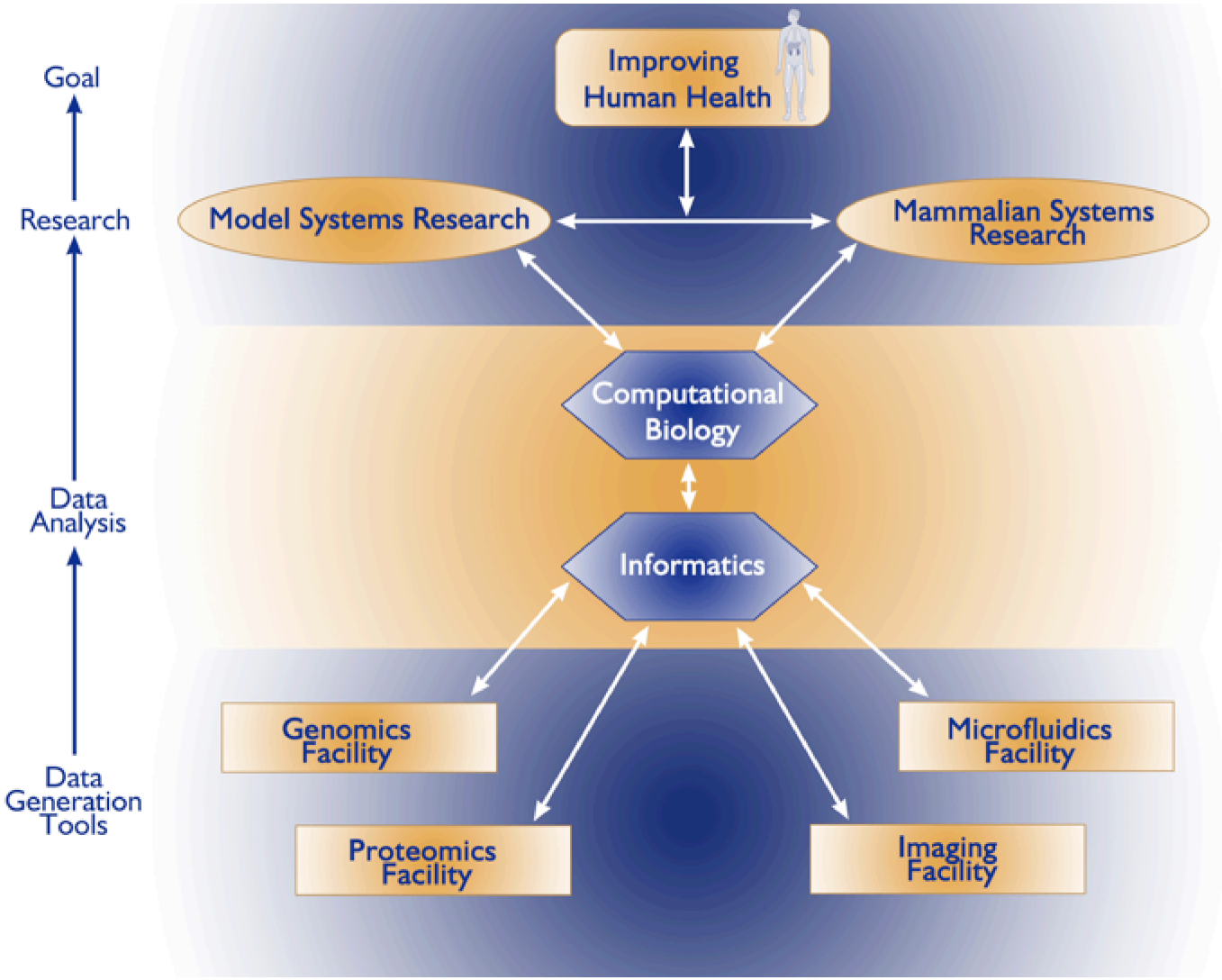
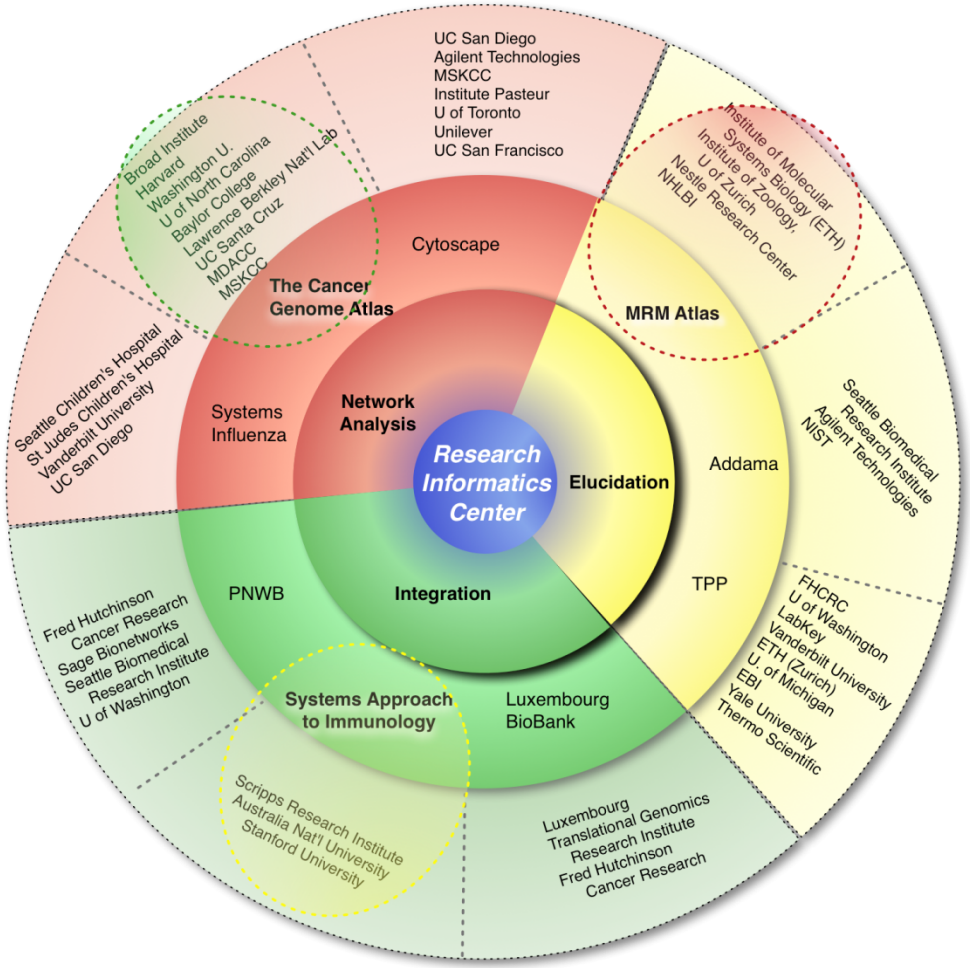
Adaptable Data Management

