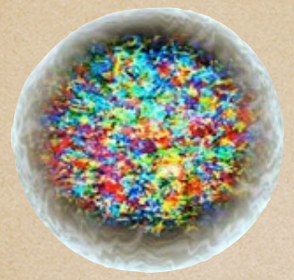
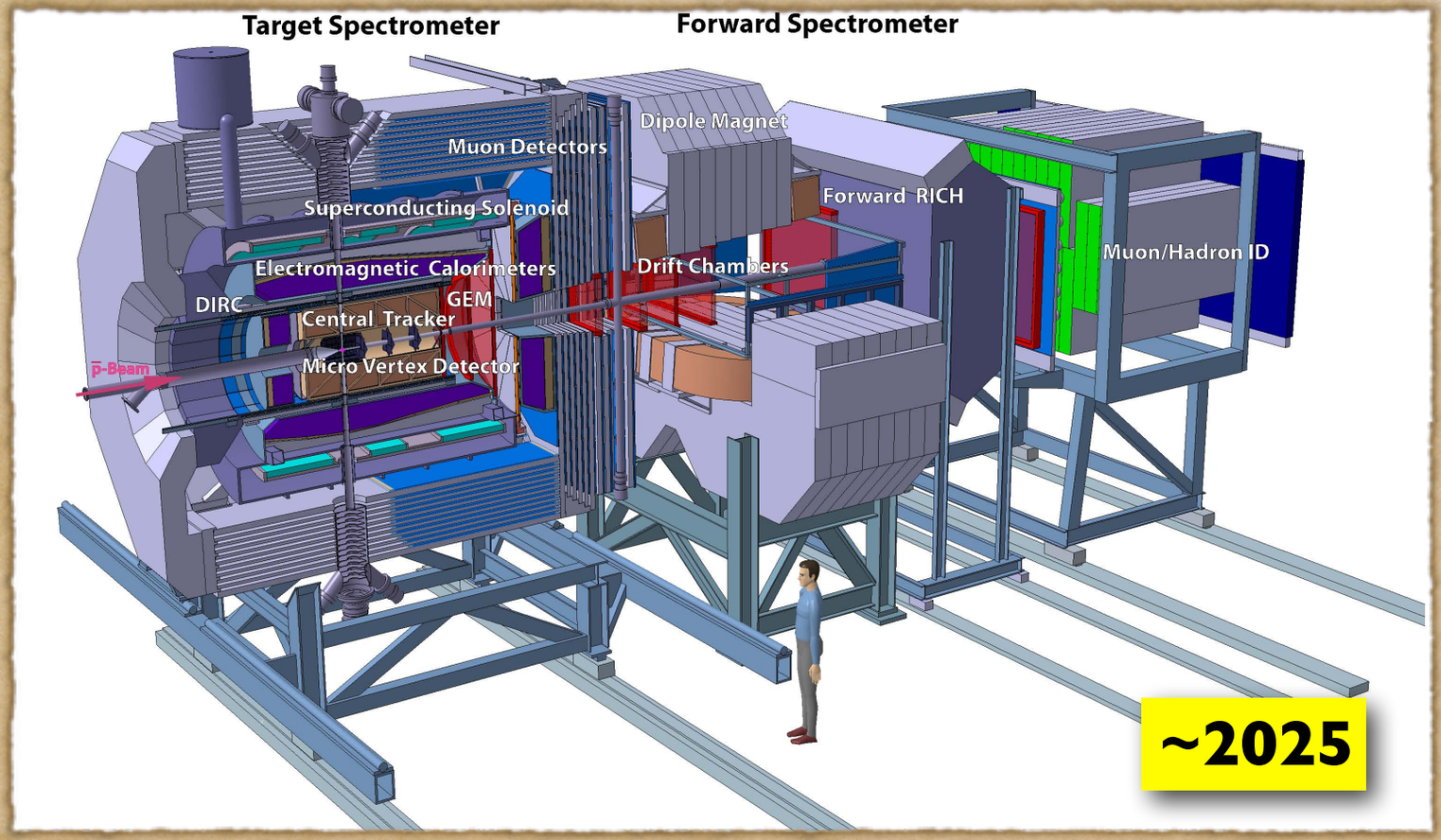
Data processing of a "typical" FAIR experiment

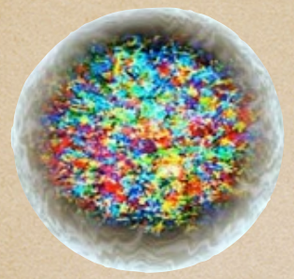
(based on some personal experiences)

