The NeuroBayes Neural Network package

M.Feindt, U.Kerzel
Phi-T, University of Karlsruhe

ACAT 05

bmb+f - Förderschwerpunkt
Elementarteilchenphysik
Großgeräte der physikalischen Grundlagenforschung

Outline

- Bayesian statistics
- Neural networks
- The NeuroBayes neural network package
  - The NeuroBayes principle
  - Preprocessing of input variables
  - Predicting complete probability density distributions
- Examples from high energy physics and industry

Bayes' Theorem (1)

Conditional Probabilities:

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Because of $P(A \cap B) = P(B \cap A)$ it follows that

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Bayes´ Theorem

Bayes' Theorem (2)

<phi-t>

Extremely important due to the interpretation A=theory B=data

Likelihood

Prior

$$P(theory|data) = \frac{P(data|theory)P(theory)}{P(data)}$$

Posterior

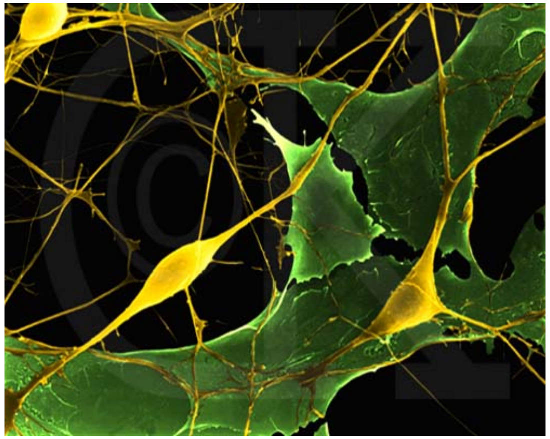
Evidence

Neural Networks (1)

- <u>Inspired by nature:</u>

  Neuron in brain “fires” if stimuli received from other neurons exceed threshold.

  (very simple model. . .)

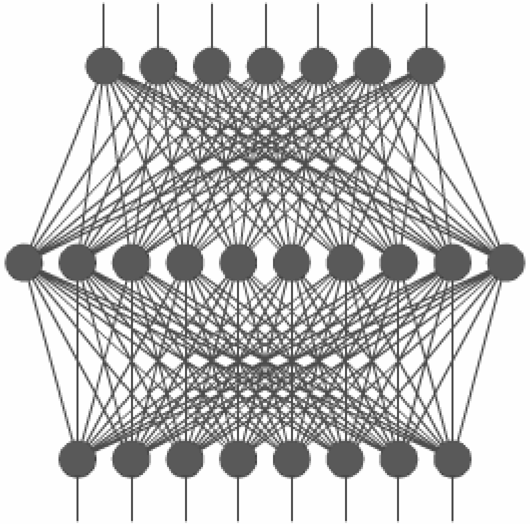


- <u>Construct Neural Network</u>

  Output of node *j* in layer *n* is given by weighted sum of output of all nodes in layer *n-1*:

$$x_j^n = g\left(\sum_k w_{jk}^n \cdot x_k^{n-1} + \mu_j^n\right)$$

$g(t)$ sigmoid function μ_j^n threshold (“bias-node”)



$\rightarrow$ information is stored in connections

Neural Networks (2)

- <u>Network training:</u>

  Minimisation of a *loss function* by iteratively adjusting the weights w^{n}_{jk} such that the deviation of the actual network output from the desired output is minimised

- <u>Loss functions:</u>
  - sum of quadratic deviations
  - entropy (max. likelihood)

Neural Networks (3)

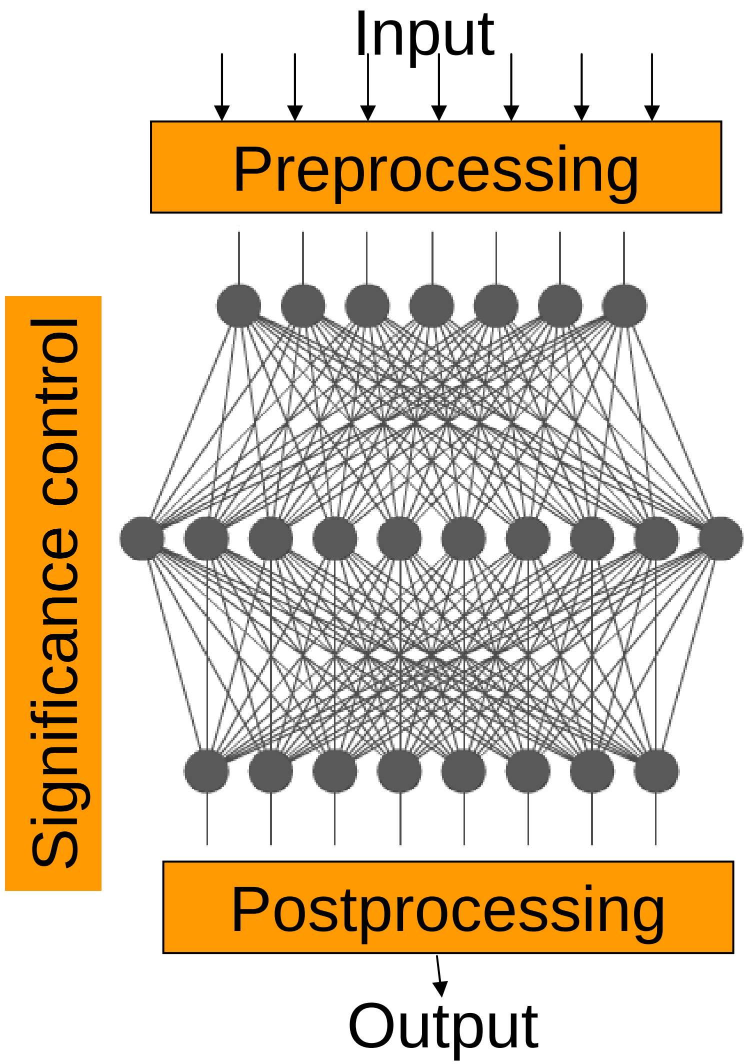
Neural Networks ...

- learn correlations between variables
- learn higher order (non-linear) correlations to training target
- do *not* require that all information is available for each input vector

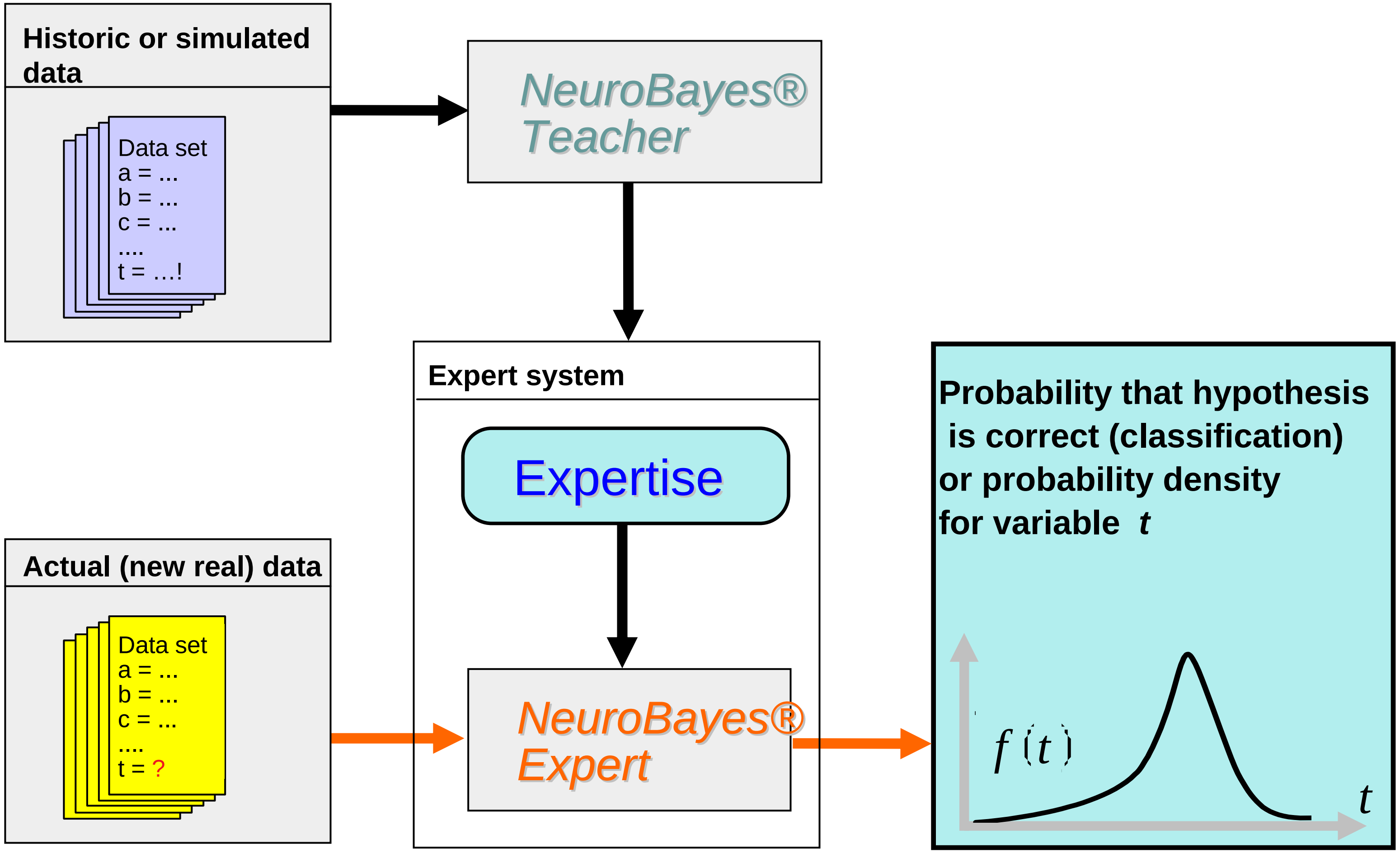
NeuroBayes principle

NeuroBayes® Teacher:
Learning of complex relationships from existing databases

NeuroBayes® Expert:
Prognosis for unknown data



How it works ...



Historic or simulated data
Data set
a = ...
b = ...
c = ...
....
t = …!
NeuroBayes®
Teacher
Expert system
Expertise
NeuroBayes®
Expert
Actual (new real) data
Data set
a = ...
b = ...
c = ...
....
t = ?
Probability that hypothesis is correct (classification) or probability density for variable t
f(t)
t

Preprocessing I

<phi-t>
Physics Information Technologies

Why preprocess input variables?

Shouldn't the network learn it all??

Yes, but ...

- Optimisation in many dimensions difficult
- Example (2D): deepest valley in Swiss Alps
  - isn't the next valley deeper?
    $\rightarrow$ difficult to find out once you're down there...
  - now try to find the minimum in $\mathcal{O}(1000)$ dimensions....

- Preprocessing: "Guide" network to best minimum

Preprocessing II

Global preprocessing:

- normalisation and decorrelation

  $\rightarrow$ new covariance matrix is unit matrix
- rotate such that first variable contains all linear information about mean, second about width, ...
- automatically recognise binary and discrete variables
- direct connection between input and output layer

  $\rightarrow$ networks learns deviations from best linear estimate
- only keep variables with stat. relevance $> 0.5n \cdot \sigma$

$\rightarrow$ completely automatic and robust !

Preprocessing III

<u>individual variable preprocessing:</u>

- variables with default value or δ function
- regularised 1d correlation to training target via spline-fit (monotonous or general continuous variable)
- ordered or unordered classes with Bayesian regularisation
- decorrelation of influence of other variables on the correlation to training target
- ...

