Challenges in Scientific Data Management

Richard Morris, Ph.D. NIH/NIAID

Age Old Story

An intelligent officer, with 10 or 12 chosen men*... might explore the whole line, even to the Western Ocean . . . and* ***return with the information acquired, in the course of two summers****.*

Jefferson's 1803 letter to Congress asking for $2,500 for the Corps of Discovery.

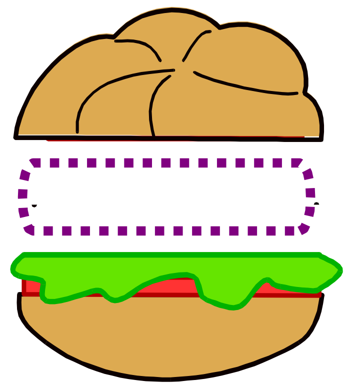
Your observations are to . . . ***comprehend all*** *the elements necessary, with the aid of the* ***usual tables*** *. . . Several copies of these as well as of your other notes,* ***should be made at leisure times,*** *and put into the care of the most trustworthy . . .* ***written on the paper of the birch****, (less liable to injury from damp than common paper).*

Jefferson, in his 1803 letter of instructions to Meriwether Lewis.

50^{th} anniversary of DNA

April 2003: completion of the sequencing of the human genome.

April 1953: Watson & Crick win Nobel Prize for description of the DNA double helix.



The Problem

Promise, Threat, and Challenge

Promise and full potential (Brent, 2000)

- abundance and rapid proliferation of genomic and proteomic data
- promised health advances from genome-based medicine
- biomedicine will be transformed by batch methods and a systems view
- emergence of biological information systems for *in silico* biology

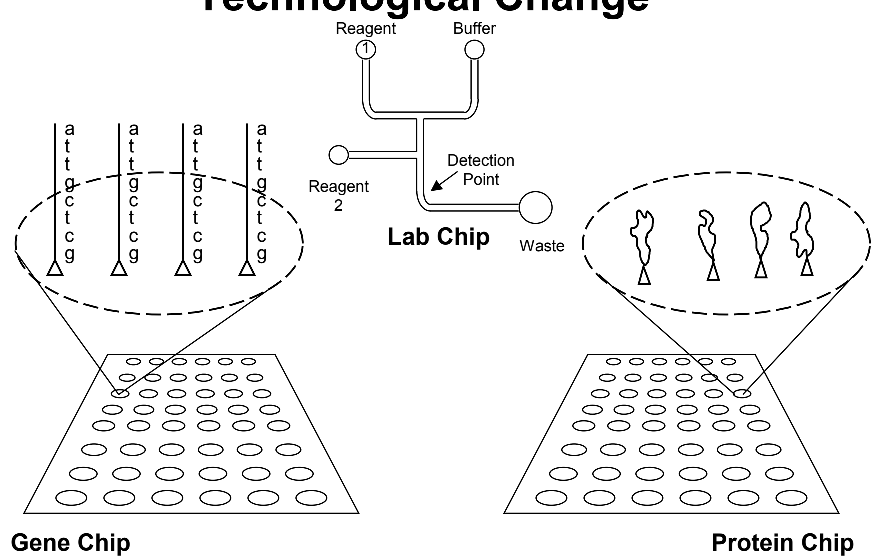
Threat (Reichhardt, 2001)

- volume -- impact of terabyte scale experimentation
- integrity -- data that are poor quality or incomparable
- access -- inability to navigate diverse, distributed reference / working data
- intractability – good data, not stored for computational purposes
- skills barrier -- scientists alienated or disenfranchised, due to paradigm shift

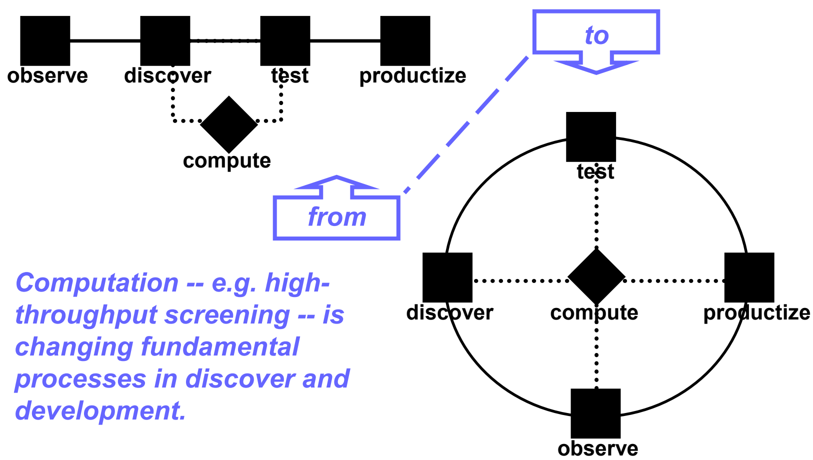
Challenges for biomedical informatics (Altman & Klein, 2002)

