The Unseen: revealing the blind production procedure and experience for NP data

Jérôme Lauret & Gene Van Buren – Brookhaven National Laboratory, Upton, NY, USA

ACAT 2021
Daejeon, South Korea

Introduction:
A unique experiment was conducted by the STAR Collaboration in 2018 to investigate differences between collisions of nuclear isobars, a potential key to unraveling one of the physics mysteries in our field: why the universe is made predominantly of matter. Enhancing the credibility of findings was deemed to hinge on blinding analyzers from knowing which dataset they were examining, necessitating efforts by the data production team to investigate and implement new (in our field) blinding practices. With nearly two decades of established machinery intended to provide open data and metadata access in STAR, the breadth of details to consider for a successful blinding process was substantial.

Meta-data:
Data-taking was conducted with full knowledge of colliding species, necessary to perform optimal communication at all levels of operation, where any mistakes were most costly. Database contents and run information were kept open while raw data and their products were restricted. It was therefore critical to prevent details in the data from being correlatable to the meta-data.

Selecting runs:
Operating conditions at the experiment varied over multiple time scales, offering a potential tell-tale if observable in the data for one species more than the other. A clear example is a dead detector channel repaired after a single fill of the collider, thus present for only one species. The excluded runs were more for one species (7% vs. 4%), but analyzers did not know exclusion criteria, and thus not which one lost more.

DSTs for analyzers:
Numerous pieces of information in the DSTs that analyzers would be given had the potential to indicate collider condition, data-taking time, or a number tied to a specific run. During production, we obscured or zeroed ("shadowed") information before filling DSTs on:

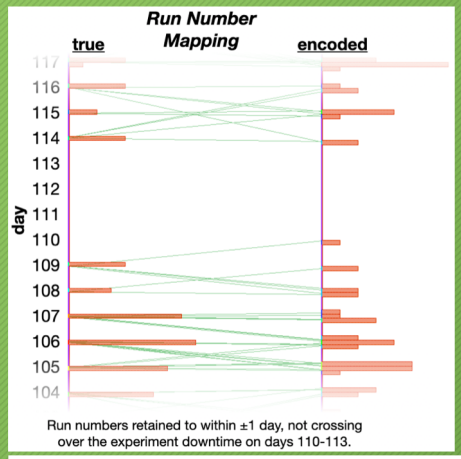
- Timestamps
- Luminosity & background rate scalers
- Details about overlapping triggers (if events matched multiple trigger criteria)
- Detector-specific trigger information
- Event number
- Run number (*encoded* : see below)
- File sequence within run (*encoded* : required for unique file names)

Non-minimum-bias triggered events were excluded from production.

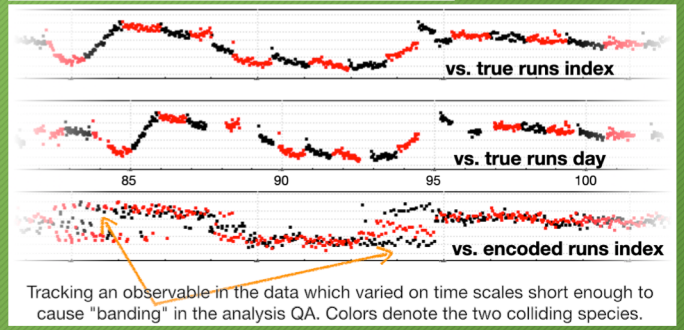
Run numbers required special consideration as some conditions varied over time scales that needed to be tracked at the analysis QA stage, using run number for chronology. Algorithm ensured 1-to-1 mapping between real and encoded numbers, while distributing the encoded numbers within the expected intervals of condition validity, estimated to be ±1 day (see upper plot). It was also considered important to distribute broadly enough to mix species, which were constant over the length of a collider fill typically lasting nearly a full day. However, a variation on a shorter time scale was eventually observed by the analysis QA (see lower plot), requiring a 3rd party to run analysis QA to understand the effect.

Unblind workforce:
These individuals were given restricted access and could not be among the blind analyzers

- Subsystem calibrations experts
- QA experts during data-taking
- Data production team
- 3rd party who could run analysis QA



Run numbers retained to within ±1 day, not crossing over the experiment downtime on days 110-113.



Tracking an observable in the data which varied on time scales short enough to cause "banding" in the analysis QA. Colors denote the two colliding species.

Workflow considerations:

- The recipe for creating random run numbers i.e. the "shadowing" code was public, but implementation was stored encrypted and only momentarily decrypted for compilation.
- Raw Data on mass storage (HPSS): by default, anyone can restore any data in STAR hence may have had access to the data prior to production. The Isobar raw data was restricted and only accessible by the "allowed" users (Unblind workforce). This was achieved by HPSS policy and a modification to the DataCarousel‡ used to restore files.
- During data production, files had to be restored on disk (local or central). To ensure absolute secrecy, a secondary (restricted) group was used on Unix file systems.
- All production logs were further protected using Unix protection / secondary group
- Three productions were processed:
  - ***Blind mixed:*** Pairs of input files were used from the same time range, one from each species, and merged into new "mixed" raw data files. 10% of events were randomly rejected to ensure minimal interpretability of run numbers from the number of events. Production occurred on the mixed sample. This served as a baseline for algorithm development. Analysis codes needed to be "frozen" at this stage.
  - ***Blind unmixed:*** The two species were produced separately and the resulting files placed in separate directories. A random portion of events were skipped to ensure obfuscation. Analyzers would re-run their frozen code unaware of which sample was which.
  - ***Unblinded:*** For final consistency (naming, all events present), a third pass was done with no filtering, no shadowing.

‡ D. Yu, J. Lauret - "Efficient Access to Massive Amounts of Tape-Resident Data", CHEP 2016 proceedings, J. Phys.: Conf. Ser. 898 082024, doi https://doi.org/10.1088/1742-6596/898/8/082024

Summary:
We have reviewed here the experience of the first-ever blind data production in high energy collider physics. What needed blinding, how to blind it, and from whom, required much consideration. Ensuring obfuscation imposed practical selections on runs to process and how events inside those runs were selected and arranged when processed, and when presented to the analyzers in the DSTs.

Despite operational efforts to avoid tell-tales in the data, some effects did arise that required additional, unplanned QA. However, clear and distinct access restriction procedures made it straightforward to deal with this. With the appropriate blinding achieved, analyzers and reviewers have been empowered to focus on the physics of interest.