PPC 2018: XII International Conference on Interconnections between Particle Physics and Cosmology

August 20 - 24, 2018 Zurich, Switzerland

Machine Learning: from Industry to Science

Fedor Ratnikov

NRU Higher School of Economics,
Yandex School of Data Analysis

Content

- Machine Learning basics: Generic Models
- ML implementations
- ML applications
- Few success stories
- From industry to science

Specific vs Generic Models

- Specific (fundamental) model
  - paradigm: first model, then data
  - assume some specific dependency *a priori* (e.g. a Nature low)
    - “fit parabola to data”
  - number of degrees of freedom corresponds to the problem
    - smaller dimensions
      - faster and more stable converging
  - extrapolatable beyond the train area
  - interpretable representation
  - can not accommodate difference between parametric model and actual data

Specific vs Generic Models

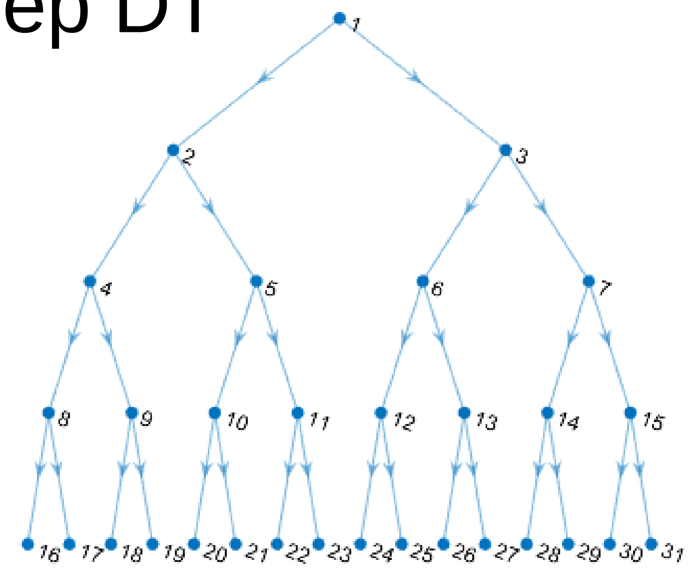
- Generic model (paradigm: first data, then model)
  - assumes nothing about dependency *a priori*
  - “find reasonable representation for data”
    - model ≡ data
  - universal - can accommodate most actual data
  - significantly over-defined problem
    - special regularisations are needed
    - less stable converging
    - significant computing resources needed
      - many speed up tricks are developed
  - extrapolation can not be trusted at all
  - hard to interpret
- Machine Learning is all about Generic Models

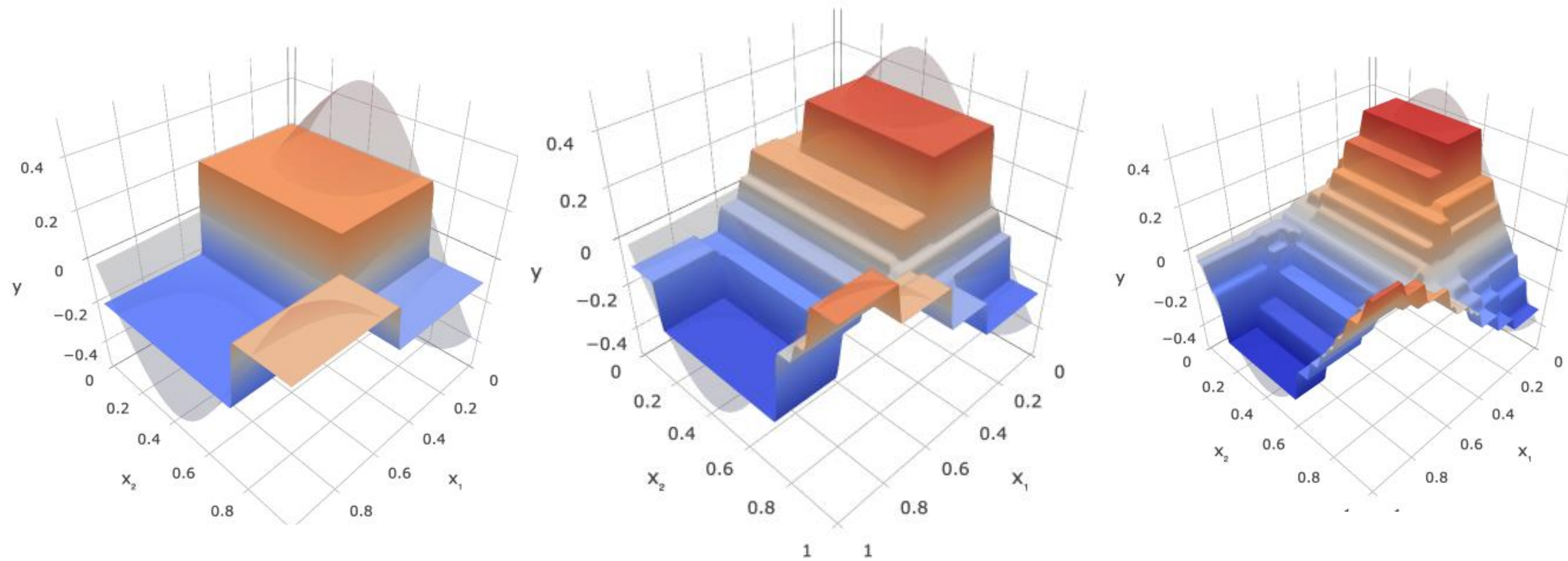
ML Basic Concept

- Machine Learning is a technical way to build generic model for the given dataset
  - inputs, outputs or goals may be very different though
    - classification
    - pattern recognition
      - text, speech, vision, etc.
    - new object generation
    - action control
      - games, driving, etc.
    - …
- Key requirement: there is a way to effectively train the model

