Deep Learning in Particle Physics Research

winnings 4 : 1

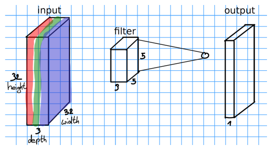
Prof. Dr. Martin Erdmann, RWTH Aachen University, Rutherford Appleton Laboratory 17-Jan-2018

Topics

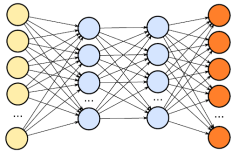
Deep neural network

from computer science to particle physics

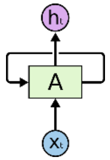
Network concepts and applications



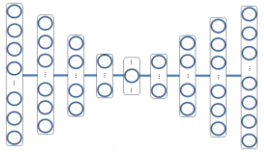
convolutional



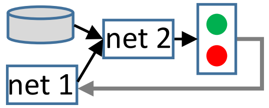
fully connected



recurrent



autoencoder



adversarial

Network quality

causality, uncertainty

Deep Learning in industry

humans

machine

SegNet University of Cambridge

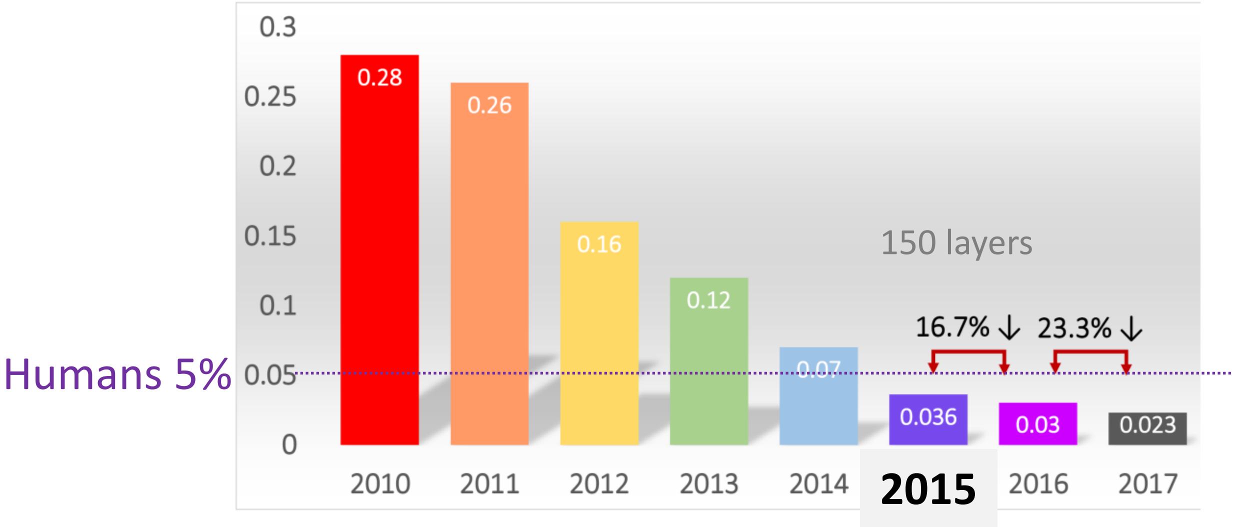
O. Russakovsky et al, arXiv:1409.0575; K. He, X. Zhang, S. Ren, J. Sunar, arXiv:1512.03385
WMW Jie Hu, Li Shen (Oxford), Gang Sun, 2017

Spectacular success

Image recognition challenge

ImageNet: 1000 categories, 1.2 million images

Classification error rate

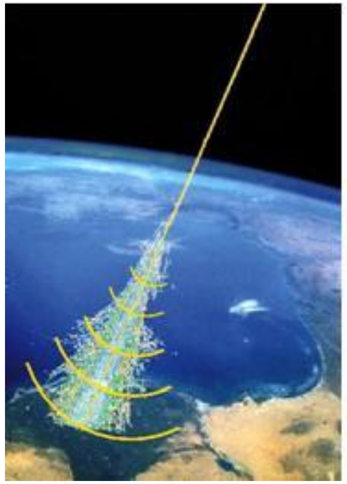


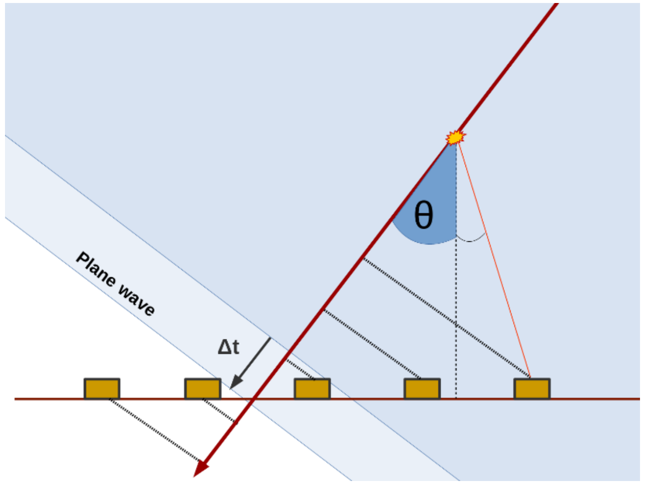
Deep learning errors < humans

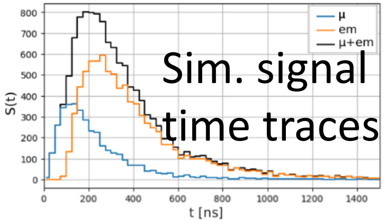
New paradigm in physics education & work ?

- Up to now: **physics laws & mathematics** are basis for algorithm development → machine learning with physicist's favorite observables
- New: **machines exploit data deeper** than physicist's algorithms so far
- **We ought to prepare for the fundamental change to include machines in our daily work**
- Next: develop **physics motivated network architectures**

Calorimeter: cosmic ray induced air showers



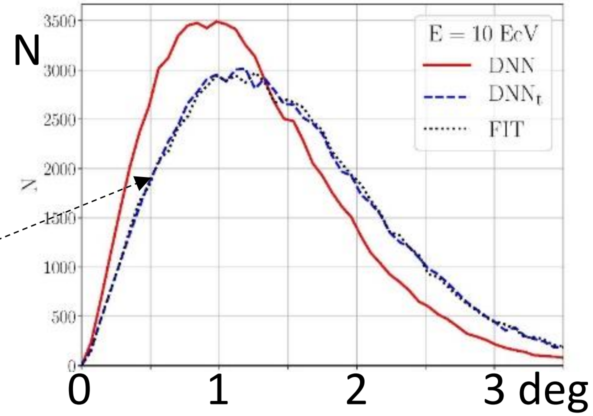




Educated physicist:

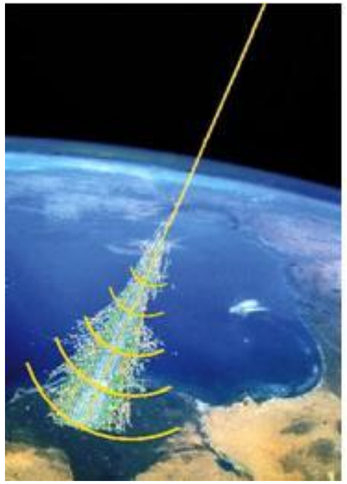
- Time offset between detectors
- Particle velocity
- Detector distance
- Plane fit of time residuals

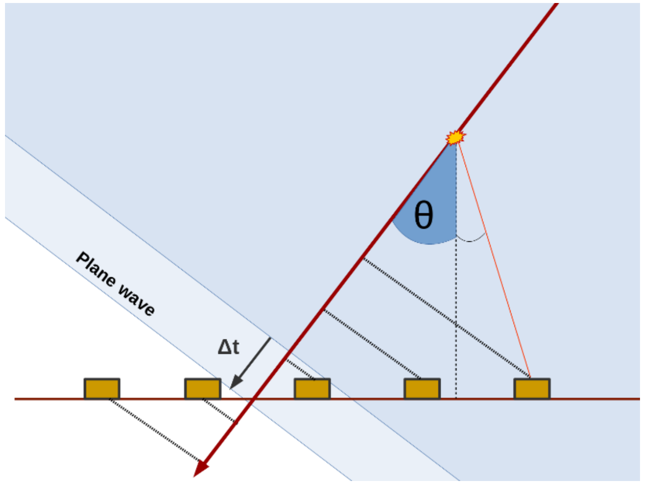
shower direction angular resolution

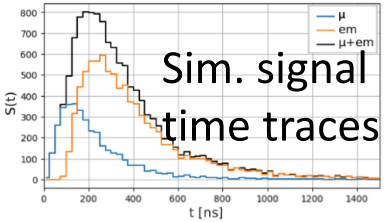


M.E., J. Glombitza, D. Walz, 10.1016/j.astropartphys.2017.10.006

Calorimeter: cosmic ray induced air showers





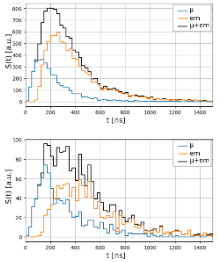


Sim. signal time traces

Educated physicist:

- Time offset between detectors
- Particle velocity
- Detector distance
- Plane fit of time residuals

RAW input data:



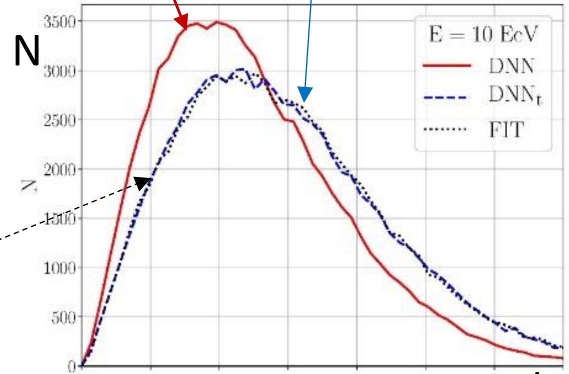
…

Deep Neural Network

- Time offset only
- Black signal traces added

shower direction angular resolution

N

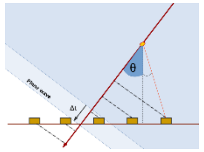


0 1 2 3 deg

Deep Neural Network learns physics from data within 3h

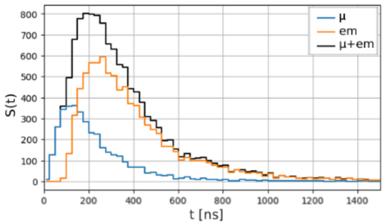
M.E., J. Glombitza, D. Walz, 10.1016/j.astropartphys.2017.10.006

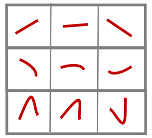
Calorimeter: cosmic ray induced air showers



