Database support for high throughput data analysis web site

BY:

Sajni Chowrira

ADVISORS:

Dr. Lawrence Hunter

Dr. Tzu L. Phang

Genes & mRNA

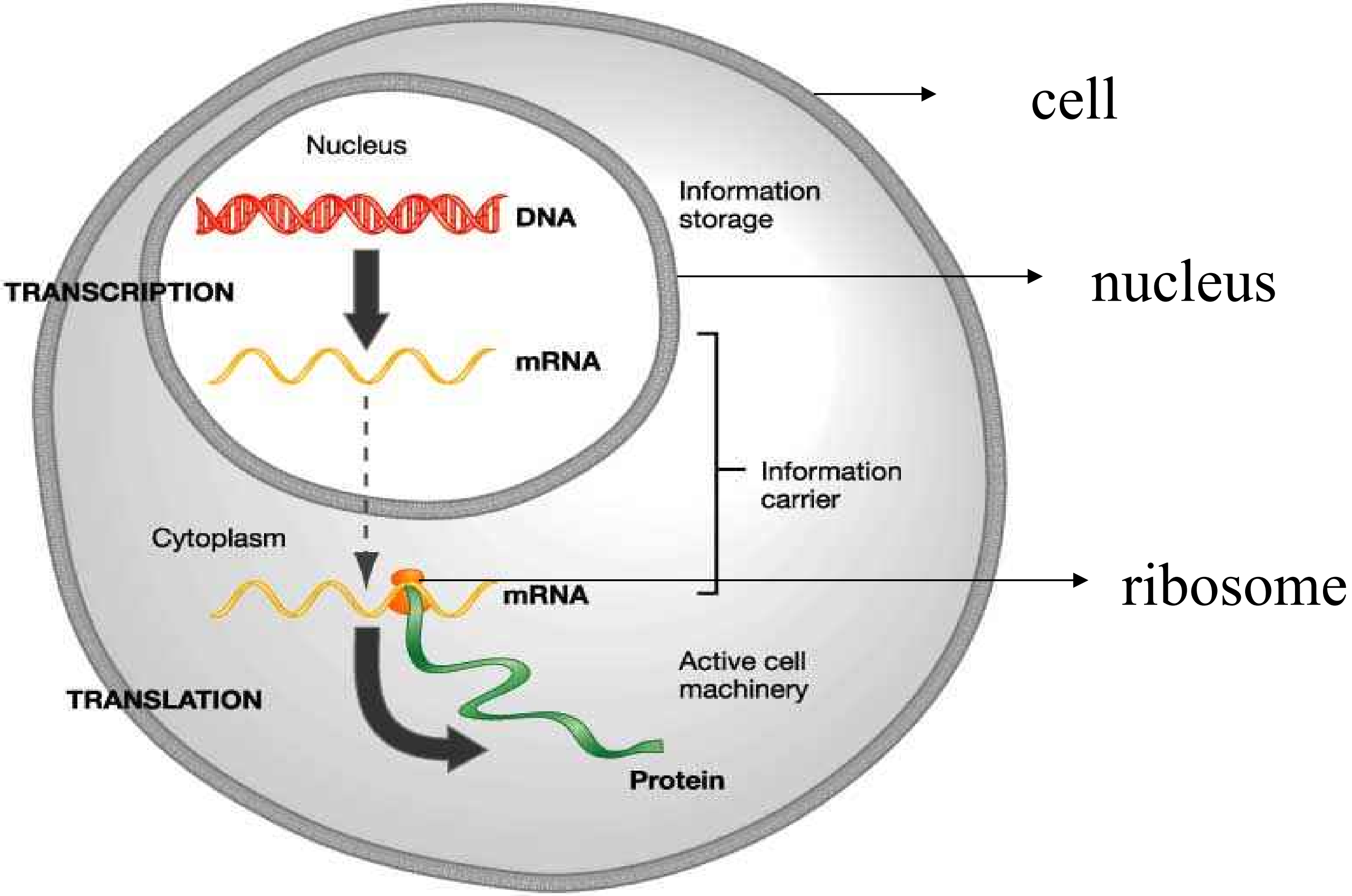
Gene: Segment of a DNA that codes for a protein

How is protein formed ?

1. DNA acts as a template to create the complementary RNA (messenger RNA) - Transcription
2. RNA carries code for protein synthesis
3. Ribosomes (a cell organelle) read this code and use it to build the protein - Translation

Genes & mRNA

DNA → mRNA → Protein



cell

nucleus

ribosome

Genes & mRNA

- All human cells contain identical genetic material
- The same genes are not active in every cell – genes can turn on and off
- Studying gene expression helps
  - understand gene behavior in diseased cell
  - develop drugs to help cell function normally again

Genes & mRNA

- Expression of a certain gene controls concentration of its corresponding mRNA
- Hence measuring mRNA concentration helps understand gene activity
- mRNA concentration can be measured using microarrays

