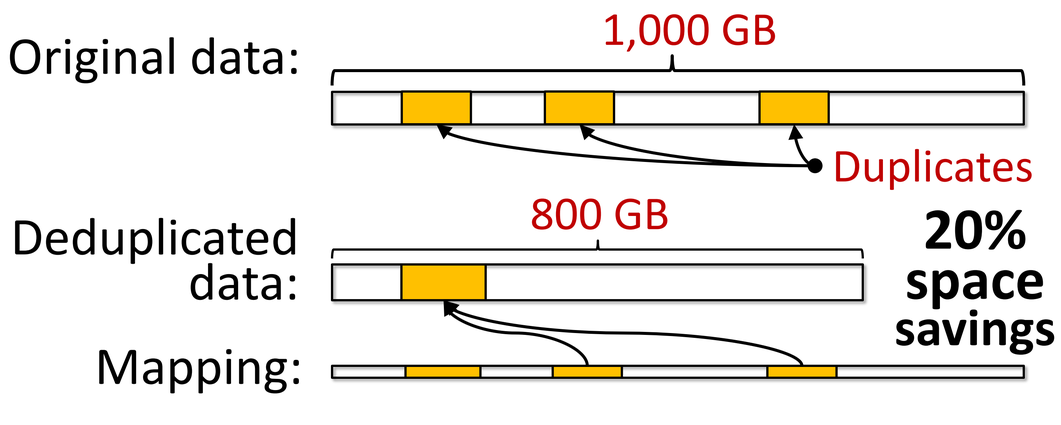
Generating Realistic Datasets for Deduplication Analysis

Vasily Tarasov[1], Amar Mudrankit[1], Will Buik[2], Philip Shilane[3], Geoff Kuenning[2]
[1]Stony Brook University, [2]Harvey Mudd College, [3]EMC Corporation

❶ Deduplication



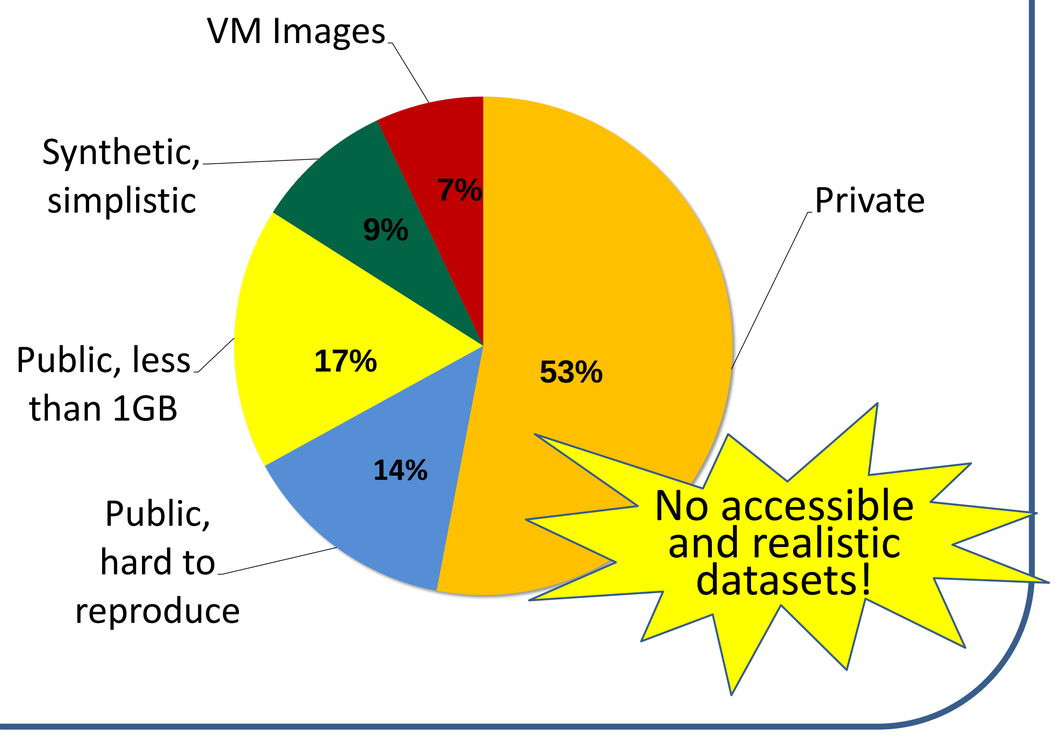
Typical Steps:

