Bioinformatics and Modeling

A. Jamie Cuticchia, Ph.D.
Director of Bioinformatics
ajc@rti.org

August 31, 2005

What is it? Why is it important?

- Bioinformatics is the use of computer hardware, software, and communications to answer biological questions
- Data is continuing to accumulate with doubling times of less than 8 months for most major data sources
- More “in silico” work is required to rank drug targets prior to initiating clinical trials
- Nearly every major NIH Roadmap Initiative has a bioinformatics component
- Consulting groups have estimated the bioinformatics market eclipsed the film industry in size in 2004 with a total market size of **$6 – 9 Billion, $23-27BB by 2009!**

Topics to be Covered

- Data Warehousing and Federation
- Modeling

Data Warehousing

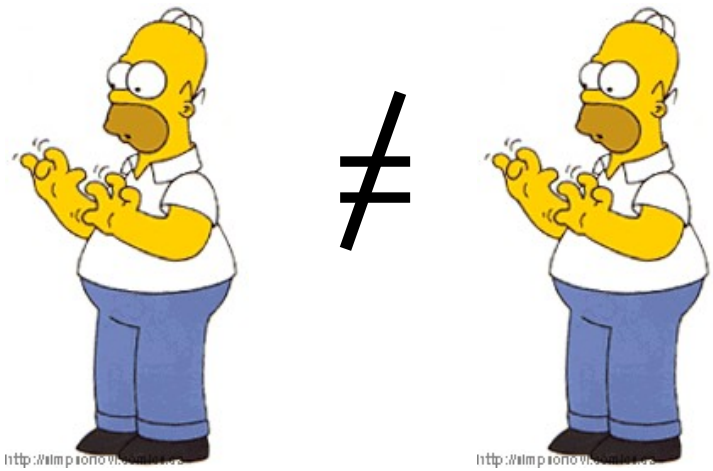
- Simply put, find a safe place to store data.
- Data can be from multiple sources.

Data Coordinating Centers

- Only a start to the process
- May need themselves to be “coordinated” with other centers
- Consistent Q/A Q/C
- Consistent Vocabulary

Do Data Across Database Point to the Same "Thing"

(a) Two Identical Records Un-matched



Simpson ≠ Homer

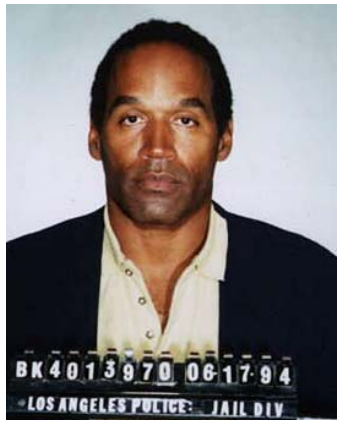
Aliases – terms with almost identical definitions

Do Data Across Database Point to the Same "Thing"

