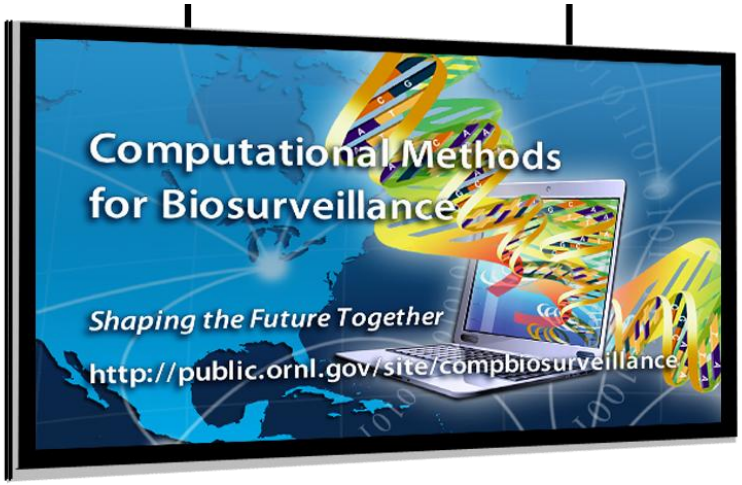
Bioinformatic Strategies: Integrating Data and Knowledge to Improve Biosurveillance



Presented by

Bob Cottingham

Biosciences Division
Oak Ridge National Laboratory

NDIA Biosurveillance Conference
August 28, 2012

Managed by UT-Battelle
for the U.S. Department of Energy

OAK RIDGE National Laboratory

Introduction and Motivation

- Biosurveillance has been primarily based on traditional and manual methods such PCR detection, and intelligence gathering and analysis.
- As science and technology advances there is increasing potential for terrorist engineered threats.
- But also increasing potential for computational means to detect both natural and synthetic threats.

What would it take to develop a standardized biosurveillance system

- Common, Standardized, Scalable
- Computational framework
- Allows rapid, efficient development of new computational analytic methods in a common, integrated data environment.
- Enable common methods for the evaluation and comparison of analytic methods to drive improved performance.
- Provide the basis for computational work environments for biosurveillance analysts.

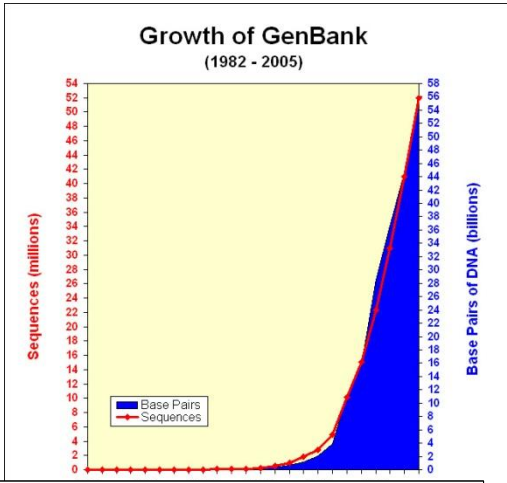
Some relevant topics to consider

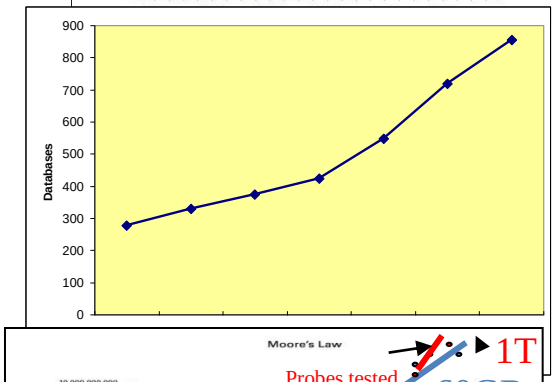
- Data intensive computing for analysis of multiple real-time data streams,
- Genomic, transcriptomic, and other *-omics* analysis of samples,
- Sensor network integration,
- Spatial analysis and visualization,
- Social network analysis

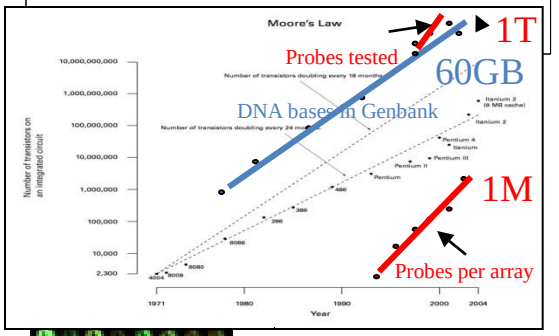
Scaling of Bio Data - Both Volume and Breadth of Data

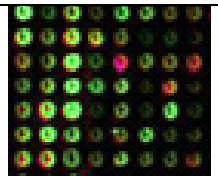
Déjà vu all over again
– Yogi Berra

- Next-next gen sequencers e.g. PacBio = 100GB/hr in 2012?
- Is storing and processing data all that's needed?
- Is computing becoming the bottleneck in research progress?
- Haven't we been saying this since early 1990s?
- Is Moore's Law keeping up?
- What about transcriptomics and omics integration?





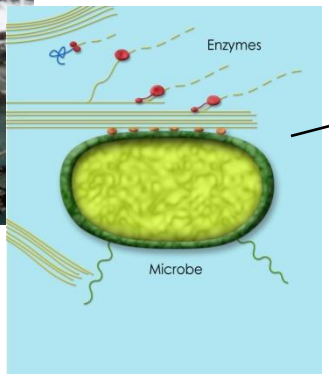




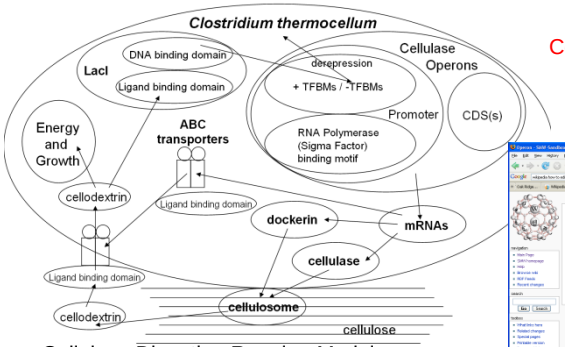
Problem: Discovery of Microbial Pathways Important to Production of Cellulosic Ethanol

Discover novel microbes

Microbial conversion of cellulose

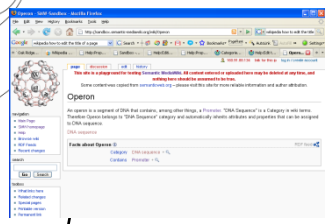


Scientific Model



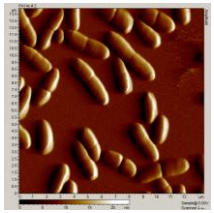
Cellulose Digestion Regulon Model

Conceptual Modeling

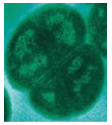


Wiki concept

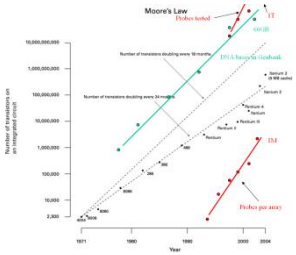
Involve All Researchers



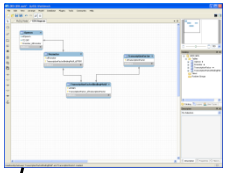
Model microbes and mutants under relevant stress conditions



