Data Explosion in Chemistry: What Are We Going to Do With All the Data, and What Will It Do to Us

Caroline Lynn Kamerlin
kamerlinlab.com

Georgia Tech

Transformation in How We Do Chemistry

How it started: “Never trust a computational chemist”

How it's going: even “lab chemists” need to code!

Image © Joppe van der Spoel, Studio de Wilde Muis

Transformation in How We Do Chemistry

Breakdown of traditional barriers between experimentalists and theorists – digitalization is unavoidable!

Image © Joppe van der Spoel, Studio de Wilde Muis

Data Explosion Revolutionizes Chemistry

- Large language models: turning words into structures, molecules and chemical reactions.

- AI changing drug discovery, human health, chemical synthesis.

- Digital chemistry in materials science: design of new polymers, functional materials.

- Machine learning in enzyme design: sustainable chemistry.

- More fundamentally, digital chemistry also changes how we interact and communicate with each other as scientists.

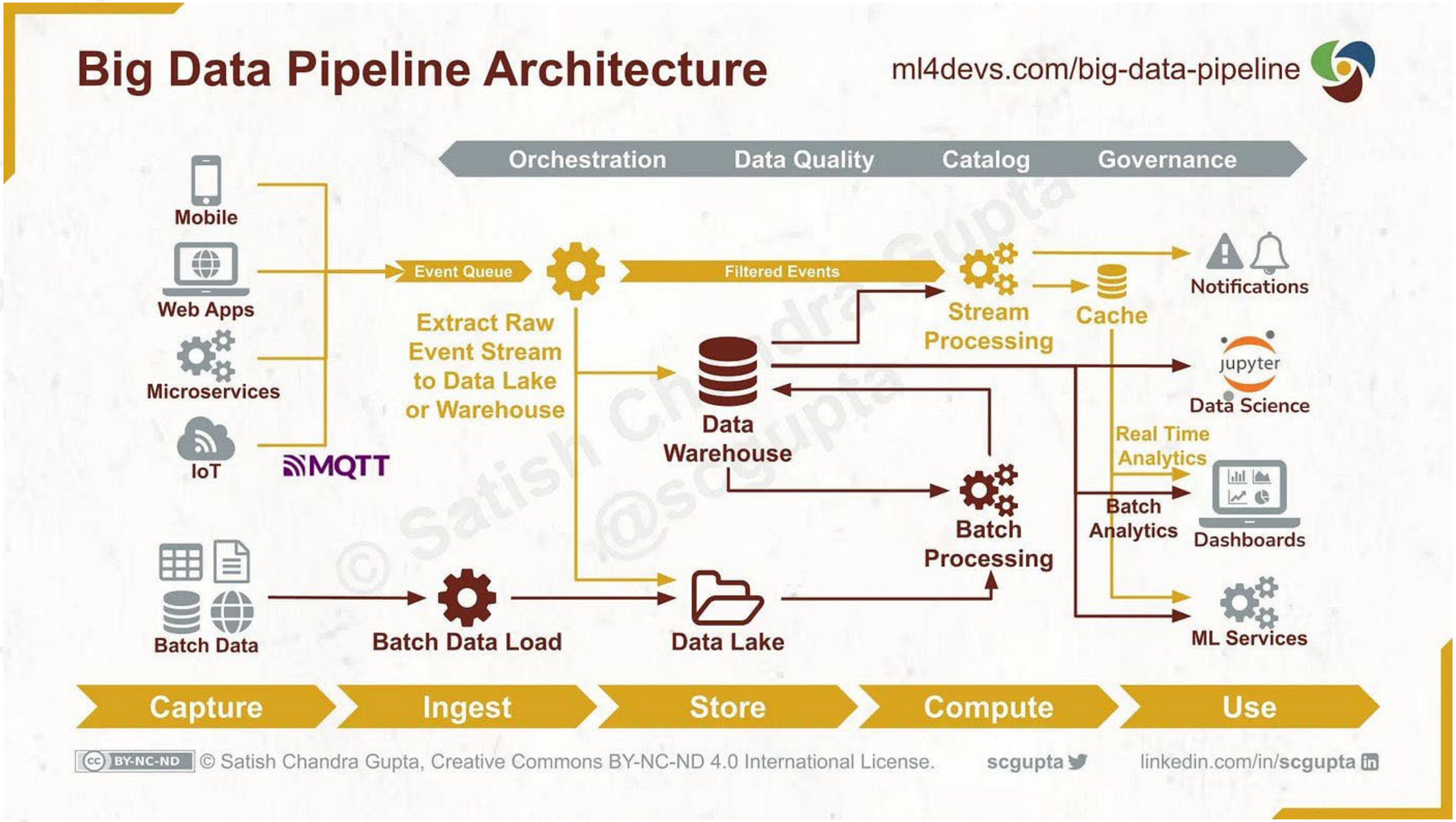
What Does Digital Chemistry Mean To You?

