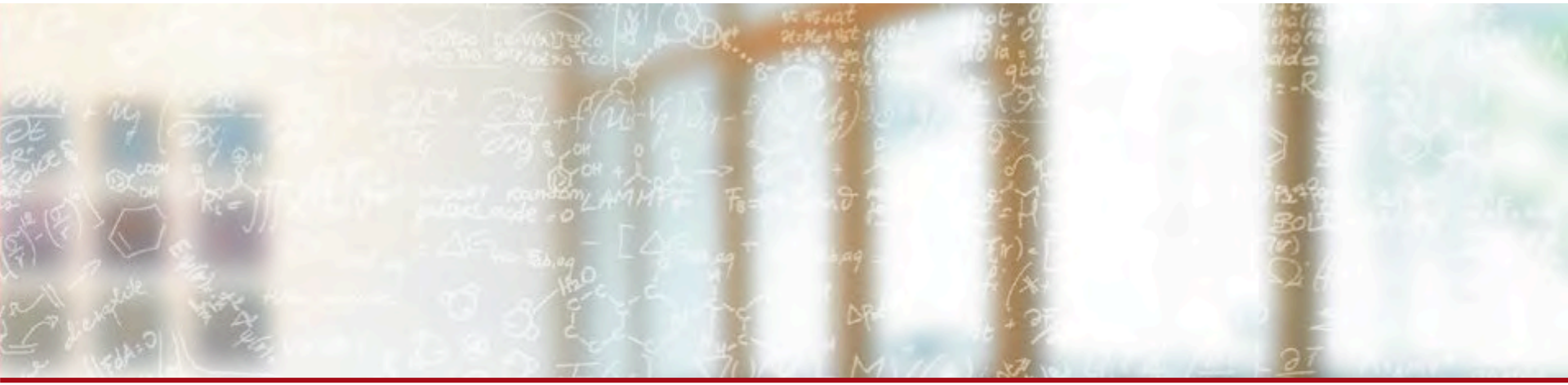
CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich



# Your Data Deserves a Permanent Identifier (PID)

hpc-ch – Storage Technologies and Data Management Workshop

Mario Valle, CSCS

October 4, 2018

# A truism: we produce a lot of data

"Our ability to capture and store data far outpaces our ability to process and exploit it. This growing challenge has produced a phenomenon we call the **data tombs**, or data stores that are effectively write-only; **data is deposited to merely rest in peace**, since in all likelihood it will never be accessed again. Data tombs also represent missed opportunities."

*Usama Fayyad – Yahoo! Research Laboratories*

# We could benefit from data use, reuse and recycle

- Discovery by browsing (a.k.a. Google science)
- Astronomy and Astrophysics Virtual Observatories (e.g., EURO-VO)
- Data reanalysis (common at CERN and in climate science)
- Find correlations between data and metadata (e.g., OMEGA project for bio-imaging of virion movement in cells)
- Providing context for other data
- Stimulate new usage patterns

# Controversial, but nonetheless a new data paradigm

- Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to settle for models at all.
- For example: Google can translate languages without actually “knowing” them.
- Every kind of Machine Learning, deep or not, do the same.

WIRED SUBSCRIBE » SECTIONS » BLOGS » REVIEWS » VIDEO » HOV

WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Illustration: Marian Bantjes

THE PETABYTE AGE:
Sensors everywhere. Infinite storage. Clouds of processors. Our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology. As our collection of facts and figures grows, so will the opportunity to find answers to fundamental questions. Because in the era of big data, more isn't just more. More is different.

THE END OF THEORY:
Essay: The Data Deluge Makes the Scientific Method Obsolete

**"All models are wrong**, but some are useful."

So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to settle for models at all.

Sixty years ago, digital computers made information readable. Twenty years ago, the Internet made it reachable. Ten years ago, the first search engine crawlers

www.wired.com/science/discoveries/magazine/16-07/pb_theory

# SNF (and other) requests about data publication…

- … but these are only the tip of the iceberg.
- Scientists want to structure the data use, reuse and recycle from the beginning, when data is created. You don’t want to attach the problem at the end, when the work is published.
- Also you don’t want only more bureaucracy or rules to comply with, without perceived benefits for your science.

# Prerequisites to make all this happens

- Data should be **discoverable** (by associated metadata or by public catalogs. Kudos to Google for its Dataset Search)
- Data should be **unambiguously** and **certainly identified** (by something that depends on data content and not location and is the basis of authorship assignment)
- Data should be **publicly accessible** and **permanent** (should not disappear when researcher moves to another university. If needed, after discovery there may be an authorization step)
- Data should be **trusted** (i.e., it is what it claim to be, authorship is clear, metadata are verified)

# In other words: data should be FAIR

FAIR data is data which meets standards of:

- **F**indability
- **A**ccessibility
- **I**nteroperability
- **R**eusability

(https://www.nature.com/articles/sdata201618)

# Another step after FAIR is Linked Open Data

The 5-stars deployment scheme for Linked Open Data proposed by Tim Berners-Lee

(https://5stardata.info/en/)

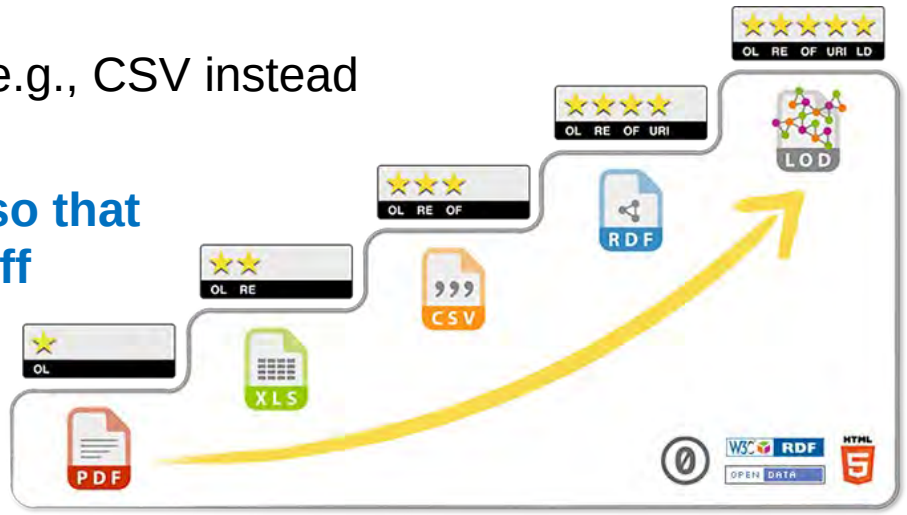
★ Make your stuff available on the Web (whatever format) under an open license

★★ Make it available as structured data (e.g., Excel instead of image scan of a table)

