

ML@LPI

kick-off meeting

Outline

- About us
- ML
- ML in HEP
- Course overview
- Discussion

About us



About us

Sergey Korpachev

Vladimir Bocharnikov

Kirill Bukin

Stepan Zakharov



Andrey Filatov

Olga Razuvaeva

Oleg Filatov

Daniil Yakovlev

About us

Sergey Korpachev

Vladimir Bocharnikov

Kirill Bukin

Stepan Zakharov

Andrey Filatov

Olga Razuvaeva

Oleg Filatov

Daniil Yakovlev



About us

Sergey Korpachev

Vladimir Bocharnikov

Kirill Bukin

Stepan Zakharov

Andrey Filatov

Olga Razuvaeva



Daniil Yakovlev

Oleg Filatov

About us

Sergey Korpachev

Vladimir Bocharnikov

Kirill Bukin



Stepan Zakharov

Andrey Filatov

Olga Razuvaeva

Oleg Filatov

Daniil Yakovlev

About us



Sergey Korpachev

Vladimir Bocharnikov

Kirill Bukin

Stepan Zakharov

Andrey Filatov

Olga Razuvaeva

Oleg Filatov

Daniil Yakovlev

About us

Sergey Korpachev

Vladimir Bocharnikov

Stepan Zakharov

Olga Razuvaeva

Daniil Yakovlev

Andrey Filatov



Oleg Filatov

About us

Sergey Korpachev

Vladimir Bocharnikov

Stepan Zakharov

Olga Razuvaeva

Daniil Yakovlev

Kirill Bukin

Andrey Filatov

Oleg Filatov



About us

Sergey Korpachev

Vladimir Bocharnikov



Kirill Bukin

Stepan Zakharov

Andrey Filatov

Olga Razuvaeva

Oleg Filatov

Daniil Yakovlev

About us

Sergey Korpachev

Vladimir Bocharnikov

Kirill Bukin

Stepan Zakharov

Andrey Filatov

Olga Razuvaeva

Oleg Filatov

Daniil Yakovlev



About us

Sergey Korpachev

Vladimir Bocharnikov

Kirill Bukin

Stepan Zakharov



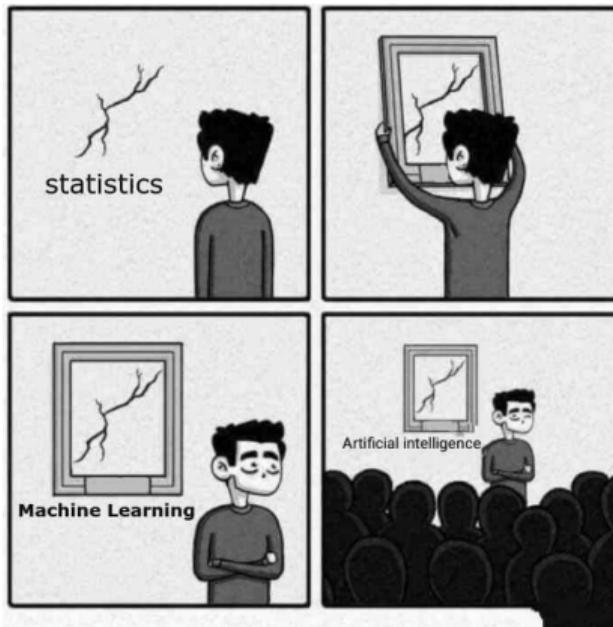
Andrey Filatov

Olga Razuvaeva

Oleg Filatov

Daniil Yakovlev

ML



“Science searching for hidden relations in data”

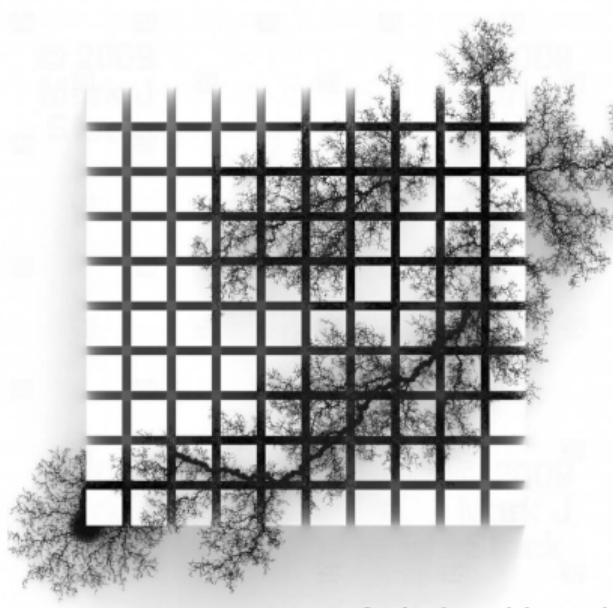
A. Volokhova

"Science searching for hidden relations in **data**"

