Particles and Cosmology

16th Baksan School on Astroparticle Physics

Machine Learning in Astroparticle Physics

Oleg Kalashev
Institute for Nuclear Research, RAS

Lecture 5

April 10-18, 2019

Applying ANN to images, time series, etc.

Translational invariance

- Handwritten digits recognition

MNIST database
("Modified National Institute of Standards and Technology")

- Classification of galaxies using images from Sloan Digital Sky Survey

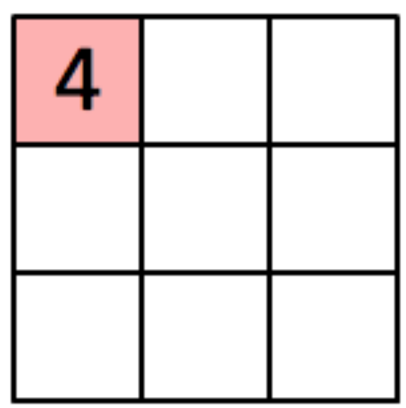
Galaxy Zoo Challenge

Convolutional ANN

The main idea: extract local features and build their maps

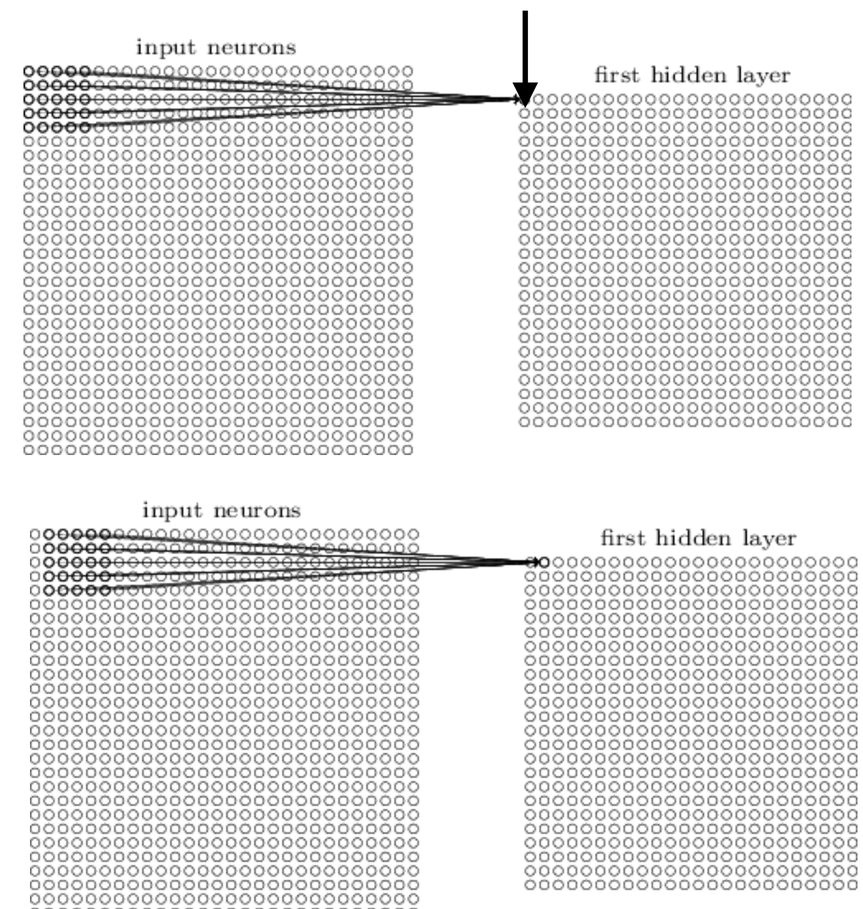
- Convolutional kernel usually has small size (compared to image)

Image



Convolved Feature

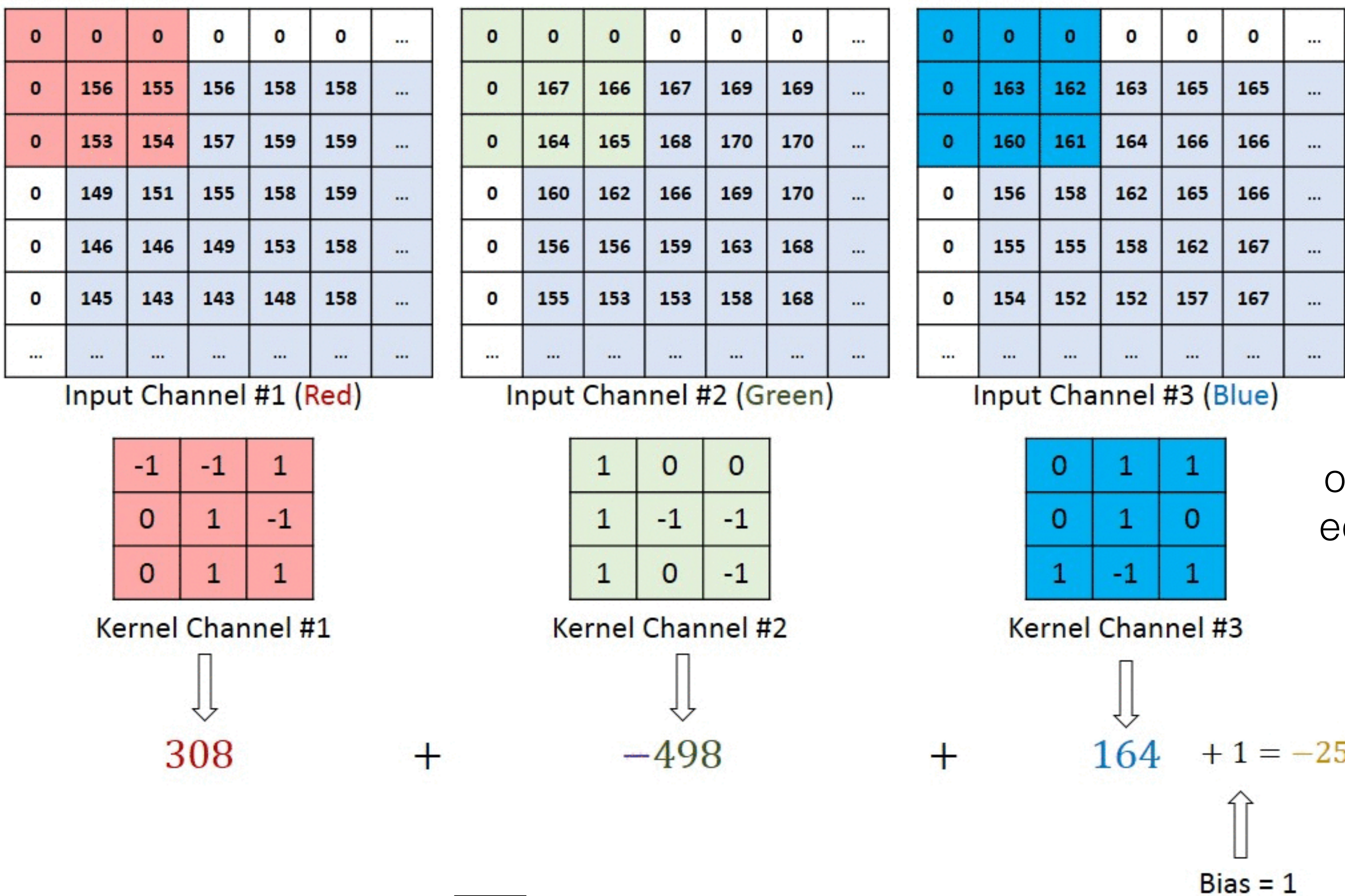
$$\sigma\left(b+\sum_{l=0}^{4}\sum_{m=0}^{4} w_{l,m}a_{j+l,k+m}\right)$$



padding: ‘valid’ (unpadded)

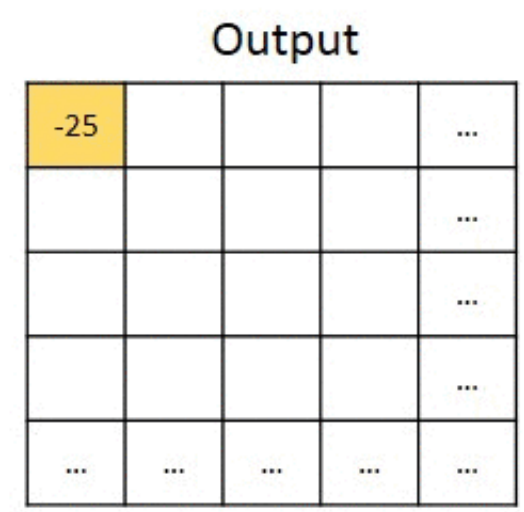
Convolutional layer

building feature maps



padding: 'same'

output image size is equal to input image size

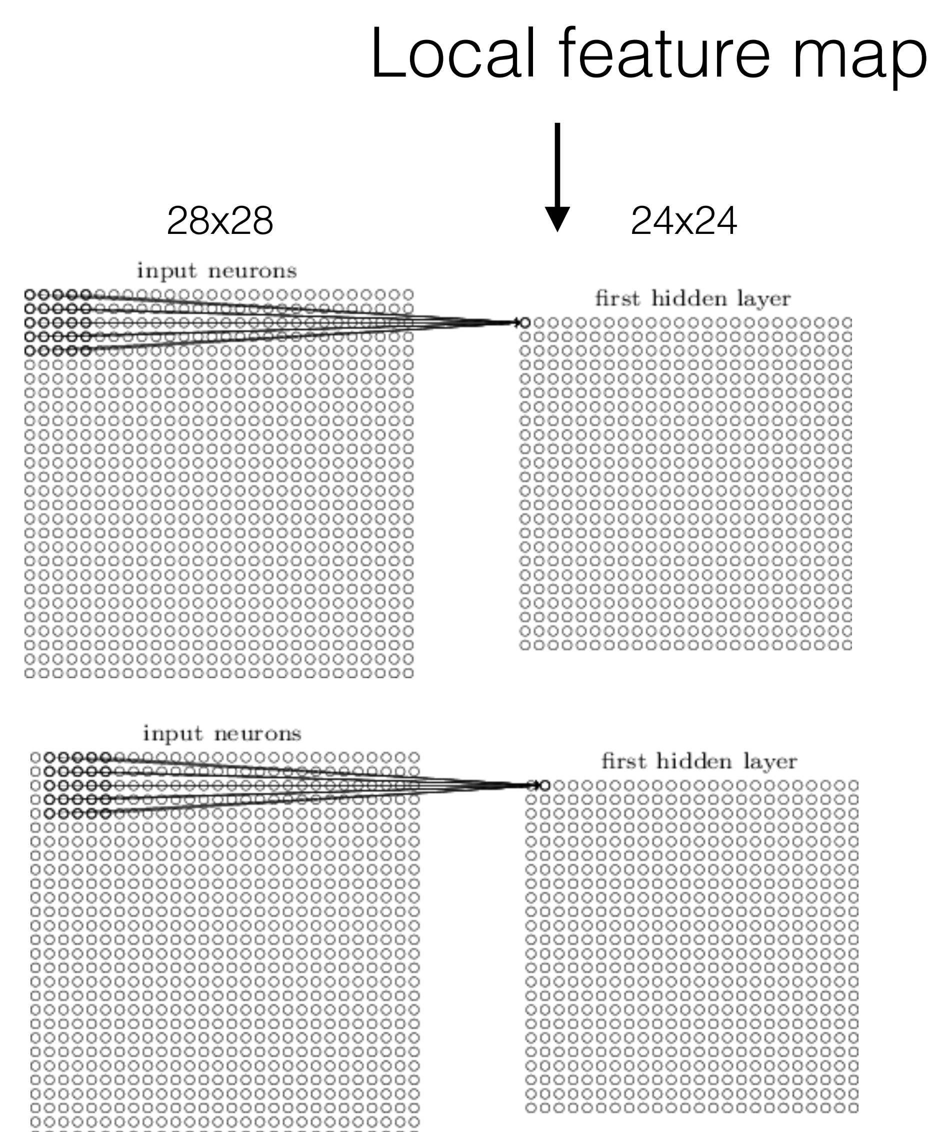


$$a^{l+1}_{d,m,n} = \sigma(b_d + \sum_{c,\alpha,\beta} W_{dc\alpha\beta} a^{l}_{c,m\delta_1+\alpha,n\delta_2+\beta})$$