Control capacity

Bayesian regularisation:

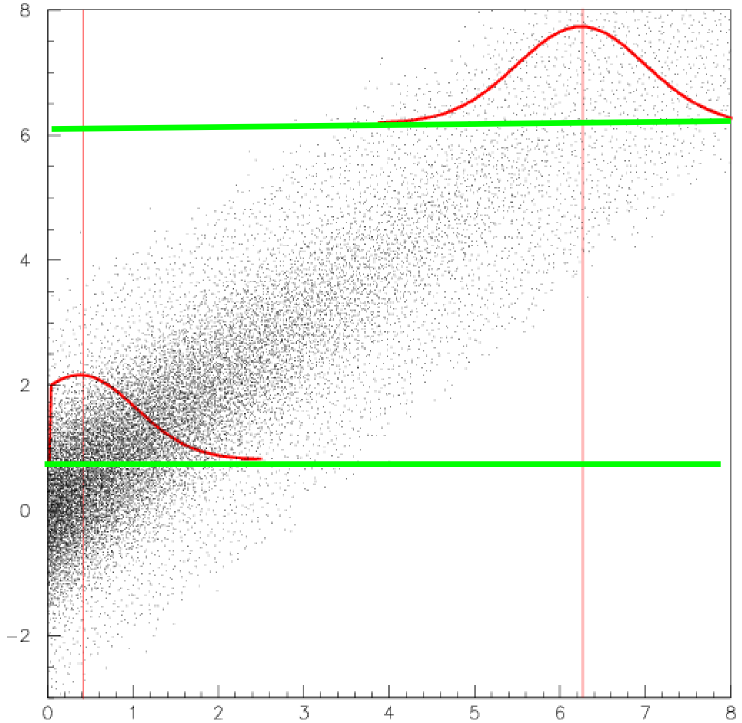
$\longrightarrow$ avoid overtraining, enhance generalisation ability

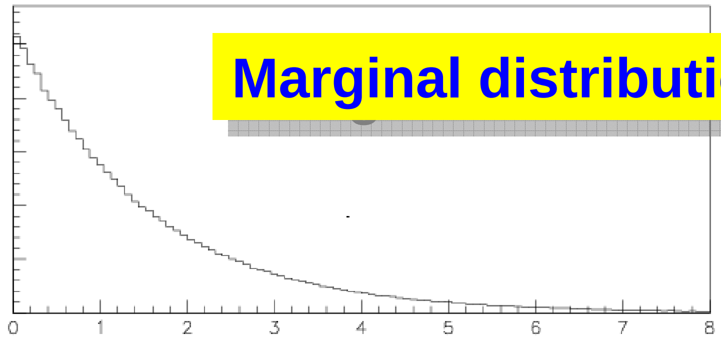
- favour small networks with small weights (“formal stabilisation”)
- separate regularisation constants for at least 3 groups of weights
- Automatic Relevance Determination of input variables
- Automatic Shape Regularisation of output nodes (shape reconstr.)
- during training:

  remove not significant weights / network nodes

$\rightarrow$ only statistically significant connections remain

Bayesian approach I





Marginal distribution f(t)

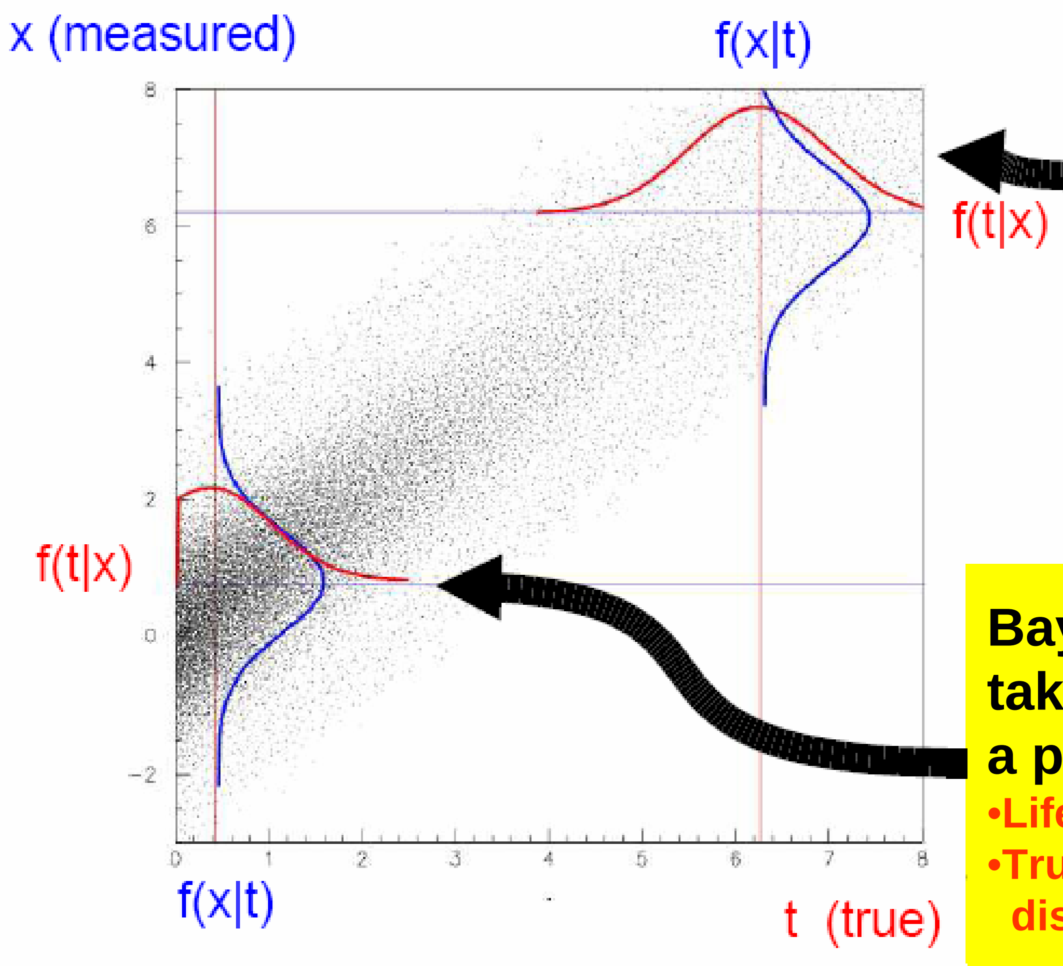
Conditional probability densities f(t|x)

Conditional probability density for a special case x (Bayesian Posterior)

Conditional probability densities f(t|x) are functions of x, but also depend on marginal distribution f(t).

Inclusive distribution (Bayesian Prior)

Bayesian approach II



Classical ansatz:
f(x|t)=f(t|x)
approximately correct at good resolution far away from physical boundaries

Bayesian ansatz:
takes into account
a priori- knowledge f(t):
- **Lifetime never negative**
- **True lifetime exponentially distributed**

NeuroBayes tasks

- **Classification:** element is part of class *A* or *B*

  particle is electron, B meson, ... or background

- **Shape reconstruction:**

  Bayesian estimator $f(t|\vec{x})$ for a single multidimensional measurement $\vec{x}$

Note:
Conditional probability density contains much more information than just the mean value, which is determined in a regression analysis.
It also tells us something about the uncertainty and the form of the distribution, in particular non-Gaussian tails.

Example: CDF

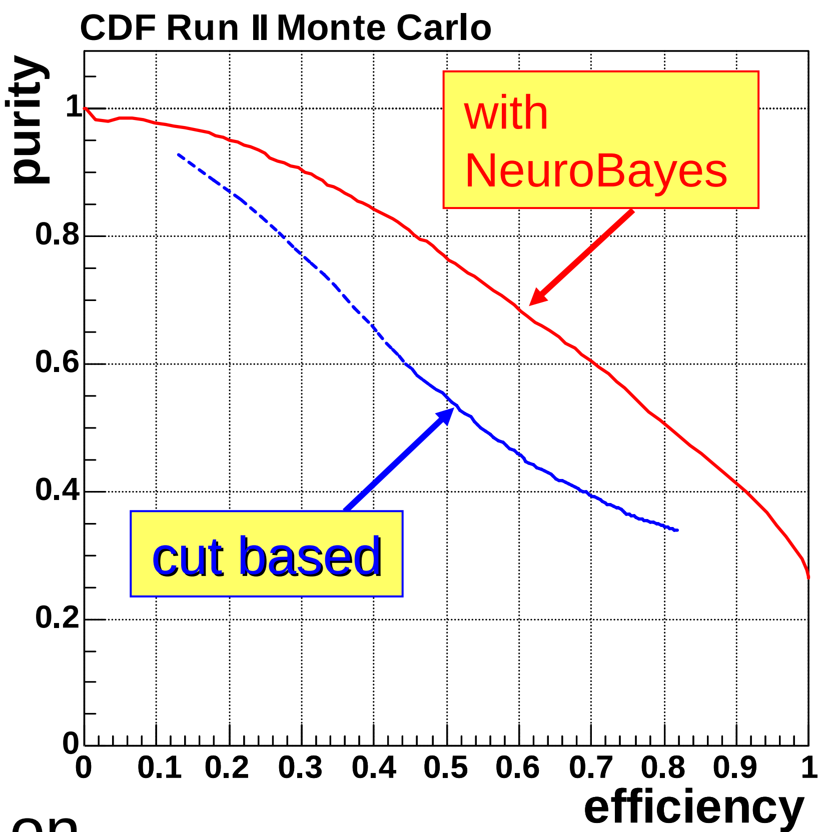
CDF Run 2:

Identify jets containing decay products of B mesons

combine correlated variables:

- jet mass
- sum of longitudinal/transverse momentum
- track originates from B decay
- ...

$\rightarrow$ huge improvement w.r.t cut on displaced tracks !



$$\text{eff} = \frac{\#\text{signal past cut}}{\#\text{ true signal}}$$

$$\text{pur} = \frac{\#\text{ true signal past cut}}{\#\text{cand. past cut}}$$

Examples (cont.)

Further examples from our Karlsruhe group:

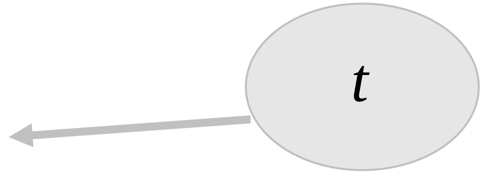
- Construct expert-system for B physics
  - B meson identification in a jet
  - particle ID (electrons, muons)
  - B meson flavour tagging (e.g. B_s mixing)
- Automated cut optimisation
- Hypotheses testing

  (e.g. determine correct assignment of quantum numbers J^{PC})
- ...

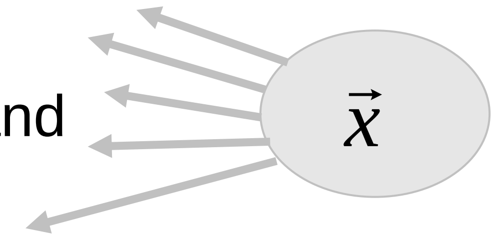
Shape reconstruction

in particle physics:

What is the probability density
of the true B energy in this event

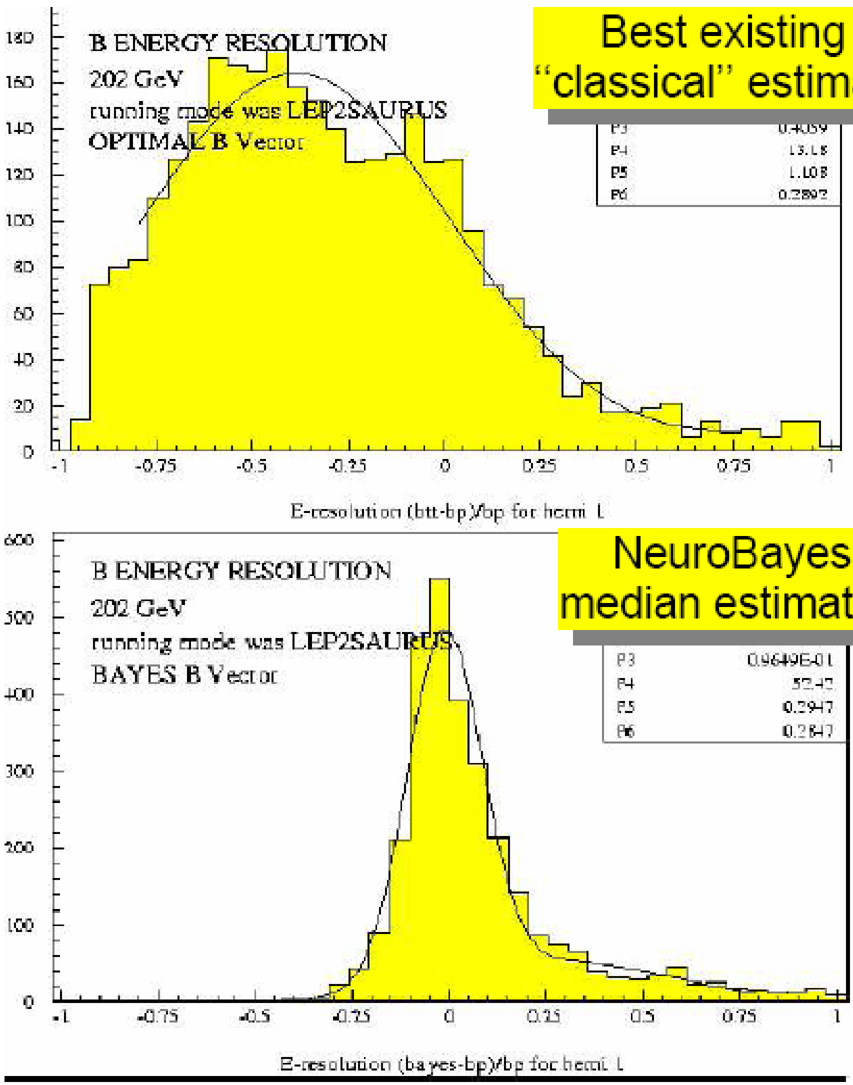


- taken with the DELPHI detector at LEP II
- at this beam energy,
- this effective c.m. energy
- these n tracks with those momenta and rapidities in the hemisphere,
- which are forming this secondary vertex with this decay length and probability,
- this number of not well reconstructed tracks, this neutral showers,
- etc pp



$$f(t \mid \vec{x})$$

Example: Delphi



Best existing "classical" estimator

NeuroBayes median estimator

B hadron energy measurement

Relative resolution of reconstructed B hadron energy in DELPHI at LEP II at 202 GeV energy

(completely inclusive)

core resolution 40% –> 10%

Technology transfer

<phi-t>
Physics Information Technologies

These methods are not only applicable in physics

<phi-t>: Foundation out of University of Karlsruhe, sponsored by exist-seed-programme of the federal ministry for Education and Research BMBF

Founding Phi-T

2000-2002 NeuroBayes®-specialisation
for economy at the University
of Karlsruhe
Oct. 2002: GmbH founded,
first industrial application
June 2003: Move into new office
199 qm IT-Portal Karlsruhe

Exclusive rights for NeuroBayes®
Juli 2004: Partnership with
2000-heads-company
msg Systems AG

Personell September 2004:
4 full time staff (all from HEP) and
a number of associated people,
Prof. consultance z.B. by Prof. Dr. Volker Blobel,
Economic/legal/marketing- expertise present

Applications in Economy

- Medicine and Pharma research
  e.g. effects and undesirable effects of drugs
  early tumor recognition
- Banks
  e.g. Credit-Scoring (Basel II), Finance time series prediction, valuation of derivates, risk minimised trading strategies, client valuation
- Insurances
  e.g. risk and cost prediction for individual clients, probability of contract cancellation, fraud recognition, justice in tariffs
- Trading chain stores: turnover prognosis

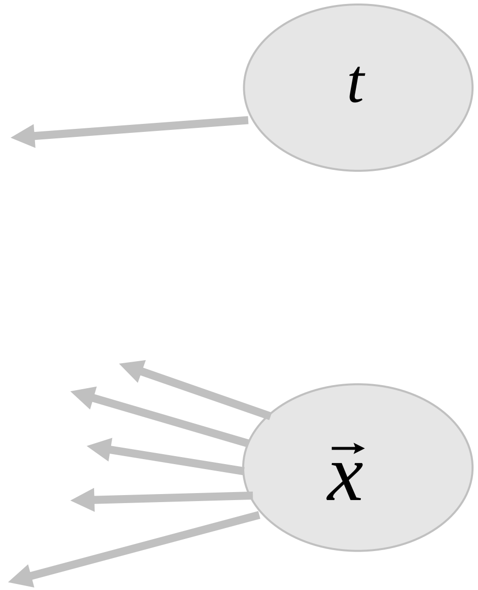
Necessary prerequisite:
Historic or simulated data must be available.

