

Digital Signal Processing & Data Science Challenge

Simon McIntosh

ITER Organization

Monday 9th December 2024

ITER International School

Nagoya, Japan

Bio:

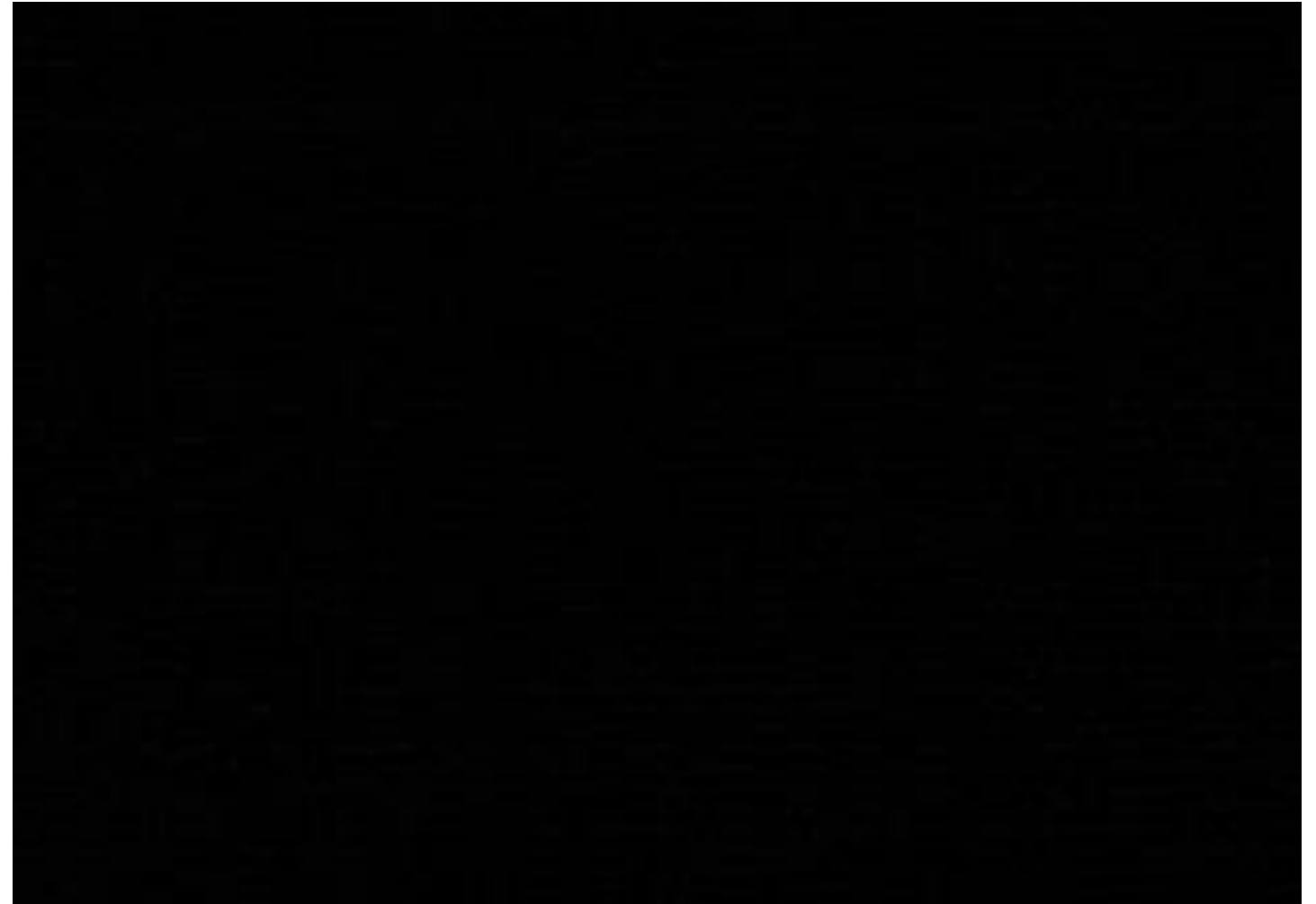
Aerospace Engineering University of Bristol

PhD Fluid Mechanics University of Cambridge

Postdoc / Lecturer University of Oxford

Staff Culham Center for Fusion Energy

Staff ITER Organization Scientific Data Processing



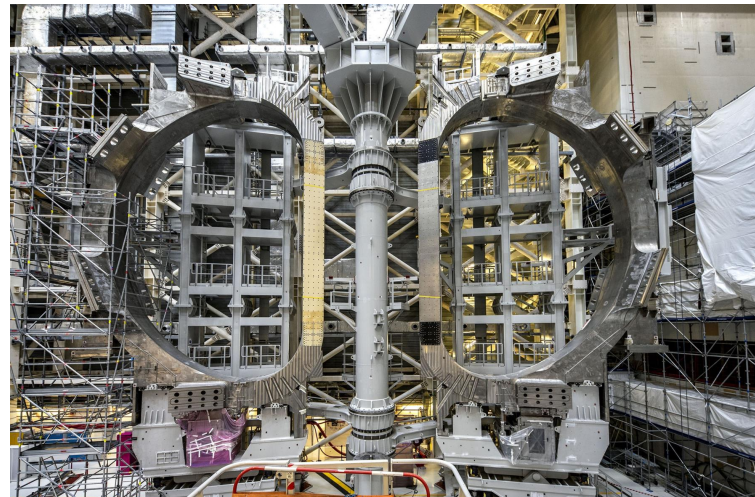
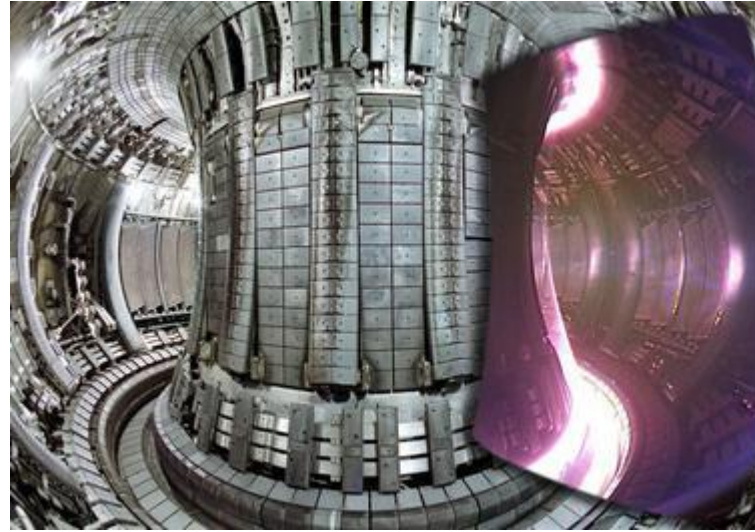
china eu india japan korea russia usa

This presentation focuses on applied Data Science for Fusion.

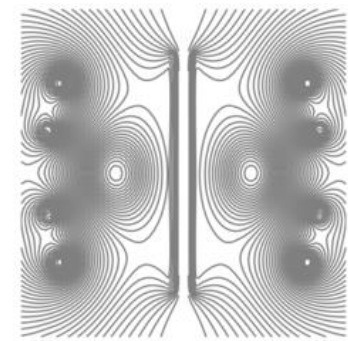
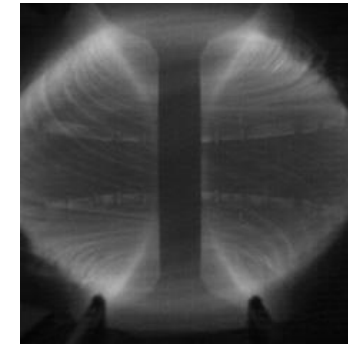
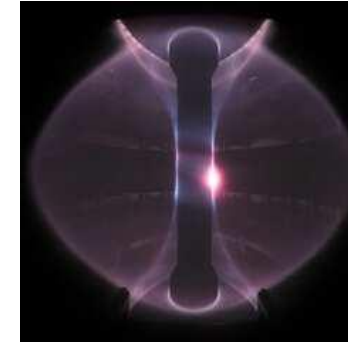
The elephant in the room.



Applied Data Science in Fusion.



Kaggle Challenges.



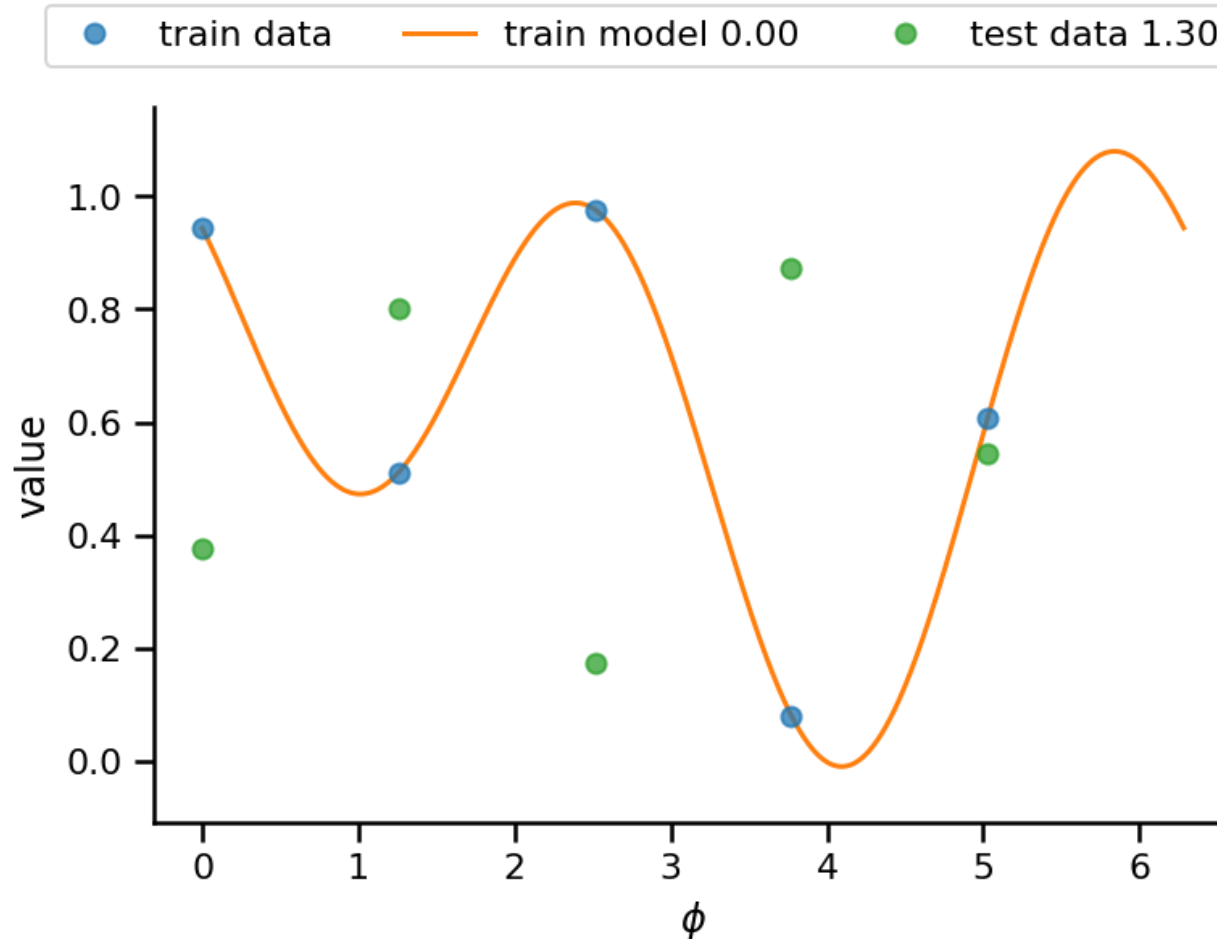
The elephant in the room is over-fitting.



