UMBC

HETEROGENEOUS DATA INTEGRATION FOR CLINICAL DECISION SUPPORT SYSTEM

Aniket Bochare - aniketb1@umbc.edu

CMSC 601 - Presentation

Date-04/25/2011

Agenda

- Introduction and Background
- Framework
- Heterogeneous Data
- Data Integration Model
- Problems Involved and Solution
- Personalized Web for Doctors and Patients
- Future Work
- Questions

Clinical Decision Support System ?

- Clinical decision support system is a computer software, which is designed to assist physicians and other health professionals with decision making tasks, as determining diagnosis of patient data.
- Clinical Decision Support systems link health observations with health knowledge to influence health choices by clinicians for improved health care pre and post diagnosis to predict diseases.
- Types

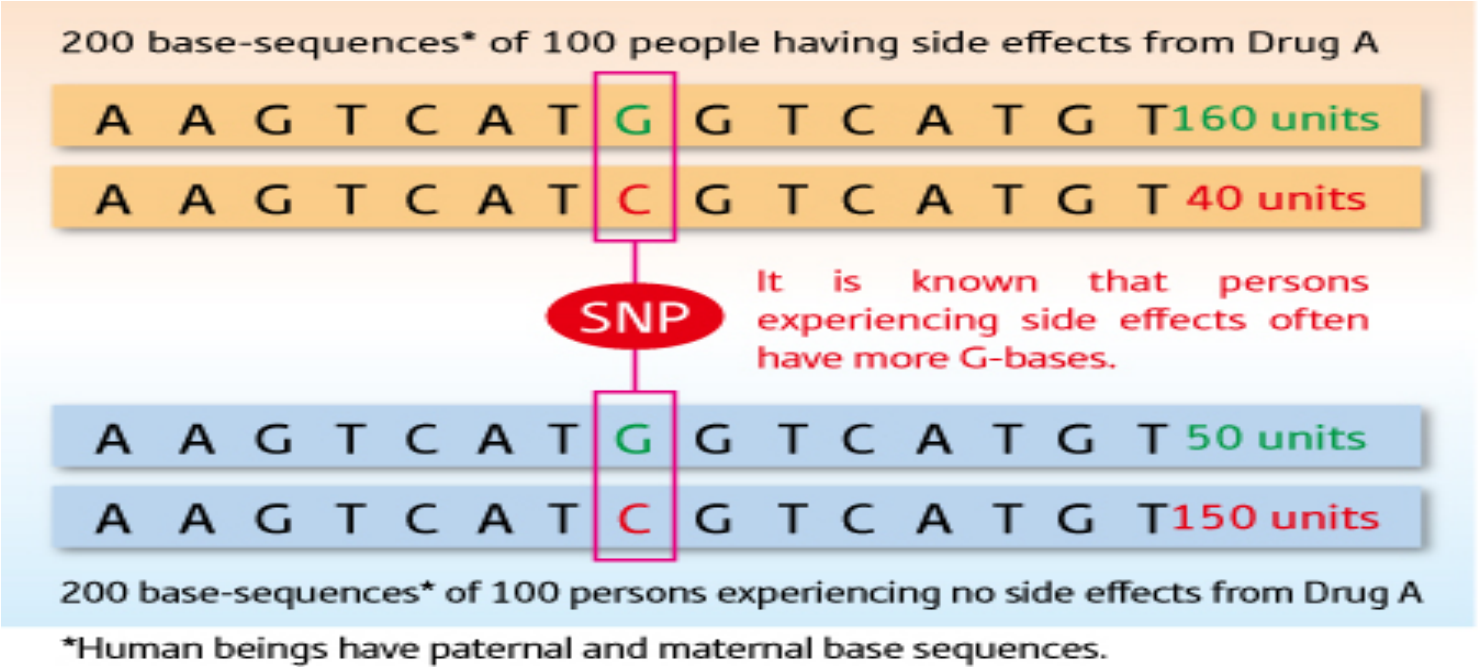
  -Knowledge Based(if-then, inference based)

  -Non-Knowledge Based(Machine Learning)

