ASU Complex Adaptive Systems Initiative
Arizona State University

Institute of Medicine
of the National Academies

A Systems-Based Approach to Cancer Informatics

Dr. George Poste
Chief Scientist, Complex Adaptive Systems Initiative
and Del E. Webb Chair in Health Innovation
Arizona State University
george.poste@asu.edu
www.casi.asu.edu

IOM National Cancer Policy Forum Workshop on
Informatics Needs and Challenges in Cancer Research
Washington, DC ● 28 February 2012

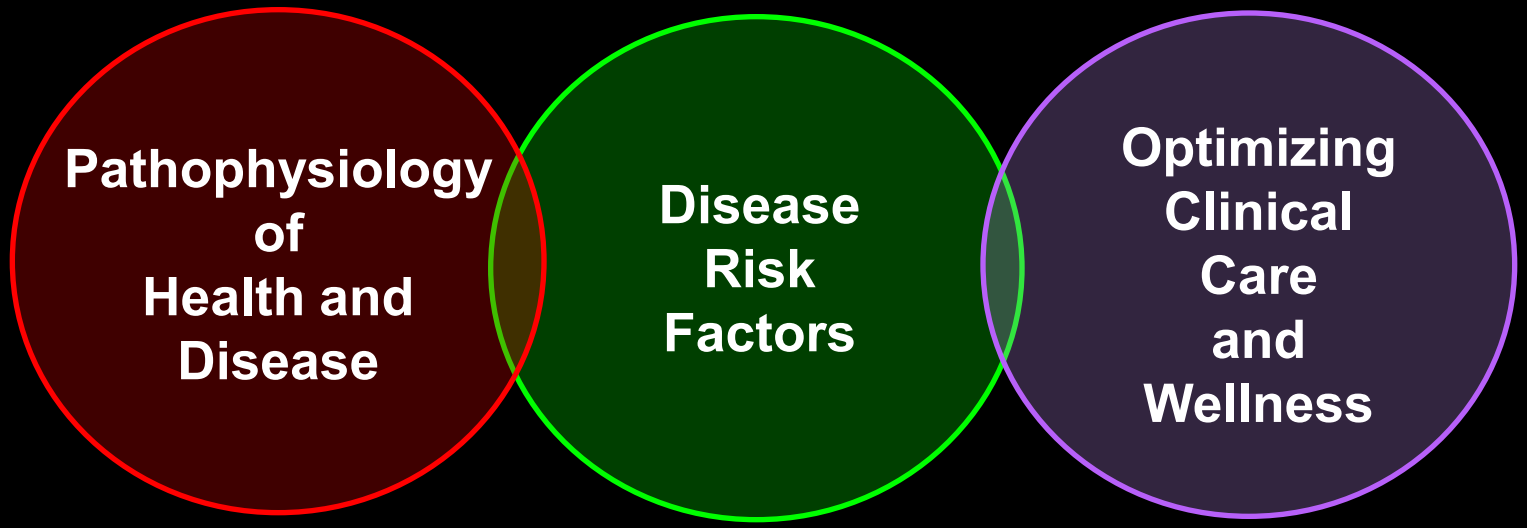
ASU COMPLEX ADAPTIVE SYSTEMS INITIATIVE
ARIZONA STATE UNIVERSITY

Slides available @ www.casi.asu.edu

Declared Interests:

- Board of Directors: Monsanto, Exelixis, Caris Life Sciences
- Scientific Advisory Board: Burrill and Co., Synthetic Genomics, Anacor
- I0M Forum on Global Infectious Diseases
- USG Activities: DoD, DHS

Knowledge Networks in Biomedicine



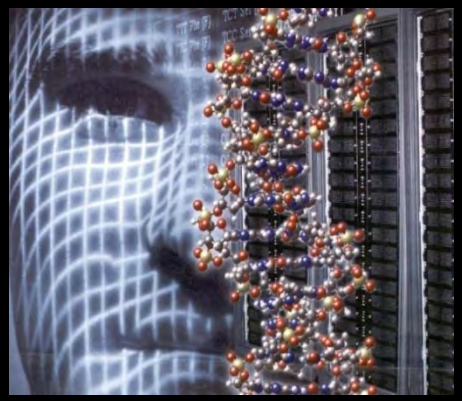
Addressing Unmet Medical Needs

Balancing Infinite Demand and Finite Resources

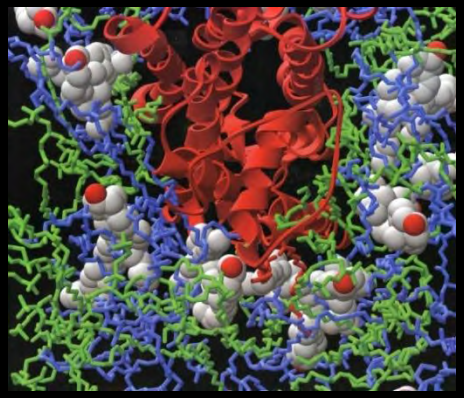
VALUE

Mapping The Molecular Signatures of Disease: The Intellectual Foundation of Rational Diagnosis and Treatment Selection

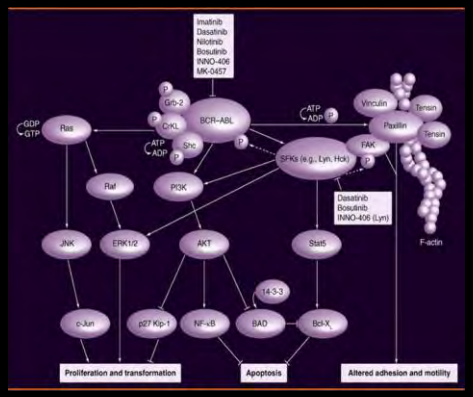
Genomics



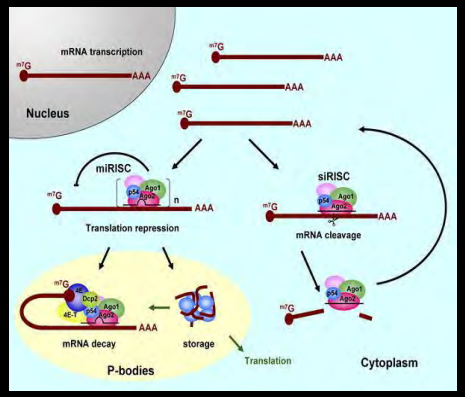
Proteomics

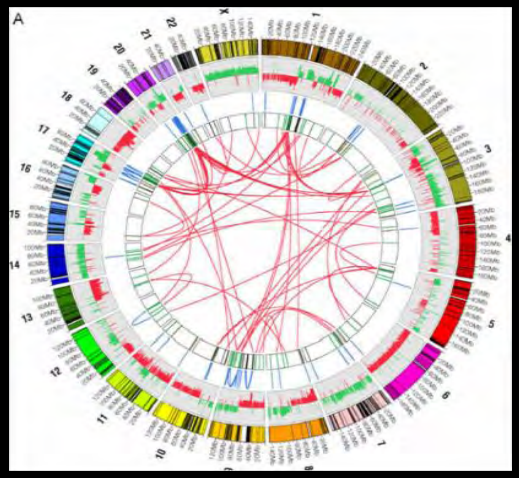


Molecular Pathways and Networks



Network Regulatory Mechanisms





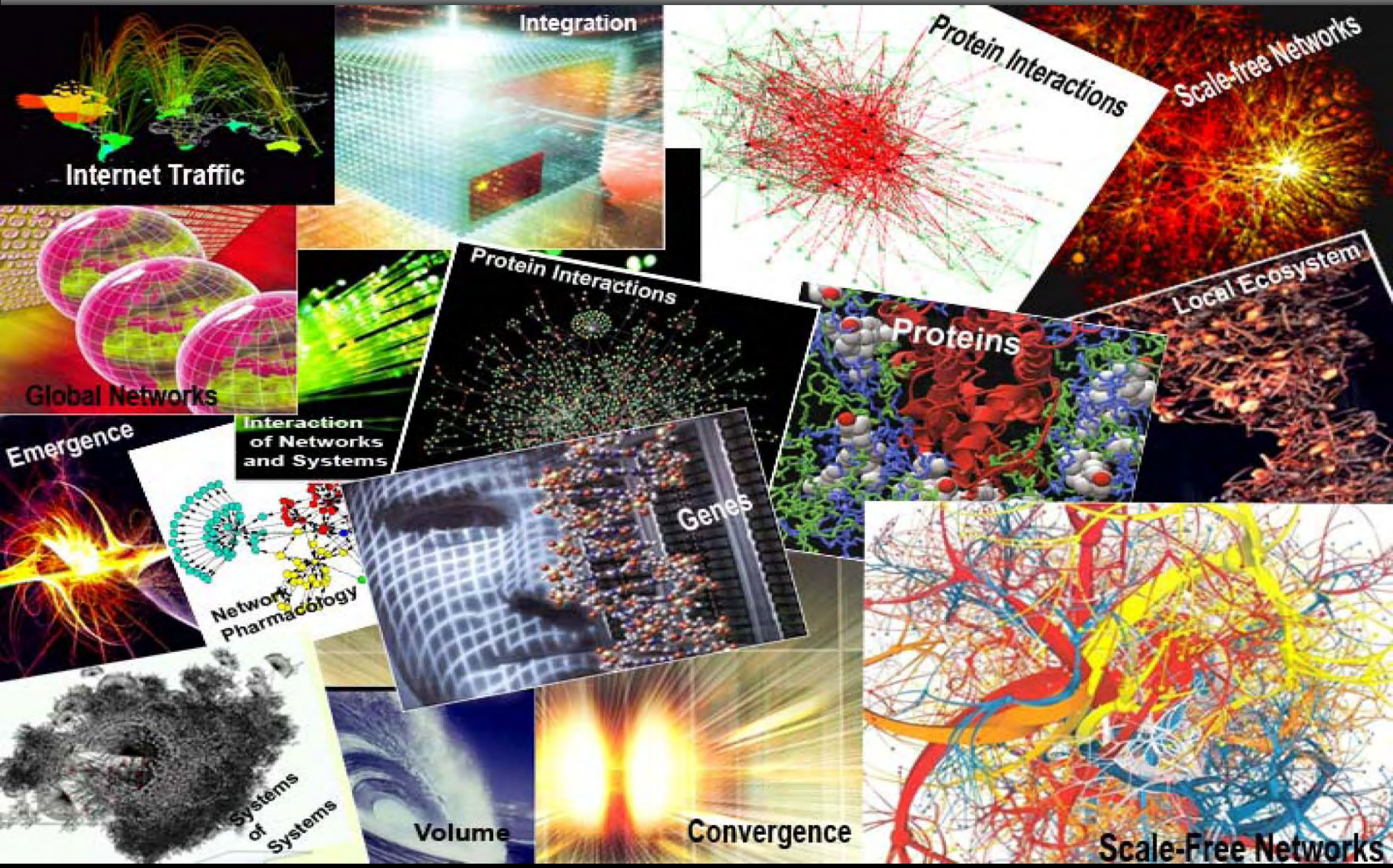
ID of Causal Relationships Between Network Perturbations and Disease

Patient-Specific Signals and Signatures of Disease or Predisposition to Disease

Critical Challenges for Biomedical R&D

- acceleration of discovery phase knowledge without parallel gains in successful clinical translation and commercial ROI
- paucity of validated biomarkers for early detection and Rx response/resistance profiling
- unacceptable high rates of failure of candidate Rx in clinical trials
- major knowledge gaps for rational discovery strategies to provide solutions for late onset chronic diseases
  - cancer, diabesity and neurodegeneration

Data: The Fastest Growing Resource on Earth

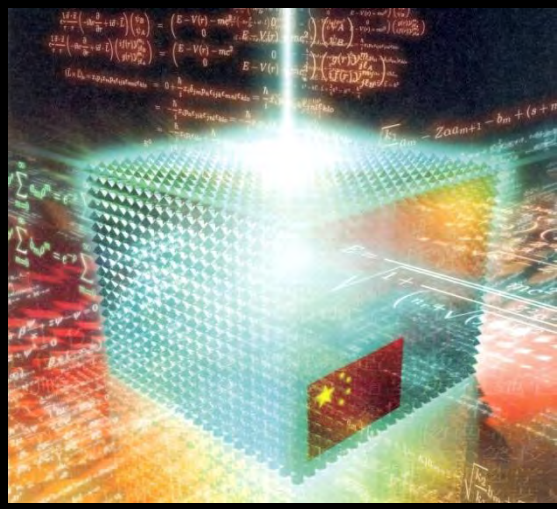


Managing “Mega-Data” in Biomedicine

volume, variety, velocity

computational scale

global networks



bench to bedside: multiscale heterogeneity

integration

Overarching Themes in Meeting the Biomedical Informatics Challenge

- Systems
- Scale, Standards and Sharing
- Software, Storage and Security
- Sustainability
- Social Issues:
  Changing Minds and Changing Behaviors

Data Production in Biomedicine

more data	but	less validated data: replication (research); fit for purpose (regulatory); authenticity (web information)
more powerful high throughput research tools	but	sample poor, high dimensionality plus inadequate analytical and statistical rigor
technology convergence and multi-disciplinary datasets	but	data integration handicapped by institutional and mental models dominated by single discipline expertise
more participants, locations and distributed data	but	pervasive lack of interoperable exchange formats, robust ontologies and vocabularies, standards for data annotation, analysis and curation, slow evolution of federated knowledge networks and open systems

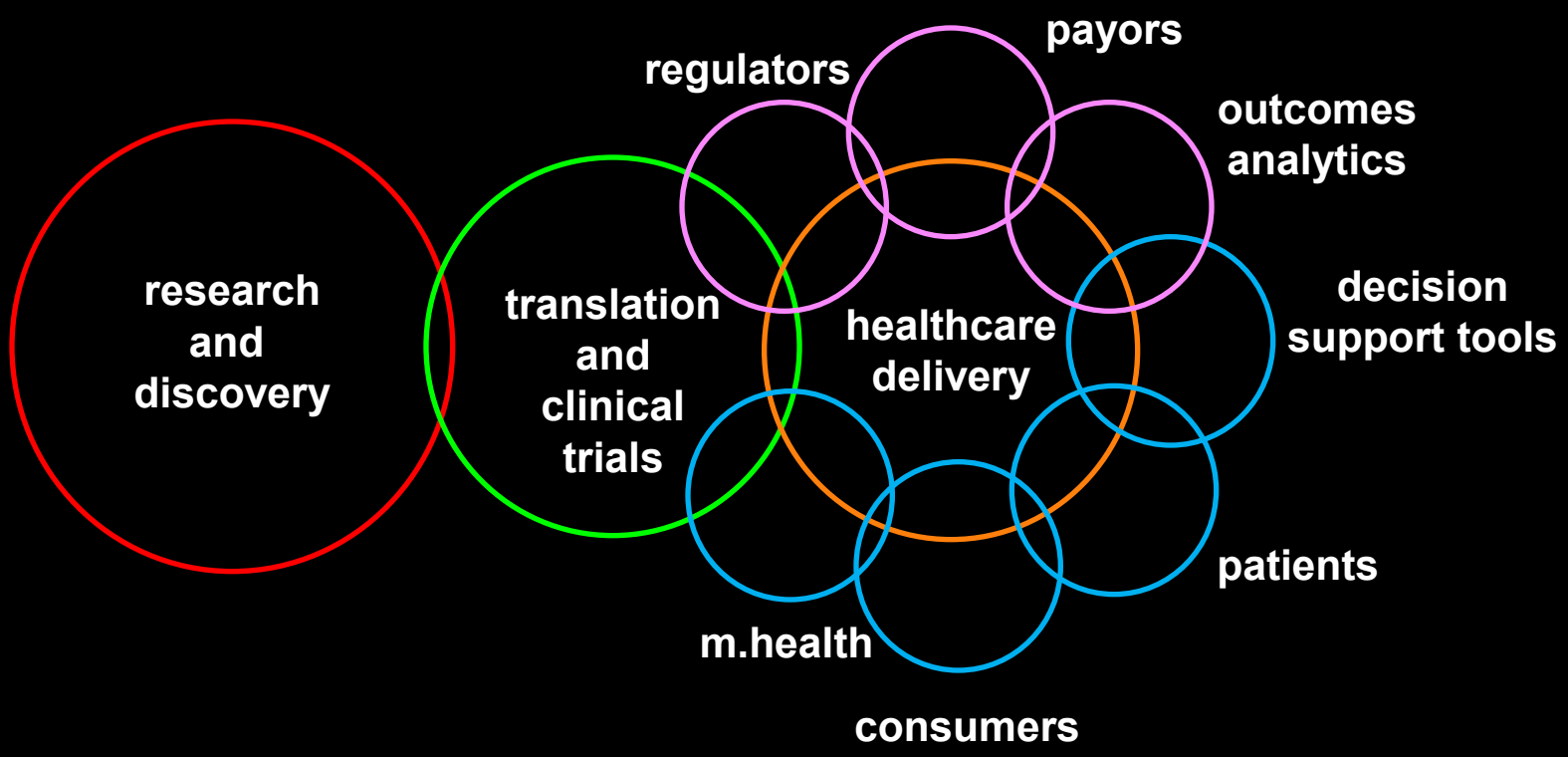
Data Analysis and Utilization in Biomedicine

more need for rapid, real-time data access	but	data trapped in balkanized and hierarchical organizational structures/databanks
more need for quantification and precision analytics	but	insufficient trained personnel for large scale data ensembles and analytics
more complexity, uncertainty and faster response:decision times	but	escalating gaps in institutional/individual cognitive and analytical capabilities
increasing rate of change	but	increasing rate at which knowledge and competencies depreciate

Informatics Needs and Challenges in Biomedical Research

- most current BIX/HIX approaches lack the agility and extensibility to meet projected needs
- need for new approaches for end-to-end system design
  - proactive articulation of system requirements and quality parameters
  - multiple end-user communities
  - omnipresent risk of new silos with accompanying waste and cost of constant failure and redesign due to inattention to facile and seamless exchange frameworks
  - omnipresent risk of simply recreating new silos with accompanying time and cost of recurrent failure and constant redesign due lack of proactive attention to task complexity (balancing system merits versus entrenched legacy silos)
  - no easy task!

