The Data Preparation Group
- Organization and Plan of Work -

ALICE Offline Week, 30 March – 01 April 2016

C. Zampolli

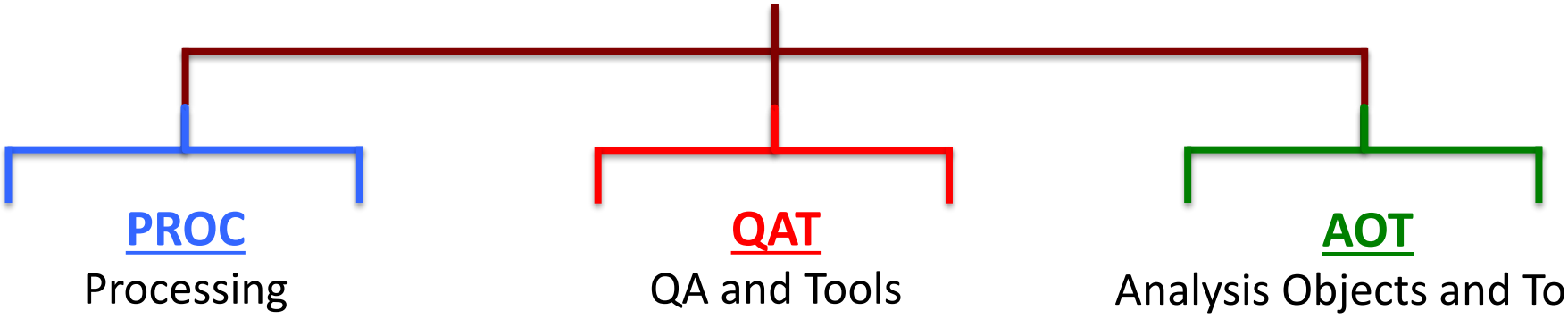
What is DPG?

- **Data Preparation Group**: new entity created to organize, plan, verify, certify, monitor the steps of the data processing

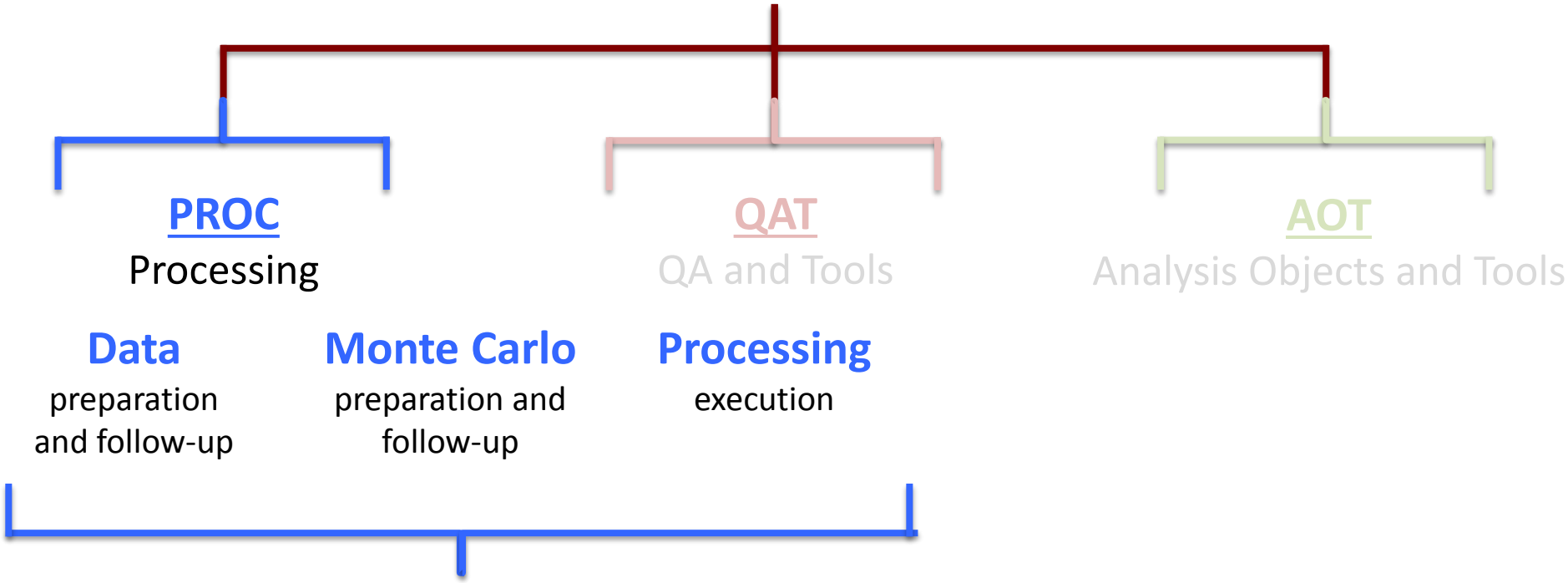
  → **DELIVERABLE: a "black box" with usable data**

