# Towards a Powerful, User-Friendly System for Controlling Access to Protected Data

Lucas Lochovsky

August 5, 2015

# Motivation

- Increasing amount of protected data being worked with in the lab
- Users with access must be controlled
- Current solutions have been developed in a relatively *ad hoc* fashion, leading to half-measure solutions

# Previously Developed Solutions

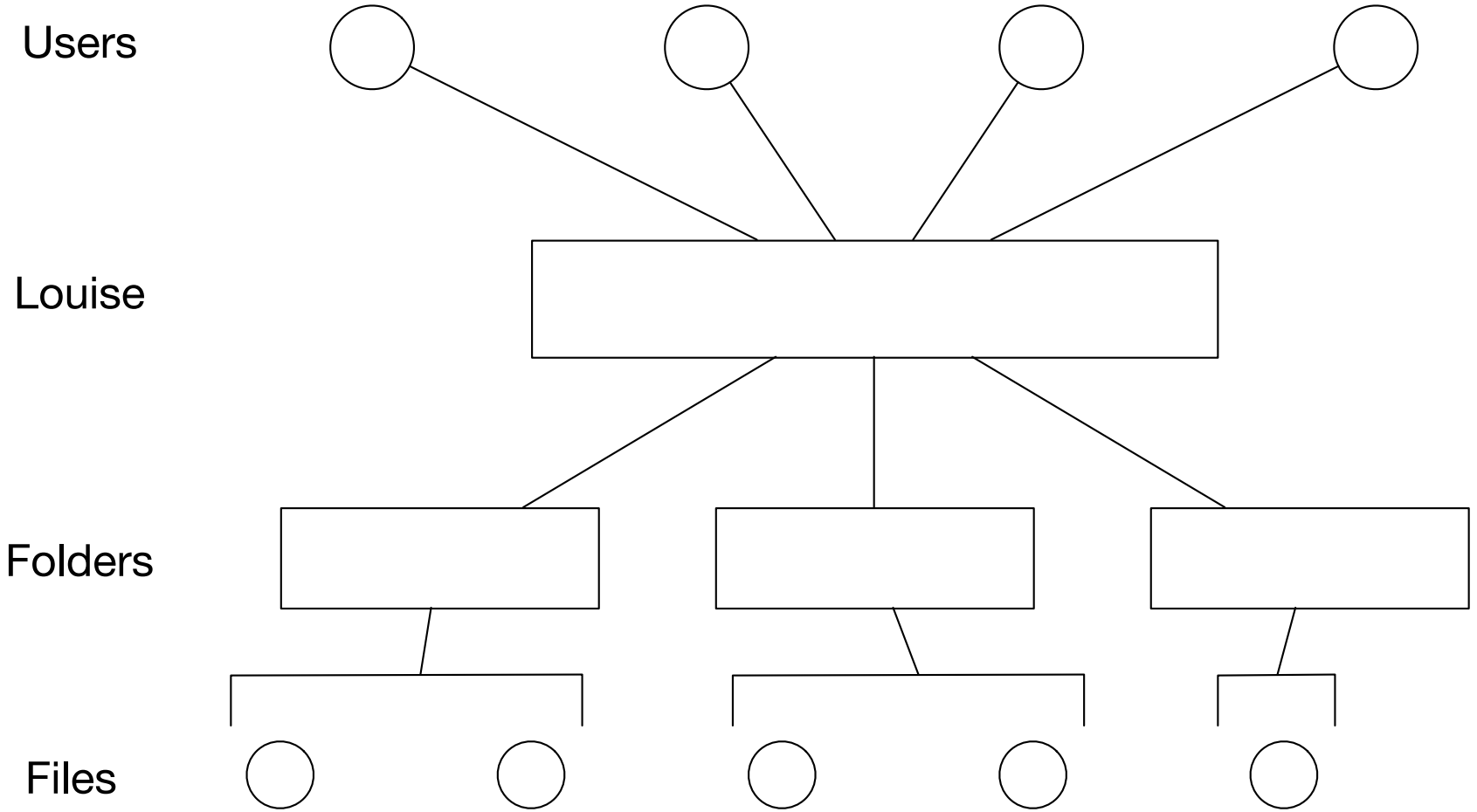
- Protected data resides on separate storage server
  - Isolated from compute nodes
  - Storage server not used
- Keep data on Louise
  - Need to rely on file permissions: define a user group that has protected data access
  - But different datasets have different users
    - Group management unwieldy: have to go through admins
  - In the end, everyone will just set file permissions to everyone in Gerstein lab out of convenience
    - Not a proper solution
- Track protected data in spreadsheet
  - Resides in a separate place from the data
    - Not used