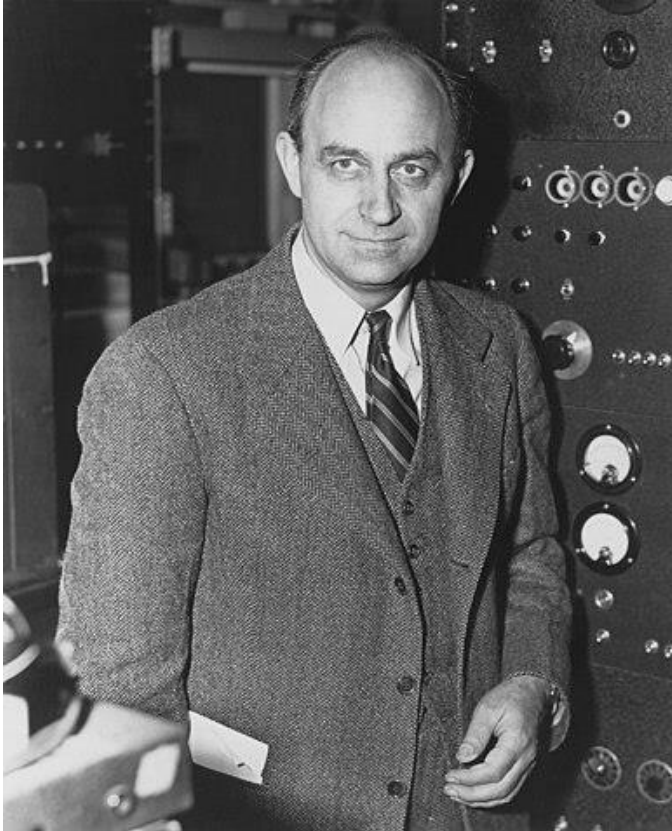
“All models are wrong, but some are useful”. George Box

Non-linear parametrizations can fit arbitrarily complexity structures.

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



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



Enrico Fermi

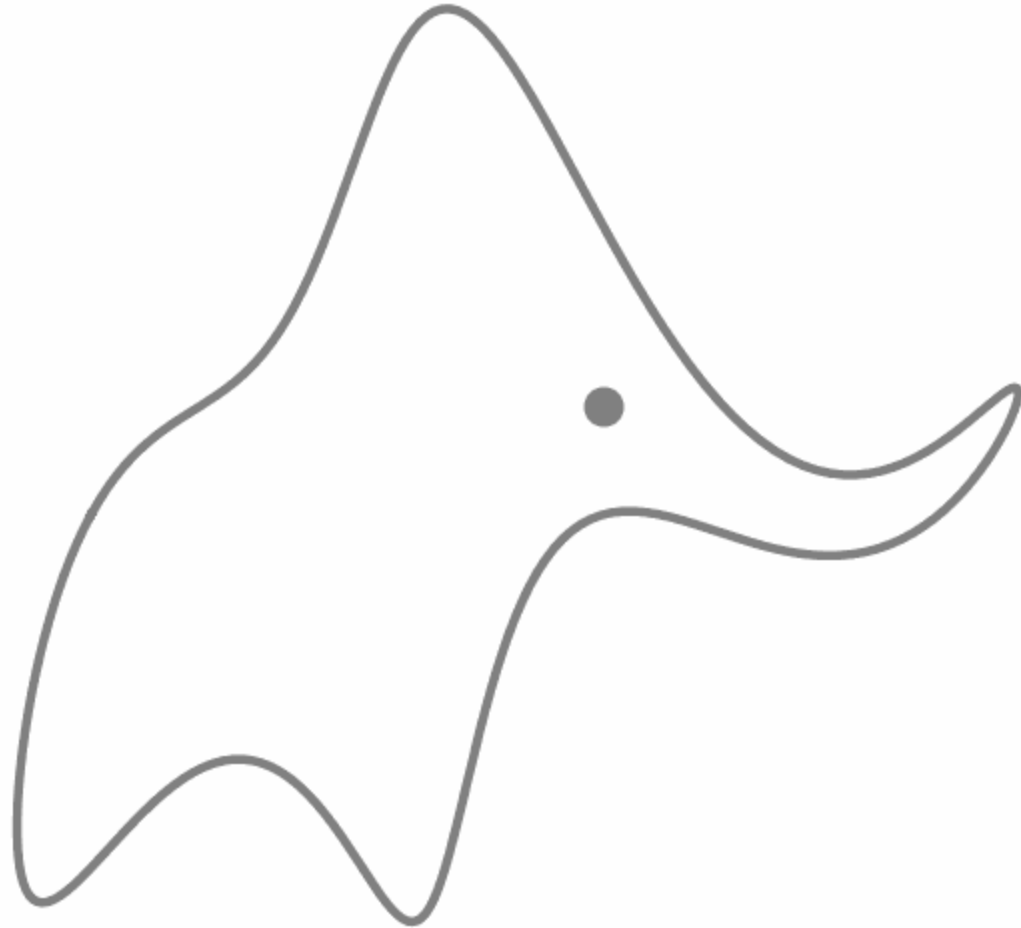


John von Neumann

Simple basis functions can generate complex shapes.

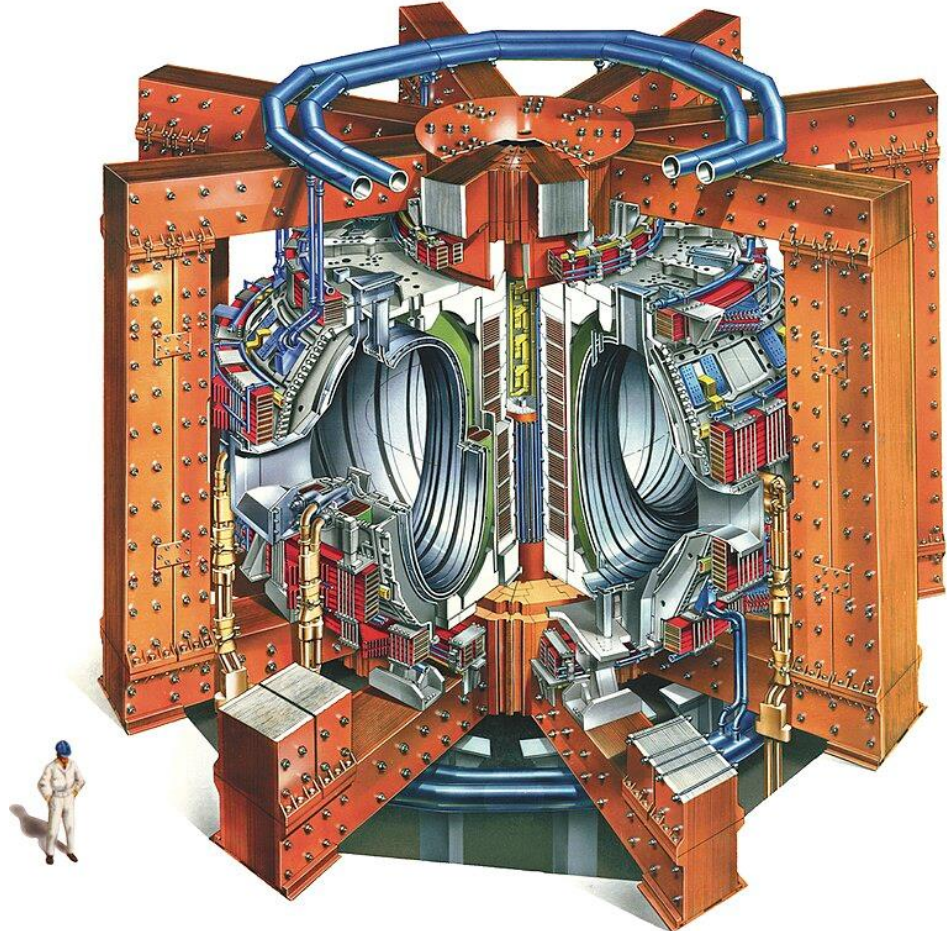
von Neumann's Elephant

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

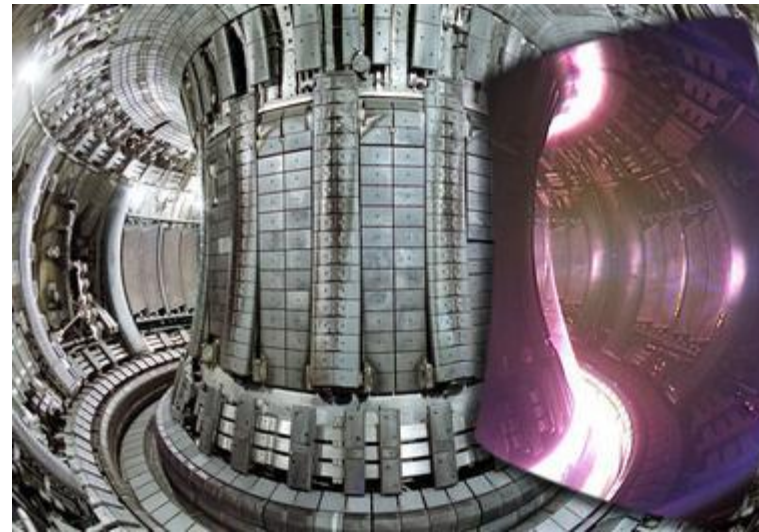


$$C_n = \begin{bmatrix} 0 + 0j \\ P_n[0] \\ P_n[1] \\ 0 + 0j \\ 0 + 0j \\ P_n[2] \\ P_n[2] \\ 0 + 0j \\ 0 + P_n[3]j \\ -P_n[1] \\ P_n[3] - P_n[0]j \end{bmatrix}$$

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 IDs.

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 occurred 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.

📁 Repositories

🔍 Find a repository...