- The DPG will function as a **glue** between the several different parties involved in the data processing: intermediate, report, follow up, collect input and feedback, contact, sort issues
  - RC, detectors, physics and PB, offline and CB, production management, Quality Assurance
- **DPG** (and BTG) will take over several of the functions of PWGPP
- **DPG** will part of the ALICE Offline Project
  - It might become a project in itself (DPP) in the long term, together with BTG and other

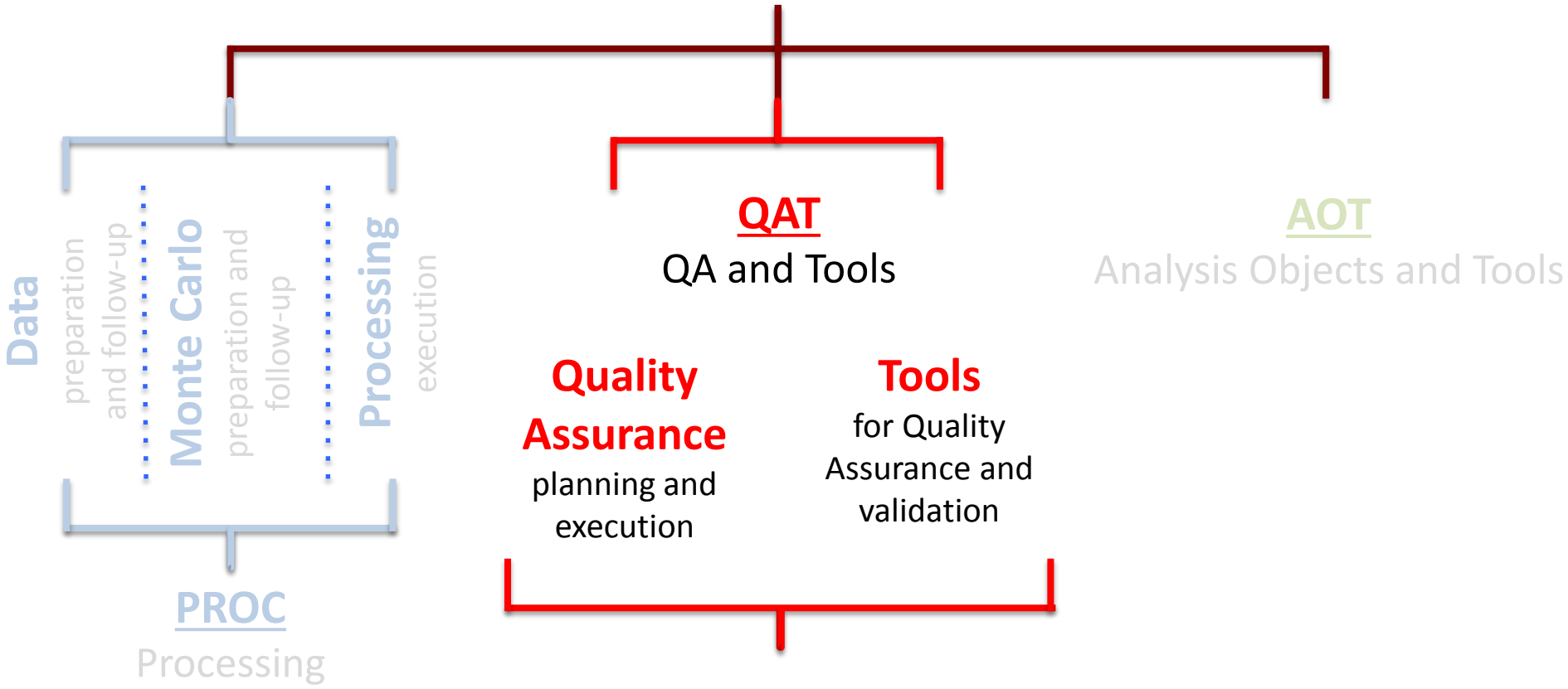
DPG Structure



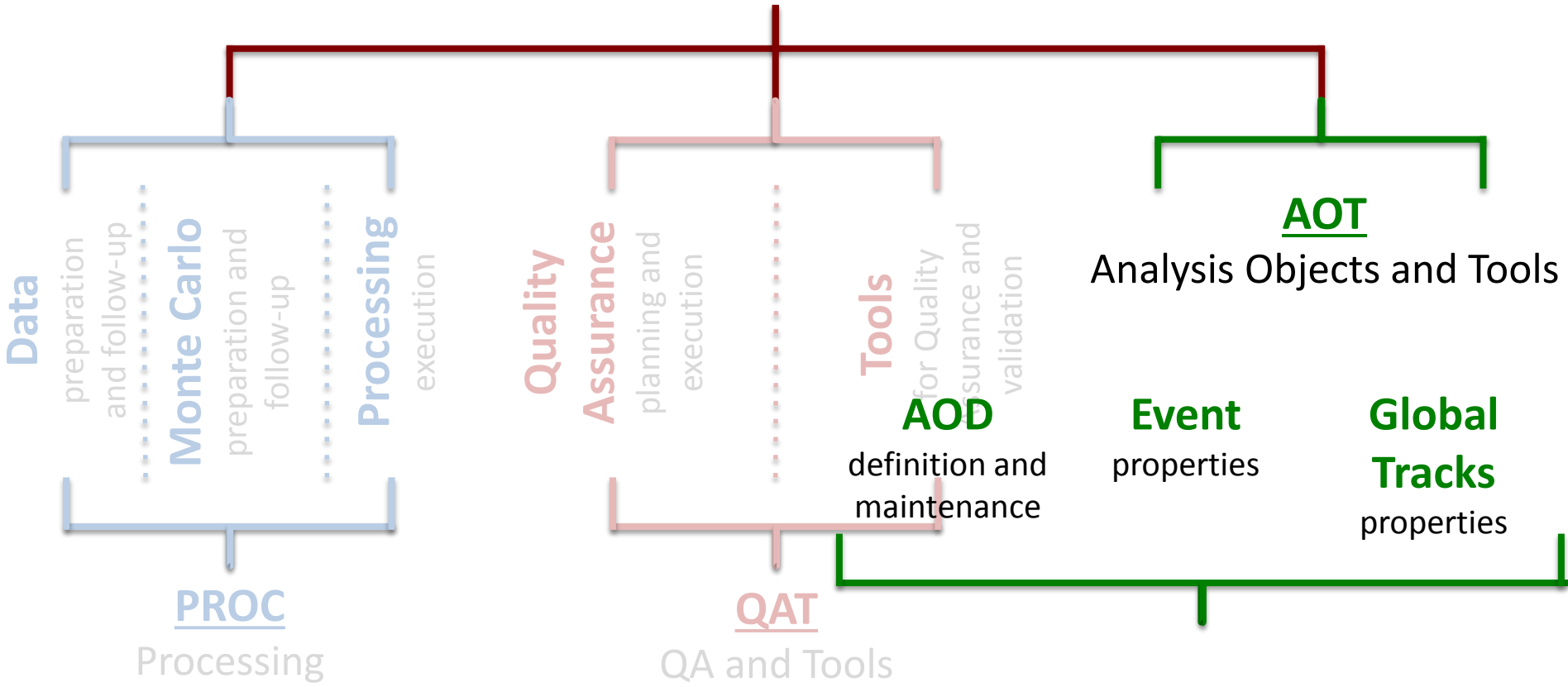
DPG Structure



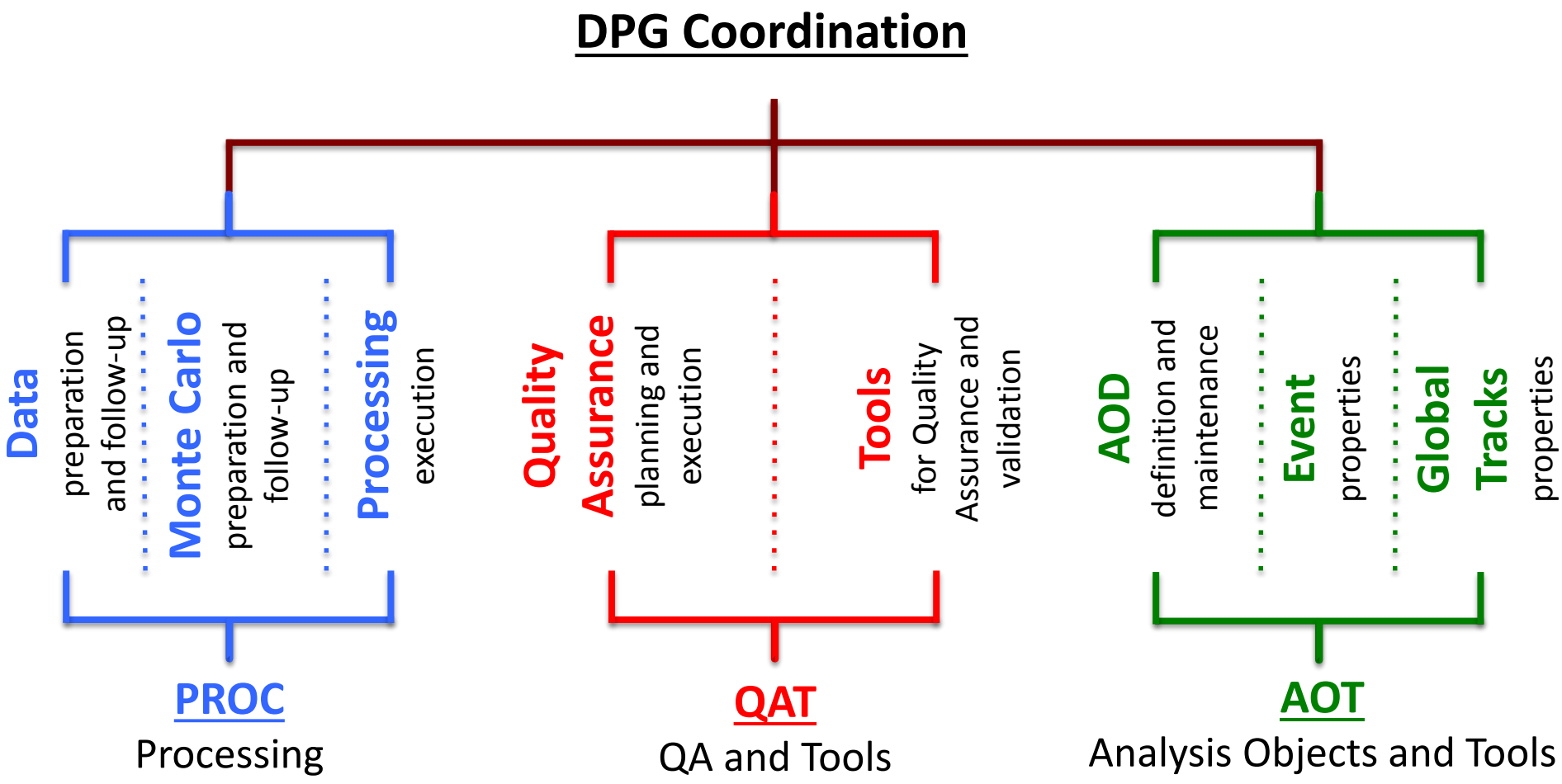
DPG Structure



DPG Structure



DPG Structure



DPG: PROC

Data

- Definition of data to be processed
- Triggers for calibration
- Definition of runs to be reprocessed
- Definition of priorities for datasets to be (re)processed with PB
- Follow-up of incidents
- Follow-up of Release validation for reco+calib
- Shuttle
- Participation in weekly RC meetings
- Calibration readiness
- Maintenance of code for reco+calib (macros + scripts) including muon_calo passes