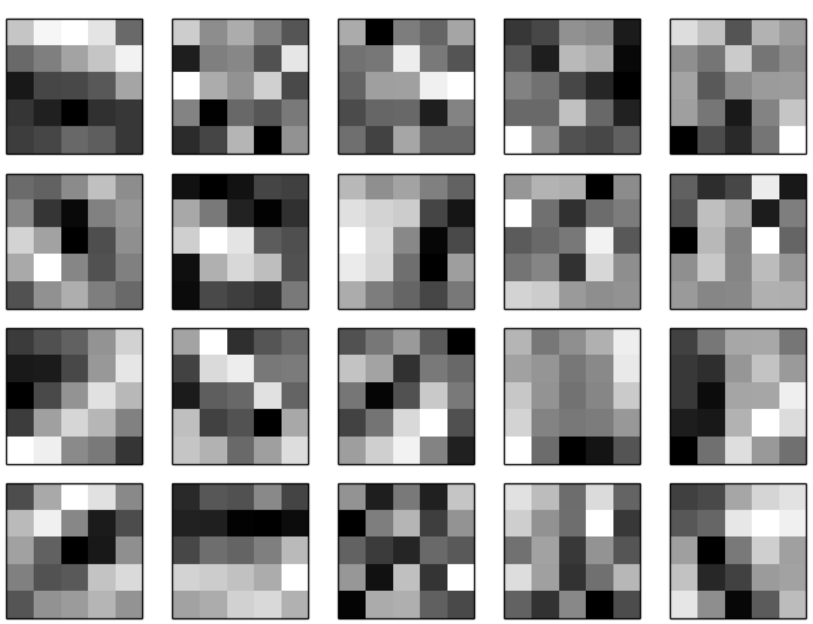
Convolutional layer
building feature maps

- Can be viewed as MLP with most of weights equal to zero and the rest of them are shared
- Number of independent weights:

5x5x*N* vs 28x28x24x24 , where
N - number of maps we want to build



How filters look like (MNIST):
(weights are color-coded)

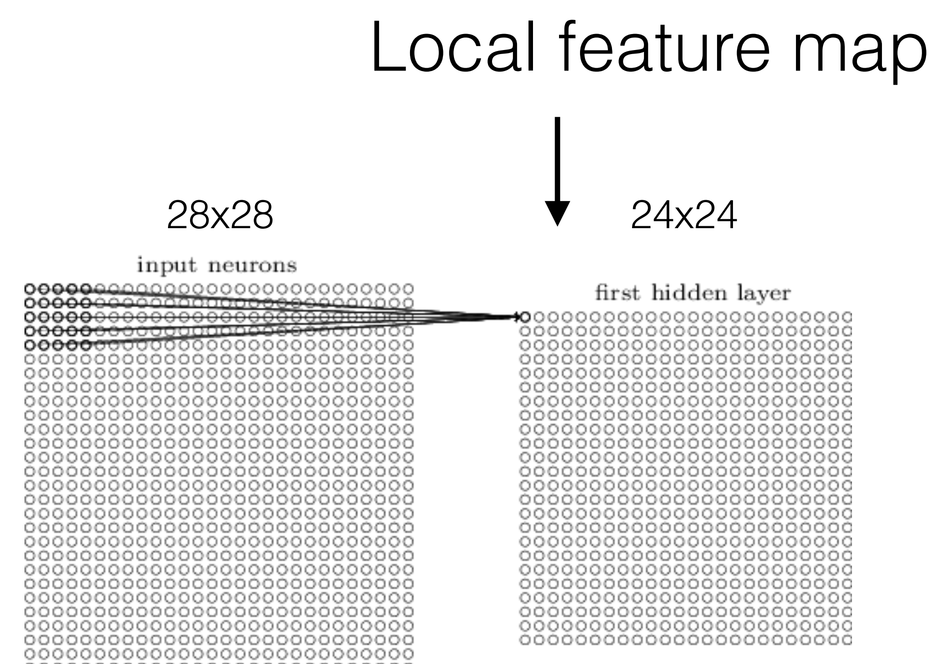


Convolutional layer

building feature maps

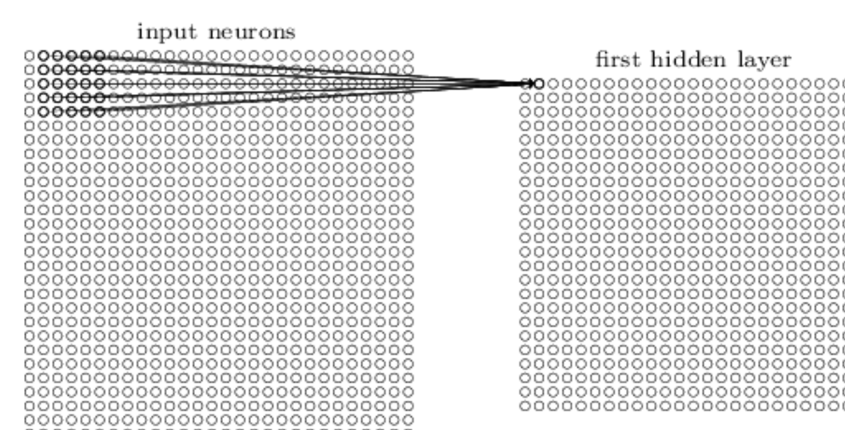
- Can be viewed as MLP with most of weights equal to zero and the rest of them are shared
- Number of independent weights:

5x5x*N* vs 28x28x24x24 , where
N - number of maps we want to build

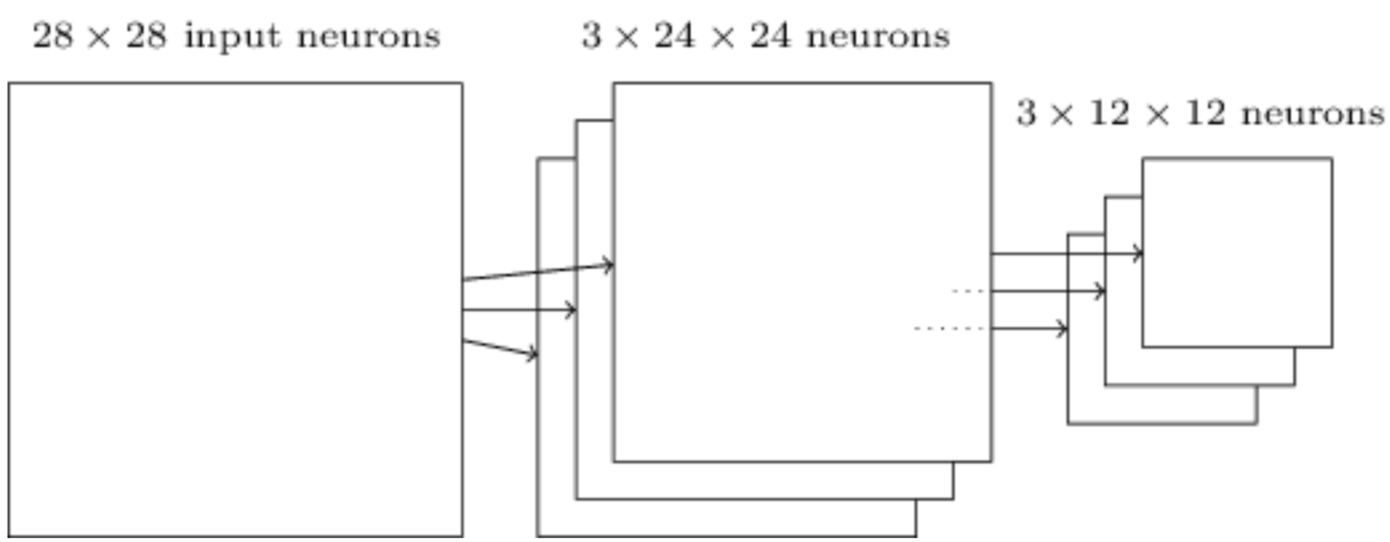


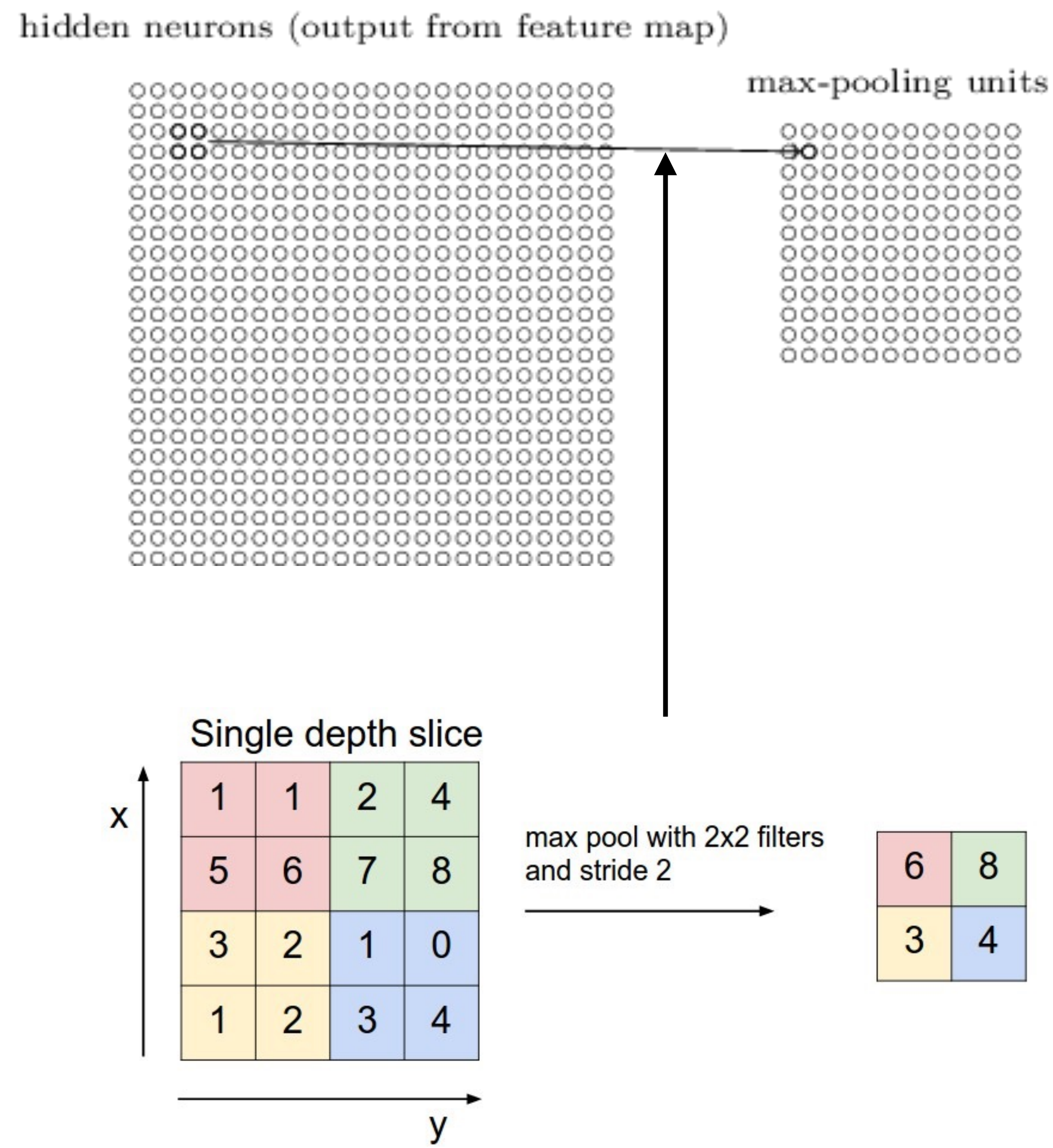
Convolutional layer parameters:

- kernel size
- stride
- padding: (number of zeros appended to the sides of the image before convolution)



