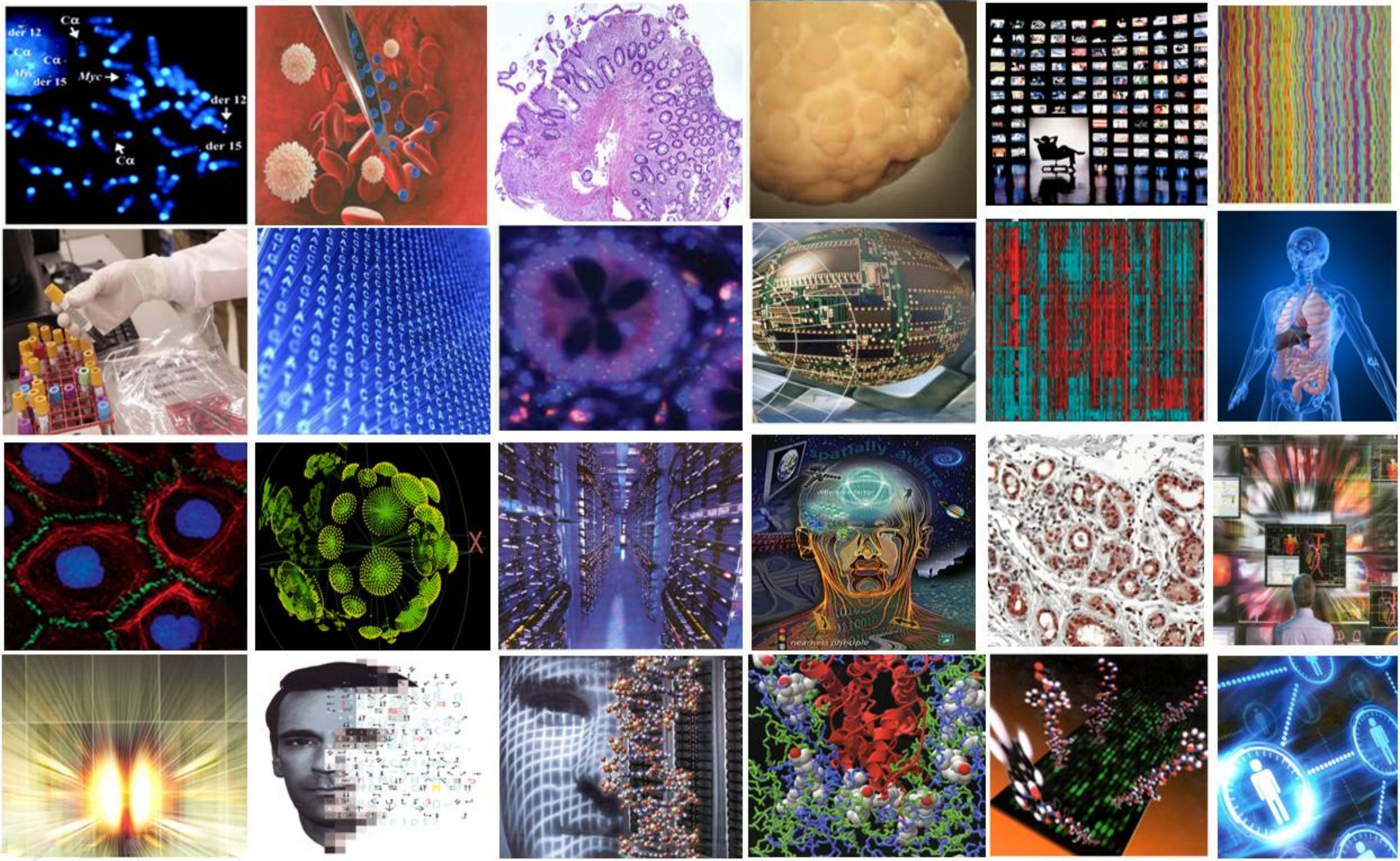
"It from Bits": Managing Massive Data as a Critical Challenge For Biomedical R&D and Healthcare Delivery

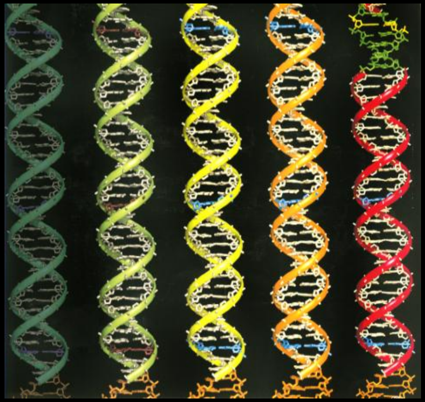
Dr. George Poste
Chief Scientist, Complex Adaptive Systems Initiative
and Del E. Webb Chair in Health Innovation
Arizona State University
george.poste@asu.edu
www.casi.asu.edu

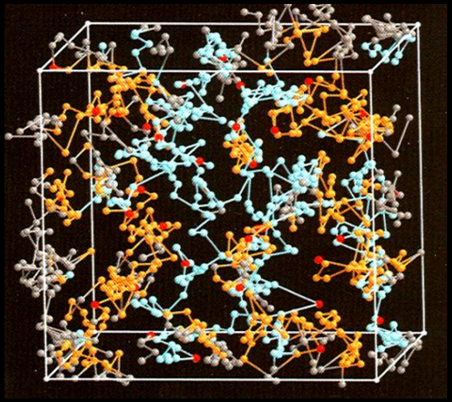
Keynote Presentation:
The Burrill Personalized Medicine Meeting
Burlingame, CA • Oct. 3-4, 2011

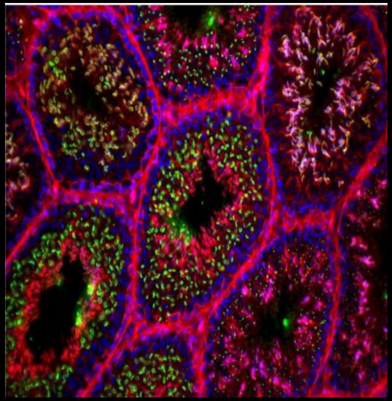
Slides available @ http://casi.asu.edu/



"It from Bits"





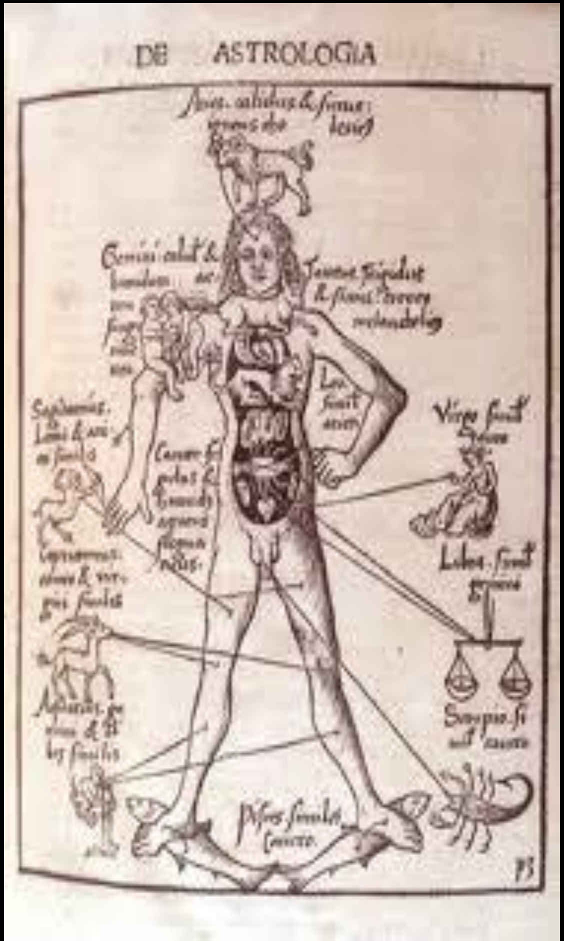


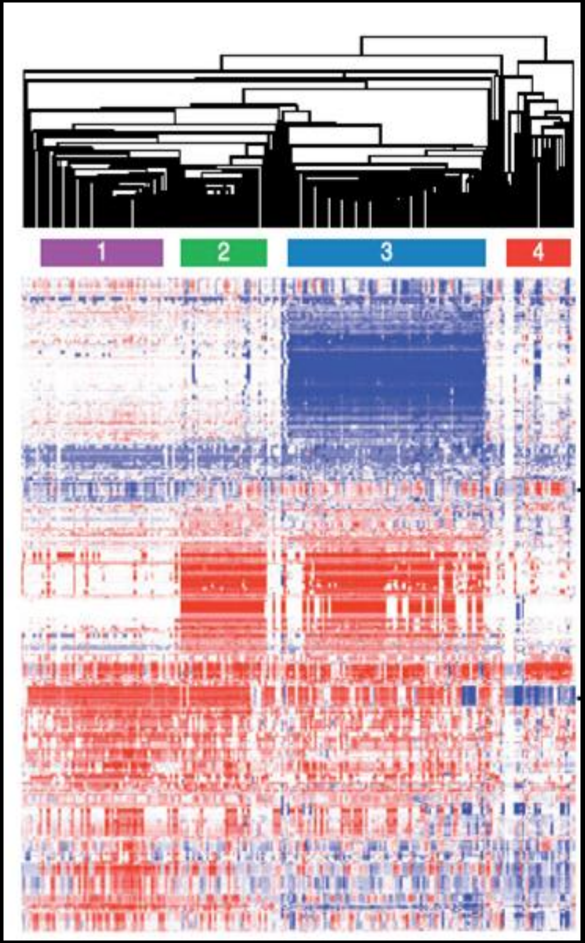
- understanding how encoded genome information creates complex multiscale biological systems ("It")

and

- defining health and disease in terms of patterns of information flow in biological information networks ("bits")

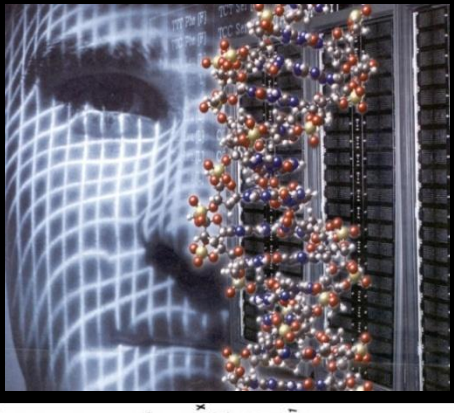
Medical Progress:
From Superstitions to Symptoms to Signatures



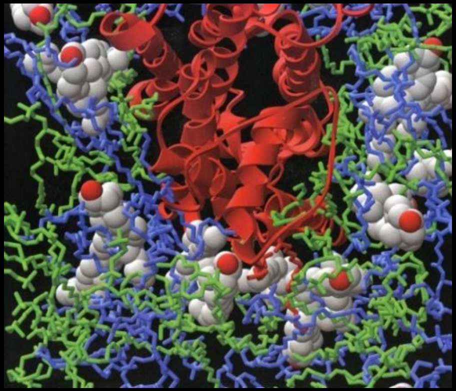


Mapping The Molecular Signatures of Disease:
The Intellectual Foundation of Rational Diagnosis and Treatment Selection

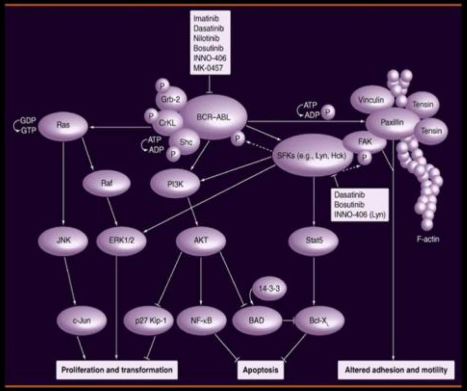
Genomics



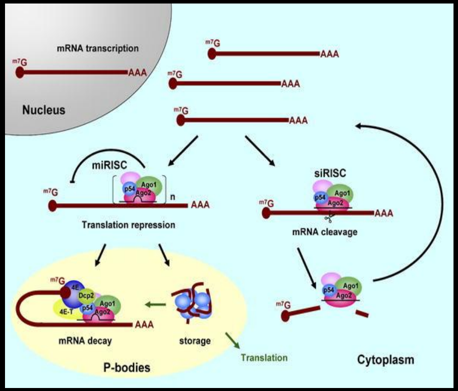
Proteomics

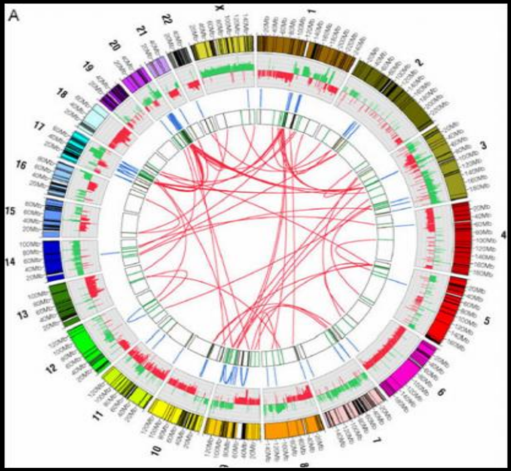


Molecular Pathways and Networks



Network Regulatory Mechanisms





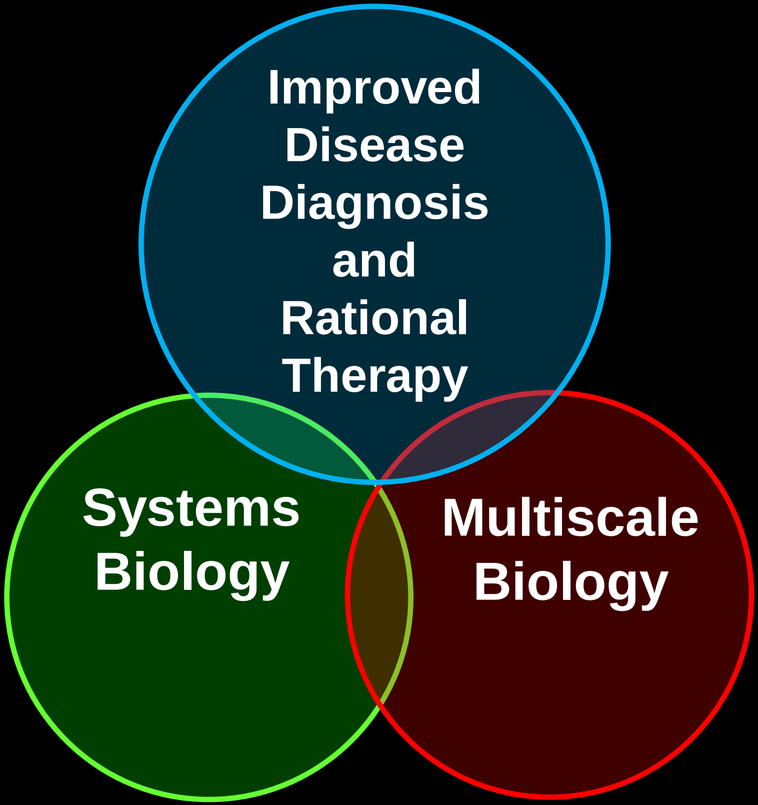
ID of Causal Relationships Between Network Perturbations and Disease

Patient-Specific Signatures of Disease or Predisposition to Disease

The Intellectual Horizons and Aspirations of Modern Biomedicine

Defining The Molecular Taxonomy of Diseases as Altered Patterns of Dysregulated Biological Networks

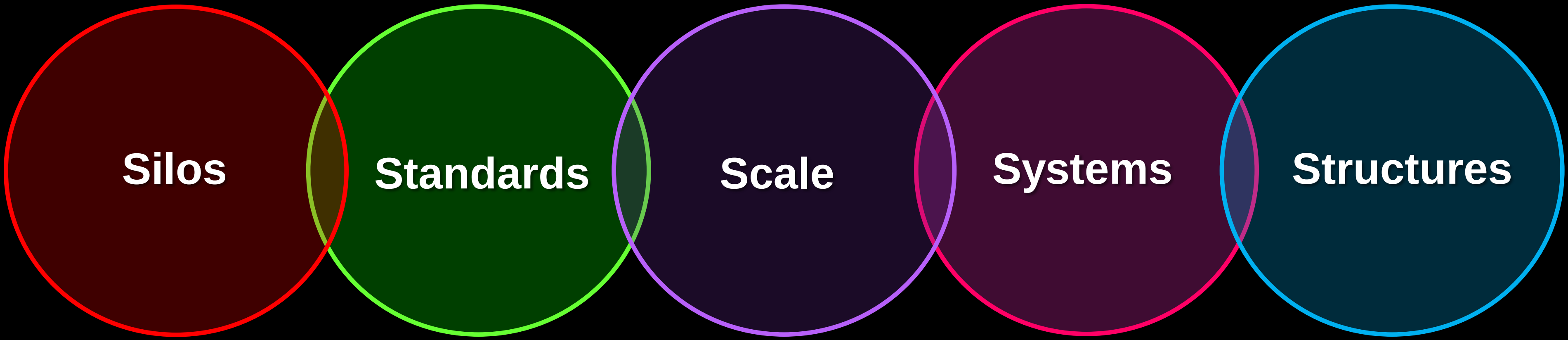
- molecular mechanisms of disease
- molecular diversity of disease arising in same organ/cell
  - disease subtyping
- germ line genome/epigenome
  - pharmacogenetic risk(s)
  - disease predisposition risk(s)



- precision diagnostics based on alterations in molecular targets/ pathways/networks
- rational (targeted) Rx selection based on underlying molecular pathology
- risk mitigation

Personalized Medicine

Challenges in Making Personalized Medicine a Reality: Key Themes



Silos

Standards

Scale

Systems

Structures

The Evolution of Biomedical R&D and Clinical Medicine as Data-and Computation-Intensive Disciplines

Advanced Computing and Knowledge Management (ACKM):

Building Critical Competencies to Sustain Progress in Biomedical R&D and Healthcare Delivery

Profiling Platforms for Mapping Molecular Networks: The Accelerating ‘Data Deluge’

Low Cost Exome- and/or Whole Genome Sequencing

life technologies™ ion torrent

illumina

454 SEQUENCING Roche

IBM

Oxford NANOPORE Technologies

Electronic Bio Sciences

华大基因 BGI

Complete genomics

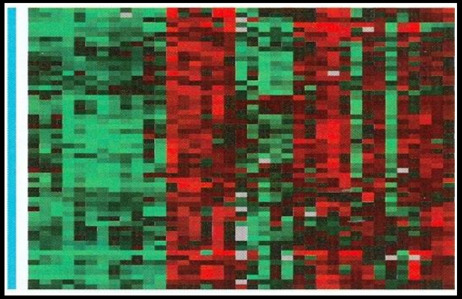
BioNanomatrix

PACIFIC BIOSCIENCES™

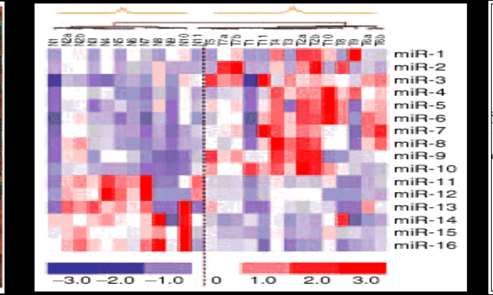
GE imagination at work

Helicos BioSciences Corporation

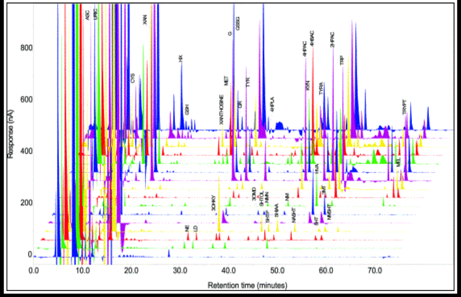
Transcriptomics



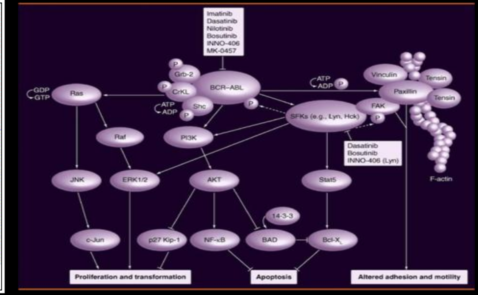
miRNAs



Proteomics



Protein Interaction Networks (PIN)