Generic Model: Decision Trees

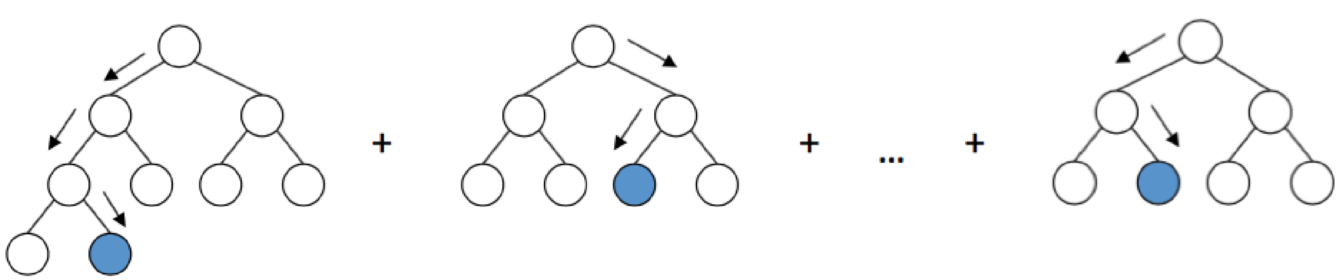
Deep DT

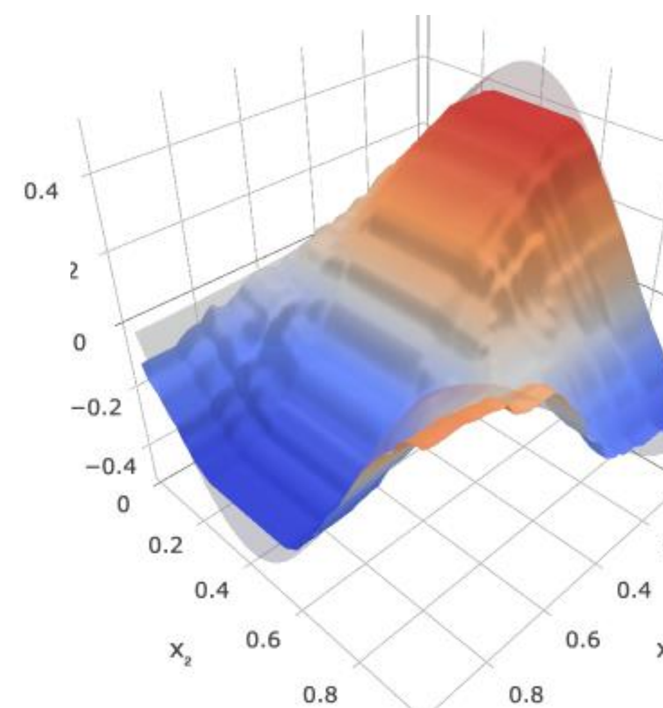




depth=2,4,6 DT

Shallow Boosted DT

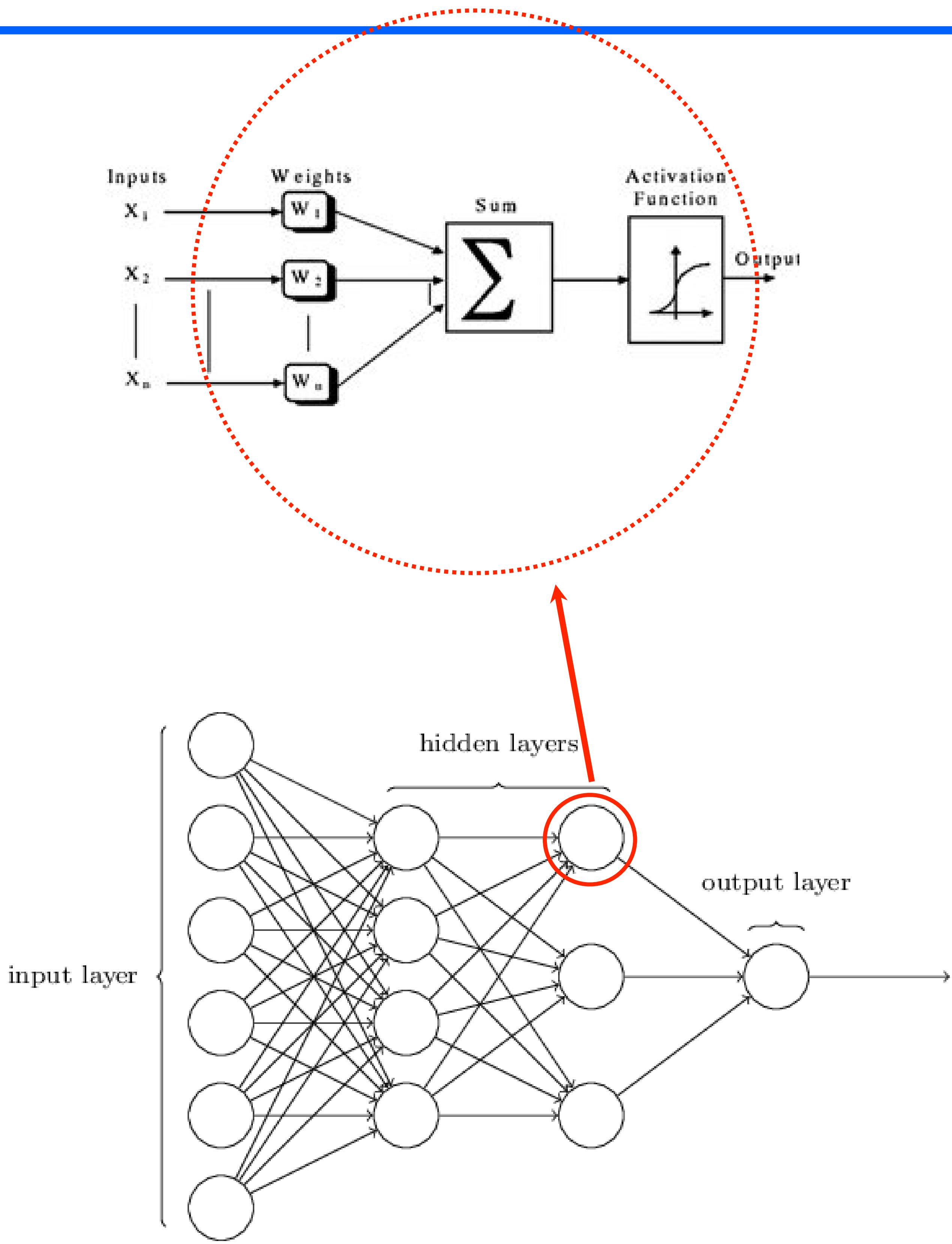


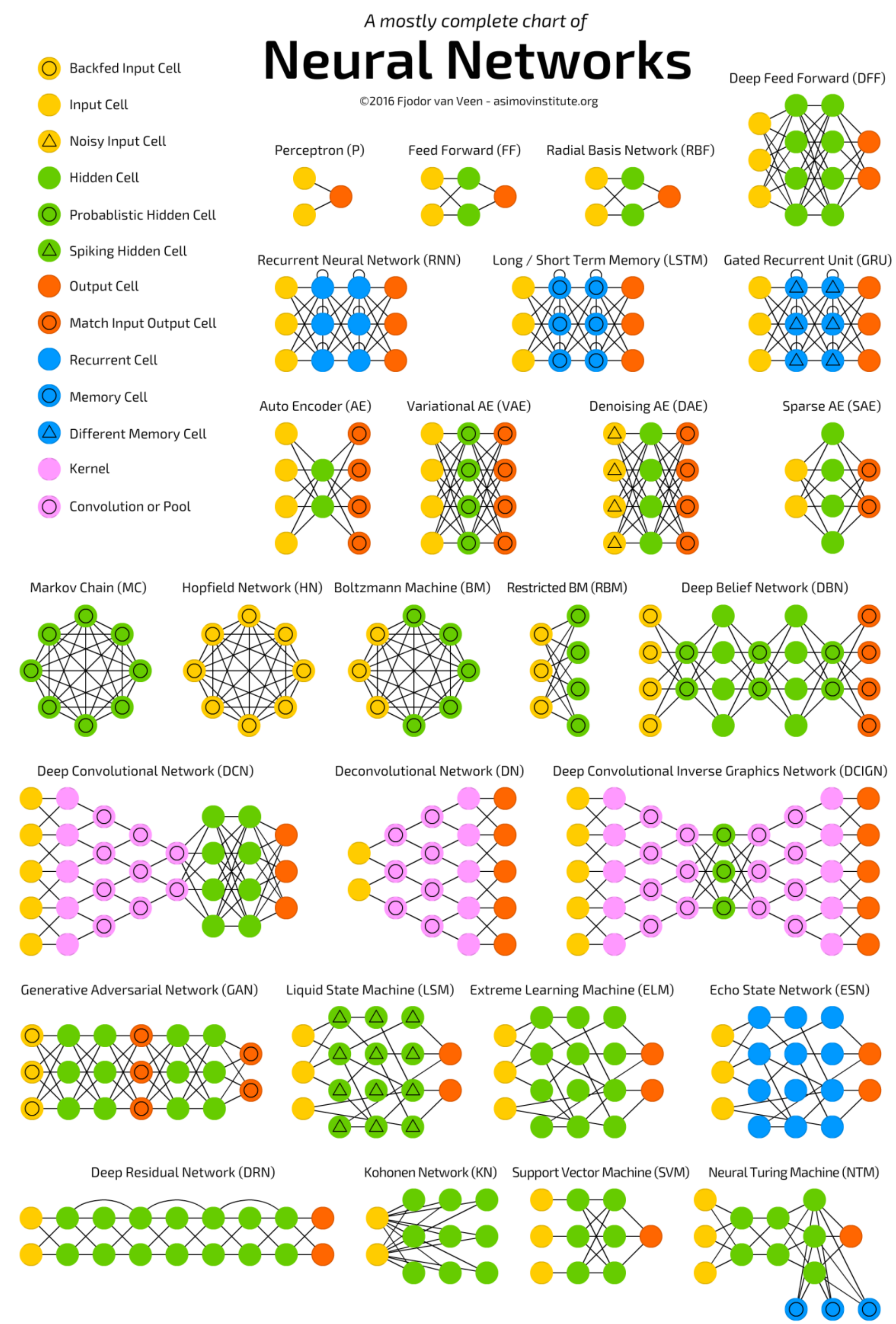


100× depth=3 BDT

http://arogozhnikov.github.io/2016/06/24/gradient_boosting_explained.html

Generic Models: Artificial Neural Networks





https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f236746

BDT vs NN

- Both have similar performance
  - as after meaningful training both represent the dataset properties
    - actual winner depends on data specifics
- NN are more flexible for building models beyond classification problems
  - special architecture (connections)
  - special additions to loss training function

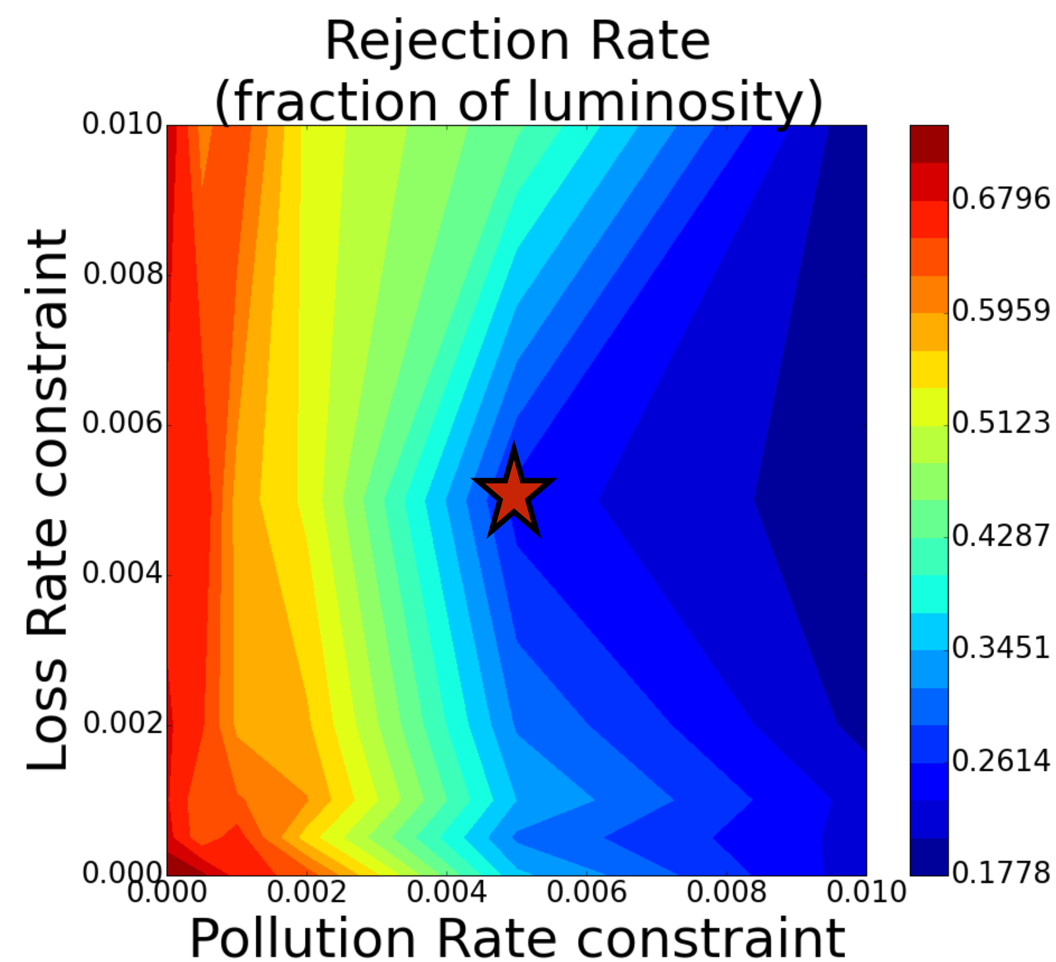
Types of Learning

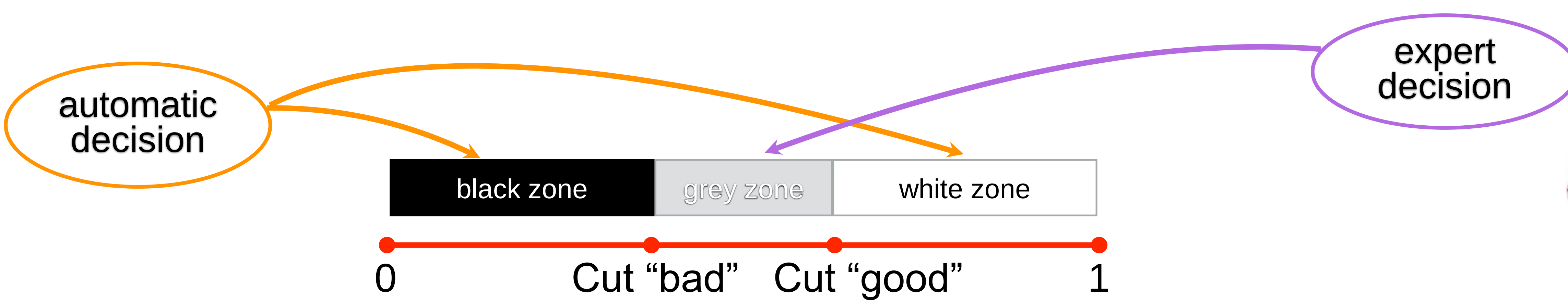
- There is well labelled data
  - we want to learn how to produce labels for more data like this
    - automation of routine
      - supervised learning
- There is labelled data, and unlabelled data which are assumed to be similar
  - we want to derive labels
    - propagation of knowledge
      - supervised learning
- There is unlabelled data with some properties
  - we want to infer some properties of this data
    - inferring new knowledge
      - unsupervised learning

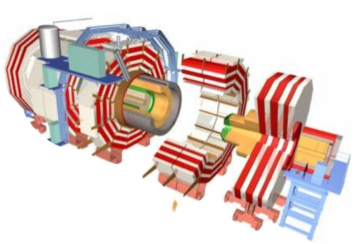
(Automation of) Data Quality Monitoring

- Data quality monitoring is boring and time consuming work
  - Extremely important to guarantee solid physics results
  - Different properties of high level physics objects are analysed
    - photons, muons, calorimeter jets, combined (particle flow) jets
      - kinematics, vertices distribution
        - ~2500 features in total
  - Build classifier to decide if data are good or bad

ManualWork = RejectionRate



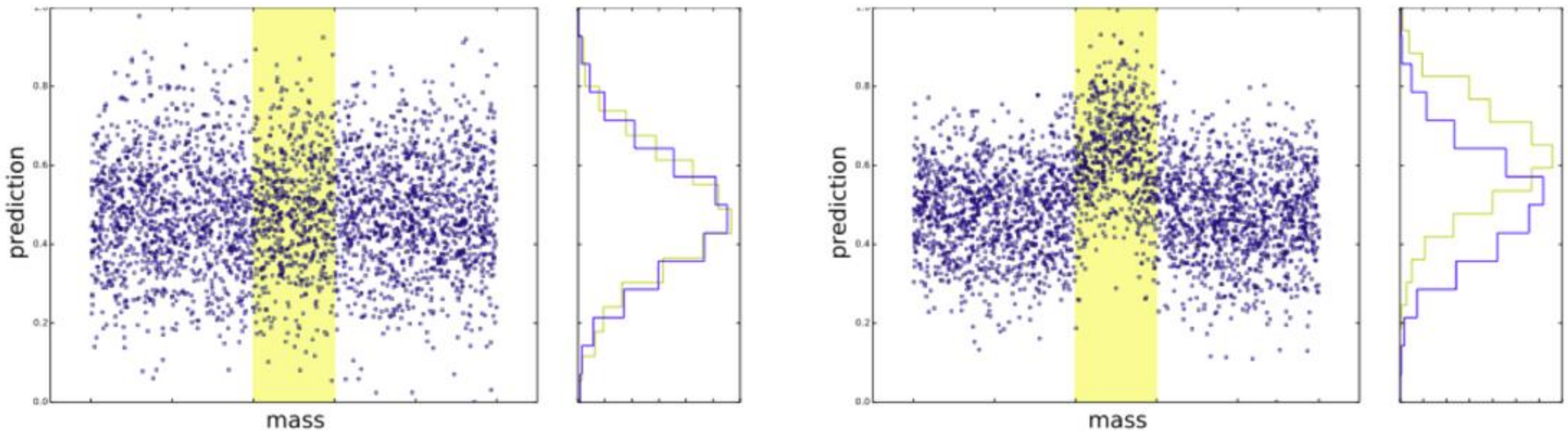




Caveat: how good the expert is?

Propagation of Knowledge

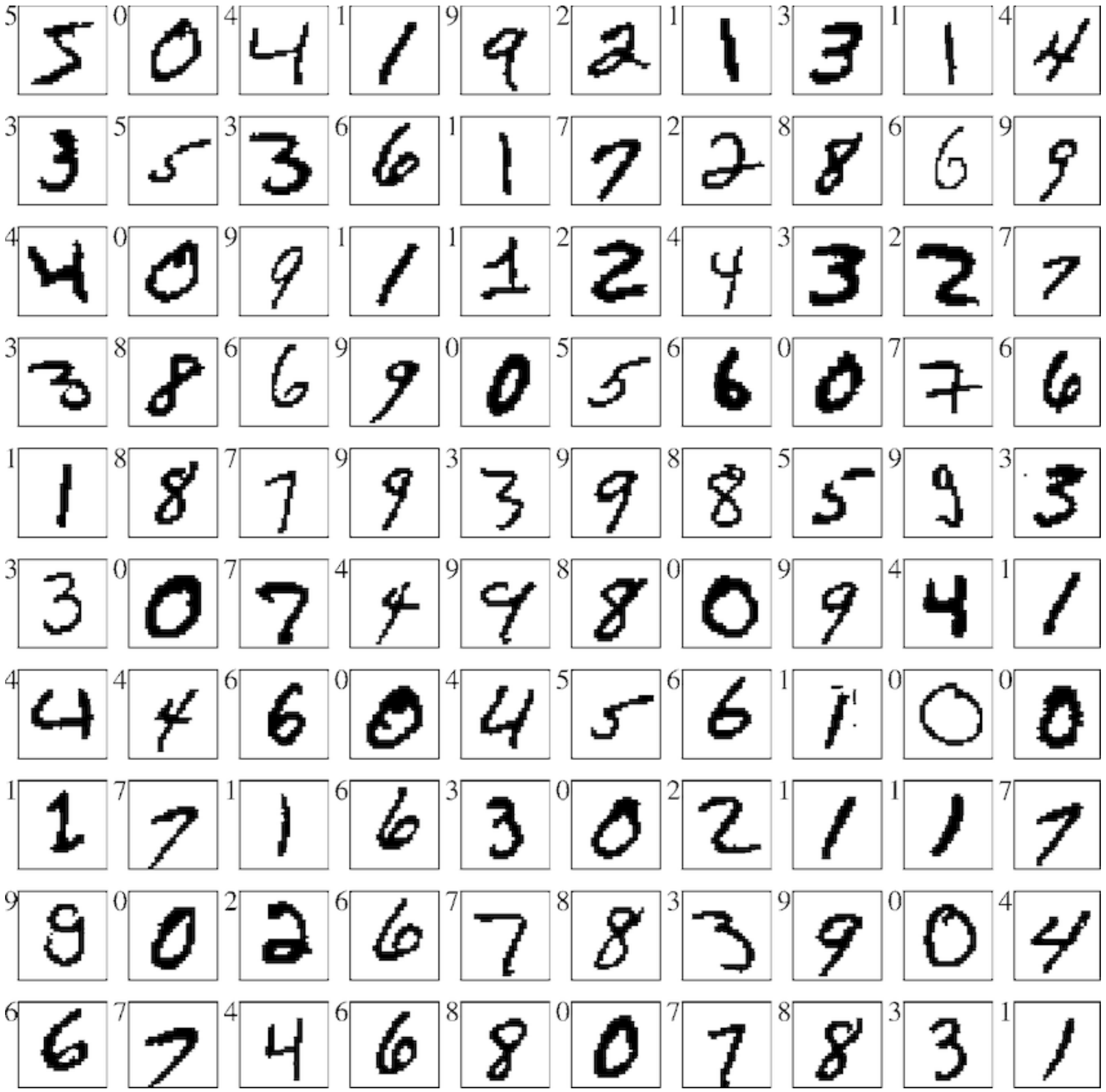
- Most common use case for HEP data analysis (MVA)
  - Train on MC or MC&DATA mixture
    - apply to data

