Looking data is always more important than ML technique

Implementation of energy estimation of gamma ray in MAGIC data analysis by Random Forest

Kazuma Ishio

MAGIC
Major Atmospheric
Gamma Imaging
Cerenkov Telescope

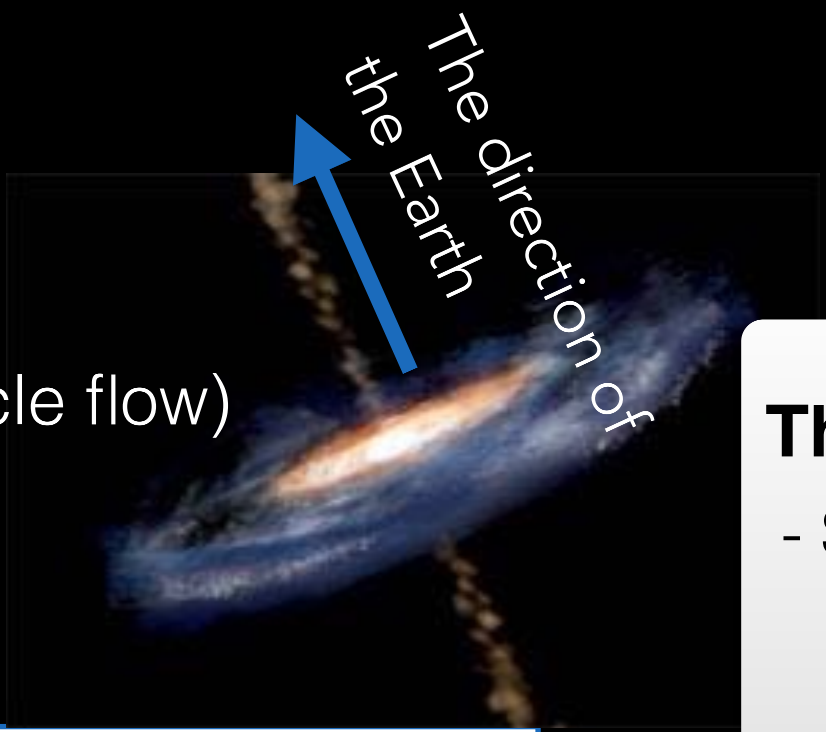
MAX-PLANCK-GESELLSCHAFT

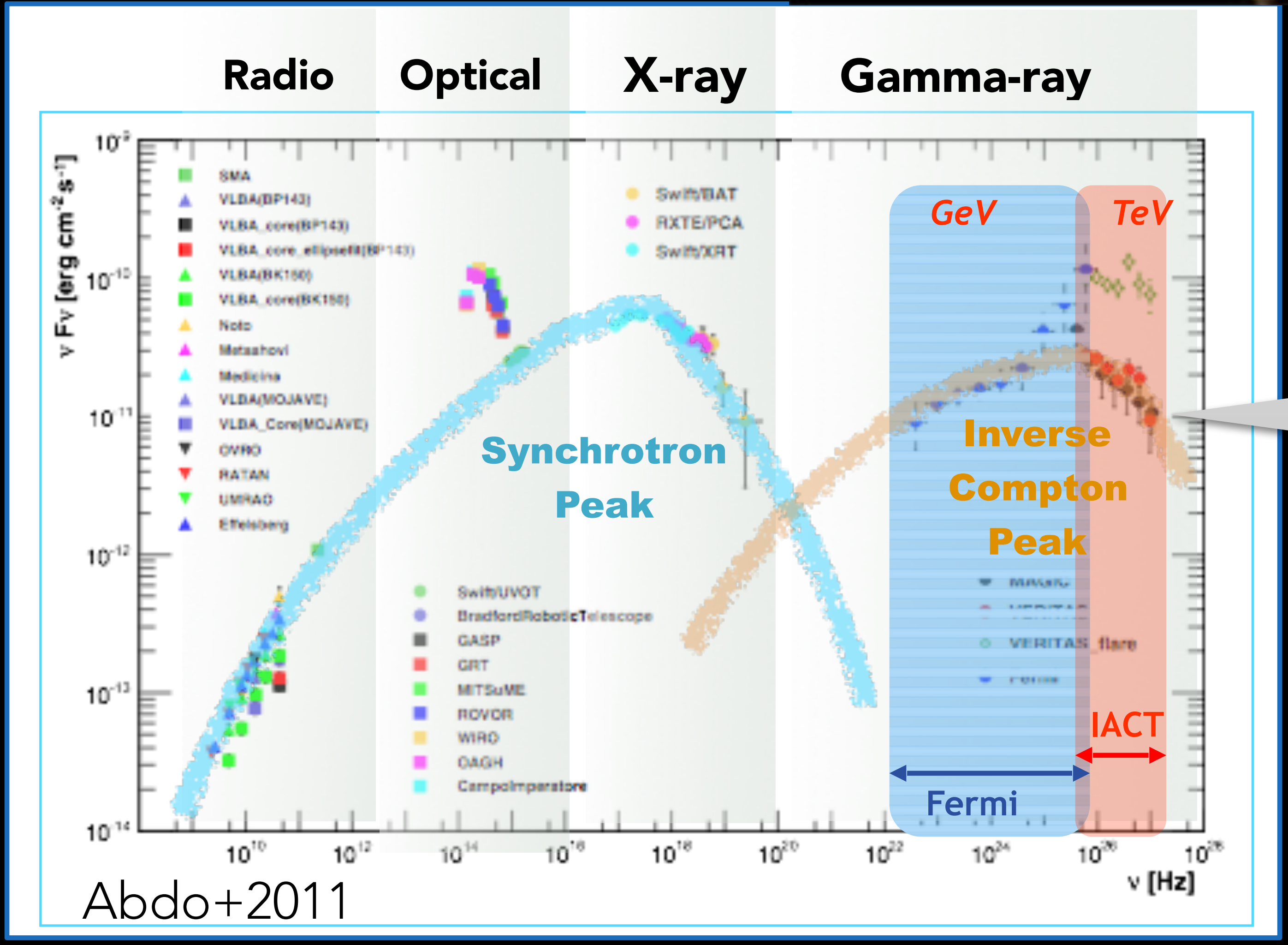
$\Delta p \cdot \Delta q \geq \frac{1}{2}\hbar$

Max-Planck-Institut für Physik
(Werner-Heisenberg-Institut)

Blazar

Central region of a galaxy
with a jet (relativistic particle flow)
towards the Earth
-> Bright in gamma ray





Abdo+2011

There may be irregularities

- Spectrum break or additional component by multiple zones or multiple emission components
- Additional component by contrubution from DM
- Cut off by absorption by extra-galactic background light
- Oscillation by transition to Axion-Like-Particle

+ Very variable

Better energy reconstruction allows for better science

MAGIC telescopes

La Palma(29°N, 18°W), asl. 2200m

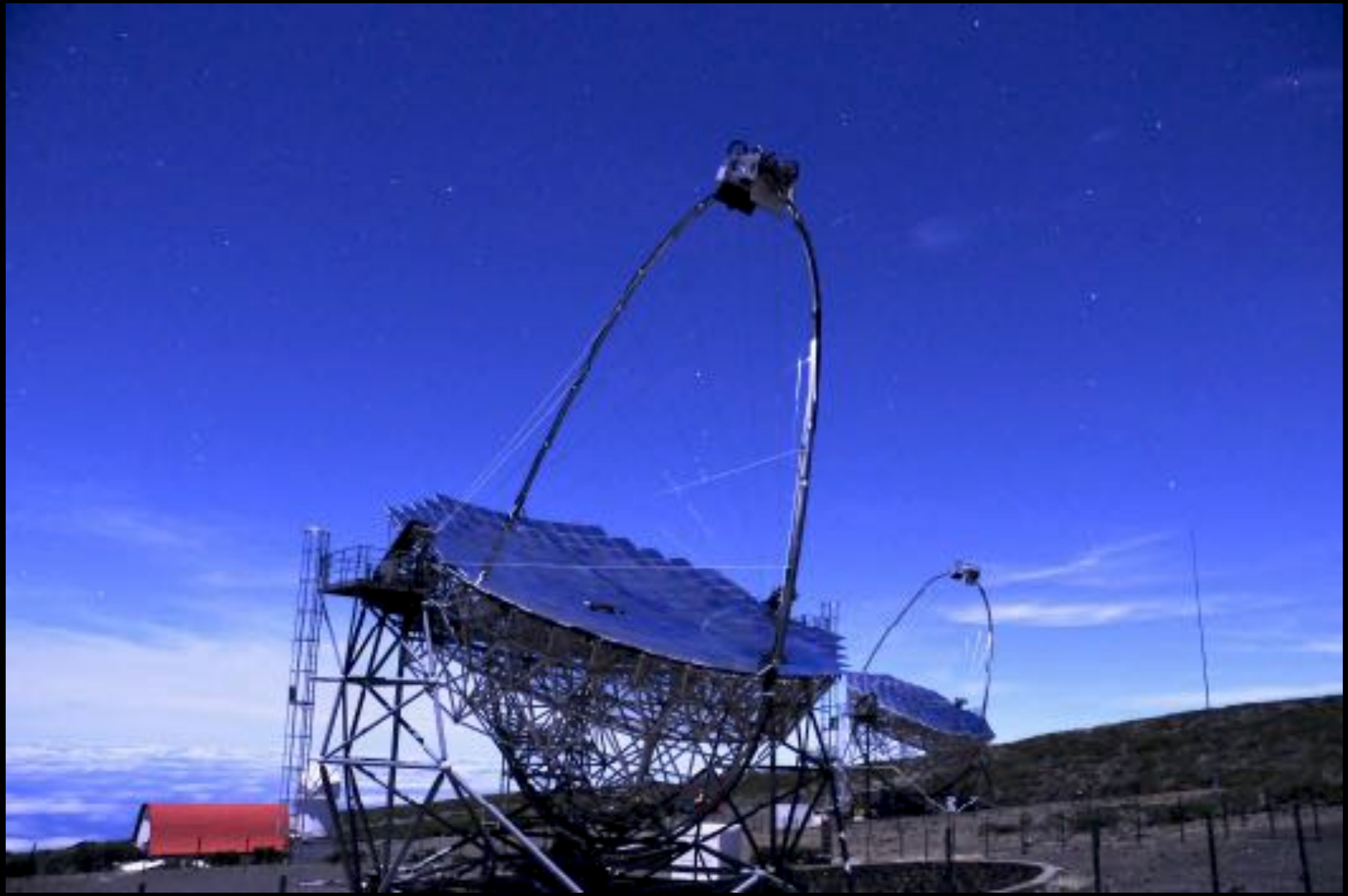
Imaging Atmospheric Cherenkov Telescope (IACT)