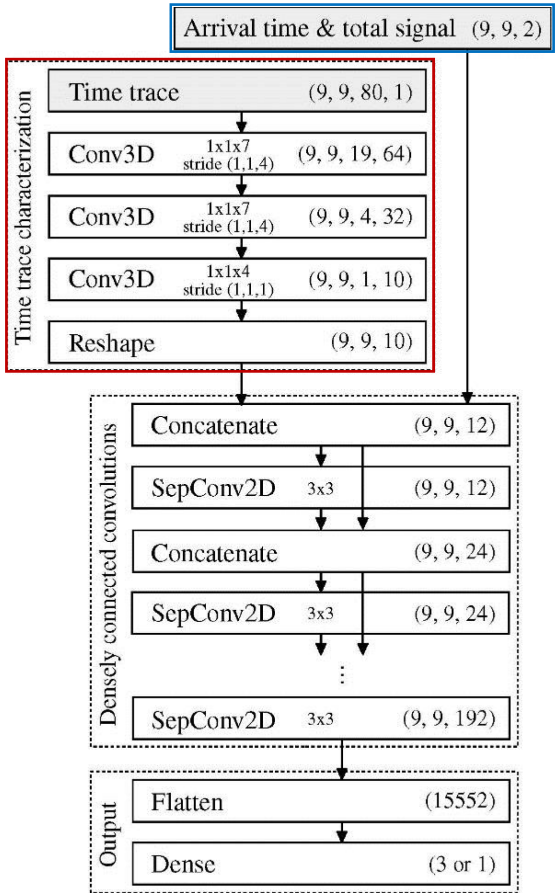
Time offset & total signal

Characterization
of signal traces





features

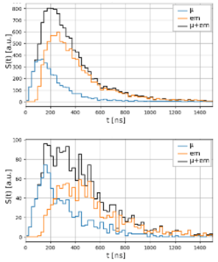


Network extracts from training data *optimized intermediate variables* suited for shower direction

120k parameters

shower direction

RAW input data:

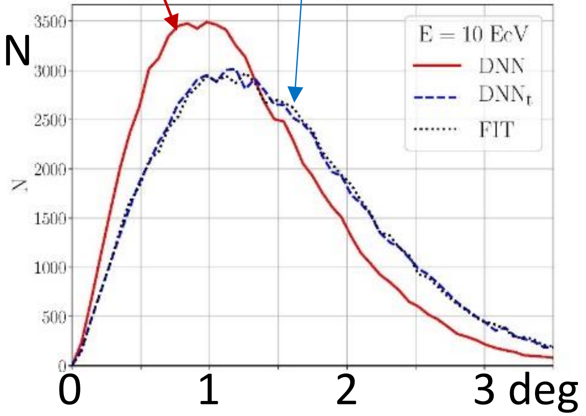


...

Deep Neural Network

- Time offset only
- Black signal traces added

shower direction angular resolution



Deep Neural Network learns physics from data within 3h

Deep Learning

Hierarchy of “features” with increasing level of complexity

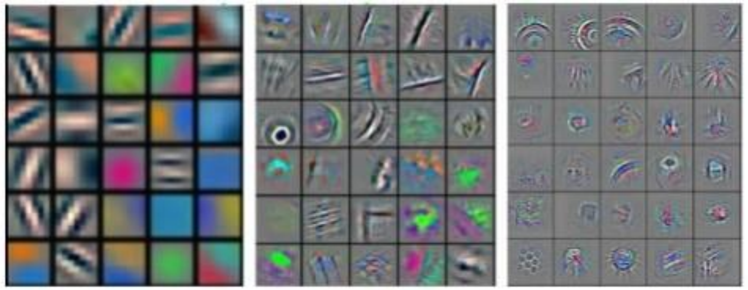


Image recognition

– Pixel→edge→texton→motif→part→object

Text

– Character→word→word group→clause→sentence→story

Speech

– Sample→spectral band→sound→…→phone→phoneme→word

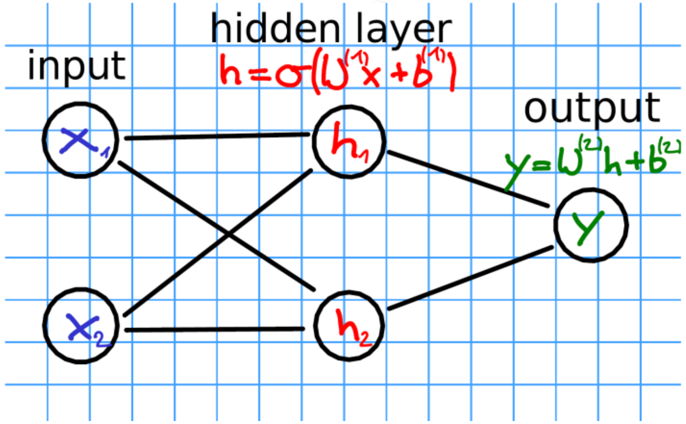
Frank Rosenblatt, Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC, 1961

Neural Networks: 2 pages reminder

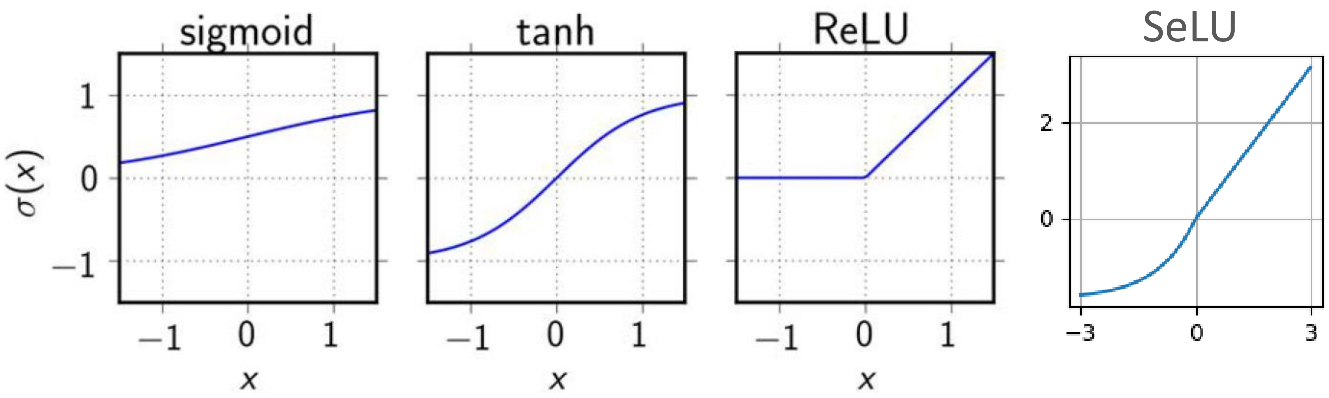
Neural Network Operations

x multi-dimensional input data
W, b to be trained
successively apply 2 operations:

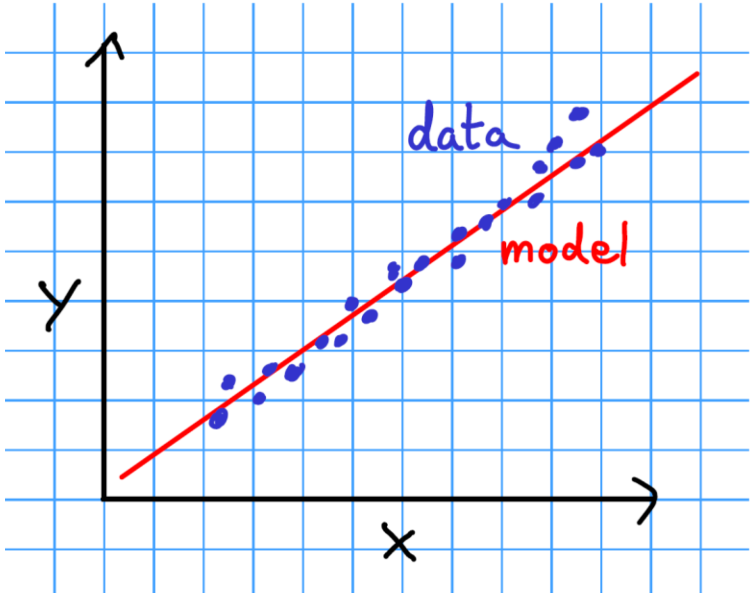
$$\mathbf{y} = \mathbf{W}\,\mathbf{x} + \mathbf{b}$$
$$\mathbf{h} = \sigma(\mathbf{y})$$



activation function: departure from linear system



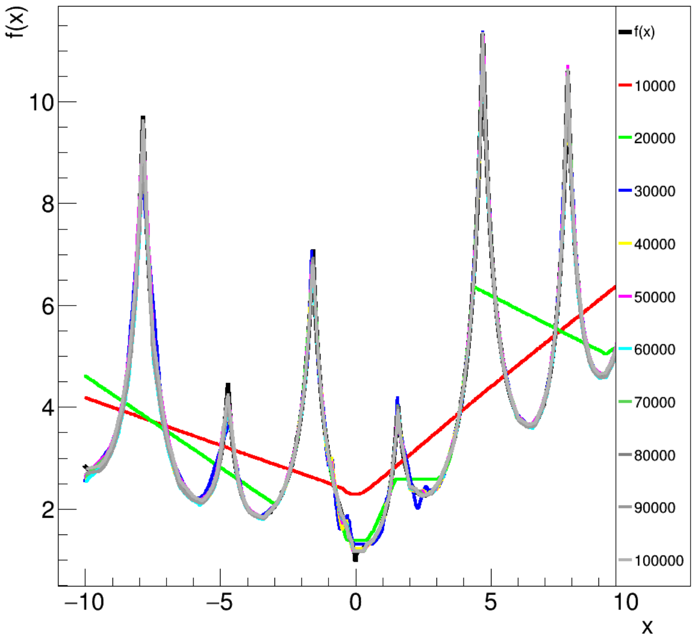
Neural Network Training



D. Walz, RWTH

- **Data** set $\{x_i, y_i\} \quad i = 1 \ldots N$
- Define **model**
  $y_{\mathrm{m}}(x; \theta) = Wx + b$ with $\theta = (W, b)$
- Define **objective** function (also called loss/cost)
  $J(\theta) = \frac{1}{N} \sum_{i=1}^{N} (y_{\mathrm{m}}(x_i) - y_i)^2$
- **Train** the model by optimizing the parameters
  $\hat{\theta} = \arg\min J(\theta)$

Neural Network: automated parameterization of arbitrary function



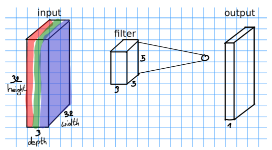
T. Quast CERN/RWTH Aachen

Fully connected neural network
7 hidden units with 20 nodes each
ReLU activation function

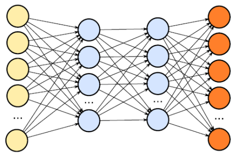
Good description (grey) of original function (black) after 10^5 training steps
(if you don‘t know the original function: don‘t want to invent a suitable parameterization yourself)

Deep Learning Concepts

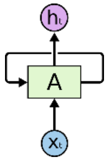
from computer science to particle physics



