PSDI

PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Connecting Infrastructures: The Physical Sciences Data Infrastructure (PSDI) in the UK

CoRDI Karlsruhe, 13th Sept 2023

Lead Team	
STFC Scientific Computing	**University of Southampton**
Juan Bicarregui	Simon Coles
Vasily Bunakov	**Nicola Knight**
Brian Matthews	Jeremy Frey
Barbara Montanari	

Engineering and Physical Sciences Research Council

UK Research and Innovation

https://www.psdi.ac.uk/

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Aim of PSDI

Data is a major driver of research in Physical Sciences, but it is fragmented

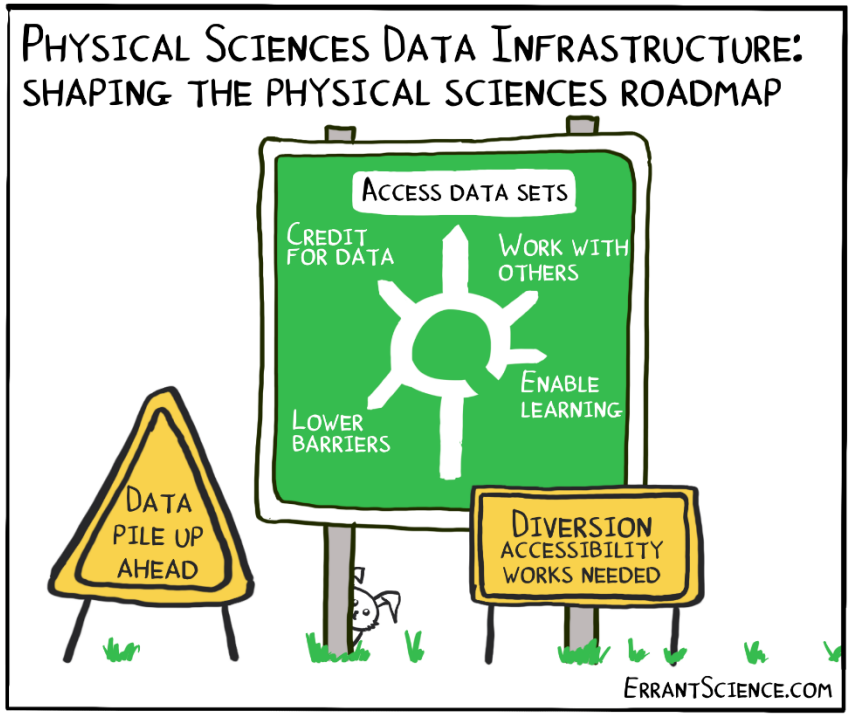
PSDI will provide

A data infrastructure that

connects existing

experimental and computational facilities

within Physical Sciences and beyond



Through PSDI:

- Find and Access to reference quality data from commercial and open sources
- Combine data from different sources
- Share data, software and models including experimental and simulation data
- Use AI to explore data
- Learn how to make the results of their research open and FAIR

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

PSDI: filling a Gap in Provision

- **Other countries** have initiatives underway **in this domain**, e.g.
  - USA: Materials Genome Initiative
  - Japan: NIMS
  - European data infrastructures, such as E-CAM, MaX and NOMAD
  - German National Research Data Infrastructure (NFDI)

- **Other domains** have initiatives underway **in the UK**, e.g.
  - EBI in Life Sciences
  - NERC Data centres in Environmental Science
  - UK Data Archive in Social Science

We are building a UK, Physical Science, Data Infrastructure

- Supporting Chemistry, Materials and related disciplines
- Traversing to and interfacing with Life, Medical, Engineering and Environmental Sciences

PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

The bigger picture in the UK

Facilities, Institutes & Hubs

Examples:

- Catalysis Hub
- CCFE
- Central Laser Facility
- Diamond
- Future Manufacturing Hub
- ISIS
- Royce Institute
- ATI

National Research Facilities

Examples:

- HarwellXPS
- NXCT
- NCS
- PSDS
- SuperSTEM
- UK High Field Solid-State NMR
- XMaS

Computational Initiatives

Examples:

- CCP5++
- CCP9
- CCP Biosim
- CCP EngSci
- CCPi
- CCP NC
- CCP NTH
- CCP QC
- CCP SAS
- CCP Turbulence
- CCP WSI
- SSI
- UK society of RSE
- HEC Biosim
- HEC Plasma
- MCC
- UKCTRF
- UKCP
- UKTC
- CoSeC
- EPSRC Tier2
- ExCALIBUR
- STFC Hartree Centre
- ARCHER

Research Institutions, Groups and Laboratories

Examples:

- Equipment Infrastructures
- Equipment Facilities
- University Labs
- ELNs
- Repositories
- Local Computing Resources

PHYSICAL SCIENCES DATA INFRASTRUCTURE

PSDI: Outline Timeline

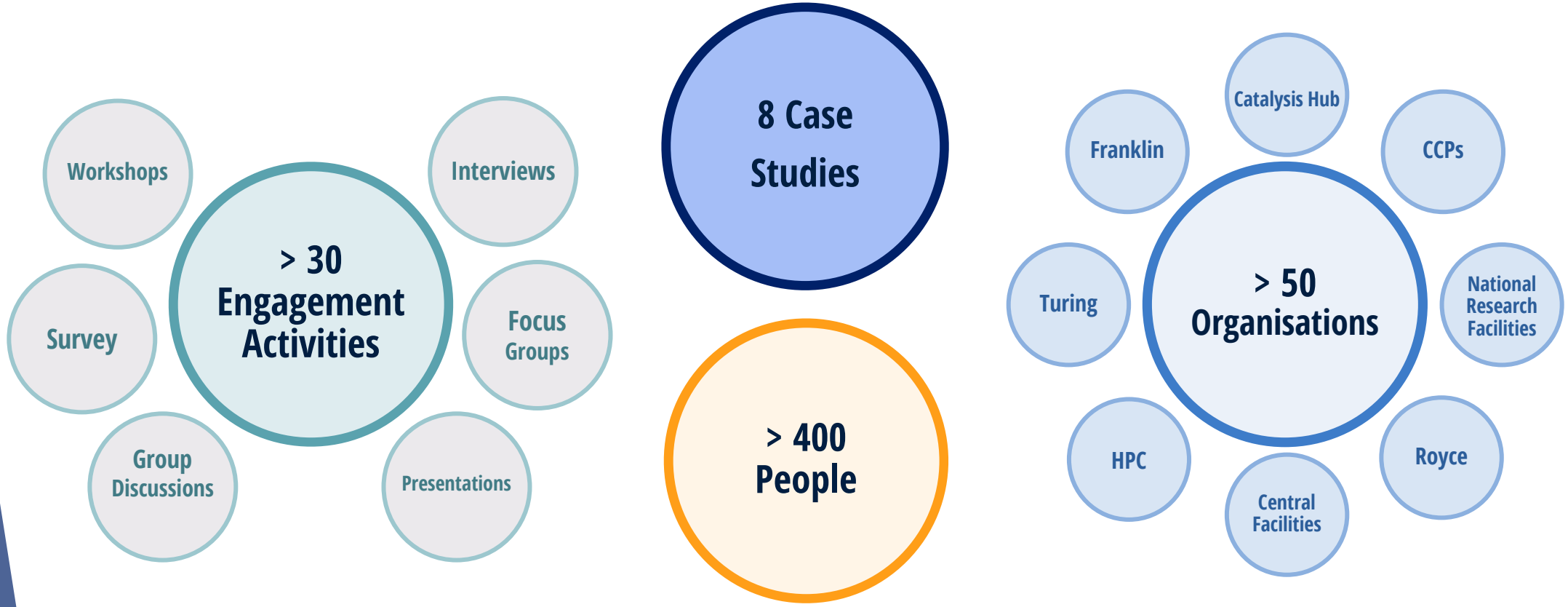
- Feb 2021: EPSRC statement of need for large-scale infrastructure
- Nov 2021 – Mar 2022: PSDI Pilot
  - Initial PSDI engagement & design including 8 case studies
- Oct 2022 – Sept 2023: PSDI Phase 1b
  - Funded as part of UKRI DRI Phase 1b
  - Begin initial development of PSDI Hub components
  - 5 pathfinders to create exemplar tools, data and services
- Oct 2023 – Mar 2025: PSDI (Phase 1c)
  - To be funded from next phase of UKRI DRI
  - Complete initial development and deploy Platform and 5 Pathfinders
  - Develop further Pathfinders
- Beyond Mar 2025 (Phase 2: Operational and Further development)

