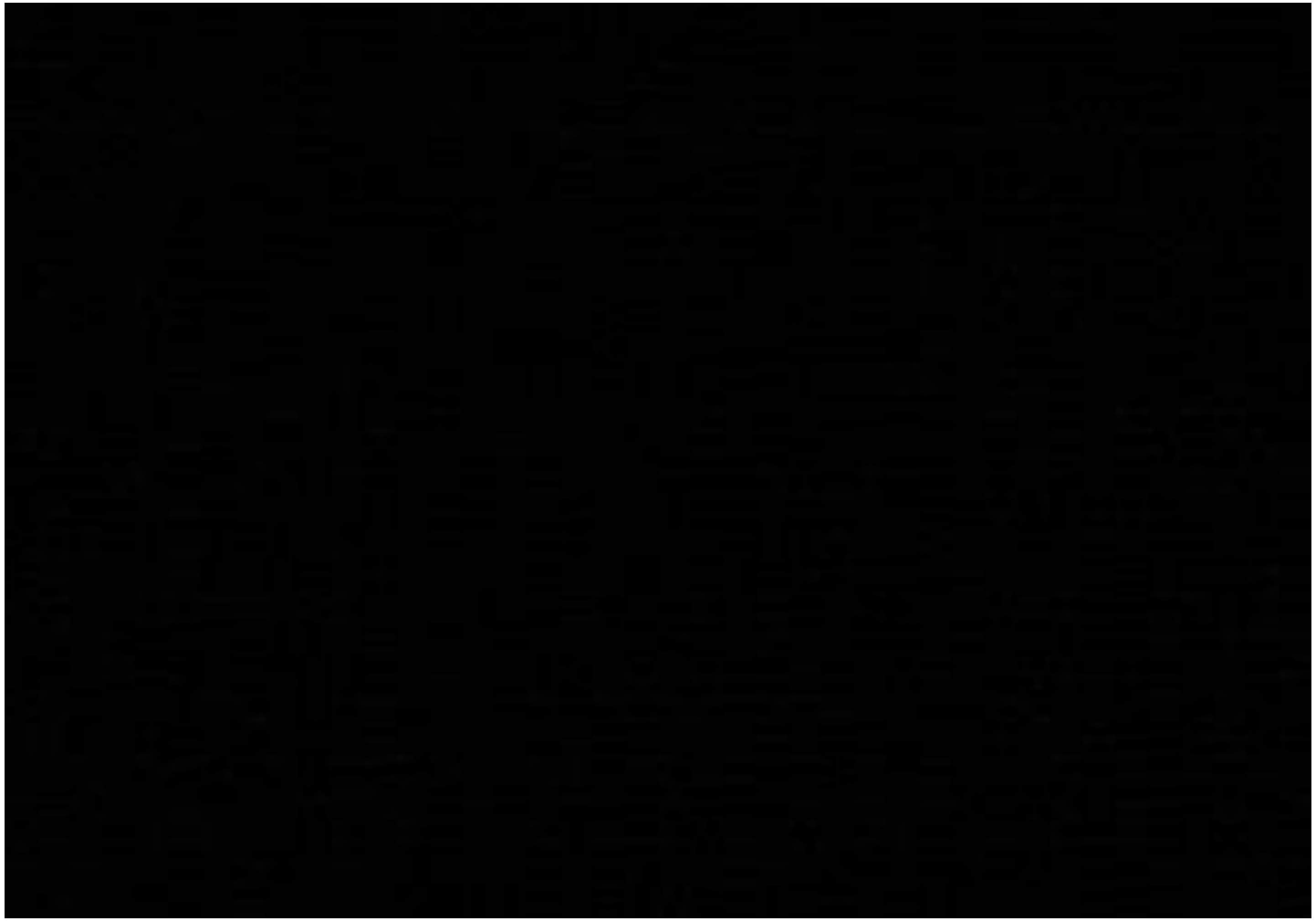
Digital Signal Processing & Data Science Challenge

Simon McIntosh
ITER Organization

Monday 9th December 2024

ITER International School
Nagoya, Japan

Bio:
Aerospace Engineering University of Bristol
PhD Fluid Mechanics University of Cambridge
Postdoc / Lecturer University of Oxford
Staff Culham Center for Fusion Energy
Staff ITER Organization Scientific Data Processing

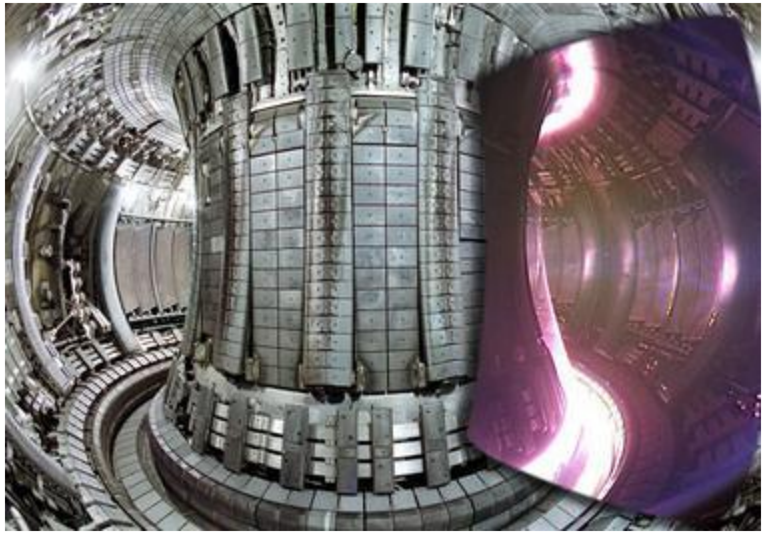


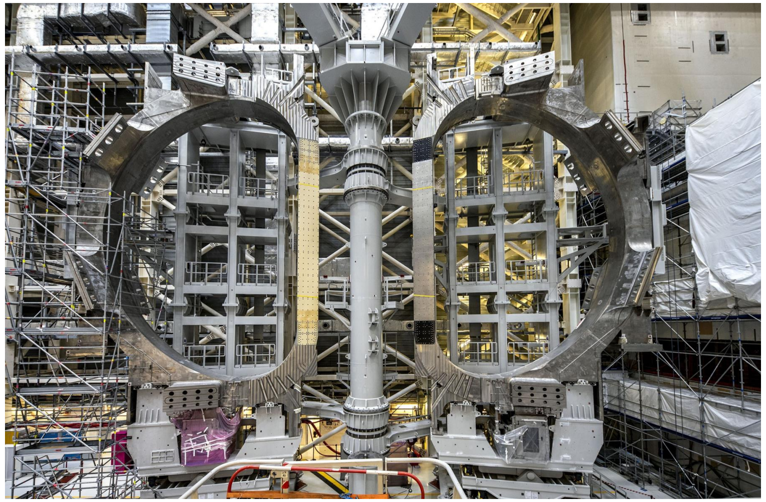
iter
china eu india japan korea russia usa

This presentation focuses on applied Data Science for Fusion.

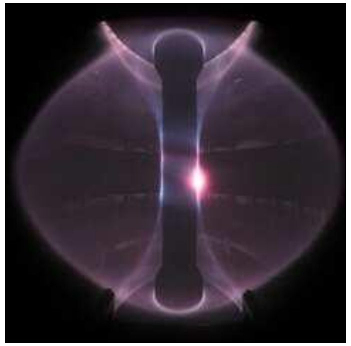
The elephant in the room.

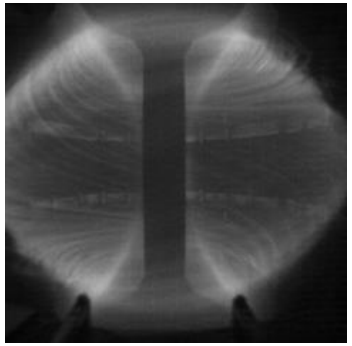
Applied Data Science in Fusion.

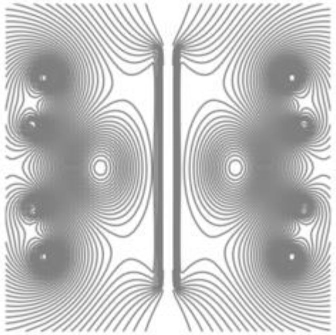




Kaggle Challenges.





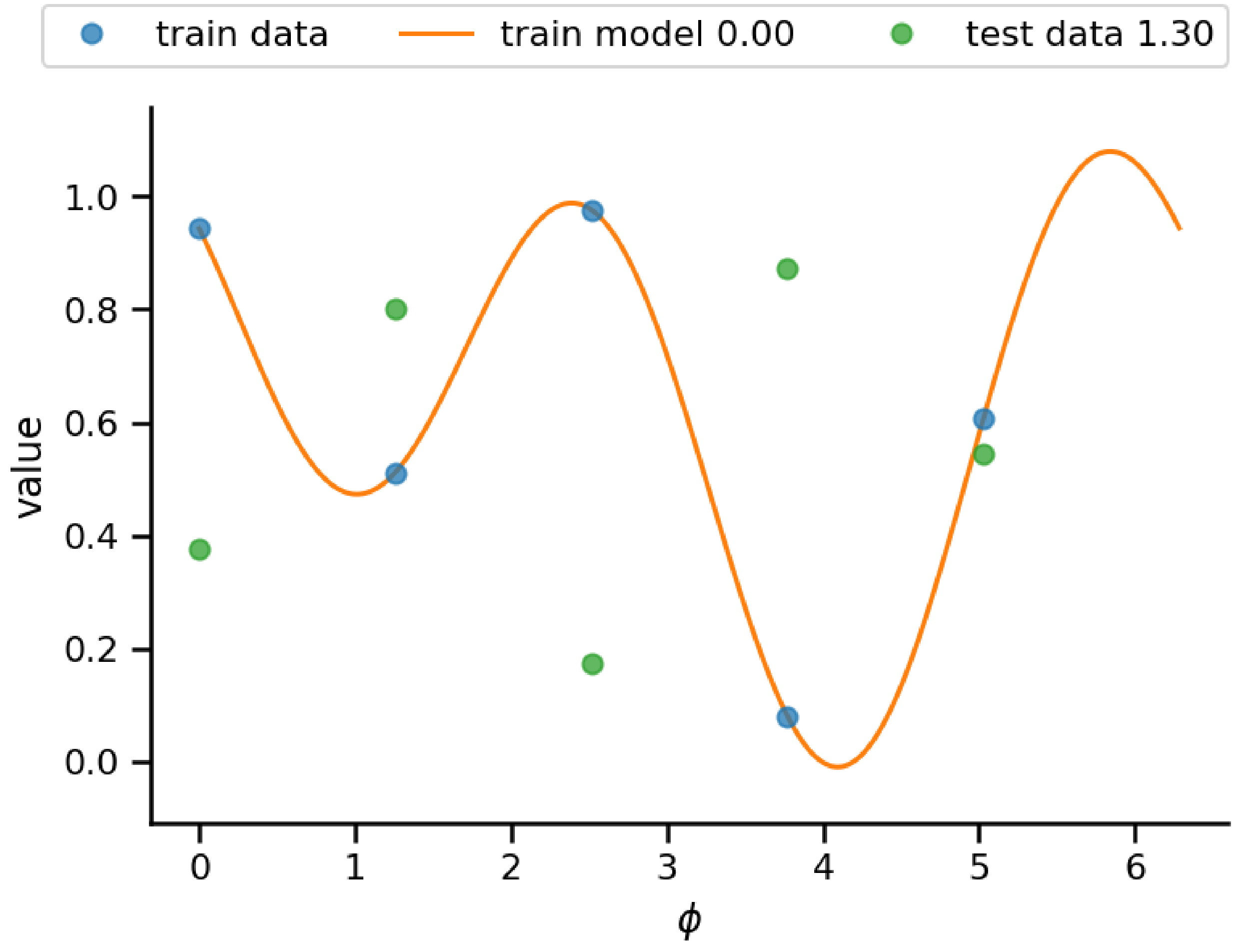


The elephant in the room is over-fitting.

