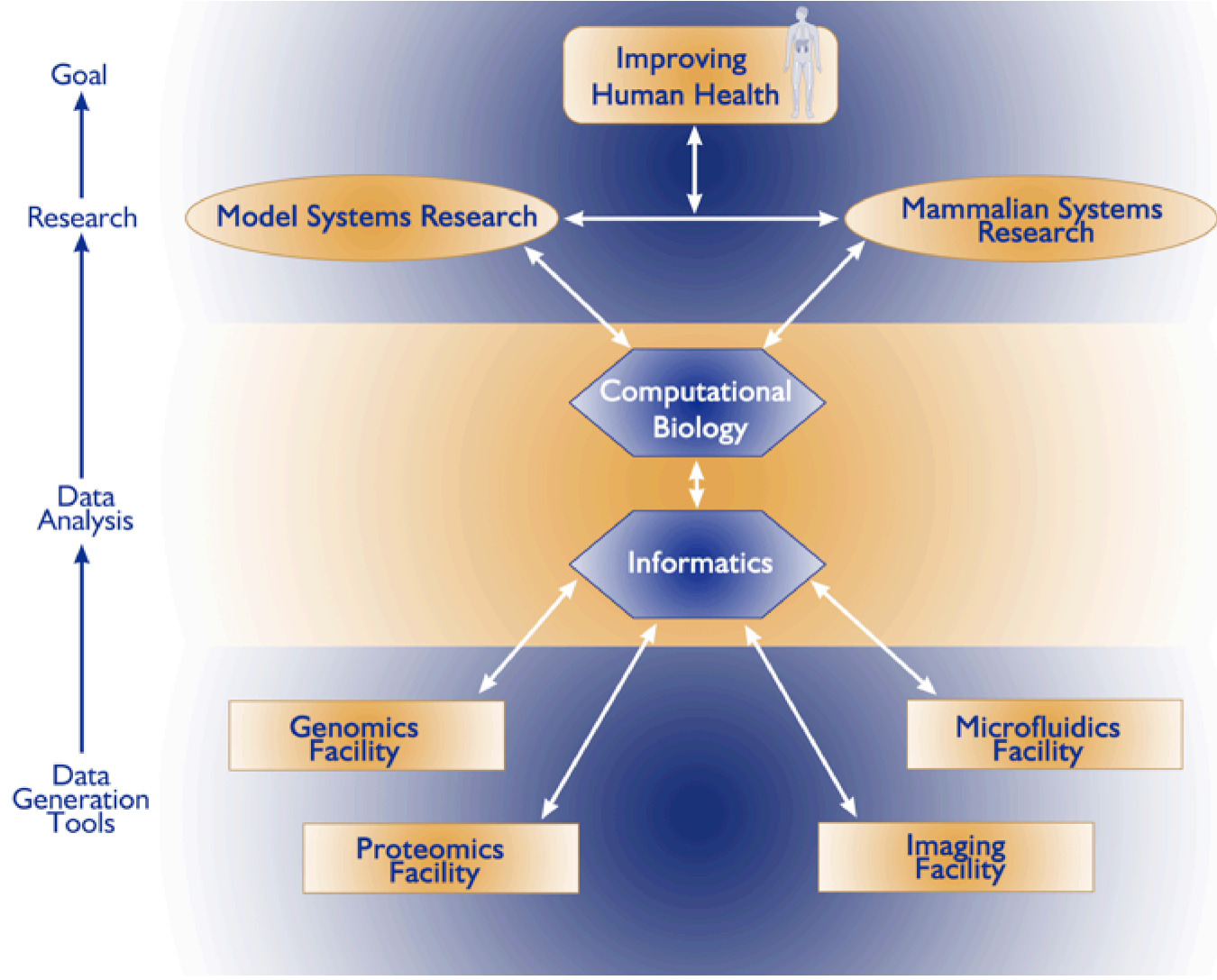


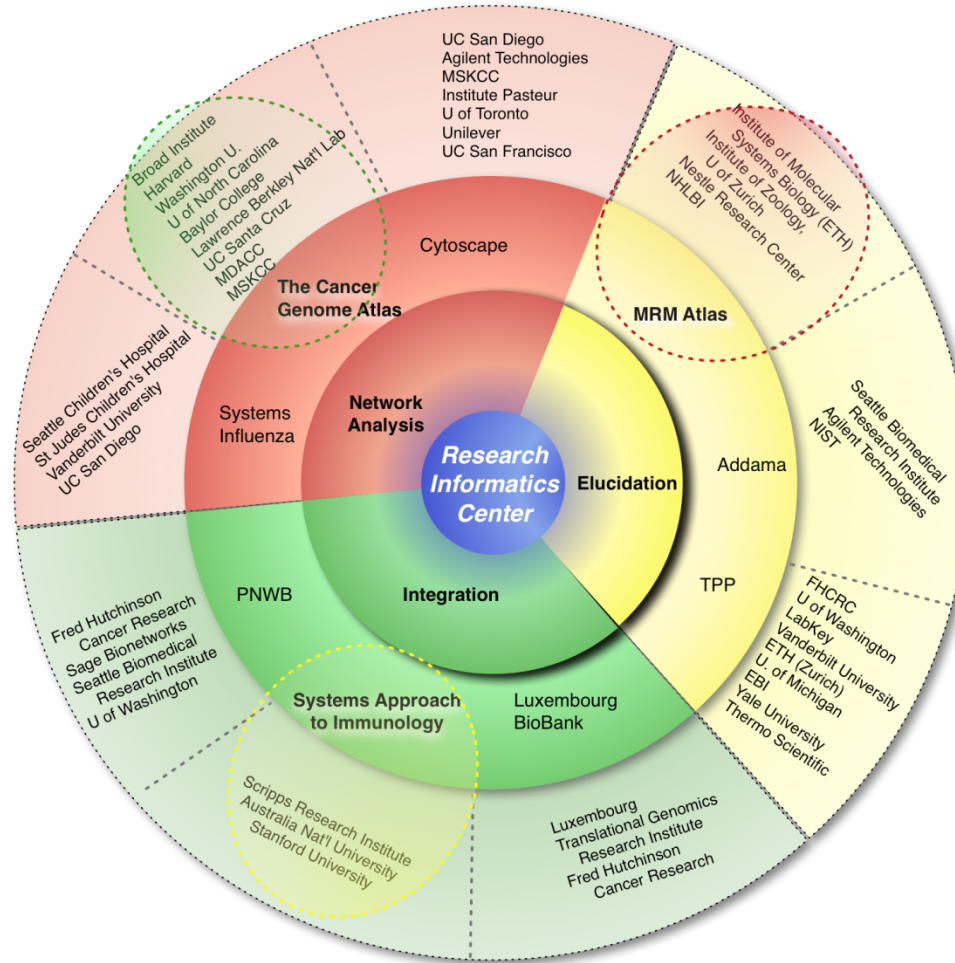
Adaptable Data Management

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

