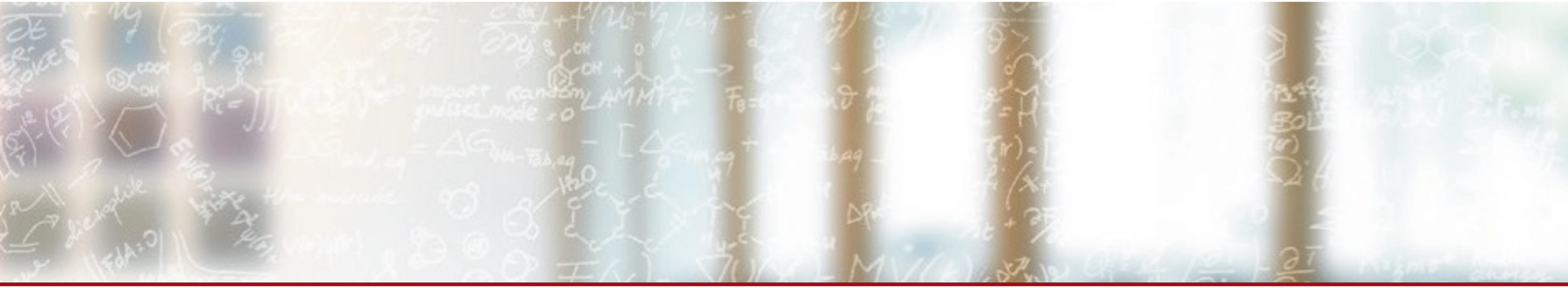
CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich



Long Term Storage Service at CSCS – An Introduction

CSCS Webinar
Mario Valle, CSCS
June 23, 2021

Webinar agenda

1. Why the Long Term Storage service is needed? (Mario Valle)
2. Technical structure of the Long Term Storage service (Giuseppe Lo Re)
3. Demo of the service (Stefano Schuppli)
4. Q&A

Volume of scientific data is not the (only) issue

Inside the CSCS tape archive

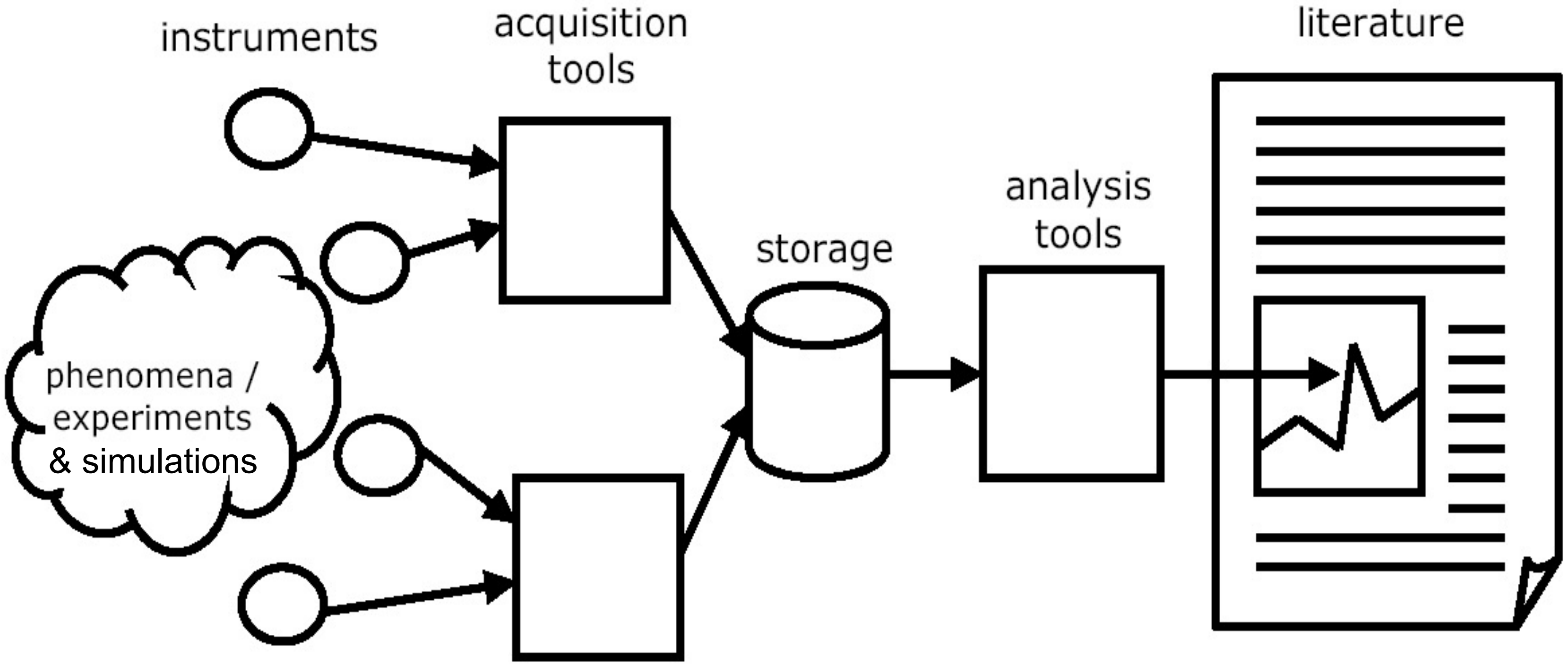
Every CSCS user remembers the Richard Hamming's admonition:

"Purpose of computation is insight, not numbers"

We produce data and more data to help our insight, but often the result is that:

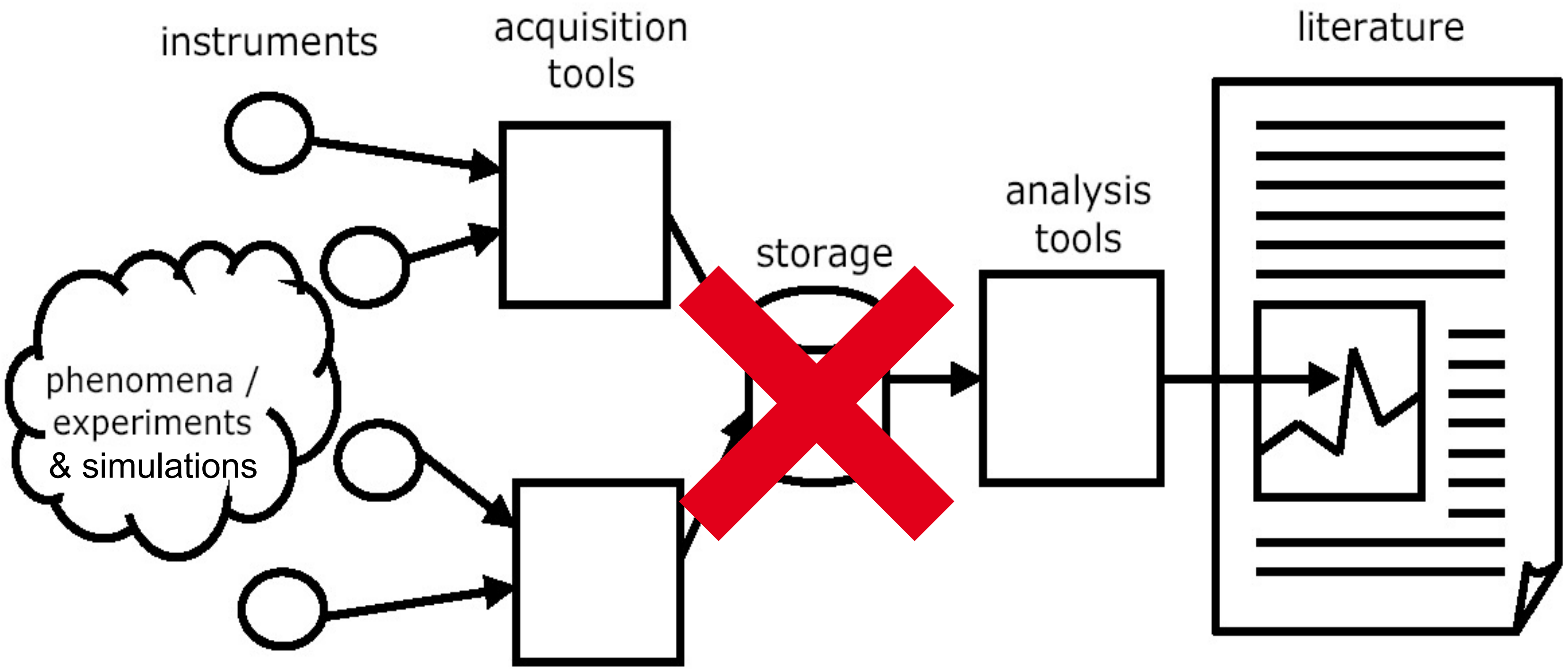
"We are drowning in data, but starving of information"

The usual scientific data lifecycle



The usual scientific data lifecycle often ends with publication

Too often, after publications, data might consciously or unconsciously be forgotten and sometimes even throw away.



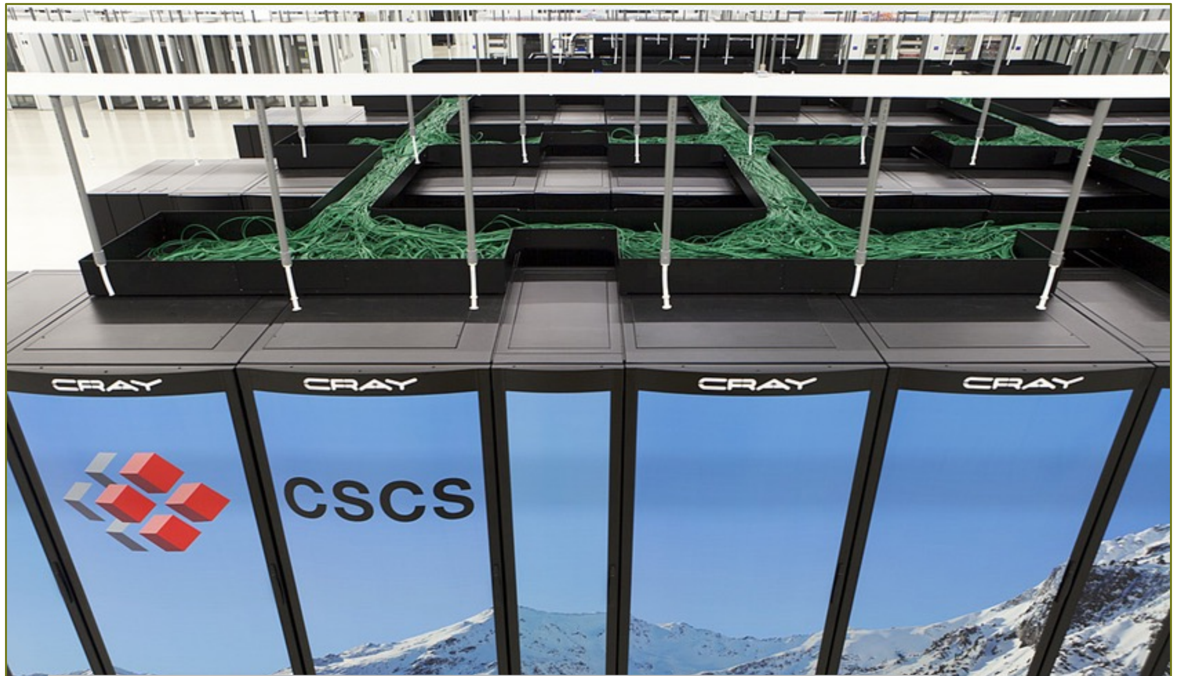
Data creation is costly

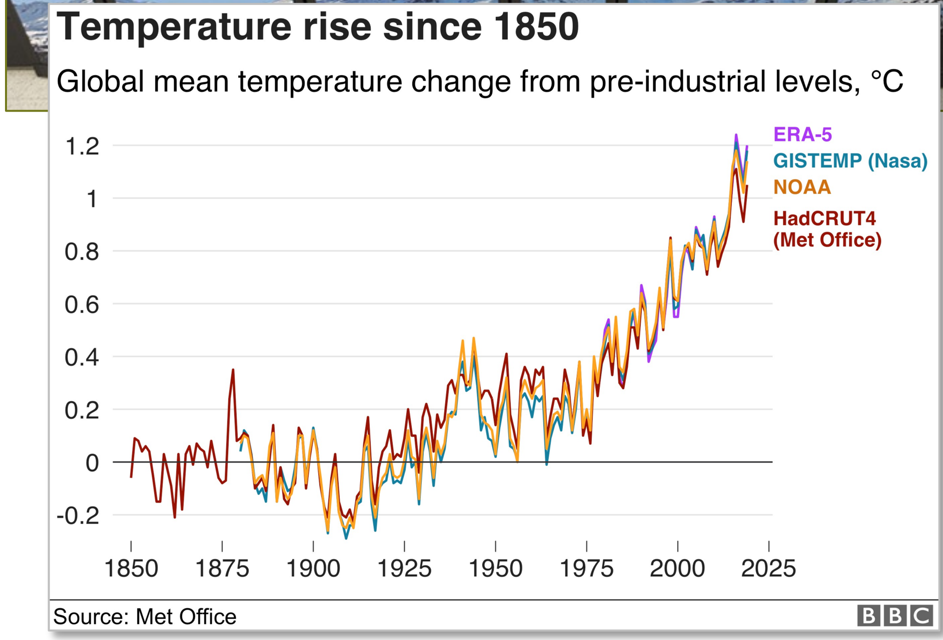
Data are costly to collect or generate:

- Compute/hours for simulation
- Telescope time
- Etc.

Some data cannot even be reproduced at will:

- Earthquake data
- Climate records
- Etc.