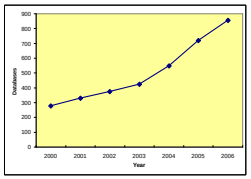
High Throughput Assay Data Sets
2nd Gen Sequencing
Microarrays
Mass Spec
Metabolomic



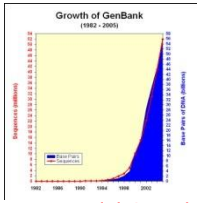
Super-Exponential Growth



Data Model



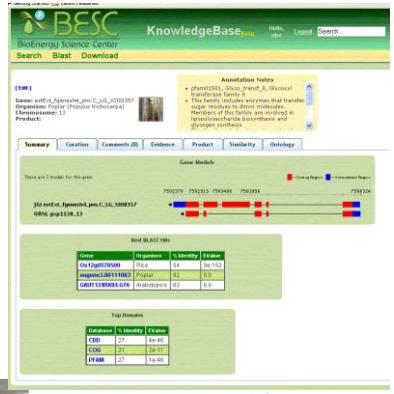
Other Bio Databases



Exponential Growth
PacBio 100GB/hr

Annotation Pipeline

Integration



Knowledgebase

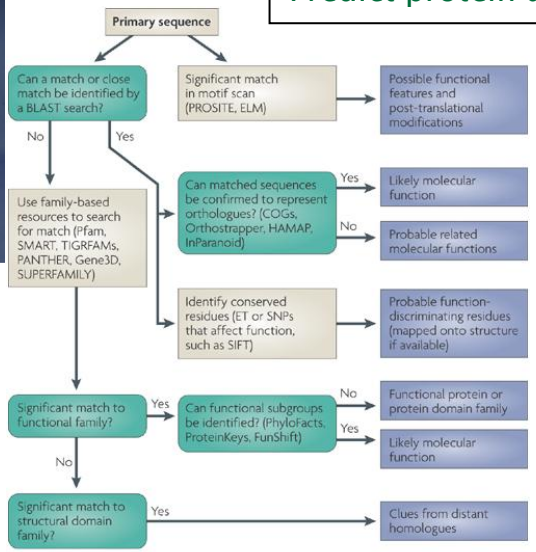
Transition Clusters > HPC

Computational Biology & Bioinformatics:

Future of Analytical Integration

Super-exponential growth: more data to analyze, more often.

Sequencing

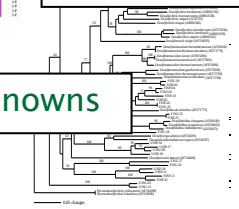


Microbial Annotation Pipeline

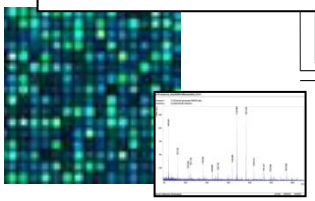
Predict protein and function

Align with all conserved protein domains >1000

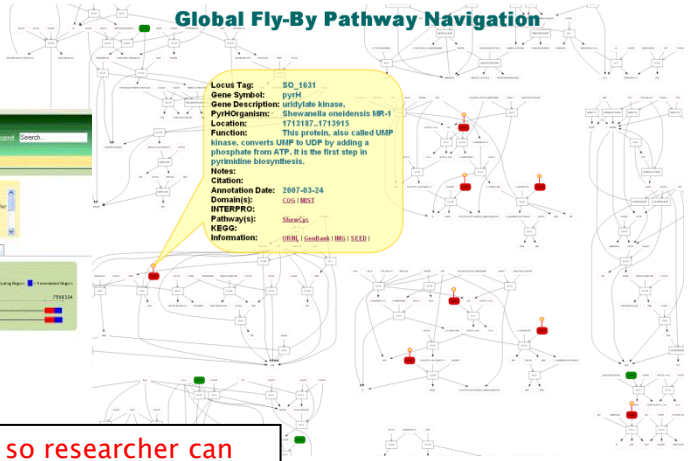
Increased data and integration increases computation and storage

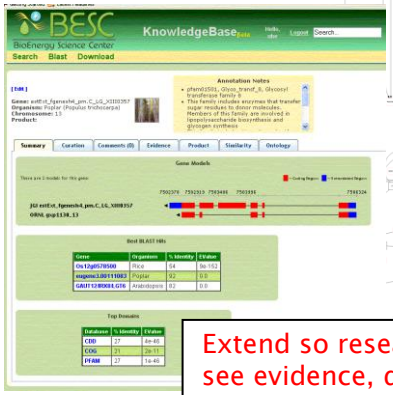


Correlate unknowns



Versioning of analysis tools, data and evidence.





Quality assessment of data, analysis tools and results.

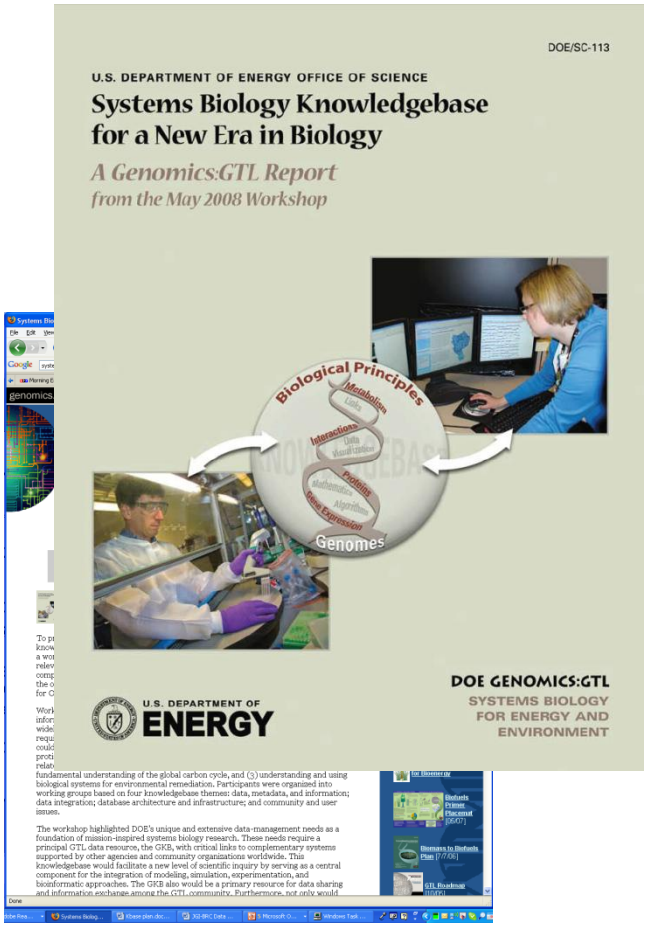
Establish open standards for assessing quality and performance of analytical methods.

Lead open source software policy.