The Need for Facile, Seamless Data Exchange Formats for Large Scale Biomedical Data Systems



The Rise of Data-Driven, Data-Enabled Science and Technology

- data changed by computing
- computing changed by data
- data are now fundamentally networked
- increasing fraction of data is ‘born digital’
- ever larger data sets become increasingly unmovable with existing infrastructure
- simulations using data and meta-analytics amplify the data metaverse

The Fourth Paradigm: Data-Driven Knowledge, Intelligence and Actionable Decisions

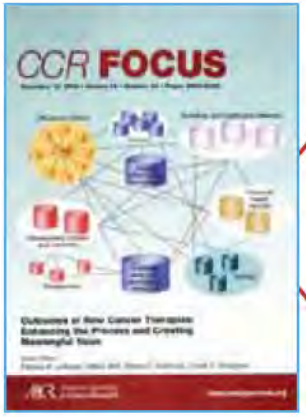
- changing the nature of discovery
  - hypothesis-driven versus hypothesis-generating unbiased analytics of large datasets (patterns, rules)
- changing the nature of explanation
  - statistical probabilities versus unitary values
- changing the cultural process of knowledge acquisition
  - large scale collaboration networks, open systems
- changing knowledge application
  - increased quantification and decision-support systems
- changing cognitive frameworks, intellectual capabilities and competencies for knowledge-intensive competitiveness in multiple domains
- changing education and training

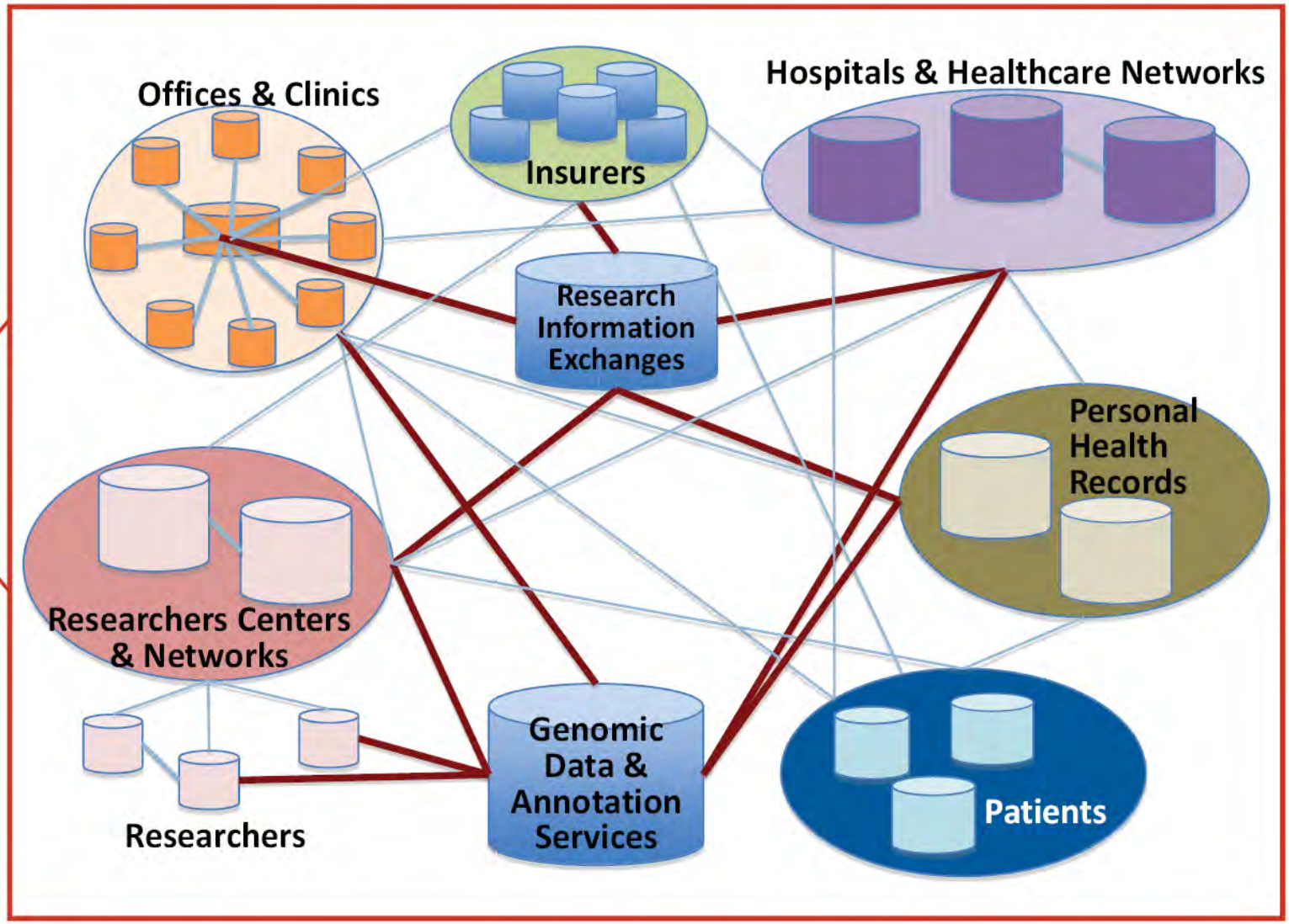
- are we building systems and infrastructure that merely support the collection of data?

or

- an integrated knowledge ecosystem that supports data validation, sophisticated analytics, evidence generation and actionable knowledge to drive a learning healthcare enterprise?

The Multiple Users and Complex Connectivities for Seamless Information Transfer in the HIT Ecosystem





W. Dalton et al, Clin Cancer Res; 16 (24) December 15, 2010

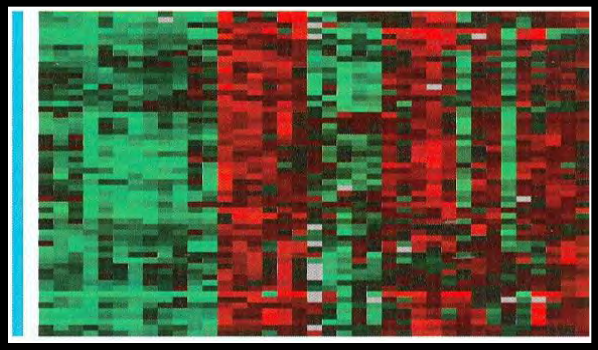
Evaluation of CaBIG by Board of Scientific Advisors (March 2011) and New Oversight Ad Hoc Subcommittee (ongoing)

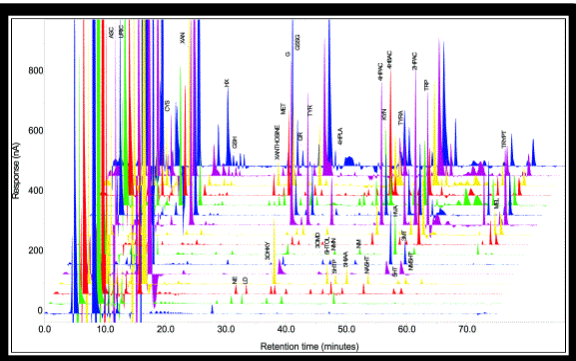
- strong support for original vision and goals
- clinical informatics tools/algorithmic advances viewed as mission critical
- technology-centric, one-size-fits all approach to data management and tensions with end user systems
- lessons learned or reversion to fragmented academic efforts divorced from support of investigational trials and improved care delivery

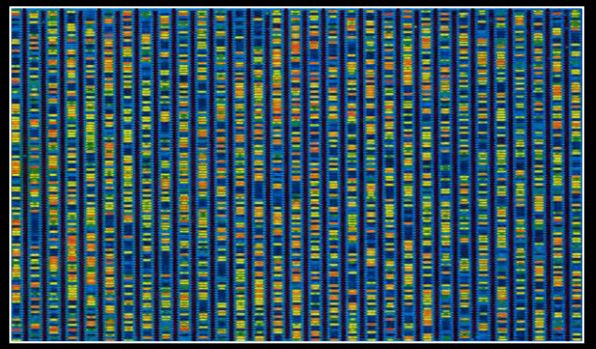
Analytical Platforms for the Elucidation of the Design and Regulation of Complex Biological Networks

Massively Parallel Biosignature Profiling

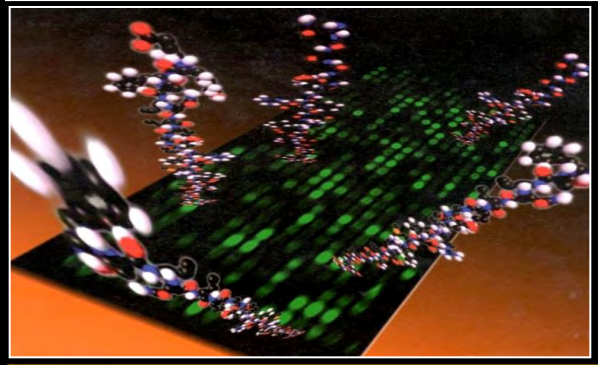
genomics	proteomics	immunosignatures

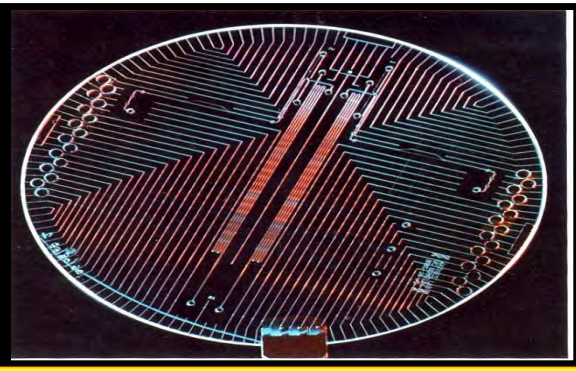


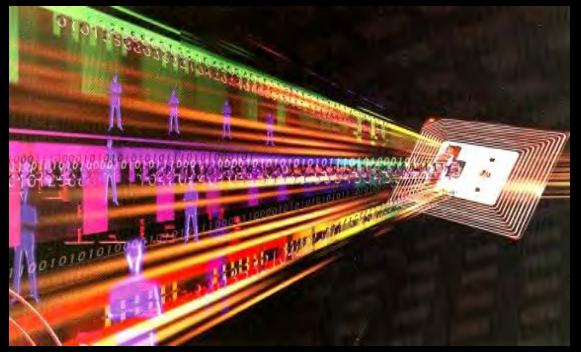




automated, high throughput multiplex assays	novel test formats and devices (POC)	complex signal deconvolution

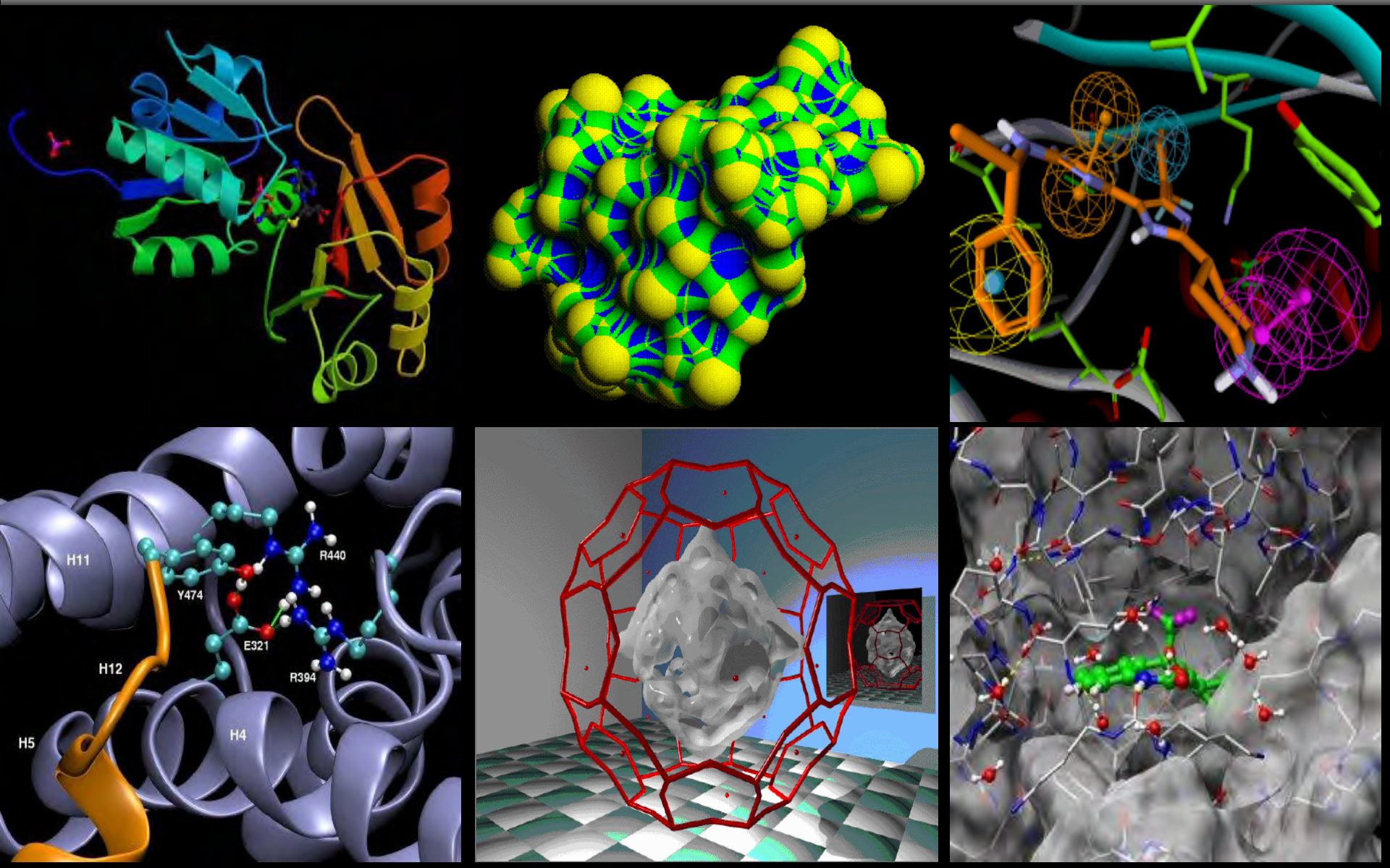






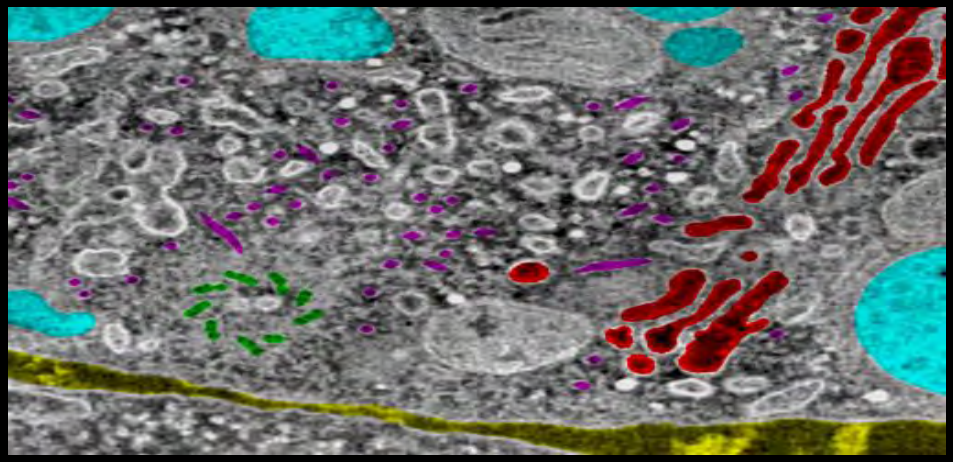
Large Datasets, Standardization and New Computational Analytics