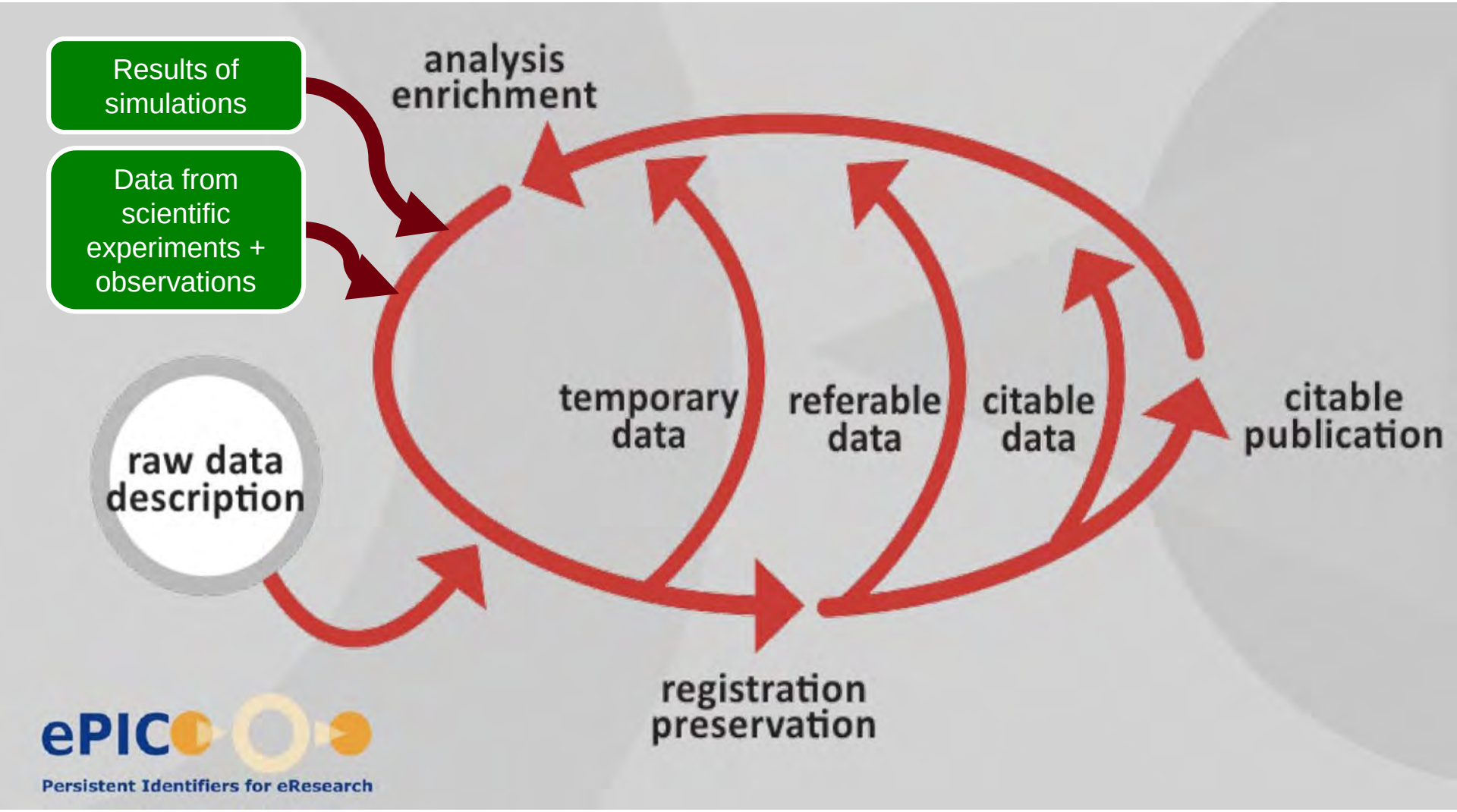
★★★ Use non-proprietary formats (e.g., CSV instead of Excel)

★★★★ **Use URIs to denote things, so that people can point at your stuff**

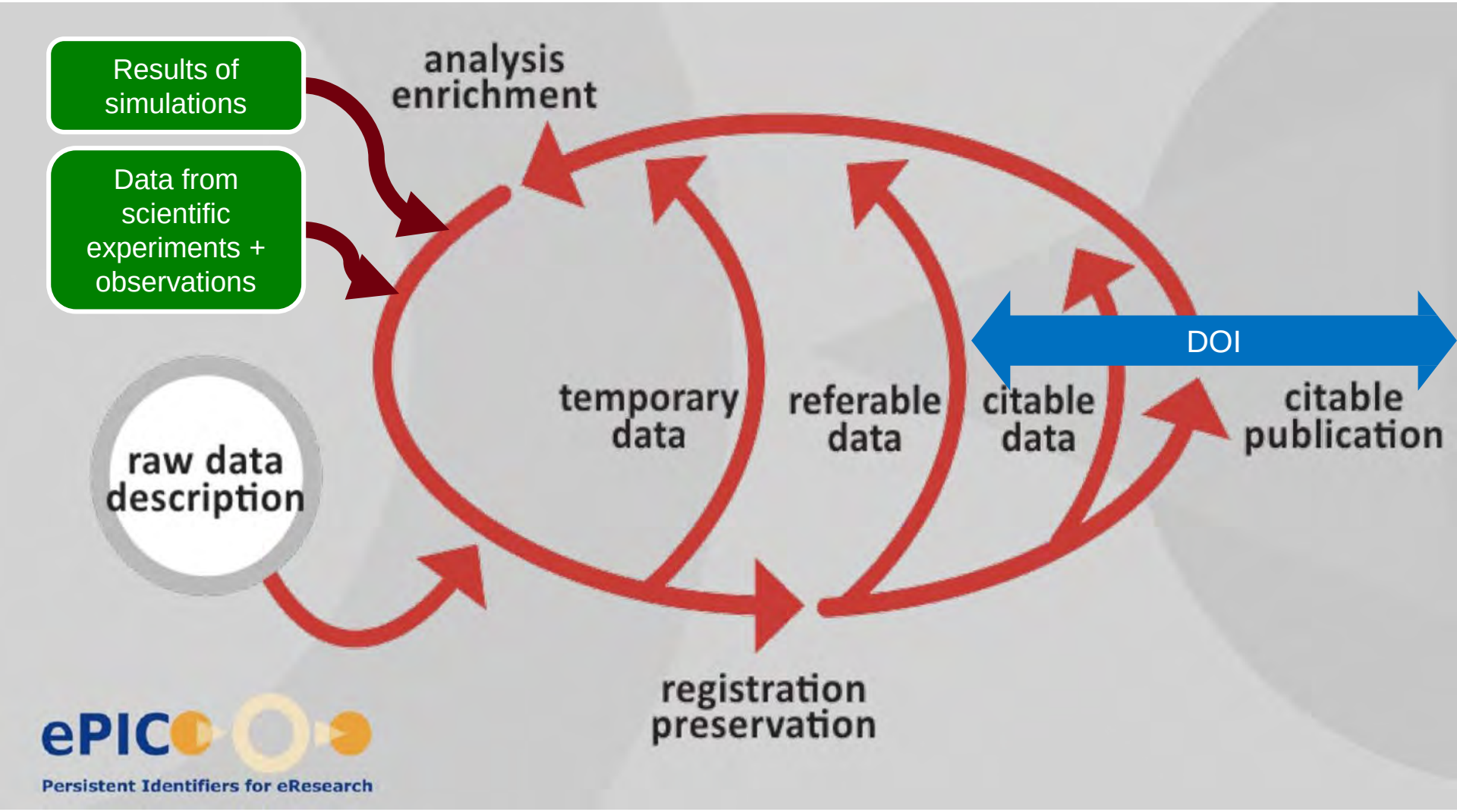
★★★★★ Link your data to other data to provide context



