

PSDI: IT services in consideration

F. Barnsley, J. Bathe, A. Belozarov, V. Bunakov,†, A. Pawula Hewage, T.L. Underwood, P. Wright

STFC, Rutherford Appleton Laboratory, Harwell Campus, Didcot, OX11 0QX

† vasily.bunakov@stfc.ac.uk

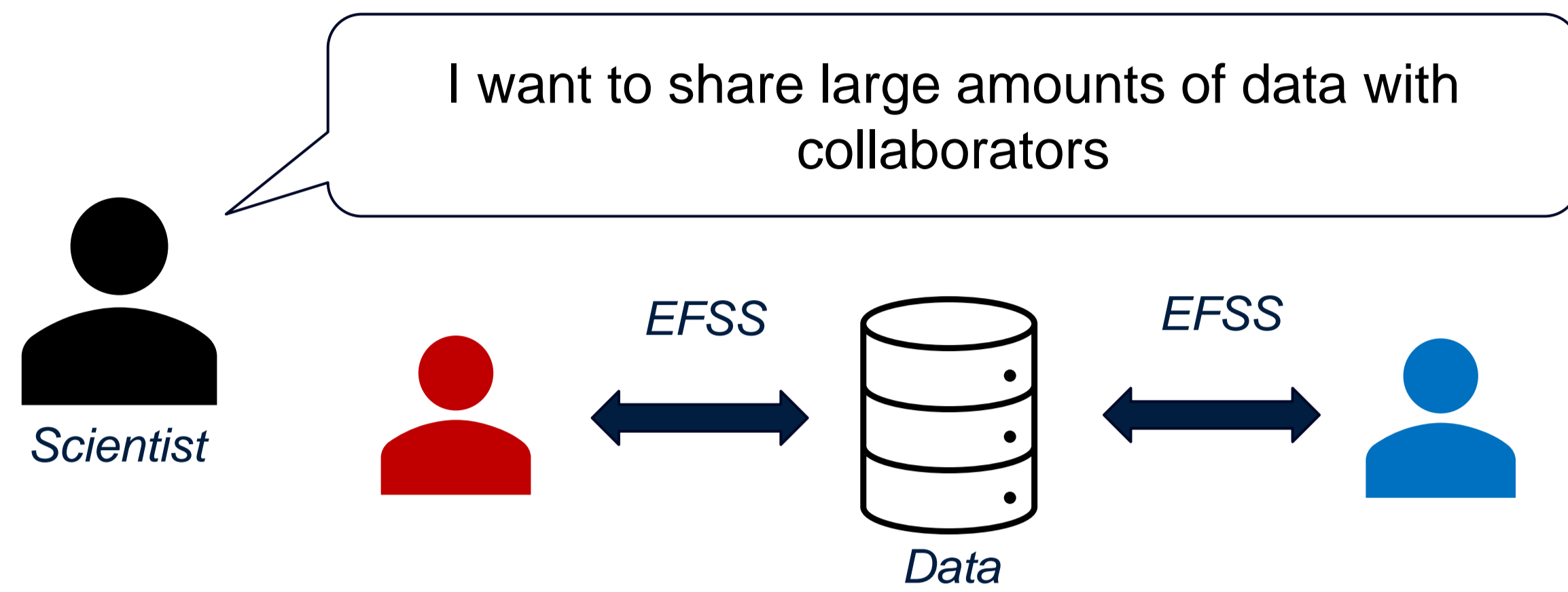


Visit our website

Abstract

The Physical Sciences Data Infrastructure (PSDI) will provide the scientific community with new capabilities for processing, storing, sharing and finding data. These capabilities will leverage existing STFC infrastructure, as well as many open-source technologies. Here we outline some of the prospective services PSDI could provide, and the technology which they might utilise.

