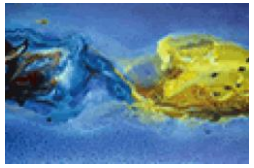
UF
UNIVERSITY of FLORIDA



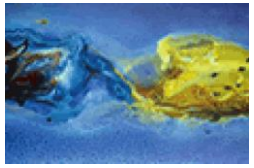
QCHS Track H

Summary

Sergei Gleyzer

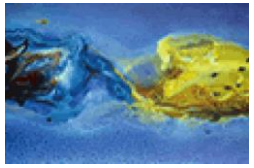
Aug 3, 2018

Keywords



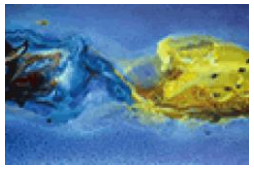
- Machine Learning
- Data Analysis
- Statistics

Statistics



- 21 presentations
- ~1/2 related to machine learning

Presentation Topics



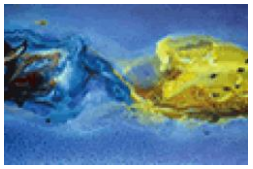
- **Machine Learning**
  - Software and Tools, Deep Learning Applications and Interpretation, Metrics, Gaussian Processes, Simulation, PDFs
- **Statistics and Data Analysis**
  - Bayesian Methods, Unfolding, Confidence Intervals, Coverage, Morphing, Managing Systematics, Anomaly Detection

A Few Open Problems

- Here is an incomplete list of open issues in the application of statistical tools to HEP analysis
  - Discovery levels: can we go Bayesian?
  - Optimization: everybody claims they did it. But what about systematic uncertainties?
  - DNNs: brute force or feature engineering?
  - Unsupervised learning and model-independent searches: can we ever safely get there?
  - Unfolding in multi-D: should we bother?

T. Dorigo

UF UNIVERSITY of FLORIDA



Deep Learning

Every standard machine learning method, even one as cryptic as AdaBoost, can be cast as an optimization problem whose goal is to minimize the average

$$R(f) = \frac{1}{N}\sum_{i=1}^{N} L(y, f(x_i, \theta)) + C(\theta)$$

of a suitable loss function $L(y, f)$ subject to some constraint $C(\theta)$.

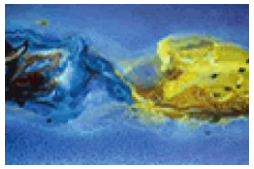
The key point to note is that this sum approximates the *functional*

$$R[f] = \int \left[\int dy L(y, f(x_i, \theta))\, p(y, x) \right] dx,$$
$$\equiv \int G(f)\, dx,$$

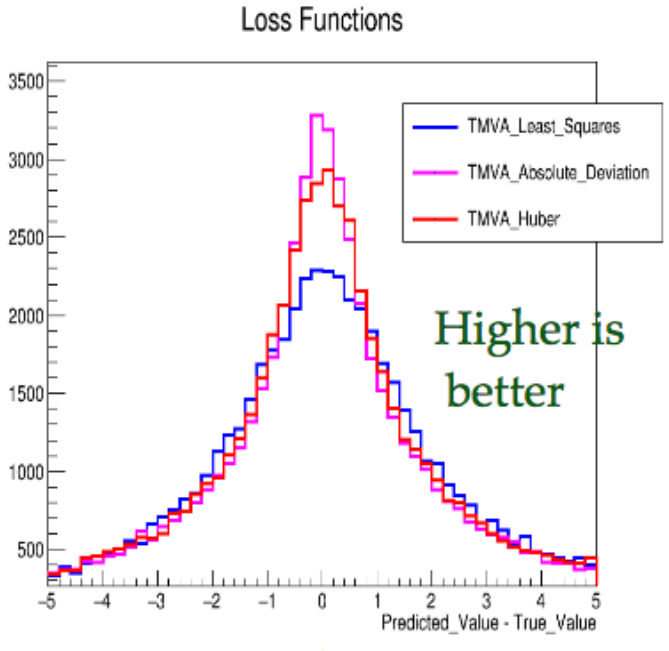
H. Prosper

where $p(y, x)$ is the probability density of the targets y and features x.

Loss Functions



- **New Regression Features:**
  - Loss function
    - Huber (default)
    - Least Squares
    - Absolute Deviation
    - Custom Function



Important for regression performance

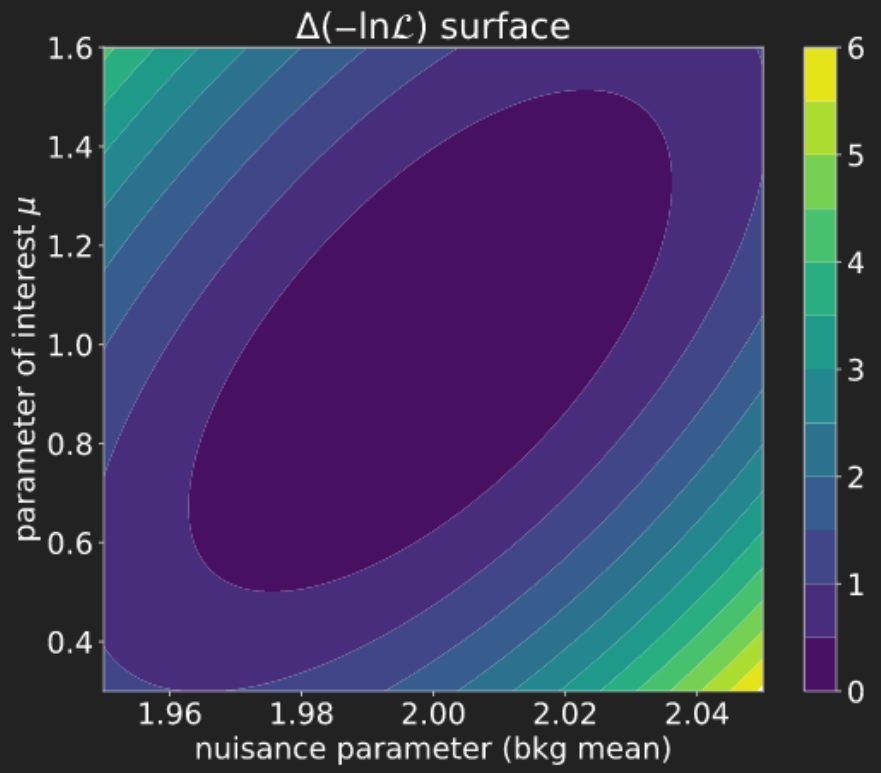
INFERENCE-MOTIVATED LOSS FUNCTION

If we expand the negative log-likelihood around minimum (e.g. Asimov $n_i = \alpha_s \cdot s_i + \alpha_b \cdot b_i$), due to Cramér-Rao bound:

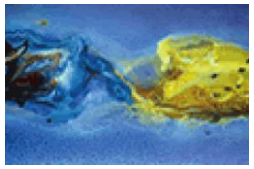
$$\text{covariance} \geq \mathbf{H}^{-1}(-\ln \mathcal{L})$$

which can be computed via autodiff. Can use as loss function directly the variance bound on the parameters of interest

$$\text{loss} \approx \text{Var}(\mu) \quad \text{(expected)}$$



UF UNIVERSITY of FLORIDA



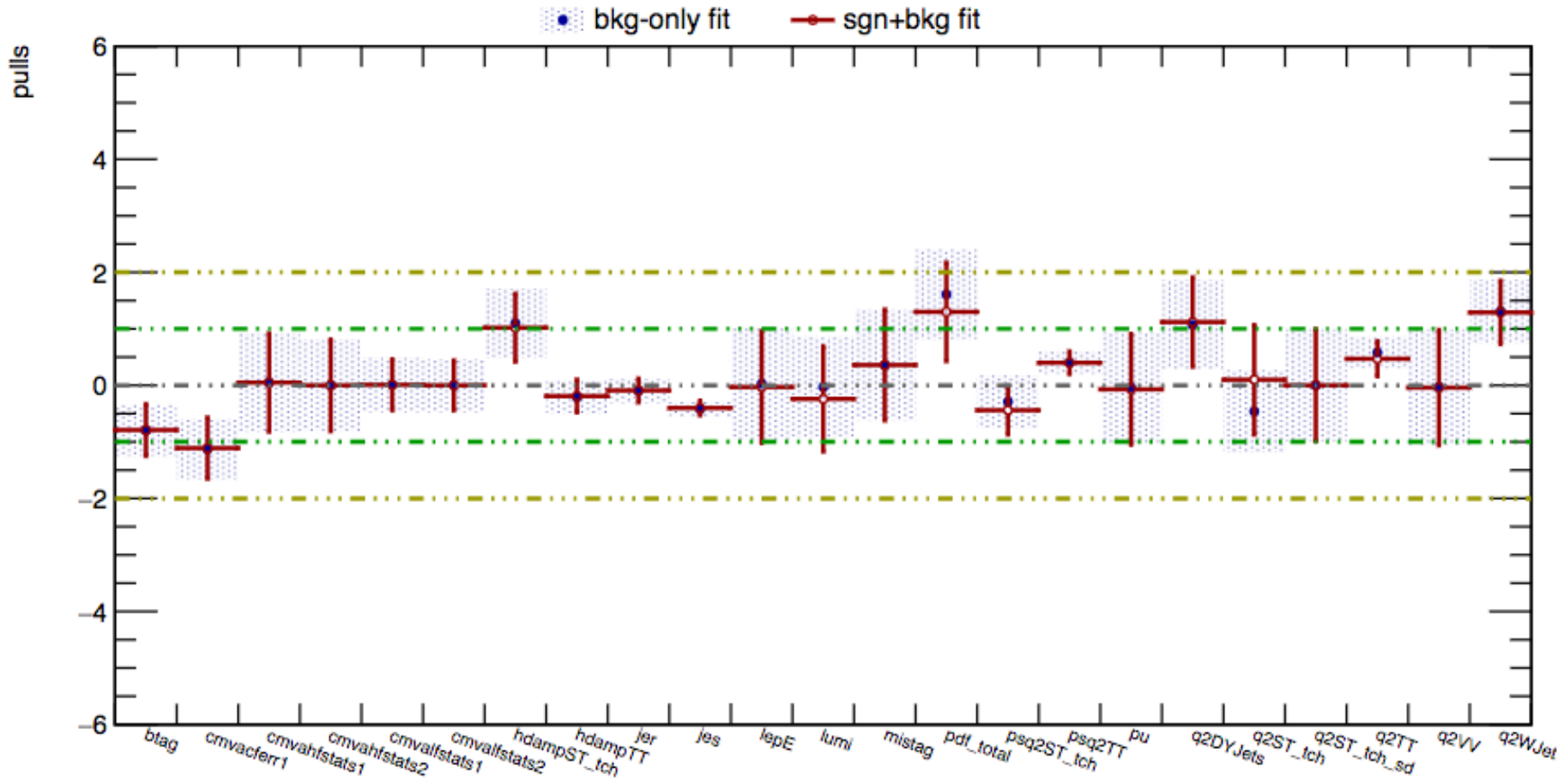
Systematic Uncertainties

Sources of uncertainties

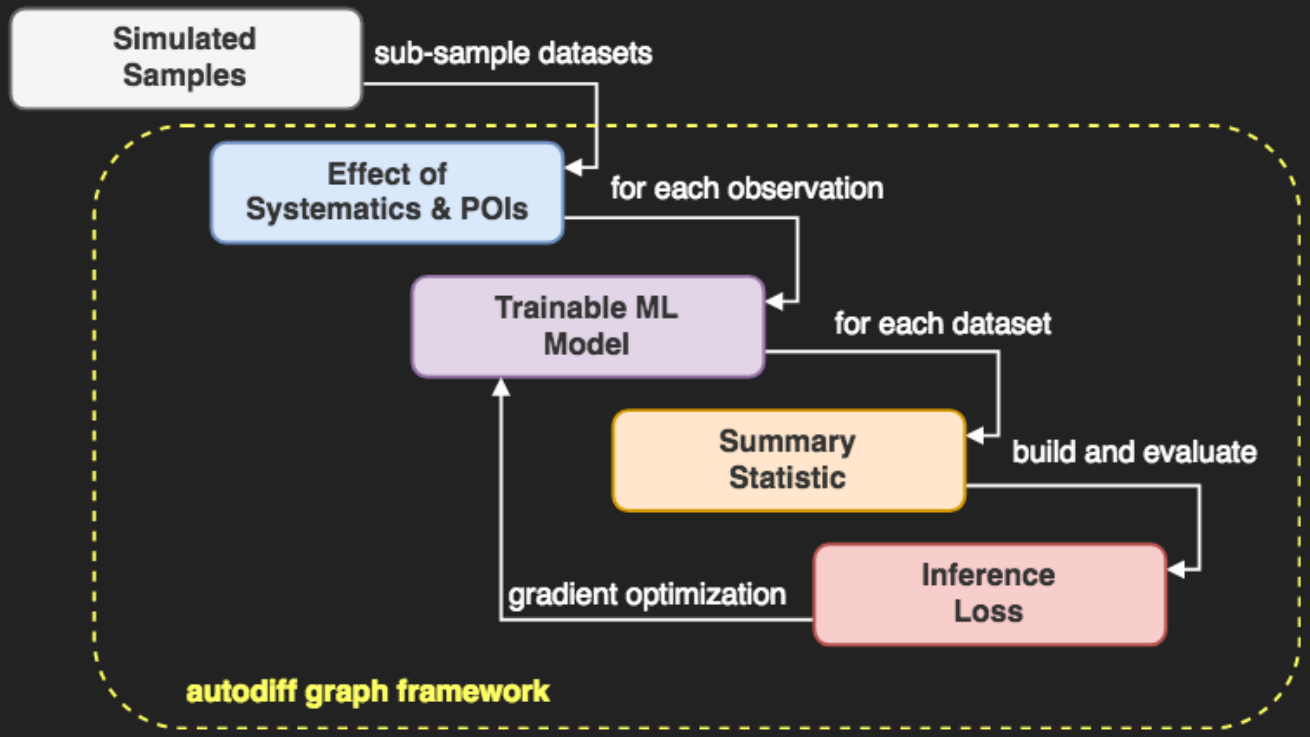
- Systematic uncertainties may affect the rate (i.e.: cross section) or shape (i.e.: distribution) of a process or both
  - Luminosity
  - Pile up modeling in simulation
  - Jet Energy Scale
  - b-tagging efficiency, mis-id, flavor dependence
  - Mu, e selection, reconstruction and trigger efficiencies
  - Theory modeling:
    - Individual cross section predictions
    - Shape and normalization due to renorm./factor. Scales
    - PDF models
    - Parton shower modeling
    - Generator choice
    - . . .
  - Monte Carlo simulation
    - Limited sample size
  - . . .

