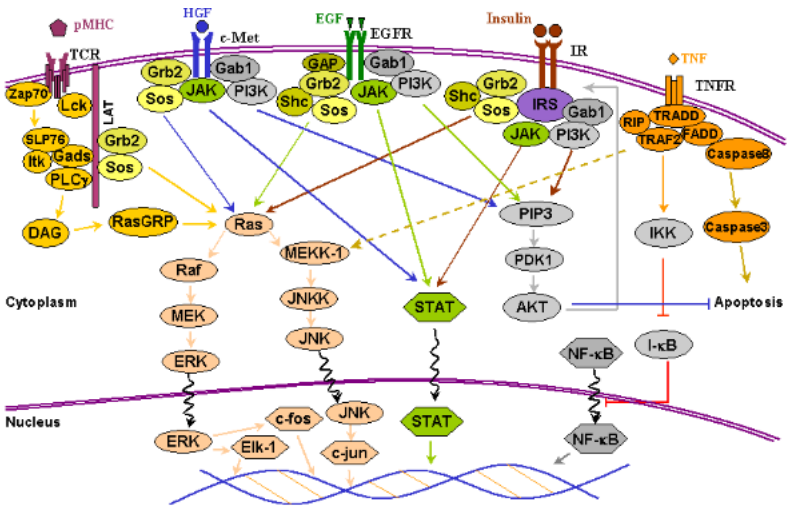
NCIP Workshop

May 31, 2012

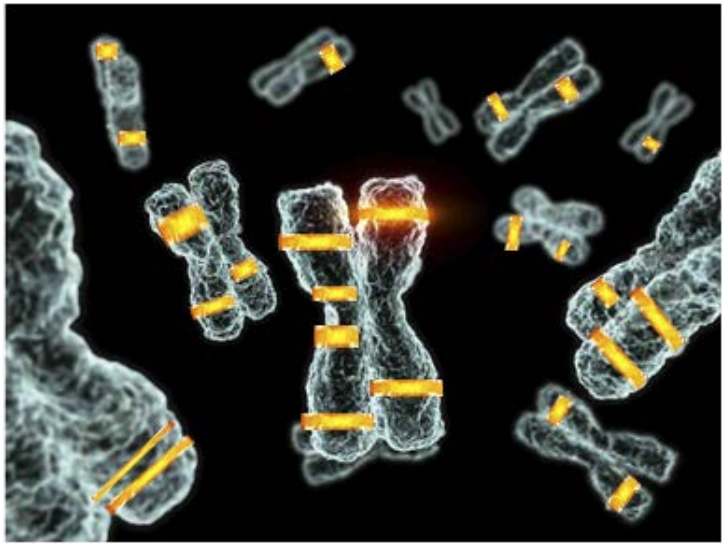
Kevin White
Institute for Genomics and Systems Biology
The University of Chicago

