Machine Learning for Magnetic Fusion Science Part I

Cristina Rea

crea@mit.edu

MIT Plasma Science and Fusion Center, Cambridge, MA, USA

Joint ICTP-IAEA School on AI for Nuclear, Plasma and Fusion Science, Trieste, May 24th- 25th, 2023

PSFC

About Me

2014 – PhD in Physics, University of Padova, Italy
2015 – Data Scientist, UniCredit, Milan, Italy
2016 – Postdoctoral Associate, MIT PSFC, Cambridge USA
2019 – Research Scientist, MIT PSFC, Cambridge USA

Research focus on **disruption physics** and **disruption warning algorithms** for fusion plasmas adopting state-of-the-art **Machine Learning** techniques.

Outline

1. Brief Fusion primer
2. The Universality theorem and brief ML taxonomy
3. Explainable deep learning vs design of interpretable models

Part I

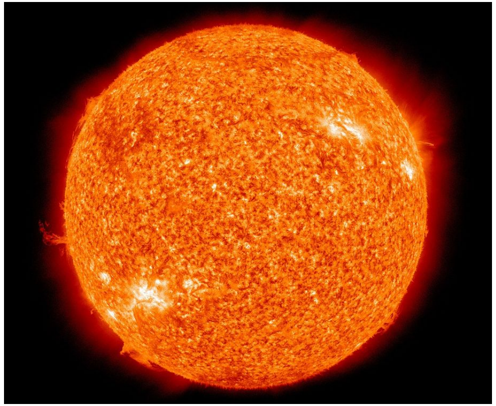
4. Domain adaptation and transfer learning
5. Current challenges and opportunities for future research
6. Conclusions

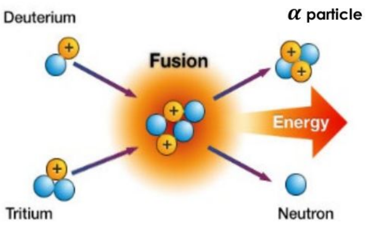
Part II

Outline

1. Brief Fusion primer
2. The Universality theorem and brief ML taxonomy
3. Explainable deep learning vs design of interpretable models
4. Domain adaptation and transfer learning
5. Current challenges and opportunities for future research
6. Conclusions

Fusion research is tackling transformational technologies to provide alternative, carbon-free electricity generation





$$D + T \rightarrow \alpha(3.5\ \mathrm{MeV}) + n(14.1\ \mathrm{MeV})$$

Lab research conducted via:

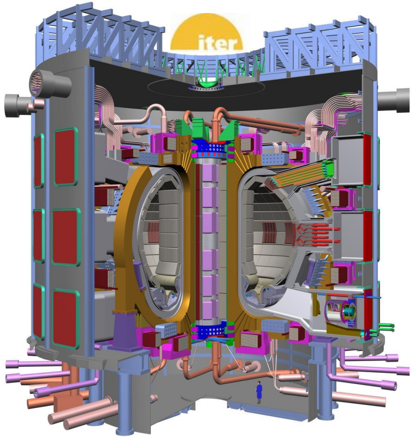
- Heating and confinement of a plasma of hydrogen isotopes via magnetic fields → **magnetic confinement**

  Mordijck's Tuesday lecture

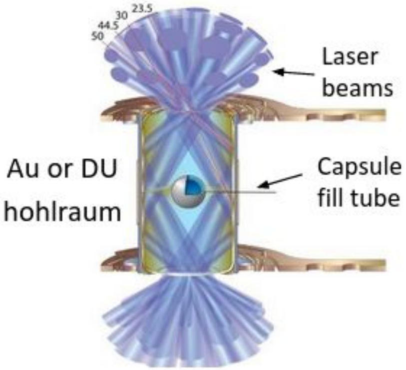
- Heating and compressing via lasers a fuel target of hydrogen isotopes → **inertial confinement**

  McClarren's Tuesday lecture

MFE



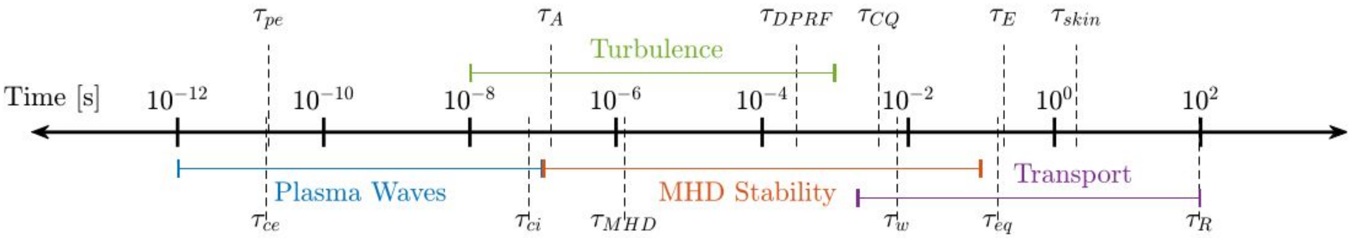
IFE



Let's take a closer look at **MFE** plasmas.

Fusion Plasmas: nonlinear phenomena really hot or really fast, hard to diagnose, lots theory & exp data, not so easy to bridge

Fusion Plasmas: nonlinear phenomena really hot or really fast, hard to diagnose, lots theory & exp data, not so easy to bridge



← K. Montes, PhD Thesis, 2021

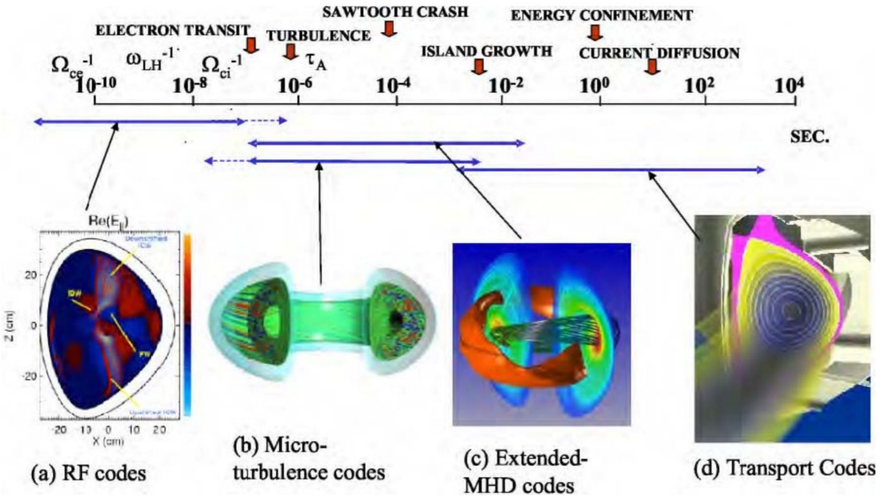
and

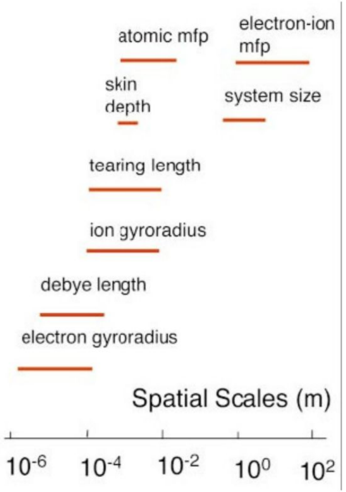
↓ L. Chacon 2022 ICTP School

Challenges in thermonuclear fusion simulation: "The tyranny of scales"

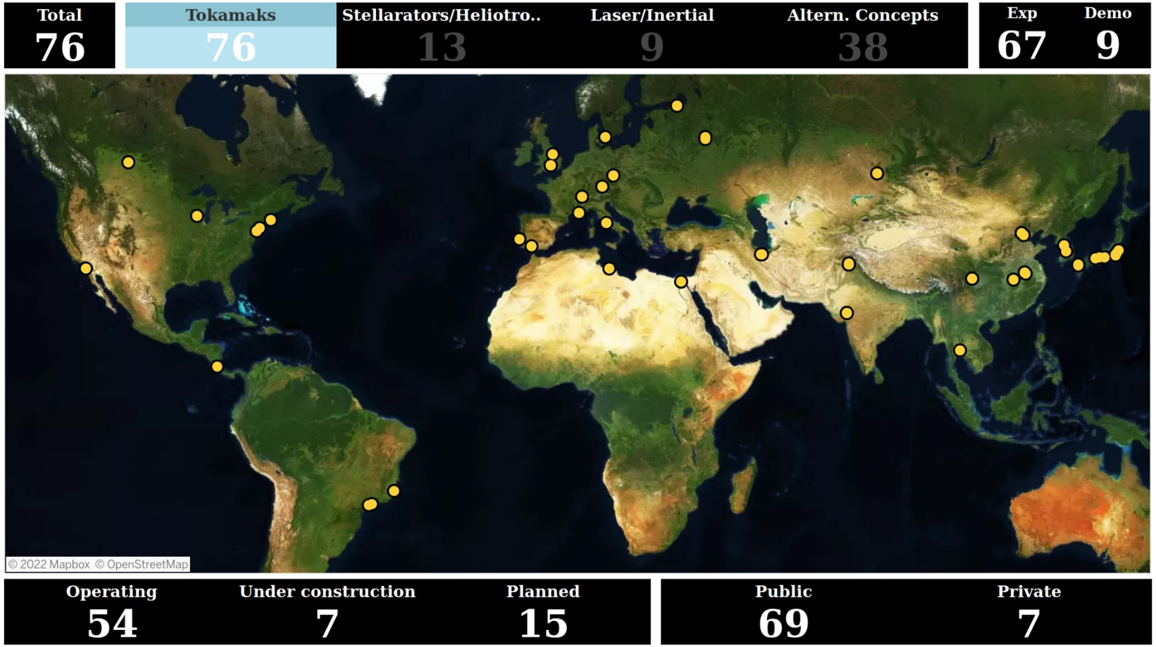
Fusion plasma dynamics spanning wide range of spatial and temporal scales

Not so easy to develop first principle solutions!





