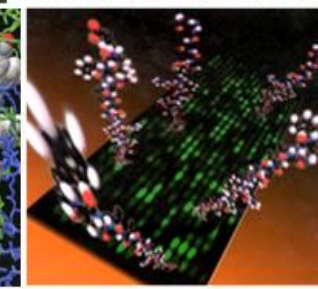
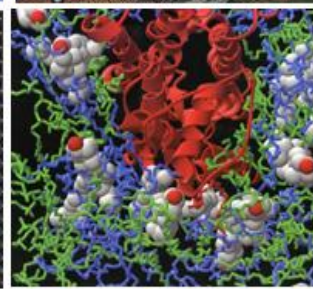
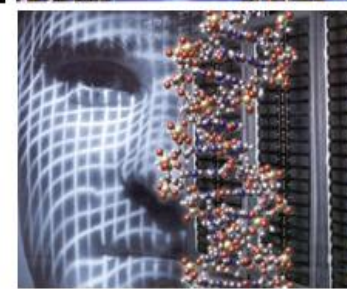
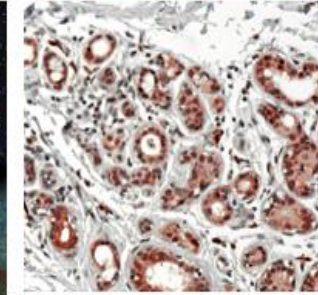
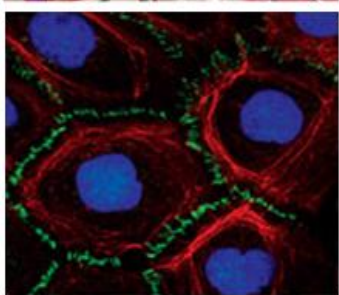
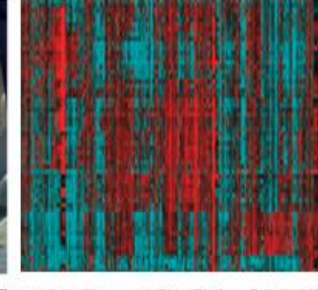
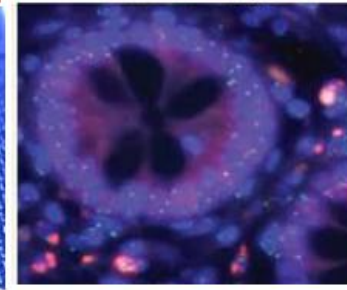
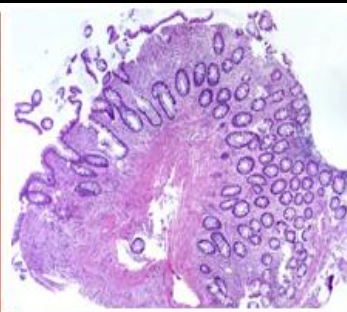
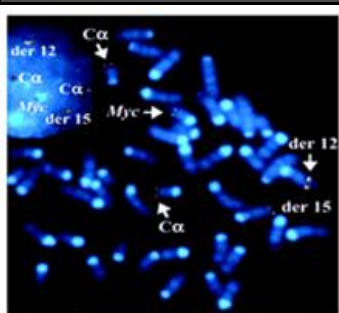


**“It from Bits”:
Managing Massive Data as a Critical Challenge
For Biomedical R&D and Healthcare Delivery**

**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**

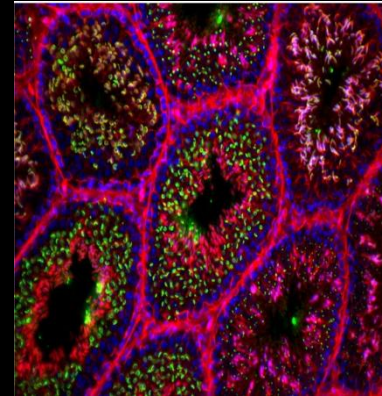
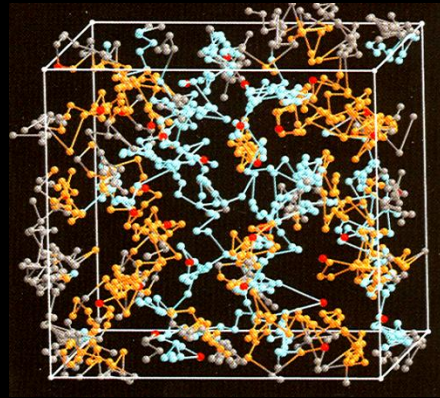
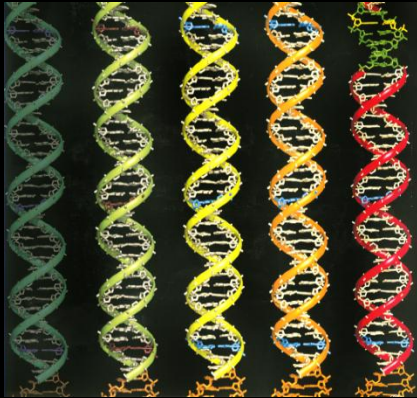
**Burrill Biotech Meeting, 24th Annual CEO Meeting
Laguna Beach, CA • Oct. 17-18, 2011**


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



“It from Bits”

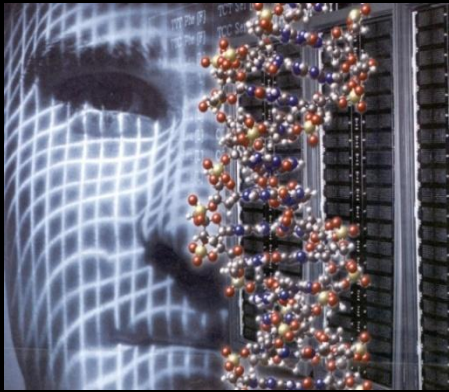
The Rise of Digital Biology



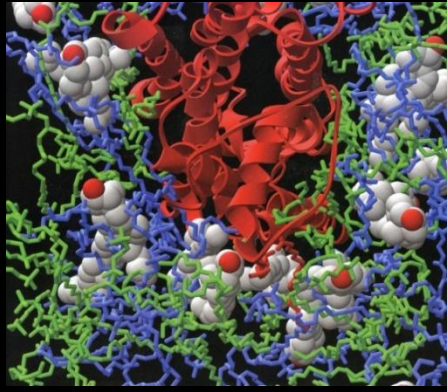
- understanding how encoded genome information creates complex multiscale biological systems (“It”)
 - defining health and disease in terms of patterns of information flow in biological information networks (“bits”)
- 
- the transformation of biomedical R&D, clinical medicine and healthcare delivery into increasingly quantitative, mechanistic data-intensive disciplines

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

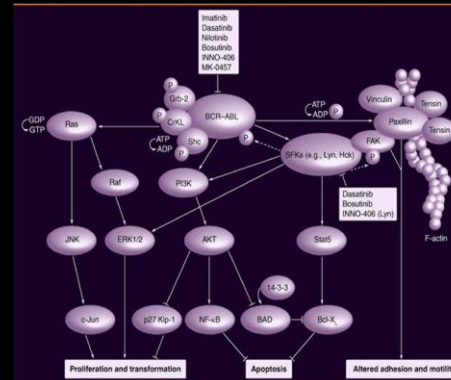
Genomics



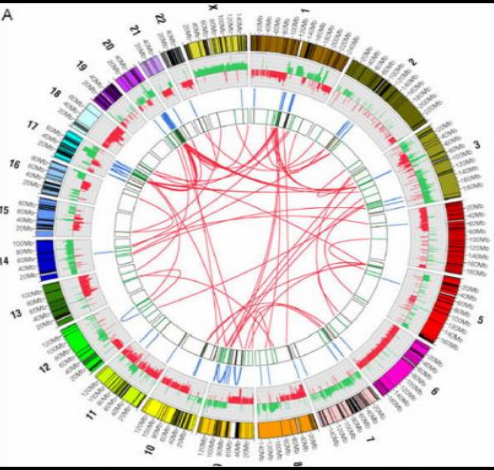
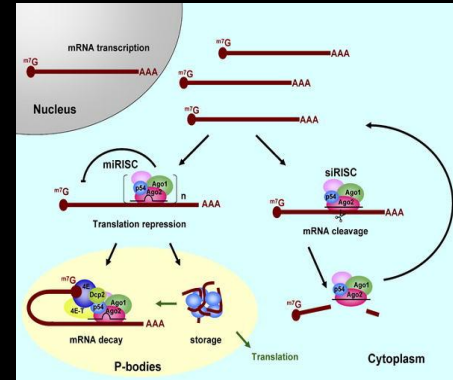
Proteomics



Molecular Pathways and Networks



Network Regulatory Mechanisms



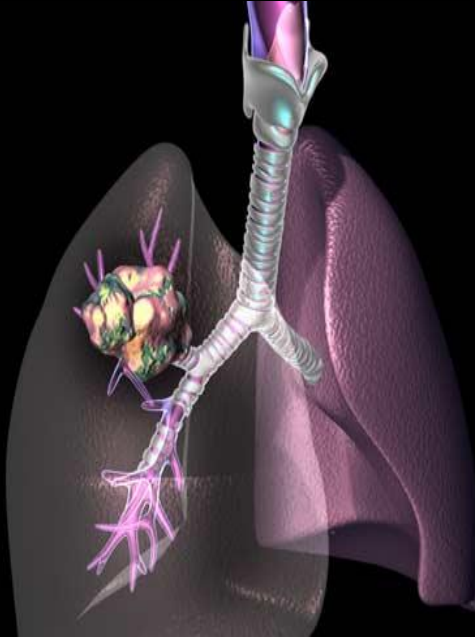
**ID of Causal Relationships Between
Network Perturbations and Disease**

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

Mapping the Molecular Signatures of Disease, Disease Subtyping and Targeted Therapy: The Right Rx for the Right Disease (Subtype)



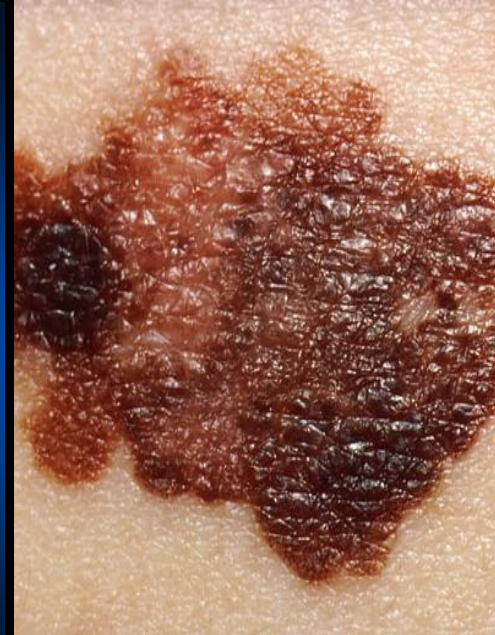
**Her-2+
(Herceptin)**



**EML4-ALK
(Xalkori)**



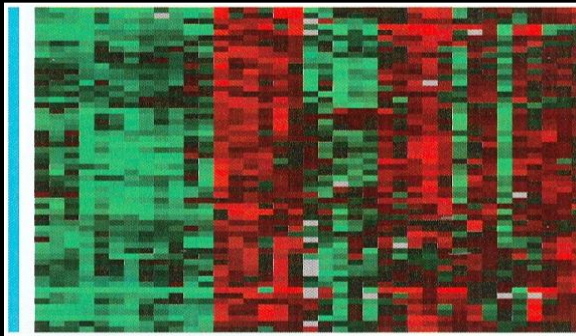
**KRAS
(Erbix)
(Vectibix)**



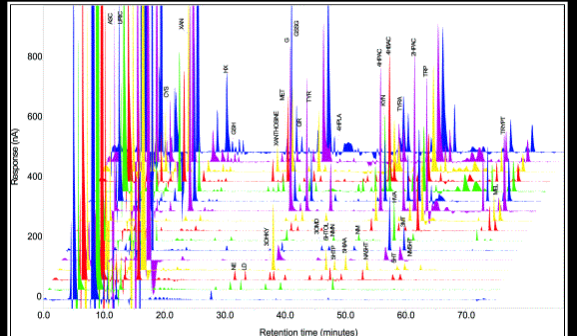
**BRAF-V600
(Yervoy)
(Zelboraf)**

Massively Parallel Biosignature Profiling

genomics



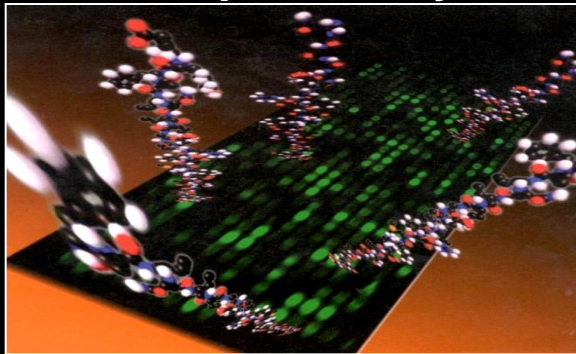
proteomics



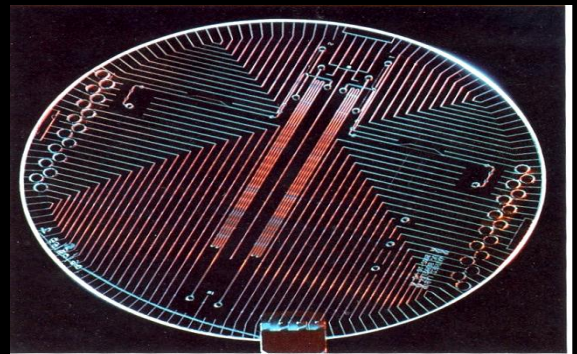
immunosignatures



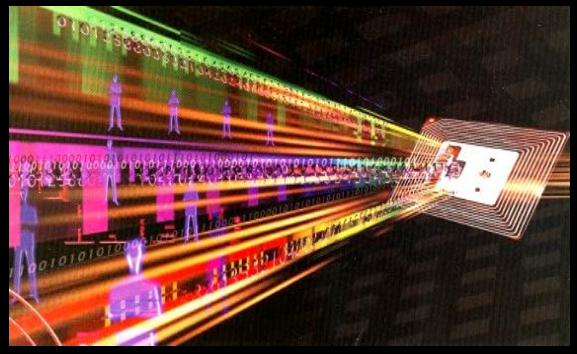
**automated,
high throughput
multiplex assays**



**novel test formats
and devices (POC)**



complex signal deconvolution



Large Datasets, Standardized Ontologies and New Computational Analytics

Sensing the First Ripples in the Approaching Massive Data Wave

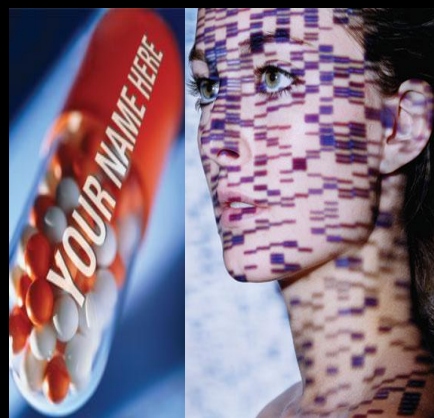
- **uncompressed human genome = 3 gigabytes**
- **current WGS sequencing efforts now generating terabytes (TB) of data from still modest number of genomes/species (and at low depth coverage)**
- **1000 Genomes Project raw data posted after 6 months was 2x Genbank deposits over previous 30 years**
- **imminent expansion of massive WGS/enriched genome datasets**
 - **ENCODE, modENCODE, TCGA, ICGSC, Human Microbiome Project**
 - **comparable datasets emerging in agriculture and metagenomics**

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

Patient Stratification For Clinical Trials



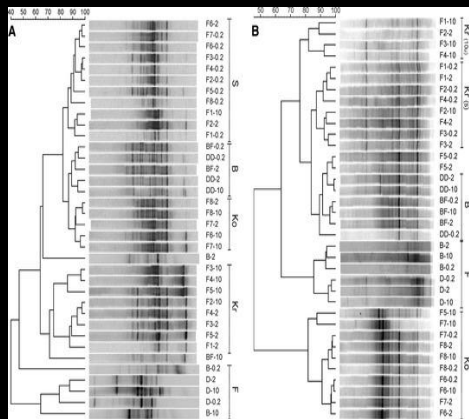
Pharmacogenomics



m.Health



Monitoring Networks



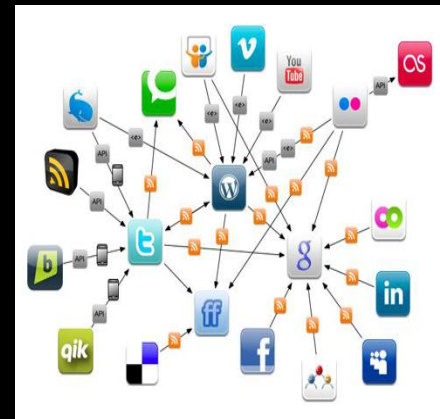
Microbial Diagnostics



Biosurveillance and Public Health



Health IT and EMRs



Social Media

Managing Massive Data

Standards for Data Reporting and Database Design

**Interoperability of Databases
Across The Continuum from Discovery to Patient Care**

New Analytics and High Performance Computing

Extracting Value from Data

- raw data is increasingly cheap
- ill-defined, non-replicable, non-standardized, statistically underpowered data is of little or no value
 - resources sink, diversion and duplication
- the file drawer problem
 - negative findings, often highly instructive, are rarely reported and journal disinterest
- maximizing value
 - from data to knowledge to relevance
 - validated, actionable, adopted and reimbursable!

