Laboratoire d'Annecy-le-Vieux de Physique des Particules

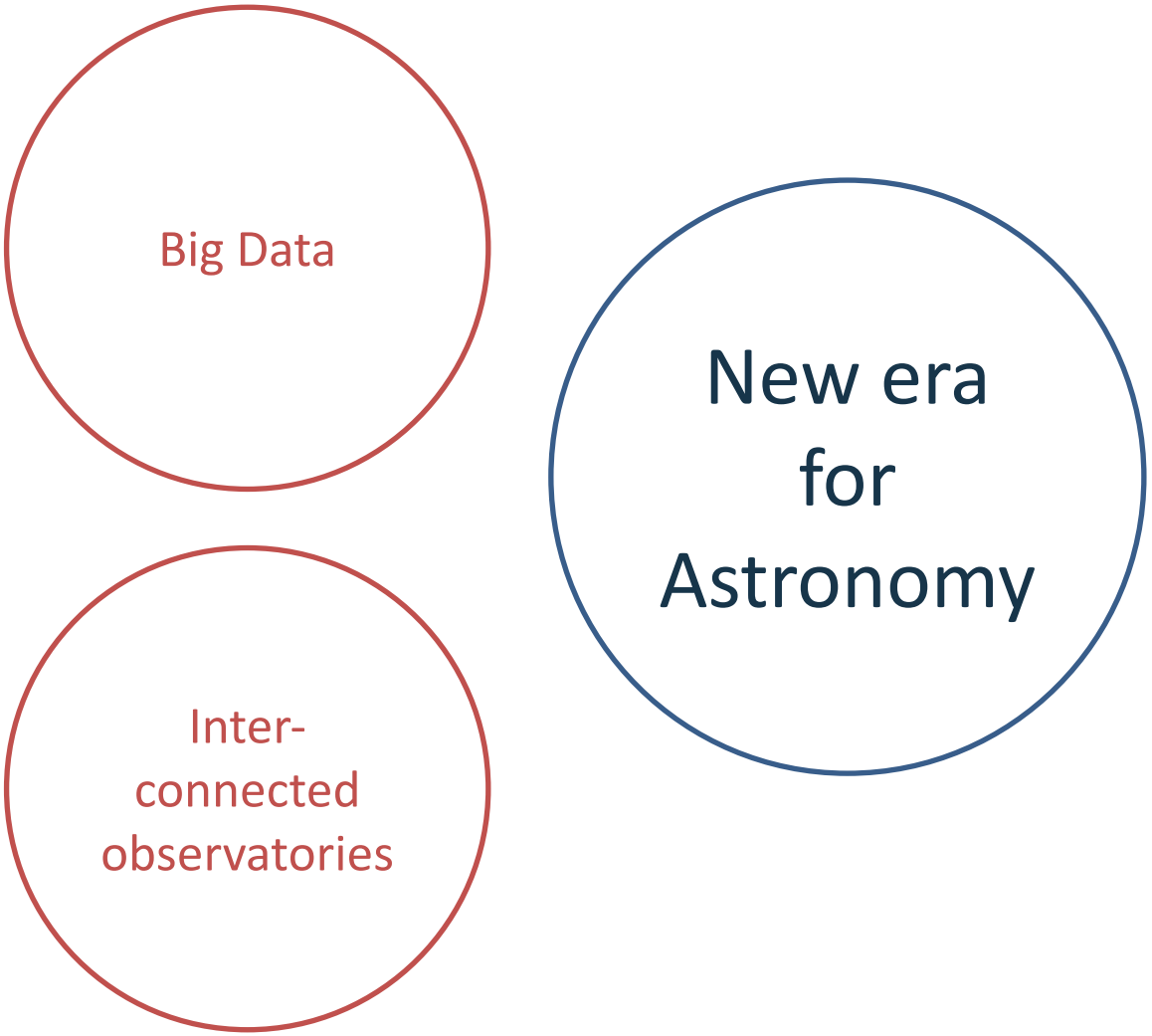
When High Performance Computing meets Astronomy - a concrete case.

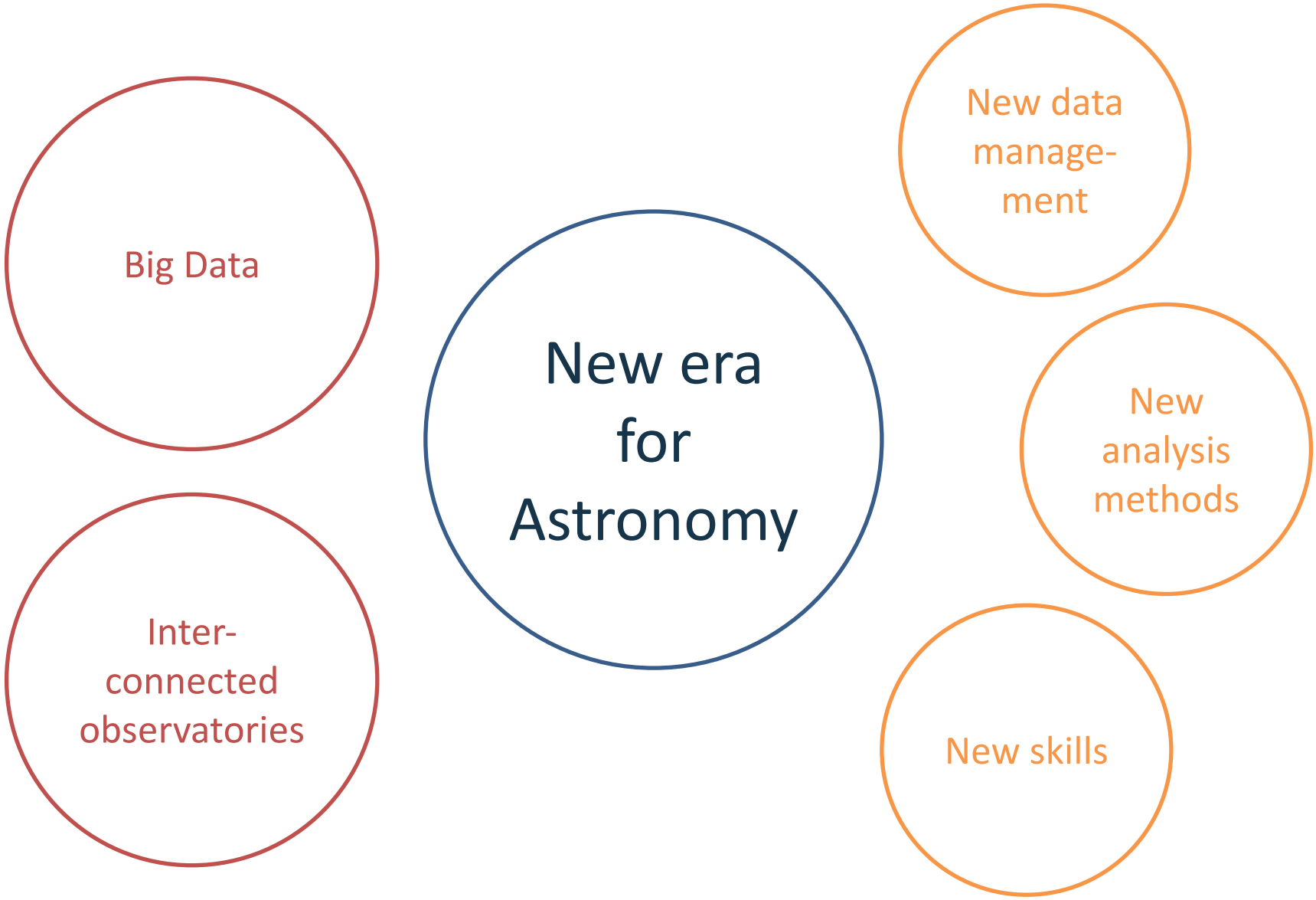
Thomas Vuillaume & Pierre Aubert,
on behalf of Gilles Maurin, Florian Gaté, Jean Jacquemier & Giovanni Lamanna

HEP software foundation workshop, LAPP, June 2017

Astronomy ESFRI & Research Infrastructure Cluster

UNIVERSITÉ SAVOIE MONT BLANC


Big Data
Inter-
connected
observatories
New era
for
Astronomy


Big Data
Inter-connected observatories
New era for Astronomy
New data manage-ment
New analysis methods
New skills

Astronomy ESFRI & Research Infrastructure Cluster

- ASTERICS – a European Horizon 2020 project
  - Joining forces, knowledge, developments and ressources in astronomy and astroparticle
  - Help ESFRI telescopes infrastructures operate together

- Astronomy (desperately) needs Data Scientists and HPC/HTC skills
  - Train physicists: e.g. 1st ASTERICS international school* @LAPP in June

  - Hire (more) computer scientists **IN** the physics teams
  - Both approaches are necessary

*https://indico.in2p3.fr/event/14227/

- Astronomy (desperately) needs Data Scientists and HPC/HTC skills
  - Train physicists: e.g. 1st ASTERICS international school* @LAPP in June

  - Hire (more) computer scientists **IN** the physics teams
  - Both approaches are necessary

- Requires **collaboration** (even for the talks…)

*https://indico.in2p3.fr/event/14227/

Today's menu

- The Cherenkov Telescope Array
  - How it works
  - Some physics
  - The data management challenges
- How to solve those challenges (spoiler alert)
  - New compression algorithm
  - HPC applied to Hillas parameters calculation
  - New approach for the reconstruction

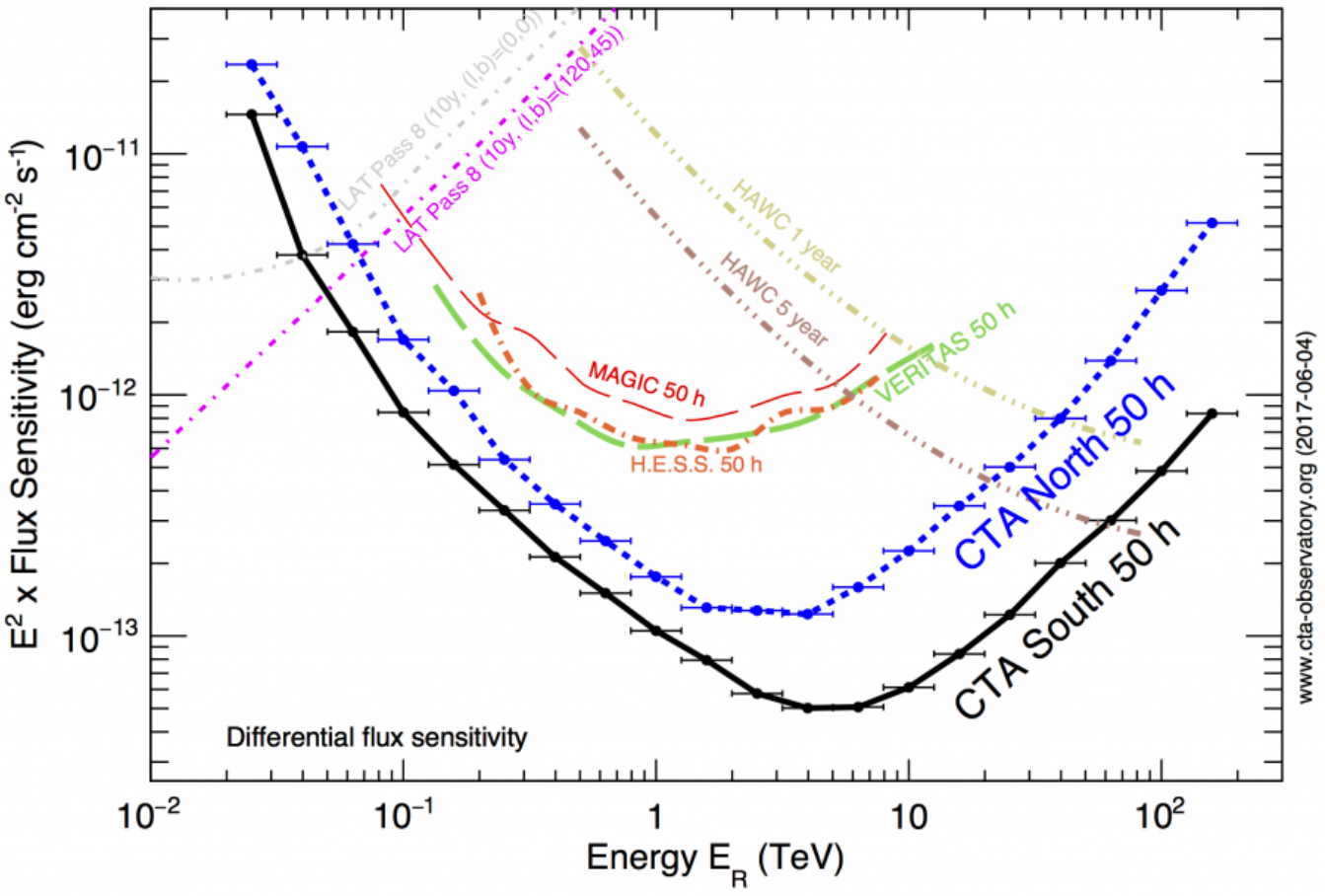
The Cherenkov Telescope Array (CTA)

- Future ground-based telescope for gamma-ray astronomy (Imaging Atmospheric Cherenkov Telescope)
- Observatory
- International collaboration: > 1300 scientists and engineers in 32 countries
- Study the HE universe: cosmic ray origins, HE astrophysical phenomena, fundamental physics and cosmology
- Pre-construction phase. Production phase in 2019

Location

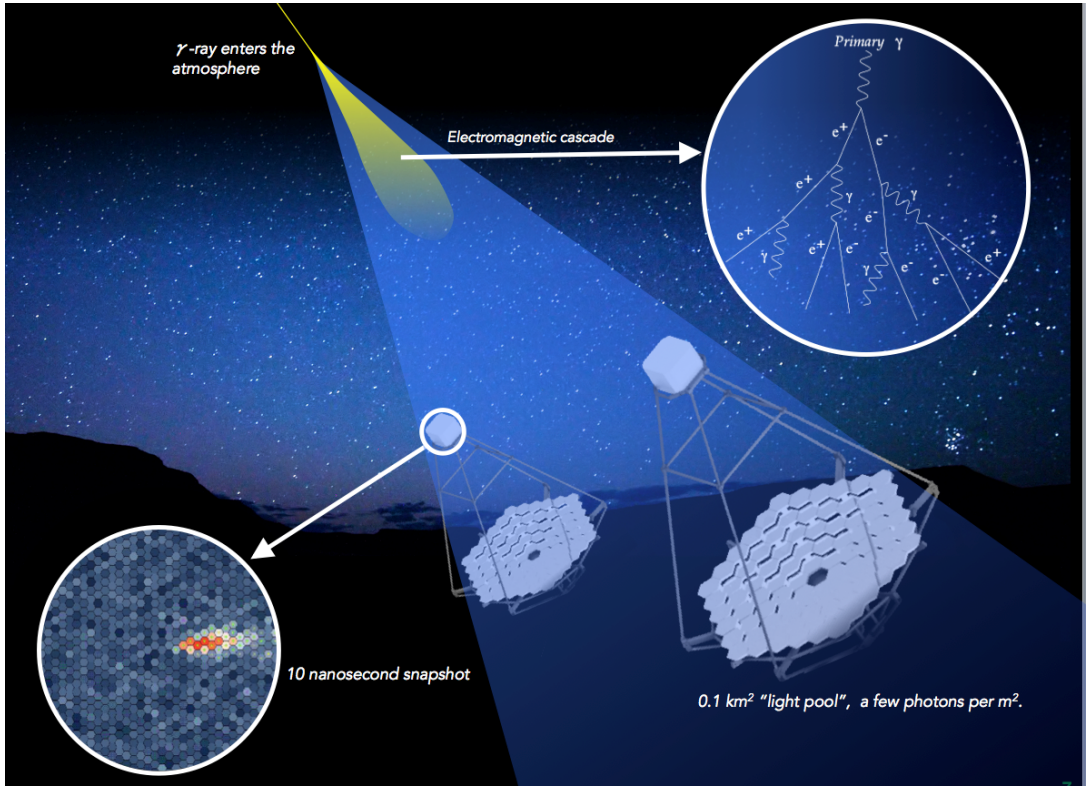
Zeuthen
DESY
Spain
La Palma
Bologna
INAF
northern hemisphere
southern hemisphere
Chile
Paranal
La Palma Map: ©2015 Google, Inst. Geogr. Nacional
Paranal Map: ©2015 Google
Array Sites
Headquarters
Science Data Management Centre

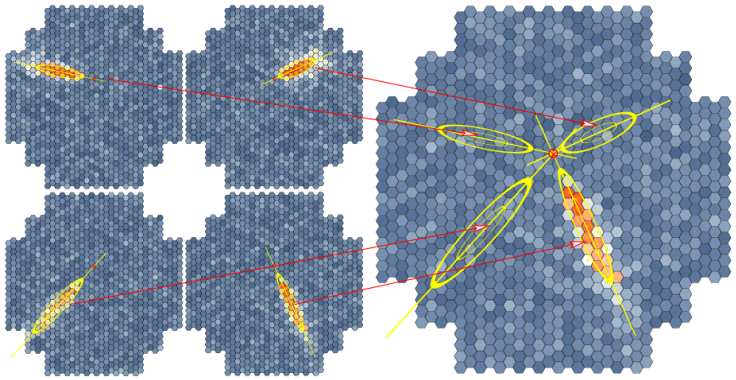
Sensitivity compared to other instruments



Lower is better

How does it work?





