



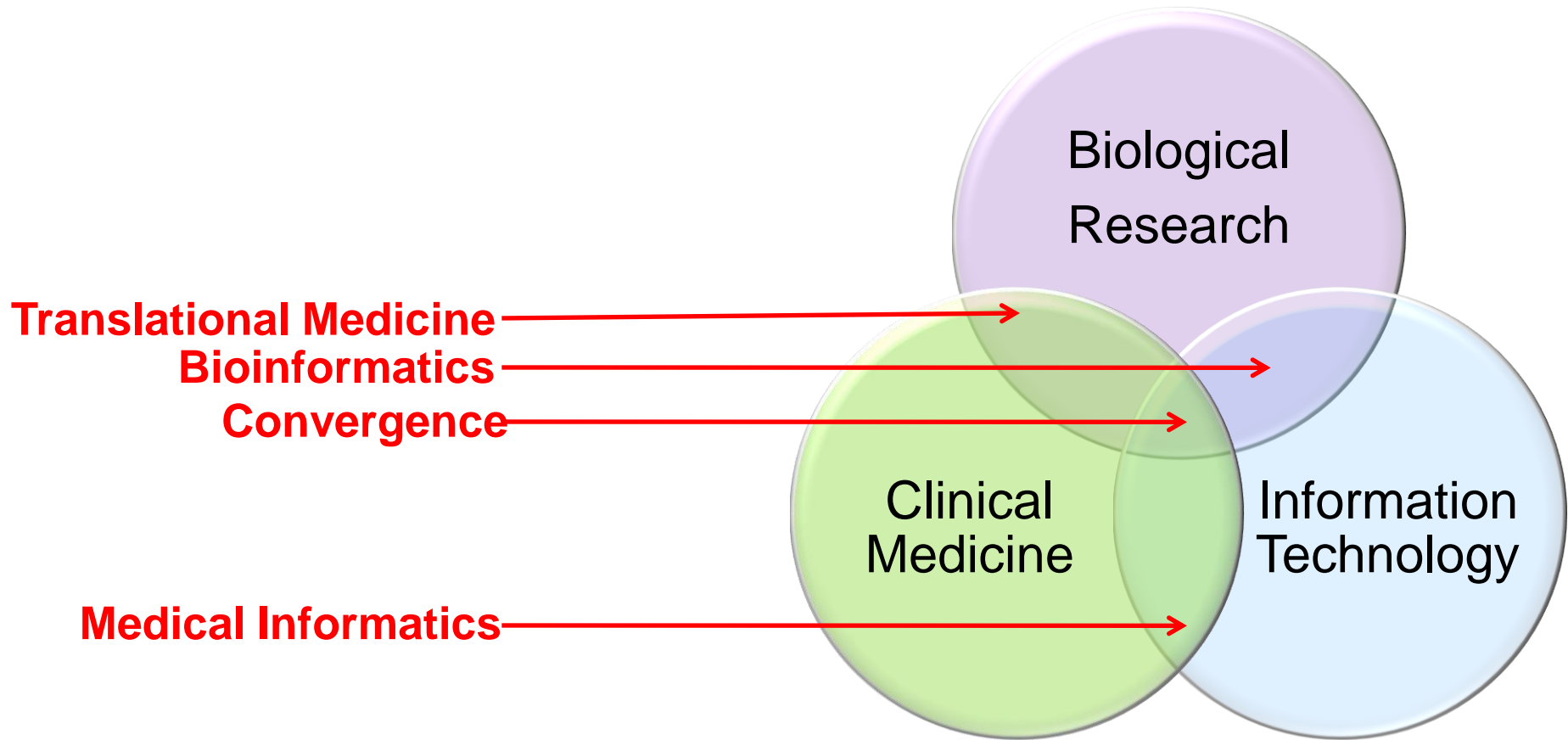
CSC

Bioinformatics and the Future of Medical Research and Clinical Practice

Robert V. House, Ph.D.
President, DynPort Vaccine Co.

