DOE Systems Biology Knowledgebase

INTEGRATION and
MODERN *for*
PREDICTIVE BIOLOGY

The KBase Platform for Dissemination of Tools and Analysis of Microbes, Plants and Their Communities: Examples from the users

Office of Biological and Environmental Research

KBase Team

KBase is a multi-institutional collaboration with participation from these laboratories and universities

Lawrence Berkeley National Laboratory

Lead institution

Argonne National Laboratory

Brookhaven National Laboratory

Oak Ridge National Laboratory

Cold Spring Harbor Laboratory

University of Tennessee

Chris Henry, ANL
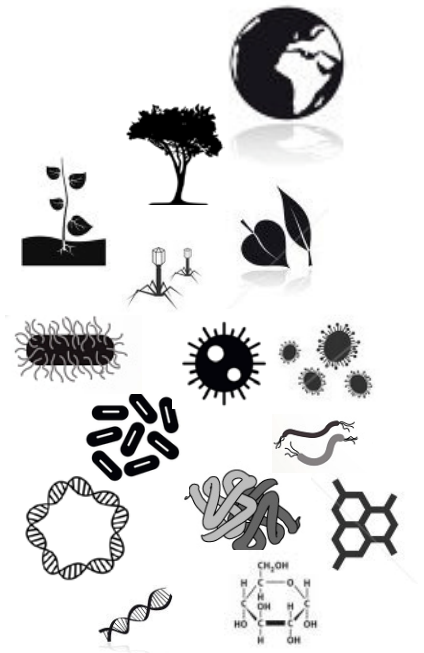
Bob Cottingham, ORNL

What is KBase?

KBase is a knowledge creation and discovery environment

KBase is a knowledge creation and discovery environment

Seamless integration of multiply-sourced data and tools in a platform supporting learning, reproducibility and collaboration
Powerful scientific framework for predicting function of biomolecules, microorganisms, plants and their communities.

Data about...

❖ KBase is designed to accelerate research about , microorganisms, plants and their communities in environmental context with emphasis on DOE goals.
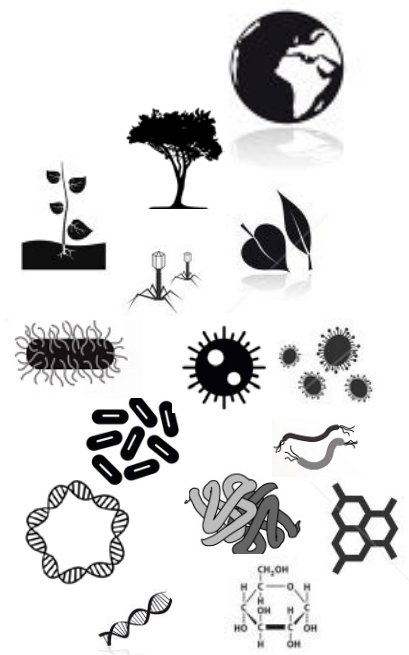
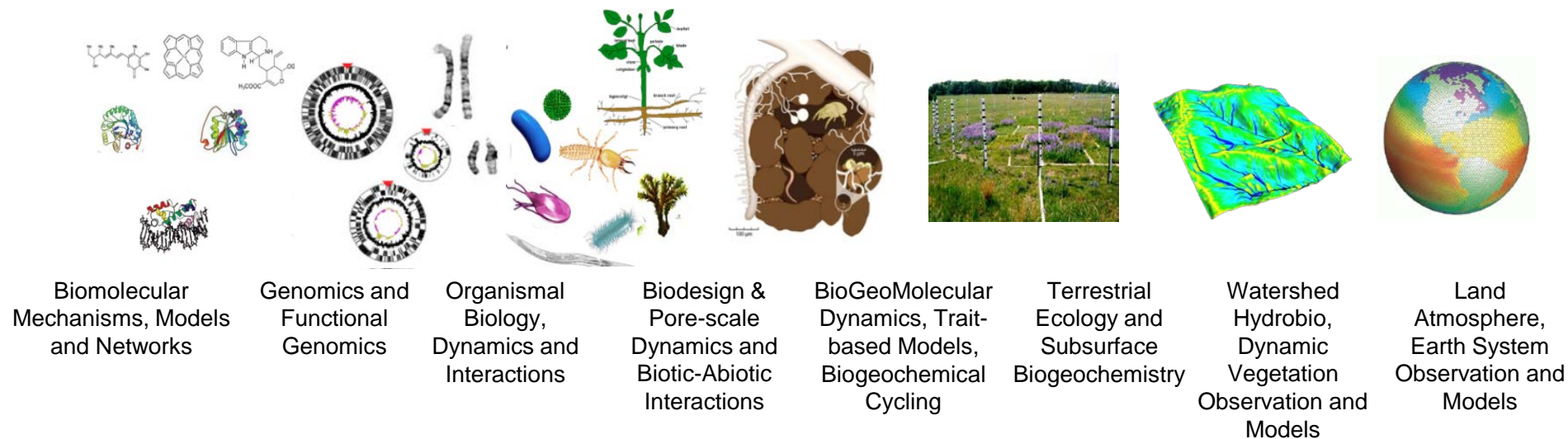# KBase is a knowledge creation and discovery environment

Seamless integration of multiply-sourced data and tools in a platform supporting learning, reproducibility and collaboration
Powerful scientific framework for predicting function of biomolecules, microorganisms, plants and their communities.

❖ KBase is designed to accelerate research about microorganisms, plants and their communities in environmental context with emphasis on DOE goals.
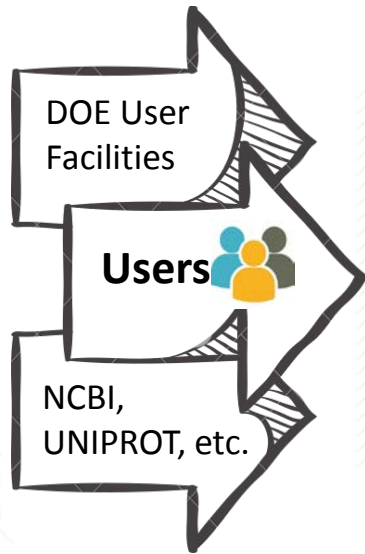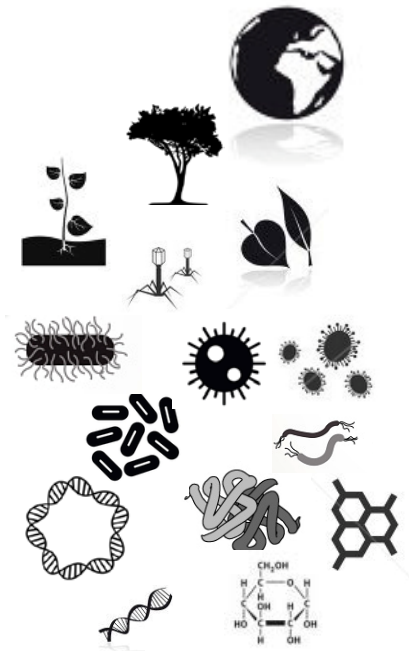
Data about...

KBase Scope of Operations

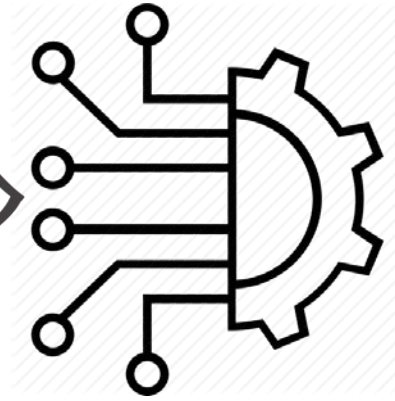| Biomolecular Mechanisms, Models and Networks | Genomics and Functional Genomics | Organismal Biology, Dynamics and Interactions | Biodesign & Pore-scale Dynamics and Biotic-Abiotic Interactions | BioGeoMolecular Dynamics, Trait-based Models, Biogeochemical Cycling | Terrestrial Ecology and Subsurface Biogeochemistry | Watershed Hydrobio, Dynamic Vegetation Observation and Models | Land Atmosphere, Earth System Observation and Models |

KBase is a knowledge creation and discovery environment

Seamless integration of multiply-sourced data and tools in a platform supporting learning, reproducibility and collaboration
Powerful scientific framework for predicting function of biomolecules, microorganisms, plants and their communities.
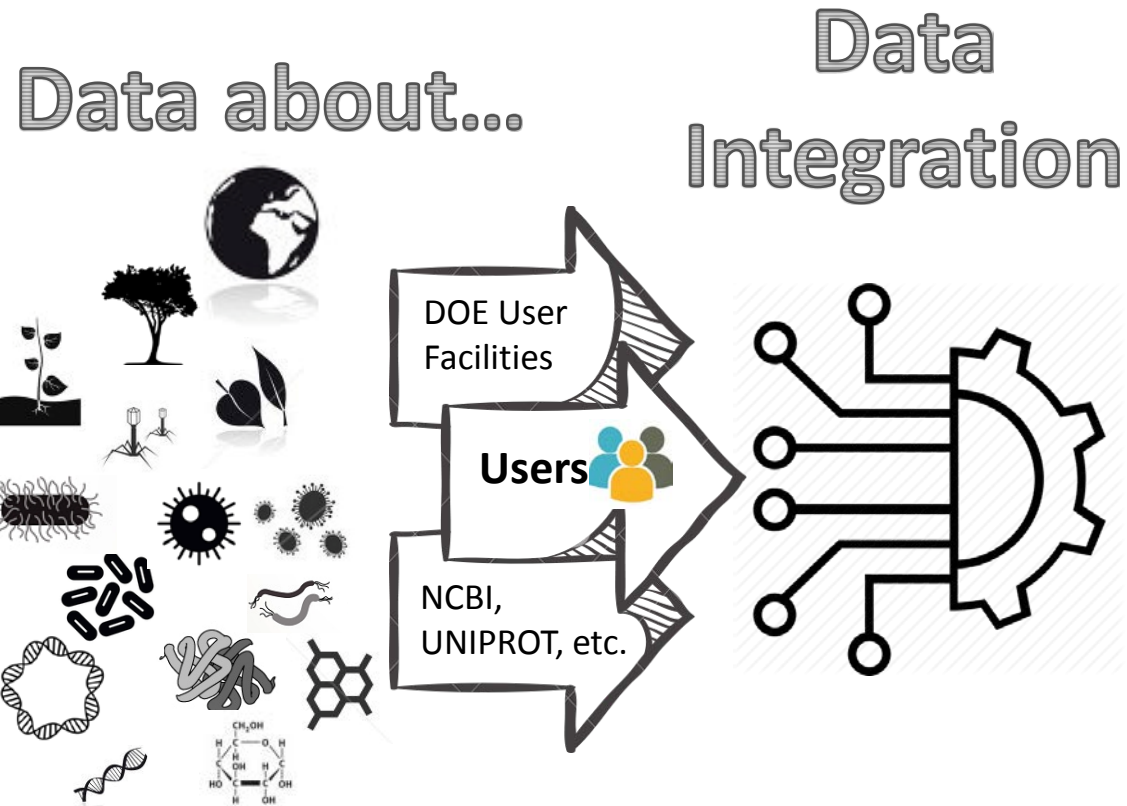
Data about...
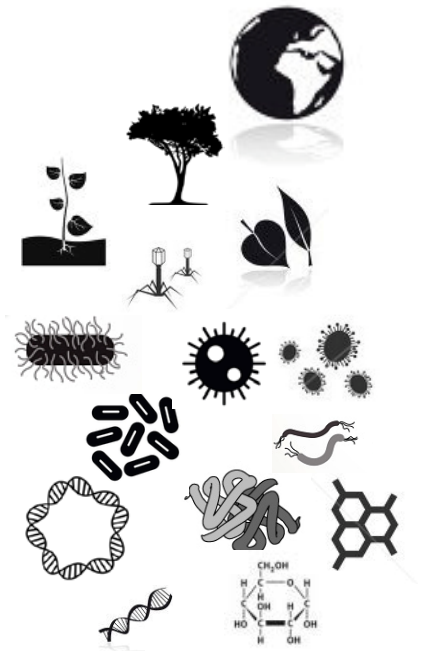
❖ KBase is designed to accelerate research about microorganisms, plants and their communities in environmental context with emphasis on DOE goals.

- Understand the biological complexity of plant and microbial metabolism and interfaces across scales spanning molecules to ecosystems.

- Scalable data processing, data analysis, machine learning, discrete algorithms, and multiscale multiphysical simulation are crucial for advancement of biological and environmental systems science.

- Innovations in representation, search, and visualization of large-scale, heterogeneous, ontologically rich primary and derived biological and contextual data (e.g., abiotic environmental information) are crucial for input to and validation of these methods.

- New architectures, data transport protocols, software libraries, and languages are necessary to create a platform for community tool development and use supporting interactive and seamless interoperation of both mid- and large-scale cluster resources and enterprise-class computing environments.

KBase is a knowledge creation and discovery environment

Seamless integration of multiply-sourced data and tools in a platform supporting learning, reproducibility and collaboration
Powerful scientific framework for predicting function of biomolecules, microorganisms, plants and their communities.

KBase is a knowledge creation and discovery environment

Seamless integration of multiply-sourced data and tools in a platform supporting learning, reproducibility and collaboration
Powerful scientific framework for predicting function of biomolecules, microorganisms, plants and their communities.

❖ KBase is designed to accelerate research through:

❖ Lowering the barrier to integration of diverse data from multiple sources
❖ Sharing data and analyses in a "persistent", transparent, reusable, computable, and reproducible format.

KBase is a knowledge creation and discovery environment

Seamless integration of multiply-sourced data and tools in a platform supporting learning, reproducibility and collaboration
Powerful scientific framework for predicting function of biomolecules, microorganisms, plants and their communities.

Data about...

Data Integration

DOE User Facilities

Users

NCBI, UNIPROT, etc.
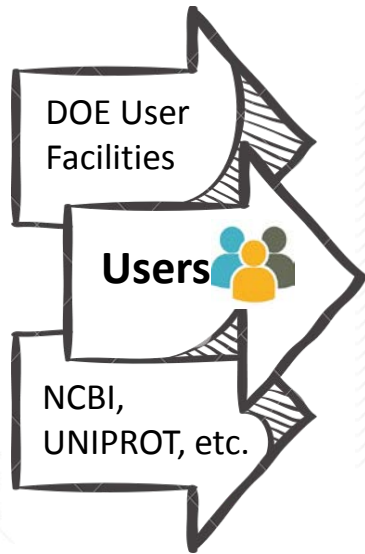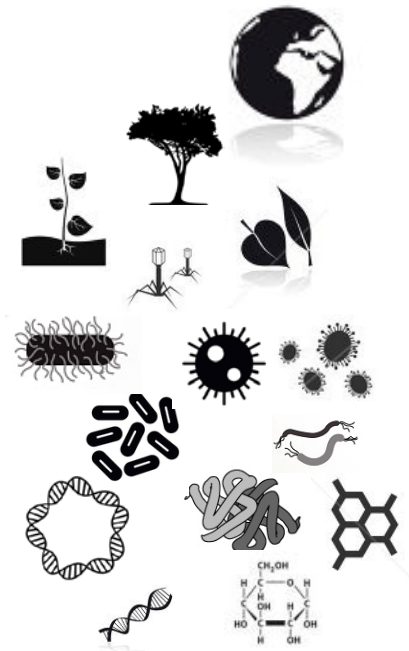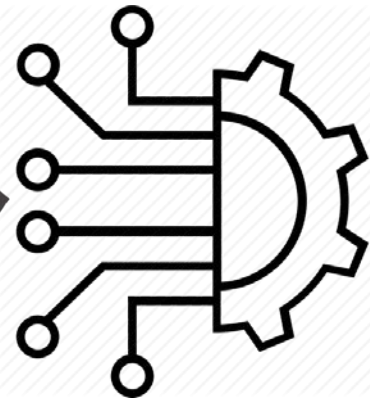
❖ KBase is designed to accelerate research through:

   ❖ Lowering the barrier to integration of diverse data from multiple sources
   ❖ Sharing data and analyses in a "persistent", transparent, reusable, computable, and reproducible format.

