable and accessible
- **Globally** embedded in a ‘**Commons**’



EOSC: **Scope**

- **Human expertise**
- **Core** resources
- **Standards, best practices**
- **underpinning technical infrastructures**
- **A web of Services**

EOSC: **Supports**

- **Open Science**
- **Open Innovation**
- **Systematic and professional data management**
- **Long term data stewardship**

EOSC: **Challenges and Observations**

- The majority of the challenges are **social** rather than **technical**
- Not just the **size of data**, but in particular **complex data** and **analytics across domains**.
- Shortage of **data experts** globally and in the European Union
- **Archaic system of rewards** and **funding** of science and innovation
- ‘**Valley of death**’ between **(e-)infrastructure providers** and **domain specialists**.
- **Short funding cycles** of **core research infrastructures** are **not fit for purpose**
- **Fragmentation** between domains causes **repetitive** and **isolated** solutions
- Distributed data sets increasingly **do not move** (**size & privacy** reasons)
- Centralised HPC is **insufficient** to support **distributed meta-analysis and learning**.
- However, the **major components** for a **first generation EOSC** are largely ‘there’
- But ‘**lost in fragmentation**’ and spread over 28 Member States.

EOSC: **Key requirements**

- **New modes** of scholarly communication
- **Modern reward** and recognition practices need to support data sharing and re-use
- **Innovative**, fit for purpose **funding schemes** for sustainable underpinning infrastructures
- Core **data experts** need to be trained and their career perspective significantly improved
- Cross-disciplinary **collaboration-specific measures** for review, funding and infrastructure
- Support for the transition from **scientific insights** towards **societal innovation**
- The EOSC needs to be developed as an **eco-system of infrastructures**
- Key Performance Indicators should be developed for the EOSC
- The EOSC should **enable automation of data processing** and thus **machine actionability** is key.
- FAIR principles

EOSC: **Policy Recommendations**

- P1: Take immediate, affirmative action in close concert with Member States
- P2: Close discussions about the ‘perceived need’
- P3: Build on existing capacity and expertise where possible
- P4: Frame the EOSC as supporting Internet based protocols & applications

EOSC: **Governance Recommendations**

- G1: Aim at the lightest possible, internationally effective governance
- G2: Guidance only where guidance is due
- G3: Define Rules of Engagement for formal participation in the EOSC
- G4: Federate the Gems across Member States

EOSC: **Implementation Recommendations**

- I1: Turn this report into an EC approved White Paper to guide EOSC initiative
- I2: Develop, Endorse and implement a Rules of Engagement scheme
- I3: Fund a concentrated effort to locate and develop Data Expertise in Europe
- I4: Install a highly innovative guided funding scheme for the preparatory phase
- I5: Make adequate data stewardship mandatory for all research proposals
- I6: Install an executive team to deal with international coherence of the EOSC
- I7: Install an executive team to deal with the preparatory phase of the EOSC

