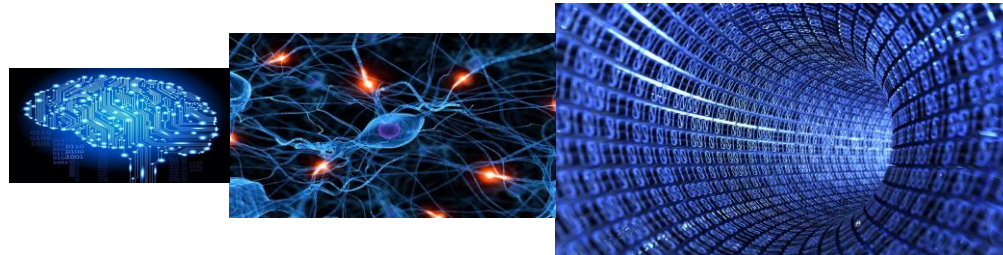


HEP-CS: Machine Learning and Algorithms

Sergei V. Gleyzer

University of Florida




S2I2 HEP-CS Workshop

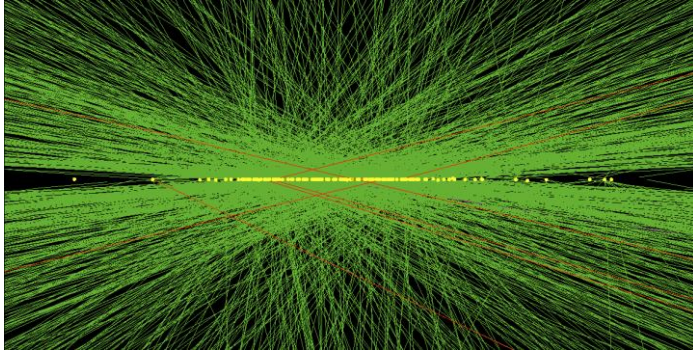
May 3, 2017

Machine Learning Session

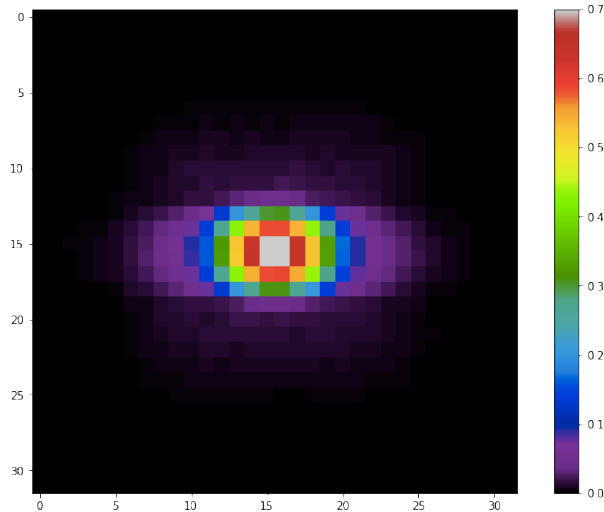


- **Machine Learning and Algorithms**
 - [Google Doc](#)
- **Participants:**
 - 50/50 HEP/CS
- **Introduction**
 - Challenges and Current Applications
- **11 Lightning Talks**  **7 HEP/ 4 CS (1 IN)**
 - Ideas, directions and questions

Current Applications

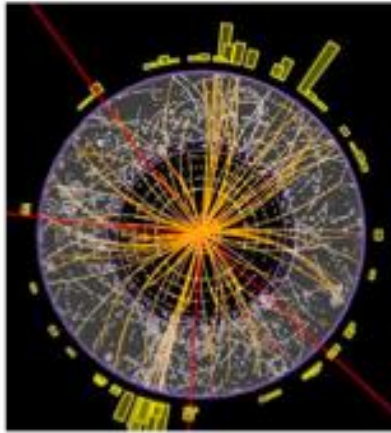


- Particle Identification
- Pattern Recognition (tracks)
- Searches for New Physics
- Data Quality Monitoring

