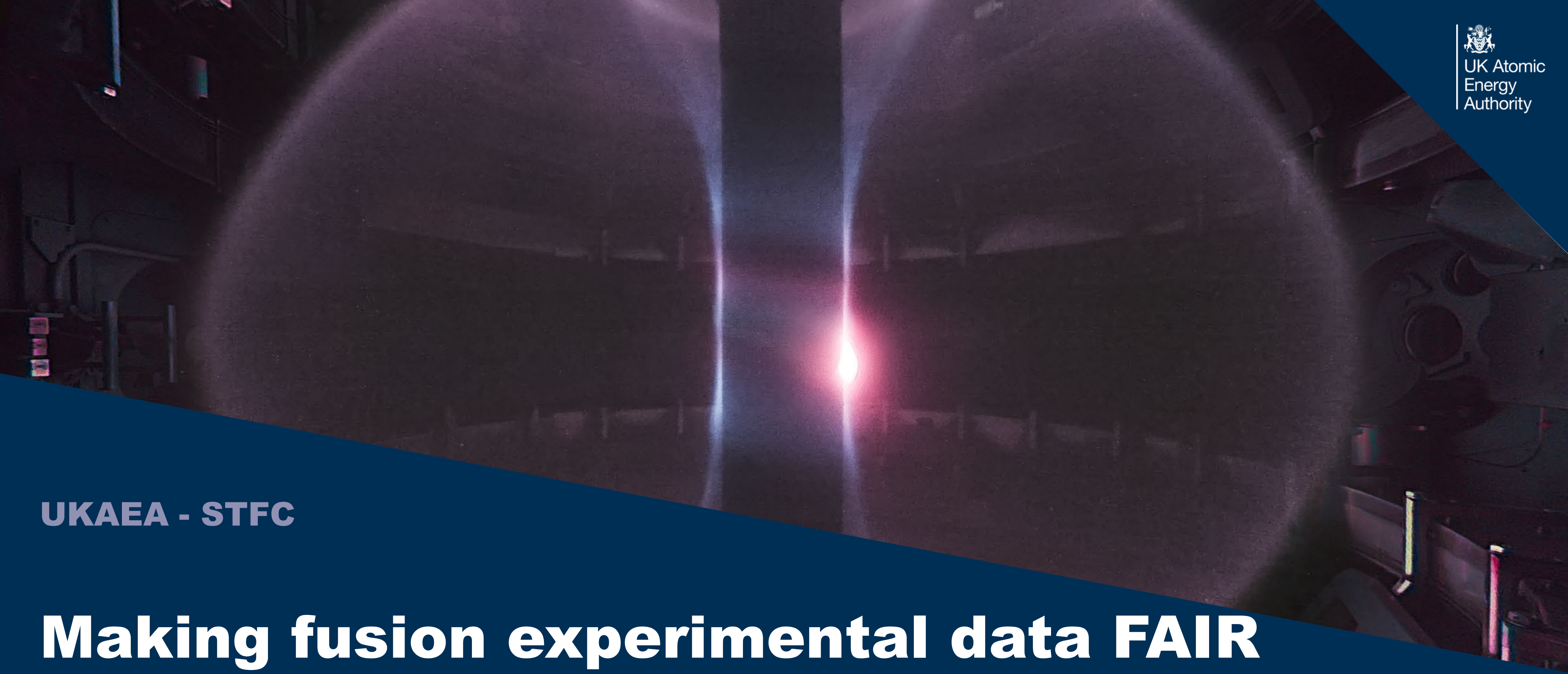


UK Atomic Energy Authority

UKAEA - STFC

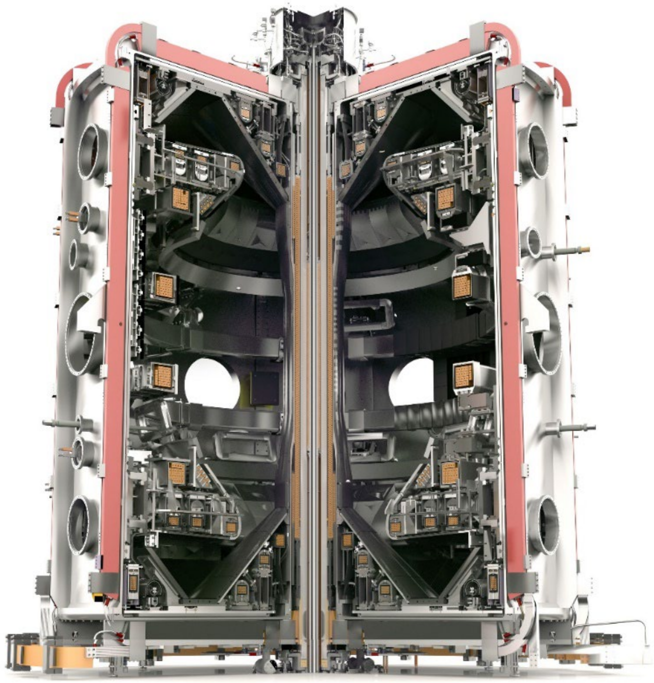
Making fusion experimental data FAIR

Samuel Jackson, Saiful Khan, Jeyan Thiyagalingam – STFC

Rob Akers, Shaun de Witt, Nathan Cummings, James Hodson, Edward Harrington – UKAEA

Special thanks to James Buchanan (UKAEA), Nicola Amorisco (STFC) and Vignesh Gopakumar (UKAEA) for also contributing slides, and to Jimmy Measures (UKAEA) for much-appreciated support