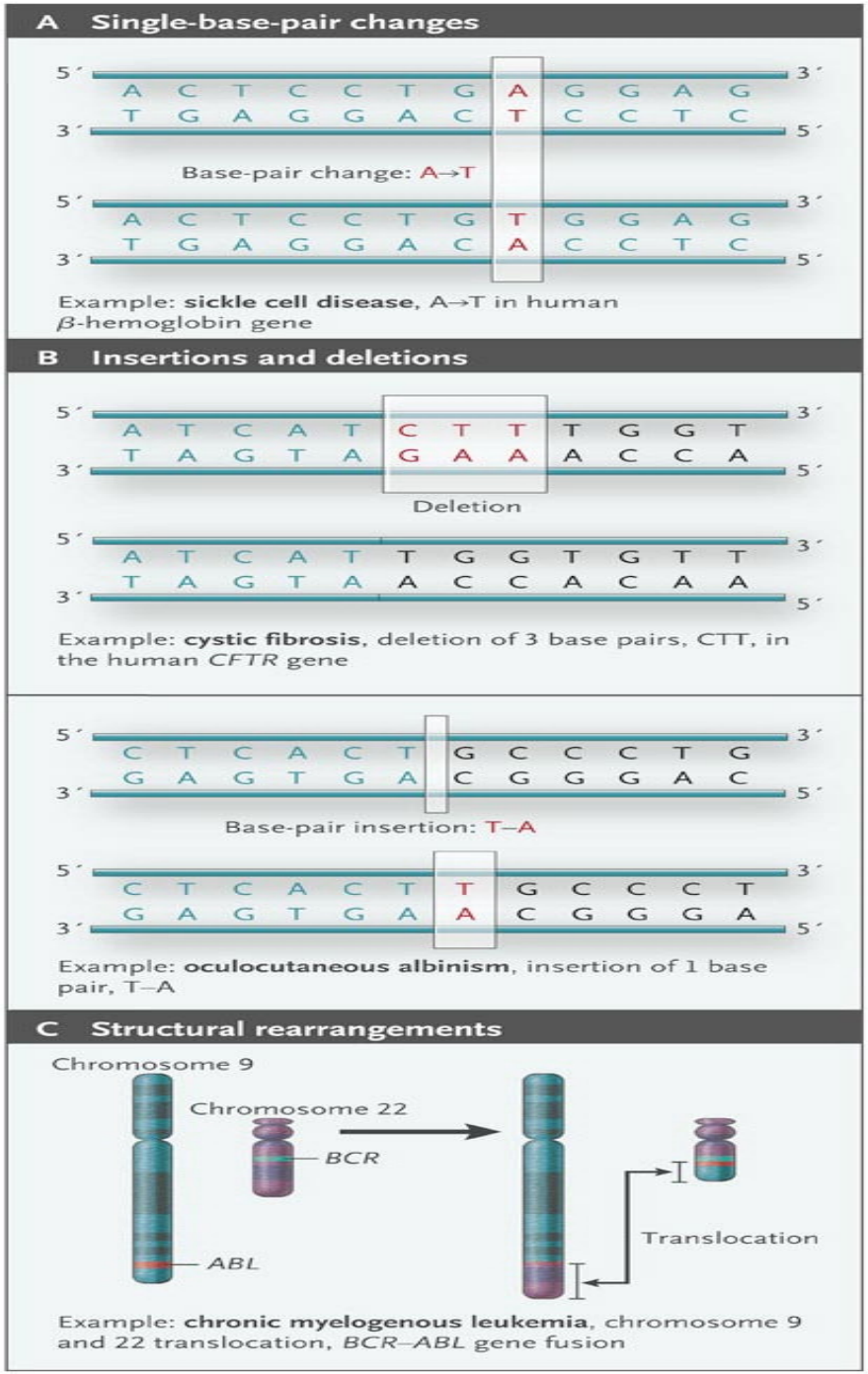
- Challenges include patient's symptoms, medical history, family history and genetics to be understood in detailed with the help of doctors.
- Choosing from methodologies like Bayesian Network, Neural network, genetic algorithms and rule based mechanisms is a challenge.

SINGLE-NUCLEOTIDE POLYMORPHISM (SNP) ?

- Variation occurring when a single nucleotide — A, T, C, or G — in the genome differs between members of a species or paired chromosomes in an individual.



