ATI Workshop Data science for Physics and Astronomy

https://indico.cern.ch/e/TuringPhysics2019

General suggestions that should be captured in the white paper:

- CS has benchmark data sets (e.g. MNIST, CFAR10(0)). It would be good to ensure that appropriately curated exemplars from the STFC science field(s) should be made available. For example the Kaggle data challenge sets (Track ML, Flavour, Higgs) and anticipated astro equivalents. These would ideally have some standard interface, but at least provision in an easily readable format would be a minimum.

Group 1(a+b) - Synergies between high-energy physics & astro experiments. Data/ dimensionality reduction

Alkistis Pourtsidou (also G6), Catherine Watkinson (also G6), Conor Fitzpatrick (also G6), Caterina Doglioni (also G4/G6), Elena Cuoco, Sam Lawrence (also G3b), Catarina Alves (also G5/6), Iacopo Vivarelli, Nachiketa (also G3b,G5) Andreas Korn, Chris Lovell

[complete braindump can be found at the very end of this document]

Common problems:

- We have too much data and we cannot store it all
- Some of our data is useless now, how to get rid of it?
    - What if it's useful in the future?
    - Can we use independent sensors (HEP: control regions where you try and remove the signal as much as possible, e.g. point sensors to the sky and know where you're pointing at, to check noise levels) to measure each of those backgrounds?
        - GW does it already, same as astro to a certain degree
- Possibility of recording different "tiers" of data if you can't keep it all:
    - All raw data (necessary for GW), large events
    - No raw data (some LHC experiments), small events but can't go back
    - Needs discussed and decided in advance
        - Also because sometimes you can collect all the data, but then no resources to (re)process it...
- "Commensality problem": different configurations of the experiment are necessary for different science cases, and are sometimes incompatible
    - This happens for LHC as well but to a different level, we have a "trigger

menu" decided in advance where we choose which categories of events to take
    - Solution: identify commonalities through discussion and data challenges with pre-full-experiment (pathfinder data or simulated data)
        - Data challenges currently work at the level of the output and final physics analysis, but we are interested in the data challenge at the hardware level
    - Prioritization requires consensus-building meetings
- What about data compression?
    - One way to do it with machine learning: autoencoders?
        - Brief lit review:
            - Kramer - Nonlinear Principal Component Analysis Using Autoassociative Neural Networks
            - Variational autoencoders
            - A presentation by Eric Wulff, Caterina's student who worked on this: Overview_of_AEs
                - Thesis/github will appear here at the end of January :)
        - There are pros and cons, and differences with PCAs (already used).
        - It would be worth to try and stick different datasets in existing networks

Facilities that have to customize their data / computing centers for multiple users/experiments

Differences between HEP, astro and gravitational waves:

- HEP
    - every collision event is independent
    - To first order, an event is either signal or noise
        - Concept of "trigger" is: throw away non-interesting events
- Astro / gravitational waves:
    - time series
    - Various superimposed signal and background frequencies
    - Trigger concept comes in with follow-up instruments

Assorted notes:

Measuring the background is a lot more challenging in astro because there are many frequencies superimposed and one may not know what they are until the end of the data analysis

Glossary:

RFI = Radio Frequency Interference = noise from mobile phones, satellites…
CSP = Central Signal Processor = receives very high data rate input from antennas and pulls it together, performs the decorrelation, equivalent of "DAQ" in HEP.
SDP = Science Data Processor = generally a group (a consortium) who process the raw data into science-ready data, e.g. images for radio (for HEP: similar to reconstruction software). Now, these are merged into the science regional centers. (for HEP: grid-equivalent of tier 2s).

CDR = Consolidated Design Review

Action items:

- Construct data challenges with pathfinder data or simulated data (SKA/MeerKAT)
    - People: Alkistis [add your name here]
    - Find what has already been done (e.g. https://www.kaggle.com/c/PLAsTiCC-2018/overview for LSST)

- Test autoencoders for compression that have been tested for HEP with different data (GW)
    - People: Caterina, Elena, Michaela? [add your name here]

- Discuss PCAs more, differences with autoencoders
    - People: Conor, Alkistis [add your name here]

Things we didn't talk about yet:

Questions of HEP to astro:

- Can you switch observation mode fast enough with humans? Or do you need / can you trust algorithms (ML)?
- Do you have a latency problem? How fast do you have data analysis in, e.g. for transients?
- Do you have software/toolkit catalogues (within experiment and outside it)
- How do you deal with monitoring when observation time is limited: how are you sure that you have everything you need working correctly at the start of the observation?

Additional notes - not discussed:

- E.g. BBC are actively working on data compression for streaming; so industry may have some insight for certain types of data. [AB]

Group 2 - Generative models and data augmentation

Gordon Yip, Davide Piras (also G10 and G13), Conor Sheehan, William McCorkindale (also 1b/3b/6), Andrew Patterson, +Nikos Nikolaou, Luke Conaboy, Stephen Menary, Darren Price, Adrian

When do we need generative methods?

Generate data if you do not have a lot of them - happens often for computational expensive data generation processes.
Clean the data (example with images of stars): using a generative method can shed light about what is important in the generation of an image.
Learn something invariant within the network, about the physics of your problem (link to explainability with Grad-CAM; [*AB added a note this alg is one of many - rarely consistent*

output from these different algorithms])
Applicable to stars, galaxies, dark matter haloes; also, to particle physics events, especially in the last step of the generation (they can help in simulating the physics we expect); also, to quantum technology, by learning the wave function of your device using a generative algorithm.

What generative methods are available?