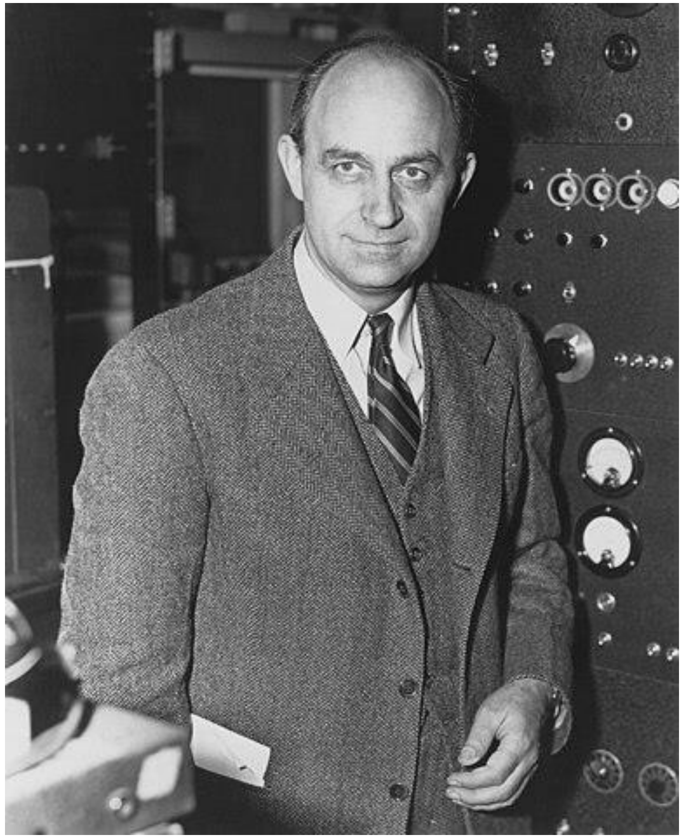
“All models are wrong, but some are useful”. George Box

Non-linear parametrizations can fit arbitrarily complexity structures.

$$Cn = \begin{bmatrix} 3.12 + 0j \\ 0.43 - 0.43j \\ 0.36 + 0.91j \end{bmatrix}$$



“With four parameters I can fit an elephant, and with five I can make him wiggle his trunk”.



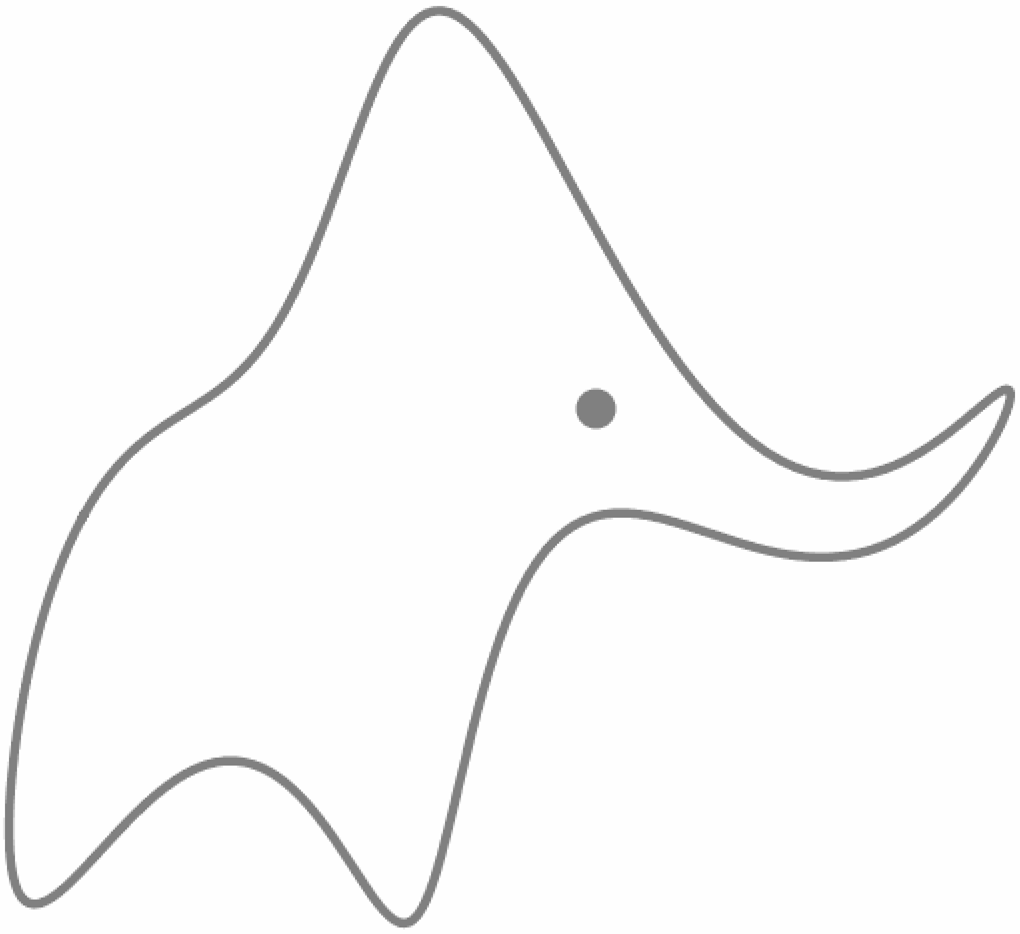
Enrico Fermi

John von Neumann

Simple basis functions can generate complex shapes.

von Neumann's Elephant

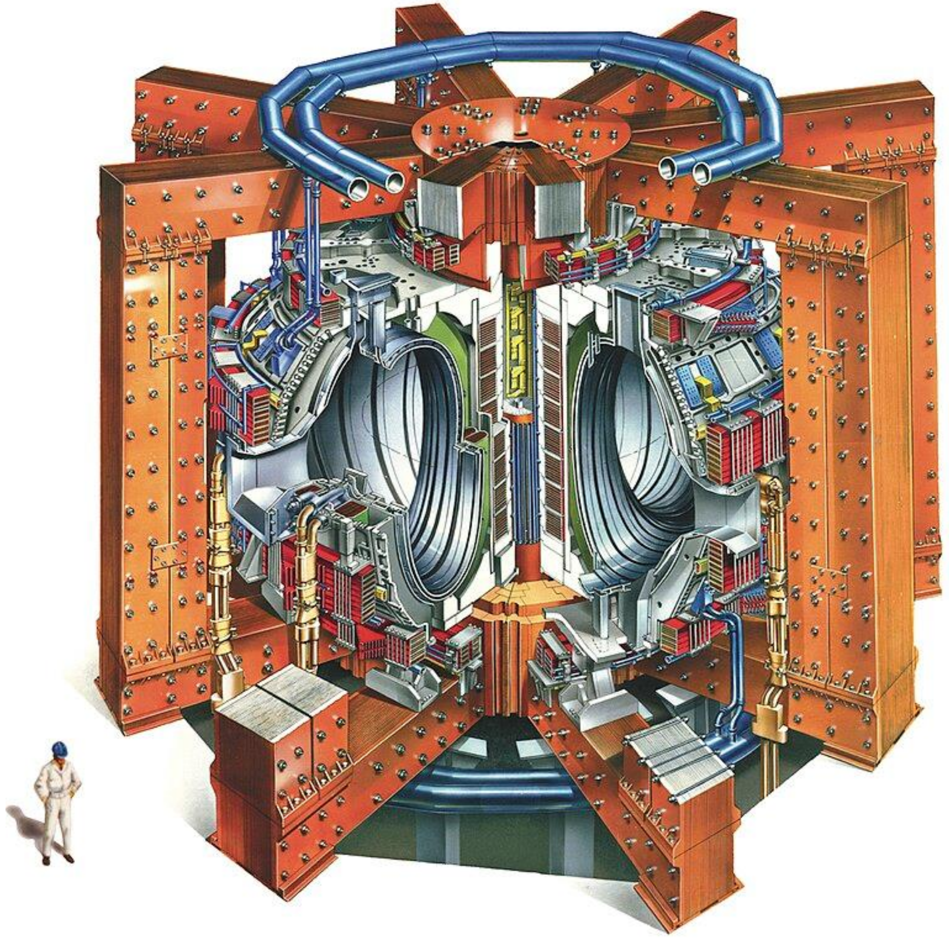
$$Pn = \begin{bmatrix} -55 + 15j \\ -9 \ - 4j \\ 0 + 7j \\ -5 \ - 11j \\ \boxed{20} \boxed{+ 1j} \end{bmatrix}$$



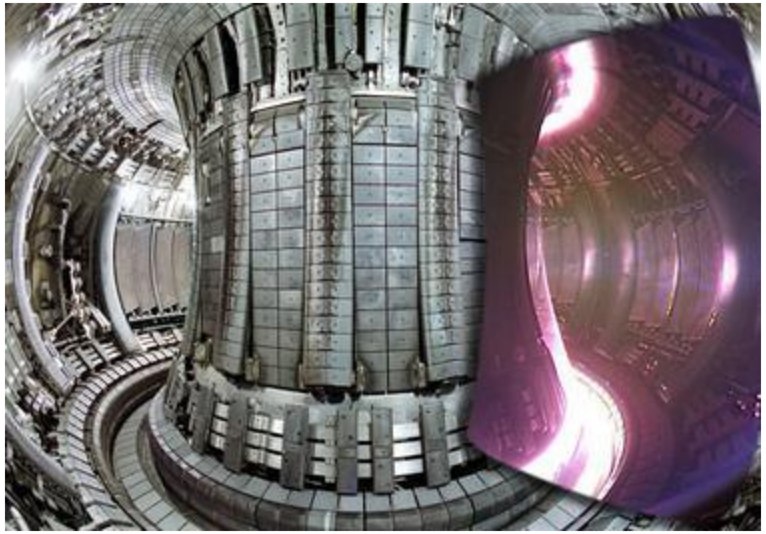
$$Cn = \begin{bmatrix} 0 + 0j \\ \boxed{Pn[0]} \\ \boxed{Pn[1]} \\ \boxed{0 + 0j} \\ 0 + 0j \\ \boxed{Pn[2]} \\ \boxed{Pn[2]} \\ 0 + 0j \\ \boxed{0 + Pn[3]j} \\ \boxed{-Pn[1]} \\ \boxed{Pn[3] - Pn[0]j} \end{bmatrix}$$

Scientific Data is a valuable product of expensive experiments.

If not properly archived the value of this data depreciates rapidly.



Capital cost ~0.5 billion 2014 US dollars



Operating cost ~200,000 euros per day

~~Code~~ Data is read more often than it is written.

~~Code~~ Data should always be written in a way that promotes readability.

The IMAS Data Dictionary defines an extensive set of attributes.

Attribute sets are grouped as Interface Data Structures IDSs.

amns_data	disruption	langmuir_probes	refractometer
barometry	distribution_sources	lh_antennas	sawteeth
bolometer	distributions	magnetics	soft_x_rays
bremsstrahlung_visible	divertors	mhd	spectrometer_mass
calorimetry	ec_launchers	mhd_linear	spectrometer_uv
camera_ir	ece	mse	spectrometer_visible
camera_visible	edge_profiles	nbi	spectrometer_x_ray_crystal
camera_x_rays	edge_sources	neutron_diagnostic	summary
charge_exchange	edge_transport	ntms	temporary
coils_non_axisymmetric	em_coupling	pellets	thomson_scattering
controllers	equilibrium	pf_active	tf
core_instant_changes	gas_injection	pf_passive	transport_solver_numerics
core_profiles	gas_pumping	plasma_initiation	turbulence
core_sources	gyrokinetics	polarimeter	wall
core_transport	hard_x_rays	pulse_schedule	waves
cyrostat	ic_antennas	radiation	workflow
dataset_description	interferometer	real_time_data	
dataset_fair	iron_core	reflectometer_profile	

Diagnostics

Heating systems

The IMAS Data Dictionary defines attributes in a tree-like structure.

Coordinates, units and descriptions are attached to these attributes.

ITER Physics Data Model Documentation for equilibrium

Description of a 2D, axi-symmetric, tokamak equilibrium; result of an equilibrium code.

Notation of array of structure indices: itime indicates a time index; i1, i2, i3, ... indicate other indices with their depth in the IDS. This notation clarifies the path of a given node, but should not be used to compare indices of different nodes (they may have different meanings).