Shape reconstruction

in investment-banking:

What is the probability density for a price change of equity A in the next 10 days…

- that made this and that price movement in the last days and weeks…
- is so much more expensive than the n-days moving average…
- but is so much less expensive that the absolute maximum…
- has this correlation to the crude oil price…
- and the Dow Jones index…
- etc. pp.



$$f(t \mid \vec{x})$$

Conclusion

- NeuroBayes is a sophisticated neural network based on Bayesian statistics
  - automated and robust preprocessing
  - advanced regularisation techniques
  - can predict complete probability density distributions on *event-by-event* basis

- Successful application in high-energy physics and industry

BACKUP

Bayesian vs. classical statistics

Classical statistics is just a special case of Bayesian statistics:

Likelihood Prior

$$P(theory|data) = \frac{P(data|theory)P(theory)}{P(data)}$$

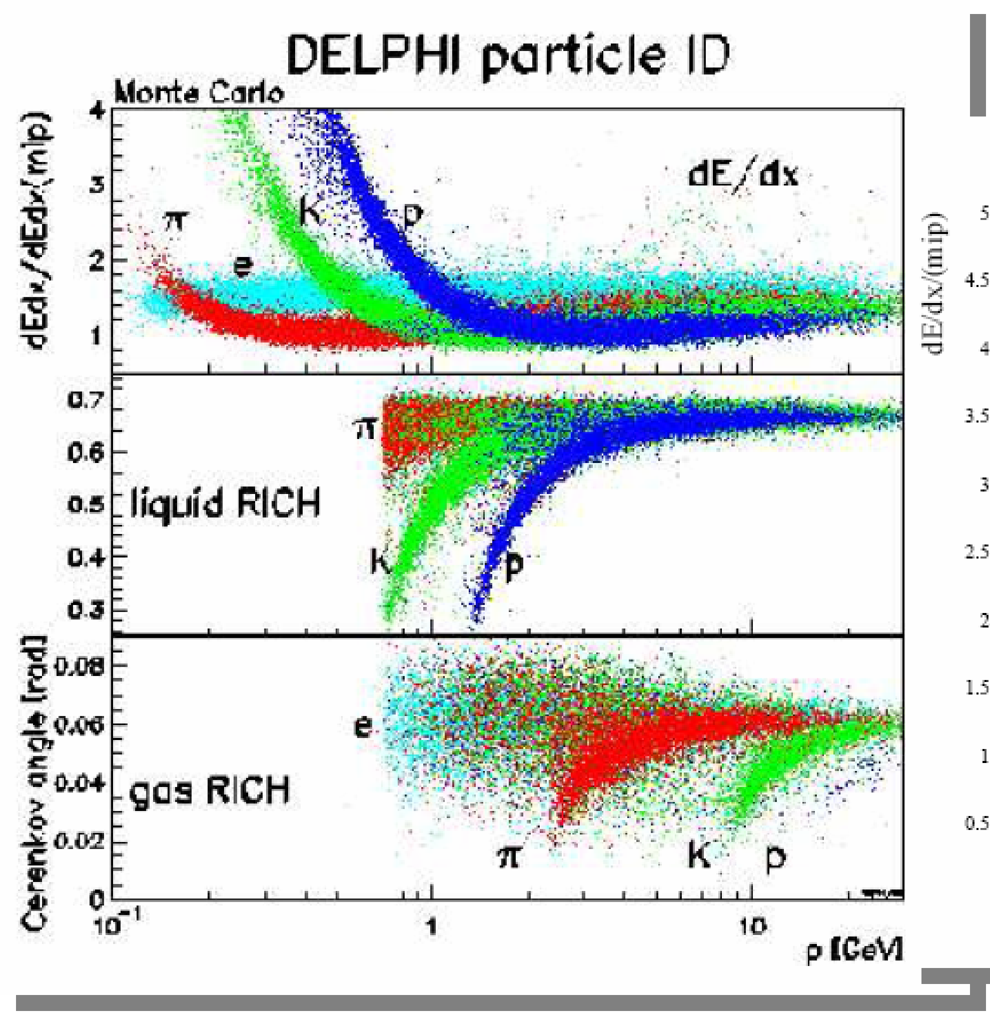
Posterior Evidence

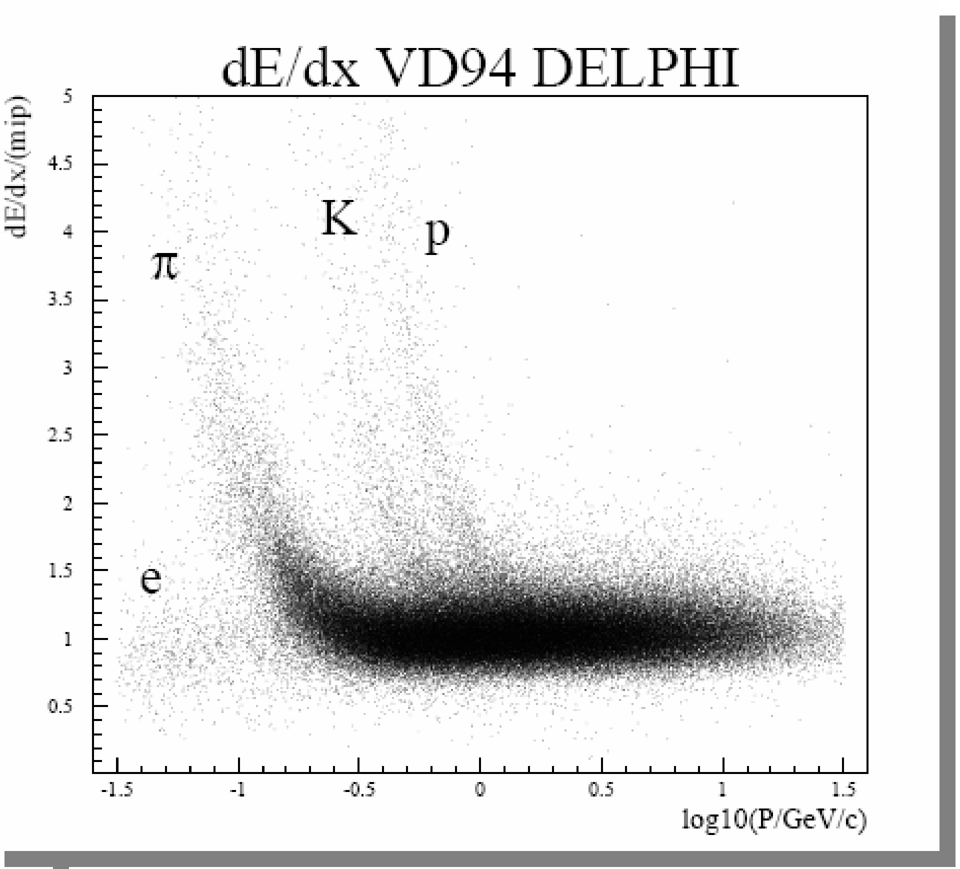
Maximising of likelihood instead of a posteriori probability means:

Implicit assumption that prior probability is flatly distributed, i.e. each value has same probability.

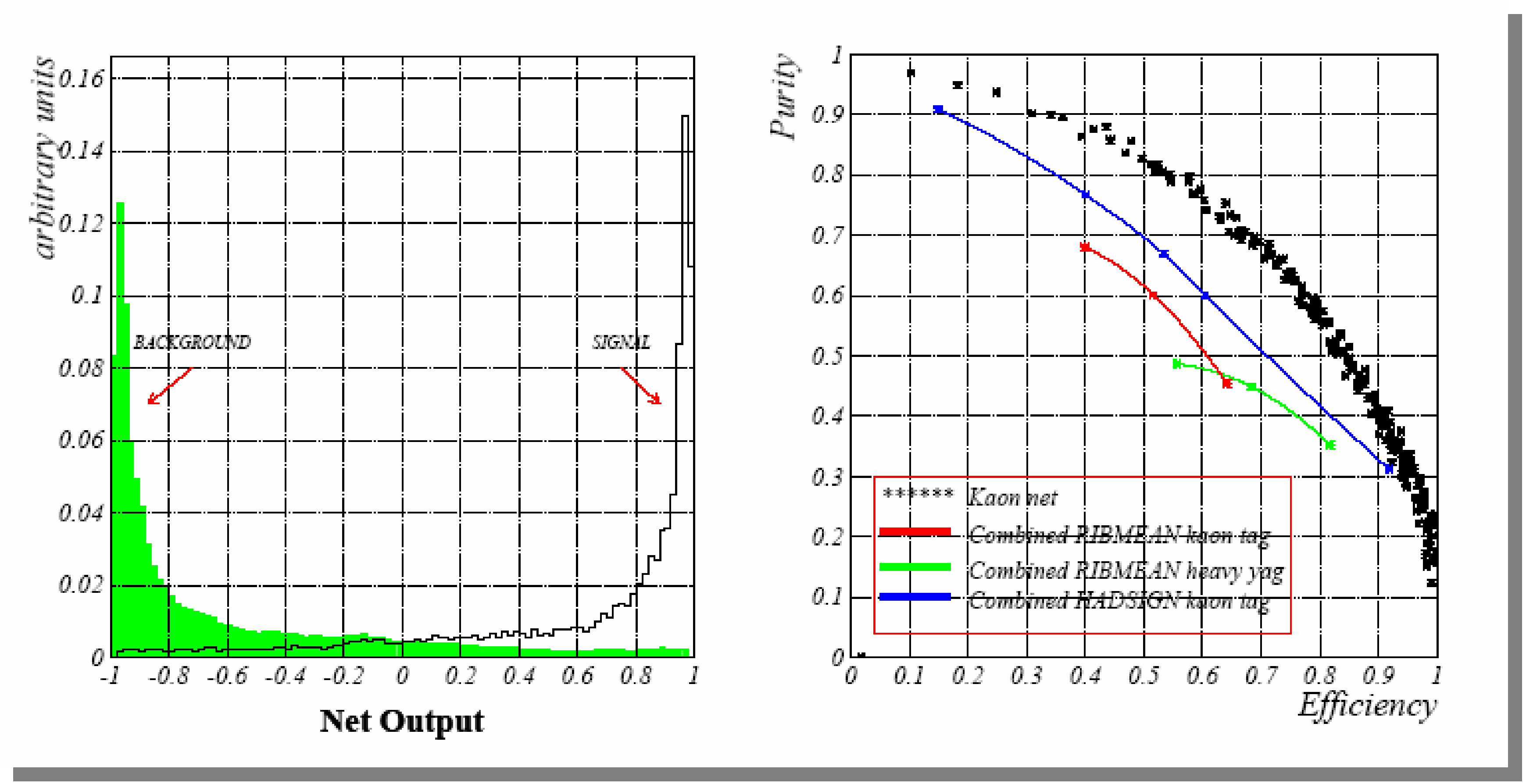
Sounds reasonable, but is in general wrong!
Does not mean that one does not know anything!

