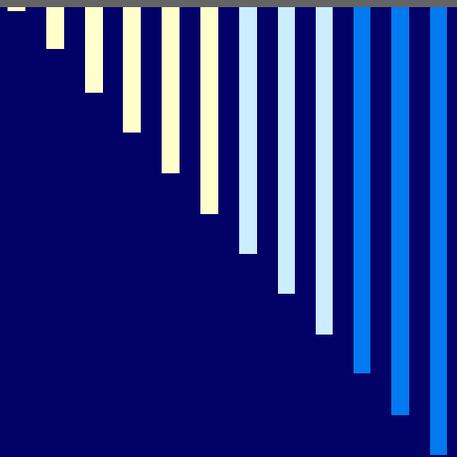


NATIONAL
CANCER
INSTITUTE[®]



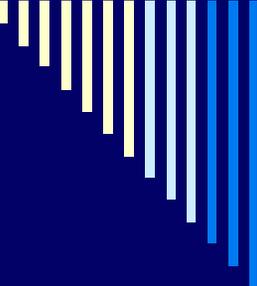
Cancer Imaging Informatics Workshop

25-27 September 2002

Bethesda, MD

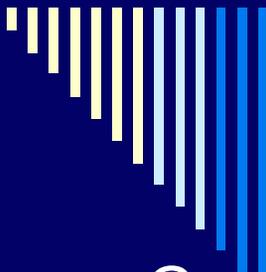
NATIONAL CANCER INSTITUTE

Biomedical Imaging Program



Outline

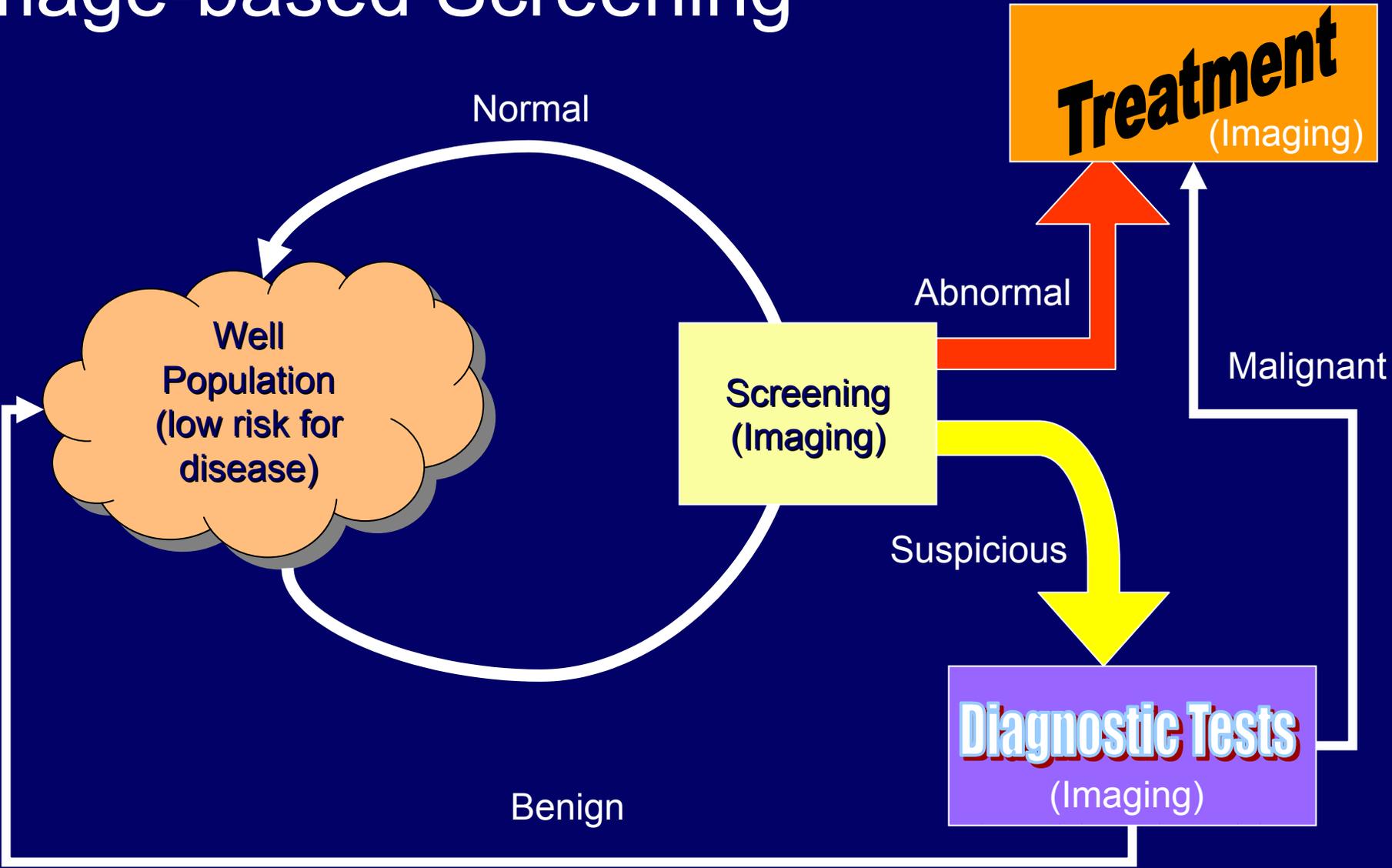
- Why?
 - Who?
 - What?
 - How?
 - Examples...
-

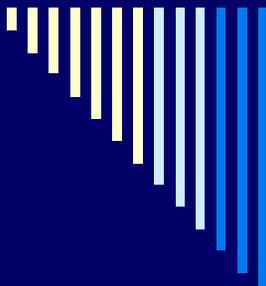


Why now?

- ❑ Growing importance of imaging in NCI's mission
 - ❑ Cancer screening by imaging
 - ❑ Proliferation of image databases (digital radiology departments; integrated healthcare enterprise)
 - ❑ Digital image teaching files for training and certification
 - ❑ Image-guided interventions (for cancer diagnosis and therapy)
 - ❑ Emerging technologies to support cancer diagnosis and therapy (CT, MRI, US, optical, CAD, microarrays, ...)
-

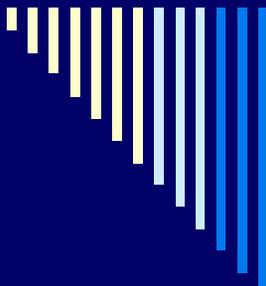
Image-based Screening





Areas of concern

- Data acquisition capabilities increase faster than infrastructure to organize and use the information we gather
 - Disconnection between in vivo images and “mainstream” biological knowledge resources (e.g., genome and text databases, among others)
 - Cancer imaging science appears to lag behind neuroscience and genomics/proteomics in the integrated information infrastructure
-



Workshop Objectives

1. To understand how cancer imaging data can best be managed to fully exploit its potential utility and synergy with existing databases (sequences, arrays, and text)
2. To promote research to predict risk, detect and diagnose cancer, select and tailor treatments, predict outcomes and follow therapy using image repositories, biological databases, and software tools
3. To accelerate the process of testing new agents and therapies using imaging as a surrogate marker of outcome, and employing standards for eClinical Trials
4. To build image repositories that are generally useful for testing and certification of diagnostic agents, especially software post-processing of cancer images for computer aided diagnosis

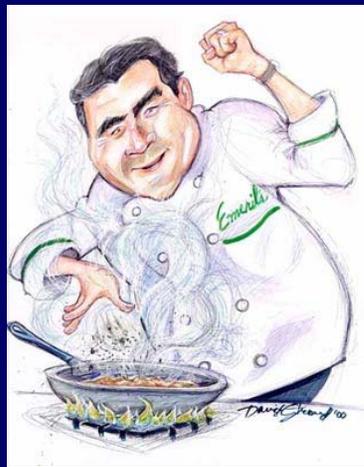
NATIONAL CANCER INSTITUTE

Biomedical Imaging Program

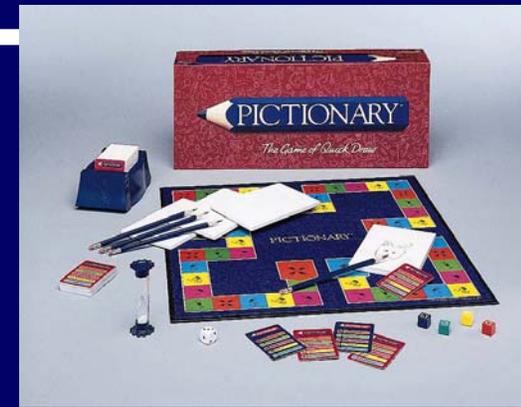
Potential consequences of status quo

- Failure to integrate data sources has serious consequences for cancer imaging science and related applications
 - Lack of tools
 - Delay in translating technical developments into clinical applications
 - Inability to address many fundamental questions
 - Understanding the cancer phenotype and its behavior, especially related to therapy
 - Barrier to innovation and marginalization of imaging





Who are you?



- World's experts in biological databases, image repositories, clinical image management, radiotherapy (image-guided) quality assurance, large database architecture and applications, grid & middleware technologies, cancer ontologies, and non-image cancer data management
 - Physicians, engineers, physicists, computer scientists, neuroscientists, ...
-

Who is here?



DICOM

CDISC

RSNA / MIRC

QARC

3D-QA Ctr

RCET

LIDC

ACRIN

NDMA

BIRN

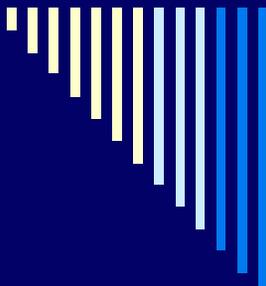
Industry (Many)

WEAR

Federal Agencies (NSF, DOD, NIST, DOE, and especially FDA)

NIH (NCI, CIT, NLM, NCRR, NIA, NINDS, NIBIB, ...)

Academia -- NCI-sponsored Cancer Centers
and Cooperative Groups

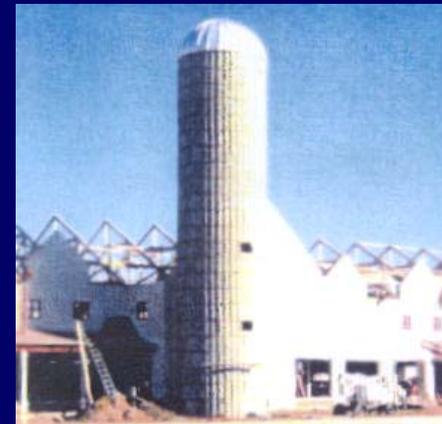


How was the workshop organized?

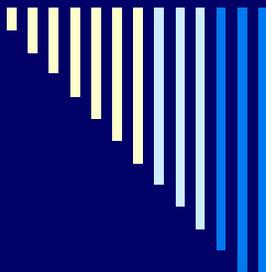
- Diversity of ideas; challenge the imaging science / image management community
 - Cancer focus
 - Motivated by cancer screening by imaging and concern about image repositories
 - Avoid loss of in vivo cancer phenotype information
 - Move cancer imaging science into the mainstream – to better reflect its growing importance in screening, diagnosis, treatment and followup of clinical cancer patients
 - Especially at the clinical trial level
 - And to move agents, devices, procedures from lab to clinic more quickly
-



Topics

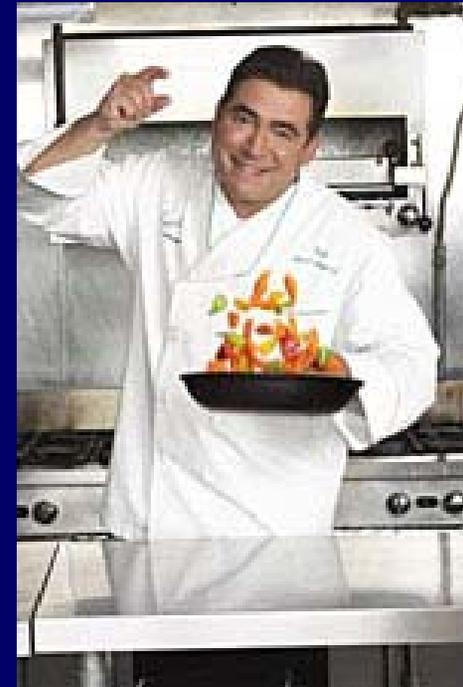


- NCBI and the Entrez system
- Very large databases in science
- Cancer (and other) imaging databases
- BIRN
- Standards and the FDA electronic submission process
- WEAR



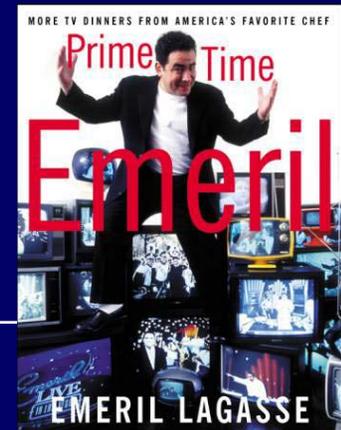
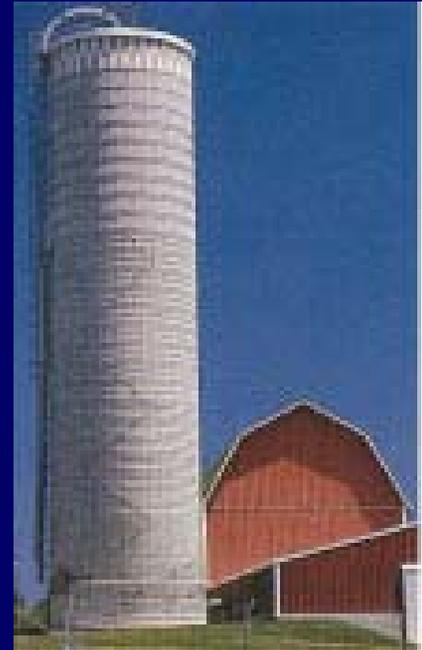
What do we want to accomplish?

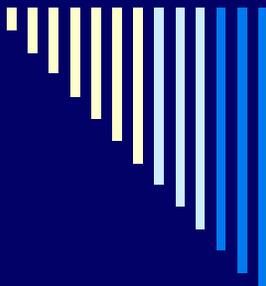
- A new vision of cancer imaging archives
- What should we do?
- Where are the opportunities?
- What have we missed?
- Which is most important, if we must choose one?
- Should we work independently, or seek existing group(s) to collaborate with us?



Friday morning is important!

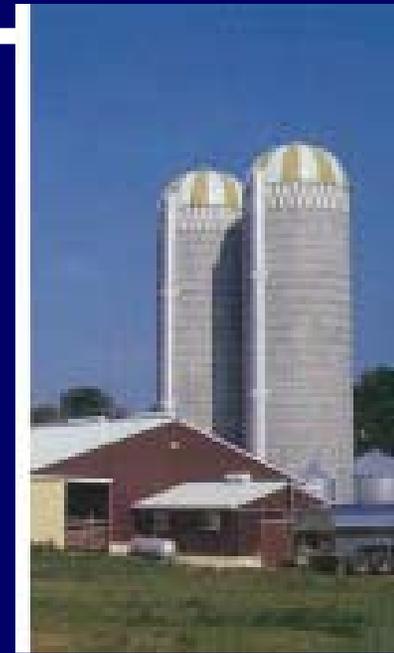
- We want your ideas. Participate!
- Input...input...input
- Try to provide a few key items and as much secondary detail as you care to provide.
- Not necessary to achieve consensus – it is sufficient to simply help us understand the controversies and alternatives
 - How can we resolve these issues?
- After Friday, we will continue to be interested in your input. If you want to add something later, please contact us.





Maintaining contact

- NIH listserver: Archive-Comm-L at <http://list.nih.gov>
- You have our e-mail addresses
- We will publish reports (including web dissemination) via <http://cancer.gov/bip>
- Remember: National Cancer Institute Biomedical Imaging Program office here in Bethesda, MD



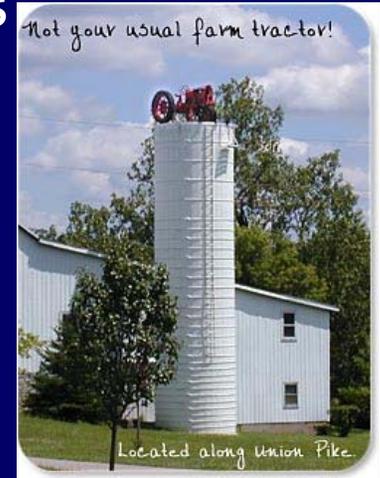
NATIONAL CANCER INSTITUTE

Biomedical Imaging Program

Selection of speakers

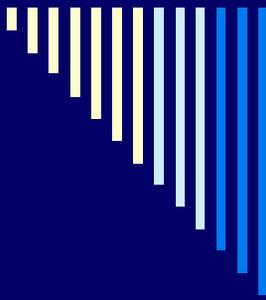
- All are world's experts in their respective areas
- Many (?most) have never have met, despite strong common interests and potential complementary expertise
- Integrative, collaborative, broad vision
- All are working at the frontiers of technology and/or cancer imaging applications

The Past



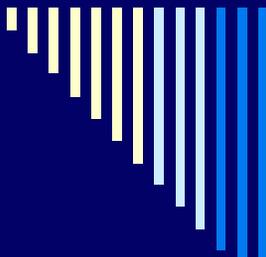
The Future





Definitions

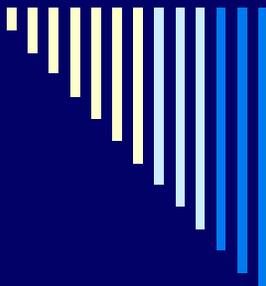
- **in·for·mat·ics** Information science
[informat(ion) + -ics.]
 - **bi·o·in·for·mat·ics** The use of computers in solving information problems in the life sciences, mainly, it involves the creation of extensive electronic databases on genomes, protein sequences, etc. Secondarily, it involves techniques such as the three-dimensional modeling of biomolecules and biologic systems.
-



Bio-informatics

- **Bioinformatics** is conceptualizing bioscientific data and applying "**informatics techniques**" (derived from disciplines such as applied mathematics, computer science and statistics) to **understand and organize the information** associated with the data on a *large scale*.