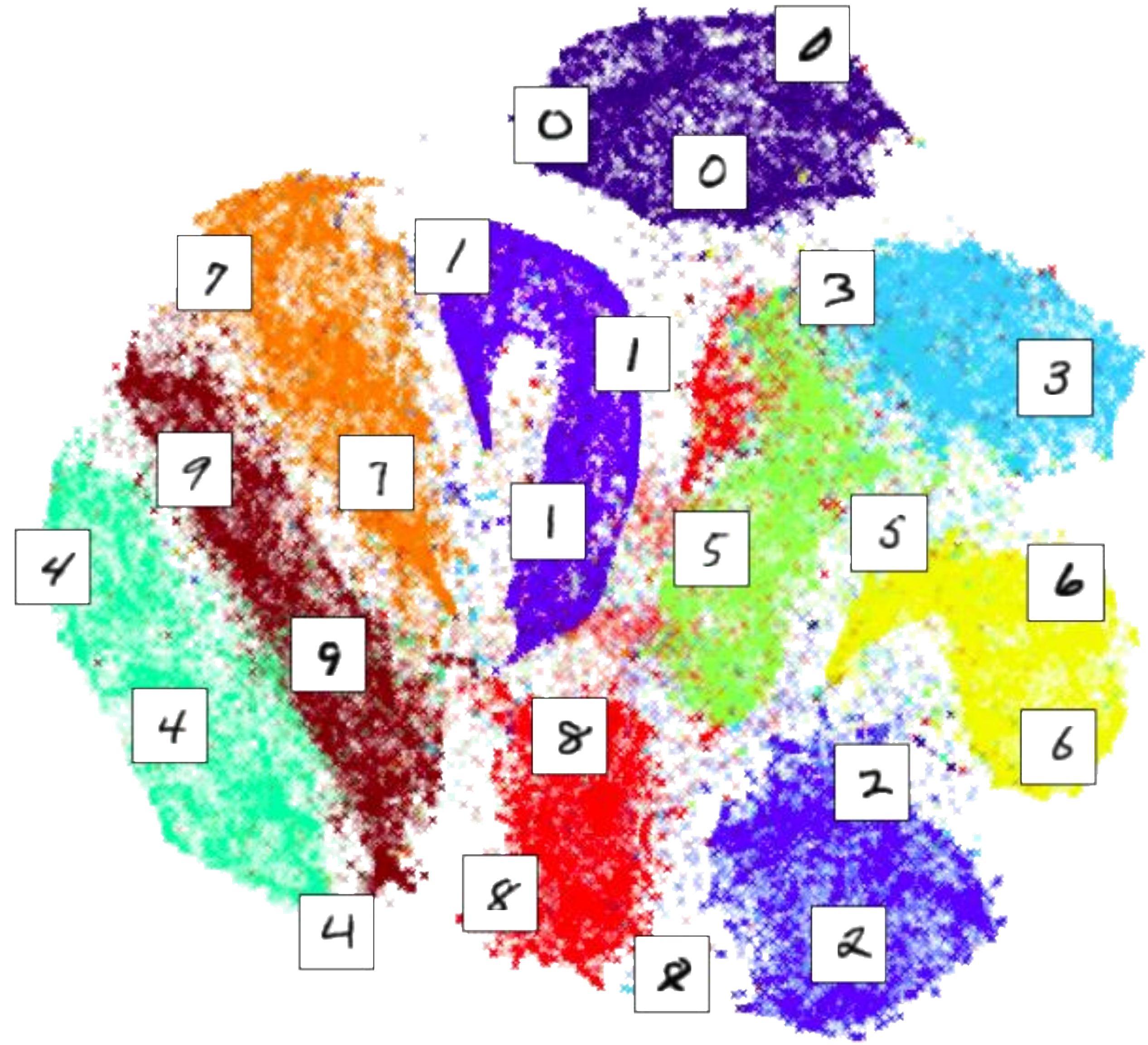


  - Need to account for (minor) difference between MC and data
    - different approaches: data doping, domain adaptation etc.
      - either approach requires control sample to test the difference
      - no ML tools to propagate difference in samples into systematics in classifier

Caveat: how to determine systematics due to sample difference

Unsupervised Learning: Dimensionality Reduction





- Reduce 28x28=784D space to 2D
  - t-SNE algorithm effectively separates classes

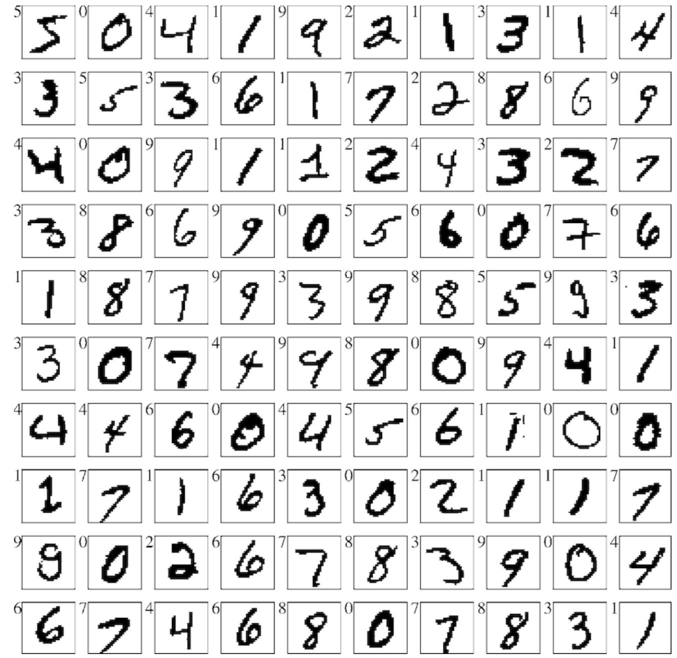
Generative Models

- Learn properties of the given dataset
  - determine domain where objects of the dataset are concentrated
  - probability distributions in the domain
- Then can generate random new but similar objects

- Unconditional
  - domain for all objects (e.g. “cat on the image”)
- Conditional
  - objects are accompanied by features

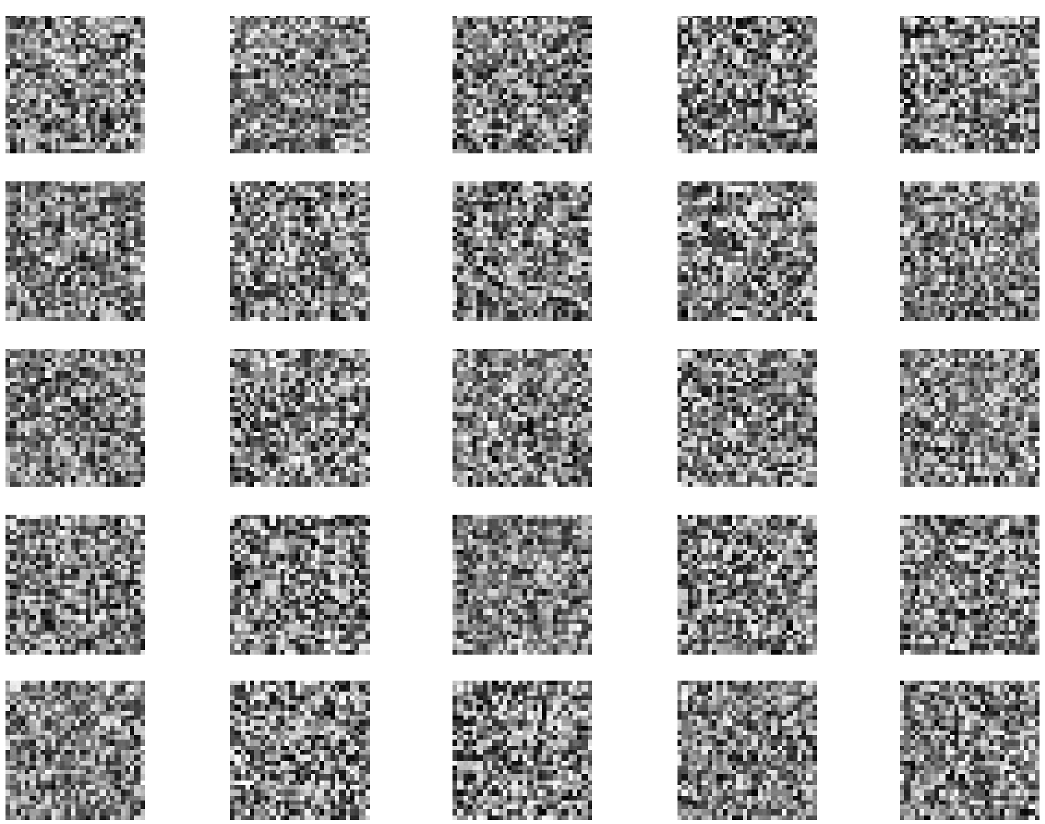
    - different domains for different conditions (e.g. “red cat”)