- Generative Adversarial Networks (GANs, with the variations WGAN, WGAN-GP). They are good from the visual perception point of view, but bad in terms of training stability. Works for very simple problems in terms of parameter space (low number of parameters, in combination with the labels - CGANs), even with large training set. Problems include mode collapse, and mainly training not perfectly converging; all in all, quite frustrating.
- Other track is Auto-Encoders (AE): starts with dimensionality reduction, with an output layer as big as the input one. Project from real space to lower dimensional space (link to data compression), and learn a lower representation of your data. Variational Auto-Encoders can be used to generate data, building on top of the AE; training is more stable, but output tends to be blurry.
- In general, other problems include “checkerboard effect”, general problem in convolutional neural networks actually. Solution is explained in this blog post.
- Bayesian networks can be an alternative, as well as probabilistic graph modelling. They do not use neural networks, but are quite computational expensive - especially for images. Nikos can maybe link some more paper or resources about this. Norman Fenton has a book on Bayesian Networks and Risk Analysis (“Risk Assessment and Decision Analysis with Bayesian Networks”).

Are they reliable enough (less than 1% accuracy)?

In short, no, but that depends on the network complexity and amount of training data available. A more complex model may be needed, but it could be easily overfit when dealing with low amount of data. Universities maybe do not have the computational resources vs e.g. Google and other big companies to tackle this problem. If you can’t beat them, join them.

How can we evaluate them and how do you quantify the uncertainty?

We dove have evaluation metrics, beyond visual inspection - this is definitely a plus with respect to, e.g., images of cats and dogs, which are usually employed in computer science examples.
Loss function is important, but how do we go beyond L2 loss (MSE)? Earth-mover distance is the one actually used in WGAN.
Depends on problem as well.

How do you deal with such big (and low amount of) data?

You can augment your dataset, by injecting noise in images; or, you can use low quality images (e.g. obtained by the GAN itself); else, you can apply general transformations to which you know the problem is symmetric, e.g. rotations or

translations - but that is not always going to work.
You can also lower the dimensionality of your data.
Another approach is using transfer learning! Ideas coming from NLP and style-transfer, which works even with low number of data points.

Summary: **we are all trying to simulate samples of an experiment or a simulation, as a function of a different set of parameters. The low amount of data, and the required high complexity of the problem, are the main hurdles.**

Outlook: **we all rely on having a big dataset of high-quality data for doing generative methods; but that is not always available. Possible way forward include:**

- **improve hardware (TPUs, bigger facilities) - poses an environmental challenge, though, and may take a while;**
- **wait for new algorithms to come out of pure ML research;**
- **use statistics and physics assumption to ease your problem, and to accurately evaluate the performance of the algorithm (we want the error bar on the results);**
- **create some shared high-quality dataset, on which pre-train models (example could be the Quijote simulations), in the same way there is a pre-trained net from the ImageNet dataset.**

A general point of fear is the bias some results may show, at high-order statistics or even with very simple problems (quantum example).

Group 3a - Dependence on ML algorithm choice and architectures

Jonathan Holdship, Katie, Ofer, Andreas Korn,Tom Stevenson, Jeyan, Ilian Iliev, Gleb, Fabrizio Salvatore, Adrian

Learning the model or architecture that works for your problem

- Repository of algorithms for specific types of problems
- Prevent community moving on without you
- Rules of thumb for architecture/hyperparameter choices

How much knowledge do you put in?

- How does type of problem decide this?
- Don't want to overfit to simulations
- Might not have enough data to feature select

Feature Importance

- Experimental approach, remove parameters

- Linearize activation functions to better trace how they contribute
- Can you train random forest and neural network? Then report importance from random forest
    - Can hit a problem if the models treat features very differently
    - Can compare distributions of features over whole dataset and only disagreements between models

How much does a blackbox matter?

- If you just want the answer - who cares?

Testing

- How do you check you covered your parameter space properly
- How do you evaluate errors over the parameter space

Interpretability

- Better visualizations
- How was decision reached
- What is the "noise" preventing clear classification? Are we able to identify features that are confusion model?

Human Learning

- How do we learn new physics from a machine learning model
- Models can learn things like Newtonian mechanics from scratch
- Question is then how do we extract that as human knowledge
- They're weights in a network but we want a principle that we can teach people
- What happens when the thing learned is beyond human capacity?

Group 3b: *explainable AI & interpretability and data visualisation (incl. ethics)*

Adrian, Heather, Darren Price, William McCorkindale (also 1b/2/3b), Roberto Trotta, Nachiketa, Joe, David Berman, Tom Stevenson, Sarah Jaffa(3b/4/6), Jeyan, Jonathan, Andreas

Raised topics:

- Poking inside, e.g. looking at information in hidden layers and then interpreting that information
- feeding in features, higher level variables vs. letting neural nets make these determinations (affected by training sample?)
    - If it’s a medical device, it needs to be *explained*
- information loss after putting it through a system

- hard to come up with toy models: we need to be able to understand them but also have them be representative of the actual problem
- explainability
- --Conceptual problems need a way to be physically realizable
- How would a clinician explain how a decision was made using these algorithms?

Establish: what are the characteristics that a system ought to have? Explainability, interpretability, etc etc

A model should be able to provide an uncertainty on its results

Should be able to characterise the stability of a model (e.g. bootstrapping by varying inputs and hoping for small/no output variation)

Common sense outcomes? Any problem should have some trivial test points, extrema (unit tests)

Or adding IN a variable that you know is irrelevant, going in, and making sure it remains irrelevant

Why are we inherently untrustful of neural networks (black boxes in general)

e.g. we trust linear differential equations, so there is a natural extension to non-linear, even though we might not *understand.*

Integrated gradient weights - to determine importance of variables in NN
We want to know why a feature was picked in some circumstances which can be an issue if we don't understand that hidden layer

Uncertainties in input data -- does architecture have awareness of our own trustworthiness on the data?
-- trust in results related to trust in features/ input data

Bias arising from irreproducible results: pre-determined statistical framework to determine "trustworthiness" of model, so as to not fall victim to playing with model until it fits expectations (e.g. with variable importances…)

What **is** trustworthiness? (can resolve by same as above? Some statistical interpretation)

5sigma vs. p-value of 0.05…. Are we just over-protecting ourselves?