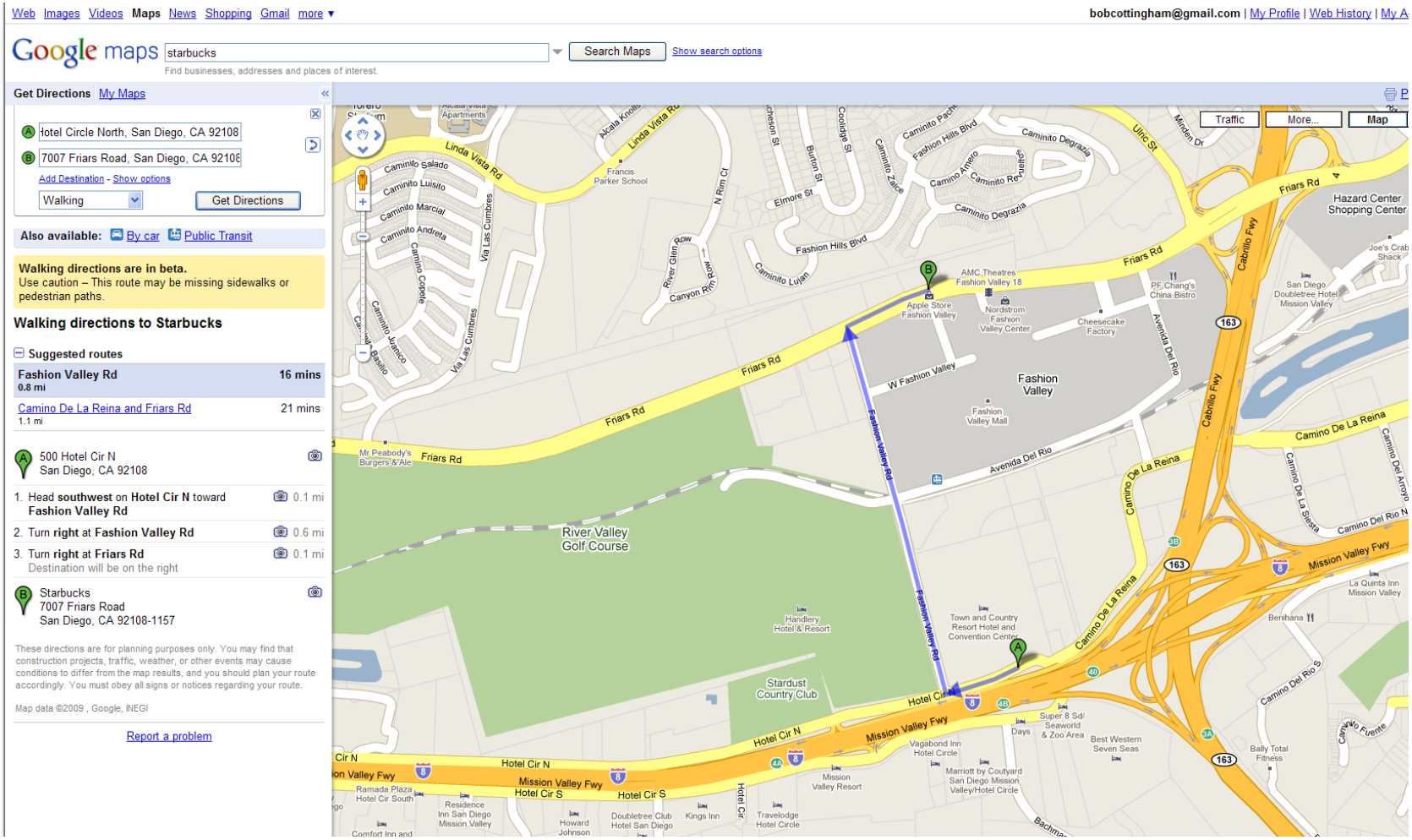
Extend so researcher can see evidence, quality and history.

Knowledgebase R&D Project Background and Objective

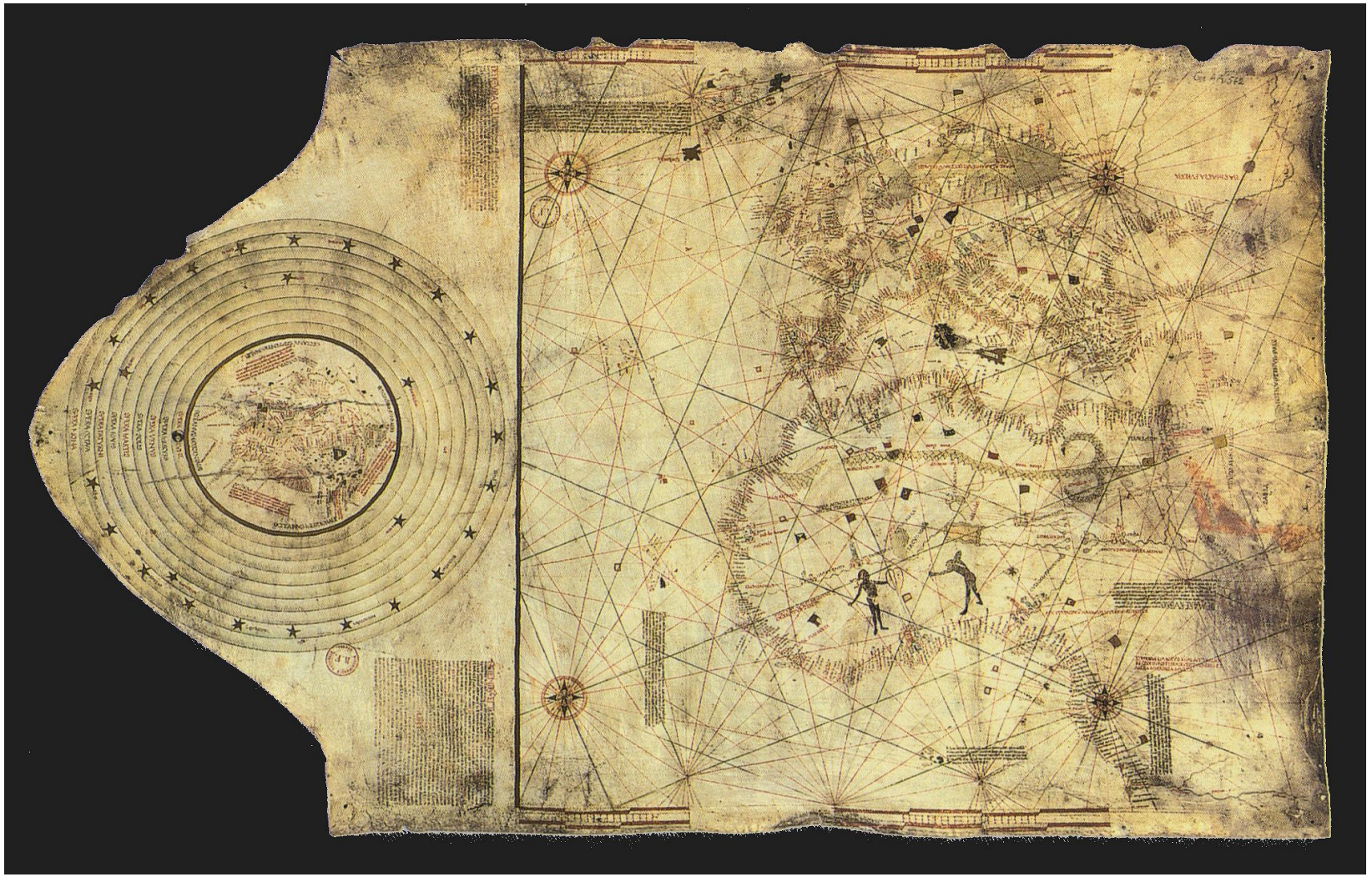
- 2008 Workshop Report
  - Historically projects developed in isolation resulting in isolated data and methods.
  - Vision: Integrating, community informatics resource enabling a broader and more powerful systems biology research effort.
- Objective: Develop an implementation path toward the vision of the DOE Systems Biology Knowledgebase.



