

**UKAEA - STFC**

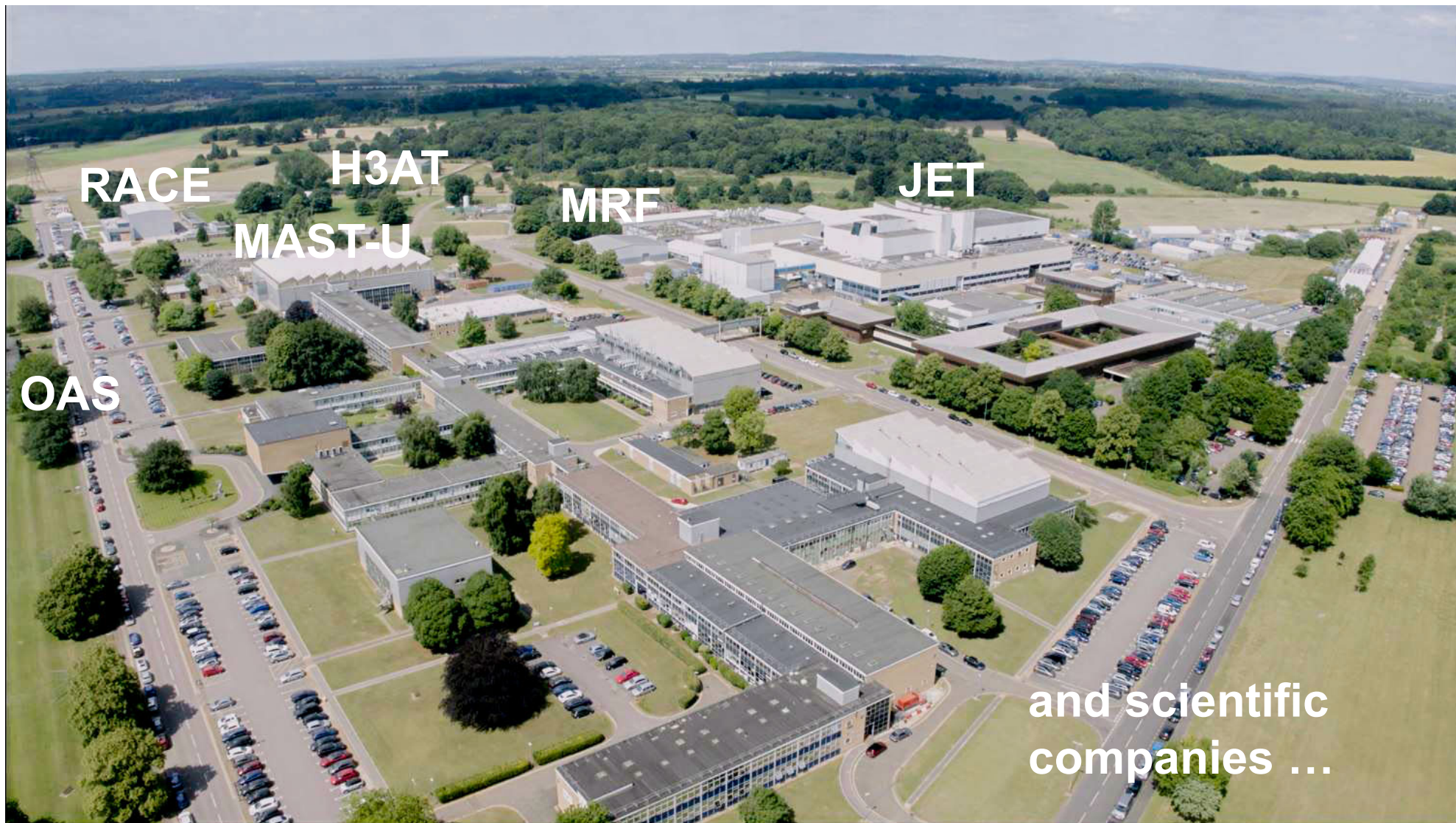
# Database development for magnetic fusion applications

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

# UKAEA - Culham Science Centre



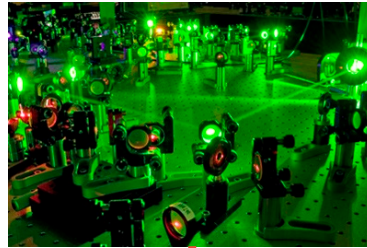
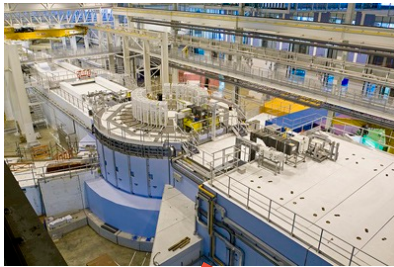
# STFC and SciML



- Based at the Rutherford Appleton Laboratory (RAL) at Harwell Campus
- Harwell Campus is home to many STFC facilities including:
  - Diamond Light Source
  - Central Laser Facility
  - ISIS Neutron and Muon Source
  - NERC and JASMIN Centres for environmental data analysis

► Overall remit is on AI for Science to support large-scale experimental facilities like Rutherford Appleton Laboratory

ISIS Neutron and Muon Source



Central Laser Facility



SciML



# Focus Areas of SciML

Research to  
Advance AI

Use AI for Science to  
improve AI itself

Use of AI  
in Science

- Use AI to understand experimental results (from facilities)

*AI for Science*

Smart  
Facilities

- Smart facilities can improve science – high quality data, faster results etc.
- It is all about embedding AI at the heart of operations

“

...the mission of the group is to explore how the use of Machine Learning and other AI technologies can help scientists analyse the vast amounts of experimental data now being routinely generated by the large experimental facilities at the Rutherford Appleton Laboratory (RAL) and STFC's Harwell campus.