Examples: DELPHI Particle ID

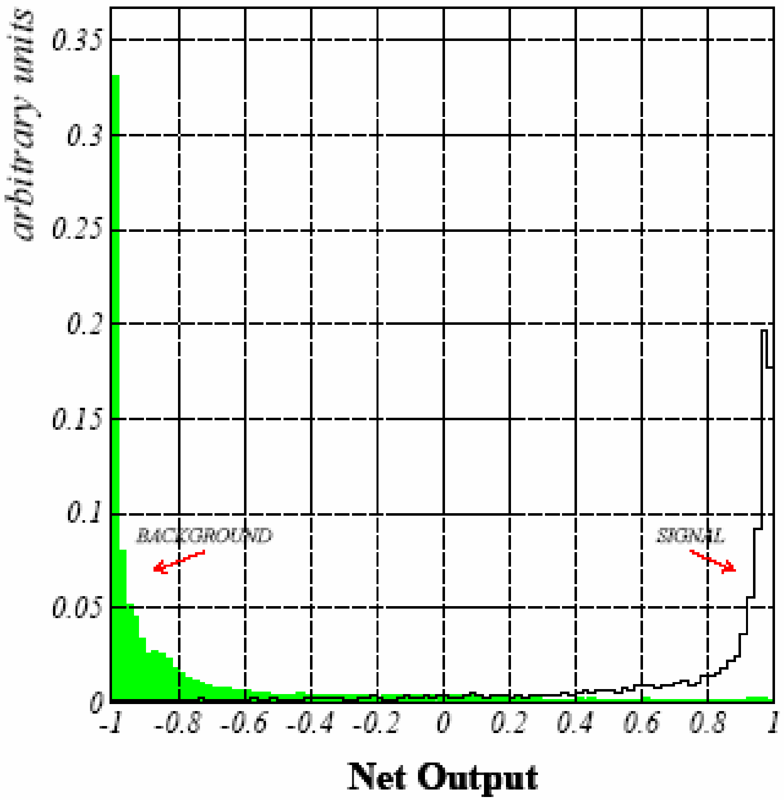


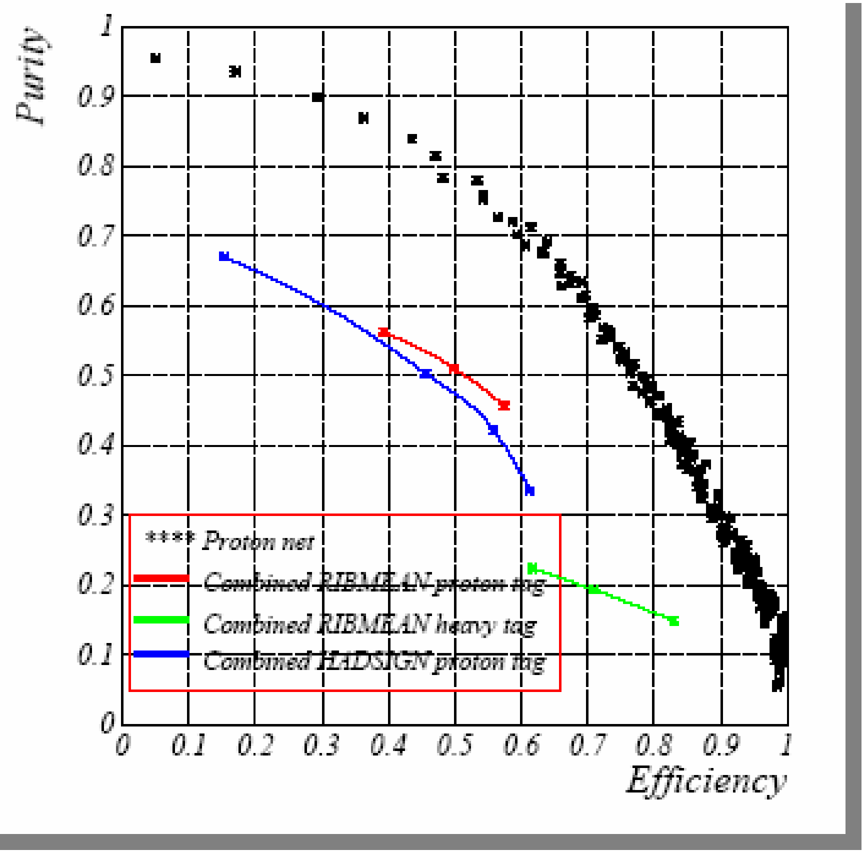


MACRIB Kaon ID



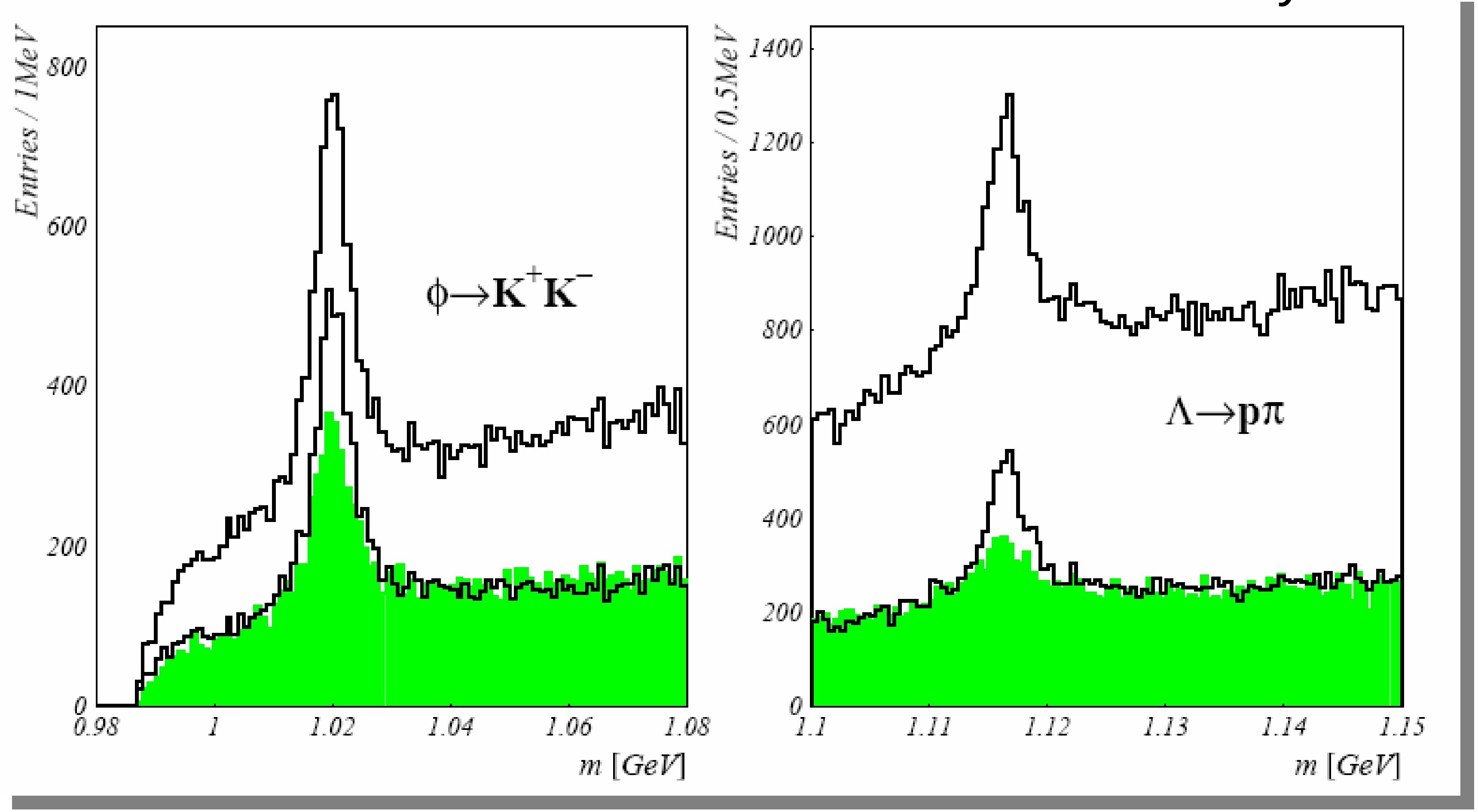
MACRIB Proton ID



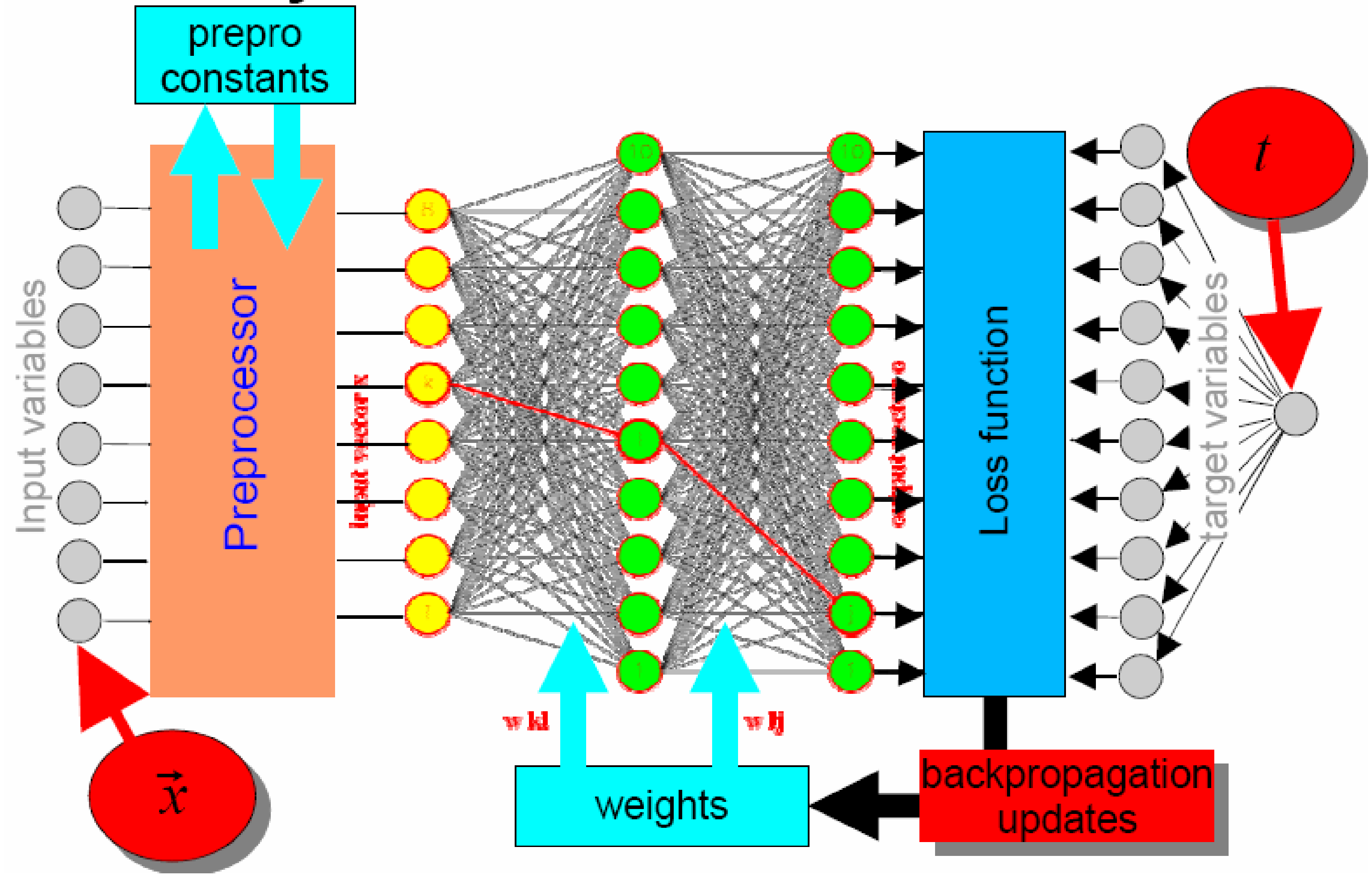


