The Data Science Challenges of Particle Physics

Kyle Cranmer,
New York University
• Experimental Particle Physicist

• Statistics Convener of ATLAS experiment at LHC

• Founder of RooStats framework (used for Higgs discovery)

• Co-lead Open Science Working Group for Moore-Sloan Data Science Environment at NYU
A harbinger for things to come

Large, Distributed Collaborations
Big Science

Complicated Sensor Environment
Big Data
Big Simulation

Scientifically Motivated Data Modeling
Big Simulation
Big Model

\[
\mathcal{L}_{SM} = \frac{1}{4} W_{\mu\nu} \cdot W^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} G_{\mu\nu} G^{\mu\nu} \\
+ \tilde{L} \gamma^\mu \left( i \partial^\mu - \frac{1}{2} g^\gamma \cdot W^\mu - \frac{1}{2} g^\gamma B^\mu \right) L + \bar{\tilde{R}} \gamma^\mu \left( i \partial^\mu - \frac{1}{2} g^\gamma \cdot Y^\mu \right) \bar{R} \\
+ \frac{1}{2} \left( i \partial^\mu - \frac{1}{2} g^\gamma \cdot W^\mu - \frac{1}{2} g^\gamma Y^\mu \right) \phi^2 - V(\phi) \\
\]

\[W, Z, \gamma, \text{and Higgs masses and couplings}
\]

\[g^\gamma(\gamma^\mu T_3) G_\mu + (G_1 L_3 R + G_2 L_5 R + h.c.)\]
Complex Models for Big Data

Max Welling
UvA

The Computational Wall: If a model has hundreds of parameters, how can we:

1) Find the parameter values that match the observations best?
2) Determine if we underfit (model too simple) or overfit (model too complex)?
3) Compare two models?

Computer simulations have become increasingly complex (e.g., weather, earthquake models)

This production run producing 360 sec of wave propagation sustained 220 Tflop/s for 24 hours on NCCS Jaguar using 223,074 cores.

The Four Paradigms

We have added big data to computer simulation, experiment and theory.

Not replaced it...

3x Exponential Growth in Machine Learning

Data Science Research Center
Max Welling
UvA

Complex Models for Big Data

Max Welling
UvA

The Computational Wall: If a model has hundreds of parameters, how can we:

1) Find the parameter values that match the observations best?
2) Determine if we underfit (model too simple) or overfit (model too complex)?
3) Compare two models?

Computer simulations have become increasingly complex (e.g., weather, earthquake models)

This production run producing 360 sec of wave propagation sustained 220 Tflop/s for 24 hours on NCCS Jaguar using 223,074 cores.

The Four Paradigms

We have added big data to computer simulation, experiment and theory.

Not replaced it...

3x Exponential Growth in Machine Learning

Data Science Research Center
Max Welling
UvA

Complex Models for Big Data

Max Welling
UvA

The Computational Wall: If a model has hundreds of parameters, how can we:

1) Find the parameter values that match the observations best?
2) Determine if we underfit (model too simple) or overfit (model too complex)?
3) Compare two models?

Computer simulations have become increasingly complex (e.g., weather, earthquake models)

This production run producing 360 sec of wave propagation sustained 220 Tflop/s for 24 hours on NCCS Jaguar using 223,074 cores.

The Four Paradigms

We have added big data to computer simulation, experiment and theory.

Not replaced it...

3x Exponential Growth in Machine Learning

Data Science Research Center
Max Welling
UvA
$L_{SM} = \sum_{i} \frac{1}{4} W_{\mu i} \cdot W^{\mu i} - \frac{1}{4} B_{\mu i} B^{\mu i} - \frac{1}{4} G_{a \mu i} G^{a \mu i}$

- Kinetic energies and self-interactions of the gauge bosons

$\sum_{i} \frac{1}{2} g_{\mu i} W_{\mu} - \frac{1}{2} g B_{\mu} L + \bar{R} \gamma^\mu Q_{\mu} L + \frac{1}{2} g' Y B_{\mu} R$