Microarray Technology

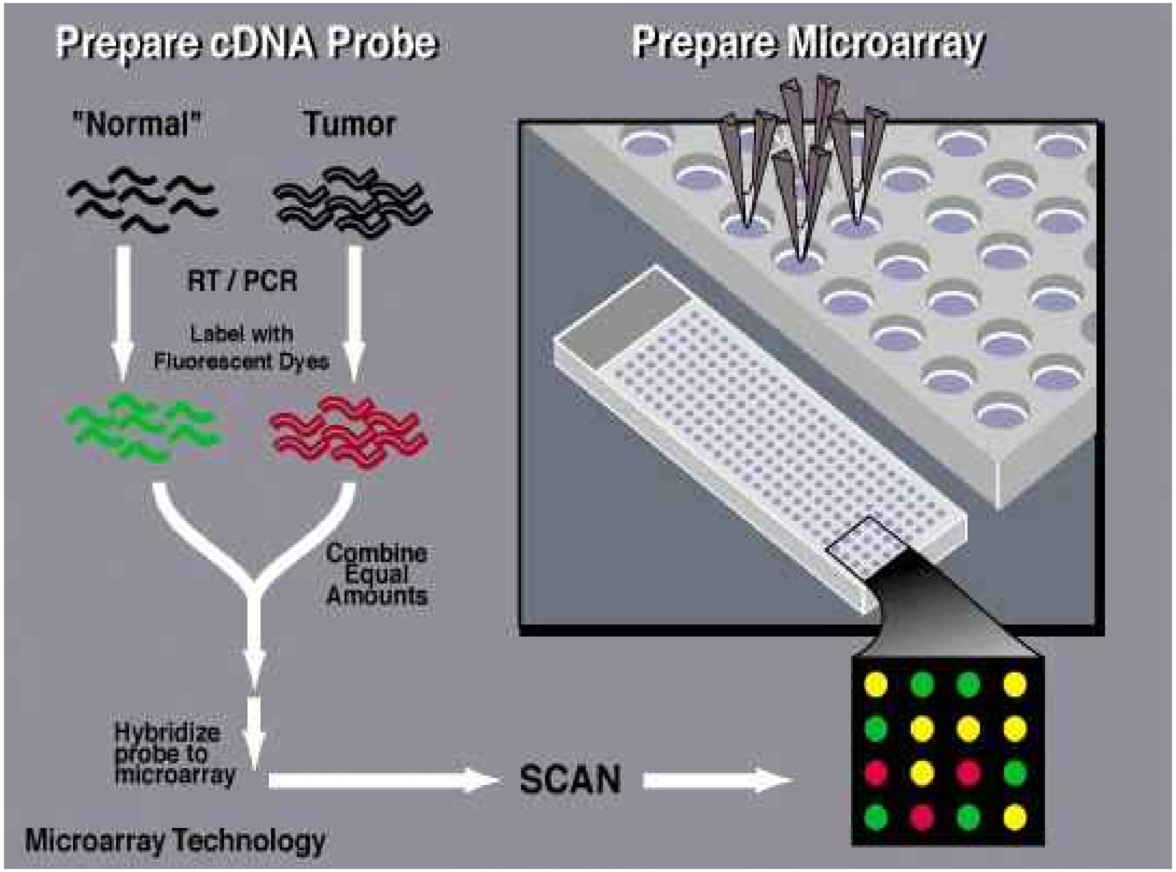
Microarray

- also known as DNA chip or gene chip
- consists of a piece of glass or plastic
- single-stranded pieces of DNA are affixed in a microscopic array (probe)
- hundreds or thousands of identical DNA molecules are affixed at each point

Microarray Technology

1. Isolate the total mRNA molecules in normal and diseased cells
2. Label mRNA molecules by attaching a fluorescent dye – one color for normal mRNA & another color for mRNA from diseased cell
3. Both extracts are washed over the microarray
4. mRNA binds to their complementary DNA sequences on the microarray (hybridization)

Microarray Technology



Microarray Technology

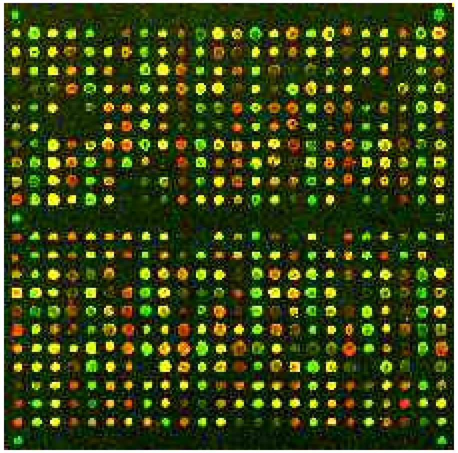
5. Wash away unhybridized material - fluorescent tags left behind
6. Use a special scanner (with laser, camera, microscope) to scan the chip
7. Laser excites the fluorescent dyes
8. Microscope and camera work together to create a digital image of the array

Microarray Technology

9. A special program is used to measure the intensities of the cells

10. It also creates a table that contains the ratios of the intensities of the 2 colors for every spot on the array.

Microarray Technology



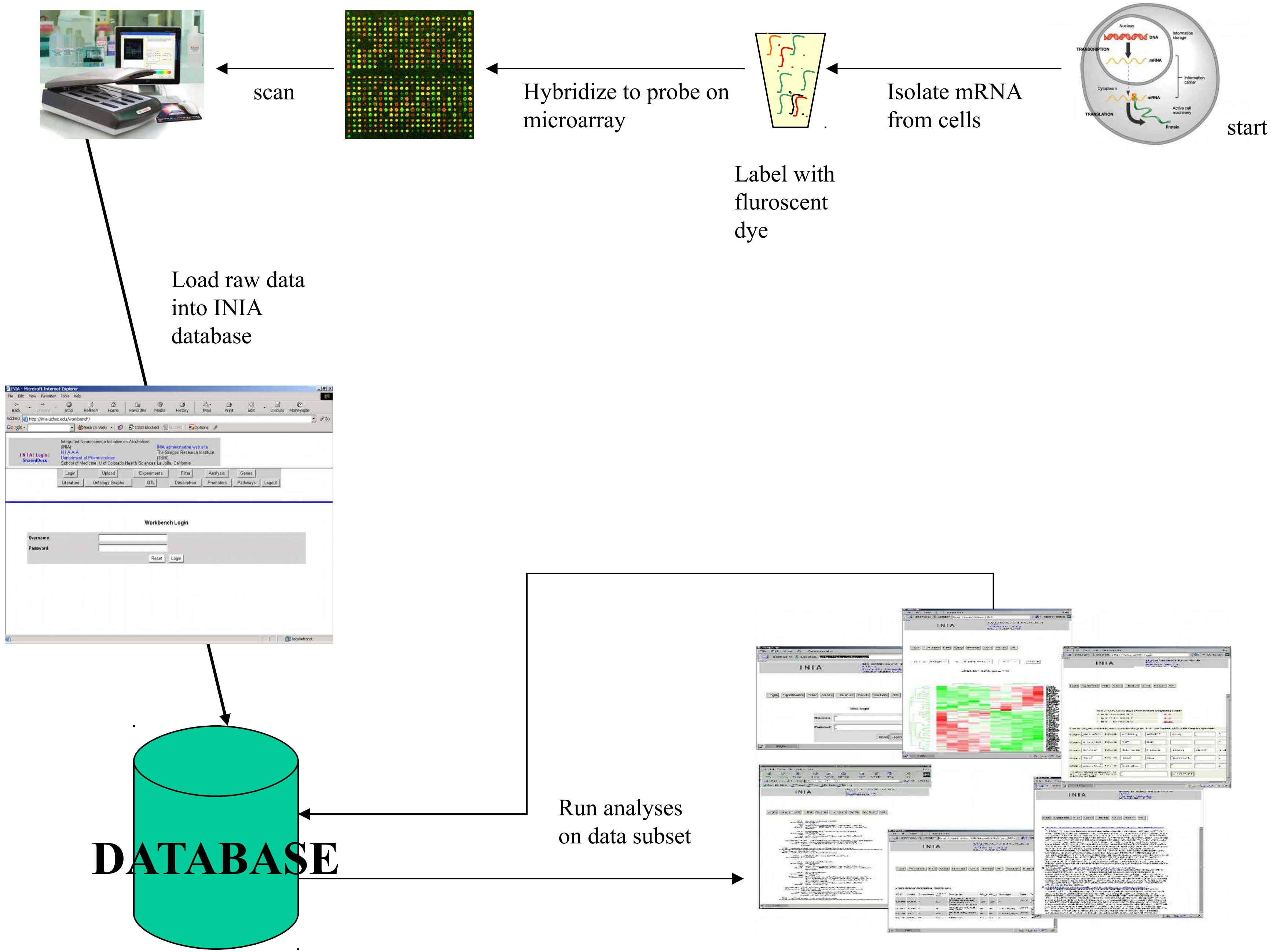
Microarray chip

Microarray Technology

- Increased fluorescence => cell recently transcribed
- Decreased fluorescence => cell ceased transcription
- intensity of the fluorescense
  - proportional to the number of copies of a particular mRNA that were present
  - indicates the activity or expression level of that gene.
- Arrays can paint a picture or "profile" of which genes in the genome are active in a particular cell type and under a particular condition.