Monte Carlo

Execution

DPG: PROC

Data

- Definition of data to be processed
- Triggers for calibration
- Definition of runs to be reprocessed
- Definition of priorities for datasets to be (re)processed with PB
- Follow-up of incidents
- Follow-up of Release validation for reco+calib
- Shuttle
- Participation in weekly RC meetings
- Calibration readiness
- Maintenance of code for reco+calib (macros + scripts) including muon_calo passes

Monte Carlo

- Definition of MC sets to be produced (together with PB)
- Development of MC configuration tools (including snapshot)
- Follow-up of MC QA
- Comparison with performance from data (tracking, PID, centrality…)
- Calibration for MC, tuning to data

Execution

DPG: PROC

Data

- Definition of data to be processed
- Triggers for calibration
- Definition of runs to be reprocessed
- Definition of priorities for datasets to be (re)processed with PB
- Follow-up of incidents
- Follow-up of Release validation for reco+calib
- Shuttle
- Participation in weekly RC meetings
- Calibration readiness
- Maintenance of code for reco+calib (macros + scripts) including muon_calo passes

Monte Carlo

- Definition of MC sets to be produced (together with PB)
- Development of MC configuration tools (including snapshot)
- Follow-up of MC QA
- Comparison with performance from data (tracking, PID, centrality…)
- Calibration for MC, tuning to data

Execution

- Running of data processing (from reco to AOD)
- Running of MC processing (from reco to AOD)
- Triggering of QA, follow-up…
- OCDB uploads
- Development of tools for OCDB upload traceability
- Maintenance of code for reco+calib (macros + scripts) including muon_calo passes

DPG: PROC

Data

- Definition of data to be processed
- Triggers for calibration
- Definition of runs to be reprocessed
- Definition of priorities for datasets to be (re)processed
- Follow-up of incidents
- Follow-up of Release validation for reco+calib
- Shuttle
- Participation in weekly RC meetings
- Calibration readiness
- Maintenance of code for reco+calib

Monte Carlo

- Definition of MC sets to be produced (together with PWG)
- Development of MC configuration tools (including anchor)
- Evaluation of MC QA
- Comparison with performance from data (tracking, PID, centrality…)
- Redefinition of MC naming scheme?

Execution

- Running of data processing (from reco to AOD)
- Running of MC processing (from reco to AOD)
- Triggering of QA, follow-up…
- OCDB uploads
- Development of tools for OCDB upload traceability

Data and Monte Carlo associated productions should go in **parallel** as much as possible

- E.g.: during data taking, when final calibrations are not yet available, MC setting up, tests… so that when all calibrations are ready, MC can start

DPG: QAT

Quality Assurance

- Organization and follow-up of QA from Data processing
- Organization and follow-up of QA from MC processing
- Validation of data and MC
- Creation of run-lists for “basic” analyses (tracklets, tracks, tracks+PID, muon, calo)
- Definition of main properties to characterize data and MC
- Report to RC

Tools

DPG: QAT

Quality Assurance

- Organization and follow-up of QA from Data processing
- Organization and follow-up of QA from MC processing
- Validation of data and MC
- Creation of run-lists for “basic” analyses (tracklets, tracks, tracks+PID, muon, calo)
- Definition of main properties to characterize data and MC
- Report to RC

Tools

- Development of regression tests
- Development of functional tests
- Development of a tool to automatically compare productions (any: data vs data, MC vs MC, data vs MC)
- Definition of QA monitoring tools (logs, databases)
- Visualization of QA information (histograms, trending…)
- OCDB regression tests?

DPG: AOT

AOD

- Definition of cuts to create AOD sets
- Follow-up of AOD creation
- AOD refiltering configuration and documentation
- Bookkeeping of AOD creation (configuration)
- Common interface between AOD and ESD
- Improvement in data structure? (for Run3)

Event Properties

Track Properties

DPG: AOT

AOD

- Definition of cuts to create AOD sets
- Follow-up of AOD creation
- AOD refiltering configuration and documentation
- Bookkeeping of AOD creation (configuration)
- Common interface between AOD and ESD
- Improvement in data structure? (for Run3)

Event Properties

- Event selection maintenance (and improvements… AliBits), including documentation, follow-up of issues
- Pileup studies with documentation (*)
- Vertexing studies (*)
- Centrality (*) documentation, calibration, automatization
- Porting Event selection to AOD (*)

Track Properties

(*) implementation is part of Offline?

