Breakout Group 2
Databases

Participant Makeup

- Computer Scientists (databases, machine language, integration)
- Image (medical image) databases
- Clinical interests
- Signals (e.g., ECG)
- Genomics
- Other

- Recognized different backgrounds and language of the various groups.

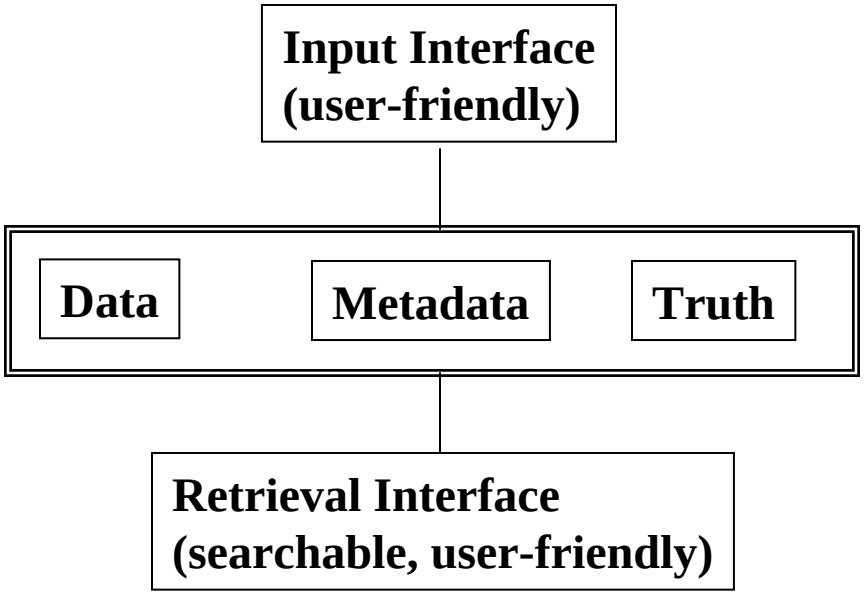
Brief Database Presentations

- Image (medical image) databases
  - Physionet -- Goldberger (NCRR funding)
  - BIRN (brain MRI) -- Kikinis
  - LIDC (lung imaging) -- McLennan (NCI funding)
  - Medical Informatics Europe -- Wittenberg (Europe partial funding)
  - caBIG (cross database integration) -- Covitz (NCI funding)
  - Genomic databases (Bermuda Accords) -- Covitz
  - FDA aspects -- Brown
  - Organism modeling (mod.org) -- audience
  - Microarray groups -- audience

Pattern after NECTAR
(National Electronic Clinical Trials And Research Network)

National Biomedical Data Center - NBDC)
(National Electronic Research Database - NERD)

Databank A <---caBIG-----> Databank B



- **Security & Quality Control**
- **Authenticity of input data**
- **Approved retrieval**
- **Continued maintenance/funding**

Note: Data includes images, electric signals, genomic data, etc.
Need to show the vision giving justification for the need of such a national resource/databank

Breakout Group 2
Databases

1. In general, need a culture change with grass roots motivators/initiators required
   a. For investigators & industry in terms of sharing databases
   b. For funding agencies in terms of funding non-hypothesis driven database development
   c. Also need to include publication recommendations similar to that which was done within the genomic community
      a. Have associate editors suggest in letter to editor
      b. Have some investigators "set the example"
      c. See the NY Times article
      d. Currently some clinical journals are evaluating this
      e. Submission would not be to the journal but rather to the national databank
      f. Vancouver group which includes JAMA etc.
      g. EFMI has signed a contract with publications in Europe
   d. Support from the roadmap initiatives; also need evidence of continued funding for the national resources
   e. Need to recognize that database development publications should be used in promotion.

2. Recognize that the ultimate national biomedical databank could be the regular deposit of all patient biomedical data -- very futuristic though

Breakout Group 2

Databases

Recommendations: Short Term < 5 years

3. Short term: Create a comprehensive inventory of existing databases with corresponding infrastructure (a database of databases)
   a. Include information on use, contributors, allowed users, lifetime, limitations, and other descriptors
   b. Include broad categories of imaging data, physiological signal data, genomics, etc.
   c. Include even restricted/limited access databases (e.g., Mayo Clinic/IBM database)
   d. Create as a "pubmed" of databases with appropriate interfaces
   e. Mechanism: RFA or contract via the NLM ?
   f. Similar to the clinical trial research inventory performed for NECTAR (which is via contract)