Johan Messchendorp, April 17, 2019

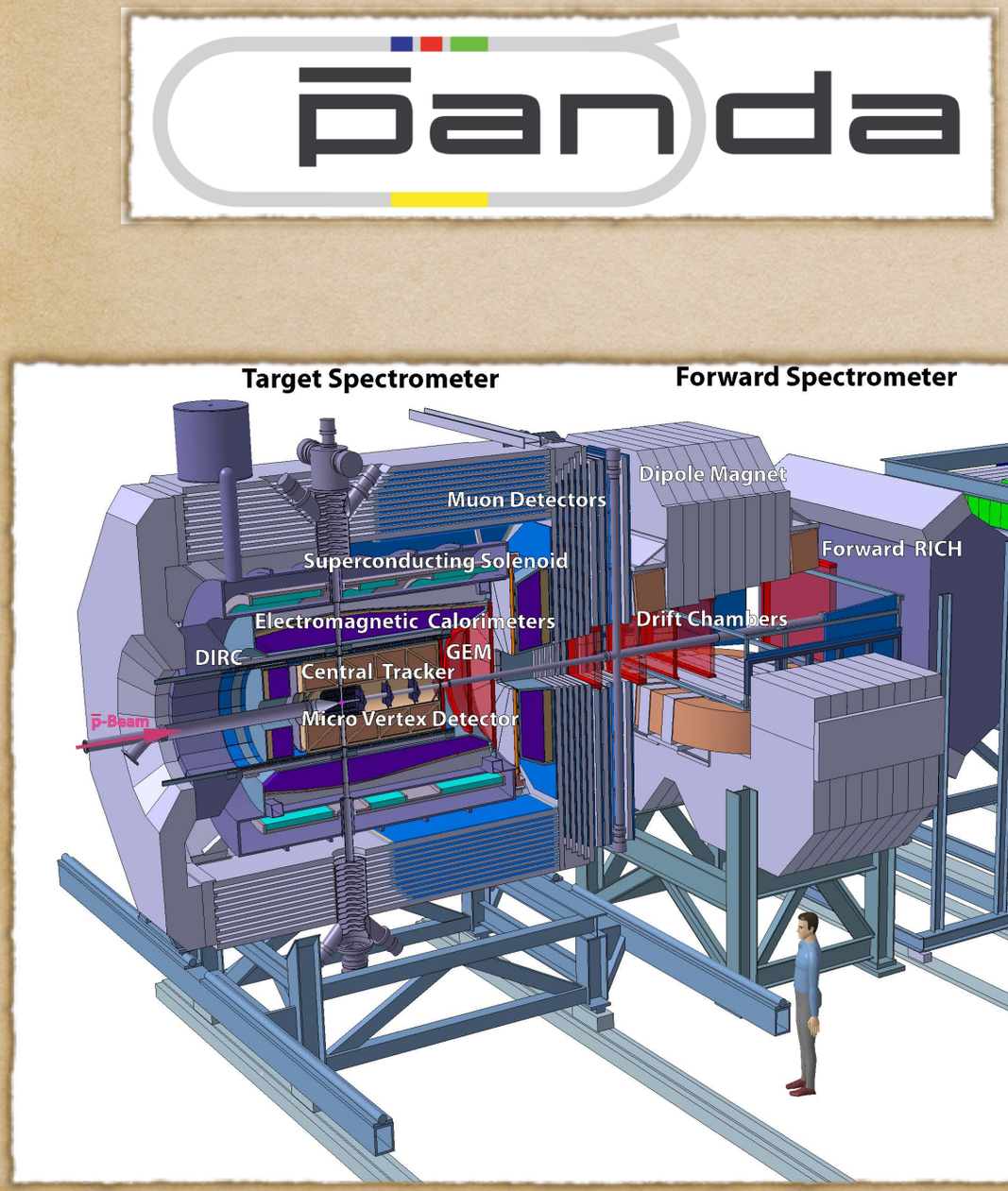


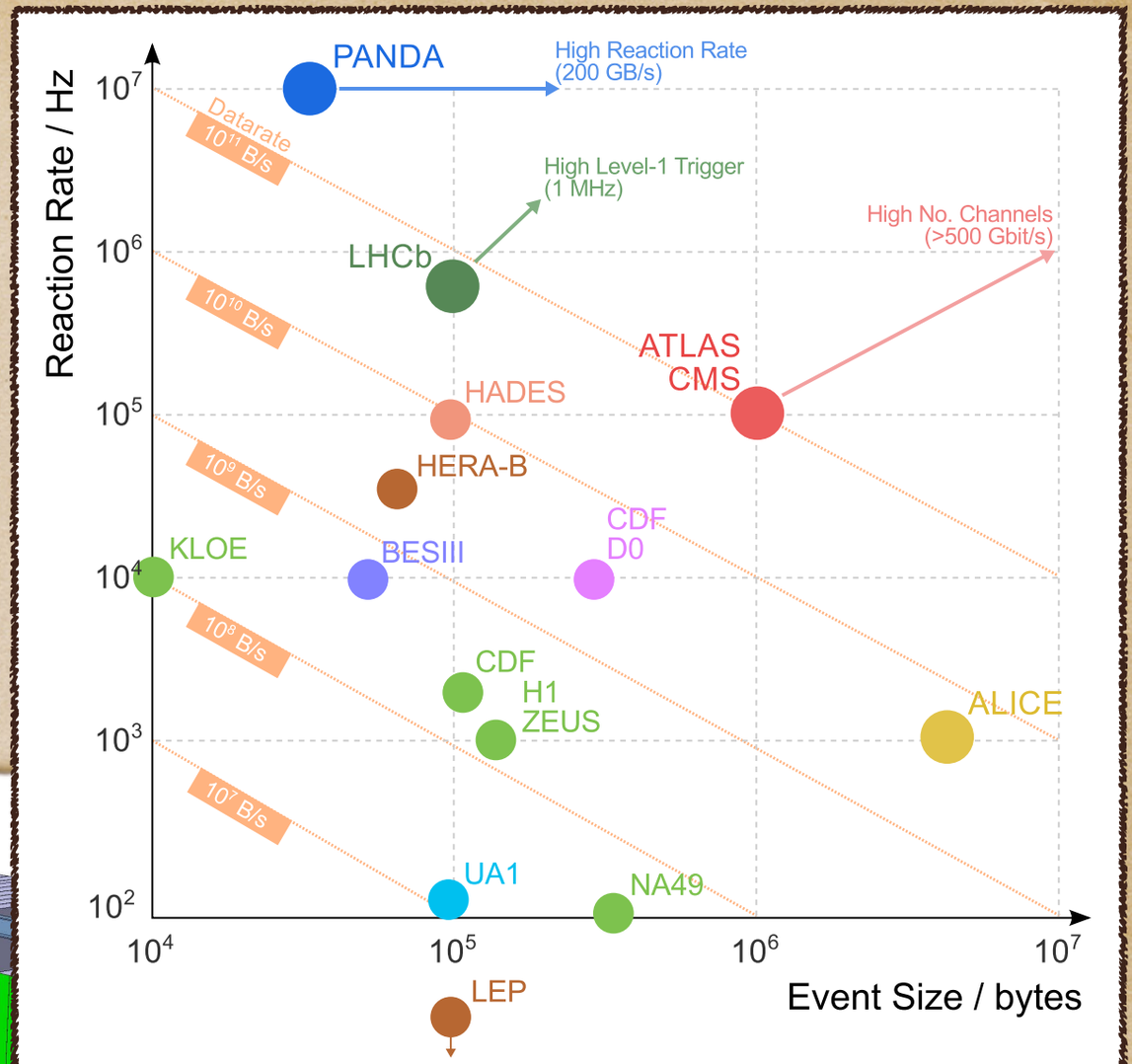
PANDA


Target Spectrometer
Forward Spectrometer
Dipole Magnet
Muon Detectors
Superconducting Solenoid
Forward RICH
Muon/Hadron ID
Electromagnetic Calorimeters
Drift Chambers
DIRC
GEM
Central Tracker
Micro Vertex Detector
p̄-Beam
~2025