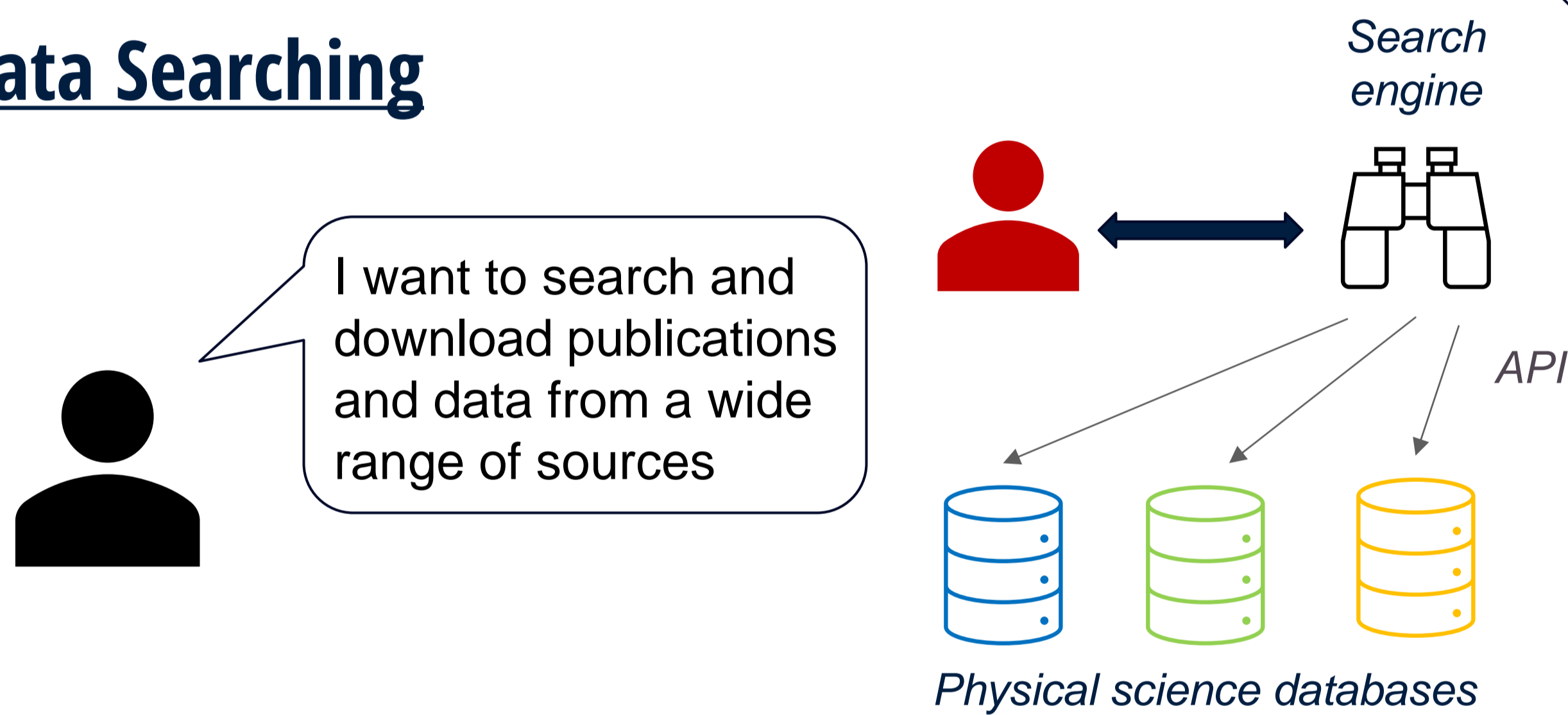
Data Sharing and Synchronisation



What could PSDI do?

- Provide *enterprise file synchronisation software* (EFSS) to facilitate:
 - Sharing large amounts of data between, e.g. HPC, Cloud, lab cluster
 - Synchronise data across multiple locations
 - We are exploring various open-source technologies, including Globus [1]
- Provide storage: we investigated various technologies, including object storage

