CMSDAS 2024

Machine Learning Short Exercise

CERN, 17-24 June 2024

Evan Armstrong Koenig, Matthias Komm, Pietro Vischia

Supported by project
RYC2021- 033305-I
funded by

The facilitators

- Evan Armstrong Koenig (University of Florida)
  - Graduate student since 2020, CMS member since 2017
  - ML contact for the B2G PAG (since 2023)
  - Developing graph neural networks for event reconstruction in multi-higgs to jet final states

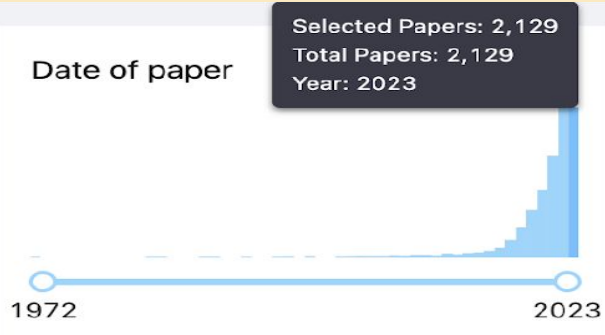
- Matthias Komm (DESY, Germany)
  - CMS member since 2010; PhD in 2017 on single top quark cross section measurements
  - CMS ML innovation convener (2019-2021)
  - Research on exotic long-lived particle decaying to displaced jets (& leptons)
    = playground for ML due to lack of any standard reconstruction

- Pietro Vischia ("Ramón y Cajal" Senior Researcher at Universidad de Oviedo and ICTEA)
  - CMS member since 2009, PhD in 2016 (IST Lisboa)
  - MLG L2 convener (2024-2026), CERM IML coordinator for CMS (2020-2024)
  - ML in analysis (charged Higgs, ttH multilepton), + anomaly detection, design of experiments, neuroscience
  - Various projects/groups: PI of NeuroMODE (neuromorphic computing for design of experiments and trigger applications), steering board of MODE (Machine-learning Optimized Design of Experiments), steering group of EUCAIF (European Coalition for AI in Fundamental physics). Past: partner node PI of AMVA4NewPhysics (Advanced MultiVAriate Analyses for New Physics)
  - Regular courses and lectures on ML (book on Stat and ML near completion, or at least this is what I told the editor :D)

