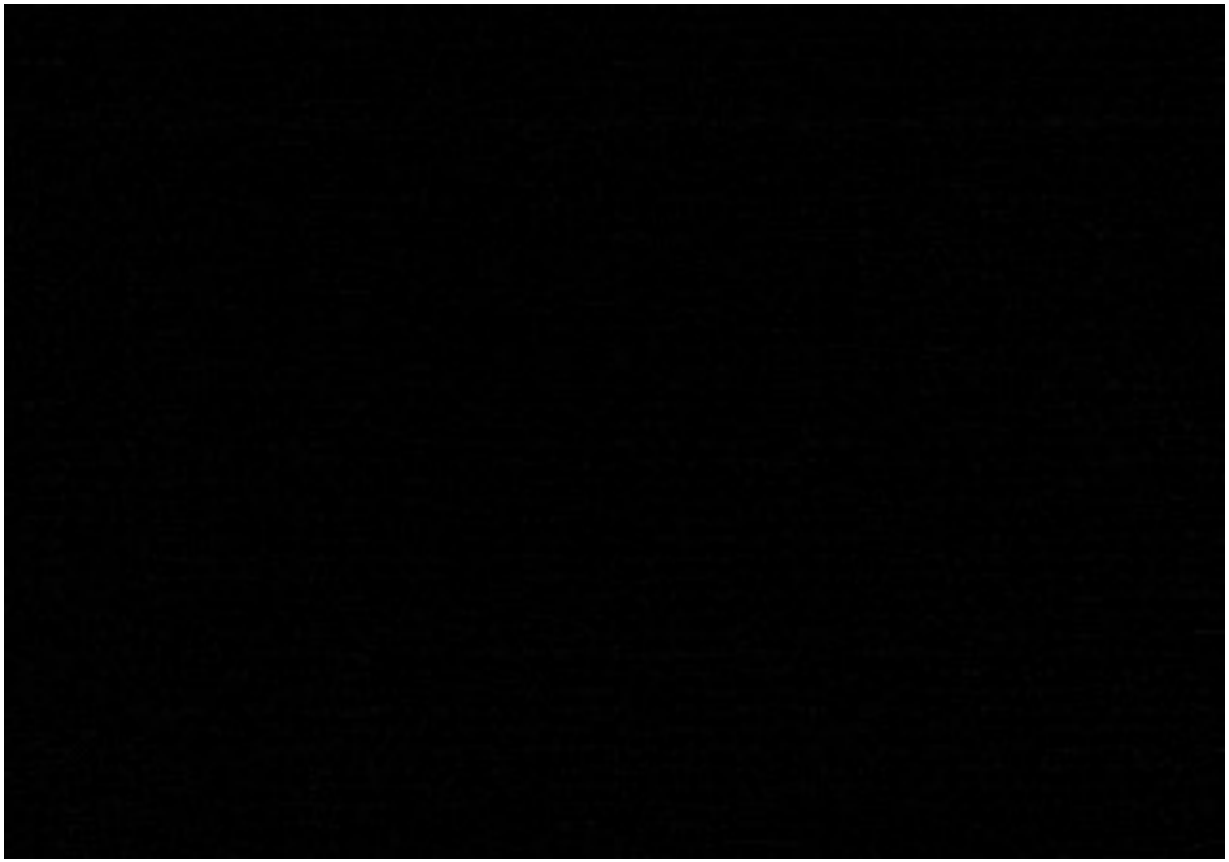
Hands-on Training - Data Science Challenge

