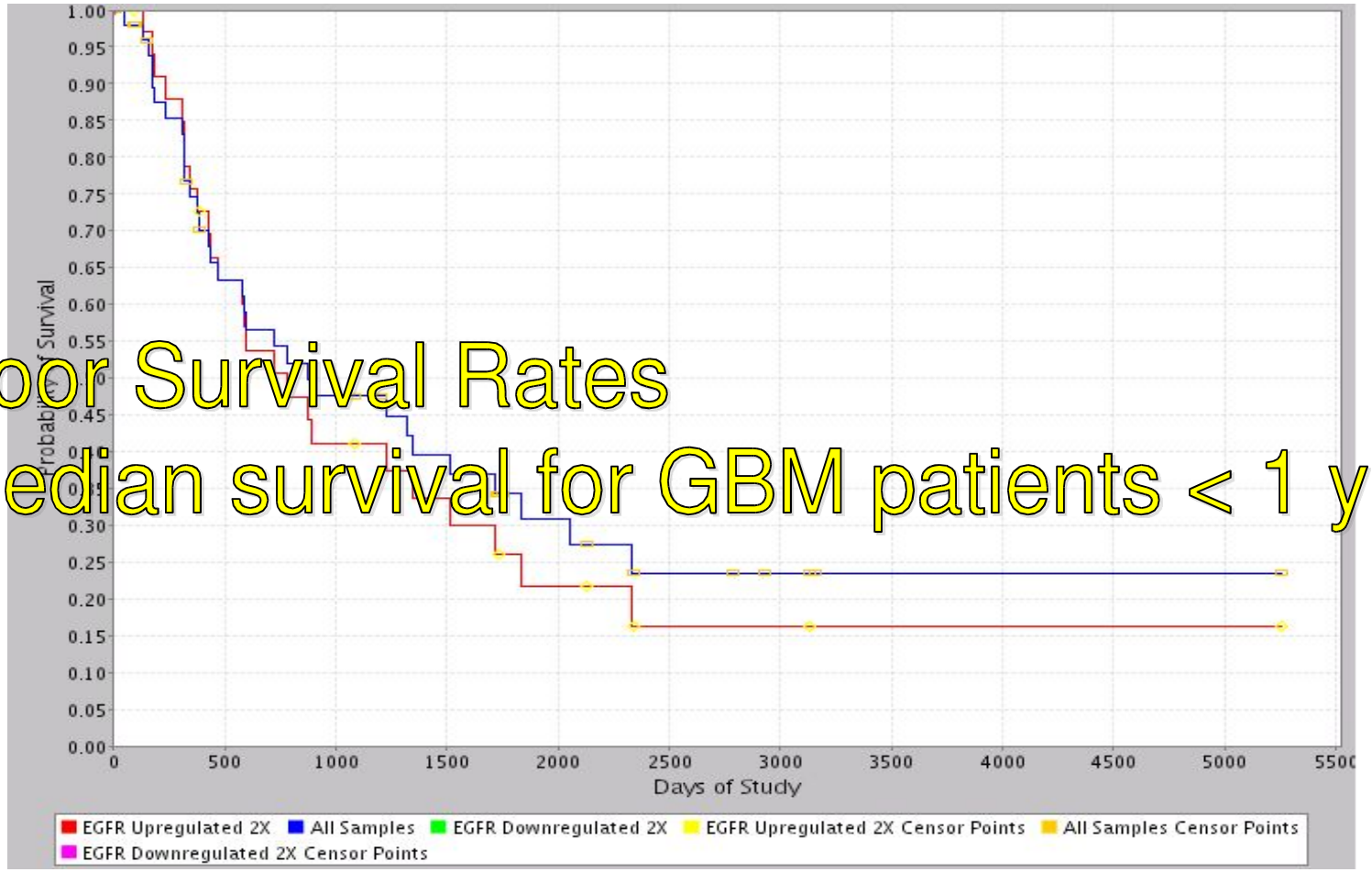
REMBRANDT:

Building a robust translational research framework for brain tumor studies

REpository of **M**olecular **BRA**in **N**eoplasia **Da**T**a**

Himanso Sahni
Center for Bioinformatics, NCI
SAIC


Debilitating Brain Tumors



Poor Survival Rates

Median survival for GBM patients < 1 yr

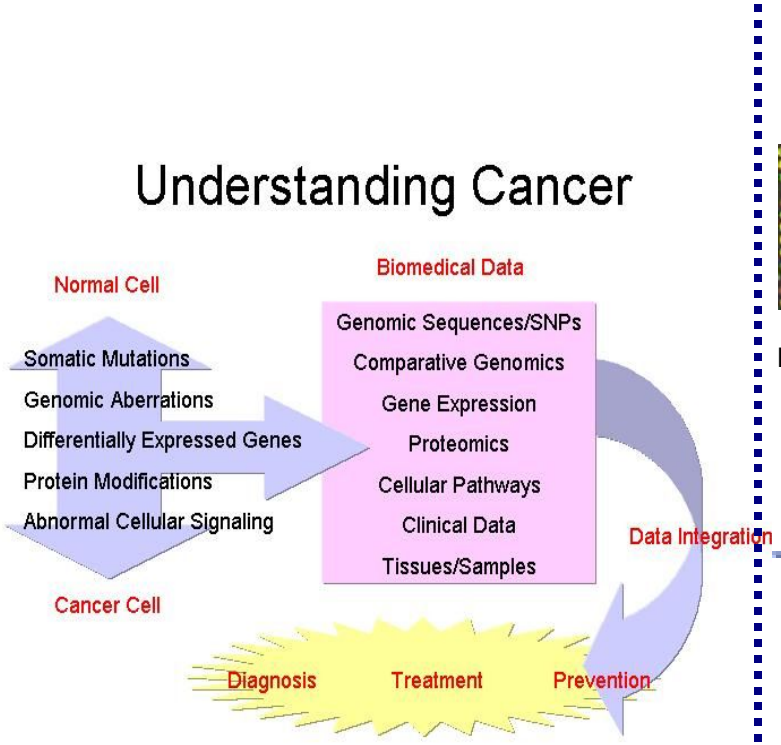

National Cancer Institute
U.S. National Institutes of Health | www.cancer.gov
REMBRANDT
home contact support
Repository for Molecular Brain Neoplasia Data.
Empowering translational research for brain tumor studies.
Let's Just do it
We are designing a robust bioinformatics knowledgebase framework called CaIntegrator that leverages data warehousing technology to host and integrate GMDI trial data. The knowledge framework will provide researchers with the ability to perform ad hoc querying and reporting across multiple domains.
Scientists will be able to answer basic scientific questions related to a patient or patient population and view the integrated data sets in a variety of scientific based contexts. Tools that link data to other annotations such as cellular pathways, gene ontology terms and genomic information will be embedded.
User Name:
Rembrandt
Password:
Submit
Reset
HOME | CONTACT | SUPPORT | NCICB HOME
FIRSTGOV

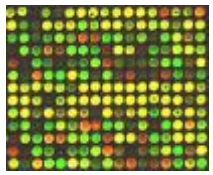
Challenges

- Few therapeutic advances in the last 3 decades
- Histopathological classifications for the heterogeneous group of tumors known as gliomas are broad and do not predict for therapeutic outcome or prognosis
- Standard therapies generally have minimal effect on long term survival

What can we do to help?

Rembrandt Knowledgebase

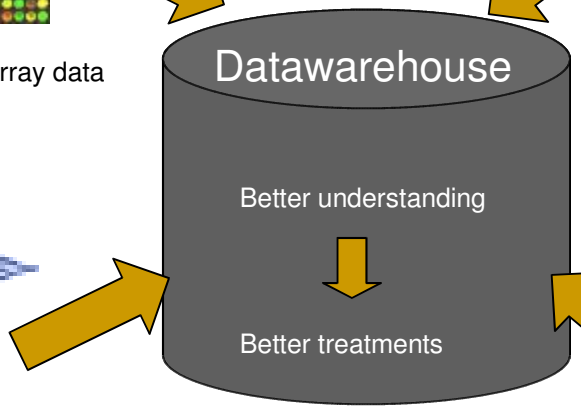


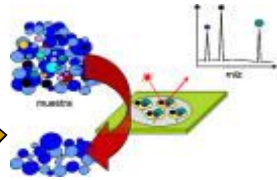


Expression array data

SNPArray data

Clinical data





Proteomics data

Concept → **Creation**

NCI's GMDI Study

Blood

Tumor

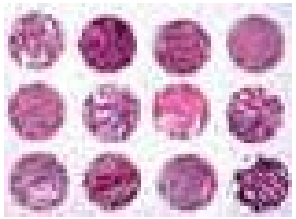
Plasma

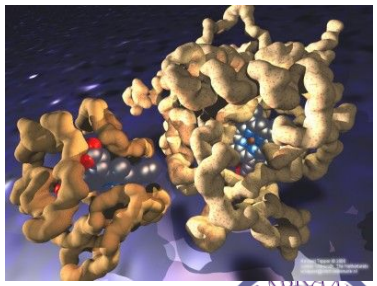
DNA

RNA

Tumor Core Punch

Proteins





Typical Rembrandt Usage Scenario

- In brain tissue from patients diagnosed with the glioblastoma multiforme (GBM) subtype of Astrocytoma, which genes in the EGF signaling pathway are over or under expressed in cancerous versus normal tissue?
- Is there a correlation between the expression and genomic (copy number) data collected from these patients?
- How did EGFR up-regulation affect survival of patients within this study?
- Of these groups of samples, which ones were obtained from patients that were males and were diagnosed between the ages of 25 and 40 yrs?