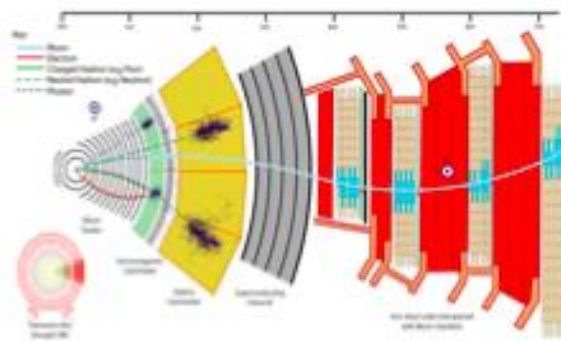


- Image Techniques
- Deep Learning
- Energy/Momentum Regression

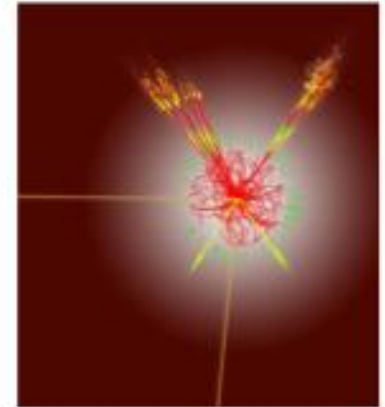
ML Applications



Tracking



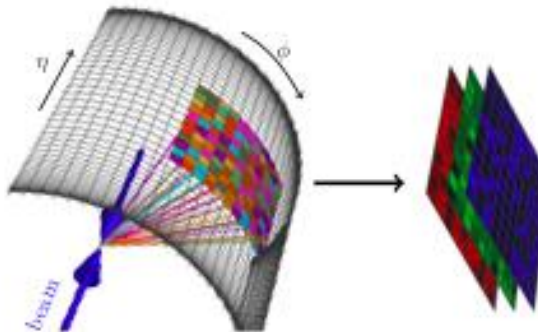
**Fast
Simulation**



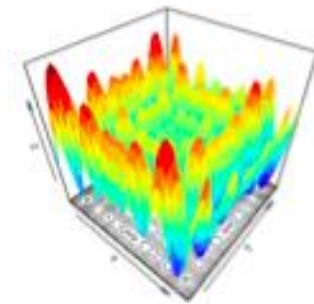
**Object
Identification**



Trigger



Imaging Calorimetry



Simulation

Lightning Talks



Variety of subjects:

- **New Trends in Machine Learning**
- **Pattern recognition for Tracking (2)**
- **ML Applications in Networking and Data Management (2)**
- **End-to-End Reconstruction and Classification with ML**

Lightning Talks



Variety of subjects:

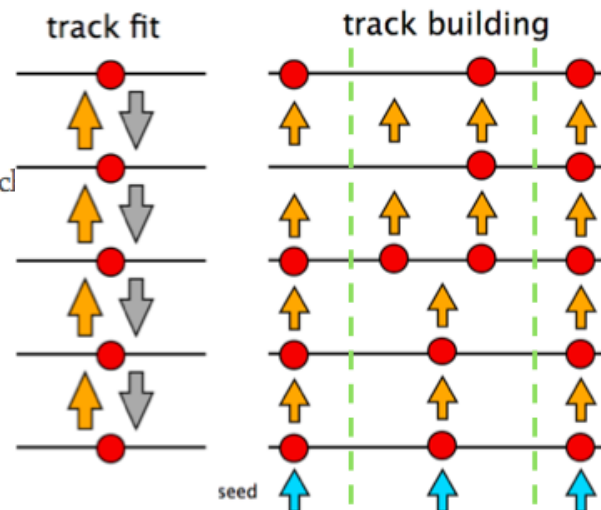
- **Optimization of ML for Physics**
- **New Algorithms (Probabilistic and Inference)**
- **Machine Learning in Simulation**
- **Machine Learning as a Service**
- **Industry Perspective**

Parallelized Kalman-Filter-Based Reconstruction of Particle Tracks on Many-Core Processors and GPUs

2nd S212 HEP/CS Workshop
May 2, 2017

G. Cerati⁴, P. Elmer³, S. Krutelyov¹, S. Lantz², M. Lefebvre³,
M. Masciovecchio¹, K. McDermott², D. Riley², M. Tadel¹, P. Witticl
F. Würthwein¹, A. Yagil¹

1. University of California – San Diego
2. Cornell University
3. Princeton University
4. Fermilab



Open ML Questions for S2I2 ML4PR

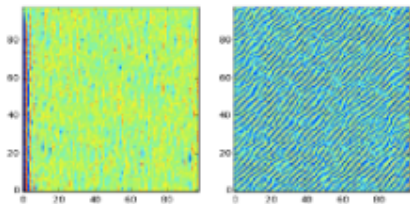
How to train on sparse images?

Incorporate tracking priors?