Critical Challenges for Biomedical R&D

- acceleration of discovery phase knowledge without parallel gains in successful clinical translation and commercial ROI
- unacceptable high rates of failure of candidate Rx in clinical trials
- major knowledge gaps for rational discovery strategies to address late onset chronic diseases
 - diabetes, cancer and neurodegeneration
- regulatory and reimbursement uncertainties for molecular diagnostics (MDx) to subtype disease and drive rational selection of companion Rx
- cost control in healthcare and future pricing/adoption of products with uncertain/limited efficacy

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

- **mapping disease-induced perturbations in biological systems and networks as the intellectual foundation and critical success factors for enhanced R&D productivity, improved clinical decisions and to promote better health outcomes**
- **creation of this knowledge resource will require new technical capabilities and organizational models to assemble and analyze biomedical data on an unprecedented scale**

Profiling Platforms for Mapping Molecular Networks: The Accelerating 'Data Deluge'

Low Cost Exome- and/or Whole Genome Sequencing



ion torrent

illumina

454
SEQUENCING

Roche

IBM

Oxford
NANOPORE
Technologies

SEQUENOM



Complete
genomics

BioNanomatrix

PACIFIC
BIOSCIENCES

imagination at work

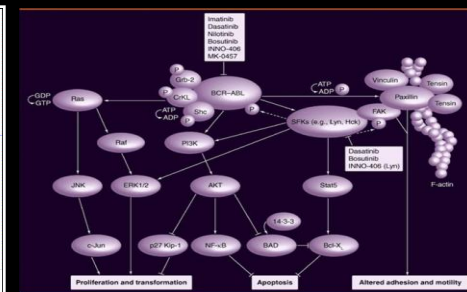
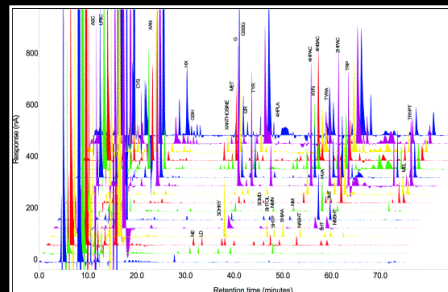
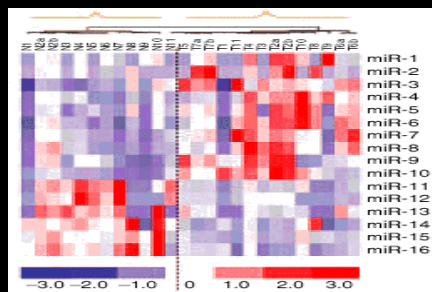
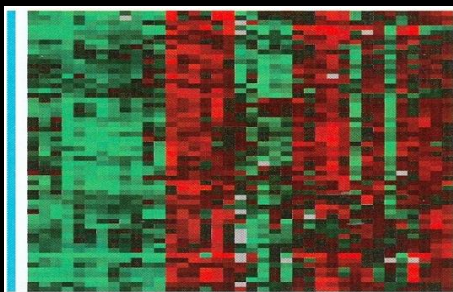
Helicos
BioSciences Corporation

Transcriptomics

miRNAs

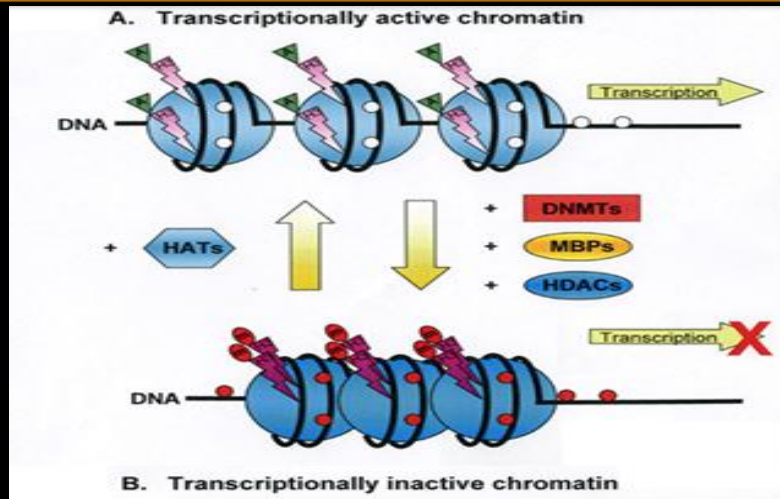
Proteomics

Protein Interaction Networks (PIN)



The Epigenome

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



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



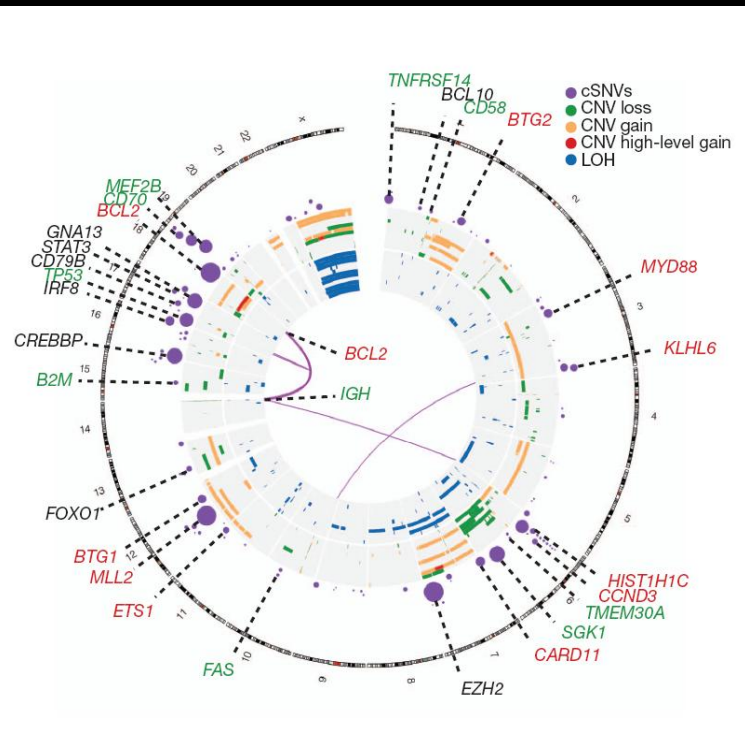
International Human Epigenome Consortium
• • • 1000 reference genomes by 2020



project blueprint

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

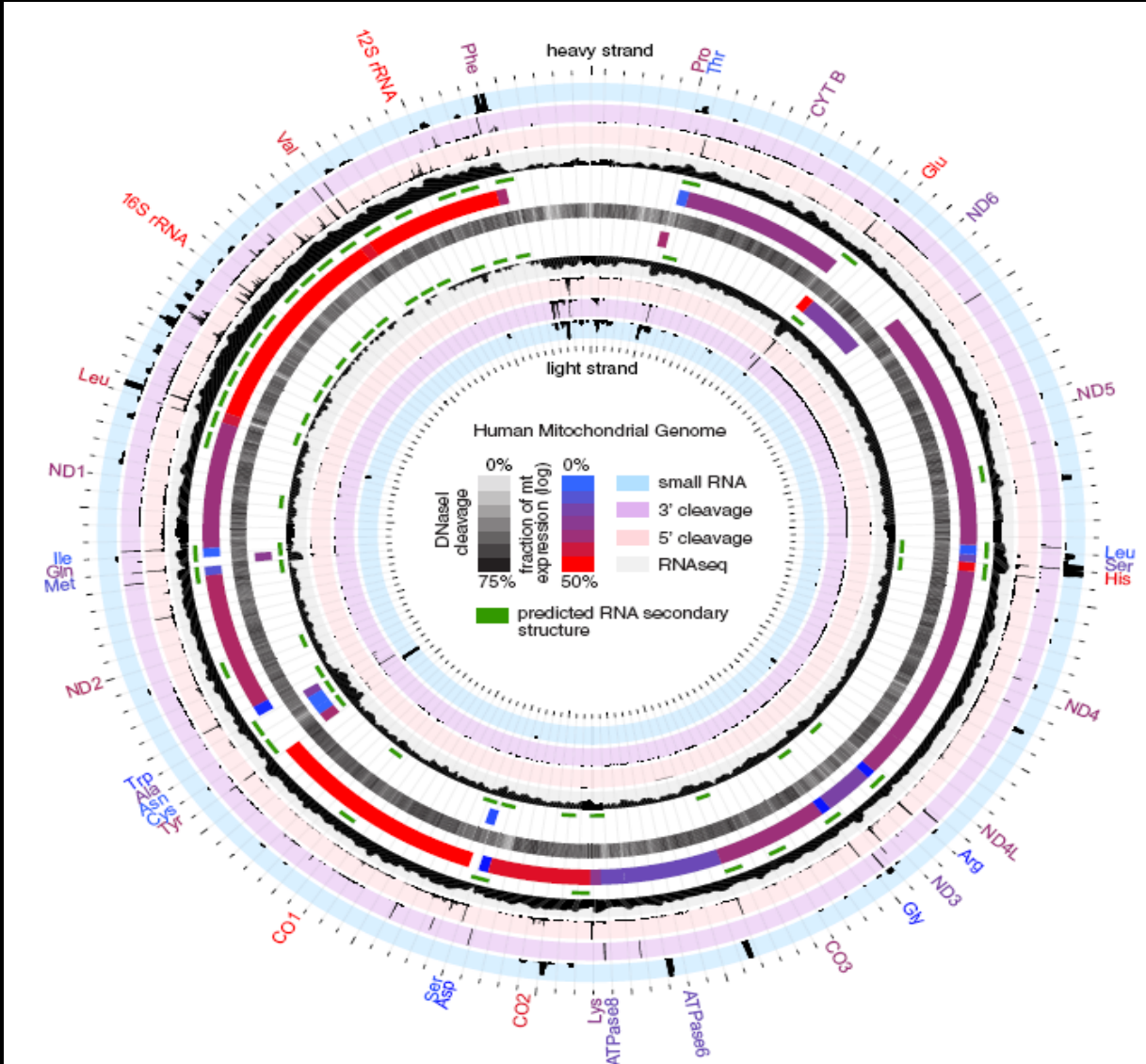
Cancer Epigenome



Nature (2011) 476, 298

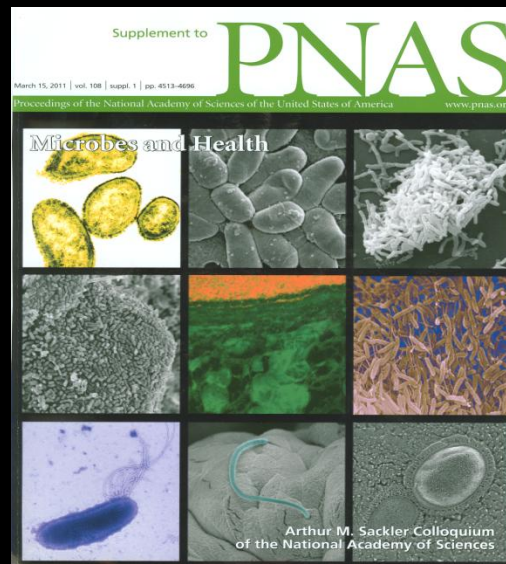
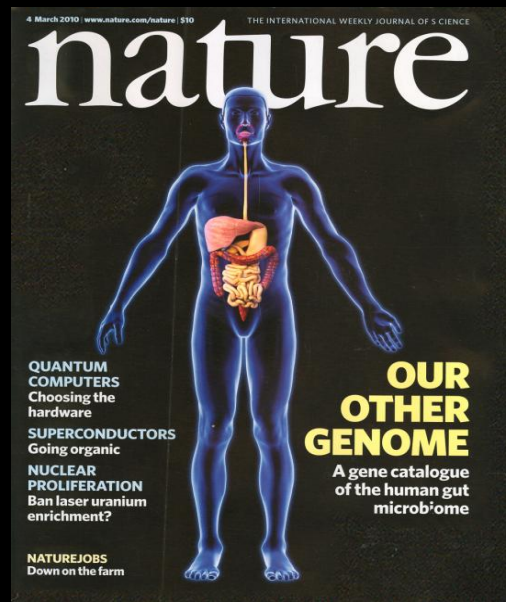
- increased methylation variation at CpG islands
 - increased variability in gene expression
- extensive mutations in genes encoding chromatin remodeling proteins and histone modifications
- hypomethylation of large genome blocks (upto several Mb) involving more than half the genome

