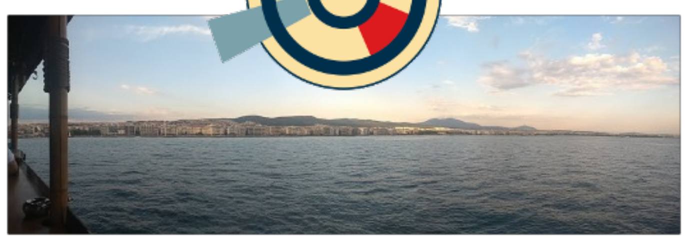
Machine learning for jets (reconstruction)

Henning Kirschenmann
(Helsinki Institute of Physics)



Machine learning for jets (reconstruction)	*Henning Kirschenmann*
Amphitheater of Central Library of Aristotle University of Thessaloniki, right in the middle of the campus, Thessaloniki	11:00 - 11:30

- track selection
- regression
- b-tagging
- q/g tagging

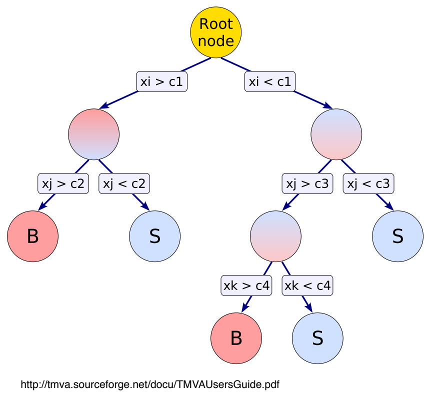
Machine learning for jets (substructure)	*Steven Schramm*
Amphitheater of Central Library of Aristotle University of Thessaloniki, right in the middle of the campus, Thessaloniki	11:30 - 12:00

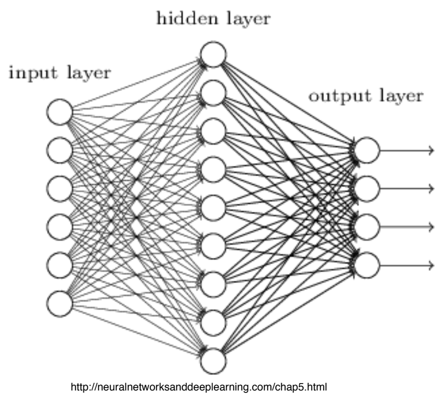
- W/Z tagging
(- Top tagging)
- Mass-decorrelated tagging
- Mass calibration

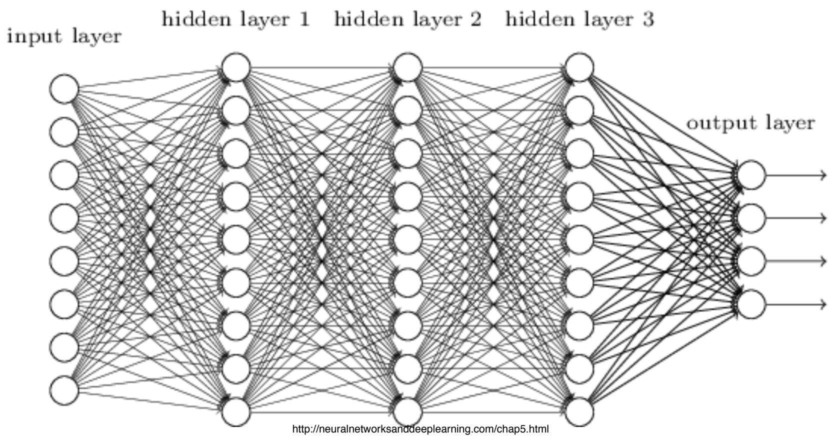
Disclaimer: Will report on what “is already there” in the experiments/around the corner, more in the pipeline…

For more details, e.g.
Machine Learning for Jet Physics (12/2017)
2nd IML Machine Learning Workshop (04/2018)
First EWSB Spring School (04/2018)

The new players



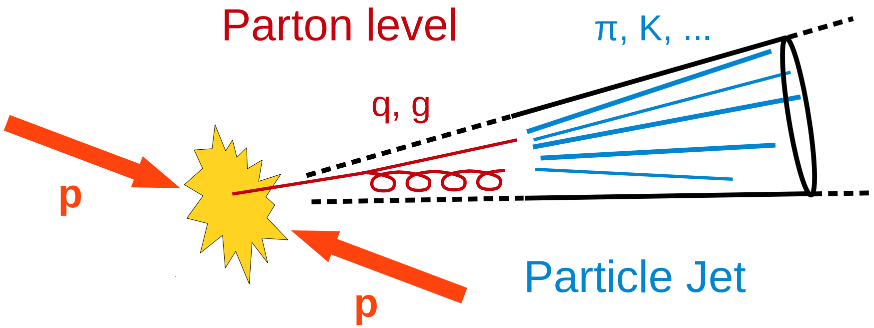


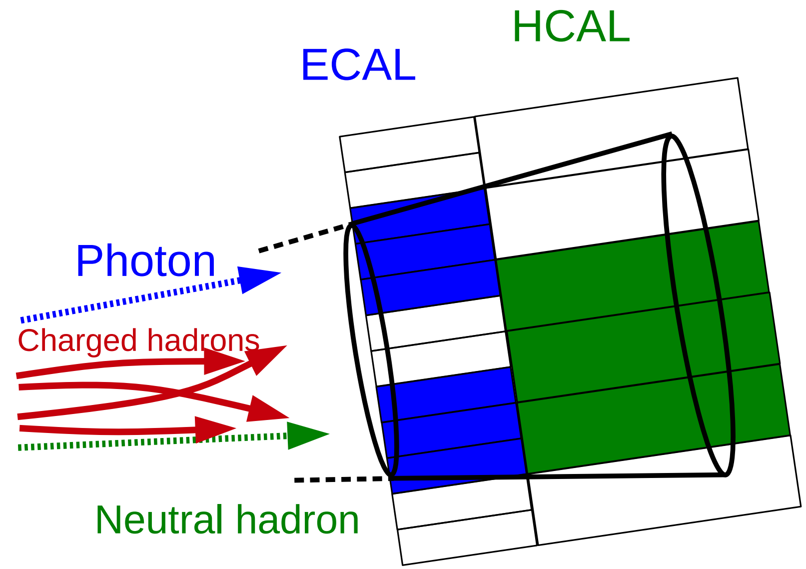


- Machine learning already used for a long time in HEP
- BDTs/shallow NNs
- TMVA/ROOT most widely used for a long time

Industry/ML community moved on

- Many open source/industry tools with huge community/big money behind them
- DNNs being adopted more and more by HEP community - can handle lower level inputs





