(Machine) Learning how to discover new particles at the Large Hadron Collider

Juan Rojo
VU Amsterdam & Theory Group, Nikhef
www.juanrojo.com, @JuanRojoC

SURF Research Bootcamp
Amsterdam, 10/6/2018

Machine Learning in high-energy physics

- **Huge, fast growing field**, with new applications being proposed every day
- Here restrict ourselves to **two representative examples**: if you want to learn more about other applications, don't hesitate to ask!
- For further **overviews of ML applications to high-energy physics** and related fields please see

☑ *Big data tools in Physics and Astronomy* (Amsterdam, https://indico.cern.ch/event/622093/)

☑ *Machine learning for Phenomenology* (Durham, https://conference.ippp.dur.ac.uk/event/660/)

☑ *Inter-Experimental LHC Machine Learning WG* (https://iml.web.cern.ch/)

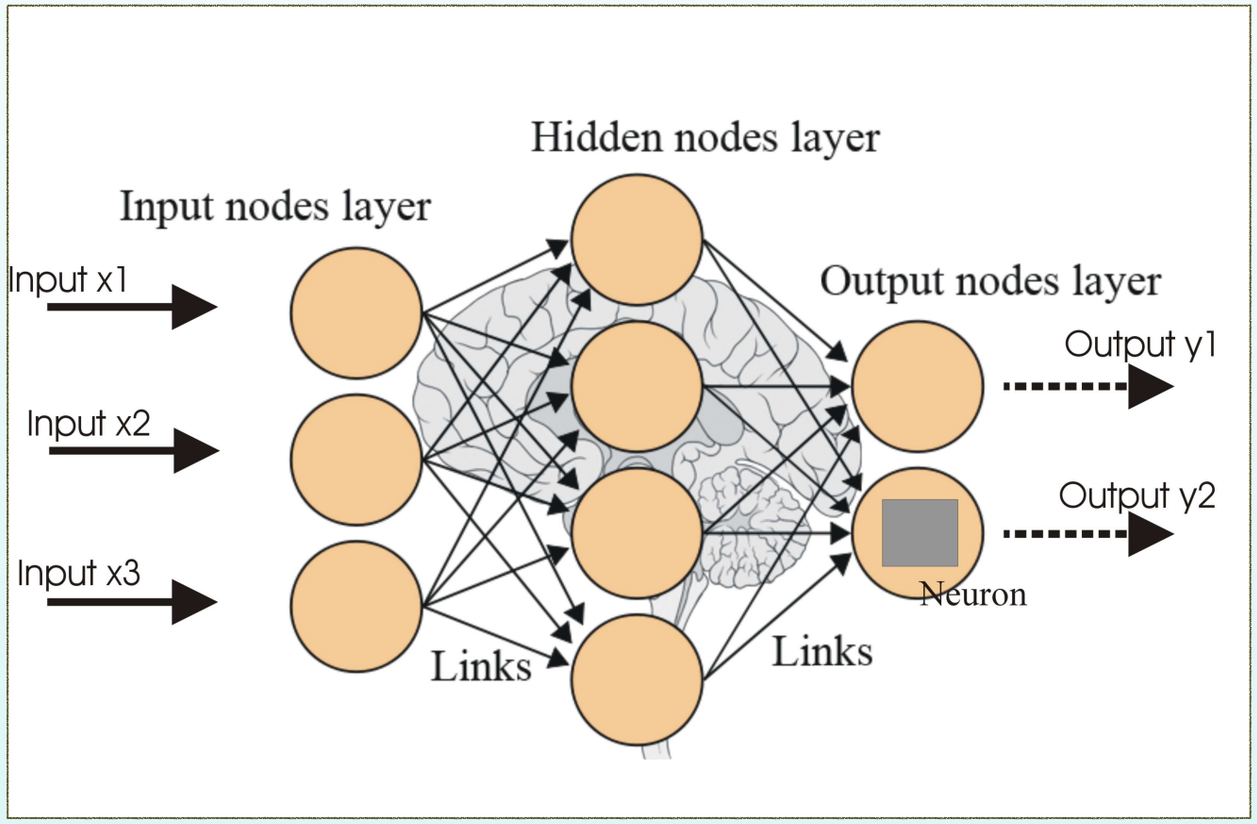
☑ *Accelerating searches for Dark Matter with Machine Learning* (https://indico.cern.ch/event/664842/)

☑ *CERN Data Science seminars* (https://indico.cern.ch/category/9320/)

Artificial Neural Networks

Inspired by **biological brain models**, **Artificial Neural Networks** (ANNs) are **mathematical algorithms** widely used in a wide range of applications, from **HEP** to **targeted marketing** and **finance forecasting**

From biological to artificial neural networks



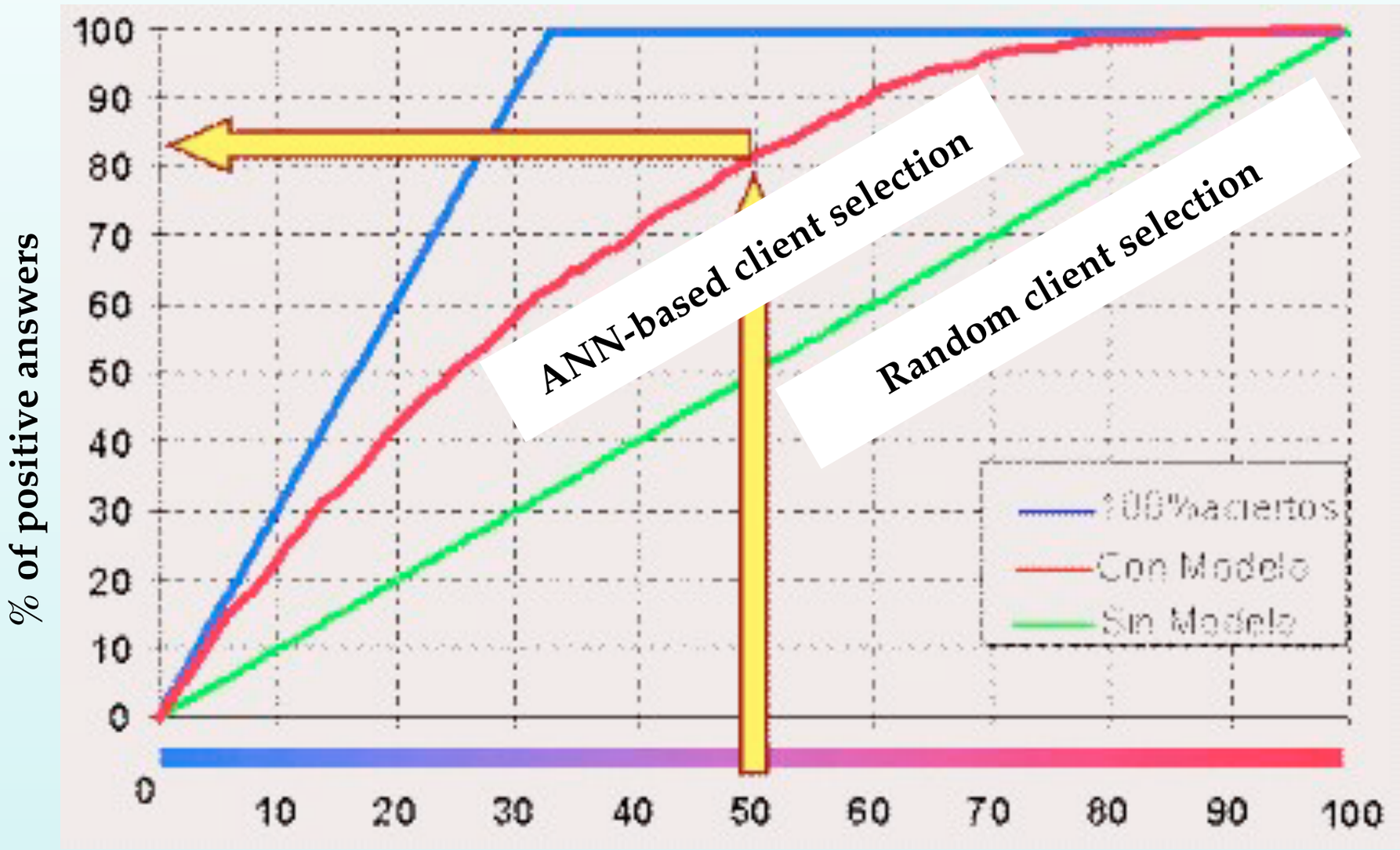
Artificial neural networks aim to excel where domains as their **evolution-driven counterparts outperforms traditional algorithms in tasks such as pattern recognition, forecasting, classification**, ...

ANNs - a marketing example

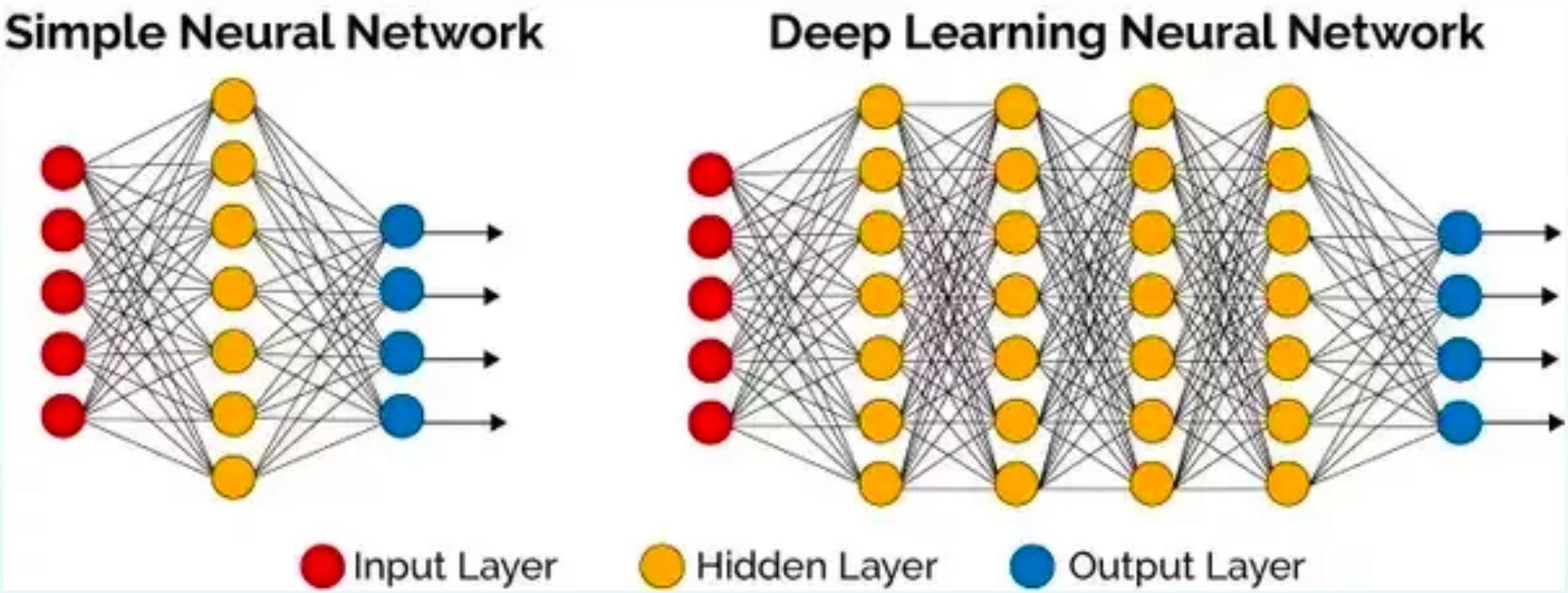
A bank wants to offer a new credit card to their clients. Two possible strategies:

- **Contact all customers**: slow and costly
- Contact **5%** of the customers, **train a ANN with their input** (savings, income, loans) and **their output** (yes/no) and use the information to c**ontact only clients likely to accept the product**

Cost-effective method to improve marketing performance!



Deep Neural Networks



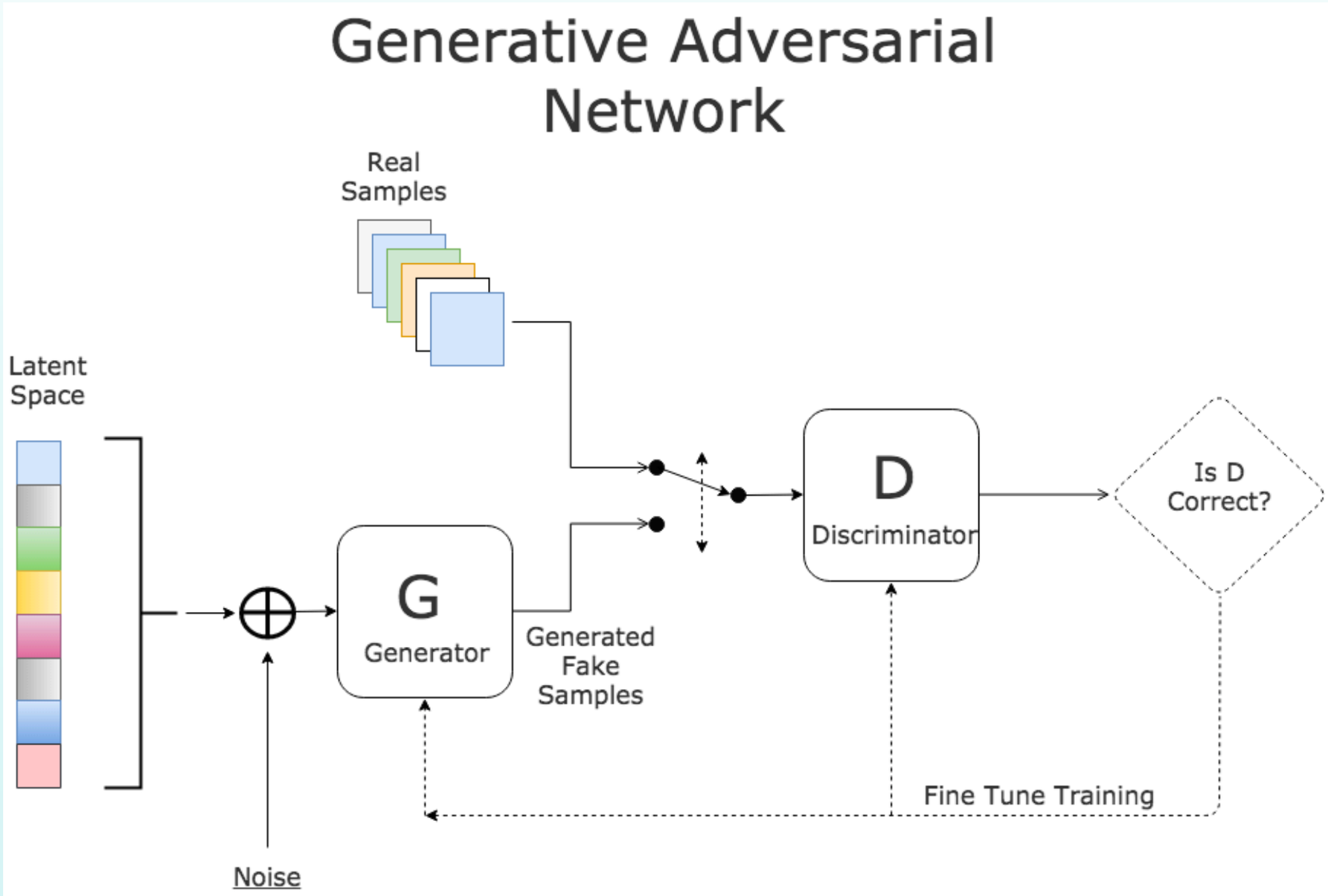
- A **Deep Neural Network (DNN)** is a standard multi-layer feed-forward perceptron with a large number of internal layers

- All types of neural nets eg **Recursive, Convolutional, Parametrised** etc can be made "deep" by adding more hidden layers