The Human Mitochondrial Transcriptome



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

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



Diet-Derived miRNAs Can Survive Digestion and Modulate Host Gene Expression



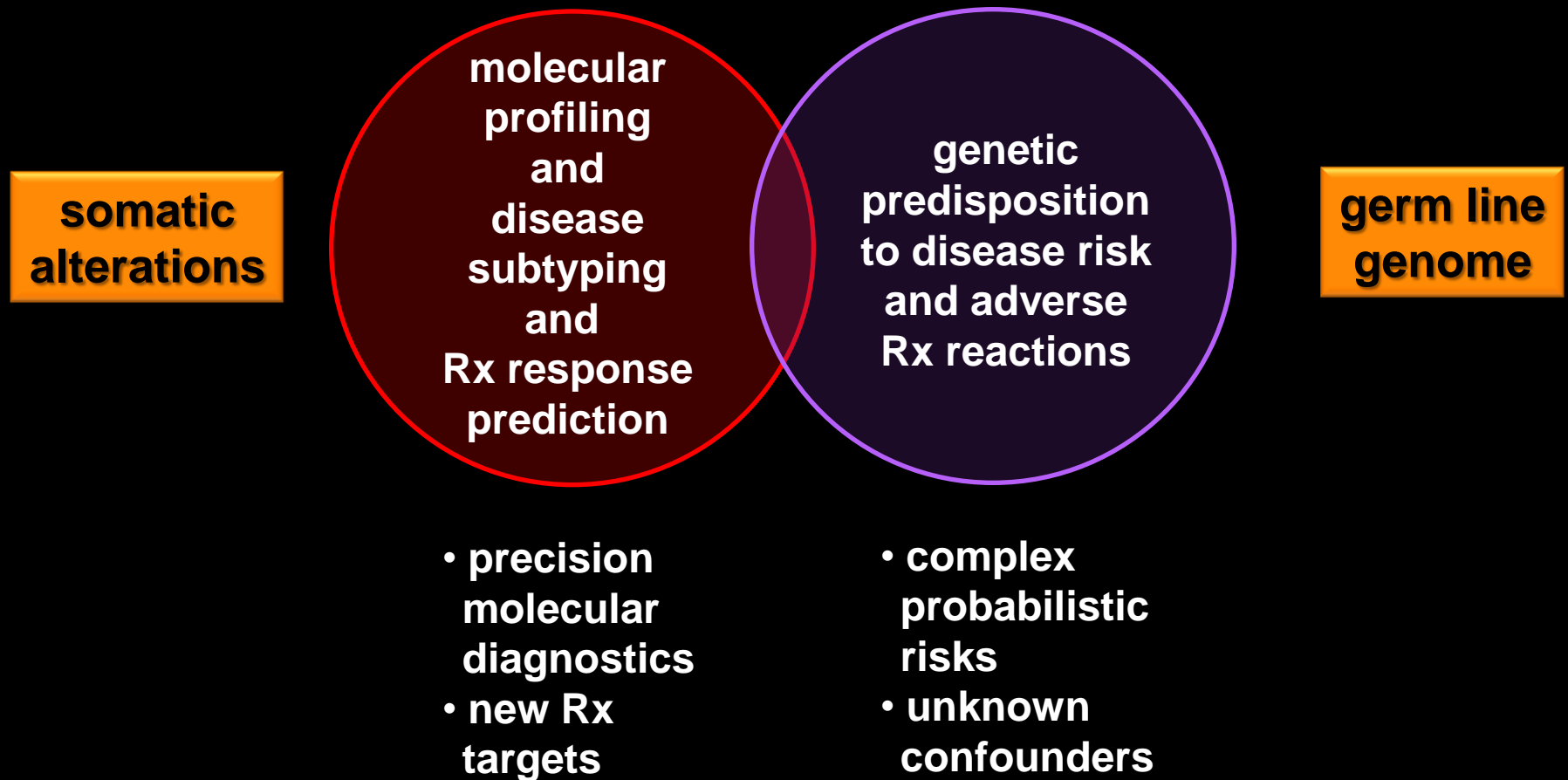
**Eating greens
Alters genes**

New Sci. (2011) 1 Oct., p. 10; C. Y. Zange et al. (2011) Cell Res. DOI: 10.1038/cr.2011.158

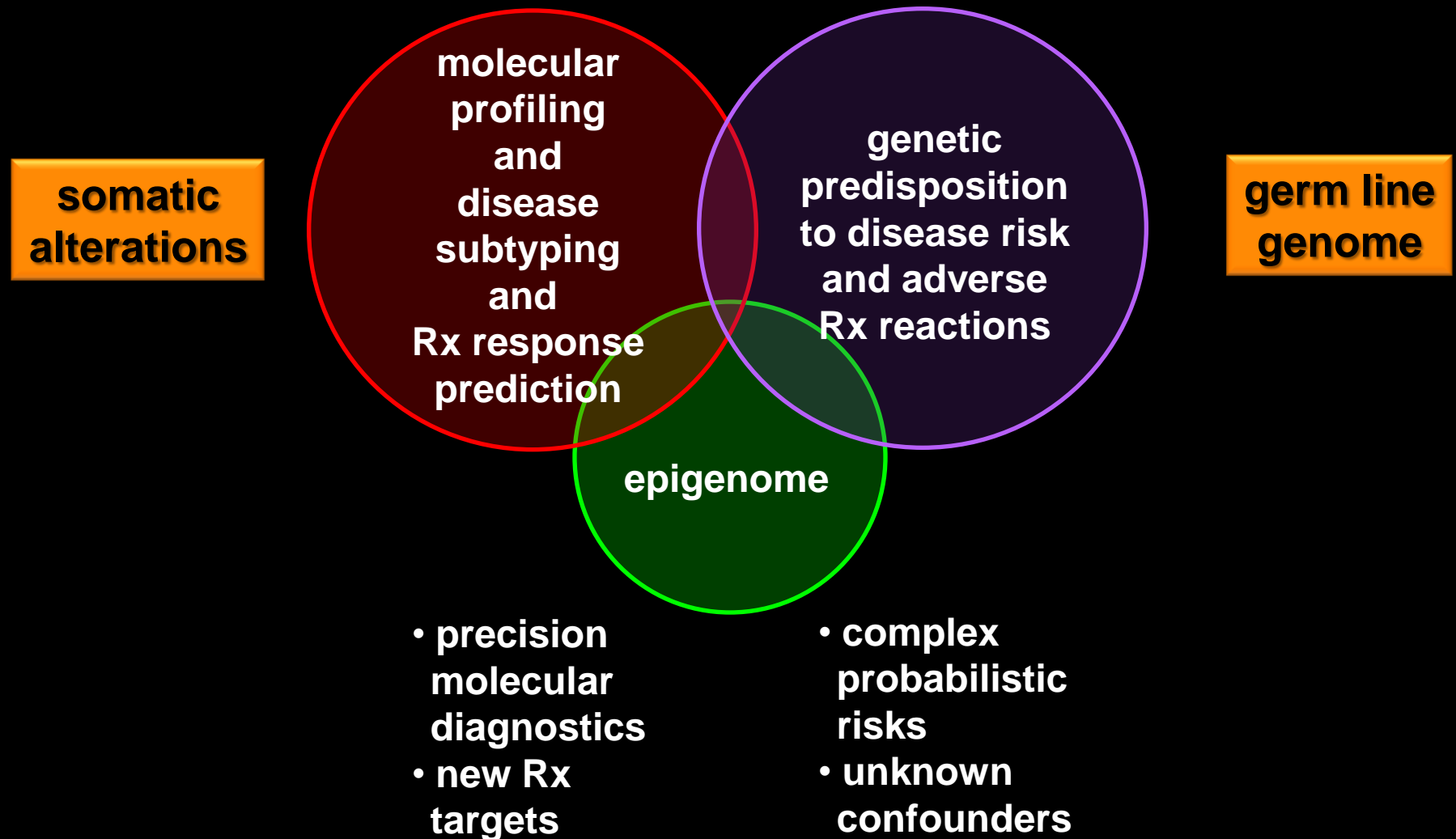
**Mapping the Human Variome:
Defining the Molecular Taxonomy of Individuality
and
Correlations With Phenotypic Traits
and Disease Processes**

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

Applications of Whole Genome and Transcriptome Sequencing



Applications of Whole Genome and Transcriptome Sequencing



**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

Current Challenges in Accurate Genome Annotation

- **ID of true variation versus machine artifacts from high error rate-and context-specific sequence errors**
- **mapping reads to reference genome(s)**
 - **misalignment of reads spanning indels**
 - **relevance of non-aligned reads**
- **persistent inaccuracies in per-base quality scores**
 - **differences with different sequencing technologies, machine cycles and sequence context**
- **need for platform agnostic consistency**
 - **sample preparation and enrichment**
 - **prerequisite for inter-operable data sourcing**

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) for clinical utility**
- **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 accessible/renewable**
 - **cancer**

The Adoption of Genome Sequencing in Clinical Diagnostics

- from research odysseys to routine clinical use (“just another lab value!”)
- early clinical applications
 - oncology
 - infectious diseases
 - hereditary cancers and individual risk assessment
 - rare diseases of suspected genetic etiology
 - HLA profiling for transplant matching
 - cardiomyopathies
 - X-linked intellectual disability
 - congenital muscular dystrophy
 - mitochondrial disorders



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

- **criteria to assess platform accuracy**
- **minimum sequencing depth for reliable clinical decisions**
- **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)**
- **RUO reagents**
- **sequencers as Class III devices?**

Next Comes The Hard Part!

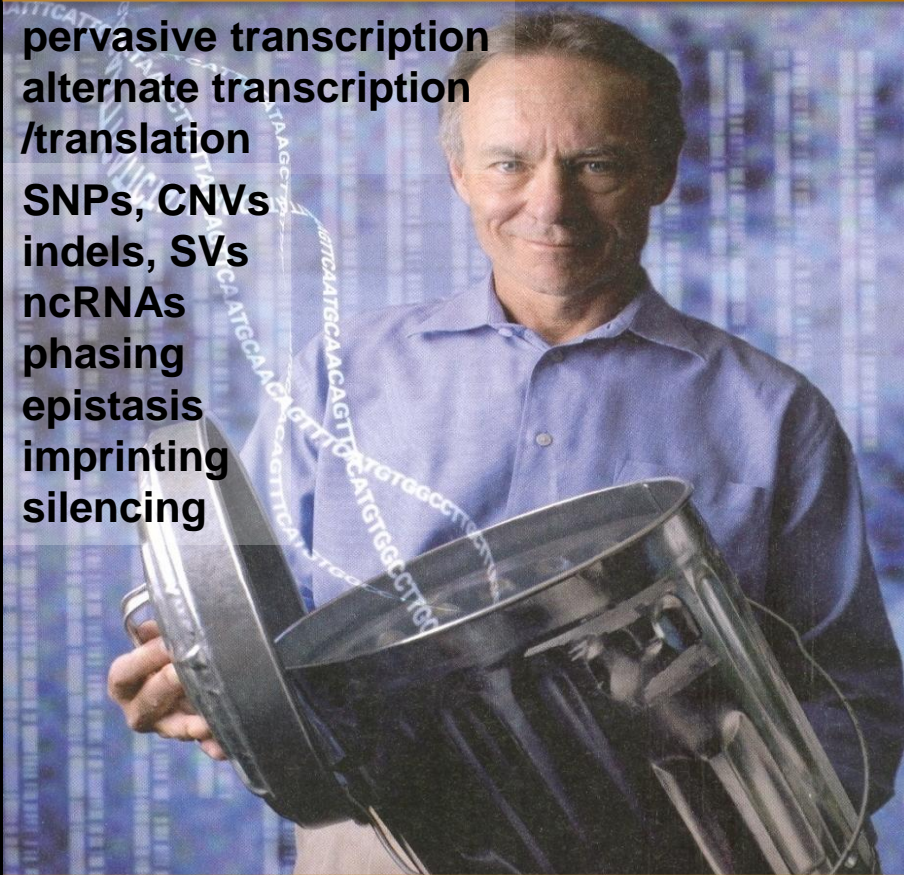
From Genotype to Phenotype

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

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

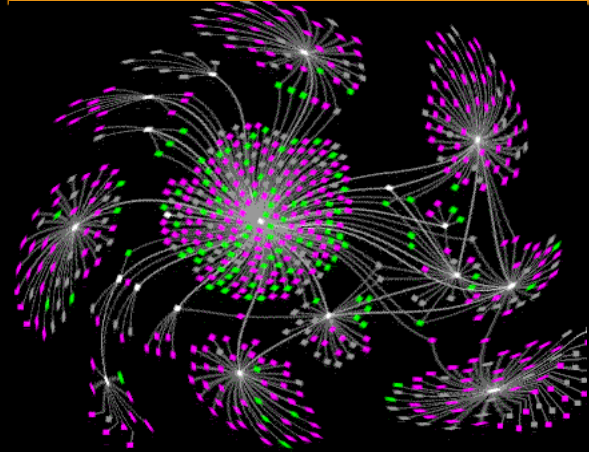
Junk No More!

pervasive transcription
alternate transcription
/translation
SNPs, CNVs
indels, SVs
ncRNAs
phasing
epistasis
imprinting
silencing



recognition of genome
organizational and regulatory
complexity

Cell-specific Molecular Interaction Networks



Disease Perturbations



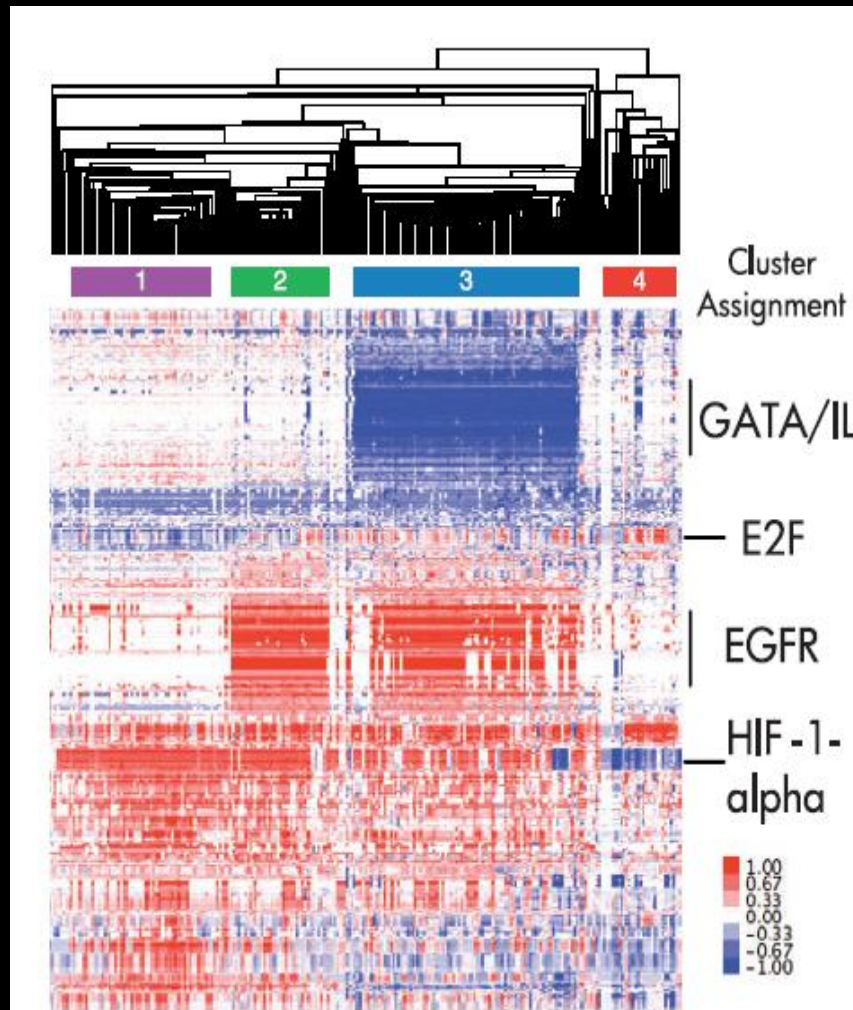
Design Rule #1 in Biology

- **knowledge of the genotype at a single locus or co-regulated sequences cannot predict the phenotype accurately**