Mapping Biological Network Dynamics in Health and Disease

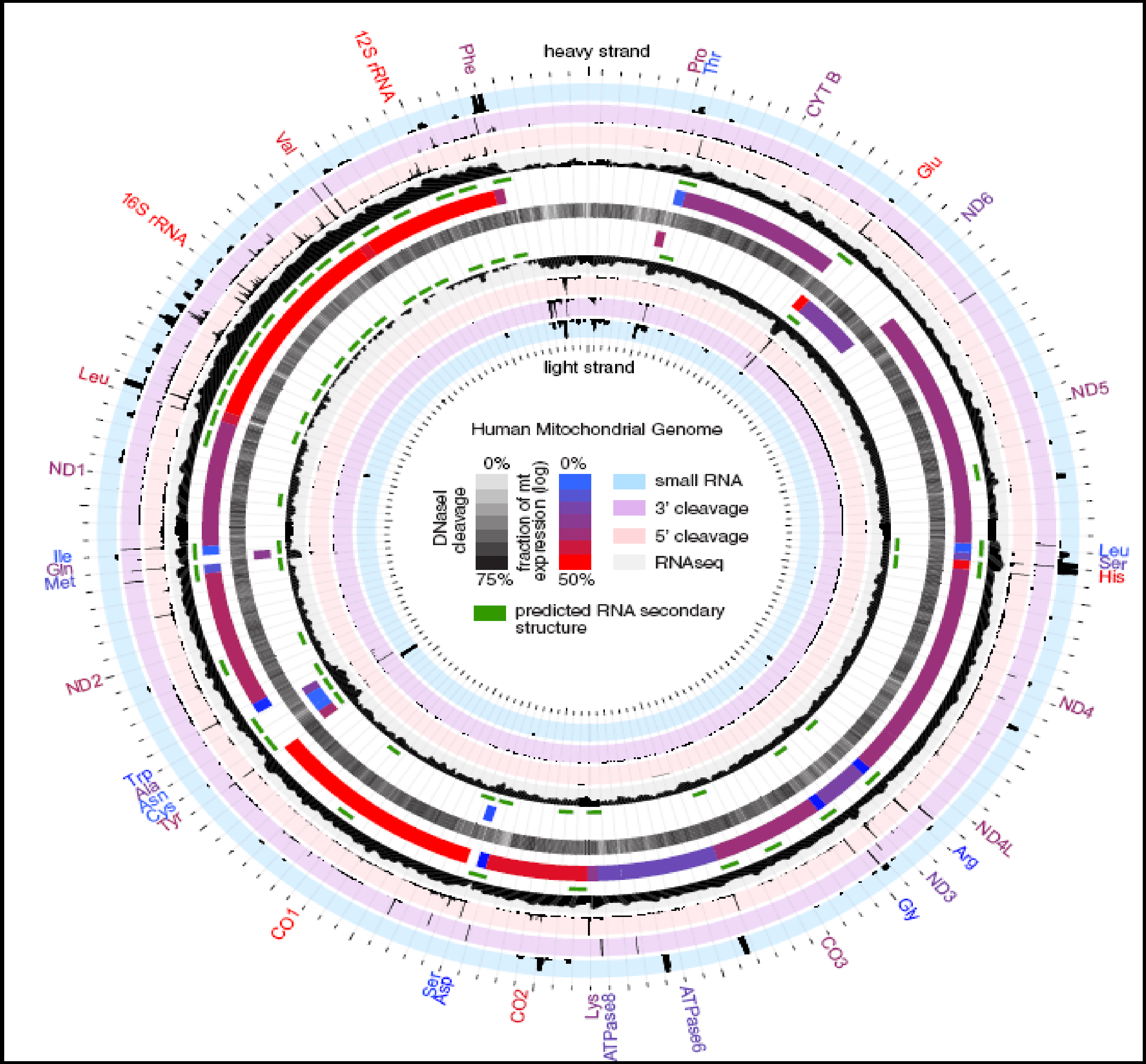
- daunting complexity of defining signals and signatures across massive combinatorial space
  - 230 different cell types + body fluids
  - pre-and post-translational gene regulation
  - SNPs, copy number variants, mutations, rearrangements
  - at least 200 PTMs and multiple PTMs in same biological pathway
  - protein expression, abundance and interactomes
  - localization, trafficking, turnover
  - dynamic range (from attomole to millimole)
  - physiological homeostasis
  - dysregulation and disease pathogenesis

Rapid Growth of Human Genome Sequencing Data

Individual Diversity

- evaluation of all combinations of two SNPs for 1 million SNPs represents nearly 500 billion possibilities
- dbSNP contains over $20x10^6$ validated SNPs
- Human Gene Mutation Database contains over 76,000 mutations from 2,900 genes
- COSMIC (The Catalogue of Somatic Mutations in Cancer) over 25,000 unique mutations
- PharmGKB dbase lists over 40 pharmacogenes and over 3,400 annotated drug-response variants

The Human Mitochondrial Transcriptome



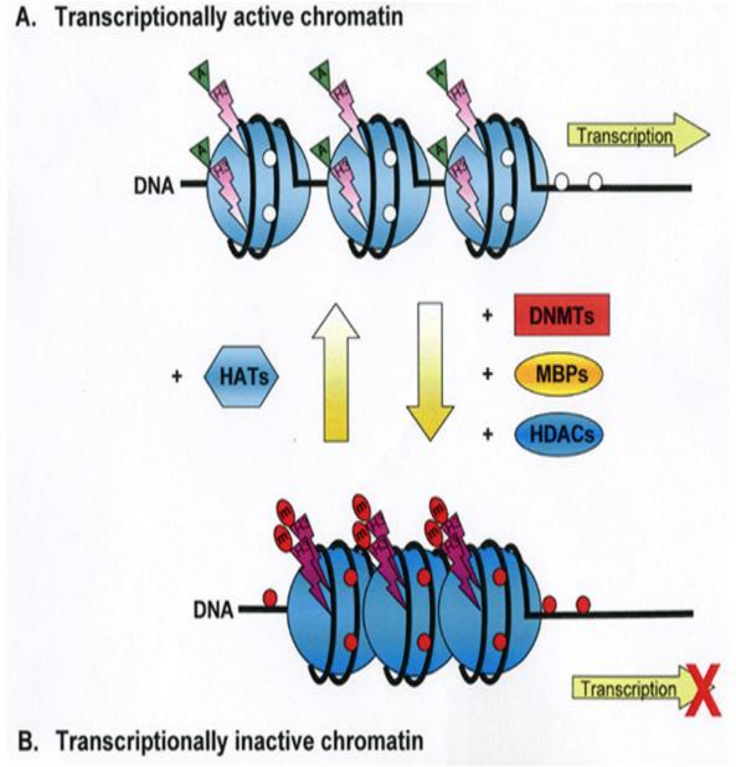
From: T. R. Mercer et al. (2011) CELL 146, 645

http://www.sciencedirect.com/science?_ob=MiamiImageURL&_cid=272196&_user=56861&_pii=S0092867411007677&_check=y&_coverDate=2011-08-19&view=c&_gw=y&wchp=dGLbVlS-zSkWb&md5=88c54aba632f066f1b3f6c67d6a7ffce/1-s2.0-S0092867411007677-main.pdf

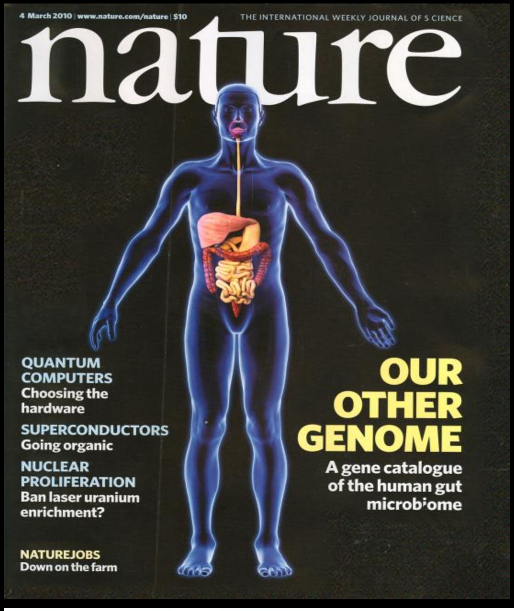
The Epigenome

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

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


A. Transcriptionally active chromatin
DNA
Transcription
+ HATs
+ DNMTs
+ MBPs
+ HDACs
DNA
Transcription
B. Transcriptionally inactive chromatin

We Are Not Alone:
Variation in the Human Microbiome as a Potential Factor in Health and Disease

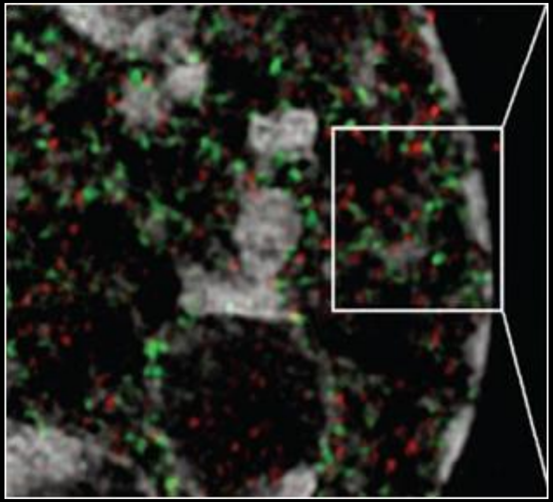

4 March 2010 | www.nature.com/nature | $10
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
nature
QUANTUM COMPUTERS Choosing the hardware
SUPERCONDUCTORS Going organic
NUCLEAR PROLIFERATION Ban laser uranium enrichment?
NATUREJOBS Down on the farm
OUR OTHER GENOME
A gene catalogue of the human gut microbiome

Supplement to PNAS
March 15, 2011 | vol. 108 | suppl. 1 | pp. 4513–4696
Proceedings of the National Academy of Sciences of the United States of America
www.pnas.org
Microbes and Health
Arthur M. Sackler Colloquium of the National Academy of Sciences

nature REVIEWS
april 2011 volume 9 no. 4
www.nature.com/reviews
FOCUS Mucosal microbiology
MICROBIOLOGY
SHIFTING THE BALANCE
Antibiotic effects on host–microbiota interactions
Host interactions with the skin microbiome
Scratching the surface

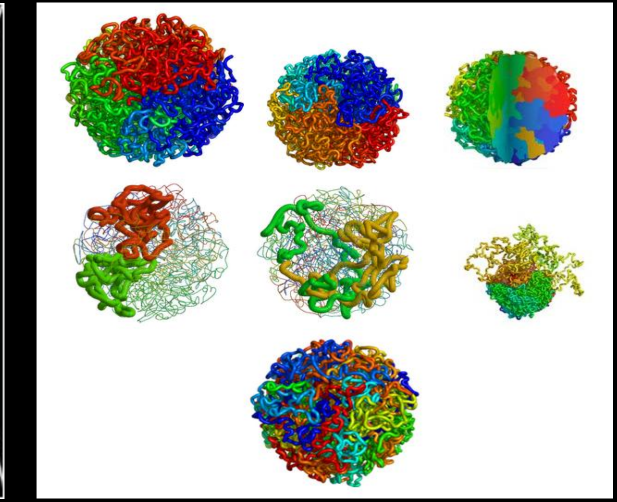
Data-Intensive Imaging and High Content Analysis of Cellular Architecture and Dynamics

Chromatin Loop Domains



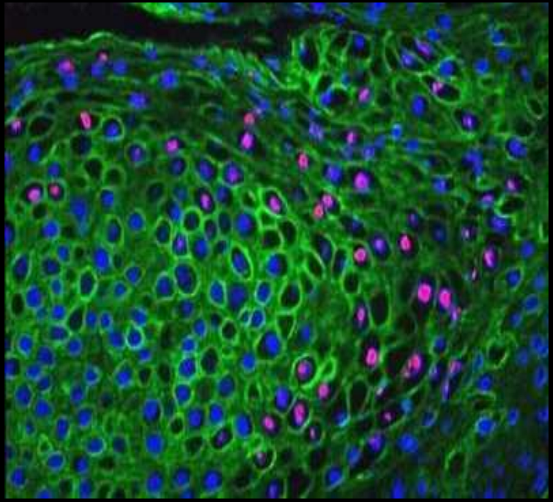
Nature (2011) 470, 292

Modeling of Nucleosome Folding

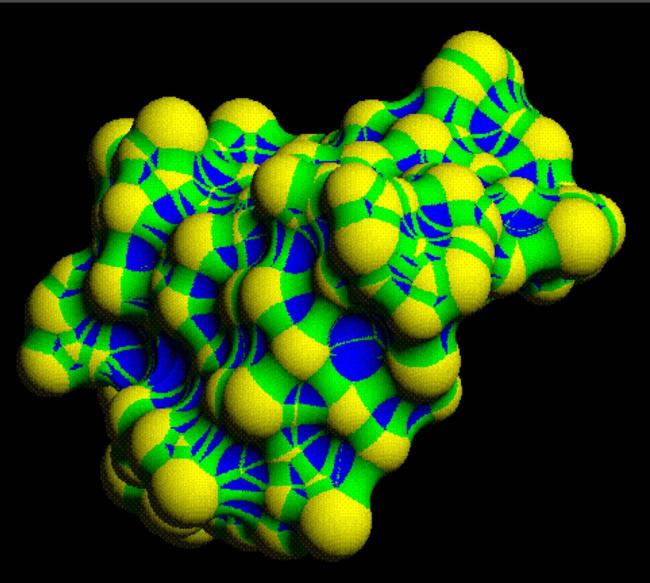


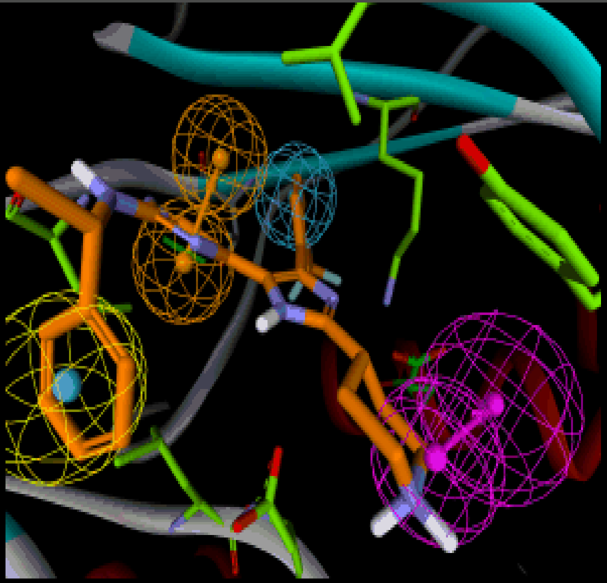
Science (2009) 326, 289

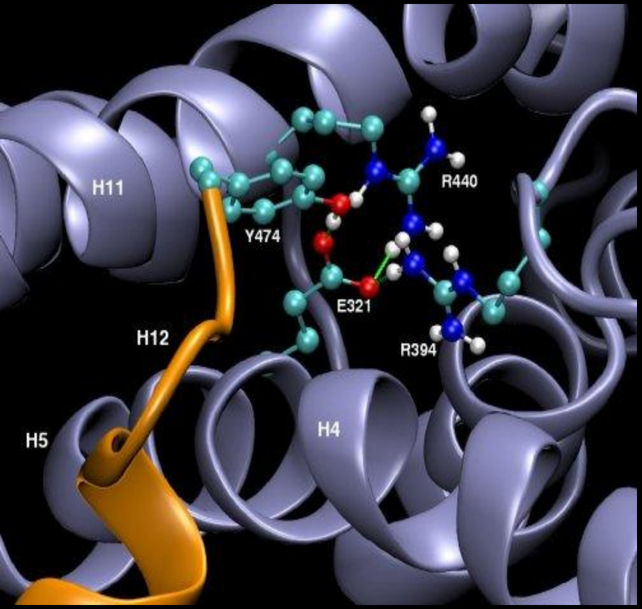
Digital Pathology

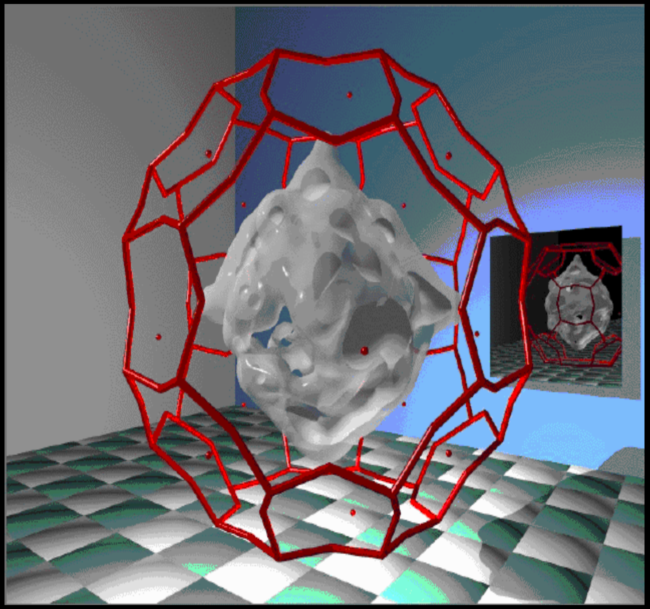


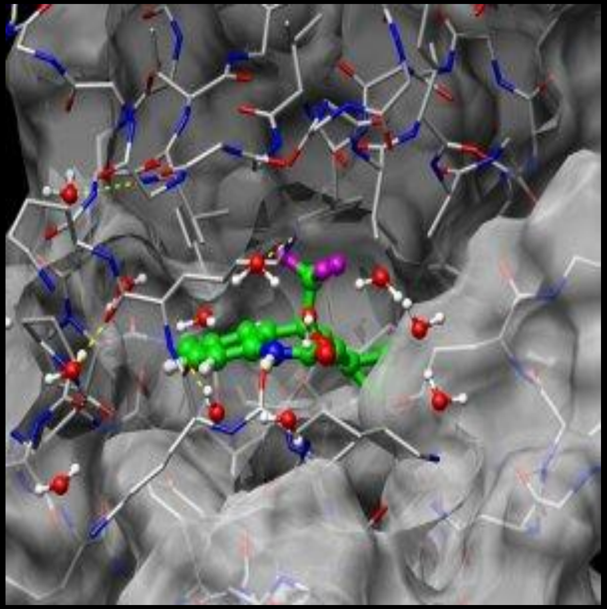
Computational Chemistry and Molecular Modeling