April 29, 2009

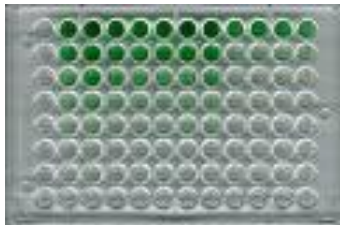
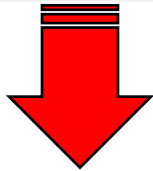
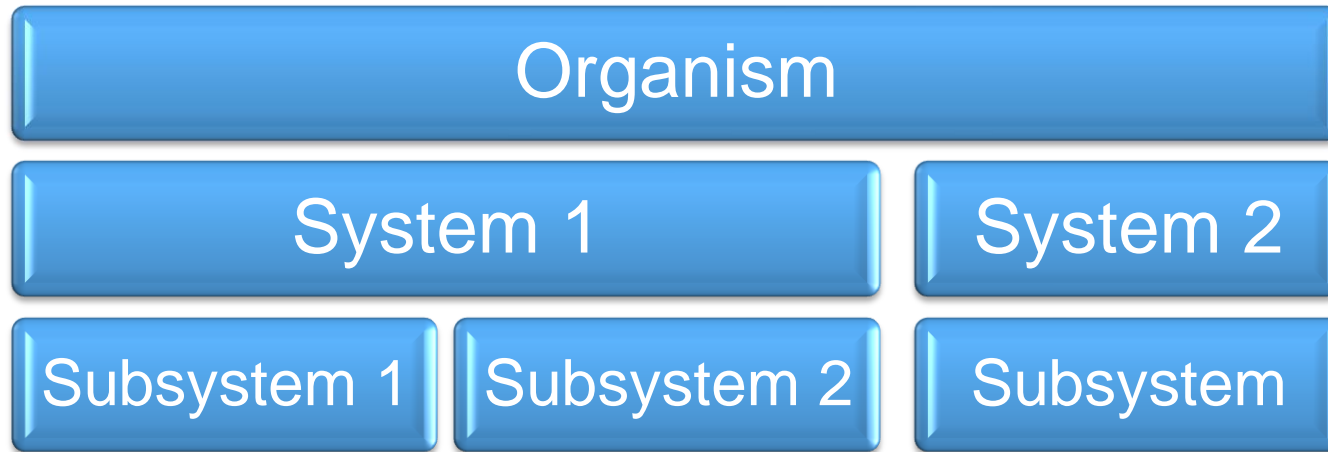
Relationship of Terminology: One Man's View



**"When I use a word, it means just what I choose it to mean -- neither more nor less."
– *Through the Looking Glass***

Why Do We Need a New Biology?

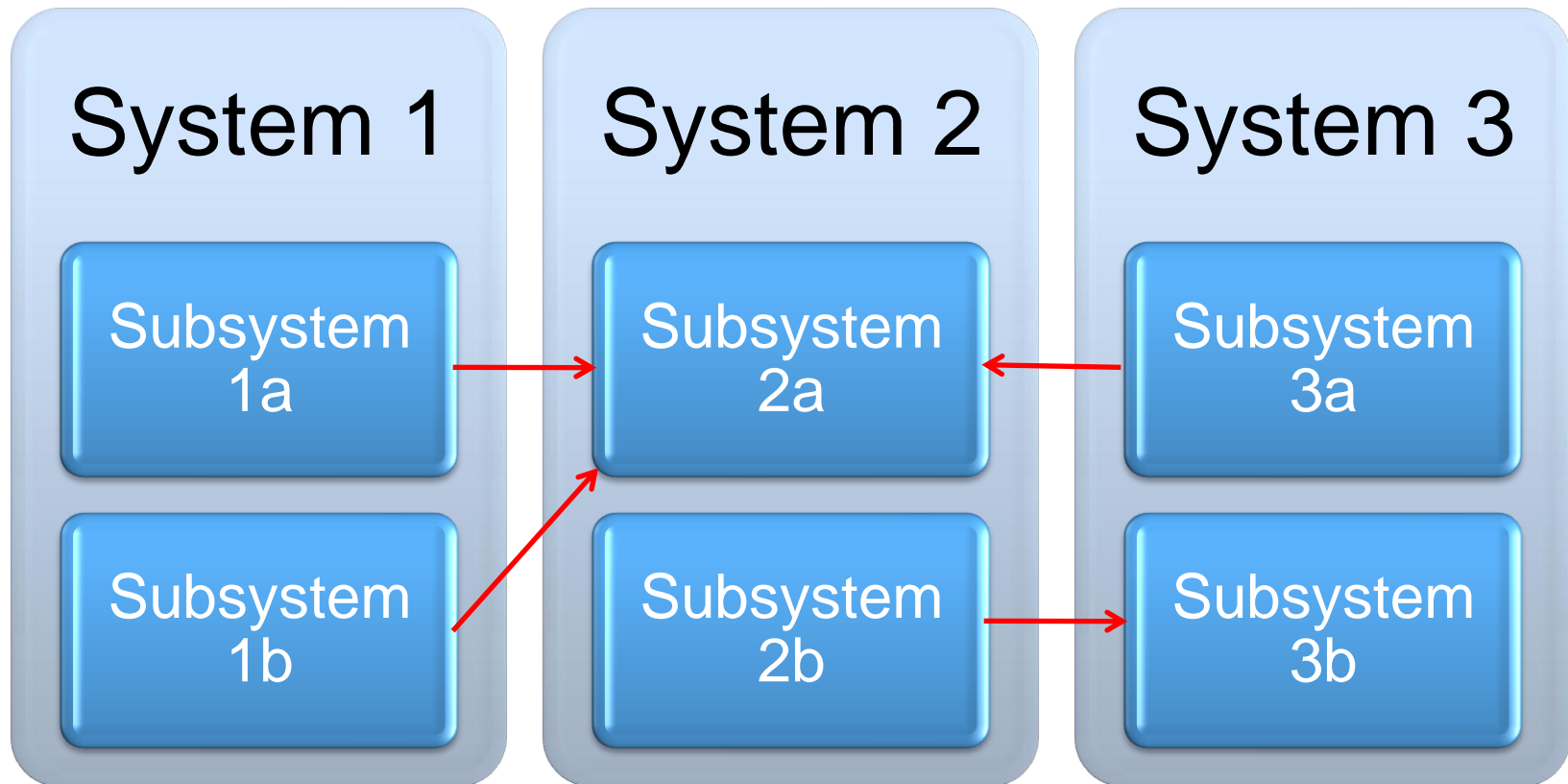
Traditional reductionist scientific method



