PSDI

PHYSICAL SCIENCES
DATA INFRASTRUCTURE

UK Physical Sciences Data Infrastructure (PSDI) initiative

26th October 2023

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

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

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Aim(s) of PSDI

Support **Data as** a major driver of research in Physical Sciences

PSDI will provide a data infrastructure that **connects existing** experimental and computational facilities within Physical Sciences and beyond

- A platform for data collection, sharing, aggregation, integration and curation

Building Bridges

- Sustaining data resources beyond lifespan of individual research projects

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)

UK catch up

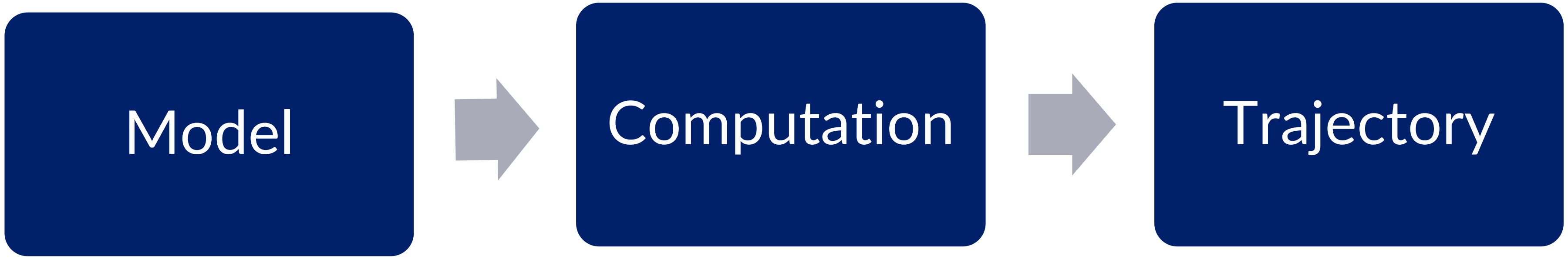
- **Other domains** have initiatives underway **in the UK**, e.g.
  - EBI in Life Sciences
  - NERC Data centres in Environmental Science
  - UK Data Archive in Social Science

Physical Sciences catch up

We are building a UK, Physical Science, Data Infrastructure

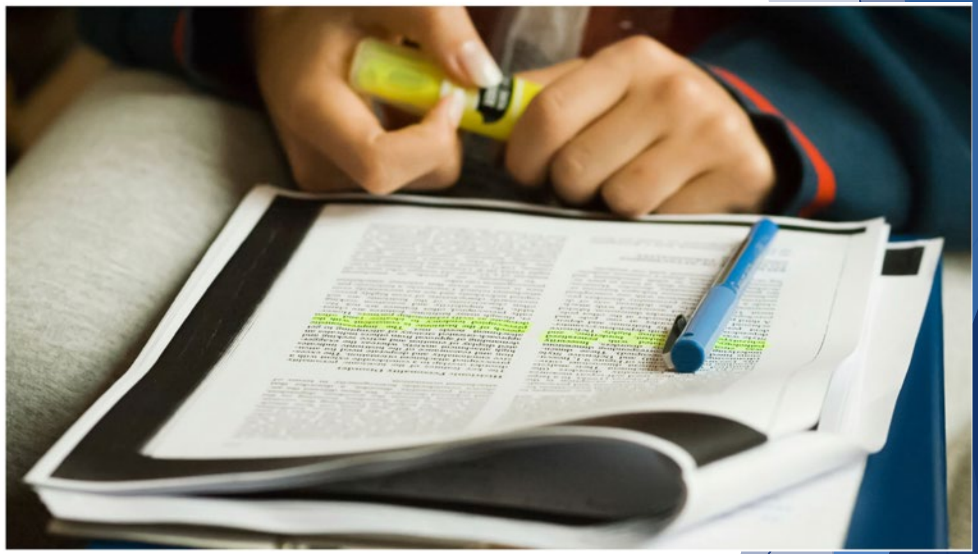
- Supporting Chemistry, Materials and related disciplines
- Traversing to and interfacing with Life, Medical, Engineering and Environmental Sciences through federated systems