Simon McIntosh
ITER Organization

Nathan Cummings
Culham Center for Fusion Energy

Thursday 22nd May 2025

Joint ICTP-IAEA Fusion Energy School
Trieste, Italy

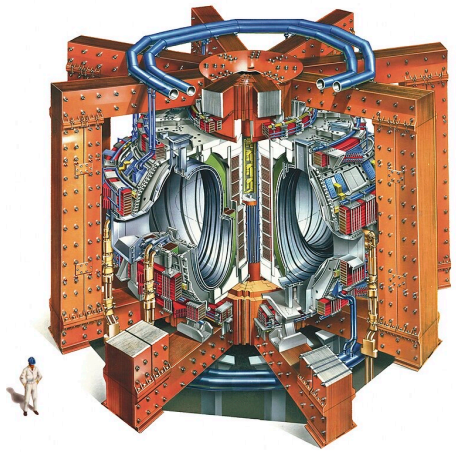


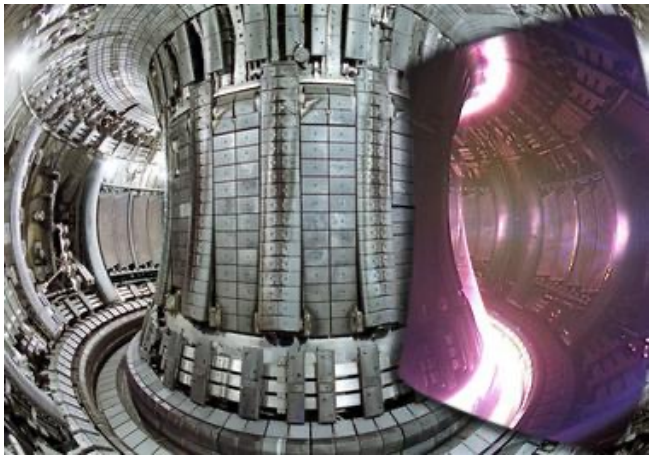
iter
china eu india japan korea russia usa

This presentation focuses on applied Data Science for Fusion.

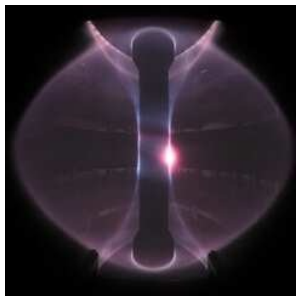
The elephant in the room.

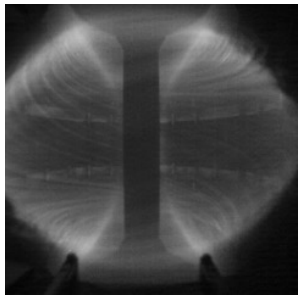
Applied Data Science in Fusion.

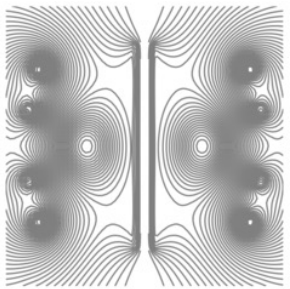




Challenge Notebooks.