By studying highly isolated subsystems, we are misrepresenting how the subsystem functions as part of the whole system. We never have the complete elephant, only an extrapolation.

To date, this has been necessary because we lacked the tools and data-handling capabilities to see the whole elephant.

Systems Biology: Inevitable



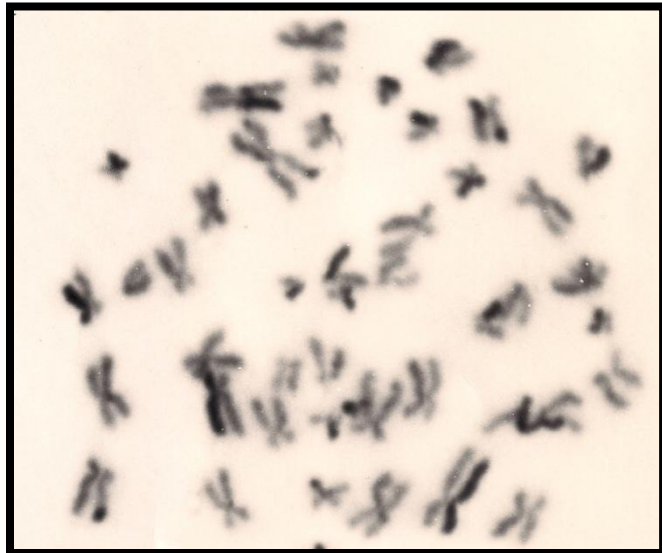
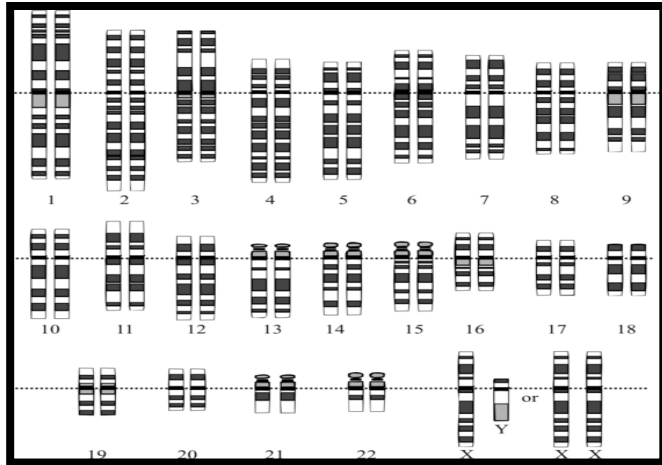
Systems biology recognizes that processes in every system have an effect on every other system. Bioinformatics provides the tools, the data-handling capacity, and the conceptual models to make this possible.

Synthetic Biology: Unpredictable

- Synthetic biology is the nexus of biology and engineering in which biological processes are reduced to their most elemental form.
- These structures and functions are then engineered using molecular biology, nanotechnology, and microfabrication, resulting in standardized “parts” which can then be reassembled into purpose-driven structures.
 - Example: BioBricks
 - Example: Nanobots
 - Example: Synthetic life
- Understanding the “templates” upon which such organic machines are built is fundamentally dependent on the tools provided by bioinformatics.
- Far-future applications of the fusion of synthetic biology with convergence (NBIC) are a Black Swan.

“Commerce is our goal...more human than human is our motto.” – Blade Runner

The Human Genome: Beauty in Complexity



- About 3 billion base-pairs (letters): A/G/C/T
- 20,000 - 25,000 genes (fewer than expected)
- Coding sequences (exons)
 - DNA sequences that specify protein production
- Noncoding sequences (so-called “junk”)
 - Retrotransposons
 - Tandem repeats
 - Interspersed repeats
- Fun fact: almost 8% of the human genome is “fossil” retrovirus (98,000 viral fragments)
- These fossils may drive evolution more than natural mutation

The genome is not really a “blueprint”. It is necessary to detect the message and ignore (or understand) the noise.

