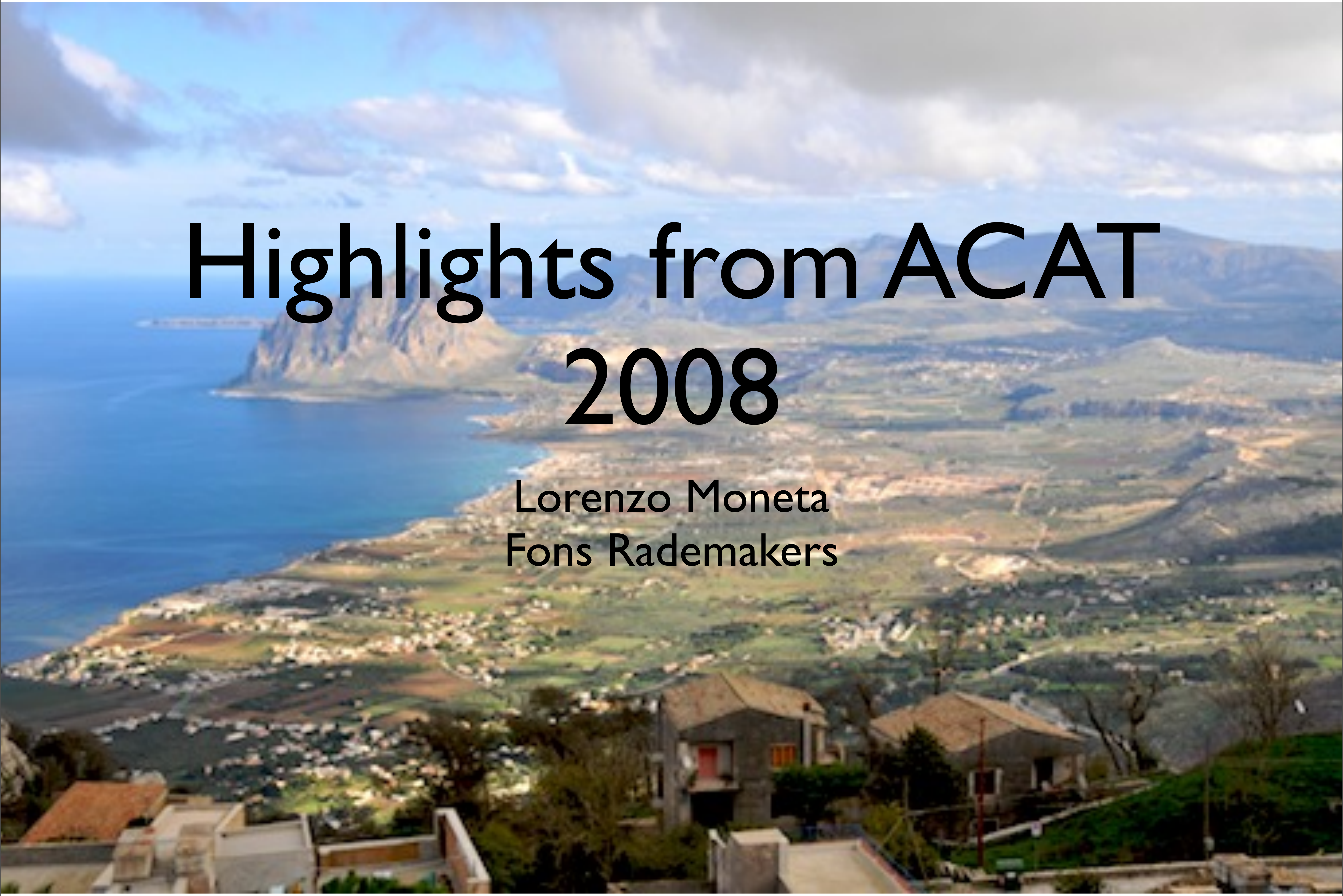
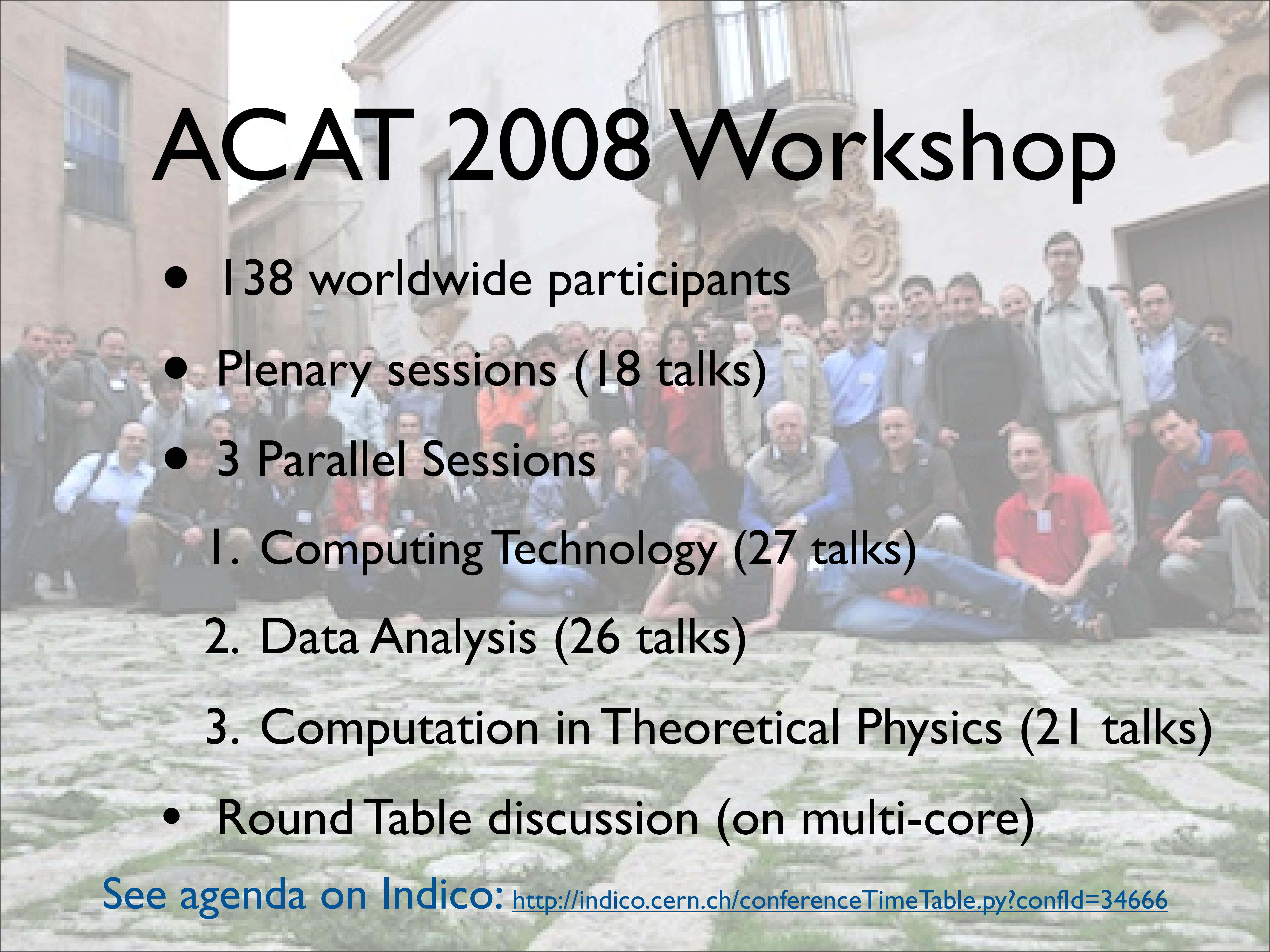


Highlights from ACAT 2008

Lorenzo Moneta
Fons Rademakers

ACAT 2008 Workshop



- 138 worldwide participants
- Plenary sessions (18 talks)
- 3 Parallel Sessions

1. Computing Technology (27 talks)
2. Data Analysis (26 talks)
3. Computation in Theoretical Physics (21 talks)

- Round Table discussion (on multi-core)

See agenda on Indico: http://indico.cern.ch/conferenceTimeTable.py?confId=34666

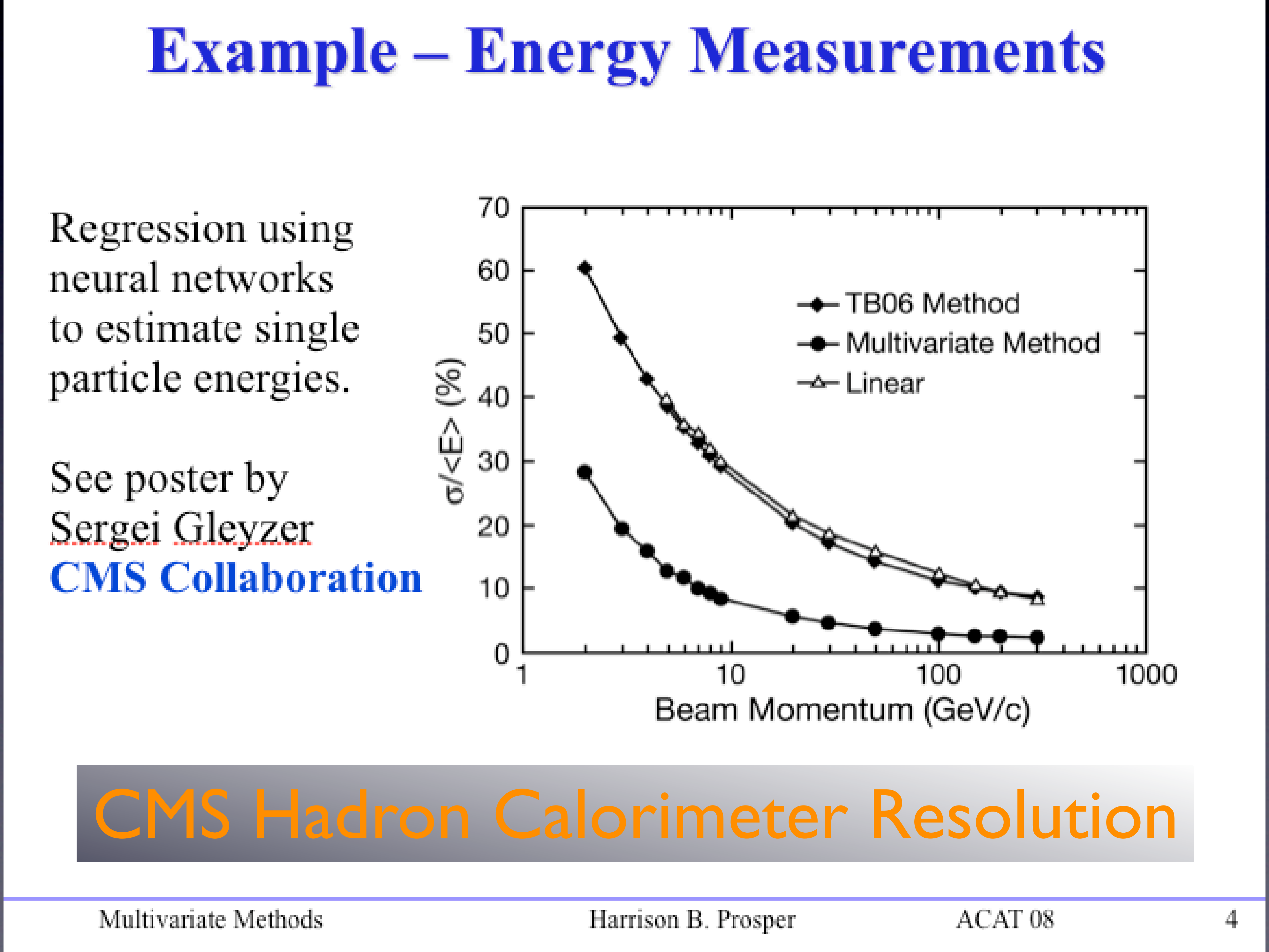
Outline

- Highlights from Data Analysis, algorithms and tools (session 2)
    - MultiVariate analysis methods
        - one plenary talk + 9 parallel talks
    - Parallelization, multi-core
- Summarize some interesting plenary talks
- Many interesting talks spanning various subjects
    - Apologize for not mentioning many of them (from Session 1 or 3)

Multi-Variate Methods

- Very nice plenary presentation from H. Prosper on MVA
- Various presentation on multivariate analysis methods in Session 2 (Data analysis, algorithms and tools)
    - MVA methods are gaining acceptance
    - getting attention from LHC experiments
- This year less emphasis on new classifiers
- Presentations on example of usage and comparisons between methods

Examples of MVA
(from H. Prosper talk)



Example – Energy Measurements

Regression using neural networks to estimate single particle energies.

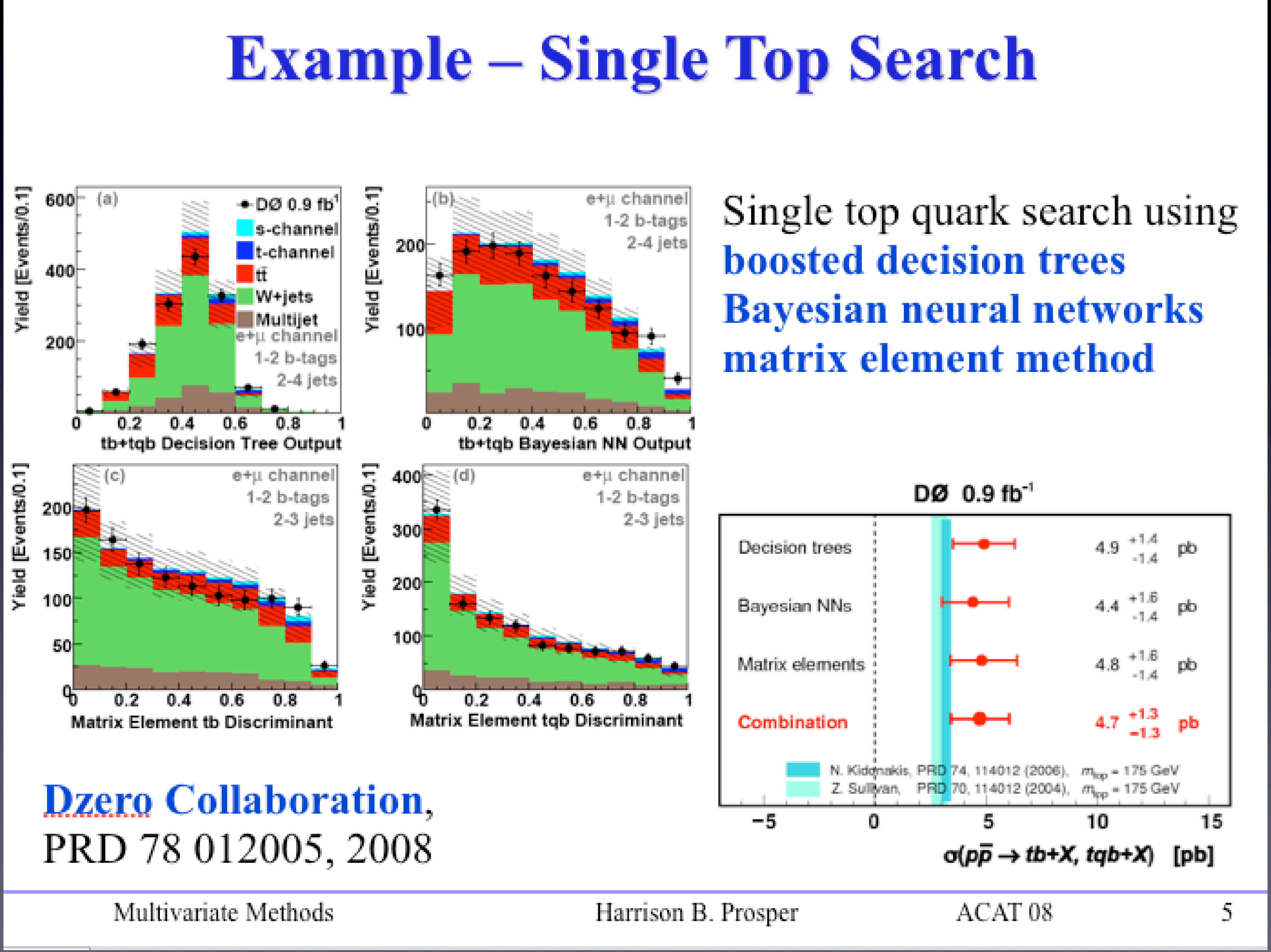
See poster by Sergei Gleyzer CMS Collaboration