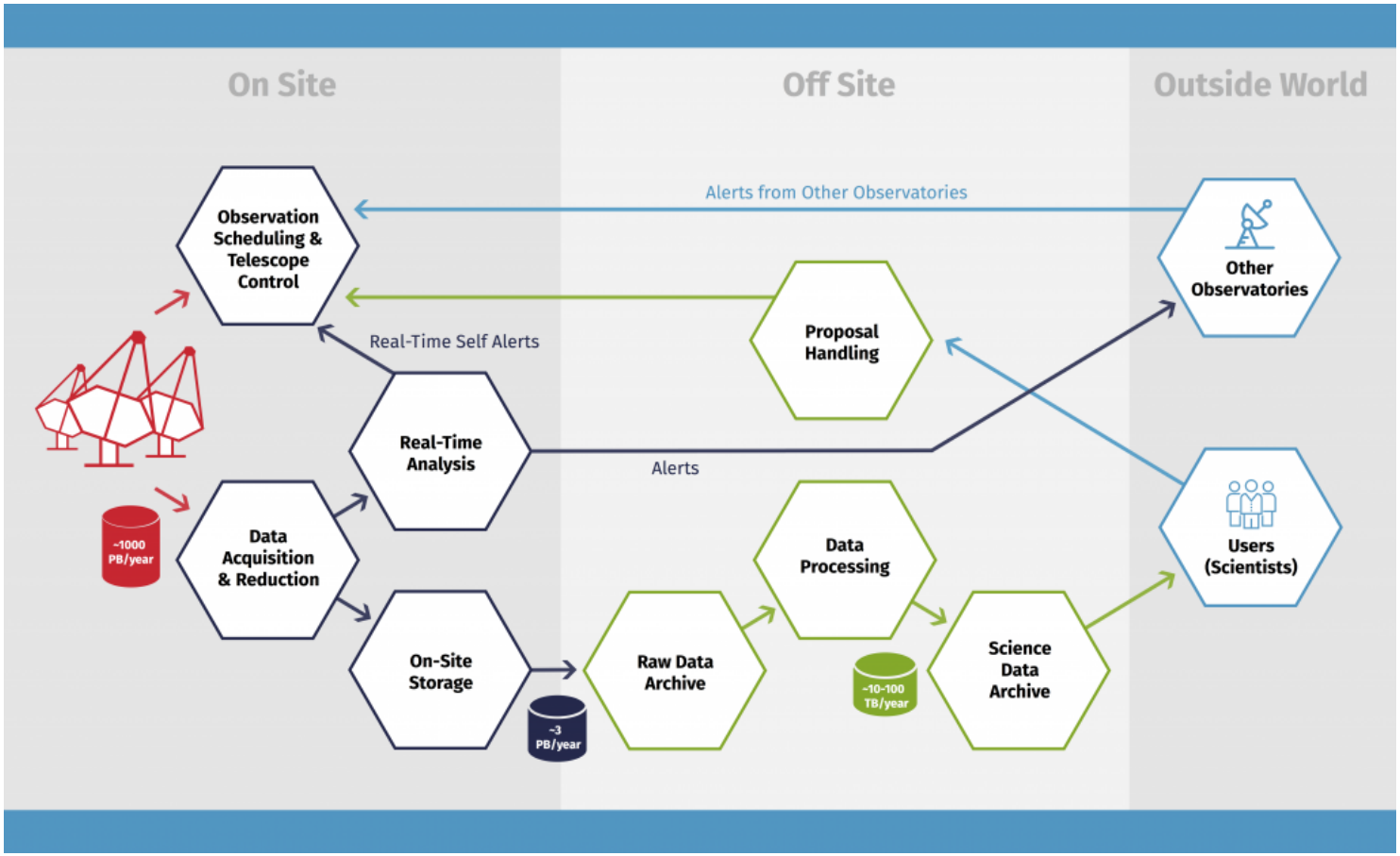
HESS figure (credit K. Bernlohr)

Stereoscopy reconstruction:

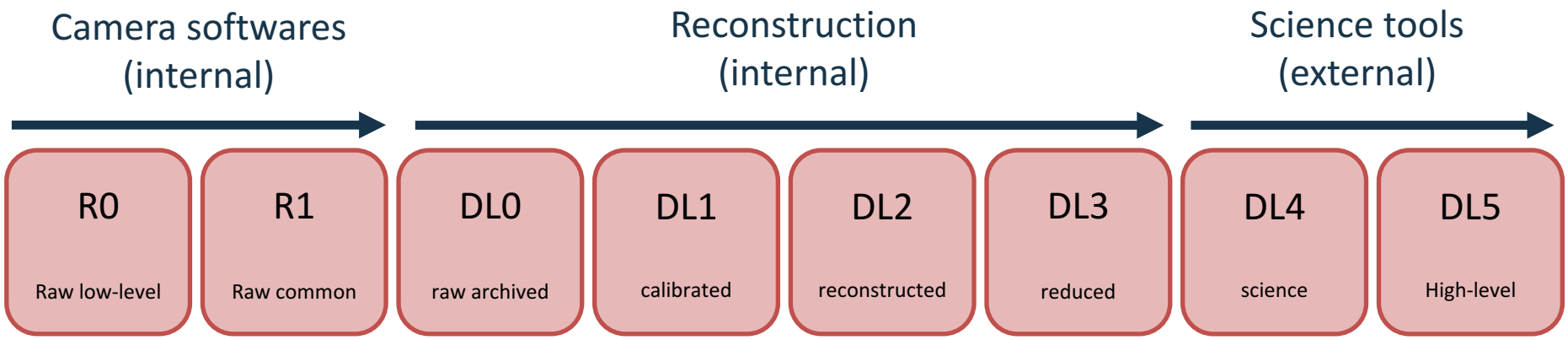
- Hillas parameters = ellipse + higher order parameters
- Geometric reconstruction for incident direction
- Machine learning for energy reconstruction and event discrimination (particle type)

Fancier reconstructions are possible !

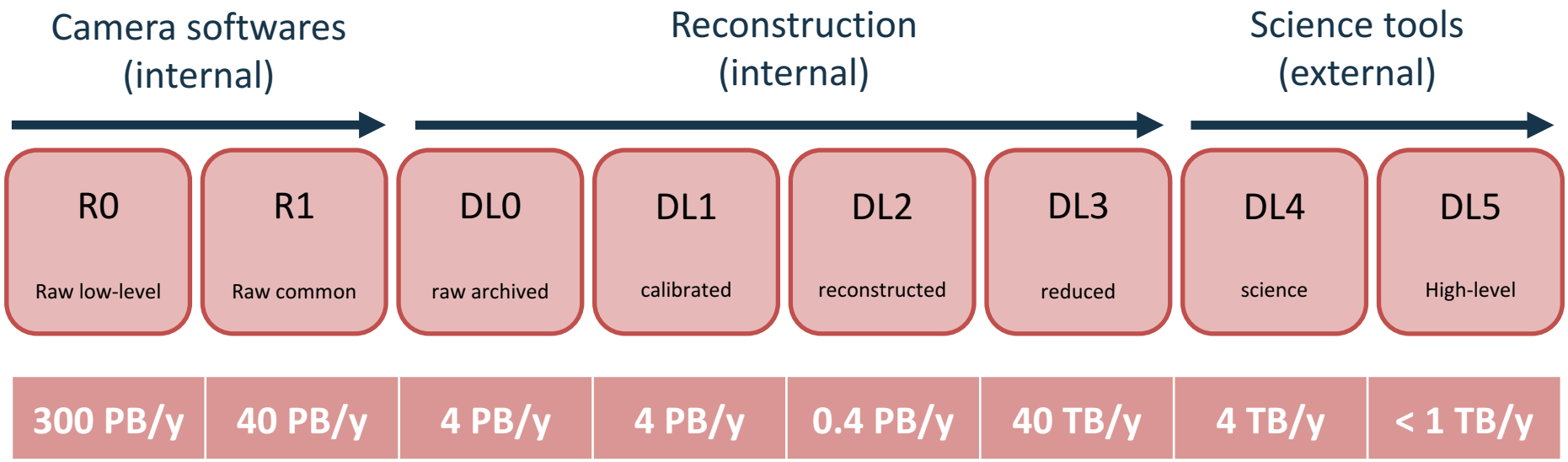
Data management



Computing model

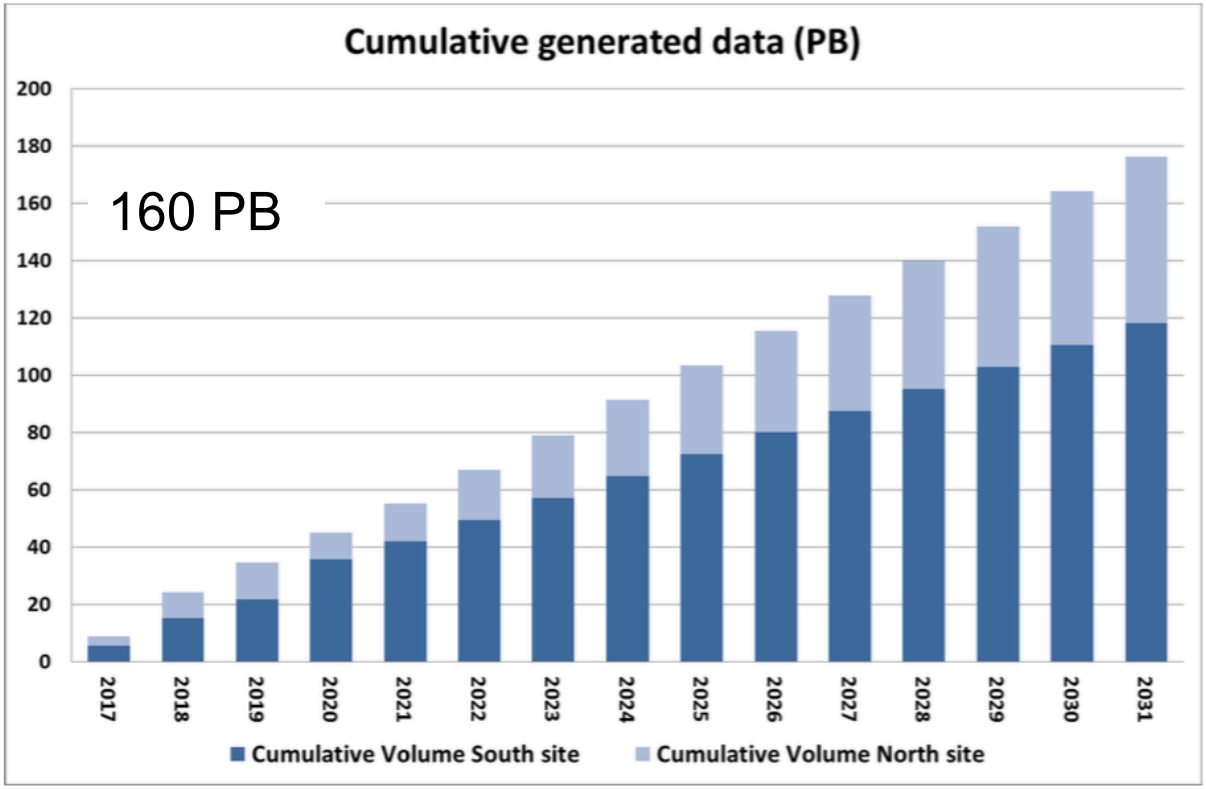


Computing model



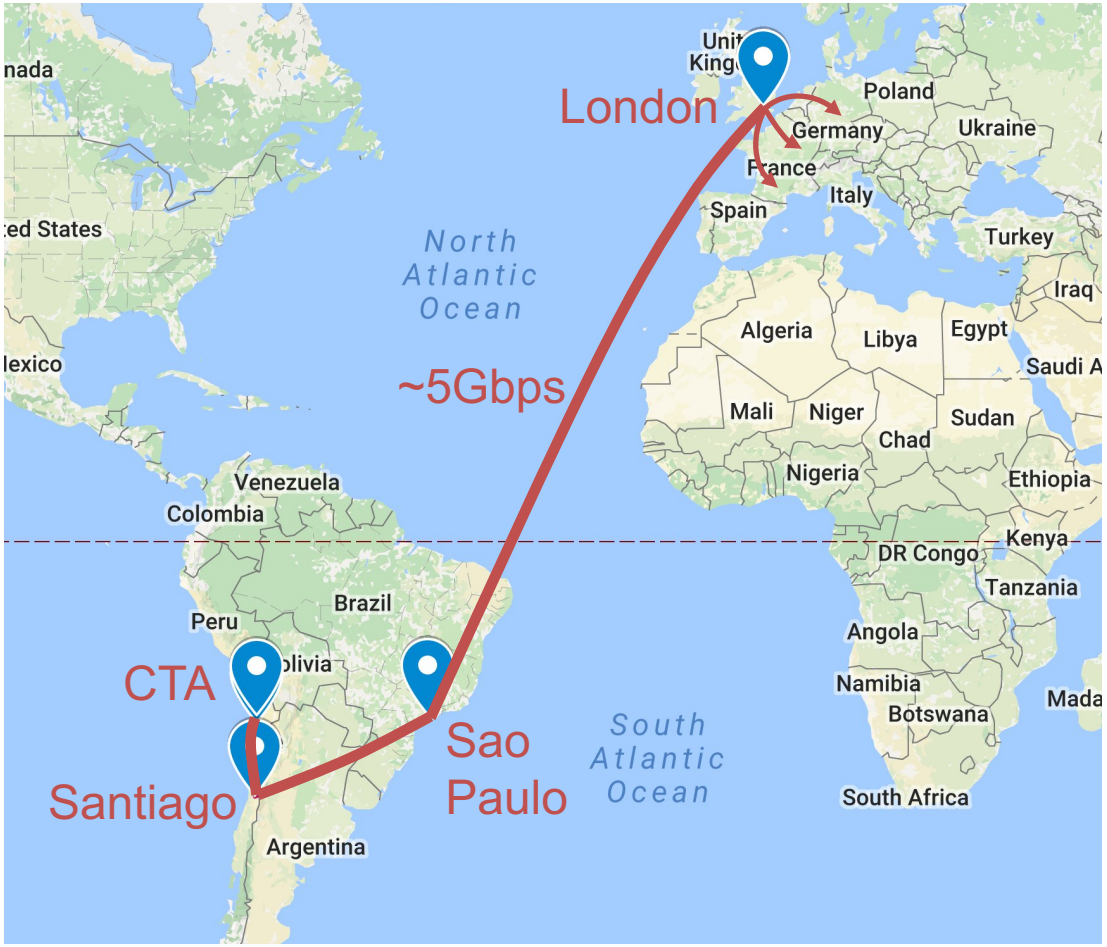
Challenges: data volumes

- Generated data:
  - Raw = 4PB/yr
  - Processed = 4PB/yr
  - Simu (MC) = 20PB
  - Technical…
  - = 12PB/yr
- Total required storage:
  - 27PB/yr (including replicas)
  - 6PB on disk, 21PB on tapes
- Data archived for 30 years of operation + 10 years



Challenges: data transfer

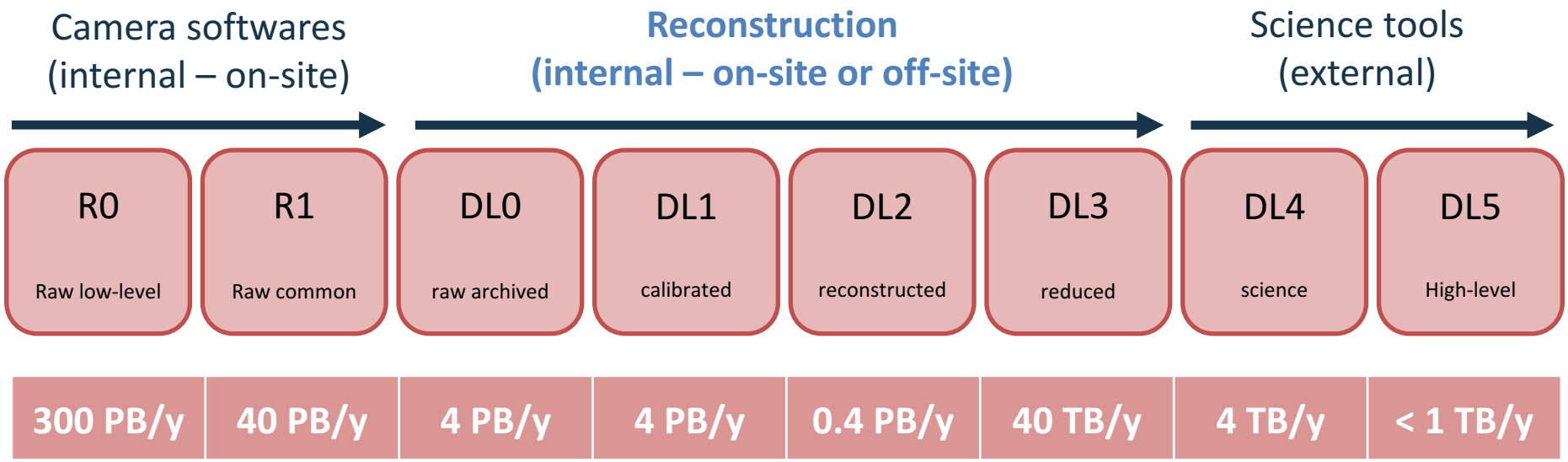
- Requirement:
  Transfer of all acquired raw data from the site to the off-site infrastructure must have a latency of < 10 days.



