



Good FAIR Practices

Erik Schultes, Barbara Magagna, Andrea Tarallo

4 December 2024

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-
Mission 4 “Education and Research” - Component 2: “From research to business” - Investment
3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”





Welcome - Recap

13:00-13:10
(Erik & Barbara)

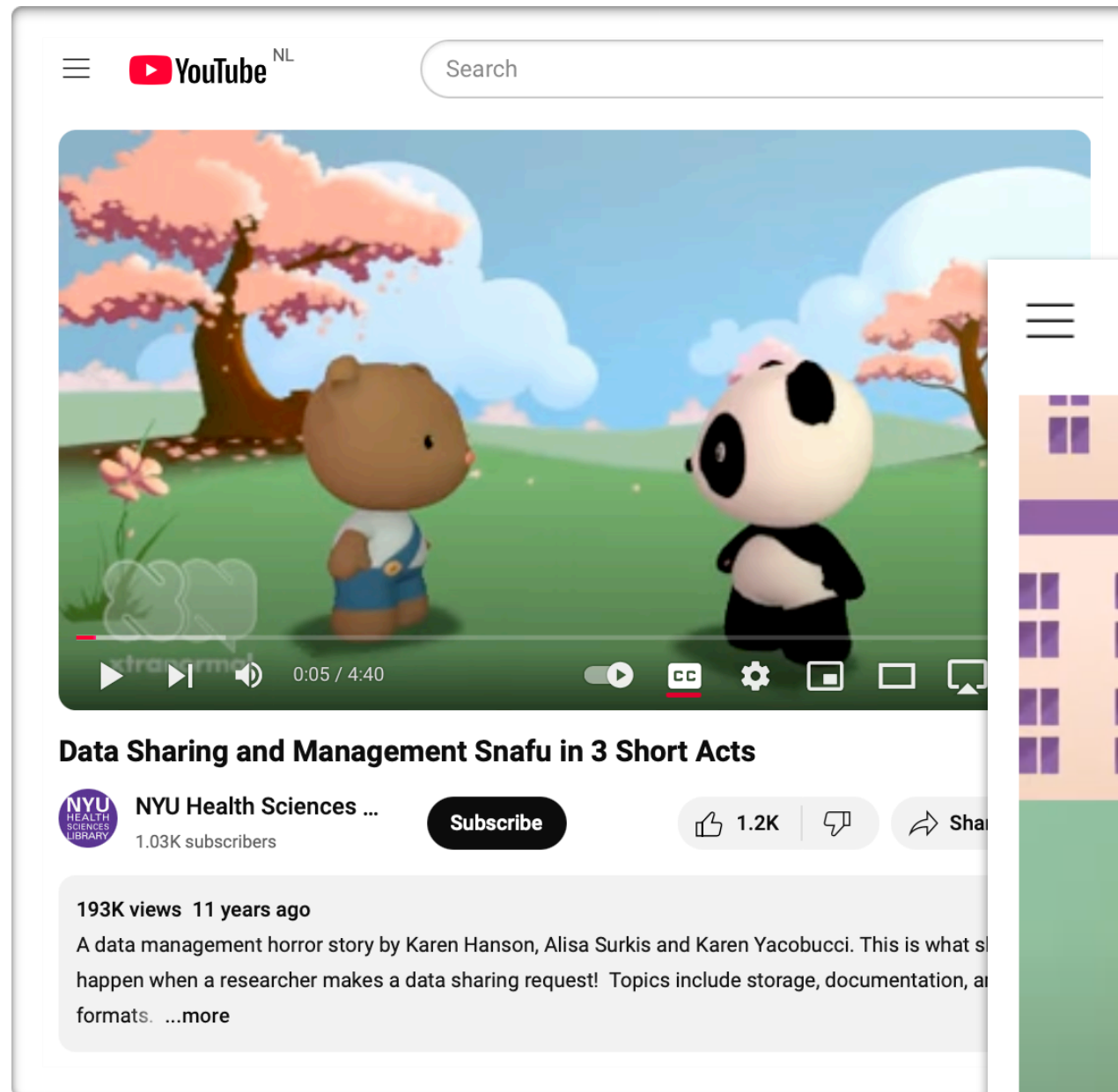


The FAIR Principles

Erik Schultes, Barbara Magagna, Andrea Tarallo
29 November 2024

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-
Mission 4 "Education and Research" - Component 2: "From research to business" - Investment
3.1: "Fund for the realisation of an integrated system of research and innovation infrastructures"



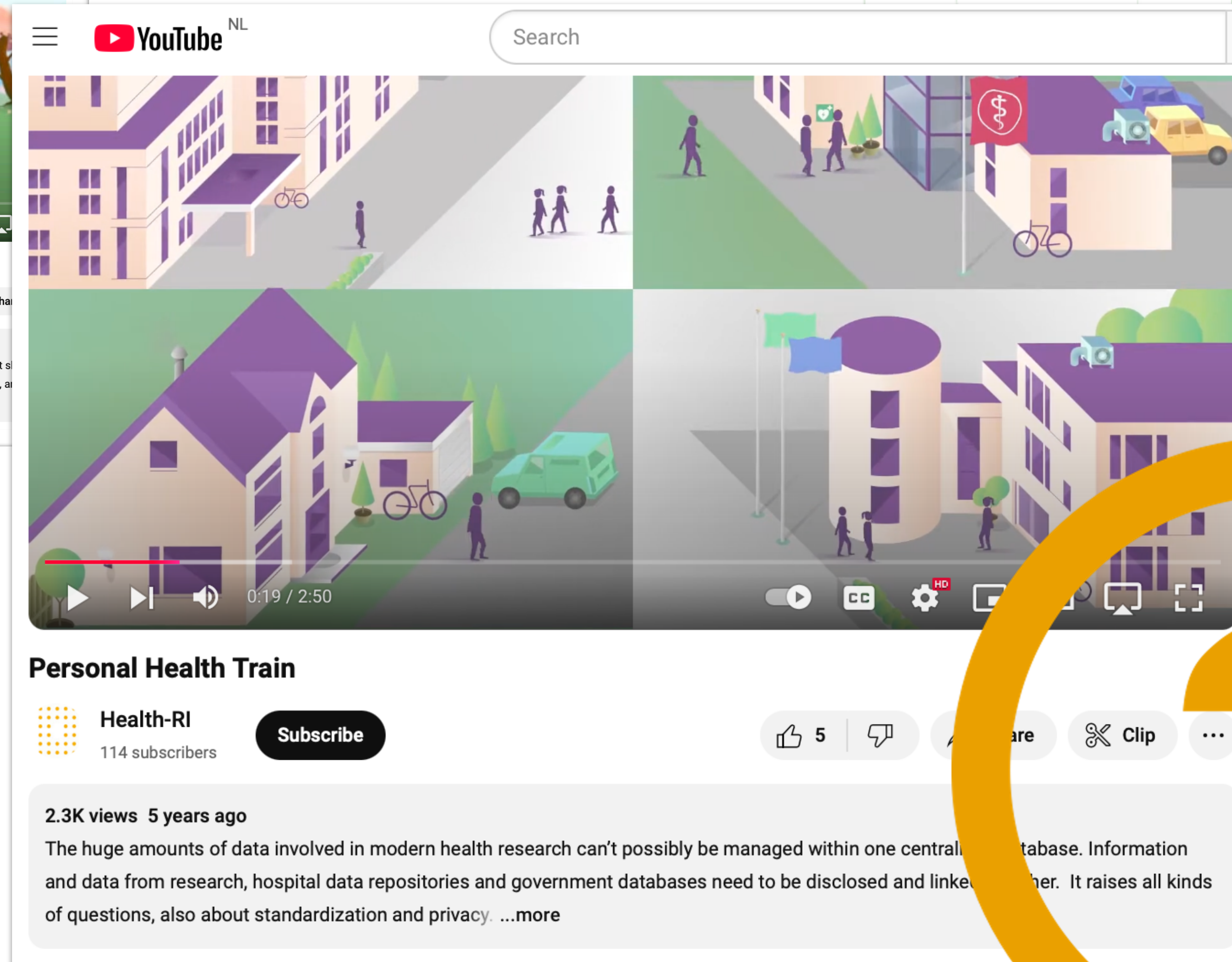


Data Sharing and Management Snafu in 3 Short Acts

NYU Health Sciences ...
1.03K subscribers

193K views 11 years ago

A data management horror story by Karen Hanson, Alisa Surkis and Karen Yacobucci. This is what happens when a researcher makes a data sharing request! Topics include storage, documentation, and formats. ...more

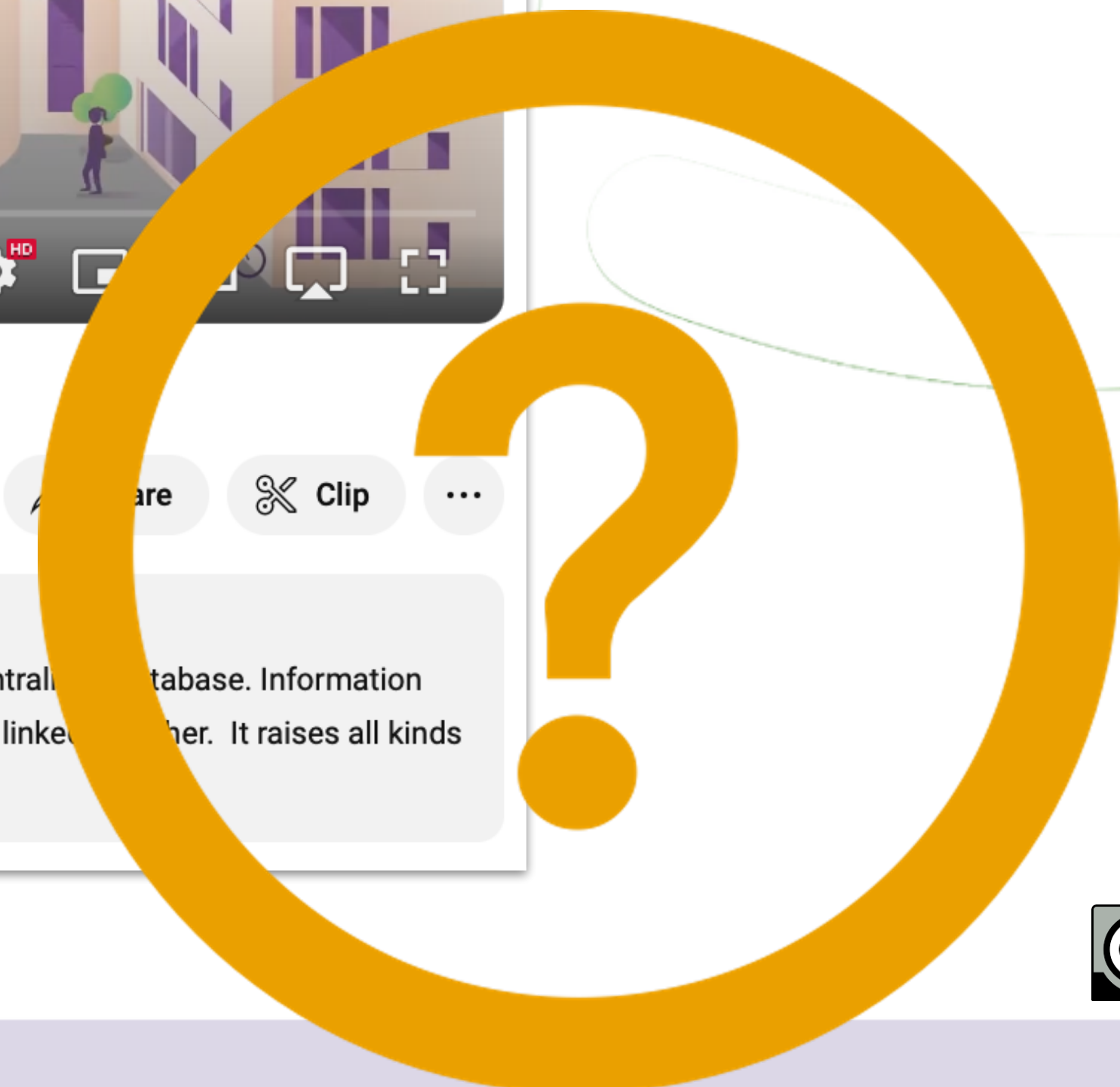


Personal Health Train

Health-RI
114 subscribers

2.3K views 5 years ago

The huge amounts of data involved in modern health research can't possibly be managed within one central database. Information and data from research, hospital data repositories and government databases need to be disclosed and linked together. It raises all kinds of questions, also about standardization and privacy. ...more



Agenda

13:00-13.10 (10 minutes)	Welcome (Erik)
13:10-13:30 (20 minutes)	Breakout - group sticky notes according to FAIR Principles
13:30-13:40 (10 minutes)	Report and discuss
13:40-14:00 (20 minutes)	Definitions: FAIR practice vs FAIR policy (Barbara)
14:00-14:30 (30 minutes)	Explore repositories of FAIR practices
14:30-14:40 (10 minutes)	Report and discuss
14:40-14:55 (15 minutes)	Break
14:55-15:10 (15 minutes)	FAIR data visiting, FAIR data train (Erik)
15:10-15:30 (20 minutes)	FAIR Hourglass (Erik)
15:30-15:45 (15 minutes)	Discussion
15:45-16:10 (25 minutes)	FAIR and AI (Erik)
16:10-16:20 (10 minutes)	Break
16:20-16:50 (35 minutes)	How can AI help in the FAIRification (discussion using FigJam)
16:50-17:00 (10 minutes)	Wrap up & Homework
17:00	Adjourn

Common Notes: https://bit.ly/FA-5-T-1_ITINERIS

Breakout

group sticky notes according to the

FAIR Principles

13:10-13:40
(4 groups)

Definitions:

FAIR practice vs FAIR policy

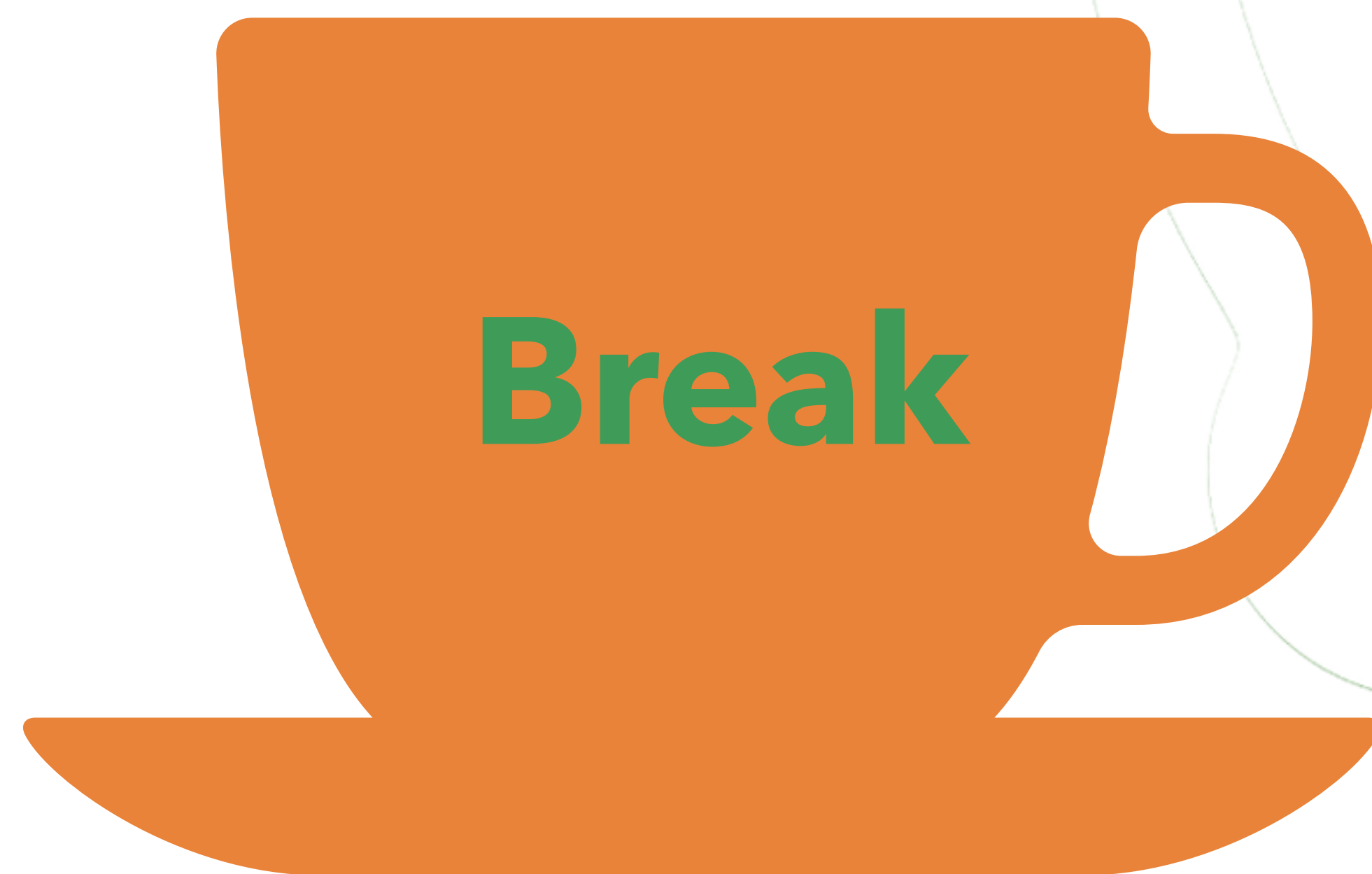
13:40-14:00
(Barbara)

Breakout

Explore repositories of FAIR practices

14:00-14:40
(3 groups)





Resume 15:30 CET

Conventional Data Sharing, FAIR data visiting & FAIR data trains

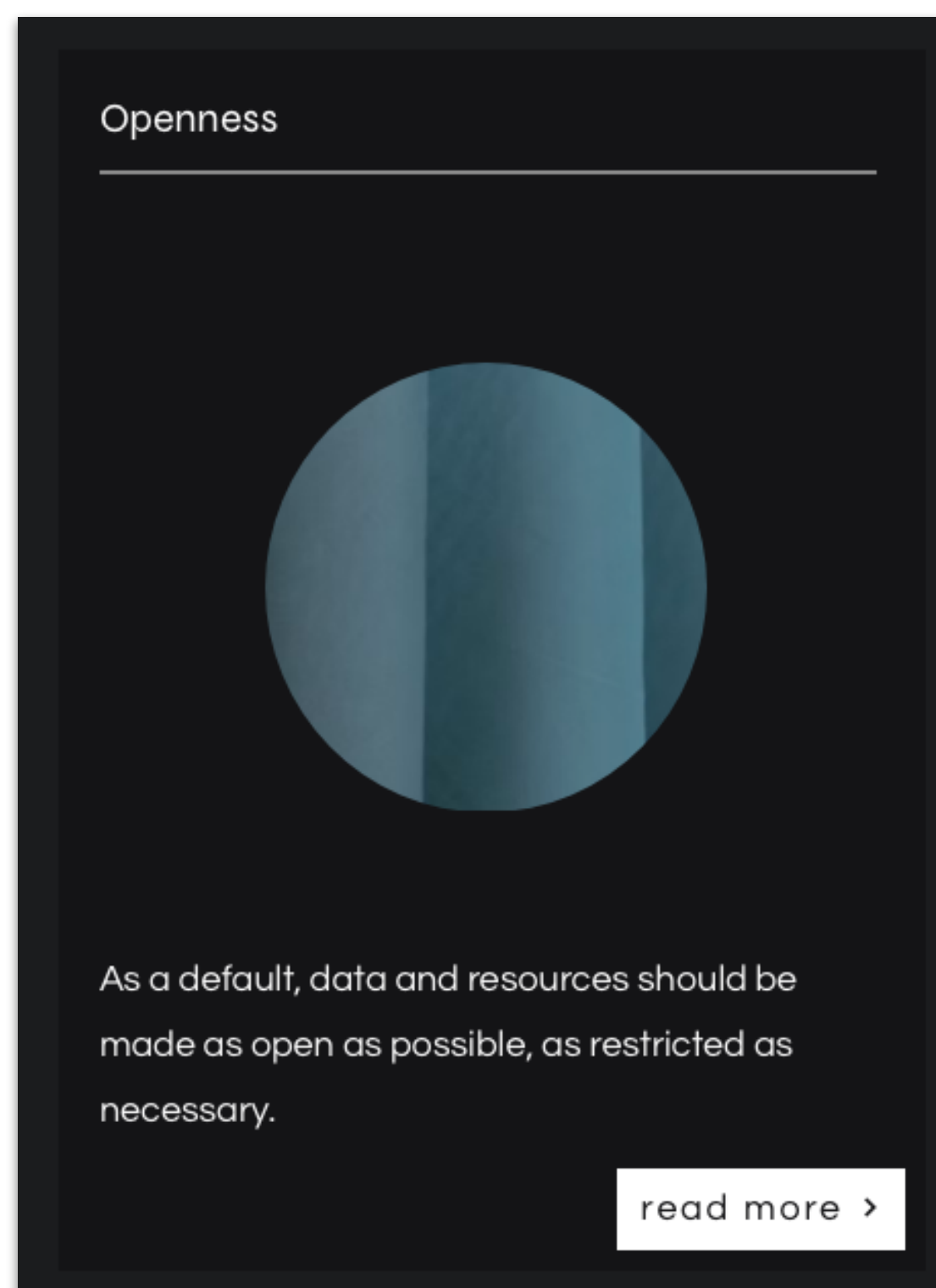
14:55-15:10
(Erik)



A1.2: the protocol allows for authentication & authorization procedure where necessary

- FAIR is not equal to ‘Open’ or ‘Free’
 - Data are often Open (OA articles) but not FAIR
 - Some data can never be Open, yet be perfectly FAIR

A1.2: the protocol allows for authentication & authorization procedure where necessary



As open as possible, as restricted as necessary
<https://www.gofair.foundation/openness>



A1.2: the protocol allows for authentication & authorization procedure where necessary

Reasons for restricting access to data:

- Personal privacy / confidentiality
- Institutional (hospital) or regional (GDPR) regulations
- Proprietary information (business considerations)
- Security

True patient story...