Many operating experimental devices for magnetically confined fusion research, more planned!



IAEA Fusion Device Information System
https://www.iter.org/of-interest/944

Huge amount of experimental and simulation data available enabling **Machine Learning applications:**

- ❑ optimization of experimental design
- ❑ real-time monitoring of proximity to instability
- ❑ trajectory planning optimization
- ❑ fast surrogates to accelerate simulations
- ❑ …

Outline

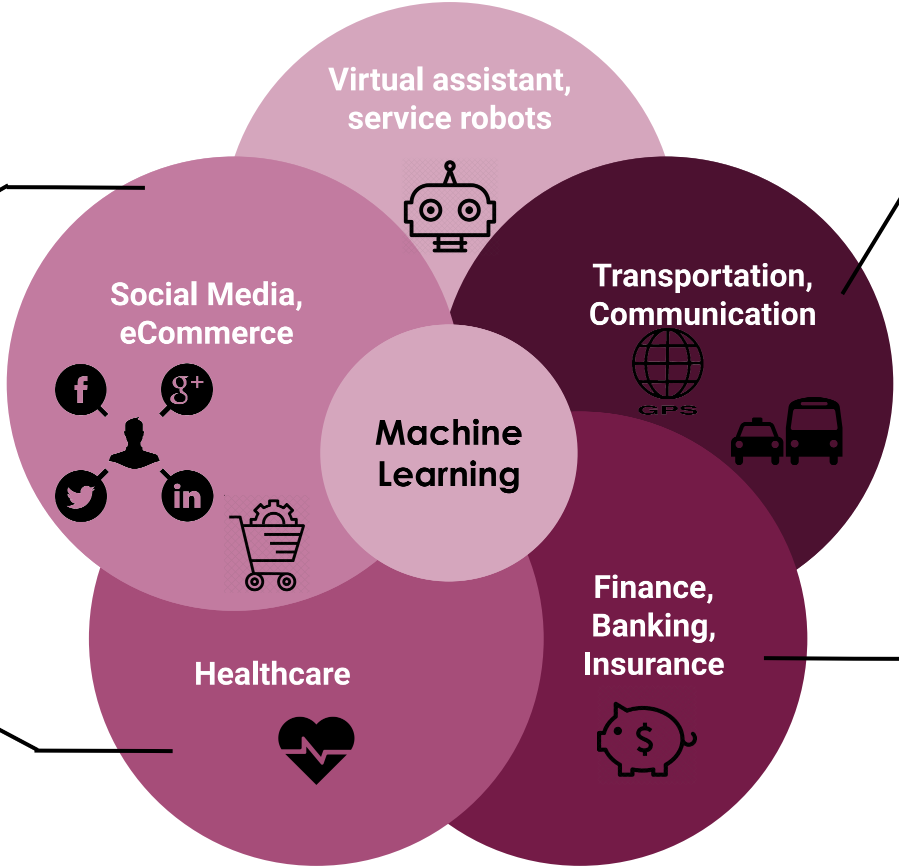
a series of useful concepts for ML practitioners

1. Brief Fusion primer
2. ~~The Universality theorem and brief ML taxonomy~~
3. Explainable deep learning vs design of interpretable models
4. Domain adaptation and transfer learning
5. Current challenges and opportunities for future research
6. Conclusions

Pervasive use of Machine Learning in everyday life, widely adopted tool in Fusion too!

customer support, product recommendation, advertising, sentiment analysis, spam filtering

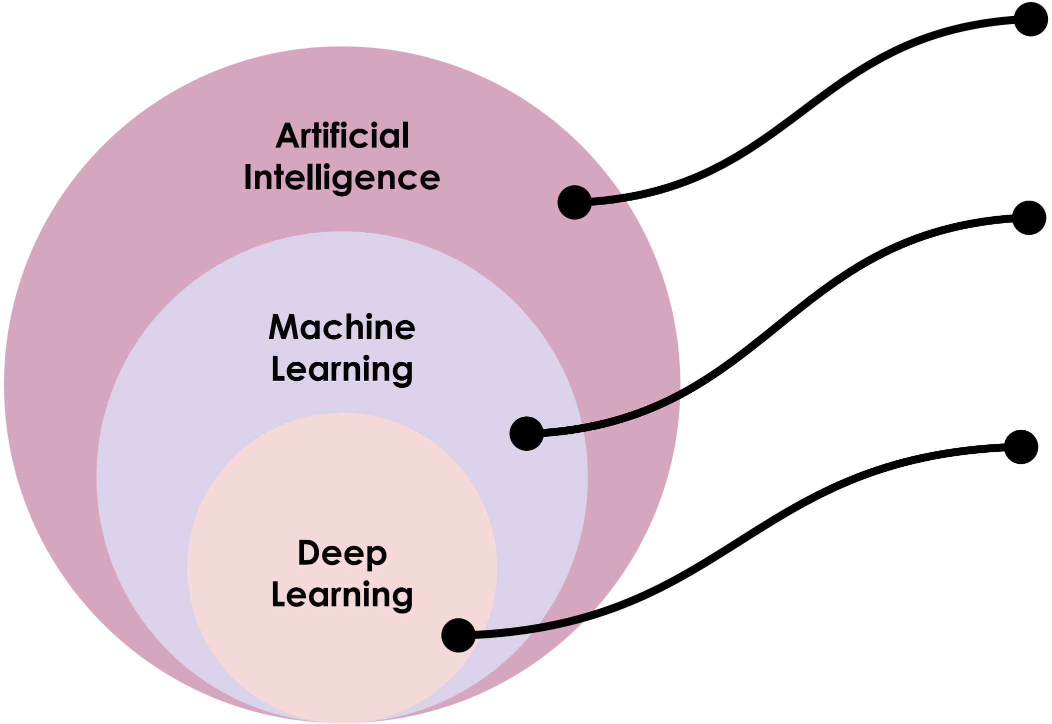
robotics surgery, therapy assessment, disease diagnosis and predictive healthcare, drug discovery, …



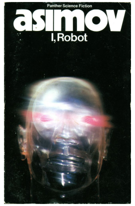
safety monitoring, air traffic control, unmanned aerial vehicles, self-driving cars, GPS, imagery,...

high frequency trading, portfolio management, fraud detection, …

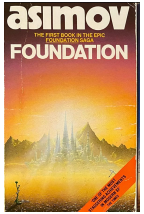
Sometimes there's confusion about terminology, too many buzzwords!



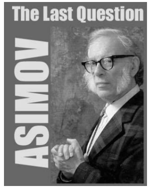
To **mimic human behavior** and functions such as **learning** and **problem solving**.



AI subset using **statistical methods** to enable **learn-from-experience** paradigm.

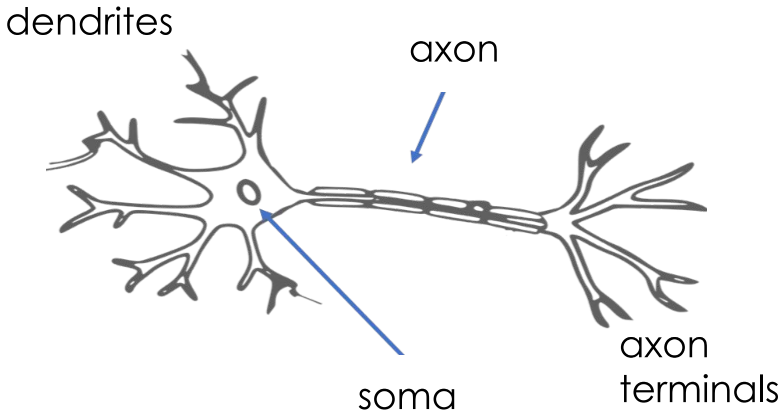


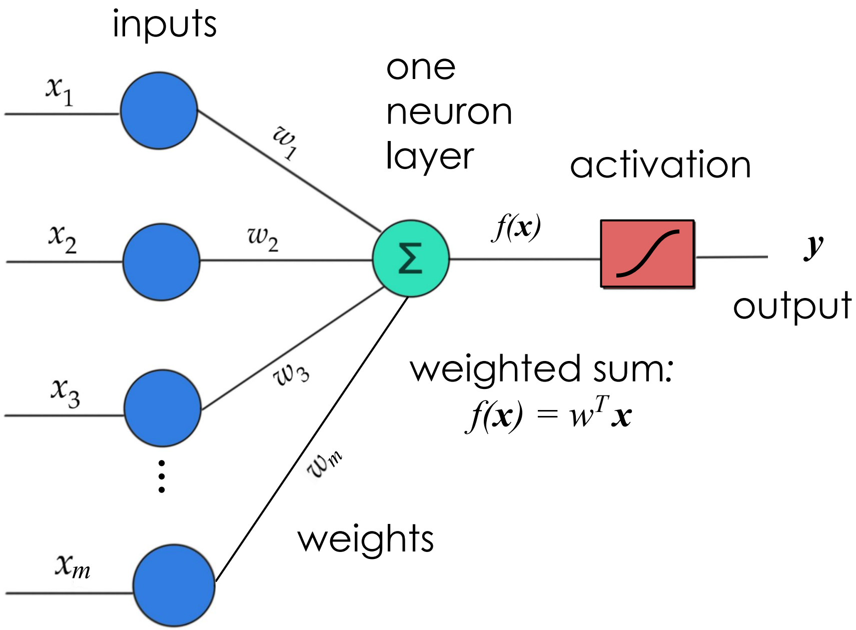
ML subset with broader **generalization** capabilities – **neural networks**.