Lifecycle status: active since version 3.1.0

Last change occured on version: 3.42.0

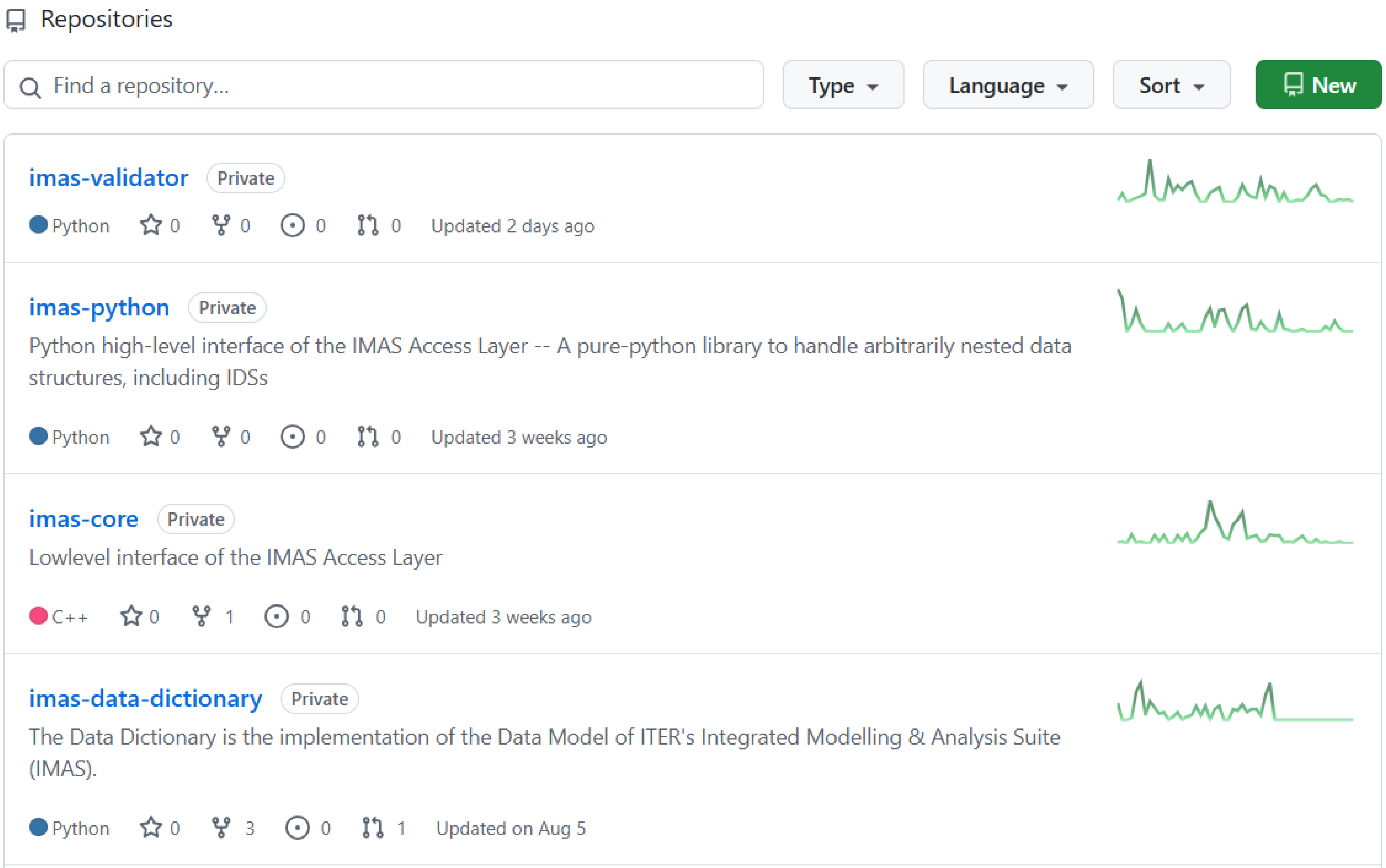
Back to top IDS list

Flat display Show/Hide errorbar nodes By convention, only the upper error node should be filled in case of symmetrical error bars. The upper and lower errors are absolute and defined positive, and represent one standard deviation of the data. The effective values of the data (within one standard deviation) will be within the interval [data-data_error_lower, data+data_error_upper]. Thus whatever the sign of data, data_error_lower relates to the lower bound and data_error_upper to the upper bound of the error bar interval.

Full path name	Description	Data Type	Coordinates
▸ ids_properties	Interface Data Structure properties. This element identifies the node above as an IDS	structure	

The IMAS Data Dictionary is in the process of being open-sourced.

Check back on github.com/ITER-Organization in the coming months.



IMAS Data is now available as self-describing netCDF files.

Two of the Data Science Challenges use a netCDF input format.

- Store IDS data in a “tensorized” form
  - Equilibrium example:
    `time_slice(i)/profiles_2d(j)/psi(k,l)`
      -> `time_slice.profiles_2d.psi(i,j,k,l)`

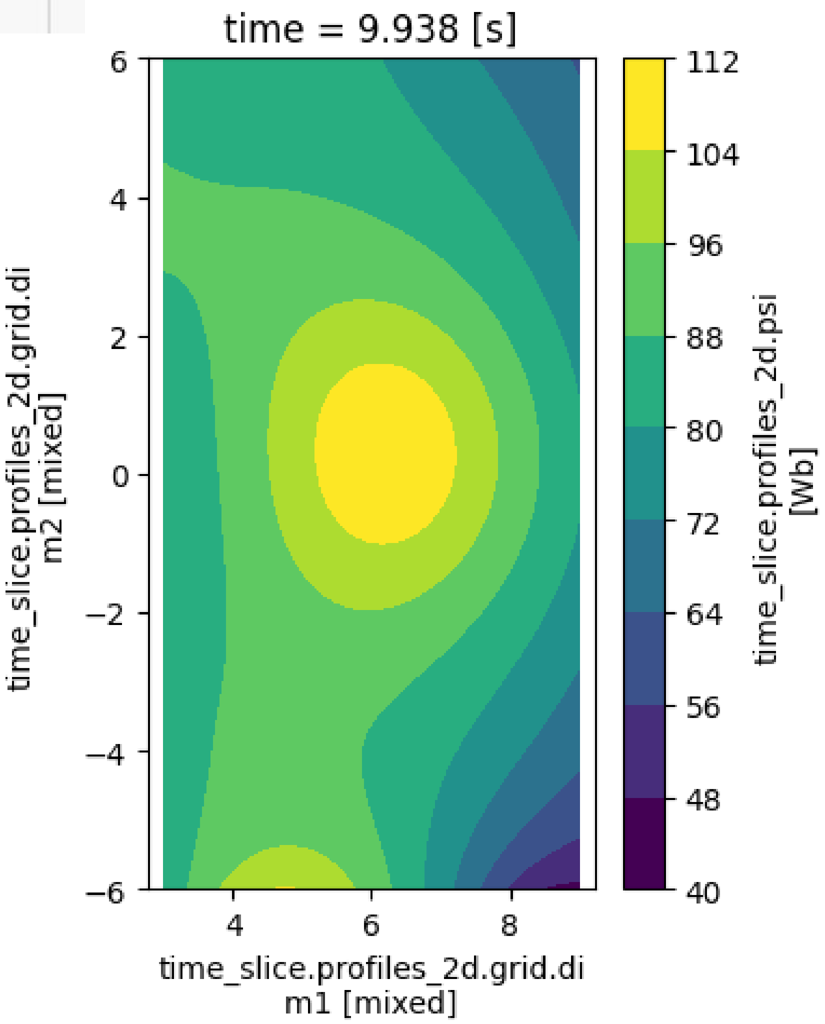
    2D data in 2 levels of AoS becomes a 4D array

- Labelled dimensions and coordinates
  - `time_slice.profiles_2d.psi(i,j,k,l)` has 4 dimensions
    1. `time` with coordinate `time`
    2. `time_slice.profiles_2d`, which is an index
    3. `time_slice.profiles_2d.grid.dim1`
    4. `time_slice.profiles_2d.grid.dim2`

- Additional metadata for
  - Units (`Wb`)
  - Documentation (`Values of the poloidal flux at the grid in the poloidal plane`)
  - Metadata follows the “CF Conventions” (developed for geosciences) as much as possible

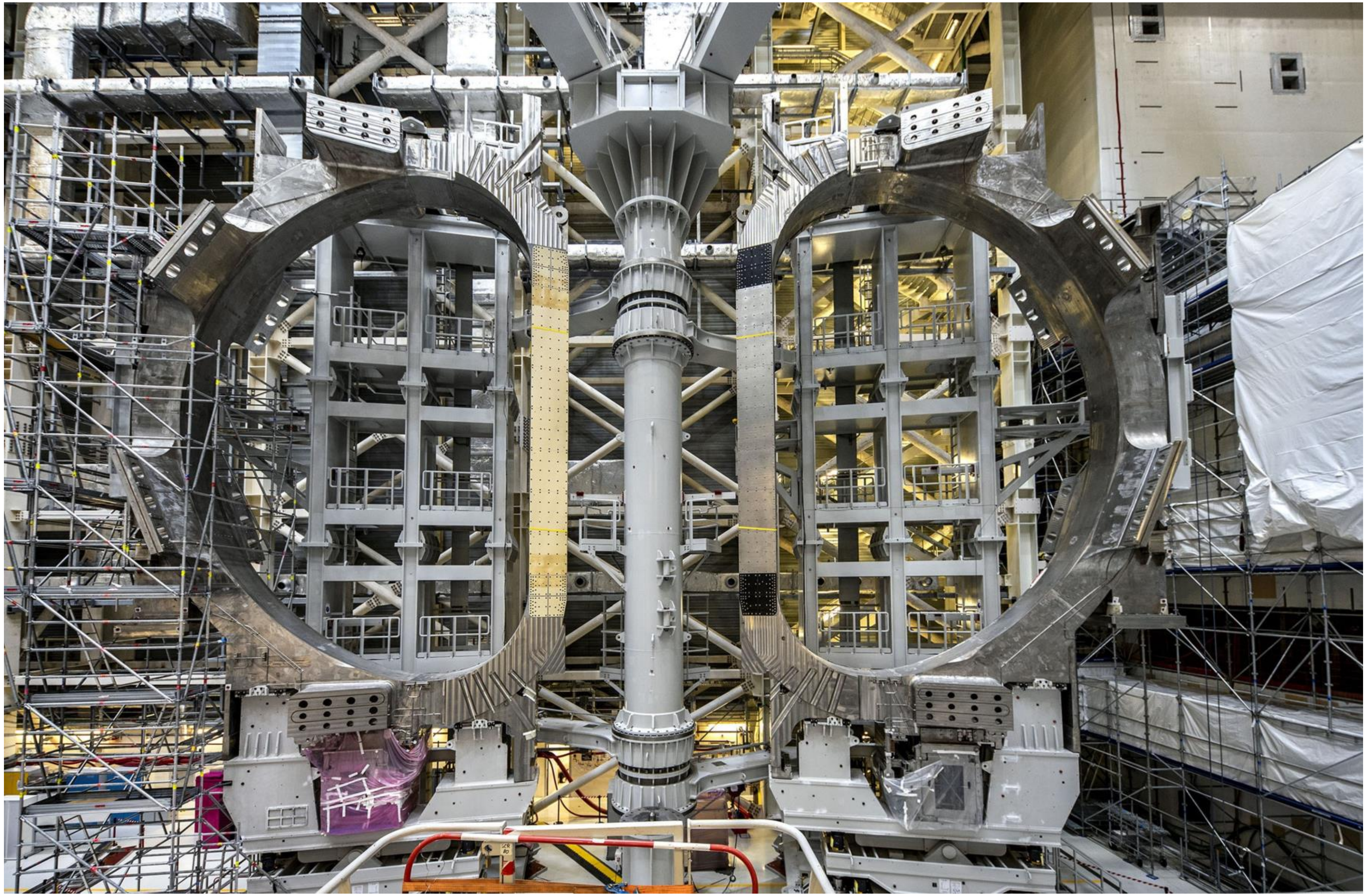
Self-describing IMAS data without a custom Access Layer (xarray).

```
psi = ds["time_slice.profiles_2d.psi"].isel({"time_slice.profiles_2d:i": 0})
```



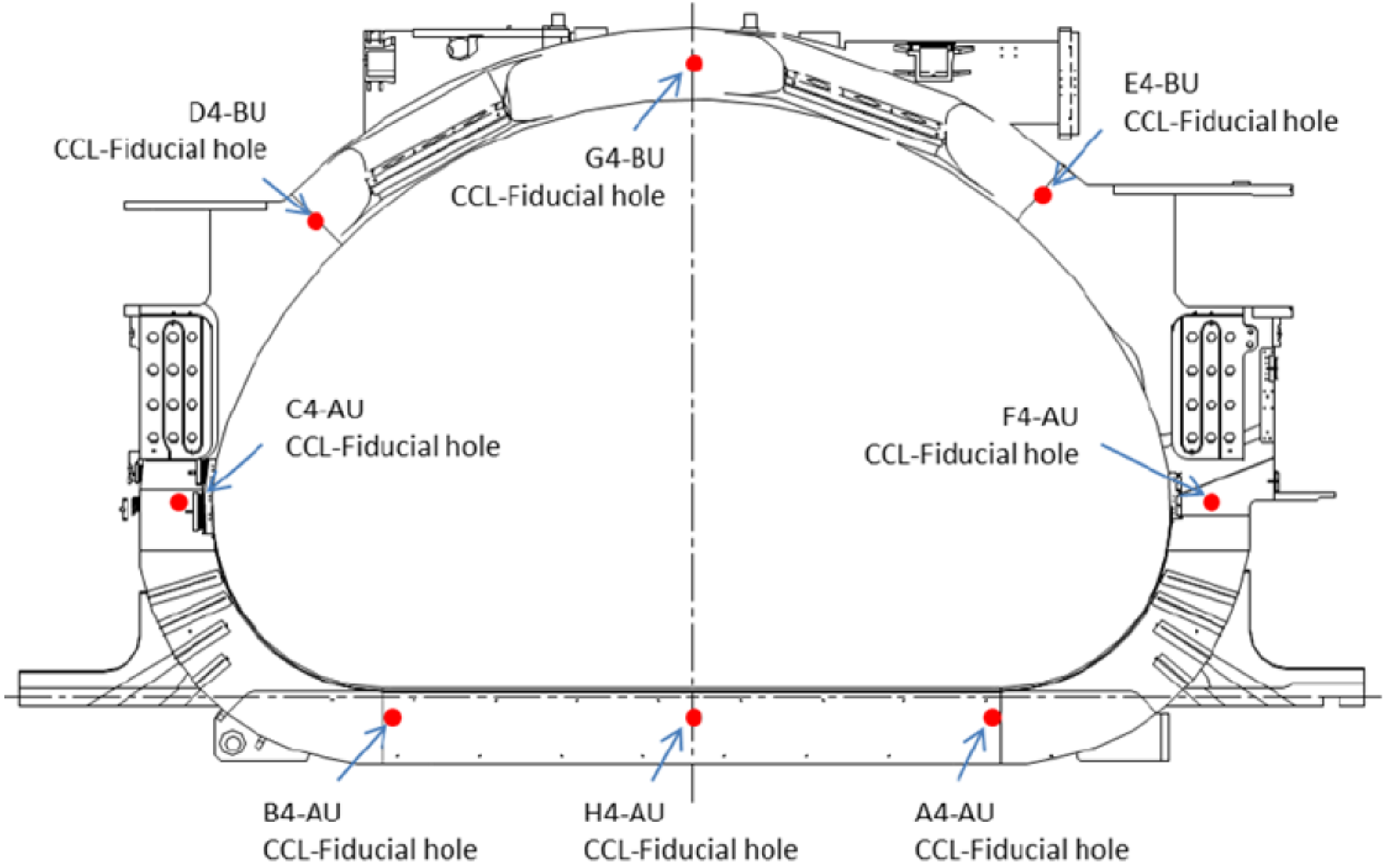
The alignment of ITER's 17 meter high, 360 tonne D-shaped Toroidal Field magnets is a feat of precision engineering.