# Proposal for a New, Well Thought Out Solution

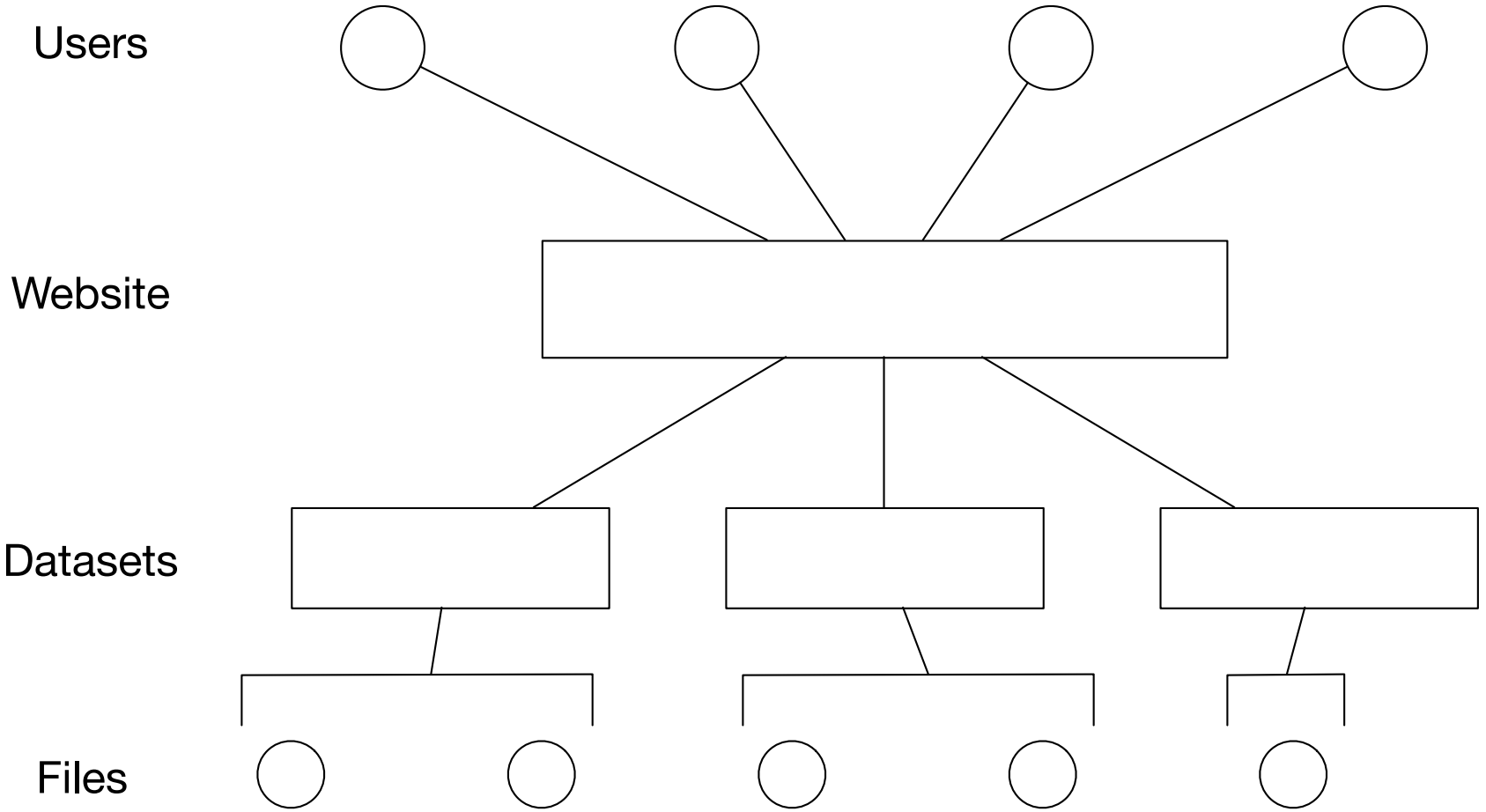
**Goals**

- Bring the management of data access closer to the data
  - Actions taken in the interface directly influence the data files
- Make the management interface easy enough that users will prefer its use to alternatives
  - No need to work out *ad hoc* access rules