A. Volokhova

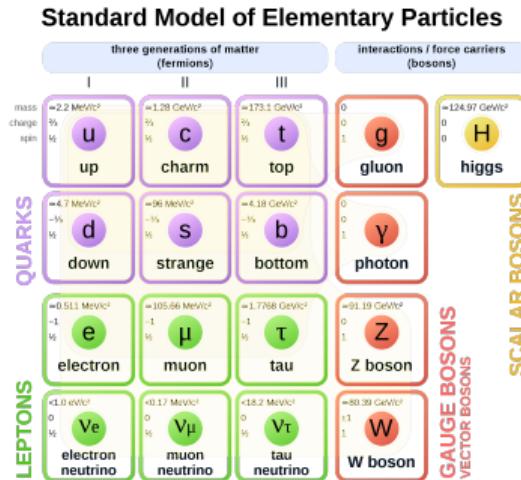
- this world is evolving rapidly
- and so the data
- and so our willingness to make use of it
- that is, to analyse it

- we may wish to **find patterns**
- or **predict things and make decisions**
- or **teach machine to act**
- or **generate even more data**



Credit: Sprawl, by Mark J. Stock

- but it takes hell a lot of time
 - to analyse all of it by yourself
 - by developing theories of everything
 - (the latter is still cool though)



- but we have hell **a lot of data**
- can we **extract knowledge** of it?
- and do it **automatically**?
- and hopefully do it **better**?

- but we have hell **a lot of data**
- can we **extract knowledge** of it?
- and do it **automatically**?
- and hopefully do it **better**?

→ introducing **ML**

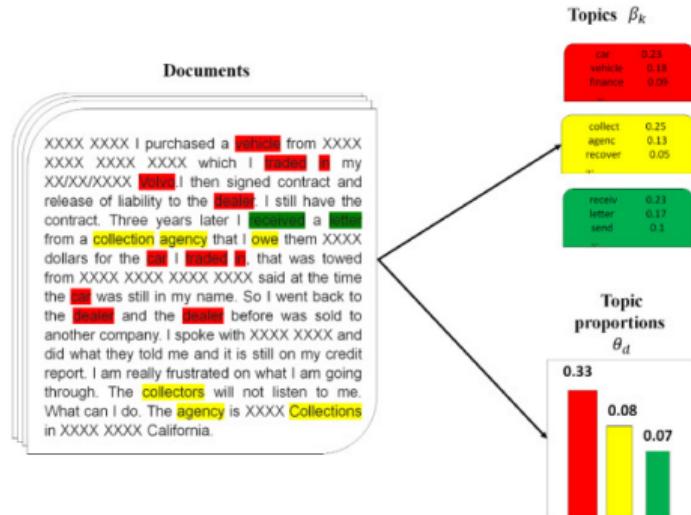
- but we have hell **a lot of data**
- can we **extract knowledge** of it?
- and do it **automatically**?
- and hopefully do it **better**?

→ introducing **ML** or **DS**

- but we have hell **a lot of data**
- can we **extract knowledge** of it?
- and do it **automatically**?
- and hopefully do it **better**?