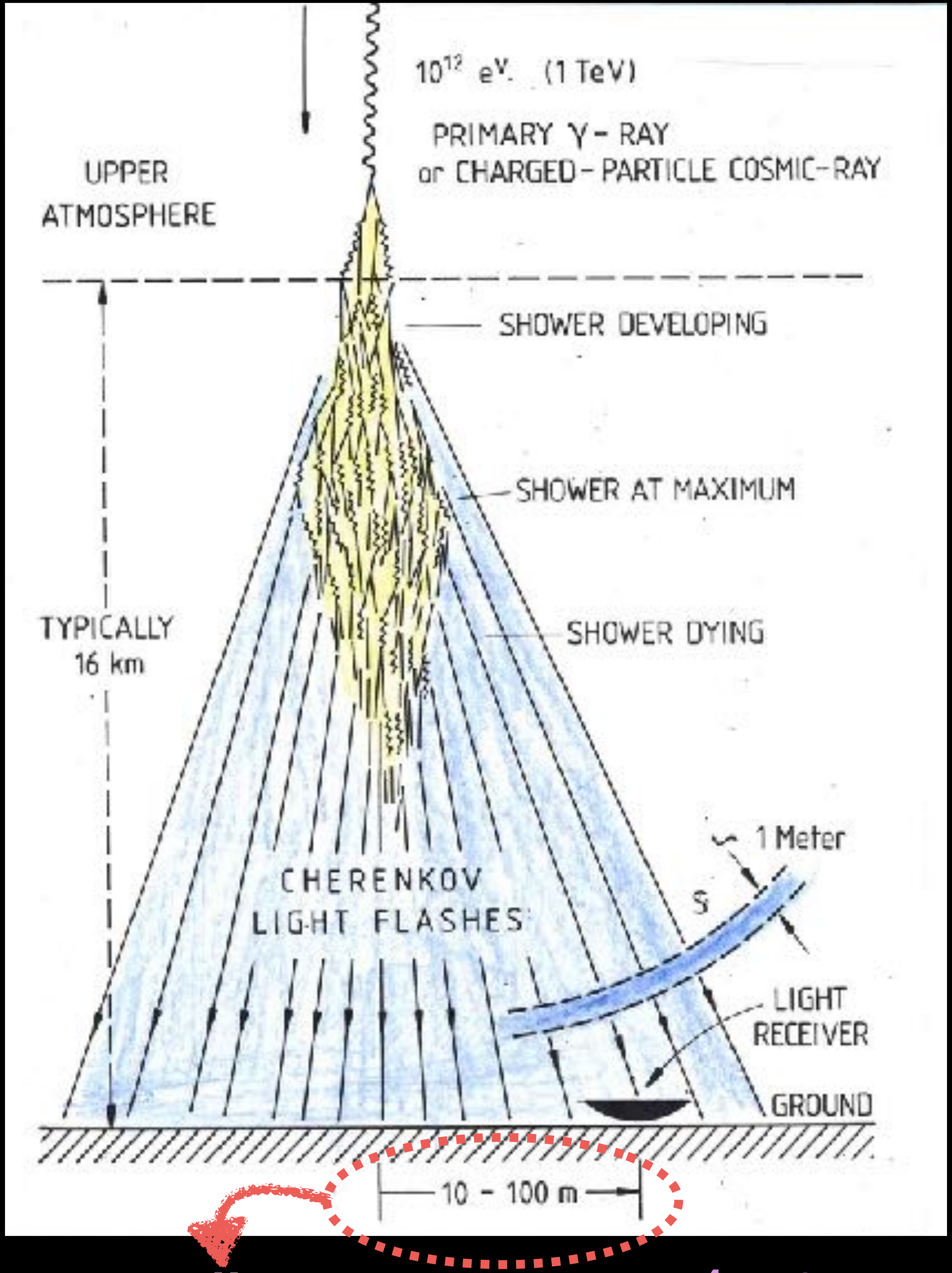
2 telescopes with

- **Dish diameter : 17m**
- **Camera FoV : 3.5deg**

Energy threshold of gamma ray : ~50 GeV

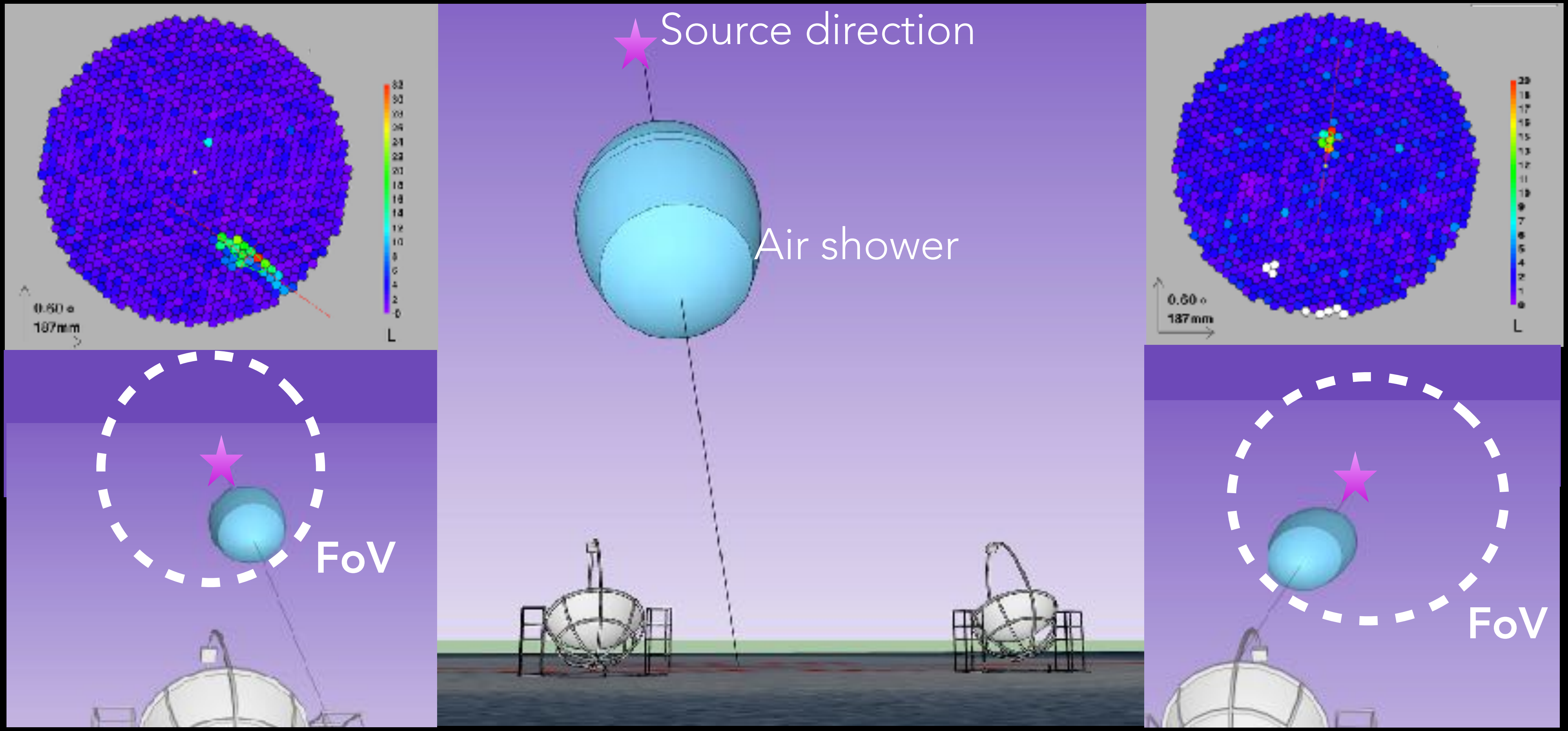
Sensitivity : ~0.7% of Crab flux above 0.2TeV





collection area ~ 10^4 m^2

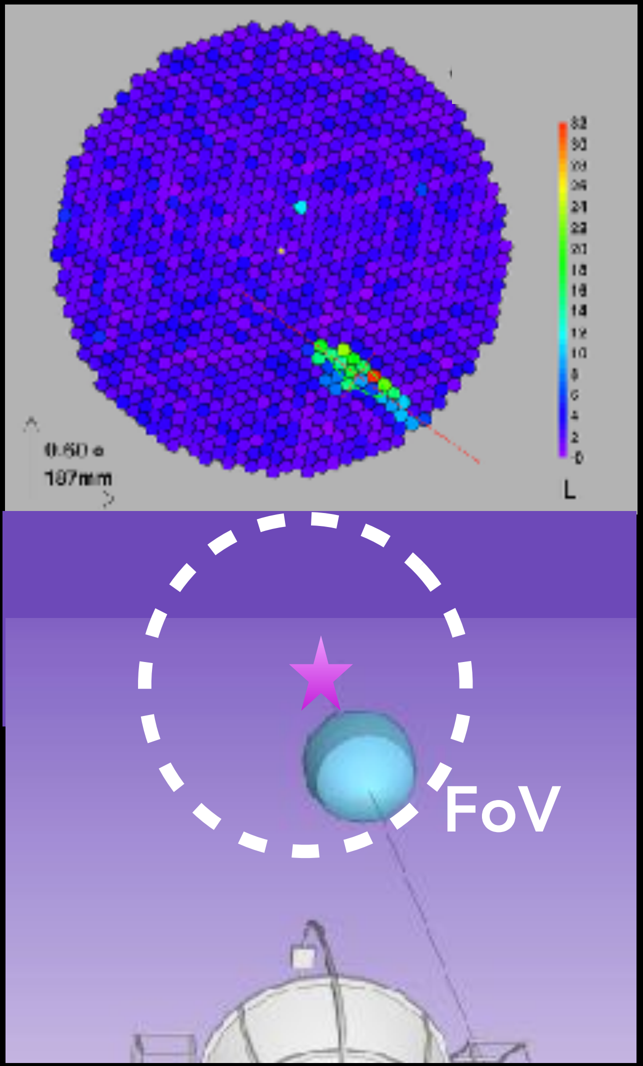
Images of the shower event



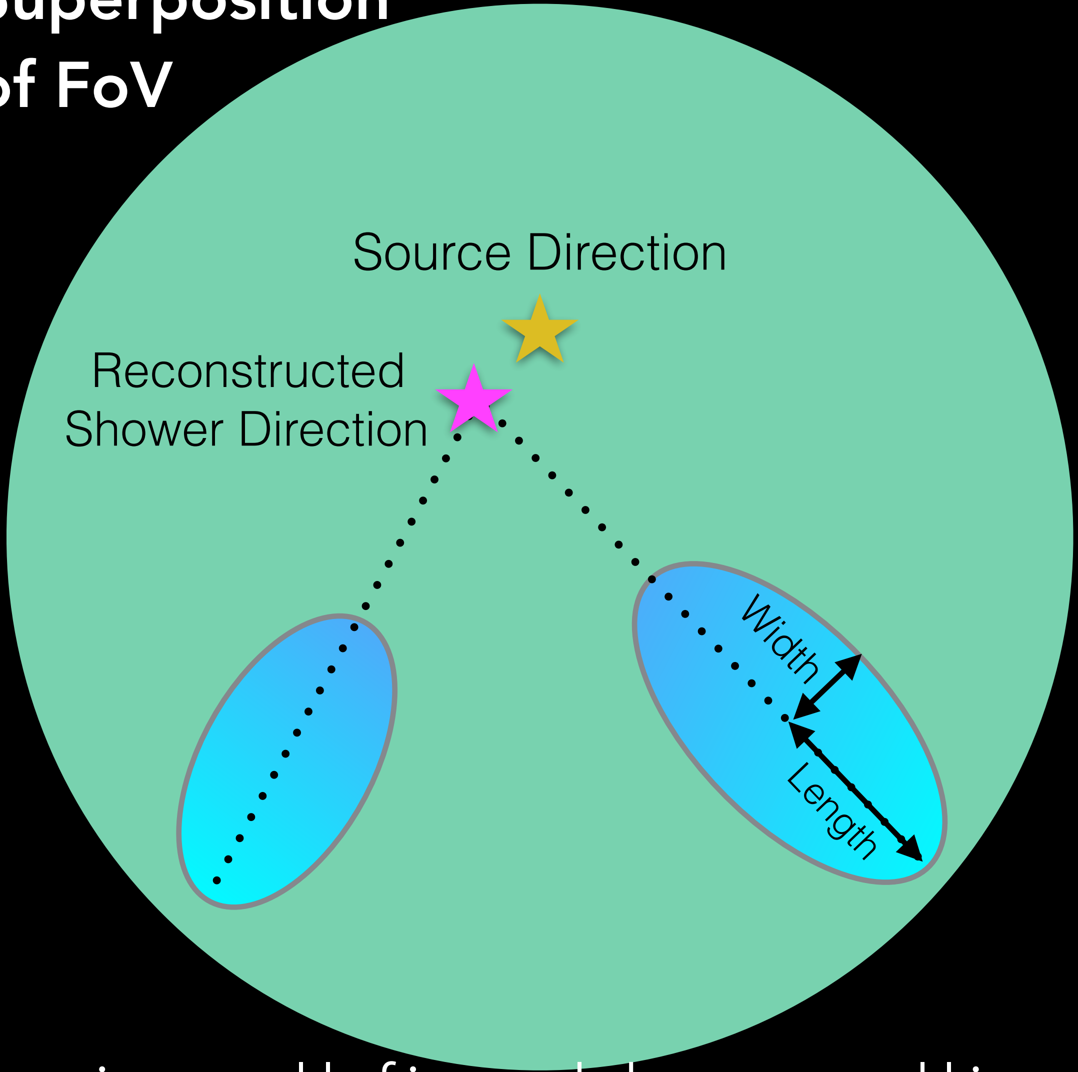
- 2 telescopes **capture the image of the shower** detecting faint Cherenkov flash within a few ns.
- From **light content** and **timing information** in each pixel.