Experimental Design to Validate Prediction



A Model of the Earth – 15th century



A Current Model of the Earth

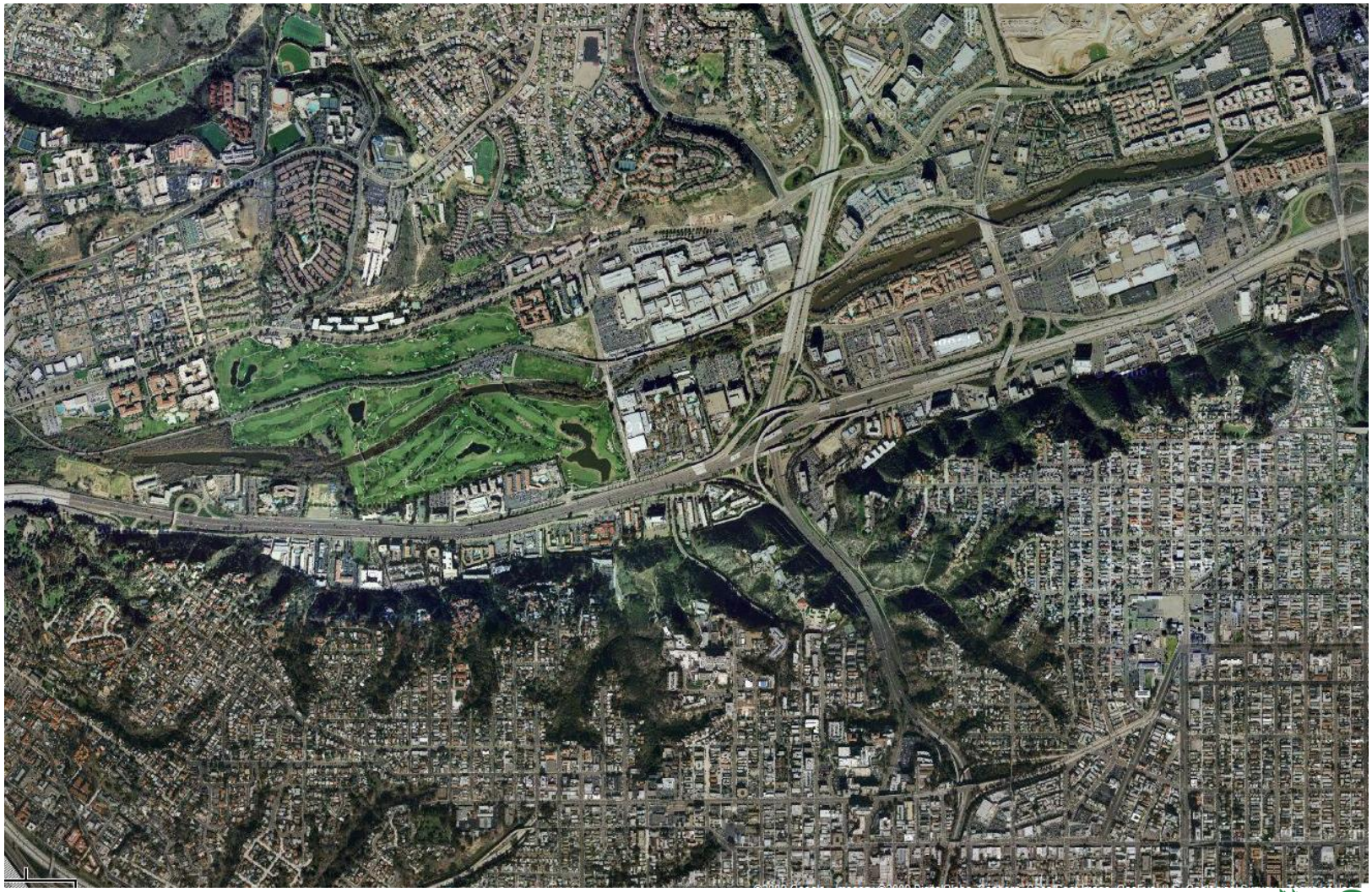
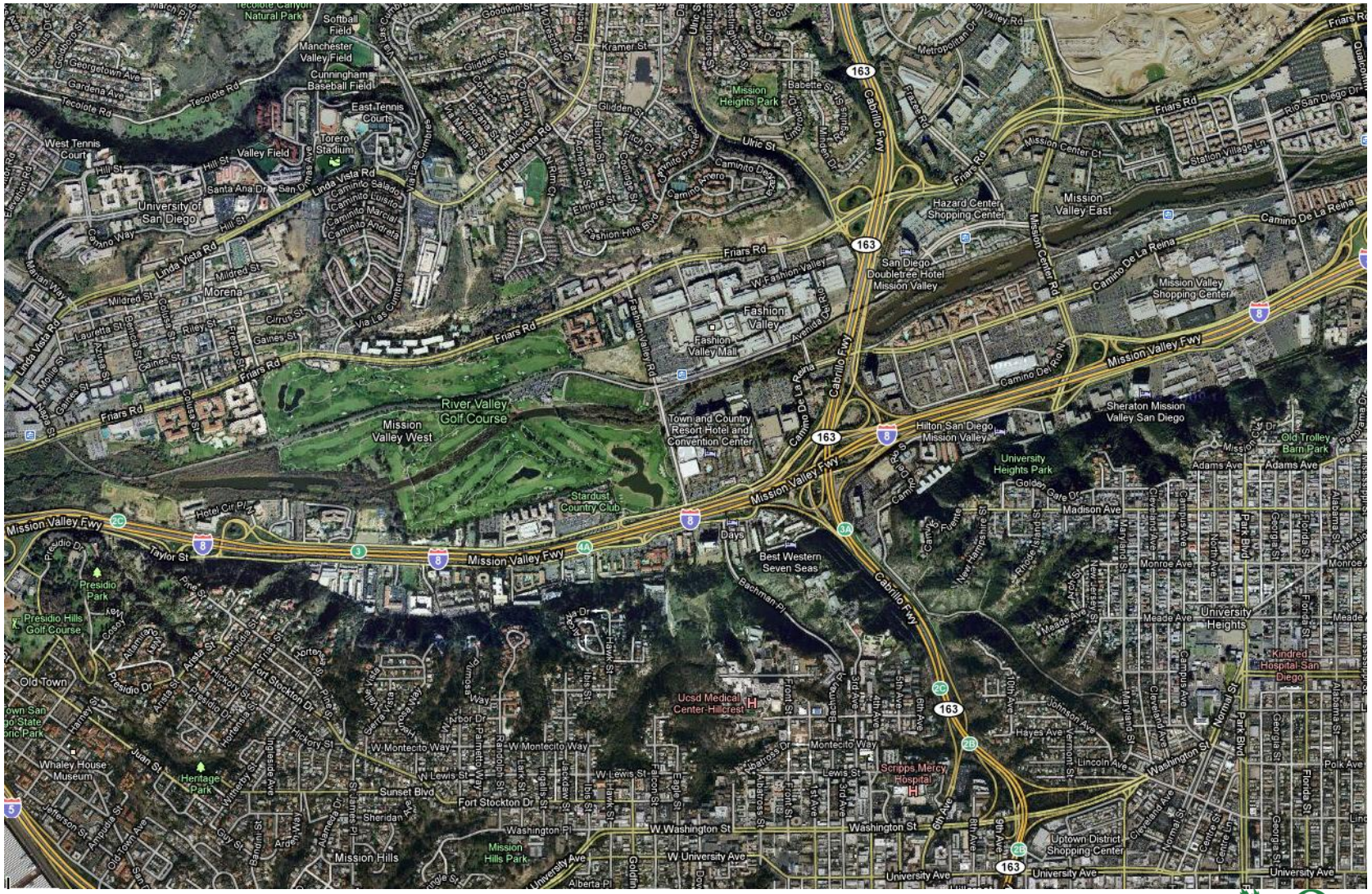
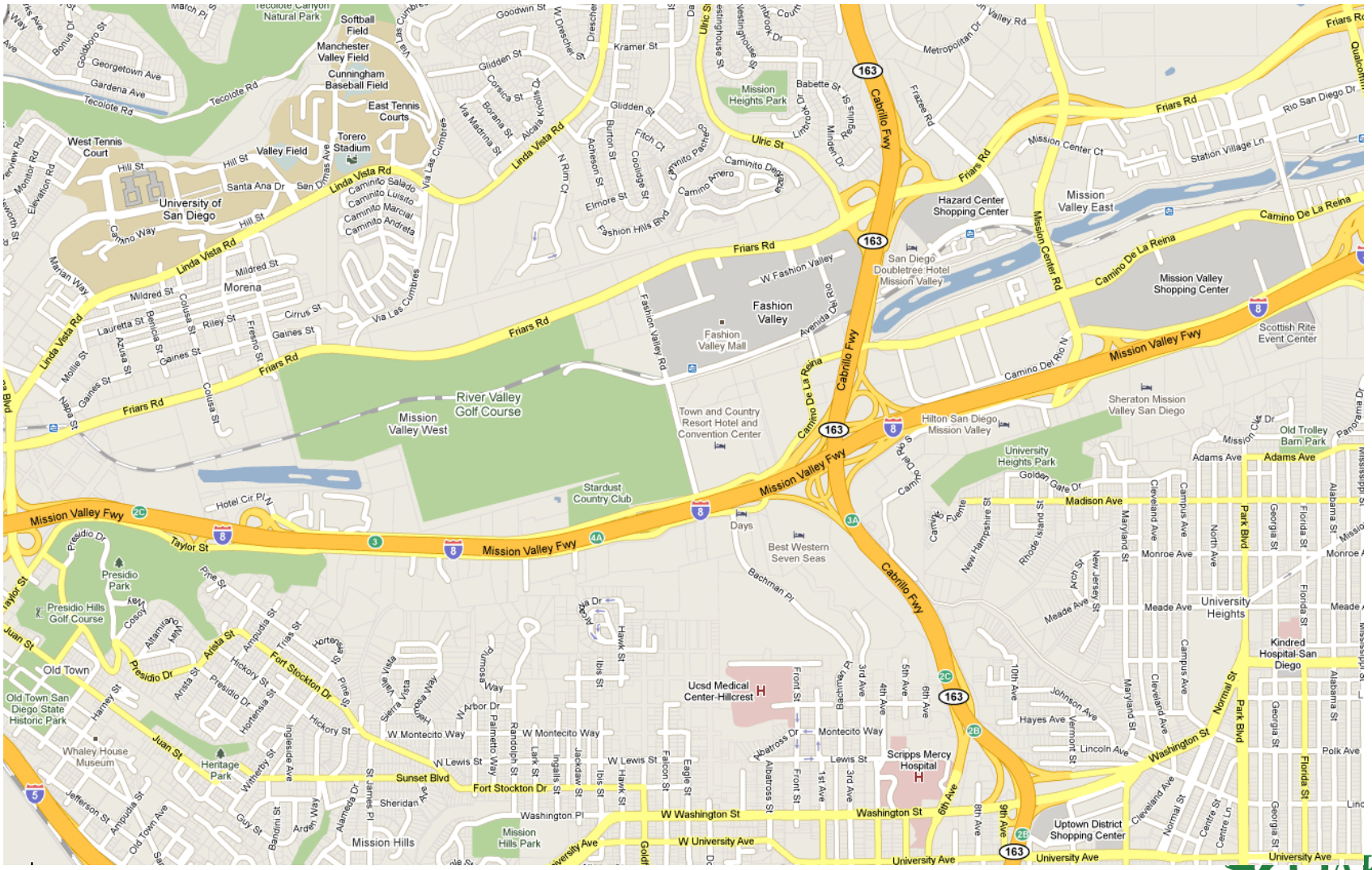