MAST/MAST-Upgrade



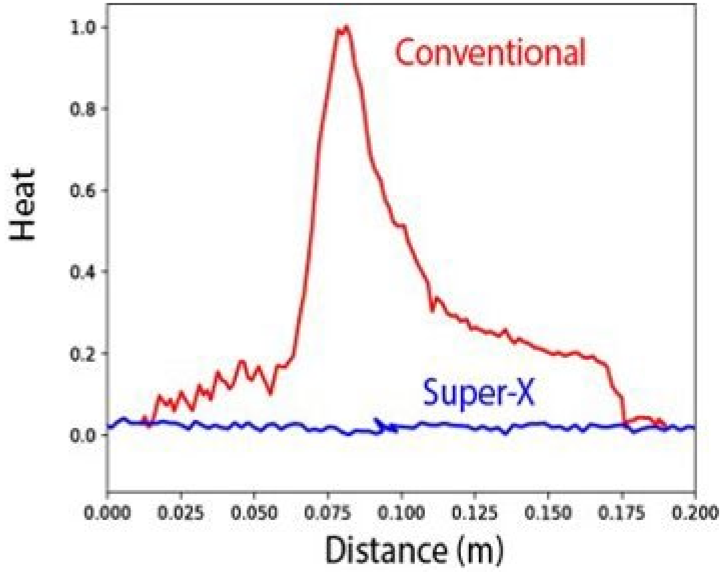
MAST-U is a *spherical tokamak,* a more compact design than JET which is less mature but has the potential to be more efficient.

MAST-U has a **super-X** divertor. A larger open structure which creates a longer path length for the plasma to reach the wall.

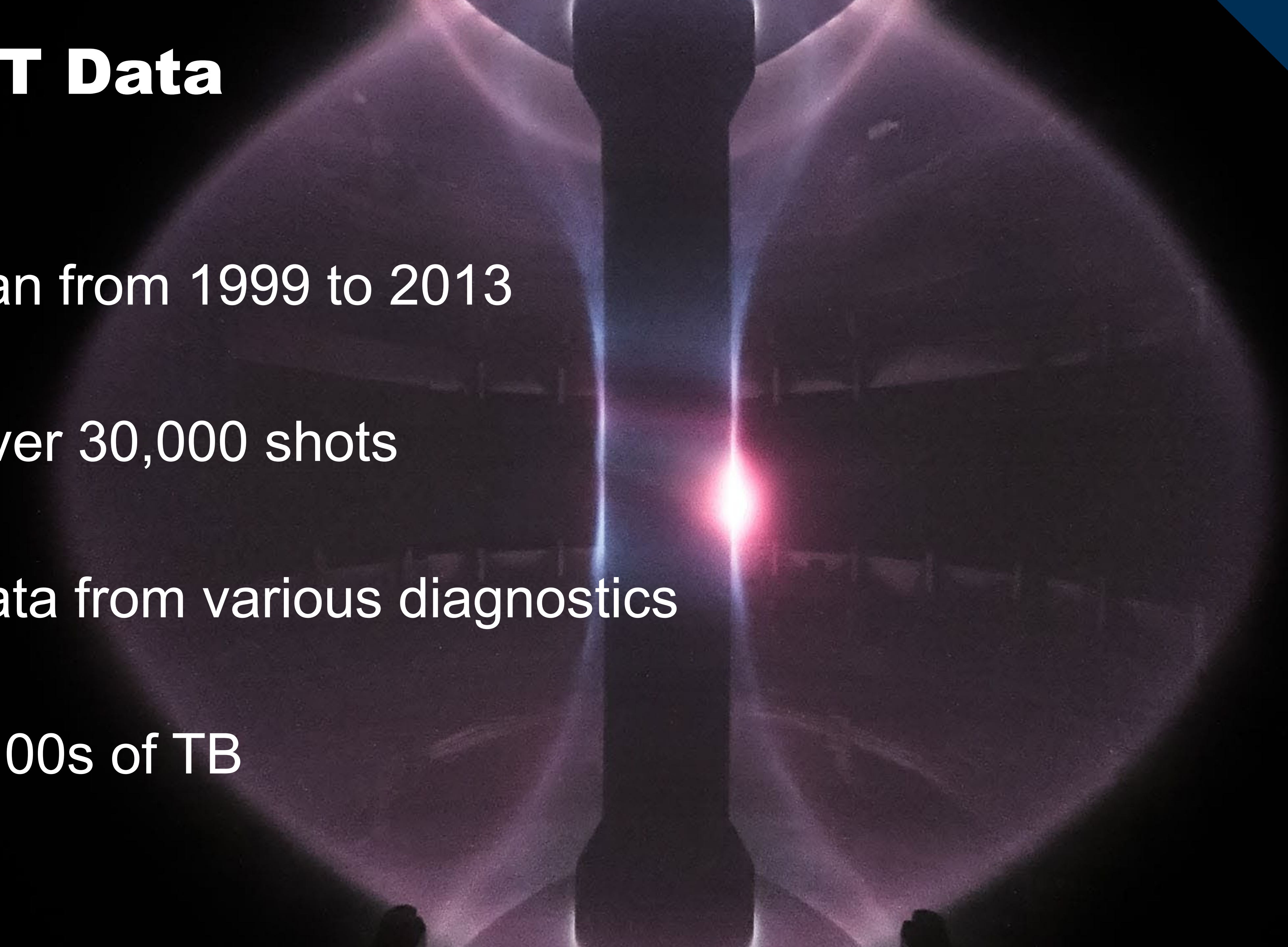
This allows the plasma to cool and be spread over a larger area leading to a lower heat flux on the wall.

