



DOE Systems Biology Knowledgebase

KBASE

Data and modeling for
predictive biology

BERAC Meeting
February 16, 2012

Adam P. Arkin, Ph.D.
Physical Biosciences Division
Lawrence Berkeley National Laboratory



U.S. DEPARTMENT OF
ENERGY

Office of
Science

- **DOE mission:** predict, control and design the biological components of energetic processes and environmental balance.
- Complex missions with rapidly expanding, intricately related diverse data types requires a mean to augment scientists' ability to:
 - Filter information
 - Focus attention
 - Ask the right questions
 - Leverage other minds

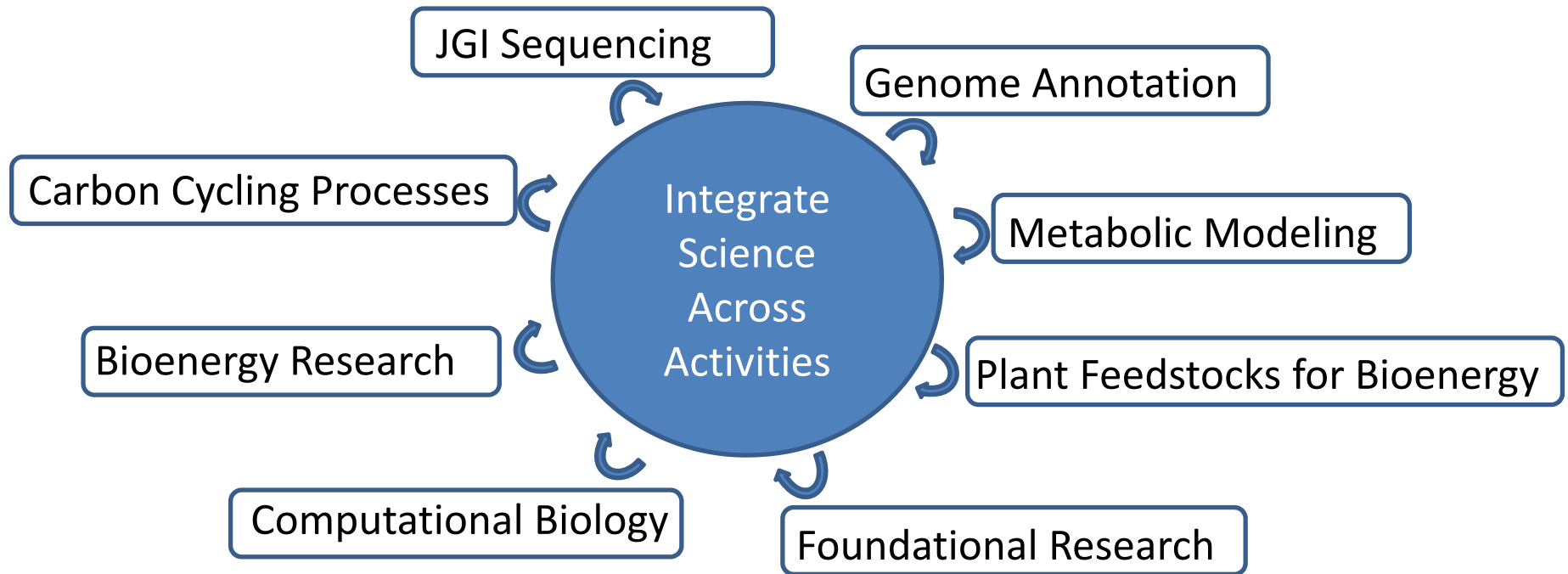




KBASE
predictive biology

DOE Systems Biology Knowledgebase

Integrate science and the science community

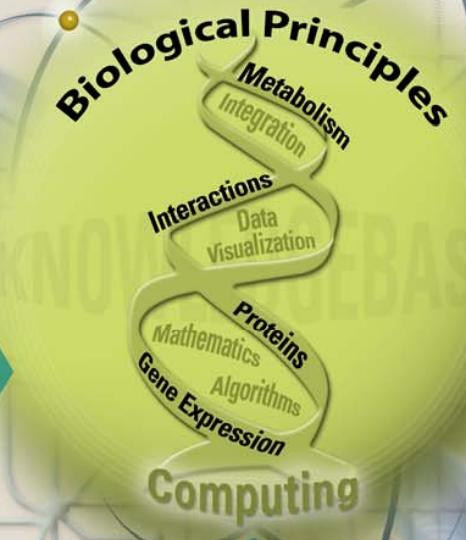


There is a tremendous wealth of data and information in the Genomic Sciences program. The [Knowledgebase \(Kbase\)](#) is an opportunity to integrate this data and information both within individual activities as well as to integrate together different activities.

DOE Systems Biology Knowledgebase

Components Enabling Its Development

Plants and Microbes
for Energy and Environment



2010 Knowledgebase R&D Project

Year-long effort funded by American Recovery and Reinvestment ACT (ARRA).

Results from this project, which was completed in September 2010, are underpinning Knowledgebase development:

- DOE Systems Biology Knowledgebase Implementation Plan
- ARRA pilot projects

User Community Data and Resources

The Knowledgebase will leverage and establish critical partnerships with other synergistic efforts:

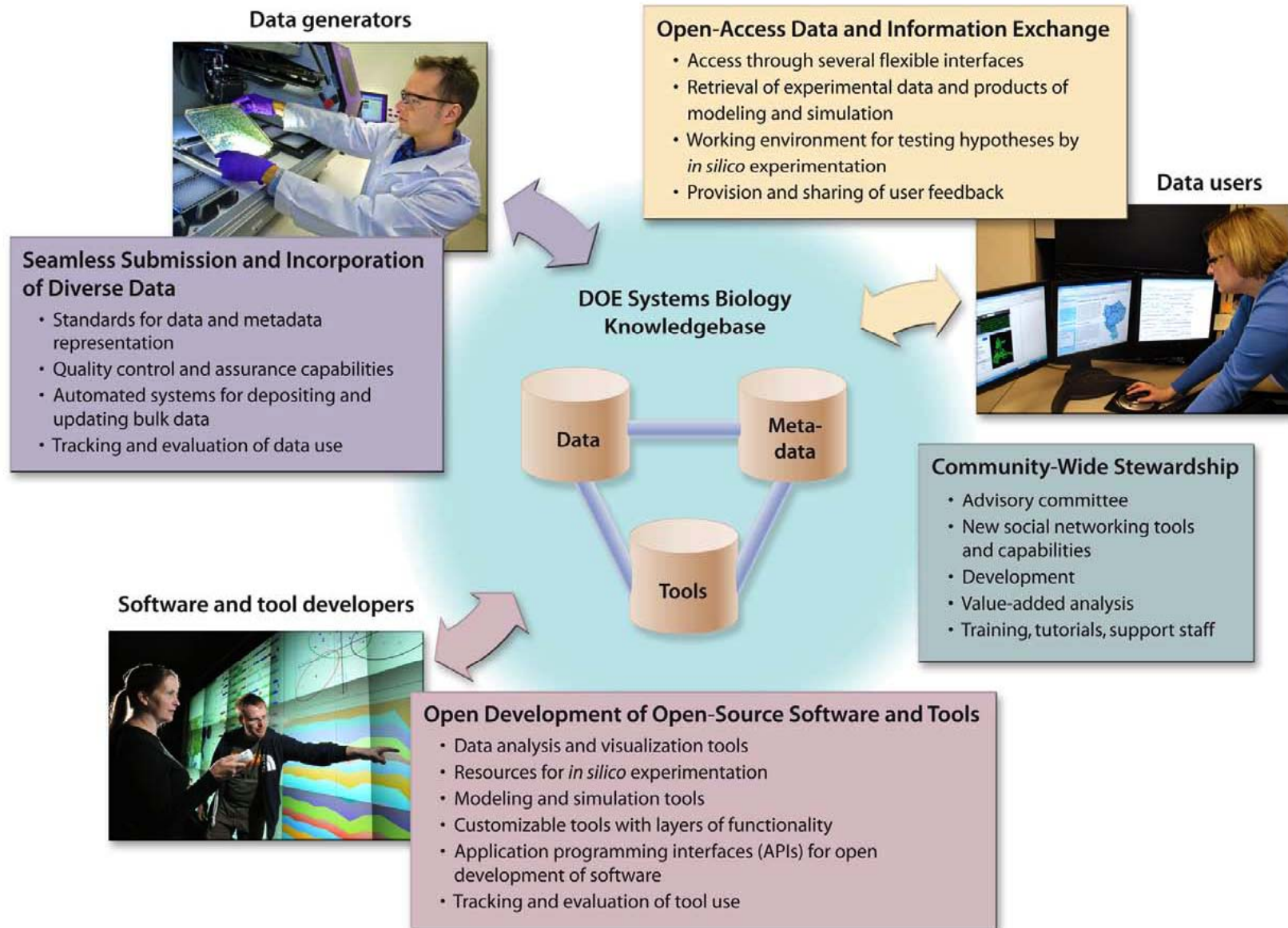
- DOE Joint Genome Institute
- National Center for Biotechnology Information
- iPlant Collaborative
- DOE Office of Advanced Scientific Computing Research
- Others

University-Led Projects to Develop Computational Biology and Bioinformatic Methods Enabling the Knowledgebase

In 2010, the DOE Office of Biological and Environmental Research awarded funding to 11 projects for basic research that will support Knowledgebase development in four areas:

- Omic data integration
- Integrated pathway reconstruction
- Genomic annotation
- Whole cellular simulations

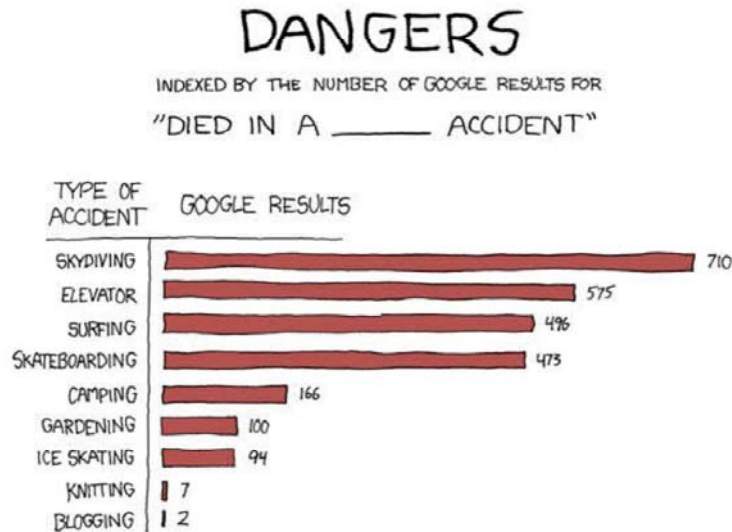
Predictive Understanding



Kbase, working to build a(n)...