Temperature rise since 1850

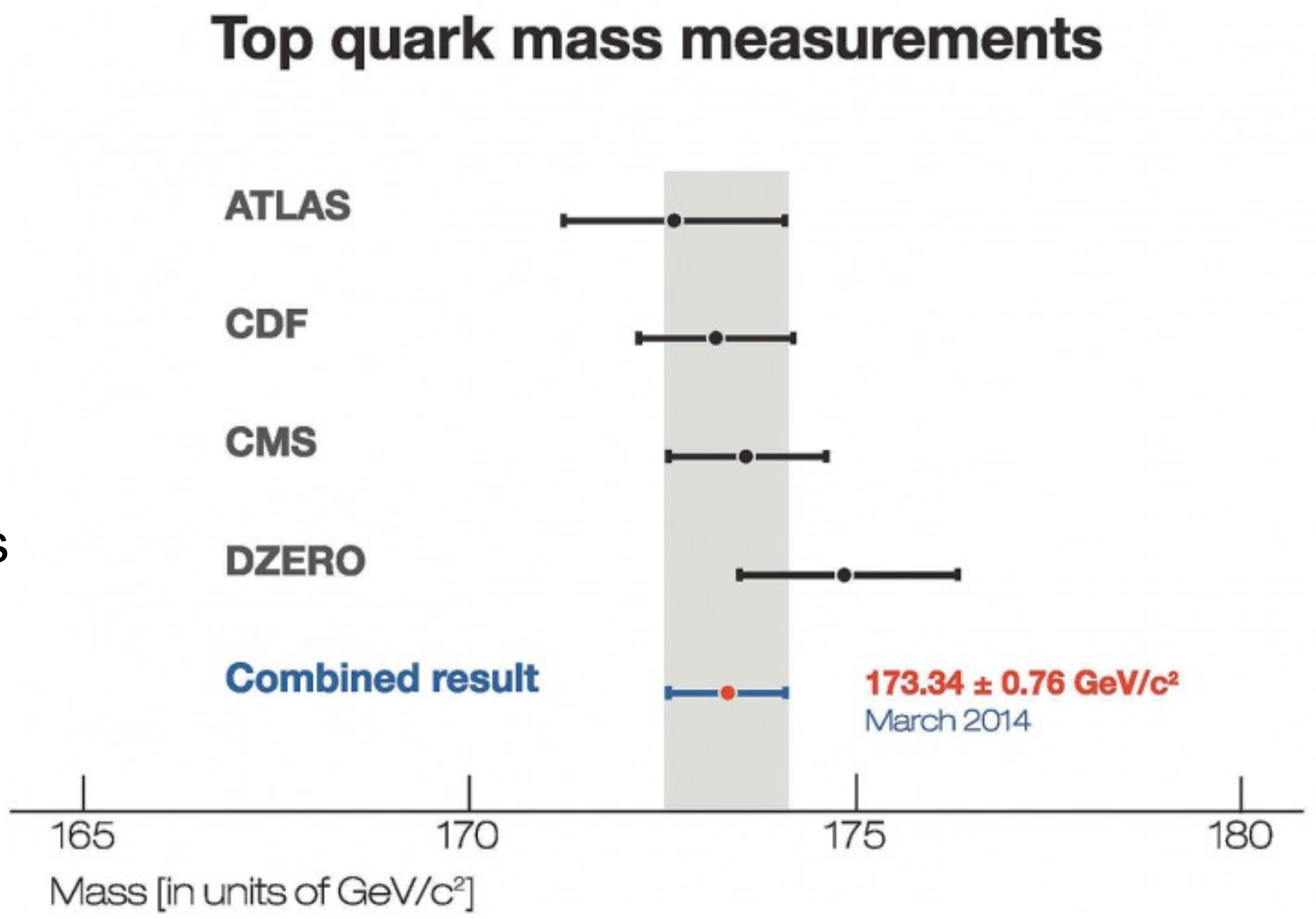
Global mean temperature change from pre-industrial levels, °C

Source: Met Office

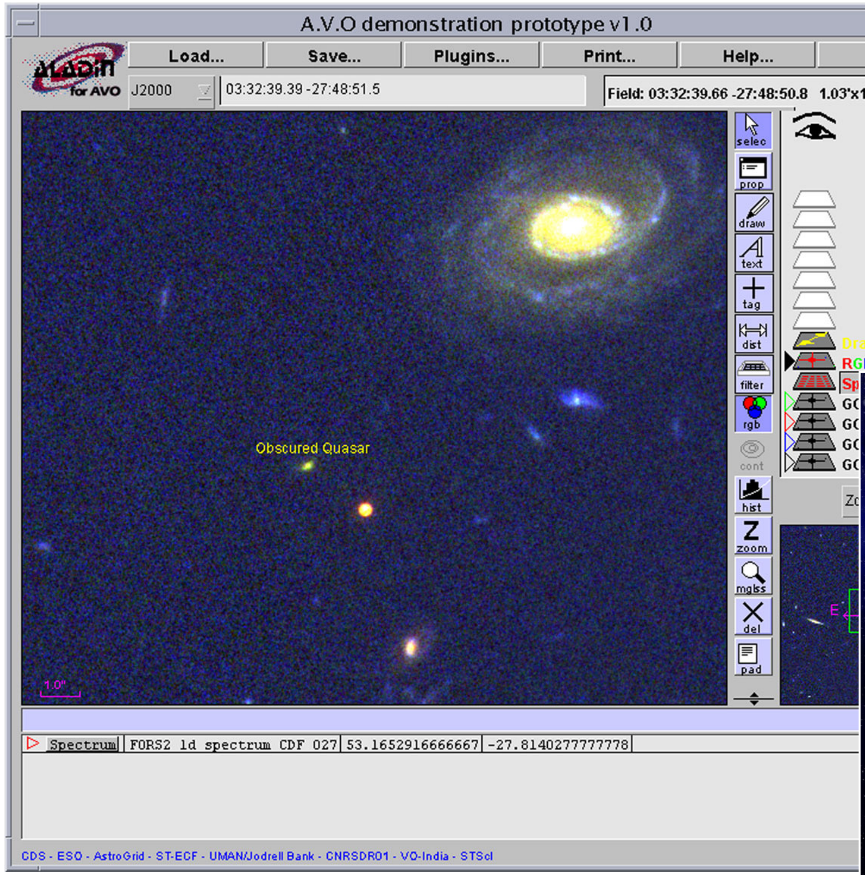
Data could be reanalyzed or combined

It is routine at CERN,
for example

Here data from CERN LHC and Fermilab DØ experiments are combined to give a better estimate of the top quark mass



Archived data could support “discovery by browsing”