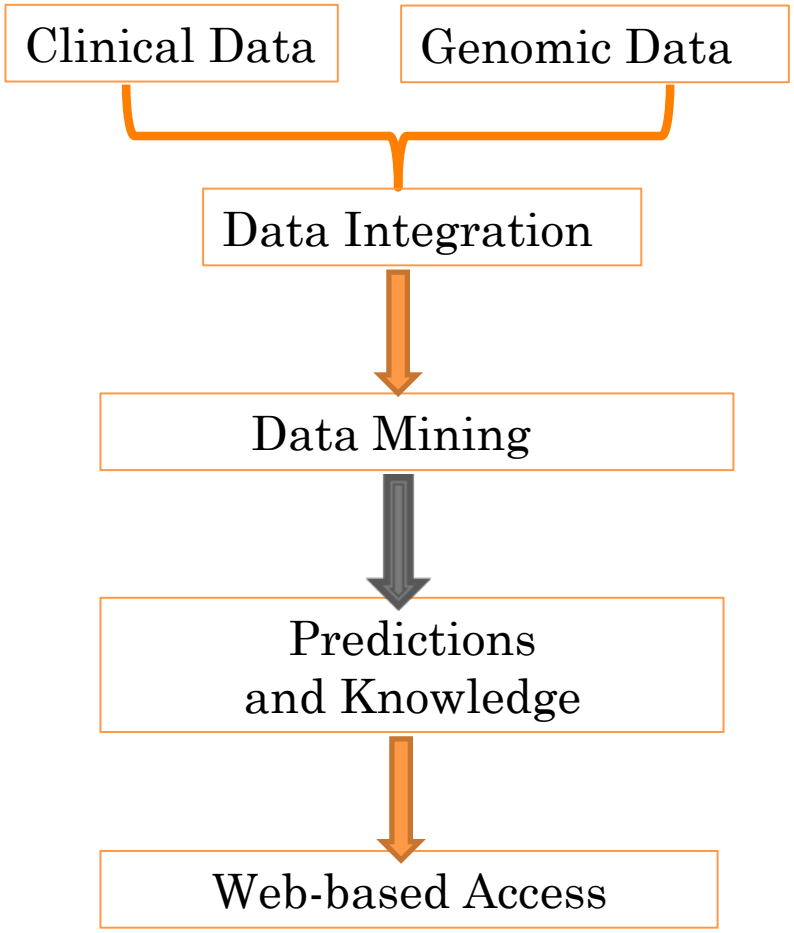
Human Genetic Variation



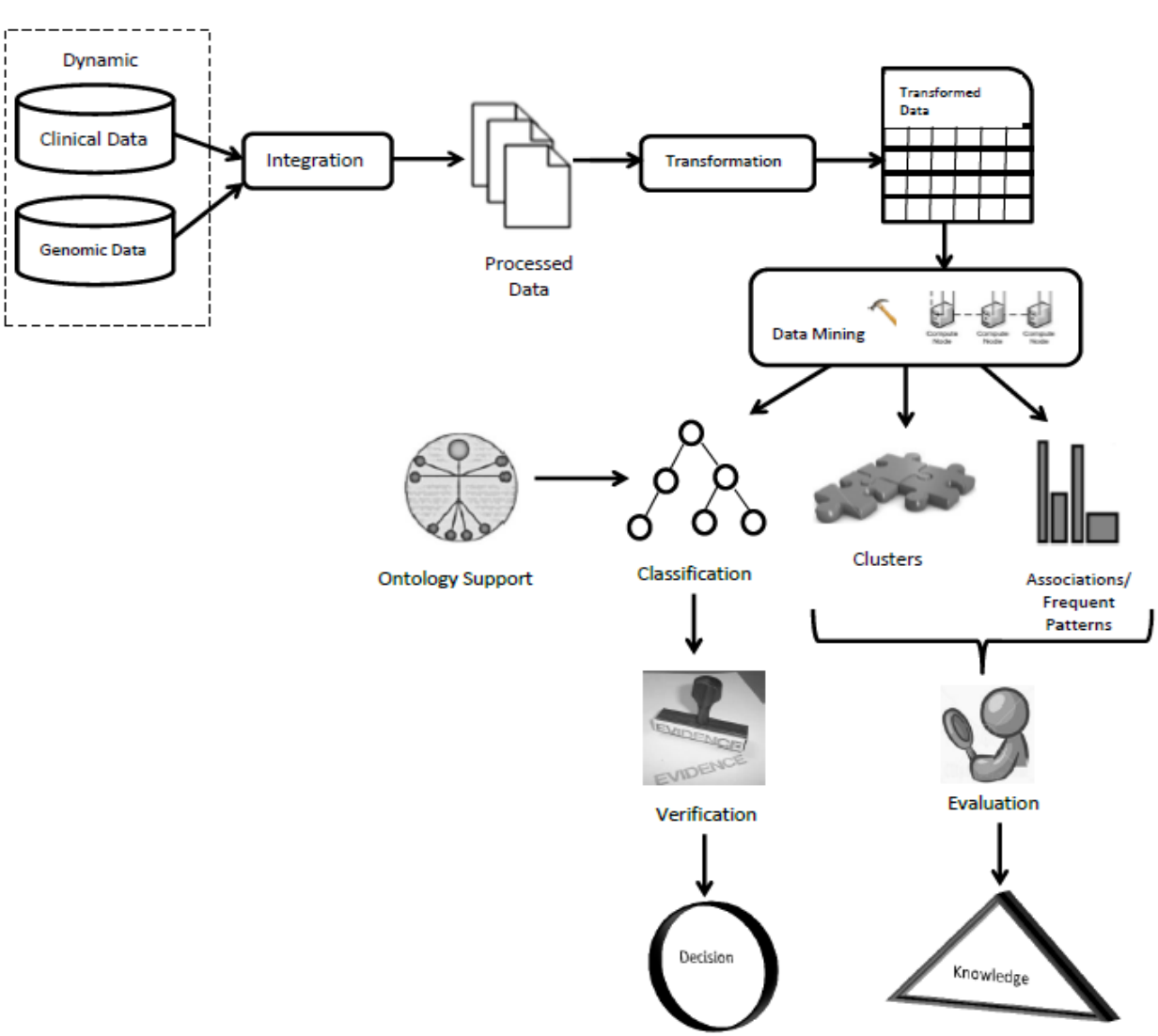
PROPOSED SOLUTION

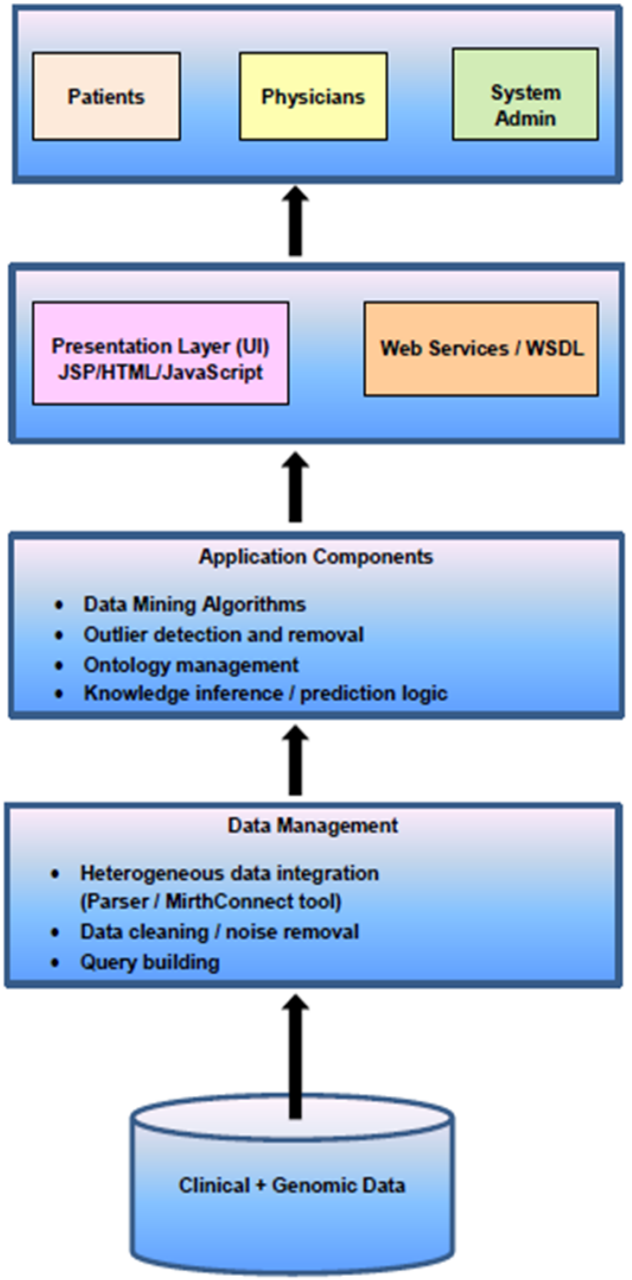
- Develop a Web-based Clinical Decision Support System that will integrate genomic, metabolic associations and data mining correlative evidence gathered by computational algorithms for prediction and knowledge discovery and will be invoked on demand at the point of care.