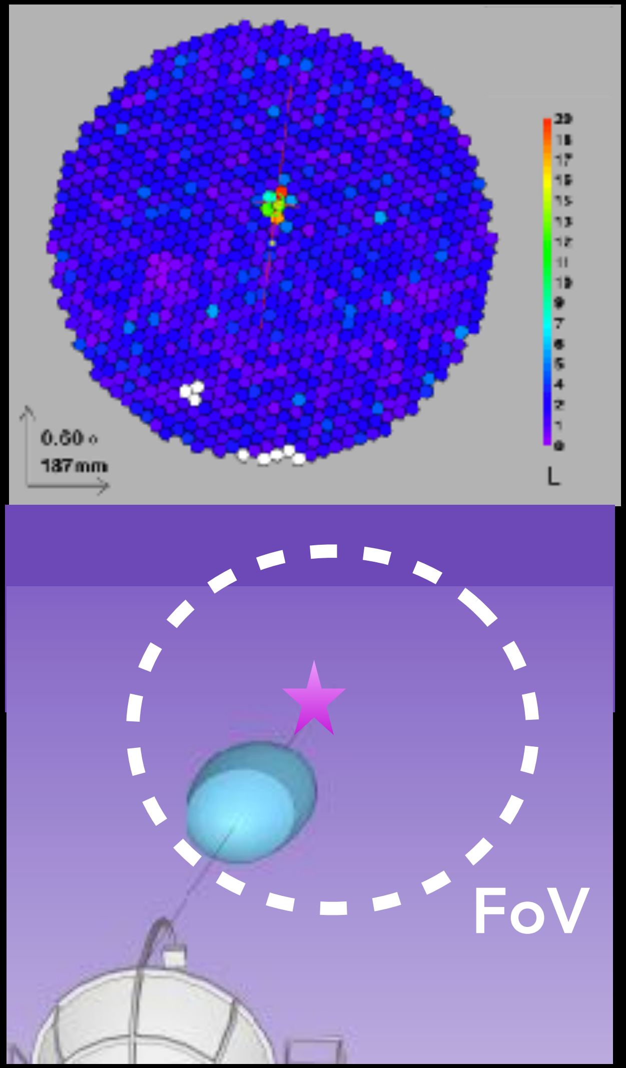
From each event we want to know the direction, the energy and the gamma-likeliness

Parametrisation of the shape



Superposition of FoV

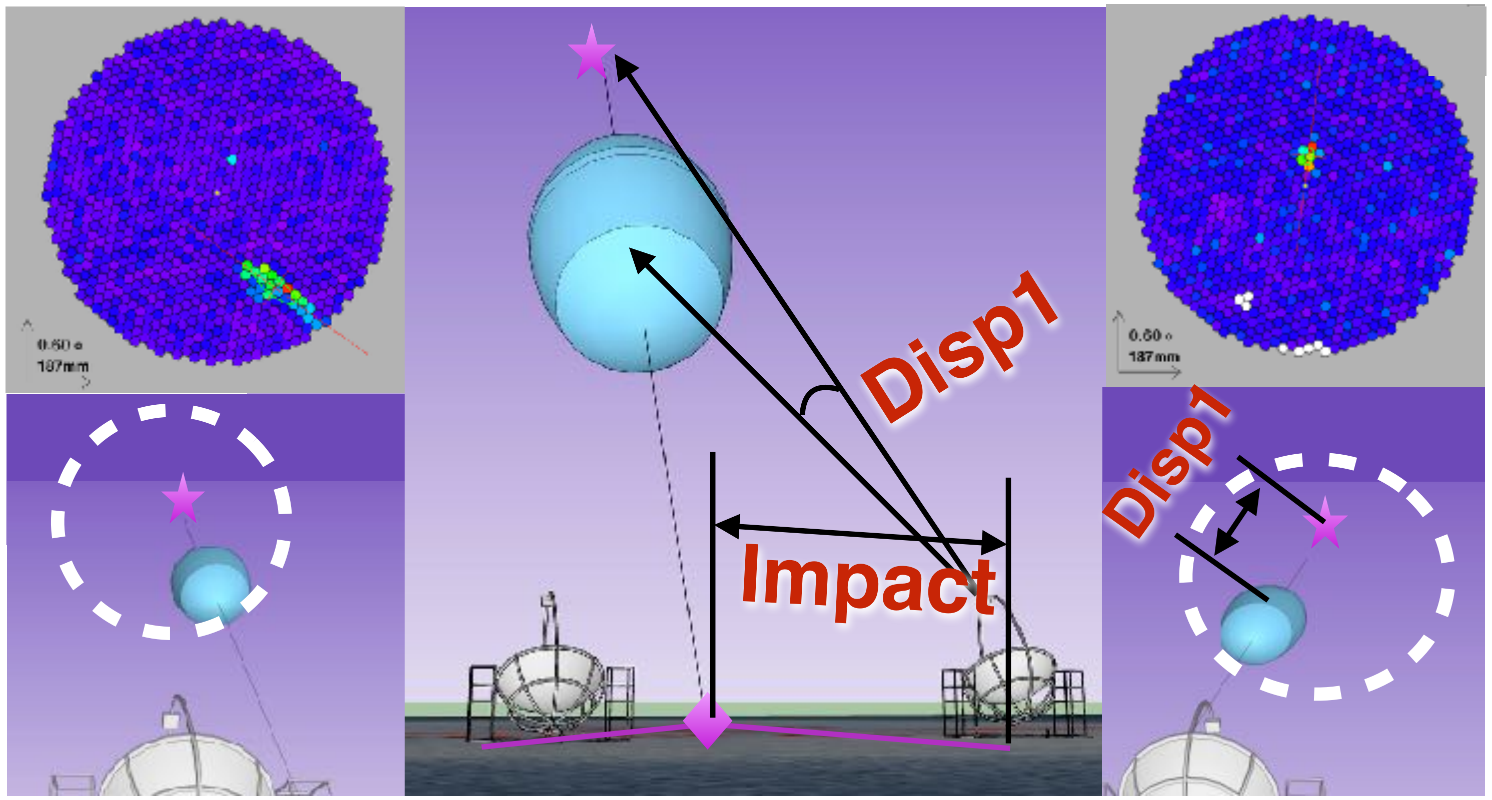




- A shower image is well fitted by an ellipse.
- Superposition of two images gives stereo information.
- The axes-intersection indicates the shower direction.
  But there is a better way called "Disp" method.

→ **The direction is very important crew also for 3 dimensional informations.** → **Better by "Disp"**

Disp: What is Disp? How is it important?



Disp

the displacement between the incoming direction and the image centroid in FoV

(displacement of incoming direction from the shower)

↓

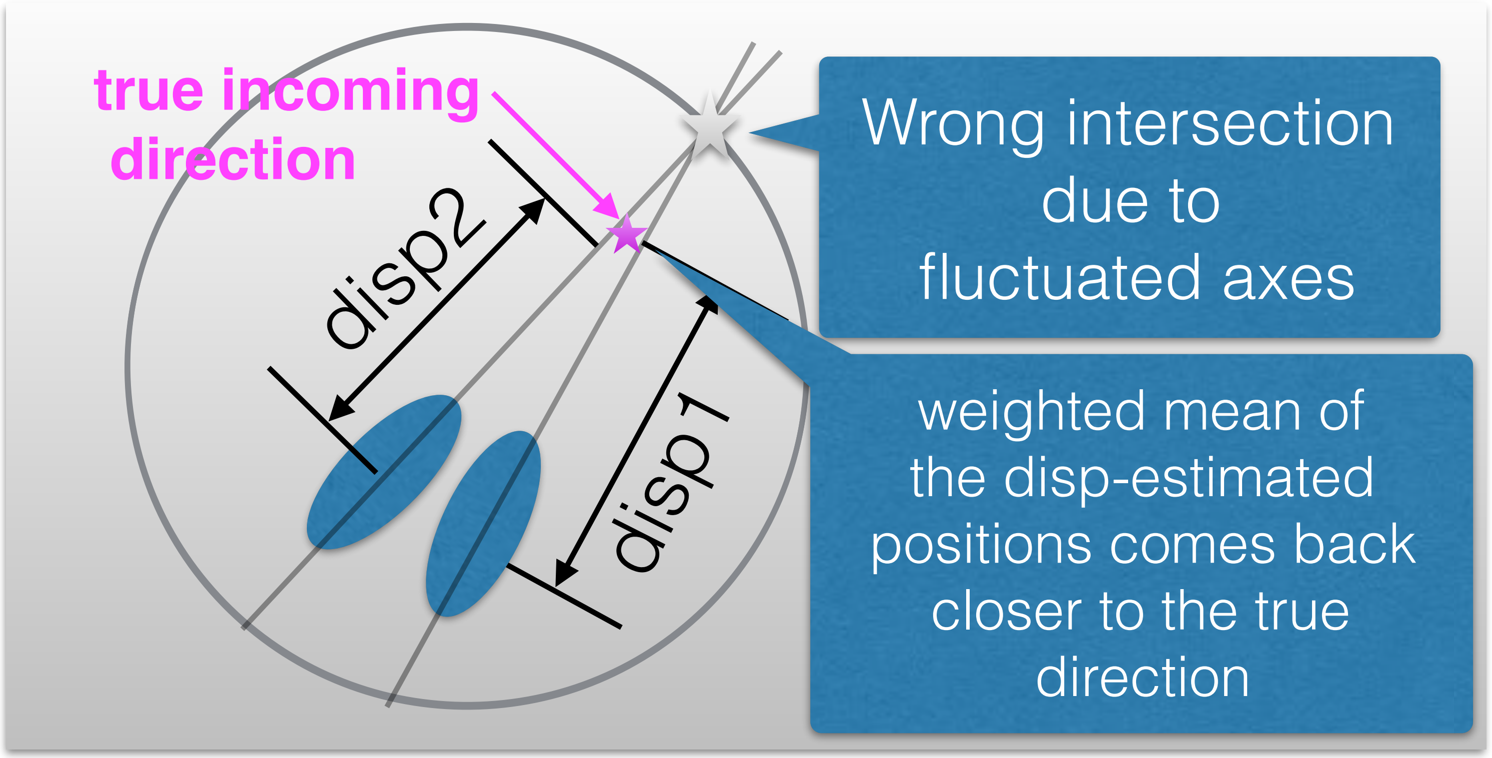
Using RandomForest, we can estimate proper “disp”.

↓

direction reconstruction performs much better.

↓

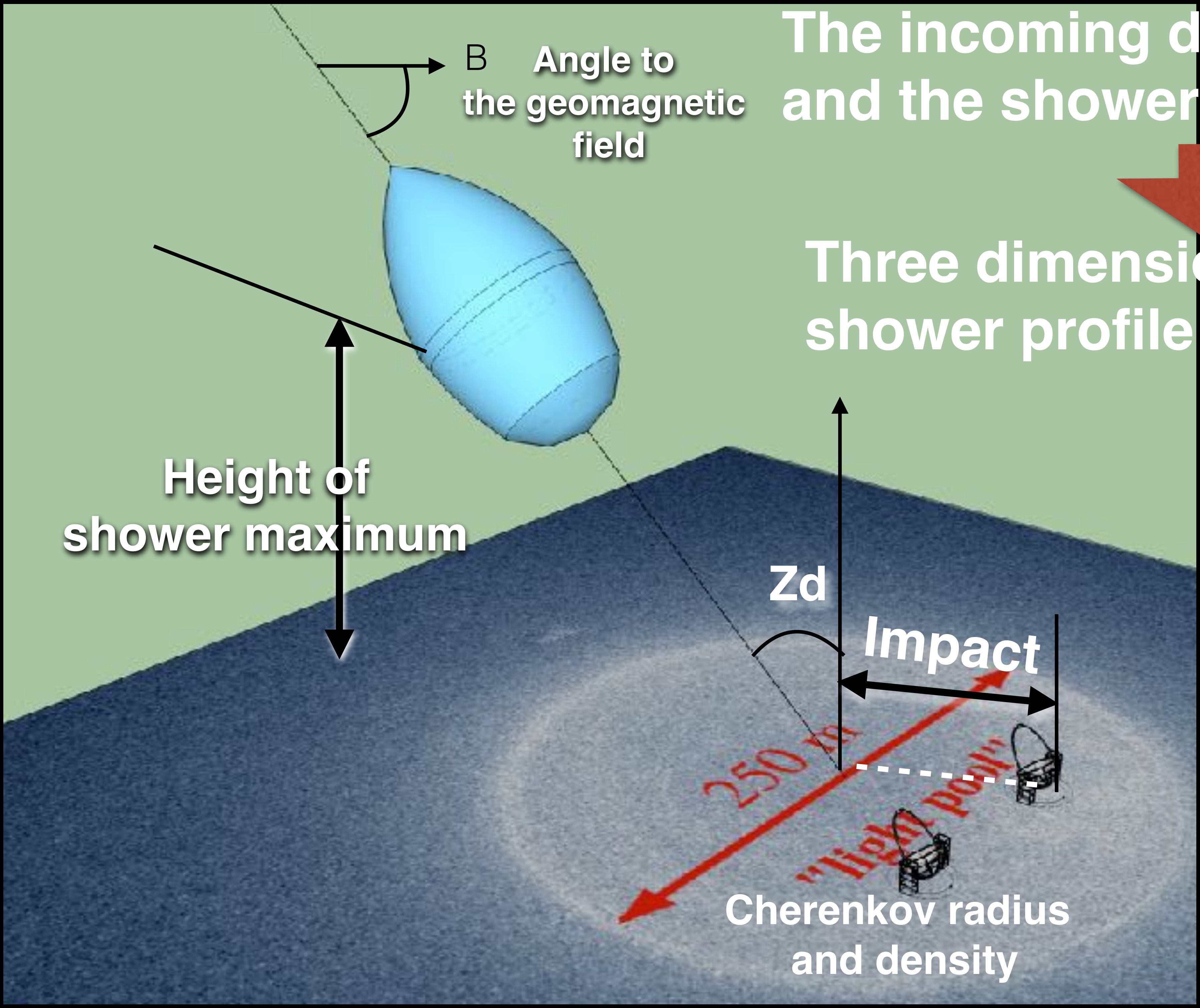
We can obtain better geometrical informations.