(b) Two Different Records Matched

Simpson

=

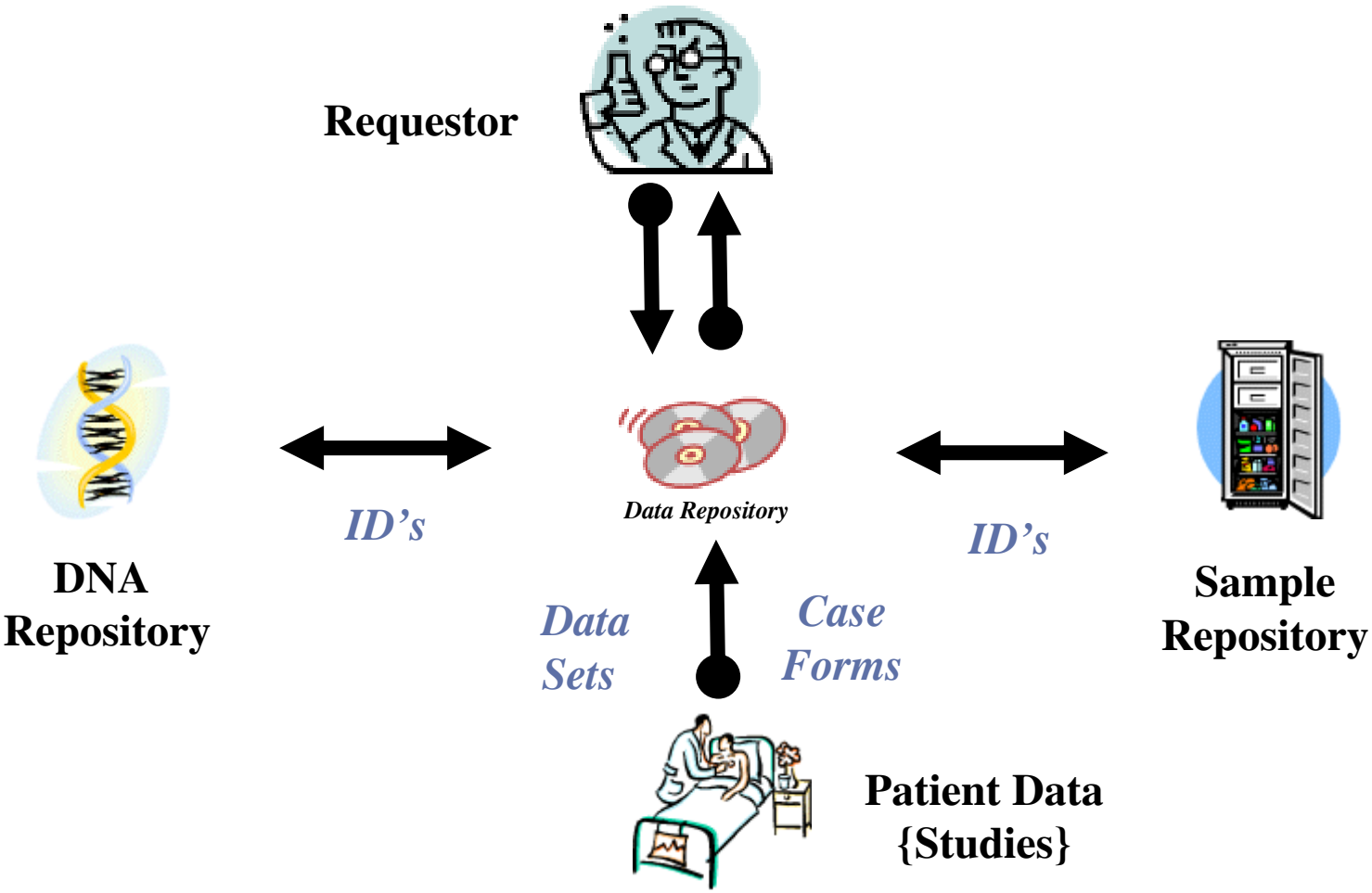


Simpson

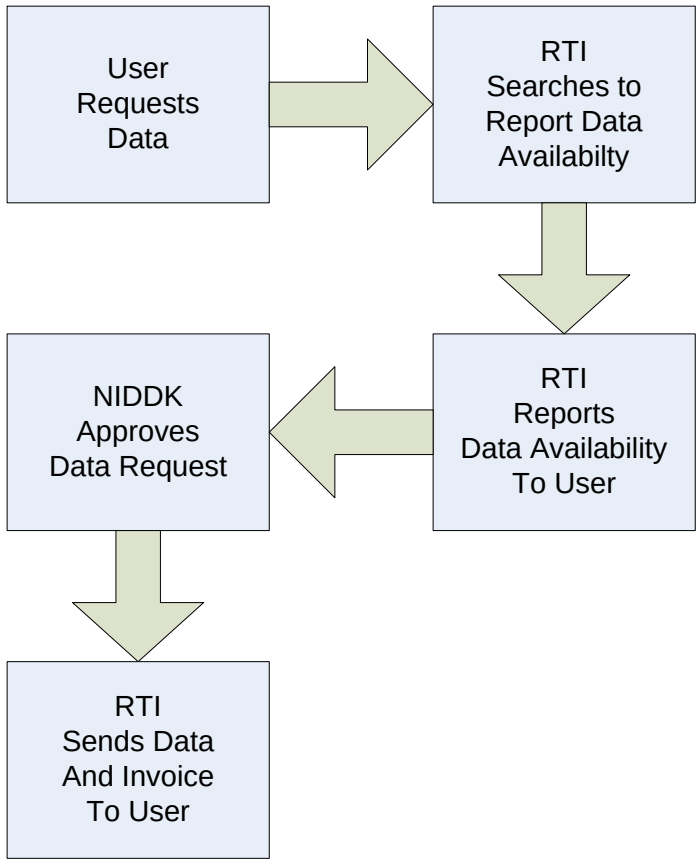
Synonyms – multiple definitions for one term

Created when vocabularies are merged

Case Study: NIDDK Central Repositories



Data Release (simplified)



