technische universität dortmund

Experimentelle Physik V
Teilchenphysik

IceCube

Searching Leading Muons with IceCube

A Machine Learning Approach

Johannes Werthebach

Machine learning: what and why?

"We are drowning in information and starving for knowledge."

John Naisbitt

[1]

Machine learning: what and why?

"We are drowning in information and starving for knowledge."

John Naisbitt

Why?

- High dimensionality
- Hidden Patterns
- Multivariate analysis

- Scientist:
  Precuts & attributes
  Knowledge / experience / bias

[1]

Machine learning: what and why?

"We are drowning in information and starving for knowledge."

John Naisbitt

Why?

- High dimensionality
- Hidden Patterns
- Multivariate analysis

- Scientist:
  Precuts & attributes
  Knowledge / experience / bias

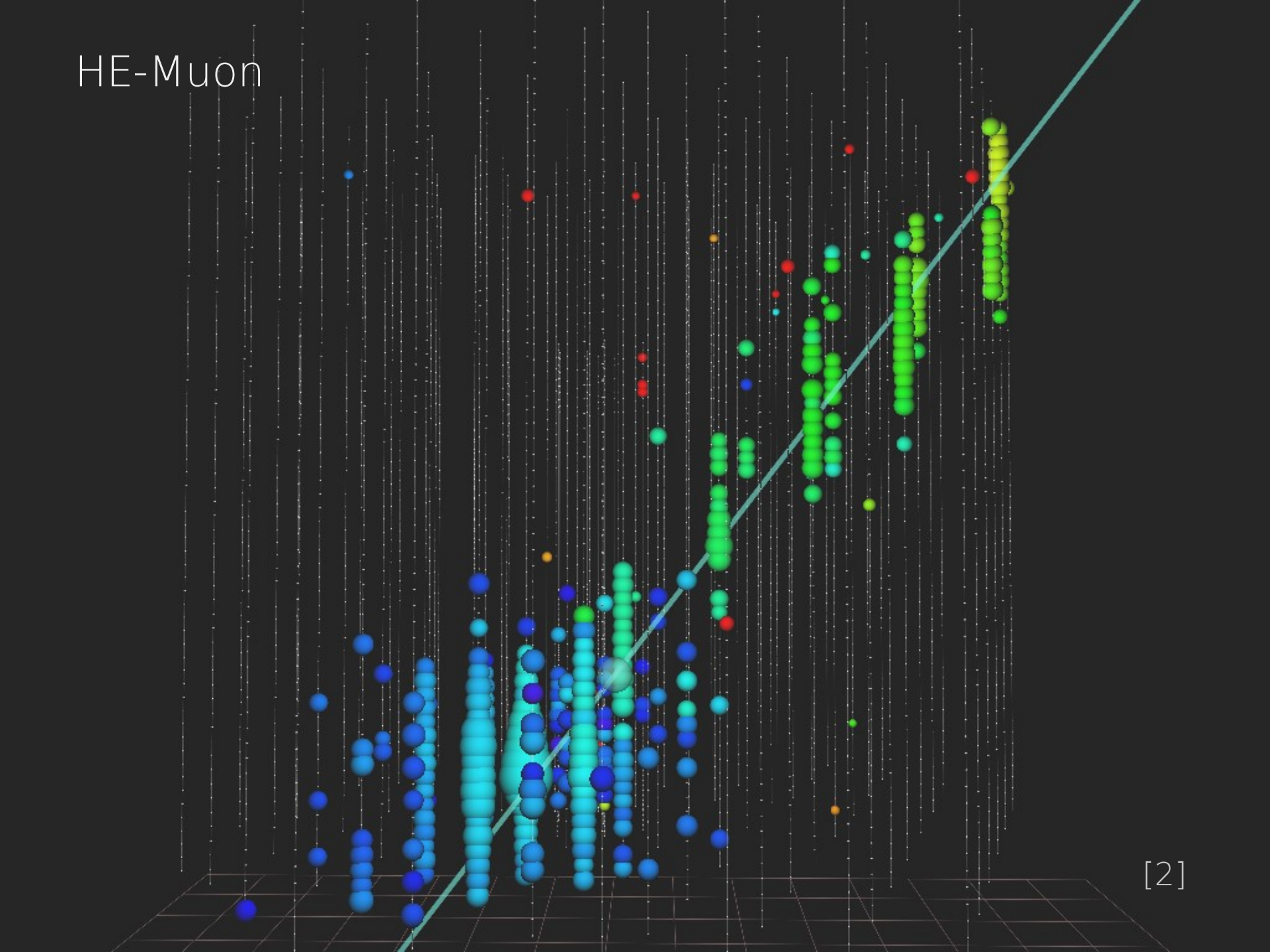
What?