Baseline Generator



unconditional

Generative Adversarial Network

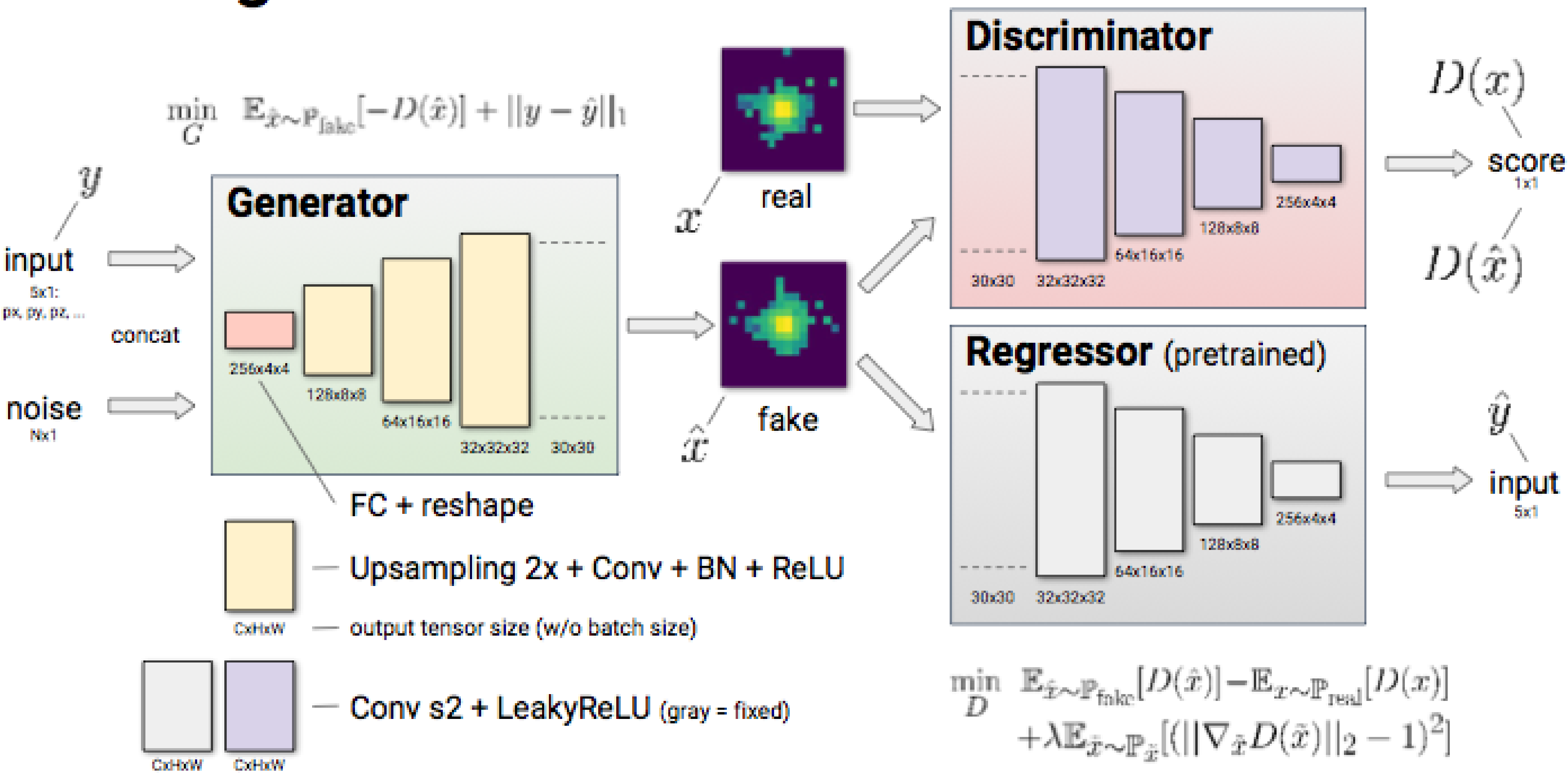


conditional

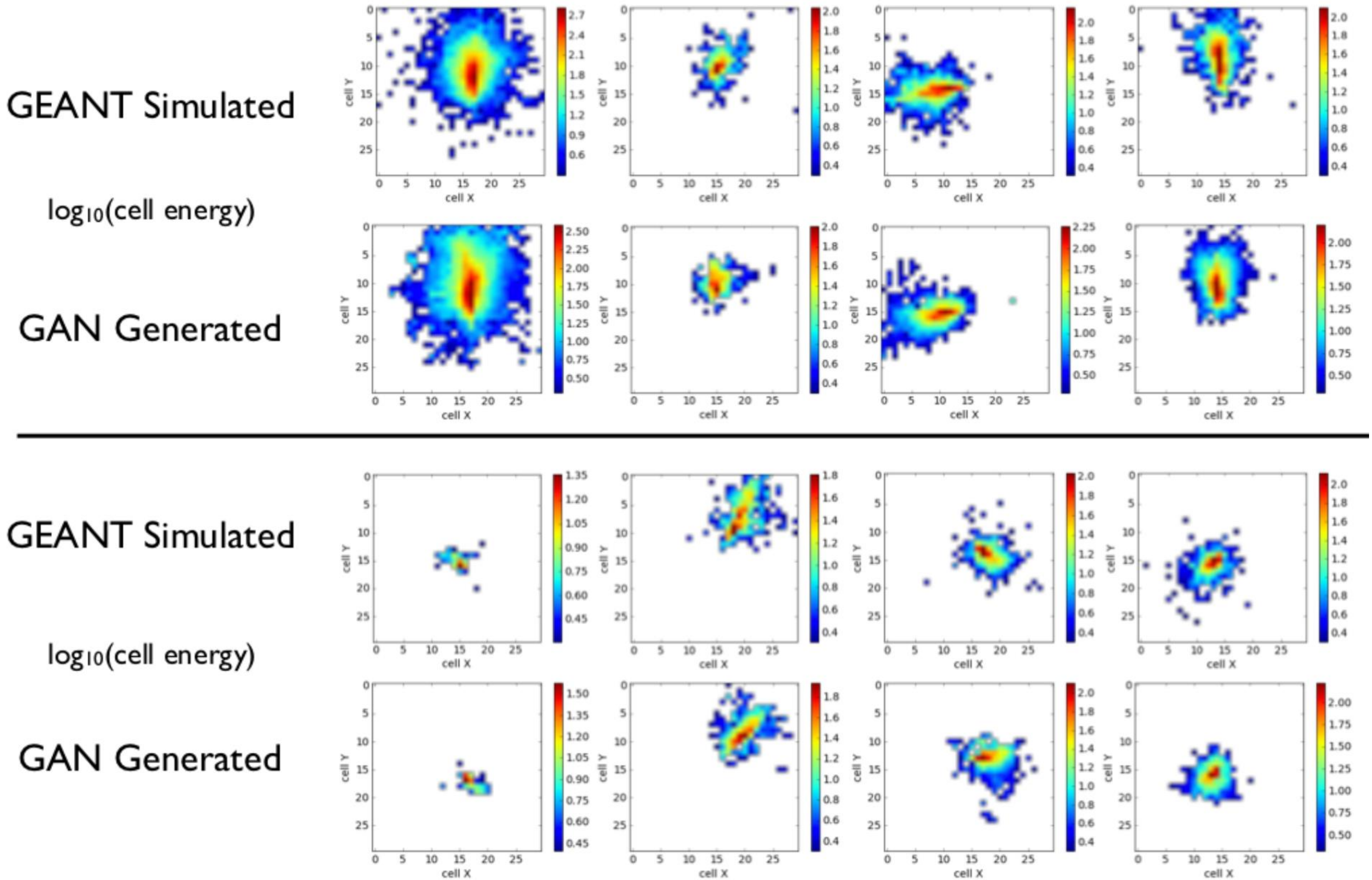
Epoch 1

Calorimeter Shower Simulation for LHC

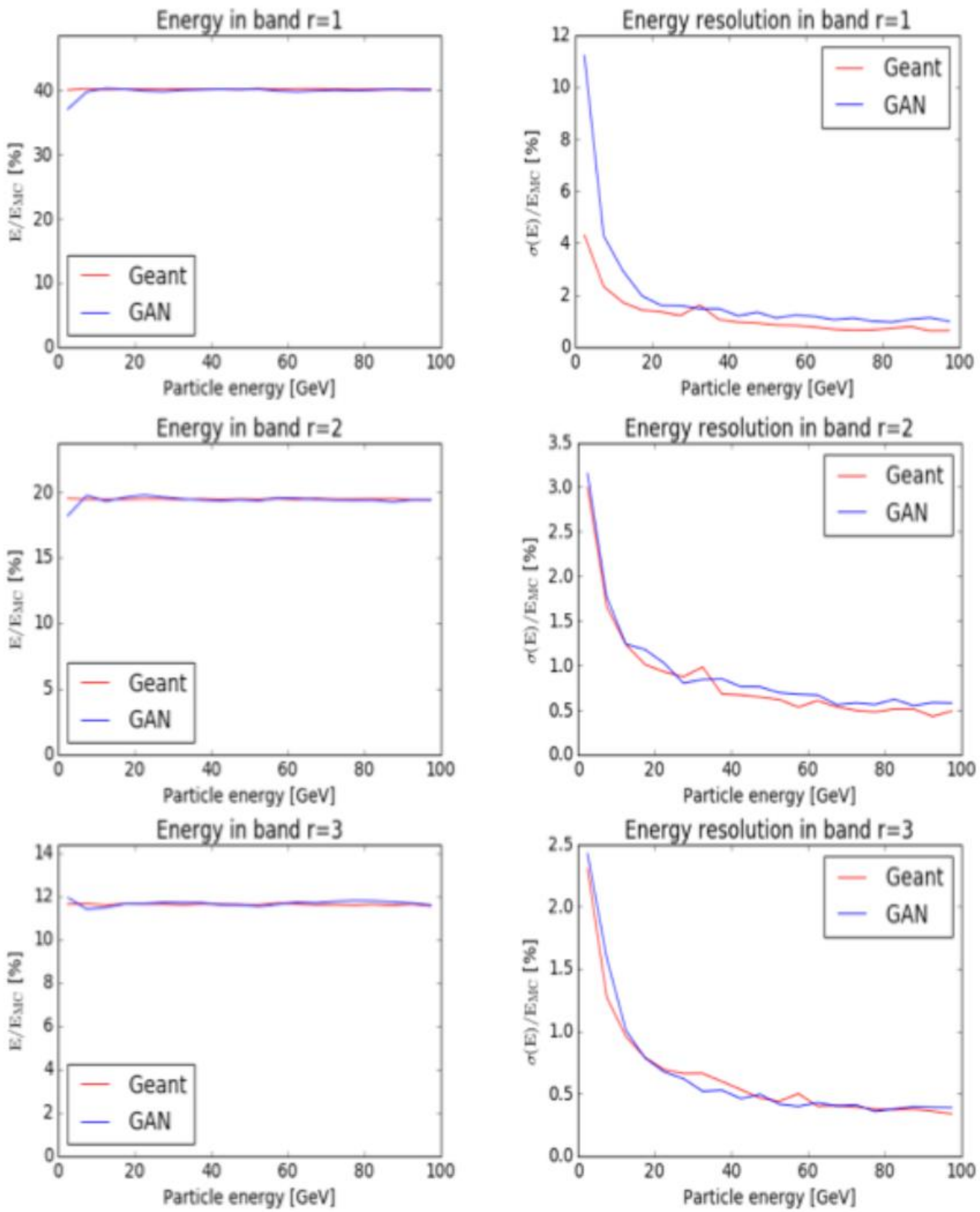
Training scheme

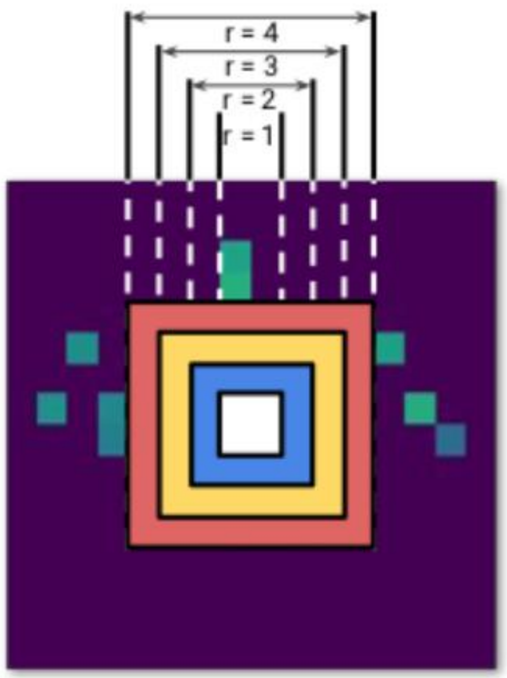


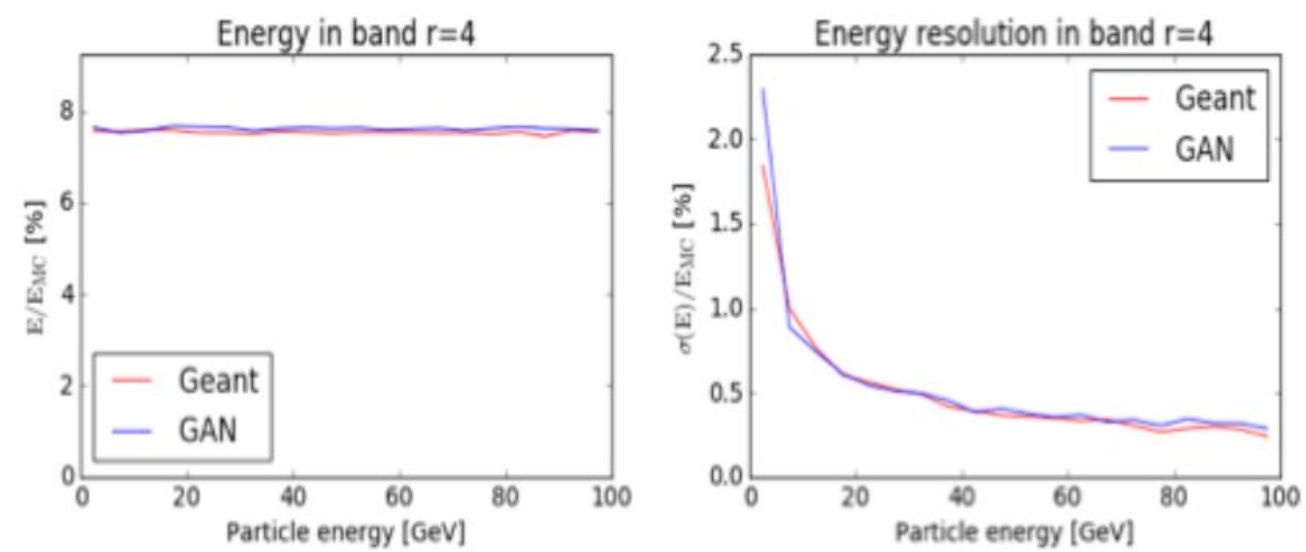
Conditional Generative Model in 5D



Performance in 1D (Energy)