Exceptionally low tolerances that are repeatable and stable over time.

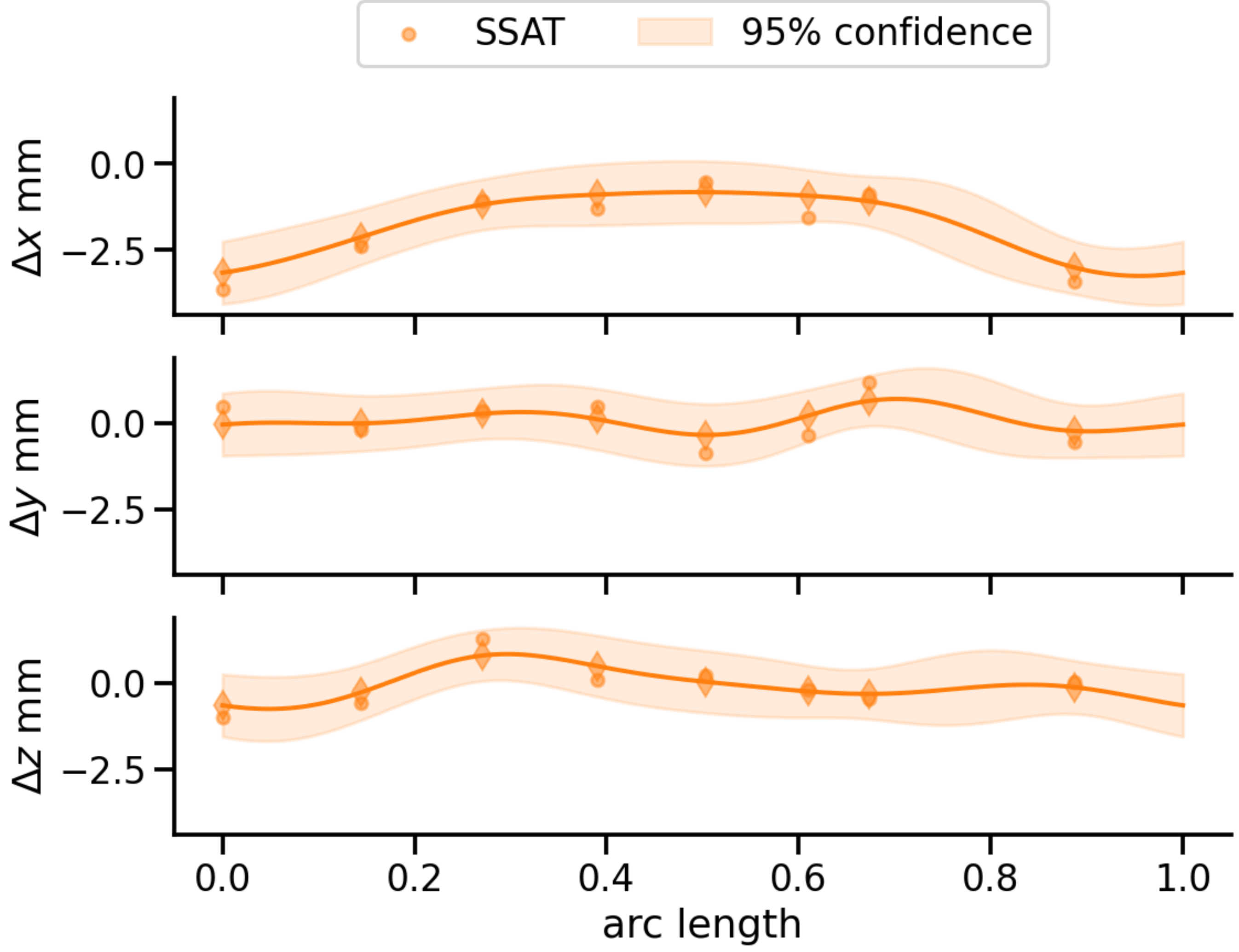


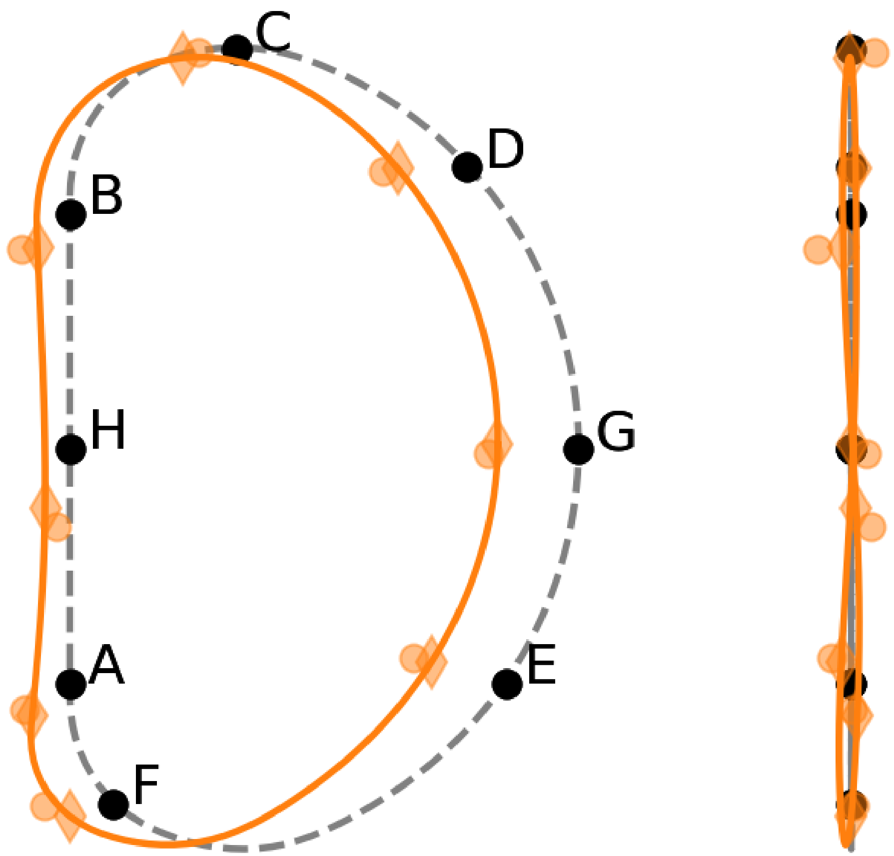
Data Science in Fusion is not restricted to the Physics domain.

Gaussian Process Regression used on ITER to reconstruct coil centerline.



New metrology for Sector #7 shared with Science Division last week.

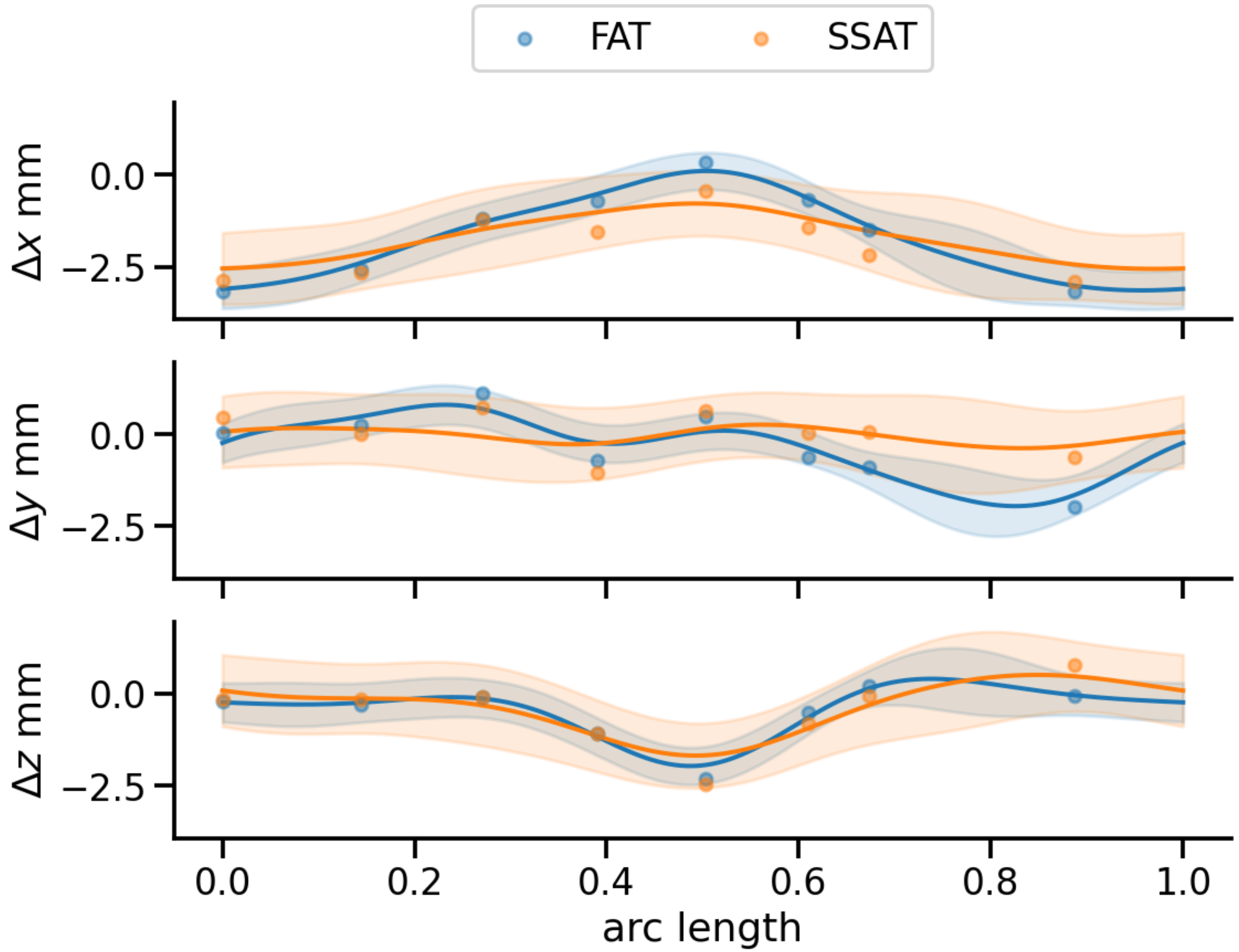


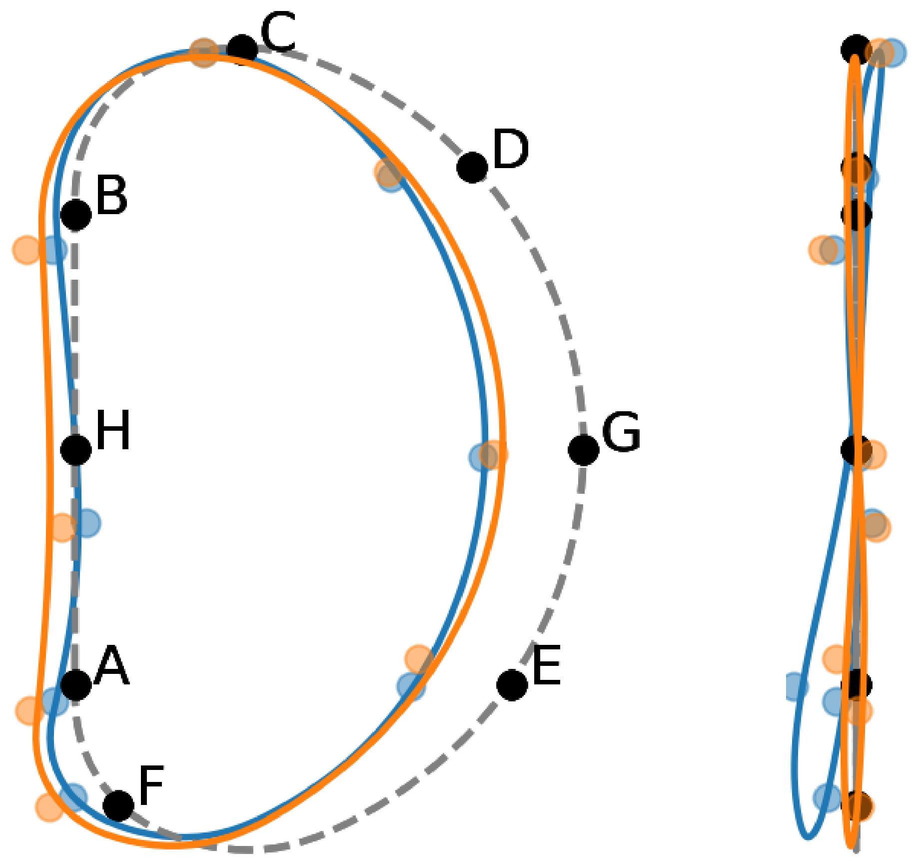


Coil alignment is based on inferred CCL positions (diamonds)