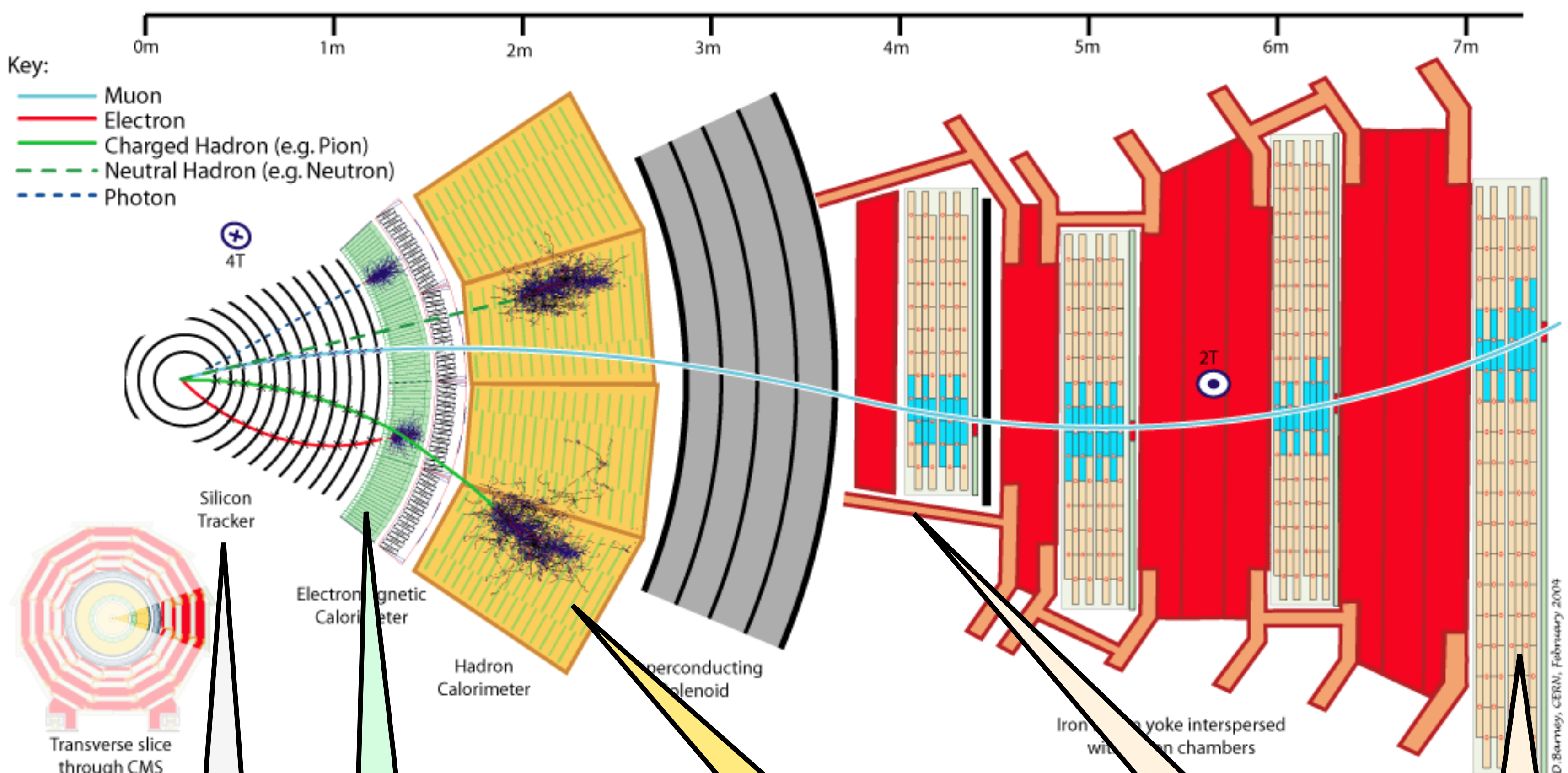
How to define flavour of [fat] jet for ATLAS/ CMS multi-classification approaches?

How to optimise low level reconstruction [in CMS]?

How to use particle flow event interpretation most efficiently?

What about DeepPFCandidates?

Particle Flow (PF) approach



Silicon Tracker

Position, momentum of charged particles : **$e^{\pm}$, $\pi^{\pm}$, $\mu^{\pm}$**

Electromagnetic Calorimeter

Position & ID, energy of **$e^{\pm}$,γ, π^{0}**

Hadron Calorimeter

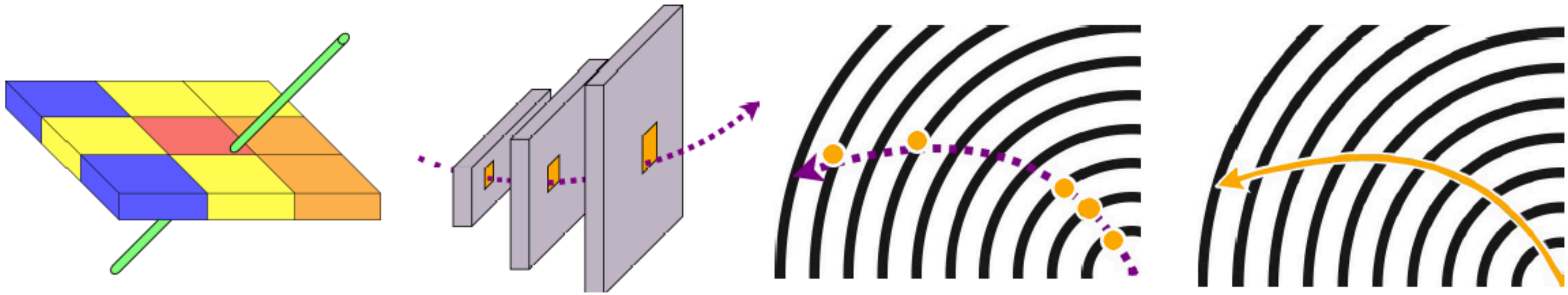
Energy of hadrons : **p, n, $\pi^{\pm}$, K ..**