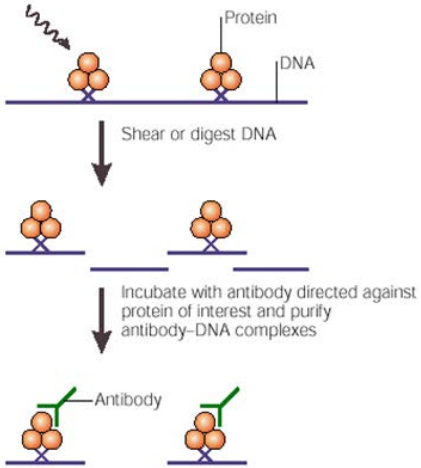
Genomics and cancer (systems) biology

Cellular state

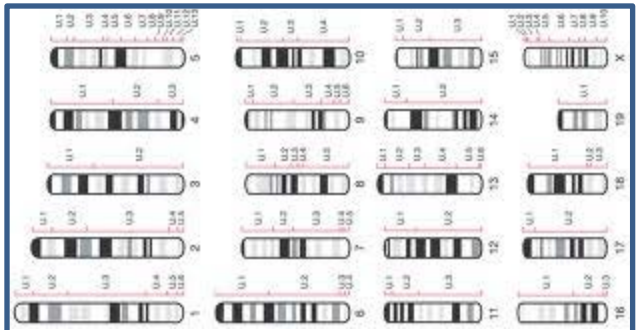


Genetic state

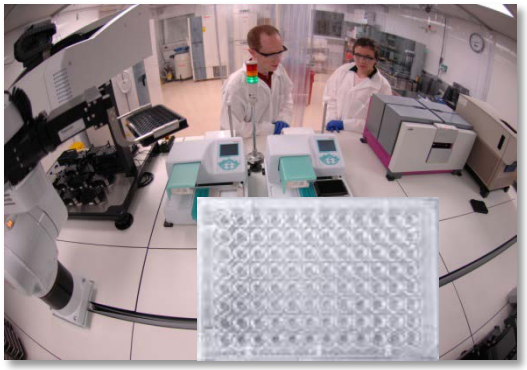




Genome regulation data

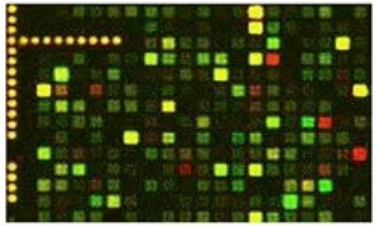


Genome variation data

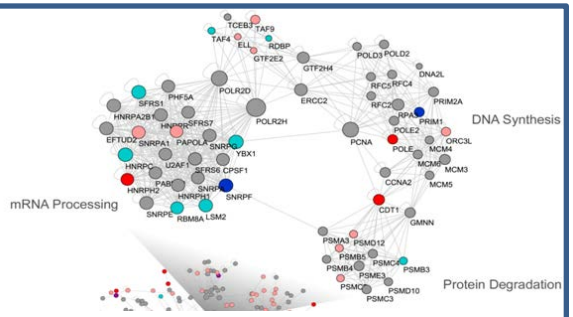


Drug screening data

Integration and building predictive models

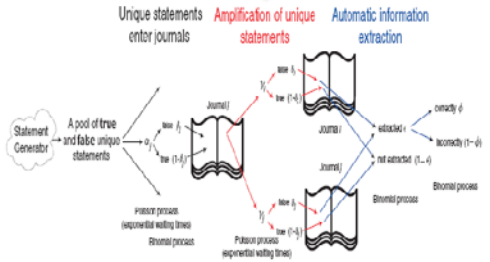


Gene expression data



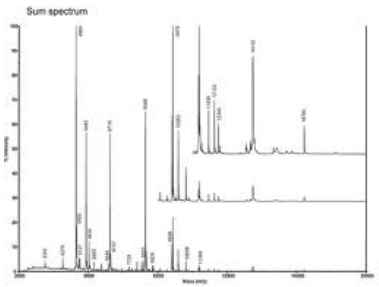
Proteome data

Etc…



Literature/text mining data

EMRs



Metabolome data

A dynamic, evolutionary framework

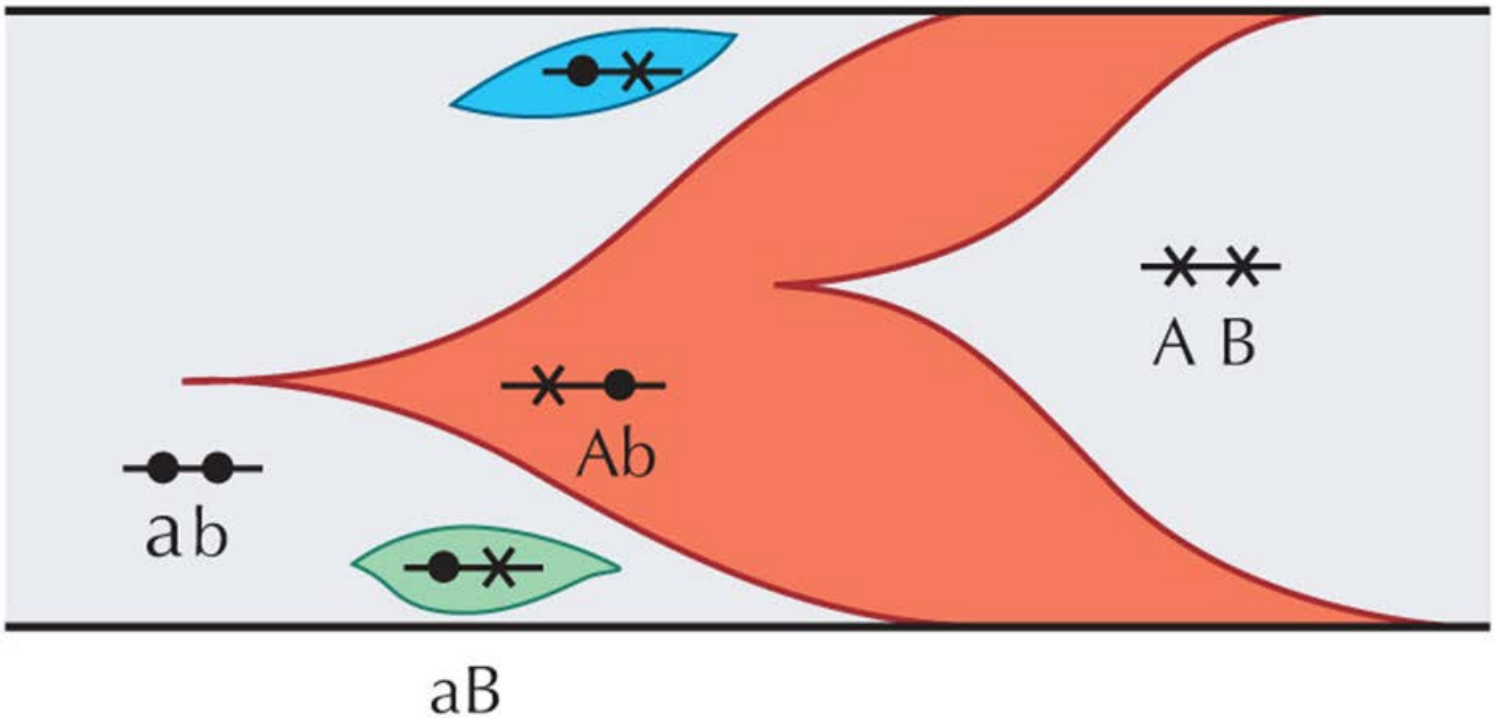
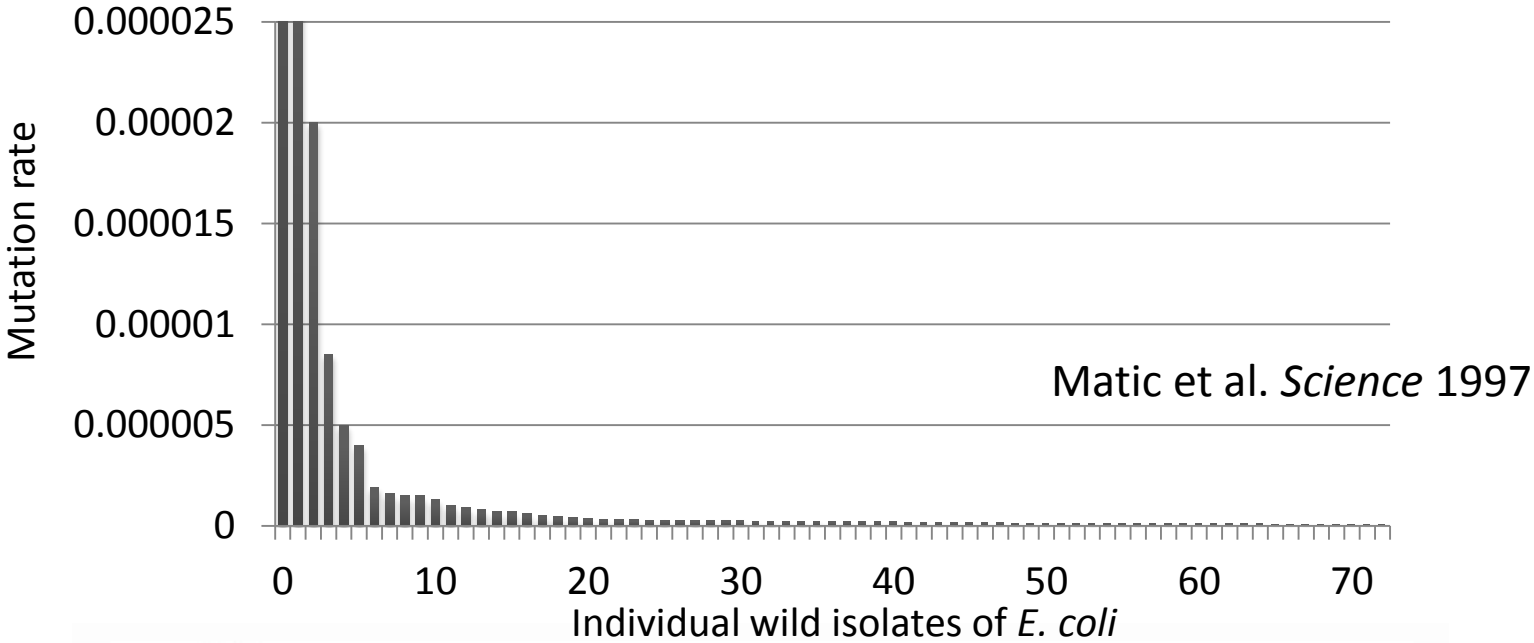


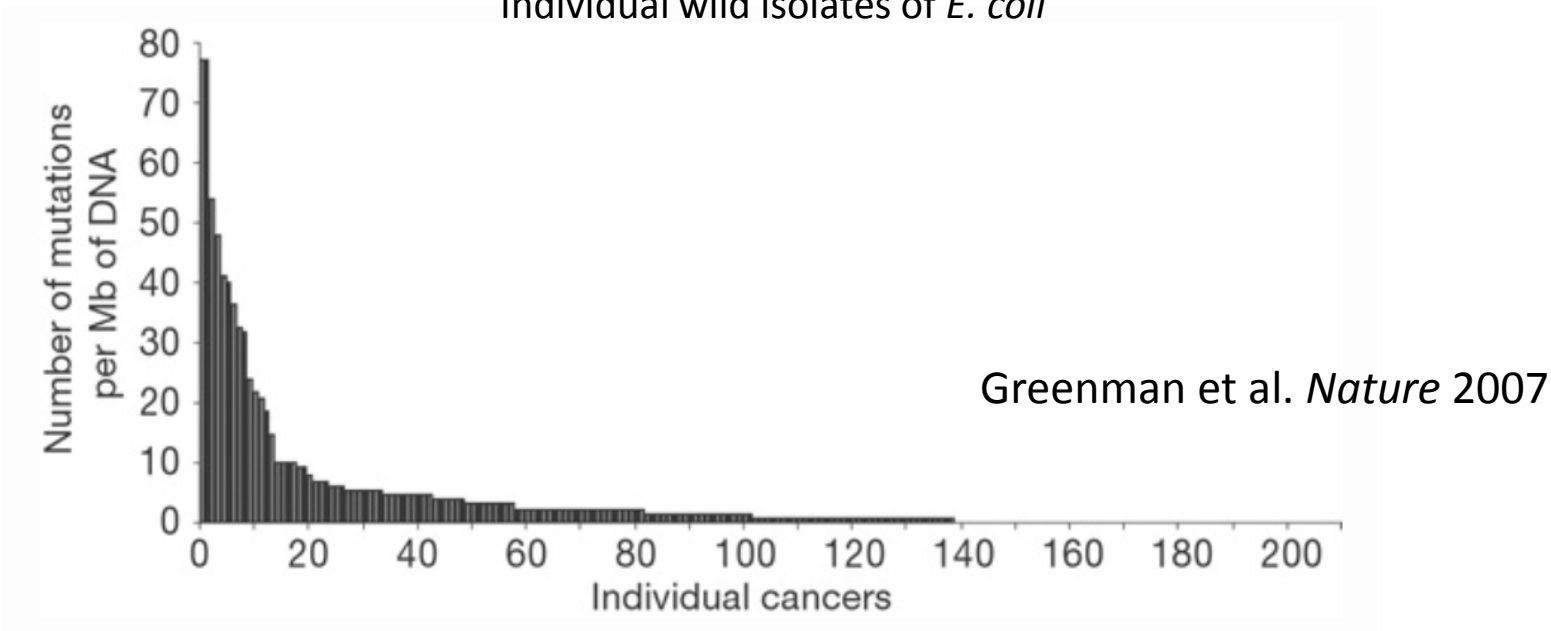
FIGURE 23.18. The Fisher–Muller argument. (*A*)

With asexual reproduction, favorable mutations must be established sequentially. For example, if allele A is destined to replace a, then any favorable alleles that occur at other loci (B, for instance) can only be fixed if they occur within a genome carrying A