03:32:39.39 -27:48:51.5

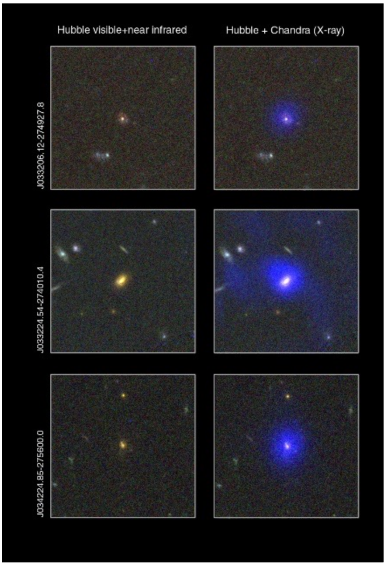
Field: 03:32:39.66 -27:48:50.8 1.03'x1.03'

```
Spectrum | FORS2 1d spectrum CDF 027 | 53.1652916666667 | -27.8140277777778
```

CDS - ESO - AstroGrid - ST-ECF - UMAN/Jodrell Bank - CNRSDR01 - VO-India - STScI

Paolo Padovani, “Discovering missing black holes: First Science from a Virtual Observatory”, AVO (2004)

https://esahubble.org/news/heic0409/



Our data could become part of some scientific data collection

NOAAServer

Access to Distributed NOAA Data and Information

NVO

NATIONAL VIRTUAL OBSERVATORY

RCSB PDB

PROTEIN DATA BANK

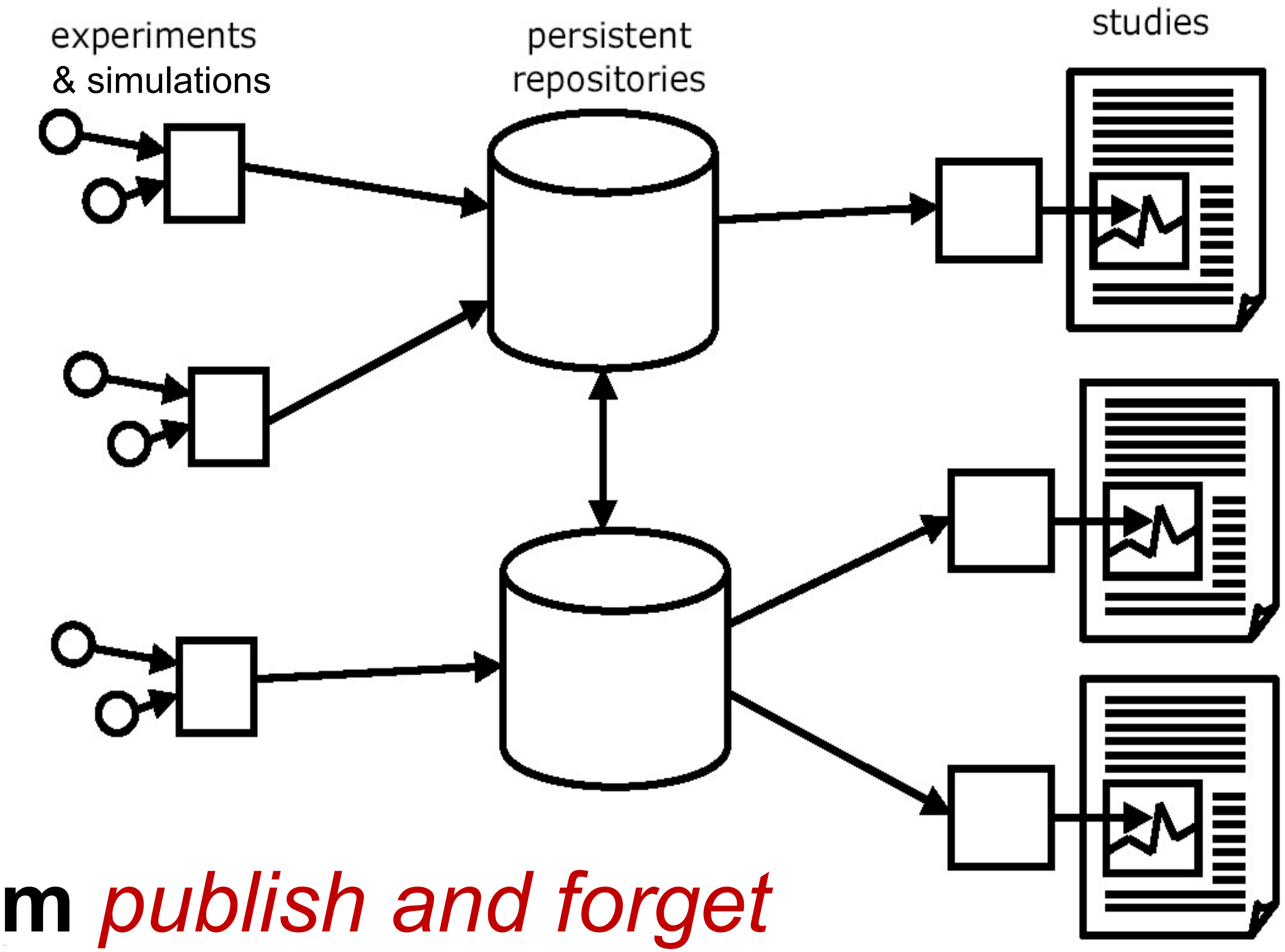
AVO

ASTROPHYSICAL VIRTUAL OBSERVATORY

- There is already a growing list of managed data collections
- They make data accessible and “recycle” existing data

HBP Human Brain Project

New trend in scientific data lifecycle



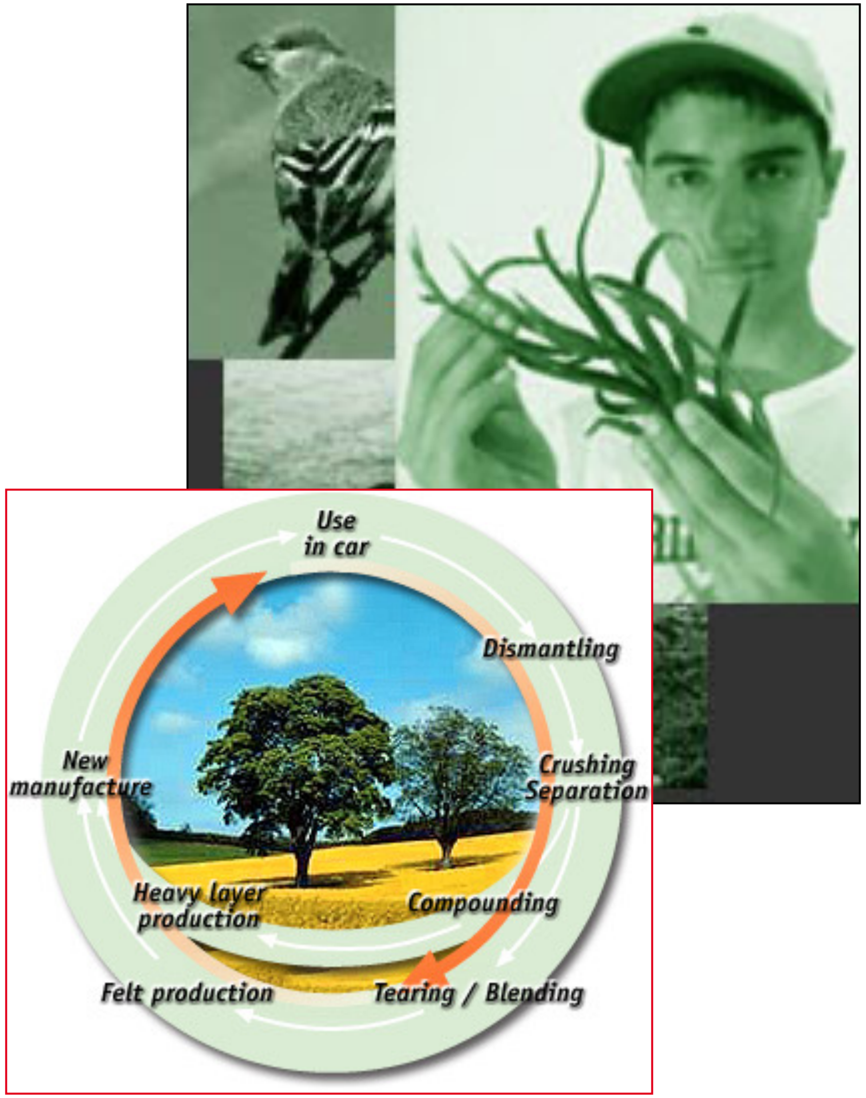
From *publish and forget*
to *use, reuse, recycle*