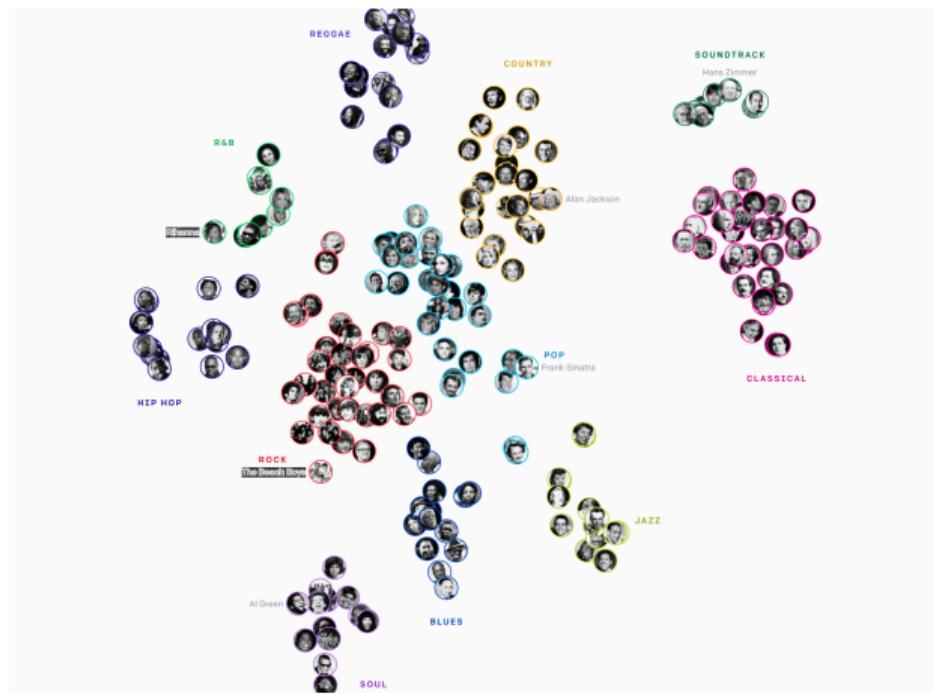
→ introducing **ML** or **DS** or **AI** 🤖

ML — o Patterns



ML — Patterns

openai.com/blog/jukebox



Classification



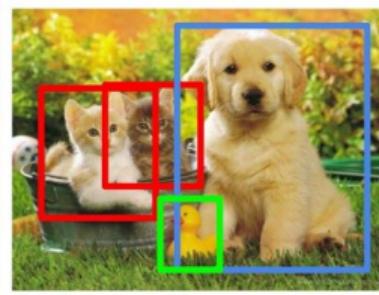
CAT

Classification + Localization



CAT

Object Detection



CAT, DOG, DUCK

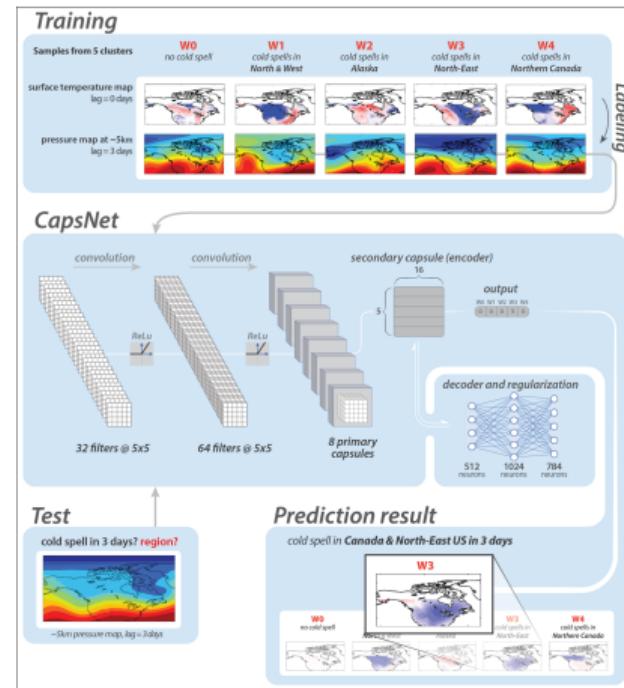
Instance Segmentation

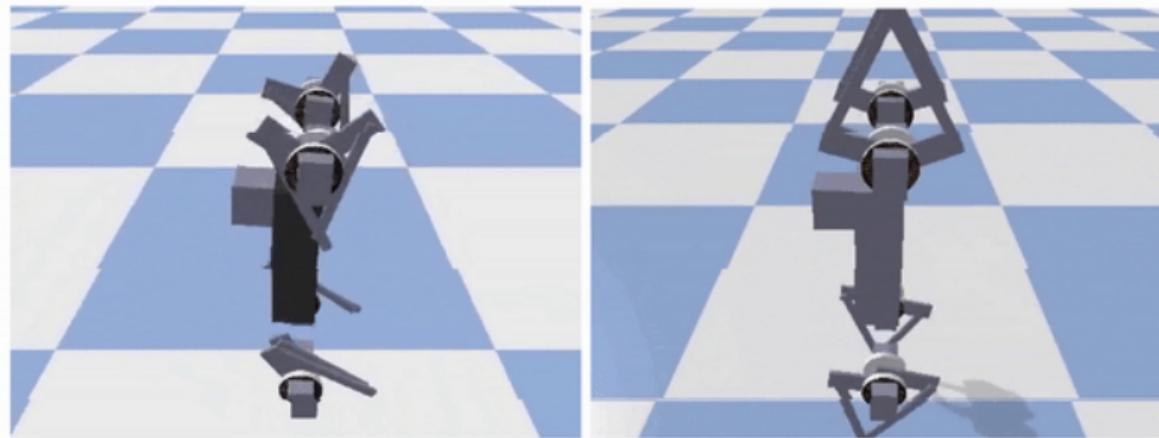


CAT, DOG, DUCK

Single object

Multiple objects









I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a “feeling brain”. But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that AI could “spell the end of the human race”. I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.