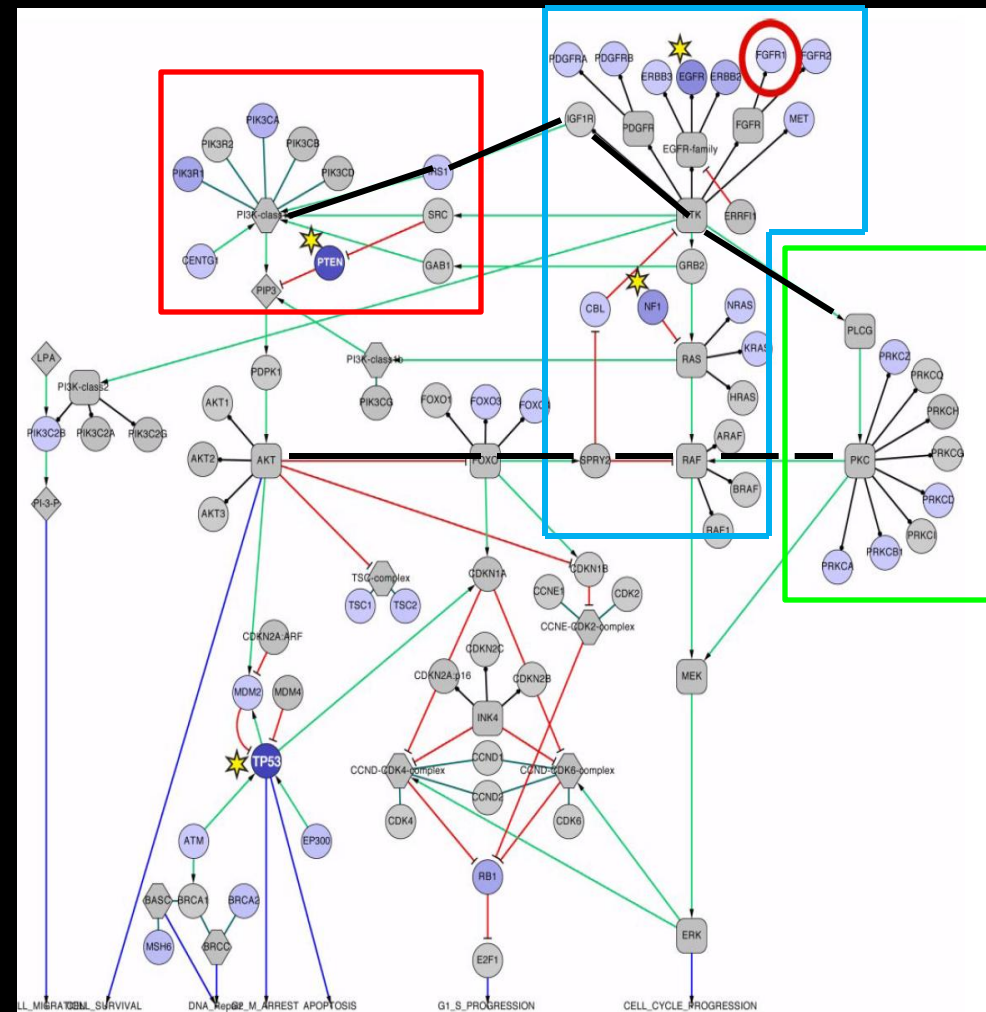
Design Rule #2 in Biology

- the expressed molecular network (“bits”) defines “it” (the phenotype)
- the phenotype and the underlying molecular networks are defined by context
 - cell-specific differences (differentiation)
 - gene-gene (epistasis) and gene-environment effects
 - epigenetic gene activation or silencing
- graded levels of network dysregulation in disease create a continuum of clinical phenotypes (minor to severe)

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



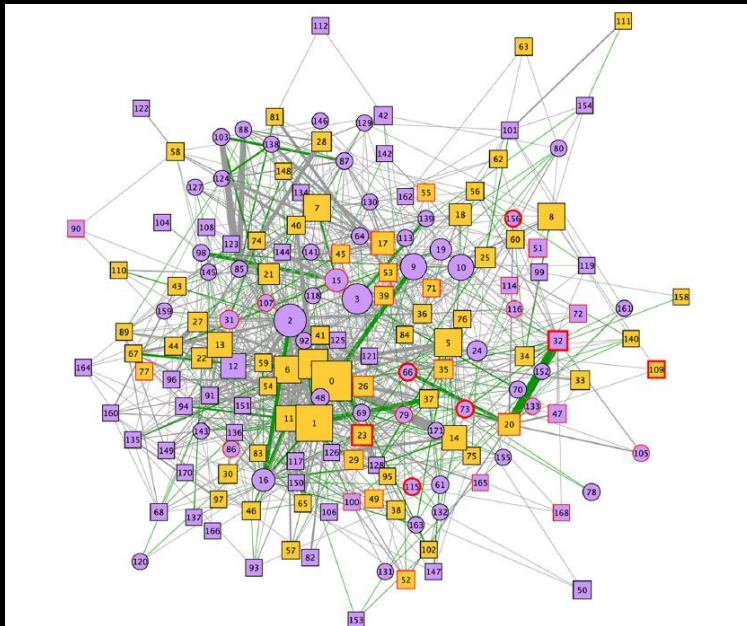
From: C. J. Vaske et al. (2011)
Bioinformatics 26, i237



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

Mapping of the 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!

Amyloid-Beta Protein as Therapeutic Target in Alzheimer's Disease

monoclonals

- gantenerumab (Roche)
- solanezumab (Lilly)
- poneznmab (Pfizer)
- bapineuzumab (J&J)
- MAB-5102A (AC Immune)

immune globulin

- octagam (Octapharma)

imaging agents

- amyvid (Lilly)
- ACC-001 (J&J/Pfizer)

vaccine

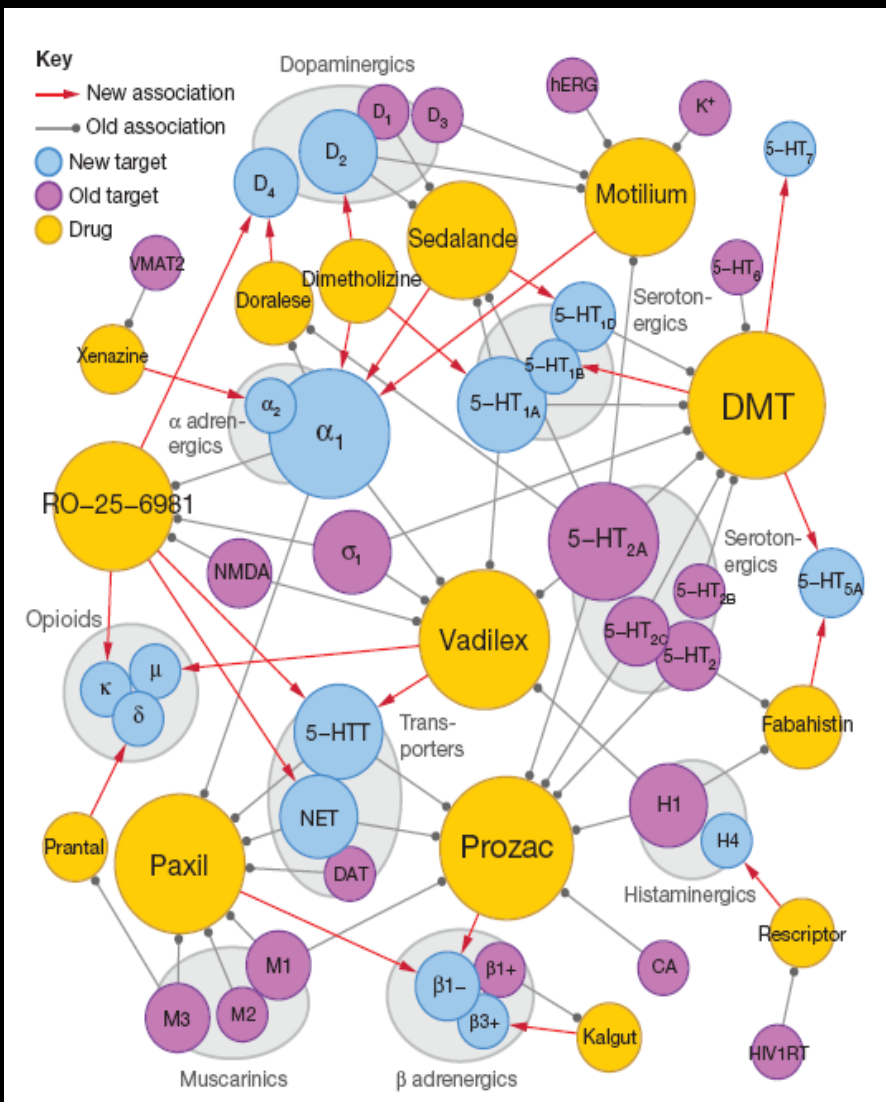
- UB 311 (United Biomedical)

small molecules

- metformin
- BMS-708113 (BMS)
- CAD106 (Novartis)
- PF 04995274 (Pfizer)
- RQ 09 (RaQualia)
- Anavex (2-73
- NP-61(Noscira)
- LY-2886721 (Lilly)
- E-2212 (Eisai)

Now Comes The Even Harder Task!:

Defining Network Pharmacology



- analysis of Rx action in context of network topologies and dynamics
- same drug: interaction with multiple targets
- same target: interaction with multiple drugs
- mapping structural chemotypes to specific pathways and subnetworks for targeted (poly)pharmacology

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”**
 - **adaptive capacity to use “by-pass” pathways to frustrate Rx effects**

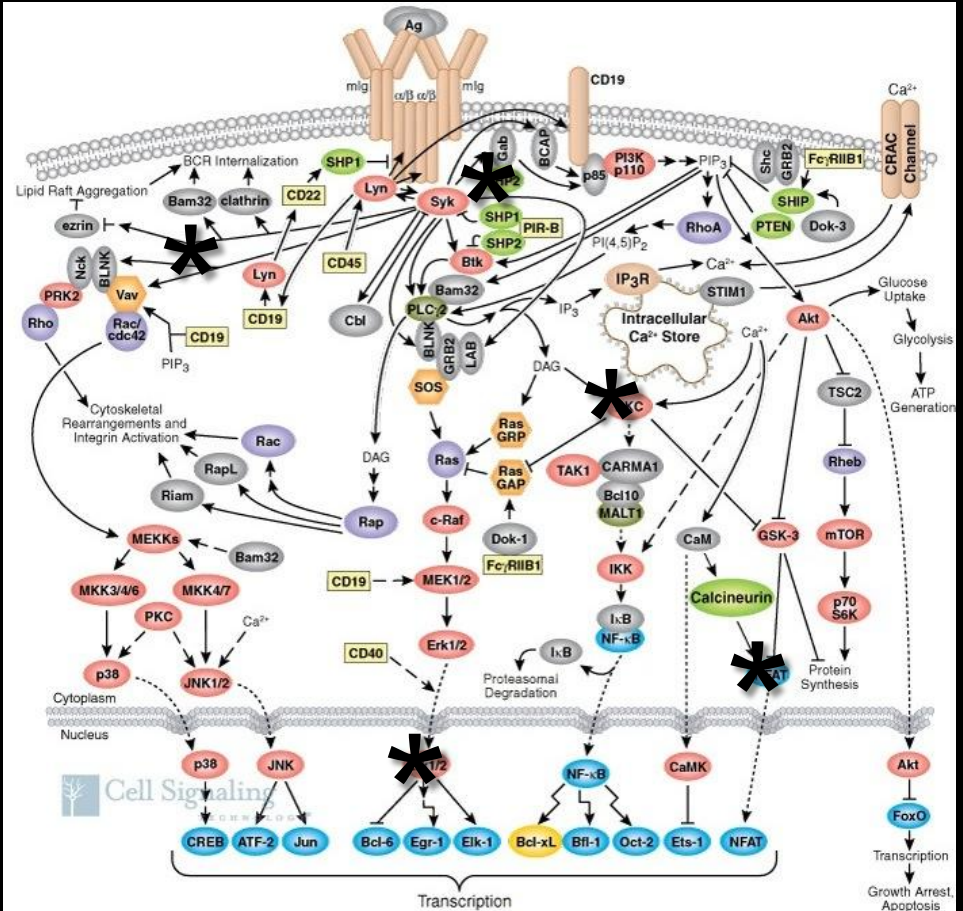
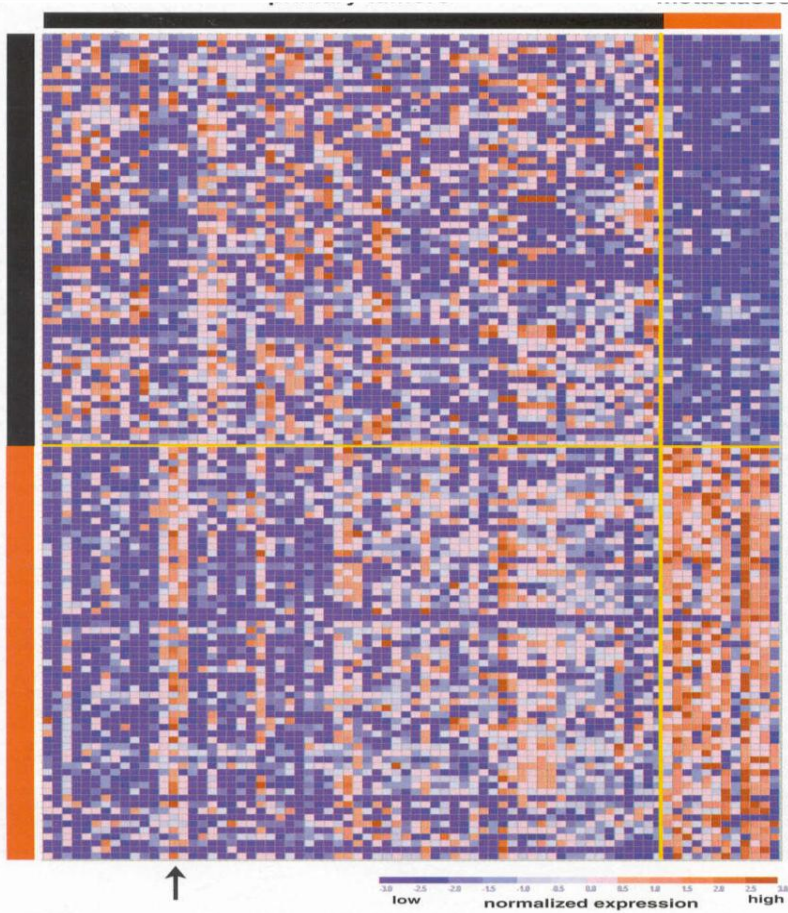
The Challenge of Mapping Network Structure and Dynamics for Improved Predictability in Selection of Diagnostic Biomarkers and Rx Targets

- **the current ‘black box’ gap in knowledge linking genotype to molecular networks (phenotype)**
- **still too often ‘flying blind’ in ID of causal vs. correlative changes in network dysregulation in disease**
 - **poor accuracy in ID of biomarkers and Rx targets**
 - **high failure rates in Rx clinical trials due to inability to define multi-target Rx actions needed to reverse network dysregulation and restore homeostasis**

Mapping Dysregulation of Biological Networks in 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**