PANDA

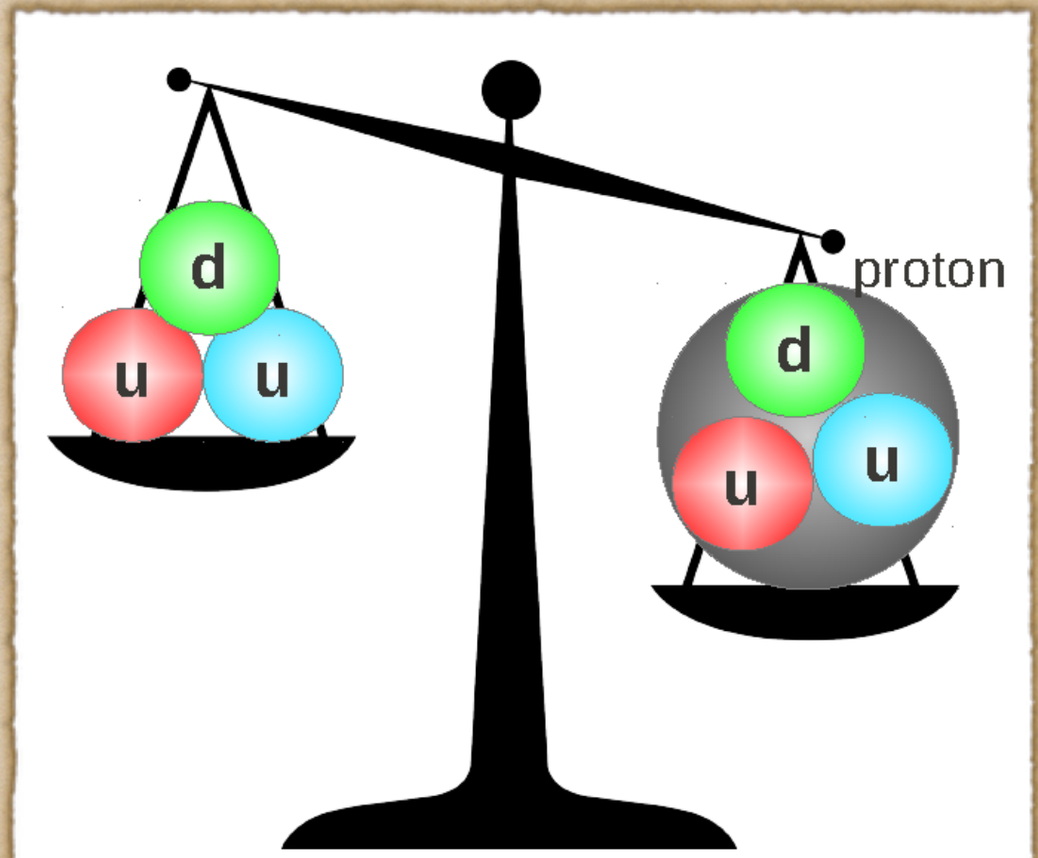

Target Spectrometer
Forward Spectrometer
Muon Detectors
Dipole Magnet
Superconducting Solenoid
Forward RICH
Electromagnetic Calorimeters
Drift Chambers
DIRC
GEM
Central Tracker
Micro Vertex Detector
p̄-Beam
~2025


Reaction Rate / Hz
Event Size / bytes
PANDA
High Reaction Rate
(200 GB/s)
High Level-1 Trigger
(1 MHz)
High No. Channels
(>500 Gbit/s)
Datarate
10^11 B/s
10^10 B/s
10^9 B/s
10^8 B/s
10^7 B/s
LHCb
ATLAS
CMS
HADES
HERA-B
KLOE
BESIII
CDF
D0
CDF
H1
ZEUS
ALICE
UA1
NA49
LEP

Physics

Origin of Mass & Color Confinement

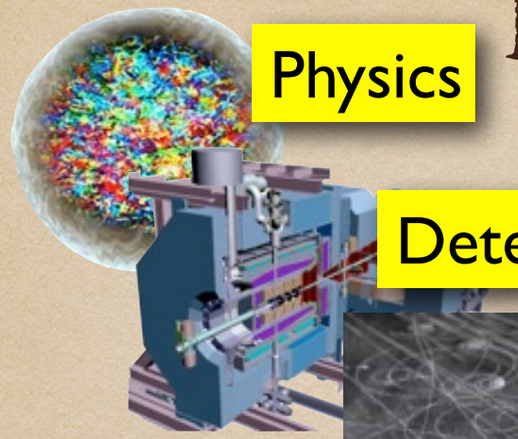
- Higgs addresses only a tiny (1%) fraction of the mass of all visible matter.
- The rest is due to the ***strong color force.***



