Research Data Management Challenges brought about by the ubiquitous use of AI

Dr. Line Pouchard

Computational Science Initiative

Brookhaven National Laboratory

BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

BNL Operates and Supports Many Data-rich Facilities

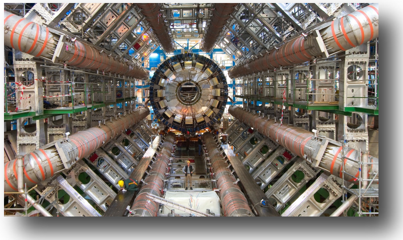
- Relativistic Heavy Ion Collider (**RHIC**)
- National Synchrotron Light Source II (**NSLS-II**)
- Center for Functional Nanomaterials (**CFN**)
- Accelerator Test Facility (**ATF**)
- LHC **ATLAS US Tier 1 Center**
- Atmospheric Radiation Measurement (**ARM**) program
- **Belle II**: Computing for B meson physics experiment
- Quantum chromodynamics (**QCD**) computing facilities for BNL, RIKEN, and U.S. QCD communities

RHIC

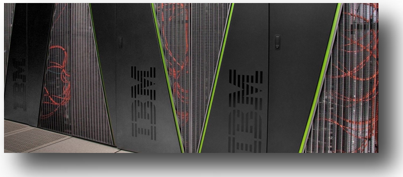
NSLS II

CFN

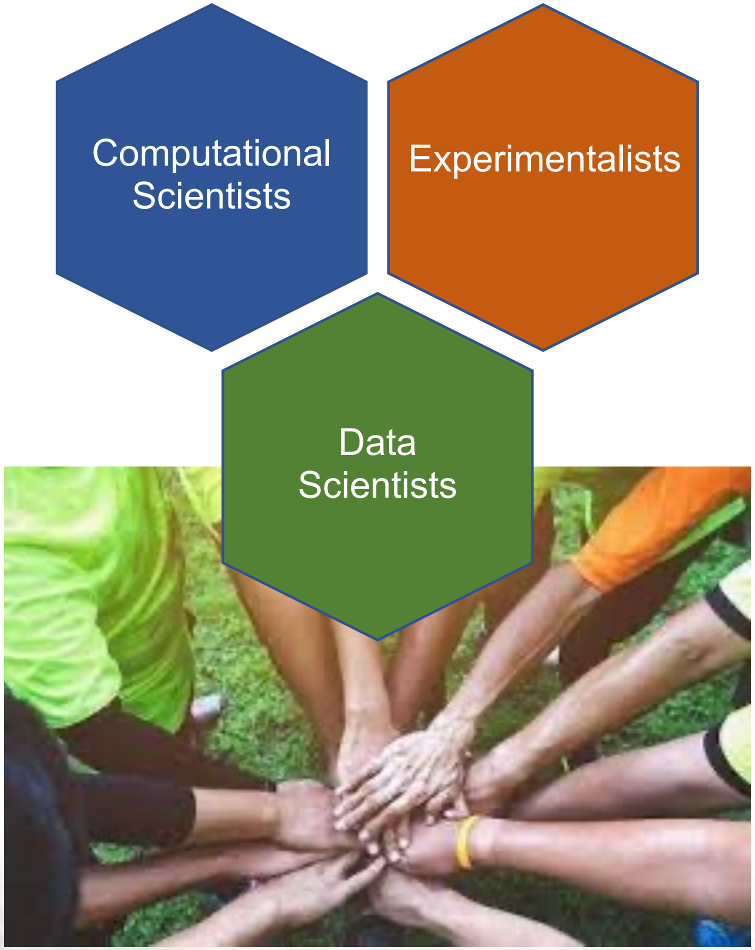
ATLAS



QCD



What does it take to build an AI application for science?



Computational Scientists

Experimentalists

Data Scientists

Identify emerging phenomena in high velocity, possibly streaming, data - Streaming Statistics, Data Mining, Machine Learning

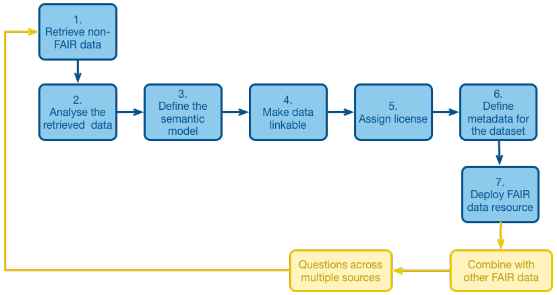
Determine what is of interest and impact, generate candidate explanations – Streaming Deductive Reasoning, Computational Models

Human-Computer collaboration to jointly adjust data collection, reasoning and insights – Science of Interaction, Cognitive Depletion Detection, Hypothesis Exchange, Adaptive Algorithms and Workflows

Evaluate the impact of possible decisions - On Demand Prediction, ML Surrogate Models

Document which decisions were taken during the analysis process to explain the results - Provenance, Explainability, Reproducibility