Wrong intersection due to fluctuated axes

weighted mean of the disp-estimated positions comes back closer to the true direction

Geometrical parameters



The incoming direction and the shower maximum direction

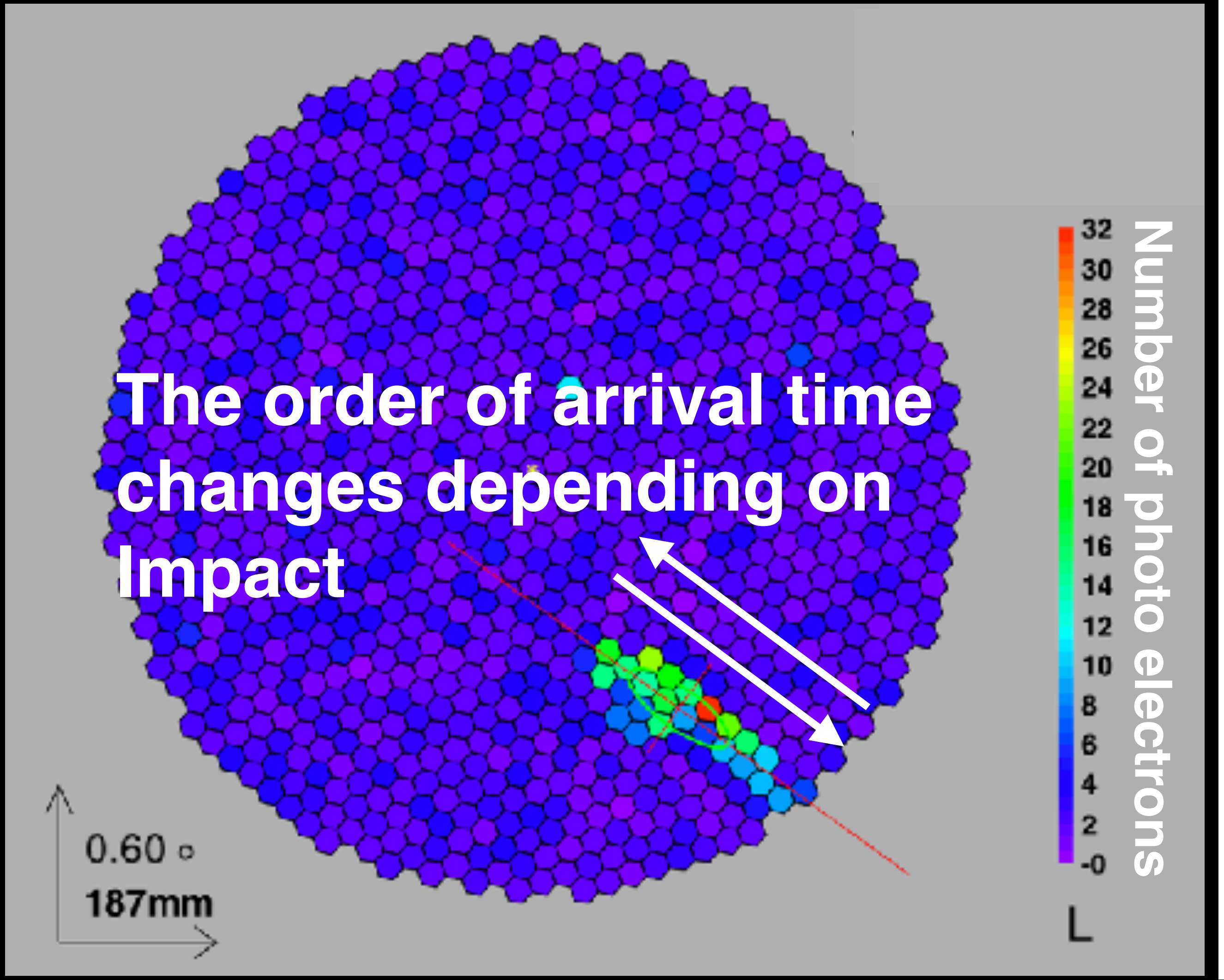
↓

Three dimensional shower profile reconstruction

Timing enriches the information

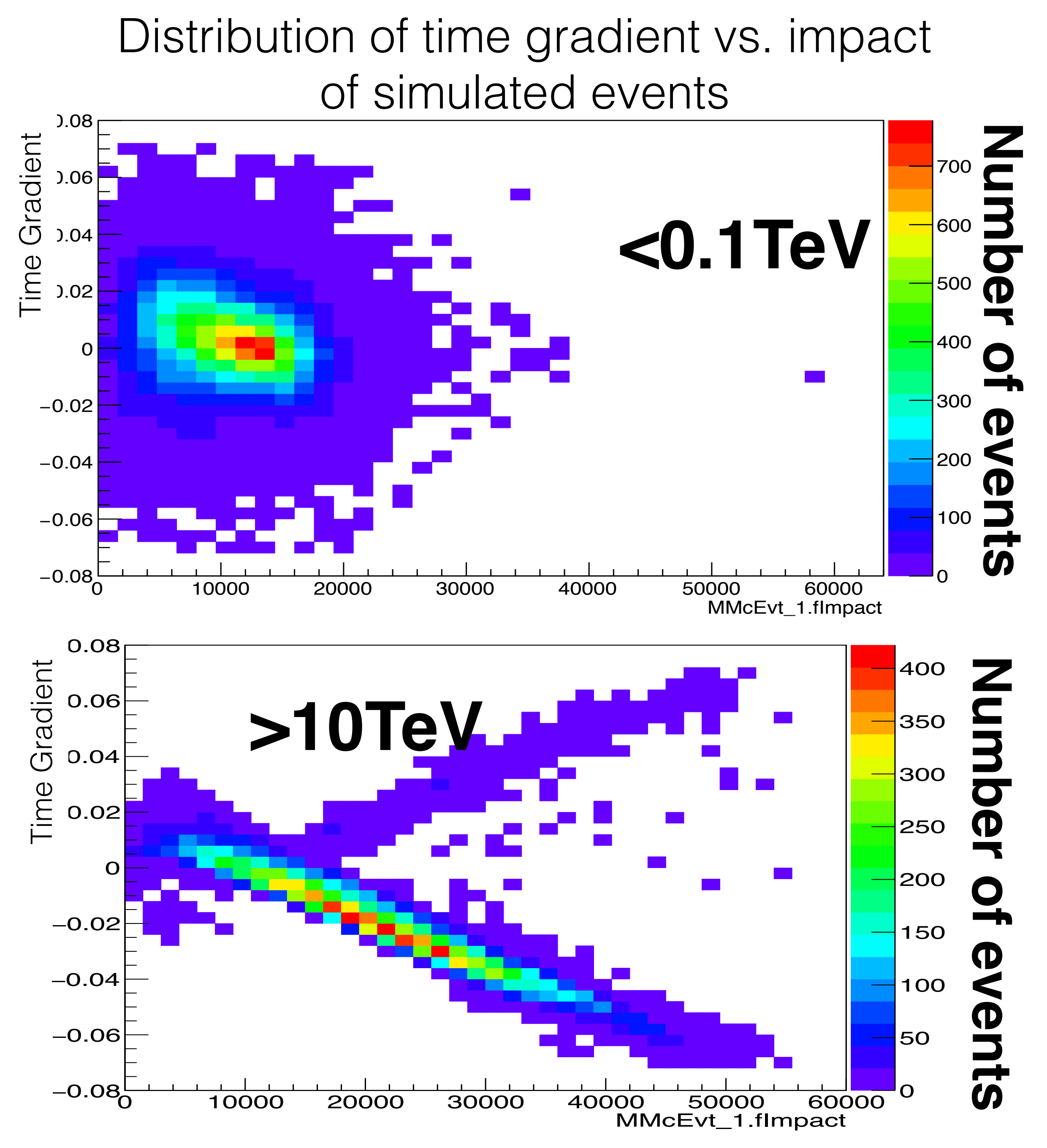
Gradient of arrival times
of phtons along shower axis

For i-th pixel,
yi : the arrival time
xi : the projected position onto major axis
—> fitted by y=ax+b, where a=TimeGradient



The order of arrival time changes depending on Impact

Very strong correlation to Impact,
but weakening in low energy



Distribution of time gradient vs. impact of simulated events

Parameters for Energy Reconstruction(1)

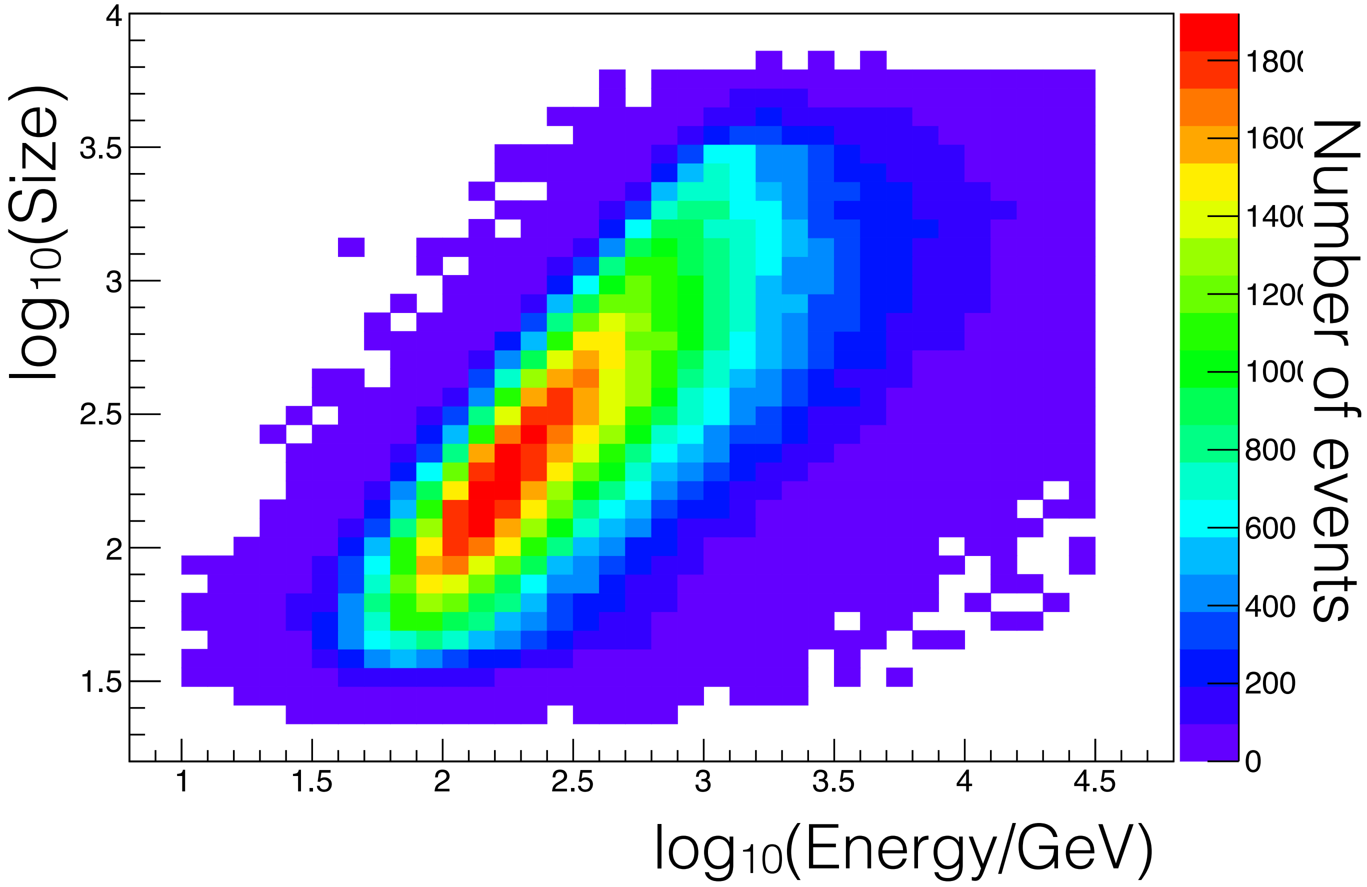
How can we estimate energy of gamma ray?