- Kinetic energies and electroweak interactions of fermions

$\frac{1}{2} \left| \left| \left( i \partial_{\mu} - \frac{1}{2} g \tau^\mu W_{\mu} - \frac{1}{2} g' Y B_{\mu} \right) \phi \right| \right|^2$

- $W^\pm, Z, \gamma, and Higgs masses and couplings

$\frac{1}{2} \left( T_{a} T_{b} \right) G_{\mu}^{a \mu} + \left( G_{1} \bar{L} \phi R + G_{2} \bar{L} \phi c R + h.c. \right)$

- Interactions between quarks and gluons

$\frac{1}{2} \left| \left| \left( i \partial_{\mu} - \frac{1}{2} g' Y B_{\mu} \right) \phi \right| \right|^2$

- Fermion masses for generations to Higgs
Non-trivial aspects of the theory have been tested to $< 1$ ppm

A unique realm for reasonable statistical exploration of a scientific theory

\[ a_\mu \,(\text{exp}) = 11\,659\,208\,(6) \times 10^{-10} \,(0.5\,\text{ppm}) \]
Overview of Predictions

1) The language of the Standard Model is Quantum Field Theory

2) Perturbation Theory, Feynman Diagrams, and Factorization are used to construct Monte Carlo simulations of the interactions

3) The interaction of outgoing particles with the detector is simulated.

4) Finally, we run algorithms on the simulated data as if they were from real collisions.
Number of collisions

expected number of scatterings = cross section [cm$^2$] x Luminosity [1/cm$^2$]

80 mb $\cdot$ 25 fb$^{-1}$ = $2 \cdot 10^{15}$ collisions

signal : background $\sim 1 : 10^9$
The steady march of progress
**ATLAS Experiment**

Run: 154822, Event: 14321500
Date: 2010-05-10 02:07:22 CEST

\[ p_T(\mu^-) = 27 \text{ GeV} \quad \eta(\mu^-) = 0.7 \]
\[ p_T(\mu^+) = 45 \text{ GeV} \quad \eta(\mu^+) = 2.2 \]
\[ M_{\mu\mu} = 87 \text{ GeV} \]

**Z\rightarrow\mu\mu candidate in 7 TeV collisions**
The steady march of progress
Top quark pair decaying to $bb \, e\mu \, E_{T,\text{miss}}$
The steady march of progress
How good is the modeling?
How good is the modeling?

\[ \int L \, dt = 1.08 \, fb^{-1} \]
\[ \sqrt{s} = 7 \, \text{TeV} \]
Complex Models for Big Data

Max Welling
UvA

Big Simulation

Computer simulations have become increasingly complex (e.g. weather, earthquake models)

This production run producing 360 sec of wave propagation sustained 220 Tflop/s for 24 hours on NCCS Jaguar using 223,074 cores.

The Computational Wall: If a model has hundreds of parameters, how can we:

1) Find the parameter values that match the observations best?
2) Determine if we underfit (model too simple) or overfit (model too complex)?
3) Compare two models?

The Four Paradigms

We have added big data to computer simulation, experiment and theory.

Not replaced it...

3x Exponential Growth in Machine Learning

Computer Power
Data Volume
Model Capacity

Data is Growing Exponentially

(next slide)
Use of Machine Learning:

Particle-Level and Event-Level
$H \rightarrow ZZ \rightarrow 4l$
Putting the Higgs back together again

Don’t believe the media: $E \neq mc^2$

What Einstein really said:

$E^2 = (mc^2)^2 + (|\vec{p}|c)^2$

Every physics student knows energy and momentum are conserved

$E_{\text{Higgs}} = E_{\text{before}} = E_{\text{after}} = \sum_i E_i$

$\vec{p}_{\text{Higgs}} = \vec{p}_{\text{before}} = \vec{p}_{\text{after}} = \sum_i \vec{p}_i$

Thus, we can estimate the mass of the Higgs with

$m_H = \sqrt{E_{\text{after}}^2 / c^4 - |\vec{p}_{\text{after}}|^2 / c^2}$
An example high-level feature

From the 16 energies and momenta measured in this system, this particular combination gives a very sharp feature.