Type ▾

Language ▾

Sort ▾

📁 New

imas-validator Private

● Python ☆ 0 🍴 0 🔄 0 📄 0 Updated 2 days ago



imas-python Private

Python high-level interface of the IMAS Access Layer -- A pure-python library to handle arbitrarily nested data structures, including IDSs

● Python ☆ 0 🍴 0 🔄 0 📄 0 Updated 3 weeks ago



imas-core Private

Lowlevel interface of the IMAS Access Layer

● C++ ☆ 0 🍴 1 🔄 0 📄 0 Updated 3 weeks ago



imas-data-dictionary Private

The Data Dictionary is the implementation of the Data Model of ITER's Integrated Modelling & Analysis Suite (IMAS).

● Python ☆ 0 🍴 3 🔄 0 📄 1 Updated on Aug 5



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

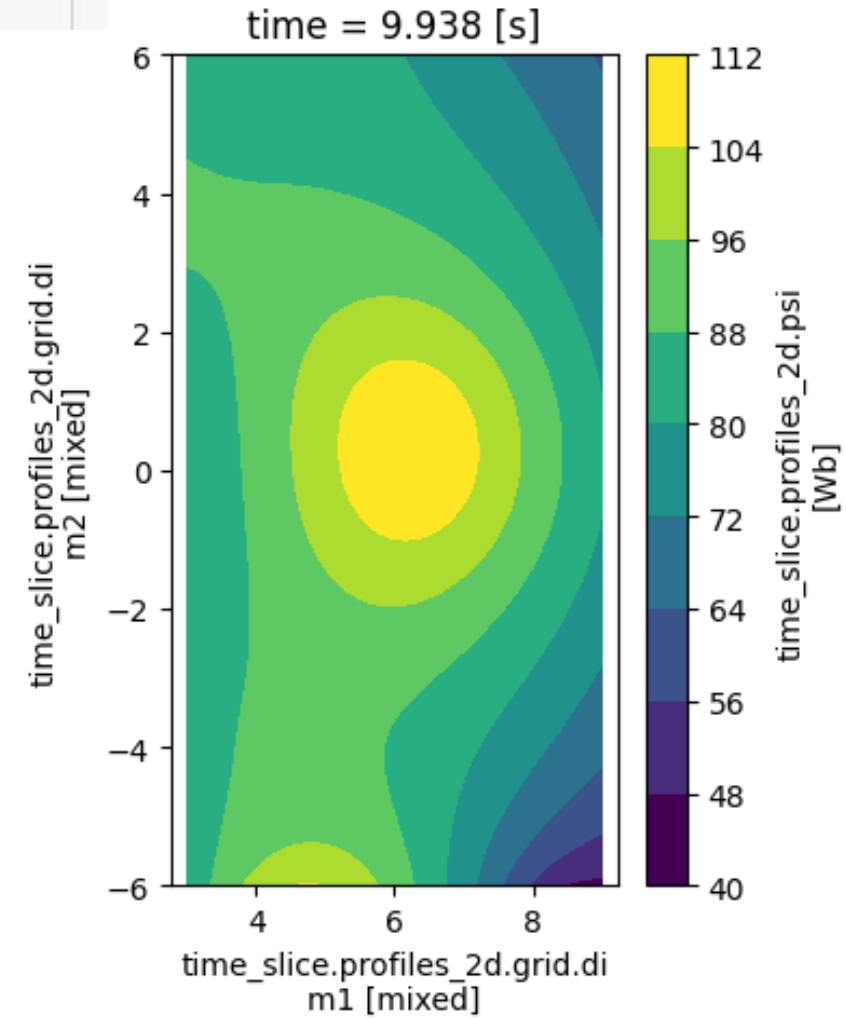
1. `time` with coordinate `time`
2. `time_slice.profiles_2d`, which is an index
3. `time_slice.profiles_2d.grid.dim1`
4. `time_slice.profiles_2d.grid.dim2`

- 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

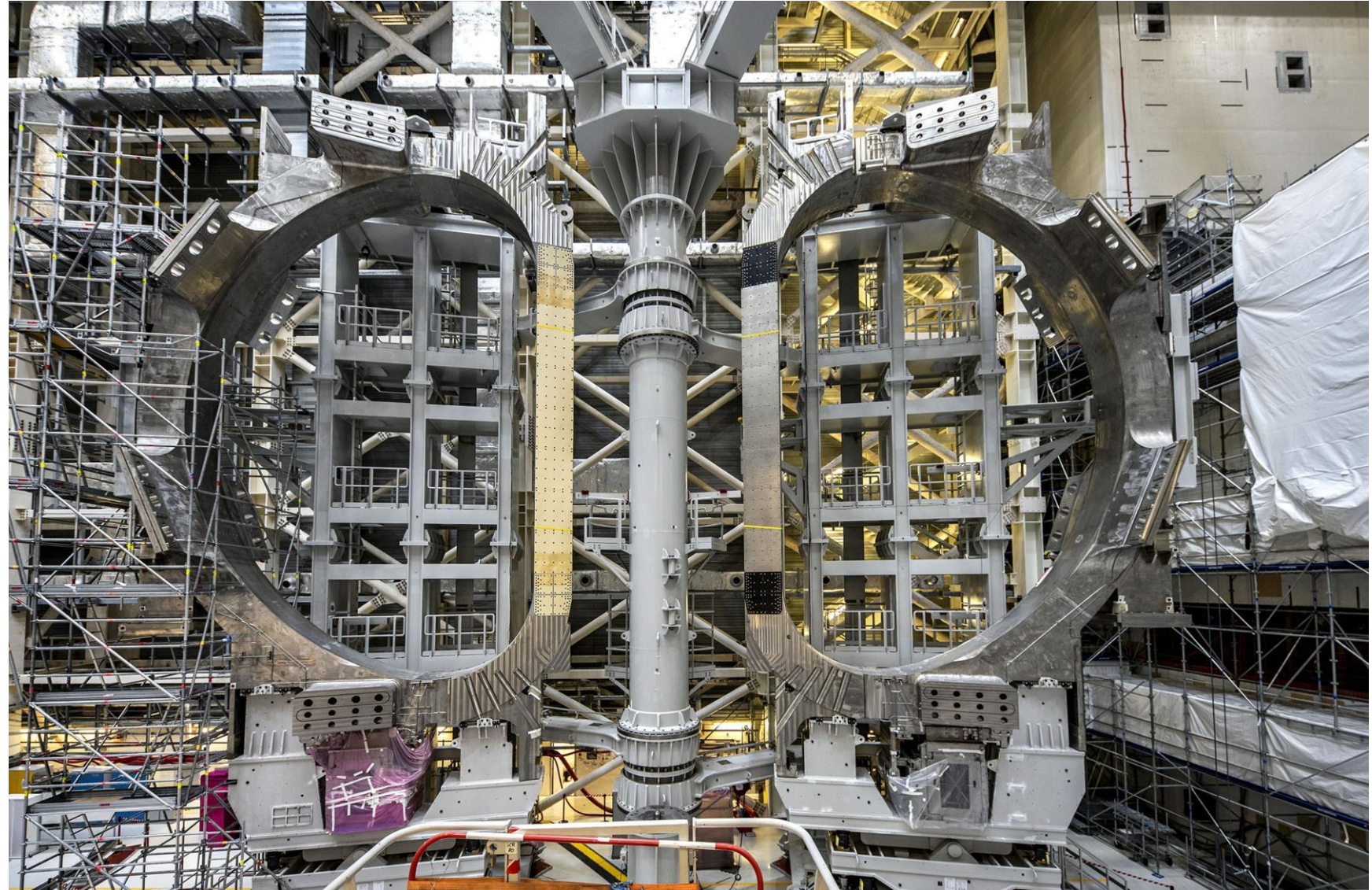
Self-describing IMAS data without a custom Access Layer (xarray).

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



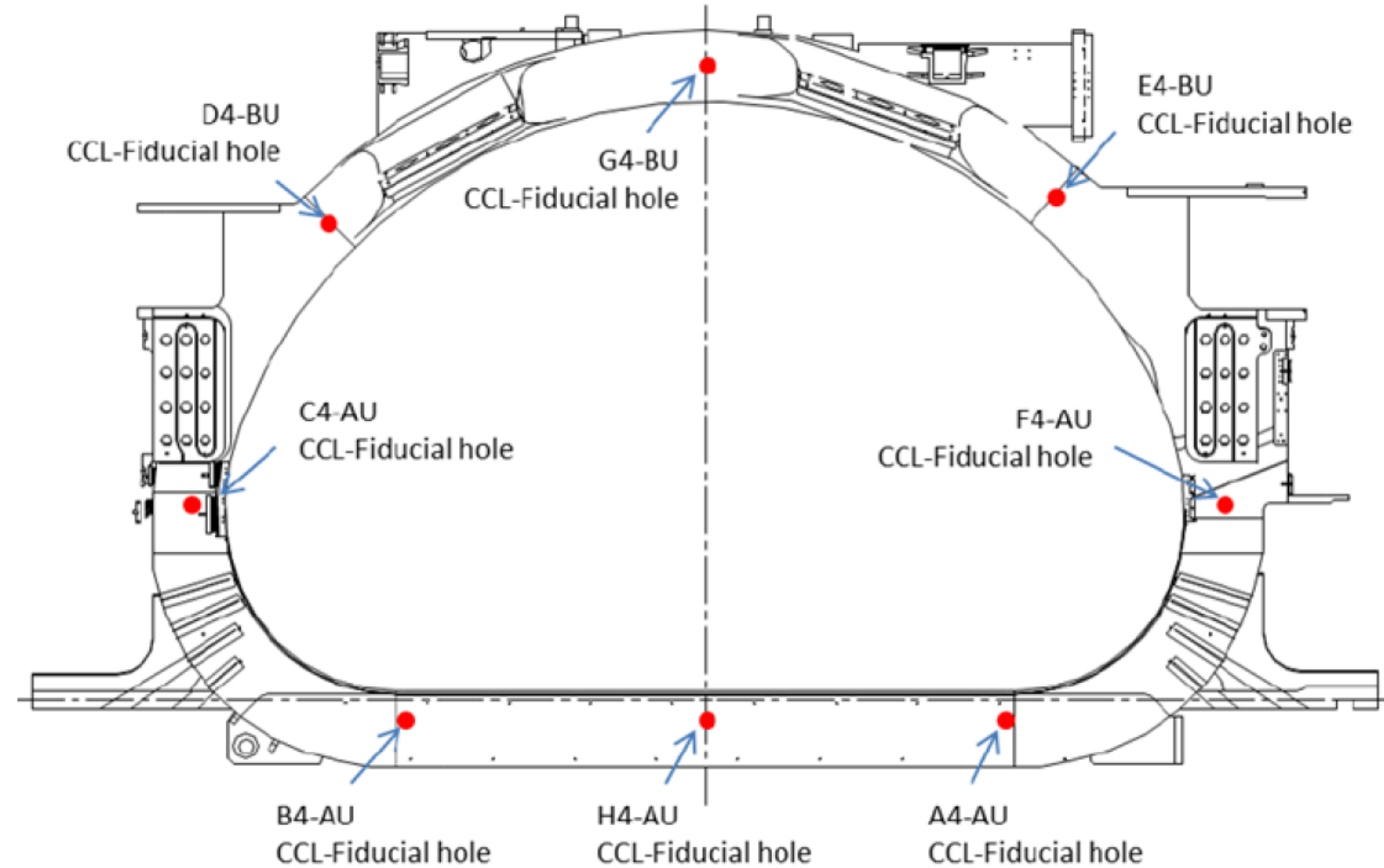
The alignment of ITER's 17 meter high, 360 tonne D-shaped Toroidal Field magnets is a feat of precision engineering.