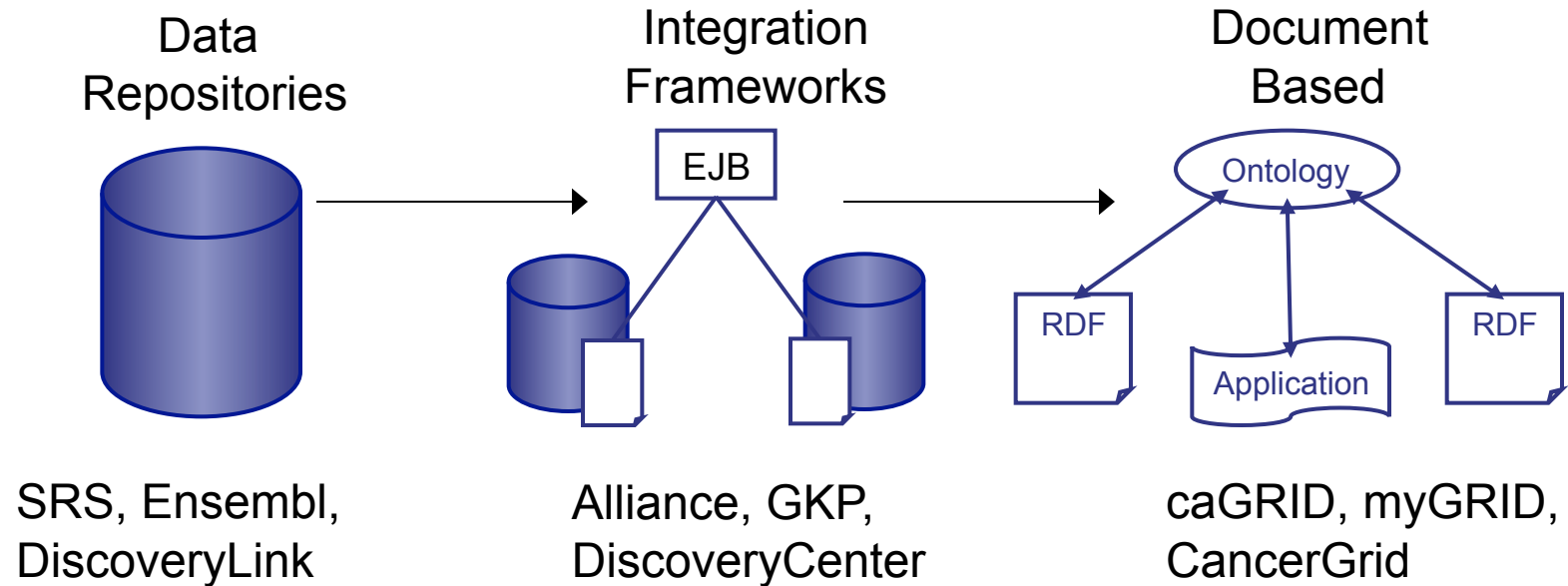


Scientific Integration Systems

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