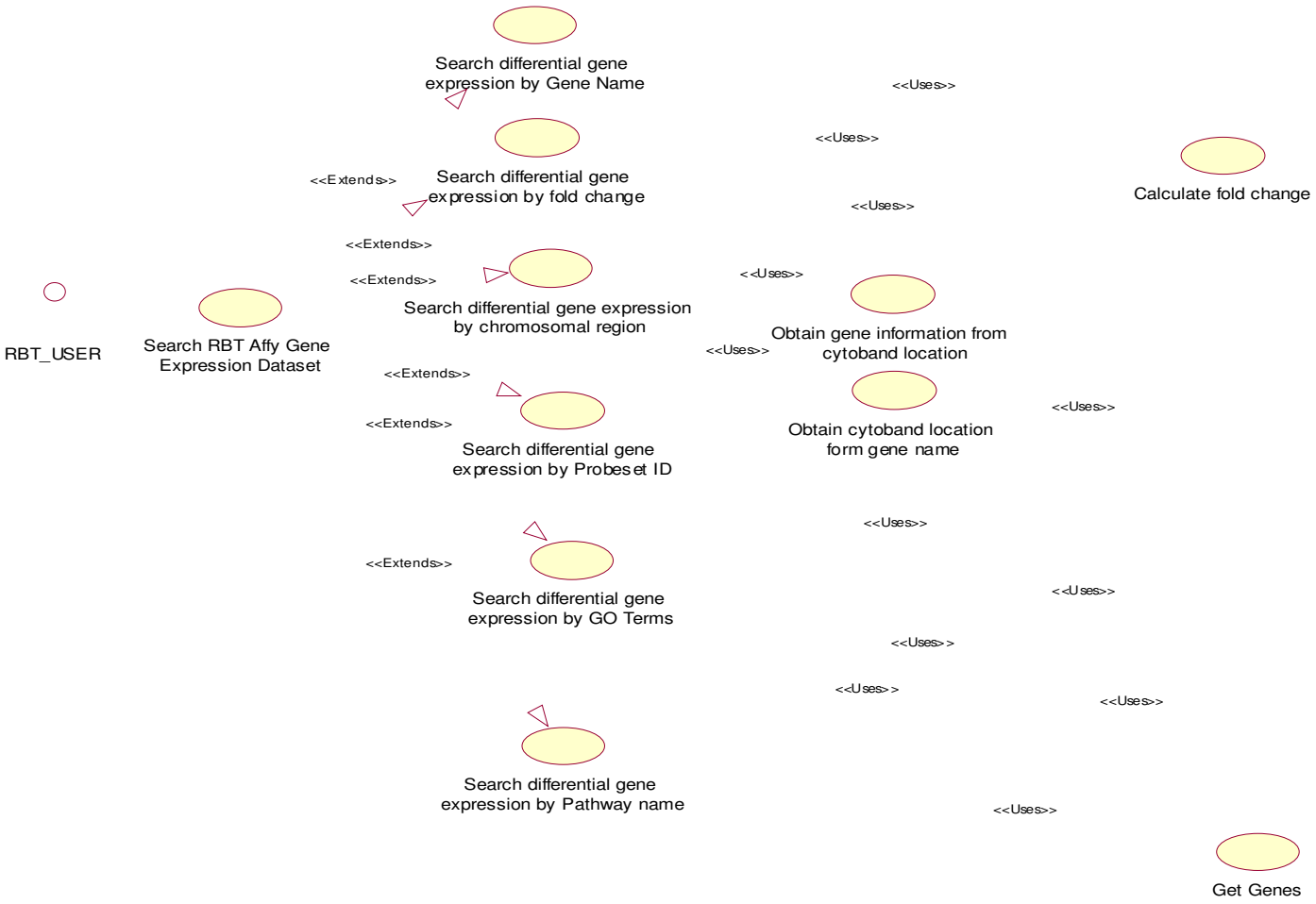
Rembrandt's Objectives

- Must support translation research use cases:
  - Build an infrastructure that provides users with the ability to create complex translational queries
  - For Example:
    - Ability to AND /OR a Gene Expression query with a Copy Number query and then further nest this within a Clinical Results Query
    - Ability to further refine the results by applying a criteria to the subset of samples grouped by high order analysis
    - Ability to apply filters to the result set for user friendly analysis.

