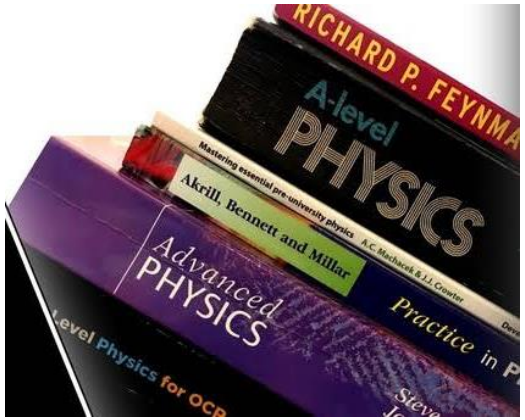
Master Degree in *Physics of Data*

Marco Zanetti,

University of Padua and INFN

The Gap

What students are taught



Skills/knowledge required in physics research

- Back in the days, students self-taught technical skills on scientific computing
- The gap is constantly increasing, that approach doesn't work anymore

Stating that loud

Physics of Data

- A seminal Nature commentary by J. Byers elaborating on how connected Physics and Data Science are

commentary

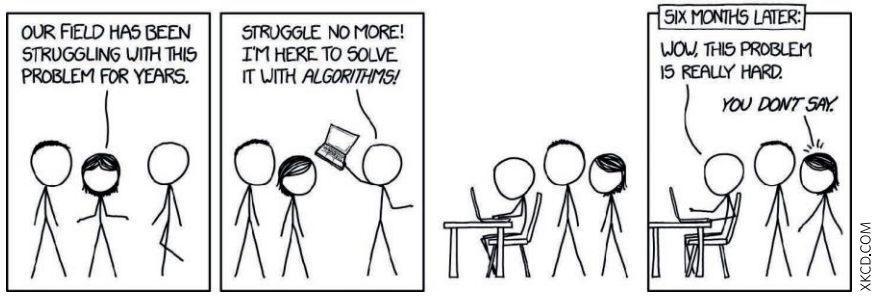
The physics of data

Jeff Byers

Physicists are accustomed to dealing with large datasets, yet they are fortunate in that the quality of their experimental data is very good. The onset of big data has led to an explosion of datasets with a far more complex structure — a development that requires new tools and a different mindset.

Ernest Rutherford is said to have famously declared that "If you need statistics then you should have done a better experiment." This observation still strikes a chord with most physicists today: complex, messy experiments tend to be viewed as bad experiments, and poor data quality is, often correctly, attributed to poor experimental design.

Then again, I guess Rutherford would have made a terrible astronomer. Astronomers have to follow a vastly different path than physicists, since in their



Manifesto

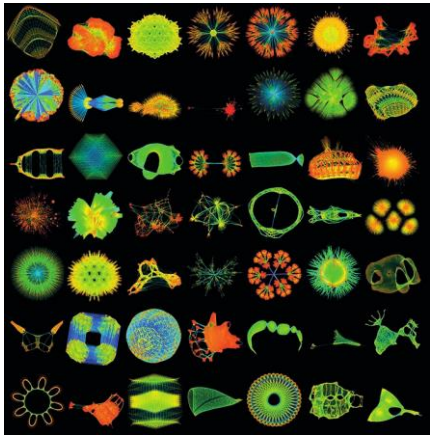
editorial

The thing about data

The rise of big data represents an opportunity for physicists. To take full advantage, however, they need a subtle but important shift in mindset.

Let's face it, we've all been there: the economist making a prediction — down to a tenth of a percent of GDP — of the effect a particular policy will have on the economy. The social psychologist drawing counter-intuitive conclusions from a study based on a tiny sample size. The web developer presenting the results of a questionable A/B test as definitive evidence that readers prefer the version of the *Nature Physics* website without a link to the current issue. "These people don't know what they're talking about," we think. "If only they knew how to analyse data properly."

Physicists, it is fair to say, like to think they understand data. They have the mathematical tools and the empirical expertise to work out the causal relationships between things. Sure, certain systems are more complicated than others — chiefly those made up of chemical, biological or social components — but a well-designed



A gallery of large graphs displaying complex data structures. Image credit: © 2011 ACM. Reproduced from T. A. Davis & Y. Hu, *ACM Trans. Math. Softw.* **38,** 1; 2011.

one such example: up to a fitting parameter that sets the energy scale, this model works insofar as it fits the experimental data accounting for the ferromagnetism in a lump of, say, iron. But it remains a caricature of reality, one that rests on multiple assumptions and approximations.