Science AAAS

Journals unite for reproducibility

Marcia McNutt
+ See all authors and affiliations

Science 07 Nov 2014
Vol. 345, Issue 6210, pp. 679
DOI: 10.1126/science.aaa1724

Article | Info & Metrics | eLetters | 📄 PDF

Reproducibility, rigor, transparency, and independent verification are cornerstones of the scientific method. Of course, just because a result is reproducible does not necessarily make it right, and just because it is not reproducible does not necessarily make it wrong. A transparent and rigorous approach, however, can almost always shine a light on issues of reproducibility. This light ensures that science moves forward, through independent verifications as well as the course corrections that come from refutations and the objective examination of the resulting data.

It was with the goal of strengthening such approaches in the biomedical sciences that a group of editors representing over 30 major journals, representatives from funding agencies, and scientific leaders assembled at the AAAS headquarters in June of 2014 to discuss principles and guidelines for preclinical biomedical research. The gathering was convened by the U.S. National Institutes of Health, *Nature*,* and *Science*.

The discussion ranged from what journals were already doing to address reproducibility and the

"...scientific journals are standing together in their conviction that reproducibility and transparency are important..."

a natureresearch journal

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

KBase is a knowledge creation and discovery environment

Data Integration

Data about...
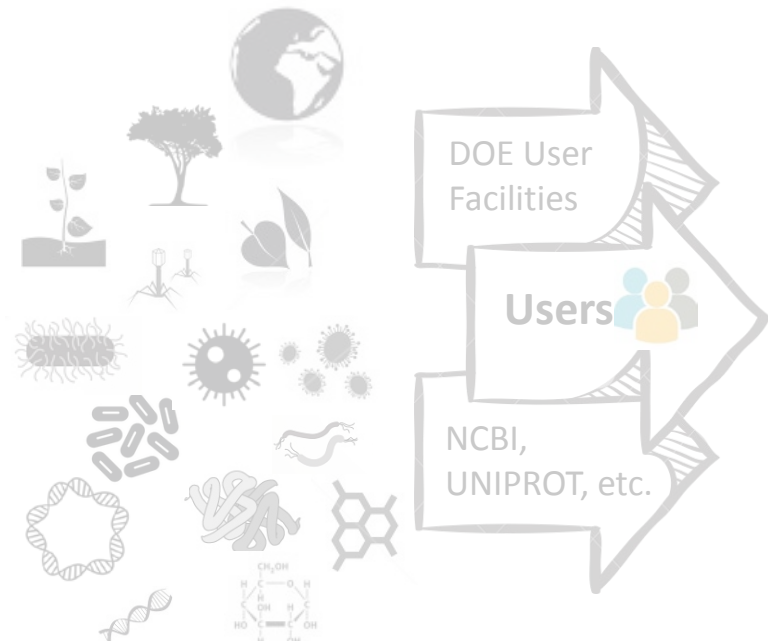
Data from all your Narratives

Data from collaborators

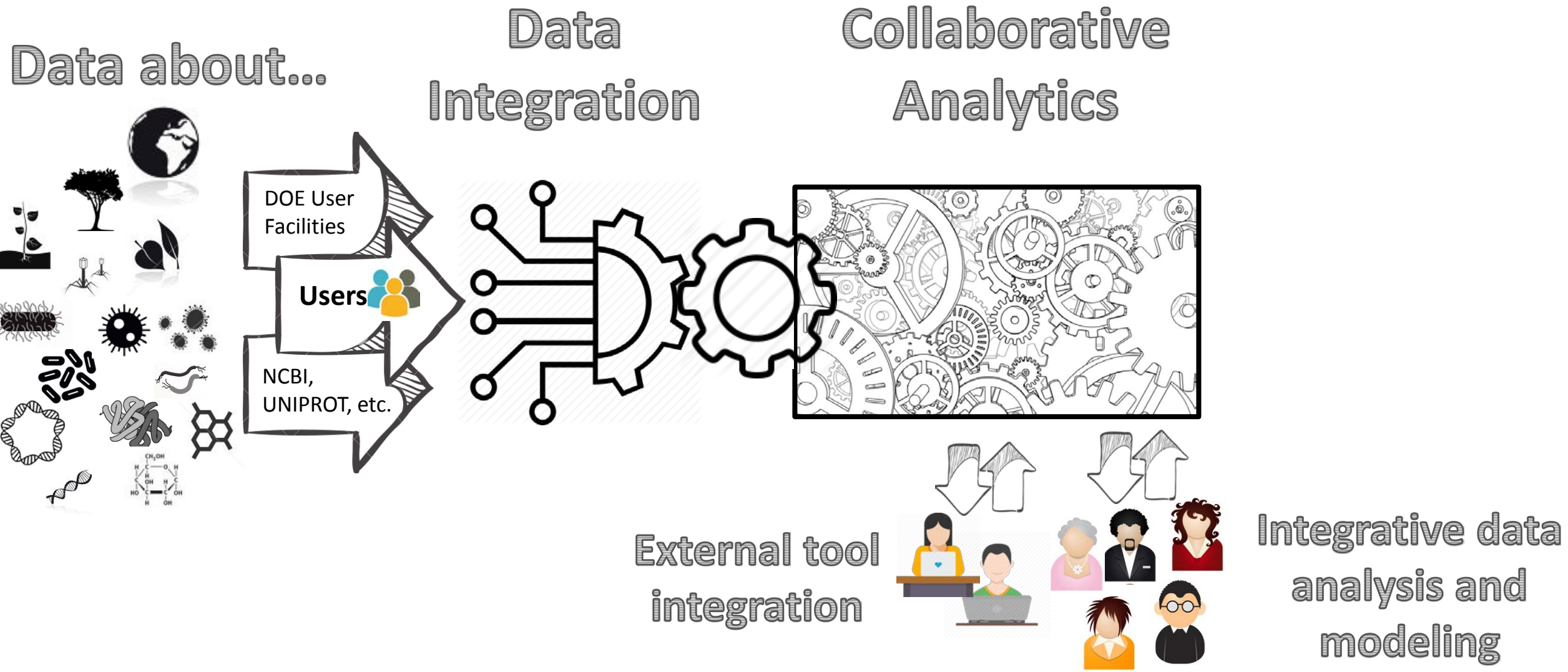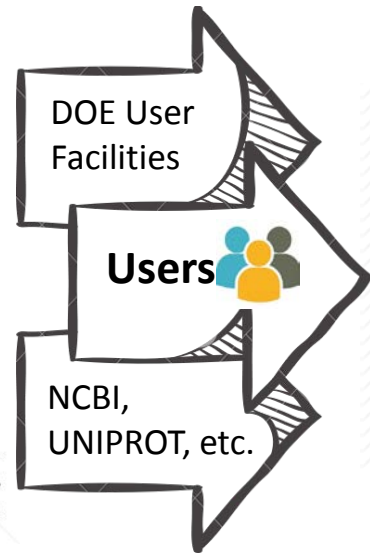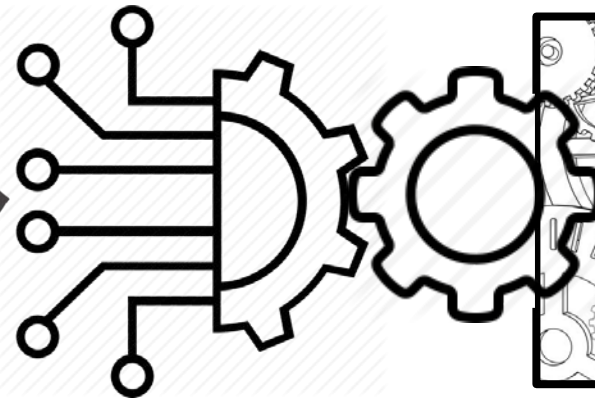Public data from KBase's reference collection

Example data

Import your own data

| Analyze | Narratives | Jobs | My Data | Shared With Me | Public | Example | Import |

DATA

Search data...   All types...   All narratives...

This Narrative has no data yet.

Add Data

Click here to access the Data Browser and begin loading data into your Narrative

Currently Uploadable Types

- Short reads
- Contigs
- Genomes
- Transcripts
- Expression
- FBA models
- Media
- Phenotype sets

Ptrichocarpa.JGI3.0_draftModel *v1*
FBAModel
Metabolic modeling of Populus trichocarpa
4 hours ago by hlh35

PlantHeterotrophicMedia *v3*
Media
Metabolic modeling of Populus trichocarpa
4 hours ago by hlh35

Populus_trichocarpa *v7*
Genome: Populus trichocarpa
Metabolic modeling of Populus trichocarpa
4 hours ago by hlh35

Ptrichocarpa.JGI3.0_gapfilledModel *v1*
FBAModel
Metabolic modeling of Populus trichocarpa
4 hours ago by hlh35

Ptrichocarpa.JGI3.0_annotate *v1*
Genome: Populus trichocarpa
Metabolic modeling of Populus trichocarpa
4 hours ago by hlh35

Escherichia_coli_O157_H7_str._EC4045 *v1*
Genome
Clostridium Species Tree
42 hours ago by hlh35

Clostridium_thermocellum_DSM_4150 *v1*
Genome
Insert_Genome_Into_Species_Tree
3 days ago by hlh35

Clostridium_thermocellum_DSM_1313 *v1*

DOE User Facilities

Users

NCBI, UNIPROT, etc.

APPS & METHODS

Assemble and Annotate Microbial Genome
v0.1.0

Build and Normalize Metagenomic Functional Abundance Data
v0.1.0

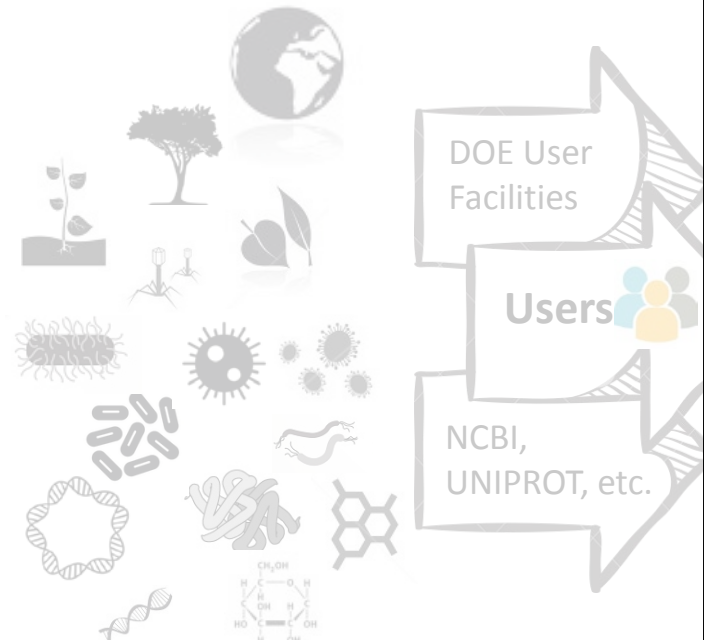Build and Normalize Metagenomic Taxonomic Abundance Data
v0.1.0

Close

KBase is a knowledge creation and discovery environment

Seamless integration of multiply-sourced data and tools in a platform supporting learning, reproducibility and collaboration
Powerful scientific framework for predicting function of biomolecules, microorganisms, plants and their communities.

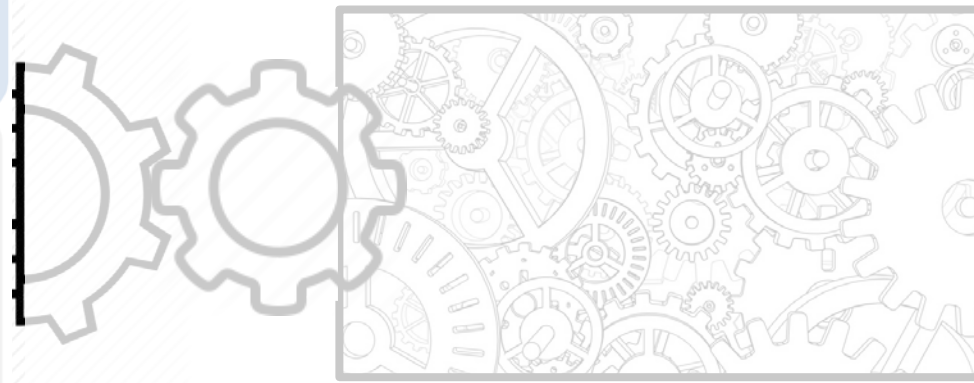# KBase is a knowledge creation and discovery environment

Seamless integration of multiply-sourced data and tools in a platform supporting learning, reproducibility and collaboration
Powerful scientific framework for predicting function of biomolecules, microorganisms, plants and their communities.

Data about...

Data Integration

Collaborative Analytics

DOE User Facilities

Users

NCBI, UNIPROT, etc.

External tool integration

Integrative data analysis and modeling

❖ KBase is designed to accelerate research through:

❖ Access to a user-extensible powerful suite of tools driving towards modeling biological function.

❖ Access to enterprise-level computational resources.

❖ A collaborative notebook based system for organizing data, analyses and conclusions.

KBase is a knowledge creation and discovery environment

Seamless integration of multiply-sourced data and tools in a platform supporting learning, reproducibility and collaboration
Powerful scientific framework for predicting function of biomolecules, microorganisms, plants and their communities.
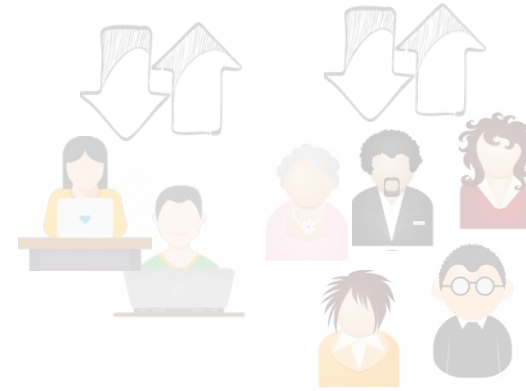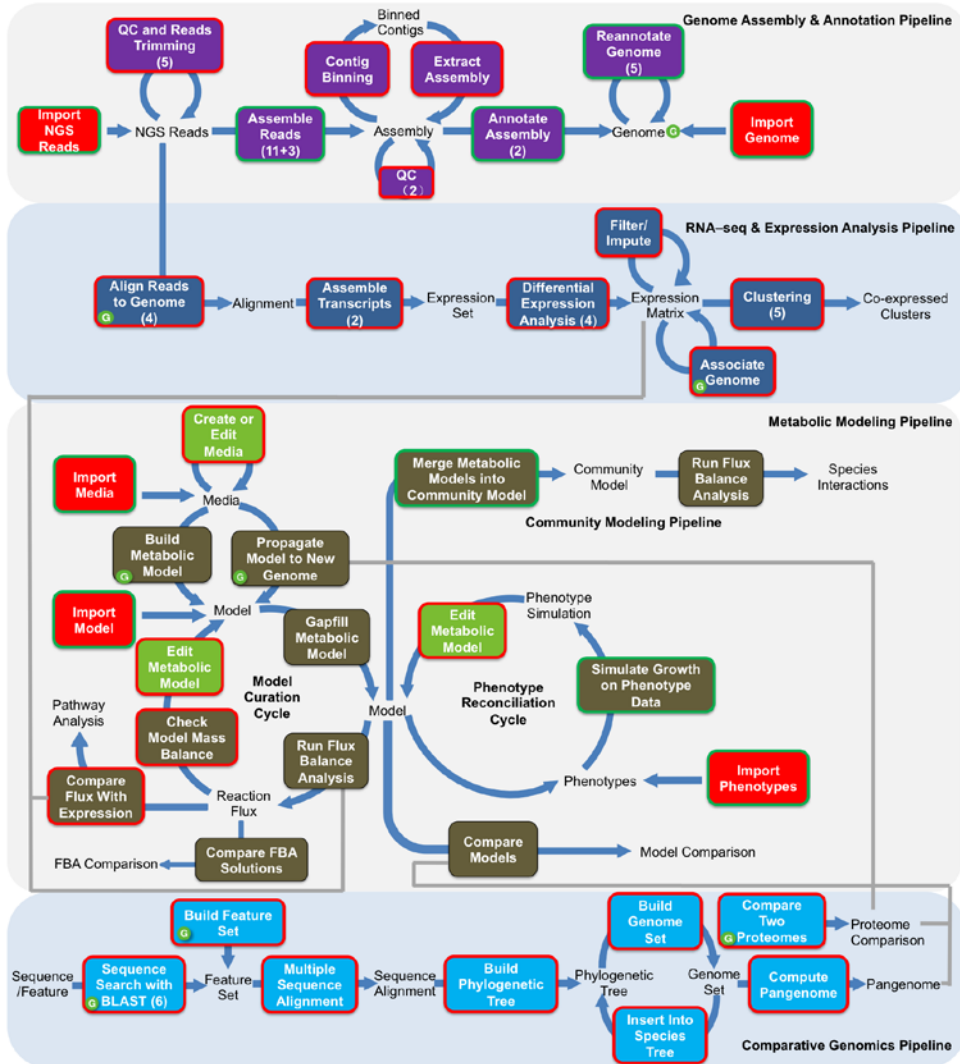

The Narrative Interface

An interactive, dynamic, and persistent document created by users that promotes open, reproducible, and collaborative science

KBase is a knowledge creation and discovery environment

Seamless integration of multiply-sourced data and tools in a platform supporting learning, reproducibility and collaboration
Powerful scientific framework for predicting function of biomolecules, microorganisms, plants and their communities.