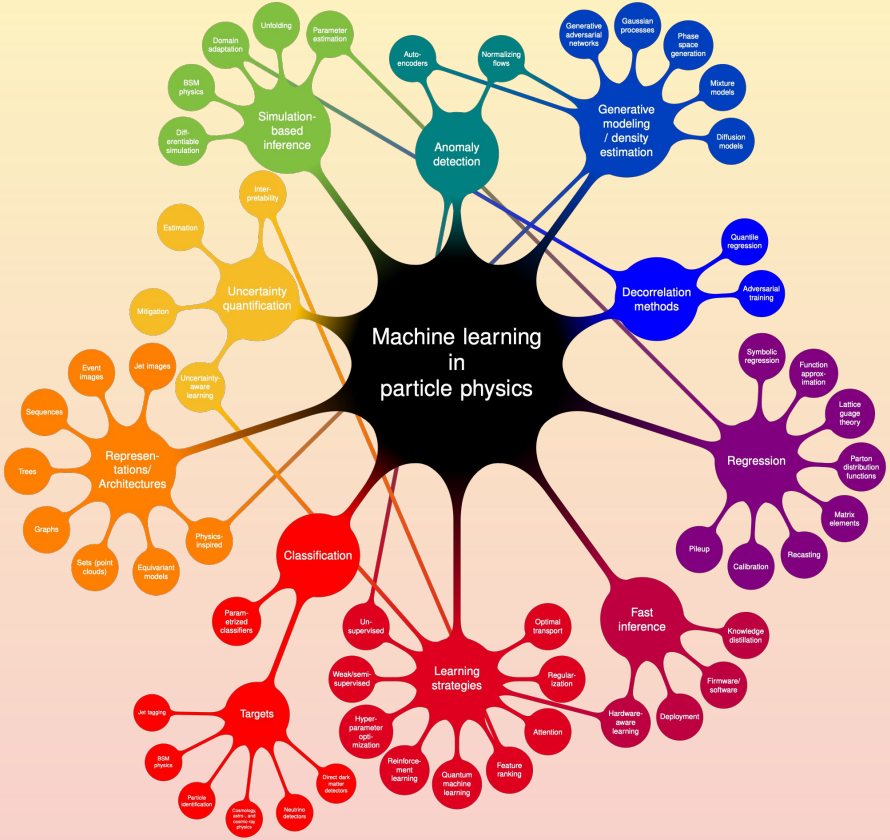
Machine Learning

The use of data to make predictions or decisions without explicit programming

Use of ML in particle physics and CMS, in particular, has grown exponentially



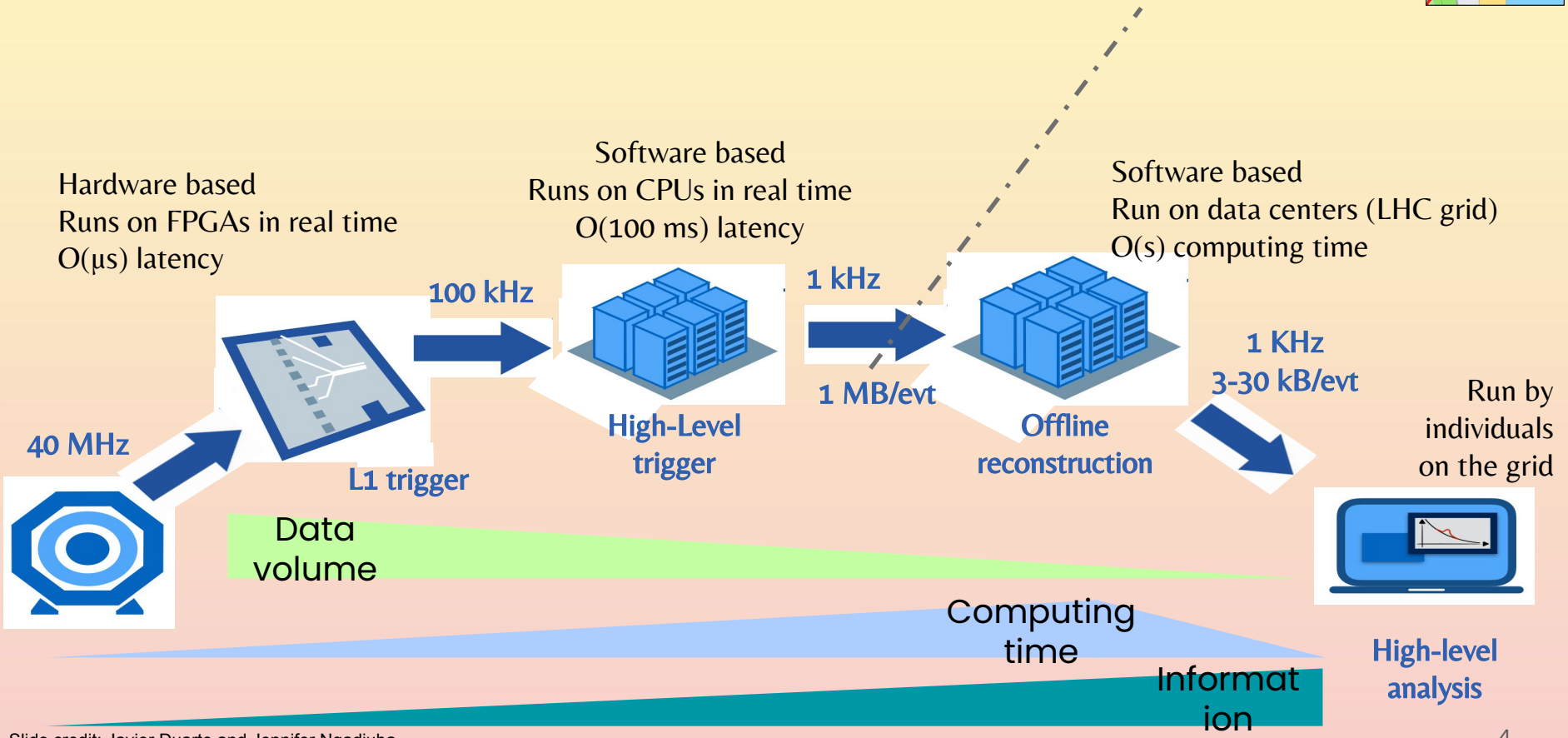
inspirehep.net



Slide credit: Javier Duarte and Jennifer Ngadiuba

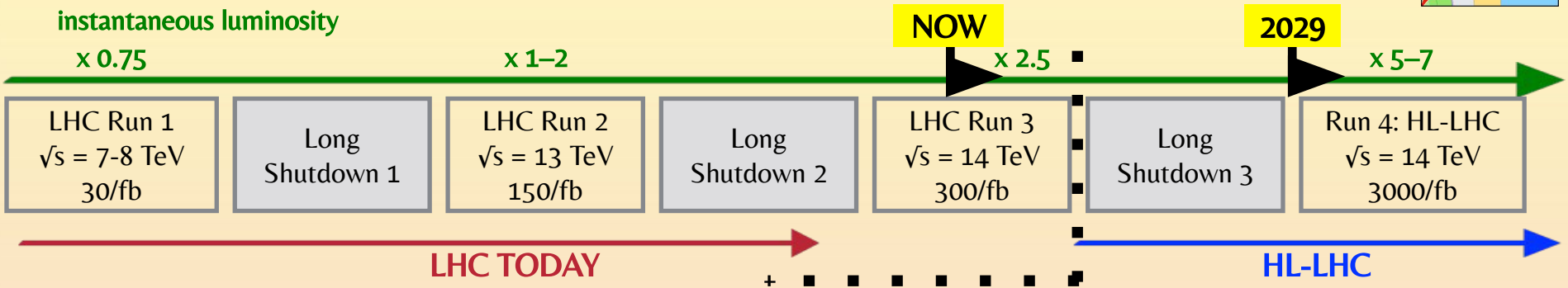
iml-wg.github.io/HEPML-LivingReview

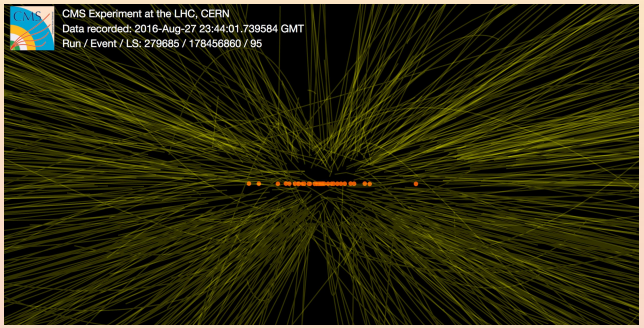
Data reduction workflow @ LHC



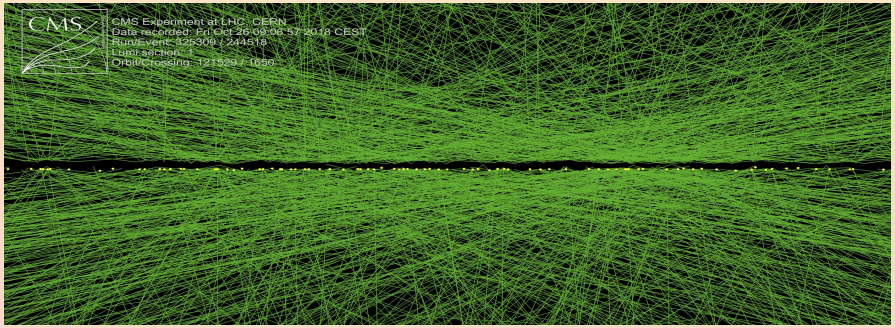
Slide credit: Javier Duarte and Jennifer Ngadiuba