Exceptionally low tolerances that are repeatable and stable over time.



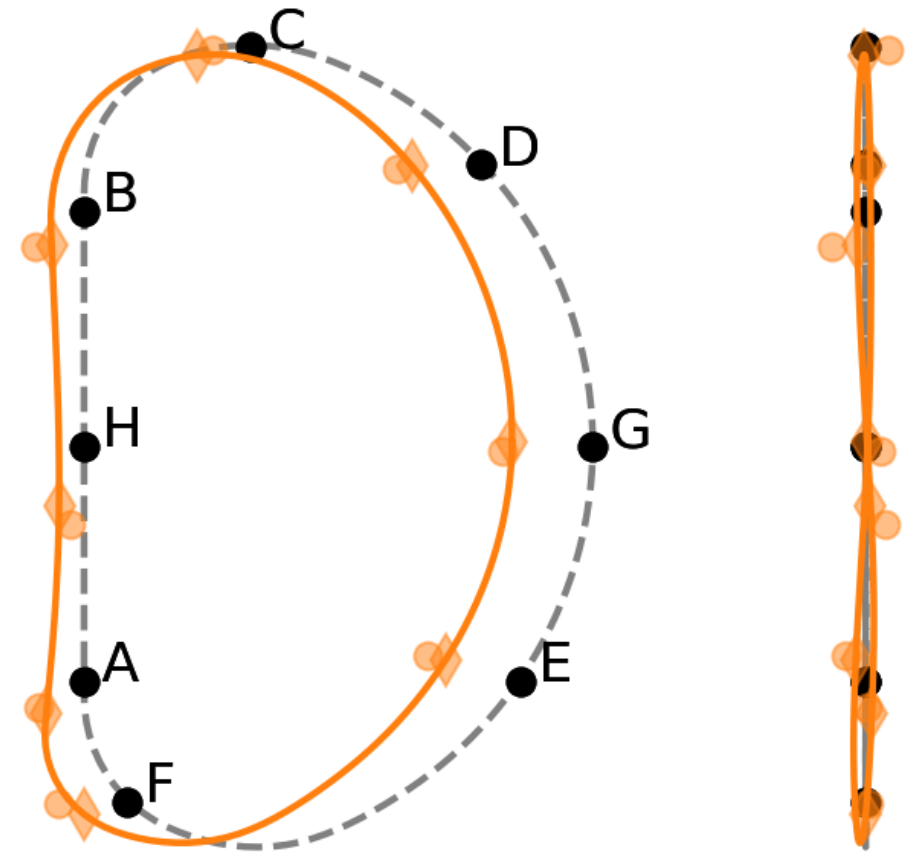
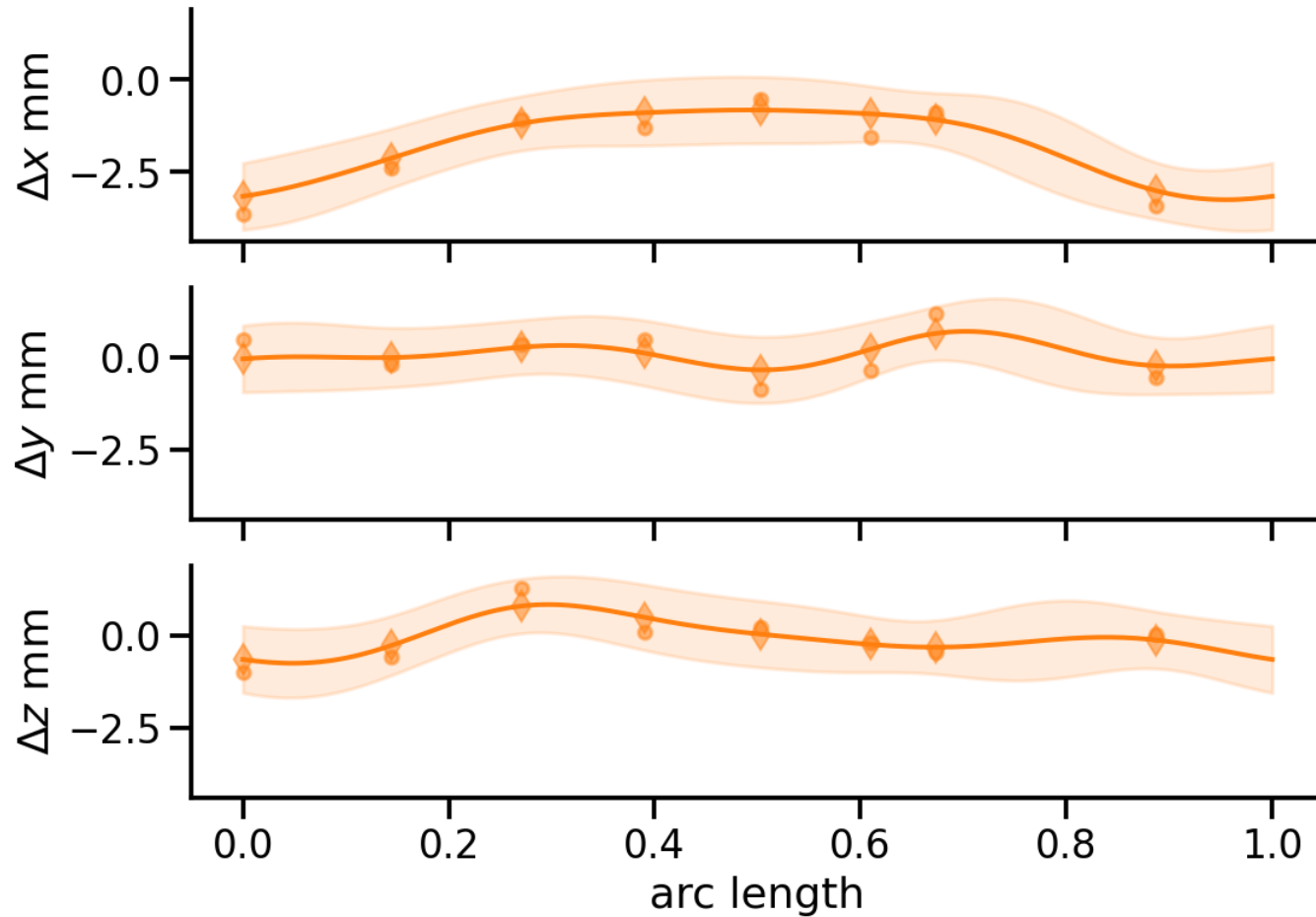
Data Science in Fusion is not restricted to the Physics domain.

Gaussian Process Regression used on ITER to reconstruct coil centerline.



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

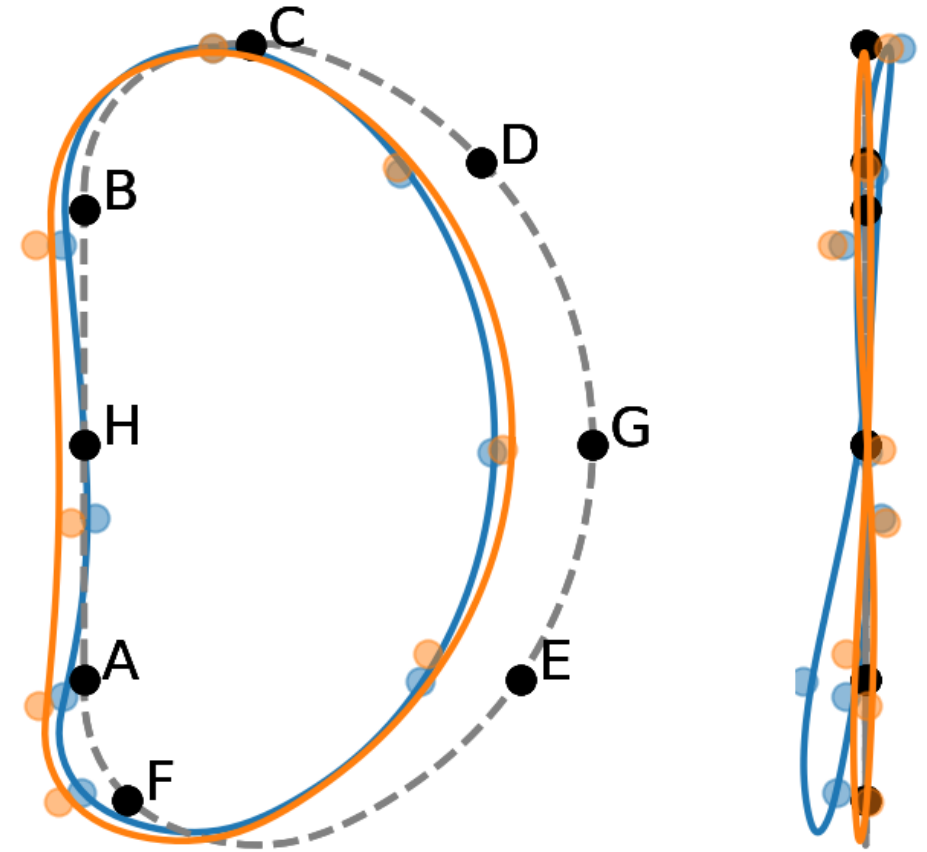
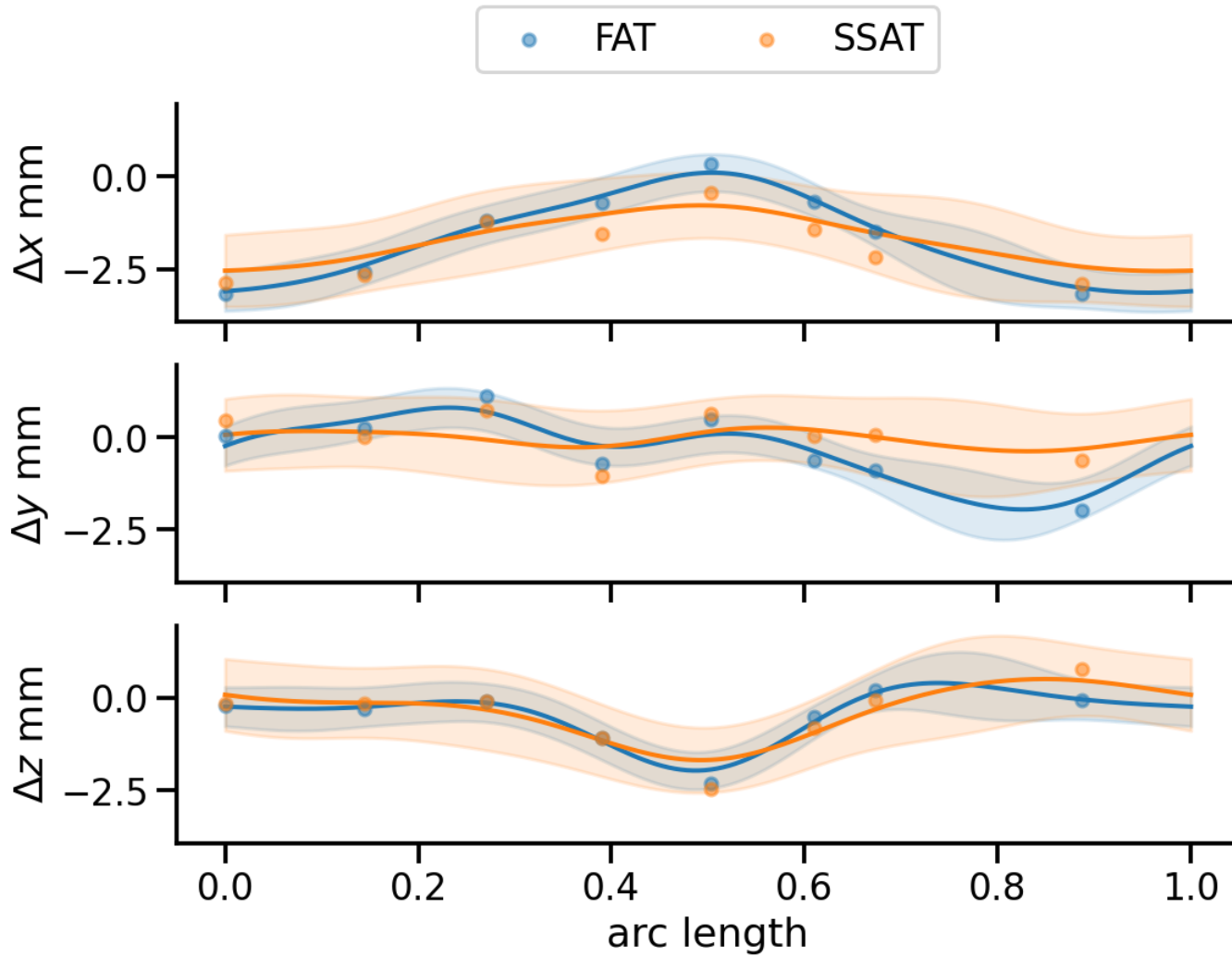
SSAT 95% confidence