- **Knowledgebase** enabling *predictive* systems biology.
- Powerful modeling framework.
- **Community-driven**, extensible and scalable **open-source** software and application system.
- Infrastructure for integration and reconciliation of algorithms and data sources.
- Framework for standardization, search, and association of data.
- Resource to enable **experimental design** and **interpretation** of results.

Data is extracted and displayed



This is the database model

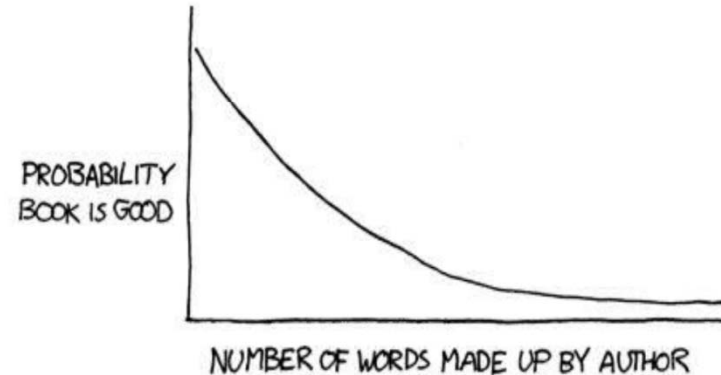
Knowledgebases should *learn* a
"model" of the data to provide
"conclusions" (hypotheses)

M-C Jenkins (<http://www.scienceforseo.com>)

Images from xkcd comics

Databases enable the rapid organization and **search** of data

Knowledge is learning & answering

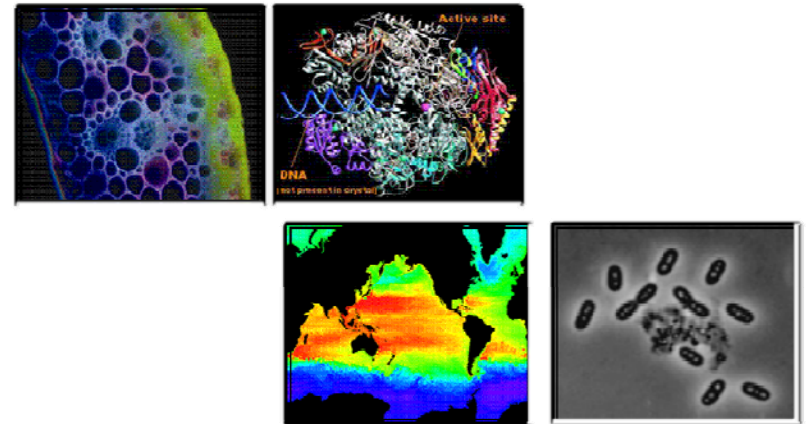


"THE ELDERS, OR FRAÄS, GUARDED THE FARMLINGS (CHILDREN)
WITH THEIR KRYTOSES, WHICH ARE LIKE SWORDS BUT AWESOMER..."

This is the knowledgebase model

Utilize existing commercial software technology and leveraging DOE internet resources (ESNet) and DOE cloud computing platforms (Magellan)

➤ ***Kbase is a framework for data collection, integration and analysis tools to enable the simplified use and modeling of large scale genome and genome enabled information***



By 2013 Deliver on the Initial Goals:

- Kbase Infrastructure firmly established at 4 laboratories including high performance and cloud computing with routine 10+ Gb/s data transfer over ESNet between all Kbase sites
- First public release includes:
 - Integration of data to reconstruct and predict metabolic and gene expression regulatory networks for up to 1,000 microbes to manipulate microbial function.
 - Integration of phenotypic and experimental data for bioenergy plants to predict metabolic and regulatory genotypes enabling manipulation of biomass properties.

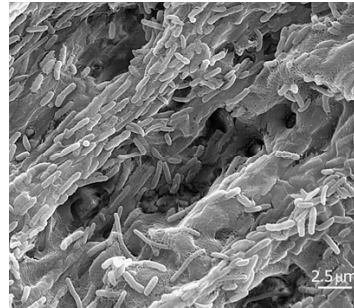


KBASE
predictive biology

DOE Systems Biology Knowledgebase

HOW ARE WE GOING TO DO IT?

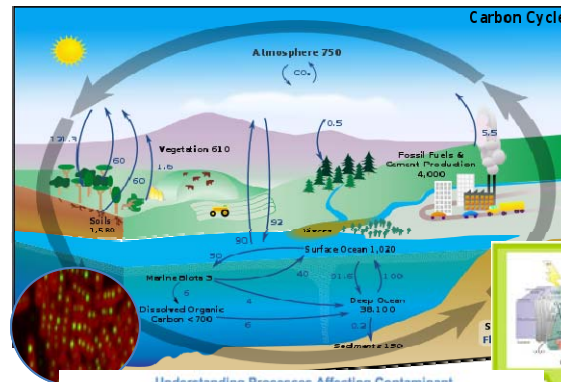
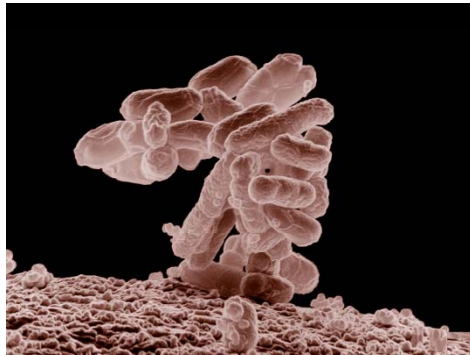
Communities



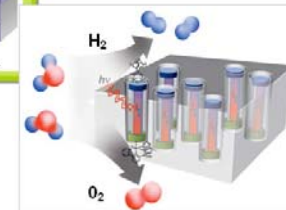
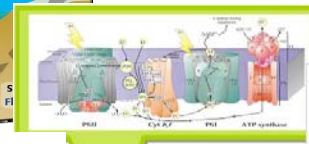
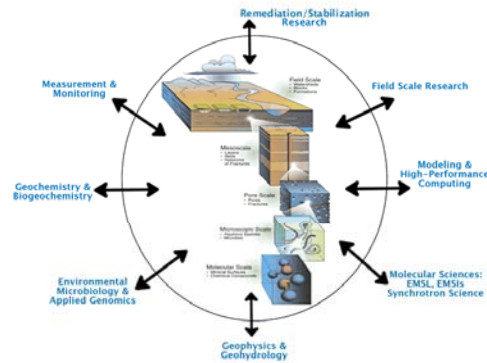
Plants



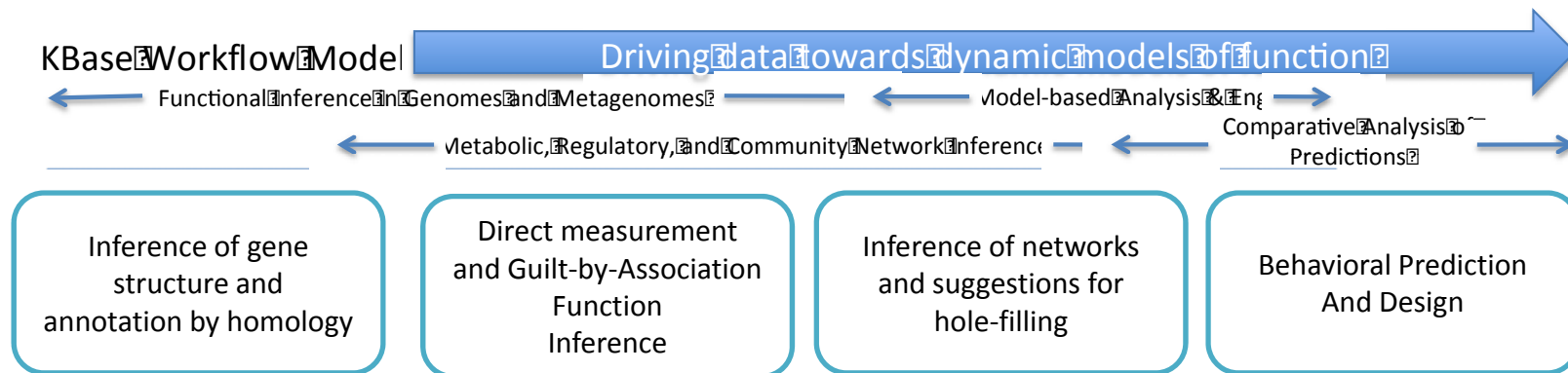
Microbes



Understanding Processes Affecting Contaminant Transport Across Scales in the Subsurface