- representing the diversity of data
- developing exchange standards; integrating data from multiple standards
- specific tasks, e.g., managing lab data, protecting sensitive info, mining lit.
- specific questions, e.g., regulatory expression, structure-based variabililty

Technological Change

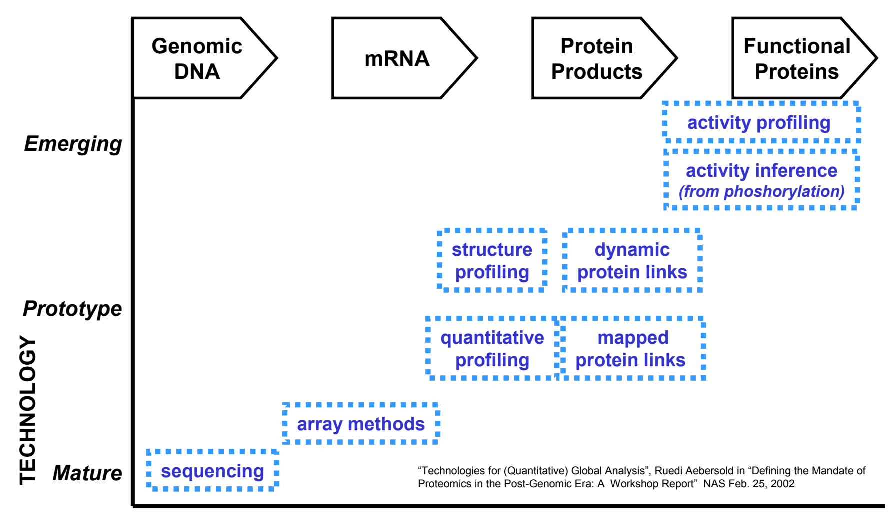


Re-invention of Discovery



Computation -- e.g. high-throughput screening -- is changing fundamental processes in discover and development.

Computation-intensive Tasks



"Technologies for (Quantitative) Global Analysis", Ruedi Aebersold in "Defining the Mandate of Proteomics in the Post-Genomic Era: A Workshop Report" NAS Feb. 25, 2002

One Group's Data Production (1.5 yrs)

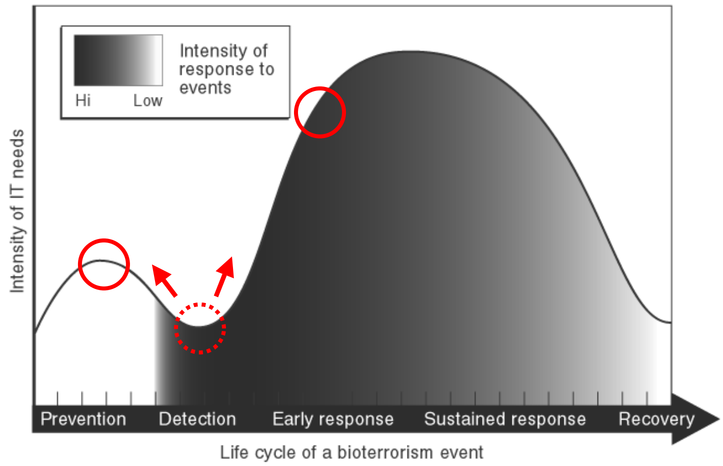
Assay Type	Specimens	Storage (in MB)
Auto Ab	742	20
SNP/Genotyping	549	5490
EliSpot	3,882	271,740
Flow Cyto / Tetramer	2,824	282
Microarray	1,672	439,000
RT-PCR	1,704	260,400
	11,373	976,932

"Informatics and Information Systems at ITN" Art Williams, 13 Jan 2002

Challenge

1. **Evidence-based Medicine:** sharing data of many types and from diverse sources; protecting confidence, privacy, property
2. **Systems Biology**: modeling complex phenomena to meaningful scale; enabling multi-disciplinary, distributed teams
3. **Biodefense**: handling large, diverse data sets; representing knowledge and discerning relevance without context
4. **Science Administration**: keeping up with workload and staying abreast of and discerning relevance of technology trends
5. **International Science**: sharing data and resources; monitoring and adapting to trends in disease, science, and policy pertinent to core mission