Dirk-Jan van der Pol (circa 2017)

- Patient gets thyroid cancer - thyroid is removed.
- 2 years later patient develops COPD (Chronic Obstructive Pulmonary Disease).
- Question: is there a causal link between removal of thyroid and COPD ?
- Patient consults doctors → “no connection”
- Patient consults PubMed → “no connection”
- Medical Odyssey begins:
 - 2 years + 50TB of data
 - patient gathers personal medical records
 - patient reads literature
- Finding: 3 proteins linking TC and COPD (CLCX12 down regulated in COPD)
- Hypothesis: Perhaps if CLCX12 levels can be stabilized, COPD may be prevented ?

True patient story...

Dirk-Jan van der Pol (circa 2017)

- v/d Pol consults HealthTap:
 - \$99 to get opinion of physician
 - 5 physicians say “no connection”
 - Another physician says “no connection, but I am wondering why you ask that question, may I call you”?
 - Patient says “yes... for \$99” ;-)



108,000 top doctors
available immediately to
help you live a healthier,
happier, longer life

True patient story...

Dirk-Jan van der Pol (circa 2017)

- v/d Pol consults Euretos:
 - Finds 5 proteins
 - 2 minutes



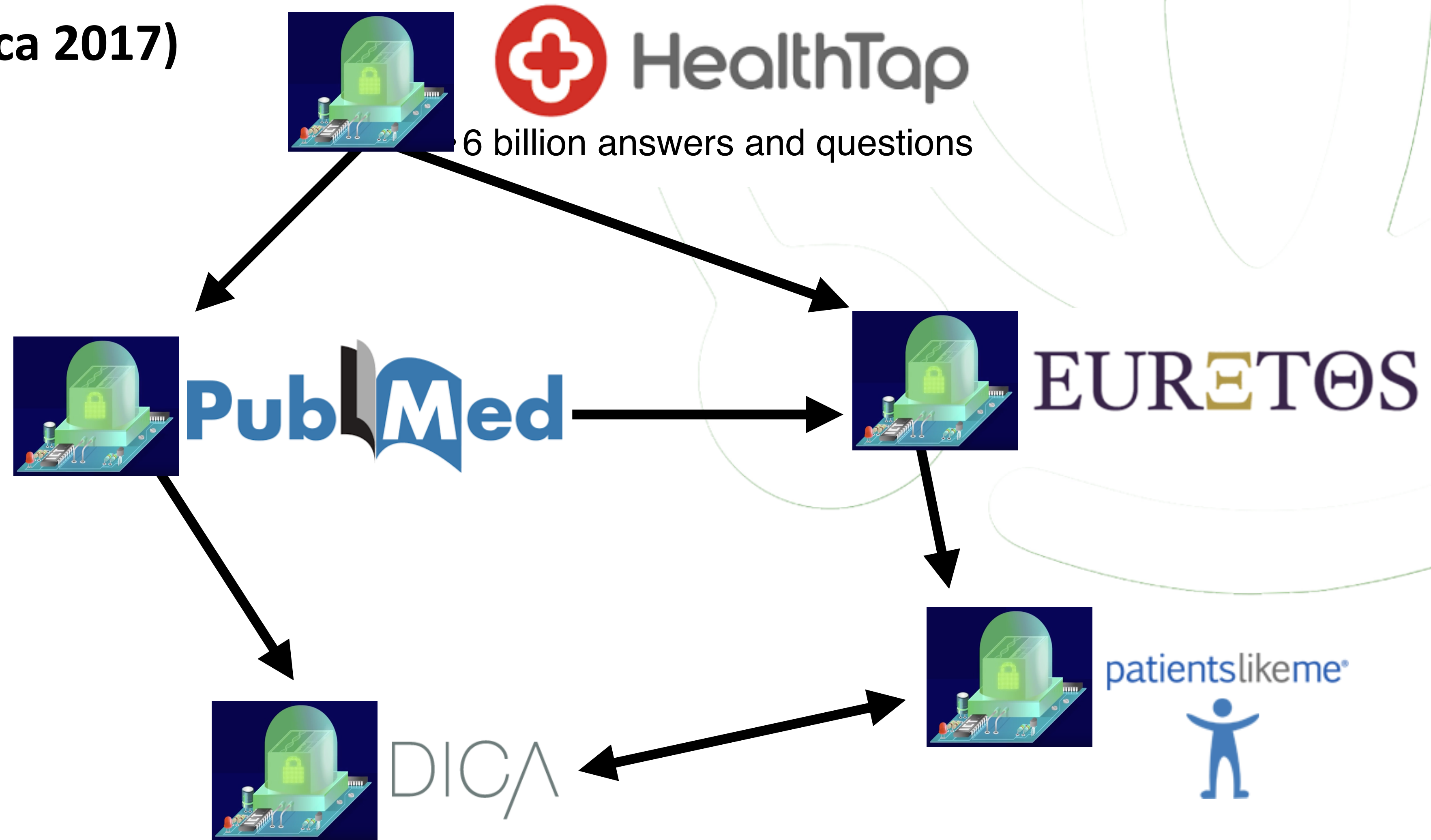
EURETOS

200+ integrated
biomedical
data resources



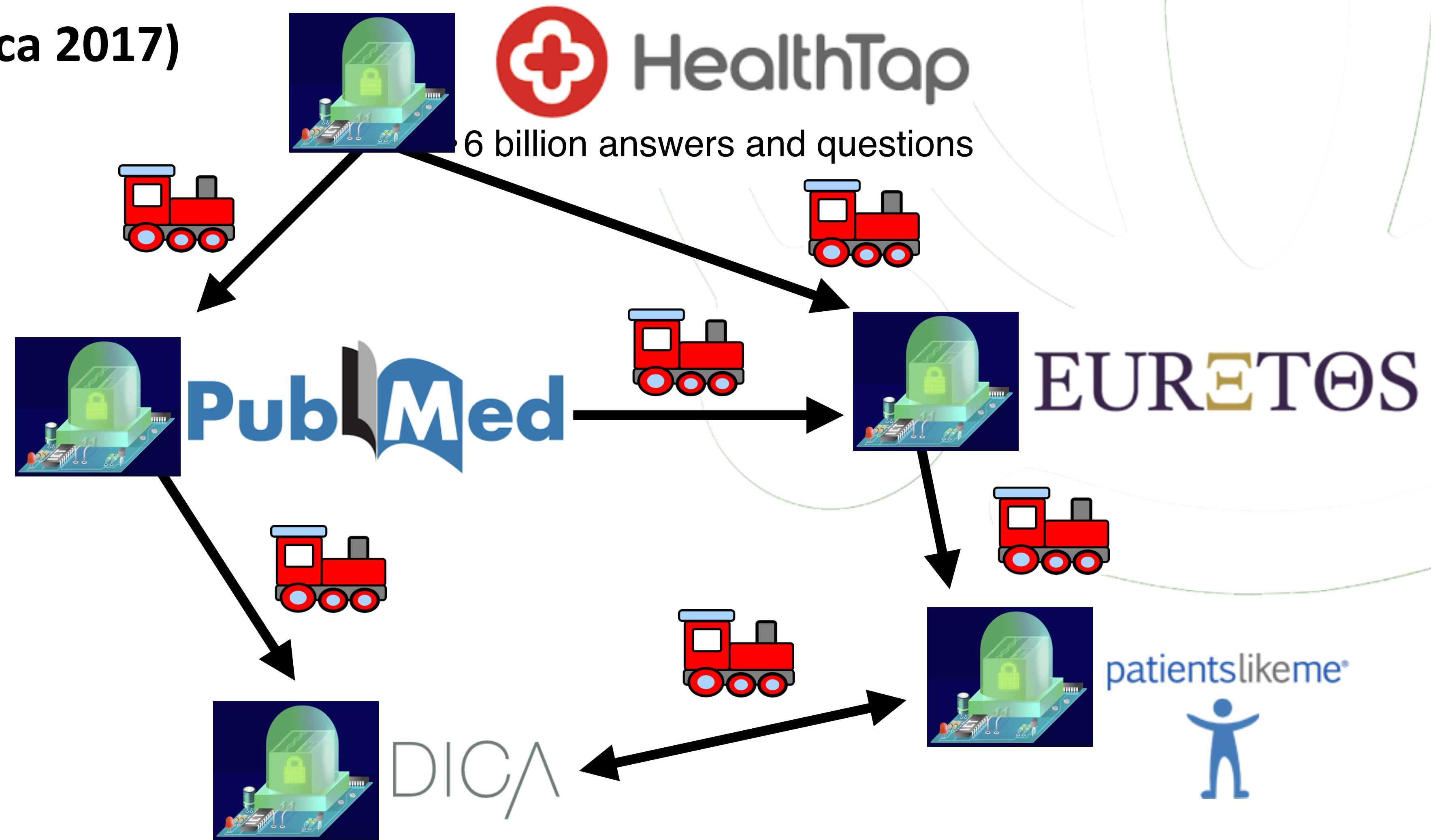
True patient story...

Dirk-Jan van der Pol (circa 2017)



True patient story...

Dirk-Jan van der Pol (circa 2017)



Many patient stories...

TED Ideas worth spreading

WATCH DISCOVER ATTEND PARTICIPATE ABOUT LOG IN

Dave deBronkart:

Meet e-Patient Dave

TEDxMaastricht · 16:31 · Filmed Apr 2011

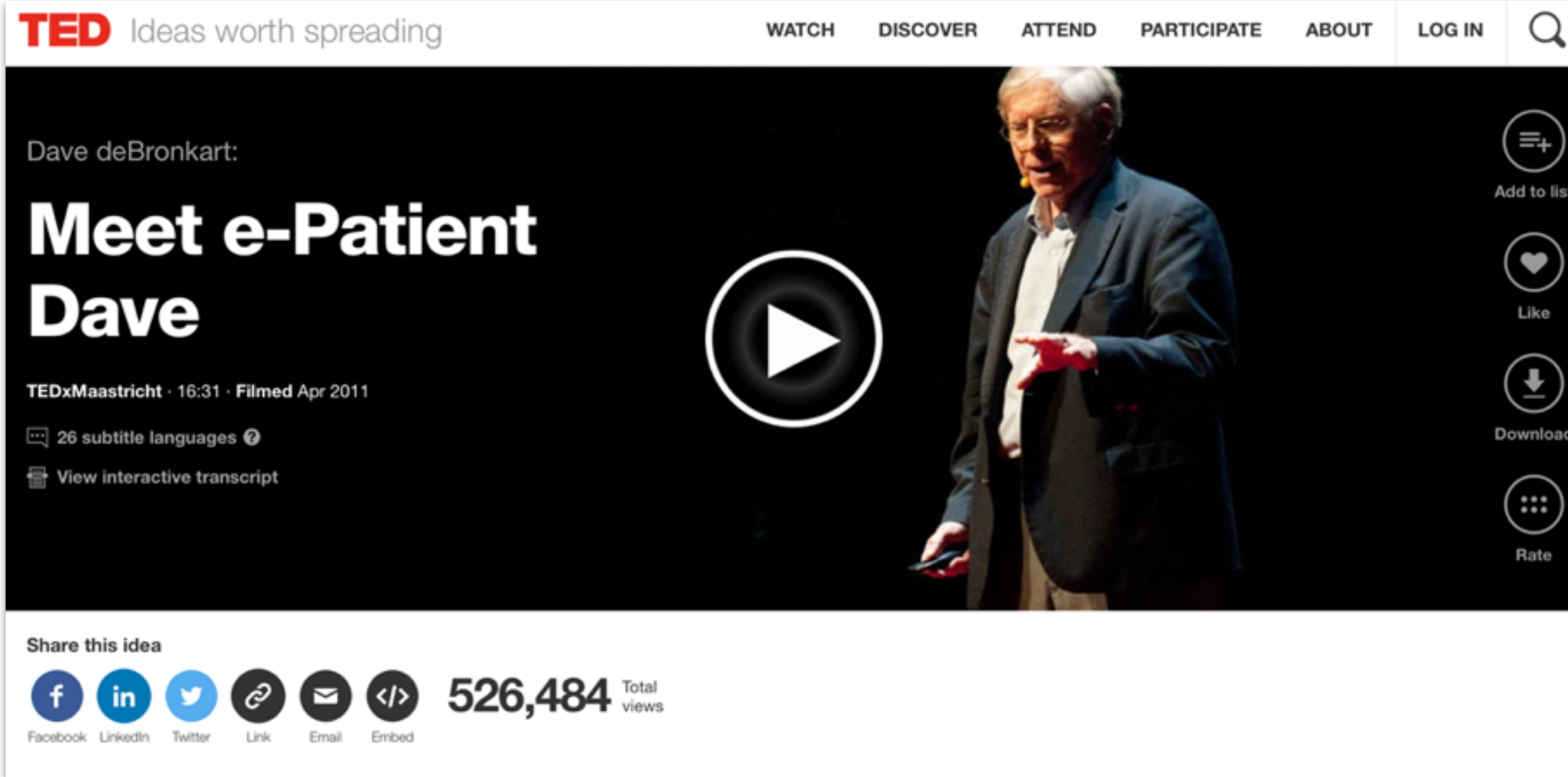
26 subtitle languages

View interactive transcript

Share this idea

526,484 Total views

Facebook LinkedIn Twitter Link Email Embed



Jack Andraka



Andraka in 2013

Born Jack Thomas Andraka
January 8, 1997 (age 20)
[Crownsville, Maryland, U.S.](#)

Residence United States

Nationality American

Fields [Cancer research, medical research, invention](#)



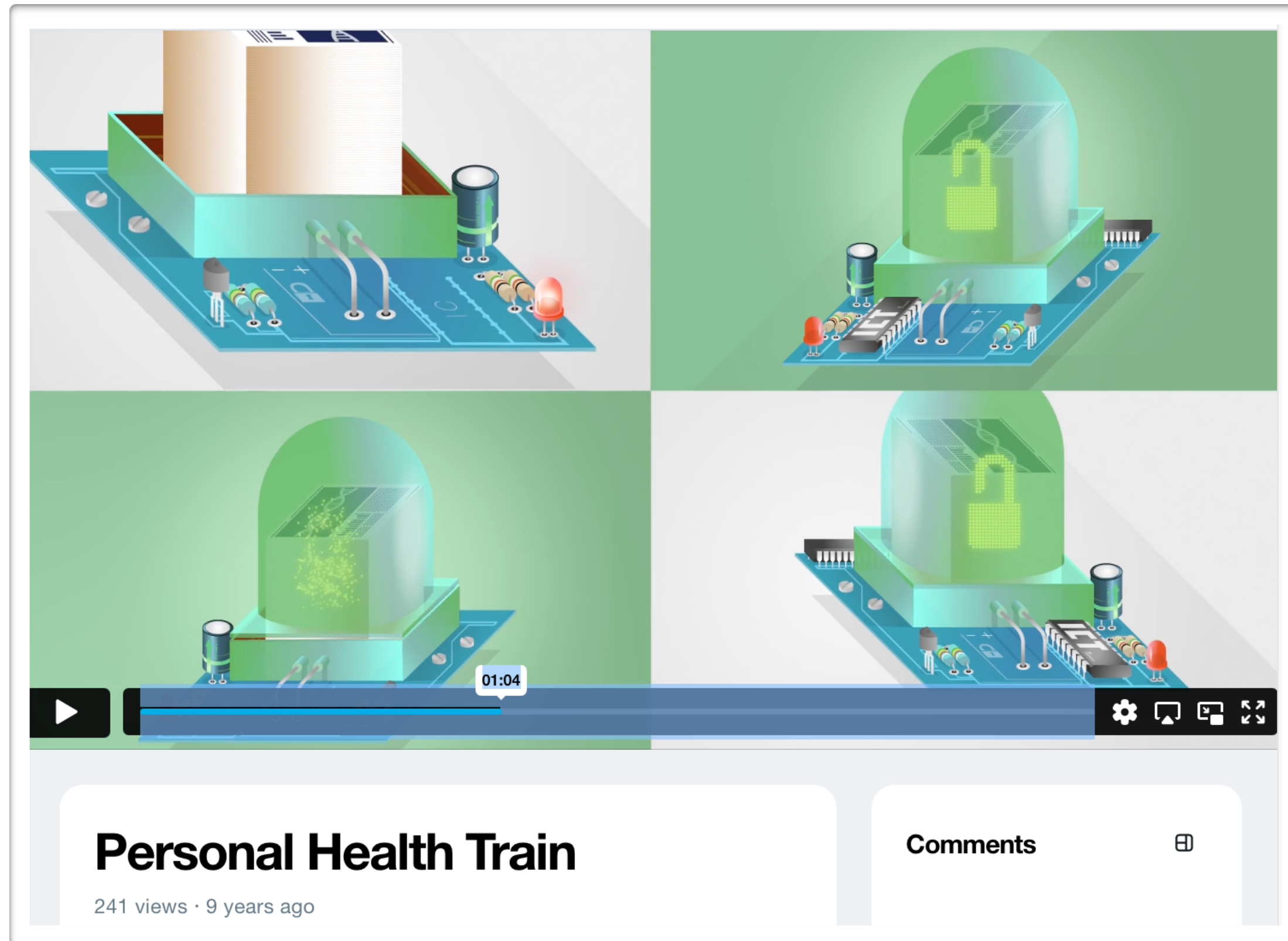
CHRISTIAN KROMME ALS SPREKER INHUREN

ONDERNEMER EN SPREKER OP GEBIED VAN
DISRUPTIEVE INNOVATIE, VISIE ONTWIKKELING EN
HUMAN CENTERED DESIGN

“Think big, start small!”



Personal Health Train



- Distributed Learning
- Edge Computing

Hi Erik,

The story of the start of the PHT is that we were working in Maastricht on the concept of distributed learning (bringing research to the data instead of vice versa) ...

In one sentence: The PHT started over a beer as a joint idea of Maastricht and Leiden as a combination of distributed learning (**trains, track, Maastricht**) and FAIR data (**stations, Leiden**).