- Good reproduction of first and second moments for cluster shape

Generative Models in Cosmology

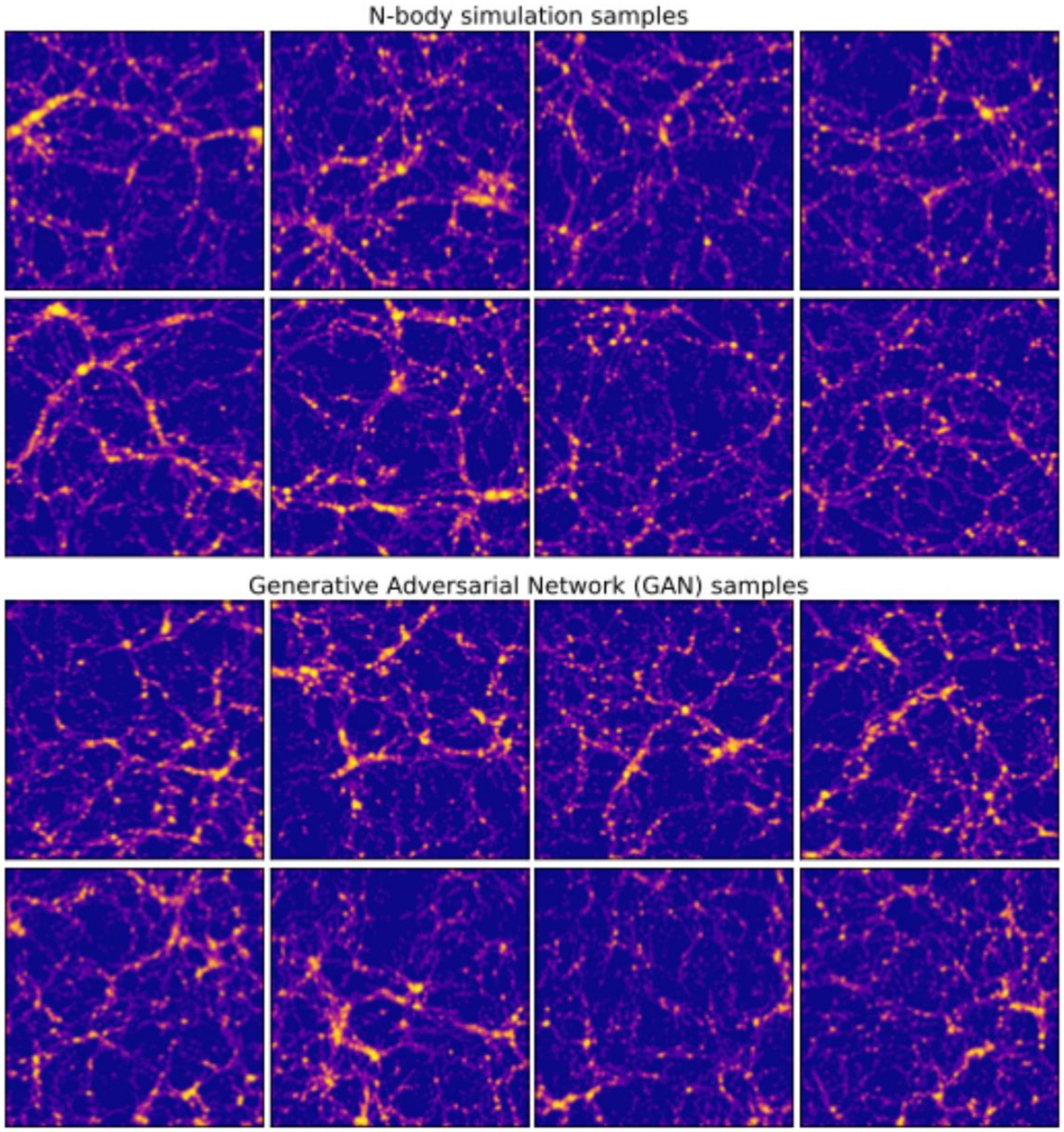


Figure 3: Samples from N-body simulation (top two rows) and from GAN (bottom two rows) for the box size of 100 Mpc. In this figure, transformation S1 with $k = 7$ was applied.

Rodríguez et al,

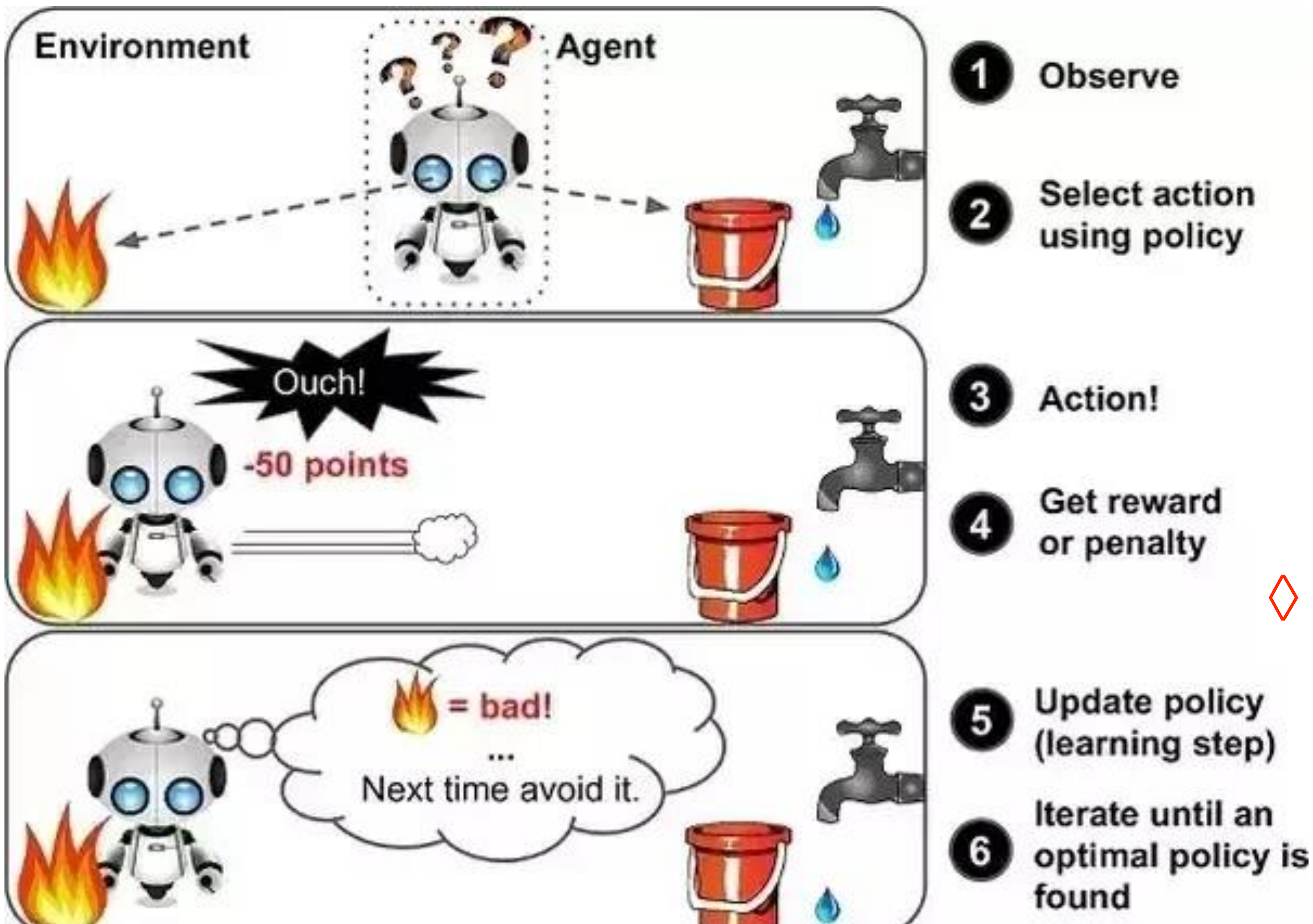
arXiv:1801.09070

Reinforced Learning

- We need to find some actions which push system in right direction
- Finding a optimal solution by probe-and-fail approach
  - probe points are chosen to optimise search speed
- Two extreme cases
  - getting system response to probe action is computationally cheap
    - classic RL
      - e.g. dynamic system control
  - calculation of one response is computationally heavy
      - e.g. optimising detector configuration
    - minimising number of probes
      - Gaussian Processes etc.

RL Basics

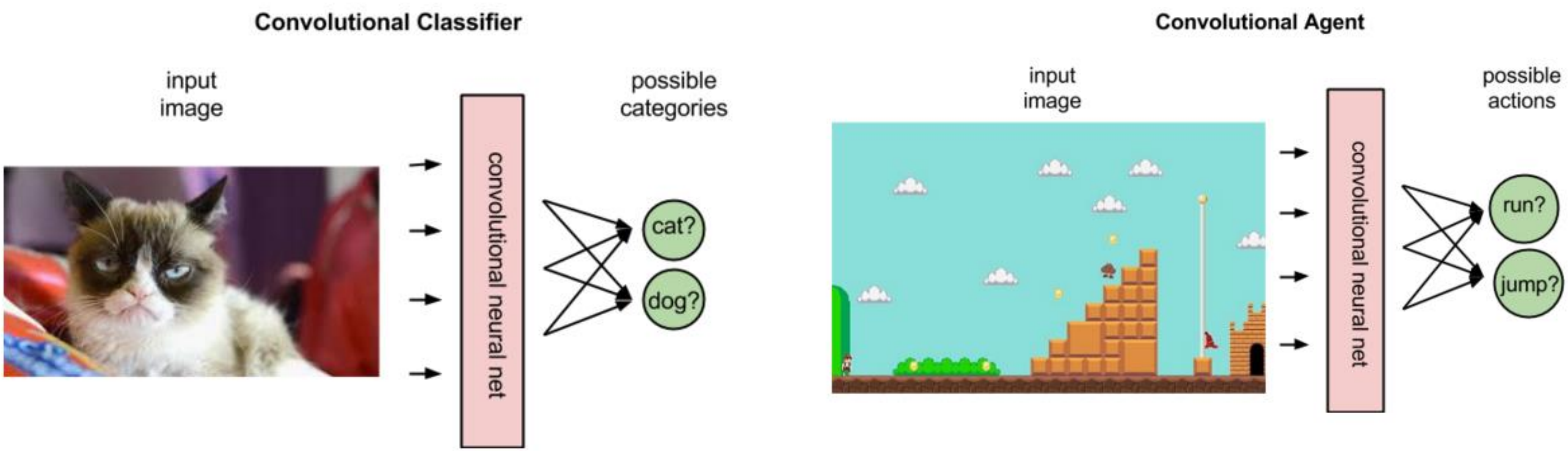
- Computer can run “groundhog day” many times at high rate
  - learn right strategy



- … and eventually learn the best action
  - e.g. to play Go or drive cars…

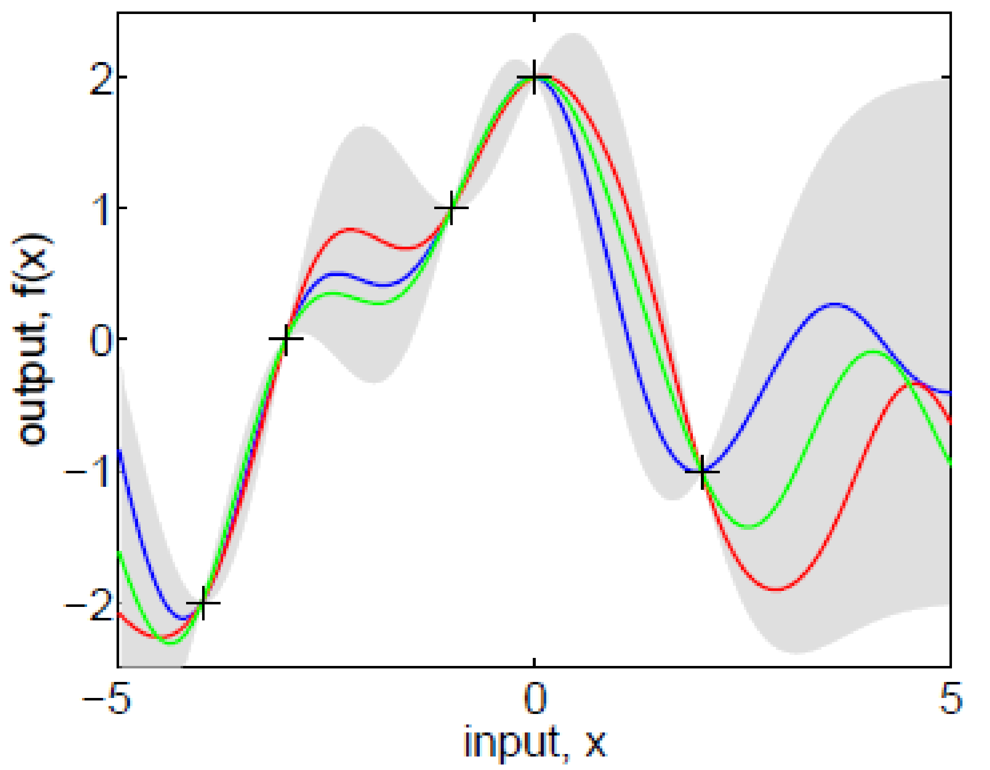
RL and NN

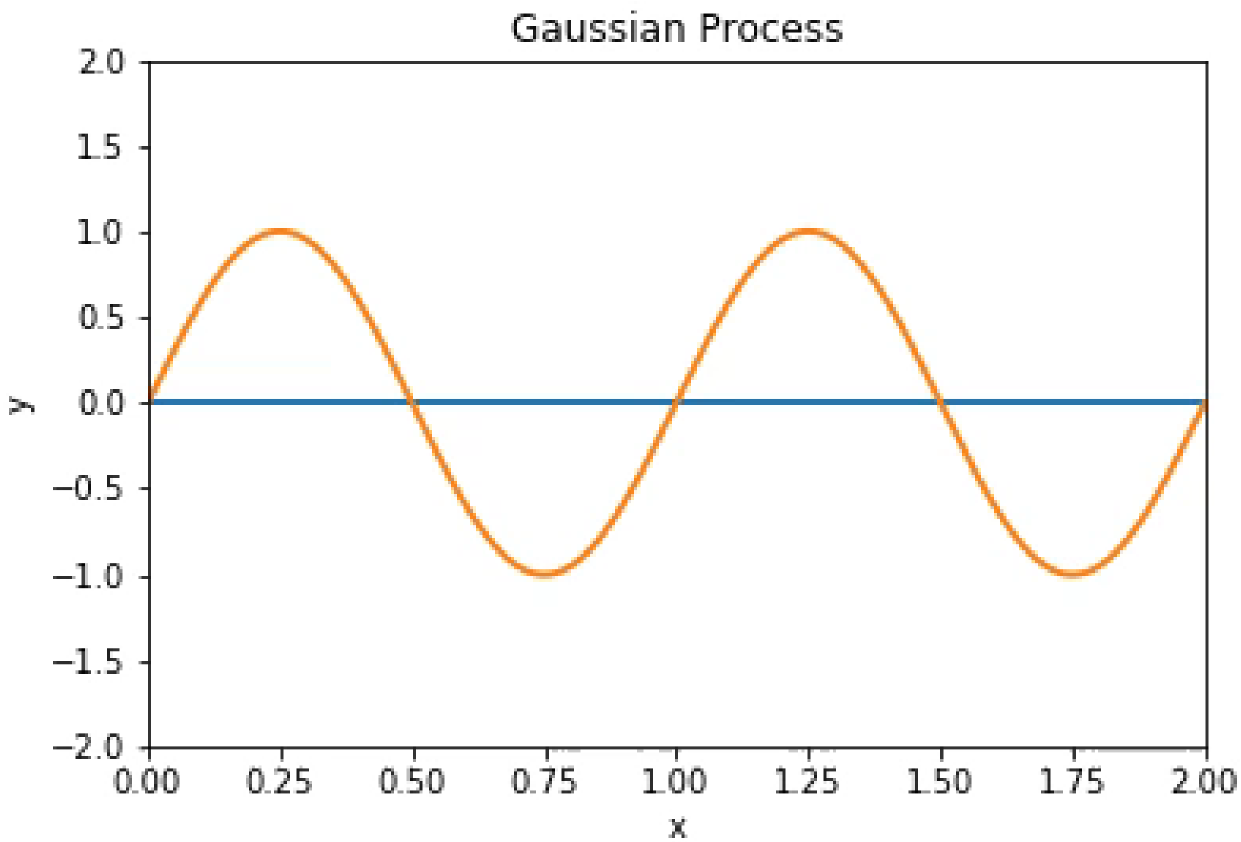
- ◊ NN is just a convenient approach to represent a dependency
  - ◊ with well established mechanisms for training