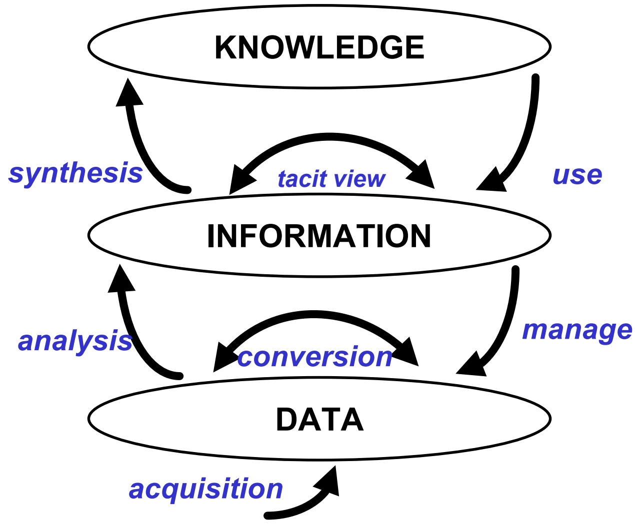
Challenge re: Biodefense



© RAND; Rippen, H.: A Framework for the IT Infrastructure for Bioterrorism, Results of the 1st Summit, 2001, DRU-2761/1-OSTP

The Solution

Solution



The Solution

PROBLEM (in detail)	**SOLUTION**
Limitations of descriptive tools	
Volume and uncertain relevance of data	
Proliferation data types	
Restricted data release and merging	
Discovery stuck in a lab; not applied	
Diversity, changeful nature of data types	
Policies re: confidence, privacy, property	
Diffusion: e.g., standards and practices	

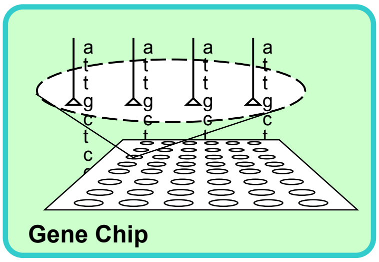
The Solution

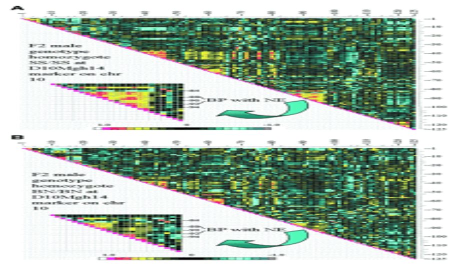
PROBLEM	SOLUTION
Limitations of descriptive tools	• iterative, tested, scaled models
Volume and uncertain relevance of data	• knowledge management
Proliferation data types	• data integration tools practices
Restricted data release and merging	• data protection / provenance
Discovery stuck in a lab; not applied	• collaboration and ubiquitous tech
Diversity, changeful nature of data types	• new data types, models, queries
Policies re: confidence, privacy, property	• codification and SW instantiation
Diffusion: e.g., standards and practices	• new tech, training and services

Envisioning Solutions

A problem well described is
. . . a problem half solved.

If data volume exceeds Moore's law . . .





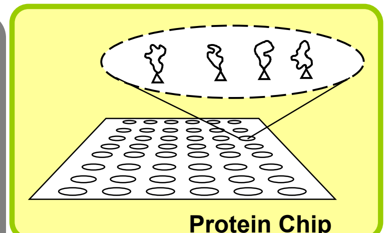
M. Stoll, et al Science, 11:2001

Myths of microarrays (Lee, 2001)

DB alternatives, not enough (Masys, 2001)

Informatics barrier in genomics (Achard, 2001)

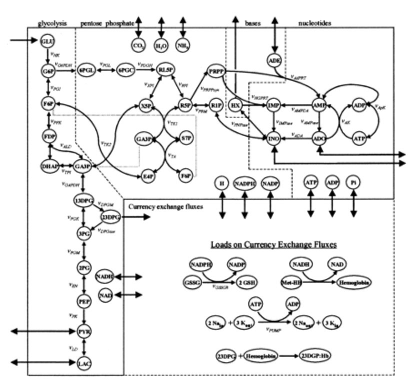
Non-specialists and experts (Gelbart, 1998)



Imperative: *Data Curation*

Biology today is quantitative and dependent on computers for:

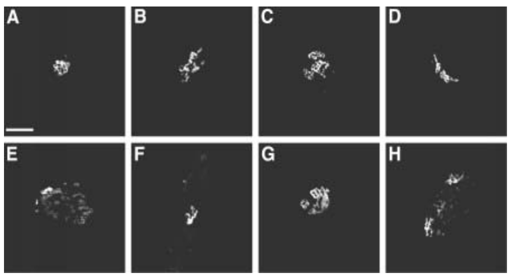
1. **Production of scientific data**
2. **Analysis of scientific data**
3. **Management of scientific data**
   - experimental design
   - data modeling
   - data curation
   - on-line collaboration
   - novel data types and queries



SJ Wiback and BO Palsson Biophysical Journal, 8:2002

If data integrity is essential . . .

