AN INFORMATICS FRAMEWORK FOR TESTING DATA INTEGRITY AND CORRECTNESS OF FEDERATED BIOMEDICAL DATABASES

Mijung Kim,[1] Tahsin Kurc,[2] Alessandro Orso,[1]
Jake Cobb,[1] David Gutman,[2] Mary Jean Harrold,[1]
Andrew Post,[2] Ashish Sharma,[2] Tony Pan, [2]
Dhananjaya Sommanna, [2] Joel Saltz[2]

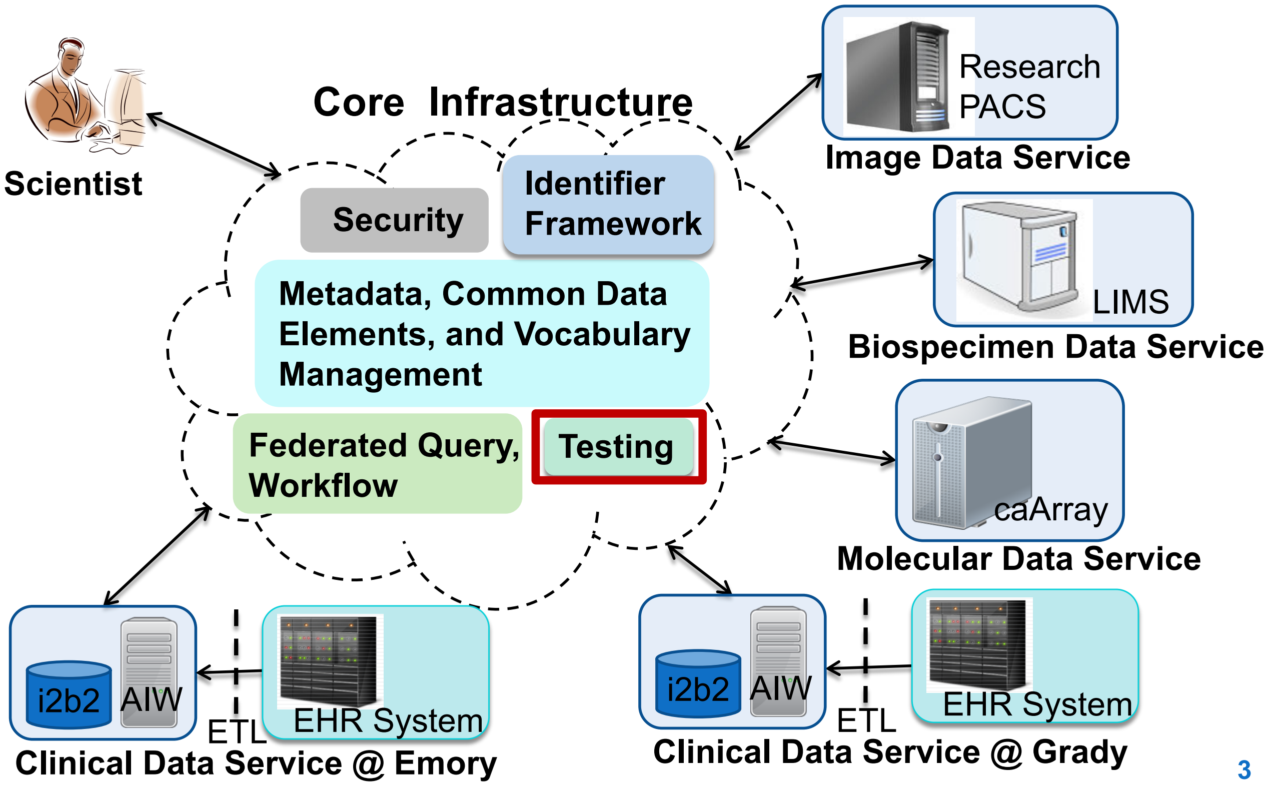
[1] College of Computing, Georgia Institute of Technology