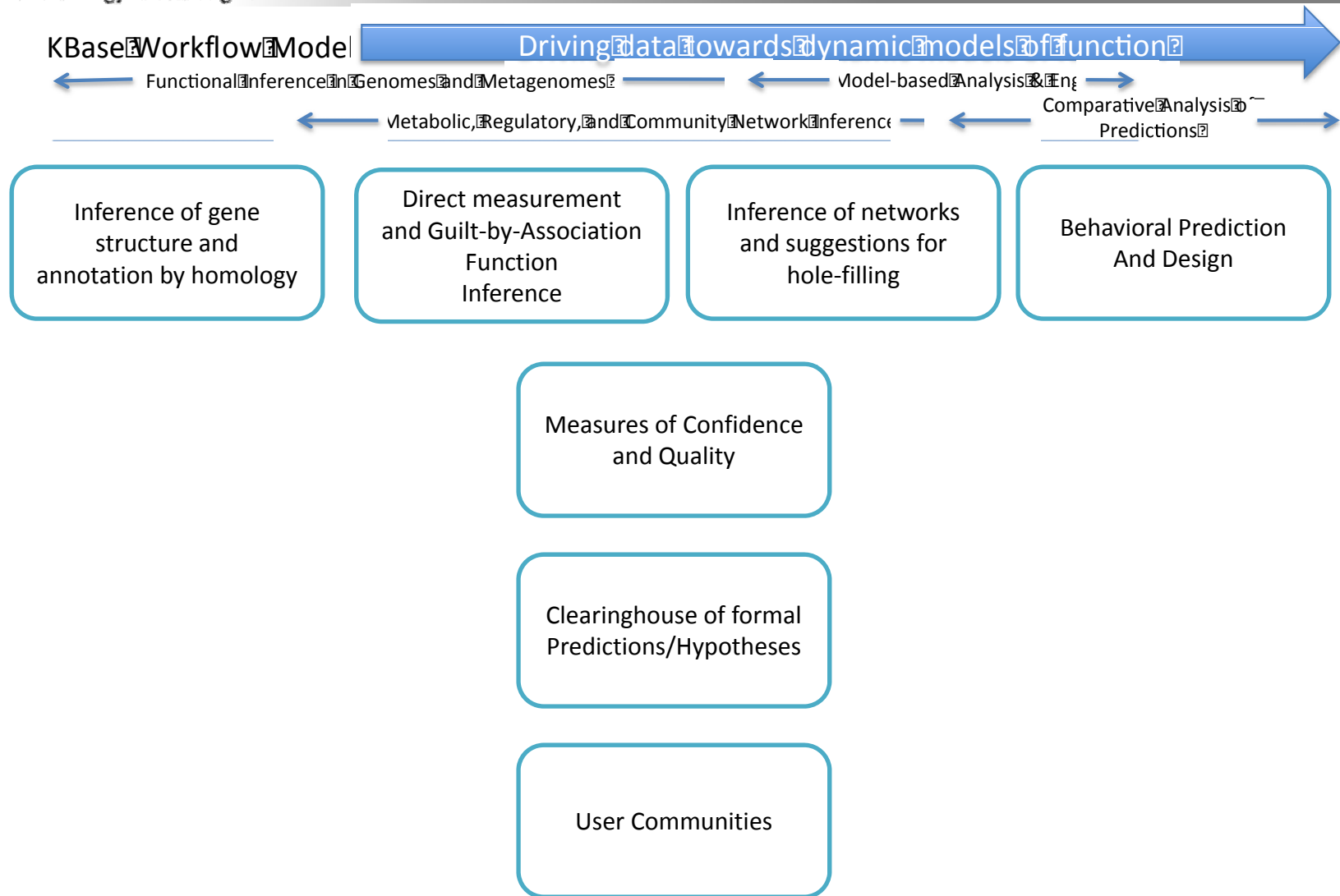
Target “Activities”





DOE Systems Biology Knowledgebase

Transformative capability



Inference of gene structure and annotation by homology

Direct measurement and Guilt-by-Association
Function Inference

Inference of networks and suggestions for hole-filling

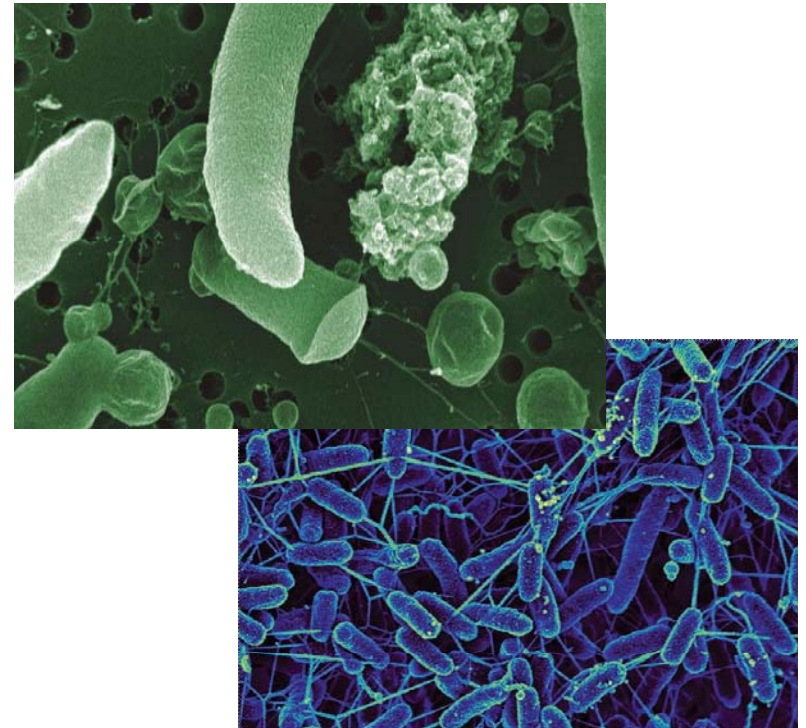
Measures of Confidence and Quality

Clearinghouse of Predictions/Hypotheses

User Communities

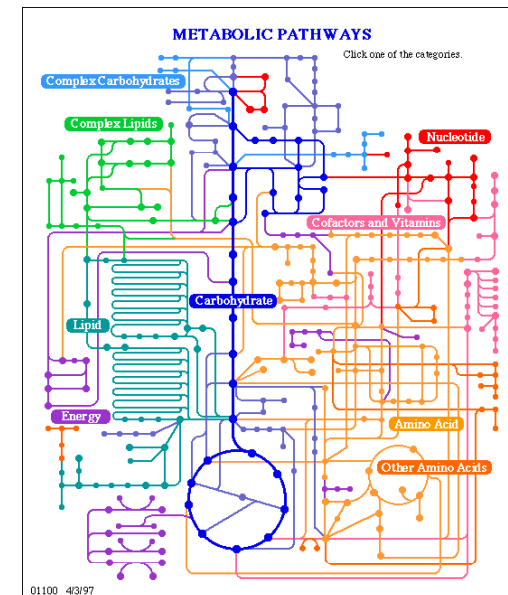
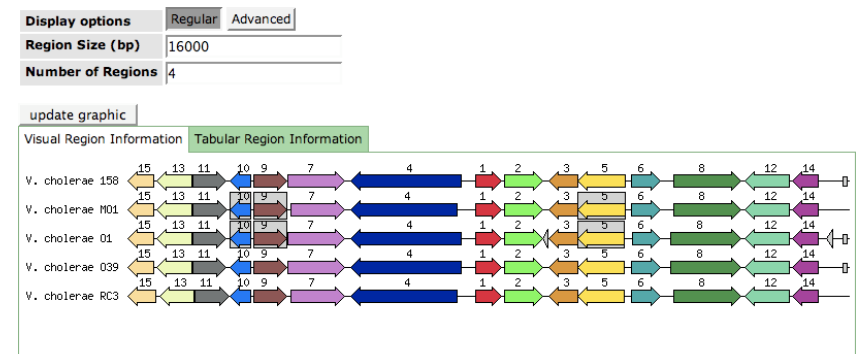
Our overall goals are to:

- Reconstruct and predict metabolic and gene expression regulatory networks to manipulate microbial function
- Vastly increase the capability of the scientific community to communicate and utilize their existing data
- Enable the planning of effective experiments and maximizing understanding of microbial system function



We propose to do this by:

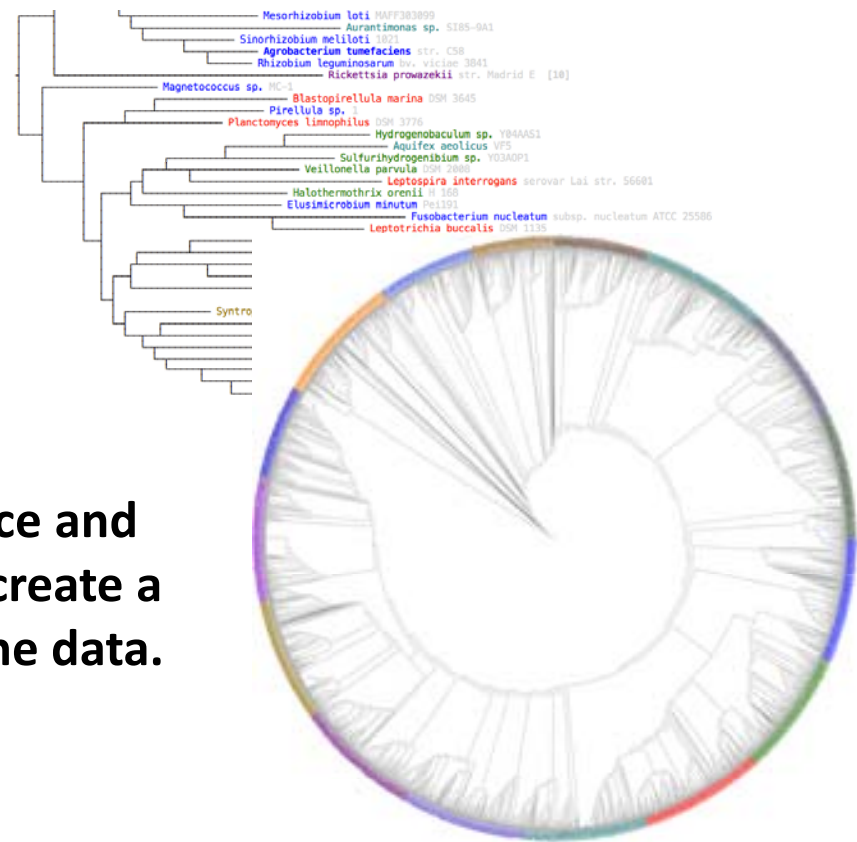
- Annotating genomes and assigning confidence
- Reconstructing metabolism and optimizing for function
- Reconstructing regulation and assessing agreement with expression data
- Integrating and standardizing -omics data from multiple data sources
- Constructing models of microbial organisms and interlinking models with data



Within 13 months, we will be able to demonstrate use of the following:

- Data integration and data model
- Next generation organism pages
- Phylogenetic tree services
- Next generation gene pages
- Metabolic modeling
- Regulatory/Transcriptional networks

A microbiologist with a genome sequence and phenotypic growth data will be able to create a metabolic model fully reconciled with the data.