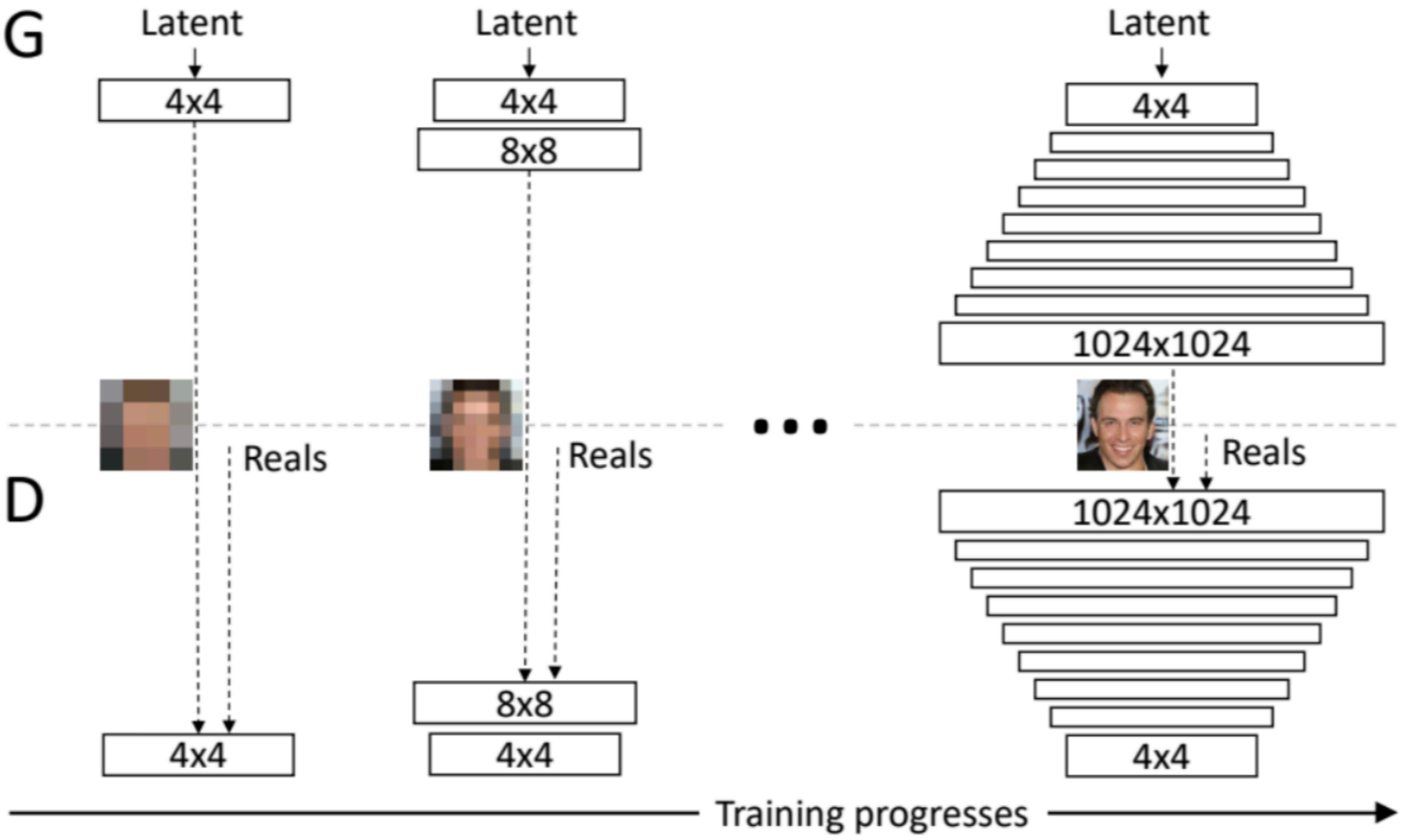
- For several applications, the **increased complexity** achieved this way leads to a significant improvement in performance

Generative Adversarial Networks

- New architecture for an **unsupervised neural network training** (unlabelled samples)
- Based on two **independent nets** that work separately and act as adversaries:
  - the **Discriminator (D)** undergoes training and plays the role of classifier, and
  - the **Generator (G)** and is tasked to generate random samples that **resemble real samples** with a twist rendering them as fake samples.



The many uses of GANs

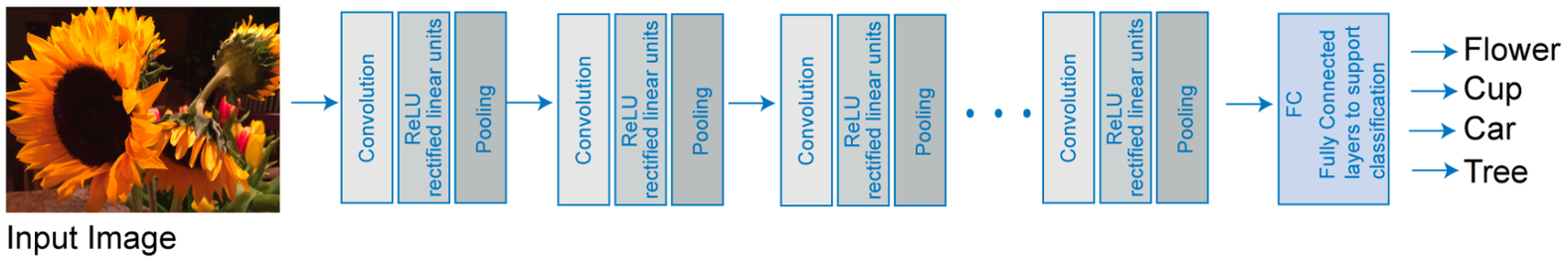


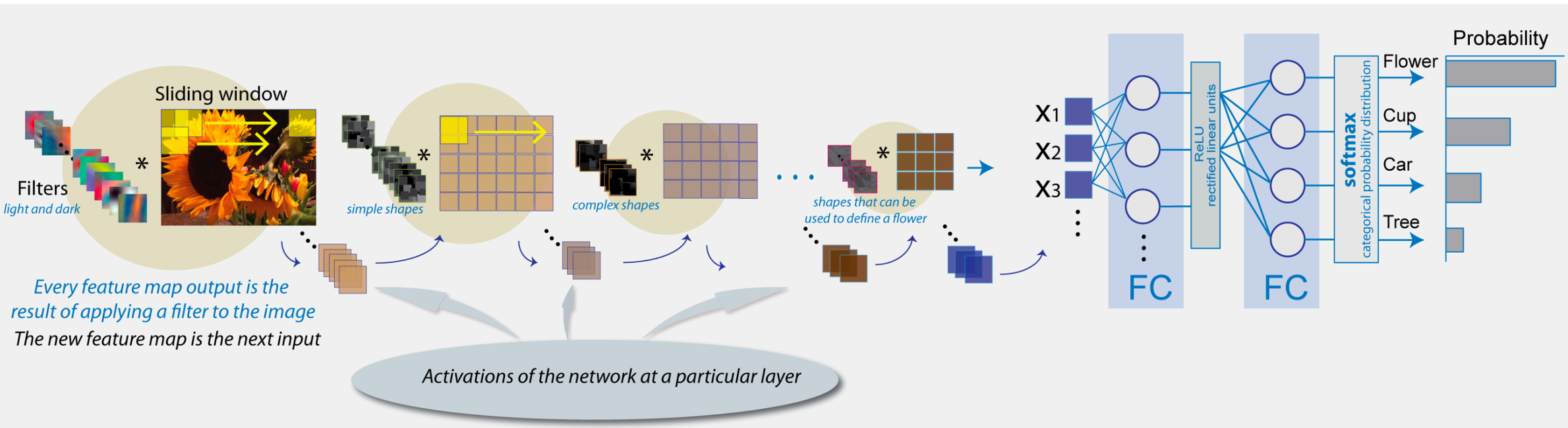
arXiv:1710.10196

Which one of these images are real and which ones are fake (generated by the GANs)?

Convolutional Neural Networks

- Convolutional Neural Networks (CNNs) have convolutional layers based on **filters**
- Each **filter** maps a group of numbers into a number, reducing the dimensionality of the data
- Specially useful for **pattern recognition** (eg for self-driving vehicles)

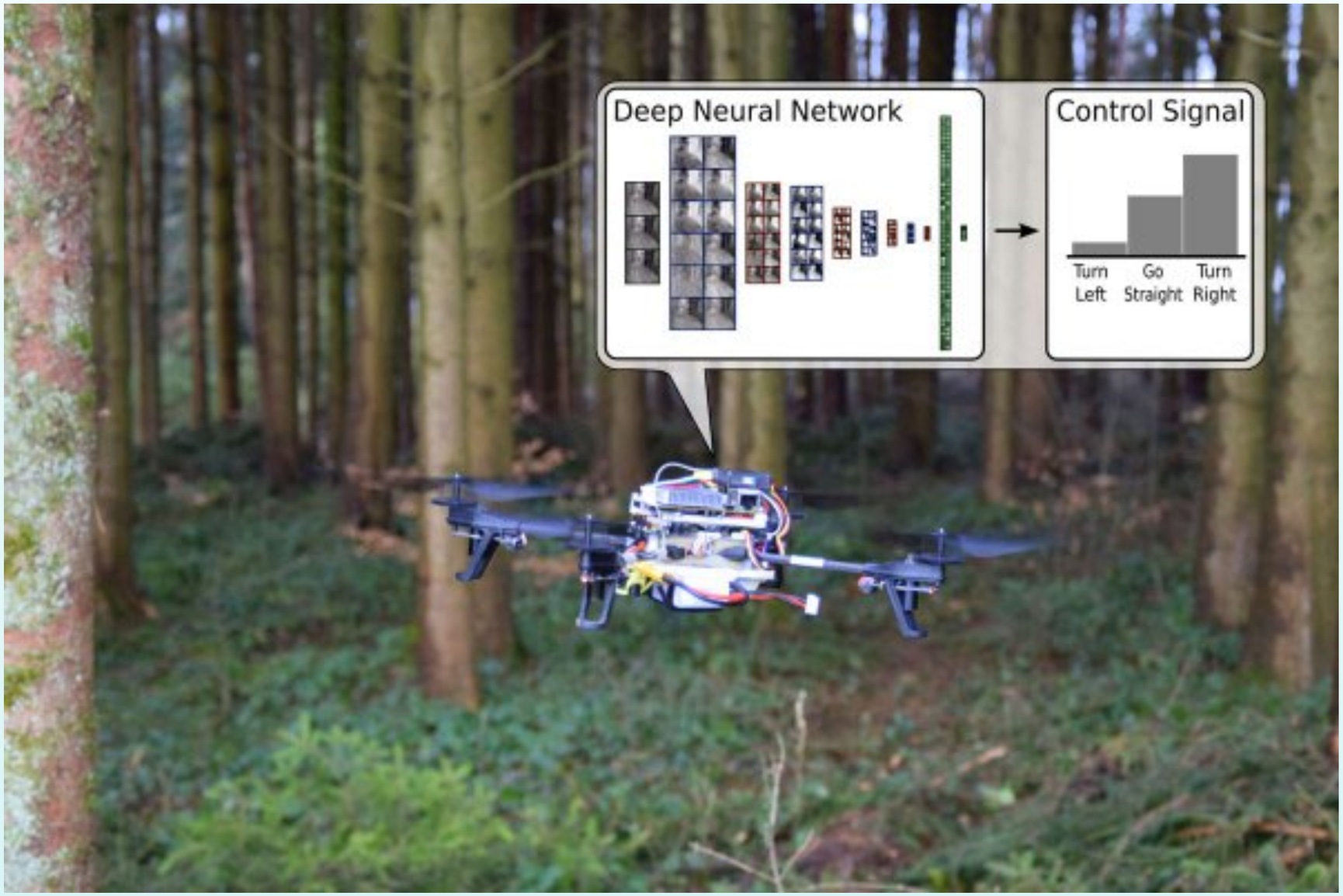




mathworks.com

Convolutional Neural Networks

- ANNs can enable an **autonomous vision-control drone** to recognise and follow forest trails
- Image classifier operates directly on **pixel-level image intensities**
- If a trail is visible, the **software steers the drone** in the corresponding direction

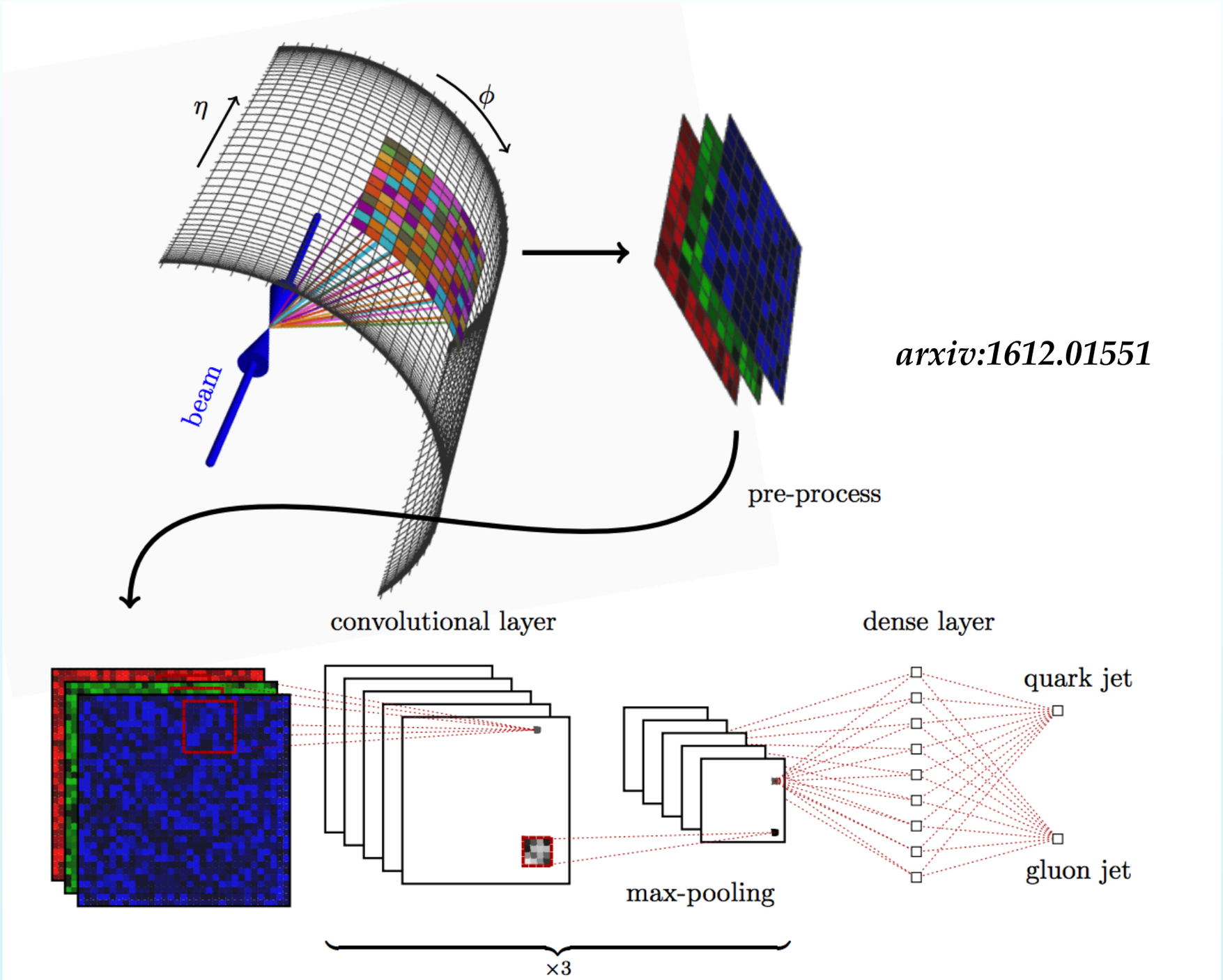


Giusti et al, IEEE Robotics and Automation Letters, 2016

Similar algorithms at work in self-driving cars!