The High-Luminosity LHC challenge





40 simultaneous collisions
per bunch crossing

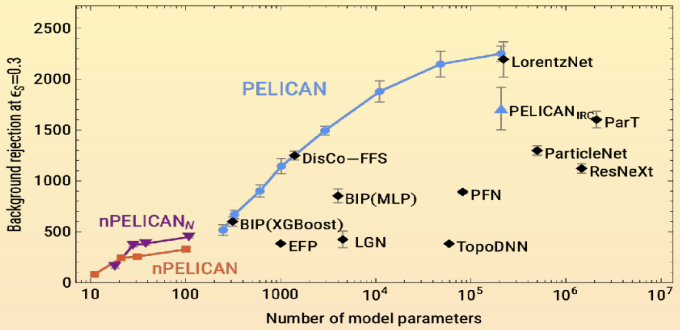


200 simultaneous collisions
per bunch crossing
+
more granular detector!

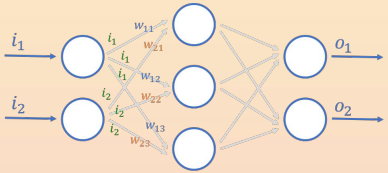
Slide credit: Javier Duarte and Jennifer Ngadiuba

ML Advantages

- **High accuracy**
  - improved physics performance:
    extract most useful info from complex datasets
  - better scaling with data complexity (HL-LHC)
  - enables novel strategies (e.g., anomaly detection)

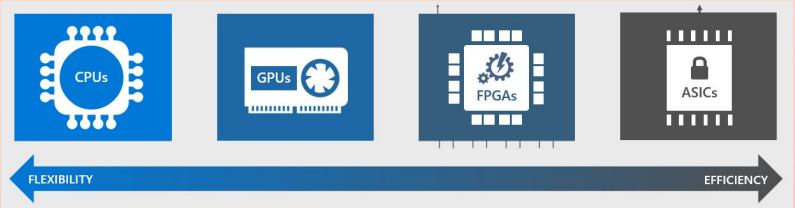


- **Fast speed**
  - reduced development time
  - matrix multiplication can be massively parallelized
  - take advantage of reduced precision



$$\begin{bmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \\ w_{13} & w_{23} \end{bmatrix} \cdot \begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = \begin{bmatrix} (w_{11} \times i_1) + (w_{21} \times i_2) \\ (w_{12} \times i_1) + (w_{22} \times i_2) \\ (w_{13} \times i_1) + (w_{23} \times i_2) \end{bmatrix}$$

- **Better portability**
  - many dedicated processors:
    GPUs, FPGAs, TPU, IPU, ...
  - large investment in tools to
    compile and optimize
    ML models for the hardware



Slide credit: Javier Duarte and Jennifer Ngadiuba

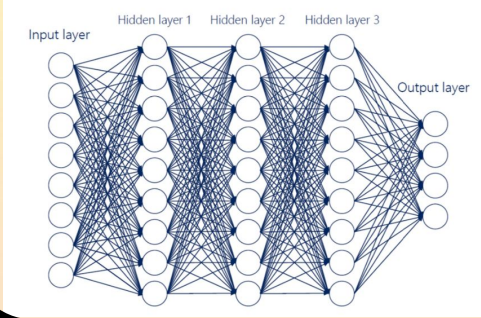
Where ML is (or will be) used in CMS

Operations	⊙ Monitoring and anomaly detection for logs ⊙ Automated decision making for grid jobs and data management
Front-end electronics	⊙ Fast ML on ASICs for data compression in Phase 2 HGCAL
Trigger	⊙ Fast ML on FPGAs for Run 3 & Phase 2 L1 trigger and 40 MHz scouting
DQM	⊙ Automated data certification ⊙ Online anomaly detection (ECAL, HCAL, muon system)
Simulation	⊙ Calorimeter/jet simulation with generative adversarial networks, variational autoencoders, normalizing flows, diffusion models
Calibrations	⊙ Jet energy corrections and scale factors
Reconstruction	⊙ Energy and mass regression (e.g., MET, photons, electrons, jets) ⊙ PU mitigation ⊙ Clustering (e.g., calorimeter, jets, vertexing) ⊙ Particle flow
Analysis / object ID	⊙ Tau leptons, heavy flavour / boosted / displace jets tagging ⊙ Event classification ⊙ Background estimation ⊙ Uncertainties evaluation

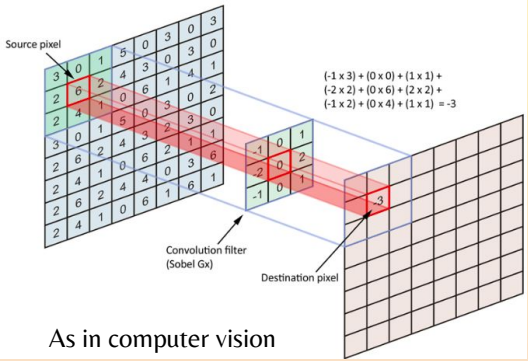
Slide credit: Javier Duarte and Jennifer Ngadiuba