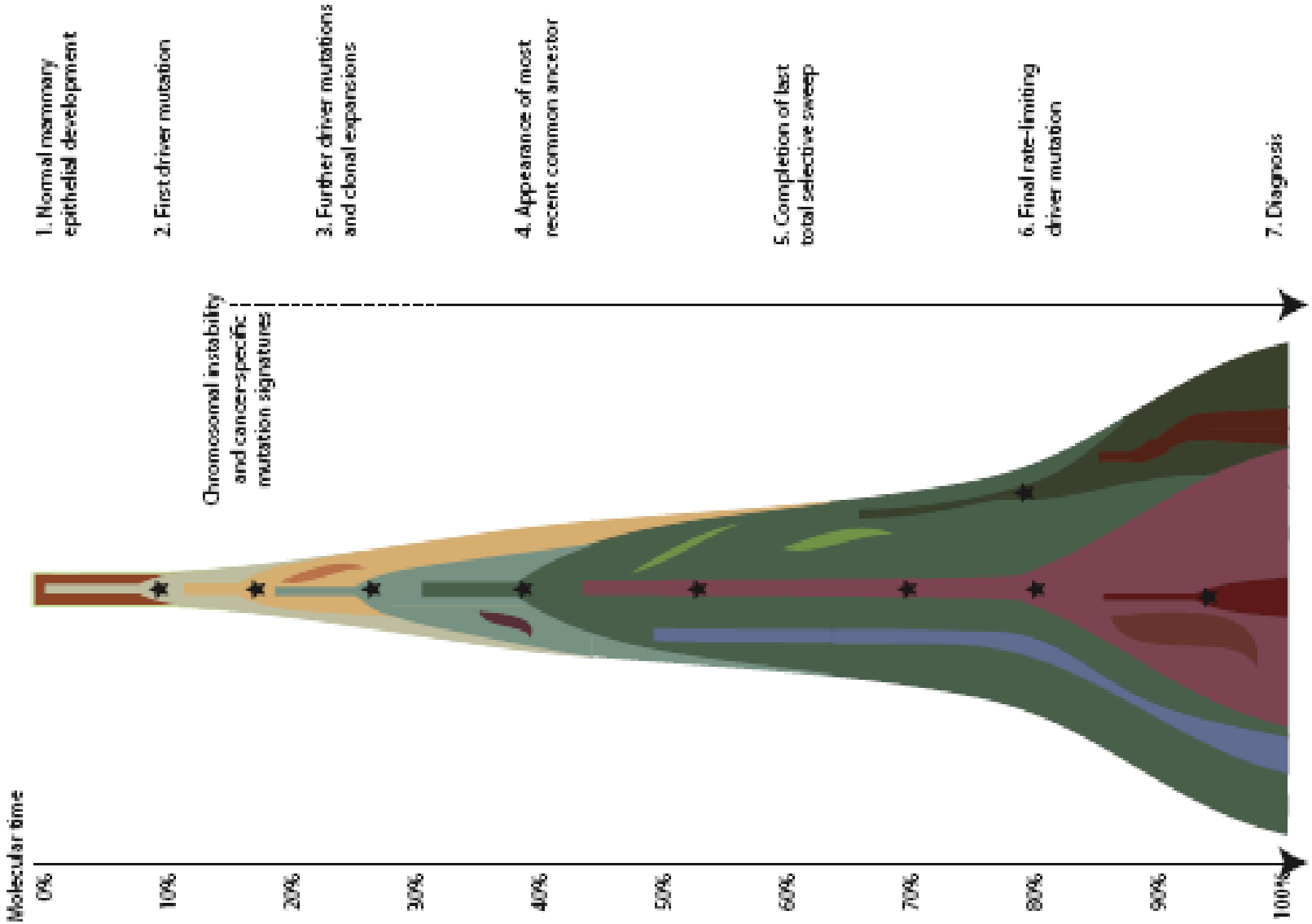
Fisher 1930; Muller 1932; Nordling 1953; Knudson 1971

Mutation rates in cancer and "simple" evolving cell populations





Deep sequencing and lineage analysis

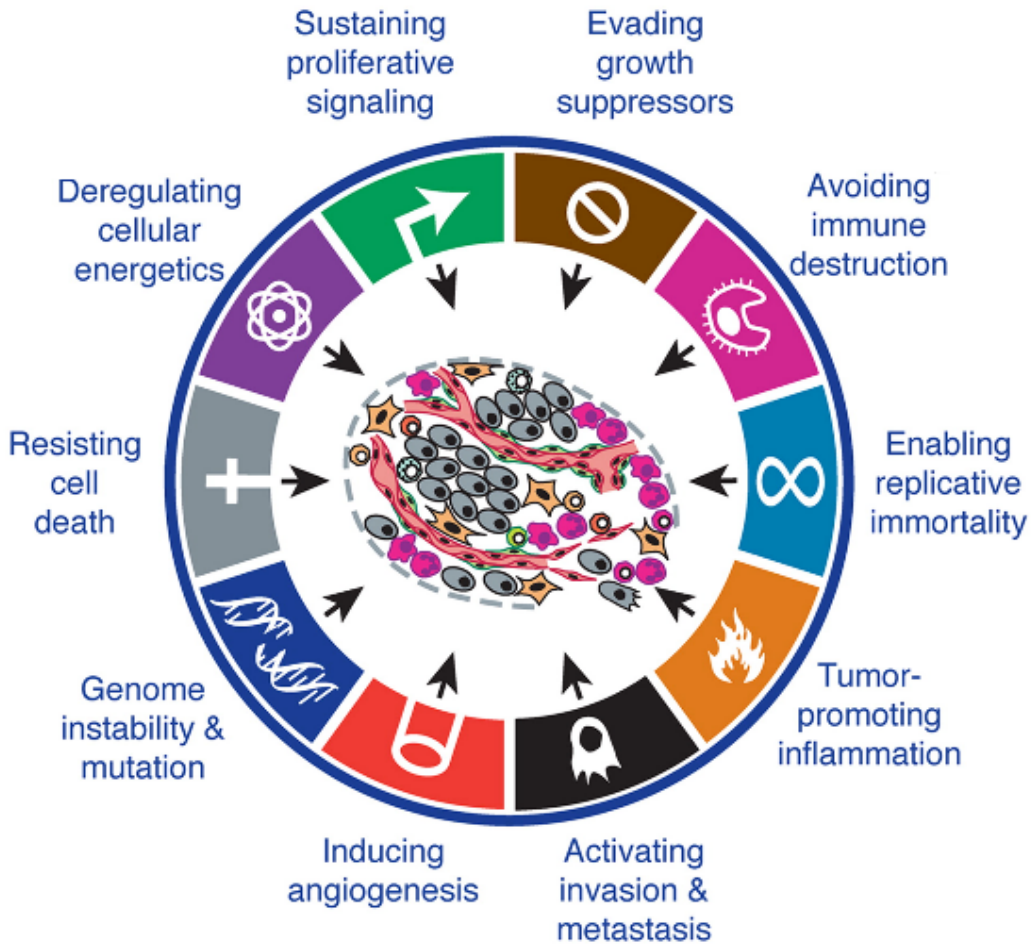


Nik-Zainal et al. *Cell* 2012

Quantitative Genetics and Cancer TCGA/TARGET and GWAS

Phenotype (P) = G + E +G*E

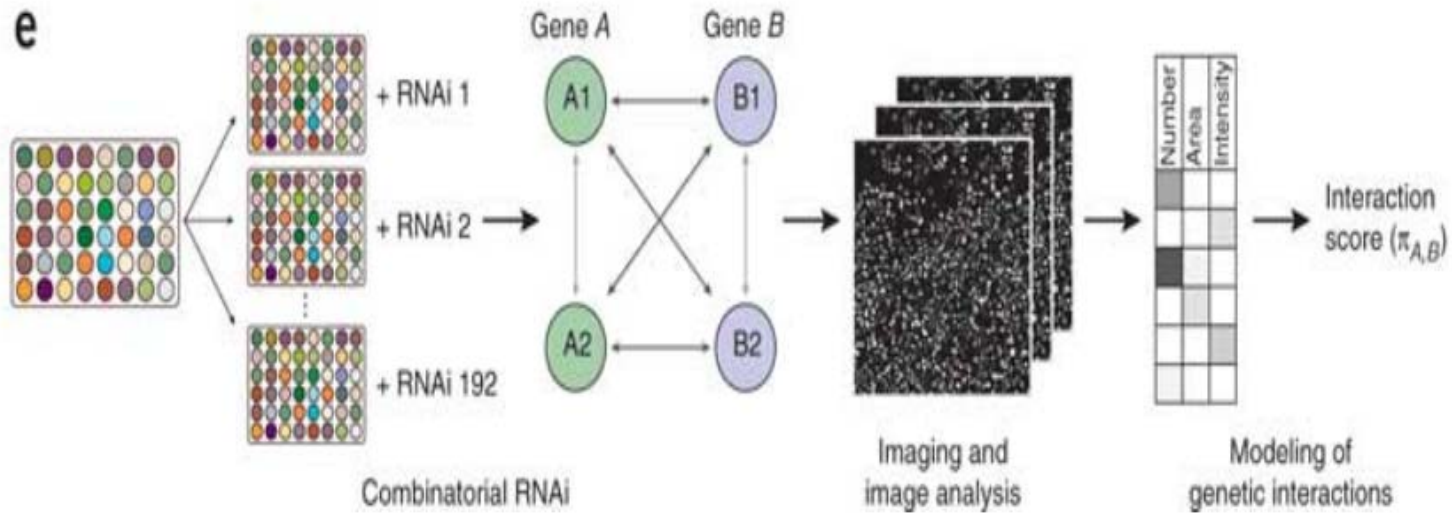
Cancer C = G + E + M + G*E + G*M + E*M + G*E*M.