1. Chunking
   - Whole-file, Fixed, Variable size, …
2. Hashing
   - SHA256, MD5, Tiger, …
3. Indexing
   - Conventional databases
   - Specialized indexes
   - ….

Performance depends on dataset!

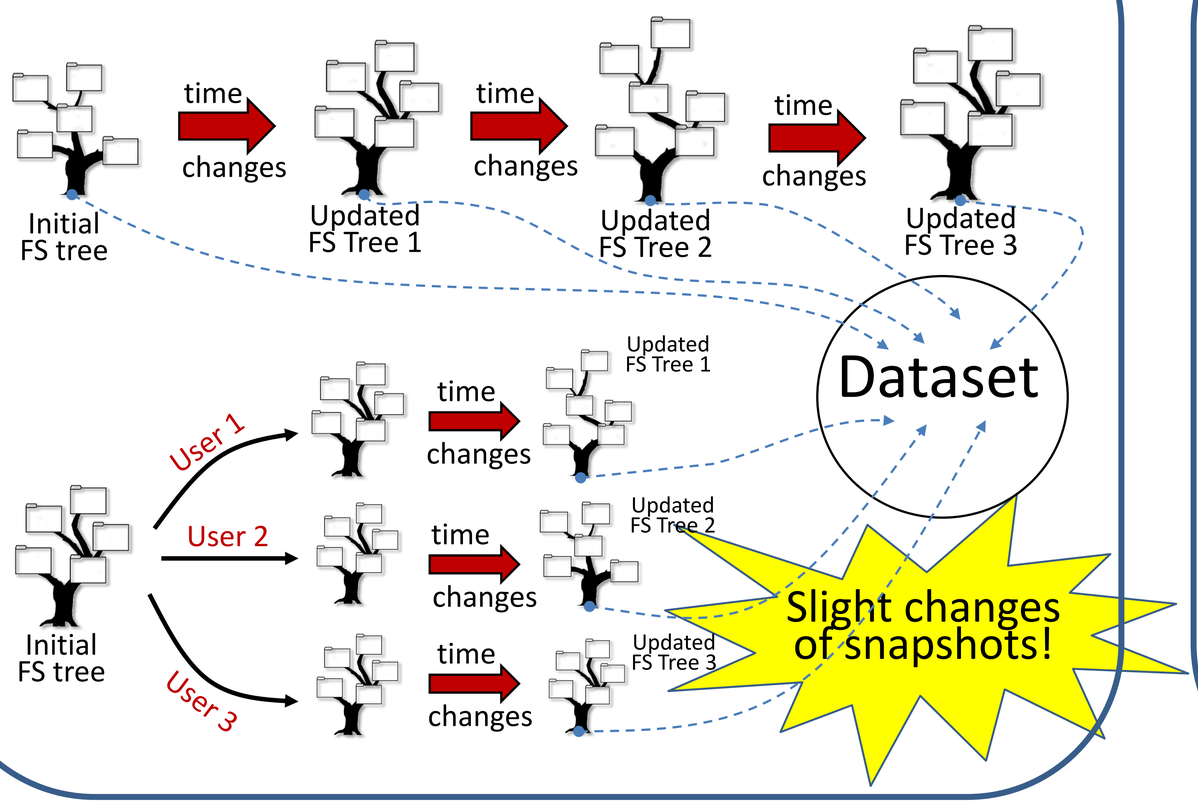
❷ Previous Datasets

120 datasets, 29 papers from, e.g., FAST, ATC conferences



No accessible and realistic datasets!