Coil alignment is based on inferred CCL positions (diamonds)

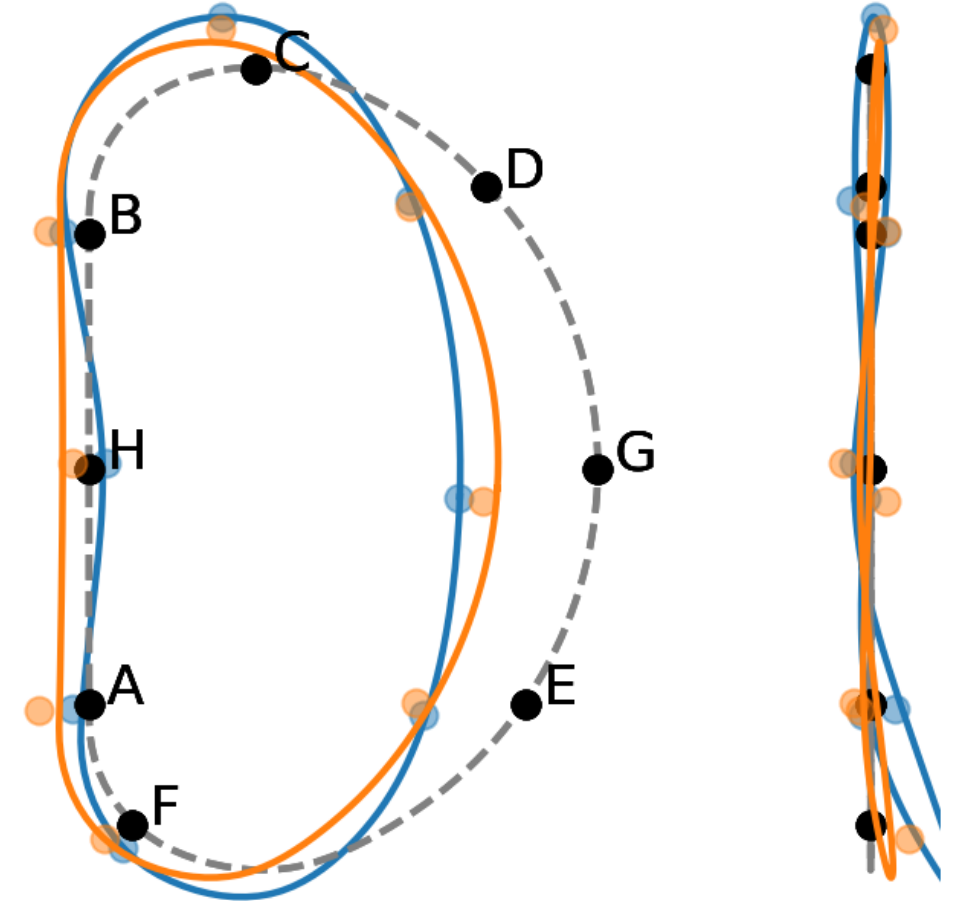
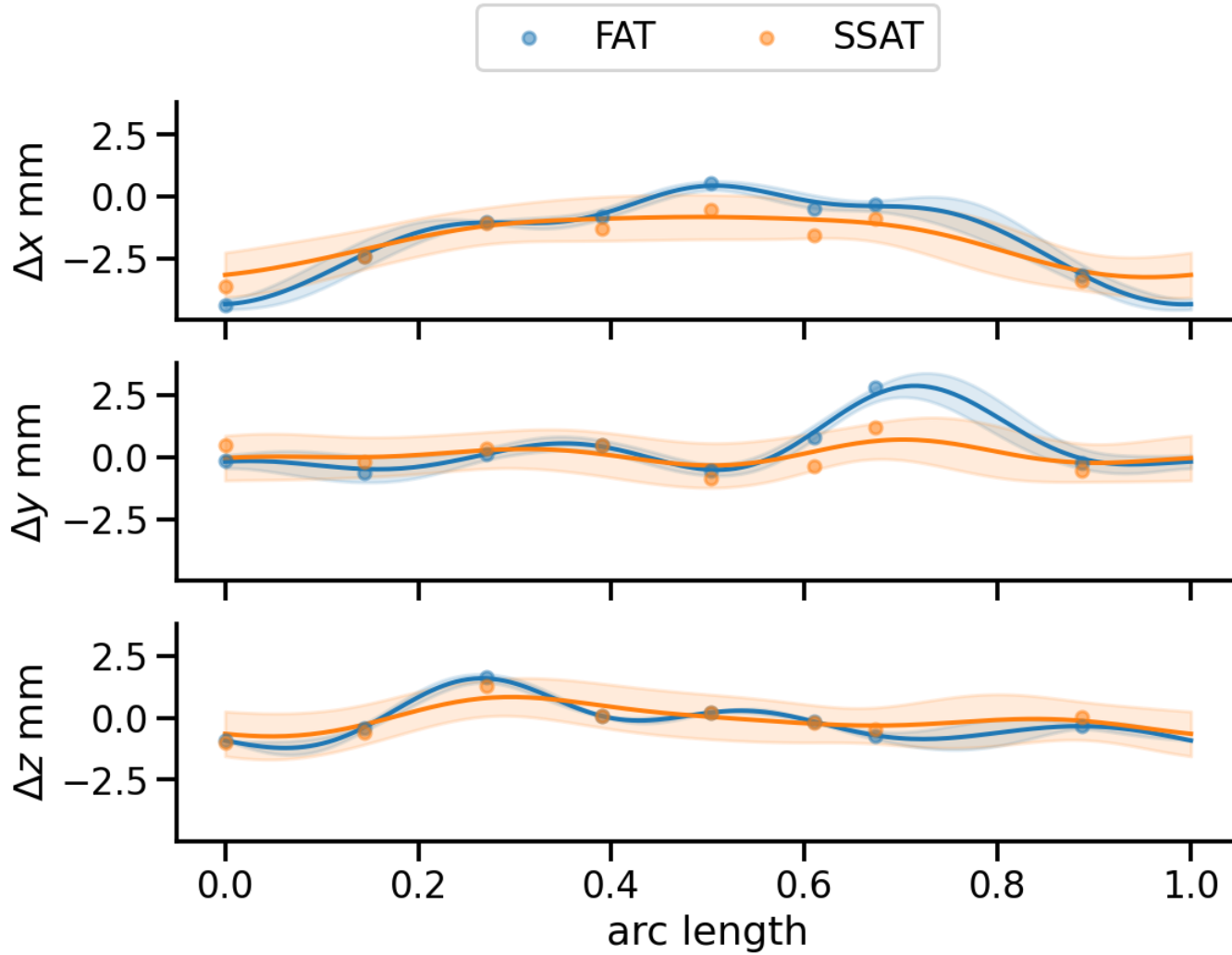
Deformation x500