- IceCube: 1 TB/day
- Kinect
- Spam, …

- Predictive learning
- **Supervised learning**

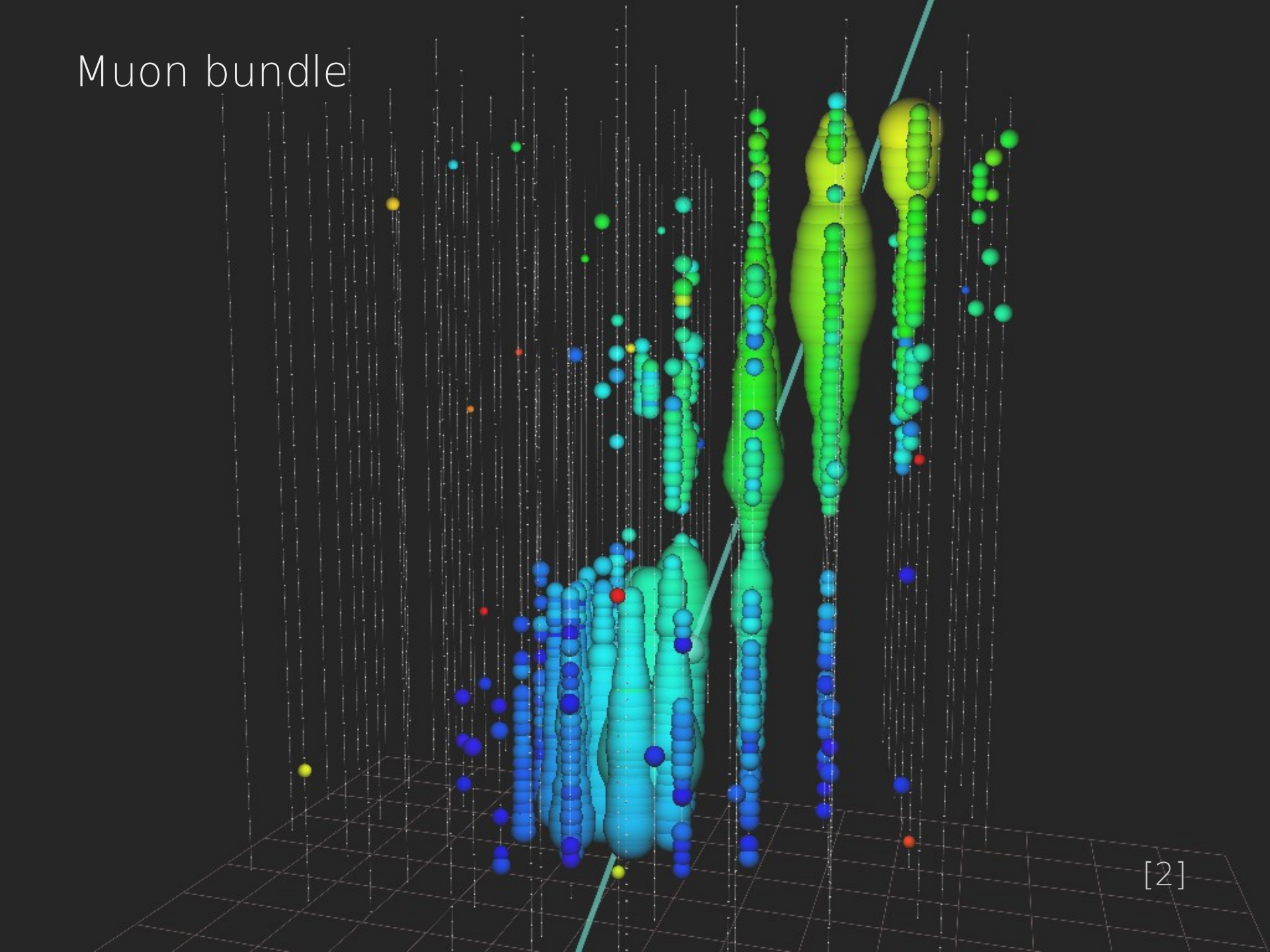
[1]