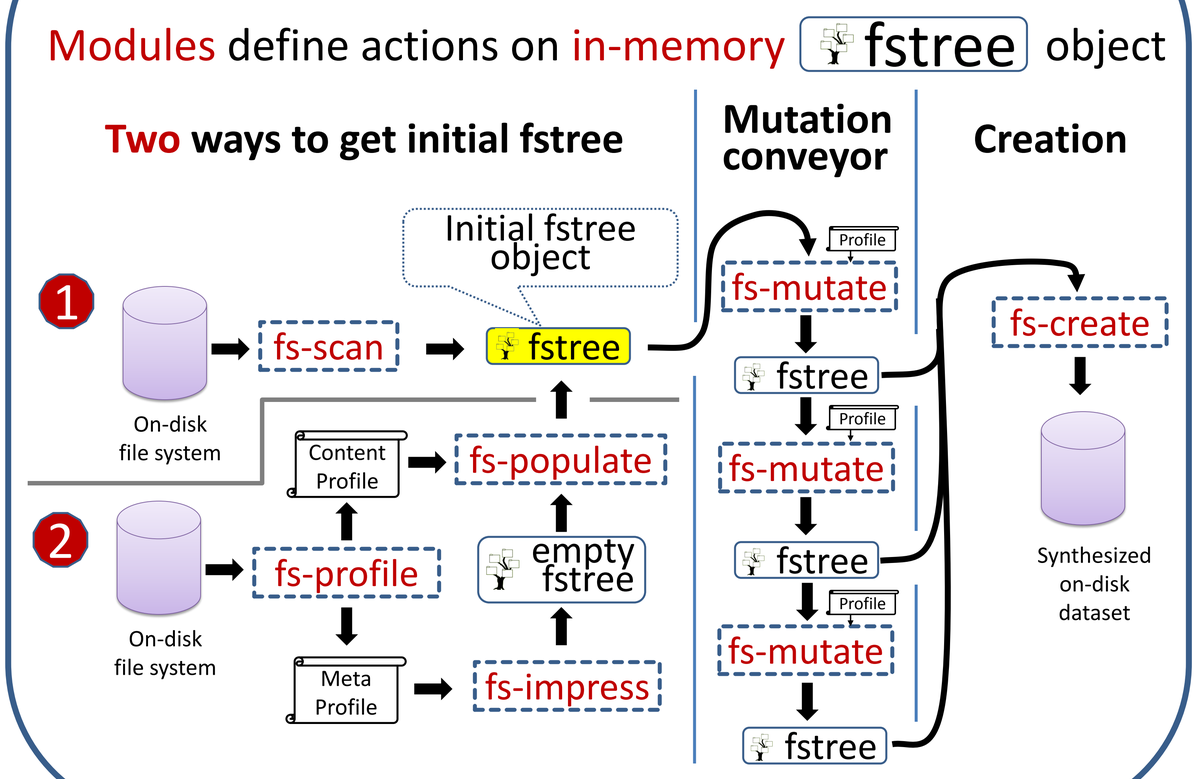
❸ Dataset Formation



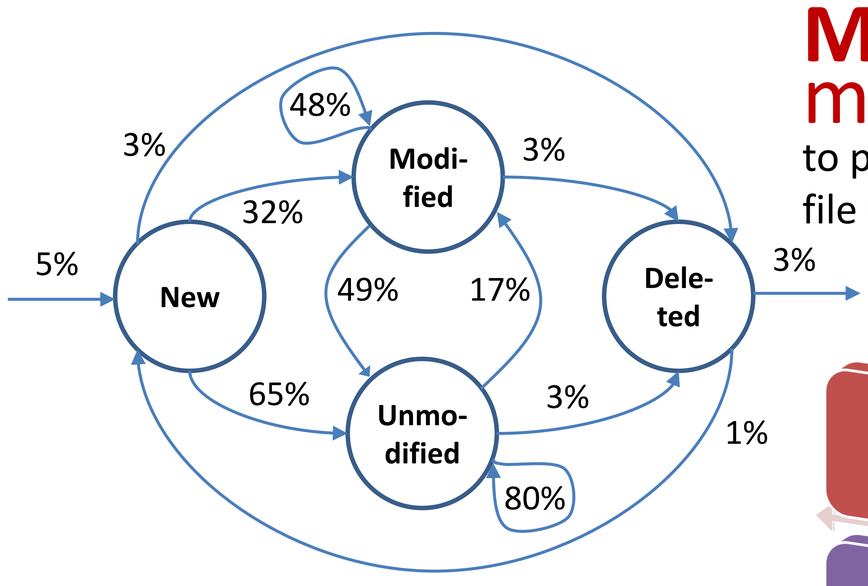
Slight changes of snapshots!