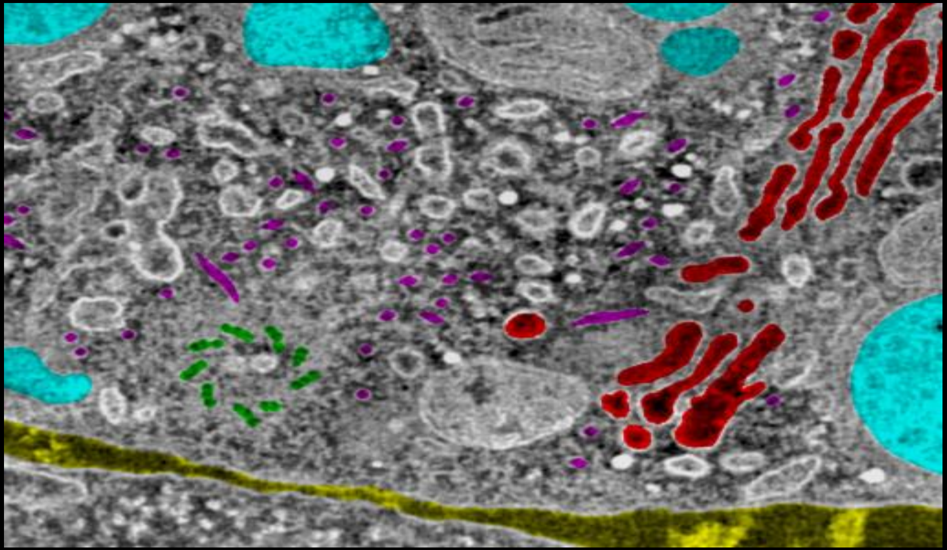


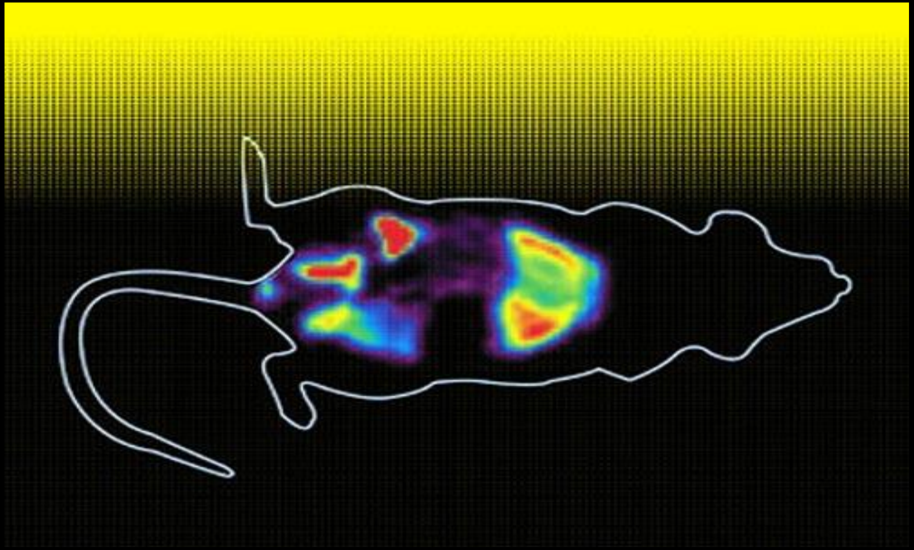


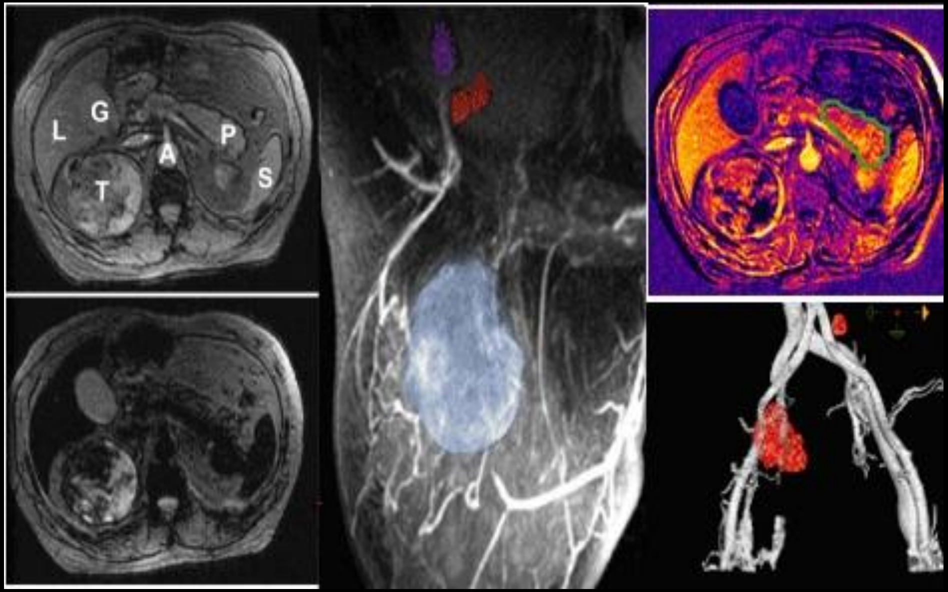


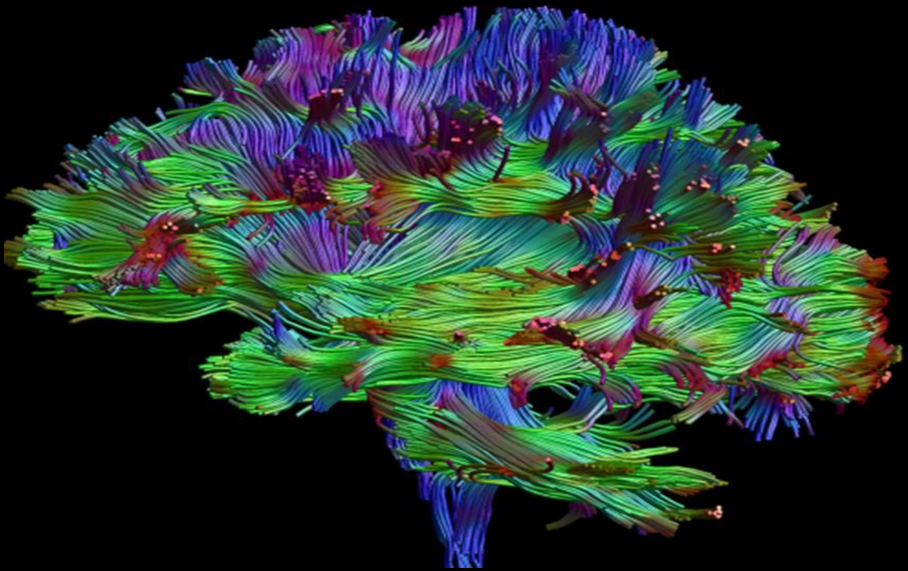


Data-Intensive Imaging Technologies









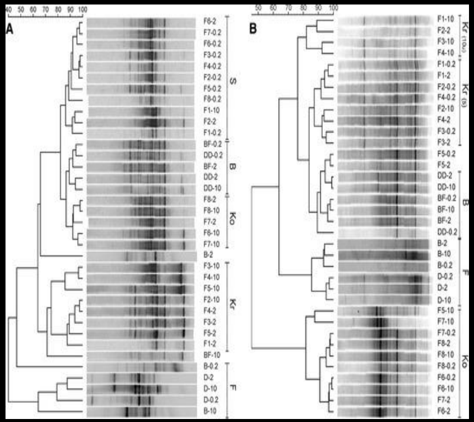
Data-Intensive Biomedical R&D and 'The Data Deluge'

Patient Stratification For Clinical Trials

Pharmacogenomics

m.Health

Remote Health Monitoring



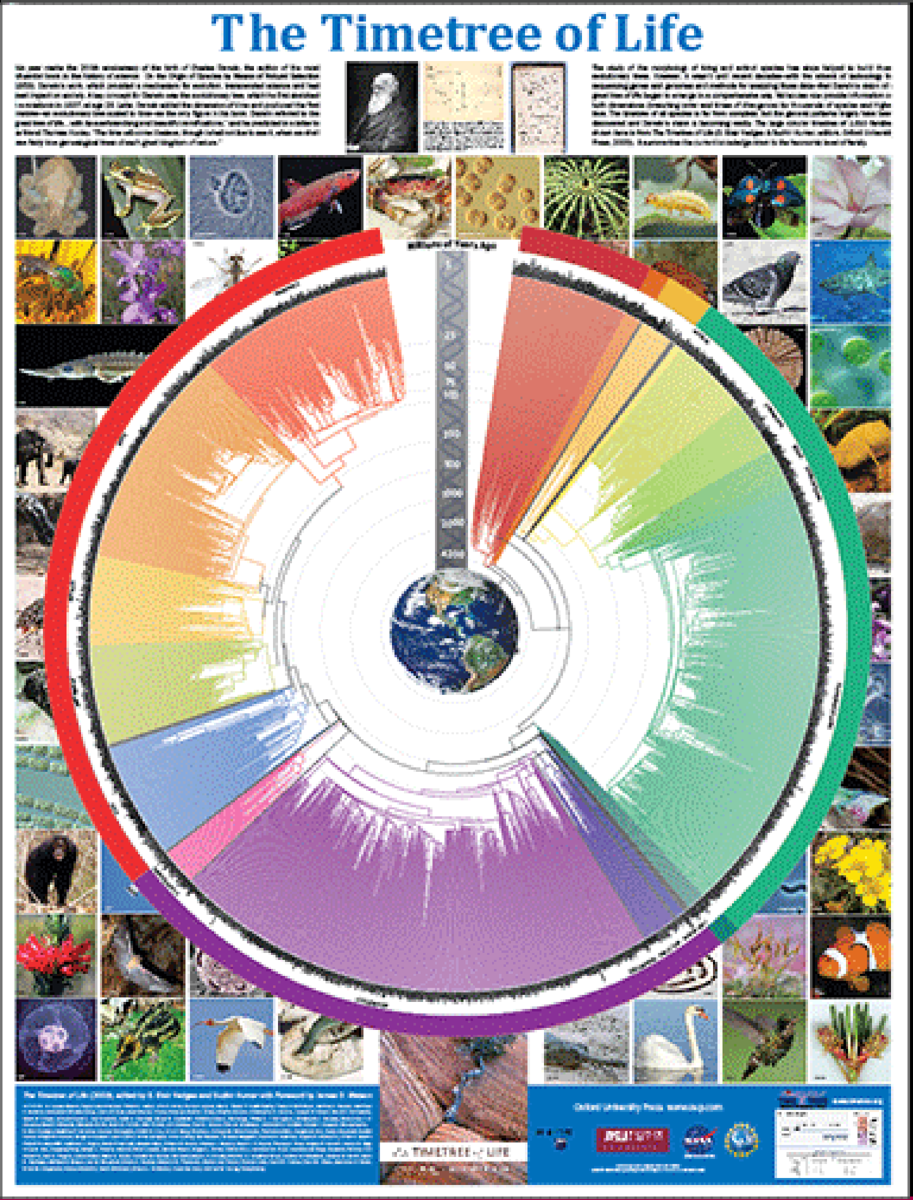
Microbial Diagnostics

Biosurveillance and Public Health

High Performance Computing

Health IT and EMRs

From Linneas to Life Codes:
Mapping Biological Diversity and a New Digital Taxonomy

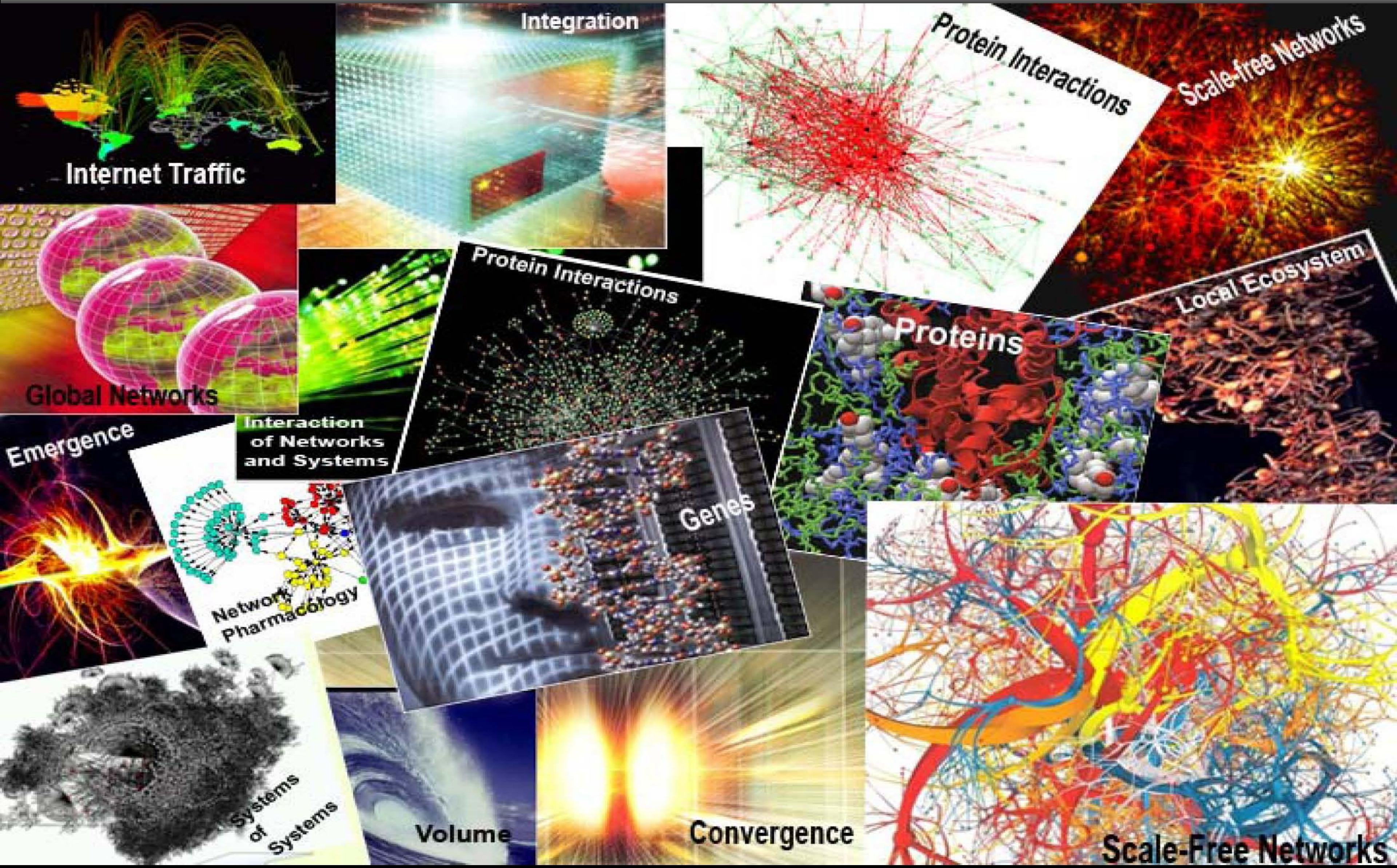


NCBI

GenBank release 185.0
14 August 2011

- 131 gigabases of data from 142 million entries of non-whole genome sequencing
- 208 gigabases from 65 million WGS
- additional 970,764 records updated

Data: The Fastest Growing Resource on Earth



Exabyte World

- projected 1800 exabytes of new global data in 2011 (10x more than 2006)
- routine multi-petabyte data sets emerging in national security and big science
- Large Hadron Collider estimated 15 petabytes/year
- smart electricity grid: 100 million customers ≡ 50 petabytes/year before compression
- Walmart: 1 MM transactions/hr = 2.5TB
- Boeing Dreamliner jet engines produce 10TB operational status/30mins
- Twitter ecosystem: 8TB data/day vs NYSE 1TB/day
- meta-data (information about information)
  - – growing as fast as data in big data environments

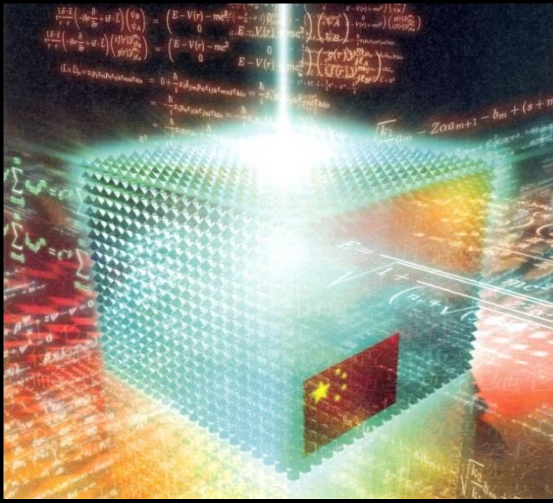
Managing “Mega-Data” in Biomedicine

volume

computational scale

global networks

bench to bedside: multiscale heterogeneity



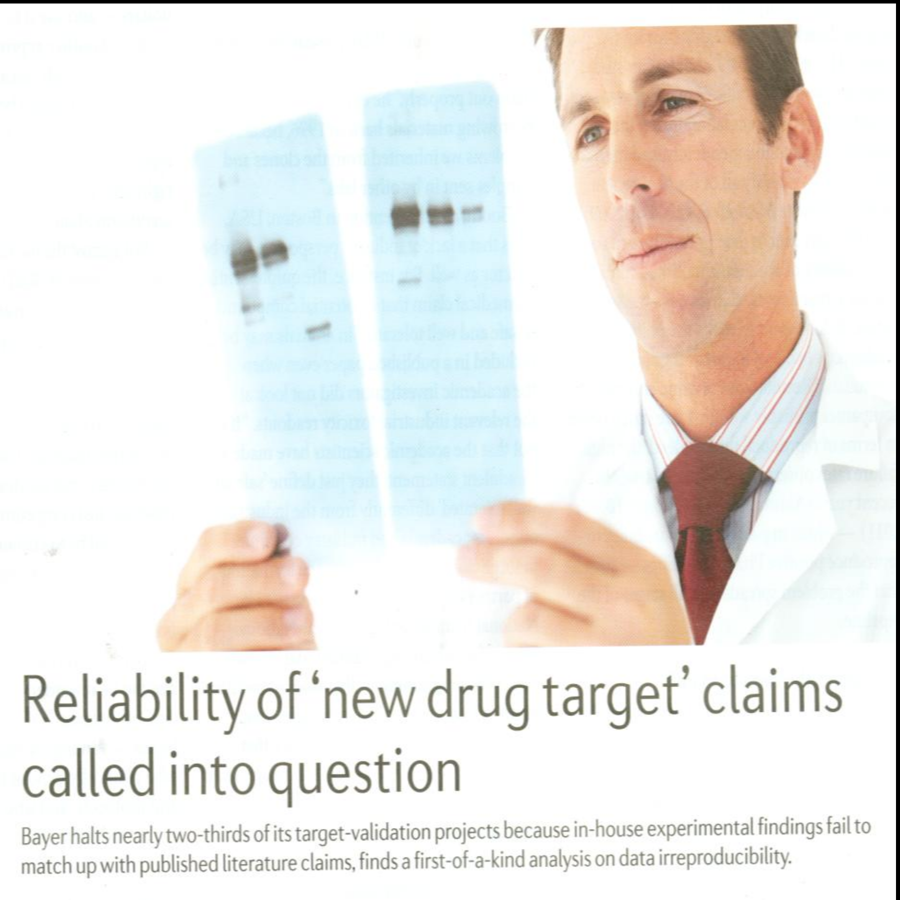
integration

Building Large Scale Biomedical Collaborative Research Networks and Commitment to Standardization of Experimental Platforms, Models and Data Reporting

Website: **IPSOGEN (**http://www.ipsogen.com/corporate/)

The Proliferation of Poorly Standardized, Non-Reproducible and Statistically Flawed Research Data

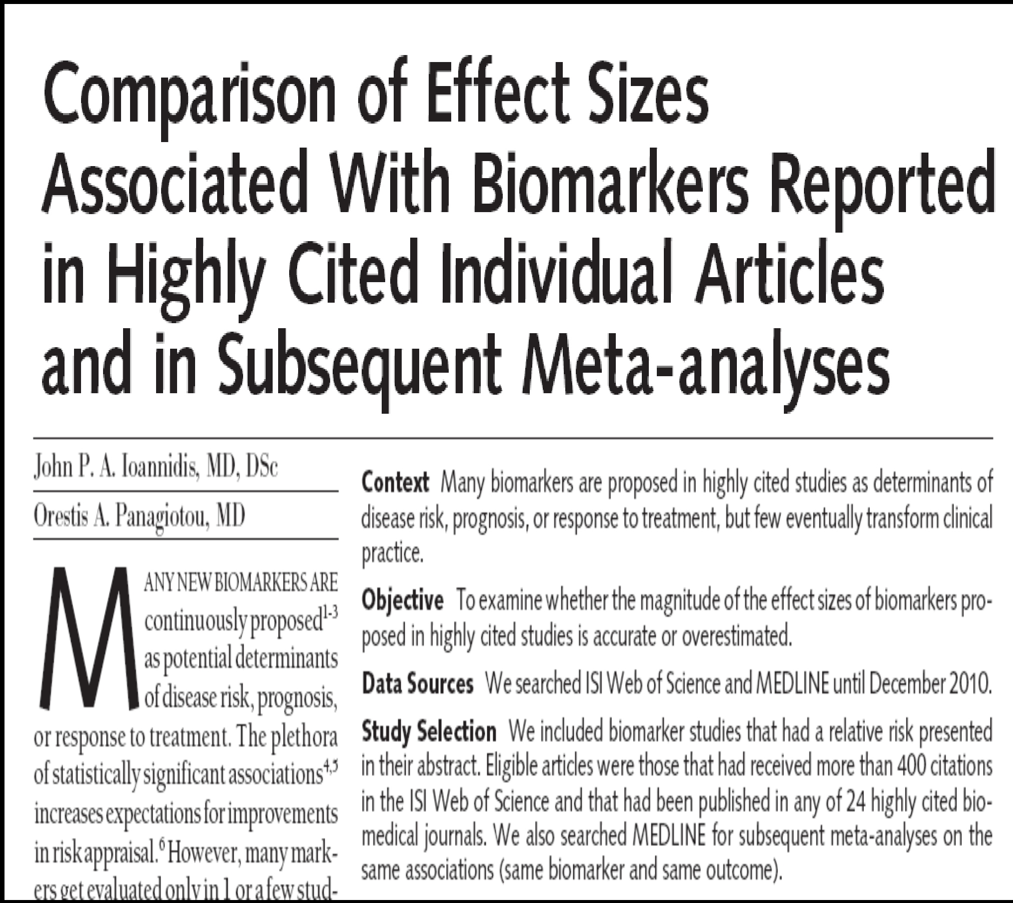
Nature Rev. Drug Disc. (2011) 10, 643



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.

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

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-

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).

Garbage Data, Fragmented Data, Selfish Data and Untapped Data: Pervasive Deficits in the Conduct and Organization of Academic Research

- poor access to rigorously annotated biospecimens from stringently phenotyped sources
- insufficient control of pre-analytical parameters and variable analytical standards
- idiosynchratic individual investigator methods
- ‘small N’ studies lacking statistical power
- chaotic data reporting formats and poor dbase interoperability
- pressure to publish and poor compliance with funding agency/journal policies on data sharing
- failure to work to industry standards

Mapping the Human Variome: Defining the Molecular Taxonomy of Individuality and Correlations With Phenotypic Traits and Disease Processes

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

The Expansion of Human Genome Sequencing Projects

1000 Genomes
A Deep Catalog of Human Genetic Variation

The Cancer Genome Atlas

HMP

Personal Genome Project
www.personalgenomes.org

ClinSeq®

International Cancer Genome Consortium

life technologies™

EUVADIS
EUROPEAN VACCINE DEVELOPMENT INSTITUTIONAL SUPPORT

华大基因
BGI

NATIONAL INSTITUTES OF HEALTH

CureSearch
Children's Oncology Group

illumina®

Complete genomics

medicure

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

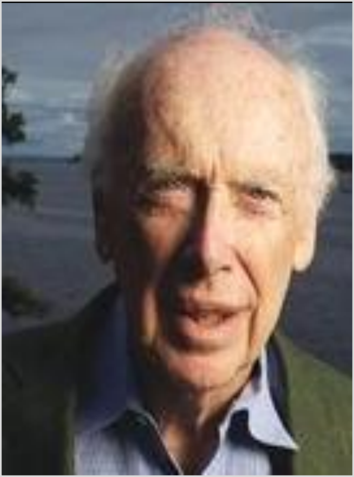
The Scale and Complexity of Human Genome Sequencing Data

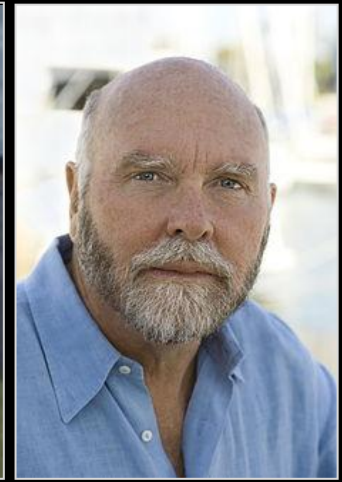
Accuracy and Comprehensiveness

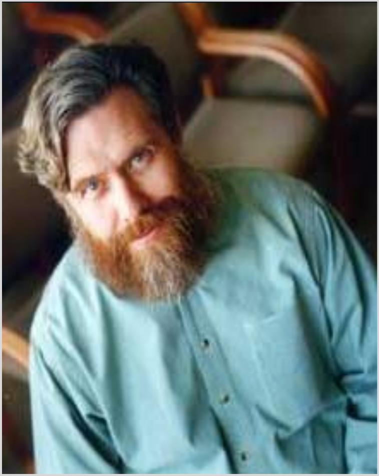
- need for consensus metrics for these parameters
- population-based studies
  - pooled samples with low depth coverage (<10x)
- personal genomes
  - greater accuracy and confidence for base calling for clinical diagnostics and care decisions
  - regulatory oversight of QA/QC and analytics algorithms
- current technologies
  - 30-40x coverage to ID 92-95% of both alleles
  - 50-100x coverage to ID 99.9% sequence and rare variants
  - final sequence with only 1 error/10^6 bases will still contain 6000 errors

What Is A Healthy Genome?