Murphy Lab - Typical Image Selection. In microscopy have developed methods and classification schema for choosing a typical image from a large set of molecular-level images in cell biology. Different methods for estimating distance between images have also been explored.



A selection of the most (A-D) and least (E-H) typical giantin images, as determined using various methods (see the reference below for details). Scale bar = 10 microm.

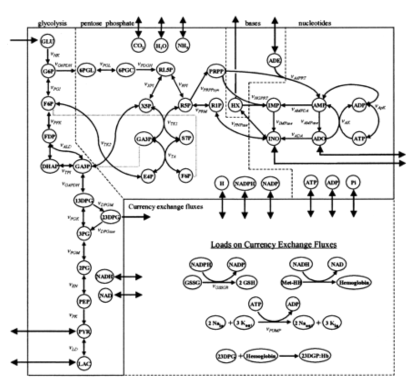
Challenges

- **Identification**
- **Quantitation**
- **Localization**
- **Disambiguation**
- **Comparison**

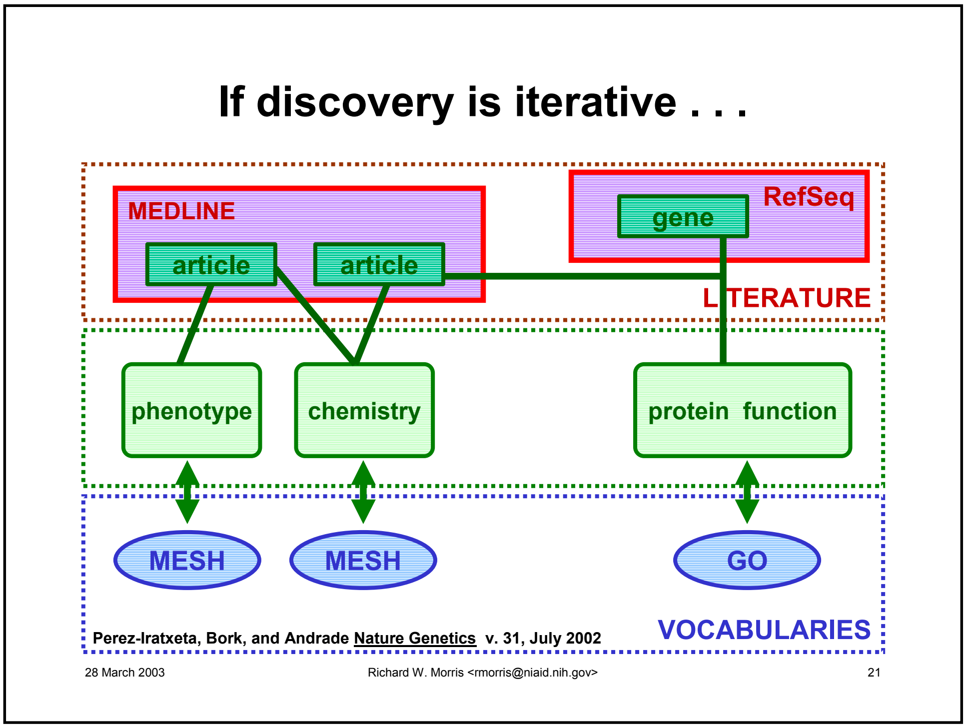
Imperative: *Manage Novel Data Types*

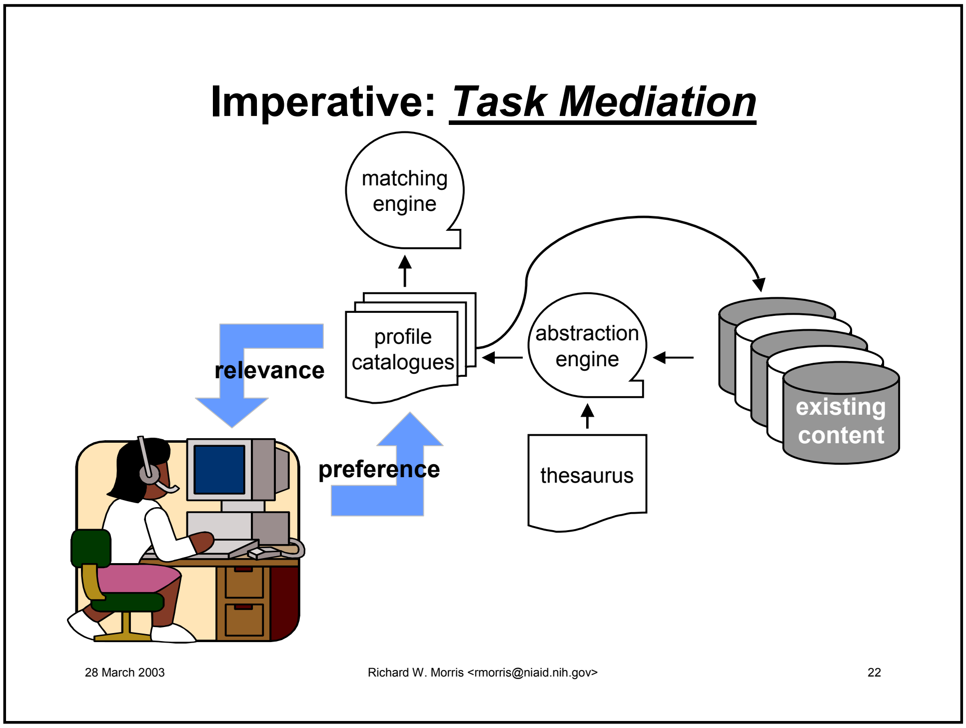
Biology today is quantitative and dependent on computers for:

1. **Production of scientific data**
2. **Analysis of scientific data**
3. **Management of scientific data**
   - **experimental design**
   - **data modeling**