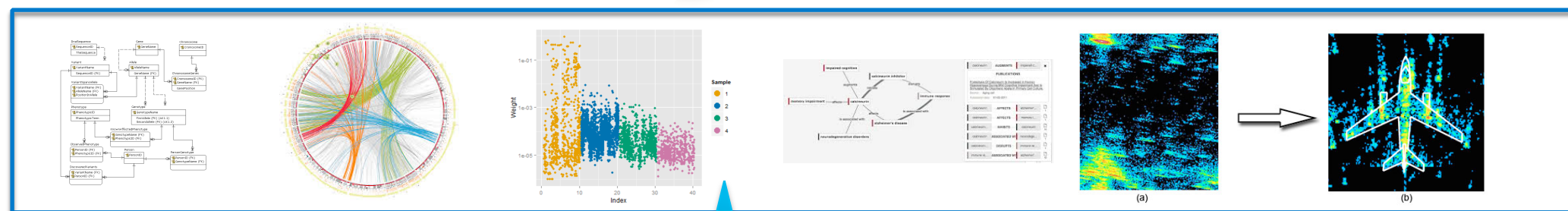


confirmational reading



Actionable Knowledge

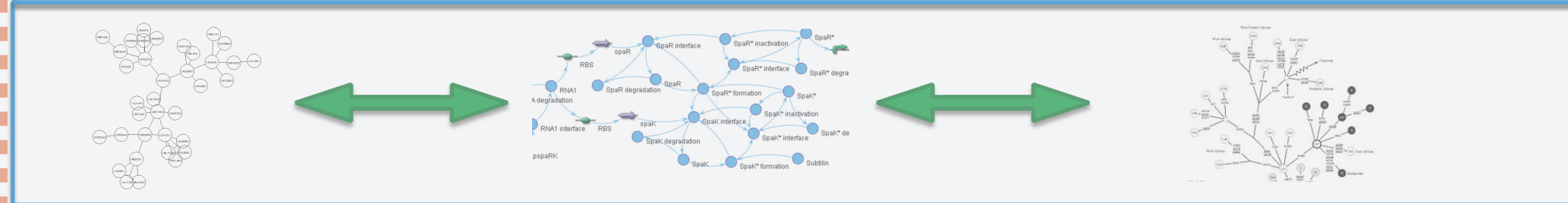
Pattern
recognition



Analysis transformation

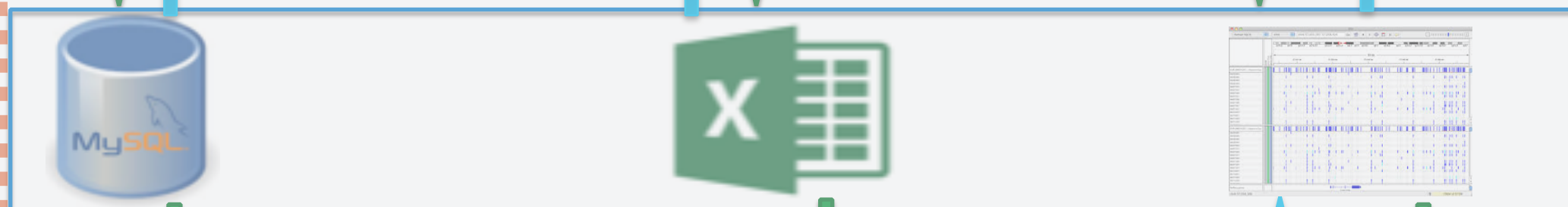
(mostly interlinking and optimisation)

EOSC API's



FAIR transformation

FAIR download
(in local format)



provenance

provenance



Hard infrastructure and repositories

FAIR (meta)data
(RDF, XML etc.)



processed data
(any storage format)

