Data Repositories

INFX 551
Winter 2018

Course Outline

Week 3
(final week in this module)

Data	Data Systems	Policy, Privacy, & Ethics
Types & Roles	Repositories	Policy
Documentation	Preservation	Privacy
Standardization	Cost Models	Ethics

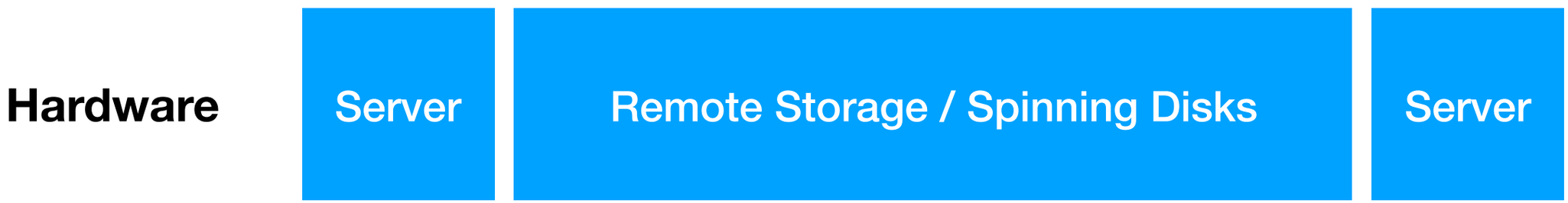
Topics

- Repository Storage and Management
    - Repository Layers
    - Repository Functions
- Repositories by type

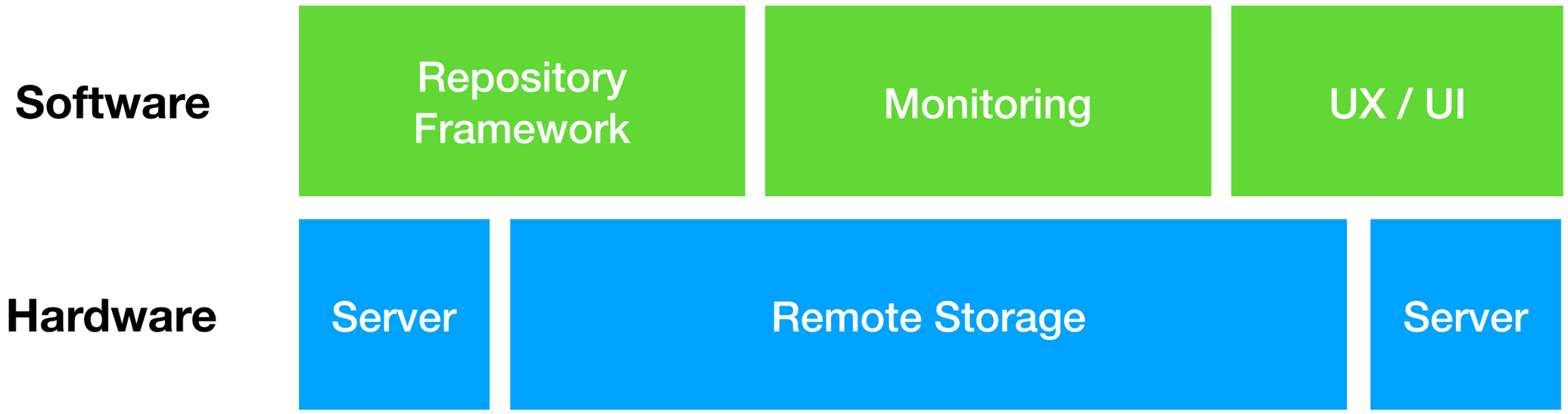
A data repository is both a digital archive for **storing and preserving data**, as well as an institution that establishes policies, and governance for **data management**.

- Storage
    - Software and Hardware for ingesting, curating, and preserving data.
- Management
    - Policies: Establish mission, goals, levels of preservation, types of data accepted, cost of deposit, etc.
    - Governance: How the repository will be managed, by whom, and through what budgets, and with what level of security.