CMS Hadron Calorimeter Resolution

Multivariate Methods Harrison B. Prosper ACAT 08 4

NN reflects better non linear sharing of energy among calorimeter towers

Examples of MVA
(from H. Prosper talk)



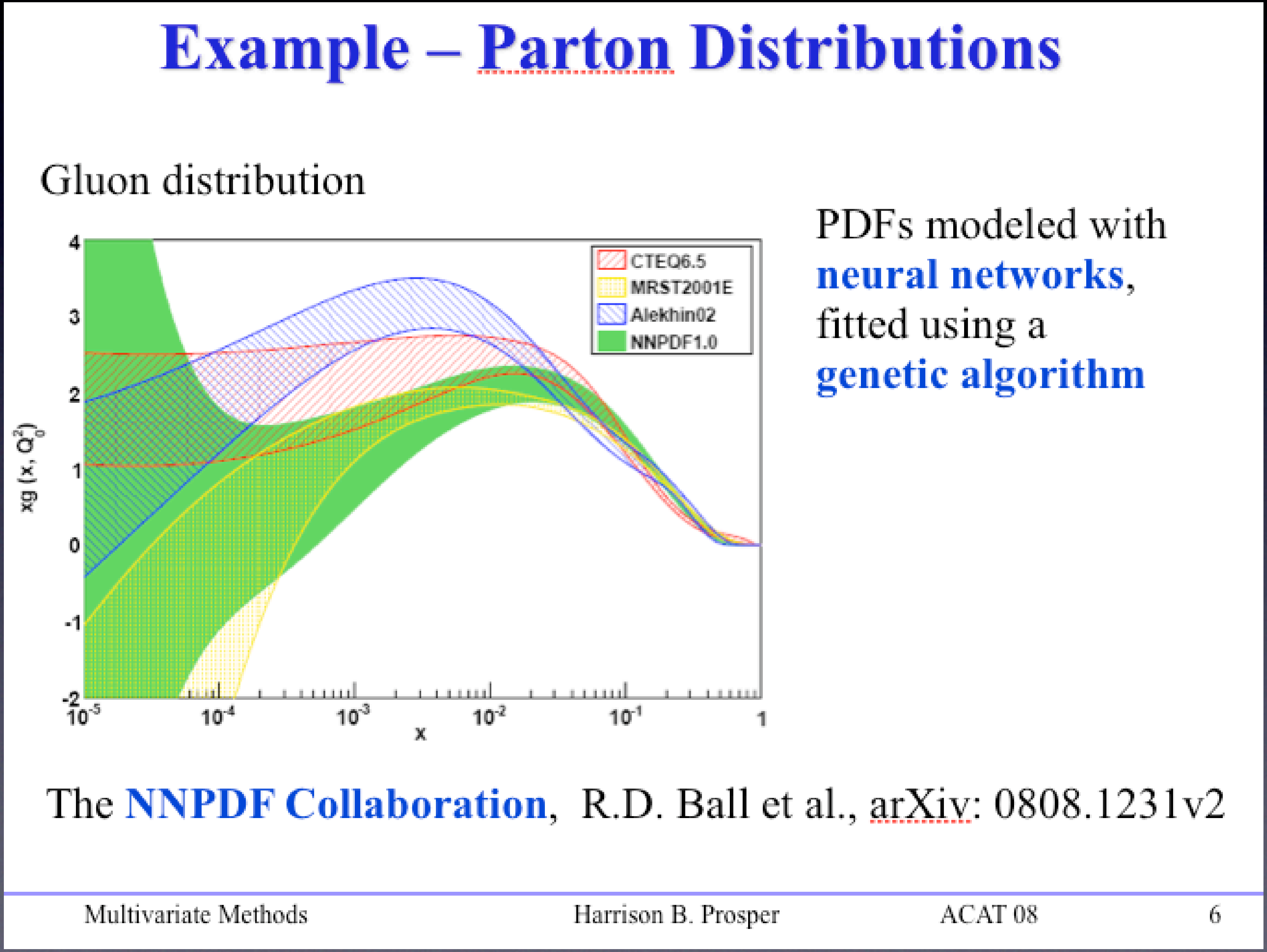
Example – Single Top Search

Single top quark search using
boosted decision trees
Bayesian neural networks
matrix element method

Dzero Collaboration,
PRD 78 012005, 2008

Multivariate Methods Harrison B. Prosper ACAT 08 5

Examples of MVA
(from H. Prosper talk)



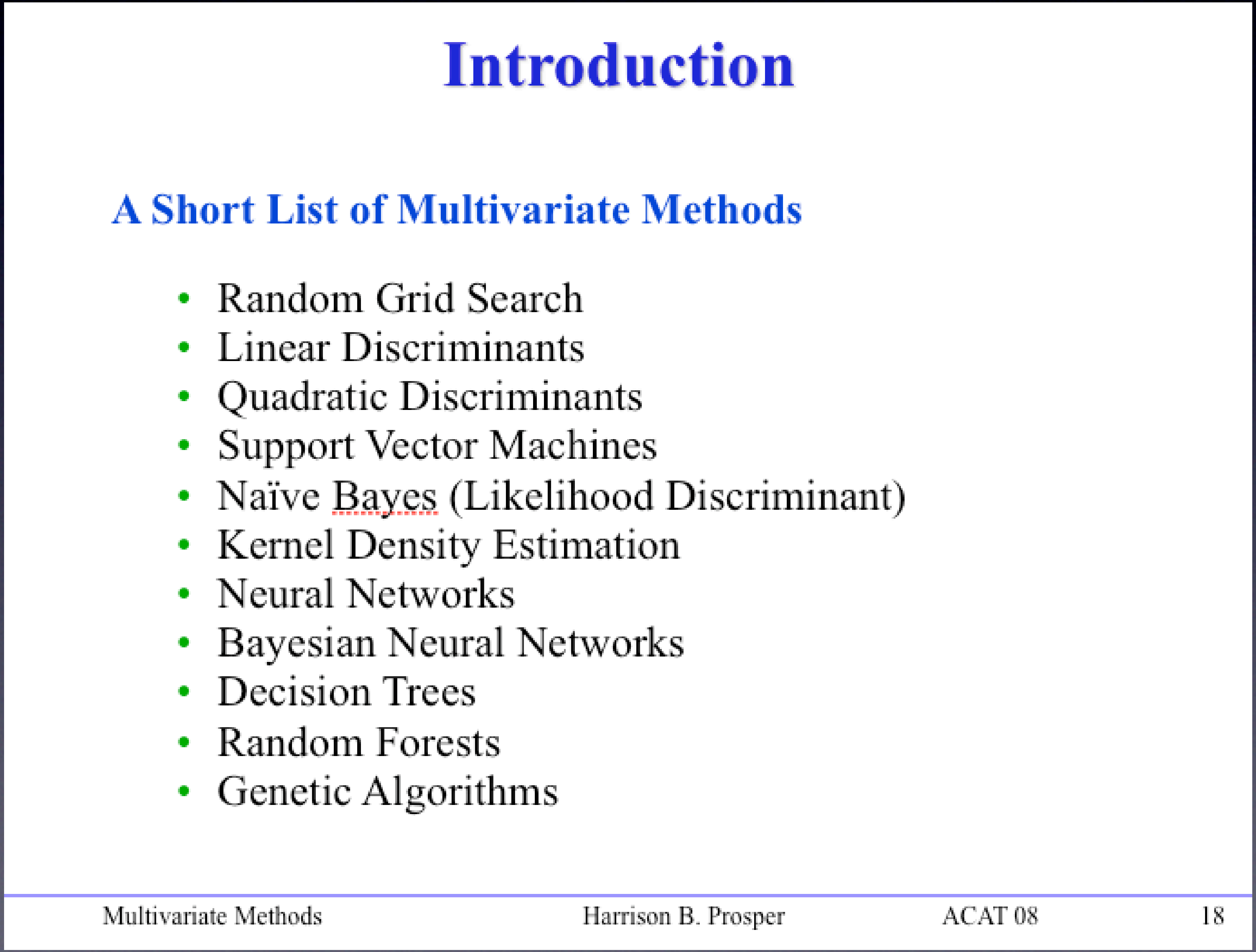
Example – Parton Distributions

Gluon distribution

PDFs modeled with **neural networks**, fitted using a **genetic algorithm**

The **NNPDF Collaboration**, R.D. Ball et al., arXiv: 0808.1231v2

Multivariate Methods in Particle Physics Today and Tomorrow (H. Prosper)



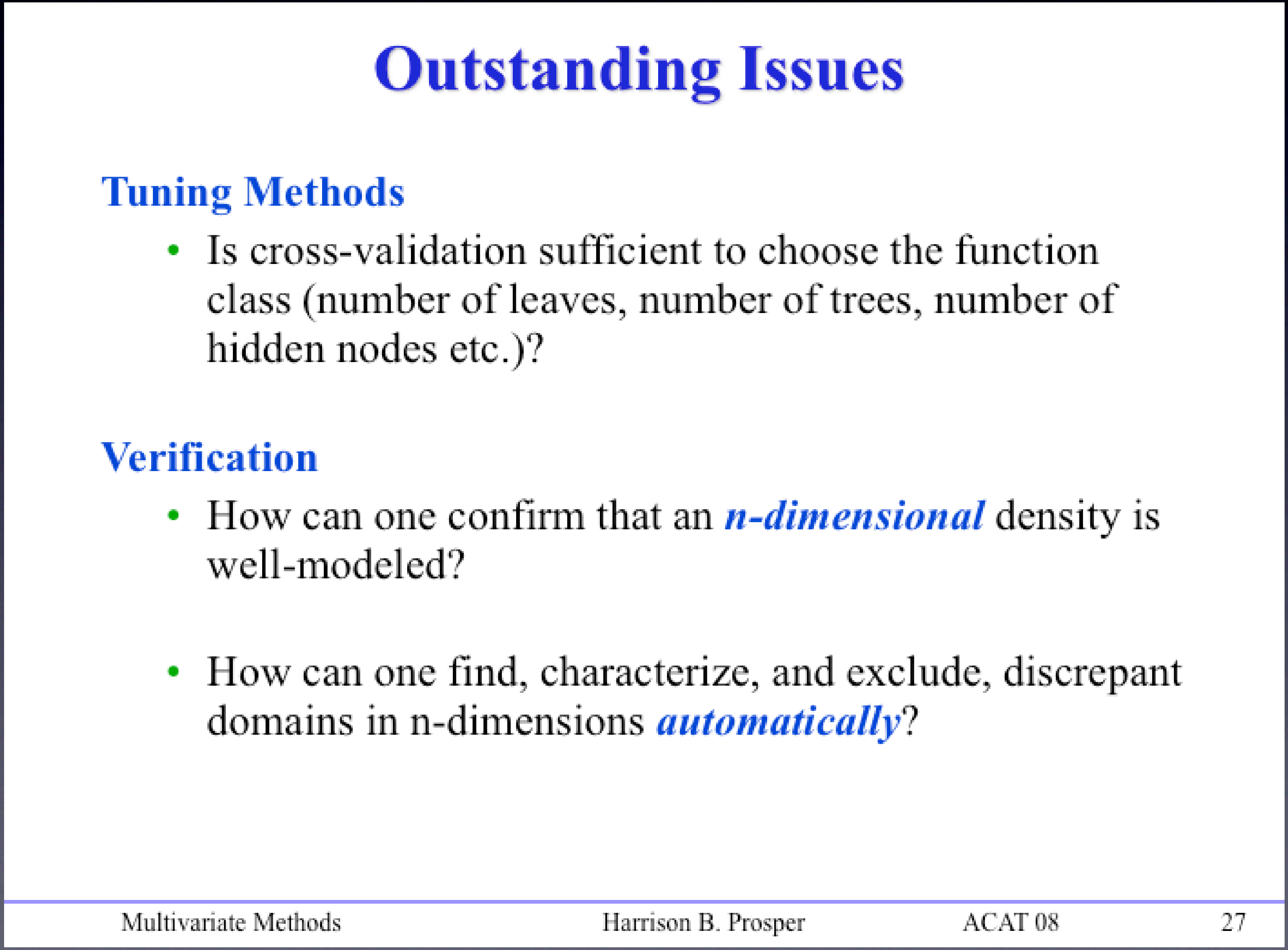
Introduction

A Short List of Multivariate Methods

- Random Grid Search
- Linear Discriminants
- Quadratic Discriminants
- Support Vector Machines
- Naïve Bayes (Likelihood Discriminant)
- Kernel Density Estimation
- Neural Networks
- Bayesian Neural Networks
- Decision Trees
- Random Forests
- Genetic Algorithms

Multivariate Methods Harrison B. Prosper ACAT 08 18

Multivariate Methods in Particle Physics Today and Tomorrow (H. Prosper)



Outstanding Issues

Tuning Methods

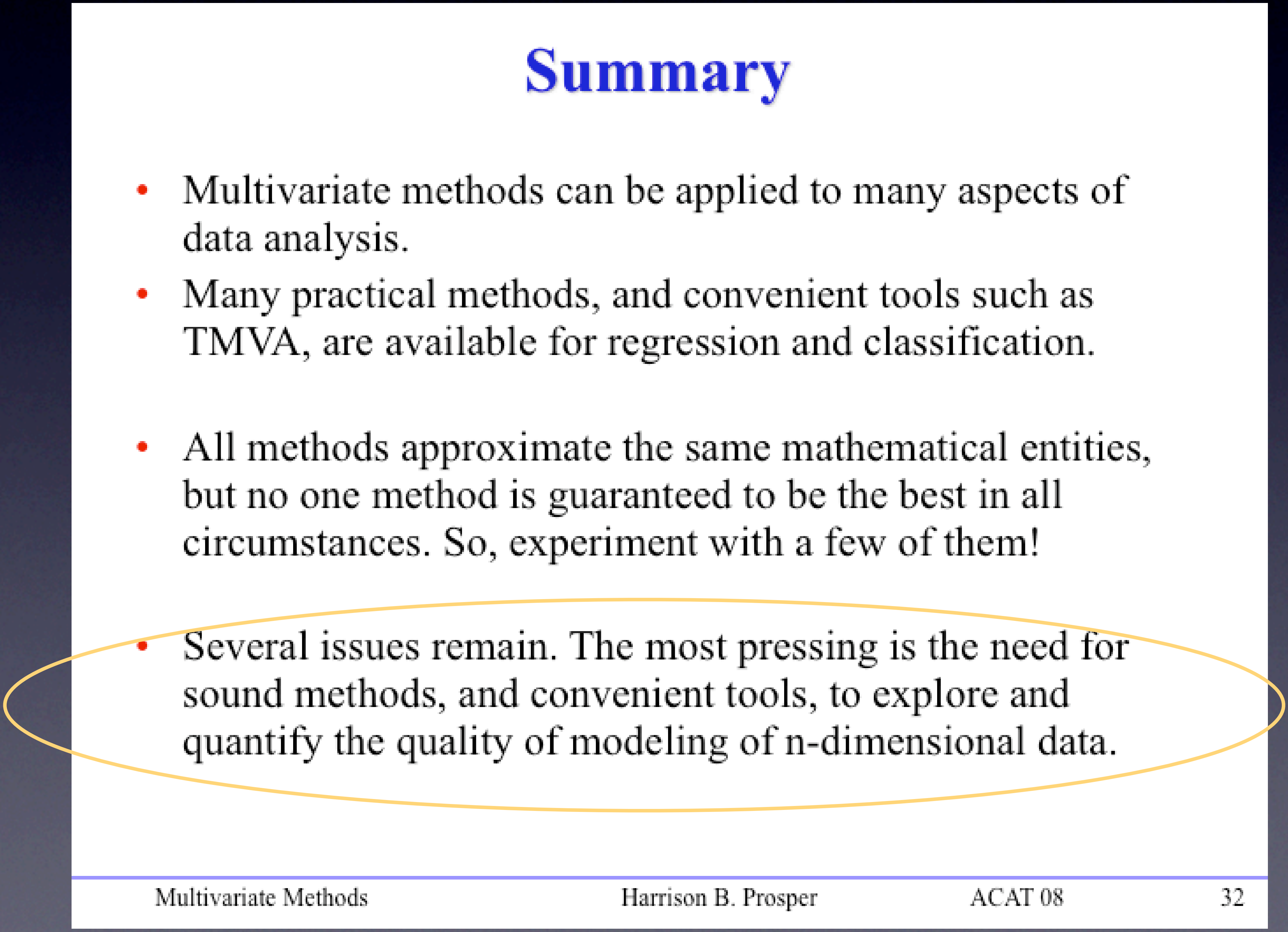
- Is cross-validation sufficient to choose the function class (number of leaves, number of trees, number of hidden nodes etc.)?