KBase is a knowledge creation and discovery environment

Seamless integration of multiply-sourced data and tools in a platform supporting learning, reproducibility and collaboration
Powerful scientific framework for predicting function of biomolecules, microorganisms, plants and their communities.

KBase is a knowledge creation and discovery environment

Seamless integration of multiply-sourced data and tools in a platform supporting learning, reproducibility and collaboration
Powerful scientific framework for predicting function of biomolecules, microorganisms, plants and their communities.

o Go from reads to metabolic models of of plants, microbes and their community in a matter of a few hours.

o Current tools(**That you can add to**) include:

✓ Bulk upload, download and execution
✓ reads management
✓ sequence quality assessment and control
✓ Comparative microbial assembly and annotation
✓ Basic metagenomic binning, assembly and annotation
✓ RNA-SEQ analysis for proks and euks
✓ Comparative genomics
✓ Metabolic modeling for isolates and communities.

External tool integration

KBase is a knowledge creation and discovery environment

❖ Integration of primary and derived products into a data model that supports human and machine learning analysis of all shared and published results across the system and automatic propagation of new results to biologically-related entities.

Data about...

DOE User Facilities

Users

NCBI, UNIPROT, etc.

Data Integration

Collaborative Analytics

Knowledge Propagation Engine

External tool integration

Integrative data analysis and modeling
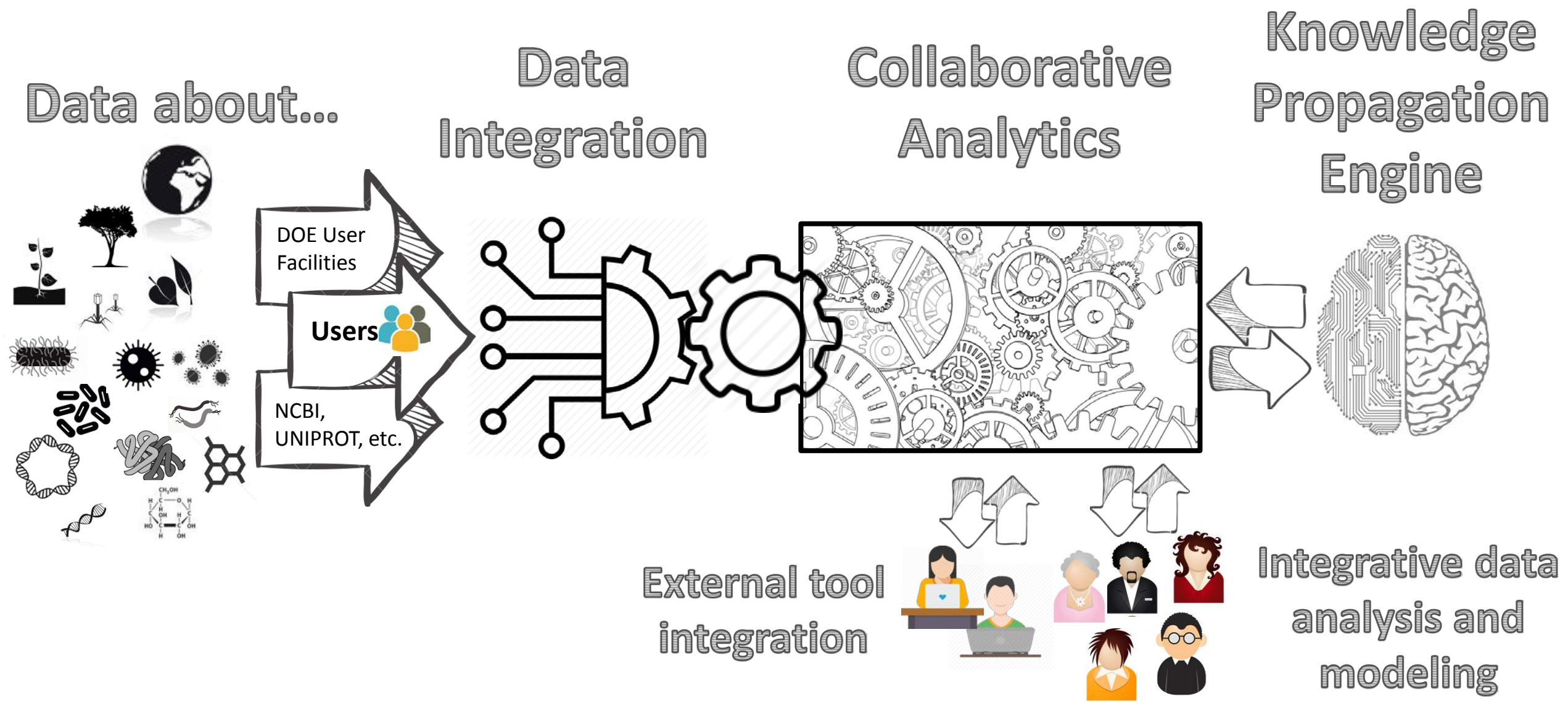
What is new in KBase since 2016?

Platform Improvements since 2016

Improved Ease of Use

- ✓ Streamlined sign-up and login process
- ✓ Fast and powerful search interface to public and user's private data in KBase
- ✓ Improved search and import of JGI public data
- ✓ Easily import large data and collections of data via Globus or drag-and-drop interface

- ✓ Narrative interface core built on Jupyter is kept current with upstream
- ✓ All previous apps converted to the SDK to improve stability
- ✓ "Docker-ized" the platform to simplify updates, improve reliability, and reduce hardware requirements
- ✓ Added initial support for HPC and parallel execution

Increased Stability and Scalability

- ✓ Introduced the KBase SDK to dramatically lower the barrier to adding new apps
- ✓ Used the SDK to develop 100+ new Apps and fill in gaps in reads management, isolate analysis, microbiome analysis, expression analysis, and modeling
- ✓ Revamped reference data, which now includes all of RefSeq and Phytozome and portions of MycoCosm

Interface to search reference, public and shared user data

- New high-speed interface to search all public data and data shared by users in KBase
- Currently limited to top-level objects (genomes, models, media, Narratives)
- Will soon expand to lower-level objects (genes, proteins, functions, reactions, compounds)
- Valuable for data/collaboration discovery (e.g., search for a species name)

Dashboard

Catalog

Search *beta*

Jobs

Account

Search: JGI | KBase - User Data, Reference Data, Features

ENIGMA   ✕ ↺ ?

Copy Selected...   View Detail ⬍

| User Data **14** | Reference Data **0** |

|◀ ◀ ▶ ▶| Page 1 of 1   Found 14 Narratives   Access: ☑ Private ☑ Public

Type	Name		Modified

📄 ENIGMA Mercury Mystery Plasmids   John-Marc Chandonia   04/20/2018
  📄 Narrative   Narrative.1523576735739   04/20/2018

Matches		Detail	
Title	**ENIGMA** Mercury Mystery Plasmids	Title	ENIGMA Mercury Mystery Plasmids

📄 ENIGMA FW-306 Core-Sediment Campaign   John-Marc Chandonia   02/21/2018
  📄 Narrative   Narrative.1519257520913   02/21/2018

Matches		Detail	
Source	**This Narrative is part of a series of narratives from the **ENIGMA** project**; to see all of them	Title	ENIGMA FW-306 Core-Sediment Campaign
Title	**ENIGMA** FW-306 Core-Sediment Campaign		

📄 ENIGMA project master narrative   John-Marc Chandonia   02/21/2018
  📄 Narrative   Narrative.1516747247669   02/21/2018

Matches		Detail	
Source	# **ENIGMA** Project ## This Narrative links to other **ENIGMA** narratives in KBase, to help project in the cells below. ### More info about the **ENIGMA** project can be found on our website: https://enigma.lbl.gov/ ### Most **ENIGMA** data that can't currently be stored in KBase are stored in our Google	Title	ENIGMA project master narrative
Title	**ENIGMA** project master narrative		

📄 ENIGMA metagenomic sequence data   John-Marc Chandonia   03/09/2018
  📄 Narrative   Narrative.1513798166246   03/09/2018

Matches		Detail	
Source	**This Narrative is part of a series of narratives from the **ENIGMA** project**; to see all of them## **ENIGMA** Metagenomic Sequences This Narrative contains metagenomic sequence data sequenced by	Title	ENIGMA metagenomic sequence data
Title	**ENIGMA** metagenomic sequence data		

ENIGMA Project

This Narrative links to other ENIGMA narratives in KBase, to help project members find data and analyses from across the project.

Other Narratives are described and linked in the cells below.

More info about the ENIGMA project can be found on our website: **https://enigma.lbl.gov/**

Most ENIGMA data that can't currently be stored in KBase are stored in our Google Drive data folder, here: **https://drive.google.com/drive/folders/0B62rJp3HQTPMbUdjOC1Nd3dkSWs**

Samples and Wells

This spreadsheet contains a sample request log for samples taken since the 100 Well Survey (see below for those data), along with records of which samples were taken. Look on the 2nd and 3rd tabs of the spreadsheet for sample records and measurements taken on each sample: https://docs.google.com/spreadsheets/d/1a_NC6vVY5Au6geVZj3Mn4eItdqpG4EppftA1uiyWkOg/edit#gid=1606165980

Metadata describing the wells (location, screen depths, etc) are in this folder: https://drive.google.com/drive/folders/0B18gfpPD5aW8b0prTnUxd2dBa0E

2017 Core Sample Pilot

Data and methods from the 2017 Core Sample Pilot Project are here: https://narrative.kbase.us/narrative/ws.26612.obj.1

100 Well Survey

Data from the 100 Well Survey and the resulting Smith et al mBio paper are here: https://narrative.kbase.us/narrative/ws.26835.obj.1

Isolates

A narrative with all the **sequenced** isolates is here: https://narrative.kbase.us/narrative/ws.24918.obj.1

Note that you can access these sequenced isolates as "Genome" objects in KBase: just go to "Add Data" and look under "shared with me" to copy any of these into a new narrative so you can explore the data

Fast Search and Import of JGI Data

➢ Co-developed with the JGI team using JAMO
➢ Currently ports raw reads (FASTQ) and assembled contigs (FASTA) – more datatypes coming
➢ Currently supports public data – the ability to search private data is coming

KBase Software Development Kit

A tool and framework for dynamically adding new Apps to KBase:

➢ All scientific tools available to KBase users run as Apps

➢ Apps can be updated and deployed directly by any developer

➢ App Catalog tracks and manages all versions of Apps

➢ Built-in Dev/Beta/Release lifecycle management

➢ Opens the door for 3rd party developers

➢ Improved data access and file handling

➢ Ability to call other KBase modules from your App

➢ HTML Reports with linked files

➢ Support for loading versioned Reference Data in Apps

➢ Web services built with the SDK
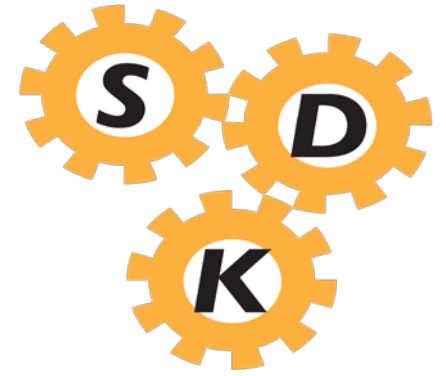
Register with KBase
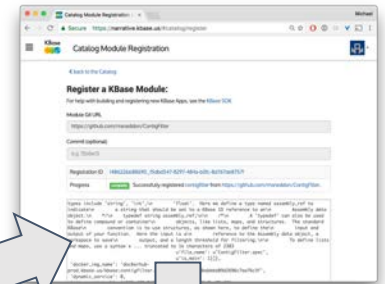
Push to GitHub
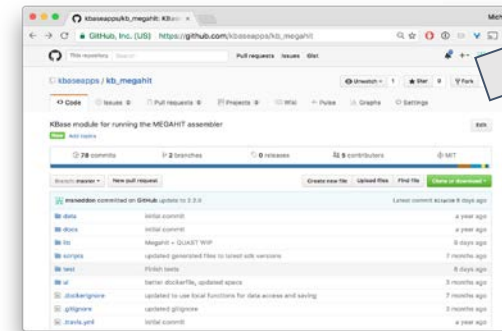
App is Available!
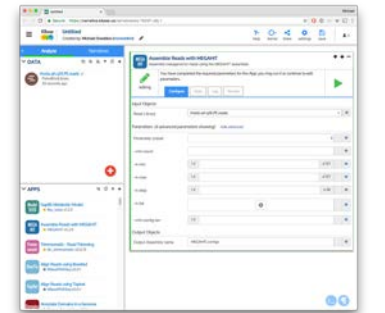
New workflows in microbiome and comparative genomics
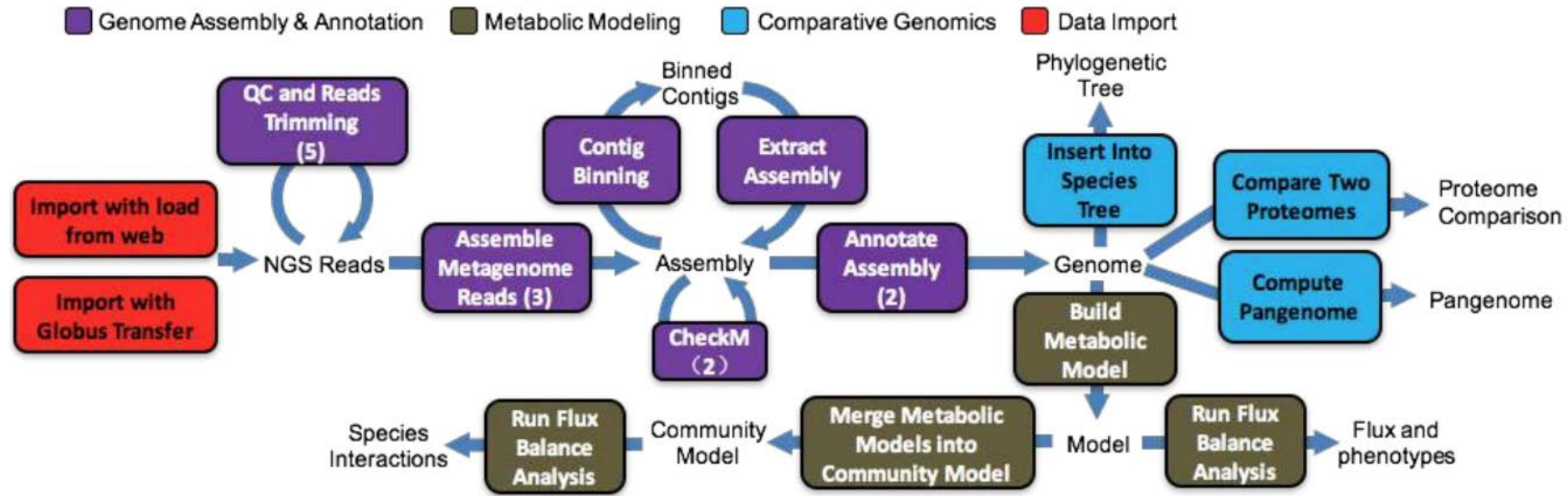
- Assemblers for metagenomics reads
- Binning of contigs into species
- Genome quality assessment

