QCHS XIII, Maynooth University
August 1-6, 2018

Statistical Methods for Physics Analysis in the XXI Century

Introduction and Goals

Tommaso Dorigo
INFN – Sezione di Padova

Motivations of this Session

- Basic Research in Fundamental Physics relies more and more on advanced statistical tools for point and interval estimation, hypothesis testing
  - a Ph.D. student at the LHC will spend more time developing and applying statistical tools than using his or her knowledge of QFT

- The use of machine-learning algorithms has become common and widespread, but the real insight of the physicists community in these tools is perfectible
  - Bertolucci's analogy: data science is like teenage sex

- There is a language barrier between physicists and statistics / data science
  - Let's break it

→ A parallel session where advanced statistics tools are discussed should be part of any major gathering of physicists doing basic research

Language Check

Physicists say	Statisticians say
Determine	Estimate
Estimate	Guess
Observable space	Population
Observe	Draw a sample
Data	Sample
Uncertainty	Error
Systematic	Nuisance parameter

Goals of this Session

- Let us try to throw a stone in a still pond, to draw attention to the problem: our community should become aware of the centrality of the issues we discuss, in all experimental work we do (and in quite a good chunk of theoretical work, too) – and consequently increase its attention to them
  - This is just one of several initiatives in this direction: Phystat, Phystat-nu...
  - But being residents of a physics conference changes the audience, increases the scope, builds bridges between communities, has more evangelicalical value

- **Goals of this session:**

  Education / information
  - get informed on current statistics practice in fundamental physics research
  - foster an increase of the statistics knowledge base of physicists
  - become aware of malpractice / suboptimal exploitation of data, find ways to steer the field in the right direction

  Development / innovation
  - identify areas where good ideas (old and new) may become drivers of significant advancements
  - Propose new use cases for existing tools borrowed from ML & data science
  - Discuss wish-list of new tools for existing use cases in our specific research area

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?
- What other topics should I add to this list ?

Organization of the Sessions

- Three afternoons: Aug 1, Aug 2, Aug 3
  - divided in six sessions by coffee breaks
- Talks are 15'+5' or 25'+5'
  - Let us try to spend well our discussion time... Please contribute with insight and meaningful questions
    - Speakers: stick to allotted time!
    - and please upload talks in advance!
    - Audience: please keep to a minimum interruptions during the presentations
- Sessions originally divided by topic: Bayesian inference, Classical statistics, Statistical learning
  - But speakers' travel plans messed this up significantly!
  - Agenda in indico still not perfectly aligned → see next slides

Today, Aug 1

14.00-14.20	T. Dorigo: Introduction and goals of the session
14.20-14.40	M. Krueger: Bayesian unfolding of charged particle pT spectra with ALICE
14.40-15.10	P. Vischia: Pseudosignificances as figures of merit: a systematic study and Bayesian solutions
15.10-15.40	H. Prosper: The Bayesian interpretation of Deep Neural Networks
15.40--16.10	Coffee break
16.10-16.30	V. Kovalenko: Determination of the quark-gluon string parameters from the data on pp, pA and AA collisions at wide energy range using Bayesian Gaussian Process Optimization
16.30-16.50	L. Brenner: ATLAS continuous signal modeling
16.50-17.20	L. Stanco: Statistics and data analysis for neutrino experiments
17.20-17.50	~~M. Stoye: Machine learning for hypothesis testing in HEP~~

Aug 2

14.00-14.30	L. Lista: Managing many simultaneous systematic uncertainties
14.30-15.00	F. Matorras: Neural networks and machine learning tools for global PDF analyses
15.00-15.30	~~L. Moneta~~ S. Gleyzer: New Machine Learning Tools in ROOT-TMVA
15.30-16.00	J. Rojo: Neural networks and machine learning tools for global PDF analyses
16.00-16.30	Coffee break
16.30-16.50	L. Graczyowsky: Using Machine Learning methods for improving data quality in the ALICE experiment
16.50-17.10	G. Strong: Recent developments in deep learning applied to open physics data
17.10-17.30	G. Kotkowski: Model independent searches for new physics via parametric anomaly detection
17.30-18.00	M. Kuusela: Unfolding: Point Estimation, Uncertainty Quantification and Future Directions
18.00-18.30	J. Pivarsky: Big data software in HEP