HE-Muon



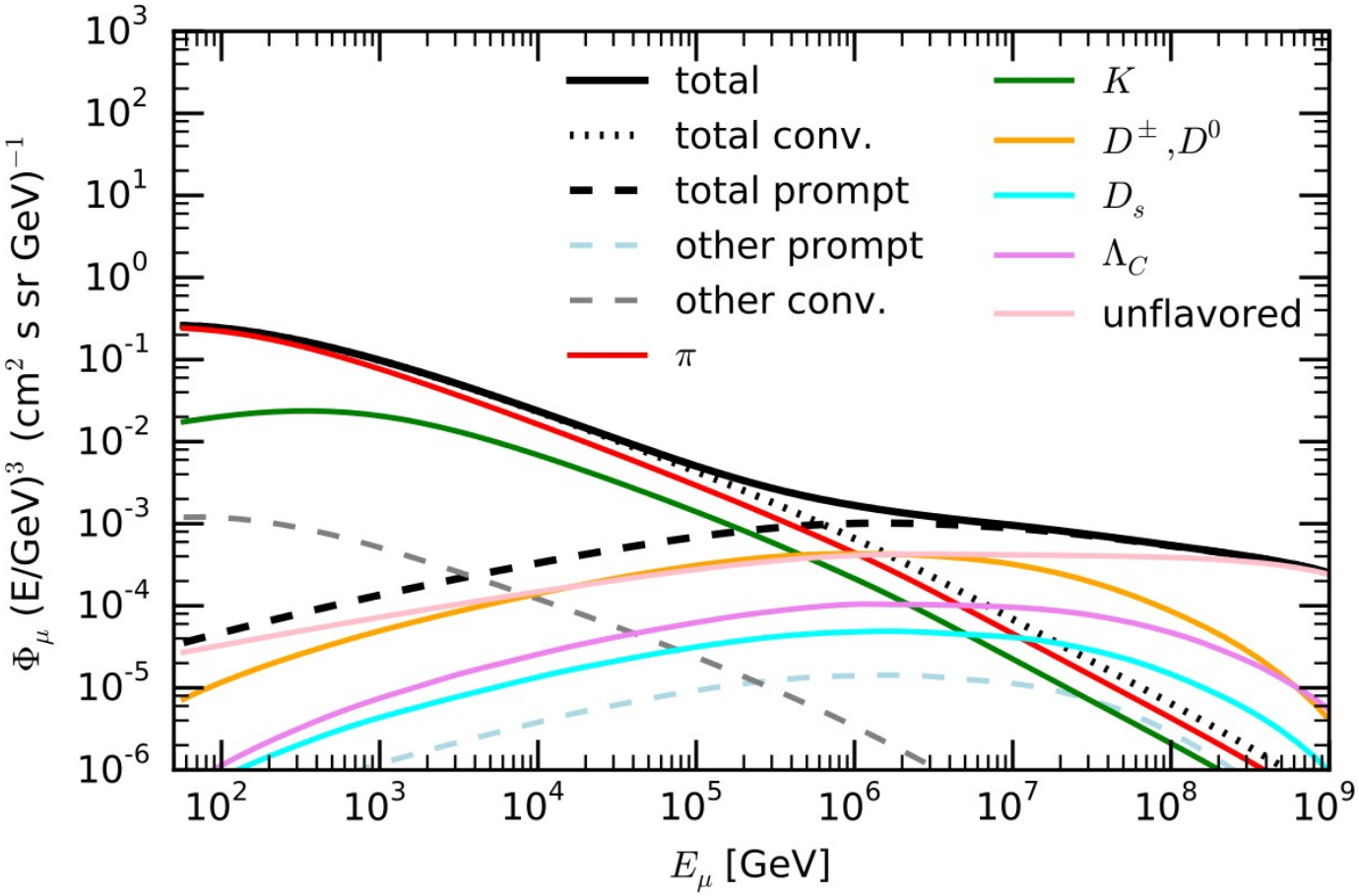
[2]