MACRIB Consequences for analyses

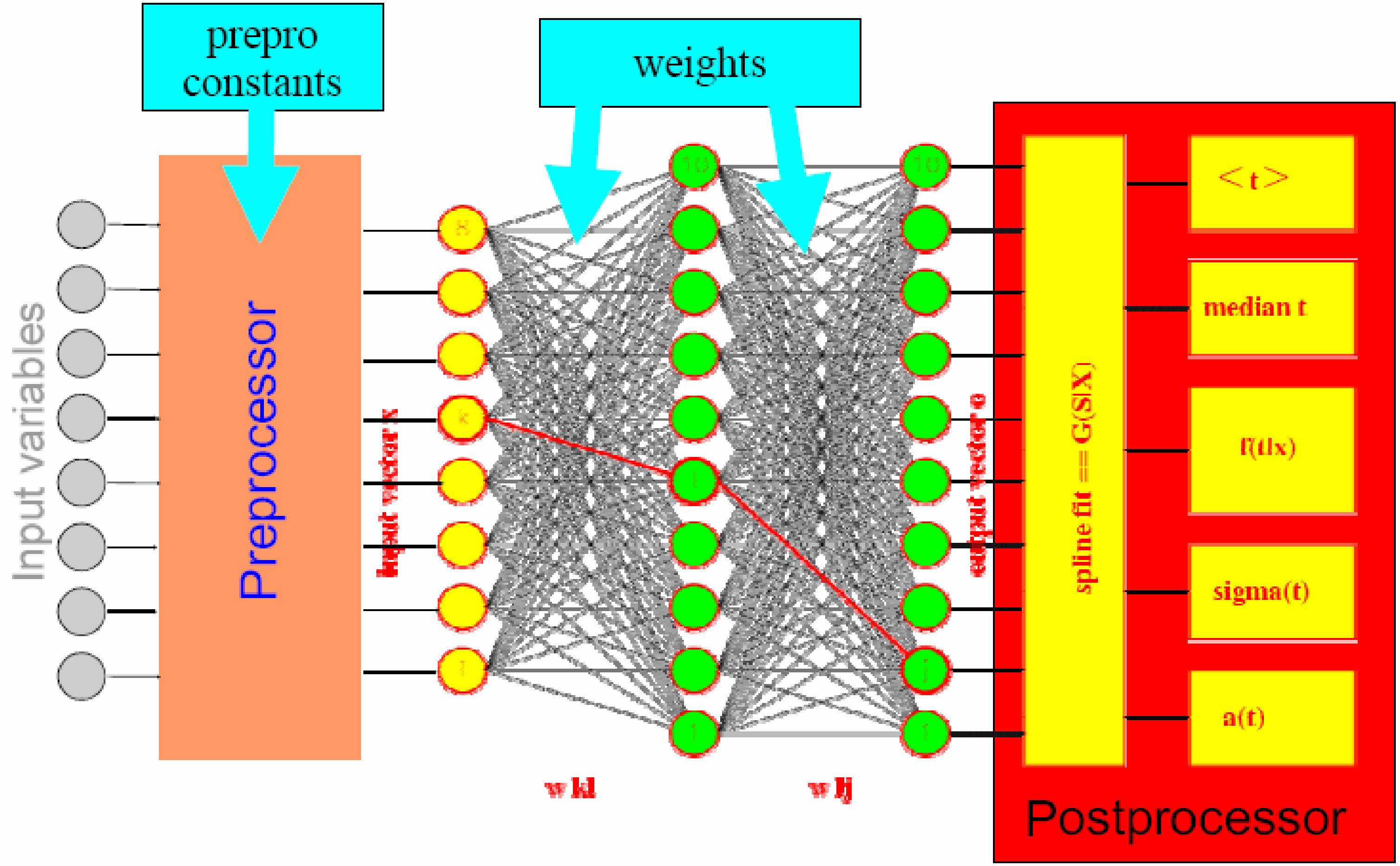
300% Phi-mesons | 300% Lambda-baryons



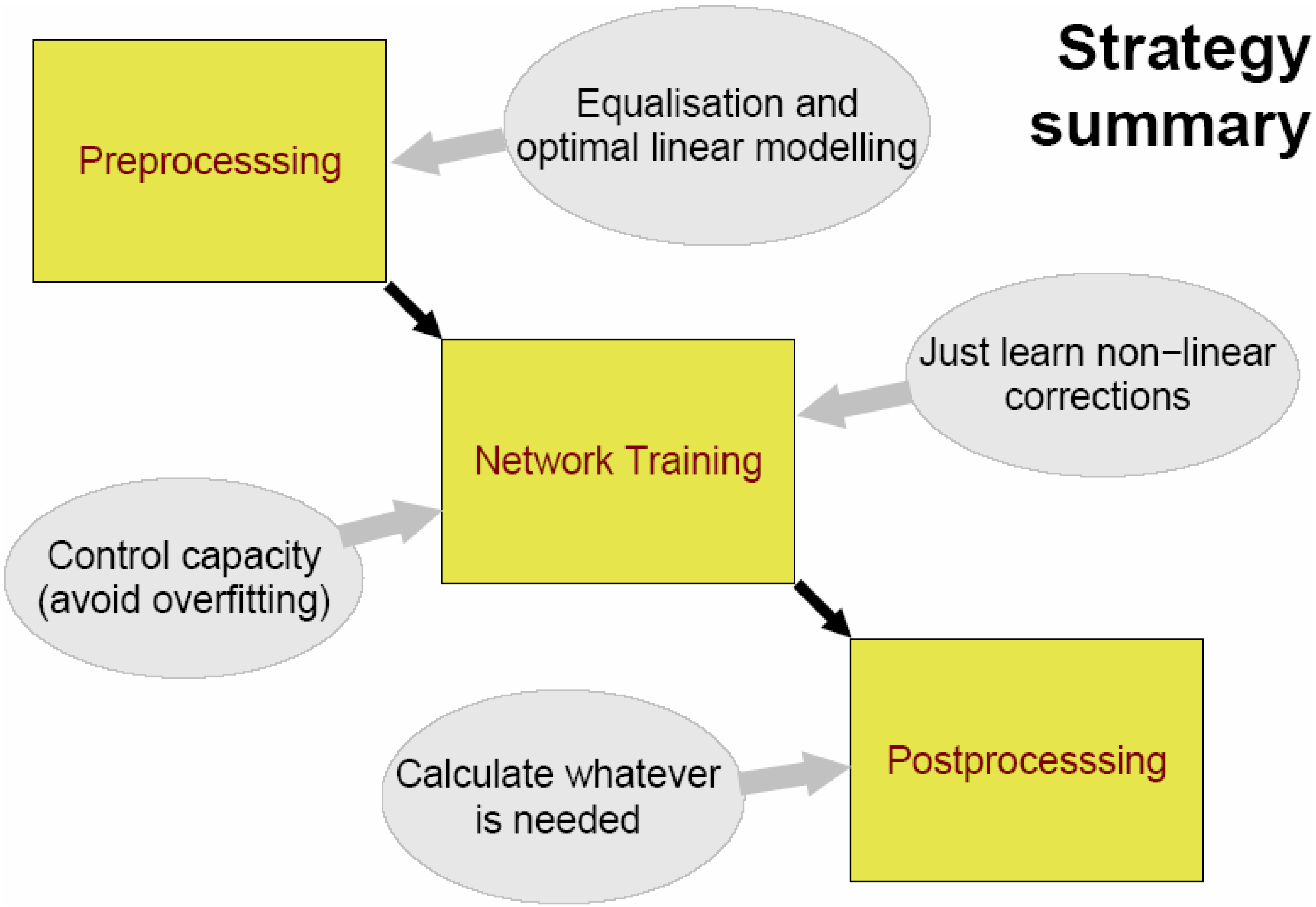
NeuroBayes Network architecture: Teacher