Results of fit (2)

- Constraint of systematic uncertainties



END-TO-END DIFFERENTIABILITY FOR LHC ANALYSES



Within this general framework, several approaches are possible, focus here is

DIRECT LEARNING OF SYSTEMATICS-AWARE SUMMARY STATISTICS

P. de Castro

Example (Absolute Loss: $L(y, f) = |y - f| \equiv \sqrt{(y - f)^2}$)

For the absolute loss,

$$G(f) = p(x) \int \sqrt{(y - f)^2}\, p(y|x)\, dy,$$

$$\frac{\partial G}{\partial f} = p(x) \int \frac{(y - f)}{|y - f|}\, p(y|x)\, dy = 0.$$

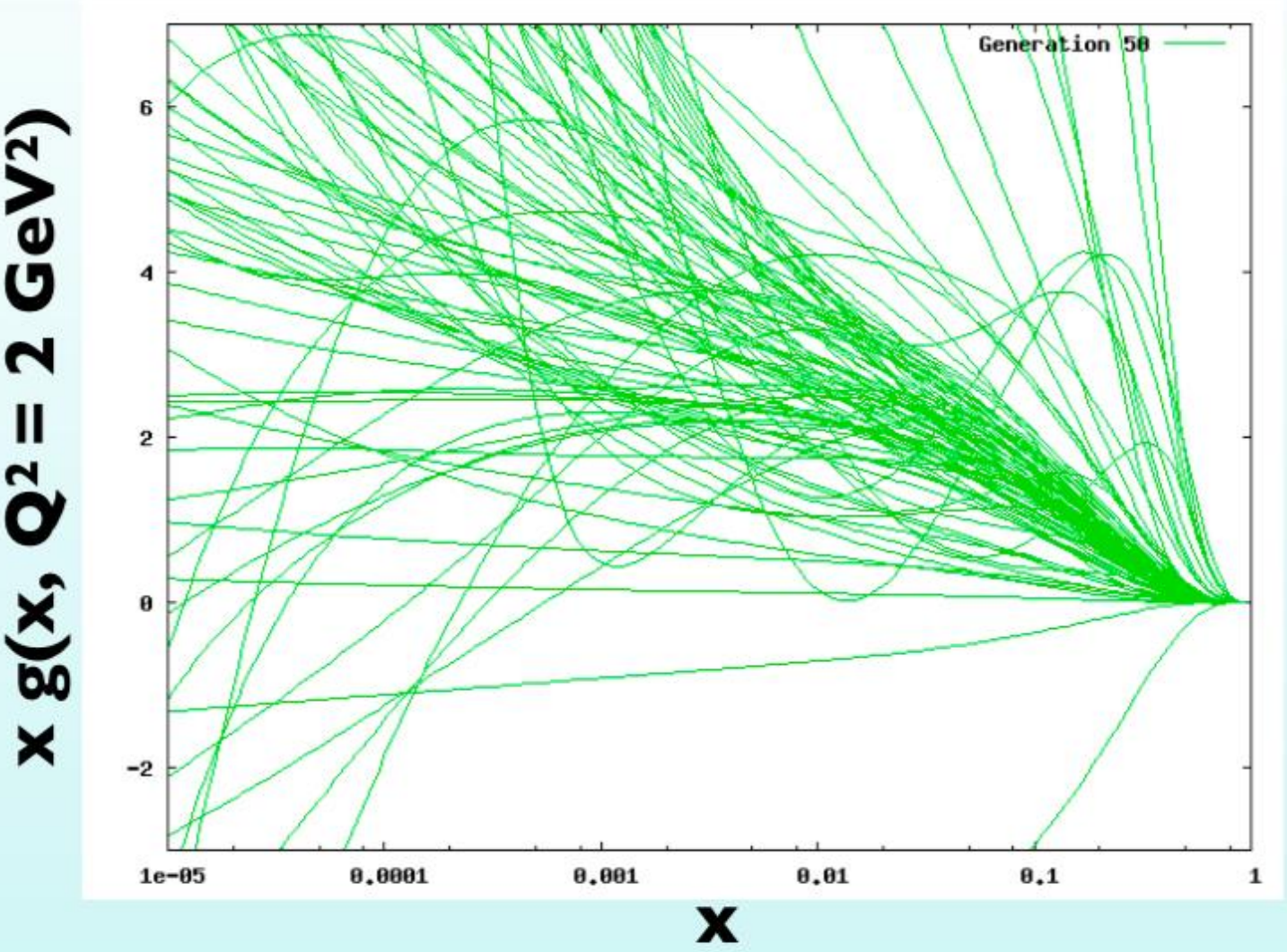
Noting that $(y - x)/|y - x| = 2H(y - f) - 1$, where $H(z)$ is the Heaviside function, $f(x, \theta)$ is the solution of $\boxed{\int_{y>f} p(y|x)\, dy = \frac{1}{2}}$.

Conclusion If 1) the training data are sufficient and 2) $f(x, \theta)$ is sufficiently flexible and 3) we use the absolute loss then $f(x, \theta)$ will approximate the *median* of $p(y|x)$.

08/02/18

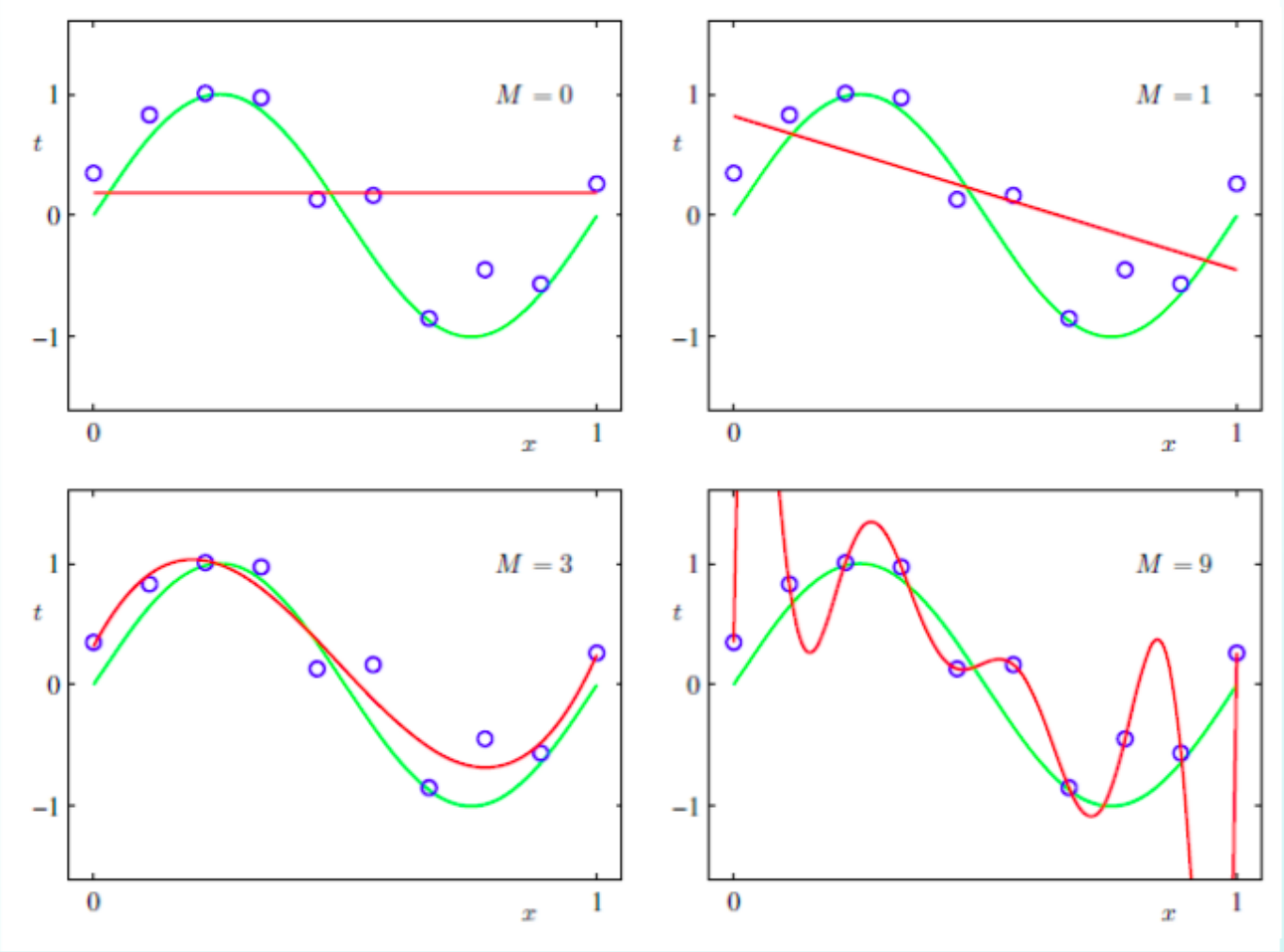
Neural network training

Starting from **random boundary conditions** for the N_{rep} replicas,
the ANN training ensures that only those functional forms **minimising the χ^2** are selected

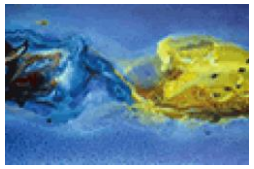


Avoiding overfitting

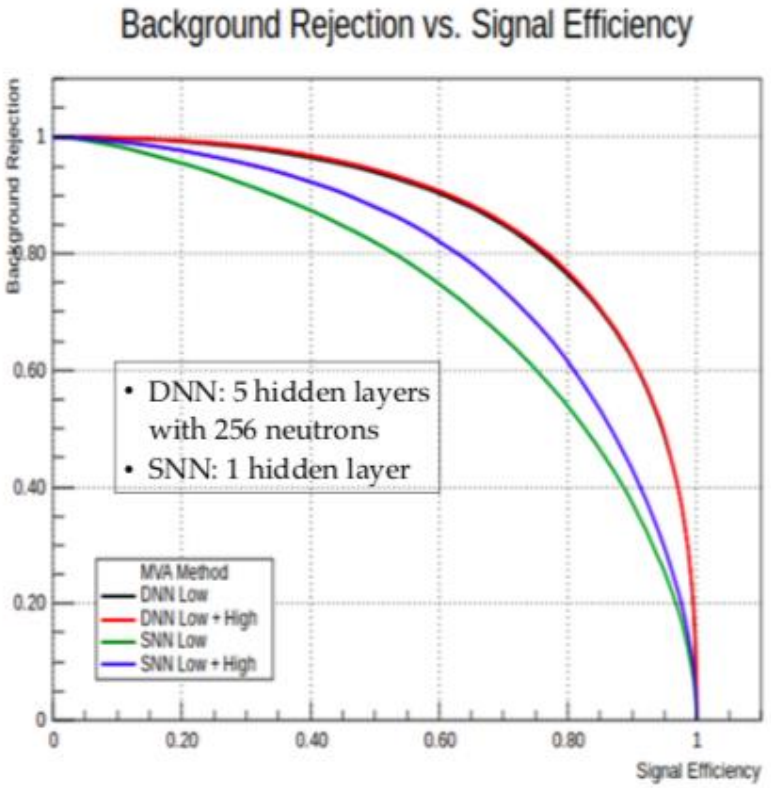
For a **flexible enough input functional form for the Parton Distributions**, one might end up **fitting statistical fluctuations** rather than the underlying physical law!



