University of Padua

Data Physics

Master Degree in *Physics of Data*

Marco Zanetti,

University of Padua and INFN

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
- The project is approved, first classes starting in October!

The name

- "Physics of Data"
  - From a seminal Nature commentary that elaborates on how connected Physics and Data Science are

commentary

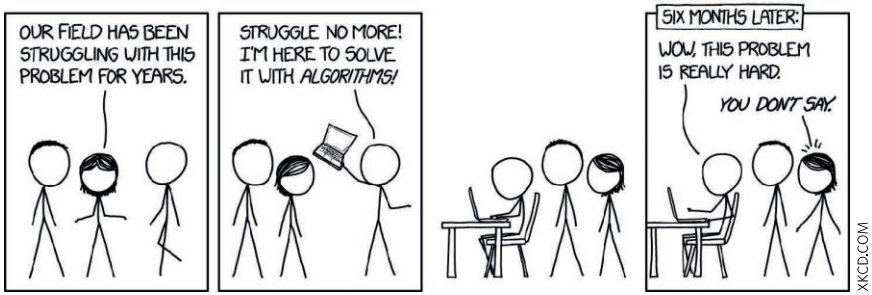
The physics of data

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

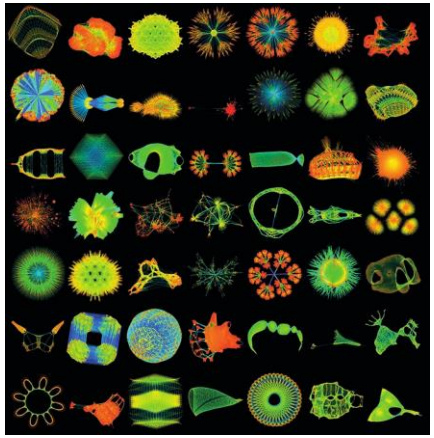
editorial

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

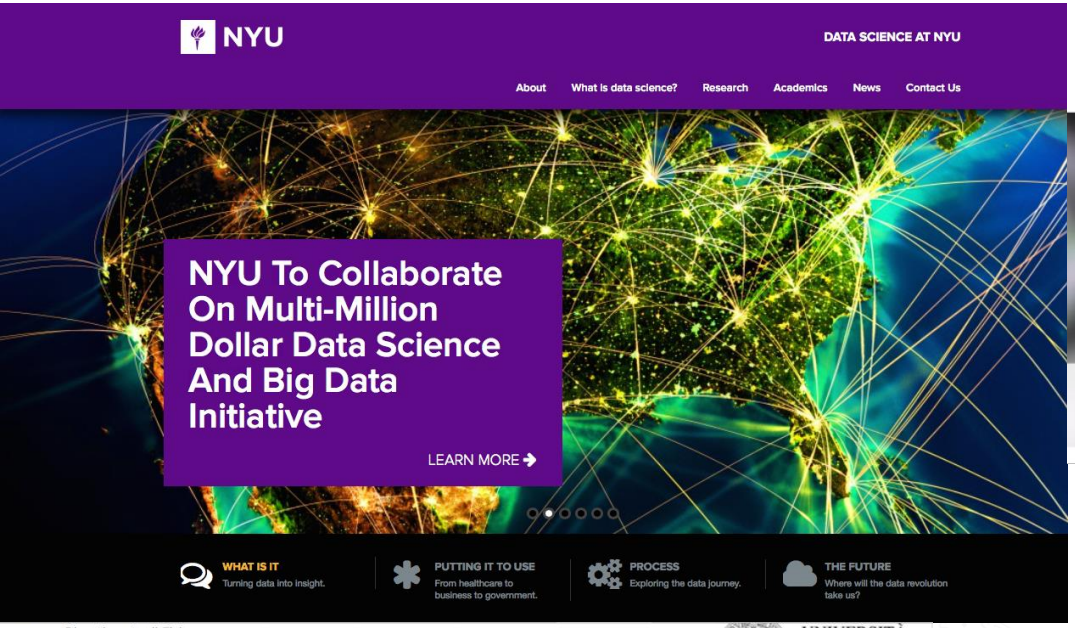
Manifesto

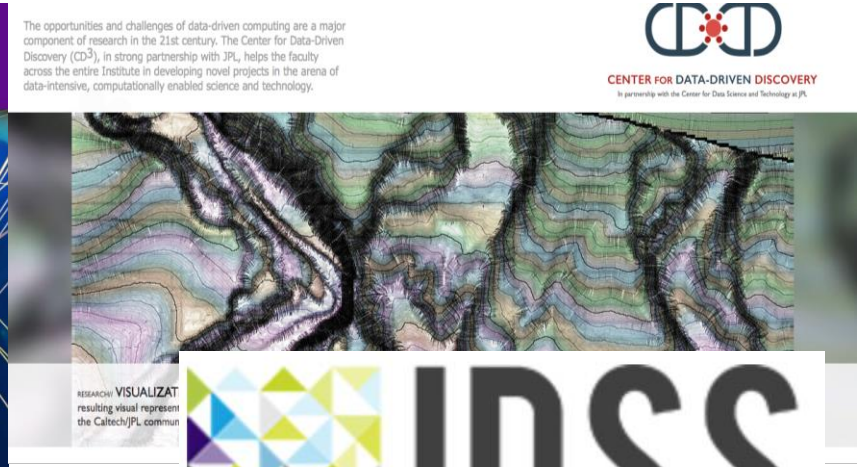