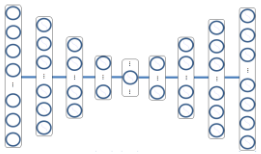
convolutional



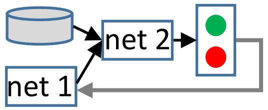
fully connected



recurrent

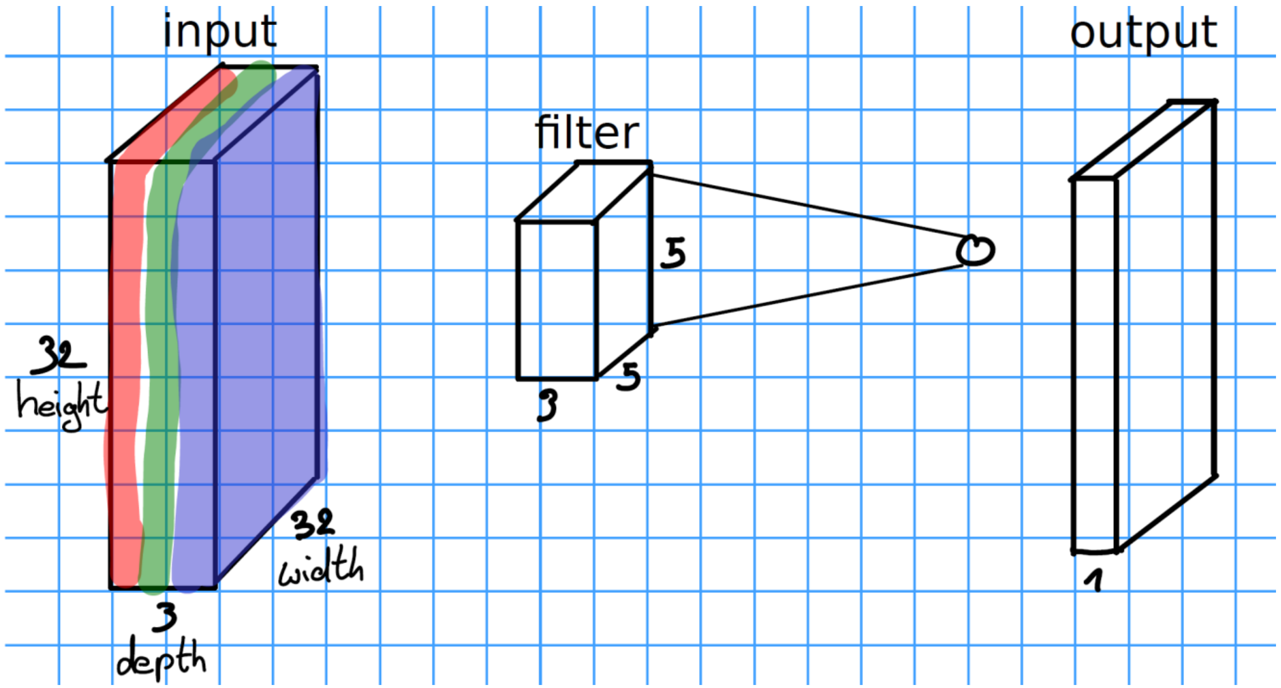


autoencoder



adversarial

Yann LeCun et al. (1989), http://ieeexplore.ieee.org/document/41400

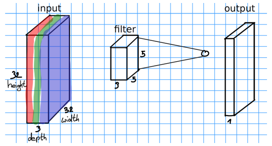


- Slide filter over input volume
- Calculate W x + b
- Apply activation function
- 1 output value at each position
- *Translation invariance*

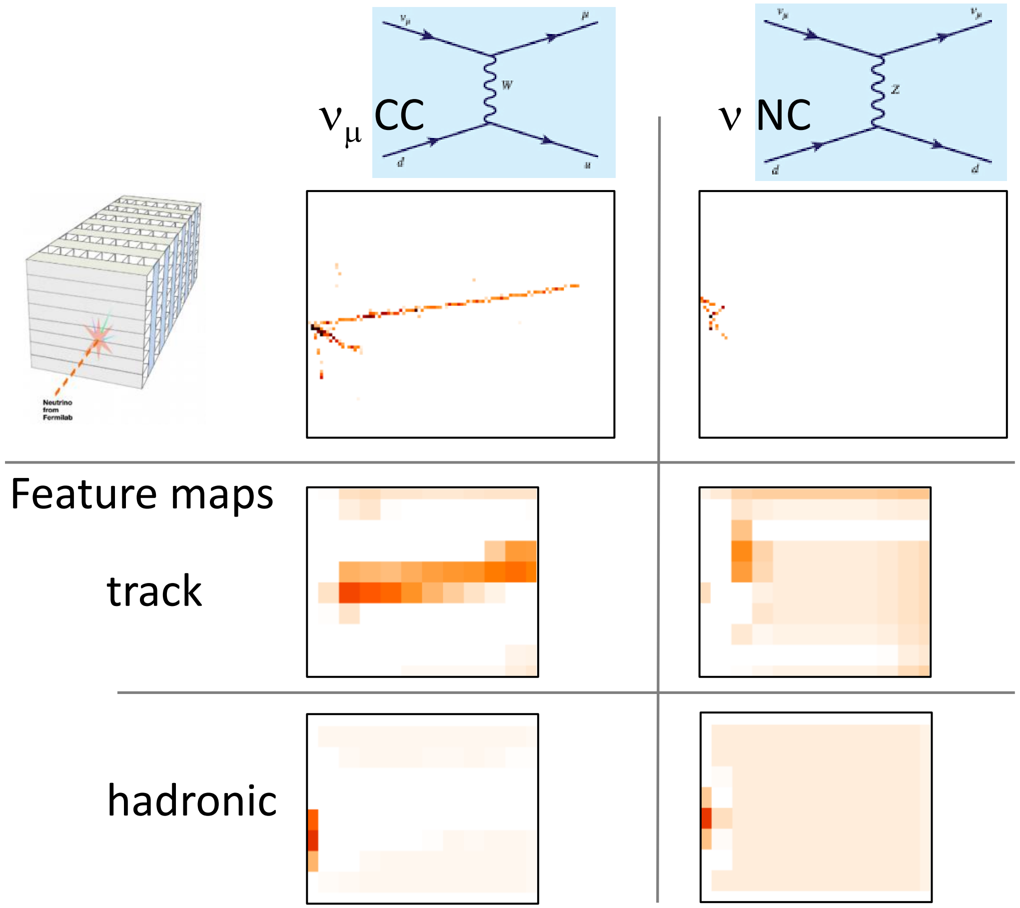
Concept 1: ***convolutional*** networks to analyse image-like data

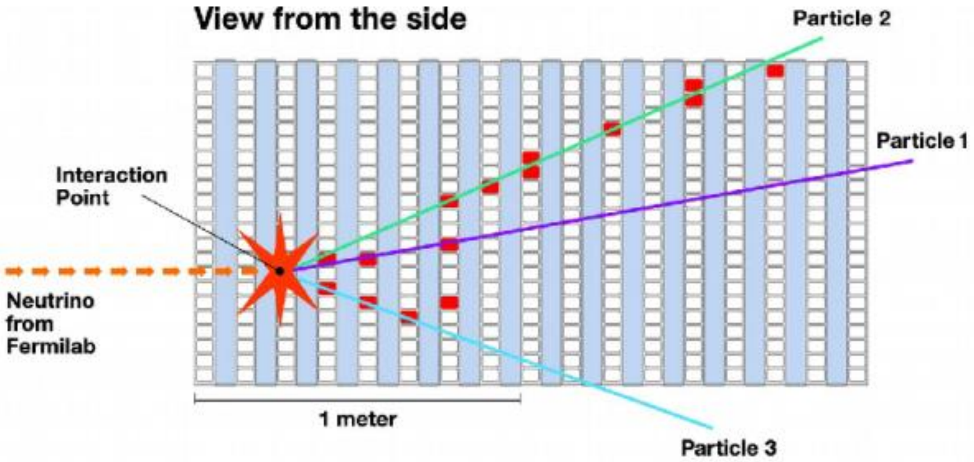
A. Aurisano et al., JINST 11 (2016) P09001

Electron neutrino identification



neural network neutrino event classifier

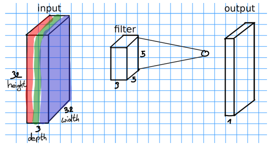


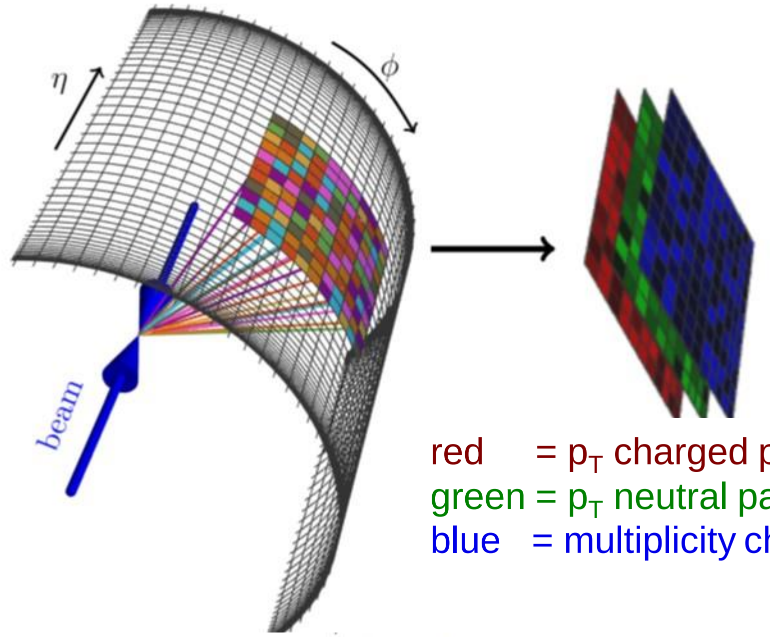


Method	ν_e efficiency (same purity)
Physicists algorithm	35%
Deep learning neural network	49%

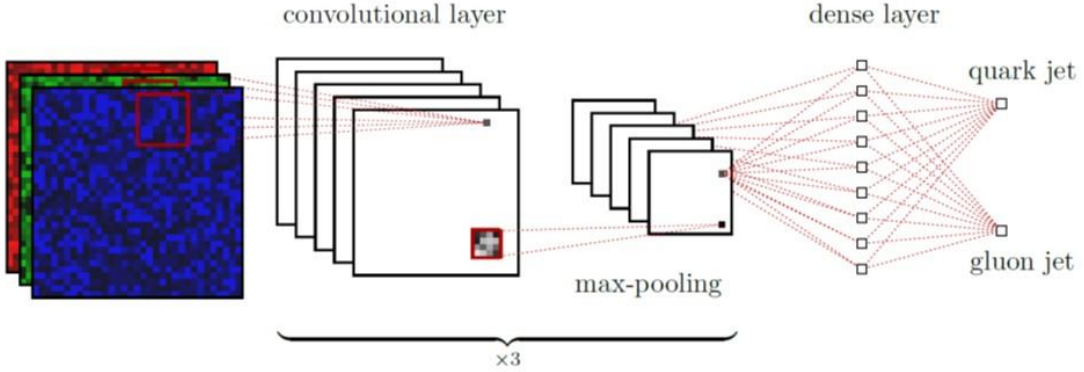
P.T. Komiske, E.M. Metodiev, M.D. Schwartz, arXiv:1612.01551