- Multiple algorithms for homology search
- Algorithms for multi-sequence alignment
- Tools for pangenome and phylogenetic analysis
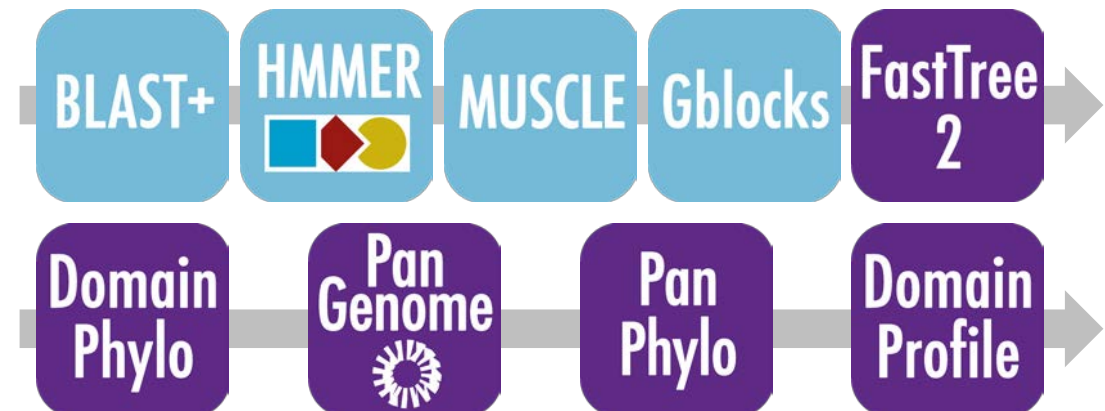
Better App through 3rd party dev and SDK

KBase FBA Model Import (Current default - internally developed)

SBML Tools Model Import (3rd party)

3rd party developer constructed a far better version of tool based on 5 years of thesis work on biochemistry integration

Import Report (current)

Import Report (3rd party)
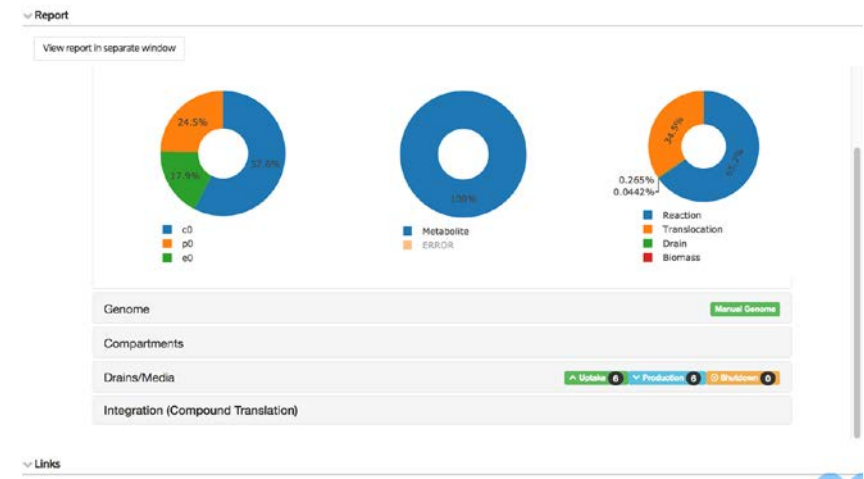
3rd party developer also generated far better visualizations of models

SDK enables external developers to wrap tools as KBase apps

BioSIMP: ML-based prediction of growth of new microbes
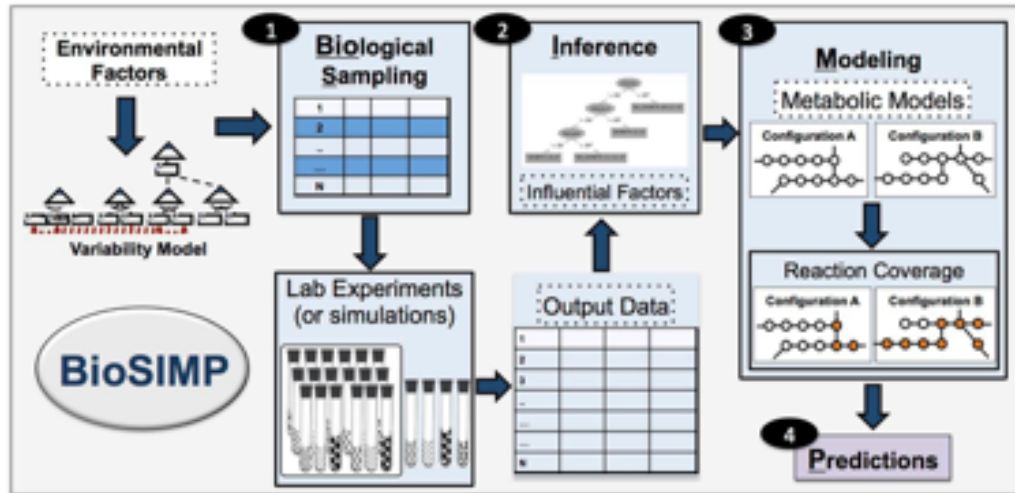
Mikaela Cashman & Jennie L. Catlett

OptStoic: Predict and design metabolic pathways
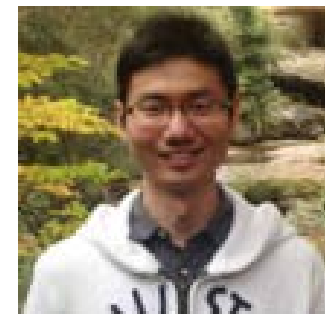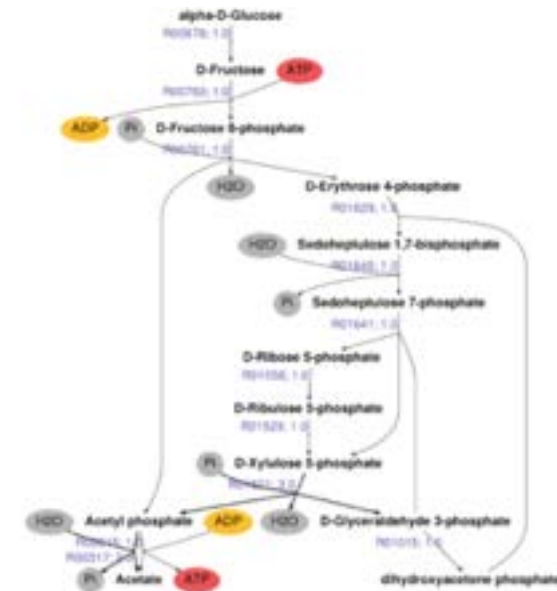
Lin Wang & Costas Maranas

SDK enables external developers to wrap tools as KBase apps

SBML Importer: Bulk import and integration of metabolic models

Filipe Liu &
Isabel Rocha
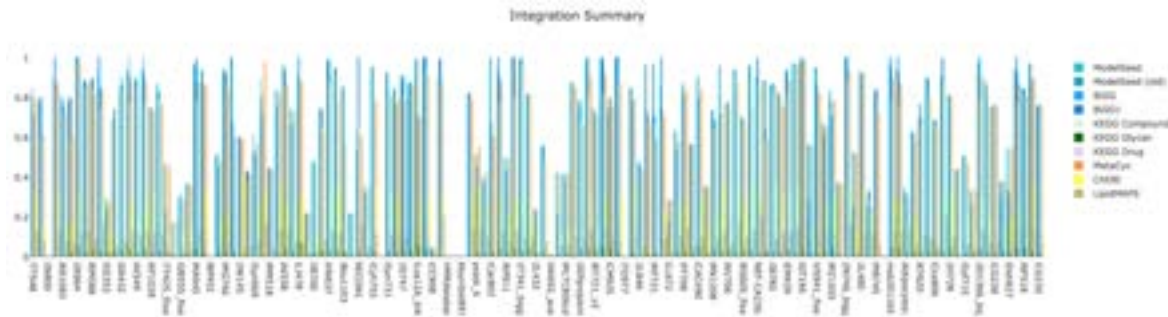
UNIVERSIDADE NOVA DE LISBOA

- Automatically selects the correct genome to associate with each model by searching the reference genomes for genes in the models

- Enables a curated integration of models with the KBase namespace for compounds, reactions, compartments, and genes

- Applied to import 118 published metabolic models into KBase

Flux Mutual Information Analysis
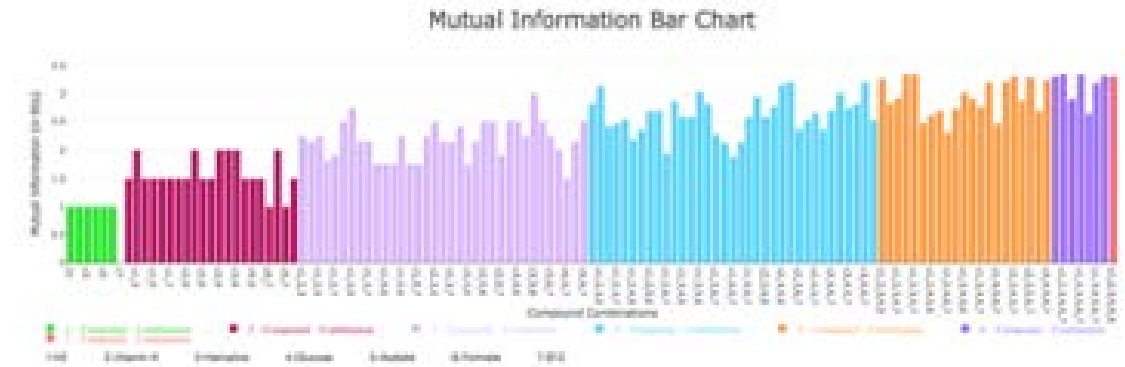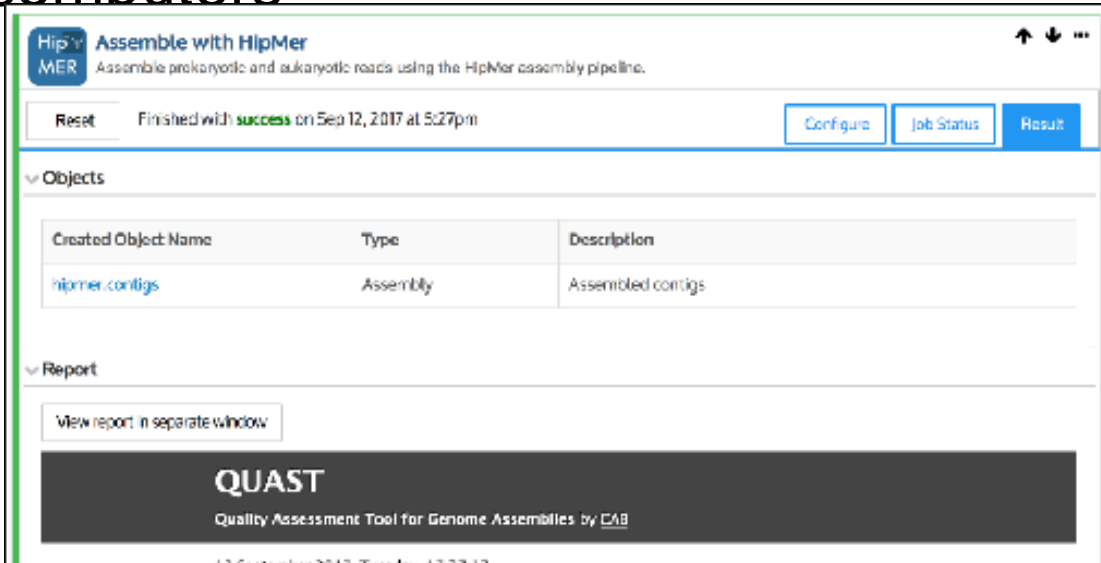
Zahmeeth Sakkaff &
Massimiliano Pierobon

UNIVERSITY OF Nebraska Lincoln

- Computes mutual information between reaction flux / cell interactome and input nutrient profile

- Aids experimentalists in identifying the combination of nutrients that reveal the greatest information about the cell by causing variations in pathways, by-products, and growth rate

Facilities are Using SDK to Add Tools

HipMer: assembly of large genomes and metagenomics on enterprise-class computers
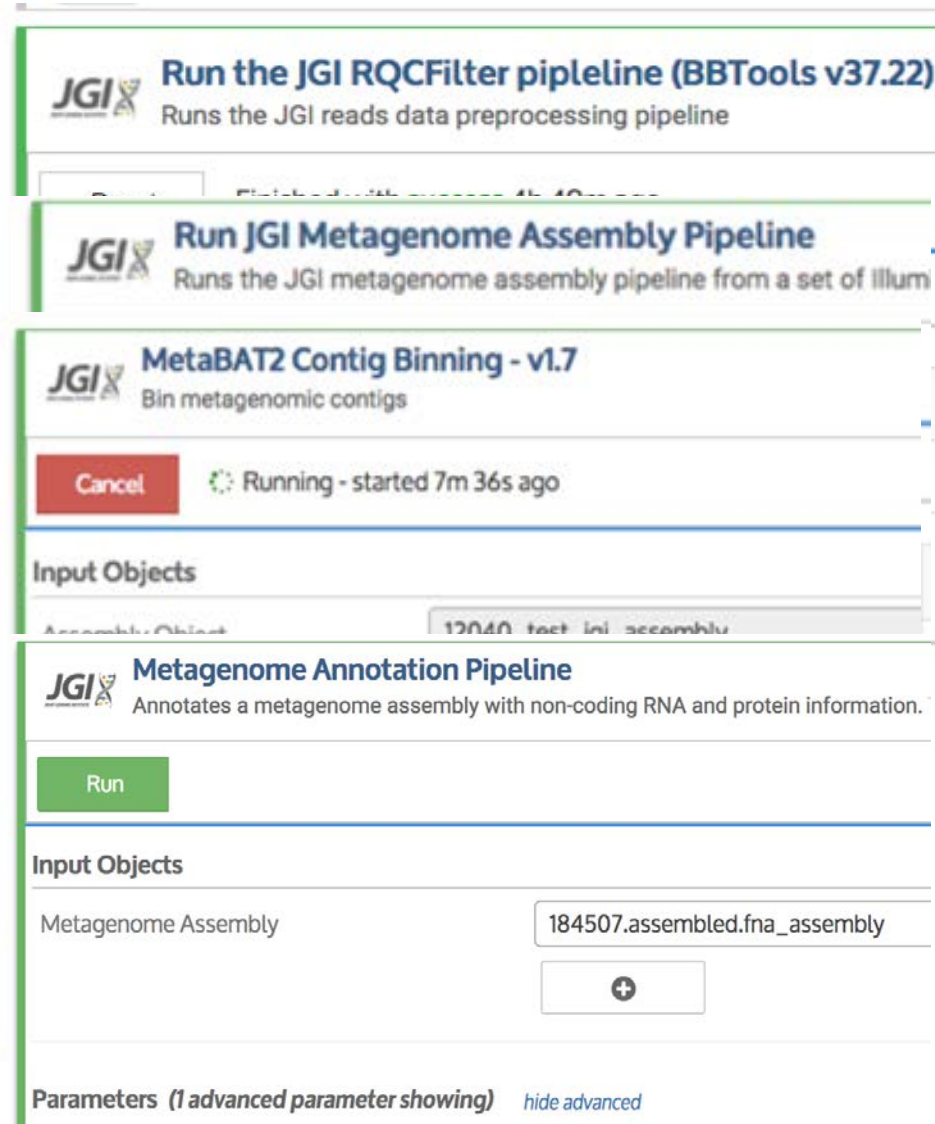
BBTools: core sequencing quality control from JGI