"Indeed, there is plenty that physicists can learn from machine learning, and in order to bridge this gap, Byers advocates for a greater exposure of physics students to statistics and probability, as well as information theory"

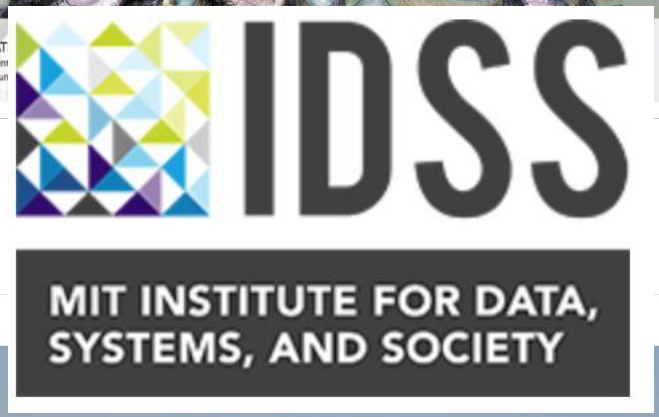
"To some, it may come as a surprise that mathematical techniques originally developed in physics, such as those required to compute the partition function, have been exported to other domains of research and developed further. It is perhaps time physicists learn about these developments and take them 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."

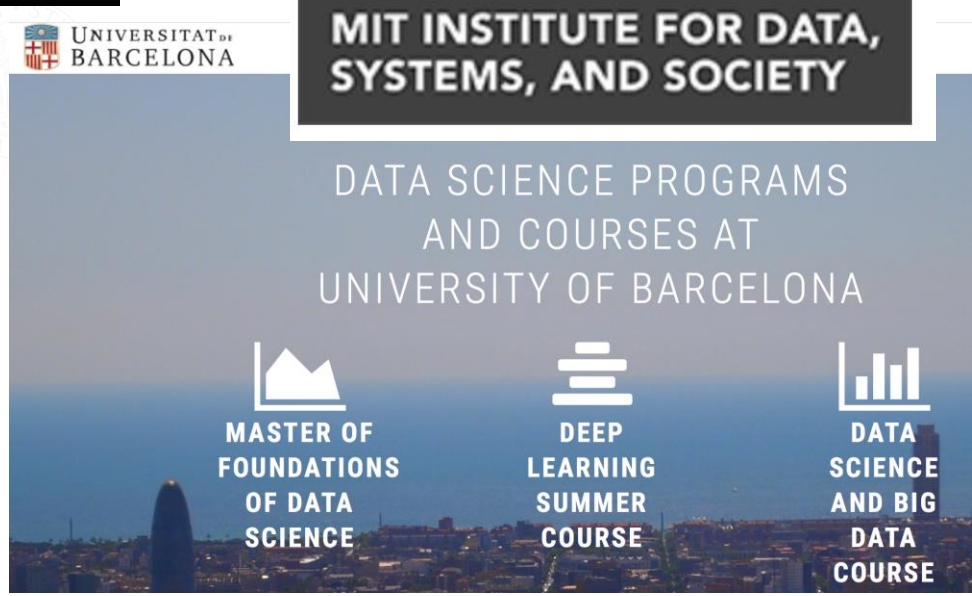
"As complexity features ever more prominently in the realm of physics, we need a new generation of physicists equipped with the tools to rise to the challenges this poses."