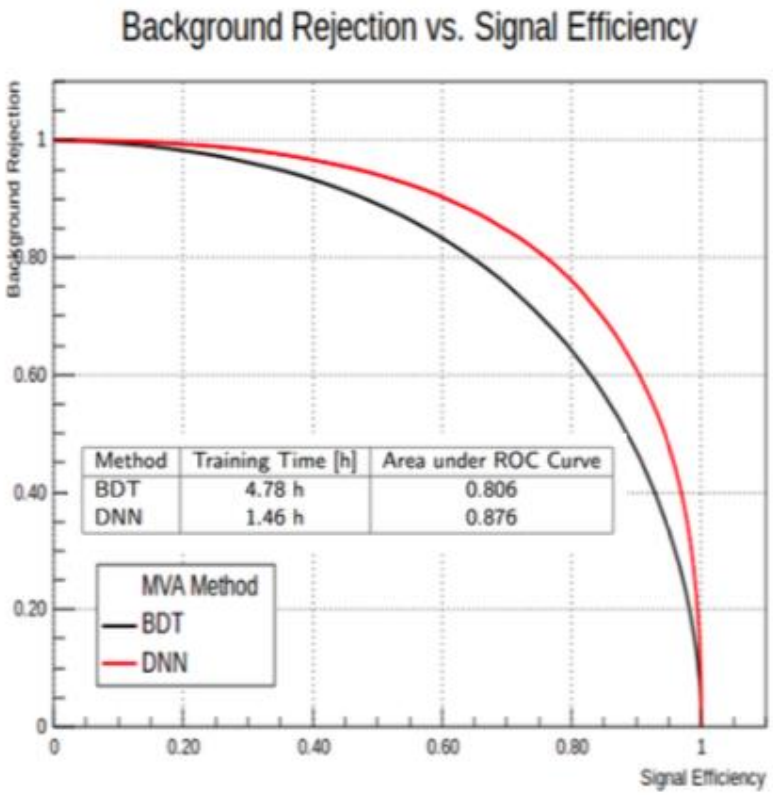
Extraction vs. Feature Engineering



DNN vs Standard ANN



DNN vs BDT



SG

Generative Adversarial Networks

- Extending the GAN architecture – provide a set of initial parameters for the generator and discriminator:
  - generator would not generate a random output, but a customized one
  - in our case: initial momenta of Monte Carlo particles

https://33milesinnewaygocounty.files.wordpress.com

Discriminator

https://thechive.files.wordpress.com

Generator

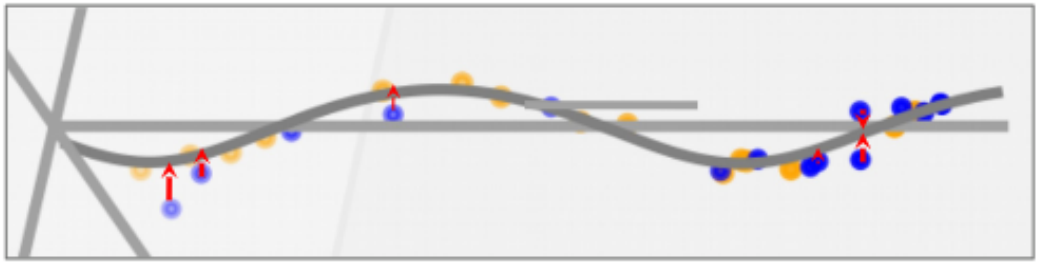
https://giphy.com/gifs/leonardo-dicaprio-catch-me-if-you-can-5leocharacters-t1h4nnWEWKfn2

Initial Parameters

Results

- Mean Squared Error (MSE) from the original helix as a quality measure
- Evaluation conducted on the separate test-set with ~15000 tracks

MSE visualisation:
Red - error
Grey- ideal helix
Orange – original clusters
Blue – generated clusters



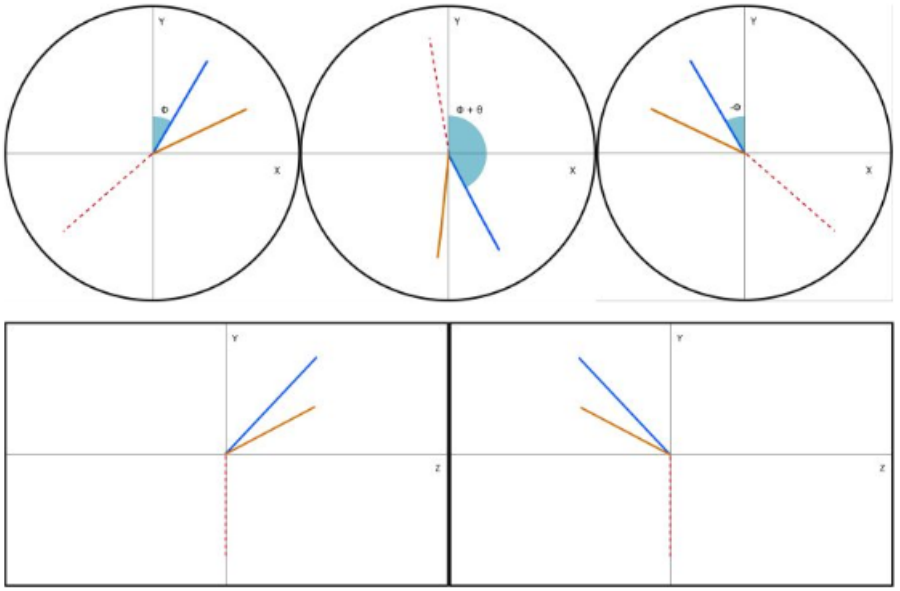
Method	Mean MSE (mm)	Median MSE (mm)	Speed-up
GEANT3	1.20	1.12	1
Random (estimated)	2500	2500	N/A
condLSTM GAN	2093.69	2070.32	100
condLSTM GAN+	221.78	190.17	
condDCGAN	795.08	738.71	25
condDCGAN+	**136.84**	**82.72**	

Learning Optimization

Data augmentation

G. Strong

- Correct application of augmentation relies on exploiting invariances within the data: domain specific
- At the CMS and ATLAS detectors, the initial transverse momentum is zero, therefore final states are produced isotropically in the transverse plane: the class of process is invariant to the rotation in azimuthal angle
- Similarly, the beams collide head on with equal energy: therefore final states are produced isotropically in Z-axis

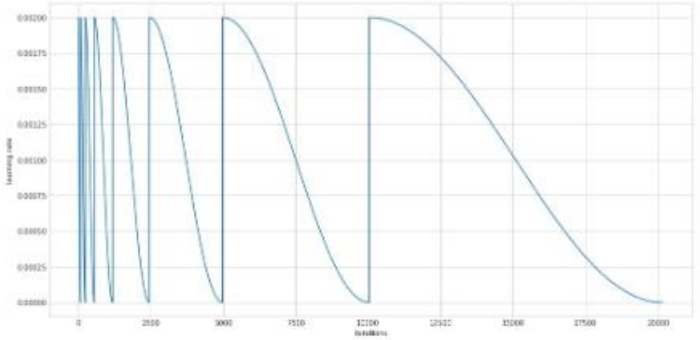


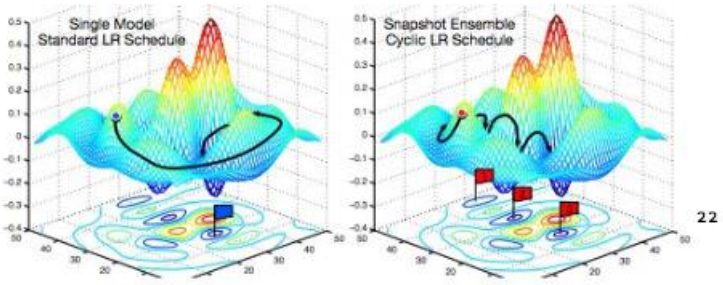
Learning Optimization

Learning-rate cycles

G. Strong

- Loshchilov and Hutter 2016 instead suggests that the LR should be decay as a cosine with the schedule restarting once the LR reaches zero
- Huang et al. 2017 later suggests that the discontinuity allows the network to discover multiple minima in the loss surface
- 2016 paper demonstrates on image and EEG classification

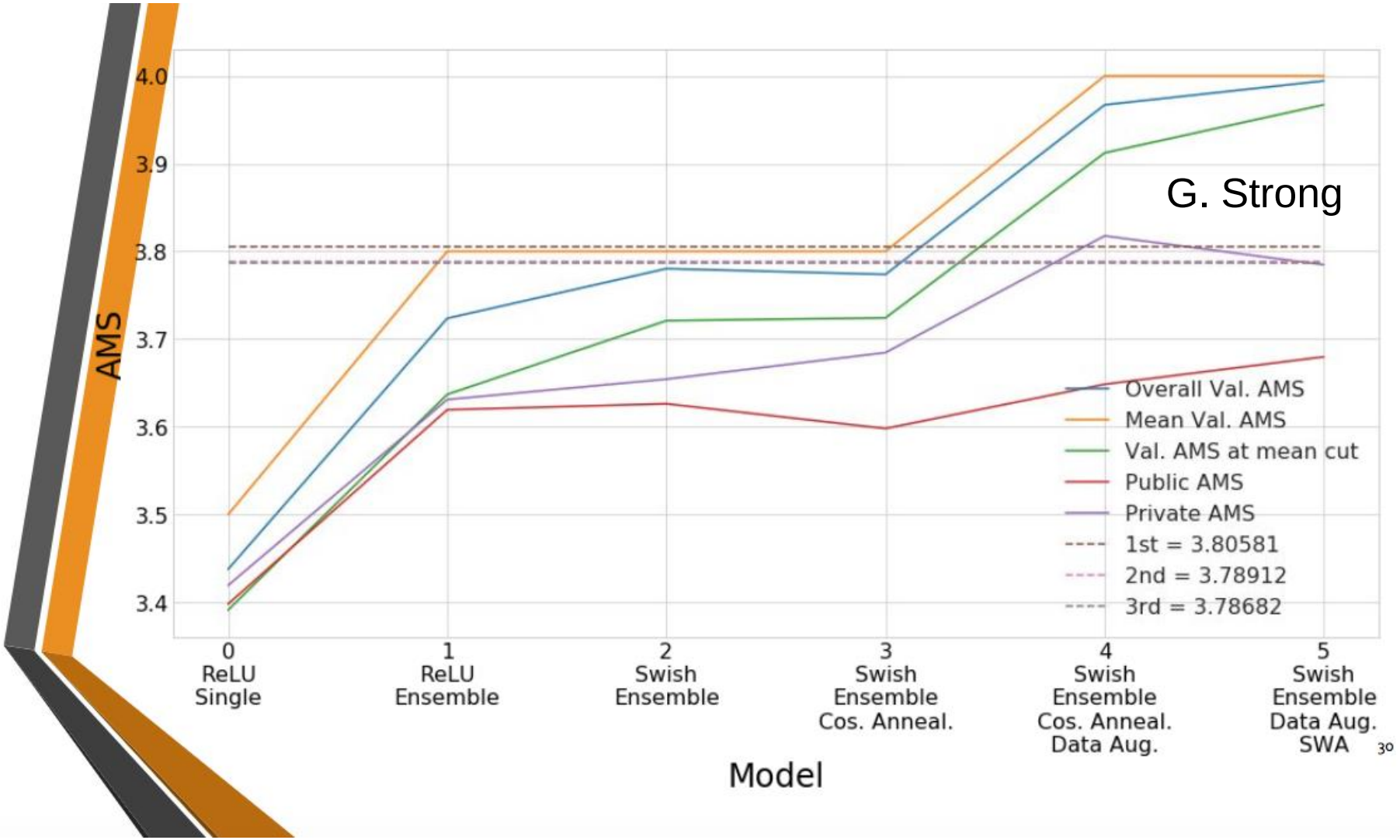


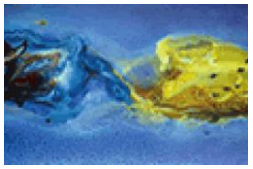


22

Lower figure - Huang et al., 2017, arXiv:1704.00109

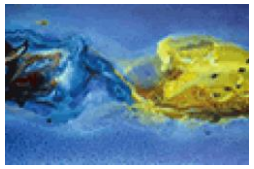
Optimization Results





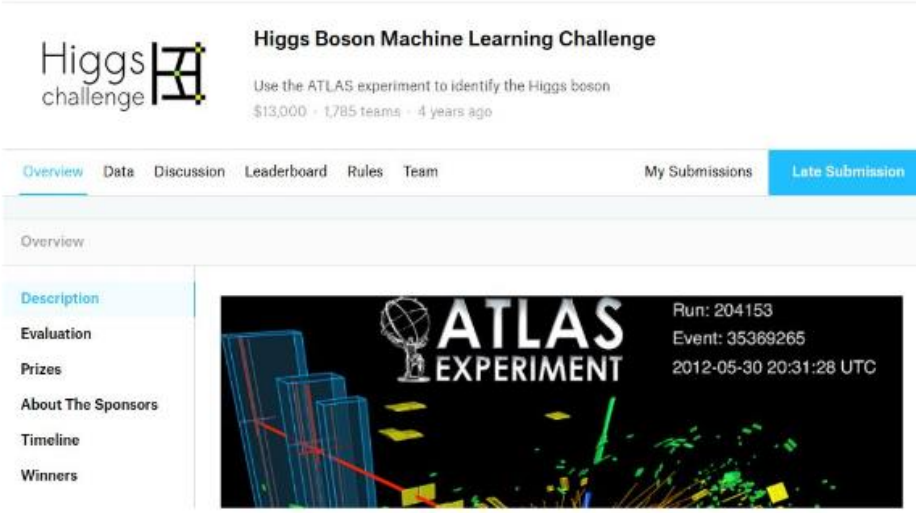
Collaborative/Open Data Science

Challenges



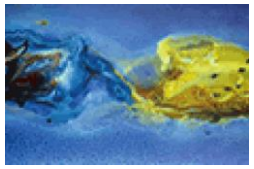
Higgs ML Kaggle Challenge

- Launched in 2014, the Higgs ML Kaggle competition was designed to help stimulate outside interest in HEP problems
- The data contains simulated LHC collision data for Higgs to di-tau and several background processes
- Participants were tasked with classifying the events in order to optimise the Approximate Median Significance
- The competition was highly successful, and helped introduce new methods to HEP, as well as produce more widely used tools, such as XGBoost