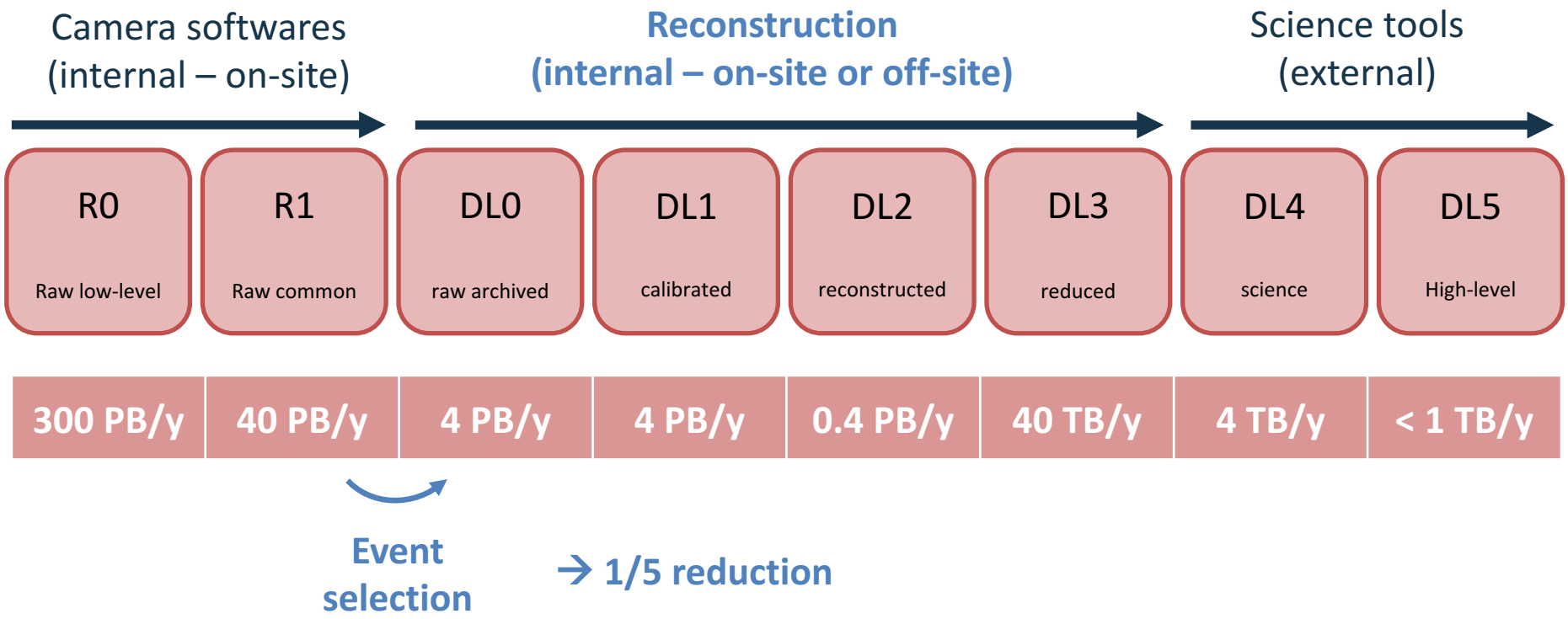
Challenges: data analysis

- Requirements:
  - *Real-time analysis - latency of less than 30s between the acquisition of an event and the generation and issuing of a science alert*
    - Peak data rate = 28GB/s
  - *On-site analysis must be run on all Events within 10 hours of the end of observations for a given night*
  - *Off-line analysis: one day of raw data acquisition must be processed in less than one day*
  - *Off-line analysis: it must be possible to reprocess 1 year of CTA data within 1 month*
    - *1.5GB/s*
  - *Complete reprocessing of all data every year*