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



Deformation x500

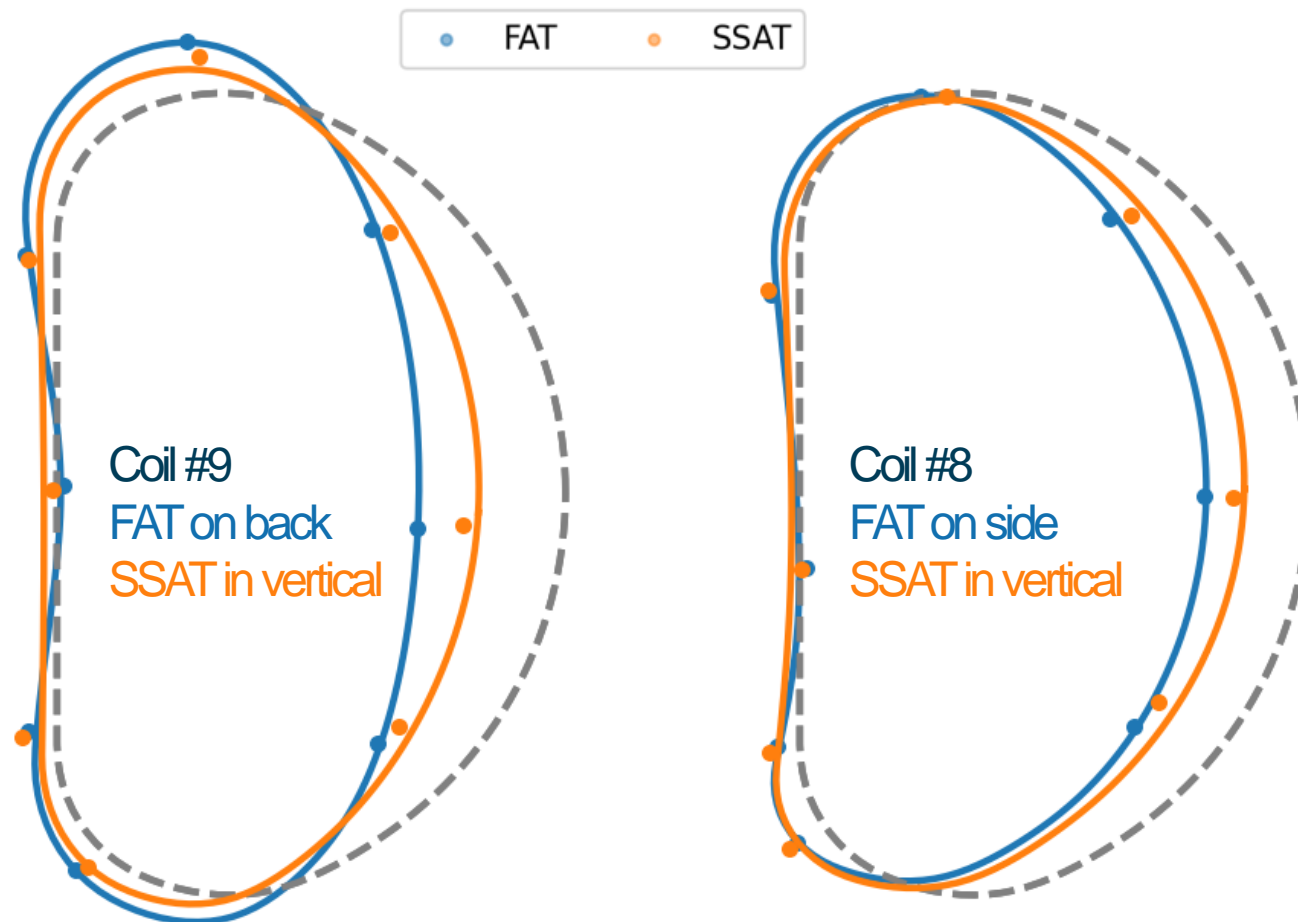
The orientation of each TF Coil affects its shape (Coil #9 EU)



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

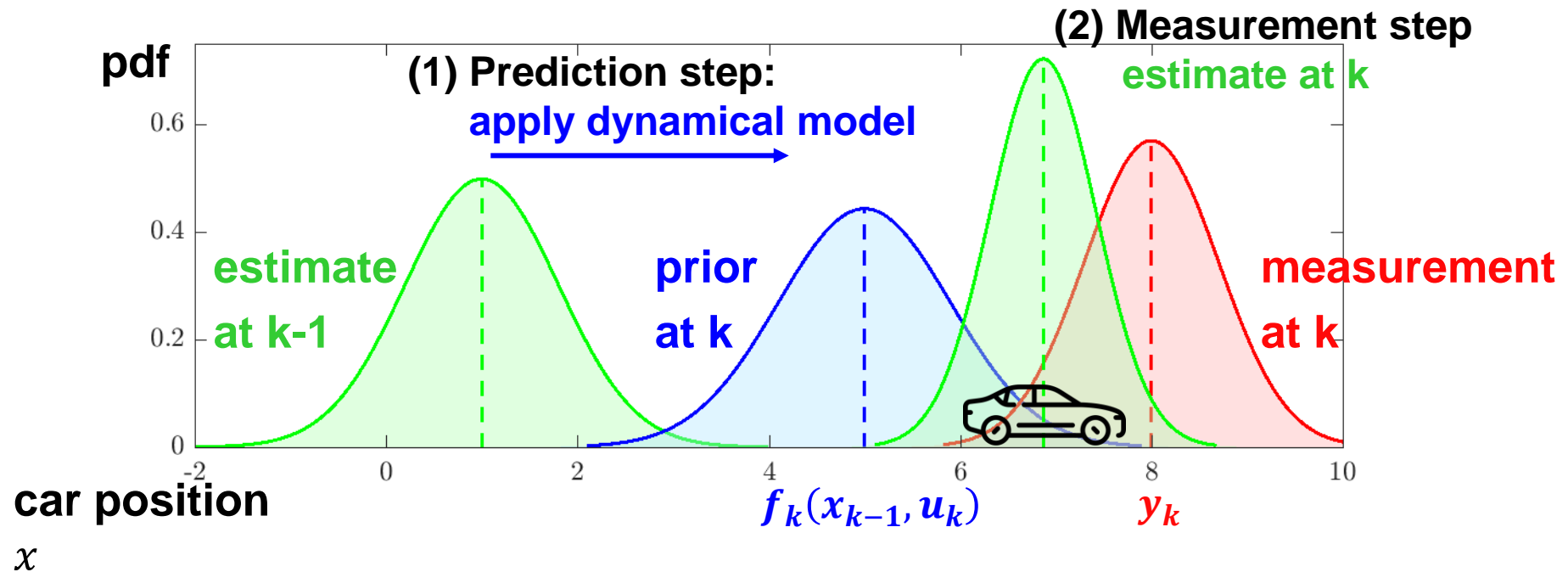
The Extended Kalman Filter: a sequential Bayesian filter

Example



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

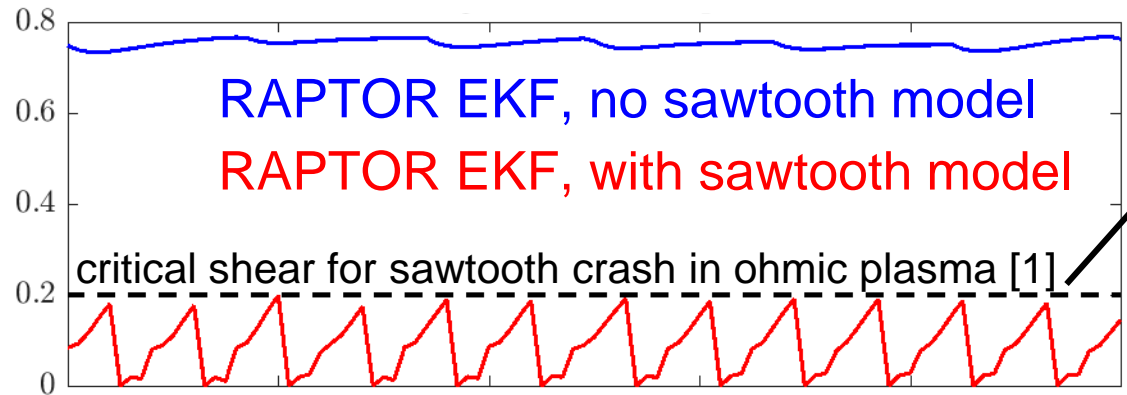
and a series of **measurements** y_k , affected by noise v_k with **measurement covariance** $R_k = E[v_k v_k^T]$



Sawtooth model allows realistic inter-measurement prediction

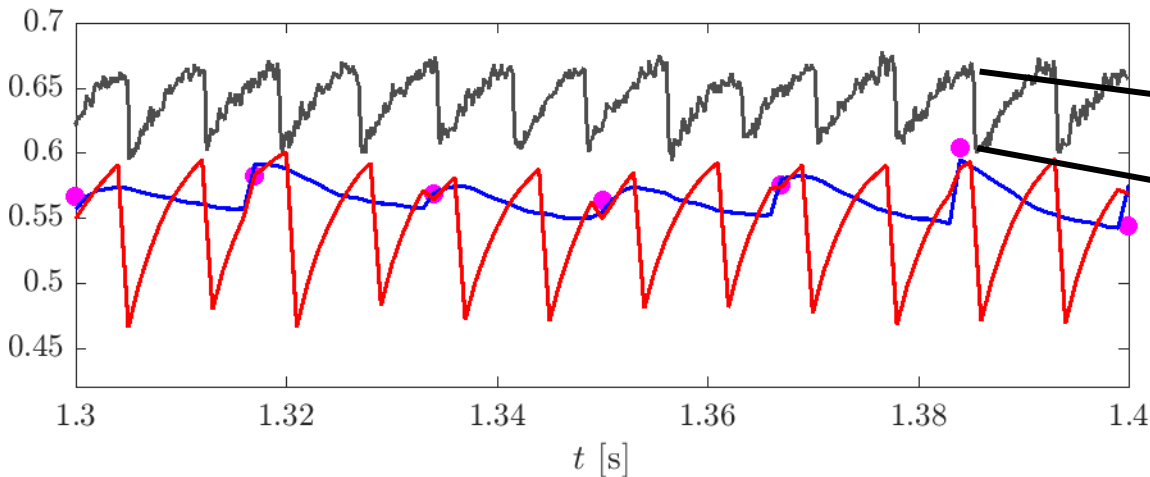
TCV 64965
(ohmic)

magnetic shear
at $q = 1$



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

central T_e [keV]



soft X ray T_{e0}

Thomson
scattering

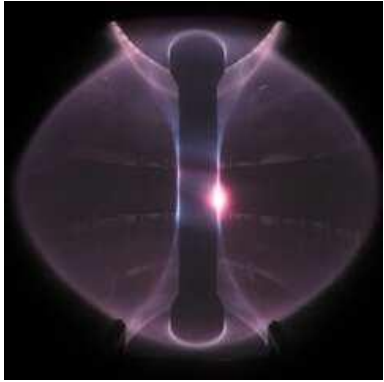
TCV

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)

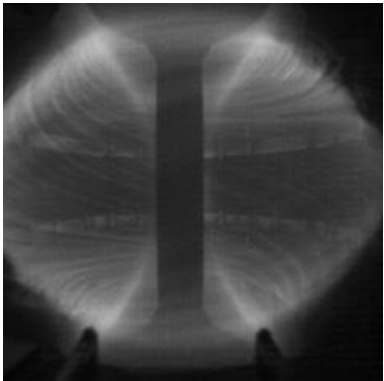
from S. Van Mulders et al, to be subm. to Nuclear Fusion