# Citing Data in Science



# Citing Data in Science



# Publications solved these problems introducing DOI

www.doi.org

HOME | HANDBOOK | FACTSHEETS | FAQs | RESOURCES | REGISTRATION AGENCIES | NEWS | MEMBERS AREA

## The DOI® System ISO 26324


This is the web site of the International DOI Foundation (IDF), a not-for-profit membership organization that is the governance and management body for the federation of Registration Agencies providing Digital Object Identifier (DOI) services and registration, and is the registration authority for the ISO standard (ISO 26324) for the DOI system. The DOI system provides a technical and social infrastructure for the registration and use of

### Resolve a DOI Name

Type or paste a DOI name, e.g., 10.1000/xyz123, into the text box below. (Be sure to enter all of the characters before and after the slash. Do not include extra characters, or sentence punctuation marks.)

SUBMIT

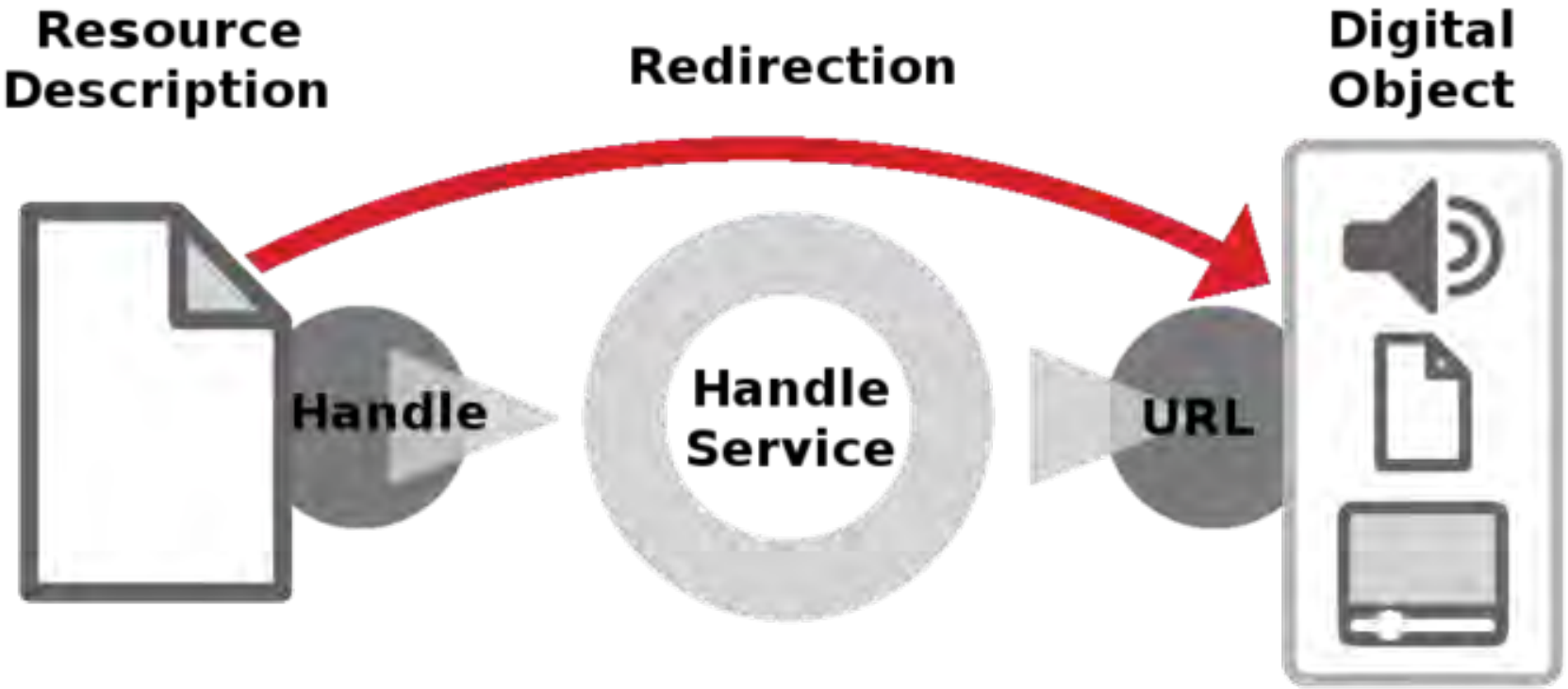
Clicking on a DOI link (try this one: https://doi.org/10.1109/5.771073) takes you to one or more current URLs or other services related to a single resource. If the URLs or services change over time, e.g., the resource moves, this same DOI will continue to resolve to the correct resources or services at their new locations.

Check the current status of the DOI system at doi.statuspage.io.

DRIVEN BY doi

Enhance the value of your content.
Join the DOI Community.
Watch a video, get the facts, and find