John Boyle
The Institute for Systems Biology

The Institute for Systems Biology



Collaborative Projects



Requirements

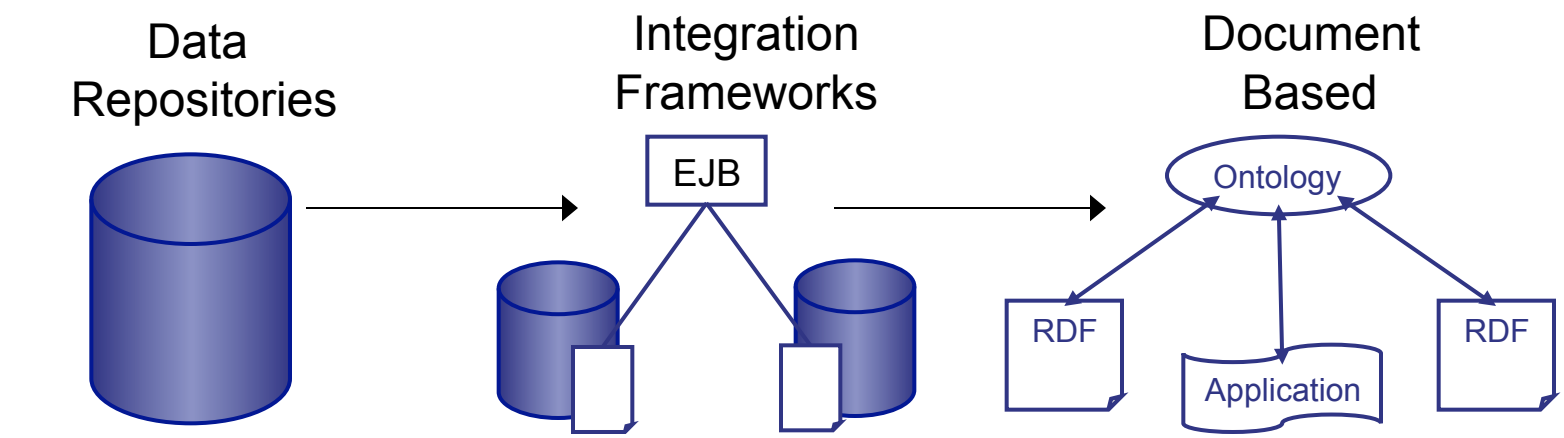
- High volume of heterogeneous data
- High volume of heterogeneous users
- Continual introduction of new data sources and technologies

- Easy to access and understand
- Interoperable and non-intrusive integration

Scientific Integration Systems

System	Technology	Generation
SRS (LionBiosciences)	External indexing of flat files	Database
DiscoveryCenter (Netgenics)	CORBA based components	Integration Framework
Alliance (Synomics)	J2EE distributed system	Integration Framework
MetaLayer (Tripos)	XML message passing	Integration Framework
DiscoveryLink (IBM)	Federated database solution	Integration Framework
GKP (Incyte)	EJB based object integration	Integration Framework
LSP (Oracle)	Embedded web services	Stateless Web Services
myGRID (EPSRC)	Ontology driven services	Stateless Web Services
caBIG (NCI)	MDA based architecture	Stateless Web Services
BioMoby (NSF)	Registry and semantic web based	Document Based
CancerGRID (MRC)	Resources using web services	Document Based
caGRID (NCI)	Web service/registry solution	Document Based
Amalga (Microsoft)	Data warehouse for content	Document Based

Integration System Evolution



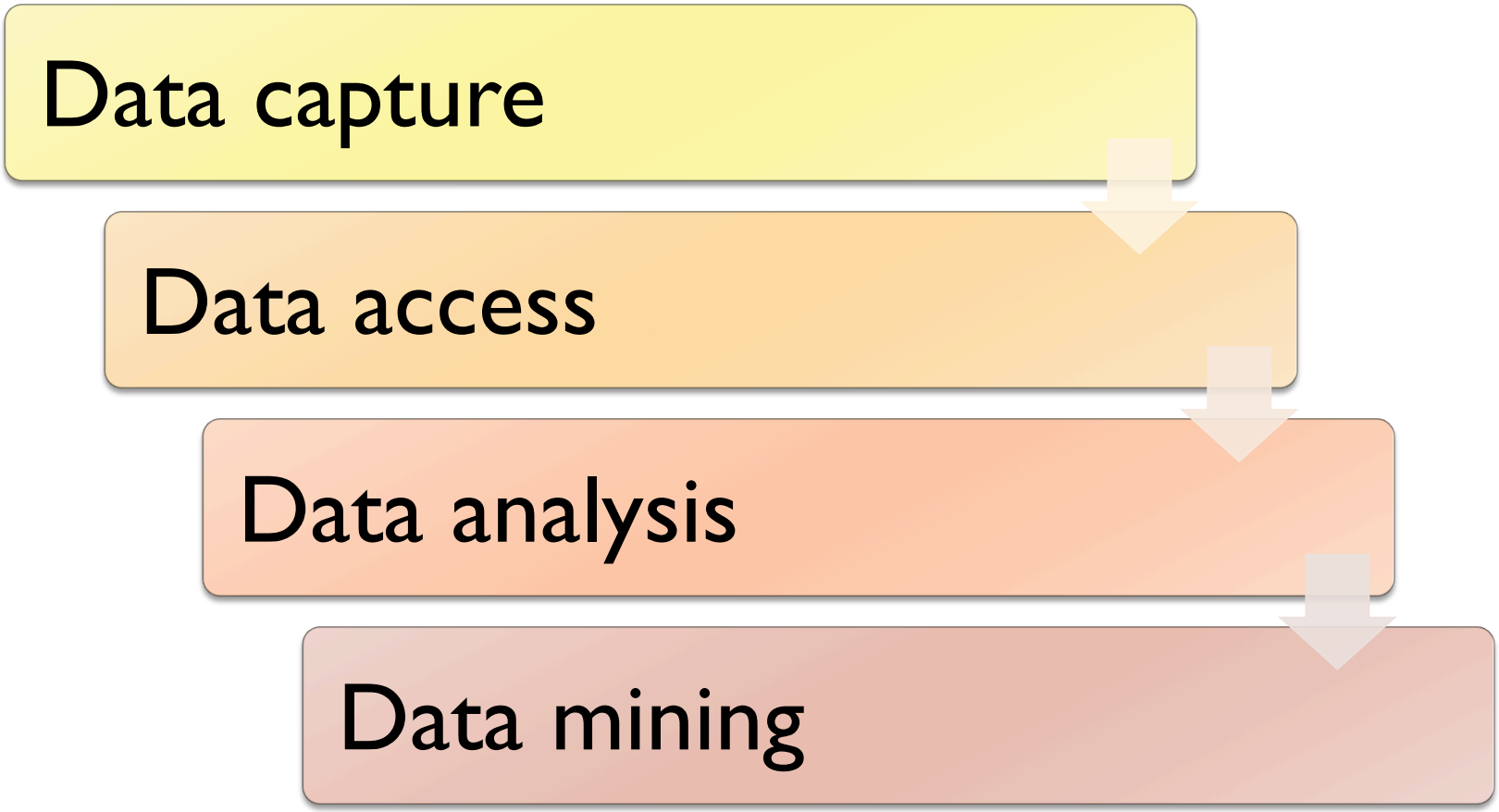
Methodology:	Components →	Frameworks →	Aspects
Technology:	Brokers →	P(M)BV →	SOAP/REST
Ideology:	RUP →	Agile →	MDD