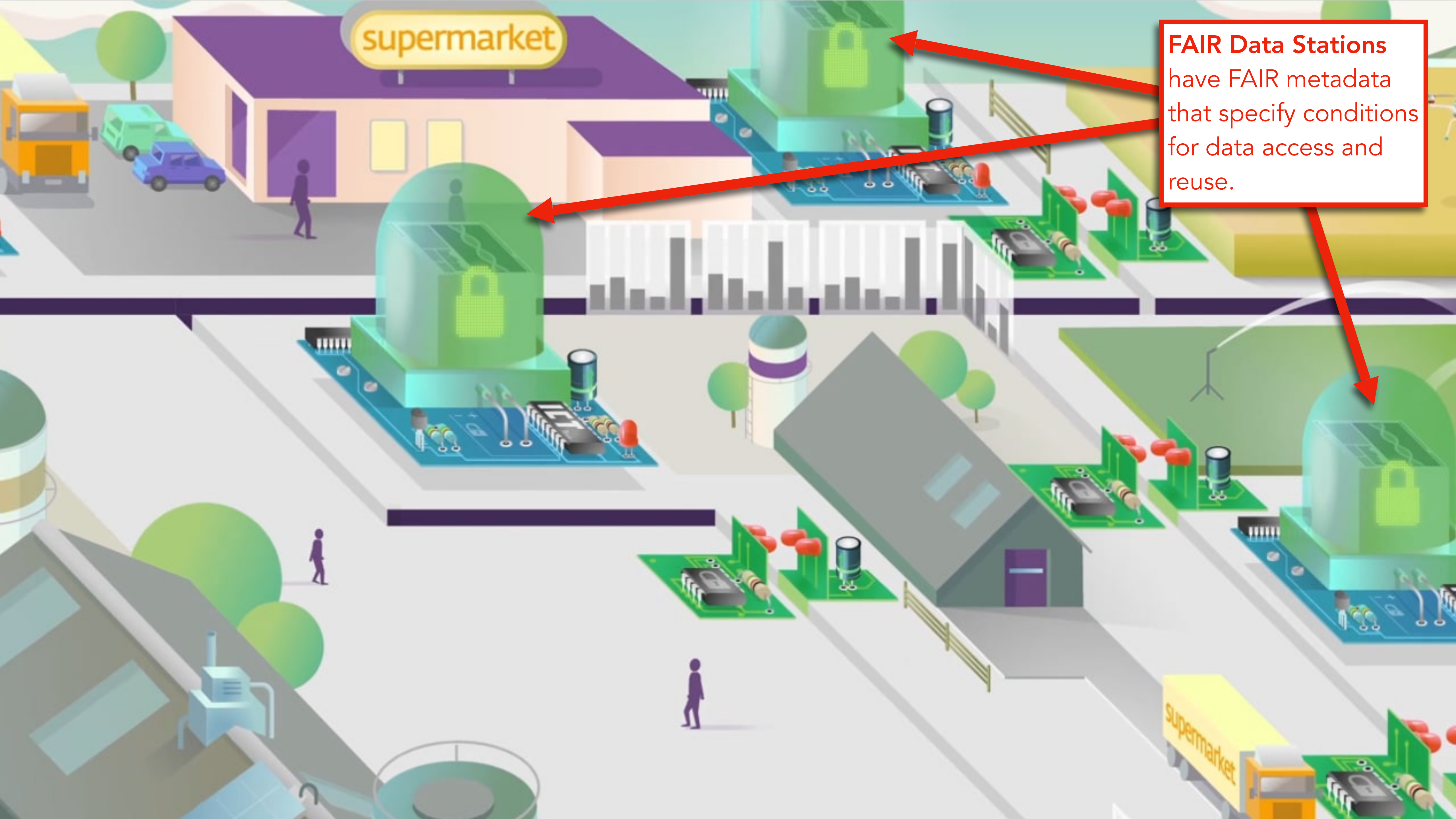
Lesson to be learned: Presenting your work at different venues and discussions over beer lead to increased serendipity in science.

Andre Dekker



supermarket

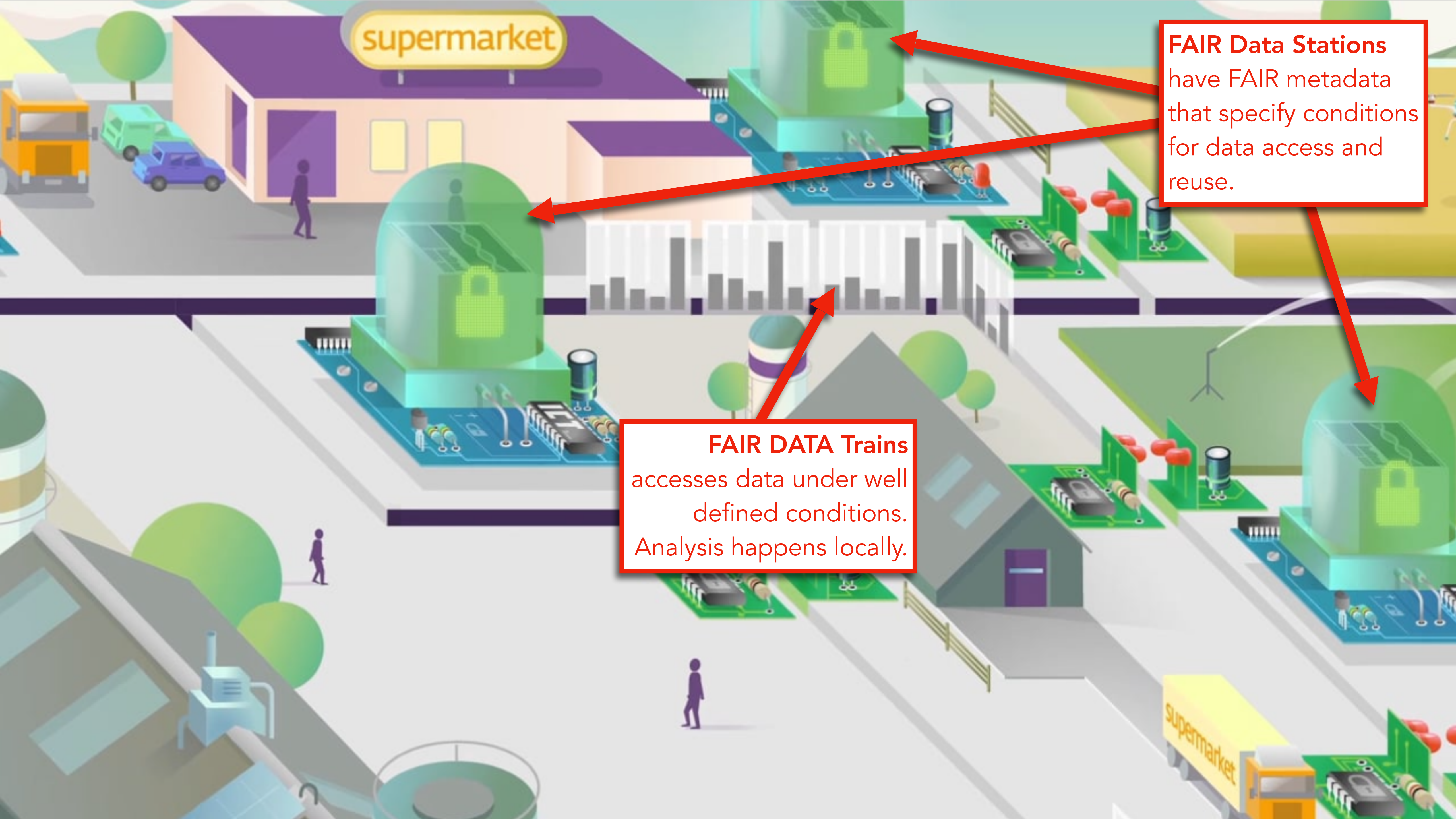
supermarket



supermarket

FAIR Data Stations
have FAIR metadata
that specify conditions
for data access and
reuse.

supermarket



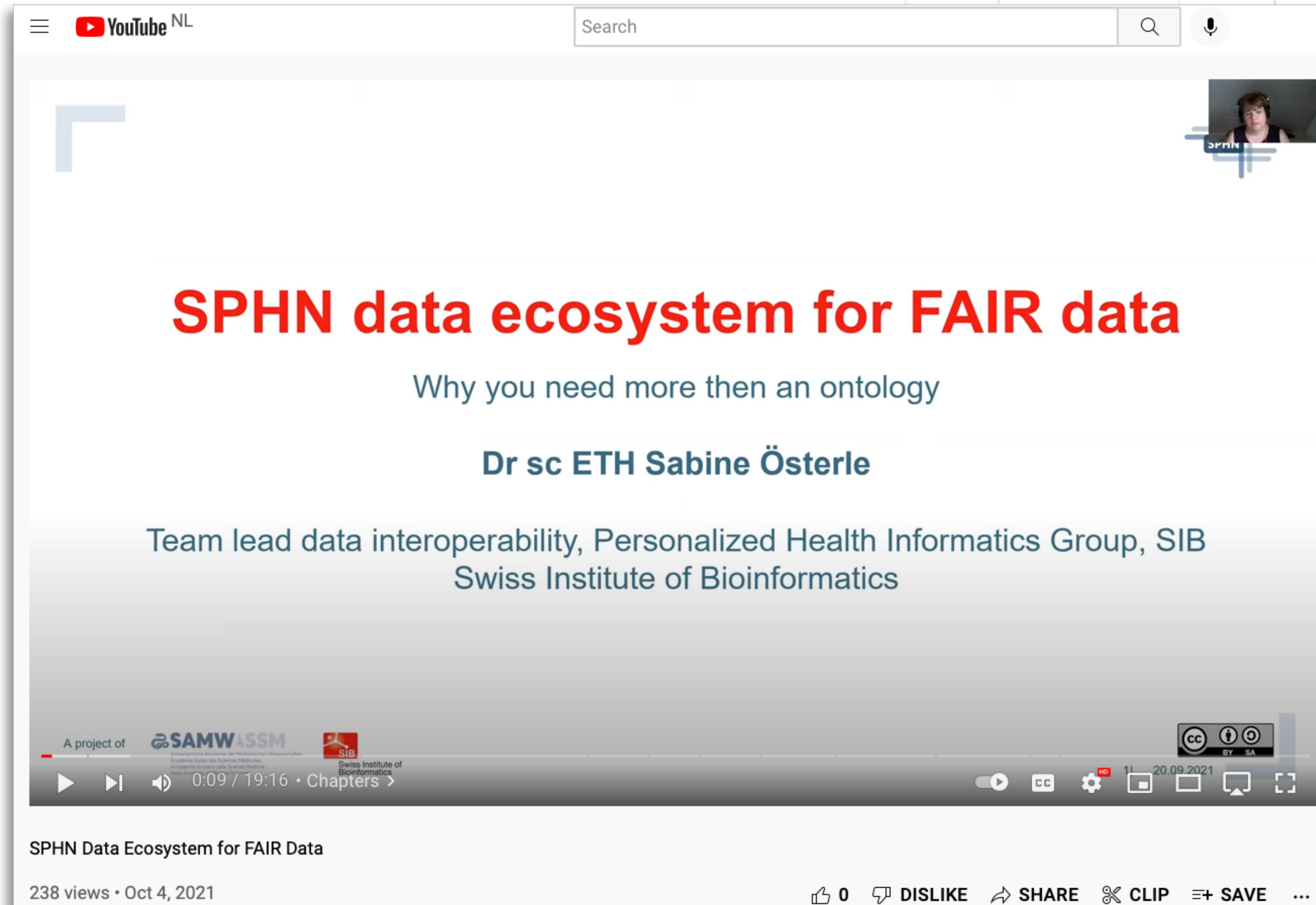
supermarket

FAIR Data Stations have FAIR metadata that specify conditions for data access and reuse.

FAIR DATA Trains accesses data under well defined conditions. Analysis happens locally.

supermarket

Personal Health Train: Implementations



The screenshot shows a YouTube video player interface. At the top, there is a search bar and a microphone icon. The video content is a presentation slide with the following text:

- SPHN data ecosystem for FAIR data** (in red)
- Why you need more than an ontology
- Dr sc ETH Sabine Österle**
- Team lead data interoperability, Personalized Health Informatics Group, SIB
Swiss Institute of Bioinformatics

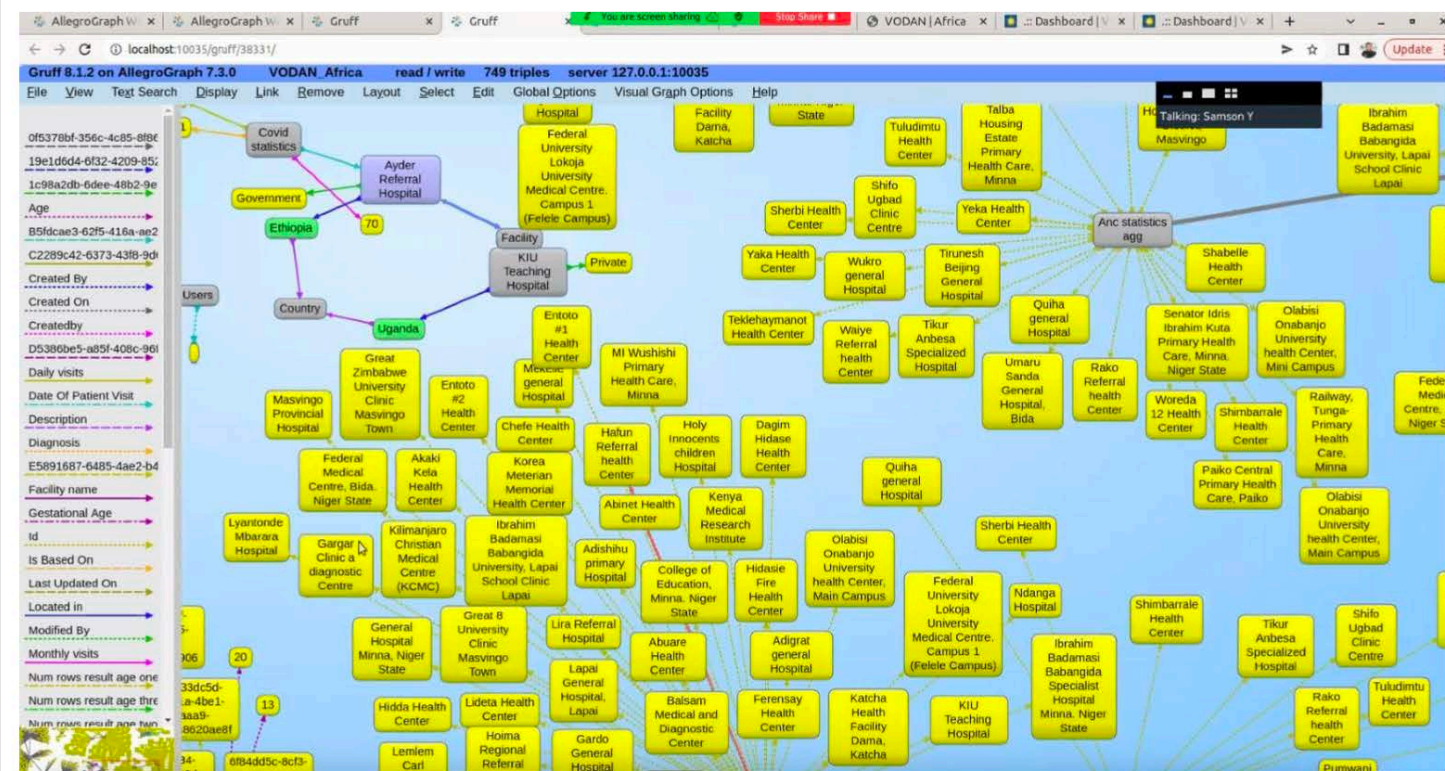
At the bottom of the slide, it says "A project of" followed by logos for SAMWASSM and SIB (Swiss Institute of Bioinformatics). The video player controls at the bottom show a progress bar at 0:09 / 19:16, a play button, and a volume icon. To the right of the controls are icons for CC BY SA, a red "1" icon, and the date "20.09.2021". Below the video player, the video title "SPHN Data Ecosystem for FAIR Data" is displayed, along with "238 views • Oct 4, 2021" and interaction buttons for Like (0), Dislike, Share, Clip, Save, and a menu icon.



Personal Health Train: Implementations



Invitation VODAN-Africa platform Presentation



When: Wednesday 23 Nov 2022

Where: LUMC P-05-034 Leiden University
12:00-14:00 CET/14:00-16:00EAT

On-line video zoom [link](#)

Programme:

- Demonstration of the data and platform
- Life demonstration of queries and analytics
- Presentation of next phase



What We Do About DMD/BMD News Care Research Advocacy Library [Donate](#)

DATE
22
11
22

TIME
13:00
17:00
GET

ONLINE

SAVE THE DATE

3RD INTERNATIONAL FAIR
MEETING ON DATA "VISITING" FOR
DUCHENNE & OTHER RARE DISEASES

WORLD DUCHENNE ORGANIZATION
UPPMD

3RD INTERNATIONAL FAIR DATA 'VISITING' FOR DUCHENNE & OTHER RARE DISEASES



Personal Health Train: Implementations

nature View all journals Search Q Login

Explore content About the journal Publish with us

nature > articles > article

Download PDF

Article | Open Access | Published: 26 May 2021

Swarm Learning for decentralized and confidential clinical machine learning

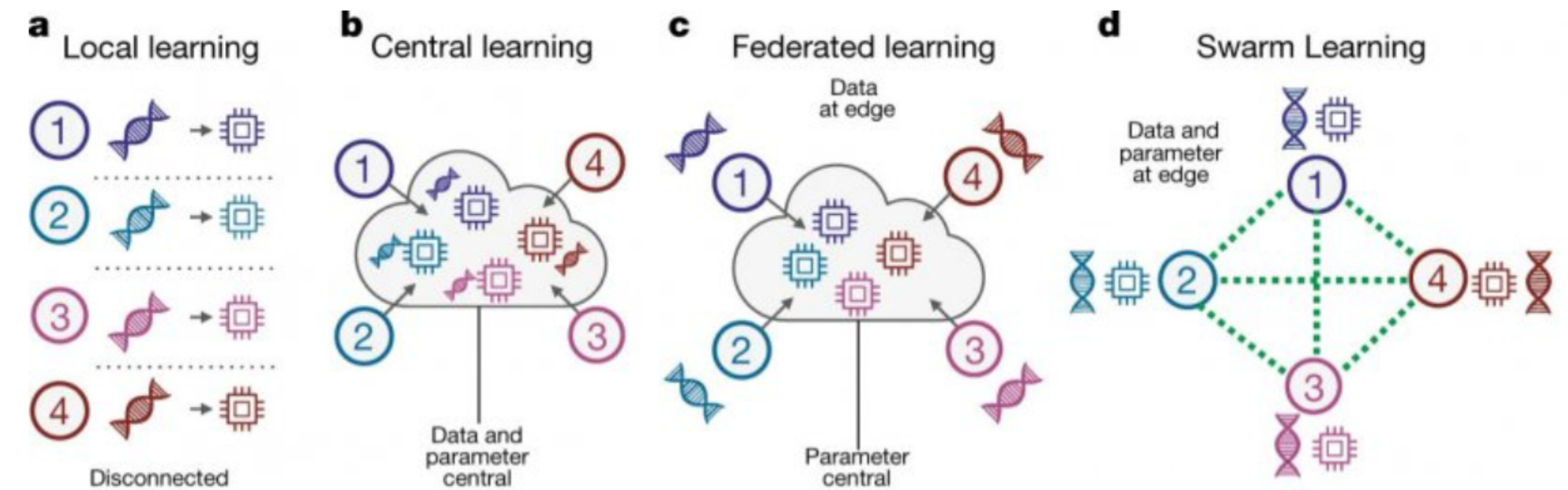
Stefanie Warnat-Herresthal, Hartmut Schultze, Krishnaprasad Lingadahalli Shastry, Sathyanarayanan Manamohan, Saikat Mukherjee, Vishesh Garg, Ravi Sarveswara, Kristian Händler, Peter Pickkers, N. Ahmad Aziz, Sofia Ktena, Florian Tran, Michael Bitzer, Stephan Ossowski, Nicolas Casadei, Christian Herr, Daniel Petersheim, Uta Behrends, Fabian Kern, Tobias Fehlmann, Philipp Schommers, Clara Lehmann, Max Augustin, Jan Rybniker, COVID-19 Aachen Study (COVAS), Deutsche COVID-19 Omics Initiative (DeCOI), ... Joachim L. Schultze

