Integrative Genomic Analysis

Sharing Data, Tools and Models

Use of bionetworks to build better maps of disease

Stephen Friend MD PhD

Sage Bionetworks (Non-Profit Organization)
Seattle/ Beijing/ San Francisco

NAS
March 10th, 2011

BETTER MAPS OF DISEASE

NOT JUST WHAT WE DO BUT HOW WE DO IT

POWER OF BUILDING A PRE-COMPETITIVE
COMMONS FOR EVOLVING
GENERATIVE MODELS OF DISEASE

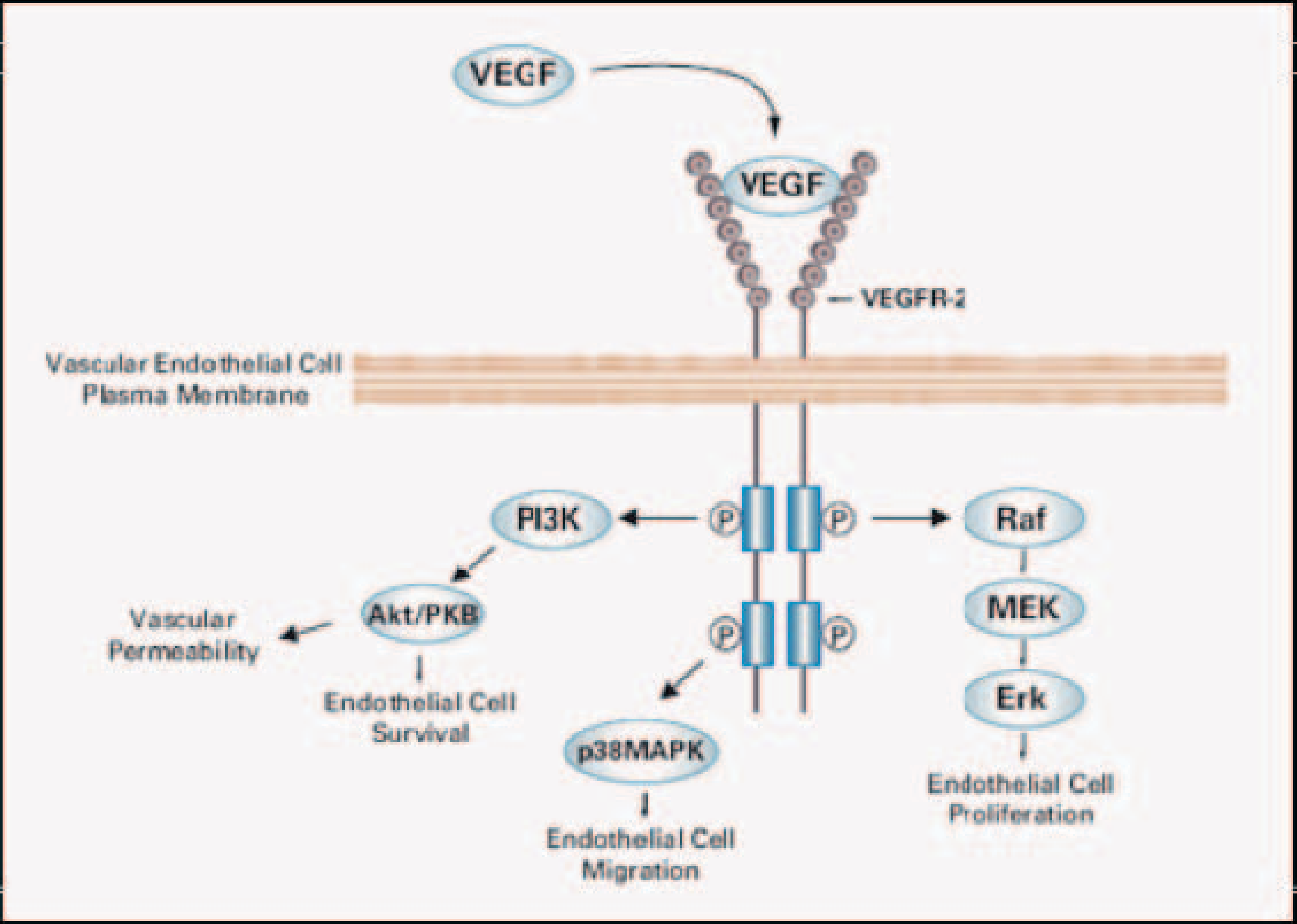
Existing approaches and issues in Drug Discovery

Current costs for drug approval- ~$1Billion – 5 -10 years

Only 6% of therapies in Phase I trials will lead to approval (CMS)

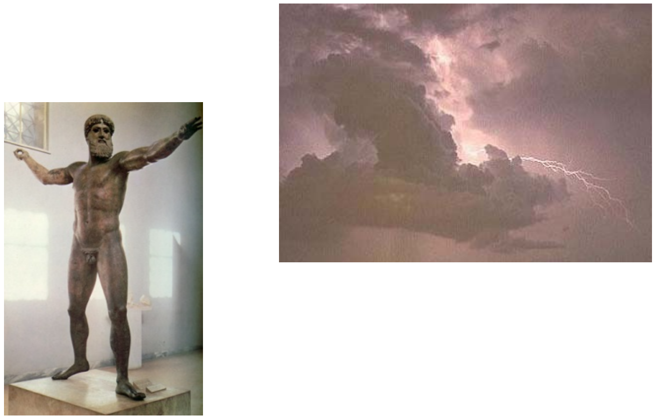
Cancer- FDA approved marketed drugs as ”standards of care” provide significant impact in only 25% of patients

VEGFR Classical Pathway

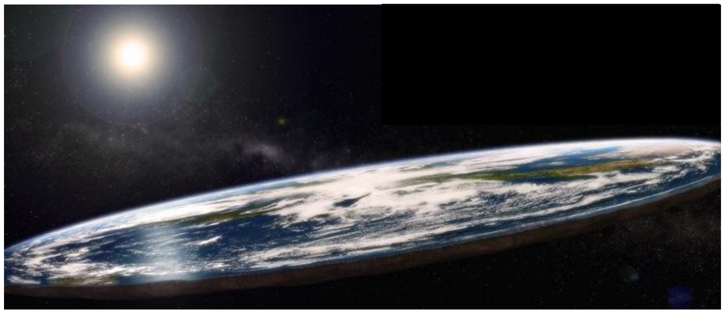


Where the data-driven network approach can combat the want our minds' have for story telling

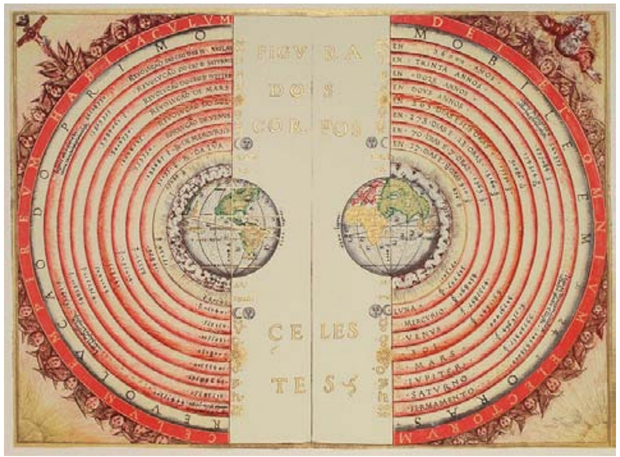
Zeus, the sky god; when he is angry he throws lightening bolts out of the sky



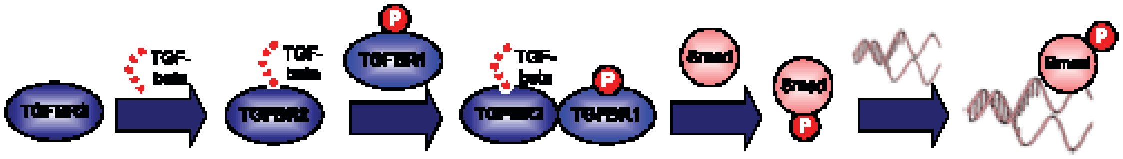
The earth is flat

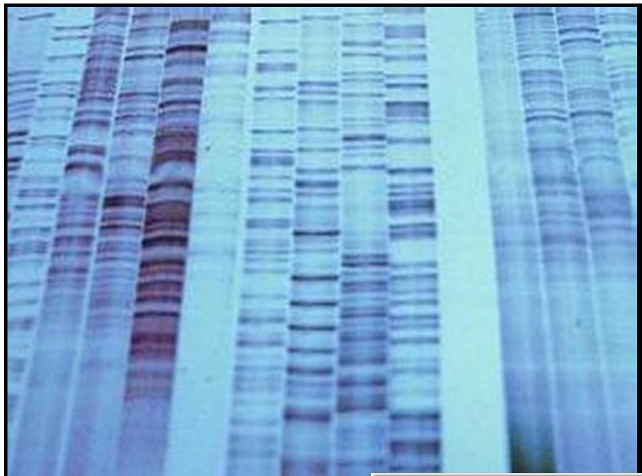


Ptolemaic astronomy: the earth is the center of the universe

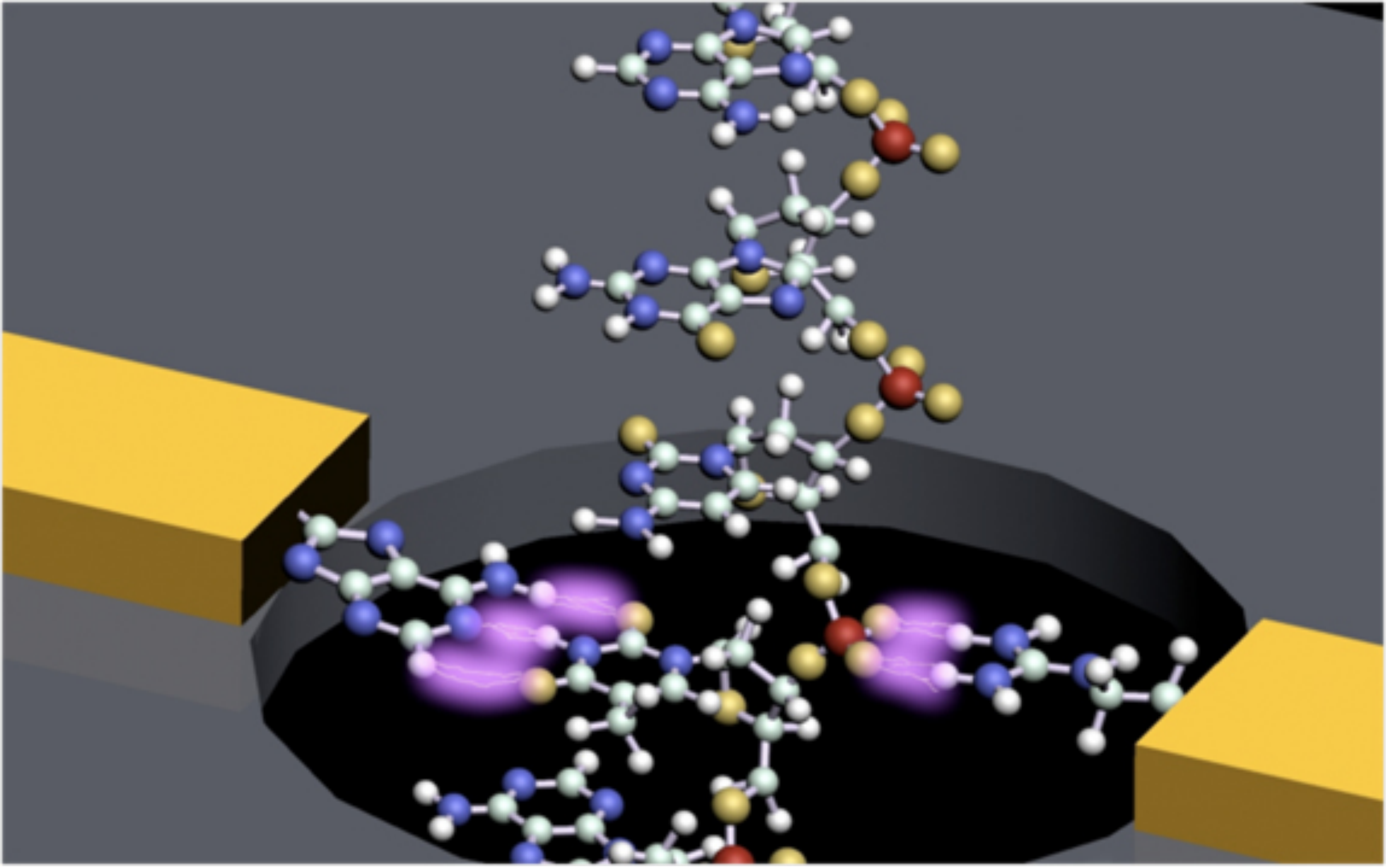


Biological processes are driven by simple linearly ordered pathways (TGF-beta signaling)