Requirements: Semantics → Integration → Ease of Use

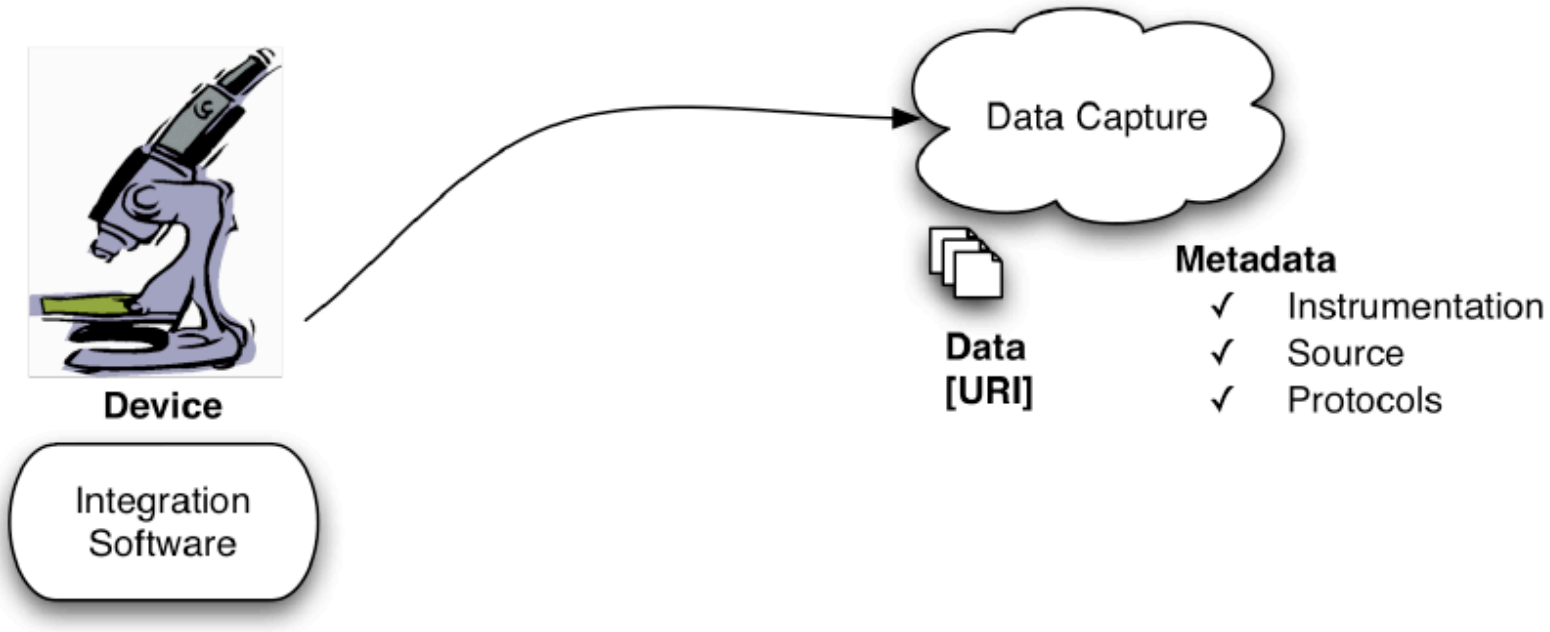
Design Rationale

- Support ad hoc development
- Easy to use and integrate
- Rapid development and deployment
- Adaptable and maintainable
- Supporting a dynamically changing environment