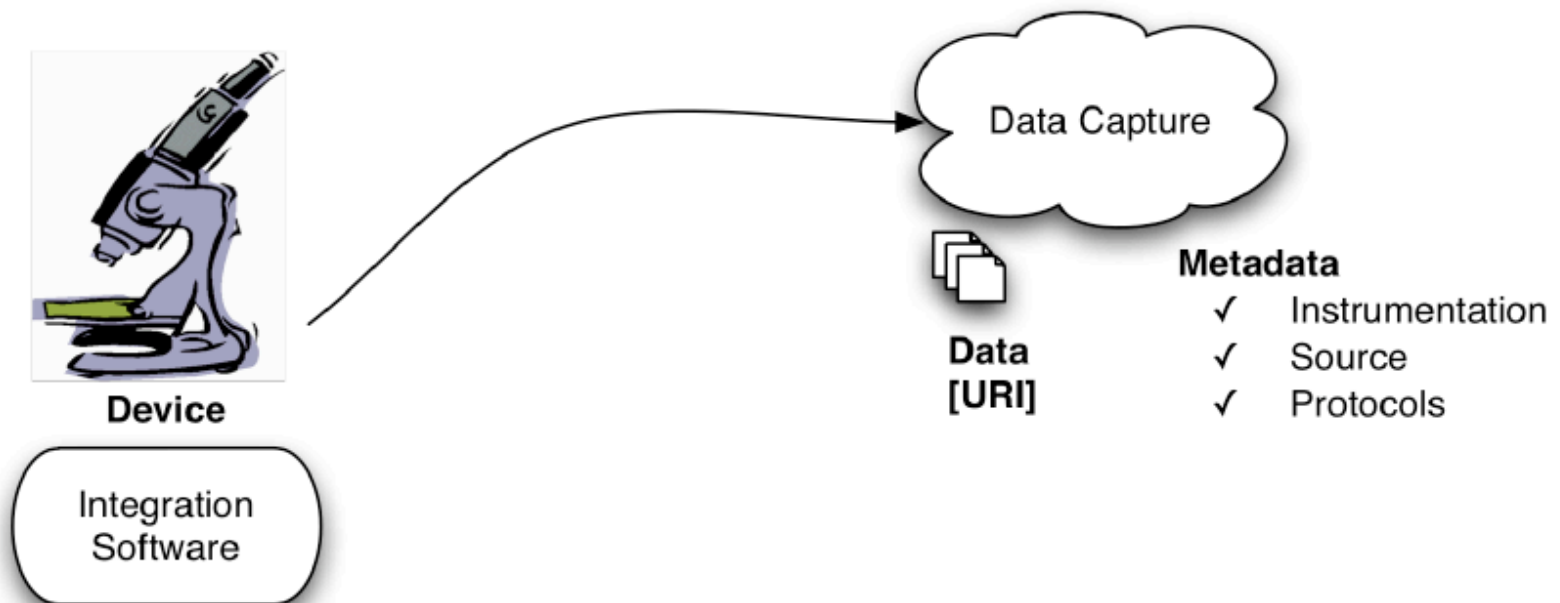
Data access

Data analysis

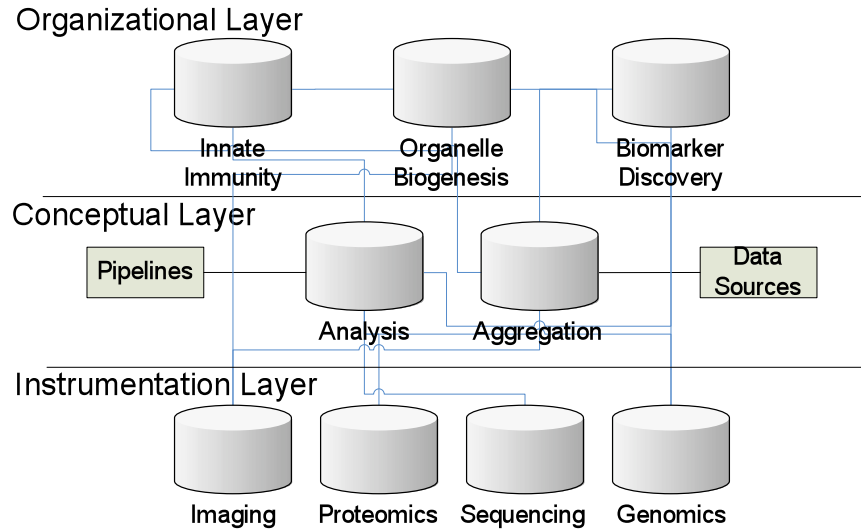
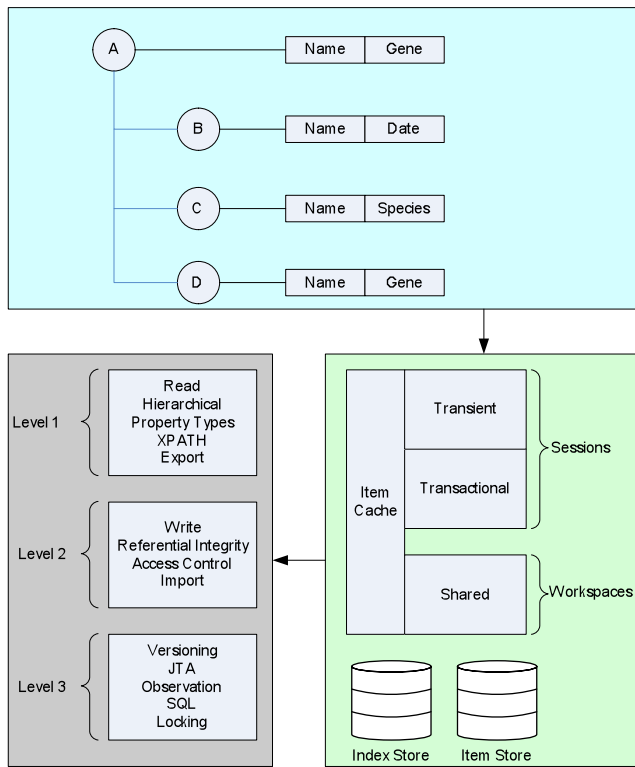
Data mining