Computational Chemistry, Molecular Modeling and Ligand-Target Design For SAR

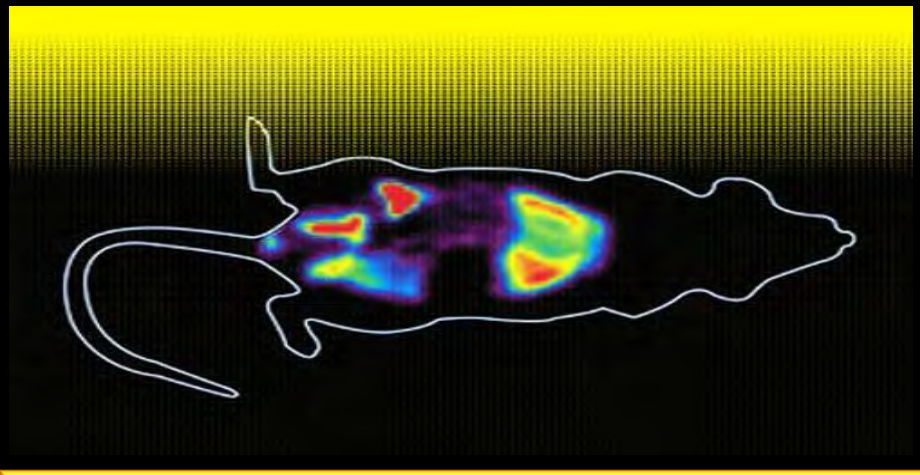


Imaging Technologies

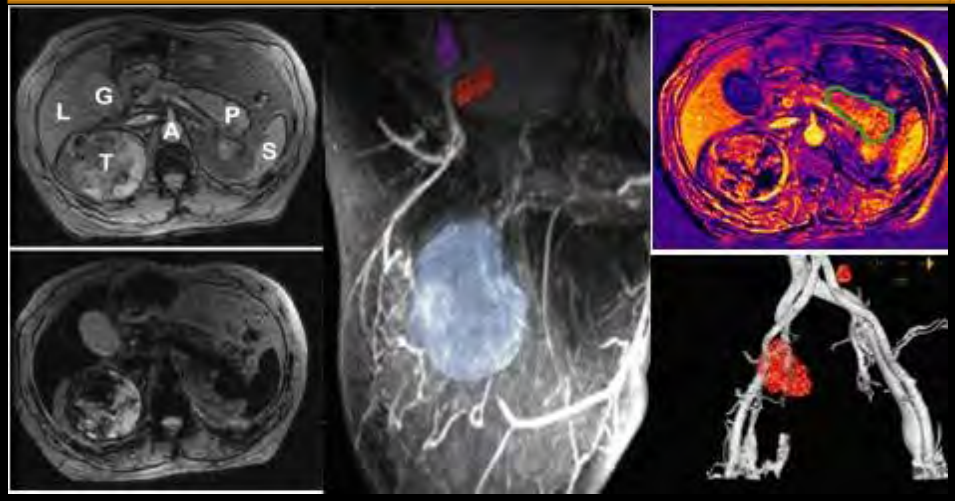
Cellular High Content Assays



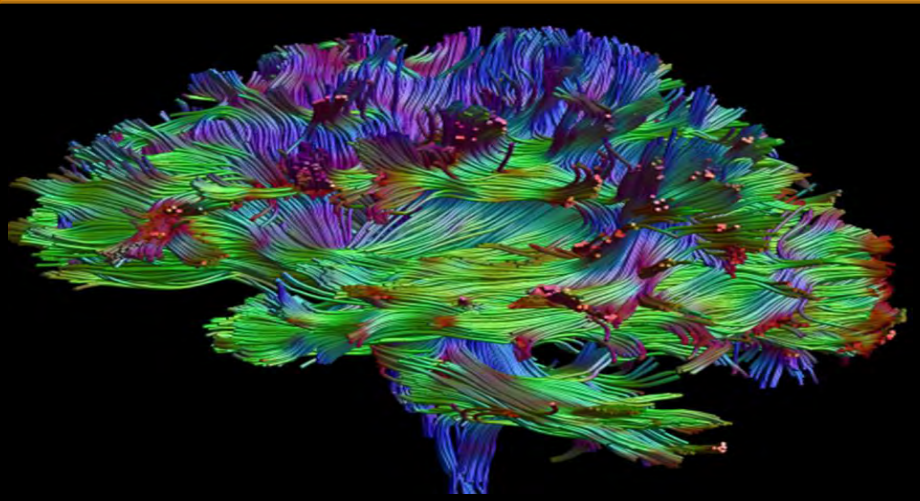
Preclinical



Clinical



Visualization

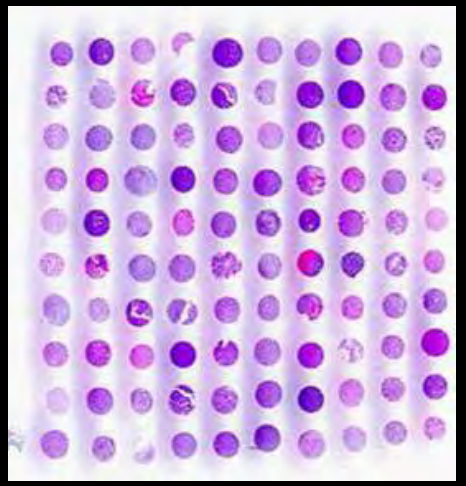


Rigorous Selection of Specimen Donors and Specimen Collection

primacy of standardized clinical phenotyping and ability to correlate biomarkers with disease stage and outcomes via stringently annotated health records

challenge of obtaining fresh tissue

poorly standardized biospecimen collection, control of pre-analytical variables and assay standards



uncertain value of legacy tissue blocks (technical/regulatory)

Quotes for Prominent Display in Every Biomarker Research Laboratory

"The technological capacity exists to produce low-quality data from low-quality analytes with unprecedented efficacy."

"We now have the ability to get the wrong answers with unprecedented speed."

Dr. Carolyn C. Compton
Director, Office of Biorepositories and Biospecimen Research
National Institutes of Health
'Institute of Medicine Workshop, July 2010'

Pervasive Problems in Biomarker Identification and Validation

The Small 'N' Problem: Bias, Overfitting, Apophenia

JAMA (2011) 305, 2200

Comparison of Effect Sizes Associated With Biomarkers Reported in Highly Cited Individual Articles and in Subsequent Meta-analyses

John P. A. Ioannidis, MD, DSc
Orestis A. Panagiotou, MD

Context Many biomarkers are proposed in highly cited studies as determinants of disease risk, prognosis, or response to treatment, but few eventually transform clinical practice.

Objective To examine whether the magnitude of the effect sizes of biomarkers proposed in highly cited studies is accurate or overestimated.

Data Sources We searched ISI Web of Science and MEDLINE until December 2010.

Study Selection We included biomarker studies that had a relative risk presented in their abstract. Eligible articles were those that had received more than 400 citations in the ISI Web of Science and that had been published in any of 24 highly cited biomedical journals. We also searched MEDLINE for subsequent meta-analyses on the same associations (same biomarker and same outcome).

MANY NEW BIOMARKERS ARE continuously proposed[1-3] as potential determinants of disease risk, prognosis, or response to treatment. The plethora of statistically significant associations[4,5] increases expectations for improvements in risk appraisal.[6] However, many markers get evaluated only in 1 or a few stud-

Failure to Work to Industry Standards

Nature Rev. Drug Disc. (2011) 10, 643

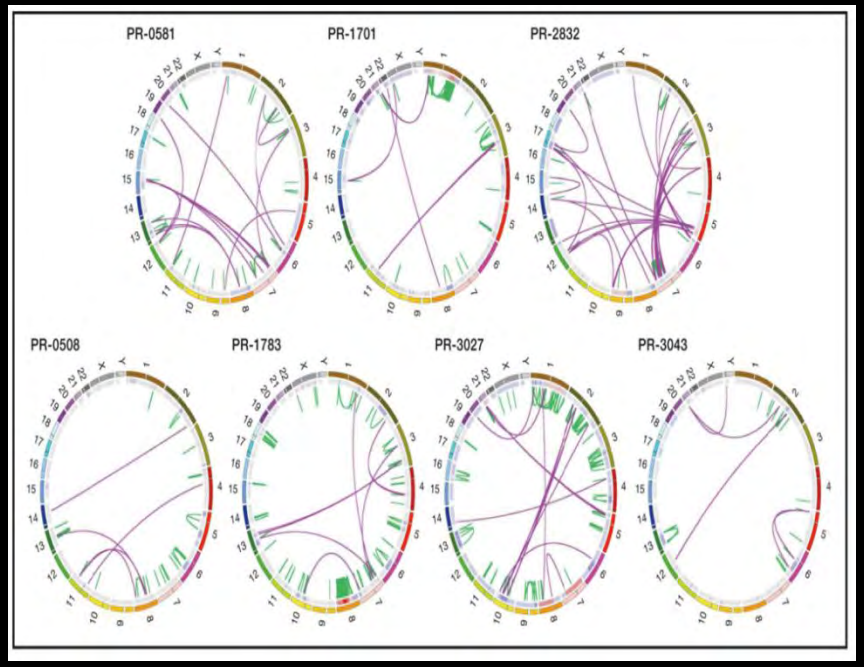
Reliability of 'new drug target' claims called into question

Bayer halts nearly two-thirds of its target-validation projects because in-house experimental findings fail to match up with published literature claims, finds a first-of-a-kind analysis on data irreproducibility.

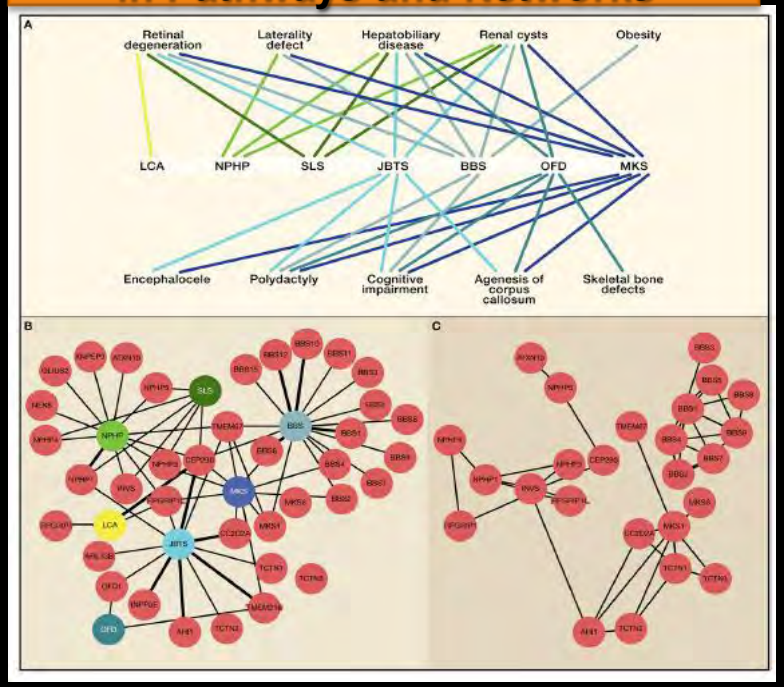
Prepare for the "Tsunami of Genomic Information" ASCO Presidential Address: Dr. George Sledge Chicago, 5 June 2011

- "the day when a patient walks into her oncologists office carrying a memory stick containing personal genomic information could be less than a decade away"
- "when data are that cheap….things will get very, very complicated"

Exome- or Whole Genome Sequencing



Disease-Associated Perturbations in Pathways and Networks



When Will Partial- and Whole Genome Sequencing Become 'Just Another Laboratory Value' in Patient Care?

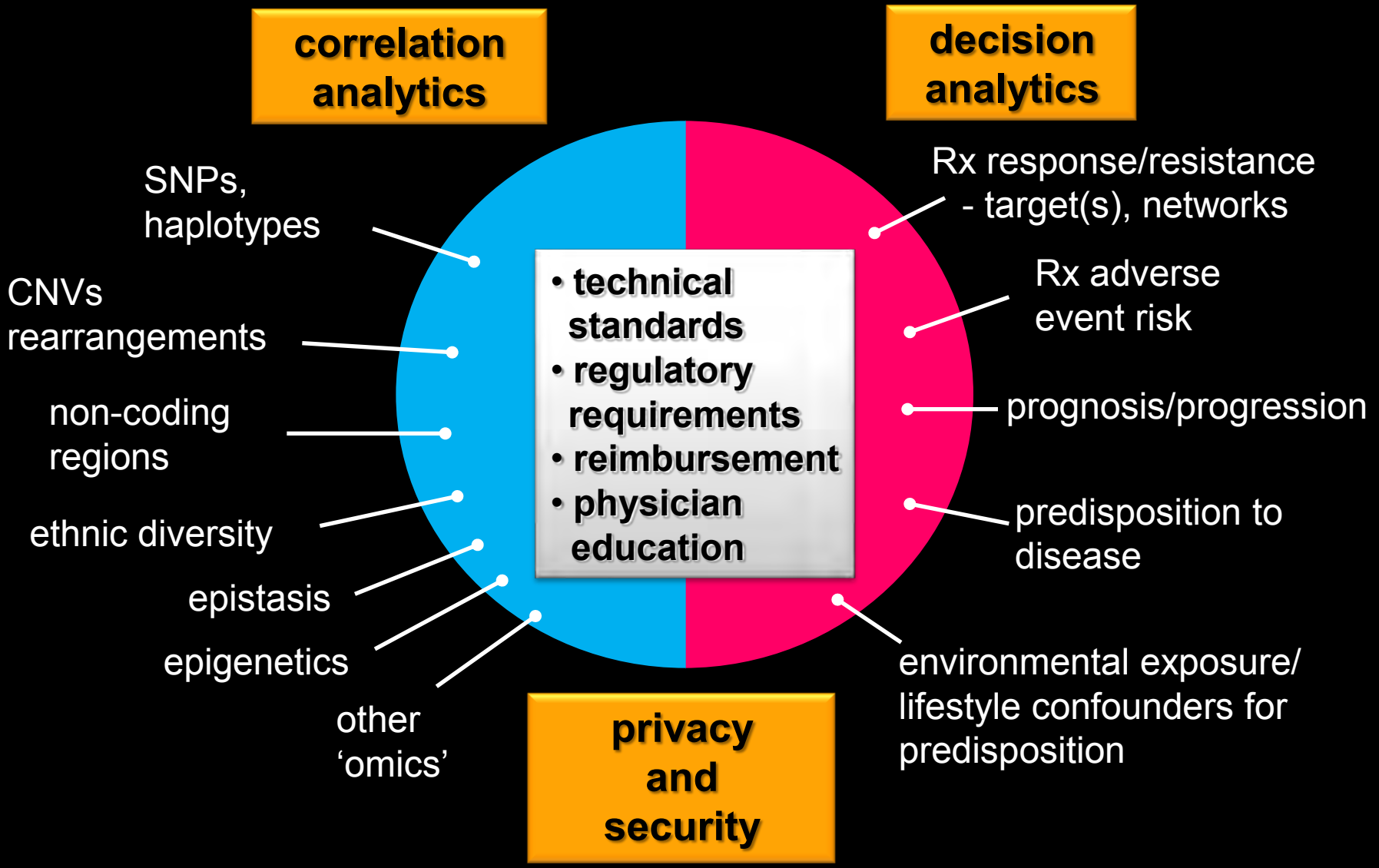
The MinION Sequencer Oxford Nanopore

- disposable USB pocket-sized sequencer
- $500-1000
- 150 Mbp sequence/hour
- smaller instrument versus pending benchtop GridION 2K machine

BGI-The Beijing Genome Institute: The World's Largest Gene Sequencing Capacity

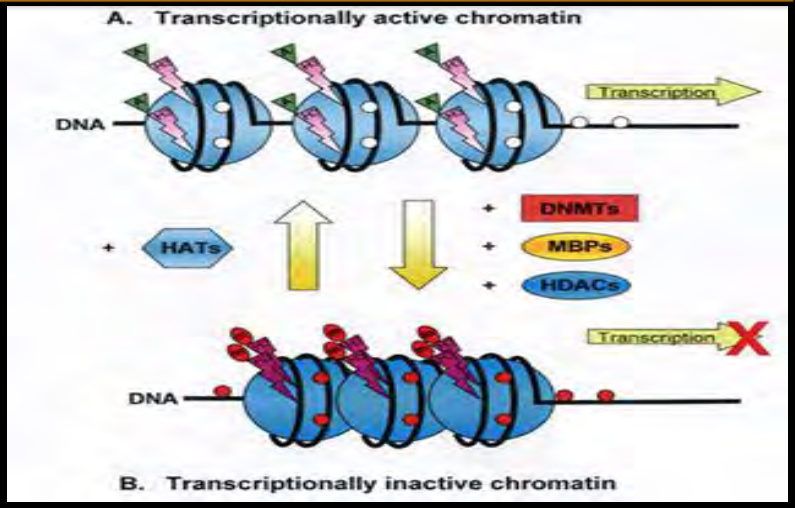
- Main Facilities in Shenzhen and Hong Kong, China
  - Branch Facilities in Copenhagen, Boston, UC Davis
- Supported by Supercomputing ~160TF, 33TB Memory
  - Large-Scale (12PB) Storage

Low Cost Whole Genome Sequencing and Molecular Medicine: Dependency on Large Scale (Massive) Data Annotation and Analytics



The Epigenome

Modulation of Gene Expression/Regulation by Environmental Factors, Xenobiotics and Rx (The Exposome)



Effect of Maternal Diet/Stress/Rx exposure on Germ Line Genome Imprinting (+ trans-three-generational?)