An Example:
Biomolecular Simulations



Model → Computation → Trajectory

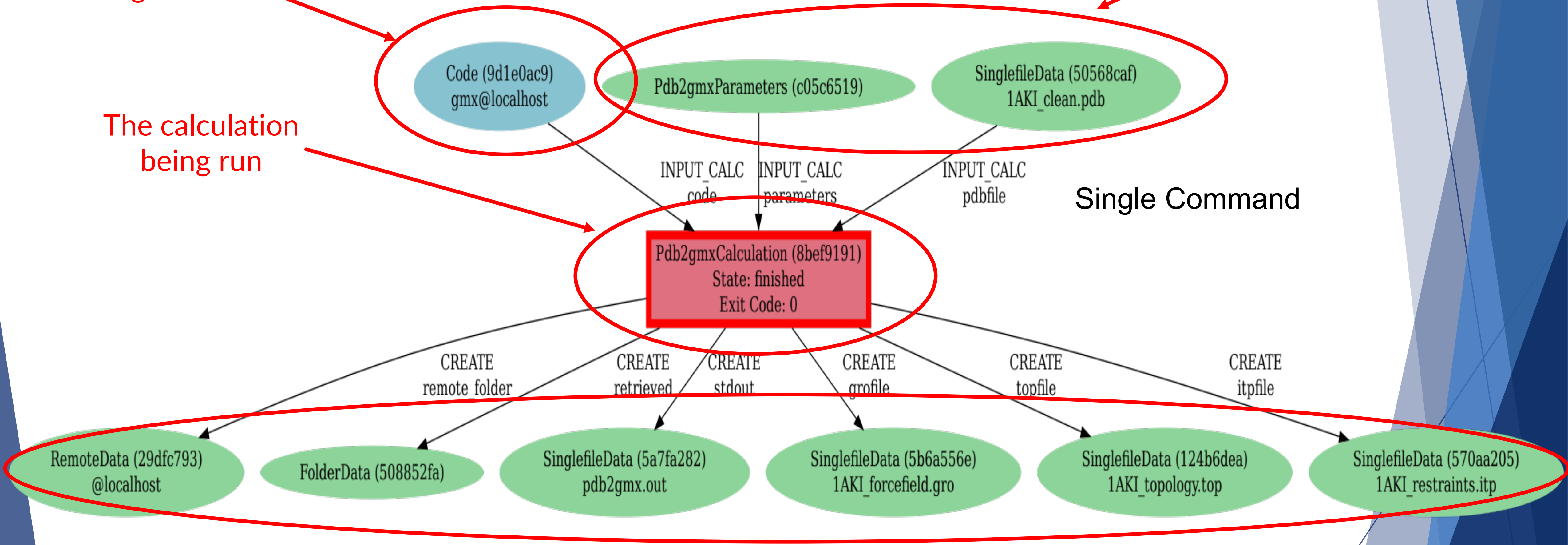
- Run 10s of simulations to generate data
- Apply know-how to extract science from data
- Publish paper



But ….