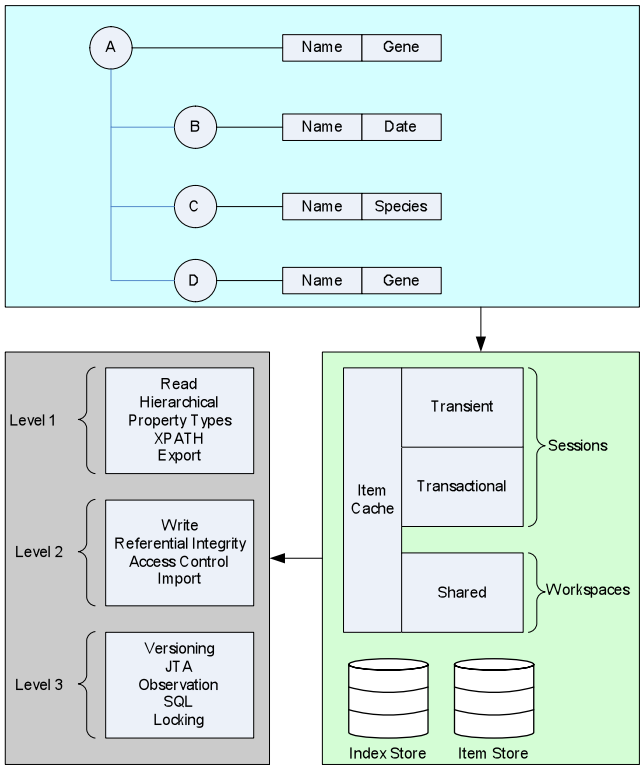
The Basics

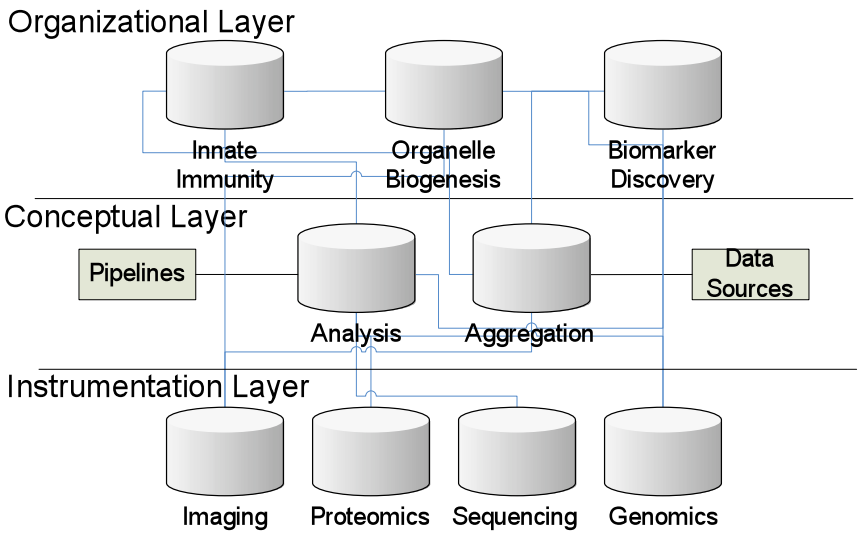


Data Capture

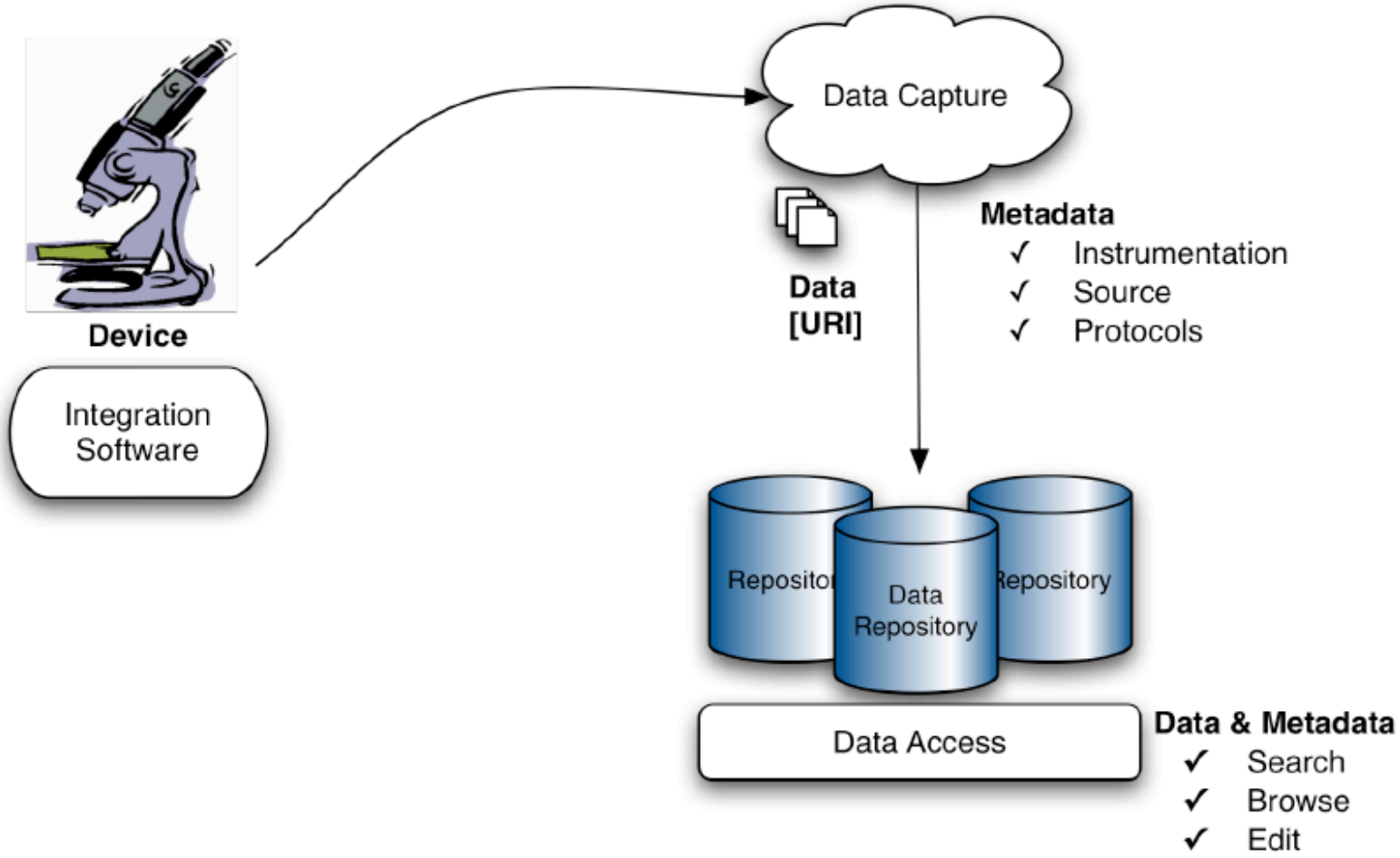


Integration ‘as is’ and ‘step-wise’