Where we bring HPC solutions



Where we bring HPC solutions

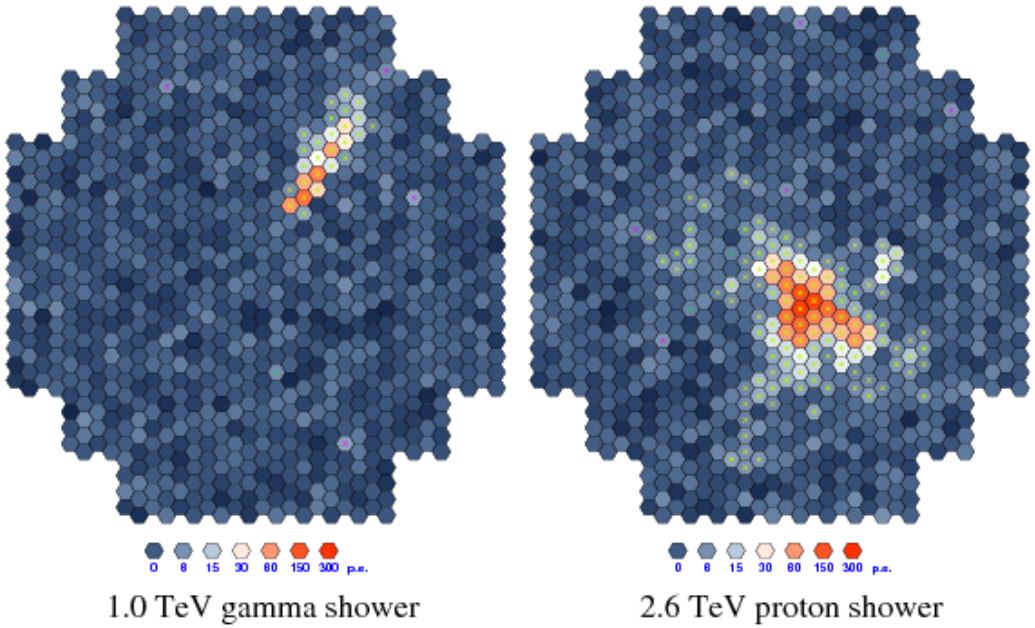


Event selection

Project starting in September with industrial and academic partners

Aim: Deep learning in real-time for data discrimination

Photon induced signal



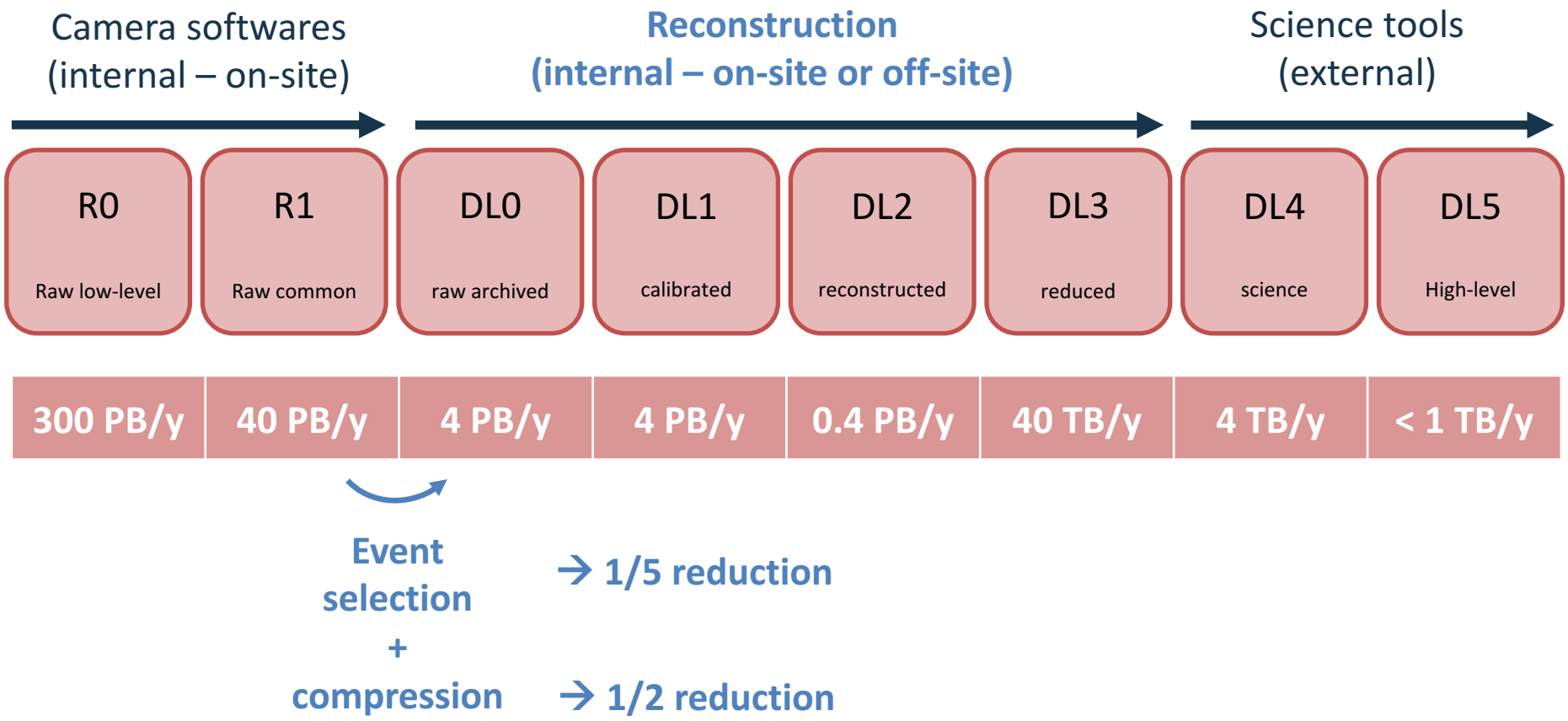
1.0 TeV gamma shower

2.6 TeV proton shower

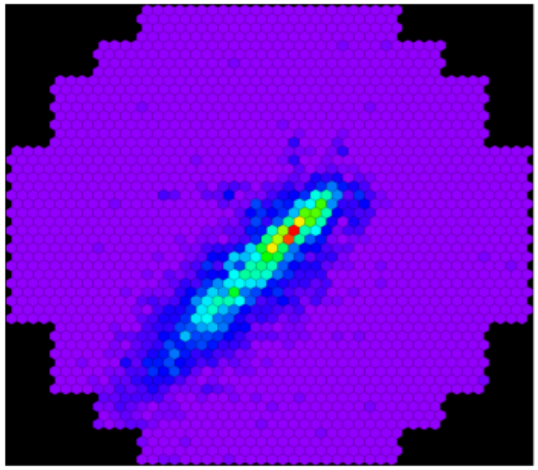
Proton induced signal

Credit: H.E.S.S. collaboration

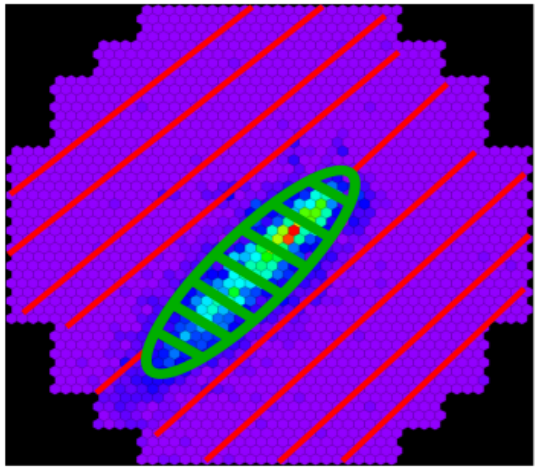
Where we bring HPC solutions



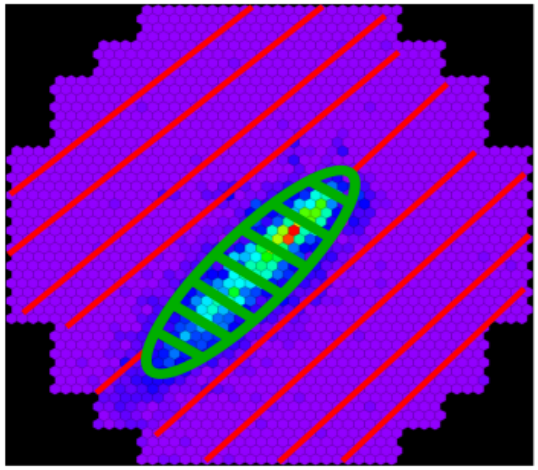
Lossless Data Compression

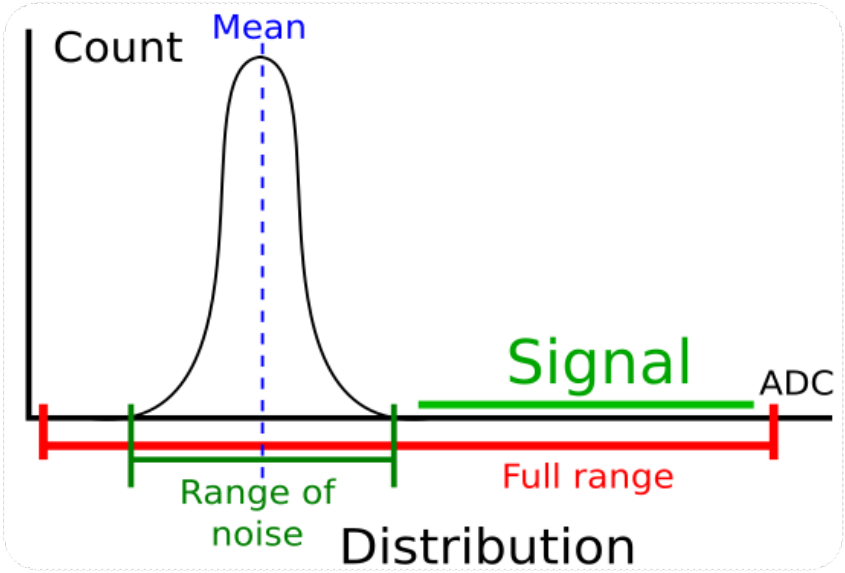


Lossless Data Compression

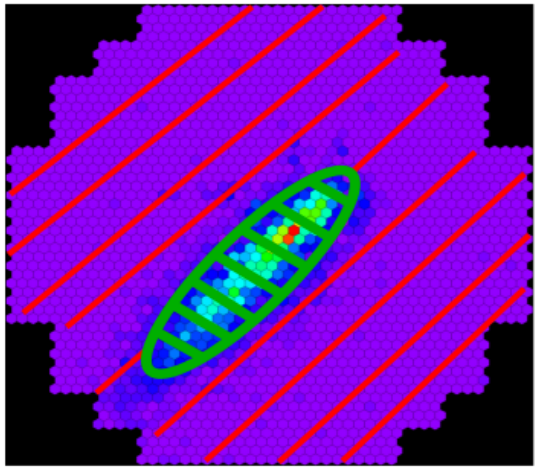


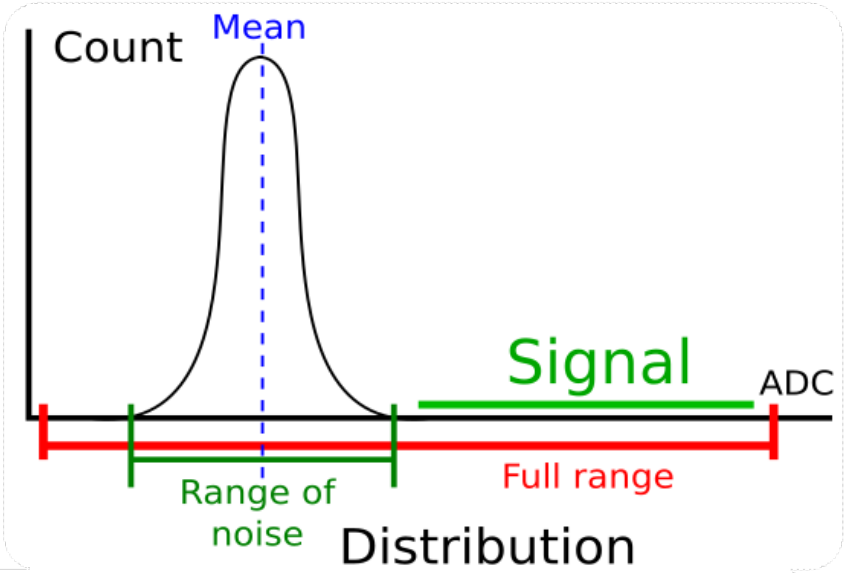
Lossless Data Compression

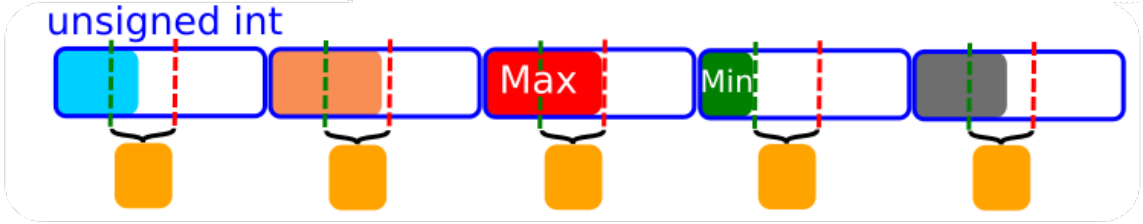




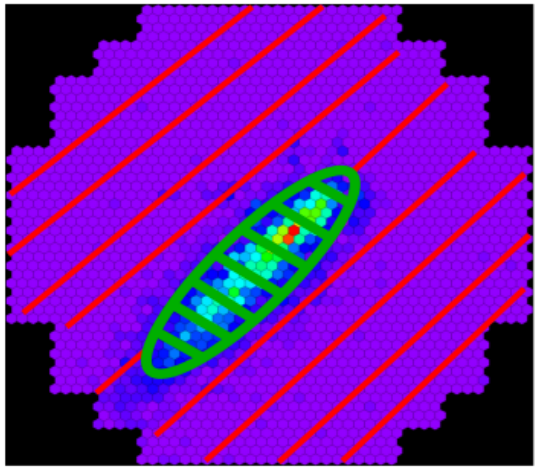
Lossless Data Compression

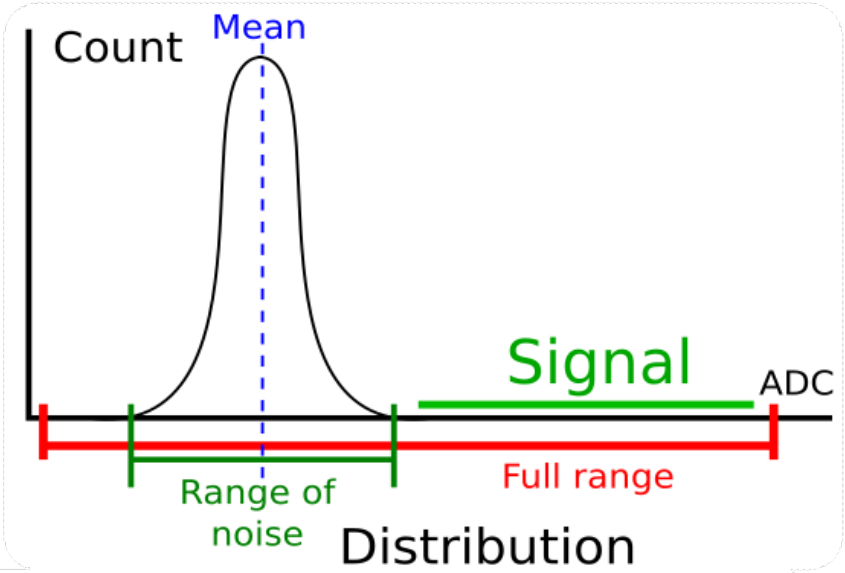


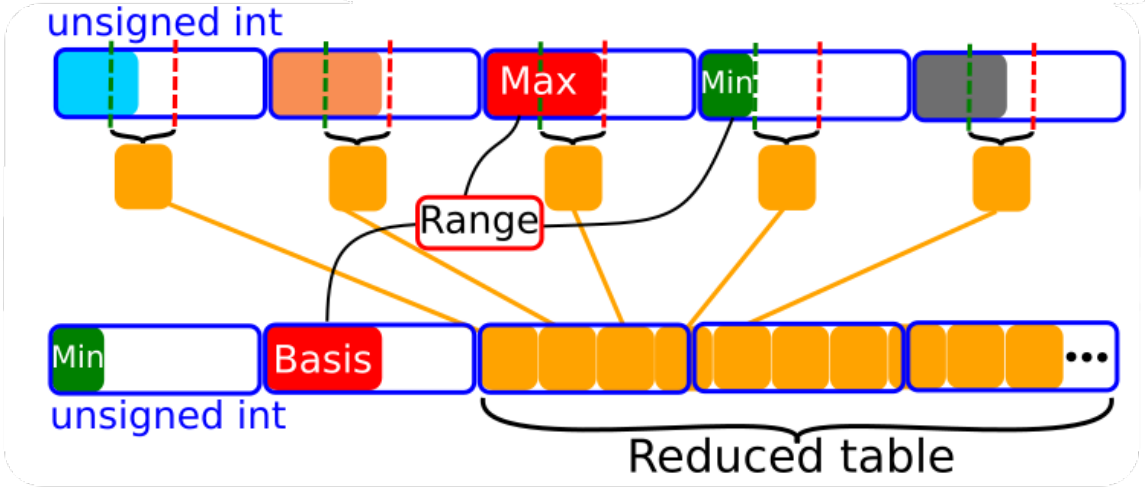




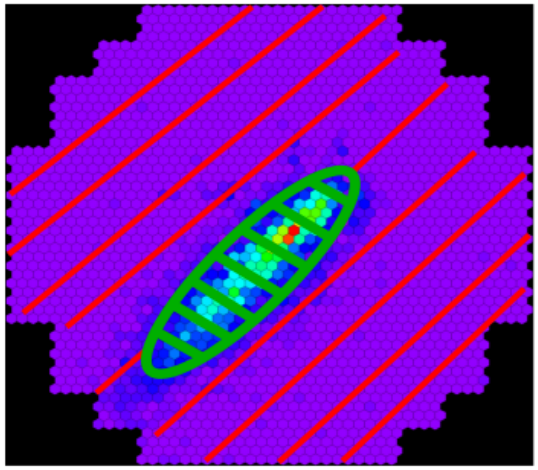
Lossless Data Compression

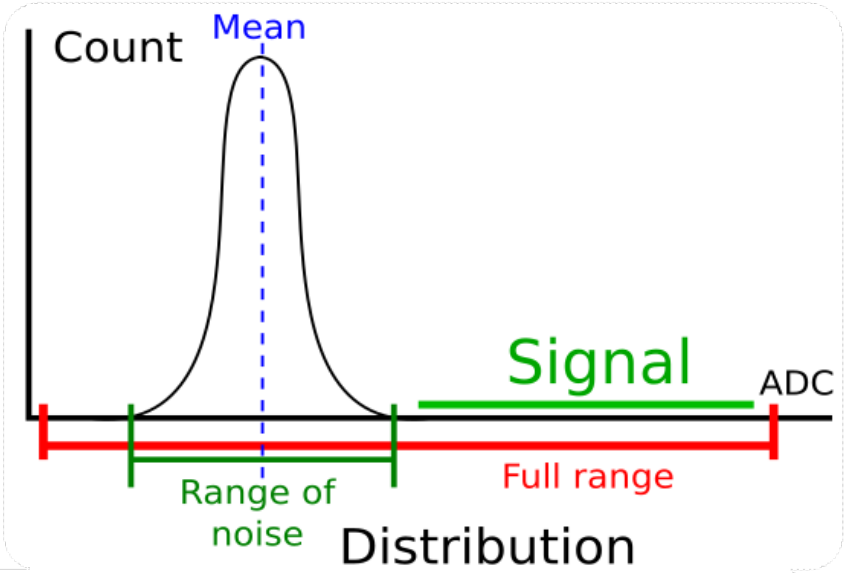




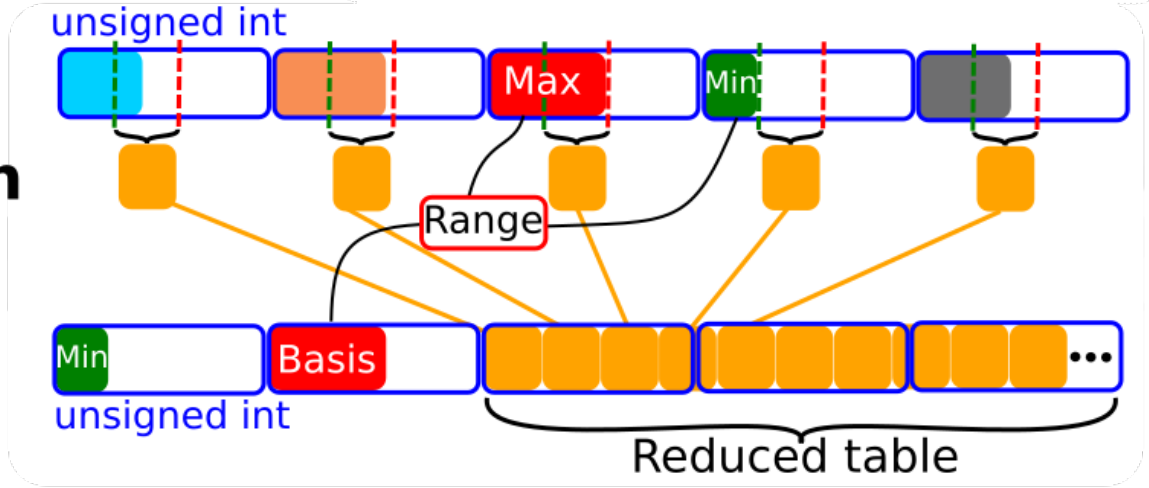


Lossless Data Compression





Publication ongoing



Raw data compression and speed

Test File
- **475 MB**
- **Up to 99% ADC values**

	Compression ratio	Time	File size (MB)
LZMA (7Z)	**4.84**	**7 min 48s**	**98**
Advanced Polynomial Compression	**3.74**	**3.7 s**	**127**

Raw data compression and speed

Test File
- **475 MB**
- **Up to 99% ADC values**

	Compression ratio	Time	File size (MB)
LZMA (7Z)	**4.84**	**7 min 48s**	98
Advanced Polynomial Compression	3.74	3.7 s	127
Advanced Polynomial Compression + LZMA	**4.84**	24.646 s	98

Raw data compression and speed

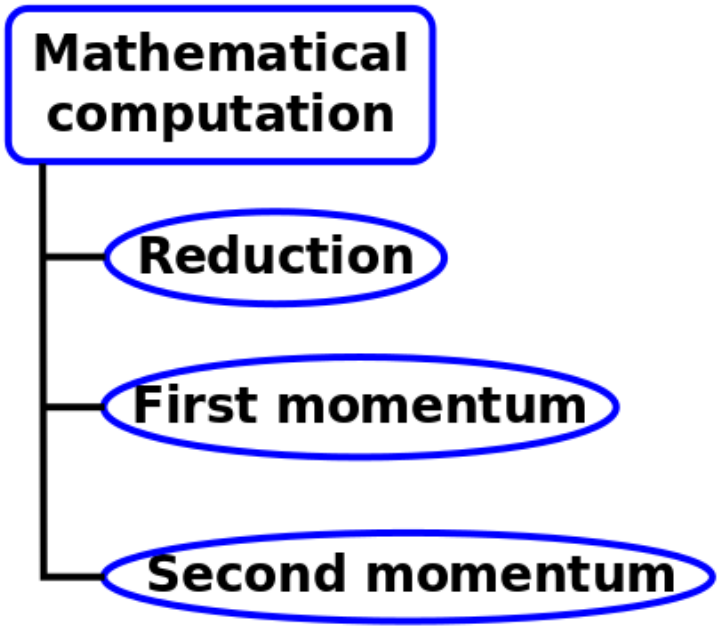
Test File
- **475 MB**
- **Up to 99% ADC values**

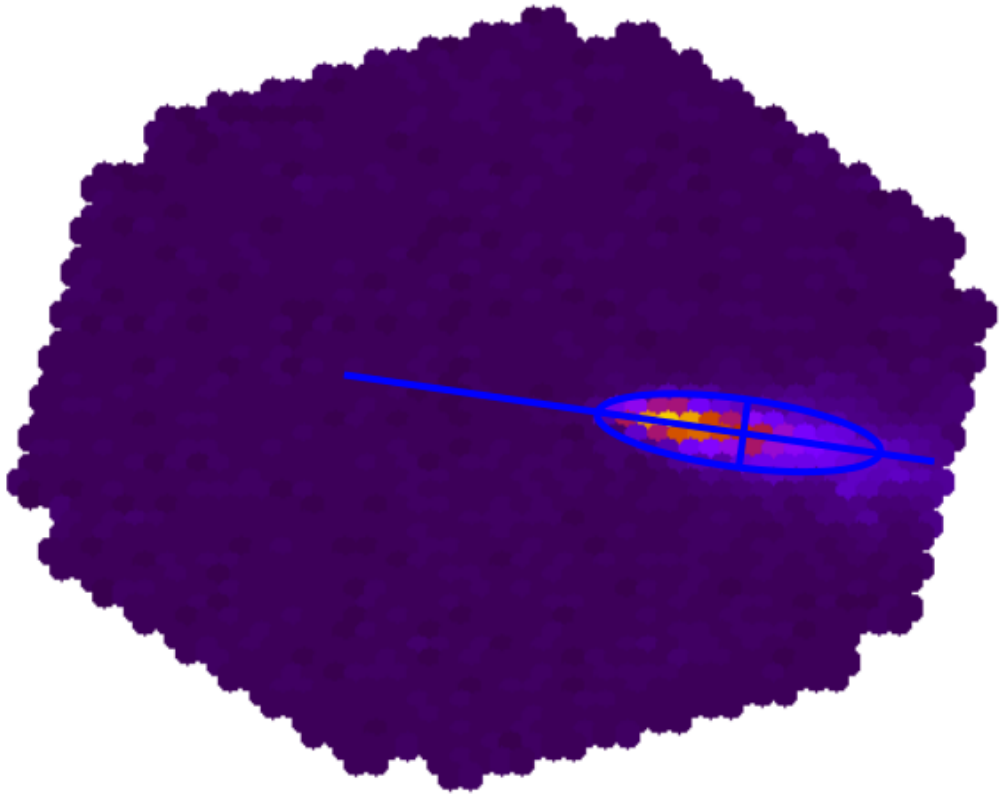
	Compression ratio	Time	File size (MB)
LZMA (7Z)	4.84	7 min 48s	98
Advanced Polynomial Compression	3.74	3.7 s	127
Advanced Polynomial Compression + LZMA	4.84	24.646 s	98

