Tools of Economic Complexity Analysis

Review Session 0A: Cluster Usage and Data Management

Motivation

- Most GL project datasets on the cluster
- Data security
- Larger datasets / computational requirements
- Inadequate documentation on cluster usage
- Evolving data management practices

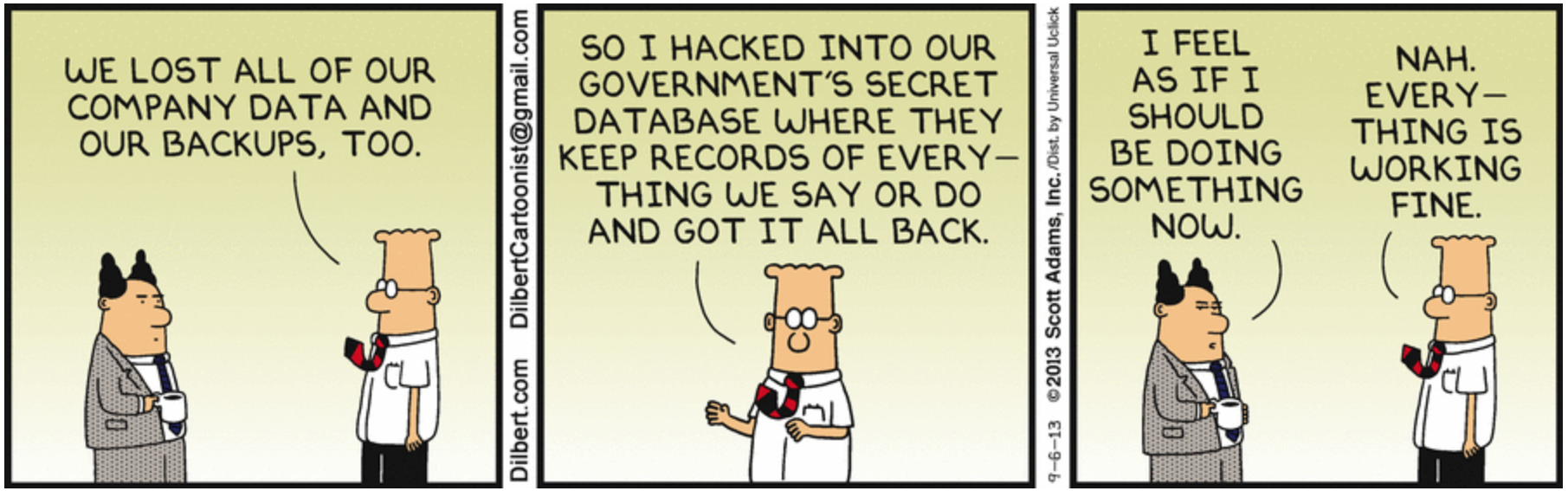
Website

- https://cid-harvard.github.io/workshop-cluster-training

Responsibilities

- For specific dataset(s) *you* work with: who is responsibile for them?
- Keep data clean, backed up, secure, documented