Convolutional Neural Networks

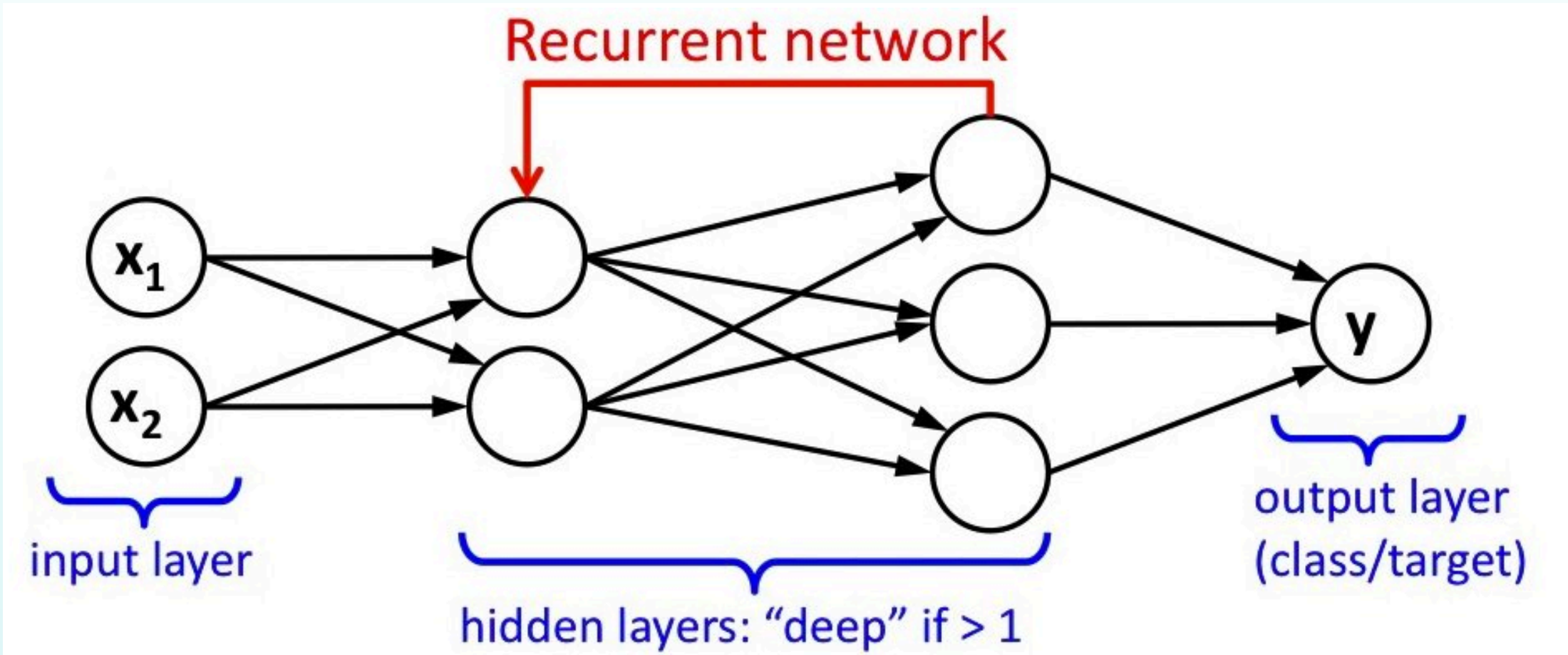
The results of the **collisions of high-energy particles** can be treated analogously to **image processing** using Convolutional Neural Networks



arxiv:1612.01551

Recurrent Neural Networks

RNNs use as inputs not just the current "training examples" but also **what they have perceived previously**: they have a **built-in notion of time ordering** useful for time-dependent functions



The output of a RNN at time time, ***y(t),*** depends both on the current input example ***x(t)*** as well as of its previous output ***y(t-1)*** (or activation states of hidden neutrons at ***t-1***)

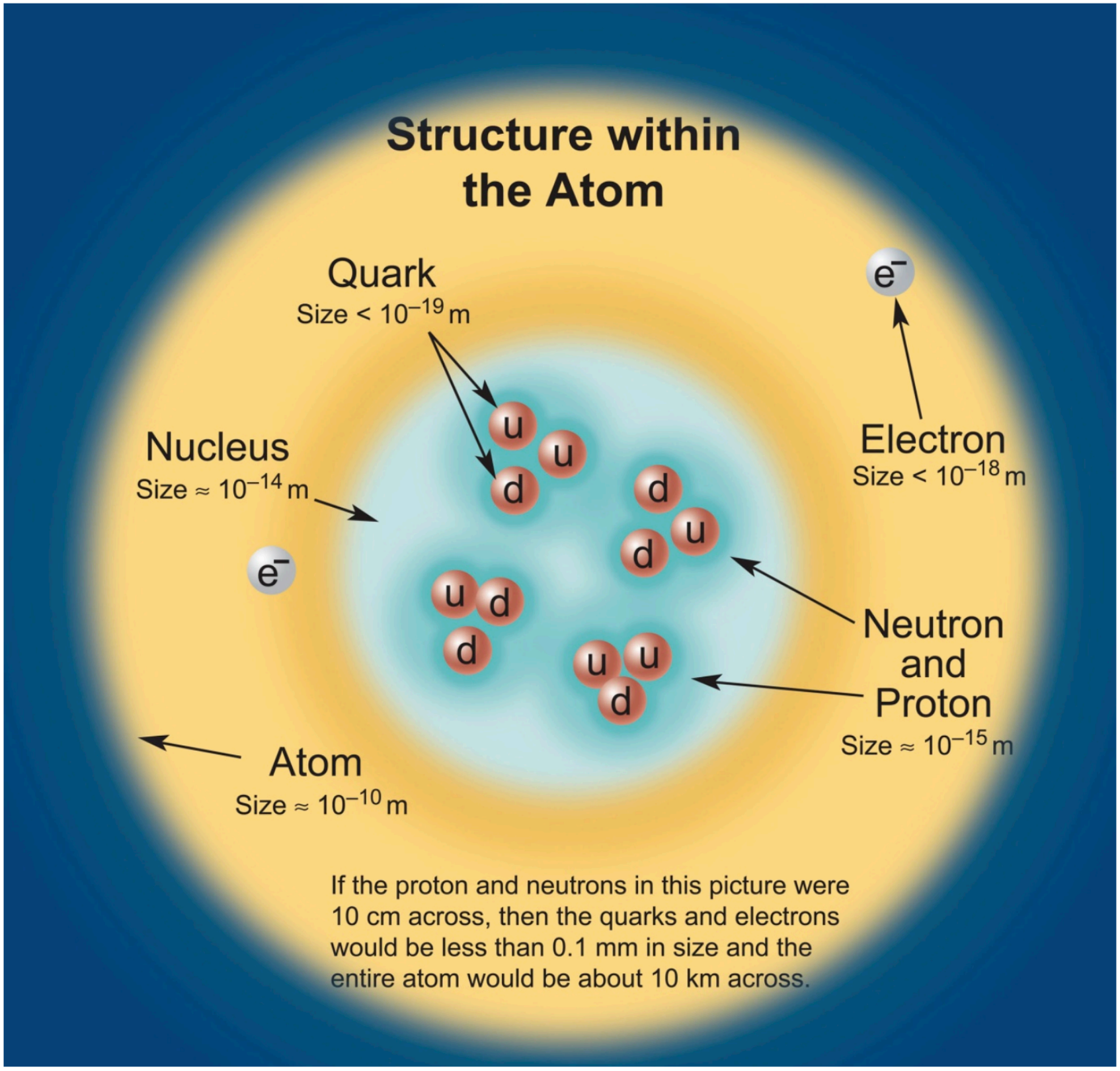
Recurrent Neural Networks

Lead to truly game-changer applications, such as **random generation of country song lyrics**

```
Tied right now
I got life now he never thought I got by the all
Going up like a house four boy
Nothing his thing out of hands
No one with the danger in the world
I love my black fire as I know
But the short knees just around me
Fun the heart couldnes fall to back
I see a rest of my wild missing far
When I was missing to wait
And if I think
It's a real tame
I say I belong is every long night
Maybe lovin' you
```

http://www.mattmoocar.me/blog/RNNCountryLyrics/

Investigating the microcosmos

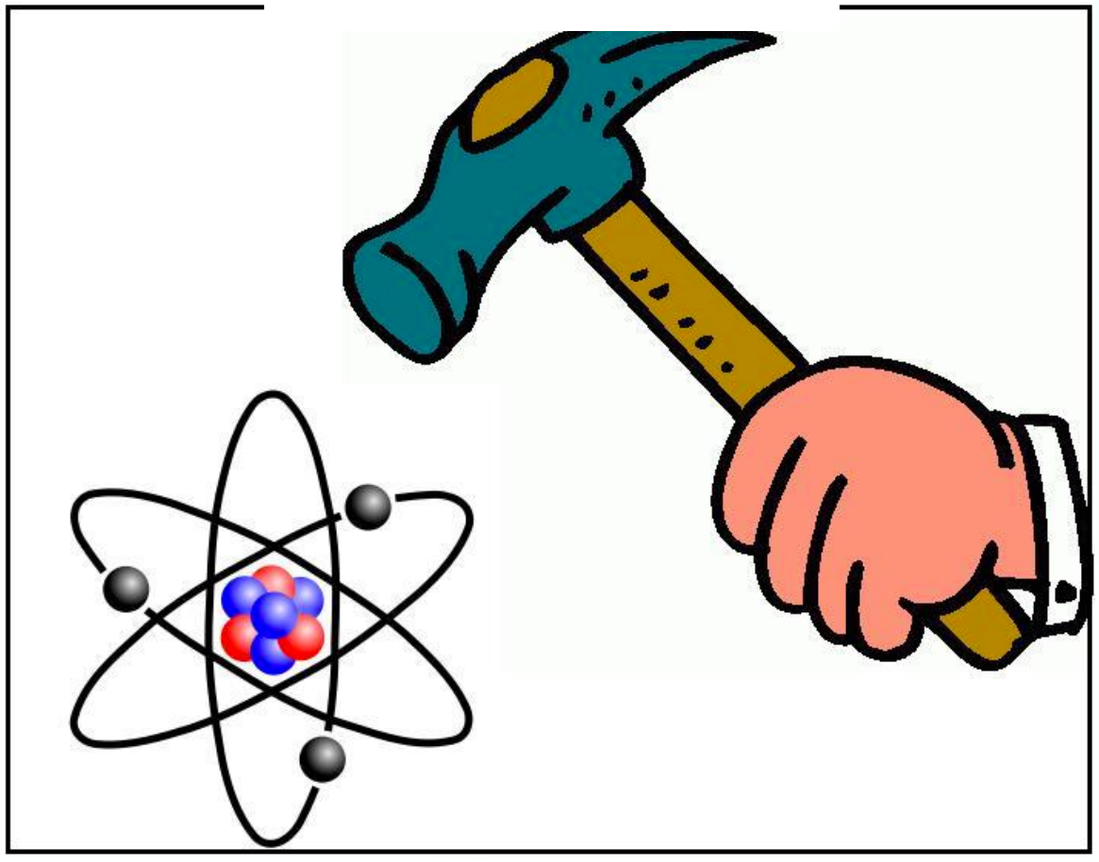


High energy colliders

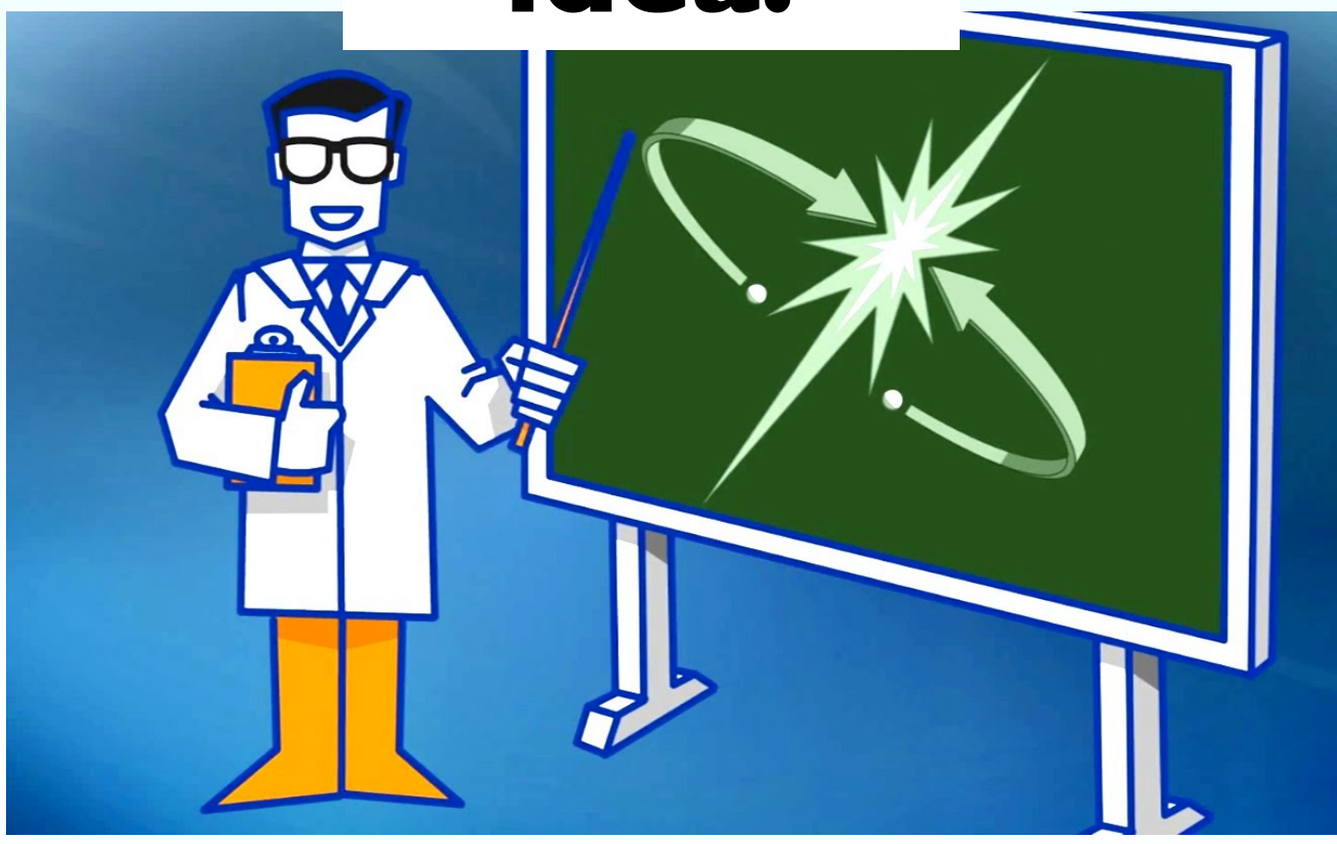
The idea behind **high-energy colliders** is very simple:

- We want to see **what is inside protons**: we need to **break them**. How we do this?
- We make protons **go very fast**, and then collide them: by looking at the r**esults of the collision,** we can understand the stuff protons are made of, if there are new particles or forces ….
- Since protons are very small, we need **extremely high energies to see inside them**

Bad idea!



Good idea!

