Determining Characteristics for Data in a Data Center

University of Reading

- Data characteristics:
  - Proportion of a given (scientific) file format
  - Compression characteristics (ratio, speeds)
  - Performance behavior when accessing file data (e.g. using alternative I/O)
- Understanding these characteristics is useful
  - Proportions of a file format to identify relevant formats
    - Starting point for optimization of format
  - Conducting what-if analysis on the scale of the data center
    - Estimate the influence storage compression has
    - Performance expectations when applying a new I/O strategy
- Existing studies use a manual selection of “data” for representing stored data
- Conducting analysis on representative data is non-trivial
  - What data makes up a representative data set?
  - How can we infer knowledge for all data based on the subset?
    - Based on file number/count (i.e., a typical file is like X)
    - Based on file size (i.e., 10% of storage capacity is like Y)