Wrap-up:

Online repository with a cheat-sheet for models

model X - common problems - useful solutions

but also granularity for different uses of algorithms

Included in the turing way? Or with a similar style of documentation

The theme of trust

When we let an algorithm determine importance, it's less "understandable" - but tend to be

more trusting when we put in higher-level information
Explainability/interpretability means different things to different people (and _needs are different)

Sources of untrustworthiness

- Bias arising from irreproducible results: pre-determined statistical framework to determine “trustworthiness” of model, so as to not fall victim to playing with model until it fits expectations (e.g. with variable importances…)
- How stable is a model - need to characterise/ensure stability, robustness against small perturbations in training/inputs/number of inputs (validation)
- Input data uncertainties? ← does the model know? (do _we_ know?)
- Over-training, cherry-picking of analysis methodologies post-facto

Ethics

Link into reproducibility discussions

Unintended consequences of open source / public algorithms
(misuse, ethical considerations, use beyond limits of applicability)

Can we live with something we don’t fully understand.

Link to software catalogue for Astrophysics and Particle physics in the ESCAPE framework:
https://projectescape.eu/services/escape-software-repository

More in general, if you want to read more about ESCAPE:
https://projectescape.eu/

As a side note, quantification and possible reduction of environmental impact of data science. Raise awareness about this.

Group 4 - Reproducible Big Data

Chris Lovell, Sarah Jaffa(3b/4/6), Emma Slade, Darren Price, Caterina Doglioni (also G1/G6)

Barriers to entry (version control, continuous integration etc.) → culture change and also link into training tomorrow

Domain jargon translation: are materials written by one community accessible by others?
Turing Way : open source, add “translations”?
https://the-turing-way.netlify.com

‘Stories’ to direct readers of the Turing Way for particular use cases, e.g. ‘My supervisor doesn’t believe in source control’

Funding for a central repository for source control?

Github large file storage, Google drive, Azure links to github (temperature: hot data is accessed frequently, cold data is not)...

Can existing infrastructures for data outputs be supported or broadened (things like hepdata https://www.hepdata.net/ and rivet https://rivet.hepforge.org/ ?) (data outputs and data analysis records are linked)
Not all experimental analyses can be re-used: even LHC data from 5+ years ago cannot be fully used. Point to good practice and expand support and expectations that data+code versioned is released publicly.

What data outputs can and should be preserved to ensure reproducibility? (Not always feasible to record/store long-term all data)

Containerisation (link code + data available publicly with computational environment)

Journal submission, credit for open source research (JOSS, volunteer based peer review)

Private repos - good or bad?(!) Good - nobody cares about your code. Bad - good for early career researchers that might be low on confidence, can still share with individuals

ArXiv a good example - what does an arXiv for code look like? (github?) can it be better? Durham has an open source text that should go into the end of a paper, and pressure to make the code open source.

Showing open GitHub repos in job interviews is advantageous

Plan

- Funding Councils
    - E.g. STFC already require that research funded by their grants is made open source - can we include reproducible in that too
    - Is there an Open Source/Reproducibility Officer within STFC that can educate and encourage PIs/Supervisors on best tools and practices for open source/reproducibility
    - Provide a 1 day course that can be disseminated through Doctoral Training Centres
        - Crucial topics: version control, continuous integration and containerization
        - Content from The Turing Way?
- PIs/Supervisors
    - Encouraging students to learn reproducible practices saves time and stress when in “code hell” scenarios or when reviewers request re-runs of analyses
    - Produce “good problem solvers with good quality code” instead of “good problem solvers with good enough code” - increases probability of students passing and being successful for jobs

- Students/ECRs
    - They already want to learn this stuff!
    - Being able to show public examples of good quality, tested, robust code in an interview is really advantageous
    - Being (as) reproducible (as possible) from the beginning will save you so much time over the course of your career

Group 5 - Deep Learning (also for Time Series)

Nachiketa (also G1+b,G3b), Ingo Waldmann (also G6),Nikos Nikolaou (also G2), Omar Jahangir, Darren Price, Maritza Soto, Elena Cuoco, Andre Freitas

Notes from discussion

Big Picture
Problems are common to different types of data (and questions - forecasting, classification)
Time-Series - We love LSTMs or do we have a choice !!

1. How do we avoid overfitting the signal (transits) - low to medium S/N ?
   a. Preprocessing is key - find the right basis for decomposition (example)
   b. This depends upon the dataset - what works for the bank may not work for stars and galaxies

2. Can we use simulated data to train ?
   Depends upon what is the question asked ?
   Time Series
   We can mathematically encode certain properties (PDF, PSD)
   a. https://arxiv.org/abs/1803.09933
   b. https://ui.adsabs.harvard.edu/abs/1995A%26A...300..707T/abstract (Gaussian, Power-Law Noise)
   c. https://ui.adsabs.harvard.edu/abs/2013MNRAS.433..907E/abstract (more general PDF very expensive / slow - many parameters and therefore convergence needs testing)
   Particle showers
   Challenge - tails
   a. GEANT simulations are very good but computationally expensive
   b. GANs and autoencoders may not provide proper representation
   Images
   Classification done “routinely” (example - DeepCEE)
- Variability in image classes

https://towardsdatascience.com/deep-learning-for-image-classification-why-its-challenging-where-we-ve-been-and-what-s-next-93b56948fcef

https://ui.adsabs.harvard.edu/abs/2018CQGra..35i5016R/abstract (Imaged based DL in noise Transients in GW)