- New parallel computational technologies in conjunction with complex database queries will be employed for personalized diagnosis.



FRAMEWORK


Dynamic
Clinical Data
Genomic Data
Integration
Processed Data
Transformation
Transformed Data
Data Mining
Ontology Support
Classification
Clusters
Associations/ Frequent Patterns
Verification
Evaluation
Decision
Knowledge


Patients
Physicians
System Admin
Presentation Layer (UI) JSP/HTML/JavaScript
Web Services / WSDL
Application Components
Data Mining Algorithms
Outlier detection and removal
Ontology management
Knowledge inference / prediction logic
Data Management
Heterogeneous data integration (Parser / MirthConnect tool)
Data cleaning / noise removal
Query building
Clinical + Genomic Data

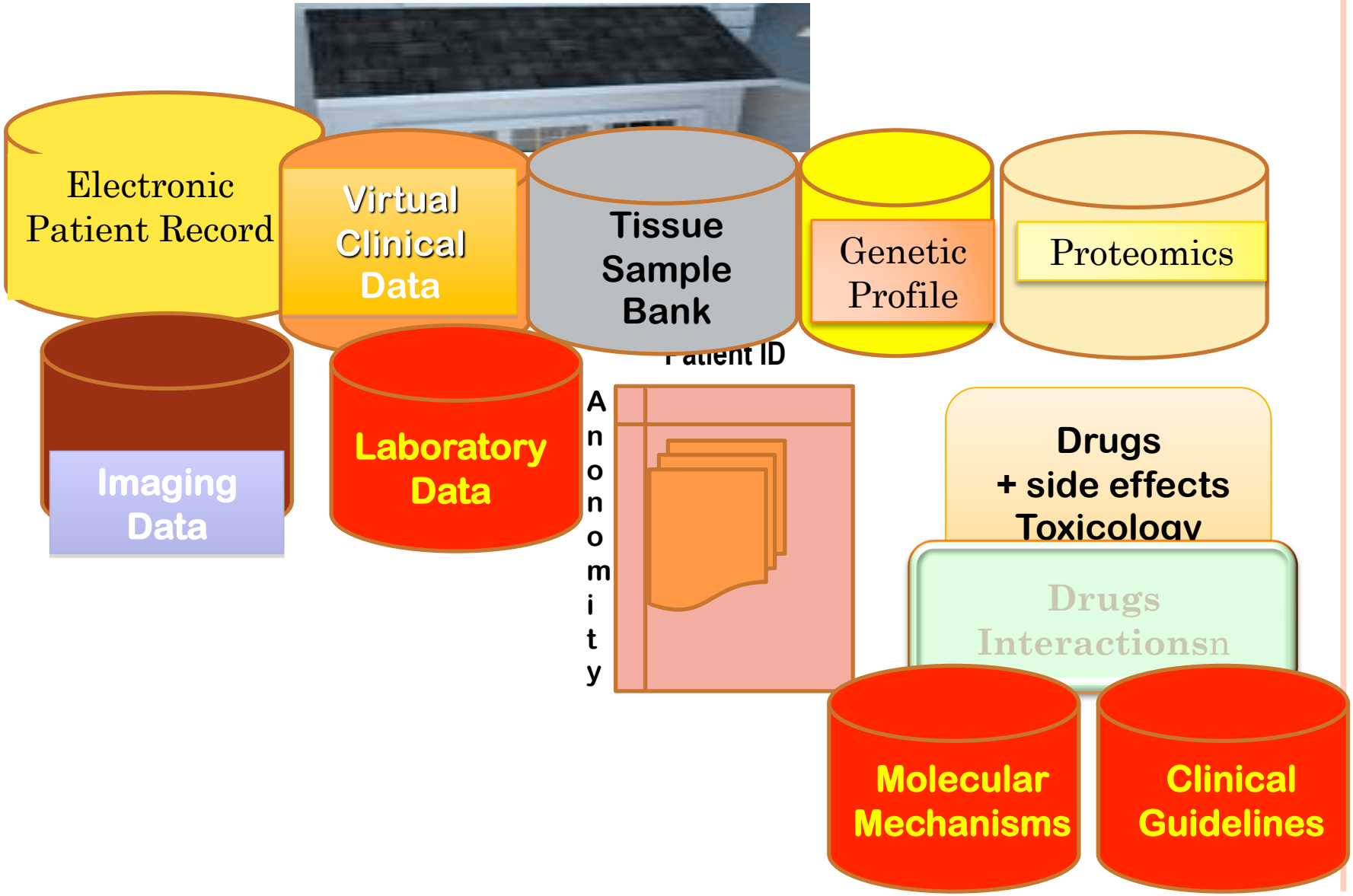
HETEROGENEOUS NATURE OF DATA

- 1) Difference in DBMS
  - Data Models (structures, constrains, query languages)
  - System Level Support (concurrency control, commit, recovery)
  - Semantic Heterogeneity
- 2) Operating System
  - File Systems
  - Naming, file types, operations
  - Transaction Support
  - Inter process Communication
- 3) Hardware
  - Instruction Set
  - Data Format & Representation
  - Configuration