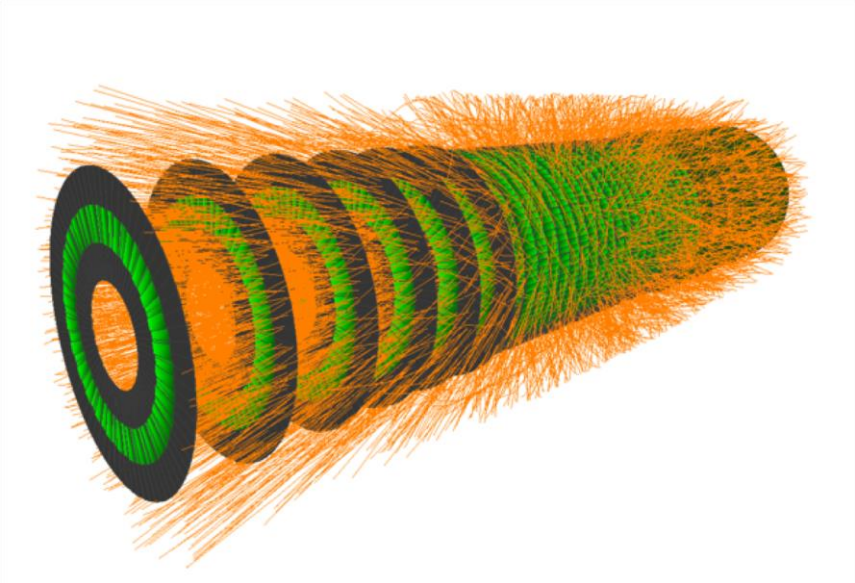


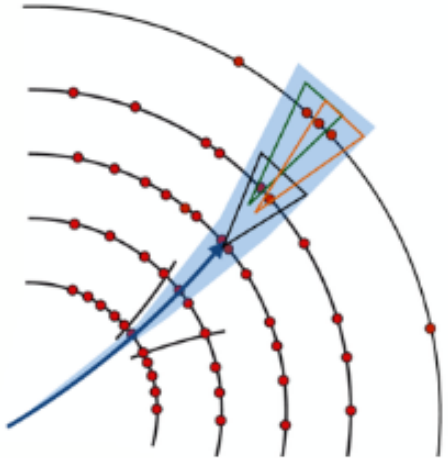
5

TrackML Challenge

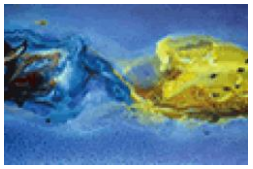


https://www.kaggle.com/c/trackml-particle-identification





UF UNIVERSITY of FLORIDA



Unsupervised Learning and Anomaly Detection

Example of the model fit

G. Kotkowski

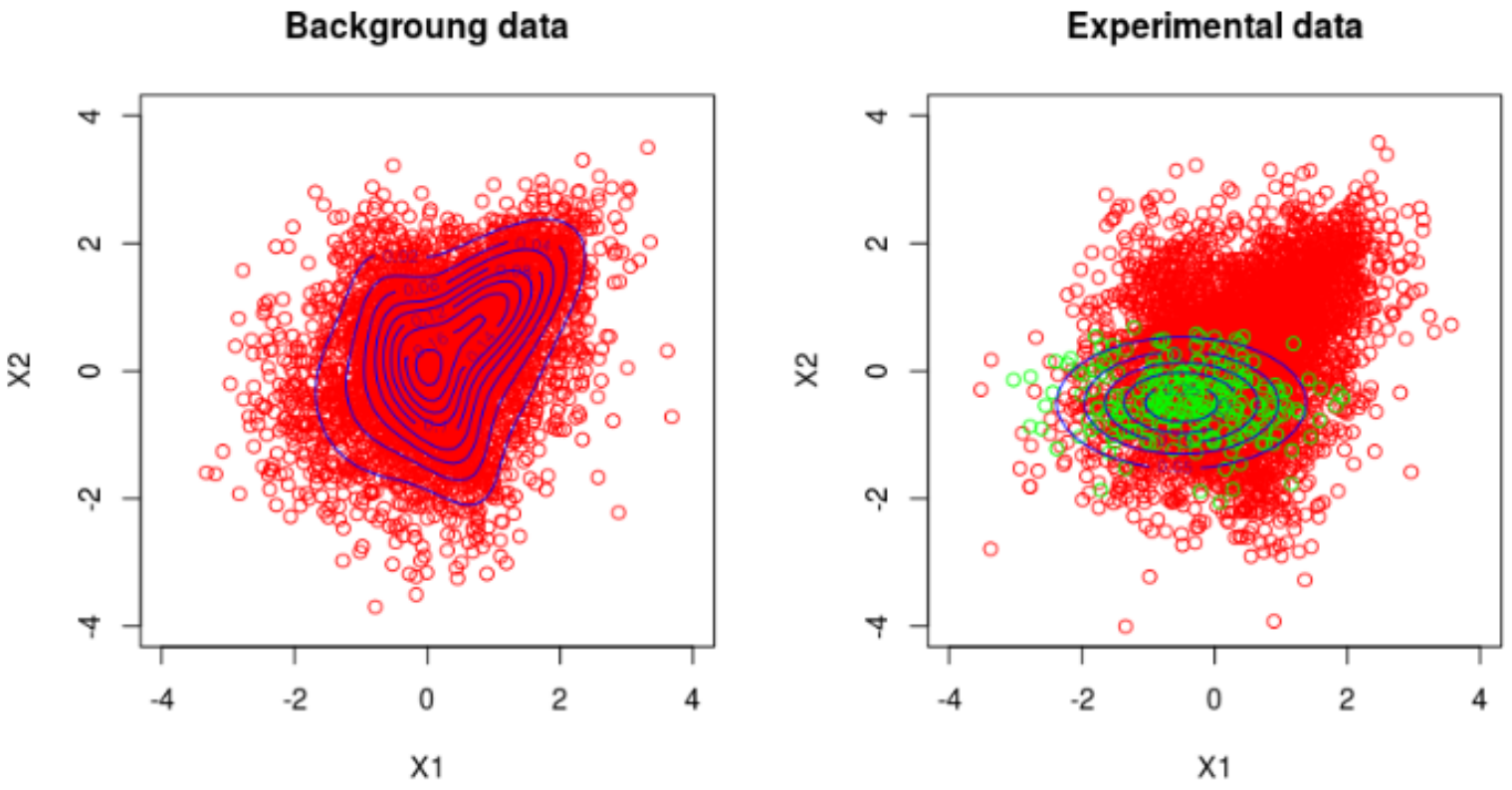
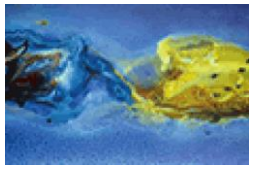


Figure: Examples of background and experimental data with the contoured background and signal distributions.

UF UNIVERSITY of FLORIDA



Figures of Merit

Binary classifier evaluation – reminder

Discrete classifiers: the confusion matrix

Binary decision: signal or background

	true class: Positives (HEP: **signal** Stot)	*true class*: Negatives (HEP: **background** Btot)
classified as: positives (HEP: **selected**)	**True Positives (TP)** (HEP: selected signal **Ssel**)	**False Positives (FP)** (HEP: selected bkg **Bsel**)
classified as: negatives (HEP: **rejected**)	**False Negatives (FN)** (HEP: rejected signal **Srej**)	**True Negatives (TN)** (HEP: rejected bkg **Brej**)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

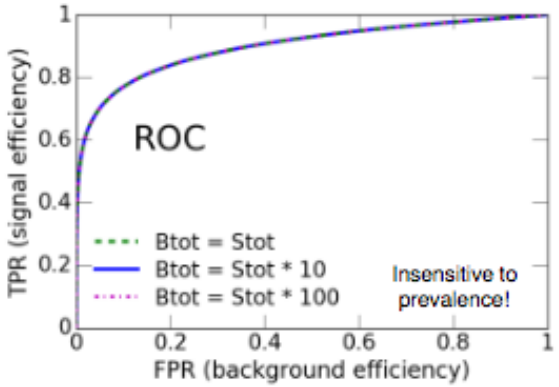
$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} = 1 - \mathrm{FPR}$$

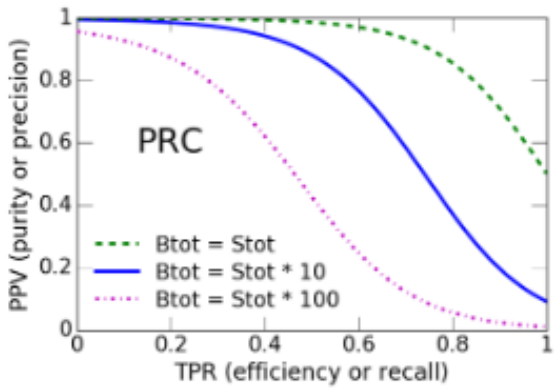
$$\text{Prevalence } \pi_s = \frac{S_{\mathrm{tot}}}{S_{\mathrm{tot}} + B_{\mathrm{tot}}}$$

Scoring classifiers: ROC and PRC curves

Continuous output: probability to be signal

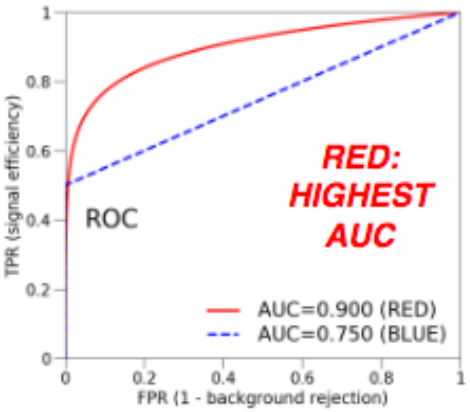
Vary the binary decision by varying the cut on the scoring classifier

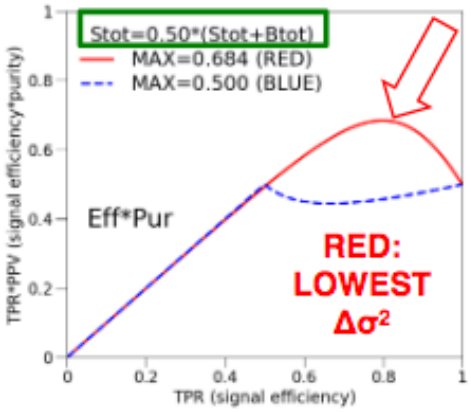


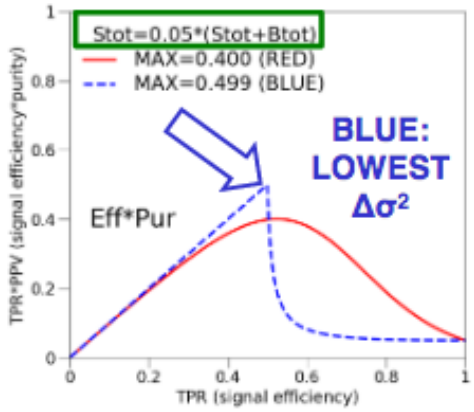


Examples of issues in AUCs – *crossing ROCs*

- Cross-section measurement by counting experiment
  - Maximize FIP1=$\varepsilon_s * \rho$ → Minimize the statistical error $\Delta\sigma^2$
- Compare two classifiers: red (AUC=0.90) and blue (AUC=0.75)
  - The red and blue ROCs cross (otherwise the choice would be obvious!)
- Choice of classifier achieving minimum $\Delta\sigma^2$ *depends on* S_{tot}/B_{tot}
  - *Signal prevalence 50%*: choose classifier with higher AUC (red)
  - *Signal prevalence 5%*: choose classifier with lower AUC (blue)
  - AUC is irrelevant – and ROC is only useful if you also know prevalence