”

# Questions

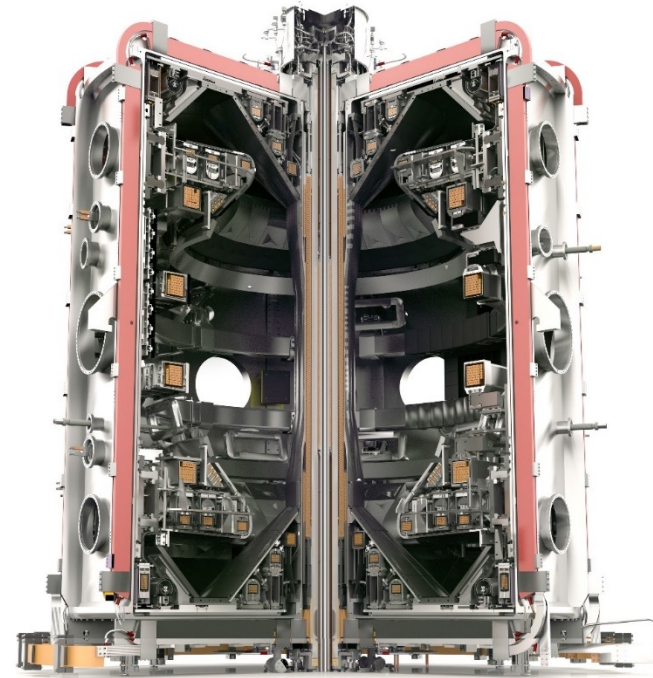
- What is the first thing you need to do machine learning?

# DATA

- Where do data come from?
  - The internet
  - Simulations
  - Physical experiments

# 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.

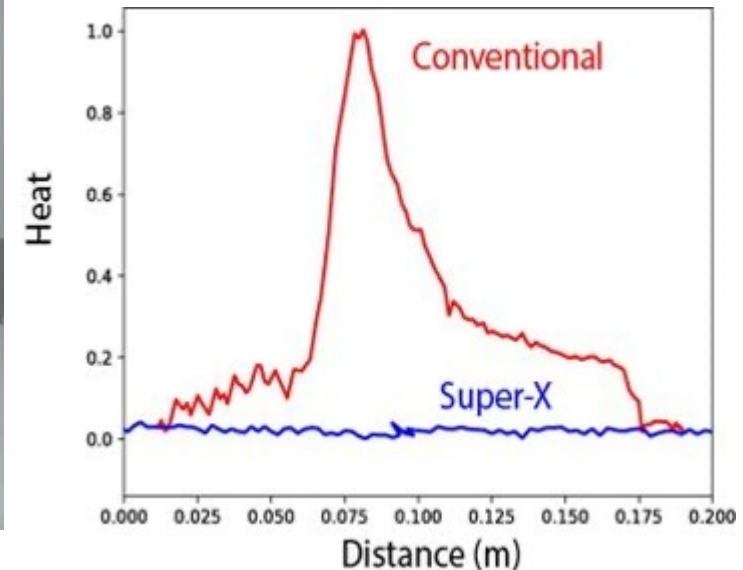
This allows the plasma to cool and be spread over a larger area leading to a lower heat flux on the wall.



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  $\sim 2.1 \times 10^{20}$  - only achieved  $1.5 \times 10^{20}$

Useful: No

Abort: No

SC: akirk  
mvalov  
jhilles  
jstorr  
rmartin  
nben

SL: akirk  
athorn

MEiC: msimm

## Reference

30117

Current Range: 750 kA

Heating: Ohmic

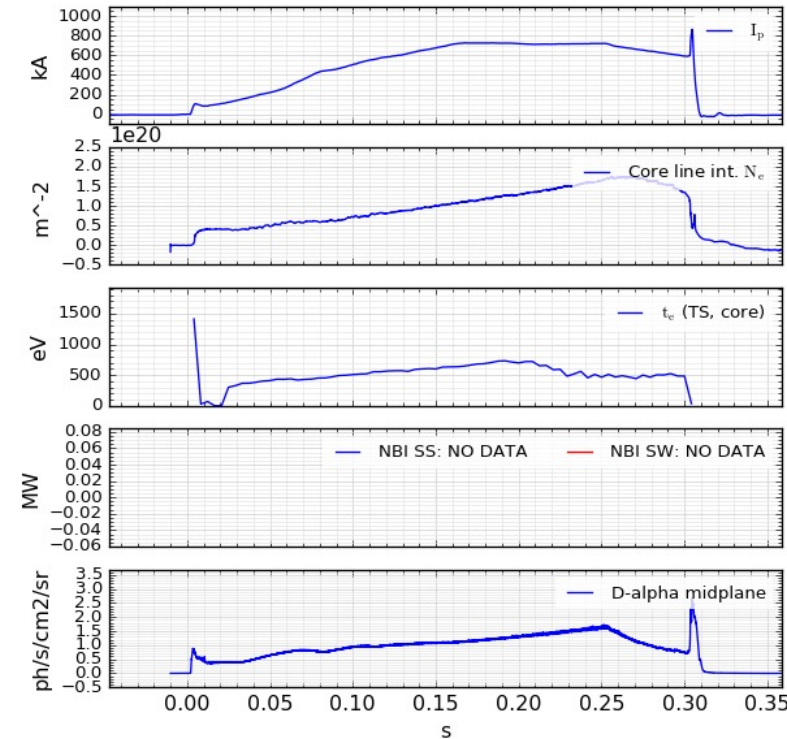
Divertor Configuration: Conventional

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
$\kappa$ max	2.01328
$T_e$ max	1412.55 eV
$n_e$ max	$4.55424 \times 10^{19} \text{ m}^{-3}$
$B_\phi$ max	-0.444581 T
$\beta$ max	3.70063
$\beta_\theta$ max	0.250728
$W_{mhd}$ max	41592 J
$\tau_E$ max	0.0549857 s