Surrogate Models

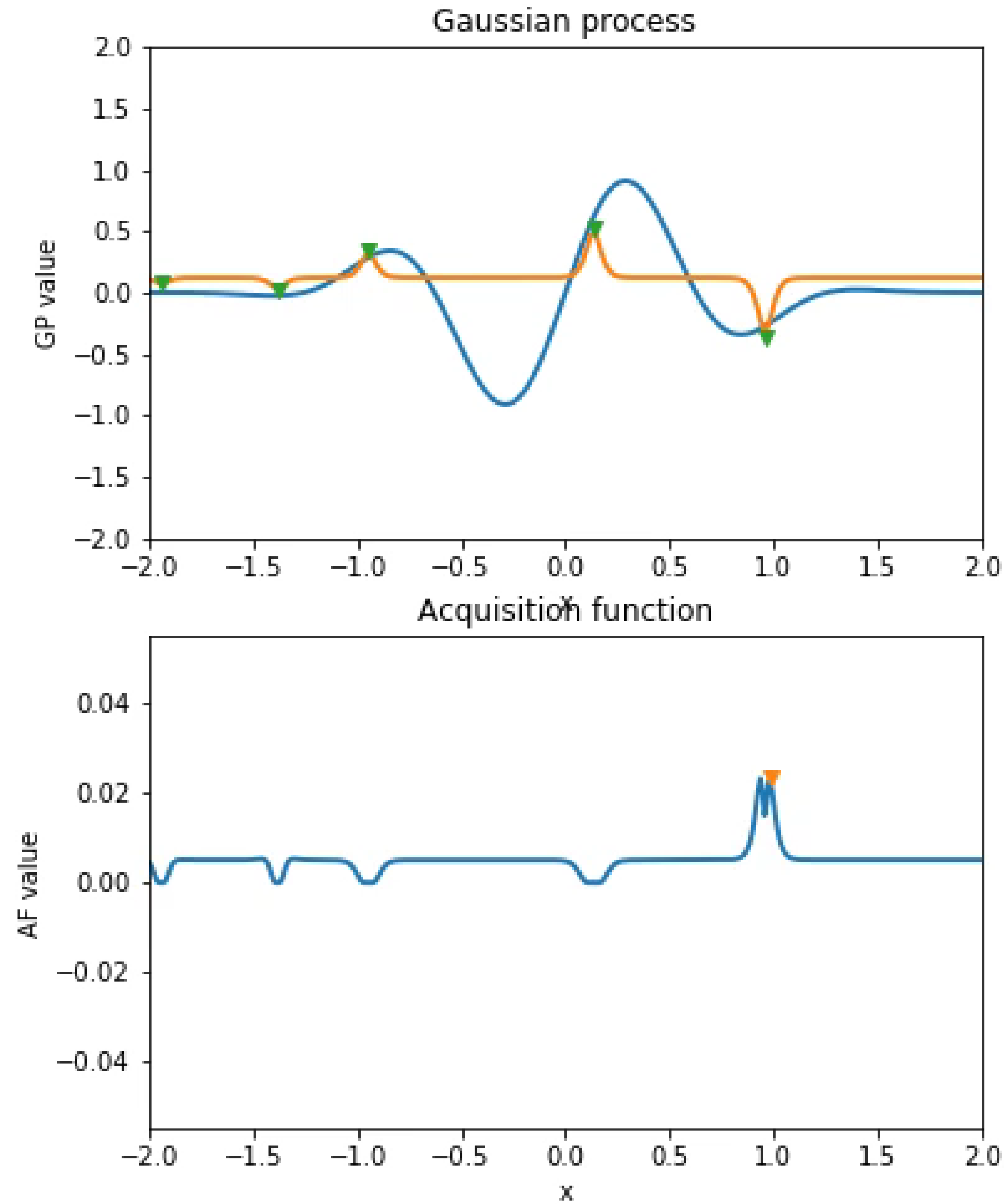
- Another extreme:
  - want to evaluate function within some multi-dimensional phase space
  - evaluation in a point is computationally expensive
  - need to dynamically optimise measurement points to provide the best sensitivity
- Typical Bayesian approach: Gaussian Processes
- Typical use case: detector optimisation
  - every measurement requires a set of MC (Geant) simulation
    - may take days



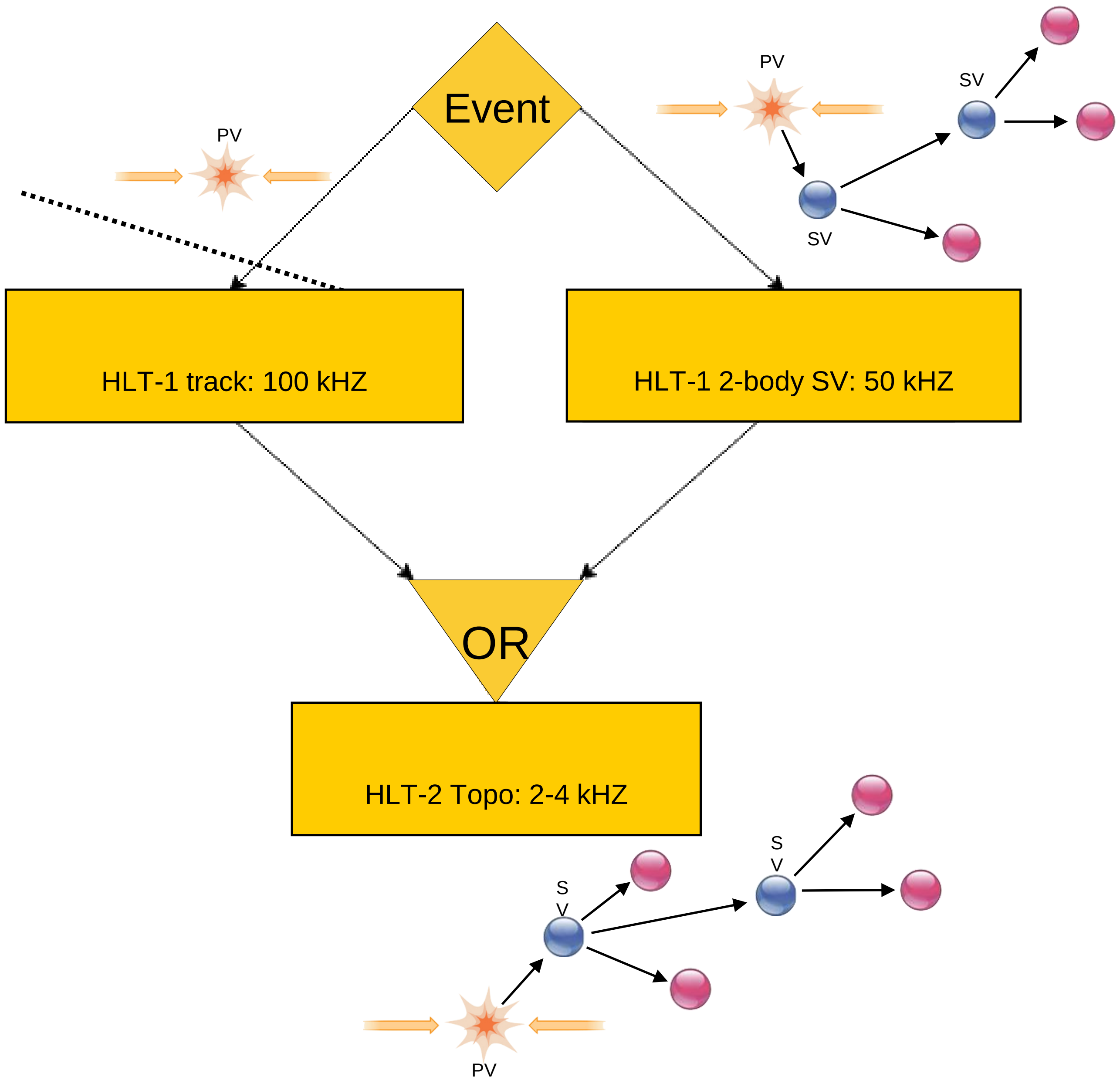


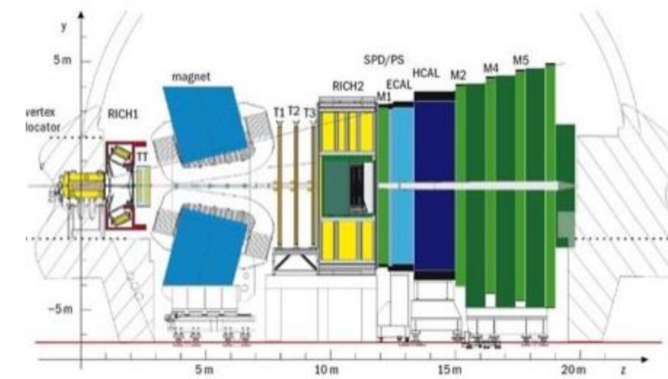
Finding Optimum

- Another extreme:
  - want to evaluate function within some multi-dimensional phase space
  - evaluation in a point is computationally expensive
  - need to dynamically optimise measurement points to provide the best sensitivity
- Typical Bayesian approach: Gaussian Processes
- Typical use case: detector optimisation
  - every measurement requires a set of MC (Geant) simulation
    - may take days