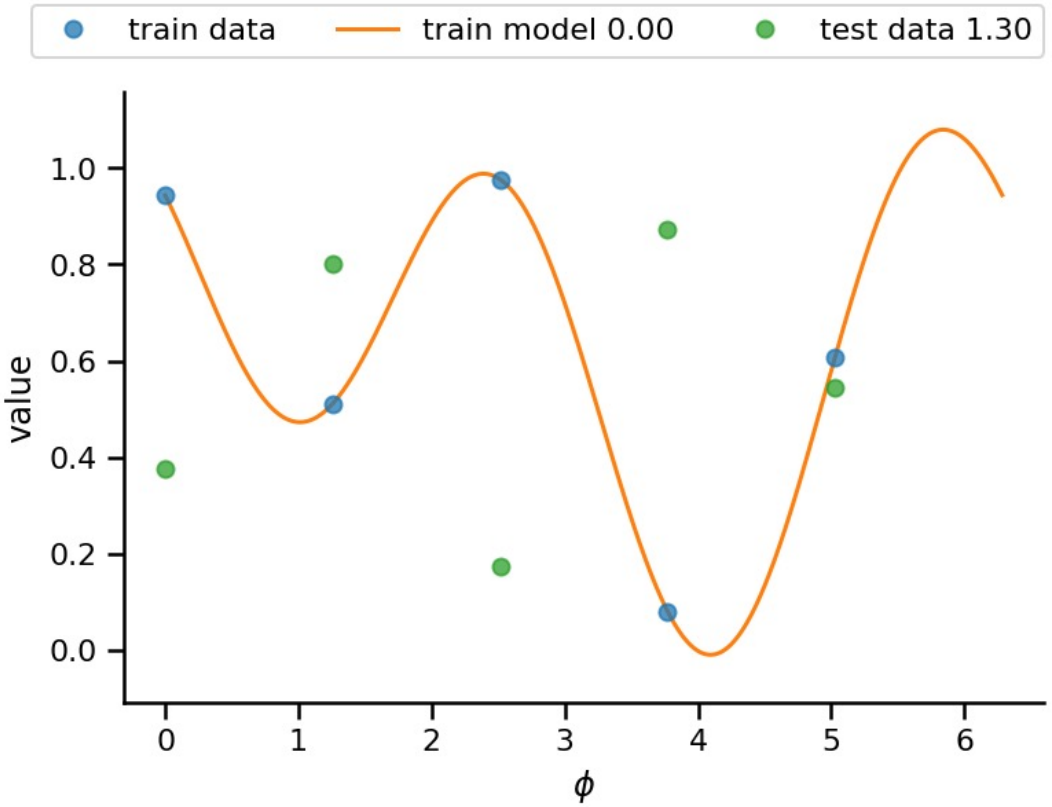


The elephant in the room is over-fitting.

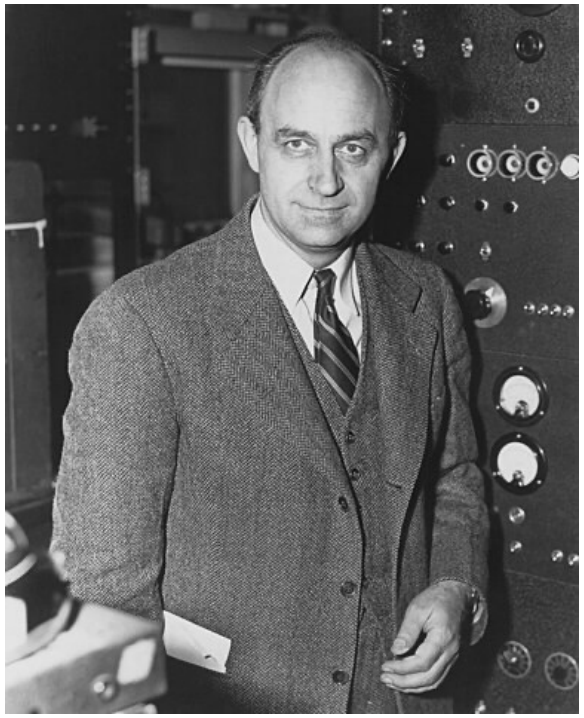
“All models are wrong, but some are useful”. George Box

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”.