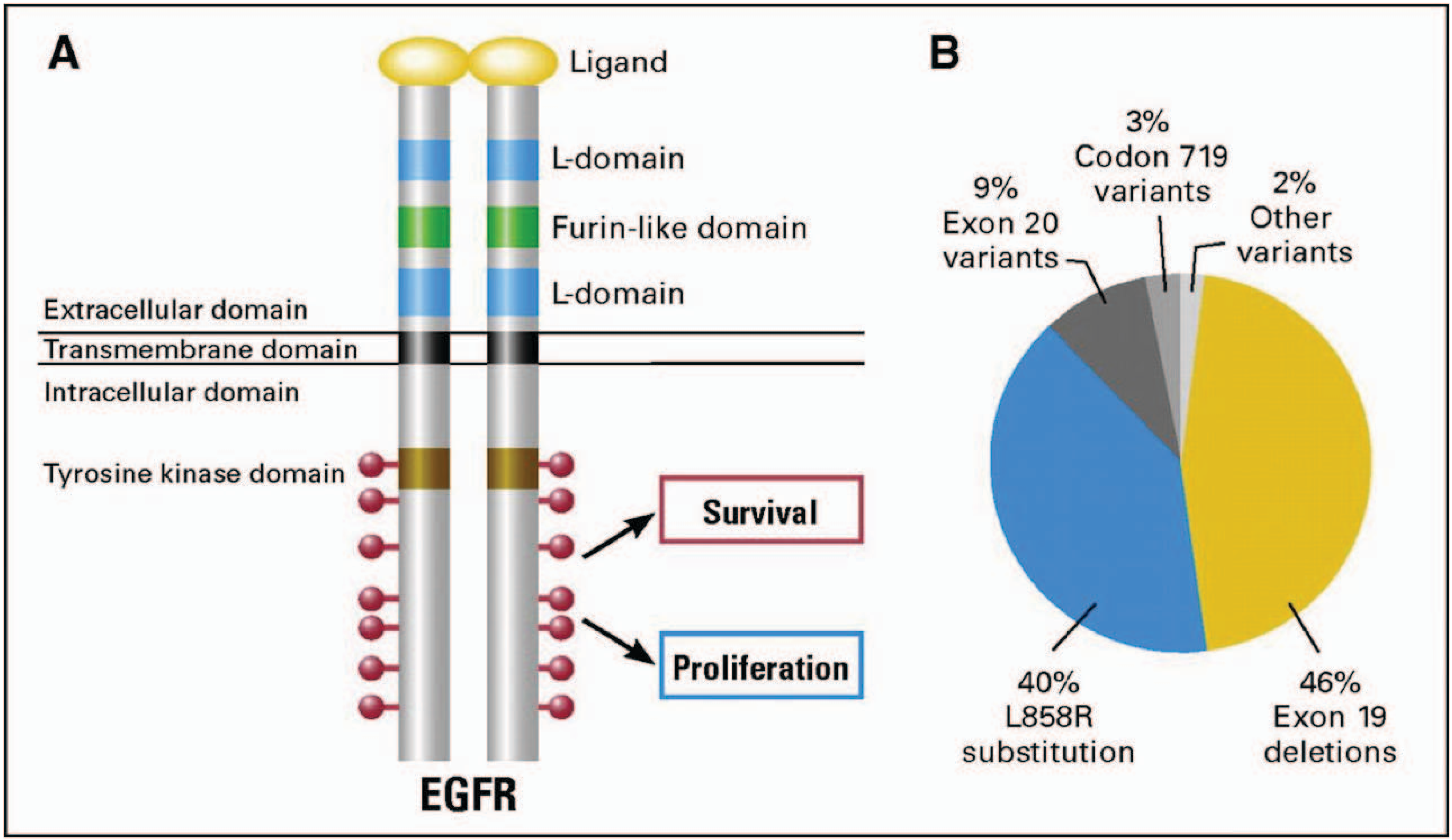




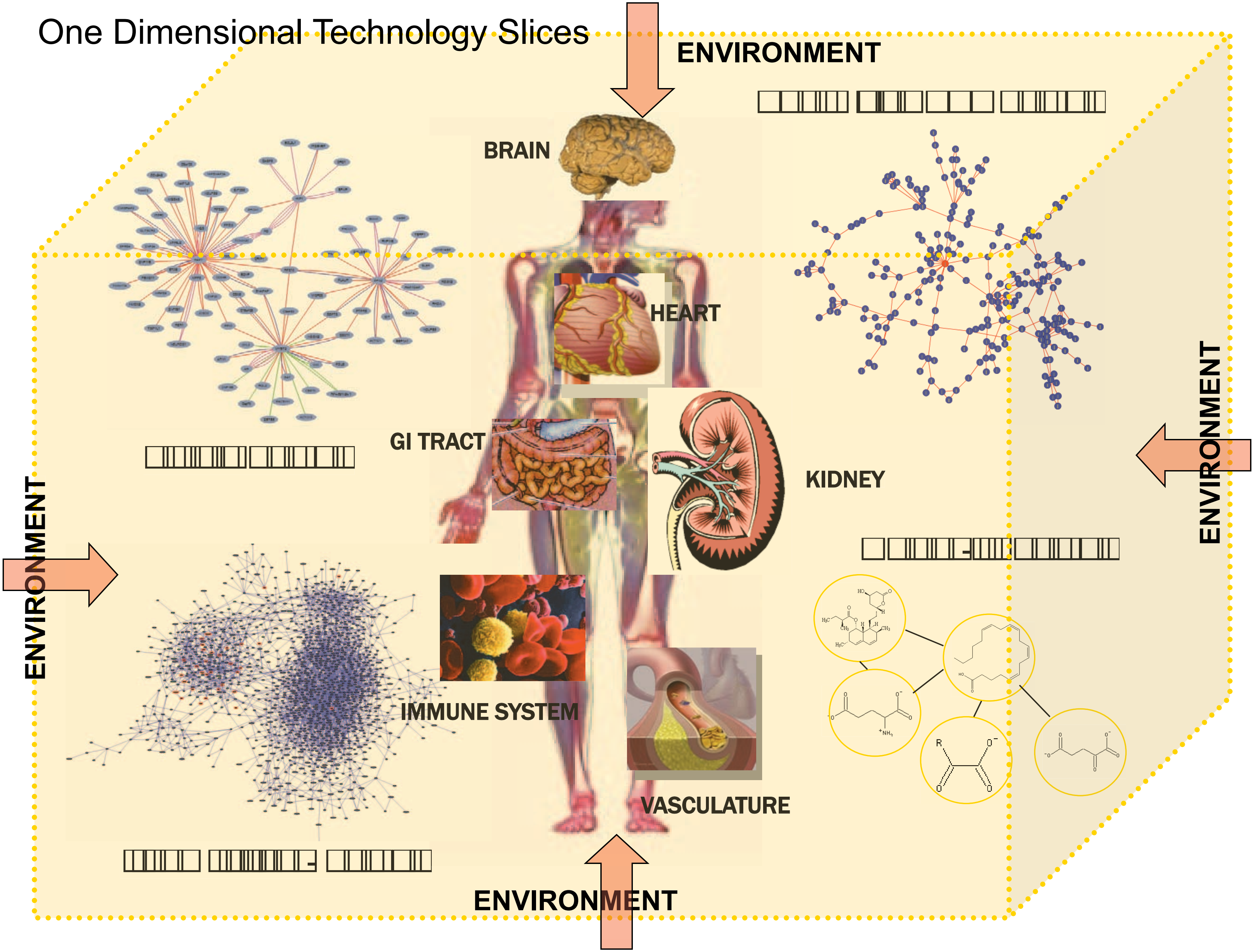
Amazing Technologies
will soon create a revolution
in our understanding of diseases
resulting in new categories, and therapies



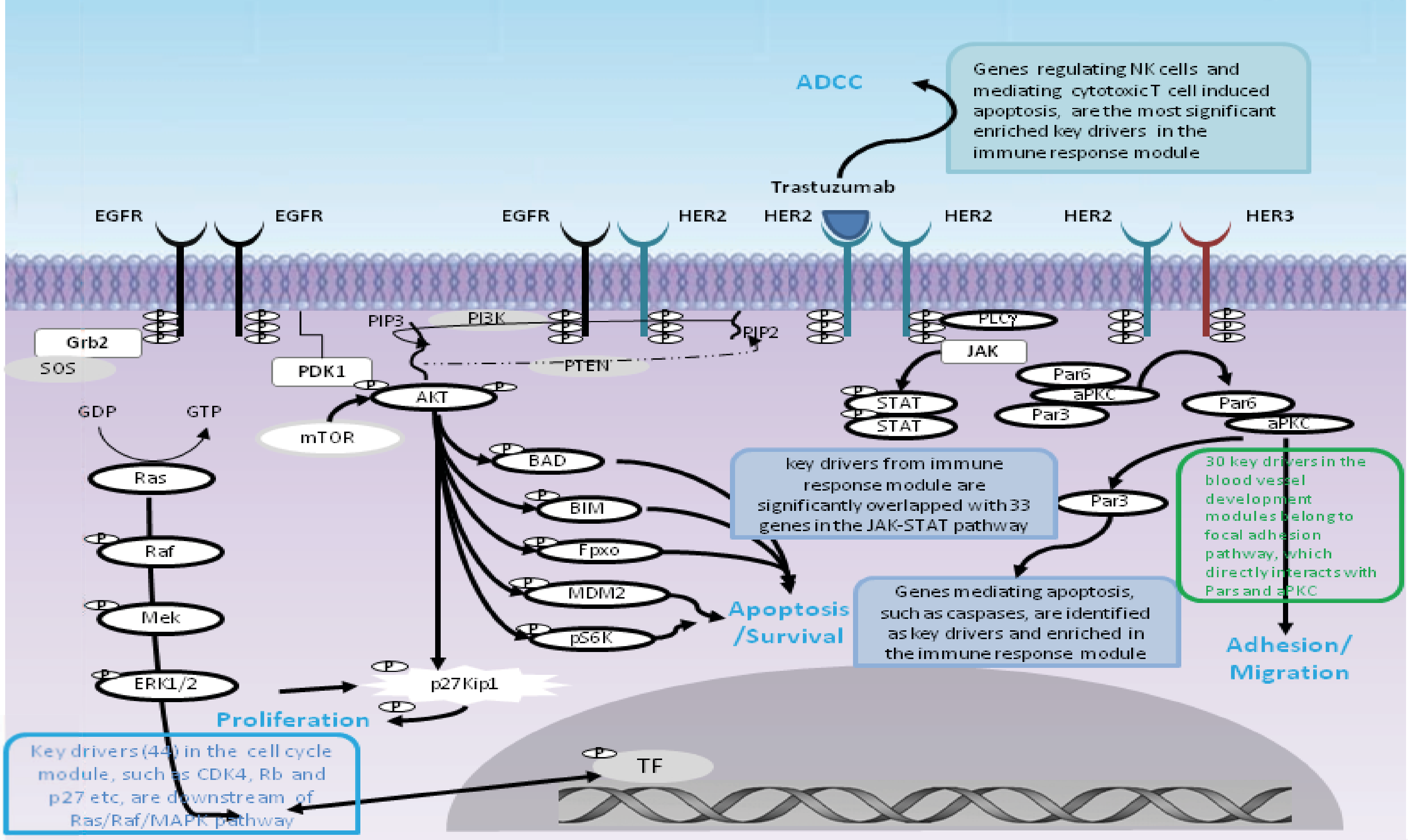
Personalized Medicine 101:
Capturing Single bases pair mutations = ID of responders



One Dimensional Technology Slices



Example of Complexity: Overlapping of EGFR and Her2 Pathways



Genes regulating NK cells and mediating cytotoxic T cell induced apoptosis, are the most significant enriched key drivers in the immune response module

key drivers from immune response module are significantly overlapped with 33 genes in the JAK-STAT pathway

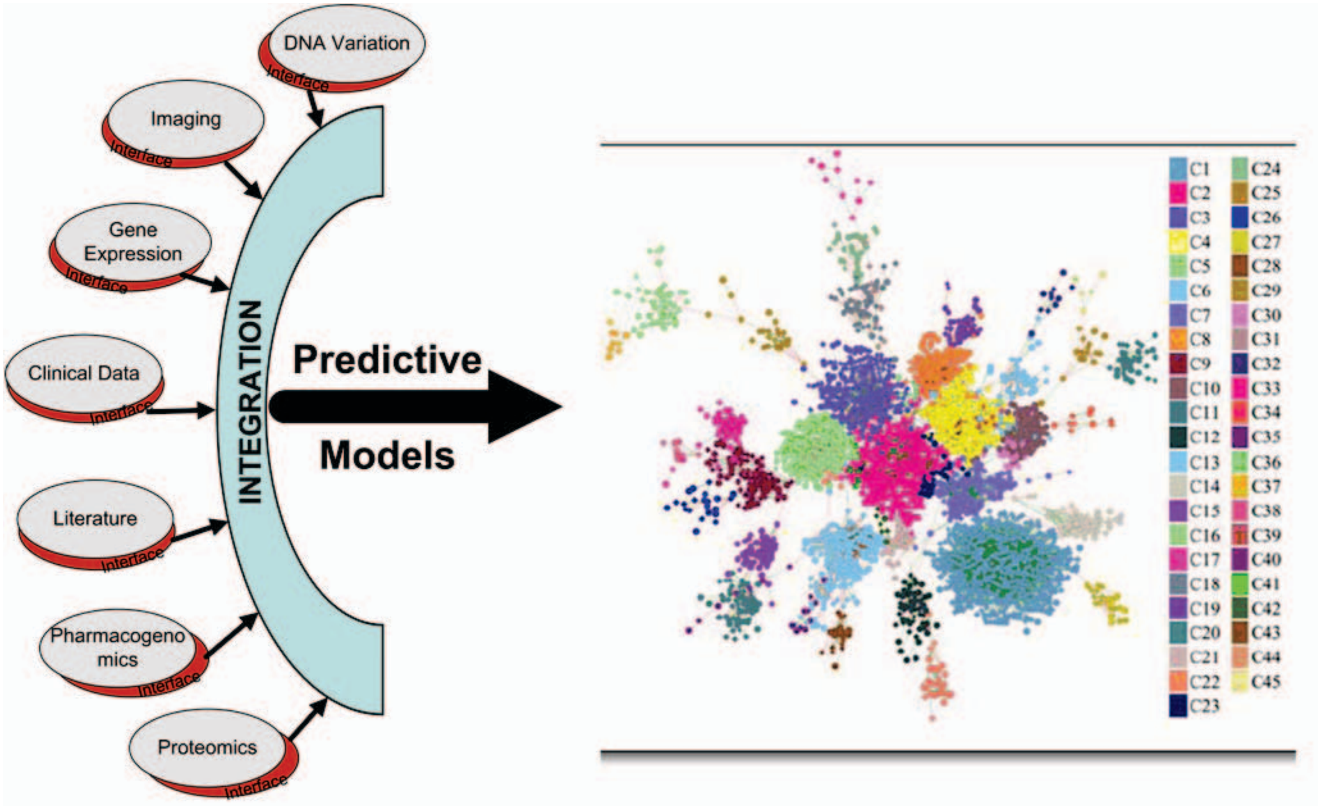
Genes mediating apoptosis, such as caspases, are identified as key drivers and enriched in the immune response module

30 key drivers in the blood vessel development modules belong to focal adhesion pathway, which directly interacts with Pars and aPKC

Key drivers (44) in the cell cycle module, such as CDK4, Rb and p27 etc, are downstream of Ras/Raf/MAPK pathway

The “Rosetta Integrative Genomics Experiment”: Generation, assembly, and integration of data to build models that predict clinical outcome

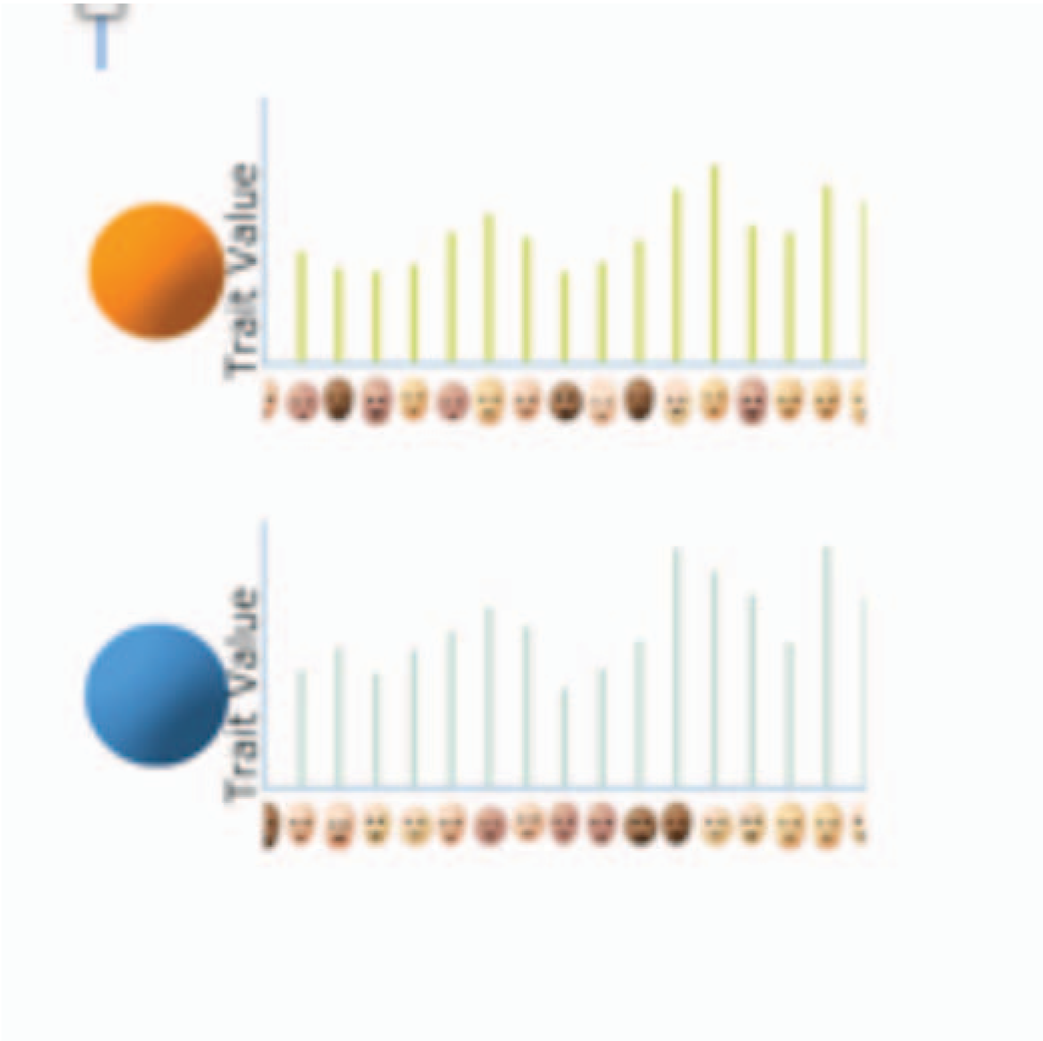
Merck Inc. Co.
5 Year Program
Based at Rosetta
Total Resources
>$150M



- **Generate data need to build**
- **bionetworks**
- **Assemble other available data useful for building networks**
- **Integrate and build models**
- **Test predictions**
- **Develop treatments**
- **Design Predictive Markers**

We need to carry out comprehensive monitoring of many traits at the population level