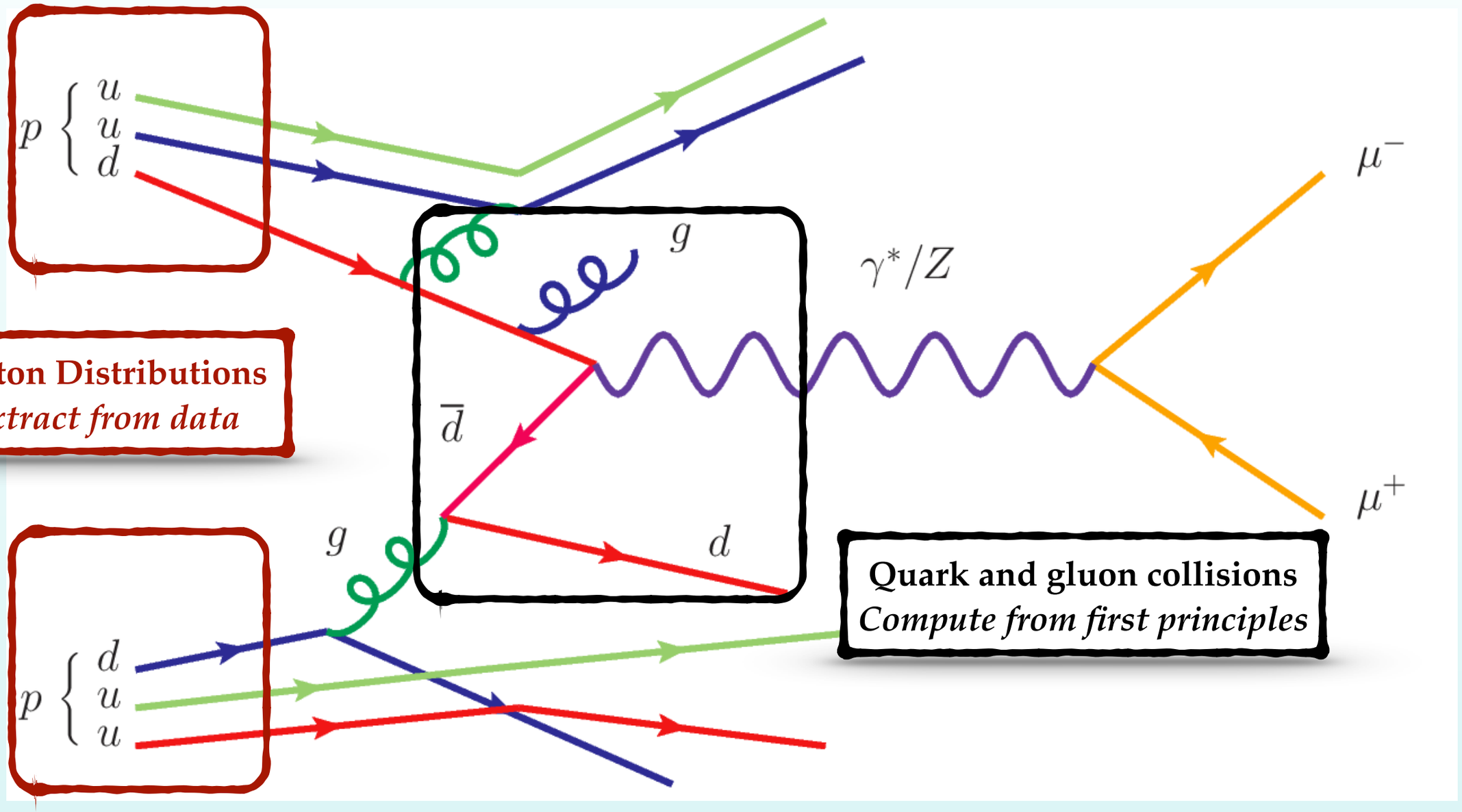


The Large Hadron Collider

- ☑ The LHC is the **most powerful particle accelerator ever build by mankind**
- ☑ Hosted by CERN in Geneva, is composed by a **27 km long tunnel** with **four gigantic detectors**

Anatomy of a proton-proton collision

High-energy **hadron colliders** involve **composite particles** (protons) with internal substructure (quarks and gluons): the LHC is actually a **quark and gluon collider!**



Calculations of **cross-sections** in hadron collisions require the combination of **perturbative cross-sections** with **non-perturbative parton distribution functions (PDFs)**

Parton Distributions

The distribution of energy that **quarks and gluons carry inside the proton** is quantified by the **Parton Distribution Functions (PDFs)**

$$g(x, Q)$$

Q: **Energy of the quark/gluon collision**
Inverse of the resolution length

g(x,Q): **Probability** of **finding a gluon inside a proton**, carrying a fraction *x* of the proton momentum, when probed with energy *Q*

x: **Fraction of the proton's momentum**

PDFs cannot be computed from first principles, and need to be **extracted from experimental data!**

ANNs as universal unbiased interpolants

ANNs provide **universal unbiased interpolants** to parametrise the non-perturbative dynamics that determines the **size and shape of the PDFs** from experimental data

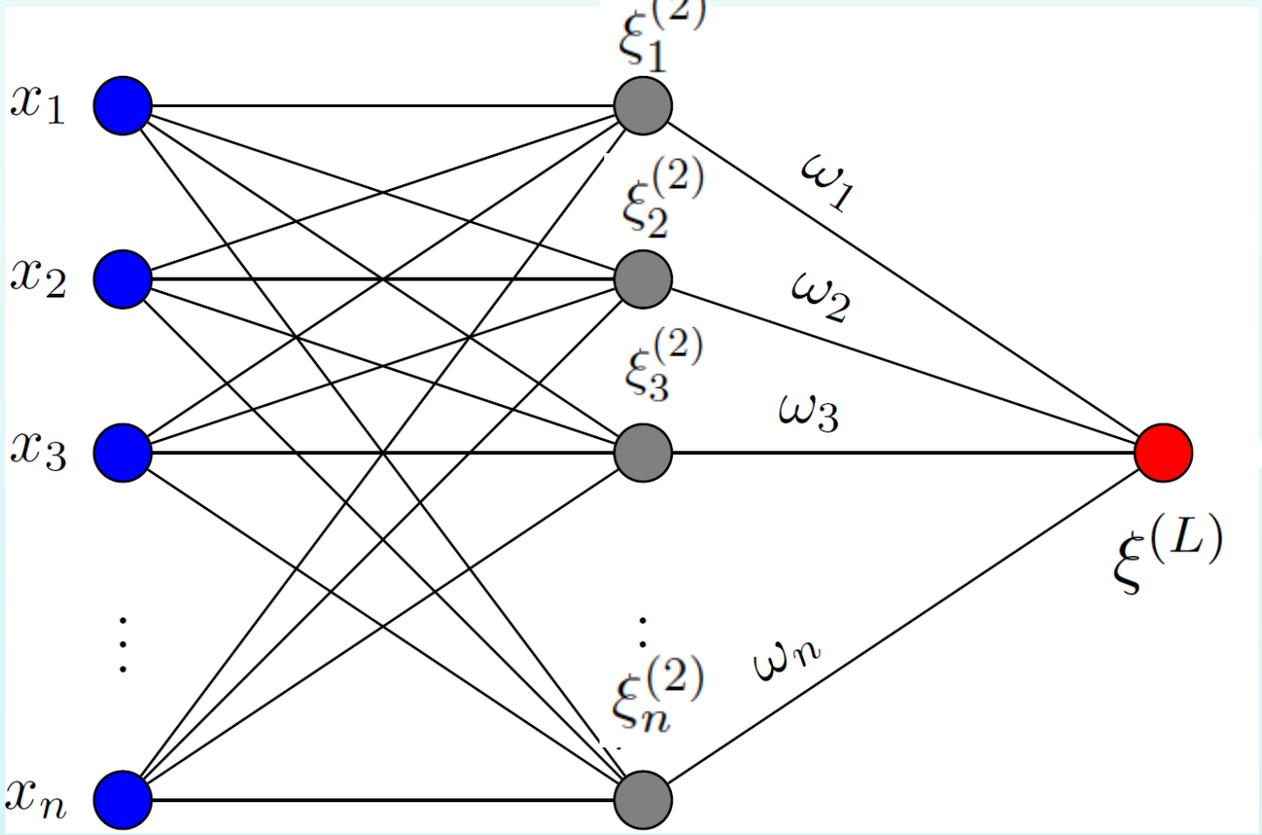
ad-hoc ansatz

Traditional approach

$$g(x,Q_0) = A_g(1-x)^{a_g}x^{-b_g}\left(1+c_g\sqrt{s}+d_g x+\ldots\right)$$

NNPDF approach

$$g(x,Q_0) = A_g \mathrm{ANN}_g(x)$$



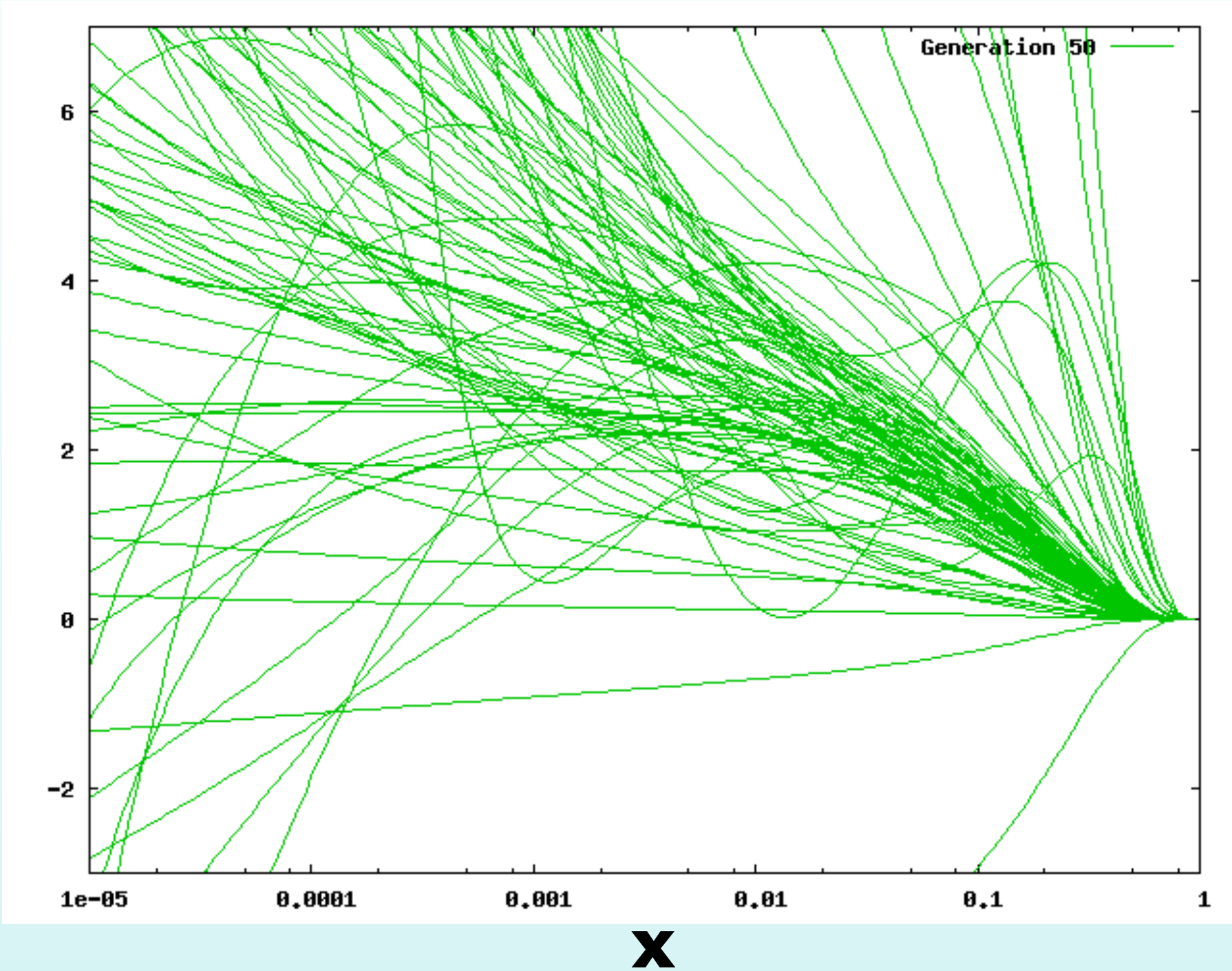
$$\mathrm{ANN}_g(x) = \xi^{(L)} = \mathcal{F}\left[\xi^{(1)},\{\omega_{ij}^{(l)}\},\{\theta_i^{(l)}\}\right]$$

$$\xi_i^{(l)} = g\left(\sum_{j=1}^{n_{l-1}}\omega_{ij}^{(l-1)}\xi_j^{(l-1)}-\theta_i^{(l)}\right)$$

- ANNs eliminate **theory bias** introduced in PDF fits from choice of *ad-hoc* functional forms
- NNPDF fits used **O(400) free parameters**, to be compared with O(10-20) in traditional PDFs. Results stable if **O(4000) parameters used!**

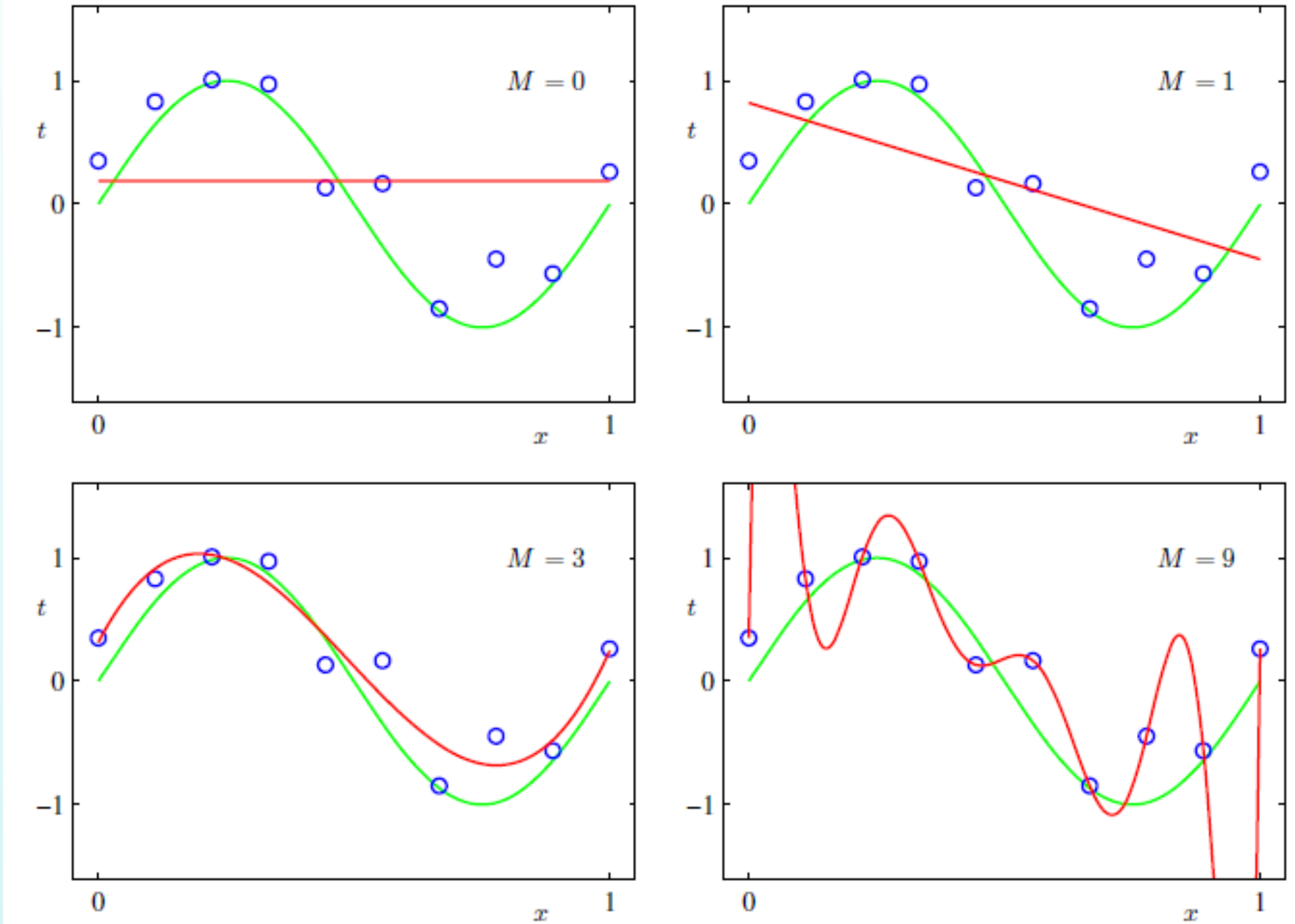
PDF Replica Neural Network Learning

The training of the Neural Networks is performed using **Genetic Algorithms**
Each **curve** corresponds to a **possible functional form** for the gluon PDF



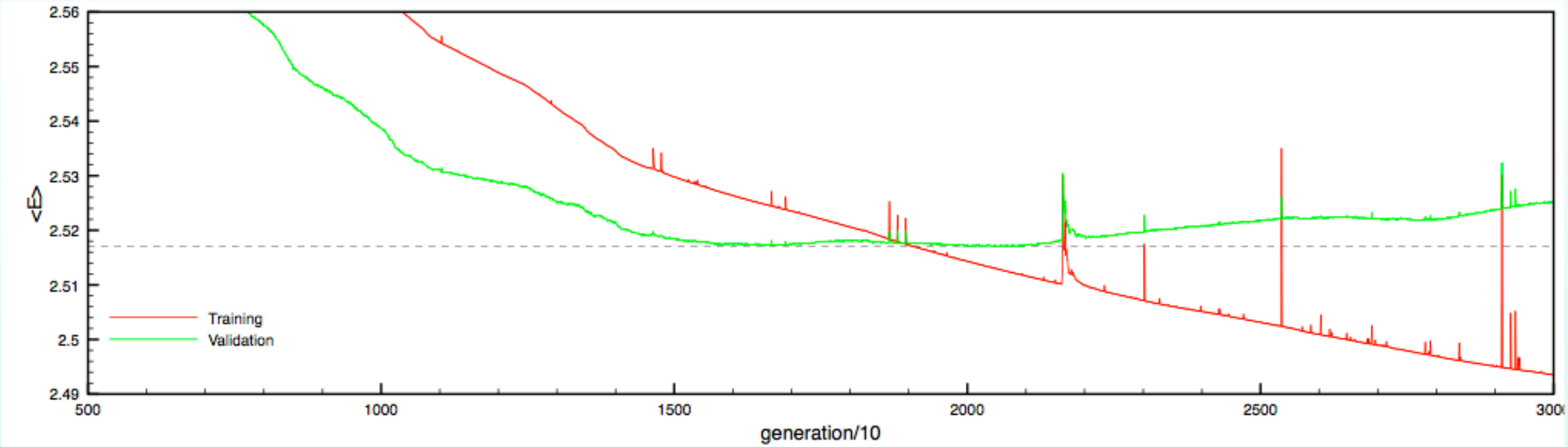
Avoiding overfitting

For a **flexible enough input functional form for the Parton Distributions**,
one might end up **fitting statistical fluctuations** rather than the underlying physical law!