Muon Chambers

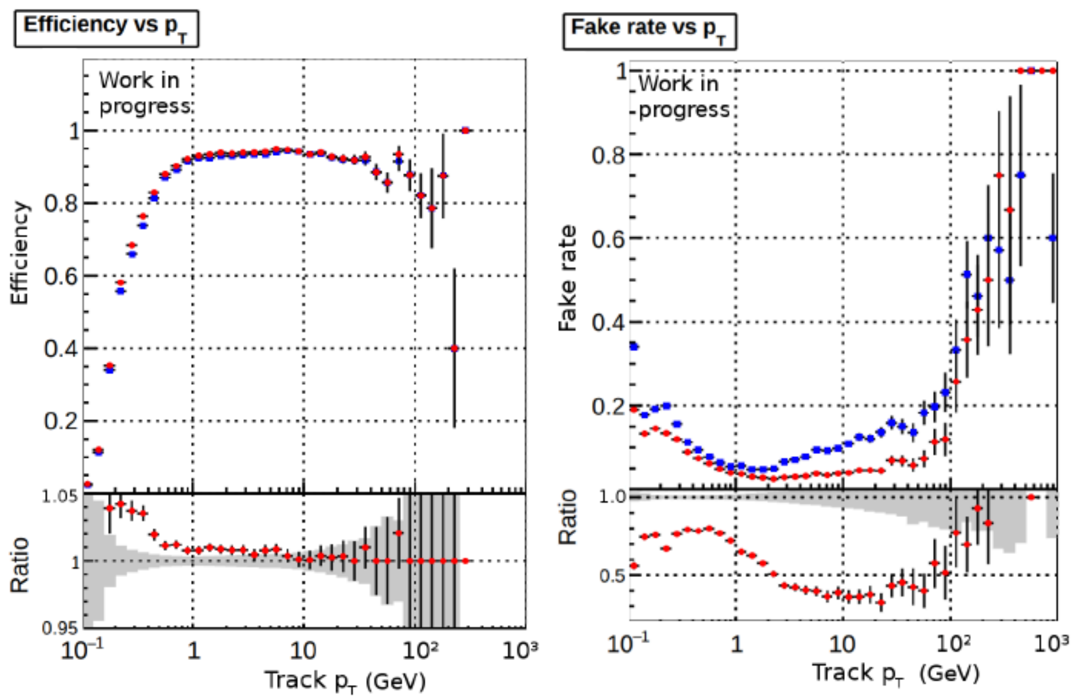
Position & momentum of **$\mu^{\pm}$**

Track selection

Tracking core of particle flow



Hit clustering **Track seeding** **Track building** **Track fitting**

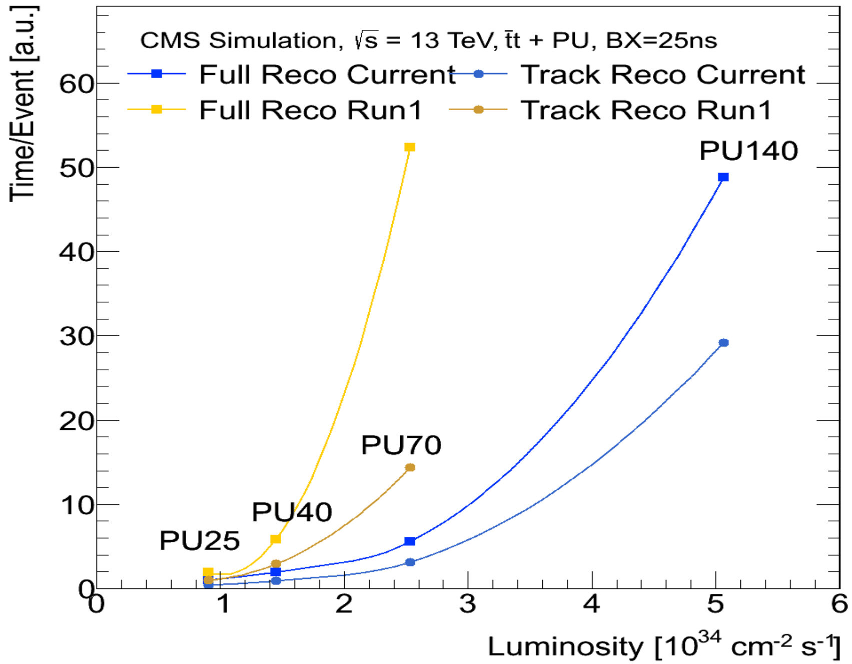


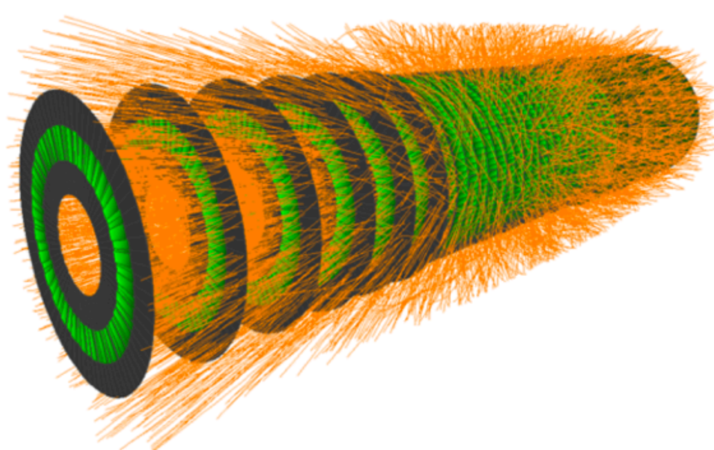
Ongoing:

- Track quality estimator
- Replacing 11 different BDTs used for each offline tracking iteration by a single DNN
- Performance promising
  - Higher efficiency/lower fake rate

[Connecting the dots, Joona Havukainen]

Tracking at HL-LHC (and Kaggle Challenge)





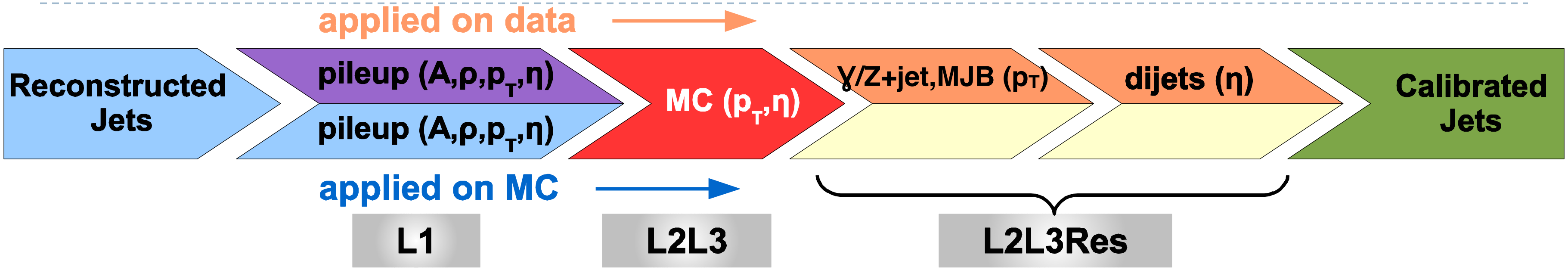
Looking ahead:

- Tracking remains huge combinatorial challenge
- No fundamental change in approach so far
- Survived with [code] optimisations, but [probably] not feasible for HL-LHC and beyond
- Kaggle challenge to collect new ideas
    - Not as “simple” in terms of ML as 2014 Higgs classification challenge

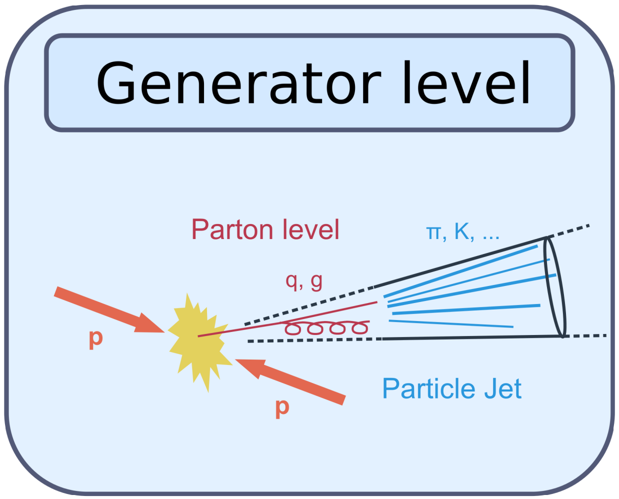
https://www.kaggle.com/c/trackml-particle-identification

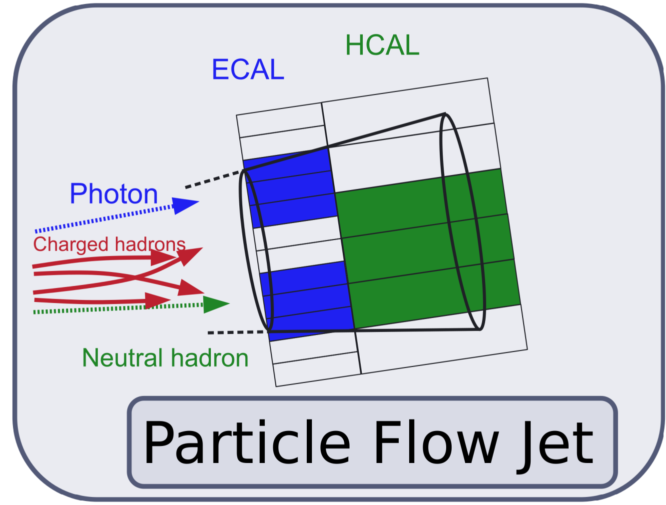
Regression

Jet energy corrections (state-of-the-art)



JEC corrects reconstructed jets - on average - back to particle level
$\langle p_{T,reco}\rangle / \langle p_{T,gen}\rangle = 1$
(vs. p_T^{gen} , η, A, pileup μ)



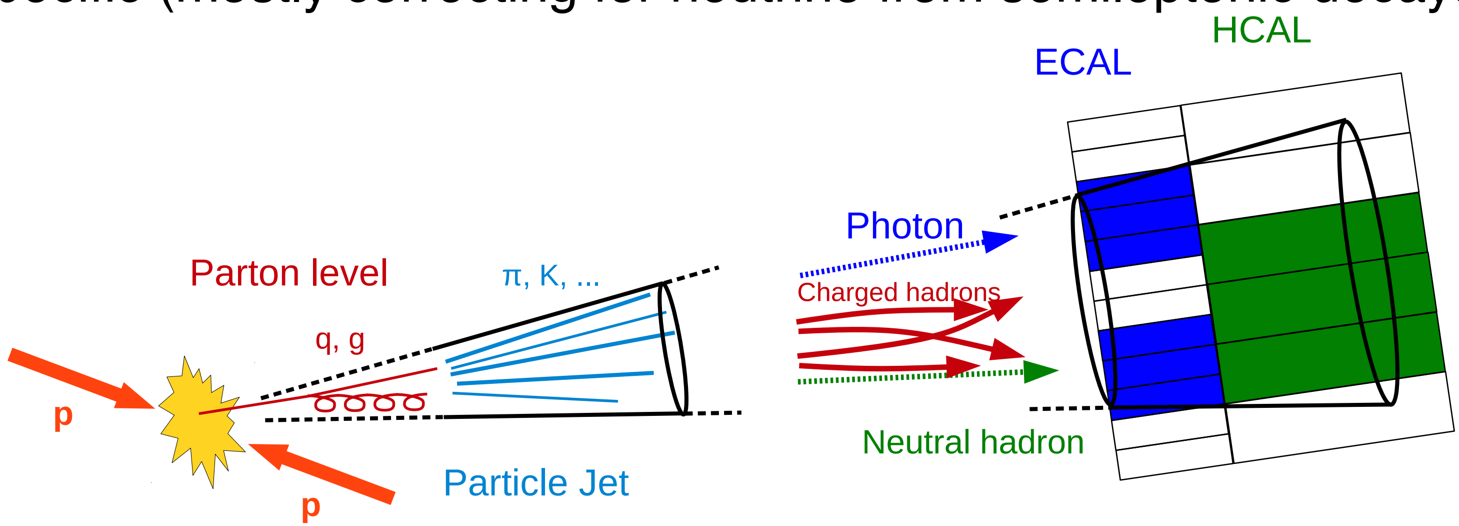


Factorized approach to JEC:

- Pileup corrections to correct for offset energy (noPU vs. PU jet matching)
- Correction to particle level jet vs. p_T and η from simulation
- Only for data: Small residual corrections (Pileup/relative and absolute) to correct for differences between data and simulation

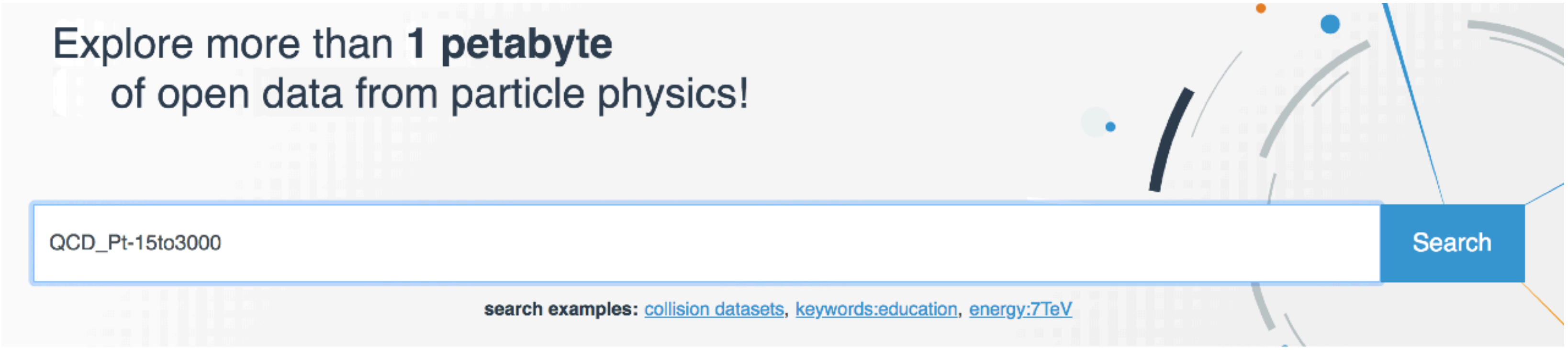
Jet energy regression

JEC so far only parametrised as a function of p_T, η, A, ρ in CMS for ~all analyses
B-jet energy regression used in some places (e.g. $H \rightarrow bbar$), analysis-specific (mostly correcting for neutrino from semileptonic decays



- Correcting for dependence on single observables: Marginal gain for PF jets (useful for calo jets, cf. ATLAS global sequential calibration)
- DNNs on low level (PF candidates/jet images) might give performance boost
- Extra challenge: Would like to have it universally applicable

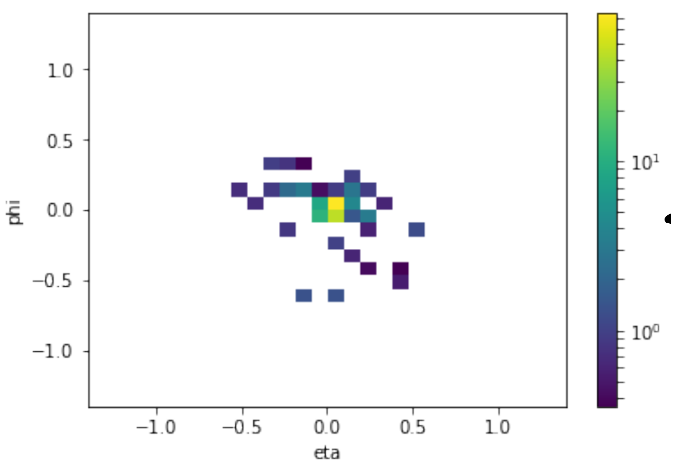
Jet energy regression

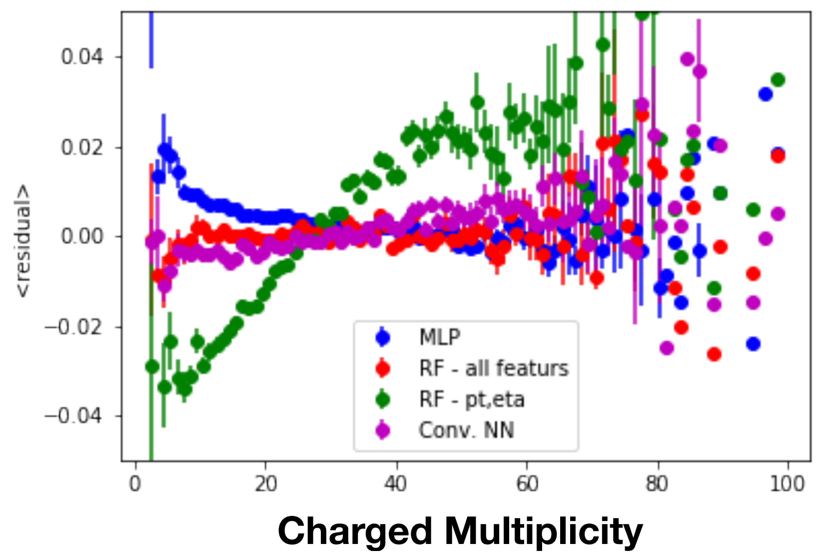


- One example: [Jet Response Prediction Using Jet Images (Machine Learning for Jet Physics Workshop)]()

Setting up a DCNN

- Creating Multiple convolutional filters
- The larger the filter the more physics it captures - reduces effect of sparsity