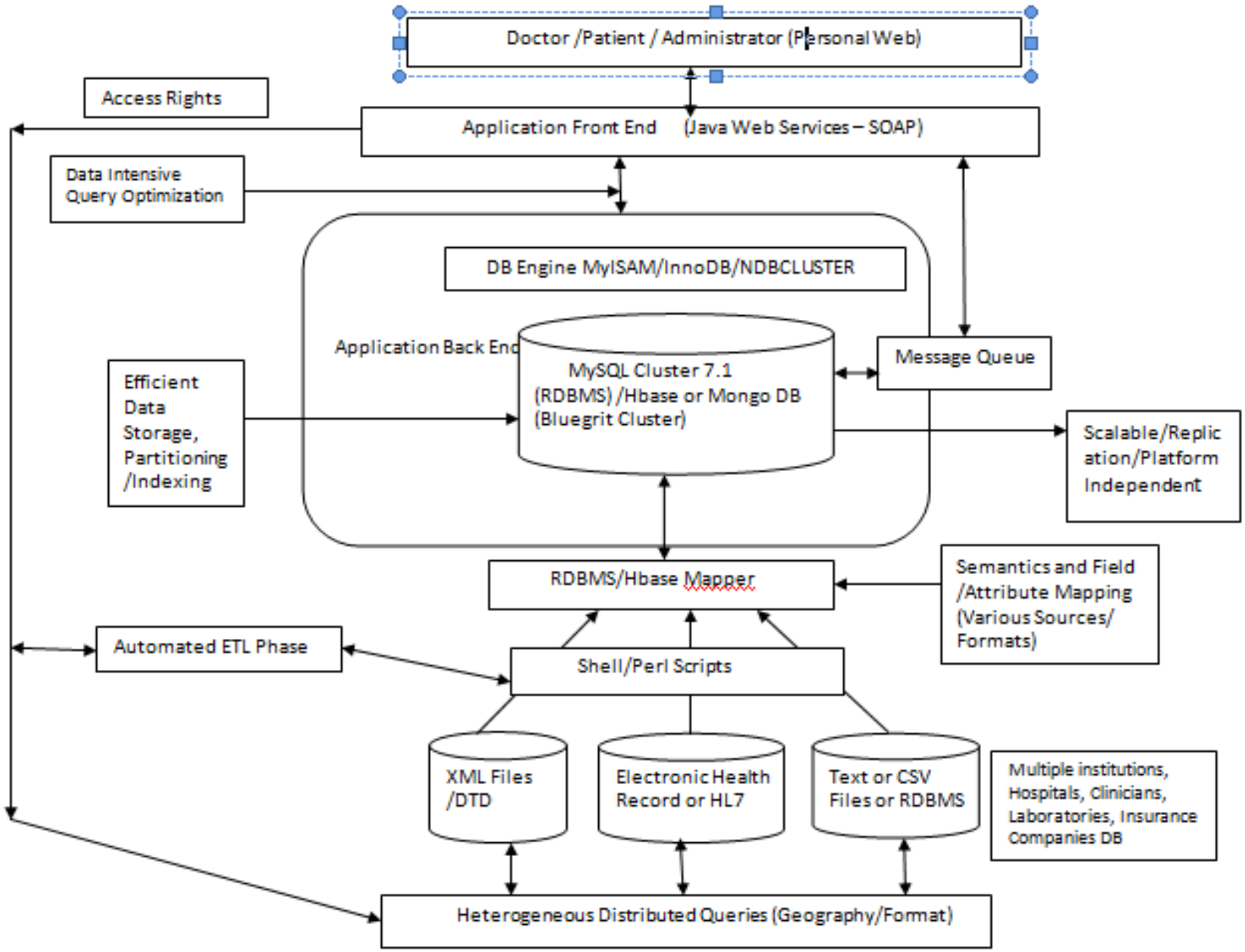
Medical Data Formats

- Clinical Data : From Clinics : HL7 (Health Level 7), XML,Text,CSV
- Genetic Data : XML, VCF, Text, CSV, etc
- Per patient :