Luscombe, N. M., Greenbaum, D., and Gerstein, M. (2001).
What is bioinformatics? A proposed definition and overview of the field.
Method. Inform. Med., 40(4):346-258.



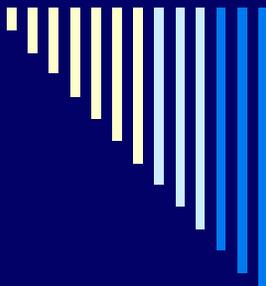
Neuroinformatics

- Neuroinformatics is:

neuroinformatics = neuroscience + informatics

... combining neuroscience and informatics research to develop and apply advanced tools and approaches essential for a major advancement in understanding the structure and function of the brain.

Neuroinformatics research is uniquely placed at the intersections of medical and behavioral sciences, biology, physical and mathematical sciences, computer science, and engineering. The synergy from combining these approaches will accelerate scientific and technological progress, resulting in major medical, social, and economic benefits.



Cancer imaging informatics

- **Cancer imaging informatics** is conceptualizing cancer image and related scientific data, and applying "**informatics techniques**" (derived from disciplines such as applied mathematics, computer science and statistics) to **understand** and **organize** the **information** associated with the data on a *large scale*.
-

John S. Silva · Marcus J. Ball · Christopher G. Chute
Keith S. Douglas · Curtis P. Langford
Jayce C. Noland · William L. Scherlis · (more)

Foreword by Richard D. Klausner

Cancer Informatics

Essential Technologies for
Clinical Trials



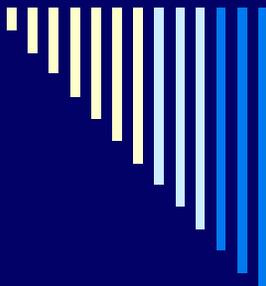
HEALTH INFORMATICS SERIES

Cancer Informatics Essential Technologies for Clinical Trials

2002 - 377 pp. 62 figs. Hardcover

John S. Silva, et al.

Springer-Verlag: New York



NCBI and the Entrez System

- Alex E. Lash, MD
 - National Center for Biotechnology Information
 - National Library of Medicine
 - National Institutes of Health
 - Bethesda, MD
-

Cancer Imaging Informatics

**Michael W. Vannier, Edward V. Staab,
and Laurence C. Clarke**

National Cancer Institute
Biomedical Imaging Program
Bethesda, MD

vannierm@mail.nih.gov





Neuroimaging Databases

The Governing Council of the Organization for Human Brain Mapping (OHBM)

These are comments written by the Governing Council of the Organization for Human Brain Mapping (OHBM), the primary international organization dedicated to neuroimaging research. The purpose of these comments is to identify and frame issues concerning data sharing within the neuroimaging community. Data sharing has become an important issue in most fields of science. The neuroimaging community is no exception, and it clearly perceives potential benefits in such efforts, as have been realized in other fields such as genomics. At the same time, such efforts can be costly (both in time and expense), and there are important factors that differentiate brain imaging from other fields and that pose specific challenges to the generation of useful neuroimaging databases. These include the rapid pace of change in brain imaging technologies; the complexity of the variables that must be specified to meaningfully interpret the results (such as the method of image acquisition, behavioral design, and subject characteristics); and concerns about participant confidentiality. These issues are outlined with the goal of framing and promoting a public discussion of the benefits and risks of data sharing, which can inform the field of neuroimaging as well as others that face similar challenges.

that obscures their full complexity. The data themselves take a variety of forms and typically are not accessible for widespread sharing and use. Making neuroimaging data more accessible for sharing would facilitate the comparison of findings across laboratories, to allow better assessment of the reliability of methods and reproducibility of results; encourage meta-analyses that explore phenomena that are not apparent in individual data sets; and give investigators who do not have access to neuroimaging facilities the opportunity to conduct research using existing data. All of these are more efficient uses of neuroimaging data, which are relatively expensive to collect.

Some challenges. These potential benefits and the success of data sharing in other com-

Appendix: Outline of Issues Related to Neuroimaging Databases

1. Data contents
 - a) Imaging data
 - b) Metadata
 - c) Data import and export
 - d) Data quality
2. Data access
3. Data ownership, credit, and confidentiality
4. Database structure
5. Interactions with the community



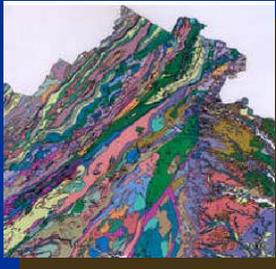
A Geoscientist's Information Integration Problem

*What is the distribution and U/Pb zircon ages of A-type plutons in VA?
How about their 3-D geometry?
How does it relate to host rock structures?*

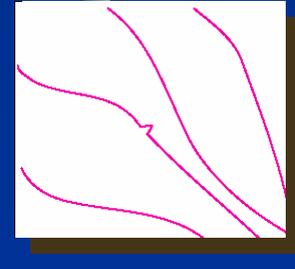
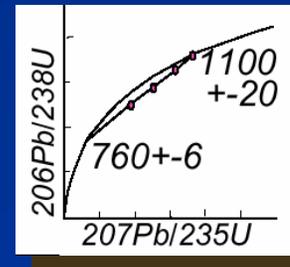
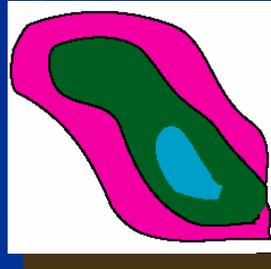


**Information
Integration**

**“Complex
Multiple-Worlds”
Mediation**



SiO ₂	72.22
CaO	0.62
K ₂ O	4.88
Ga	21.1
Sr	72.6



*Geologic Map
(Virginia)*

GeoChemical

*GeoPhysical
(gravity contours)*

*GeoChronologic
(Concordia)*

*Foliation Map
(structure DB)*

A Neuroscientist's Information Integration Problem

What is the cerebellar distribution of rat proteins with more than 70% homology with human NCS-1? Any structure specificity? How about other rodents?

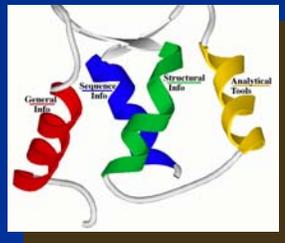


**Information
Integration**

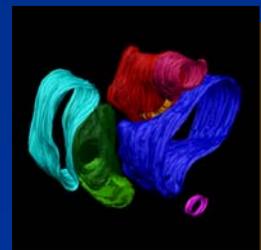
**“Complex
Multiple-Worlds”
Mediation**



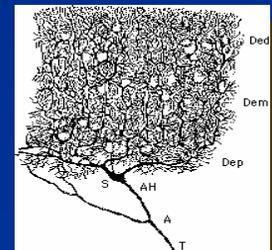
*protein localization
(NCMIR)
(NCMIR)*



*sequence info
(CaPROT)
(CaPROT)*



*morphometry
(SYNAPSE)
(SYNAPSE)*



*neurotransmission
(SENSELAB)
(SENSELAB)*

The Problem

- Most biological knowledge is stored in databases
- Creation, expansion, and integration of these databases has become central to the advancement of biology and medicine
- Many databases are isolated “silos”
- Medical imaging is unique in that there are few publicly accessible databases, links to mainstream biological knowledge collections are absent, and there are few (software) tools available that allow you to use them



Why are imaging databases important?

- Images contain the phenotype
- In other fields (e.g., astronomy, geoscience, neuroscience, ...), the integration of image (and other) databases has had a revolutionary effect
 - Coalescence of the scientific community
 - Open the field to rapid technological advancement
 - Possible to address questions that could not otherwise be answered (e.g., trans-species, multiscale, ad hoc group collaboration)



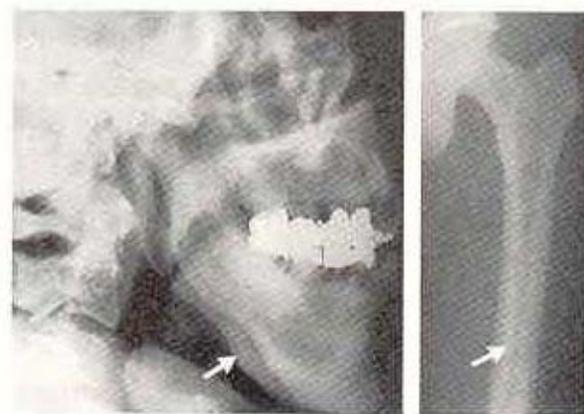
This Week in the Journal

May 16, 2002

High Bone Density Due to an *LRP5* Mutation

Osteoporosis can be caused by a loss-of-function mutation in the gene for low-density lipoprotein receptor-related protein 5 (*LRP5*). In this study, the authors, reasoning that a gain-of-function mutation in the same gene might be associated with high bone density, performed biochemical and genetic analyses of a kindred with high bone density, a prominent mandible, and torus palatinus. Genetic analysis revealed an *LRP5* mutation, the substitution of valine for glycine at codon 171, that segregated with the clinical findings. In vitro studies demonstrated that the defect in *LRP5* resulted in changes in signaling events with other molecules that normally interact with this receptor-related protein, resulting in increased bone density.

The findings suggest that molecules that interact with LRP5 may provide targets for the treatment of osteoporosis.



The New England Journal of Medicine

Copyright © 2002 by the Massachusetts Medical Society

VOLUME 346

MAY 16, 2002

NUMBER 20



HIGH BONE DENSITY DUE TO A MUTATION IN LDL-RECEPTOR-RELATED PROTEIN 5

LYNN M. BOYDEN, PH.D., JUNHAO MAO, PH.D., JOSEPH BELSKY, M.D., LYLE MITZNER, M.D., ANITA FARHI, R.N.,
MARY A. MITNICK, PH.D., DIANQING WU, PH.D., KARL INSOGNA, M.D., AND RICHARD P. LIFTON, M.D., PH.D.

ABSTRACT

Background Osteoporosis is a major public health problem of largely unknown cause. Loss-of-function mutations in the gene for low-density lipoprotein receptor-related protein 5 (*LRP5*), which acts in the Wnt signaling pathway, have been shown to cause osteoporosis-pseudoglioma.

Methods We performed genetic and biochemical analyses of a kindred with an autosomal dominant syndrome characterized by high bone density, a wide and deep mandible, and tibiae relatively

OSTEOPOROSIS is a major public health problem, and its prevalence is increasing.¹⁻³ In the United States, nearly 1 million fractures occur annually in people over the age of 65 years, the majority of which are due to osteoporosis.^{1,4} Osteoporotic fractures are associated with substantial morbidity, and the estimated rate of death in the first year after a hip fracture is 25 to 30 percent.^{5,6}

Bone mass, a major determinant of the risk of os-

gle propeller of the low-density lipoprotein (LDL) receptor in humans, mice, rats, pigs, hamsters, and rabbits. Moreover, glycine is also found at this position in the first propeller of the *Drosophila melanogaster* LDL-receptor-related protein homologue, *arrow*. In addition, glycine is present at this position in a wide range of other YWTD propellers, including those in other LDL-receptor-related proteins, as well as those in the epidermal growth factor precursor, the very-low-density lipoprotein receptor, and the vitellogenin receptor in fruit flies and mosquitos (protein sequences are available at <http://www.ncbi.nlm.nih.gov/entrez>). The evolutionary conservation of this glycine residue is strong evidence of the functional importance of its mutation in our kindred.

Molecular Studies

If this mutation indeed causes gain of LRP5 function and increased Wnt signaling, downstream target genes in the Wnt signaling pathway should show increased expression in vivo. A direct transcriptional target of Wnt signaling is the extracellular matrix protein fibronectin.³¹ Fibronectin levels were markedly elevated in the affected members of our kindred, with

es an autosomal dominant disorder characterized by high bone density, torus palatinus, and a wide, deep mandible.

Our in vitro and in vivo studies show that the *LRP5*_{V17I} mutation increases Wnt signaling. The mutation impairs antagonism of Wnt signaling by Dkk-1 in vitro, and the levels of fibronectin, a downstream target of Wnt signaling, are increased in vivo in patients with this mutation. These findings indicate that unopposed Wnt signaling due to loss of action of a

Protein sequences are available at

ENTREZ

It is striking that the same mutation is associated with nonsyndromic high bone mass in one family and syndromic high bone mass in the other. These findings suggest that alleles of other genes or environmental factors influence phenotypic manifestations of the mutation and that other phenotypes in kindreds with autosomal dominant high bone mass may also arise from

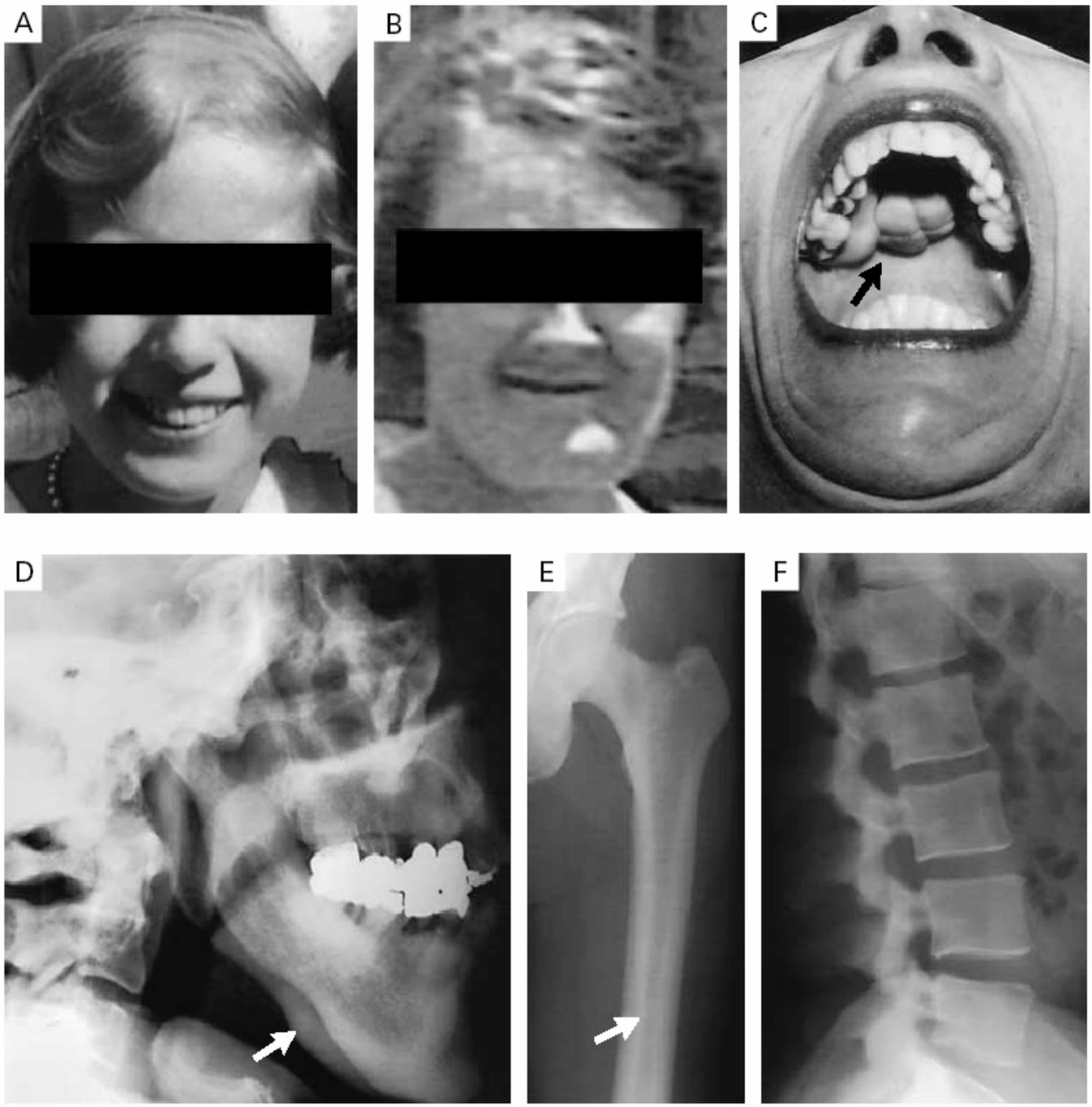


Figure 1

Figure 1. Clinical and Radiographic Features of Affected Members of the Kindred.

Entrez-PubMed - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media Refresh Print Mail Stop

Address <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi> Go Links >>

NCBI PubMed National Library of Medicine NLM

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books

Search PubMed for Go Clear

Limits Preview/Index History Clipboard Details

- Enter one or more search terms, or click [Preview/Index](#) for advanced searching.
- Enter [author names](#) as smith jc. Initials are optional.
- Enter [journal titles](#) in full or as MEDLINE abbreviations. Use the [Journal Browser](#) to find journal titles.

PubMed, a service of the National Library of Medicine, provides access to over 12 million MEDLINE citations back to the mid-1960's and additional life science journals. PubMed includes links to many sites providing full text articles and other related resources.

Bookshelf Additions

Developmental Biology, 6th ed.
by SF Gilbert

Surgical Treatments - Evidence Based and Problem-Oriented
edited by RA Holzheimer & JA Mannick

and NCBI's **Genes and Disease.**

These three resources are now available for interactive searches on Bookshelf.

