Data is growing rapidly:
IDC study estimates 1 ZB in
2010 growing to 35 ZB in 2020

Backup workloads have unique properties
compared to Primary workloads

- Much larger file size
- Many more files per directory
- Shorter lived files
- Higher churn (20% of data freed and rewritten every week)

Metadata traces collected from over
700 TB of backup data representing
various dataset types and sizes

Best deduplication with 4KB chunks, but 8KB is often sweet
spot for data-structure sizes and cleaning

Backup != Primary