Proposed approaches
- propose extensive testing with simulations (example)
- look at variety of datasets (different domains)
- explore overlap with transfer learning
- explainability - medical community is interested

3. Encode physics into the loss function - force the physical constraints
   - take cues from NLP
   - take cues from Materials/Chemistry
4. Can we perform sensitivity tests ? Sensitivity to network architecture and data ?
5. How do we gauge complexity and "match" the complexity of the forecasting method (eg network) ? How to express complexity as a mathematical function ?
   - preprocessing - dimensionality reduction (eg images)
   - open
   - topological data analysis

Way forward - **Topics for White pape**r on challenges in deep learning

1. Low S/N ratio overfitting - preprocessing is key example - Banking Turing Data Group
2. Encoding physics or constraints into the ML algorithm
   a. Find ways to encode in loss function
   b. Physics motivated features which "must" be predicted / surrogate data (physics based high dim surrogate modeling, Label-free Physics Supervision)
3. Quantify complexity of data and find appropriate (matching complexity) deep learning frameworks

Specific notes / references

- Exoplanets - done parametrically ; use LSTMs - problem overtraining
Gaussian processes - not fast
- Deep Auto Regressive network - Amazon
- ***S/N overfitting problem - Find the correct basis for decomposition*** Fourier decomposition / Hermite decomposition (Hermite model - generalisation of Poisson but probability is not preserved, Finance - fitting N coefficients) (ref : https://zenodo.org/record/2557809#.XeZAWpP7SRs)
- O(N J^2) - J no. of terms to add - Celerite - Mckee
https://celerite.readthedocs.io/en/stable/
https://gpytorch.ai/
https://arxiv.org/abs/1806.08305
- Ben Pope et al., - GAN
- Trained loss functions => insensitive to choice
- LSTMs for characters - write thesis
- when will the WIMP interact with XENON

Group 6 - Unsupervised learning and anomaly detection

William McCorkindale (also 1b/2/3b), *Catarina Alves (also G1/5),* Sotiria, *Conor Fitzpatrick*, Ingo Waldmann (also G5), *Darren Price,* Omar Jahangir, Gordon Yip(G2/G6), Marcella Bona, Adrian, Caterina Doglioni (also G1/G4), Sarah Jaffa(3b/4/6), Heather Russell, Jonathan Holdship, Andrew Patterson

Ordered notes (braindump at the end):

Definitions of anomaly detection not always consistent across fields.
Anomalies can be wanted (a new rare process) or unwanted (a system glitch, bad data, a transaction that should not happen).

In general, anomaly detection means spotting patterns that deviate from the norm —> one needs a ”history” (training sample) to start with.

Discussed autoencoders as anomaly detection algorithms for unsupervised learning.
Reconstruct through a ”bottleneck”

Example in high energy physics from this talk:
- want to spot collision events that leave a signature in the detector that is different with respect to our “background” and comes from a new physics process
- method: autoencoders (or others)
- challenge: distinguishing new physics from detector glitches
- what would be useful: set of rules for understanding why something was flagged that is learned from the system

Example in cosmology:
- want to spot rare classes of transients (supernova, kilonova, crazy stuff that we don’t know if it exists)
- method: iterative classification, run classifier and get reconstruction error, maybe it’s because your autoencoder was not encapsulating all the variability of your data. Select those that were “not too bad” (criterion: reconstruction error), and feed back into the training.
- challenge: will deal with 10k events per night, we need to find which ones are rare enough and worth looking at them for the next instrument

Next steps:

HEP people talk to each other about unsupervised searches in ATLAS :)

Darkmachines unsupervised searches: anyone interested?
https://darkmachines.org → look for **Collider searches and unsupervised: or supervised or not-yet-thought-off learning**

Conference in 2020: https://indico.cern.ch/event/852857/overview

LHC olympics dataset: https://indico.cern.ch/event/809820/page/19002-lhcolympics2020
LSST PLAsTiCC dataset (warning: time-ordered series):
https://www.kaggle.com/c/PLAsTiCC-2018/

Are there other datasets out there?

GW real data:
https://www.gw-openscience.org/about/
https://www.kaggle.com/elenacuoco/the-gravitational-waves-discovery-data
https://www.kaggle.com/tentotheminus9/gravity-spy-gravitational-waves

- Dataset from DarkMachines will be dumped on CERN Open Data page.

Could we do something for the IPA conference (with satellite meeting: https://indico.cern.ch/event/862409/) - https://indico.cern.ch/event/837621/

Group 7: Inference and regression (“modelling”)

Marcella Bona, Serena (also G3a), Benjamin Joachimi, Alkistis Pourtsidou, Catherine Watkinson, Andre Freitas, Roberto Trotta

- Gaussian process -> built-in uncertainties
- Benjamin: will get some literatures for Marcella :)
- Go for a physical model as much as possible
- Connecting diverse datasets to all possible model configurations
- Number of model parameters - which inference method to use?
- Marginalisation techniques: calibration/detector data -> informative priors (Jeffreys priors, flat priors, Gaussian uncertainties)
  -> analytic marginalisation -> Laplace approximation (Gaussian) -> integration
- Forward modelling vs. inverse problems
- Challenges: many-parameter numerical inference; forward-modelling with large parameter spaces; likelihood-free inference / empirical likelihood

Group 7: day2 -> Inference and regression (“modelling”) (after coffee, Enigma)