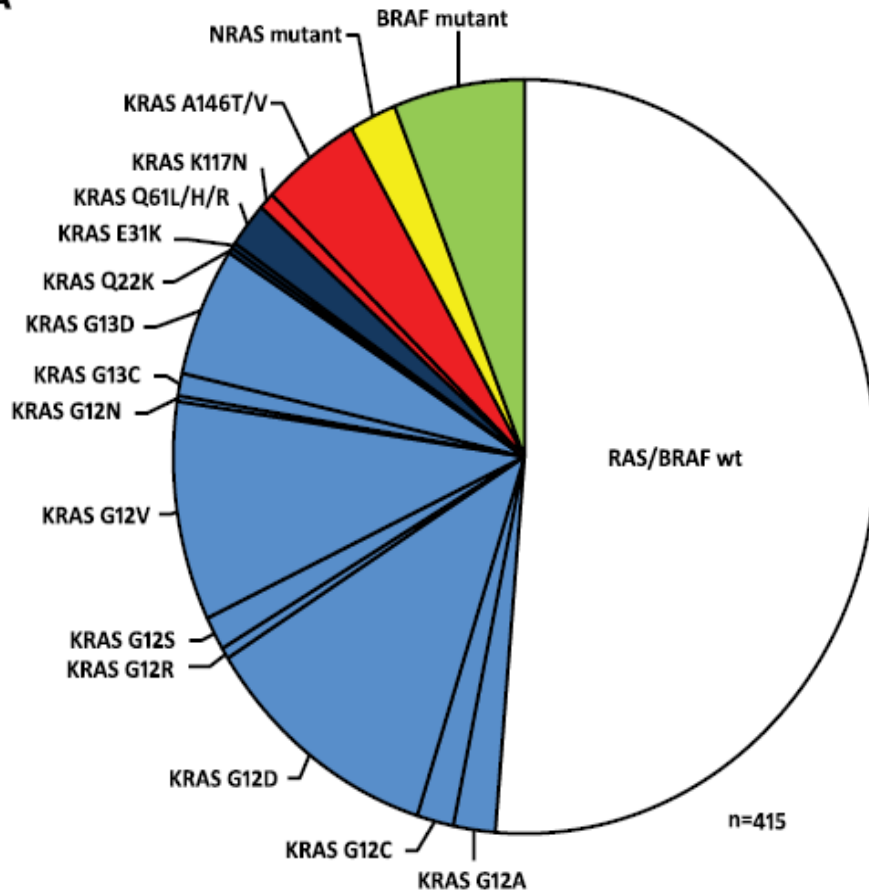


**From: N. Wagle
et al. (2011)
J. Clin. Oncol. 29, 3085**

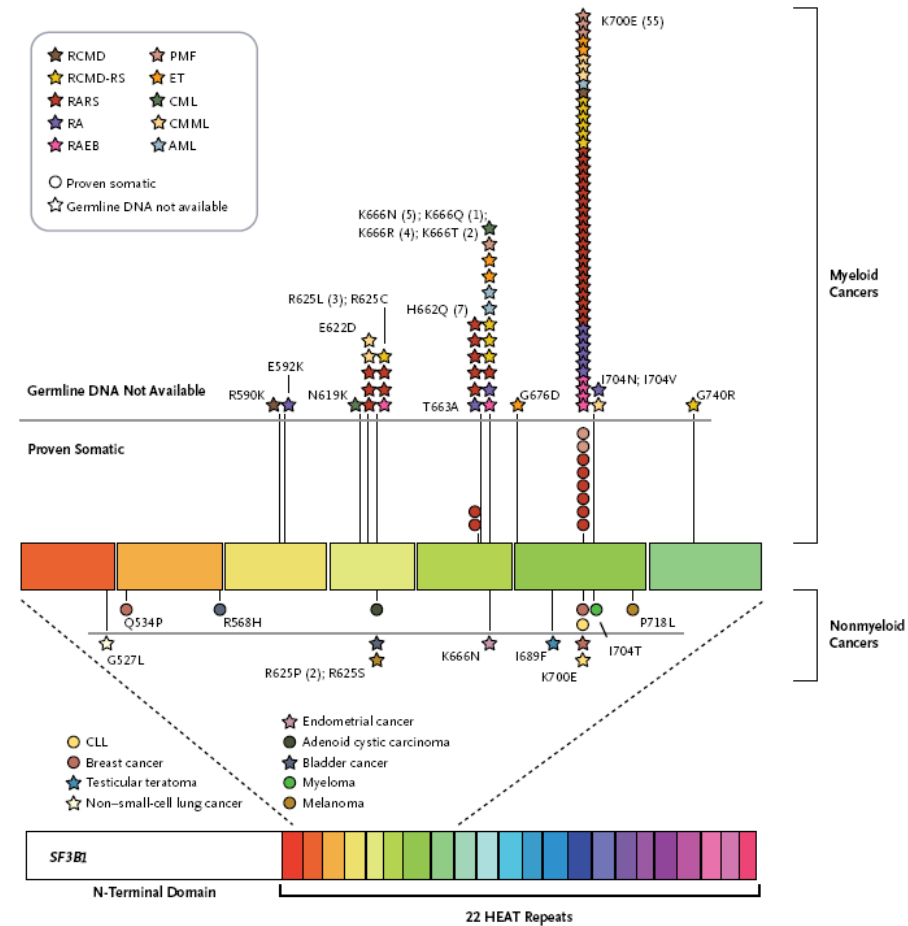
Sequence-Based Profiling of Diversity of Mutation Repertoire and Design of Rx-Intervention Strategies in Stratified Patient Cohorts

Colorectal Cancer

A



Myelodysplastic Disorders



Cancer Res (2011)

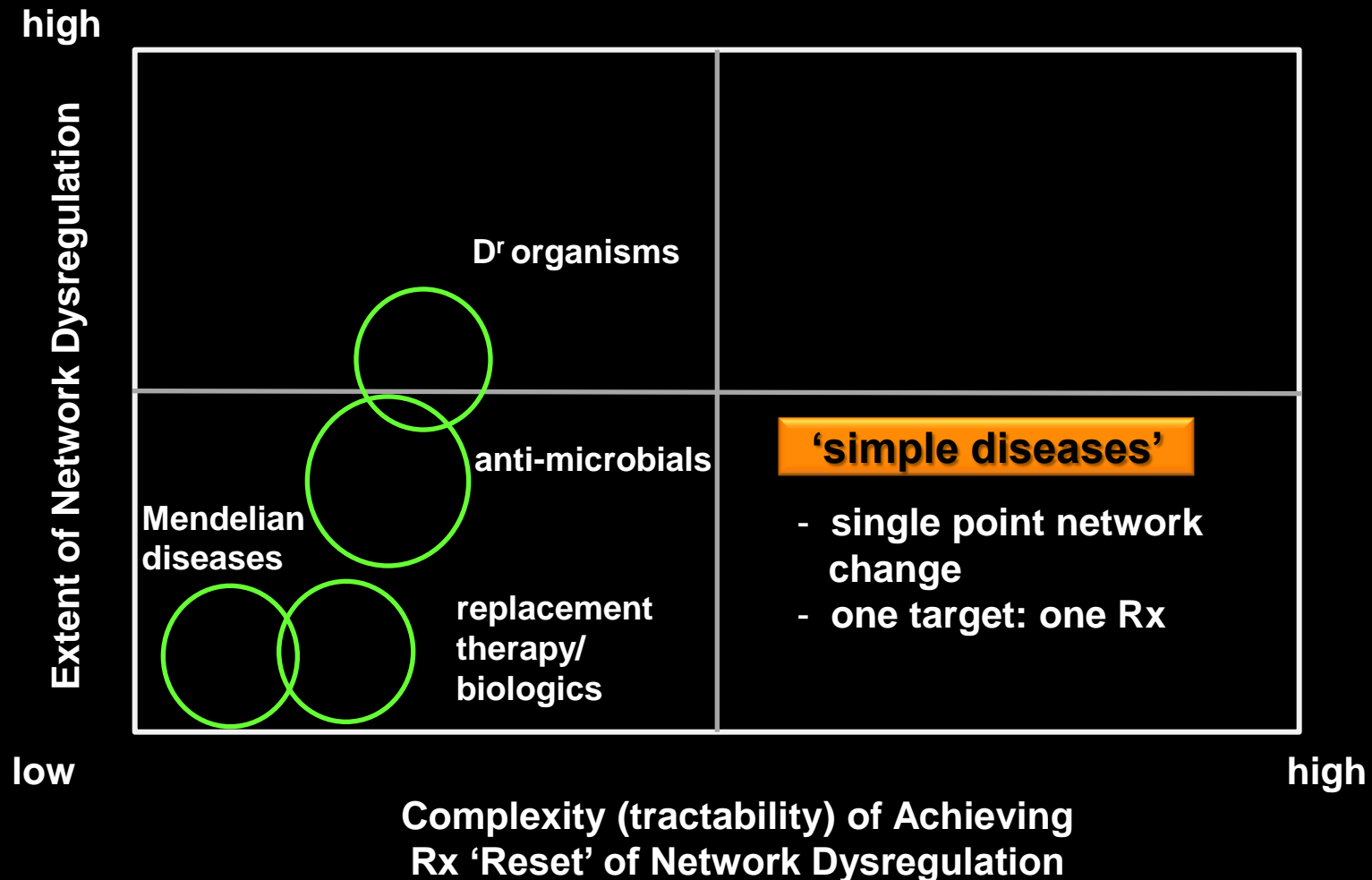
doi: 10.1158/0008-5472.CAN-10-0192

NEJM 2011 365, 1384

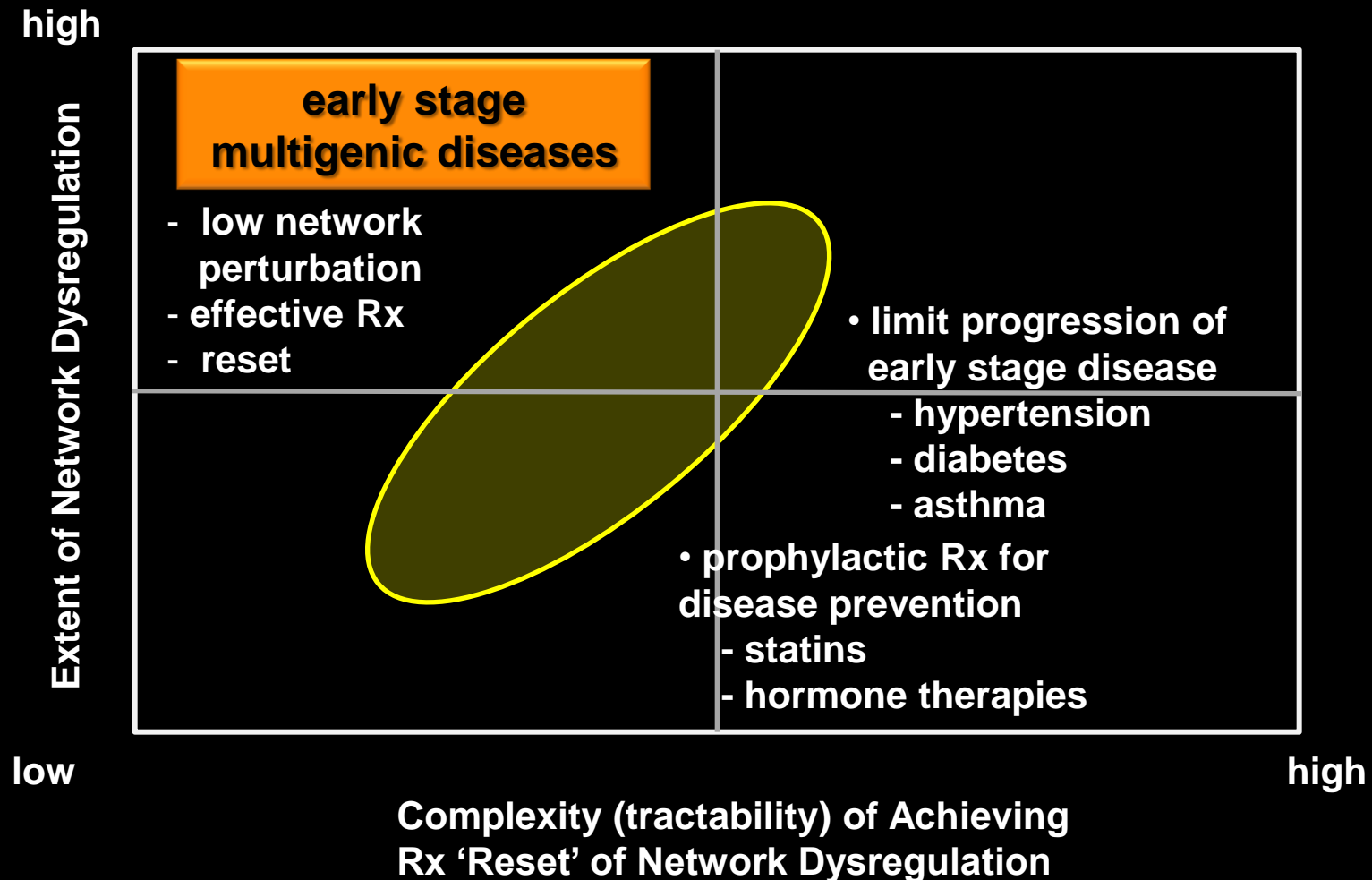
Network Pharmacology

- **elucidation of definitive network ‘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?)/regulatory challenges**
 - **controlled multi-target promiscuity in a single moiety**
- **does chronic progression in complex, multigenic diseases amplify module/subnetwork dysregulation?**
 - **greater complexity of multi-target homeostatic Rx ‘reset’**
 - **role of Rx in driving selection of variants with Rx-resistance (‘escape’) pathways (e.g. oncology)**

The Relationship Between The Scale of Network Dysregulation in Disease and Technical Complexity (POS) for Successful Rx-Driven 'Homeostatic Reset'



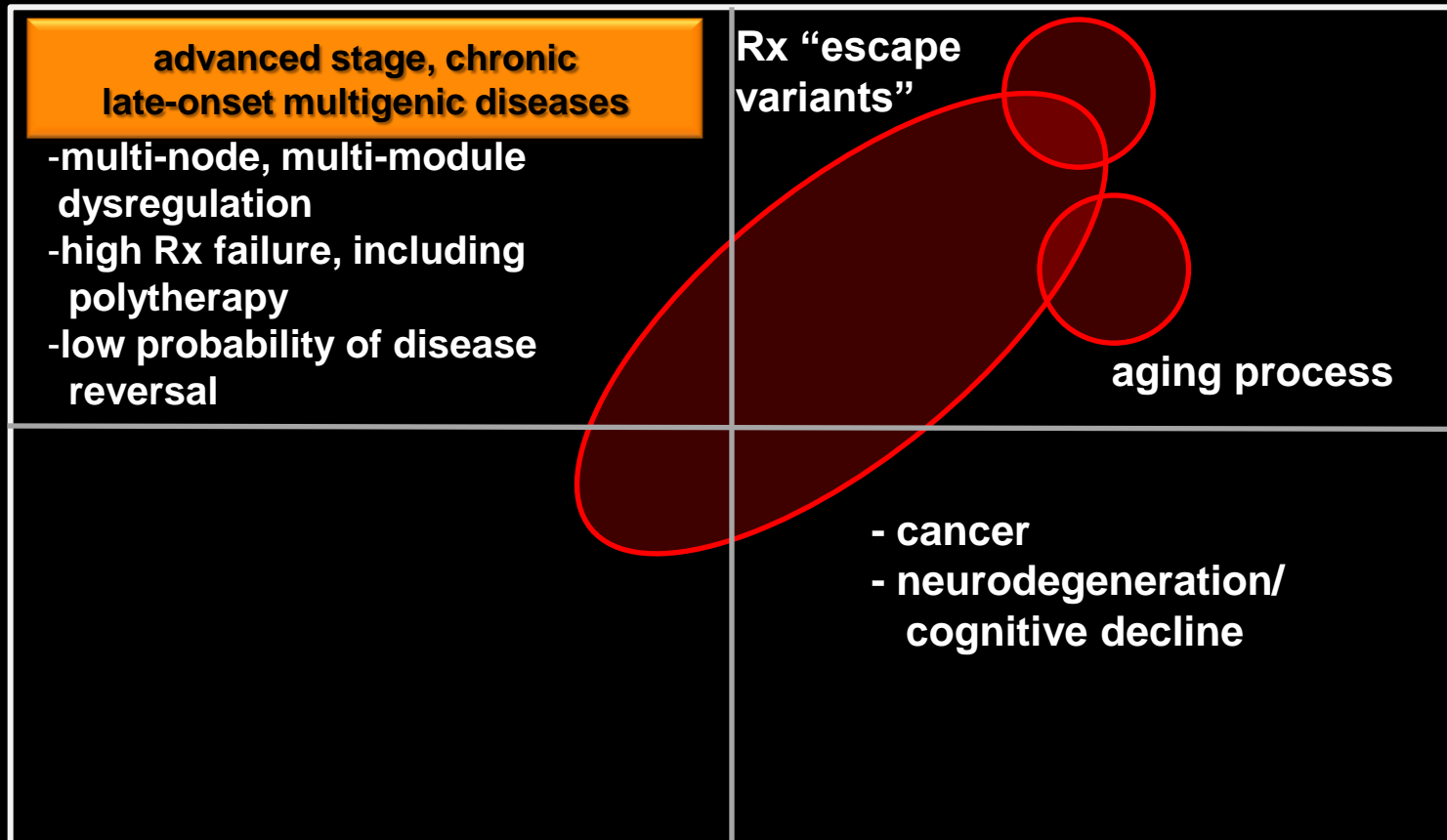
The Relationship Between The Scale of Network Dysregulation in Disease and Technical Complexity (POS) for Successful Rx-Driven 'Homeostatic Reset'