# Base of any handle system (e.g., DOI)



# DOI comes with an established set of metadata

doi2bib — give us a DOI
and we will do our best to get you the BibTeX entry

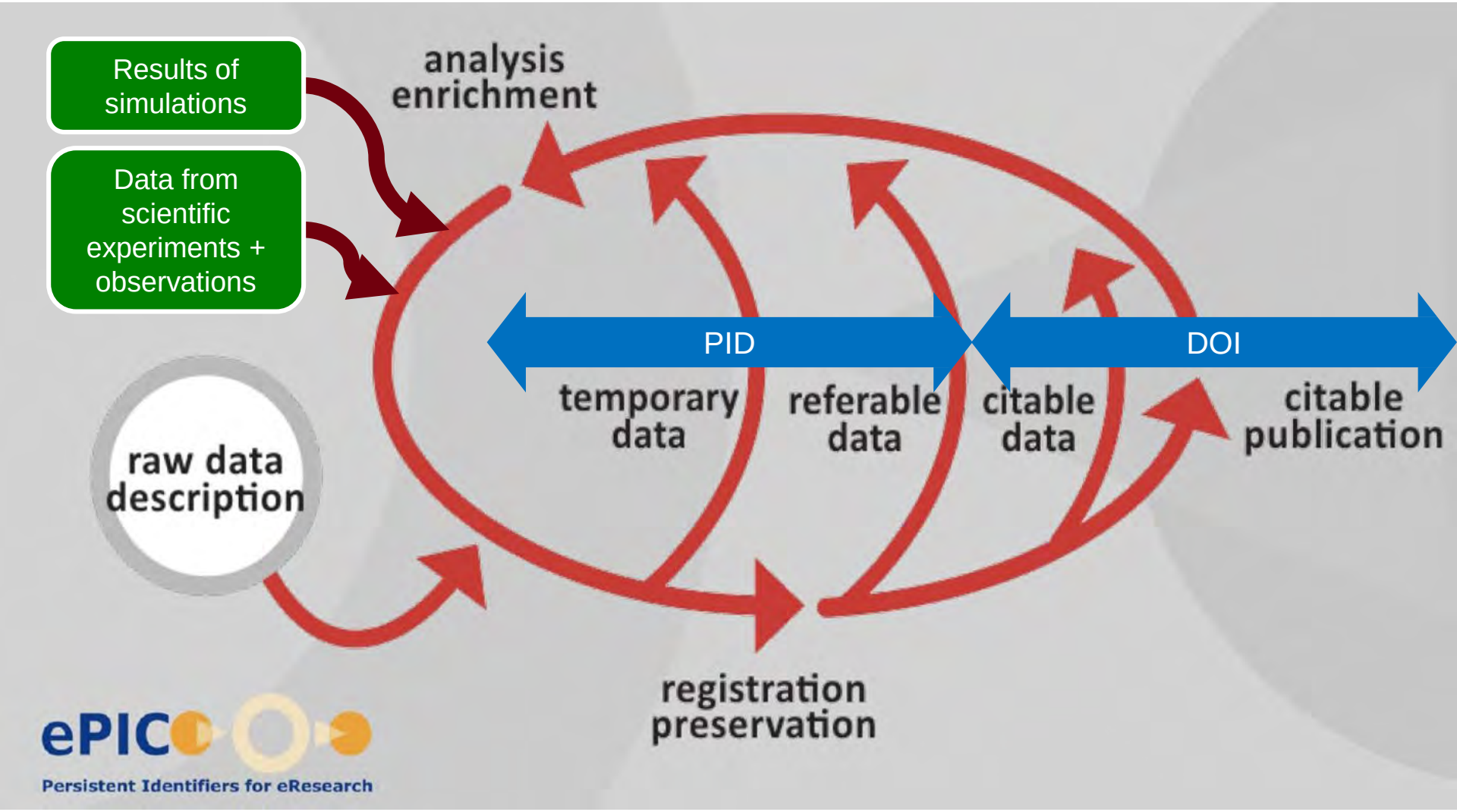
10.1107/S0108767310026395 get BibTeX

```
@article{Valle2010,
  doi = {10.1107/s0108767310026395},
  url = {https://doi.org/10.1107/s0108767310026395},
  year  = {2010},
  month = {aug},
  publisher = {International Union of Crystallography ({IUCr})},
  volume = {66},
  number = {5},
  pages = {507--517},
  author = {Mario Valle and Artem R. Oganov},
  title = {Crystal fingerprint space {\textendash} a novel paradigm for studying crystal-s
  journal = {Acta Crystallographica Section A Foundations of Crystallography}
}
```

https://doi.org/10.1107/s0108767310026395

Copy Bib to Clipboard Copy URL to Clipboard

# Citing Data in Science

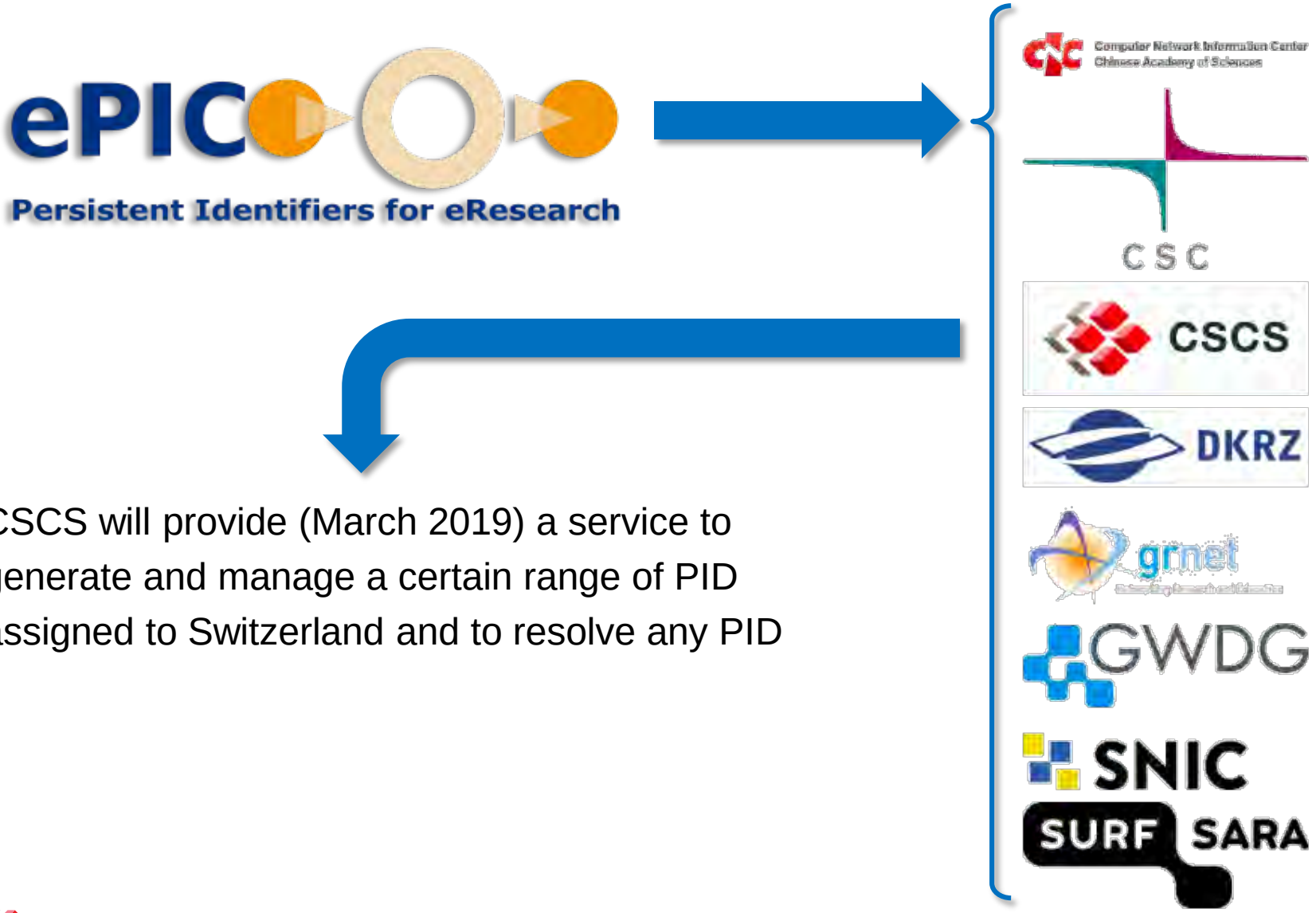


# Permanent Identifiers (PID) to cover the rest


https://www.pidconsortium.eu/

- A Permanent Identifier (PID) identifies data objects regardless of their location, associate metadata to them and claim authorship.
- The PID infrastructure provides, at least, the following services:
  - Create PID and keep track of them.
  - Resolve a PID to the corresponding location.
- The ePIC consortium members provides this infrastructure ensuring its trustfulness and stability.