Nature 594, 265–270 (2021) | Cite this article

97k Accesses | 99 Citations | 479 Altmetric | Metrics

Abstract

<https://healthcare-in-europe.com/en/news/ai-with-swarm-intelligence-to-analyse-medical-data.html>

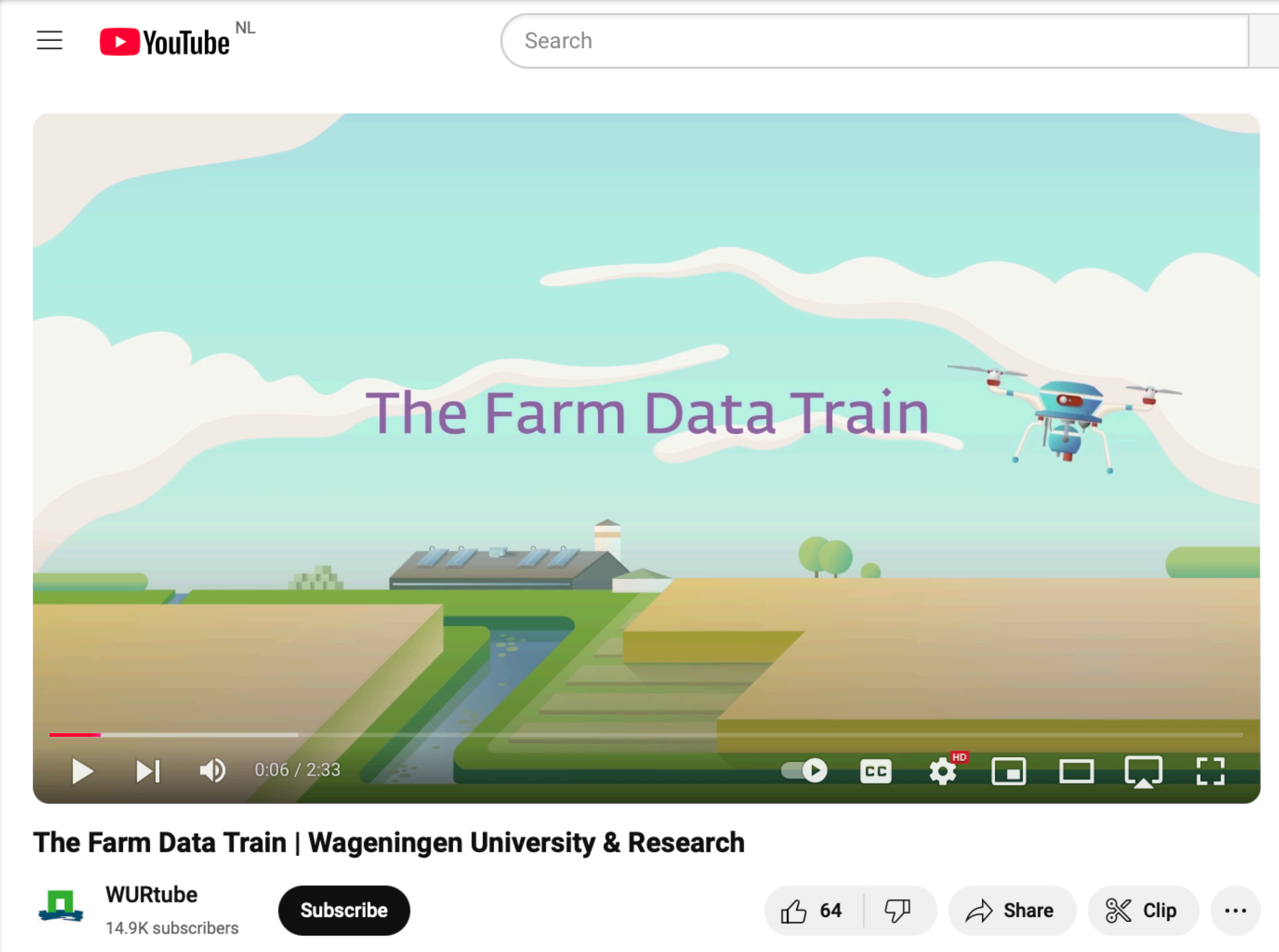


The illustration compares the different learning concepts: a) local learning with data and computation at different, disconnected locations; b) cloud-based machine learning; c) Federated learning, with data being kept with the data contributor and computing performed at the site of local data storage and availability, but parameter settings orchestrated by a central parameter server; d) Swarm Learning without the need for a central custodian

Image source: Warnat-Herresthal et al., Nature 2021 (CC BY 4.0)



Farm Data Train



The Farm Data Train | Wageningen University & Research

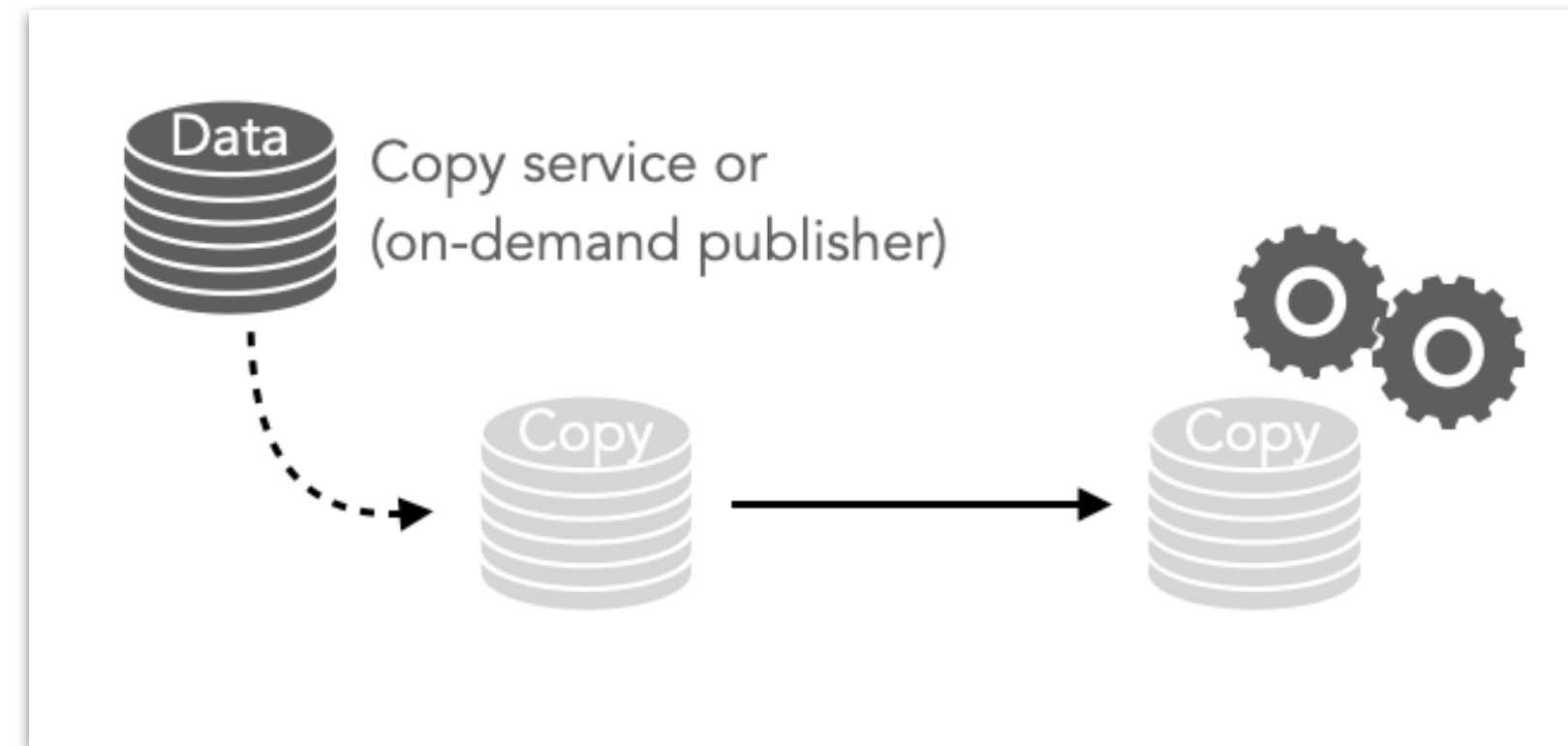
WURtube
14.9K subscribers

64 | Share | Clip



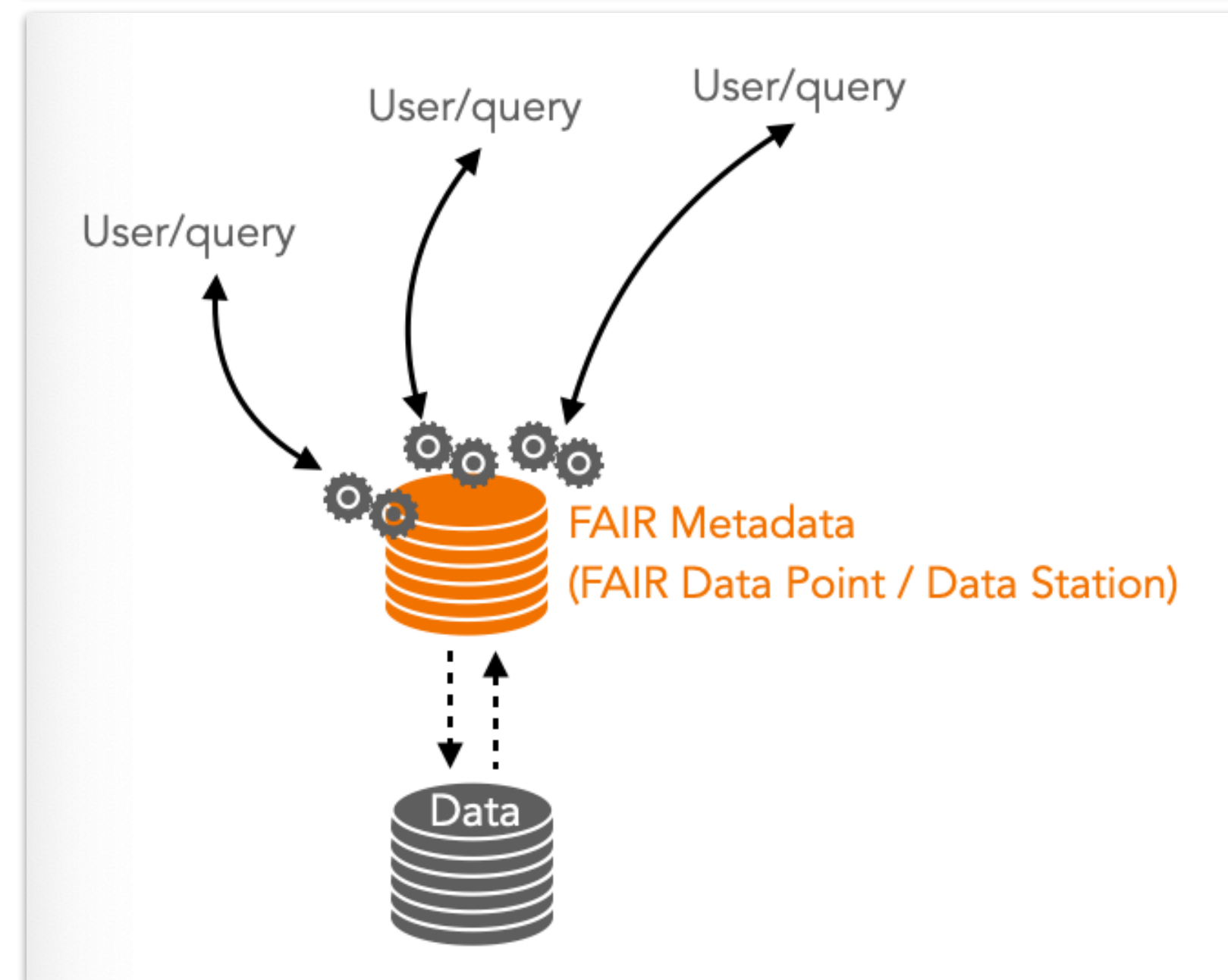
FAIR Data Train & FAIR Data Visting

An alternative to "data centers" and "cloud" environments
<https://registry.ern-euro-nmd.eu/the-way-forward-is-not-data-sharing-but-data-visiting/>



Data Sharing:

- copy and transport data
- analysis happens elsewhere
- data is out of your control



Data Visitation:

- data stay local (are not copied)
- analysis happens locally
- data access is known

FAIR Data Train & FAIR Data Visting



LIFES Networking Meeting - November ...

 **LIFES Institute**
249 followers
14h · 🌐 [+ Follow](#) [...](#)

Today was another great networking meeting @LIFES with over 40 people joining in Leiden and online. We discussed what is needed to make distributed data & learning systems interoperable and decided on jointly ...more



 Hans Roubos and 23 others 6 reposts

  Like  Comment  Repost  Send

Today was another great networking meeting @LIFES with over 40 people joining in Leiden and online. We discussed what is needed to make **distributed data & learning systems** interoperable and decided on jointly developing specifications and testing them on existing implementations of LIFES members. Great to see such enthusiasm and commitment in the LIFES network!

Special thanks to the presenters: [Luiz Olavo Bonino da Silva Santos](#), [Hans Niendieker](#), [Jildau Bouwman](#), [Dyonne Van der Leer - de Mari](#) and [Joachim Schultze](#).



FAIR Hourglass

15:10-15:30
(Erik)

Common Patterns in Revolutionary Infrastructures and Data

Peter Wittenburg, Max Planck Computing and Data Facility

George Strawn, US National Academy of Sciences

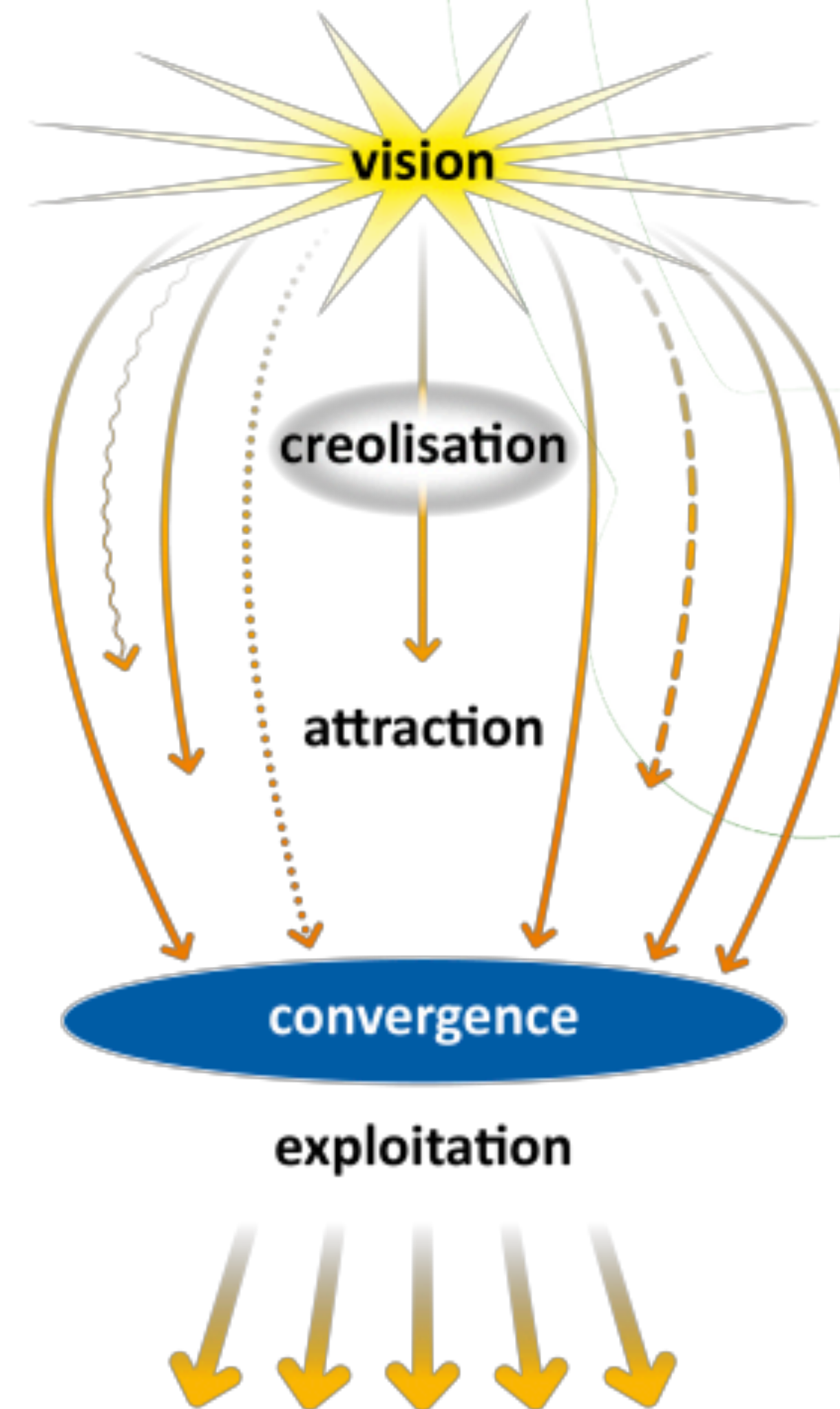
February 2018

1. Summary

Societies have seen large infrastructures emerge when new technologies become available. From history we see that such infrastructures can have a huge influence on all aspects of societal life. Moreover, some patterns appear to reoccur in the evolution of such infrastructures. *Early visions* about the possibilities of a new technology lead to a phase of *creolization*¹ of approaches resulting in a deeper knowledge of the technology's pros, cons and limitations. A huge "solutions space" emerges and fragmentation results. Some solutions are more *attractive* than others, but a final phase transition occurs where the experts converge towards broadly accepted principles and specifications that lead to *exploitation and standardization*.

It appears that the "data infrastructure" is evolving into such a large infrastructure, with a potentially large influence on societies, industry and science. In order to gain new insights about complex relationships in nature, societies and minds, by integrating data from different silos we have seen an explosion of (non-interoperable!) solutions for data management, access and processing, i.e. we have entered a phase of *creolization*. Also, we have an increasingly clear view of the current inefficiencies in working with data. These inefficiencies retard innovation and broad participation, which will become even more important as billions of smart devices produce the data deluge of the Internet of Things. Stakeholders have begun looking for steps toward *convergence* that would increase efficiency without hampering innovation.

Comparing the evolution of the data infrastructure with the evolution of the infrastructures of electrification, computer networking and of information networking (WWW), we can observe that, despite all initiatives already taken, we have not reached convergence on a set of universals that would boost developments and create a momentum towards an efficient and interoperable data infrastructure. We propose that such a set of universals could be based on the concepts of "Digital Objects" (DOs), persistent identifiers (PIDs), and metadata (including data typing). These concepts could greatly reduce current inefficiencies in data processing and open the way towards automatic processing. In particular, the Core Data Model of the Research Data Alliance (RDA) provides a design for a universal Digital Object Access Protocol (DOAP, comparable to IP for the Internet or HTTP for the Web) which can interconnect the many organizations of data in use today, such as cloud systems, files systems, SQL databases, no-SQL databases and so forth. The agreement on fairly simple but potentially universal commonalities such as PIDs, DOs, and a DAOP could create the confidence for many developers to invest in data infrastructure building. We believe that it is time to take this step towards convergence.