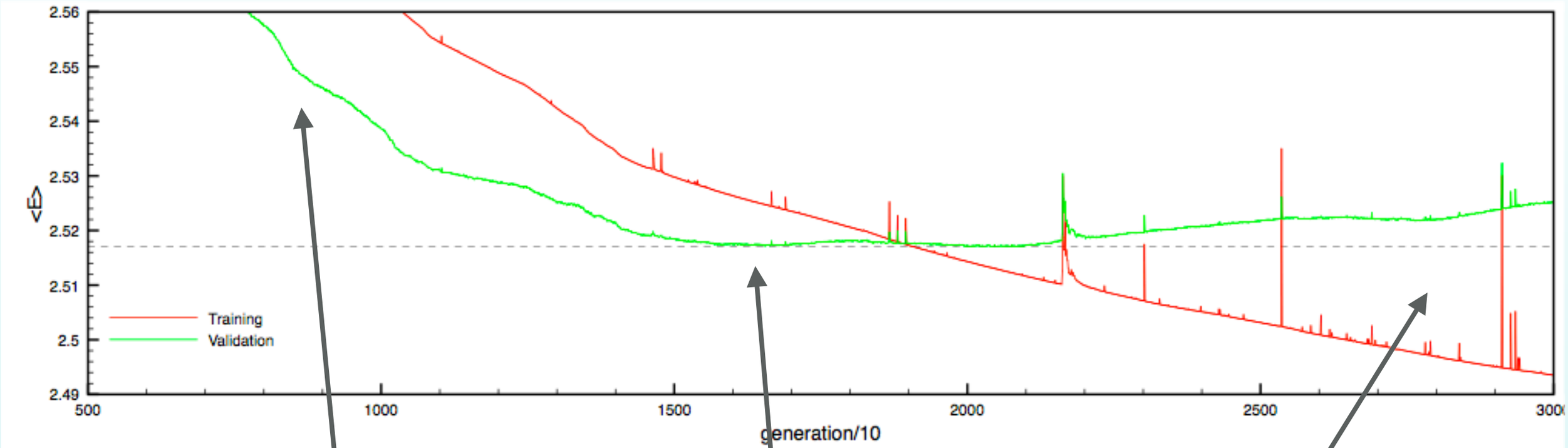
Avoiding overfitting

- Separate the input measurements into a **training** and a **validation** sample



Avoiding overfitting

- Separate the input measurements into a **training** and a **validation** sample
- The validation sample is never trained, only used to monitor the quality of the fit to the training sample
- The optimal stopping point is at the **global minimum of the validation χ^2**



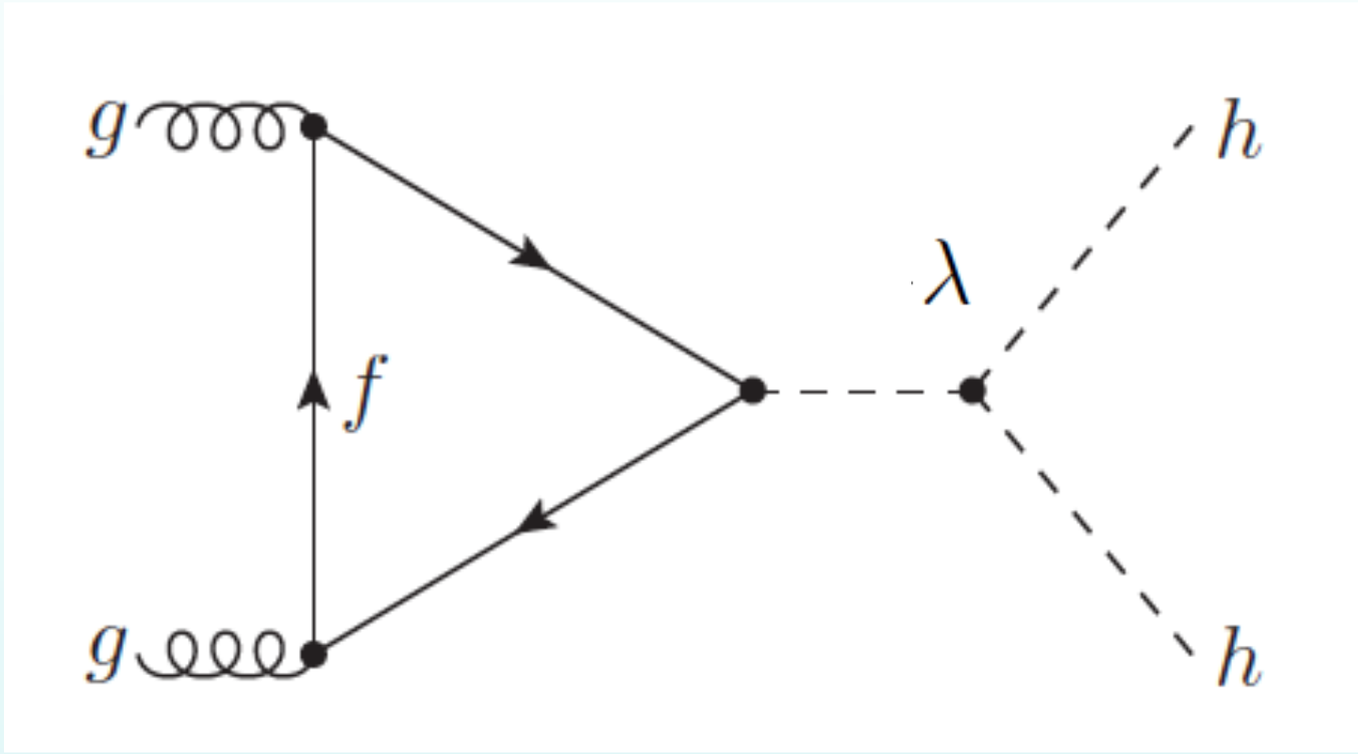
Undertrained

Optimal stopping point

Overtrained

Probing Electroweak Symmetry breaking

- **Current measurements** (couplings in single Higgs production) probe **Higgs potential close to minimum**
- Double Higgs production essential to **reconstruct the full Higgs potential** and clarify EWSB mechanism
- The Higgs potential is *ad-hoc*: **many other EWSB mechanisms conceivable**



Higgs mechanism **Coleman-Weinberg mechanism**

$$V(h) = m_h^2 h^\dagger h + \frac{1}{2}\lambda (h^\dagger h)^2$$

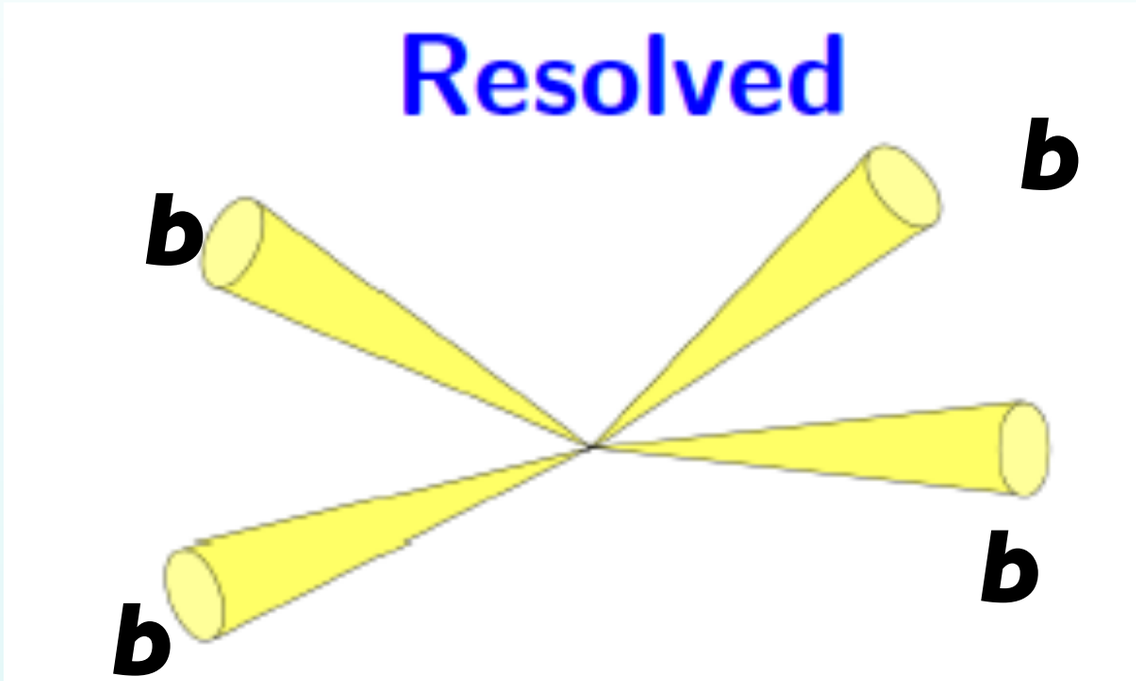
$$V(h) \to \frac{1}{2}\lambda (h^\dagger h)^2 \log\left[\frac{(h^\dagger h)}{m^2}\right]$$

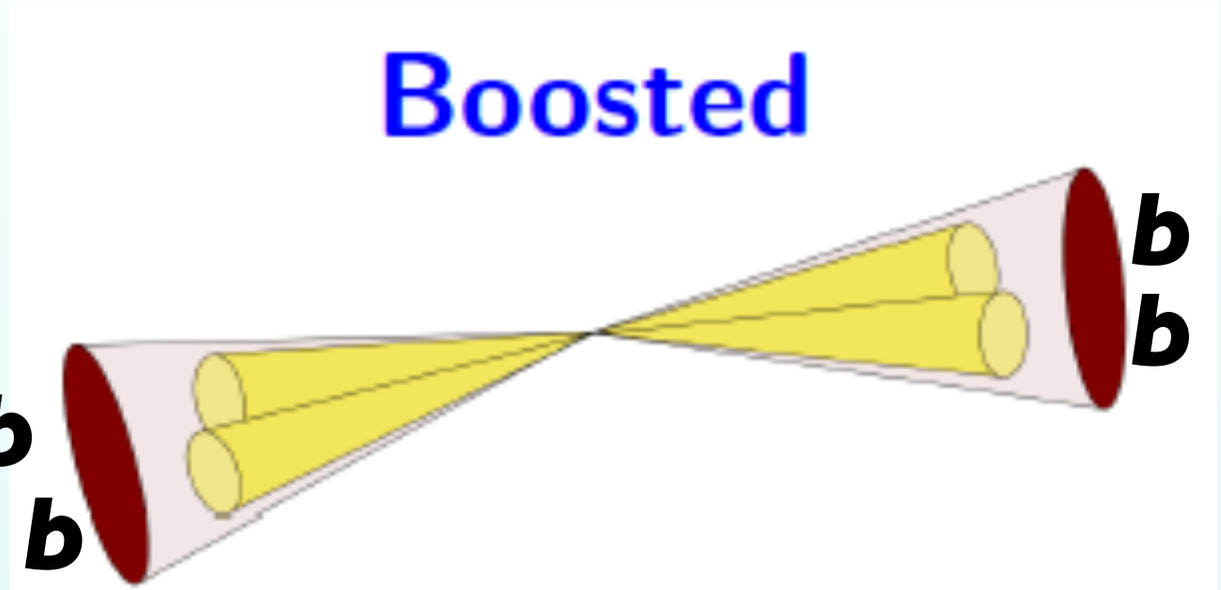
Each possibility associated to **completely different EWSB mechanism**, with crucial implications for the **hierarchy problem**, the structure of quantum field theory, and **New Physics at the EW scale**

Arkani-Hamed, Han, Mangano, Wang, arxiv:1511.06495

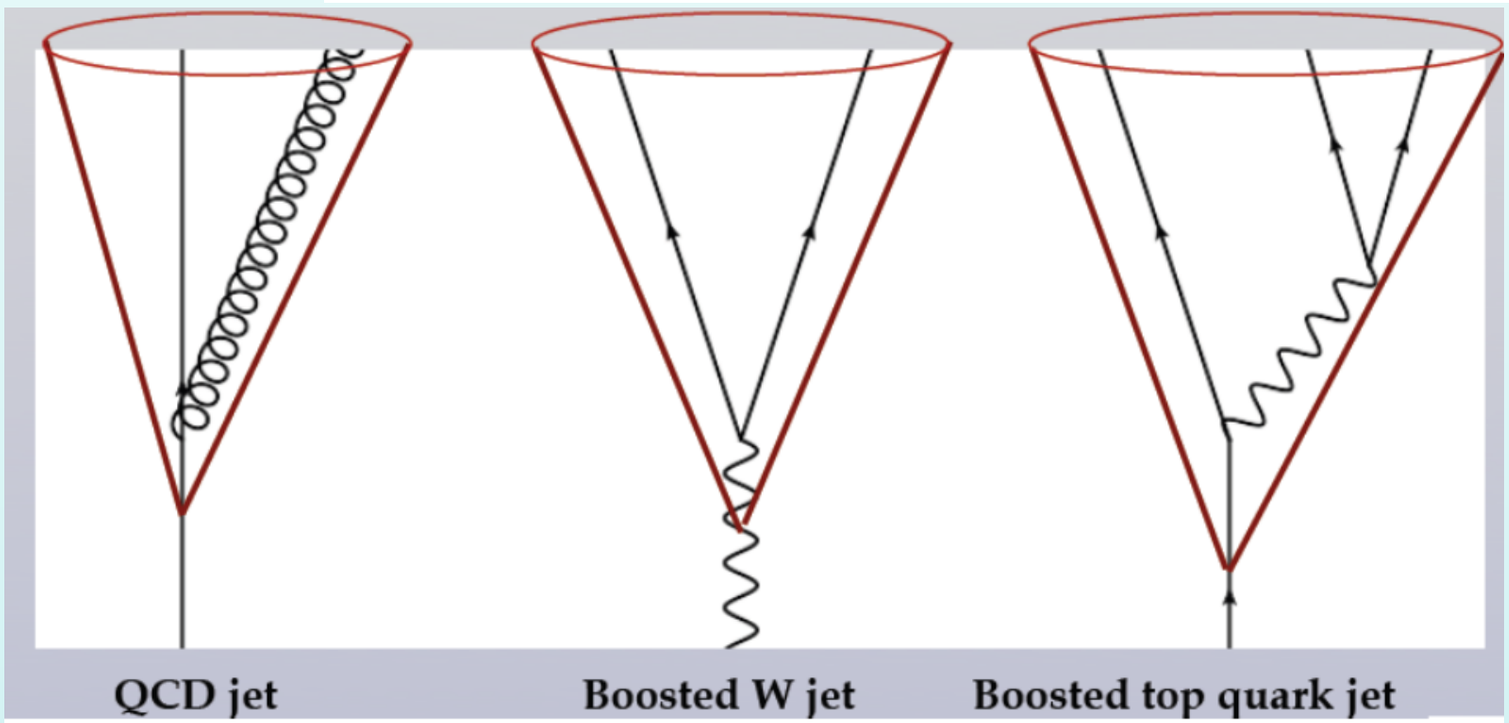
hh->bbbb: selection strategy

- Exploit **4b final state**: highest signal yields, but **overwhelming QCD background** (by orders of magnitude!)
- Carefully chosen selection strategies ensure that **all relevant event topologies can be reconstructed**