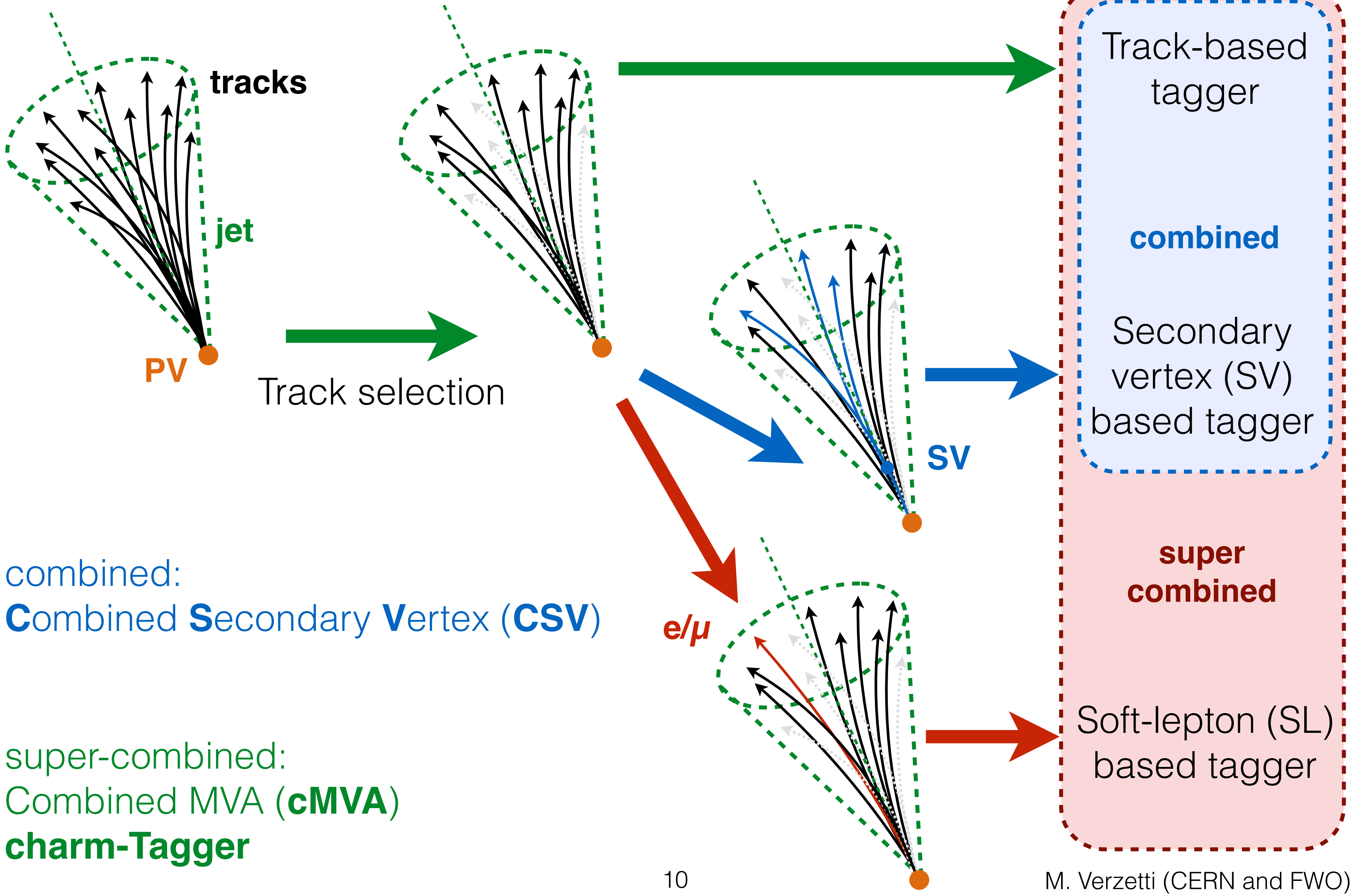




- Network captures dependence of response on observables sensitive to fragmentation

Heavy flavor/b-tagging

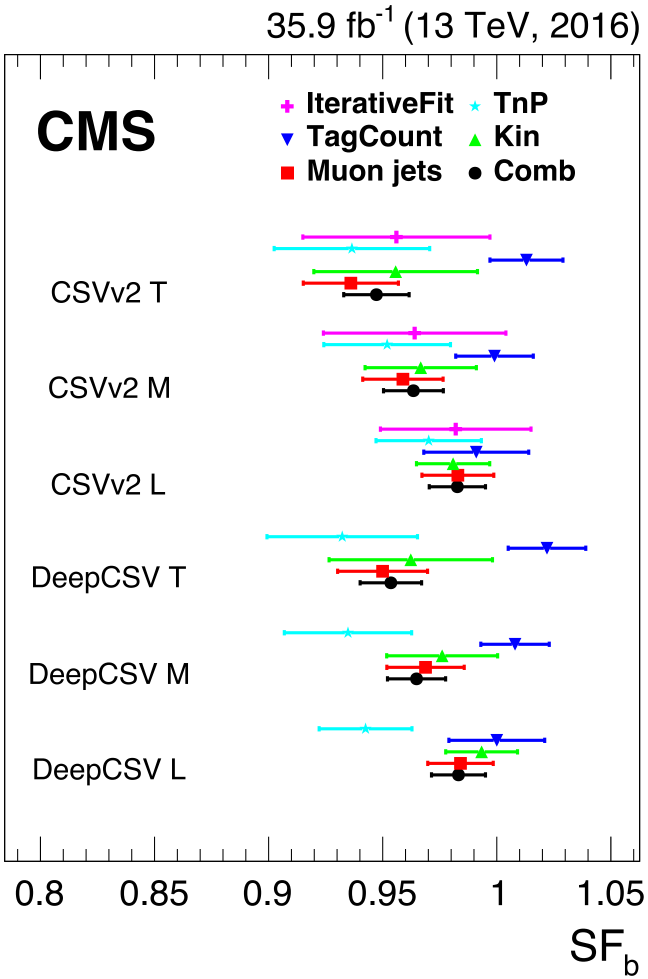
HF tagging @ CMS — arXiv:1712.07158

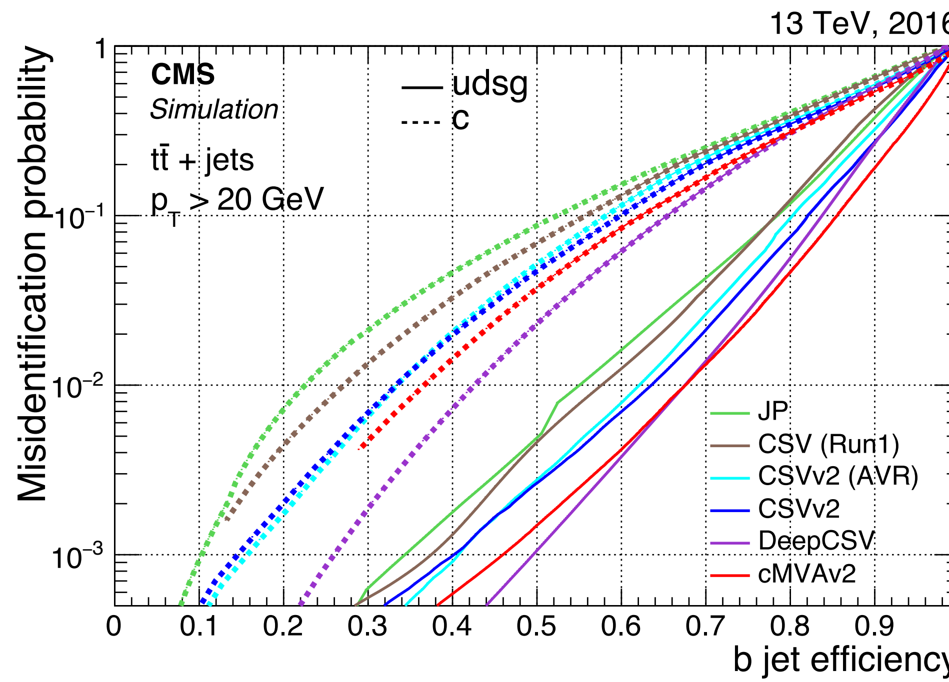


combined:
Combined **S**econdary **V**ertex (**CSV**)

super-combined:
Combined MVA (**cMVA**)
charm-Tagger

Heavy flavour tagging (b and c)