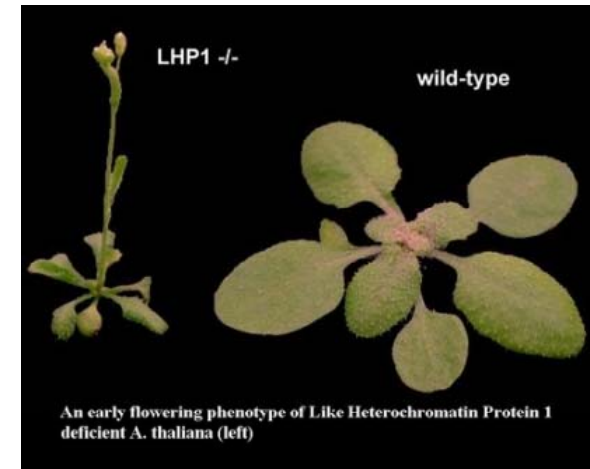


Our overall goals:

In order to extract knowledge from the wealth of high-throughput data in plant biology we need the ability to meaningfully integrate data.

Therefore we aim to:

- Deploy Kbase capability that will allow for interactive, data-driven analysis and exploration across multiple -omics experiments.
- Provide researchers in plant sciences access to comprehensive datasets from high-throughput experiments together with relevant analytical tools and resources.
- Provide a platform for researchers to analyze their own experimental data, and have these results incorporated into the data exploration framework.



We will accomplish this by using:

- Advance storage and indexing strategies for fast but persistent retrieval of large-scale genomics data
- Massively parallel processing of raw data using cloud compute resources
- Interactive charting libraries
- Using controlled vocabularies and ontologies for describing and storing genomic data



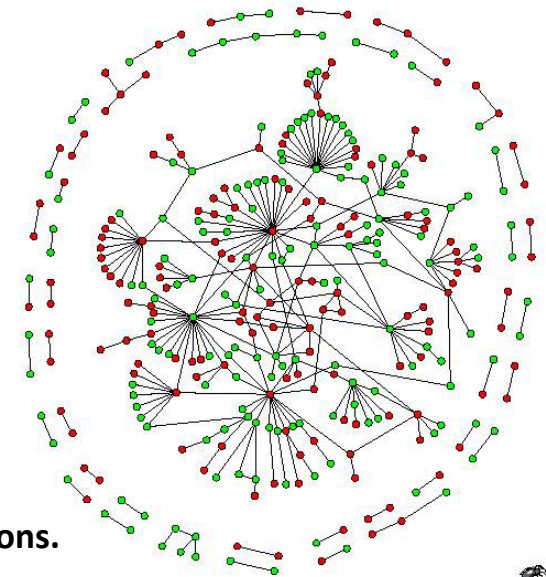
Within 13 months, we will have the following capabilities:

Genotyping Workflow

- Create a workflow for rapidly converting sequencing reads into genotypes
- Demonstrate more than 100-fold speedup over serial version by leveraging capabilities of KBase cloud
- Workflows will be developed as part of the KBase CyberInfrastructure

Data exploration: Linking of gene targets from phenotype and genotype studies with co-expression, protein-protein interaction, and metabolic models

- Allow users to narrow candidate gene lists based on these integrated data types
- Recommender system based on “guilt-by-association” principle: identify other genes, expression datasets, etc. associated with the user-selected group of genes
- Project genetic variations and network edges onto metabolic pathways
- Visualize co-expression network and node Interactions among subnetworks correlated to phenotype of interest



Plants work is cross-cutting with microbial and communities data and predictions.

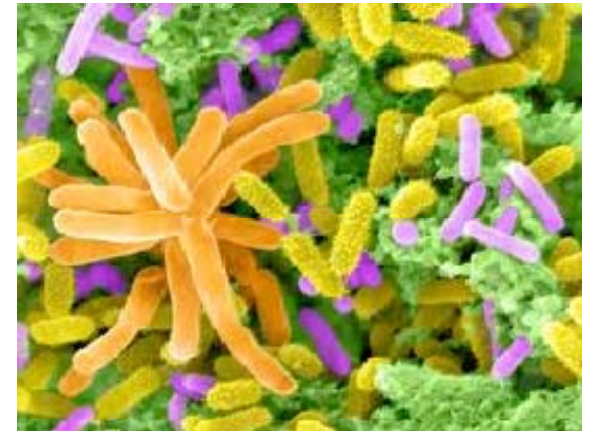


There has been an explosion of metagenomics data:

- Systems biology is driven by the ever-increasing wealth of data
- Metagenomics is >90% of the data
- Computation needs to be smarter

Our overall goal is to build a Kbase metagenomic platform that provides:

- Scalable, flexible analyses, link physiological and metadata sets to metagenomic sequences
- Data QC and GSC compliant data and standards for data collection
- Enable modeling of metabolic processes within a community
- Predict microbial growth in isolation and in a community



Within 13 months, we will have the following capabilities:

Metagenomic Experimental Design Wizard

The foundation laid will enable researchers to perform *in silico* experimentation and hypothesis testing

Bioprospecting

- Find communities with similar alpha diversity
- Find communities in similar biomes
- Locate novel proteins (unknowns)
- Suggest functions (based on metadata) that might be encoded in “abundant unknowns”
- Identify optimal candidates for screening



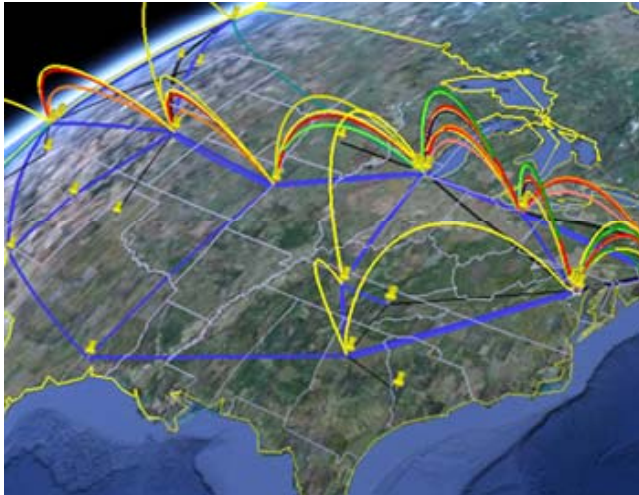
Communities work is cross-cutting with microbial and plant data and predictions.

KBASE is providing robust and scalable infrastructure to support new science by offering...

- High speed data transfers will enable better response times
- Access to back end storage systems for large data sets
- Remote compute services for HPC, cluster and cloud based
- Workflow support
- Web UX access to data and computing
- Persistent and transient data management capabilities
- Support for users, teams, projects and cross talk

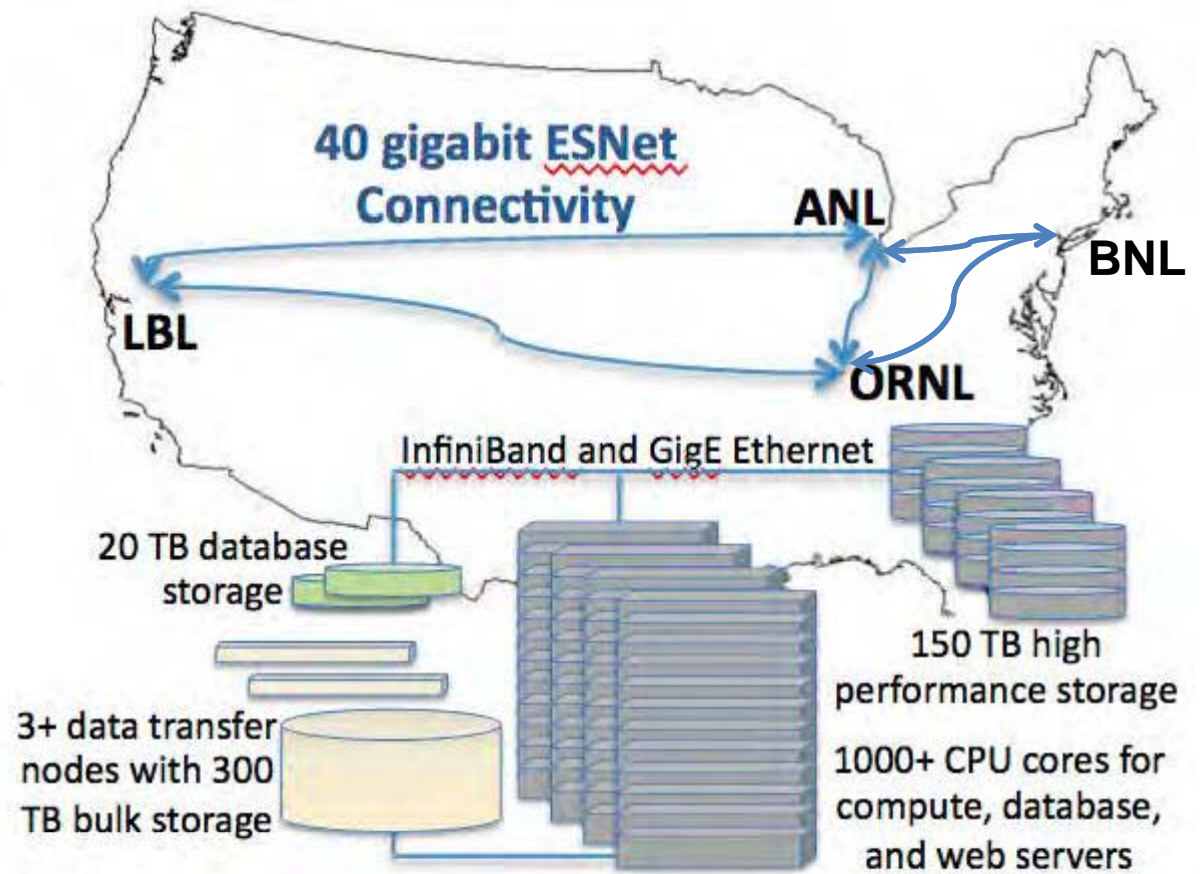


Energy Sciences Network (ESNet)



Kbase leverages ESNet for 10+ Gb/s data transfer between all nodes

- ESnet backbone (ESnet4) is anational 10 Gbps optical circuit infrastructure
- ESnet shares its optical network with [Internet2](#)
- ESnet's IP network functions as a Tier 1 internet service provider

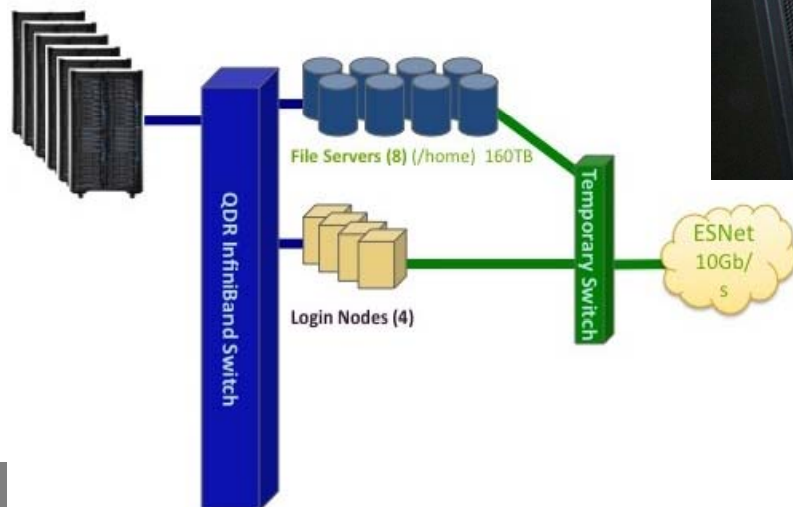


Argonne Magellan Hardware

Built on DOE Magellan Cloud Magellan Goals:

Establish a nationwide scientific mid-range distributed computing and data analysis testbed.