[2] Center for Comprehensive Informatics, Emory University

Problem Definition

- Support systematic testing of data integrity and correct operation in a federated database environment
- Federated Database Environment
  - Heterogeneous data sources
  - Autonomously created and managed
- Efforts for Resource Federation
  - caBIG (cancer Biomedical Informatics Grid)
  - CVRG (CardioVascular Research Grid)
  - NHIN (Nationwide Health Information Network)
  - CTSAs (Clinical and Translational Science Awards)
  - Shrine (i2b2 Shared Health Research Information Network)

Federated Environment



Core Infrastructure

Scientist

Security

Identifier Framework

Metadata, Common Data Elements, and Vocabulary Management

Federated Query, Workflow

Testing

Research PACS

Image Data Service

LIMS

Biospecimen Data Service

caArray

Molecular Data Service

i2b2 AIW

ETL

EHR System

Clinical Data Service @ Emory

i2b2 AIW

ETL

EHR System

Clinical Data Service @ Grady

Use Case: In Silico Brain Tumor Research Center

- A research center for in silico study of brain tumors
  - Collaboration among four institutions
  - **Goal: Better disease classification and study of disease progression**
  - Initial focus on Gliomas
- Systematically execute in silico analyses (experiments) using complementary data types
  - Integration and correlation of clinical data and analysis results from omics, radiology imaging, and microscopy imaging data
  - Data from TCGA and Rembrandt projects as well as partner institutions

Examples of Issues Encountered

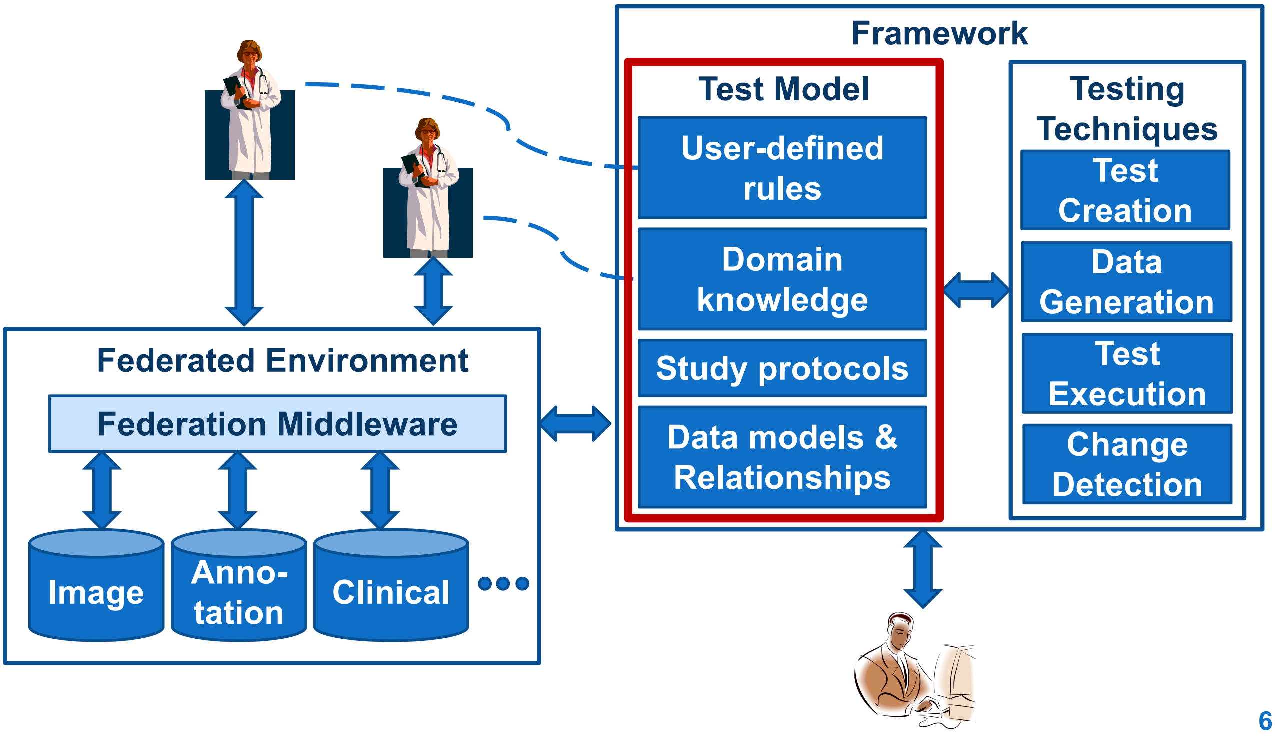
- Violation of existence constraints
  - Not all images for slides used in manual annotations were available
  - Some patients had image data but no mRNA data
  - Data in molecular datasets with patient identifiers was