Conventional

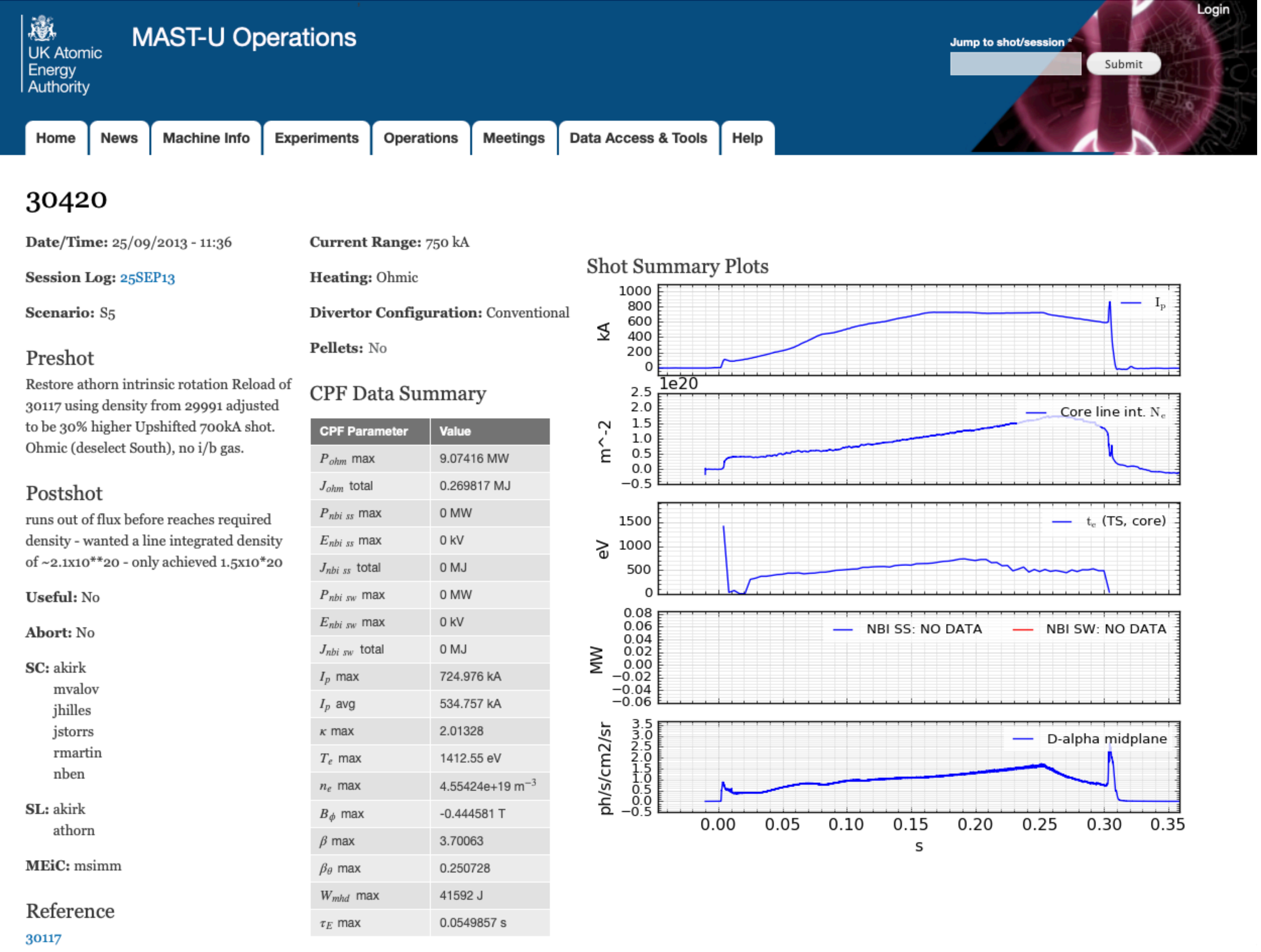
Super-X



MAST Data



- Ran from 1999 to 2013
- Over 30,000 shots
- Data from various diagnostics
- ~100s of TB



30420

Date/Time: 25/09/2013 - 11:36

Session Log: 25SEP13

Scenario: S5

Preshot

Restore athorn intrinsic rotation Reload of 30117 using density from 29991 adjusted to be 30% higher Upshifted 700kA shot. Ohmic (deselect South), no i/b gas.

Postshot

runs out of flux before reaches required density - wanted a line integrated density of ~2.1x10**20 - only achieved 1.5x10*20

