Parallel Session 2: "Data Collection, Reduction and Analysis"

Lamar Moore, Toby Perring, Celine Durniak, Juan Carmona Loaiza and Iary Davidzon.

Main Themes

Data Management

Trusting the process

Community Building and Engagement (Sharing the Knowledge)

Data Management

- Should we be using consistent formats?
- Standards for representing the data?
- Could this inform future experiments?
- How do we know when we can throw data away?

Data Management

Solutions

- European Open Science Cloud (EOSC) with F-indable A-ccessible I-nteroperable R-eusable guiding principles for improving data management.

Prevailing Issues

- Knowing when to throw away data in the neutron world is difficult since enough may not be known apriori to select interesting regions within the raw data.

Trusting the process

- What is going on inside the black box?
- I have model A and model B. I want the probability of A vs the probability of B vs the probability of neither...with errors!

Trusting the process

Solutions

- Uncertainties for NN can be provided using dropout techniques.
- Uncertainties for trees can be determined using NGBoost.
- SHAP (SHapley Additive exPlanations) was suggested for understanding which features contribute to a model's prediction.

Prevailing Issues

- See next slide!

Community Building

- As an outsider looking into other domains it seems as though everything is solved or at least everyone is miles ahead. How do we access this expertise?
- How do people get access to infrastructure which enables them to try out these techniques?
- How can we share ideas, techniques, scripts etc.

Community Building

Solutions

- Workshops such as these help with knowledge sharing!
- Open data should help.
- EOSC-enabled Infrastructures like EGI and EU-DAT.

Prevailing Issues

- We need more hands-on/practical workshops.

Useful Links

- European Open Science Cloud https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
- PANOSC (Photon and Neutron Open Science Cloud) www.panosc.eu
- Data Management https://www.nature.com/articles/sdata201618
- Dropout Techniques for Estimating Uncertainty https://www.cs.ox.ac.uk/people/yarin.gal/website/blog_3d801aa532c1ce.html
- SHAP Values https://github.com/slundberg/shap
- NGBoost https://stanfordmlgroup.github.io/projects/ngboost/
- European Grid Infrastructure https://www.egi.eu/services/
- European Data Infrastructure https://eudat.eu/catalogue

Questions to the panel?