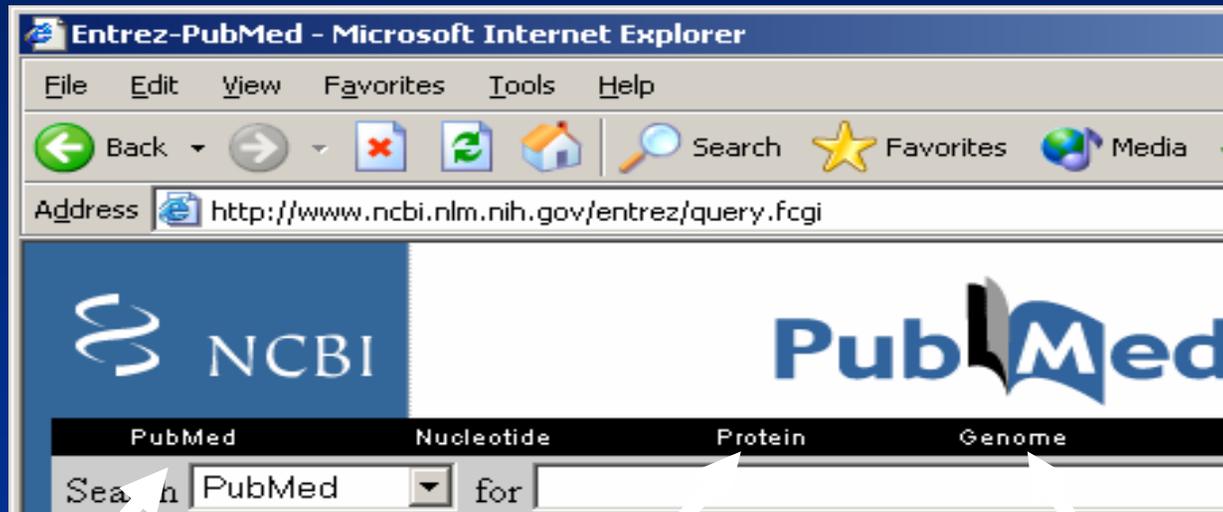
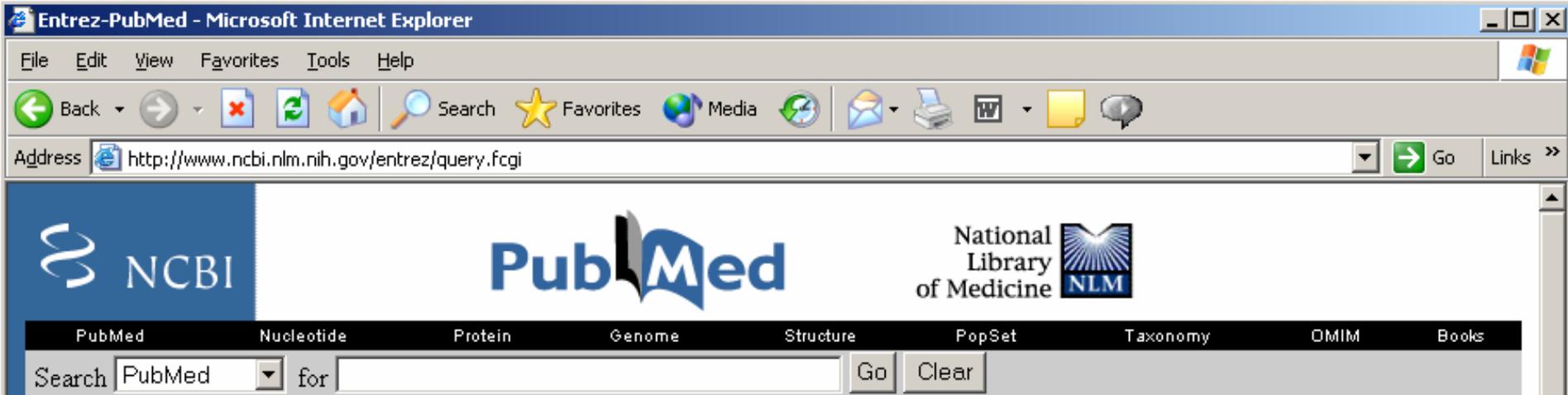
Try LinkOut

 [LinkOut](#) is a jumping-off point from PubMed citations to relevant resources on the web, such as, full-text articles, library holdings, commentaries, author biographies, practice guidelines, consumer health information, and research tools. [Example](#) LinkOut resources are available.

About Entrez
Text Version
Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities
PubMed Services
Journal Browser
MeSH Browser
Single Citation
Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
Cubby
Related Resources
Order Documents
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central
Privacy Policy

Done Internet

Entrez
PubMed

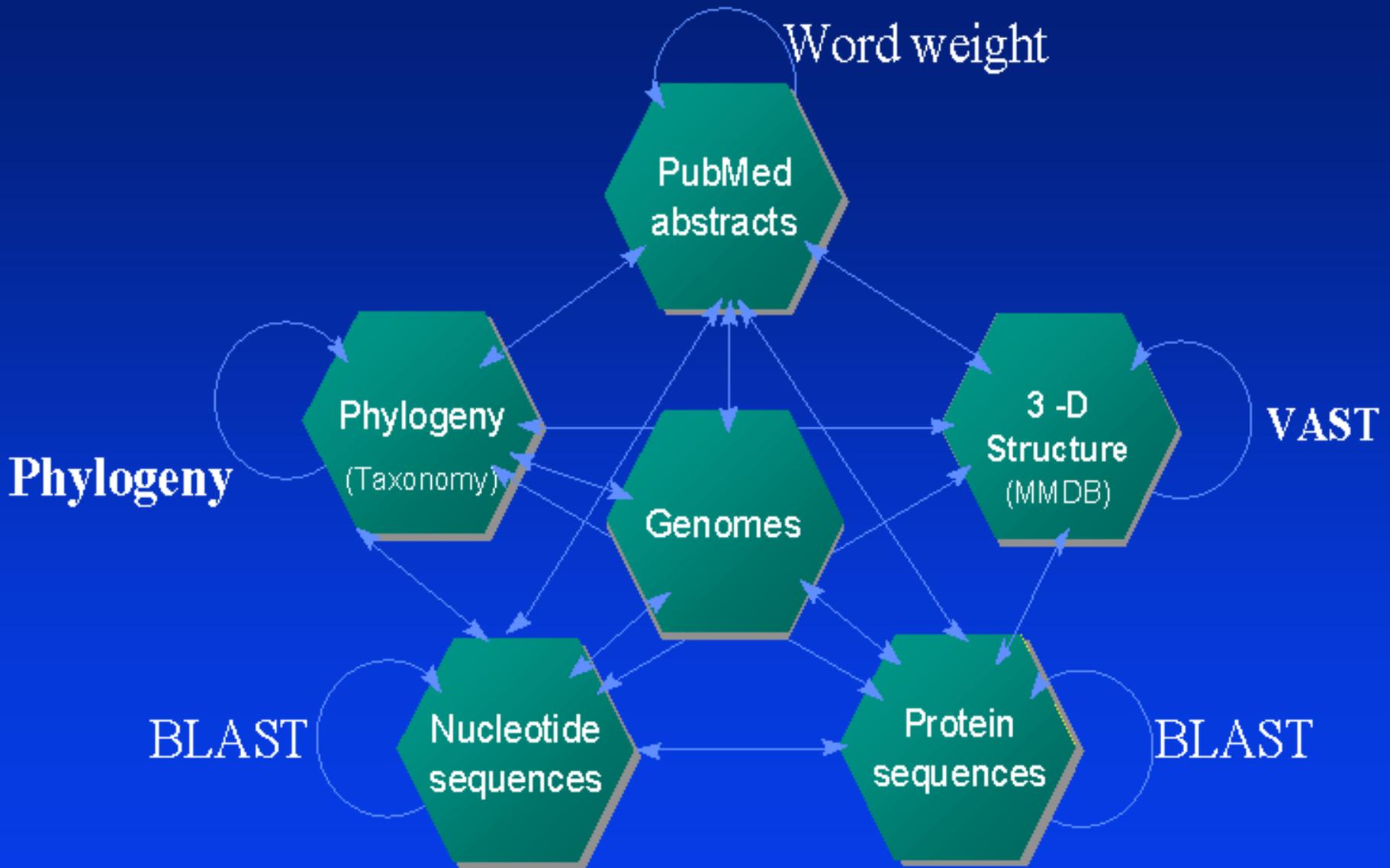


Medline

PDB =
Protein
Data Base

Genbank =
All known sequences

Entrez: Neighboring and Hard Links

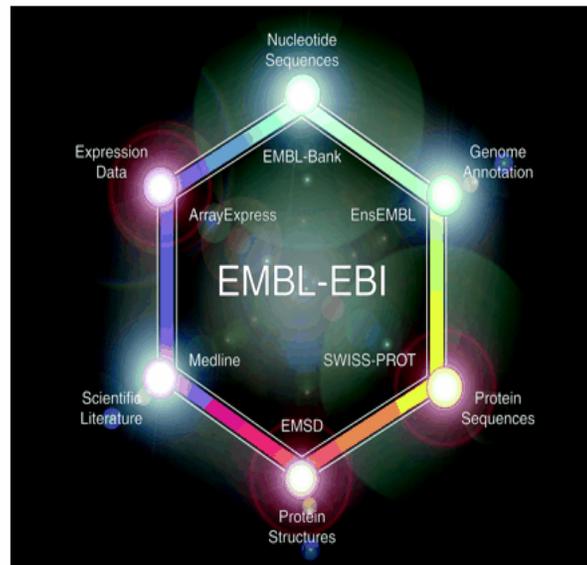


Volume 30, No. 1 – January 2002 – Database Issue

Nucleic Acids Research

[HOME](#) [HELP](#) [FEEDBACK](#) [SUBSCRIPTIONS](#) [ARCHIVE](#) [SEARCH](#)

About the Cover



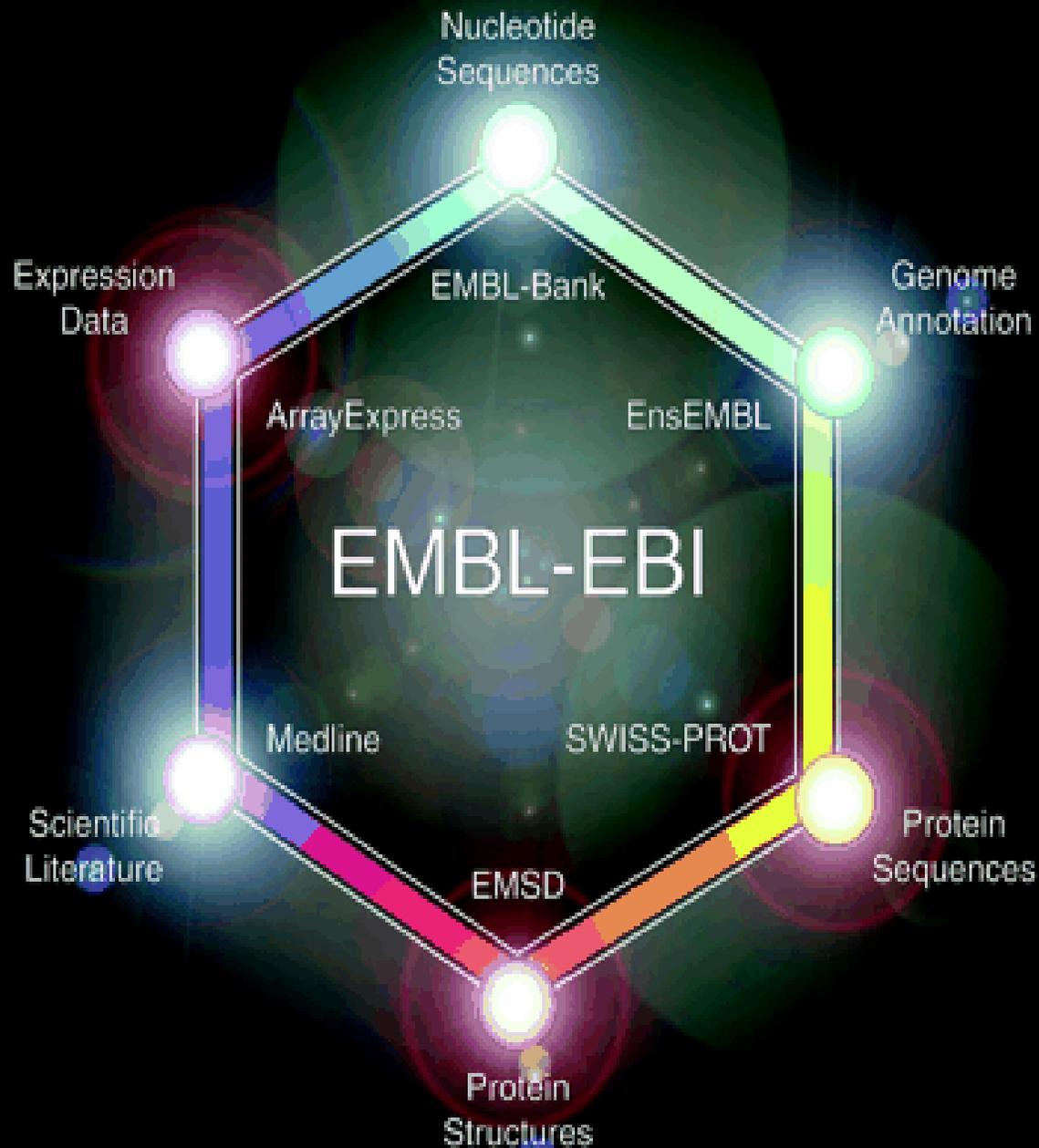
Nucleic Acids Research

Cover: Design by Guenter Stoesser and Tom Oinn. The EMBL-Outstation European Bioinformatics Institute (EBI) is a centre for research and services in Bioinformatics. The Institute manages databases of biological data including nucleotide and protein sequences, macromolecular structures, microarray and other types of data. For further details see the following papers in this issue. Stoesser *et al.*, The EMBL Nucleotide Sequence Database, *Nucleic Acids Res.* (2002) **30**, [21-26](#); Hubbard *et al.*, The Ensembl genome database project, *Nucleic Acids Res.* (2002) **30**, [38-41](#); Fredman *et al.*, HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources, *Nucleic Acids Res.* (2002) **30**, [387-391](#).

[\[Table of Contents\]](#)

[HOME](#) [HELP](#) [FEEDBACK](#) [SUBSCRIPTIONS](#) [ARCHIVE](#) [SEARCH](#)

Copyright © 2002 Oxford University Press.



Cover

Nucleic
Acids
Research

Jan 2002

Annual
Database
Issue

This year,
112
databases
were
reported.

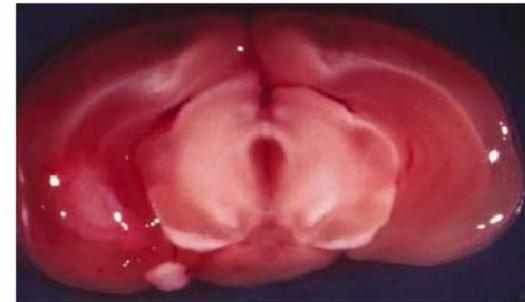
news feature

Information overload

The tide of genetic data threatening to swamp researchers has led a 'data warehousing' firm to tune in to science. Carina Dennis charts its move from airlines and banks to biology.



Checking in: a powerful database used to track airline reservations will soon be tracing gene expression in mouse brains (right).



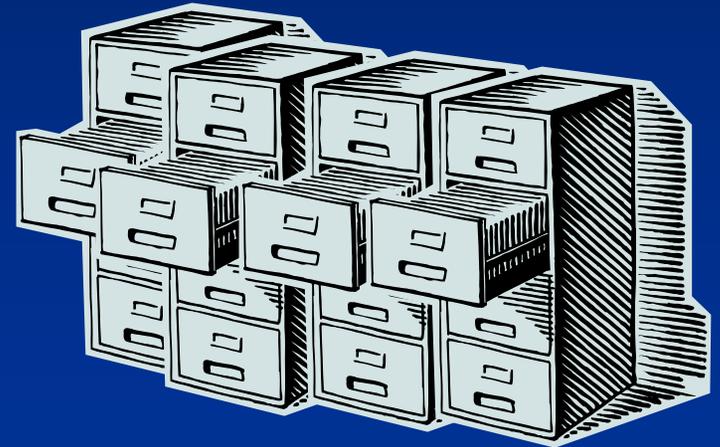
What do biologists interested in gene expression have in common with airlines, banks and food retailers? Data, data ... and yet more data. A range of high-throughput techniques such as DNA microarray analyses are providing biologists with sets of data that dwarf anything they have ever dealt with before. And in this rising tide of information, a company that supplies data 'warehouses' has spied an opportunity.



Silo of Data



=



Silos

Text

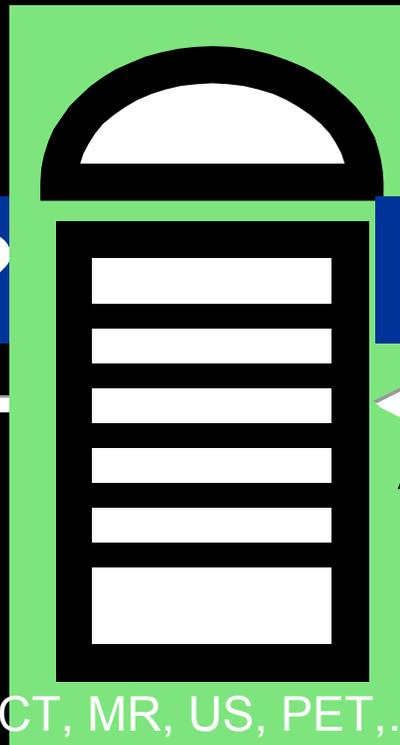
Medical
Literature



Full Text

Dicom

Diagnostic
Images



CT, MR, US, PET, ...