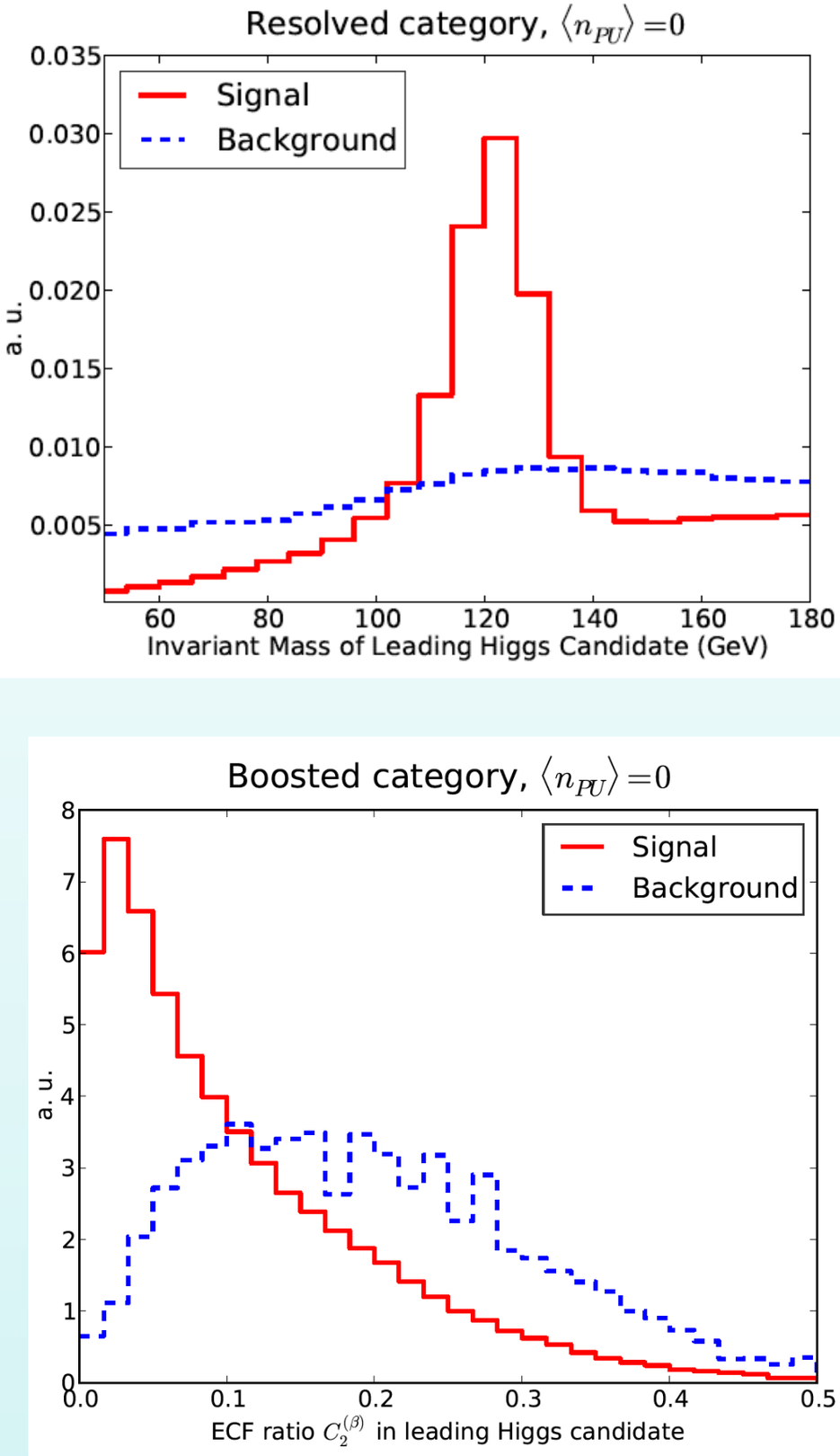


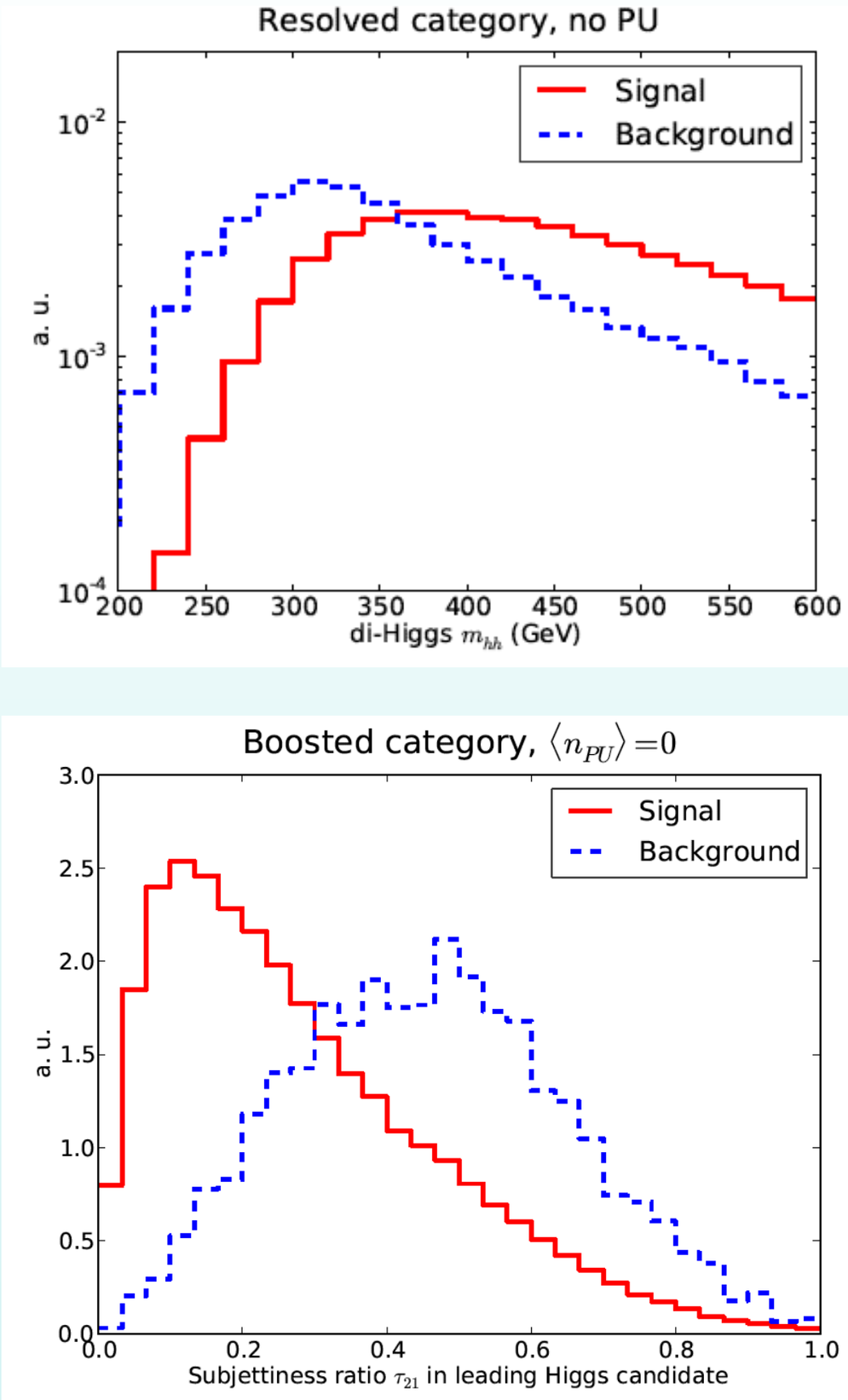


Recent progress in **jet substructure** techniques important to reduced QCD background in the **boosted regime**

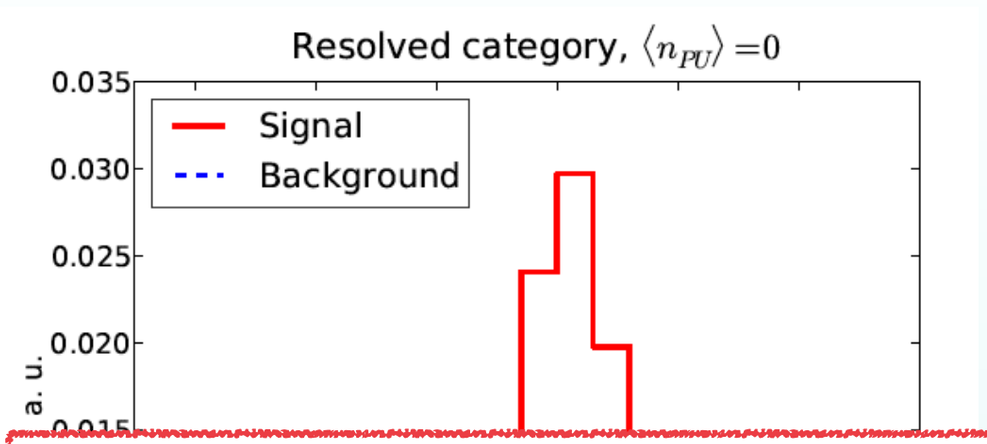


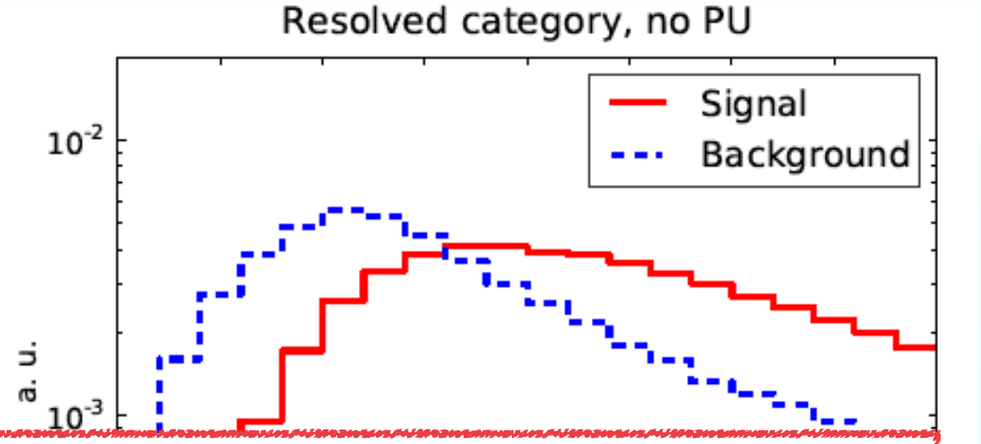
di-Higgs kinematic distributions





di-Higgs kinematic distributions

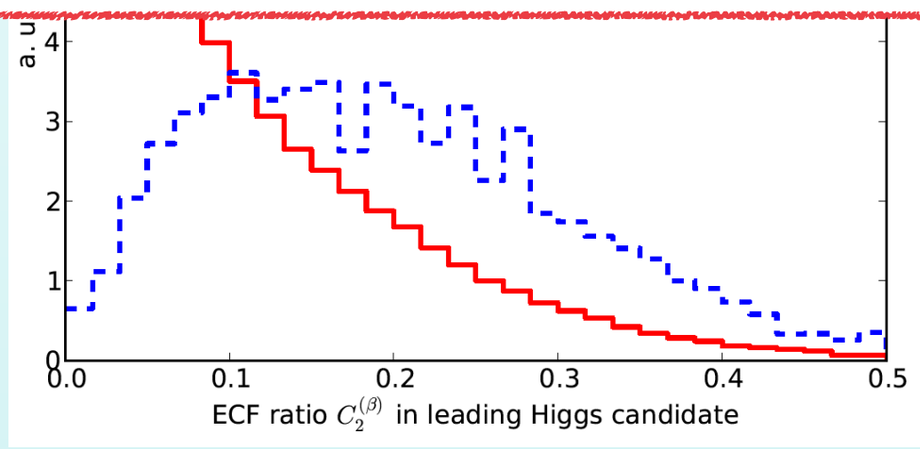


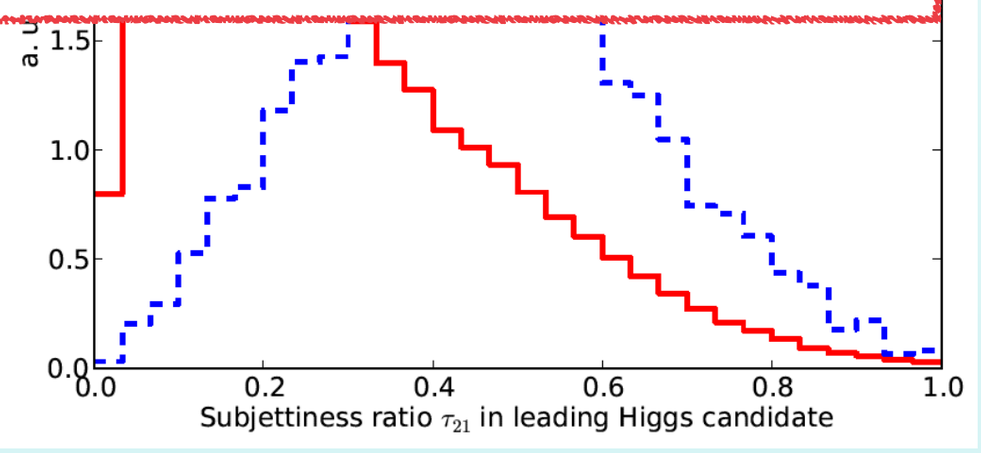


Many kinematic variables can be used to **disentangle signal and background**

How do we select which ones to use? And the optical cuts? And the cross-correlations among variables?

We don't need to! Use **ML methods to identify automatically** the combination of kinematical variables with the highest discrimination power!





Multivariate techniques

Input

Higgs pT

Higgs m

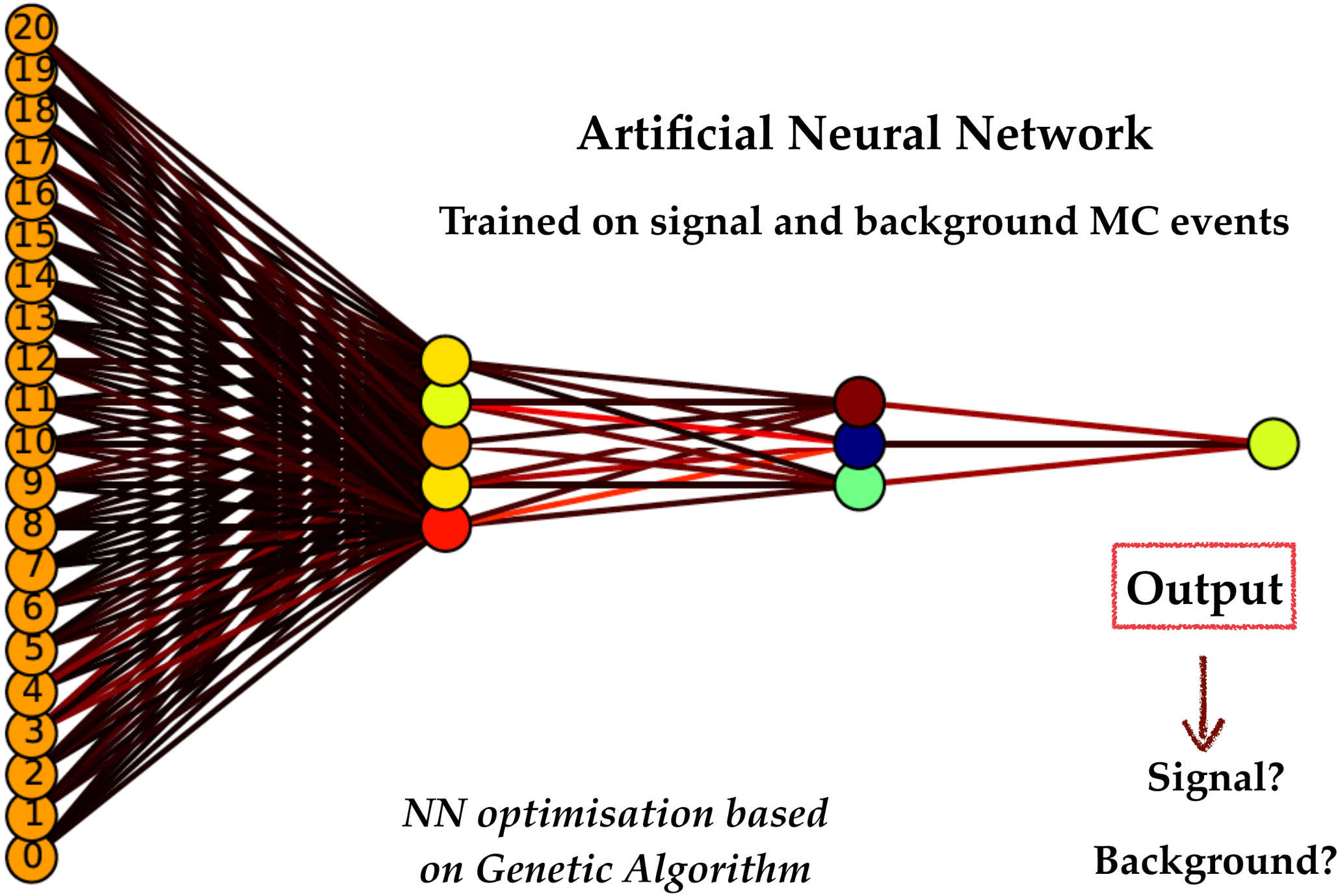
di-Higgs m

ECF

τ_{12}

Subjet pT

…..