Same Compression ratio

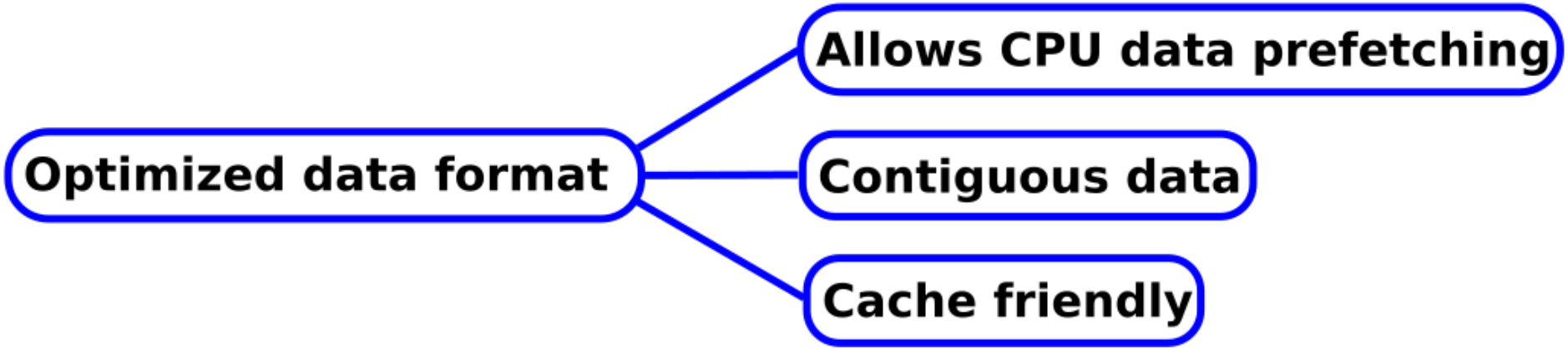
19 × faster

Hillas method



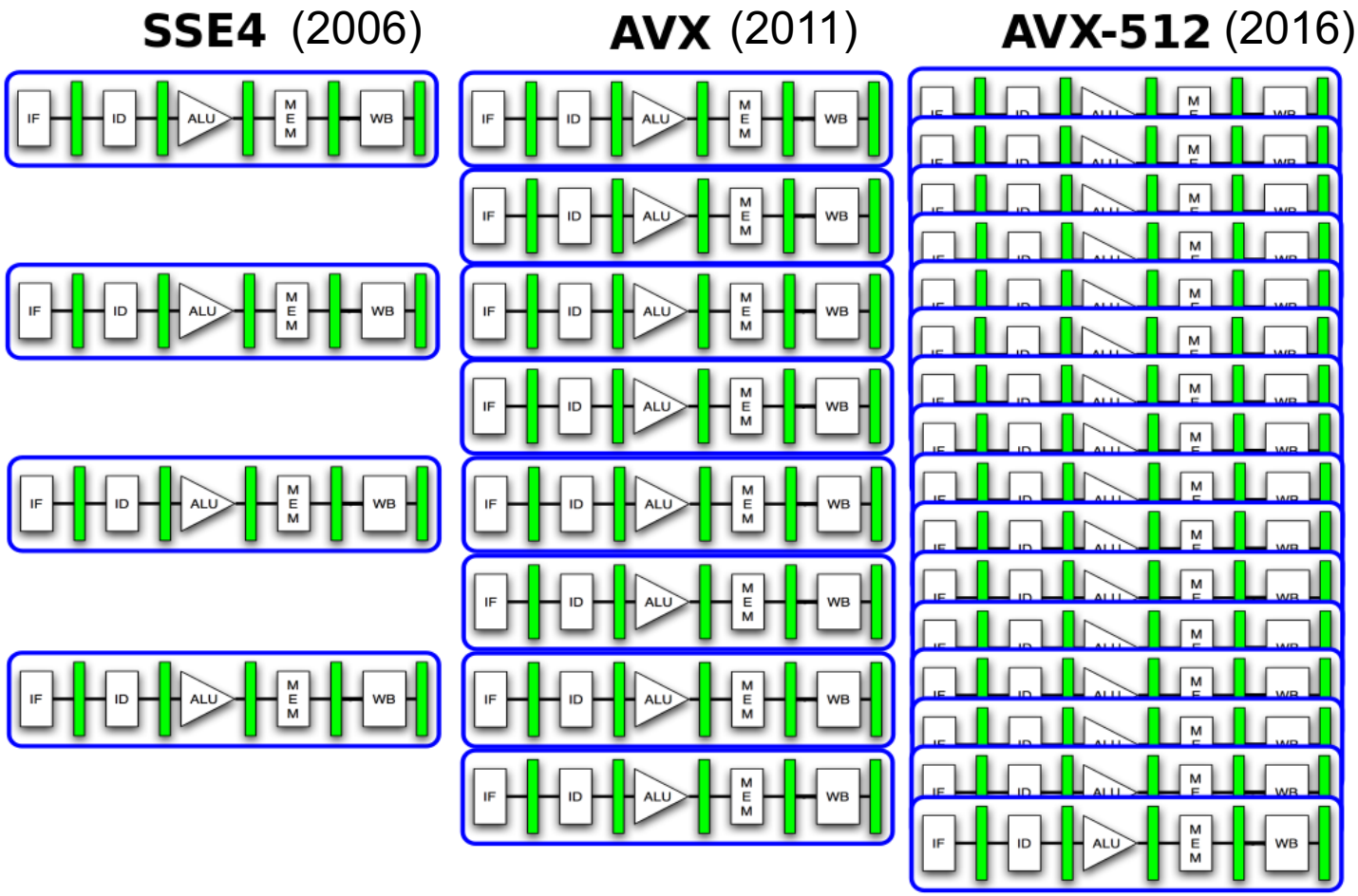


Hillas with optimized data format



	Speed (cy/el)	Speed up
Initial data format (H.E.S.S. ROOT)	2125.5	1
Optimized data format	53.1365	**40.0**

CPU Architecture



Reduction Optimization

Reduction : Sum of all vector elements

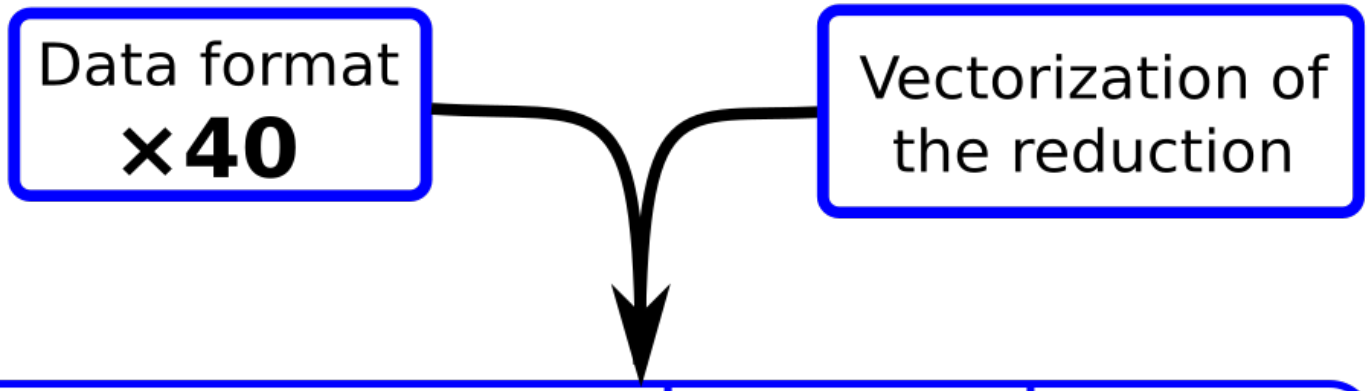
	Speed (cy/el)	Speed up
Classical	**2.69842**	1
Vectorized (GCC, SSE4)	0.702845	3.8

Reduction Optimization

Reduction : Sum of all vector elements

	Speed (cy/el)	Speed up
Classical	**2.69842**	1
Vectorized (GCC, SSE4)	0.702845	3.8
Intrinsic Vectorized SSE4	0.226675	11.9
Intrinsic Vectorized AVX	**0.11379**	**23.7**

Hillas with optimized data format



	Speed (cy/el)	Speed up
Initial data format	2125.5	1
Optimized data format	53.1365	**40.0**
+ Intrinsic Vectorized SSE4	6.39931	332
+ Intrinsic Vectorized AVX	**2.98499**	**712**

Hillas with optimized data format

Data format ×40

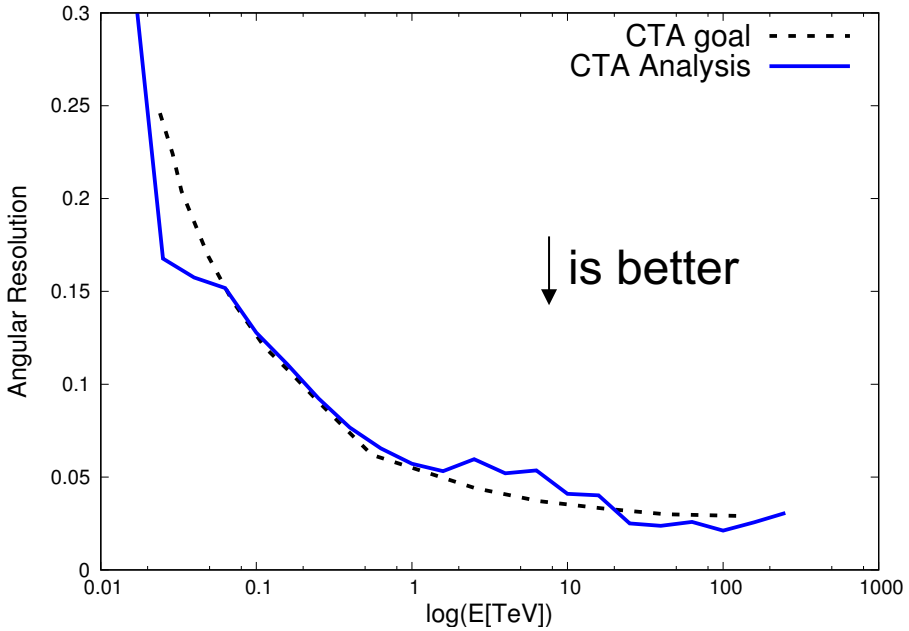
Vectorization of the reduction

	Speed (cy/el)	Speed up
Initial data format	2125.5	1
Optimized data format	53.1365	**40.0**
+ Intrinsic Vectorized SSE4	6.39931	332
+ Intrinsic Vectorized AVX	**2.98499**	**712**