Pooling layer



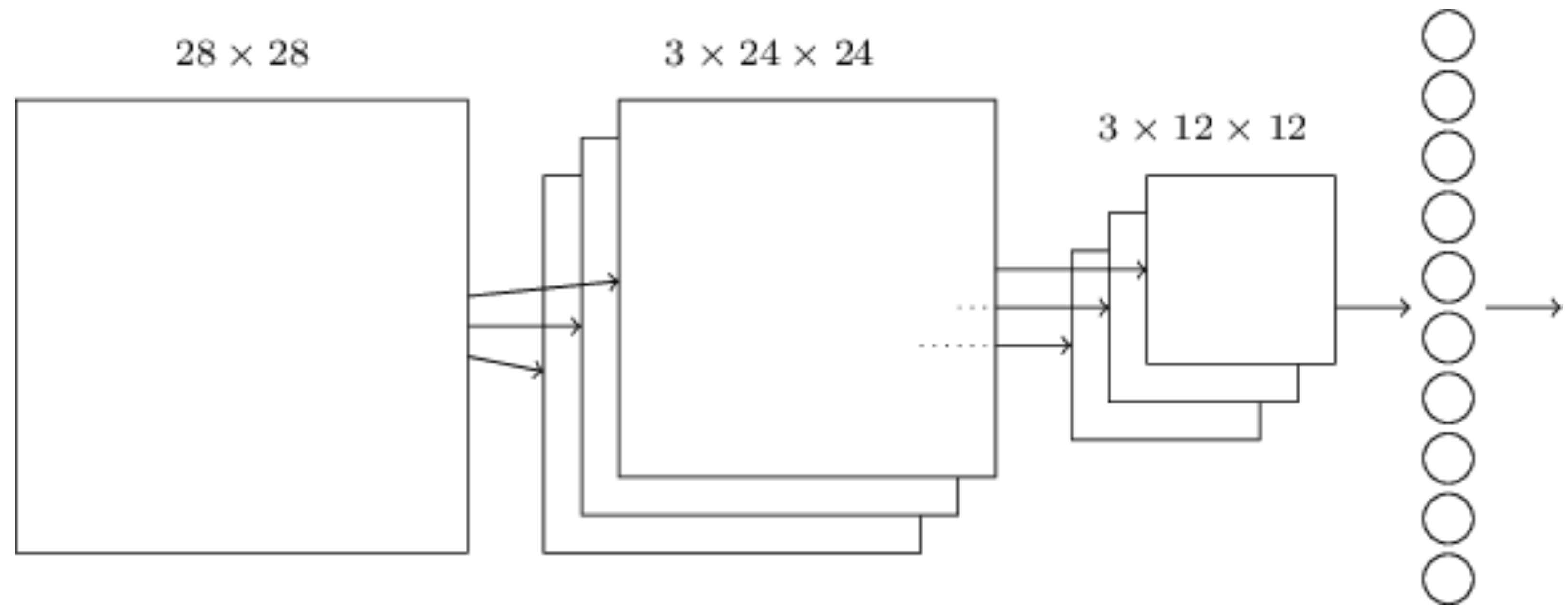


operation:
- **maximum**
- average
- minimum

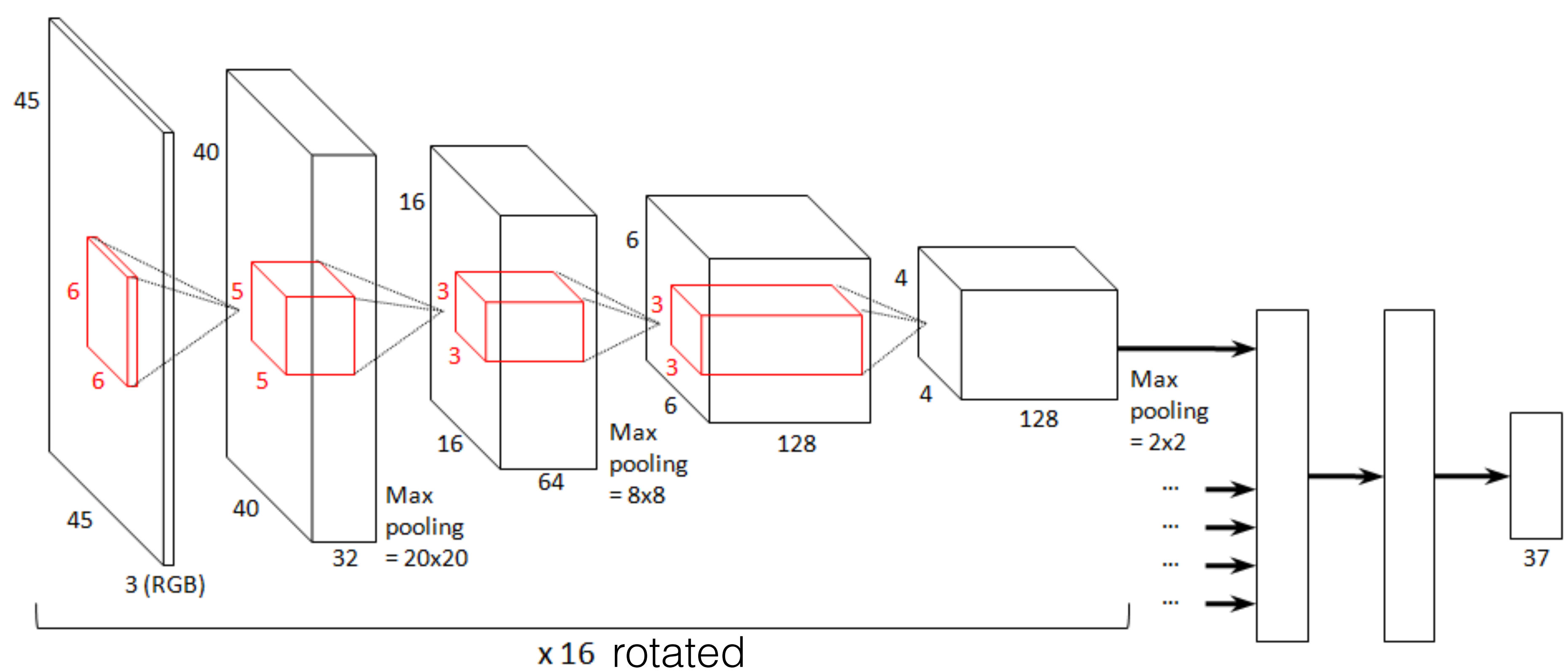
Local translational invariance is achieved via convolution-pool combination

Convolutional ANN architecture

minimal



Galaxy zoo challenge winner



Galaxy Zoo

Manual classification of galaxies using images from Sloan Digital Sky Survey

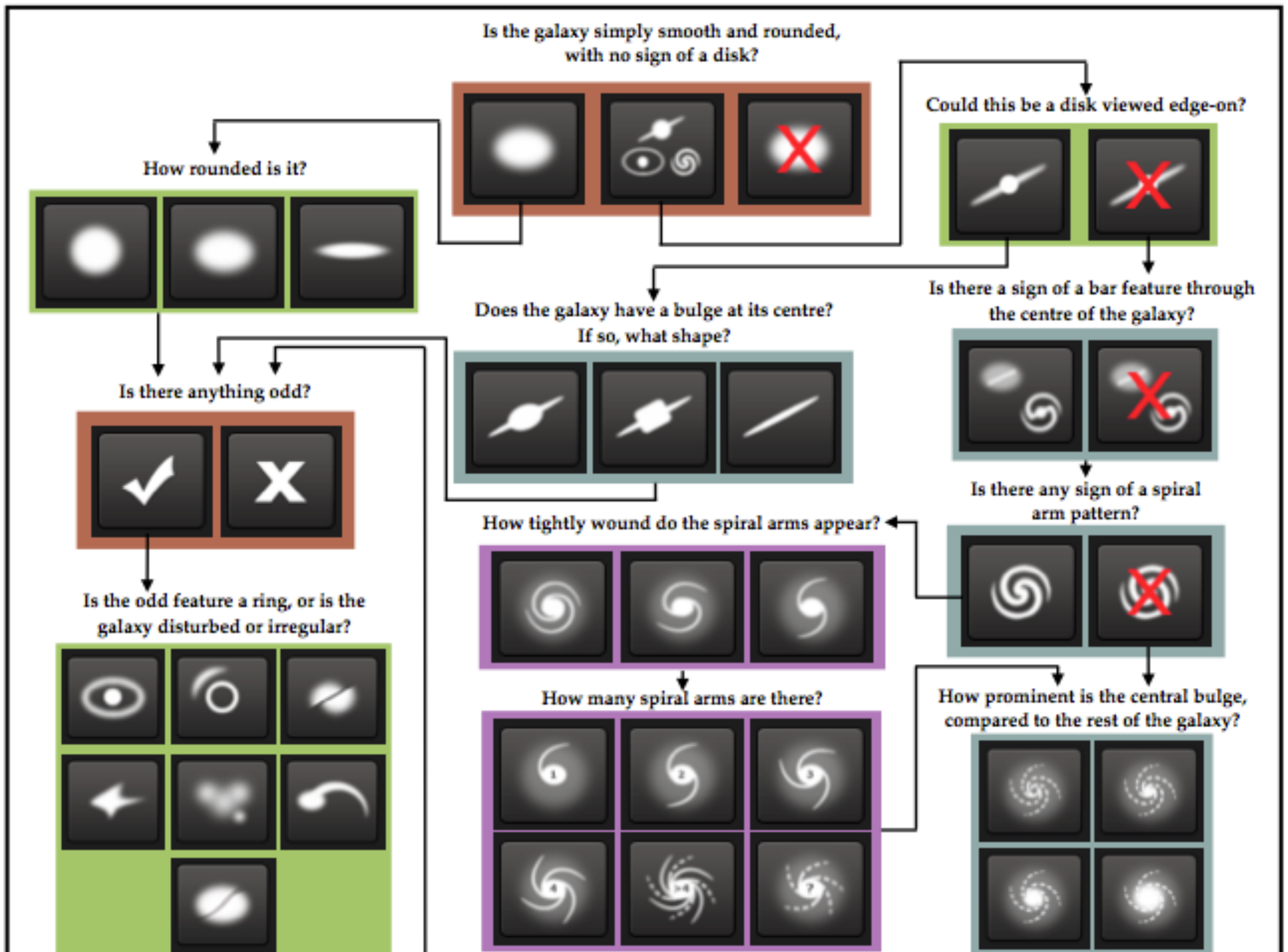


Figure 1. Flowchart of the classification tasks for GZ2, beginning at the top centre. Tasks are colour-coded by their relative depths in the decision tree. Tasks outlined in brown are asked of every galaxy. Tasks outlined in green, blue, and purple are (respectively) one, two or three steps below branching points in the decision tree. Table 2 describes the responses that correspond to the icons in this diagram.