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

Provenance map of a Single Command in a Simulation



The computer being used

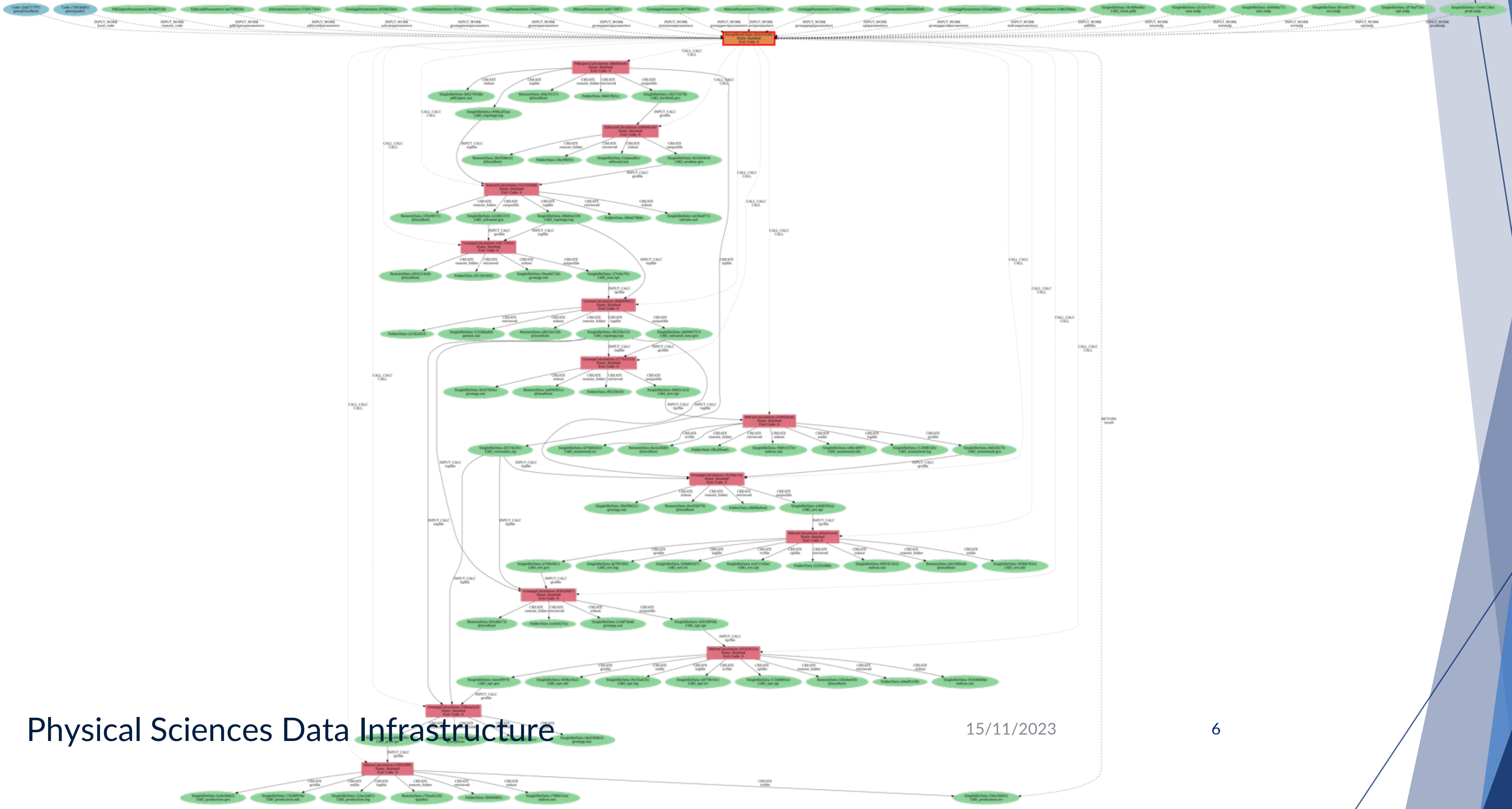
The inputs

The calculation being run

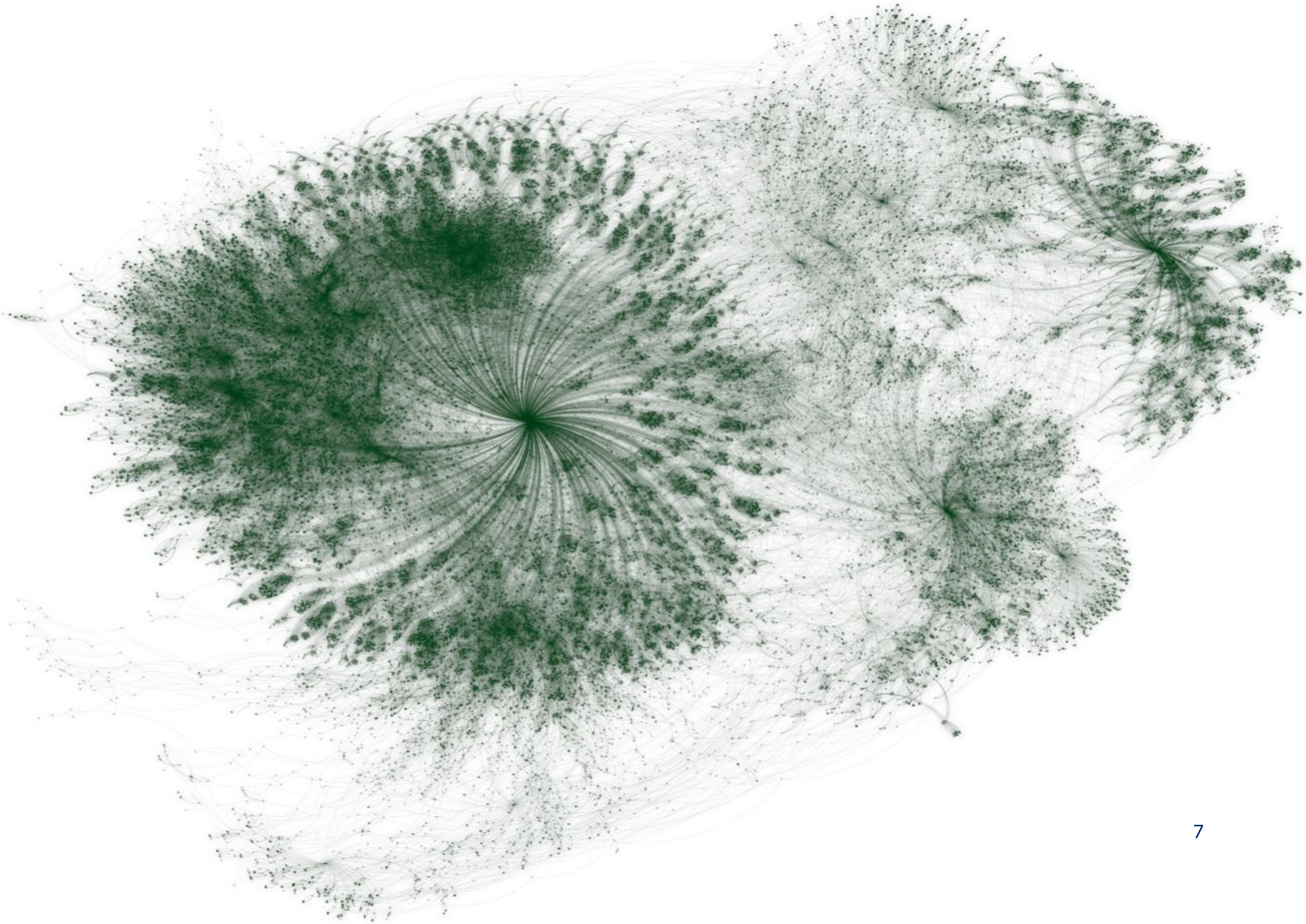
Single Command

The outputs

Creating a model… (not yet a simulation)



An Entire Study



PSDI PathFinder on Research Process Orchestration

Main aim is to improve data practices in domain – align with FAIR principles

- Prototype tools to **capture full data provenance** for model creation, simulation and analytics (FAIR)
- Prototype infrastructure tools to **store, access, find and share** data (FAIR)
- **Collect** and Integrate existing small scale, disparate data sources
- Maintain **compatibility** with other data initiatives (EBI, EU and US)
- Link **computational and experimental** data sources
- “I” (FAIR) **Integrations** *not yet in scope* of this pathfinder (excellent projects in CCPBioSim)

James Gebbie & Jas Kalayan

Pilot Recommendations

13 recommendations in 4 areas:

Connecting existing infrastructures

3 Recommendations: connecting existing research data services, beyond the lifespan of individual projects, co-operation and co-creation between all stakeholder organisations

Best Use of Data

4 Recommendations: developing a toolkit for publishing, access to provenanced data, tools for reproduceable data processing, support for transforming data to knowledge