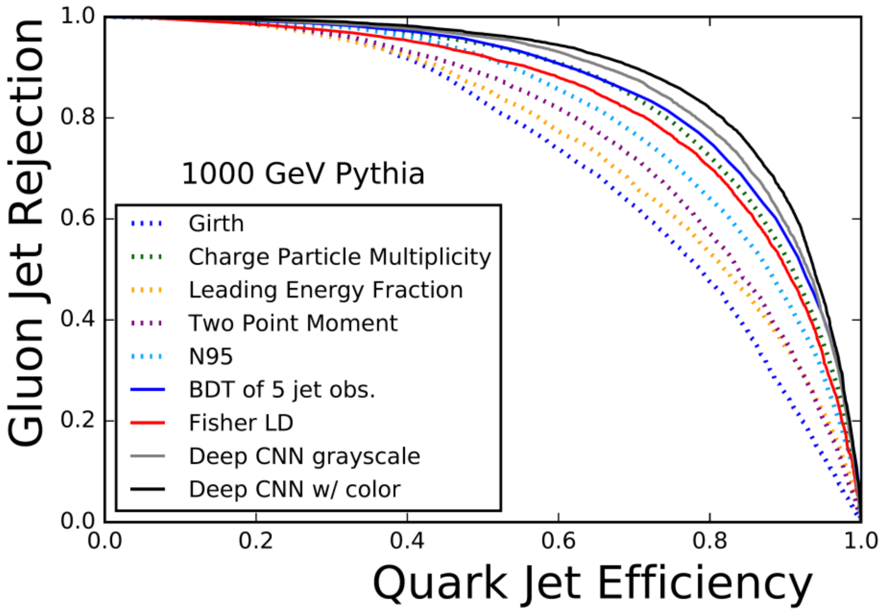
Distinguish quark & gluon jets





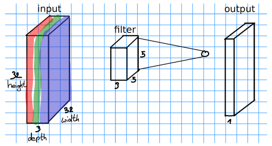
red = p_T charged particles
green = p_T neutral particles
blue = multiplicity charged particles

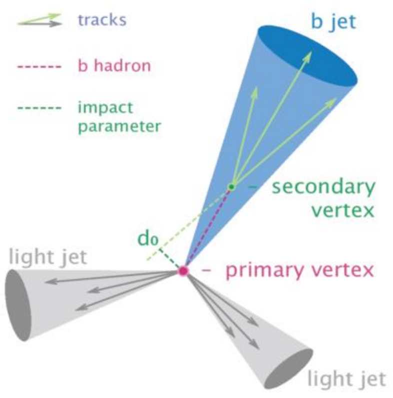


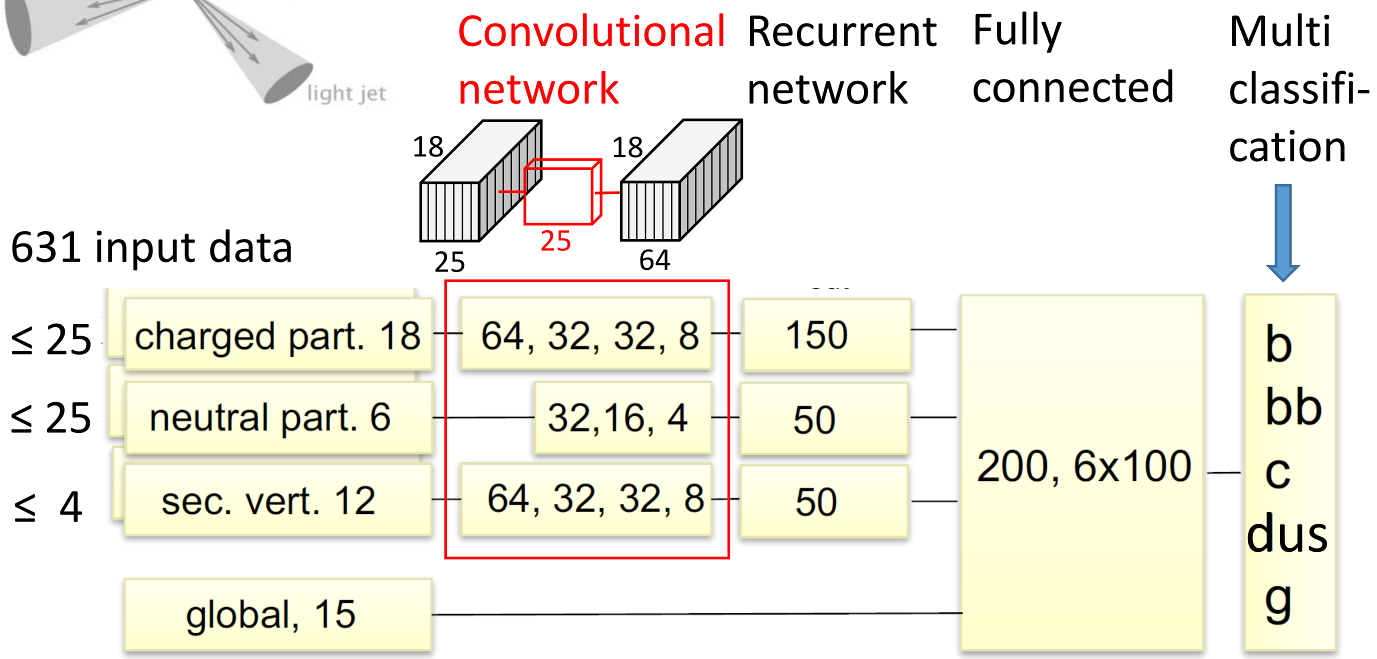


Best performance by convolutional network (black)

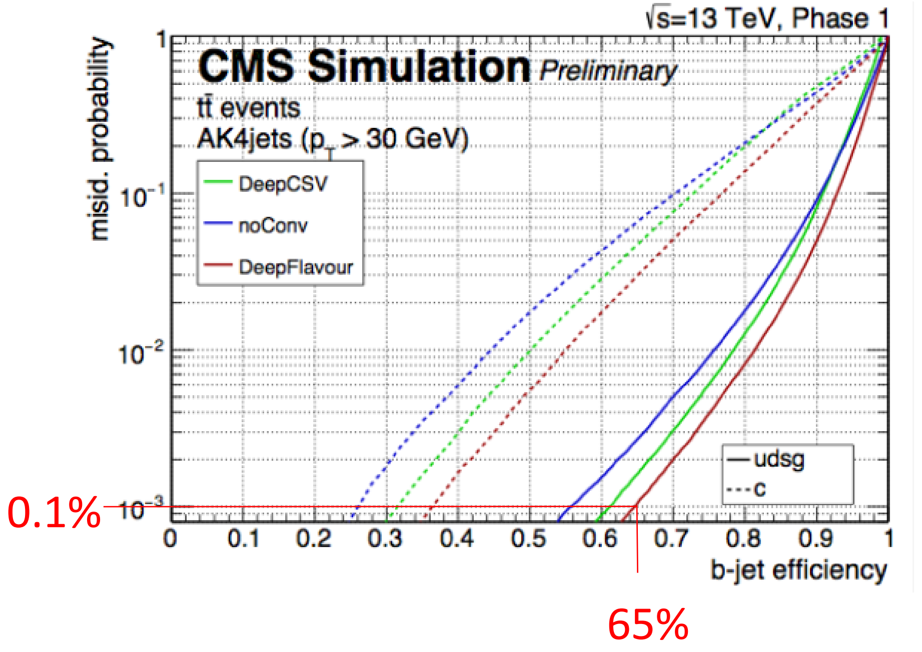
CMS jet flavor tagging



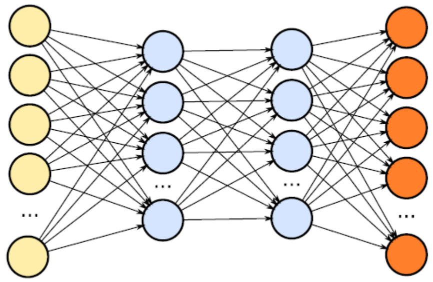




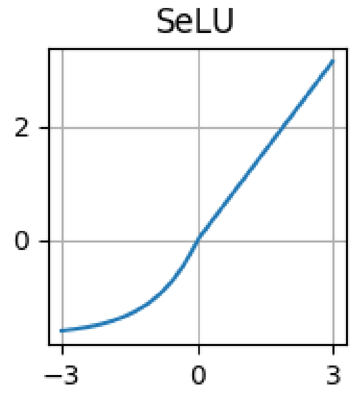
250k parameters



Brillant improvement with high impact on physics exploitation of data



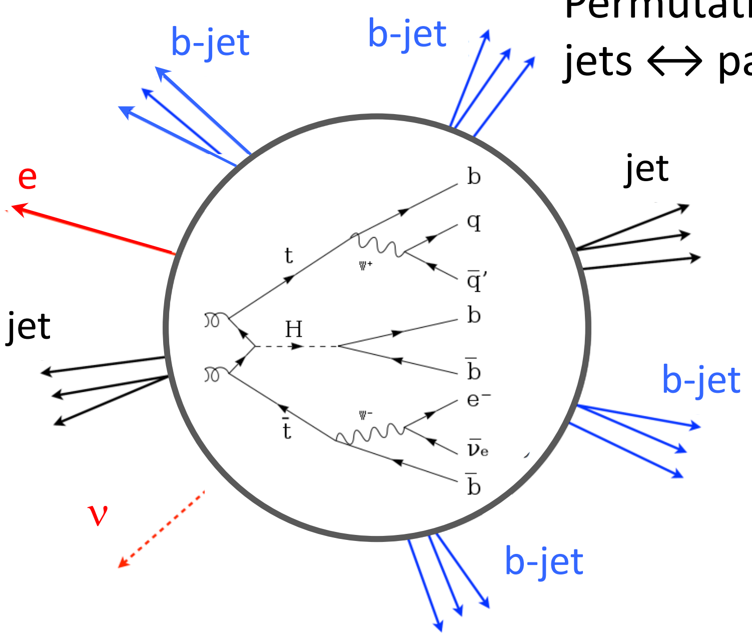
Self-Normalizing
Neural Networks



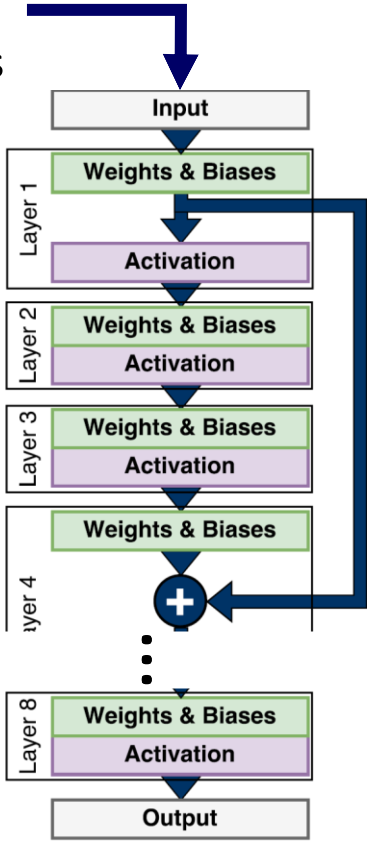
Concept 2: ***fully connected*** networks to analyse fixed length data