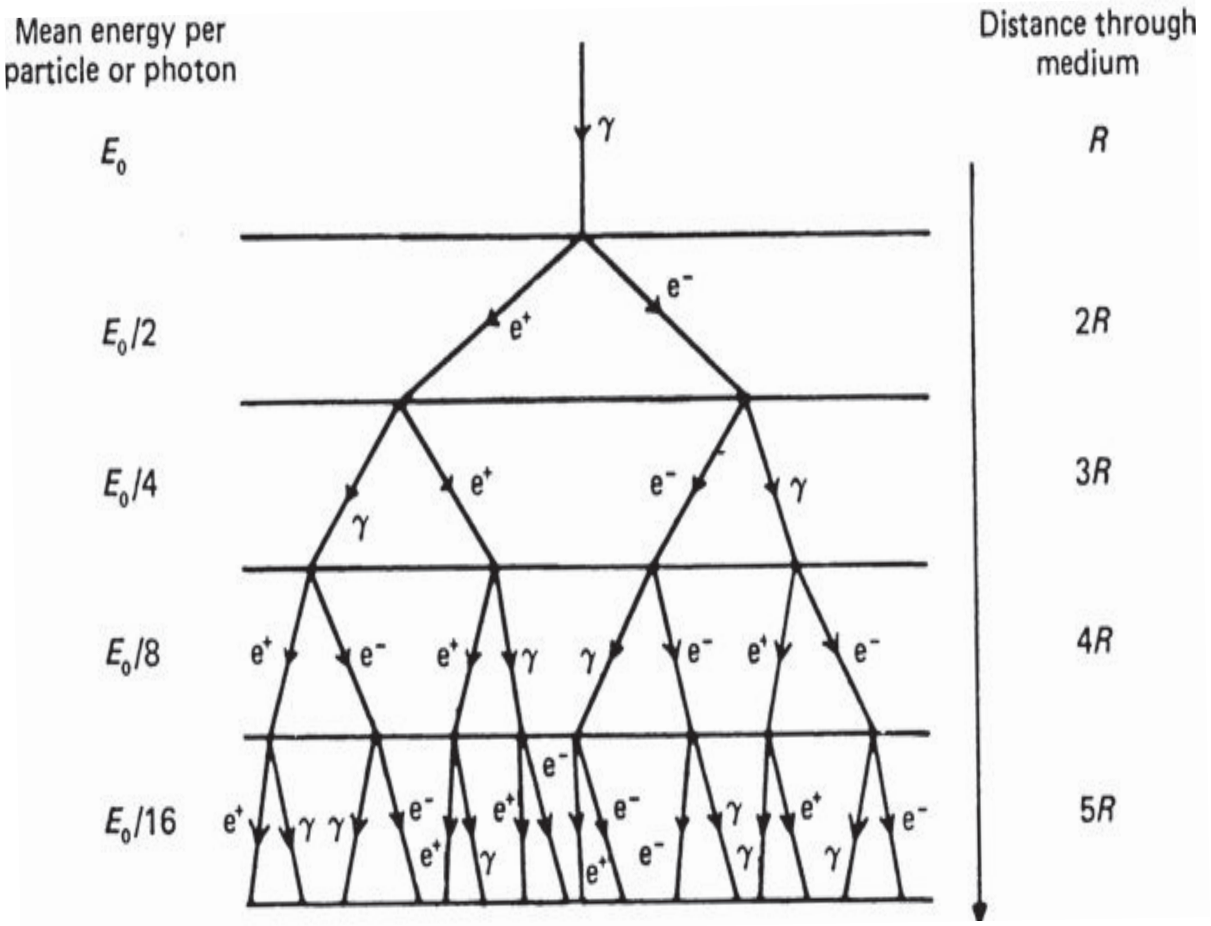
-> Energy ∝ amount of secondary particles ∝ Amount of photons

-> The brighter, the higher the energy!

Size is almost **proportional to energy**

The Size distribution of simulated events



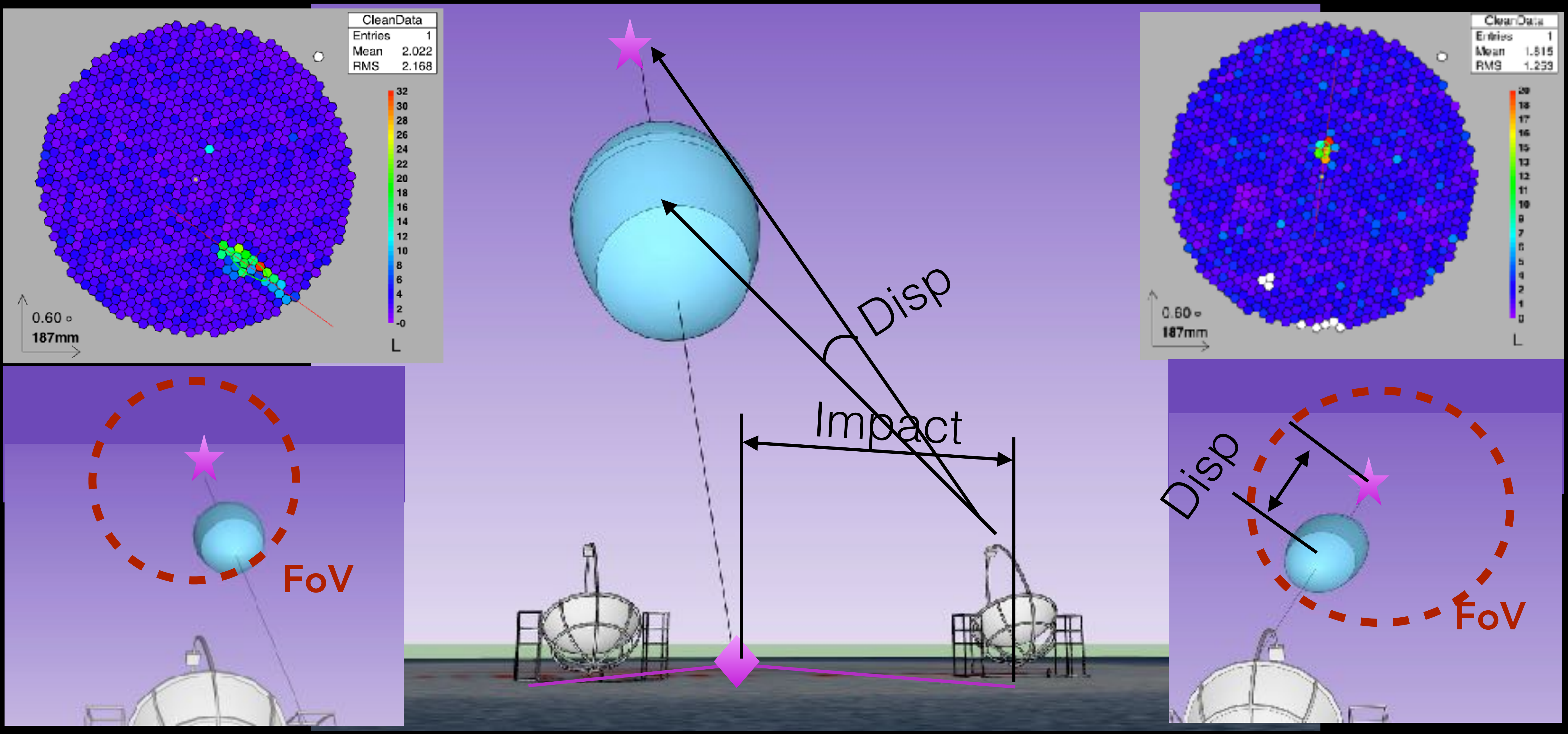


Maximum number of particles is proportional to the original energy

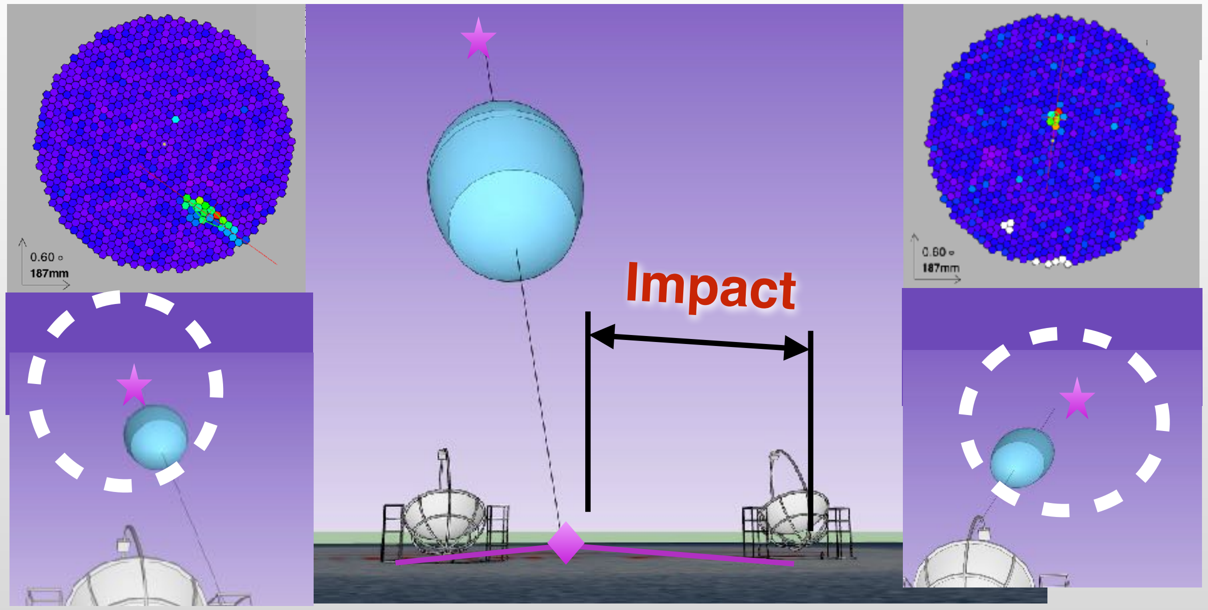
Parameters for Energy Reconstruction(2)

Main factor for the correction:
The closer, the brighter. Thus geometrical corrections are needed.
Offset on the ground and along the axis can be interpreted as **Impact and Disp**

And the other contributions…



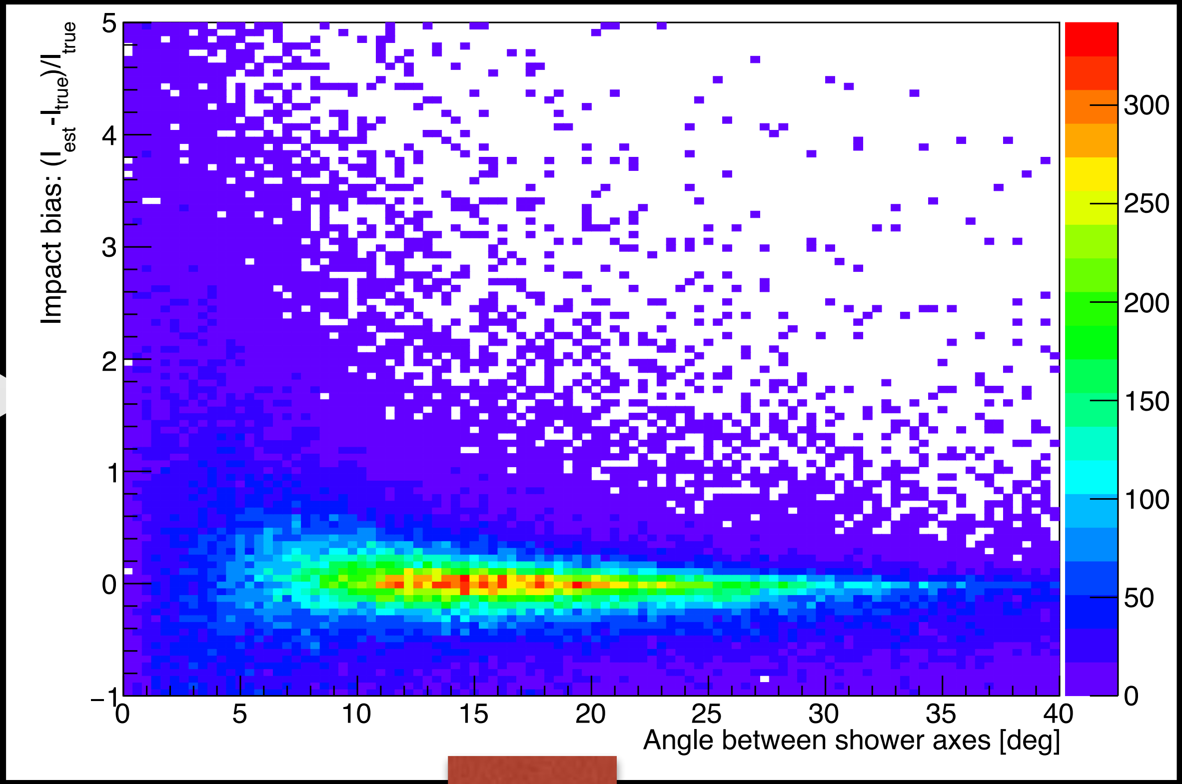
Impact itself should be cleared



Classical method

Intersection of two shower axes projected on the ground

-> **poor reconstruction** especially **when** the two axes meet by **too small angle**