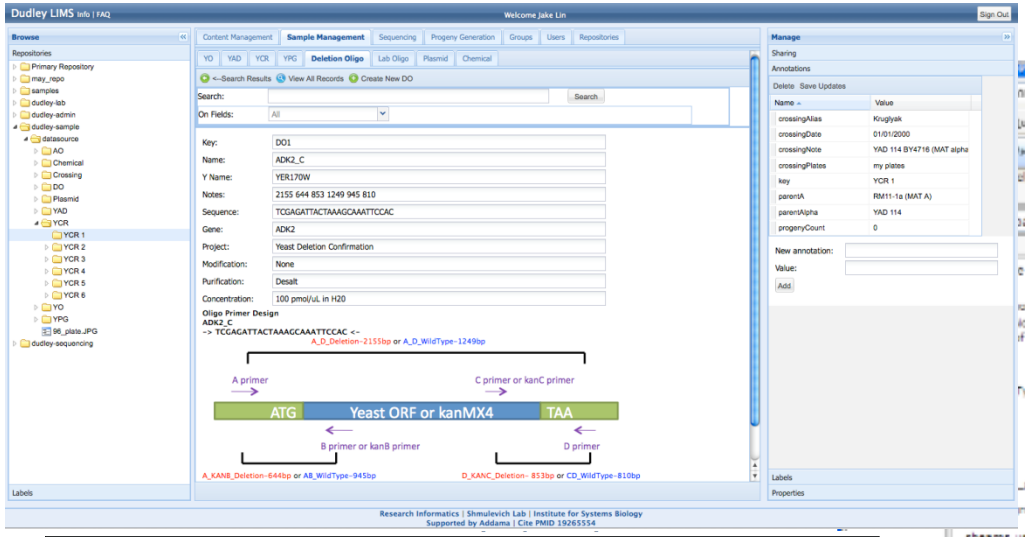


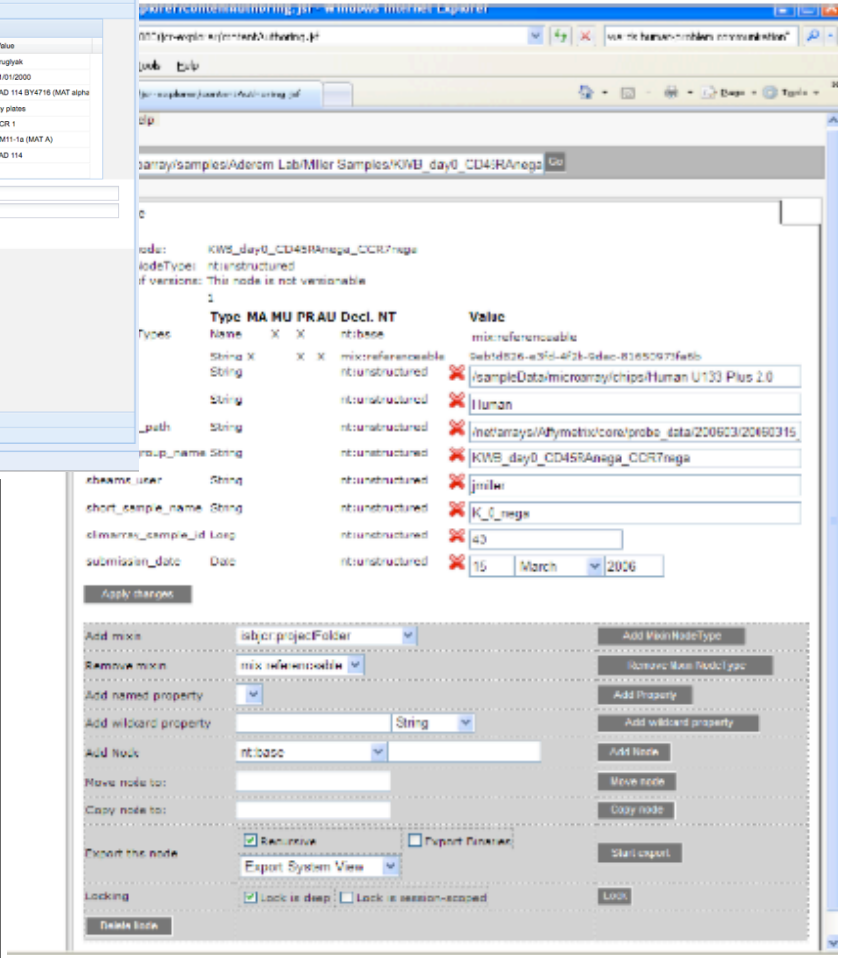


Data Access


Device
Integration Software
Data Capture
Data
[URI]
Metadata
✓ Instrumentation
✓ Source
✓ Protocols
Repository
Data Repository
Repository
Data Access
Data & Metadata
✓ Search
✓ Browse
✓ Edit

Easy to access and understand





This site will provide access to data processing pipelines, results from pipeline runs, and data visualization and integration under the the NIAID Contract "Systems Approach to Immunity and Inflammation" (HHSN272200700038C). The site sits behind an ISB firewall, and includes data not released on the corresponding publicly-accessible site: www.systemsimmunology.org

Download Processed Data

	Pipeline	Cell Type	Description	Chip	Date Run	Download Files
1	**Exon: *Gene-Level***	All	All Stims	Affymetrix Mouse Exon 1.0 ST	Sat Aug 22 03:20:33 PDT 2009	Data_Matrix.tsv expression_set.tsv probe_intensity_repCombined.MADs.tsv Data_Matrix.xls
2	**Exon: *Gene-Level***	BMDM	All Stims	Affymetrix Mouse Exon 1.0 ST	Sun Aug 23 02:50:28 PDT 2009	Data_Matrix.tsv expression_set.tsv probe_intensity_repCombined.MADs.tsv Data_Matrix.xls
3	**Exon: *Gene-Level***	Dendritic Cells	All Stims	Affymetrix Mouse Exon 1.0 ST	Fri Aug 07 12:24:45 PDT 2009	Data_Matrix.tsv expression_set.tsv probe_intensity_repCombined.MADs.tsv Data_Matrix.xls

Process Arrays

Run a normalization pipeline using only the experimental conditions of your choice.

- Affymetrix Mouse 3' Expression Array Pipeline
- Affymetrix Mouse Exon GeneLevel Expression Array Pipeline