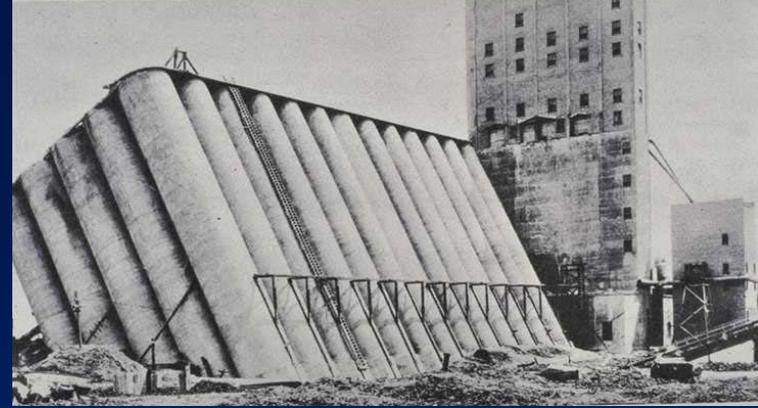
HL-7

Clinical
Records



H

Data Silos



=





Building a Nation from a Land of City States

Lincoln D. Stein

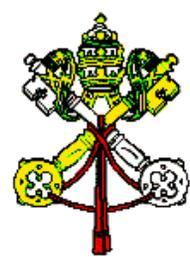
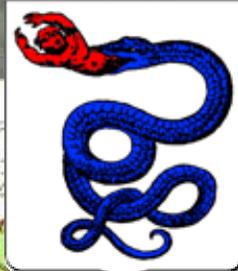
Cold Spring Harbor Laboratory

Italy in the Middle Ages



Italy in the Middle Ages

ITALY
in the XI Century



Effect on Trade & Technology

- ◆ Italian city states had
 - Different legal & political systems
 - Different dialects & cultures
 - Different weights & measures
 - Different taxation systems
 - Different currencies
- ◆ Italy generated brilliant scientists, but lagged in technology & industrialization

Italy, 1796

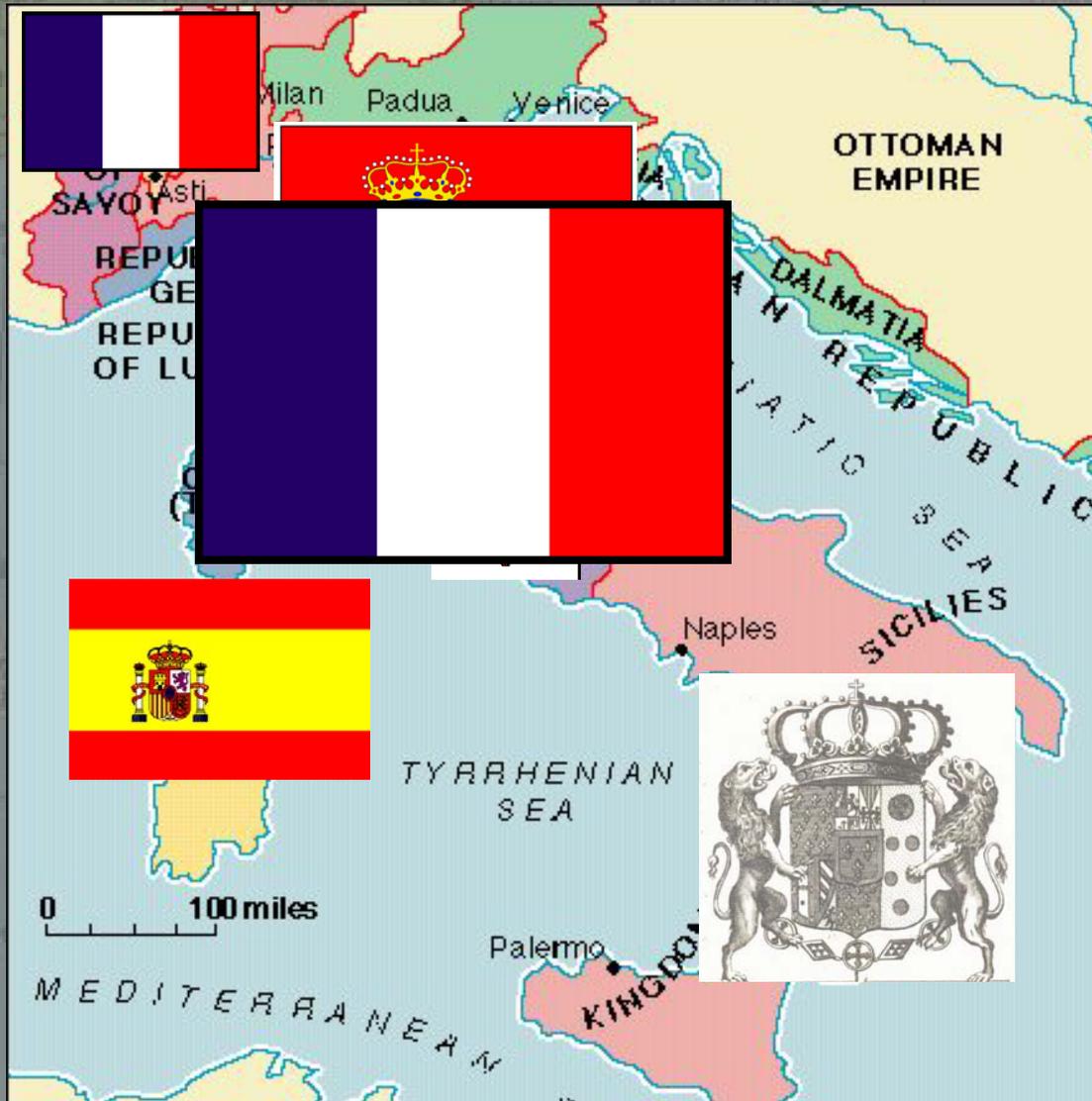


ITALY in the XI Century

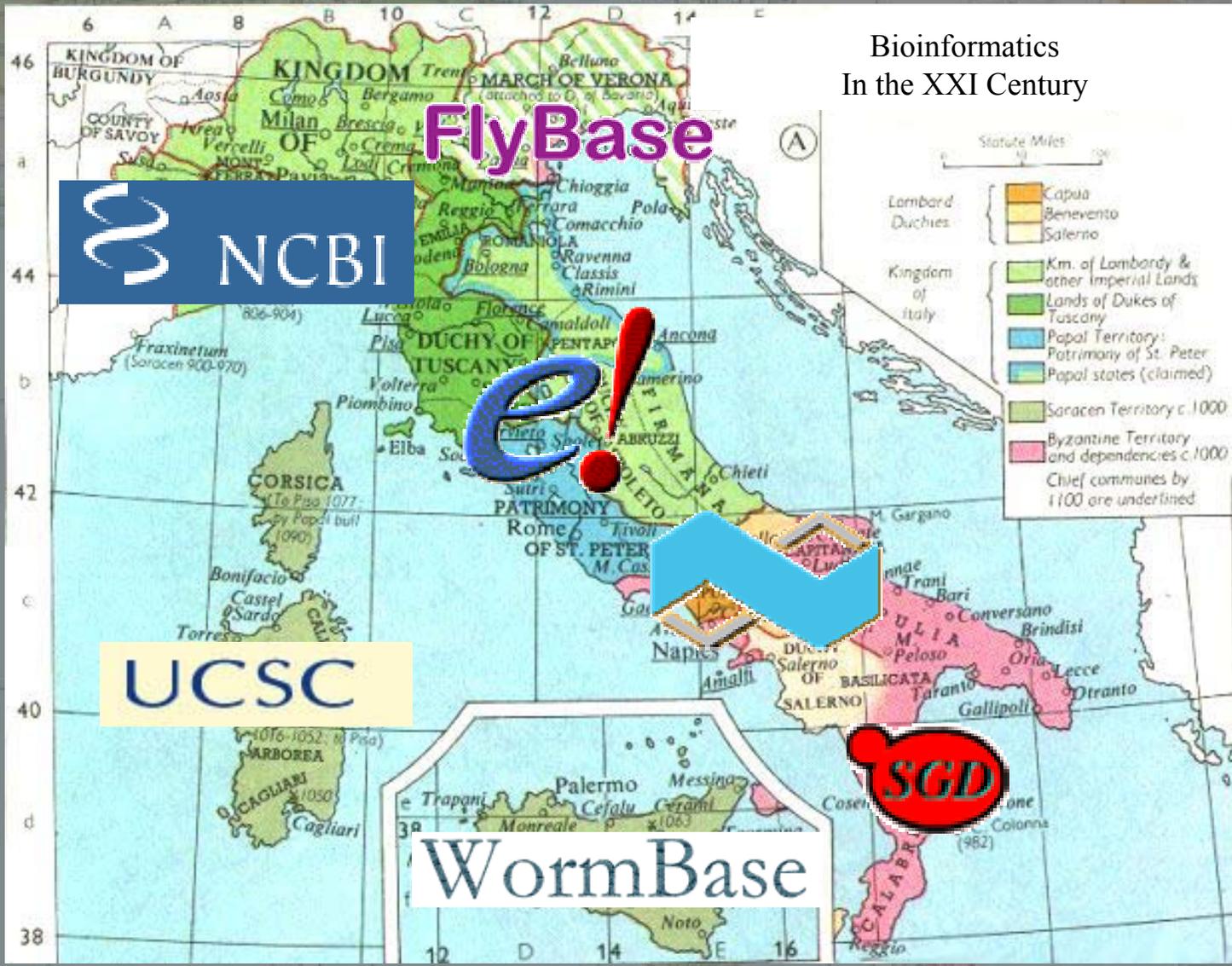
View Map

- Green
- Orange
- White
- Yellow
- Light Green
- Blue
- Light Blue
- Light Green
- Light Green
- Orange
- Light Green
- Light Green
- Light Green

Italy, ca 1820



Bioinformatics, ca. 2002



Making Easy Things Hard

Give me all human sequences
submitted to GenBank/EMBL last
week.



Lots of ways to do it

- ◆ Download weekly update of GenBank/EMBL from FTP site
- ◆ Use official network-based interfaces to data:
 - NCBI toolkit
 - EBI CORBA & XEMBL servers
- ◆ Use friendly web interfaces at NCBI, EBI

Creating a Bioinformatics Nation

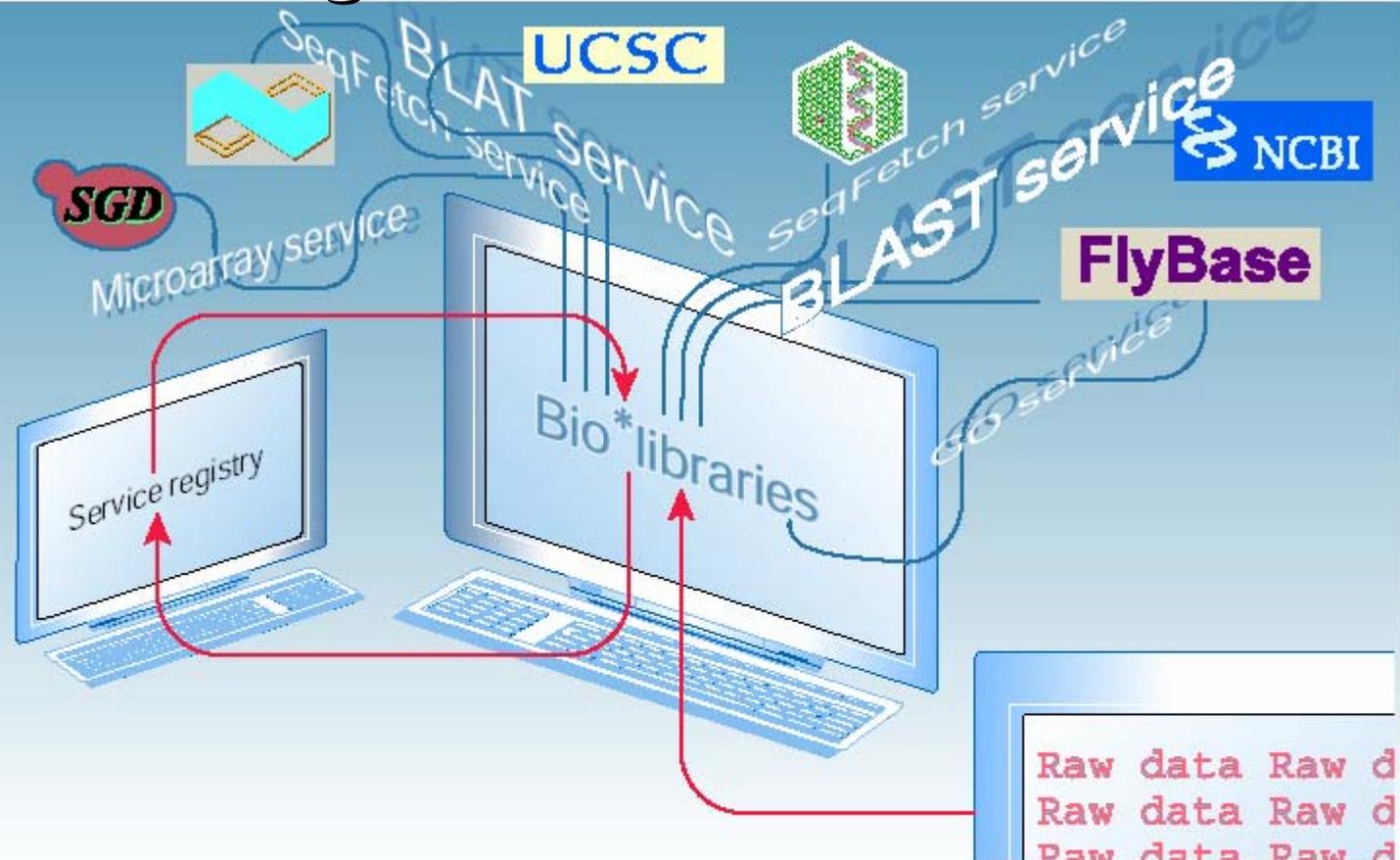


Figure 1 Moving towards a bioinformatics nation. Because each data provider (such as Flybase and UCSC) publishes data in an idiosyncratic form, the Bio* software package (Bio* libraries) was created to massage data into a standard internal format. Unfortunately, Bio* needs to be fixed each time a provider changes its formats. A web-services world would build on the successes of the Bio* projects by defining standard interfaces to various types of computations and data formats. The Bio* libraries can be written to recognize these interfaces, allowing them to interoperate easily with all data providers. A service registry would let data providers enter an electronic 'address book', allowing the Bio* libraries to locate and interact with new data sources automatically.

Special Supplement – Nature Genetics

September 2002

A user's guide to
the human genome

nature
Genetics
volume 32 september 2002



Genomic empowerment: the importance of public databases

Harold Varmus

Over the past twenty five years, a mere sliver of recorded time, the world of biology — and indeed the world in general — has been transformed by the technical tools of a field now known as genomics. These new methods have had at least two kinds of effects. First, they have allowed scientists to generate extraordinarily useful information, including the nucleotide-by-nucleotide description of the genetic blueprint of many of the organisms we care about most—many infectious pathogens; useful experimental organisms such as mice, the round worm, the fruitfly, and two kinds of yeast; and human beings. Second, they have changed the way science is done: the amount of factual knowledge has expanded so precipitously that all modern biologists using genomic methods have become dependent on computer science to store, organize, search, manipulate and retrieve the new information.

Thus biology has been revolutionized by genomic information and by the methods that permit useful access to it. Equally importantly, these revolutionary changes have been disseminated throughout the scientific community, and spread to other interested parties, because many of those who practice genomics have made a concerted effort to ensure that access is simplified

teaching many of the principles of biological design, including evolution, gene organization and expression, organismal development and disease; and in part because those who work on

...all modern biologists using genomic methods have become dependent on computer science to store, organize, search, manipulate and retrieve the new information.

Thus biology has been revolutionized by genomic information and by the methods that permit useful access to it.

into publishable form, has spurred efforts to share other kinds of research tools and has encouraged the notion of making the scientific literature freely accessible through the Internet. The contribution of scientists in many countries to the sequencing of many genomes, including the human genome, has inspired efforts to develop gene-based sciences—from basic genomics to

- **Foreword - Power to the people** - A D Baxevanis & F S Collins
- **Perspective - Genomic empowerment: The importance of public databases**
- H Varmus

- **User's Guide**

- **Question 1**
How does one find a gene of interest and determine that gene's structure? Once the gene has been located on the map, how does one easily examine other genes in that same region? pp 9 - 17

- **Question 2**
How can sequence-tagged sites within a DNA sequence be identified? pp 18 - 20

- . . .

- **Question 12**
How does a user find characterized mouse mutants corresponding to human genes? pp 66 - 69

- **Question 13**
A user has identified an interesting phenotype in a mouse model and has been able to narrow down the critical region for the responsible gene to approximately 0.5 cM. How does one find the mouse genes in this region? pp 70 - 73

(protein/DNA) Sequence Data and Molecular Biology Journals

- Prior to publication, peer-reviewed molecular biology journals require that the authors deposit their data sets in a publicly-accessible archive and obtain an Accession Number.
- The Accession Number is included with the publication (both printed and electronic form)
- In many cases, the software tools used to analyze the sequence data are in the public domain





INSIGHT - Imaging Tools from the Visible Human Project

Terry S. Yoo, Ph.D.

Office of High Performance Computing and
Communications
National Library of Medicine

What is it?: Insight

- A toolkit for registration and segmentation.
- A common Application Programmers Interface (API).
- A validation model for segmentation and registration.
- Open-source resource for future research.