Data Science Challenges of the ITER International School 2024



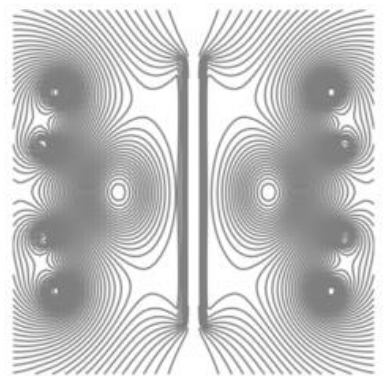
MAST Plasma Current

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

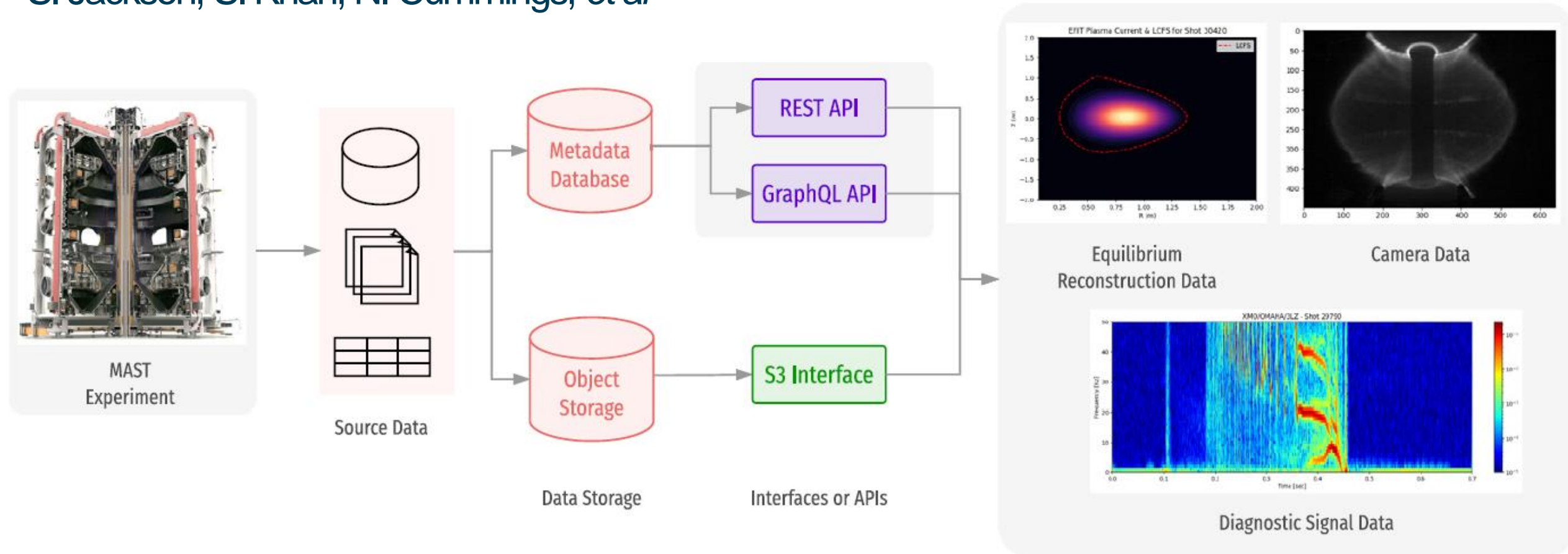
Infer two-dimensional poloidal flux maps produced by the EFIT++ equilibrium reconstruction code from a diverse set of diagnostic measurements.



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



Parquet



Zarr

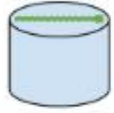


DuckDB

Legacy/
domain-
specific
formats



Data access



fsspec



INTAKE



kerchunk

Data API



pandas



xarray

RAPIDS



Polars



CuPy

Awkward
Array



GRAPHBLAS

Data processing

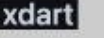
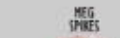
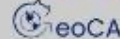


Numba



dask

Your domain-
specific code



Visualization



hvPlot



Bokeh

matplotlib



plotly



Datashader

User interface



jupyter



Panel



jupyterhub

Packaging

CONDA

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



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



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



- 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.

☰ **kaggle**

+ Create

🏠 Home

🏆 Competitions

📁 Datasets

🤖 Models

📄 Code

💬 Discussions

📖 Learn

∨ More

📌 Your Work

∨ VIEWED

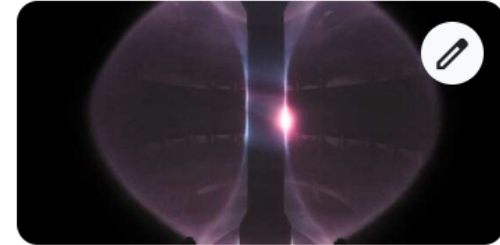
🖼️ MAST Plasma Curre...

🖼️ MAST Equilibrium

📄 View Active Events



MAST Plasma Current



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

- Overview**
- Data
- Code
- Models
- Discussion
- Leaderboard
- Rules
- Team
- Submissions
- Settings

Overview



This competition is the first of three Data Science challenges proposed for the ITER International School 2024.

Goal: infer a one dimensional time-series waveform of plasma current recorded by the Mega Ampere Spherical Tokamak MAST from the provided set of discrete magnetic field measurements.

Data for this competition has been made available by the open-source MAST Data Catalog. I thank the curators of this FAIR dataset, Samuel Jackson, Nathan Cummings, Saiful Khan, and the wider MAST community for this initiative.

Start

9 days ago

Close

4 days to go



Competition Host



Simon McIntosh

Prizes & Awards

Kudos [Edit](#)

Does not award Points or Medals

Participation

0 Entrants

0 Participants

0 Teams

0 Submissions

Tags

Mean Absolute Percentage Error

Accessing and submitting data from a Kaggle Notebook is simple.

notebook2757a6b67a Draft saved Share Save Version 1

File Edit View Run Settings Add-ons Help

+ [Icons] Run All Code (35m) H D D C P U R A M [Icons]

```
mape = sklearn.metrics.mean_absolute_percentage_error(y_test, y_pred)
print(f"MAPE {mape:1.3f}")
mae = sklearn.metrics.mean_absolute_error(y_test, y_pred)
print(f"MAE {mae:1.3f}")
```

MAPE 5.023
MAE 9.959

+ Code + Markdown

```
[11]: submission = pd.DataFrame(pipe.predict(test), columns=["plasma_current"])
      submission.index.name = "index"
      submission.to_csv("submission.csv")
```

```
[12]: sns.set_context("notebook")
      axes = plt.subplots(figsize=(8, 6))[1]


      sort_index = np.argsort(X_test.time)
      _X_test = X_test.iloc[sort_index]
      _y_test = y_test.iloc[sort_index]
      for shot_index in np.unique(X_train.shot_index):
          index = X_test.shot_index == shot_index
```

Output (172KiB / 19.5GiB)

- /kaggle/working
 - submission.csv

Table of contents

Submit to competition

 MAST Plasma Current

LATEST SCORE	BEST SCORE	DAILY SUBMISSIONS
-	-	0 / 5 used

[Submit](#) [More options]

Session options

Schedule a notebook to run

Accessing and submitting data locally is also straight forward.

The screenshot shows the Kaggle website interface. On the left is a navigation sidebar with options: Create, Datasets, Models, Code, Discussions, Learn, More, Your Work, and a list of viewed items including 'MAST Plasma Volu...', 'MAST Plasma Curre...', 'MAST Equilibrium', and 'MAST Plasma Curre...'. The main content area is titled 'Submit to Competition' and features a search bar at the top. Below the search bar, there's a header for 'SIMON MCINTOSH · COMMUNITY PREDICTION CC' with options for 'File Upload' and 'Notebook'. The competition title 'MAST Plasma Current' is prominently displayed, along with a description: 'Infer plasma current produced by CCFE's M data.' Below the title are tabs for 'Overview', 'Data', 'Code', 'Models', and 'Dis'. The 'Overview' tab is active, showing a goal: 'infer a one dimensional time-series waveform Tokamak MAST from the provided set of discrete' and a note that data is available via FAIR dataset from Samuel Jackson and Nathan Cummings. A large central box contains a file upload icon and the text 'Drag and drop file to upload (e.g., .csv, .parquet, .zip, .gz, .7z, .tar)'. Below this box is the word 'or' and a 'Browse Files' button. A user profile icon is visible in the top right corner.