So, be a data ecologist

Data is like a natural resource:

- Try to use it better
- Do not waste or pollute
- Recycle
- Assure quality
- Preserve for future generations

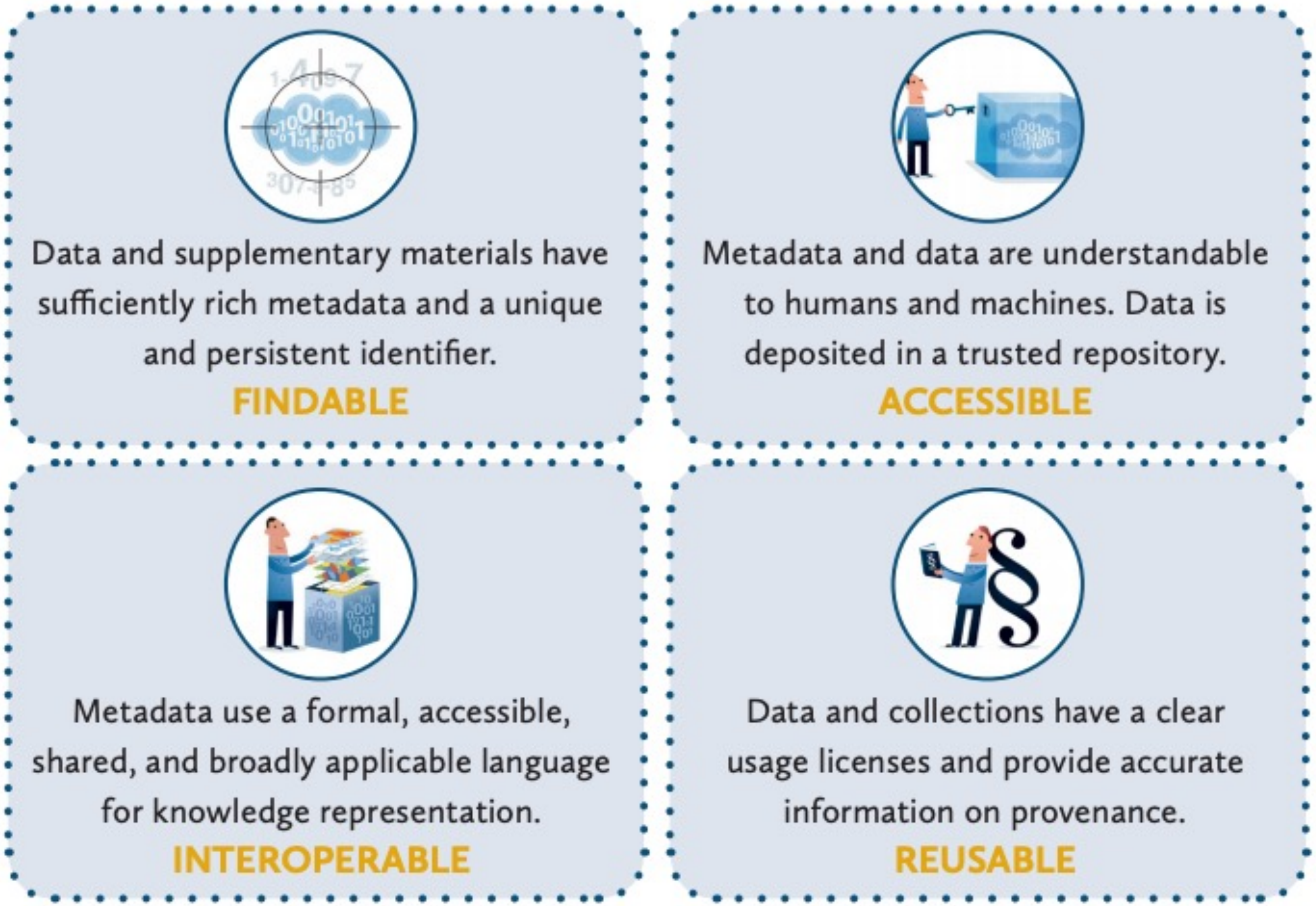
Scientific journals and funding agencies started this change by requiring data associated to papers/project to be preserved and made available



FAIR principles

The principles emphasize machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

https://www.go-fair.org/



Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.

FINDABLE

Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.