Decoding the Language of Life

DNA (INSTRUCTIONS)



RNA (MESSAGE)



PROTEINS (BUILDING BLOCKS)



METABOLISM (PROCESS)

GENOMICS



TRANSCRIPTOMICS



PROTEOMICS



METABOLOMICS

DNA SEQUENCE



GENE MESSAGE



PROTEIN
STRUCTURE
(1°, 2°, 3°)



METABOLITES

This is highly abbreviated and subject to constant revision. There are now dozens of “omics”, all with massive amounts of data.

Any “system of systems” can similarly be parsed into discrete, yet interconnected, information flows.

What's in the Toolbox?

- Data generation tools:
 - Genomics
 - Sequencing technologies, including next-generation/high-throughput
 - Brute force approach, regardless of technology
 - Tells us the exact sequence of AGCT in the genome, little else
 - Billions of data points per organism, many different organisms of interest
 - Data storage and retrieval becomes nontrivial
 - Error control important for all downstream data analyses

What's in the Toolbox?

- Data generation tools:
 - Transcriptomics
 - Mostly microarray technology
 - DNA is unzipped and copied (cDNA), copy is bound to a substrate (microarray)
 - Tagged RNA sequences are hybridized to cDNA and genes are identified
 - There are at least tens of thousands of potential genes
 - Genes must be deduced from genomic sequence and not all gene probes have been made, so gaps exist in our knowledge.

What's in the Toolbox?

- Data generation tools:
 - Proteomics
 - Standard protein chemistry tools including 2D gel electrophoresis, chromatography, etc.
 - Immunoproteomics (physicochemical separation in a matrix followed by probing with labeled antibodies) or protein microarrays
 - First level of discrimination of actual physiological function
 - Metabolomics
 - Primary approaches at present are targeted analysis, metabolic fingerprinting and metabolic profiling
 - Thousands of proteins known
 - Protein function is subject to 3D conformation (added complexity)
 - Data stream is potentially enormous when secondary variables are included such as genetic predisposition, species differences, concomitant environmental exposure and potential toxicity (for example)

What's in the Toolbox?

- Data analysis tools:
 - **Prospective:** Generally open-source (but increasingly proprietary) software to parse data into discrete blocks of information.
 - Image analysis for reading microarrays
 - Sequence databases to determine genes from sequence analysis (e.g., BLAST)
 - There are many competing options, but very little standardization
 - **Retrospective:**
 - Data warehousing
 - Database cleaning
 - Representing missing data points
 - Accommodation for noise in the system
 - Creation of logical access to the various forms of data
 - Including off-line data and metadata

What's in the Toolbox?

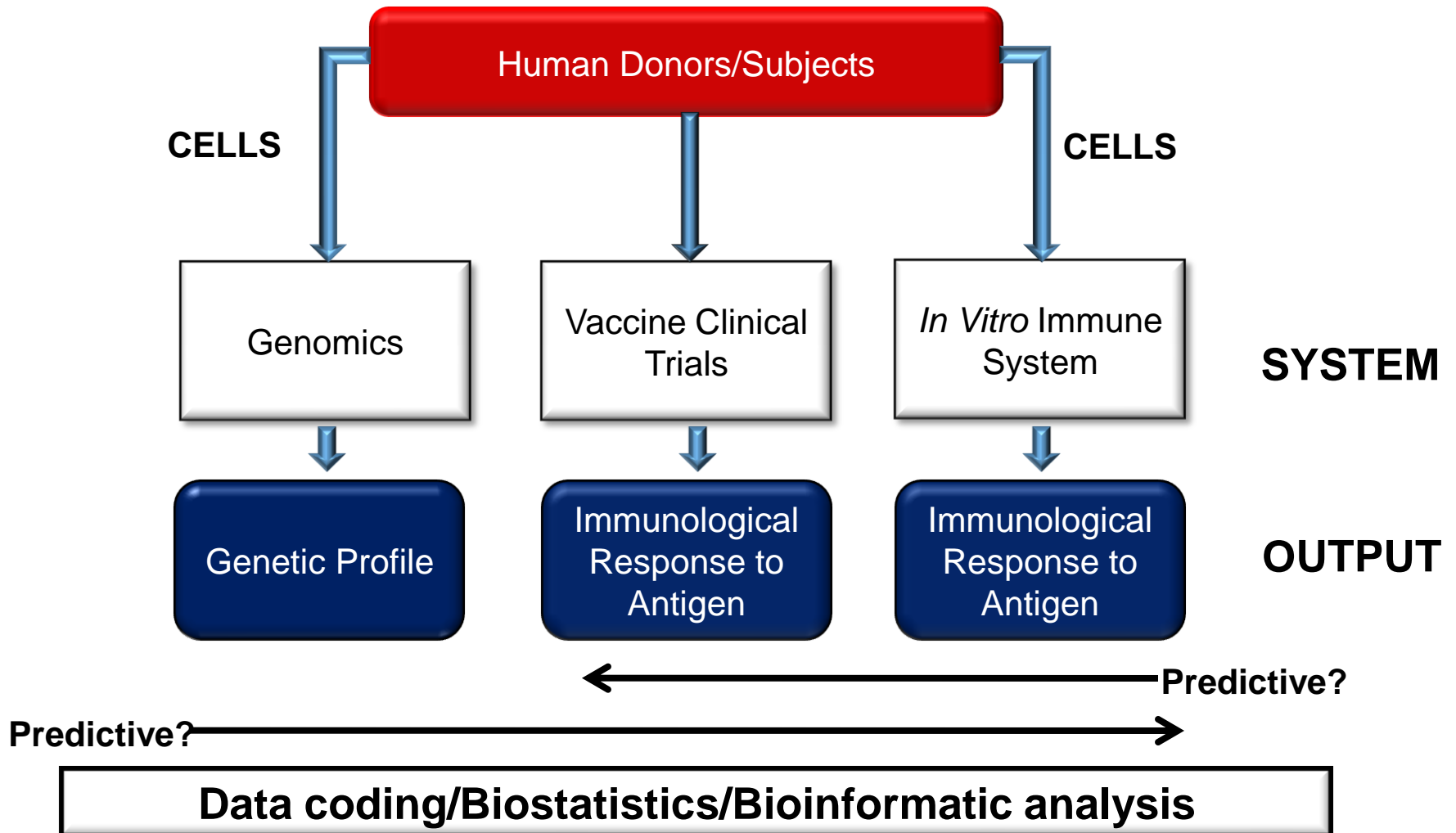
- Data analysis tools:
 - **Retrospective (cont):**
 - Data mining
 - Classification
 - Regression
 - Clustering
 - Summarization
 - Dependency modeling
 - Change and deviation detection
 - **“Social networking”**
 - Peer-to-peer data sharing and collaborations
 - Massive data bandwidth necessary for real-time data sharing in some cases
 - Not as effective as it could be due to lack of single (or limited few) standard(s)
 - This is one of the most pressing issues (i.e., opportunities) in bioinformatics