# CSCS is part of the ePIC consortium (since Sept. 2018)




CSCS will provide (March 2019) a service to generate and manage a certain range of PID assigned to Switzerland and to resolve any PID

# Structure of a PID

- A PID is a string with the following structure:
  - `<PREFIX>/<SUFFIX>`
- `<PREFIX>`
  - `21.nnnnn`
  - Where “21.” identifies a PID (note that DOI starts with “10.”)
  - “nnnnn” five digits identifying the namespace (could be composed by country and institution IDs for example, but in general it is opaque)
- `<SUFFIX>`
  - Can be any unique string inside the namespace. But preferred as: `PRE-0000-0000-0000-0-POST`
  - An optional PRE UTF-8 string
  - An UUID with check digit (Universally Unique Identifier. It enables distributed systems to uniquely identify information without significant central coordination).
  - An optional POST UTF-8 string

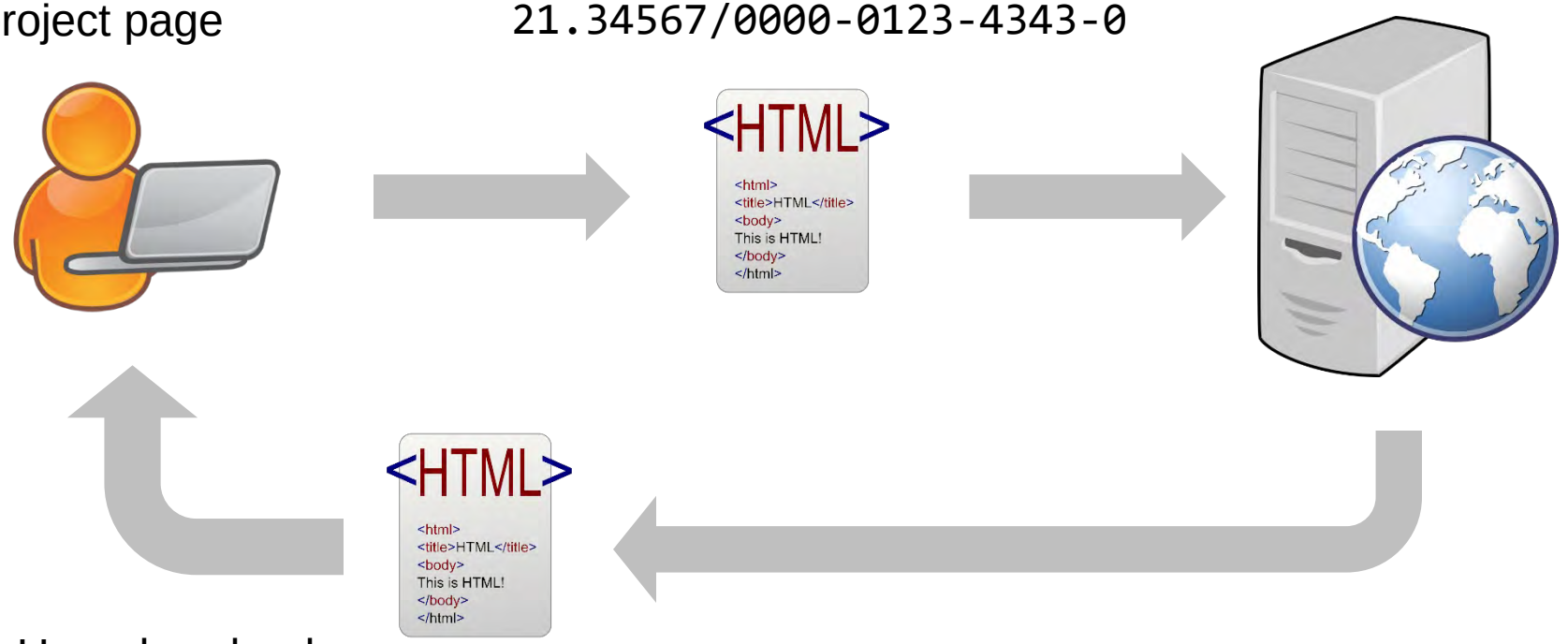
# Temporary or Test PID

- We can generate and manage not only permanent PID, but also temporary (or test) PID
- DOI does not have this capability
- Only difference: the `<PREFIX>` format is `21.Tnnnnn`
- The differences between Permanent and Temporary PID are:
  - A Permanent PID should always resolve to an URL. If the corresponding data has been removed, it should resolve to a page that states the data is missing. The PID itself could never be deleted.
  - A Temporary PID instead could be deleted anytime.

# PID Resolution

User access some project page

User clicks on a PID present there:
`21.34567/0000-0123-4343-0`

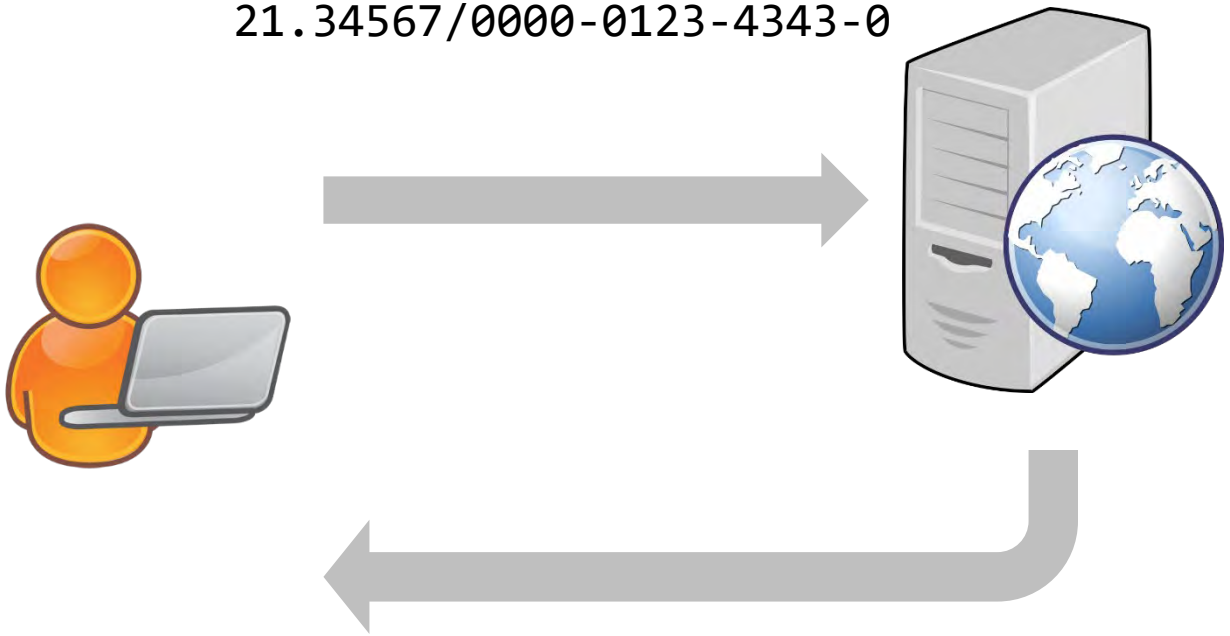



User download or access the data file from the page

Resolver returns and redirect the user to:
`https://cscs.ch/data/proj1/file.html`