Adapted from medium.com/@StepUpAnalytics

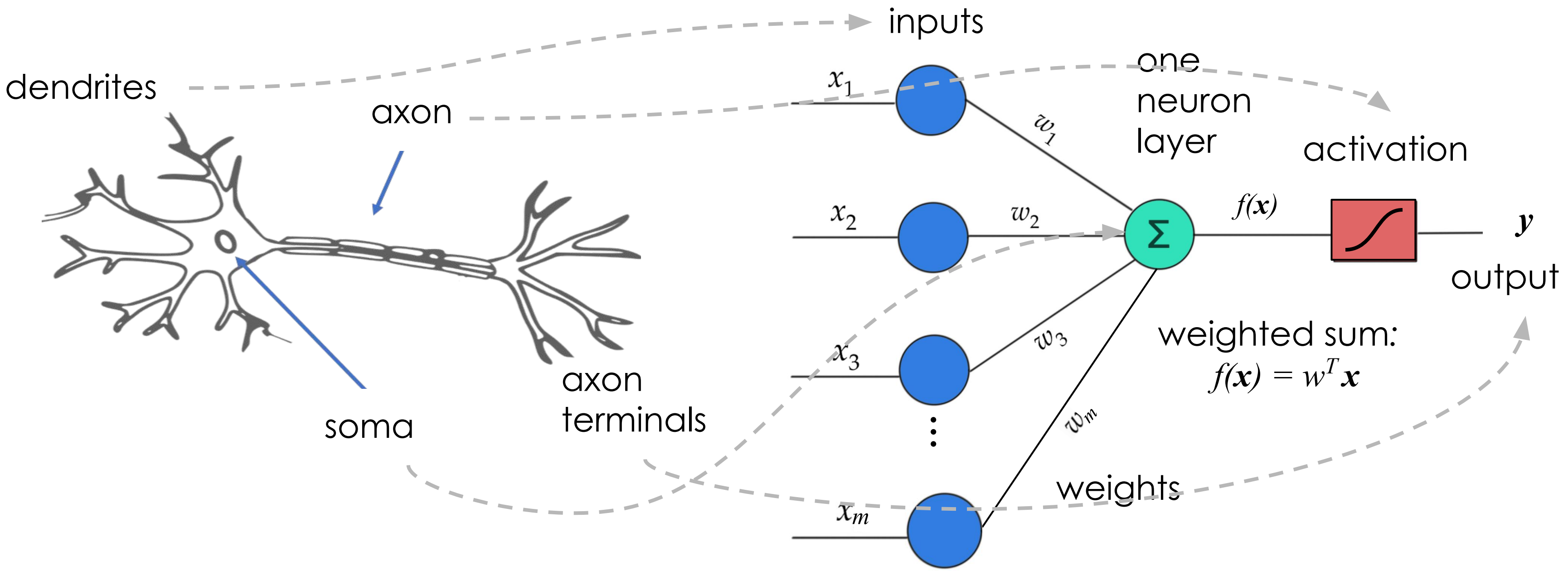
From biological to artificial neurons: the computational graph





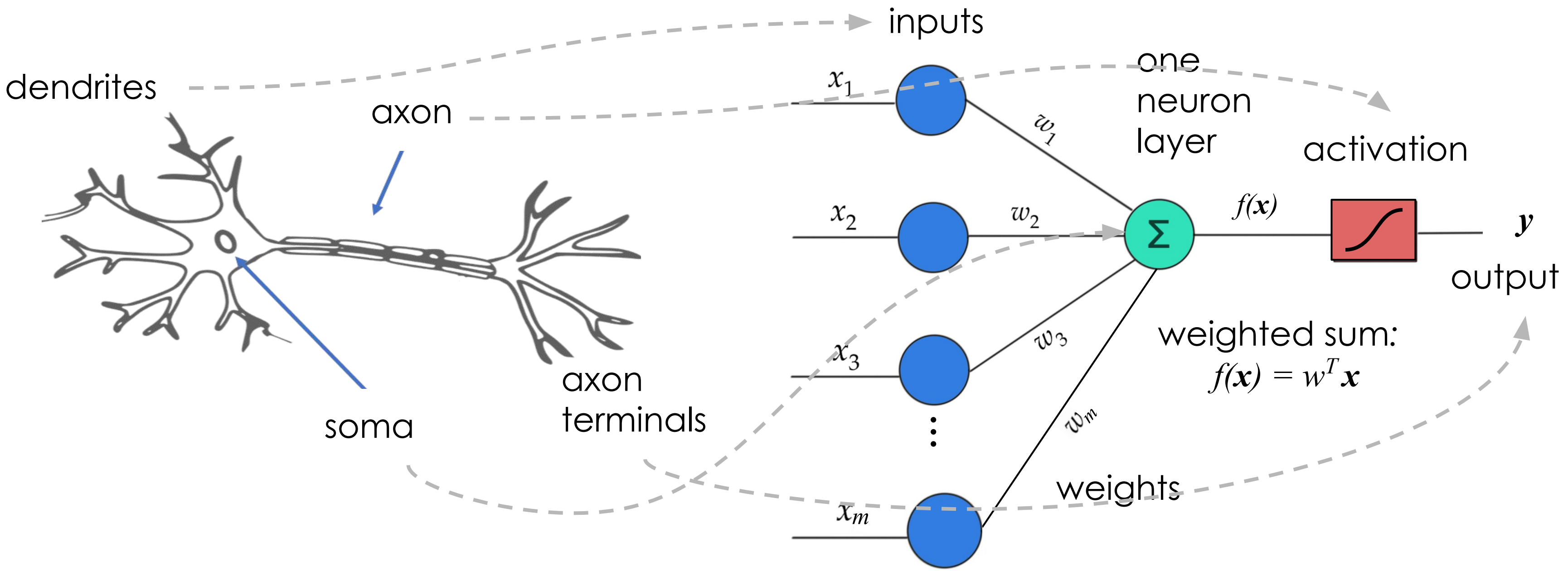
Credits: M. Kuchera 💖

From biological to artificial neurons: the computational graph



Credits: M. Kuchera 💖

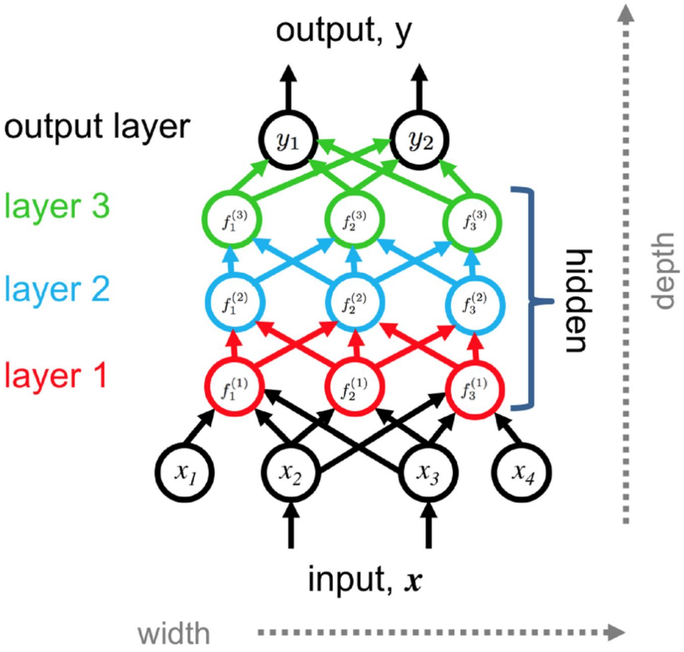
Artificial neurons can <u>represent</u> any function with arbitrary accuracy



Credits: M. Kuchera 💖

The Universality Theorem: for <u>any arbitrary</u> $f(x)$, there is <u>always</u> a network that can <u>approximate</u> it

$$y \approx f(x) = f^{(4)}(f^{(3)}(f^{(2)}(f^{(1)}(x))))$$



Deep neural network example, adapted from B. Spears et al PoP 2018

Caveats:

- Increasing the depth can improve the <u>approximation</u>.