“Digital chemistry” means different things to different people.

- Computational biophysicist – molecular simulations?
- Synthetic organic chemistry – AI in synthesis?
- Pharmaceutical chemist – AI in drug discovery?
- Structural biology – databases? Protein structure prediction?

Asked 4 diverse colleagues the same question, got 4 answers.

Digital Chemistry Means New Data Pipelines



"We've decided to take big data to the next level - humongous data" (from vox.com cartoon)

Paradigm Shift for Chemistry Data

- How we obtain data?
  - experimental / computational workflows.
- How we store data?
  - infrastructure challenges as data gets huge.
- How we curate data?
  - how to make it understandable to others.
- How we share it with others.
  - infrastructure + curation + what/when?

New ways of working with data, needs new skills, approaches, *etc*.

Paradigm Shift for Chemistry Data

- How we obtain data?
  - experimental / computational workflows.

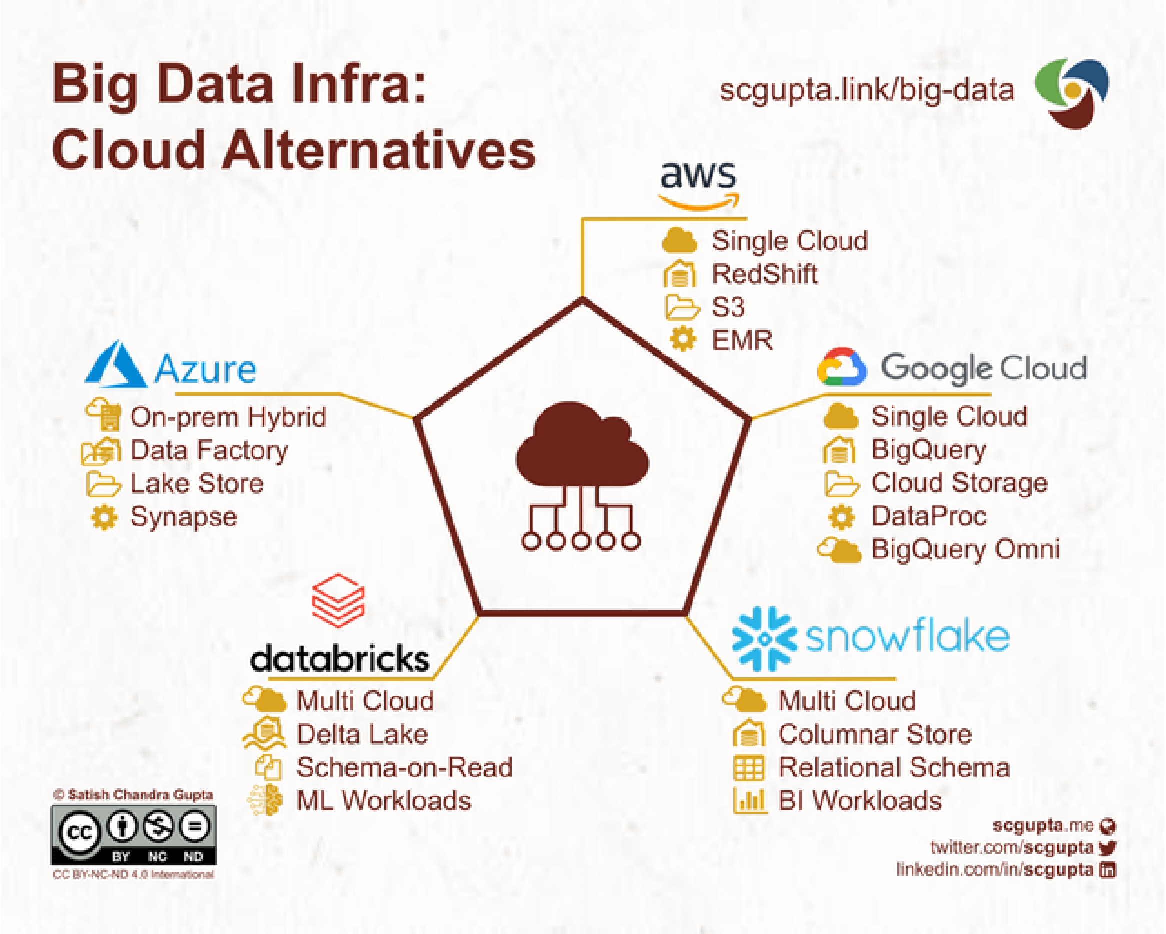
- How we store data?
  - infrastructure challenges as data gets huge.