INIA

- Integrated Neuroscience Initiative on Alcoholism
- website for the Neuroinformatics Core of INIA - developed by members of the Department of Pharmacology, University of Colorado Health Sciences Center
- The goal of the neuroinformatics core is to create an integrated repository of neuroscience data, ranging from molecules to behavior, for collaborative research on alcoholism



Nucleus
DNA
Information storage
TRANSCRIPTION
mRNA
Information carrier
Cytoplasm
mRNA
TRANSLATION
Active cell machinery
Protein

start

Isolate mRNA from cells

Label with fluroscent dye

Hybridize to probe on microarray

scan

Load raw data into INIA database

INIA - Microsoft Internet Explorer

Address http://inia.uchsc.edu/workbench/

File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Favorites Media History Mail Print Edit Discuss MoneySide
Go
Google Search Web 1150 blocked AutoFill Options

I N I A | Login | SharedDocs

Integrated Neuroscience Initiative on Alcoholism (INIA)
N I A A A
Department of Pharmacology
School of Medicine, U of Colorado Health Sciences

INIA administrative web site
The Scripps Research Institute (TSRI)
La Jolla, California

Login Upload Experiments Filter Analysis Genes
Literature Ontology Graphs QTL Descriptron Promoters Pathways Logout

Workbench Login

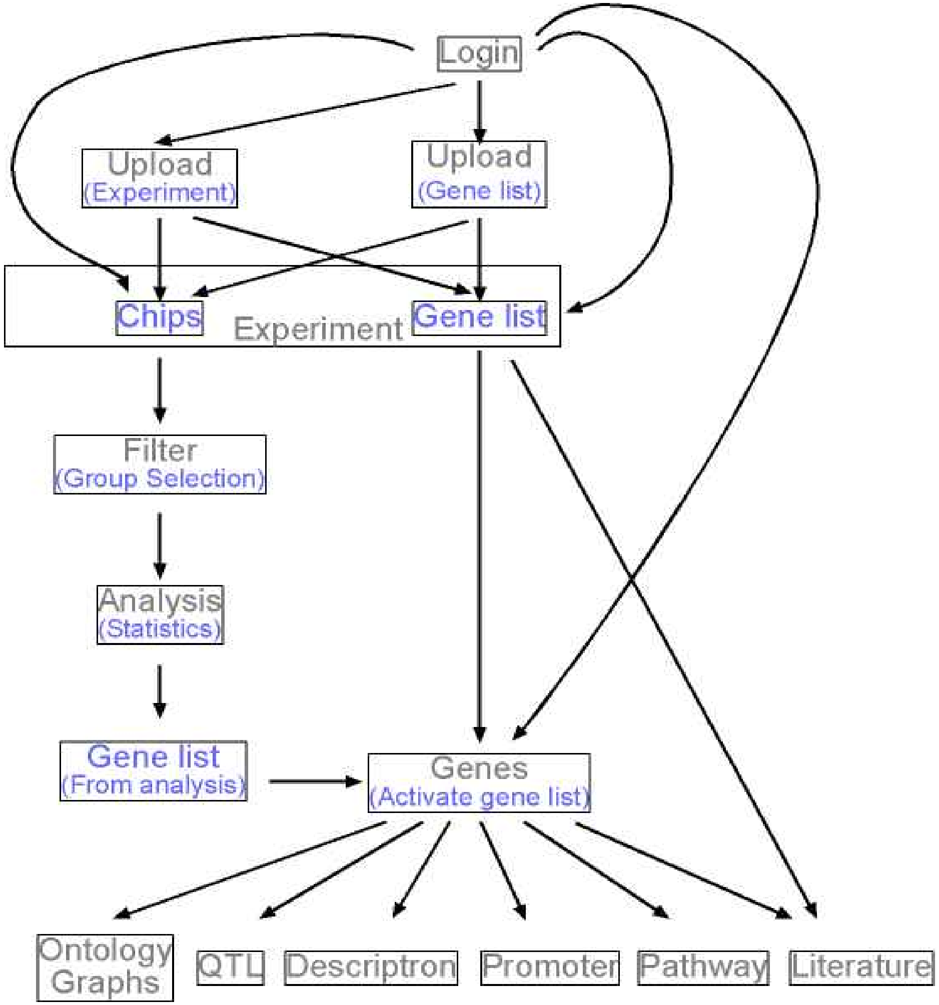
Username:

Password:

DATABASE

Run analyses on data subset

INIA



General analysis flow for the INIA workbench

MGED Standards

Need to standardize the recording & reporting of microarray-based gene expression data

- A single microarray experiment can produce several million pieces of data
- gene expression data is very complex it cannot be interpreted without additional information about the conditions under which they were generated
  - biological material, experimental design, array composition and design, hybridization conditions

MGED Standards

- experiments use different platforms and designs, the results could differ
  - in units
  - in format

Microarray Gene Expression Data (MGED) Society has defined standards for microarray data annotation and exchange . Example: MIAME

MIAME

- **M**inimum **I**nformation **A**bout a **M**icroarray **E**xperiment
- outlines the minimum information that should be reported
- enable its unambiguous interpretation and reproduction of an experiment
- Requires 6 pieces of information to be described for any published microarray-based gene expression experiment

MIAME

1. Experimental design – title, author information, contacting the author, a link to the online publication
2. Array design – details of the manufacturer, platform type, surface and coating specifications of array
3. Samples (biological substance for which gene expression is being measured) – source of the sample and how it was obtained, treatment information if the sample was treated before being used

MIAME

4. Hybridization – conditions under which the hybridizations were carried out, instruments and procedures used and protocols used for the hybridization, blocking and washing.
5. Measurements – This is where the results of the experiment are described
6. Normalization controls – normalization algorithm and strategy used, steps involved in preparing the hybridization extract

Problems: Molecular Biology Experiments

- Produce large volume of data
- Data hard to interpret
- Data differ greatly
- Data relationships difficult to model
- New data types constantly evolving
- Analyses produce new data types
- Results are difficult to manage