<http://magellan.alcf.anl.gov/>
<http://science.energy.gov/ascr>



Current config at 252nodes
Expected to reach ~700 nodes in FY12

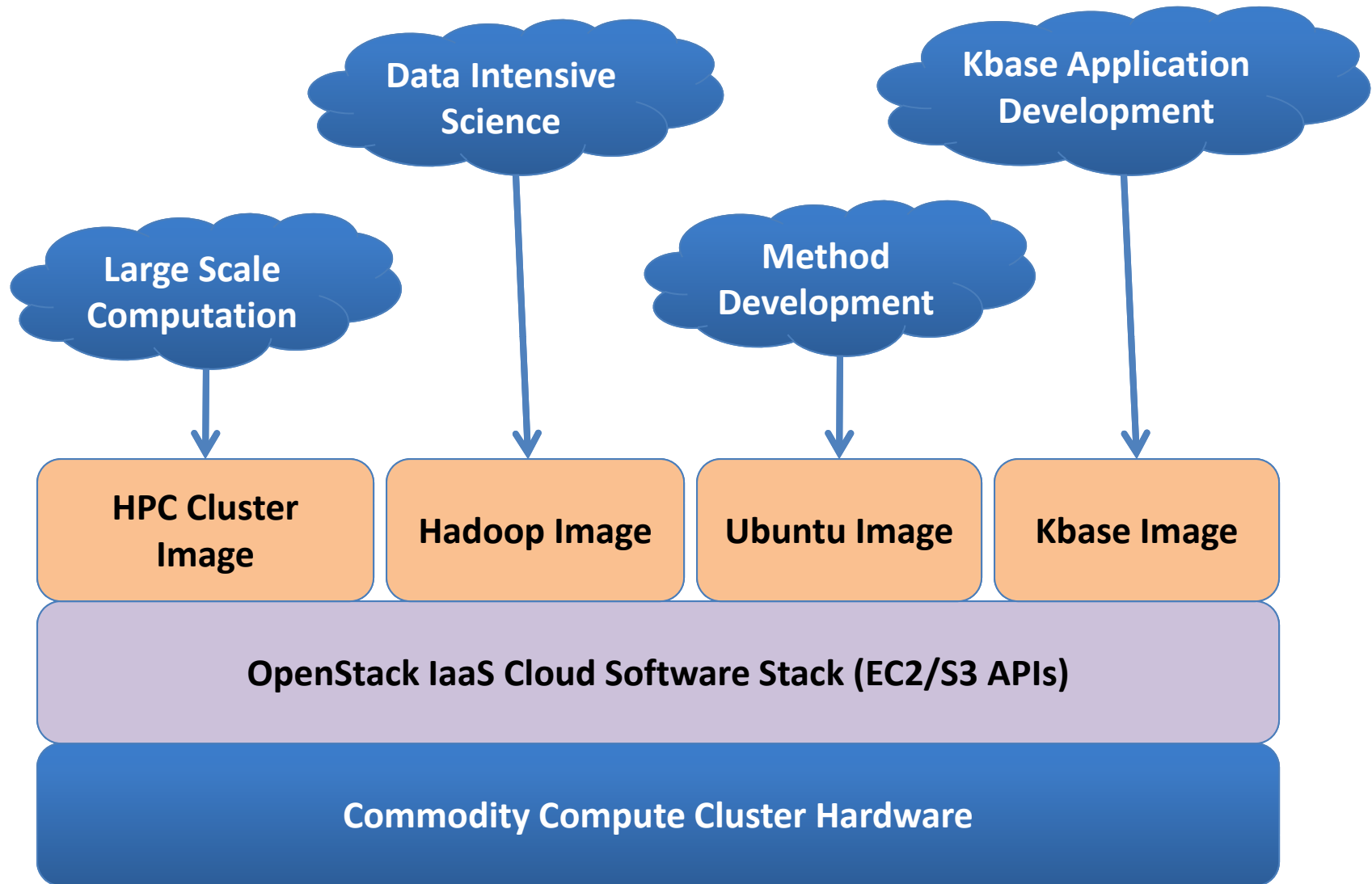
Smaller system housed at ORNL with
around ~70 nodes



