Data Analysis and Interpretation: Charge and Scope

- **Charge:** Look at the future of analysis at scales up to that of the HL-LHC
  - Future analysis models
  - Sociology of data access in groups
  - Challenges and opportunities surrounding analysis code development and infrastructure
  - Approaches to preserve and disseminate knowledge

- **Scope:** Current and future analysis techniques and the toolkits that support them
  - Consider how factors like data volume, type and parallelism of data access, and number of parallel analyses impact the ability to carry out a comprehensive physics program.

- **Not a goal:** Reach community consensus
- **Goal:** Create a roadmap for R&D projects
  - Experiments can develop an informed plan for the HL-LHC era (pick and choose)
  - Create a framework for contacts and discussions to work together on similar aspects of the R&D roadmap

Key challenges and opportunities

- Baseline analysis model: utilize successive stages of data reduction, in the end analyzing a compact dataset with quick real time iteration
  - No consensus on where the line sits between managed production-like analysis processing and individual analysis
  - Maybe intermediate steps and the output of a final “laptop” dataset will not be necessary in the future

- Challenge for all analyses: reduce the “time to insight” while exploiting the maximum possible scientific potential of the data
  - **Interactivity**: Important for analysis, enables quick turn-around for “dialog with data”
  - **Elasticity**: Many analyses have common deadlines defined by conference schedules
  - **Heterogeneity**: Heterogeneous computing hardware can be exploited to reduce the “time to insight”

HEP Analysis Ecosystem

- ROOT and its ecosystem both dominate HEP analysis
  - Impacting the full event processing chain, providing foundation libraries, I/O services, etc.

- This is a significant advantage for the HEP community compared over other science disciplines
  - Provides a fully integrated and validated toolkit
  - Enabling the community to talk a common analysis language
  - Improvements and additions become quickly available to the whole community

- However: open source analysis tools landscape used in industry is evolving very fast
  - Surpasses the HEP efforts both in total investment and size of community

- ROOT should evolve from a toolkit to a toolset
  - Community feels a more modular setup and installation is needed
  - Bridges to external open source analysis tools like external machine learning and deep learning toolkits
  - Ferries to shuttle data efficiently to these external tools from the ROOT foundation layer

- Need evolution of the community to sustain the ecosystem
  - Software life cycle and sustainability: modules, bridges and ferries become dependencies
  - Maintain the ecosystem: more effort, distributed more broadly

Analysis Languages

- Reconstruction and other performance-critical code ➜ C++
- Analysis code: python developed into the language of choice
  - Outside the HEP community and more and more within
- Python should become a first class language in HEP (like C++)
- Functional or declarative programming model
  - Instead of defining the “how”, the analysts would declare the “what” of their analyses
    - Essentially removing the control of the event loop
    - Leave it to underlying services and systems to optimally iterate over events
  - High-level approach: more freedom in optimizing the utilization of diverse forms of computing resources

Analyzing Data

- Minimizing the “time to insight”: I/O performance becomes driving factor
  - Data is made available to the community for analysis after most of the CPU-intensive processing steps are completed centrally

- Disk space is usually the key concern of the experiment computing models
  - Disk is the most expensive computing hardware

- Continue R&D in file formats, compression algorithms, and new ways of storing and accessing data for analysis
  - Investigate optimizing the storage systems and data representations together
  - Facilitate utilization of new additional storage layers like SSD storage and NVRAM-like storage
    - These have different characteristics compared to spinning disk

- Access to non-event data for analysis (cross section values, scale factors, tagging efficiencies, …)
  - ROOT TTree enabled easy storage of event data
  - Need a similar way of storing and accessing non-event information during analysis

Analysis Models

- Baseline analysis model: subsequent data reduction steps
  - Minimize “time to insight” for a large number of parallel analyses by optimizing reusability of intermediate outputs