Kathy Yelick
CRD, NERSC, LBNL
EECS, UCB

Dan Rokhsar
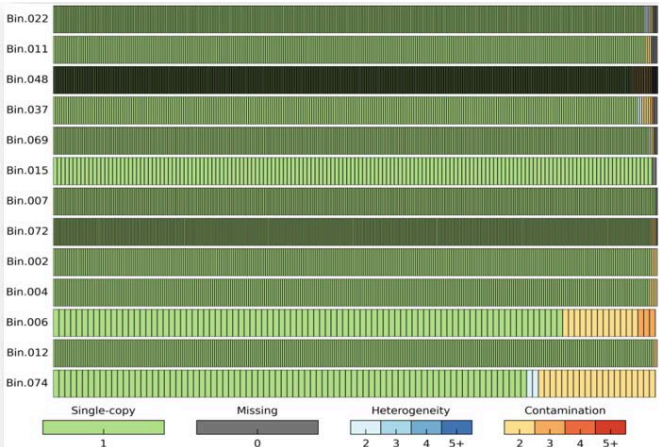EGSB, LBL
MCB, UCB

Jeff Froula
JGI, LBL

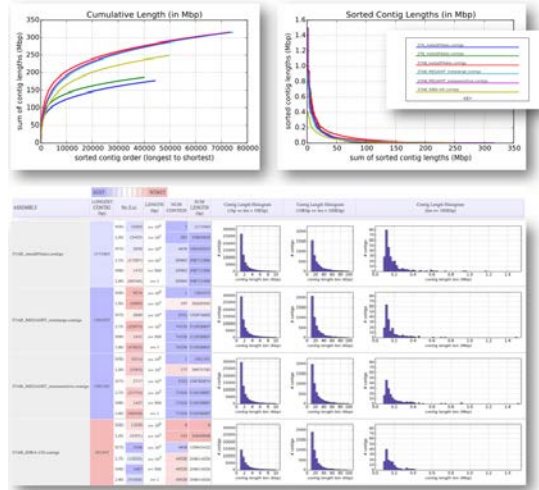# KBase is providing access to increasingly sophisticated visualization: New this year
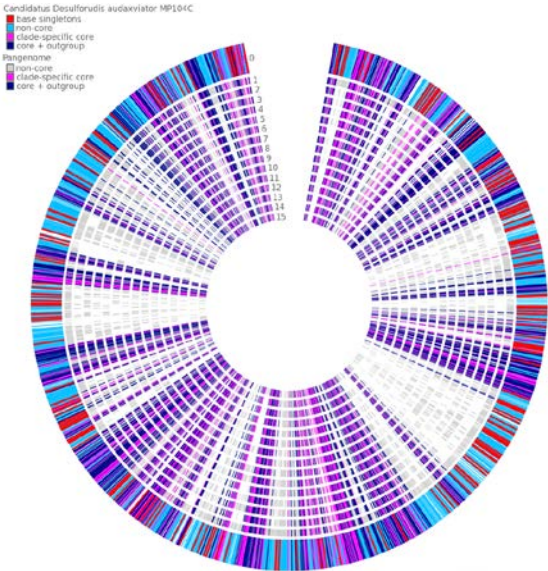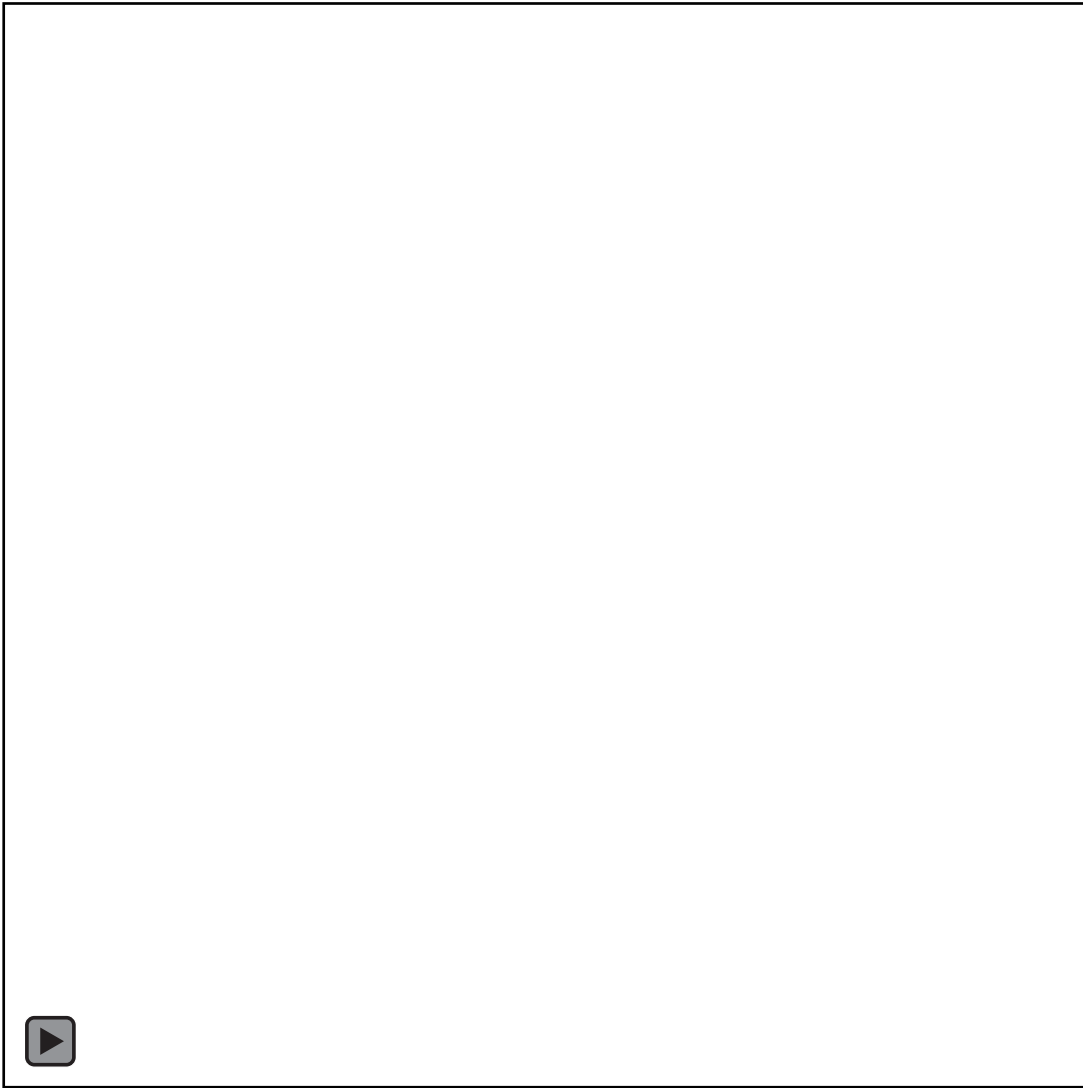

Multi genome alignment

Pangenome Accumulation

QUASTand assembly statistics

CheckM single copy
gene analysis

Network visualization of metabolic interactions
between species in a microbiome

Growing number of analysis apps
(>120 in Production, >55 in Beta, many in development)

Users are taking many paths through the system

Expression apps

Assembly apps

Comparative genomics and annotation apps

Metabolic modeling apps
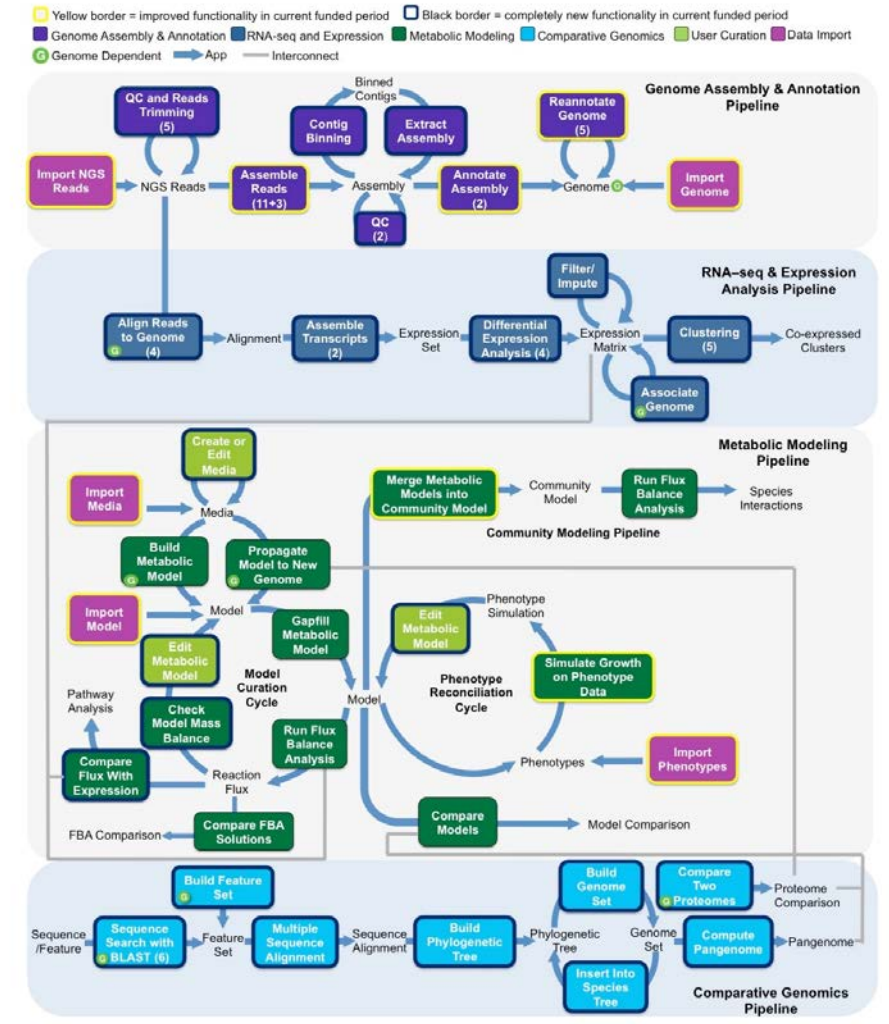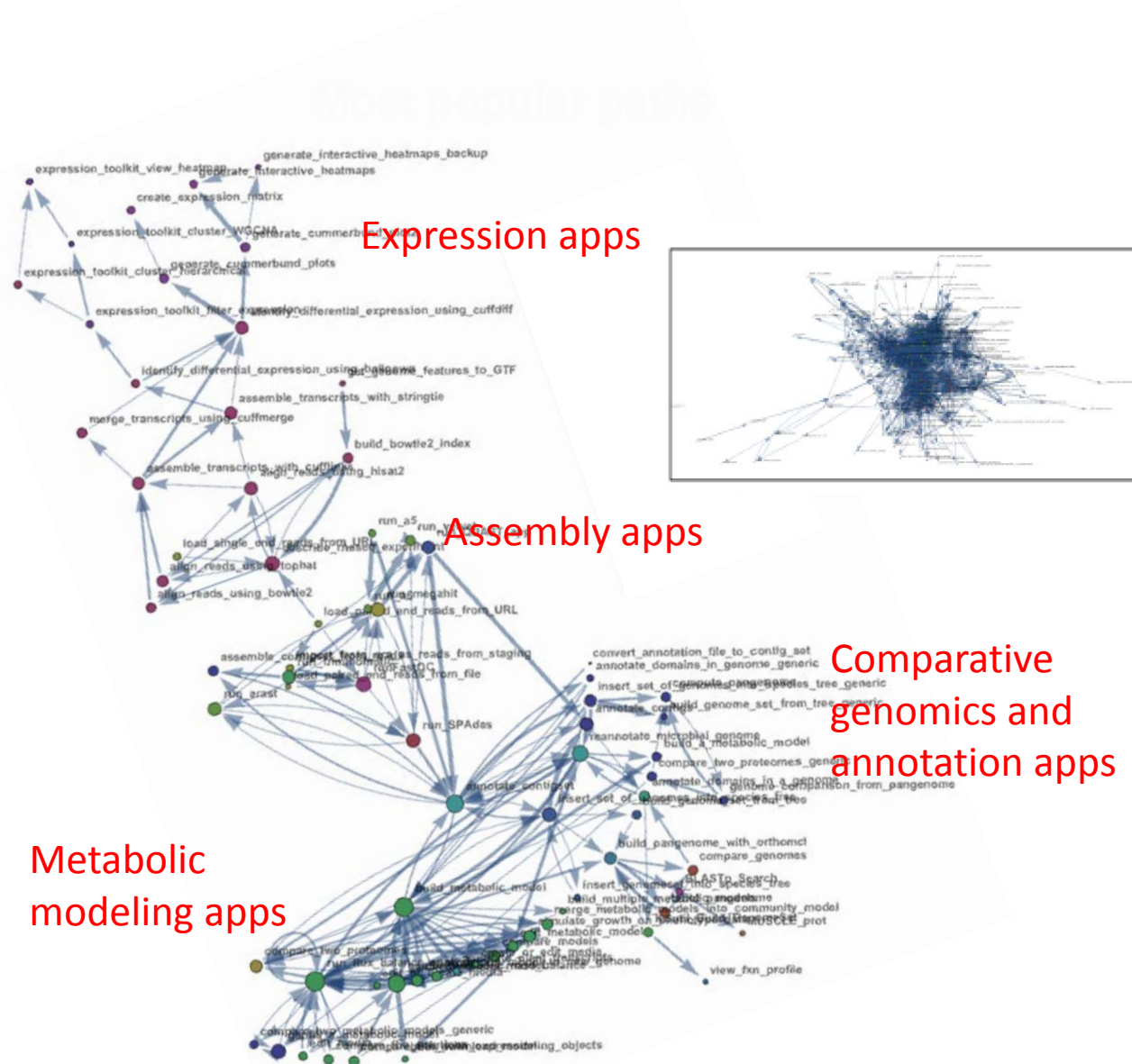
Discovery of genes associated with indole-acetate production in *Pseudomonas SBW25*