Galaxy Zoo Challenge

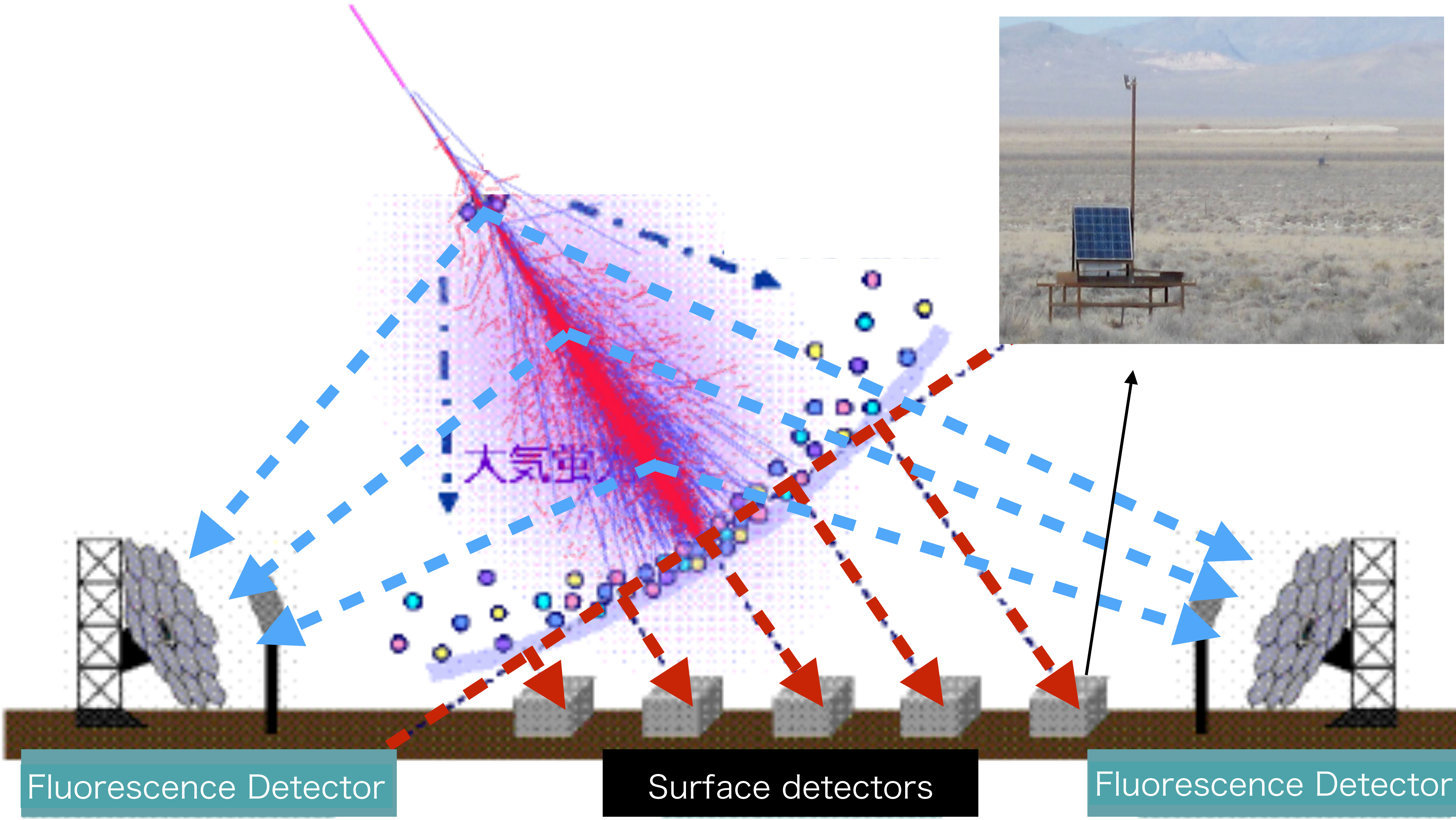
purpose: using image as input, we want to predict how Galaxy Zoo users (zooites) classify the image, i.e. predict vector of probabilities for all classes

training set: color images 424x424 along with vectors of probabilities for more than 60000 galaxies

https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge

Time to open jupyter notebook

UHECR Detection Methods



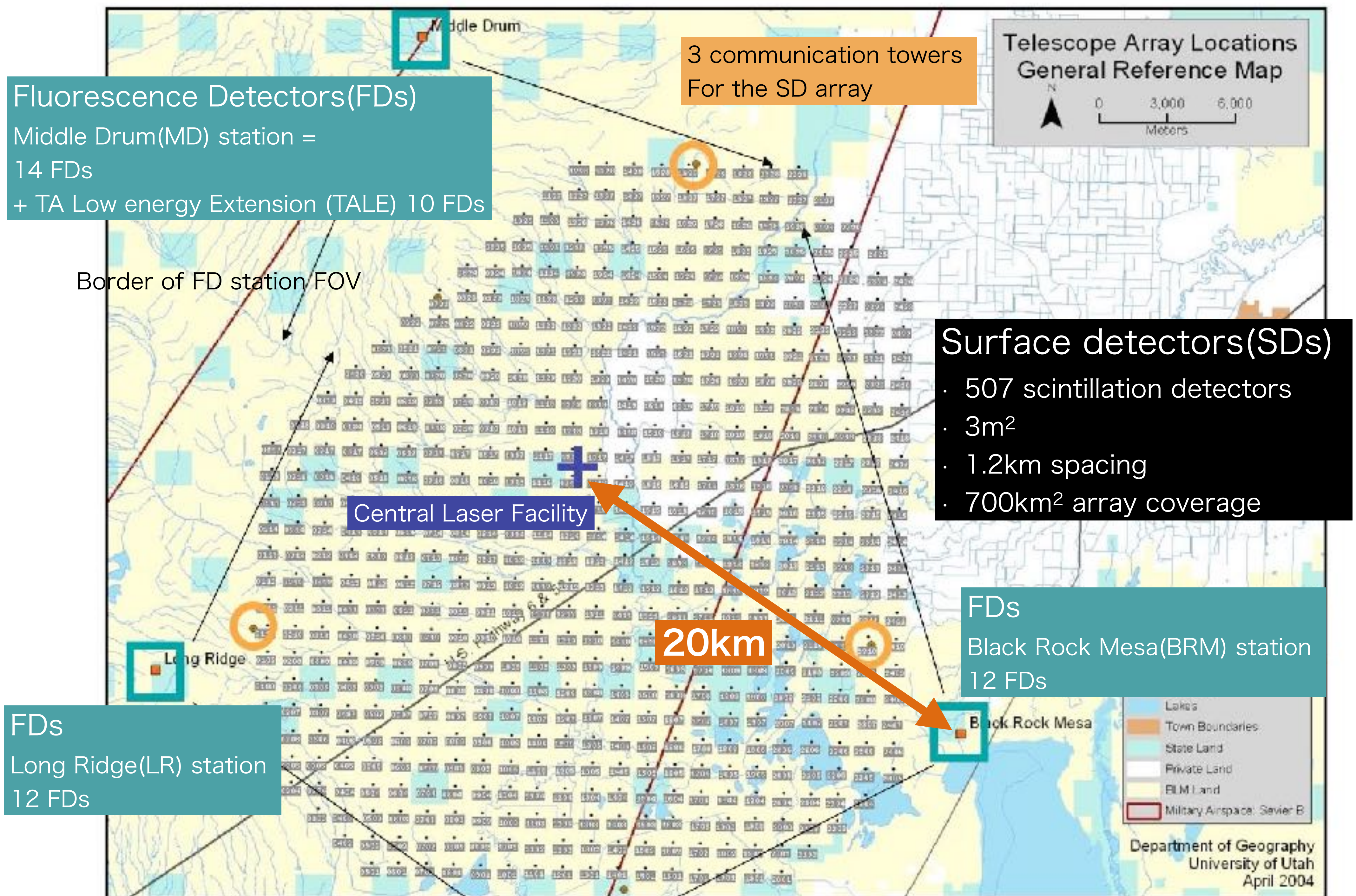
Flourescence detectors:
Duty cycle ~ *10%*

Surface detectors:
Duty cycle ~ *95%*

Telescope Array

- The biggest experiment in the northern hemisphere (Utah, USA).

USA,
Russia,
Japan,
Korea,
Belgium



Event reconstruction

standard parametric approach

- LDF

$$f\left(r\right)=\left(\frac{r}{R_m}\right)^{-1.2}\left(1+\frac{r}{R_m}\right)^{-(\eta-1.2)}\left(1+\frac{r^2}{R_1^2}\right)^{-0.6}$$

$$R_m=90.0\ \mathrm{m},\ R_1=1000\ \mathrm{m},\ R_L=30\ \mathrm{m},\quad \eta=3.97-1.79\left(\sec\left(\theta\right)-1\right),$$

$$r=\sqrt{\left(x\mathrm{core}-x\right)^2+\left(y\mathrm{core}-y\right)^2},$$

- Timing

$$t_r=t_o+t_{plane}+a\times\left(1+r/R_L\right)^{1.5}LDF\left(r\right)^{-0.5}$$

$$LDF\left(r\right)=f\left(r\right)/f\left(800\ \mathrm{m}\right)\qquad S\left(r\right)=S_{800}\times LDF\left(r\right)$$

Free parameters:

$x_{core},\ y_{core},\ \theta,\ \phi,\ S_{800},\ t_0,\ a$

Observables:

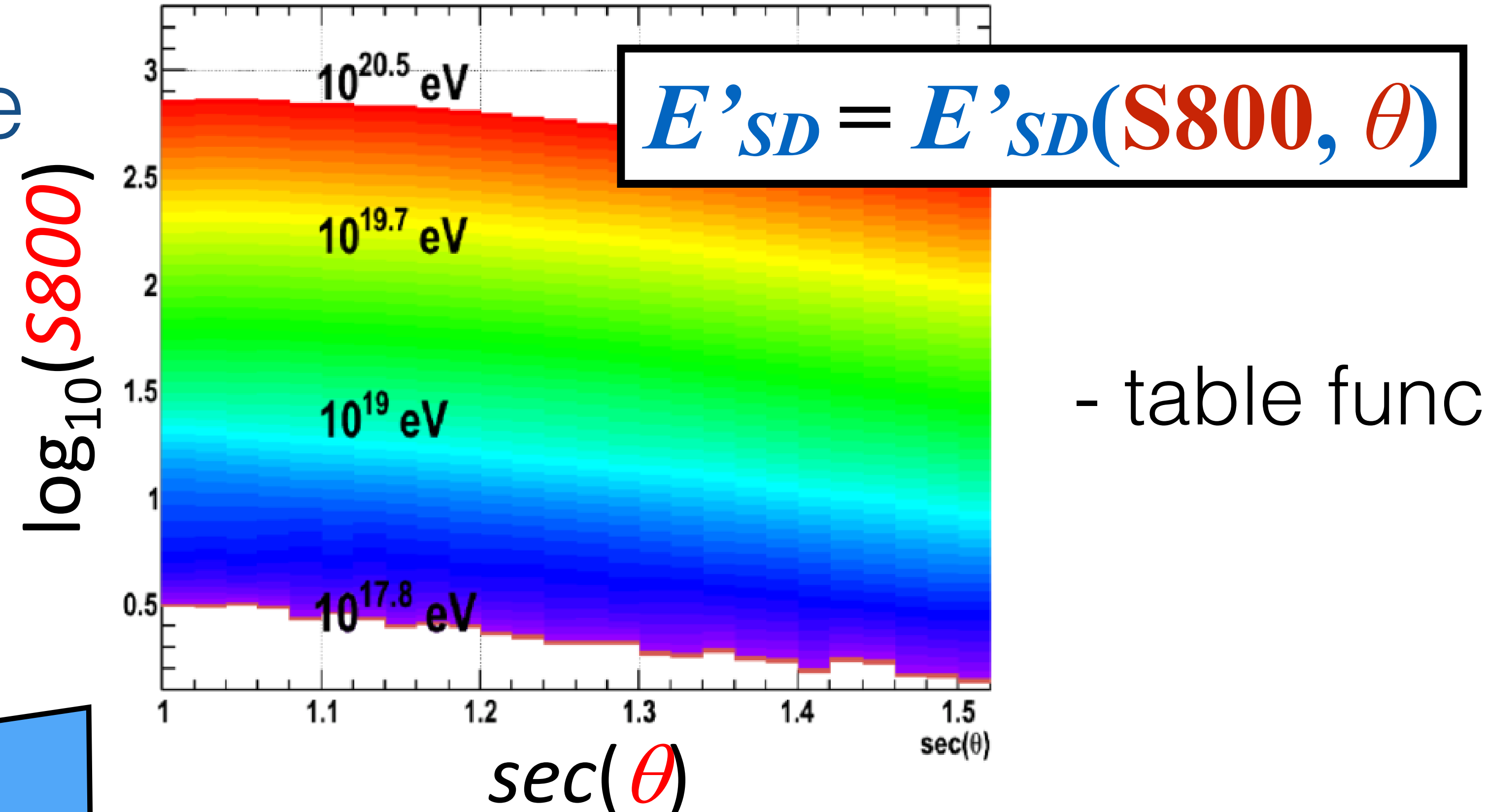
t_r - detector time

S_r - detector integral signal

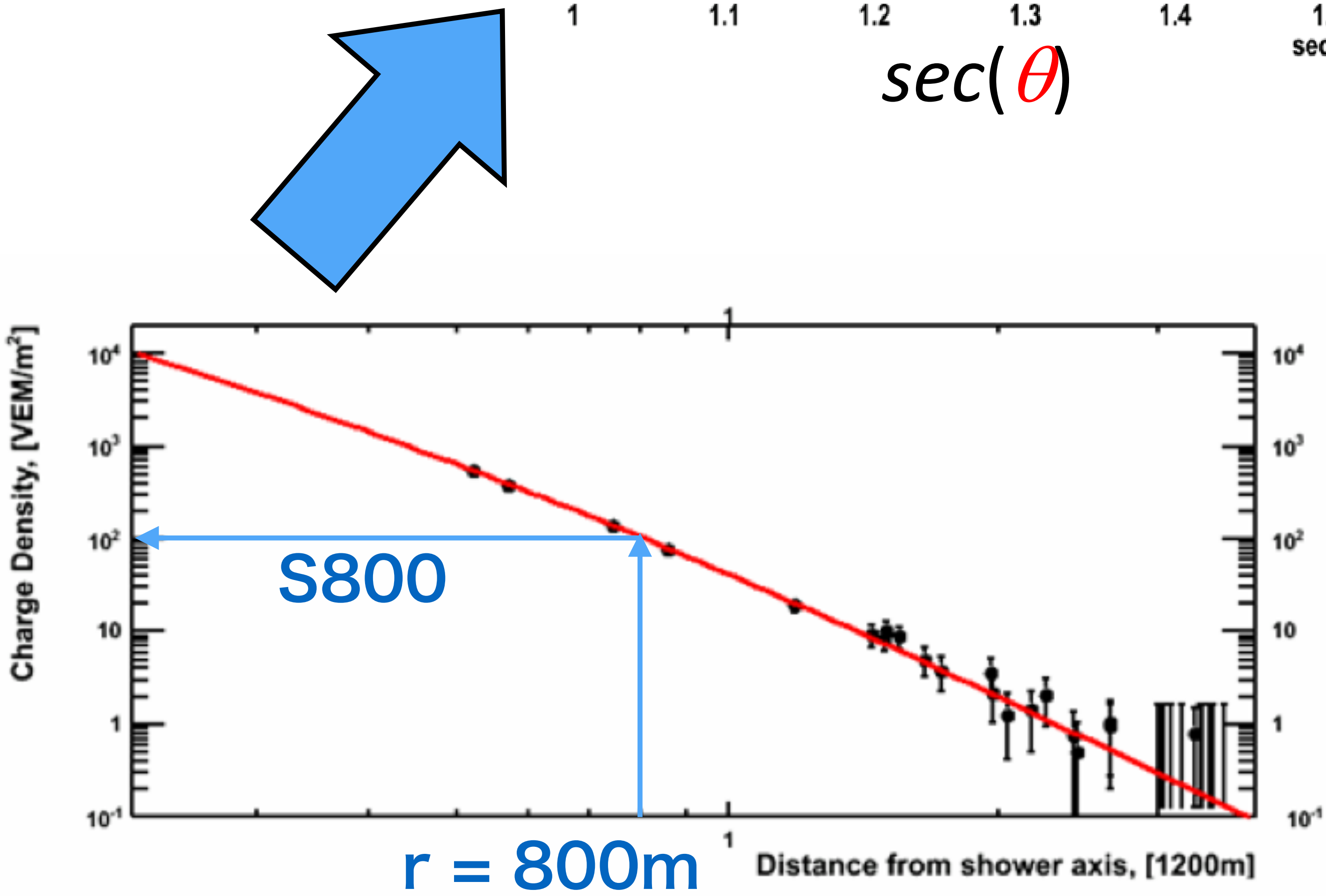
Event reconstruction

standard parametric approach

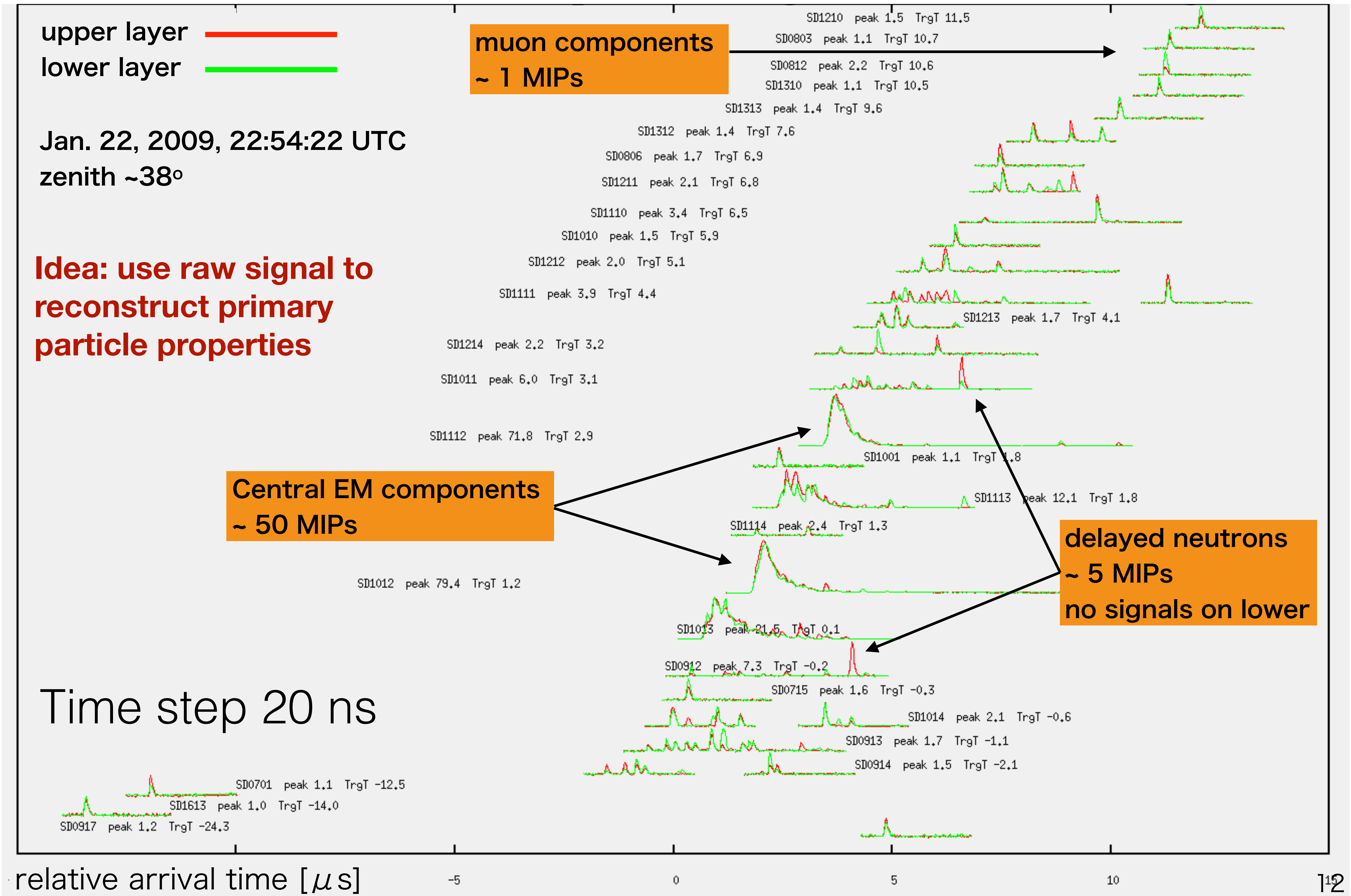
Energy estimate



- table function



Sample event



SD reconstruction NN architecture

event data

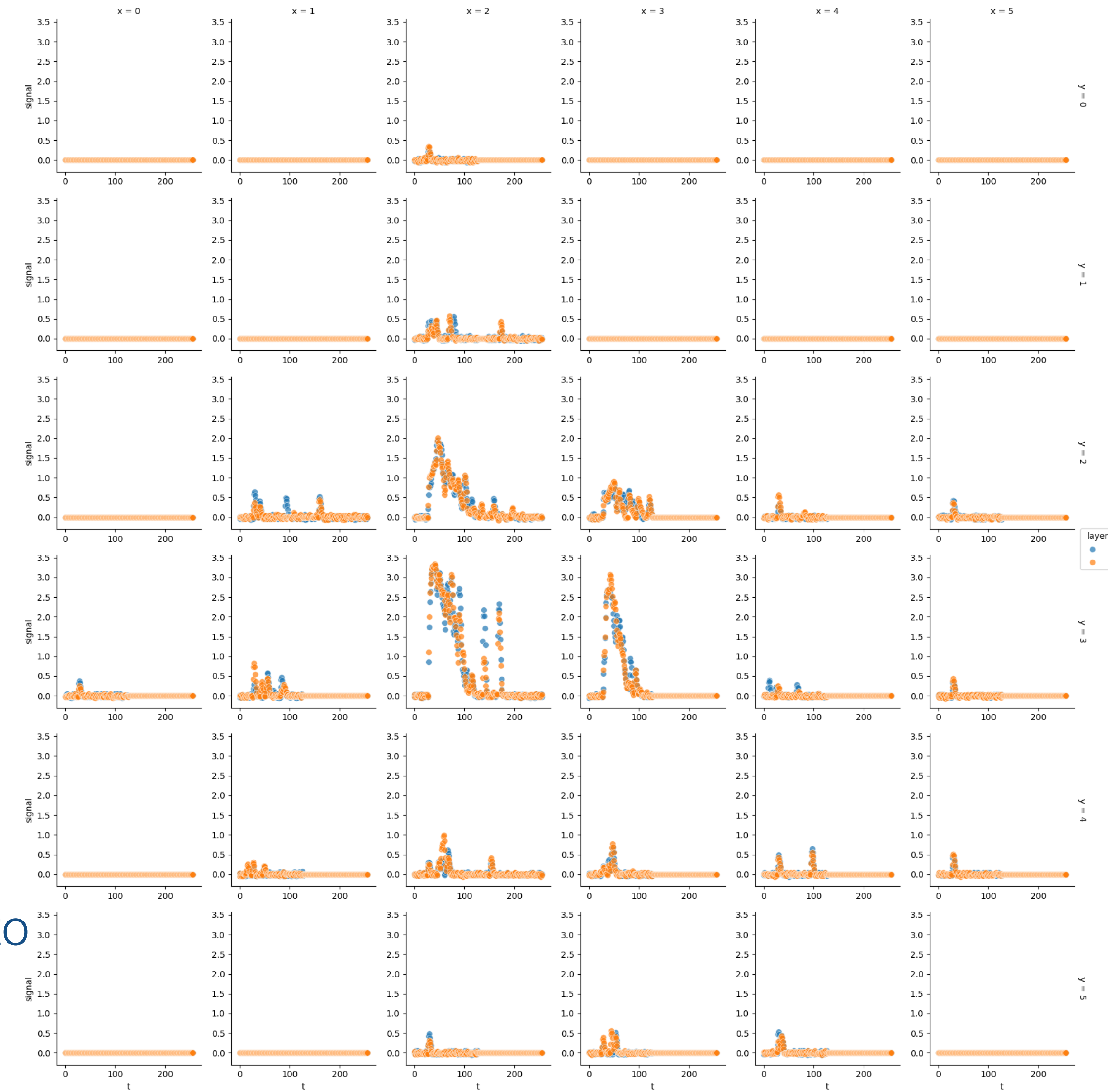
Dimensions:
(N,N,T,2)