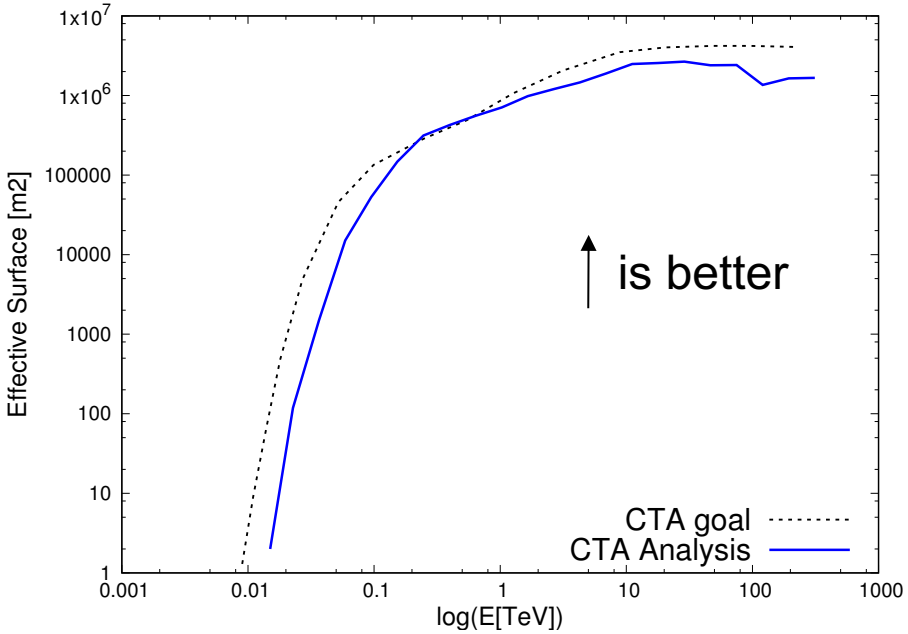
Reduction without optimization 2.69842

HPC reconstruction based on Hillas parameters

Angular resolution



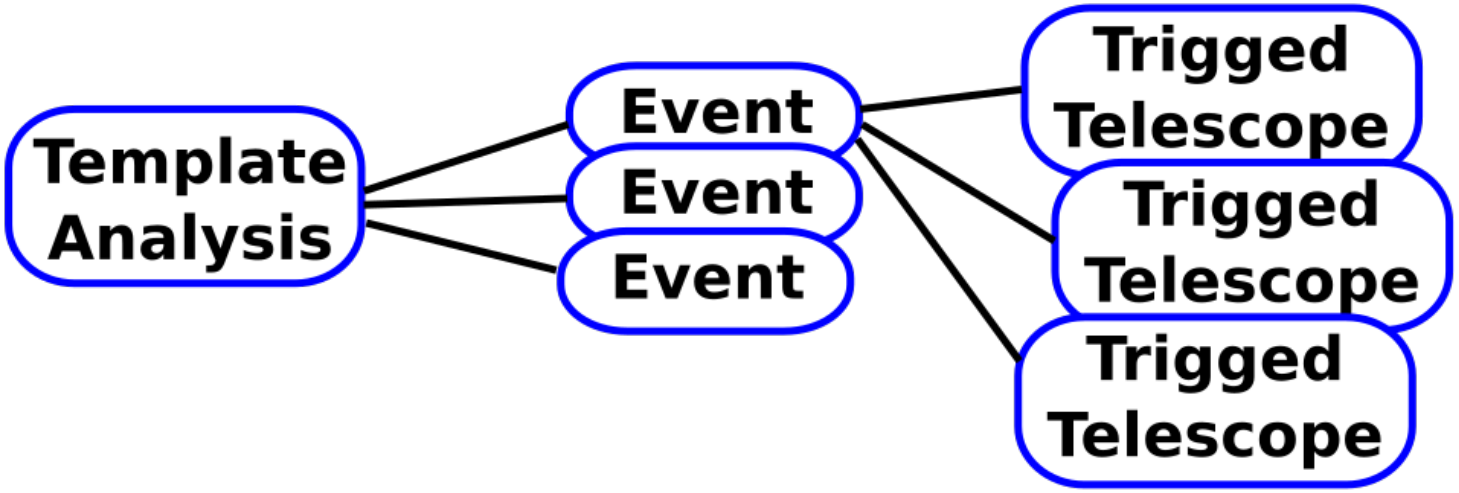
Effective area



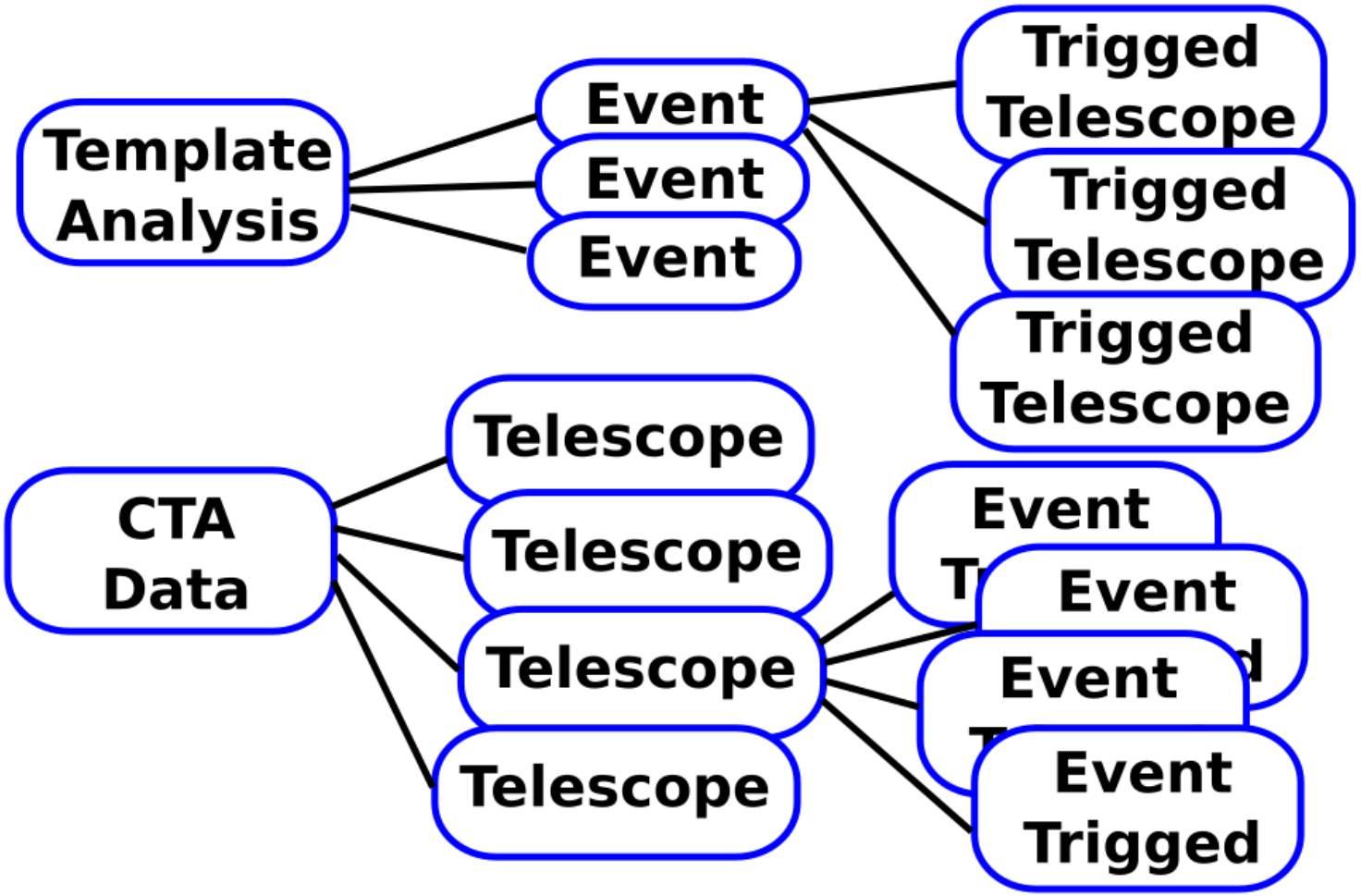
- physics results almost in agreement with CTA requirements (but no fine tuning yet)
- HPC reduces computation time by 600 compared to current analysis (called in Python)
- Open-source library available to download
  https://gitlab.in2p3.fr/CTA-LAPP/PLIBS_8.git

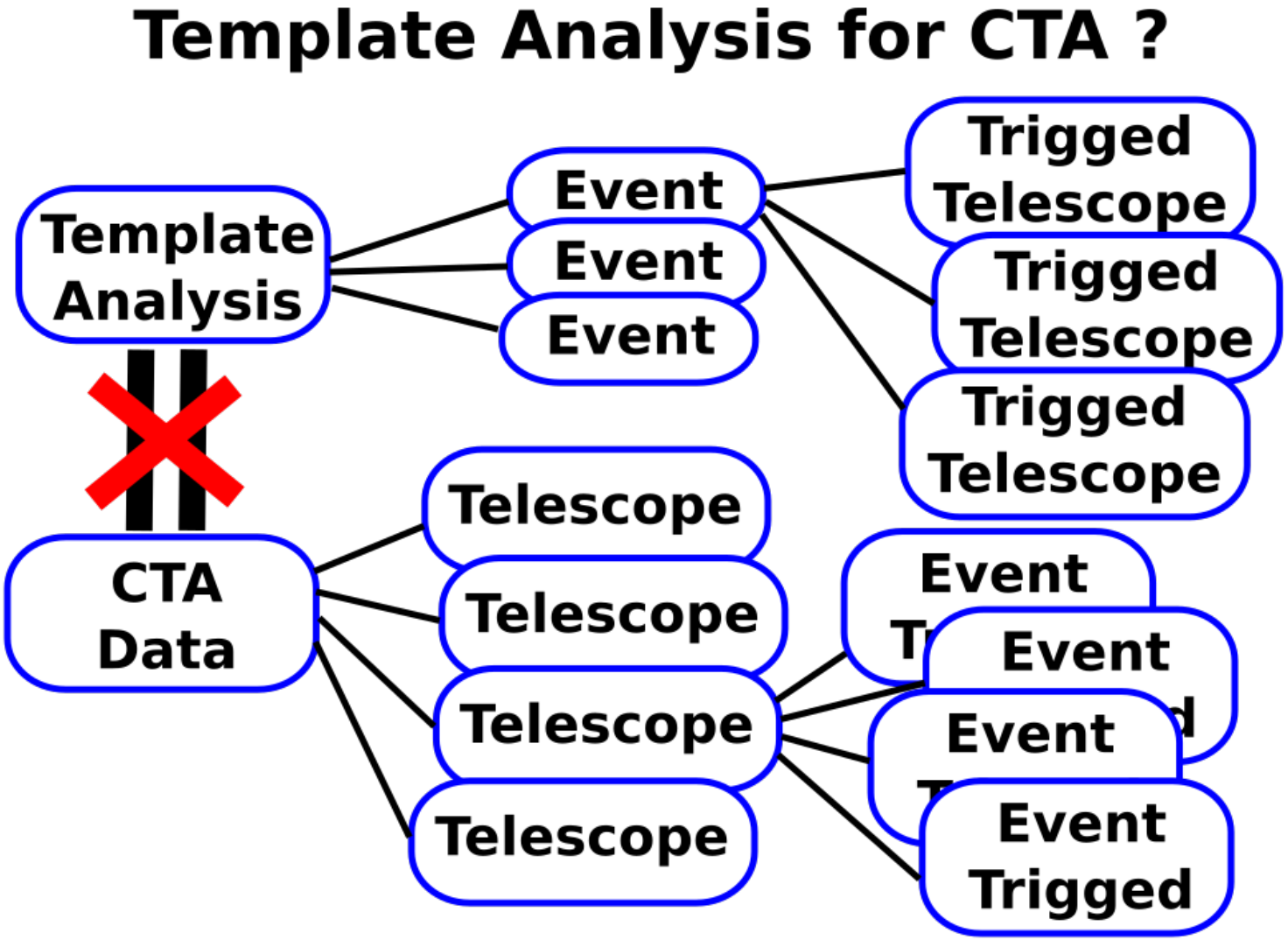
Template Analysis for CTA ?

Template Analysis for CTA ?

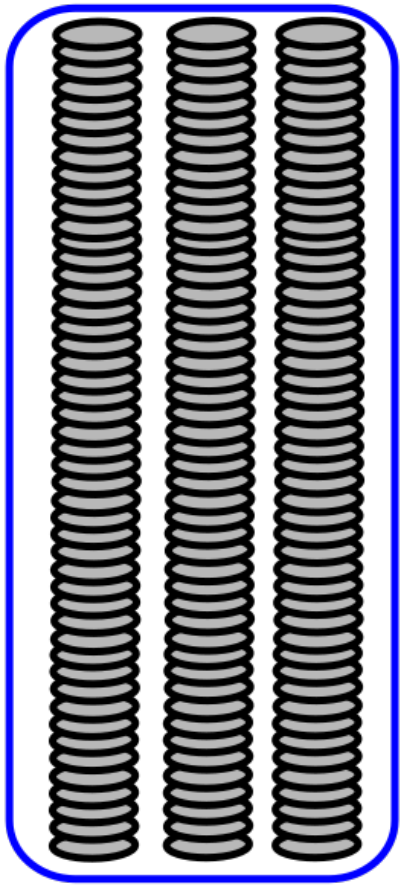


Template Analysis for CTA ?

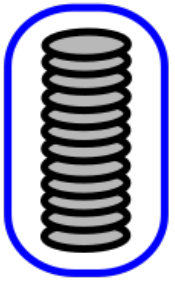



Template Analysis for CTA ?
Template Analysis
CTA Data
Event
Event
Event
Telescope
Telescope
Telescope
Telescope
Trigged Telescope
Trigged Telescope
Trigged Telescope
Event
Event
Event
Event Trigged

CTA Data Sort ?

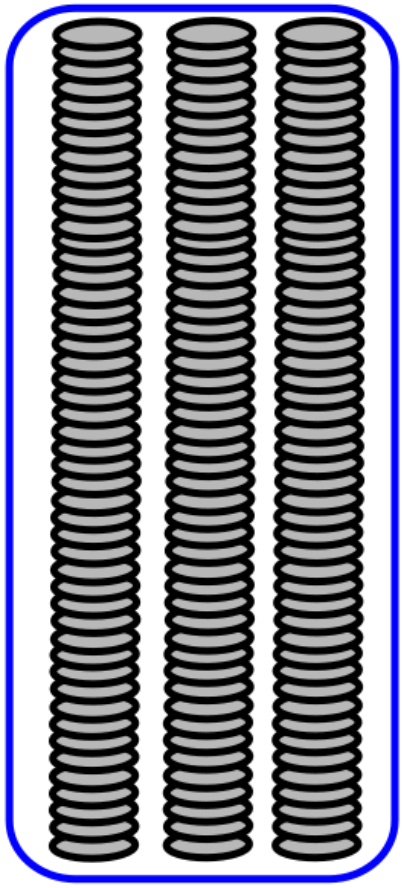


Very hard sort

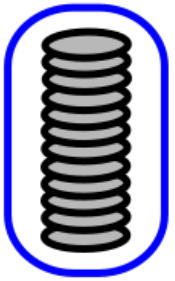


Easy sort

CTA Data Sort ?