   - **data curation**
   - **on-line collaboration**
   - **novel data types and queries**

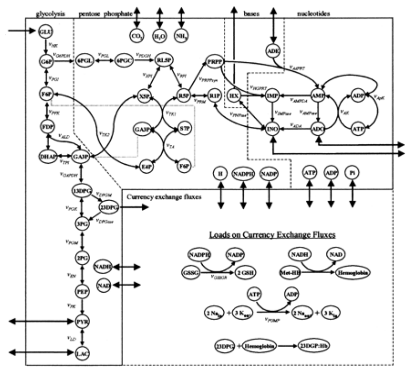


SJ Wiback and BO Palsson Biophysical Journal, 8:2002


If discovery is iterative . . .
MEDLINE
article
article
RefSeq
gene
LITERATURE
phenotype
chemistry
protein function
MESH
MESH
GO
Perez-Iratxeta, Bork, and Andrade Nature Genetics v. 31, July 2002
VOCABULARIES
28 March 2003
Richard W. Morris <rmorris@niaid.nih.gov>
21


Imperative: Task Mediation
matching engine
profile catalogues
abstraction engine
thesaurus
existing content
relevance
preference
28 March 2003
Richard W. Morris <rmorris@niaid.nih.gov>
22

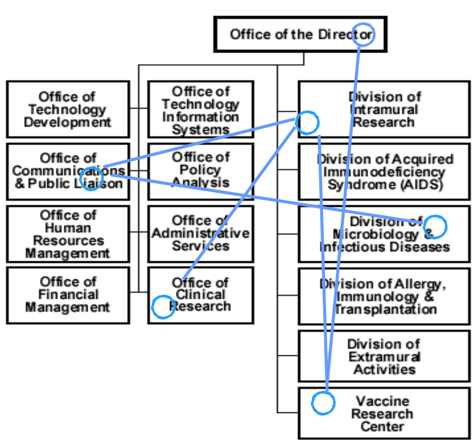
If “break the model” is the game . . .



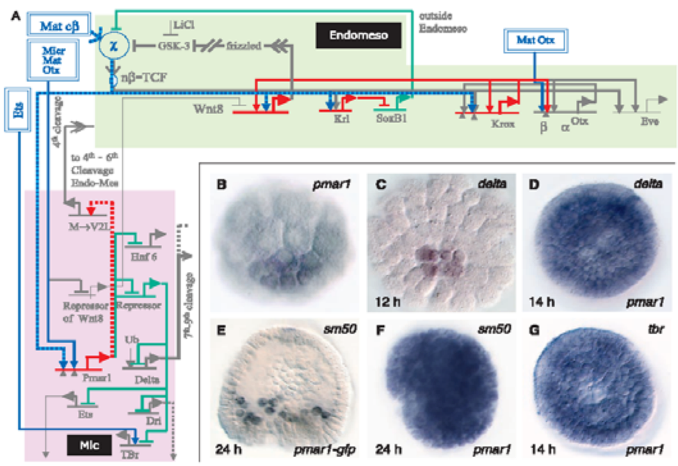
SJ Wiback and BO Palsson Biophysical Journal, 8:2002

Imperative: *Communities of Practice*

NATIONAL INSTITUTE OF ALLERGY & INFECTIOUS DISEASES



If data integration is essential . . .