Early Members of the 3 Gigabyte WGS Club





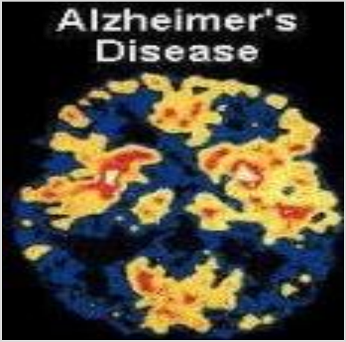


Consumer Genomics: Hype or Personal Freedom?

23andMe

Navigenics

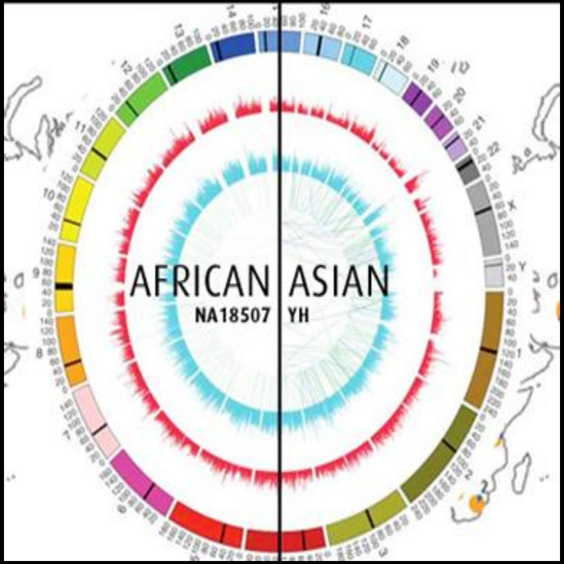
Analysis of Probabilistic Risk(s)



Interpretation of the Functional Effects of Variants

- prediction of deleterious missense SNPs
  - significant fraction in non-coding regulatory regions
- new methods needed to evaluate functional effects of synonymous and intronic SNPs, insertions, deletions and CNVs
- ‘phasing’
  - ID on which of the two chromosomes a variant is located
- emerging evidence of widespread RNA and DNA sequence differences (RDD) in human transcriptome
  - RNA sequences do not match DNA
  - some RNA variants at RDD sites translated into proteins
  - unknown role in biology and/or disease

The Perils of Assembly and Annotation: Total Base Pair Discrepancies in Published WGS of Han Chinese (YH) and Yoruban (Y) Individuals Versus Reference Genome



- C. Alkan et al. (2011) Nature Methods 8, 61
- de novo assemblies were 16.2% shorter than reference genome
- 420.2 megabases of common repeats and 99.1% duplicated sequences missing
- over 2,337 coding exons completely missing
- 136,613 bp in Y genome had high sequence identity to EpsteinBarr virus
- Y sequence generated from cell line vs YH from blood DNA

The Pervasive Problem of Poor Standardization and Annotation in Large Scale 'Global Profiling' of Genomes and Gene Expression

"The expertise and motivation to sequence genomes to a high quality are disappearing….
…if genome manuscripts' can now be published without accounting for the 20% that is missing
…what incentive remains to spend the additional effort and cost to sequence genomes well?"

"The balance between quantity and quality of genomes (must be) reestablished."

C. Alkan et al. (2011) Nature Methods 8, 61

"The excitement of having more and more tools always brings us back to the very important question of having to validate or replicate. I worry that that's getting lost as everyone gets so excited about the next really cool tool."

Dr. Stephen Chanock
Chief, Translational Genomics, NCI
Genome Technology April 2011 p. 31

The Cost of Sequencing
Versus
The Cost of Computational Analysis and Storage

- the $1000 genome,
  - the $? analysis and interpretation cost
  - the $? storage, retrieval and security costs
- turn around time (TAT) and analysis for clinical value cost
- regulatory and reimbursement policies

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 with differences to reference set
  - 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 reacquirable/renewable
  - cancer

“Clinical Grade” Genome Sequencing: Ready for Prime Time?

Disparate Views on Timing

National Human Genome Research Institute

“…pie in the sky ideas about what the clinician
could do if they had at their disposal
a genome and a set of analytical tools.”

….we don’t know how to do any of this.”

….it will take years, if not a decade or so,
to implement widely and effectively.”

Dr. Les Biesecker
Chief, National Human Genome Research Institute, Genetic Disease Branch
IOM Symposium 2011

“Next generation sequencing is truly
changing the way we’re treating patients.”

Francois Ferre
CEO, Althea Dx
Genome Web 22 March 2011

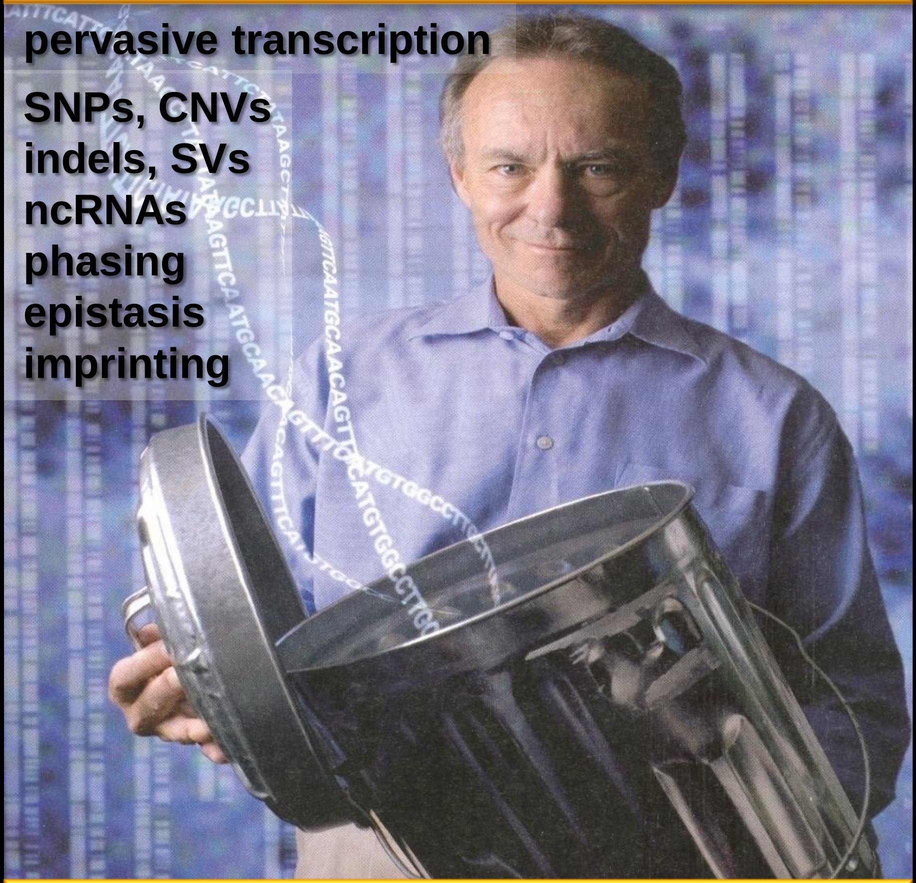
Regulatory Issues in Genome Sequencing for Clinical Decisions

- 23 June 2011 workshop
- accuracy, depth of coverage, validation set, impact of pre-analytic/analytic variables
- CLIA/CAP facilities
- sequencers as Class III devices?
- RUO materials
- source computer code(s) for analytical algorithms
- performance thresholds and QA/QC requirements for error detection (instrumentation + analytics)

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

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

Junk No More!

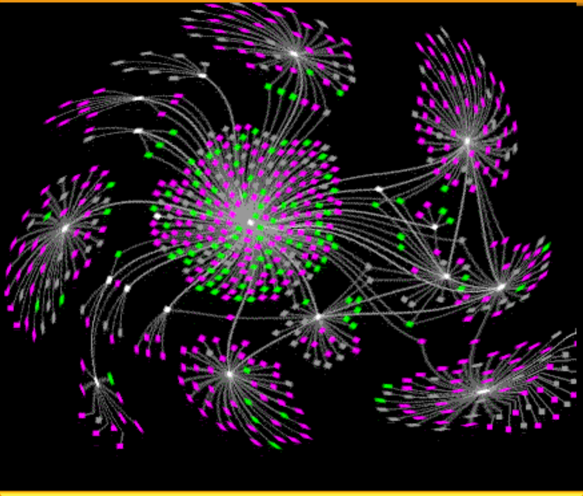


pervasive transcription

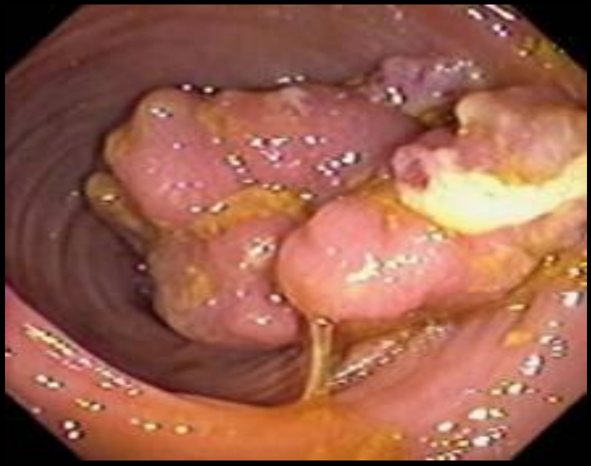
SNPs, CNVs
indels, SVs
ncRNAs
phasing
epistasis
imprinting

recognition of increasing organizational and regulatory complexity

Molecular Interaction Networks



Disease Perturbations



Mapping Information Flow in Biological Systems: Modules, Pathways, Circuits, Subnetworks, Networks and Systems

- “An Algebra for Theoretical Genetics”
  Ph.D. Thesis, MIT (1939)
- “A Mathematical Theory of Communication”
  Bell System Technical Journal (1945)
- “It is obvious from the analysis of these (bacterial genetic regulatory) mechanisms that their known elements could be connected into a wide variety of ‘circuits’ endowed with any desired degree of stability.”

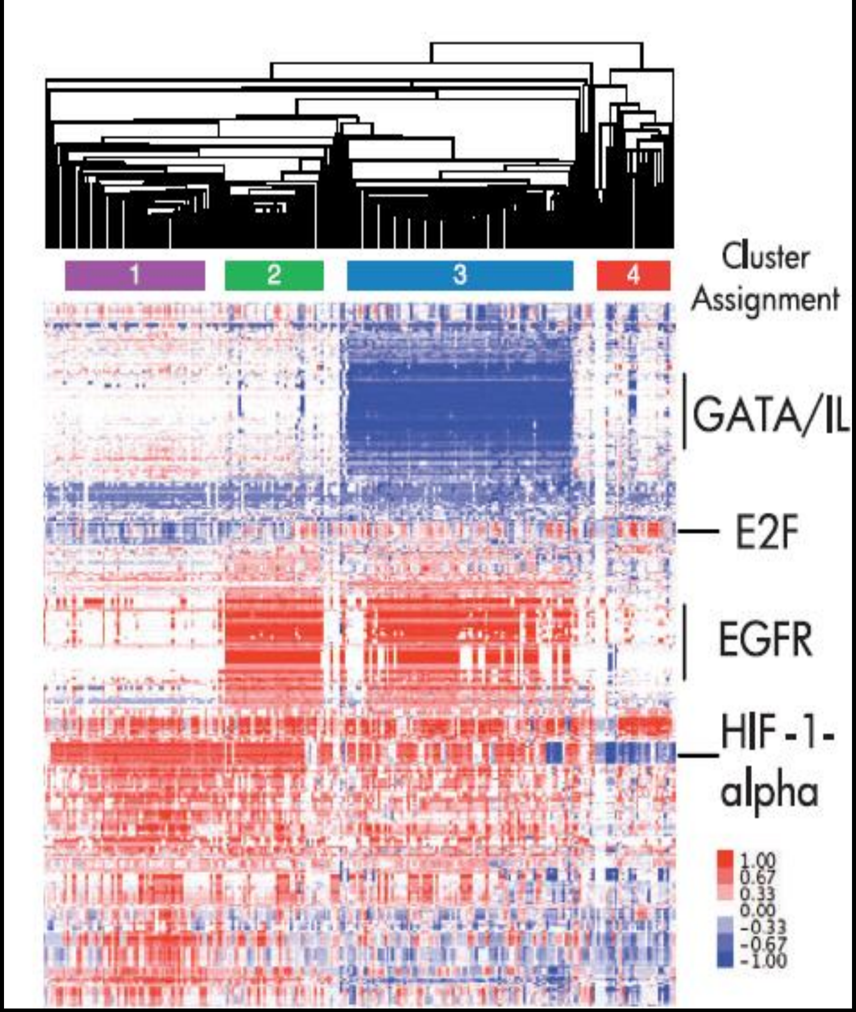
François Jacob and Jacques Monod (1962)
Cold Spring Harb. Symp. Quant. Biol. 26, 193

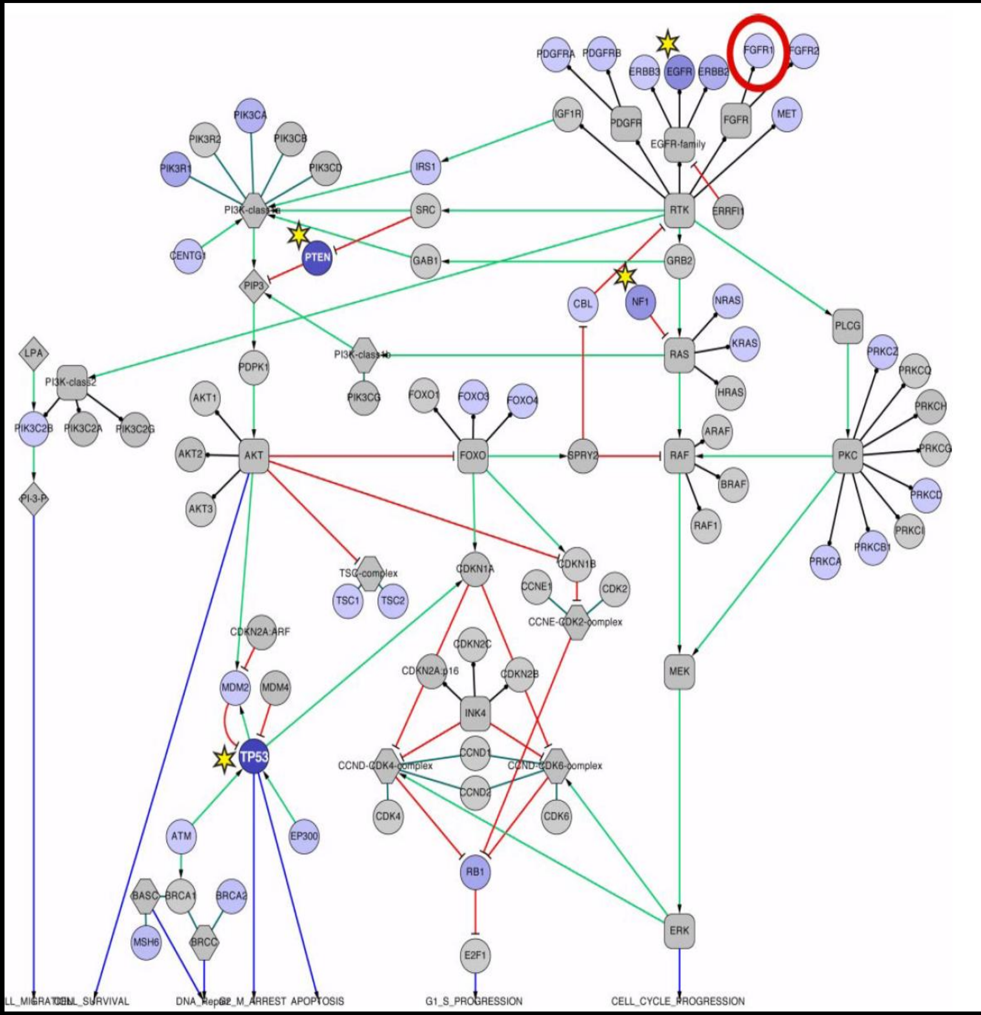
The Non-linear Relationship Between Genotype and Phenotype: Coding for Complex Molecular Networks (Modules, Pathways, Subnetworks)

- the current ‘black box’ of major knowledge voids that create poor predictability in accurate ID and selection of biomarkers/Rx targets with resulting high failure rates in clinical trials

- has the gap between basic science and realizable therapeutic applications widened?
- how can systems complexity be deconvoluted to identify tractable approaches for new molecular diagnostics, targeted therapy out and disease risk predisposition assessment?

Mapping Modules, Pathways and Subnetworks in Biological Systems: The TCGA Glioblastoma Multiform Dataset and Pathway Analysis





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

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

Molecular Interaction and Signaling Pathway Resources

BIOCARTA

In stke
Signal Transduction Knowledge Environment

PANTHER
Classification System

KEGG
Kyoto Encyclopedia of Genes and Genomes

BioGRID

METACYC™

Human Protein Reference Database

REACTOME

ChiBe

MINT

NATIONAL CANCER INSTITUTE
nature PathwayInteractionDatabase

IMID

PATIKA

PathCase SB
featuring models from BioModels

BETA
WIKIPATHWAYS
Pathways for the People

PID

Pfam

CST
COMPUTER SIMULATION TECHNOLOGY

PC Pathway Commons

VISANT

NETPATH
Custom Internet & Business Data Program

HUMANCYC™

INTACT
TECHNOLOGY

PathScan® ELISA

Mapping Pharmacological "Interaction Space" in Biological Pathways and Subnetworks

DrugBank

Therapeutic Targets Database

BIDD
Bioinformatics and
Drug Design group

SUNSET
MOLECULAR

PUBLISHERS OF WOMBAT
WOrld of Molecular BioAcTivity

PDTD [Potential Drug Target Database]

CTSA Pharmaceutical Assets Portal

chem2bio2rdf

Reducing The Failure Rate of Investigational Drugs in Clinical Trials

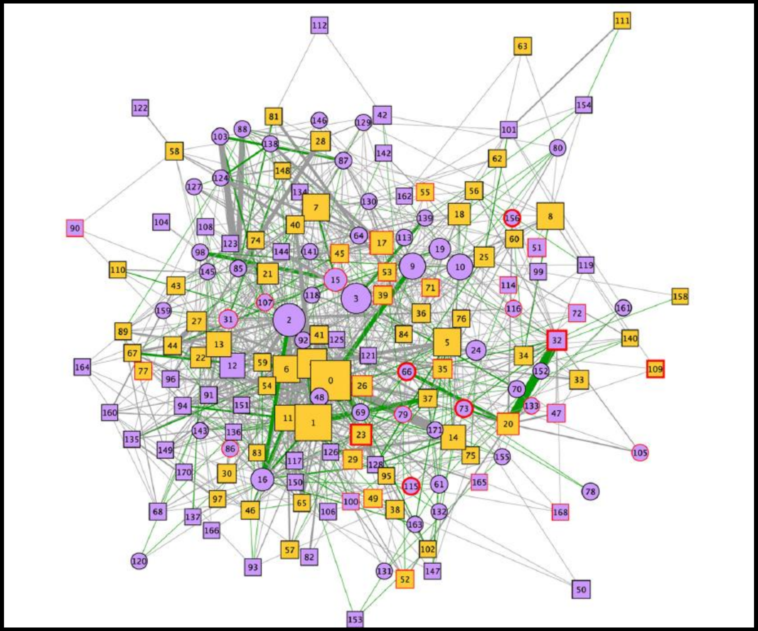
- targeted therapies, YES!

but

- improved success requires targeting network modules, pathways and subnetworks **not single targets**
- complexity of linked and overlapping modules and pathway "cross-talk"
  - long range pleiotropic effects
  - weak indirect effects

Mapping of Protein Interaction Network in Alzheimer's Disease (AD)

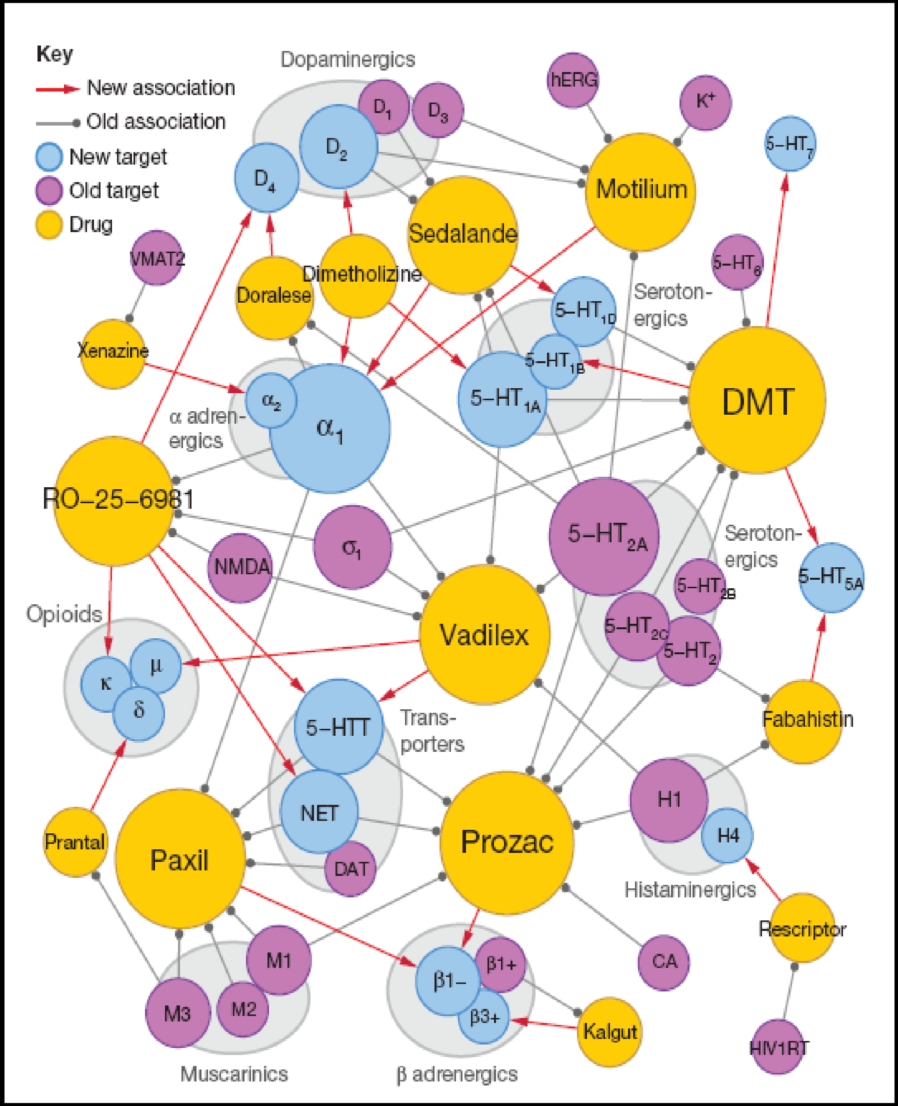
From: M. Soler-Lopez et al. (2011) Genome Res. 21, 364



- 200 high confidence 2P interactions
  - 8 confirmed AD – related genes
  - 66 additional candidates
  - 31 in chromosome regions containing putative susceptibility loci
  - 17 dysregulated in AD

Place Your Bets!