- ➢ Hamiltonian Monte Carlo -> stan and pyMC3 and
- ➢ PySTAN https://towardsdatascience.com/an-introduction-to-bayesian-inference-in-pystan-c27078e58d53
- ➢ Approximate Bayesian Computing (ABC)

Symbolic Learning

- ➢ https://github.com/deepmind/graph_nets
- ➢ https://towardsdatascience.com/introduction-to-bayesian-networks-81031eeed94e
- ➢ Bayesian Network (DAGs)
  -https://www.sciencedirect.com/topics/computer-science/bayesian-networks
  - Jon Williamson (health warning - initial qualitative discussions go on and on !) http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.138.2670&rep=rep1&type=pdf
- ➢ Causality - Principal reference : http://bayes.cs.ucla.edu/PRIMER/

Group 8: Transfer learning

Sotiria Fotopoulou, Nachiketa (also G1+b,G3b,G5), William Mc, Difu Shi, Andre Freitas

Algorithm architecture

- Neural networks
  - Pre-train, replace end layers
  - Time-series - transfering models using physics knowledge / constraints
- Unsupervised learning: limited to same problem but with different data; use trained model as prior, e.g. number of clusters and rough centers
- Decision trees?

Limited to subset of problems:

- input dimensions must stay the same
- predictive property must not be too different
- assumes first half of model learns some statistically useful hidden representation of the data

Applications

- Map imaging from one survey to another, e.g. galaxy morphology SDSS -> DES
- Transfer between photometric systems
- Liaise between models and data
- Airport scanners? (Transfer between data from different airports or systems etc.)

Successful example - nuclear fusion paper 2018
Failure example - ImageNet to X-ray Galaxy Clusters

Is there a way to mathematically quantify ‘transferability’?
Reports of failed applications would be useful!

Papers
Galaxy morphologies
Fusion experiments

Algorithms

PySTAN
https://towardsdatascience.com/an-introduction-to-bayesian-inference-in-pystan-c27078e58d53

Group 9: Inhomogeneous and imbalanced data

Questions/Problems

- Weak signal in a big background
- Perhaps simulations are not echoing what happens as it may be too homogeneous eg will produce similar amounts of signal and background samples
- We know unbalances exist but not knowing where they come from hurts our simulations
- Easier to correct these inhomogeneities in astro
    - Took spectroscopic images and saw which ones did not have much data etc
    - Can use mapping to find where these holes are: ‘mapping the unknown’ -> self organising maps
        - Dimensionality reduction without something like PCA
- Medical has one of the messiest datasets and unbalanced
    - Data not shared between GPs and hospitals so there are multiple data streams that don’t necessarily interact with each other
    - Lots of diagnostic info is codified but there are multiple opinions from doctors confusing the data
- HEP is purely statistical: we don’t know when an event is Higgs unless we work on the test statistics found
- In principle HEP our simulations mirror our actual processes
    - Idea of where our particle interacted and what energy it had
    - Go through steps of stripping away layers to find what kind of particles we might have eg cutting, logical reasoning etc

- (1 vs all) vs (1 vs many): usually 1 vs all works better
- If data is inhomogenous how do we deal with it: do we take a small data set or consider everything
    - when there are gaps in the data can use gaussian processes and Regression
    - adaptive models
    - we can simulate from the fitted gaussian processes and fill in the gaps
- doesn't necessarily rely on model being paramaterized
- classifying is fine but making inferences can be difficult

Wrap Up

- Need better data!
- Gaussian processes can be used in order to fill in the gaps
- Using self organized maps can help in astro
- Sharing data can help with filling in those gaps

Group 10: Tension/significance/consistency/decision theory

Questions and points of discussion:

- How do you define tension, the significance of it?
- How do you make your analysis robust - blinding?
- Particle physics more frequentist approach, Astro mostly Bayesian
- Importance of priors and importance of systematics
- 2-point systematics problem in particle. Different choices give very different results.
- Theoretical systematics very similar between astro and hep because of use of perturbation theory!
- How does one decide on significance if the parameter space is very large? HEP often uses CLS. Q: What does high dimensionality mean in astro? E.g. size of data vector 200, and 10 parameters. Works well with Gaussian distributions, but can get nasty...very tricky -numerically- with correlated data. PyMC3 and STAN useful for this analysis.

- HEP: generally accepted statistical framework, so everyone agrees on what is meant by x significance
- Avoid reporting single numbers; provide pdfs, at least for back-up
- Cosmology: example of Hubble constant discrepancy
- Prior dependence: both HEP and Astro use informative priors from previous experiments and theory; no such thing as uninformative prior (e.g. neutrino mass hierarchy)
- Galaxy evolution: many choices to be made, only get a qualitative idea of scatter by trying a selection of options; similar problem in perturbation theory in both HEP and cosmology
- ATLAS: Statistics Committee oversees stats choices in ATLAS analysis and checks if they are sound
- Both HEP and cosmology blind - different strictness and approaches to blinding; implications for computational cost

Conclusions:

- Challenge: robustness against analysis choices and analyst's beliefs -> blinding
- Challenge: how to quantify or marginalise over theory uncertainty when that

uncertainty is not well characterised or parameterisable?

- You can only assess tension/significance if you have reliable error bars -> important for industry applications: quantify statistical uncertainty
- No one-fits-all for tension/significance analysis; provide pdfs rather than just single numbers where possible

Group 11: Data science training solutions (plenary)

Discussion on CDTs

- Est 120 CDT participants for Schools run by UCL (similar for Sussex).
- Combination of industry and academic lectures + hands on sessions.
- Industry style varied significantly.
- Industry contribution valued by participants.
- Broad scope; would have benefited from having more focus
- Kavli school is 6 weeks long that involved industry and academia lectures; followed by an extended data camp; the aim being to solve a problem and publish some results.
- Sussex school had more freedom; the topics one could study with choice was good as participants can opt to study something “useful”. Downside comes from potential clashes.
- Archival is an important aspect. Professional recordings important; we need to understand how to fund this.
- Pre-learning with recordings? + more focused hands-on in person?
- Hands on extended work at the end of a training session is valuable to bring together the topics discussed and ensure that participants can come together and test their understanding.