APPLICATIONS

Building a Better Vaccine

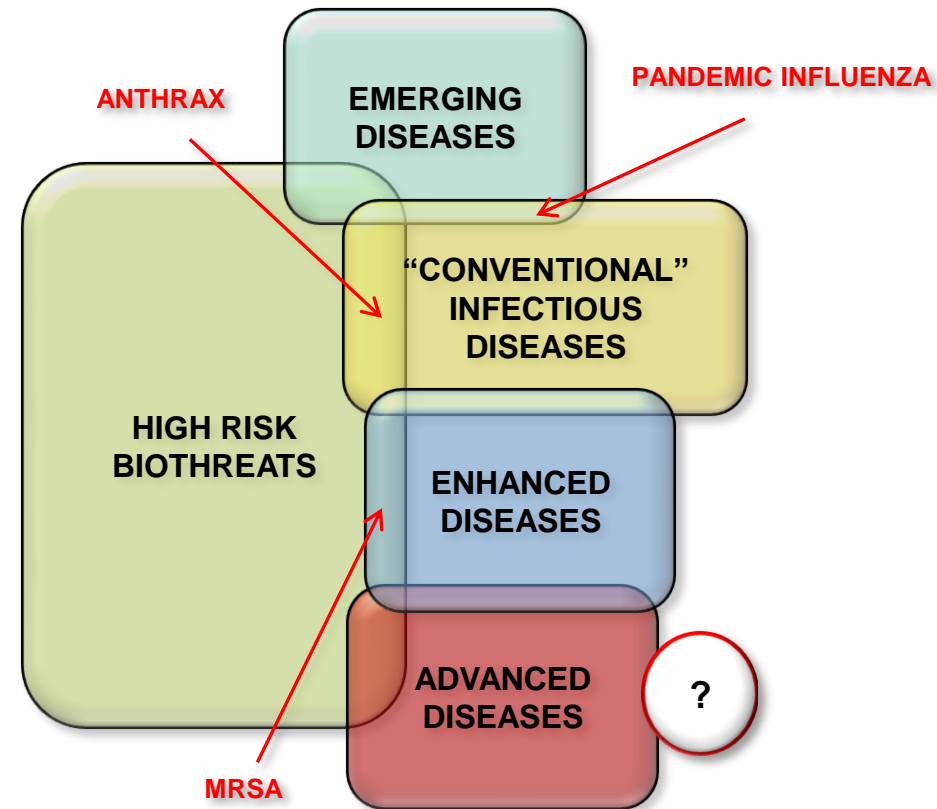
- **Old School:** Kill or inactivate your bug of choice, inject it, and hope that the immune system is tricked into believing that it has survived a real infection.
- **New School:** Using various genomic and proteomic tools, predict in advance which portions of a disease organism cause disease, which portions of the bug that the immune system will recognize, and how the two will interact.
 - Design a vaccine based only on these highly specific interactions (a process known as “reverse vaccinology”).
- **New School Advanced Placement:** Use genomic tools to predict how humans will react to any give vaccine before they ever receive it.
 - Increase efficacy
 - Decrease adverse effects