DPG: AOT

AOD

- Definition of cuts to create AOD sets
- Follow-up of AOD creation
- AOD refiltering configuration and documentation
- Bookkeeping of AOD creation (configuration)
- Common interface between AOD and ESD
- Improvement in data structure? (for Run3)

Event Properties

- Event selection maintenance (and improvements… AliBits), including documentation, follow-up of issues
- Pileup studies with documentation (*)
- Vertexing studies (*)
- Centrality (*) documentation, calibration, automatization
- Porting Event selection to AOD (*)

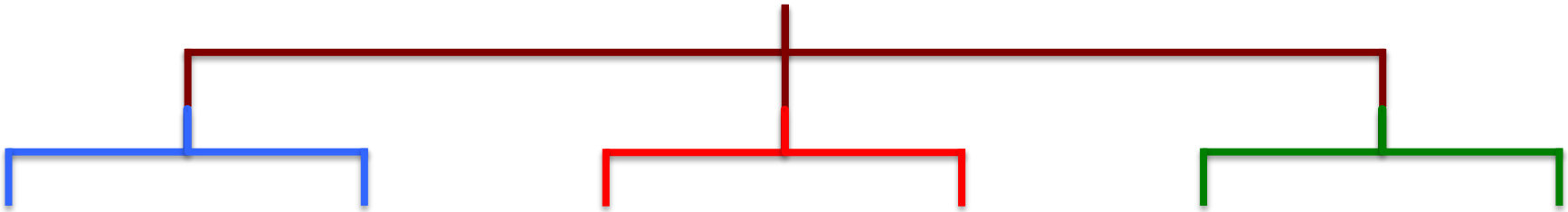
Track Properties

- Track cuts studies and recipes per periods
- Evaluation of common systematics due to track cuts variation
- PID performance studies

(*) implementation is part of Offline? Distinction between AliRoot and AliPhysics

DPG meetings

DPG Coordination



PROC
Processing
→ Internal meetings

QAT
QA and Tools
→ Internal meetings and usual QA meetings

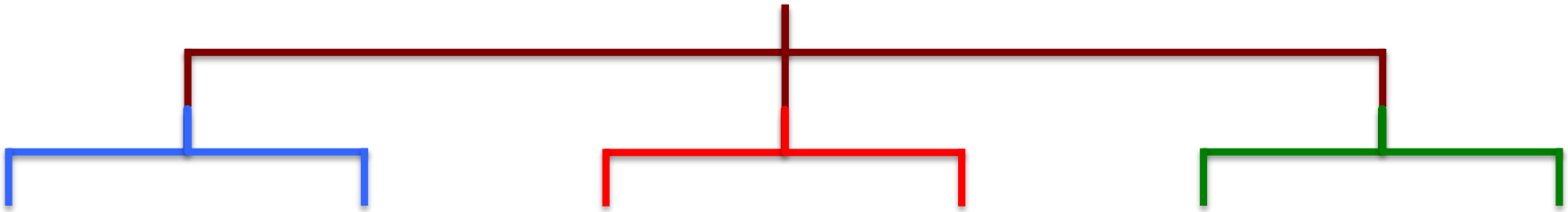
AOT
Analysis Objects and Tools
→ Internal meetings

Weekly closed meetings to discuss:

- Acceptable/unacceptable defects (to be proposed to PB)
- Run lists
- Status of and readiness for productions
- Datasets definition
- Collection of bug and problem reports
- Production prioritization (to be discussed with PB)
- Status of the release for production