Need database that handles all the above

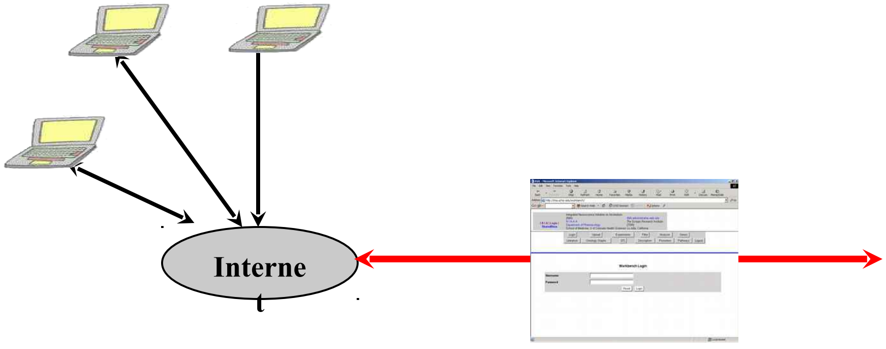
Neocore XMS

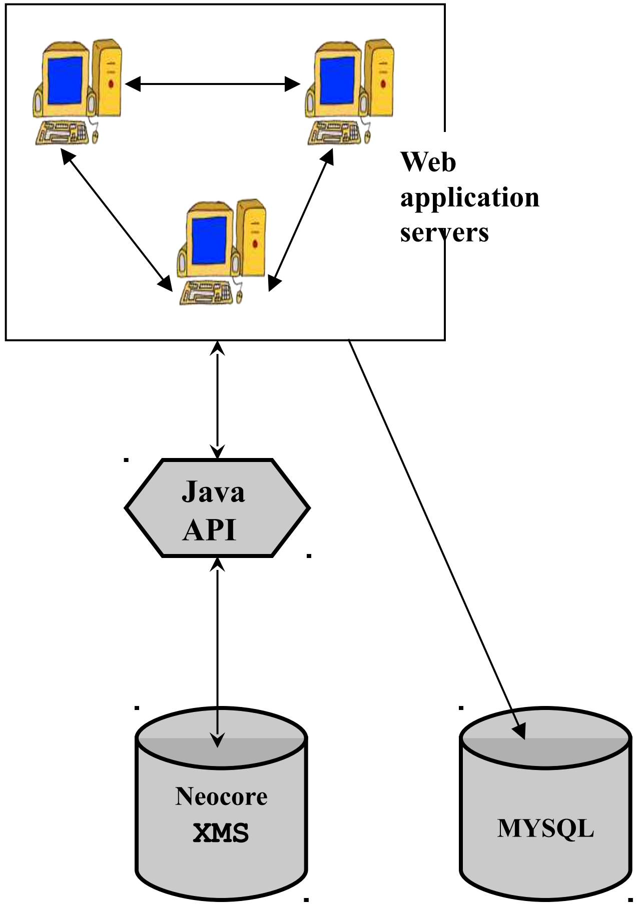
- XML Information Management System
- Not a relational database
- Stores Information = Data + Context
- XML in – XML out
- Stores hierarchical forms of data expressed in XML
- Information can be queried, modified and deleted

Neocore XMS

- Can be accessed through APIs like Java, C++, etc.,
- Self-constructing – requires no design
- Query language – subset of Xpath
- Self-constructing – requires no design