Data Science Masters in the world









Boundary conditions

- Masters in Data Science and its hybridization spreading all around the world
  - Any explicitly Physics oriented? We are definitely among the firsts
- Initiative supported by Italian Ministry of Research and Education
  - 5 new Masters in Data Science funded last year
- We investigated the impact on the territory (regional and national) ➔ very positive feedback from companies and research centers!
  - Many partnerships have been established


24 ORE IMPRESA & TERRITORI | L'università diventa digitale: boom di informatica e percorsi...

INDUSTRIA 4.0

L'università diventa digitale: boom di informatica e percorsi sui big data

Generalities

- Classes are in English
- Admission
  - Max 40 students (30 EU + 10 non EU)
  - Bachelor in Physics or sufficient credits in Physics subjects
    - Physics labs and theory up to QM and Special Relativity
- Mandatory internship in a private company or research institute
- Close collaboration with related masters (“*Physics*”, “*Computer Science*” and “*Data Science*”)

Details

- 3 Semesters of classes, each of 48 h (6 CFU)
- 6 mandatory classes of new institution
- Within the master, possibility of specializing in one of the three main research area in Physics
  - Physics of fundamental Interactions, physics of matter, physics of the universe
- Several pure tech classes (from the Computer Science school)
  - E.g. Machine Learning
- Several classes with large fraction of total hours dedicated to lab activities
- Compulsory internship

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	

Lab of Computational Physics

- Laboratory for Computational Physics, targeting the learning and practicing of modern techniques for data analysis and processing
- 2 parts, on two semesters, 98 h total
- Part 1:
  - Python ecosystem of scientific libraries, dataset manipulation, data visualization, data properties and features extraction and statistical assessment; Monte Carlo methods, simulations
- Part 2:
  - Advanced Machine Learning
  - Lab experiences divided per research area
    - E.g. analysis of LHC data, Plank, genomic datasets
- Infrastructure
  - 120 desktops lab
  - INFN cloud and HPC resources