Single 30 month development project

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Pilot Phase at a Glance

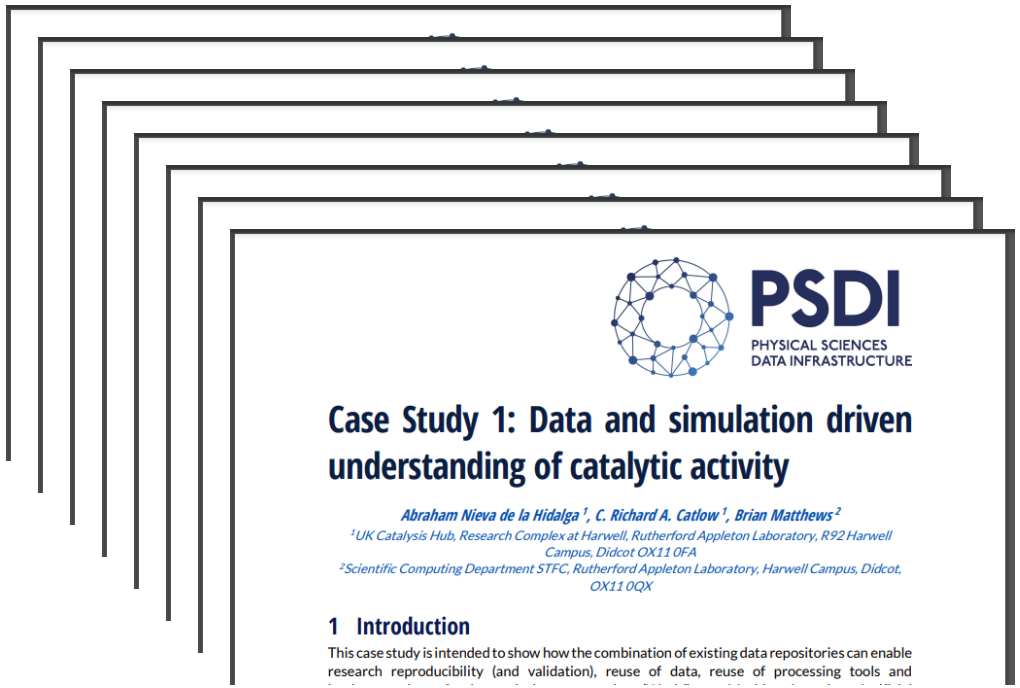
Catalysis
Spectroscopy
Structural Representation
ELNs
Materials
Data Curation



PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Pilot Outputs

You can find the PSDI *pilot* final report on the PSDI website….



…and the reports from each case study

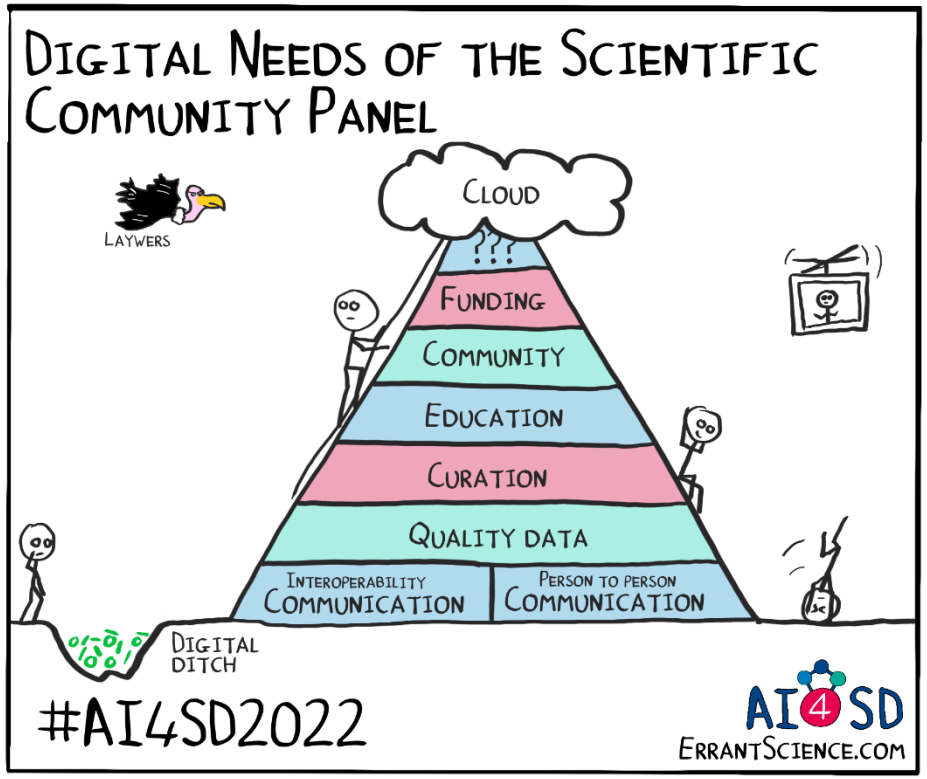
Available on PSDI website: psdi.ac.uk
And the PSDI Zenodo Community

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Current Work - Phase 1b

Platform, Pathfinders and Hub

- **Platform**
  - Requirements Analysis
  - Capacity Planning
  - System Architecture design
  - Component testing
  - Beginning Build
- **Pathfinder project**
  - Initial development and implementation
- **Hub**: Communications, Governance, Planning,…



PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Our areas

- Physical Sciences is a broad area
- Initial focus areas:
  - Catalysis
  - Biomolecular simulation
  - Materials
  - Atomistic simulations
  - NMR
  - Process Recording
  - Data Collections

NFDI Alignment

NFDI4Chem

DAPHNE 4NFDI

NFDI4Cat
NFDI for Catalysis-Related Sciences

FAIRmat

NFDI MatWerk

NFDi4ing

PUNCH 4NFDI

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

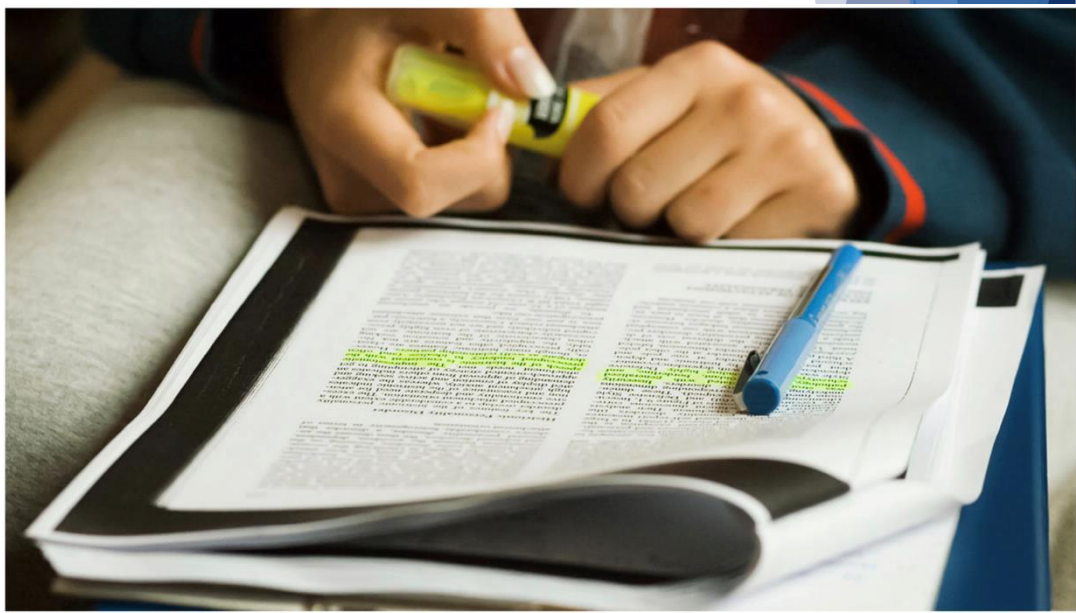
An Example: Biomolecular Simulations

Model → Computation → Trajectory

- Run 10s of simulations to generate data
- Apply know-how to extract science from data
- Publish paper