Best Use of People

4 Recommendations: co-ordination for community activities and input, community training and support, professionalisation for data roles, governance structure for PSDI

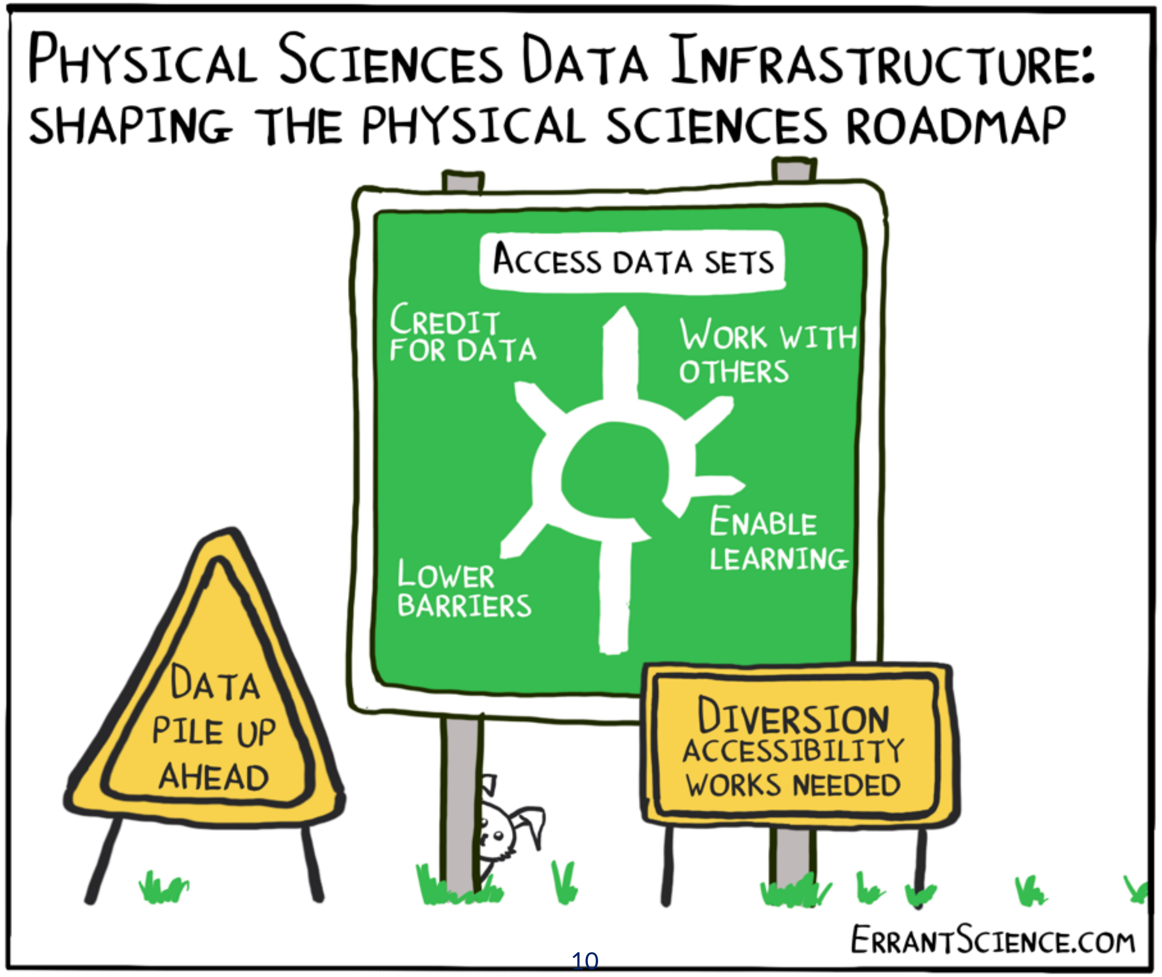
Best Use of Technology

2 Recommendations: services to connect existing provision (data and services), adopt existing technologies

Full recommendations at: https://www.psdi.ac.uk/the-pilot/recommendations
Outputs available via www.psdi.ac.uk and PSDI zenodo community