# Physical Modelling of Data

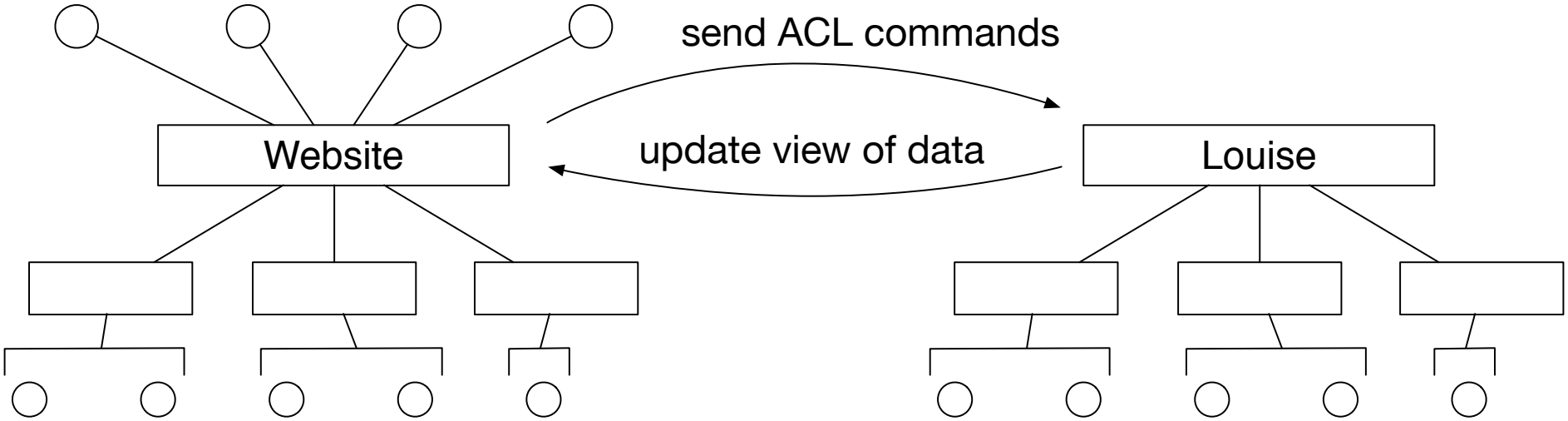


# Conceptual Modelling of Data



# Proposal

- Users interact with the data through the conceptual model, which directly influences the physical model



# Website Mockup

## Gerstein Lab Protected Data Management

MOCKUP

- Prostate cancer
  - USERS
    - ll426
    - jz435
  - FILES
    - BI_375923.bam
    - BI_735878.bam
- Breast cancer
  - USERS
    - mj123
    - kl234
  - FILES
    - KS_36784.bam
    - KS_35675.bam
- Glioma
  - USERS
    - ty647
    - pw758
  - FILES
    - HMS_3472.bam
    - HMS_2347.bam

# Data Schema

- User entity
  - Username
  - Password
  - Real name
  - Datasets curated
  - Datasets with access privileges

# Data Schema

- Dataset entity
  - Name
  - Files and Location (path)
  - Curator
  - Accessors
  - Source (link to publication, accession number)
  - Access restrictions (incl. expiration date)
  - Tags: Add datasets to categories that allow for useful groupings

# Access Control Lists (ACL)

- Owners of files can dynamically add and remove permissions of individual users on files
- Bypasses having to work with Unix groups (requires admin privileges)

```
[ec2-user@ip-172-30-0-223 ~]$ getfacl myfile.txt
# file: myfile.txt
# owner: ec2-user
# group: ec2-user
user::rw-
user:user1:rw-
group::rw-
mask::rw-
other::r--

[ec2-user@ip-172-30-0-223 ~]$ setfacl -m "u:user1:rwx" myfile.txt
[ec2-user@ip-172-30-0-223 ~]$ getfacl myfile.txt
# file: myfile.txt
# owner: ec2-user
# group: ec2-user
user::rw-
user:user1:rwx
group::rw-
mask::rwx
other::r--

[ec2-user@ip-172-30-0-223 ~]$
```

# Current List of Functions

- User signs in/signs up, and has an interface into the datasets
- User downloads a new dataset, initiates a new dataset record and becomes the dataset's curator
- Curator adds/removes accessor on access list, accessor receives message to confirm
  - Message includes reason for addition/removal
- Accessor requests access/removal, curator receives message to confirm
  - Message includes reason for addition/removal

# Currently Planned Features

- Messaging system
  - Users in the same group can discuss data without involving anyone else
  - Useful for private data
  - Members of group are always up-to-date: Always know you are reaching exactly the people who can help and no one else
- Put Change History on things (like MS Word's Track Changes)
- Notifications when restricted dataset access is about to expire
  - Timely preparation of renewal application

# Development Timetable

- Create design
- Survey users on validity of design, and gather new requirements
- Incorporate new requirements into design
- Deploy and maintain
- Product evangelism