Useful: No

Abort: No

SC: akirk
mvalov
jhilles
jstorrs
rmartin
nben

SL: akirk
athorn

MEiC: msimm

Reference

30117

Current Range: 750 kA

Heating: Ohmic

Divertor Configuration: Conventional

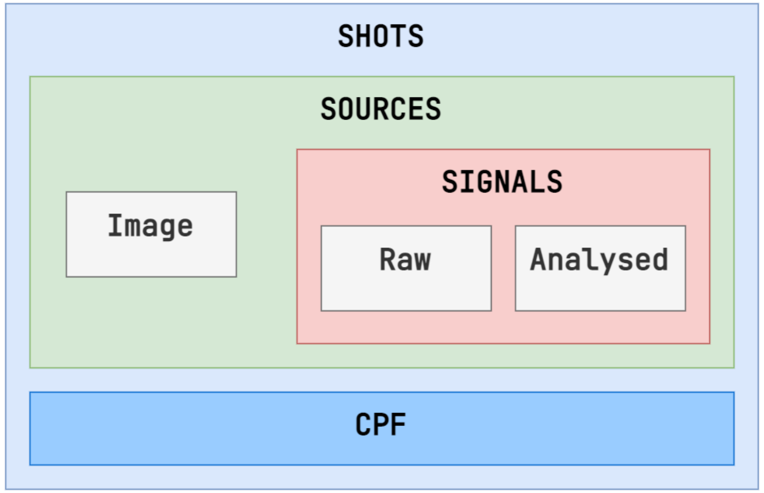
Pellets: No

CPF Data Summary

CPF Parameter	Value
P_{ohm} max	9.07416 MW
J_{ohm} total	0.269817 MJ
$P_{nbi\ ss}$ max	0 MW
$E_{nbi\ ss}$ max	0 kV
$J_{nbi\ ss}$ total	0 MJ
$P_{nbi\ sw}$ max	0 MW
$E_{nbi\ sw}$ max	0 kV
$J_{nbi\ sw}$ total	0 MJ
I_p max	724.976 kA
I_p avg	534.757 kA
κ max	2.01328
T_e max	1412.55 eV
n_e max	4.55424e+19 m^{-3}
B_ϕ max	-0.444581 T
β max	3.70063
β_θ max	0.250728
W_{mhd} max	41592 J
τ_E max	0.0549857 s

Shot Summary Plots

- Different types of data:
  - Time-series data
  - Video data (2D+t)
- Metadata
  - Descriptions
  - Units
  - Authors
  - Dates & times
  - Etc..