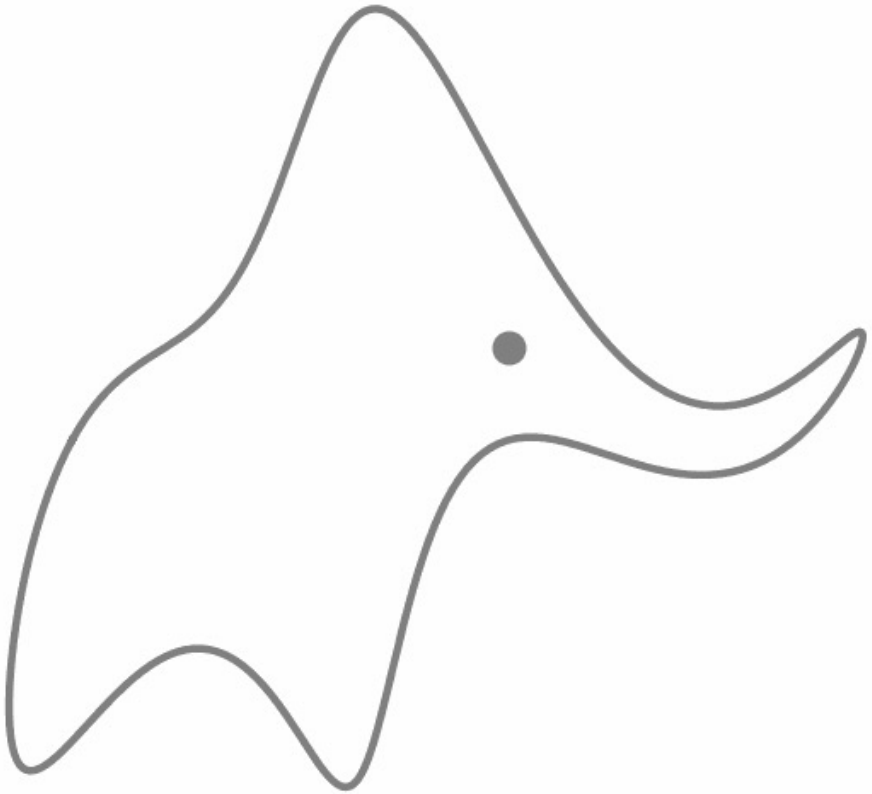
Enrico Fermi

John von Neumann

Simple basis functions can generate complex shapes.

von Neumann's Elephant

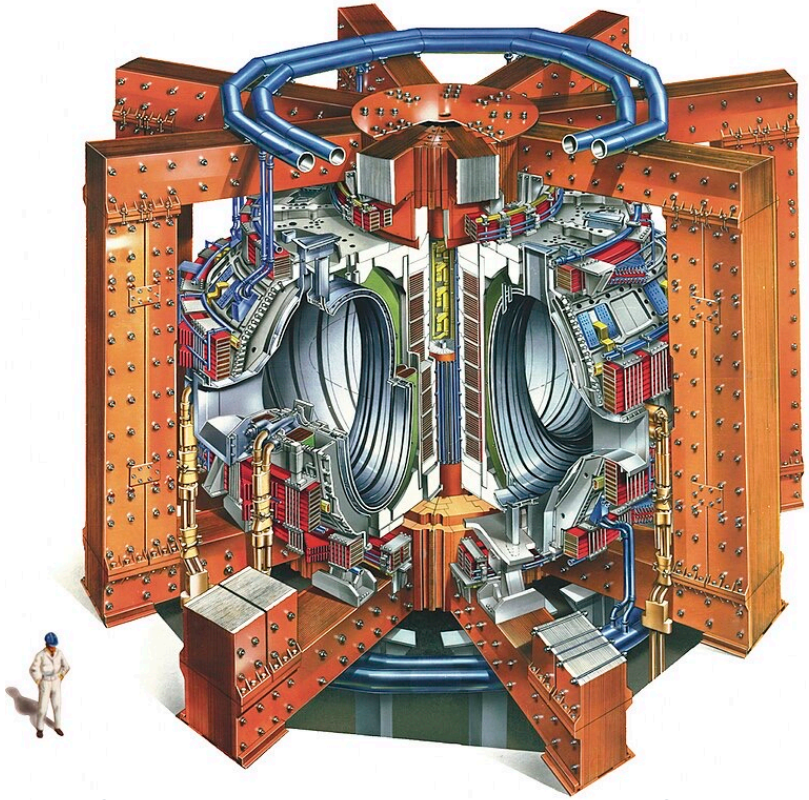
$$Pn = \begin{bmatrix} -55 + 15j \\ -9 - 4j \\ 0 + 7j \\ -5 - 11j \\ 20 + 1j \end{bmatrix}$$