Data Searching

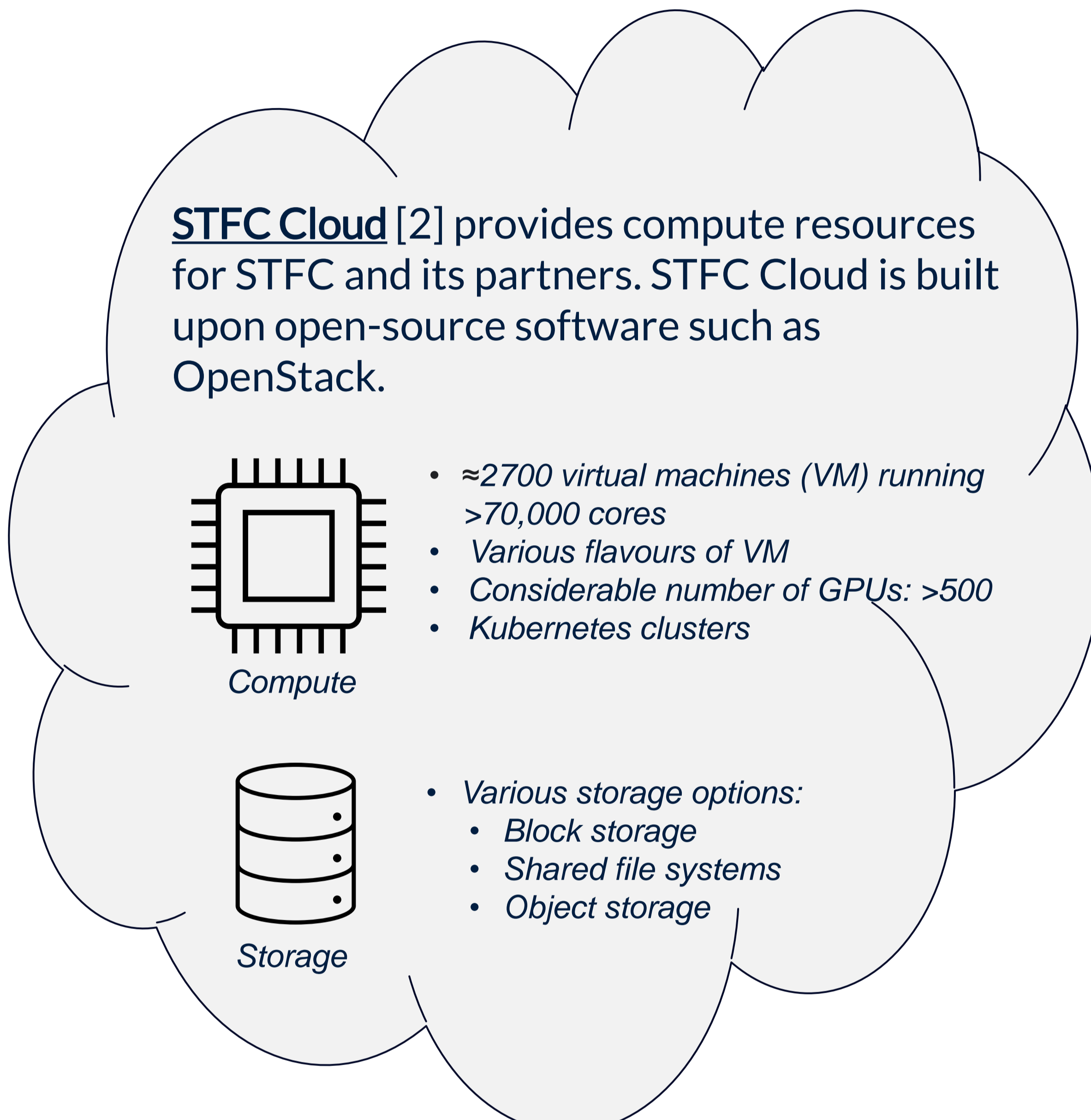


What could PSDI do?

- Provide a search engine indexing data from a wide range of databases in the physical sciences
 - Use open-source search engine: OpenSearch
- Provide an API for querying and accessing data
 - We've looked into OPTIMADE [3] for materials science databases