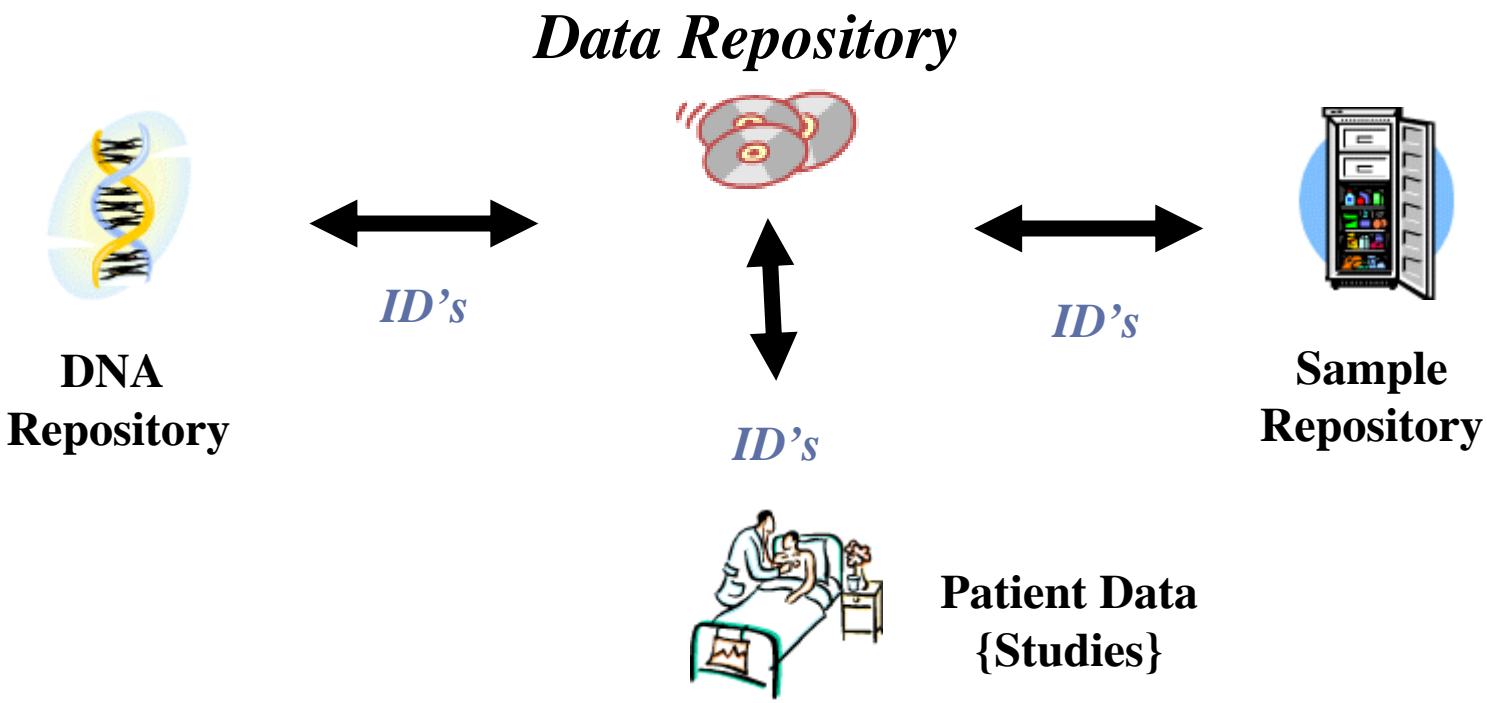
Two Challenges

1. To keep all the repositories “in synch”
2. To attempt to consolidate via a meta-dictionary identical data types

Synching Data

- Automated
  - All three databases are linked via the Internet and when one gets updated they all do
- File Push
  - At regular times files are sent to the Data Repository with identifiers to link samples to patients
- Chaos
  - Phone calls, data in all different formats (spreadsheets, word documents, SAS), finger pointing