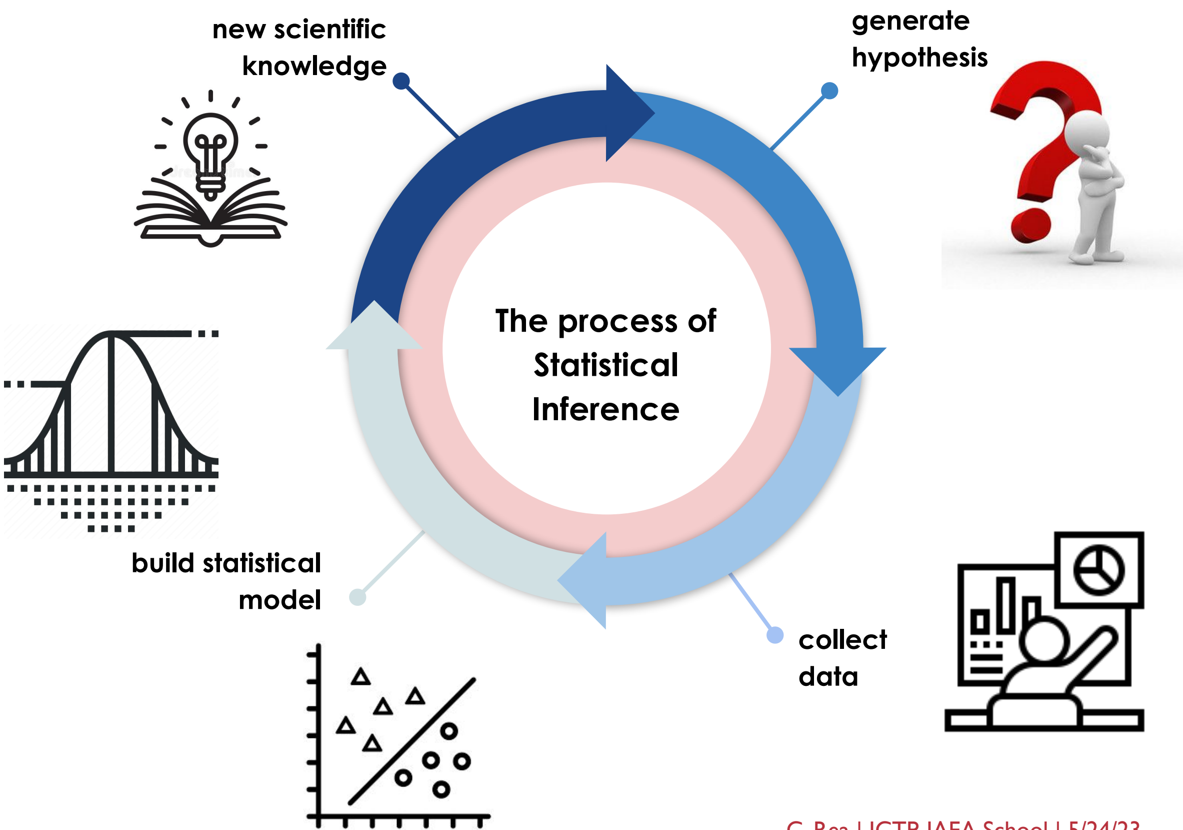
  $| y - f(\mathbf{x}) | < \varepsilon$

- Activation must be <u>continuous</u>.

Neural networks provide *nonlinear* **mapping from inputs to outputs**, or a way to **represent your data** through *function approximation and estimation*.

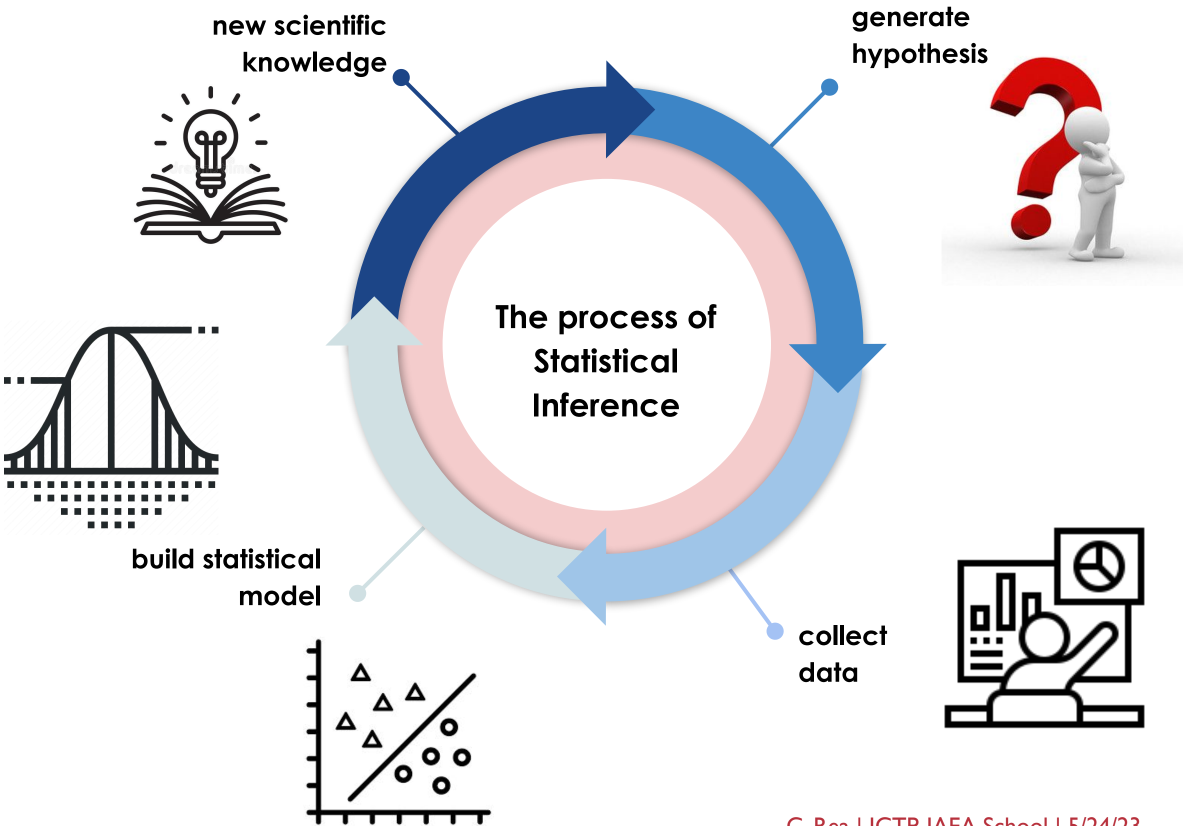
Statistical inference to learn representations from available data

Existing challenges in evaluating the mapping adherence to ground truth for ML models

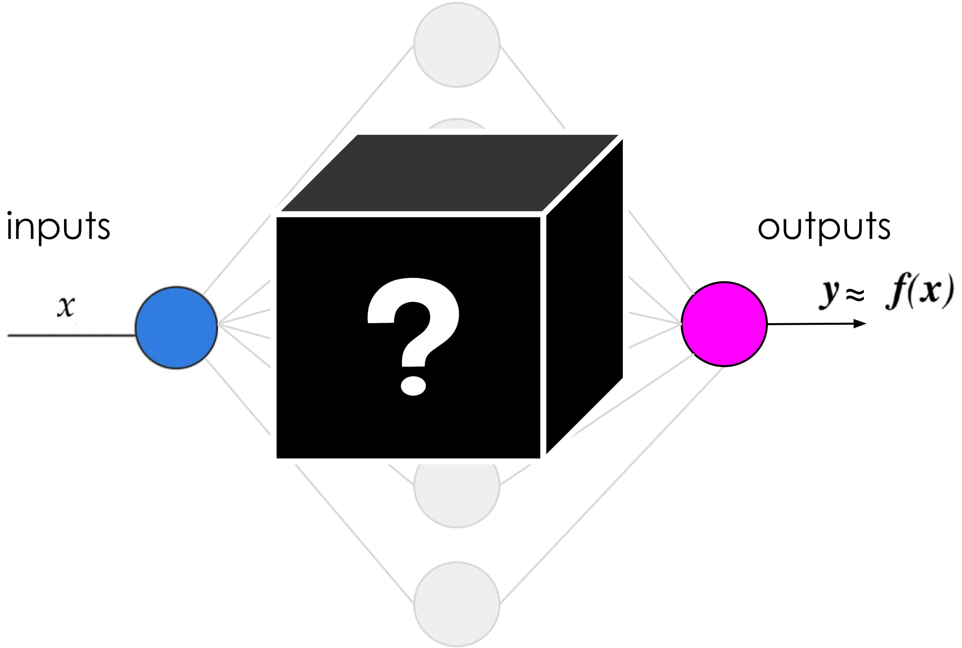


Statistical inference to learn representations from available data

https://xkcd.com/1838/



Introducing the black box: the issue with high-stakes decision making …



Black box as either

- function **too complicated** for human to comprehend or
- function that is **proprietary**

C. Rudin, Nat Mach Intell 1, 206–215 (2019)

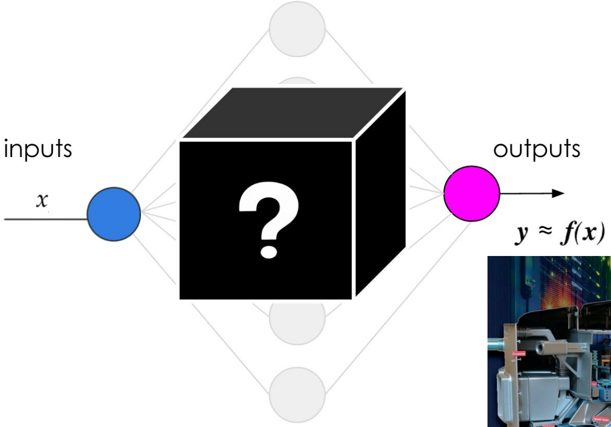
Implications:

- lack of transparency and accountability,
- troubleshooting challenges.