Verification

- How can one confirm that an ***n-dimensional*** density is well-modeled?

- How can one find, characterize, and exclude, discrepant domains in n-dimensions ***automatically***?

Multivariate Methods in Particle Physics Today and Tomorrow (H. Prosper)

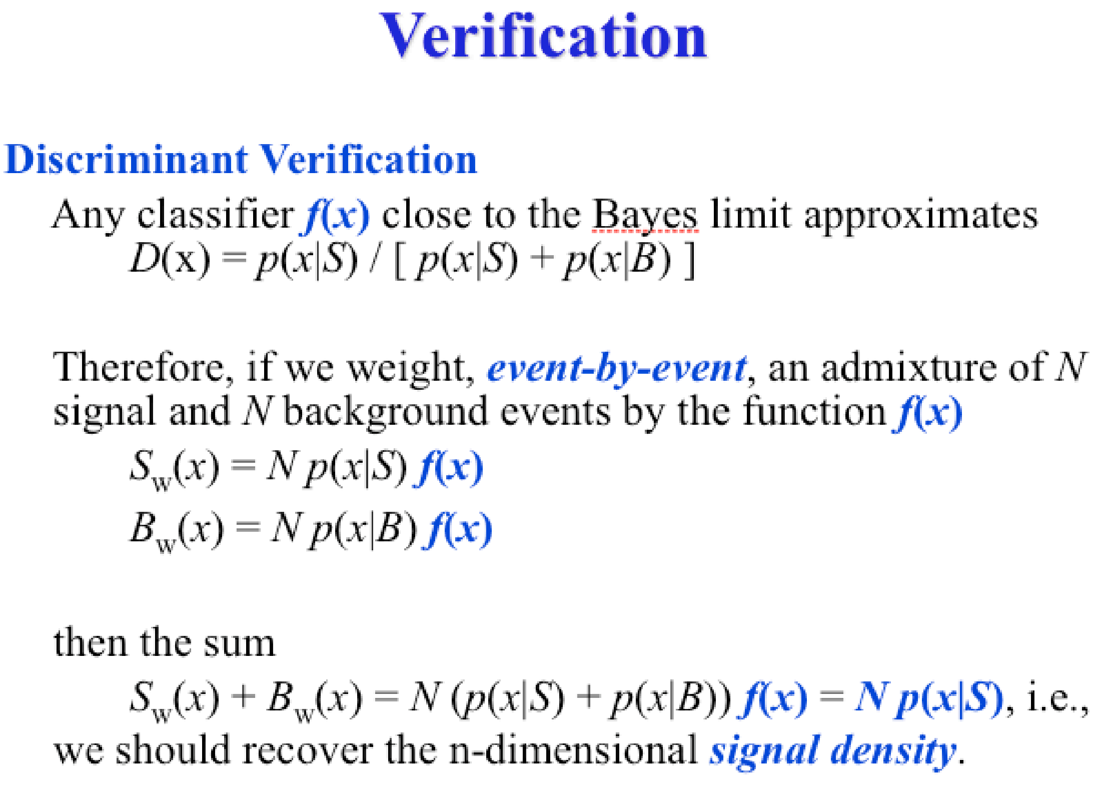


Summary

- Multivariate methods can be applied to many aspects of data analysis.
- Many practical methods, and convenient tools such as TMVA, are available for regression and classification.
- All methods approximate the same mathematical entities, but no one method is guaranteed to be the best in all circumstances. So, experiment with a few of them!
- Several issues remain. The most pressing is the need for sound methods, and convenient tools, to explore and quantify the quality of modeling of n-dimensional data.

Multivariate Methods Harrison B. Prosper ACAT 08 32

Multivariate Methods in Particle Physics Today and Tomorrow (H. Prosper)



Verification

Discriminant Verification

Any classifier $f(x)$ close to the Bayes limit approximates

$D(\mathrm{x}) = p(x|S) \,/\, [\, p(x|S) + p(x|B)\,]$

Therefore, if we weight, ***event-by-event***, an admixture of N signal and N background events by the function $f(x)$

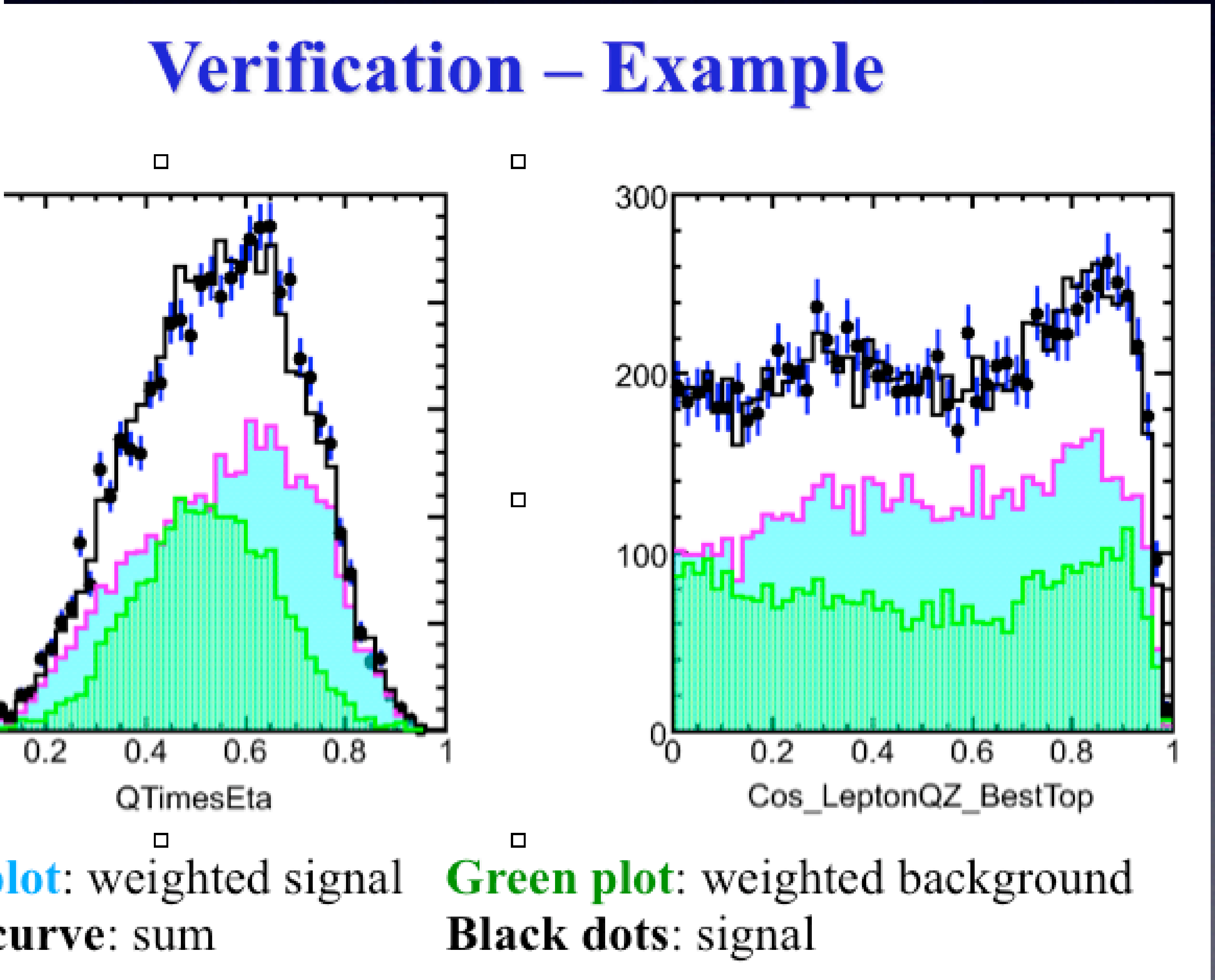
$S_w(x) = N\, p(x|S)\, f(x)$

$B_w(x) = N\, p(x|B)\, f(x)$

then the sum

$S_w(x) + B_w(x) = N\,(p(x|S) + p(x|B))\, f(x) = N\, p(x|S)$, i.e.,

we should recover the n-dimensional ***signal density***.



Verification – Example

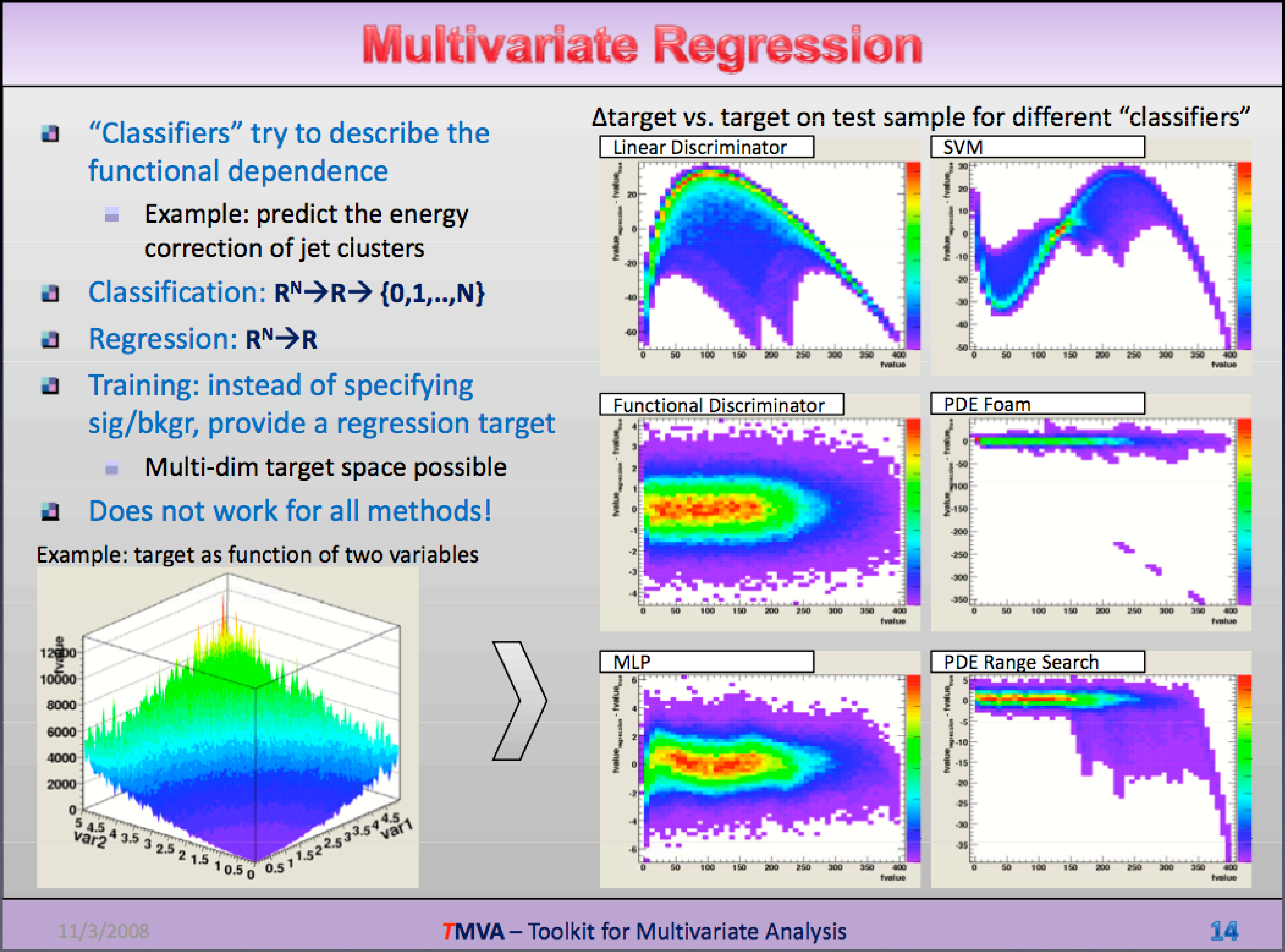
Cyan plot: weighted signal **Green plot**: weighted background

Black curve: sum **Black dots**: signal

TMVA

- Most popular packages in the community
- New version next year with various nice new features:
  - Support for regression in addition to classification
  - multi-class classification
  - automatic tuning using cross validation
  - generic boost or bag of any classifier
  - new method PDE-FOAM (separate talk)

TMVA (J. Stelzer)