~sufficient statistic

\[ m_H = 130 \text{ GeV} \]

\[ m = (129.72 \pm 0.03) \text{ GeV} \]

\[ \sigma = (1.78 \pm 0.03) \text{ GeV} \]

fraction outside \( \pm 2\sigma \): 19%
The observation in the 4l channel

![Plot showing the distribution of the four-lepton reconstructed mass in the full mass range for the sum of the 4e, 2e2µ, and 4µ channels. Points with error bars represent the data, shaded histograms represent the backgrounds, and the unshaded histogram the signal expectation for a mass hypothesis of $m_H = 126$ GeV. Signal and ZZ background are normalized to the SM expectation, Z+X background to the estimation from data. The expected distributions are presented as stacked histograms. No events are observed with $m_{4l} > 800$ GeV.]

Table 9: The number of observed candidate events compared to the mean expected background and signal rates for each final state. Uncertainties include statistical and systematic sources. The results are given integrated over the full mass measurement range $m_{4l} > 100$ GeV and for 7 and 8 TeV data combined.

<table>
<thead>
<tr>
<th>Channel</th>
<th>4e</th>
<th>2e2µ</th>
<th>4µ</th>
<th>ZZ background</th>
<th>Z+X background</th>
<th>All backgrounds</th>
<th>$m_H = 500$ GeV</th>
<th>$m_H = 800$ GeV</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 TeV</td>
<td>5</td>
<td>13</td>
<td>8</td>
<td>1.0 ± 0.1</td>
<td>0.8 ± 0.2</td>
<td>1.8 ± 0.2</td>
<td>3.0 ± 0.4</td>
<td>0.7 ± 0.1</td>
<td>4</td>
</tr>
<tr>
<td>8 TeV</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>1.1 ± 0.1</td>
<td>0.9 ± 0.2</td>
<td>2.0 ± 0.2</td>
<td>5.2 ± 0.5</td>
<td>0.7 ± 0.1</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 10: The number of observed candidate events compared to the mean expected background and signal rates for each final state. Uncertainties include statistical and systematic sources. The results are integrated over the mass range from 121.5 to 130.5 GeV and for 7 and 8 TeV data combined.

<table>
<thead>
<tr>
<th>Channel</th>
<th>4e</th>
<th>2e2µ</th>
<th>4µ</th>
<th>ZZ background</th>
<th>Z+X background</th>
<th>All backgrounds</th>
<th>$m_H = 500$ GeV</th>
<th>$m_H = 800$ GeV</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 TeV</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>0.3 ± 0.1</td>
<td>0.2 ± 0.0</td>
<td>0.5 ± 0.1</td>
<td>1.1 ± 0.1</td>
<td>0.8 ± 0.2</td>
<td>1</td>
</tr>
<tr>
<td>8 TeV</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.3 ± 0.1</td>
<td>0.2 ± 0.0</td>
<td>0.5 ± 0.1</td>
<td>0.7 ± 0.1</td>
<td>0.9 ± 0.2</td>
<td>0</td>
</tr>
</tbody>
</table>

---

Kyle Cranmer (NYU)  
Paris-Saclay Center for Data Science  
June 30, 2014
ML techniques performed poorly unless these high-level features were supplied. Deep learning techniques can discover them.
$H \rightarrow \gamma \gamma$
The observation in the 2 photon channel

ATLAS

Data 2011+2012
SM Higgs boson $m_H = 126.8$ GeV (fit)
Bkg (4th order polynomial)

