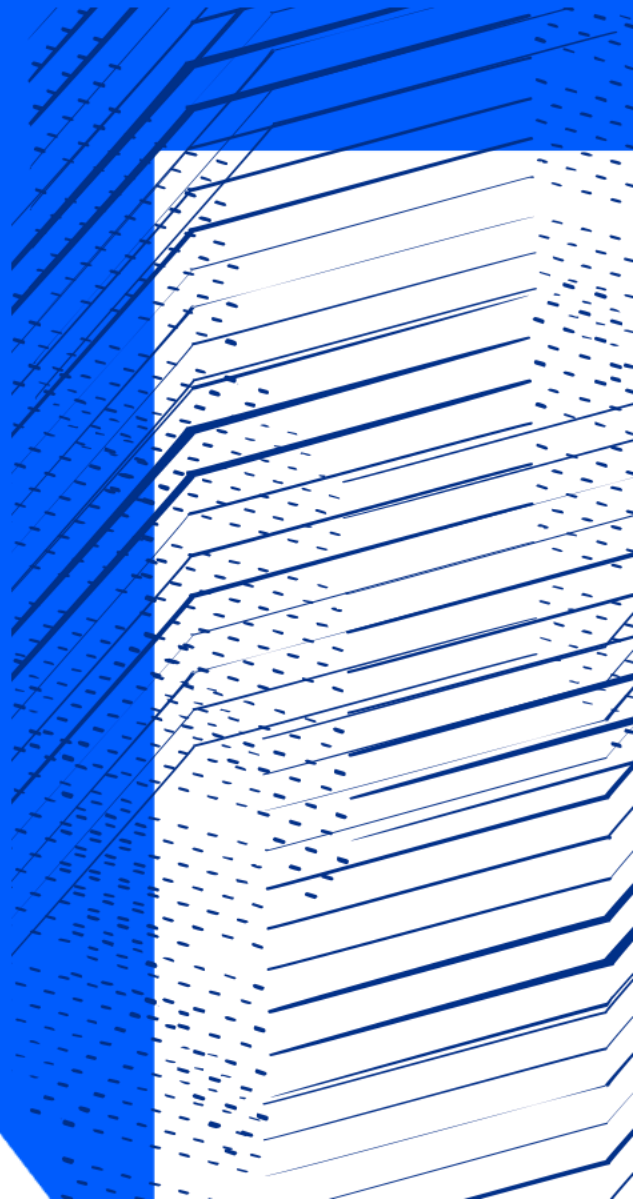
Science and Technology Facilities Council