Multivariate Regression

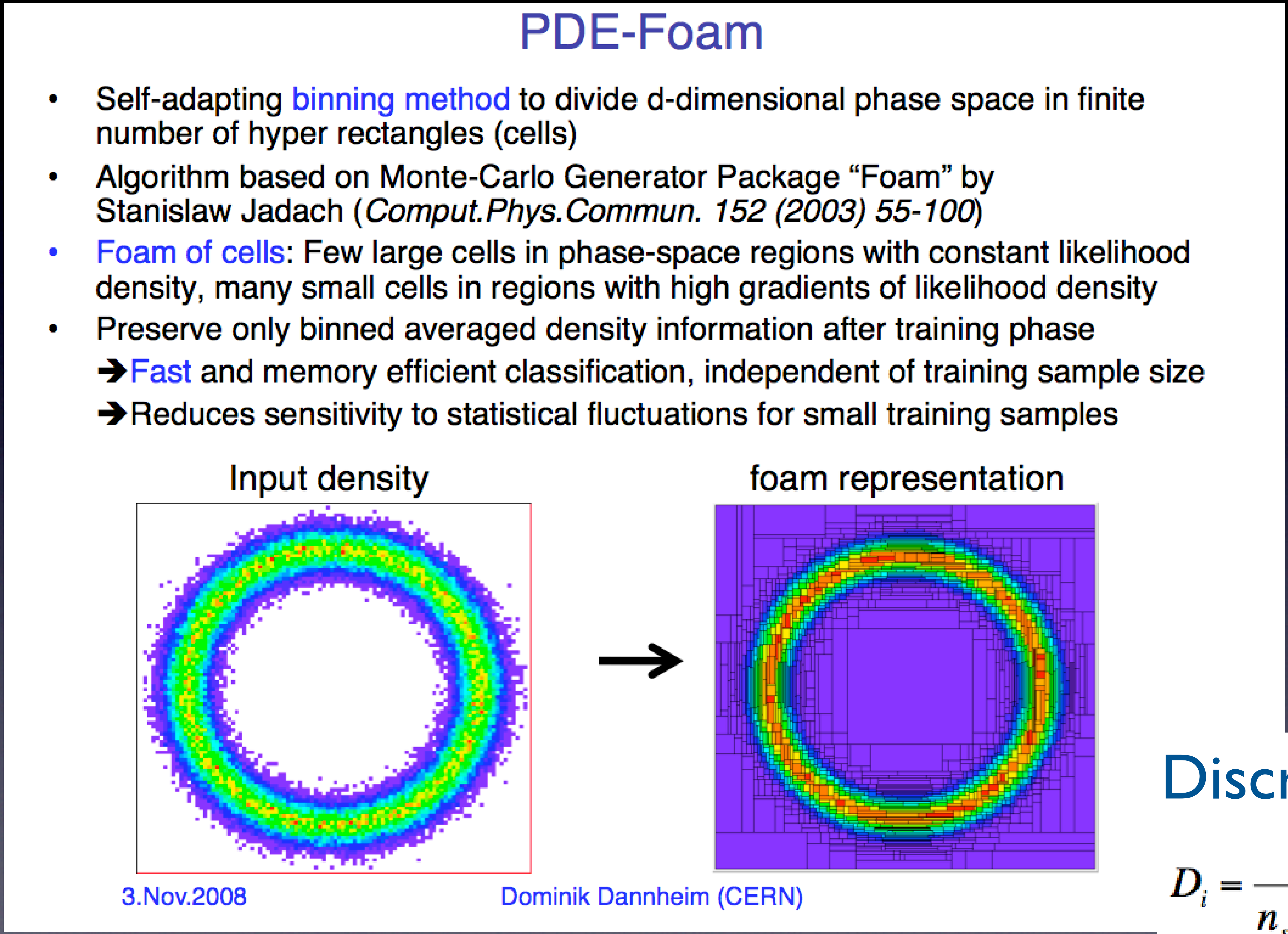
- “Classifiers” try to describe the functional dependence
  - Example: predict the energy correction of jet clusters
- Classification: **$R^N \rightarrow R \rightarrow \{0,1,..,N\}$**
- Regression: **$R^N \rightarrow R$**
- Training: instead of specifying sig/bkgr, provide a regression target
  - Multi-dim target space possible
- Does not work for all methods!

Example: target as function of two variables

Δtarget vs. target on test sample for different “classifiers”

11/3/2008 TMVA – Toolkit for Multivariate Analysis 14

PDE FOAM (D. Dannheim)



PDE-Foam

- Self-adapting binning method to divide d-dimensional phase space in finite number of hyper rectangles (cells)
- Algorithm based on Monte-Carlo Generator Package “Foam” by Stanislaw Jadach (*Comput.Phys.Commun. 152 (2003) 55-100*)
- Foam of cells: Few large cells in phase-space regions with constant likelihood density, many small cells in regions with high gradients of likelihood density
- Preserve only binned averaged density information after training phase

➔Fast and memory efficient classification, independent of training sample size

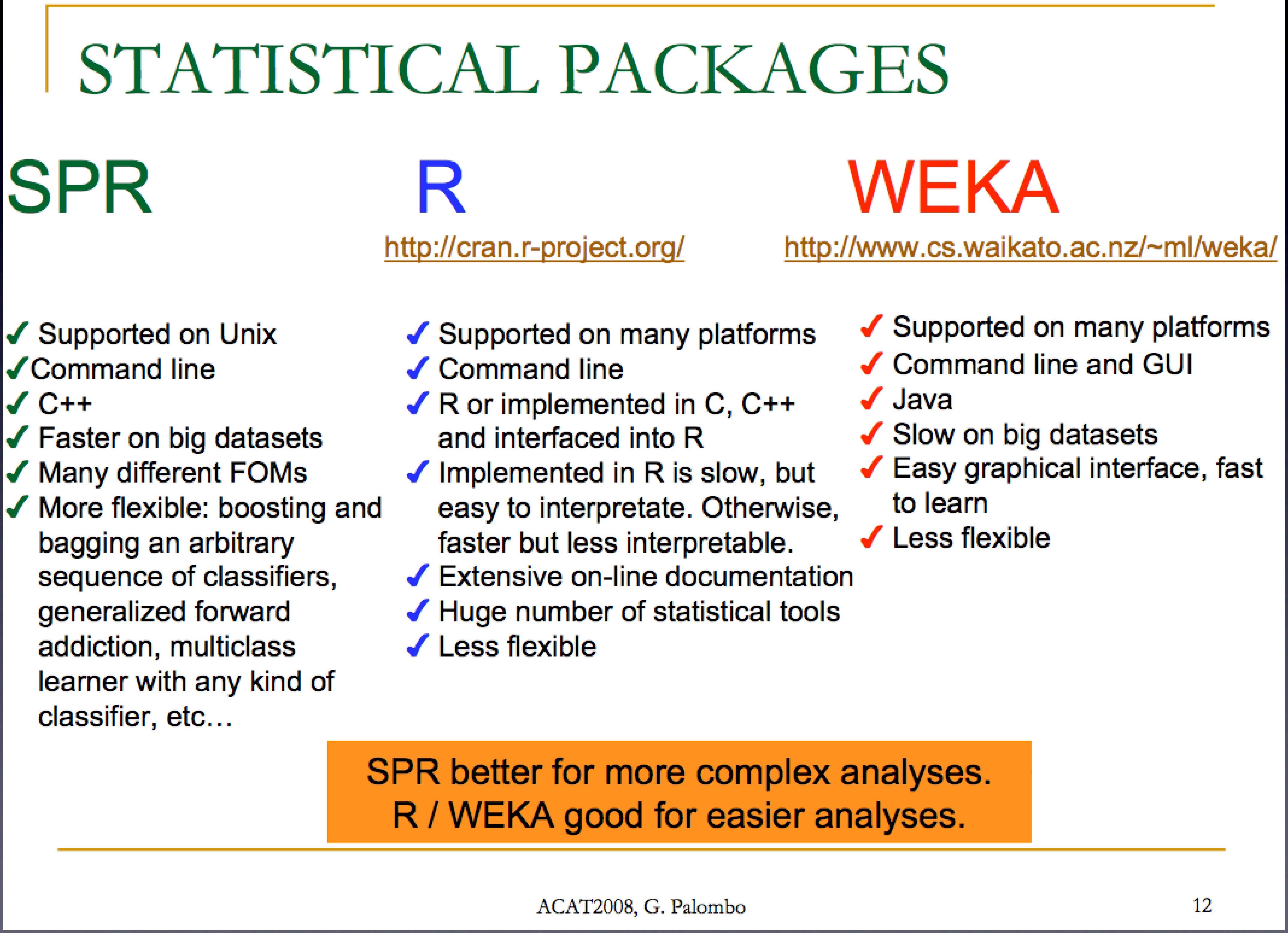
➔Reduces sensitivity to statistical fluctuations for small training samples

3.Nov.2008 Dominik Dannheim (CERN)

Discriminant

$$D_i = \frac{n_{sig,i}}{n_{sig,i} + c \cdot n_{bg,i}}$$

Others MVA Packages



STATISTICAL PACKAGES

SPR

- ✓ Supported on Unix
- ✓ Command line
- ✓ C++
- ✓ Faster on big datasets
- ✓ Many different FOMs
- ✓ More flexible: boosting and bagging an arbitrary sequence of classifiers, generalized forward addiction, multiclass learner with any kind of classifier, etc…

R

http://cran.r-project.org/

- ✓ Supported on many platforms
- ✓ Command line
- ✓ R or implemented in C, C++ and interfaced into R
- ✓ Implemented in R is slow, but easy to interpretate. Otherwise, faster but less interpretable.
- ✓ Extensive on-line documentation
- ✓ Huge number of statistical tools
- ✓ Less flexible

WEKA

http://www.cs.waikato.ac.nz/~ml/weka/

- ✓ Supported on many platforms
- ✓ Command line and GUI
- ✓ Java
- ✓ Slow on big datasets
- ✓ Easy graphical interface, fast to learn
- ✓ Less flexible

SPR better for more complex analyses.
R / WEKA good for easier analyses.

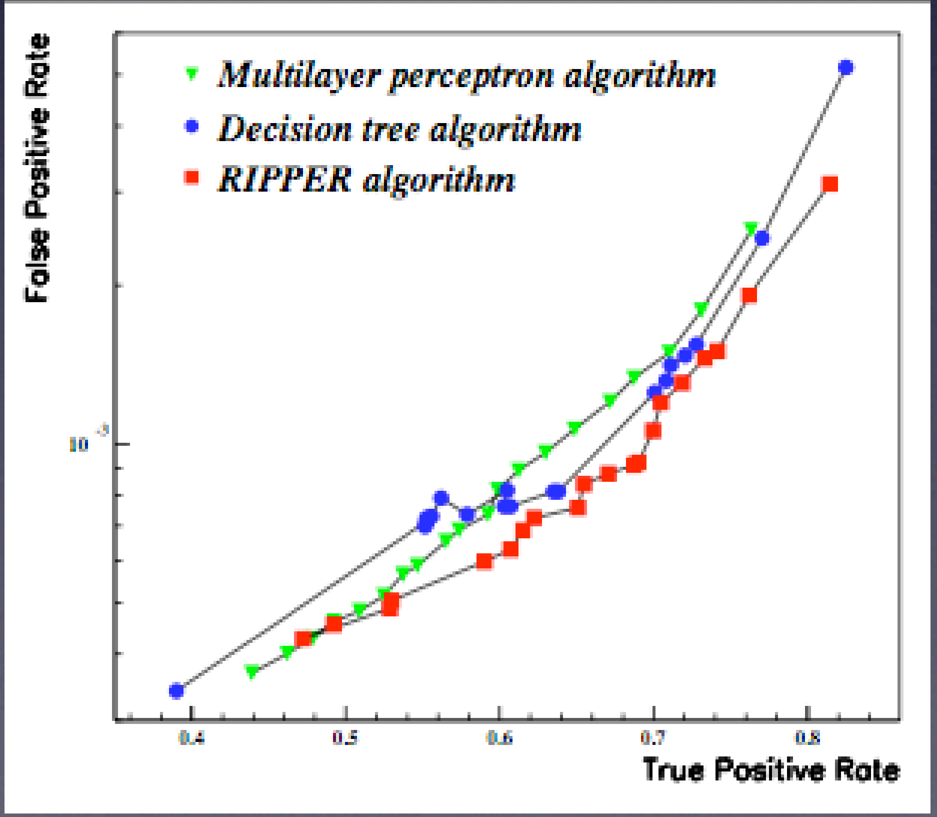
ACAT2008, G. Palombo

12

RIPPER (R. Britsch)

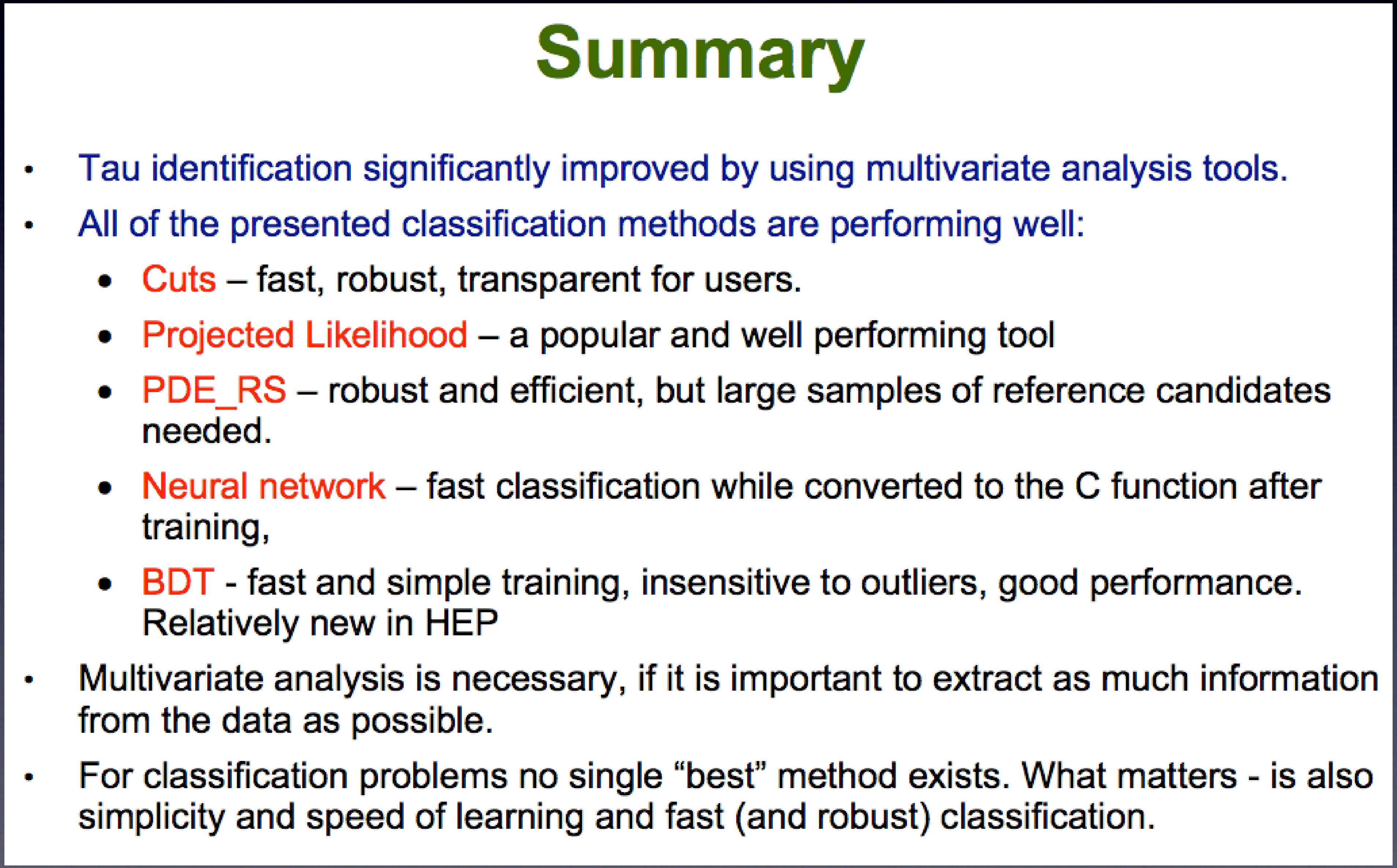
- RIPPER algorithm
  - classification of events using collection of if-then rules (direct rule based classifier)
- Instance weighting according to cost
  - assign cost to wrongly (or corrected) classified instances
- Use Bagging