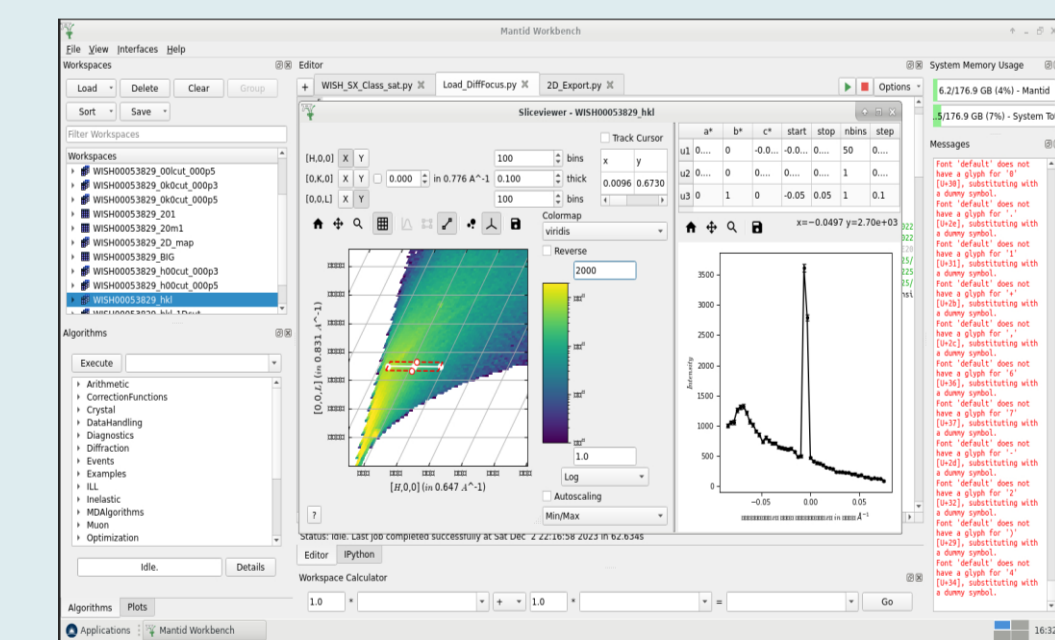
Storage and Compute underpinning PSDI

STFC's Scientific Computing Department (SCD) has a long history of providing computing infrastructure and services for users to support cutting-edge science. SCD provides infrastructure for national experimental facilities (ISIS, Diamond, CLF) and external partners such as CERN. PSDI will exploit and build upon existing infrastructure and expertise within SCD.