## Shot Summary Plots

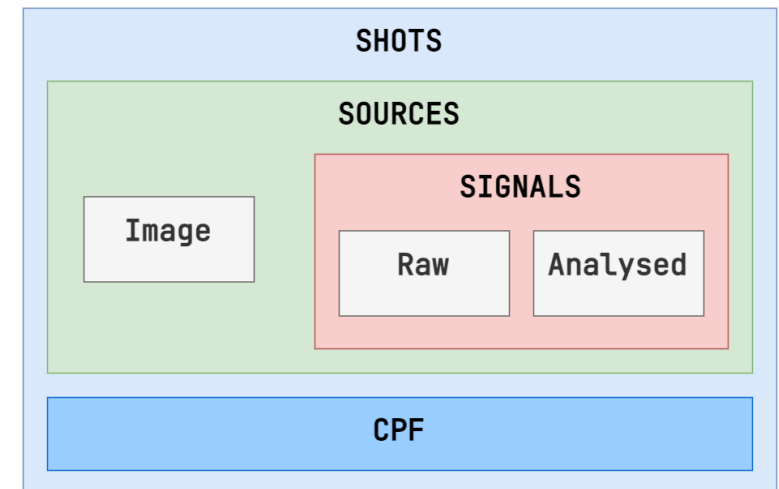


## Different types of data:

- Time-series data
- Video data (2D+t)

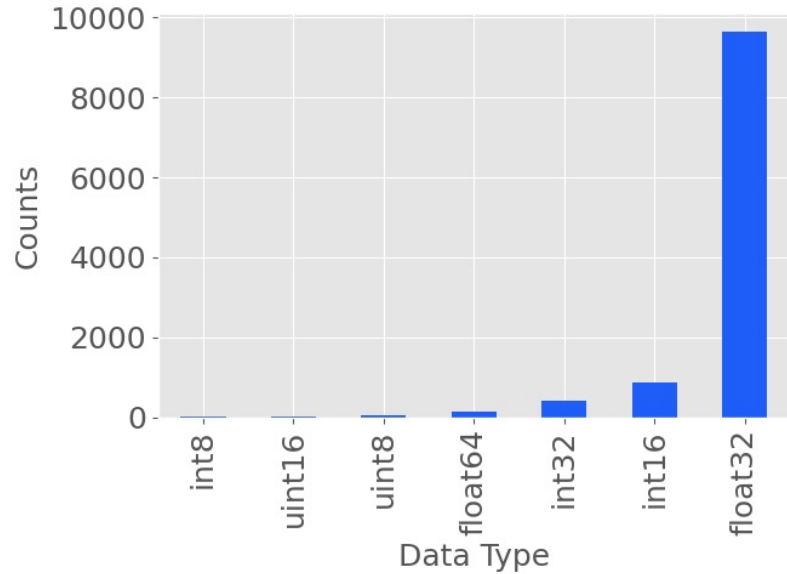
## Metadata

- Descriptions
- Units
- Authors
- Dates & times
- Etc..

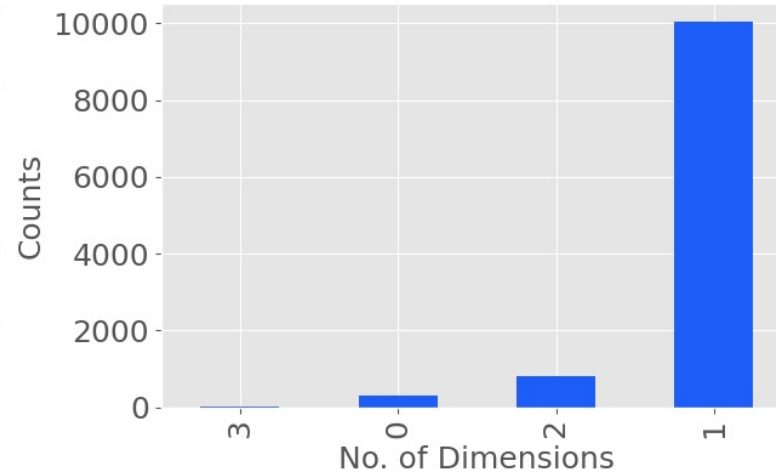


# MAST Data

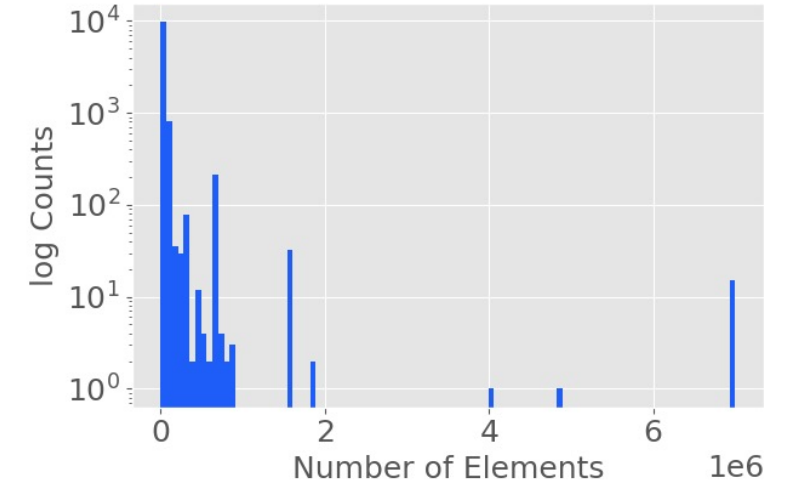
## A single shot



**~11000**  
Signals



**~7 Billion**  
Data Points



**~7 GB**  
(Uncompressed)

# MAST Data Access Web

- Published data
- Not designed for
- Non-published data by request only
- Not automatable