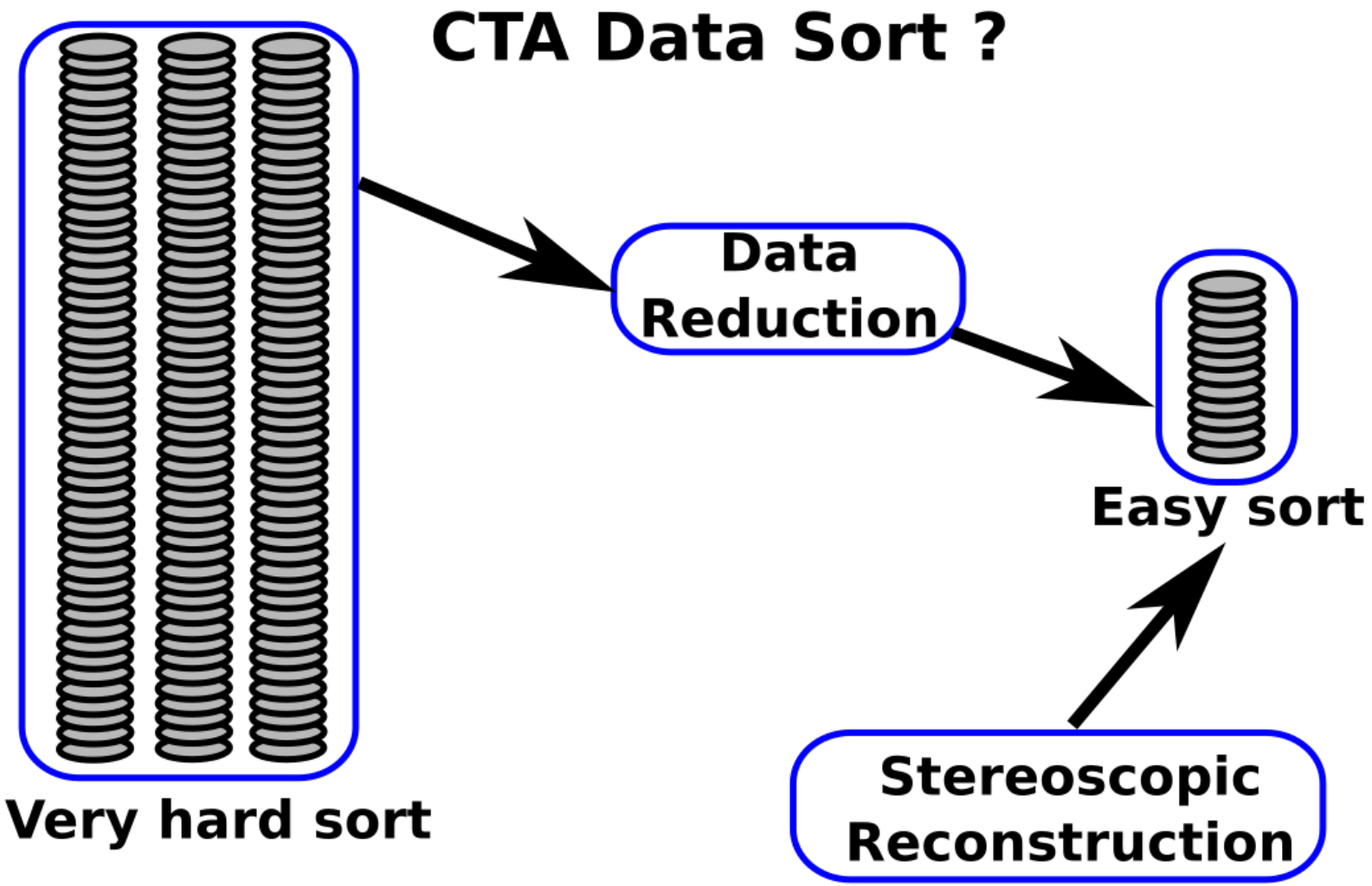


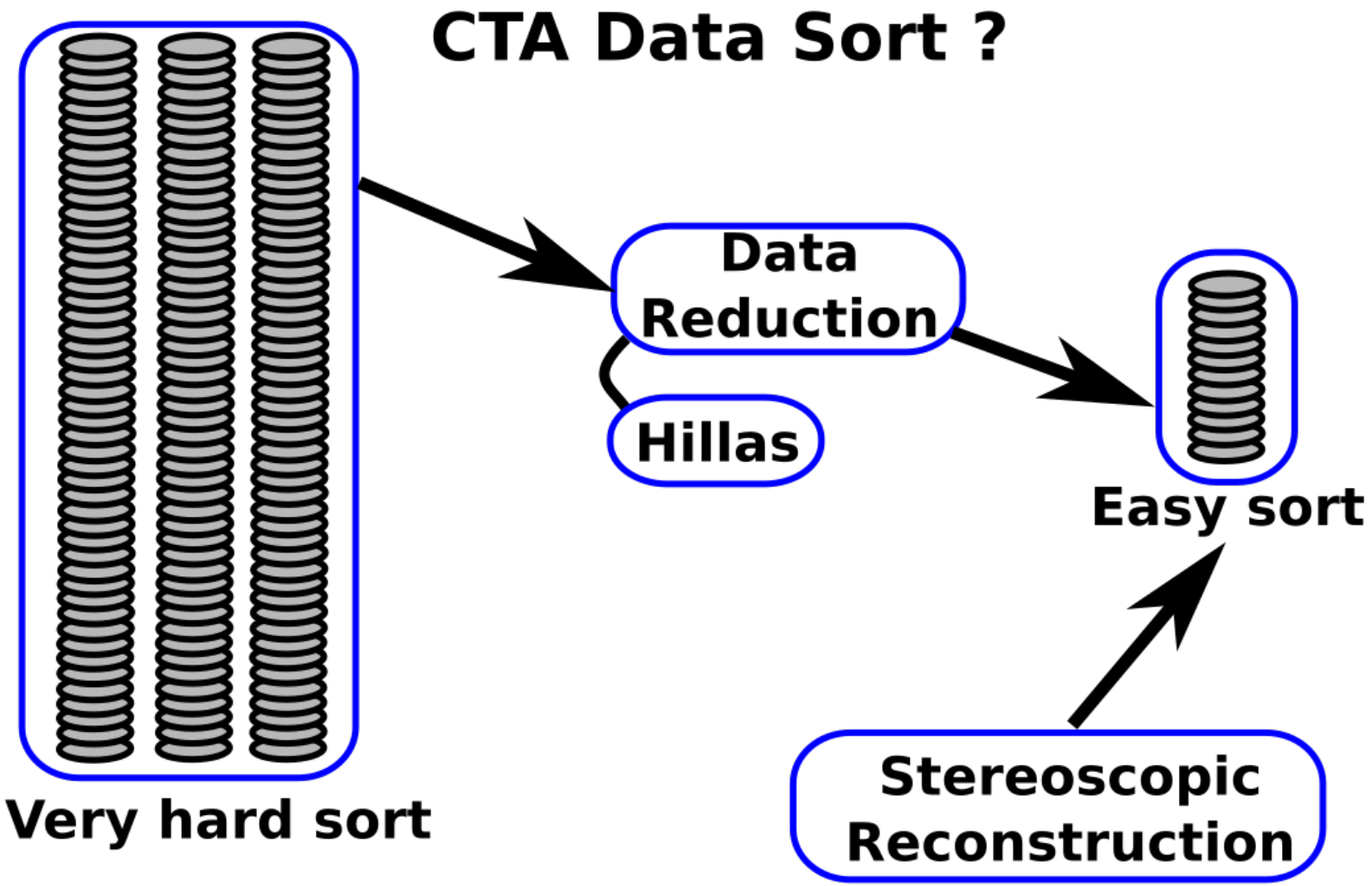
Very hard sort

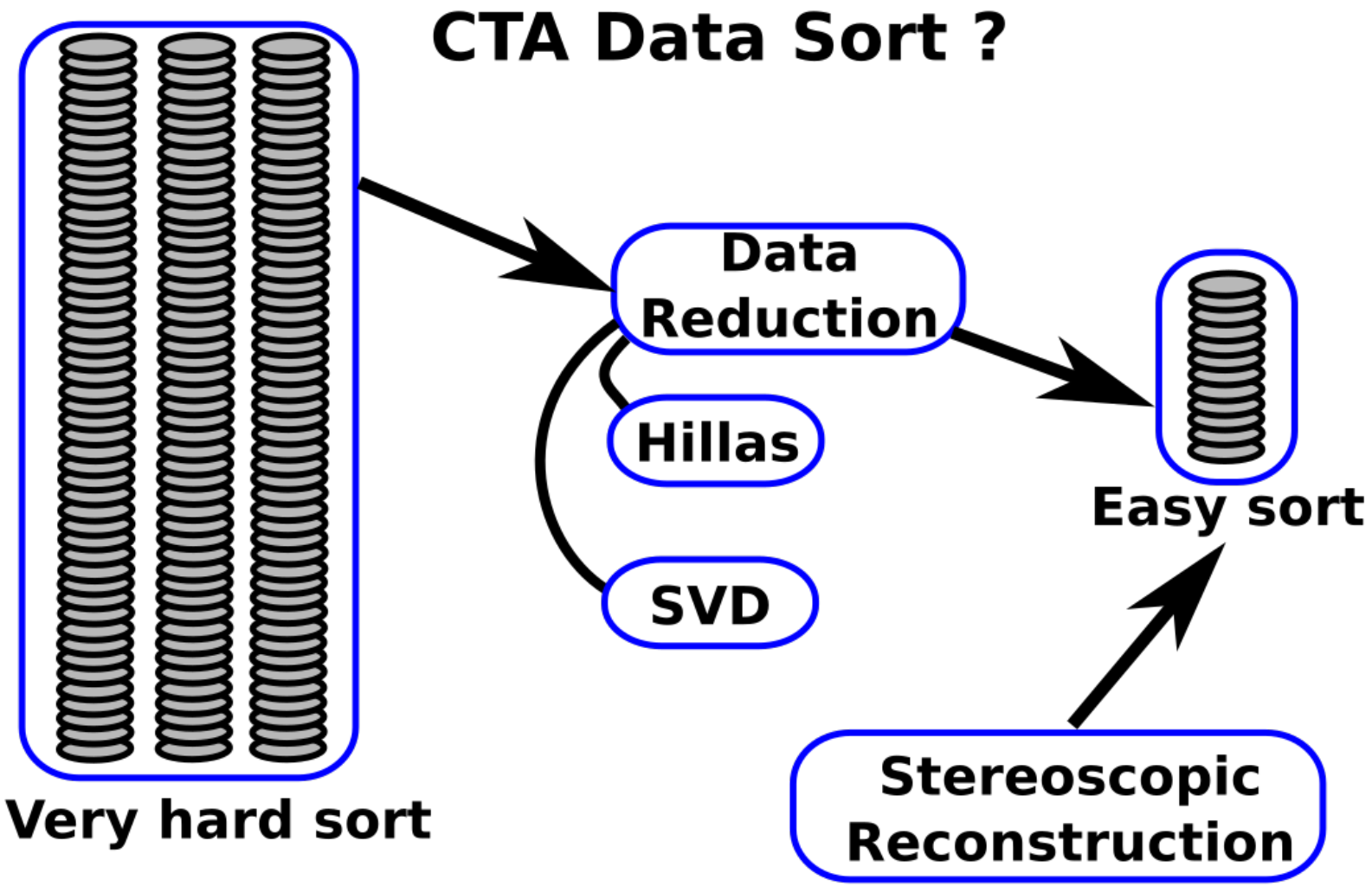


Easy sort

Stereoscopic Reconstruction

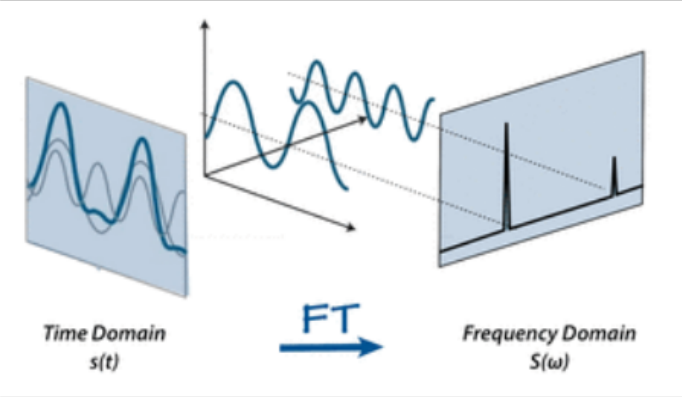

CTA Data Sort ?
Data Reduction
Easy sort
Very hard sort
Stereoscopic Reconstruction


CTA Data Sort ?
Data Reduction
Hillas
Easy sort
Very hard sort
Stereoscopic Reconstruction


CTA Data Sort ?
Very hard sort
Data Reduction
Hillas
SVD
Easy sort
Stereoscopic Reconstruction