Side topic, from a training perspective; having community experts who were available to go to for advice on how to do the right thing. The issue of research software engineer has been recognised, but how does the career progression for these people work, how to fit in with metrics that the community need to work with?

Sometimes having a broader remit could help in cognition. This should be a requirement.

Group 12: Industry/Public stakeholder engagement

- No link between insight and how to put things into practice
- If not in a teaching hospital, the link with academia is not there
- NHS has launched an AI initiative but in practice how is that communicated to

clinicians? What are the results?

- Officially a clinician should be able to get data from GP but in reality this doesn't happen
    - There is not a trusted person that can share this data in the community; there is no sharing agreement
- Structural changes happen every year because of the politicised nature of the NHS
- Every council has their own data analytics strategy
- Who do I go and ask for this data
- Have had to develop a common lexicon between research and the commercial data community
- CDTs offer work-place placements
- For a lot of what we do, data is inputted by practitioners and experts
    - This data is then centrally collected
    - Most of what is done with it is billing
- There doesn't seem to be people using the data again to improve data collection or modelling
- The biggest thing is python which the nhs doesn't use
- Data collection to have a look at how the system is utilized doesn't get fed into that same system to improve it or make changes
- What are the ethical problems associated with using AI in the medical realm?
- We want to be able to share data as that is where the interest comes from
    - Linking data from different facets of a hospital (maternity, orthopedics etc) is something that hospitals just don't do
- High dim feature space that is sparsely populated->patient data
- Python being taught to many scientists now
- Psychological Science Accelerator is trying to achieve similar things, and they are looking to (working with) high energy physicists: https://psysciacc.org/people/
- Medical data curation is a problem: messy and not standardized
- Curation at the moment is very labour intensive
- Adopting a Turing approach in industry in order to get an improvement value
- Data is shared from science as it's publically funded
- We need to have databases that we can query or get data from
    - There's no point in using excel if that's all you're using
    - We don't have to use AI but we do absolutely need clean data and tidy databases
- Commercial realm is better at visualization
- There needs to be a way of accessing data but also visualizing it because that's what humans need
- Instilling good curation of data from day 1 is a valuable skill

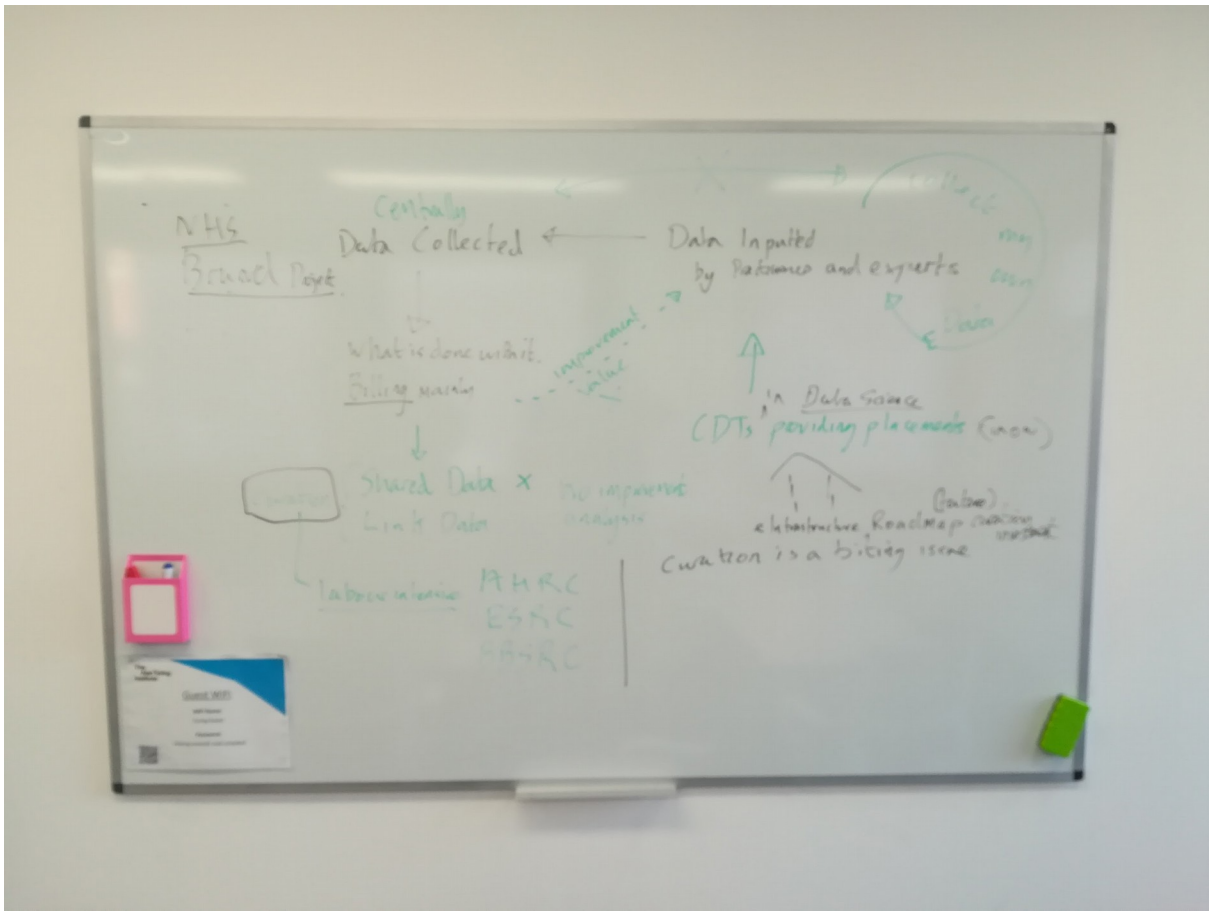

Figure 1: Sketch of engagement with various stakeholders in a cycle of working (both academic and industry development).