IHEC
International Human Epigenome Consortium

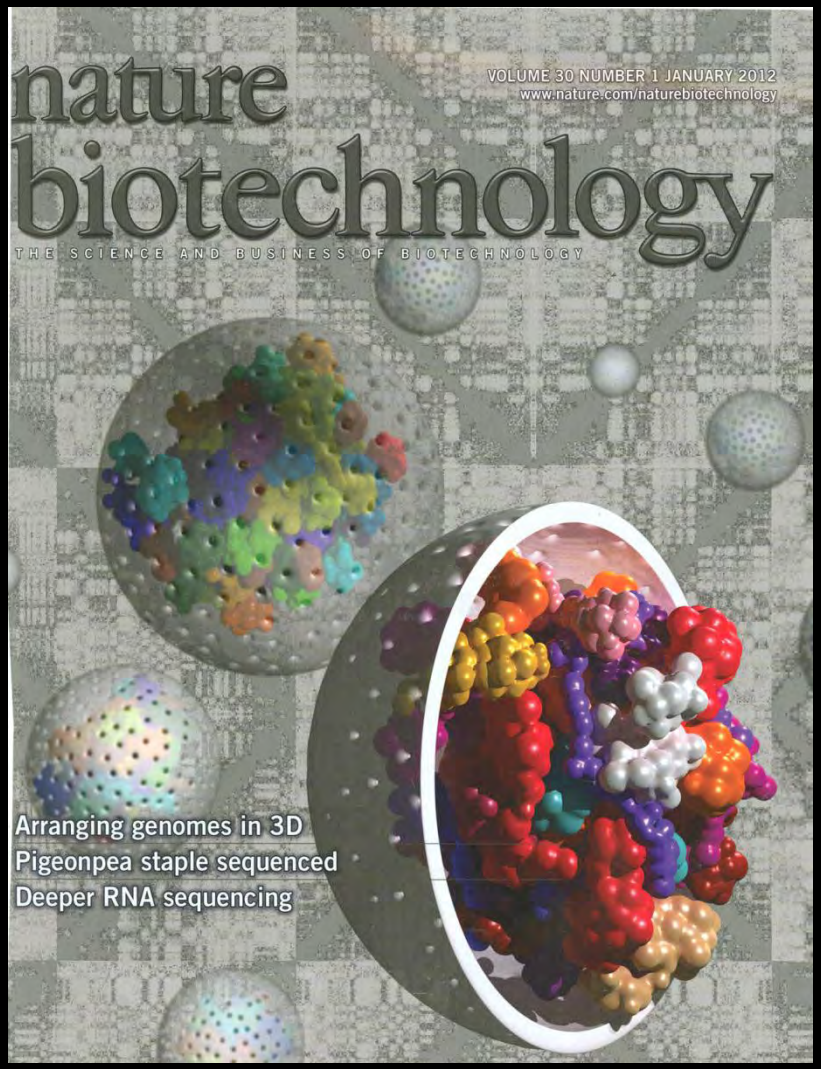
International Human Epigenome Consortium

- 1000 reference genomes by 2020

project blueprint

- launch September 2011 with €30-million
- map epigenome in 60 human blood cell classes and neoplastic counterparts

Understanding the 3D Genome in Cancer



- higher order chromatin architecture and landscape of chromosomal alterations
  - G. Fudenberg et al. (2011) Nature Biotech. 29, 1109
- influence of DNA replication timing and long-range DNA interactions on mutational landscape
  - De and Michor (2011) Nature Biotech. 29, 1103

What is A Complete and Accurate Analysis of Genome Sequence, Architecture and Regulation?

Performance Comparison of WGS Platforms
(H.Y.K. Lam et. al. 2012 Nature Biotechnol. 30, 78)

- sequencing of blood and saliva samples from same individual on Illumina and Complete Genomics Platforms at 76x coverage
- only 88.1% SNVs concordant ≡ 10,000's platform-specific calls in exons and intergenic regions
- need to supplement with exome sequencing to fill gaps in detection of coding variants
- only 26.5% indels concordant
- implications for use of WGS data for clinical decisions/regulatory submissions

Review of Validation Issues for Clinical Use of Genome Sequencing 23 June 2011

- minimum sequencing depth for reliable clinical decisions
- appropriate validation sample sets to evaluate platform accuracy
- metrics for quality of sequence assembly and alignment algorithms
- standardization of pre-analytical variables (e.g. preparation of libraries, extraction and quality control of nucleic acids, capture methods, amplification)
- source computer code(s) for analytical algorithms
- sequencers as Class III devices?

The Data Storage Challenge: The Price of Sequencing is Falling Faster Than Computer Storage Costs and Availability

- data 'triage'
  - store only data deemed relevant and/or different to reference genome(s)
  - risk of bias/ignorance about value of discarded data elements
- data compression and 'loss of precision'
  - different compression methods depending on desired end use/reuse needs
- unmapped reads cannot be compressed using current alignment frameworks
  - 10-40% of reads remain unmapped to traditional reference genomes
  - 60-70% for short RNA sequencing reads
- many samples may not be accessible/renewable for resequencing
  - cancer

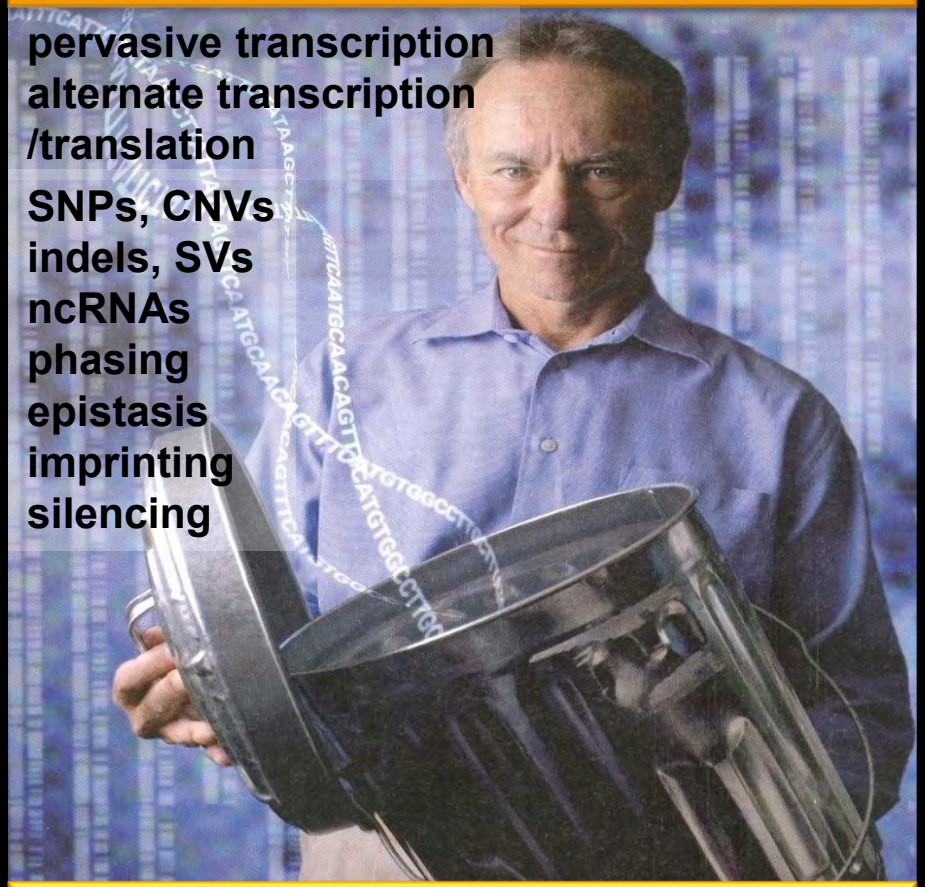
From Genotype to Phenotype

Integration of Gene Expression and Genome Sequencing Data With The Dynamics of Biological Pathways and Networks

Identification of Causal Correlations Between Genome Alterations and Disease and/or Predisposition to Disease

Individual Variation, Genome Complexity and the Challenge of Genotype-Phenotype Prediction

Junk No More!

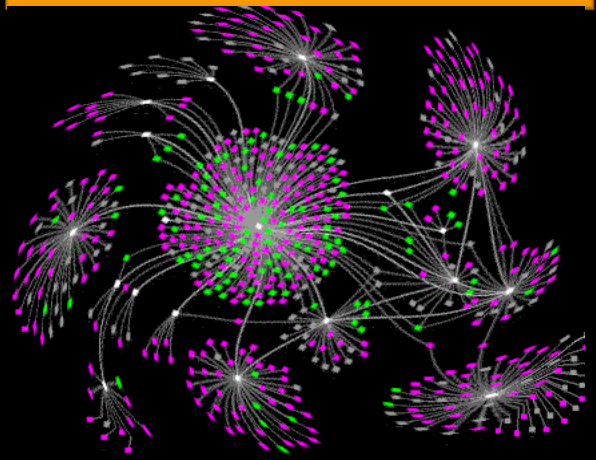


pervasive transcription
alternate transcription /translation

SNPs, CNVs
indels, SVs
ncRNAs
phasing
epistasis
imprinting
silencing

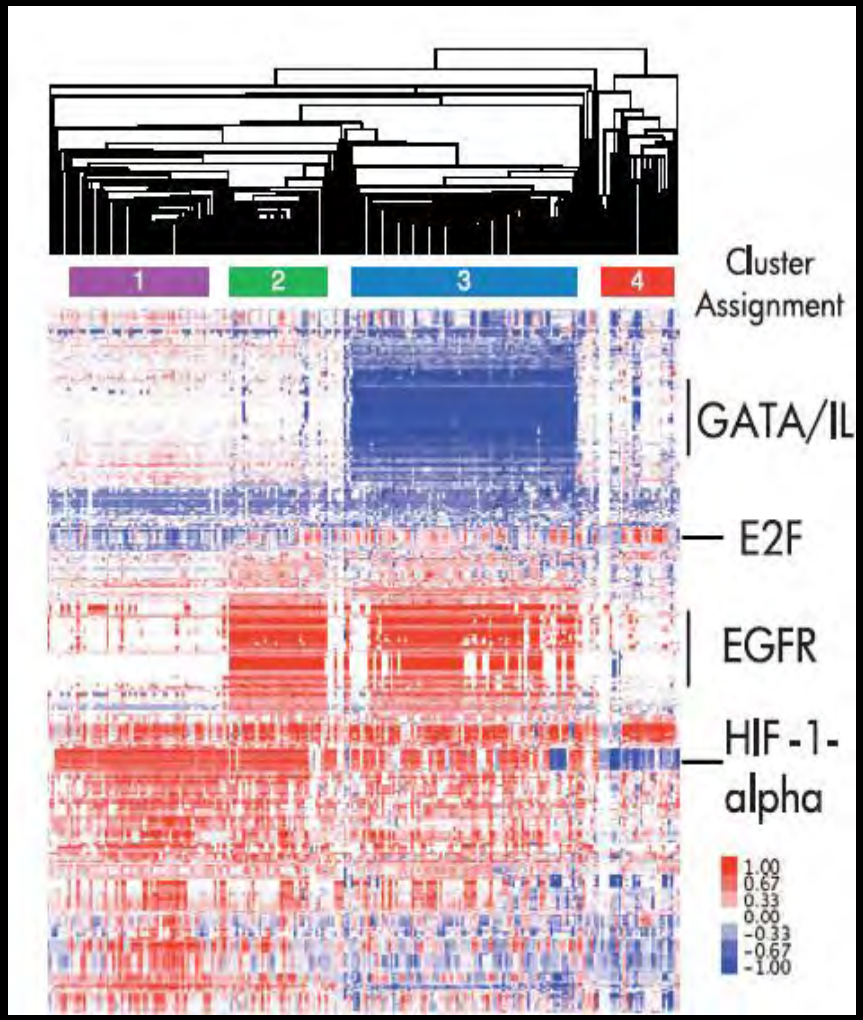
recognition of complexity of genome organization and regulation

Cell-specific Molecular Interaction Networks

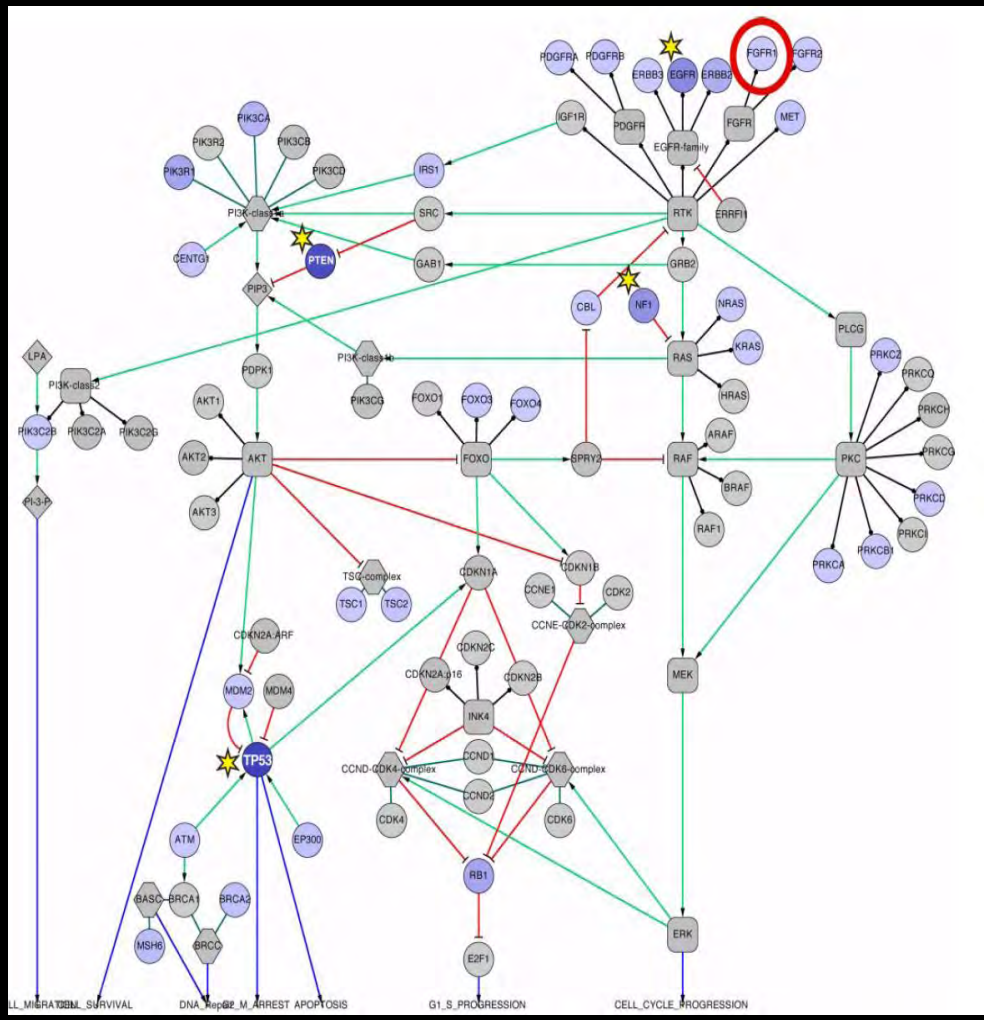


Network Perturbations in Disease

Mapping Pathways, Modules and Subnetworks in Biological Systems: The TCGA Glioblastoma Multiform Dataset and Protein Interaction Networks