Class		Type	Description	Filename	Format	Size	Pass	Signal Count	Download or Request Data	Identifier	Download Information
abm	Analysed		multi-chord bolometers	abm0238.18	IDA3	3	0	21	<a href="#">Request Data</a>	P/20/400	<a href="#">Download</a>
adg	Analysed		Plasma Edge Density gradient from the linear Dalphi camera	adg0238.18	IDA3	1	0	4	<a href="#">Request Data</a>	P/20/140	<a href="#">Download</a>
aga	Analysed		molecular deuterium pressure, neutral gas pressure, Gas Injection/Fueling	aga0238.18	IDA3	2	3	16	<a href="#">Request Data</a>	P/20/224	<a href="#">Download</a>
ahx	Analysed		Hard X-rays	ahx0238.18	IDA3	1	0	6	<a href="#">Request Data</a>		

# 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/mustrgrsvr1/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'

```

# Exercise


- Either on your own, or with the person next to you:
  - Find a dataset online (scientific data ideally)
  - Access the data
  - Plot something that shows that you know what the data mean
- Answer 3 questions about them:
  - Where are the data from?
    - Who/how/when/(why?)
  - How do you know what the data are/represent?
  - What are you allowed/not allowed to do with the data?

15 minutes (ish)... **GO!**

- 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](#)

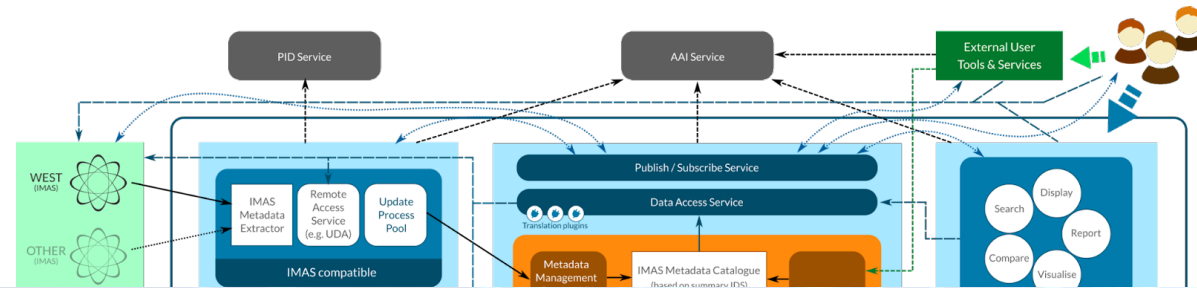
[Scientific Data](#) **3**, Article number: 160018 (2016) | [Cite this article](#)

**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.

# Experimental data and FAIR – why should you care?

- AI/ML is easier when you're FAIR
- Findable – easily locate the data needed for training sets etc...
- Accessible – fosters collaboration
- Interoperable – easily work with different workflows & analysis tools