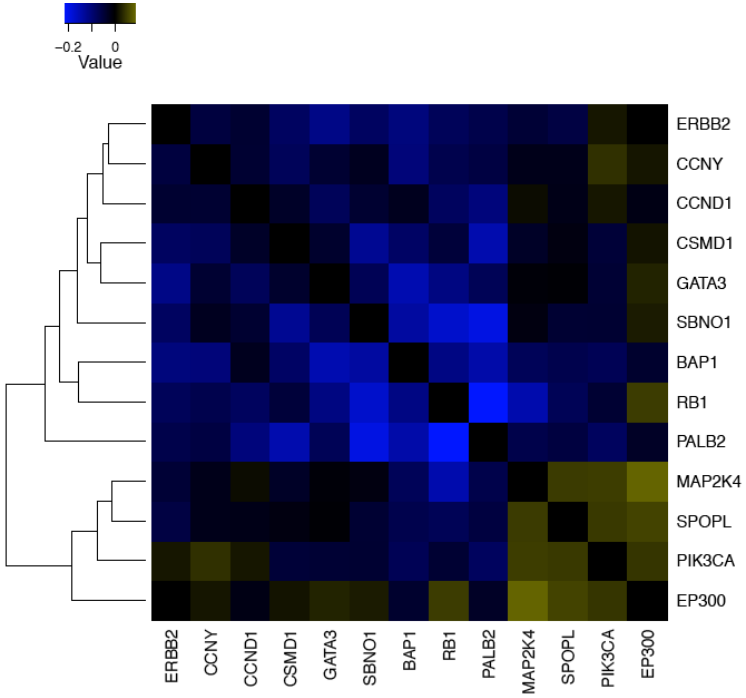
Hanahan and Weinberg Cell 2011

- **What combinations and order of mutations?**
- **What germ line – somatic interactions?**
- **Contribution of germ line variation to tumor microenvironment?**

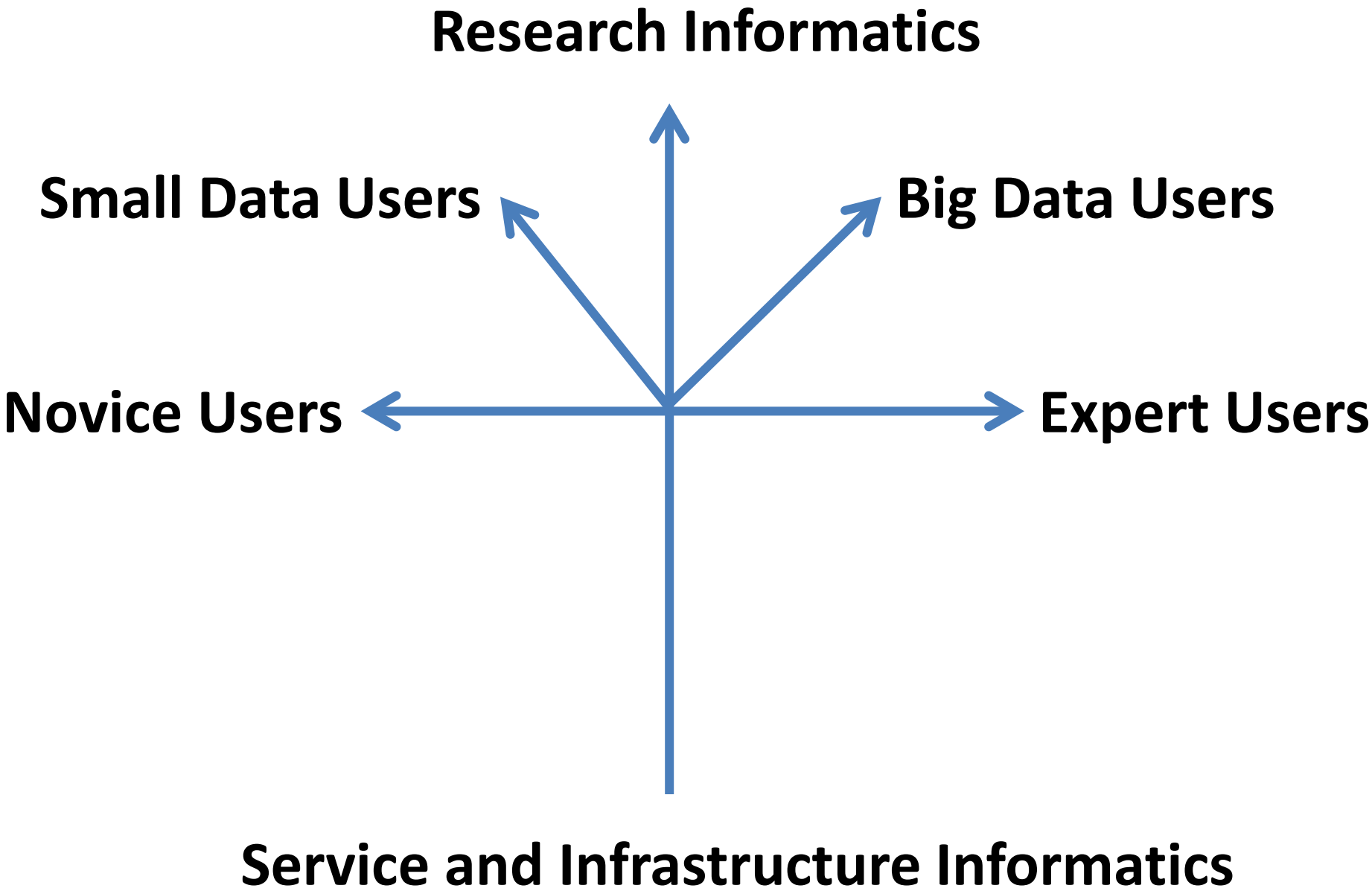
Genetic Interaction Screening: Validating recurring allelic combinations from TCGA