From: C. J. Vaske et al. (2011)
Bioinformatics 26, i237

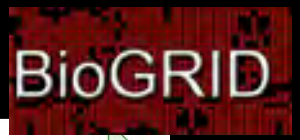


From: J. H. Morris et al. (2010)
Molec. Cell. Proteomics 9, 1703

Protein-Protein Interaction (PPI) and Pathway Databases

BIND: Berkeley Internet Name Domain

Database of Interacting Proteins
THE DIP DATABASE



DroID: The Comprehensive Drosophila Interactions Database

InterPro is an integrated database

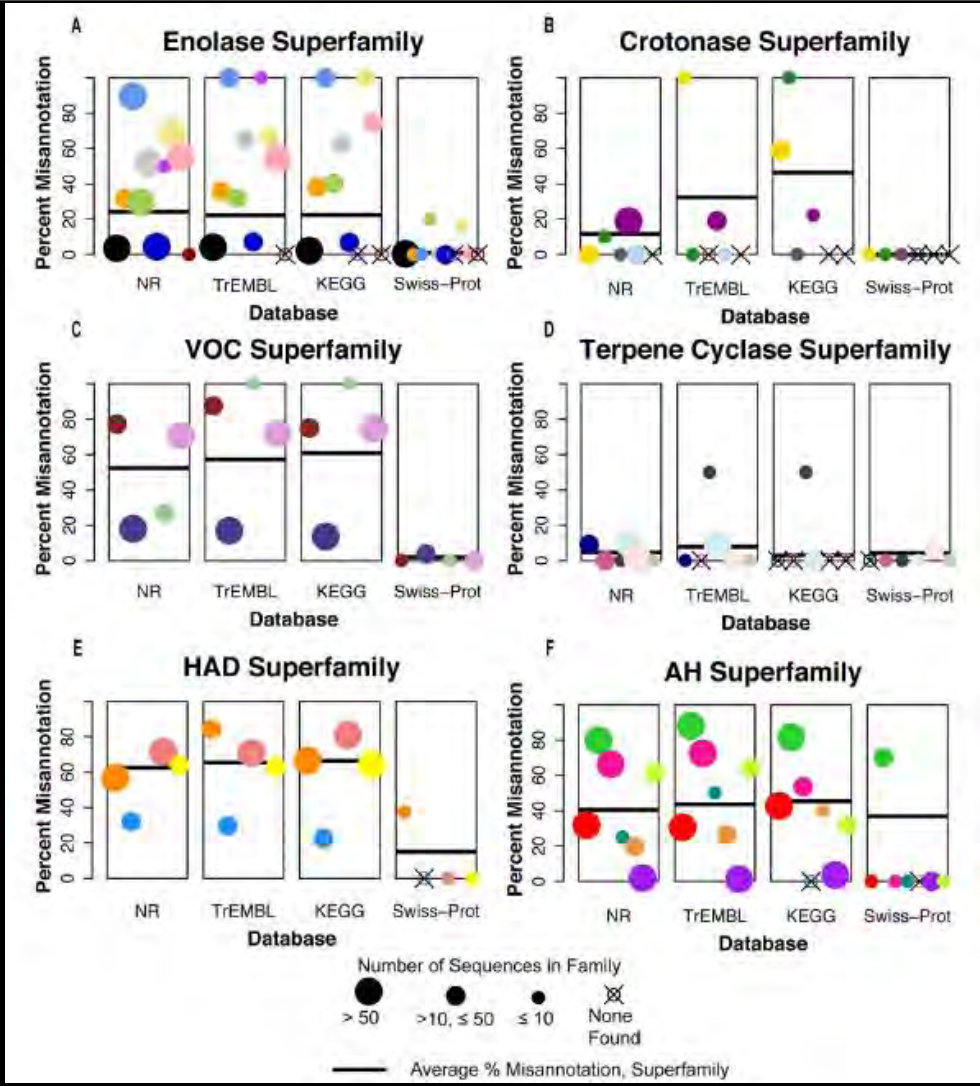
PARADIGM

Molecular INTeraction database

Massachusetts Institute of Technology

the Gene Ontology

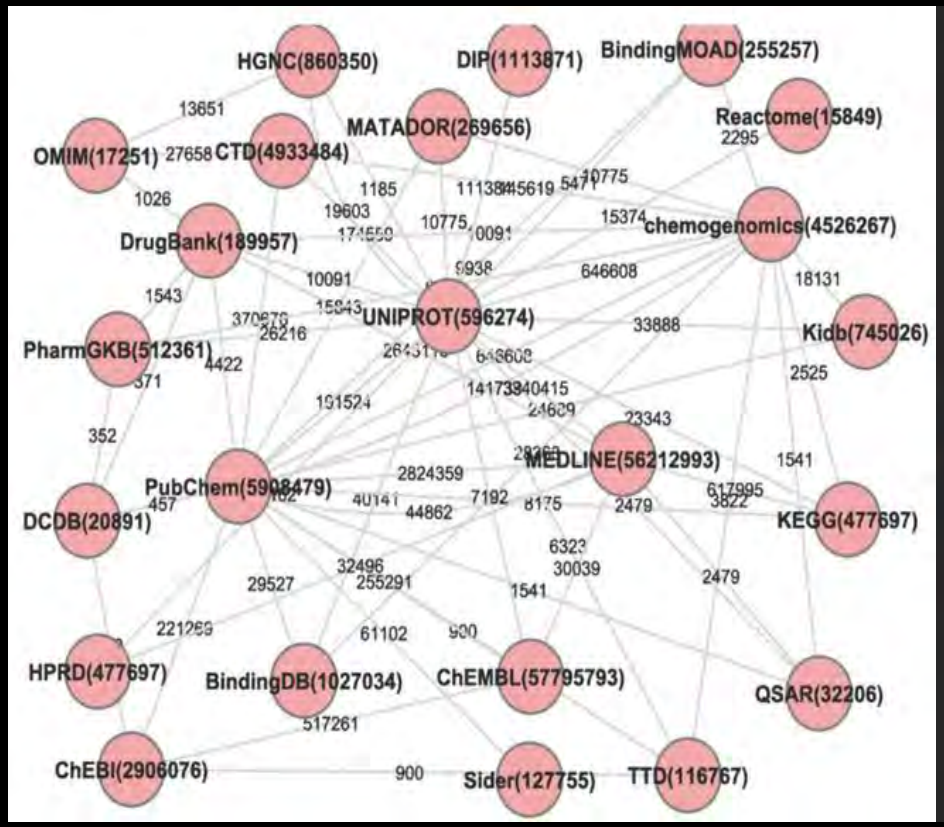
Percent Misannotation in a Series of Protein Superfamilies in Large Primary Databases (GenBank NR, TrEMBL), Secondary Databases (KEGG) and Highly Curated Databases (UniProt/Swiss-Prot)



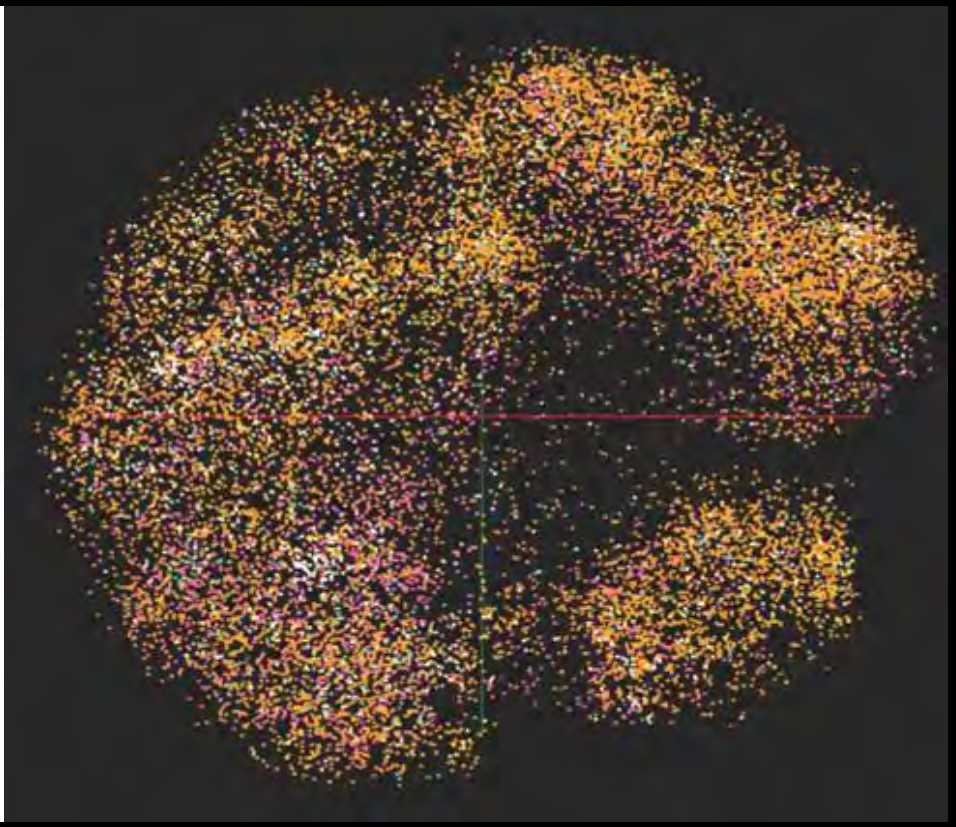
From: A. M. Schnoes et al. (2011) PLoS Comp. Biol. 5,(12) e1000605

Big Data in Drug Discovery

Chem2Bio2RDF



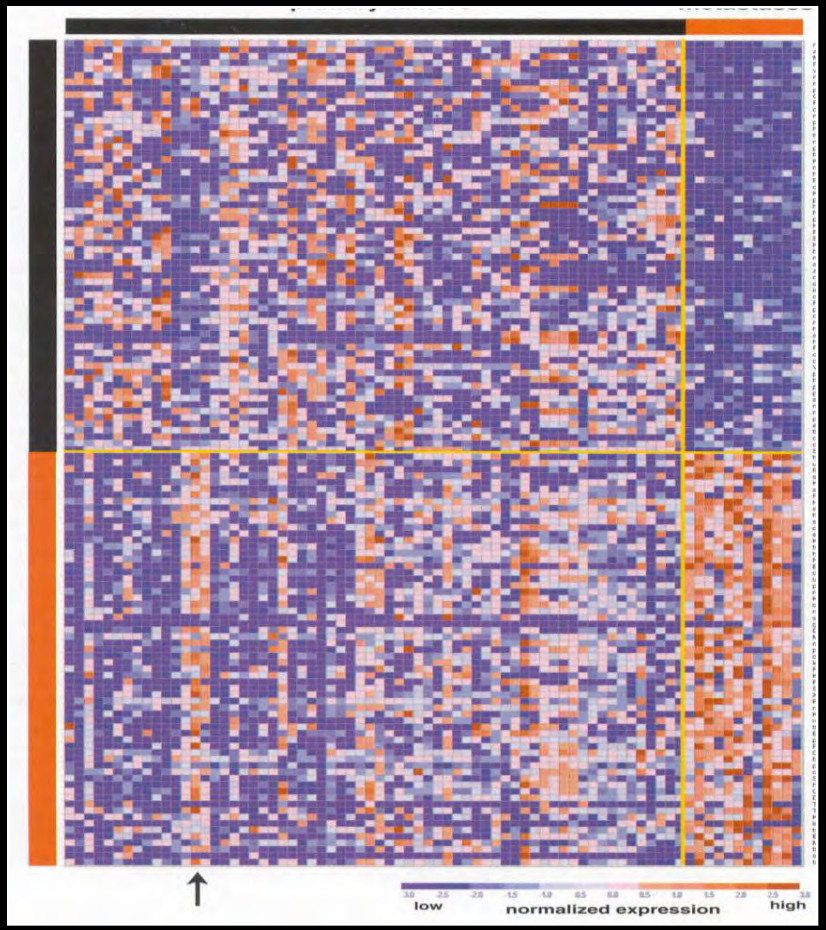
Mapping Large Scale Chemoinformatics Space



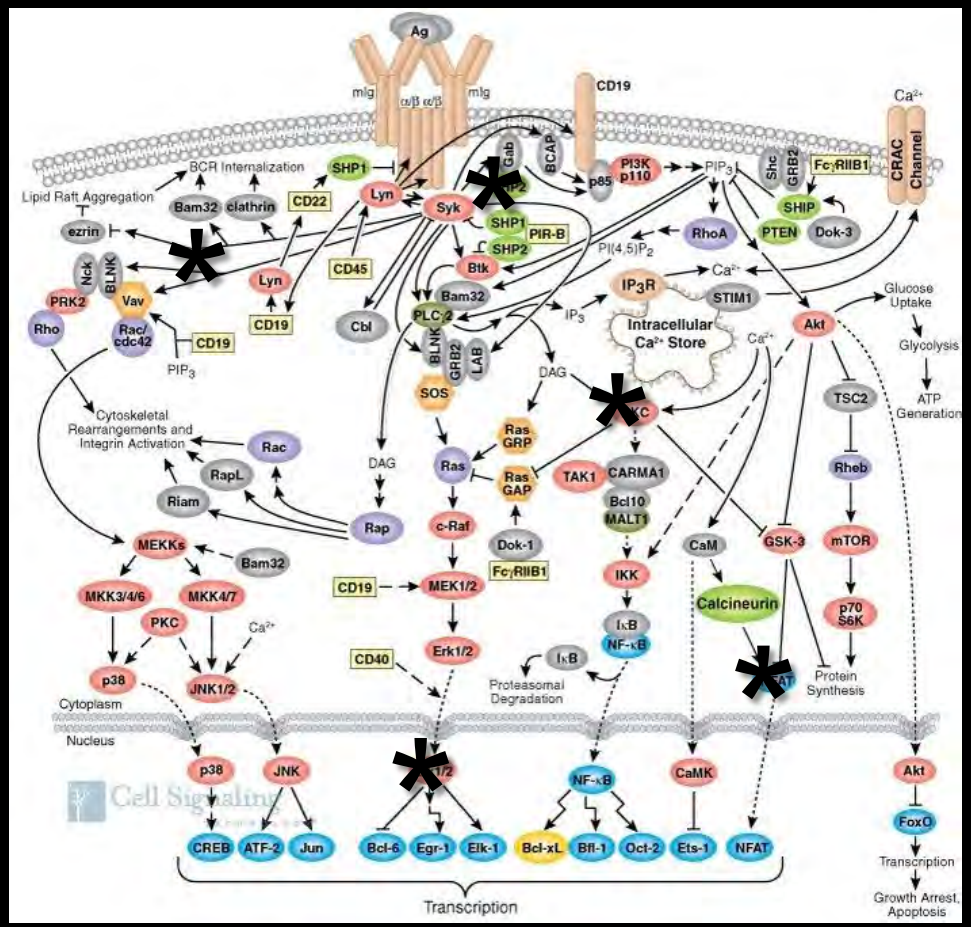
D. J. Wild (2010) Indiana Univ.

Mapping Dysregulation of Biological Networks in Disease

Disease Profiling to Identify Subtypes (+ or - Rx Target)



ID Molecular Targets for Rx Action and Blockade of "By Pass-Rx Escape/Resistance" Pathways



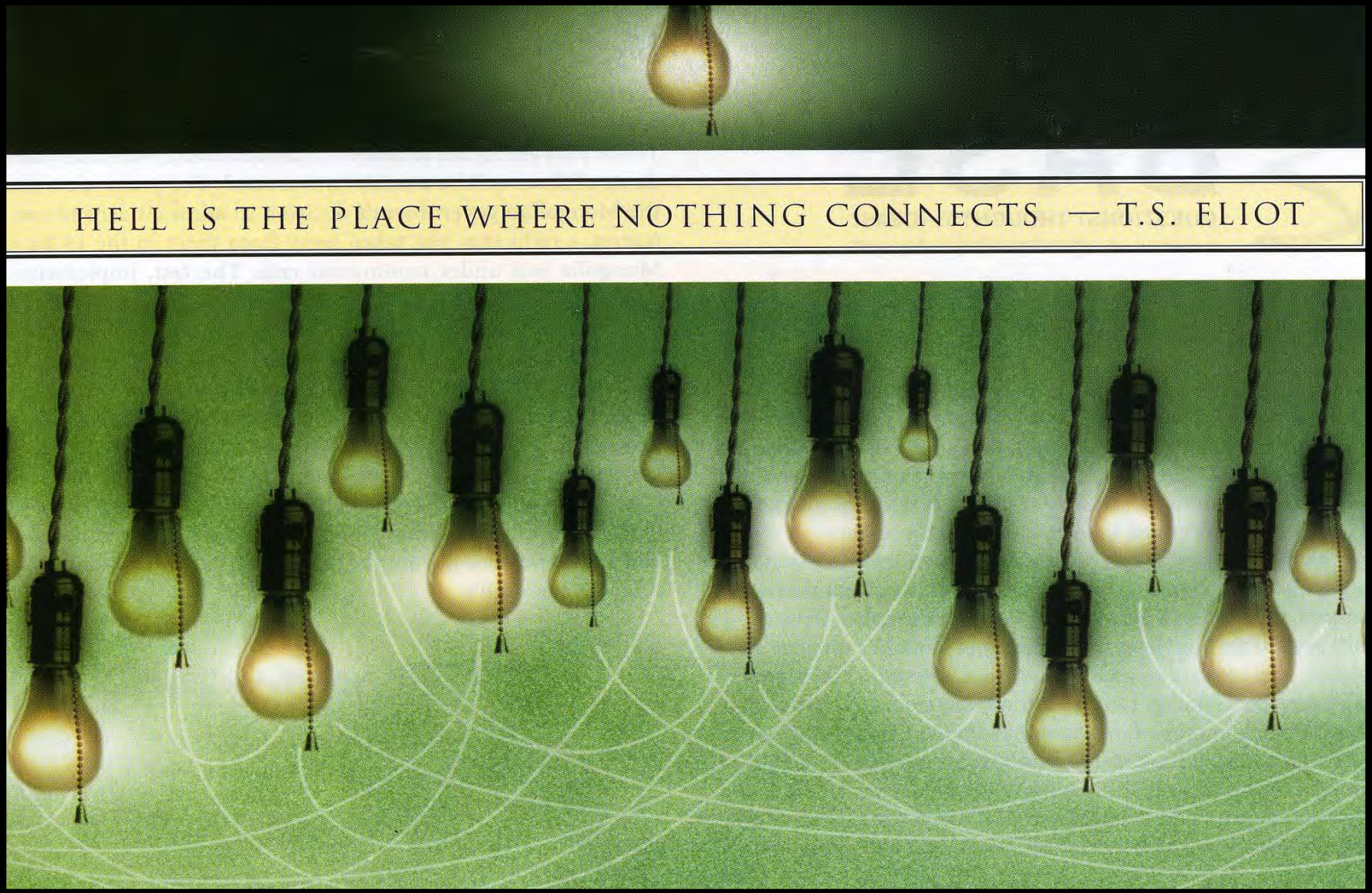
Initial Response (A/B) of BRAF-V600 Positive Metastatic Miliary Melanoma After 15 Weeks Therapy with Vemurafenib (Zelboraf® - Roche) Followed by Rapid Recurrence of Rx-Resistant Lesions with MEKI C1215 Mutant Allele After 23 Weeks Therapy

From: N. Wagl
et al. (2011)
J. Clin. Oncol. 29,

Network Pharmacology

- elucidation of definitive network ‘chokepoints’ as optimum targets
  - subvert adaptive cellular options to use alternate compensatory “escape” pathways
- the design challenge for multi-target polypharmacology
  - multi-agent therapy (patient tolerance?)
  - controlled multi-target promiscuity in a single moiety (low feasibility?)
- a disturbing question
  - at what point does level of network dysregulation eclipse feasible Rx “homeostatic reset”?