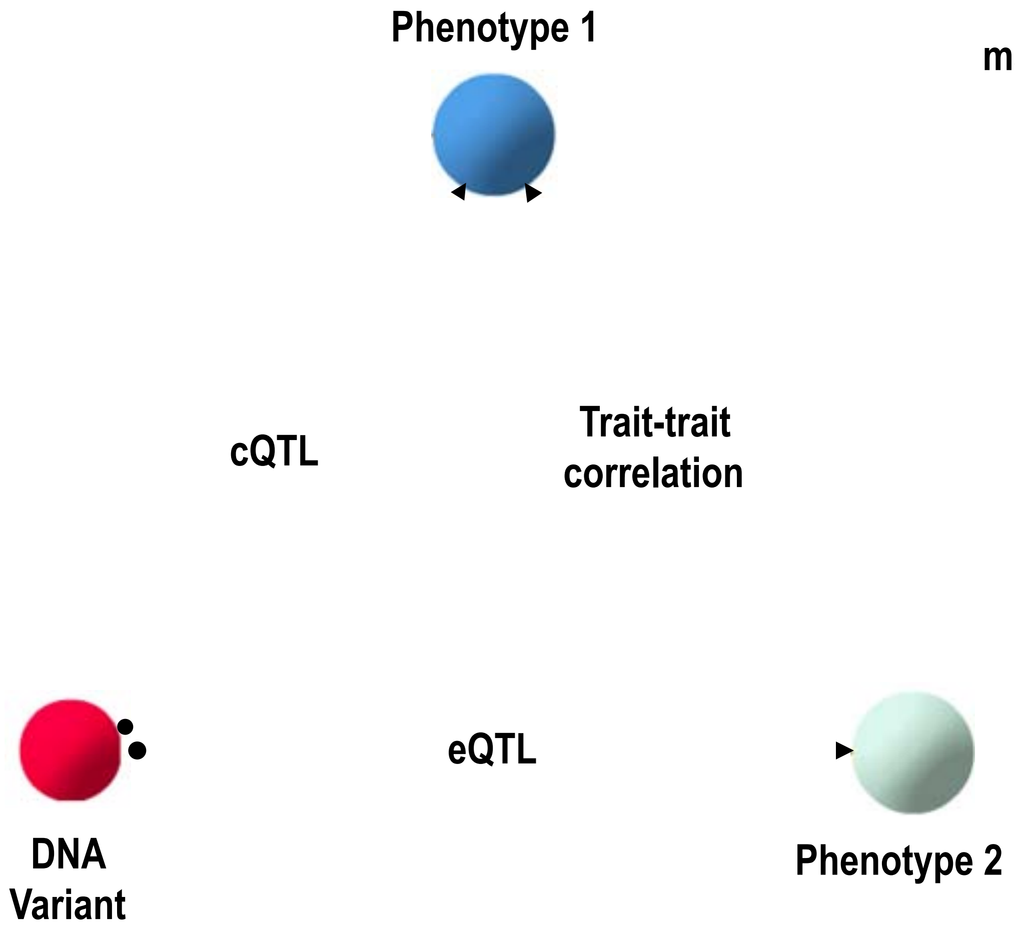
Monitor disease and molecular traits in populations



● **Putative causal gene**

● **Disease trait**

Leveraging DNA variation as a systematic perturbation source



Phenotype 1

cQTL

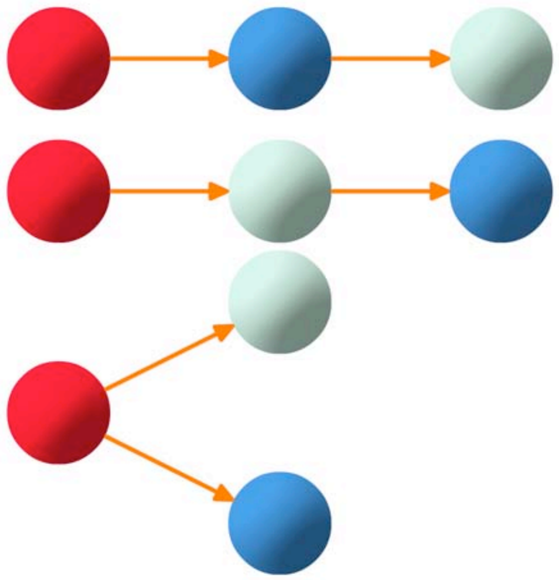
Trait-trait correlation

DNA Variant

eQTL

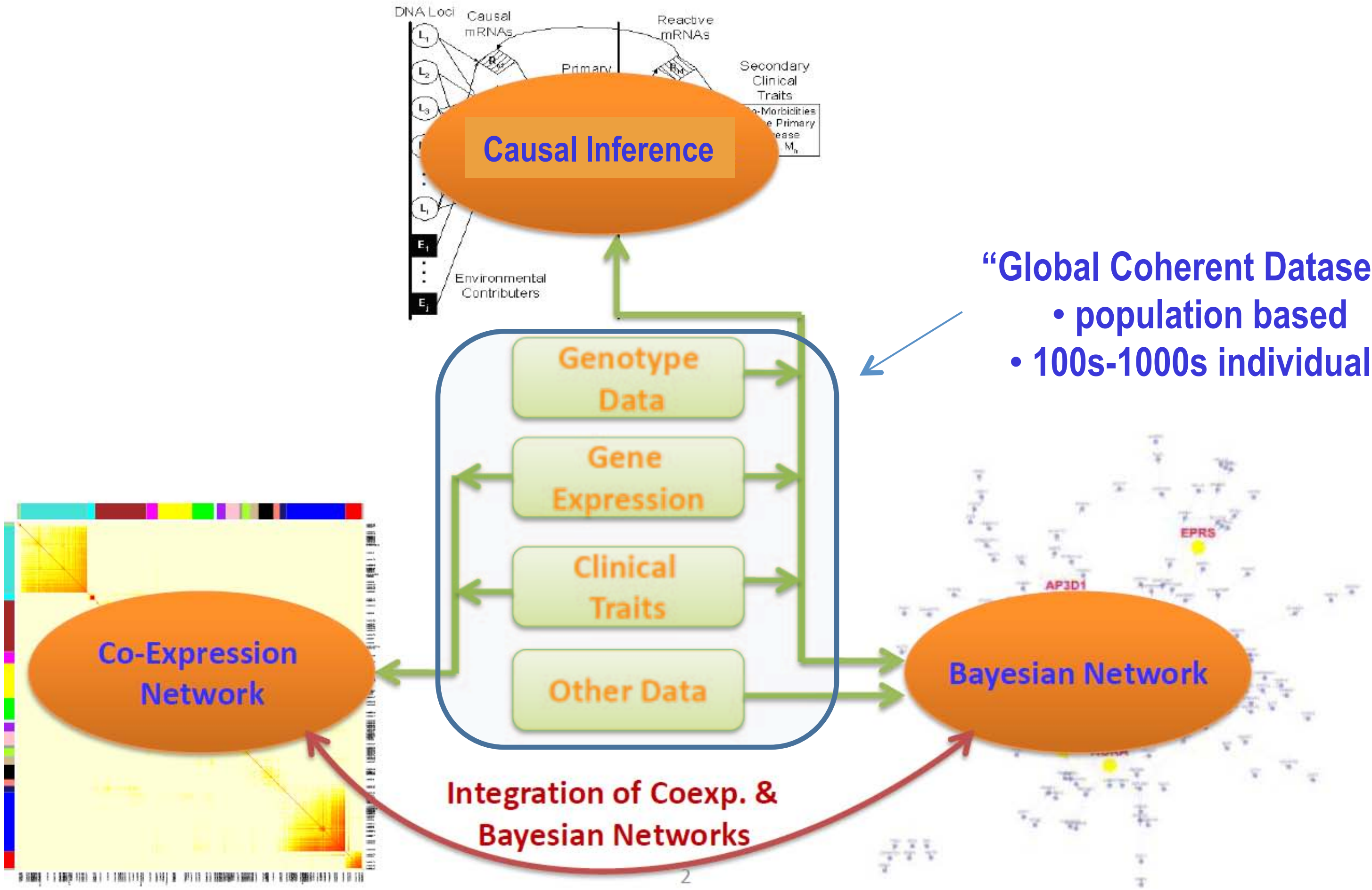
Phenotype 2

Given this triangle relationship, we can mathematically infer the most likely relationship



Integration of Genotypic, Gene Expression & Trait Data

Schadt et al. Nature Genetics 37: 710 (2005)
Millstein et al. BMC Genetics 10: 23 (2009)



Causal Inference

Genotype Data

Gene Expression

Clinical Traits

Other Data

Co-Expression Network

Bayesian Network

Integration of Coexp. & Bayesian Networks

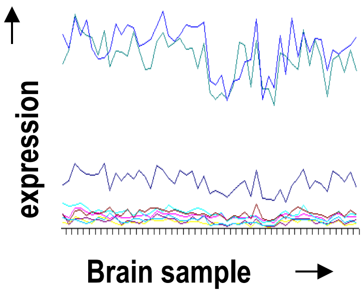
“Global Coherent Datasets”
- population based
- 100s-1000s individuals

Chen et al. Nature 452:429 (2008)
Zhang & Horvath. Stat.Appl.Genet.Mol.Biol. 4: article 17 (2005)

Zhu et al. Cytogenet Genome Res. 105:363 (2004)
Zhu et al. PLoS Comput. Biol. 3: e69 (2007)

Constructing Co-expression Networks

Start with expression measures for ~13K genes most variant genes across 100-150 samples



Establish a 2D correlation matrix for all gene pairs

	1	2	3	4
1	1	0.8	0.2	-0.8
2	0.8	1	0.1	-0.6
3	0.2	0.1	1	-0.1
4	-0.8	-0.6	-0.1	1

Correlation Matrix

Note: NOT a gene expression heatmap

Define Threshold eg >0.6 for edge

	1	2	3	4
1	1	1	0	1
2	1	1	0	1
3	0	0	1	0
4	1	1	0	1

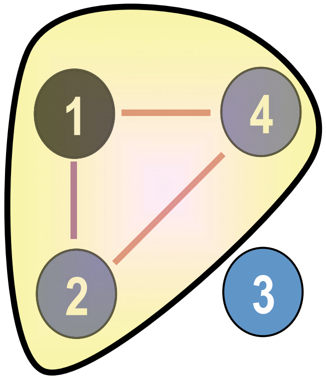
Connection Matrix

Hierarchically cluster

	1	2	4	3
1	1	1	1	0
2	1	1	1	0
4	1	1	1	0
3	0	0	0	1

Clustered Connection Matrix

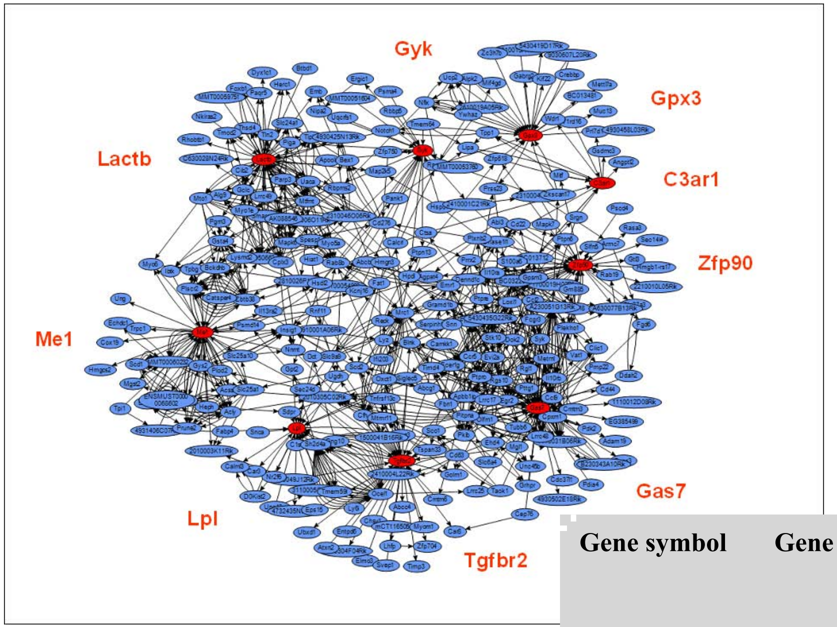
Identify modules



Network Module

sets of genes for which many pairs interact (relative to the total number of pairs in that set)

Probabalistic Models- Rosetta



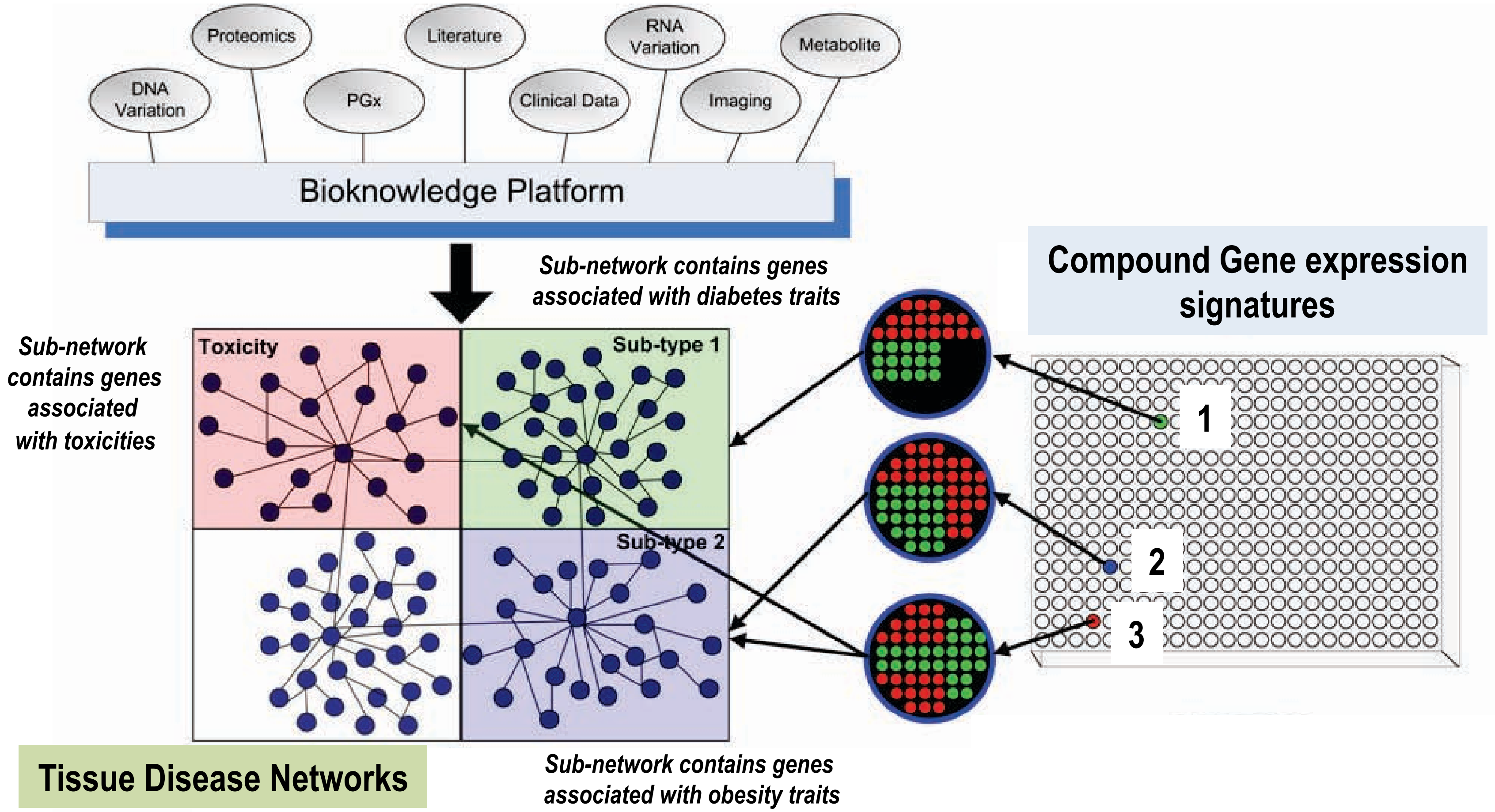
Networks facilitate direct identification of genes that are causal for disease

Evolutionarily tolerated weak spots

Gene symbol	Gene name	Variance of OFPM explained by gene expression*	Mouse model	Source
Zfp90	Zinc finger protein 90	68%	tg	Constructed using BAC transgenics
Gas7	Growth arrest specific 7	68%	tg	Constructed using BAC transgenics
Gpx3	Glutathione peroxidase 3	61%	tg	Provided by Prof. Oleg Mirochnitchenko (University of Medicine and Dentistry at New Jersey, NJ) [12]
Lactb	Lactamase beta	52%	tg	Constructed using BAC transgenics
Me1	Malic enzyme 1	52%	ko	Naturally occurring KO
Gyk	Glycerol kinase	46%	ko	Provided by Dr. Katrina Dipple (UCLA) [13]
Lpl	Lipoprotein lipase	46%	ko	Provided by Dr. Ira Goldberg (Columbia University, NY) [11]
C3ar1	Complement component 3a receptor 1	46%	ko	Purchased from Deltagen, CA
Tgfbr2	Transforming growth factor beta receptor 2	39%	ko	Purchased from Deltagen, CA