Aug 3

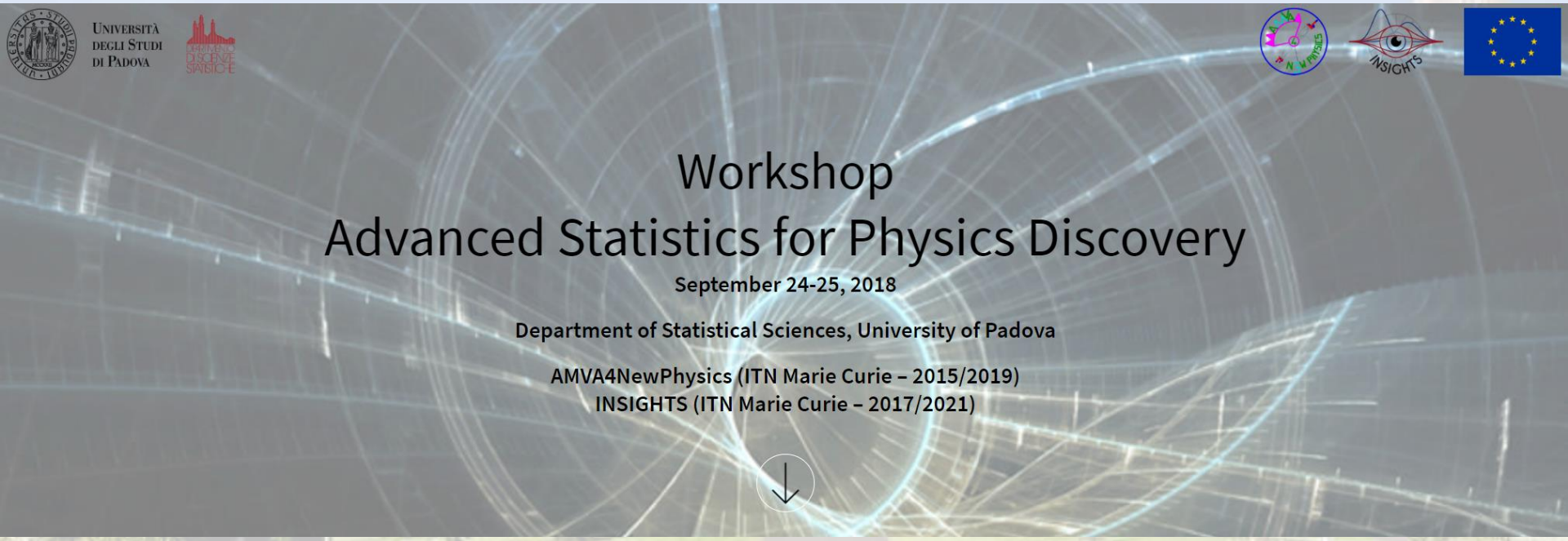
14.00-14.30	A. Ustyuzhanin: Networked data-science for research, academic communities and beyond
14.30-15.00	P. De Castro: Direct learning of systematics-aware summary statistics
15.00-15.30	A. Valassi: Fisher information metrics for binary classifier evaluation and training
15.30-15.40	Discussion
15.40-16.10	Coffee break
16.10-16.30	A. Di Florio: Convolutional Neural Network for Track Seed Filtering at the CMS HLT
16.30-17.00	A. Read: TBD
17.00-17.30	M. Mozer: Statistics in HEP: ideals vs reality
17.30-17.50	S. Gleyzer: Concluding remarks

And next Month...

I would like to advertise a 2-day workshop in Padova, organized by the AMVA4NewPhysics and INSIGHTS EU networks:

Advanced Statistics for Physics Discoveries

September 24-25, Botanical Garden, Padova (Italy)



With flash talks, a panel session, and a poster session for young participants on the evening of Sep 24

See web page of event:
http://aspd.stat.unipd.it/