SHOTS

SOURCES

Image

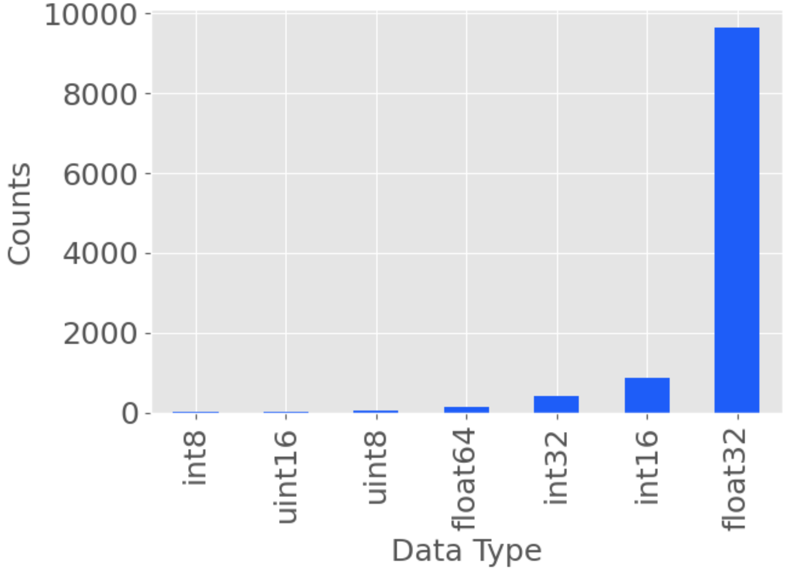
SIGNALS

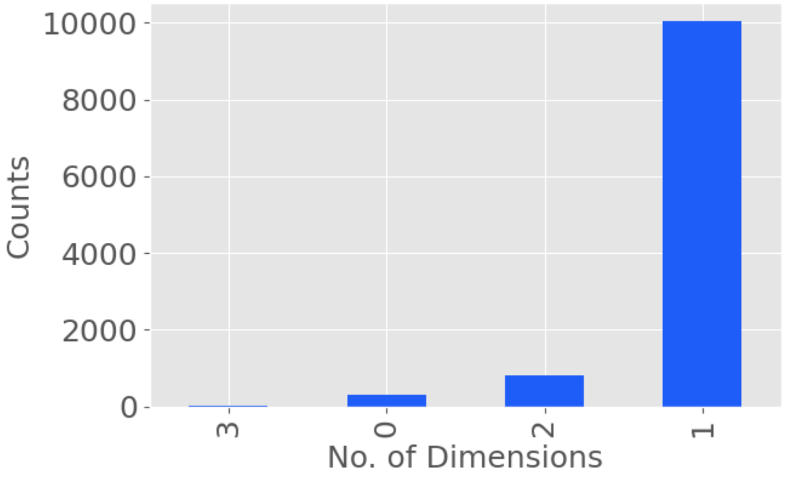
Raw

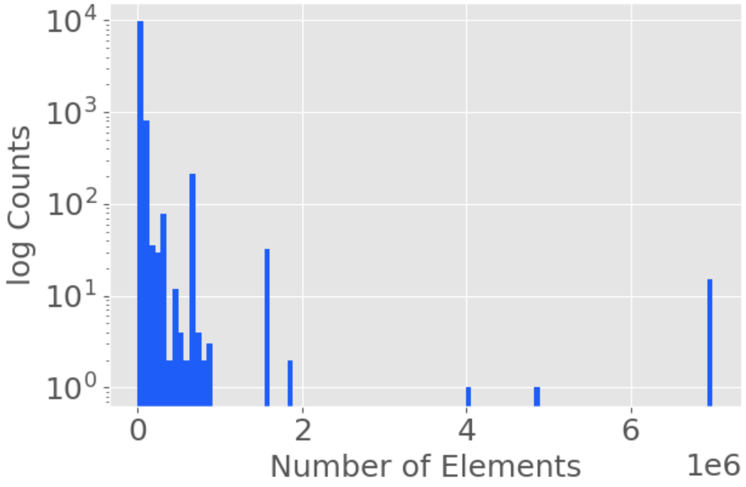
Analysed

CPF

MAST Data
A single shot







~11000

Signals

~7 Billion

Data Points

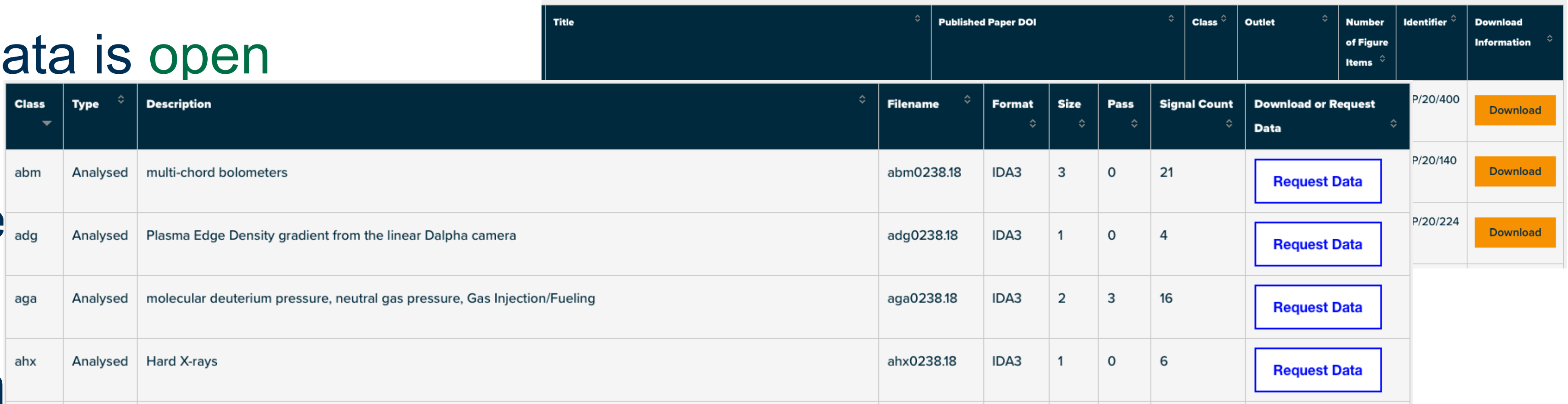
~7 GB

(Uncompressed)

MAST Data Access

Web

- Published data is open
- Not designe
- Non-publish
- Not automatable
- No licence

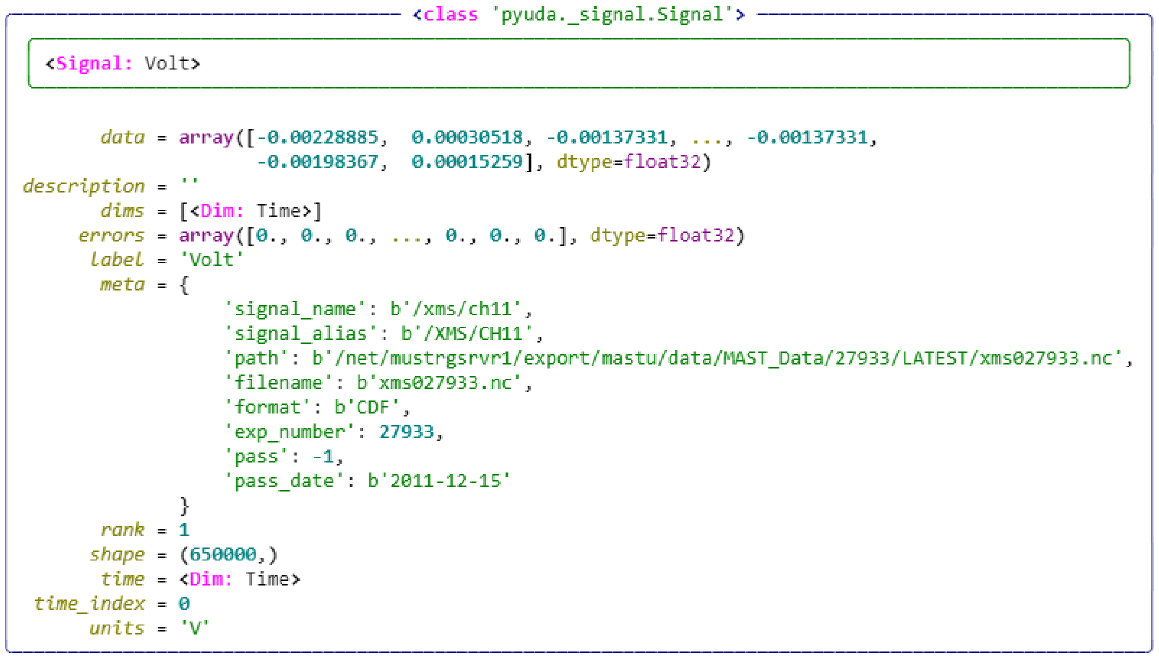


Title	Published Paper DOI	Class	Outlet	Number of Figure Items	Identifier	Download Information
					P/20/400	Download
					P/20/140	Download
					P/20/224	Download