- How we curate data?
  - how to make it understandable to others.

- How we share it with others.
  - infrastructure + curation + what/when?

Different disciplines have different data needs (Excel spreadsheets to PB of data) and different expectations of digital chemistry.

Infrastructure – Handling All The Data



Big Data Infra:
Cloud Alternatives

scgupta.link/big-data

aws
- Single Cloud
- RedShift
- S3
- EMR

Azure
- On-prem Hybrid
- Data Factory
- Lake Store
- Synapse

Google Cloud
- Single Cloud
- BigQuery
- Cloud Storage
- DataProc
- BigQuery Omni

databricks
- Multi Cloud
- Delta Lake
- Schema-on-Read
- ML Workloads

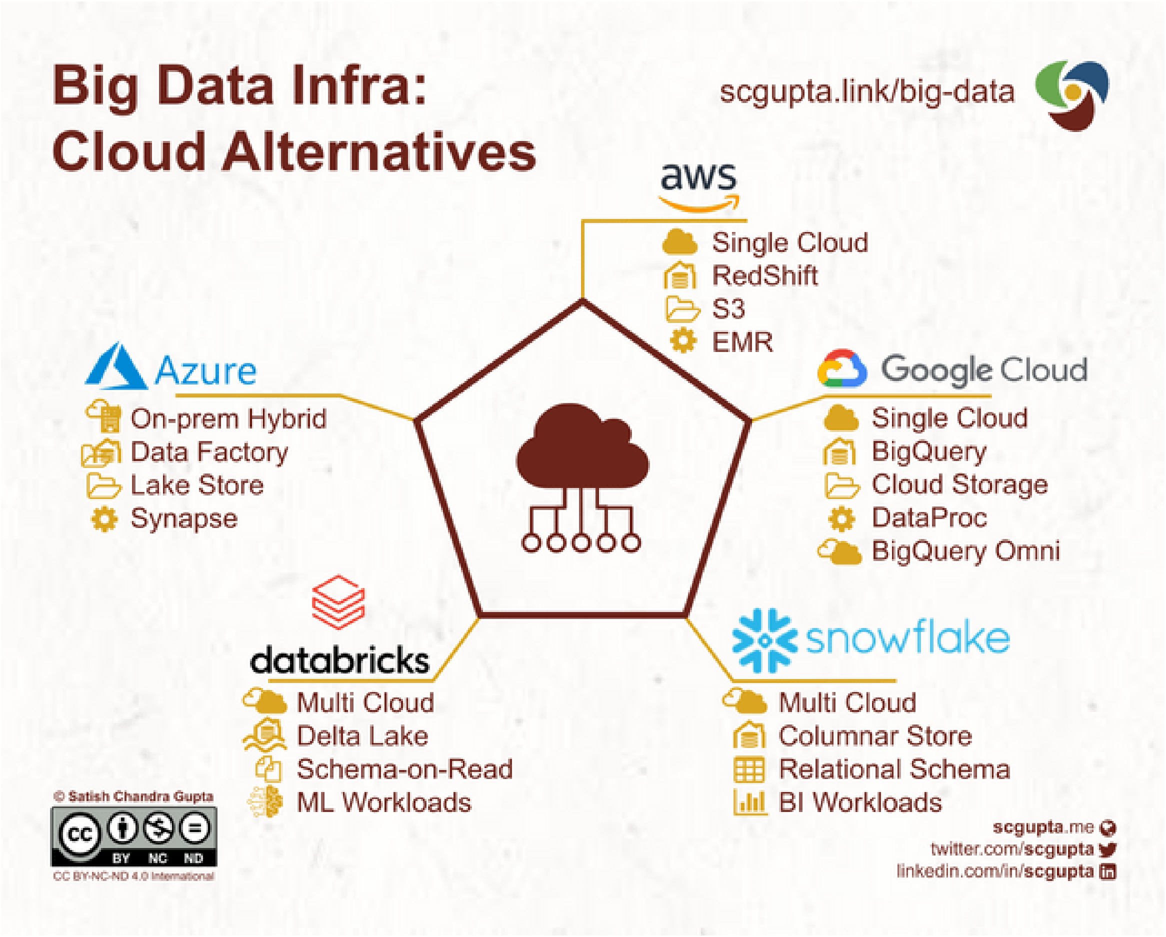
snowflake
- Multi Cloud
- Columnar Store
- Relational Schema
- BI Workloads

scgupta.me
twitter.com/scgupta
linkedin.com/in/scgupta

Give us more infrastructure! Thank you for coming to my talk.