Nat Genet (2005) 205:370

Map compound signatures to disease networks



Sub-network contains genes associated with diabetes traits

Sub-network contains genes associated with toxicities

Compound Gene expression signatures

Tissue Disease Networks

Sub-network contains genes associated with obesity traits

Compound 1: Drug signature significantly enriched in subnetwork associated with diabetes traits

Compound 2: Drug signature significantly enriched in subnetwork associated with obesity traits

Compound 3: Drug signature significantly enriched in subnetwork associated with obesity traits BUT also in subnetwork associated with toxicities

Extensive Publications now Substantiating Scientific Approach
Probabilistic Causal Bionetwork Models

- **>80 Publications from Rosetta Genetics / Sage Group (~30 scientists) over 6 years including high profile papers in PLoS Nature and Nature Genetics**

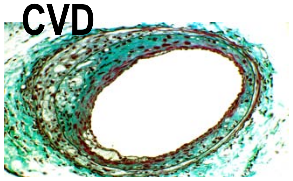
"*Genetics of gene expression surveyed in maize, mouse and man.*" **Nature**. (2003)

"*Variations in DNA elucidate molecular networks that cause disease.*" **Nature**. (2008)

"*Genetics of gene expression and its effect on disease.*" **Nature**. (2008)

"*Validation of candidate causal genes for obesity that affect...*" **Nat Genet**. (2009)

….. Plus 10 additional papers in Genome Research, PLoS Genetics, PLoS Comp.Biology, etc



***"Identification of pathways for atherosclerosis."* Circ Res. (2007)**

"*Mapping the genetic architecture of gene expression in human liver.*" **PLoS Biol**. (2008)

…… Plus 5 additional papers in Genome Res., Genomics, Mamm.Genome

"*Integrating genotypic and expression data …for bone traits…*" **Nat Genet**. (2005)

"*..approach to identify candidate genes regulating BMD…*" **J Bone Miner Res**. (2009)

***"An integrative genomics approach to infer causal associations ..."* Nat Genet. (2005)**

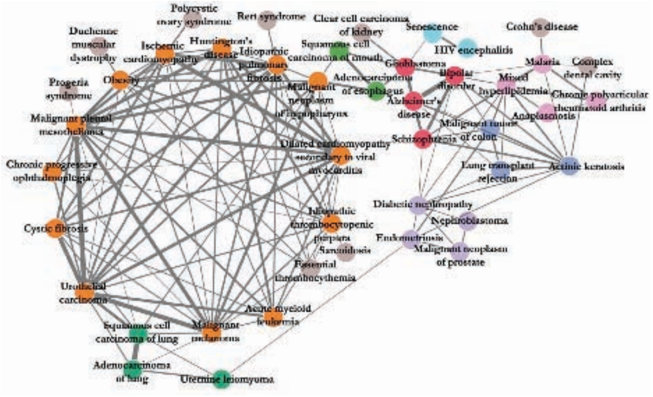
***"Increasing the power to detect causal associations*… "PLoS Comput Biol. (2007)**

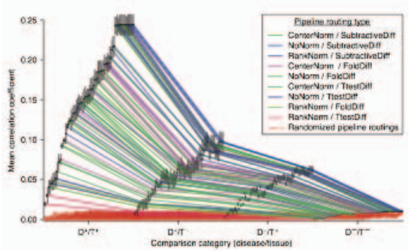
"*Integrating large-scale functional genomic data ...*" **Nat Genet.** (2008)

…… Plus 3 additional papers in **PLoS Genet., BMC Genet.**

Exploring the Global Landscape of Human Disease Through Public Data- approaches taken by Atul Butte

Public data enables quantitative disease relationships


Polycystic ovary syndrome
Rett syndrome
Clear cell carcinoma of kidney
Senescence
Crohn's disease
Duchenne muscular dystrophy
Ischemic cardiomyopathy
Huntington's disease
Idiopathic pulmonary fibrosis
Squamous cell carcinoma of mouth
HIV encephalitis
Malaria
Complex dental cavity
Progeria syndrome
Obesity
Malignant neoplasm of oropharynx
Adenocarcinoma of esophagus
Glioblastoma
Bipolar disorder
Mixed hyperlipidemia
Chronic polyarticular rheumatoid arthritis
Malignant pleural mesothelioma
Alzheimer's disease
Anaplasmosis
Malignant tumor of colon
Schizophrenia
Dilated cardiomyopathy secondary to viral myocarditis
Chronic progressive ophthalmoplegia
Lung transplant rejection
Actinic keratosis
Diabetic nephropathy
Nephroblastoma
Idiopathic thrombocytopenic purpura
Cystic fibrosis
Endometriosis
Sarcoidosis
Malignant neoplasm of prostate
Essential thrombocythemia
Urothelial carcinoma
Acute myeloid leukemia
Squamous cell carcinoma of lung
Malignant melanoma
Adenocarcinoma of lung
Uterine leiomyoma



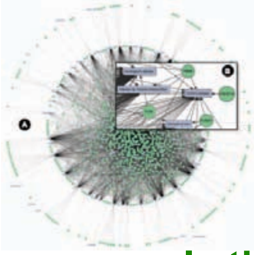
Joel Dudley et al.. Molecular systems biology (2009) vol. 5 pp. 307

High quality signals exist in public data

Differences

Commonalities

Plasma proteome networks

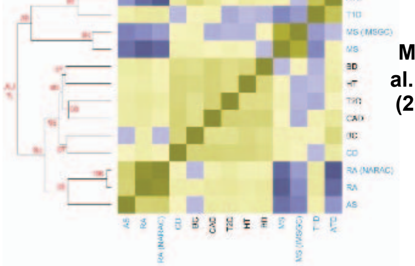


Joel Dudley and Atul Butte. Pacific Symposium on Biocomputing (2009) pp. 27-38

Which biomarkers best discriminate diseases?

Is there a blood biomarker for general pathology?

Genetic architecture of autoimmune diseases

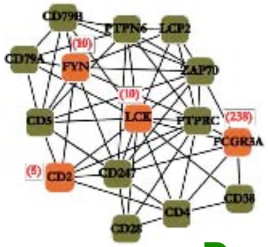


Marina Sirota et al. PLoS genetics (2009) vol. 5 (12) pp. e1000792

Are there genetic “switches” for autoimmunity?

Is there a common autoimmune susceptibility variant?

Functional gene module networks

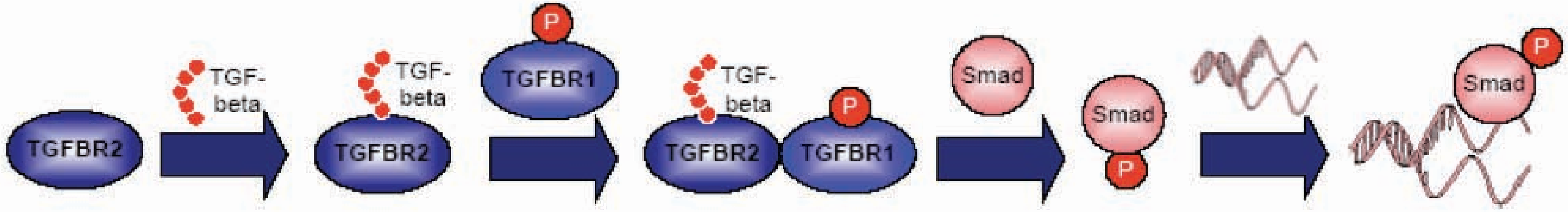


Silpa Suthram et al. PLoS computational biology (2010) vol. 6 (2) pp. e1000662

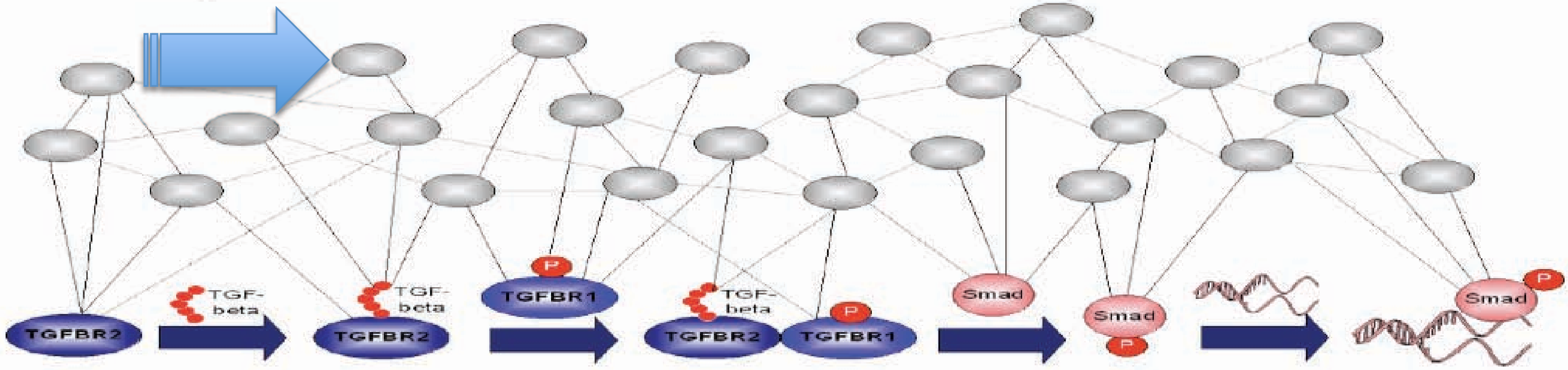
Which modules are unique to metabolic diseases?

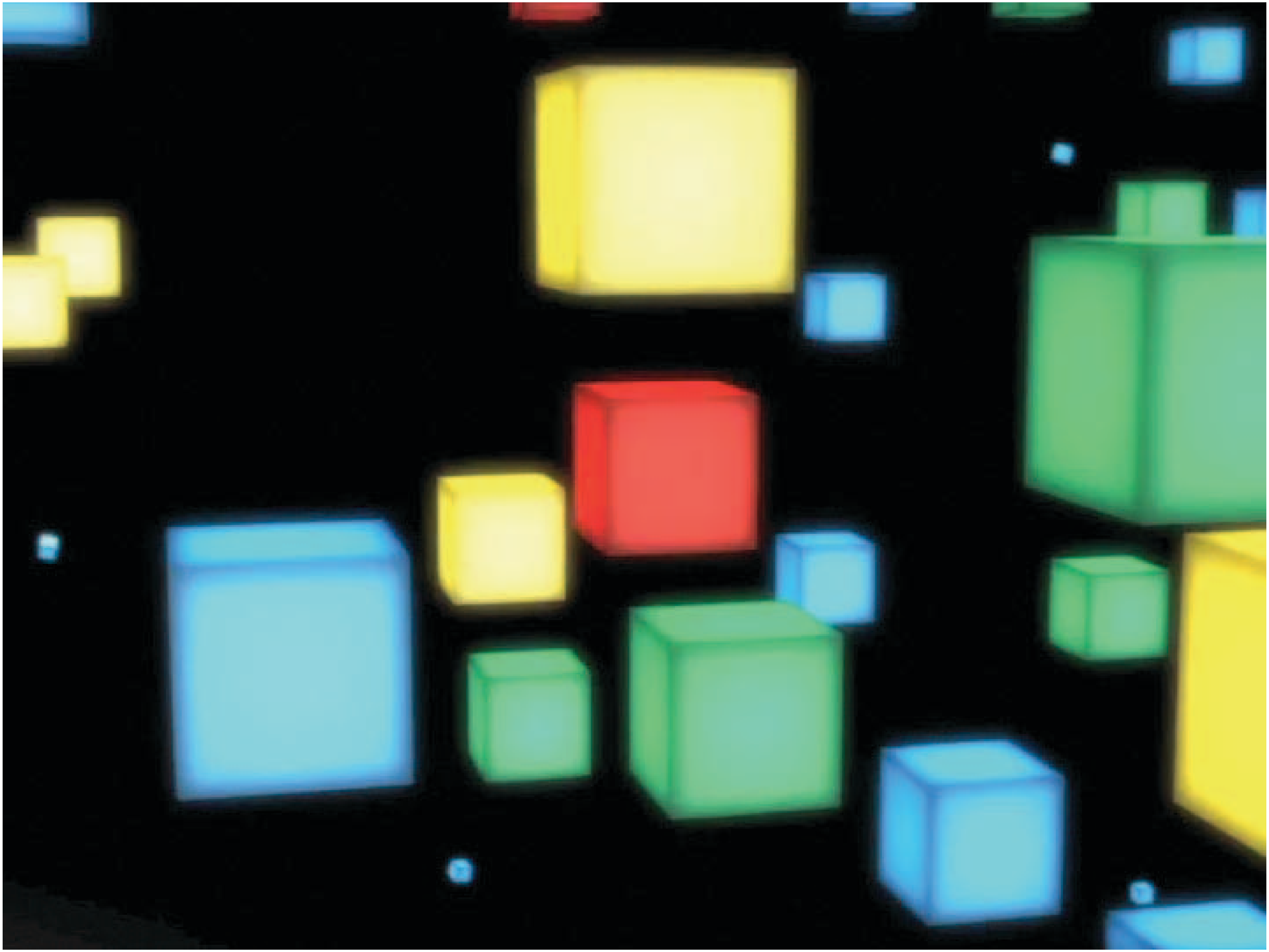
Do common modules harbor pluripotent drug targets?

The way we like to think:

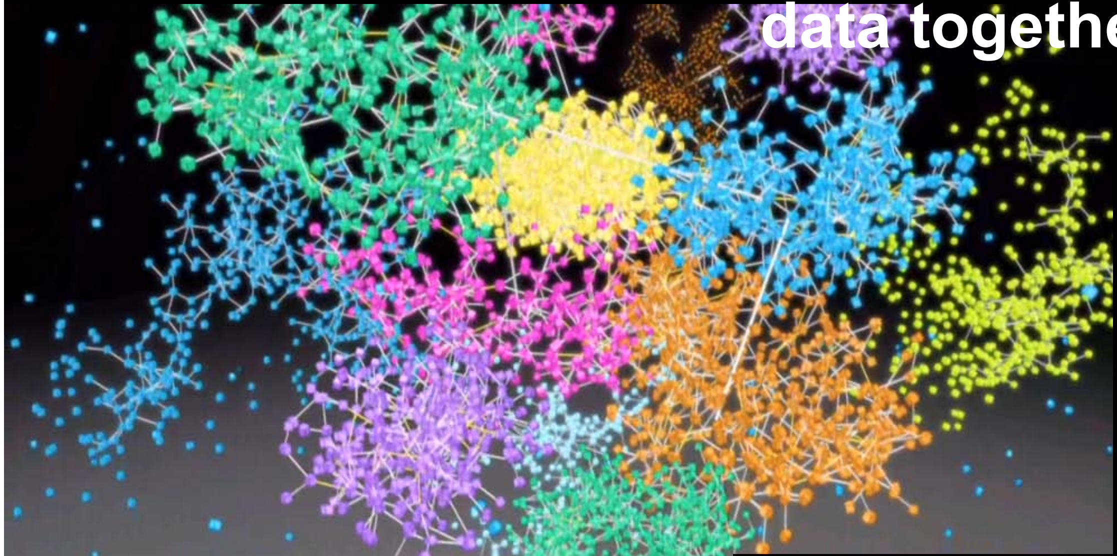


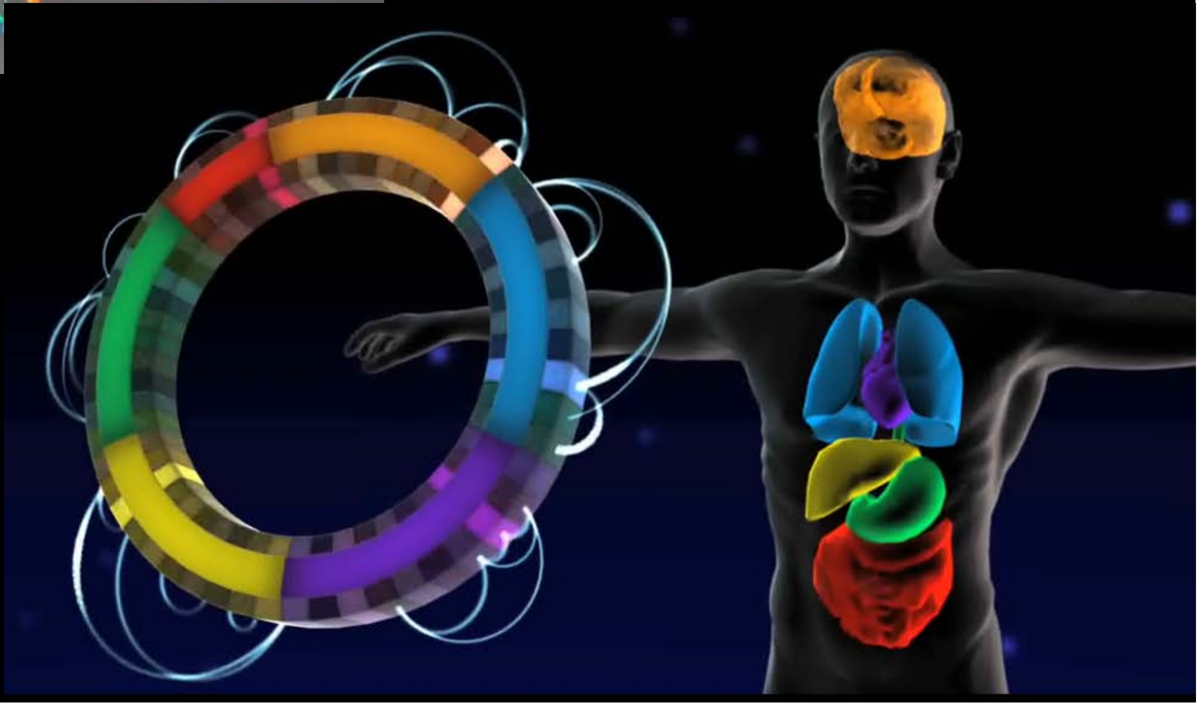
The way it is:





We are about to be able to build maps of disease based on alterations found in multitudes of patients if we can bring the data together



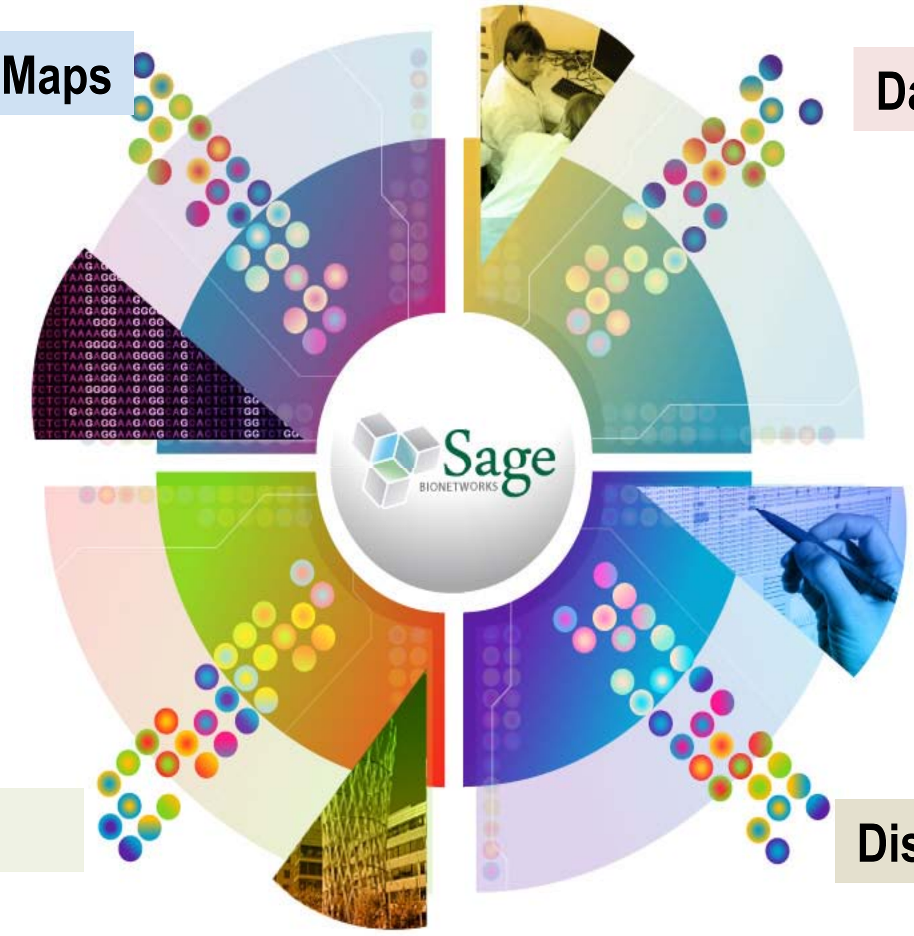


Recognition that the benefits of bionetwork based molecular models of diseases are powerful but that they **require significant resources**

Appreciation that it will **require decades** of evolving representations as real complexity emerges and needs to be integrated with therapeutic interventions

Sage Mission

Sage Bionetworks is a non-profit organization with a vision to create a “commons” where integrative bionetworks are evolved by contributor scientists with a shared vision to accelerate the elimination of human disease



Building Disease Maps

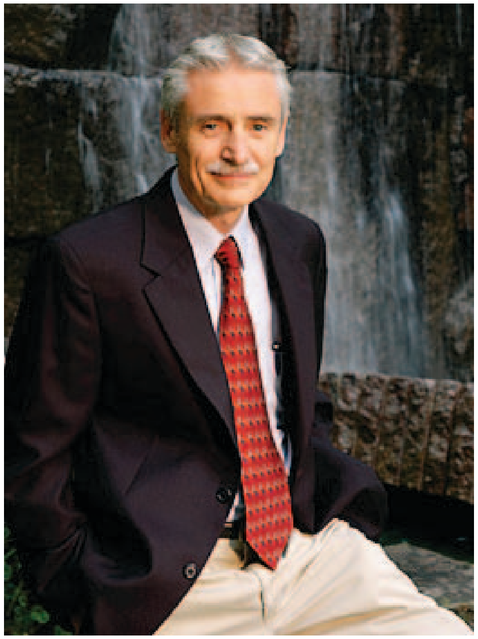
Data Repository

Commons Pilots

Discovery Platform

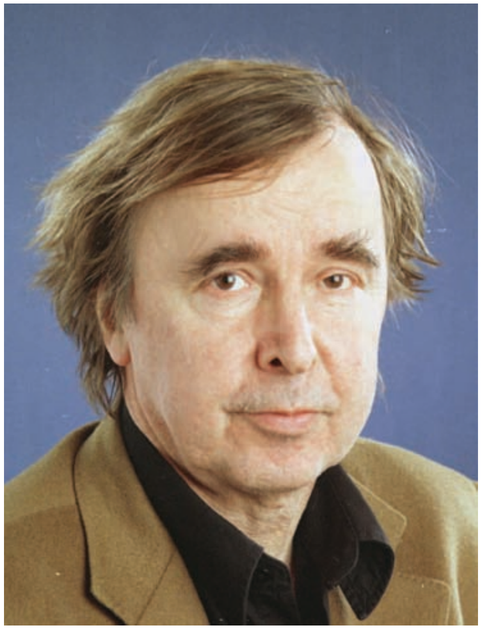
Sagebase.org

Board of Directors- Sage Bionetworks



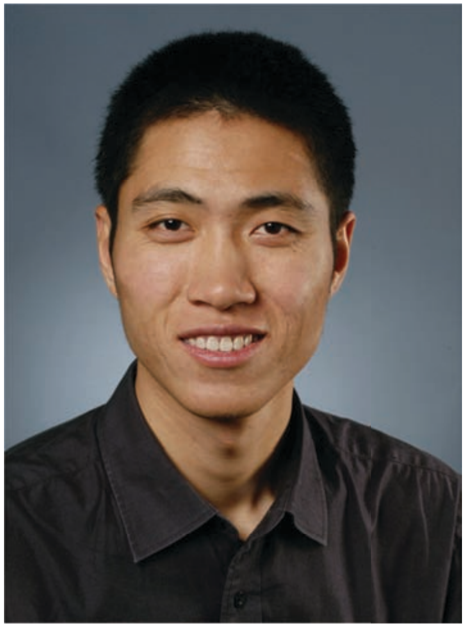
Lee Hartwell

Nobel Laureate
Co-Founder Rosetta



Hans Wizgell

ExPresident Karolinska
Head SAB Rosetta



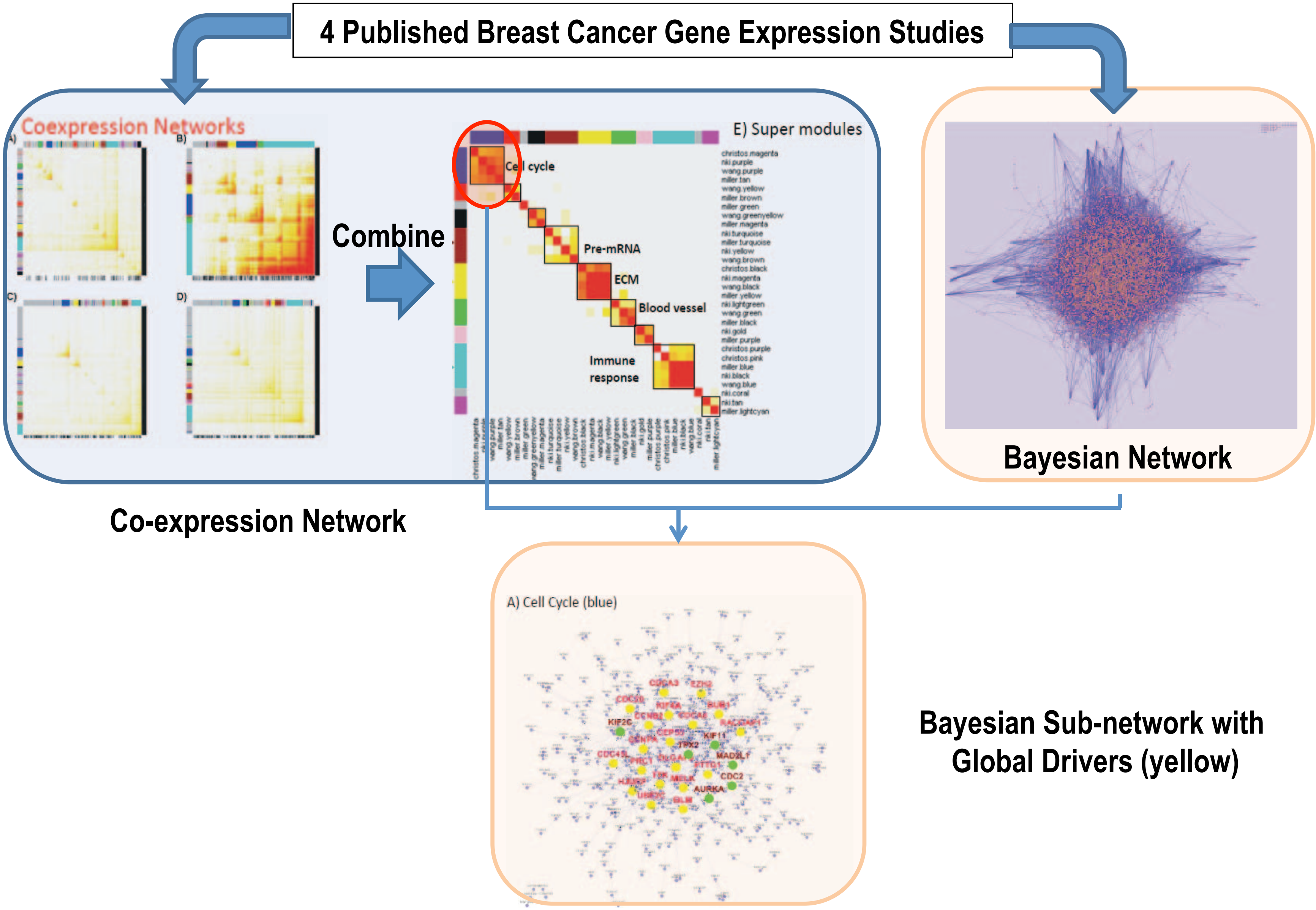
WangJun

Executive Director
BGI

Jeff Hammerbacher

CEO Cloudera
Built and Headed
Facebook
Data Architecture

Example 1: Identification of Molecular Drivers of Breast Cancer



4 Published Breast Cancer Gene Expression Studies

Combine

Co-expression Network

Bayesian Network

Bayesian Sub-network with Global Drivers (yellow)

Example 2. The Sage Non-Responder Project in Cancer

Purpose:

- To identify Non-Responders to approved drug regimens so we can improve outcomes, spare patients unnecessary toxicities from treatments that have no benefit to them, and reduce healthcare costs

Leadership:

- Co-Chairs Stephen Friend, Todd Golub, Charles Sawyers & Rich Schilsky

Initial Studies:

- AML (at first relapse)
- Non-Small Cell Lung Cancer
- Ovarian Cancer (at first relapse)
- Breast Cancer
- Renal Cell
- Multiple Myeloma

Example 3: Clinical Trial Comparator Arm Partnership (CTCAP)

Genetic Alliance

Sage Bionetworks

Johnson & Johnson

Pfizer

gsk GlaxoSmithKline

Roche

Lilly

MERCK

QUINTILES

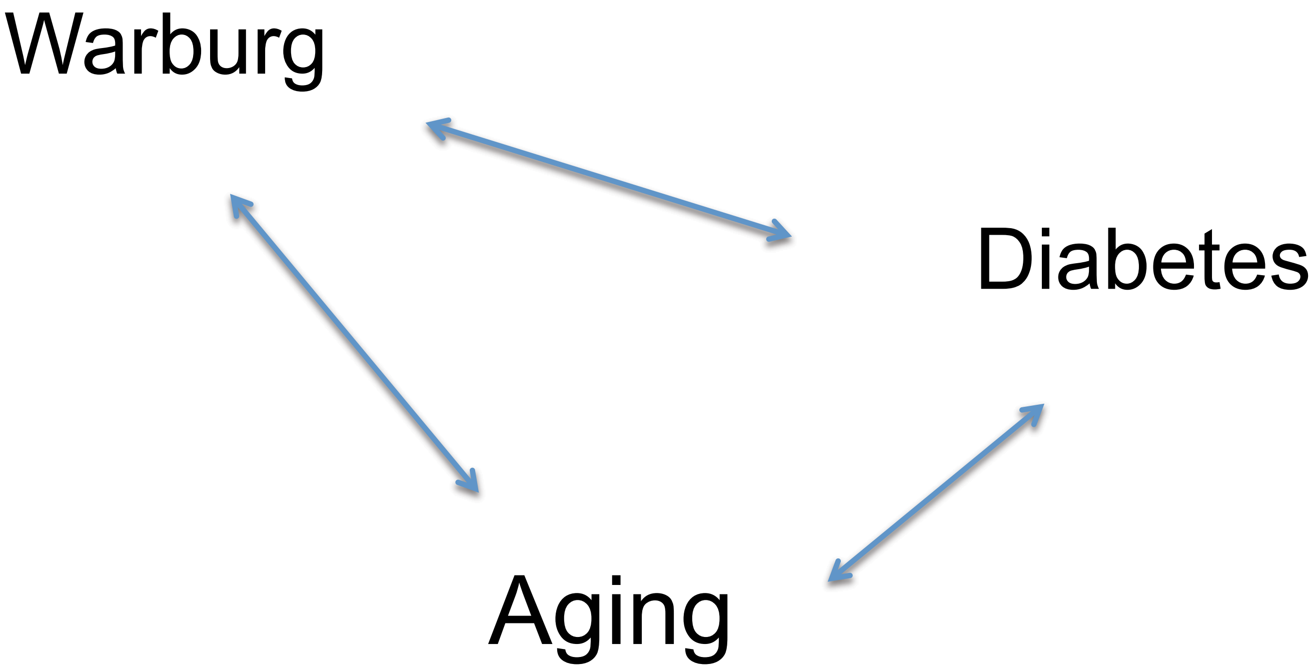
medco

- **Description**: Collate, Annotate, Curate and Host Clinical Trial Data with Genomic Information from the Comparator Arms of Industry and Foundation Sponsored Clinical Trials: Building a Site for Sharing Data and Models to evolve better Disease Maps.
- **Public-Private Partnership** of leading pharmaceutical companies, clinical trial groups and researchers.
- **Neutral Conveners**: Sage Bionetworks and Genetic Alliance [nonprofits].
- **Initiative to share existing trial data** (molecular and clinical) from non-proprietary comparator and placebo arms to create powerful new tool for drug development.