ML is Data Representation

High level/tabular: fully connected NN, BDTs

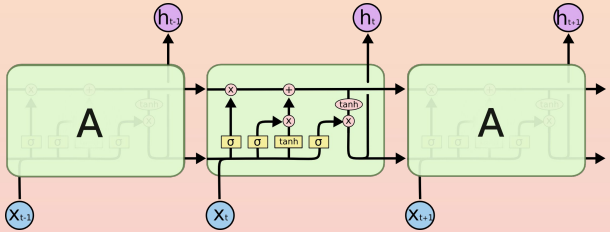


Regular grid: convolutional NN



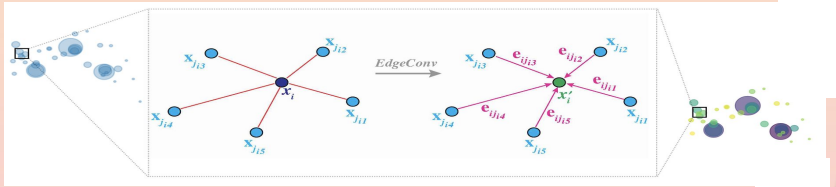
As in computer vision

Ordered sequence/time series: recurrent NN, transformers



As in natural language processing

Point cloud: Deep sets, graph NN, transformers



As in social media analysis

Slide credit: Javier Duarte and Jennifer Ngadiuba

Types of learning

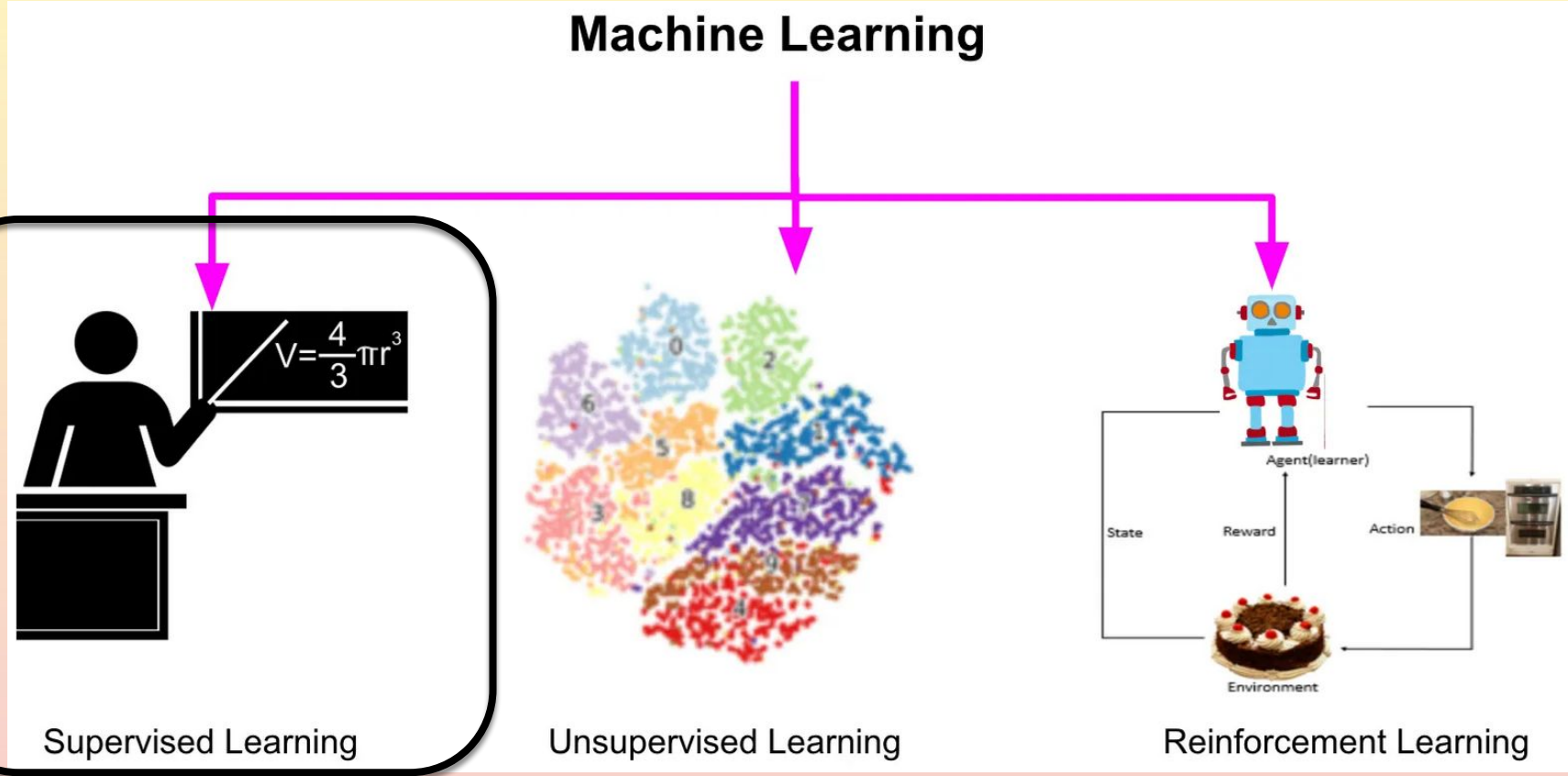


Image by Renu Khandelwal

ML Frameworks

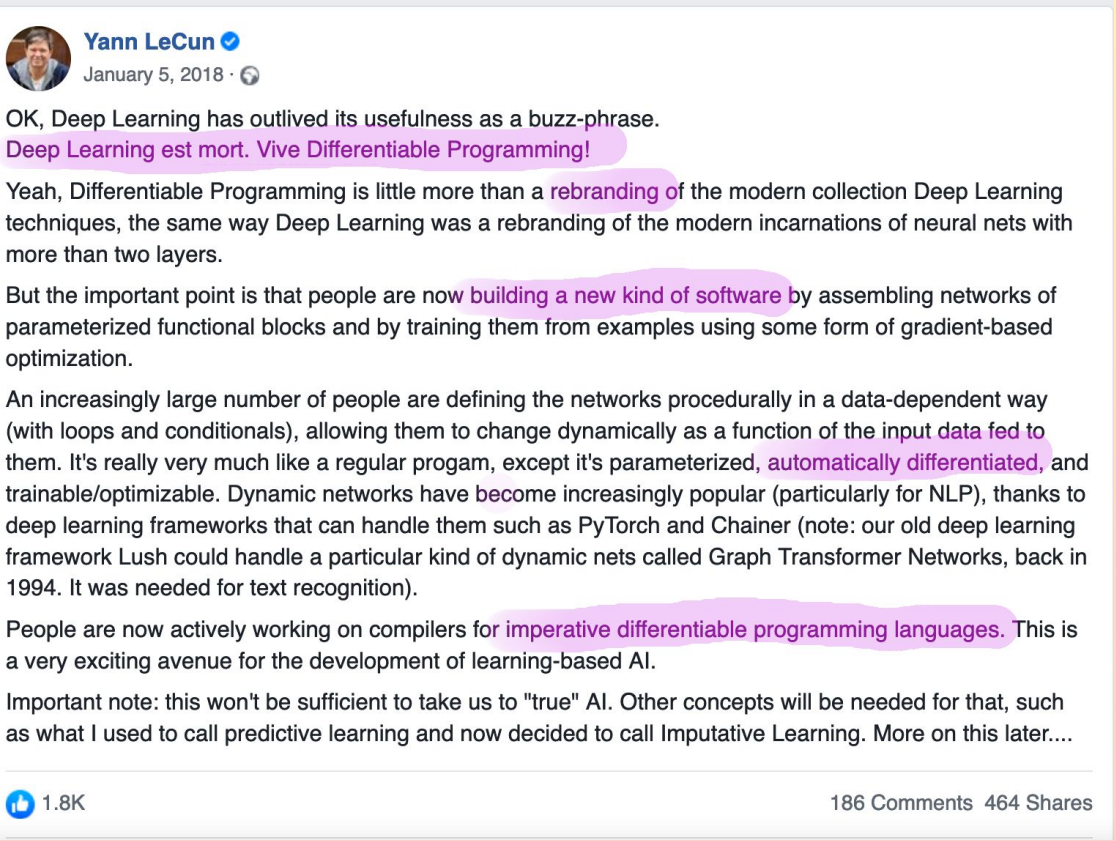
- **Keras/TensorFlow**
  - Conventionally one of the most beginner-friendly formats
  - Inference supported for production in CMSSW
- **PyTorch (Lightning)**
  - More “pythonic” way of building models (especially models within models)
  - Inference support in CMSSW forthcoming...
- **ONNX/ONNXRuntime**
  - ONNX intended to be a universal exchange format (can convert from all libraries)
  - For inference only; supported for CMSSW
- **Flax/JAX**
  - Newest library based on “autograd”
- **XGBoost**
  - For training boosted decision trees (fast)

Slide credit: Javier Duarte and Jennifer Ngadiuba

Automatic Differentiation

Execute **differentiable functions (programs)** via ***automatic differentiation***

Wanna know more? Come at
https://indico.cern.ch/event/1380163/



Yann LeCun
January 5, 2018 ·

OK, Deep Learning has outlived its usefulness as a buzz-phrase.
Deep Learning est mort. Vive Differentiable Programming!

Yeah, Differentiable Programming is little more than a rebranding of the modern collection Deep Learning techniques, the same way Deep Learning was a rebranding of the modern incarnations of neural nets with more than two layers.

But the important point is that people are now building a new kind of software by assembling networks of parameterized functional blocks and by training them from examples using some form of gradient-based optimization.

An increasingly large number of people are defining the networks procedurally in a data-dependent way (with loops and conditionals), allowing them to change dynamically as a function of the input data fed to them. It's really very much like a regular progam, except it's parameterized, automatically differentiated, and trainable/optimizable. Dynamic networks have become increasingly popular (particularly for NLP), thanks to deep learning frameworks that can handle them such as PyTorch and Chainer (note: our old deep learning framework Lush could handle a particular kind of dynamic nets called Graph Transformer Networks, back in 1994. It was needed for text recognition).

People are now actively working on compilers for imperative differentiable programming languages. This is a very exciting avenue for the development of learning-based AI.

Important note: this won't be sufficient to take us to "true" AI. Other concepts will be needed for that, such as what I used to call predictive learning and now decided to call Imputative Learning. More on this later....