- Reusable – descriptive metadata provides context

# Towards FAIR for MAST data Metadata

- Metadata stored in relational database
  - Made **open** (F)
  - Queryable by SQL, ORM (e.g., SQLAlchemy) etc... (A, I)
  - Develop REST API
  - Graphical frontend
- Relevant metadata fields labelled according to **FAIR** vocabularies, e.g., Dublin Core (I)
- Contains URIs and paths for time-series data (F)
- **Richly populated** with plasma parameters/tagged with event flags (mode transitions/disruptions/etc) that we can iteratively populate. (F)

# Towards FAIR for MAST data

## Time-series/signal data

- Stored on HPC system at Cambridge (CSD3)
  - free and open format (A, I)
- Access control
  - UKAEA staff/approved collaborators to access all data
  - public access limited to **validated** data (A)
- Eventually...
  - Map existing data to a common standard (I)

# Considerations for 'open'

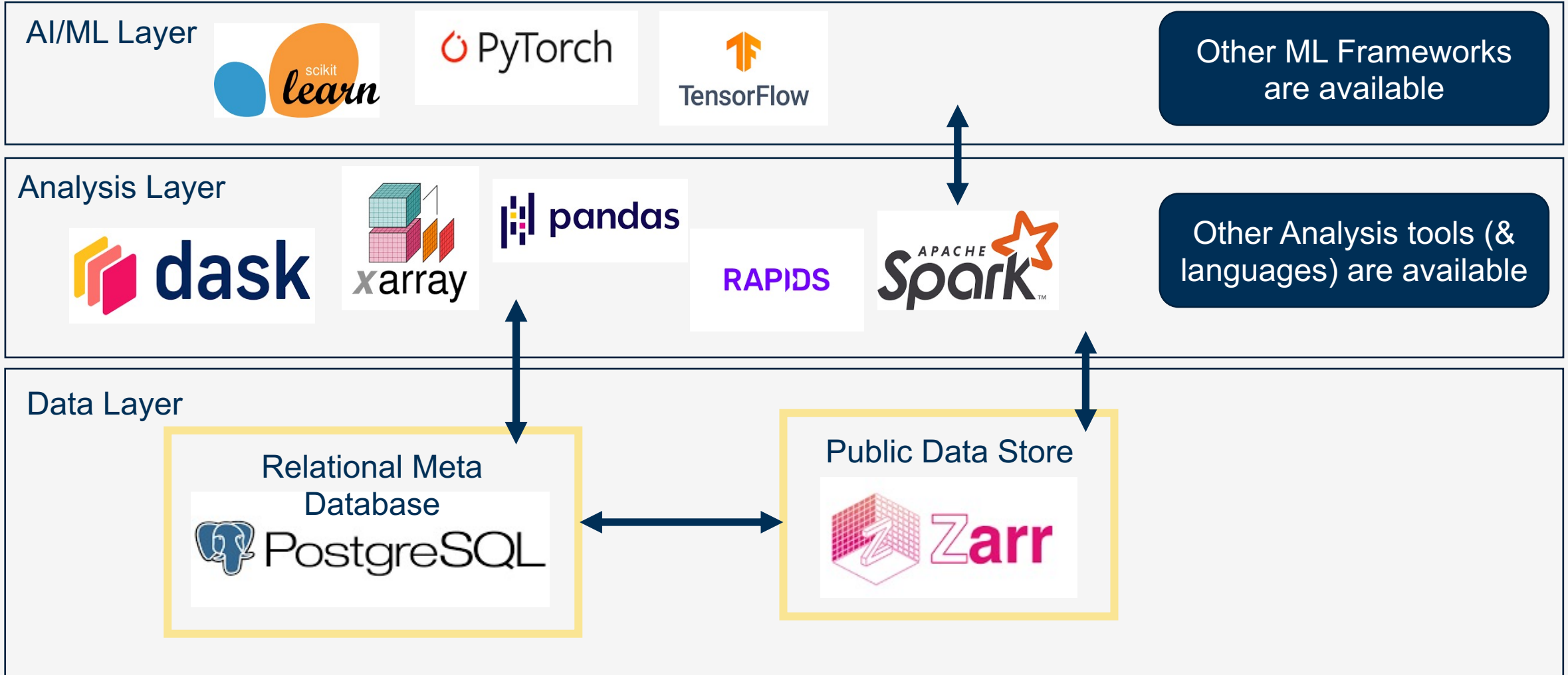
- Data could be **misinterpreted**
  - Mitigated by providing **rich metadata**
- Could be **misused**
  - Data licencing/disclaimers
- Researcher's work could be '**scooped**'
  - Research embargo and authenticated access to time-series data
- More data can be validated and **opened up** in time

# Towards FAIR for MAST data

## Project Goals

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

# FAIR MAST data for AI Architecture



# System Components



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



- 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.



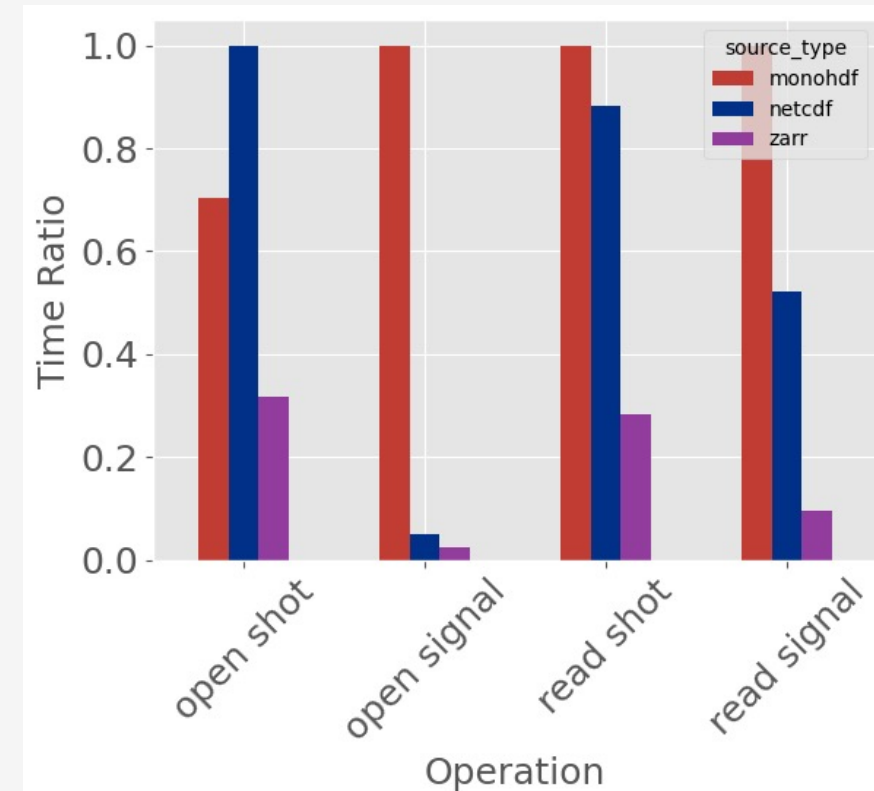
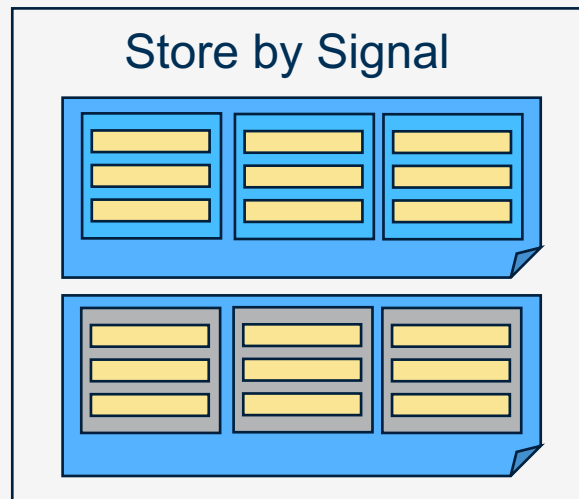
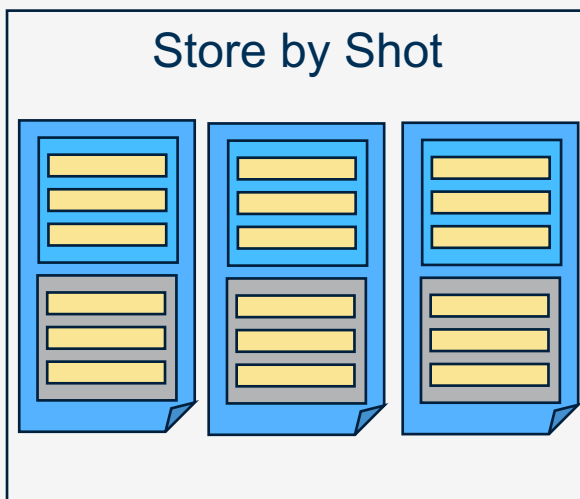
# Files Vs. ArrayDBMS

- **File-based storage**
  - Store data in a binary format in a file.