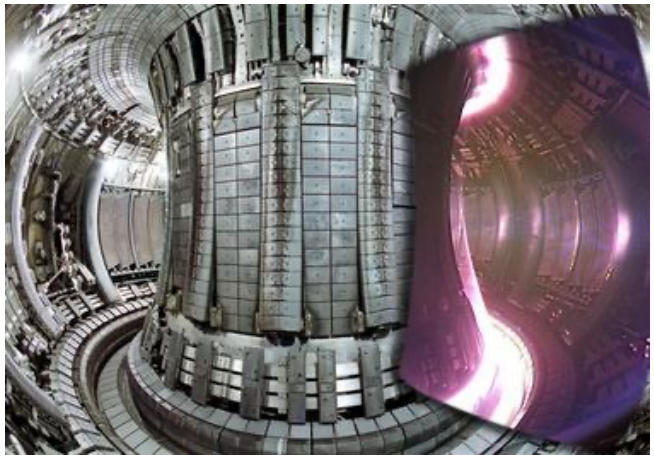
$$Cn = \begin{bmatrix} 0 + 0j \\ Pn[0] \\ Pn[1] \\ 0 + 0j \\ 0 + 0j \\ Pn[2] \\ Pn[2] \\ 0 + 0j \\ 0 + Pn[3]j \\ -Pn[1] \\ Pn[3] - Pn[0]j \end{bmatrix}$$

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

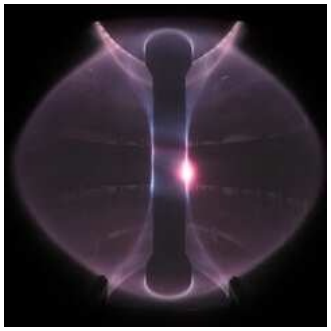


Operating cost ~200,000 euros per day

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

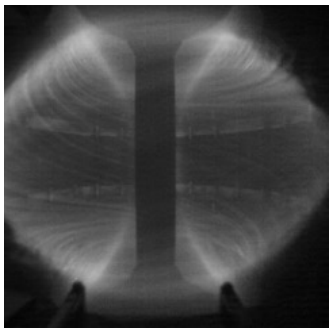
~~Code~~ Data should always be written in a way that promotes readability.

Data Science Challenges of the ITER International School 2024



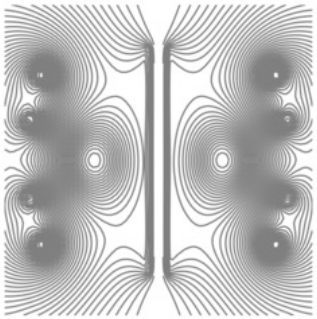
MAST Plasma Current

Infer plasma current produced by CCFE's Mega Ampere Spherical Tokamak from discrete magnetic diagnostic data.
https://github.com/Simon-McIntosh/data-science-challenges