Physics

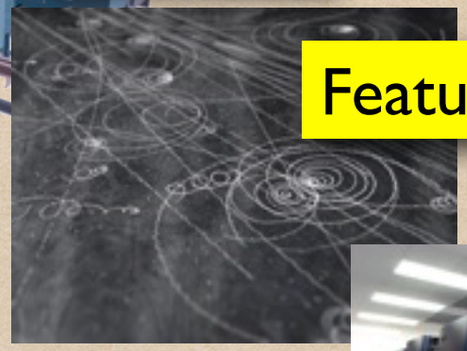
Data crunching

Data crunching



Physics

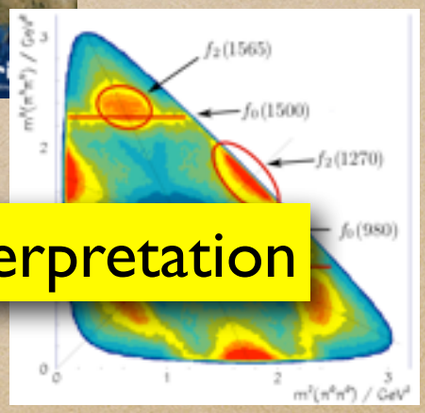
Detector



Feature extraction

Online computing

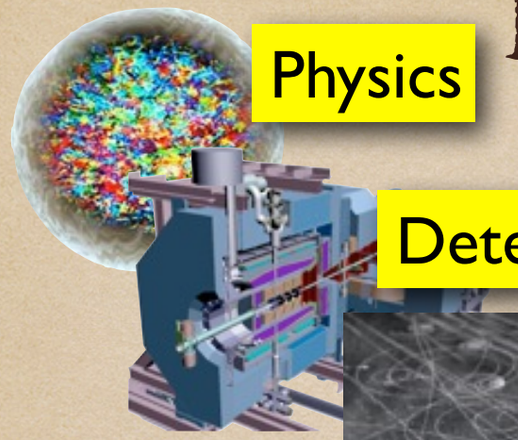
Distributed computing



Data interpretation

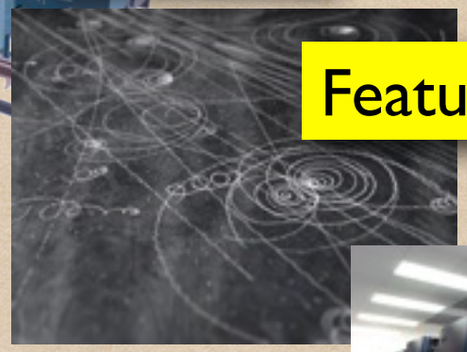
Data crunching

Physics

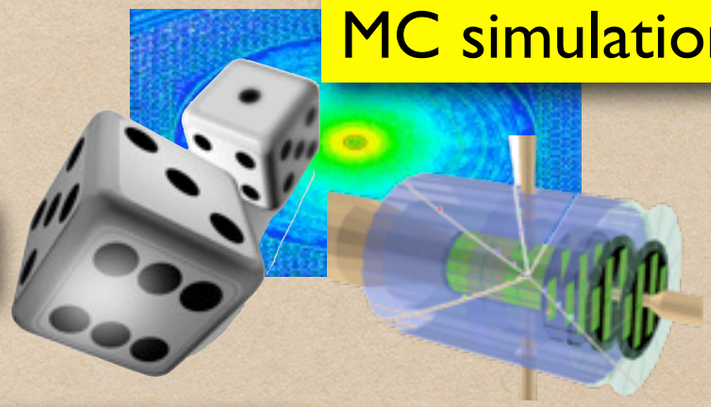


Detector

Feature extraction



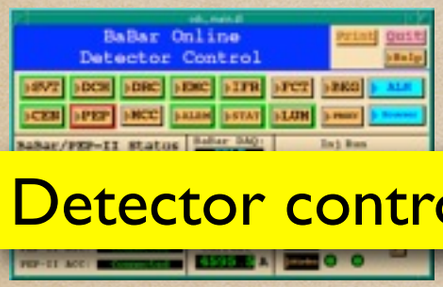
MC simulations



Online computing

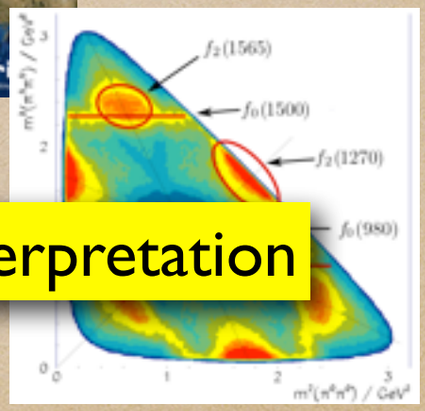
Distributed computing

Detector control



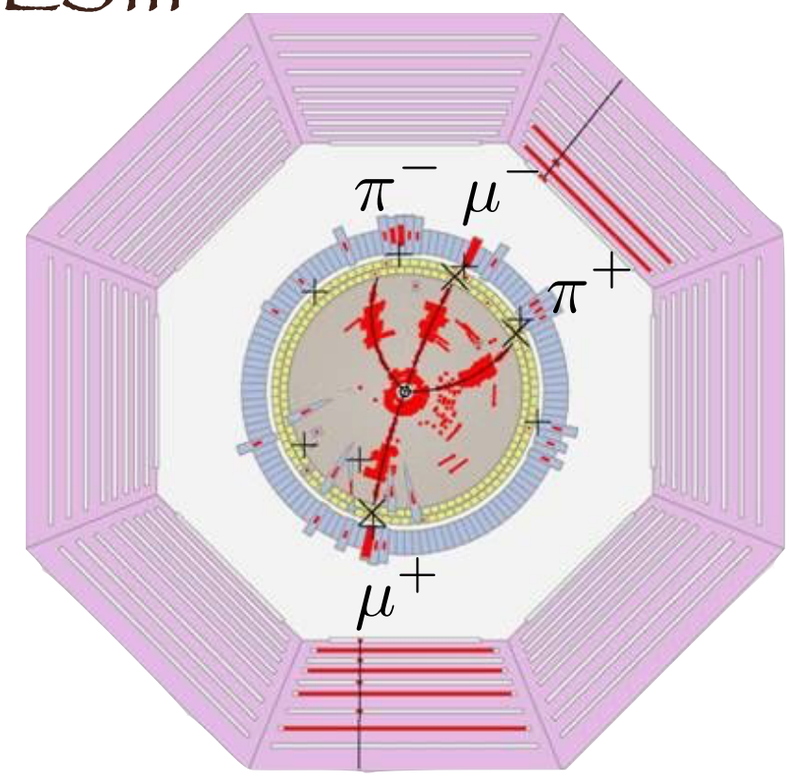
Database management

Data interpretation



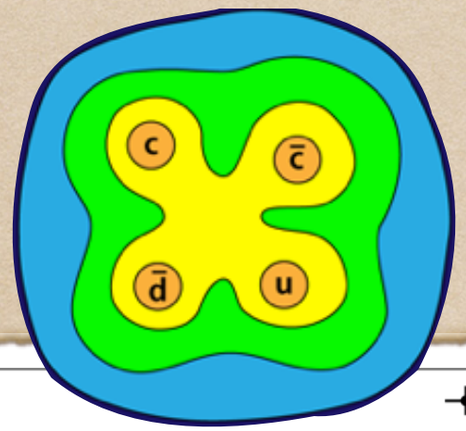
Basic analysis ingredients

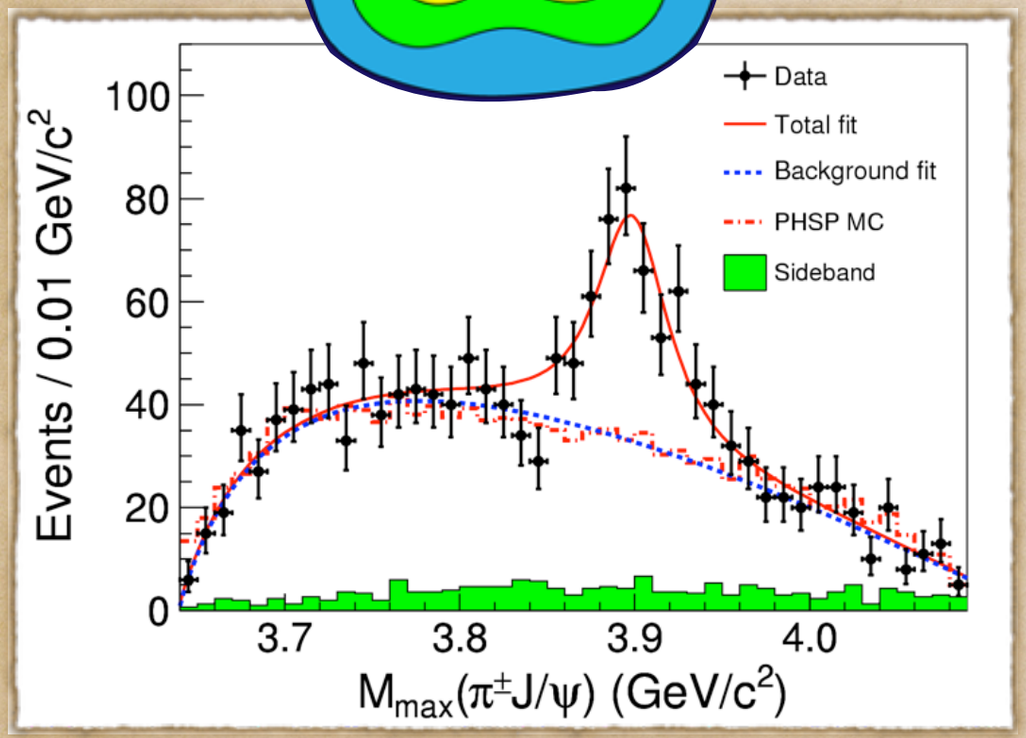
BESIII



- **Clustering/pattern recognition:** identify tracks from sensor data
- **Kalman fitter/vertex finder:** reconstruct momenta and decay vertex of each track
- **Multi-variate analysis/Machine Learning:** determine particle-type probabilities of each track, background reduction
- **Kinematic constrain fitting:** determine event topologies, improve resolutions
- **Multi-parameter fitting (Partial Wave Analysis)**: determine properties of new discovered particle
- **Monte Carlo simulations**: study efficiencies, analysis techniques, systematic errors, etc.

Basic analysis ingredients





- **Clustering/pattern recognition:** identify tracks from sensor data
- **Kalman fitter/vertex finder:** reconstruct momenta and decay vertex of each track
- **Multi-variate analysis/Machine Learning:** determine particle-type probabilities of each track, background reduction
- **Kinematic constrain fitting:** determine event topologies, improve resolutions
- **Multi-parameter fitting (Partial Wave Analysis)**: determine properties of new discovered particle
- **Monte Carlo simulations**: study efficiencies, analysis techniques, systematic errors, etc.

Challenges

Challenges

- Computational complexity: machine learning, fitting, iterative procedures, Monte Carlo studies, etc..

Challenges

- Computational complexity: machine learning, fitting, iterative procedures, Monte Carlo studies, etc..
- Data rates "high": up to 10 MHz.

Challenges

- Computational complexity: machine learning, fitting, iterative procedures, Monte Carlo studies, etc..
- Data rates "high": up to 10 MHz.
- Data processing "slow": ~second/event.