M. E., B. Fischer, M. Rieger, JINST 12 (2017) P08020

Jet-parton assignment in ttH events

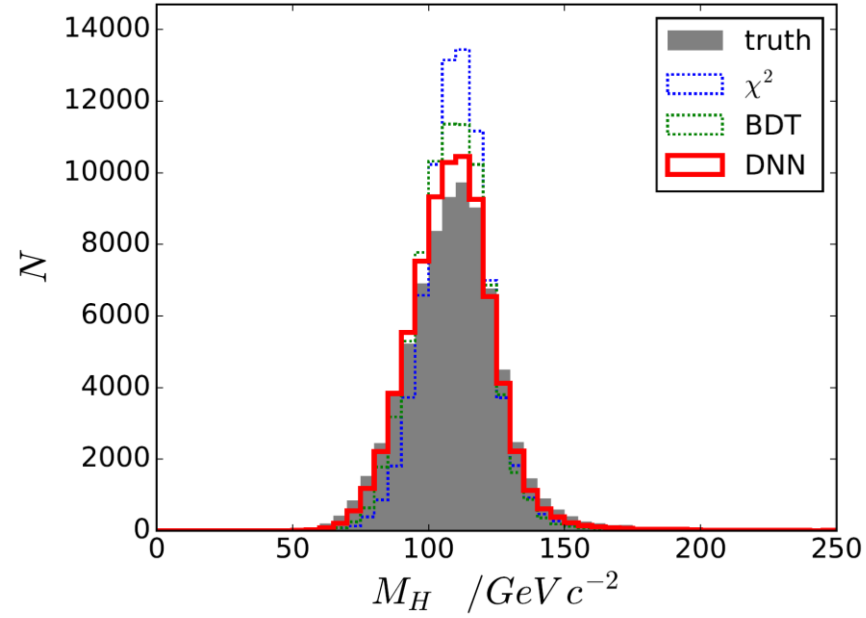


Permutations
jets ↔ partons



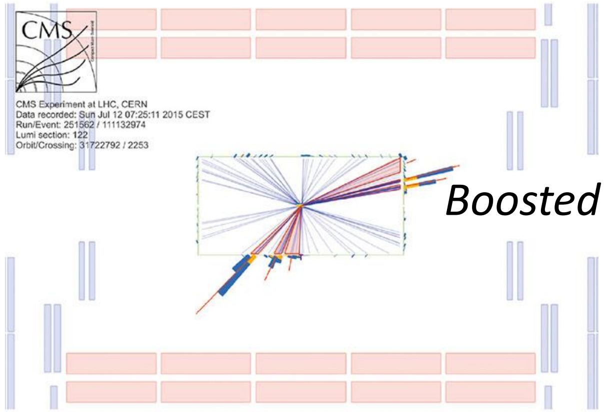
Method	$\varepsilon(t\bar{t}H)$ / %
χ^2 method	37
Boosted Decision Trees	45
Deep Learning	52

2,000k parameters

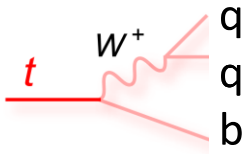


- Network distribution approximates the best possible reconstruction
- Best results: **low level** (4-momenta) **+ high level** (decay angle,…) variables

Top tagging featuring Minkowski space



Boosted top quark → fat jet



1) Combinations of particles

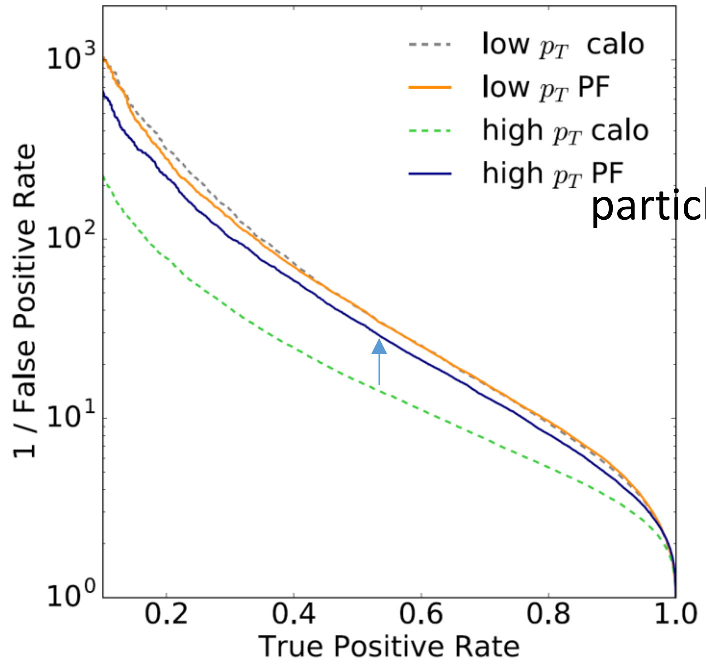
$$\tilde{k}_{\mu,j} = k_{\mu,i}\, C_{ij} = \begin{pmatrix} E & E & \cdots & E \\ px & px & \cdots & px \\ py & py & \cdots & py \\ pz & pz & \cdots & pz \end{pmatrix} \left(\text{single particles} \quad \overbrace{\text{combinations of particles}}^{\text{trainable}} \right)$$

2) Extract features from particle combinations respecting Minkowski metric

$$d^2_{jm} = (\tilde{k}_j - \tilde{k}_m)_\mu \, g^{\mu\nu} \, (\tilde{k}_j - \tilde{k}_m)_\nu$$

$$\hat{k}_j = \begin{pmatrix} m^2(\tilde{k}_j) \\ p_T(\tilde{k}_j) \\ w^{(E)}_{jm} \, E(\tilde{k}_m) \\ w^{(d)}_{jm} \, d^2_{jm} \end{pmatrix}$$

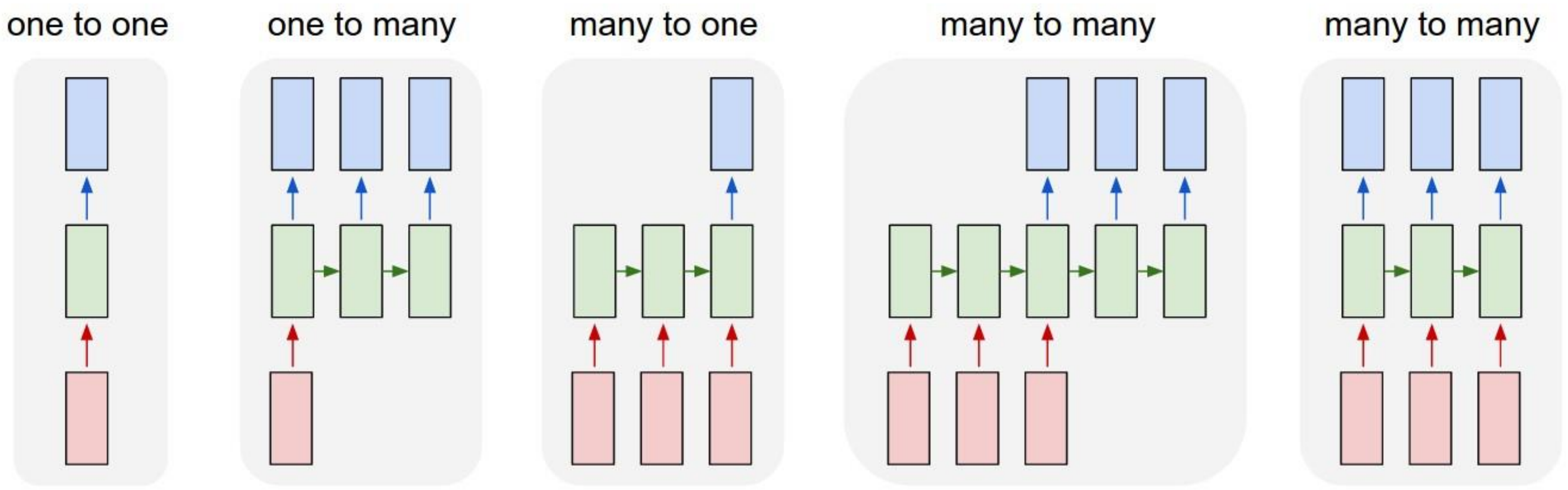
trainable



particle flow

Efficiency of top quark identification

- Include Minkowski metric for 4-vectors
- Let network decide on relevant particle combinations



Fully connected,
Convolutional
networks

Image: Kaparthy

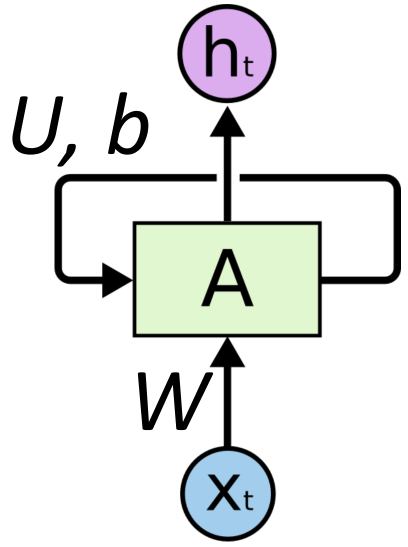
Concept 3: ***recurrent*** networks
to analyse input data of variable size

D.E. Rumelhart, G.E. Hinton, R.J. Williams, Nature 323(9-Oct-1986)533
S. Hochreiter, J. Schmidhuber, Neural Computation 9(1997)1735

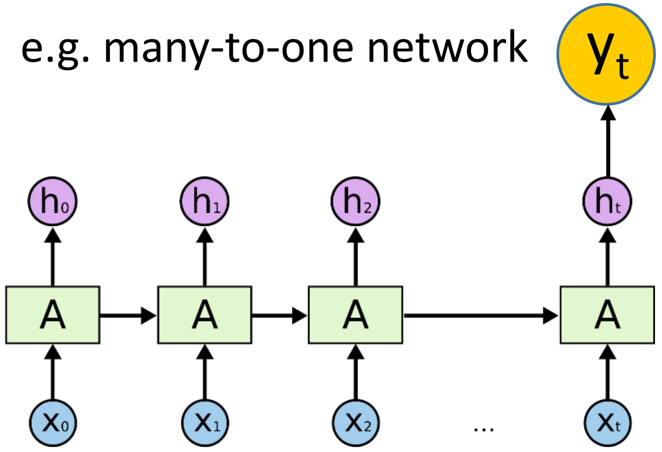
Recurrent networks

Hidden state depends on previous state

$h^{(t)} = \tanh\left(U\, h^{(t-1)} + W\, x^{(t)} + b \right)$



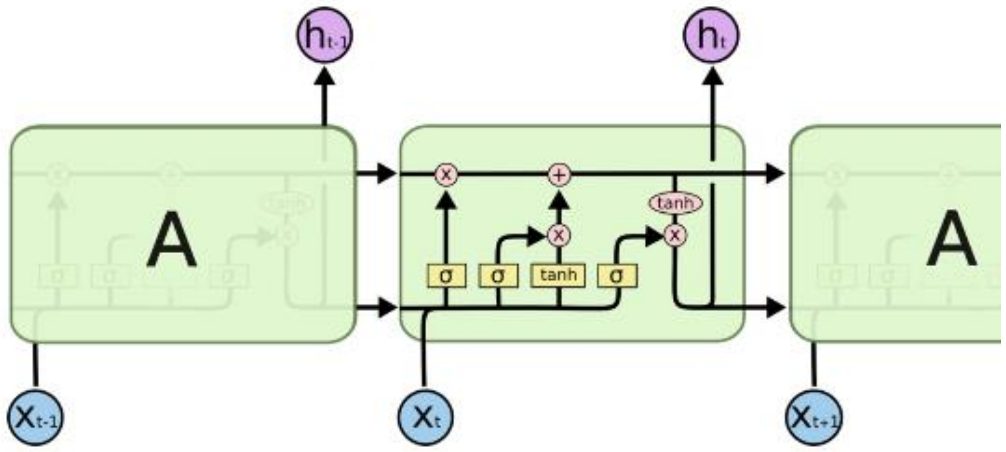
e.g. many-to-one network



Efficient parameter sharing: same *W, U, b* for all input data

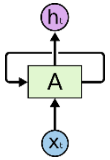
Long **S**hort **T**erm **M**emory

E.g. detecting correlations between tracks on input far away from each other through a **cell memory**