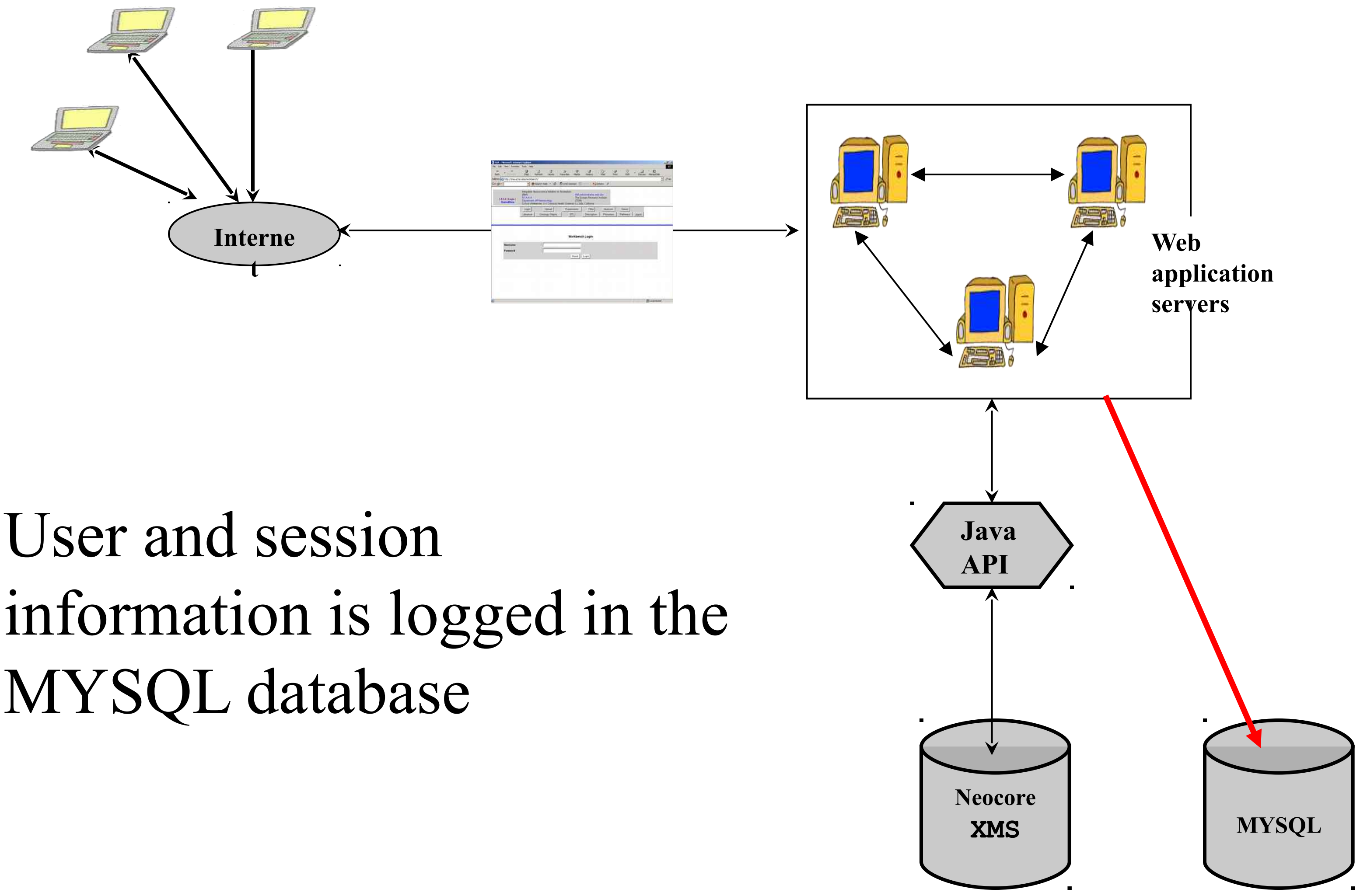
Design & Implementation





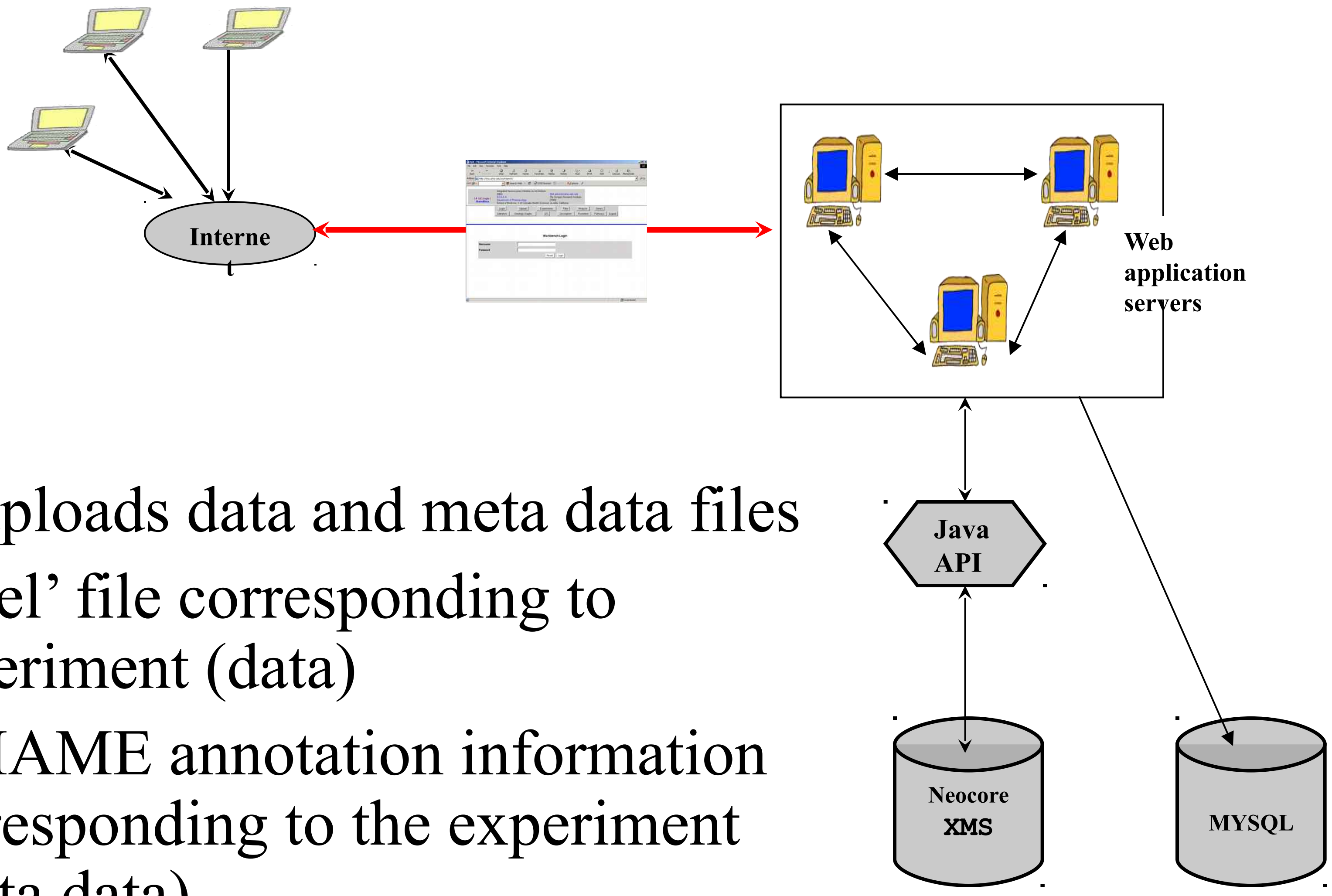
User enters login information and gains access into application

Design & Implementation



User and session information is logged in the MYSQL database

Design & Implementation



User uploads data and meta data files

–'.cel' file corresponding to experiment (data)

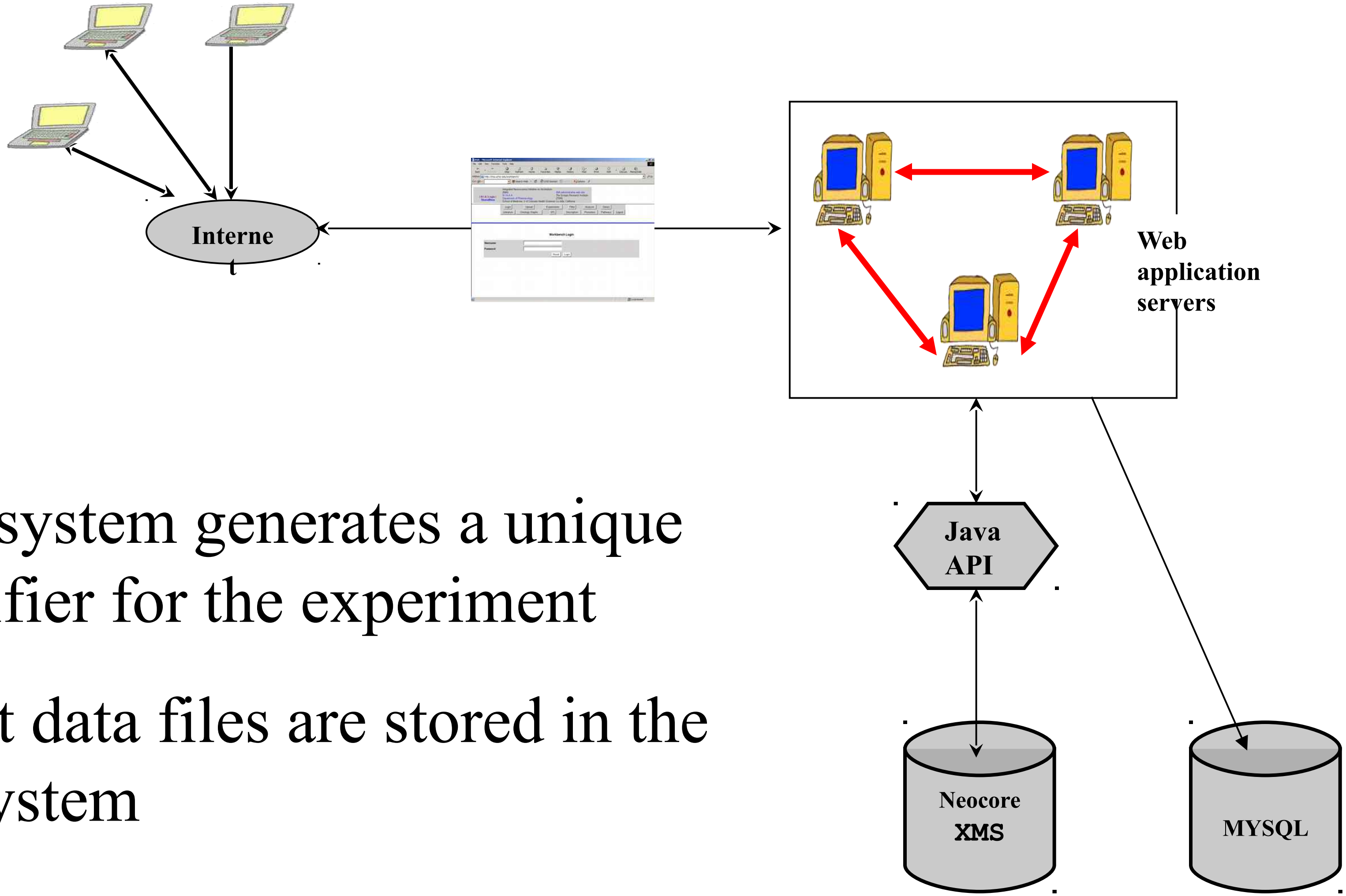
–MIAME annotation information corresponding to the experiment (meta data)