Next-generation Vaccine Development

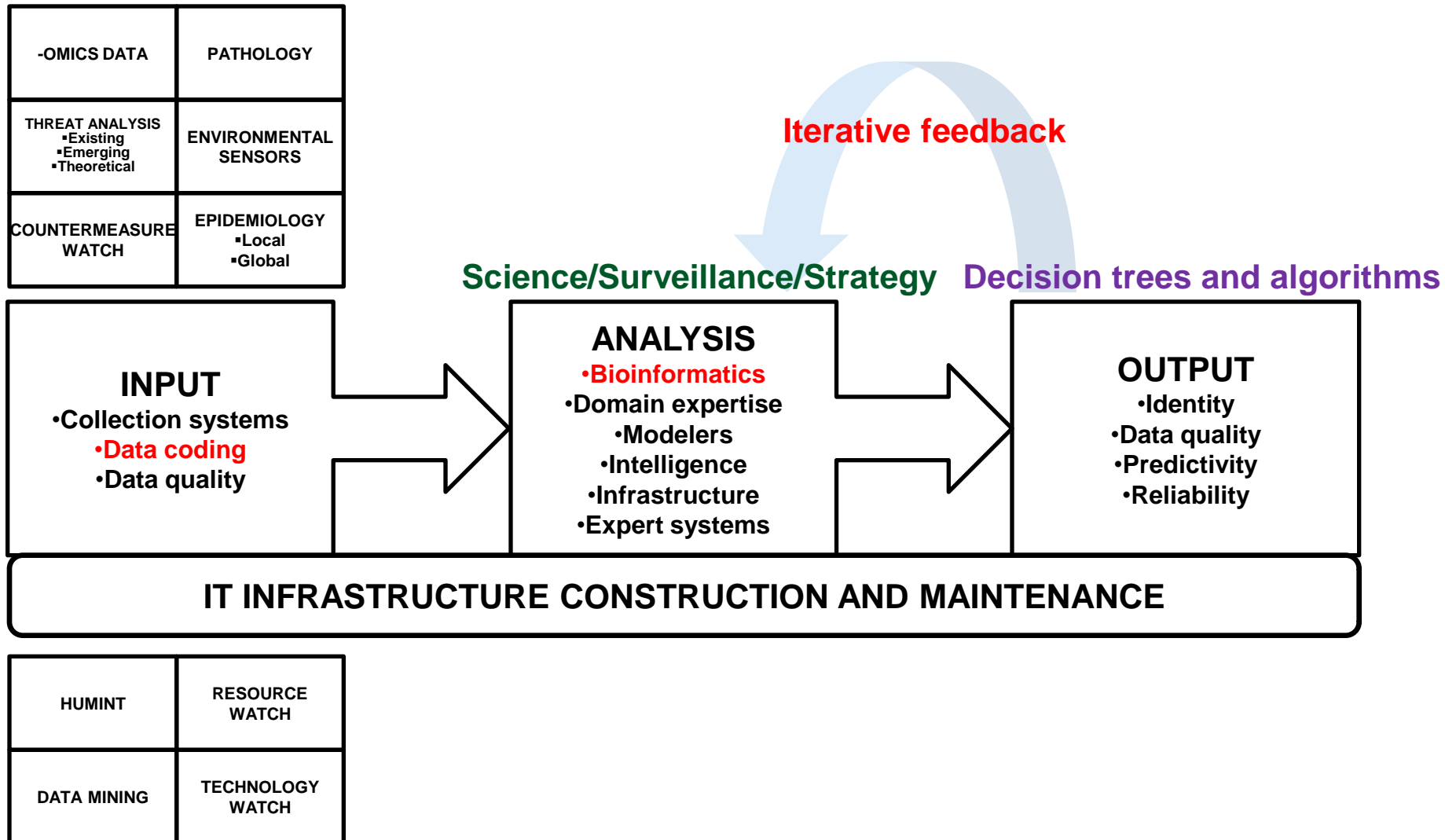


Responding Better to Emerging/Re-emerging Diseases

- A diversity of threats:
 - Emerging diseases
 - MRSA, Hantavirus, Marburg
 - Re-emerging diseases
 - Pandemic influenza
 - Biodefense concerns
 - Traditional agents
 - Enhanced agents
 - Advanced agents
- The role of bioinformatics:
 - Molecular forensics (bioterrorism)
 - Tracing known agents (Amerithrax)
 - Identifying unknown agents (bioengineered weapons)
 - Molecular epidemiology (unknown natural diseases)
 - Disease modeling (all)