Rembrandt's Objectives (cont'd)

- Allow users to view the results by easily pivoting between the various dimensions:
  - Grouped by Disease
  - Grouped by Patient / Sample
  - Grouped by Genes for Gene Expression or Cytogenic Location for Copy Number
  - View Associated Annotations
  - Time Course View (future)

Gene Expression Search Use cases



Rembrandt's caBIG objectives

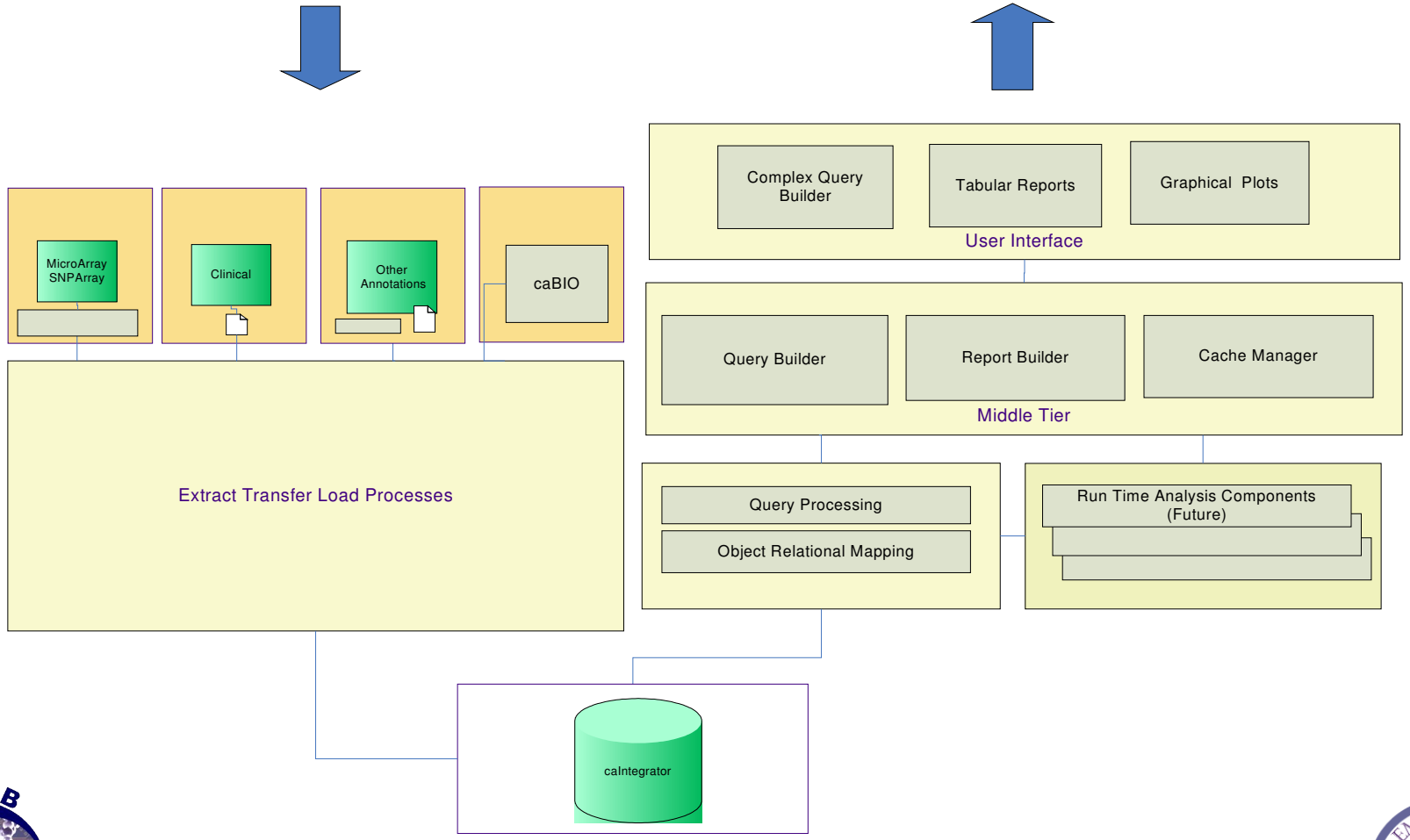
- Aligns with NCI's caBIG (cancer Biomedical Informatics Grid) principles:
  - Open source
  - Open access
  - Syntactic and Semantic interoperability
  - Federated access
- Leverage NCICB and caBIG Infrastructure Components
  - caCORE Infrastructure (caBIO, EVS, caDSR)
  - caARRAY gene expression data repositories and analysis tools
  - C3D Clinical Informatics System
  - caBIG Infrastructure being delivered by caBIG workspaces