Prototypes & over-specified solutions

Generalization

Influential early movers (Investments)

Common Patterns in Revolutionary Infrastructures and Data

Peter Wittenburg, Max Planck Computing and Data Facility

George Strawn, US National Academy of Sciences

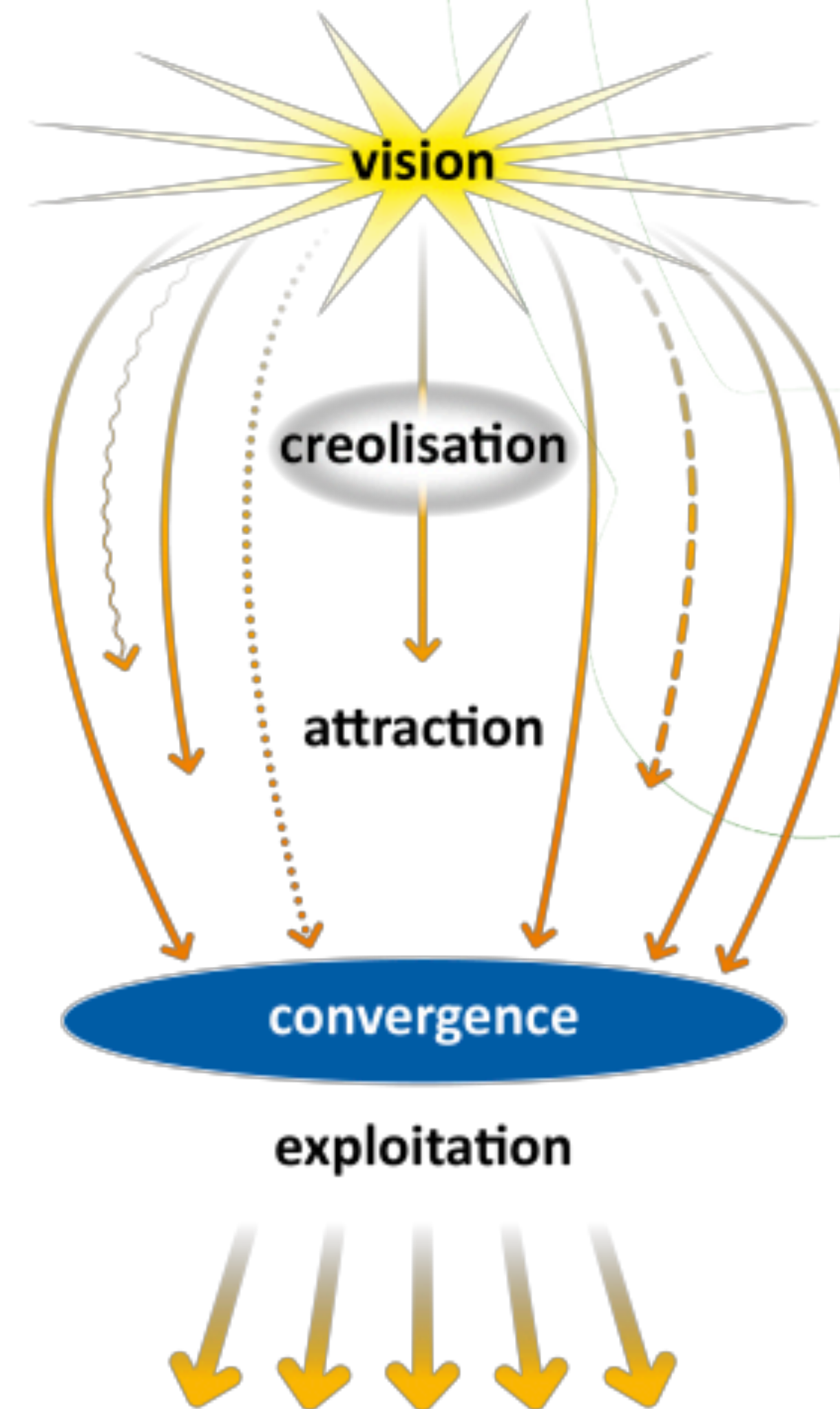
February 2018

1. Summary

Societies have seen large infrastructures emerge when new technologies become available. From history we see that such infrastructures can have a huge influence on all aspects of societal life. Moreover, some patterns appear to reoccur in the evolution of such infrastructures. *Early visions* about the possibilities of a new technology lead to a phase of *creolization*¹ of approaches resulting in a deeper knowledge of the technology's pros, cons and limitations. A huge "solutions space" emerges and fragmentation results. Some solutions are more *attractive* than others, but a final phase transition occurs where the experts converge towards broadly accepted principles and specifications that lead to *exploitation and standardization*.

It appears that the "data infrastructure" is evolving into such a large infrastructure, with a potentially large influence on societies, industry and science. In order to gain new insights about complex relationships in nature, societies and minds, by integrating data from different silos we have seen an explosion of (non-interoperable!) solutions for data management, access and processing, i.e. we have entered a phase of *creolization*. Also, we have an increasingly clear view of the current inefficiencies in working with data. These inefficiencies retard innovation and broad participation, which will become even more important as billions of smart devices produce the data deluge of the Internet of Things. Stakeholders have begun looking for steps toward *convergence* that would increase efficiency without hampering innovation.

Comparing the evolution of the data infrastructure with the evolution of the infrastructures of electrification, computer networking and of information networking (WWW), we can observe that, despite all initiatives already taken, we have not reached convergence on a set of universals that would boost developments and create a momentum towards an efficient and interoperable data infrastructure. We propose that such a set of universals could be based on the concepts of "Digital Objects" (DOs), persistent identifiers (PIDs), and metadata (including data typing). These concepts could greatly reduce current inefficiencies in data processing and open the way towards automatic processing. In particular, the Core Data Model of the Research Data Alliance (RDA) provides a design for a universal Digital Object Access Protocol (DOAP, comparable to IP for the Internet or HTTP for the Web) which can interconnect the many organizations of data in use today, such as cloud systems, files systems, SQL databases, no-SQL databases and so forth. The agreement on fairly simple but potentially universal commonalities such as PIDs, DOs, and a DAOP could create the confidence for many developers to invest in data infrastructure building. We believe that it is time to take this step towards convergence.

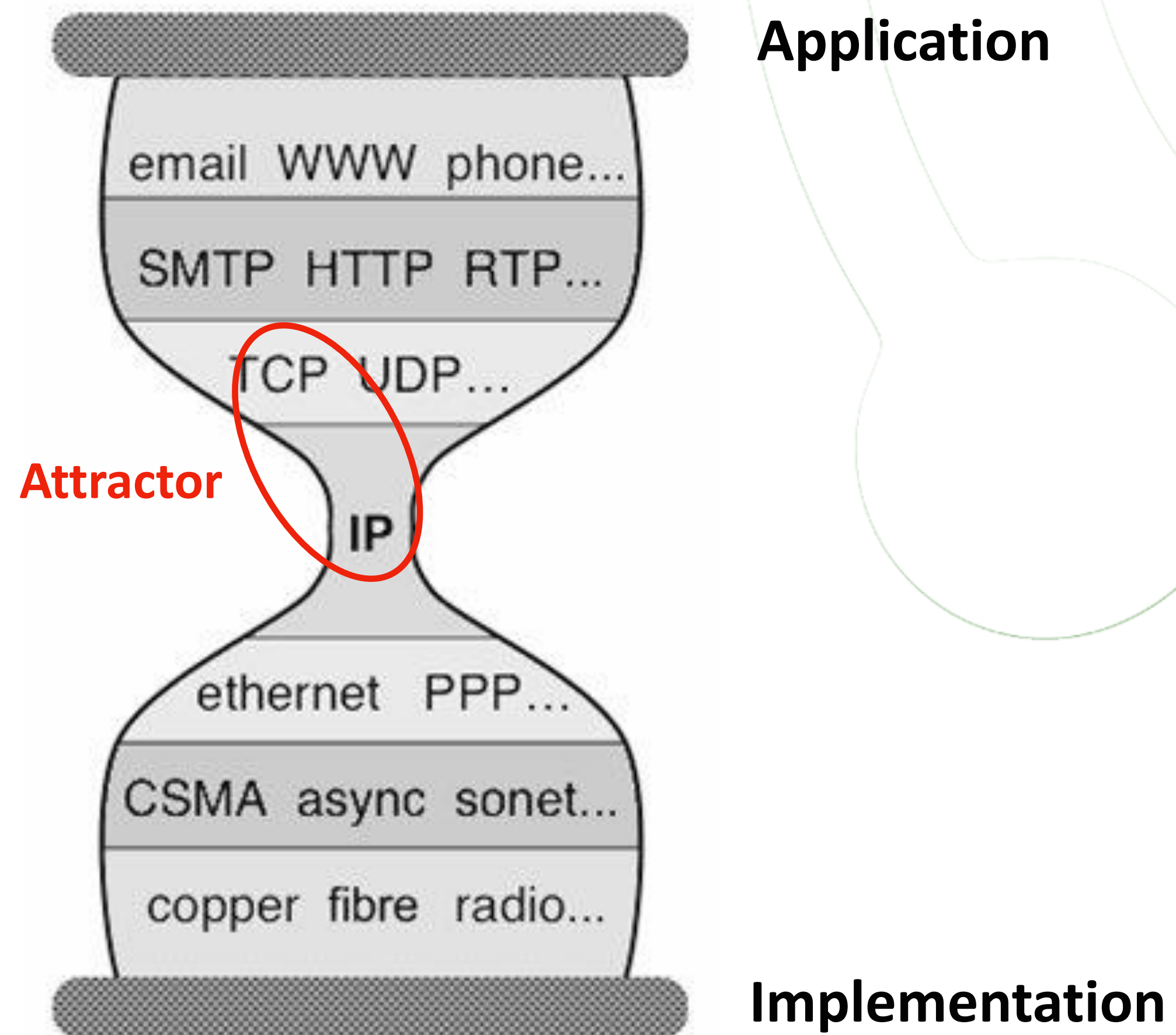


Interoperation of computer networks - an "internet"

TCP/IP

1995 - private sector provided the critical mass

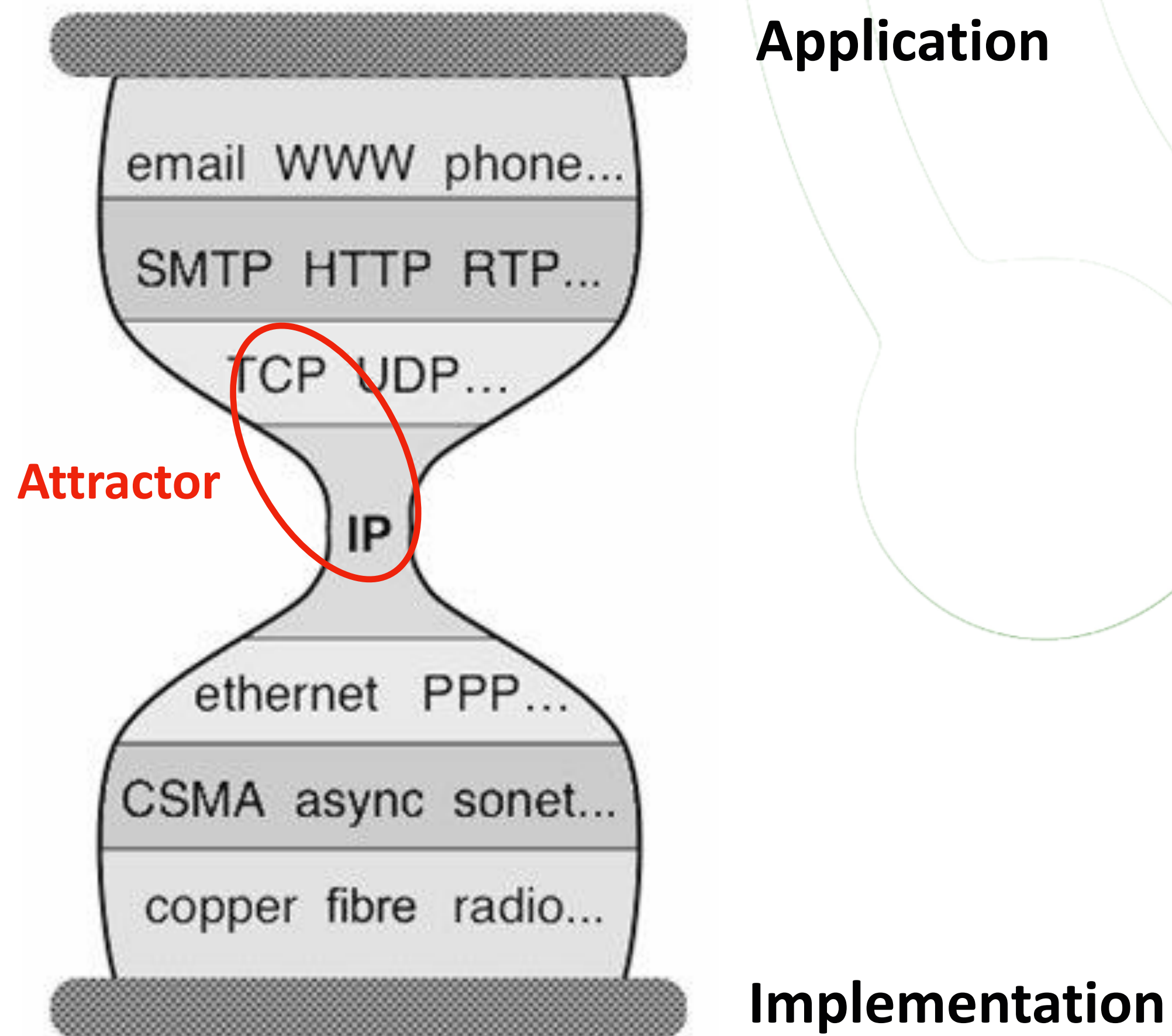
Hourglass Architecture of the Internet



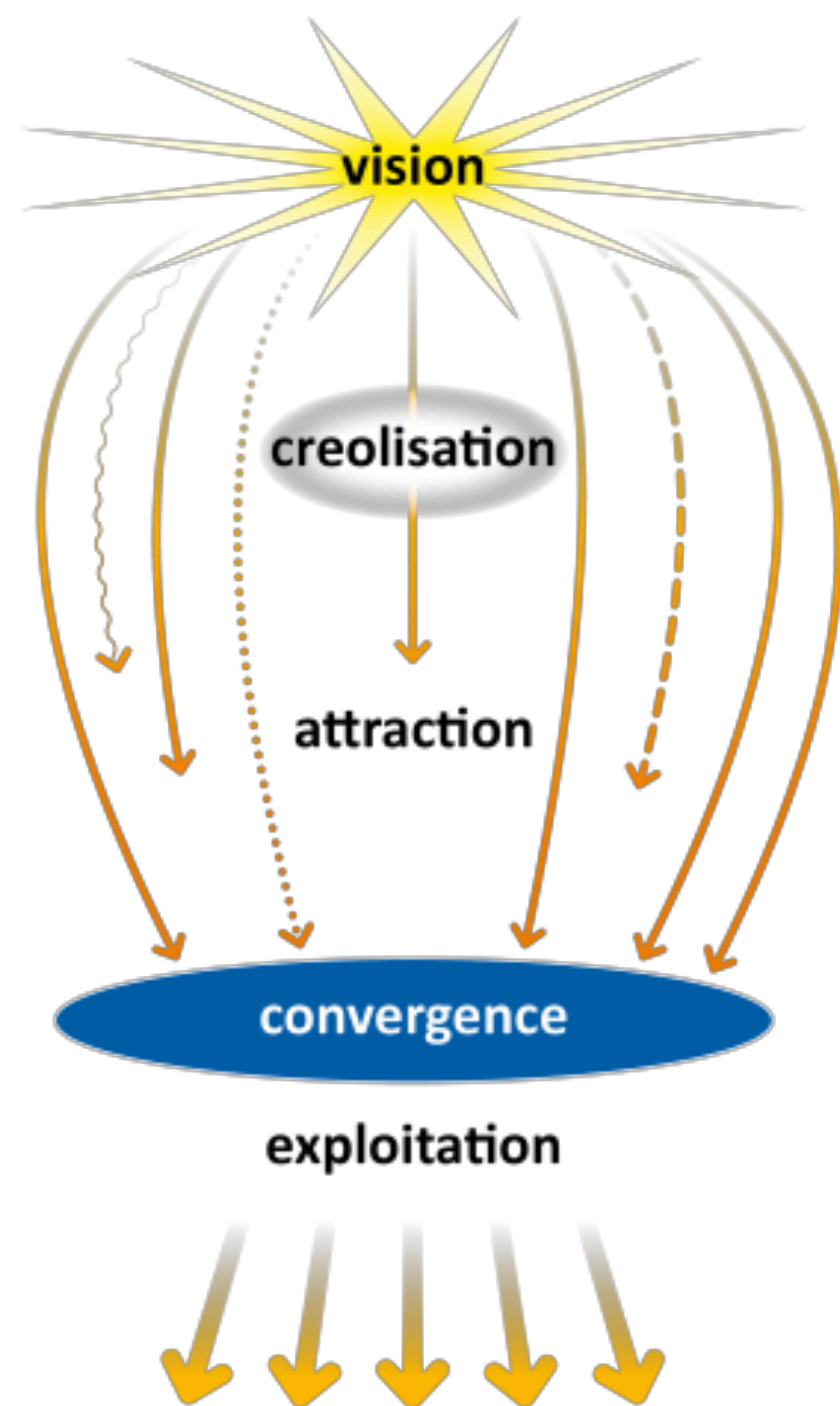
Hourglass Architecture of the Internet

A recipe for convergence

- Minimal standard
- Voluntary participation
- Critical mass of users
- Rough consensus, running code
- 2 implementations of a spec



Hourglass Architecture of the Data-Internet?



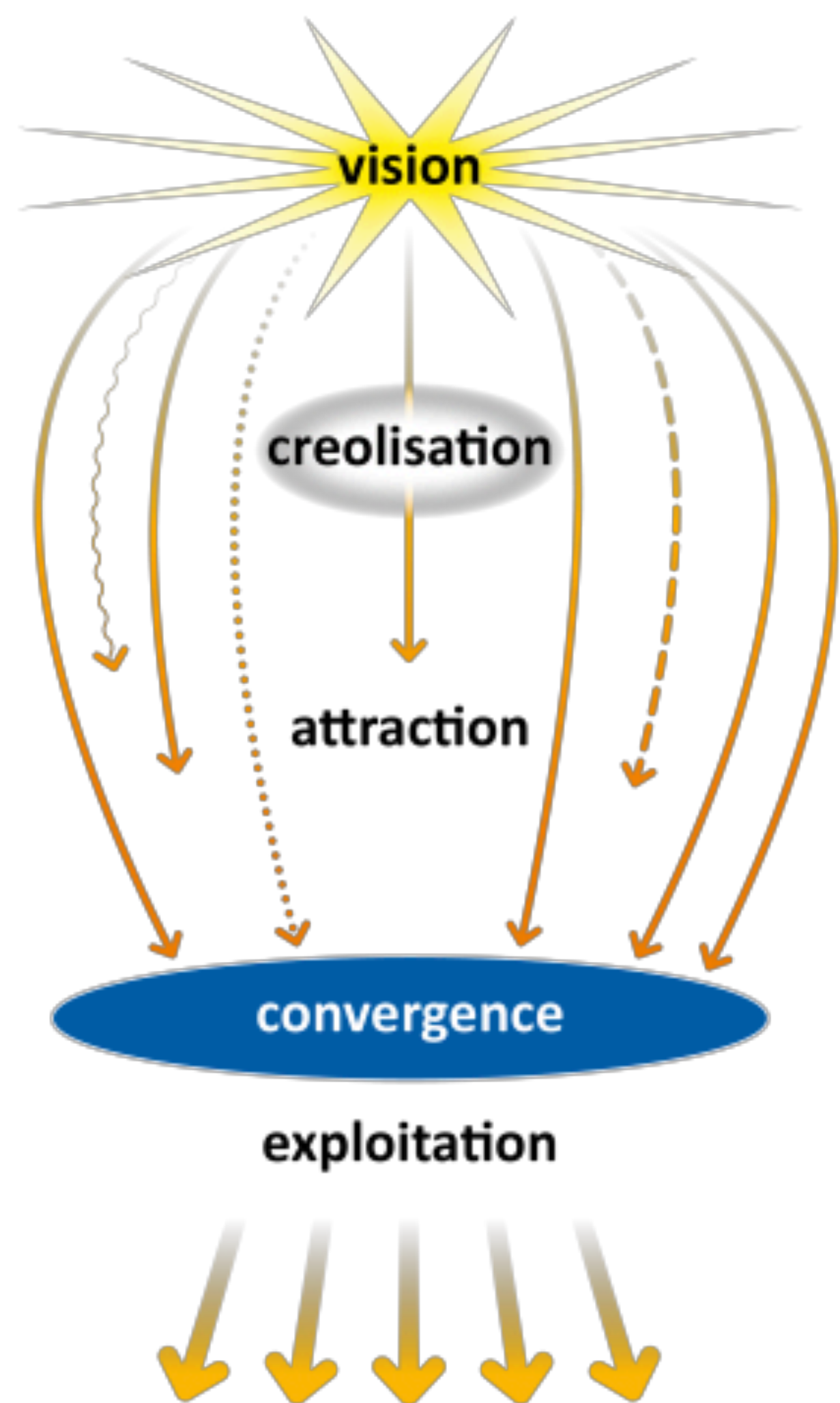
Prototypes &
over-specified solutions



Generalization

Influential early movers
(Investments)

Hourglass Architecture of the Data-Internet?