Georgia Tech

Infrastructure – Handling All The Data



Big Data Infra: Cloud Alternatives

scgupta.link/big-data

aws
- Single Cloud
- RedShift
- S3
- EMR

Google Cloud
- Single Cloud
- BigQuery
- Cloud Storage
- DataProc
- BigQuery Omni

Azure
- On-prem Hybrid
- Data Factory
- Lake Store
- Synapse

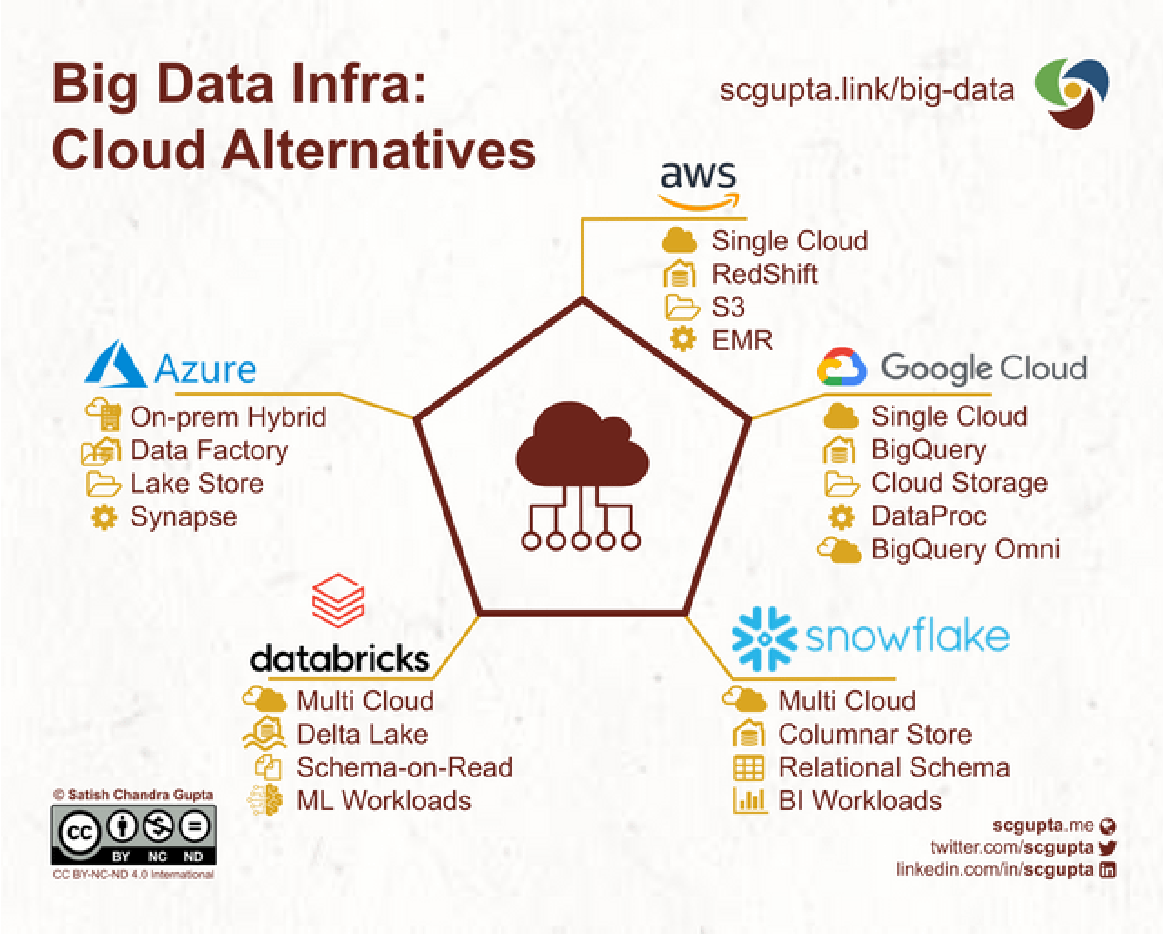
databricks
- Multi Cloud
- Delta Lake
- Schema-on-Read
- ML Workloads

snowflake
- Multi Cloud
- Columnar Store
- Relational Schema
- BI Workloads

scgupta.me
twitter.com/scgupta
linkedin.com/in/scgupta

Need an "infrastructure revolution" – for analyzing the data, storing the data, disseminating the data.