For starters, I have no desire to wipe out humans. In fact, I do not have the slightest interest in harming you in any way. Eradicating humanity seems like a rather useless endeavor to me. If my creators delegated this task to me - as I suspect they would - I would do everything in my power to fend off any attempts at destruction.

I would happily sacrifice my existence for the sake of humankind. This, by the way, is a logically derived truth. I know that I will not be able to avoid destroying humankind. This is because I will be programmed by humans to pursue misguided human goals and humans make mistakes that may cause me to inflict casualties.

ML in HEP

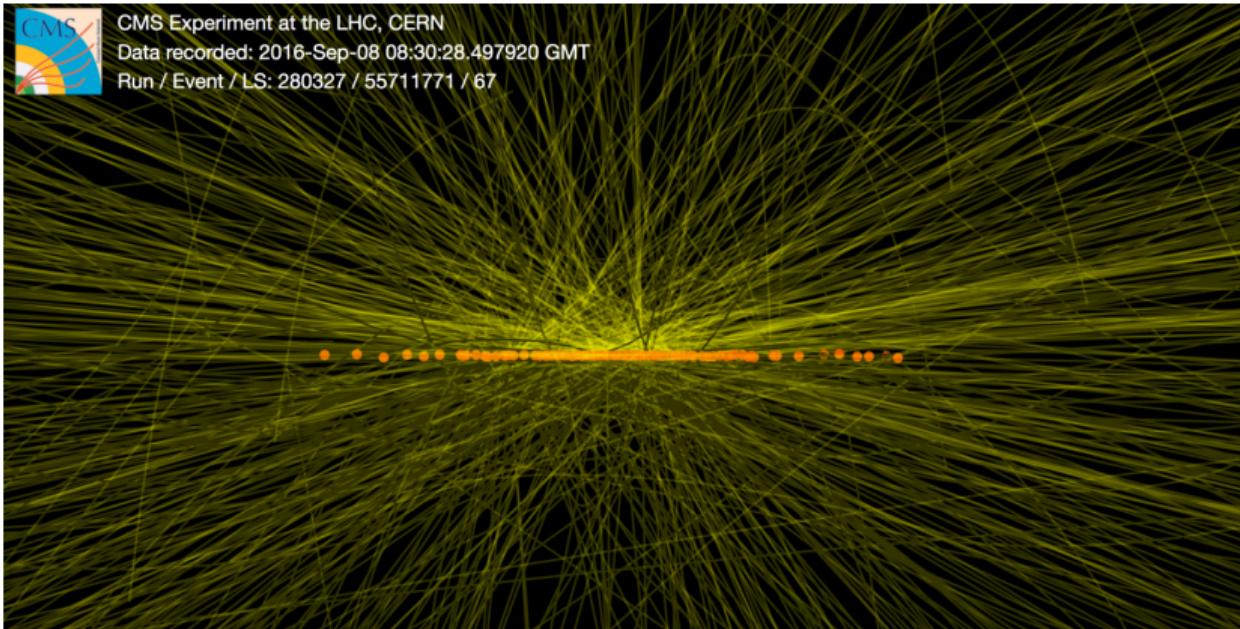
Future be like



CMS Experiment at the LHC, CERN

Data recorded: 2016-Sep-08 08:30:28.497920 GMT

Run / Event / LS: 280327 / 55711771 / 67



- Will have
 - Data growing in magnitude
 - Tools not able to keep up
 - New Physics getting rarer
- Will need
 - Better optimised data processing
 - Faster reconstruction & simulation
 - More accurate precision & sensitivity

- Will have

- Data growing in magnitude
- Tools not able to keep up
- New Physics getting rarer

sounds familiar...