  - Use parallel & distributed tools to process
  - User defines how to query & process the data
- **ArrayDBMS**
  - Store data in a database structure
  - Query with SQL-like syntax
  - We define how to query & process the data
  - Maintenance & resource overhead
  - Limited support for ragged data


















# File Storage & Chunking

- How should we structure the data?
- By shot does not facilitate **horizontal** access
  - Hierarchical view is human centric
  - Can be recovered from meta data
  - Slower access
- Alternative: store by signal for faster access
- NetCDF vs. Zarr vs. HDF
  - Chunking to **minimise I/O**
- **Assumption:** users will want to access a *few signals* and but data from *many shots*



# 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 

# Use Cases

- Potential use cases for such a platform are myriad
- Simple statistical analysis
- NLP on shot summaries
- Data validation/anomaly detection
- Experiment design
- Real-time control
- ... Ideas?

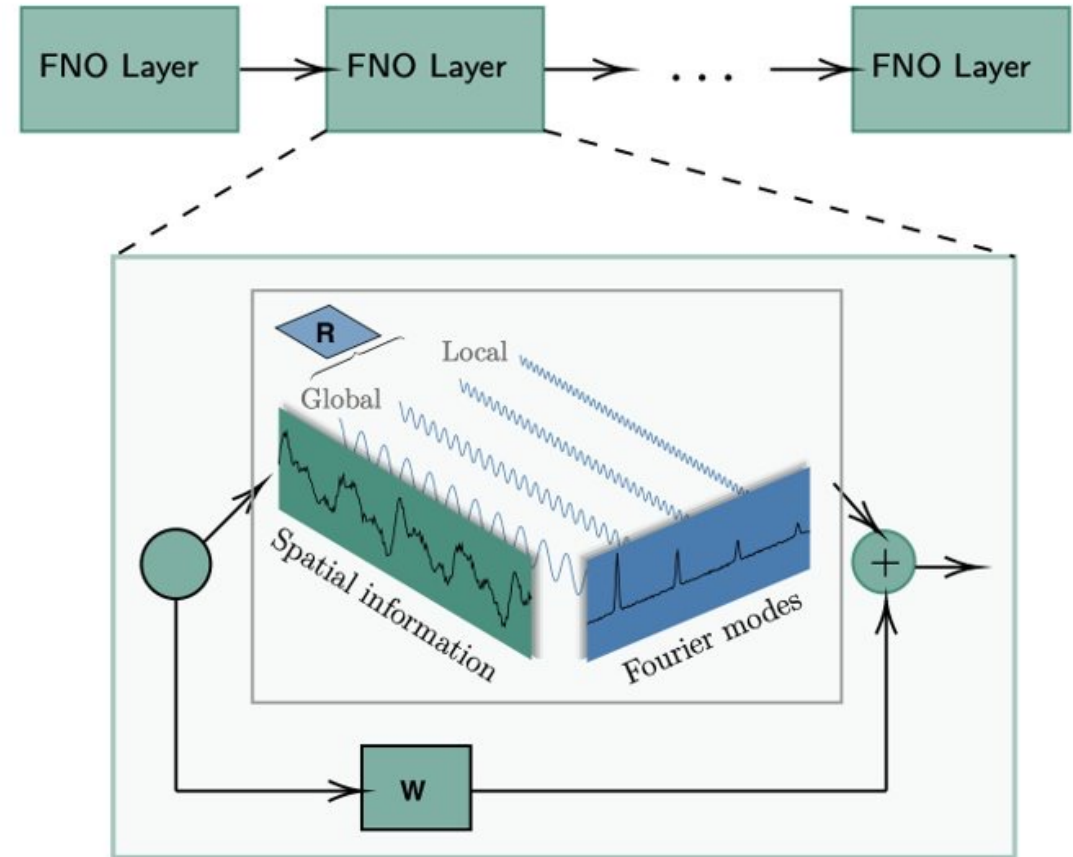
# Use Cases

## Digital Twins for Plasma Prediction

- Experimental Information capturing the plasma dynamics from the MAST FAIR database can be distilled into AI/ML models.
- Once trained these models could be run simultaneously with the plasma shot in the Tokamak allowing us to perform real-time forecasting.
- Considering the data abides by certain physics, we could modify the models to be informed by them.

### Fourier Neural Operator (FNO):

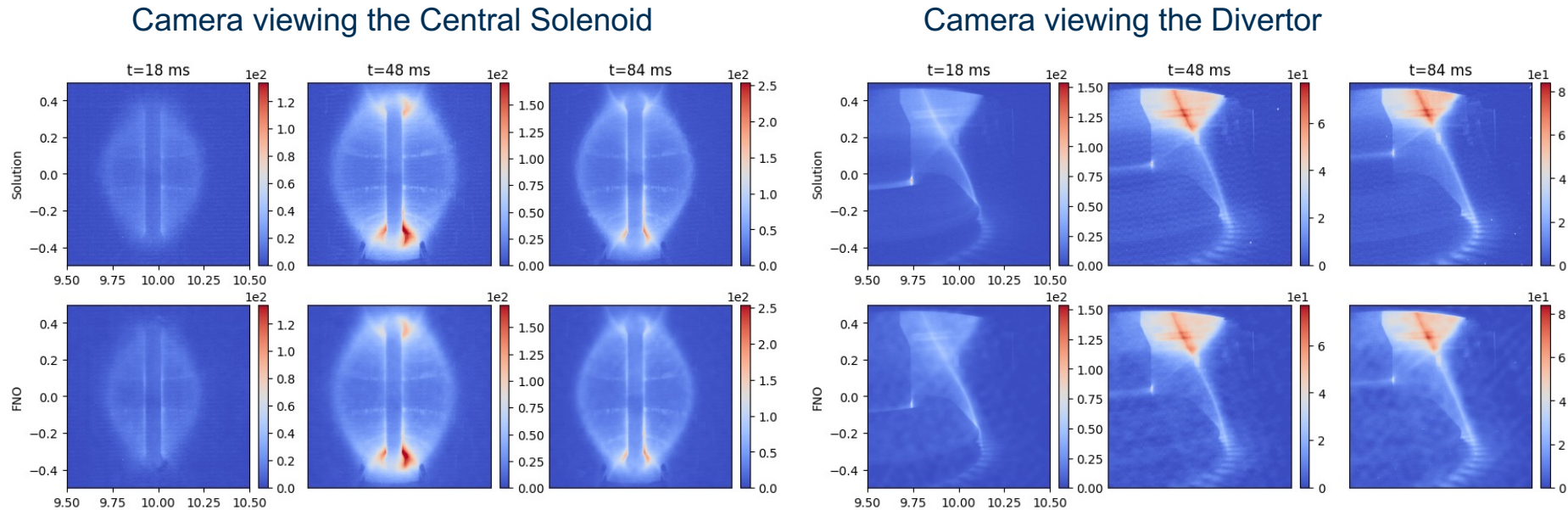
FNO is a neural network configuration that extracts the Fourier modes found within the data, allowing for better learning of the global and local features.