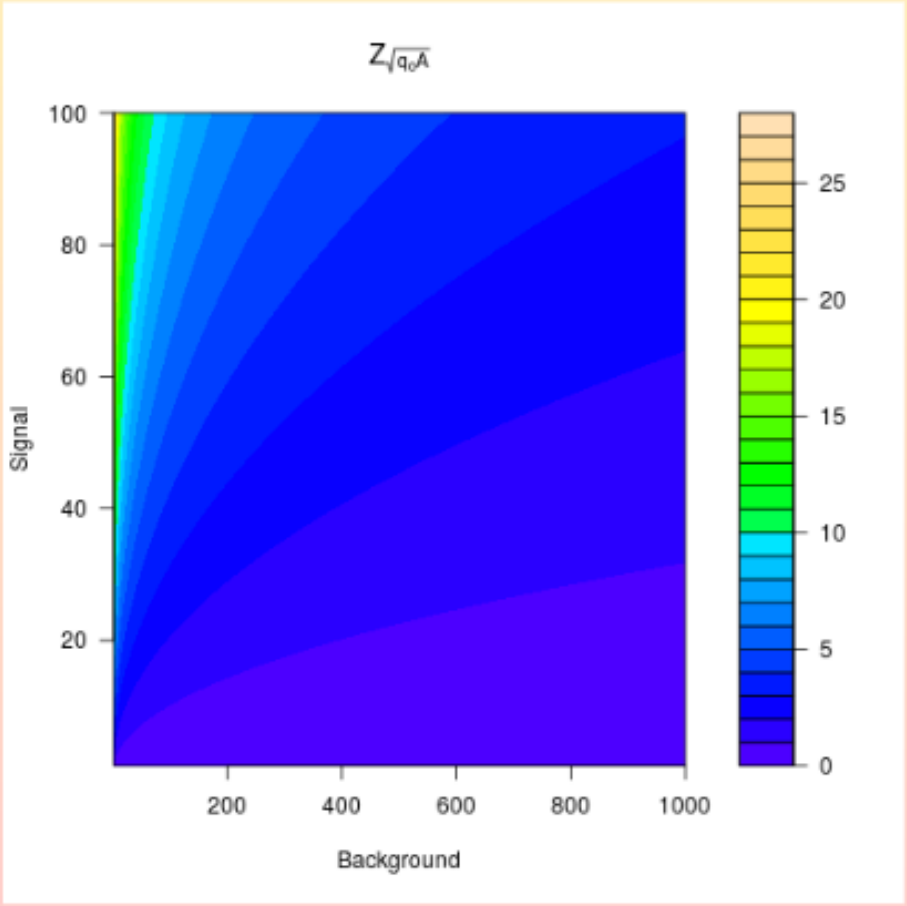


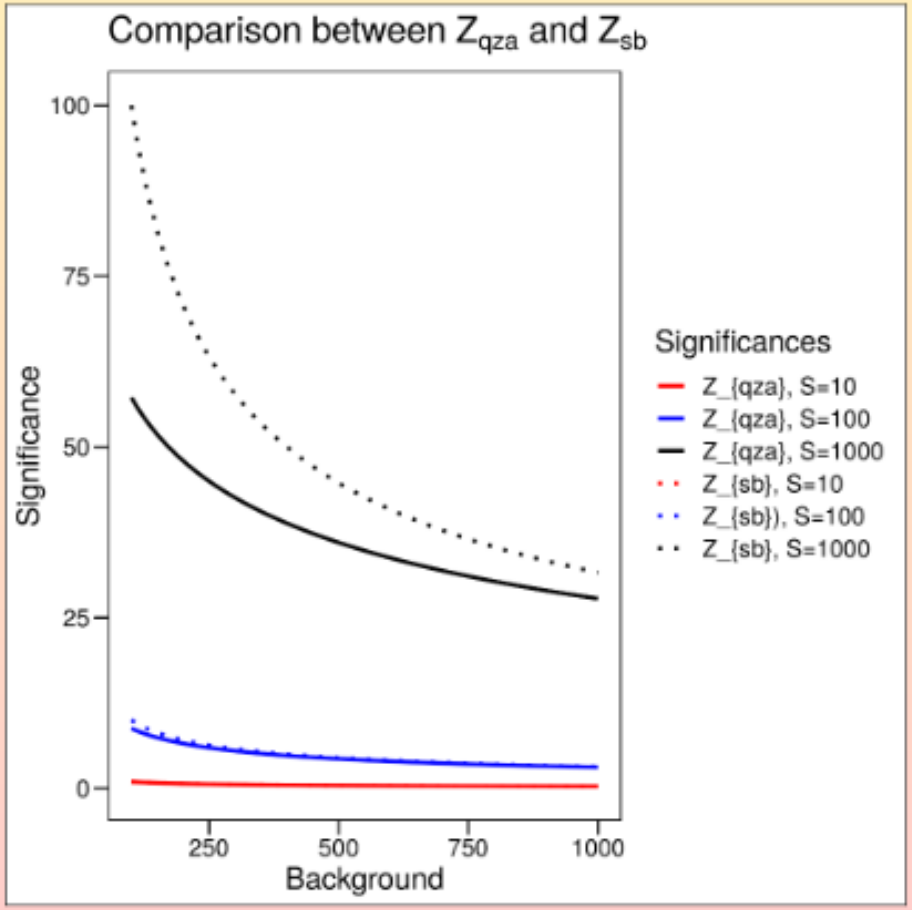


	FIP1	AUC
Range in [0,1]	YES	YES
Higher is better	YES	NO
Numerically meanigful	YES	NO

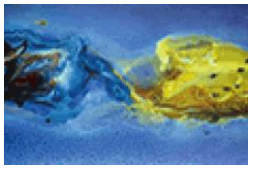
When $\Delta B \sim 0$ is negligible

- Approximation of the Cowan-Cranmer-Gross-Vitells asymptotic formula for known B
  $\sqrt{q_{0,A}} := \sqrt{2((S+B)ln(1+\frac{S}{B}) - S}$
- Its expansion in $ln(\frac{S}{B})$ is $\frac{S}{\sqrt{B}}(1+\mathcal{O}(\frac{S}{B}))$
- The $\frac{S}{\sqrt{B}}$ is hence good only for $S << B$
  - In literature has been used in general for large $S+B$, hence failing when $S \sim B$

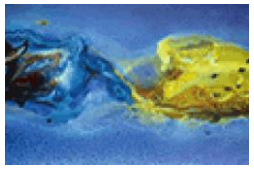




UF UNIVERSITY of FLORIDA

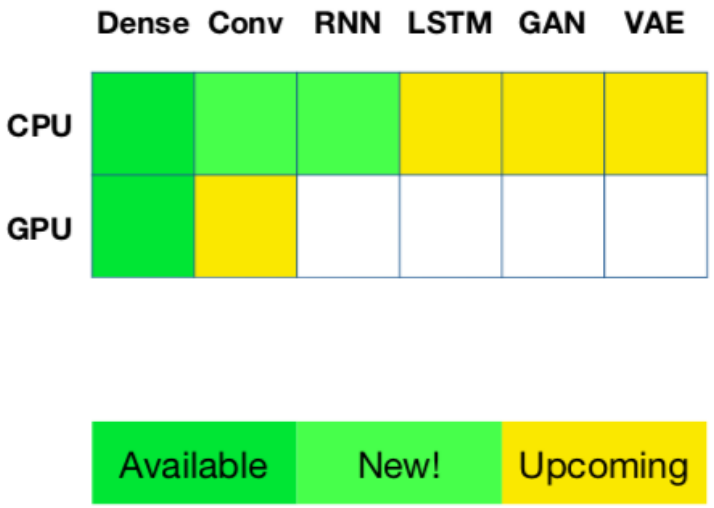


Software and Tools

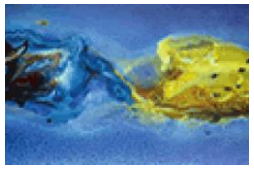


Status deep learning library

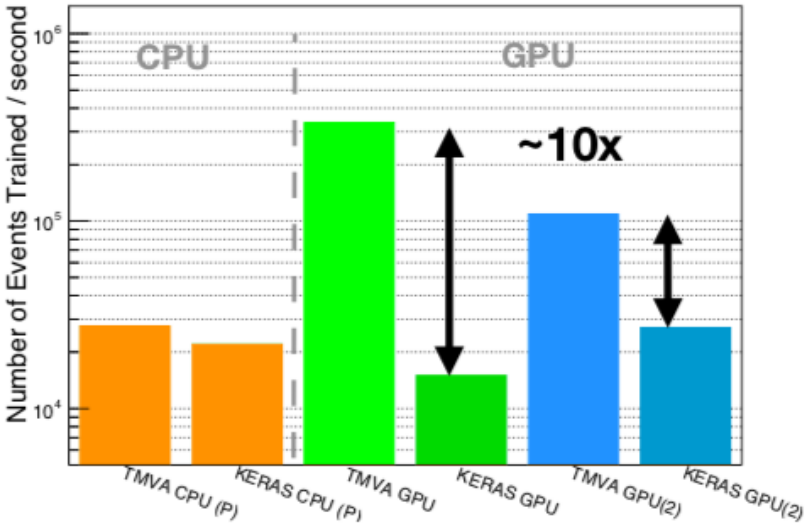
- Deep learning library since 2016
- Recent additions
  - Convolutional and recurrent layers
- Development ongoing!



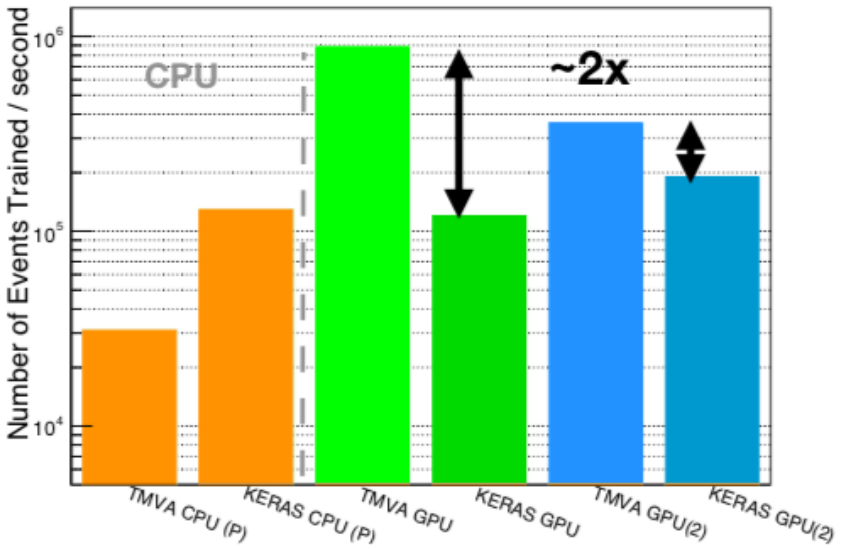
TMVA



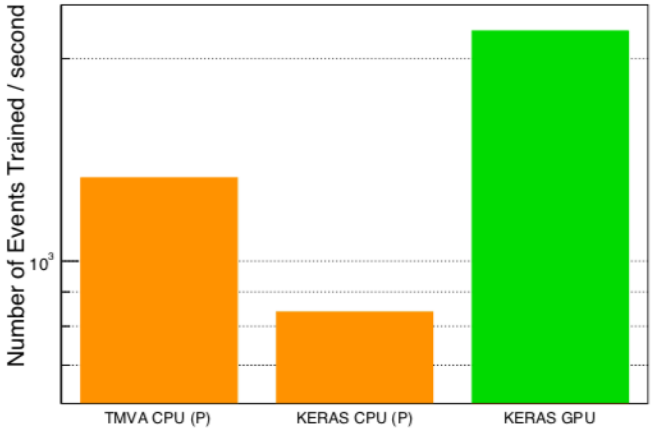
Batch size 100



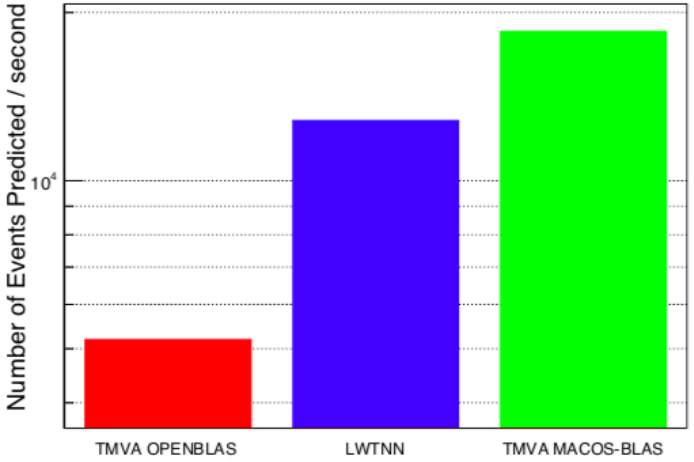
Batch size 1000



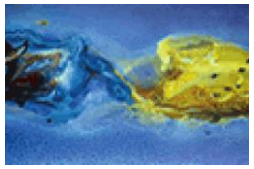
2 Conv Layer - 12 3x3 filters - 32x32 images - batch size = 32



Prediction Time (5 Dense Layers - 200 units)

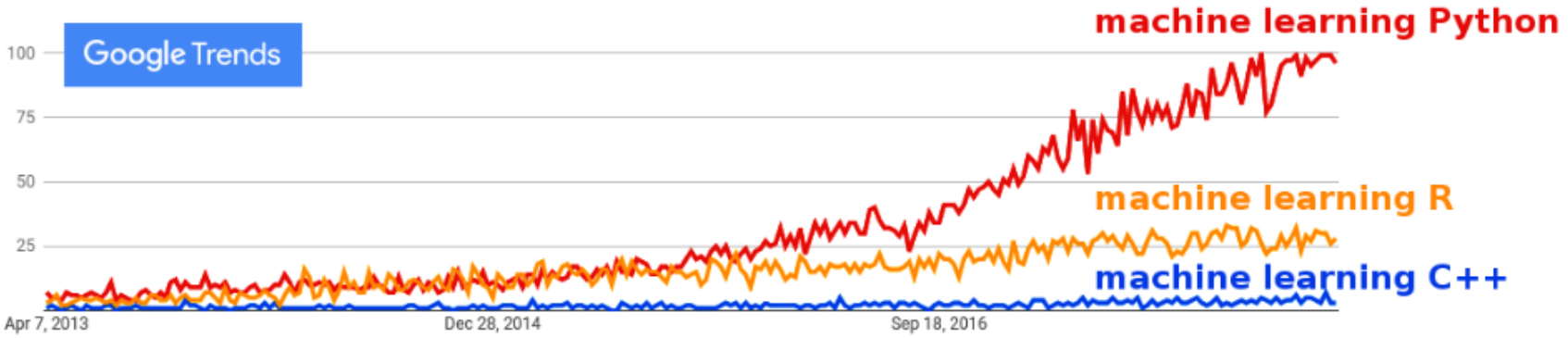


Python in ML

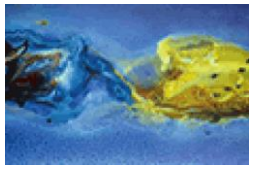


Particularly machine learning

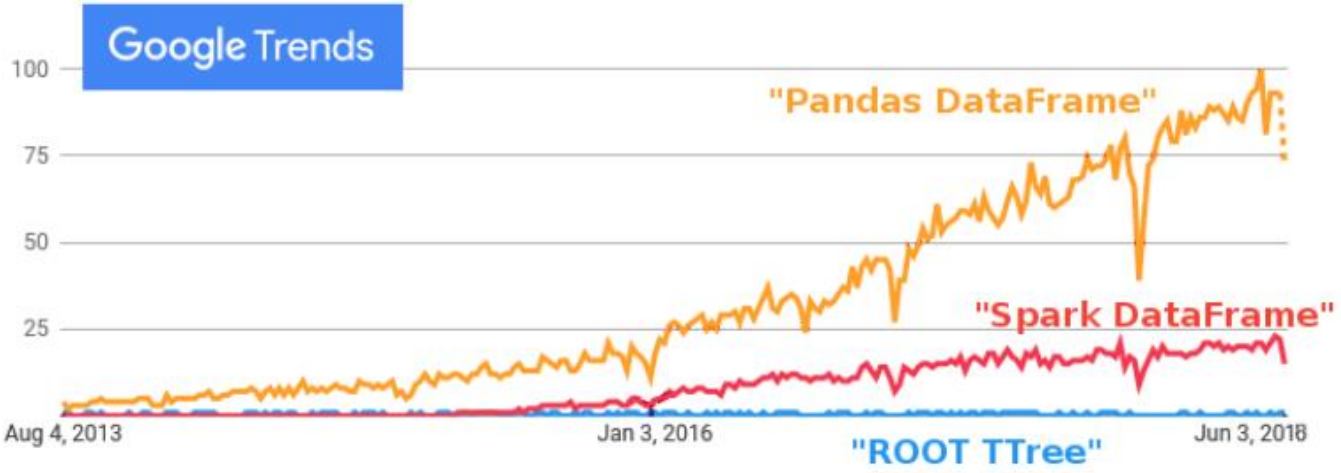
scikit learn · PYTORCH · Keras · TensorFlow · Caffe2 · GLUON · Chainer · Microsoft CNTK · Lasagne · ONNX · Cesium · dmlc XGBoost