# PID Resolution

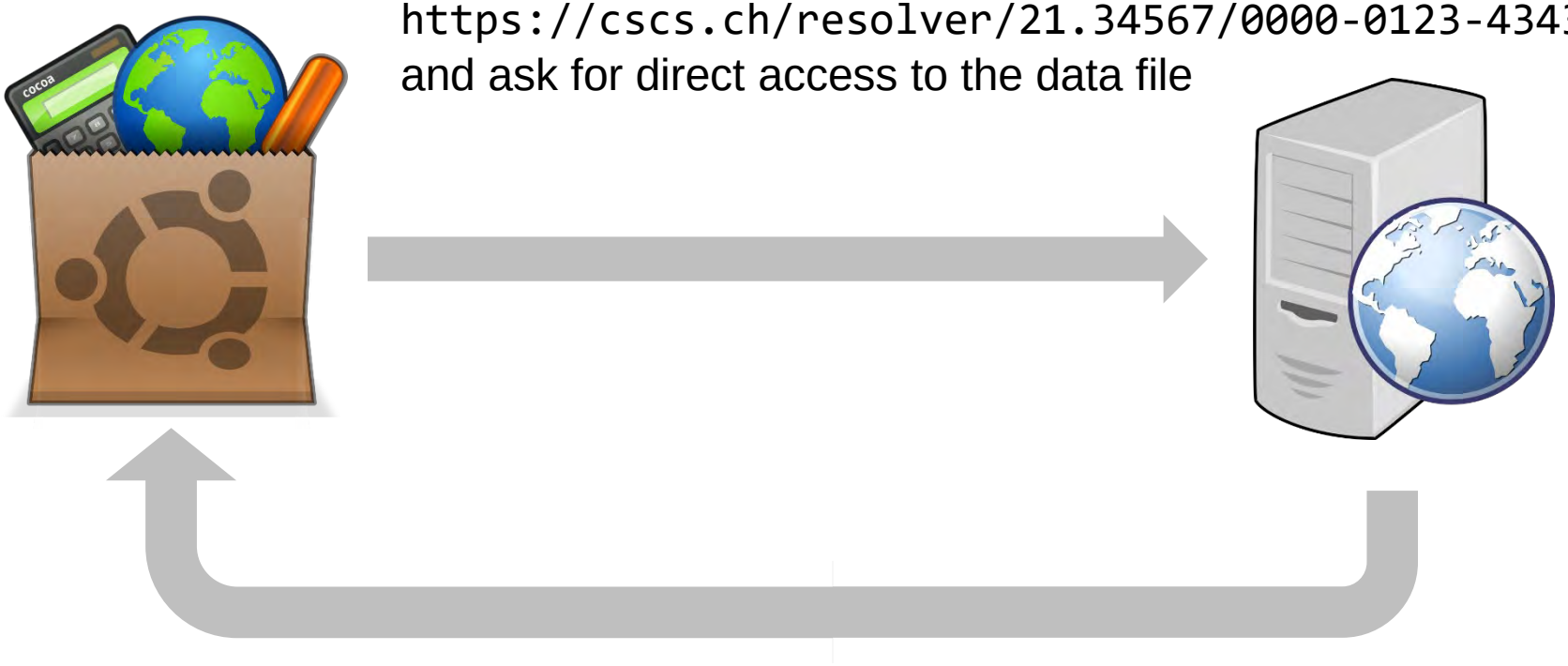
User enter a PID on the resolver web form:
`21.34567/0000-0123-4343-0`



Resolver returns:
`https://cscs.ch/data/proj1/file.html`

# PID Resolution from API

One application accesses resolver API via a GET request:
`https://cscs.ch/resolver/21.34567/0000-0123-4343-0`
and ask for direct access to the data file




Application accesses the data file

Resolver returns by content negotiation:
`https://cscs.ch/data/proj1/file.dat`

# CSCS has a roadmap to comply to ePIC consortium requirements


CSCS

Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

# CSCS PID levels of service

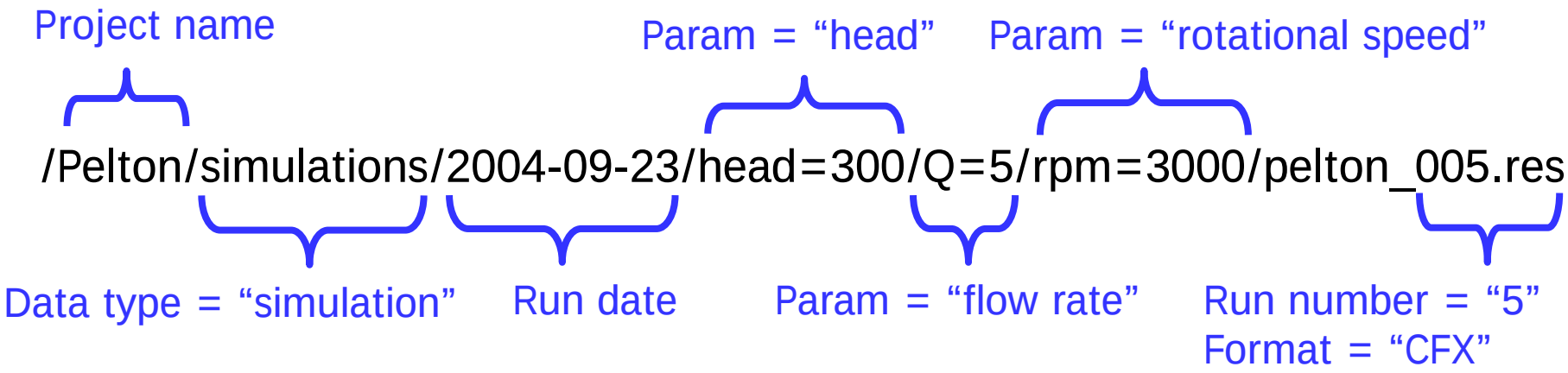
- **Level 1 – Basic PID creation/resolution**
  - March 2019
  - PID creation initially in a CSCS namespace, plan to provide institution-specific namespaces
  - Resolution for any issued PID (not only from CSCS)
  - User editing of resolved URL and minimal metadata
  - Documentation and support
- **Level 2 – Storage at CSCS**
  - Tentatively June 2019
  - CSCS provides a public, permanent storage space
  - Data ingested with a Dropbox-like mechanism (user deposits the file in a directory, and receives a PID for it).