# Use Cases

## FNO - predicting plasma evolution

**FAST Cameras on MAST:** The Tokamak is fitted with cameras that capture the evolution of the plasma within the visible spectrum. These cameras are crucial diagnostic tools that provide the first line of visual inference in the control room. By using FNOs, we can predict in real-time the evolution of plasma across the central solenoid and at the divertor.



# Data-intensive, unsupervised disruption prediction on MAST

Nicola C. Amorisco, Stan Pamela et al.

Plasma disruptions limit performance and lifespan of tokamak/diagnostics.

First-principles prediction is challenging,  
AI approach promising for mitigation/avoidance.

**Data-intensive approach:** a variety of plant diagnostics and signals to predict disruption events.  
Raw data stream of **~1000D / 3.2ms** used in predictions.

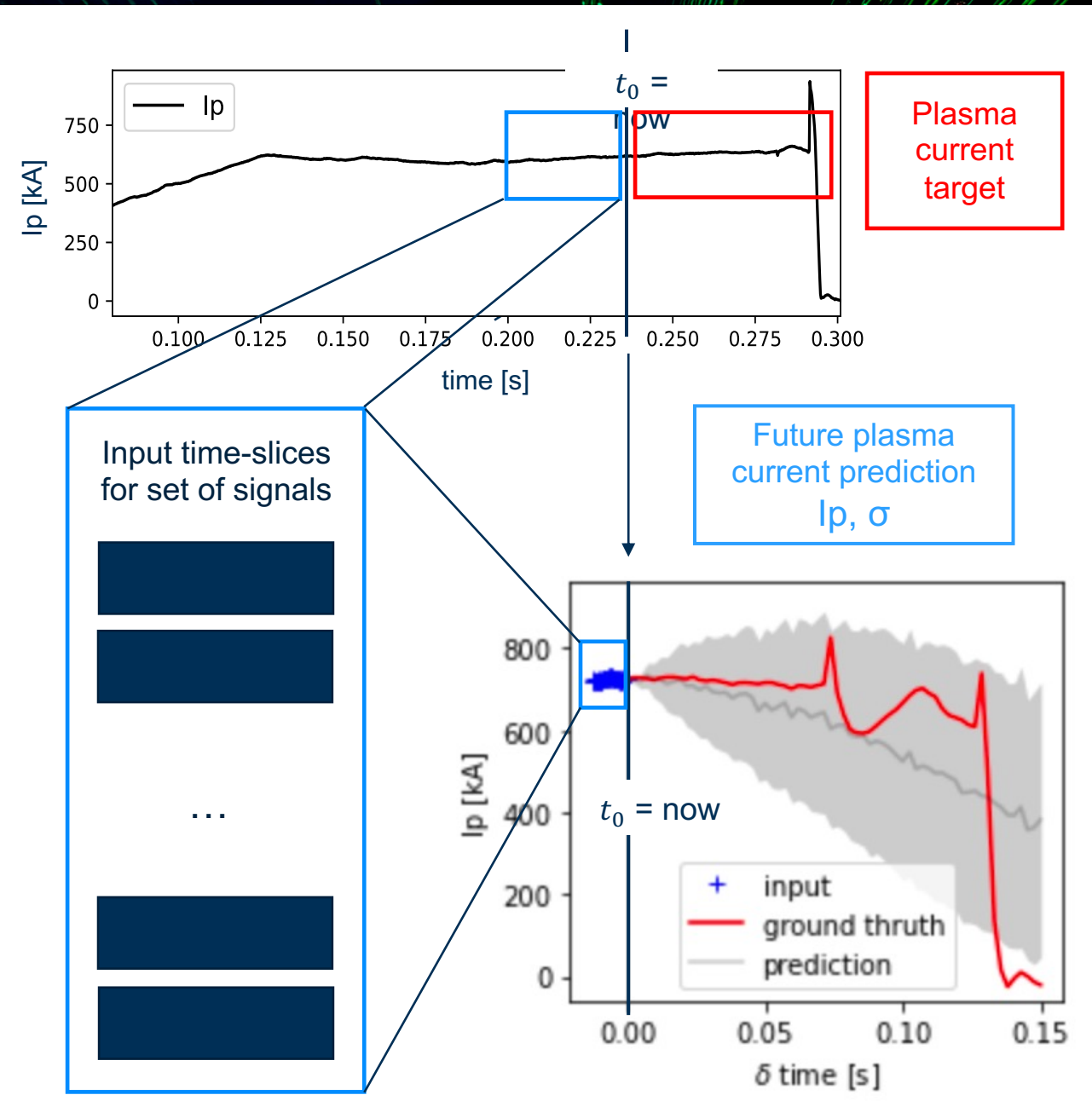
0D signals:

- Plasma current, Solenoid current
- NBI total power
- Electron density
- Set of equilibrium properties from EFIT

1D signals:

- Thomson scattering density and temperature profiles
- Arrays of magnetic probes
- Outboard Mirnov coils (STFT)

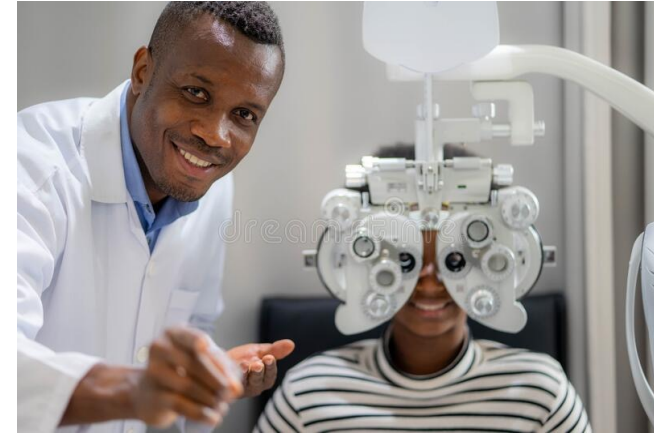
Data for ~18k MAST shots.



# Use Cases

## Google Optometrist

- Fusion reactors have **large, complex operational spaces** to explore.
- It can also be hard to write down formally what **good** looks like...
- Google worked with fusion start up TAE technologies to develop the **Google Optometrist**.
- Optometrist aims to efficiently explore the space while optimising a **hidden utility model**.
- It does this by **asking the operator** to rate shots against each other.
- This approach allowed TAE to identify a **new operating regime** on their device with lower power loss.

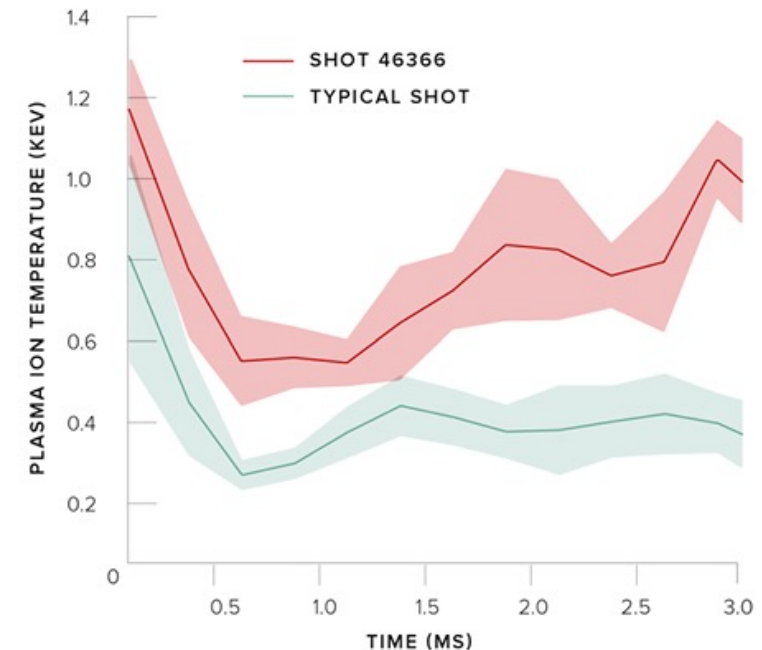