View Additional Pipeline Results

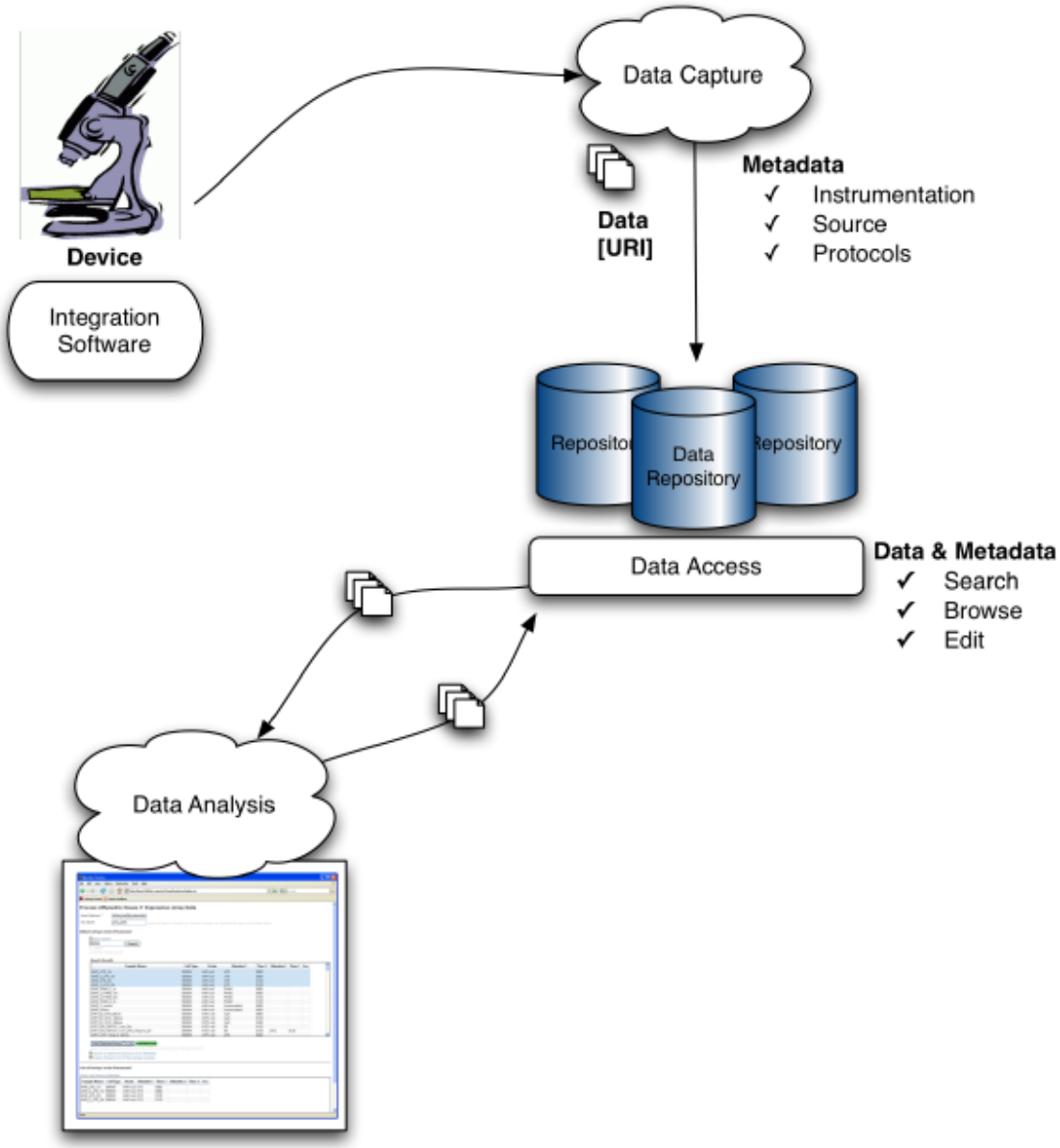
- Access all recent pipeline runs

Useful Tools

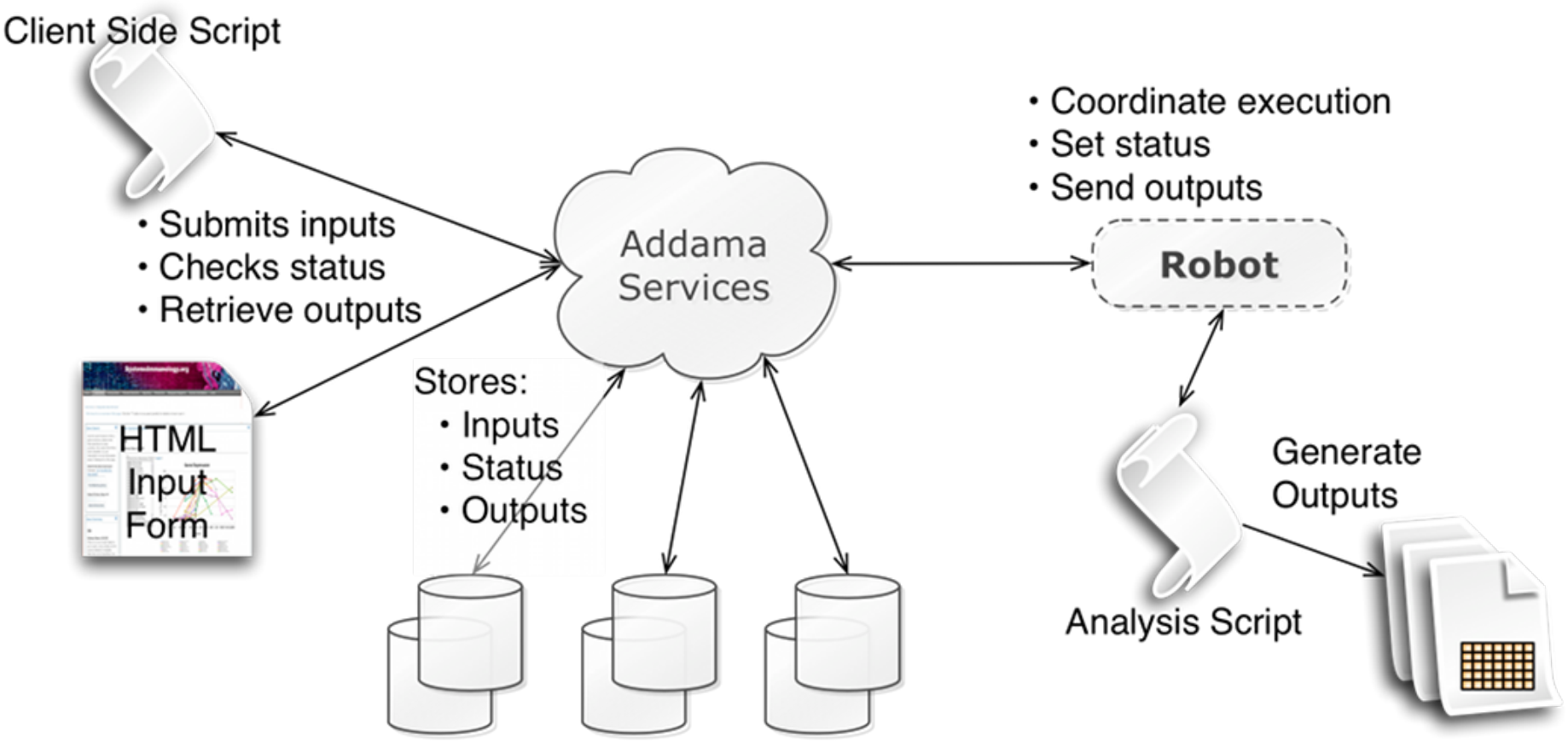
- Mouse Analysis Resources

If you have questions about this page, please email immunoinformatics@systemsbiology.org

Data Analysis



Support Ad-hoc Analysis



Addama Web Automation Research Informatics Systems Biology

JCR Browse Search

Analyzing ChIP-Seq Peaks with R and MACS

ChIPSeq Data File:

File Format: BED

Sample Name from JCR: Select Sample Name

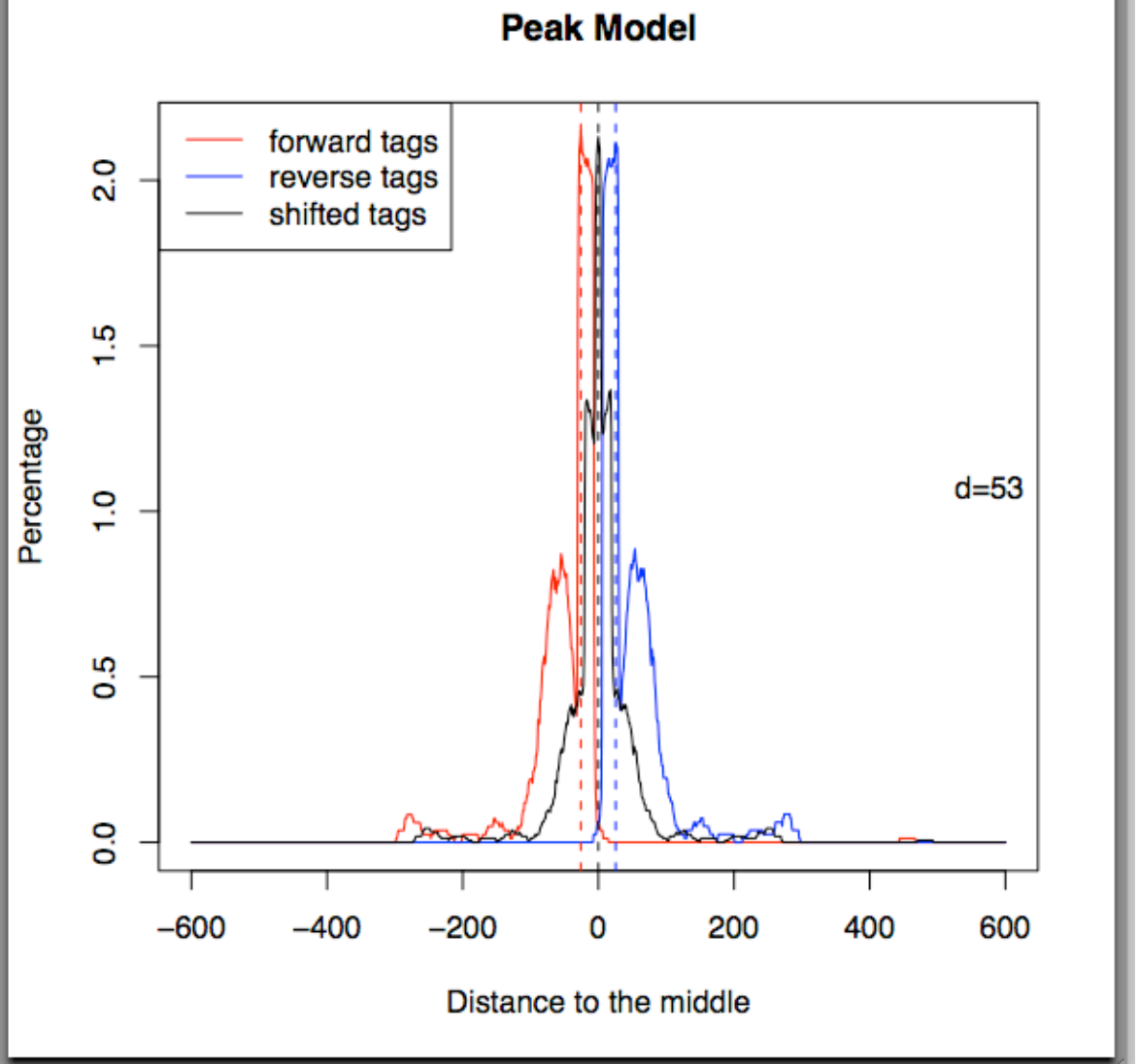
MACS Optional Parameters README:

ChIP-Seq Automation Results

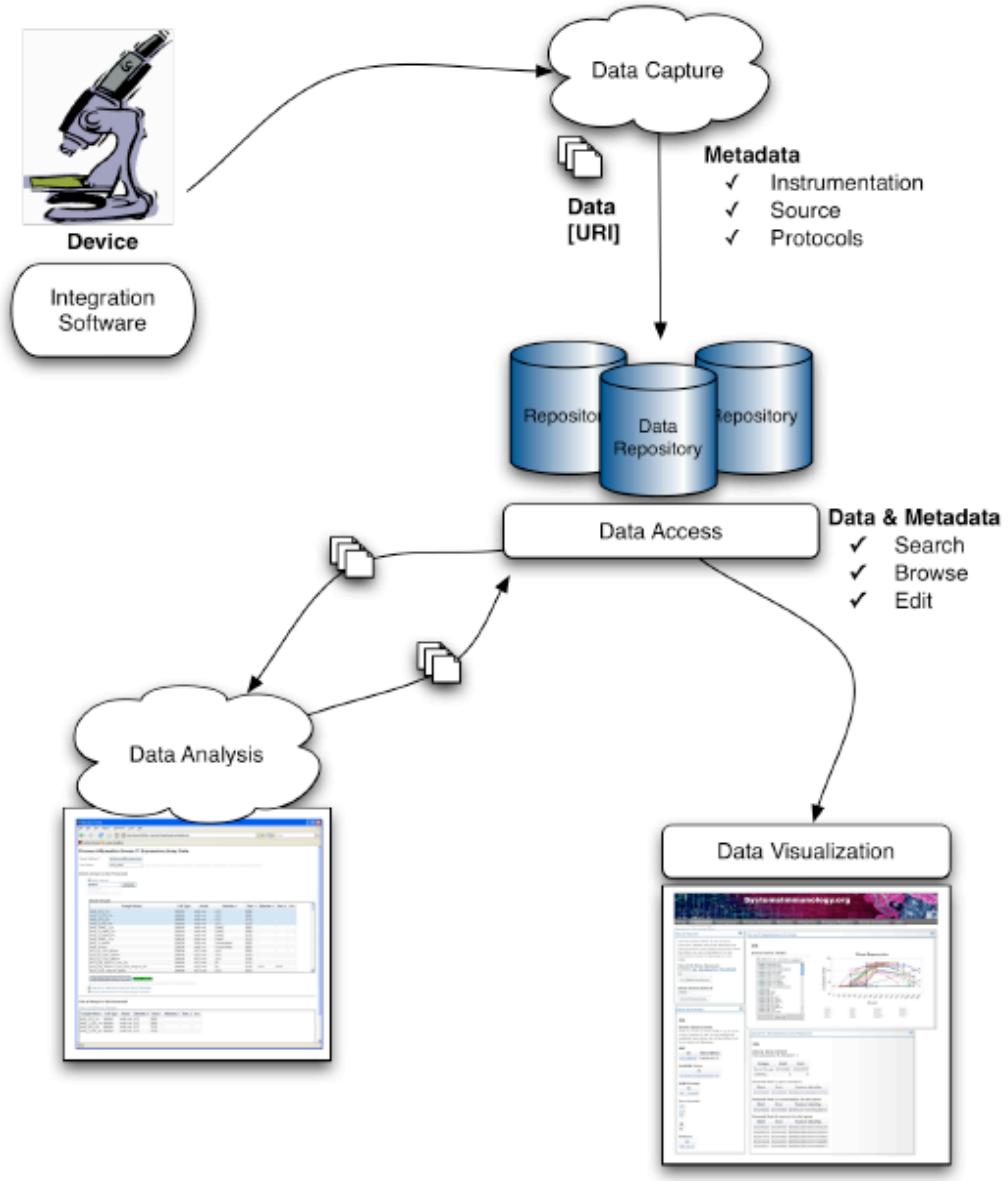
Status: Completed at: 06/17/2009 09:35:28
Execution time (mn:ss) 0:9
MACS outputs meta posted to JCR path:/addama-rest/s
View Peak Distributions Model PDF
Peak Model R Input
Peak Distributions .BED for Genome Browsers
Peak Distributions .XLS
+ Download JSON
Inputs:
ChIP-Seq File: s_1_eland_multi_200k.txt
Selected File Format: ELANDMULTI
Sample Name: sample-1007 Desc-> yeast-unknown-x2
MACS Optional Parameters: --mfold=16

macs_model-3.pdf (1 page)

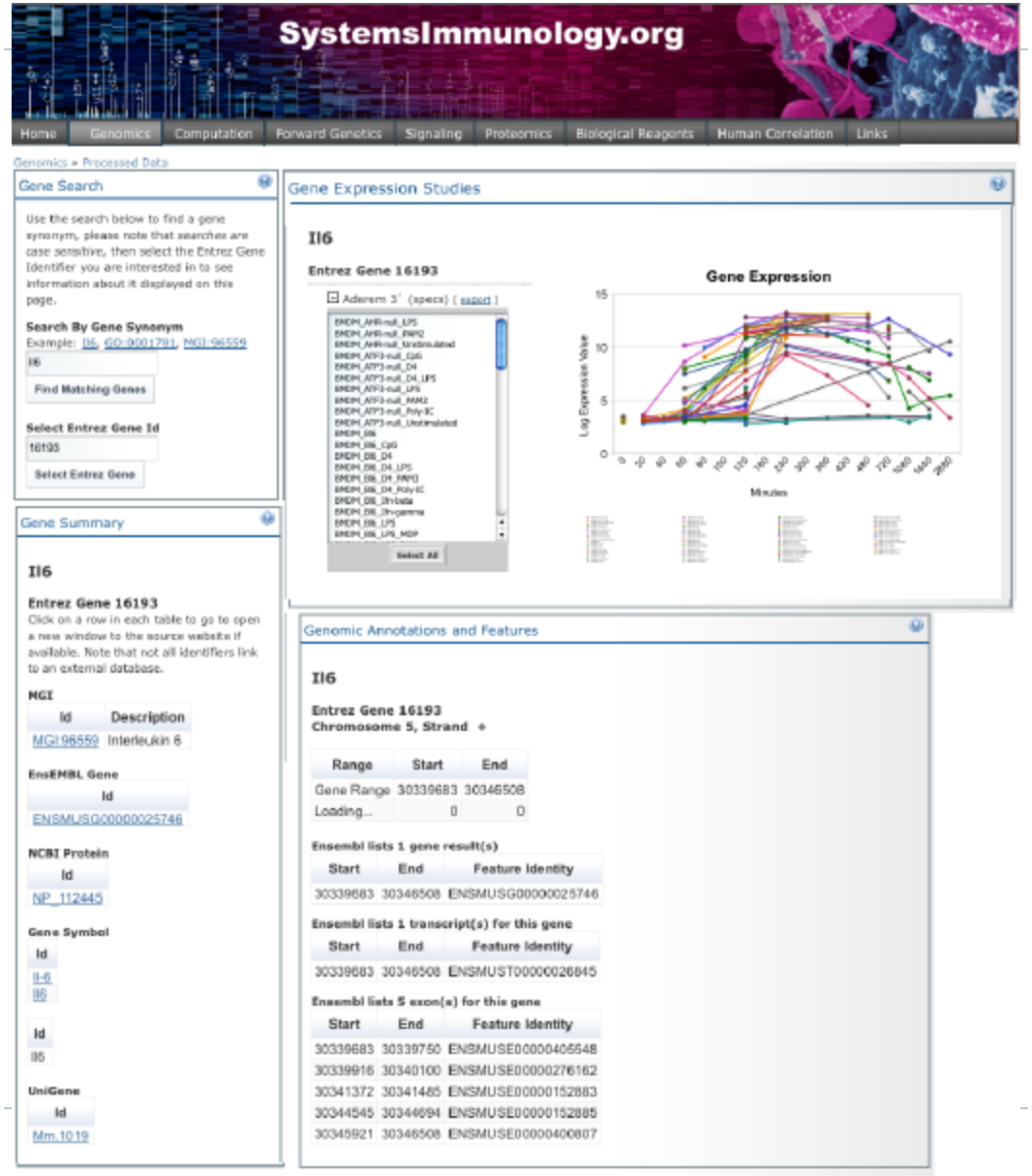
Previous Next Zoom Move Text Select Sidebar Search