AI introduces new challenges to curation and RDM



1. Retrieve non-FAIR data

2. Analyse the retrieved data

3. Define the semantic model

4. Make data linkable

5. Assign license

6. Define metadata for the dataset

7. Deploy FAIR data resource

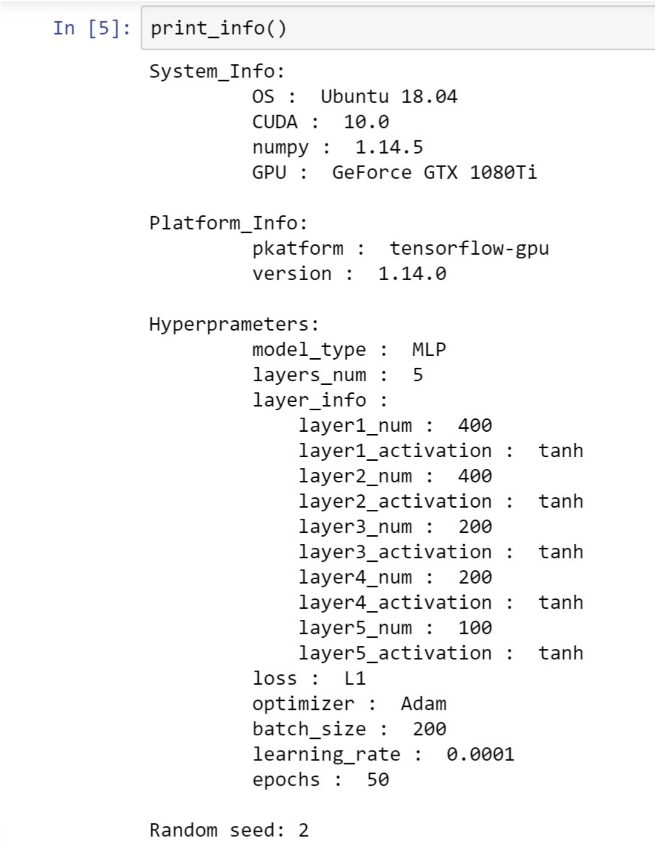
Combine with other FAIR data

Questions across multiple sources

- FAIR is designed for data, not data and software
- New criteria and/or Research Objects to trace:
  - Training sets, models, hyperparameters, calibrations
  - Size of training sets and availability are challenges
- Lack of precision and accuracy in metadata compromises quality
  - Datasets with incomplete records are/should not be used
  - Datasets with incorrect records introduce errors
- Lack of adequate datasets in some disciplines despite abundance
  - Not enough diversity of datasets in large dbs
  - Not enough diverse datasets to build robust training models

Additional challenges

- What do we store? what do we annotate?
  - training models, initialization scripts, hyperparameters, network structure, decision points – AI is a black box
- How do we annotate decision points made by a black box?
  - documentation, metadata, provenance
  - which variables affect reproducibility and replicability?
  - method, data, experiment?
- Machine Learning changing platforms and environments:
  - TensorFlow, PyTorch, LightGBM, etc.
- Availability of compute environment:
  - specialized architectures affect results
- Use of surrogate models –
  - to what extent do they provide some measure of reproducibility?



```
In [5]: print_info()

        System_Info:
                OS :  Ubuntu 18.04
                CUDA :  10.0
                numpy :  1.14.5
                GPU :  GeForce GTX 1080Ti

        Platform_Info:
                pkatform :  tensorflow-gpu
                version :  1.14.0

        Hyperprameters:
                model_type :  MLP
                layers_num :  5
                layer_info :
                    layer1_num :  400
                    layer1_activation :  tanh
                    layer2_num :  400
                    layer2_activation :  tanh
                    layer3_num :  200
                    layer3_activation :  tanh
                    layer4_num :  200
                    layer4_activation :  tanh
                    layer5_num :  100
                    layer5_activation :  tanh
                loss :  L1
                optimizer :  Adam
                batch_size :  200
                learning_rate :  0.0001
                epochs :  50

        Random seed: 2
```

U.S. DEPARTMENT OF ENERGY

Line Pouchard, PhD, Virtual Residency Workshop, June 9, 2021

BROOKHAVEN NATIONAL LABORATORY

RDM must apply to data AND software

- amount and type of compute resources:
  - increasing trend to integrate AI specific systems at HPC facilities
  - running ML codes on HPC systems
  - RDM to apply to data AND software

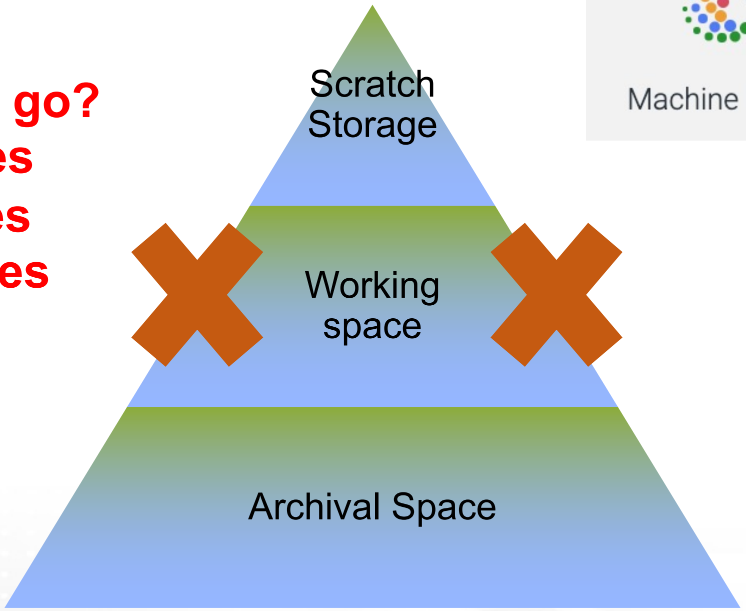
re3data.org

OpenML beta_2

Machine learning, better, together

Where will the AI provenance go?

- **specialized repositories**
- **publishers' repositories**
- **institutional repositories**



Scratch Storage

Working space

Archival Space

DLHub