The Long Tail of Science

Alastair Dewhurst

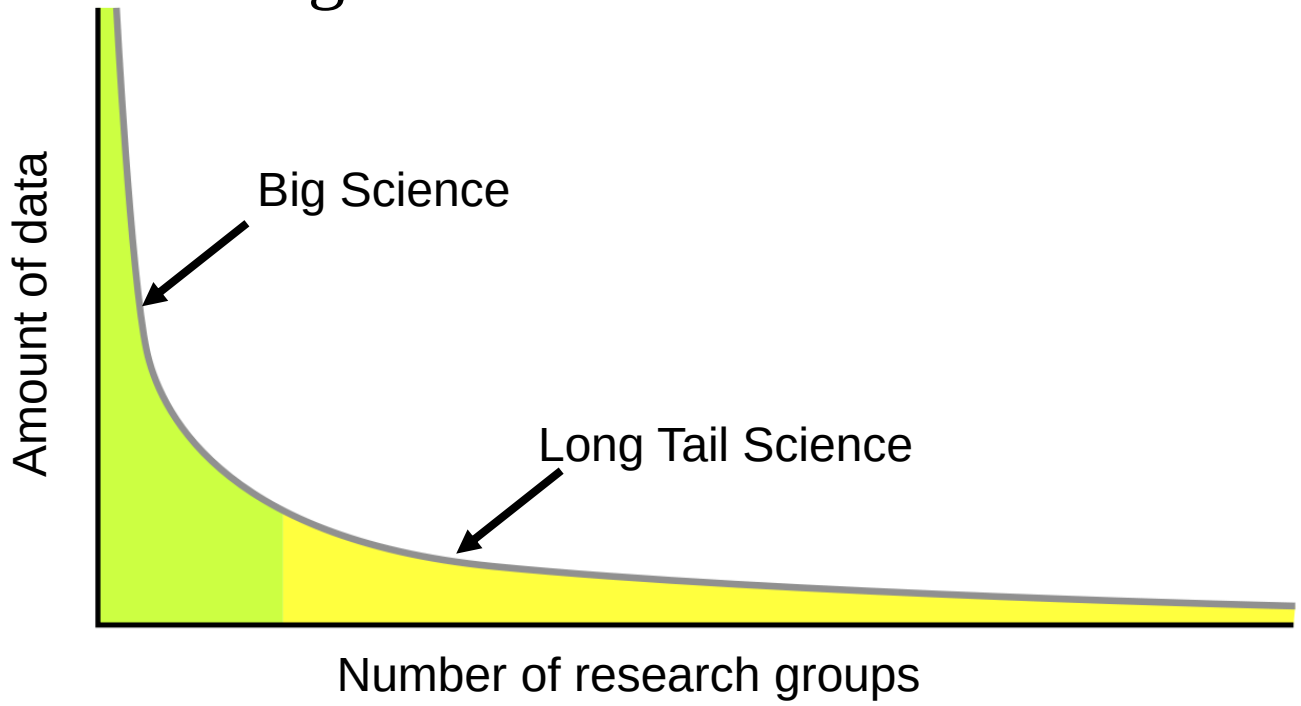


Introduction

- This session is aimed at new communities.
  - No prior knowledge is expected.
- Feel free to ask questions.
  - If you can't understand this session then we are failing in what we are trying to do to deliver to the community.
  - I will keep an eye on chat, if you raise your hand during a talk, I can interrupt at the end of the slide to ask it.
- Not many slides to allow plenty of time for questions/discussion.

What is the long tail of science?

The long tail of science refers to the large number of individual researchers and small laboratories who do not have access to dedicated computational resources and online services to manage and analyse large amounts of data.



What does this imply?

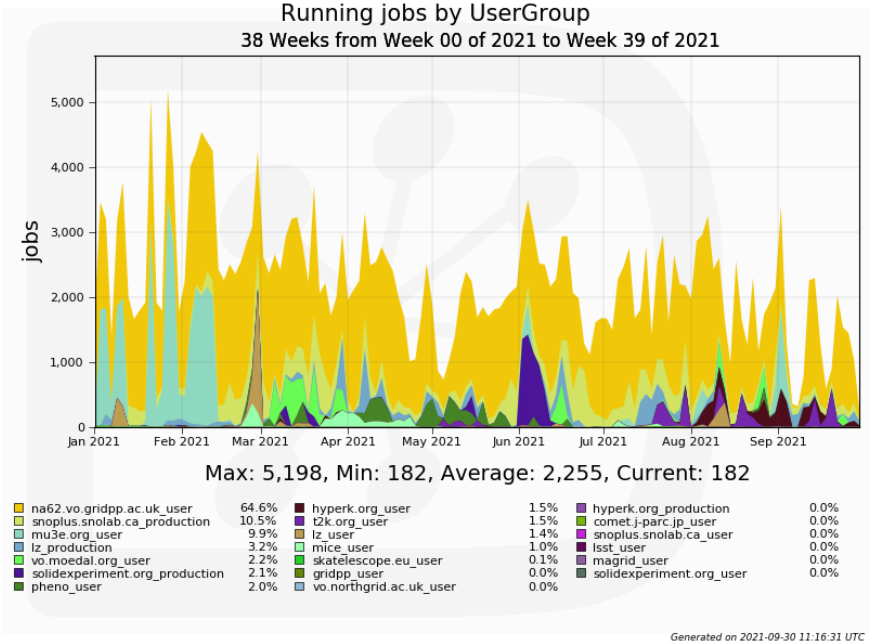
- Constraints are primarily on effort:
  - No effort to manage services
  - No expertise
  - Possible long periods of inactivity.

Requirements are a subset of those for a large VO:

- Data:
  - < 10 million files
  - < 10 PB data
- QoS:
  - Archival (Tape) - Their data is just as precious as the LHCs.
  - High Performance Disk - e.g. Parallel File System behind HPC
  - Data Sharing - Share results with others, often publicly.
- Ease of use:
  - Documentation
  - Clients easy to install / use
  - Integrated with their university / institutes login account.

Example: Multi-VO DIRAC

- In 2015 a Multi-VO instance of DIRAC was deployed at Imperial College, London.
  - Aim was to allow multiple smaller communities to use Grid Resources.
  - Many countries have setups to allow small experiment job submission.
- It has been successful!
  - 10+ communities run regularly.
  - 2000+ jobs running on average.
  - 6 million HS06 days used in 2021.
- Experiments often manage their data in other ways.
  - We hope Rucio can provide tools they need.



Example: ExaTEPP bid

- Recently I was part of a grant proposal for the lattice QCD community which included funding for Rucio. Extracts from the bid:
- Adapt the Rucio system for lattice simulation data. In the lattice field theory context, work is needed for Rucio to handle lattice-specific metadata (ideally arbitrary and extensible).
- The aim is to demonstrate data management and curation for lattice field theory data satisfying FAIR principles, supporting modern authentication and access, with data searchable and accessible according to physics application.
- A variety of storage types will be investigated, including the currently used Ceph-based systems, but others like Storj decentralised storage with S3 access.

Is our strategy correct?

- In the rest of the session you will hear from experts in data management about how they believe we will provide data management for the long tail of science.
- Do you think we are correct?
- Have we missed anything?
- What can be done better?

Science and
Technology
Facilities Council

Questions?