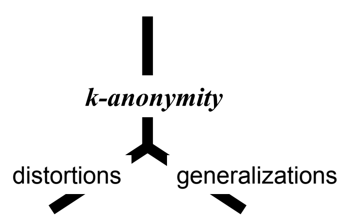


Institute of Systems Biology 2003

Imperative: ***Anonymity & Provenance***

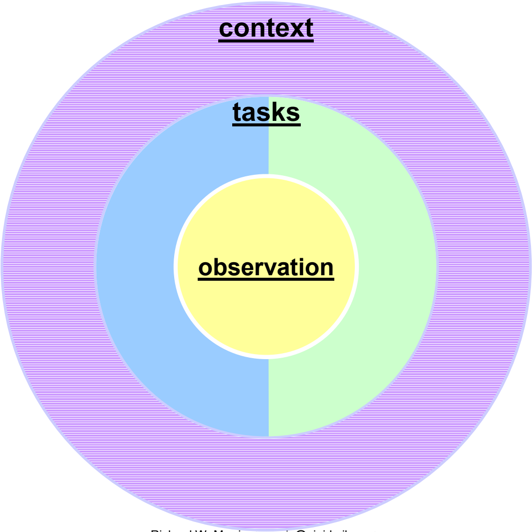
L. Sweeney Lab -- ***Datafly*** maintains anonymity in medical data by automatically generalizing, substituting and removing information as appropriate without losing many of the details found within the data.

1. A ***data holder*** declares specific attributes;
2. Groups and ranks a subset of attributes;
3. Weights those used for linking (to subject);
4. Specifies min-max anonymity levels; and
5. Ranks attributes to be distorted.



Scrub is an approach to locating and replacing personally-identifying information in unrestricted text that extends beyond straight search-and- replace procedures, while minimizing risk of confidentiality loss.

Possible Framework

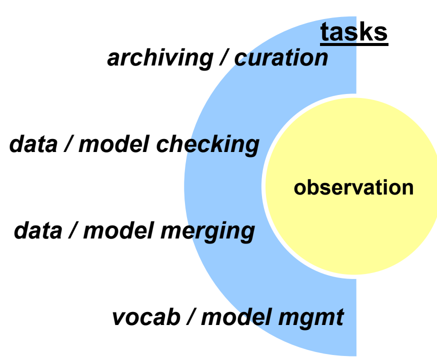


Enabling Data Capture

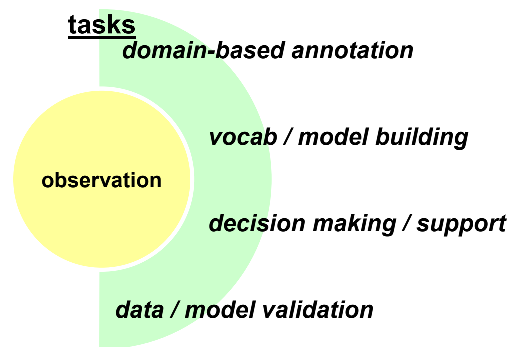
AIMS

- **accuracy of models and records (instances of model-using)**
- **rigor in reference resources (e.g., to extend observation)**
- **support iterative, multidisciplinary, collaborative processes**