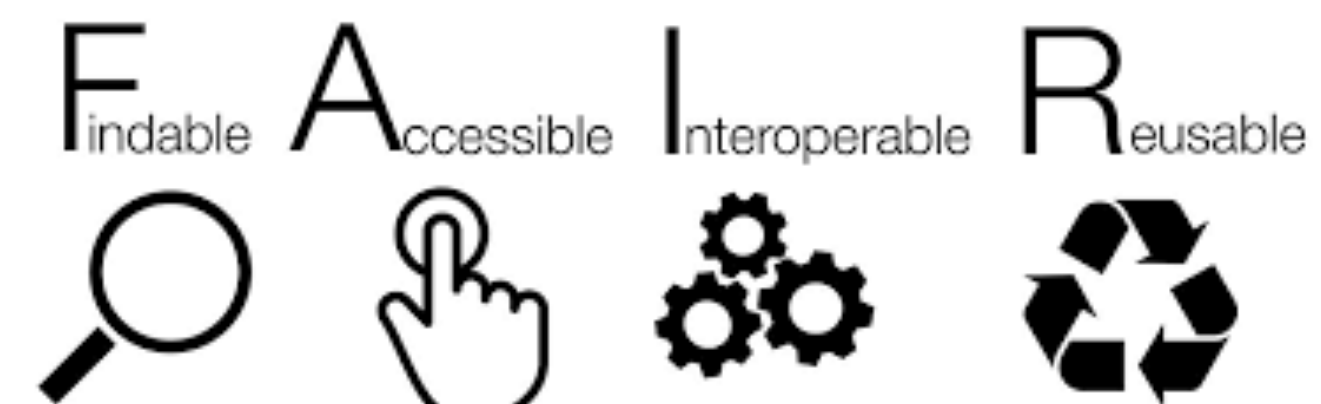


Prototypes & over-specified solutions

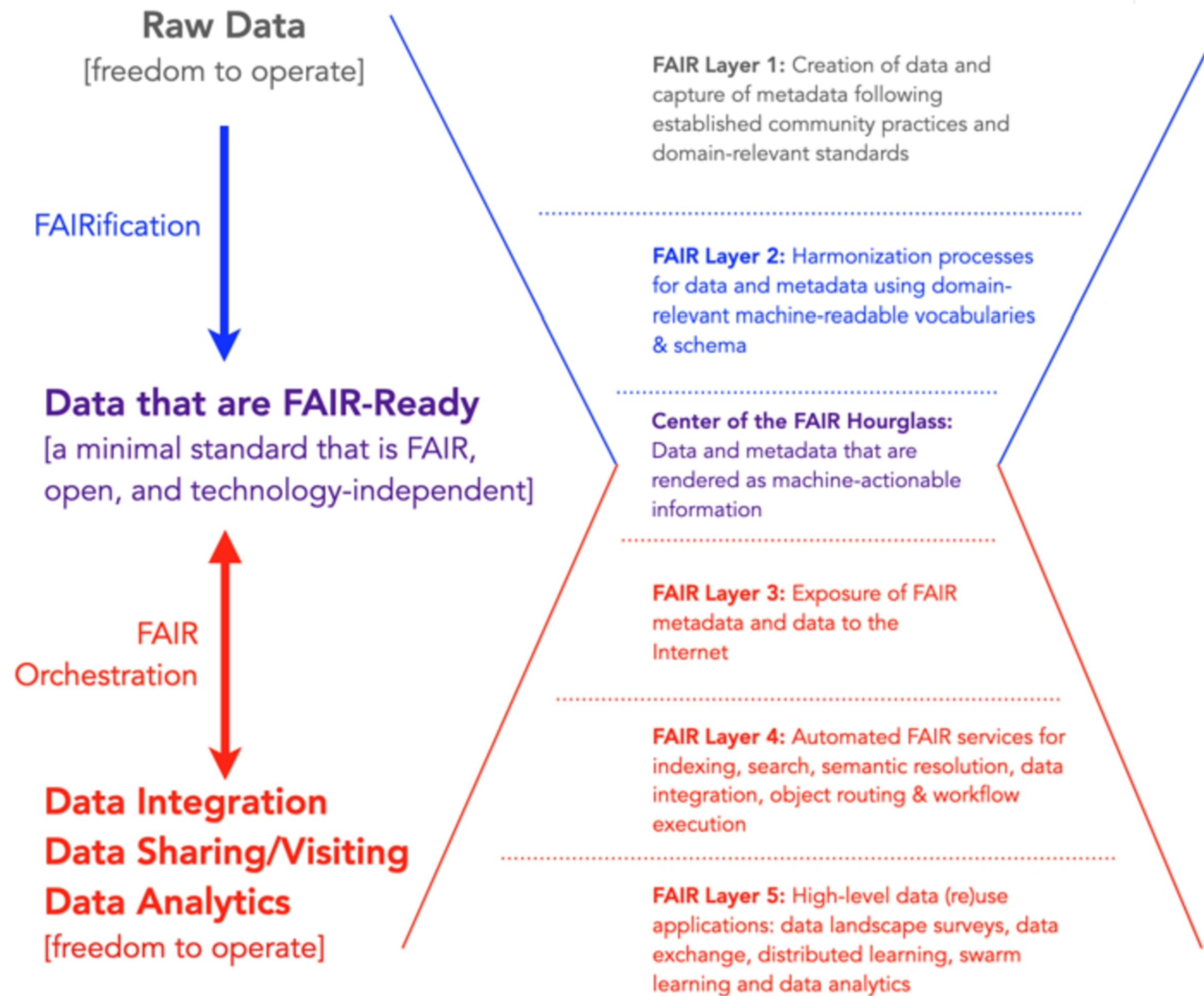


Generalization

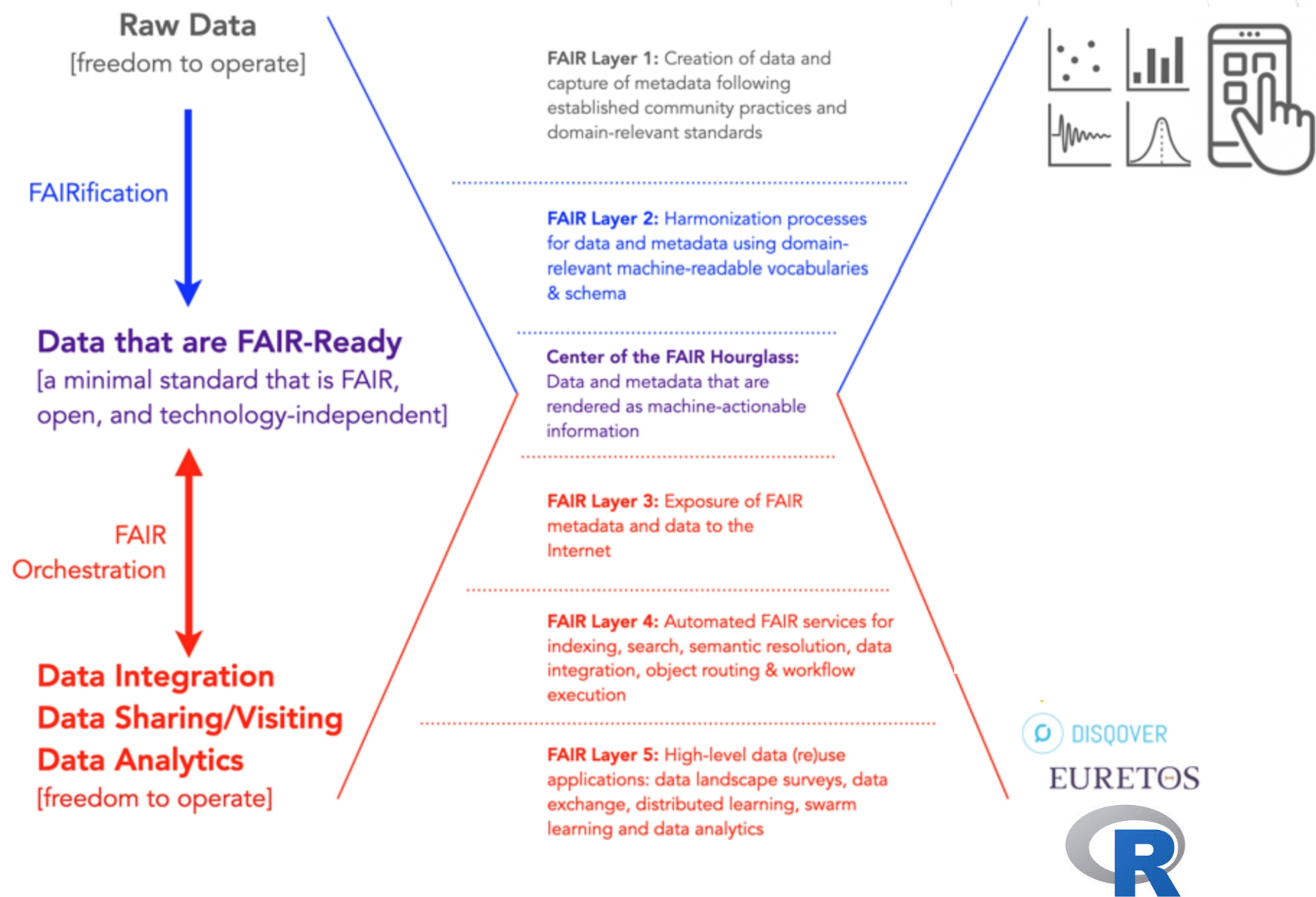
Influential early movers (Investments)



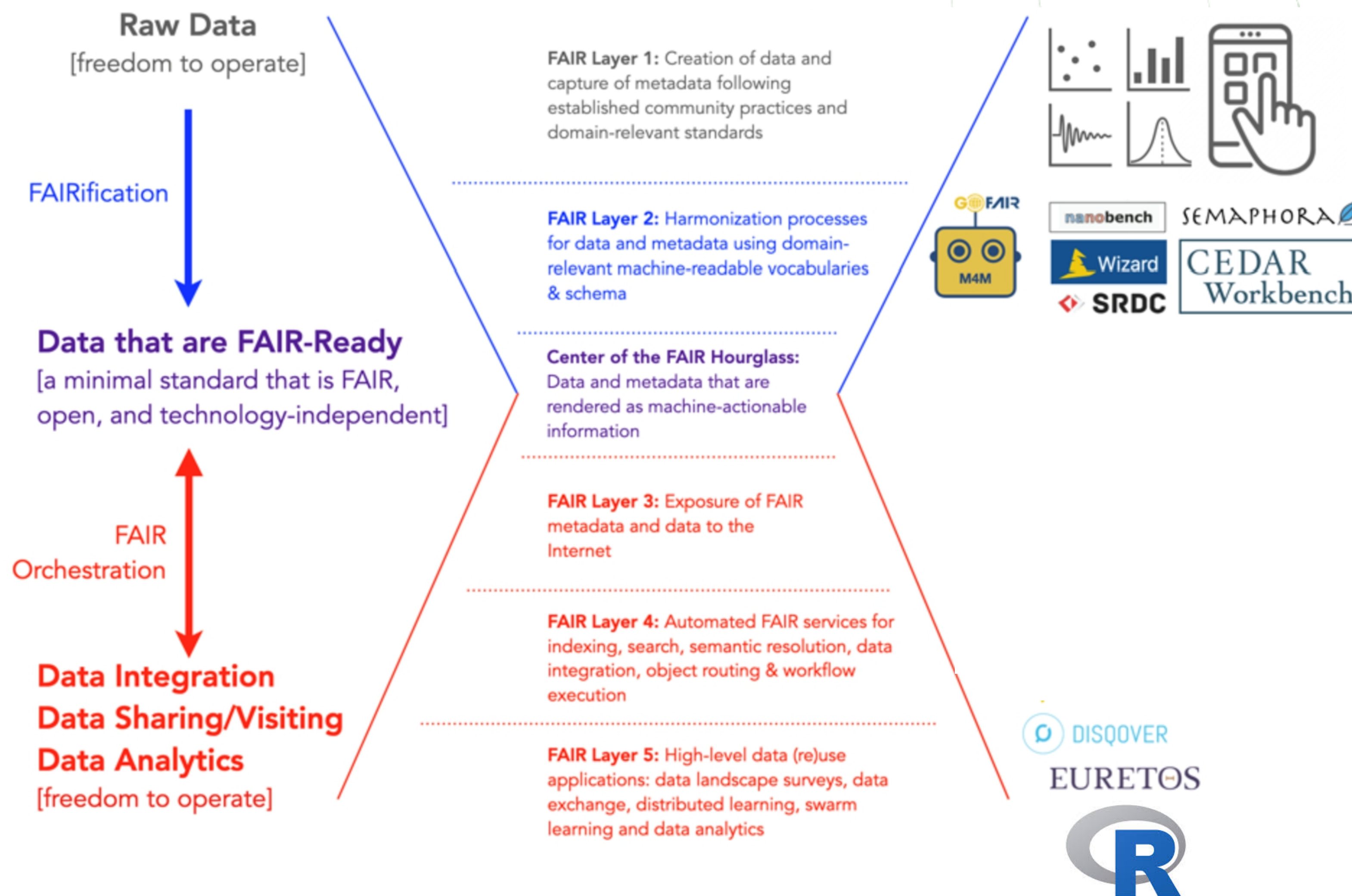
Hourglass Architecture of the Data-Internet?



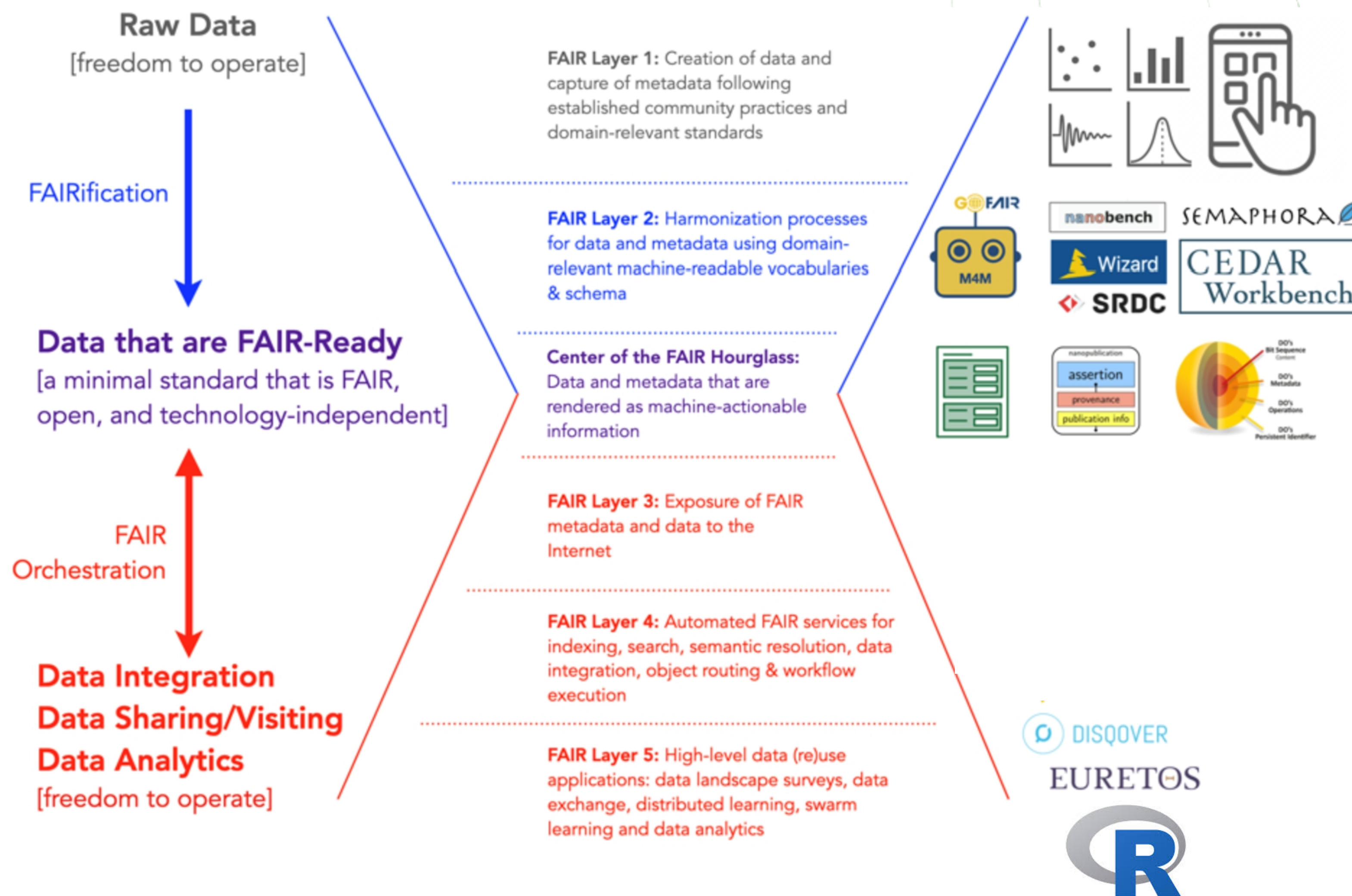
Hourglass Architecture of the Data-Internet?