Pandas



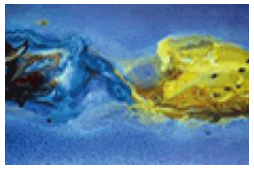
Pandas is a bigger thing than Spark



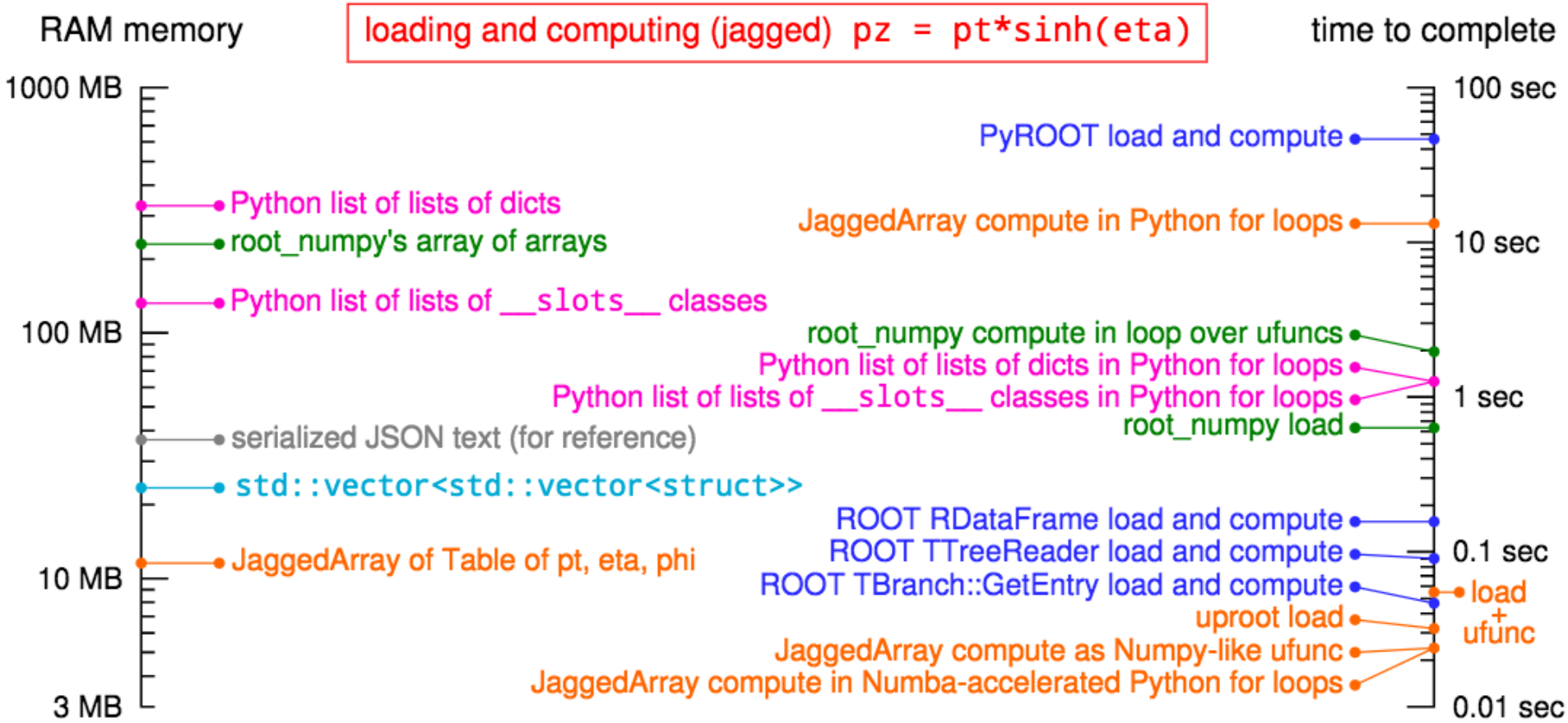
pandas

$$y_i t = \beta' x_{it} + \mu_i + \epsilon_{it}$$

Columnar Arrays



Loading and computing columnar arrays is fast



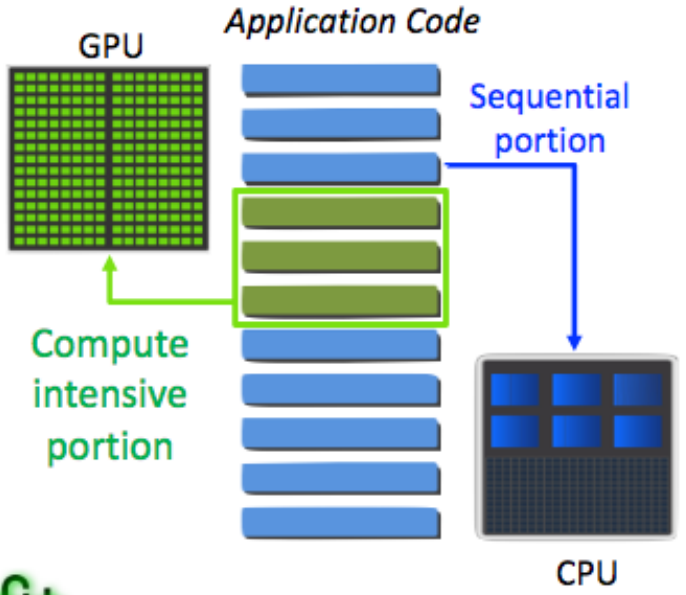
RooStats

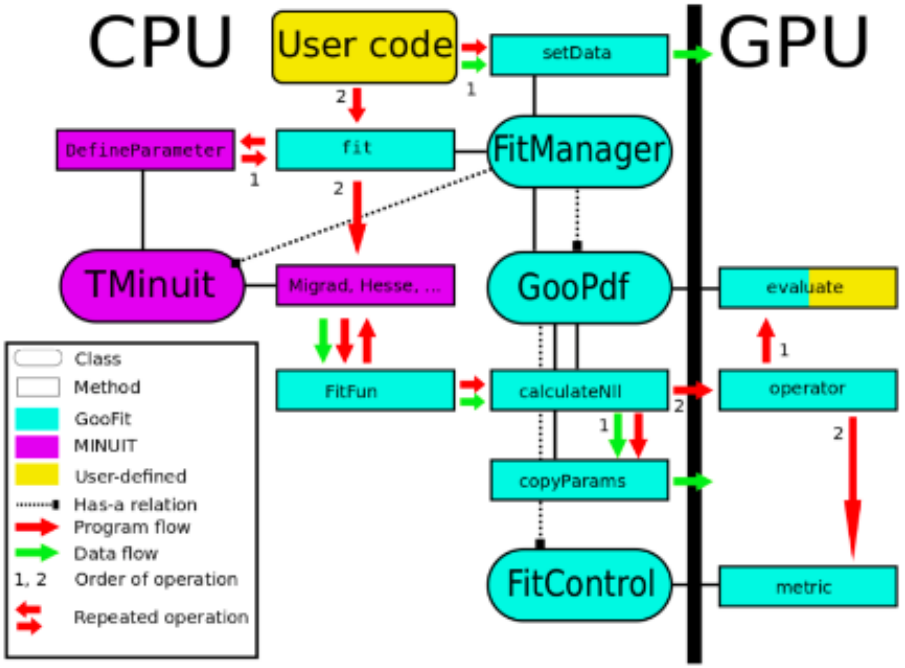
- Most of the methods adopted in High Energy Physics are implemented in the RooStats C++ framework
- Convenient modeling of PDF via RooFit package
  - PDFs from templates determined from ROOT histograms (`RooHistPdf` class)
  - PDF models and data with parameter definition stored in a convenient file format (`RooWorkspace`)
- Asymptotic approximations available, allow to save CPU time avoiding intensive toy Monte Carlo generation
  - G. Cowan et al., Eur.Phys.J.C71:1554,2011

GPU Computing in HEP analysis: the GooFit framework

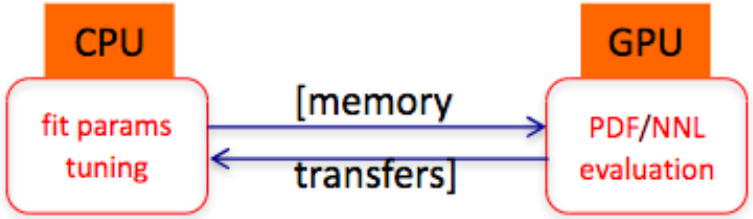
- Hetherogeneous GPU-acccelerated computing is the use of a Graphics Processing Unit to accelerate scientific applications (among other apps).

We explored the capabilities of GPU compuiting in the context of the 'end-user HEP analyses' by using *GooFit*.





Goofit is a data analysis tool for HEP, that interfaces ROOT/RooFit to CUDA parallel computing platform on *nVidia* GPU. It also supports OpenMP.



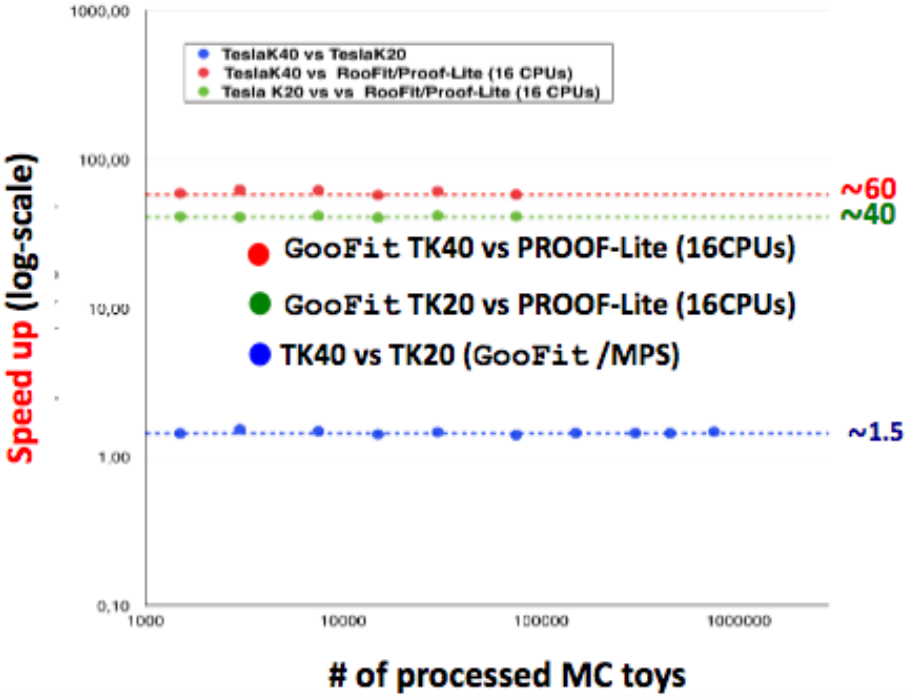
From the user's perspective? Applications simply run significantly faster! How much faster ? It depends - of course - on the application... We tested it firstly with the estimation of the local significance of a known signal.

Since v2.0 **Goofit** is completely integrated in python through **PyBindings** and it can run within jupyter notebooks that makes its use even easier.