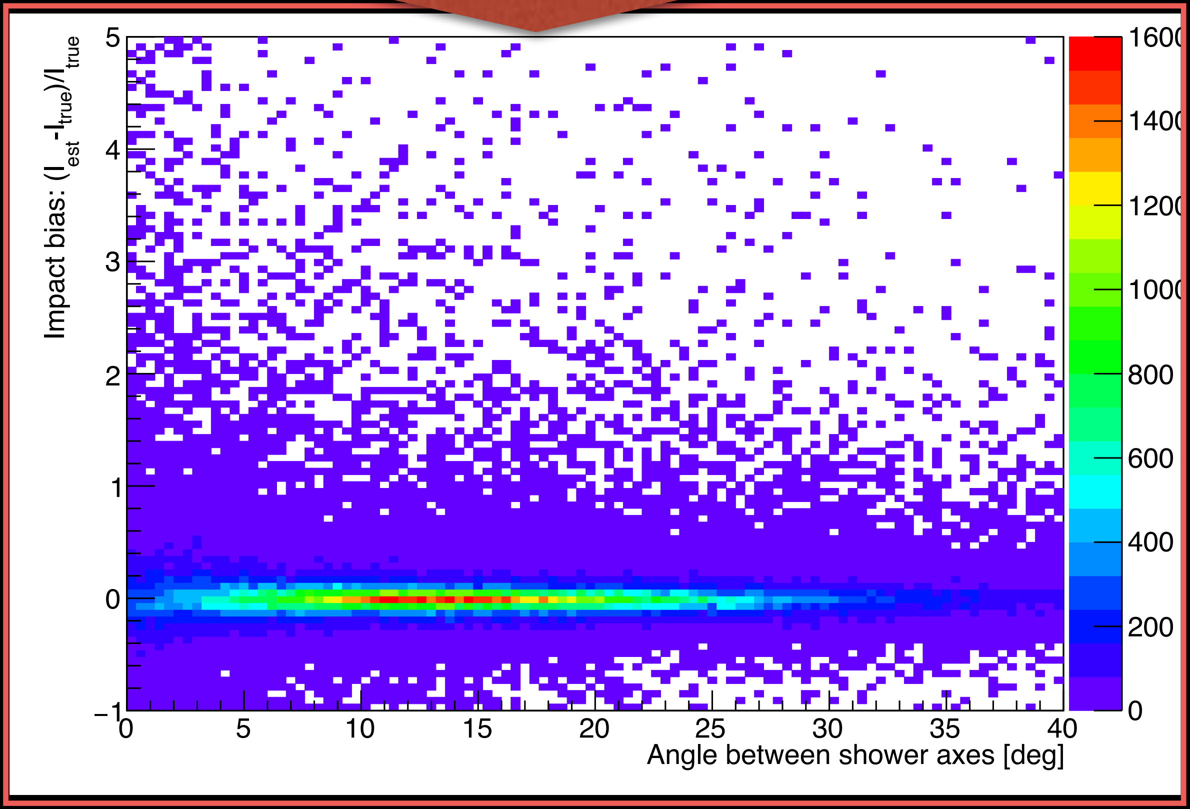


The impact from Disp

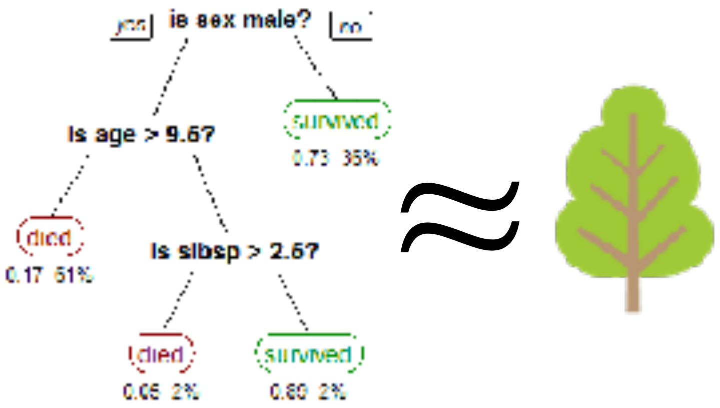
🙂 **much better in low Zd (general) observation**

😑 **huge fluctuation in high Zd**

-> (better to use only time gradient)



Random Forest (RF)



What is Random Forest?

- A sort of “Machine Learning” technique
- It consists of large number of **Decision Trees**
- Randomisation of trees by “**Bagging**”

=bootstrap aggregating

Why Random Forest?

- Robust

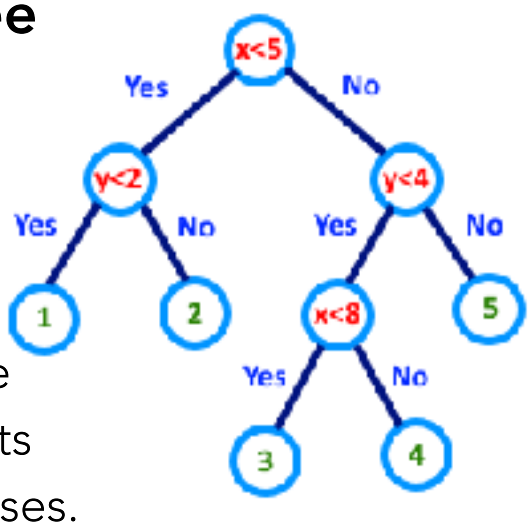
  Less risk to have bias in the estimation.
- Visible

  Larger capability to control parameters.
  Simpler structure and visibility of importance of the parameters
- Light & fast

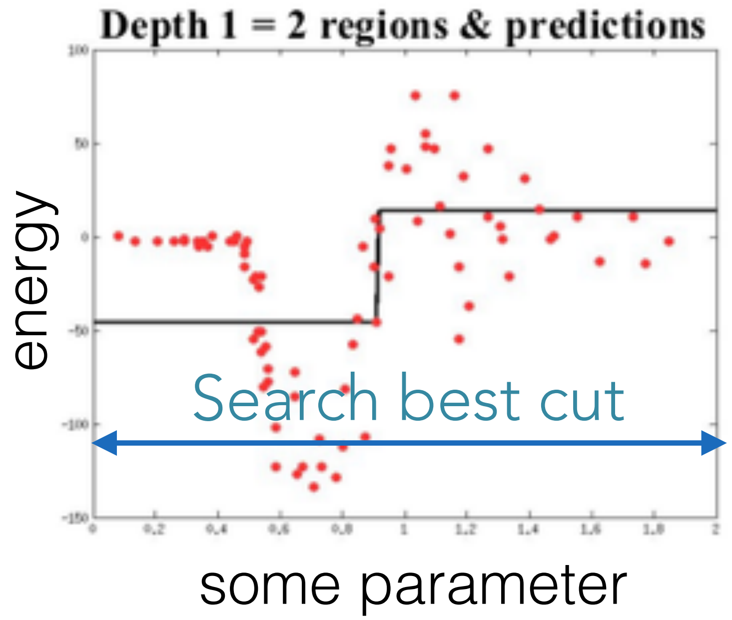
  less computing stress & parallel processing

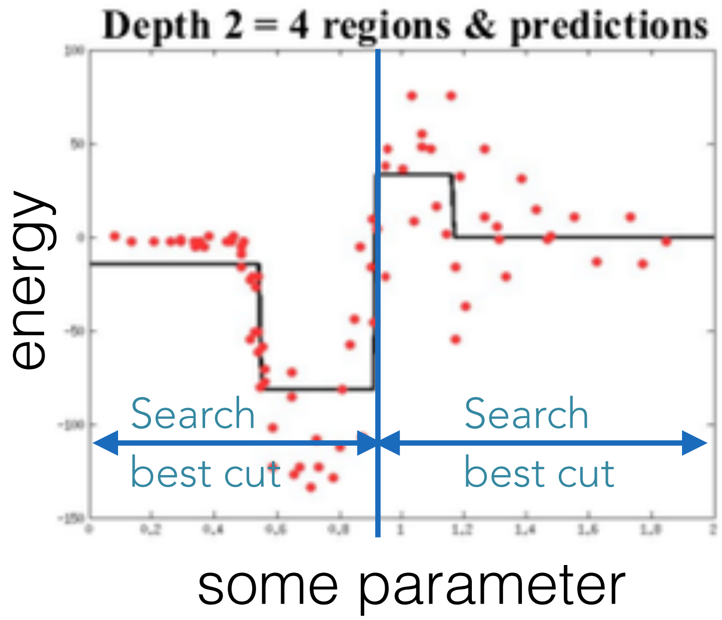
Generating a tree for regression

Decision tree



A decision Tree classifies events by energy classes.





Supervision data is given

Search best cut

The distributions are separated at minimum of the mean variance σ^2.

$$\sigma^2(E) = \frac{1}{N_L + N_R}(N_L\sigma_L^2(E) + N_R\sigma_R^2(E))$$

Side in the separation

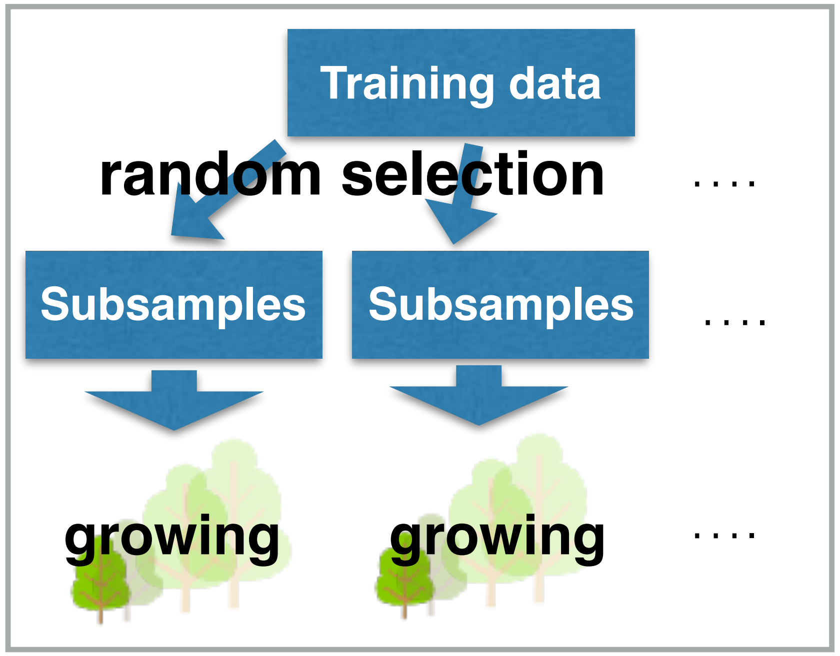
	Left	Right
Number of events	N_L	N_R
Variance	σ_L^2	σ_R^2

The best separation power is searched among different parameters

Estimation value of an energy class

Ei (the energy in class i) is determined as the average of Ni events in final nodes

Using many trees in RF : Bagging



Training data

random selection ....

Subsamples | Subsamples ....

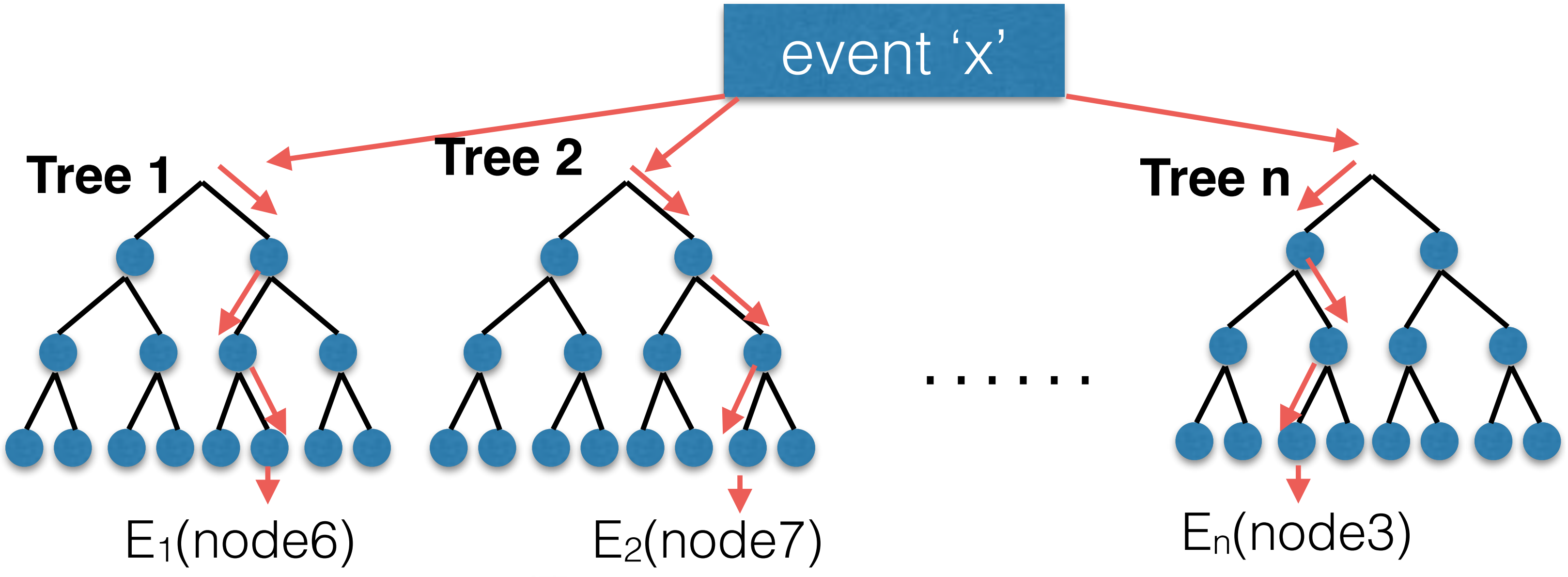
growing | growing ....