Example 4: The Sage Federation

- Founding Lab Groups
  - Seattle- Sage Bionetworks
  - New York- Columbia: Andrea Califano
  - Palo Alto- Stanford: Atul Butte- (Joel Dudley is here)
  - San Diego- UCSD: Trey Ideker
  - San Francisco: UCSF/Sage: Eric Schadt
- Initial Projects
  - Aging
  - Diabetes
  - Warburg
- Goals: *Share all datasets, tools, models*

  Develop interoperability for human data



Warburg

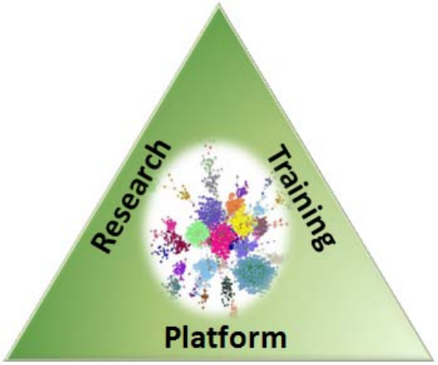
Diabetes

Aging

THE FEDERATION

Butte Califano Friend Ideker Schadt

vs



Community Systems Biologists

Evolving Disease Models

Repository

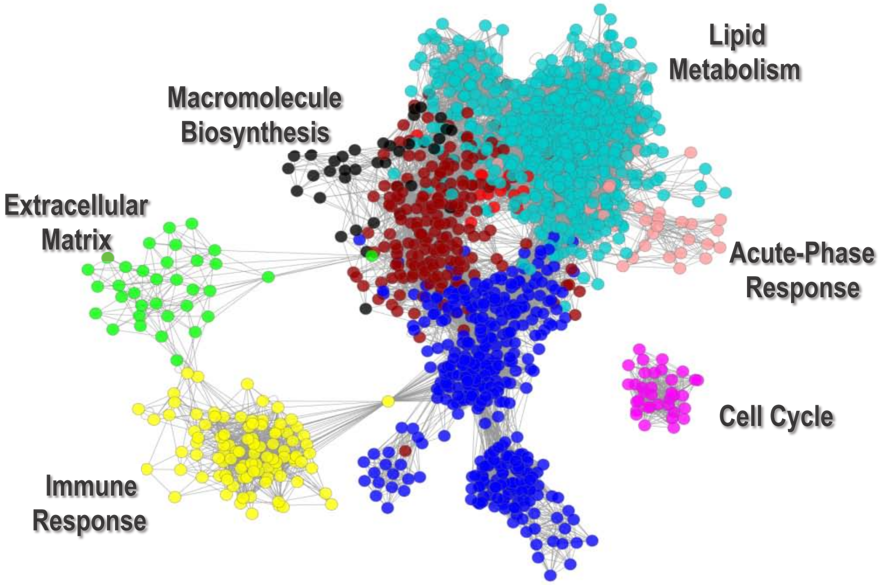
Sage Bionetworks: Platform

GLOBAL COHERENT DATASETS

A data set containing genome-wide DNA variation and intermediate trait, as well as physiological phenotype data across a population of individuals large enough to power association or linkage studies, typically 50 or more individuals. To be coherent, the data needs to be matched with consistent identifiers. Intermediate traits are typically gene expression, but may also include proteomic, metabolomic, and other molecular data.

See http://www.sagebase.org/commons/repository.php

MODELS



TOOLS

Key Driver Analysis (KDA) Tool (R package/Cystoscape plug in)

http://sagebase.org/research/tools.php

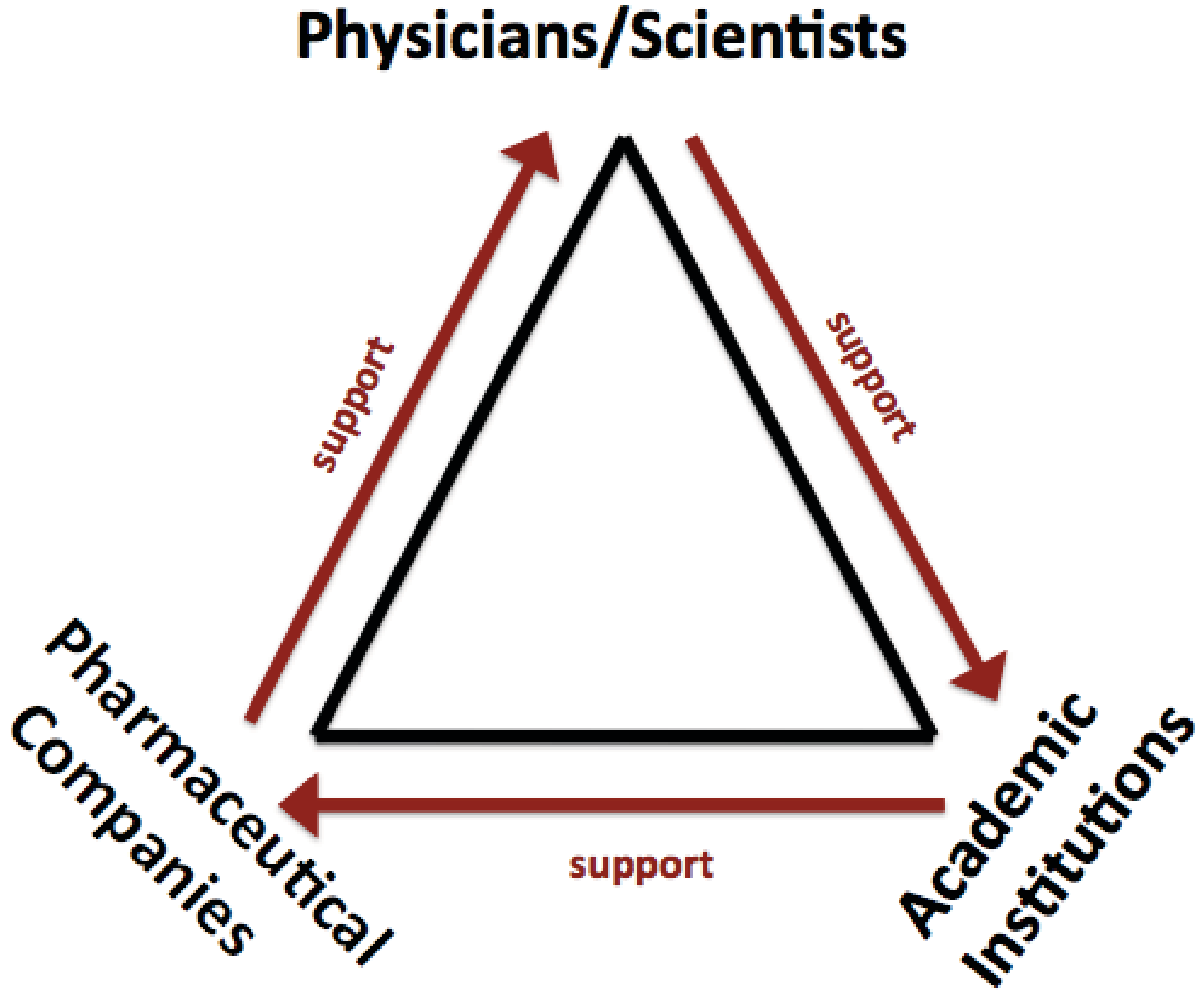
NOT JUST WHAT BUT HOW

Clinical/genomic data
are accessible but minimally usable

Little incentive to annotate and curate
data for other scientists to use

We still consider much clinical research as if we were “hunter gathers”- not sharing

The Medical Industrial Complex



Physicians/Scientists

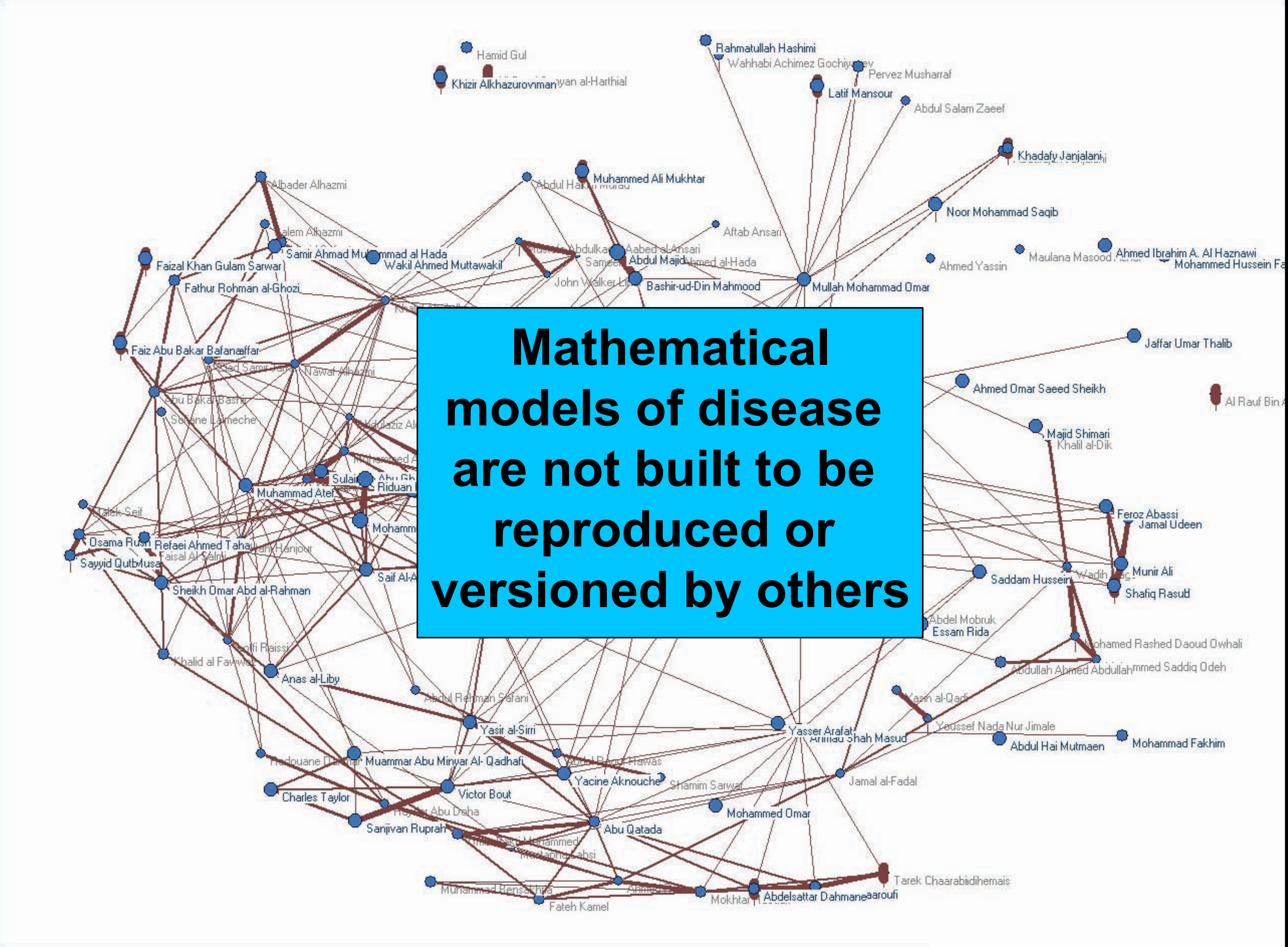
support

support

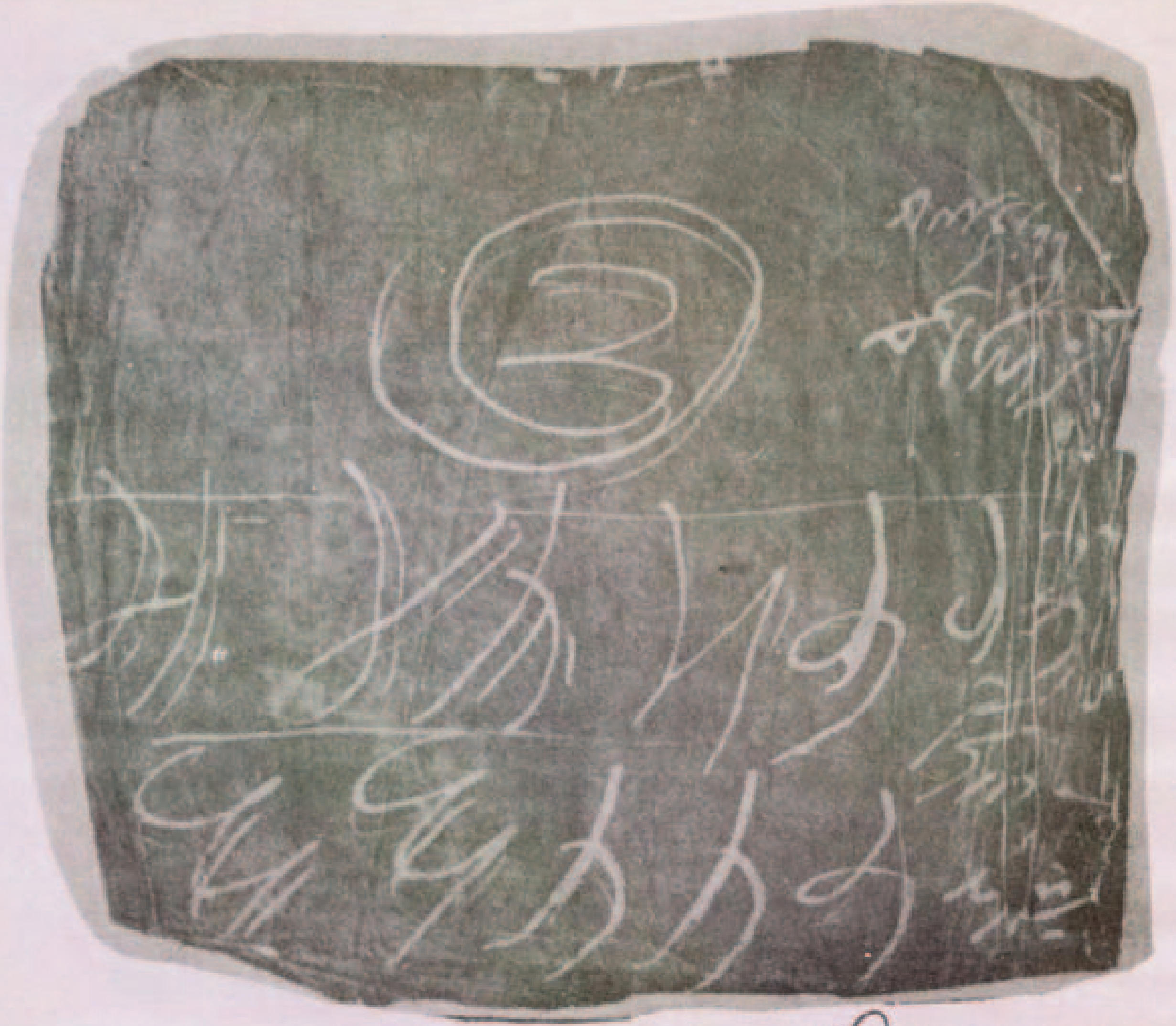
Pharmaceutical Companies

support

Academic Institutions



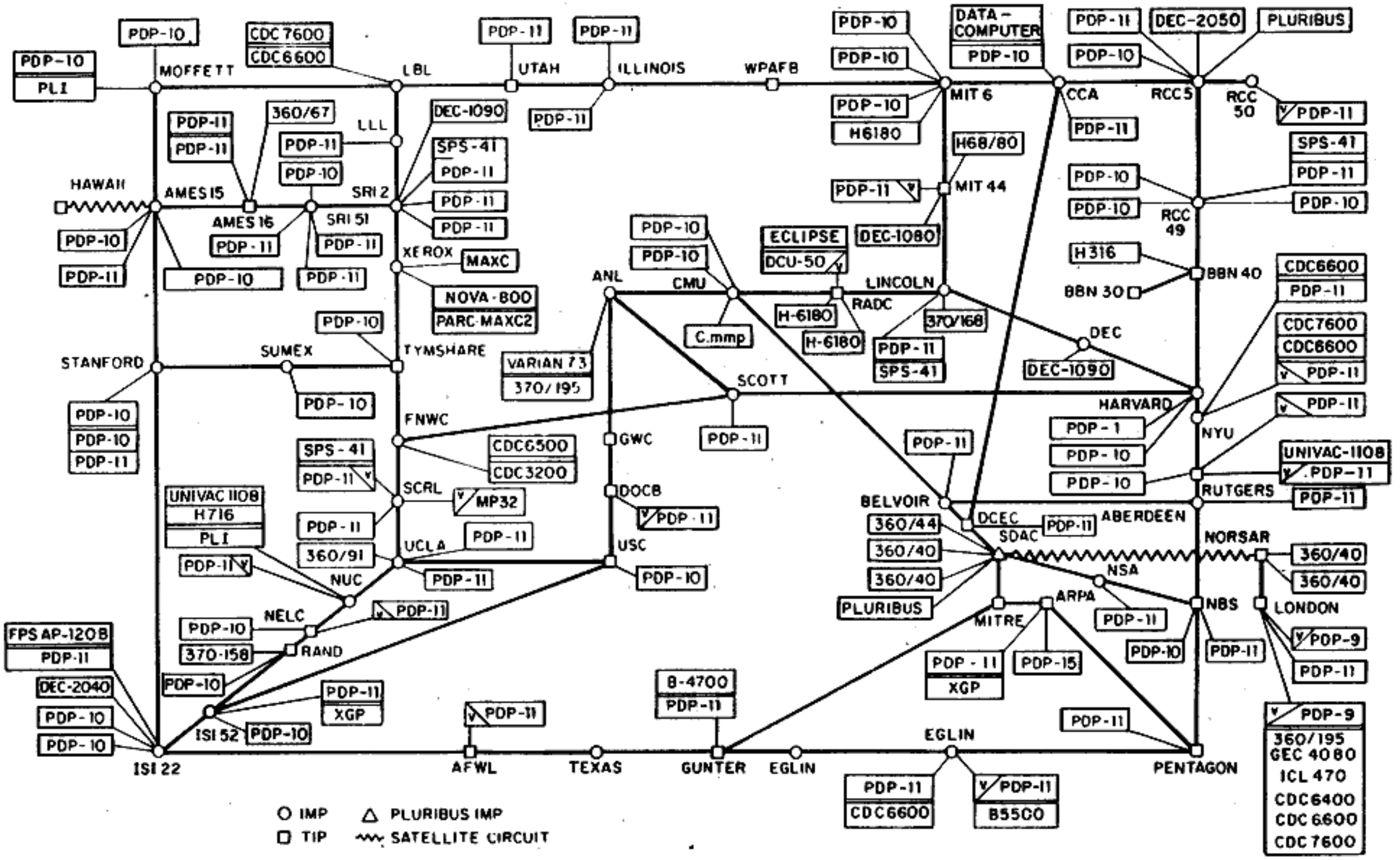
Mathematical models of disease are not built to be reproduced or versioned by others