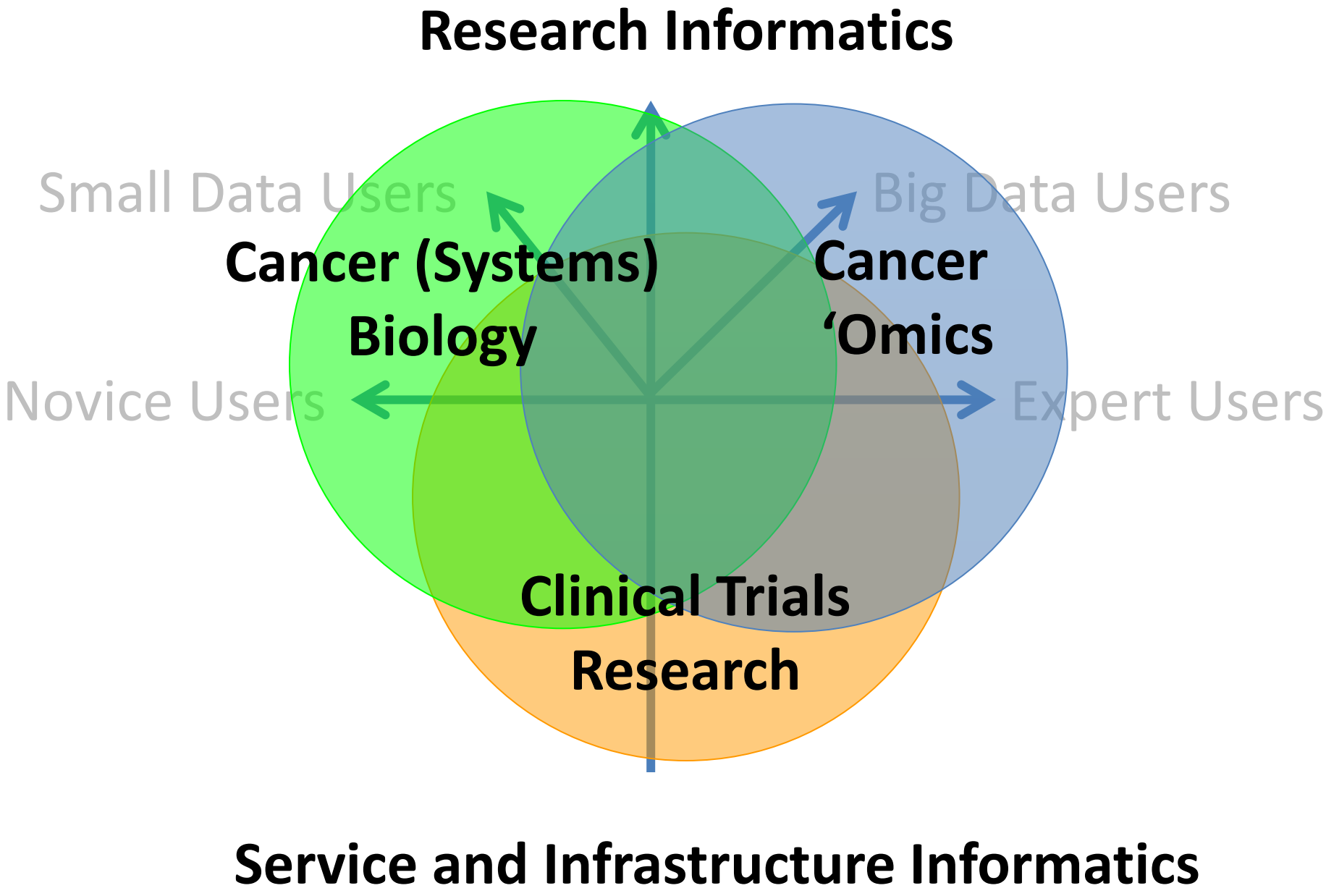


Horn et al. ***Nat. Methods*** 2011

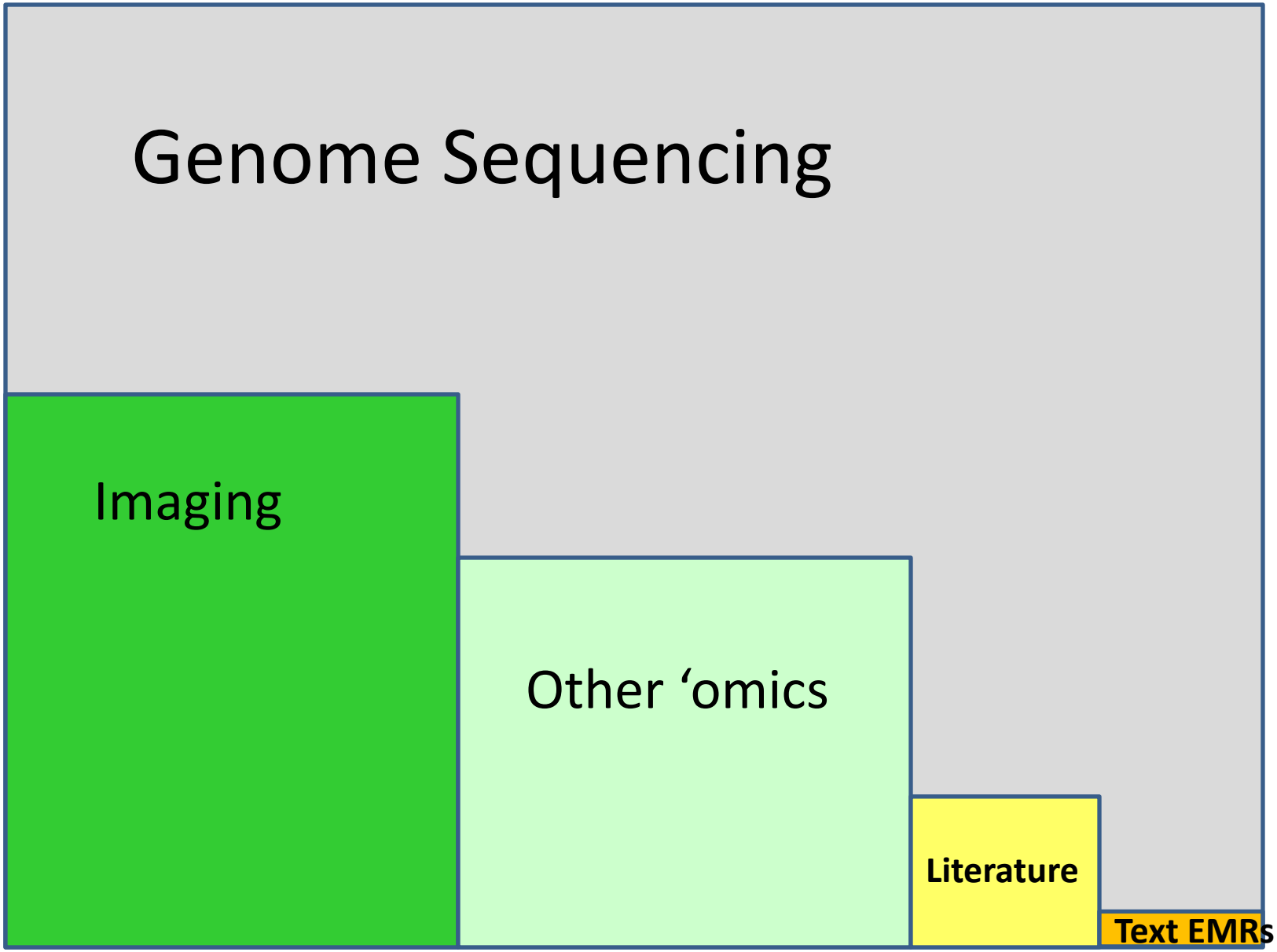


Wang et al., in prep

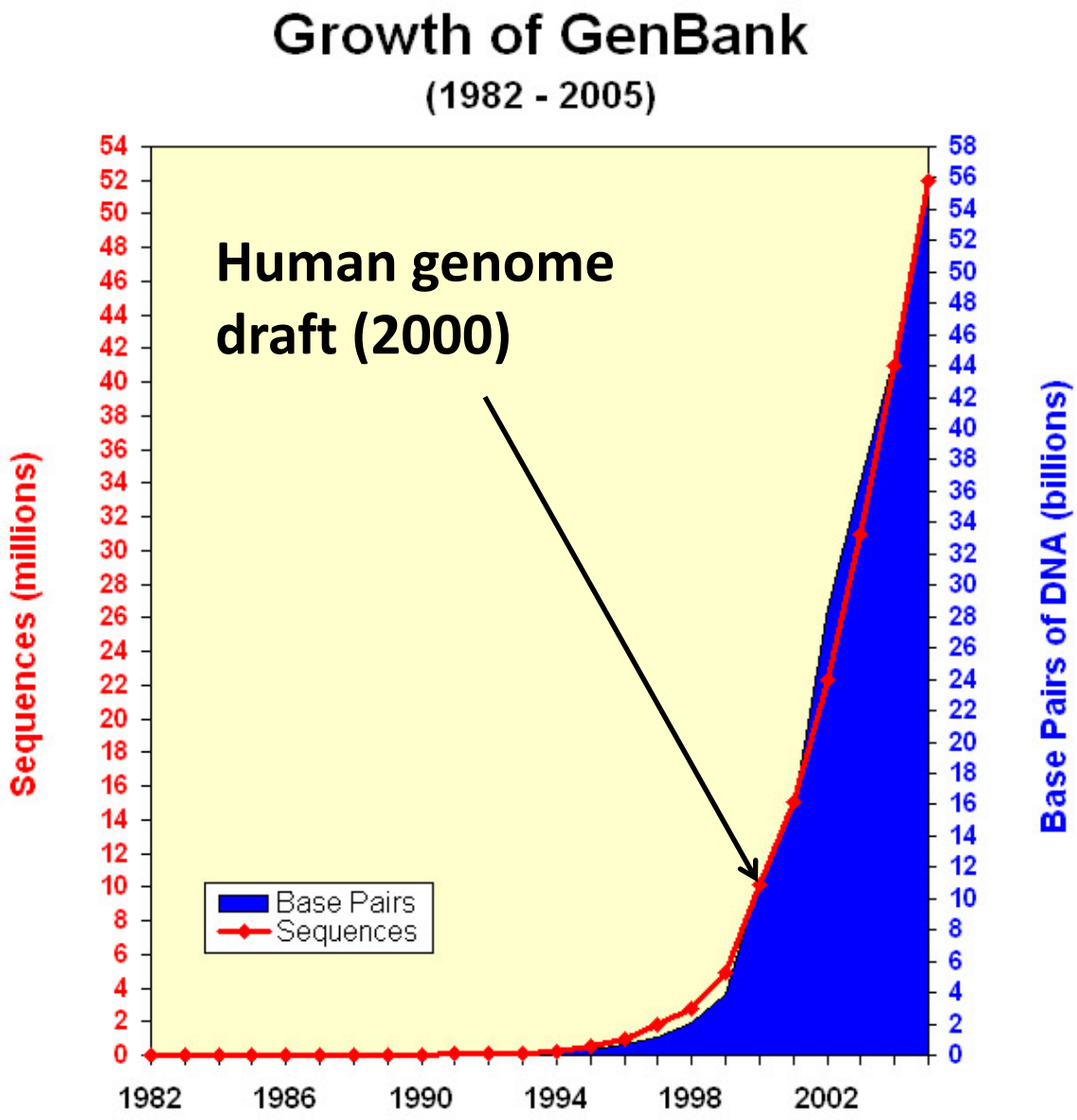

Research Informatics
Small Data Users
Big Data Users
Novice Users
Expert Users
Service and Infrastructure Informatics


Research Informatics
Small Data Users
Big Data Users
Cancer (Systems) Biology
Cancer 'Omics
Novice Users
Expert Users
Clinical Trials Research
Service and Infrastructure Informatics