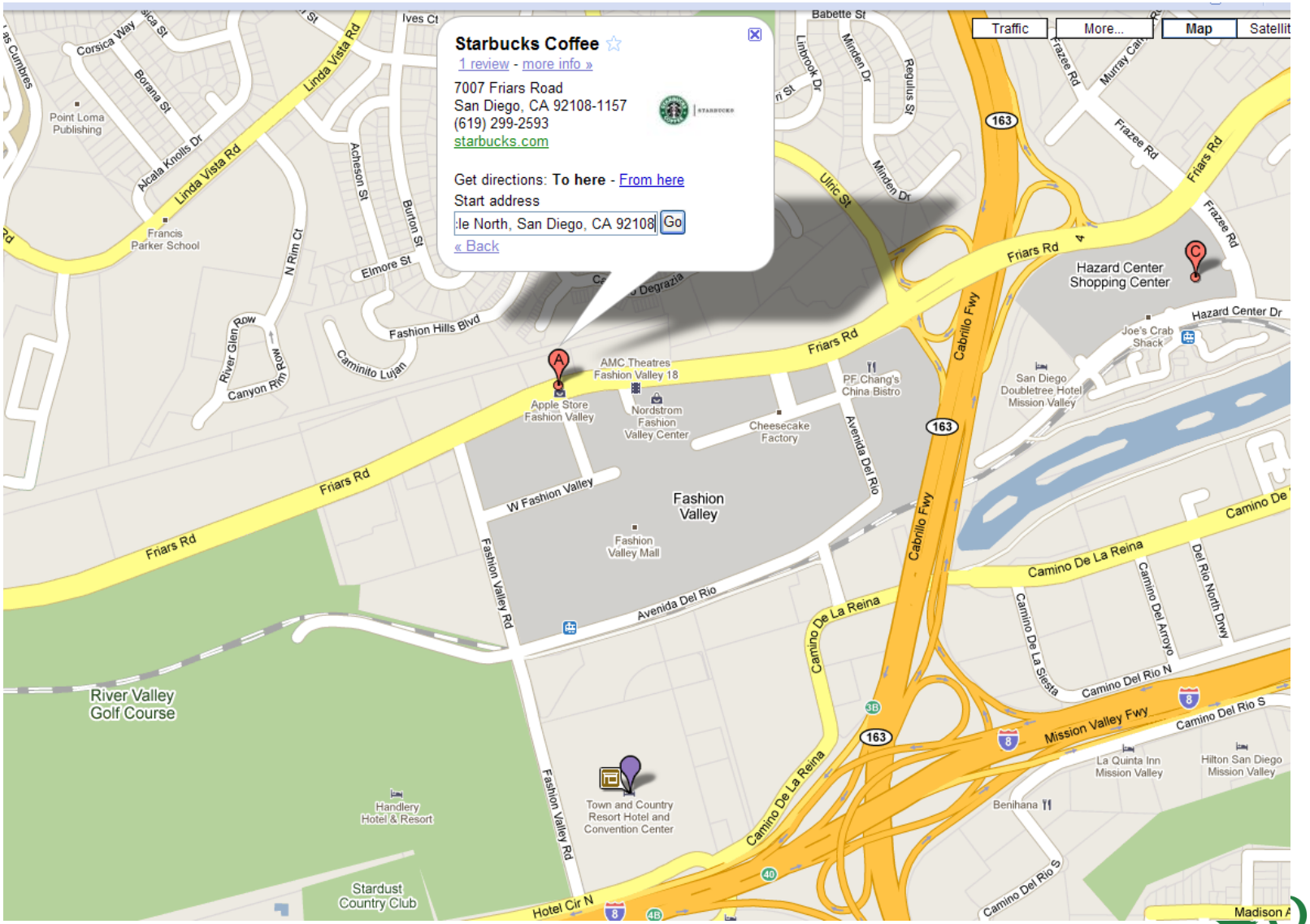
Image with Annotation



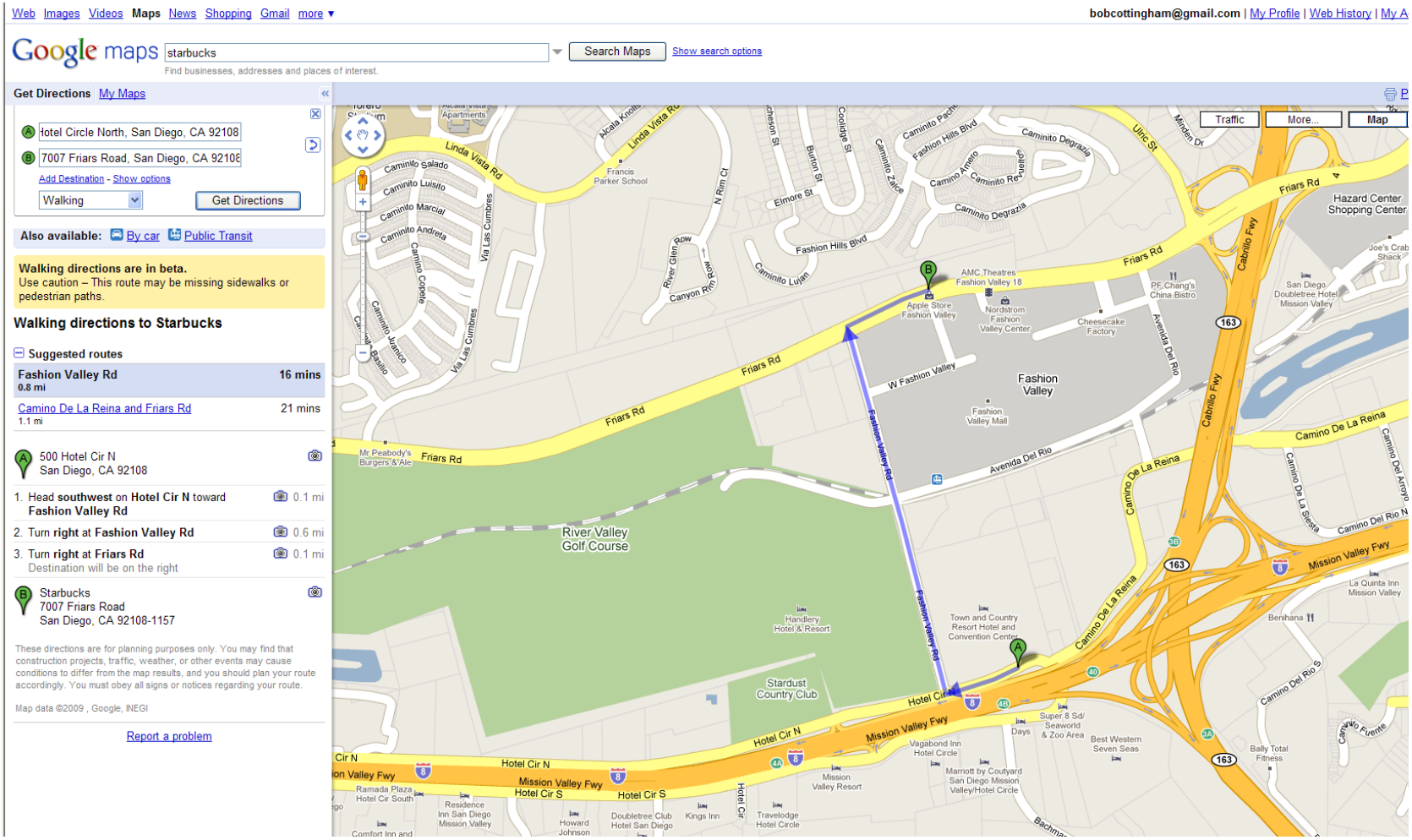
User Interface on Computational Model



Research Function with User Interface



Experimental Design to Validate Prediction



Could something like that exist for biosurveillance?

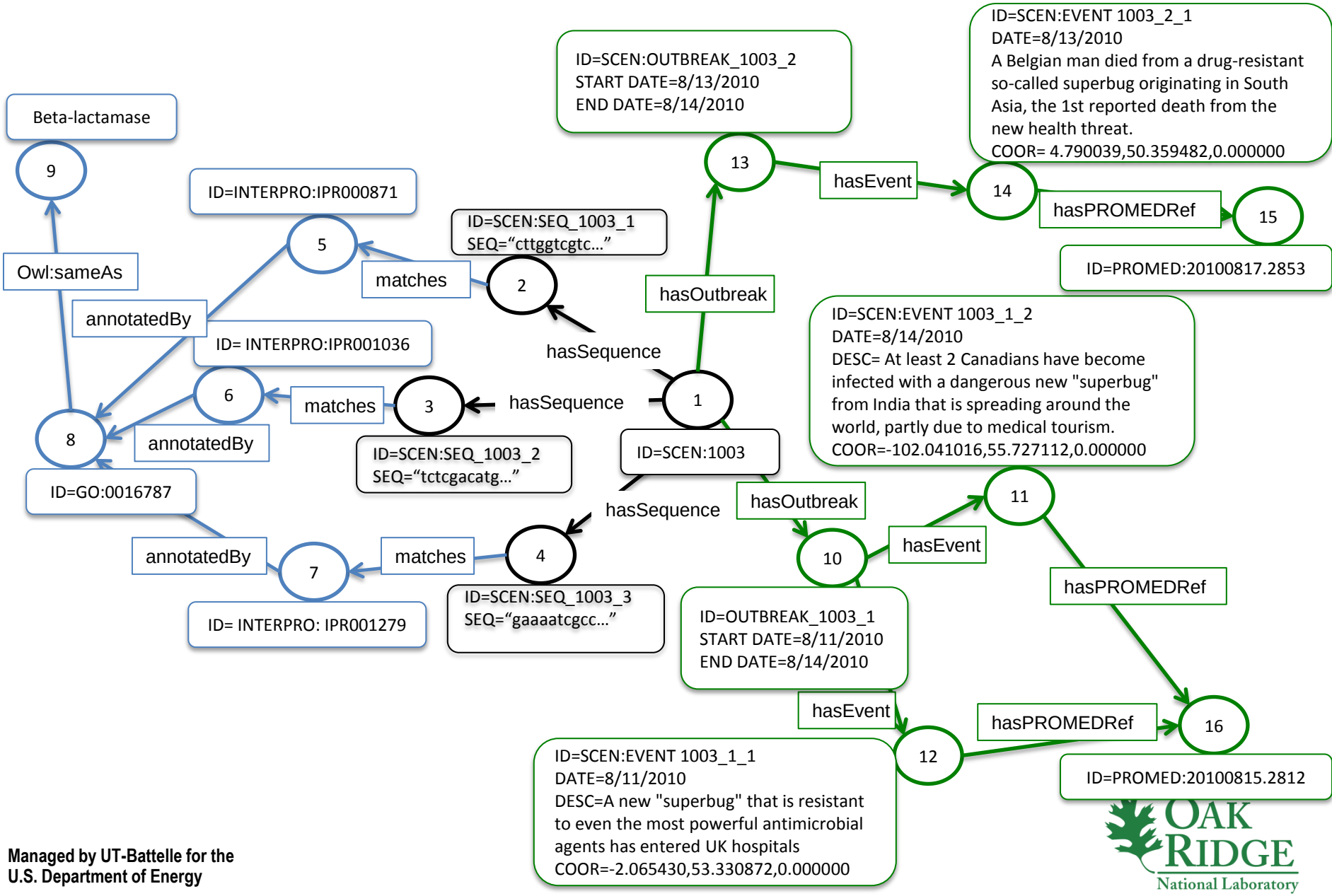
- The ability to collect, communicate and analyze biological-related data is rapidly changing.
- Diverse sources of data can be used in predicting and mitigating an outbreak.