By contrast, discriminative models do not provide a mechanism for how the data might have been generated. Instead, by using a set of techniques that are best thought of as a supervised learning approach, these treat the experimental data as a direct input, which is then used to iteratively improve the model that fits it. Byers refers to this prescription as "letting the model fluctuate around the data", an approach that is possible thanks Bayes's theorem.

Indeed, there is plenty that physicists can learn from machine learning, and in order to bridge this gap Byers advocates for a greater exposure of physics students

Nature editorial expressing the principles upon which we designed the project

Manifesto

"Indeed, there is plenty that physicists can learn from machine learning, and in order to bridge this gap, Byers advocates for a greater exposure of physics students to statistics and probability, as well as information theory and computing"

"To some, it may come as a surprise that mathematical techniques originally developed in physics, such as those required to compute the partition function, have been exported to other domains of research and developed further. It is perhaps time physicists learn about these developments and take them back. The upshot of learning to work with messy data is that there are countless interesting problems to address in this way, and there are countless companies that are willing to pay scientists that are able to do so."

"As complexity features ever more prominently in the realm of physics, we need a new generation of physicists equipped with the tools to rise to the challenges this poses."

The project in a nutshell

A Master Degree (2 years) aiming at combine the classic competences of a Physicist with those of a Data Scientist

Twofold training goal

- Match the job market requests
- Provide physicists with tools to effectively tackle modern science challenges

A chance for us too to consolidate and formalize a research line

- Its multidisciplinary nature is of paramount importance

The project started in 2018, first students graduating next year

Educational principles

A sound and advanced education focused on a given physics area:

- Fundamental interactions, Statistical Mechanics, Astrophysics/cosmology, Medical Physics, etc.

Learn by doing is our mantra

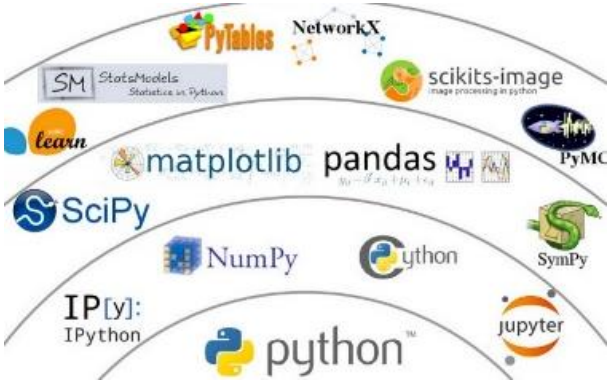
- Each class includes lab activities
- In each class students are assigned a research project upon which the final score is based

Mandatory internship in a private company or research institute/foreign university

Course Pillars (1)

Laboratory of Computational Physics

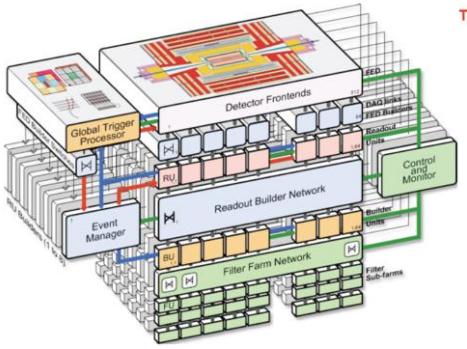
- Targeting the learning and practicing of modern techniques for data analysis and processing
- 96 hour total
- Mod. A:
  - Python ecosystem of scientific libraries, dataset manipulation, data visualization, features extraction and statistical assessment; Monte Carlo methods, simulations.
- Mod. B:
  - Applications of dvanced Machine Learning
  - Lab experiences divided per research area
    - E.g. analysis of LHC data, Plank, genomic datasets
    - Supported by experts in various fields

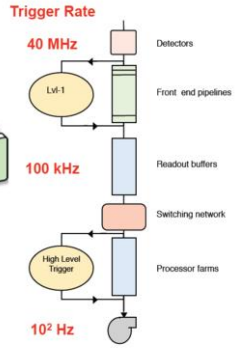


Course Pillars (2)

Management and Analysis of Physics Datasets

- The data flow from sensors to end users
- 96 hours total
- Mod A:
  - Data Acquisitions systems, trigger systems; controls; networks and communication protocols
- Part 2:
  - HPC; management of large datasets, distributed computing; big data and data analytics tools
- All with practical activities and tests in lab
  - Coordinated and partially shared with Lab. of Computational Physics





Course Pillars (3)