Class	Type	Description	Filename	Format	Size	Pass	Signal Count	Download or Request Data
abm	Analysed	multi-chord bolometers	abm0238.18	IDA3	3	0	21	Request Data
adg	Analysed	Plasma Edge Density gradient from the linear Dalpha camera	adg0238.18	IDA3	1	0	4	Request Data
aga	Analysed	molecular deuterium pressure, neutral gas pressure, Gas Injection/Fueling	aga0238.18	IDA3	2	3	16	Request Data
ahx	Analysed	Hard X-rays	ahx0238.18	IDA3	1	0	6	Request Data

MAST Data Access
UDA

- Interfaces for c, c++, FORTRAN, IDL, Java and Python
- External access is difficult
- Data is accessed ‘vertically’
- Does not expose *all* metadata
- Performs data corrections ‘on-the-fly’
- Not optimized for AI/ML



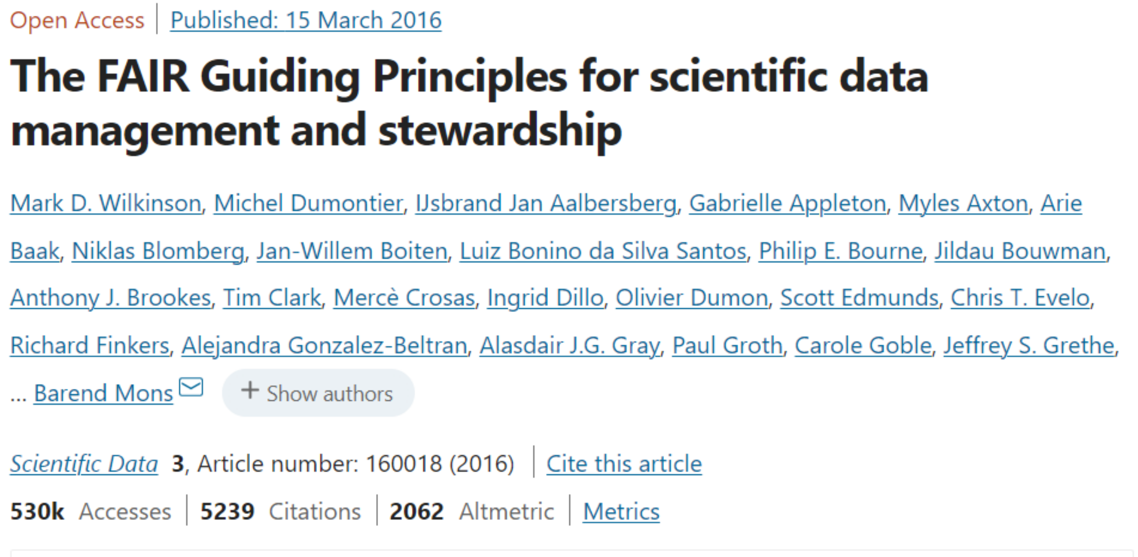
```
<class 'pyuda._signal.Signal'>

<Signal: Volt>


       data = array([-0.00228885,  0.00030518, -0.00137331, ..., -0.00137331,
                     -0.00198367,  0.00015259], dtype=float32)
description = ''
       dims = [<Dim: Time>]
     errors = array([0., 0., 0., ..., 0., 0., 0.], dtype=float32)
      label = 'Volt'
       meta = {
                  'signal_name': b'/xms/ch11',
                  'signal_alias': b'/XMS/CH11',
                  'path': b'/net/mustrgsrvr1/export/mastu/data/MAST_Data/27933/LATEST/xms027933.nc',
                  'filename': b'xms027933.nc',
                  'format': b'CDF',
                  'exp_number': 27933,
                  'pass': -1,
                  'pass_date': b'2011-12-15'
              }
       rank = 1
      shape = (650000,)
       time = <Dim: Time>
 time_index = 0
      units = 'V'
```

FAIR

- Findable
- Accessible
- Interoperable
- Reusable



Open Access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

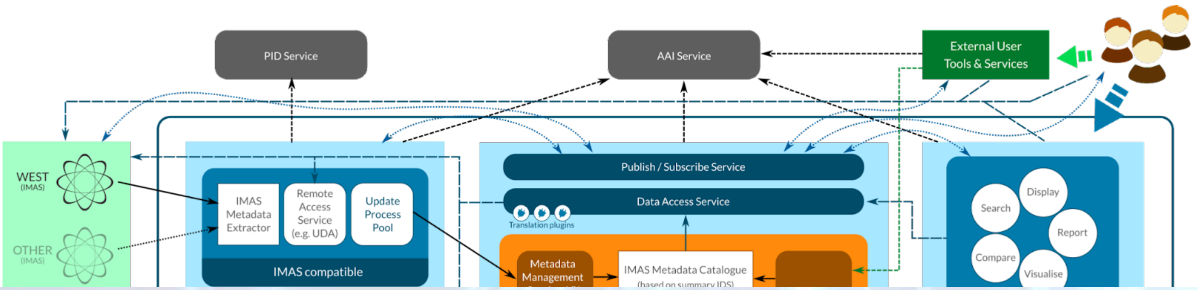
Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, … Barend Mons ✉ + Show authors