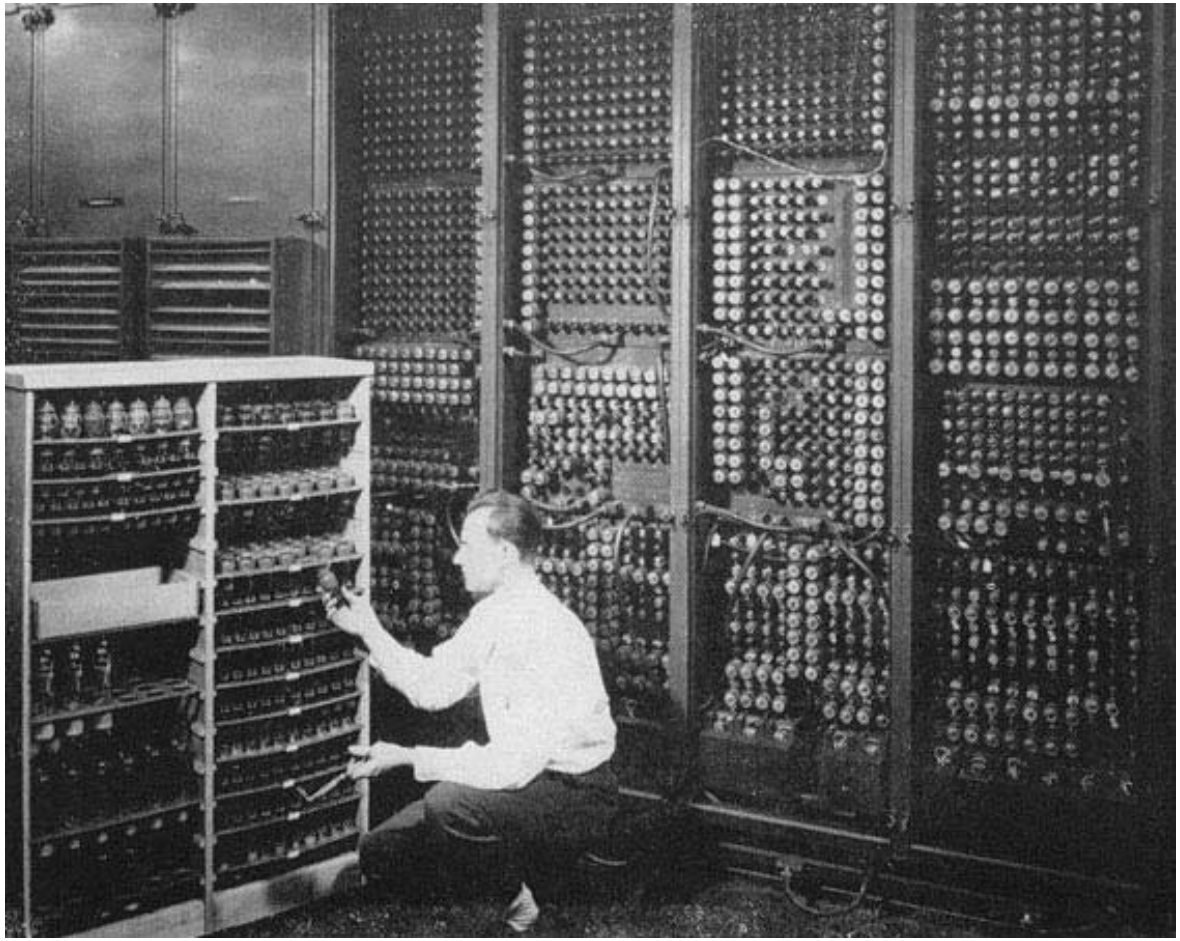
Central Identification



(Im)Perfect Example

Study_name	Rep1_count	Data_count	Diff
A	4245	4029	216
B	5148	5072	76
C	6311	6306	5
D	3664	3667	-3
E	4	4	0
F	947	947	0
Ga 1388	1950	1950	0
Gb 562			

Data Sets



Common Variables from SAS Files

Identify common variables using SAS PROC CONTENTS files

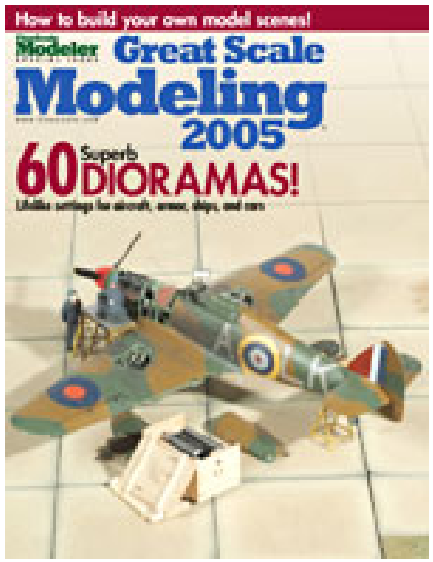
The SAS System 10:12 Monday, June 20, 2005 51
The CONTENTS Procedure