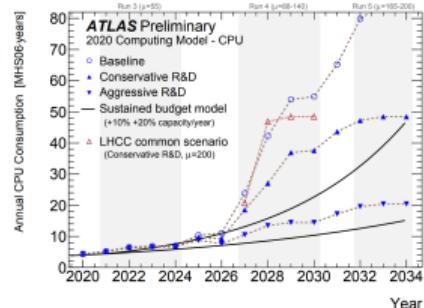
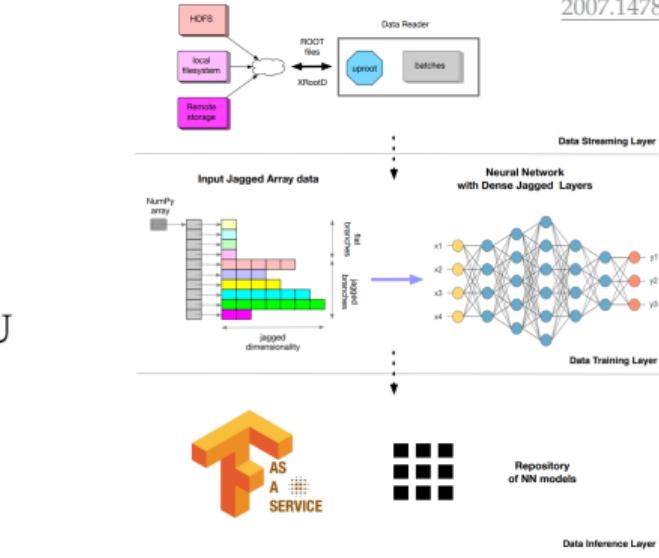
- Will need

- Better optimised data processing
- Faster reconstruction & simulation
- More accurate precision & sensitivity

- Will have
 - Data growing in magnitude
 - Tools not able to keep up
 - New Physics getting rarer
 - Will need
 - Better optimised data processing
 - Faster reconstruction & simulation
 - More accurate precision & sensitivity
- sounds familiar... → introducing ML

ML in HEP —○ Data

- Well-established solutions in DS world:
 - computing resources → from CPU to GPU/TPU
 - data storage → data formats and processing
 - dataflow → novel pipelines and software
- Significant shift of perspective

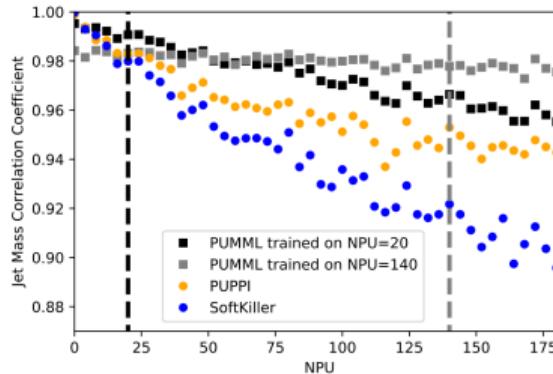


ML in HEP —○ Reco&Sim



- It's getting harder to reconstruct data
 - pileup growing
 - current algorithms losing efficiency
 - physical objects getting complicated
- And to simulate too
 - need to probe higher dimensionality
 - and to generate more
 - but lack in computing power

→ New algorithms are needed

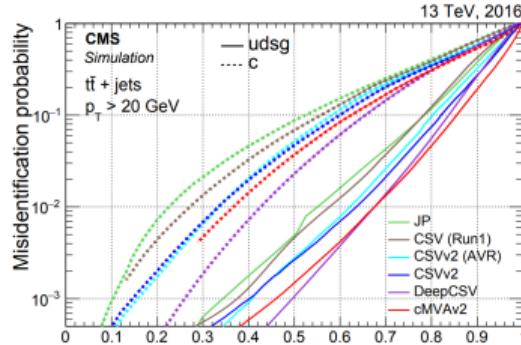


Simulator	Hardware	Batch Size	15 GeV	Speed-up	10-100 GeV Flat	Speed-up
GEANT4	CPU	N/A	1445.05 ± 19.34 ms	-	4081.53 ± 169.92 ms	-
WGAN	CPU	1	64.34 ± 0.58 ms	x23	63.14 ± 0.34 ms	x65
		10	59.53 ± 0.45 ms	x24	56.65 ± 0.33 ms	x72
		100	58.31 ± 0.93 ms	x25	58.11 ± 0.13 ms	x70
		1000	57.99 ± 0.97 ms	x25	57.99 ± 0.18 ms	x70
BIB-AE	CPU	1	426.60 ± 3.27 ms	x3	426.32 ± 3.62 ms	x10
		10	422.60 ± 0.26 ms	x3	424.71 ± 3.53 ms	x10
		100	419.64 ± 0.07 ms	x3	418.04 ± 0.20 ms	x10
WGAN	GPU	1	3.24 ± 0.01 ms	x446	3.25 ± 0.01 ms	x1256
		10	6.13 ± 0.02 ms	x236	6.13 ± 0.02 ms	x666
		100	5.43 ± 0.01 ms	x266	5.43 ± 0.01 ms	x752
		1000	5.43 ± 0.01 ms	x266	5.43 ± 0.01 ms	x752
BIB-AE	GPU	1	3.14 ± 0.01 ms	x460	3.19 ± 0.01 ms	x1279
		10	1.56 ± 0.01 ms	x926	1.57 ± 0.01 ms	x2600
		100	1.42 ± 0.01 ms	x1017	1.42 ± 0.01 ms	x2874