Muon bundle



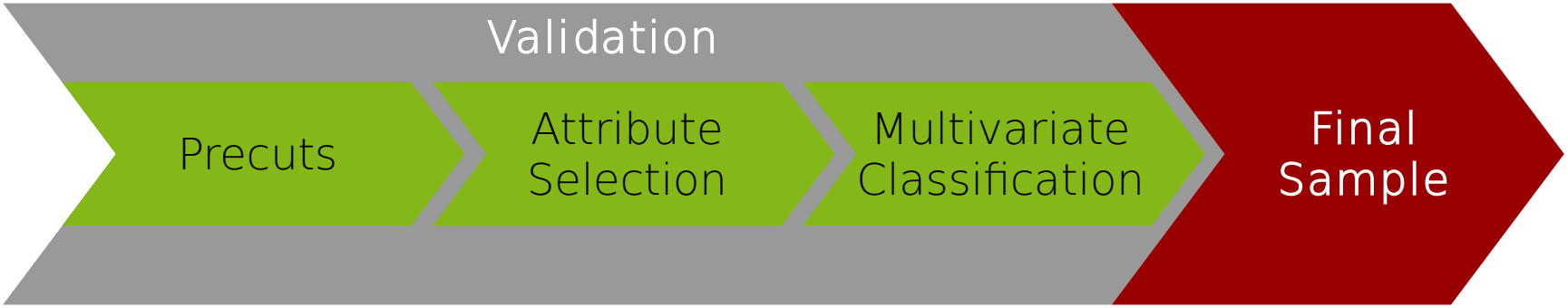
[2]

The atmospheric muon flux



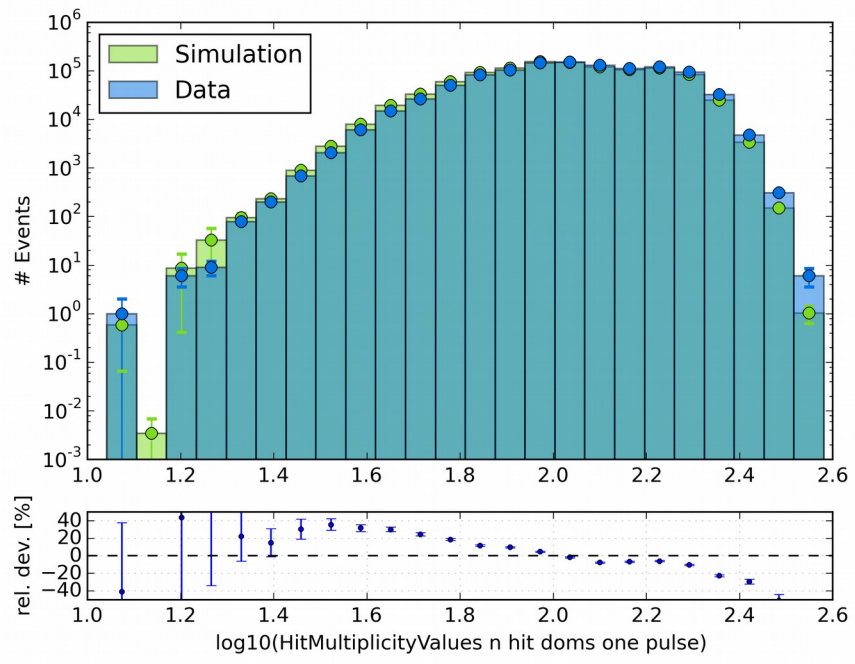
[3]