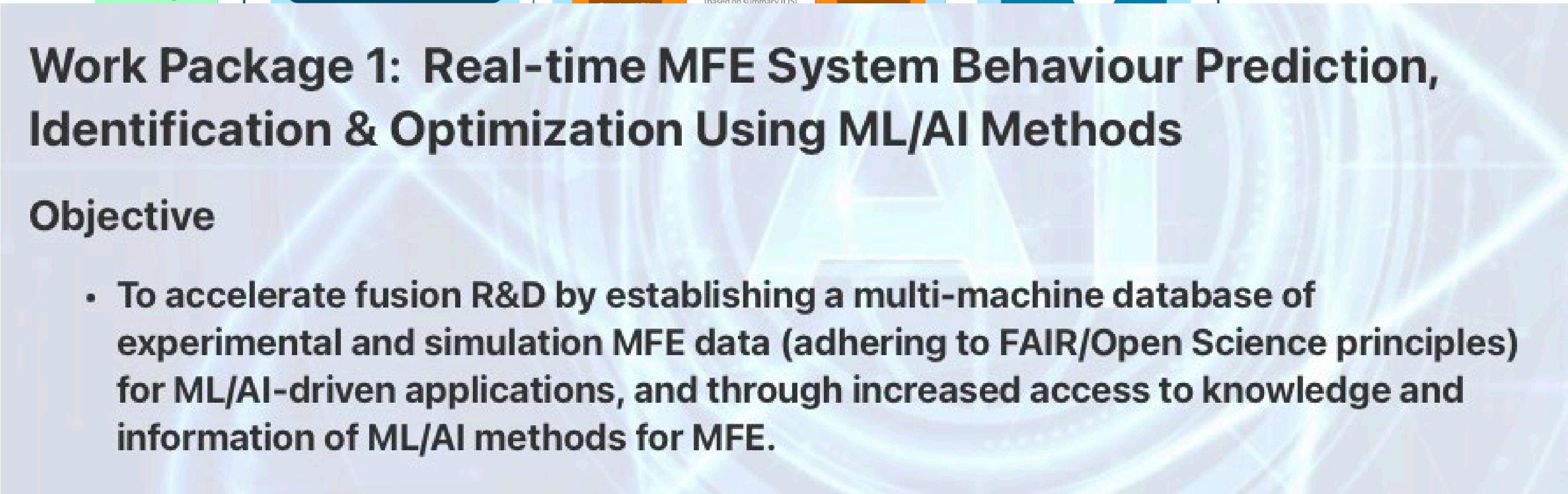
Scientific Data **3**, Article number: 160018 (2016) | Cite this article

530k Accesses | **5239** Citations | **2062** Altmetric | Metrics

Recent FAIR initiatives in fusion

- FAIR4fusion
- EUROfusion DMP
- IAEA - CRP





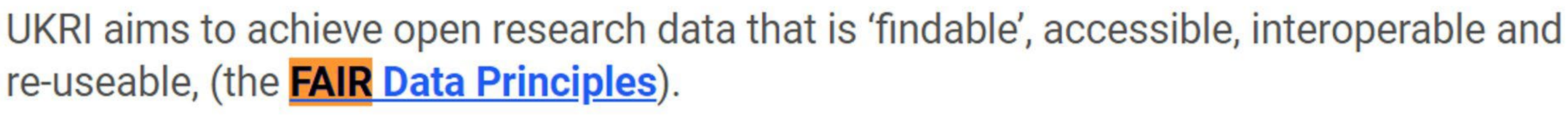
Work Package 1: Real-time MFE System Behaviour Prediction, Identification & Optimization Using ML/AI Methods

Objective

- To accelerate fusion R&D by establishing a multi-machine database of experimental and simulation MFE data (adhering to FAIR/Open Science principles) for ML/AI-driven applications, and through increased access to knowledge and information of ML/AI methods for MFE.

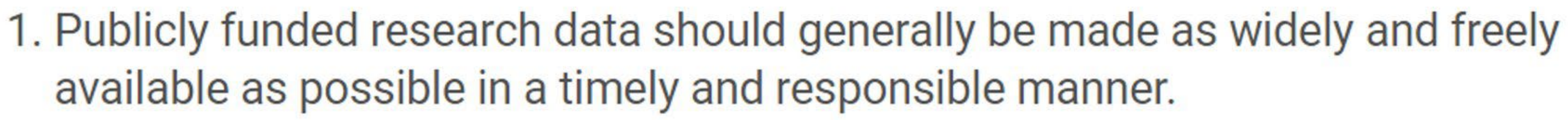
Why should MAST data be FAIR?

- UK Research and Innovation (UKRI) open access policy



UKRI aims to achieve open research data that is 'findable', accessible, interoperable and re-useable, (the FAIR Data Principles).

- Engineering and Physical Science Research Council (EPSRC) research data policy



1. Publicly funded research data should generally be made as widely and freely available as possible in a timely and responsible manner.

Sources: https://www.ukri.org/what-we-offer/supporting-healthy-research-and-innovation-culture/open-research/
https://www.ukri.org/about-us/epsrc/our-policies-and-standards/policy-framework-on-research-data/principles/

Experimental data and FAIR – why should you care?

- Data processing, analysis and sharing is easier when you're FAIR
- Findable – easily locate the data needed for simulations etc…
- Accessible – fosters collaboration
- Interoperable – easily work with different workflows & analysis tools
- Reusable – descriptive metadata provides context

Considerations for ‘open’

- Data could be misinterpreted
  - Mitigated by providing rich metadata
- Could be misused
  - Data licencing/disclaimers
- Researcher’s work could be ‘scooped’
  - Research embargo and authenticated access to time-series data
- More data can be validated and opened up in time

Towards FAIR for MAST data

Project Goals

- Data must be easily findable through the metadata
- Data must be in exposed in an interoperable format
- Prioritise performance optimisation for data-intensive workflows (e.g. AI/ML)
- Minimise loading and transferring data
- Support analysis codes/libraries and ML/AI frameworks
- Support larger-than-memory & parallel computation
- Be publicly accessible