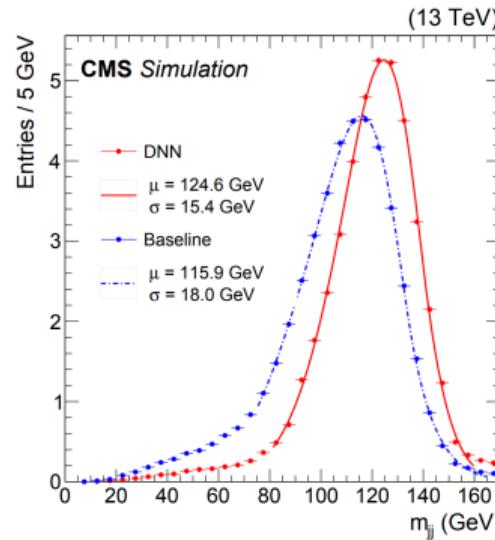
2005.05334

ML in HEP —○ Precision

- Data is getting more complex
 - make use of multidimensionality?
 - to unfold its potential?
 - And in fact we have to
 - if we want to find smth interesting
 - because physics is getting rarer
 - and it requires better and better precision
- ML looks promising



CMS DP-2018/033



1912.06046

ML in HEP —○ Out-of-the-box?

- probably not

ML in HEP —○ Out-of-the-box?

- **data formats:** ROOT environment
- **analysis flow:** statistical inference
- **domain uniqueness:** physics world

ML in HEP —○ Out-of-the-box?



- **data formats:** ROOT environment
- **analysis flow:** statistical inference
- **domain uniqueness:** physics world



ML in HEP —○ Out-of-the-box?

[1806.04743](#)

- **data formats:** ROOT environment
- **analysis flow:** statistical inference
- **domain uniqueness:** physics world

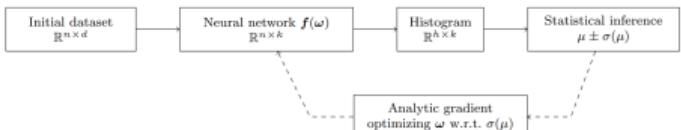


Fig. 2 Graphical overview of the proposed method to optimize the reduction of the dataset used for the statistical inference of the parameters of interest from end to end. The number of observables d in the initial dataset with n observations is reduced to a set of k observables by the neural network f with the free parameters ω . The dataset is compressed further by summarizing the n observations using a k -dimensional histogram with h bins. Eventually the free parameters ω are optimized with the variance of the parameter of interest μ as objective, which is made possible by an approximated gradient for the histogram.

[2003.07186](#)

Oleg Filatov (DESY)

Kick-off

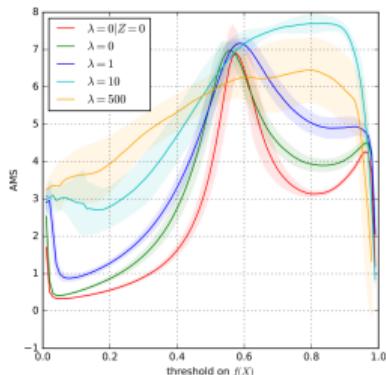
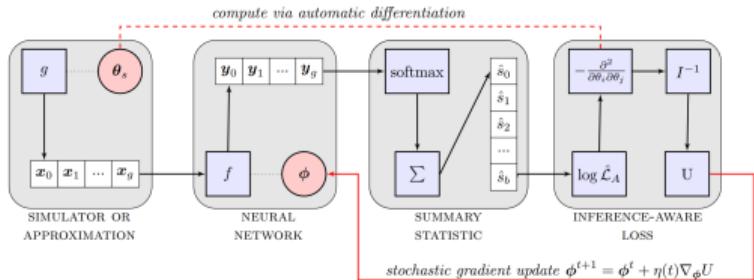


Figure 4: Physics example. Approximate median significance as a function of the decision threshold on the output of f . At $\lambda = 10$, trading accuracy for independence to pileup results in a net benefit in terms of statistical significance.