1.8K 186 Comments 464 Shares

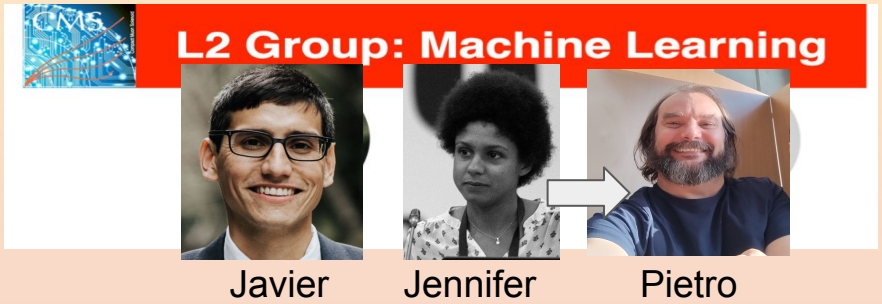
Screenshot of Yann LeCun's facebook post

ML Group

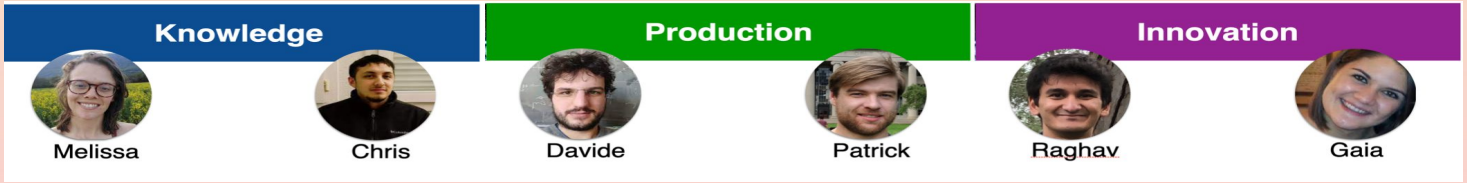
- State-of-the art machine learning methods are increasingly being deployed in CMS
- Many opportunities to integrate/develop proof of concepts methods into the experiment with many challenges ahead to bring the most promising into production
- CMS ML group was created to supervise this effort: cms-phys-conveners-MLG@cern.ch

Our goal: enable, support, guide and foster ML developments in computing, POGs, PAGs

Forum (bi-weekly):
https://indico.cern.ch/category/12412/
L3 subgroup meeting (weekly, rotating):
https://indico.cern.ch/category/12413/



L3 topical groups



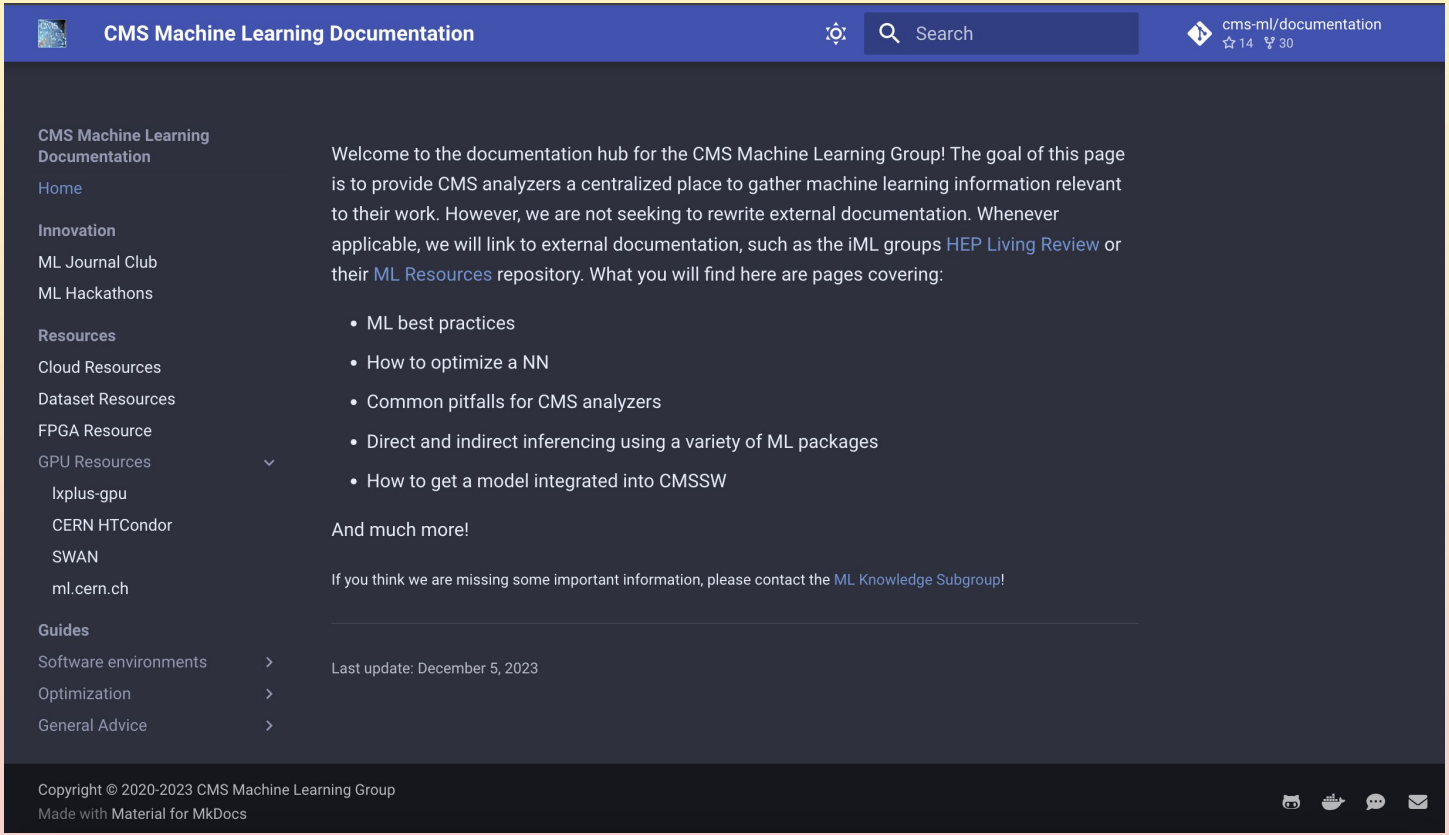
+ contacts to POG/PAGs and external initiatives