Example: LHCb MC $\Lambda \rightarrow p_{+} + \pi_{-}$



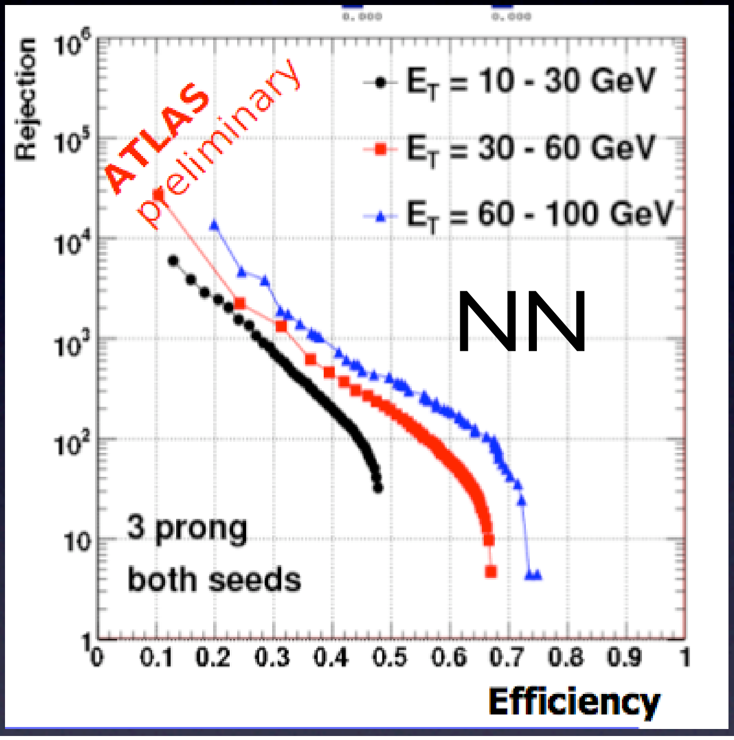
Examples of MVA Applications

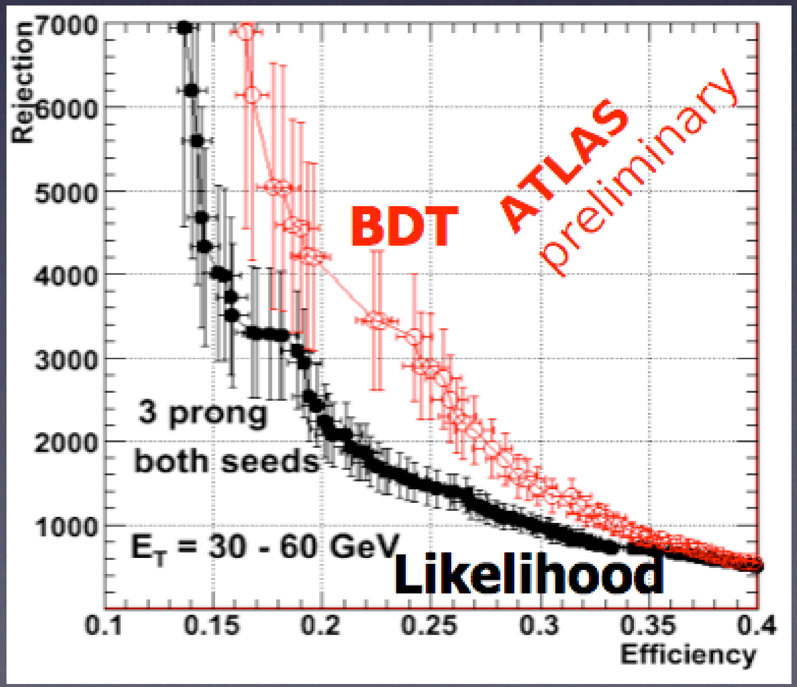
Tau Identification in ATLAS using MV tools (M. Wolter)



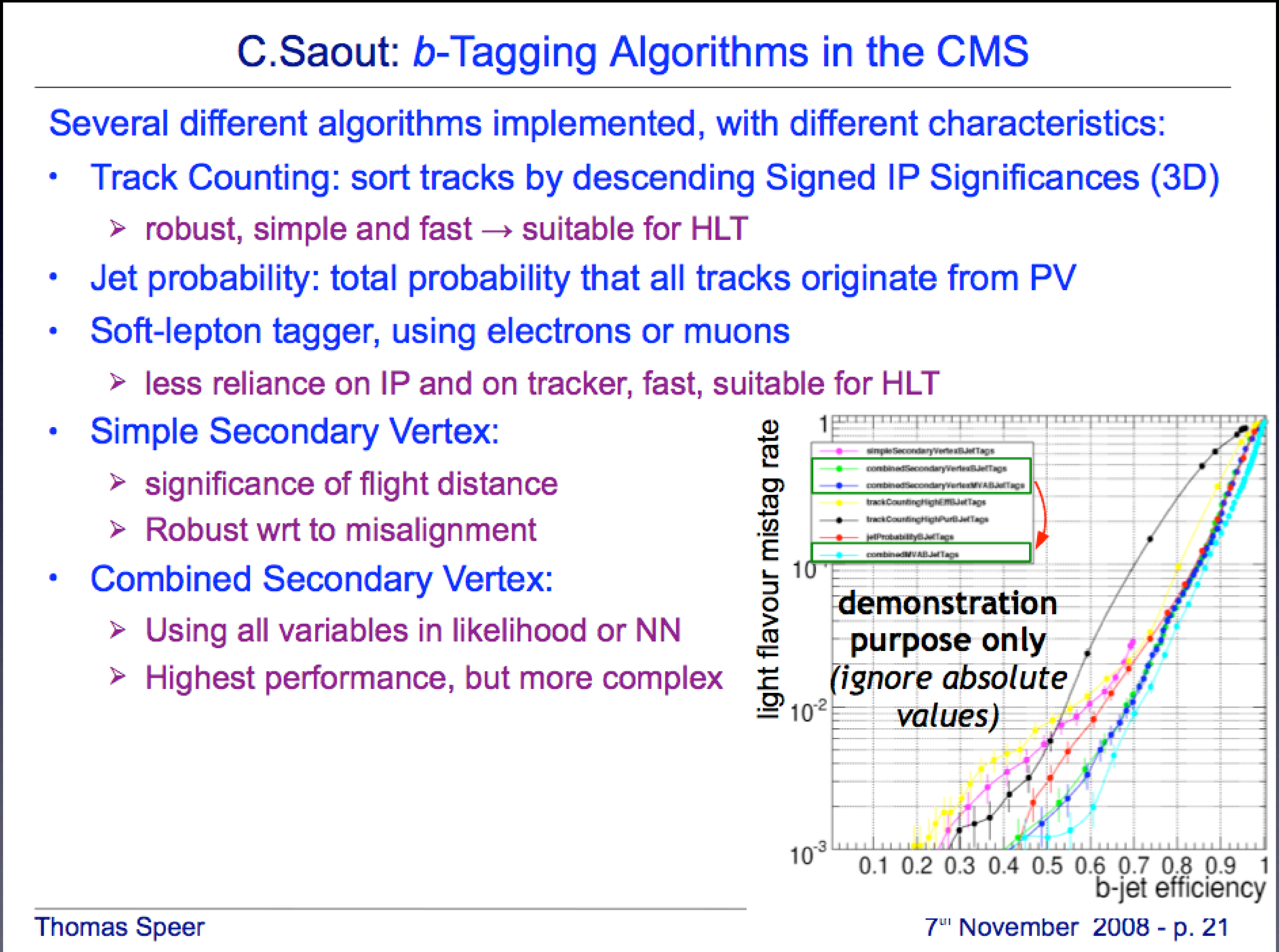
Summary

- Tau identification significantly improved by using multivariate analysis tools.
- All of the presented classification methods are performing well:
  - Cuts – fast, robust, transparent for users.
  - Projected Likelihood – a popular and well performing tool
  - PDE_RS – robust and efficient, but large samples of reference candidates needed.
  - Neural network – fast classification while converted to the C function after training,
  - BDT - fast and simple training, insensitive to outliers, good performance. Relatively new in HEP
- Multivariate analysis is necessary, if it is important to extract as much information from the data as possible.
- For classification problems no single “best” method exists. What matters - is also simplicity and speed of learning and fast (and robust) classification.





Examples of MVA Applications



C.Saout: *b*-Tagging Algorithms in the CMS

Several different algorithms implemented, with different characteristics:

- Track Counting: sort tracks by descending Signed IP Significances (3D)
  - ➢ robust, simple and fast → suitable for HLT
- Jet probability: total probability that all tracks originate from PV
- Soft-lepton tagger, using electrons or muons
  - ➢ less reliance on IP and on tracker, fast, suitable for HLT
- Simple Secondary Vertex:
  - ➢ significance of flight distance
  - ➢ Robust wrt to misalignment
- Combined Secondary Vertex:
  - ➢ Using all variables in likelihood or NN
  - ➢ Highest performance, but more complex

light flavour mistag rate
1
10^{-1}
10^{-2}
10^{-3}
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
b-jet efficiency
simpleSecondaryVertexBJetTags
combinedSecondaryVertexBJetTags
combinedSecondaryVertexMVABJetTags
trackCountingHighEffBJetTags
trackCountingHighPurBJetTags
jetProbabilityBJetTags
combinedMVABJetTags

demonstration purpose only *(ignore absolute values)*

Thomas Speer

7th November 2008 - p. 21

Paradigm (S. Gleyzer)
A Decision Making Framework for HEP

Decision making tool based on critical information

Relevant PARADIGM criteria

Relative Variable Importance RVI

Useful for :

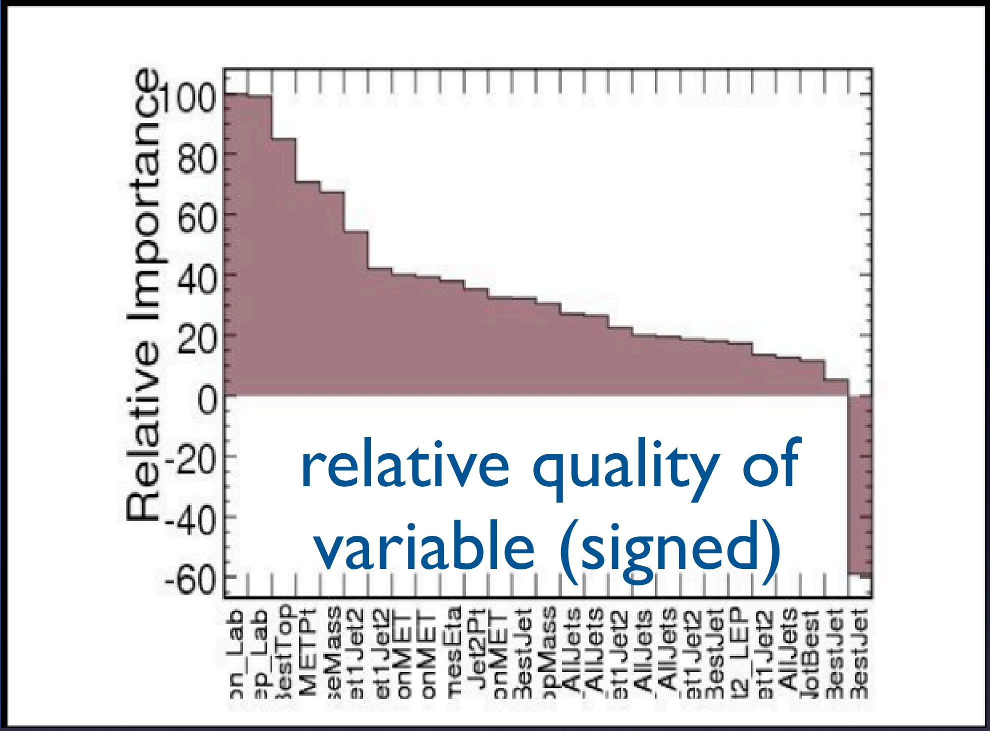
Variable Selection
Classifier Improvement

Global Loss Function GF

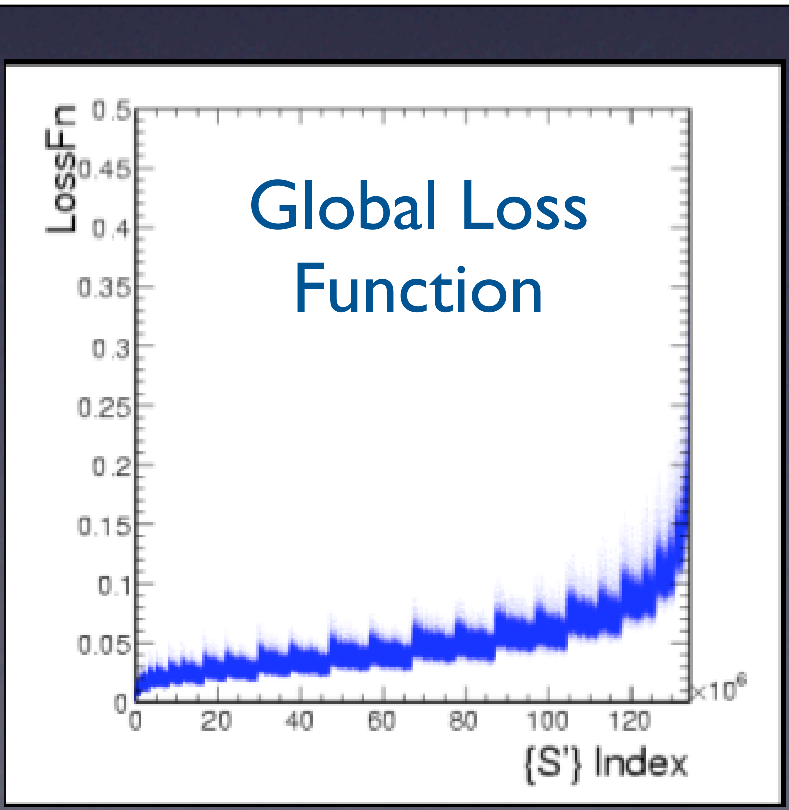
Useful for :

Variable Reduction
Classifier Selection

Lower the GF: lower loss of classification power from removing a subset {S'} from {V}