ACCESSIBLE

Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

INTEROPERABLE

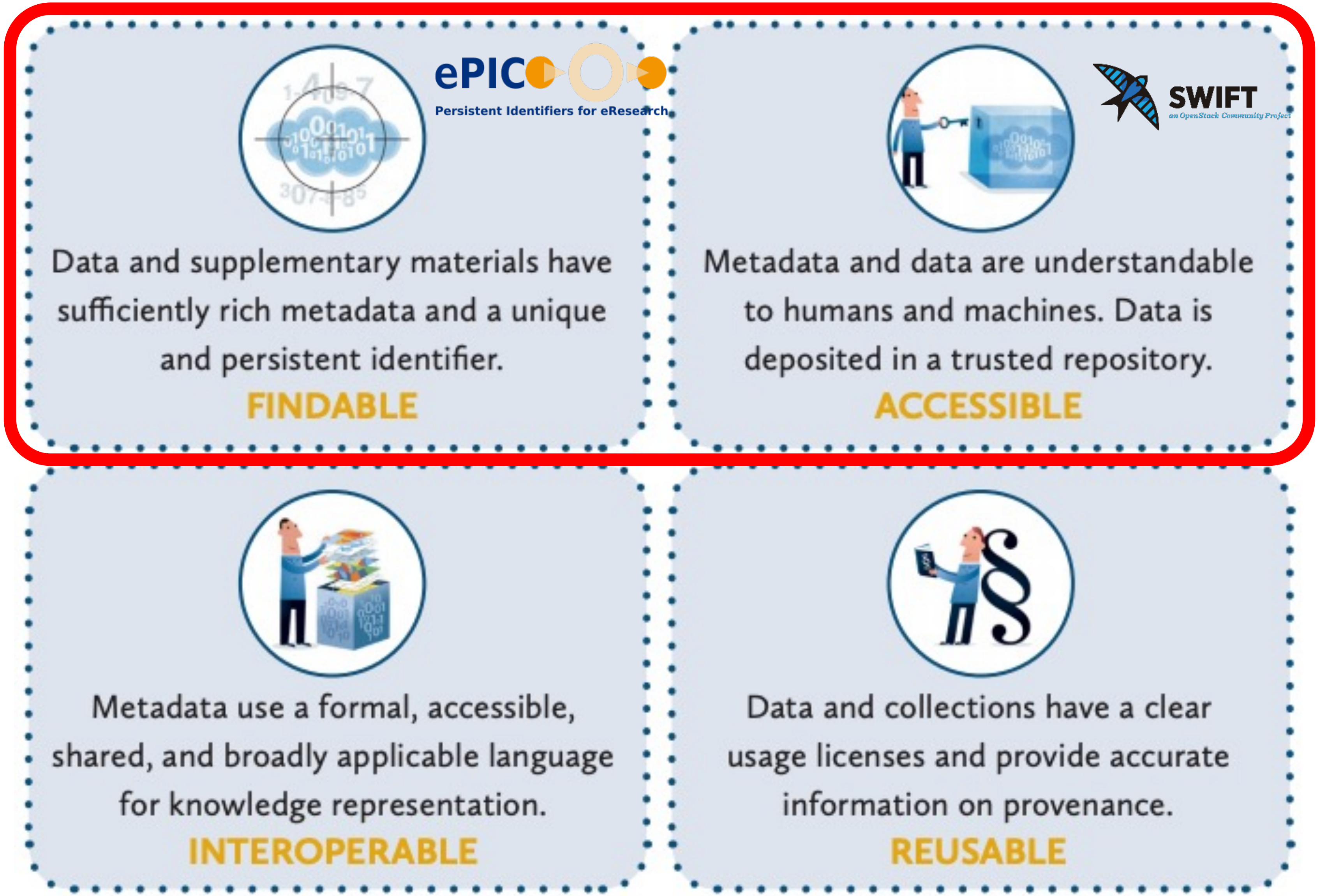
Data and collections have a clear usage licenses and provide accurate information on provenance.

REUSABLE

FAIR and CSCS data services

The four FAIR Principles, together with increasing requirement from funding agencies to have the data produced by their funded research publicly available, has motivated CSCS to offer its:

Long-Term Storage (LTS) Service


ePIC
Persistent Identifiers for eResearch
SWIFT
an OpenStack Community Project

Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.

FINDABLE

Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.

ACCESSIBLE

Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

INTEROPERABLE

Data and collections have a clear usage licenses and provide accurate information on provenance.

REUSABLE

Findability by the CSCS PID service

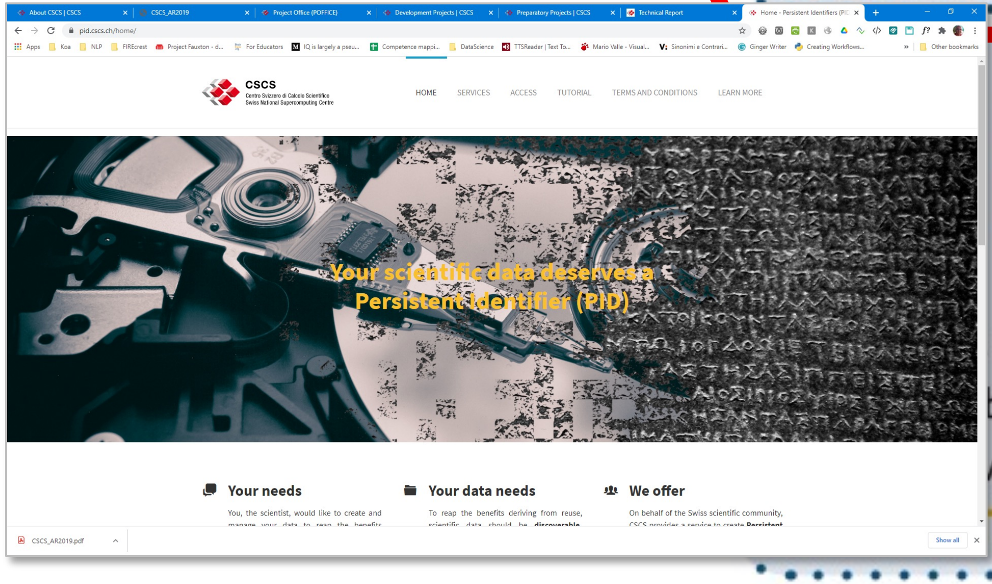
To make LTS data findable, CSCS provides a service to generate and manage a certain range of **Persistent Identifiers (PID)** assigned to Switzerland by the ePIC consortium and to resolve them.

Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.

FINDABLE

Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.

ACCESSIBLE



Your scientific data deserves a Persistent Identifier (PID)

Your needs

You, the scientist, would like to create and manage your data to reap the benefits

Your data needs

To reap the benefits deriving from reuse, scientific data should be **discoverable**

We offer

On behalf of the Swiss scientific community, CSCS provides a service to create **Persistent**