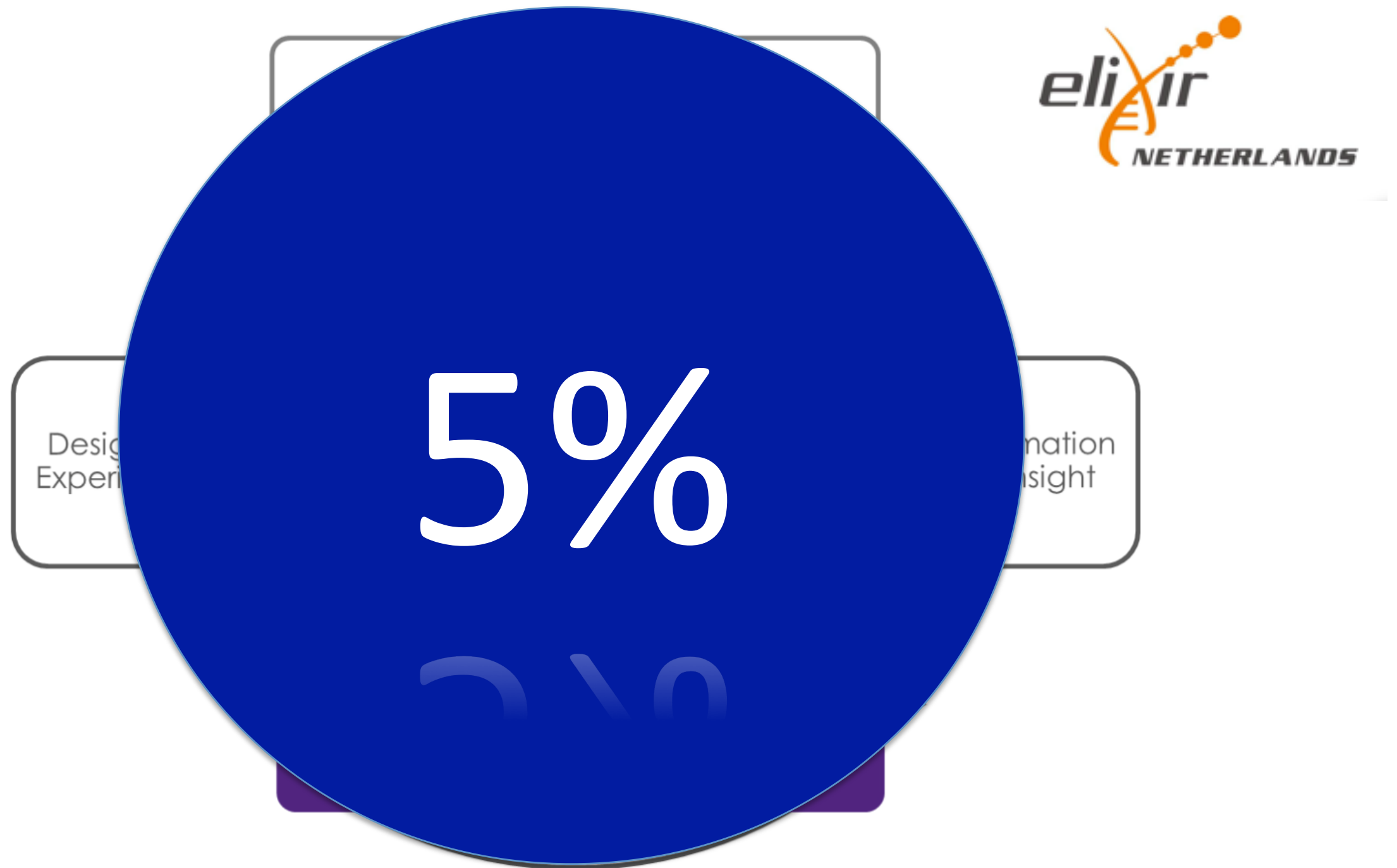


Raw data
(many formats)

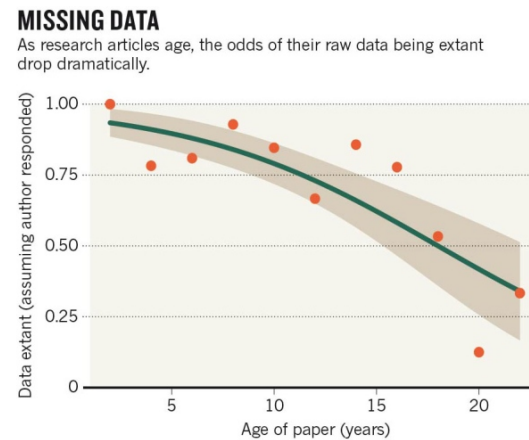
5 min.
of

Conclusions

The Data Stewardship Cycle



Malpractices.....



‘supplementary data’

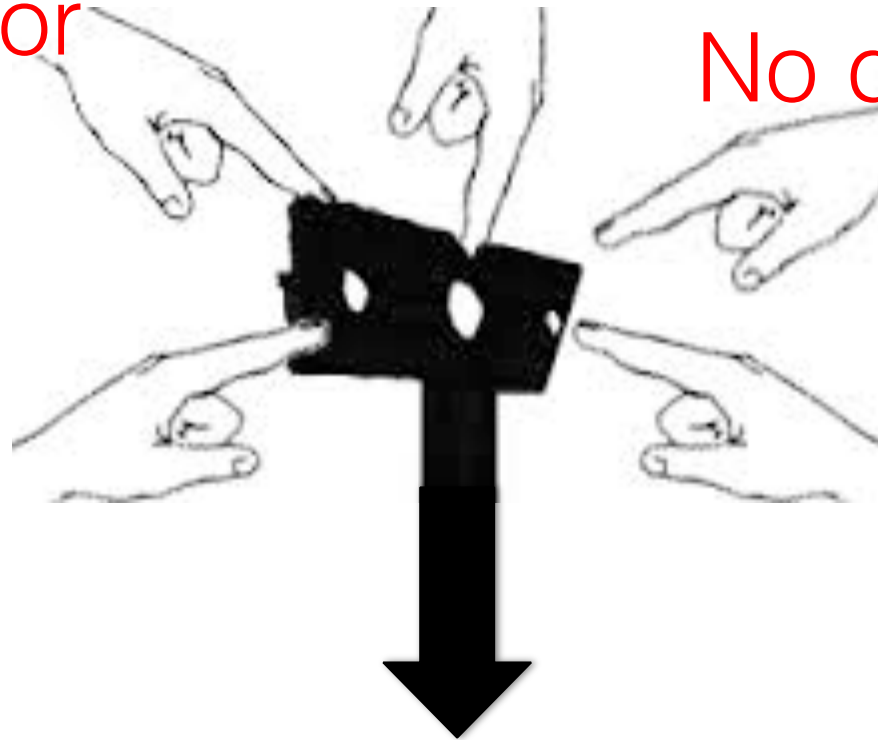
Journal Impact Factor

No data stewardship plan

Ignore Altmetrics

Obstruct Tenure
Data Experts

Knowledge Sharing Impaired



YOU ARE HERE



CWA

Open PHACTS

ODEX4all

FDG

FAIRdom, CEDAR, EUCAT, PHT, FDG, etc. etc

FORCE11, JDDCP, FAIR

EXCOR

EOSC

ELIXIR

2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020

95% MS

5% EC

Data Management Plans

Mandatory for Research Projects H2020 & Member States

Long-Term Data Stewardship

How to finance ESFRI's and EBI SIB type + infra
Mainly private for reliability

Interoperability Backbones, Standards, Procedures

Mainly H2020 + ESFRI-type domain expertise



2.51 min.

of

PHT

The Personal Health Train