Challenges

- Computational complexity: machine learning, fitting, iterative procedures, Monte Carlo studies, etc..
- Data rates "high": up to 10 MHz.
- Data processing "slow": ~second/event.
- Data volume "large": few PBytes/year after filtering.

Challenges

- Computational complexity: machine learning, fitting, iterative procedures, Monte Carlo studies, etc..
- Data rates "high": up to 10 MHz.
- Data processing "slow": ~second/event.
- Data volume "large": few PBytes/year after filtering.
- User community "large", O(10³) and divers.

Challenges

- Computational complexity: machine learning, fitting, iterative procedures, Monte Carlo studies, etc..
- Data rates "high": up to 10 MHz.
- Data processing "slow": ~second/event.
- Data volume "large": few PBytes/year after filtering.
- User community "large", O(10³) and divers.
- "Software" development has a "low reputation".

How to deal with the challenges?

How to deal with the challenges?

- Centralized "online" processing of detector data with dedicated hardware/software: event selection -> "raw data" (RAW).

How to deal with the challenges?

- Centralized "online" processing of detector data with dedicated hardware/software: event selection -> "raw data" (RAW).
- Centralized calibration and reconstruction of raw data to "analysis objects" (ESD/AOD).

How to deal with the challenges?

- Centralized "online" processing of detector data with dedicated hardware/software: event selection -> "raw data" (RAW).
- Centralized calibration and reconstruction of raw data to "analysis objects" (ESD/AOD).
- Centralized production and pre-processing of MC data.

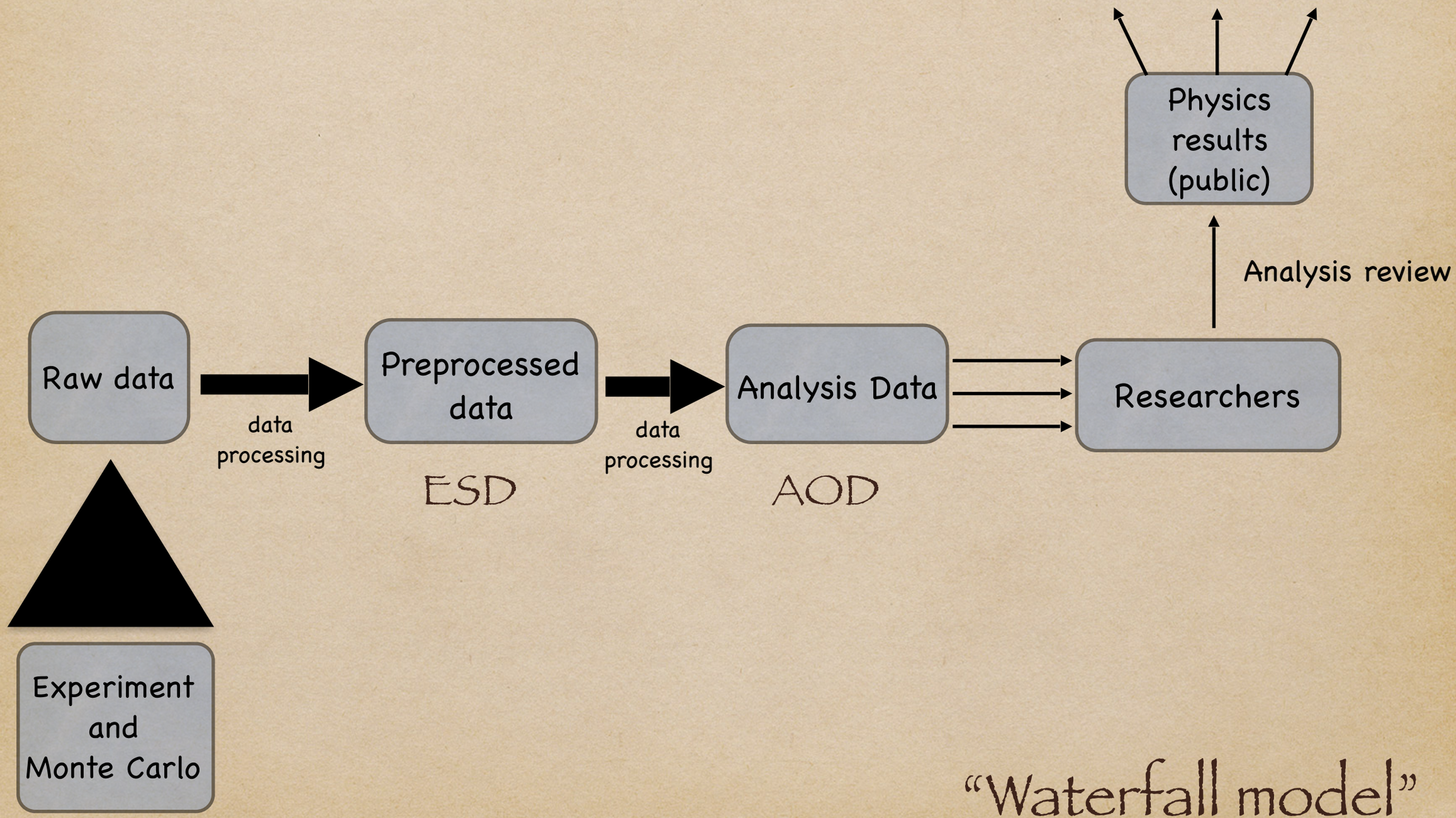
How to deal with the challenges?

- Centralized "online" processing of detector data with dedicated hardware/software: event selection -> "raw data" (RAW).
- Centralized calibration and reconstruction of raw data to "analysis objects" (ESD/AOD).
- Centralized production and pre-processing of MC data.
- Usage of common software framework (version control) and training of users to handle AOD to study their channel of interest.

How to deal with the challenges?