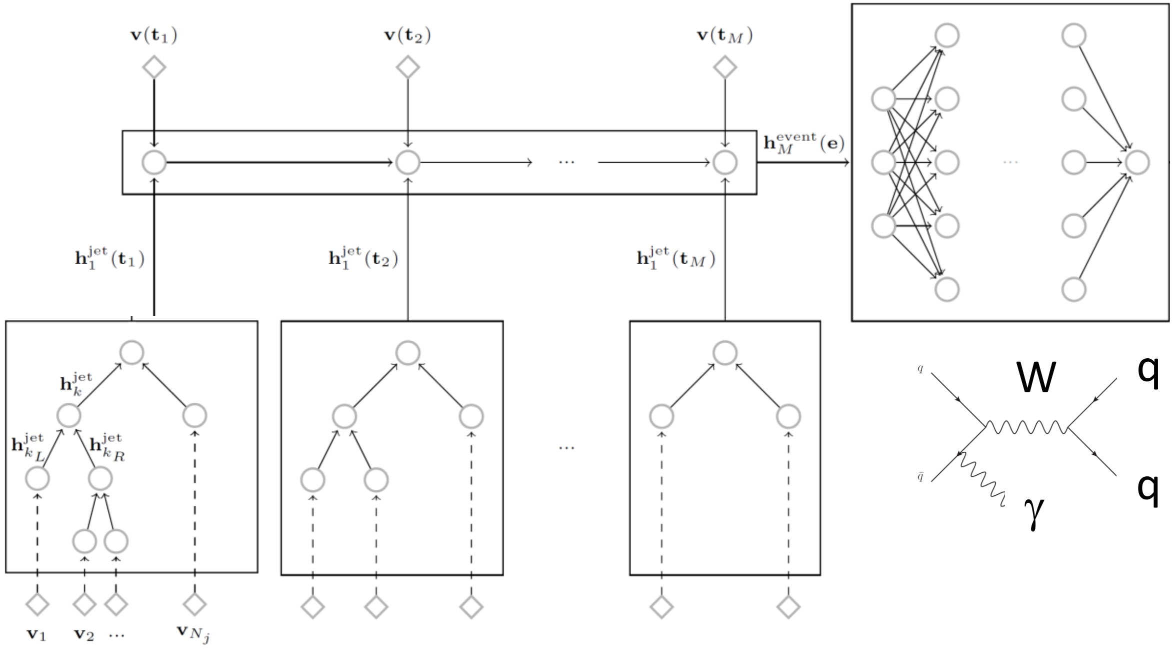


Jet formation & fat jet identification

Recurrent network combines jets of event

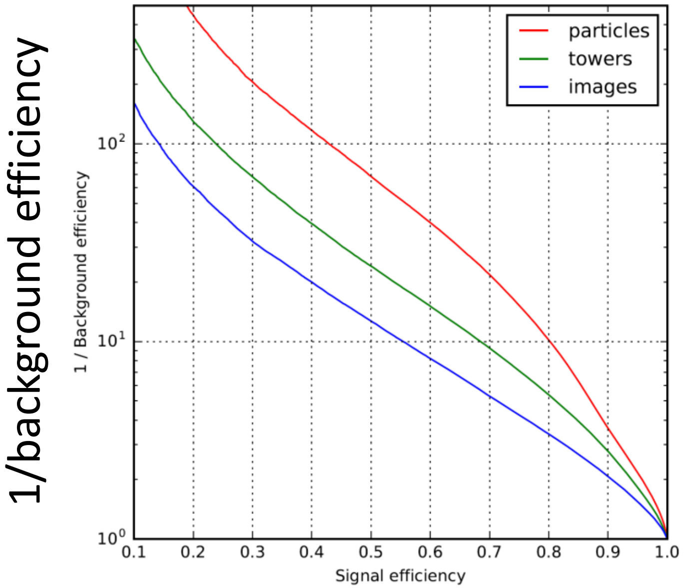


Fully connected: fat W jet or not



network combines particle 4-momenta to jets (recursive)

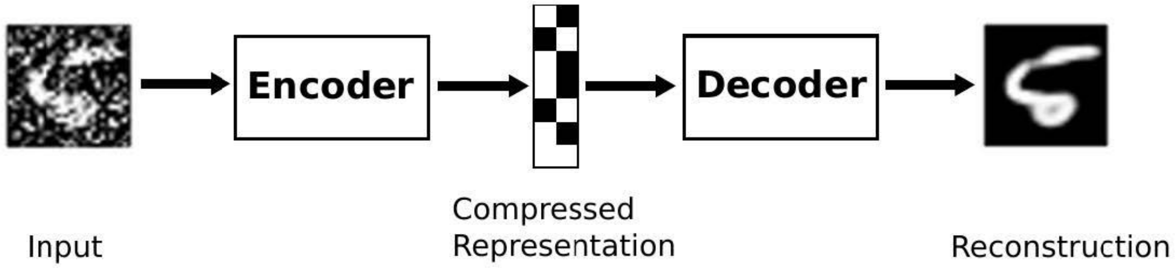
30k parameters

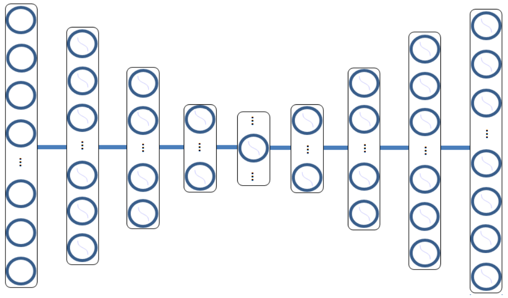


Efficiency of finding boosted W fat jet

Physics motivated:
mapping 4-vectors into network
to build & process jets

D.H. Ballard, Proc. 6th Nat. conf. on Artificial intelligence, Seattle, Washington, USA, Vol.1 (1987) 279
P. Vincent, H. Larochelle, Y. Bengio, P.-A. Manzagol, Proc. 25th Int. Conf. on Machine Learning, Helsinki, Finland, (2008) 1096





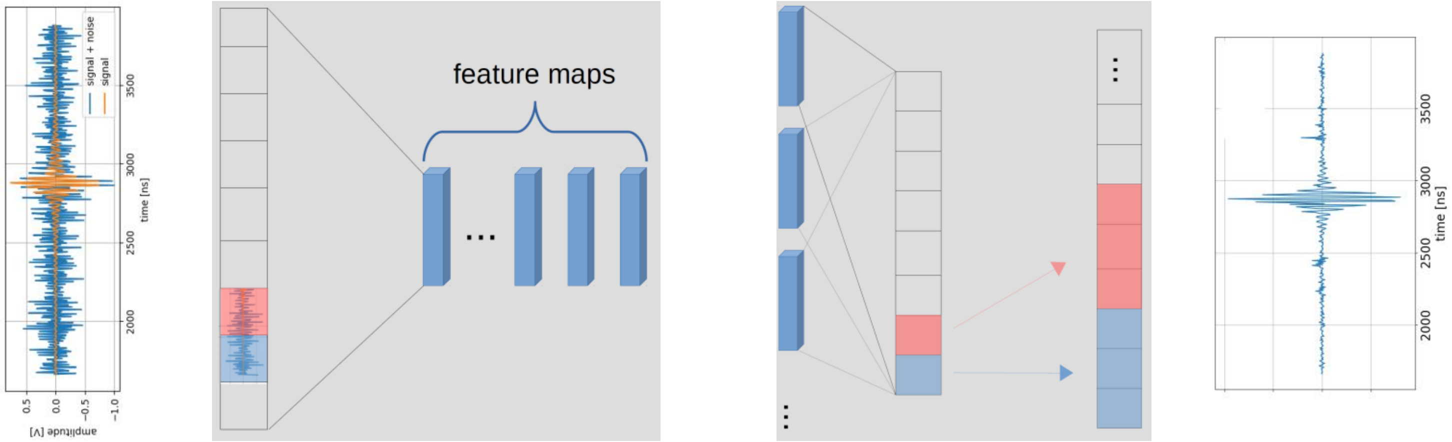
Concept 4: ***autoencoder*** networks to extract signal from noisy data

F. Schlüter, M.E., D. Walz, RWTH Aachen

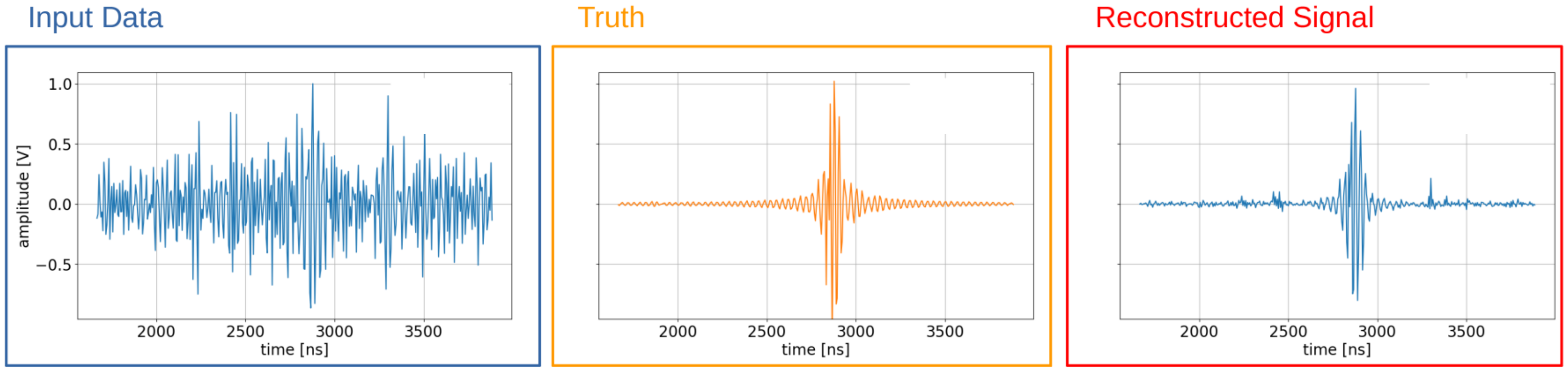
Radio signal from cosmic ray air showers