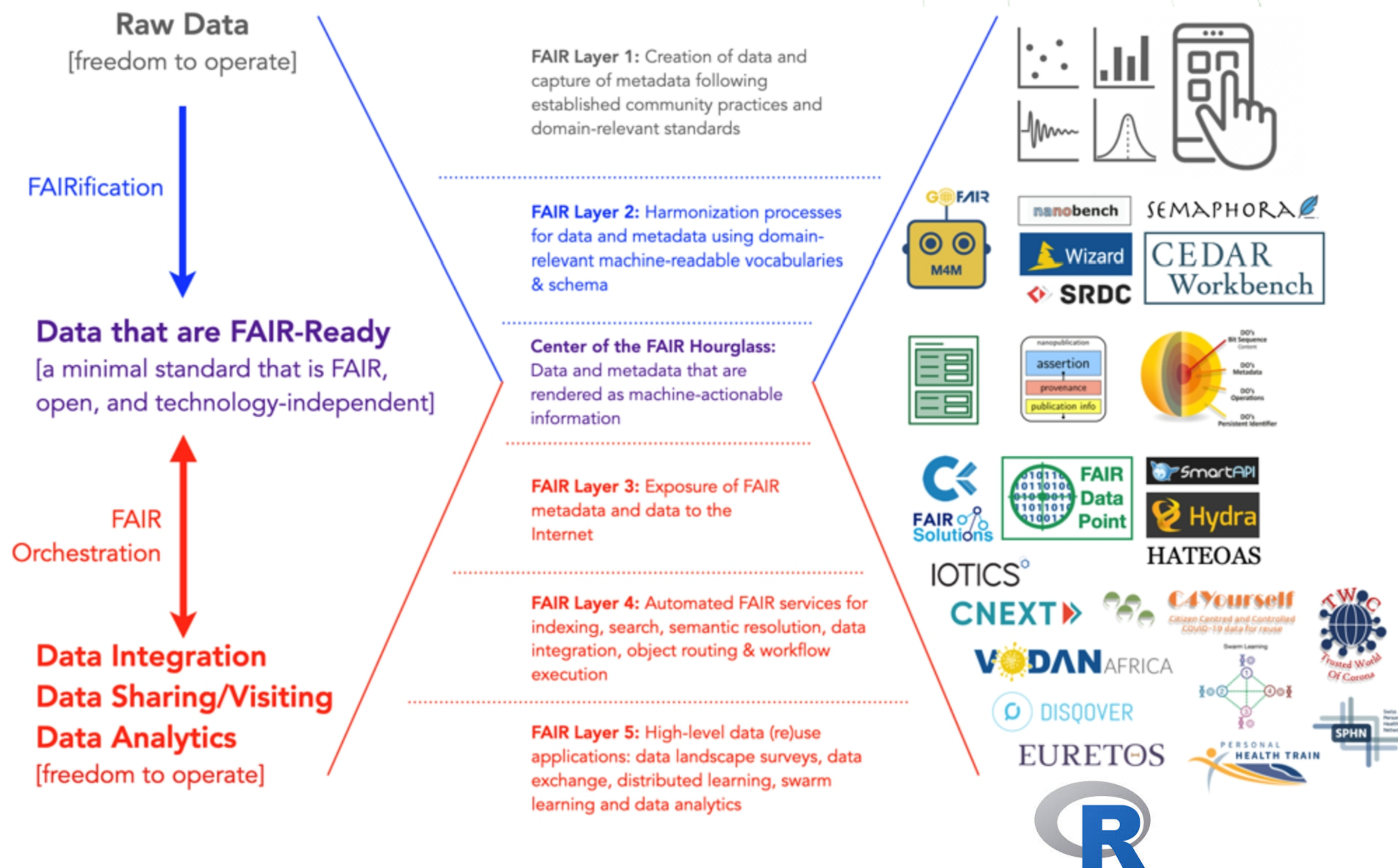
Hourglass Architecture of the Data-Internet?



Hourglass Architecture of the Data-Internet?



Hourglass Architecture of the Data-Internet?



Hourglass Architecture of the Data-Internet?



The FAIR hourglass: A framework for FAIR implementation Cite

Article type: Research Article
Authors: Schultes, Erik

Affiliations: FAIR Implementation Lead, GO FAIR Foundation, Poortgebouw Noord, Rijnsburgerweg 10, 2333 AA Leiden, The Netherlands

Correspondence: [*] Corresponding author. E-mail: eriks@gofair.foundation.

Keywords: Data management, data stewardship, FAIR principle, knowledge of theories underlying fair implementation, fair evaluation of repositories for data deposition

DOI: 10.3233/FC-221514

Journal: [FAIR Connect](#), vol. 1, no. 1, pp. 13-17, 2023

Received 30 November 2022 | **Accepted** 16 December 2022 | **Published:** 9 January 2023

[Get PDF](#)

Abstract

The FAIR Hourglass provides a framework to organize two general phases of FAIR implementation: FAIRification (top) and FAIR Orchestration (bottom). The center of the hourglass represents the use of widely agreed-upon open, minimal standards ensuring machine-actionability. Since the FAIR Guiding Principles were published in 2016 [17], there has been a diverse deployment of methods, workflows and technologies towards their instantiation: Bring Your Own Data events, the FAIR Cookbook, the RDM Kit, RO-Crates, workflows based on Jupyter Notebooks, FAIR Digital Objects of various sorts and incremental upgrades to existing repositories. Regardless of method and technology, FAIR implementation has a number of universal features. Taking inspiration from the “hourglass” architecture of the Internet [18] these common implementation activities can be organized into a “FAIR hourglass”, providing a framework making approaches to FAIR implementation more systematic, reproducible, and scalable.

<https://content.iospress.com/articles/fair-connect/fc221514>

The image shows a screenshot of a Vimeo video player. The video title is "231025_557_The FAIR hourglass: A framework for the practical implementation of FAIR data and services.mp4". The video has 26 views and was uploaded 1 year ago. The video content features a "Welcome!" message and a circular graphic for "SciDataCon 2023" organized by CODATA and the World Data System, as part of the "International Data Week" (A Festival of Data) from October 23-26, 2023, in Salzburg. The video player includes a progress bar, volume control, and a "Join for free" button.

<https://vimeo.com/879788085>



Discussion

15:30-15:45

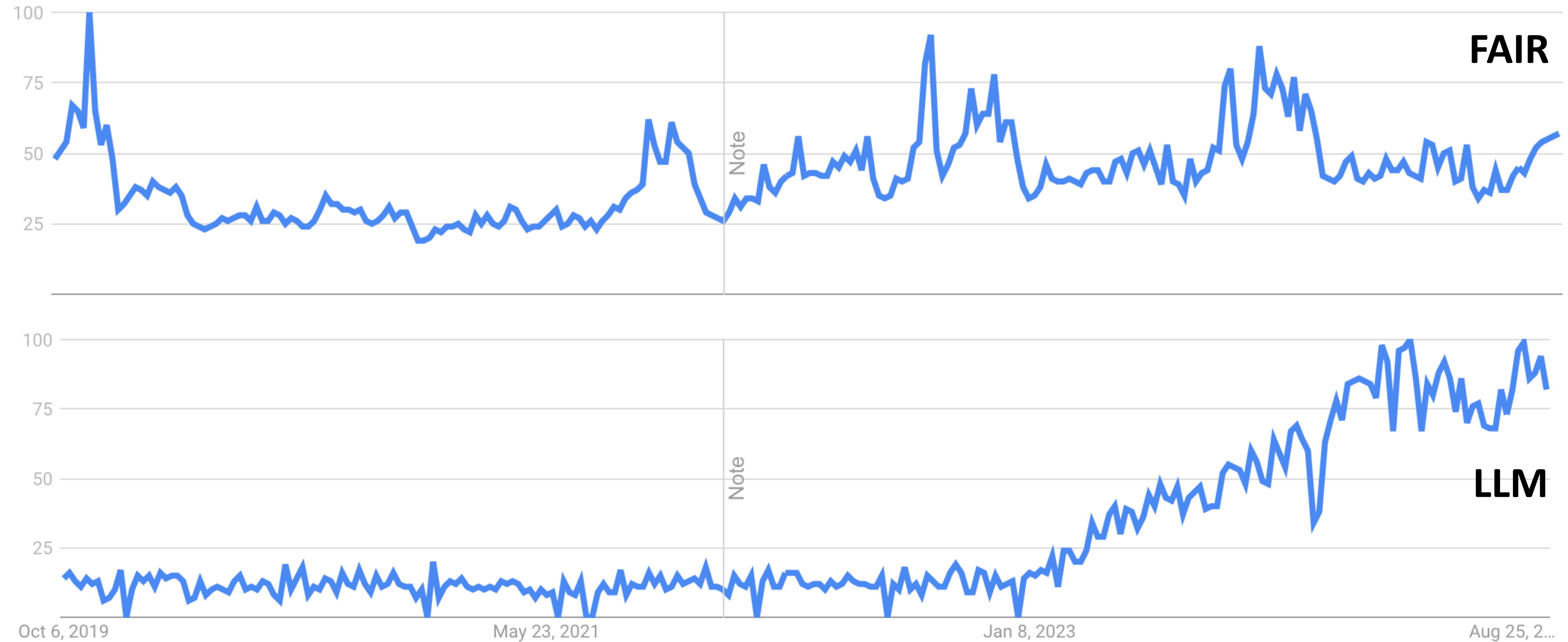


FAIR for AI

AI for FAIR

15:45-16:10
(Erik)

Trending: AI (especailly LLMs)



Trending: AI (especailly LLMs)

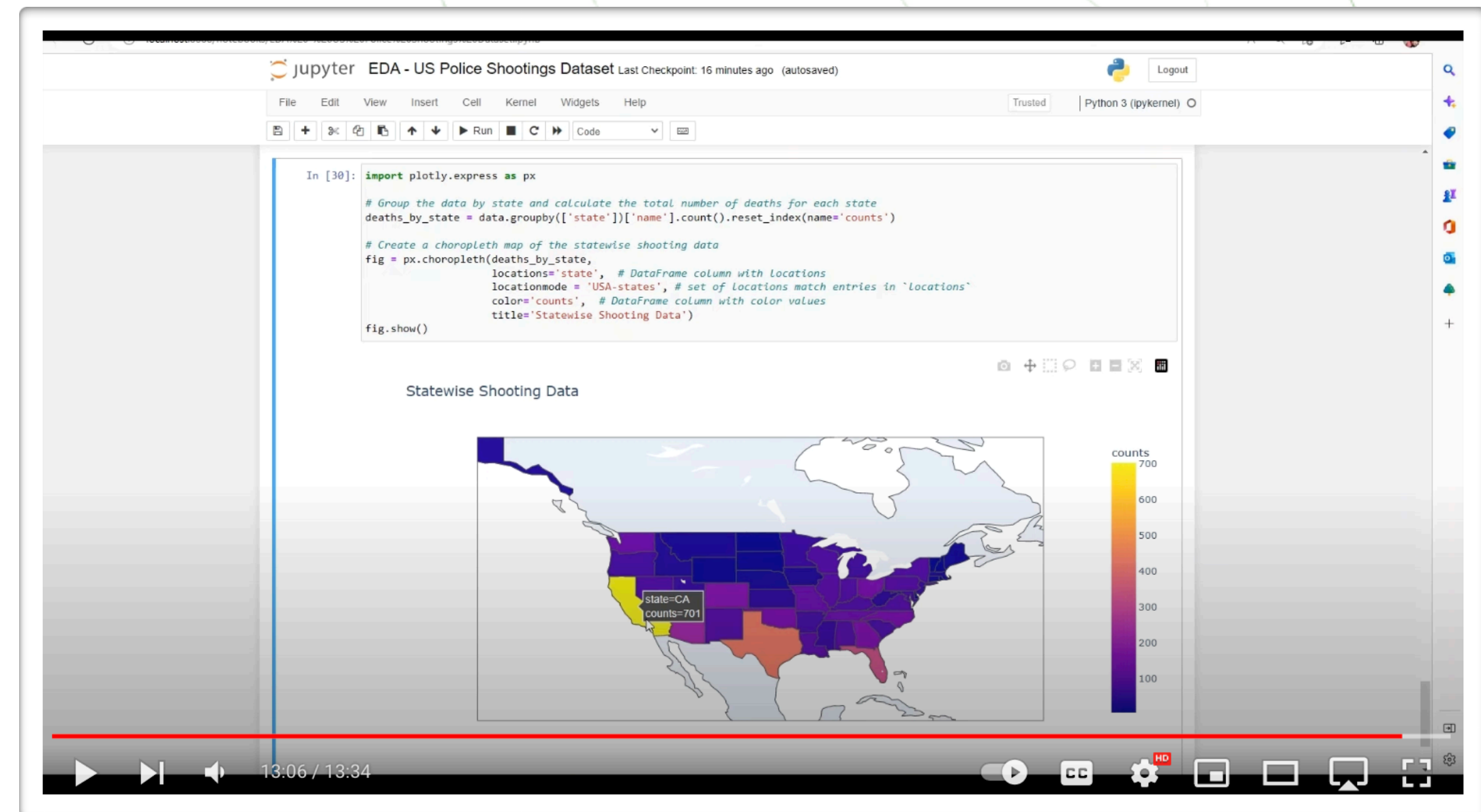
ChatGPT + Noteable (Jupyter) = Mind-blowing!

<https://www.youtube.com/watch?v=otk2AA6hGgU>



I asked ChatGPT to do Exploratory Data Analysis with Visualizations

<https://www.youtube.com/watch?v=787VOGoqBzk>



Trending: AI (especailly LLMs)



A new 'AI scientist' can write science papers without any human input. Here's why that's a problem

Published: August 20, 2024 10.29pm CEST

Wes Cockx & Google DeepMind / Better Images of AI, CC BY

Email X (Twitter) Facebook 260 LinkedIn Print

Scientific discovery is one of the most sophisticated human activities. First, scientists must understand the existing knowledge and identify a significant gap. Next, they must formulate a research question and design and conduct an experiment in pursuit of an answer. Then, they must analyse and interpret the results of the experiment, which may raise yet another research question.

Can a process this complex be automated? Last week, [Sakana AI Labs announced](#) the creation of an "AI scientist" – an artificial intelligence system they claim can make scientific discoveries in the area of machine learning in a fully automated way.

Author



Karin Verspoor
Dean, School of Comp Technologies, RMIT University

Disclosure statement

Karin Verspoor receives funding from Australian Research Council, the M Research Future Fund, the National Medical Research Council, and BV. She is affiliated with BioGrid Australia. She is a co-founder of the Australian Alliance for Artificial Intelligence in Healthcare.

<https://theconversation.com/a-new-ai-scientist-can-write-science-papers-without-any-human-input-heres-why-thats-a-problem-237029>



Trending: AI (especailly LLMs)

Why New AI Tools Like ChatGPT Need Knowledge Graphs

Irene Polikoff

Blog > Guest Posts

Why should you combine ChatGPT with Knowledge Graphs?

This post is republished by permission of ONLIM and originally appeared at their website. ONLIM is part of our partner ecosystem that powers the next generation of content and data management applications for many of the world's most knowledge-intensive enterprises – all the way from Life Sciences and Financial Services to Publishing, Government and Industry.

March 2, 2023 4 mins. read Marc Isop

[in](#) [t](#) [f](#)

Search Medium

Sign up Sign In

Write

LLM's Closing the KG Gap

Dean Allemang · Follow

12 min read · Mar 23

Documentation API reference

Sign Up

- Topics
- More
- RESOURCES
 - Documentation
 - API reference
 - Help center
- CATEGORIES
 - Announcements
 - API
 - Plugin Development

Welcome to the OpenAI Developer Forum!

What to know before posting a new question:

1. Search the forum for similar topics - the question might have been discussed before.
2. If the question relates account issues (e.g., billing and login issues), please contact us through our [Help Center](#)
3. Please be kind and helpful in conversations!

Introducing OntoGPT: A toolkit for working with ontologies, knowledge bases, and GPT

Community api

cjmungall

Jun 5

BY NC SA

Trending: AI (especailly LLMs)

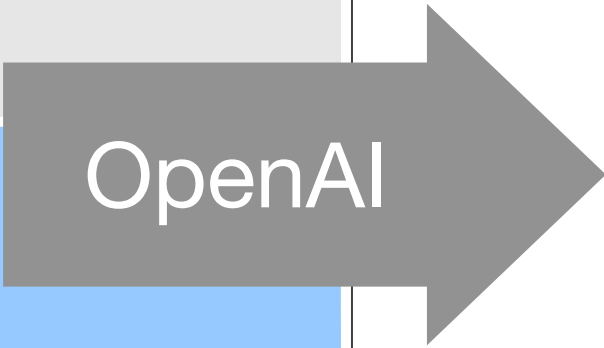
```
@prefix this: <http://purl.org/np/RAanXiiQmSiaZU7mXmSu6Xhaj01TRckSoiFh14cE50YtM> .
@prefix sub: <http://purl.org/np/RAanXiiQmSiaZU7mXmSu6Xhaj01TRckSoiFh14cE50YtM#> .
@prefix np: <http://www.nanopub.org/nschema#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix orcid: <https://orcid.org/> .
@prefix dcat: <https://www.w3.org/ns/dcat#> .
@prefix fip: <https://w3id.org/fair/fip/terms/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix npx: <http://purl.org/nanopub/x/> .

sub:Head {
  this: np:hasAssertion sub:assertion ;
  np:hasProvenance sub:provenance ;
  np:hasPublicationInfo sub:pubinfo ;
  a np:Nanopublication .
}

sub:assertion {
  sub:declaration a fip:FIP-Declaration ;
  fip:considerations "no actions made here, implicit in internet access." ;
  fip:declared-by <http://purl.org/np/RAge51gbVKMxb0zu5CUpFaCiwg71UXa_5QGHrW68zaq2g#SeaDataNet-CDI> ;
  fip:declares-current-use-of <http://purl.org/np/RAm6j12F_ysCwqHyEEJjb7vdEnZXMtNyIXSjsbNU3In8#HTTP> ;
  fip:refers-to-question fip:FIP-Question-A1.1-D ;
  dcat:endDate "2022-12-31"^^xsd:date ;
  dcat:startDate "2022-01-01"^^xsd:date .
}

sub:provenance {
  sub:assertion dct:creator orcid:0000-0001-9214-3217 .
}

sub:pubinfo {
  sub:signature npx:hasAlgorithm "RSA" ;
  npx:hasPublicKey
  "MIGFMA0GCsGSIb3DQEBAQUAA4GNADCBiQKgQCokFb7Lk1Xy8B7BuvjFF9b0JNk0WxH3Xerr4H3v/VCIrVshED13B2VekbMMdRe+T0Pyc8A0Ld9IAL2VFVVD/T0mmChk7Pw7DQjI96mweXhE7f4g0WGSUT5Pxxmof
  ;
  npx:hasSignature
  "ViYFZ+FRkHrh4JablPS8DFoWZjaNwLHZRYuU4Qa3AIa5GBV8LVkJPFPg7p8/xz6fWw0Jzpej8R19ETamb6n9Tvo5QwApqcoI0ZnpIGfe9b9gpS1mXx+UkHY/gS68sIS10+YYed5oLWXeGJFEqGiHk4qi/m9DrLi
  ;
  npx:hasSignatureTarget this: .
  this: dct:created "2023-04-25T04:51:06Z"^^xsd:dateTime ;
  dct:creator <https://fip-wizard.ds-wizard.org> ;
  dct:license <https://creativecommons.org/publicdomain/zero/1.0/> ;
  prov:wasDerivedFrom <https://w3id.org/fip/wizard/95145c3b-6d96-444f-a99f-fc5acabb6d38> .
}
```



- This nanopublication contains a **declaration** that the entity **SeaDataNet-CDI** is currently using the technology **HTTP** with no additional **considerations**, for the **period** January 1, 2022 to December 31, 2022.
- The declaration refers to the FAIR self-assessment questionnaire, specifically question **A1.1-D**.
- The assertion was created by the **ORCID** <https://orcid.org/0000-0001-9214-3217>.
- The full nanopublication was **created on** April 25, 2023 by the **FIP Wizard** at <https://fip-wizard.ds-wizard.org>.
- It is **licensed** under CC0 public domain dedication.
- The nanopublication was derived from the FIP Wizard **session** with id [95145c3b-6d96-444f-a99f-fc5acabb6d38](https://w3id.org/fip/wizard/95145c3b-6d96-444f-a99f-fc5acabb6d38).

In summary, this nanopublication records the declaration that SeaDataNet-CDI uses HTTP technology for the year 2022, linking this to the FAIR self-assessment.

Thanks: Rado Vleugel



The original idea behind FAIR

...the FAIR Principles put specific emphasis on **enhancing the ability of machines to automatically find and use the data**, in addition to supporting its reuse by individuals.

<https://www.nature.com/articles/sdata201618>

Computability

- *FAIR data* use knowledge representation languages and controlled vocabularies that reduce/eliminate ambiguity.

Trustworthiness

- *FAIR data* have (rich) provenance which provides evidence for the source of data. Provenance can include uncertainties and error estimation. Trust also requires large amounts of metadata.

Equitability

- *FAIR data* make explicit the conditions for reuse. Equitability can be technologically ensured, when data and metadata are FAIR

The original idea behind FAIR

FAIR

Find me **all** known **low molecular weight inhibitors of the Human p65 Protein**. Separate the list based on those that were **found in curated databases**, from those that were **found in self-deposited data archives**. Keep track of the **license and citation information** for each one. If data is relevant, but **not public**, please provide the **contact information for the person** I need so I can request the data.

Mark had imagined agents that would keep track of their journey and the metadata associated with each successful data discovery (a problem currently under investigation in an EOSC task force co-chaired by Mark).

What should be returned:

- Lists of URLs pointing to inhibitors
- Reuse information.