Advanced Statistics for Physics Analysis

- An overview of the most important statistical methods for large data sets and complex system analysis
- All in the R analysis framework
- Topics :
  - R basic concepts
  - Bernoulli theorem and Central Limit Theorem
  - Inference: general ideas and asymptotic results (large sample size).
  - Fits as special case of parametric inference
  - Monte Carlo methods

Additional courses

- Ad hoc course on *Machine Learning* (48h), providing the theoretical basis
- *Complex systems* as example of theoretical physics relying and exploiting data to elaborate models
- *Quantum information and computing*

- Several physics topics addressed in their “data science” declination:
  - Astrostatistics, life data epidemiology, etc
- Engineering classes (optional):
  - Network science, signal processing, game theory, computer vision, etc

Generalities

- Classes are in English (obviously)
- Admission
  - Bachelor in Physics or sufficient credits in Physics subjects
    - Physics labs and theory up to QM and Special Relativity
  - So far max 40 students (30 EU + 10 non EU), limit will be removed as of next year
- Close collaboration with in masters in Physics, Math, Engineering, Statistics, etc.
- Several high-profile visiting professors from all over the world (CERN, Penn State, Washington, Sacley, etc.)
  - Skilled teaching assistants from companies and research institutes

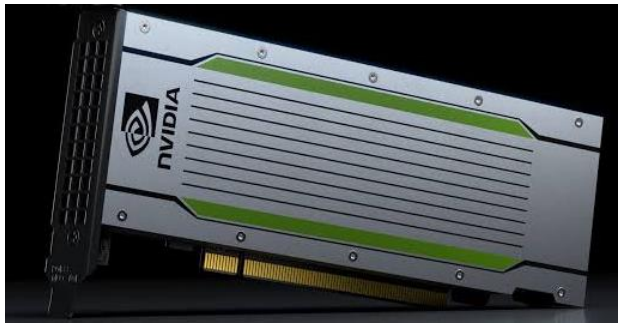
Teaching tools and infrastructure

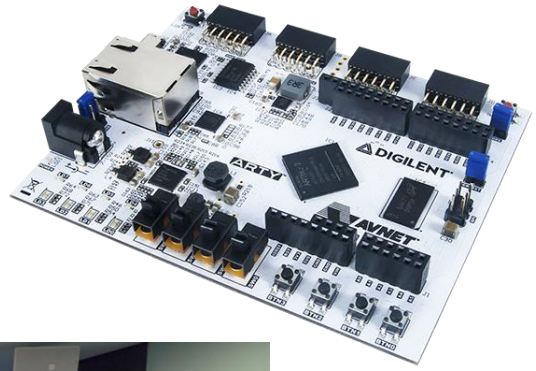
To be effective in such an educational goal, proper teaching tools and computing resources must be available

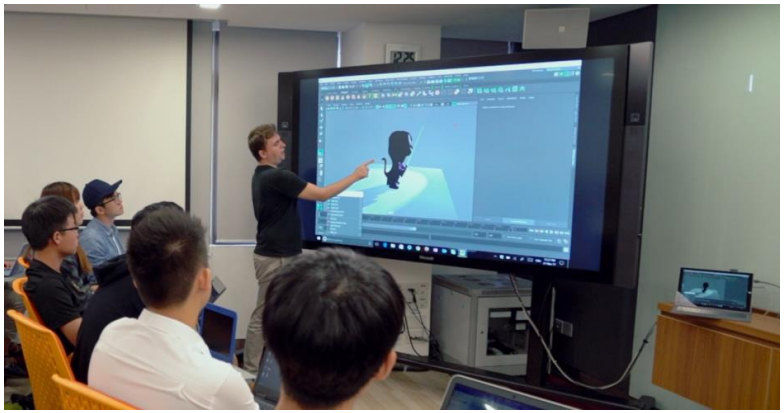
GitHub

jupyter

cloudveneto







Internship at research centres

- Padova Neuroscience Center (Padova): Computational Neurosciance, Imaging, Data Analysis.
  - Corbetta / Suweis
- CNISM (Padova): Physics of Matter, Biological Physics
  - Seno, Maritan
- CNR (Roma) : Network applied to Economy/Finance and Neuroscience
  - Gabrielli / Baiesi, Suweis
- CNR (Padova): Plasma physics, network controllability and modelling
  - Sattin / Baiesi, Suweis
- CNRS (Paris-Saclay): Disordered systems, Computational Physics
  - Rosso/ Baiesi