Through feature engineering?

Constrained training?

D. Anderson



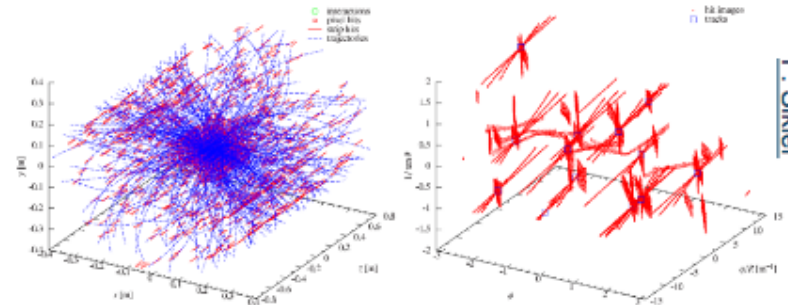
How to interpret network behaviour?
Stability against e.g. miscalibrations?

P. Calafiura

DNN performance (bandwidth, latency, scaling)

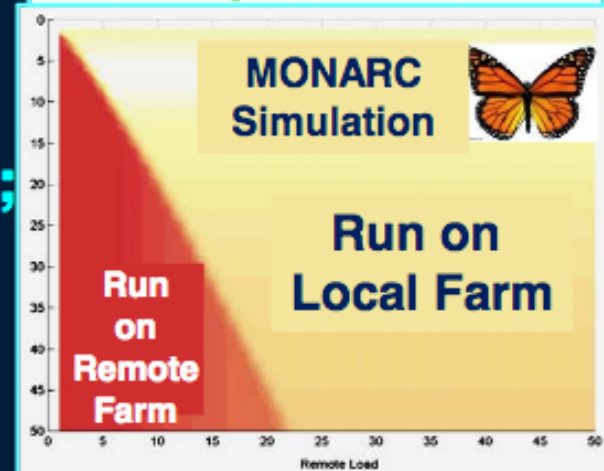
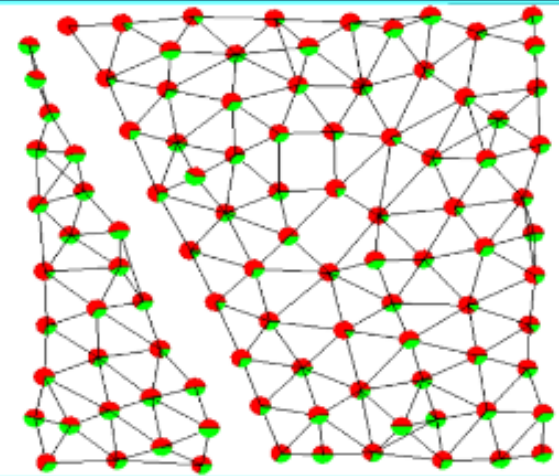
DNN optimization ("zipping", weight precision)

DNN hardware deployment (FPGAs and dedicated)



Key Developments from the HEP Side: Machine Learning, Modeling, Game Theory

- **Applying Deep Learning + Self-Organizing systems methods to optimize LHC workflow**
- **Unsupervised: to extract the key variables and functions**
- **Supervised: to derive optima**
- **Iterative and model based: to find effective metrics and stable solutions [*]**
- **Reinforced: according candidate metrics**
- **Complemented by modeling and simulation; game theory methods [*]**
- **Progressing to real-time agent-based pervasive monitoring**
- **Application to CMS Workflow**



Self-organizing neural network
for job scheduling in
distributed systems

[*] [T. Roughgarden](#) (2005). *Selfish routing and the price of anarchy*

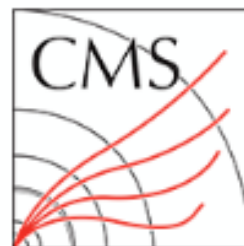
H. Newman

Exploring End-to-End Deep Learning for Event & Object Classification

Michael Andrews^{1,2}, Manfred Paulini^{1,2}, Sergei Gleyzer^{1,3}, Barnabas Poczos⁴

¹CMS, ²Carnegie Mellon University-Physics, ³University of Florida, ⁴Carnegie Mellon University-ML