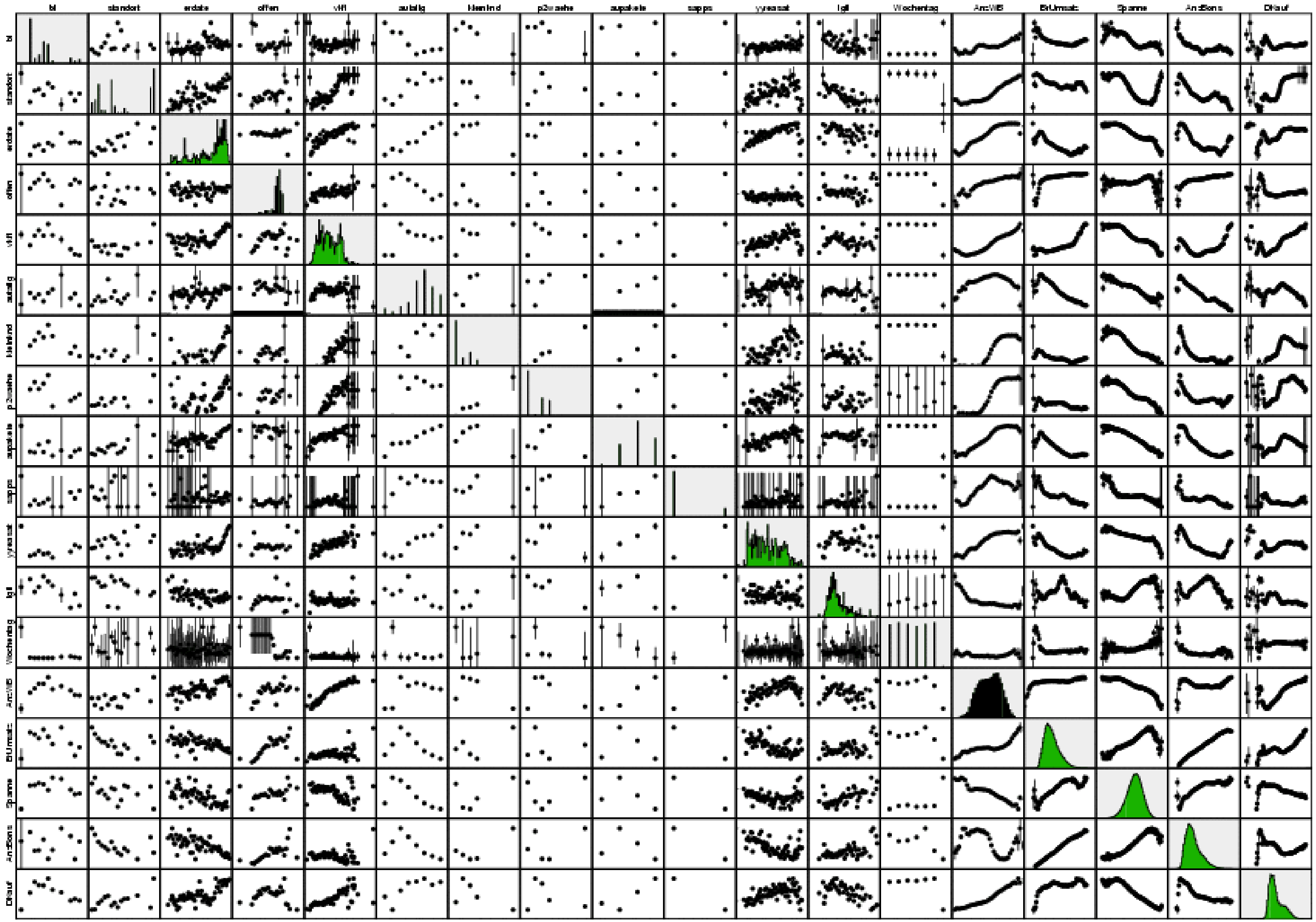
NeuroBayes network architecture: Expert



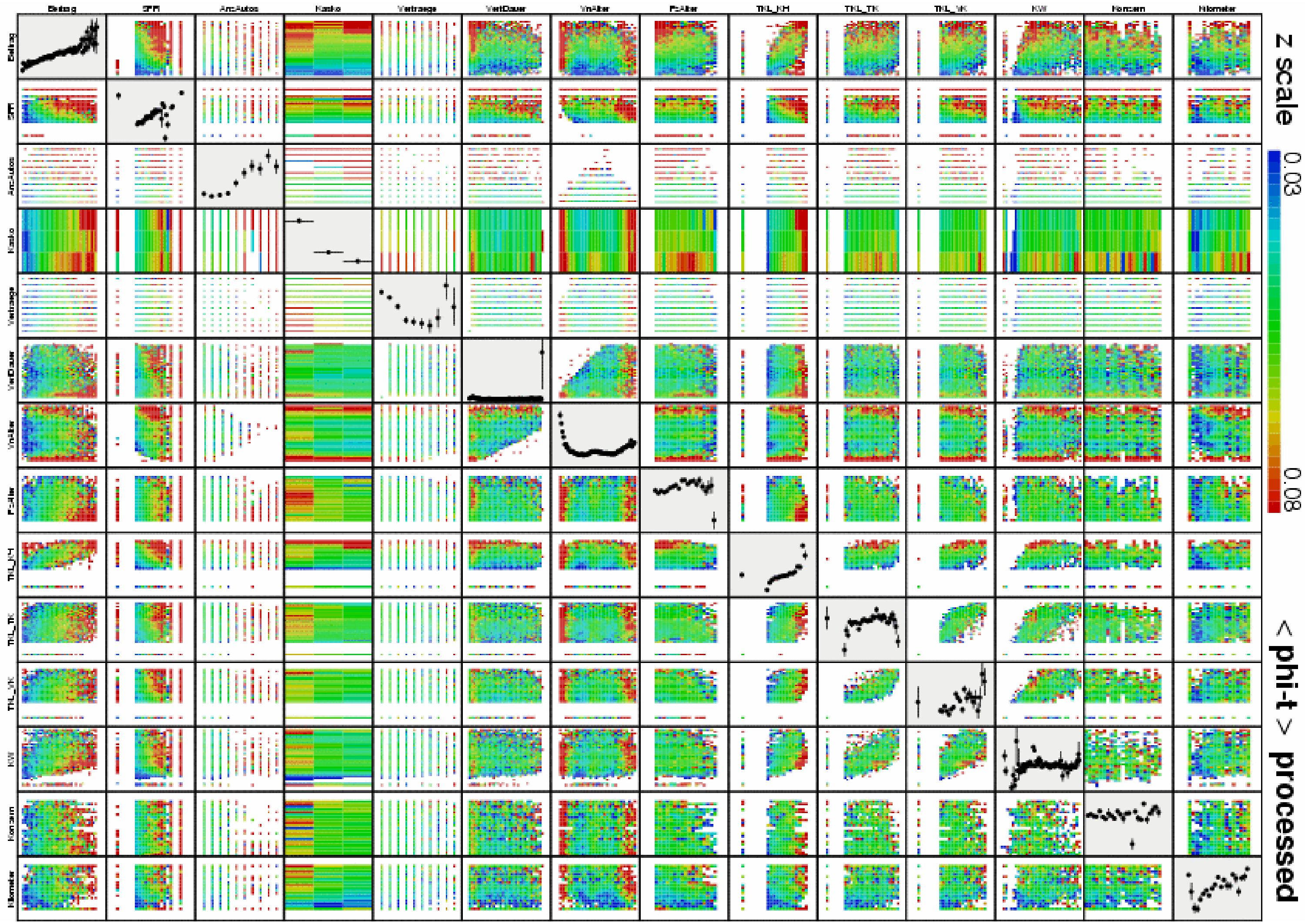
Strategy summary



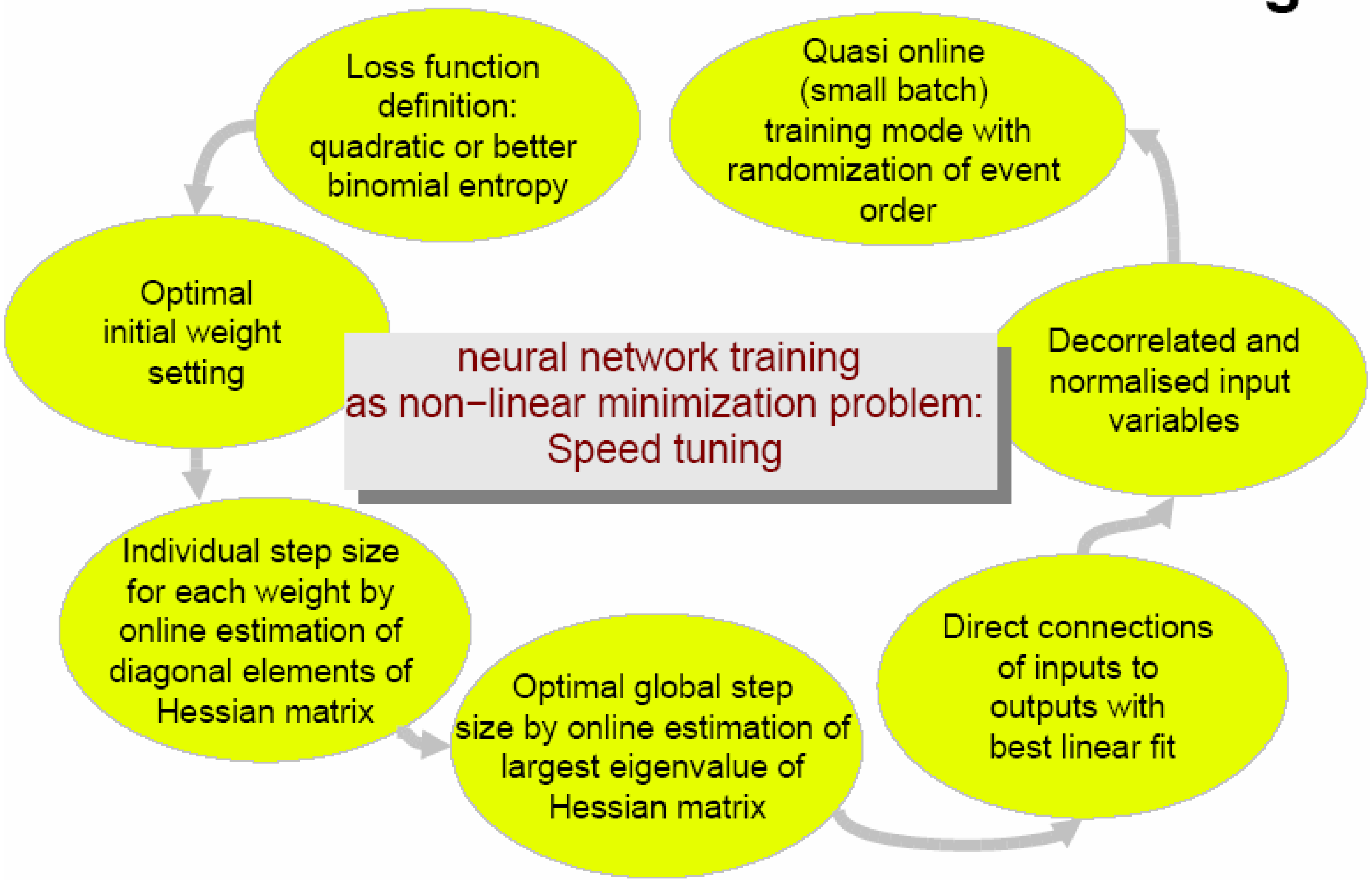
Ramler-plot (extended correlation matrix)



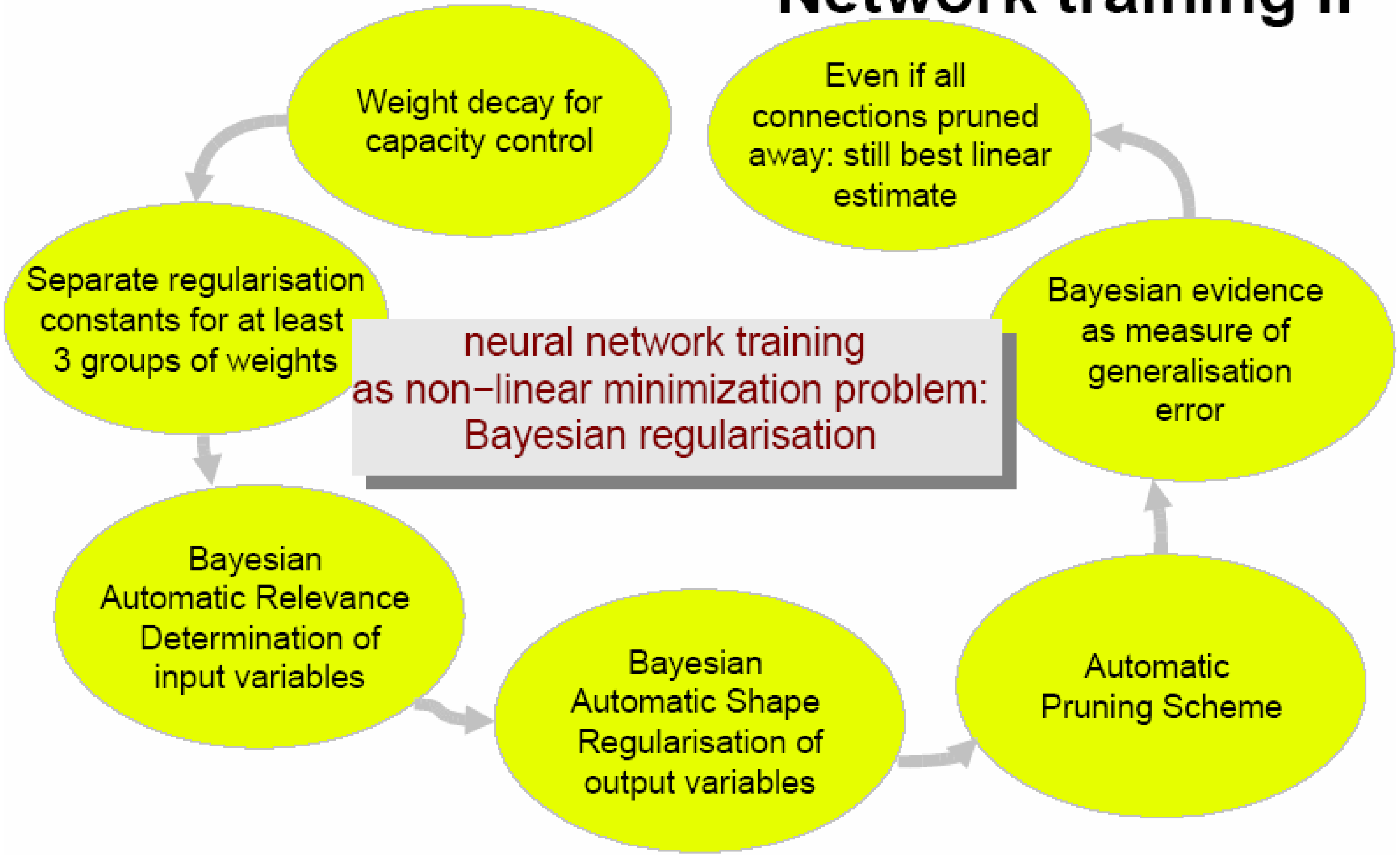
Ramler-II-plot (visualize correlation to target)



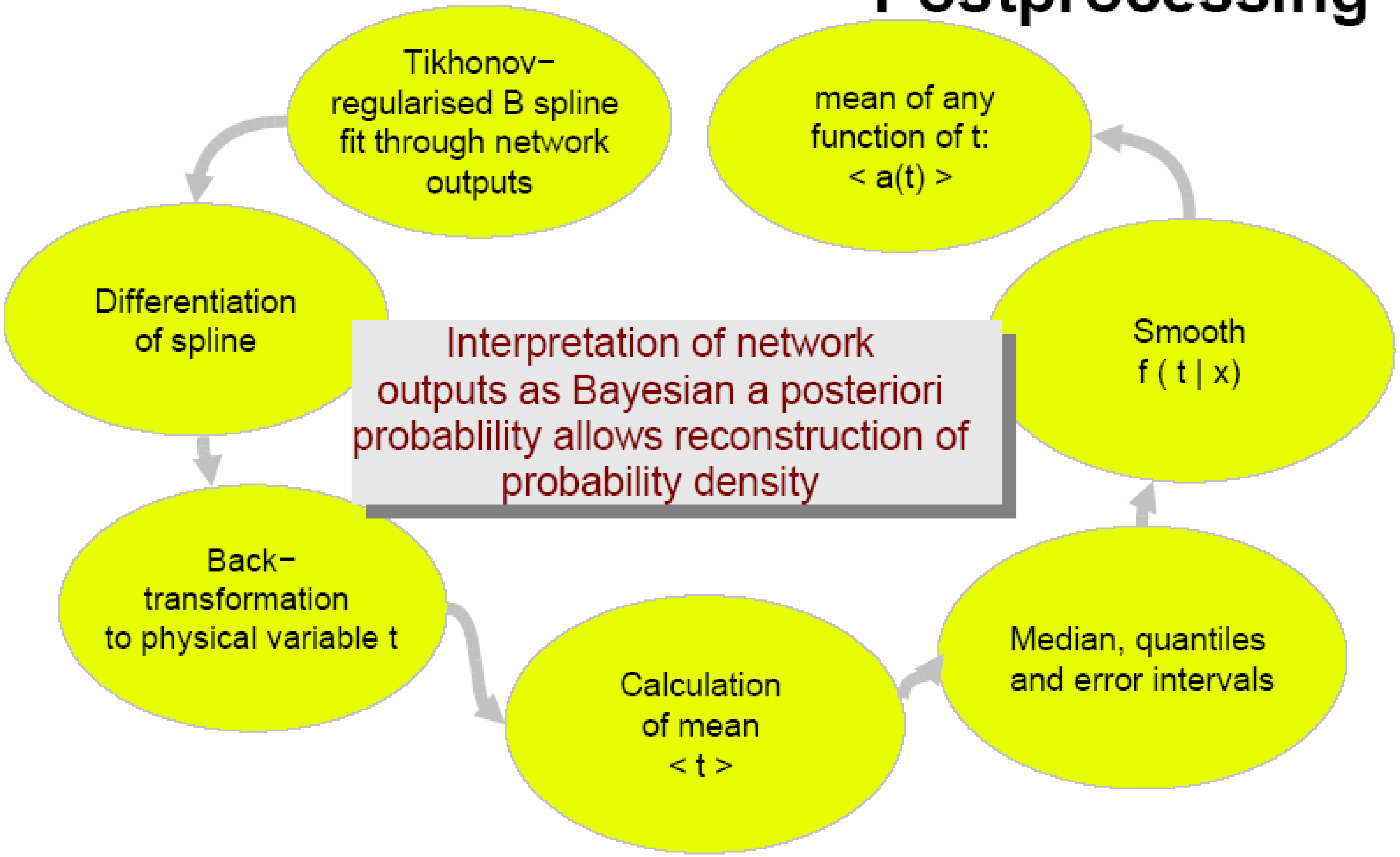
Network training I



Network training II



Postprocessing



Shape reconstruction details I

NeuroBayes solution ansatz

Discretize f(t) into N intervals of same area by equalisation (nonlinear transformation t –> s)

↓

Train a neural network with N output nodes to the N binary decisions:
The true t is larger than / lower than threshold i

↓

Fit smooth function (cubic spline) through N net outputs:
= cumulated conditional probability in transformed variable s