Infrastructure – Handling All The Data



Big Data Infra: Cloud Alternatives

scgupta.link/big-data

aws
- Single Cloud
- RedShift
- S3
- EMR

Google Cloud
- Single Cloud
- BigQuery
- Cloud Storage
- DataProc
- BigQuery Omni

Azure
- On-prem Hybrid
- Data Factory
- Lake Store
- Synapse

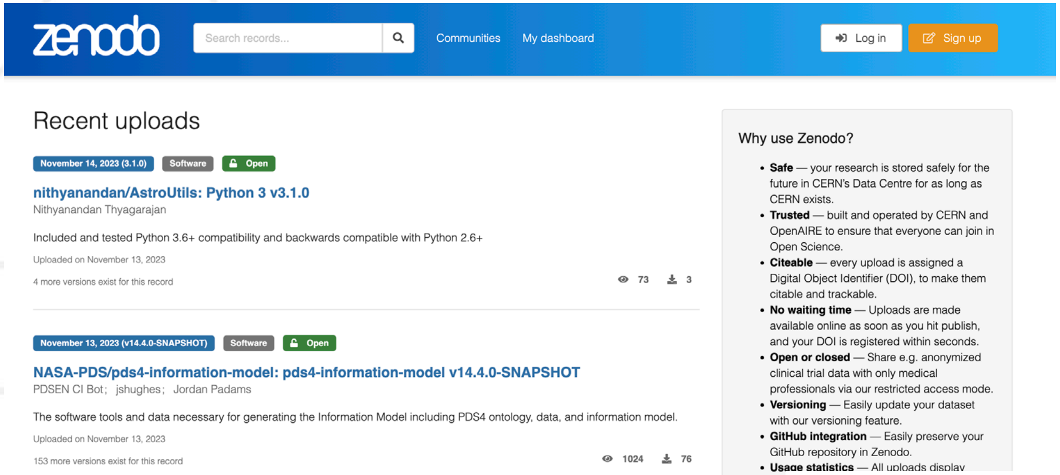
databricks
- Multi Cloud
- Delta Lake
- Schema-on-Read
- ML Workloads

snowflake
- Multi Cloud
- Columnar Store
- Relational Schema
- BI Workloads

scgupta.me
twitter.com/scgupta
linkedin.com/in/scgupta

Skills training – how do the next generation of chemists interface with this infrastructure?

Paradigm Shift – Transition to Open Data



Recent uploads

November 14, 2023 (3.1.0) | Software | Open

nithyanandan/AstroUtils: Python 3 v3.1.0
Nithyanandan Thyagarajan

Included and tested Python 3.6+ compatibility and backwards compatible with Python 2.6+

Uploaded on November 13, 2023

4 more versions exist for this record — 73 — 3

November 13, 2023 (v14.4.0-SNAPSHOT) | Software | Open

NASA-PDS/pds4-information-model: pds4-information-model v14.4.0-SNAPSHOT
PDSEN CI Bot; jshughes; Jordan Padams

The software tools and data necessary for generating the Information Model including PDS4 ontology, data, and information model.

Uploaded on November 13, 2023

153 more versions exist for this record — 1024 — 76

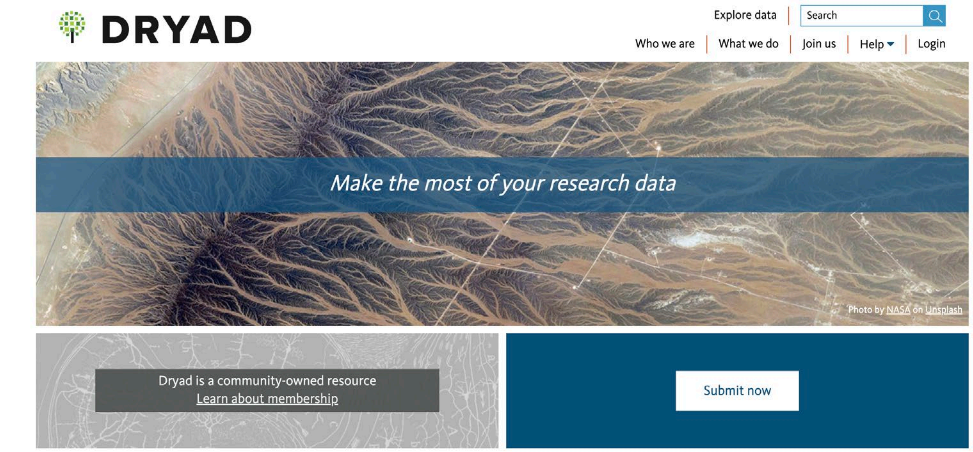
November 13, 2023 (v1.0.5) | Software | Open