See https://cabig.nci.nih.gov/

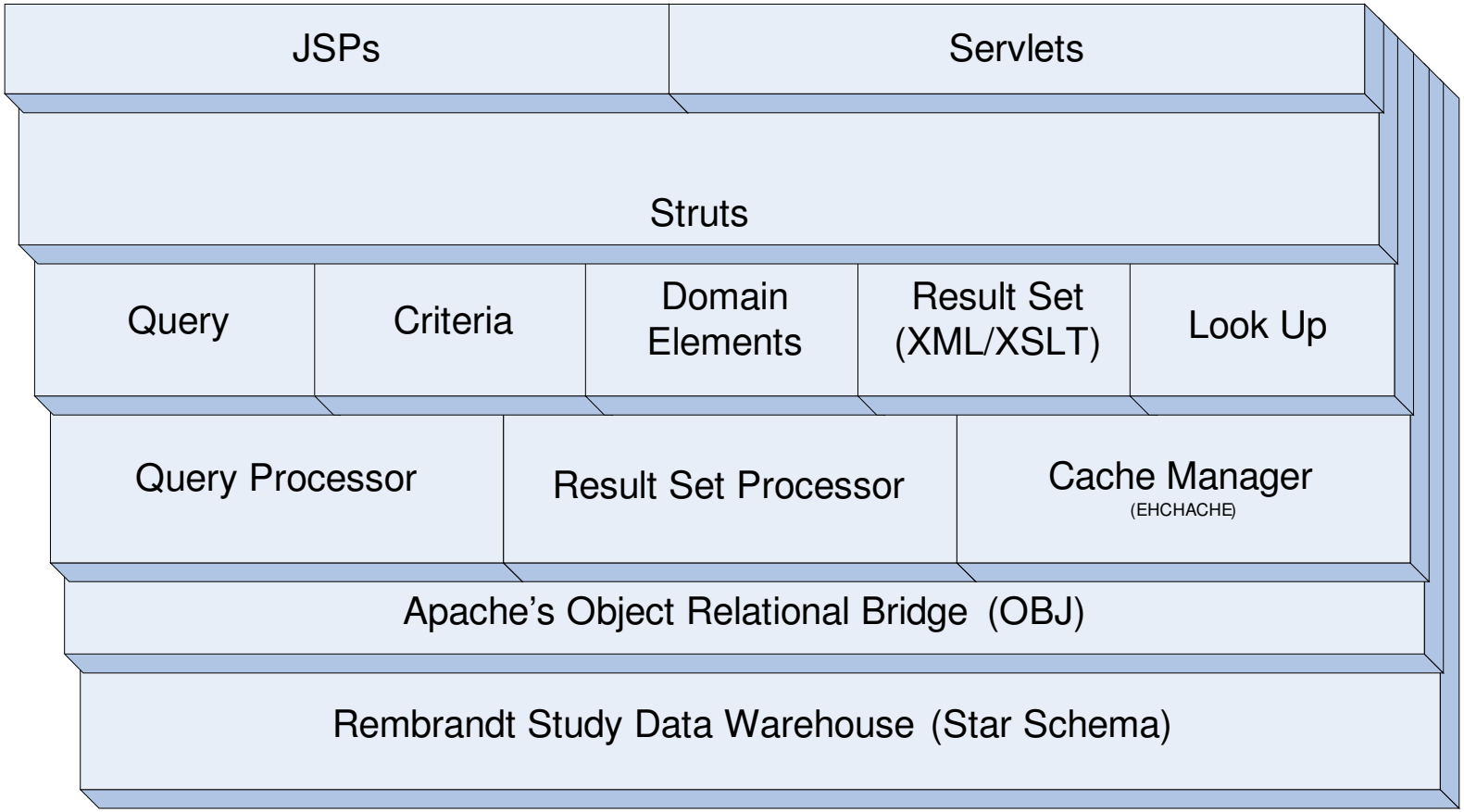
Rembrandt Technical Objectives

- Build a scalable high performance application
  - Tiered Architecture
  - Abstraction / Model View Controller
- Support Strong Type Checking & Validations
- “Fast” Queries
- User Friendly Interface
- Groundwork for a robust translational research framework