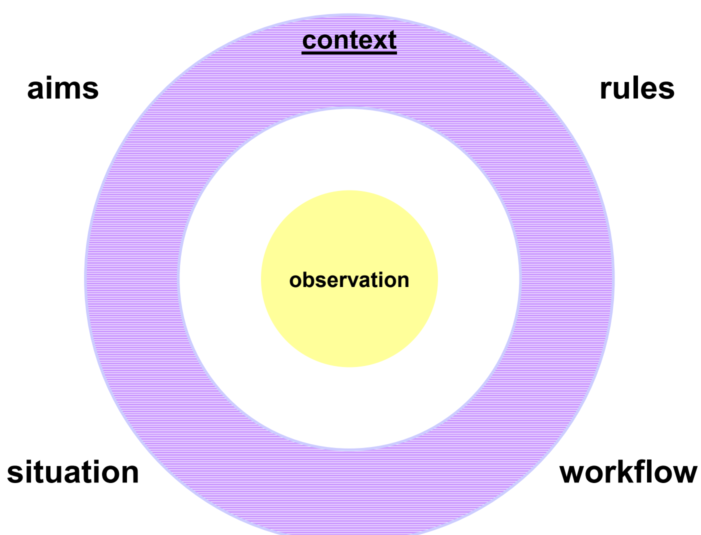
Enabling Tasks

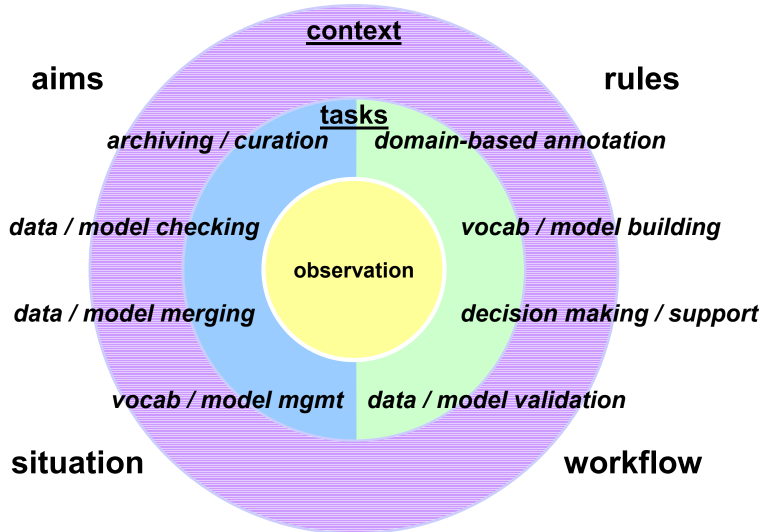


Enabling Tasks



Reifying Context




context
aims
rules
tasks
archiving / curation
domain-based annotation
data / model checking
vocab / model building
observation
data / model merging
decision making / support
vocab / model mgmt
data / model validation
situation
workflow

end

Resources

Knowledge representation (Sowa, 2000). Focus on the development of schema to support quantitative biology, computer modeling, and systems theory.

Linguistic structures (D. Searls) – Like the genome, language expresses complex relations systematically, e.g., Carroll's doublets, syzygies, segments, symmetries

Pragmatism (B. or C. Pierce) – Due to the inherent vagueness of symbols, we need to enable the process of achieving agreement btwn. interlocutors. We start with Kant's assumption that something corresponds to our generalizations ~the world.)

Relevance (D. Sperber) – Adheres to principles and invites us to question the "presumption of relevance". Not every message makes sense in every context.

Methaphor (G. Lakoff) – Concepts that are rooted in physical reality. These higher-order concepts lend coherence and utility to our symbols. They extend ability to see and agree on what we see.

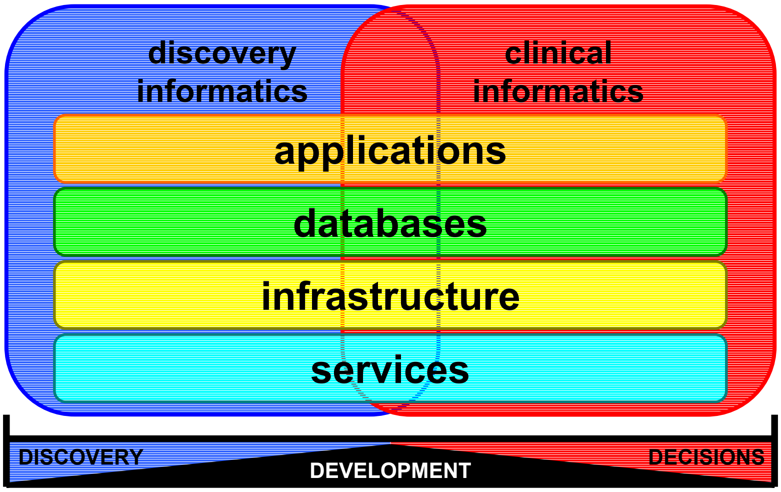
– ***metaphoric imagery*** (A. Miller) – essential element of scientific revolutions
– ***methaphoric grammar*** (A. Ortony) – relations systematically expressed
– ***methaphoric capacity*** (J. Searle) --rich meaning "Sally is a block of ice."