  - Clinical data and genomic data not available in most cases and integration with the available is a challenge.

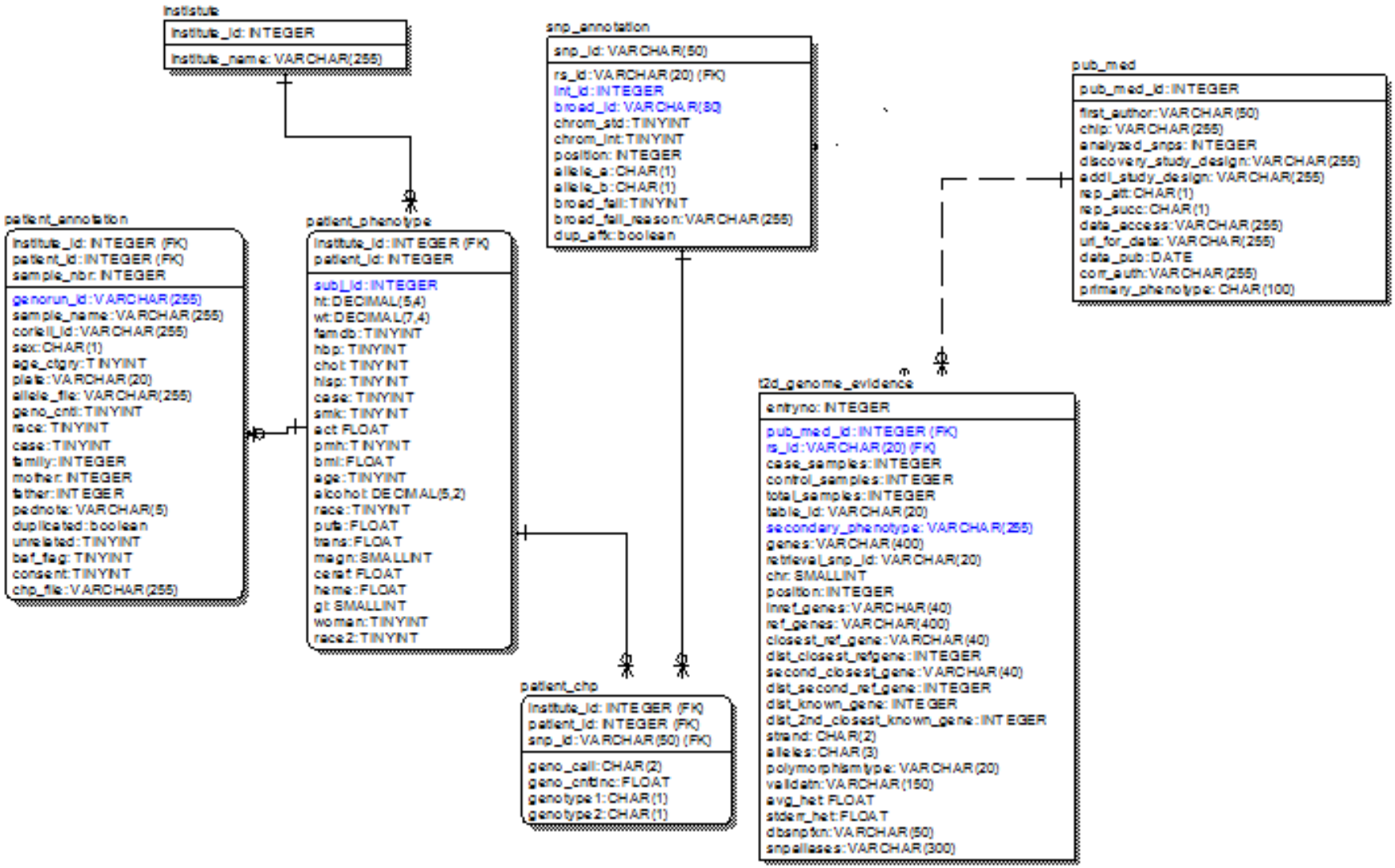
Existing and Potential Data Bases in Healthcare