A New Research Program

- Image Segmentation
 - multivalued (multimodal) data
- Image Registration
 - rigid and deformable registration
- Validation
 - Generation of mathematical models as test data
 - Acquisition of validation datasets from medical scanners



Why is it?

- Segmentation and Registration recognized as major research areas for medical image research.
- Common platforms may encourage communication and dissemination of research results.





Who's sponsoring it?

INSIGHT

- National Library of Medicine (NLM)
- National Institute for Dental and Craniofacial Research (NIDCR)
- National Institute of Neurological Disorders and Stroke (NINDS)
- National Institute of Mental Health (NIMH)
- National Eye Institute (NEI)
- National Science Foundation (NSF)
- National Institute for Deafness and Other Communication Disorders (NIDCD)
- National Cancer Institute (NCI)





Contractors and Subcontractors

INSIGHT

GE CRD

Bill Lorensen

MathSoft

Vikram Chalana

U Penn

Dimitris Metaxas

Harvard BWH*

Ron Kikinis

U Penn*

Jim Gee

Columbia U.*

Celina Imielinska

Kitware

Will Schroeder

UNC-CH

Stephen Aylward

U Tennessee

Ross Whitaker

U Pittsburgh*

George Stetten

U Utah*

Ross Whitaker

* indicates a subcontractor.



An Open Source Initiative

- Encourages high-level technical communication.
- Provides conventions (vs. standards) for interoperable software development.
- Establishes a baseline for improvement.
- Opens the field to “beginners.”
- Creates common ground for product growth.
 - example: the creation of HTML enabled Web-based internet development
 - originally part of a broader Government sponsored initiative (incl. gopher, WAIS, etc.)



NIH Draft Statement on Sharing Research Data

March 2002

NIH Office of Extramural
Research



NIH Draft Statement on Sharing Research Data

Proposed Effective Date: January 1, 2003

- NIH will ***expect*** the timely release and sharing of final research data for use by other researchers.
- NIH will ***require*** applicants to include a plan for data sharing or to state why data sharing is not possible.



What do we mean by data?

- We mean final research data necessary to validate research findings.
- Research data do not include:
 - laboratory notebooks
 - partial data sets
 - preliminary analyses
 - drafts of scientific papers
 - plans for future research
 - communications with colleagues
 - physical objects, such as gels or laboratory specimens



Statement will apply to:

- Intramural scientists
- Extramural scientists seeking
 - Grants
 - Cooperative agreements
 - Contracts



What Kind of Research Does This Apply To?

- Data generated with support from the NIH, including:
 - Basic research
 - Clinical studies
 - Surveys
 - Other types of research
- Unless human research participants' identities cannot be protected
- Especially important to share:
 - Unique data sets that cannot be readily replicated
 - Large, expensive data sets



Will NIH Provide Support for Data Sharing?

- Yes
- In grant application – budget and budget justification
- Administrative supplements



Caveats for Studies Including Human Research Participants

- Investigators need to be cautious with
 - Studies with very small samples
 - Studies collecting very sensitive data
- However, even these data can be shared *if*
 - Safeguards exist to ensure confidentiality and protect the identify of subjects



What is Meant by Timely?

- No timeline specified
 - Will vary depending on nature of the data collected
- Investigators who collected the data have a legitimate interest in benefiting from their investment of time and effort
- Therefore, they could benefit from first and continuing use but not from prolonged exclusive use



Why Share?

- Extends NIH policy on sharing research resources
- Reinforces open scientific inquiry
- Encourages diversity of analysis and opinion
- Promotes new research
- Supports testing of new or alternative hypotheses and methods of analysis
- Facilitates the education of new researchers
- Enables the exploration of topics not envisioned by the initial investigators
- Permits the creation of new data sets from combined data



How to Share Data

- Provide in publications
- Share under the investigator's own auspices
- Place data sets in public archives
- Put data on a web site
- Place in restricted access data centers or data enclaves
- Other ways?



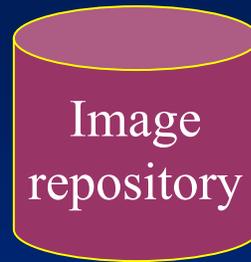
What Will NIH Applicants Need to Do?

- Include a data sharing plan in application
 - Statement of how data will be shared
 - If not, why not
- Where in application
 - End of research plan
 - Budget, budget justification if asking for funds
 - Significance, if creating an important scientific resource

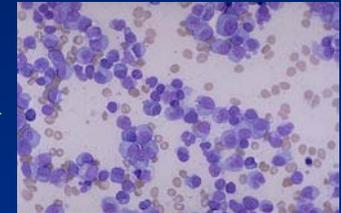


Why “tag” images with concepts?

- Retrieving images by concept
 - Different words for different people
 - term used to catalog is not term used to search
- Multi-resource knowledge acquisition (Medline, image repository, patient database)
- Knowledge mining
 - “Are there unknown relationships between image findings and clinical history?”



“myeloma”



Concepts as mediators

synonyms

Myeloma, Plasma 
Myelomatosis
Plasmacytic myeloma
MULTIPLE MYELOMA
MYELOMATOSIS
KAHLER BZZOLO DISEASE
Kahler's disease
MYELOMATOSIS MULTIPLE
[M]Plasma cell myeloma

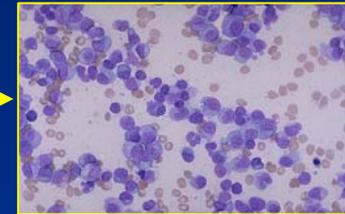
“myeloma”



UI: C0026764



Myelom, Plasmazell [GER]
Mieloma multiplo [ITA]
Mieloma de plasmocitos [POR]



Visualization and Annotation tool = NeuArt

Atlas Viewer

File Edit Atlas View Go To Query Selection Window Help

Image

Zoom In Zoom Out Move

Query

Identity Structure Intensity

Selection

Rectangle Circle Freehand

Boolean Operations

()

AND OR NOT

Atlas

Jones Rhesus

Atlas Orientation

Coronal

Atlas Level

684

Cursor Position

419,792

Zoom

2X

Signal Intensity

Annotations

- Atlas Level 684
 - Brain
 - Lateral geniculate nucleus
 - P-layers
 - Thalamus

View On

Delete ALL

Remove

Update View

Label Manager

Unlink Maintain state Show

Prev Atlas Prev Data Next Data Next Atlas

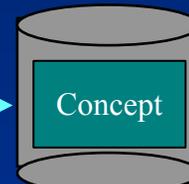
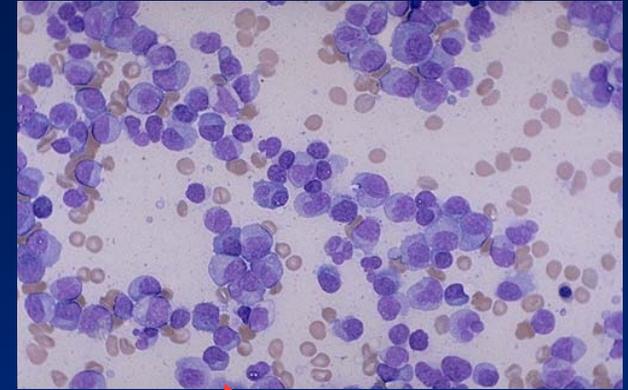
M. Hogarth, UC Davis

Example

XML

```
<image-object>
  <image-id>
    190991
  </image-id>
  <image-location>
    http://images.ucdavis.edu
  </image-location>
  <region-selected>
    <coordinates>
      ...
    </coordinates>
  </region-selected>
  <concept>
    <concept_id>
      <term-system>
        UMLS2000
      </term-system>
      <id-value>
        C0026764
      </id-value>
    </concept_id>
  </concept>
</region-selected>
</image-object>
```

Image



Terminology
repository

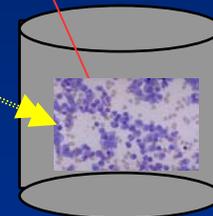
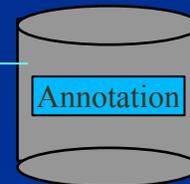
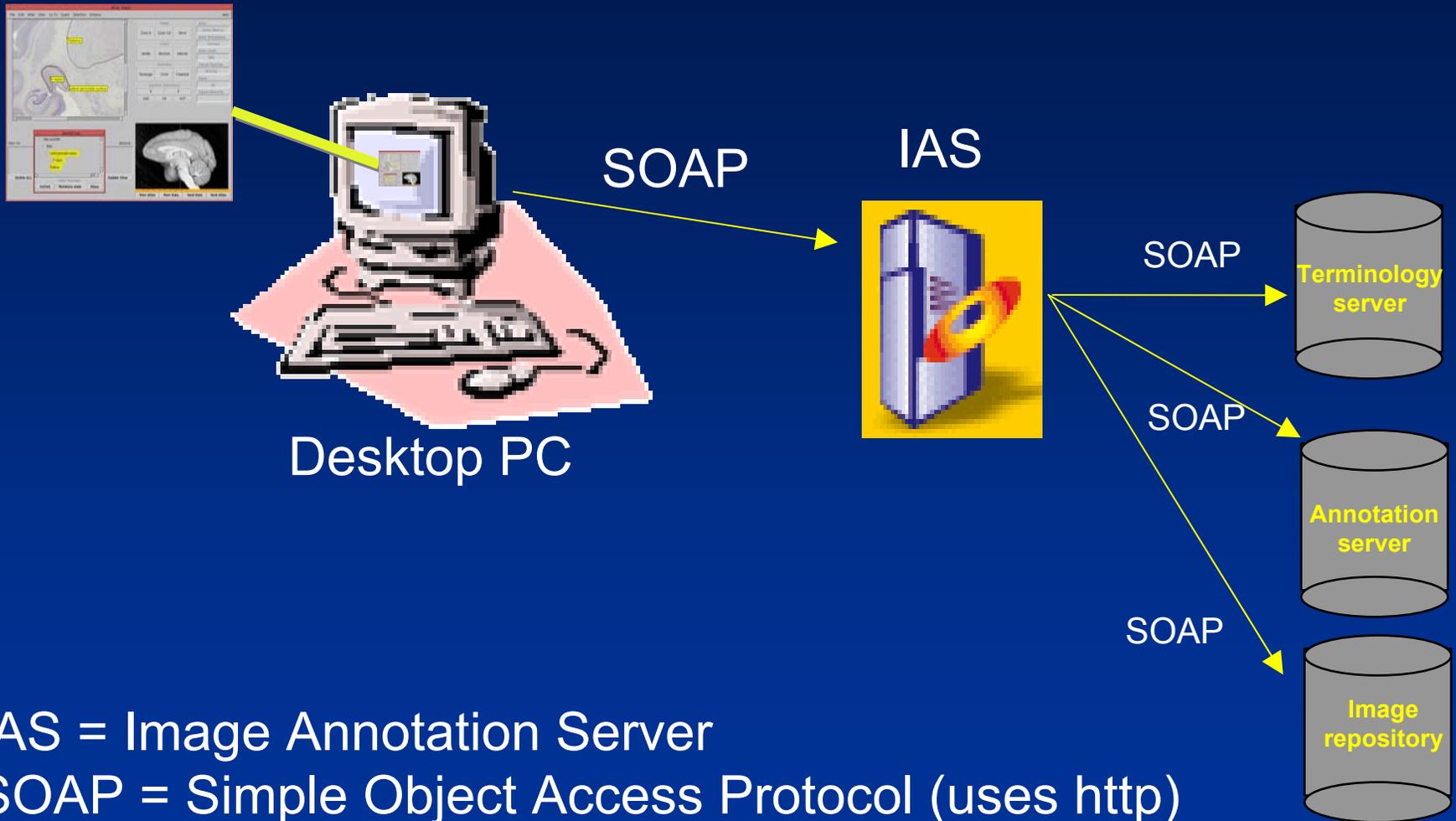


Image
repository



Annotation
repository

HBP Image Annotation Server



IAS = Image Annotation Server

SOAP = Simple Object Access Protocol (uses http)



Grid Computing

"The Grid"

- Coined in 1990's to denote a proposed distributed computing architecture.
- "Flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources"

From "The Anatomy of the Grid"

- Resource Sharing
 - ❏ Computers, Storage, Sensors, Networks, Scientific Instruments
 - ❏ Sharing is highly controlled -- Providers & Consumer define
 - ✌ What is shared
 - ✌ Who is allowed to share
 - ✌ Conditions for sharing
- Coordinated problem solving
 - ❏ Beyond client-server: distributed data analysis, visualization, computation, collaboration
- Similar to the Power Grid, Faucets (Water supply), Nationwide Phone System.

Online Access to Scientific Instruments

Advanced Photon Source

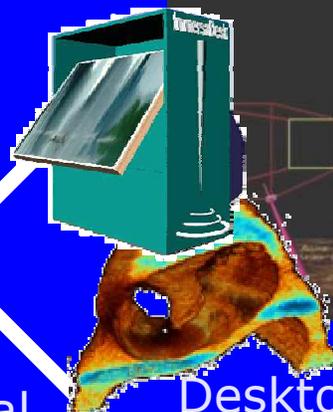


Real-time collection



Tomographic reconstruction

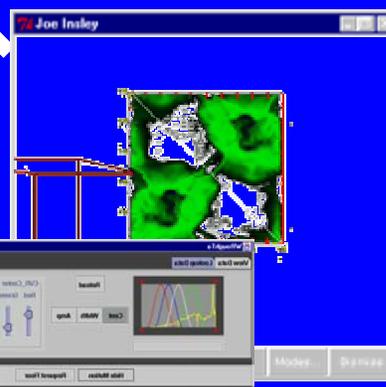
Wide-area dissemination



Archival storage



Desktop & VR clients with shared controls



DOE X-ray Grand Challenge: ANL, USC/ISI, NIST, U.Chicago

Major Grid Projects

Name	URL & Sponsors	Focus
BlueGrid	IBM	Grid testbed linking IBM laboratories
DISCOM	www.cs.sandia.gov/discom DOE Defense Programs	Create operational Grid providing access to resources at three U.S. DOE weapons laboratories
DOE Science Grid	sciencegrid.org DOE Office of Science	Create operational Grid providing access to resources & applications at U.S. DOE science laboratories & partner universities
Earth System Grid (ESG)	earthsystemgrid.org DOE Office of Science	Delivery and analysis of large climate model datasets for the climate research community
European Union (EU) DataGrid	eu-datagrid.org European Union	Create & apply an operational grid for applications in high energy physics, environmental science, bioinformatics

iVDGL: International Virtual Data Grid Laboratory



www.ivdgl.org

National Digital Mammography Archive

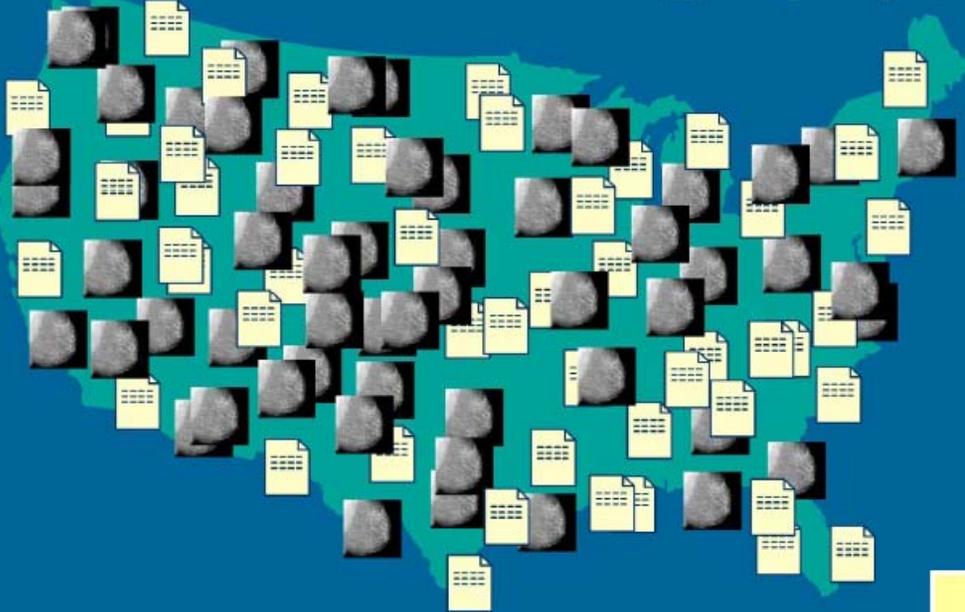
(NGI) Next Generation Internet Demonstration Project

- Three Applications -

1. Archive Storage and retrieval for clinical use
2. Teaching File repository for Radiology Departments
3. Computer Assisted Diagnosis (CAD) as a service