ਪੋਥੀਆਂ ਗੁਰੂ ਹਰਿਸਹਾਇ ਵਾਲੀਆਂ ਵਿਚੋਂ ਇਕ ਭਾਗ ਦਾ ਆਰੰਭਕ ਸਫ਼ਾ ੴ ਸਤਿਨਾਮ—ਬਾਬਾ ਨਾਨਕ

Lack of standard forms for sharing data and lack of forms for future rights and consentss

ARPANET LOGICAL MAP, MARCH 1977

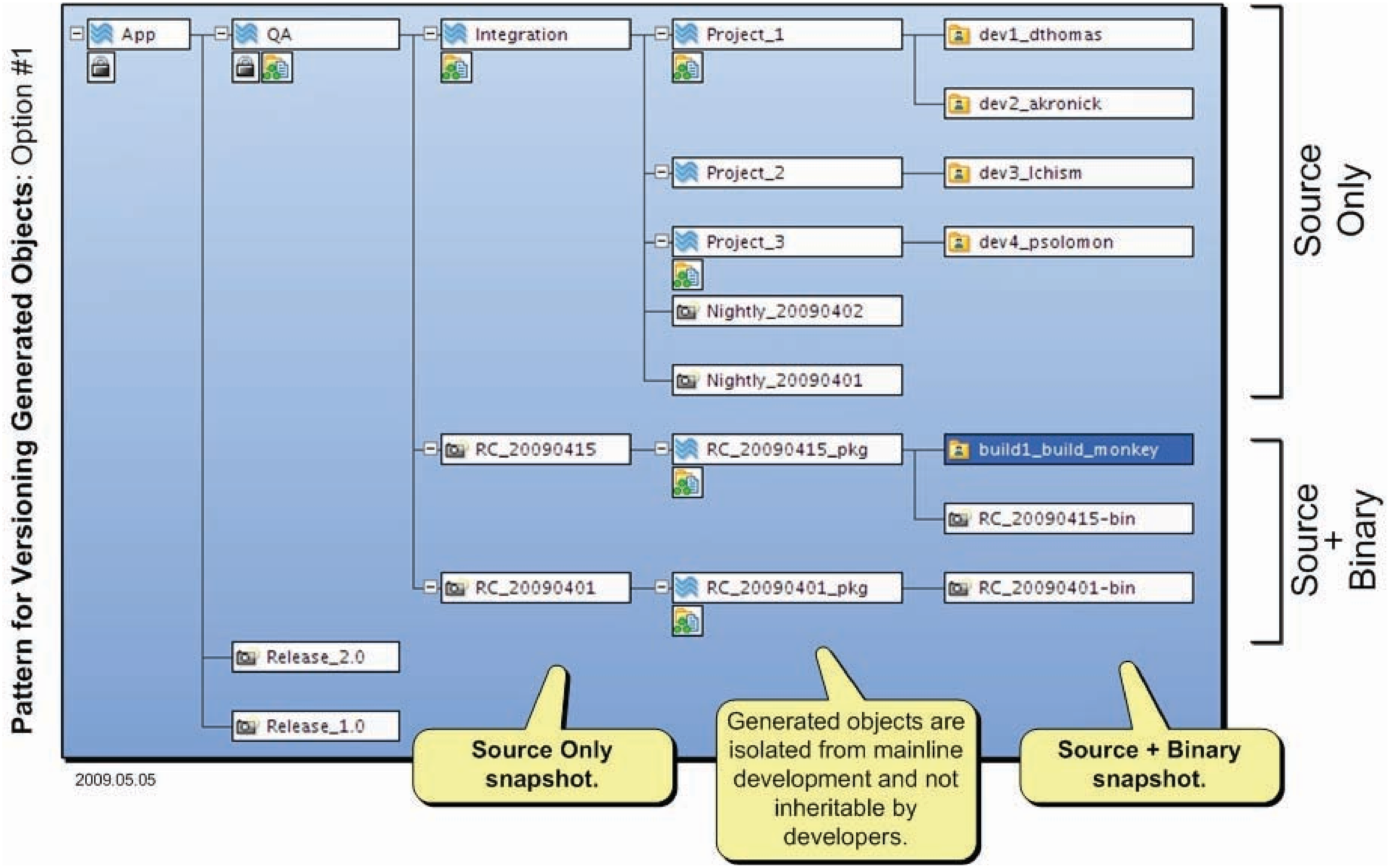


(PLEASE NOTE THAT WHILE THIS MAP SHOWS THE HOST POPULATION OF THE NETWORK ACCORDING TO THE BEST INFORMATION OBTAINABLE, NO CLAIM CAN BE MADE FOR ITS ACCURACY)

NAMES SHOWN ARE IMP NAMES, NOT (NECESSARILY) HOST NAMES

sharing as an adoption of common standards..
Clinical   Genomics   Privacy   IP

Why not share clinical /genomic data and model building in the ways currently used by the software industry (power of tracking workflows and versioning



Pattern for Versioning Generated Objects: Option #1

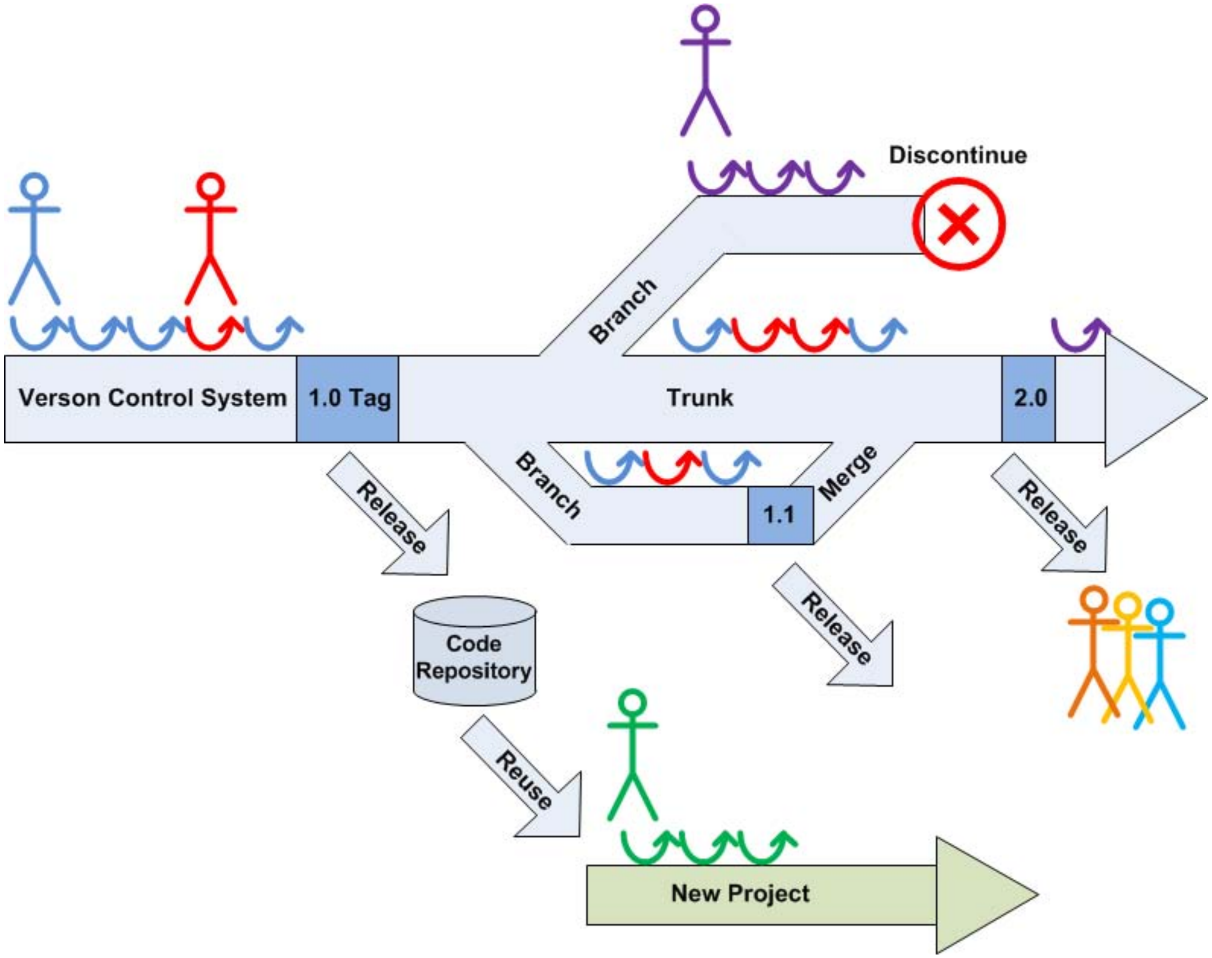
Source Only snapshot.

Generated objects are isolated from mainline development and not inheritable by developers.

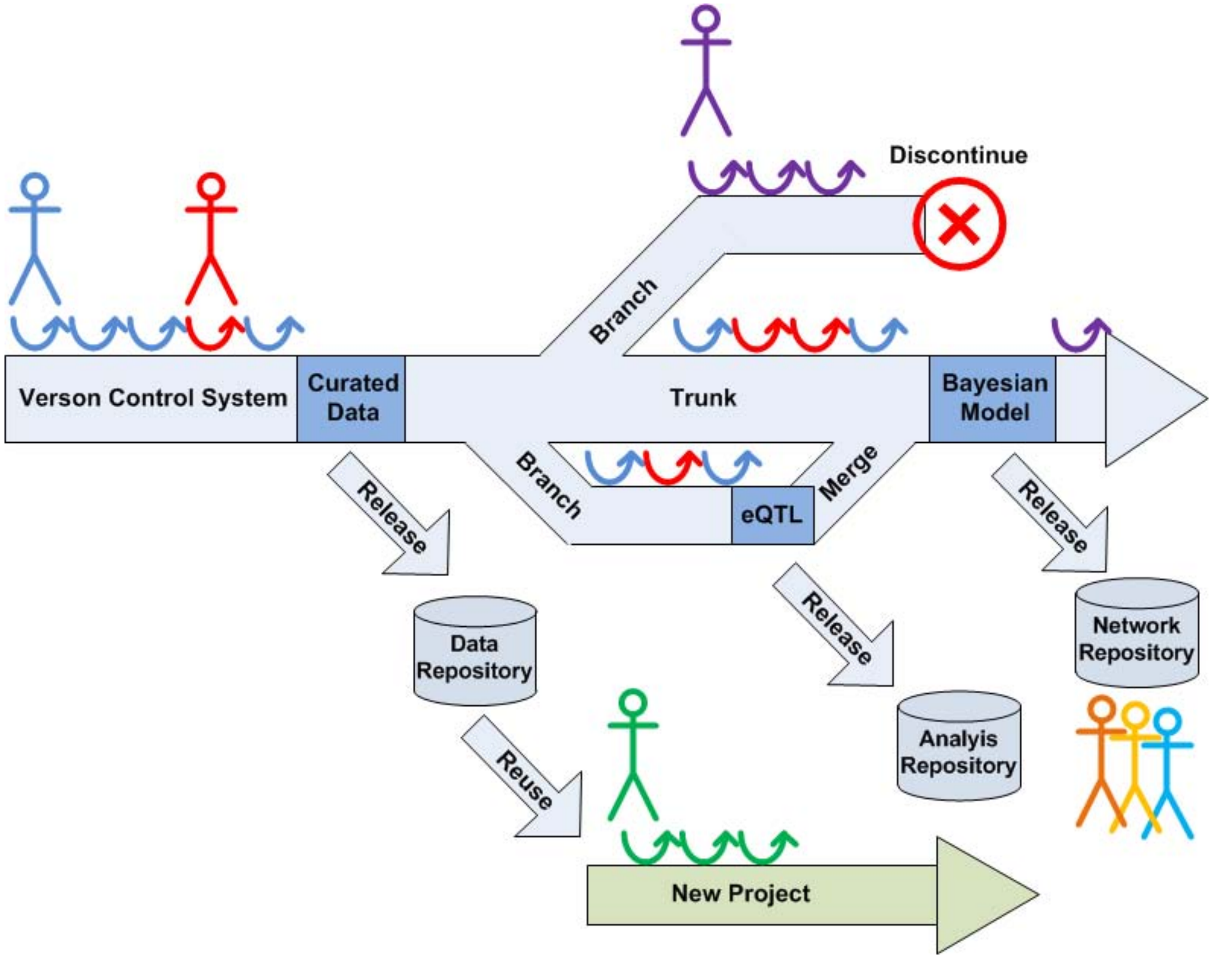
Source + Binary snapshot.