Deformation x500

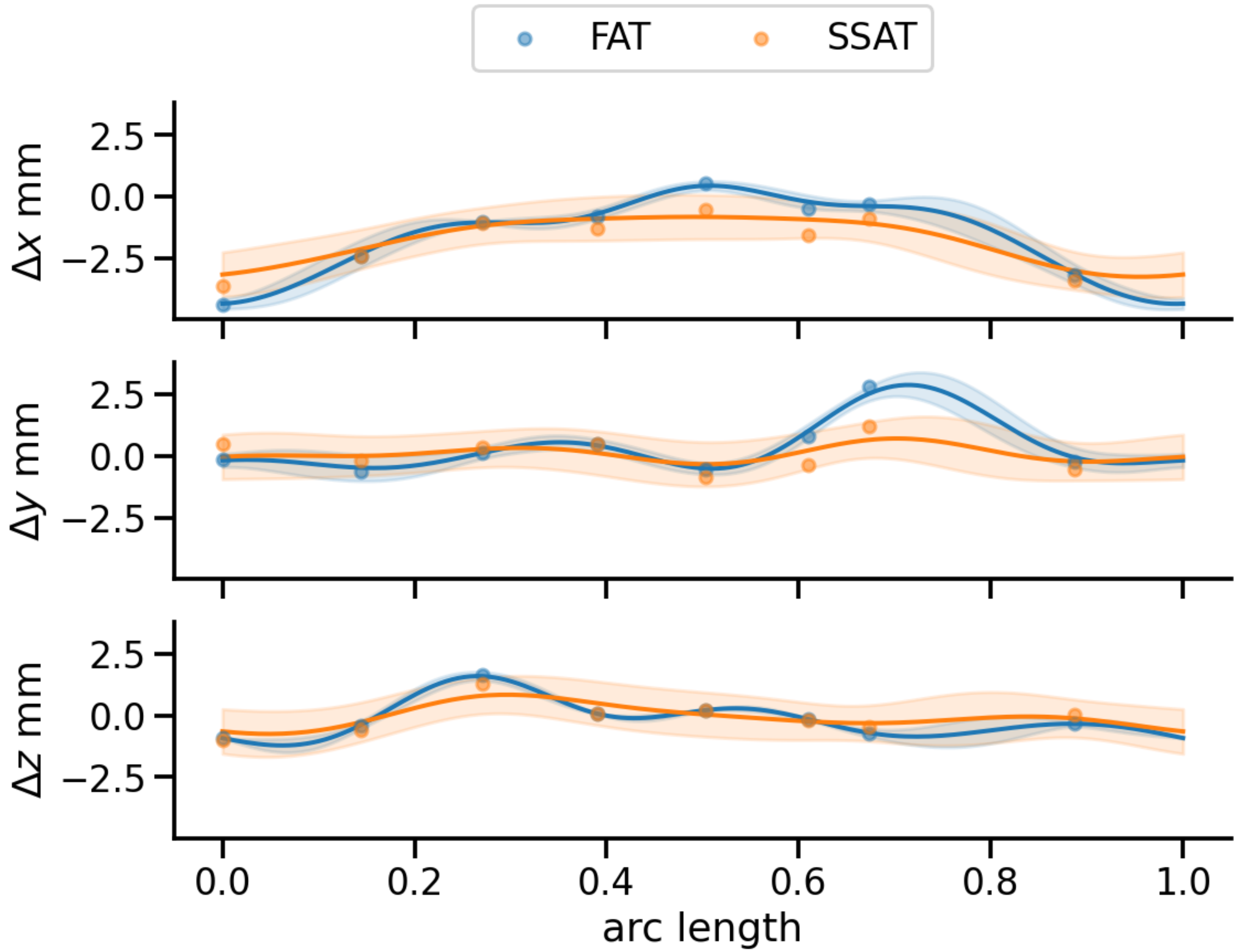
The orientation of each TF Coil affects its shape (Coil #8 Japan)

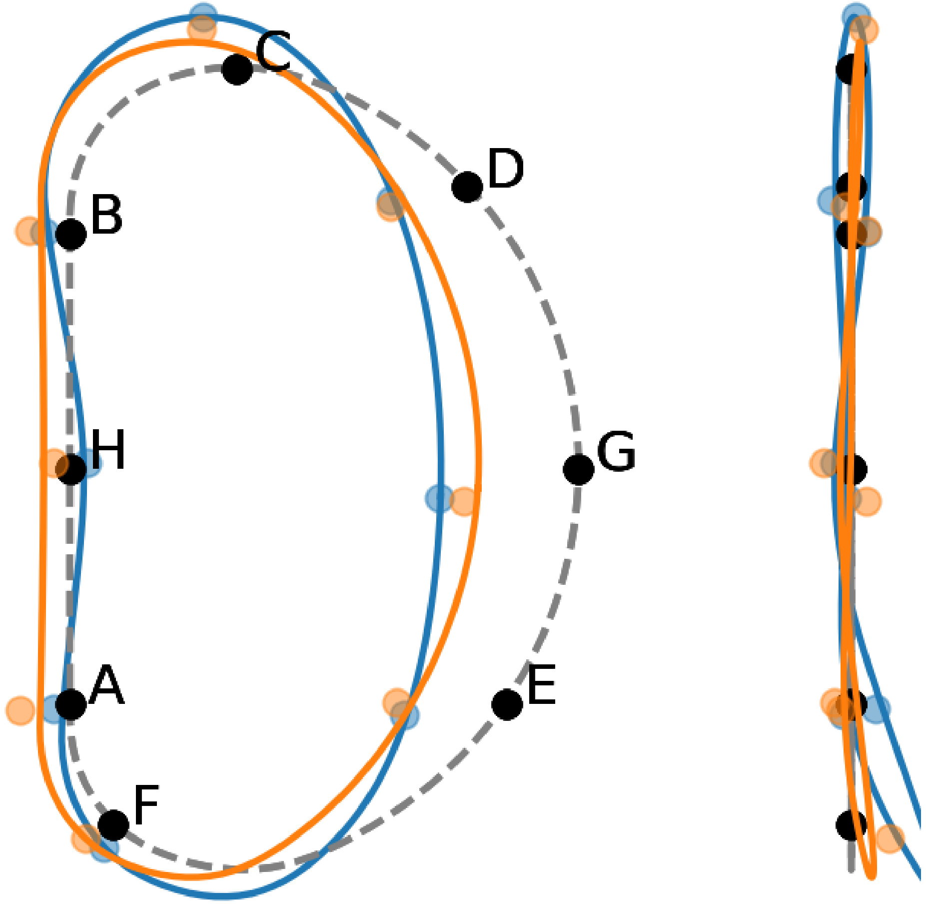




Deformation x500

The orientation of each TF Coil affects its shape (Coil #9 EU)

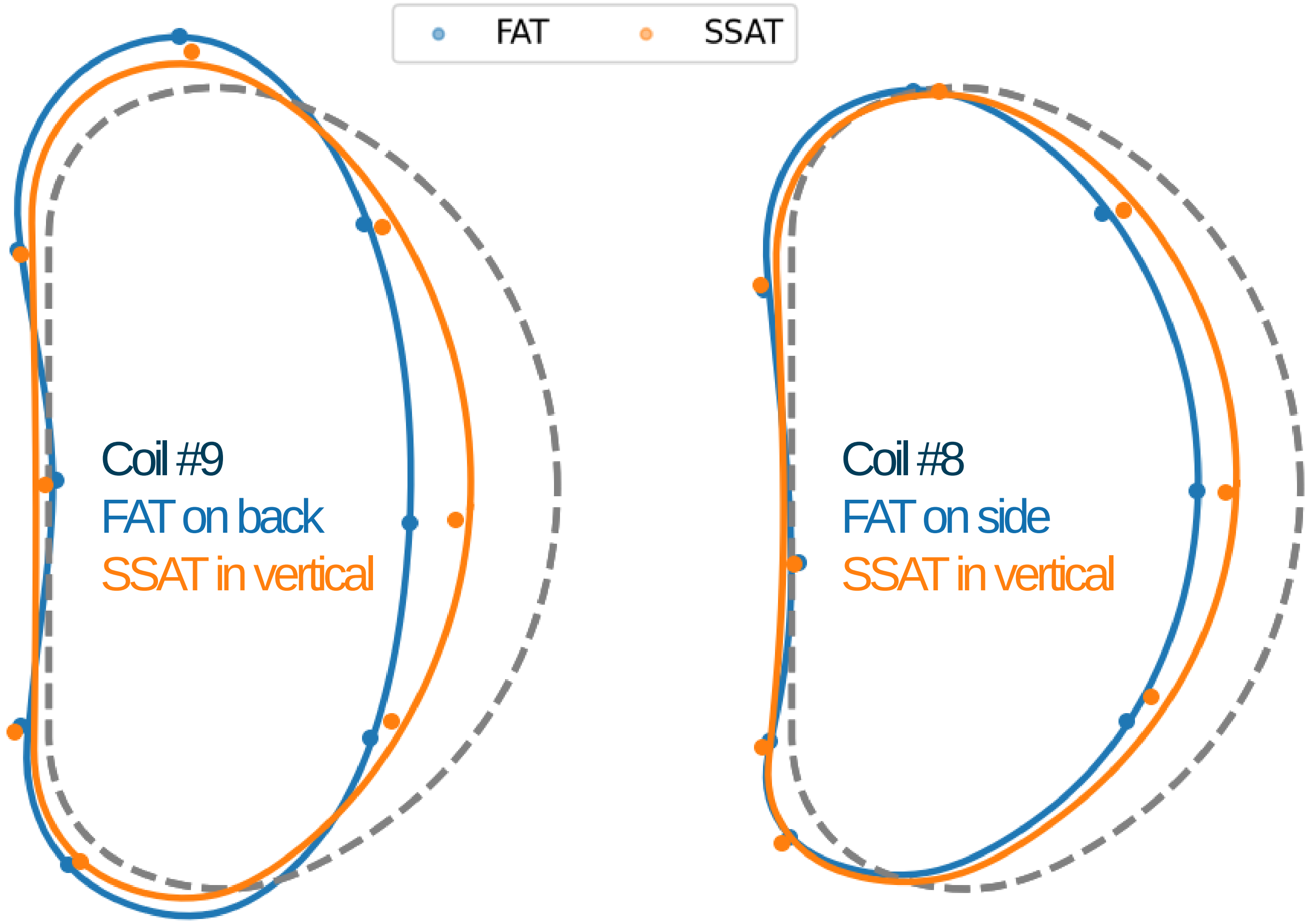




Deformation x500

Metrology of TF Coils in the vertical improves EU-JA agreement

Coil metrology carried out with a common orientation reduces the magnitude of the 'vendor' error field