Group 13: Data science implementation (code & hardware challenges)

Heather Russell, Andreas Korn, Andrew Patterson, Catarina Alves, Davide Piras (also G2 and G10), Sarah Jaffa, Gordon Yip, Conor Sheehan, John Holdship

notes:

N-body simulations - gadget(2,3,4), open source, C(++?) ⇒ outputs into tensorflow/python

Almost all python - with the exception of TMVA (domain specific code) in c++

Hardware? GPUs, TPUs, FPGAs…

Google collab: jupyter notebook hosted with access to GPUs and TPUs - but there are memory and time limits.

Access to terabytes of storage + week-long computing time?

- Amazon cloud, can bid for time/space. Amazon also has a scientific computing

division, so some support.

- nvidia - can apply for a GPU (supervisor/PI has to apply (can get one GPU per project) not PhD students)
- Can apply to dirac (Talk to Gordon's supervisor, Jeremy Yates)

How to know when and where hardware is relevant? GPUs/TPUs during training. If you need to process lots of data with a trained algorithm, can use an FPGA.

If you have a problem, talk with the community! Someone will have encountered it and solved it already, e.g. too long to train -> use gpu.

Can take output parameters from a trained neural net and import them into a different solution -> e.g. onto an FPGA

Exascale HPCs - calls for use-cases?
Long-term investment in resources+infrastructure so that our projects have long-term support (rather than relying on transient commercial products)
Astrophysical fluid dynamics models? Cosmology/galaxy/star formation/planet formation simulations. (Planets are the current "buzzword" in astro hydrodynamics simulations).
Detector simulation?

Need strategies for evolution of processing frameworks, etc.

Azure: easy to learn tool for machine learning
https://azure.microsoft.com/en-us/services/machine-learning/
Makes it easy to visualize the steps, and it makes suggestions
Useful for learning, communicating

Software:

Are there benefits to having dedicated software engineers on specific projects vs. postdocs assigned to specific projects?
But as a postdoc if you spend too much time on coding/learning enough to be this dedicated software person, then you end up forced out -- two jobs with differently aligned goals and outputs
Bid for person-time from a dedicated staff member? (National pools of RSEs / following Turing model)
Continuity + documentation are important for any large codebase
Needs both RSEs and postdocs for differently scoped projects
And RSEs who understand how code would need to be written

Braindump of group 1

Topic of discussion: how do we all do data reduction? Start from examples.

Q [astro]: At which point of the data reduction do you choose, in HEP?
A [HEP]: all the time, at any point, any chunk…

[astro] This is a highest priority stuff we want to do with SKA. Some of the data reduction has to be done on the fly, in close-to-real-time because we simply don't have enough storage.

Q [grav]: are you only interested in transients? If so, how can you throw away data?

A [astro]: we don't, but if we had frequency channels that were dominated by RFI (RFI = Radio Frequency Interference = noise from mobile phones, satellites…)
we would get rid of those and we would like to do that on the fly. What we do is keep the chunk of data in its entirety, then downweight in the data analysis, but you still keep them. This is not necessarily sustainable when you record petabytes and petabytes…

[astro] time and frequency averaging of the data needs RFI-removal done beforehand.

[grav] are you throwing away vs cleaning data?

[astro] RFI too messy to remove without also removing the data.

[grav] independent sensors? We can acquire independently noise with sensors that e.g. just monitor the power supply that introduces noise.

[astro] we don't understand instruments as well as others, but we could try this eventually.

[grav] kinds of noise: EM, seismic…each measures things separately. We have a short signal in time, and we throw out where there isn't enough signal.

[astro/HEP] in pulsar/timing, is there a method with which you can reduce the quantity of data that only produces the result you want? Eg sample a frequency range +/- of this, then monitor for a shift?

[astro] this is not an event, other things can happen overlayed on the same frequency. This is like a 3D dataset. This is like in SKA, because there are many many goals. "Commensality problem" = same observation that does pulsars && other things. So you can't easily tell one group that they will not do their observation because someone else will do theirs instead.

[HEP] we have the same problem. Solution: take same events with different kinds of information for different people, or categories of different events ("trigger menu"). Different ways of "observing" in

[astro] We observe in different ways, time assigned for different science goals.

[missed a bit]

[HEP] antenna gain calibration?

[astro] you want to do this all at the end, because otherwise we risk committing already to “ancient algorithms” not scalable. So you want to have data as raw as possible.

[HEP] we do certain things by dropping raw data

[astro] we do too, similar with visibility

[grav] online vs offline calibration. Each pipeline will have its own data reduction. Others do filtering, preprocessing…

[astro] how early did you do this?

[grav] a while ago...iterative process. We started with the on-the-fly calibration (someone thought we could just do the calibration online and release the data off already cleaned to be analysed) but in the end it didn’t quite work. This is for the fast alert system, otherwise it doesn’t work. Then we do offline calibration that is better.

[astro] the telescope is going to be built in stages, antennas stitched together. ⅓ of SKA is MeerKAT, it’s already working and we can do science verification data. Would you try to do data challenges with the actual data?