DPG: Deliverable

DPG Coordination



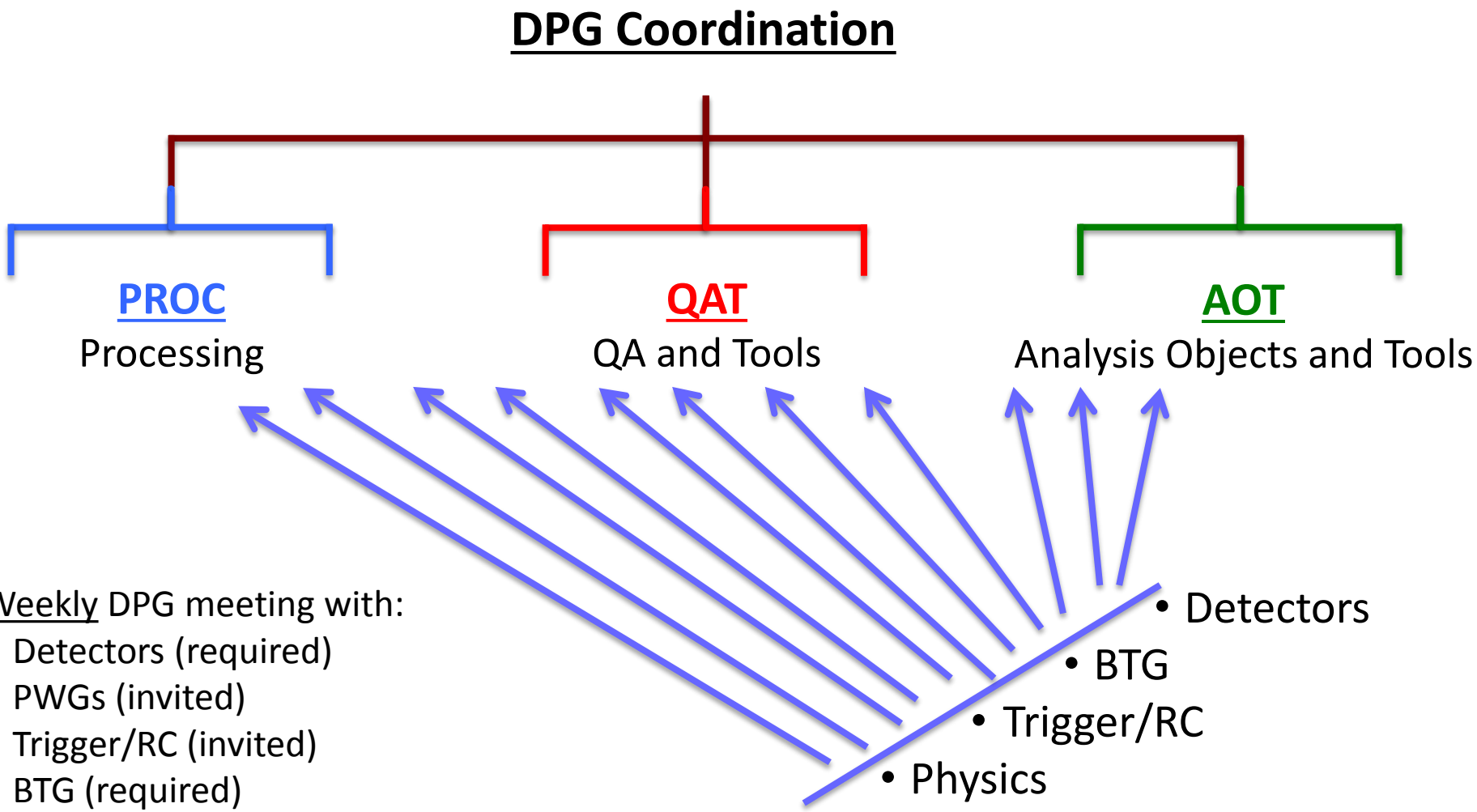
PROC
Processing

QAT
QA and Tools

AOT
Analysis Objects and Tools

- Database to bookkeep productions (MC ←→ data, run lists, AODs) with corresponding settings (aliroot versions used to produce data/needed to analyze data, macros, configurations…), usability and quality/performance needs to be defined and maintained
  - Extension of RCT functionalities
  - Now often only at PWG level → not shared among everyone!
  - Needed also for past productions

DPG and the "others"



Weekly DPG meeting with:

- Detectors (required)
- PWGs (invited)
- Trigger/RC (invited)
- BTG (required)
- HLT (invited)

Role of the Detectors

- 1 contact per detector
  - Should report about the status of the corresponding detector in terms of calibration, alignment, reconstruction, simulation, PID
    - QA will be covered in the QA meetings
  - Will start with those for Computing Board
    - The work could be delegated to one expert of the above activities

Role of the Detectors

- 1 contact per detector
  - Should report about the status of the corresponding detector in terms of calibration, alignment, reconstruction, simulation, PID
    - QA will be covered in the QA meetings
  - Will start with those for Computing Board
    - The work could be delegated to one expert of the above activities
- 1 week of time for feedback, if none, PB will be informed

To be added afterwards, indicates feedback

	Production name					
Activity / Detector	Calib	Align	Reco	PID	Sim	QA
det1	✗	✓	✗	✗	✗	✗
det2	✓	✓	✓	✓	✗	✗

Default!