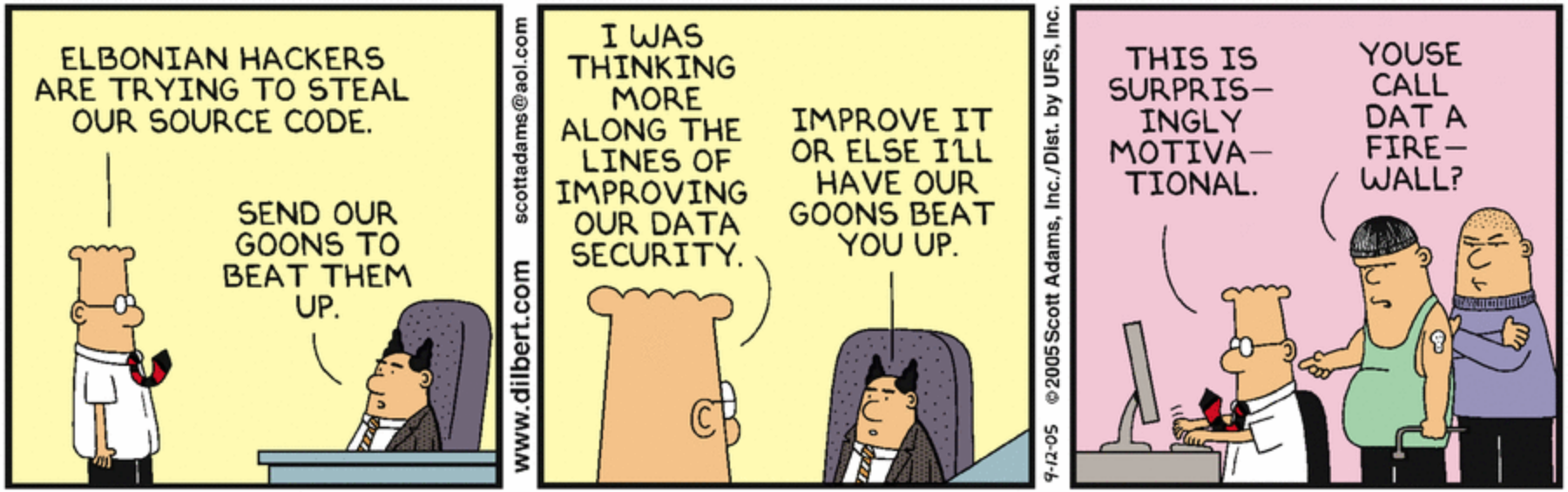
Data Backup Strategy



Access and security

- Cluster security discussions
- Data security levels: http://security.harvard.com/dct

Data Security Strategy

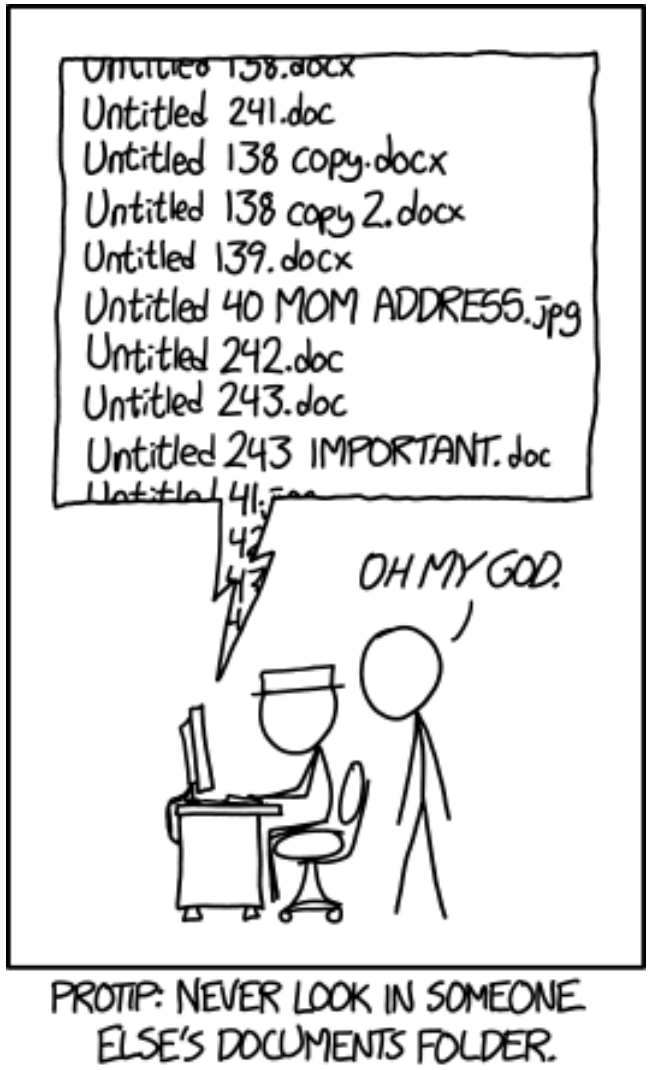


Ethics and Privacy

- *Your* responsibility
- Does your project need an IRB review?
- Do you need / have an IRB ethics certification?

Folder Structures

- Raw: Immutable
- Processed: Clean, centrally available
- Intermediate: Irrelevant
- Personal folders: Derivatives of "Processed"



Cluster Training

Website: https://cid-harvard.github.io/workshop-cluster-training