Management and Analysis of Physics Datasets

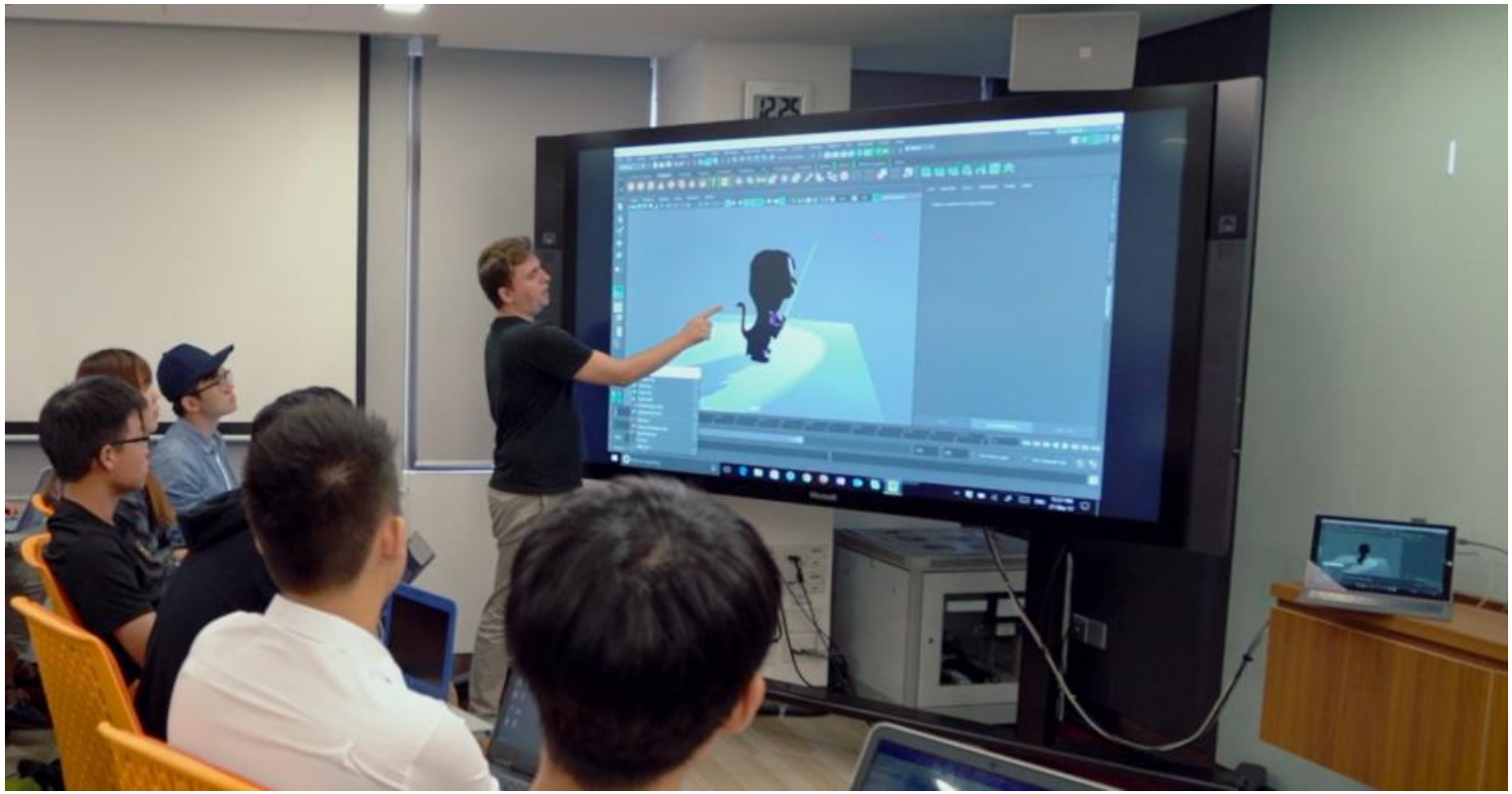
- The data flow from sensors to end users
- 2 parts, on two semesters, 98 h total
- Part 1:
  - Data Acquisitions systems, trigger systems; controls; networks and communication protocols
- Part 2:
  - HPC; management of large datasets, distributed computing; big data and data analytics tools
- All with practical activities and tests in lab

Applied Statistics for Physics Analysis

- Detailed review and application of most relevant statistical tools for complex physics datasets
  - Parameters/point estimation, coverage
  - Bayesian statistics
  - Monte Carlo Markov Chains, Bayesian networks, multivariate analysis
- Applications and examples with modern tools, in particular R and scikit-learn; several lab activities

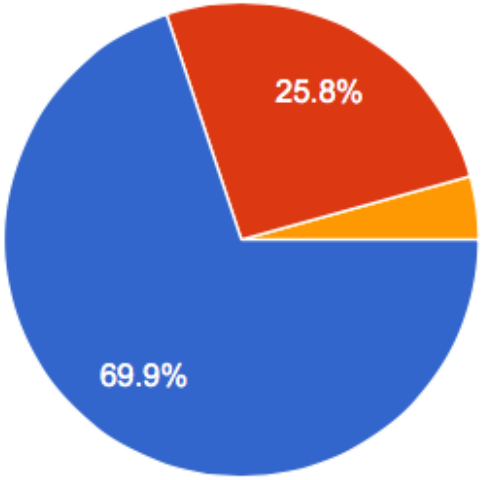
Teaching tools

- Classes broadcasted, recorded and made available online
- Modern teaching tools (e.g. Surface Hub from Windows)



Students' opinion

What's your opinion about the new Physics of Data Master Degree?



- I'm interested, I'll consider it for my academic career
- I'm interested, but NOT for my academic career
- I'm NOT interested

Getting started

- Publicity
  - In the socials (check us out on Facebook and Twitter)
  - Leaflets and posters (check your mailbox!)
- Series of colloquia
  - The Physics of Data, J. Byers (April 19th)
  - Data Science in High Energy Physics, M. Pierini (May 31st)
  - Data Science in Astrophysics and Cosmology, S. Szalay (June 6th)
- Looking forward to collaborate!
  - Already a couple of visiting professors from CERN and Penn State, **more would be welcome!**

University of Padua

Data Physics

01000001011001110110111
00110010101
110011
0110
010
1