S2I2 HEP/CS Workshop, 2017-MAY-02

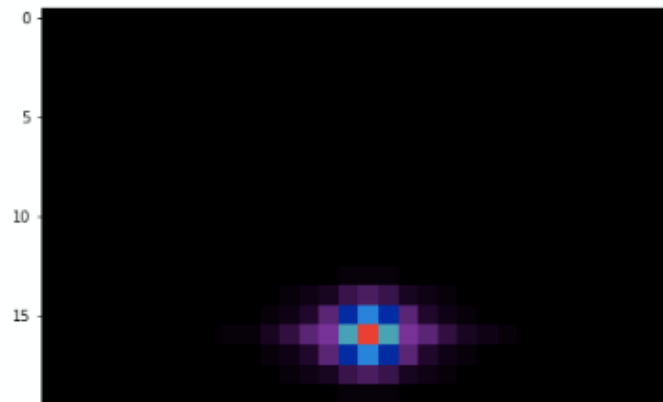


End-to-End Deep Learning



Photon-Induced EM Shower

mean energy distribution over 10k events



Network Type	Fully-Connected	Fully-Connected	CNN	CNN+LSTM
Inputs	Shower Shape Variables	Flattened Image	Stacked Images	Image Sequence
ROC AUC	0.708	0.770	0.806	0.799

Simulation

- One of the features HEP may be able to offer is we have very large, very high quality simulation sets.
 - Untold hours of effort have been devoted to making our simulations both very realistic and very detailed.
 - This is a rich playground not only for physics, but for algorithm development - it is possible to take slices of simulation that are very complex and hide and show relationships at a wide variety of levels.
 - And we have a lot of simulation! Plus huge simulated sets from different versions of our physics models...
 - We are very worried about domain differences between our data and our simulation. How do we manage this?
 - There are, of course, a lot of tricks for managing bias in training, but quantifying it is crucial for us.

G. Perdue

Learning from Industry



Meghan Kane
meghaphone.com

Software Engineer @ SoundCloud 📍 Berlin 🇩🇪

Math, CS @ MIT, 2012 🎓 🇺🇸





Code Quality

Testing (lots): infrastructure, coverage, and education

CI: each codebase can have its own

Monitoring: make it easy to see health of systems

- prometheus, dashboards, slack integrations, downtime as KPI, track on-call incidents

Tech debt: needs to be prioritized

Learn from mistakes: post mortems, no blame culture

M. Kane



How Did (Do) We Get Here (There)?

- Occam's razor: valuing simplicity & scalability
- working smart > working hard
- learning from peer companies
- connecting with Open Source Community
- postmortem culture, transparency
- investing in internal learning & development



Upskilling & Enabling Innovation

people problems are harder than technical problems

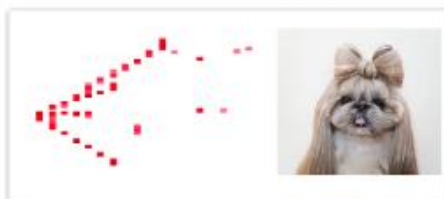
- short term: keep existing systems running
- medium / long term: upskilling people so that we can tackle the problems more elegantly and dynamically

how do we do it && stay current with industry & academia?

journal club, internal moves, tech talks, 20% hacker time, demos, open houses

M. Kane

Accounting for training sample biases



Typical issue is how to show robustness in data.

- ★ *Data driven tests*
- ★ *Training sample composition (to minimize biases which you know of a priori)*



and rough performance...

- ★ *Overall accuracy*
- ★ *Behavior of loss functions, etc*



How do we find the biases we have introduced in our training?



F. Psihas

Ensuring dependencies on the physics



She would know this is not what doggies look like in nature.



How can we make sure these algorithms incorporate the physics that we know?

- ★ *There are some alternatives out there i.e. GANs trained in data.. but this matters for any algorithm*
- ★ *Simple tests on an individual basis*

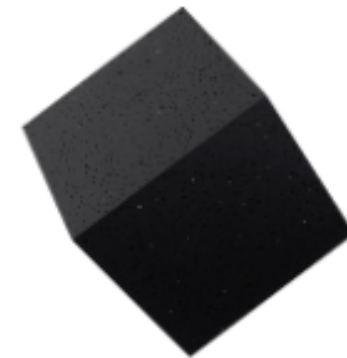
Can we develop tools to universally optimize (NOT TUNE) for the physics we understand?