$\sqrt{s} = 7$ TeV $\int L dt = 4.8$ fb$^{-1}$

$\sqrt{s} = 8$ TeV $\int L dt = 20.7$ fb$^{-1}$

Kyle Cranmer (NYU)
Particle Identification

\[ e^+ + e^- \rightarrow \bar{t}t \rightarrow b\bar{b}W^+W^- \]

\[ \gamma \]

\[ \pi^0 \]

\[ u \]

\[ u \]
Machine Learning in $H\rightarrow\gamma\gamma$

Sensor → Particle → Event → Dataset

- **EM Cluster** (RAW Energy, Shower Shape, Local/Global Coords)
- **Reconstructed Tracks**
- **ECal and HCal Deposits**
- **Primary Vertex Reconstruction**
- **Conversion Reconstruction**
- **Isolation Sums**
- **Selected Primary Vertex**
- **Primary Vertex Probability MVA**
- **Primary Vertex Selection MVA**
- **Di-photon Mass Resolution Estimate**
- **Mass-Factorized Kinematics**
- **Di-photon MVA**
- **Per-Event Mass Resolution Estimate**
- **Per-Event Primary Vertex Probability**
- **Categorized Mass Fits**
- **Categorize and Count MVA**

*MVA = BDT implemented in TMVA (Deep networks being used for particle identification)*
Reaching out to the ML community

Higgs Boson Machine Learning Challenge

Monday, May 12, 2014
$13,000 • 837 teams
Monday, September 15, 2014

Dashboard

Home
Data
Make a submission

Information
Description
Evaluation
Rules
Prizes
About the Sponsors
Timeline

Competition Details » Get the Data » Make a submission

Evaluation

The evaluation metric is the approximate median significance (AMS):

\[
AMS = \sqrt{2 \left( (s + b + b_r) \log \left( 1 + \frac{s}{b + b_r} \right) - s \right)}
\]

1 11 Gábor Melis * 3.80573 32 Thu, 26 Jun 2014 06:14:34 (-0.2h)

359 49 Jeje 3.25012 4 Sat, 21 Jun 2014 01:11:13

simple TMVA boosted trees 3.24954

360 49 Xiaohu SUN 3.24954 3 Tue, 03 Jun 2014 13:14:47
Statistical Modeling for Higgs Discovery
I will represent PDFs graphically as below (directed acyclic graph)

- eg. a Gaussian $G(x|\mu, \sigma)$ is parametrized by $(\mu, \sigma)$
- every node is a real-valued function of the nodes below

Clearly related to Graphical Models, but not the focus here.
Total distribution is a mixture model with components corresponding to various signal and background interactions:

\[ f(x) = \frac{1}{\nu} \sum_{i \in \text{interactions}} \nu_i f_i(x), \quad \nu = \sum_{i \in \text{interactions}} \nu_i \]

\[ \begin{array}{c}
\text{Events / 5 GeV} \\
10^6 \\
10^5 \\
10^4 \\
10^3 \\
10^2 \\
10^1 \\
10^{-1} \\
10^{-2} \\
10^{-3} \\
0 \\
50 \\
100 \\
150 \\
200 \\
250 \\
\end{array} \]

\[ \int L \, dt = 35 \, \text{pb}^{-1} \]

\[ \sqrt{s} = 7 \, \text{TeV} \]

\[ H \rightarrow \text{ee} + \nu\bar{\nu} \quad (m_H = 400 \, \text{GeV}) \]

\[ \nu_i \]
Incorporating Systematic Effects

Tabulate effect of individual variations of sources of systematic uncertainty

- typically one at a time evaluated at nominal and “± 1 σ”
- use some form of interpolation to parametrize $p^{th}$ variation in terms of nuisance parameter $\alpha_p$
Visualizing the model for one dataset
After parametrizing each component of the mixture model, the pdf for a single channel might look like this.
RooFit’s Workspace now provides the ability to save in a file the full likelihood model, any priors you might want, and the data necessary to reproduce likelihood function.

Gives flexibility in later statistical analysis (frequentist vs. bayesian) and handles for detailed meta-analysis.
Collaborative Statistical Modeling
Collaborative Statistical Modeling
REPRODUCIBILITY PROBLEM

Not possible for others to reproduce results from paper.