Design & Implementation

Sample meta data

```
<Chip_Description>C57 11012 line WT5</Chip_Description>
<Date_of_experiment>5/9/2001</Date_of_experiment>
<Phone>303-315-1064</Phone>
<E-mail>Sanjiv.Bhave@uchsc.edu</E-mail>
<Species>Mus musculus</Species>
<Gender>Male1</Gender>
<Age>10 weeks</Age>
<Genetic_background>WT littermate</Genetic_background>
<Organ>Brain</Organ>
<Manufacturer>Affymetrix</Manufacturer>
<Gene_Array_ID>MG U74Av2.0</Gene_Array_ID>
<RNA_extraction_method>Trizol followed by Rneasy mini kit</RNA_extraction_method>
<Amont_of_RNA_used_in_microgram>10 micorgram </Amont_of_RNA_used_in_microgram>
<Total_or_mRNA>Total RNA</Total_or_mRNA>
<cDNA_or_cRNA_preparation_methods_used>Affymetrix protocol</cDNA_or_cRNA_preparation_methods_used>
<Phenotype_data>Anxiety</Phenotype_data>
<Electrophysiological_data>EEG10</Electrophysiological_data>
<Biochemical_data>Phosphorylation</Biochemical_data>
<Physiological_data>Hormonal Response WT5</Physiological_data>
```