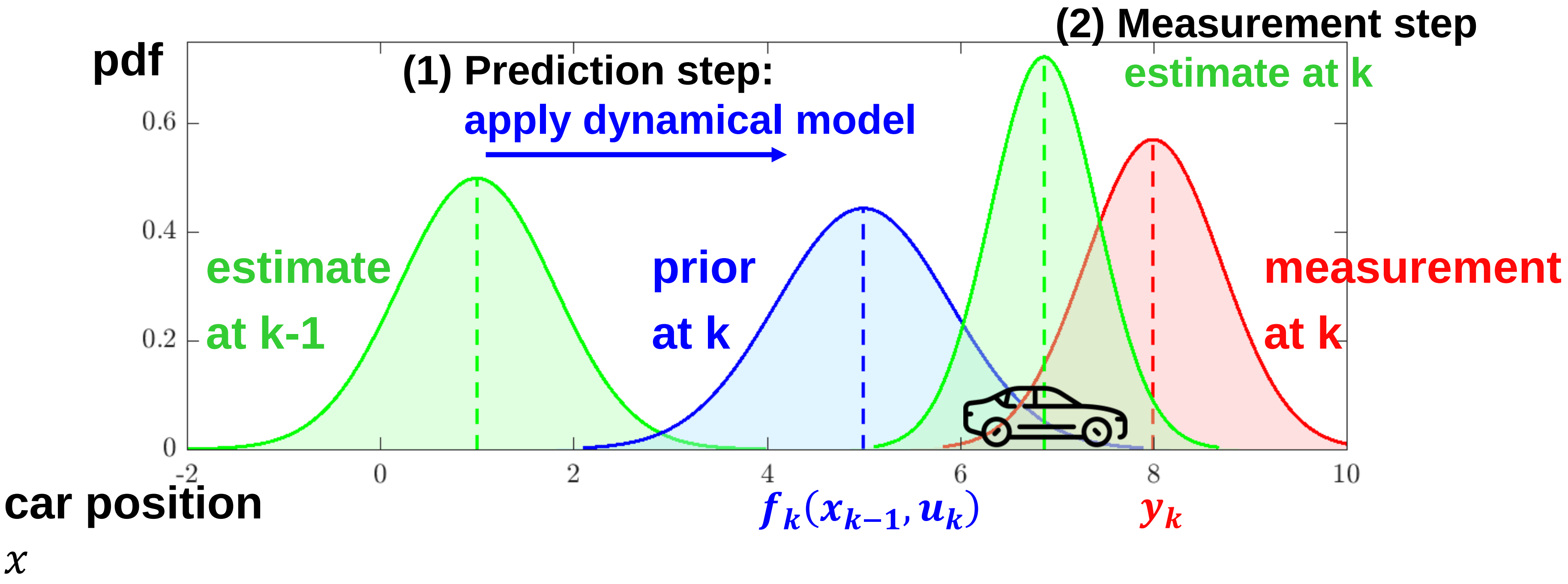
Deformation x500

The Extended Kalman Filter: a sequential Bayesian filter

Example 

Dynamic model for a car's position $\boldsymbol{x_k} = \boldsymbol{f_k}(\boldsymbol{x_{k-1}}, \boldsymbol{u_k}) + \boldsymbol{w_k}$ with actuators u_k and **linearization** $\boldsymbol{F_k}$, affected by noise $\boldsymbol{w_k}$ with **model uncertainty covariance** $\boldsymbol{Q_k} = \boldsymbol{E}[\boldsymbol{w_k}\boldsymbol{w_k^T}]$

and a series of **measurements** $\boldsymbol{y_k}$, affected by noise $\boldsymbol{\nu_k}$ with **measurement covariance** $\boldsymbol{R_k} = \boldsymbol{E}[\boldsymbol{\nu_k}\boldsymbol{\nu_k^T}]$

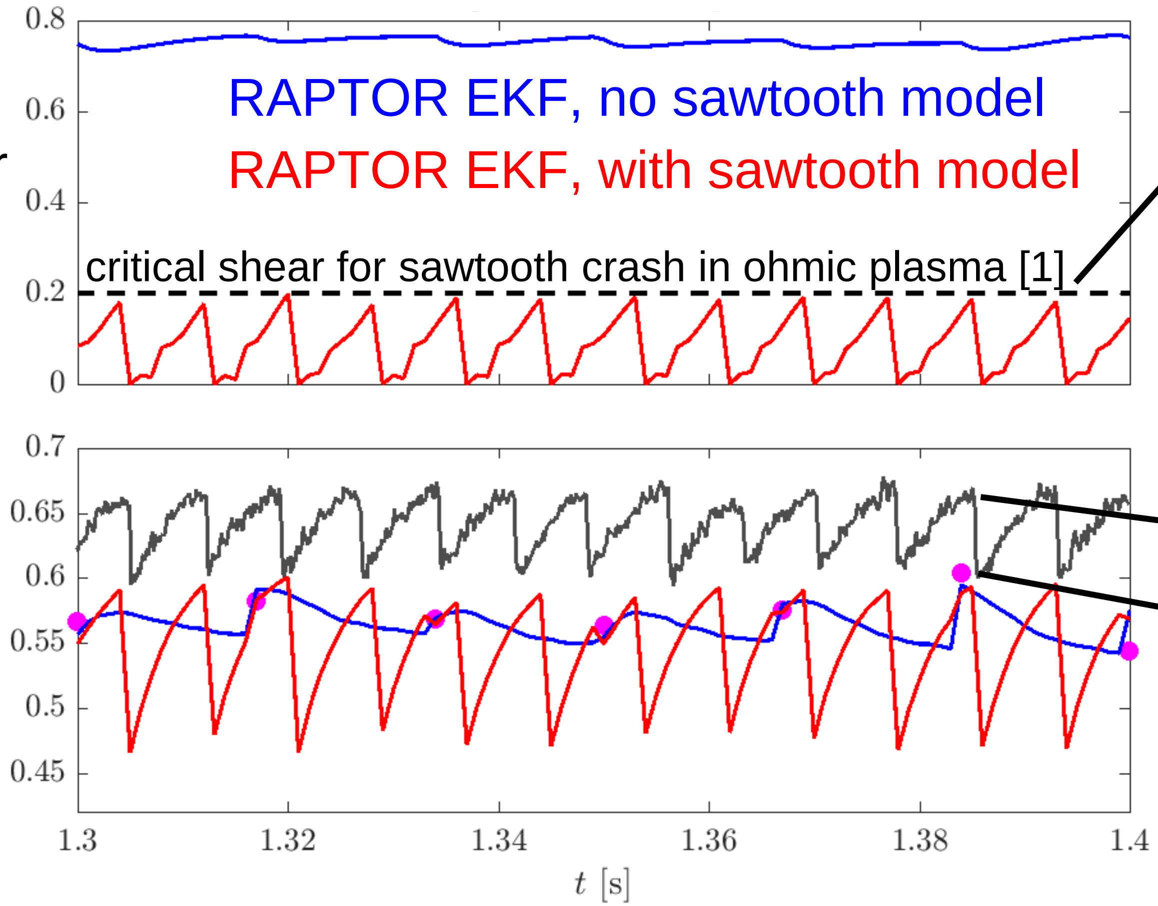


Sawtooth model allows realistic inter-measurement prediction

TCV 64965
(ohmic)

magnetic shear at $q = 1$

central T_e [keV]



sawooth crash triggered when **modelled magnetic shear** at $q = 1$ reaches $s_{crit} = 0.2$

soft X ray T_{e0}

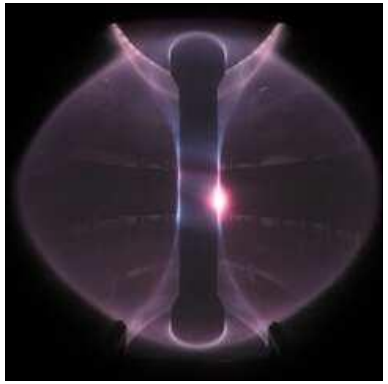
Thomson scattering

Adaptive EKF scheme: RAPTOR **model parameters** are continuously adapted based on the past measurements, allowing for realistic predictions for time points between measurement points

[1] O. Sauter et al, Varenna (1999)

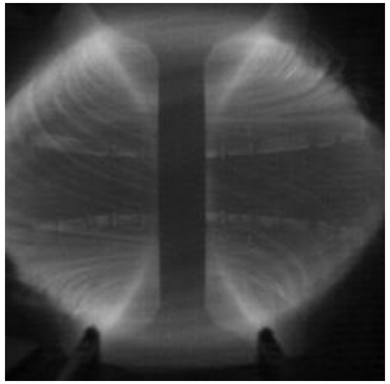
from S. Van Mulders et al, to be subm. to Nuclear Fusion

Data Science Challenges of the ITER International School 2024



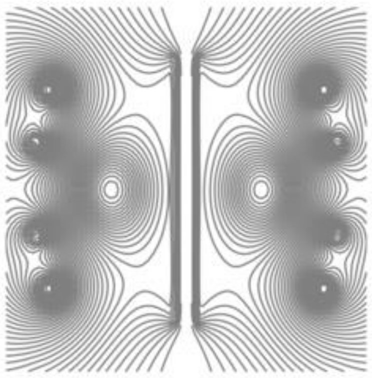
MAST Plasma Current

Infer plasma current produced by CCFE's Mega Ampere Spherical Tokamak from discrete magnetic diagnostic data.



MAST Plasma Volume

Infer the volume of plasmas produced by the CCFE's Mega Ampere Spherical Tokamak using frames from a wide-angle visible spectrum camera.



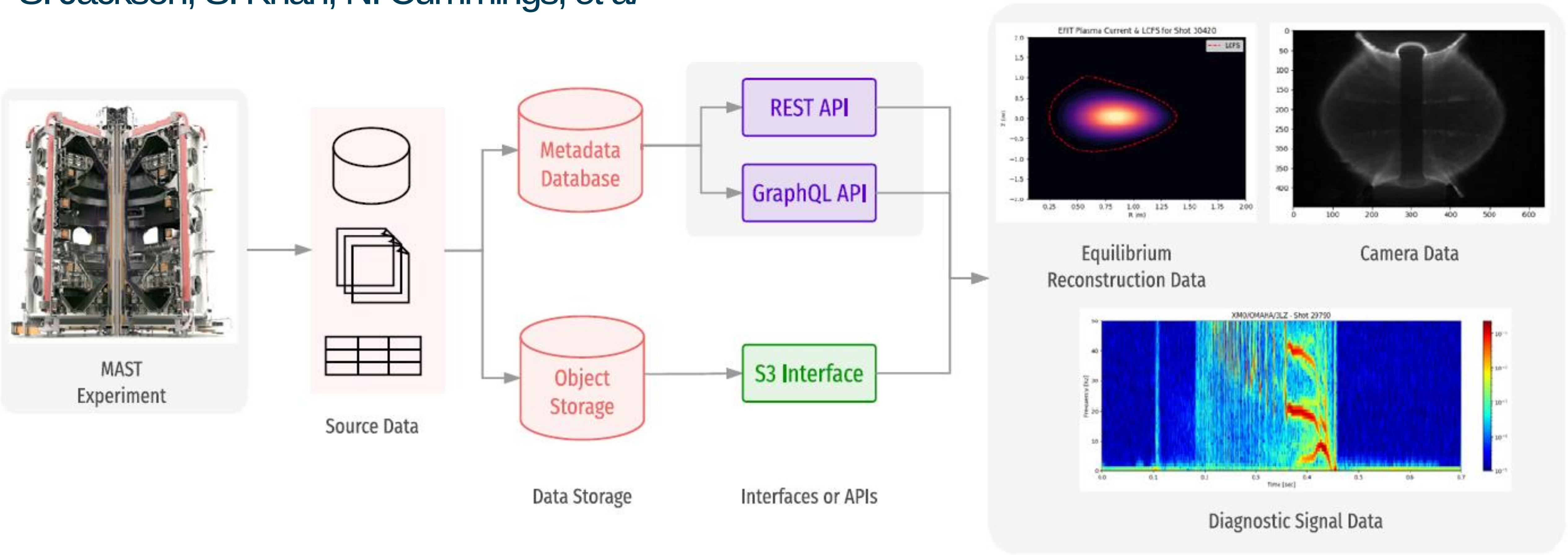
MAST Plasma Equilibrium

Infer two-dimensional poloidal flux maps produced by the EFIT++ equilibrium reconstruction code from a diverse set of diagnostic measurements.

The Data Science Challenges have been built on top of FAIR-MAST

A fusion device data management system.

S. Jackson, S. Khan, N. Cummings, *et al*



Pandata is a modern Python data-analytics stack.