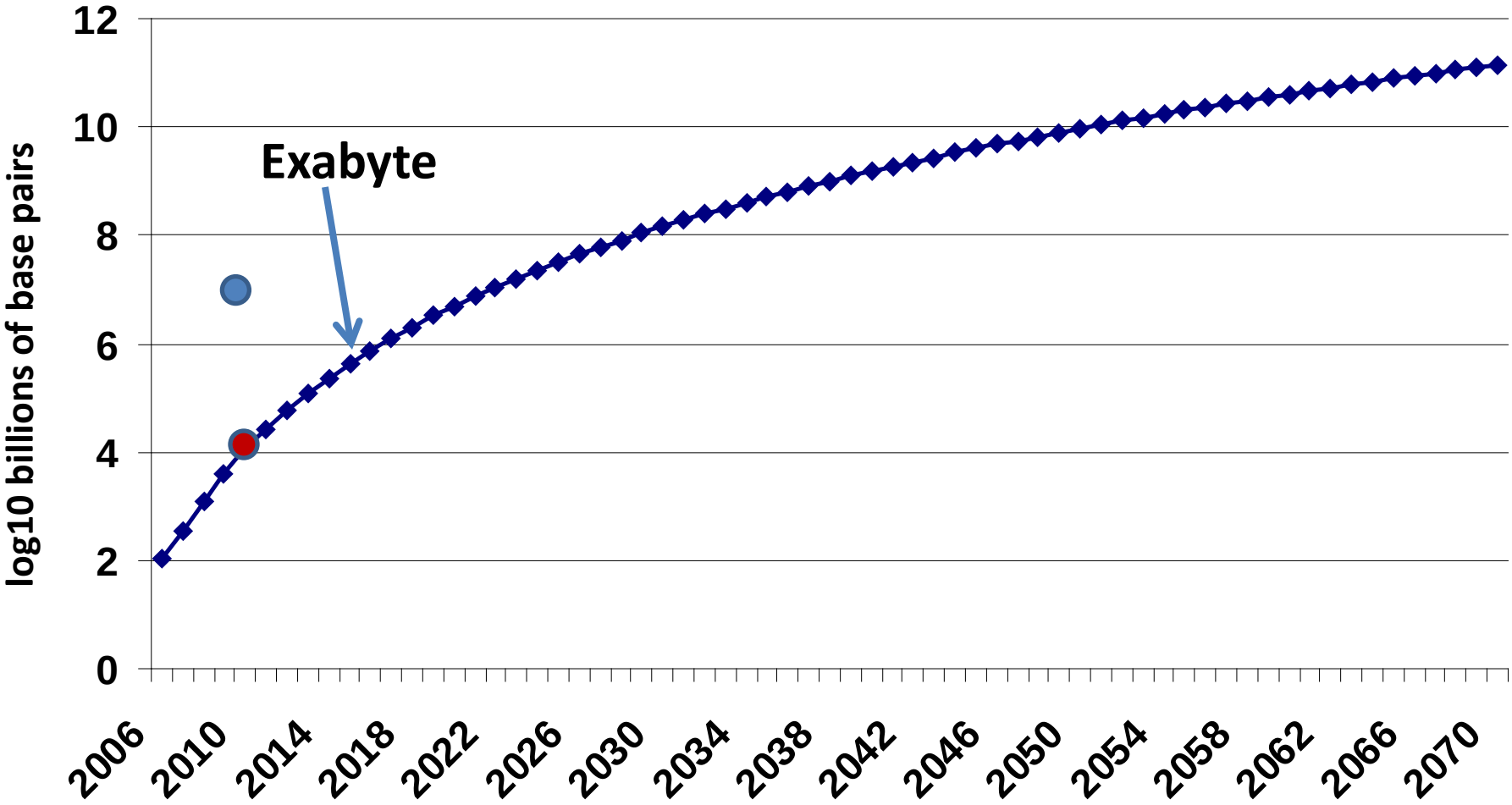
Data and Scale



The Root of 21st Century Genome Science
Exponential Growth of DNA sequencing



Projected sequencing capabilities world-wide



(estimates calculated in 2007 by K.P.W. for U.S. National Science Foundation advisory board presentation)

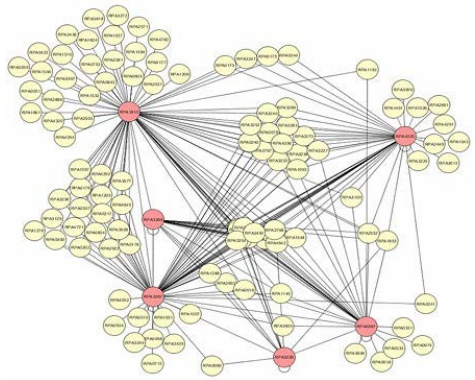
Google’s computing infrastructure uses 260 MW / hour.

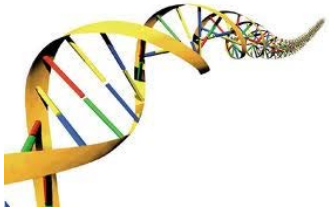
Google manages hundreds of PB of data.

NCBI

National Center for
Biotechnology Information

In 2012 NCBI had >2PB of data. But genome sequencing data is growing exponentially.



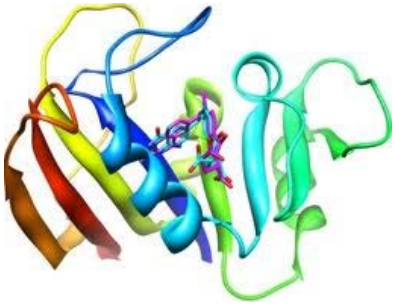


Elastic Clouds

Goal: Minimize cost of virtualized machines & provide on-demand.

Large Data Clouds

Goal: Maximize data (with matching compute) and control cost.