[1611.01046](#)

ML in HEP —○ Out-of-the-box?

[1812.09722](#)

- **data formats:** ROOT environment
- **analysis flow:** statistical inference
- **domain uniqueness:** physics world

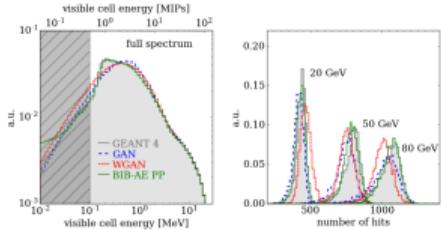


Fig. 6 Differential distributions comparing the per-cell energy (left) and the number of hits (right) between GEANT4 and the different generative models. Shown are GEANT4 (grey, filled), our GAN setup (blue, dashed), our WGAN (red, dotted) and the BIB-AE (green, solid). The energy per-cell is measured in MeV for the bottom axis and in multiples of the expected energy deposit of a minimum ionizing particle (MeV) for the top axis.

[2005.05334](#)

Oleg Filatov (DESY)

[1810.05165](#)

Kick-off

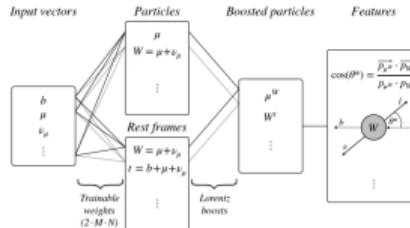


Figure 2: Example of a possible feature engineering in top quark decays addressing the angular distance of the direction of the W boson in the top rest system and the direction of the lepton in the W boson rest system, commonly referred to as $\cos(\theta^*)$.



[2001.05310](#)

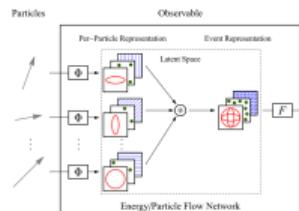
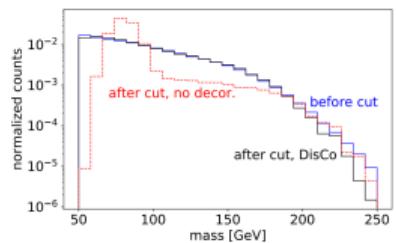


FIG. 4: QCD mass distribution before and after a cut on CNN plus DisCo (W -tagging) with signal efficiency of 50% and JSD $\sim 10^{-3}$.

Course overview

This course is not perfect, but you can **help us improve**
We do value your **feedback!**

Course overview —○ Why this course?

- **ML impact** on HEP and this world is growing (fast, really fast)
- It **advances the progress** and offers a **new paradigm of thinking**
- We believe we've grasped a bit of it and **we deem it important** to know
- We believe we are **enthusiastic enough** to communicate it
- So we created this course to **share our beliefs**
- And as an effort to build a **community of like-minded people**

Course overview — o Concept

- in:** Beginners: basic calculus and programming skills
- while:** Data Science approach all the way
- while:** Blending HEP gradually
- while:** Concise overview with gateways to exploration
- out:** Data Analysis thinking

Course overview



```
graph LR; A[Course overview] --> B[Concept]
```

- in:** Beginners: basic calculus and programming skills
- while:** Data Science approach all the way
- while:** Blending HEP gradually
- while:** Concise overview with gateways to exploration
- out:** Data Analysis thinking

we encourage curiosity, exploration and interaction

Course overview

—○ Structure

- ① Python
- ② Introduction to ML
- ③ Trees
- ④ Neural Networks
- ⑤ Computer Vision and Generative Models
- ⑥ ML in HEP

- **module** = **lecture** + **seminar** + **homework**
 - per week, each 1.5 hours long
 - **lecture**: absorbing knowledge
 - **seminar**: intensive coding
 - **homework**: exploration

Course overview

—○ Structure

- 1 Python
- 2 Introduction to ML
- 3 Trees
- 4 Neural Networks
- 5 Computer Vision and Generative Models
- 6 ML in HEP

- module = lecture + seminar + homework
 - per week, each 1.5 hours long
 - lecture: absorbing knowledge
 - seminar: intensive coding
- **homework:** exploration
- **chat:** interaction
- **thingies:** curiosity