- Reduce the need for storing intermediate steps and increase interactivity of analyzing large amounts of data
  - Future hardware infrastructure and new technologies might make intermediate steps unnecessary.
  - Output of a final “laptop” dataset could not be necessary using these new approaches

- Late stage analysis will neither be entirely local nor entirely remote
  - Need to support both, e.g. support syncing from remote services (like notebook based analysis-as-a-service) into a local laptop environment to support ‘airplane mode’
  - Analysis should scale easily from laptop to cloud/grid resources to special resources
  - Automatic failure recovery is of great importance: most time is currently spent on recovering the last 1%

- Reproducibility needs to be taken very seriously from the start
  - Heterogeneous hardware and diverse set of techniques and technologies complicates reproducibility
  - Reproducibility needs to be a fundamental component of the system as a whole

Analysis Roadmap

Community topics

1. Sustainability of HEP analysis ecosystem
2. Reproducibility

Short term R&D

3. Python
4. Toolkit ➜ Toolset: modular installation and setup

Medium term R&D

5. Bridges and ferries
6. File formats
7. Impact of heterogeneous hardware technologies on analysis
8. Non-event data TTree

Long term R&D

9. Analysis models
10. Functional or declarative analysis programming approaches

Practical Consideration for Progress in the WG Area

- Ecosystem is a big advantage but needs continued support
  - Community should come together and invest more effort that is also more broadly distributed

- Modularity is the key to easy integration of new components and ease of installation
  - Analysis toolset will not only be distributed centrally, users will want to install pieces of it themselves
  - Community should define rules for life cycle of modules of the ecosystem

- Many industry trends: some short lived, some with more long term application
  - Need to stay vigilant and define metrics for success of analysis using new technologies
  - Need to be open for even newer technologies down the road and retain ability to integrate them and use them

- Analysis depends on developing and trying out new ideas
  - Increases in data volumes threaten to reduce interactivity further ➜ time to “experiment” with the data increases
  - Finding new, fast, interactive ways to analyze large data volumes ➜ Enable creativity of community and scientific success

Commonality and Leveraging S&C beyond HEP domain

- We are already using a common analysis toolkit across all experiments
  - Challenges
    - Sustainability of ecosystem
    - Maintenance and lifecycle of components

- Industry has a wide variety of analysis tools and continues to invest heavily
  - All these open source technologies we could use
  - Key is easy integration and maintaining/providing what is unique to our field in our toolset

Cross-cutting Elements

- Analysis touches many areas and WGs
  - Computing Models, Facilities, and Distributed Computing WG
  - Data Access and Management WG: data management is the key enabler of analysis
  - Event Processing Frameworks WG: event data model
  - Data and Software Preservation WG: reproducibility

- Synergistic activities
  - Analysis model development and testing could be done together with data management R&D
  - Industry and other science fields are good partners in all analysis investigations
    - HEP has to offer very attractive use cases that push industry systems

CWP Chapter Status and Plans

- What is the status of the CWP Chapter for this working group? Are the key ideas and R&D in place?
  - CWP chapter: http://tinyurl.com/y9xrhphx
  - First comments received, light on concrete R&D but roadmap exists

- What additional work is required to get the prose in good shape for a viable CWP chapter and for others outside of your WG to read and comment?
  - Comments and improvement of the draft

- How do you plan to complete your chapter?
  - Use parallel sessions to improve draft, check for missing content, discuss cross-cutting topics with other WGs
  - Follow up after workshop and complete

- What do you expect to accomplish by the end of this workshop?
  - Finalize content and discussions so that what only remains is to polish the draft

Auxiliary Material

Primary Activities

- Group Meeting
  - 1. March 2017: https://indico.cern.ch/event/617708/

- Amsterdam Analysis Ecosystem Workshop
  - HEP Analysis Ecosystem Workshop Report
  - https://indico.cern.ch/event/613842
  - Many thanks to the organizers and everyone who participated

Data Analysis and Interpretation WG

Primary organizers of the WG are:

1. Oliver Gutsche
2. Mark Neubauer