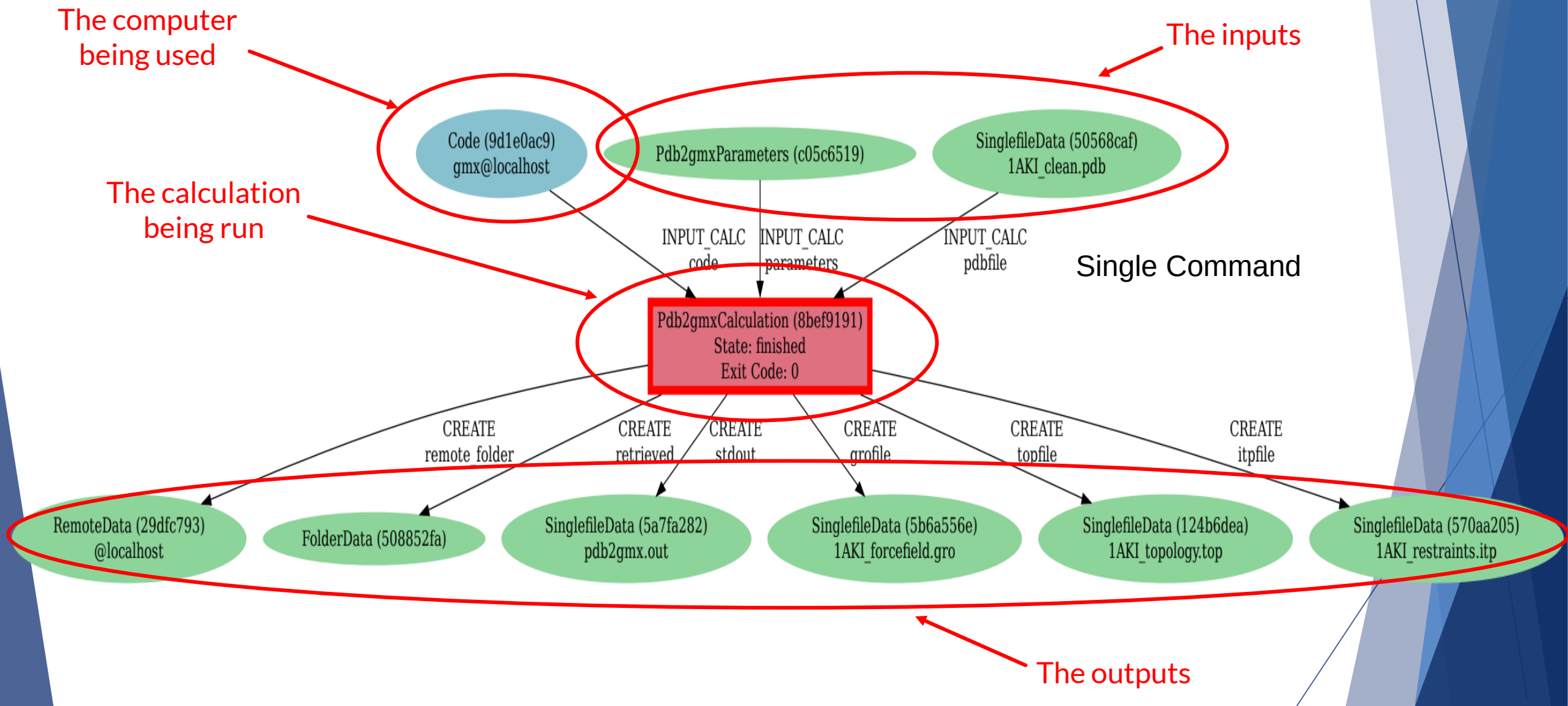
But ….

- Paper does not include all details needed to **repeat** simulation
- Citations do not give **credit** for *all* resources used



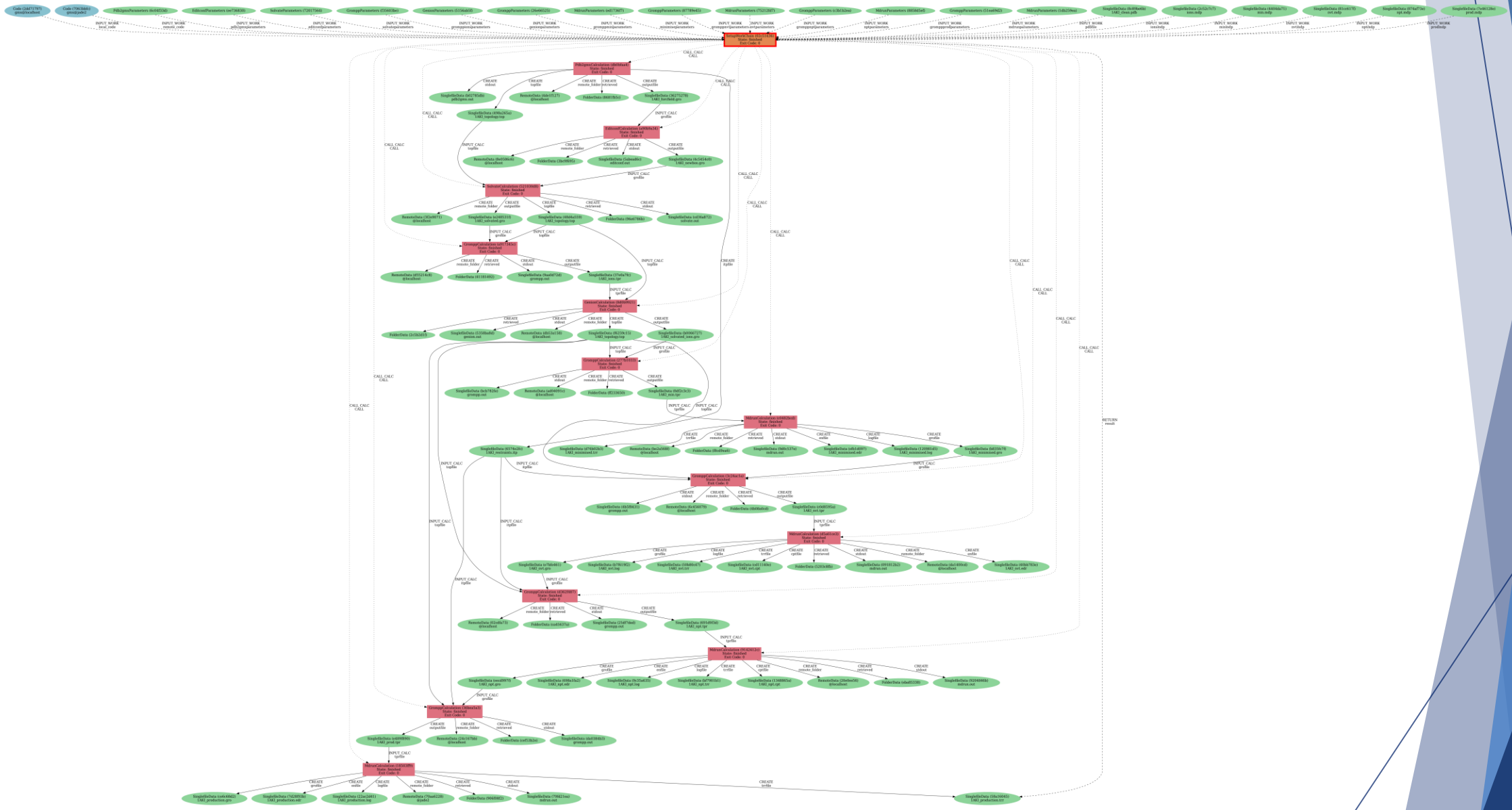
PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Provenance map of a Single Command in a Simulation