*Hypothetical search application on FAIR data.
Mark Wilkinson, CBGP-UPM/INIA, Madrid
2017*

The original idea behind FAIR?

Google inhibitors p65

All Images Shopping Videos News More Settings Tools

About 544.000 results (0,51 seconds)

RelA/NFkB p65 Inhibitors: Novus Biologicals
<https://www.novusbio.com/inhibitors/rela-nfkb-p65>
RelA/NFkB p65 Inhibitors available through Novus Biologicals. Browse our RelA/ NFkB p65 Inhibitor catalog backed by our Guarantee+.

Inhibiting NF-kB Activation by Small Molecules As a Therapeutic ...
<https://www.ncbi.nlm.nih.gov/NCBI/Literature/PubMedCentral/PMC>
by SC Gupta - 2010 - Cited by 345 - Related articles
May 21, 2010 - ... pathway (Fig 3). Table 1. A list of small molecules as inhibitors of NF-kB pathway Blocking NF-kB activation by inhibitors of p65 acetylation.

Suppression of p65 phosphorylation coincides with inhibition of ... - NCBI
<https://www.ncbi.nlm.nih.gov/pubmed/16163708>
by J Hu - 2005 - Cited by 26 - Related articles
Mol Carcinog. 2005 Dec;44(4):274-84. Suppression of p65 phosphorylation coincides with inhibition of IkappaBalpha polyubiquitination and degradation.

Identification of a p65 peptide that selectively inhibits NF-kappa B ...
<https://www.ncbi.nlm.nih.gov/pubmed/14711835>
by Y Takada - 2004 - Cited by 157 - Related articles
Jan 7, 2004 - Identification of a p65 peptide that selectively inhibits NF-kappa B activation induced by various inflammatory stimuli and its role in ...

NF-kB - Wikipedia
<https://en.wikipedia.org/wiki/NF-kB>
NF-kB is a protein complex that controls transcription of DNA, cytokine production and cell Concerning known protein inhibitors of NF-kB activity, one of them is IFRD1, which represses the activity of NF-kB p65 by enhancing the HDAC- mediated deacetylation of the p65 subunit at lysine 310, by favoring the recruitment of ...

2017

Google inhibitors p65

Last updated Download format Usage rights Topic Provider Free Saved datasets

56 datasets found

BIOGRID CURATED DATA FOR PUBLICATION: Inhibitor of...
thebiogrid.org
zip
Updated Apr 29, 2013

Genome-wide TNFM-NM-1-induced p65 binding before...
ebi.ac.uk
Updated Nov 25, 2012

BIOGRID CURATED DATA FOR PUBLICATION: Transcriptional...
thebiogrid.org
zip
Updated Mar 17, 2003

SIRT1 Activators Suppress Inflammatory Responses...
plos.figshare.com
tiff
Updated May 31, 2023

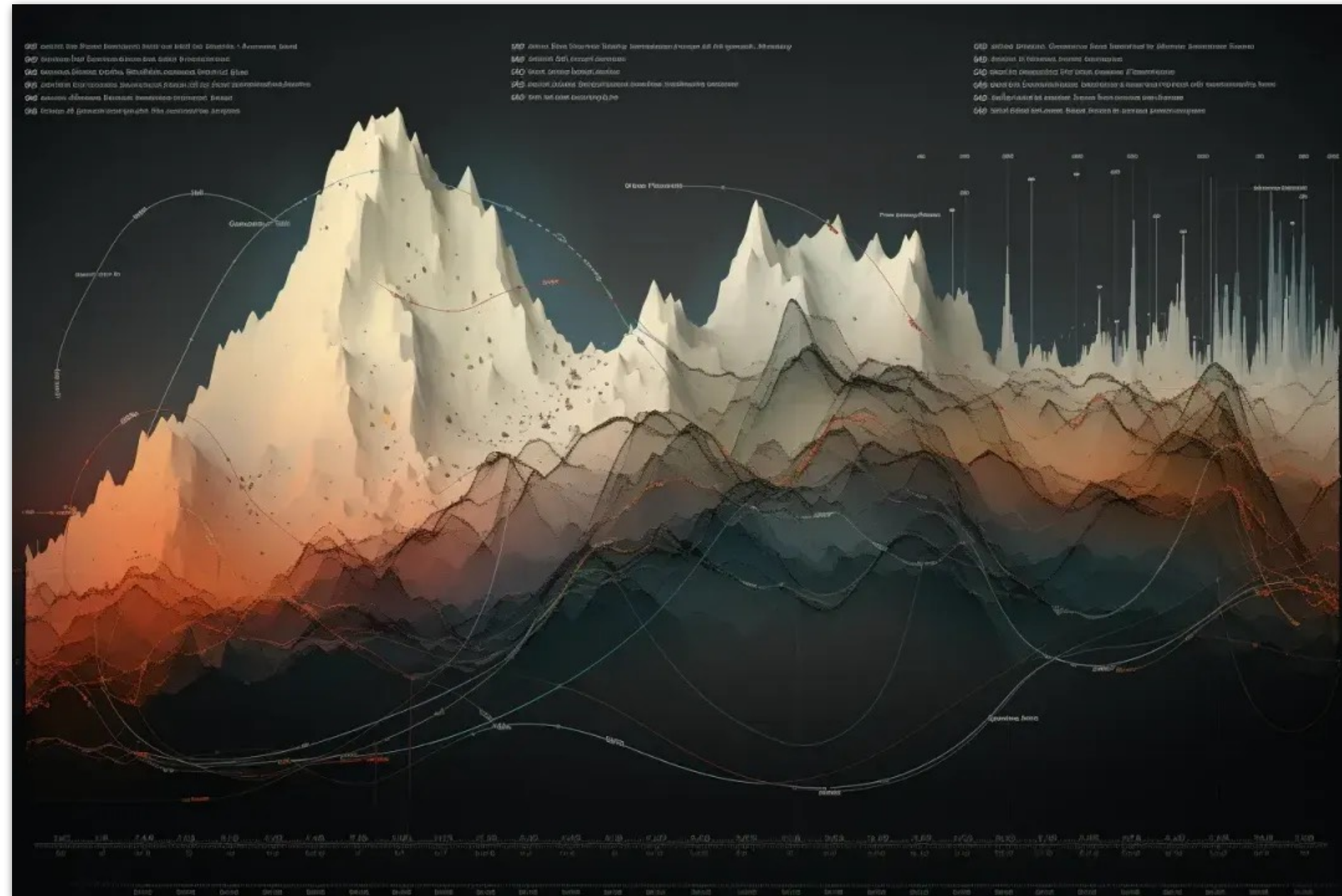
BIOGRID CURATED DATA FOR PUBLICATION: Inhibitor of growth 4 induces NF κ B/p65 ubiquitin-dependent degradation.
Related Article
Explore at: thebiogrid.org
zip
Dataset updated Apr 29, 2013
Dataset authored and provided by BioGRID Project
License MIT License
License information was derived automatically
Description
Protein-Protein, Genetic, and Chemical Interactions for Hou Y (2013):Inhibitor of growth 4 induces NF κ B/p65 ubiquitin-dependent degradation. curated by BioGRID (<https://thebiogrid.org>); ABSTRACT: As a tumor suppressor protein, the inhibitor of growth 4 (ING4) has an important role in many cellular processes including cell cycle progression, proliferation, apoptosis, DNA damage response, tumor angiogenesis and contact inhibition. Here, we report that ING4 functions as an E3 ubiquitin ligase to induce nuclear factor- κ B (NF κ B)/p65 degradation. The plant homeodomain finger of ING4 interacted with p65 to undergo robust ubiquitination and degradation. ING4 bound to p65 and delivered the Lys-48-linked polyubiquitin to Lys-62 residue of p65, leading to ubiquitination of p65 and degradation. Lys-62 residue of p65 was required for ING4-mediated ubiquitination of p65 and degradation. Further analysis shows that C239 of ING4 was critical for ING4-induced p65 degradation.

2020

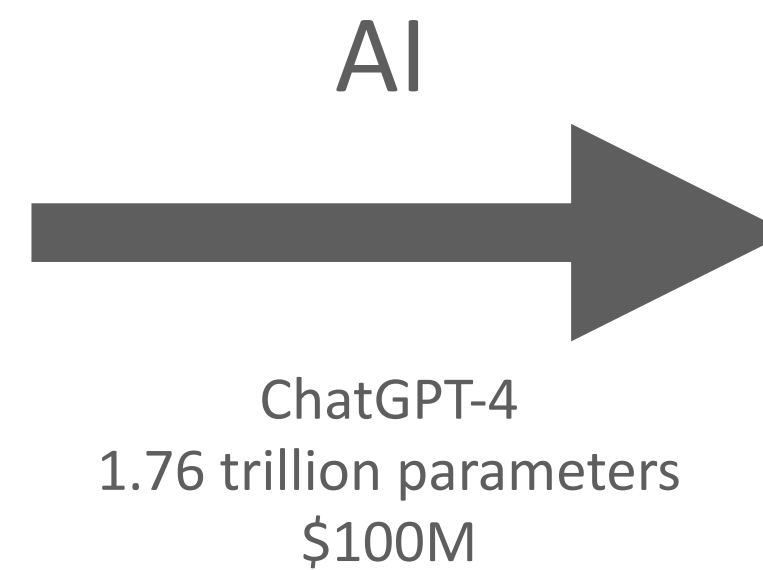


LLMs

Existing data input to train/validate AI



<https://commoncrawl.org/blog/host--and-domain-level-web-graphs-july-august-and-september-2024>



New data output from AI



https://miro.medium.com/v2/resize:fit:1400/format:webp/1*B7Npyd_WSEWSiJeQt4WrZA.jpeg

1. The Web: websites, articles, blogs, forums
2. Books: e.g., Project Gutenberg
3. Academic Papers: e.g., arXiv, PubMed
4. Reference Works: e.g., Wikipedia
5. Code Repositories: e.g., GitHub
6. Social Media: e.g., Reddit and social media platforms
7. News Articles
8. Government and Legal Texts: Laws, regulations, court rulings, governmental reports

Content generated using OpenAI's ChatGPT-4

LLMs can output in both human- & machine-actionable formats

1. Text
2. Knowledge graphs
3. Code
4. Images
5. Audio / Video
6. ...



LLMs

Question 1: ChatGPT-4 answers Mark Wilkinson's query from 2017...



ChatGPT follows a complex, natural language command.

ChatGPT resolves ambiguity:

- *p65* - RelA, NF-κB
- *Inhibit* - target, modulate

Reality check:

It seems that Chat is good at identifying potential data sources, but isn't going much further than that... APIs will still be pretty opaque to an LLM! Data needs to be well-modeled and linked.

Finding known **low molecular weight inhibitors of the Human p65 protein** (also known as RelA, part of the NF-κB complex) can be sourced from curated databases and self-deposited data archives.

Curated Databases:

1. **ChEMBL**: A well-established database of bioactive molecules, ChEMBL includes several small molecule **inhibitors** targeting p65/RelA. These molecules are indexed based on their interaction data, activity, and relevance in drug discovery. You can access and download compound data under an open license from the ChEMBL site (EMBL-EBI).
2. **PubChem**: This public resource houses chemical compounds, including those with **inhibitory** action on p65. Detailed molecular properties, including structure and bioactivity, are available. Licensing for data is open for academic use.
3. **SMMDB**: The SMMDB (Small Molecule Modulator Database) contains detailed information on small molecules **modulating** targets like p65 in neurological and other diseases. It is a manually curated resource, providing PubMed links and compound details, including molecular weight, 3D structures, and inhibition data **[15†source]** **[18†source]** .

Self-deposited Data Archives:

1. **ZINC Database**: This includes user-submitted and curated molecules, many of which **target** NF-κB, including p65. Researchers can deposit their findings here, making it a mixture of community-contributed data. Usage of the database is open, but some entries may require contacting the depositor for further licensing details.
2. **IPAD-DB**: Although primarily focused on Alzheimer's disease, the IPAD-DB contains data on molecules that **modulate** related inflammatory pathways, including NF-κB. Licensing details vary, but access is generally open **[17†source]** .

Citation and License Information:

- Most curated databases like **ChEMBL** and **PubChem** offer open access with clear academic use licenses. **SMMDB** is similarly freely accessible for non-commercial research, while data in **ZINC** may have different licensing based on individual submissions.
- For direct inquiries or requesting non-public data, each database typically provides the depositor's contact, especially in **ZINC** and **IPAD-DB**, where self-deposited entries may have restricted access.

For the most comprehensive set of inhibitors targeting p65, I recommend starting with **ChEMBL** and **PubChem** for licensed, reviewed data, and exploring **ZINC** or **IPAD-DB** for community-deposited resources.



LLMs

Question 2: Is ChatGPT-4 FAIR?



I dialogued with ChatGPT-4 to assess the FAIRness of the tool, much as we do with many other resources in GO FAIR.

Note: in this capacity, ChatGPT-4 is both a resource to be evaluated, as well as an automated FAIR assessment tool.

Reality check:

The answer for F1 fails the relevance test fairly quickly. A lot of detail but not answering your question. F2 tends to do the same thing, listing criteria but not always responding to them, instead just defining them.

<https://docs.google.com/document/d/163LxY5lg6s2U0EOF2U2Pg8oivXDiVneUjZ97tXm-NCc/edit?usp=sharing>

ChatGPT assisted FAIRification of ChatGPT

Erik Schultes

28 September 2024

ChatGPT-4 was asked to evaluate itself for each of the FAIR Principles. This is the first step to creating the FIP for ChatGPT-4. The nanopublications supporting the ChatGPT-4 FIP might be created directly by the LLM, obviating the use of a FIP Wizard or NanoDash. See summaries of existing FIPs created by ChatGPT-4 in section R1.3.

Blue font indicates text that is written by a human (Erik), red font indicates text written by a machine (ChatGPT-4). Green font indicates code written by a machine (ChatGPT-4).

Much of the content herein was generated using OpenAI's ChatGPT-4. **ChatGPT** is based on the **GPT-4 architecture**. Specifically, this version is designed with enhanced conversational abilities, offering more accurate and nuanced responses compared to previous versions like GPT-3. It includes various improvements in reasoning, factual correctness, and contextual understanding. Additionally, it supports a wide range of functionalities, including generating detailed explanations, assisting with complex tasks, and offering multimodal capabilities (if enabled).

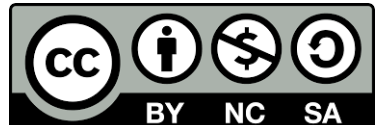
Table of Contents

Principle F1	2
Principle F2	3
Principle F3	10
Principle F4	11
Principle A1.1	11
Principle A1.2	12
Principle A2	13
Principle I1	15
Principle I2	17
Principle I3	18
Principle R1.1 & Principle R1.2	20
Principle R1.3	22



These results lay the ground work for the FAIRification of the tool itself. For example:

- I asked also for JSON-LD regarding metadata (and got it without hesitation).
- Can we use this document to build a FIP for ChatGPT-4?
- Can we ask ChatGPT-4 to FAIRify itself?
- Can we ask ChatGPT-4 to FAIRify itself, aligned with a given FIP?
- With all FIPs?



LLMs

Might we conclude that AI has solved (or obviated the need for) FAIR?

- In the web of 1.76 trillion parameters, does ChatGPT-4 contain, if only **implicitly**, all the information required to orchestrate FAIR operations on arbitrary data?
- Given the vast training and validation data sets used to create AI, has the need for **explicitly** making FAIR data been obviated?
- If the machine is acting (reliably) on data, is the data by definition machine-actionable?
- Do we invest in making data more FAIR, or do we invest in making machines more intelligent (more discerning).
- Who better to make data machine-actionable than a machine?

LLMs

Might we conclude that AI has solved (or obviated the need for) FAIR?

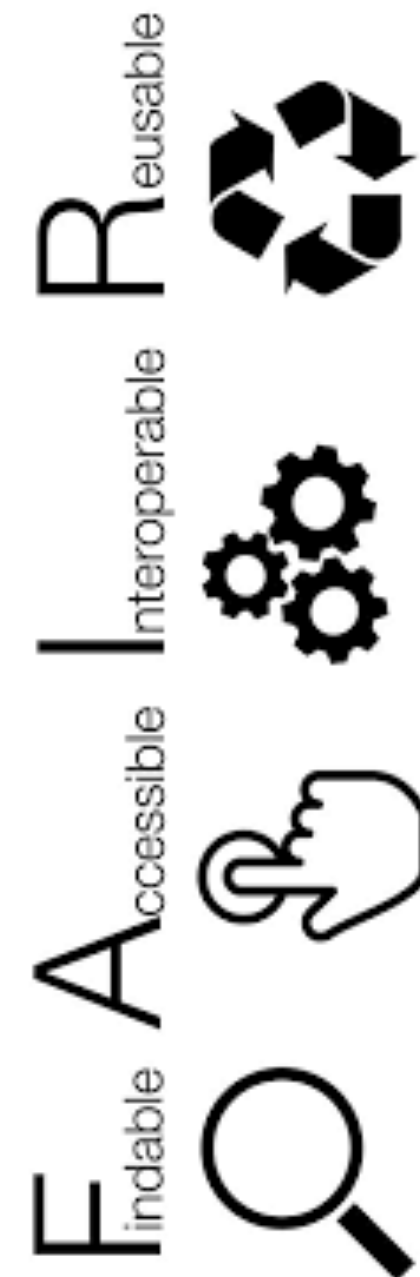
Key Considerations:

- **Filtering:** The data collected from these sources are heavily filtered to remove noise, low-quality text, and duplicates.
- **Bias and Ethical Concerns:** Training on public data can introduce biases present in the original content (e.g., social biases, regional biases), which is why researchers try to apply techniques to mitigate these issues.
- **Copyrighted Material:** While efforts are made to use publicly available content, there are ongoing debates and concerns about training on data that may come from copyrighted works, especially when it is scraped without explicit permission.

Content generated using OpenAI's ChatGPT-4

LLMs

Might we conclude that AI has solved (or obviated the need for) FAIR?



Computability

Trustworthiness

Equitability

Key Considerations:

- **Filtering:** The data collected from these sources are heavily filtered to remove noise, low-quality text, and duplicates.
- **Bias and Ethical Concerns:** Training on public data can introduce biases present in the original content (e.g., social biases, regional biases), which is why researchers try to apply techniques to mitigate these issues.
- **Copyrighted Material:** While efforts are made to use publicly available content, there are ongoing debates and concerns about training on data that may come from copyrighted works, especially when it is scraped without explicit permission.

Content generated using OpenAI's ChatGPT-4



FAIR for AI for FAIR

Legacy
Data

AI issues: Noise, Bias, Copyright

AI



FAIR for AI for FAIR

Legacy
Data

FAIRification

FAIR
Data

- Less noise
- More Trust
- More Equitably

Enhance

FAIR
AI

FAIR for AI for FAIR

Legacy
Data

FAIRification

FAIR
Data

- Less noise
- More Trust
- More Equitably

Enhance

FAIR
AI

Orchestration of FAIR data:
AI does F, A, I, and R

FAIR for AI for FAIR

Legacy Data

FAIRification

FAIR Data

- Less noise
- More Trust
- More Equitably

Enhance

FAIR AI

Orchestration of FAIR data:
AI does F, A, I, and R

Outputs data
in FAIR formats

FAIR for AI for FAIR

Legacy Data

FAIRification

FAIR Data

- Less noise
- More Trust
- More Equitably

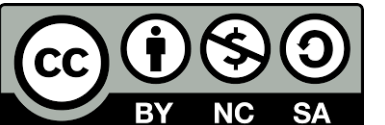
Enhance

FAIR AI

Orchestration of FAIR data:
AI does F, A, I, and R

Outputs data
in FAIR formats

Automated
FAIRification





Resume 16:20 CET

How can AI help in FAIRification

16:20-16:50
(discussion using FigJam)



Wrap up & Homework

16:50-17:00
(Erik & Barbara)





FAIR
Well!

THANKS!

Erik Schultes, Barbara Magagna, Andrea Tarallo
4 December 2024

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-
Mission 4 "Education and Research" - Component 2: "From research to business" - Investment
3.1: "Fund for the realisation of an integrated system of research and innovation infrastructures"