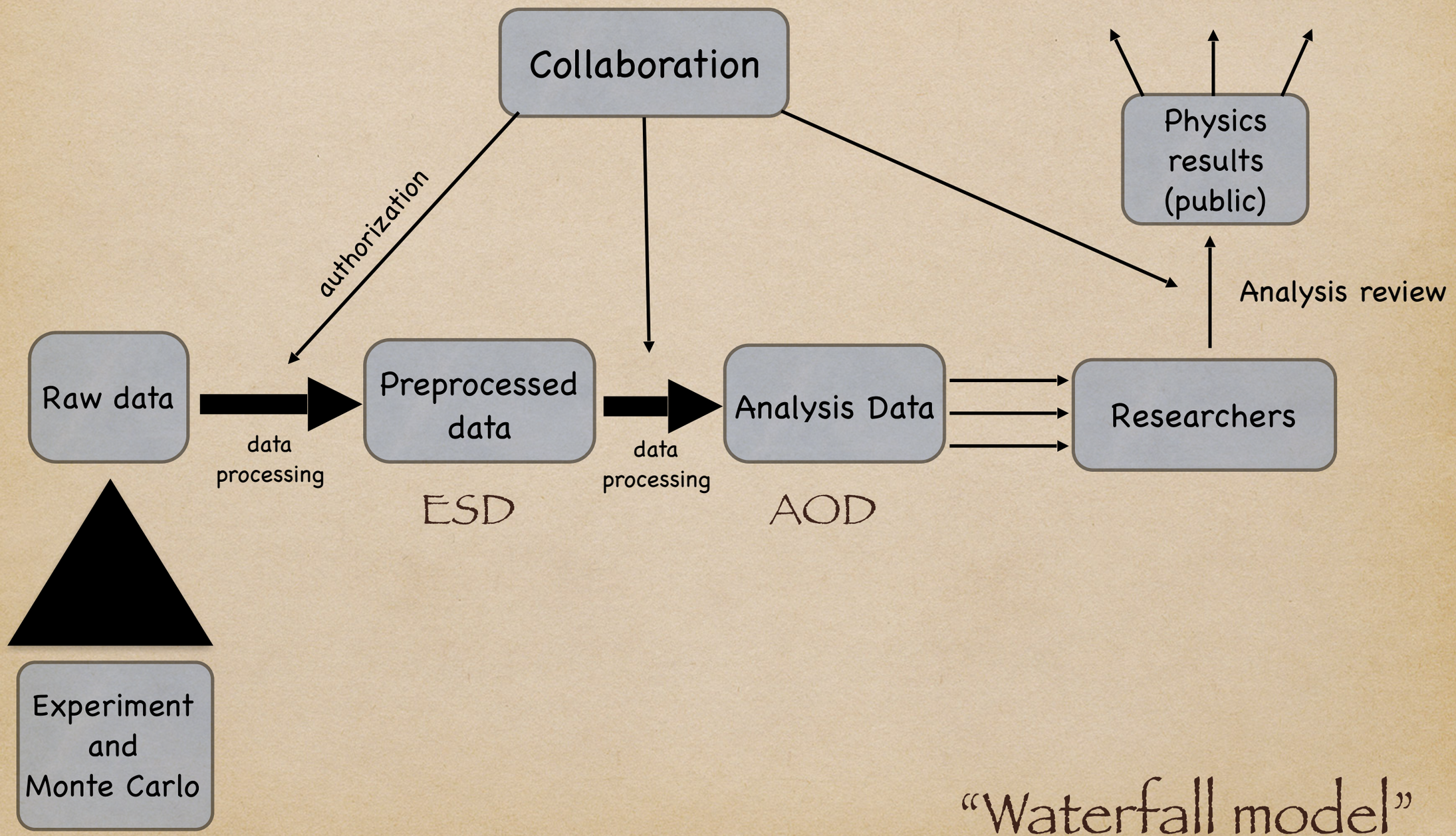
- Centralized "online" processing of detector data with dedicated hardware/software: event selection -> "raw data" (RAW).
- Centralized calibration and reconstruction of raw data to "analysis objects" (ESD/AOD).
- Centralized production and pre-processing of MC data.
- Usage of common software framework (version control) and training of users to handle AOD to study their channel of interest.
- Decentralized and distributed data of users. Validation via internal reviewing process based on analysis memos, leading to publication.

How to deal with the challenges?



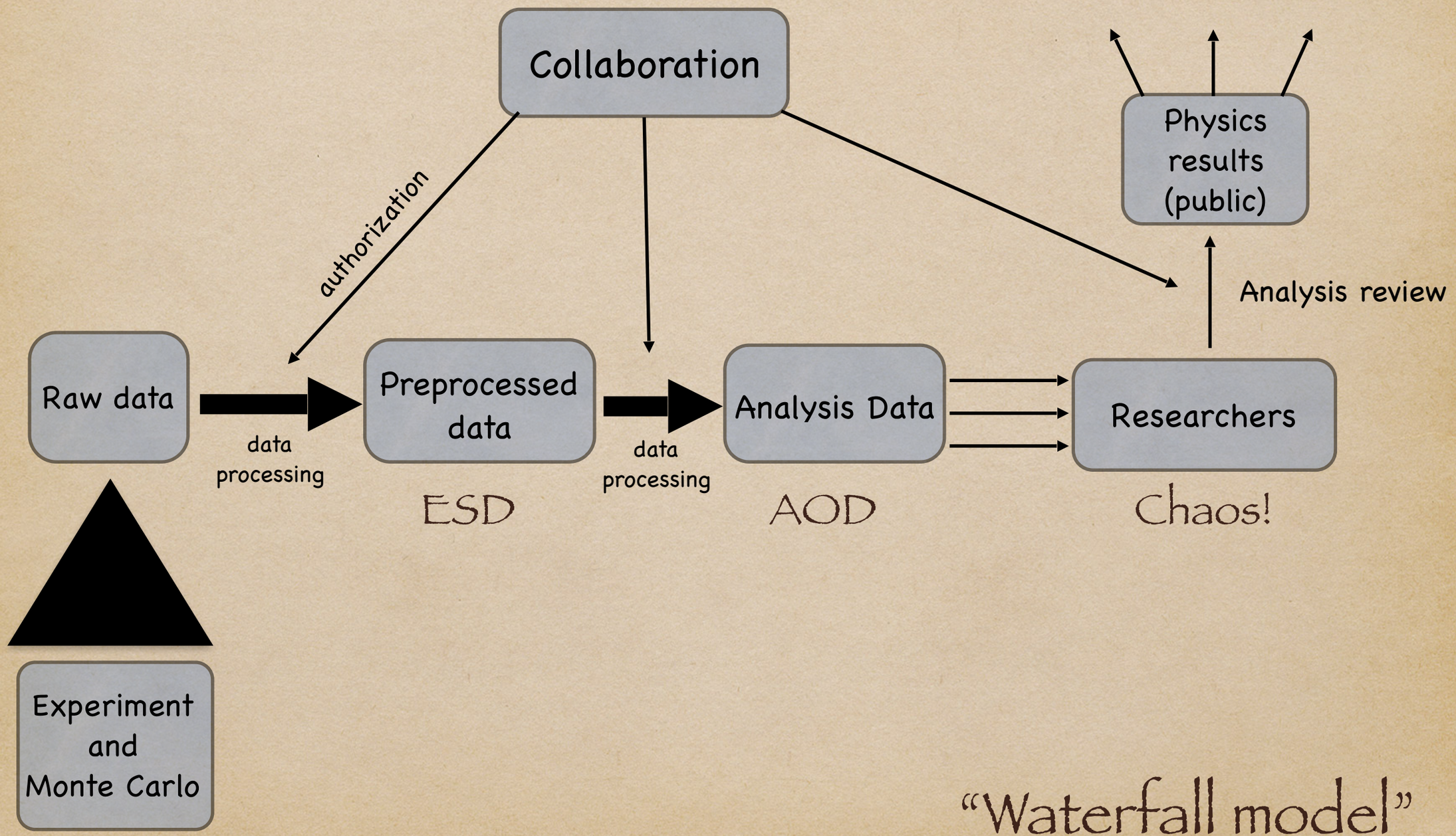
"Waterfall model"

How to deal with the challenges?