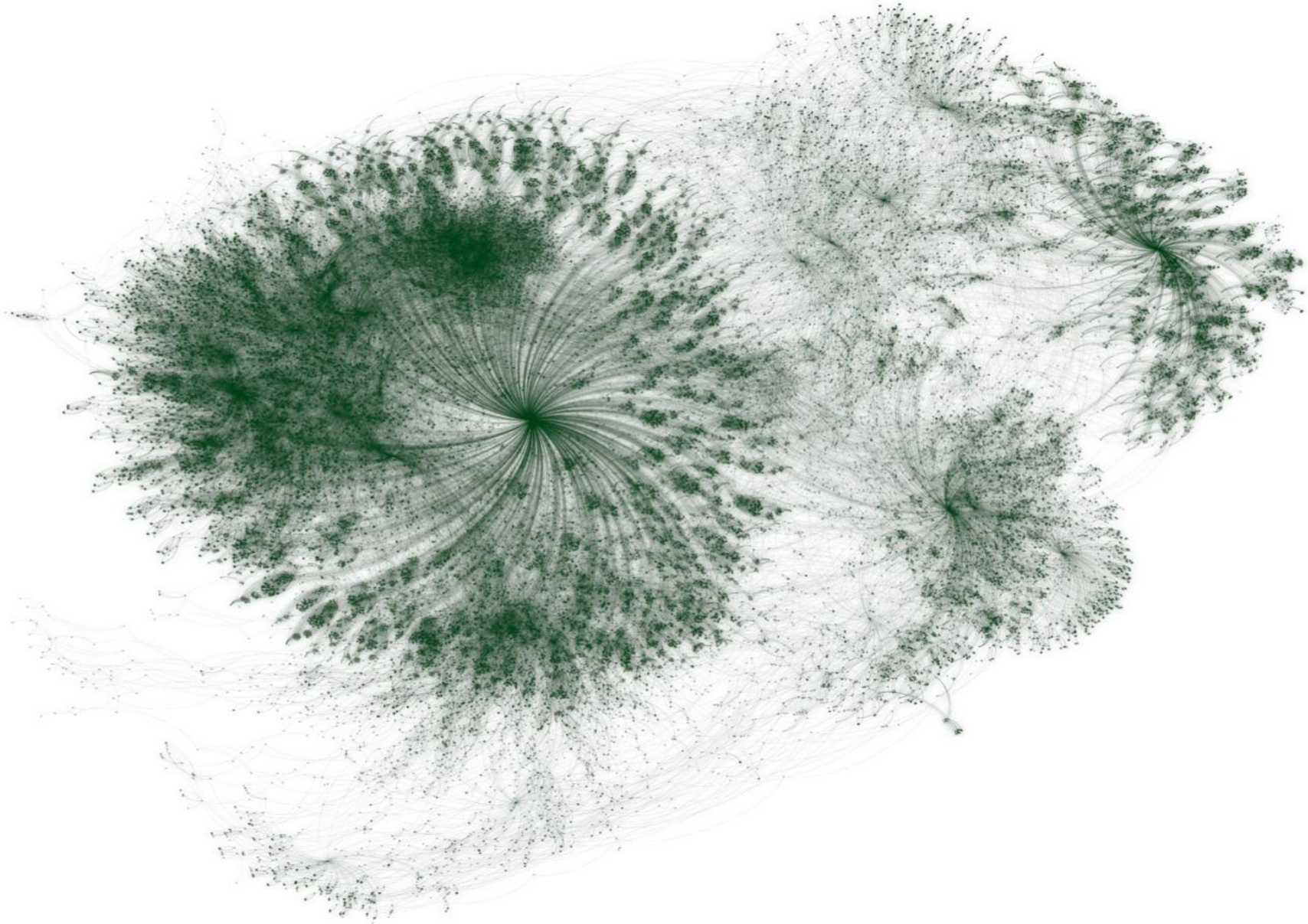
PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Creating a model... (not yet a simulation)



PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

An Entire Study



PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Process Orchestration PathFinder: User Environment Prototype

- Built prototype to **capture full data provenance** using AiiDA
- Building on GROMACS software (70% of users in UK HPC Biosim Consortium)
- Designed to mimic working with native package (command line driven)
- Simple to install and setup our plugin

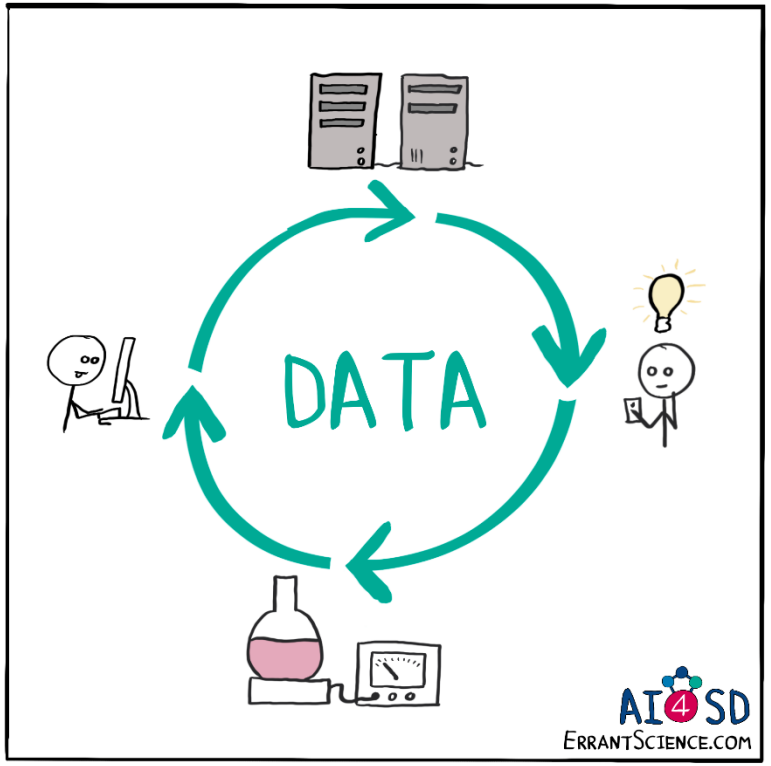
Normal command:
gmx pdb2gmx -f prot.pdb -ff oplsaa -water spce -o prot.gro -p prot.top -i prot.itp

Capture provenance with AiiDA:
gmx_pdb2gmx -f prot.pdb -ff oplsaa -water spce -o prot.gro -p prot.top –i prot.itp

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

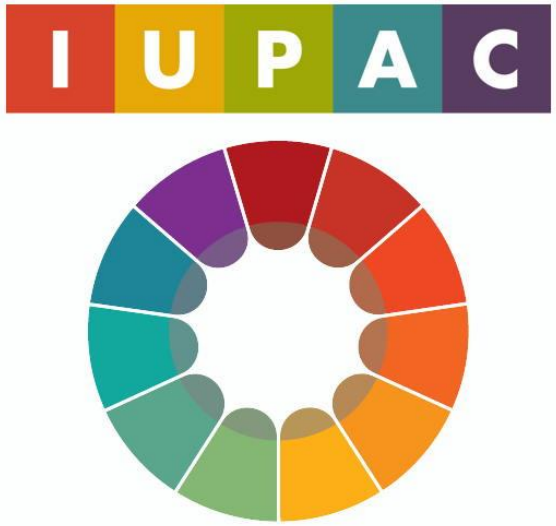
Collaboration Topics

- Metadata
  - Standards
  - Collection
- Ontologies
- Standards
- Best practices
- Training
- Tool development
- Publishing & Supplementary Data



PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

International Collaboration (among others)