The Relationship Between The Scale of Network Dysregulation in Disease and Technical Complexity (POS) for Successful Rx-Driven 'Homeostatic Reset'

high

Extent of Network Dysregulation



low

high

Complexity (tractability) of Achieving
Rx 'Reset' of Network Dysregulation

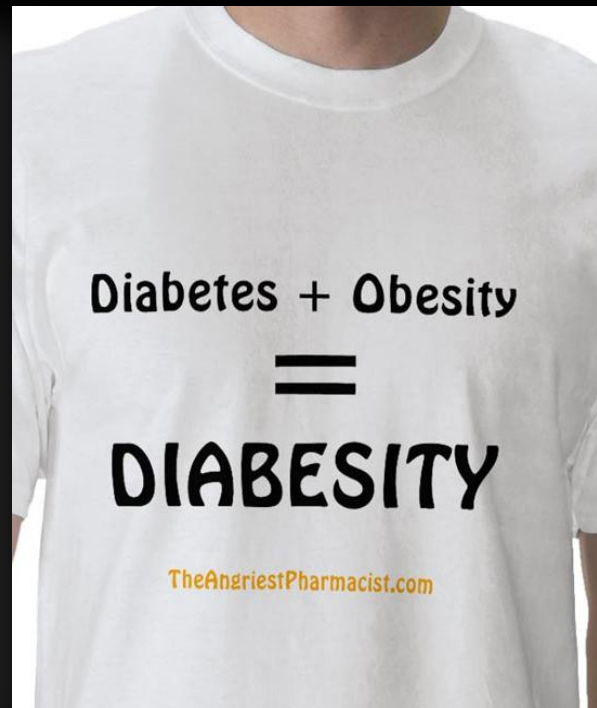
Opportunities and Challenges Posed by New Diagnostics for Ever Earlier Detection of Major Diseases

Cancer Detection Before Metastasis



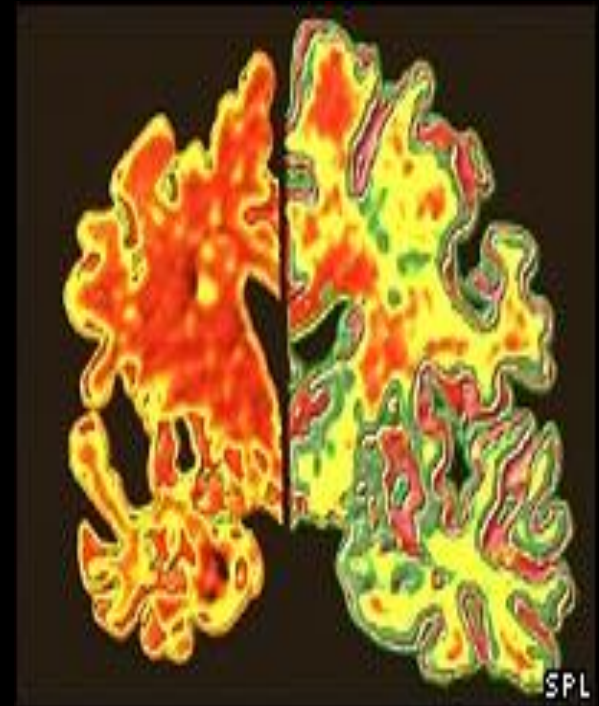
Early Diagnosis and Curative Surgery

**Cardiovascular/
Metabolic Diseases**



Lifestyle Changes and/or Rx to Limit Risk

Neurodegenerative Diseases



The Dilemma of Early Diagnosis Without Rx

Now Comes the Hardest Part of All!

**Moving Downstream Beyond Discovery:
The Escalating Scale and Complexity of the Data Stream**

**Driving Molecular Medicine and IT-Centric Capabilities Into
Routine Clinical Practice**

Overcoming Gaps in Physician Knowledge of Molecular Medicine and a Paper-Centric Healthcare System

- **90% of Americans lack confidence in their clinicians ability to understand and use genetic information**
 - http://www.cogentresearch.com/news/Press%20Releases/CGAT_2010
- **professional cultural vulnerability/reluctance to acknowledge**
- **refuge in outdated SOC/guidelines that fail to integrate much new molecular profiling data**
- **protracted deliberations by professional societies/boards**
- **less than 4% of 8967 ACGME programs relate to genetic expertise (JAMA 2011 306, 1015)**
- **MD curriculum/CME challenges**
- **generational gap in IT use/facileness and resistance to computerized decision-support tools**

Managing “Mega-Data” in Biomedicine

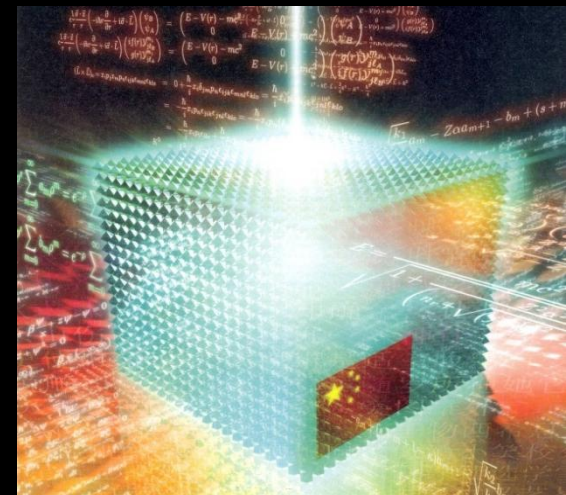
volume



computational scale



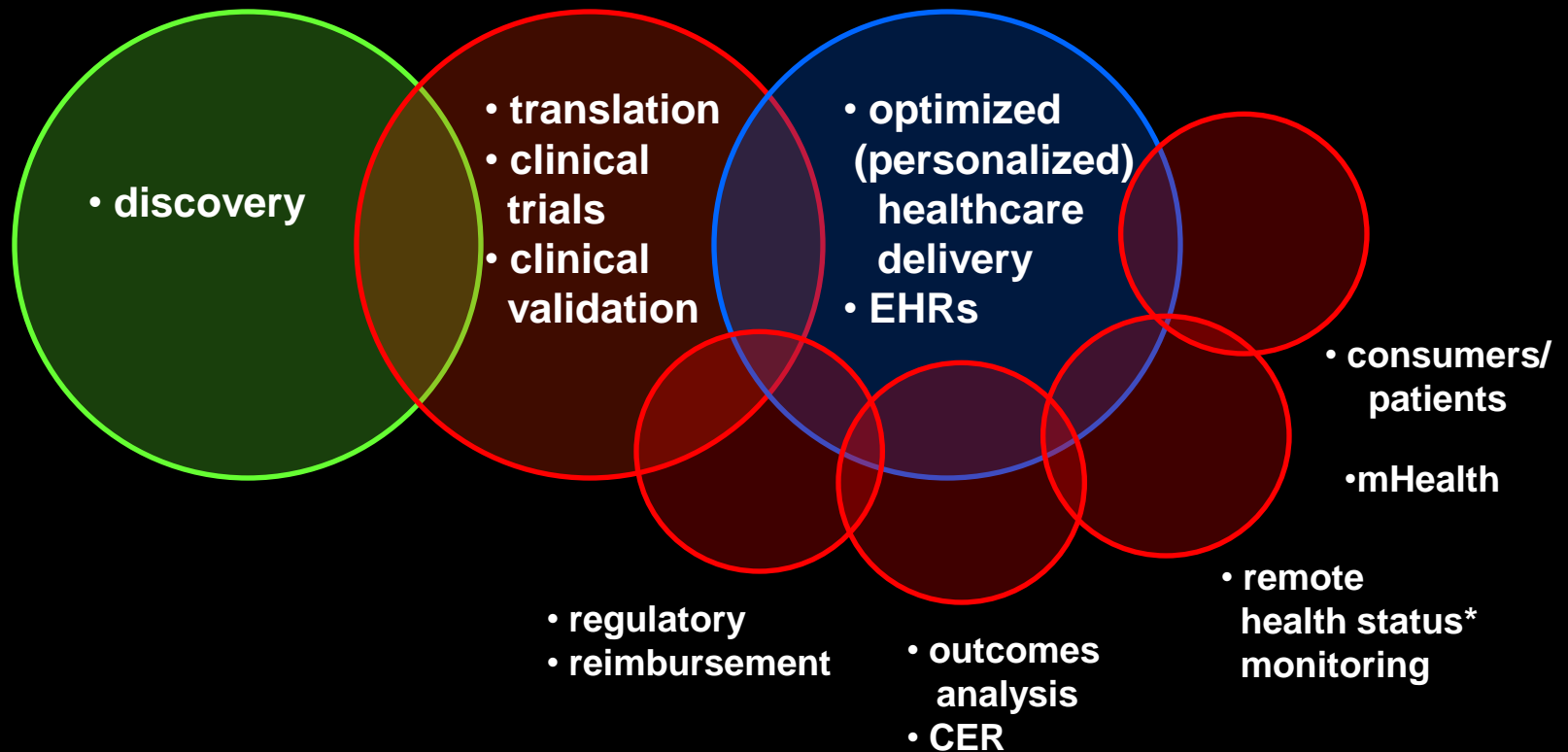
global trials and new markers



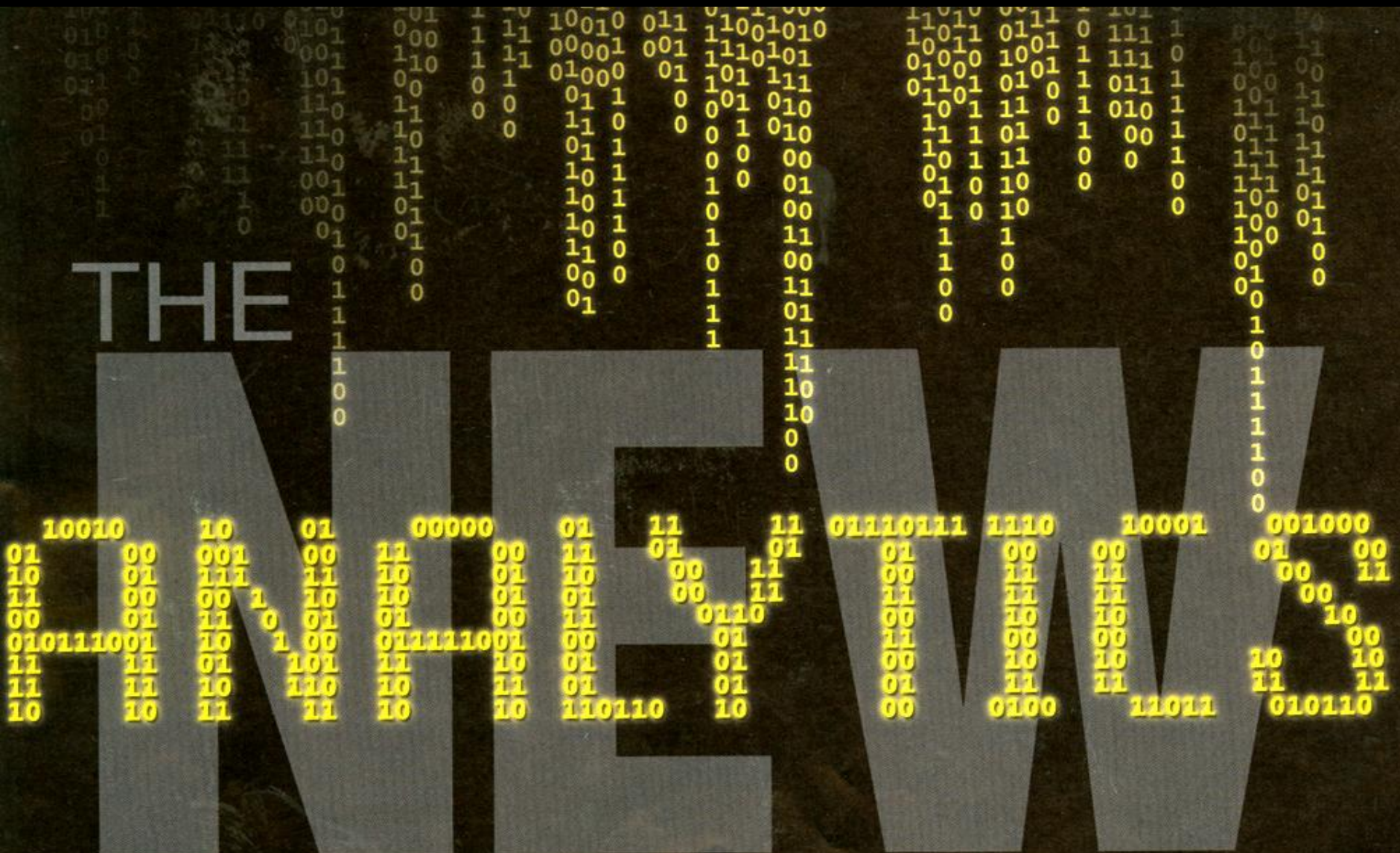
bench to bedside: multiscale heterogeneity

integration

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



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**

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



- 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)

What Is?