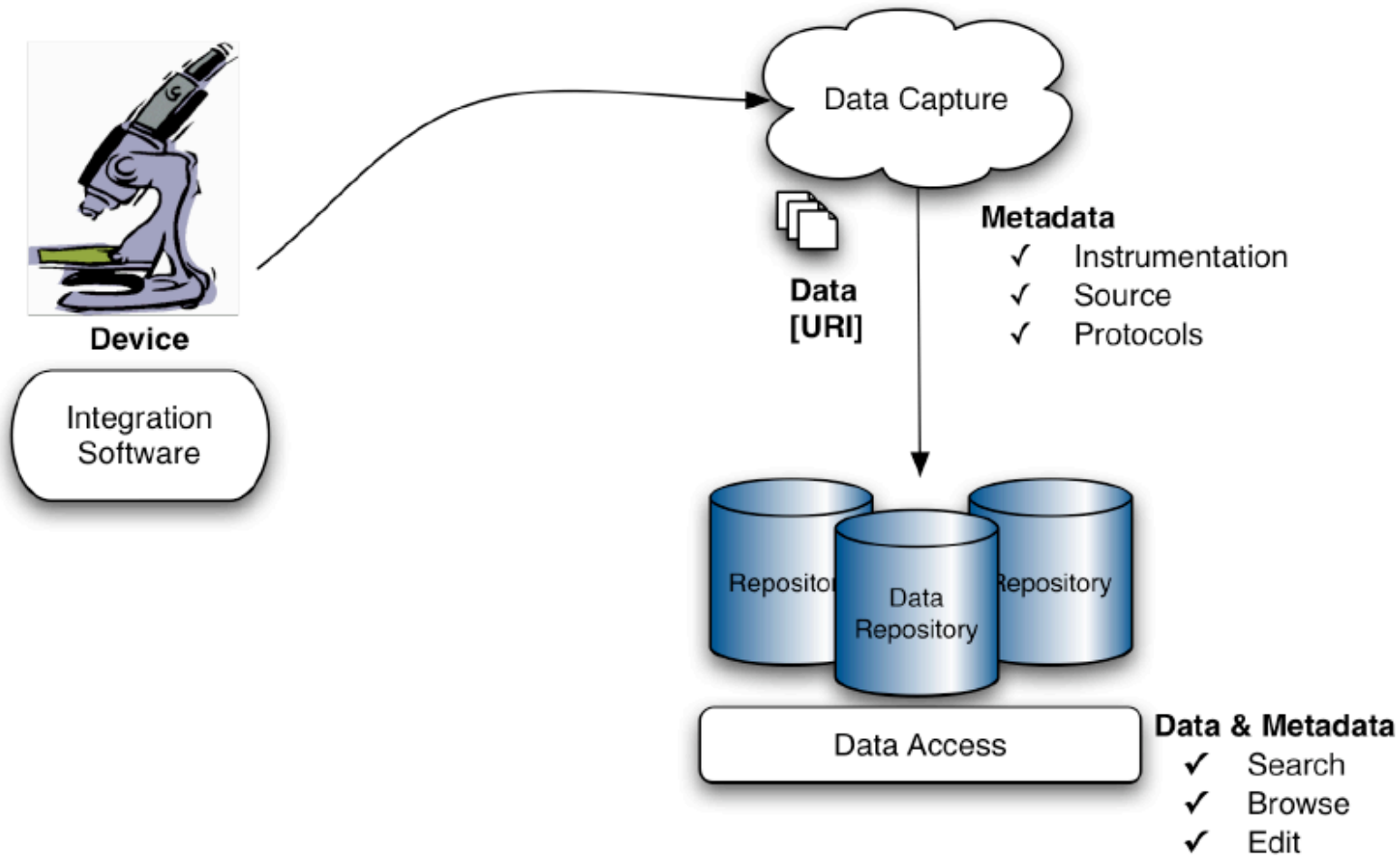
Data Capture



Integration 'as is' and 'step-wise'



Data Access



Easy to access and understand

The screenshot shows the Dudley LIMS software interface. The main window displays sample management details for a sample named 'ADK2_C'. The 'Notes' field contains the sequence 'TCGAGATTACTAAGCAATTCCAC'. Below the notes, a diagram illustrates the primer design for a 'Yeast ORF or kanMX4' gene. The diagram shows a green bar representing the gene with 'ATG' at the start and 'TAA' at the end. Four primers are indicated: 'A primer' (forward), 'B primer or kanB primer' (reverse), 'C primer or kanC primer' (forward), and 'D primer' (reverse). The sequence 'TCGAGATTACTAAGCAATTCCAC' is shown above the gene, with arrows indicating the primer binding sites. The diagram also includes coordinates for deletions: 'A_KAN_Deletion-644bp or All_WildType-945bp', 'A_D_Deletion-2155bp or A_D_WildType-1249bp', and 'D_KAN_Deletion-853bp or CD_WildType-810bp'.

The screenshot shows a web browser displaying a data table and a form for sample management. The table lists various properties and their values for a sample named 'KWS_day0_CD45RAnege_CDOR7nega'. The table has columns for 'Types', 'Name', 'X', 'X', 'nt:base', and 'Value'. The values include 'mix:referenceable', 'Sdbid826-a3f0-4f3b-9d4c-81650979f6b', '/sampleData/microarray/chips/Human_U133_Plus_2.0', 'Human', '/metarrays/Affymatrix/core/probe_data/200603/20060315', 'KWS_day0_CD45RAnege_CDOR7nega', 'jrnile', 'K_0_nega', and '15 March 2006'. Below the table, there is a form with fields for 'Add mix', 'Remove mix', 'Add named property', 'Add wildcard property', 'Add new', 'Move node to', 'Copy node to', 'Export the node', 'Locking', and 'Delete form'.