Breakout Group 2

Databases

Recommendations: Short term < 5 years

4. Short term: Aim to incorporate and create shared grid-based national databanks for depositing existing and new data
   a. Centers for databank development (roadmap) -- similar to the feasibility studies for NECTAR (by contract) (learn from other databases)
   b. Fund limited number of feasibility/demonstration projects for a limited number of specific-type databases (roadmap with NIH and FNIH)
      - Example of a specific database is the LIDC database -- focused on a specific task (e.g., ECG, CAD)
      - Relate to tasks in Breakout Session 1
   c. Fund limited number of feasibility/ demonstration project for a nonspecific-type database (roadmap with NIH and FNIH)
      - Example of a non-specific database is pubmed
      - Might be all patient cases with annotation from records from two hospitals for two years
      - Might be a general imaging database
      - Test with various "appropriate clinical questions"
   d. Potentially mechanism could be like a non-hypothesis driven R21/R33 with R21 on 4c and the R33 on 4b; done by contract like NECTAR**?**
   e. Each would incorporate the development of attributes from all 4 issues described in recommendations # 6-9 (all four sections)
   f. Requires multidisciplinary team due to broad range of attributes (ultimate users, developer, software, hardware)

Breakout Group 2

Databases

Recommendations: Long term > 5 years

5. Long Term: Aim to incorporate and create a shared national databank for depositing existing and new data
   a. Incorporation of existing databases where possible into a national databank housed on national grounds such as NIH
   b. Means for incorporation of "private" databases, on which a publication was based, into the national databank
   c. Creation of more data for the databank
      - Specific (e.g., LIDC) versus nonspecific (e.g., pubmed) databases
      - Requires multidisciplinary teams especially in determining "truth" characteristics
   d. Would incorporate the developed attributes from the short term feasibility projects for this national databank
   e. Single physical warehouse on NIH grounds vs. Federation of databanks/registry (distributed databases and/or data warehouses)
   f. Funding of new data/databases from grants
   g. Funding of main infrastructure via contract with NIH and FNIH, etc.

Breakout Group 2

Attributes to be considered & developed for databanks in general

6. Contributor agreements and rewards for contributing
   a. Credit databases in research publications
   b. Credit system for those who contribute
   c. Debit system for those who don't contribute but want to use (grant fees, university research)
   d. Line item in RO1 budgets to help in the continued maintenance of the national resource

Breakout Group 2

Attributes to be considered & developed for databanks in general

7. Databank elements/entry descriptors
    1. Appropriate standard sematics/annotation/CDEs/ontologies from controlled vocabulary
        a. Biomedical objects
        b. Common data elements (CDEs)
        c. Controlled vocabularies
    a. Databanks contain data, metadata, and sometimes outcome truth
    b. Metadata (examples) & clinical reason
        – Clinical info, structured reports
        – Associated image data, genomic data
        – Diagnostic or therapeutic outcome data
    c. Treat the semantics/annotation of “truth” as another descriptor of the database entry
    d. Use layered truth, I.e., e.g., actionable region --> lesion --> cancerous lesion
    e. Characteristics of the data acquisition system (e.g.., physical characteristics of an imaging system)
    f. Methods to handle changing metadata over time (updates or new entries)

Breakout Group 2

Attributes to be considered & developed for databanks in general

8. Databank infrastructure (not just list retrieval; need intelligent extraction [semantics/annotation])
   a. Input interface
   b. Internal organization (note needs to be able to handle image data)
   c. Intelligent retrieval based on -- First searchable annotation -- Then intelligent feature extractions
   d. Retrieval (web-based?, others?)
   e. Open source
   f. Quality control (authenticity of input data/metadata/truth, integrity of maintained data)
   g. Integrity of database development to include ethical standards
   h. Handling of IRB & HIPAA issues; and associated road blocks
   i. Security (privacy issues, varying limited access rights for input, browsing, and retrieval)
      – Ardais example
   j. Links to source, e.g., the clinical trial from which data/images came
   k. Flexible/dynamic/expandable/scaleable/robust database
   l. Flexible data entry including new modalities; changing truth, expandable
   m. Ongoing maintenance (bugs, new metadata, elimination of old, curation)
   n. Oversight and advisory committees
   o. Ability to reuse data
   p. Customer involvement and support
   q. Linkage with FDA?

Breakout Group 2

Attributes to be considered & developed for databanks in general

9. Interoperability (linking among) databases
   a. Being considered under NCI's cancer Biomedical Informatics Grid (caBIG) initiative and can be translated
   b. Common language/structures/ontologies (e.g., UMLS)

Breakout Group 2

Databases: Additional Issues

10. Need to consider ownership of the national resource databanks
    1. Physical warehouse should be a national resource on national grounds
    2. An alternative is a federated distributed warehouse -- concern on ownership and continued maintenance
    3. Note that this may be the first national database with private individual data
    4. Recognize the potential hazards of the databank

11. Toolkits, open source in addition to those in # 9

12. National IRB?

13. Learn from non-biomedical databases (National Space Science Data Center)

14. May need to link with separate databases
    a. CMAP -- cancer molecular analysis project
    b. NIAP - national image & analysis project ???
    c. NCRR/PhysioNet -- complex physiologic signals
    d. Genomic project ?