Cause data inconsistencies!!

expected/known progression of disease for some patients

- Incorrect temporal dependencies
  - Some patients were in one study, then were recruited to the other study

Testing Framework Overview



Framework

Test Model

User-defined rules

Domain knowledge

Study protocols

Data models & Relationships

Testing Techniques

Test Creation

Data Generation

Test Execution

Change Detection

Federated Environment

Federation Middleware

Image

Anno-tation

Clinical

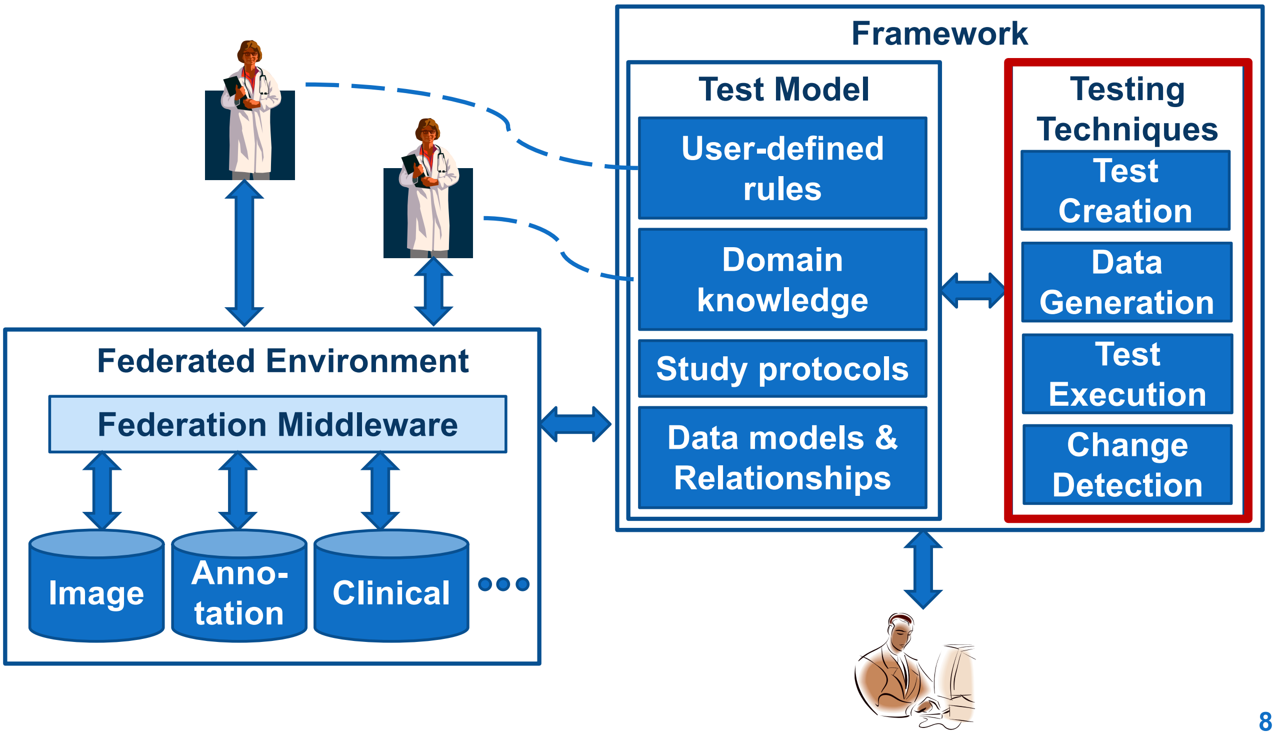
•••

Test Model

- User-defined rules
  - **"days to death" value in Clinical database should not change.**
  - *(Clinical/Patient/days_to_death) → immutable*
- Domain Knowledge
  - **Stage X should not follow Stage Y for disease A.**
  - ∀*t2 > t1* ⇒ *diseaseA.stage(Clinical/Exam/status)[t1] < diseaseA.stage(Clinical/Exam/status)[t2]*
- Study protocols

  In-silico brain tumor study must contain (1) *MR Data, (2) Microscopy Data, (3*) Patient survival data, and (4)mRNA data
- Data models & Relationships
  - **Attribute Gender in Image database has the same value as Attribute Sex in Clinical database.**
  - *(Image/Patient/Gender, Clinical/Patient/Sex) → sameValue*

Testing Framework Overview



Testing Techniques

- Test Creation
  - ○ Analyze the test model
  - ○ Identify relevant data elements
  - ○ Generate testing requirements and test cases
- Data Generation
  - ○ Generate synthetic datasets to test critical but rarely-violated rules and private data
- Test Execution
  - ○ Run tests periodically and on demand
  - ○ Report test outcome
- Change Detection
  - ○ Detect changes
  - ○ Identify effects of changes
  - ○ Execute relevant test cases

Current State

Type of Dataset	Data Management System
Neuroimaging Data	
Radiology images	Virtual PACS, xNAT
Manual annotations	AIME
Molecular Data	
mRNA, miRNA, methylation data, gene-expression data	in-house developed database with file system for data files
Clinical Data	
Clinical data, specimen data	i2b2, in-house developed database
Pathology Data	
Whole slide microscopy images, image metadata	caMicroscope
Microscopy image analysis results	PAIS