Silos Subvert Solutions: Protecting Turf and Sustaining the Status Quo



Representation of Datasets and Abstractions

- controlled vocabularies and formal ontologies
- minimal information checklists and open source repositories
- algorithms and source code for analytical tools
- exchange formats and semantic interoperability
- cross-domain harmonization/integration/ migration/sharing
  - community-driven (eg. SMBL.org, BioSharing catalogue), industry-driven (eg. Pistoia Alliance), regulatory-driven (eg. CDISC, Sentinel), clinical (eg. HL7), HITECH (EHR/MU) reimbursement (CPT, ICD), legal (HIPAA, GINA)

The Only Valuable Data is Validated, Actionable Data

Elucidation of Information Flow and Dysregulation in Biological Networks as Foundation for Precision Diagnostics, Patient Profiling and Rational Therapy and/or Risk Mitigation

- application of automated and robotic assays and high throughput production suites and advanced machine learning tools for analysis/decision

- development of new mathematical, statistical and computing tools for analysis and modeling of non-linear phenomena in complex networks

- modeling and simulation of biological networks of escalating complexity

What Is?
The Evolution of Computation Capabilities for Natural Language Q&A in Large Datasets

Jeopardy 16 February 2011

- IBM's Watson
  - 2880 CPUs
  - natural language query processing
- prelude to Q&A systems for biomedicine beyond keyword IR searches

WELLPOINT | Health.Care.Value.™

NUANCE®

What Is?
The Evolution of Computation Capabilities for Natural Language Q&A in Large Datasets

- IBM's Watson
  - 2880 CPUs
  - natural

y

What is When?

R

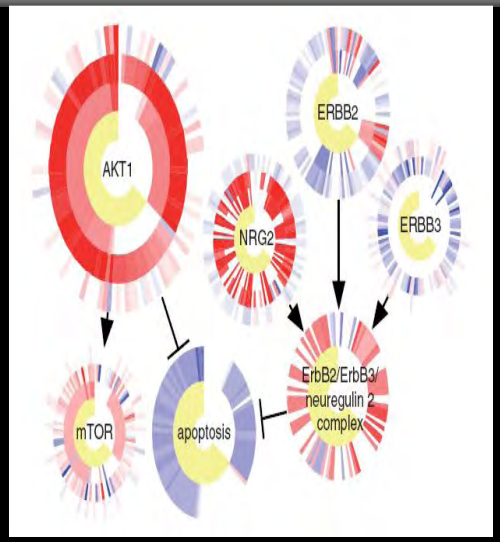
searches

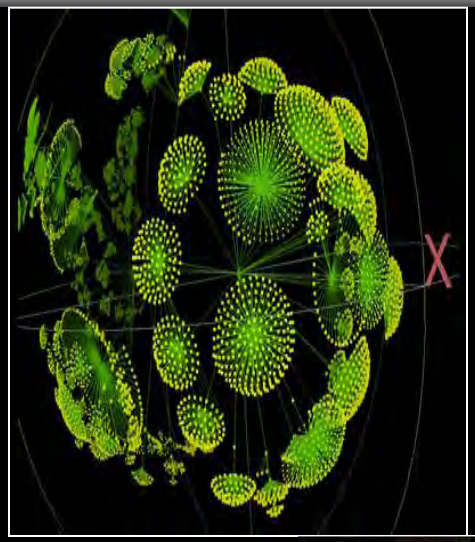
Jeopardy 16 February 2011

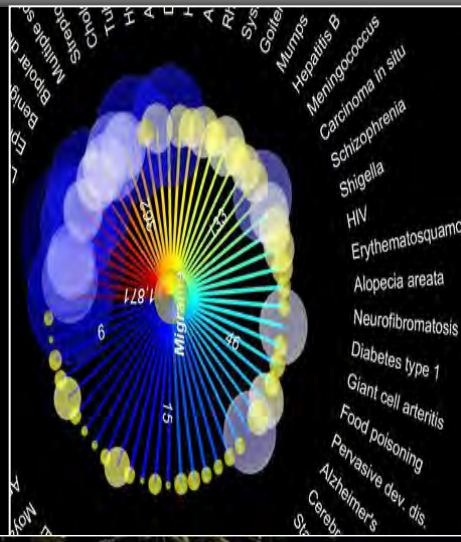
WELLPOINT | Health.Care.Value.™

NUANCE®

New Visualization Tools, Interactive Interfaces and Rapid Customization Formats







Cognitive Cartography: Dynamic and Static Division of Labor

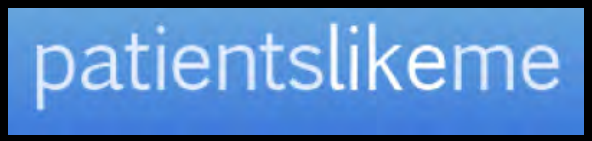
static

- conventional collaborations
- traditional social-professional preferences and hierarchies and related institutional and career constraints
- reluctance to share data
- minimum patient input

dynamic

- web-based collaborations
- fluid, unbounded populations of diverse participants with many unanticipated productive inputs
- capture of latent information/expertise via crowd sourcing
- open source data and new extended communities

Open Data Systems and Crowd Sourcing in Biomedical R&D



Public Availability of Published Research Data in High-Impact Journals

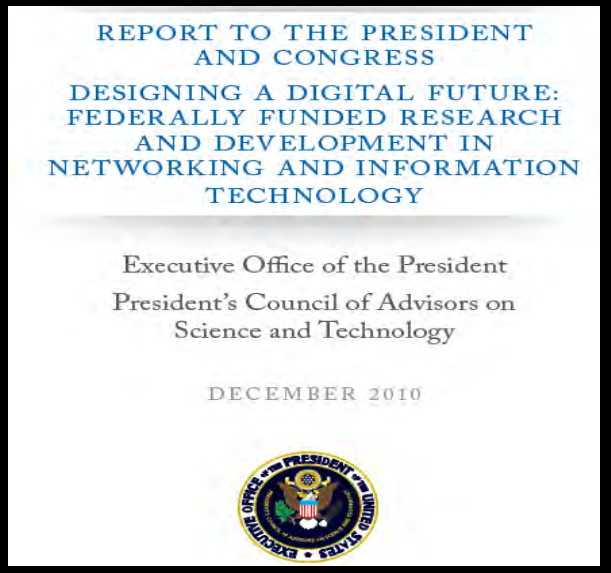
A. Alsheikh et al. (2011) PLoS ONE 61, e24357

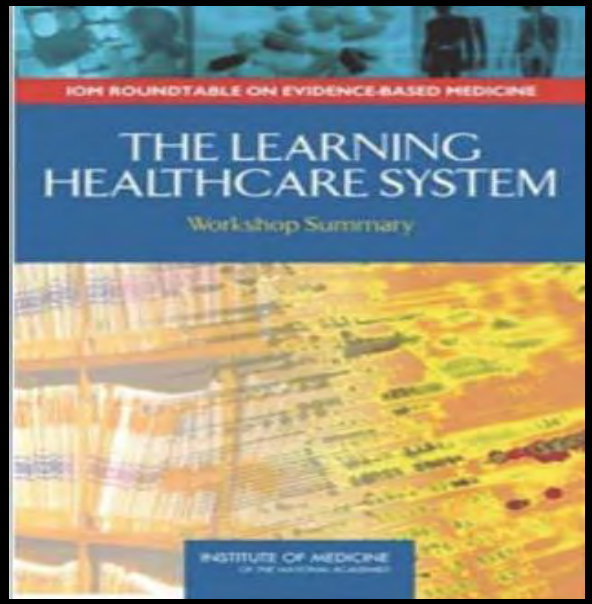
- review of 500 papers (2009) in 50 journals with highest impact factor
- variable editorial policies about data sharing and deposition
- only 47/500 (9%) deposited full primary raw data online
- 73% did not deposit microarray data
- 59% did not comply with data access instructions of the journal/funding agencies

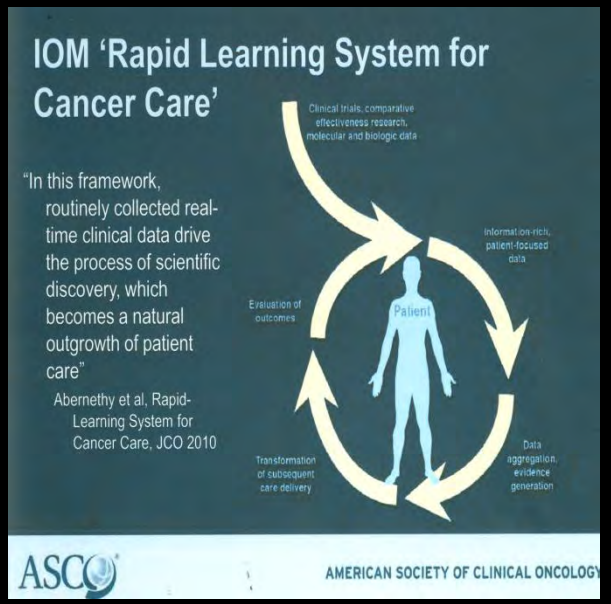
New Incentives and Resources for Researchers in Era of Data Intensive Science

- credit, attribution and citation for construction of annotated datasets used by others
- funding, standards and organization of repositories for 'supplementary materials'
  - obligate/open deposition of publically funded data
  - full disclosure of raw data and code
- funding agencies to support cost of comprehensive data management, movement and storage resources as core element of modern research capabilities and cyberinfrstructure
- metadata access control and certification requirements
- resources for content protection and preservation
- digital rights management

Now Comes the Hardest Part: Driving Molecular Medicine and IT-Centric Capabilities Into Routine Clinical Practice







JAMA (2011) 305, 767

National Cancer Institute Begins Revamp of Clinical Trials Cooperative Program

Bridget M. Kuehn

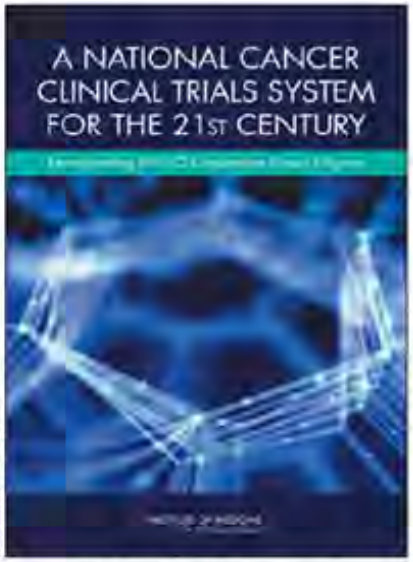
With tighter deadlines for the launch of new trials, new technology to increase transparency and streamline data management, and plans to consolidate the number of groups in its Clinical Trials Cooperative Group Program, the National Cancer Institute (NCI) has moved swiftly to implement some of the changes recommended by the Institute of Medicine (IOM) in April 2010. But questions remain about whether recommended funding increases and other changes will occur.

STATE OF CRISIS

The Clinical Trials Cooperative Group Program plays a key role in conducting trials unlikely to be undertaken by industry. In a report commissioned by the NCI, the IOM cautioned that the 55-year-old program was reaching "a state of crisis." The report (http://www.iom.edu/ncicancertrials) outlined how a cumbersome structure, poor reimbursement for those conducting the trials, and other problems were making it difficult for the program to adapt to changes in cancer research.

The program is organized into 10 groups, which work with more than Group, last spring the NCI began implementing some changes to increase the efficiency of the program.

Among the changes implemented in the past 8 months were the establishment of new scheduling targets that dramatically cut the suggested timeline for moving a trial from conception to approval. Previously, the median time to take a trial



National Academies Press

Recommendations from the Institute of Medicine are helping to reshape the national cancer clinical trials program.

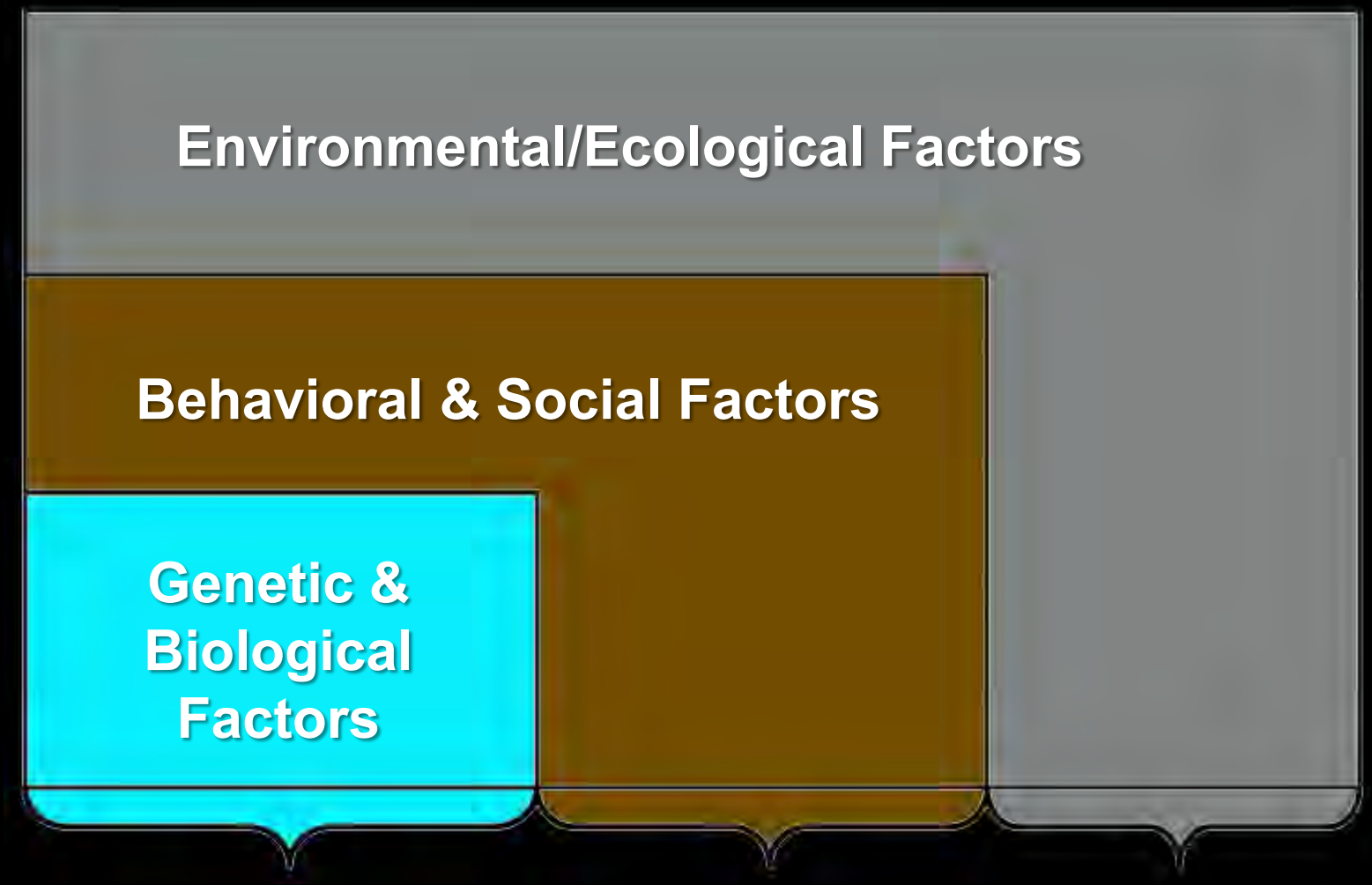
APPROVAL PROCESS CHANGED

To achieve this efficiency, the NCI has made several changes to the approval process, Doroshow explained. For example, when a trial proposal is submitted, contractors for the NCI will make any formatting changes necessary for the document to meet NCI requirements rather than sending back a list of suggested changes for the investigators to complete. Also, once a panel of experts has reviewed the proposal and identified any scientific issues that need to be addressed, a teleconference is arranged within a week to allow the investigators, statisticians, and NCI staff to either work out the issues or decide not to pursue the study. A similar teleconference is held to address issues identified by the NCI's national institutional review board (IRB). These changes have greatly reduced the amount of time investigators and NCI staff spend corresponding about a proposed trial. For example, the IRB teleconferences have helped cut the time to receive IRB approval from 120 to 40 days.

"We're just not allowing the process to be bogged down," Doroshow said.

Investigators can also now monitor the status of their proposal in much the same way delivery companies enable customers to track packages. The NCI

The Challenge of the Capture of Comprehensive Information Relevant to Disease Risk, Progression and Outcomes



The Design Challenge for Next Generation HIT Systems

- design of dynamic EHRs versus minimum value of digital duplication of current static paper formats
- most EHRs today are not designed to support secondary use of data to inform research/translational medicine
- lack of harmonized data standards in different disciplines/delivery systems as handicap to data meta-analytics outside of original capture institution
- urgent need for comprehensive clinical data integration formats
  - current and planned RCTs
  - observational data from primary care provider and patient self-reported data
  - SEER (surveillance, epidemiology and end results) data
  - m.health/sensor net data remote and health status monitoring
  - payer datasets

The Cancer Journey

Goldberg et al / Am J Prev Med 2011;40(5S2):S187–S197

Patients, Friends, Family on a Cancer Journey
Personas for Cancer.gov

Figure 3. For the National Cancer Institute's www.cancer.gov, the map of consumer personas considers both technology and health literacy, but also differing needs at various places along the cancer journey.