Rembrandt Current Architecture



Another Architecture Perspective

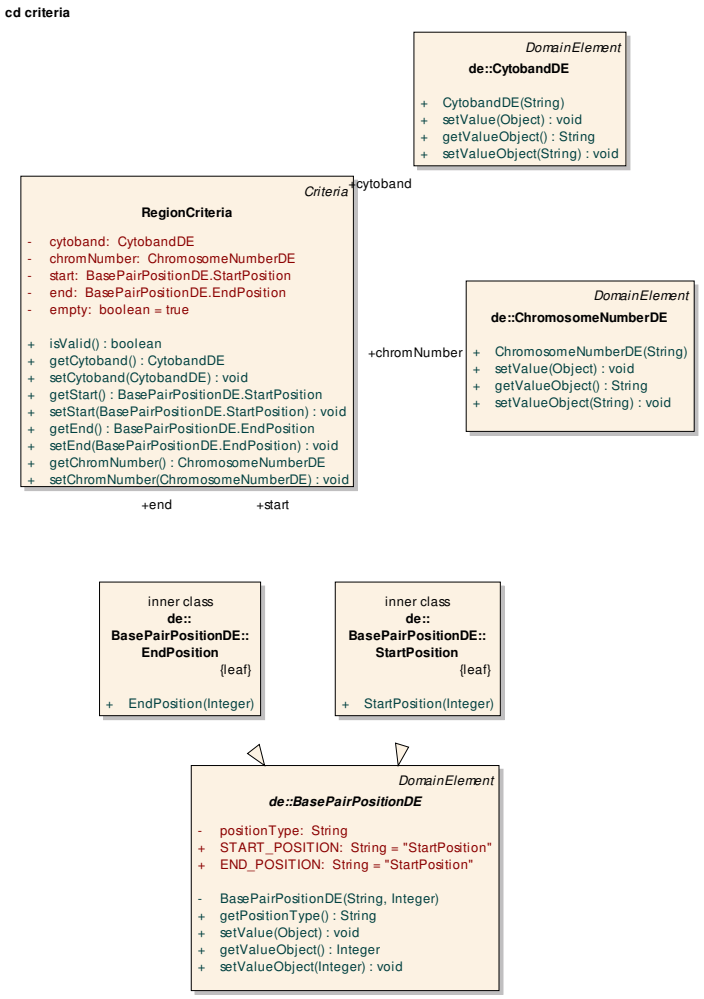


Query & Retrieval Objects :
Support Strong Type Checking & Validations

- Such as Query, View, Criteria, Domain Element objects
  - Abstracts presentation logic from the query helper objects
  - Provides the ability to nest cross domain queries (AND/OR)
  - Is strongly typed
  - Can validate itself

Example: Criteria Objects

- **Criteria Object**
  - Consist of DomainElements
  - Provide Generic Cross Domain Filters
  - Each Criteria can validate itself
  - For e.g.: RegionCriteria
    - Consists of ChromosomeNumberDE, CytobandDE, BasePairPositionDEs for start & end positions.
    - Is used in both Gene Expression and Comparative Genomic domain queries



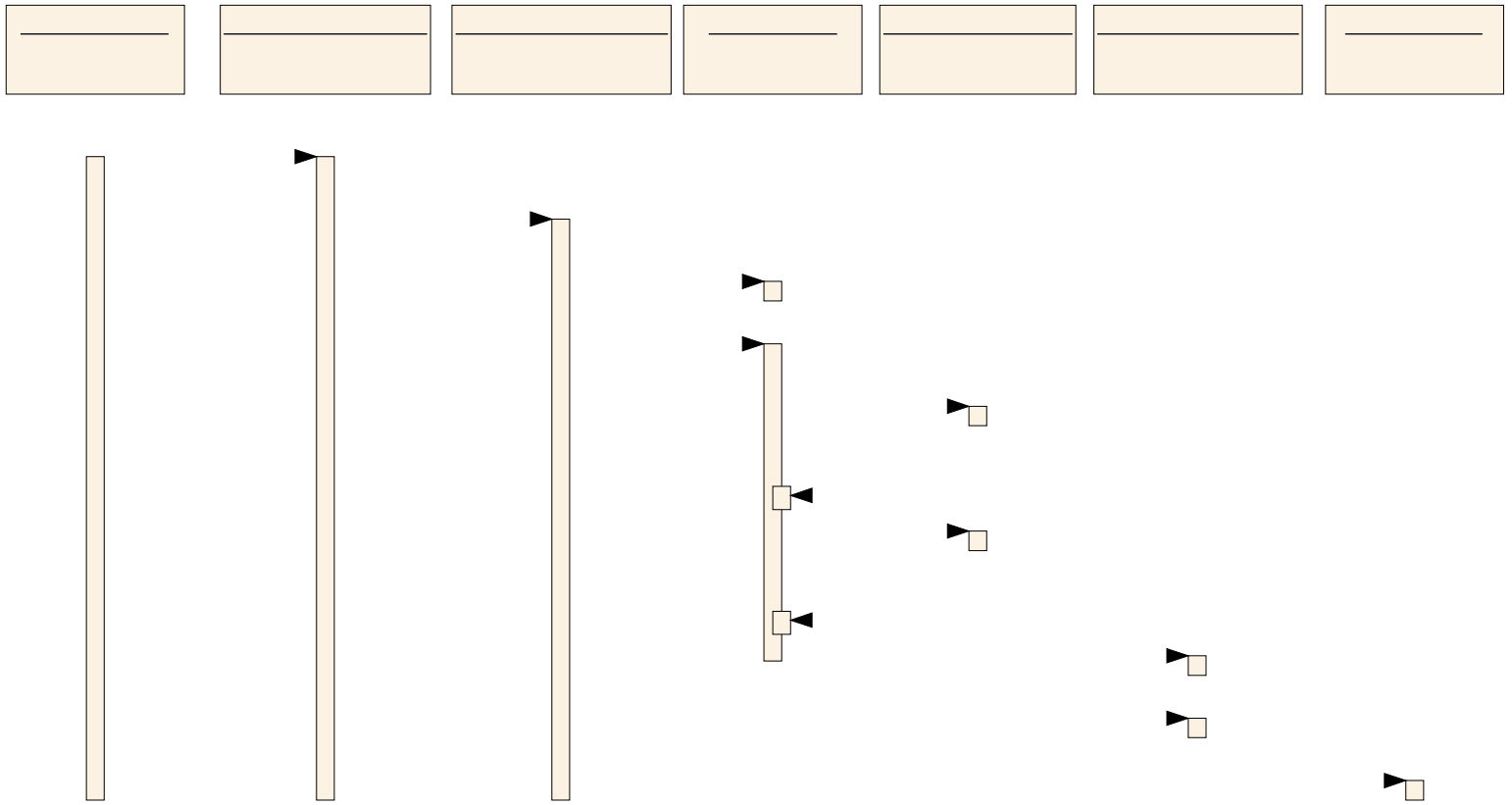
Agnostication can result in Obfuscation…