Dataset Characterization

- Should be collected in a summary table

Production (data should be used as anchor name)											
Activity / Dataset	Tracklets	Global tracks	Global tracks PID	Muon tracks	Calo tracks	Vtx	Pileup	Ev. selection	Track (barrel) selection	Centrality	Certification
pass1	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
pass2	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
MC1	…	…	…	…	…	…	…	…	…	…	…
MC2	…	…	…	…	…	…	…	…	…	…	…

Default!

Dataset Characterization

- Should be collected in a summary table

Production (data should be used as anchor name)											
Activity / Dataset	Tracklets	Global tracks	Global tracks PID	Muon tracks	Calo tracks	Vtx	Pileup	Ev. selection	Track (barrel) selection	Centrality	Certification
pass1	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
pass2	✓	✓	✓							✗	✗
MC1	…	…	…							…	…
MC2	…	…	…							…	…

Summary of what the dataset can be used for:

- Tracklets
- Global tracks
- Global tracks + PID
- Muon
- Calo
- …

Need to define the criteria for certifications

Reprocessing

- Based on accurate planning
  - Decision of what new features to include made when the latest reprocessing starts together with PB
    - Proposals from detector groups
- Happening at a fixed time of the year
  - E.g. February
    - Developments during the year, together with data taking
    - Readiness by December
    - January left to fix last “problems”, and test
  - Whatever is not ready, will be not waited for
    - In agreement with PB
- How will this impact Run2? Maybe mainly for Run3…

Channels

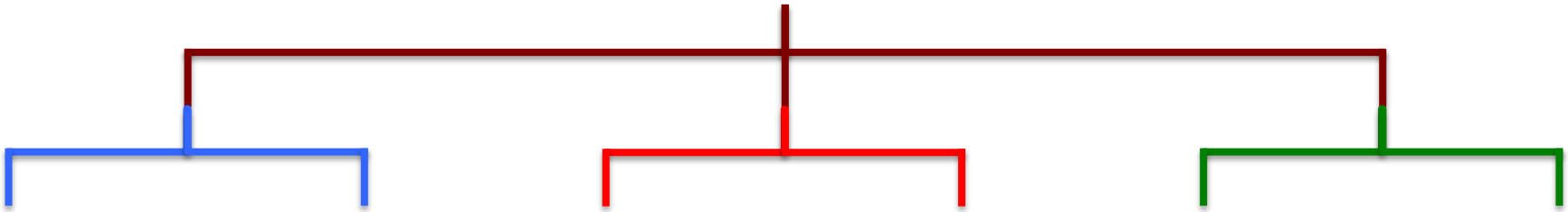
- Communication should be made as efficient as possible
- Repetition should be avoided
- Private communication is a killer
- Common documents (e.g. Google docs), JIRA, HyperNews, Twiki, databases
  - Interface between JIRA and common database

Tasks

- Some of the tasks within DPG can be assigned in form of shifts
  - QA activities
  - AOD production
  - Database maintenance
- Need to involve Institutes → institutional responsibilities
- Need documentation to ensure continuity and maintainability

DPG People

DPG Coordination Chiara Zampolli + deputy(ies)



PROC
Processing
Roberto Preghenella
Catalin-Lucian Ristea (*)
Chiara Zampolli

QAT
QA and Tools
Marie Germain
Jacek Otwinowski (*)

AOT
Analysis Objects and Tools
People currently involved in equivalent PWGPP activities

Especially here more people needed to undertake tasks

(*) Institutional responsibilities → for long-term tasks, this is a key!

Summary and Conclusion

- **Data Preparation Group** within Offline will have the role to organize the data and Monte Carlo processing, monitor the quality, prepare the tools and provide the end user with the input for their analysis
- Development will be outside the scope of the DPG
- Several topics will be dealt with within DPG
  - Needs collaboration and communication between people
  - People will be assigned to tasks and will be expected to report
- The proposed structure will start being operational now
  - Experience will tell us in which direction to go to improve it, what is missing, what is not needed…
  - The first period will be the most demanding, but it should converge to a stable running mode