2009.05.05

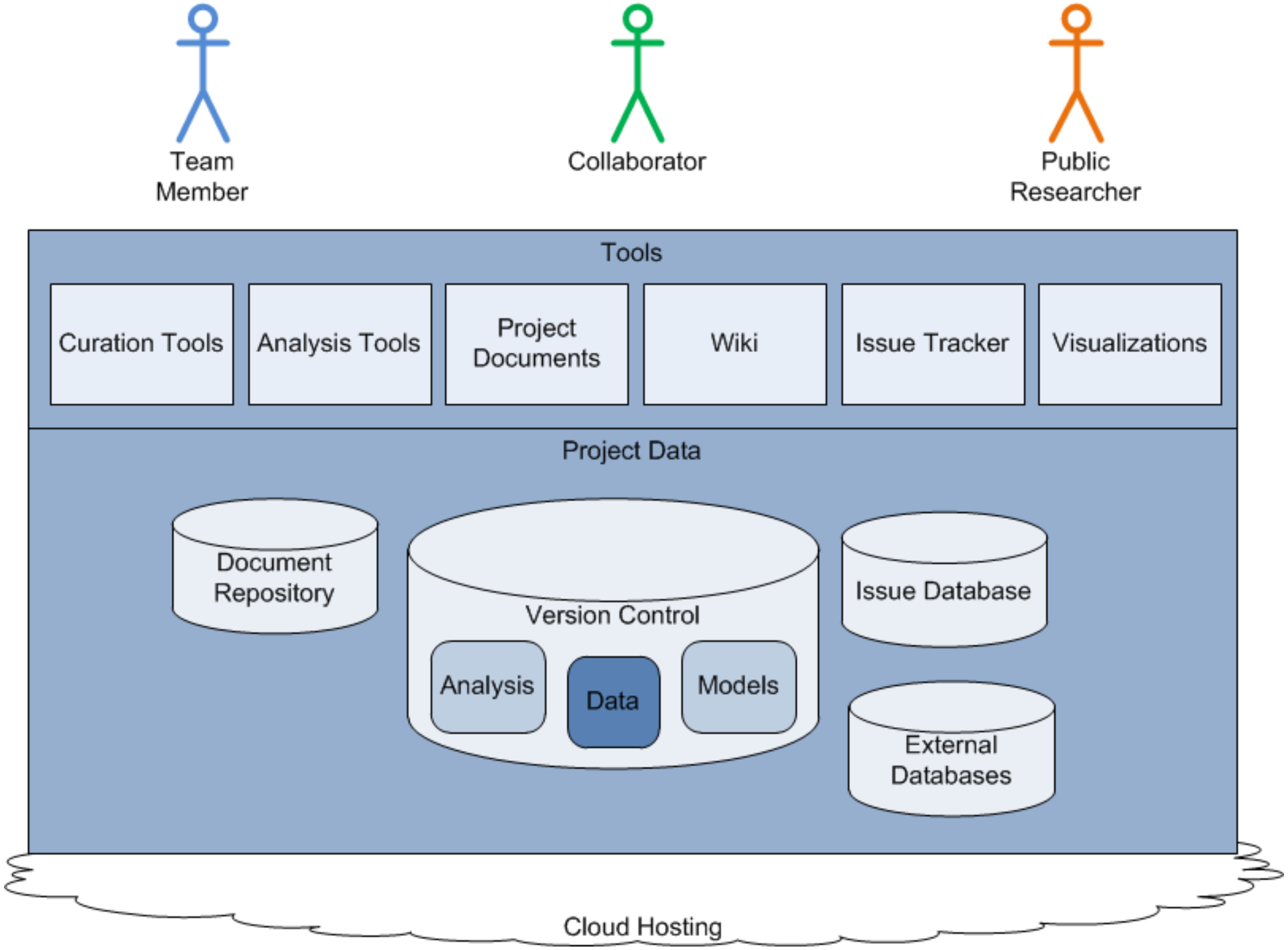
Evolution of a Software Project



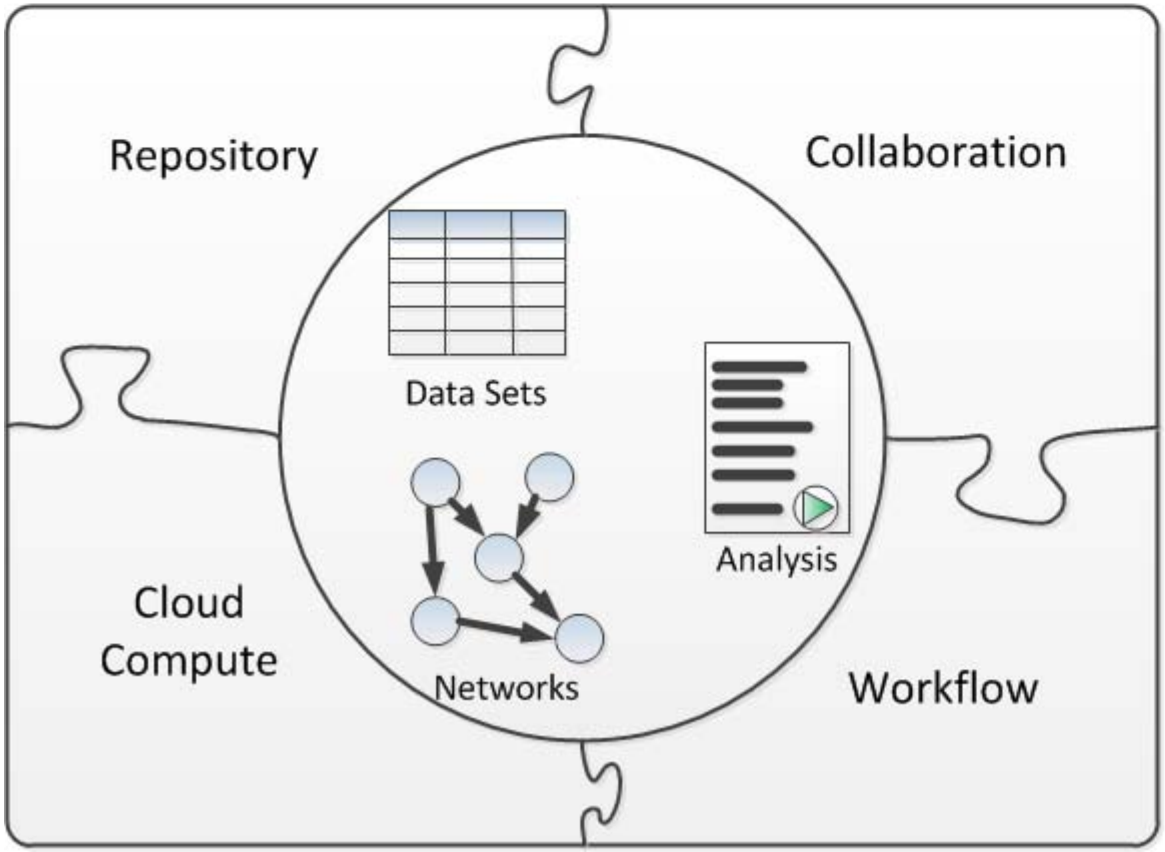
Evolution of a Biology Project



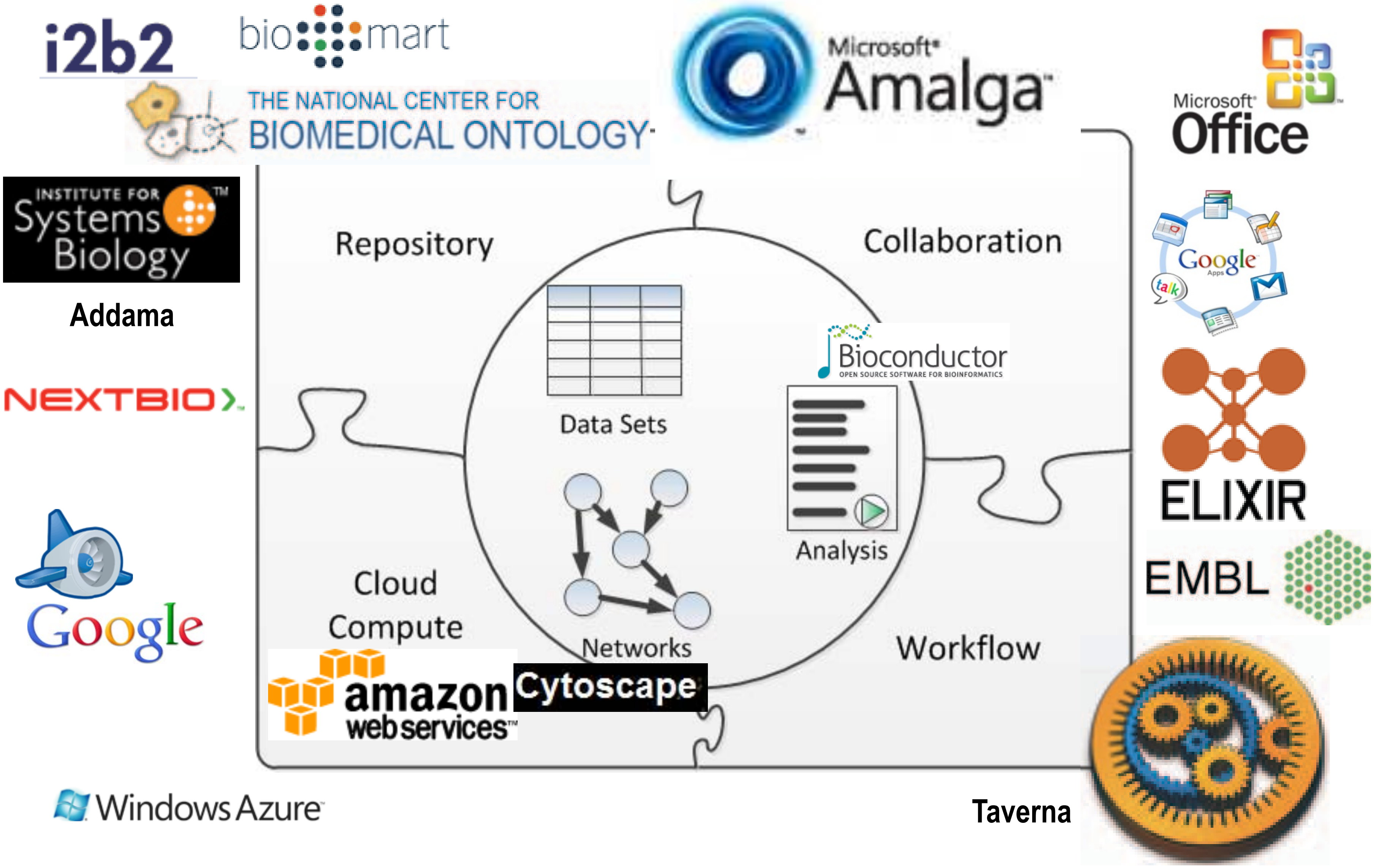
Biology Tools Support Collaboration

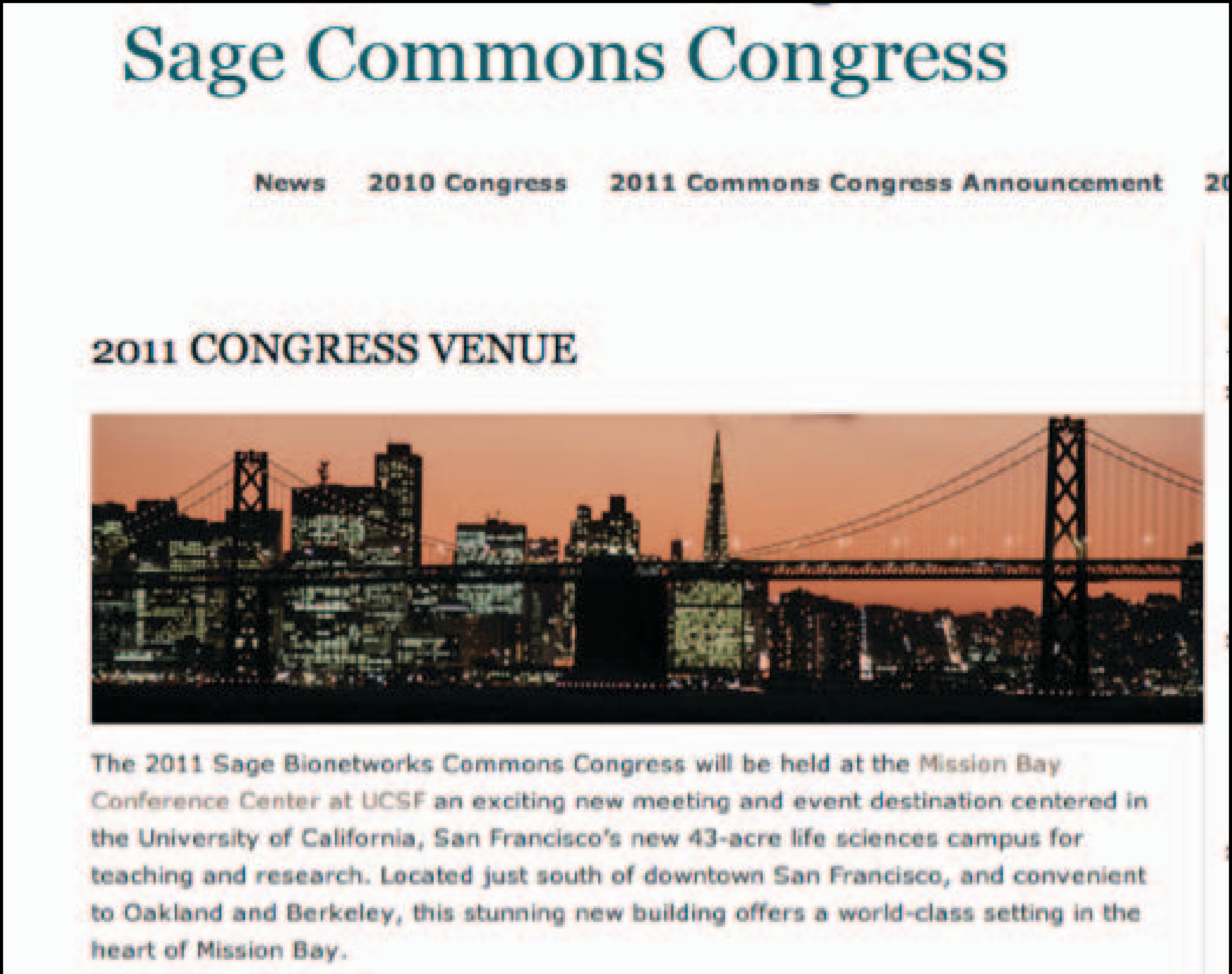


Platform Functional Areas



Potential Supporting Technologies





Sage Commons Congress

2011 CONGRESS VENUE

The 2011 Sage Bionetworks Commons Congress will be held at the Mission Bay Conference Center at UCSF an exciting new meeting and event destination centered in the University of California, San Francisco's new 43-acre life sciences campus for teaching and research. Located just south of downtown San Francisco, and convenient to Oakland and Berkeley, this stunning new building offers a world-class setting in the heart of Mission Bay.

Aprill 15-16 this will be broadcast LIVE as completely oversubscribed

Google: Sage Commons Congress as time approaches

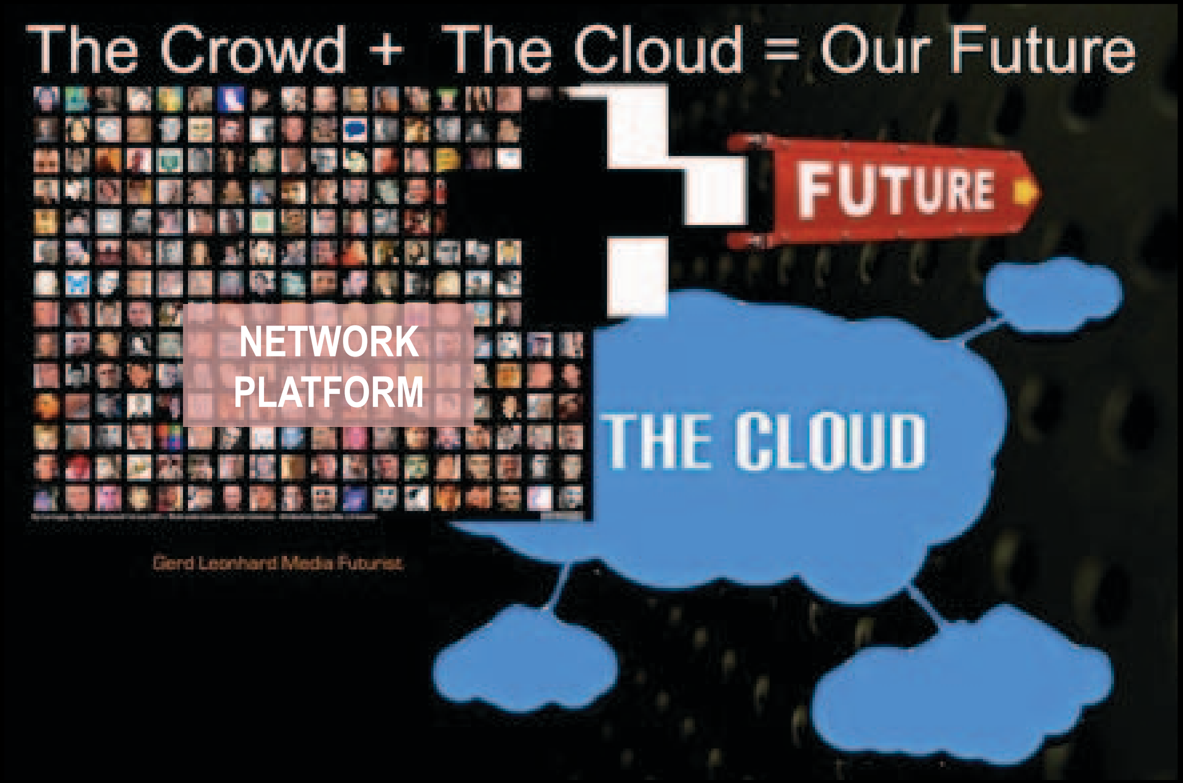
Who will build the datasets/ models capable of providing powerful safety and efficacy insights?

Institute

PI

Post Docs

Grad Students


The Crowd + The Cloud = Our Future
FUTURE
NETWORK PLATFORM
THE CLOUD
Gerd Leonhard Media Futurist

Patients Physicians Citizens Knowledge Experts

BETTER MAPS OF DISEASE

NOT JUST WHAT WE DO BUT HOW WE DO IT

POWER OF BUILDING A PRE-COMPETITIVE
COMMONS FOR EVOLVING
GENERATIVE MODELS OF DISEASE
USING A PUBLIC PRIVATE PARTNERSHIP