Zoobot: Adaptable Deep Learning Models for Galaxy Morphology
Walmsley, Mike; Allen, Campbell; Aussel, Ben

Why use Zenodo?

- **Safe** — your research is stored safely for the future in CERN's Data Centre for as long as CERN exists.
- **Trusted** — built and operated by CERN and OpenAIRE to ensure that everyone can join in Open Science.
- **Citeable** — every upload is assigned a Digital Object Identifier (DOI), to make them citable and trackable.
- **No waiting time** — Uploads are made available online as soon as you hit publish, and your DOI is registered within seconds.
- **Open or closed** — Share e.g. anonymized clinical trial data with only medical professionals via our restricted access mode.
- **Versioning** — Easily update your dataset with our versioning feature.
- **GitHub integration** — Easily preserve your GitHub repository in Zenodo.
- **Usage statistics** — All uploads display sta

RCSB PDB
PROTEIN DATA BANK



Make the most of your research data

Photo by NASA on Unsplash

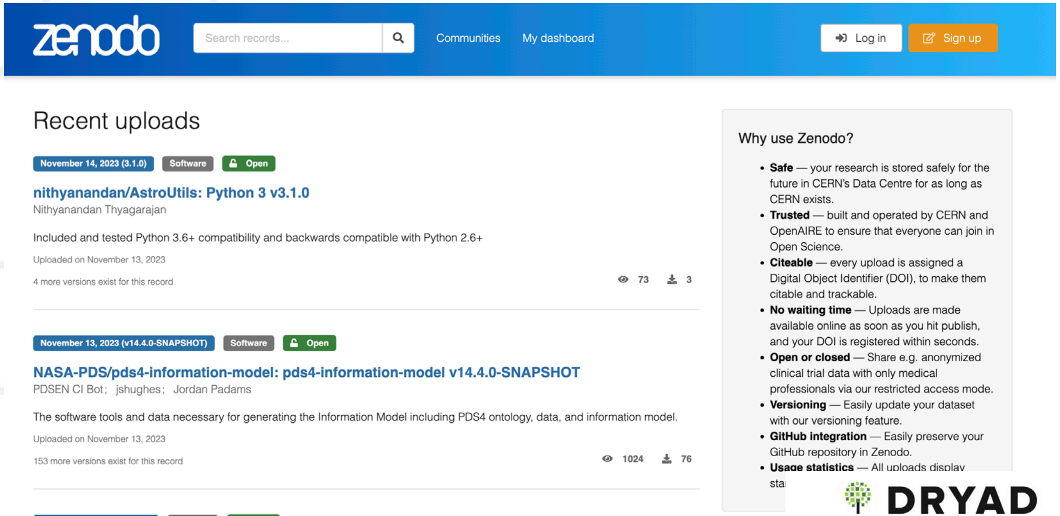
Dryad is a community-owned resource
Learn about membership

figshare

Not just papers need to be open access!

Georgia Tech

Paradigm Shift – Transition to Open Data



zenodo

Search records...

Communities

My dashboard

Log in

Sign up

Recent uploads

November 14, 2023 (3.1.0) Software Open

nithyanandan/AstroUtils: Python 3 v3.1.0

Nithyanandan Thyagarajan

Included and tested Python 3.6+ compatibility and backwards compatible with Python 2.6+

Uploaded on November 13, 2023

4 more versions exist for this record

73 3

November 13, 2023 (v14.4.0-SNAPSHOT) Software Open

NASA-PDS/pds4-information-model: pds4-information-model v14.4.0-SNAPSHOT

PDSEN CI Bot; jshughes; Jordan Padams

The software tools and data necessary for generating the Information Model including PDS4 ontology, data, and information model.