Waveform

detector
layers

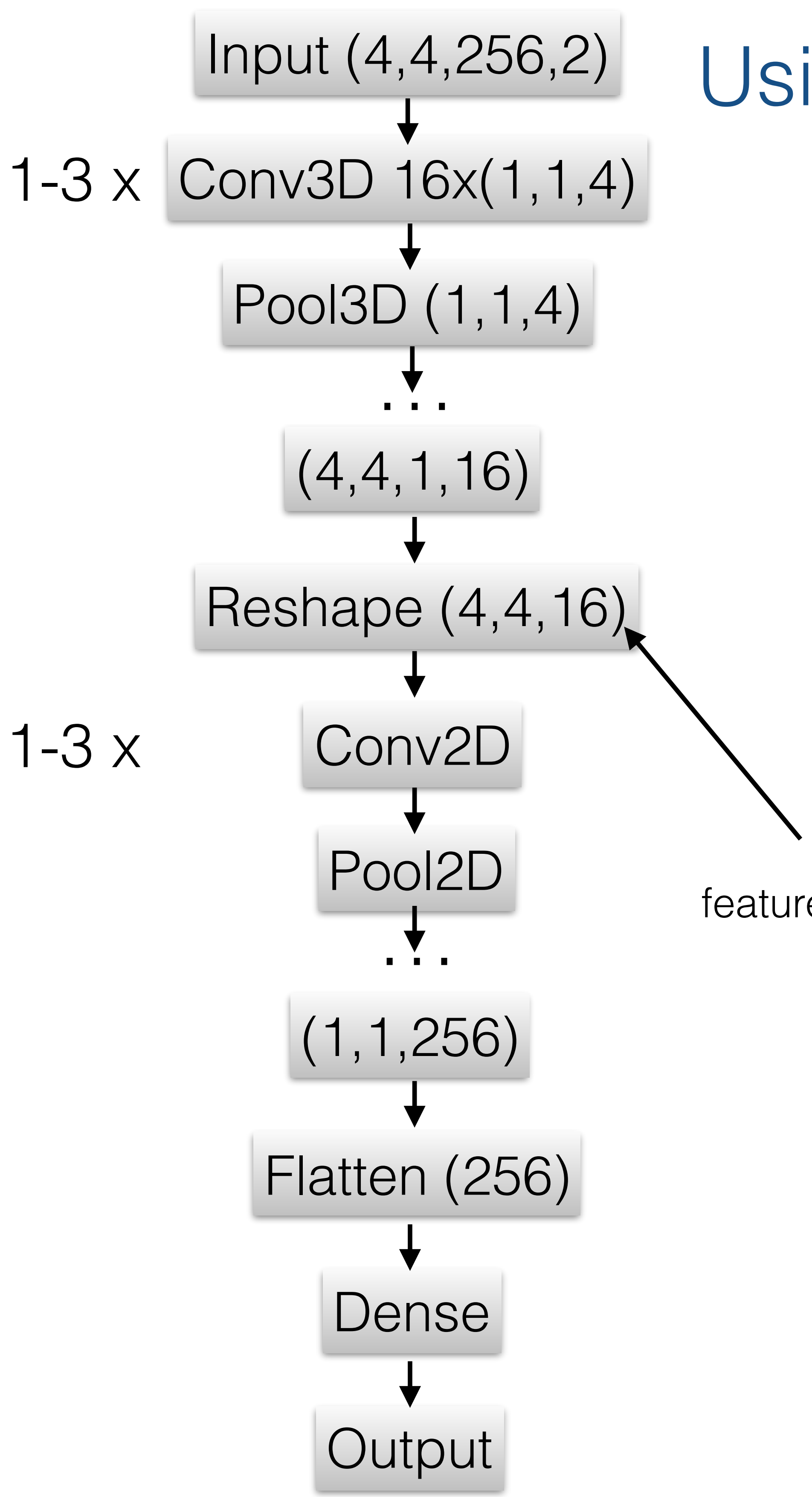
N=4-8, T=128-256

Standard SD reconstruction is used to center image around shower core



SD reconstruction NN architecture

Using 1D + 2D convolutions

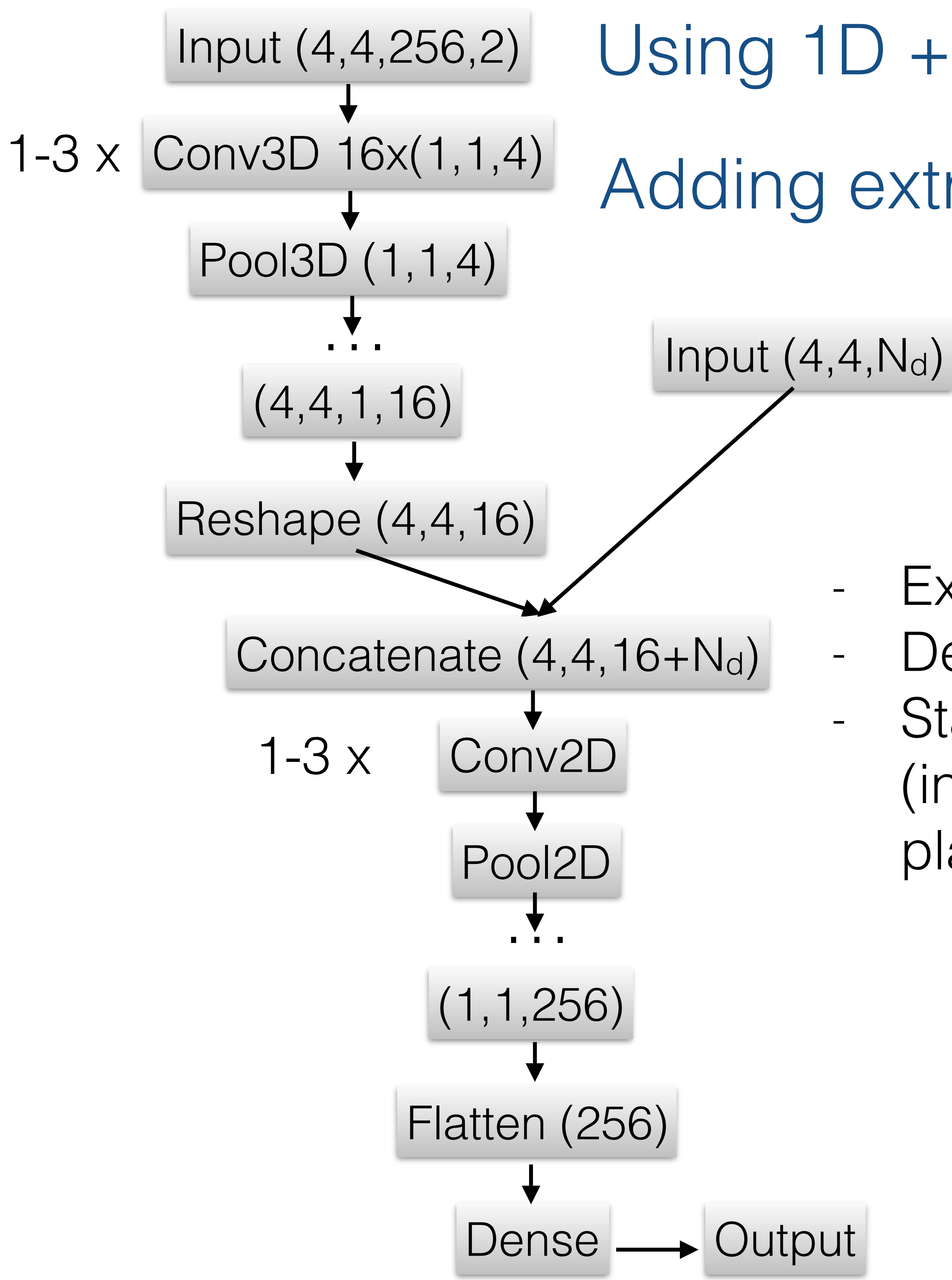


feature vectors for each detector

SD reconstruction NN architecture

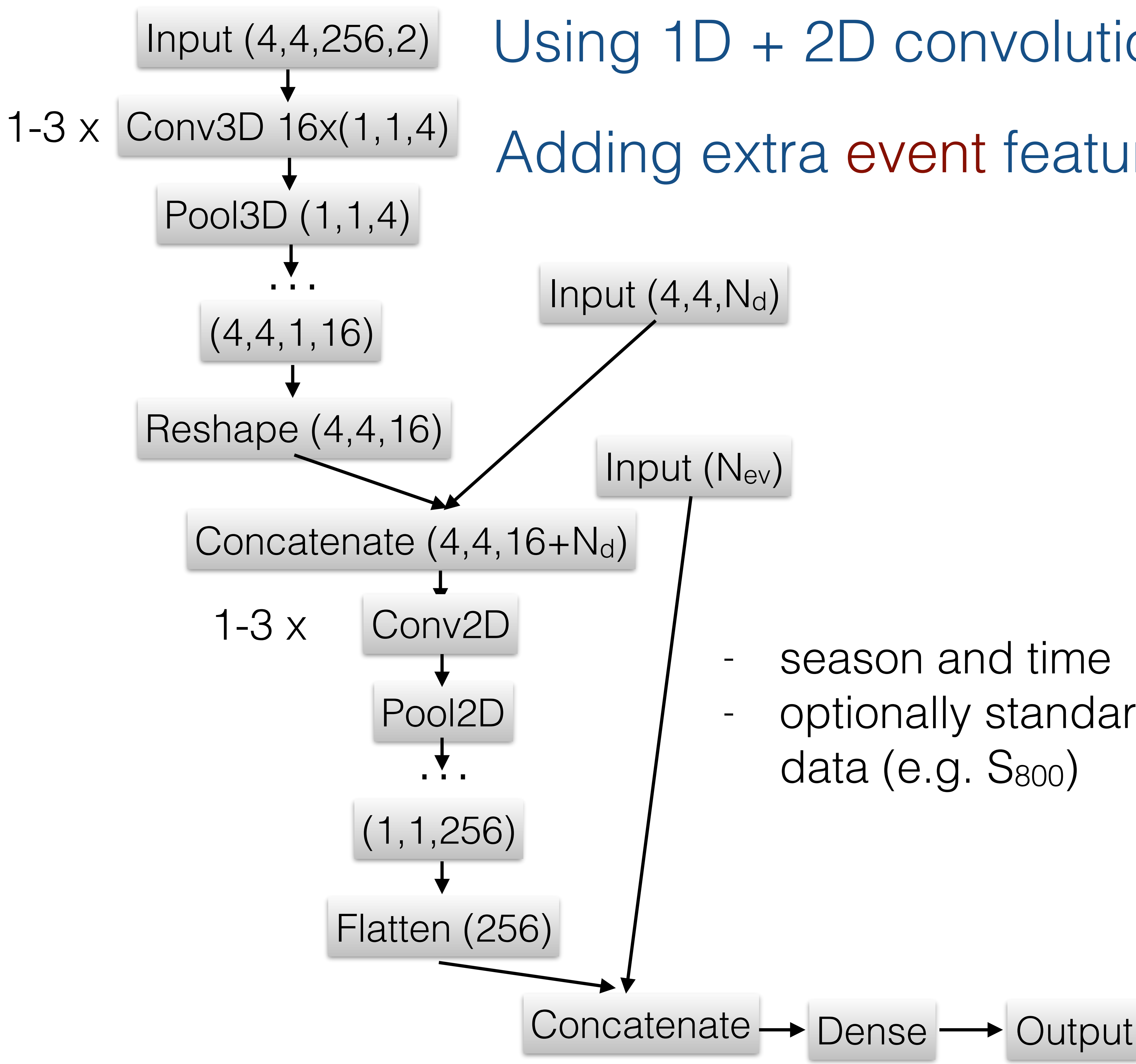
Using 1D + 2D convolutions

Adding extra detector features:



- Exact detector position
- Detector state (on/off/saturated)
- Standard reconstruction parameters (integral signal, timing relative to plane front)