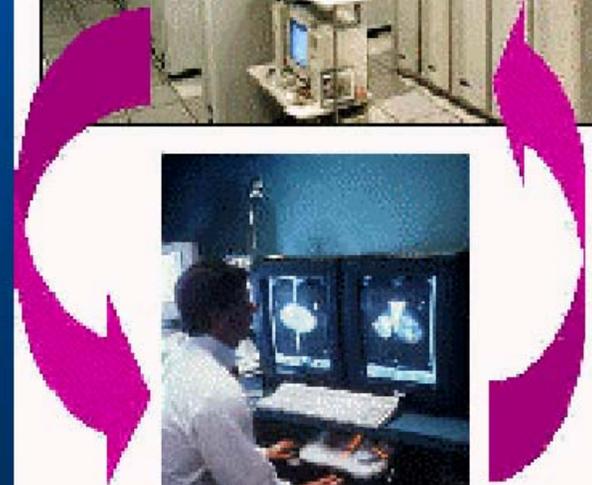
Managing Large Scale Data

DIGITAL Mammography

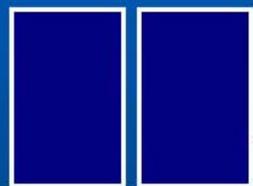


Highly Distributed Source

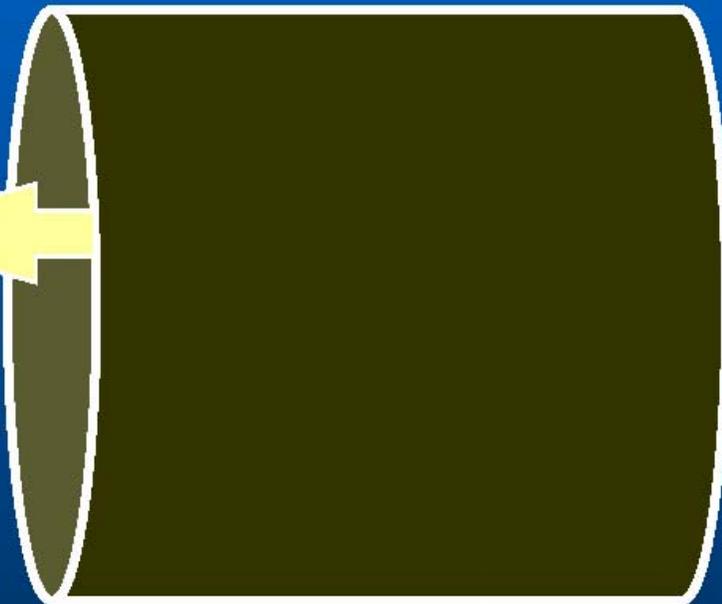
Hierarchical Storage and Indexing



Major Components



Hospital
Portal
Systems



Security Layer
VPN, qmp

RadAR:
Large Scale
Storage

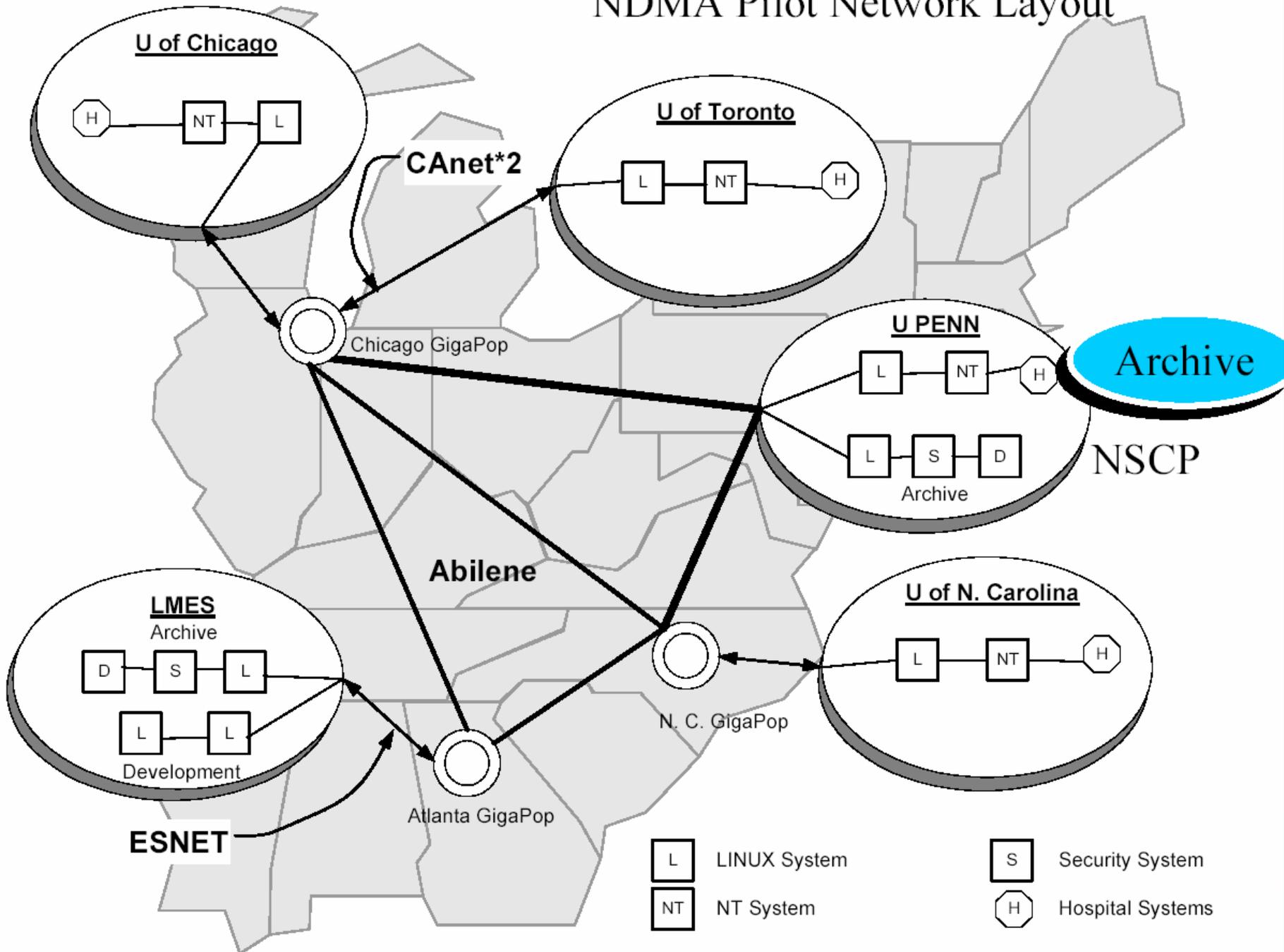
Index

**Audit
and Log**

**Search
and Retrieve**

Mine

NDMA Pilot Network Layout





BIRN – Biomedical Informatics Research Network

BIRN – NCRR

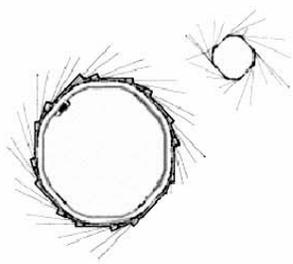
National Center for Research Resources

National Institutes of Health

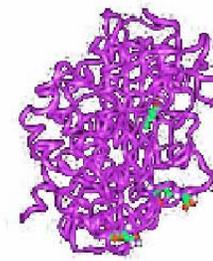
Biomedical Informatics Research Network—BIRN

- Integrating data from different brain mapping research sites
 - UCSD, UCLA, Caltech, Duke, Mass General, Harvard
 - Mouse and human brain
- BIRN Data/Knowledge Grid
 - High-speed networking
 - Access to distributed data
 - Semantic mediation
 - Intra-species and inter-species queries
 - Visualization and analysis tools





Synchrotrons

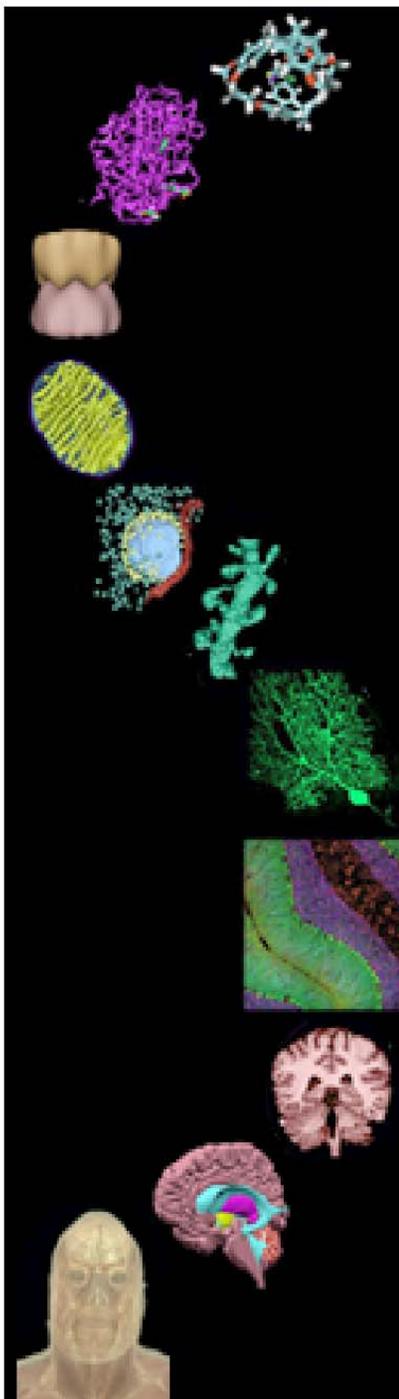
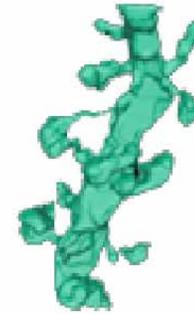


Molecules

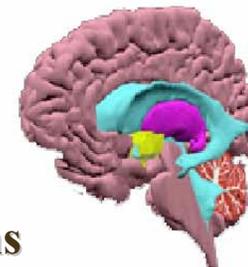
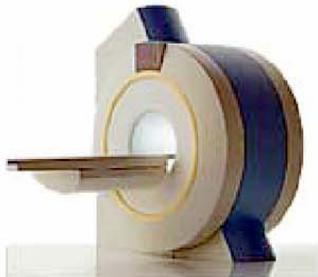


Microscopes

**Macromolecular
Complexes,
Organelles, Cells**



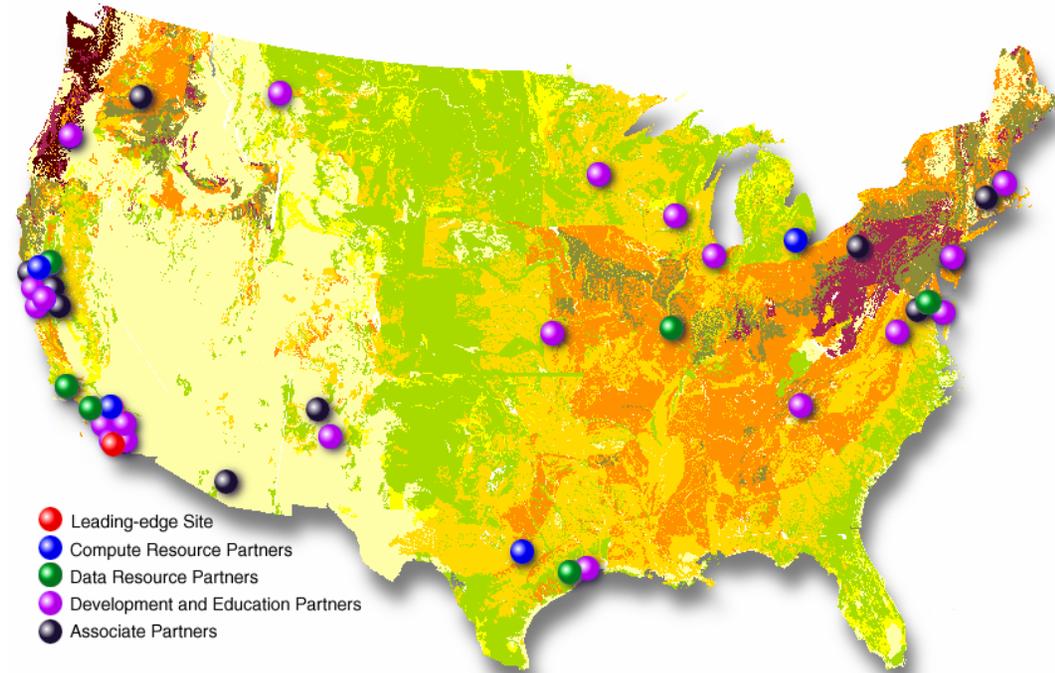
**Organs, Organ
Systems, Organisms**



Magnetic Resonance Imagers

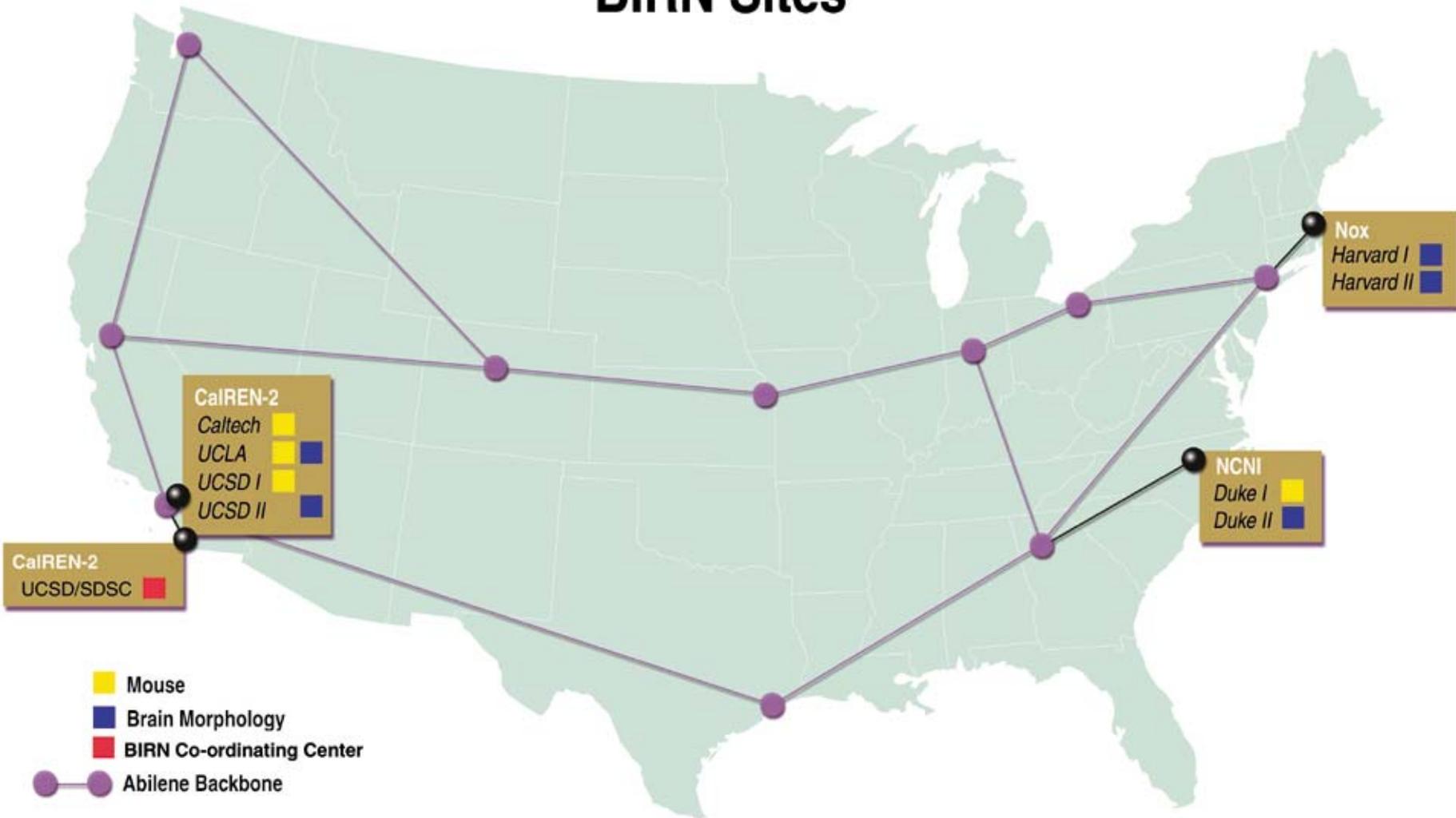
NPACI - National Partnership for Advanced Computational Infrastructure (NSF)

- ~50 partner sites
- shared compute resources
- high-speed networks
- Computational science efforts in “thrusts”
 - Neuroscience
 - Molecular Science
 - Earth Systems Science
 - Engineering
- Enabling Technology Thrusts
 - Resources (TeraFlops, High Performance Networks, Data Caches)
 - Metacomputing (Grid Tools - Middleware)
 - Interaction Environments (Visualization - Science Portals)
 - Data-Intensive Computing (Databases - Data Migration - Knowledge Eng.)