❹ Action Modules

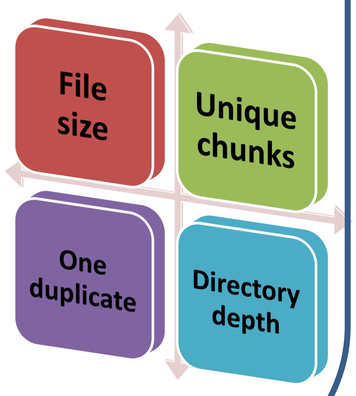
Modules define actions on in-memory fstree object



❺ M&D Model

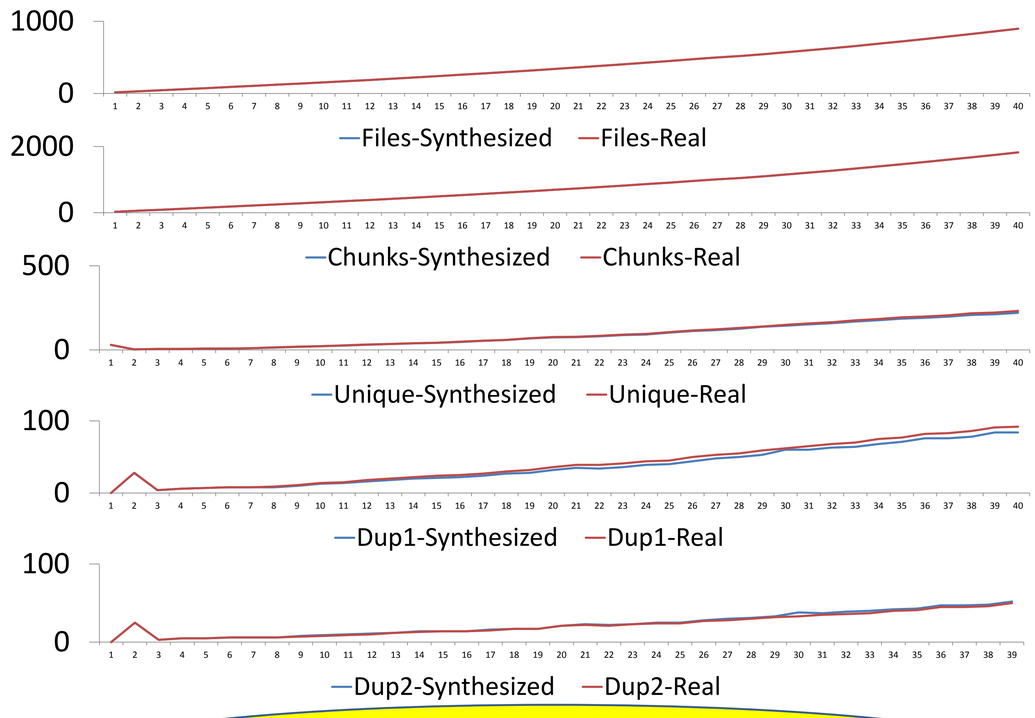


Markov model
to preserve active file set history



Multi-dimensional distribution
to select file and perform appropriate action on it

❻ Preliminary Results



2.6.0 – 2.6.39 Kernels dataset: within 10% accuracy!

10th USENIX Conference on File and Storage Technologies (FAST), February 14-17, 2012. San Jose, CA