Data storage	Data access	Data API	Data processing	Visualization	User interface
Parquet	fsspec	pandas	Numba	hvPlot	jupyter
Zarr	INTAKE	xarray	dask	Bokeh	Panel
DuckDB	kerchunk	RAPIDS	*Your domain-specific code*	matplotlib	jupyterhub
Legacy/ domain-specific formats		Polars	Asari	plotly	
HDF		CuPy	GeoCAT	Datashader	
FITS		Awkward Array	Xarray Spatial		
GRIB		GRAPHBLAS	MEG SPIKES		
netCDF			xdart		
STAC			Dask-ML		
COG			dask-image		
			Panel-Chemistry		
			raijin		
			Xradar		

Packaging

CONDA

You will find the following three packages useful for the challenges.

pandas

- Columnar data
- Very good indexing
- Suggested for writing submission.csv files

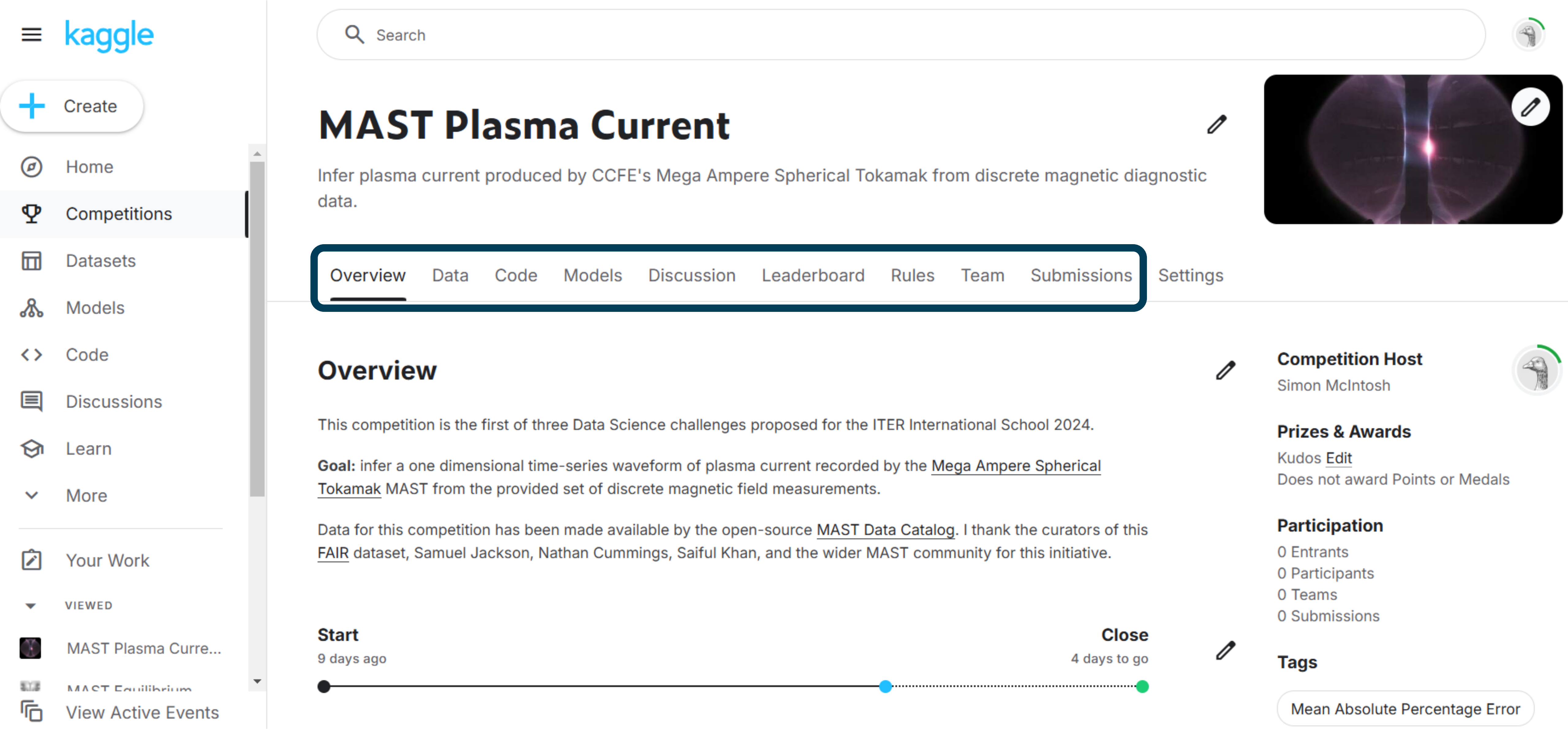
xarray

- Can open netCDF files
- Very good indexing
- Support for n-D arrays
- Supports labelled data

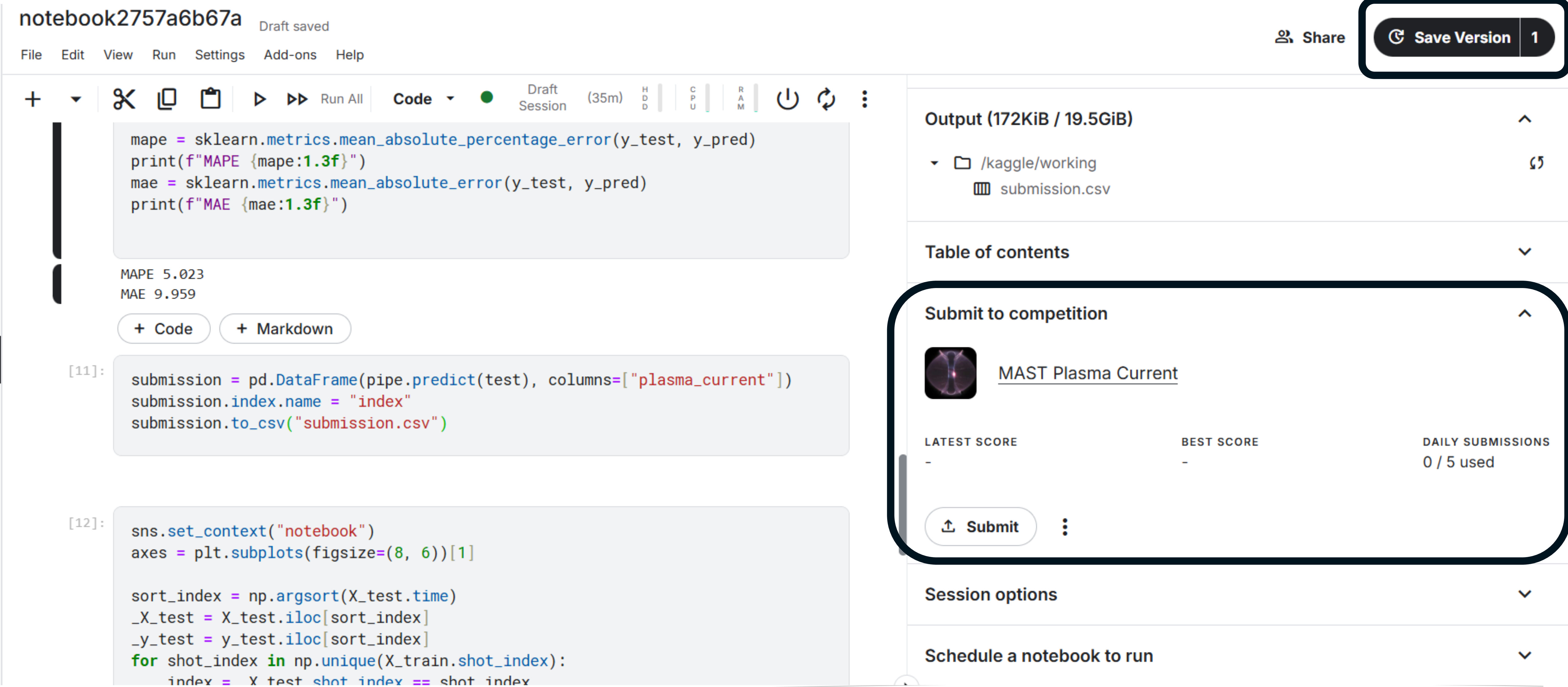
scikit learn

- Good entry ML library
- Fast learning curve
- Consistent API
- fit(X_train, y_train)
- predict(X_test)

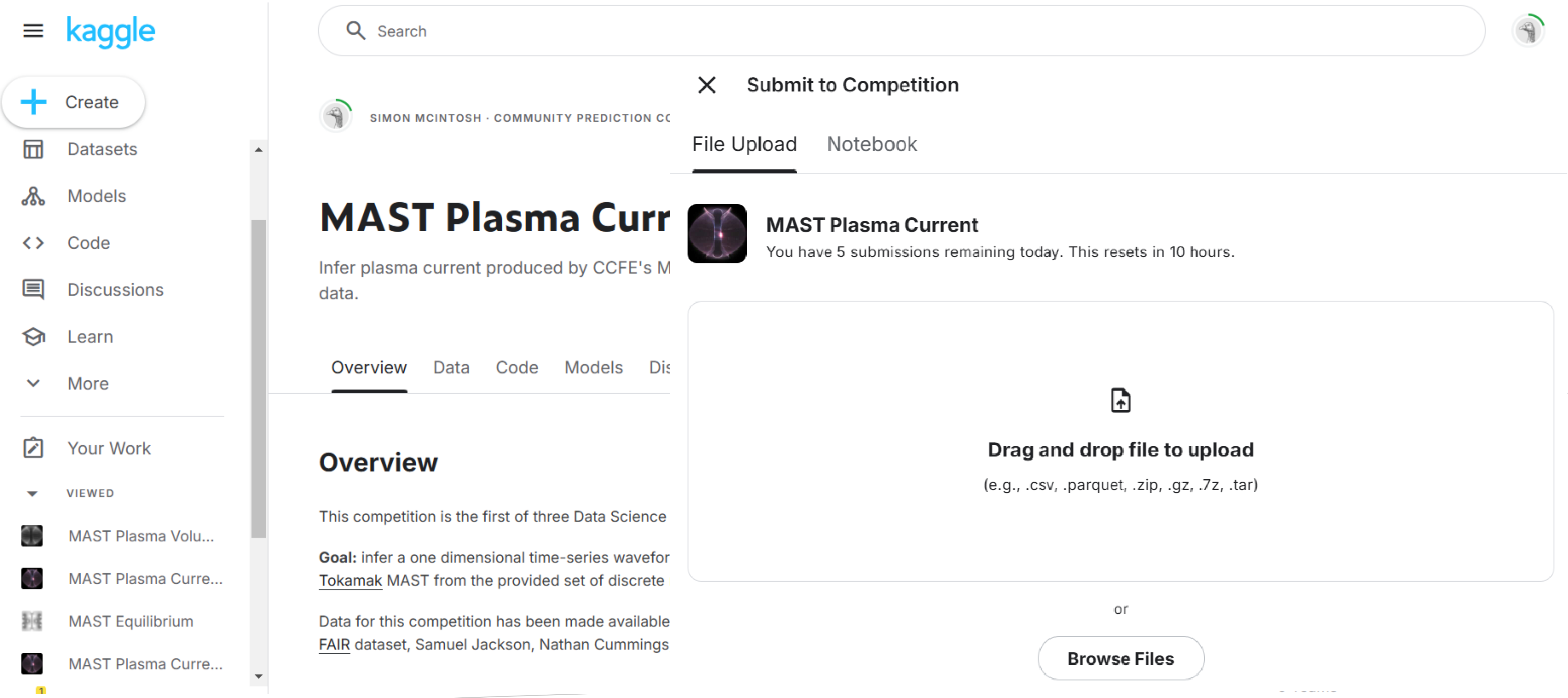
The Data Science Challenges will run on the Kaggle platform.



Accessing and submitting data from a Kaggle Notebook is simple.



Accessing and submitting data locally is also straight forward.



Each challenge uses a different evaluation metric.

See the Overview tab, Evaluation Section for further details.

Evaluation

Submissions are evaluated on Mean Absolute Percentage Error between the predicted and observed plasma current.

Submission File

For each `index` in the test set, you must predict a value for the `plasma_current` variable. The file should contain a header and have the following format:

```
index,plasma_current
0,-4.993814753222239
1,-2.9837154151294385
2,-5.1550966427939215
3,-4.030642466070503
4,-3.3313901825856647
5,-4.605478179129648
6,-4.566414377376589
etc.
```

The Data Science Challenges will close at 11pm this Thursday.

Submissions will be ranked using a private leaderboard.