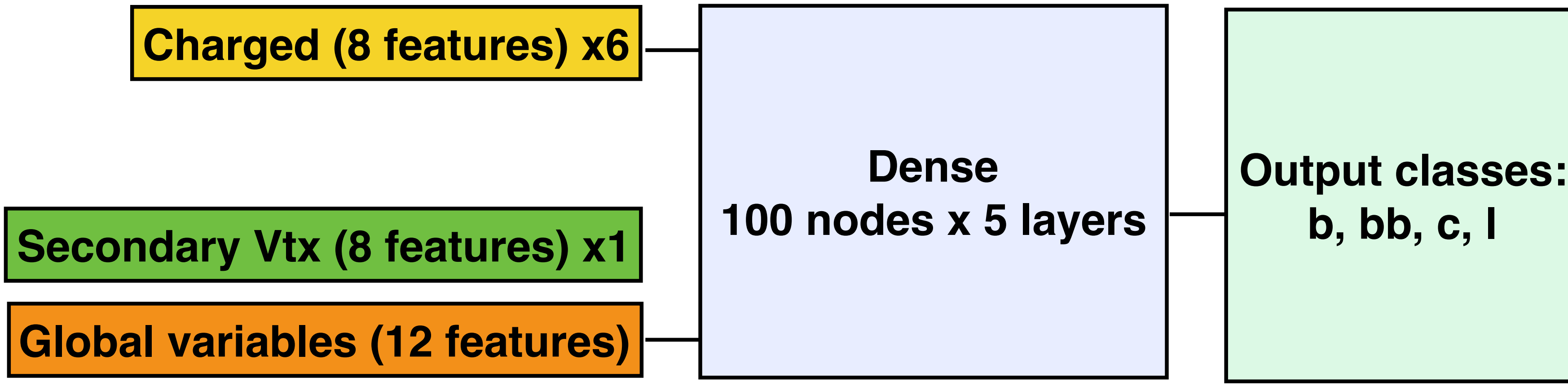




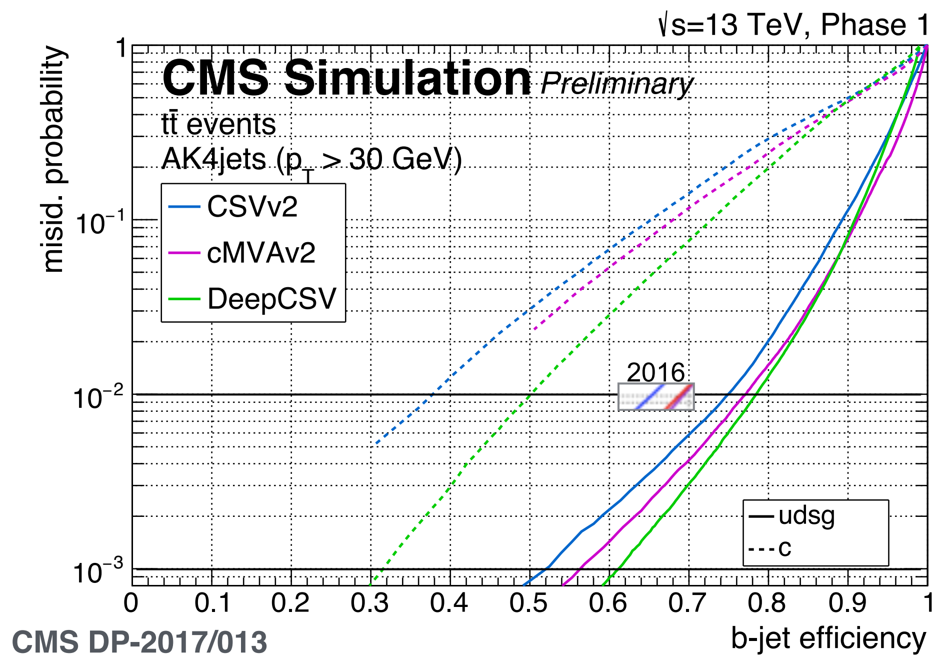
Towards end of 2016:

- New tagger coming into the game
- DeepCSV - using ~same information as CSVv2, but performing significantly better; SFs fine

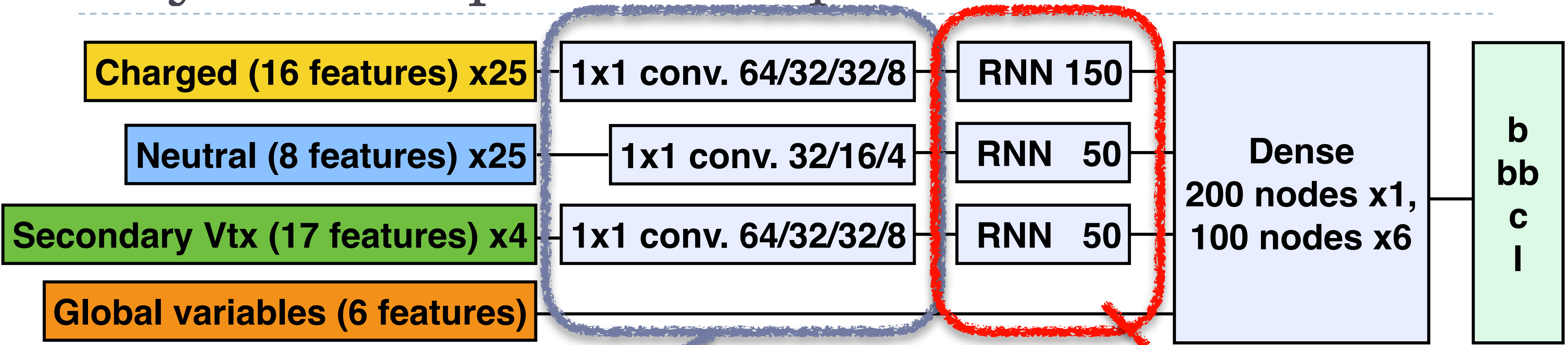
DeepCSV



- Current standard b-tagging algorithm in CMS
- Also used at High Level Trigger
- Significant gain without using [much] more input than previous taggers



Beyond DeepCSV: DeepJet

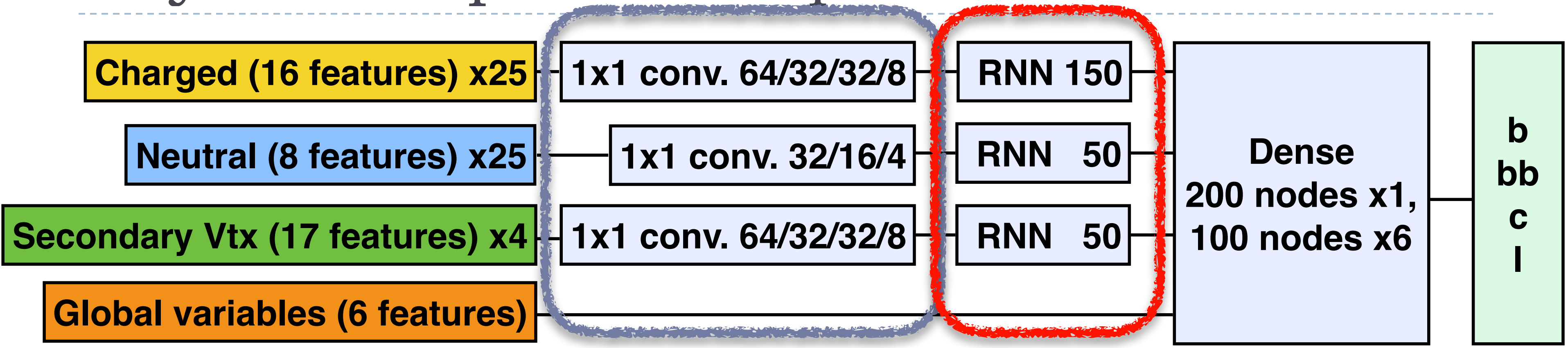


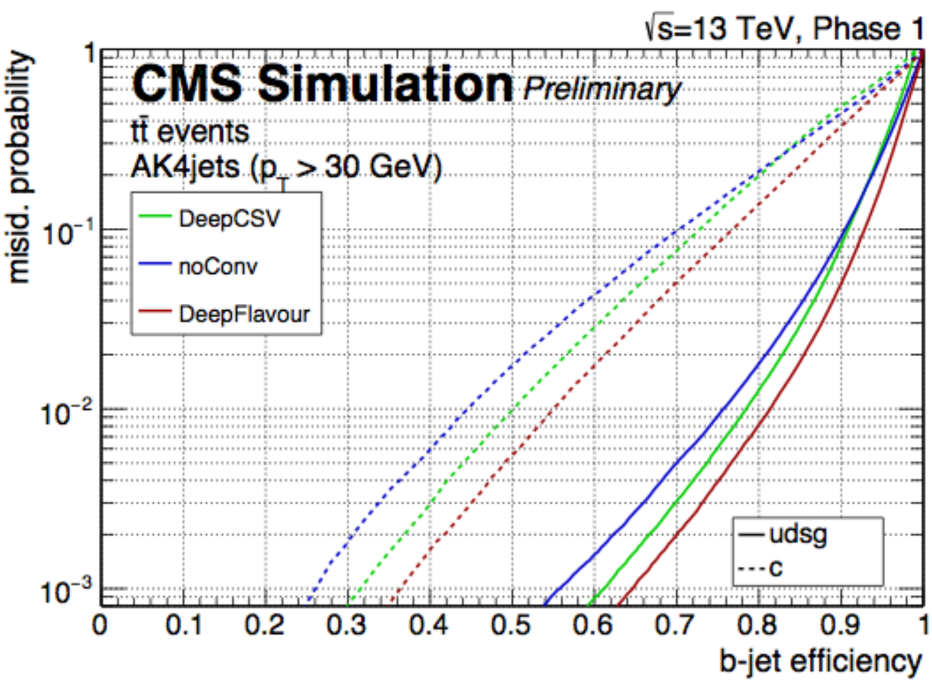
Convolutional layers progressively learn a more compact feature representation (automatic feature engineering)