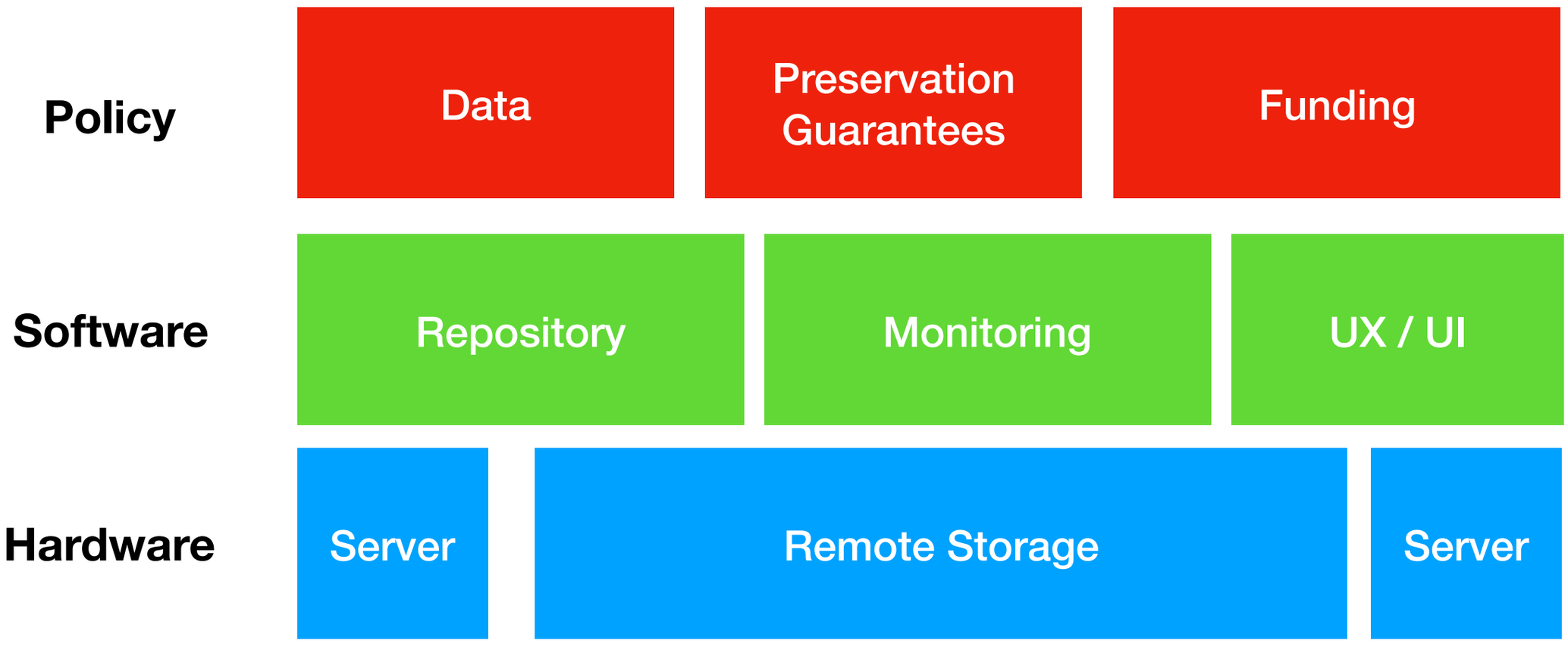
Repository Layers



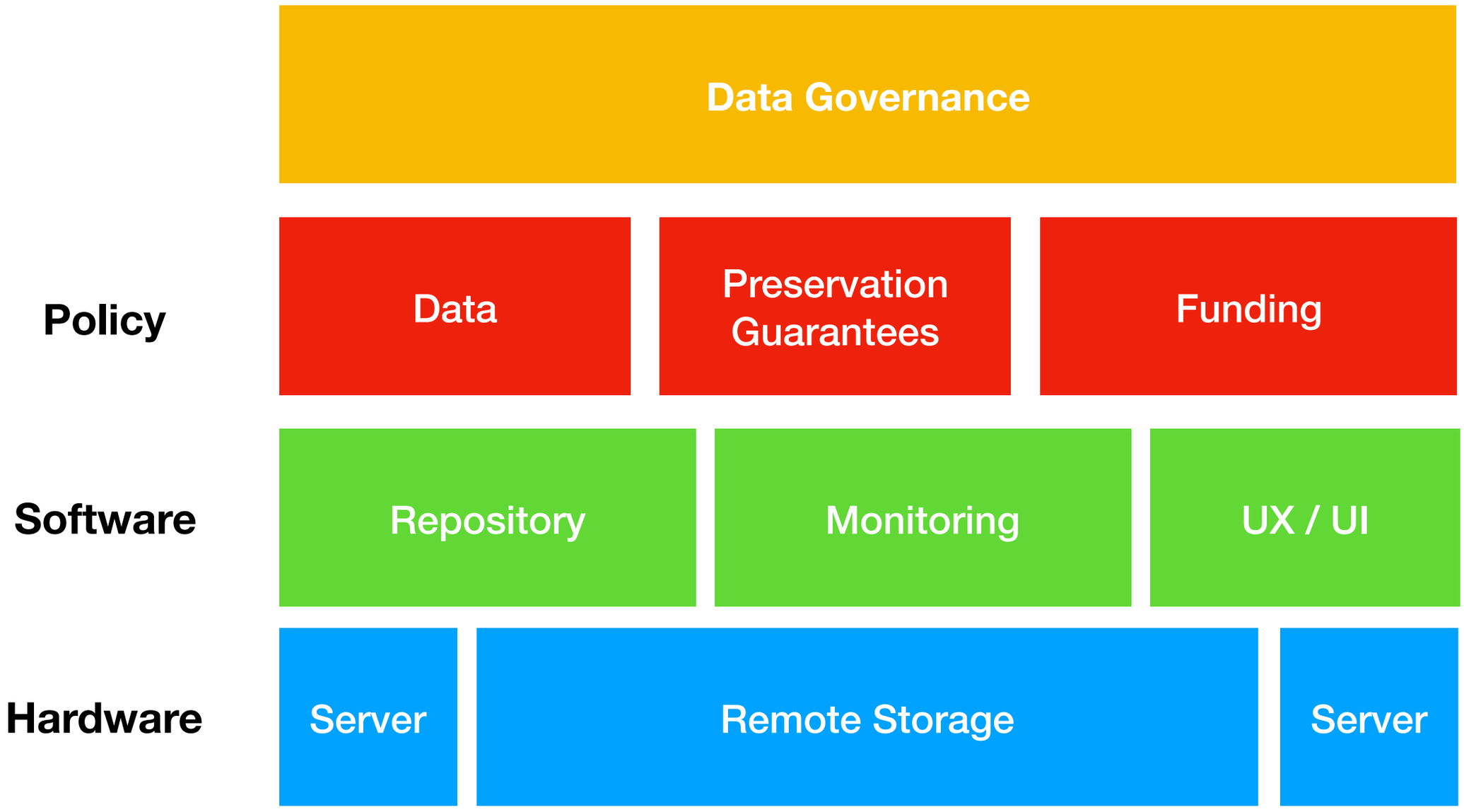
Repository Layers



Repository Layers



Repository Layers



Fedora™
A DIGITAL REPOSITORY FRAMEWORK

DSPACE

Research Data

The Dataverse Project

Open Data

Socrata

ckan

Software

Repository Framework	Monitoring	UX / UI

Hardware

Server	Remote Storage	Server

Repository Frameworks offer three basic services

- **Submit** : Users have a web-interface to upload files, create metadata, sign deposit agreements.

- **Archive**: Ingest (pull data into software environment to structure data), characterize data (e.g. using JHOVE), create a unique identifier for data, create additional metadata / documentation, writes to database(s) for storage.

- **Disseminate**: Index metadata, create catalog, publish data pages, enable data to be downloaded, require users to sign data usage agreements.

Type	Role	Example
Discipline / Domain	**Specific:** Optimize curation services and storage of specific types of domain data	ICPSR (social science) GenBank (DNA) British Geology Service
Institutional	**Generic:** Offer basic services to diverse institutional members	DSpace@MIT ResearchWorks @ UW NIH Data Commons
Public / Open	**Generic**: Offer broad in topic, little curation before data storage	Dataverse at Harvard Zenodo (CERN) FigShare
Private	**Specific**: Non-public, optimized around specific data at private organization	Microsoft Data Lake through AZURE