Network Pharmacology

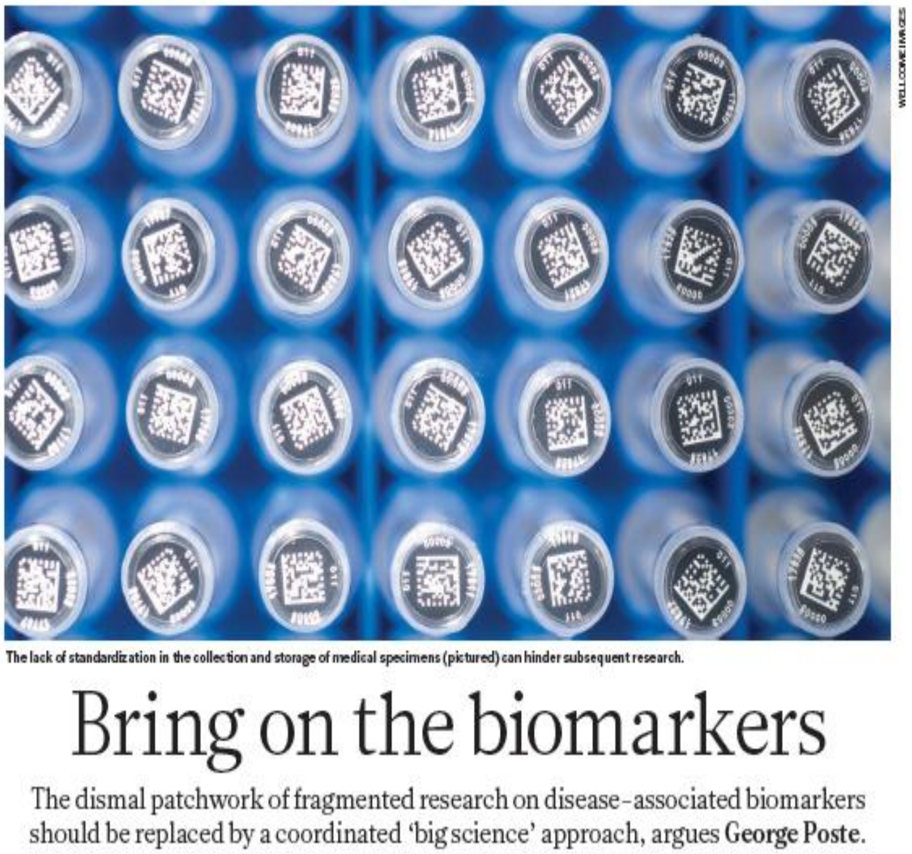


From: M. J. Keiser et al. (2011) Nature 462, 180

- same drug: interaction with multiple targets
- same target: interaction with multiple drugs
- mapping structural chemotypes to pathways and subnetworks for targeted (poly)pharmacology

Network Based Perturbations in Disease: Implications for Biomarker Discovery and Validation

G. Poste (2011) Nature 469, 156



WELLCOME IMAGES

The lack of standardization in the collection and storage of medical specimens (pictured) can hinder subsequent research.

Bring on the biomarkers

The dismal patchwork of fragmented research on disease-associated biomarkers should be replaced by a coordinated 'big science' approach, argues George Poste.

- 'publish and vanish': poor productivity due to failure to evaluate network effects
- if disease involves multi-loci perturbations in modules/ subnetworks then multiple parameters will need to be measured
  - different multiplex biomarkers in different disease subtypes
  - changes in biomarker profile with disease progression/Rx response/Rx resistance

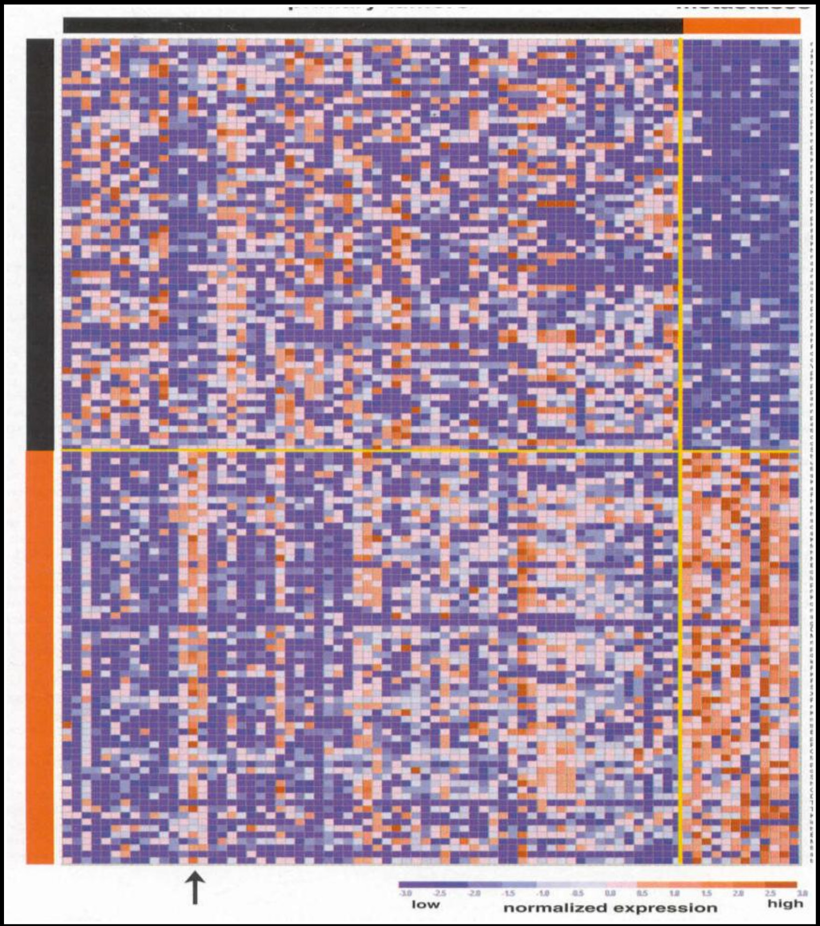
Understanding Complexity

"It takes a network
to stop a network."

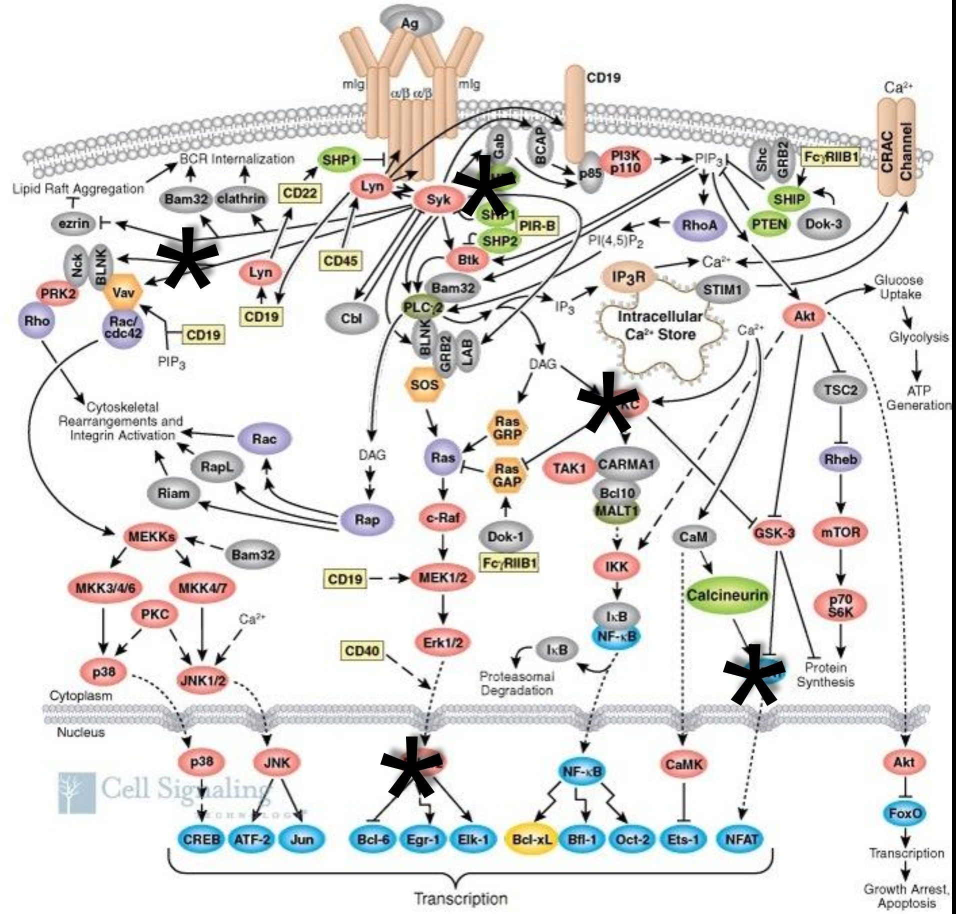
Lt. Gen. S. McChrystal
(on combating the Al-Qaeda and Taliban in Afghanistan)

Understanding the Internal Circuit Diagrams of Cells and Identification of the Disruption(s) Caused by Disease

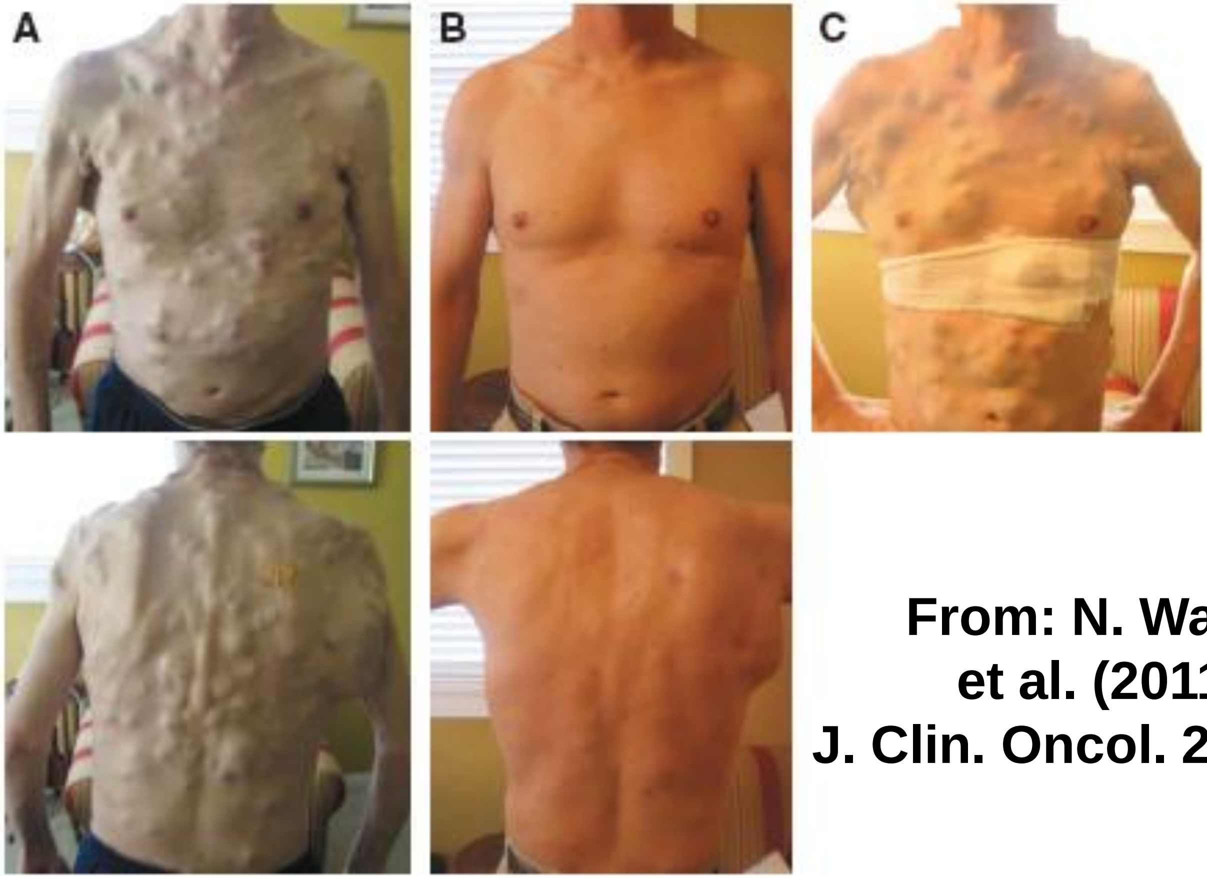
Disease Profiling to Identify Subtypes (+ or - Rx Target)



ID Molecular Targets for Rx Action and Blockade of Compensatory "By pass" 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. Wagle
et al. (2011)
J. Clin. Oncol. 29, 3085

http://jco.ascopubs.org/content/29/22/3085.full.pdf+html

Network Pharmacology

- elucidation of definitive ‘chokepoints’ as optimum targets
  - subvert adaptive cellular options to use alternate compensatory pathways
- the design challenge for multi-target polypharmacology
  - multi-agent therapy (patient tolerance?)
  - controlled multi-target promiscuity in a single moiety
- does chronic progression in complex, multigenic diseases amplify pathway/subnetwork dysregulation?
  - greater complexity of multi-target Rx for homeostatic ‘reset’
  - role of Rx in driving selection of variants with Rx-resistance (‘escape’) pathways (e.g. oncology)

Improving the Productivity and Proficiency of Biomedical R&D and Clinical Medicine: An Unavoidable (But Essential) Transition to Data-and Computation-Intensive Methods

An Unavoidable Transition to Data-and Computation-Intensive Methods

Current Era

- a high opinion, low robust information content world
- “silos” of research/clinical activities and slow adoption of “systems-based” cross-disciplinary integration
- proliferation of poorly standardized and fragmented data, semantic anarchy and incompatible databases
- poor predictability of the behavior of complex biological networks and accompanying ‘rude shocks’
  - clinical trial failures (biomarkers, Rx)
  - inaccurate diagnosis and flawed clinical decisions
  - highly variable treatment selection and uncertain clinical outcomes
  - extravagant waste and risk to patients

An Unavoidable Transition to Data-and Computation-Intensive Methods

Current Era

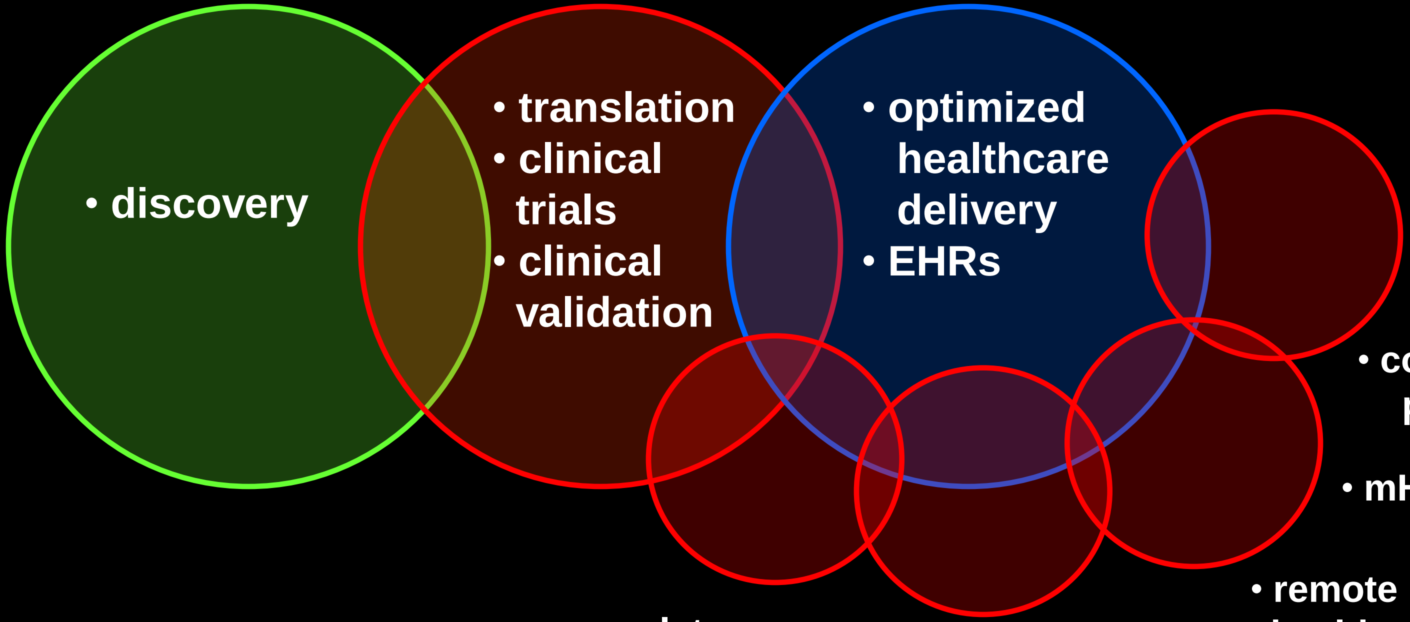
The Imperative to Move from Descriptive Phenomenology
to
Mechanism-Based Knowledge of the
Behavior of Complex Biological Networks
to
Achieve Precision Diagnosis, Rational Targeted Rx Design
and Improved Clinical Outcomes

An Unavoidable Transition to Data-and Computation-Intensive Methods

Pending Era

- “massive data”: automated, high throughput, massively parallel profiling tools to study body function
- technology acceleration and convergence (life sciences, engineering, telecommunications, computing)
- adoption of stringent analytical QC/QA and consistent ontologies for data reporting
- new annotation tools for ‘big data’ and interoperable dbase designs for facile cross-disciplinary/cross-sector integration
- new computational architectures and services for mega-data and machine-based mining and analysis

The Imperative for Integrated Inter-Operable ACKM Capabilities Across the Full Continuum from Discovery to Patient Care