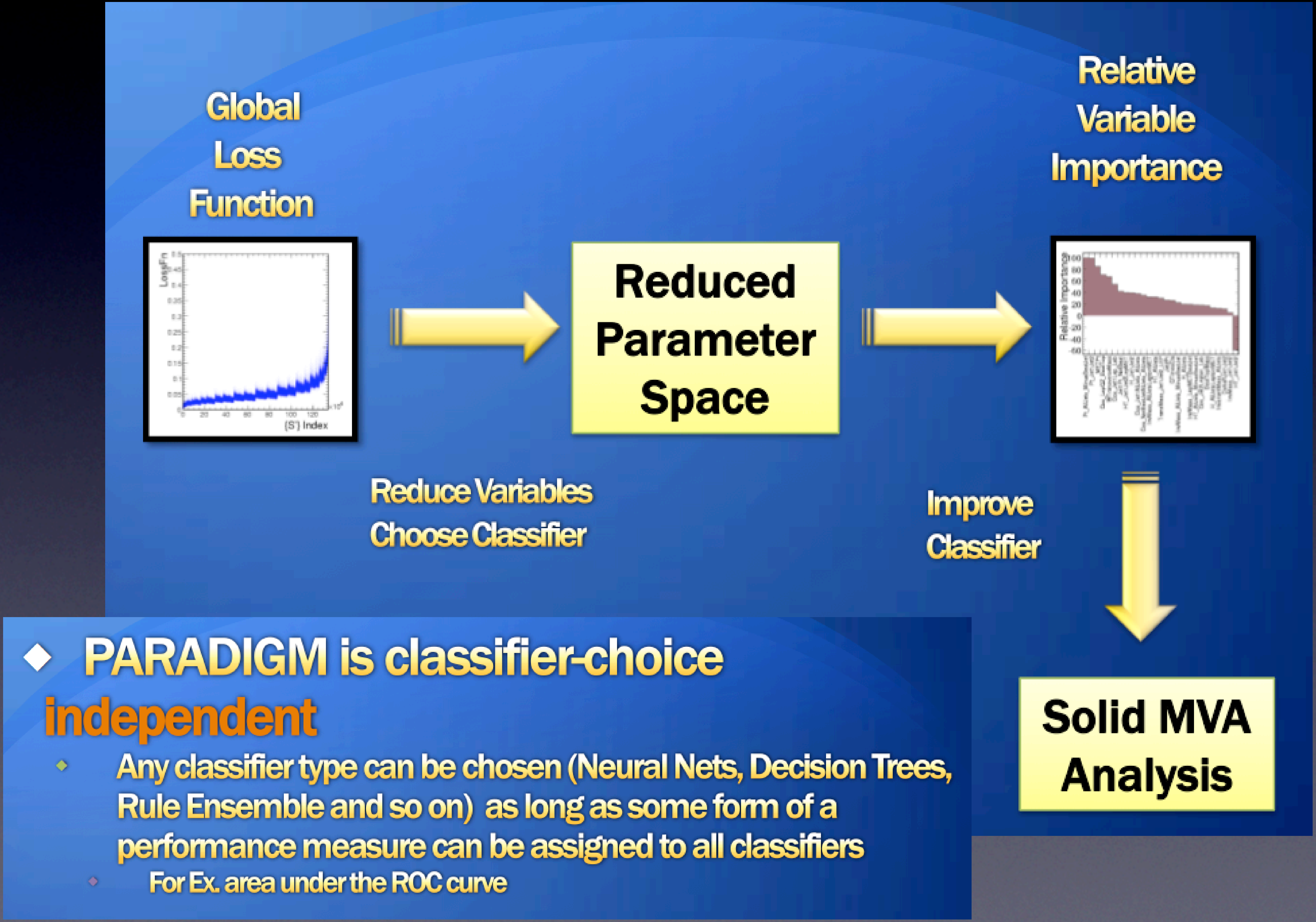


relative quality of variable (signed)



Global Loss Function

Paradigm (S. Gleyzer)



Global Loss Function

Reduced Parameter Space

Relative Variable Importance

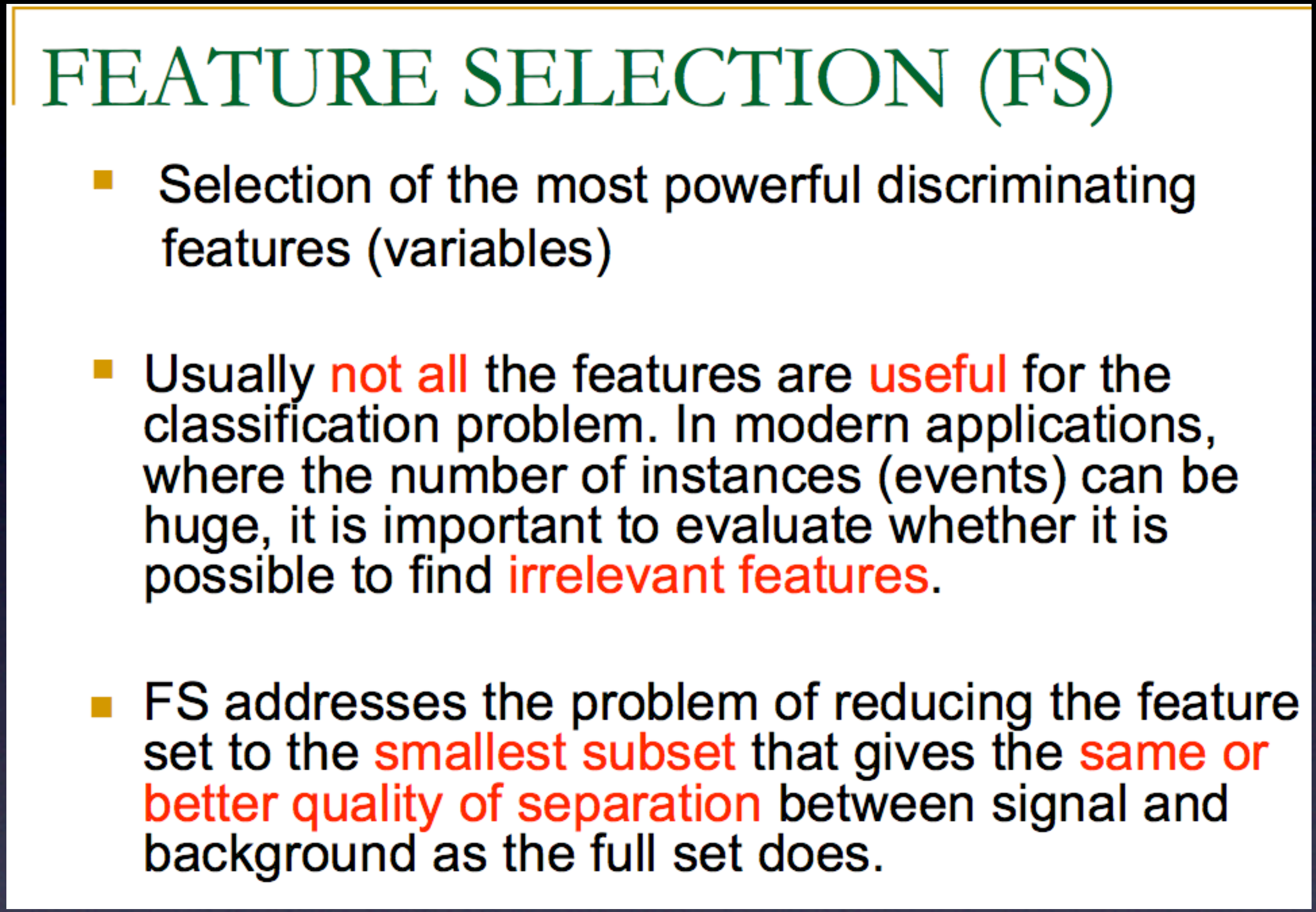
Reduce Variables
Choose Classifier

Improve Classifier

Solid MVA Analysis

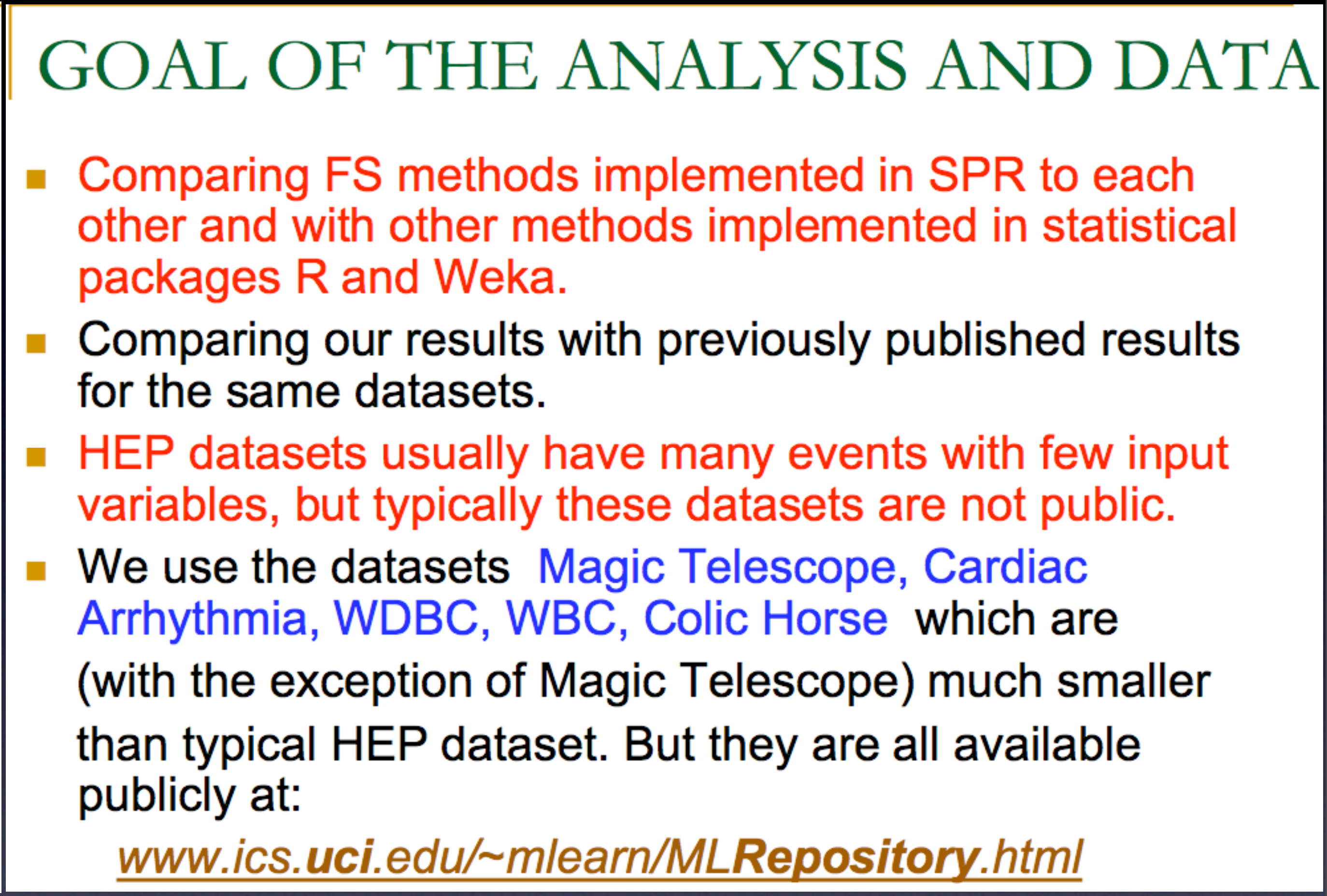
- PARADIGM is classifier-choice independent
  - Any classifier type can be chosen (Neural Nets, Decision Trees, Rule Ensemble and so on) as long as some form of a performance measure can be assigned to all classifiers
    - For Ex. area under the ROC curve

Feature Selection Algorithms
(G. Palombo)



FEATURE SELECTION (FS)

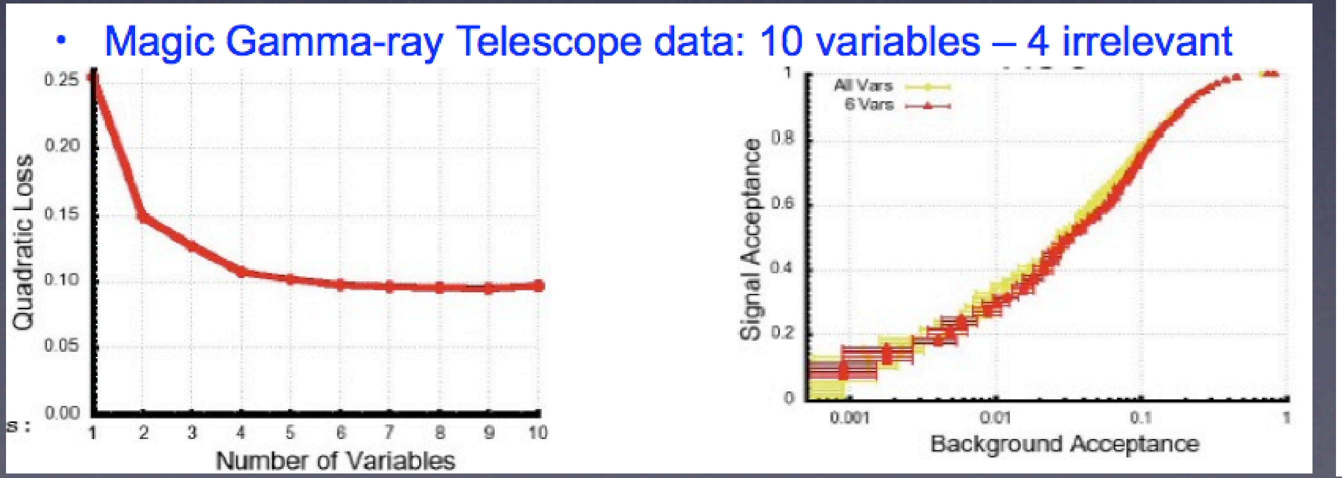
- Selection of the most powerful discriminating features (variables)
- Usually not all the features are useful for the classification problem. In modern applications, where the number of instances (events) can be huge, it is important to evaluate whether it is possible to find irrelevant features.
- FS addresses the problem of reducing the feature set to the smallest subset that gives the same or better quality of separation between signal and background as the full set does.



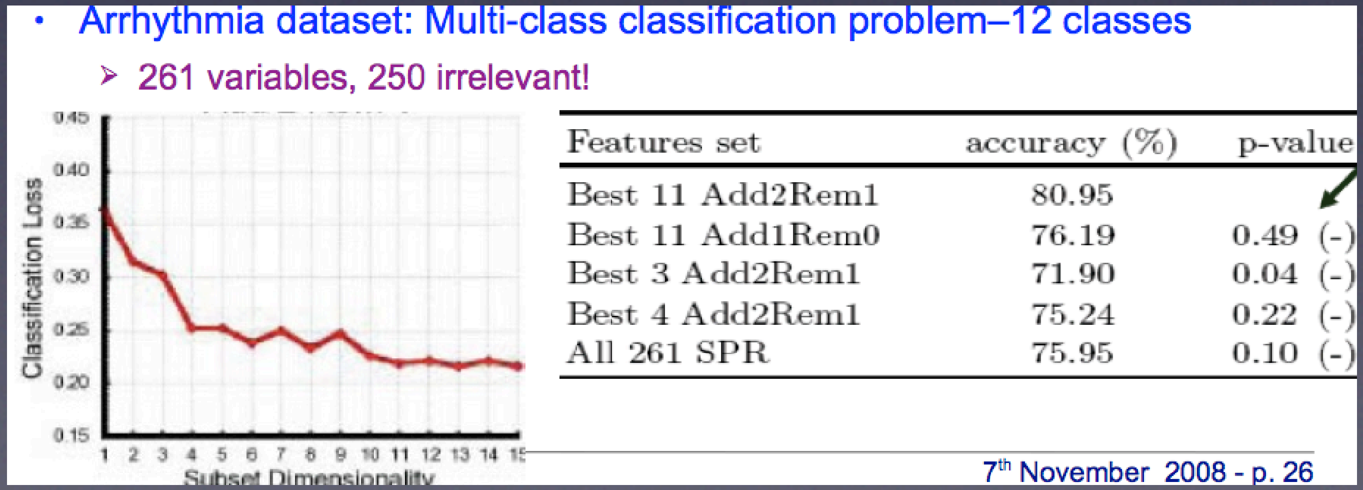
GOAL OF THE ANALYSIS AND DATA

- Comparing FS methods implemented in SPR to each other and with other methods implemented in statistical packages R and Weka.
- Comparing our results with previously published results for the same datasets.
- HEP datasets usually have many events with few input variables, but typically these datasets are not public.
- We use the datasets Magic Telescope, Cardiac Arrhythmia, WDBC, WBC, Colic Horse which are (with the exception of Magic Telescope) much smaller than typical HEP dataset. But they are all available publicly at:

  www.ics.uci.edu/~mlearn/MLRepository.html



- Magic Gamma-ray Telescope data: 10 variables – 4 irrelevant

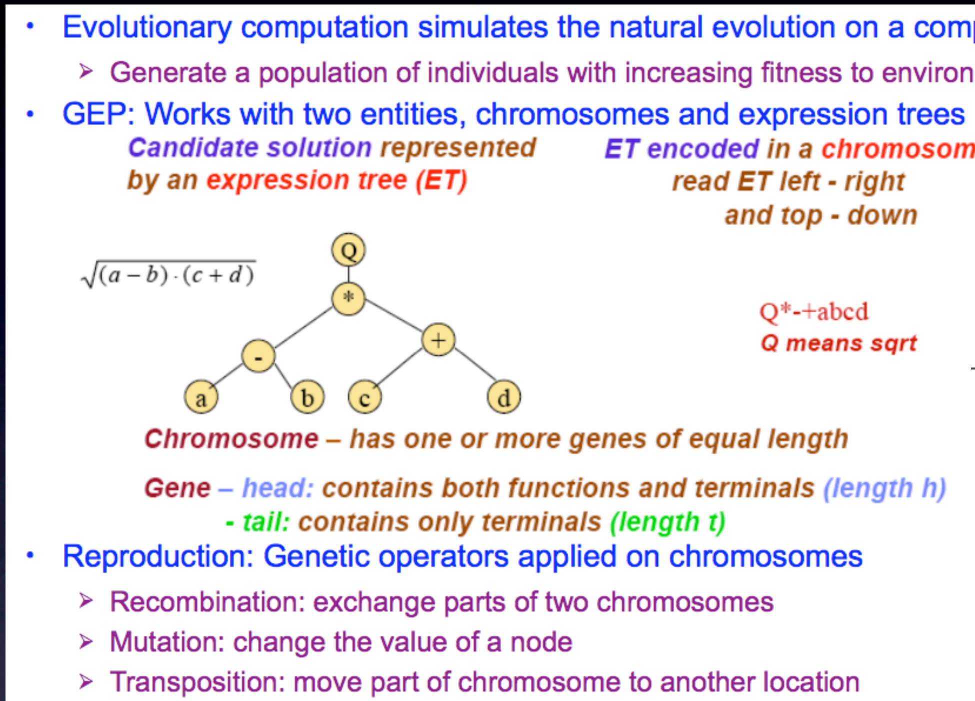


- Arrhythmia dataset: Multi-class classification problem–12 classes
  - 261 variables, 250 irrelevant!

Features set	accuracy (%)	p-value
Best 11 Add2Rem1	80.95	
Best 11 Add1Rem0	76.19	0.49 (-)
Best 3 Add2Rem1	71.90	0.04 (-)
Best 4 Add2Rem1	75.24	0.22 (-)
All 261 SPR	75.95	0.10 (-)

7th November 2008 - p. 26

Gene Expression Programming
(L. Teodorescu)



- Evolutionary computation simulates the natural evolution on a computer
  - ➢ Generate a population of individuals with increasing fitness to environment
- GEP: Works with two entities, chromosomes and expression trees

Candidate solution represented by an expression tree (ET)

ET encoded in a chromosome: read ET left - right and top - down

$\sqrt{(a-b)\cdot(c+d)}$

Q*-+abcd

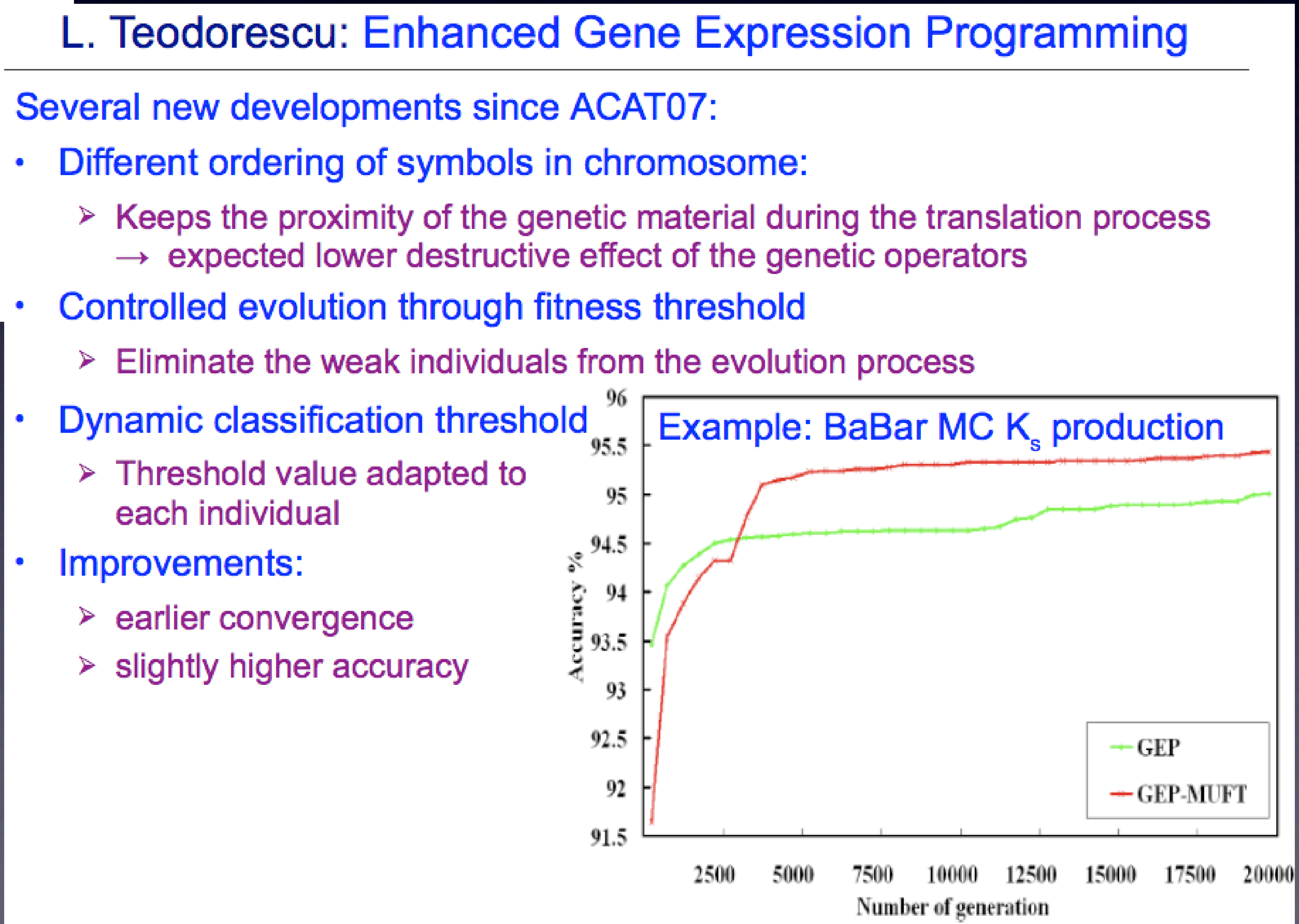
Q means sqrt

***Chromosome** – has one or more genes of equal length*

***Gene** – head: contains both functions and terminals (length h)*

- tail: contains only terminals (length t)

- Reproduction: Genetic operators applied on chromosomes
  - ➢ Recombination: exchange parts of two chromosomes
  - ➢ Mutation: change the value of a node
  - ➢ Transposition: move part of chromosome to another location



L. Teodorescu: Enhanced Gene Expression Programming

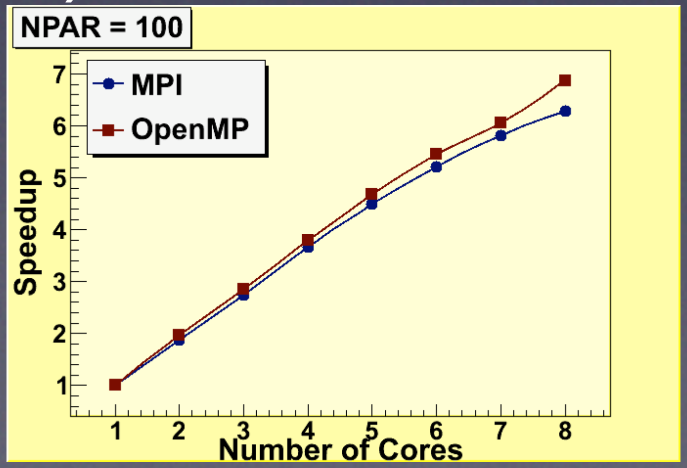
Several new developments since ACAT07:

- Different ordering of symbols in chromosome:
  - ➢ Keeps the proximity of the genetic material during the translation process → expected lower destructive effect of the genetic operators
- Controlled evolution through fitness threshold
  - ➢ Eliminate the weak individuals from the evolution process
- Dynamic classification threshold
  - ➢ Threshold value adapted to each individual
- Improvements:
  - ➢ earlier convergence
  - ➢ slightly higher accuracy

Example: BaBar MC K_s production

Data Analysis

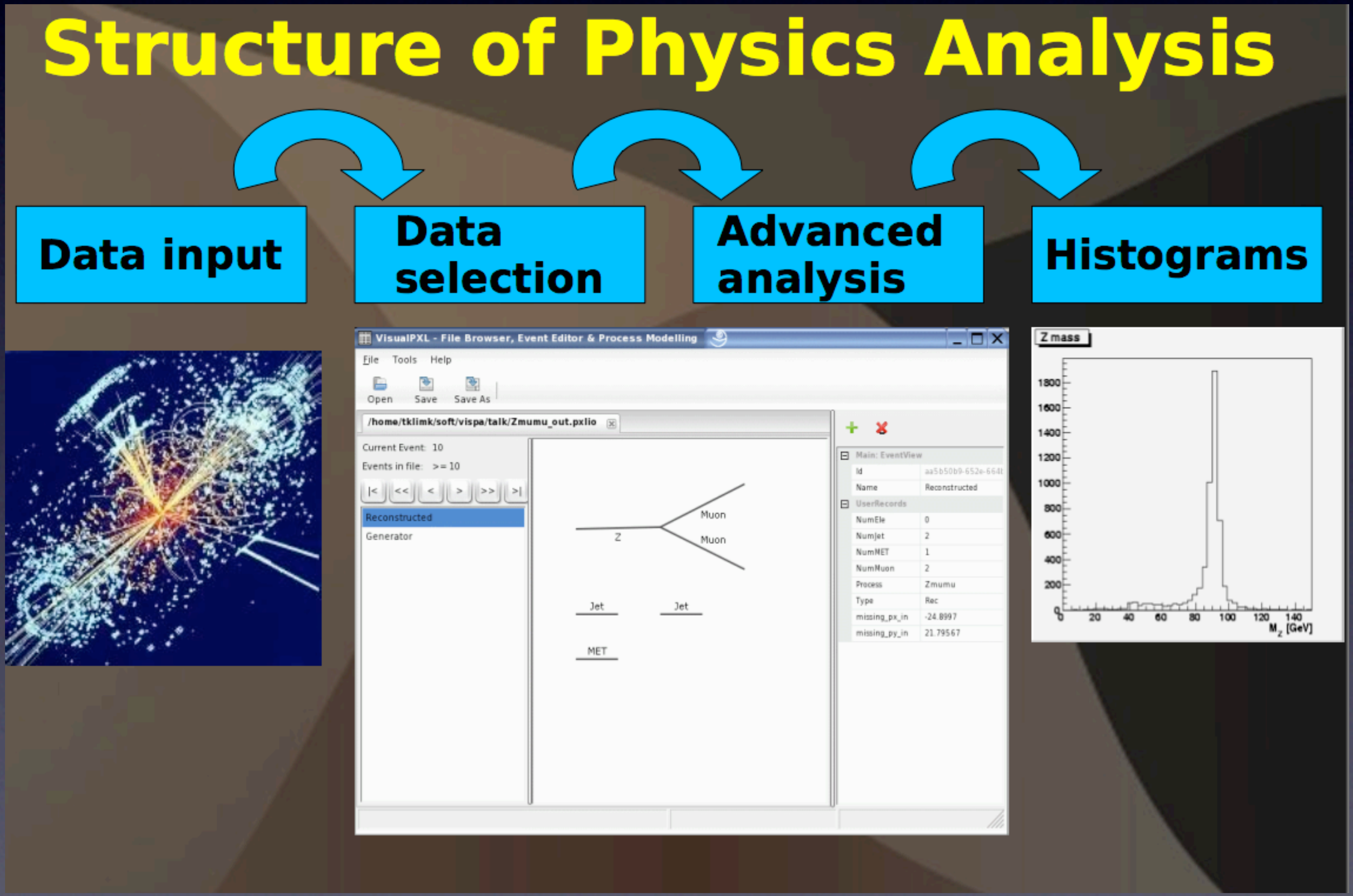
- Improvements in ROOT fitting and minimization (L.M.)
- Goodness of fit test for weighted histograms (poster by N. Gagunashvili)
- Parallelization of fitting and minimization (A. Lazzaro)
  - split likelihood calculation (in RooFit) and/or derivative calculation in Minuit2
  - used MPI and/or openMP



Visual Physics Analysis

(Tatsiana KLIMKOVICH)

- VISPA: a novel concept for visual physics analysis



Structure of Physics Analysis

Data input → Data selection → Advanced analysis → Histograms

/home/tklimk/soft/vispa/talk/Zmumu_out.pxlio

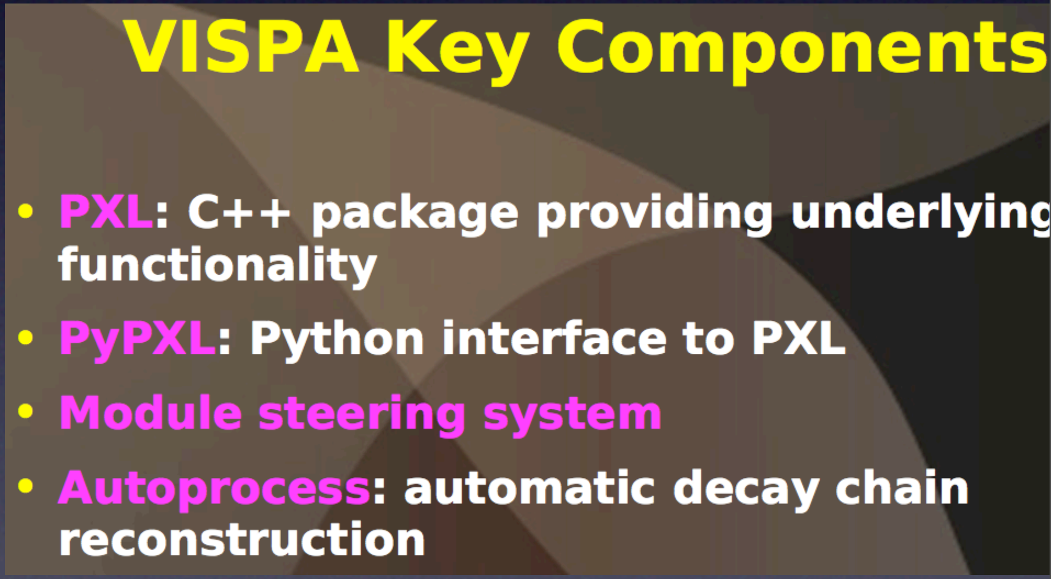
Current Event: 10

Events in file: >= 10

Reconstructed

Generator

Main: EventView	
Id	aa5b50b9-652e-664
Name	Reconstructed
UserRecords	
NumEle	0
NumJet	2
NumMET	1
NumMuon	2
Process	Zmumu
Type	Rec
missing_px_in	-24.8997
missing_py_in	21.79567