MAST Plasma Volume

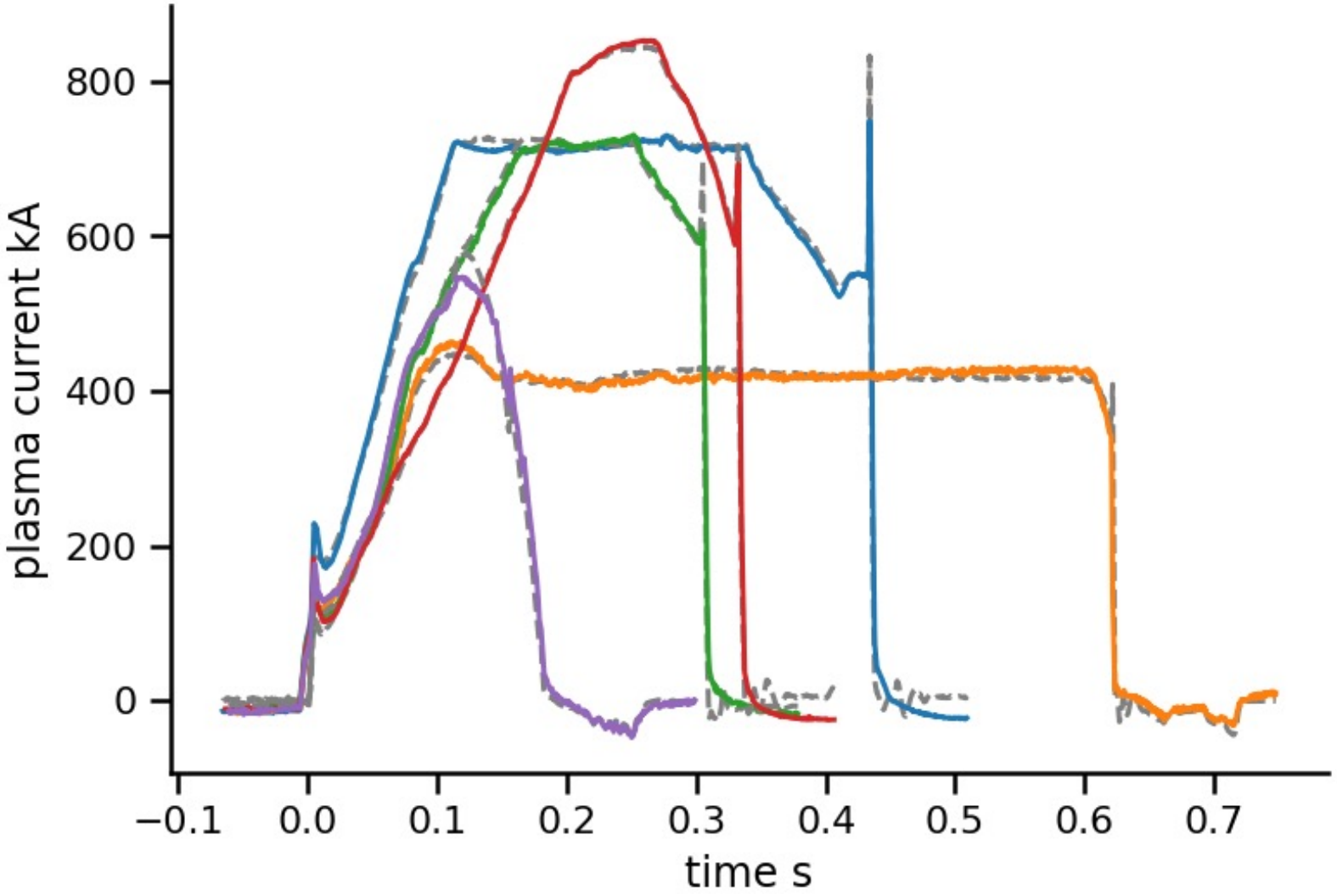
Infer the volume of plasmas produced by the CCFE's Mega Ampere Spherical Tokamak using frames from a wide-angle visible spectrum camera.