Gyorgy Babnig, ANL
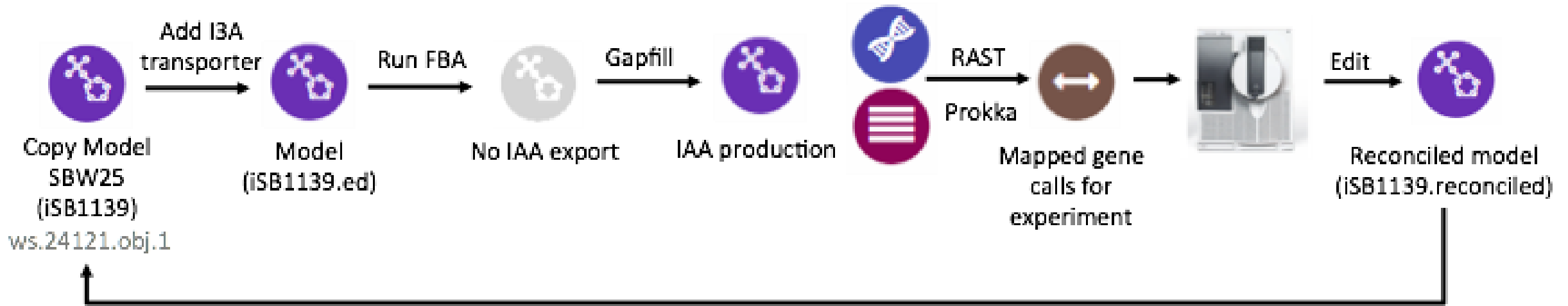
- KBase identified key roles involved in production of indole-acetate (IAA) in SBW25
- KBase enabled easy integration of multiple annotation sources (RAST, Prokka, BIGG) to propose gene candidates for each step of IAA synthesis
- KBase Narrative: https://narrative.kbase.us/narrative/ws.27990.obj.1

Annotating and modeling variation in metabolic function among *Pseudomonas*

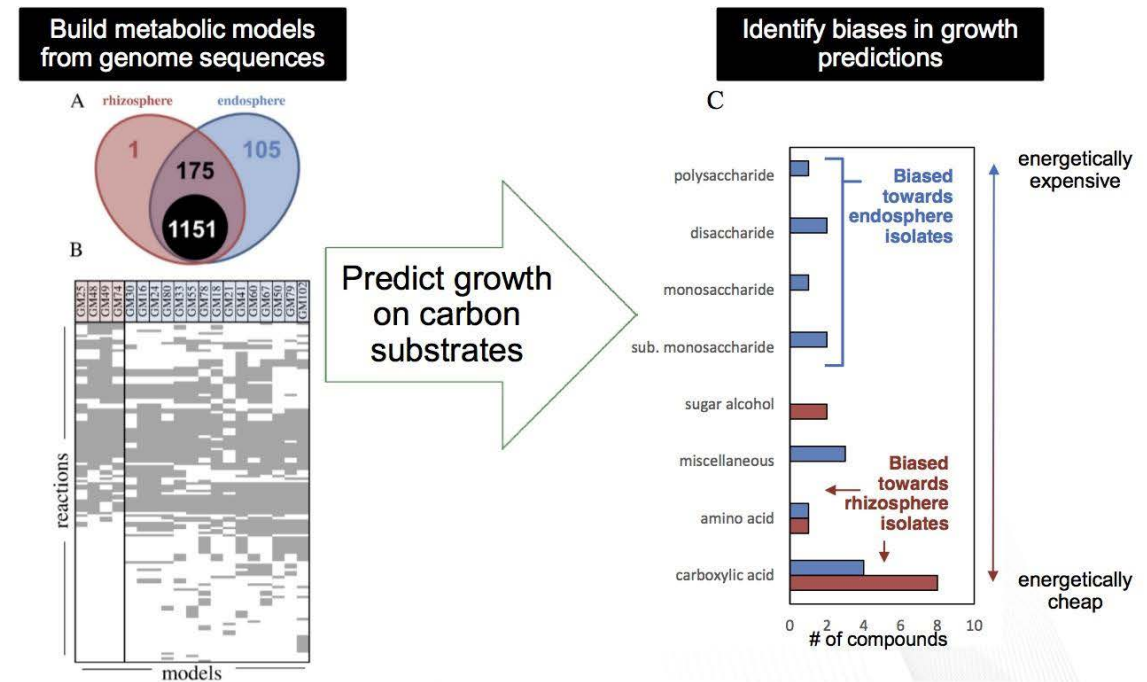*Collin Timm, Johns Hopkins University*

- KBase consistently annotated 19 strains of *Pseudomonas* from rhizosphere
- KBase models predicted how annotation variations lead to phenotype variations
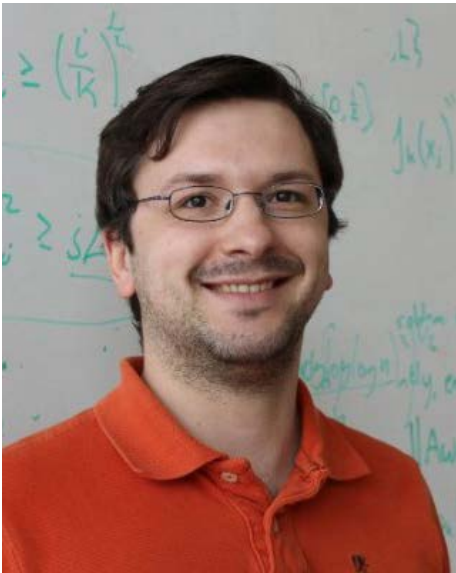
"KBase let me focus on the science, not technical difficulties with implementation of tools"

"KBase enables rapid generation and comparison of metabolic models for genome-sequenced bacteria."

Metabolic models predict substrate bias based on isolation environment

Build metabolic models from genome sequences

Identify biases in growth predictions

Predict growth on carbon substrates

Model comparison summary and predictions for *Pseudomonas* isolates. A. Number of reactions in models. B. Reaction distribution in endosphere and rhizosphere isolates C. Compound groups bias

*Timm *et al*, 2015, *Frontiers in Microbiology*

OAK RIDGE
National Laboratory

Predicting metabolic interactions between an autotroph and heterotroph

Hyun-Seob Song, PNNL

- Unique integration of assembly, annotation, transcriptomics, and modeling all in one platform (KBase)
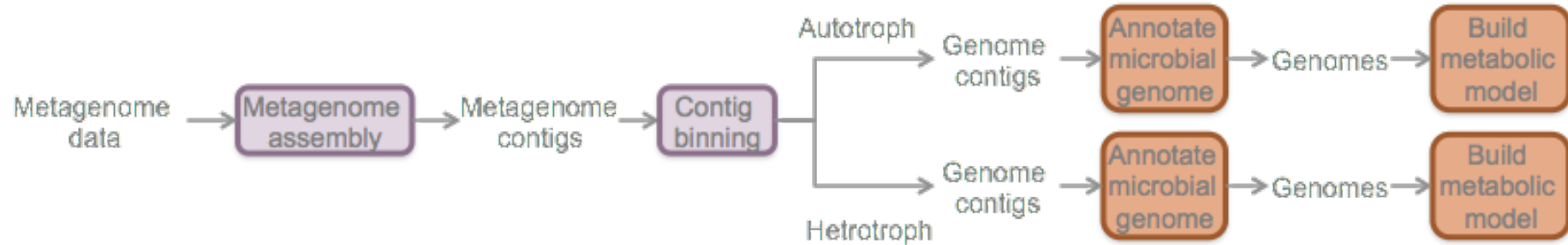
- Predicted metabolite exchanged between species in a simple microbiome

- Explored multiple modeling and gapfilling mechanisms and identified the best alternatives

Cyanobacterium
Integration of a curated model/genome of a closely related species (*Synechocystis* sp. PCC 6803) (Nogales et al., 2012)
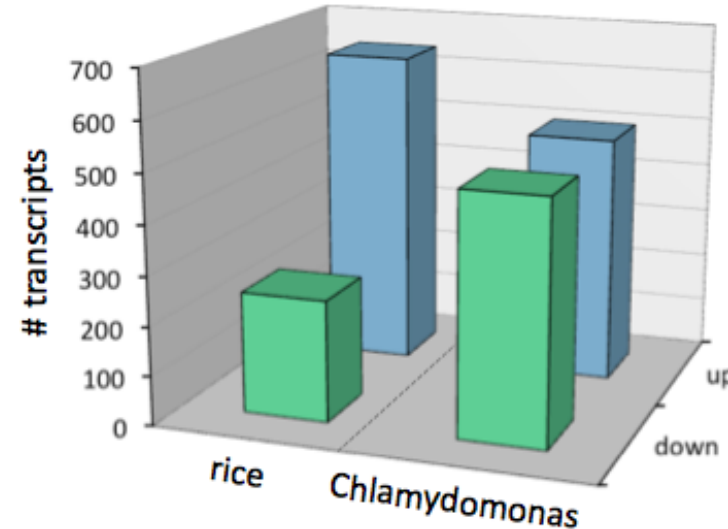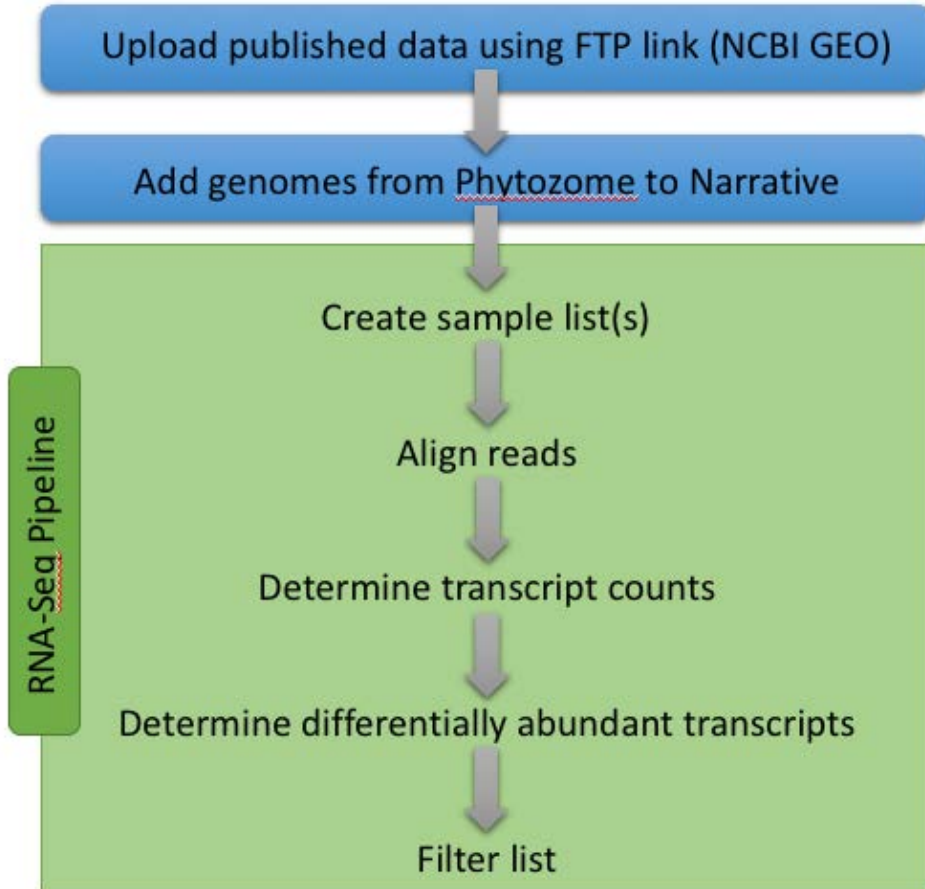
Heterotroph
The ModelSEED algorithm

Compartmentalized community network
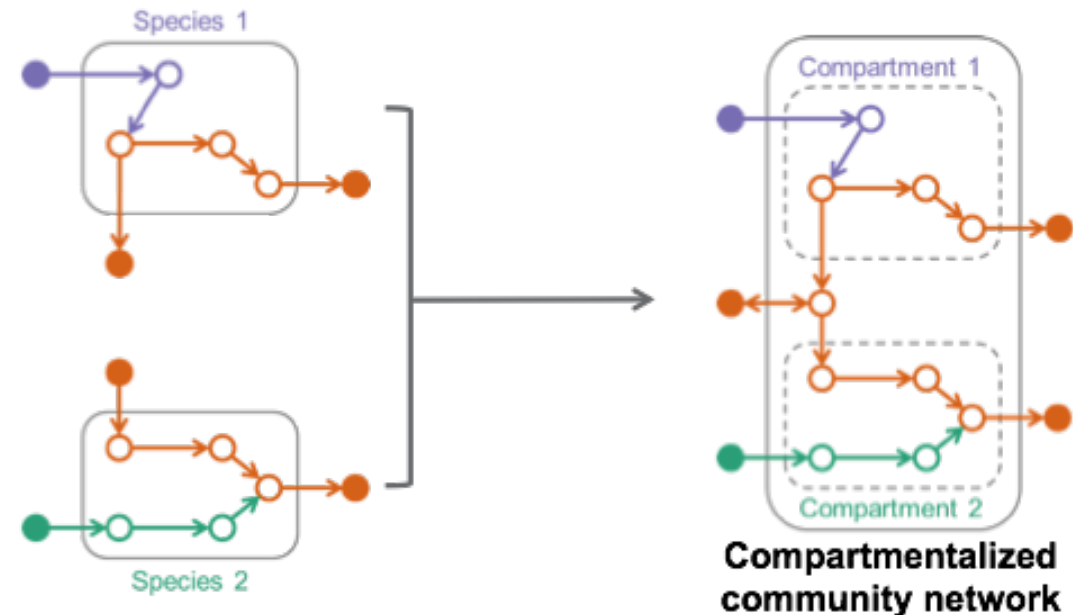
# Uncovering differential response to Zn levels in plants

Crysten E. Blaby-Haas, BNL

- Identifying differentially expressed genes in Chlamydomonas in presence of perturbed Zn levels
- Comparing with differential expression in similar conditions in other plants

- Identified clusters of genes with similar expression profiles and mapped clusters to functions
- Identified clusters of genes with evolutionarily conserved responses
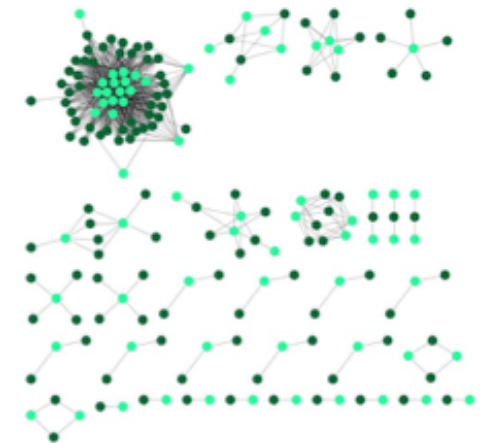
Examples of conserved responses:

Transport:
ZIP-type Zn transporters
ABC transporters
P-type ATPases
Amino acid transporter
Potassium transporter
Sulfate transporter

Zn-dependent enzymes:
Dehydrogenases
Carbonic anhydrases
Hydrolases

chlorophyll degradation

Signaling:
AP2-TFs
MYB-TFs
B-BOX-TFs
kinases

Growth in KBase use

More users, more Narratives, more publications

- Total number of users > 4000 (active usage >2500)