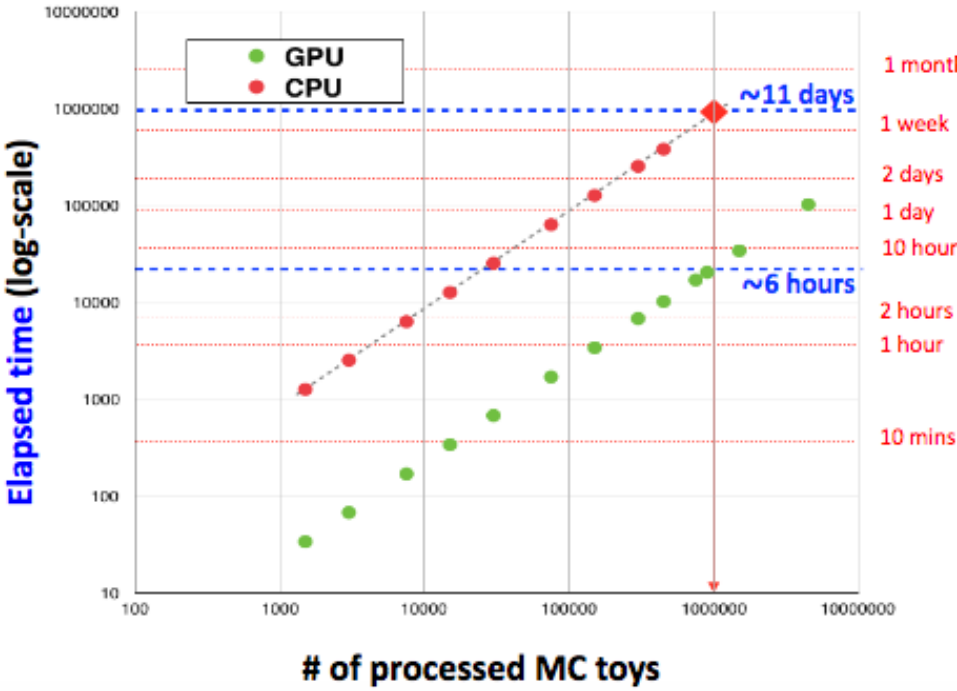
A first use case: GooFit performances

- **The optimized *GooFit* applications running, by means of the MPS, on GPUs, hosted by the servers used in the presented test, has provided a striking speed-up performance with respect to the *RooFit* application parallelized on multiple CPUs by means of *PROOF-Lite*.**

- A first performances' comparison is carried out on both the servers hosting both type of GPUs (TK20 & TK40) as a function of the # of pseudo-experiments produced keeping constant the number of workers/processes.

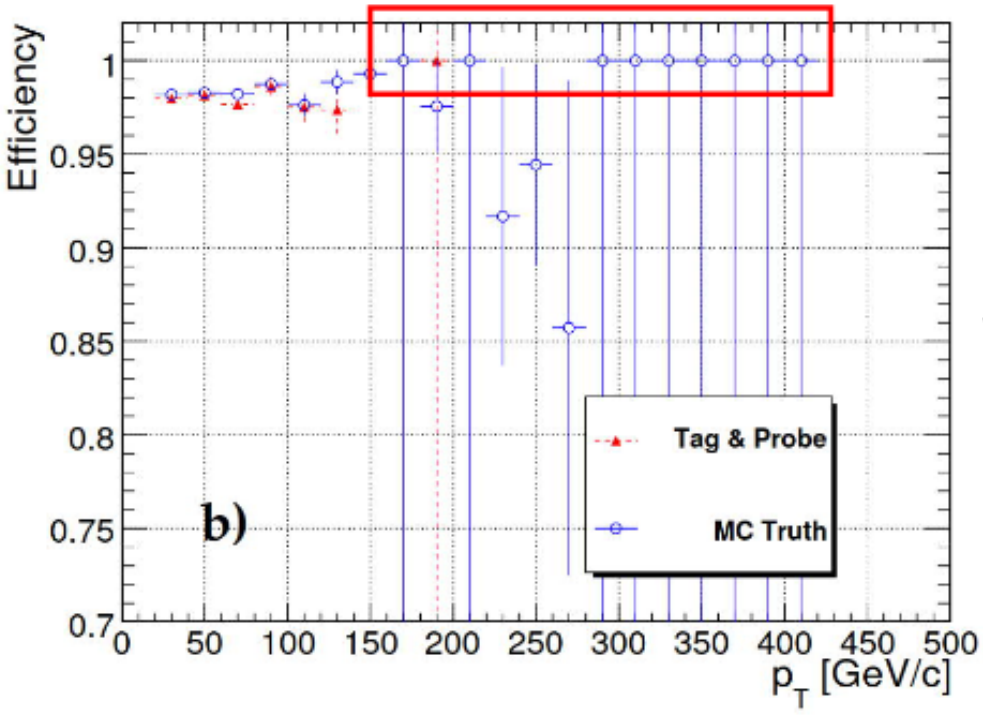


- A second comparison is done from the point of view of the end-user/analyst having at disposal **72 CPUs and 3 GPUs (1 TK40 & 2 TK20) on 2 servers**



Issue: awareness of limitations

- TH1::Divide assumes uncorrelated errors

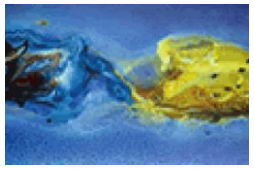


M. Mozer

Plot made public as CMS physics analysis summary

- Got a lot of help from TEfficieny => easy to use interface for all reasonable intervalshistograms
- Succesfully erradicated: poor error estimates for efficiencies

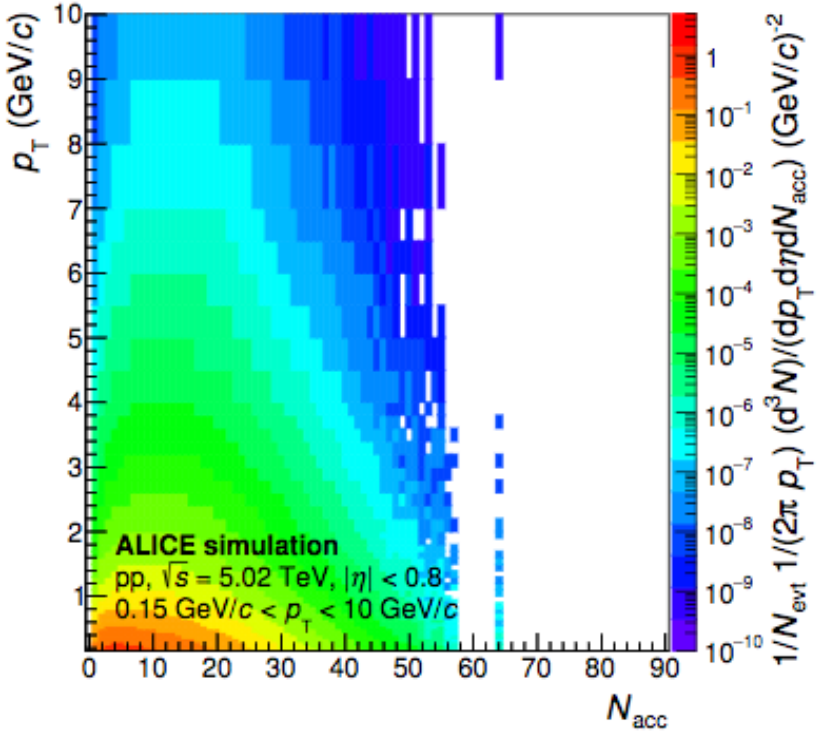
UF UNIVERSITY of FLORIDA



Unfolding

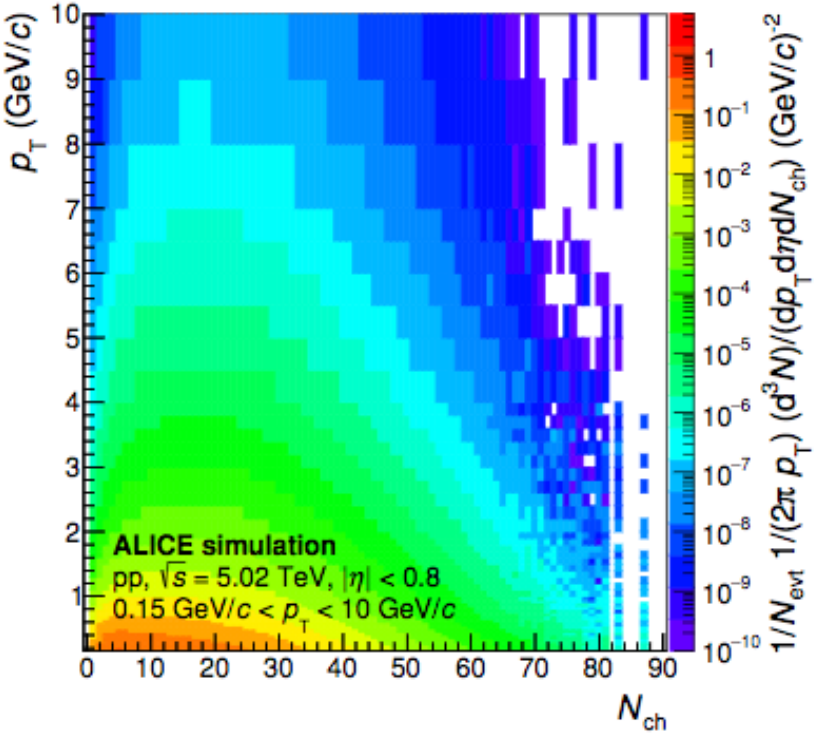
Measured and Unfolded p_T Spectra

Measured p_T spectra



ALI-SIMUL-145107

Unfolded p_T spectra



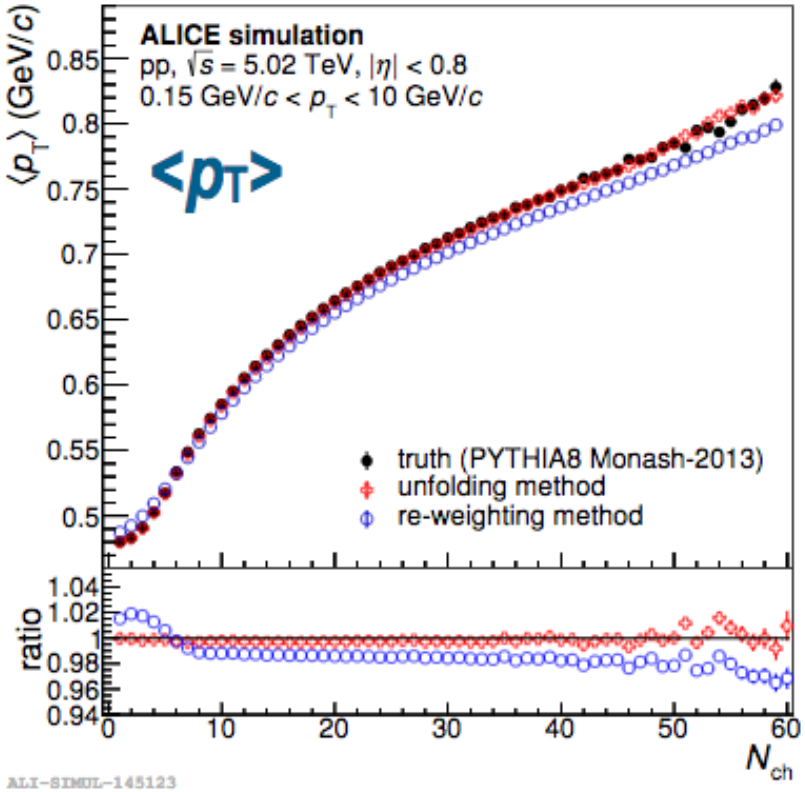
ALI-SIMUL-145111

- Multiplicity dependent charged-particle p_T spectra up to $N_{ch} \approx 80$
- Best possible resolution ($\Delta N_{ch} = 1$)

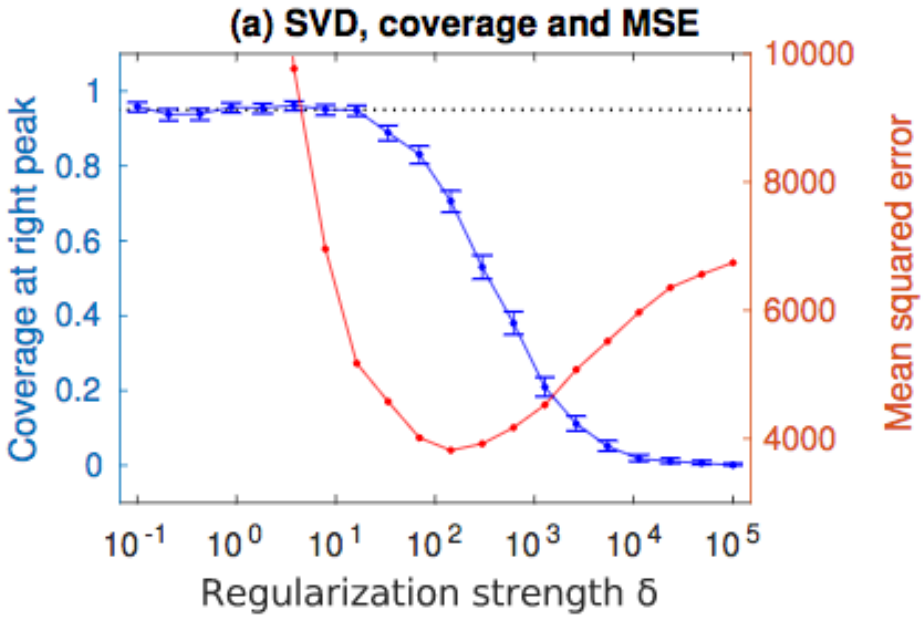
MC Studies

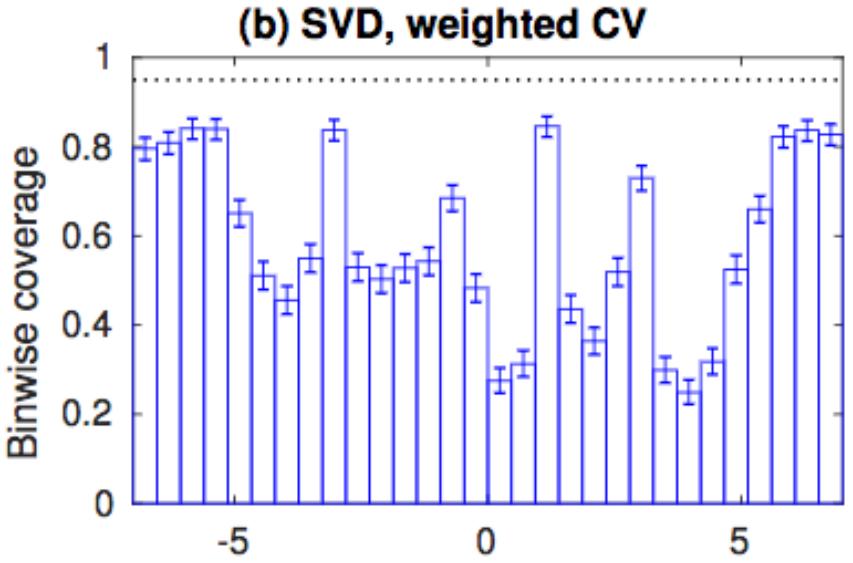
Closure test:

- Unfolding of p_T spectra from MC
- Comparison with MC truth p_T-spectra
- Difference: Important indicator for systematic uncertainty of procedure



Undercoverage of existing methods





- Optimal point estimation $\neq$ optimal uncertainty quantification
  - In terms of the uncertainties, standard methods for choosing δ tend to regularize too heavily
- Similar conclusions hold for other common methods (D'Agostini, TUnfold,...)