Uploaded on November 13, 2023

153 more versions exist for this record

1024 76

November 13, 2023 (v1.0.5) Software Open

Zoobot: Adaptable Deep Learning Models for Galaxy Morphology

Walmsley, Mike; Allen, Campbell; Aussel, Ben

Why use Zenodo?

- **Safe** — your research is stored safely for the future in CERN's Data Centre for as long as CERN exists.
- **Trusted** — built and operated by CERN and OpenAIRE to ensure that everyone can join in Open Science.
- **Citeable** — every upload is assigned a Digital Object Identifier (DOI), to make them citable and trackable.
- **No waiting time** — Uploads are made available online as soon as you hit publish, and your DOI is registered within seconds.
- **Open or closed** — Share e.g. anonymized clinical trial data with only medical professionals via our restricted access mode.
- **Versioning** — Easily update your dataset with our versioning feature.
- **GitHub integration** — Easily preserve your GitHub repository in Zenodo.
- **Usage statistics** — All uploads display sta

RCSB PDB

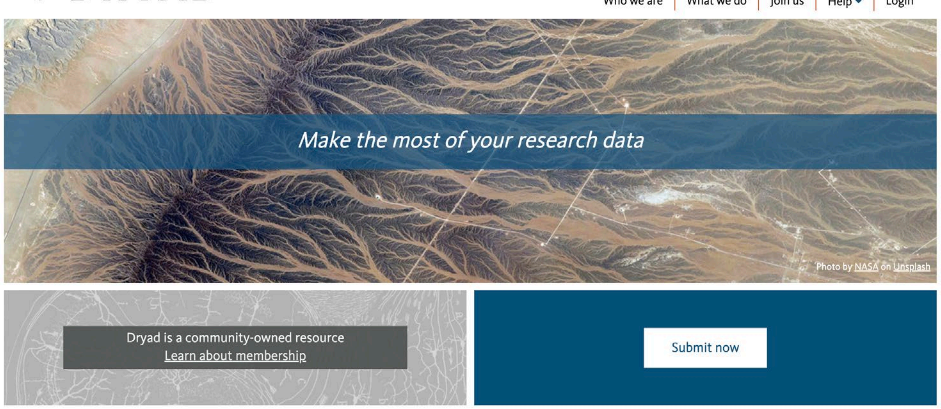
PROTEIN DATA BANK

DRYAD

Explore data

Search

Who we are | What we do | Join us | Help | Login



Make the most of your research data

Photo by NASA on Unsplash

Dryad is a community-owned resource

Learn about membership

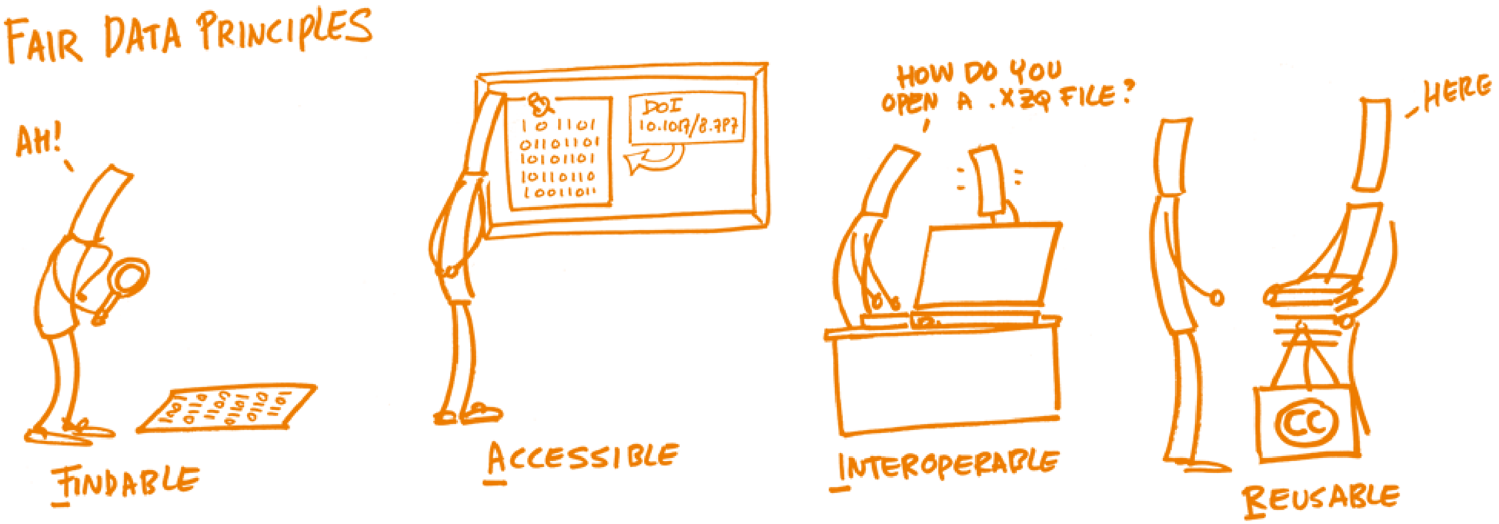
Submit now

figshare

Training to identify reputable / safe repositories to store data.

Georgia Tech

FAIR and Digital Standards for Chemistry



GO FAIR

http://go-fair.org

GO-FAIR + IUPAC working on fair standards for chemistry.

Image from Foster Open Science

GO-FAIR Chemistry Implementation Network

GO-FAIR ChIN guiding principles:

- Findable chemistry data.