Encoder

Decoder



Noise from data + signal from simulation

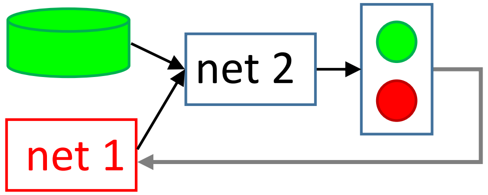


Signal energy & frequency spectrum approx. conserved

I.J. Goodfellow et al, arXiv:1406.2661

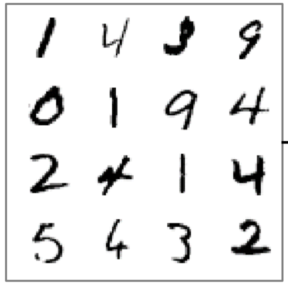

Sofia Vallecorsa ACAT 2017

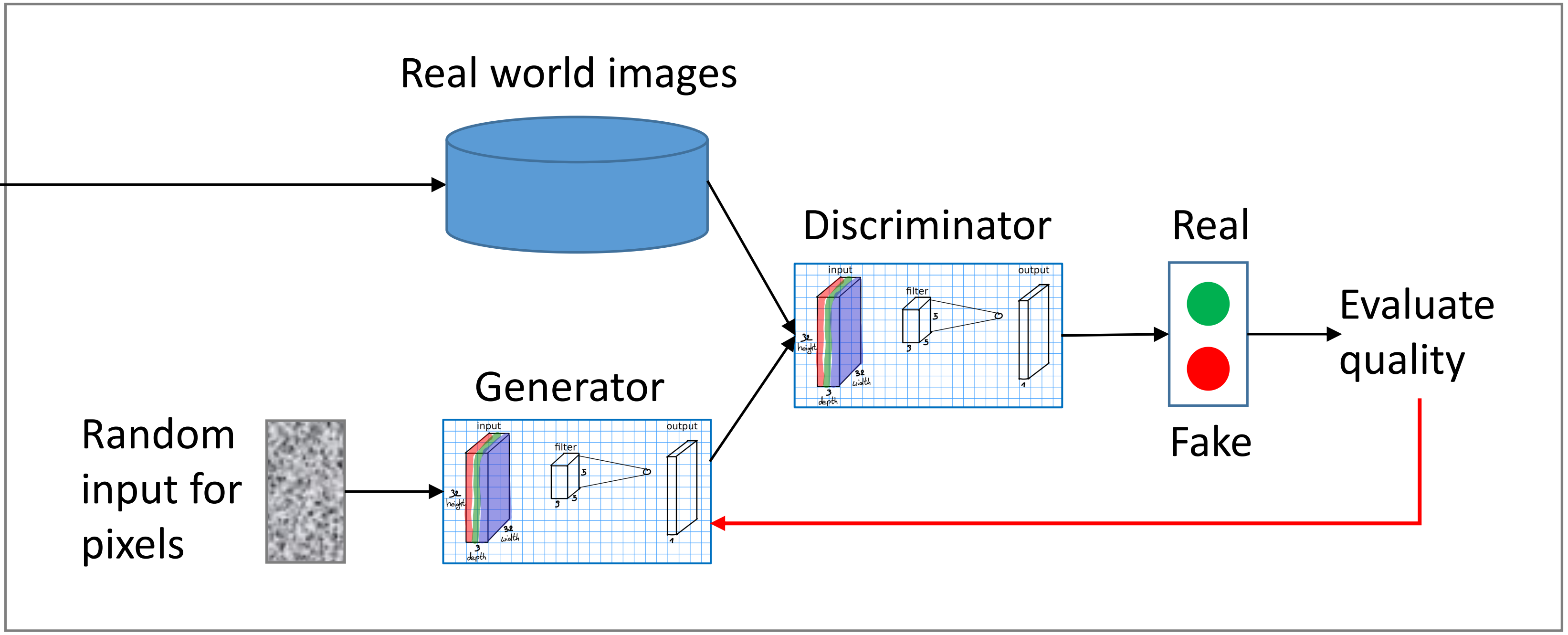
- Counterfeiter: shows police **real & fake money**
- Police: learns to distinguish & gives feedback
- Counterfeiter: new fake money based on feedback
- Iterate until police is fooled

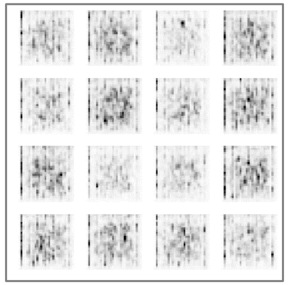


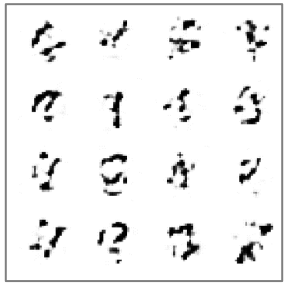
Concept 5: Adversarial network to selectively modify network output

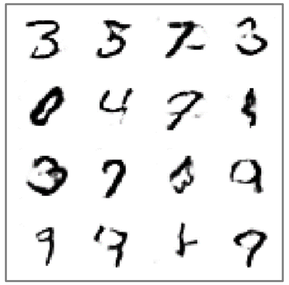
Principle of **G**enerative **A**dversarial **N**etworks





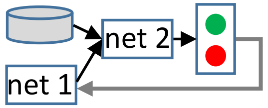

1st try

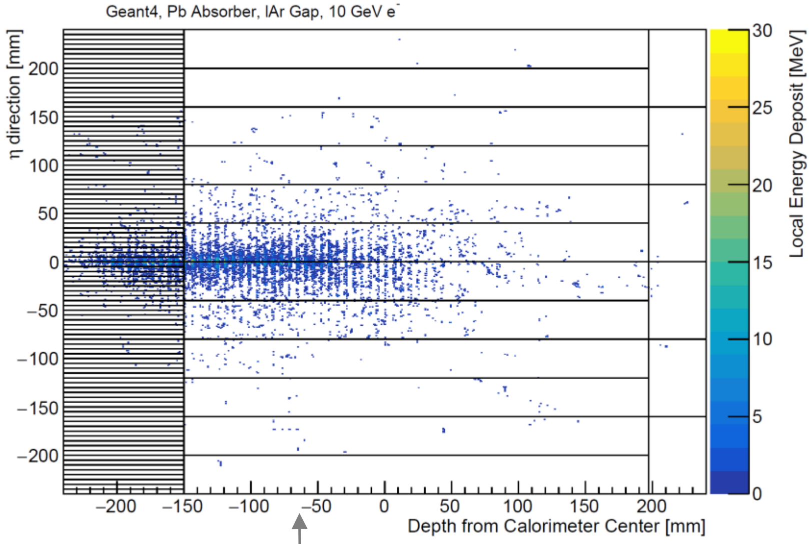

2nd try


4th try

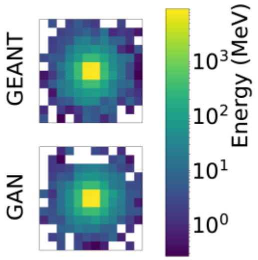
from RWTH course for masterstudents
Deep Learning in Physics Research
Summer term 2017
S. Schipmann, D. Walz, M.E., U. Klemradt

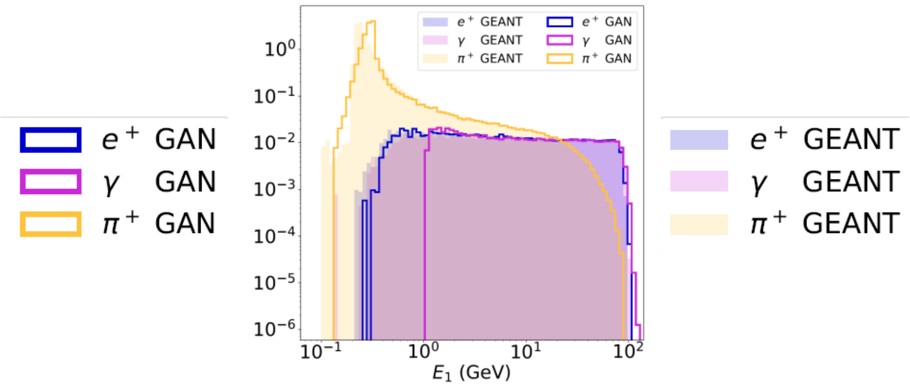
Calorimeter simulations: CaloGAN





Electron
in central
calorimeter



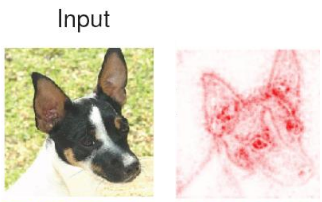


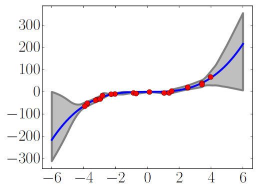
Energy uniform between 1 GeV and 100 GeV

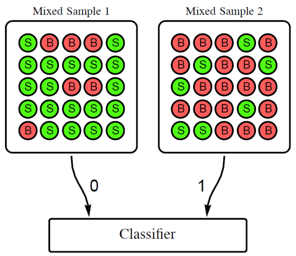
Generation Method	Hardware	milliseconds/ shower
GEANT4	CPU	**1772**
CALOGAN	CPU	2.03
	GPU	**0.012**

- Simulation time faster by 10,000
- Massive parallel computing architecture required (GPU)
- ACAT 2017: Configurable framework planned for fast simulation → Integration in GEANT V

Research on network causality, uncertainty, data labels



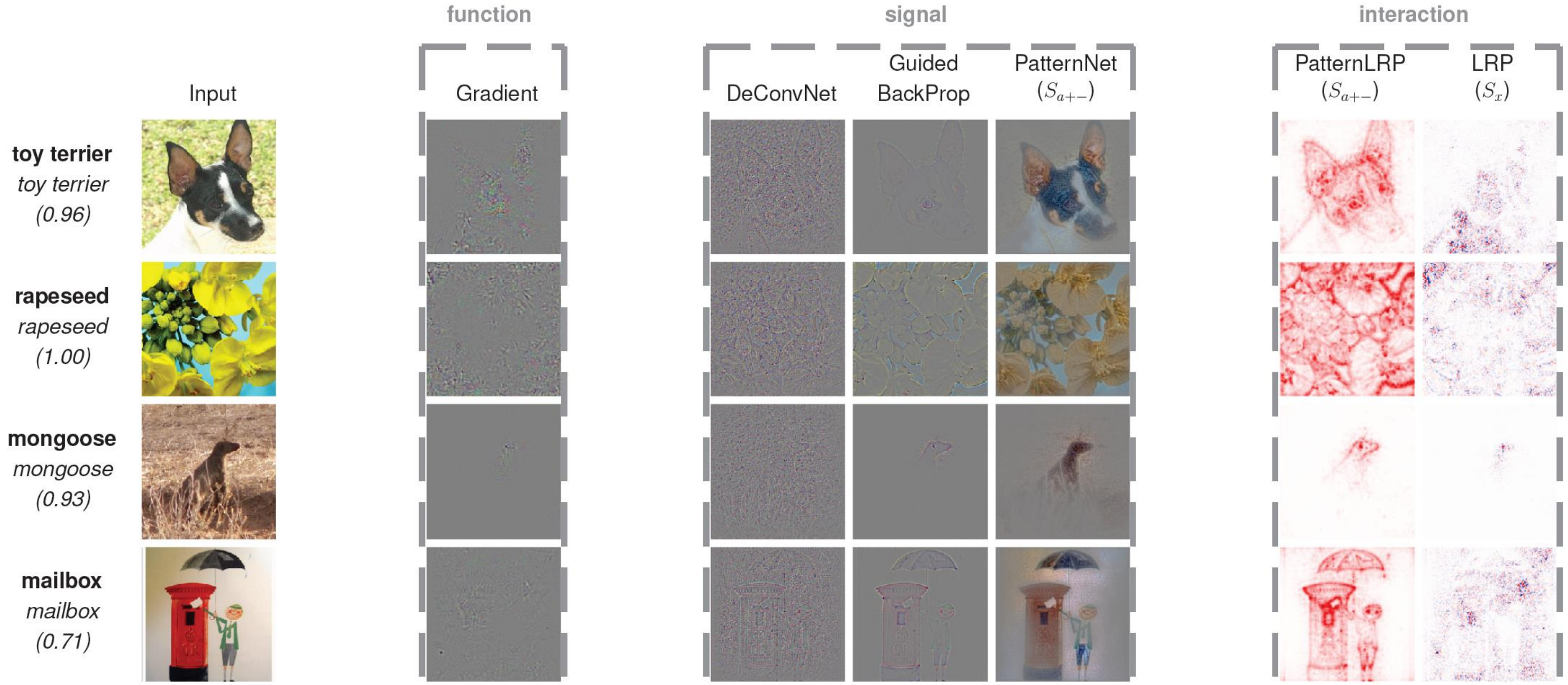




P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, S. Dähnear, arXiv:1705.05598