↓

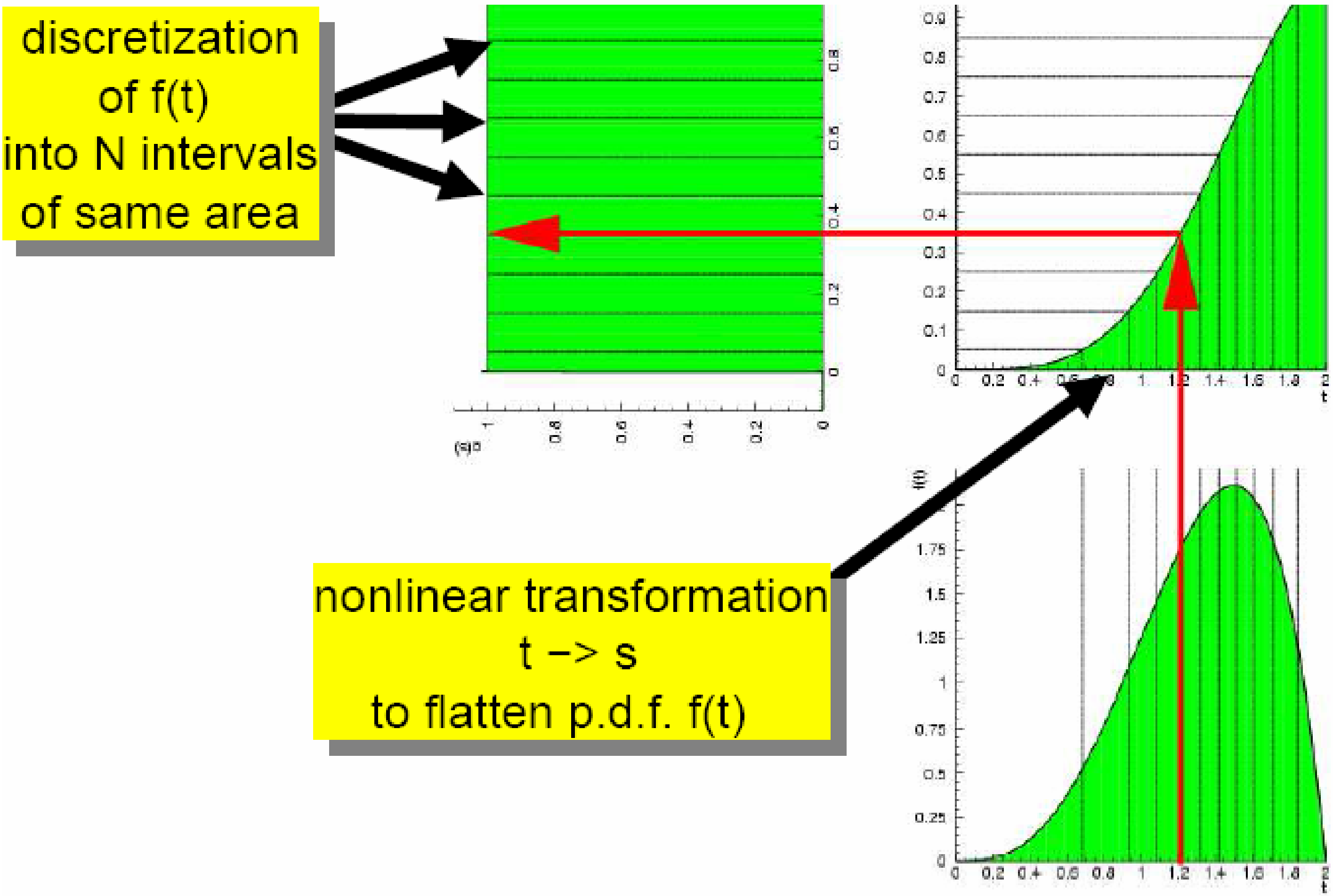
Analytic differentiation returns probability density function in transformed variable s

↓

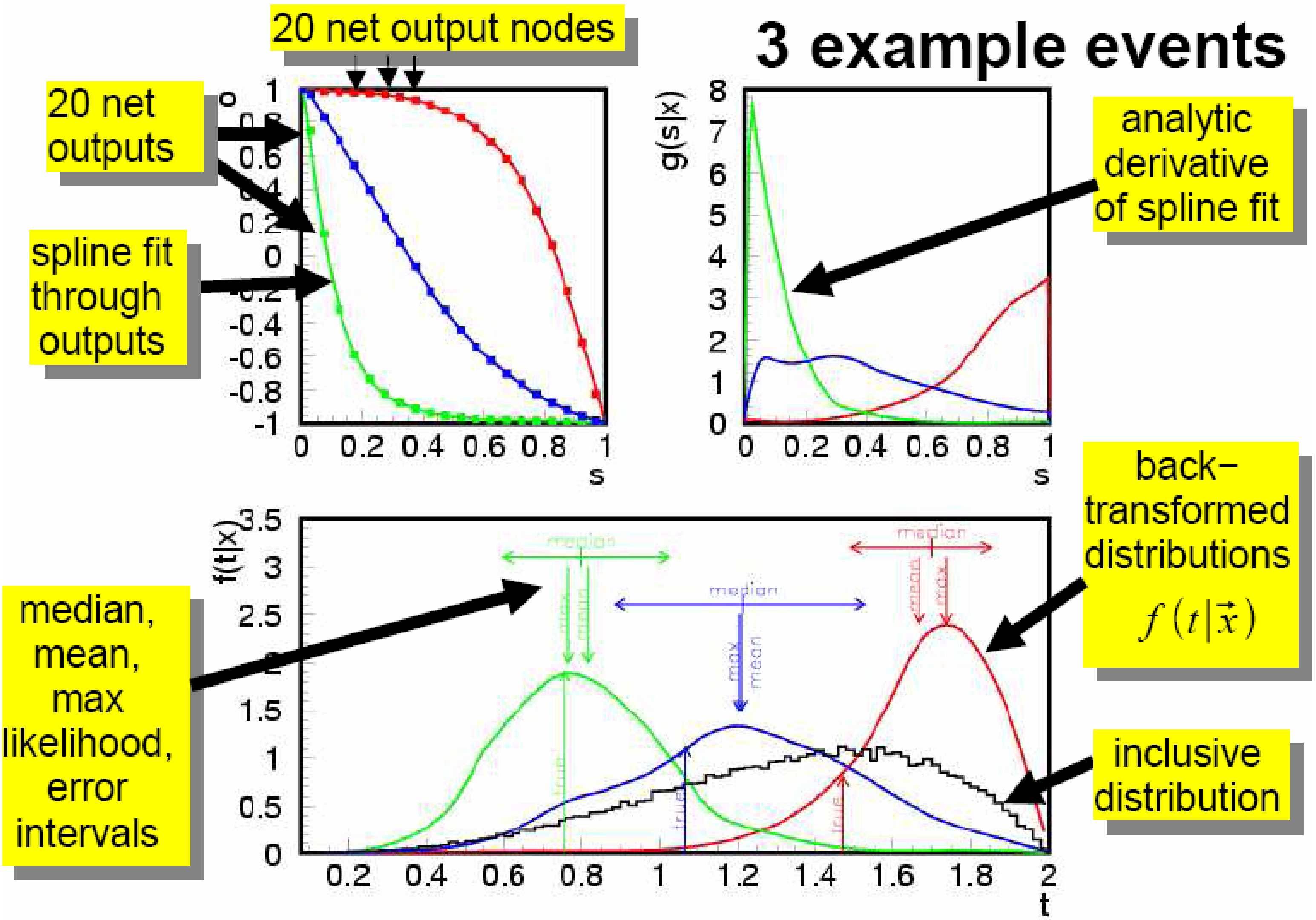
Back transformation to variable t returns f(t|x)

Shape reconstruction details II

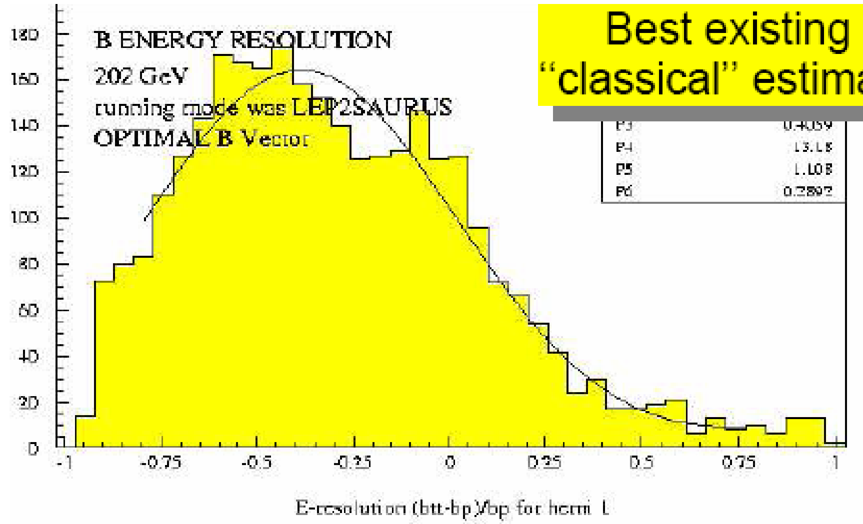
Equalisation and discretisation

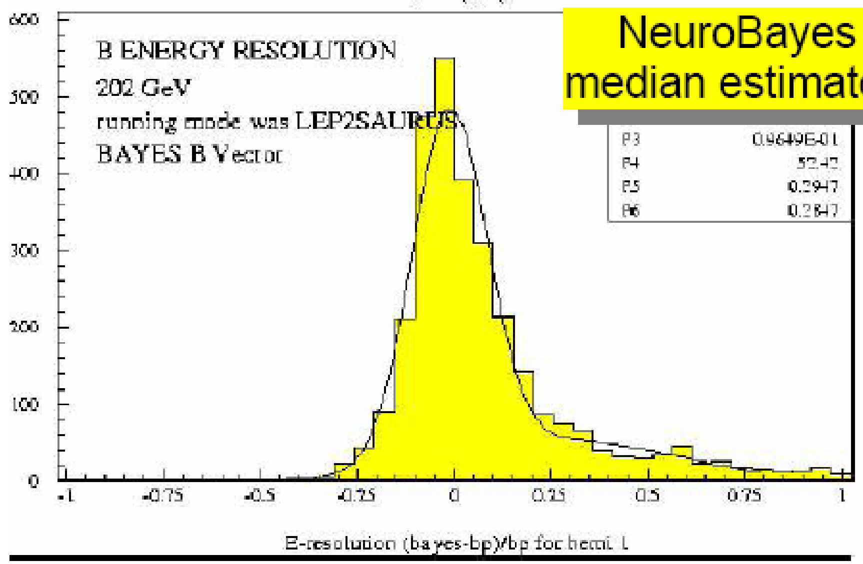


Shape reconstruction details III



Shape reconstruction example





B hadron energy

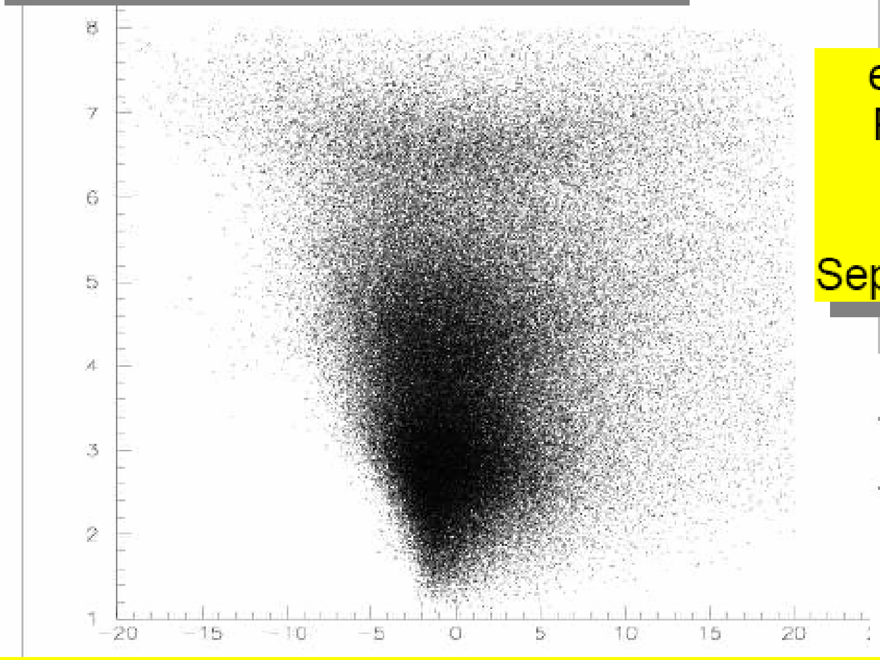
Relative resolution
of reconstructed
B hadron energy
in DELPHI at LEP II
at 202 GeV energy