Responding to Unknown Biological Threats: Big Picture



Realizing the Dream of Personalized Medicine

- Current medical treatments are based on anticipated mean response in large populations, so optimal efficacy must be balanced with minimal toxicity (sometimes an uneasy truce).
- Examples of personalized medicine include:
 - Individualized diagnosis and treatment of cancer
 - Currently the most fully realized form of individualized medicine
 - Example of HER-2 receptor and susceptibility to breast cancer
 - Improving drug efficacy through the use of pharmacogenetics and pharmacogenomics
 - Improving drug safety through the use of toxicogenomics and systems biology
 - “Fusion” technologies (theragnostics)
- Truly personalized medicine (that is, individual-specific) may never be practical due to economics, but great improvements can still be made.
 - Example: disease predisposition based on SNPs

How Will This Change Clinical Practice?

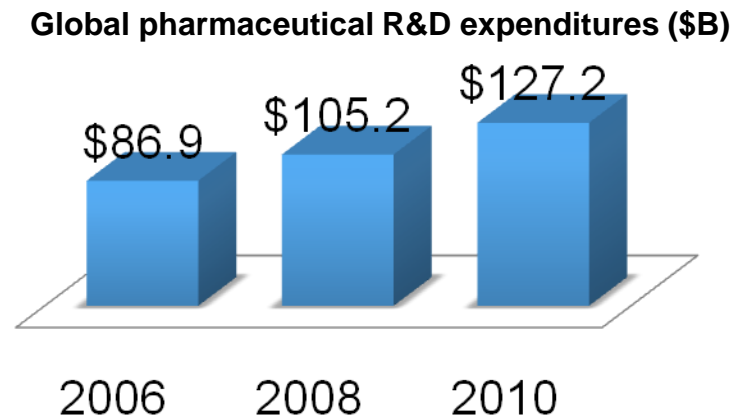
- Novel applications such as theragnostics will create a continuum between diagnosis and treatment of diseases, improving outcome.
- Therapies will be based on genetic characteristics and their consequences, not on broad physiological processes.
 - Understanding of the language of life will facilitate understanding of drug metabolism based on genetic background, protein-drug interactions, etc.
 - Susceptibility to certain diseases can be better predicted by considering genetic factors in the disease process (better diagnosis based on genetics rather than relying only upon physical observations and “one size fits most”).
 - Therapies can be tailored to specific populations, improving efficacy and safety.

Building a Bioinformatics Industry: Economic Factors (1)

- The global bioinformatics industry has grown at a double-digit growth rate in the past and is expected to follow the same pattern in the next several years (2009 to 2012).
- Currently, the U.S. remains the largest market in the world, but India and China have the fastest growth rate.
- The biggest opportunity will be in the drug discovery sector. Bioinformatics reduces the overall drug development timeline by 30% and the annual cost by 33% due to fast development of tools and software.
- Given that the development lifecycle for a new drug or biologic comprises 12 to 15 years and costs approximately a billion dollars, there is significant incentive to reduce the time necessary to develop products.

Building a Bioinformatics Industry: Economic Factors (2)

- Major U.S. pharmaceutical companies are expected to increase their R&D expenditures in the future; a sizable portion of this spending is expected to go toward bioinformatics.
- Global pharmaceutical R&D expenditures in 2006 were \$86.9B, and were expected to rise to \$105.2B in 2008, a 21% increase over two years. Moreover, expenditures in 2010 are predicted to rise to \$127.2B (another 21% increase in just two years).
- The worldwide value of bioinformatics is expected to grow to \$3.0 billion in 2010, at an average annual growth rate of 15.8%.



Building a Bioinformatics Industry: Economic Factors (3)

- Currently estimated at \$717 million, the content market will almost double to \$1.4 billion by 2010. Specialized databases will form the major part of bioinformatics content market; the share of specialized databases in the total content market will increase from 67.6% in 2005 to over 75% in 2010.
- The fastest growing market is expected to be analysis software and services. The segment is estimated to grow at an AAGR of 21.2% from \$444.7 million in 2005 to \$1.2 billion in 2010.
- A potentially huge market may exist in building and maintaining the IT Infrastructure to support bioinformatics (data storage, communications, etc.).

Caveat: Given recent economic events, it is unknown how this trend will hold up.