- *Challenge: Making Rembrandt dB agnostic using a standard Object Relational Mapping (ORM) layer AND still create high performance queries.*
  - Currently using Apache's Object Relational Bridge (OJB) as the ORM layer.(http://db.apache.org/ojb/)
  - All ORMs provide great abstraction but may not help produce the most efficient SQL.
  - Custom implementations or extending frameworks can become a maintenance nightmare.

High Performance Query Processing

- Multi-threaded Query Processing:
  - All queries are constructed and executed in parallel on separate threads from Java server side
- Dimensional Result Set Processing
  - All result set dimensions are reconstituted in Java server side
- For example:
  - The entire Chromosome 7 (1 and 15854551 bp)
  - Able to retrieve about 51,000 fact records plus all associated annotations and display results for all 51 samples in 20 sec.

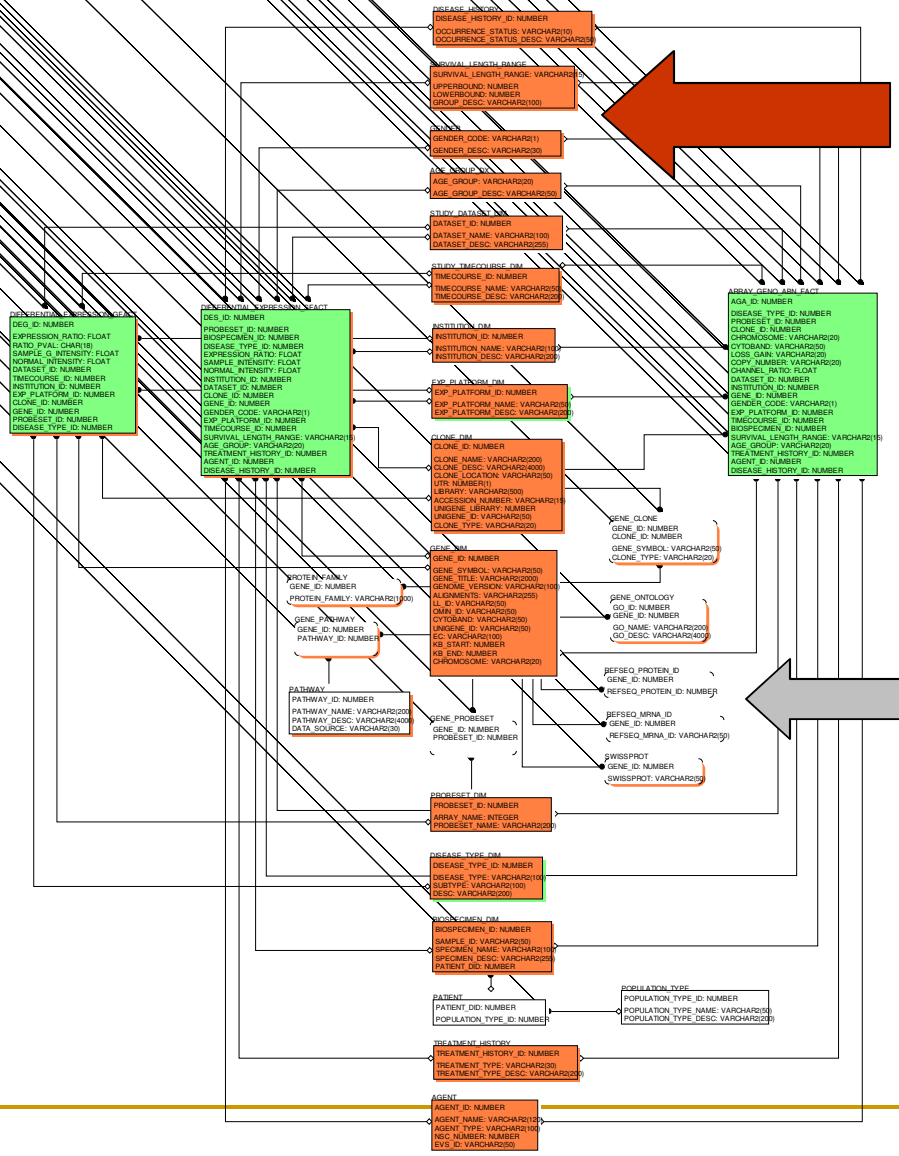
Multi-threaded Query Processing in Java



Rembrandt Data Warehouse Schema

- Highly de-normalized, query optimized star schema
  - The Fact tables contain all the pre-calculated data points based on various scientific algorithms.
  - The dimension tables contain study relevant data points, such as clinical information, genomic annotation information, etc.
  - Lookup tables and mapping tables provide static general information, such as gender, etc.

Rembrandt Data Warehouse Schema



Dimensions

(PROBESET_DIM,
CLONE_DIM,
DISEASE_DIM, etc)

Fact Tables

(DIFFERENTIAL_GENE_SFACT ,
DIFFERENTIAL_GENE_GFACT ,
ARRAY_GENO_ABN_FACT)

Lookup/Mapping Tables

Caching Strategy

- *Challenge: Provide the ability for users to quickly view reports in a different dimension and easily retrieve previously executed reports*
  - Executed reports are cached for each user session
  - Provides performance and scalability
  - Using EHCHACHE (http://ehcache.sourceforge.net/)

Report Transformation Using XSLT

Challenge: User friendly reports

- Generate XML from Result set Objects using Dom4J (http://www.dom4j.org)
- Apply XSLT to render the reports
- Allows us to provide the users with ability to
  - Filter/ Highlight data
  - Sorting of results
  - Pagination
  - CSV Generation
  - XML Import/Export
  - Multiple “Styled” views per study
  - XHTML compliance