SD reconstruction NN architecture

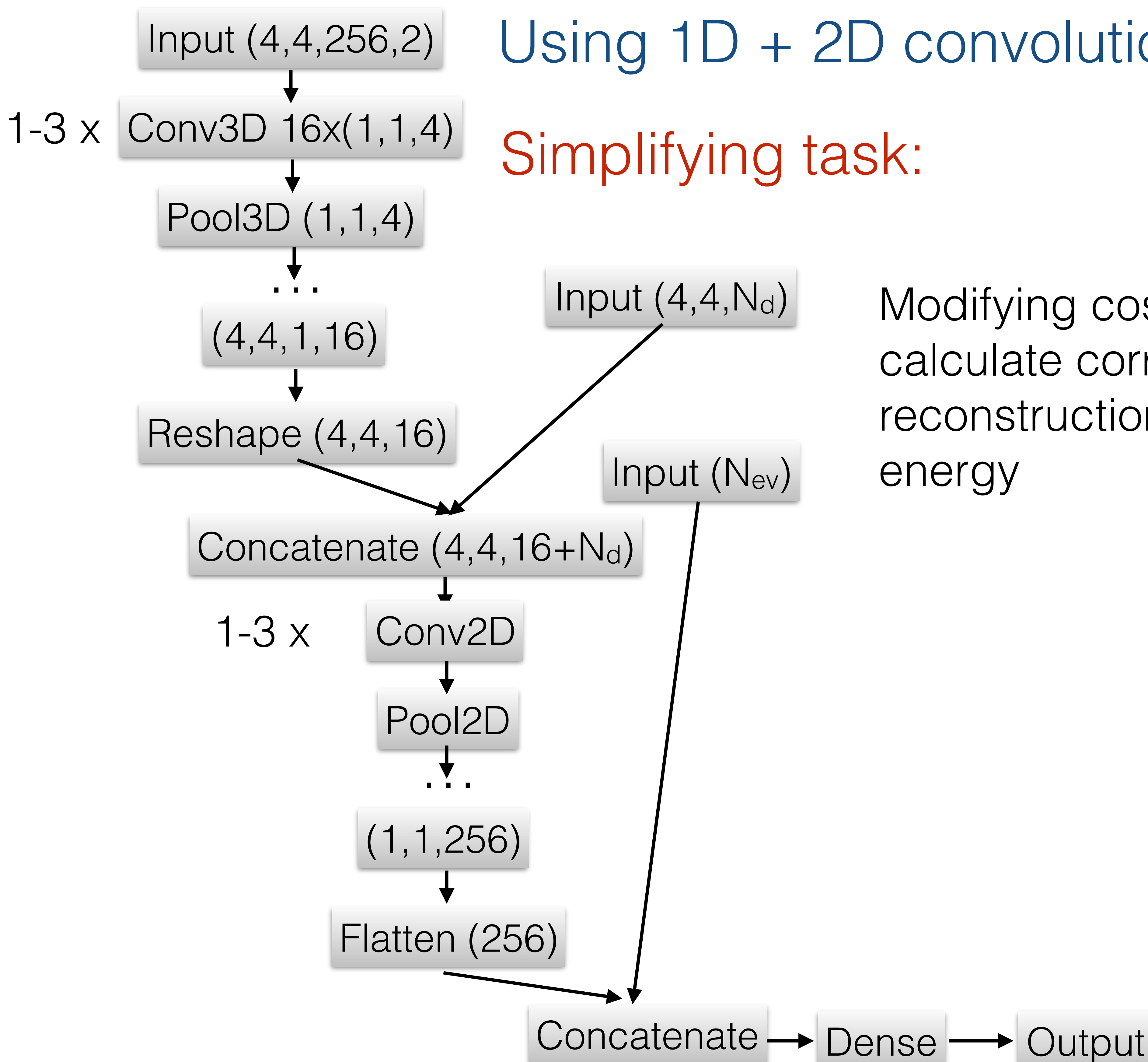


Using 1D + 2D convolutions

Adding extra event features:

- season and time
- optionally standard reconstruction data (e.g. S_{800})

SD reconstruction NN architecture



Using 1D + 2D convolutions

Simplifying task:

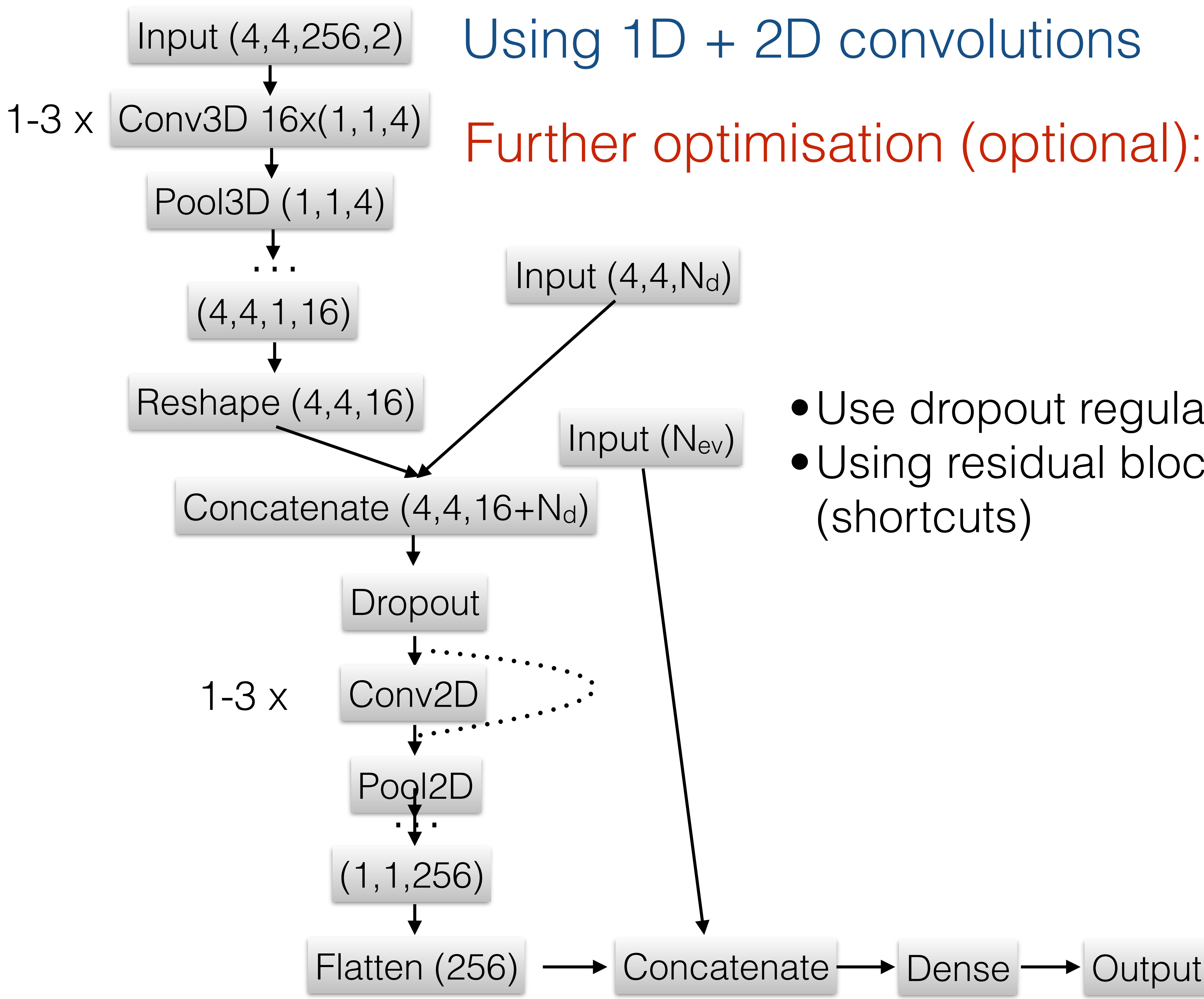
Modifying cost function - we calculate correction to standard reconstruction for angles and energy

SD reconstruction NN architecture

Using 1D + 2D convolutions

Further optimisation (optional):

- Use dropout regularization
- Using residual blocks (shortcuts)



Training the model

- Minimizing mean square error
- Adaptive learning rate (adadelta optimizer arxiv 1212.5701)
- Number of training samples ~ 10^6 (100 GB data) - do not fit into RAM). hdf container is used и generator API in keras
- Number of weights to learn 10^5 - 10^6
- Regularization to avoid overfitting:
  - L2
  - dropout
  - noise layers
- Optimizing network architecture hyper-parameters (hyperopt package)
- Hardware: NVIDIA GTX-1080-Ti GPU
- Instruments: python, numpy, tensorflow, keras, h5py

How to see that model does job

in presence of unavoidable uncertainty

Explained variance score

$$EV(y, \hat{y}) = 1 - \frac{Var(y - \hat{y})}{Var(y)}$$

y - true value of quantity being predicted (in our case, error of parametric reconstruction)

$\hat{y}$ - model estimate of y

How to see that model does job

in presence of unavoidable uncertainty

Explained variance score

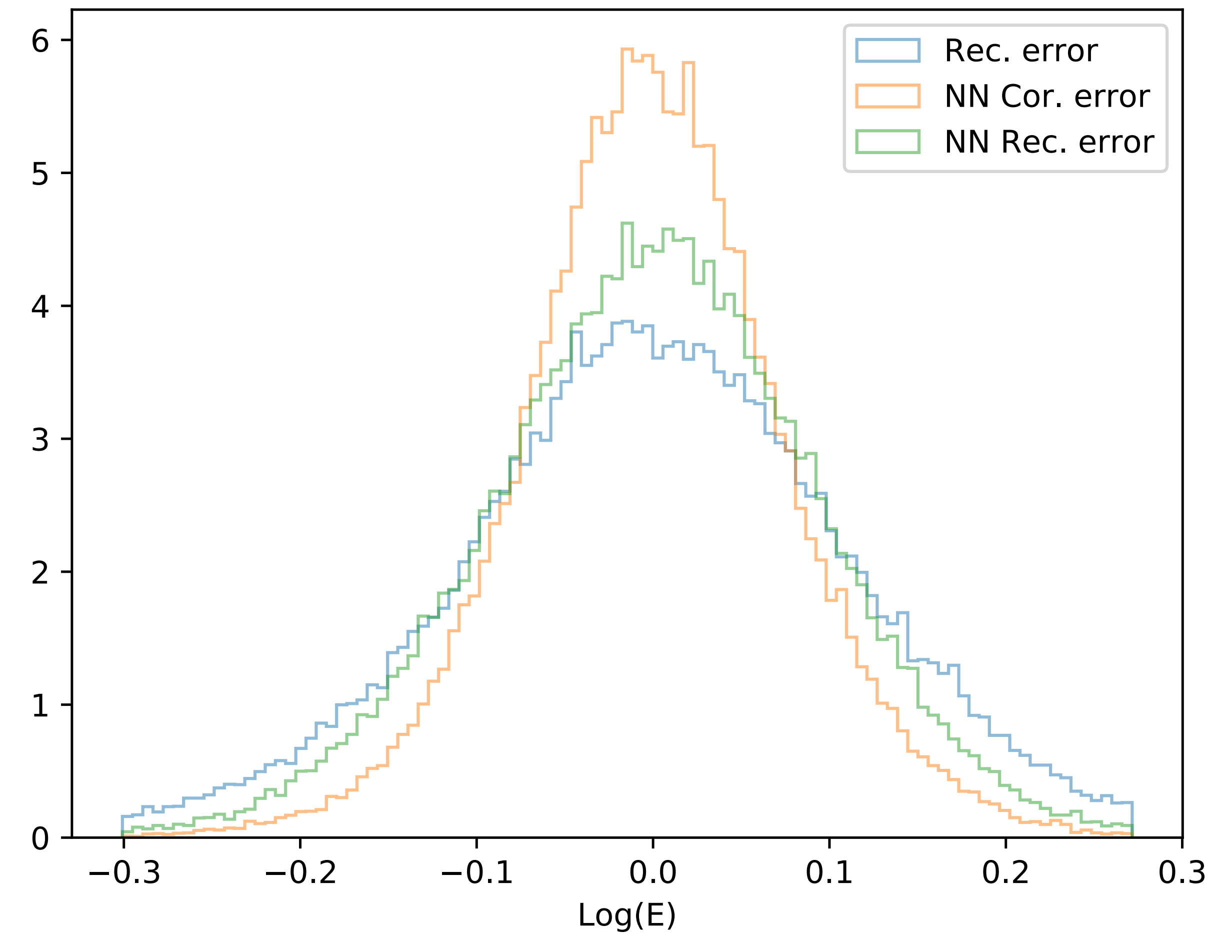
$$EV(y, \hat{y}) = 1 - \frac{Var(y - \hat{y})}{Var(y)}$$

y - true value of quantity being predicted (in our case, error of standard reconstruction)

$\hat{y}$ - model estimate of y

More visually:

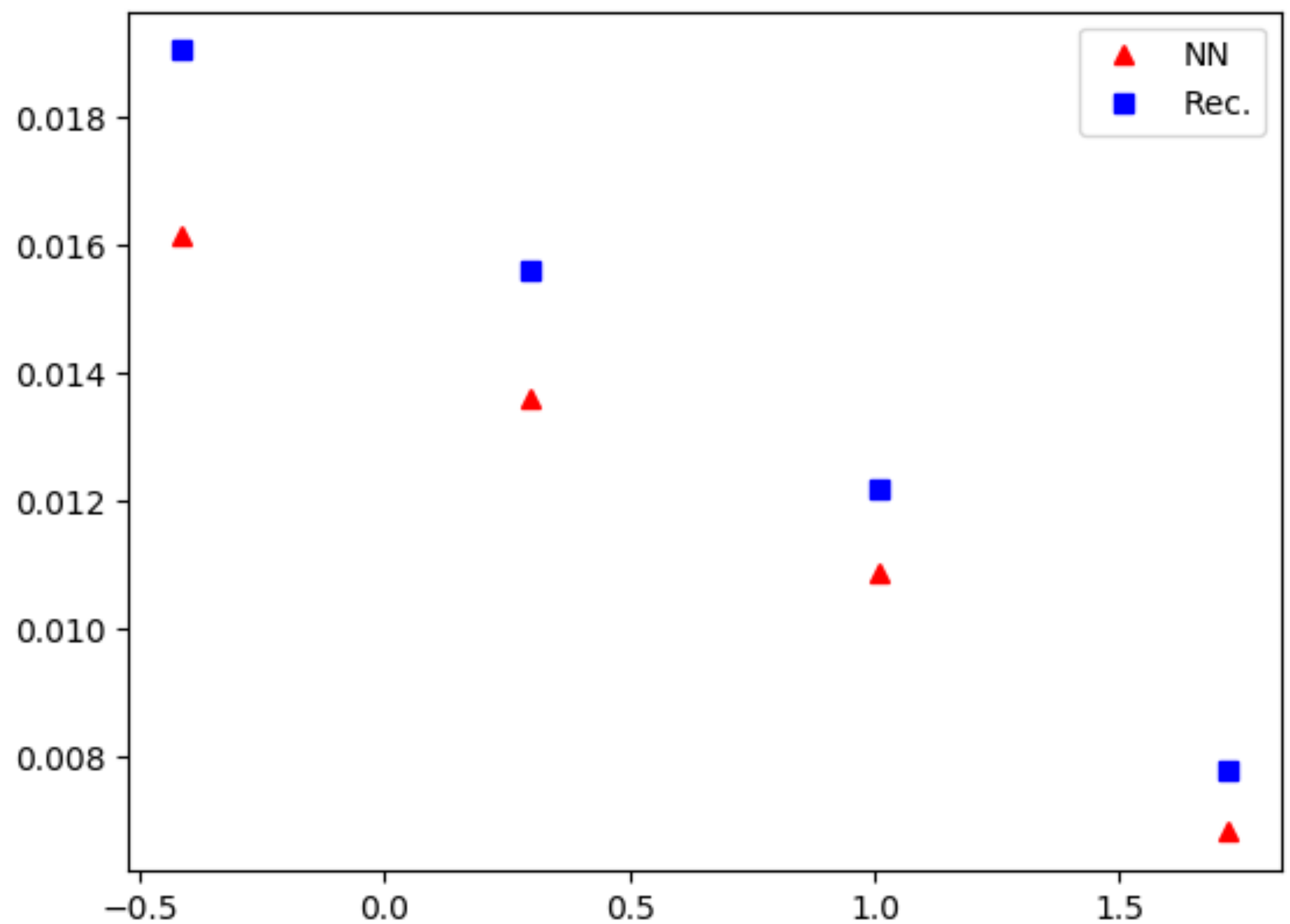
Compare error distribution in two approximations



Preliminary results

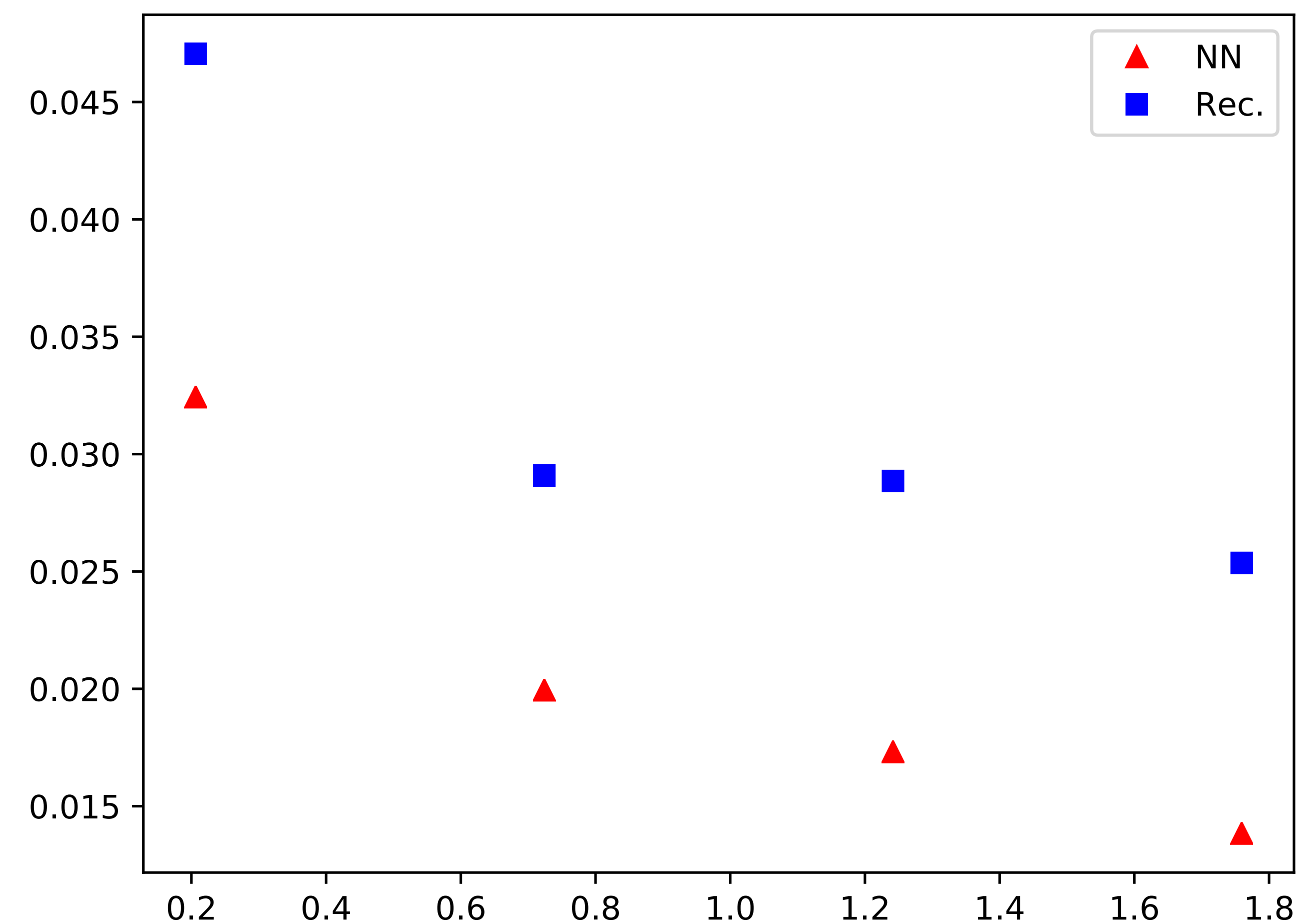
Zenith angle reconstruction errors

$\Delta cos(\theta)$ nuclei



$log(E/EeV)$

$\Delta cos(\theta)$ photons

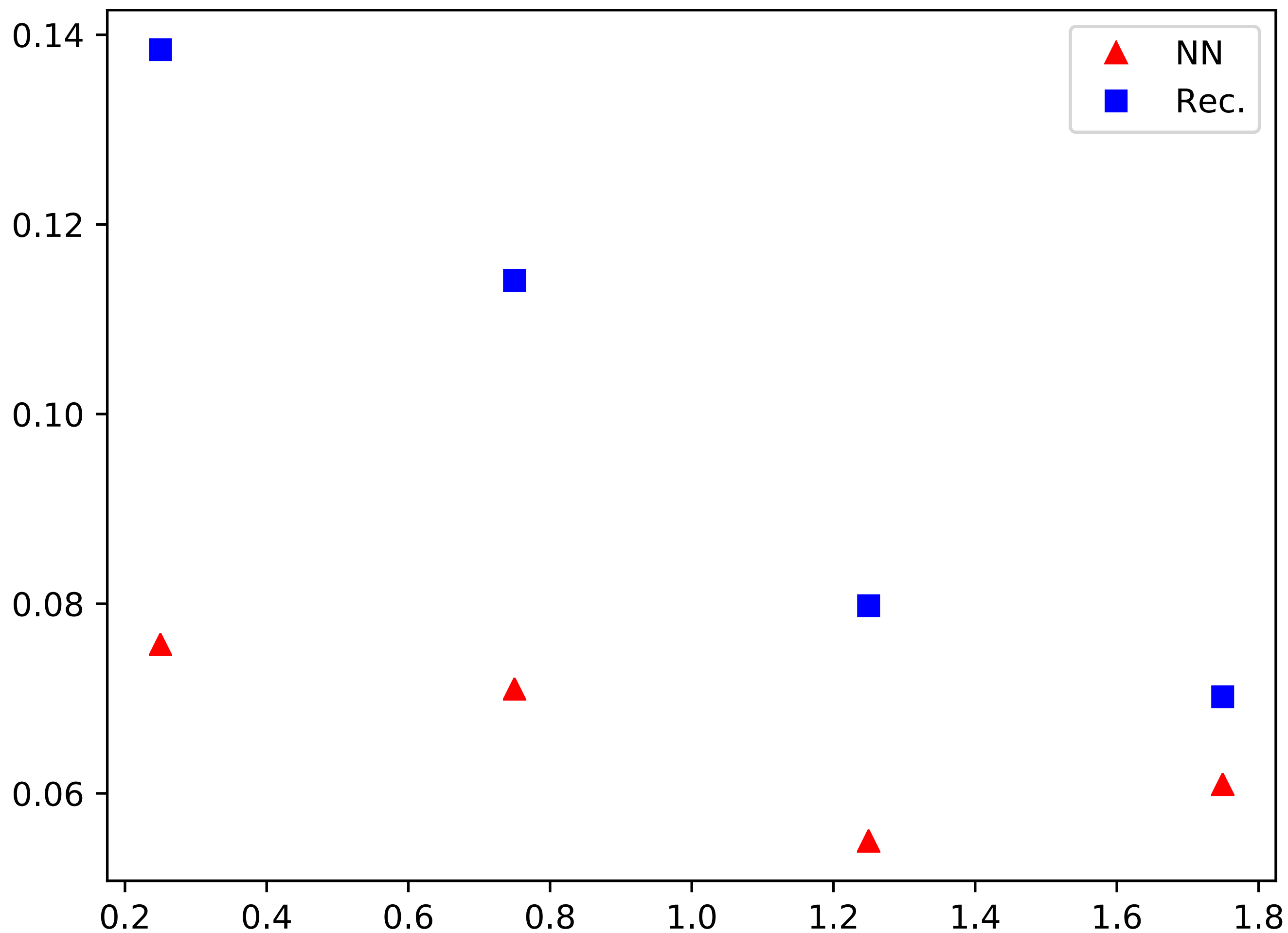


$log(E/EeV)$

Preliminary results

Energy reconstruction errors (nuclei primaries)

$\Delta E/E$



$log(E/EeV)$

EAS modelling

- MC: CORSIKA
- HE hadronic interactions: QGSJETII-03 (QGSJETII-04 in preparation)
- LE hadronic interactions: FLUKA
- EM processes: EGS4
- Detector response: GEANT4
- Event sampling:
    - Energy sampling E^{-1}
    - Mass composition: H, He, N, Fe (1:1:1:1)
    - Isotropic primary flux with zenith angles < 45 degrees
    - Standard energy spectrum reconstruction cuts applied