# CSCS PID levels of service (cont.)

- **Level 3 – Metadata search**
  - Not planned yet
  - The user could associate an ample set of metadata to a PID
  - The user can run queries on metadata to obtain a list of PID
- **Level 4 – Scientific Use Cases**
  - On going
  - Consultancy on specific Scientific Use Cases and HPC projects related to large amount of data
- **Level 5 – Future requirements**
  - On going
  - CSCS will track evolution of PID to be prepared and to implement new functionalities and services

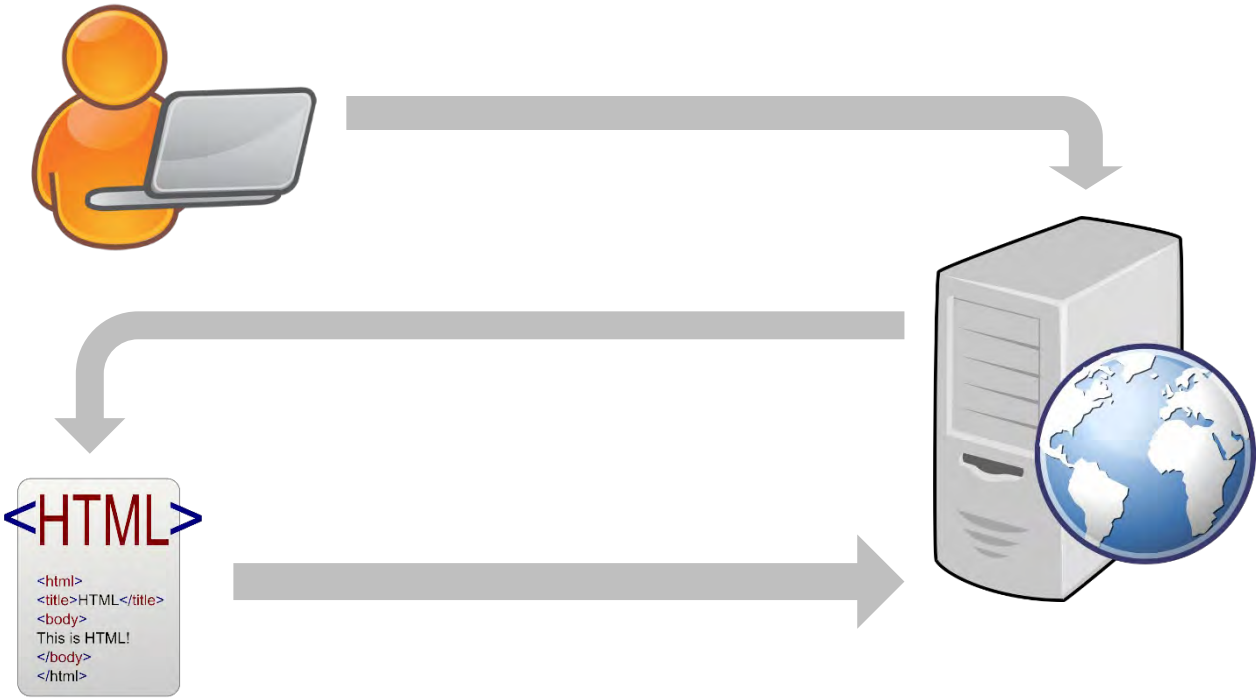
# A detour on the importance of metadata

- Project people try hard to record somewhere useful information about their data
- When those information are defined, are stored using very “ad-hoc” methods, like in a file name and path:



# PID Metadata Search & Resolution

User searches for PIDs on the resolver web form by entering:
`project=Climate&date=2009-09-09&var=ozone`




The metadata catalog returns a list of PID

As before the user selects and retrieves the data file it is interested in

# Some technicalities on metadata storage

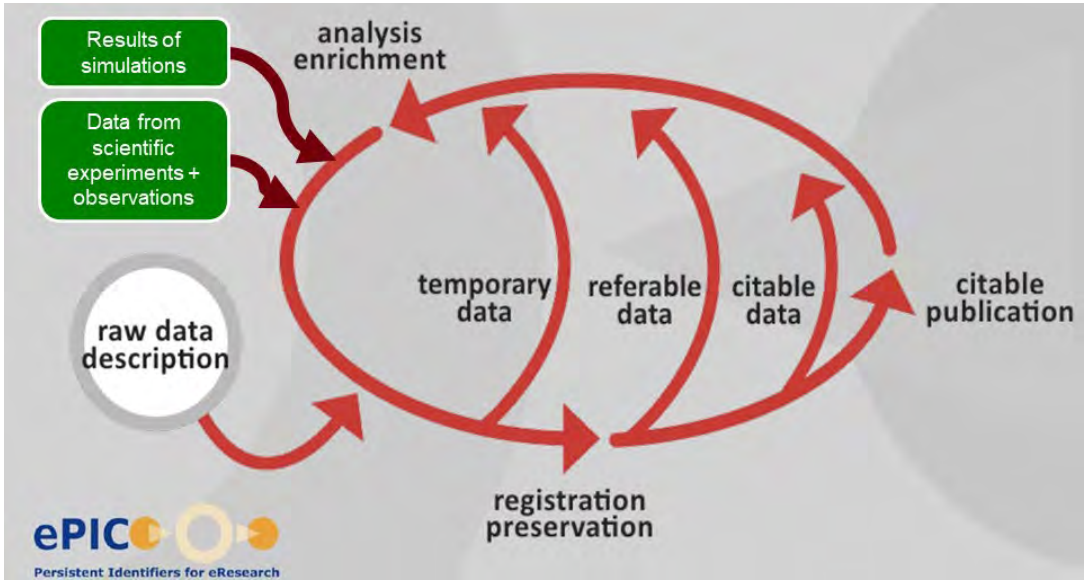
- How metadata are stored could influence how they are used in applications
- SQL database (e.g., Postgress, MySQL)
  - Fixed schema
  - Tricks to store unlimited K/V pairs
    (TABLE mdataKey: key, mdataValue: value – many-to-many)
  - Query: SQL
- NoSQL database (e.g., MongoDB)
  - No schema
  - Metadata are JSON objects {pid: pid1, key1: value1, key2: value2, …}
  - Query: db.pids.find({key: value})
- Triple store (aka RDF databases e.g. Apache Jena)
  - Triples (<subj> <property> <object>) plus ontology (private or shared?)
  - Things identified by URI. **URI ⇔ https://resolver.cscs.ch/PID**
  - Query: SPARQL