[grav] yes, when we started to do some data challenges and building the interferometers, we started already working with this data to clean it, try to identify noise...build an end-to-end pipeline

[HEP] this means there is some iteration. So the next thing you would automate tasks and get pieces on the pipeline. That’s why the “CDR-style” of ideas in astro is quite rigid, in HEP we redesign things along the way (after building consensus).

[astro] Data challenges in SKA: simulate data as realistic as possible, then have competition between offline teams. Maybe we could do those things for the pre-image data.

[HEP] This seems good also to interest people in things that aren’t “the final analysis” and raise their profile, as we sometimes have problem finding people to do this kind of tasks because everyone wants to get to the physics (and this is physics too)

[astro] Maybe could talk to Anna about doing this kind of campaign.

%%%

[HEP] In some LHC experiments, 95% of the data is written out with the reduced data (standard pipeline) and throwing away raw. But in some others that’s not the case.

[HEP] psychological factor: don’t want to do too much data reduction on day 1 because you

turn on the experiment and you may want to go back.

[astro] In astro there is no end date when you can take data unless the experiment breaks. Get science you want, first, with raw data. Then go more crazy later on.

[HEP] the risk is compensated by the much more physics that one can do with those data reduction technique.

[astro] if we get 10x improvement in some physics then we can also show the physics improvements in the bigger picture because this is ⅓ of the data.

[HEP] “we give you the data with 2x finer binning” would be more useful?

[astro] it depends…

[missed some piece]

[HEP] now the data is all calibrated when we get it. We haven’t made changes that made the data incompatible with software analysis tools...we don’t want to do that.

[astro-LSST] models depend on cadence, so if you change the observation conditions then you need to retrain models.

[astro-Euclid/SKA] things are done extremely differently. Euclid are based on experiments that were done in the past. Online 2022 - data release 2027 we already know exactly what the input is going to be and what we need. In MeerKAT / SKA, we don’t know what we’re looking at. So this is a problem that we don’t know what your final data is going to look like, you don’t know what the user needs.

The reason why we can’t “prepare in advance” in SKA is that it’s general purpose and there are 50 different science cases wanting different things.

Examples: cosmology in SKA. Drop the resolution, be like a giant collection of single dishes and see half the sky. People who do galaxy they do super high resolution. So the experiment has completely incompatible ways of operating.

[HEP] is there enough overlap between groups that you develop data products for use cases A, B, C?

[astro] this was the commensality problem, send requirements and how do we want the data and what binning…it gets into a political problem. Would be nice to be making everyone happy. But that’s why pathfinder data is important. Hopefully things will become obvious later on.

%%%

[grav] what is the data format?

[astro] the final user gets python array, or fits: https://en.wikipedia.org/wiki/FITS.

[astro?] different people working on different datasets (SKA) can you use the instrument for a given amount of time?

%%%

[HEP] autoencoders for compression?

[HEP] some interesting results from HEP. There is literature.

[Caterina Doglioni needs to add here some papers…Michaela Lawrence will also add hers]
Original autoencoder paper: Kramer - Nonlinear Principal Component Analysis Using Autoassociative Neural Networks
Variational autoencoders (if you are interested)

[grav] it may not work because you're losing variations. PCA only works for linear cases. You are only encoding the lower dimensionalities.

Full raw for 24h. Then run it through AE. Then look at a box of events that have been reconstructed.

[HEP] They're not capturing the full variance but that is where the interesting physics is going to be.

Braindump of group 6

How do you know if an anomaly really is an anomaly?
[warning! In some fields anomaly could just mean "something different", where in others it could exclusively be something you don't want, like sensor/detector malfunction]

Other ways of calling anomaly detection:

- Outlier detection
- Feature detection
- Novelty detection

Anomalies in HEP:

- Bumps (anomalies in the statistical sense, like the Higgs boson discovery)
- Single products of collision events that look different wrt others
- Anomaly

Anomalies in data science: birds sitting on a sensor - looks like anomalous data!
Remove "events" from your non-normal distribution to make it look normal, and those events are anomalous.

Use this as something that needs to be ignored.

Anomaly detection in monitoring → unwanted.

Feature detection for interesting (astro-)physics phenomena → wanted

Supervised / unsupervised.

In data science, anomalies are usually an unsupervised problem. Fit a model, what we expect should happen and if it doesn't happen then that's an anomaly. Some anomalies are "bad entries", like -1, corrupted data, and that is very easy to detect

Challenges:

- HEP: how do you distinguish wanted and unwanted anomalies?

This talk discusses anomaly detection in HEP:
https://indico.cern.ch/event/778133/contributions/3245543/attachments/1767499/2870414/DarkMachine_Nov18.pdf

Auto-encoders: good for identifying, in a broad sense, something that does not look like what you "input" (e.g. qcd input makes susy look anomalous => decoder provides a large 'reconstruction' error)

⇒ then you can proceed, afterwards, with studying and classifying the "anomalous" events.

Cosmology - classifying supernovae - classifying what we know exists vs. what we didn't already know exists (anomalies!)

"Baskets" of anomalies
Iteratively get anomalies - then feedback to model uninteresting anomalies…

Take basket, then check content: are those the anomalies that you're interested in?
How to:
Run classifier and get reconstruction error, maybe it's because your autoencoder was not encapsulating all the variability of your data. Select those that were "not too bad" (criterion: reconstruction error), and feed back into the training. Do this until you get a good separation.

Some contamination is ok - really want to find the super rare anomalies, so missing some not so rare ones would be ok. If it's not iteratively, you might not get the best separation [is this a smooth transition??? Or is it chaos!]