Leaderboard

Raw Data | Refresh

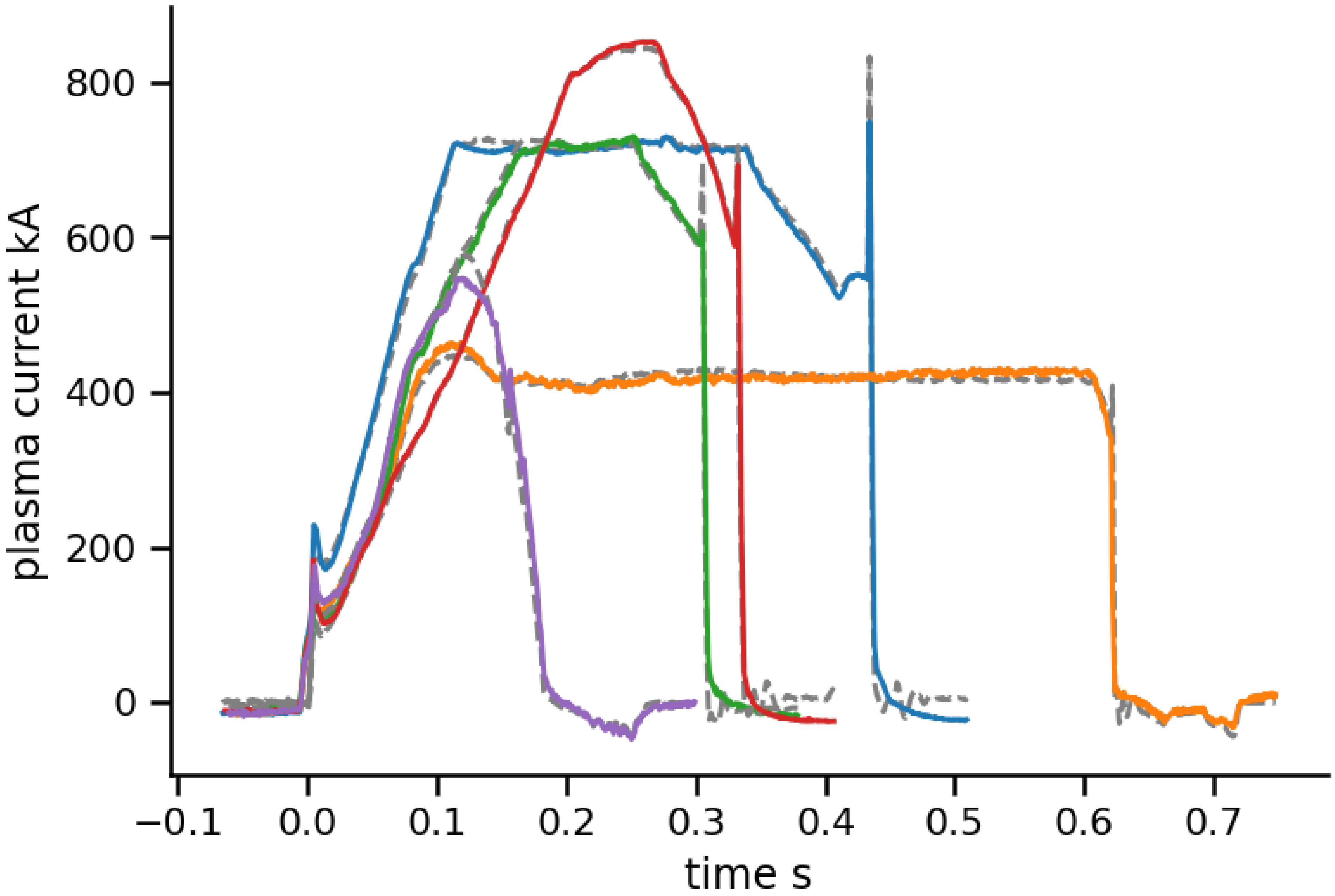
Search leaderboard

Public | Private

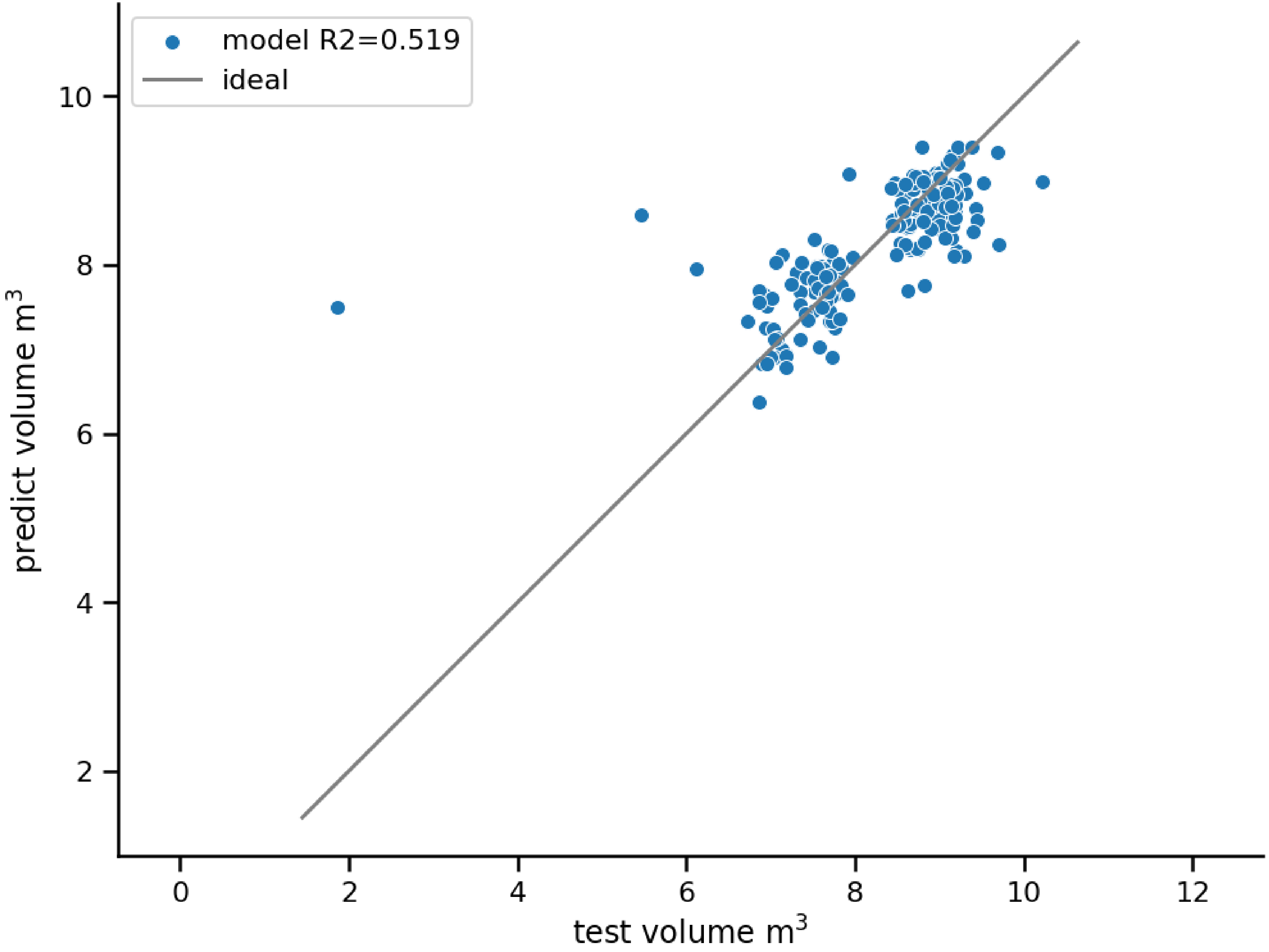
This leaderboard is calculated with approximately 44% of the test data. The final results will be based on the other 56%, so the final standings may be different.

#	Team	Members	Score	Entries	Last
	linear_regression.csv		3.77688		

Challenge #1 MAST Plasma Current

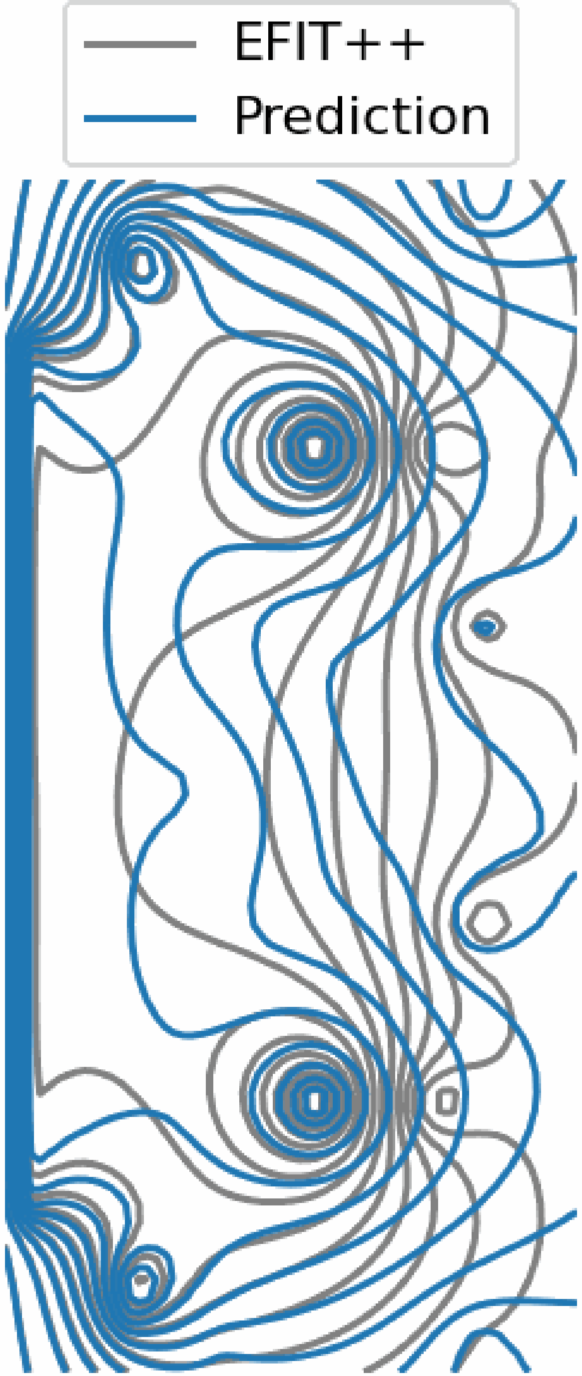


Challenge #2 MAST Plasma Volume



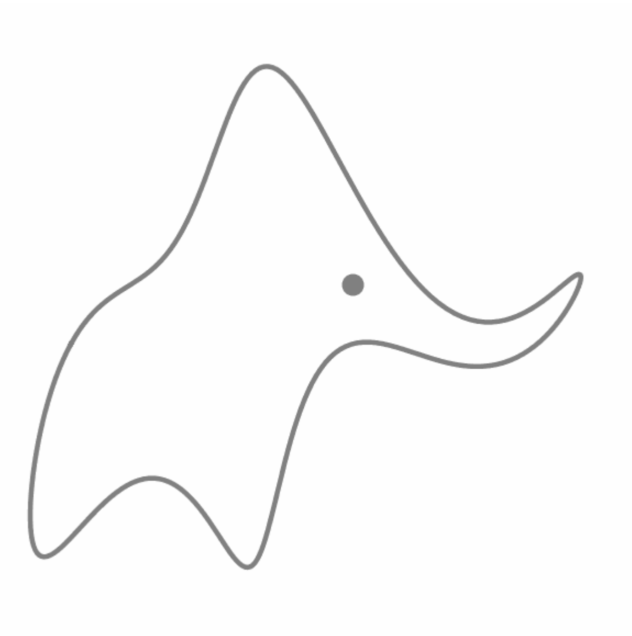
Challenge #3 MAST Plasma Equilibrium

```
Data variables:
    center_column            (center_column_channel, time) float64 16kB ...
    coil_currents            (coil_currents_channel, time) float64 19kB ...
    coil_voltages            (coil_voltages_channel, time) float64 13kB ...
    flux_loops               (flux_loops_channel, time) float64 19kB ...
    outer_discrete           (outer_discrete_channel, time) float64 26kB ...
    saddle_coils             (saddle_coils_channel, time) float64 13kB ...
    dalpha_mid_plane_center  (time) float64 3kB ...
    dalpha_mid_plane_wide    (time) float64 3kB ...
    dalpha_tangential        (time) float64 3kB ...
    hcam_l                   (hcam_l_channel, time) float64 58kB ...
    hcam_u                   (hcam_u_channel, time) float64 58kB ...
    ne                       (time, major_radius) float64 210kB ...
    ne_core                  (time) float64 3kB ...
    pe                       (time, major_radius) float64 210kB ...
    te                       (time, major_radius) float64 210kB ...
    te_core                  (time) float64 3kB ...
    shot_index               (time) float64 3kB ...
    magnetic_flux            (time, z, major_radius) float64 14MB ...
    tcam                     (tcam_channel, time) float64 58kB ...
```

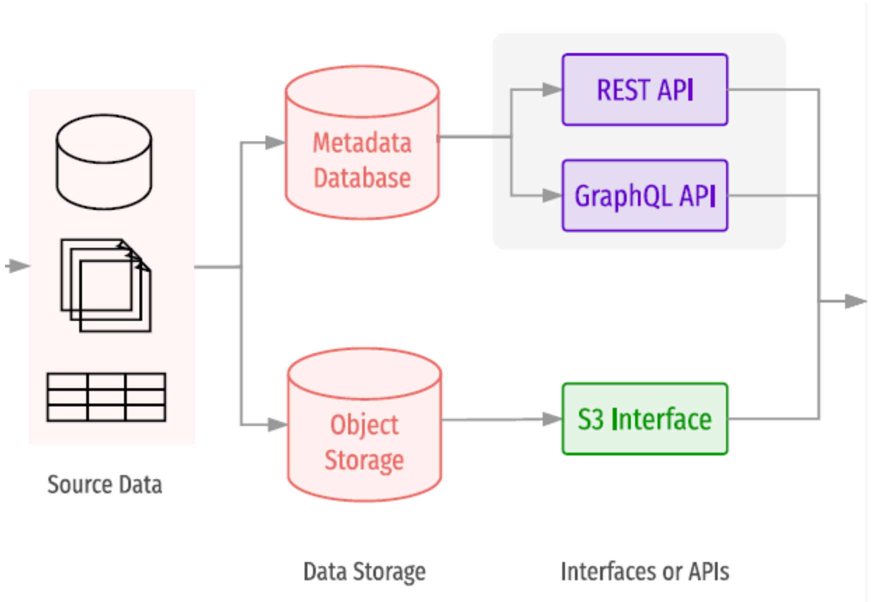


In summary this talk has warned you of the dangers of overfitting and has given you the opportunity to learn more via the challenges.

Remember the elephant.



Data Science Challenge facilitated by FAIR data and open-source tools.



Doing is often the best way to learn. Good luck!