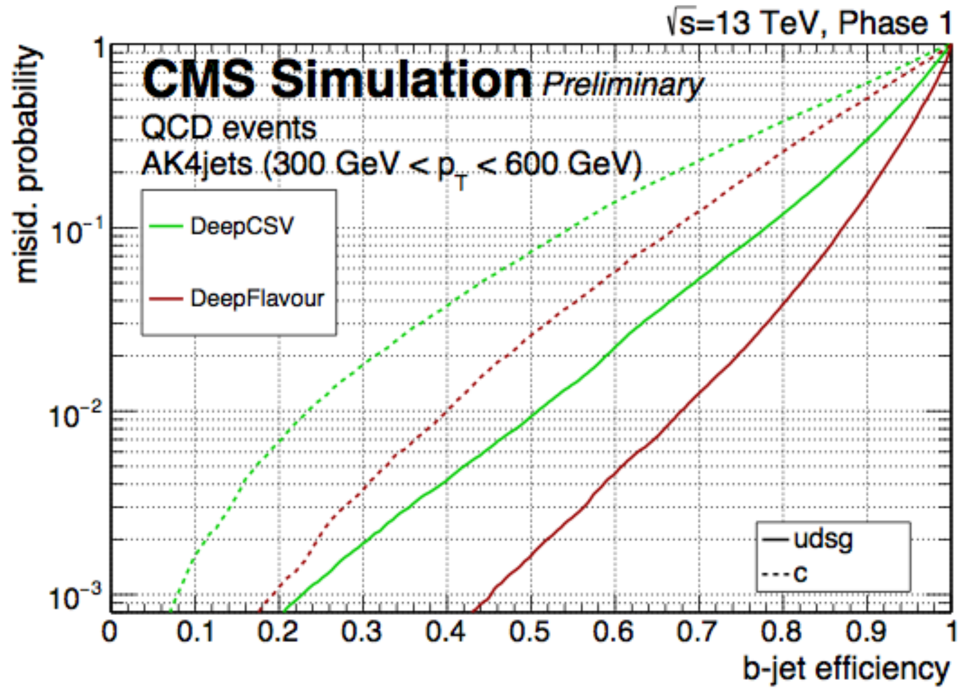
The recurrent layers (LSTM) builds a "summary" of the information contained in each set of feature types

- Starting directly from PFCandidate level, not using track selection from [Deep]CSV

Beyond DeepCSV: DeepJet

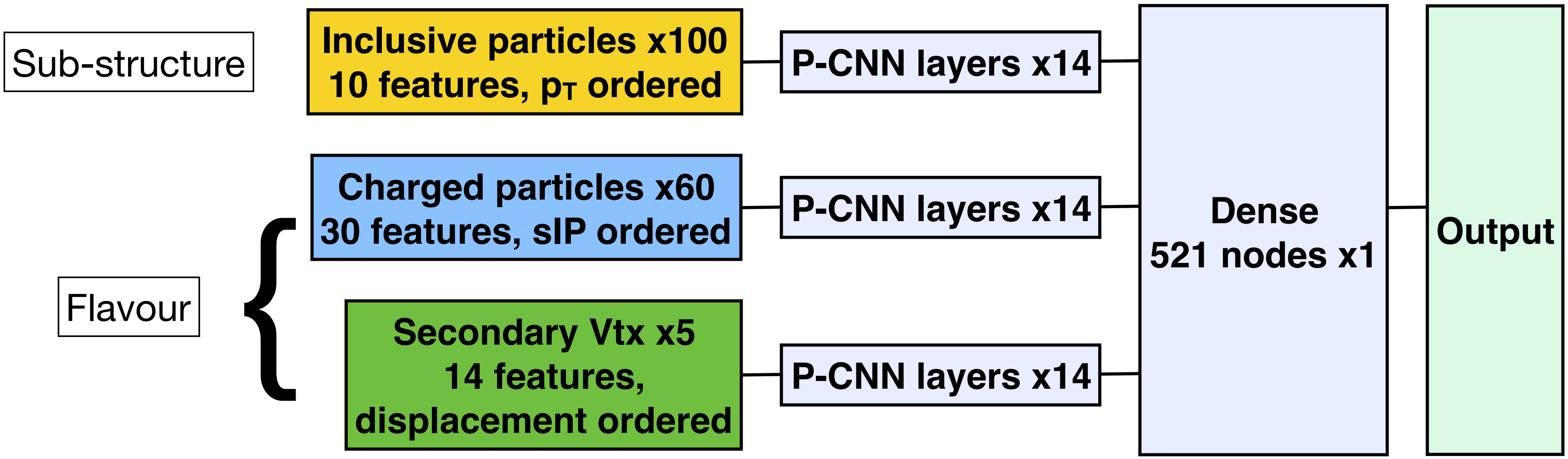






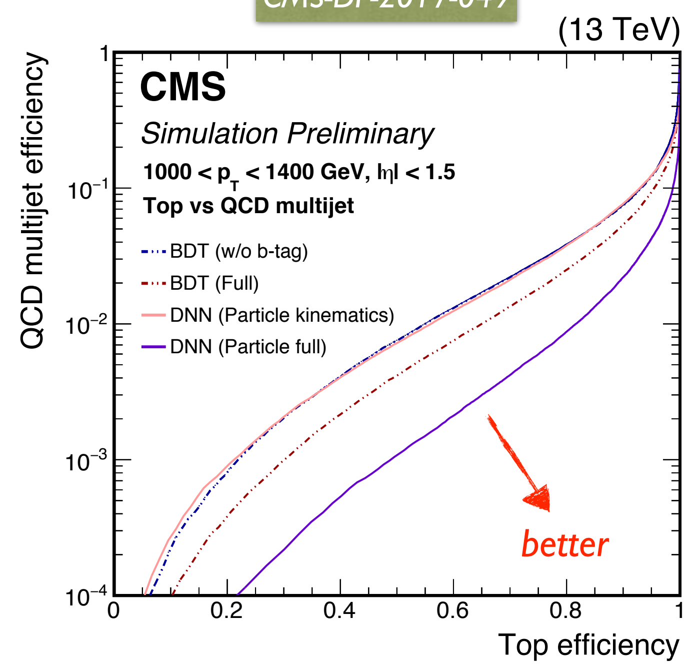
- Additional information not utilised without convolutional layers
- Huge gain where track selection was suboptimal

Going towards W/Z/H/top tagging



- Can use particle-based networks also for tagging of fat jets (many substructure taggers already on the market)
- [Again] way more inputs than for narrow jet b-tagging
- Network architecture inspired by imaging (RNNs computationally too expensive)
- ~4 times reduction in QCD multi jet misidentification for same efficiency

CMS-DP-2017-049



Quark/gluon tagging

What is a Quark Jet?

From lunch/dinner discussions

10:40 **Quark - gluon discrimination**

Speaker: Andrzej Konrad Siodmok (Polish Academy of Sciences (PL))

Ill-Defined ↓ **Well-Defined**

What people sometimes think we mean ↑ **What we mean**

Quark as noun

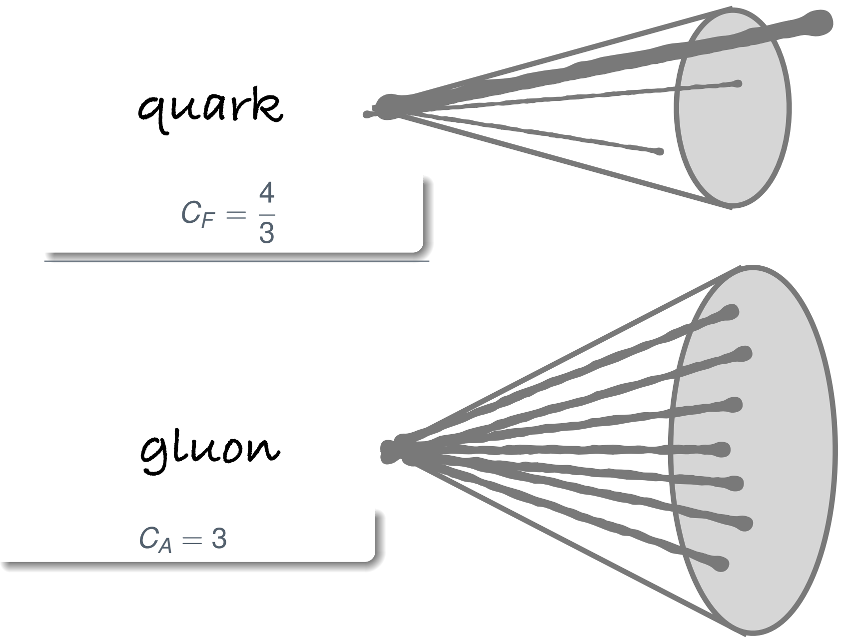
- A quark parton
- A Born-level quark parton
- The initiating quark parton in a final state shower
- An eikonal line with baryon number 1/3 and carrying triplet color charge
- A quark operator appearing in a hard matrix element in the context of a factorization theorem