Twiki for more info

Slide credit: Javier Duarte and Jennifer Ngadiuba

CMS ML Documentation

Managed by the MLG “Knowledge” team



cms-ml.github.io/documentation

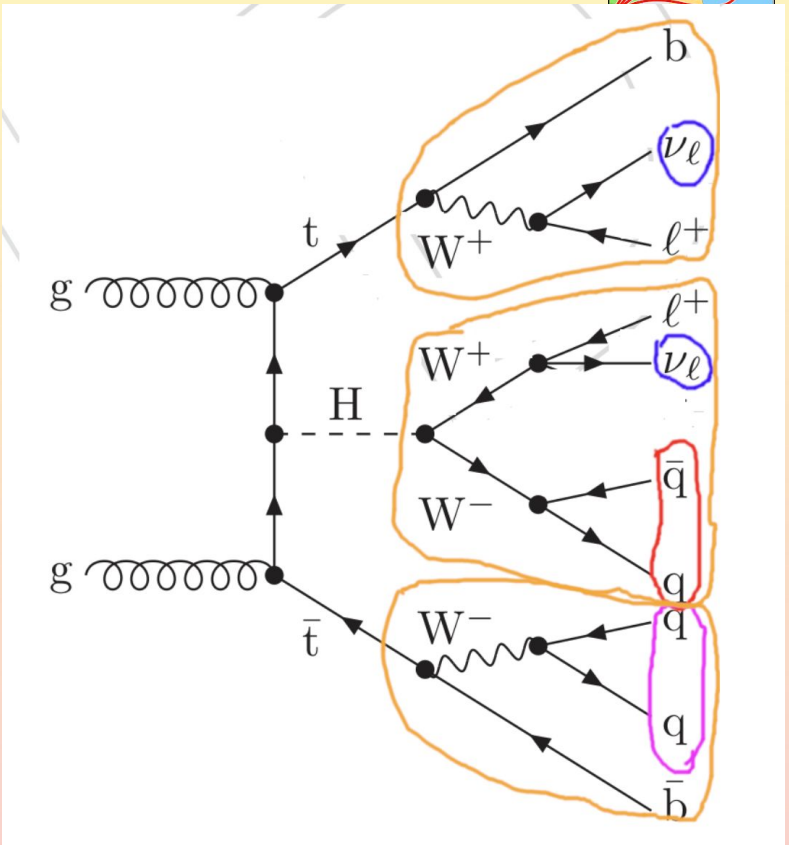
Slide credit: Javier Duarte and Jennifer Ngadiuba

The CERN IML

- CERN Interexperimental Machine Learning Working Group
  - IML organizes monthly meetings on a variety of subjects. These meetings are often topic-oriented (focusing on a certain ML technique) and may include external experts. Each spring IML organizes an annual workshop typically comprised of roughly 300 participants, which includes invited data scientist's talks, submitted talks, and tutorials.
  - IML also serves as entry point to find LHC specific machine learning resources, such as software solutions for machine learning starting from the common ROOT file format. We build a forum for community driven summaries of software solutions, announce LHC tailored trainings/school, and list relevant papers and people involved. We can help finding temporary hardware resources (GPUs) for tests. We are currently building up a database with benchmarks datasets and challenges in order to better enable testing new methods in our domain against previous ones.
- Write iml.coordinators@cern.ch if you are interested in presenting or want to propose a topic!
- Useful links
  - IML meetings → https://iml.web.cern.ch/meetings
  - IML mailing list → https://iml.web.cern.ch/forum
- Current coordinators
  - Anja Butter (TH), Stefano Carrazza (TH), Fabio Catalano (ALICE), Julián García Pardiñas (LHCb), Lorenzo Moneta (SFT), Pietro Vischia (CMS, to be replaced soon), Daniel Whiteson (ATLAS)

Today's tutorial: the dataset

- Associated production of a top quark pair and a Higgs boson
  - Main irreducible background: ttW
  - Other backgrounds: we provide Drell-Yan
- Multilepton final state
  - ambiguity in lepton and jet assignment makes it difficult to reconstruct the Higgs
  - Some Ws may be off shell
- Classification
  - easy to distinguish ttH or ttW from DY
  - Difficult to distinguish ttH from ttW
- Regression
  - Difficult to regress Higgs quantities (e.g. pT)
- Ntuples from HIG-23-015