Success Stories: LHCb Trigger



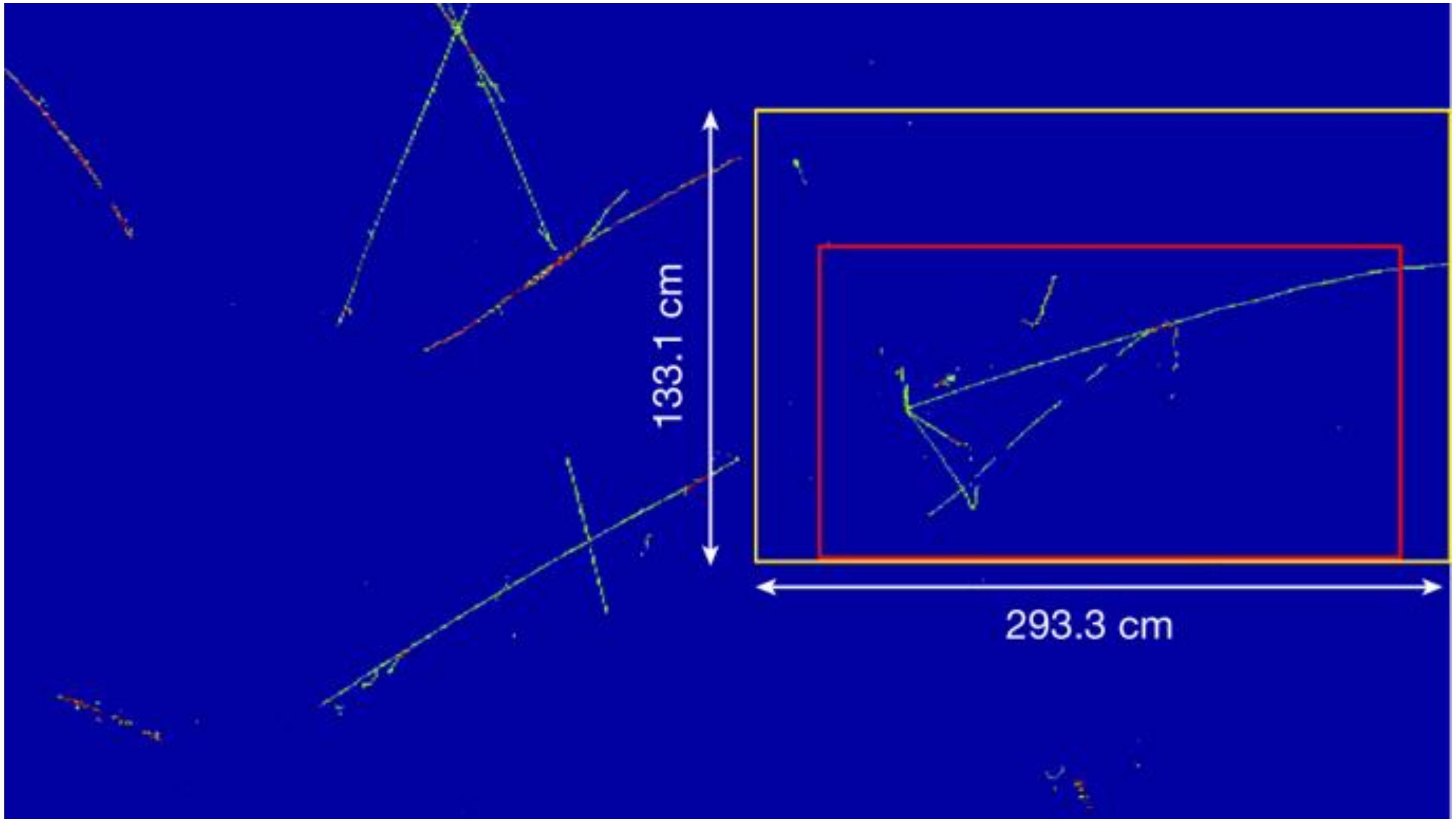


mode	2.5 kHz	4. kHz
$B^0 \to K^*[K^+\pi^-]\mu^+\mu^-$	1.64	1.72
$B^+ \to \pi^+K^-K^+$	1.59	1.65
$B_s^0 \to D_s^-[K^+K^-\pi^-]\mu^+\nu_\mu$	1.14	1.47
$B_s^0 \to \psi(1S)[\mu^+\mu^-]K^+K^-\pi^+\pi^-$	1.62	1.71
$B_s^0 \to D_s^-[K^+K^-\pi^-]\pi^+$	1.46	1.52
$B^0 \to D^+[K^-\pi^+\pi^+]D^-[K^+\pi^-\pi^-]$	1.40	1.86

- Aggressively re-optimised topological trigger for Run II operation
- Gain 10%..70% efficiency for different channels!

SS: MicroBooNE Neutrino Selection

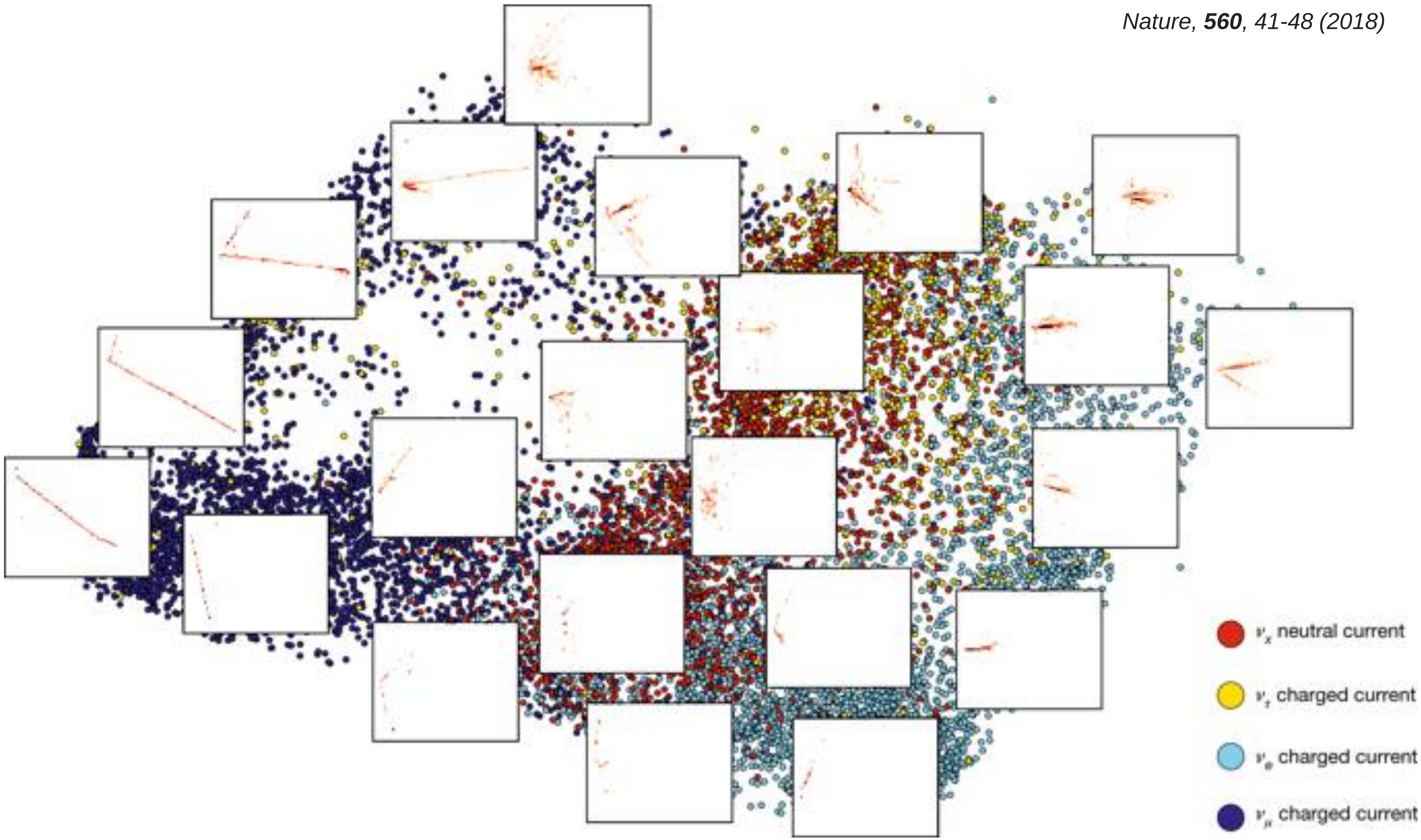
Nature, ***560***, *41-48 (2018)*



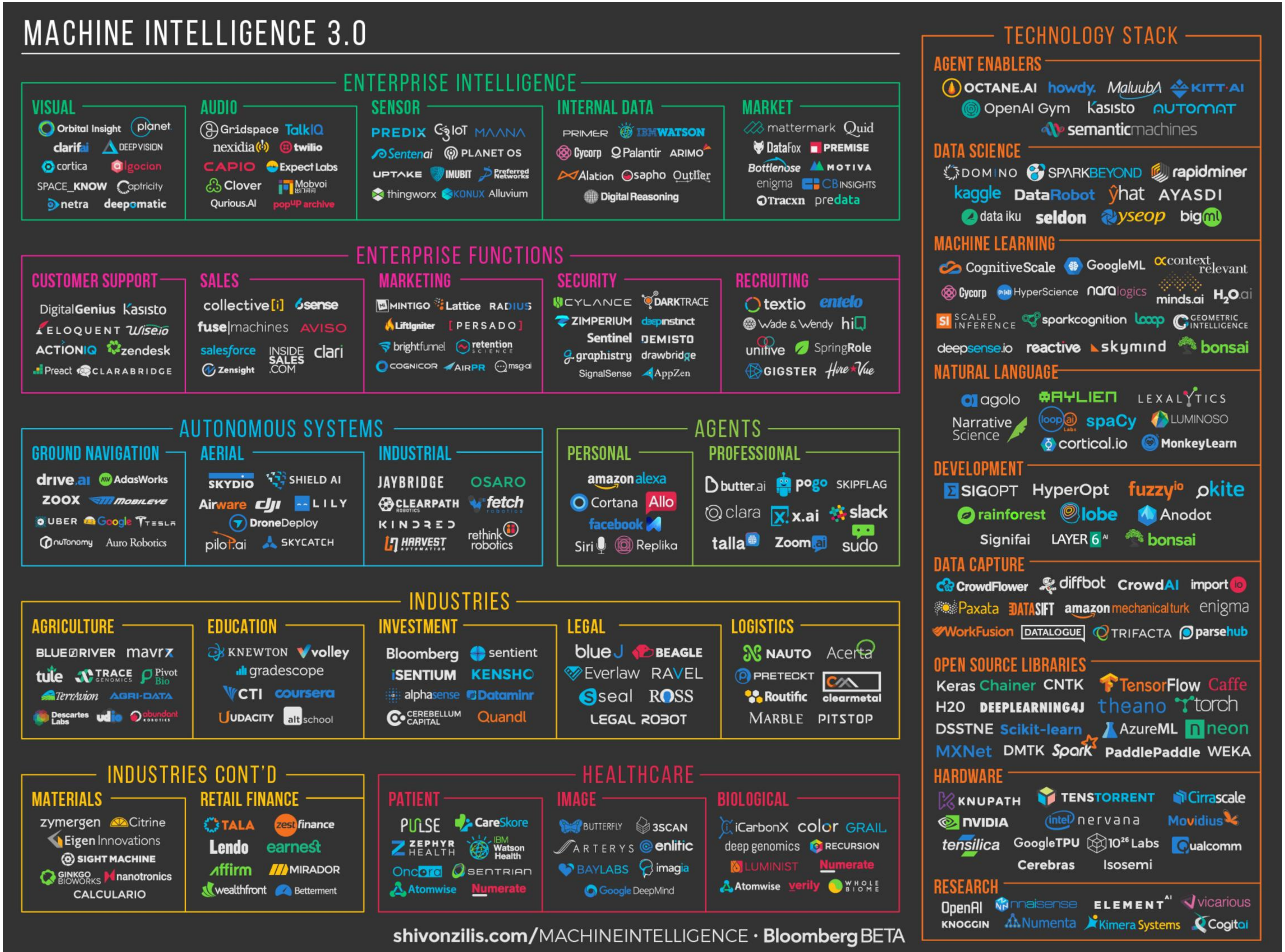
◊ CNN effectively predicts bounding box containing interacting neutrino