High-stakes decision making:

- healthcare,
- criminal justice,
- child welfare screening,
- self-driving cars,
- …

High-stakes decision making and the parallelism with the fusion context



Fusion energy systems:
Any ML-based decision needs to be **trusted** and **justified**, or *licensed* → high-stake decisions!

Science discovery → Reconciliation with physical understanding, key ingredient to advance fusion research.

explainable predictions

vs

interpretable models

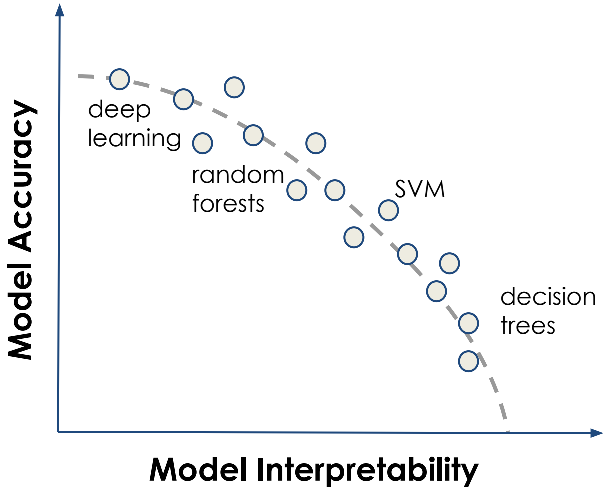
D. Humphreys et al, 2020 Advancing Fusion with Machine Learning Research Needs Workshop Report, J. Fusion Energy 39 123–55

Outline

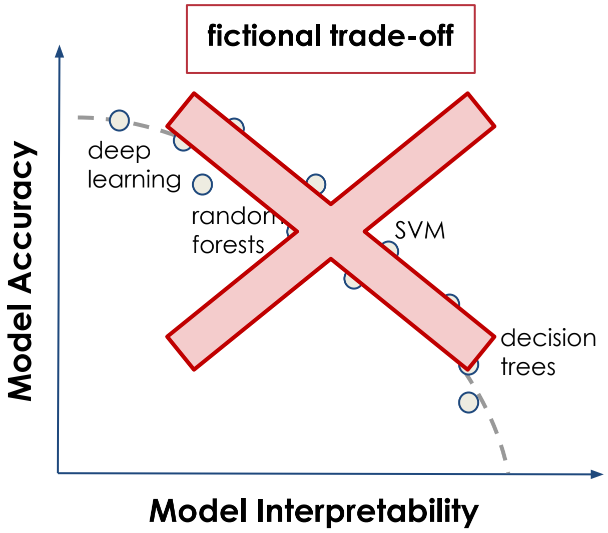
1. Brief Fusion primer
2. The Universality theorem and brief ML taxonomy
3. Explainable deep learning vs design of interpretable models
4. Domain adaptation and transfer learning
5. Current challenges and opportunities for future research
6. Conclusions

https://xkcd.com/2541/

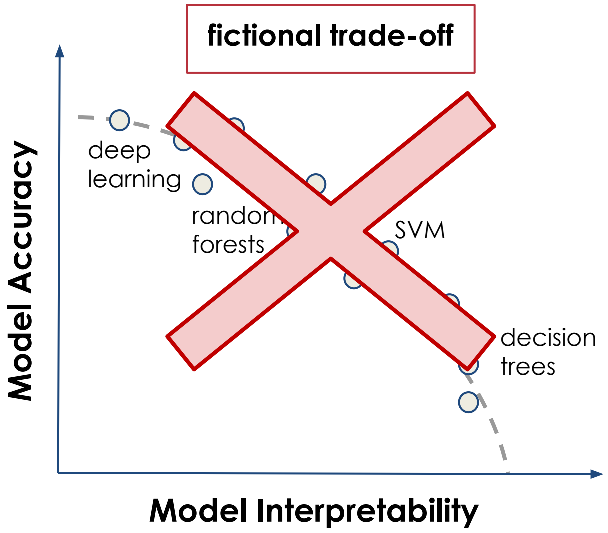
Common perception of accuracy vs interpretability trade-off



More interpretable and simpler models can be as accurate as black boxes



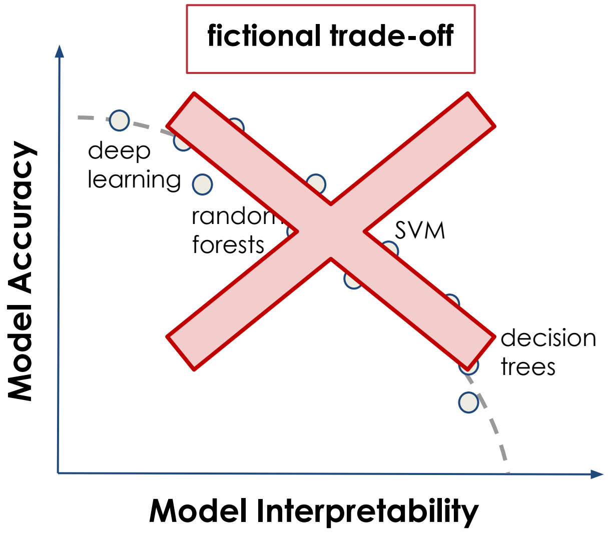
More interpretable and simpler models can be as accurate as black boxes



No unique "interpretability" definition:

- It's algorithm dependent – e.g., possibility to inspect reasons.
- It's domain dependent – e.g. sparsity not good for natural image classification.

More interpretable and simpler models can be as accurate as black boxes

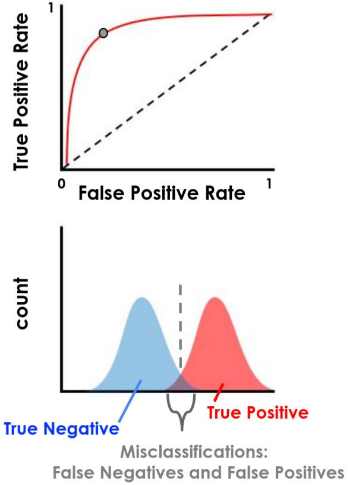


No unique "interpretability" definition:

- It's algorithm dependent – e.g., possibility to inspect reasons.
- It's domain dependent – e.g. sparsity not good for natural image classification.

What about accuracy definition?

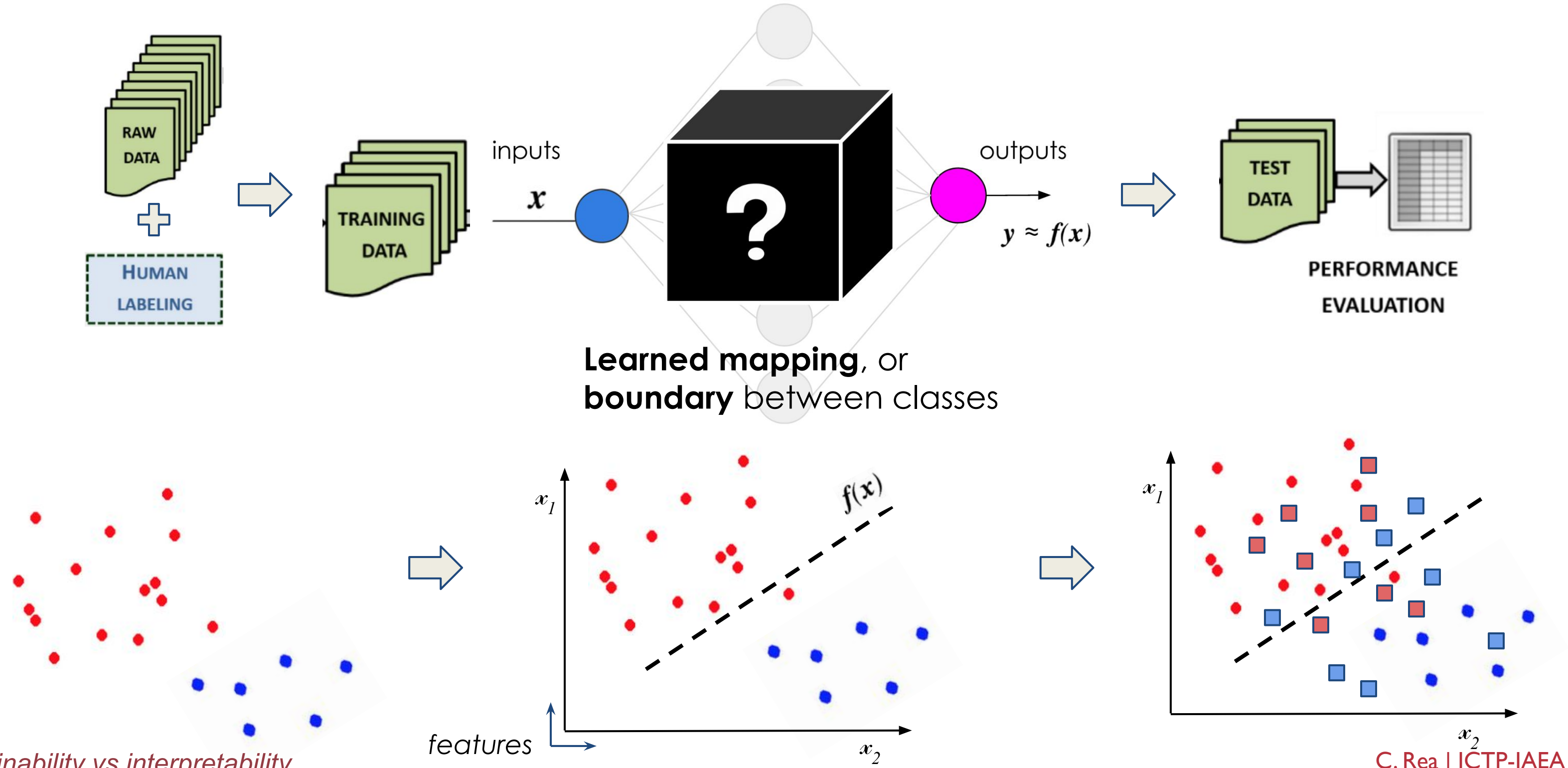
- Typically well-defined – e.g., counting statistics of misclassifications, root mean squared error, …