- When a bioterrorism event unfolds, we must rapidly collect, analyze, and filter diverse information to enable the best response and decision-making.
- Information could come from search engines like Google, from scientific labs, field detectors, handheld devices, social networks, blogs, over the counter sales, traditional media or virtually any person or entity that shares information on the Web.

Data Intensive Science – The 4th Paradigm

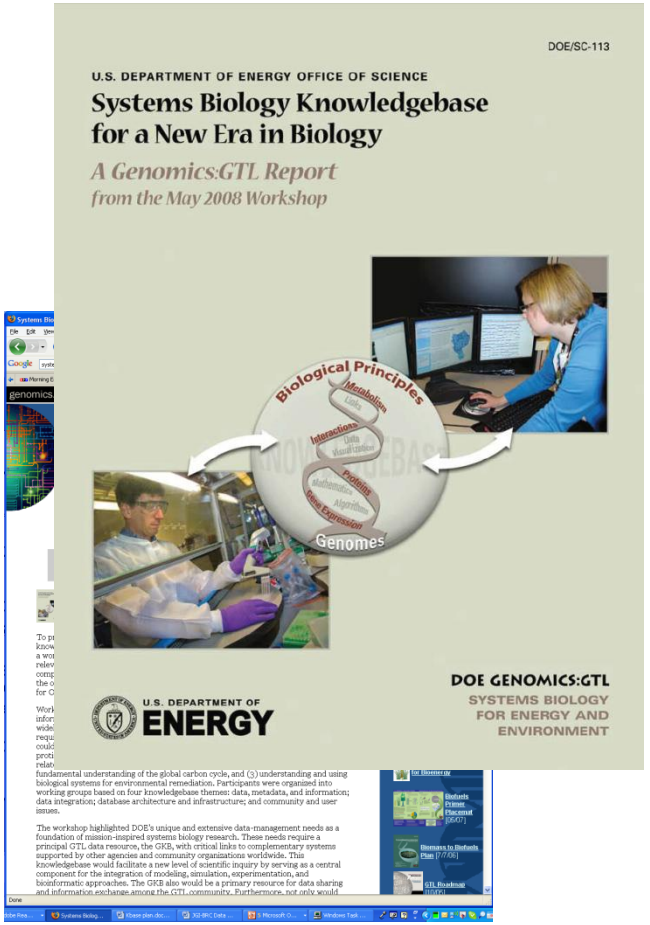
- Experimental, Theoretical, Computer Simulation, … next Data Intensive.
- Data discipline based on databases, schemas, ontologies – scientific community generally lacks understanding of these topics.
- Data intensive science requires specialized skills and analysis tools.
- Each piece of data needs have its associated ontological and semantic information.
- Search, analysis and reuse is supported by standard vocabularies.
- IT industry building huge "cloud services" with high bandwidth, low cost storage and computing. No prominent bio examples yet.
- Future scientific progress in biology depends on how well the community acquires the necessary expertise in database, workflow, visualization and cloud computing techniques.

Scenario Driven Data Modeling SDDM:

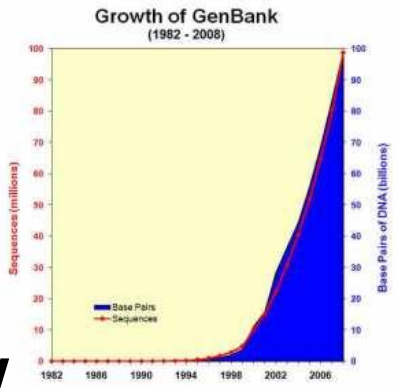


Knowledgebase R&D Project Background and Objective

- 2008 Workshop Report
  - Historically projects developed in isolation resulting in isolated data and methods.
  - Vision: Integrating, community informatics resource enabling a broader and more powerful systems biology research effort.
- Objective: Develop an implementation path toward the vision of the DOE Systems Biology Knowledgebase.

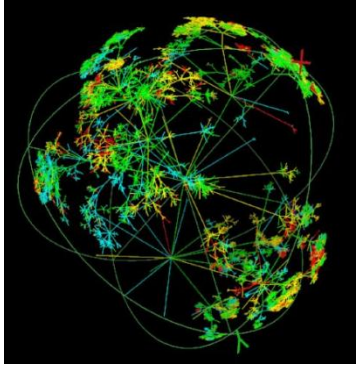


The Problem



- Biological research projects are becoming larger and more complex, both experimentally and computationally.
- To succeed there is an increasing need to cooperate and standardize.
- Technological advances continue to produce exponentially more and diverse types of data.
- A new kind of computational infrastructure is needed for the overall scientific effort to be productive and successful.

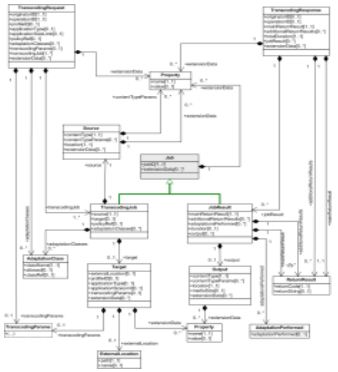
Kbase Mission



- To provide a large-scale, open community computational capability for systems biology research data management and analysis.
- Promote openness and sharing of data, code and computational infrastructure.
- Address problems of large scale data management and processing.
- Utilize computational techniques required for community effort at data integration, open development, large scale resource sharing and to meet other research community policies and objectives.

How Would Kbase Be Different?

- How would Kbase be different?
  - Integrate across projects and laboratories
  - Implies a community research effort
  - More standardized approach
  - More mature software engineering approach
  - More involvement of non-informatics researchers
- Reference models – technical:
  - Open source development, e.g. Linux
  - iPhone Apps, Google Apps, Facebook Apps
  - Google Maps cloud computing with smart phone app
  - Wikipedia shared community reference resource



Kbase Guiding Principles

- Science drives Kbase development.
- Community effort integrating data and methods across multiple laboratories to improve research productivity.
- Open access – data and methods are available for anyone to use with perhaps a limited embargo policy.
- Open contribution – data and source code managed in an open environment and can be contributed by anyone with an editorial / peer review process.
- Distributed data and methods, Kbase is not a single, centralized, monolithic system.