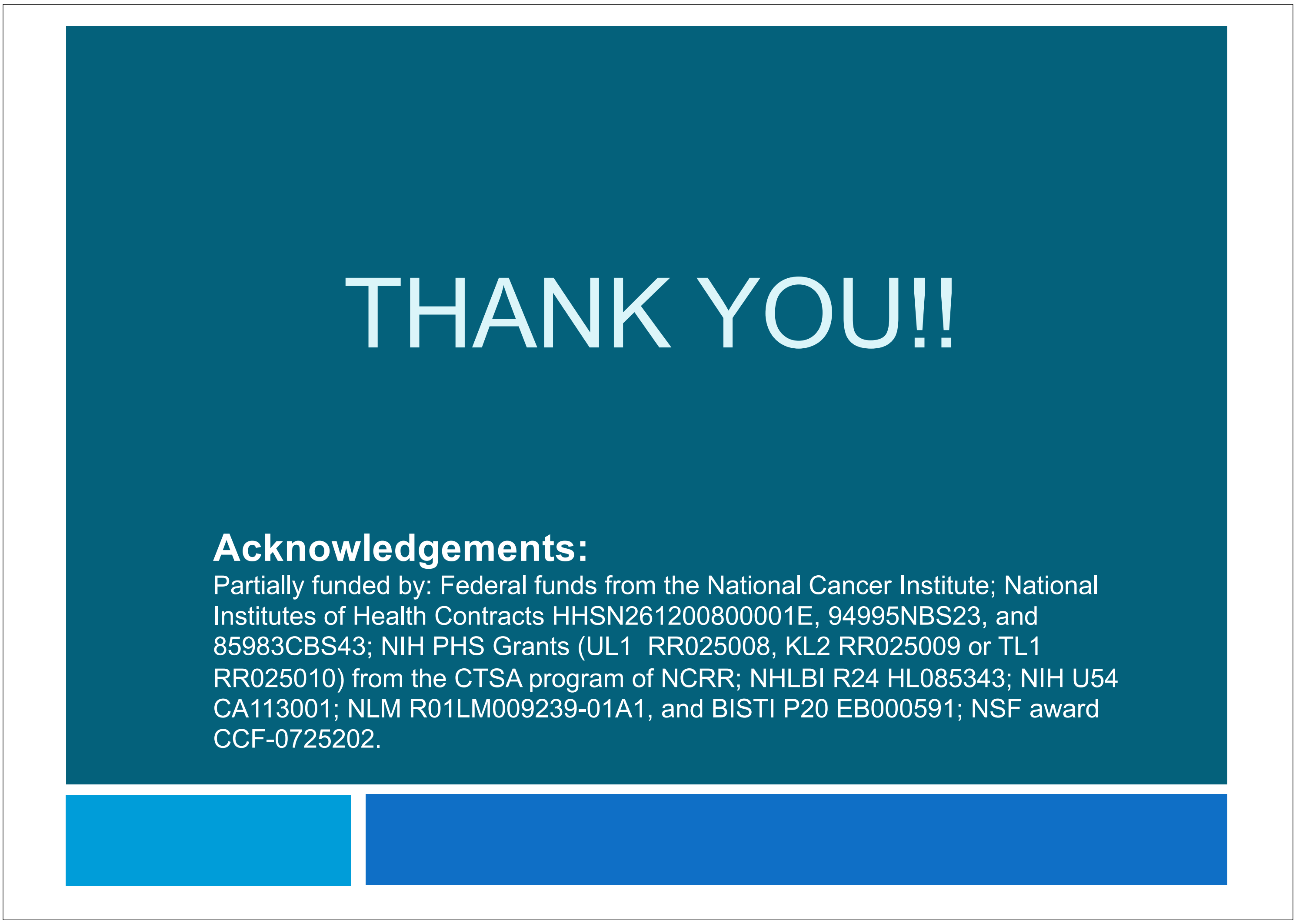
Example Rule (in OWL/SWRL)

- If a patient has molecular data, the patient must have clinical data
- *(Molecular/Genomic/patient_id, Clinical/Patient/patient_id) → existIn*

```
<owl:Class rdf:ID=“Molecular.Genomic.patient_id">
    <rdfs:subClassOf rdf:resource=“ontology.owl#Column"/>
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty>
          <owl:ObjectProperty rdf:ID=“existIn"/>
        </owl:onProperty>
        <owl:someValuesFrom>
          <owl:Class rdf:about="#Clinical.Patient.patient_id"/>
        </owl:someValuesFrom>
      </owl:Restriction>
    </rdfs:subClassOf>
</owl:class>
```

Conclusion

- Challenges in federated environments
  - Errors are inevitable
  - Developing customized and one-off solutions is expensive and inefficient
- Our work contributes a middleware framework
  - Test Model: High-level, rule-based representation of expected state
  - Testing Techniques
    - Generate test cases using the test model
    - Execute the test cases
    - Detect changes



THANK YOU!!

Acknowledgements:

Partially funded by: Federal funds from the National Cancer Institute; National Institutes of Health Contracts HHSN261200800001E, 94995NBS23, and 85983CBS43; NIH PHS Grants (UL1 RR025008, KL2 RR025009 or TL1 RR025010) from the CTSA program of NCRR; NHLBI R24 HL085343; NIH U54 CA113001; NLM R01LM009239-01A1, and BISTI P20 EB000591; NSF award CCF-0725202.