ML systems' prediction accuracy measured on new test data …

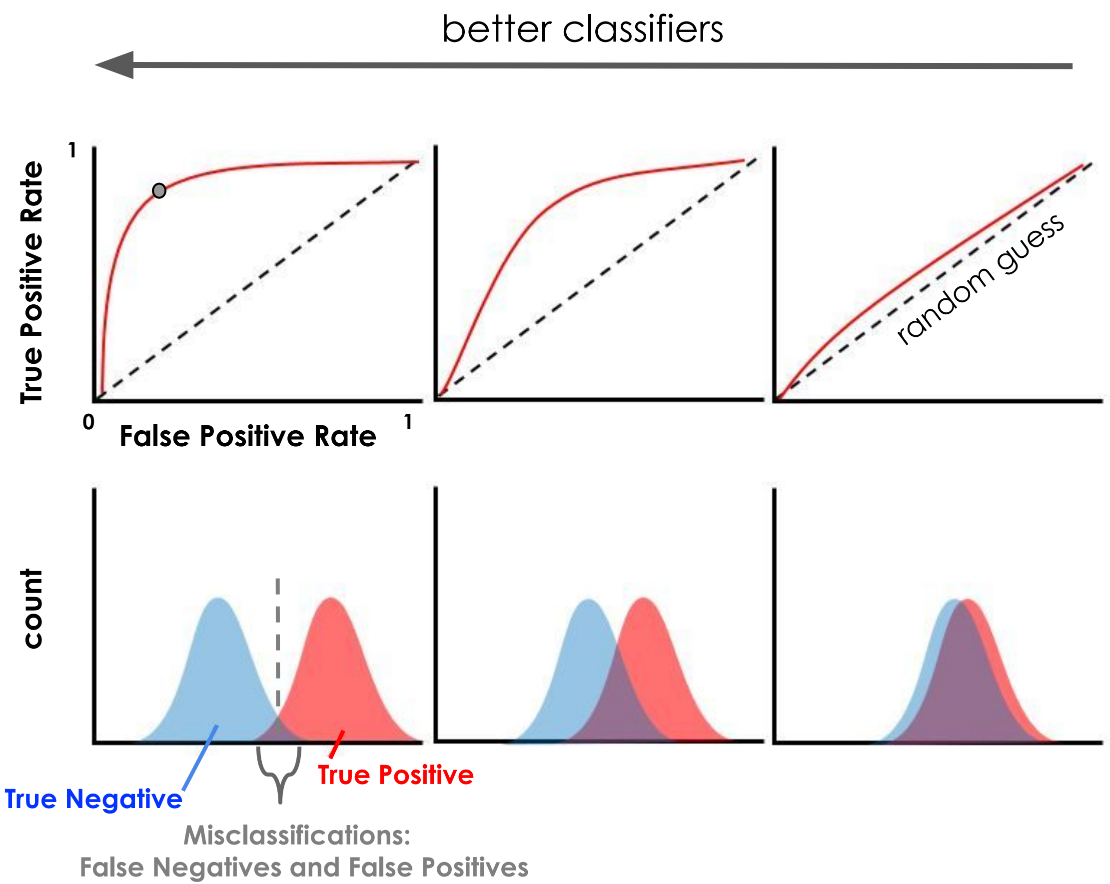
Simplified **supervised ML classification workflow**:
2D example (blue vs red)

Adapted from A. Pau et al,
Nuclear Fusion, 59(10):106017, 2019



Learned mapping, or
boundary between classes

… by counting how many times the trained classifier is right or wrong!



True Positive Rate:

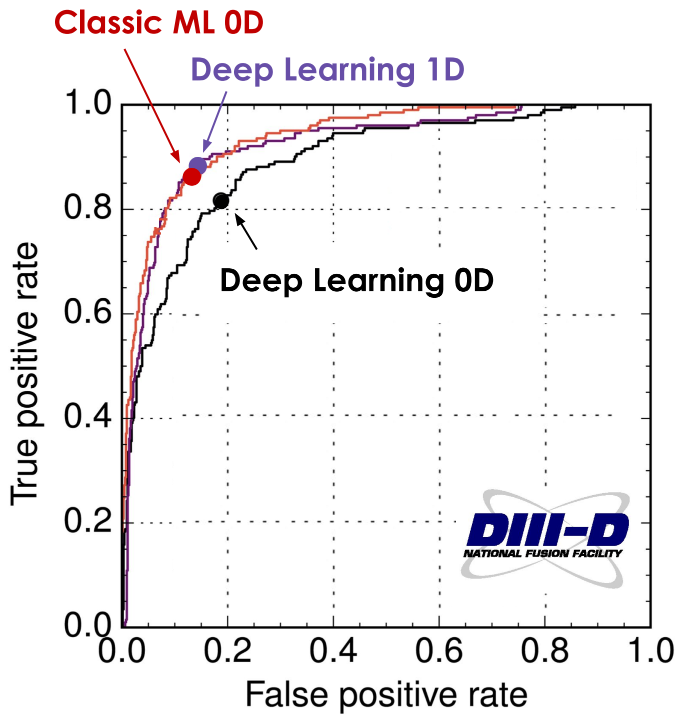
$$\frac{\text{\# correct } \textbf{positive} \text{ classifications}}{\text{total \# of } \textbf{positive} \text{ samples}}$$

False Positive Rate:

$$\frac{\text{\# wrong } \textbf{positive} \text{ classifications}}{\text{total \# } \textbf{negative} \text{ samples}}$$

ML models of varying complexity can have comparable performances

Fusion



Adapted from J. Kates-Harbeck et al., Nature 568, 526–531 (2019)

- **Rashomon Effect**:
  a multitude of models with approximately the minimum error rate exists, for many problems[1] (also in Fusion!).

  As long as a large Rashomon set exists, it is likely that some are interpretable[2,3], *maybe hard to develop*.

[1] L. Breiman et al, 2001 Statistical Science 16 199–231
[2] C. Rudin et al., 2022 Stat. Surv. 16 1–85
[3] Semenova et al, 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT'22) arXiv:1908.01755

ML models of varying complexity with comparable performances

Fusion

https://en.wikipedia.org/wiki/Rashomon

<u>Fun fact</u>:
Rashomon term inspired by 1950 Kurosawa's movie!

- **Rashomon Effect**:
  a multitude of models with approximately the minimum error rate exists, for many problems[1] (also in Fusion!).

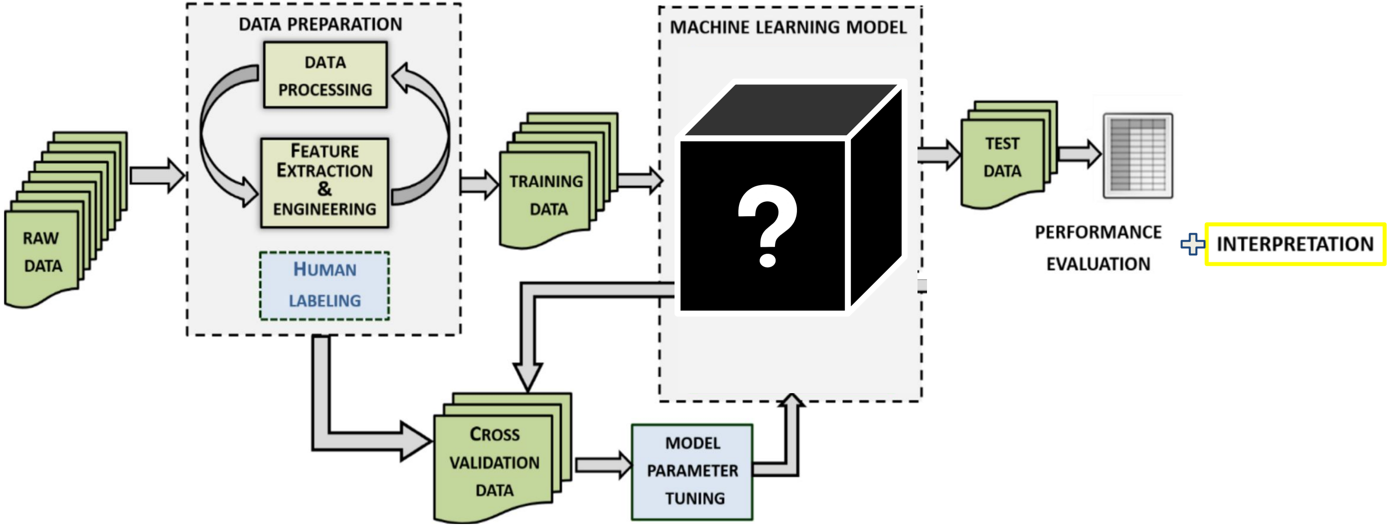
  As long as a large Rashomon set exists, it is likely that some are interpretable[2,3], *maybe hard to develop*.