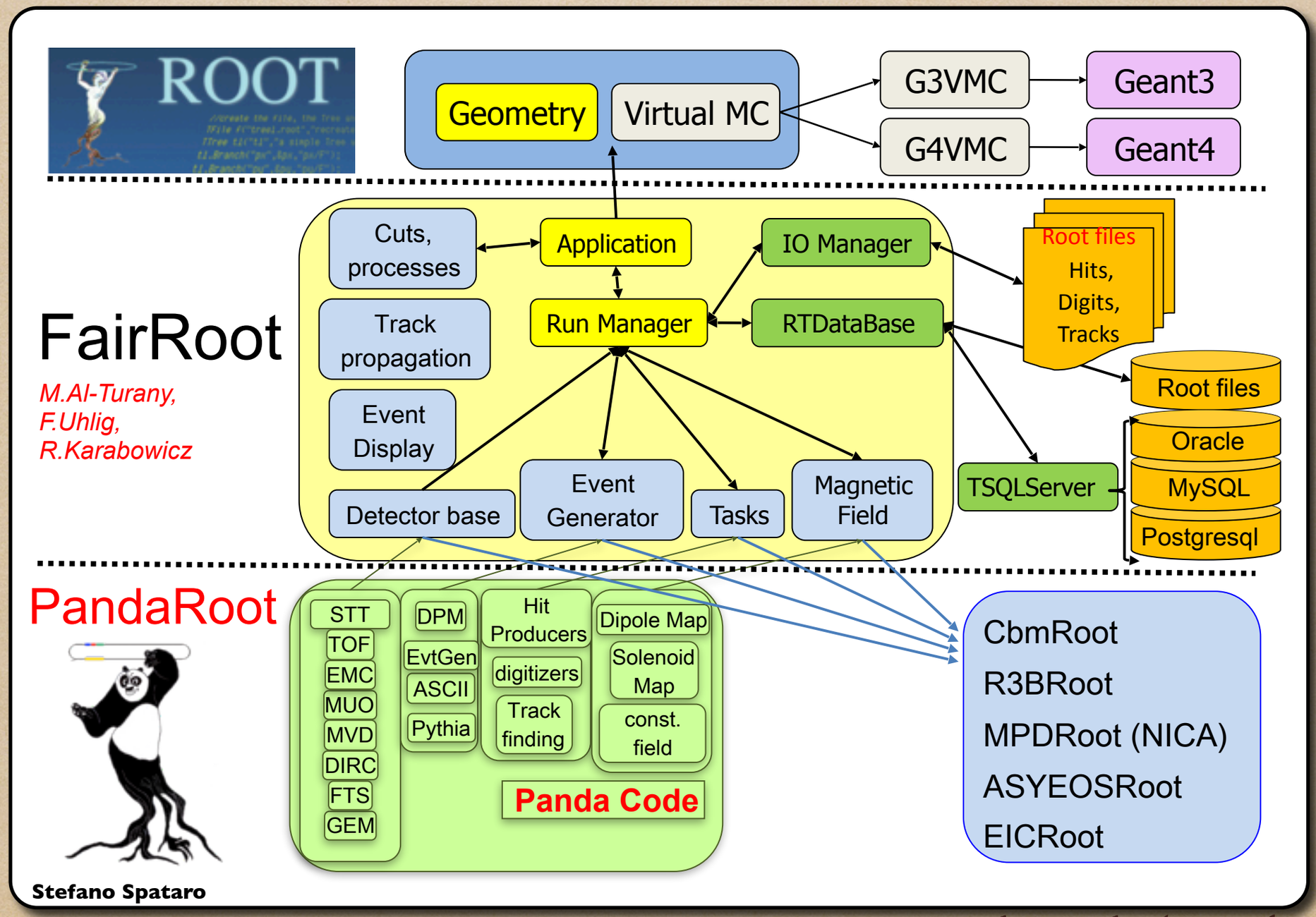
"Waterfall model"

How to deal with the challenges?



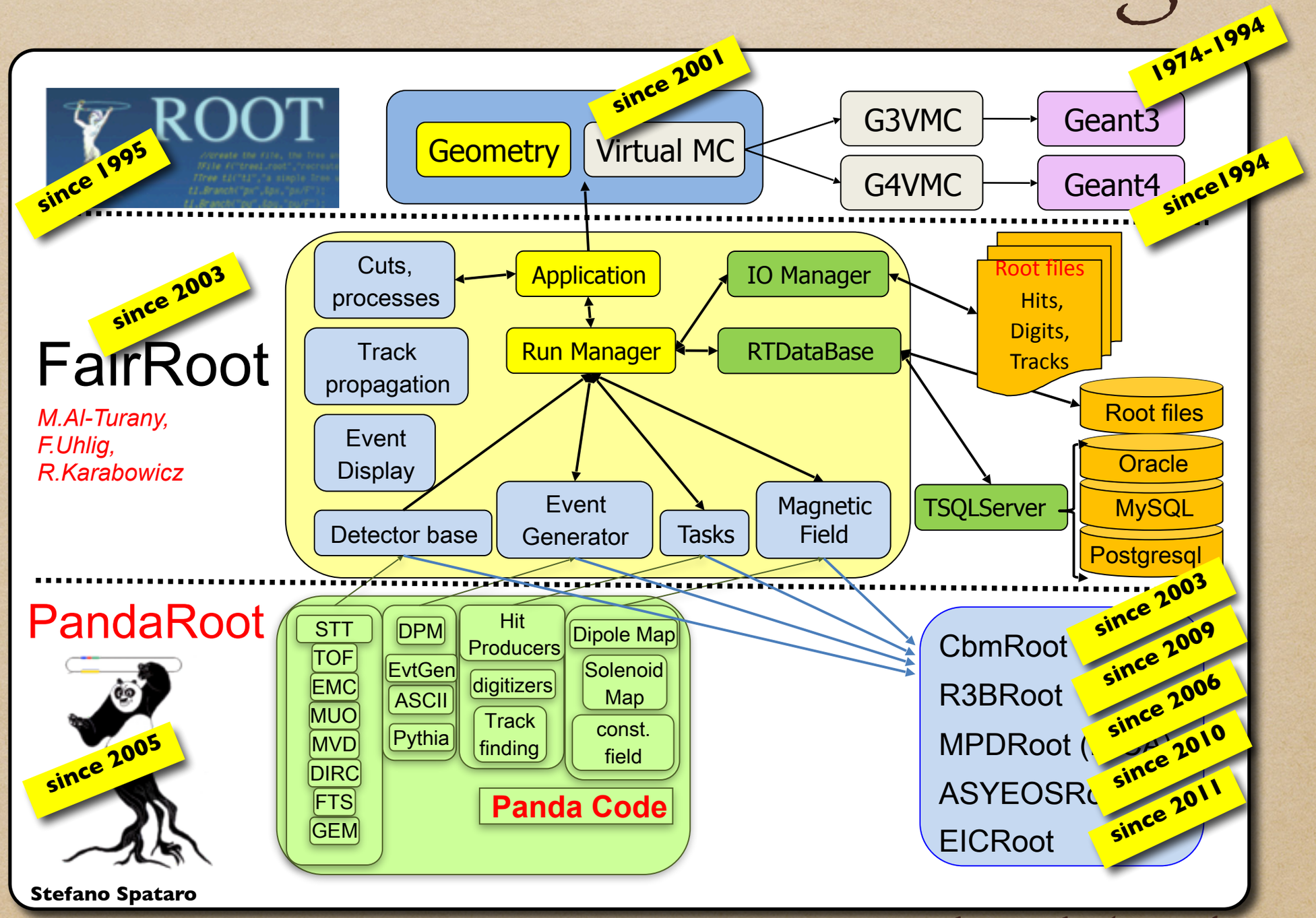
"Waterfall model"

How to deal with the challenges?