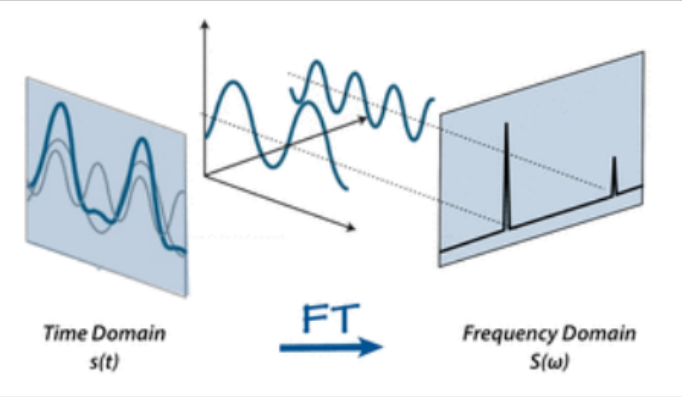
Singular Value Decomposition

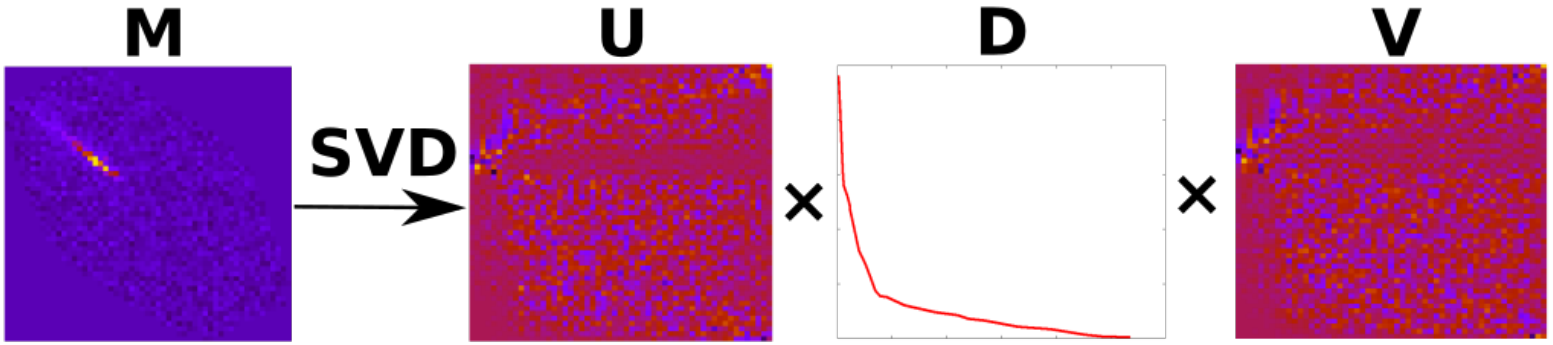
Analogy with Fourier Transform



Singular Value Decomposition

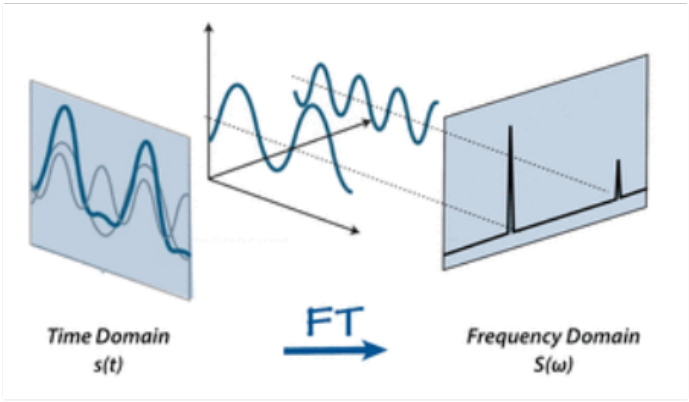
Analogy with Fourier Transform



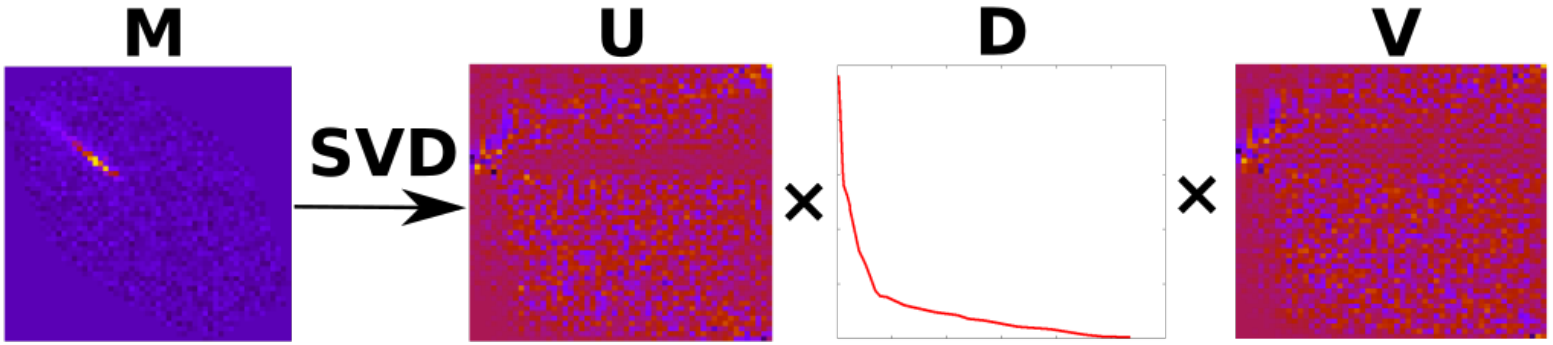


Singular Value Decomposition

Analogy with Fourier Transform

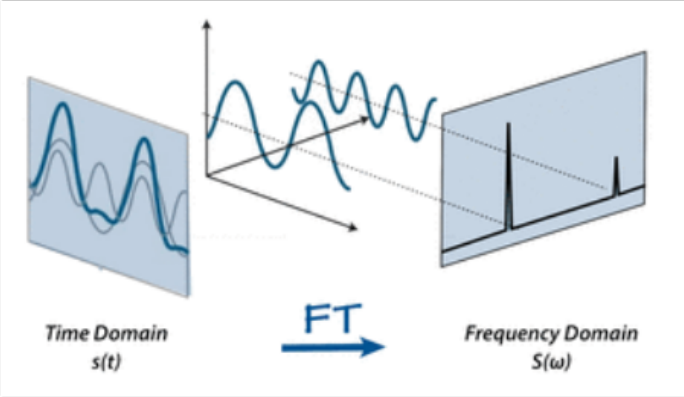


Scalable
Already vetorized



Singular Value Decomposition

Analogy with Fourier Transform



Scalable
Already vetorized

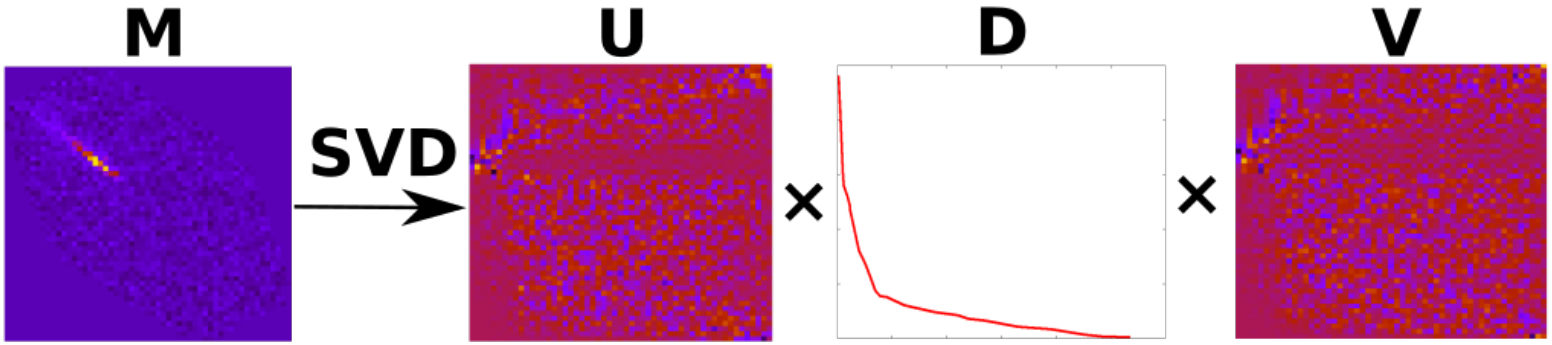
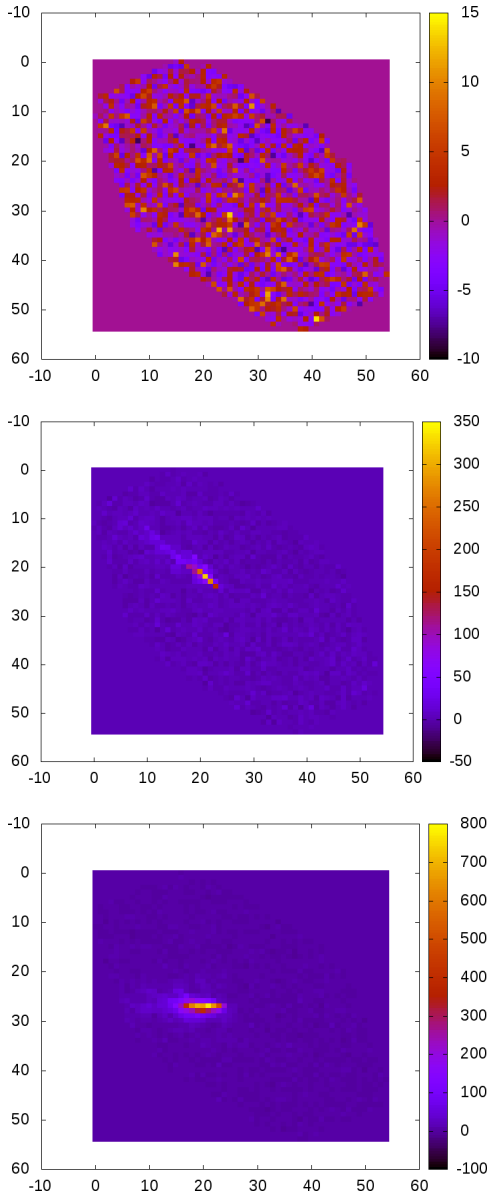
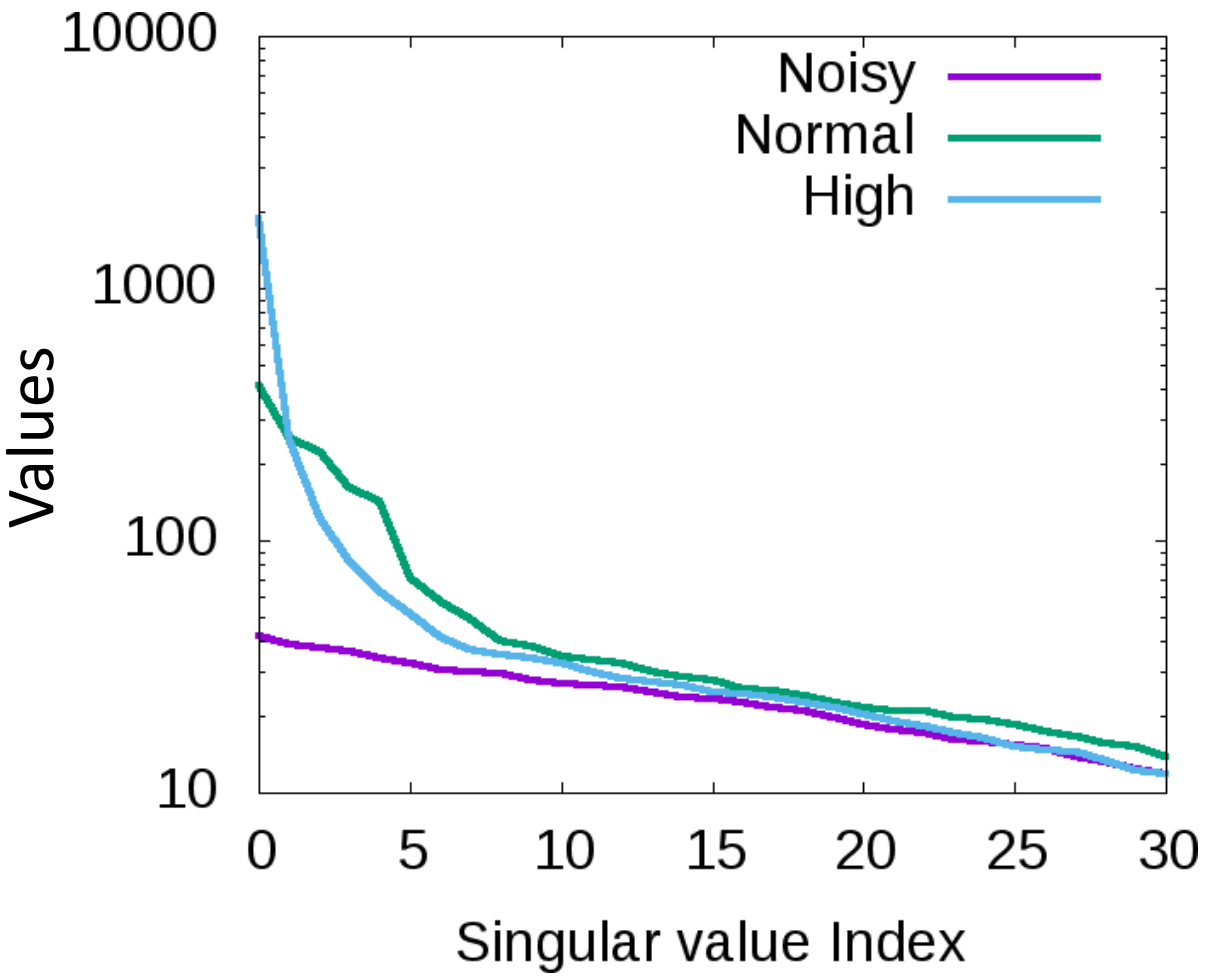
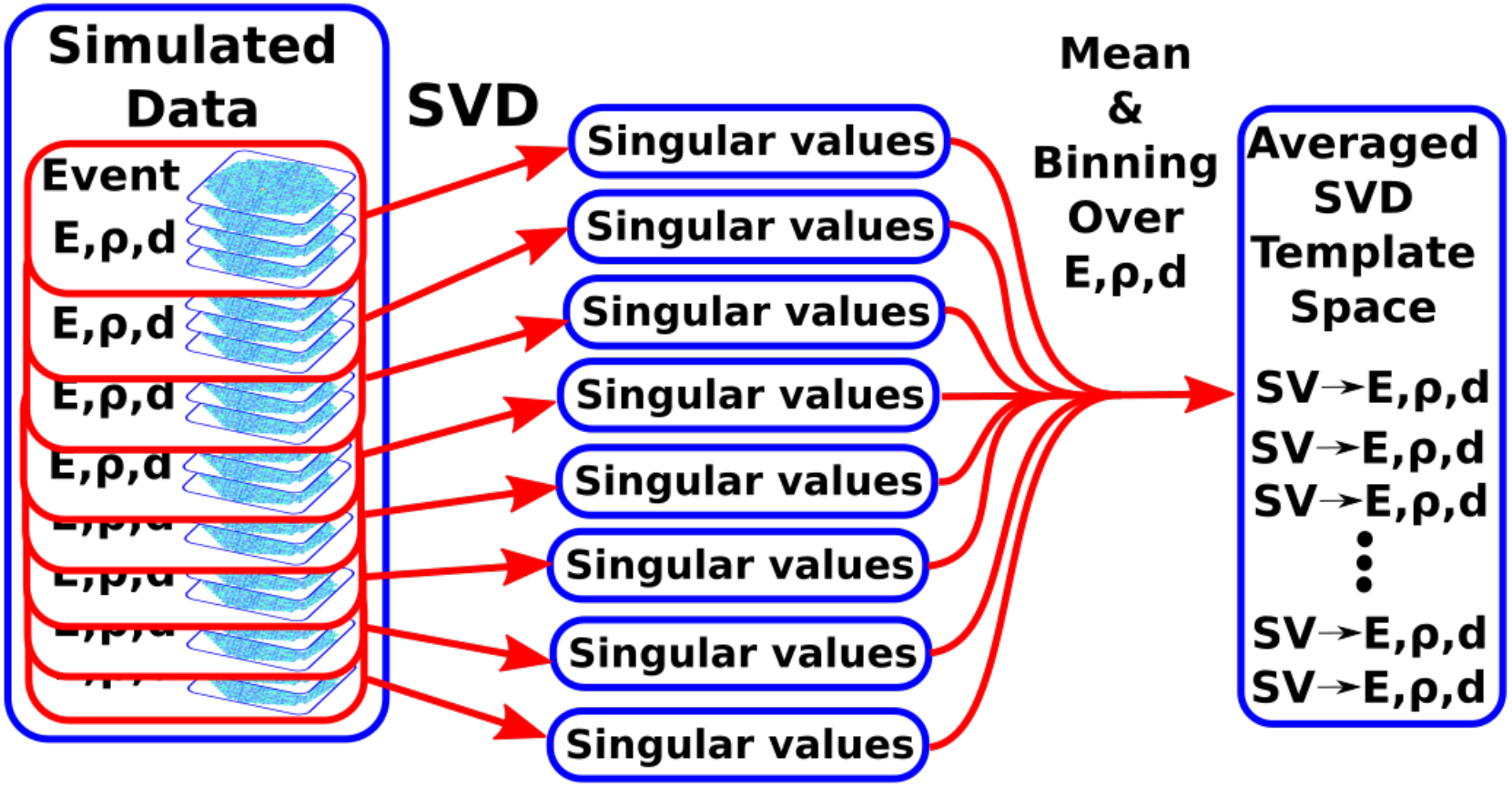


Image Search : Google Images, Wolfram

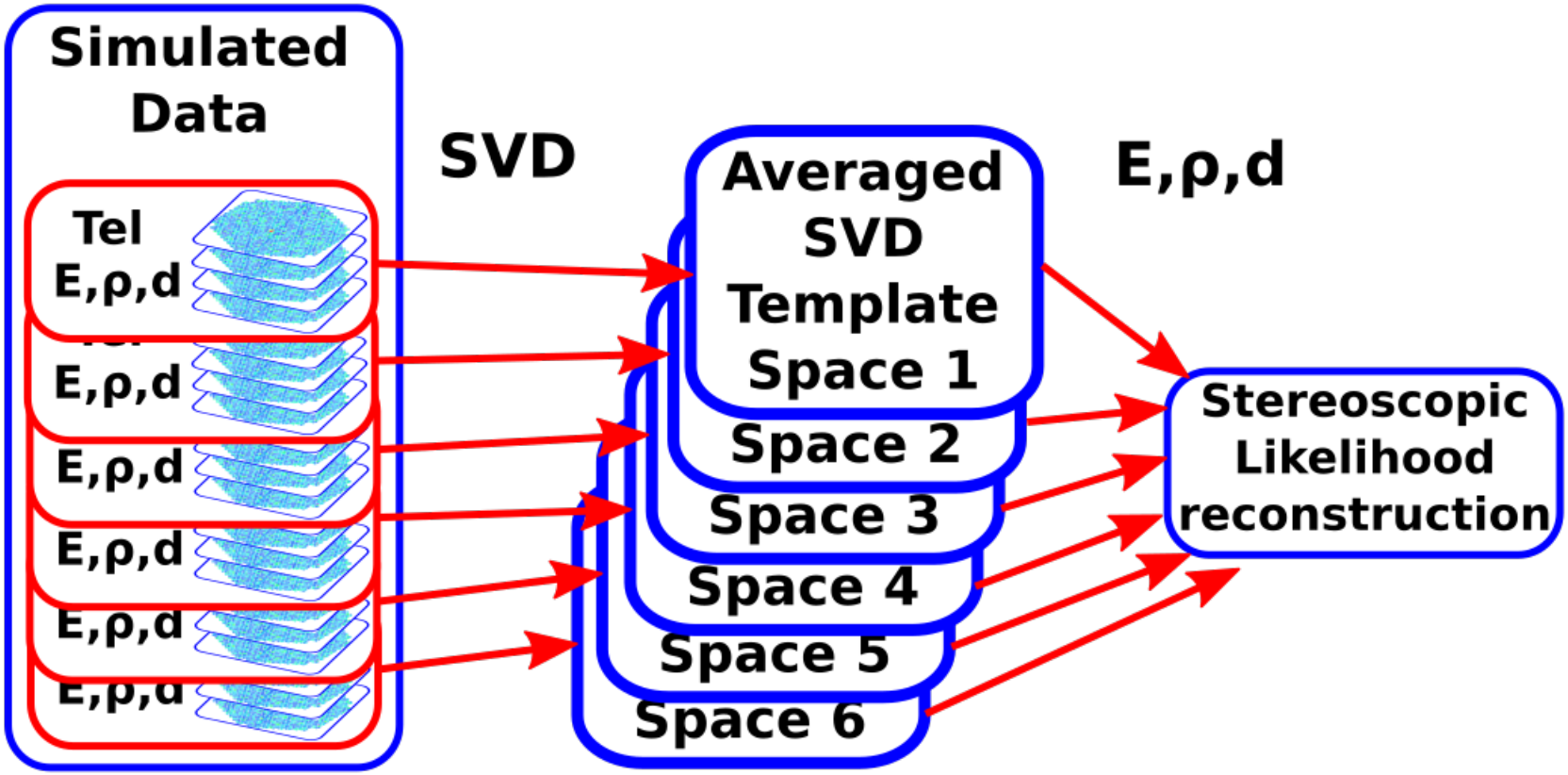

-10 0 10 20 30 40 50 60
15 10 5 0 -5 -10
350 300 250 200 150 100 50 0 -50
800 700 600 500 400 300 200 100 0 -100


10000
1000
100
10
Values
0 5 10 15 20 25 30
Singular value Index
Noisy
Normal
High

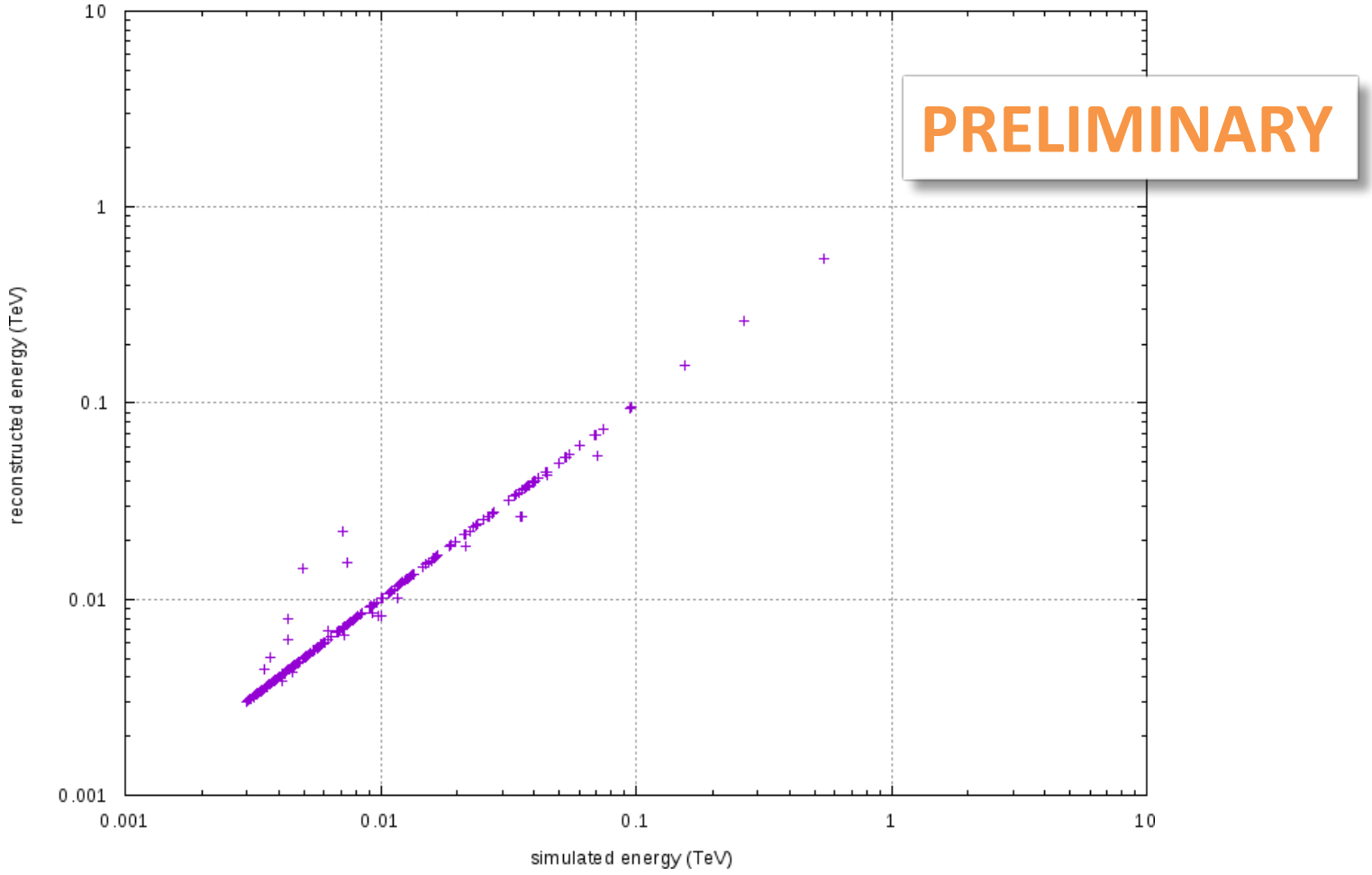
SVD Analysis template creation



SVD Template Analysis



SVD energy reconstruction



- Astronomy is entering the big data era and facing new challenges (data volumes, data rates, real-time…)

- High Performance Computing to the help
  - Compression
  - Vectorization for Hillas based reconstruction
  - New reconstruction methods (e.g. SVD)
    - Scalable and vectorised
    - Hopefully better physics

ASTERICS is a project supported by the European Commission Framework Programme Horizon 2020 Research and Innovation action under grant agreement n. 653477