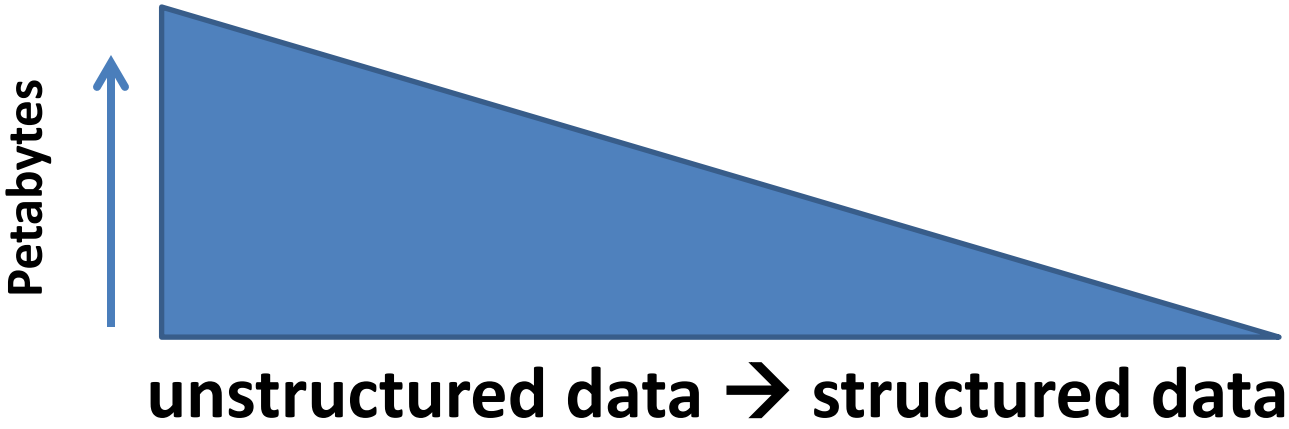


HPC

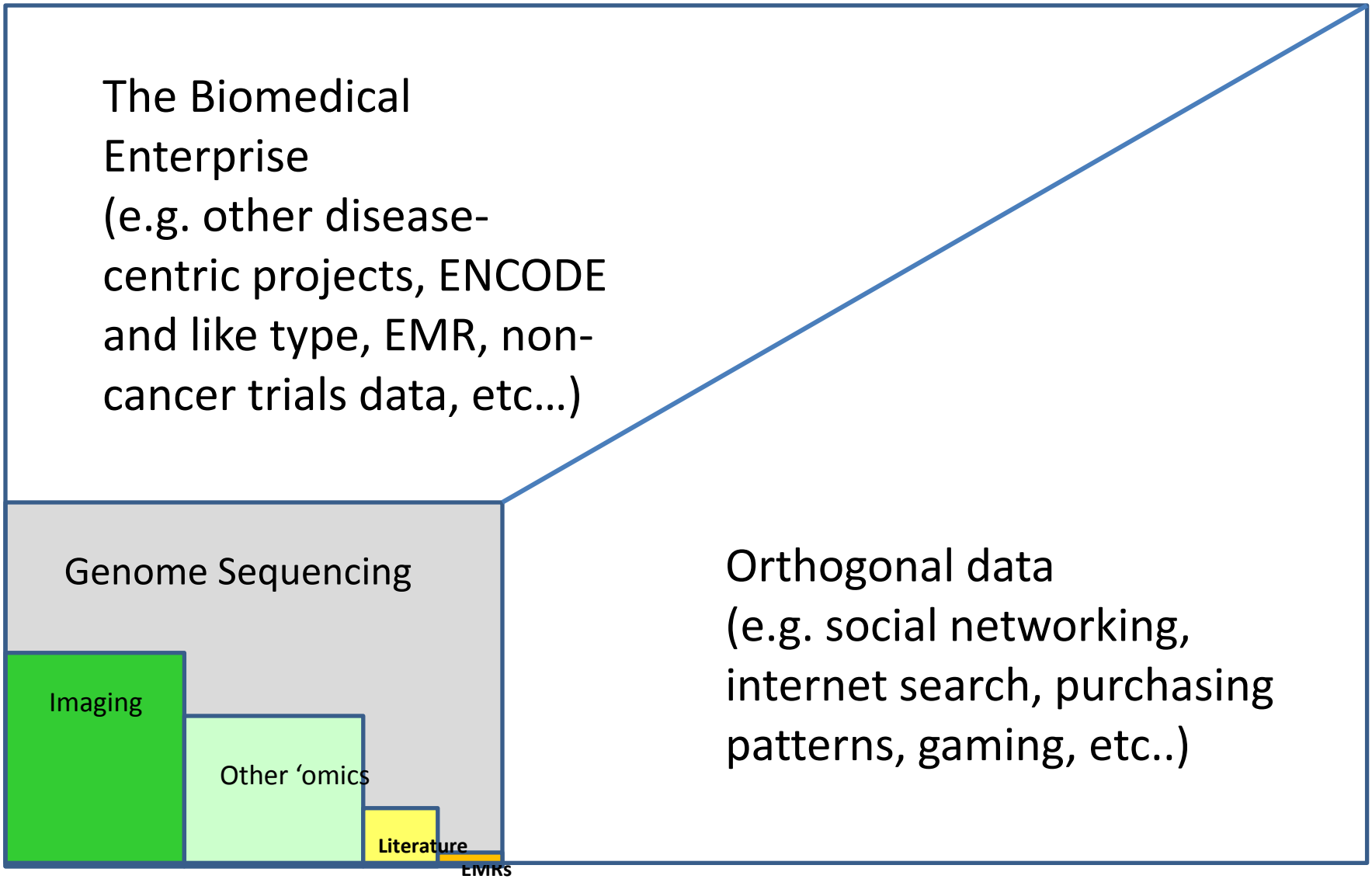
Goal: Maximize processing

Big Data Biology

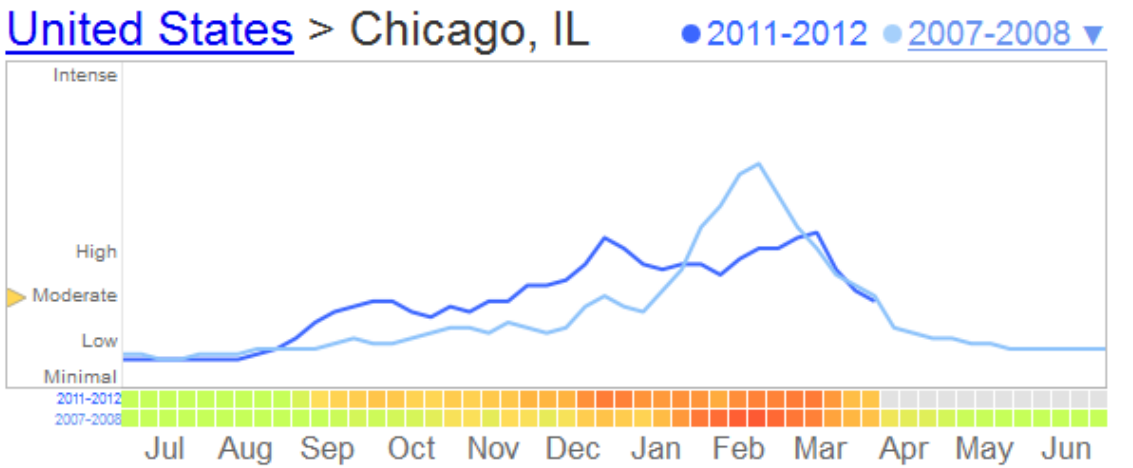
Think of big data biology as the academic discipline studying how to extract useful information from all available data relevant to biology.



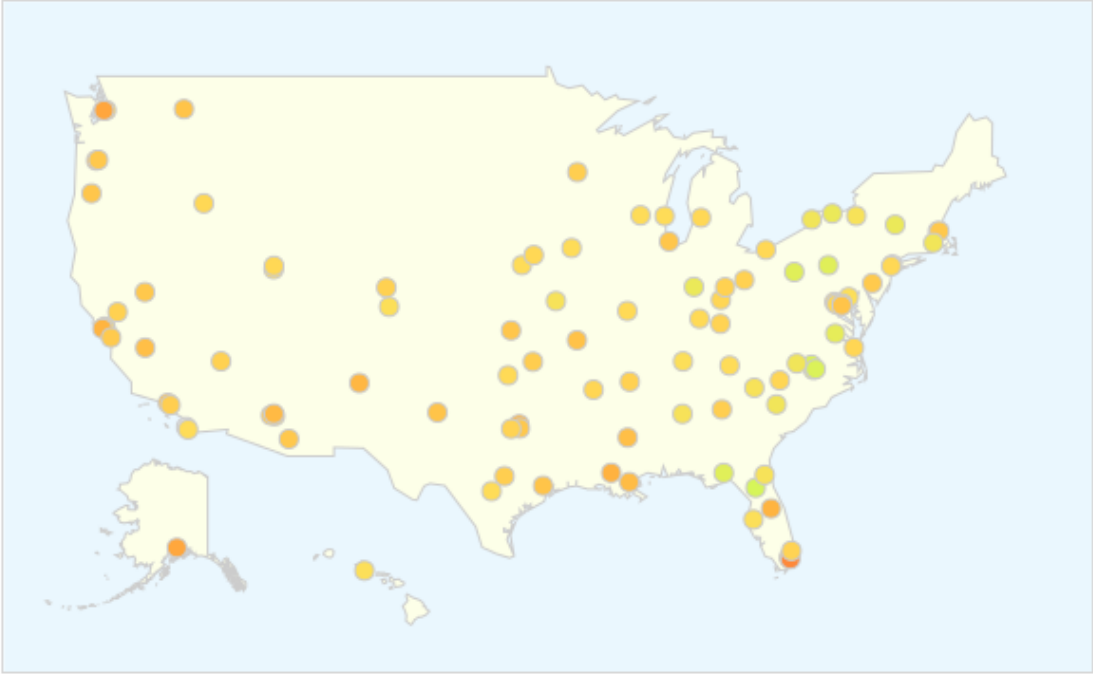
All data potentially relevant to cancer



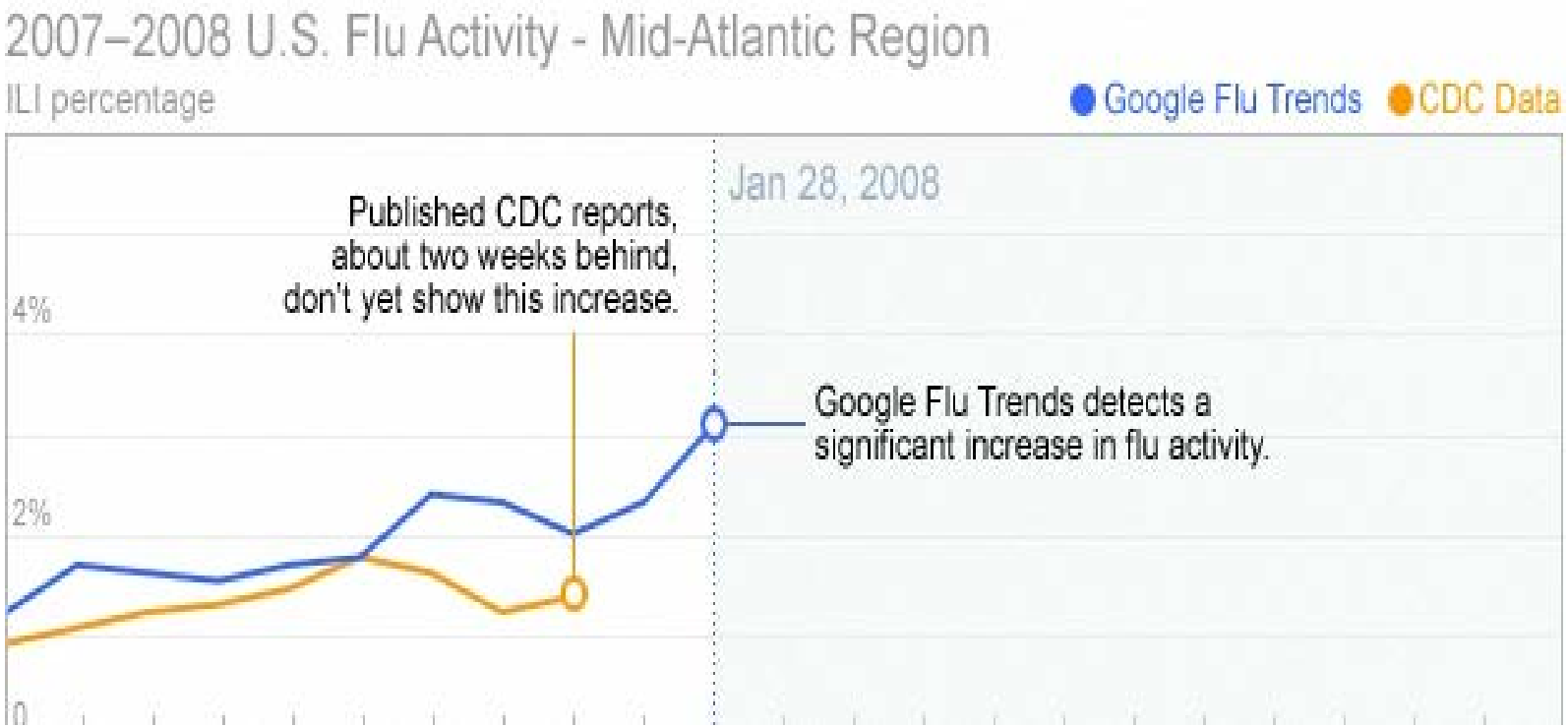
Google Flu



States | **Cities** (Experimental) - Click on a city below to chart the flu trend a