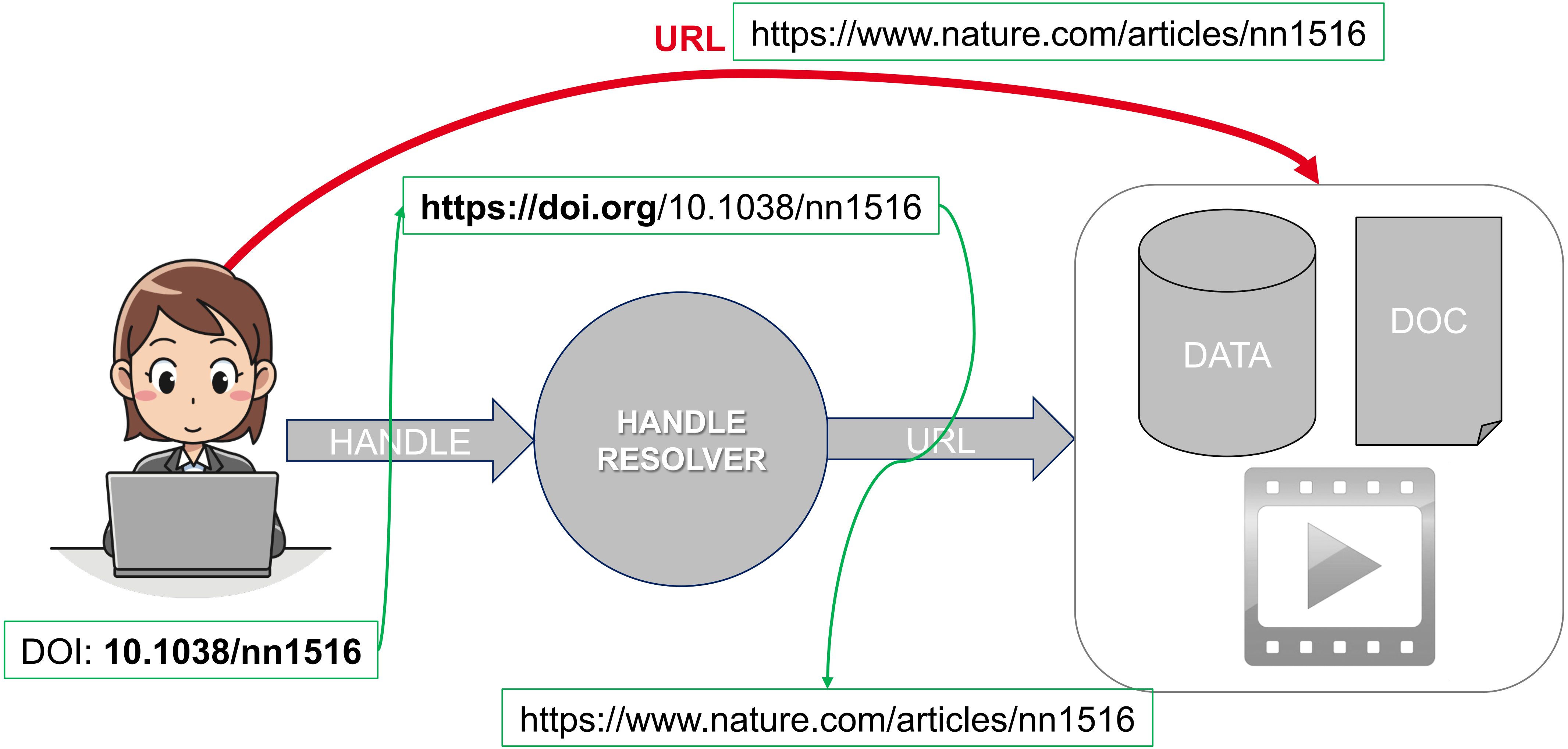
use a formal, accessible, broadly applicable language wledge representation.

TEROPERABLE

Data and collections have a clear usage licenses and provide accurate information on provenance.

REUSABLE

Base of every PID system (DOI, Handle, URN, ARK, PURL, ISBN…)



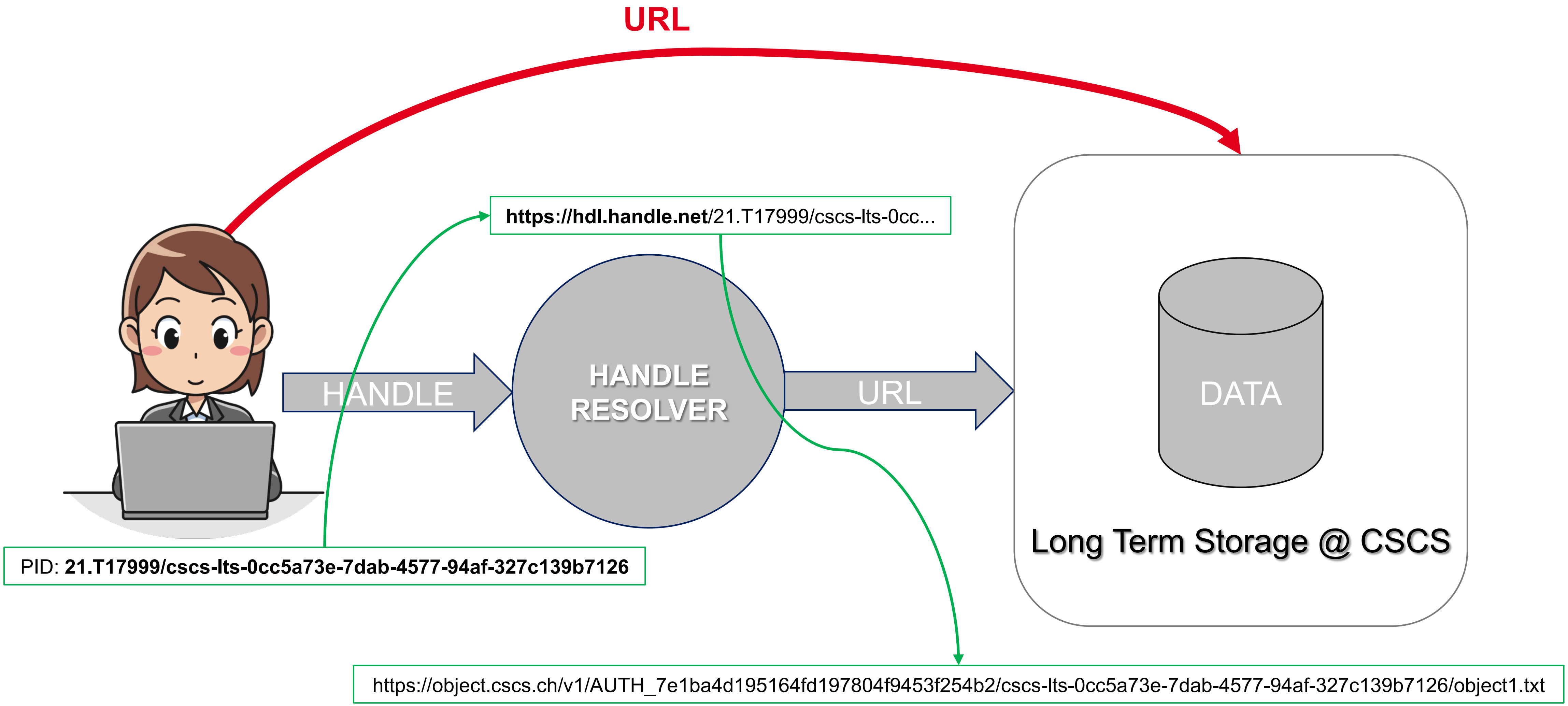
URL https://www.nature.com/articles/nn1516

https://doi.org/10.1038/nn1516

DOI: **10.1038/nn1516**

https://www.nature.com/articles/nn1516

Base of every PID system (DOI, Handle, URN, ARK, PURL, ISBN…)

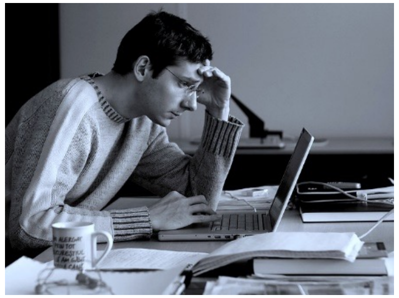


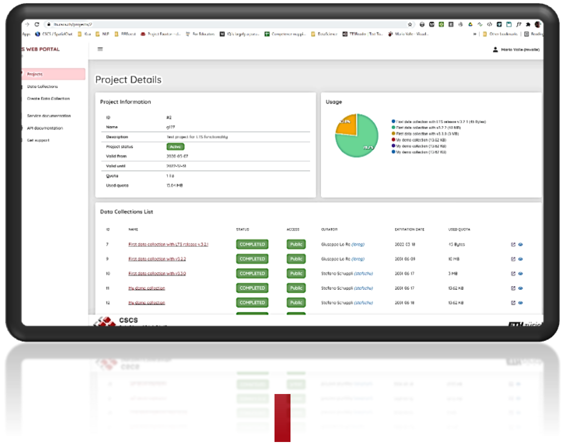
https://hdl.handle.net/21.T17999/cscs-lts-0cc...