(completely inclusive)

core resolution
40% –> 10%

Error estimates

NeuroBayes mean error estimator



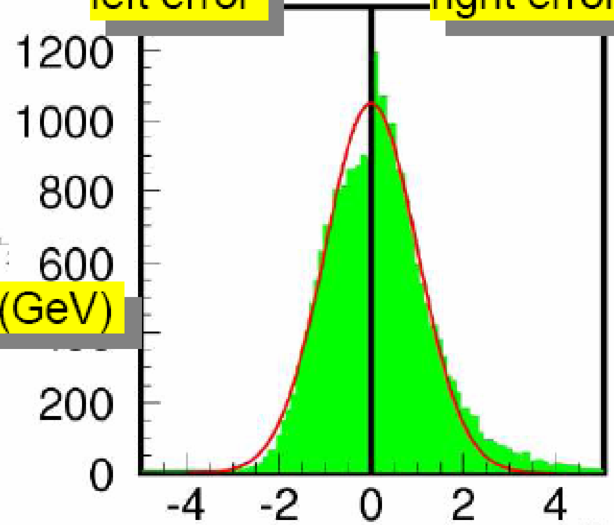
NeuroBayes median estimator – true energy (GeV)

error estimates make sense!
Pulls are almost Gaussians
of width 1

Separate left and right uncertainty

left error

right error



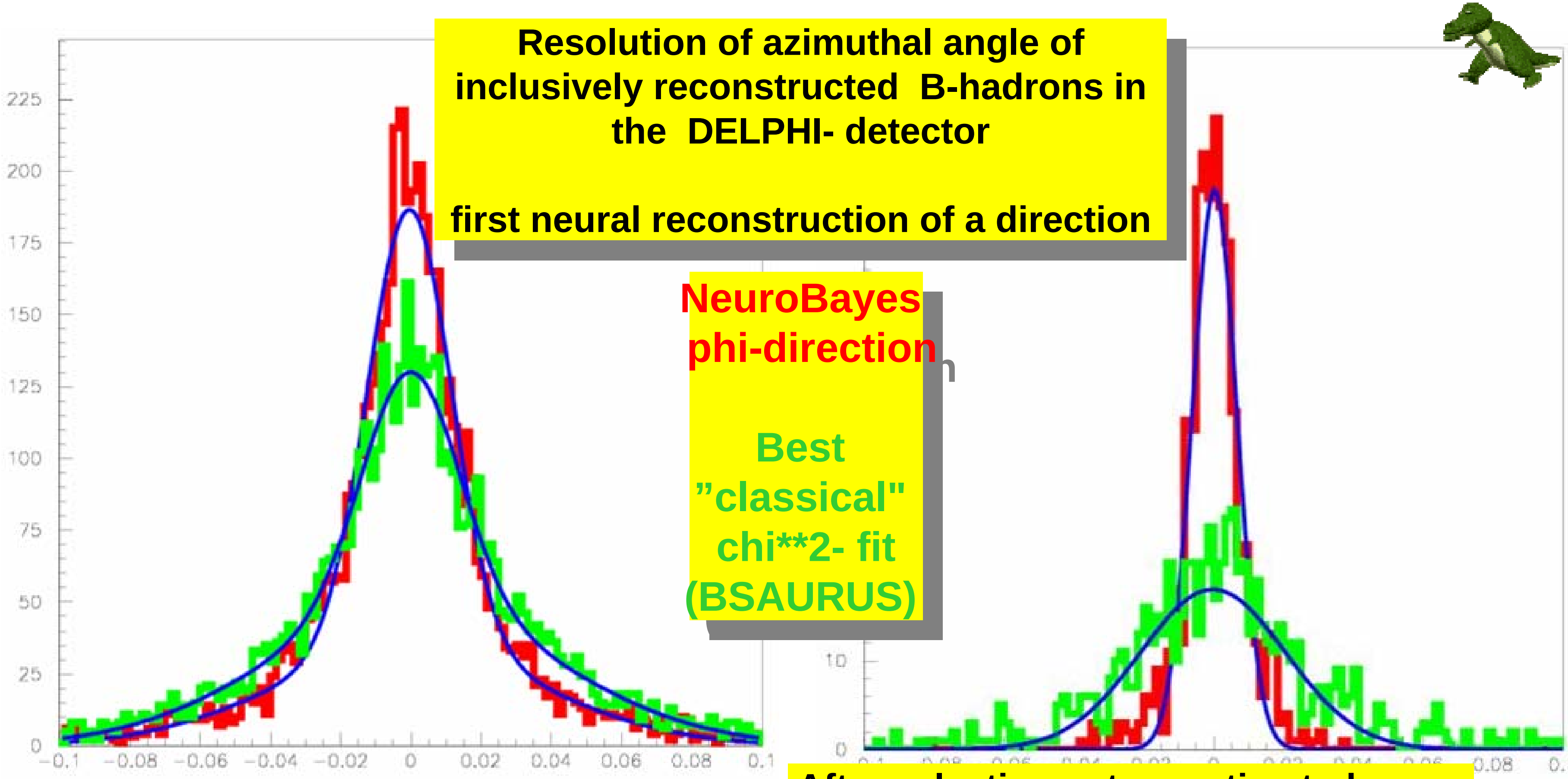
pull

Resolution of reconstructed
B hadron energy in DELPHI at LEP I
(completely inclusive)

Direction of B-mesons (DELPHI)

Resolution of azimuthal angle of inclusively reconstructed B-hadrons in the DELPHI- detector

first neural reconstruction of a direction



NeuroBayes phi-direction

Best "classical" chi2- fit (BSAURUS)**

No selection:
Improved resolution

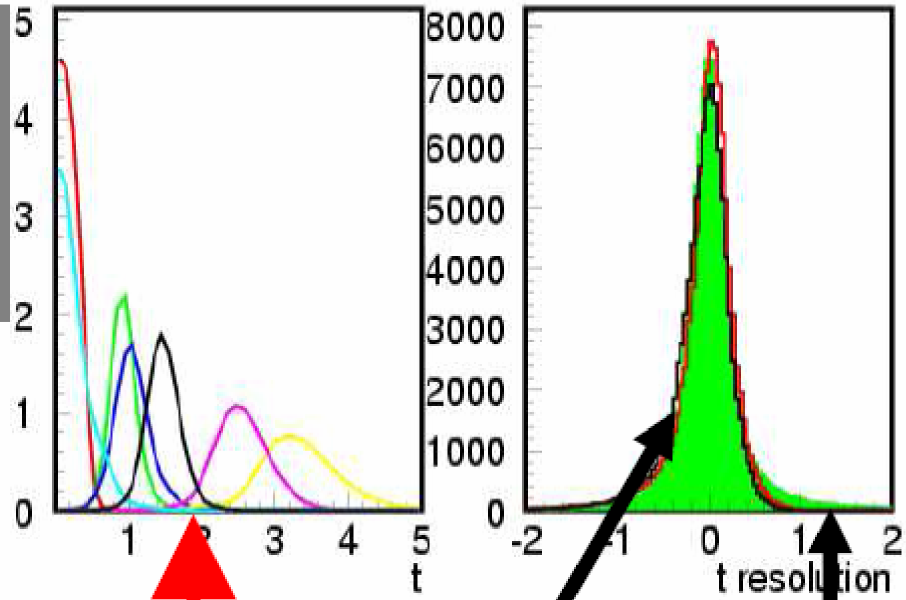
After selection cut on estimated error:
Resolution massively improved, no tails
==> allows reliable selection of good events

automatic error propagation

toy experiment :
measure (with errors)
•decay length d
•momentum p

and train for proper time t:

$$t = \frac{m}{c} \cdot \frac{d}{p}$$

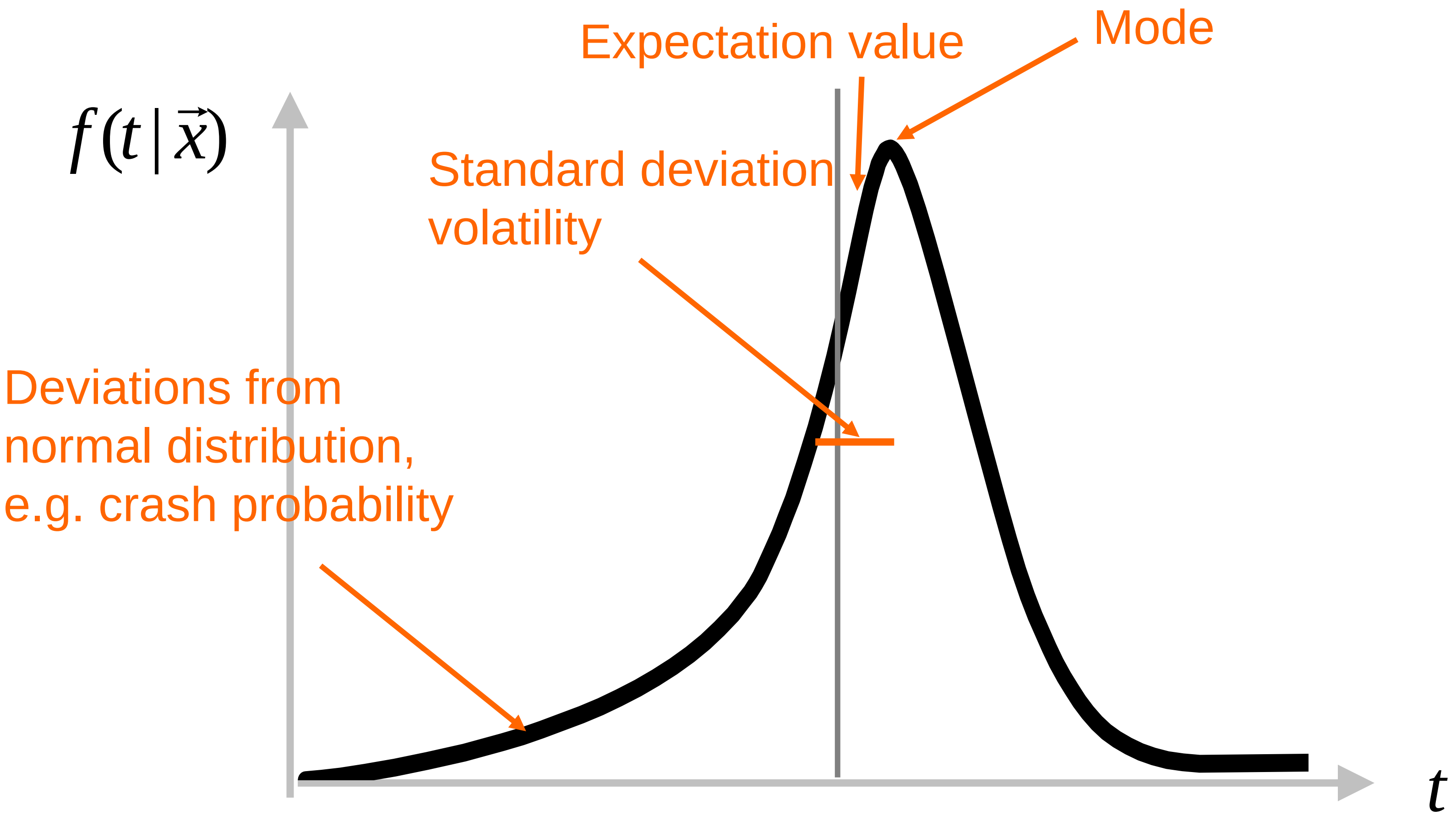


Max likelihood estimate
median estimate

classical approach:
tail from negative
lifetimes

Result :
networks learns automatically from data :
• that it should divide d by p
• how it should propagate errors
• true lifetimes are never negative (although both measured d and p can be)

Shape reconstruction in finance



Risk analysis for a car insurance BGV

Results for the Badischen Gemeinde-Versicherungen:

since May 2003: radically new tariff for young drivers!

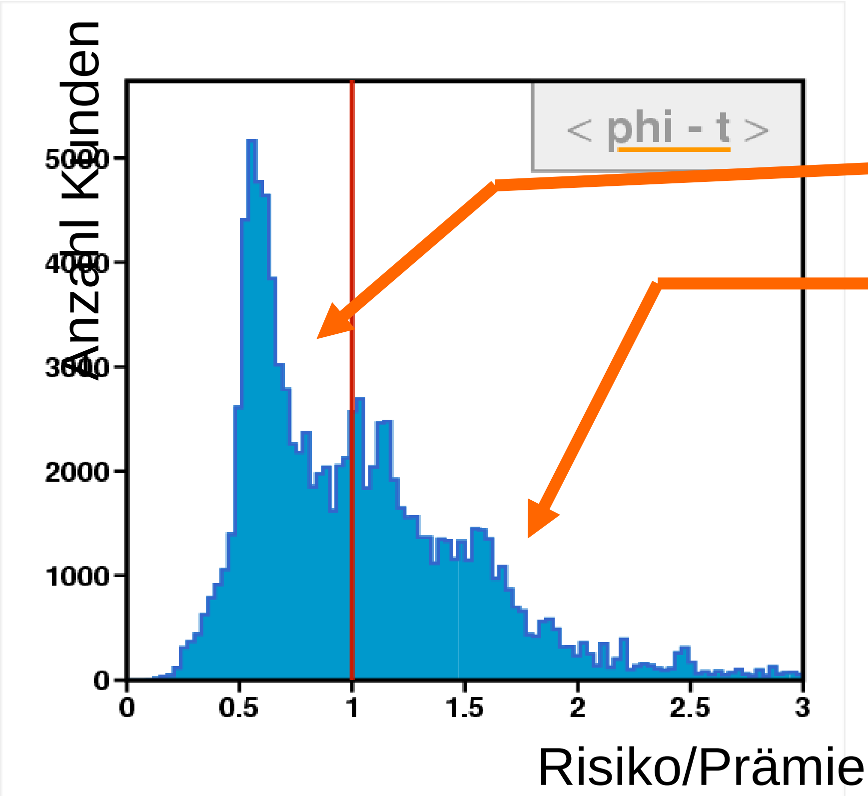
New variables added to calculation of the premium.
Correlations taken into account.

Risk und premium up to a factor of 3 apart from each other!
Even probability distribution of height of can be predicted

Premature contract cancellation also well predictable

The “unjustice“ of insurance premiums

Ratio of the accident risk calculated using NeuroBayes®
to premium paid (normalised to same total premium sum):



The majority of customers (with low risk) are paying too much.

Less than half of the customers (with larger risk) do not pay enough, some by far not enough.
These are currently subsidised by the more careful customers.

Prediction of contract cancellation

The prediction really holds:

Test on a a new statistic year