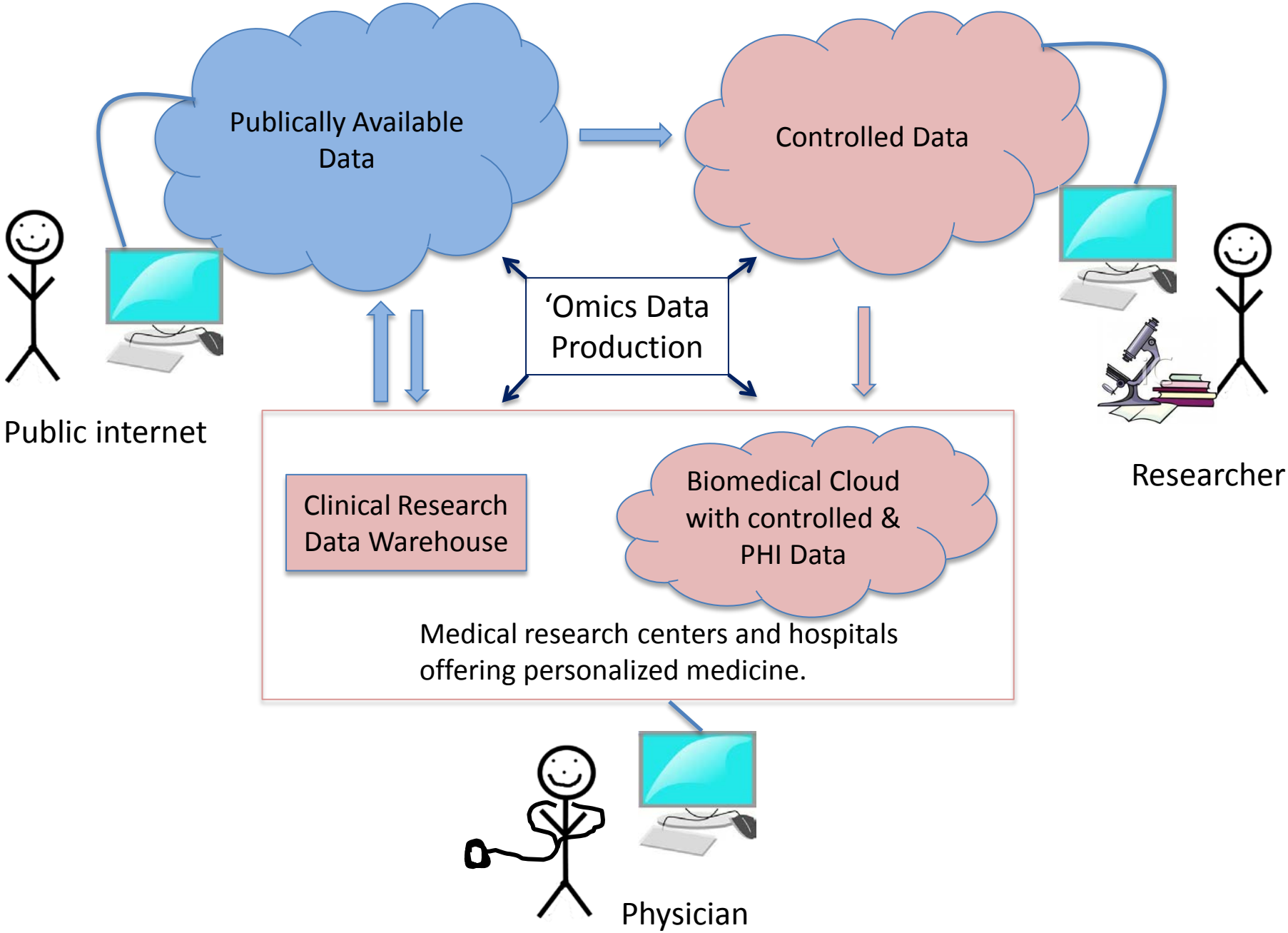
Google Flu vs. Centers for Disease Control Monitoring



Sedentary lifestyle and cancer risk

Cancers are associated with being overweight and/or sedentary lifestyle:

Breast cancer among postmenopausal women
Colorectal cancer
Endometrial cancer
Esophageal adenocarcinoma
Kidney cancer
Pancreatic cancer


Publically Available Data
Controlled Data
'Omics Data Production
Public internet
Researcher
Clinical Research Data Warehouse
Biomedical Cloud with controlled & PHI Data
Medical research centers and hospitals offering personalized medicine.
Physician

Users

Bob Grossman, IGSB/U.Chicago

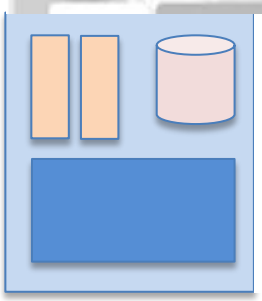
Institute for Genomics & Systems Biology

Local 'Omics Data Production

Bionimbus Cloud

Bionimbus is a cloud-based system for managing, analyzing and

www.bionimbus.org



Scales to tens of thousands/millions of genomes (exabyte sized data)

10Gbps

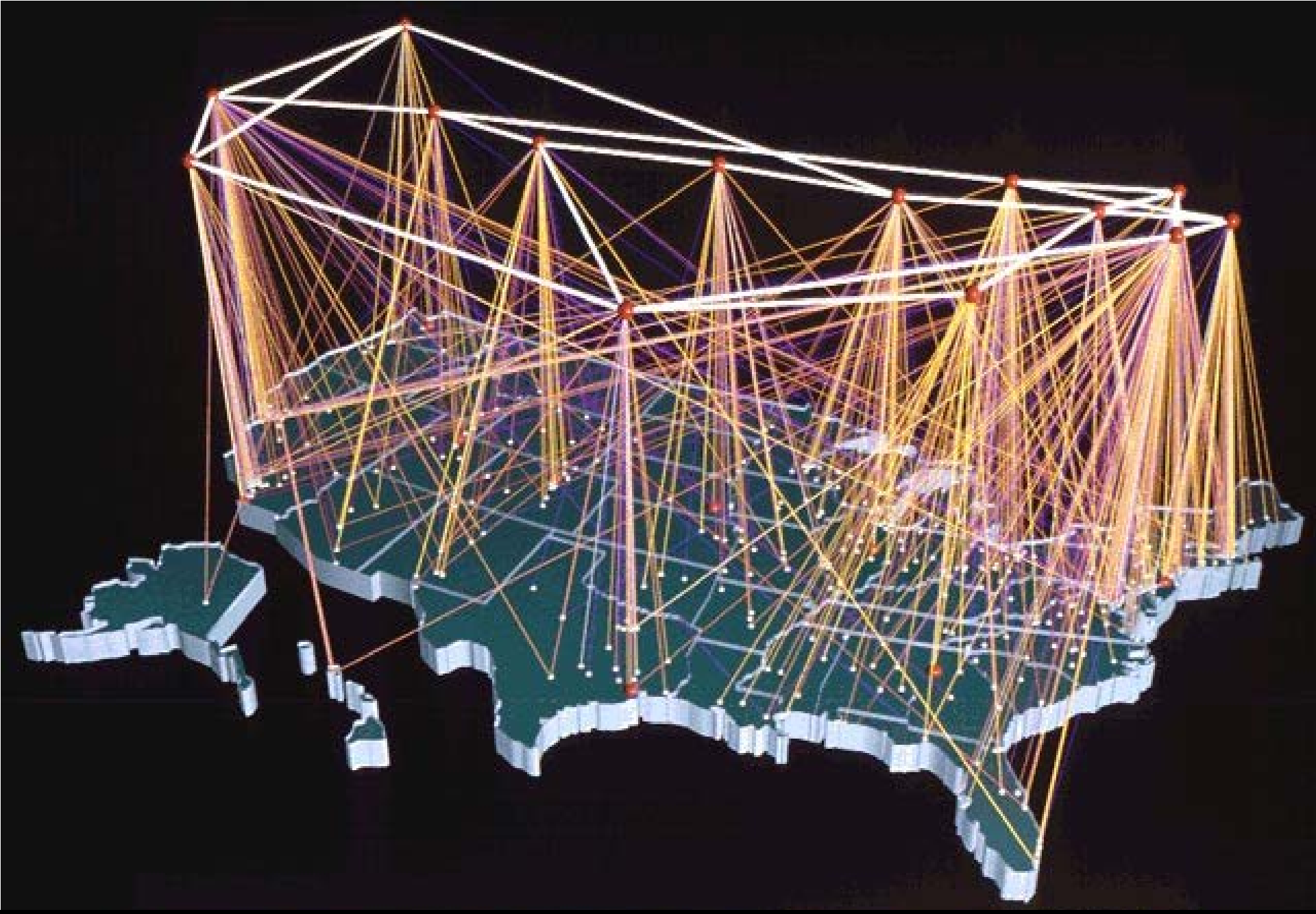
illumina Complete genomics

Genome Centers

Beijing Institute of Genomics, Chinese Academy of Sciences

NCBI

amazon web services™



United States Fast Internet Backbone

Eight requirements for a Biomedical Cloud:

1. *Appropriate security.*
2. *Secure communications with private clouds.*
3. *On-demand and scalable storage.*
4. *On-demand and scalable analysis.*
5. *Scalable ingestion of data.*
6. *Support data liberation.*
7. *Peer with other private and community clouds.*
8. *Peer with public clouds.*

Grossman and White, JIM 2012