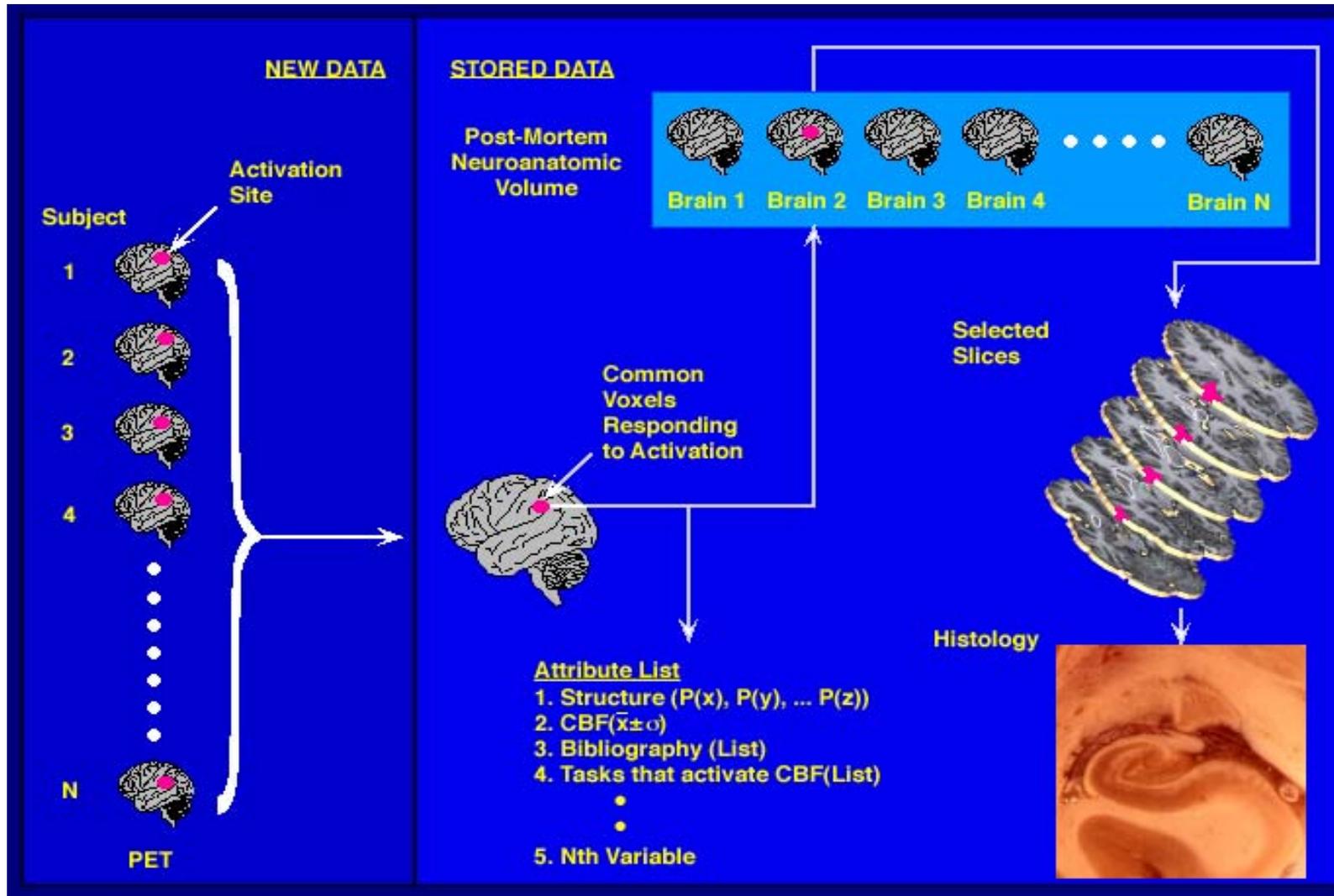
THE BIOMEDICAL INFORMATICS RESEARCH NETWORK LAYOUT FOR PHASE 1

BIRN Sites



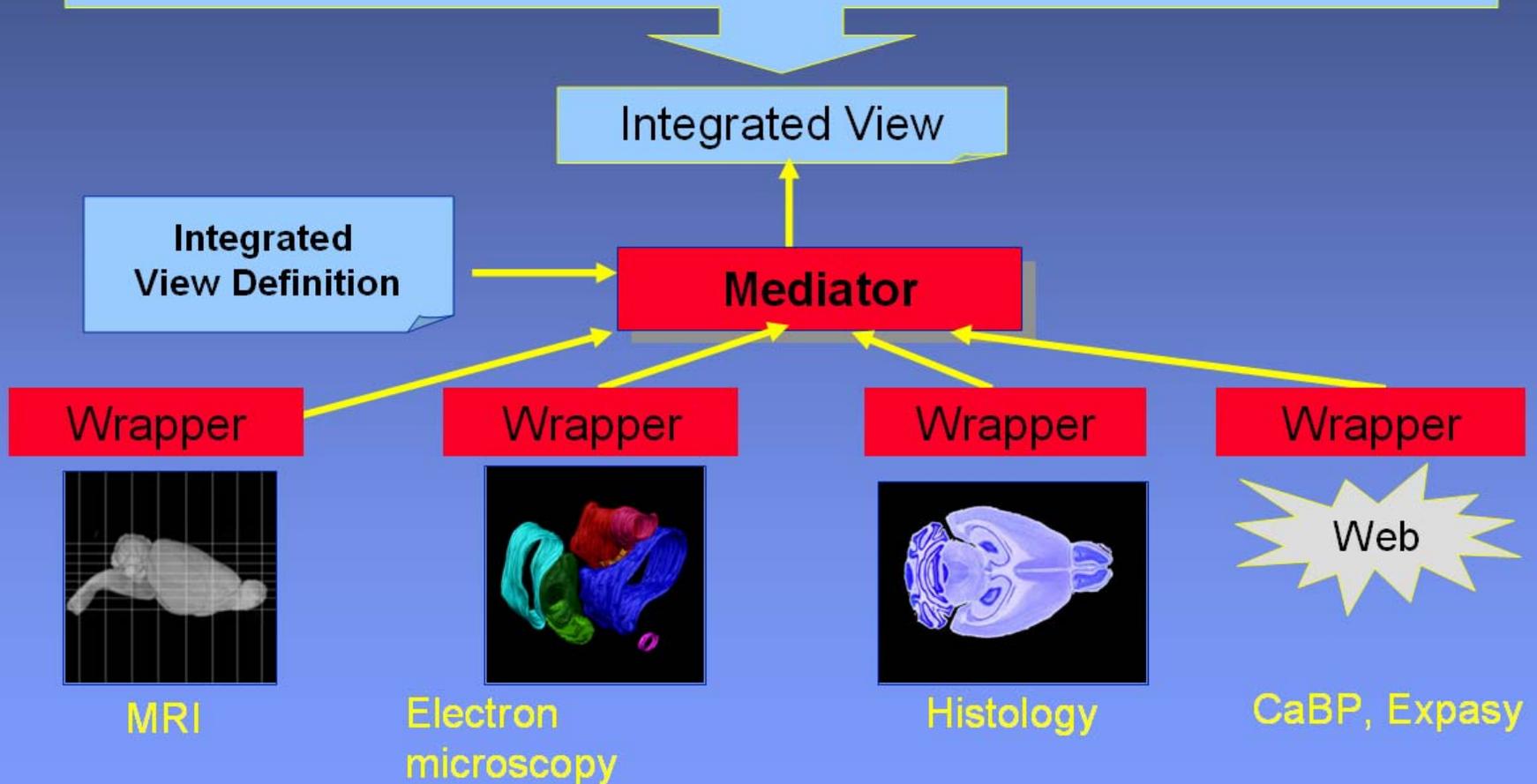
Federate Emerging Databases

Infrastructure to relate, combine & produce meta data



Example of BIRN Federation

Are there changes in axon diameter, and/or number, in the optic nerve of EAE animals, before the development of gross structural changes?



Standards for Information Interchange –

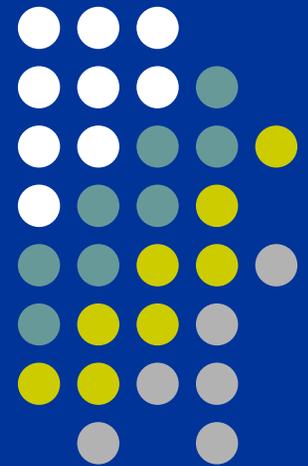
**The basis for multidisciplinary
collaboration**



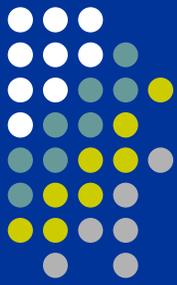
Clinical Data Standards

Building consensus within the industry towards standards for exchanging electronic data:

‘Speaking the same language’ to achieve more efficient and higher quality clinical trials



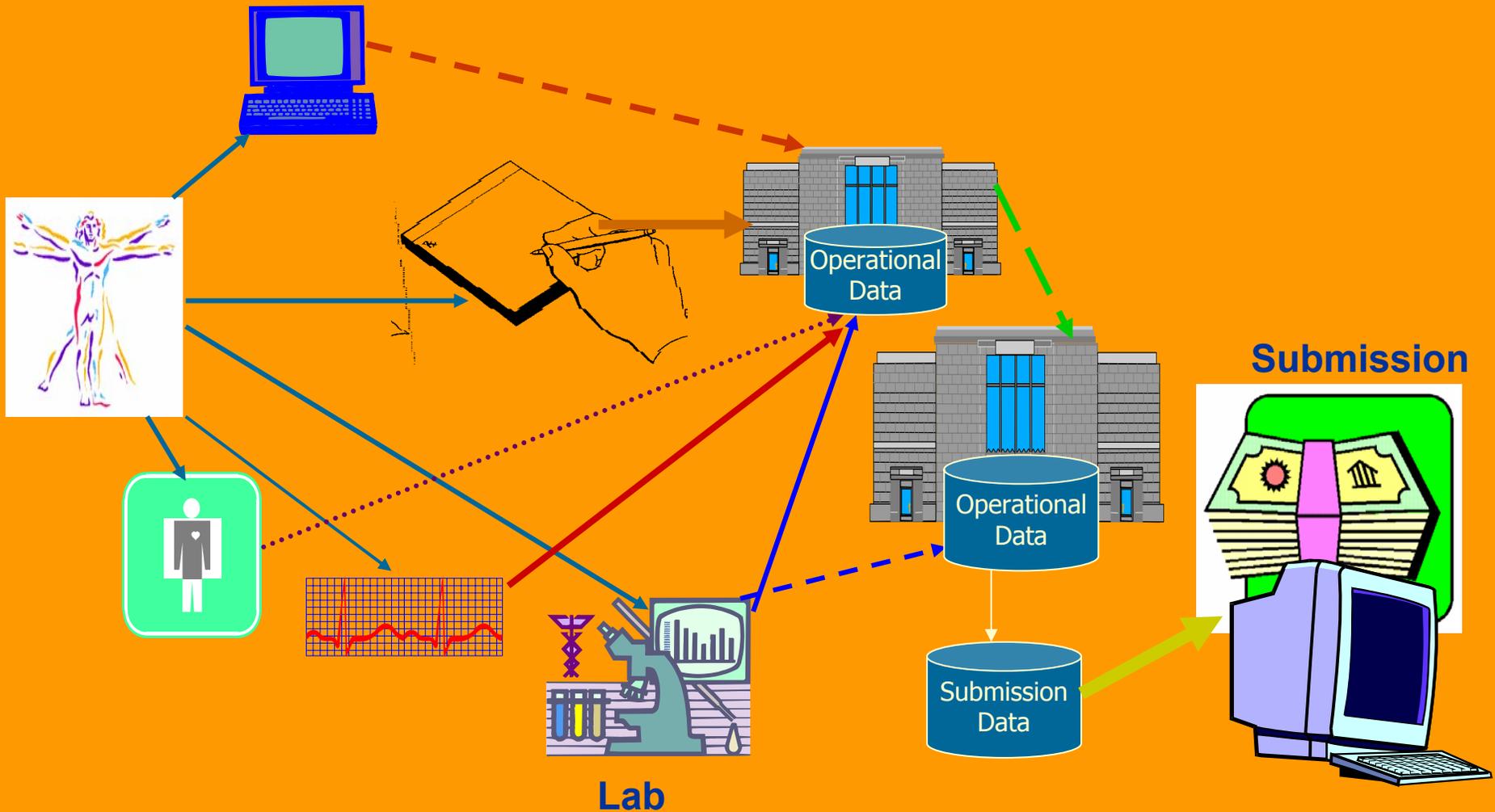
Building consensus within the industry towards standards for exchanging electronic data – Outline



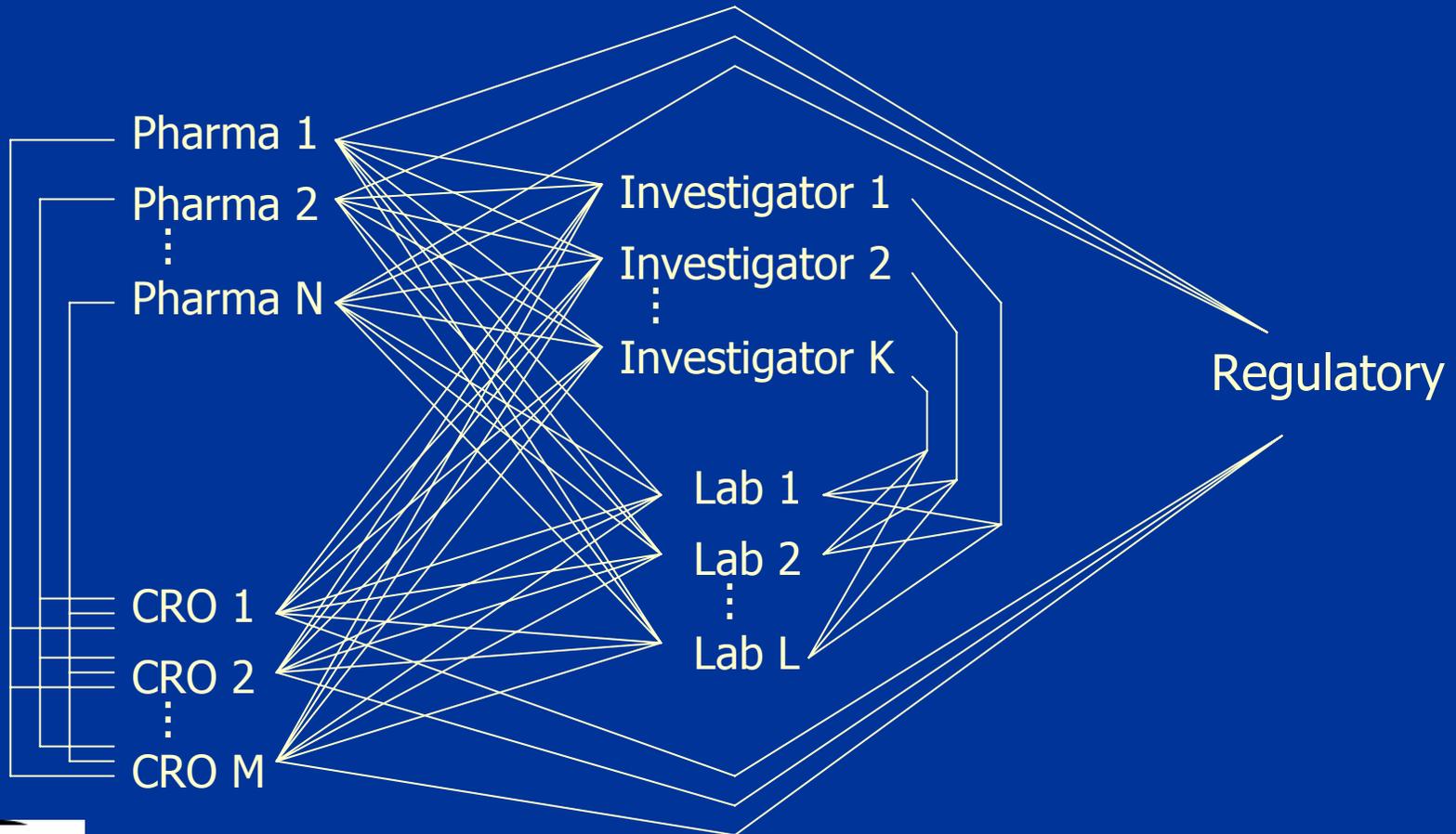
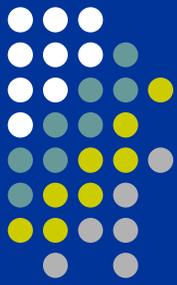
- Benefits of Standards and Potential Value
- What is CDISC?
 - Principles of CDISC
 - Organization of CDISC
- CDISC Progress
 - CDISC Models
 - Implementation of CDISC Models

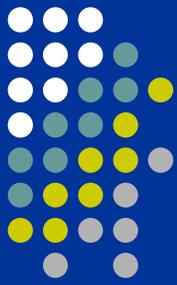


A Case for Data Standards



Current State: Costly and Time-consuming





CDISC Value: Cost of Clinical Data Interchange in Clinical Trials

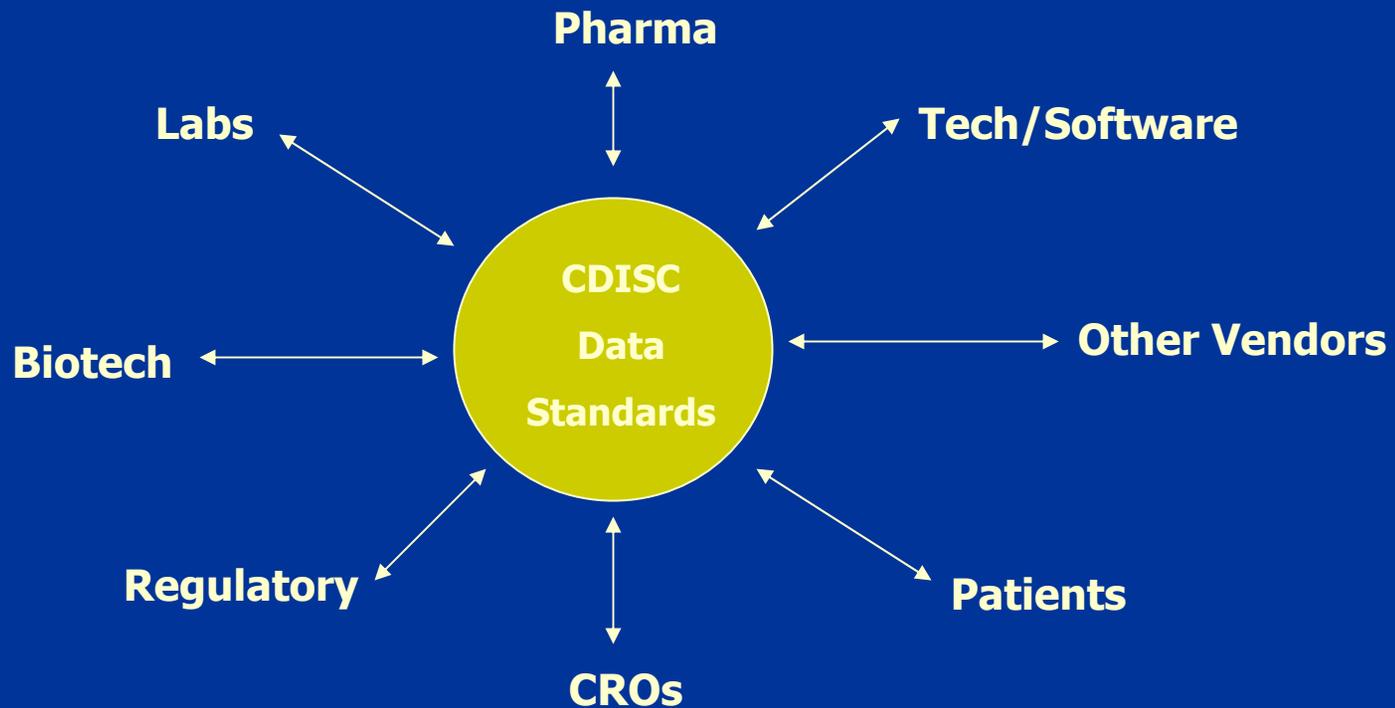
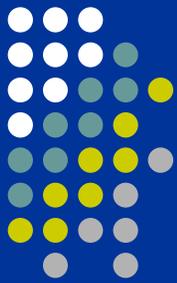
- ~7,000-8,000 clinical studies/year*
- ~ 30 % outsourced and 5-10% EDC
- Estimated cost of \$35,000 for EDC transfers, \$25,000 for CRO data transfers, and \$10,000 for lab data transfers
- **Conservative Annual Cost to the Industry:**
\$156 million

NOTE: The costs incurred with development partners or merged companies sharing data and the cost of preparing data for eSubmissions are not addressed in this set of calculations, nor are other costs such as training, planning or equipment.