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

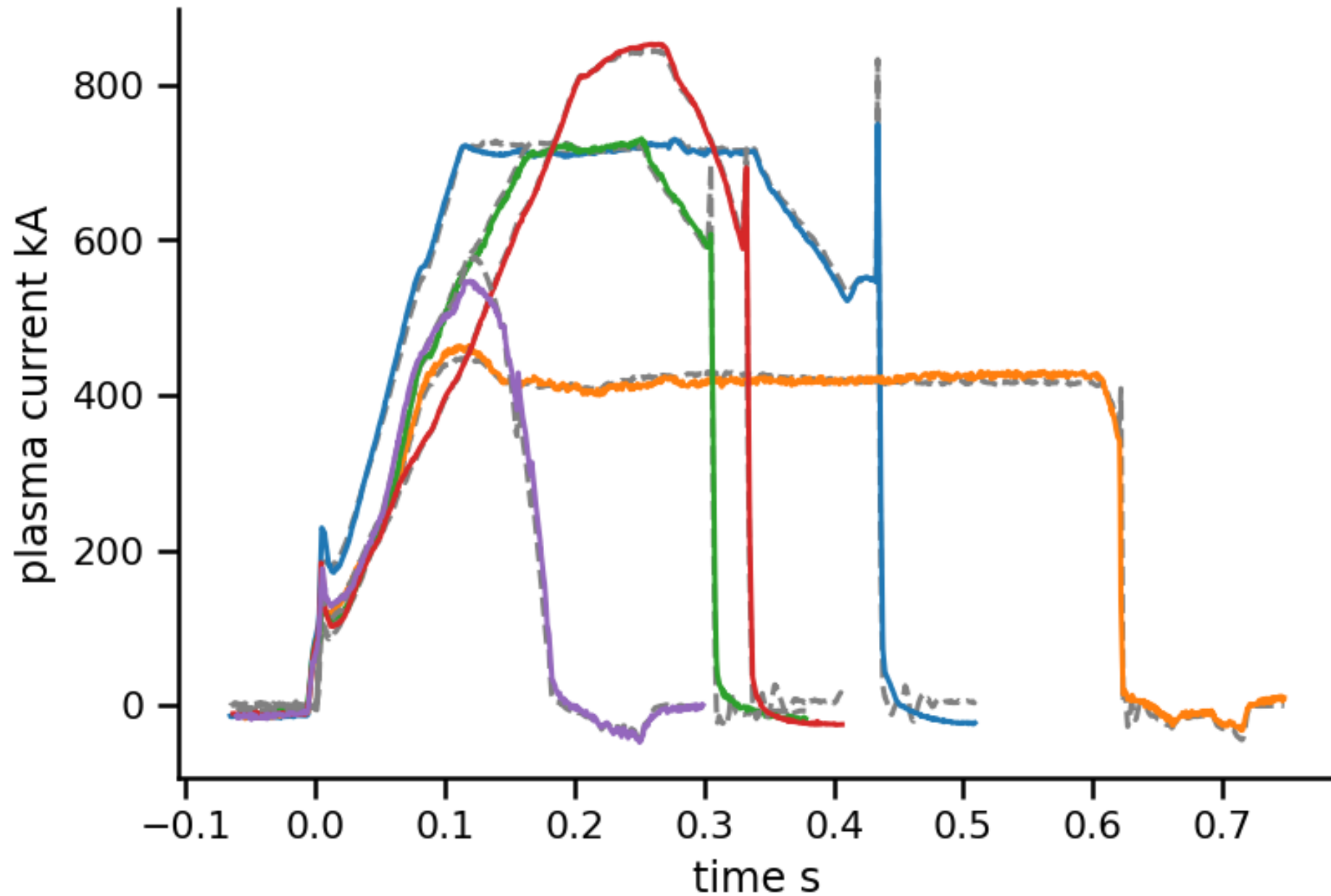
 Search leaderboard

Public Private

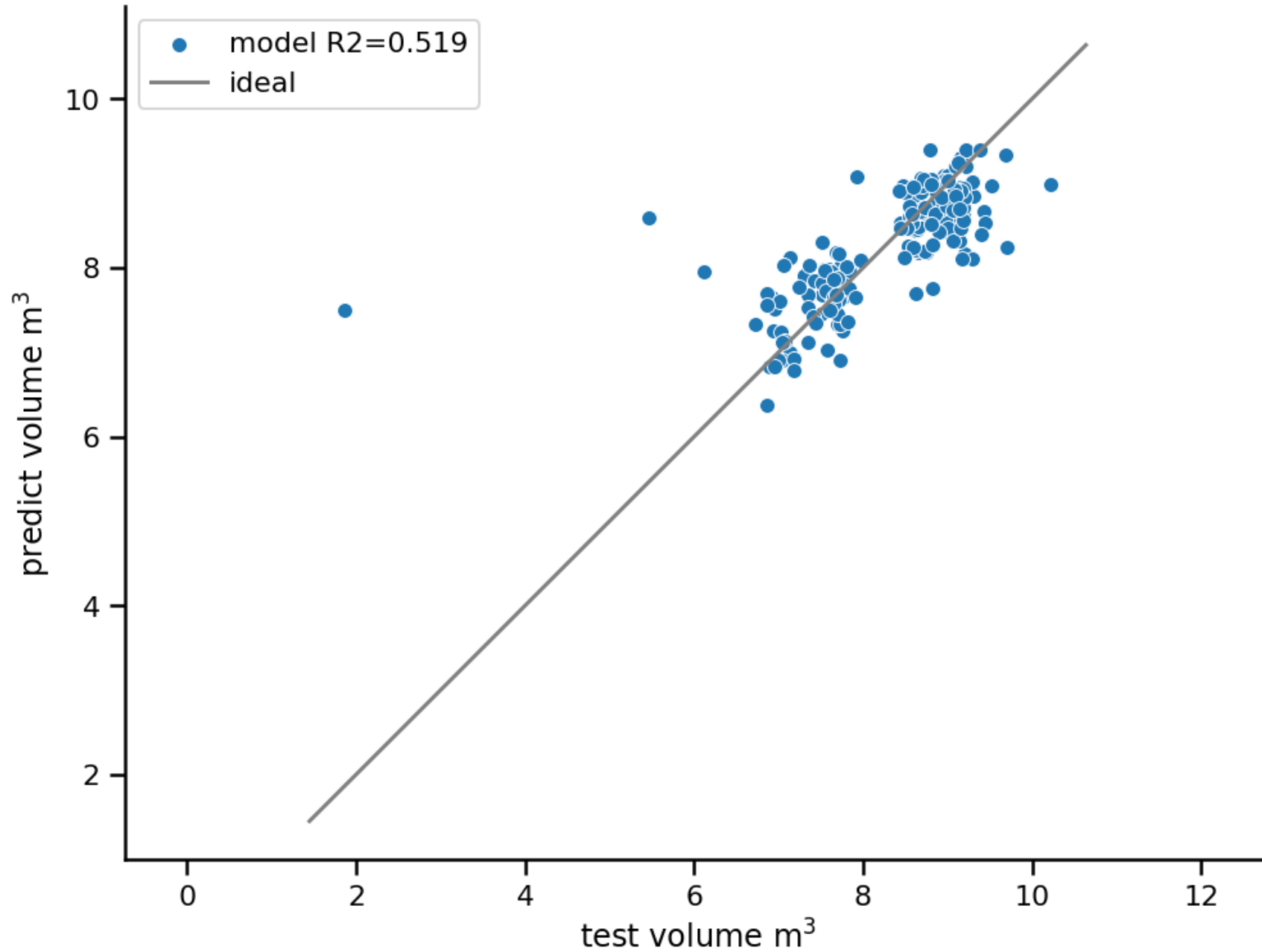
This leaderboard is calculated with approximately 44% of the test data. The final results will be based on the other 56%, so the final standings may be different.

#	Team	Members	Score	Entries	Last
	linear_regression.csv		3.77688		

Challenge #1 MAST Plasma Current



Challenge #2 MAST Plasma Volume

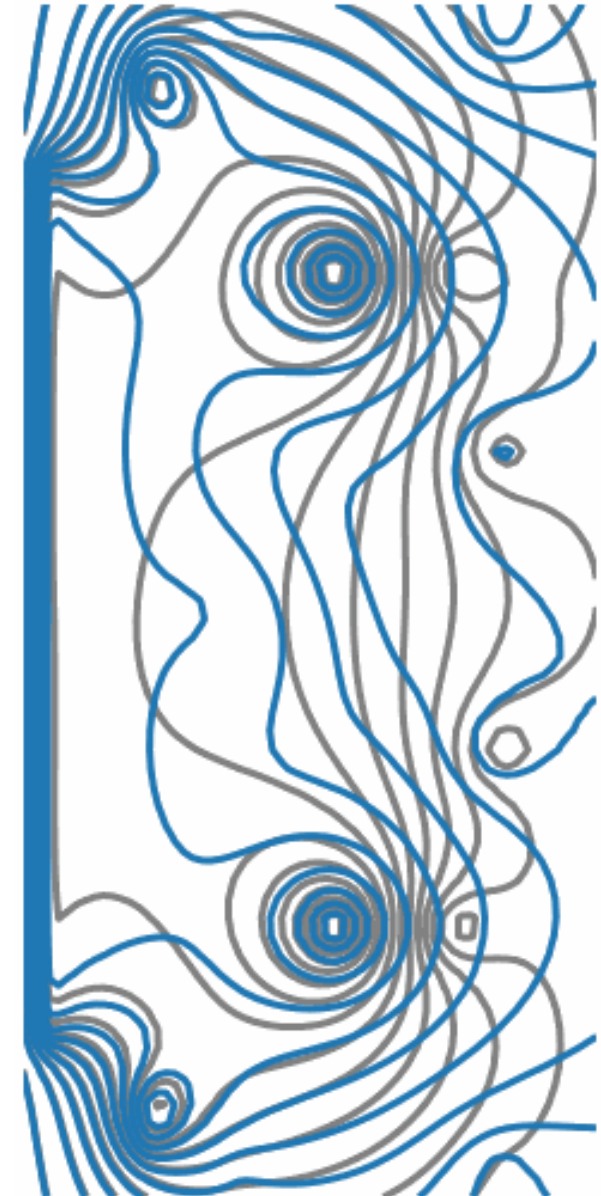


Challenge #3 MAST Plasma Equilibrium

Data variables:

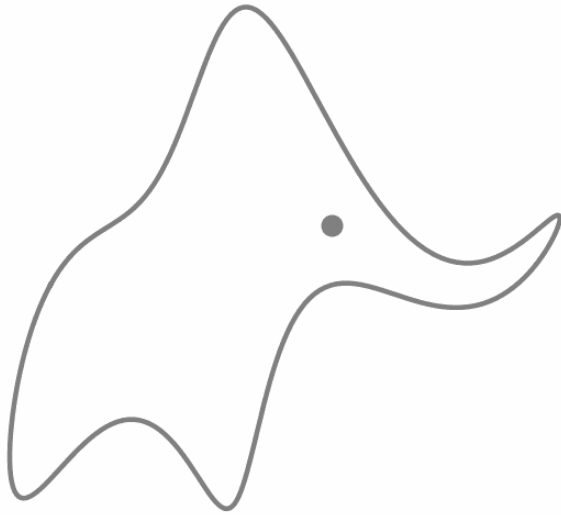
center_column	(center_column_channel, time) float64 16kB ...
coil_currents	(coil_currents_channel, time) float64 19kB ...
coil_voltages	(coil_voltages_channel, time) float64 13kB ...
flux_loops	(flux_loops_channel, time) float64 19kB ...
outer_discrete	(outer_discrete_channel, time) float64 26kB ...
saddle_coils	(saddle_coils_channel, time) float64 13kB ...
dalpa_mid_plane_center	(time) float64 3kB ...
dalpa_mid_plane_wide	(time) float64 3kB ...
dalpa_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 ...

— EFIT++
— Prediction

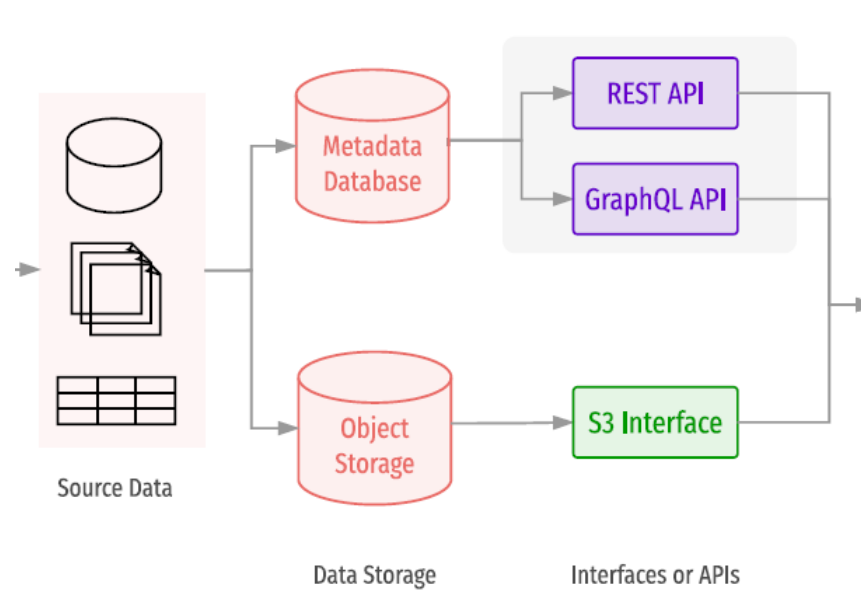


In summary this talk has warned you of the dangers of overfitting and has given you the opportunity to learn more via the challenges.

Remember the elephant.



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



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