Bagging (Bootstrap aggregating)

https://en.wikipedia.org/wiki/Bootstrap_aggregating

A technique to create a set of subsamples with **random selection**

Trees can grow on different data, thus **they will give different answers!**



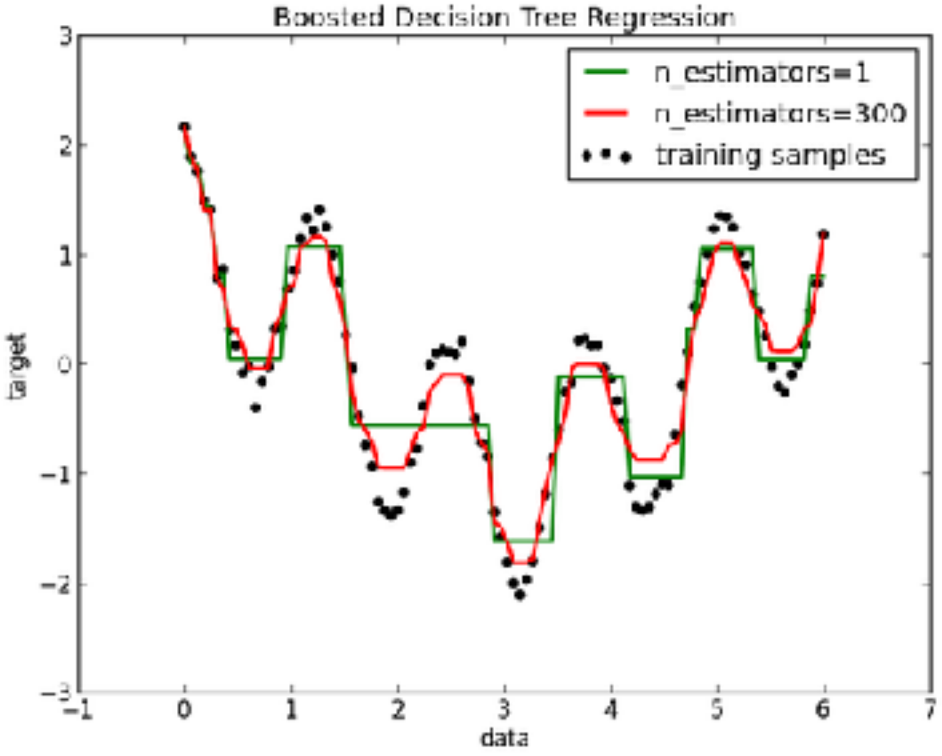
event ‘x’

Tree 1 **Tree 2** **Tree n**

......

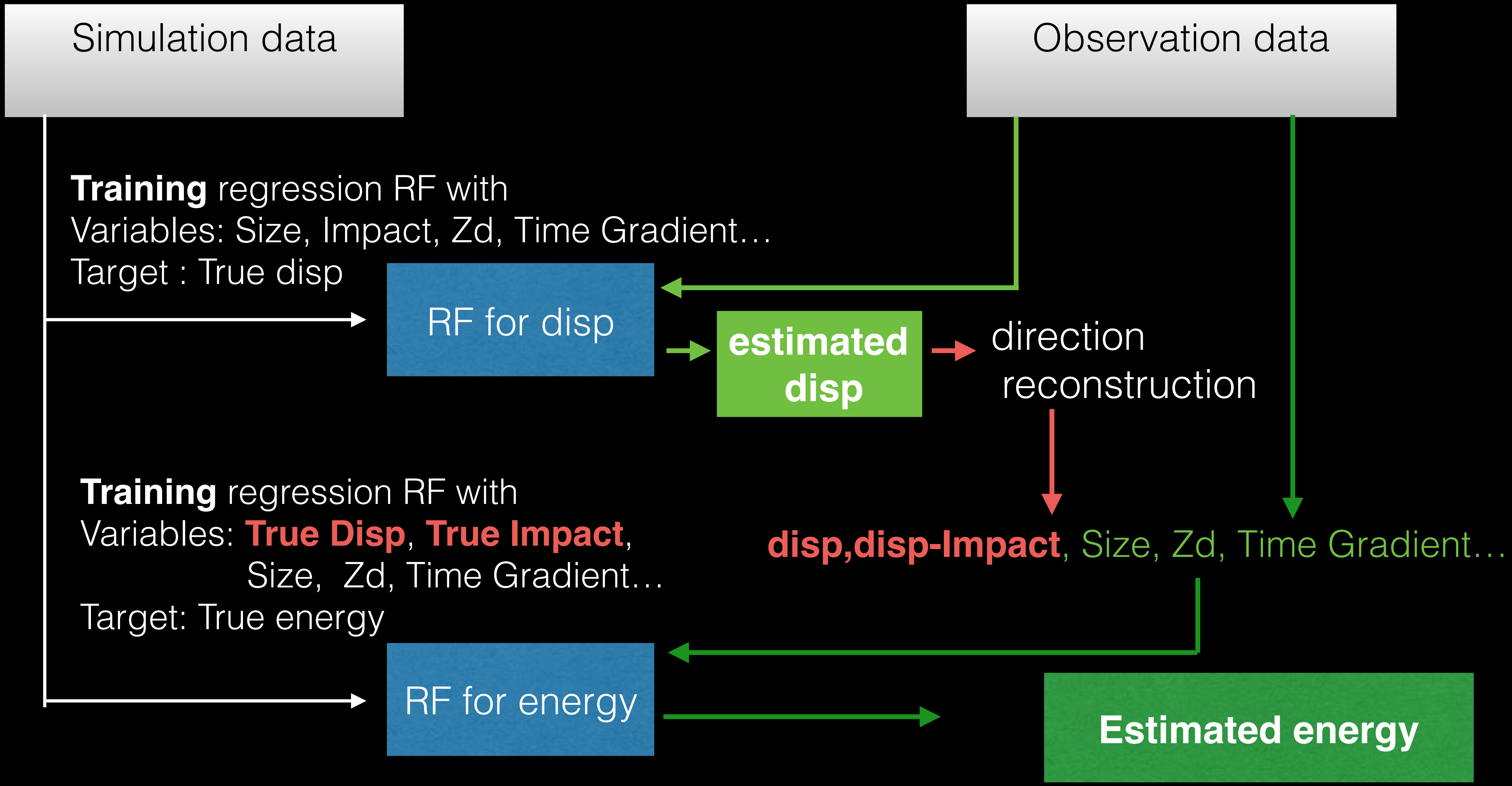
E_1(node6) E_2(node7) E_n(node3)

$$E_{est}(X) = \frac{1}{n}\sum_{j}^{n} E_i \text{ -th last node in j-th tree } (X)$$



Each tree is biased, but the average follows true value well!

Analysis flow



Simulation data

Observation data

Training regression RF with
Variables: Size, Impact, Zd, Time Gradient…
Target : True disp

RF for disp

estimated disp

direction reconstruction

Training regression RF with
Variables: **True Disp**, **True Impact**,
Size, Zd, Time Gradient…
Target: True energy

disp,disp-Impact, Size, Zd, Time Gradient…

RF for energy

Estimated energy

In the training, not only the true energy, but also true Impact and Disp are feeded
=> RF gains additional information!
-> Need to be aware of the possible bias

Implementation in the official package

- Tree configurations

  Configurable dependent on the demand (accuracy vs. limited resource)

  Number of parameters to be fed : 15

  Number of events at the last node : 3 ~ 5 (let them very biased)

  Number of trees : > ~ 100

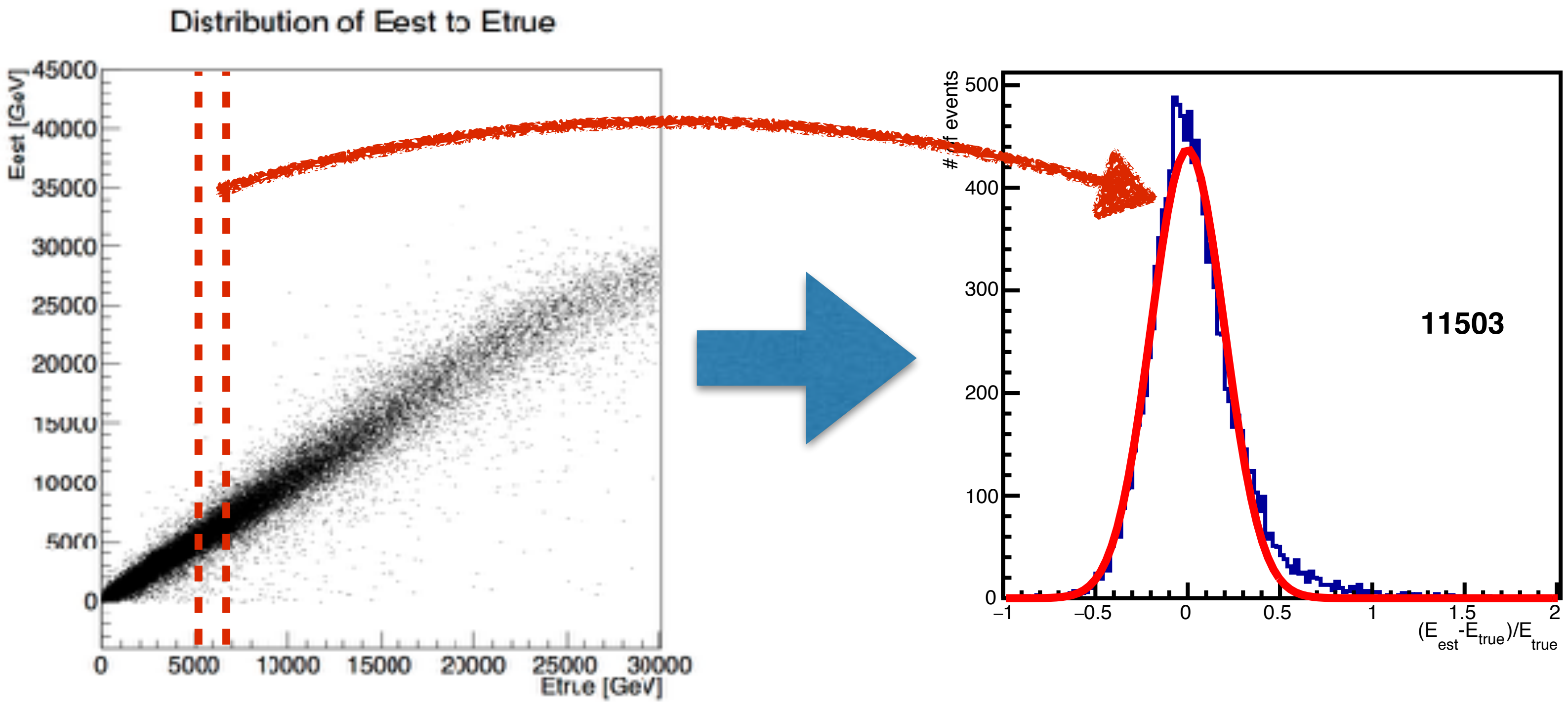
  Number of trials in each branch generation : ~ 4

  Cutcondition on the supervision data: No cut

- Integrated in the official analysis package
  - Reuse of the existing RF classes for disp and background rejection in the c++ official package called “MARS”.
  - Better performance than the former energy estimation functionality using Look-Up Table.