- discovery

- translation
- clinical trials
- clinical validation

- optimized healthcare delivery
- EHRs

- consumers/ patients

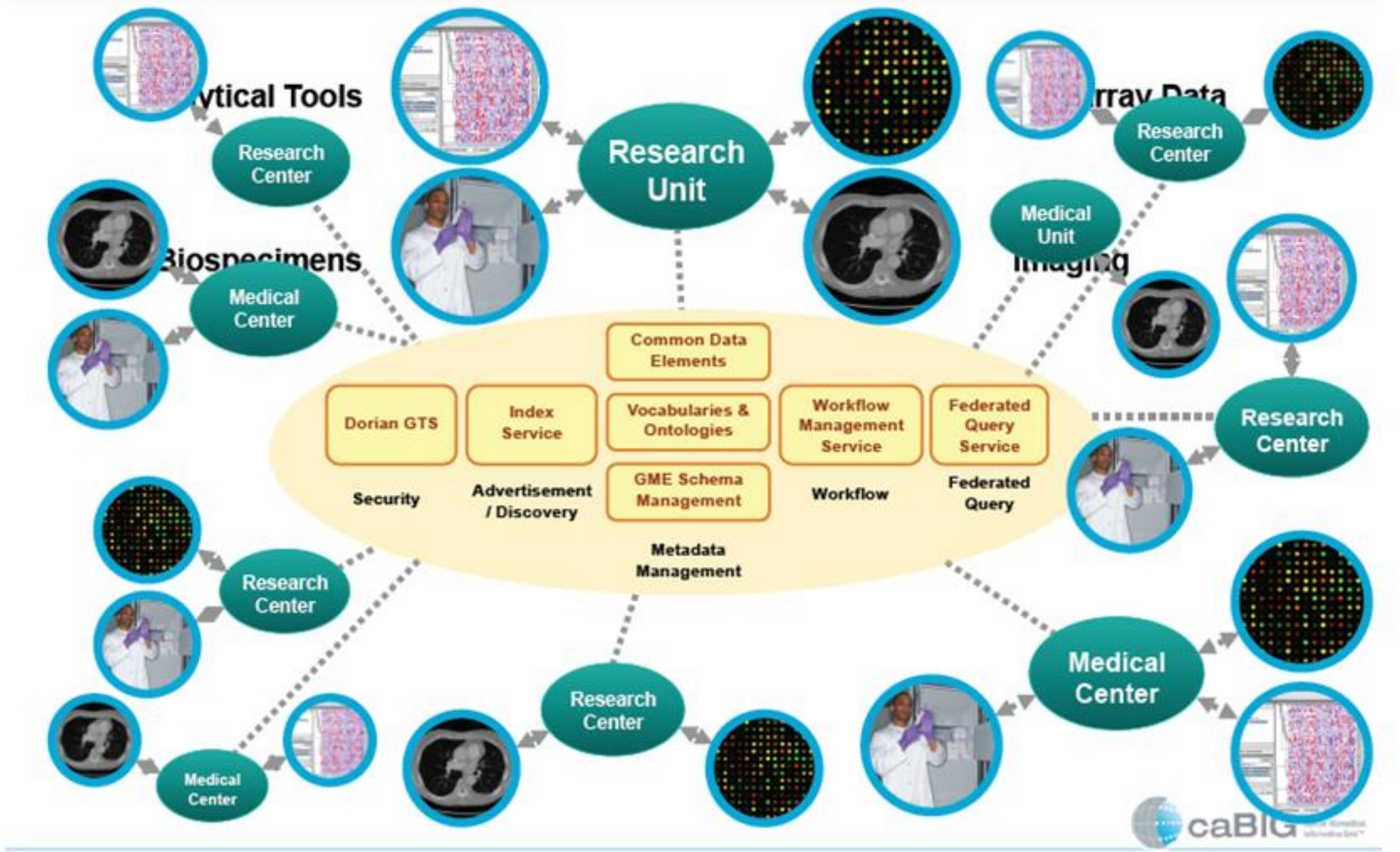
- mHealth

- remote health status monitoring

- regulatory
- reimbursement

- outcomes analysis
- CER

IT-enabled Ecosystems for Discovery and Translational Research



Adapted from: K. Buetow, NCI 2011
http://sagecongress.org/Presentations/Buetow.pdf

Data Integration and Exchange Standards

- chaotic state of discovery stage semantics, standards
- limited research dbase inter-operability with industry/regulatory standards for clinical trials
- leveraging existing HL7 standards for clinical trials
  - Draft Standards for Trial Use (DSTU)
  - CDISC, ICH
- Digital Imaging and Communications in Medicine (DICOM)
- seamless federation with healthcare system and reimbursement databases
  - CPT, ICD (USA)
- certification of compliance with proposed HITECH EHR Standards (HIMSS, AHIMA)

The Painfully Slow Adoption of Sophisticated Electronic Tracking Systems in Healthcare for Improved Patient Care

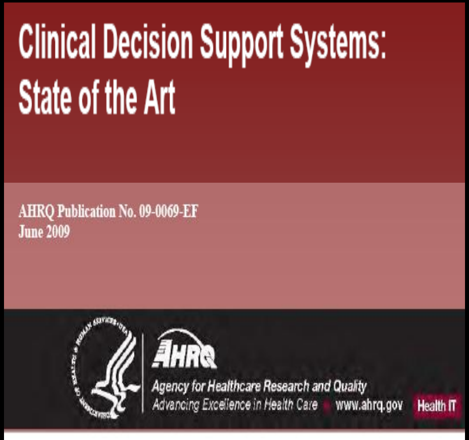
"In one 24 hour period,
465 children were admitted to the hospital with fever.
Their fever-like symptoms were recorded in the EMR in 278 ways."

Dr. C.B. Forrest, Professor of Pediatrics
Children's Hospital of Philadelphia
cited in: Registries-RemedyMD (2011)

The Children's Hospital of Philadelphia®

A Learning Healthcare System

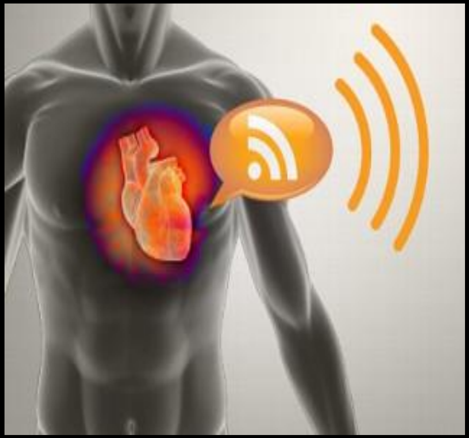
Proliferation of Clinical Computational Systems

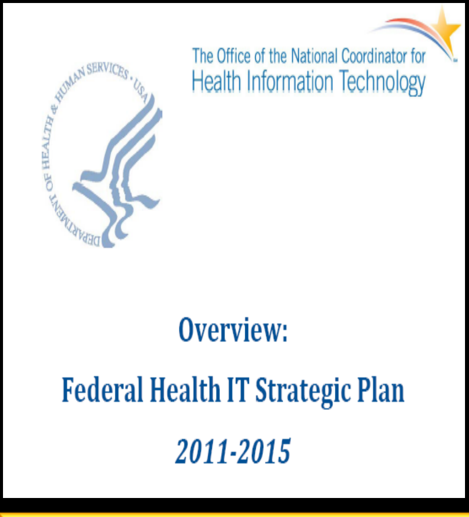


Clinical Decision Support Systems:
State of the Art

AHRQ Publication No. 09-0069-EF
June 2009

AHRQ
Agency for Healthcare Research and Quality
Advancing Excellence in Health Care www.ahrq.gov Health IT



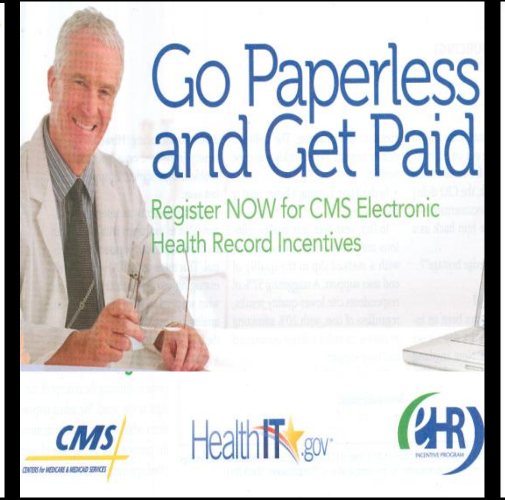


The Office of the National Coordinator for
Health Information Technology

Overview:

Federal Health IT Strategic Plan

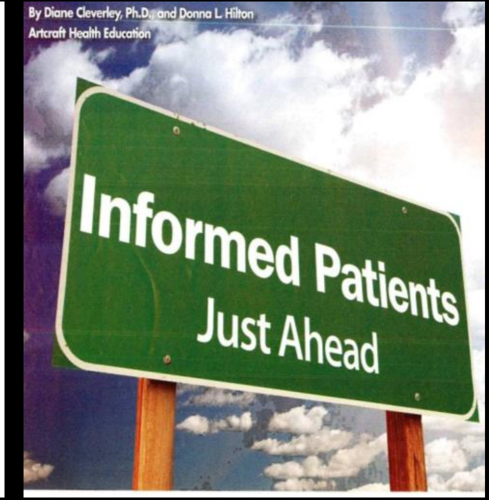
2011-2015



Go Paperless
and Get Paid

Register NOW for CMS Electronic
Health Record Incentives

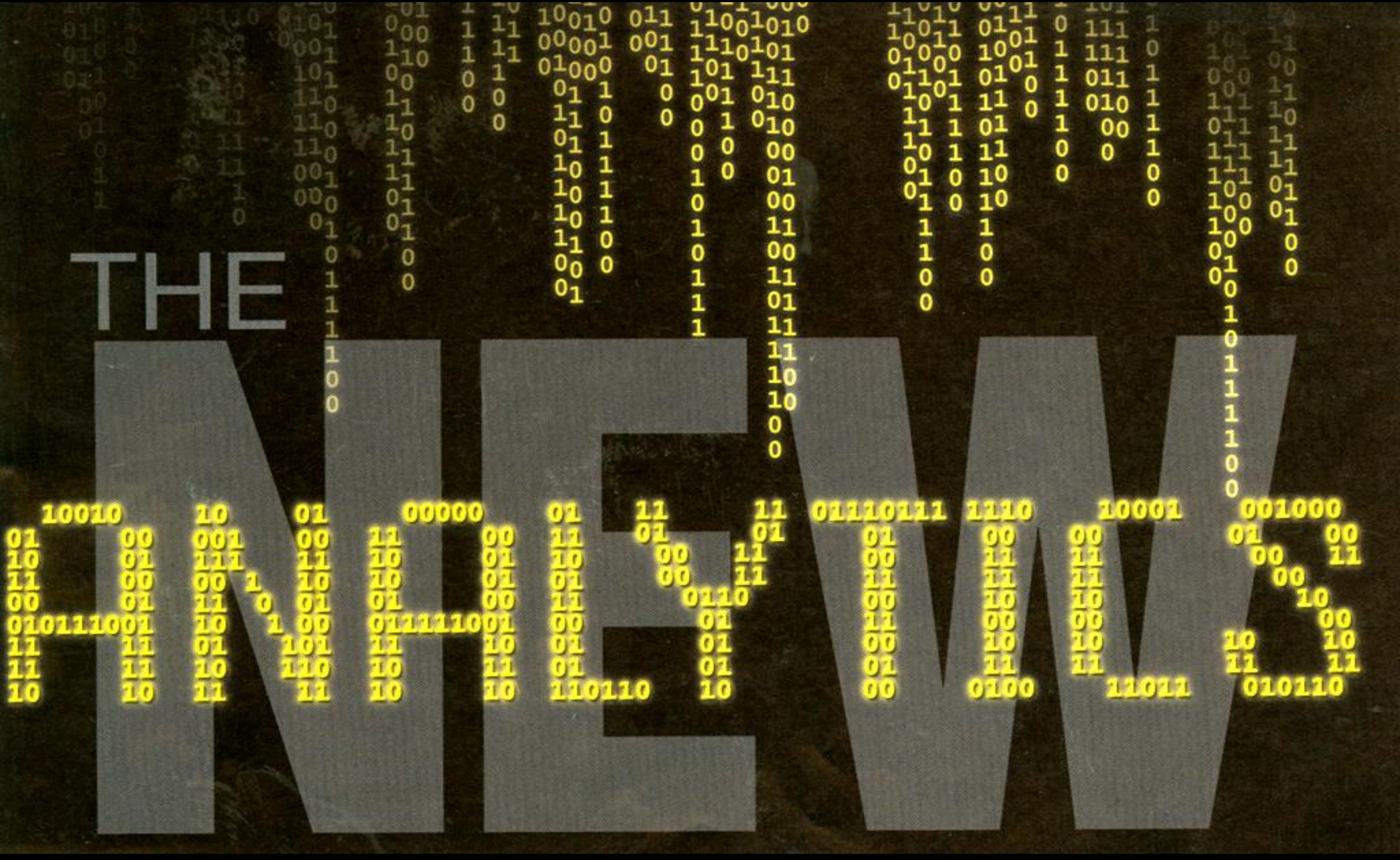
CMS CENTERS for MEDICARE & MEDICAID SERVICES HealthIT.gov EHR INCENTIVE PROGRAM



By Diane Cleverley, Ph.D., and Donna L. Hilton
Artcraft Health Education

HITECH Mandates	Incentives	EHR and Smart Cards	Informed Consumers/Patients

The Only Valuable Data is Validated, Actionable Data



Mining EHRs to Identify Disease Correlations with Molecular Profiling Datasets and Improved Clinical Stratification (Phenotyping) of Patient Cohorts

Aetna

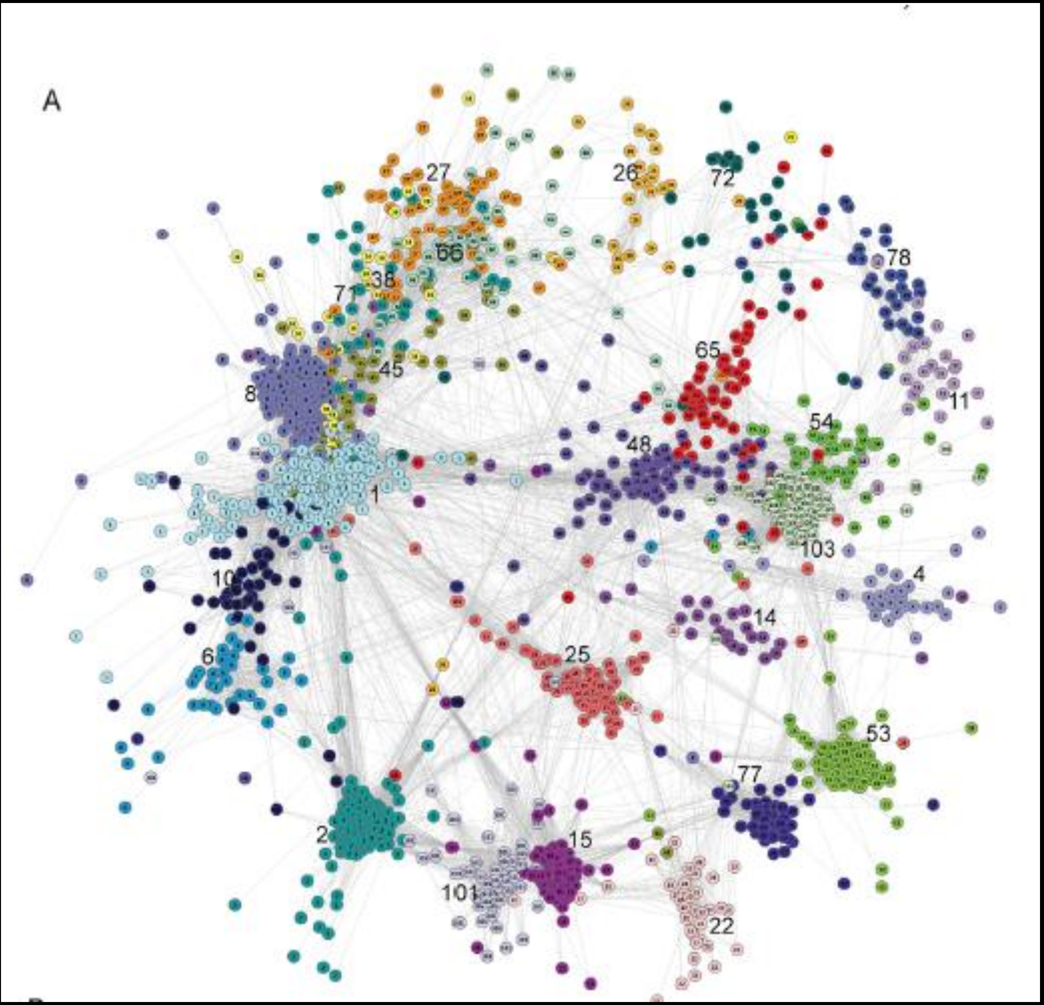
- 18.688 million medical members
- 13.953 million dental members
- 10.410 million pharmacy members

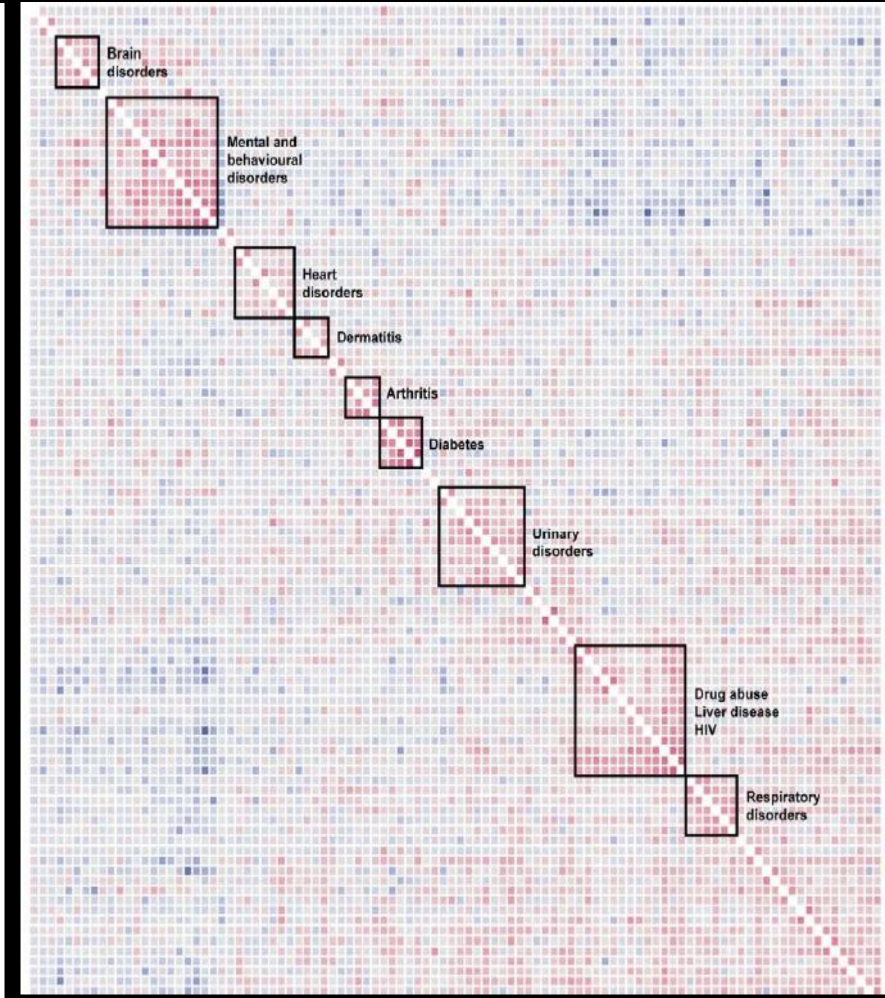
- 966,000 healthcare professionals
- 543,000 primary care doctor specialists
- 5,200 hospitals

- 71 billion health records
- 75 TB storage (50% occupied)

From: Health Data Sept. 2011

Mapping of 26 Clusters in a Phenotype Networks in 1497 Danish Psychiatric Patients By Combing Structured (ICD Codes) and NLP Processing of Medical Records





From: F. S. Roque et al. (2011) PLoS Comp. Biol. 7, e1002141

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 questions
- Prototype for intelligent systems for biomedicine beyond keyword IR searches

WELLPOINT | Health.Care.Value.℠

NUANCE®

The Status Quo is Neither Desirable Nor Sustainable

- continued undisciplined, free form tagging and idiosyncratic researcher/clinician-centric approaches
  - fragmented data silos and poor database interoperability/integration

versus

- adoption of highly engineered ontologies and formal logic
  - reference structures, vocabularies and semantics
  - automated machine-based algorithmic or rules-based scoring

OBO Foundry Ontologies

The Open Biomedical Ontologies

Cell Ontology (CL)

the Gene Ontology

Gene Ontology (GO)

Foundational Model of Anatomy

ZFIN

Zebrafish Anatomical Ontology

ChEBI

Chemical Entities of Biological Interest (ChEBI)

Disease Ontology (DO)

Plant Ontology™

PO

Consortium

Plant Ontology (PO)

SO

The Sequence Ontology Project

Sequence Ontology (SO)

Ontology for Clinical Investigations (OCI)

The Open Biomedical Ontologies

Common Anatomy Reference Ontology

The Open Biomedical Ontologies

Environment Ontology

OBI

Ontology for Biomedical Investigations

Phenotypic Quality Ontology (PATO)

PIR

A UniProt CONSORTIUM MEMBER

Protein Information Resource

Protein Ontology (PRO)

The Relation Ontology

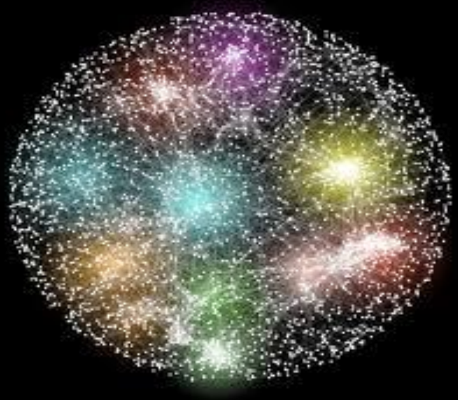
OBO Relation Ontology

RNA Ontology Consortium

RNA Ontology (RnaO)