Impediments to Growth are Also Opportunities (1)

- The bioinformatics market is fractured:
 - Characterized by numerous individual companies catering to the particular needs of drug developers
 - Fewer companies focused on providing integrated solutions to broader R&D requirements
- The lack of standardized applications addressing R&D issues:
 - Limited the growth of bioinformatics and inhibited its development into a full-fledged industry
 - Lack of integration between the various players in the bioinformatics business model (software vendors, database providers, etc.)
 - Lack of integration between the internally developed systems of drug companies and technologies provided by outside vendors

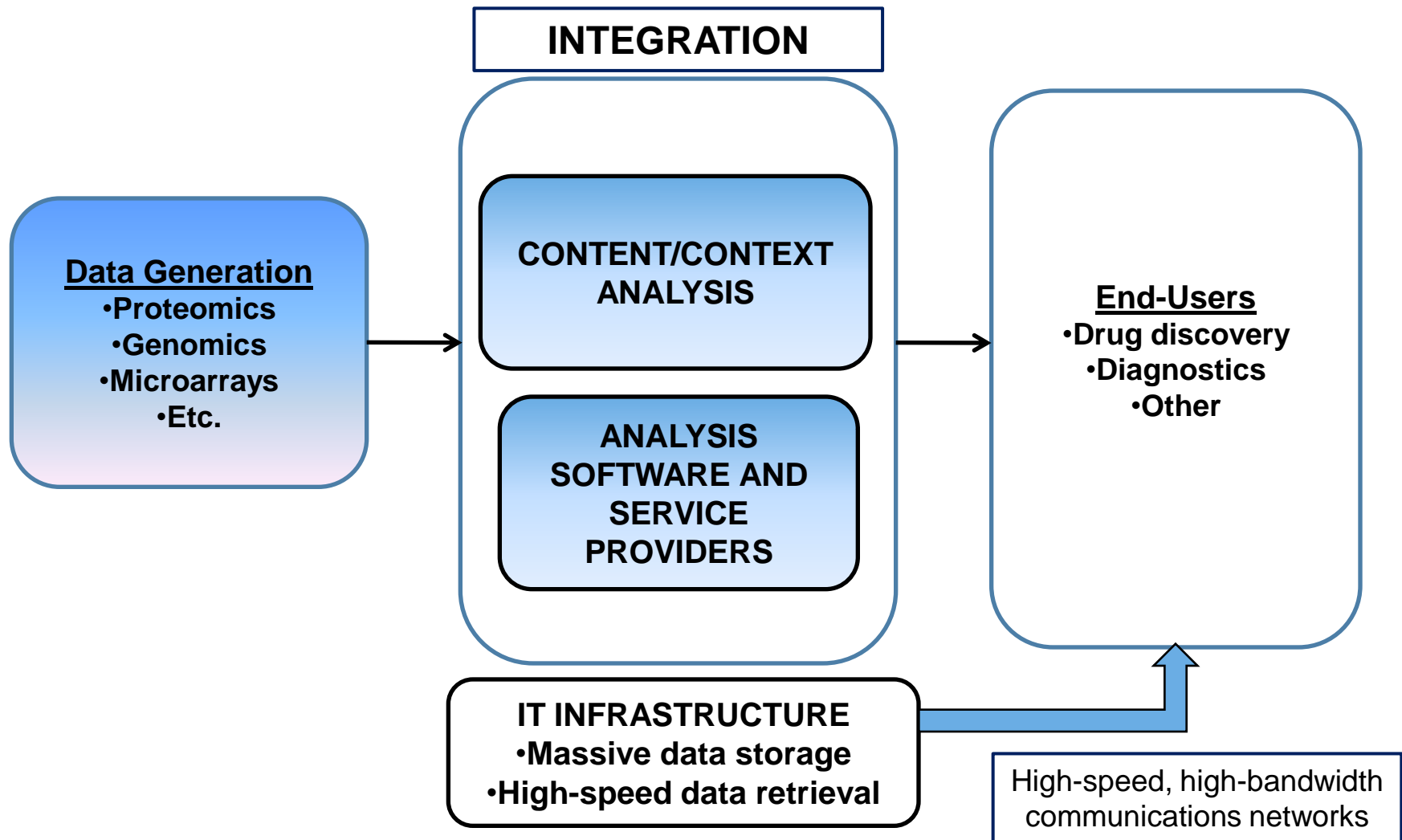
Based on a report from The Larta Institute

Impediments to Growth are Also Opportunities (2)

- The market isn't large enough to support the number of companies involved:
 - The market is fractured and niche-oriented such that standardization and scale associated with an industry will be difficult to establish.
 - Smaller companies are unable to provide integrated application. As a consequence, standardization cannot happen.
- The bioinformatics market is yet to mature and create a consistent, predictable, profitable sector for itself:
 - The industry is one with relatively low barriers to entry and increasing competition from larger established IT companies.
 - Only the fittest companies that address the standardization and integration issues will survive.

Based on a report from The Larta Institute

What is Needed Now?



Some Final Thoughts in No Particular Order

- As we gain more – and more detailed – information, the world becomes more complicated, not less so. Each new data point asks multiple new questions.
 - New questions demand better tools.
- As our tools become more powerful, we will develop novel ways to combine them to create new realities.
 - These new realities are totally unpredictable, but appear inevitable in hindsight.
- Creation of data is easy, creation of knowledge is the tricky part.
 - Aggregation, integration and synthesis are paramount.

Bioinformatics

Robert V. House, Ph.D.

301-607-5028

rhouse2@csc.com