Supervised learning: Analysis chain

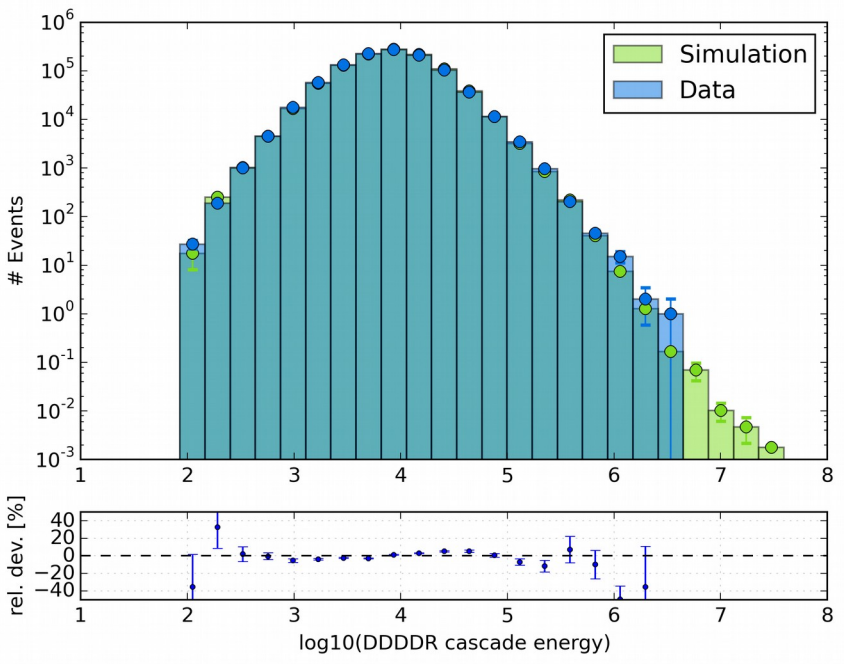


Validation

Part I:



bad



Good [2]

Attribute Selection

Attribute Selection

Min. Redundancy Max. Relevance (mRMR)

Attributes may be correlated

Optimize:

max. relevance
min. redundancy

$$\max_{x_j \in X - S_{m-1}} \left[I(x_y, c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j, x_i) \right]$$

[4]

Multivariate Classification

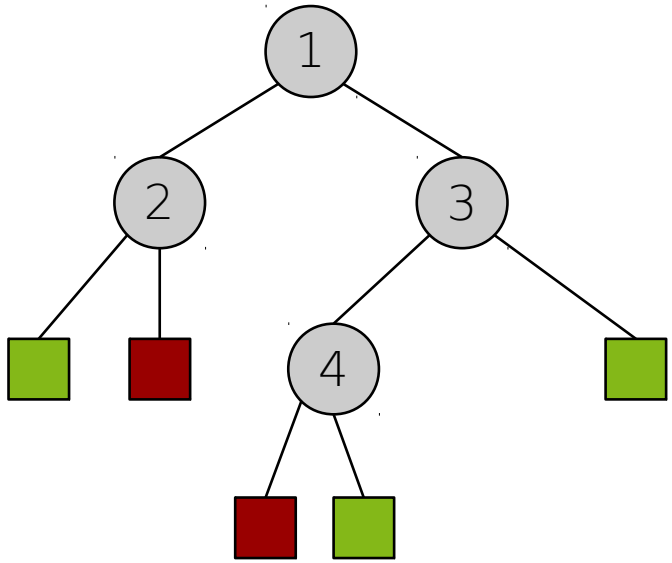
Multivariate Classification

Decision Tree

- Nodes, Leafs
- Best cut of all attributes in each node
- Leafs contain probabilities

Building a tree from top to bottom,

cutting on one attribute at each node.



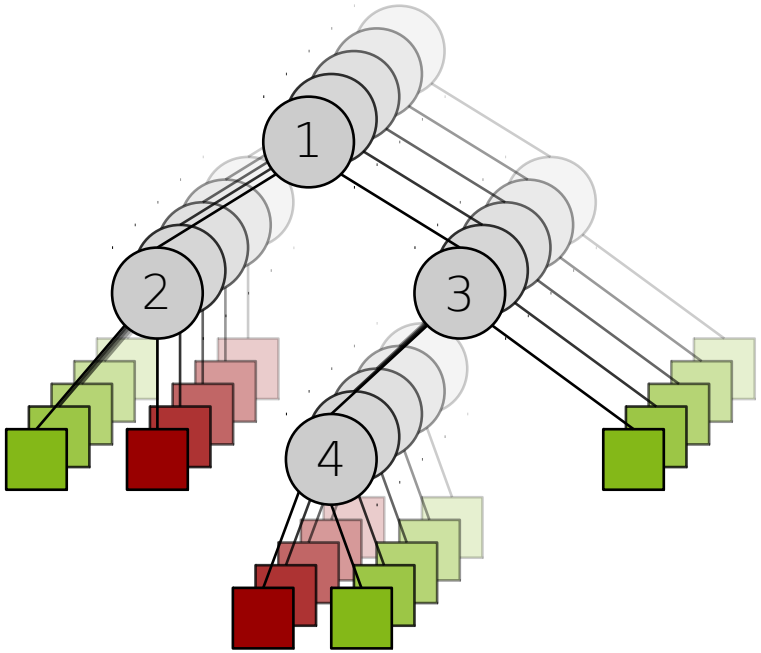
[5]