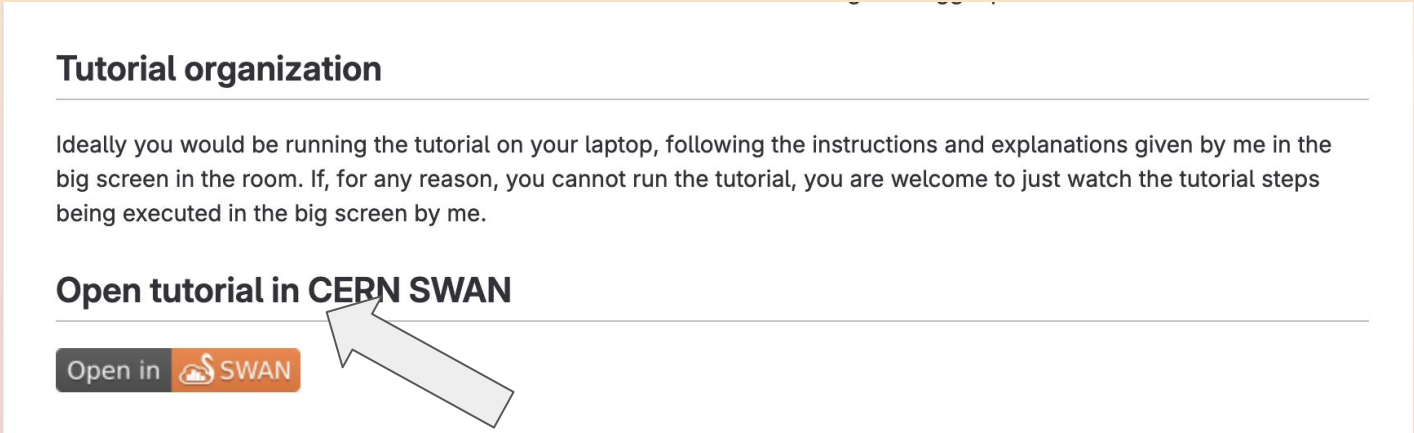
Today's Tutorial

- Go to https://gitlab.cern.ch/cmsdas-cern-2024/short-ex-mlg
- Scroll down to the README.md
- Click on "Open tutorial in CERN SWAN" (see figure below)
- Suggested path: go through exercise #1 (data inspection, BDTs), then choose one among the others, or 2+3+compare results, or 4+5+compare results:
  - #2: binary classification
  - #3: multiclass classification

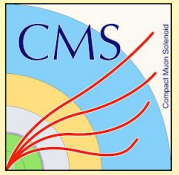
  } They build on top of each other

  - #4: regression using a dense neural network
  - #5: regression using a convolutional neural network

  } A bit more advanced



Learning Resources

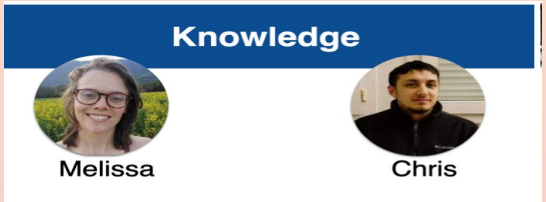


- All the main libraries come with a large set of tutorials
  - (see e.g. https://keras.io/, https://scikit-learn.org/, https://pytorch.org/, https://www.tensorflow.org/)
- The CMS ML documentation: cms-ml.github.io/documentation
- Many free courses
  - (e.g. https://coursera.org, https://www.fast.ai/, https://developers.google.com/machine-learning)
- Many good books
  - (e.g. https://www.deeplearningbook.org/, https://hastie.su.domains/ElemStatLearn/ , https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/index.html
- Other online resources
  - (e.g. https://www.youtube.com/watch?v=Ow25mjFjSmg)
  - Many tutorials on youtube (warning: quality may vary. A good one: https://www.youtube.com/@sentdex)
- **Make sure you keep learning!** ML is a fast evolving field, what is cutting-edge today will be ~obsolete in a couple of years from now

More info in the next slides (for offline consumption)

ML Subgroup: Knowledge

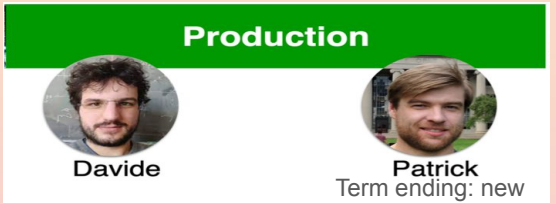
- Knowledge subgroup collects, maintains, and disseminates knowledge of ML algorithms in the CMS collaboration
  - Development and maintenance of CMS ML benchmarks, comparing and tracking the performance of algorithms, platforms, and ML frameworks on a set of benchmark CMS ML applications in reconstruction, simulation, trigger and computing
  - Consolidate tutorials and lectures on ML for CMS with the School Committee
  - Maintain list of in-house experts in various ML topics.
  - Prepares the ML part of the analysis questionnaire in collaboration with the Statistics Committee and documents good ML practices
- Documentation webpage: https://cms-ml.github.io/documentation/
  - Contributions welcome: https://github.com/cms-ml/documentation/pulls



Slide credit: Javier Duarte and Jennifer Ngadiuba

ML Subgroup: Production

- Production subgroup delivers production-level training and inference for CMS ML algorithms
  - Development and maintenance of ML training and inference workflows in the CMS software stack, including model inference and support for external frameworks such as TensorFlow, PyTorch, ONNX, and MXNet
  - Work closely with CMS framework experts and all relevant O&C groups, for example, overseeing framework-related aspects and relevant software/computing groups
  - Development of training infrastructure to satisfy the needs of as many collaborators as possible

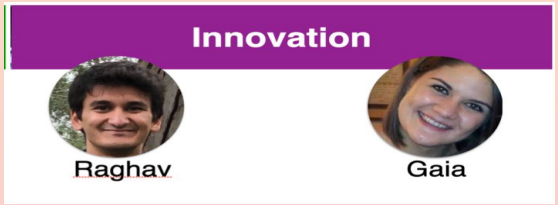


Term ending: new convener to be announced soon

Slide credit: Javier Duarte and Jennifer Ngadiuba

ML Subgroup: Innovation

- Innovation subgroup identifies and applies new ML techniques to CMS challenges while working closely with CMS groups in areas where ML is expected to have a significant impact: reconstruction, trigger, simulation, and beyond
  - Discuss the relevance of new techniques and help with the adaptation and implementation of specific models
  - Develop specific methods for CMS that will lead to technical publications
  - Lead the organization of ML-oriented hackathons and challenges to help identify new applications and ML techniques in CMS
  - Organize CMS internal journal club to discuss new and relevant ML results from inside and outside of particle physics



Slide credit: Javier Duarte and Jennifer Ngadiuba