Artificial Neural Network

Trained on signal and background MC events

NN optimisation based on Genetic Algorithm

Output

Signal?

Background?

Caveat: in a measurement, training of classifier should be done on ***real data based on control regions***

Multivariate techniques

The optimisation of the classifier is based on the minimisation of the **cross-entropy function**

Number of MC events used for the training

$$E(\{\omega\}) \equiv -\log\left(\prod_{i}^{N_{\mathrm{ev}}} P(y'_i|\{k\}_i, \{\omega\})\right)$$

$$= \sum_{i}^{N_{\mathrm{ev}}} \left[y'_i \log y_i + (1 - y'_i) \log (1 - y_i)\right]$$

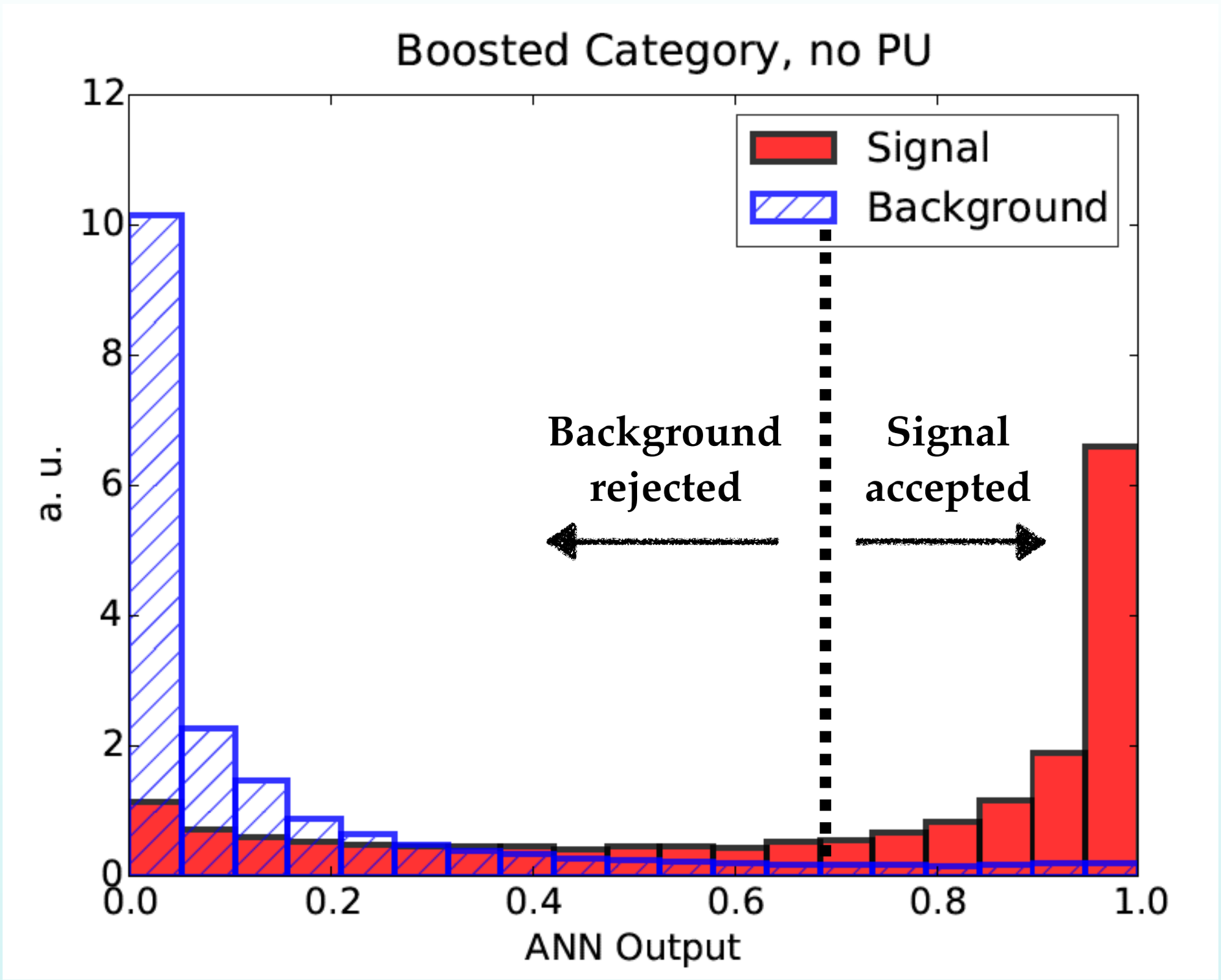
True classification of event: y'_i=0 for background, y'_i=1 for signal

Probability that the event i originates from signal process

aims to achieve the **best possible separation** between signal and background events

Multivariate techniques

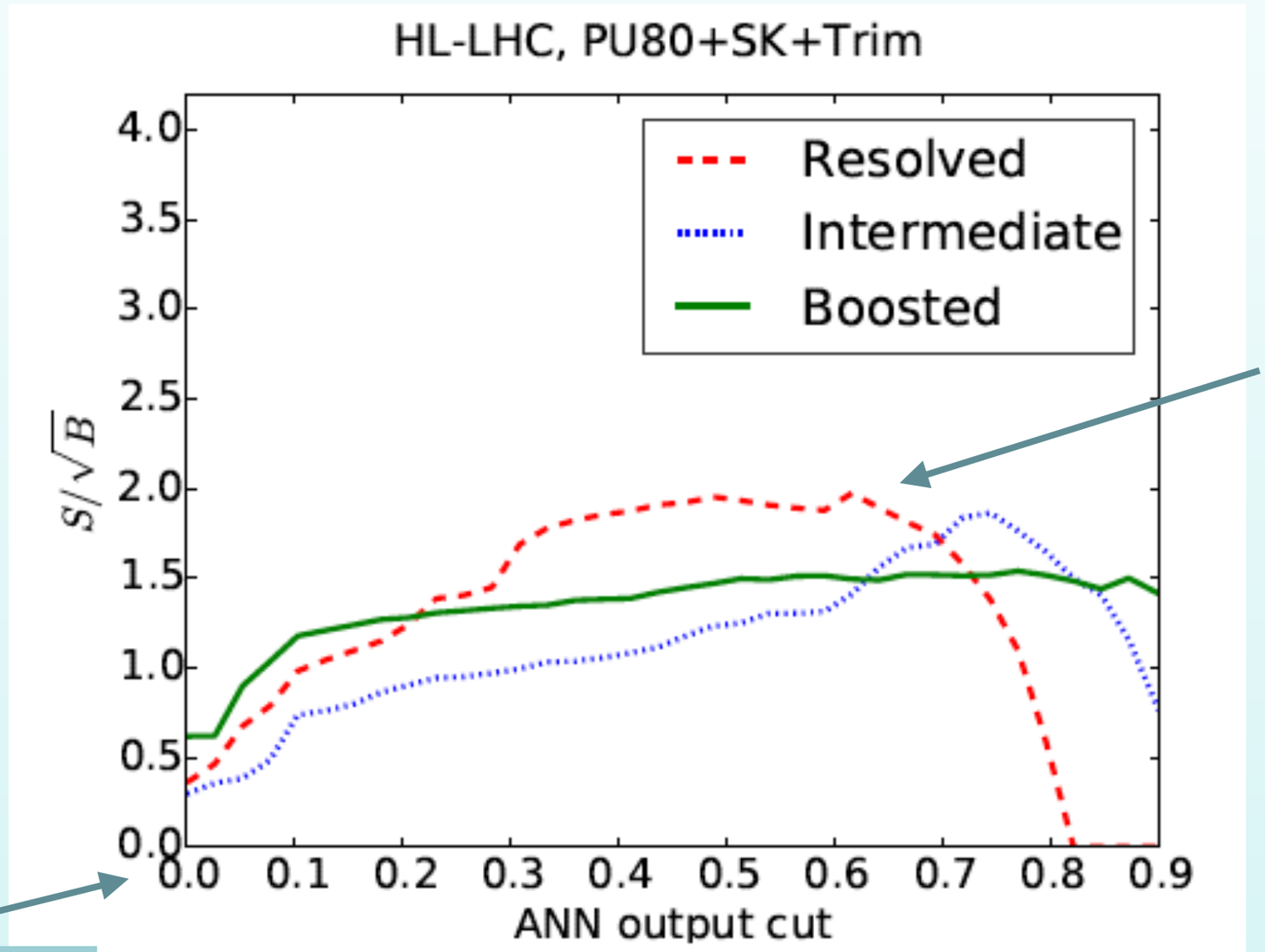
Combining information from all kinematic variables in MVA: **excellent signal/background discrimination**



Discovering Higgs self-interactions

ML techniques allow to **substantially improve the signal significance** for this process **observe Higgs pair production in the 4b final state** at the HL-LHC. Observation (maybe discovery) within reach!

$$\left(\frac{S}{\sqrt{B_{4b}}}\right)_{\text{tot}} \simeq 4.7\ (1.5)\,, \quad \mathcal{L} = 3000\ (300)\,\text{fb}^{-1}$$

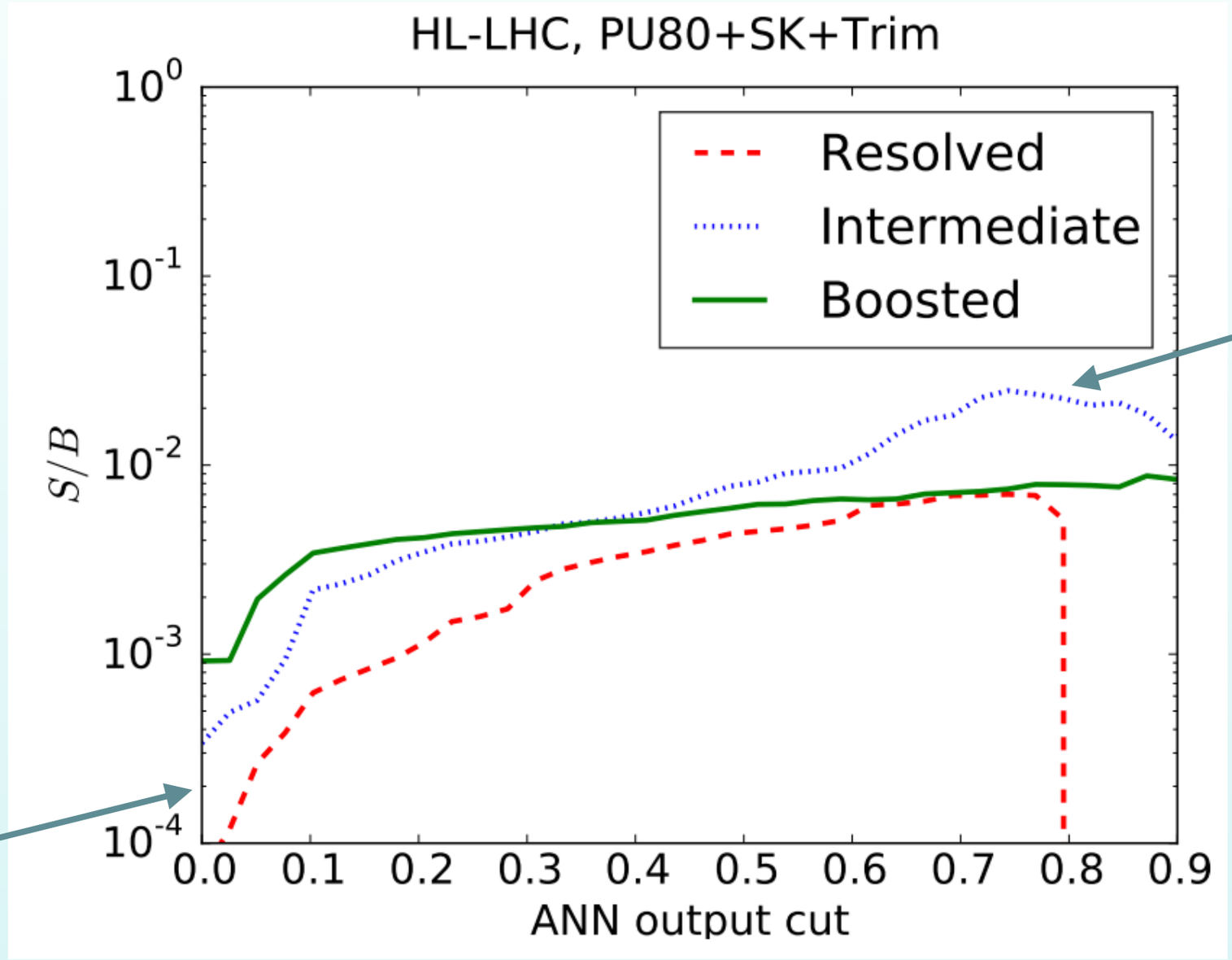


Post MVA

Pre-MVA

Discovering Higgs self-interactions

ML techniques allow to **substantially improve the signal significance** for this process **observe Higgs pair production in the 4b final state** at the HL-LHC. Observation (maybe discovery) within reach!