Note: outdated sketch!

How to deal with the challenges?



Note: outdated sketch!

How to deal with the challenges?

AliceO2
http://alice-o2.web.cern.ch/

CbmRoot
https://fair-center.eu/for-users/experiments/cbm.html

PandaRoot
https://panda.gsi.de/

R3BRoot
https://www.gsi.de/r3b

FairShip
http://ship.web.cern.ch/ship/

SofiaRoot

AsyEosRoot

MPDRoot
http://mpd.jinr.ru

ExpertRoot
http://er.jinr.ru/

EnsarRoot
http://igfae.usc.es/satnurse/ensarroot.html

ATTPCRootv2
https://github.com/ATTPC/ATTPCROOTv2

BNMRoot
http://mpd.jinr.ru

ALFA

- . . .
- FairMQ
- Topology Editor
- DDS (Dynamic Deployment)

FairRoot

- Testing
- Parameter/ Condition Management
- Geometry
- Detector Response
- Magnetic Field
- Event Generators

Software Stack

- ROOT
- Boost
- Simulation Engines
- VGM
- CMake
- Network (ZeroMQ, nanomsg, OFI libfabric)
- Serialization (Protocol Buffers, msgpack, FlatBuffers, ROOT)
- Google Test
- . . .

2018

M. Al-Turany

Limitation of computing model (and where ESAP could help)

Limitation of computing model (and where ESAP could help)

- Analysis methods and results by individual researchers not optimally shared. Reproducibility?

Limitation of computing model (and where ESAP could help)

- Analysis methods and results by individual researchers not optimally shared. Reproducibility?
- Lack of a common user-data management.

Limitation of computing model (and where ESAP could help)

- Analysis methods and results by individual researchers not optimally shared. Reproducibility?

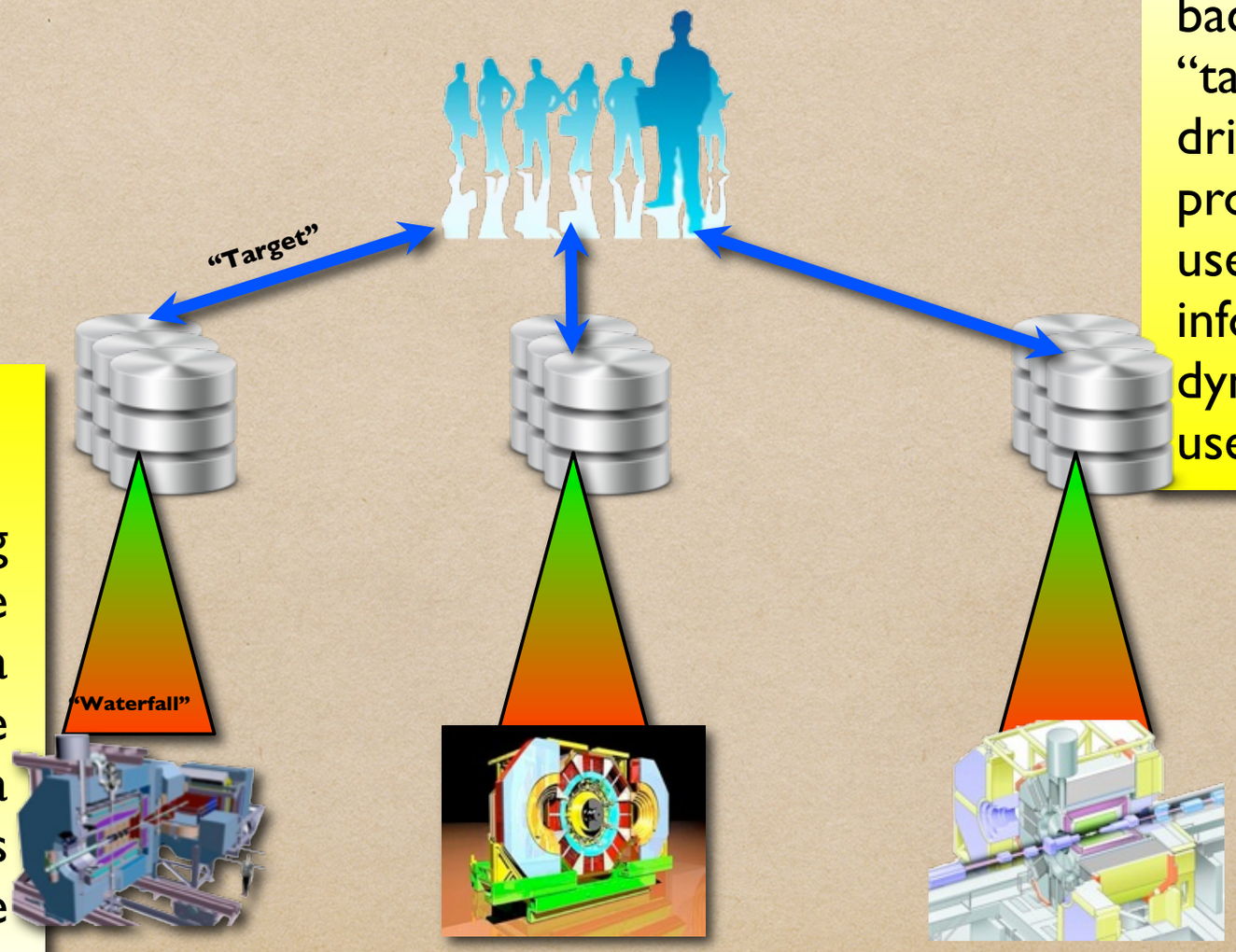
- Lack of a common user-data management.

- Analysis workflow concentrated to one single experiment, no real export/import mechanism from other experiments foreseen.

Limitation of computing model (and where ESAP could help)

- Analysis methods and results by individual researchers not optimally shared. Reproducibility?
- Lack of a common user-data management.
- Analysis workflow concentrated to one single experiment, no real export/import mechanism from other experiments foreseen.
- Missing the concept of a modern information system.

Limitation of computing model (and where ESAP could help)



"Target"

backward chaining
"target" architecture
driven by user query
process on-the-fly
users pull data
information system
dynamic archive
user data is sacred

"Waterfall"

forward chaining
"tier" architecture
driven by raw data
process in pipeline
operators push data
results in releases
static archive
raw data is obsolete

FAIR test case

FAIR test case

- Deploy ESCAPE-EOSC on analysis data!

FAIR test case

- Deploy ESCAPE-EOSC on analysis data!
- "Realistic" Monte Carlo data available.

FAIR test case

- Deploy ESCAPE-EOSC on analysis data!
- "Realistic" Monte Carlo data available.
- Exploit "data lake" prototype (WP2) between GSI and RuG.

FAIR test case

- Deploy ESCAPE-EOSC on analysis data!
- "Realistic" Monte Carlo data available.
- Exploit "data lake" prototype (WP2) between GSI and RuG.
- Good news: we have time for proof-of-principle study!