From: L. Goldberg et al. (2011) Am. J. Prev. Med.

Proactive Engagement of Patient Communities in Investigational Clinical Trials and Observational Outcomes Studies

- 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.
- Neutral Conveners: Sage Bionetworks and Genetic Alliance [nonprofits].

Partnering for Cures | A FasterCures Meeting

CYCORE CYber-infrastructure for COmparative Effectiveness REsearch

CW PHS CENTER FOR WIRELESS & POPULATION HEALTH SYSTEMS

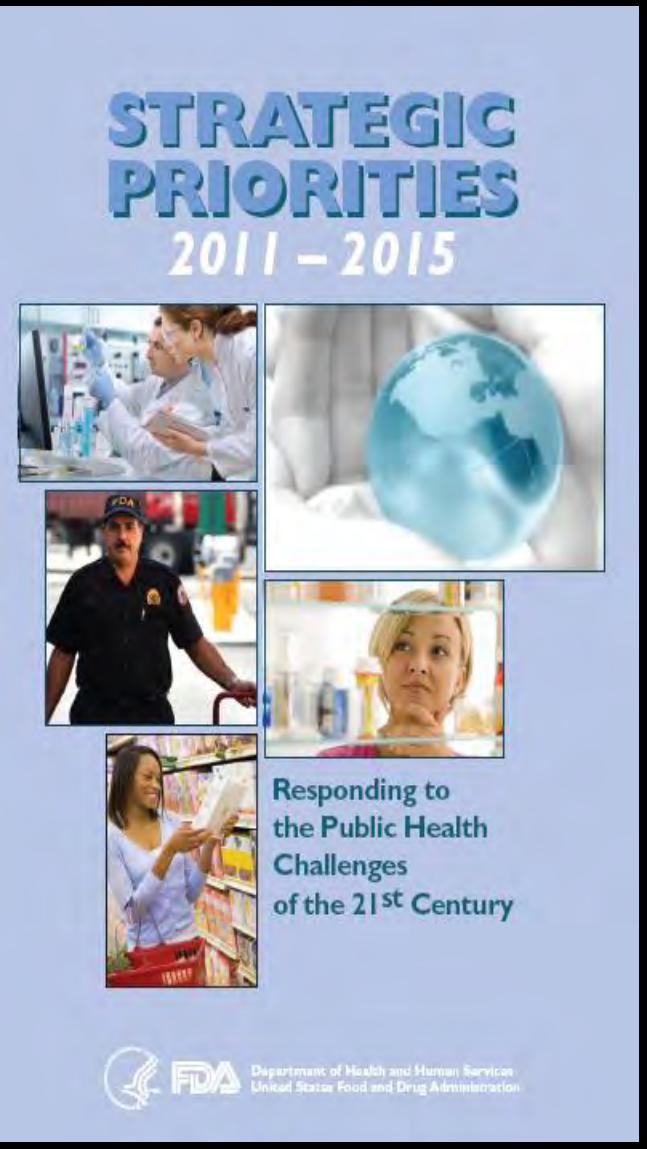
PURPOSE

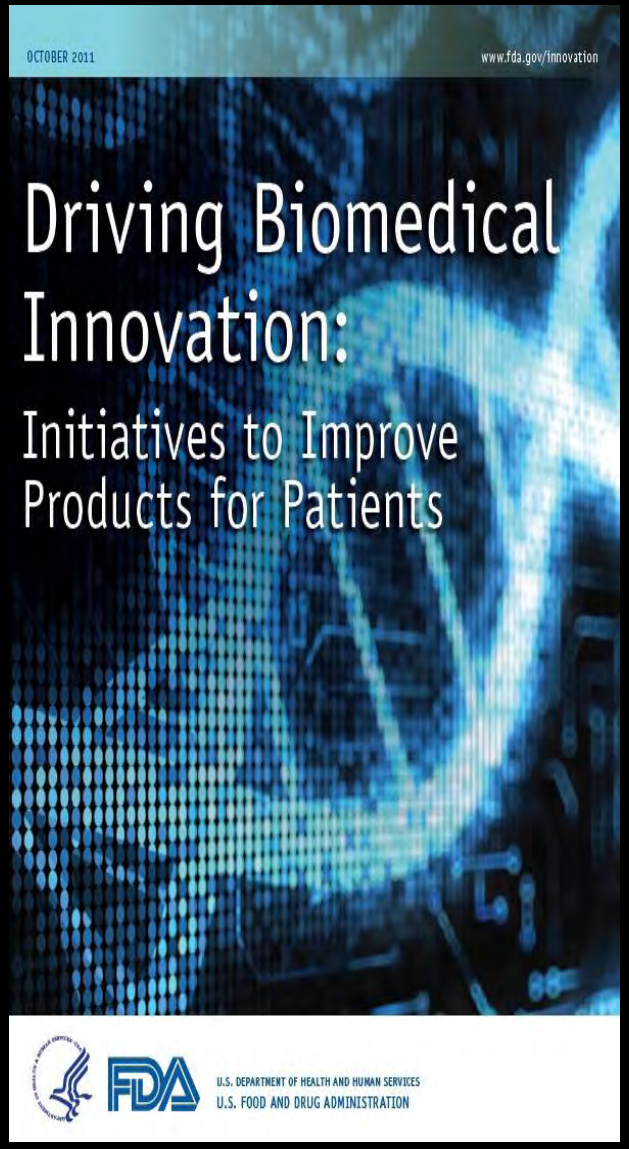
To improve **cancer-related comparative effectiveness research** by better capturing data on physiological, behavioral and psychological status from research participants at home and as the go about their daily lives.

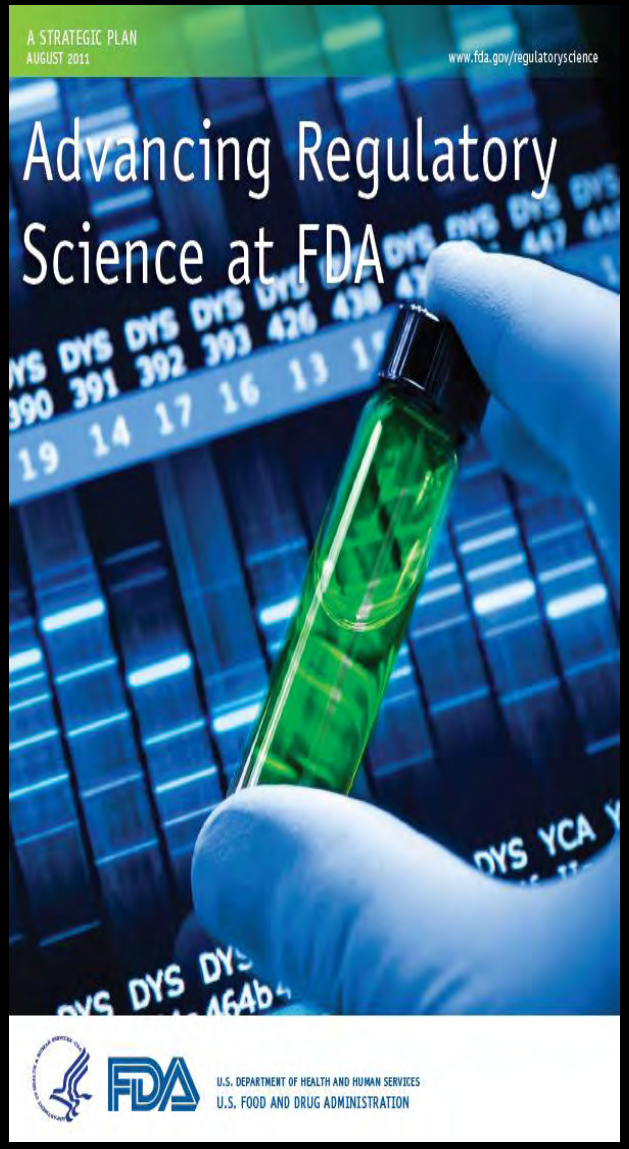
THE UNIVERSITY OF TEXAS MD Anderson Cancer Center Making Cancer History®

CYCORE

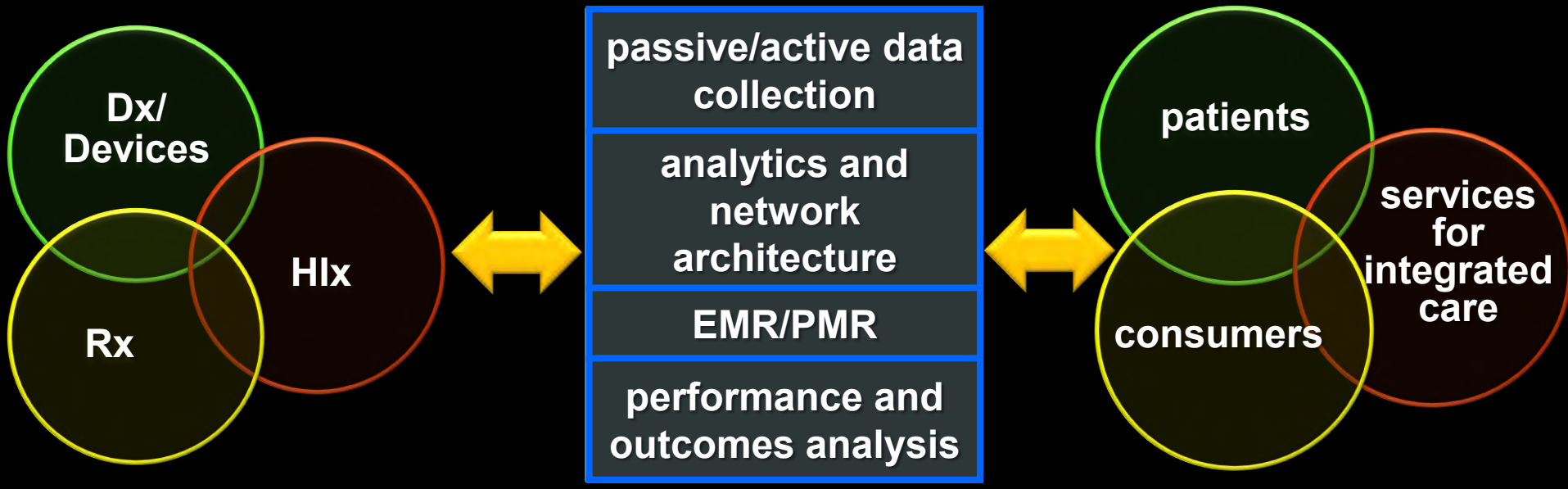
Regulatory Science







A New Healthcare Ecosystem Arising From Convergence of Technologies and Markets



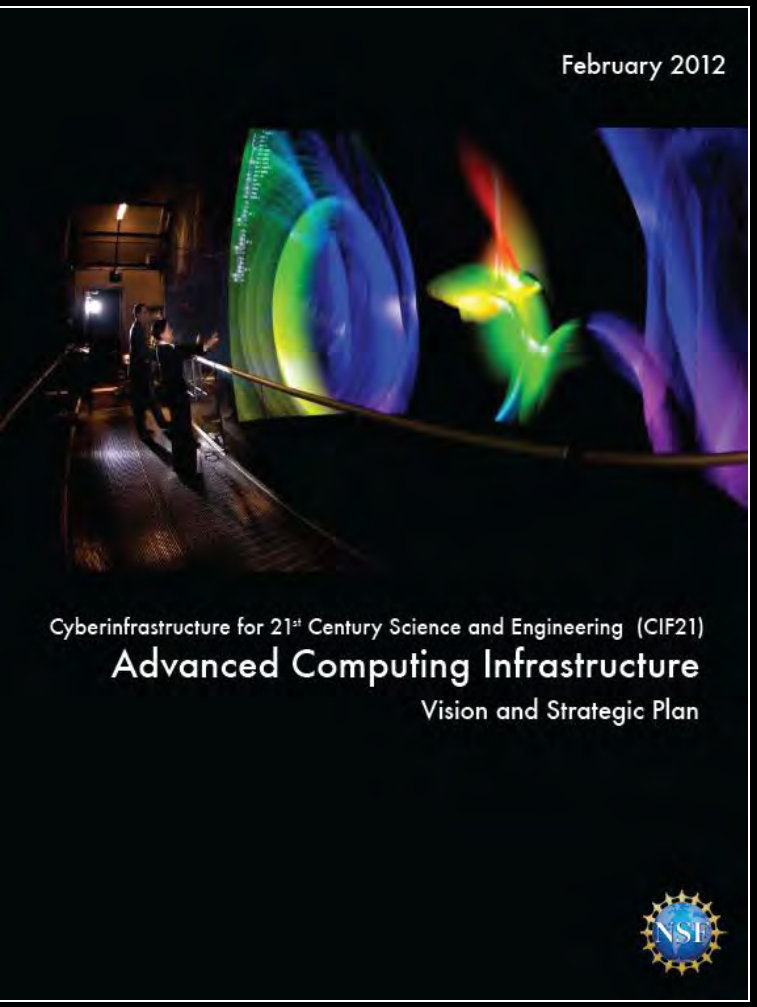
Integrated Technology Platforms

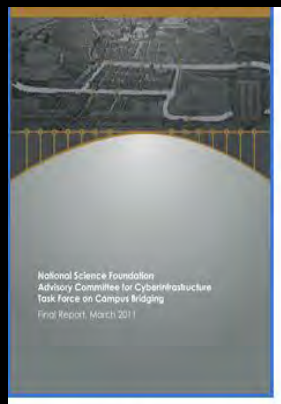
Data Mining and Integration Services

Increasingly Targeted Care and Efficient Use of Finite Resources

National Science Foundation
Office of Cyberinfrastructure (OCI)

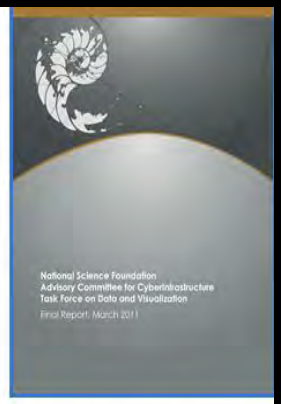
Planning for Big Data and Cyberinfrastructure in e.Science





Campus Bridging

Cyberlearning & Workforce Development



Data & Visualization

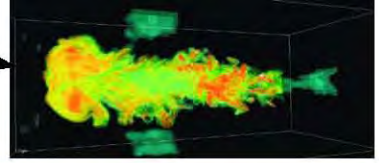
Grand Challenges

HPC

Software for Science & Engineering

http://www.nsf.gov/od/oci/taskforces/index.jsp

Cyberinfrastructure for High Performance Computing (HPC) and Cloud Computing (CC) for Large Scale Biomedical Datasets



Science Portals: collaboration and problem

Web Services and Application building services

Grid Services: secure and uniform access and management for distributed resources