Figure 4: Fits for 2-parameter benchmark models probing different coupling strength scale factors for fermions and vector bosons: (a) Correlation of the coupling scale factor $\kappa_F$ and $\kappa_V$, assuming no non-SM contribution to the total width; (b) Correlation of the coupling scale factors $\lambda_{FV} = \kappa_F / \kappa_V$ and $\kappa_{VV} = \kappa_V \cdot \kappa_V / \kappa_H$ without assumptions on the total width.

$\sqrt{s} = 7\,\text{TeV}, \int L dt = 4.8\,\text{fb}^{-1}$
$\sqrt{s} = 8\,\text{TeV}, \int L dt = 5.8-5.9\,\text{fb}^{-1}$

ATLAS Preliminary

+ SM
× Best fit
- $-2 \ln \Lambda(\kappa_V,\kappa_F) < 2.3$
--- $-2 \ln \Lambda(\kappa_V,\kappa_F) < 6.0$
Note, data record itself has 4 citation
Reproducing derived results from original paper!
CODE AS A RESEARCH PRODUCT

GitHub → Zenodo → INSPIRE

Mathematica → figshare → INSPIRE

Supplementary Material for "A Novel Approach to Higgs Coupling Measurements"
(2013) figshare.
RECASTING

same-sign leptons+2jets

$Λ^2$

$C^2/Λ^4 = 10$

Use 4f effective operators (LL, LR, RR) modes

Many models predict ss tops (esp. to explain CDF top A $173 \text{ fb}$)

Squark pairs +WW, ZZ modes

CDF RunII Preliminary

Events

Obs-Exp

$H_T [\text{GeV}]$

$\bar{q}q 200 \text{ GeV}/c^2$

$\bar{q}q 200 \text{ GeV}/c^2$

$H_T [\text{GeV}]$

$H_T [\text{GeV}]$
\[ \mathcal{L}_{\text{SM}} = \frac{1}{4} W_{\mu\nu} \cdot W^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} G_{\mu\nu} G^{\mu\nu} \]

kinetic energies and self-interactions of the gauge bosons

\[ + \mathcal{L}^{\text{q}}(i \partial_\mu - \frac{1}{2} g \gamma \cdot W_\mu L + R \gamma_\mu (i \partial_\mu - \frac{1}{2} g' Y B_\mu) R) \]

kinetic energies and electroweak interactions of fermions

\[ + \frac{1}{2} \left( (i \partial_\mu - \frac{1}{2} g \gamma \cdot W_\mu) \phi \right)^2 - V(\phi) \]

\[ W^\pm, Z, \gamma, \text{and Higgs masses and couplings} \]

\[ + g'' (q T^a T^b) G_a^{\mu\nu} + (G_1 \bar{L}_e R + G_2 \bar{L}_\mu R + h.c.) \]

interactions between quarks and gluons

fermion masses and couplings to Higgs
Review of Challenges and Possible Research Topics
Challenges & possible research areas

The complexity of our statistical models is growing exponentially, starting to need new algorithms to deal with them or principles for simplifying them

- graphical models, automatic differentiation, distributed processing, …
- better optimization & sampling algorithms
- optimal statistical procedures subject to computational constraints (link)

Interpolation of distributions based on simulated samples with different parameter settings a weak point

- experimental design, response surface interpolation, Gaussian processes, …
- complication: samples often not statistically independent

Machine learning + computer simulations

- Most analyses either use computer simulations of the detector or ad hoc parametrized models.
- Little use of machine learning to learn the expensive computer simulation
Most discussion with statisticians has focused on hypothesis testing and confidence intervals for final results. Many interesting problems up-stream

- **exception**: machine learning for selecting candidate signal events
- **barriers**: collaborations do not openly share data, requires some semi-formal agreement
- **progress**: movement towards open access (link to policy)

**Importance sampling for rare events in simulation**

- The simulation of our detectors is very computationally challenging and we use brute force to populate tails in cases where we can do something smarter

**Particle physics is a unique arena for data science**

- well posed questions in an extreme setting
- lots of data, complicated sensor environment, strong theoretical basis

**Congratulations and best wishes to**