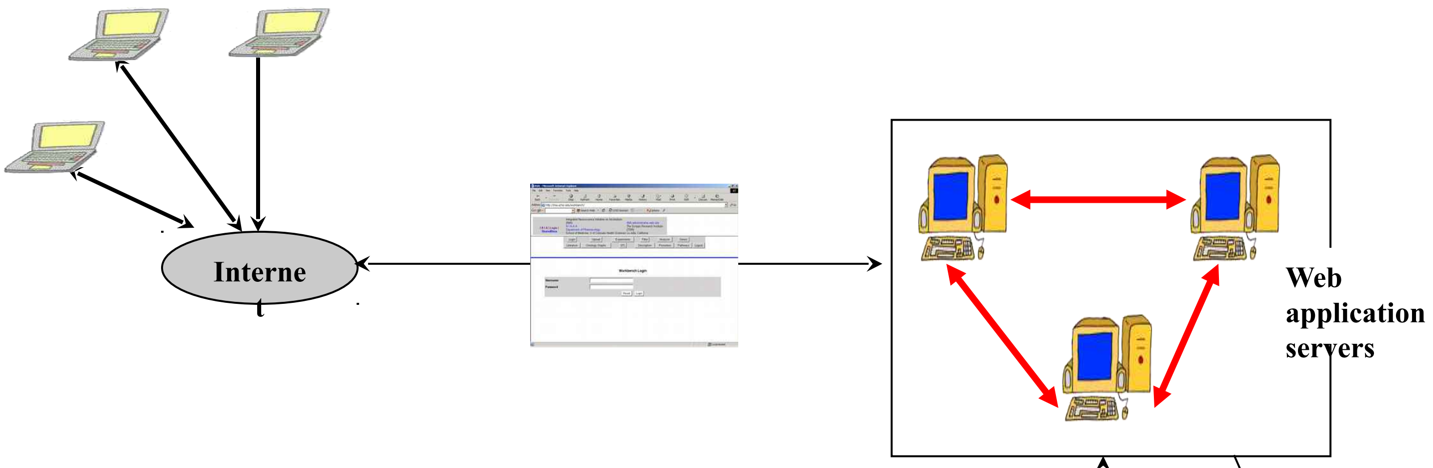
Design & Implementation

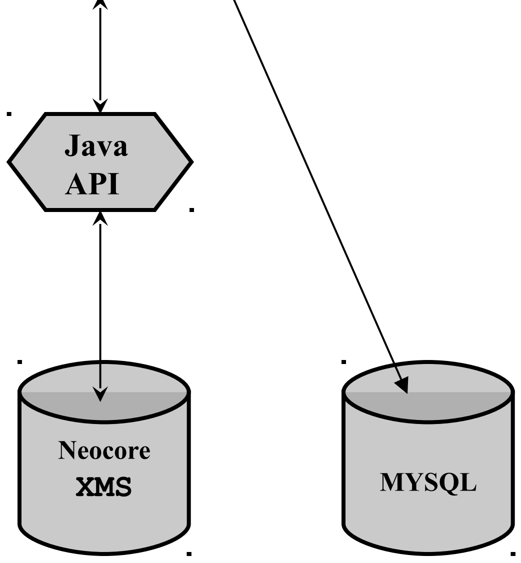


- The system generates a unique identifier for the experiment

- Input data files are stored in the file system

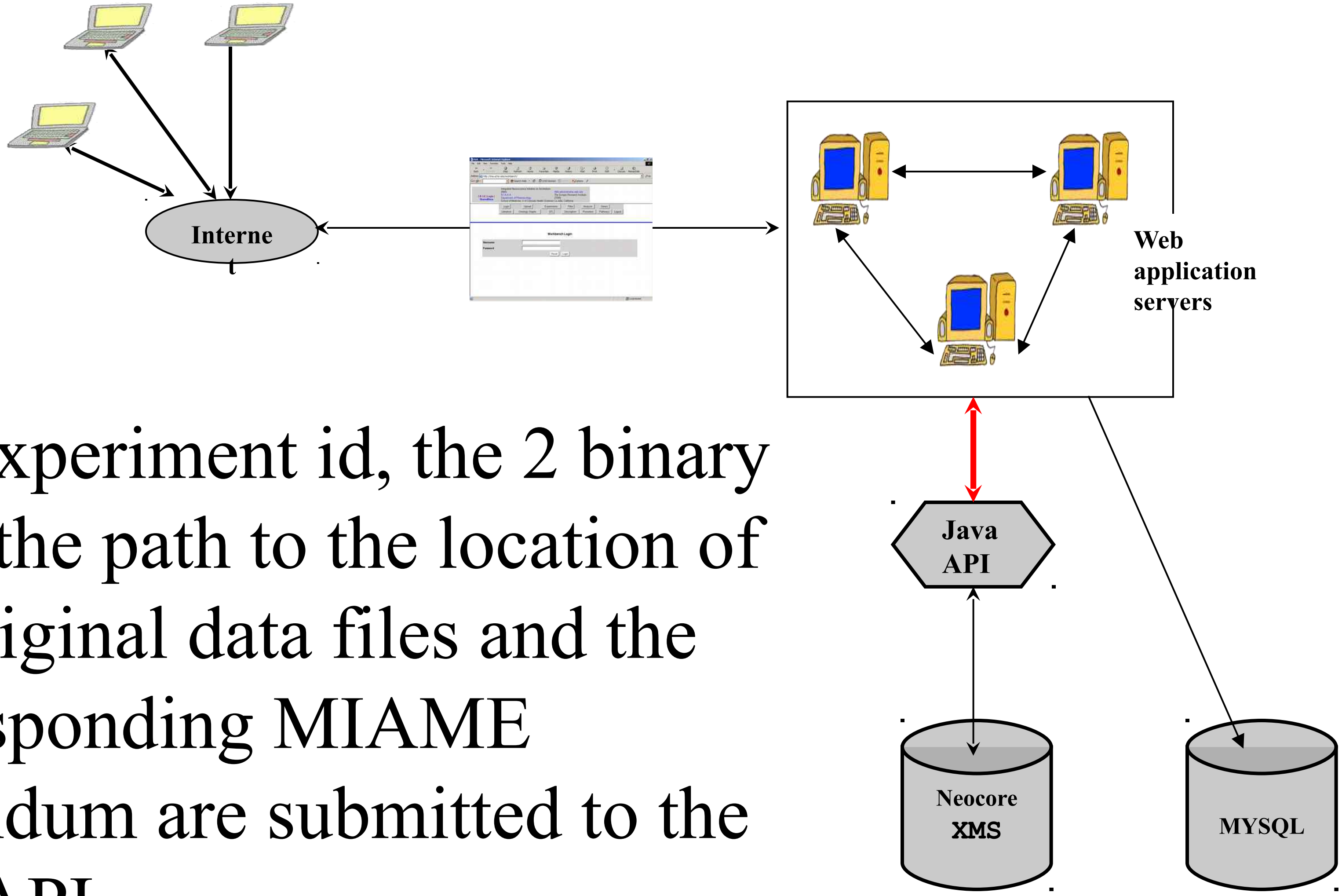
Design & Implementation





- Input data run is through a filter and statistical analyzer
- 2 binary files are produced

Design & Implementation



The experiment id, the 2 binary files, the path to the location of the original data files and the corresponding MIAME addendum are submitted to the Java API.

Design & Implementation
Data Encoding

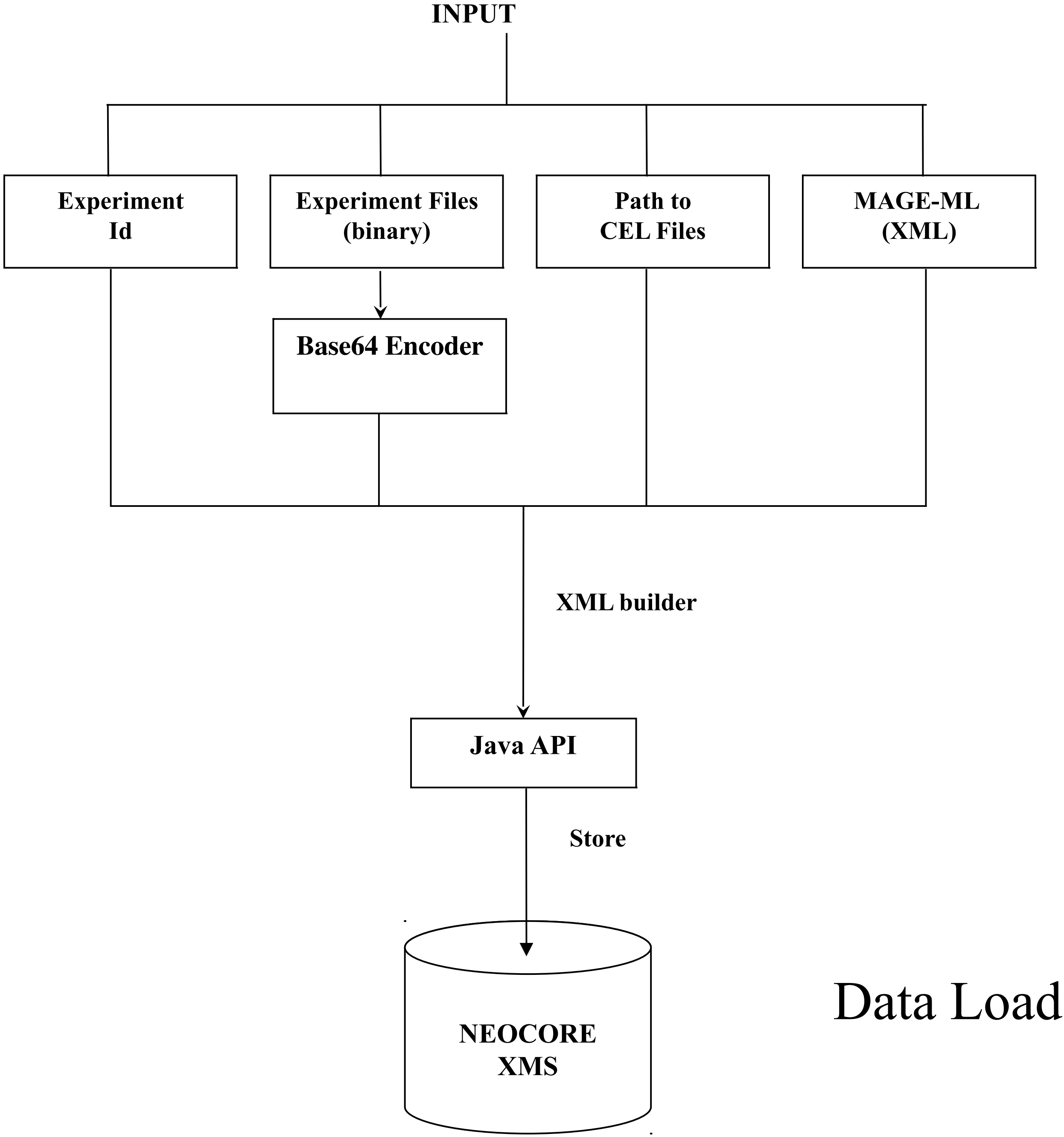
- UTF-8 (Unicode Transformation Format – 8) is the default encoding for XML
- Hence binary data to be stored in an XML document needs to be in a UTF-8 compliant character set
- Used a base-64 encoder to encode input data file

Design & Implementation Data Encoding

- Encoded data is larger in size
  - Base64 encoding divides three bytes of the binary data into four bytes of ASCII text
  - the size of the encoded data is about a third larger than the original
- Data may compressed before storage
  - GZIP compression – utility provided by Java

Design & Implementation Data Loading

- XML document built by enclosing following in appropriate tags
  - Experiment Information
  - Encoded and zipped data
  - Meta data
- Document is loaded into NeoCore
- Document id is returned by NeoCore if transaction was successful



INPUT

Experiment Id

Experiment Files (binary)

Path to CEL Files

MAGE-ML (XML)