Supercomputing and Large-Scale Storage

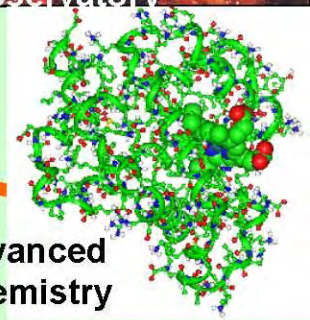
experiment

theory

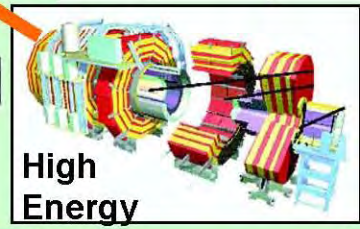
simulation

ESnet: High Speed Networking

Computing and Storage of Scientific Groups

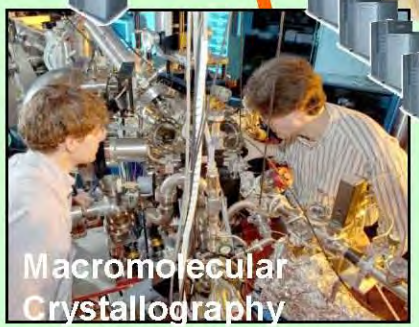


Advanced Chemistry



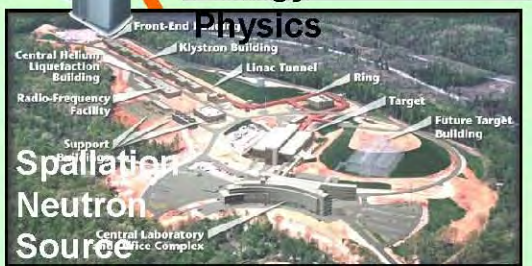
High Energy Physics

Advance d Engine Design



Macromolecular Crystallography

Advanced Photon Source

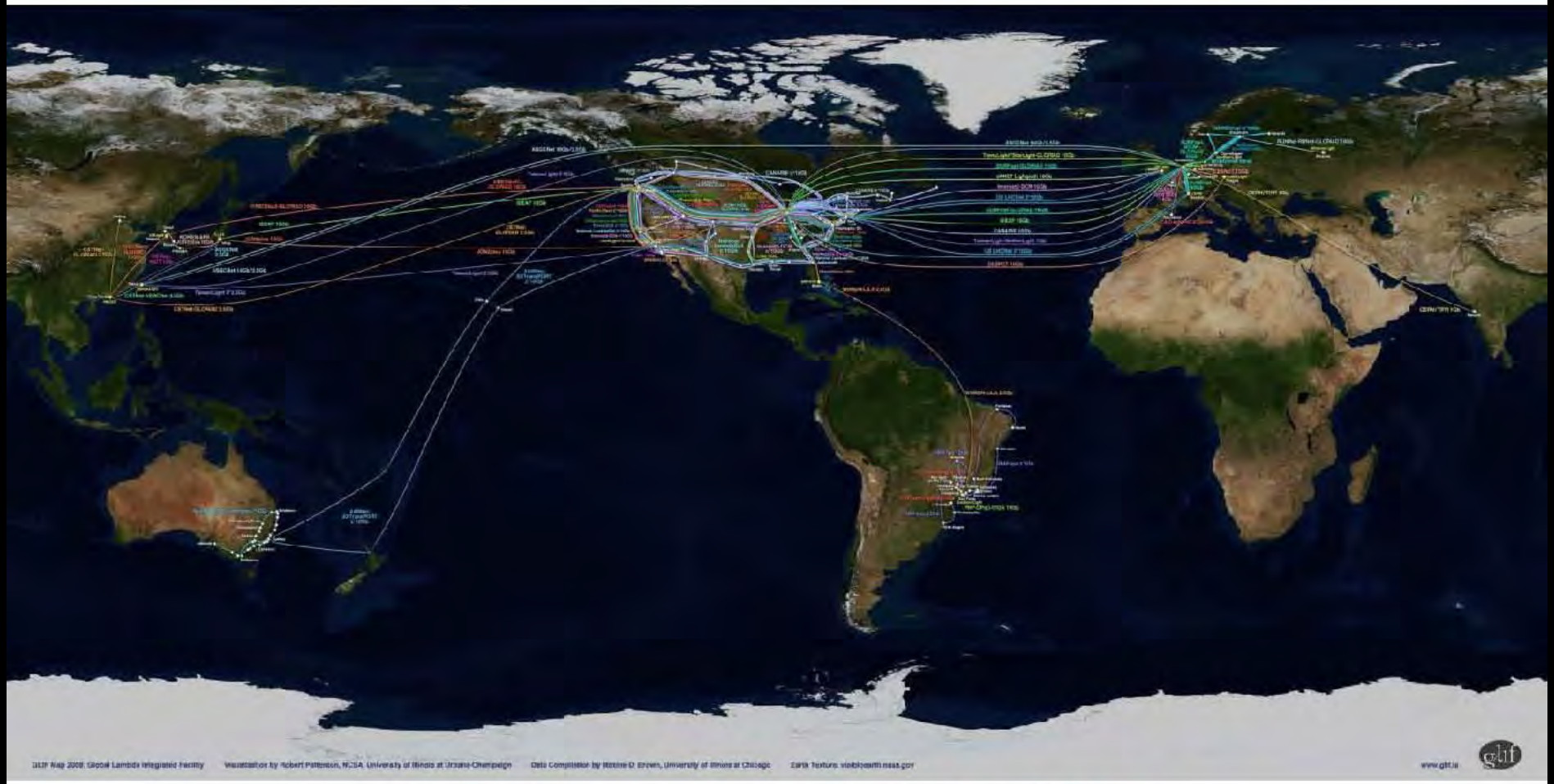


Spallation Neutron Source

Infrastructure Implications for Institutions in an Era of Data-Intensive Biomedical Research

- 'latency' is the enemy in collaboration and data transfer
- isolated "islands" (labs/centers) connected to Internet at 10 megabits/sec
- last mile/last 100 feet problem
- need for routine connectivity to optical networks with 10,000 megabit/second transfer
- "10G is the new 1G":

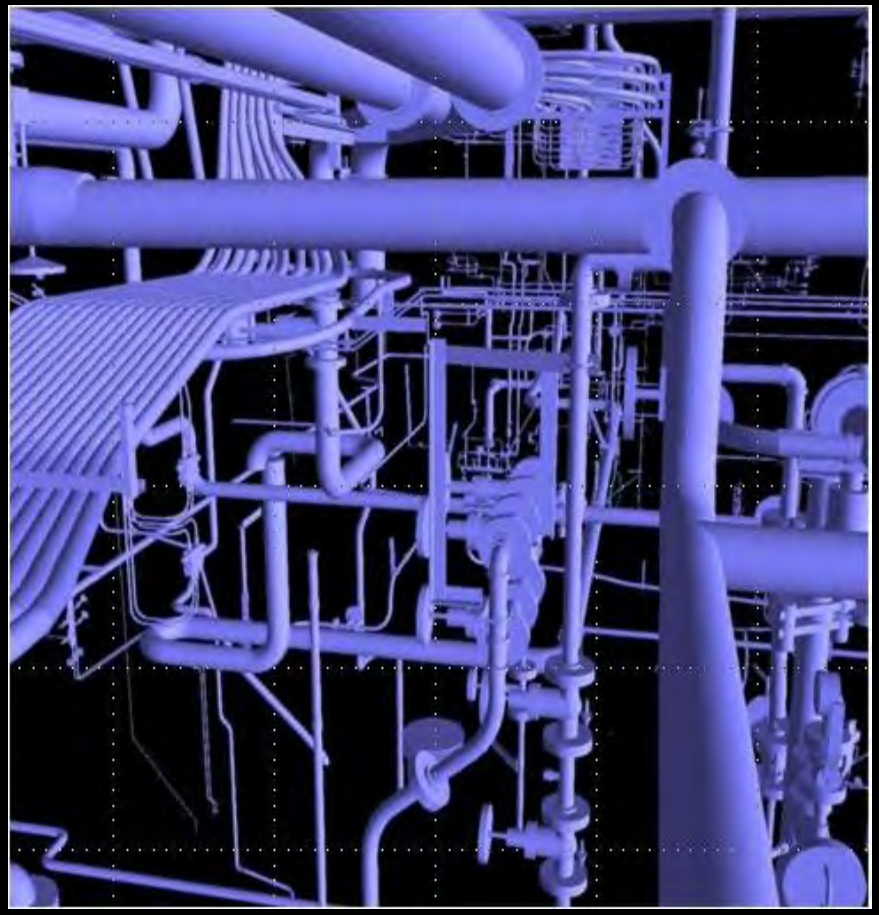
Global Innovation Centers are Being Connected with 10,000 Megabits/sec Clear Channel Lightpaths



100 Gbps Commercially Available
Research on 1 Tbps; 50 Tbps By 2020

Source: Maxine Brown, UIC and Robert Patterson, NCSA

Not All Pipes Are Created Equal



Key Principles in an Era of Data-Intensive Computing

- a pending neo-Malthusian digital divide
  - growing imbalance between end user population and their ability to embrace data scale and complexity
  - institutions unable to access and analyze large data sets will suffer ‘cognitive starvation’ and relegation to competitive irrelevance
- harnessing massive data and underlying computational competencies will demand new ecosystems for productive science and technology
- understanding the structure of information and its productive application/customization will emerge as a critical institutional competency

Commercial Cloud Computing Services

Amazon Elastic Compute Cloud (Amazon EC2) - *Beta*

Cloud Computing (CC) for Biomedical Data

- no single business model for CC adoption
- on-demand access to large scale, economically competitive, virtual supercomputing without overheads/expertise needed for large on-site clusters
- flexibility
  - low-and high-intensity users and efficient management of major fluctuations in internal demand(s)
- concerns over security, reliability, IP and regulatory compliance
  - varied attitudes of companies, depending on subject domains, content (public vs open) and regulation (HIPPA, EHRs)
- standards for public domain datasets and CC analytics

BGI Cloud on the Horizon

- "Amazon is slow"
  Evan Xiang, BGI Shenzhen
  Bio-IT World August 2011 p.8

- launch of new platforms
  - Hecate: de novo assembly
  - Gaea: SOAP, BWA, Samtools, Dindel, reals-FS algorithms

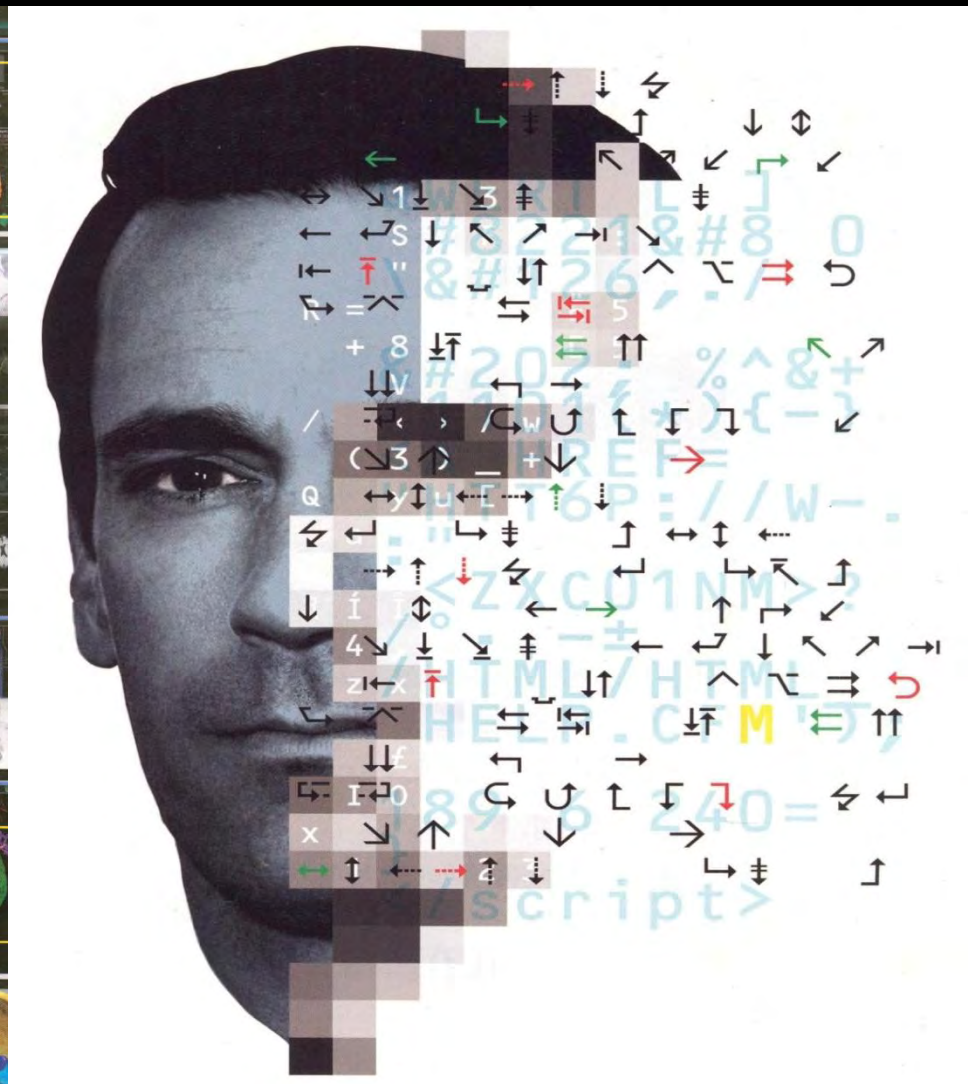
- November 2011 launch of new journal with BioMed Central
  - 'big data' studies
  - host citable public datasets on BGI cloud
  - each with permanent digital object identifiers

21st Century Knowledge Networks
versus
20th Century Organizations

From Silos to Systems
Changing Minds and Changing Behaviors

- technology is only the enabler
- emergence of new organizational structures, alliances and business models
- engage and educate multiple constituencies with long entrenched behaviors
- healthcare space will be increasingly decentralized for data generation but increasingly centralized for data analytics/decision support
- from episodic patient encounters to increasingly real-time monitoring and new e.interactions
- new business opportunities in customized health services and health broker/concierge services

Technology Acceleration and Convergence: The Escalating Challenge for Professional Competency, Decision-Support and Future Education Curricula



A Framework for Action

Adapting to the Scale and Logistical Complexity of Translational Medicine

- single investigator awards and incremental progress (at best) and excessive duplication → high risk, high reward projects with potential for radical, disruptive innovation
- single discipline focus and career rewards → obligate assembly of diverse expertise for multi-dimensional engagement
- funding agencies ill-prepared to review inter- and cross-disciplinary research → new study sections with broader expertise, including industrial input
- 'data tombs' of siloed datasets with minimal standardization, diverse ontologies and poor inter-operability → large scale, standardized, inter-operable open-source databases with professional annotation, analytics and curation

Coordination of the Complex Interactions Required to Build a Productive Translational Medical Research Capacity

Government

- promulgation of standards and centralized orchestration of resources (national/international) and enforce obligate data sharing
  - biospecimens and biorepositories
  - 'omics' analytics reference standards
  - informatics platforms (BIX, HIX)
  - recruitment of relevant case:control patient cohorts
- proactive design of regulatory frameworks to address new technologies
  - complex multiplex 'omics assays
  - new clinical design protocols (I-SPY, BATTLE)
  - m.Health and remote health status monitoring
  - review process for MDx/Rx combinations
  - new CER tools/metrics

Forging the Complex Interactions Required to Build a Productive Translational Medical Research Capacity

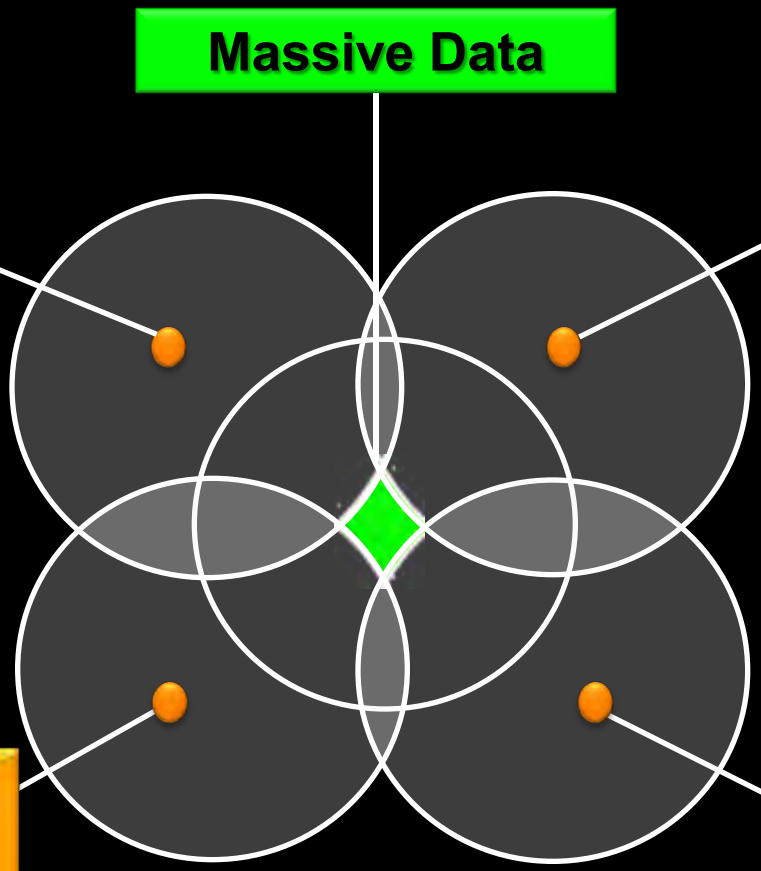
Industry

- greater recognition of value and participation in pre-competitive, open-source 3P networks/consortia
  - drive adoption of analytical and data standards
  - elucidation of network dysregulation in major diseases
- more proactive role in shaping new trans-disciplinary education/training/employment opportunities
  - translational medicine
  - large scale dbase analytics
  - new analytics/models for non-linear dynamics in complex systems
  - health economics outcomes/systems modeling

Adapting to Data-Intensive, Data-Enabled Biomedicine

Biomedical R&D and Clinical Medicine:
An Unavoidable and Painful (But Essential) Transition

Managing Massive Data:
Disruptive Changes and New Products, Services and Partnership Models



- silos subvert solutions
- multi-disciplinary scale
- pre-competitive consortia
- 3P alliances

- data standards
- federated dbases
- open-source networks
- novel mining analytics
- security

- cyberinfrastructure investment
- centralized vs. decentralized IT capabilities
- training of biomedical informatics professionals

- new organizational structures
- new funding (public) and investment priorities (private) to reflect scale and drive standards

Disruptive Change and the Complex Interactions Required to Improve Translational Biomedical Capabilities

courage

- to declare that radical change(s) is needed versus safe propagation of the status quo

resilience

- combating denial and deflection by entrenched constituencies

competitiveness and new participants

- disruptive change arises at the margins or at convergence points between previously separate sectors
- voice of patients, payers and new industrial participants as increasingly influential drivers of e.Science and HIT

accountability and responsibility

- improved ROI from public and private funding
- urgent societal and economic imperatives

Slides available @ http://casi.asu.edu/