System Components

PostgreSQL

- High performance look-up
- Postgres is horizontally scalable
- Easily convertible to an in-memory Data Frame
- Tabular data only

Zarr

- Efficient binary storage
- Supports multi-dimensional array data
- Supports larger than memory tools
  - Dask, Spark etc…
- Multiple language support
  - C++, Fortran, Python, Matlab, R, etc…
- Multiple backends
- Multiple compression options
- Not easily searchable

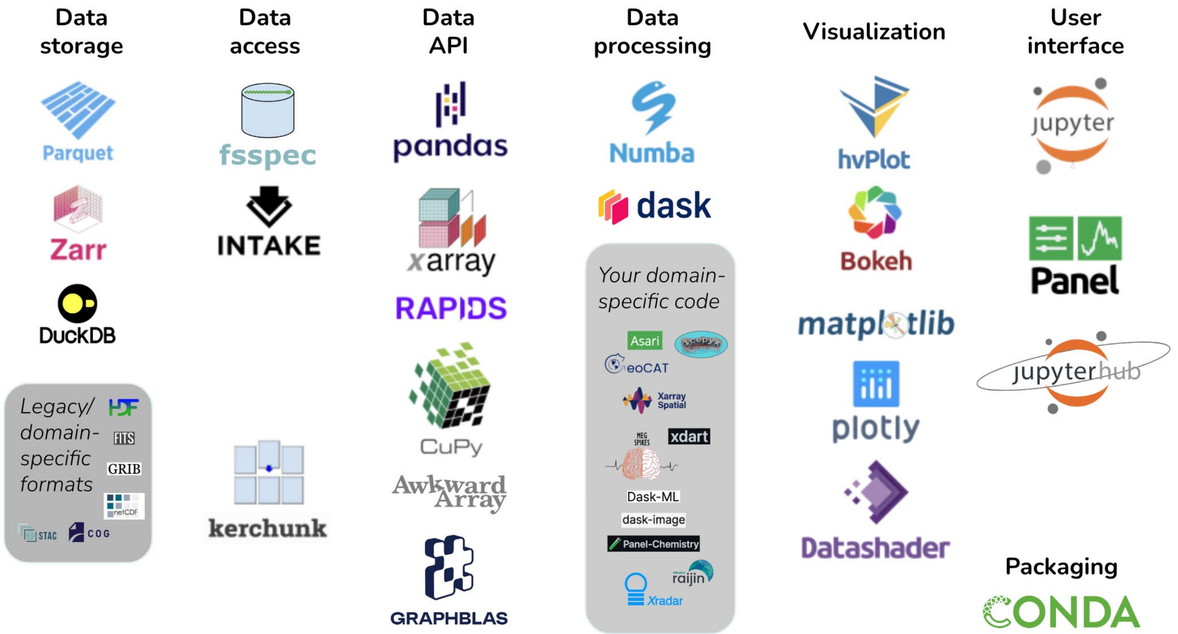
A2. Metadata are accessible, even when the data are no longer available

Separate meta-database can support an embargo period and authentication.

FAIR Checklist

- Findable
  - (Meta)data are assigned a globally unique and persistent identifier ✓
  - Data are described with rich metadata ✓
  - Metadata clearly and explicitly include the identifier of the data they describe ✓
  - (Meta)data are registered or indexed in a searchable resource ✓
- Accessible
  - (Meta)data are retrievable by their identifier using a standardised communications protocol ✓
    - The protocol is open, free, and universally implementable ✓
    - The protocol allows for an authentication and authorisation procedure, where necessary ✓
  - Metadata are accessible, even when the data are no longer available ✓
- Interoperable
  - (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. ✓
  - (Meta)data use vocabularies that follow FAIR principles ✓
  - (Meta)data include qualified references to other (meta)data ✓
- Reusable
  - (Meta)data are richly described with a plurality of accurate and relevant attributes ✓
    - (Meta)data are released with a clear and accessible data usage license ✓
    - (Meta)data are associated with detailed provenance ✓
    - (Meta)data meet domain-relevant community standards ●

Pandata
Scalable open-source analysis stack



- Domain independent: Maintained, used, and tested by people from many different backgrounds
- Efficient: Run at machine-code speeds using vectorized data and compiled code
- Scalable: Run on anything from a single-core laptop to a thousand-node cluster
- Cloud friendly: Fully usable for local or remote compute using data on any file storage system
- Multi-architecture: Run on Mac/Windows/Linux systems, using CPUs or GPUs
- Scriptable: Fully support batch mode for parameter searches and unattended operation
- Compositional: Select which tools you need and put them together to solve your problem
- Visualizable: Support rendering even the largest datasets without conversion or approximation
- Interactive: Support fully interactive exploration, not just rendering static images or text files
- Shareable: Deployable as web apps for use by anyone anywhere
- OSS: Free, open, and ready for research or commercial use, without restrictive licensing

Source: https://conference.scipy.org/proceedings/scipy2023/pdfs/james_bednar.pdf

Summary

- Some MAST data is already open
- Not optimised for data-intensive applications
- We are developing an AI/ML friendly database/service for MAST data
- By adhering to FAIR principles, we can maximise the scientific utility of our data

Observations/discussion points regarding nuclear data

- Licence
  - Data
  - Software

- Metadata separate from data
  - Perhaps less of a concern
  - Self-describing formats

- Versioning
  - Data
  - Software

- Format/Data Model
  - GNDS?

- Integration with data processing pipelines
  - PanData stack

- Continuous Delivery for databases

UK Atomic
Energy
Authority

Thank for your time and attention

Questions and comments are most welcome