Collaboration with NFDI

PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

Information sharing on key people within projects across NFDI / PSDI

Strengthening the connection and making it easier to collaborate

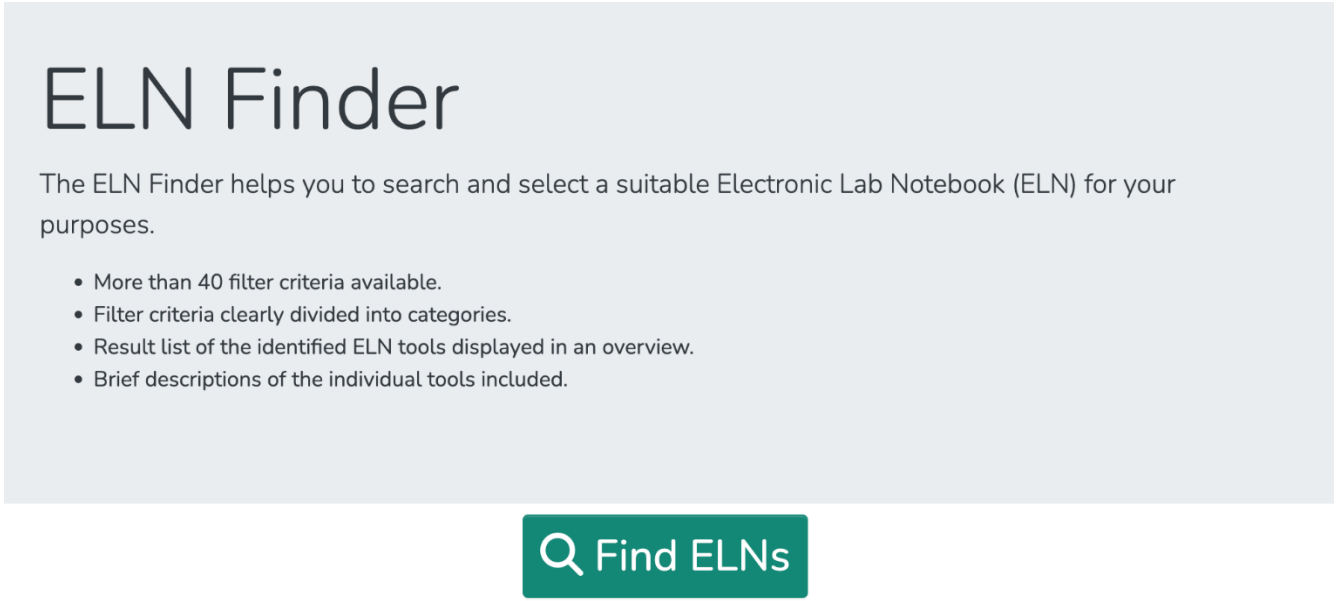
If you want to get involved with this please contact Nicola Knight (PSDI) or John Jolliffe (NFDI4Chem)

PSDI	Topic	NFDI4Chem	NFDI4Cat	DAPHNE4NFDI	FAIRmat	NFDI MatWerk	PUNCH4NFDI
PSDI Website Twitter LinkedIn Zenodo Community YouTube	Online Resources	NFDI4Chem Knowledge Base NFDI4Chem Website Twitter LinkedIn ELN Finder Terminology Service Overview of NFDIs Proposal Instagram Search Service		DAPHNE4NFDI website Proposal			
Project Coordinator, Nicola Knight, email [illegible]	Management and Coordination	[illegible]		[illegible]			
[illegible] WP2 Lead: Nicola Knight, email	Skills & Training	[illegible]		[illegible]			
[illegible] WP2 Lead: Nicola Knight, email	Outreach & Networking + representation in international initiatives	[illegible]		[illegible]			
	Storing, Sharing and making Data available to others (Repositories, Archives etc.)	[illegible]		[illegible]			
[illegible]	ELNs / Smartlab / tools for capturing and managing data / tools / workflows	[illegible]		[illegible]			

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

ELN Finder

https://eln-finder.ulb.tu-darmstadt.de



- Collaboration of PSDI Process recording work with ELN Finder in NFDI4Chem
- Eliminates duplication of services and effort

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

See what we are up to

- We are running a webinar series
  - Tues 3rd Oct 2pm – Experimental Data Capture in catalysis: producing publish ready data from processing and analysis processes
  - Oct – Exact date TBC - Process orchestration in Biomolecular Simulation

    Register now: https://www.psdi.ac.uk/events/
- Check out our website
  - www.psdi.ac.uk
- Find us on social media
  - Twitter / X: @PSDI_UK
  - LinkedIn: PSDIUK
- Follow our YouTube Channel: youtube.com/@PSDI_UK
- Join our JISCmail mailing list via our website footer: PSDI

PSDI

PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Any Questions?