DOE Systems Biology Knowledgebase

Establishing A Systems Biology Modeling Framework

Data generators

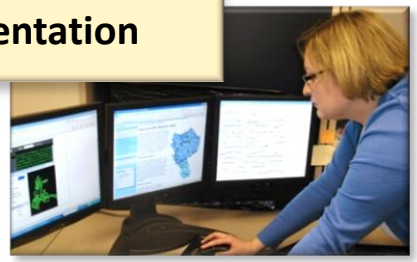
Seamless Submission and Incorporation of Diverse Data

- Standards for data, metadata
- Quality control and assurance
- Automated data handling

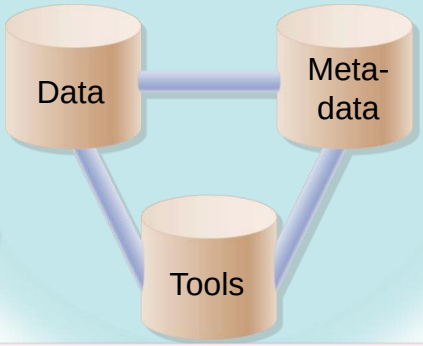
Open-Access Data and Information Exchange

- Flexible user interfaces
- Easy data retrieval
- Environment for *in silico* experimentation

Data users



DOE Systems Biology Knowledgebase



Community-Wide Stewardship

- User, Standards, and Advisory committees
- Value-added analysis
- Training, tutorials, and support

Software and tool developers

Open Development of Open-Source Software and Tools

- Analysis and visualization
- *In silico* experimentation
- Tracking and evaluation of tool use

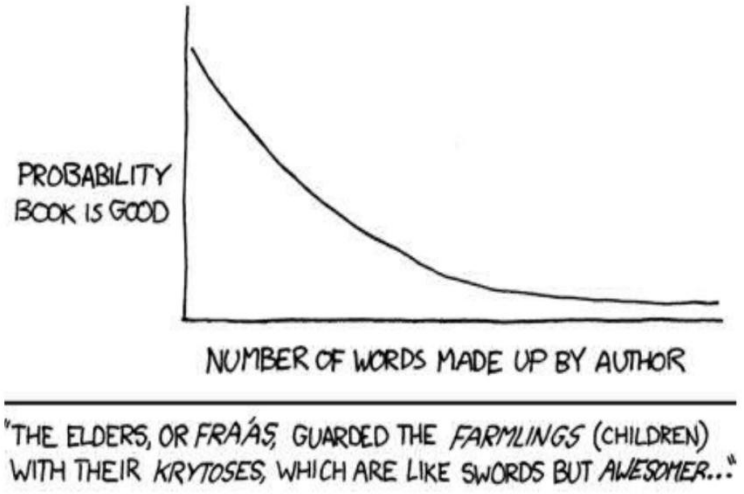
What is a Knowledgebase?

Data is extracted and displayed

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

This is the database model

Knowledge is learning & answering



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

This is the knowledgebase model

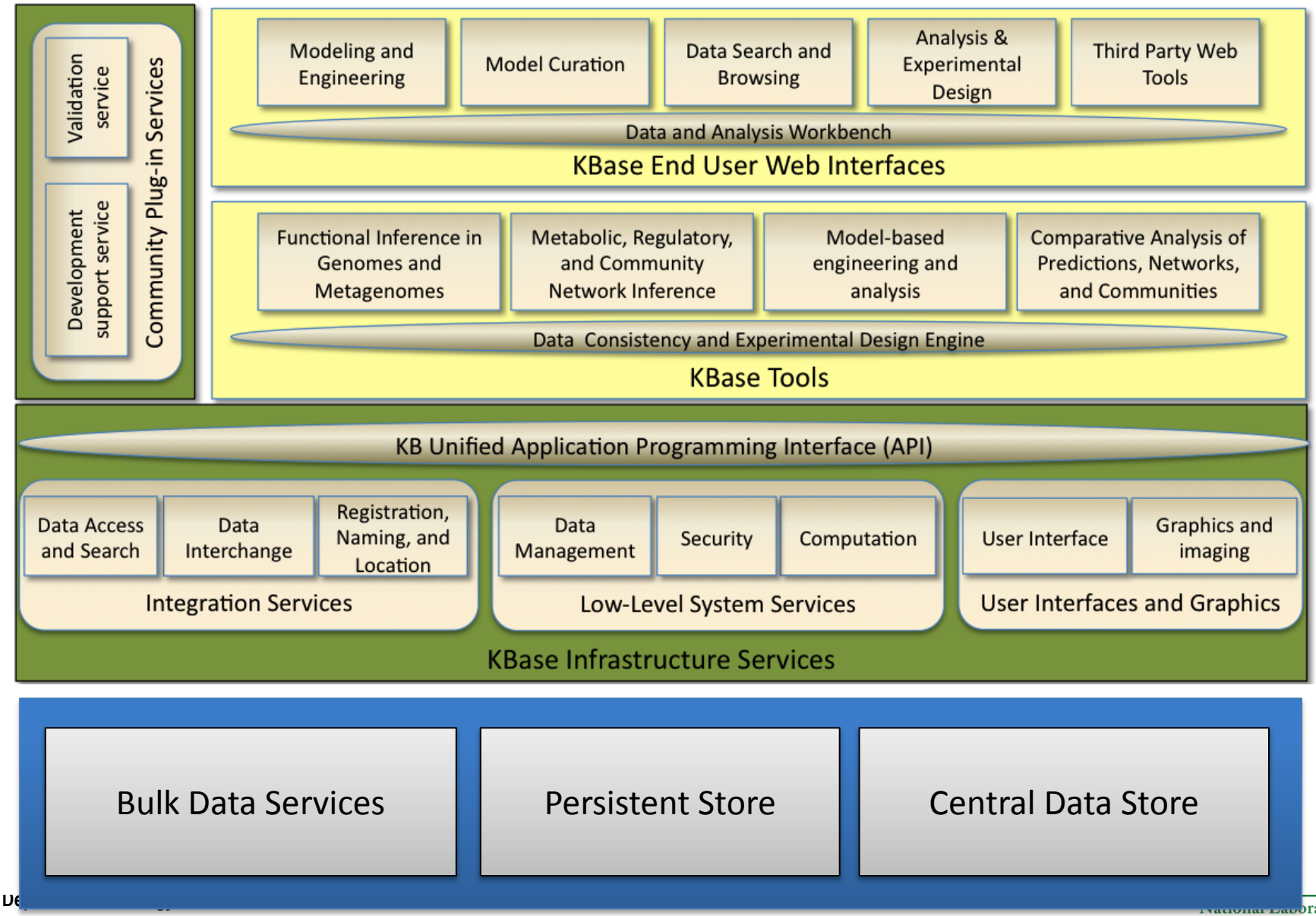
M-C Jenkins (http://www.scienceforseo.com)

Knowledge = Models

- The Knowledgebase should learn models from data and human interaction.
- Models and their parts should carry information about the quality of their data, annotations, and predictions.
- Data, protocols, algorithms and models should be subject to both calculated and community quality assessment.

KBASE
predictive biology
DOE Systems Biology Knowledgebase

The KBase Infrastructure and Services



KBASE predictive biology
DOE Systems Biology Knowledgebase

Concept: KBase User Experience

KBASE predictive biology
DOE Systems Biology Knowledgebase

Search | Narrative

Log Off

T | A | E | H | script | publish

12 narratives/10 hypotheses
2 genomes
1 metagenome
1 model
183 data uploads

New narrative | New team

Team management
- Metal reduction project
- Chemotaxis project

Data management upload

Narrative management
- Chemotaxis Study

blog

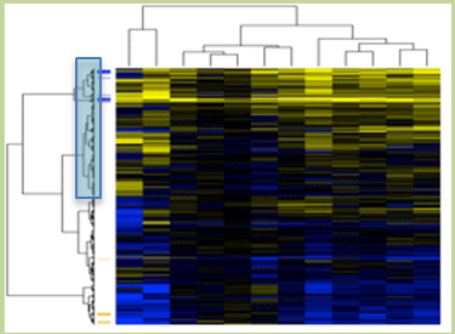
Sensory_genes: gs_aeb123456
Flagellar_genes: gs_aeb123457

[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. Hmmm... 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()



-> -> Add(High_Expression, stdin)

data | function

search

C: ClusterMe Clustering

...

Processing

Concept: Interactive Community Knowledge