- Over 7000 Narratives created by external users

- Monthly unique logins > 150

- 40 Publications citing KBase (2015-present) in their methods

KBase Publications and Mentions per Year (April 6, 2018)
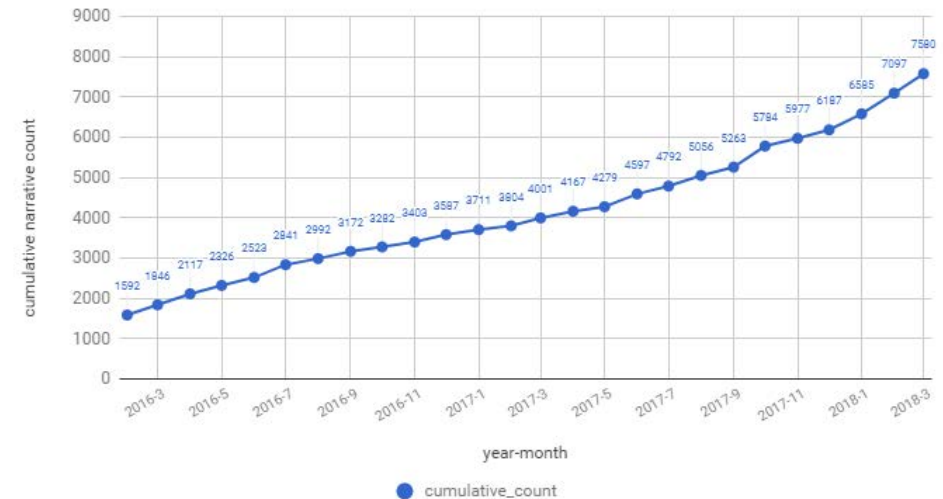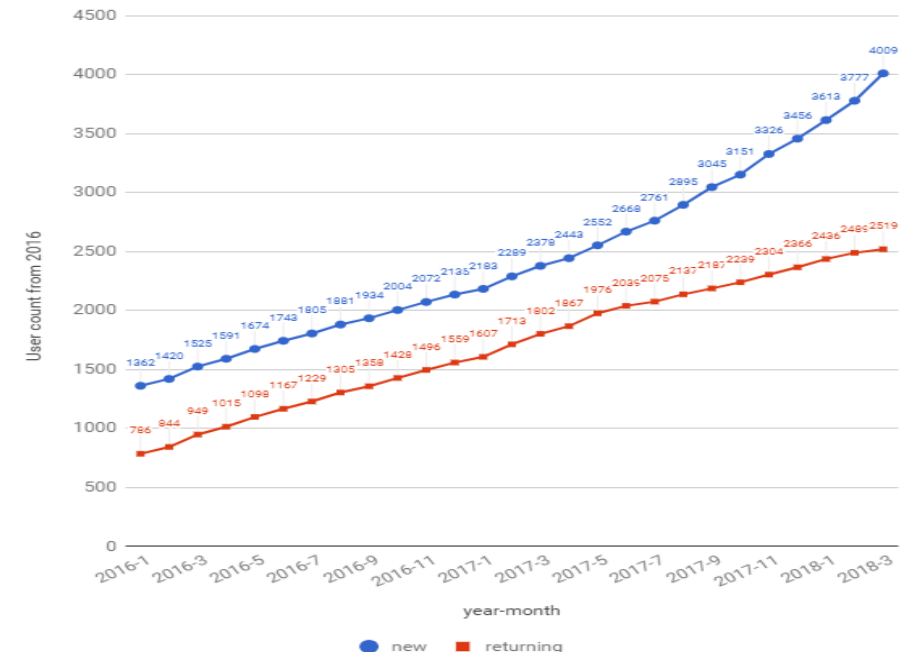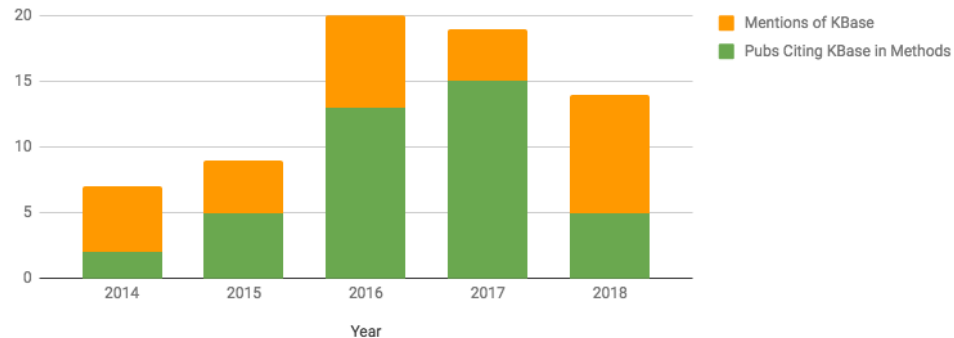
User Engagement Efforts

2017 and 2018 to Date:

- User Engagement Events in 26 locations
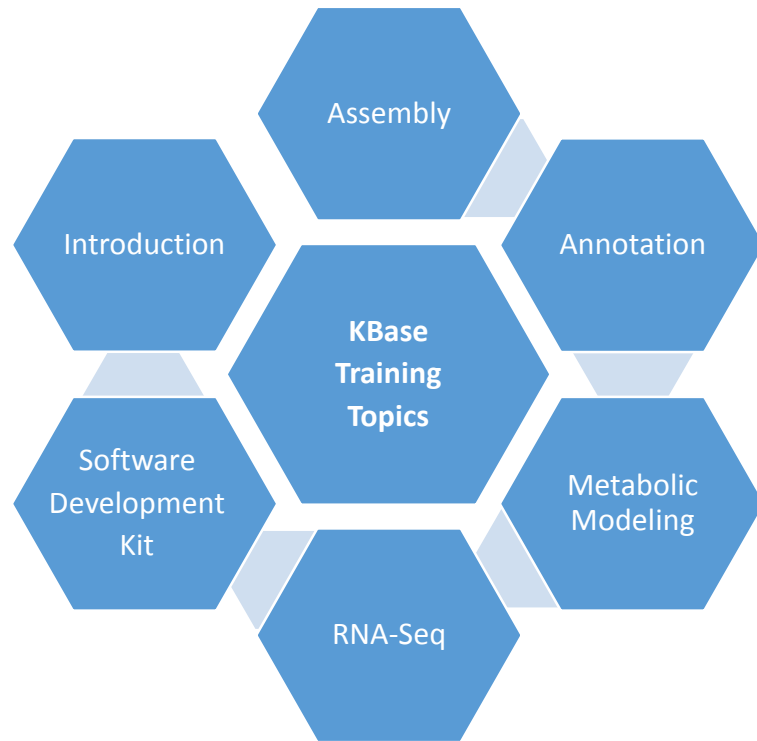- Total attendance of more than 960

Future:

- Increased focus on recurrent scheduled webinars and remote events
- Targeted events with SFAs and other key stakeholders

We can work with you to help your work and teams

➢ We can provide workshops and webinars at your site
➢ Learn how to use KBase to help organize your projects data and analyses
➢ We have support for scientist/developer interchange
➢ We are starting a **User Working Group** where we help organize strategically important science and development by engaged groups using KBase infrastructure.

Assembly

Introduction

Annotation

KBase Training Topics

Software Development Kit

Metabolic Modeling

RNA-Seq

Online Tutorials

Webinars

Individualized Sessions

Ways to learn how to be a KBase power user

Join us: engage@lists.kbase.us

Transition to New Stage

Base functionality has been introduced into the system to support the following areas:

- Assembly, Annotation, Comparative Genomics, Metabolic Modeling ,Microbiome Analysis, Gene Expression Analysis
- Bulk data management including improved upload from NCBI, FTP, local machines
- Robust search of KBase (and public data at JGI)

System is stable enough that 3rd parties can start bringing in their own data and tools with some assistance from KBase staff

- Ready for co-development with other DOE Facilities
- Future expansion in scientific capabilities will be community-powered
- User Working Groups (UWGs) will be formed to target specific scientific areas of interest
- Goals are to promote cultivation of power users, incorporation of high value data sets, and sharing/access to tools of interest to the scientific community

Project focus will be towards user growth and knowledge engine development

Collaboration with JGI

Strategic achievements:

- Established joint JGI/KBase leadership group to coordinate efforts
- Generated a 4-year roadmap with milestones to mark collaboration progress
- Working closely with Dr. Igor Grigoriev on metabolic models of select fungal microbes
- Incorporating reference data from Phytozome
- Working on Metagenome Assembled Genomes with Dr. Emiley Eloe-Fadrosh and IMG
- Initiated co-development activities starting with shared service to compute genome homologies across JGI and KBase public data

Tactical progress:

Developed a prototype data conduit between JGI and KBase (Pull to KBase)
- Enables users to search JGI databases captured in JAMO via search interface in KBase
- Supported data types can be transferred directly into KBase Narratives

Wrapping JGI-developed analysis tools and pipelines as KBase apps
- BBtools (JGI reads data preprocessing pipeline)
- MetaBat (JGI Metagenome binning tool)
- HipMer (high-performance assembler for paired-end reads)
- Significant progress on metagenome annotation pipelines

Collaboration with JGI - MycoCosm

- KBase users now able to access published MycoCosm KBase reference data
- Over 130 MycoCosm genomes ported into KBase
- New app to construct genome-scale models of fungal genomes

Joint JGI-KBase Roadmap

Collaboration with EMSL

Four categories of activity:

Integrating tools with KBase
- MFAPipeline tool for computing flux predictions from labeling patterns determined from NMR data
- NW-chem for prediction of chemical properties from structure

Supporting search and import of data from EMSL into KBase
- Enabling import of metabolomics/NMR data initially as metabolite/value pairs and chemical formula/value pairs
- Enabling import of proteomics data as protein/value pairs and peptide/value pairs

Scientific collaborations
- EMSL and PNNL are working on some fungal systems and will join the KBase/Mycocosm collaboration to annotate, model, and study their genomes
- Applying NWchem to predict thermodynamic properties for compounds in the KBase biochemistry and integrate this data in metabolic models

Codevelopment of infrastructure
- Developing a service to capture, store, represent, search, and compare experimental conditions, which will be shared and linked to datasets at JGI, EMSL, and KBase

3 year roadmap:

Year 1:
- MFAPipeline
- Fungal modeling collaboration

Year 2:
- NWChem
- Experimental conditions service
- NMR data import
- Proteomics data import

Year 3:
- Metabolomics data import
- NWchem prediction of thermodynamic properties

Scaling Computation with DOE HPC Facilities

KBase will leverage ASCR Compute Resources at NERSC to:

> - Quickly deal with high demand or to enable large-scale bulk analysis
> - Enable large-scale precomputation and value added analysis of reference data and user data to enable Knowledge Propagation
> - Support HPC-enabled applications such as HipMER in HipMCL
> - **Support JGI co-development activities around homology, taxonomy, environmental "similarity", ID-Mapping**

KBase's Execution Framework leverages NERSC's Shifter container technology to seamlessly run KBase applications.

Establishing new User Working Groups (UWGs)

KBase will work with the broader community to expand scientific functionality via User Working Groups - thematic subgroups of BER researchers that will help organize and coordinate the development of data and analyses within and across their programs in defined areas of interest.

Goals of UWG is via KBase:
- Organize a concerted community effort to expand KBase functionality and data resources in a tightly focused research area
- Build a user community in each major research area to spread awareness of KBase capabilities and support new user training
- Obtain organized user feedback on KBase functionality and data in a variety of focus areas
- Work together to design new tools and workflows, and facilitate new scientific collaborations

Three initial kernel UWGs under development based on community input (likely to evolve over time):
- Metabolism: metabolic modeling, metabolomics, cheminformatics, integrating chemistry and biology
- Functional Genomics: RNA-seq, TN-seq, proteomics, GWAS, genome annotation, discovery of gene function
- Microbiome: amplicon analysis, metagenome analysis, predicting species interactions within a microbiome
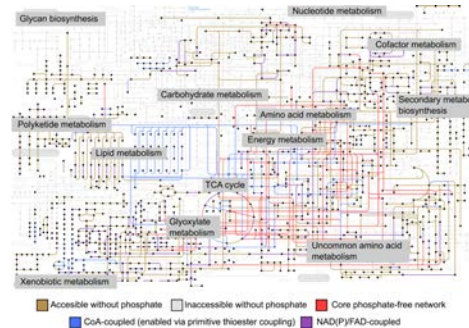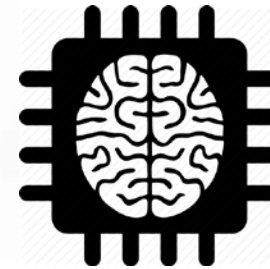
Metrics of success:
- Scientific publications emerging from interactions within UWG
- Demonstrated use of UWG tools, data, and workflows by the KBase user community
- Presentations at major conferences

Office of Biological and Environmental Research

U.S. DEPARTMENT OF
ENERGY
Office of Science

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Future Work

KBase Roadmap Synopsis 2018-2020 (v1)

Key Areas of Focus
- **Collaboration with DOE Facilities**
  - **Major Co-development Efforts with**
    - JGI
    - EMSL

- **Platform Development**
  - **Search**
  - **Knowledge Engine**
  - **Scalable Compute (HPC)**

- **Scientific Method Expansion**
  - **User Working Groups**
    - **Metabolism**
    - **Functional Genomics**
    - **Microbiome**

- **Ongoing Support and Maintenance**
  - **Development and Operations**
  - **General Outreach**

- **Management and Publications**

Projects, Groups and Labs

- Allows projects to organize their people, data and analyses

- Allows them to track metrics of progress and use

- Allows them to publish and track external citation and utility

KBase is a knowledge creation and discovery environment

❖ Integration of primary and derived products into a data model that supports human and machine learning analysis of all shared and published results across the system and automatic propagation of new results to biologically-related entities.

Data about...

Data Integration

Collaborative Analytics

Knowledge Propagation Engine

DOE User Facilities

Users

NCBI, UNIPROT, etc.

External tool integration

Integrative data analysis and modeling
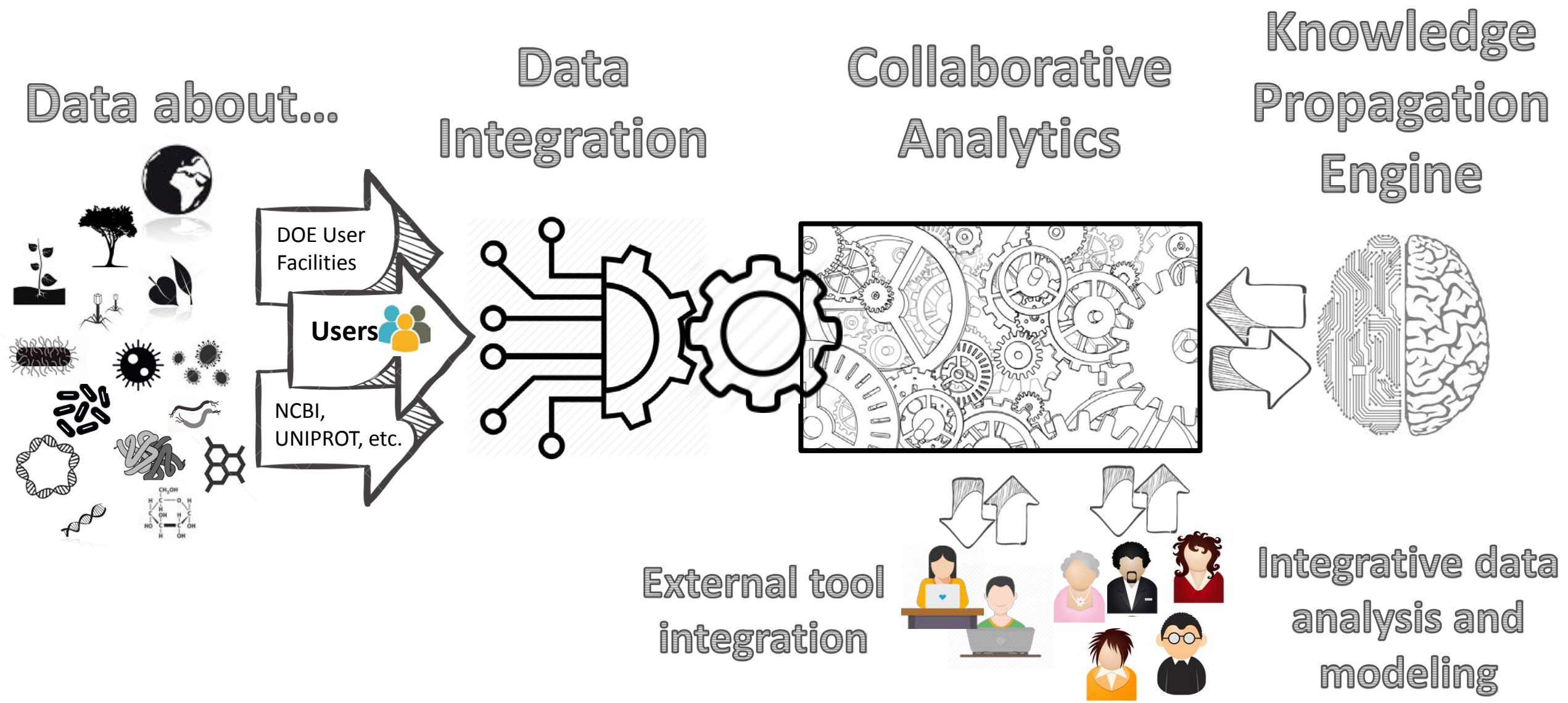
KBase is a knowledge creation and discovery environment

❖ Supports sharing to individuals, groups and the world.