PID: **21.T17999/cscs-lts-0cc5a73e-7dab-4577-94af-327c139b7126**

https://object.cscs.ch/v1/AUTH_7e1ba4d195164fd197804f9453f254b2/cscs-lts-0cc5a73e-7dab-4577-94af-327c139b7126/object1.txt

Long term storage at CSCS access through PID





Project Details

Project Information

ID	#2
Name	g177
Description	Test project for LTS functionality
Project status	Active
Valid from	2020-05-07
Valid until	2022-12-31
Quota	1 TB
Used quota	13.04 MB

Usage

- First data collection with LTS release v.3.2.1 (45 Bytes)
- First data collection with v3.2.2 (10 MB)
- First data collection with v3.3.0 (3 MB)
- My demo collection (13.62 KB)
- My demo collection (13.62 KB)
- My demo collection (13.62 KB)

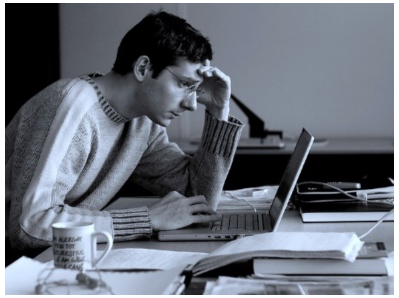
Data Collections List

ID	NAME	STATUS	ACCESS	CURATOR	EXPIRATION DATE	USED QUOTA
7	First data collection with LTS release v.3.2.1	COMPLETED	Public	Giuseppe Lo Re (loreg)	2022 03 18	45 Bytes
9	First data collection with v3.2.2	COMPLETED	Public	Giuseppe Lo Re (loreg)	2031 06 09	10 MB
10	First data collection with v3.3.0	COMPLETED	Public	Stefano Schuppli (stefschu)	2031 06 17	3 MB
11	My demo collection	COMPLETED	Public	Stefano Schuppli (stefschu)	2031 06 17	13.62 KB
12	My demo collection	COMPLETED	Public	Stefano Schuppli (stefschu)	2031 06 18	13.62 KB

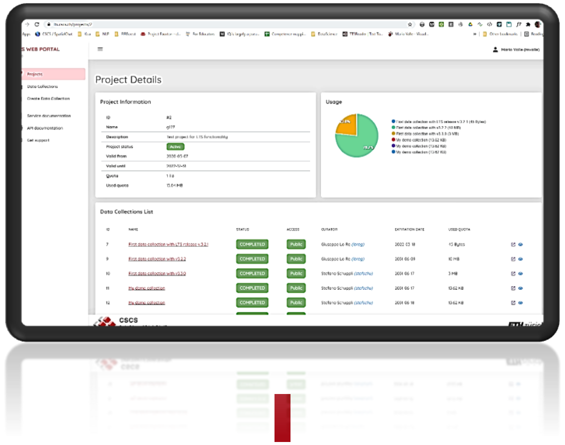
CSCS Long-term storage portal

Access simulation results using PID

Long term storage at CSCS shields users from technology changes



Access simulation results using the same PID



S WEB PORTAL

- Projects
- Data Collections
- Create Data Collection
- Service documentation
- API documentation
- Get support

Mario Valle (mvalle)

Project Details

Project Information

ID	#2
Name	g177
Description	Test project for LTS functionality
Project status	Active
Valid from	2020-05-07
Valid until	2022-12-31
Quota	1 TB
Used quota	13.04 MB

Usage

- First data collection with LTS release v3.2.1 (45 Bytes)
- First data collection with v3.2.2 (10 MB)
- First data collection with v3.3.0 (3 MB)
- My demo collection (13.62 KB)
- My demo collection (13.62 KB)
- My demo collection (13.62 KB)

Data Collections List

ID	NAME	STATUS	ACCESS	CURATOR	EXPIRATION DATE	USED QUOTA
7	First data collection with LTS release v.3.2.1	COMPLETED	Public	Giuseppe Lo Re *(lorег)*	2022 03 18	45 Bytes
9	First data collection with v3.2.2	COMPLETED	Public	Giuseppe Lo Re *(loreg)*	2031 06 09	10 MB
10	First data collection with v3.3.0	COMPLETED	Public	Stefano Schuppli *(stefschu)*	2031 06 17	3 MB
11	My demo collection	COMPLETED	Public	Stefano Schuppli *(stefschu)*	2031 06 17	13.62 KB
12	My demo collection	COMPLETED	Public	Stefano Schuppli *(stefschu)*	2031 06 18	13.62 KB

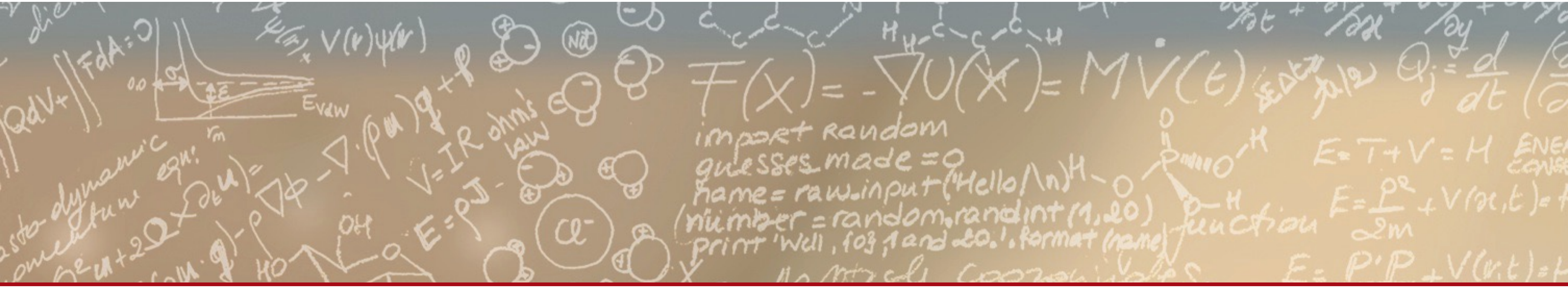
CSCS Long-term storage portal

Data migration

CSCS

Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich


F(X) = -∇U(X) = MV(t)
import Random
guesses_made = 0
name = raw_input("Hello\n")
number = random.randint(1, 20)
print 'Well, {0}, I and 20.'.format(name)
E = T + V = H
V = IR ohm's law
E = ρJ

Thank you for your attention.

cscsch