KBASE
predictive biology

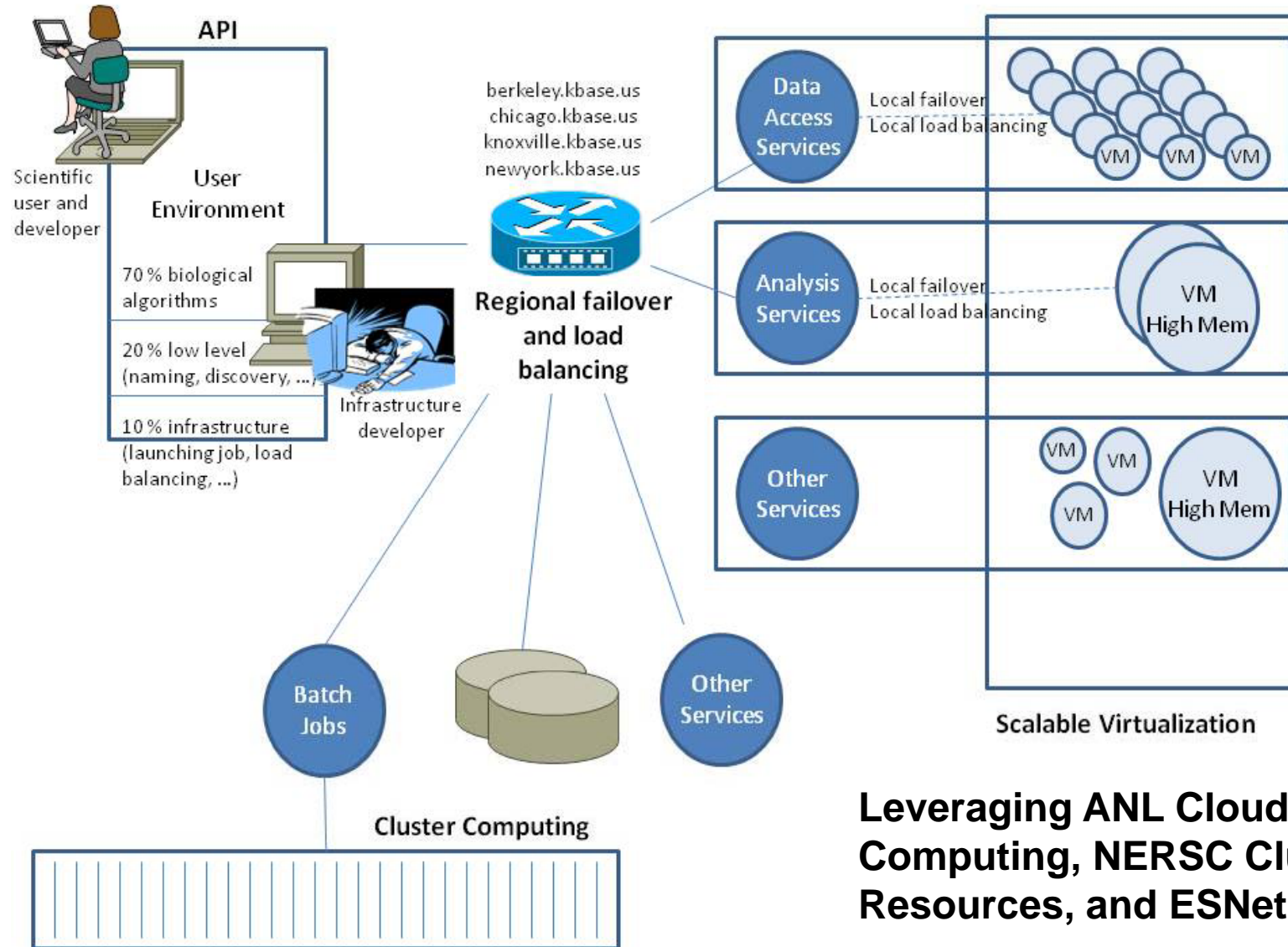
DOE Systems Biology Knowledgebase

The Kbase Cloud Architecture





The Kbase Infrastructure



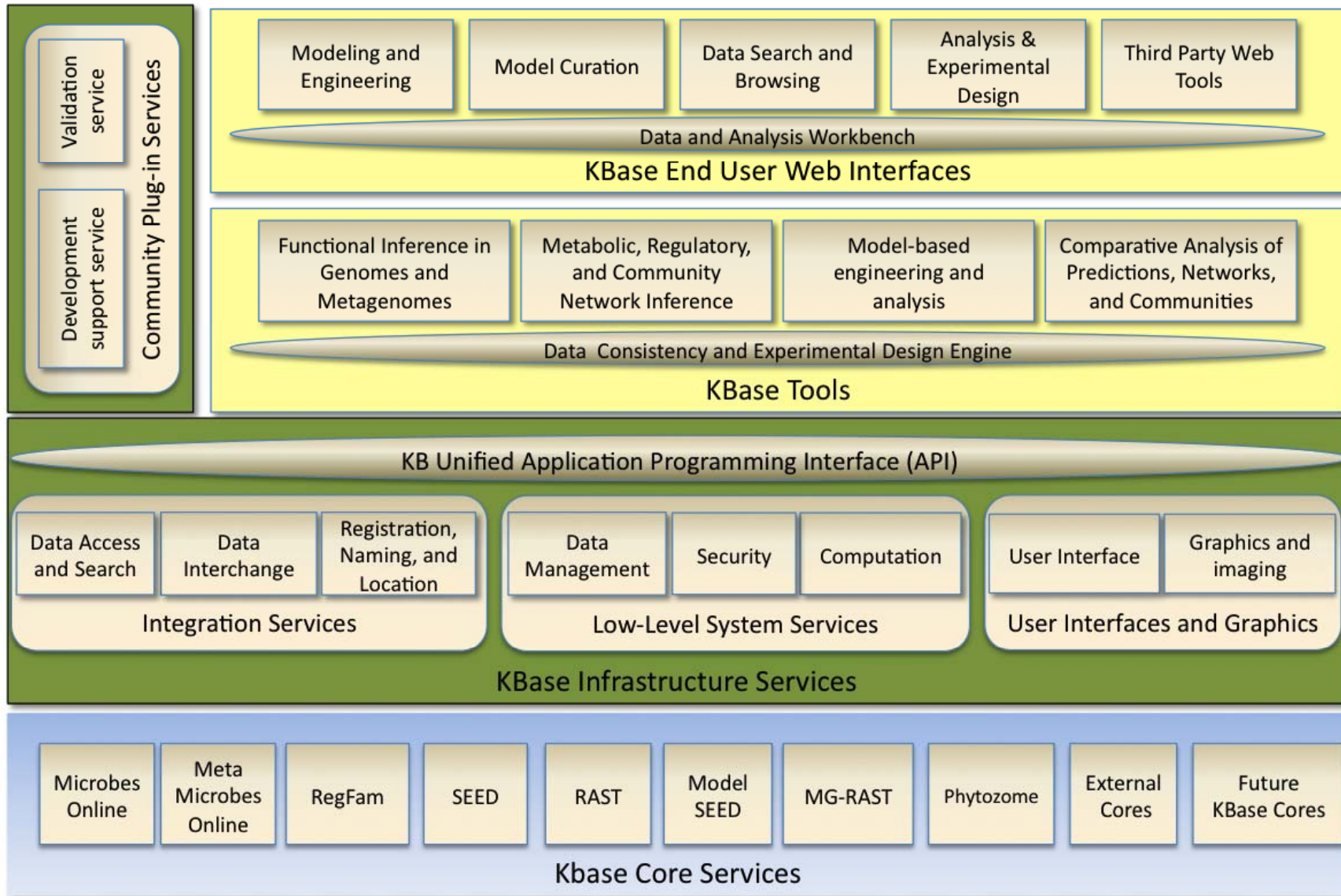
Leveraging ANL Cloud Computing, NERSC Cluster Resources, and ESNet



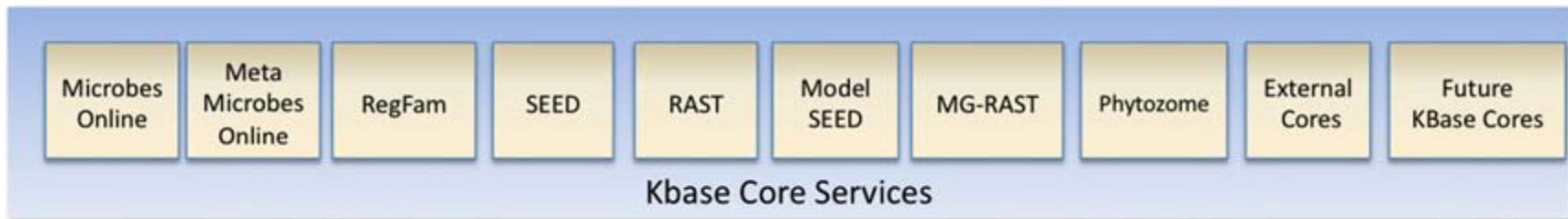
KBASE
predictive biology

The Kbase Infrastructure and Services

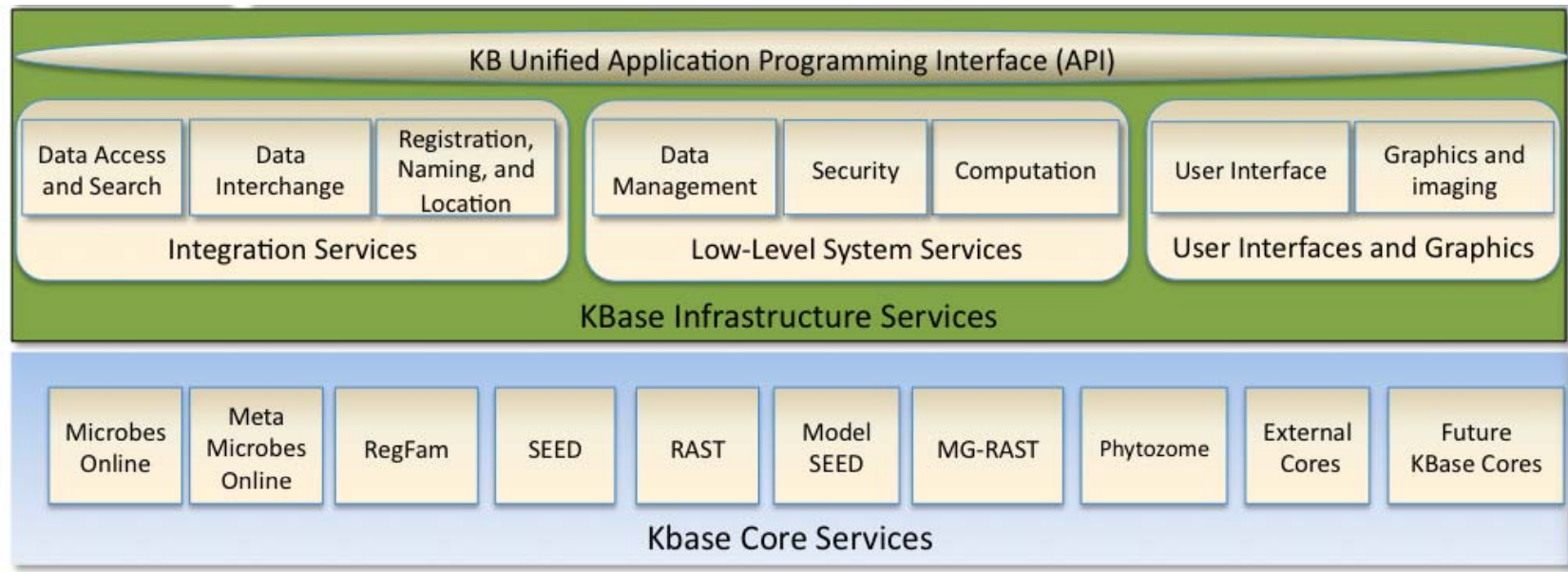
DOE Systems Biology Knowledgebase



- *KBase will integrate initial f core services and enable KBase Unified Application Programming Interface (API)*
- *The KBase Unified API is the primary programming target for KBase tools and applications*
- *KBase Unified API provides*
 - Access to core functions and data
 - Low-level systems services
 - User interface and graphics services and
 - Integration services

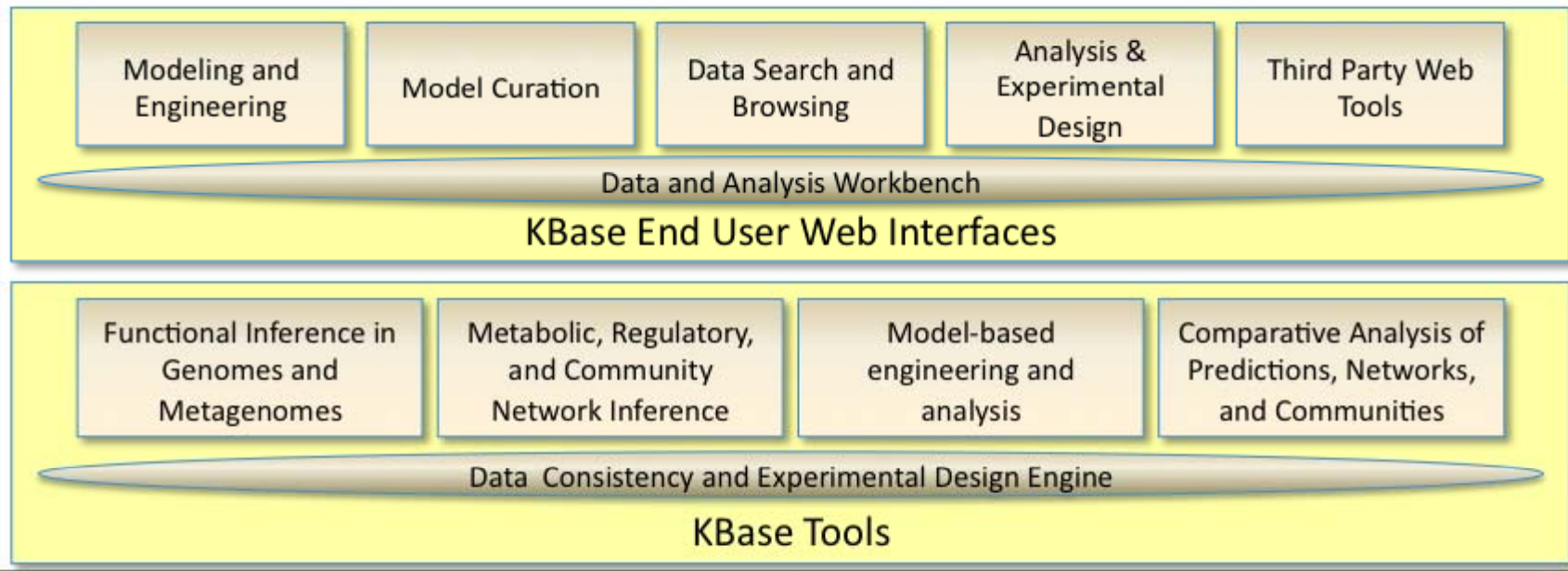


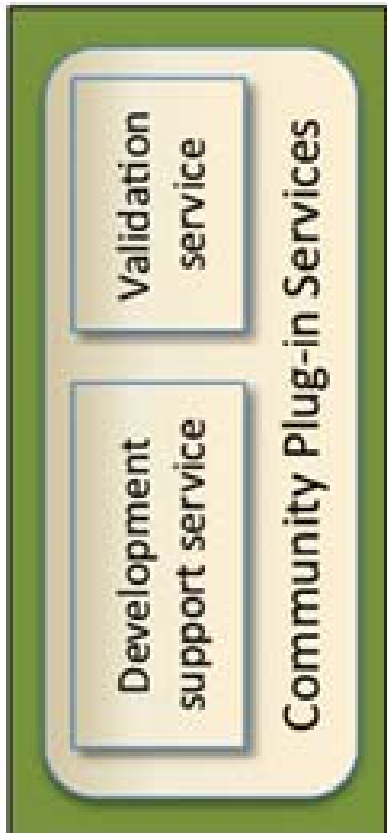
- Uniform view of KBase core services
- Access to data management, security and computational services in the KBase infrastructure
- Rich set of standard graphics and UX components for building web applications with single look and feel