Causality

Different methods to investigate network decisions



Back-projections to input space

Heatmap: influence of pixel on output

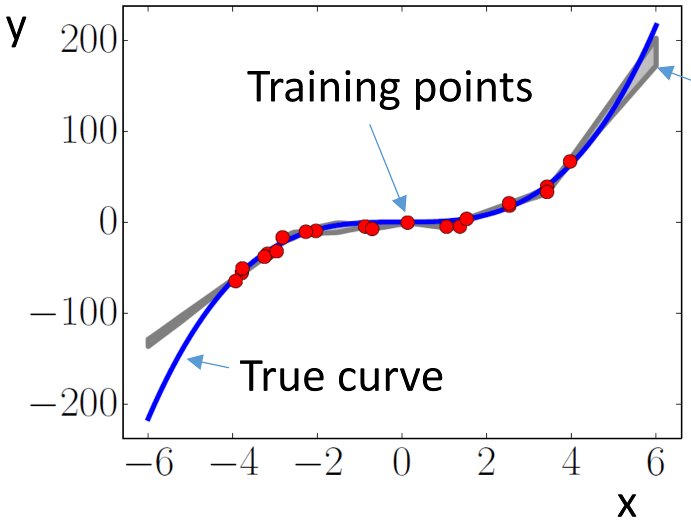
Important progress opening black box

B. Lakshminarayanan, A. Pritzel, C. Blundell, arXiv:1612.01474

Predictive uncertainty estimation

Use ensembles of deep neural networks (*deep ensembles*)

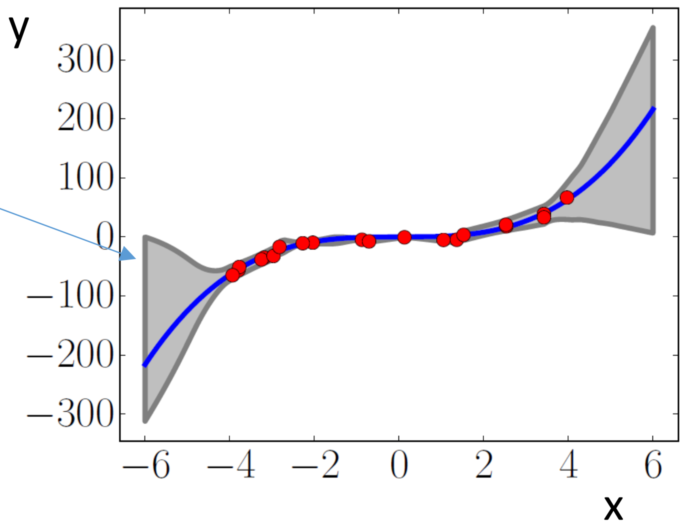
Simple averaging



Predicted mean
& std.dev.

Adversarial training

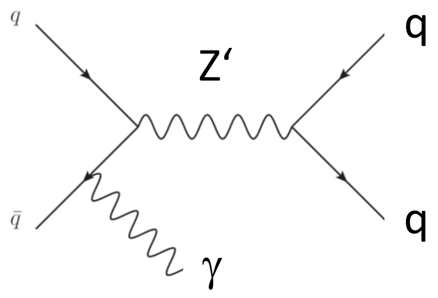
x‘ = x + perturbation



Uncertainty prediction from ensembles of networks improves with dedicated training

C. Shimmin at al., arXiv:1703.03507

Decorrelation



Problem: model jet mass of background processes
→ decorrelate
fat jet & jet mass

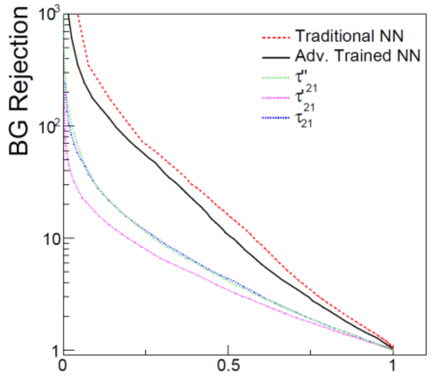
classifier:
Fat jet or QCD?

adversary: infer jet mass from classification output

Fully connected 3 layer 300 nodes	→	Fully connected 1 layer 50 nodes	→ 10 jet mass bins

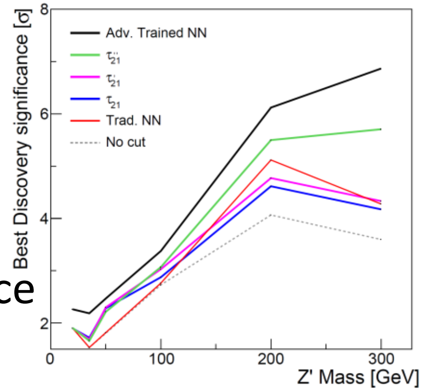
$$L_{\text{tagger}} = L_{\text{classification}} - \lambda L_{\text{adversary}}$$

Best: Traditional network



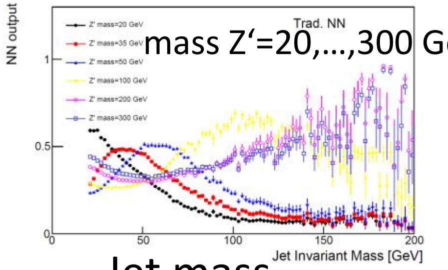
Best: Adversarial network

Discovery significance

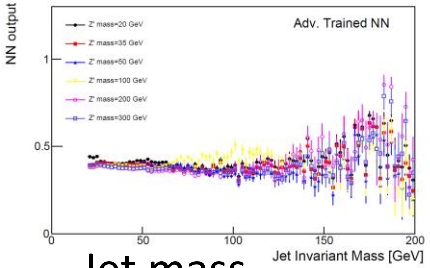


NN output

mass Z‘=20,…,300 GeV



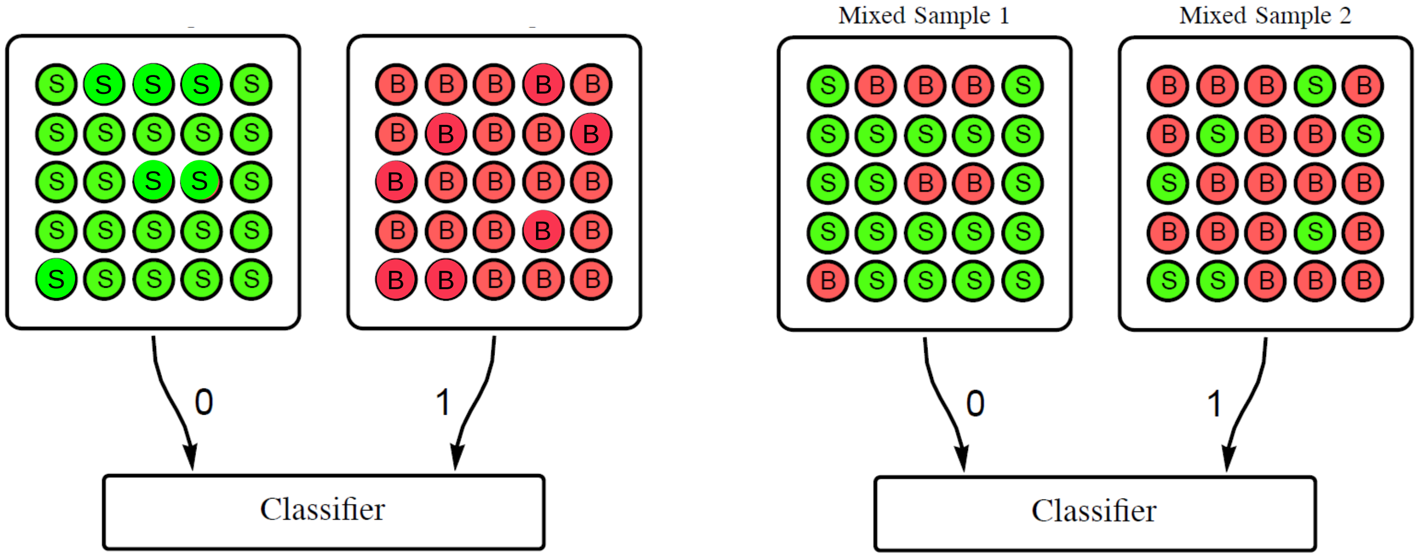
Jet mass



Jet mass

Decorrelation by 2nd network reduces systematic dependency, improves discovery significance

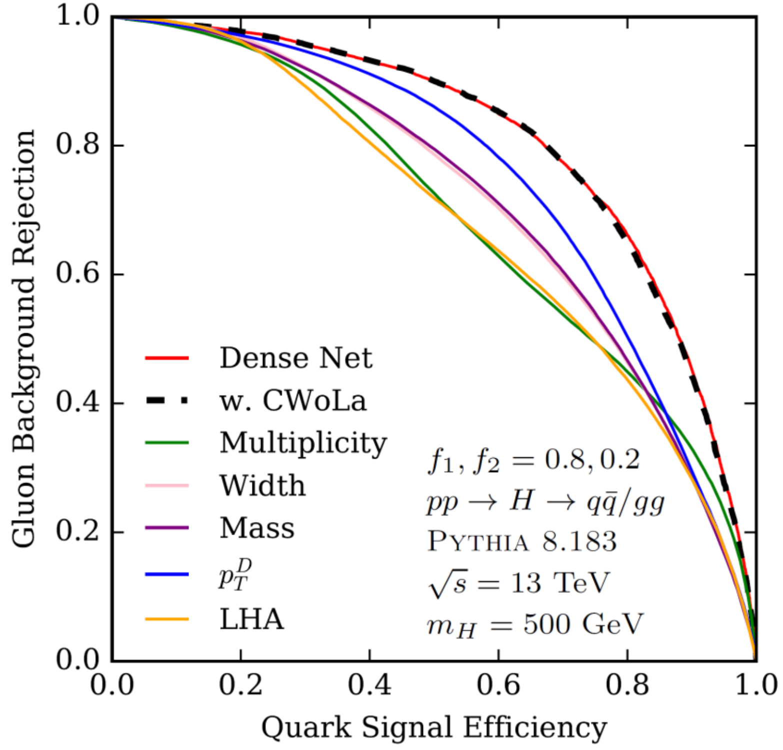
Classification without labels



$$p_{M_1}(\vec{x}) = f_1\, p_S(\vec{x}) + (1 - f_1)\, p_B(\vec{x})$$
$$p_{M_2}(\vec{x}) = f_2\, p_S(\vec{x}) + (1 - f_2)\, p_B(\vec{x})$$

Likelihood

$$L_{M_1/M_2} = \frac{p_{M_1}}{p_{M_2}} = \frac{f_1\, p_S + (1 - f_1)\, p_B}{f_2\, p_S + (1 - f_2)\, p_B} = \frac{f_1\, L_{S/B} + (1 - f_1)}{f_2\, L_{S/B} + (1 - f_2)} \quad , \quad f_1 \neq f_2$$



Step towards training with data

Summary on deep learning

- **Machines exploit physics contained in data deeper than before**
- **Modeling particle physics into deep network architecture**
- **Investigations of causality, stability, uncertainties**