[1] L. Breiman et al, 2001 Statistical Science 16 199–231
[2] C. Rudin et al., 2022 Stat. Surv. 16 1–85
[3] Semenova et al, 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT'22) arXiv:1908.01755

A simple, interpretable, and accurate model *should* exist, maybe (computationally) hard to develop

Supervised <u>and</u> interpretable ML classification workflow:

Adapted from A. Pau et al, Nuclear Fusion, 59(10):106017, 2019

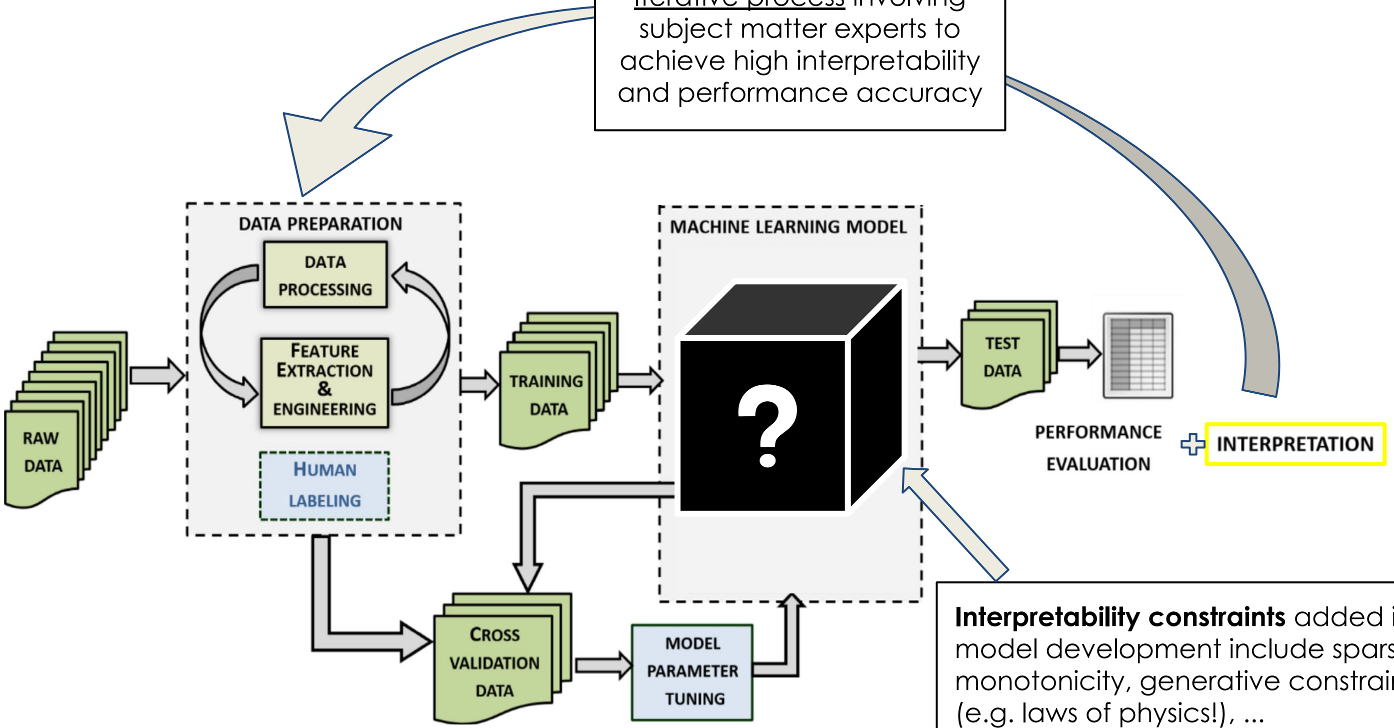


A simple, interpretable, and accurate model *should* exist, maybe (computationally) hard to develop

Supervised and interpretable ML classification workflow:

Adapted from A. Pau et al, Nuclear Fusion, 59(10):106017, 2019

Iterative process involving subject matter experts to achieve high interpretability and performance accuracy



Interpretability constraints added in model development include sparsity, monotonicity, generative constraints (e.g. laws of physics!), ...

Models interpretable by design vs black boxes that can be "explained"

existing dichotomy

Interpretable by design vs **Black boxes**

- Ensemble methods,
- Convolutional Neural Networks,
- Recurrent Neural Networks,
- Autoencoders,
- …

why did we get the output we got? → *post-hoc explainability*

explanation by

simplification → linear surrogates

feature relevance → saliency maps

additional models …

Image classification explanation through saliency maps

Test image

- Here's a dog 🐶

Image classification explanation through saliency maps

- Actually a Siberian husky!

<u>Similar</u> saliency maps to explain <u>different</u> image labels!

- Or maybe a transverse flute?
  - **Same image features relevant for different classes.**

C. Rudin, 2019 Nat. Mach. Intell. 1 206–15

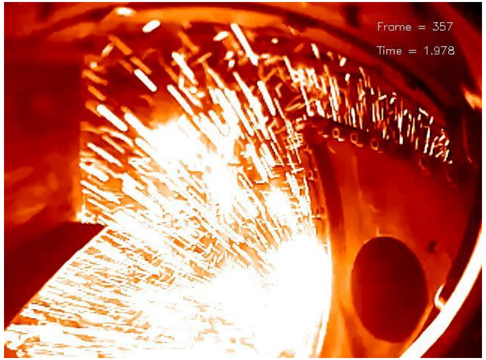
<u>Similar</u> saliency maps to explain <u>different</u> image labels!

Saliency shows **where** the network is looking, and might not convey **why** the black box predicted what it did

- Explanation for why the image contains Siberian husky is the same for why it might contain a transverse flute…
  - **Ambiguity**!

C. Rudin, 2019 Nat. Mach. Intell. 1 206–15

Tokamak disruptions challenge path to burning plasma

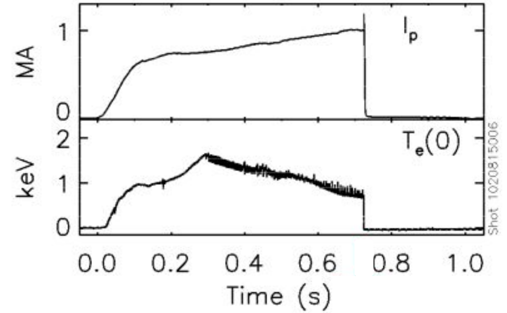


Visible camera view of RE beam hitting Alcator C-Mod first wall

JET runaway electrons damage. https://www.iter.org/newsline/-/2234

Major disruption → final **loss of control** evolving on **timescales of milliseconds**:

- Fast drop I_p leads to loss of confining poloidal field.
- Fast I_p transient causes large induced voltages, currents, forces.
- Rapid thermal losses cause surface damage.



How to take care of disruptions:

- **Accept** the damage and live with it.
- **Mitigate** the damage by injecting massive gas or shattered pellets.
- **Avoid** altogether by detecting precursors & steer plasma away from disruptive boundary.

Disruption Prediction, Avoidance, and Mitigation (DPAM)

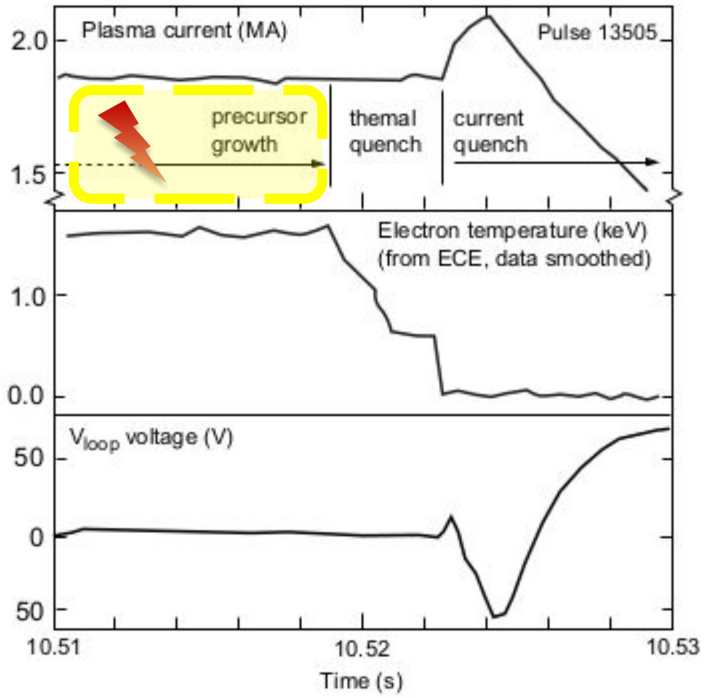
DPAM strategy mandatory when scaling to reactors.