Informatics and High Performance Computing

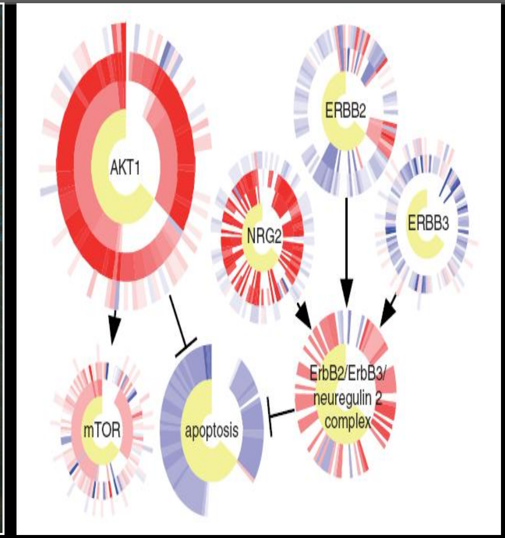
- modeling and simulation of biological networks of escalating complexity

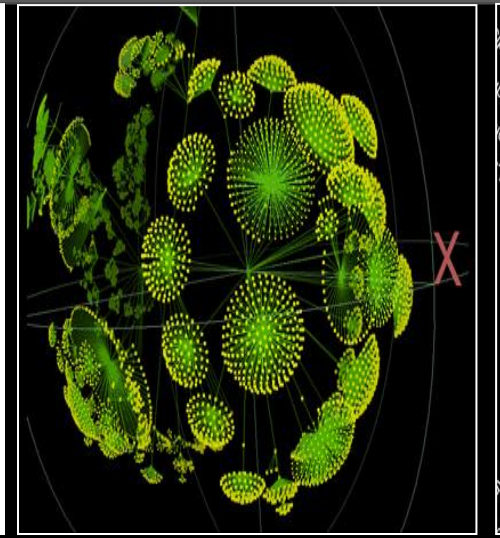


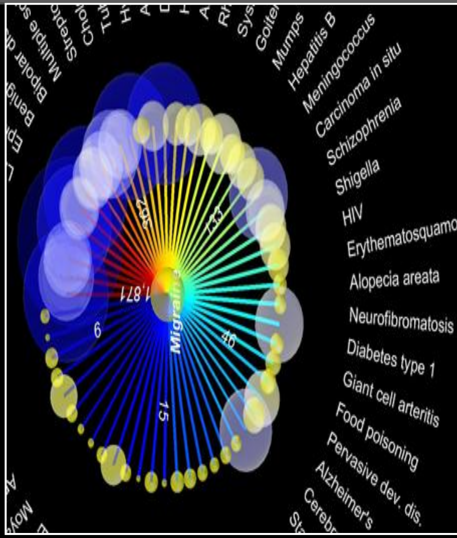
- development of new mathematical, statistical and computing tools for analysis and modeling of non-linear phenomena in complex networks

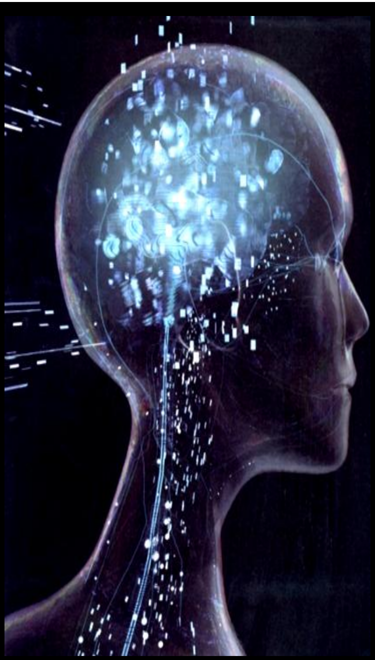
- application of advanced machine learning tools, avatars, robots and automated data production suites for customized data to promote optimum decision-actions

New Visualization Tools, Interactive Interfaces and Rapid Customization Formats

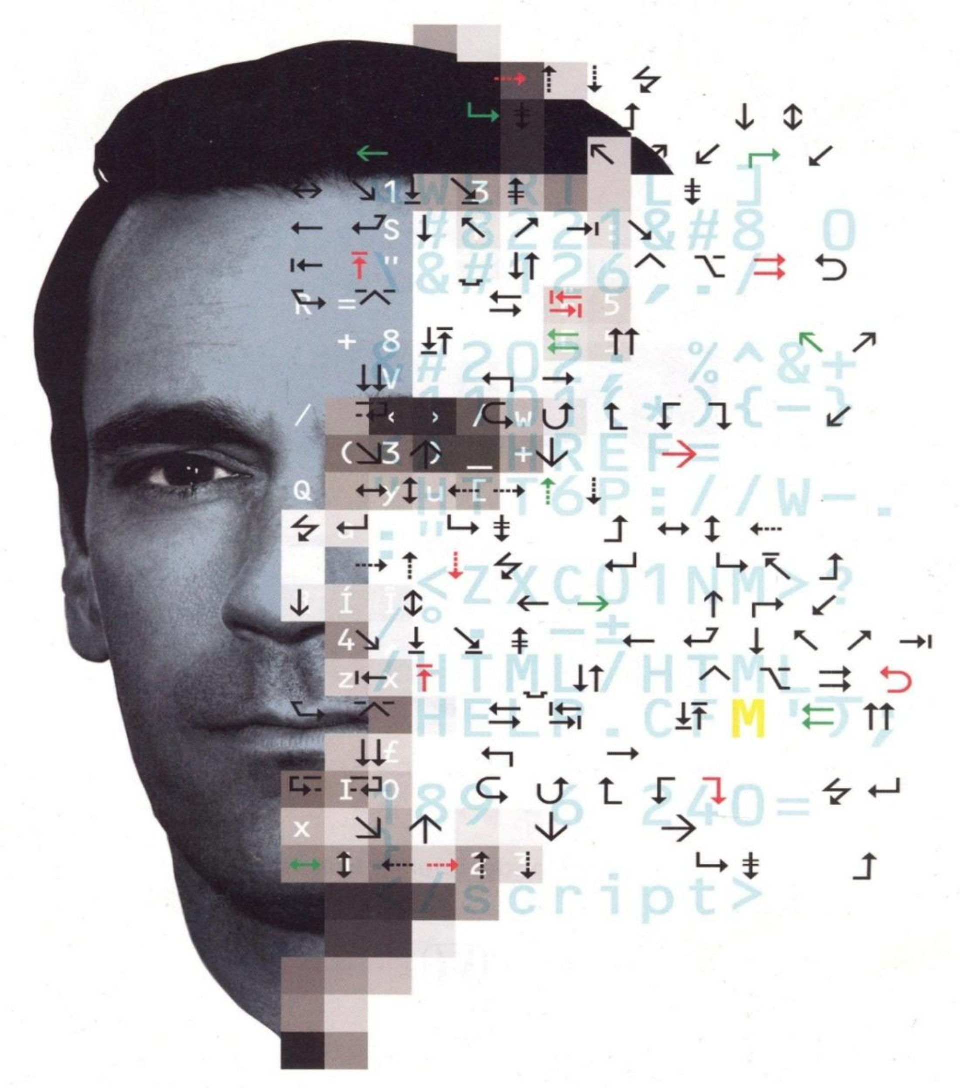




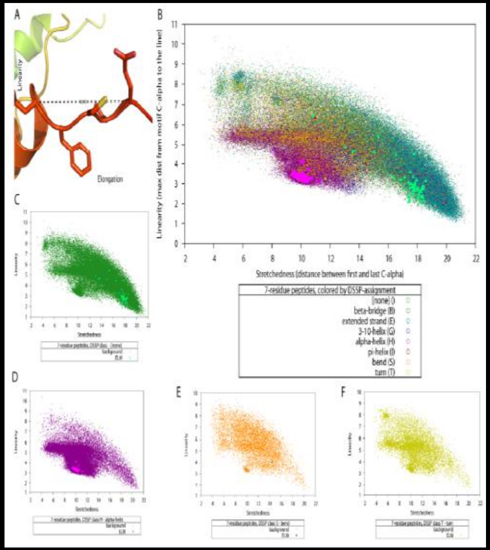


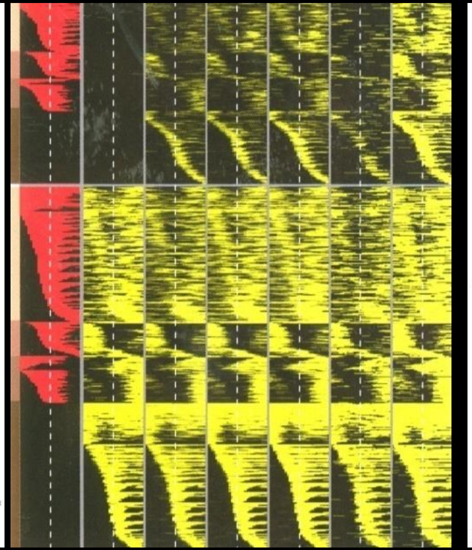


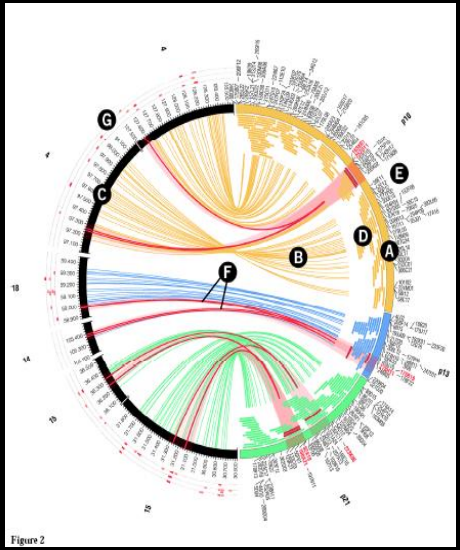
Technology Acceleration and Convergence: The Escalating Challenge for Professional Competency, Decision-Support and Future Education Curricula

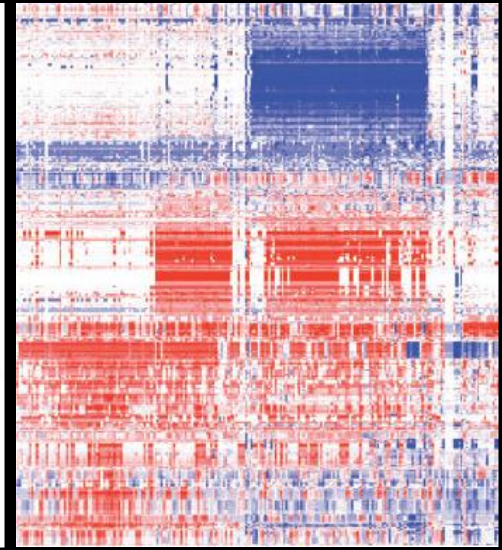


The Imminent Collapse of the Genome Informatics Ecosystem?







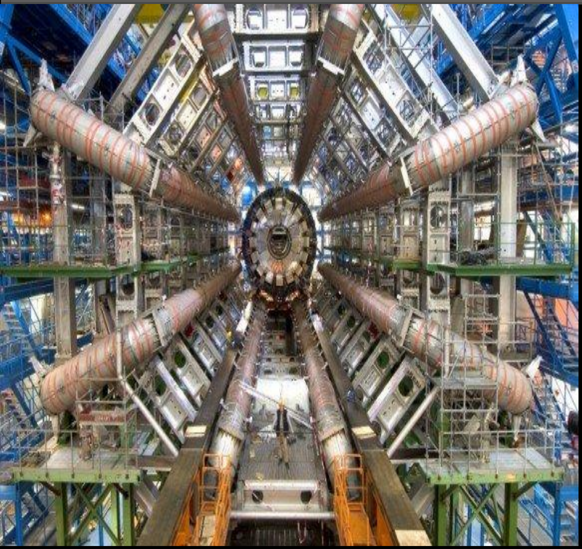


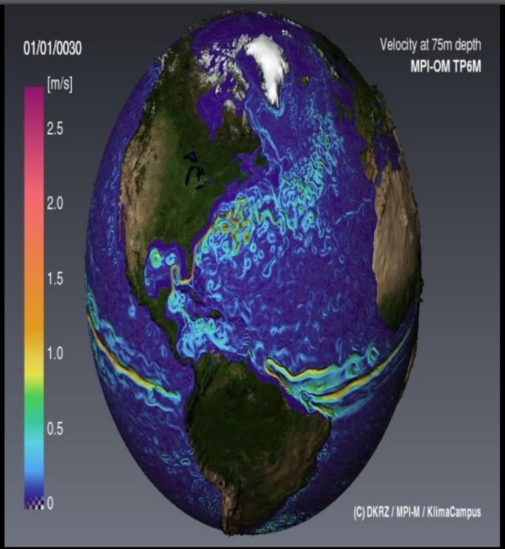
- Moore’s Law
  - # transistors/circuit board doubles c.18 months
- Kryder’s Law
  - hard disk capacity doubles c.12 months
- Butter’s Law
  - cost of sending a bit of information over optical network halves every 9 months
- NGS
  - sequence data doubles every 6 months (other ’omics will follow)

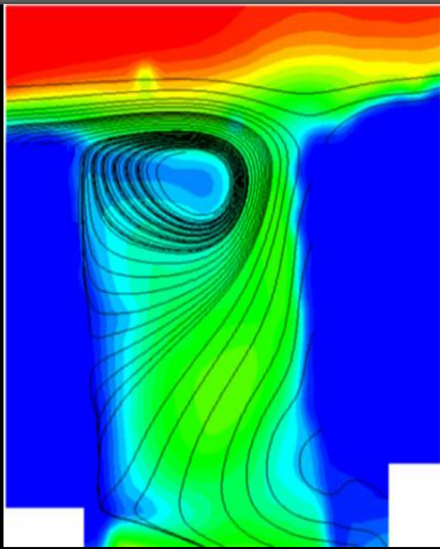
Managing Massive Data: Scale, Storage and Cost

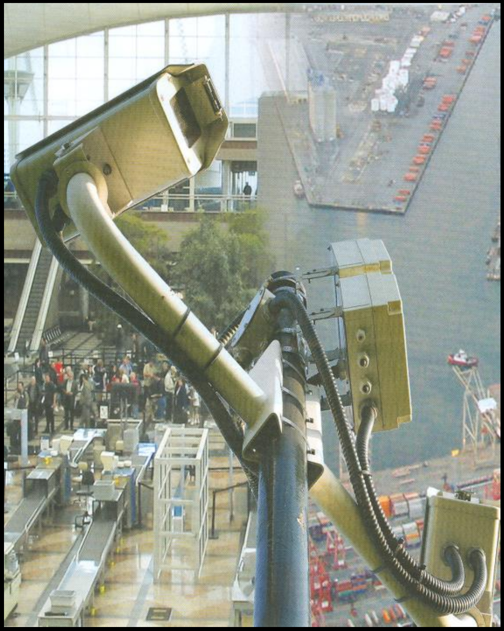
- the ‘lottabyte’ problem
- widespread strategic failure to assess scale and cost requirements for high performance computing (HPC) needs
- intelligent total-cost-of-ownership analysis
  - cost per gigabyte and user allocated charges
  - platform diversity (supercomputers to smartphones)
  - physical versus virtual storage
  - impact of new technologies (solid-state drives, deduplication)
  - data retention policies
  - cloud services versus internal operations

Managing Big Data in Biomedicine:
Learning Precedents from Other Research Domains and Corporate Capabilities









Google™

amazon®

Microsoft®

FedEx® Express

ups®

DHL

To The Cloud:

GSA

TechAmerica FOUNDATION

Biocomputing in the Cloud

- shared ‘virtual commons’ for elastic (burst), scalable scientific computing
  - CaaS, SaaS, IaaS
- internet not designed for transfer of large scale datasets (terabyte +)
  - typical research lab. connectivity 1-10 gigabit/sec (0.25 to 1.25 gigabytes/sec) requires 1 week/1 day to transfer 100 gigabyte NGS file
  - more cost-effective to FedEx hard drive to CC provider
  - requirement to constantly move same datasets (and refresh them) and mirror them in multiple local storage systems

BGI Cloud on the Horizon

华大基因
BGI

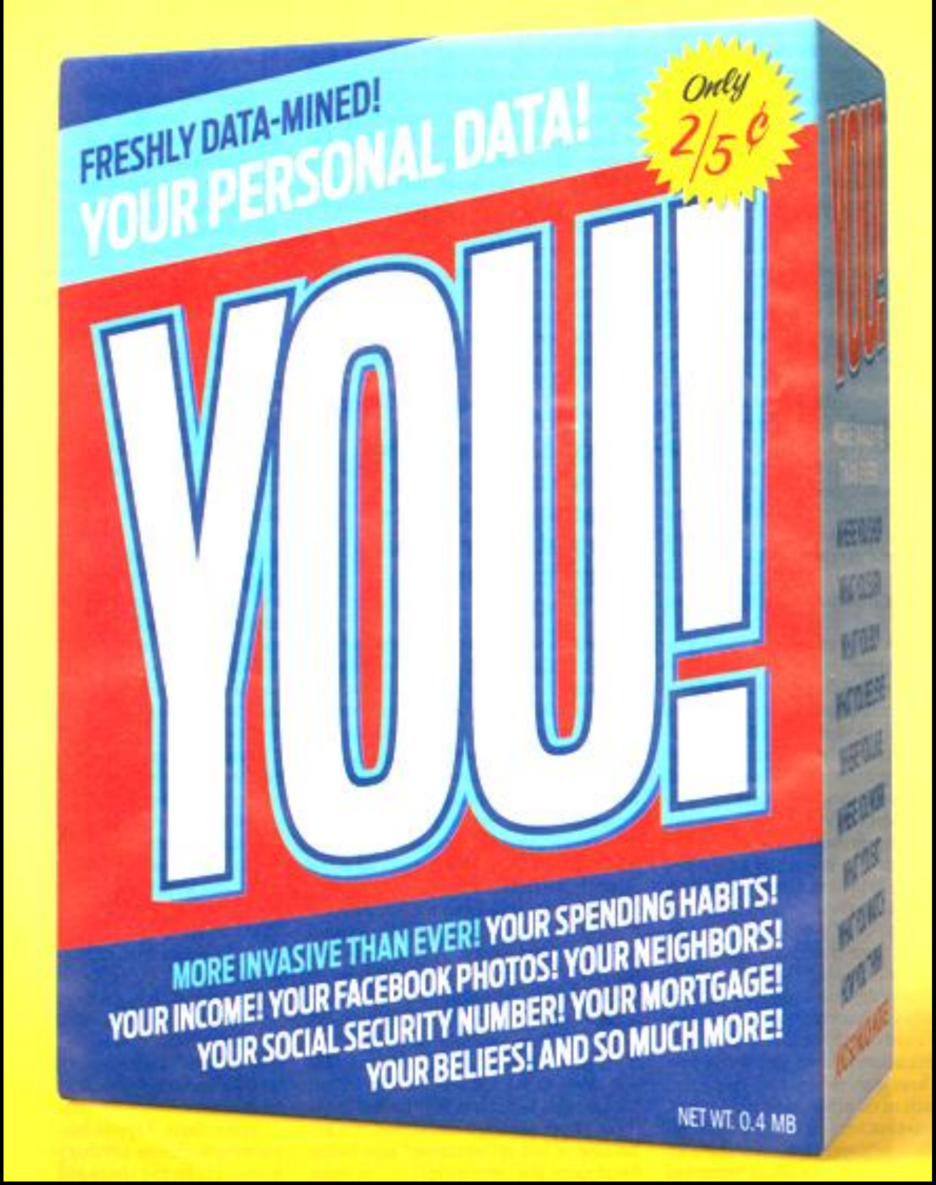
- “Amazon is slow”
  Evan Xiang, BGI Shenzhen
  Bio-IT World August 2011 p.8

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

(GIGA)^n
SCIENCE

- 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

Personal Privacy In An Era of Pervasive Computing and Multi-Parameter Profiling

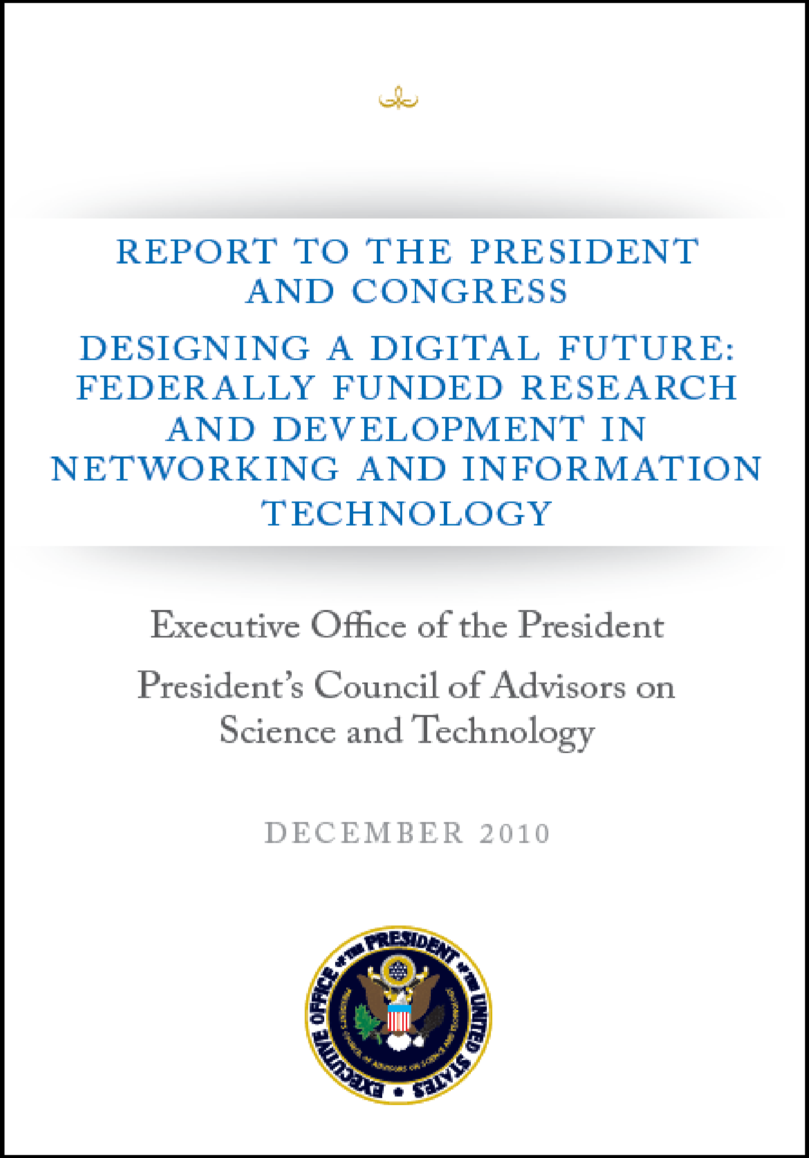

FRESHLY DATA-MINED!
YOUR PERSONAL DATA!
Only 2/5¢
YOU!
MORE INVASIVE THAN EVER! YOUR SPENDING HABITS! YOUR INCOME! YOUR FACEBOOK PHOTOS! YOUR NEIGHBORS! YOUR SOCIAL SECURITY NUMBER! YOUR MORTGAGE! YOUR BELIEFS! AND SO MUCH MORE!
NET WT. 0.4 MB
YOU!