  - Browser Compatibility (various styles based on user agent)
- XSLT uses XPath to define the matching patterns for transformations

Groundwork for a robust translational research framework

- *Challenge: Lay the foundation for a clinical genomic framework that…*
  - Integrates Clinical data with Experimental data
  - Provide researchers with the ability to perform complex ad hoc querying, real time analysis and reporting across multiple domains.
  - Generic enough to support other similar clinical genomic studies such as I-SPY

Other similar studies …
I-SPY Trial

***I**nvestigation of*
***S**erial Studies to*
***P**redict*
***Y**our*
***T**herapeutic*
***R**esponse with*
***I**maging*
***A**nd*
*mo**L**ecular analysis*

Courtesy: Laura Esserman, Director, UCSF CF Buck Breast Care Center

Goals for future releases…

- caBIG Silver Level compliance
  - Clinical Genomic Object Model
  - Domain-based Clinical Genomic Object API
- Gateway Portal that provides links to other NCICB/caBIG transaction systems for study based data submission

Goals for future releases…(cont'd)

- Package a suite of utilities that can be applied to other similar translational projects
  - Database creation utilities
  - Data retrieval utilities
  - Transformation/Pre-processing utilities
  - Data loading utilities
  - Higher-order analysis components
  - Visualization components

- **NCICB Development team**
  - Dave Bauer
  - Ram Bhattaru
  - Alex Jiang
  - Ryan Landy
  - James Luo
  - Ying Long
  - Subha Madhavan
  - Kevin Rosso
  - Himanso Sahni
  - Prashant Shah
  - Nick Xiao
  - Dana Zhang
- **NCICB Advising team**
  - Scott Gustafson
  - Sharon Settnek
  - Carl Schaefer
  - Mervi Heiskanen
  - Sue Dubman
  - Peter Covitz
  - Ken Buetow
- **NOB/CCR/NCI**
  - Howard Fine
  - JC Zenklusen
  - Yuri Kotliarov
  - Tracy Lively
- **NINDS**
  - Bob Finkelstein
- **UCSF**
  - Laura Esserman

Contact Information:

- Application: http://rembrandt-db.nci.nih.gov
- Project: http://rembrandt.nci.nih.gov
- Project Manager: madhavas@mail.nih.gov
- My Email: sahnih@mail.nih.gov

Demo

Application: http://rembrandt-db.nci.nih.gov
Informational Site: http://rembrandt.nci.nih.gov

A more in-depth demonstration of the application will be presented by

Subha Madhavan

Tuesday, June 28th 3.00 PM to 4.00 PM
at Lasalle conference room.