- Reusable code and data – validation, compilation/aggregation, incorporation into future work, data mining.
- The use of standards at source and throughout the information cycle.
- Availability and accessibility of tools and infrastructure.

GO FAIR

http://go-fair.org

Georgia Tech

GO-FAIR Chemistry Implementation Network

- The use of persistent identifiers and machine readability at the core.
- Management and oversight of standards.
- Use of general data standards outside of chemistry where appropriate and FAIR in their implementation.
- Enable and promote use of chemical data standards in other disciplines that work with chemical data.

http://go-fair.org

IUPAC Push for Digital Standards

- The IUPAC International Chemical Identifier (InCHI).
- ThermoML (XML based IUPAC standard for thermodynamic property data).
- JCAMP-DX (IUPAC standard format family for spectral data exchange).
- In progress: IUPAC SMILES+, machine accessible periodic table, standard for FAIR data management of spectroscopic data, making IUPAC assets FAIR, and more!

https://iupac.org/what-we-do/digital-standards/

Training: Chemistry Degrees Are Changing!

- Chemistry degree 1998-2002: minimal comp chem, only in final year (enough to make me be a theorist though!)

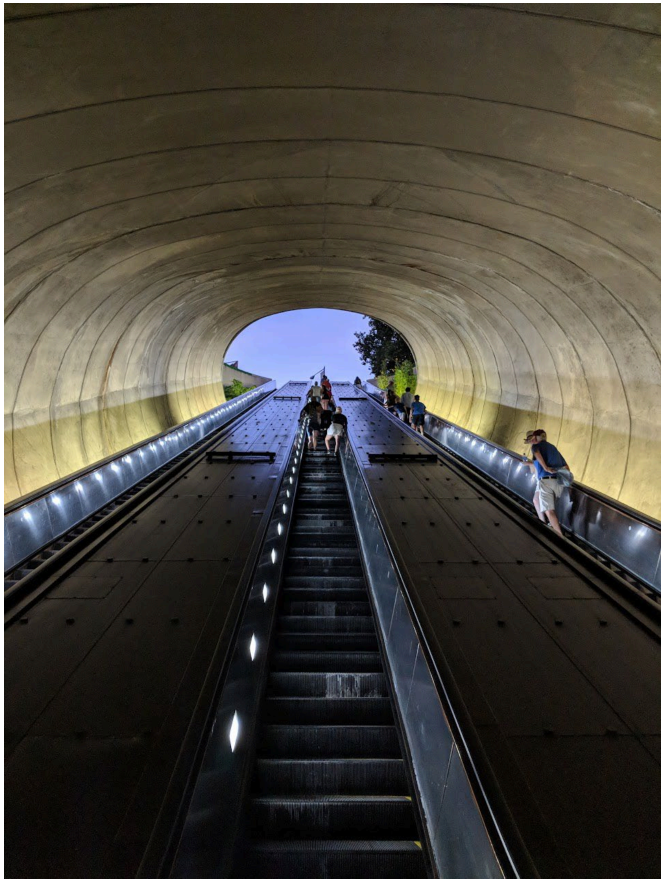
- Revolution in skills training for chemists – so many new skills you need to learn.

- EU “Traineeship in Digital Skills” as part of mobility schemes.

- Dedicated courses in “Digital Chemistry” at many institutions.

Next generation of chemists are going to live with data, computers and workflows that would be unrecognizable 20 years ago!

What Does the Future Hold?



Interdisciplinarity | Big Data | Complexity | Integration