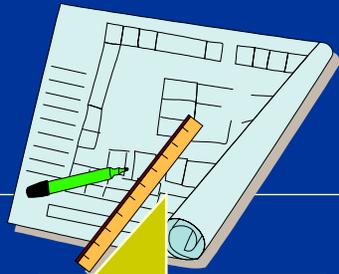
CDISC White Paper on Website
*Source: CenterWatch (v.7, issue 11)



Desired State



Where are CDISC standards being used?



Data Import
& Transfers
Specifications



Design of
Datasets & Analysis
Tables/Listings

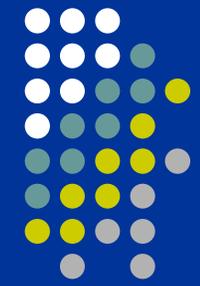
One Consistent Data Standard
Applied Across Systems & Processes

Setup

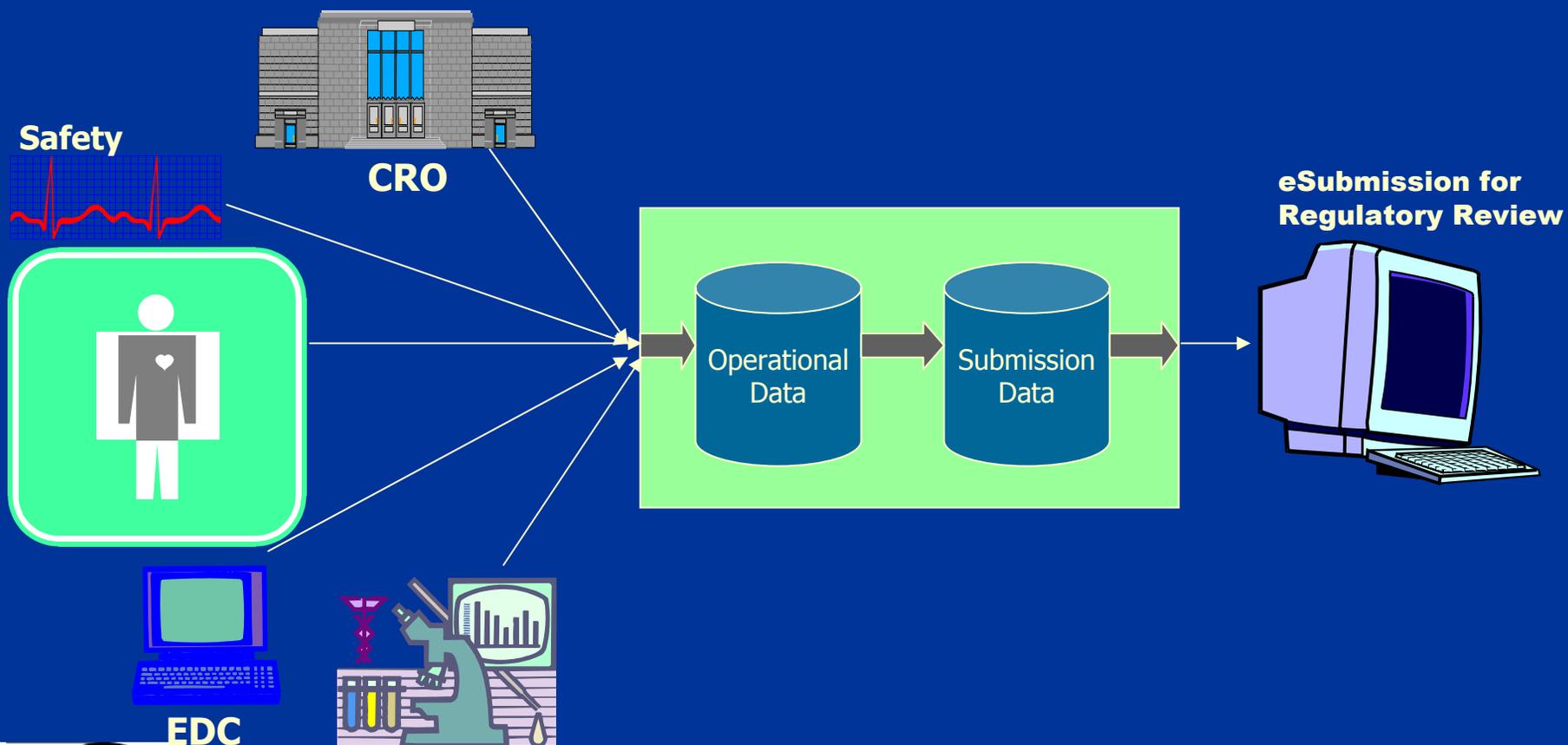
Clintrial

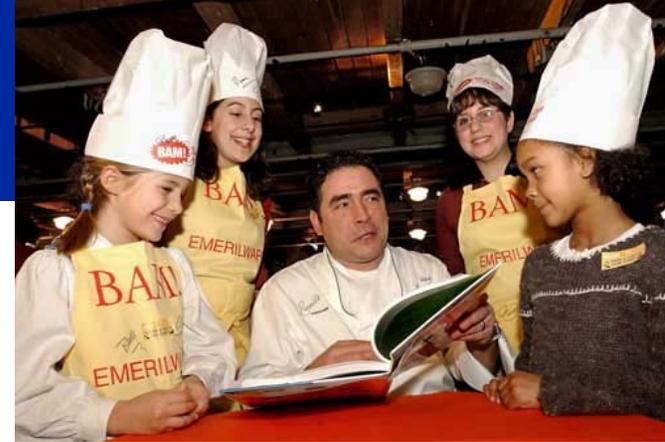
Submission





Standards to Enable Seamless Flow of Data from Patient to Reviewers





Emerging Tools for Building Integrated Scientific Data Resources

Joe Futrelle
**National Center for
Supercomputing Applications**

Overview

- **Vision: A Digital Library of Scientific Data**
- **How to Integrate Scientific Data**
- **New Technologies for Data and Metadata: XML and friends**
- **Current Scenarios, Projects and Technology**



Vision: a Digital Library of Scientific Data

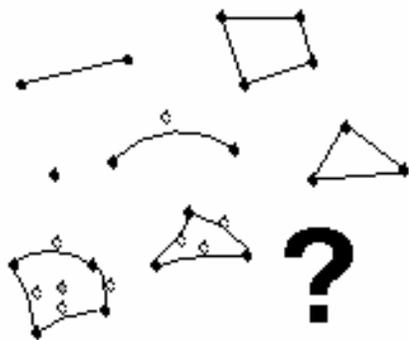
- **Contents**
 - scientific literature
 - data used in studies
 - software used to do the studies
- **Services**
 - digital publishing
 - retrieval of data based on scientific criteria
 - remote analysis and visualization
 - access to computational resources
 - ability to link data from different studies and disciplines together to do new studies

This is a Really Hard Problem

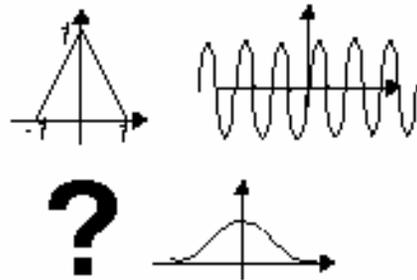
- **Scientific data is exploding in**
 - resolution
 - complexity
 - heterogeneity
 - volume
- **It's not enough to just turn every science data collection into a website**
 - large data sets cannot go “the last mile”
 - a digital library of science data will integrate many (1000's) of collections
 - data management tools must work across collections

Describing Data Is Challenging

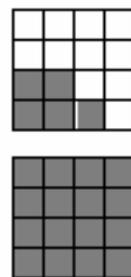
Element Types



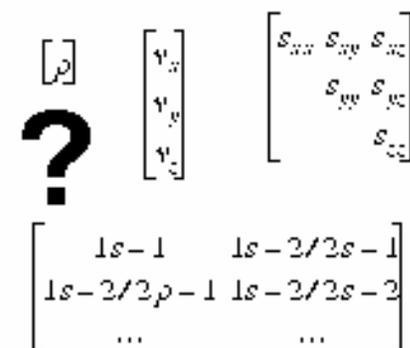
Basis Functions and Interpolation Schemes



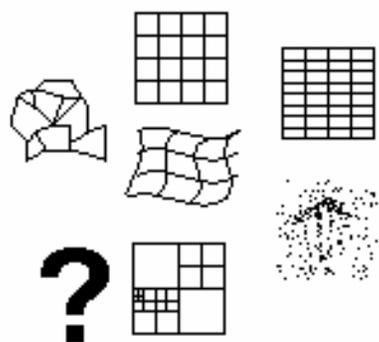
sparse and dense fields



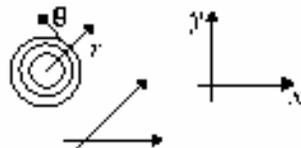
Field value types



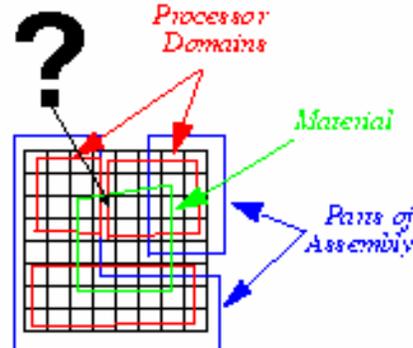
Mesh Types



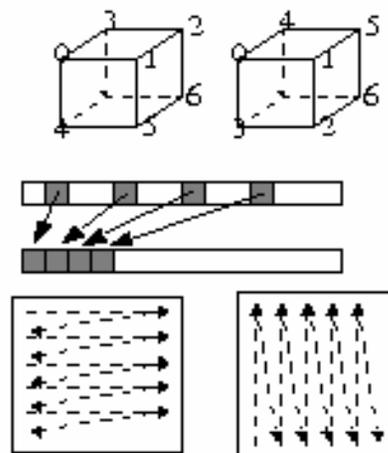
Coordinate Systems



Mesh Decompositions



Storage Conventions And Data Structures



Compression



How to Integrate Scientific Data

- **Generate integrated use scenario**
 - **input from scientific community**
 - represent ~100 groups of researchers with common scientific specialization
 - informal
 - **more than one data collection across**
 - discipline or sub-discipline (e.g. wavelength in radio astronomy, species in biology, process in chemical engineering)
 - scientific data type (e.g. satellite swath, genetic sequence, sensor trace)
 - access modality (e.g. browsing, search, visualization, simulation)

How to Integrate Scientific Data (cont.)

- **Develop data and metadata models to enable the scenarios**
 - identify community-wide data semantics
 - formal, incremental process
 - ongoing review and documentation
 - target key semantics for scenarios
 - use extensible data modeling technologies (e.g. XML, RDF, HDF) to implement data models
- **Link scenarios to build network of data services**
- **Other concerns**
 - security
 - intellectual property
 - data preservation

New Technologies for Data and Metadata

- **What's the difference between data and metadata?**
 - Metadata is data that describes other data (e.g. a card catalog)
 - Within an item in a collection of information:
 - "Data" grows as the *amount* of information in the item grows
 - "Metadata" grows as the *complexity* of information in the item grows
 - All metadata is data but not all data is metadata
- **Why does it matter?**
 - Data and metadata have different usage patterns and performance implications

New Technologies: XML

- **XML is really a set of closely-related technologies, including**
 - XML: generalized markup
 - XLink and URI: interobject reference and linking
 - XML-Schema: document model definition
 - XSL: transformation and presentation
 - RDF: metadata and and inference
 - XQuery: retrieval from XML documents
 - SOAP: remote procedure calling
- **Key commonalities:**
 - draft standards from WWW consortium
 - text-based
 - extensible/portable

New Technologies: XML

- **Suitable for metadata and "light data"**
- **Structured**
- **Hierarchical**
 - Limited graph-like relationships (e.g. ID's)
- **Portable across**
 - languages
 - operating systems
- **Becoming ubiquitous**
 - standard parser API's (DOM, SAX)
 - parsers available in all major languages, platforms

Tomorrow

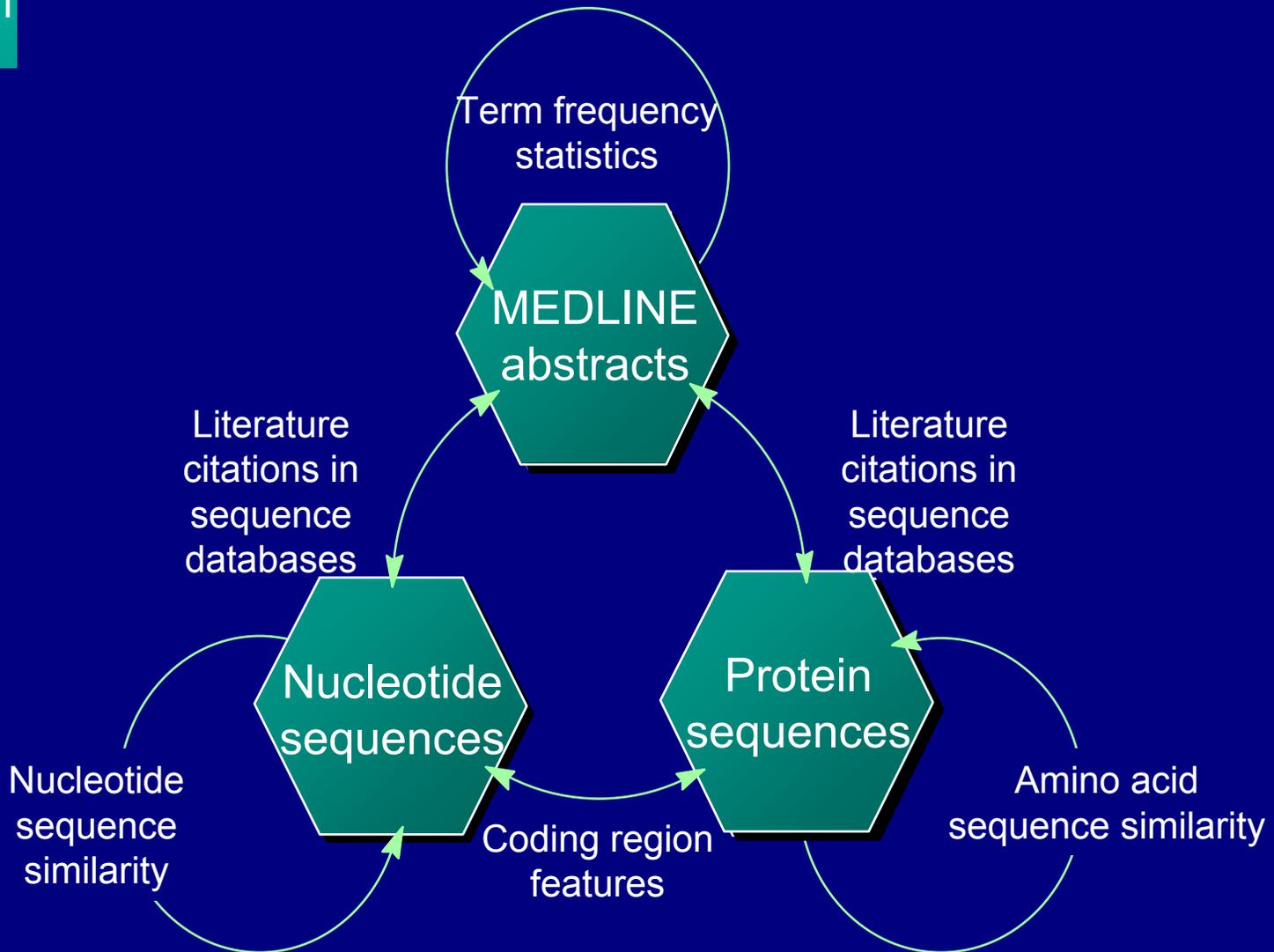
- 8 AM – Presentation by Jim Gray, Microsoft Research on Databases
- 9 AM – 1 PM ... Many image databases, architectures, and applications
- 2 PM – 3:30 PM ... Emerging standards, FDA e-submission process, and WEAR
- 3:30 - ??? PM ... Breakout sessions



The End



Entrez: Pathway to Discovery



Entrez Increases Discovery Space