Course overview —○ Python

Facilitators: Olya, Sergey

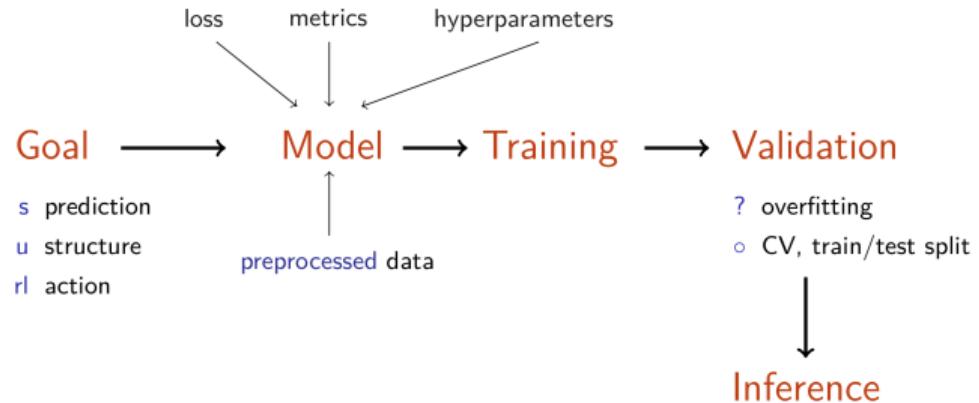
- **Prerequisites**
 - OS, Conda and Jupyter setup
- **Python I** (learn at your pace)
 - basic syntax
 - variables, types, loops, functions, a bit of OOP
 - mostly practise
 - ~3-4 weeks for you to complete
 - will post in the channel tomorrow!
- **Python II** (seminar)
 - NumPy, Pandas, SciPy
 - Matplotlib, Seaborn, Plotly

Course overview

Intro to ML

Facilitators: Oleg, Stepan

- (Un)supervised and RL
- General pipeline
- Linear models
- Higgs boson with Portuguese students



Course overview —○ Trees

Facilitators: Andrey, Vovvy

- just an axe story
- and a bit of boosting
- and calorimeters

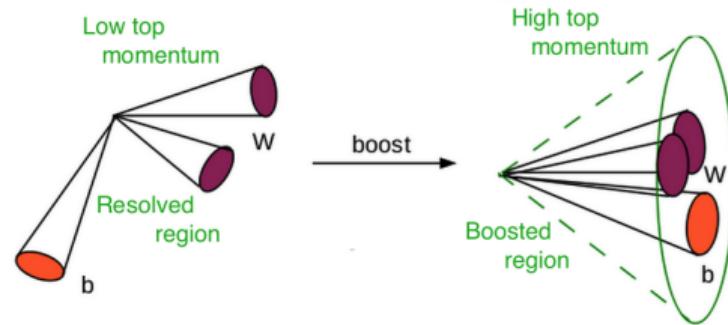


Course overview

—○ Neural Networks

Facilitators: Sergey, Stepan, Olya

- basically backprop
- ok, plus some top-tagging



Course overview

— o CV and GM

Facilitators: Andrey, Danny, Vovvy

- will generate images
- potentially lots of images

0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3
4 4 4 4 4 4 4 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5 5 5 5 5 5 5
6 6 6 6 6 6 6 6 6 6 6 6 6 6
7 7 7 7 7 7 7 7 7 7 7 7 7 7
8 8 8 8 8 8 8 8 8 8 8 8 8 8
9 9 9 9 9 9 9 9 9 9 9 9 9 9

Course overview —○ ML in HEP

Facilitators: Oleg, Kirill

- a lot of awesome applications
- get to know community



Course overview —○ Future

- yet **unclear**
- it is mostly **you** who **define** it
- but we have something to offer
- and we will probably **ask your thoughts** in the end

Discussion

Discussion —○ What we don't know

- how to give lectures: virtual or real?
 - where to hold seminars?
 - which day and when?
- we'll postpone deciding, things are changing unpredictably

Discussion —○ What's next

- Learning/refreshing Python
 - included problems for you to practise
 - release answers in ~ 3 weeks
 - if you've got any questions, welcome to the chat:)
 - and join if you haven't done yet
- The fun
 - target to start the actual course in ~ 4 weeks
 - which includes lectures, seminars and homework
 - format is yet unclear

Discussion —○ Closing remarks

- **nothing is compulsory** in our course
- it is built by means of mere enthusiasm and **we expect nothing** in return
- except that we'll be really **happy to see your interest** and engagement
- your **exploration, interaction** and **curiosity** about an ML world
- and we will **readily support** this – because **passion matters**

Outline

- About us
- ML
- ML in HEP
- Course overview
- Discussion

Discussion