This site will provide access to data processing pipelines, results from pipeline runs, and data visualization and integration under the NIAD 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	Affymatrix 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	Affymatrix 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	Affymatrix 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 Medians Data Matrix

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

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

View Additional Pipeline Results

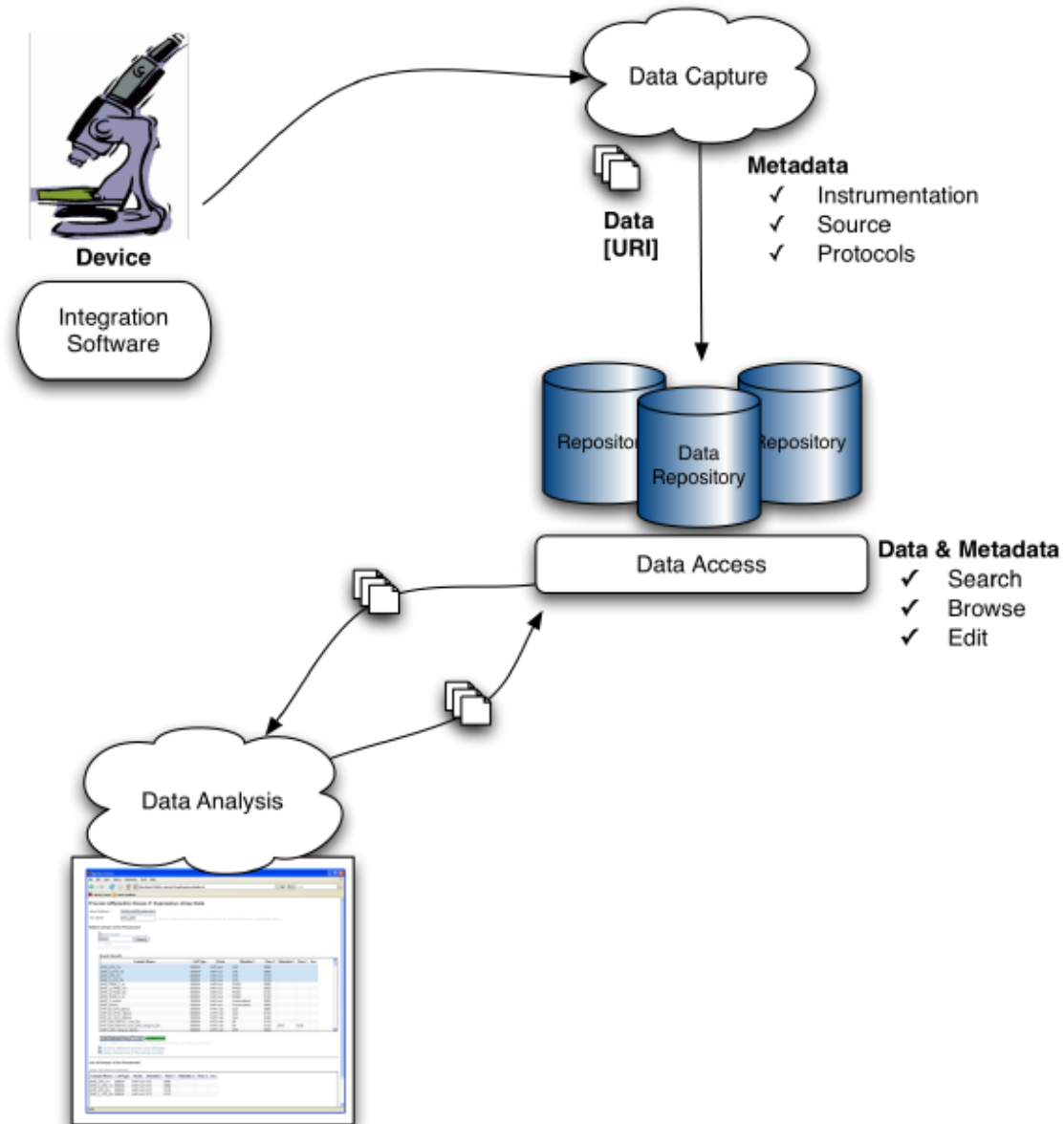
- Access all recent pipeline runs

Useful Tools

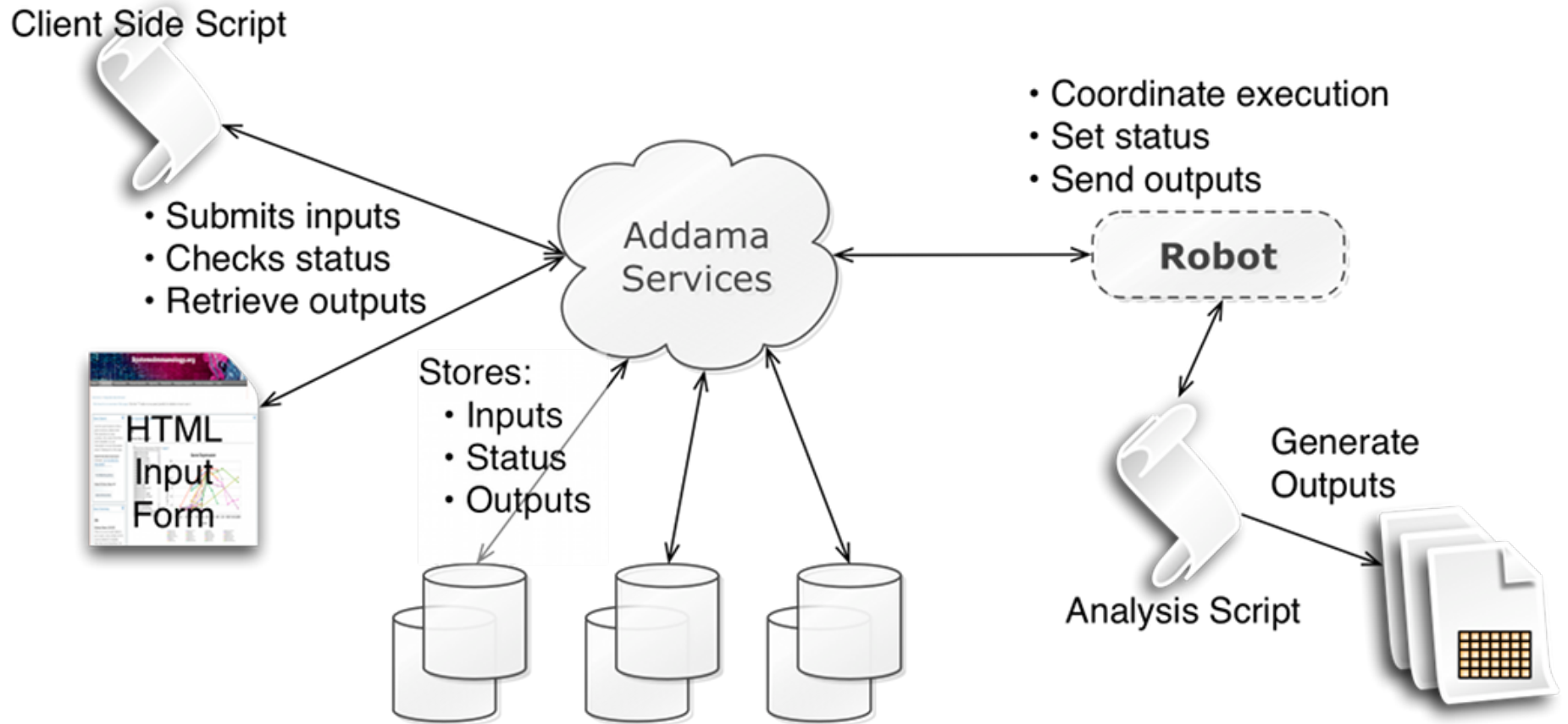
- Mouse Analysis Resources

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

Data Analysis



Support Ad-hoc Analysis



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](#)

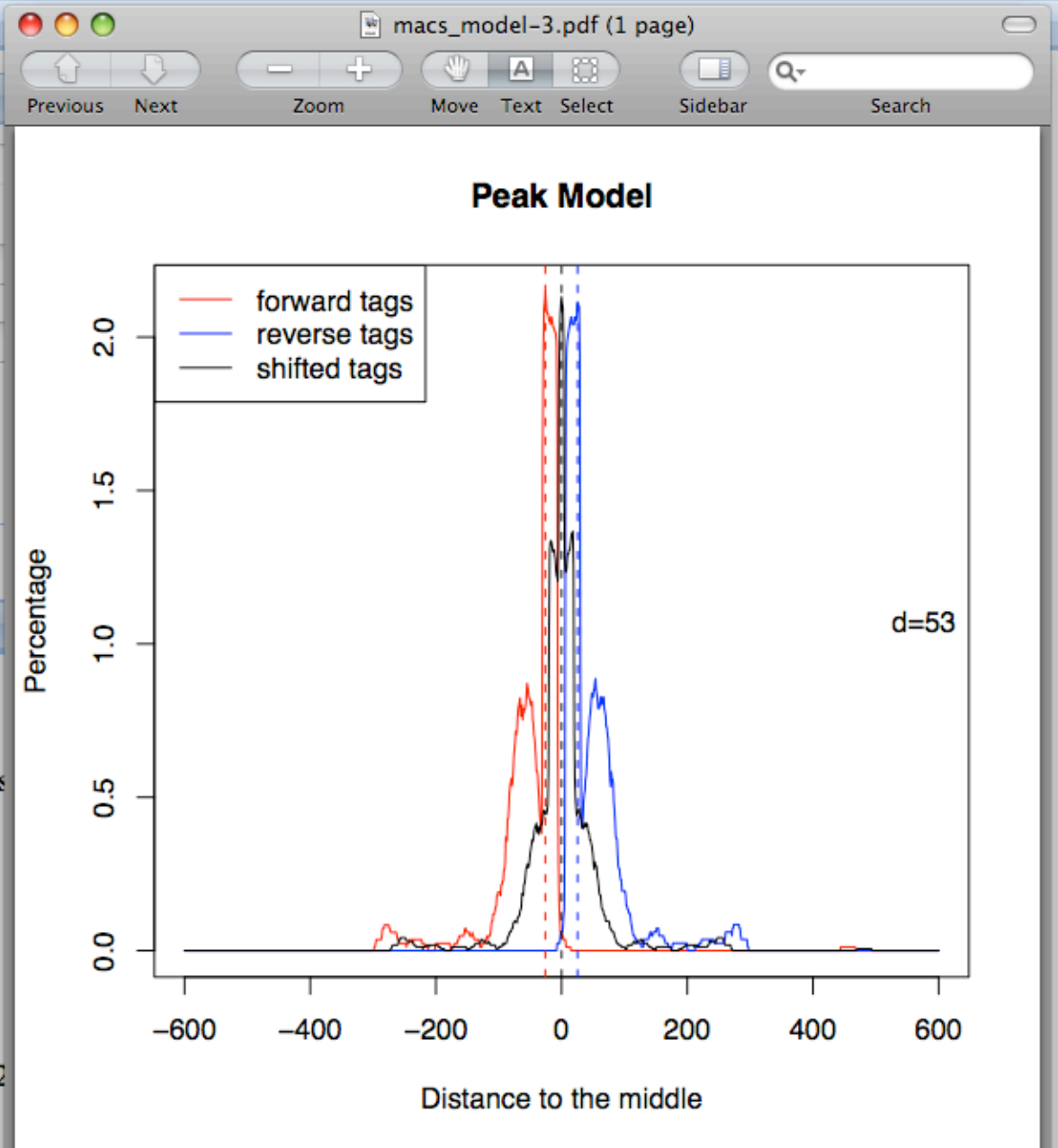
Inputs:

ChIP-Seq File: s_1_eland_multi_200k.txt

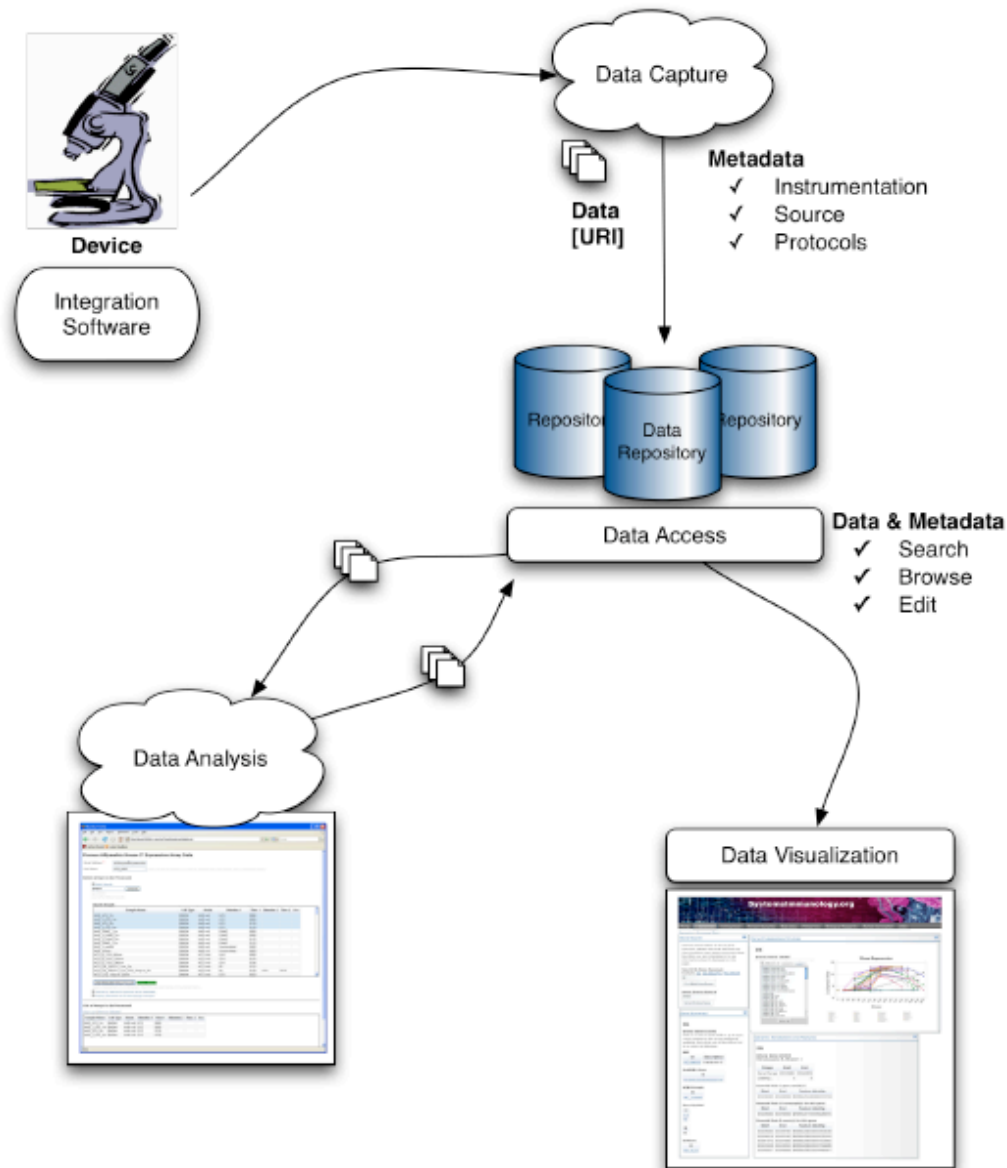
Selected File Format: ELANDMULTI

Sample Name: sample-1007 Desc-> yeast-unknown-x2

MACS Optional Parameters: --mfold=16



Data Visualization



Data Visualization

SystemsImmunology.org

[Home](#) [Genomics](#) [Computation](#) [Forward Genetics](#) [Signaling](#) [Proteomics](#) [Biological Reagents](#) [Human Correlation](#) [Links](#)

Genomics » Processed Data

Gene Search

Use the search below to find a gene synonym, please note that searches are case sensitive, then select the Entrez Gene Identifier you are interested in to see information about it displayed on this page.