F. Psihas

Edward: A library for probabilistic modeling, inference, and criticism

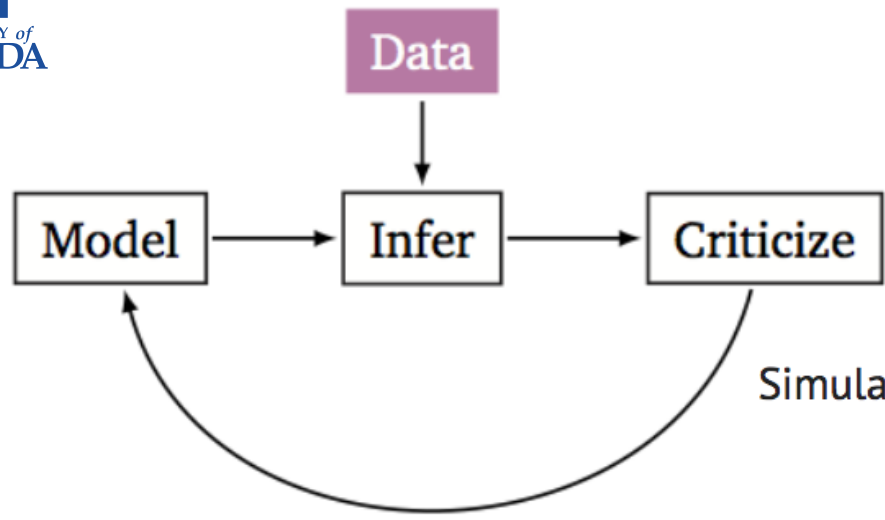
Dustin Tran, David M. Blei
Columbia University

Matt Hoffman, Rif A. Saurous, Eugene Brevdo,
Kevin Murphy
Google Brain



edwardlib.org

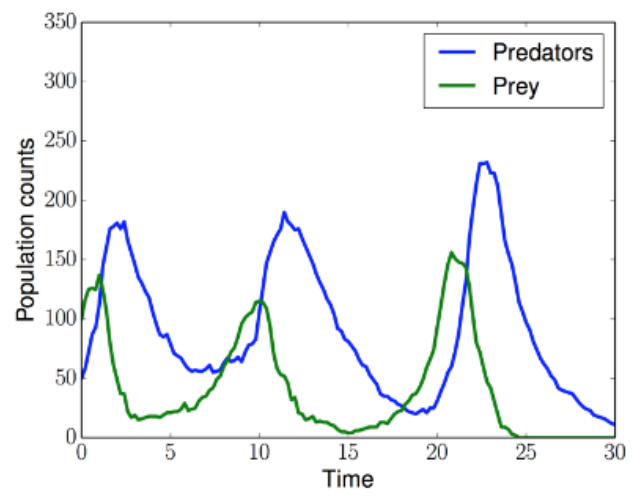
D. Tran



Simulator-based Model in Population Ecology

Edward is a library designed around this loop

[Box 1



D. Tran

Other Fantastic Ideas



- **ML as a Service and clouds**
- **ML for Data Management**
- **Others ideas in discussion and questions in live notes**

Suggestion(s) from CS



- **Present the problem without the solution**
- **Allow ideas and early collaboration**

Plan ahead

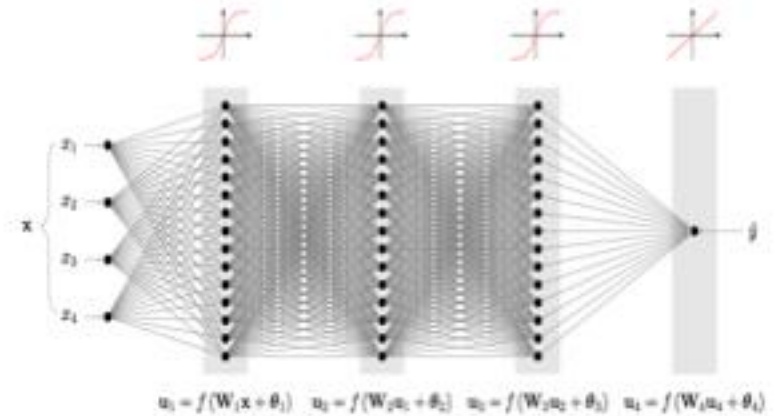


- **Agreed to put all the lightning talk ideas online in one place**
 - **To enable CS-collaboration, ideas directly**
- **Common participation in upcoming CWP-ML Part III:**
 - **DS@HEP 2017, FNAL, May 12, 2017**



Thank You

Understanding Scientific Collaboration



HEP + CS = ...
(S2I2)

Use Machine Learning