```

Define meta-parameters;
Choose reference settings;
while Experiment running do
  Construct proposal;
  Run plasma shot;
  Human evaluation of pair of shots;
  Update reference settings;
end

```

**Algorithm 1.** The Optometrist Algorithm.





“

*The MCF and plasma physics community should copy the Pangeo project and its software ecosystem, because geosciences have already solved many of the software problems we still struggle with.*

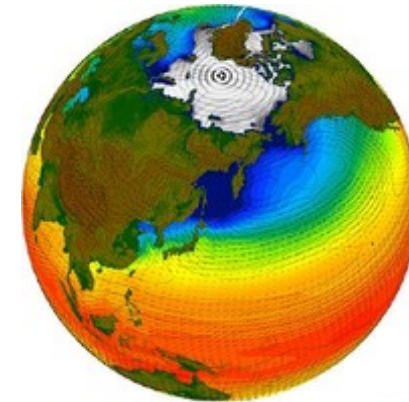
”

Dr Tom Nicholas - RSE @ Columbia University, NY / XArray core developer

# Learning from other communities

## Pangeo

- Pangeo is a software community to enable big data geoscience research
- Collection of open-source tools to handle large, complex spatio-temporal data
- Many tools are **domain agnostic**:
  - Larger than memory computation support
  - Scalable/parallel computation
  - Multi-dimensional array manipulation
  - Visualisation tools



hvPlot

**PANGEO**

# Learning from other communities

## Similarity to Pangeo

- We are designing this database along the same ethos:
  - *Solve a general problem*
  - *Clearly define a scope*
  - *Avoid duplication*
  - *(Consume and) produce standard data container*
  - *Avoid data I/O where possible*
  - *Operate Lazily (or in our case, facilitate lazy operation)*

# Summary

- Some MAST data is **already open**
- **Not optimised** for data-intensive applications
- We are developing an **AI/ML friendly** database for MAST data
- By adhering to **FAIR principles**, we can maximise the scientific utility of our data
- Lessons can be **learned** for other fields like geoscience's Pangeo community

# What can you do?

- When developing software tools
  - Pangeo's "best practices"
  - As open as possible
  - Licence
  - DOCUMENT
- When using data
  - Assess the **FAIR**ness
  - Feedback to producers
  - Recommend to others

**Thank you!**