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

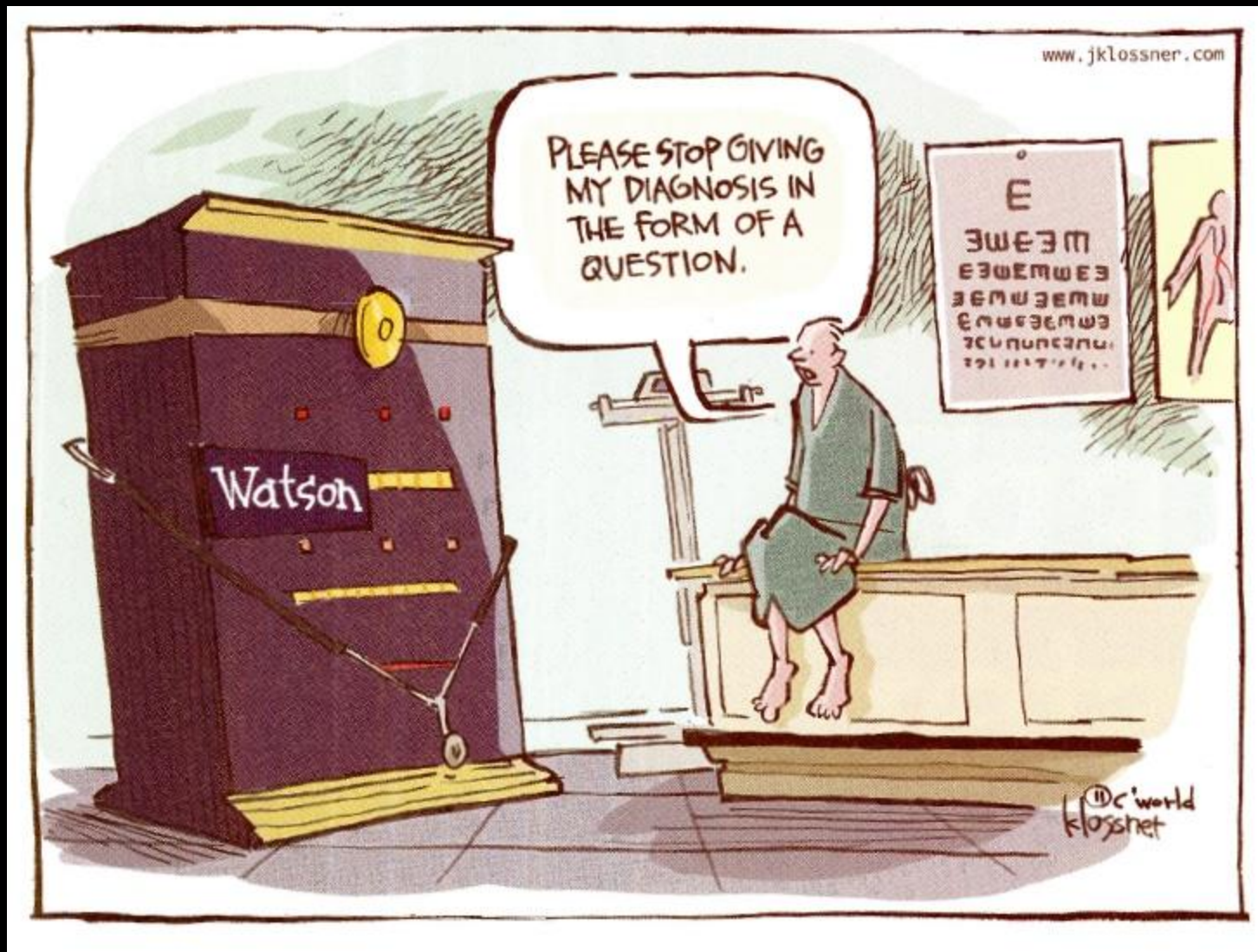


Jeopardy 16 February 2011

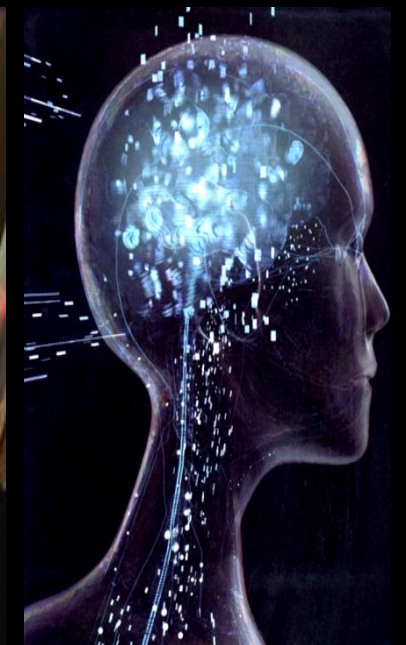
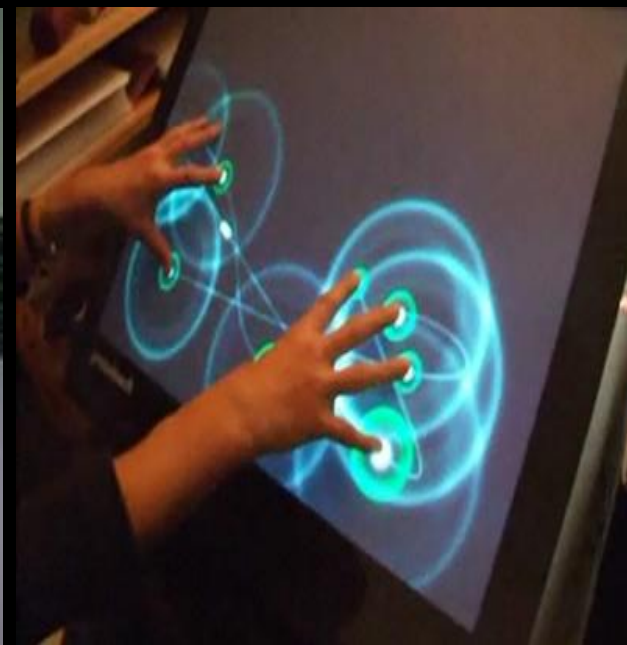
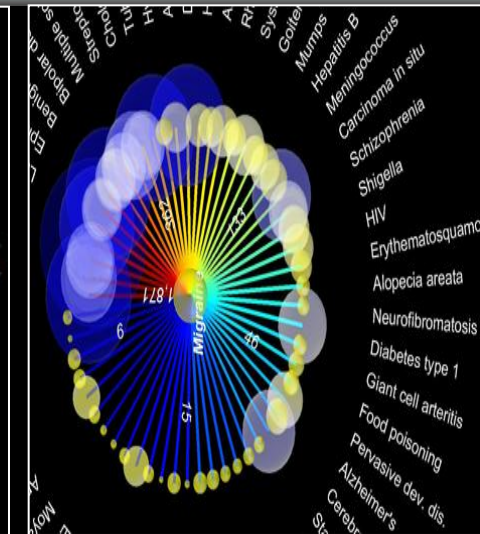
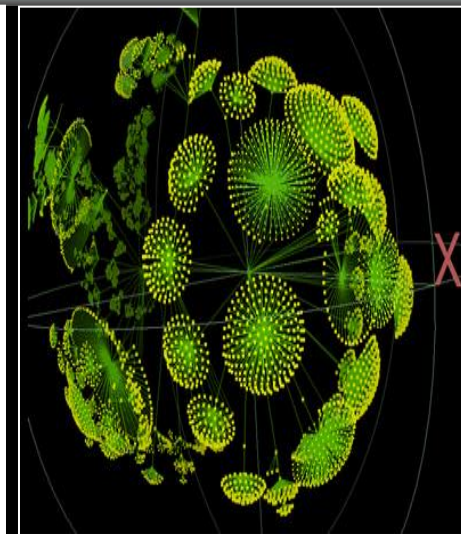
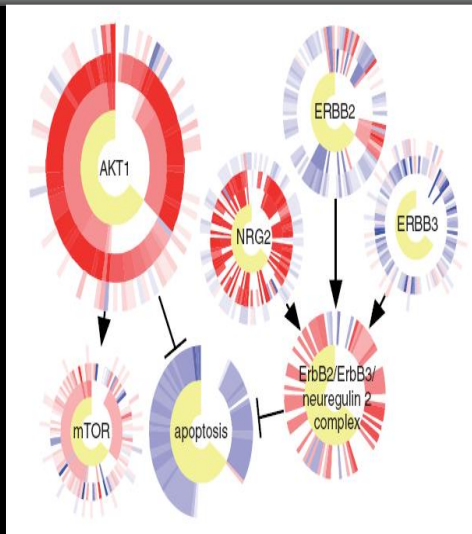
- IBM's Watson
 - 2880 CPUs
 - natural language questions
- prelude to Q&A systems for biomedicine beyond keyword IR searches

 **WELLPOINT** | Health.Care.Value.™


NUANCE



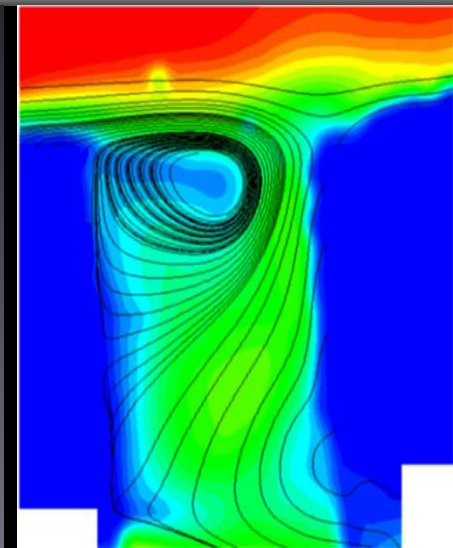
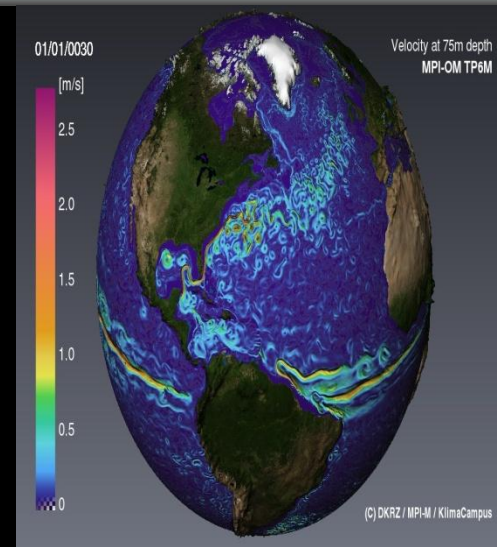
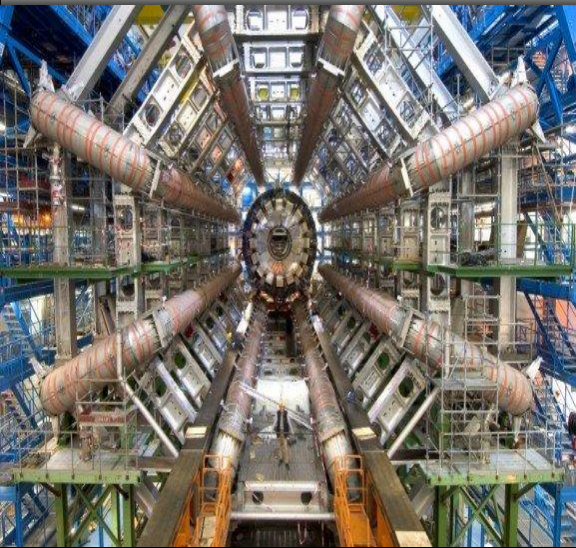
New Visualization Tools, Interactive Interfaces and Rapid Customization Formats



Planning for Rapid Growth in Data Volume and Integration of Distributed Heterogeneous Datasets

- the lottabyte challenge
 - terabytes, exabytes, zettabytes, yottabytes
- growth in volume outstripping the dropping price of physical media
- end-to-end storage strategies
 - scale, cost, location, access, security
- total-cost-of-ownership analysis for storage
 - cost per gigabyte
 - balance between physical and virtual storage
 - data retention policies
- public, private and hybrid ‘bursting’ cloud options
- security: encryption at rest/in flight

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



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



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



REPORT TO THE PRESIDENT AND CONGRESS

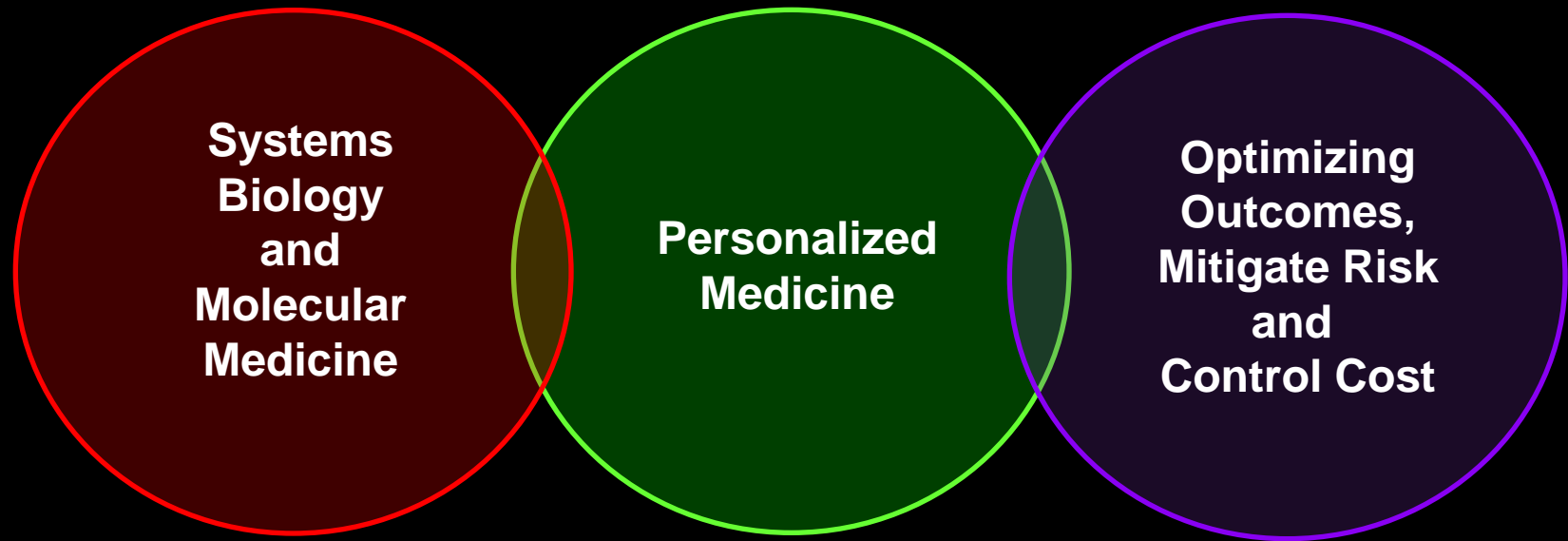
DESIGNING A DIGITAL FUTURE: FEDERALLY FUNDED RESEARCH AND DEVELOPMENT IN NETWORKING AND INFORMATION TECHNOLOGY

Executive Office of the President
President's Council of Advisors on
Science and Technology

DECEMBER 2010



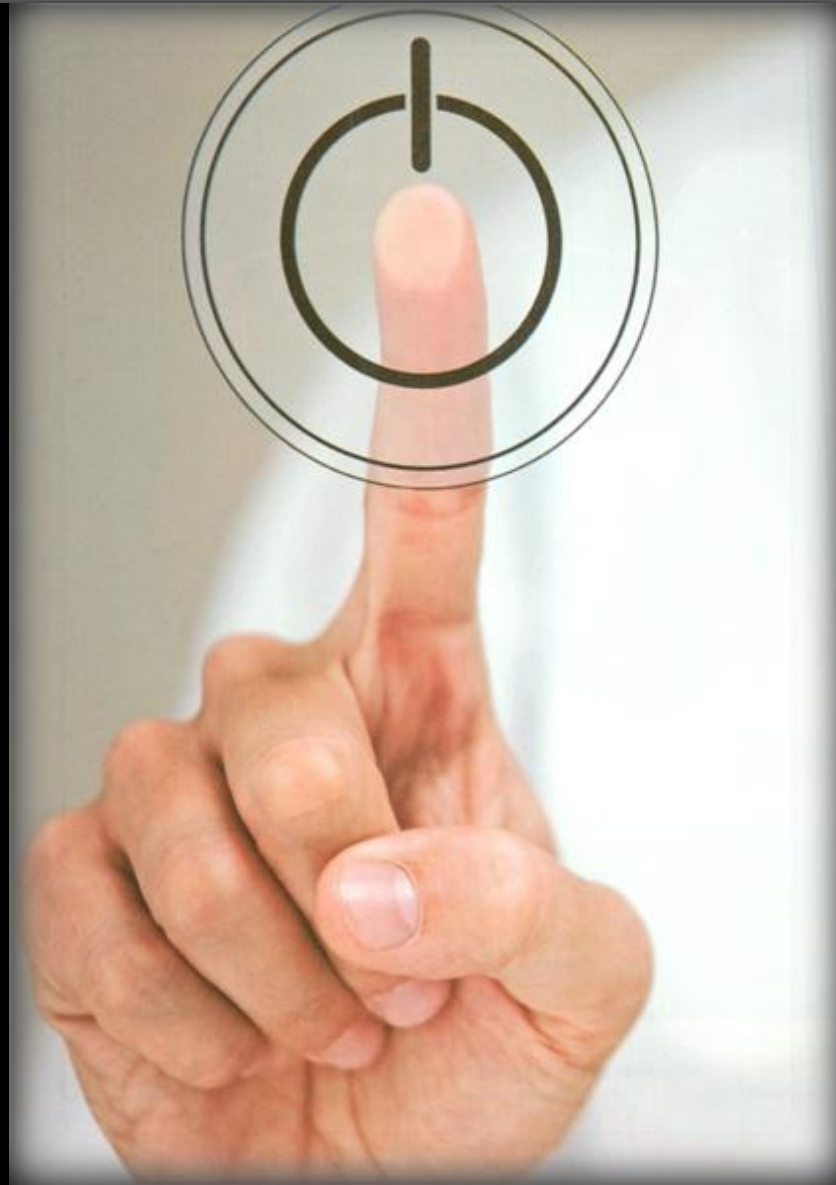
Digital Biology, Medicine and Healthcare Delivery