Building a Kbase User Experience

- Most KBase code development will be layered on top of the KBase Unified API
- Supports a layered architecture consisting of collections of KBase Tools and End User Web Interfaces
- Each layer extends the API by providing additional support for high-level functions

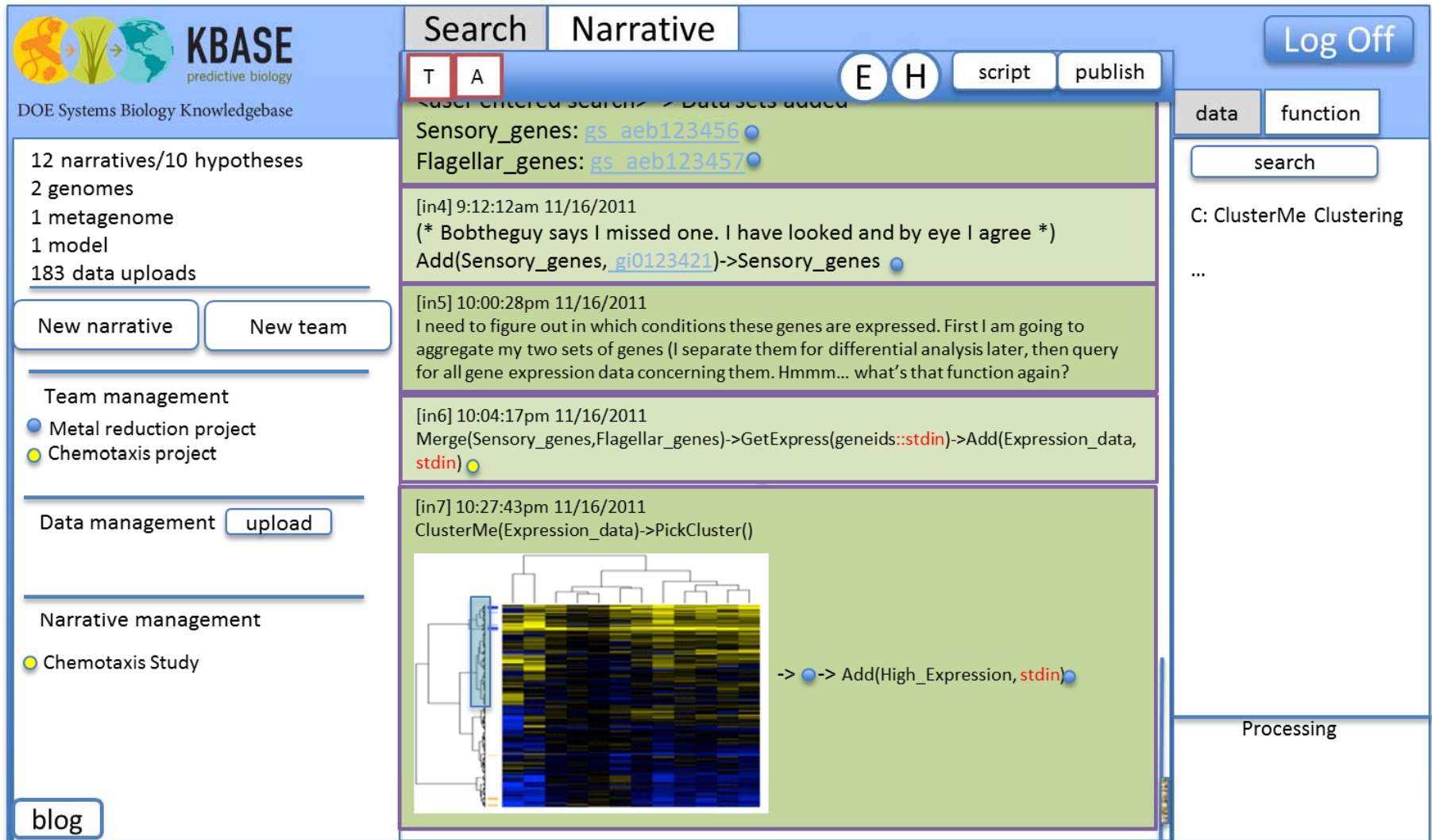




Two ways to work with KBase

- Extend from below with additional core services that ***extend the KBase API***
- Extend from above by plugging in applications that ***use the KBase API***

Concept: Kbase User Experience



The screenshot displays the KBASE user interface. At the top left is the KBASE logo and the text "DOE Systems Biology Knowledgebase". The main navigation bar includes "Search" and "Narrative" tabs, along with buttons for "E", "H", "script", "publish", and "Log Off".

On the left sidebar, there is a summary of content: "12 narratives/10 hypotheses", "2 genomes", "1 metagenome", "1 model", and "183 data uploads". Below this are buttons for "New narrative" and "New team". Further down are sections for "Team management" (listing "Metal reduction project" and "Chemotaxis project"), "Data management" (with an "upload" button), and "Narrative management" (listing "Chemotaxis Study"). A "blog" button is at the bottom left.

The central "Narrative" area shows a search result for "Sensory_genes: gs_aeb123456" and "Flagellar_genes: gs_aeb123457". It contains a series of narrative steps:

- [in4] 9:12:12am 11/16/2011: (* Bobtheguy says I missed one. I have looked and by eye I agree *) Add(Sensory_genes, gi0123421)->Sensory_genes
- [in5] 10:00:28pm 11/16/2011: I need to figure out in which conditions these genes are expressed. First I am going to aggregate my two sets of genes (I separate them for differential analysis later, then query for all gene expression data concerning them. Hmm... what's that function again?)
- [in6] 10:04:17pm 11/16/2011: Merge(Sensory_genes,Flagellar_genes)->GetExpress(geneids::stdin)->Add(Expression_data, stdin)
- [in7] 10:27:43pm 11/16/2011: ClusterMe(Expression_data)->PickCluster()

Below the [in7] step is a heatmap visualization showing gene expression data with a dendrogram on the left. A blue box highlights a cluster of genes, and an arrow points to the command: `-> Add(High_Expression, stdin)`.

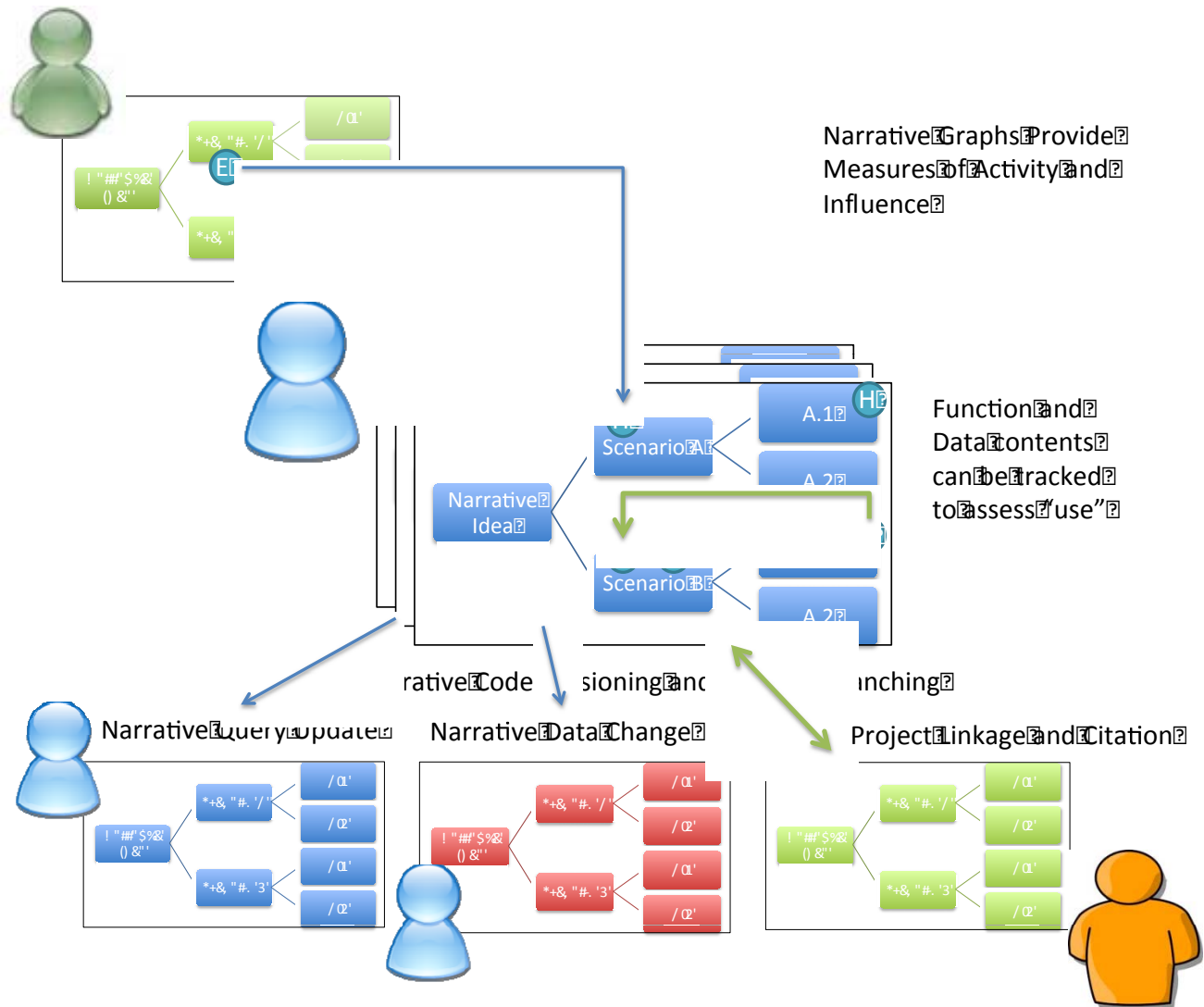
On the right sidebar, there are tabs for "data" and "function", a "search" button, and a section titled "C: ClusterMe Clustering" with an ellipsis below it. At the bottom right, a "Processing" status is shown.