Bibliography

- Achard F, Vaysseix G, Barillot E. XML, bioinformatics and data integration. Bioinformatics. 2001 Feb;17(2):115-25.
- Allan A, Edenfeld D, Joyner WH, Kahng AB, Rodgers M, Zorian Y, "2001 Technology Roadmap for Semiconductors" Computer Jan 2002, pp. 42.
- Altman RB, Klein TE. Challenges for biomedical informatics and pharmacogenomics.Annu Rev Pharmacol Toxicol. 2002;42:113-33.
- Baetjer, Howard Jr. Software as Capital: an Economic Perspective on Software Engineering IEEE Computer Society: Piscataway, NJ: 1998
- Baxevanis AD. The Molecular Biology Database Collection: 2002 update. Nucleic Acids Res. 2002 Jan 1;30(1):1-12.
- Blaha, Michael. " Data Warehouses and Decision Support Systems " Computer, Dec 2001, pp. 38.
- Brent R. Genomic biology. Cell. 2000 Jan 7;100(1):169-83.
- Brusic V, Zeleznikow J, Petrovsky N. Molecular immunology databases and data repositories. J Immunol Methods. 2000 Apr 21;238(1-2):17-28.
- Ellis LB, Attwood TK. Molecular biology databases: today and tomorrow.Drug Discov Today. 2001 May 1;6(10):509-513.
- Ellis LB, Kalumbi D. Financing a future for public biological data. Bioinformatics. 1999 Sep;15(9):717-22.
- Ellis LB, Kalumbi D. The demise of public data on the web? Nat Biotechnol. 1998 Dec;16(13):1323-4.
- Gelbart WM. Databases in genomic research. Science. 1998 Oct 23;282(5389):659-61.
- Gifford DK. Blazing pathways through genetic mountains. Science. 2001 Sep 14;293(5537):2049-51.
- Heath LS, Ramakrishnan N The Emerging Landscape of Bioinformatics Software Systems. Computer, Jul 2002, Vol. 35, Number 7, pp. 41.
- Huberman BA and Hogg T "Protecting privacy while revealing data" Nature Biotechnology, vol. 20, April 2002 pp. 332.
- Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C, Gama-Castro S. The EcoCyc Database. Nucleic Acids Res. 2002 Jan 1;30(1):56-8.
- Lee JK. Analysis issues for gene expression array data. Clin Chem. 2001 Aug;47(8):1350-2.
- Lipman, DJ Input/Output of High Throughput Biology: Experience of the National Center for Biotechnology Information. In Firepower in the Lab: Automation in the Fight Against Infectious Diseases and Bioterrorism, by Layne SP, Beugelsdijk TJ, Patel CKN editors. JH Press: Wash DC 2001
- Masys DR. Database designs for microarray data. Pharmacogenomics J. 2001;1(4):232-3.
- Miller AI Imagery in Scientific Thought: Creating 20th-Century Physics The MIT Press: Cambridge: 1986.
- Reichhardt T. It's sink or swim as a tidal wave of data approaches. Nature. 1999 Jun 10;399(6736):517-20.
- Sowa, John F. Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks/Cole Thomas Learning: Albany, 2000
- Searls DB. The language of genes. Nature. 2002 Nov 14;420(6912):211-7.
- Searls DB. From Jabberwocky to genome: Lewis Carroll and computational biology. J Comput Biol. 2001;8(3):339-48.
- Searls DB. Linguistic approaches to biological sequences. Comput Appl Biosci. 1997 Aug;13(4):333-44.
- Segel LA, Jager E, Elias D, Cohen IR. A quantitative model of autoimmune disease and T-cell vaccination: does more mean less? Immunol Today. 1995 Feb;16(2):80-4.

My Own Program

BISC Program



Overview

Rationale
- computers = instruments of pattern recognition and data integration
- needed because
  - can/must make better use of legacy data
  - increasing data volume from genomics and proteomics
  - integration imperatives: systems biology & evidence-based medicine

Project management details
- contract because not funding technology development
- outsourcing for access to best practices and novel capacity
- co-PIs – immunologist and engineer – to blend perspectives
- contractors:
  - Team 1: RTI, Duke, IBM
  - Team 2: Grumman, UTHSC-SW, Kevric

Lessons learned
- Immune Histocompatibility Network: working vs. public data
- Pediatric Renal Transplant: disaggregating HS and IP issues
- Immune Tolerance Network: importance of experimental design

What problem are we trying to solve?

Current BISC focus
- Autoimmunity Centers of Excellence
- Immune Tolerance Network
- Inner-City Asthma Consortium
- International Histocompatibility Working Group
- Cooperative Clinical Trials in Pediatric Renal Transplantation

Possible focus for BISC in the future
- Multiple Autoimmune Diseases Genetics Consortium
- Population Genetic Analysis Program: Immunity to Vaccines/Infections
- Biodefense Partnerships: Vaccines, Adjuvants, Therapeutics, Diagnostics, Resources
- Genomics and Proteomics of Transplantation
- Translational Research on Human Immunology and Biodefense
- Immune Epitope Database and Analysis Program
- Imaging Technologies and Mathematical Models of Immunity

Possible other influences
- Rethinking data sharing means changes for data coordinators
- Future administrative improvements: Enterprise Information Management System
- Desire to draw on archives for developing new training initiatives