Current Work –
Platform, Pathfinders and Hub

- **Platform**
  - Requirements Analysis
  - Capacity Planning
  - System Architecture design
  - Component testing
  - Beginning Build
- **“Pathfinders”**
  - PF1: Experimental data capture
  - PF2: Process Recording
  - PF3: Building Data Collections
  - PF4: Process Orchestration
  - PF5: Data to Knowledge
- **Hub**: Communications, Governance, Planning,…


PHYSICAL SCIENCES DATA INFRASTRUCTURE:
SHAPING THE PHYSICAL SCIENCES ROADMAP
ACCESS DATA SETS
CREDIT FOR DATA
WORK WITH OTHERS
ENABLE LEARNING
LOWER BARRIERS
DATA PILE UP AHEAD
DIVERSION ACCESSIBILITY WORKS NEEDED
ERRANTSCIENCE.COM

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

PSDI Hub
Core Activities & Services

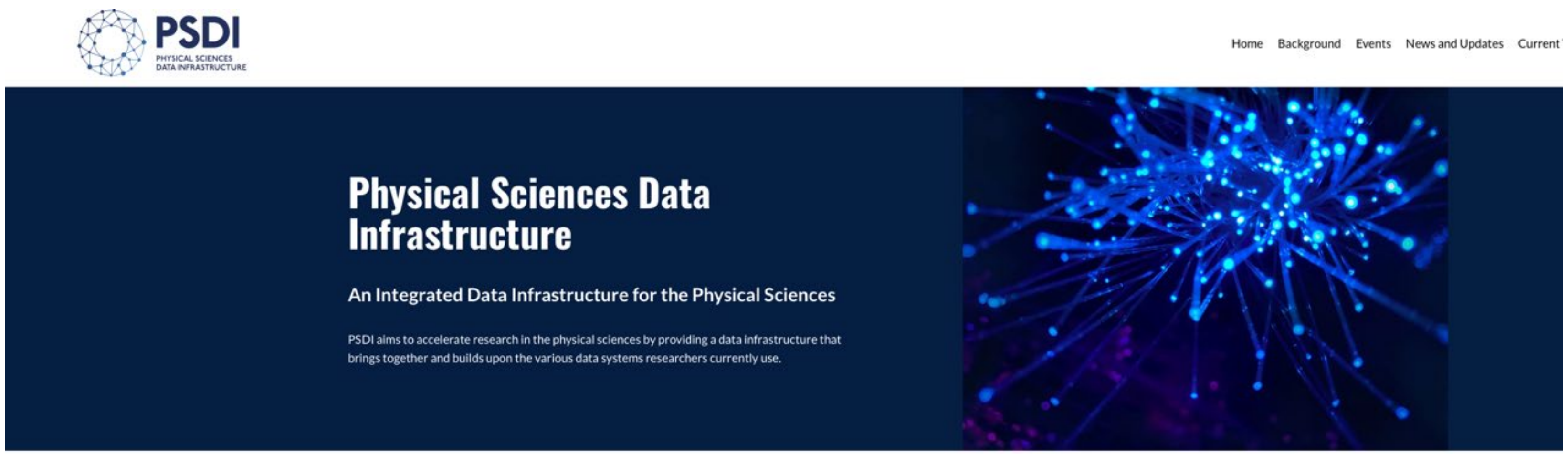
- Management, Governance &coordination
- Core data infrastructure components
- Communications and Engagement
- Training

International Collaboration

Research and data is not bounded by international borders!

Alignment with other ongoing and developing international projects

CODATA, RDA, WorldFAIR engagement (among others)



Physical Sciences Data Infrastructure

An Integrated Data Infrastructure for the Physical Sciences

PSDI aims to accelerate research in the physical sciences by providing a data infrastructure that brings together and builds upon the various data systems researchers currently use.

www.psdi.ac.uk @PSDI_UK @PSDI_UK PSDIUK

Any Questions?

Please do contact our researchers directly

They just love to talk about our work!