Massive Data

**New Technology Platforms, Databases and Analytics,
Infrastructure, Competencies and Business Models**

Reset



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

Strategic Aspirations

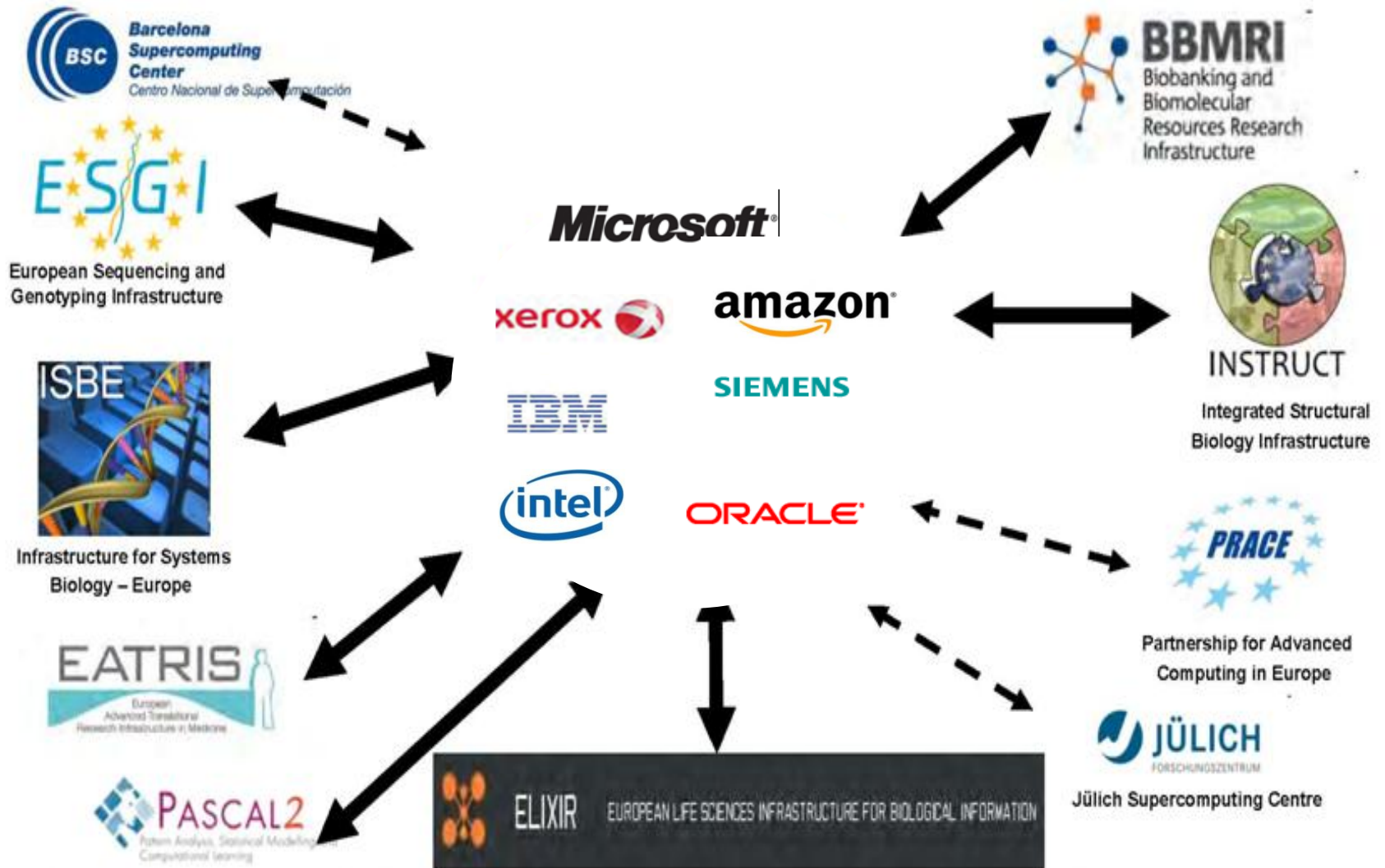
- **systems-based approaches will define physiology and pathology in terms of molecular information networks**
- **comprehensive knowledge of the topologies, dynamics and (dys)regulation of molecular networks will increase the predictability and productivity of all aspects of biomedicine**
 - **R&D strategies for Dx/Rx**
 - **clinical decisions and outcomes**
 - **risk mitigation and sustainable health costs**

Mobilization of Multi-disciplinary Scale, Pre-competitive Consortia and Private-Public Partnerships to Bring Light to the Current Black-Box of Genotype-Phenotype Relationships

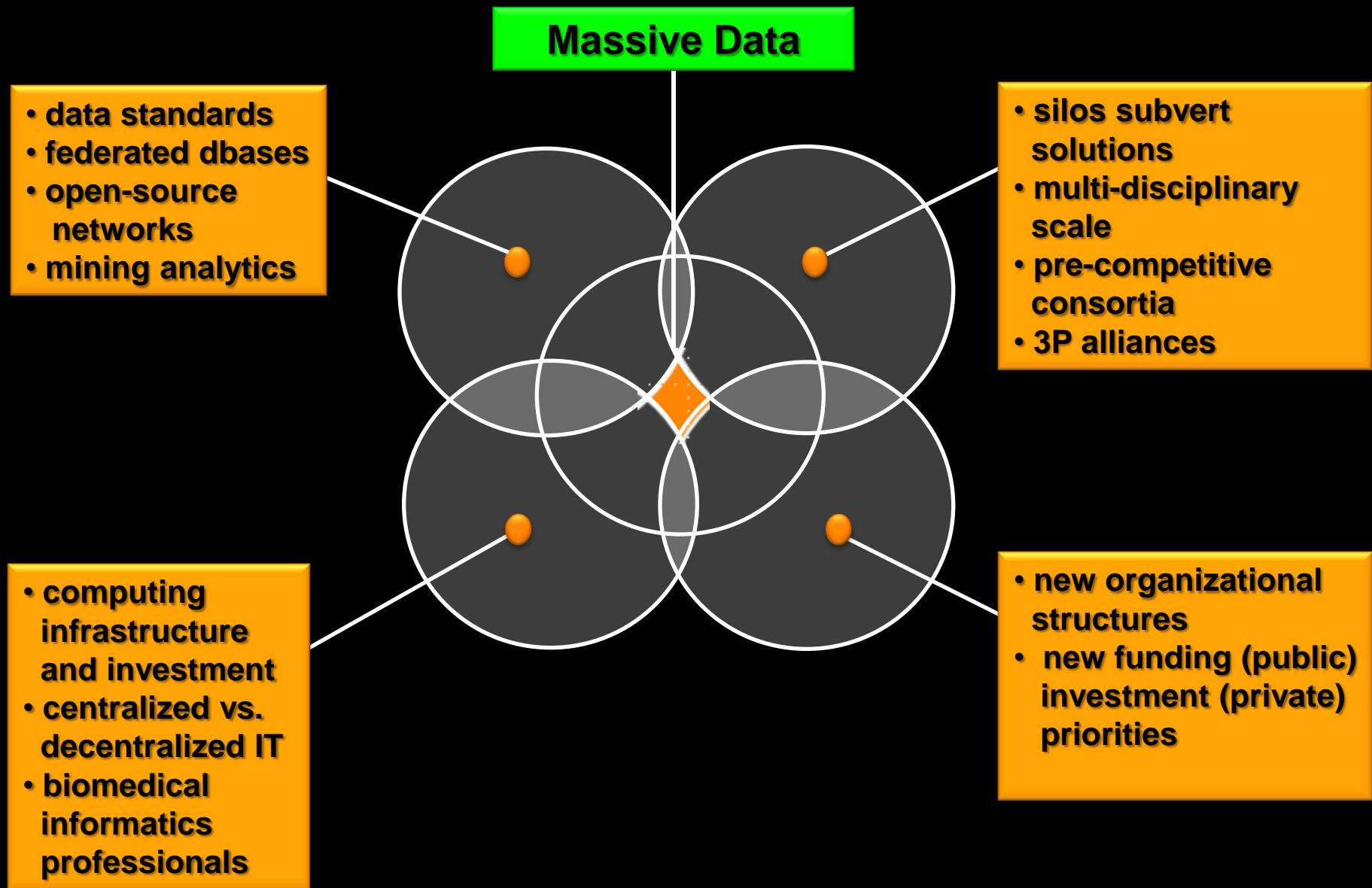
- **‘flying blind’ in understanding biological network dynamics and disease-associated perturbations**
 - limited prediction of system behavior
 - unacceptable high failure rates in clinical trials



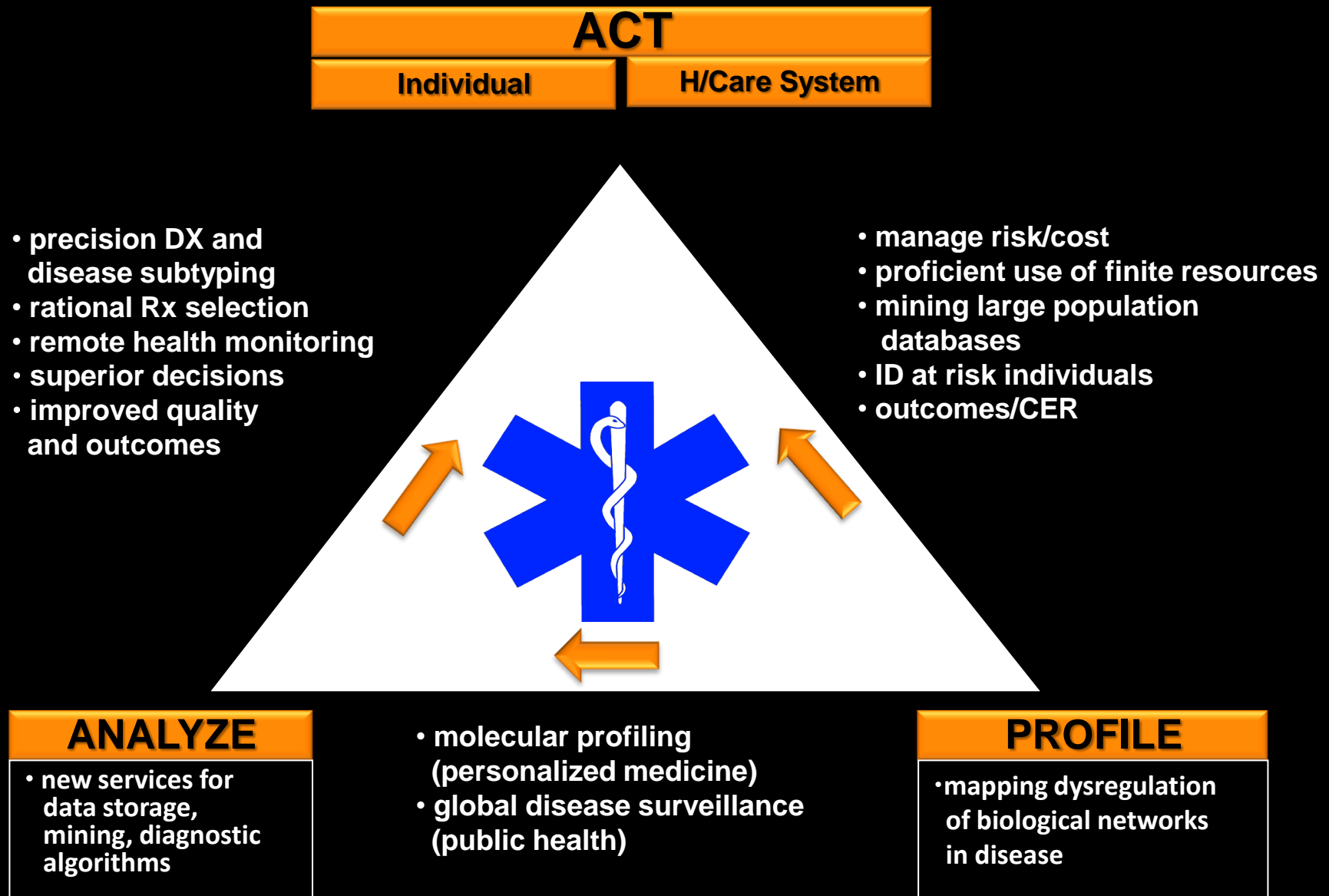
- **define ‘rules’ for biological network behavior and dysregulation in disease**
- **problem complexity unlikely to be solved by single companies or highly fragmented (‘siloed’) academic initiatives**
- **new organizational models**
 - multi-disciplinary, multi-institution, multi-sector
 - 3P alliances
- **multi-partner precompetitive consortia aligned to major disease challenges**
 - robust base for subsequent competitive ‘D’ process
 - Dx, Rx, PDx and HIT



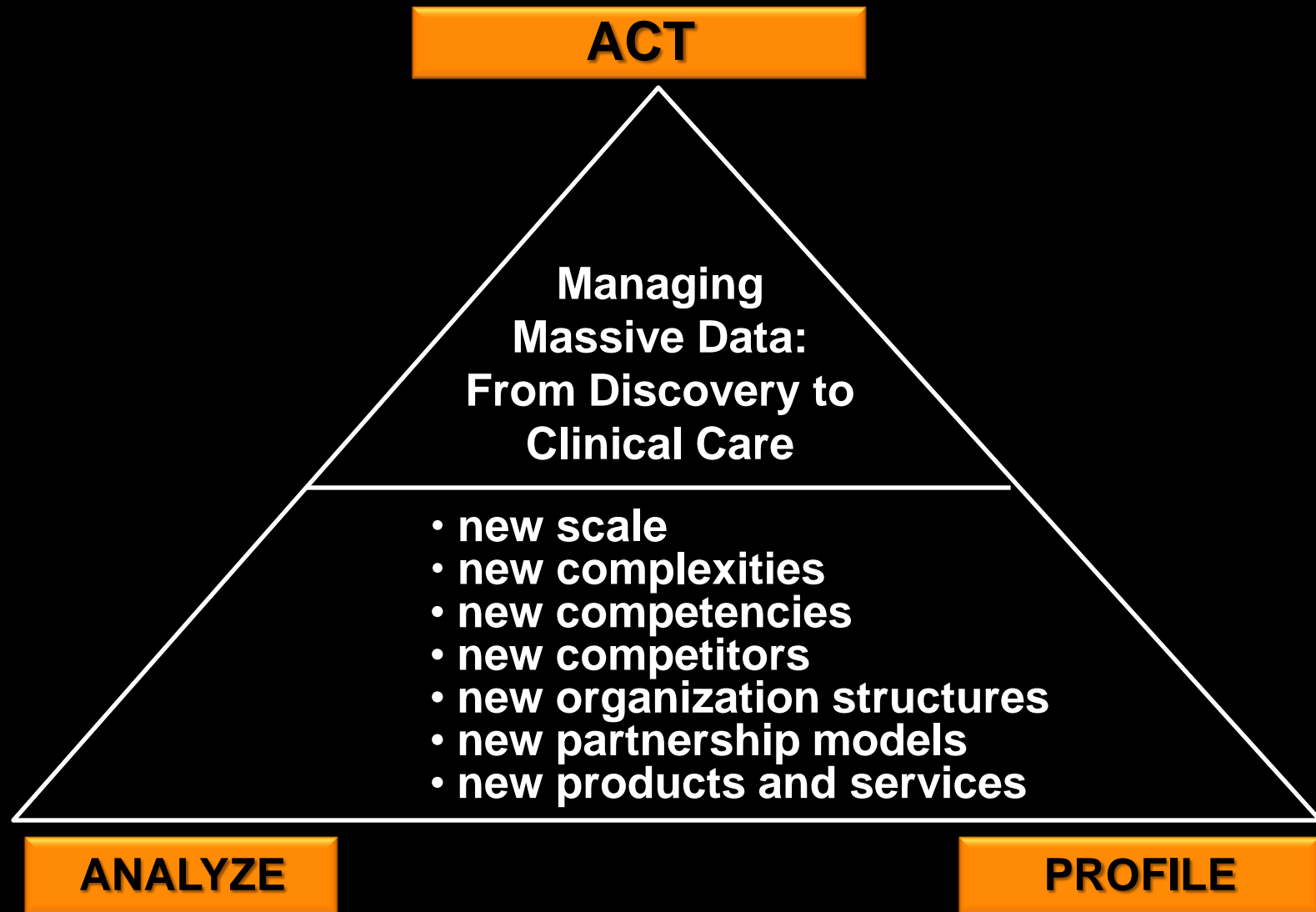
Managing Massive Data: Disruptive Changes and New Products, Services and Partnership Models



Managing Massive Data and Driving New Value Propositions in Biomedical R&D and Healthcare Delivery



Managing Massive Data and Driving New Value Propositions in Biomedical R&D and Healthcare Delivery



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