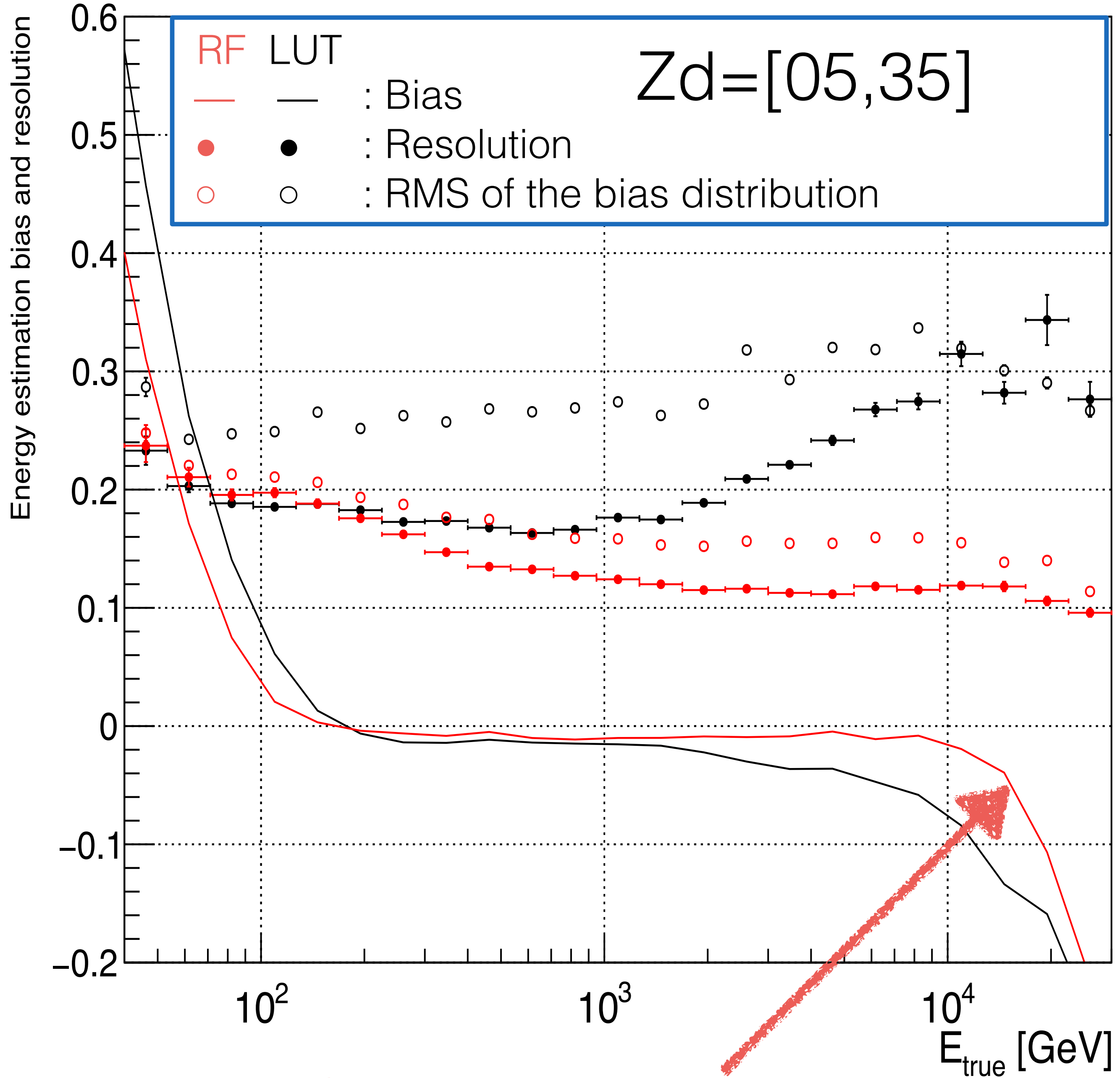
Evaluation

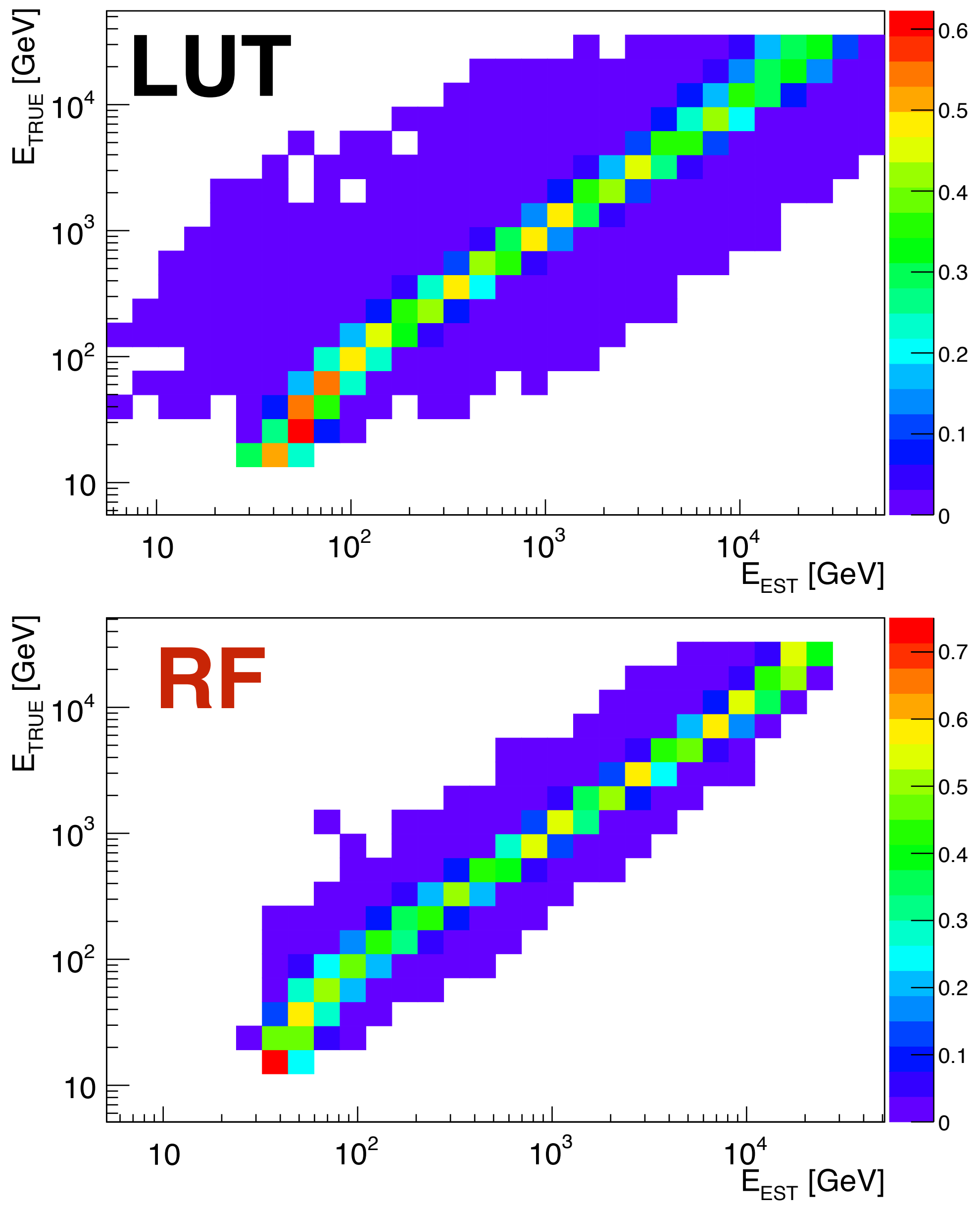
- **The performance comparison** with
  the former official strategy(Look-Up table)
- **Deviation of**
  the **asigned energy(Etrue) from the estimated energy (Eest)**
  in the simulation data
- **Mass tendencies** are evaluated
  from the deviation of distribution (Eest - Etrue)/Etrue,
  - Gaussian fit -> Mean of the distribution := **bias**
  - Gaussian fit -> Width of the distribution := **resolution**
  - **RMS** of the distribution



Significant Performance Improvement

The energy resolution: from ~ 20% (below 100 GeV),down to ~ 12%(above 1 TeV).
The improvement marks more than 50 % reduction of resolution above 10 TeV.

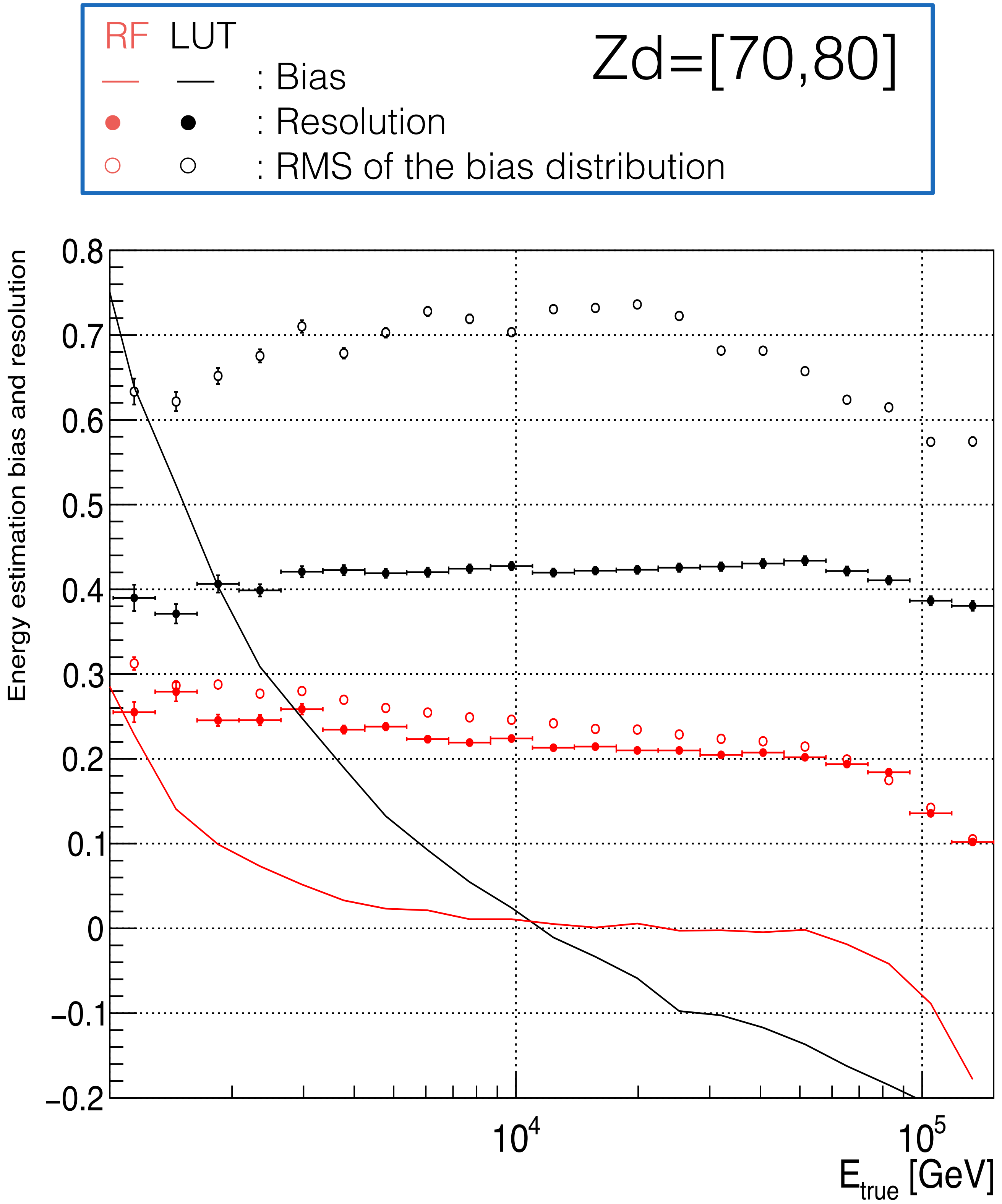




Due to lack of the events beyond simulated energy !

(energy range of MC = [10GeV, 30TeV])

Definitive Performance Improvement



In case of
high Zd observations,
it becomes obvious

What we gain from the improvement

- Significant improvement can give us
  - More accurate spectrum within limited statistics
    e.g. better estimation in power law index for entire spectrum
  - Better sensitivity to the structure in the spectrum
    Cutoff, spectral break and additional components
  - Lower energy threshold

  —> Now I am quantifying those effects

- With the accuracy we will be able to evaluate
  the systematics in the simulation can be evaluated
  (e.g. the atmosphere)

Summary

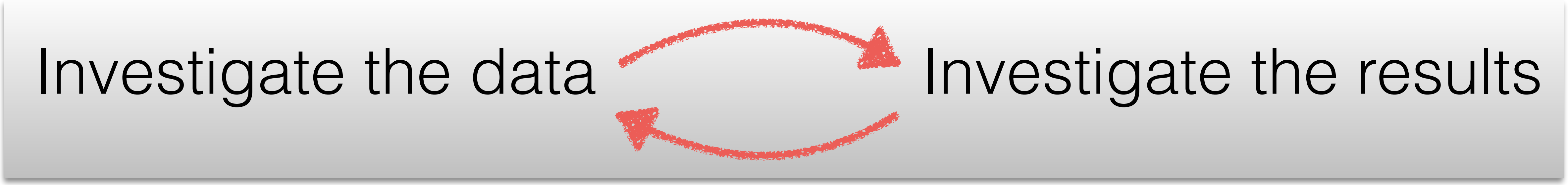
- **An accurate reconstruction of the energy** of the incoming gamma ray boosts **discovery and scientific potential of the data**
  - Variability of spectrum indicates source dynamics.
  - Additional information will be obtained from additional structure information in the spectrum.
- **Significant improvement by the new energy estimation**

  The energy resolution: from ~ 20% (below 100 GeV),down to ~ 12%(above 1 TeV).

  The improvement marks more than 50 % reduction of resolution above 10 TeV.

  It yields an improvement in the gamma-ray capabilities of the MAGIC telescopes
  - More accurate spectrum within limited statistics
  - Better sensitivity to the structure in the spectrum
  - Lower energy threshold

  -> Evaluation ongoing (and sanity checks too)
- **I implemented the new energy estimation in standard MAGIC analysis software** and it is widely used as the official strategy.



Investigate the data

Investigate the results

Thank you !

Essential point of Bagging

The steps to create subsamples with N_{select} events from N_{tot} events of training data:

- Choose numbers randomly from $[1, N_{tot}]$ and list them.
- Repeat for N_{repeat} times (Listing may duplicate)
- Select the events with the event number listed (NOTE that $N_{repeat} \geqq N_{select}$)

Example:

If you throw a dice for 6 times, you will see just ~4 numbers

If $N_{repeat} = N_{tot}$ and N_{tot} is large,

N_{select}/N_{tot} becomes $1 - \frac{1}{e}$

→ Subsample size is always ~63 % of total RANDOMLY chosen