Time
March 21, 2011

Homomorphic Encryption: Creating a Secure Cloud?

- commercial objective to build secure virtual vault for medical data
- mathematical proof of feasibility by IBM in 2009 (Craig Gentry)
- cloud-based tools for statistical analysis of encrypted data without decryption
- results remain fully encrypted for interpretation only by data owner’s key(s)
- any escape/leak/hack would only be in encrypted form
- next phase to increase speed and use in search and indexing functions

Development of Vanguard Capabilities in ACKM: A Fundamental Requirement for Sustained Competitiveness



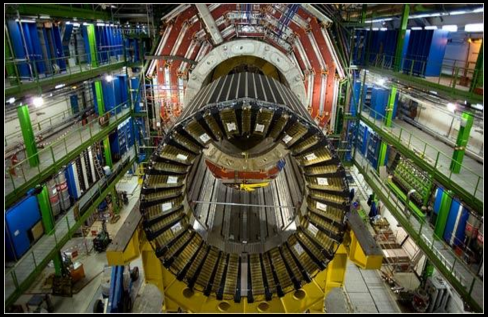
REPORT TO THE PRESIDENT AND CONGRESS

DESIGNING A DIGITAL FUTURE: FEDERALLY FUNDED RESEARCH AND DEVELOPMENT IN NETWORKING AND INFORMATION TECHNOLOGY

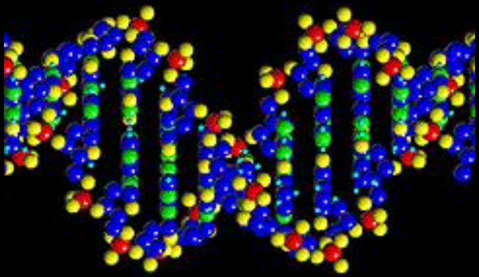
Executive Office of the President

President's Council of Advisors on Science and Technology

DECEMBER 2010



"If the scientific community can justify
billions of dollars, 100MW of power
and thousands of staff in order
to fire tiny particles that most people
have never heard of around a big ring of magnets
for a fairly narrow science purpose
that most people will never understand…..



…..then how come we can't make
the case for facilities needing half the resources
that can do wonders for a whole range
of science problems and industrial applications?"

Andrew Jones
Vice-President, Numerical Algorithms Group
HPC Wire 29 August 2011

Major New Initiatives in Supercomputing for Analysis of Research and Clinical Molecular Profiling Data

Chan Soon-Shiong
Institute for Advanced Health (CSS Institute)
and
National Coalition
for Health Integration

NantWorks

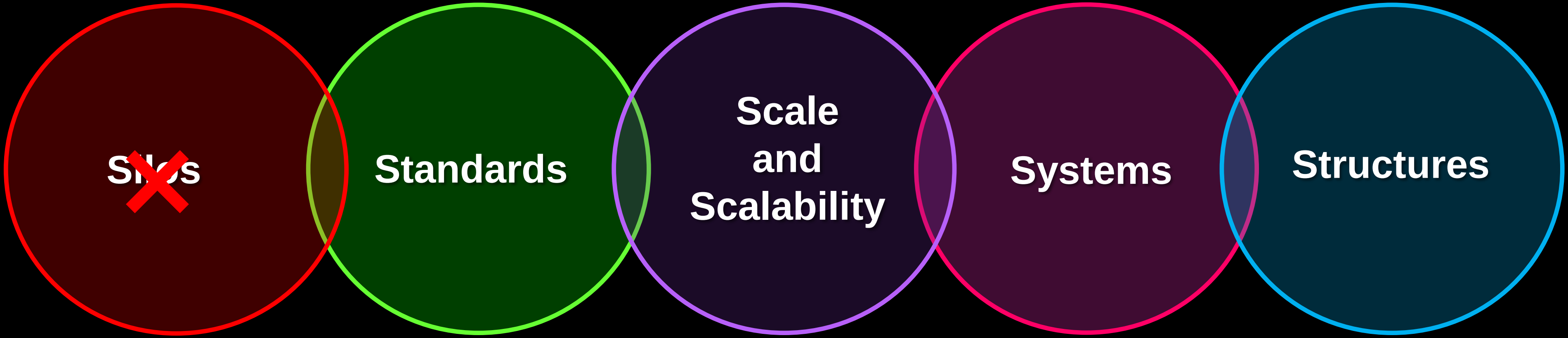
NLR National LambdaRail

ASU

MAYO CLINIC

Healthcare Transformation institute

The Principal Forces Shaping Progress in Biomedical R&D and Clinical Practice in An Era of Data-Intensive Methods



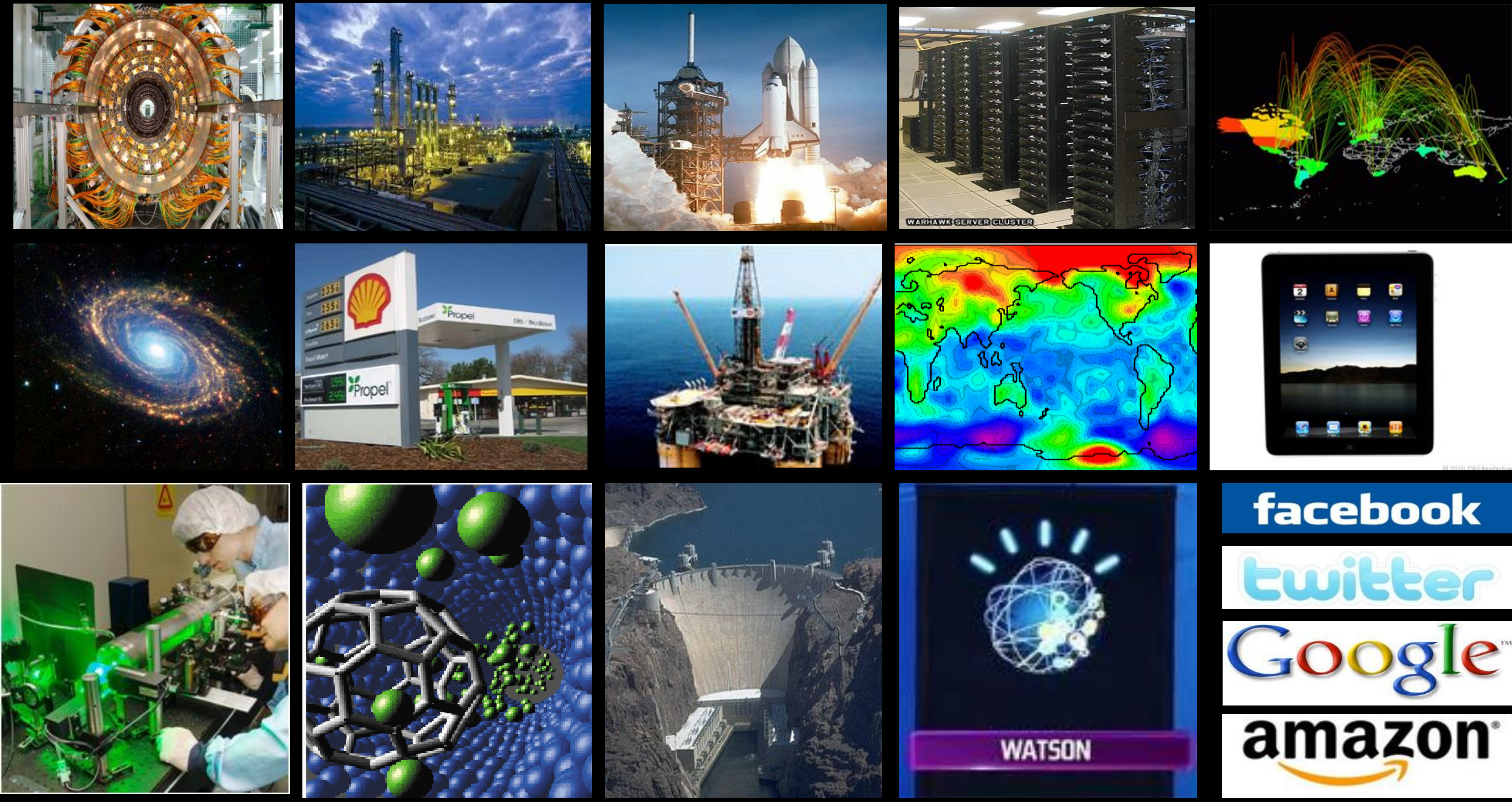
- new complexities
- new competencies
- new computation-intensive requirements
- new consumer and communication services
- new competitors
- new cross-sector alliances and business models
- new cultures and organizational structures

The Sociology of Integration of Computational Science as a Core Component of Biomedical R&D

- bridging three cultures
  - biomedical specialities, software engineering and scientific computing
- new ‘hybrid’ competencies/specialities
- building sufficient expertise (individuals/communities)
  - training, funding, incentives, rewards
- designing workflows and interfaces for e.science and federated virtual research environments
- increasing dependence/contribution on open-source datasets
- mapping data provenance in large multi-source datasets
- life in ‘the perpetual beta’

Technological Innovation in Different Domains:
A Study in Cultural, Methodological and Organizational Contrasts

Physics	Chemistry	Engineering	Computing	Telecommunications and Digital Media



Technological Innovation in Different Domains:
A Study in Cultural, Methodological and Organizational Contrasts

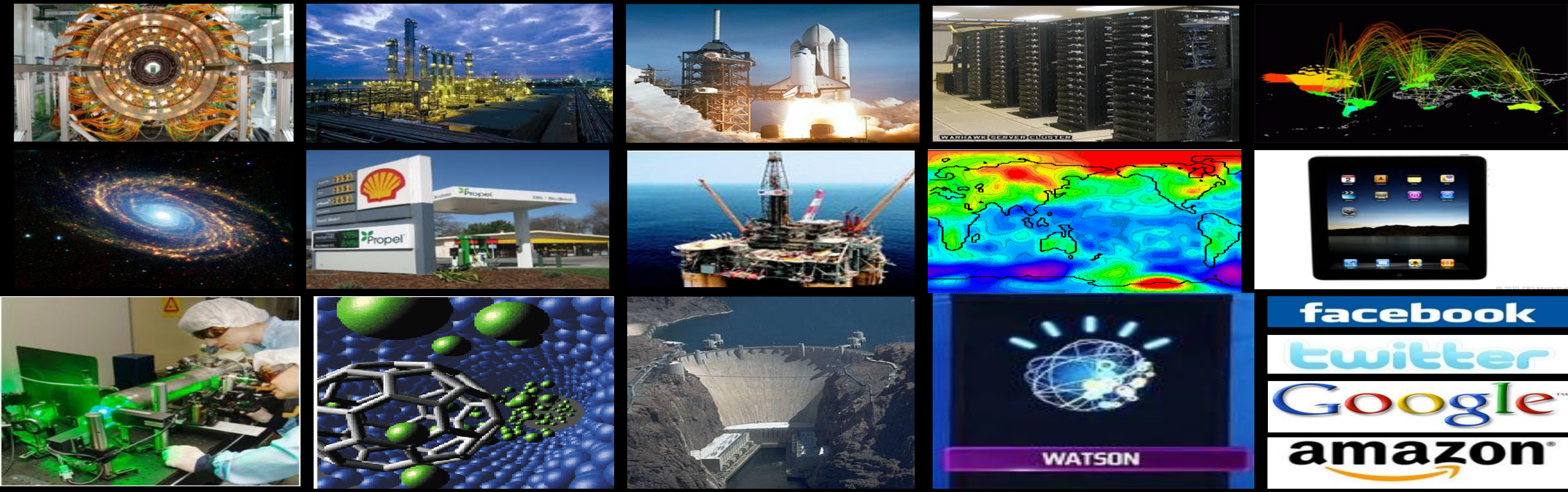
Physics

Chemistry

Engineering

Computing

Telecommunications and Digital Media



increasingly mature knowledge of underlying design laws and principles

mechanism-based, standardized methods and highly predictable outcomes

problem-oriented and focus on robust solutions

organizational/institutional structures address systems scale and complexity

Biomedical R&D and Clinical Medicine Are Methodological and Cultural Outliers in the Science and Technology Universe

Performance Parameters to Archive High Levels of Predictable Product Performance	Current Predominant Culture/Methods in Biomedicine
● mature knowledge of underlying design laws and principles	● largely phenomenological experiments/clinical interventions (‘black box’)
● mechanism-based, standardized methods and predictable outcomes	● largely descriptive data, poor standardization, replication and low predictive power
● problem-oriented and focus on solutions using integrated cross-disciplinary teams	● hypothesis-based, organizational silos and inefficient transfer processes
● organizational/institutional frameworks address systems scale and complexity	● ‘cottage industry’ era and slow adoption of systems-based approaches and requisite scale

Reset

An Unavoidable (But Essential) Transition to Data-and Computation-Intensive Methods

Strategic Needs

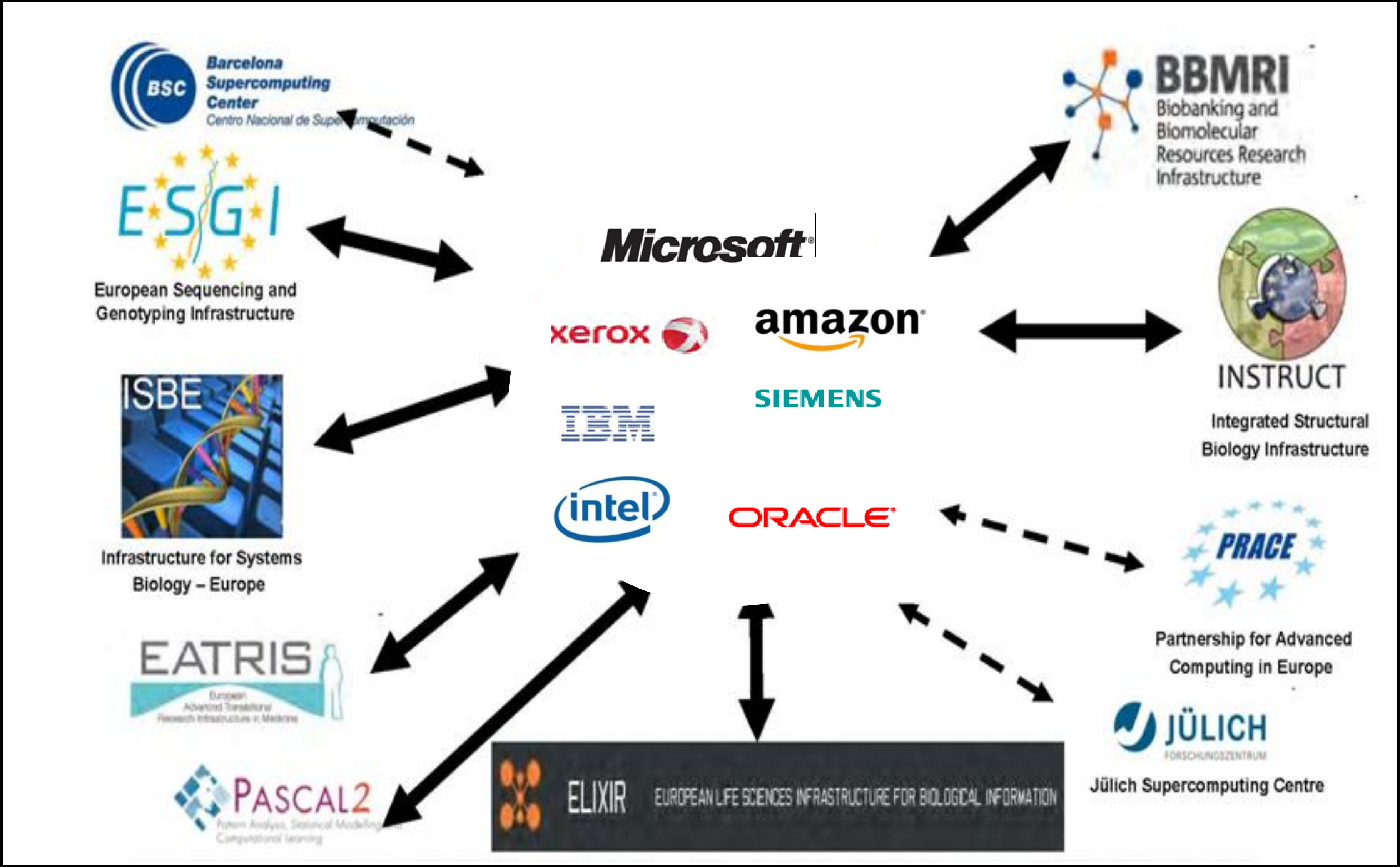
- systems-based approaches to define physiology and pathology in terms of molecular information networks
  - multiscale and hierarchical (cells to person)
  - spatiotemporal breadth (psec to lifespans)
- comprehensive knowledge of the topologies, dynamics and (dys)regulation of molecular networks to increase the predictability and productivity of all aspects of biomedicine
  - reduced failure rate of Dx/Rx assets in clinical trials
  - improved clinical decisions and health outcomes
  - risk mitigation and sustainable health

Changing the Sociology and Organization of Biomedical R&D

- ‘flying blind’: current ‘voids’ in understanding biological network dynamics and disease-associated perturbations
  - limited prediction of system behavior
  - unacceptable high failure rates in clinical trials

- pre-competitive 3P consortia to define ‘rules’ for biological network behavior
  - ‘connectome dynamics’
  - ID optimum loci for Rx
  - multiplex biomarkers for Dx
  - patient stratification
    - enrichment/ adaptive trials
    - rational Rx selection and improved outcomes
    - new analytical services and data aggregators

IT Future of Medicine



Barcelona Supercomputing Center

Centro Nacional de Supercomputación

BSC

E S G I

European Sequencing and Genotyping Infrastructure

ISBE

Infrastructure for Systems Biology – Europe

EATRIS

European Advanced Translational Research Infrastructure in Medicine

PASCAL2

Pattern Analysis, Statistical Modelling and Computational Learning

Microsoft

xerox

amazon

SIEMENS

IBM

intel

ORACLE

ELIXIR

EUROPEAN LIFE SCIENCES INFRASTRUCTURE FOR BIOLOGICAL INFORMATION

BBMRI

Biobanking and Biomolecular Resources Research Infrastructure

INSTRUCT

Integrated Structural Biology Infrastructure

PRACE

Partnership for Advanced Computing in Europe

JÜLICH

FORSCHUNGSZENTRUM

Jülich Supercomputing Centre

The Rise of Data-Intensive Biomedicine: Disruptive Change and New Value Drivers for Improved R&D Productivity and Healthcare Quality

Productivity

- biomedical R&D and healthcare delivery are under siege
  - conceptually limited, operationally inefficient, wasteful and economically unsustainable
- radical (disruptive) changes are required across the entire continuum from discovery to product development to patient care
- mapping disease-induced perturbations in biological systems and networks as a unifying theme for improved R&D strategies, rational clinical decisions and improved health outcomes
- realization of this objective will require new technical capabilities and organizational models to assemble and analyze biomedical data on an unprecedented scale

The Complexity Challenge

- new capabilities for engaging escalating complexity
  - new public funding priorities for large scale, team-based research
  - new education and training curricula
  - new 3P consortia to solve current major knowledge ‘voids’ in network dynamics that result in high product failure rates in clinical development, poor clinical decisions and less than optimal outcomes
- new opportunities to exploit technology convergence and data-centric platforms to build new business models and alliance networks

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