KBASE
predictive biology

DOE Systems Biology Knowledgebase

Concept: Interactive community knowledge



KBase will allow users to...

- Keep pace with the rate of data generation.
- Use interrelated data of different types.
- Develop analytical tools.
- Perform rapid comparisons across multiple genomes.
- Allow free and open access to KBase data, models and simulations.
- Grow partners, developers, data generators, or users.
- Coordinate with other researchers to target what experimental work is needed, and to share results.
- Learn how to use KBase features of models and move into predictive systems biology.



KBASE
predictive biology

DOE Systems Biology Knowledgebase

BUILDING THE KBASE COMMUNITY

The KBase Team:



Collaborators:

- Cold Spring Harbor Laboratory
- University of California, Davis
- Hope College in Michigan
- University of Illinois at Urbana-Champaign
- Yale University
- The Joint Genome Institute
- InterBRC Knowledgebase
- Several University Knowledgebase Projects

→ New Partners and Collaborators from Users and Stakeholders

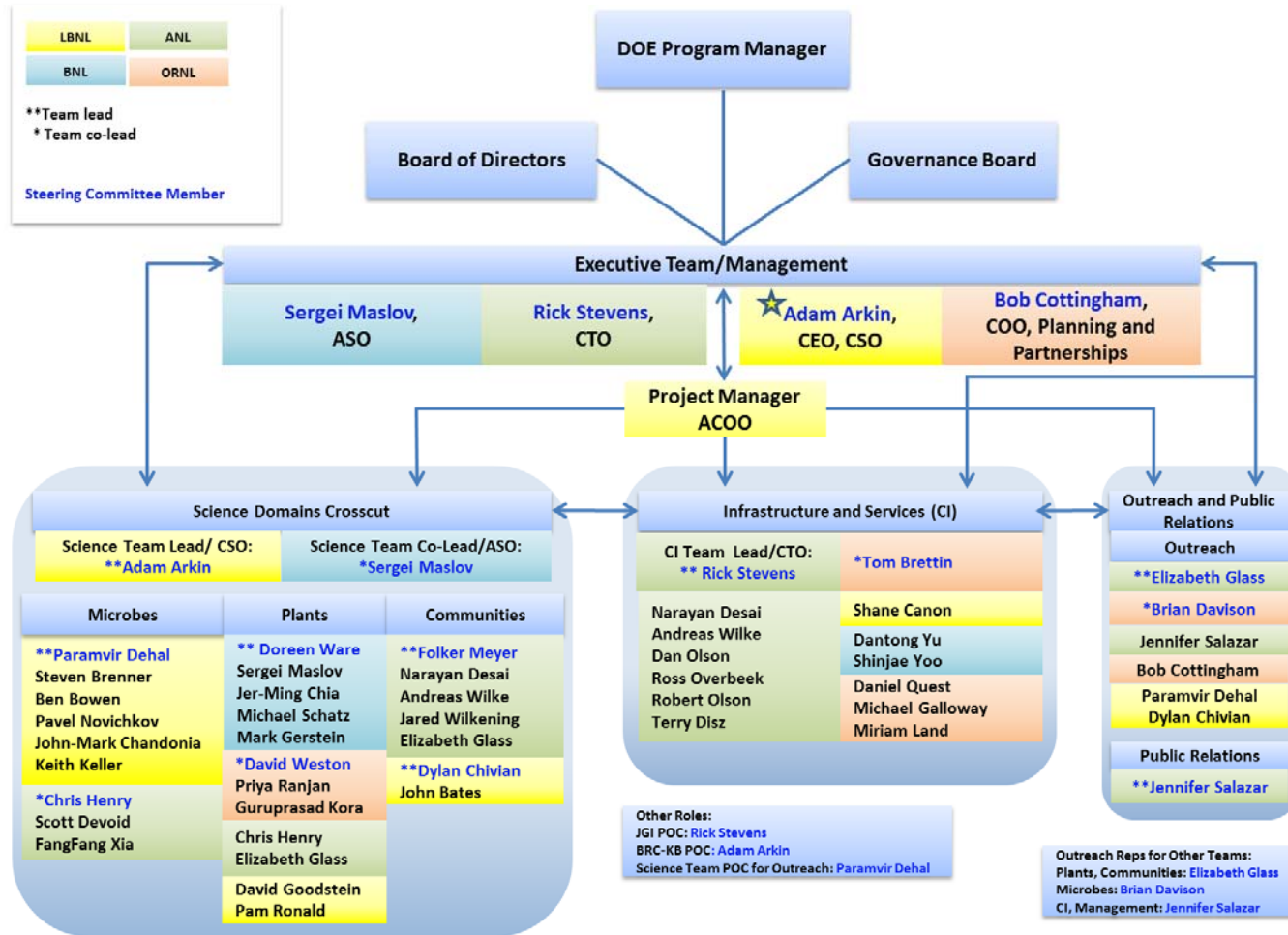
Funding: This work is supported by the U.S. Department of Energy, Office of Biological and Environmental Research under Contract DE-AC02-06CH11357.



KBASE
predictive biology

There is an org-chart but...

DOE Systems Biology Knowledgebase



Formal project management under 413-lite



DOE Systems Biology Knowledgebase

We are nucleators.

The Kbase core team is creating the initial infrastructure to:

- Allow effective and easy data federation and centralization.
- Lower the bar to adding and accessing new algorithms and analyses
- Enable nimble use of of advanced DOE network and compute resources
- Create a community framework for development and biological knowledge sharing
- Enable *their own research and yours!*

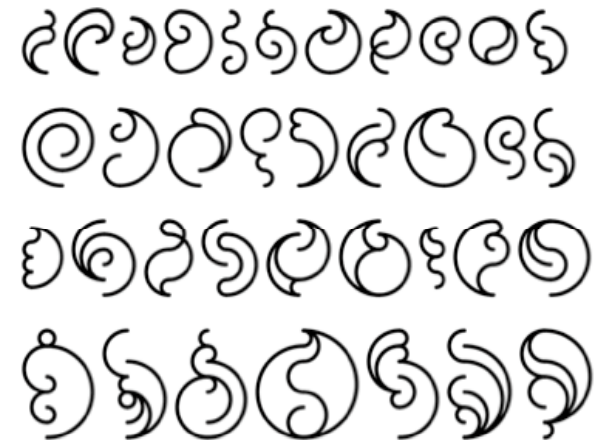
Everyone should be a contributor!

KBASE:

- A. Professional Computational Biologists
- B. Data generators and basic analysts
- C. Knowledge Seekers
- D. Knowledge Generators

Therefore we aim to:

- Create a powerful framework for programmatic access to data and functions of Kbase. (Users A,B)
 - Ultimately provide stubs for use in PERL, PYTHON, R, MATLAB, Galaxy, etc.
- Create a set of packaged “Widgets” that make placement and recognizable display of Kbase “functions” on web pages (or within perhaps other apps), easy and identifiable. (Users B)
- Create a “simplified” portal for search and aggregation of data for data consumers and Knowledge Seekers. (Users C,D)
- *Create a innovative platform for knowledge creation, evolution and sharing.*



instances of “minimum inventory/maximum diversity” systems, a term coined by Peter Pearce in his book, *Structure in Nature Is a Strategy for Design* (MIT Press, 1978).

We will build a diverse community of users by...

- Creating a unique a powerful resource for people to share new analyses and data and the use of these to produce new science.
- Advertising and informing via:
 - Online communication
 - Conference workshops
 - Press releases
- Interacting with the community and fostering relationships
- Collecting feedback from the community to ensure we are meeting their needs
- Measuring our impact
- Participating in standards groups



Visit us at kbase.science.energy.gov

Contact us at outreach@kbase.us



KBASE
predictive biology

DOE Systems Biology Knowledgebase

WHEN WILL WE DO IT?



DOE Systems Biology Knowledgebase

KBase Release/Build Timeline

- Feb 2012 – Development release (internal target)
 - debug release engineering, prototype deployments, initial data models and data loads, non-unified API, performance testing, architecture refinement
- May 2012 – Alpha release (internal target, limited invited testing)
 - draft tutorials, v0.0 database loads, draft API (performance and ubiquity unified prototypes), draft UI library, domain workflow drafts, cloud and cluster services
- Aug 2012 – Beta Release (early adopter beta testing)
 - workflow function complete, API refinement, v1.0 database loads, prototype plugin interfaces, prototype galaxy support, performance debugging
- Nov 2012 – Production Release Candidate (public beta testing)
 - draft website, draft documentation, full functionality API, draft UI v1.0, database loads v2.0, significant number of external beta test users
- Feb 2013 – KBase Production Release
 - public website, unified API, initial production UI, database loads v3.0 (microbes, community, plant databases), production demonstration workflows, replication and fail over services,

Outreach Timeline?



For more information, and to follow our progress, please visit

<http://kbase.science.energy.gov/>

<http://outreach.kbase.us>



KBASE
predictive biology

DOE Systems Biology Knowledgebase

Thank you!

Visit us at kbase.science.energy.gov

Contact us at outreach@kbase.us