Quark as adjective

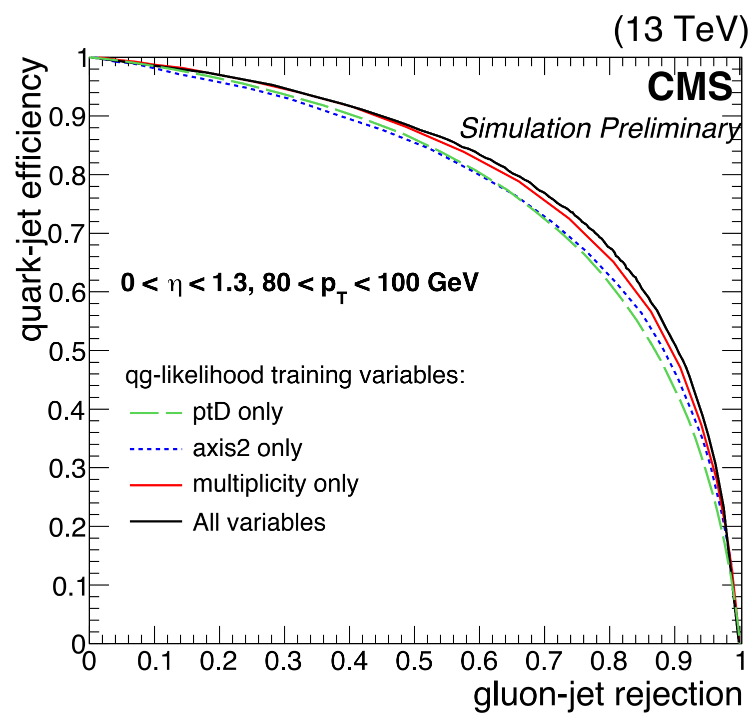
- A parton-level jet object that has been quark-tagged using a soft-safe flavored jet algorithm (automatically collinear safe if you sum constituent flavors)
- A phase space region (as defined by an unambiguous hadronic fiducial cross section measurement) that yields an enriched sample of quarks (as interpreted by some suitable, though fundamentally ambiguous, criterion)

In practice, experiments use some parton (quark/gluon) and/or hadron (b/c-tagging) flavour definition. Discussion on more consistent “truth labelling” ongoing in CMS

Quark-gluon discrimination in CMS



Quark-Gluon likelihood:
Known from Run1 - 3 input observables



PYTHIA: More discrimination
Herwig: Less discrimination
Data: Somewhere in between

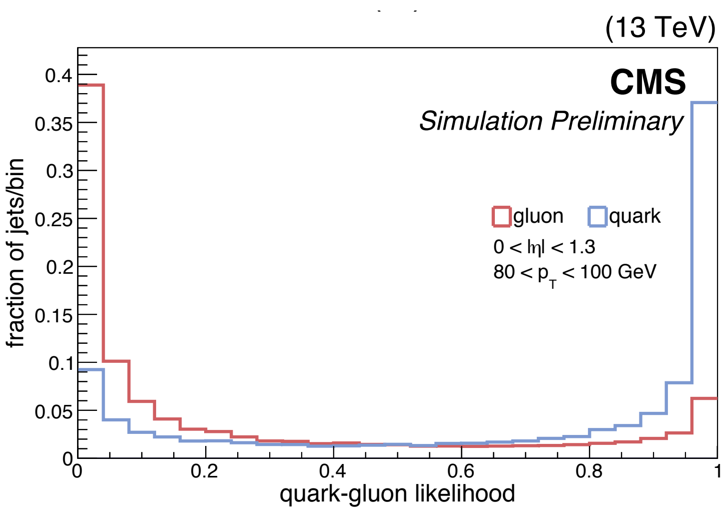
CMS DP -2016/070 CMS DP -2017/027

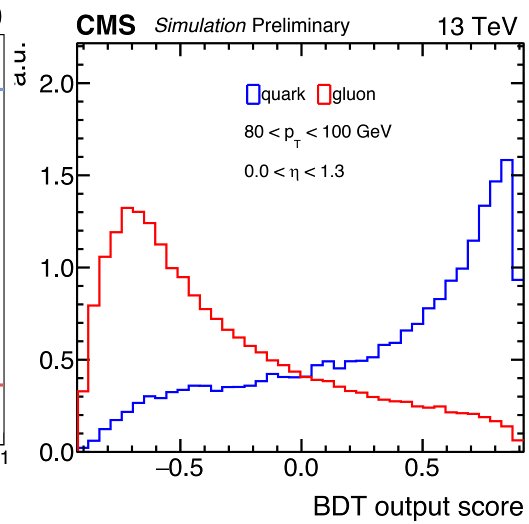
Quark-gluon discrimination in CMS

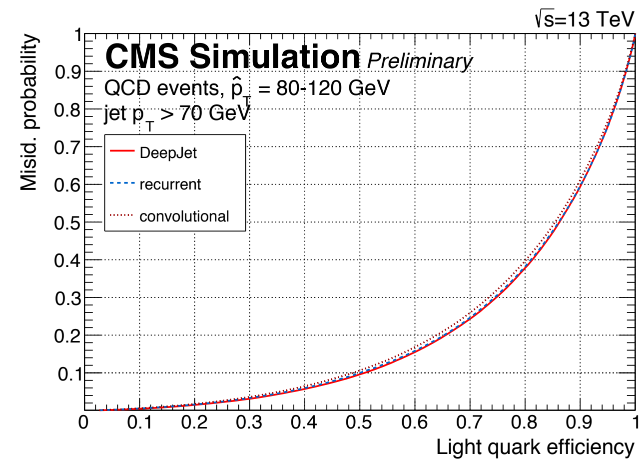
Quark-Gluon likelihood: Known from Run1 - 3 input observables (ptD,multiplicity,σ_2)

BDT - adding two more observables (σ_2, $\Sigma\log(pT/\Delta R)$ / jet pT)

DeepJet (same as for b-tagging)



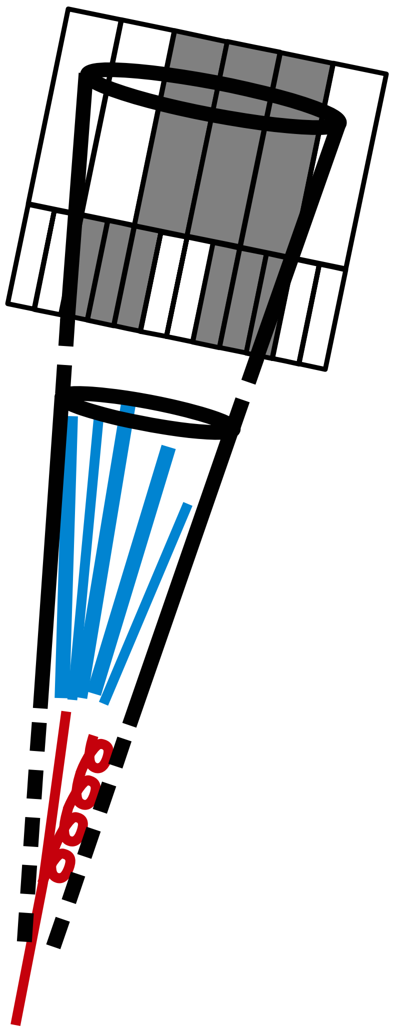




Many papers/ideas on the topic; particularly interesting to train without need for "truth labeling" or multi-classification a la DeepJet

CMS DP -2016/070 CMS DP -2017/027

Conclusions



Machine learning for jets (reconstruction) - status quo

- Has always been there
- BUT: Still took only first couple of steps towards adopting Deep Neural Networks in “production mode”
- Particle-based “brute force” low level taggers doing very well
- Regression difficult(?)

Machine learning for jets (reconstruction) - future

- Exciting times - many ideas and approaches haven’t been explored, yet
- Many new results around the corner (check e.g. BOOST2018)
- If you already know the optimal solution there is no point in ML

Backup