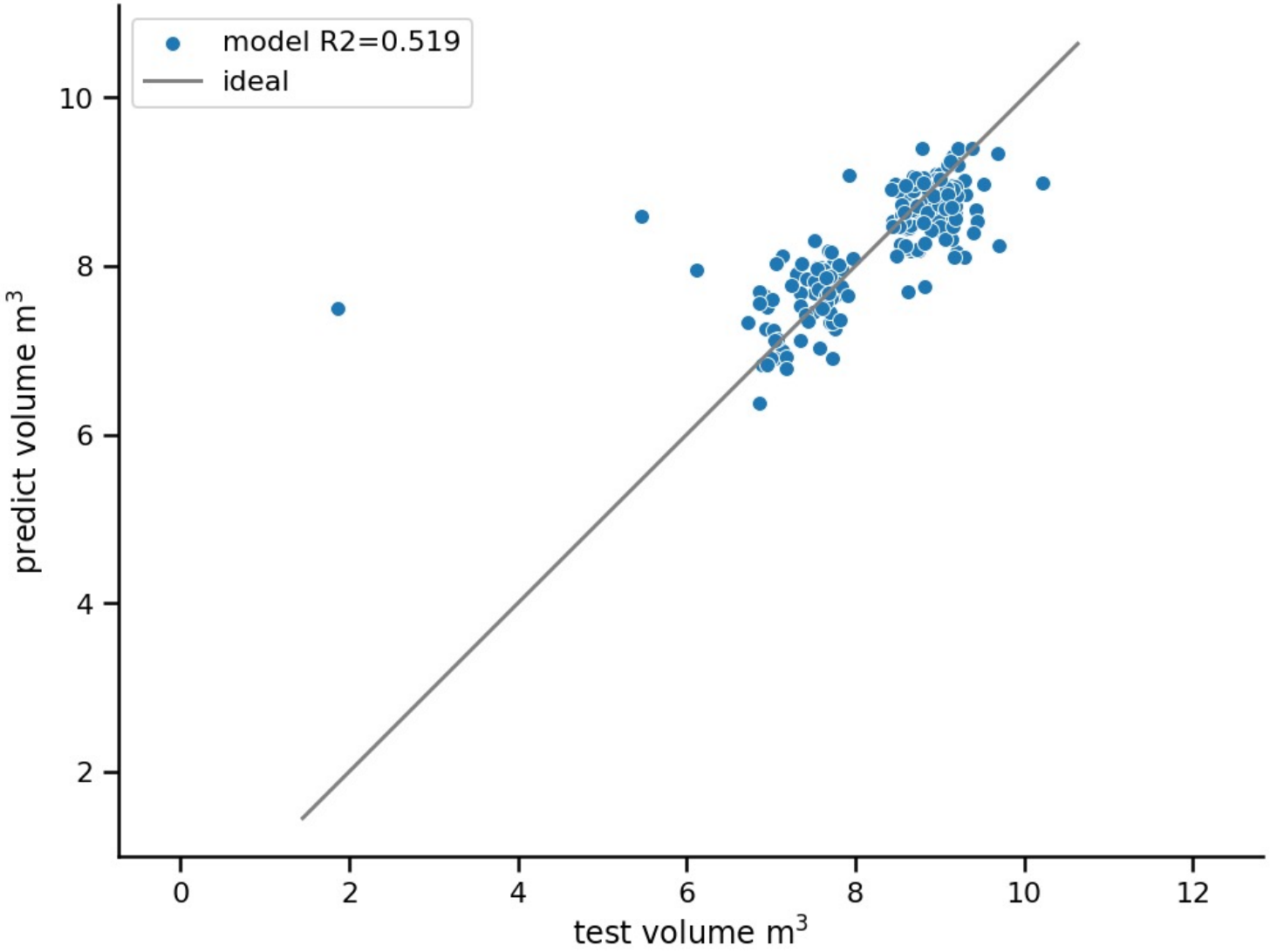
MAST Plasma Equilibrium

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

Challenge #1 MAST Plasma Current

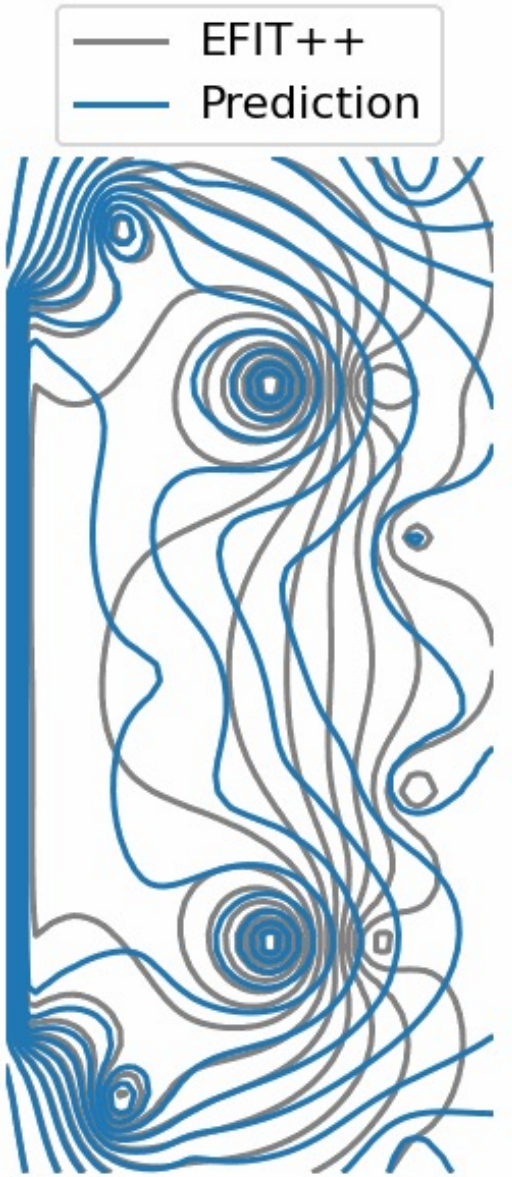


Challenge #2 MAST Plasma Volume



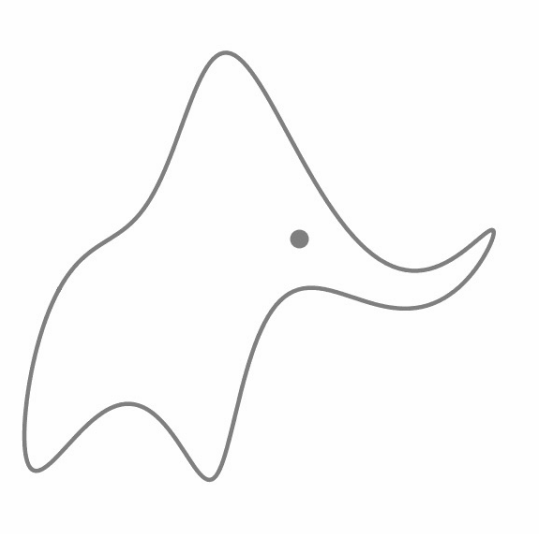
Challenge #3 MAST Plasma Equilibrium

```
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 ...
```

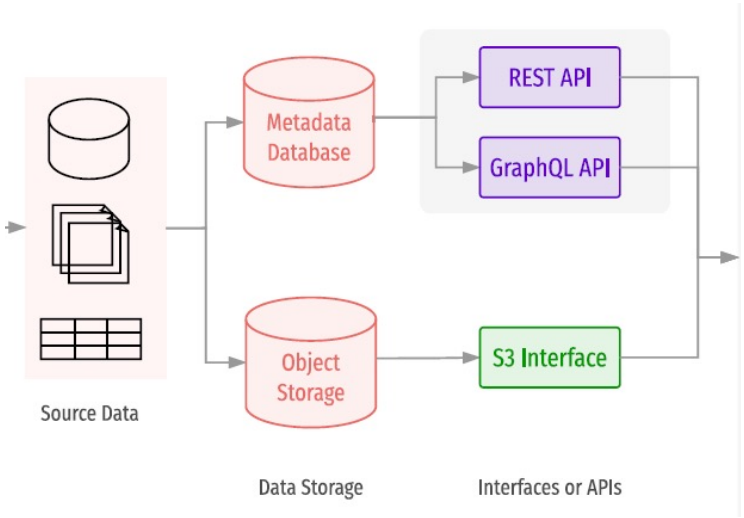


In summary this talk has warned you of the dangers of overfitting and has given you the opportunity to learn more via the challenges.

Remember the elephant.

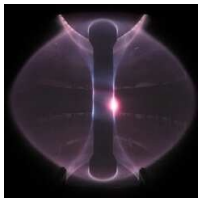
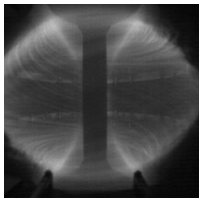
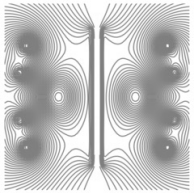


Data Science Challenge facilitated by FAIR data and open-source tools.



pandas xarray scikit learn

Practical application is considered the best way to learn.

“Tell me and I forget,

teach me and I remember,

involve me and I learn.”