❖ Discovery of results and constant update of "knowledge" from integration and analysis.
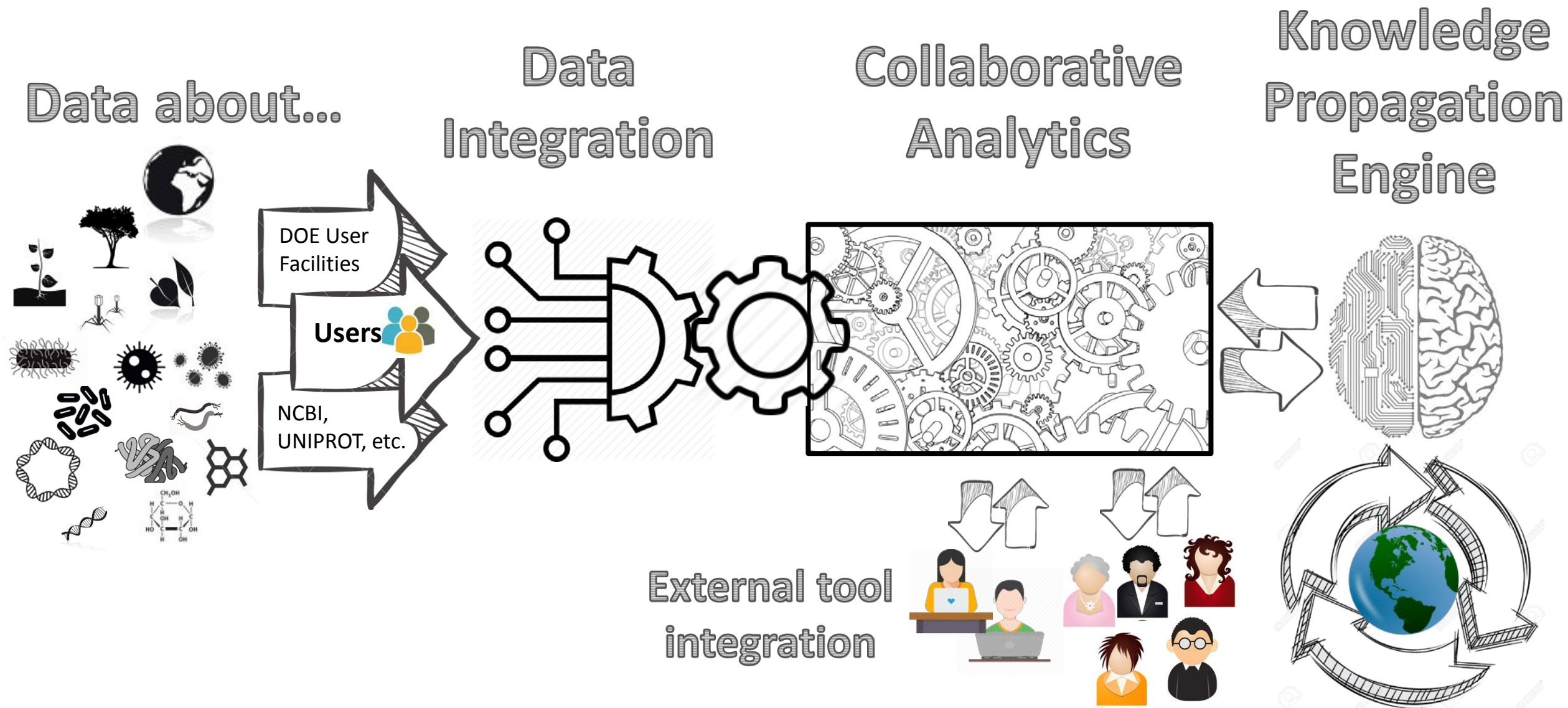
Biological Inquiry
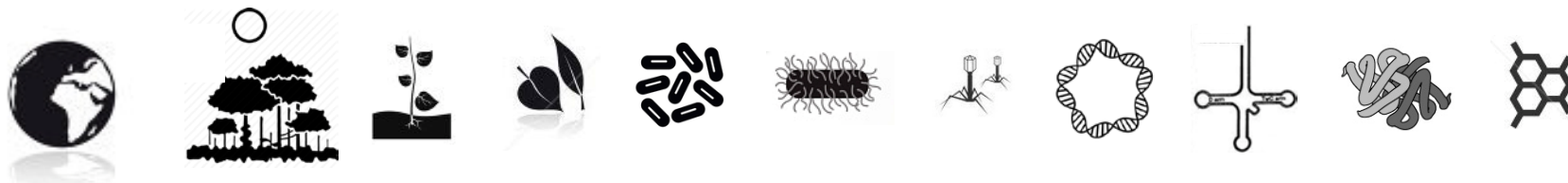
Scientific Community

KBase Users

Reference Source Data

Narrative Development

Reference-Source Function

KBase Inference

Gene name

Orthologs

Expression

Fitness

KBase APPS
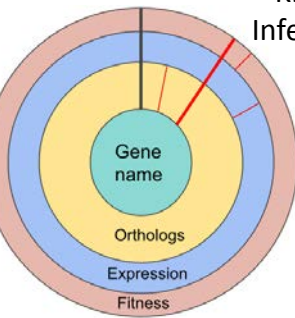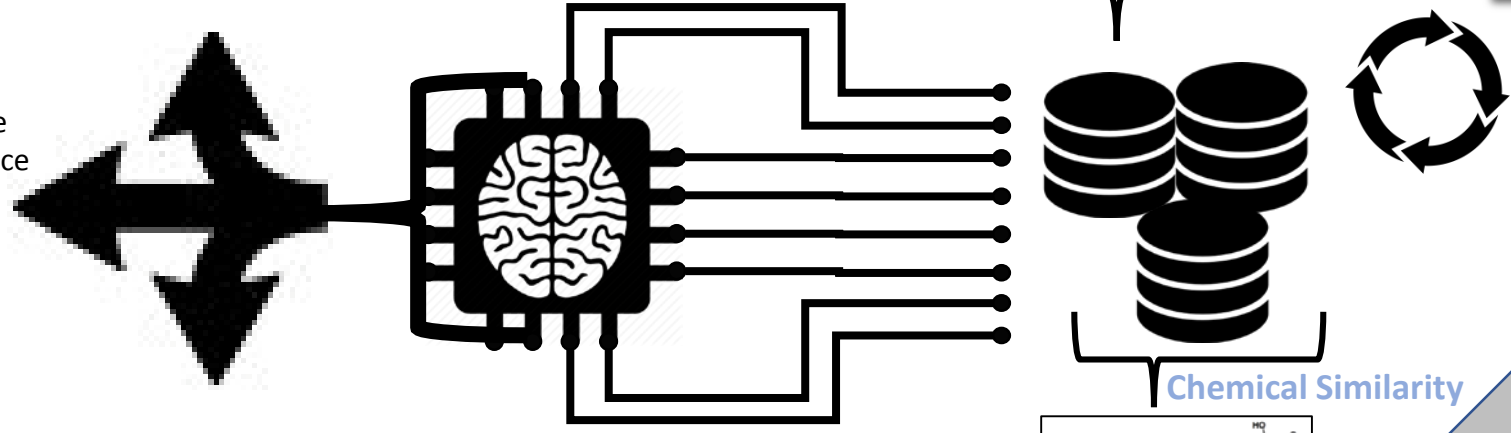
KBase Data and Apps

Chemical Similarity

Biclustering

Functional enrichment

Motif finding
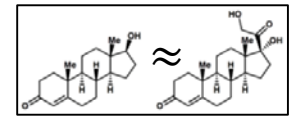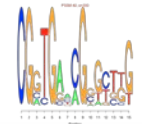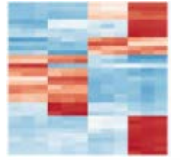
Taxonomy
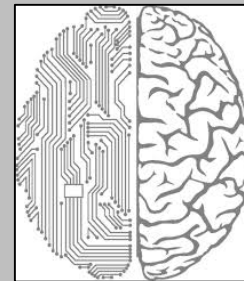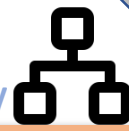
Homology

RESKE

❖ **R**elation **E**ngine
❖ **S**earch
❖ **K**nowledge **E**ngine

o **Feeds**
o **Badges**

o **Evidence-based reannotation**
o **Propagation by orthology**
o **Metabolic model update**

Informers

Propagators

Meta-analyzers

Connectors

Knowledge Engine

Make new discoveries in biological function through integration of data and results shared by users across the system

Incremental development plan:
1. Advanced search of all data and results across system
2. Relation Engine
3. Knowledge Engine

Knowledge Engine Architecture

Dashboard

Catalog

Narratives

Search Data

Account

Feeds 3 new

Genes

|◄ ◄ ► ►| 239 to 252 of 625   Filter genes

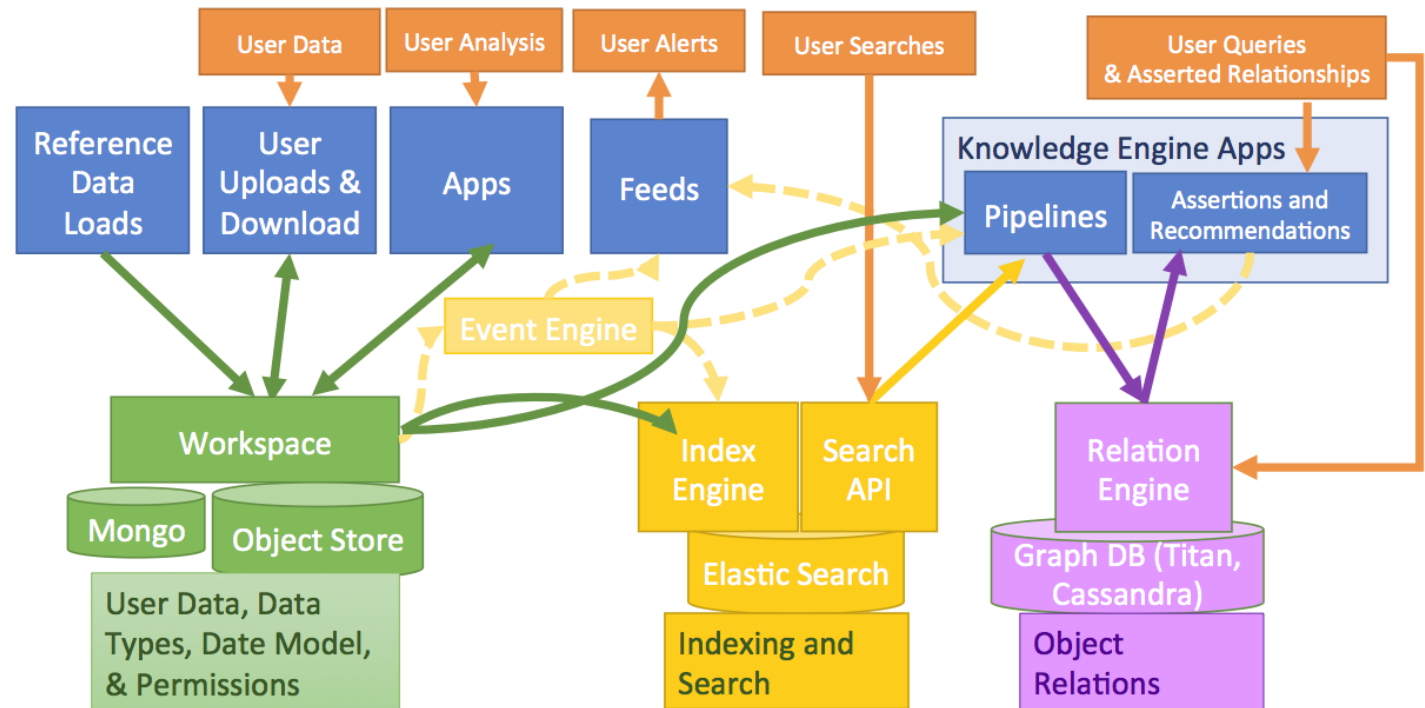| Gene	Distance	User term	Ortholog term
PGA1_c13170	0.43	acetoin catabolic process GO:0045150	nitrogen fixation GO:0009399
WP_044041547.1	0.42	signal transduction by protein... GO:0023014	C4-dicarboxylate transport GO:0015740
cobA	0.42	oxidation-reduction process GO:0055114	glycerol transport GO:0015793
PGA1_RS02130	0.42	phosphorylation GO:0016310	cell wall organization GO:0071555
WP_014881553.1	0.42	phosphorylation GO:0016310	cell wall organization GO:0071555
PGA1_c19160	0.42	oxidation-reduction process GO:0055114	phenylacetate catabolic proc... GO:0010124
WP_014881857.1	0.42	alpha-amino acid metabolic p... GO:1901605	cardiolipin biosynthetic process GO:0032049
PGA1_RS09815	0.42	oxidation-reduction process GO:0055114	Gram-negative-bacterium-ty... GO:0043165
ftsW	0.42	cell wall organization GO:0071555	coenzyme biosynthetic process GO:0009108
PGA1_RS04735	0.42	cell wall organization GO:0071555	coenzyme biosynthetic process GO:0009108
PGA1_c35370	0.42	signal transduction by protein... GO:0023014	heme transport GO:0015886
PGA1_RS03090	0.42	DNA repair GO:0006281	antibiotic transport GO:0042891
WP_014879421.1	0.42	establishment of localization i... GO:0051649	magnesium ion transmembra... GO:1903830
PGA1_c09510	0.42	cell wall organization GO:0071555	coenzyme biosynthetic process GO:0009108

Functional Profile
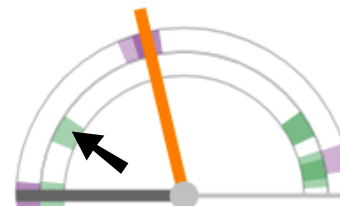
Tooltip

Type	Fitness
Term	protein secretion by the type II secretion system
Id	GO:0015628
p-value	0.0361
distance	0.156

Legend

GO Term	Id	Distance	p-value
User			
acetoin catabolic process	GO:0045150		
Ortholog			
nitrogen fixation	GO:0009399	0.43	0.004
Fitness			
cell motility	GO:0048870	0.83	0.007
acetoin catabolic process	GO:0045150	0.00	0.015
regulation of bacterial-type flagellum assembly	GO:1902208	0.93	0.015
protein secretion by the type II secretion system	GO:0015628	0.16	0.036
regulation of transcription, DNA-templated	GO:0006355	0.95	0.037
Expression			
nitrogen fixation	GO:0009399	0.43	0.004
acetoin catabolic process	GO:0045150	0.00	0.004
alpha-amino acid metabolic process	GO:1901605	0.40	0.017
organic substance transport	GO:0071702	0.93	0.021

Thank you!

KBase is a multi-institutional collaboration

Thank you!

Questions?

INTEGRATION and
MODELING *for*
PREDICTIVE BIOLOGY