-----Variables Ordered by Position-----

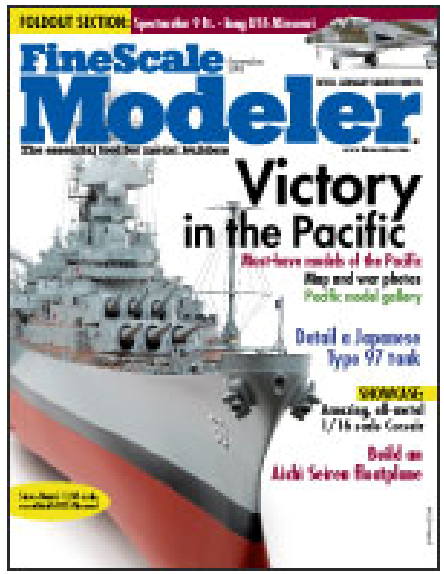
#	Variable	Type	Len	Format	Informat	Label
1	Barcode	Char	10	$10.	$10.	Participant ID
2	DN_Summary	Num	8			Meets diabetic nephropathy criteria
3	N_Summary	Num	8			Meets nephropathy criteria
4	DM_Summary	Num	8			Meets diabetes criteria
5	Dial_Trans	Num	8			Dialysis or transplant patient
6	Ht_cm	Num	8			Height used for BMI calculation
7	Wt_kg	Num	8			Weight used for BMI calculation
8	BMI_kgm2	Num	8			Body Mass Index (kg/m2)
9	CIDR	Num	8			Genotyped in first CIDR shipment
10	Proband	Num	8			Proband or not
11	Ret_Substudy	Num	8			Enrolled in Retinopathy Ancillary Study
12	Ace	Num	8			Taking ACE Inhibitors
13	Age	Num	8			Age at Enrollment
14	DM_Age	Num	8			Age at Diabetes Onset

Comparing Studies

- Compare studies by identifying common variables
- Are several clinical trials measuring the same things?
- Similarities of several trials could be summarized with <500 common variables
- Are studies measuring and entering data the same way?
- Different units, different assays
  - This would require a lot more records


How to build your own model scenes!
Great Scale
Modeling
2005
60 Superb DIORAMAS!

Modeling


FOLDOUT SECTION:
FineScale
Modeler
Victory
in the Pacific
Detail a Japanese
Type 97 tank
SHOWCASE:
Build an

Concepts in Bioinformatics for Modeling

Computing Support and Spill-over Capabilities

- Furnish/maintain LINUX clusters or similar “big iron”
- Develop/maintain a web portal

Modeling Support

- Data collection activities to support model parameter estimation and validate model predictions
- Model refinement activities
- Model architecture assessments

Portal

- Public and private web access to resources

Computing Support and Spill-over Capabilities

Central Resources might provide:

- Development & limited production
- Transparent access to external supercomputers
- Usage statistics collection across all systems
- Simulation experiment data warehouse
- Specialized tools (GIS, analysis, visualization).

Model Types

- Big ones
- Small ones
- Agent Based
- Equation Based

“Spillover” Concept

- Prepares for “surge” capacity.
- Created a change of concept for HPC:
  - No longer the “final stop” for computing but rather,
  - the entry point to larger computing facilities

Methods Used to Spillover

- Implement spillover feature – transparently adding nodes to a cluster for research
- Provide a capacity to add hundreds to thousands of nodes and the associated storage capacity
- Maintain a “owned” cluster as the focal point
- Add features to the “owned” cluster to enhance redundancy and continuous operation
- Provide flexibility to balance needs and budget

Methods Used in Spillover

Develop a system of standard workload scheduler queues across all resources that standardize the user interface

- Queues tailored to program type and capacity needs
- Queues to support development environment
- Queues to support production environment
- Queues to support surge environment

Model Support Infrastructure

Objectives

- Collect and maintain data, and software to support model development on a High Performance computing environment
- Capture models and place into a professional, maintainable and secure production environment

General Modeling Support

Computer Science Methods to create a system

- Define a vision of what the system does
- Develop Use case scenarios
- Develop system requirements
- Develop a prototype system
- Test-release andTest-release
- Finalize

Developing a Model Repository

Purpose

- **Develop a model release environment for production**
- **Place models and model results into a DBMS**
  - **Retrieve model results**
  - **The code that generated results using a query tool**
  - **Rerun model**
  - **Generate summary reports**

Model Repository Attributes

- **Secure**
- **Controlled**
  - **Working with frozen code**
  - **Releasing models of known pedigree**
  - **QA/QC benefits**
  - **Can repeat runs without rerunning model**

Sample Model Release Process

Overview

- Model Development – verified and frozen
- Generate results
- Load results into data warehouse
- Retrieve results later via Queries from a relational database that resides on the portal
- Retrieve models via Queries from a relational database