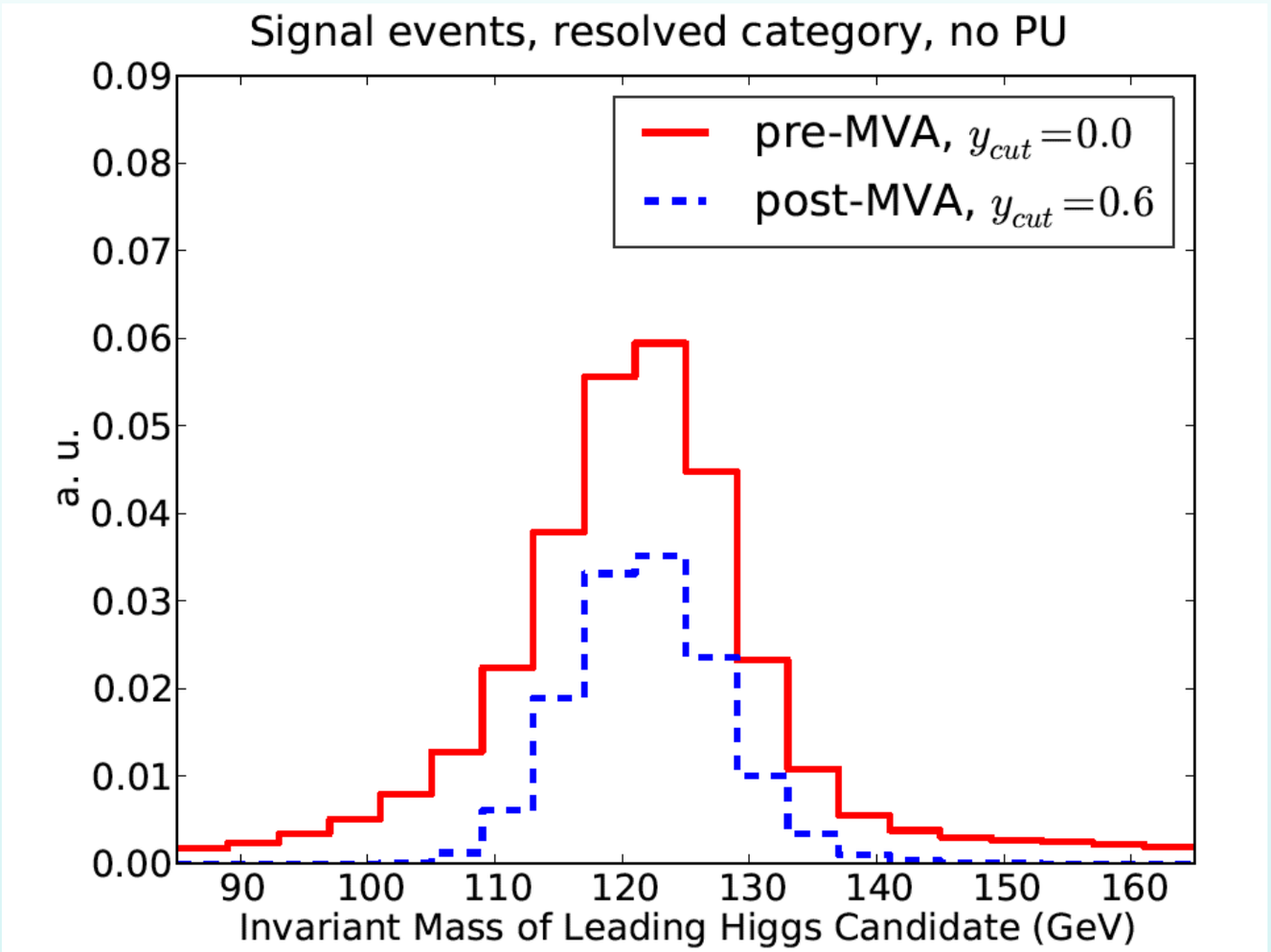
Post MVA

Pre-MVA

Need to ensure also a high enough signal/background ratio, else experimental systematic errors would kill the signal significance

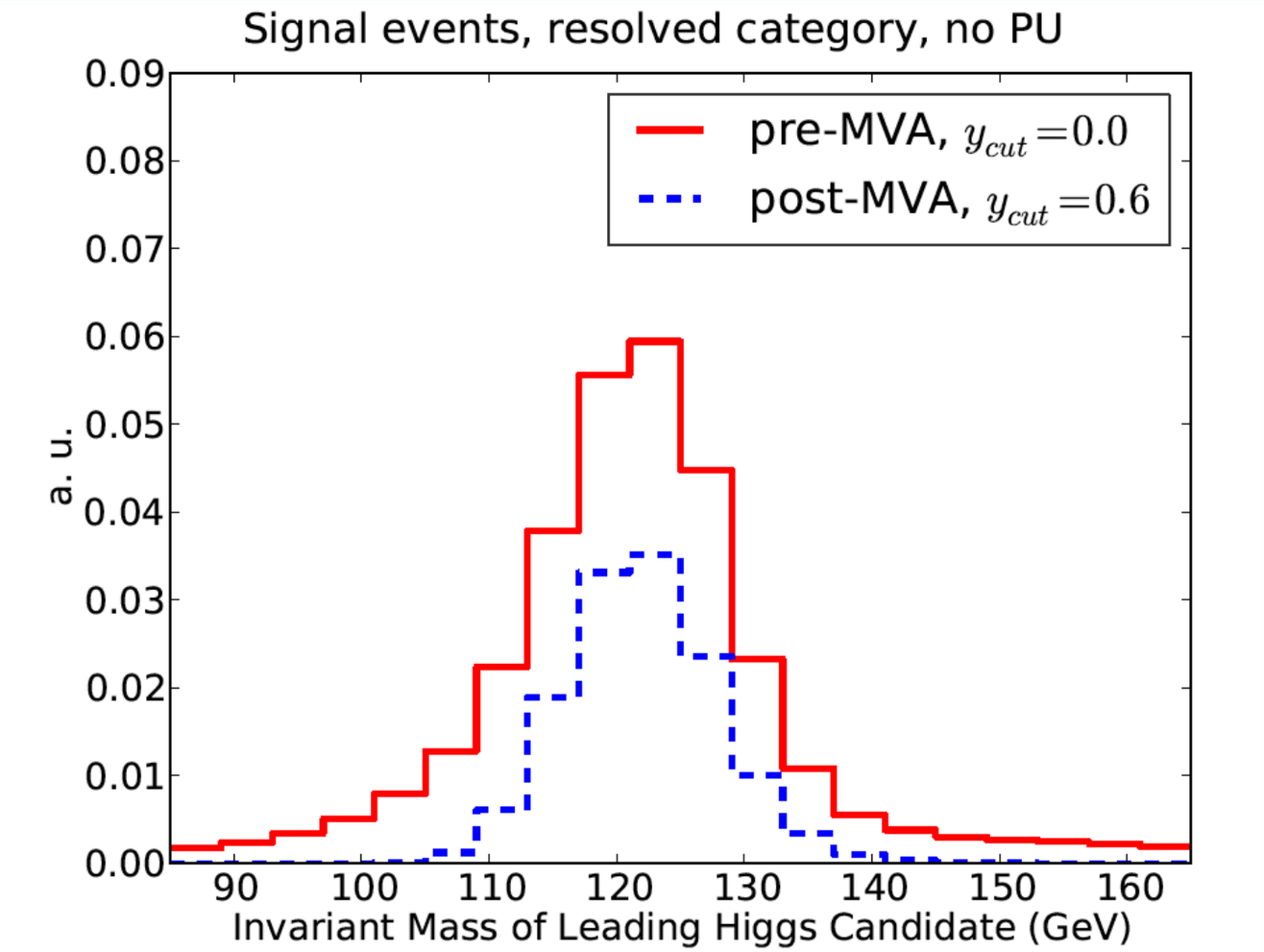
Opening the Black Box

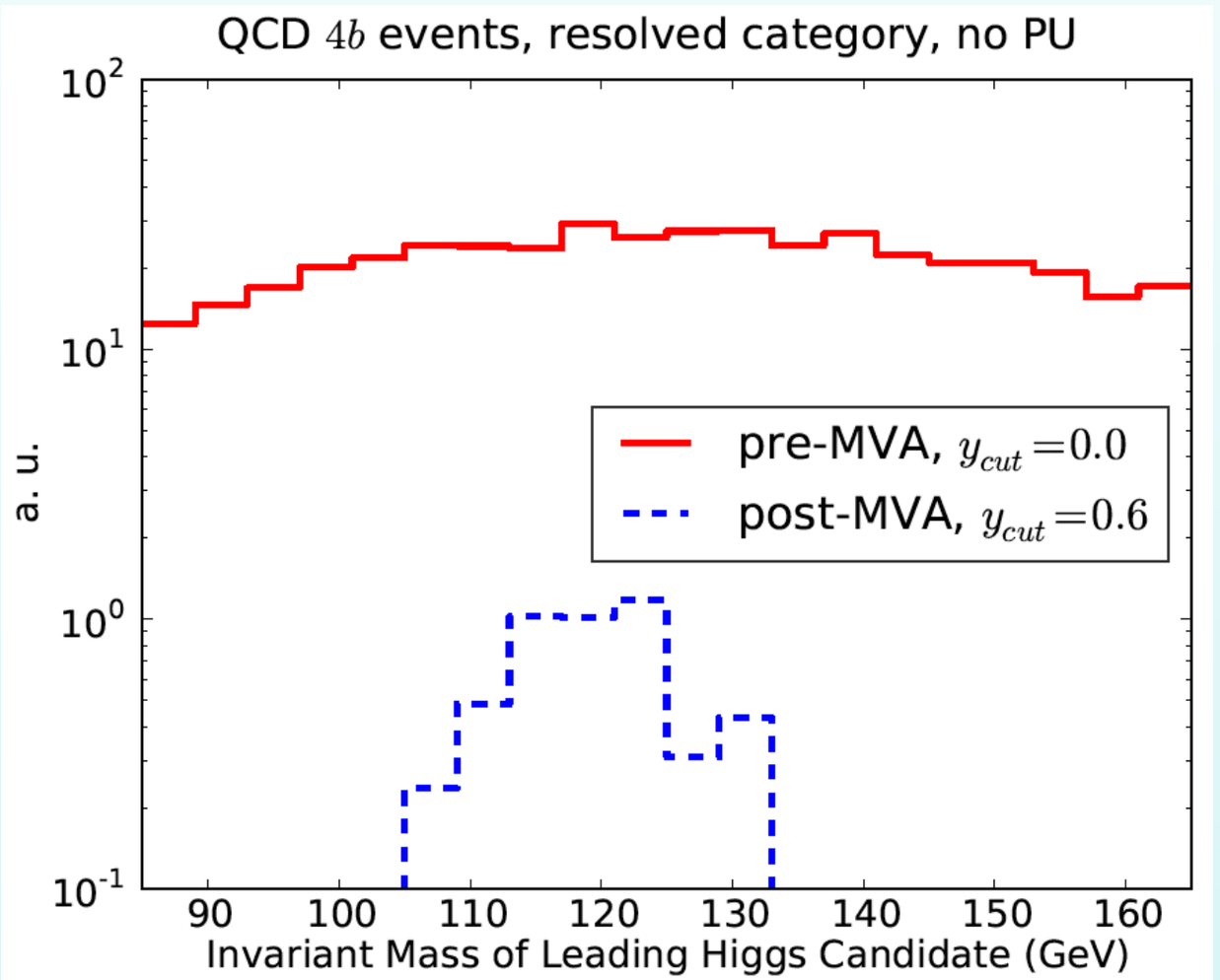
- ANNs are sometimes criticised by being **black boxes**, with little understanding of what happens inside them
- But ANNs are simply a **set of combined kinematical cuts**, nothing mysterious in them!
- Kinematic distributions **after and before the ANN cut** allow determining the **effective kinematic cuts** being optimised by the MVA, which would allow a cut-based analysis



Opening the Black Box

- ANNs are sometimes criticised by being **black boxes**, with little understanding of what happens inside them
- But ANNs are simply a **set of combined kinematical cuts**, nothing mysterious in them!
- Kinematic distributions **after and before the ANN cut** allow determining the **effective kinematic cuts** being optimised by the MVA, which would allow a cut-based analysis





The MVA sculpts a Higgs peak in the QCD background!

Take-away message

Proficiency in Machine Learning applications requires a deep understanding of both the physical problem being addressed as well as of the inner workings of the specific algorithms used!

ANNs and LHC phenomenology

- **Machine Learning algorithms** are already **transforming our world**, from the way we move, shop and heal ourselves, to our understanding of what makes us unique as humans

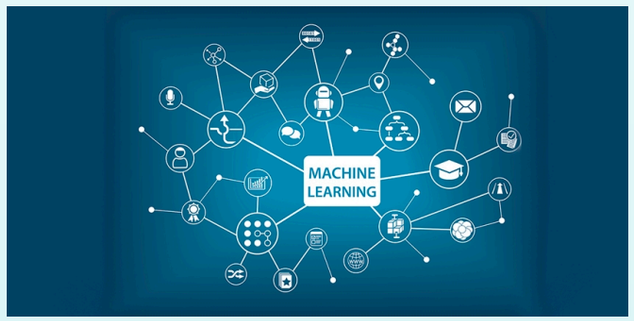
- In the context of **High Energy Physics, ML tools are ubiquitous**, from event selection deep in the detector chain (triggering) to bottom-quark tagging and automated BSM models classification (and exclusion)

- Avoid using ML tools as black boxes: a detailed understanding of both the **physical and the algorithmic aspects of the problem is essential**

Higgs self-interactions

QCD-aware NNs For jet physics

The structure of the proton at the LHC

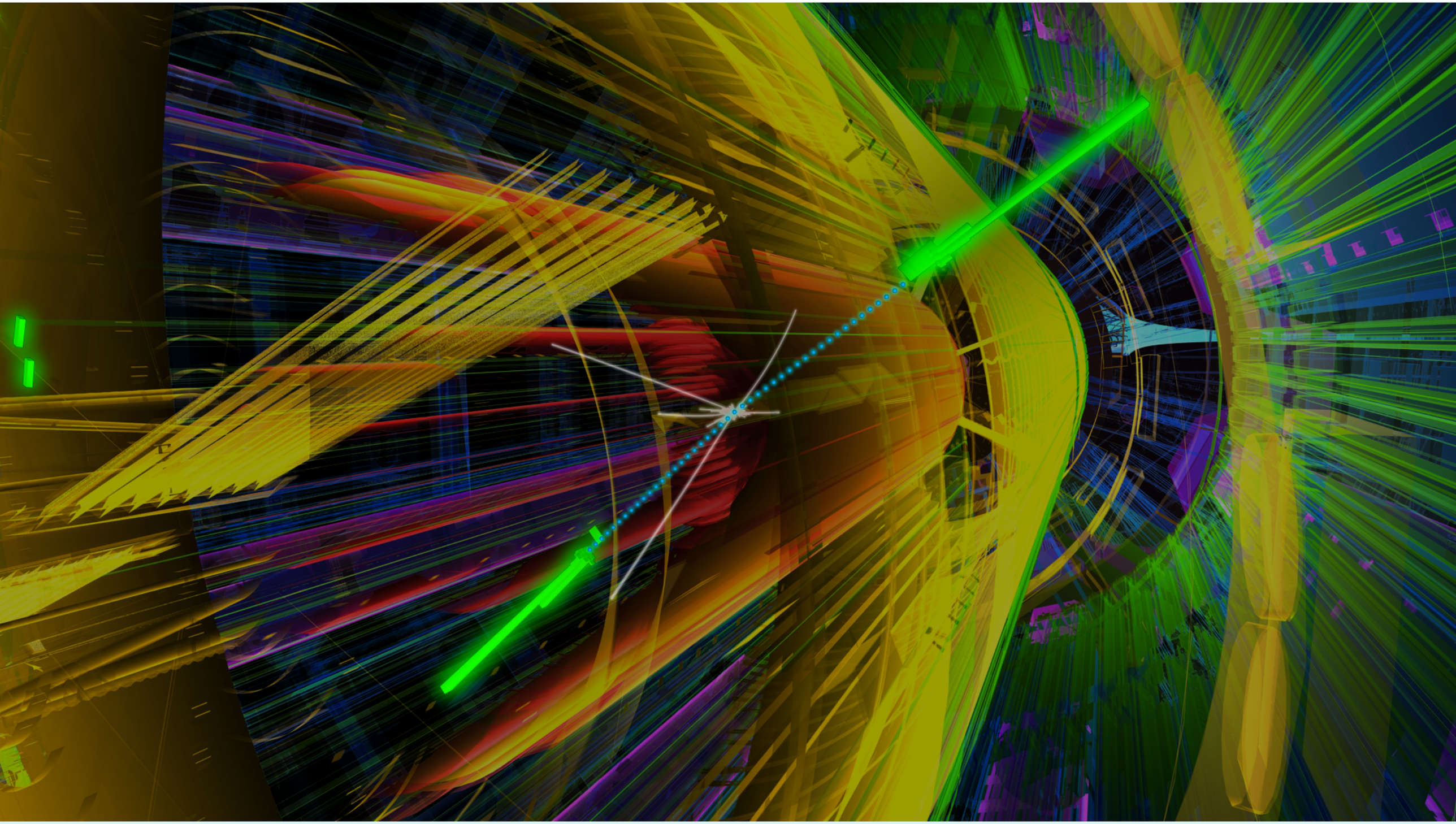


Automated bSM exclusion limits

Boosting bSM searches

HEP detector simulation

Fascinating times ahead at the high-energy frontier!



Ready to be exploited with our Machine Learning toolbox!

Fascinating times ahead at the high-energy frontier!

Ready to be exploited with our Machine Learning toolbox!