Search By Gene Synonym
Example: [IL6](#), [GO:0001791](#), [MGI:96559](#)

Select Entrez Gene Id

Gene Expression Studies

II6

Entrez Gene 16193

Adierem 3' (specs) [[esastd](#)]

EMDP_AIRI-mut_LPS

EMDP_AIRI-mut_ENR2

EMDP_AIRI-mut_Antibiotatad

EMDP_AIF3-mut_G60

EMDP_AIF3-mut_D4

EMDP_AIF3-mut_D4_LPS

EMDP_AIF3-mut_LPS

EMDP_AIF3-mut_MM2

EMDP_AIF3-mut_Poly-IC

EMDP_AIF3-mut_Unstimulated

EMDP_IL6

EMDP_IL6_CpG

EMDP_IL6_D4

EMDP_IL6_D4_LPS

EMDP_IL6_D4_MM2

EMDP_IL6_D4_Poly-IC

EMDP_IL6_3hrbeta

EMDP_IL6_3hrgamma

EMDP_IL6_LPS

EMDP_IL6_LPS_MDP

Gene Summary

II6

Entrez Gene 16193

Click on a row in each table to go to open a new window to the source website if available. Note that not all identifiers link to an external database.

MGI

Id	Description
MGI:96559	Interleukin 6

Ensembl Gene

Id
ENSMUSG00000025746

NCBI Protein

Id
NP_112445

Gene Symbol

Id
IL6
IL6

UniGene

Id
Mm.1019

Genomic Annotations and Features

II6

Entrez Gene 16193

Chromosome 5, Strand +

Range	Start	End
Gene Range	30339683	30346508
Loading...	0	0

Ensembl lists 1 gene result(s)

Start	End	Feature Identity
30339683	30346508	ENSMUSG00000025746

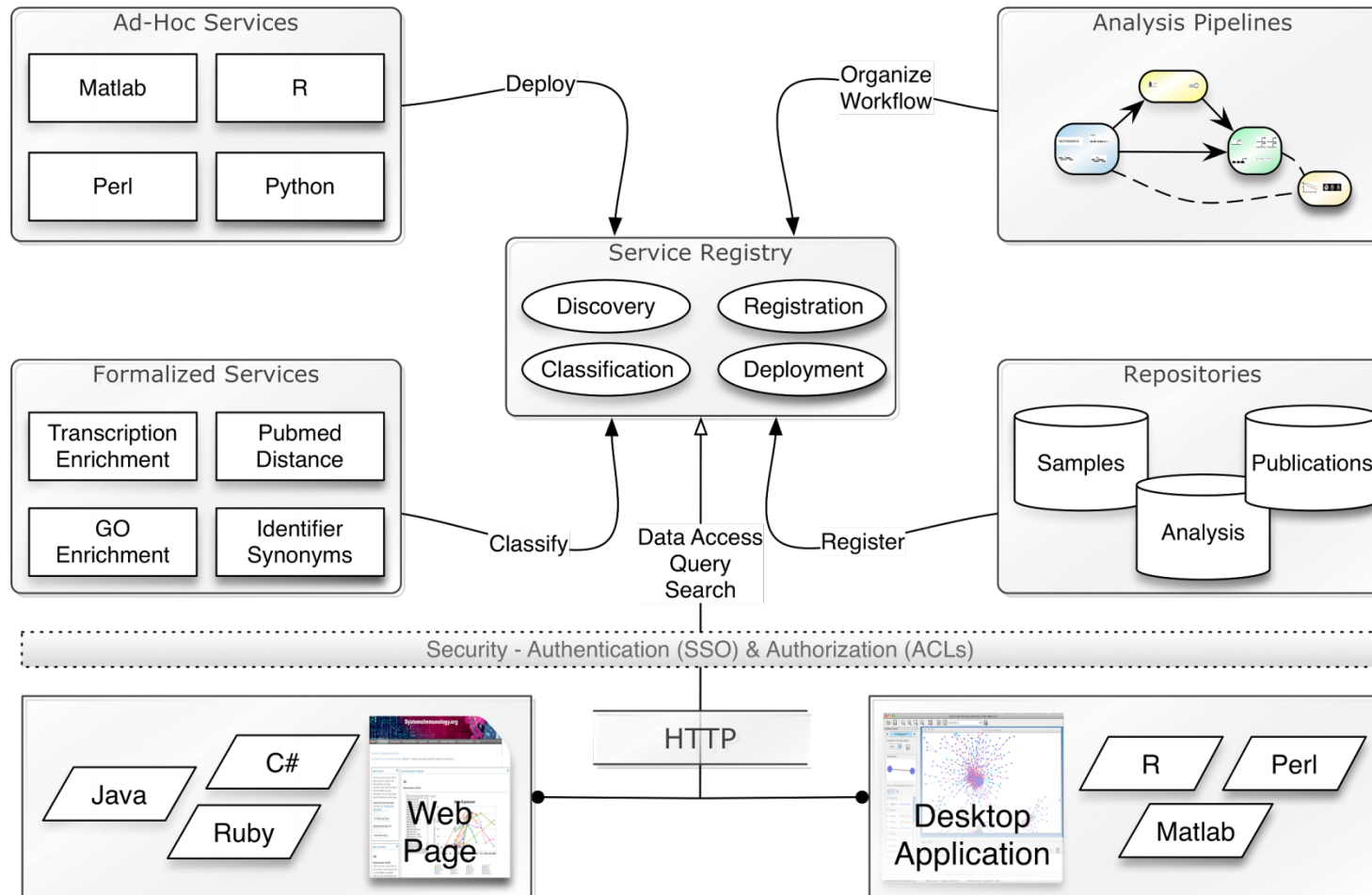
Ensembl lists 1 transcript(s) for this gene

Start	End	Feature Identity
30339683	30346508	ENSMUST00000026845

Ensembl lists 5 exon(s) for this gene

Start	End	Feature Identity
30339683	30339750	ENSMUSE00000406548
30339916	30340100	ENSMUSE00000276162
30341372	30341485	ENSMUSE00000152883
30344545	30344694	ENSMUSE00000152885
30345921	30346508	ENSMUSE00000406807

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