Base64 Encoder

XML builder

Java API

Store

NEOCORE XMS

Data Loading

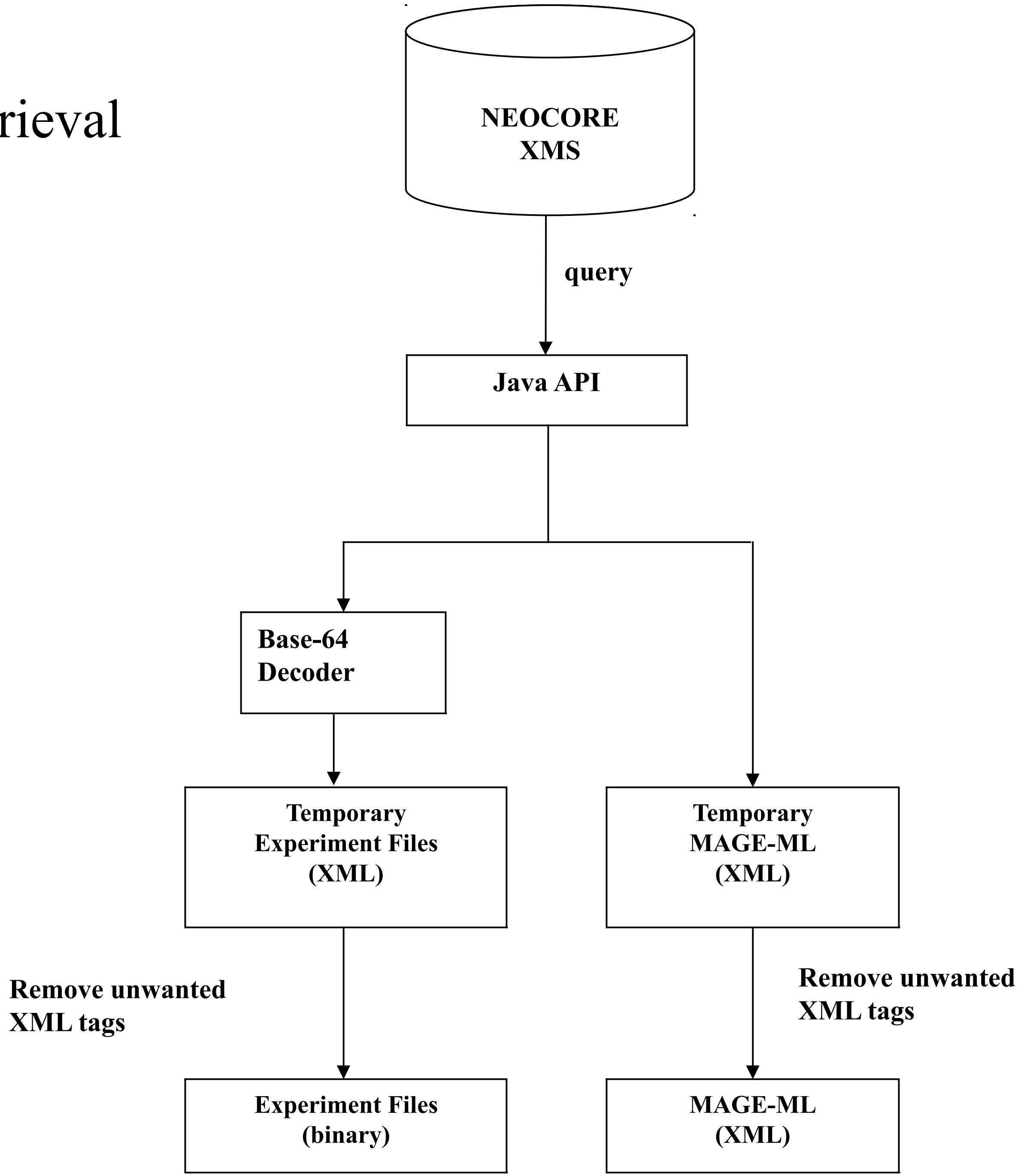
Design & Implementation
Data Retrieval

- The submitter and other users are allowed to retrieve experiment information for further analysis or verification.
- The experiment and session identifier are passed to the XMS via the Java API.
- Before being retrieved, the binary files are decoded from base-64 format.
- The binary files and MIAME annotation information corresponding to the experiment are returned as XML documents.

Design & Implementation Data Retrieval

- The XML tags are removed and the resulting files are returned to the user.

Data Retrieval



NEOCORE XMS

query

Java API

Base-64 Decoder

Temporary Experiment Files (XML)

Temporary MAGE-ML (XML)

Remove unwanted XML tags

Remove unwanted XML tags

Experiment Files (binary)

MAGE-ML (XML)

References

[1] "ArrayExpress Microarray Standards." April 20, 2004. European Bioinformatics Institute. April 24, 2004 <http://www.ebi.ac.uk/arrayexpress/Standards/ index.html>

[2] Brazma, Alvis, et al. "Minimum information about a microarray experiment (MIAME) – toward standards for microarray data." Nature Genetics. Dec. 2001 Vol. 29: 365-71.

[3] DeFrancesco, Laura. "MIAME begets MAGE." The Scientist. September 17, 2002. BioMed Central. April 24, 2004. <http://www.biomedcentral.com/news/ 20020917/02/>

[4] Direen, Harry. "Re: NeoCore Indexes." E-mail to Chowrira, Sajni. 10 May 2004.

[5] Direen, Harry. XML for Information Processing. Xpriori, LLC. ____ ____ ________________________

[6] "Information Management Solutions for eXtreme Productivity". 2004. Xpriori, LLC. March 10, 2004. < http://www.xpriori.com/>

[7] Kelley, Kevin. Base64.java. Vers. 1.3. 10 KB. September 22, 2000. Starlight Software Co. March 30, 2004 <http://kevinkelley.mystarband.net/java/ Base64.java>

References

[8] Knudsen, Steen. A Biologists Guide to Analysis of DNA Microarray Data. ____New York: John Wiley & Sons, Inc., 2002.

[9] “MGED Mission Statement.” September 26, 2003. European Bioinformatics Institute. April 24, 2004 <http://www.mged.org/Mission/index.html>

[10] Morin, Randy. "How to Base64." Online article. 29 April 2004. <http://www.kbcafe.com/articles/HowTo.Base64.pdf>

[11] Ness , Scott A. “KUGR Microarrays and Genomic Facility.” Keck-UMN Genomics Resource. 15 May 2004. <http://hsc.unm.edu/som/micro/Genomics/ basics.html>

[12] Phang, Tzulip, et al. “Interactive workbench for high-throughput molecular biological data exploration”. Department of Pharmacology, Univerity of Colorado Health Sciences Center. 2004.

[13] Shi, Leming. “DNA Microarray (Genome Chip).” January 7, 2002. www.Gene-Chips.com. April 22, 2004 < http://www.gene-chips.com/ >

[14] Spellman, Paul T., et al. “Design and inplementation of microarray gene expression markup language (MAGE-ML).” Genome Biology. 2002 Vol.3 No.9 : research0046.1-0046.9.

References

[15] NeoCore XMS: System Administration Guide. Xpriori, LLC. ___

[16] NeoCore XMS: System Programming Guide: Xpriori, LLC. __

[17] Tschabitscher, Heinz. "How Base64 Encoding Works." About.Com. 28 April 2004. <http://email.about.com/library/misc/bl_base64_enc_table.htm>

"Unified Information Management Using NeoCore® XMS." Xpriori, LLC. White paper. March 12, 2004 < http://www.xpriori.com/White_Paper-Unified_Information_Management.pdf>.