VISPA Key Components

- **PXL: C++ package providing underlying functionality**
- **PyPXL: Python interface to PXL**
- **Module steering system**
- **Autoprocess: automatic decay chain reconstruction**

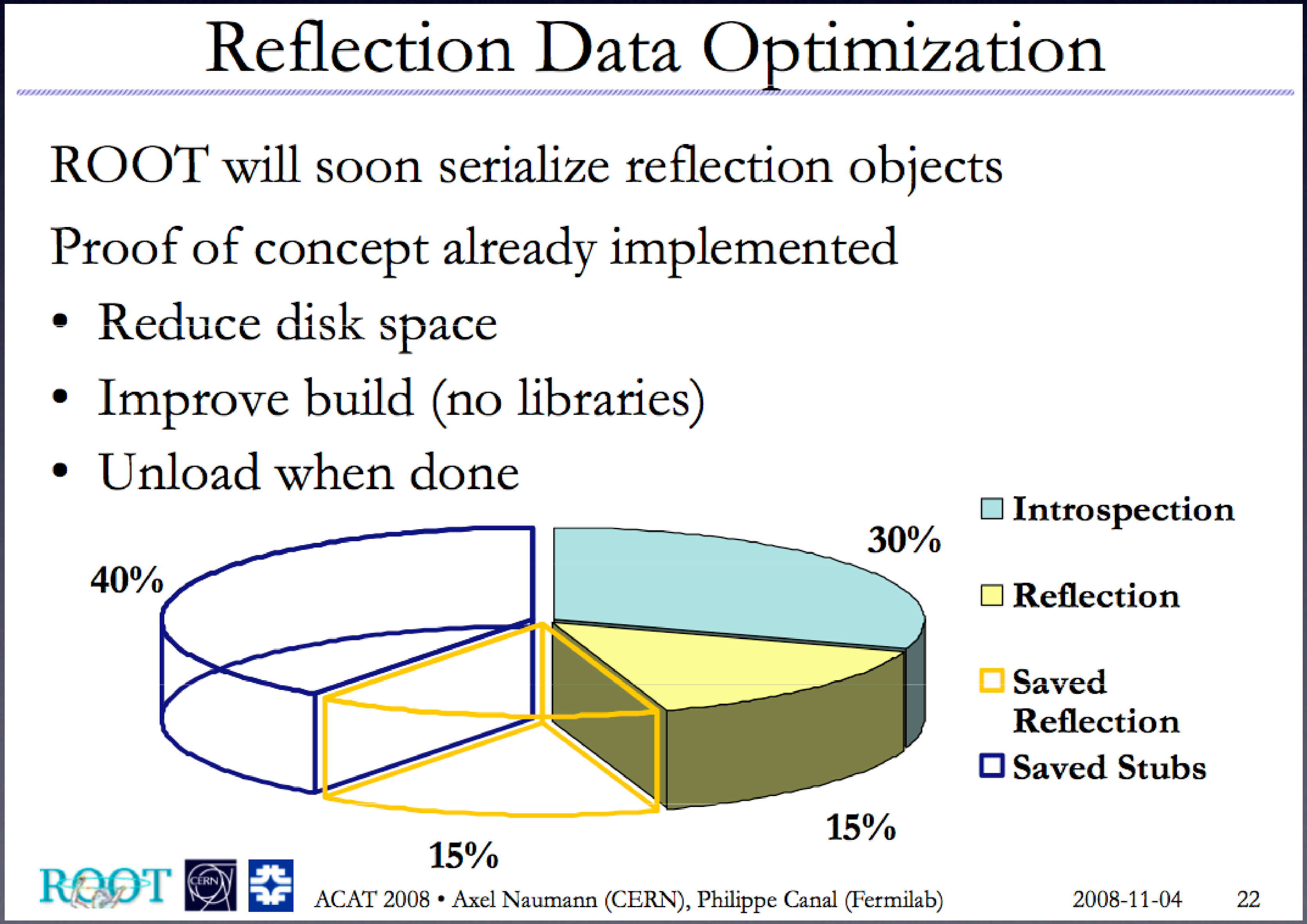
- use GUI to design analysis
- can export to Python (create Python analysis modules)

Languages, Interpreters, etc..

- T. Johnson: Java based software for High-Energy and Astro-physics
- C. Lattner: Introduction to the LLVM Compiler System
- Axel N.: The role of interpreters in HPC
- Axel N.: C++ and Data

C++ and Data
(Axel Naumann)

- Overview of C++ reflection and dictionary
  - Dictionary size and coming optimizations



Reflection Data Optimization

ROOT will soon serialize reflection objects

Proof of concept already implemented

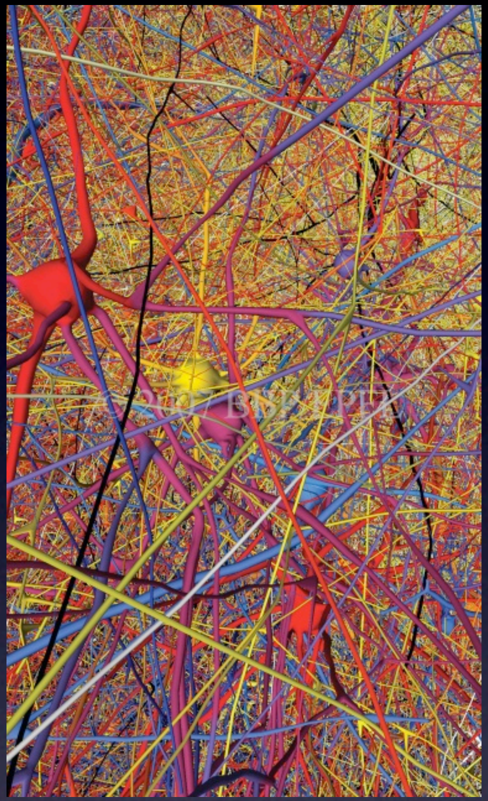
- Reduce disk space
- Improve build (no libraries)
- Unload when done

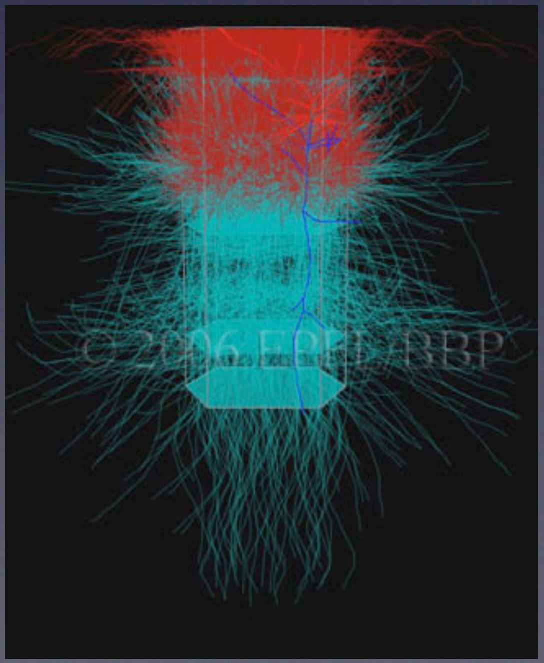
ACAT 2008 • Axel Naumann (CERN), Philippe Canal (Fermilab) 2008-11-04 22

Session 3
(Computation in Theoretical Physics)

- Geant4 talks (V.N.Ivanchenko)
    - Recent Progress of Geant4 Electromagnetic Physics and Readiness for the LHC Start
    - Hadronic Physics in Geant4: Improvements and Status for LHC Start
- MC generators in CMS (P. Bartalini)
- LCG Generators (M. Kirsanov)
- Statistics (S. Bityukov)
    - Two approaches to combine significances

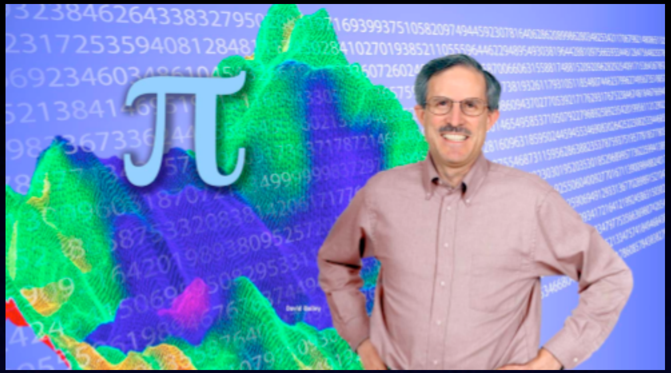
Blue Brain Project
(F. Schürmann)





- EPFL project in collaboration with IBM
- Reverse-engineer the mammal brain
    - reconstruct and simulate neurons based on large amount of experimental data
        - model connections and build neural circuits
        - simulate microcircuit (brain activity)
            - massive parallel computation
- use supercomputer (IBM Blue-gene)

High-Precision Arithmetic and Mathematical Physics (D. Bailey, LBL)

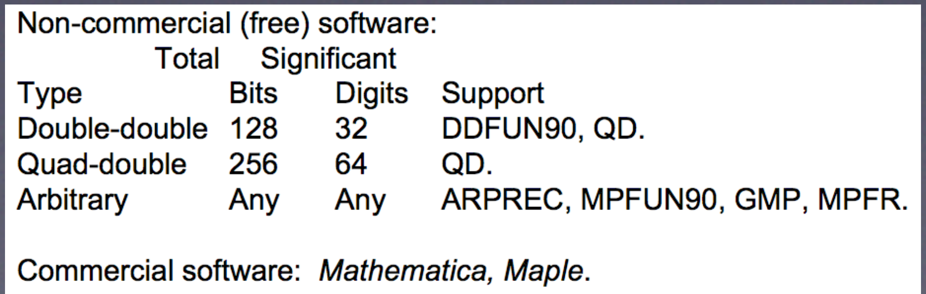


$$\pi = \sum_{n=0}^{\infty} \frac{1}{16^k} \left(\frac{4}{8k+1} - \frac{2}{8k+4} - \frac{1}{8k+5} - \frac{1}{8k+6} \right)$$

see http://en.wikipedia.org/wiki/Bailey-Borwein-Plouffe_formula

- Examples of applications requiring more extended precision (double-double or quad-double)
  - supernovae simulations
  - climate modeling
  - planetary orbit calculations
  - Coulomb N-body atomic system simulations

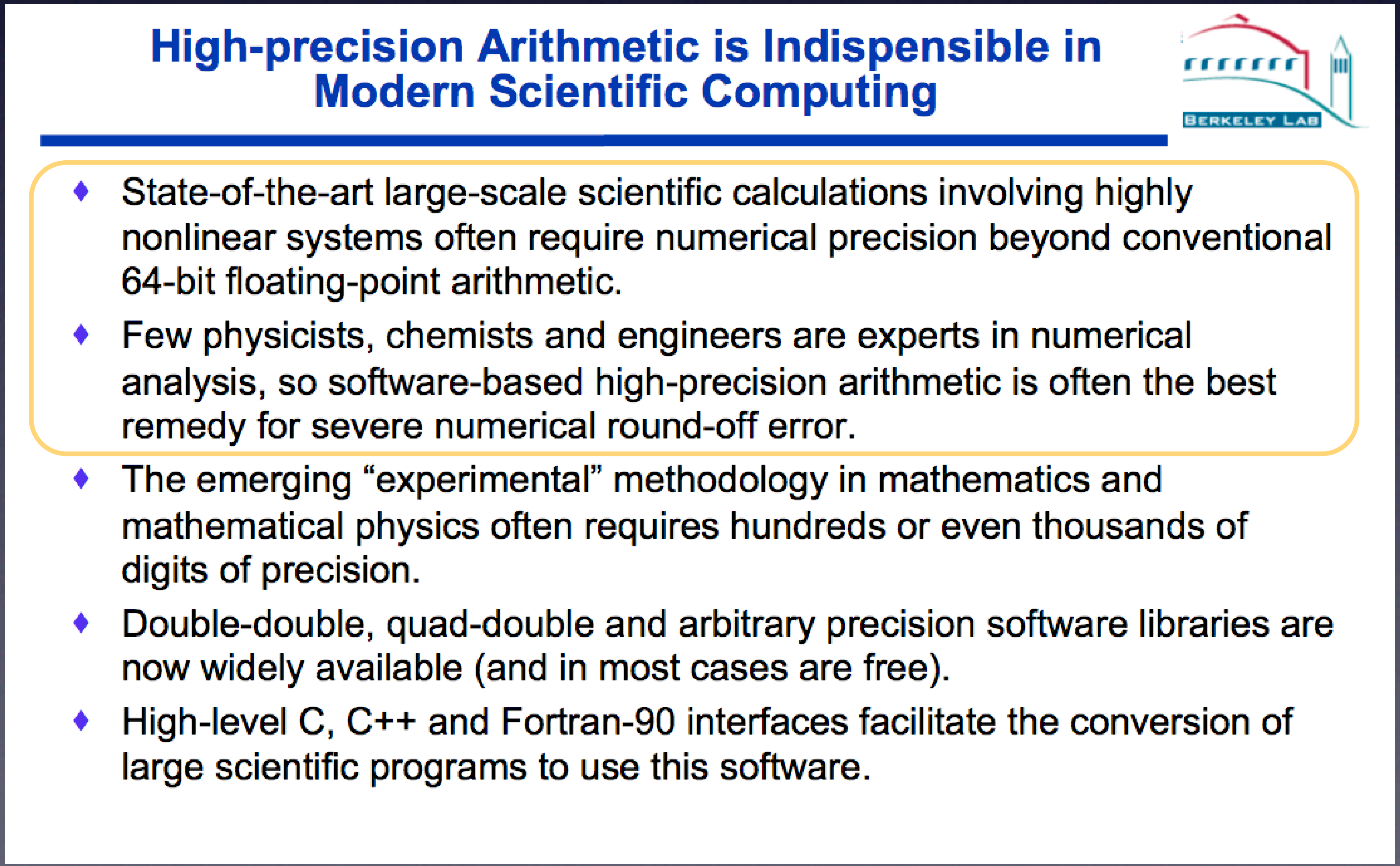
High precision software



Non-commercial (free) software:

Type	Total Bits	Significant Digits	Support
Double-double	128	32	DDFUN90, QD.
Quad-double	256	64	QD.
Arbitrary	Any	Any	ARPREC, MPFUN90, GMP, MPFR.

Commercial software: *Mathematica*, *Maple*.



High-precision Arithmetic is Indispensible in Modern Scientific Computing

- ♦ State-of-the-art large-scale scientific calculations involving highly nonlinear systems often require numerical precision beyond conventional 64-bit floating-point arithmetic.
- ♦ Few physicists, chemists and engineers are experts in numerical analysis, so software-based high-precision arithmetic is often the best remedy for severe numerical round-off error.
- ♦ The emerging “experimental” methodology in mathematics and mathematical physics often requires hundreds or even thousands of digits of precision.
- ♦ Double-double, quad-double and arbitrary precision software libraries are now widely available (and in most cases are free).
- ♦ High-level C, C++ and Fortran-90 interfaces facilitate the conversion of large scientific programs to use this software.