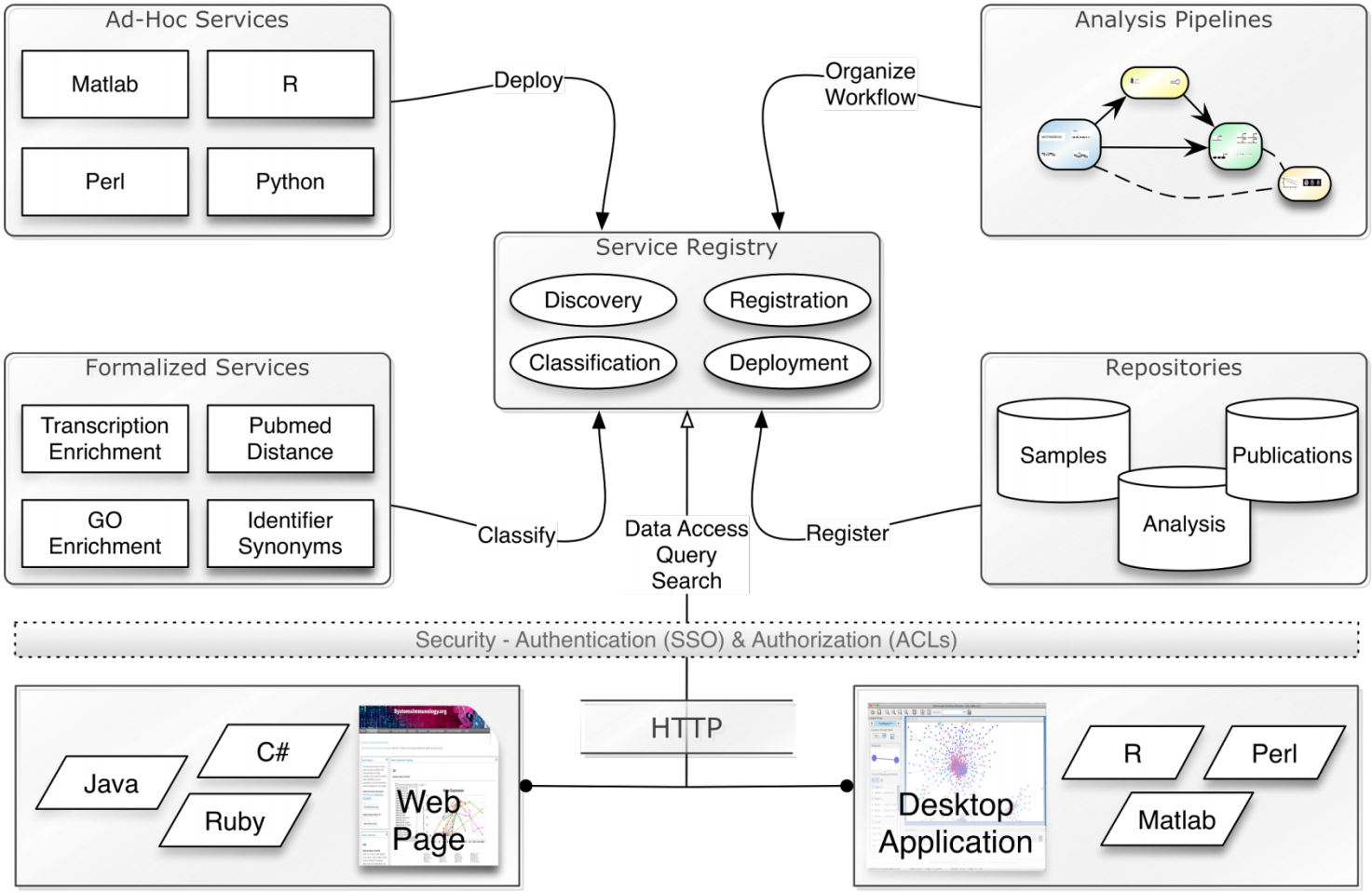
Data Visualization



Data Visualization



Adaptable Data Management



Requirements for Research

- Need to be adaptable
- Need to allow for flexible deployment
- Need to support step wise integration
- Need to support non-intrusive integration
- Needs to be easy to use and understand
- Needs to interoperable, robust, loosely coupled, standardized and maintainable

Use Suitable Technologies

- Access
  - REST/JSON
  - Google Data Sources
- Enterprise Systems
  - Message Queues
  - SOA/ROA
- Deployment
  - GAE/GBT
  - EC2/S3
- Performance
  - Hadoop
  - GPUs
- SchemaFree:
  - JCR
  - CouchDB
- Programming:
  - IoC/AOP
  - DSL
- Web Development
  - Rails
  - GWT
- Shared Spaces
  - Apache Cloud
  - WAVE

Conclusions

- Research is dynamic and evolving
- Innovation is key
- Ad hoc, unplanned, rapid development is the norm
- Adaptable systems are needed
- Need new approaches to information management, retrieval and visualisation.

References

Related References

- J. Boyle, H. Rovira, C. Cavnor, D. Burdick, S. Killcoyne, I. Shmulevich, "Adaptable Data Management for Systems Biology Investigations", BMC Bioinformatics Vol. 10, No. 79, 2009.
- J. Boyle, C. Cavnor, S. Killcoyne, I. Shmulevich, "Systems biology driven software design for the research enterprise", BMC Bioinformatics, Vol. 9, No. 295, 2008.

Related Web Sites:

- informatics.systemsbiology.net
- www.addama.org