- FBK (Trento): Complex System, Machine Learning
  - De Domenico / Suweis
- FBK (Trento): Biomedicin, Envirnoment, High Energy Physics, Machine Learning
  - Cristoforetti / Zanetti
- FBK (Trento): IoT, Big Data, Data Science
  - Salvadori / Zanetti
- ARPAV (Rovigo): Environmental physics, statistics, machine learning
  - Menini / Garfagnini
- Azienda Ospedaliera (Padova): Machine Learning, Imaging
  - Cipriani (Cardiology) / Zanetti, Suweis
  - Cipriani (Patological anatomy) / Suweis

- ENS (Paris): Disordered systems, Computational Physics
  - Ros / Baiesi
- IFISC (Palma de Majorca): Complex Systems
  - Meloni / Suweis
- IMT (Lucca): Network applied to Economy/Finance and Social Networks
  - Caldarelli / Suweis
- ICTP (Trieste): Ecological modelling
  - Grilli / Maritan
- HiT (Padova): Computer Vision, Machine Learning
  - Ballan / Suweis
- CiG (Bologna): Metagenomics, Machine Learning
  - Castellani / Suweis

- JPARK (Tokai): Neutrino Physics
  - Takahashi / Collazuol
- INFN (Padova): High energy and Neutrino Physics, Imaging, Machine Learning
  - Zanetti, Collazuol, Garfagnigni
- CERN (Geneva): High energy physics, big data, machine learning, computing infrastructure
  - Campana (Computing infrastructure) / Zanetti
  - Meschi (Data Acquisition systems) / Zanetti
  - Canali (Big Data analytics) / Zanetti
  - Roncarolo (Beam diagnostic) / Zanetti

Internship at companies

GVT S.P.A.

ENGINSOFT

CMD

ENGINEERING

UNOX®

pwc

FONDAZIONE BRUNO KESSLER

accenture

P&G

e-novia
THE ENTERPRISES FACTORY

arpav
Agenzia Regionale per la Prevenzione e Protezione Ambientale del Veneto

OMNYS

BUSITALIA NORD
GRUPPO FERROVIE DELLO STATO ITALIANE

UBS

FATER
Growing for a better life

Students performances/feedback

PROS

- Students satisfaction is very high
- Goal of merge physics and data science competences generally reached
- Students are a resource!
  - Several of their research projects are worth publishing on journals

CONS

- Students get fascinated by technology, a few preferring it over physics
- Mastering powerful tools, some students are reluctant to specialize on a given physics research domain
  - And only a few are leaning towards HEP

Summary

- We believed the gap between what standard physics education provides and what is required by modern research need to be bridged.
- The master degree in Physics of Data is our solution to that
- Excellent results from first students
  - All their projects available on their git accounts
- Check us up on the socials (Facebook and Twitter), we are looking forward to new collaborations!

BACKUP

First Year, First Semester

	Course	Description
mandatory	Laboratory of Computational Physics, Part. 1	Python scientific libraries, extraction of statistical properties from large datasets, Monte Carlo simulations
mandatory	Management and analysis of physics datasets, Part. 1	The flow of data from sensors to the end user. DAQ and trigger systems, controls, network technolgy.
mandatory	Theoretical Physics	Basics of relativistic and non relativistic Quantum Field theory
mandatory	Machine Learning	From general Statistical Learning to Deep Learning
optional	*One course from Master Degree in Physics*	
optional	Nuclear Physics	Courses specializing on a specific field of research in Physics
optional	Theoretical physics of the fundamental interactions	
optional	Solid State Physics	
optional	Statistical mechanics	
optional	The physical universe	
optional	General relativity	

First Year, Second Semester

	Course	Description
mandatory	Statistical Mechanics of Complex Systems	Complex system theory
	Advanced statistics for physics analysis	Advanced frequentist and Bayesian statistics applied to Physics Analyses
	Laboratory of Computational Physics, Part. 2	Large and complex datasets analysis: cases from HEP, Astro-Cosmo and statistical mechanics
	Management and analysis of physics datasets, Part. 2	The flow of data from sensors to the end-user. Parallel processing, distributed computing, big data analytics
optional	*One course among (6 CFU)*	
	Subnuclear Physics	Courses specializing on a specific field of research in Physics
	Structure of Matter	
	Cosmology	
	Relativistic Astrophysics	
	Quantative Life Science	
	Network modelling	Networks from a statistical mechanics perspecitive

Second Year, First Semester

mandatory

optional

Course	Description
Information Theory and Computation	Classical and quantum information theory
Four course among (12 CFU)	
Astro-statistics and cosmology	Bayesian statistics applied to Astrophysics and Cosmology
Quantum Information and Computing	Tech-oriented courses
Computational NeuroScience	
Digital Signal Processing	
Game Theory	
Network Science	