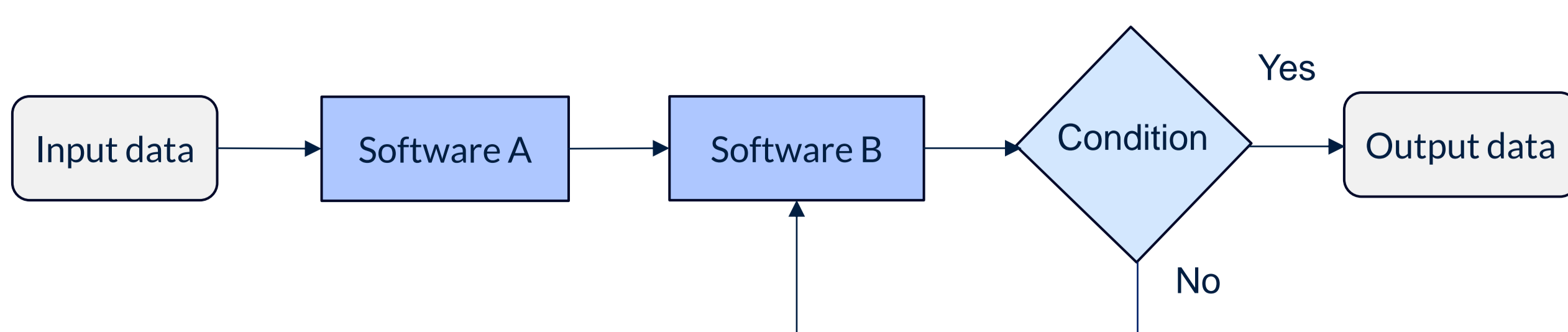
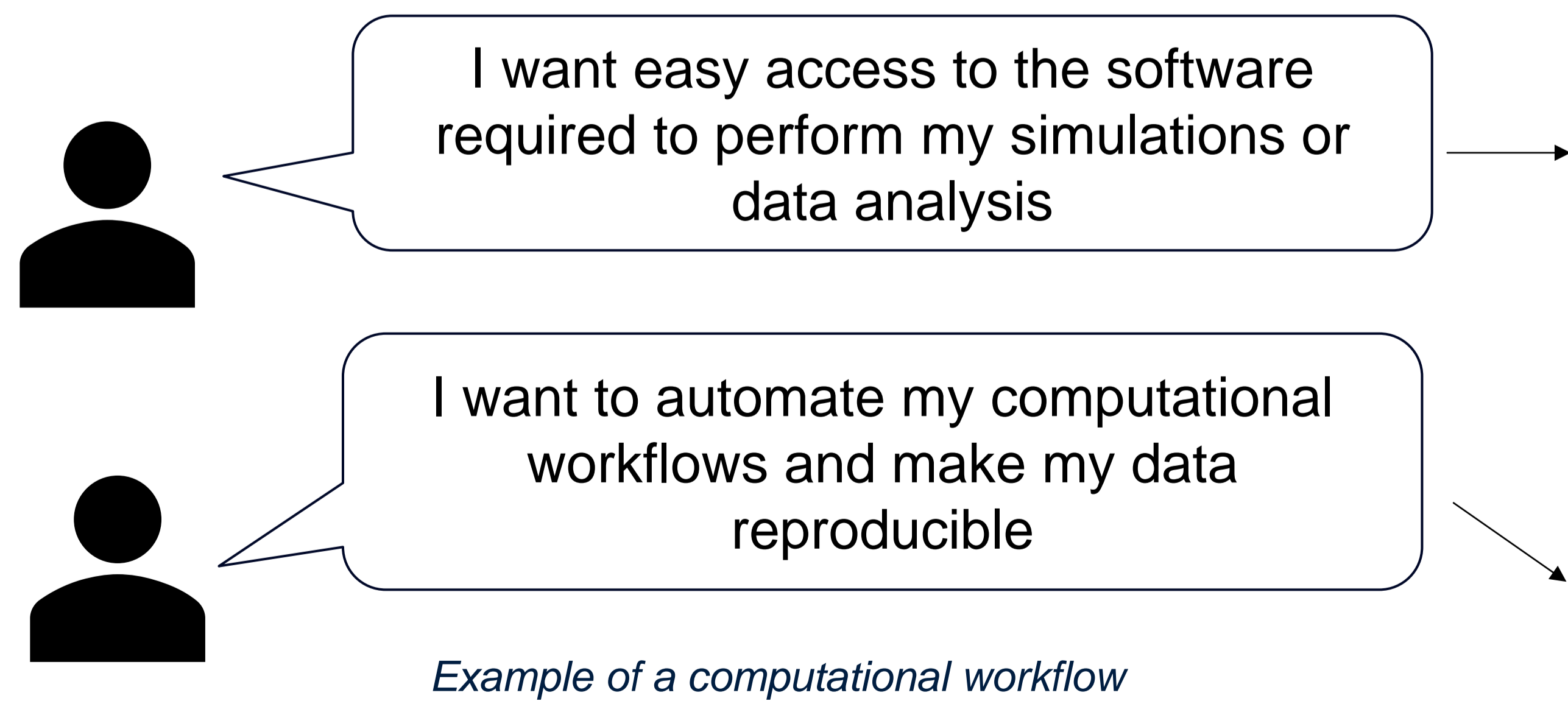
Existing capabilities: DAaaS platform

The DAaaS platform [4] enables STFC facility (ISIS) users to analyse their experimental data through a web browser. Users can choose from various software environments to perform their analysis.



Snapshot of DAaaS GUI

Data Processing and Reproducibility



What could PSDI do?

- Provide an app store where software could be easily obtained by users
 - This could be libraries of Docker images, VM images or configuration scripts
- Provide a platform where users can execute software, and store and share data
 - The DAaaS platform is a possible approach

What could PSDI do?

- Provide tools for orchestrating workflows and capturing data provenance
- We've investigated the workflow engines Galaxy [5] (for analysing experimental data) and AiiDA [6] (for executing simulations)

[1] <https://www.globus.org/>

[2] <https://www.iris.ac.uk/portfolio/stfc-cloud/>

[3] <https://www.optimade.org/>

[4] <https://www.scd.stfc.ac.uk/Pages/Dynamic-Infrastructure-Group.aspx>

[5] <https://galaxyproject.org/>

[6] <https://www.aiida.net/>