# The unpleasant side of PIDs

- The ePIC CSCS membership costs. Ergo, PID generation will cost (not too much)
- But don’t forget that if you pay for something, you attach higher value to it. The Italian saying: *“Lavoro, guadagno, spendo, pretendo”* (I work, I earn, I spend, I demand) is in action here
- Not yet defined what will cost and how much
- Besides this, permanent storage obviously will cost (but this is independent from PID)

# Scientific use cases

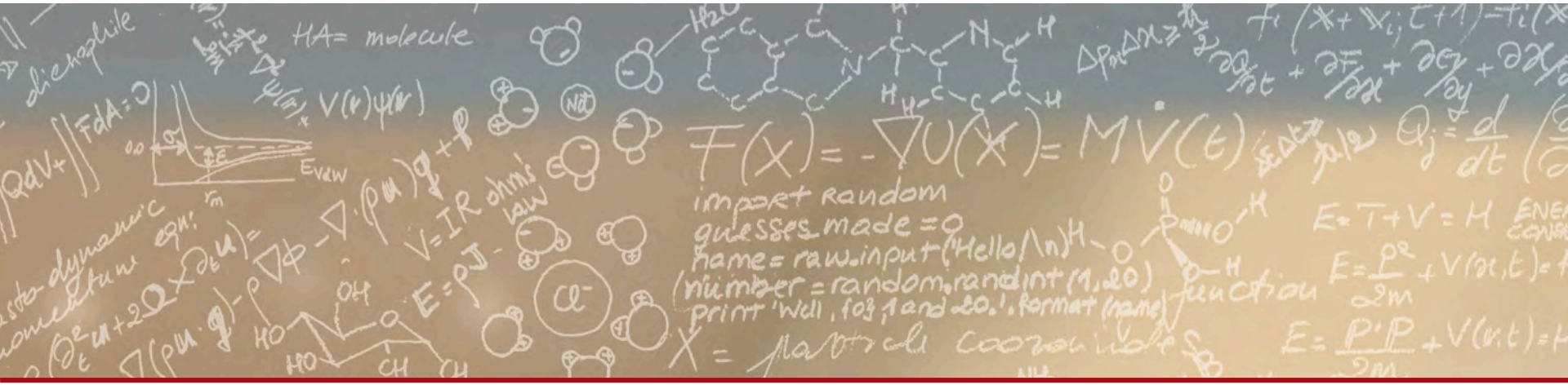
- Still searching good, real life use cases
- Integration with Provenance tracking
- Link component of an experiment in a Laboratory Notebook
- Integration with Workflow management
- Data publication
- Long term storage, migration from disk to tape (or openBIS → Repositories)
- Substituting custom references for data fragments (e.g., database record)



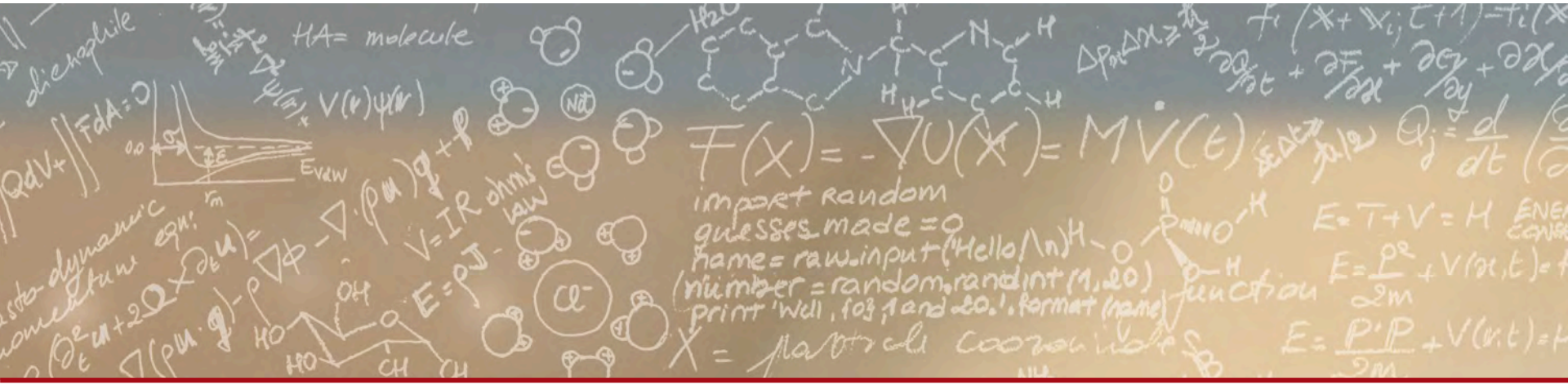
# Few cultural problems to overcome...

# Creating awareness and community

- I’m the point of contact for PID ideas, suggestions and project specific requests
- I want to create awareness and hopefully create a Swiss community interested in this aspect of data management

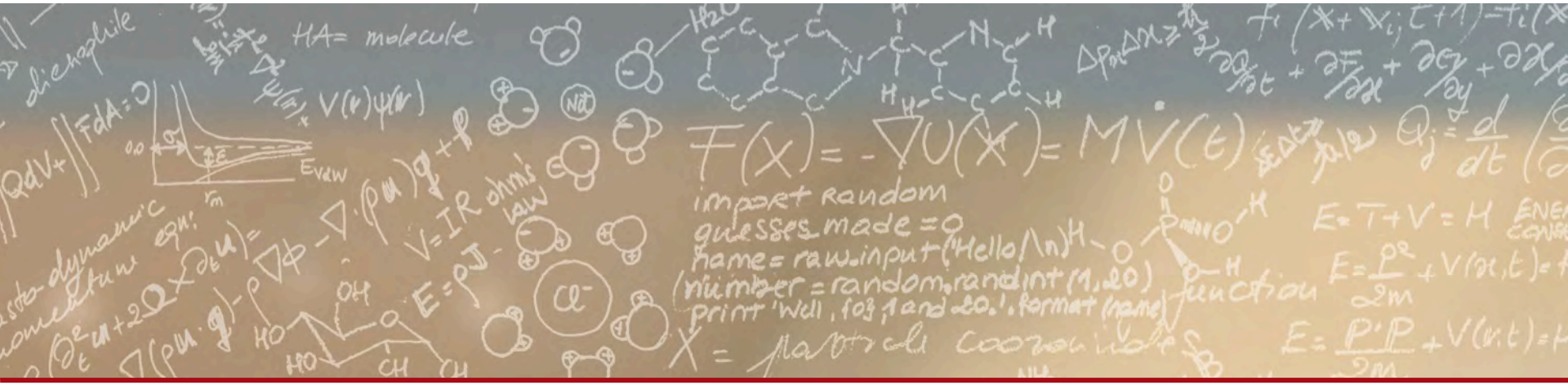



# Thank you for your attention!




# Now we have some time, so I am…


… awaiting your
valuable contributions:
questions, curiosities, ideas,
something that resonates with your research…




# Now we have truly finished.
# Thank you for your contributions!