Lehnen M. et al 2016 "Plasma disruption management in ITER", 2016 IAEA Fusion Energy Conf. EX/P6-39

Predictive algorithms need to be employed (continuously) throughout the discharge.

Mitigation, as emergency response, is triggered as last resource to mitigate disruption consequences.

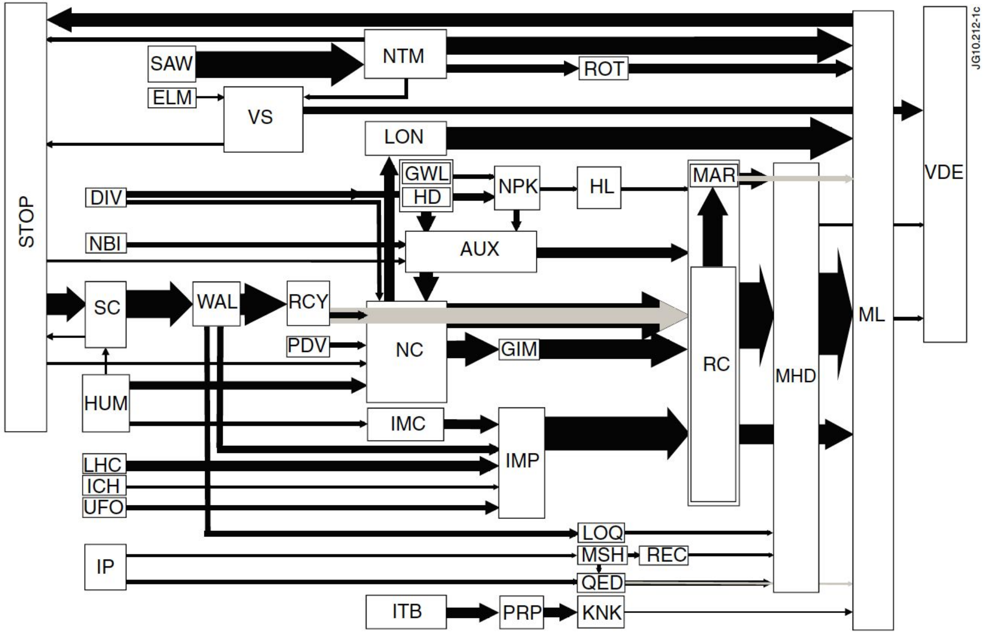
Avoidance needs timely identification of precursors' growth: <u>not an easy task</u>.



ITER Physics Expert Group on Disruptions, Plasma Control, and MHD (1999) Nucl. Fusion 39 2251

Statistical studies show complex chains of events:

possible disruptive chains of events



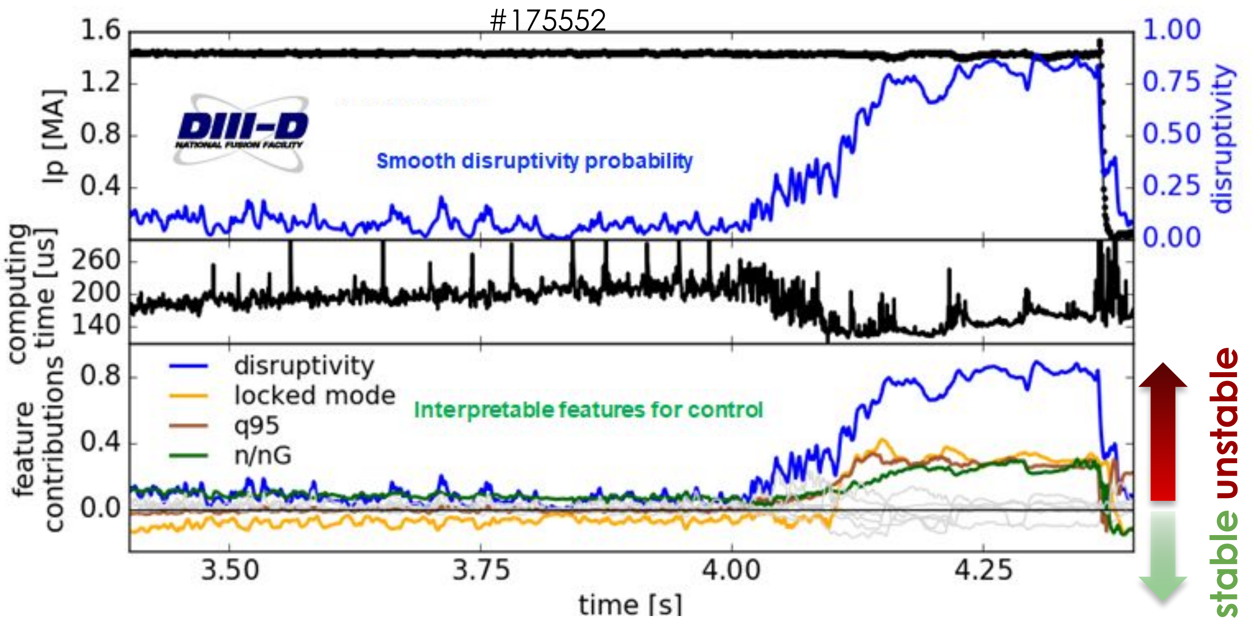
- Similar **statistical studies** not always available across different tokamaks.
- Need **timely identification of precursors** to allow the **plasma control system** (PCS) to take proper action.

Wealth of experimental data from different tokamaks enables Machine Learning applications.

Statistics of the sequence of events for ~10yrs of unintentional disruptions at JET: width of the connecting arrows is the frequency of event occurrence.

De Vries et al. NF 51 (2011) 053018 "Survey of disruption causes at JET"

Explainable ML predictions for real-time proximity to instability



caution: accuracy of local explainability metrics

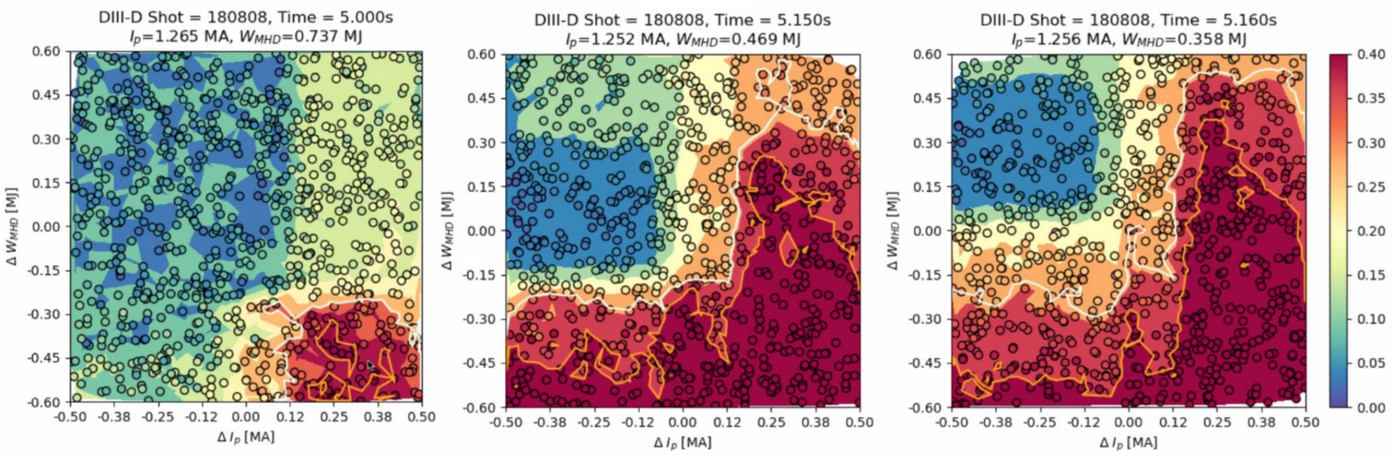
C. Rea et al, Nucl. Fusion 59 (2019) 096016
C. Rea et al, 2021 IAEA EX/P1–25,
J. Barr et al, Nucl. Fusion 61 (2021) 126019

- Identification of stability boundaries in real-time.
- Local explainability metrics leveraged inside PID controllers to modify plasma trajectory in real-time.

Identification of safe operating region through fast ML enables trajectory planning

- ML simulations evaluated by sampling from 2D operational regime variations:

time



high danger

safe region

caution: sufficiently detangled input space, robustness of sim predictor

- Goal: leverage ML-driven optimization to identify trajectory across operational space and in real-time control systems.

M. D. Boyer, C. Rea, M. Clement, Nucl. Fusion 62 (2022) 026005

DIII-D NATIONAL FUSION FACILITY

End of Part I