Binwise coverage, $\lambda^{\mathrm{MC}} = 0$

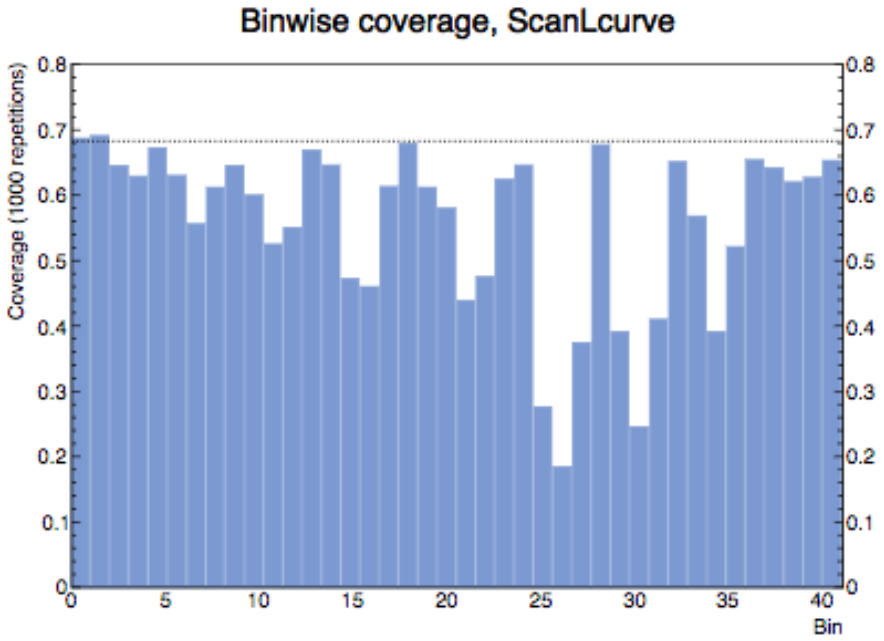


Figure: L-curve

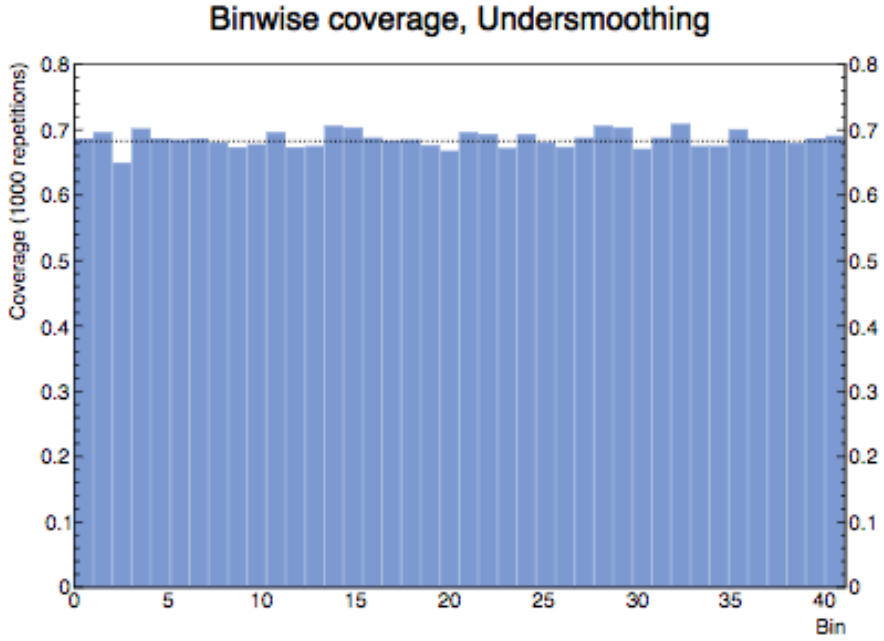
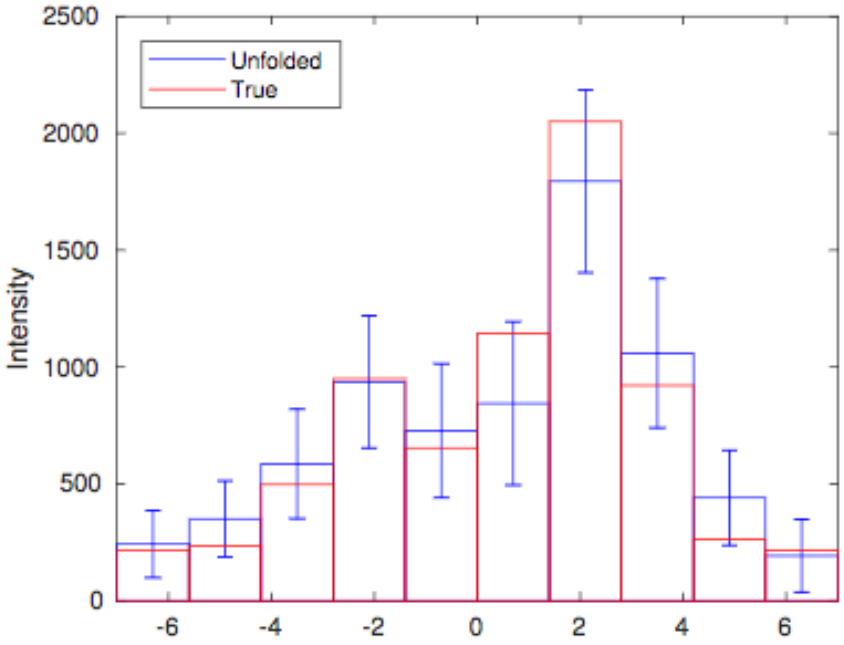
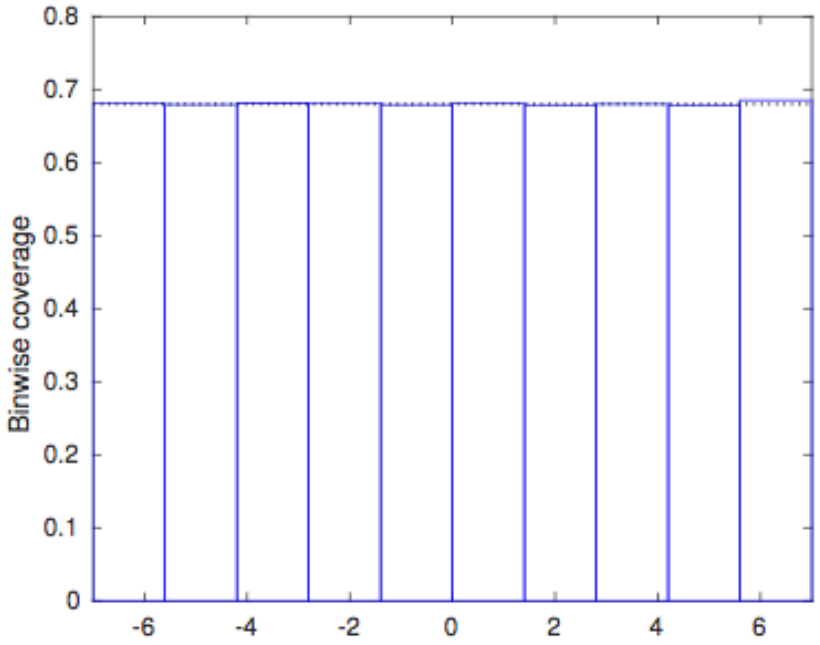


Figure: Undersmoothing

Wide bins via fine bins, perturbed MC





Wide bins via fine bins gives both correct coverage and intervals with reasonable length

Other Topics Covered

- Confidence Intrervals for Linear Poisson F. Matorras
- History of CLs A. Read
- Statistics for IN Frontier L. Stanco
- Statistics: Idealism and Reality M. Mozer
- Networked Data Science A. Ustyuzhanin
- Gaussian Processes for Q/G string parameters V. Kovalenko
- Signal Morphing L. Brenner

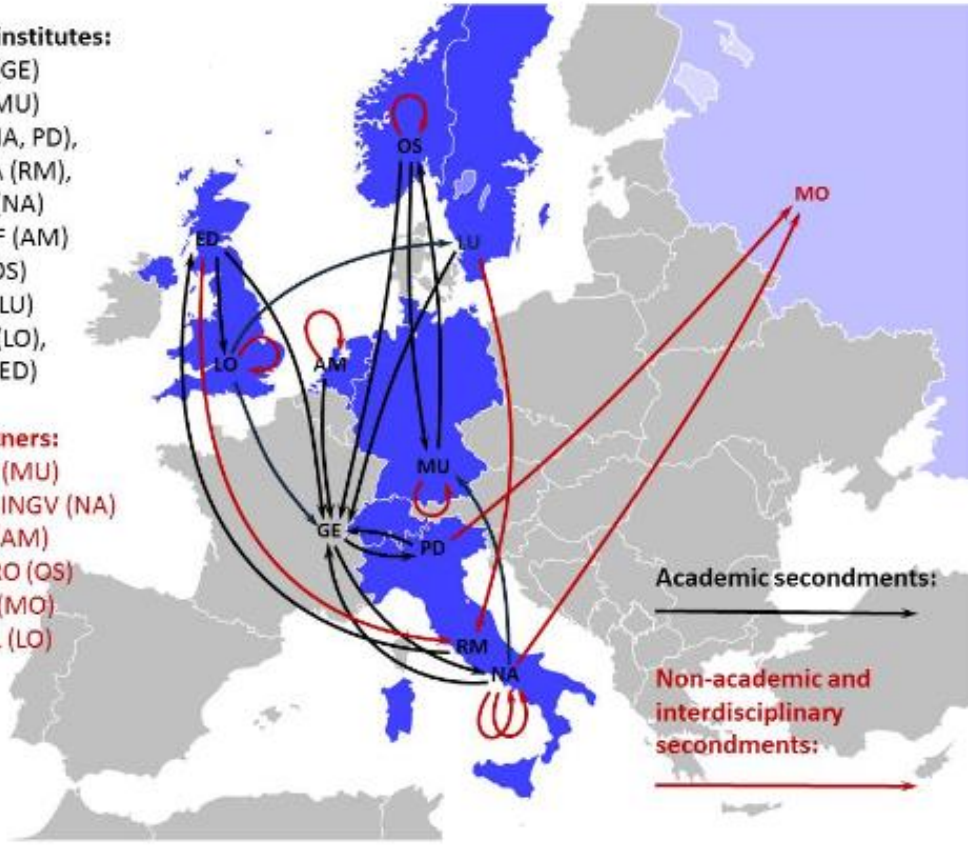
Insights

- International Training Network of Statistics for High Energy Physics and Society

- INSIGHTS is a 4-year Marie Sklodowska-Curie Innovative Training Networks project for the career development of 12 Early Stage Researchers (ESRs) at 10 partner institutions across Europe.
- INSIGHTS is focused on developing and applying latest advances in statistics, and in particular machine learning, to particle physics
- CERN is part of the network with deep interconnection with the ROOT development team

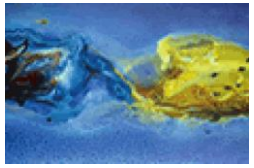
ESR hosts institutes:
CH: CERN (GE)
DE: MPP (MU)
IT: INFN (NA, PD), PANGEA (RM), UNINA (NA)
NL: NIKHEF (AM)
NO: UIO (OS)
SE: LUND (LU)
UK: RHUL (LO), UNIED (ED)

Other partners:
DE: C2PAP (MU)
IT: DCOM, INGV (NA)
NL:KPMG (AM)
NO: CICERO (OS)
RU: YNDX (MO)
UK: FISCAL (LO)



https://www.insights-itn.eu/

Summary



- Many exciting results and ideas
- Expanding number of machine learning applications
- Great progress and an opportunity to reexamine things for LHC Run 3/DUNE
- Thanks to all for making Track H a success (special thanks to….