Multivariate Classification

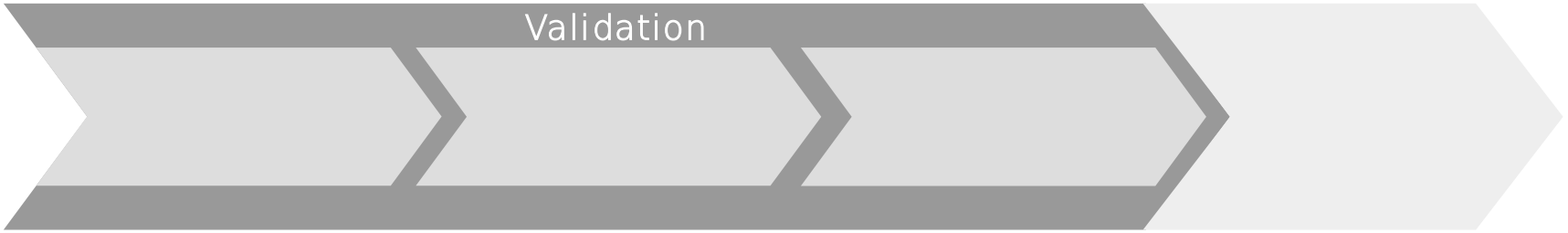
Random Forest

- Boostrap sample Z of size n from training data
- Grow random tree:
  - Select m attributes at random
  - Pick best attribute/split-point
  - Split node into two daughter nodes

→ Output ensemble of trees



[6]



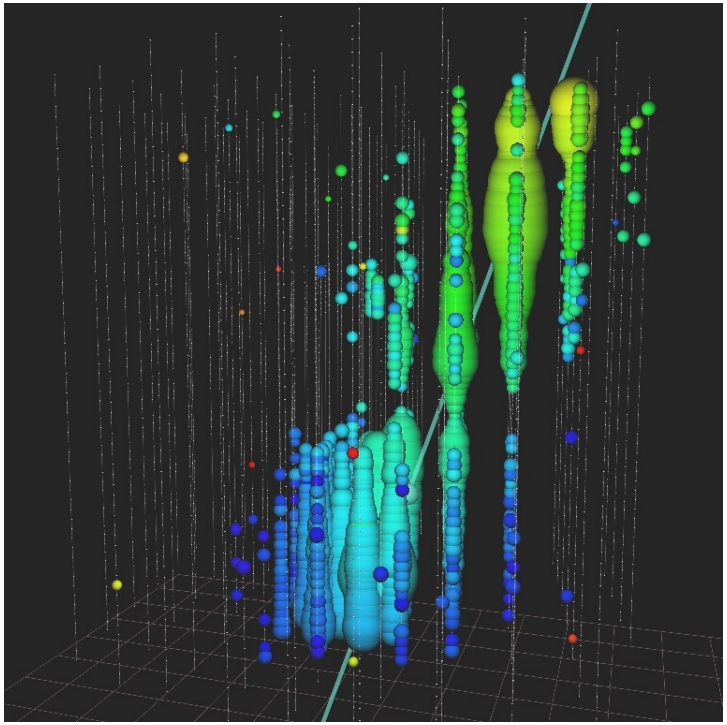
Part II:

X - Validation

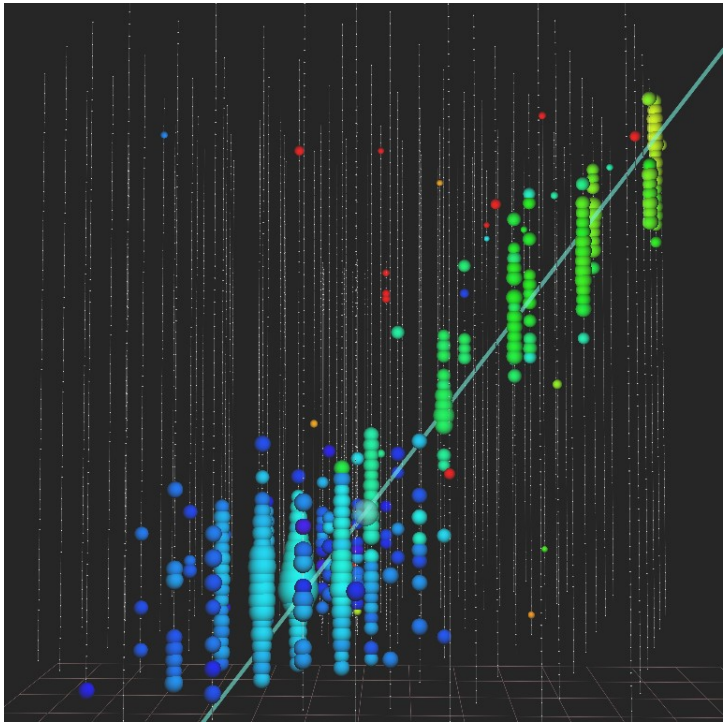
Test	Train	Train	Train	Train
Train	Test	Train	Train	Train
Train	Train	Test	Train	Train
Train	Train	Train	Test	Train
Train	Train	Train	Train	Test

[7]

Muon bundle



HE-Muon



References

[1] Kevin P. Murphy. *Machine Learning - A Probabilistic Perspective*. 2012

[2] T. Fuchs *Private Communication*. 2015

[3] Anatoli Fedynitch, et al. *A new version of the event generator Sibyll*. PoS(ICRC2015)

[4] H. Peng, F. Long, and C. Ding. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. 2005

[5] B. P. Roe, et al. *Boosted decision trees as an alternative to artificial neural networks for particle identification*. 2005

[6] L. Breiman. *Random Forests*. 2001

[7] A. W. Moore. *Cross-validation for detecting and preventing overfitting*. http://www.cs.cmu.edu/~./awm/tutorials.html

[8] Stuart Russell, Peter Norvig. *Artificial Intelligence – A modern Approach*. 1995

[9] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2013

[10] A. Criminisi, J. Shotton and E. Konukoglu. *Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning*. *2011*

Multivariate Classification

How to choose a cut?

- Information gain

$$I(A) = H\left(\frac{p}{p+n}, \frac{n}{p+n}\right) - EH(A)$$

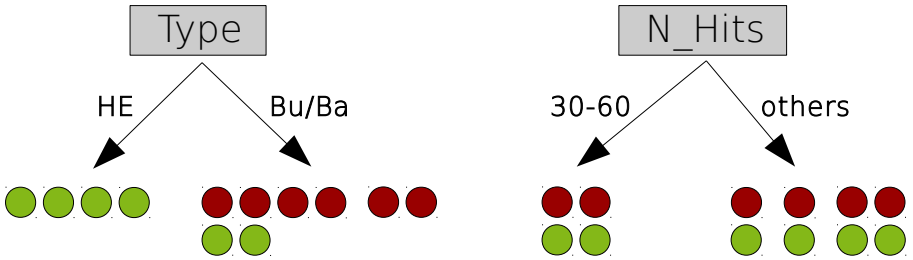
$$EH(A) = \sum_{i=1}^{k} \frac{p_i + n_i}{p+n} H\left(\frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i}\right)$$

$$H\left(\frac{p}{p+n}, \frac{n}{p+n}\right) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$

[8]

Multivariate Classification

How to choose a cut?



Convention: $p = n = 6, \quad H\left(\frac{6}{12}, \frac{6}{12}\right) = 1\,bit$

$$I(\text{Type}) = 1 - \left[\frac{4}{12}H(1,0) + \frac{8}{12}H\left(\frac{2}{8}, \frac{6}{8}\right)\right] = 0.541\,bit$$

$$I(\text{N\_Hits}) = 1 - \left[\frac{4}{12}H\left(\frac{2}{4}, \frac{2}{4}\right) + \frac{8}{12}H\left(\frac{4}{8}, \frac{4}{8}\right)\right] = 0\,bit$$

Example	Feature		Label
	Type	N_Hits	
X_1	HE	0-10	T
X_2	Bundle	10-30	F
X_3	HE	30-60	T
X_4	Bundle	10-30	T
X_5	Bundle	0-10	F
X_6	HE	>60	T
X_7	Balloon	30-60	F
X_8	HE	10-30	T
X_9	Bundle	30-60	F
X_{10}	Bundle	>60	F
X_{11}	Balloon	10-30	F
X_{12}	Bundle	30-60	T