SS: NOvA Event Selection

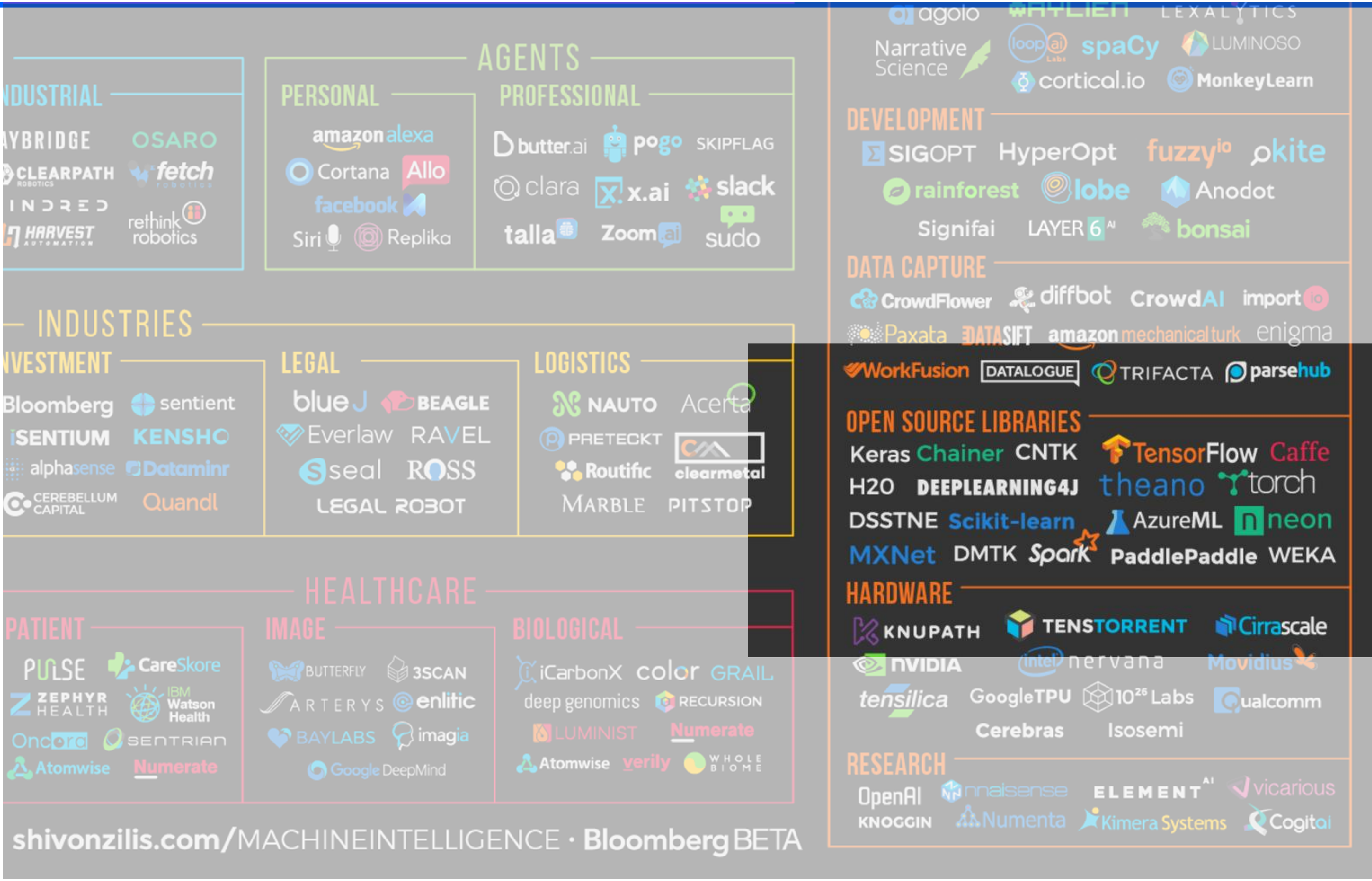
Nature, ***560***, *41-48 (2018)*



◊ Classification of different types of different neutrino interactions


MACHINE INTELLIGENCE 3.0
ENTERPRISE INTELLIGENCE
VISUAL
Orbital Insight
planet
clarifai
DEEP VISION
cortica
gocion
SPACE_KNOW
Captricity
netra
deepomatic
AUDIO
Gridspace
TalkIQ
nexidia
twilio
CAPIO
Expect Labs
Clover
Mobvoi
Qurious.AI
popUp archive
SENSOR
PREDIX
C3 IoT
MAANA
Sentenai
PLANET OS
UPTAKE
IMUBIT
Preferred Networks
thingworx
KONUX
Alluvium
INTERNAL DATA
PRIMER
IBM WATSON
Cycorp
Palantir
ARIMO
Alation
sapho
Outlier
Digital Reasoning
MARKET
mattermark
Quid
DataFox
PREMISE
Bottlenose
MOTIVA
enigma
CB INSIGHTS
Tracxn
predata
ENTERPRISE FUNCTIONS
CUSTOMER SUPPORT
DigitalGenius
Kasisto
ELOQUENT
Wise.io
ACTIONIQ
zendesk
Preact
CLARABRIDGE
SALES
collective[i]
6sense
fuse|machines
AVISO
salesforce
INSIDE SALES .COM
clari
Zensight
MARKETING
MINTIGO
Lattice
RADIUS
Liftigniter
PERSADO
brightfunnel
retention science
COGNICOR
AIRPR
msg.ai
SECURITY
CYLANCE
DARKTRACE
ZIMPERIUM
deepinstinct
Sentinel
DEMISTO
graphistry
drawbridge
SignalSense
AppZen
RECRUITING
textio
entelo
Wade & Wendy
hiQ
unitive
SpringRole
GIGSTER
HireVue
AUTONOMOUS SYSTEMS
GROUND NAVIGATION
drive.ai
AdasWorks
ZOOX
MOBILEYE
UBER
Google
TESLA
nuTonomy
Auro Robotics
AERIAL
SKYDIO
SHIELD AI
Airware
dji
LILY
DroneDeploy
pilot.ai
SKYCATCH
INDUSTRIAL
JAYBRIDGE
OSARO
CLEARPATH
fetch robotics
KINDRED
HARVEST AUTOMATION
rethink robotics
AGENTS
PERSONAL
amazon alexa
Cortana
Allo
facebook
Siri
Replika
PROFESSIONAL
butter.ai
pogo
SKIPFLAG
clara
x.ai
slack
talla
Zoom.ai
sudo
INDUSTRIES
AGRICULTURE
BLUE RIVER
mavrx
tule
TRACE GENOMICS
Pivot Bio
TerrAvion
AGRI-DATA
Descartes Labs
udio
abundant robotics
EDUCATION
KNEWTON
volley
gradescope
CTI
coursera
UDACITY
altschool
INVESTMENT
Bloomberg
sentient
iSENTIUM
KENSHO
alphasense
Dataminr
CEREBELLUM CAPITAL
Quandl
LEGAL
blue J
BEAGLE
Everlaw
RAVEL
seal
ROSS
LEGAL ROBOT
LOGISTICS
NAUTO
Acerta
PRETECKT
clearmetal
Routific
MARBLE
PITSTOP
INDUSTRIES CONT'D
MATERIALS
zymergen
Citrine
Eigen Innovations
SIGHT MACHINE
GINKGO BIOWORKS
nanotronics
CALCULARIO
RETAIL FINANCE
TALA
zestfinance
Lendo
earnest
affirm
MIRADOR
wealthfront
Betterment
HEALTHCARE
PATIENT
PULSE
CareSkore
ZEPHYR HEALTH
IBM Watson Health
Oncora
SENTRIAN
Atomwise
Numerate
IMAGE
BUTTERFLY
3SCAN
ARTERYS
enlitic
BAYLABS
imagia
Google DeepMind
BIOLOGICAL
iCarbonX
color
GRAIL
deep genomics
RECURSION
LUMINIST
Numerate
Atomwise
verily
WHOLE BIOME
TECHNOLOGY STACK
AGENT ENABLERS
OCTANE.AI
howdy.
Maluuba
KITT.AI
OpenAI Gym
Kasisto
AUTOMAT
semanticmachines
DATA SCIENCE
DOMINO
SPARKBEYOND
rapidminer
kaggle
DataRobot
yhat
AYASDI
data iku
seldon
yseop
bigml
MACHINE LEARNING
CognitiveScale
GoogleML
context relevant
Cycorp
HyperScience
naralogics
minds.ai
H2O.ai
SCALED INFERENCE
sparkcognition
Loop
GEOMETRIC INTELLIGENCE
deepsense.io
reactive
skymind
bonsai
NATURAL LANGUAGE
agolo
AYLIEN
LEXALYTICS
Narrative Science
loop ai Labs
spaCy
LUMINOSO
cortical.io
MonkeyLearn
DEVELOPMENT
SIGOPT
HyperOpt
fuzzy.io
kite
rainforest
lobe
Anodot
Signifai
LAYER 6 AI
bonsai
DATA CAPTURE
CrowdFlower
diffbot
CrowdAI
import.io
Paxata
DATASIFT
amazon mechanical turk
enigma
WorkFusion
DATALOGUE
TRIFACTA
parsehub
OPEN SOURCE LIBRARIES
Keras
Chainer
CNTK
TensorFlow
Caffe
H20
DEEPLEARNING4J
theano
torch
DSSTNE
Scikit-learn
AzureML
neon
MXNet
DMTK
Spark
PaddlePaddle
WEKA
HARDWARE
KNUPATH
TENSTORRENT
Cirrascale
NVIDIA
intel nervana
Movidius
tensilica
GoogleTPU
10[26] Labs
Qualcomm
Cerebras
Isosemi
RESEARCH
OpenAI
nnaisense
ELEMENT AI
vicarious
KNOGGIN
Numenta
Kimera Systems
Cogitai
shivonzilis.com/MACHINEINTELLIGENCE · Bloomberg BETA

Domain Relevant for Science



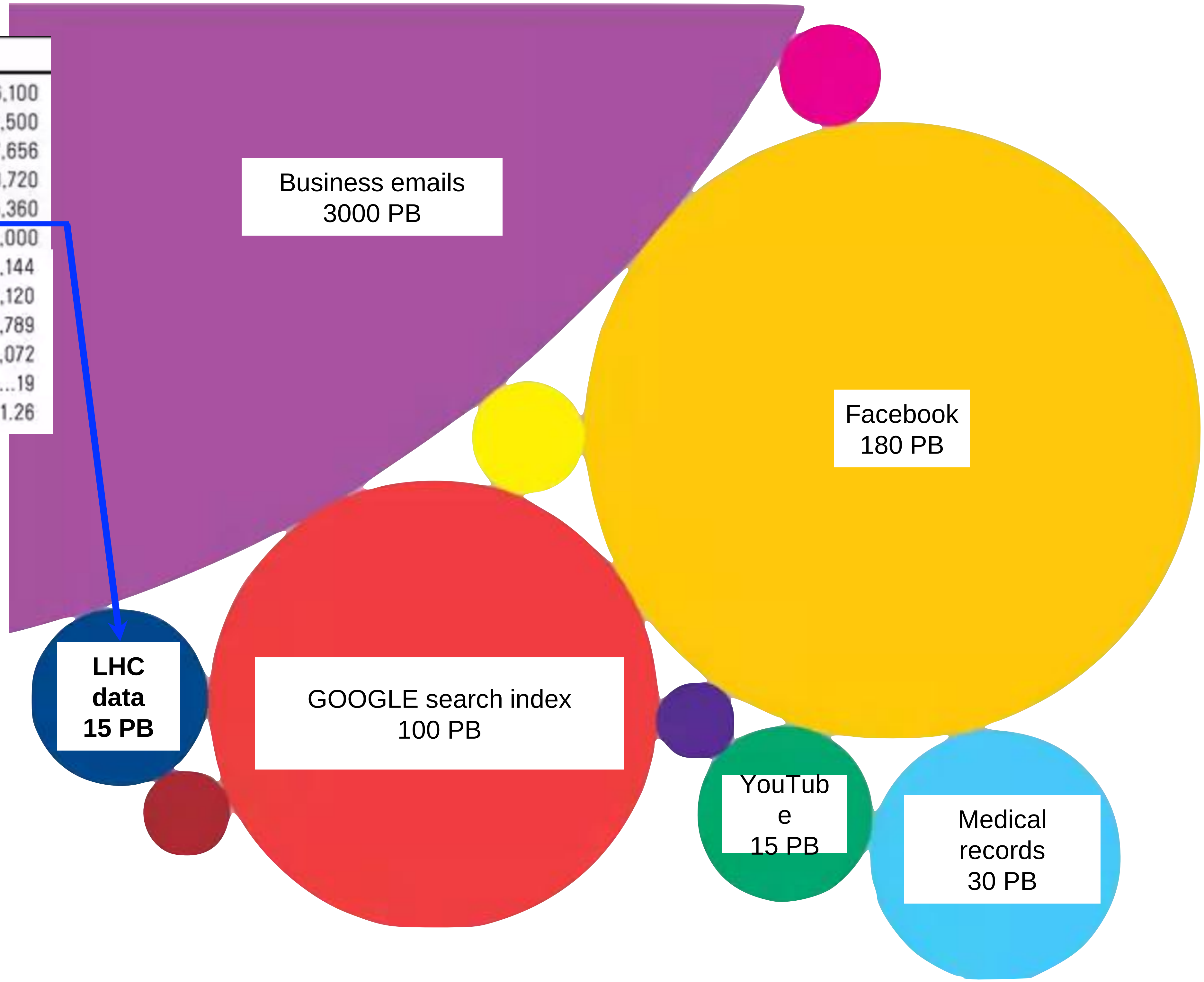
Big Data Big Players

data collected in 2012

адаптировано из http://www.wired.com/2013/04/bigdata/

Size of data sets in terabytes

Business email sent per year	2,986,100
Content uploaded to Facebook each year	182,500
Google's search index	97,656
Kaiser Permanente's digital health records	30,720
Large Hadron Collider's annual data output	15,360
Videos uploaded to YouTube per year	15,000
National Climactic Data Center database	6,144
Library of Congress' digital collection	5,120
US Census Bureau data	3,789
Nasdaq stock market database	3,072
Tweets sent in 2012	19
Contents of every print issue of WIRED	1.26



Evolution of HEP x ML Engineering

Layer	HEP (Circa 2013)	HEP (Circa 2018)	Industry
Data Layer	ROOT Files	ROOT Files	DB / HDFS etc.
Loading Layer	Ad hoc ROOT ETL logic	Numpy / HDF5 Converters / Loaders	Numpy / HDF5 Converters / Loaders
Training Layer	TMVA	Keras, TensorFlow, PyTorch, XGBoost, scikit-learn, …	Keras, TensorFlow, PyTorch, XGBoost, scikit-learn, …
Serving Layer	Deployment Target (TMVA)	Deployment Target (lwtnn, TensorFlow, TMVA wrappers)	Deployment Target (TensorFlow Serving, SageMaker, etc.)

11

Luke De Oliveira, 2018

Learning Curve

- ML is mostly empirical science now
  - there are different approaches and tricks that proved to work
    - … for some particular problems and cases
  - the success is driven by finding right combination of
    - architecture
    - data for training
    - loss function
    - training procedure
- Hands on experience of solving different problems is vital to find right combination and solve complex problems

Approaches to Accommodate Expertise

- Inside out
  - By including ML techniques into research stack
    - quality of results is boosted by using state of the art data processing technology
    - **young people learn skills that are**
      - **beneficial for scientific research**
      - **highly valued on the labor market**
- Outside in
  - There are plenty of romantic ML experts that happily contribute into fundamental study of the Universe
    - there are different approaches to such cooperation
      - more in A.U. presentation

TMVA as a ML Wrapper for Particle Physics

ROOT
Data Analysis Framework

Google Custom Search

Download Documentation News Support About Development Contribute

Home » First Steps With ROOT » Processing data with ROOT

TMVA

The Toolkit for Multivariate Data Analysis with ROOT (TMVA) is a ROOT-integrated project providing a machine learning environment for the processing and evaluation of multivariate classification, both binary and multi class, and regression techniques targeting applications in high-energy physics. The package includes:

- Hands-on experience with TMVA is hard to sell to industry
  - like is hard to sell RooStats - R and/or Matlab are valued
- Mainstreams are Scikit-learn, Keras, TensorFlow, PyTorch…

Strongly encourage to avoid domain specific wrappers

- avoid marginalising of the expertise for young people in the field

Summary

- Machine Learning technologies proved to be applicable to broad range of different tasks in modern society
  - In science as well
- Imported experience and expertise can significantly boost modern researches
- Many software and hardware solutions are developed
  - re-use, not re-invent
- Hands-on ML expertise is attractive for young researchers, and is highly valuable both inside and beyond the academy