Data Model


Doctor /Patient / Administrator (Personal Web)
Access Rights
Application Front End (Java Web Services – SOAP)
Data Intensive Query Optimization
DB Engine MyISAM/InnoDB/NDBCLUSTER
Application Back End
MySQL Cluster 7.1 (RDBMS) /Hbase or Mongo DB (Bluegrit Cluster)
Message Queue
Efficient Data Storage, Partitioning /Indexing
Scalable/Replication/Platform Independent
RDBMS/Hbase Mapper
Semantics and Field /Attribute Mapping (Various Sources/ Formats)
Automated ETL Phase
Shell/Perl Scripts
XML Files /DTD
Electronic Health Record or HL7
Text or CSV Files or RDBMS
Multiple institutions, Hospitals, Clinicians, Laboratories, Insurance Companies DB
Heterogeneous Distributed Queries (Geography/Format)

Database Schema



EXTRACT, TRANSFORM, LOAD

- Extracting from various data formats eg: relational or non –relational.

  -Parsing, Validation, etc

- Transformation according to target DB or business rules.

  -Automated column selection

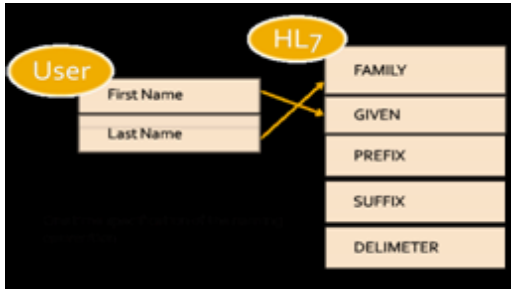
  -Mapping fields

  -Encoding values

  -Merge and Look up

  - Transposing or pivoting.

  - Proper data validation according to target schema.



- Load phase involves the proper structured files into to target DB.

PROBLEMS

- Data processing on voluminous data demands advanced computational resources.
- Deployment and continuous monitoring of database on Bluegrit Server.
- Extraction of medical data from various sources, formats and correlating them is a challenge.
- Loading stage needs proper formatting of incoming data to application database(eg XML or HL7 to RDBMS)
- Integrity among the databases should be maintained along with redundant data to be merged on the basis of various attributes.
- Automation of the ETL phase is a tough task.
- Scalability of the data model to accommodate more information.(eg :no. of keys,indexes,partitions,tables, architecture, platform)

SOLUTIONS

- Use of MySQL Cluster version deployed on Bluegrit with 1 mgt node and 2 data nodes to speed up processing.
- The data can be correlated using metadata files , online journals and consultation with doctors.
- Extraction of the data from various data sources can be achieved after getting controlled access to data sources.
- ELIXIR and MirthConnect tools can help to convert data from XML/HL7 to RDBMS respectively.
- Automation of extraction, transform and load phase with usage of scripts in conjunction with the tools.
- Migration to Hbase/MongoDB at later stage since the data is huge and not relational most of the times.

Personal web for Doctors & Patients

- Doctors and patients should be able to access the data from different institution with single login.
- Doctors should be able to map various data sources .
- Add, modify and delete existing records (controlled access).
- Access to patients historical clinical and genetic traits needs data to be archived.
- Access to recent journals and papers by doctors to make decisions needs integration of various journal DBs in future.
- Personalized pages needs doctors/patients for alerts, reminders,etc.

FUTURE TASKS

- Optimize performance to handle more patient's data from different institutions and various diseases for the web application.
- Automate tasks for variety of routine tasks related to the database(especially ETL).
- Testing various features of data model with respect to integration of modules.
- Test the compatibility of application and scalability on various platforms with respect to database, operating system.

Questions