Record Linkage

11:35 – 12:04 (Sharp!)

Rich Pinder

Los Angeles Cancer Surveillance Program

rpinder@usc.edu

NAACCR Short Course

Central Cancer Registries: Design, Management and Use

Presented at the NAACCR Annual Meeting

June 10, 2005

Introduction

- **One of the primary functions of a cancer registry is to bring together information describing the same individual from a variety of data sources.**
- **Humans can conduct this 'bringing together' (called record linkage) manually by visually comparing records from two separate sources.**

Introduction, cont

- **Approach becomes time consuming and tedious for a cancer registry**
  - **large volume of records cause manual methods to become inefficient and unworkable**

Introduction, cont

- **Multiple notifications of the same cancer likely from use of multiple sources of information**
  - **efficient record linkage procedures on same individual very important**
  - **failure in record linkage process results in missed cases and /or duplicate registrations**
- **Research questions often require linking external data against the registry**
  - **allows hypothesis testing not available using other methods**
  - **efficiency also a key feature**

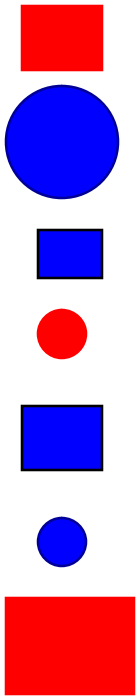
Introduction, cont

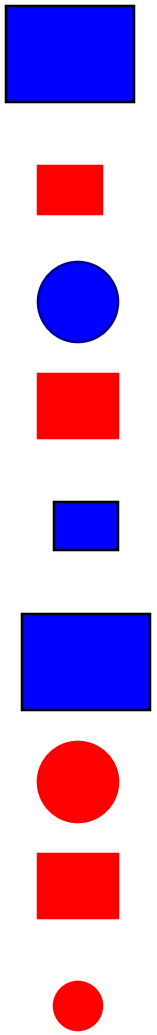
- **Technological advances in computer systems and programming techniques**
  - **economically feasible to perform computerized record linkage between large files quickly and reasonably accurately**
- **Best practices approaches**
  - **Cancer registries have wealth of experience from others to learn from**

Objectives

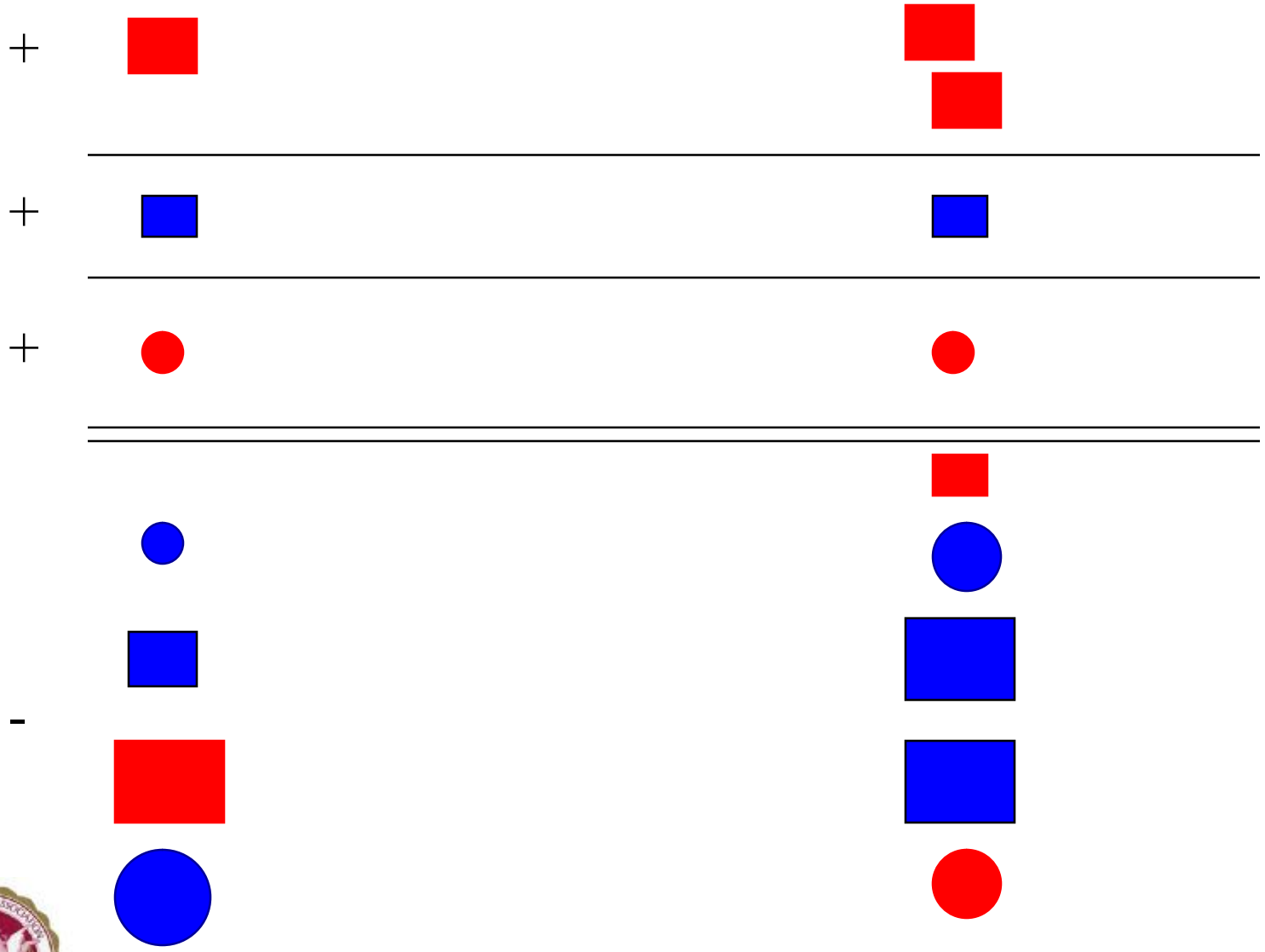
- **Introduce the concept of 'record linkage'**
- **Why it is important in the Registry**
- **Simple examples – the 'Deterministic' approach**
- **Adding some Smarts to the system – the 'Probabilistic' approach**
- **Software example – Link Plus**

- **Two ‘files’ :**



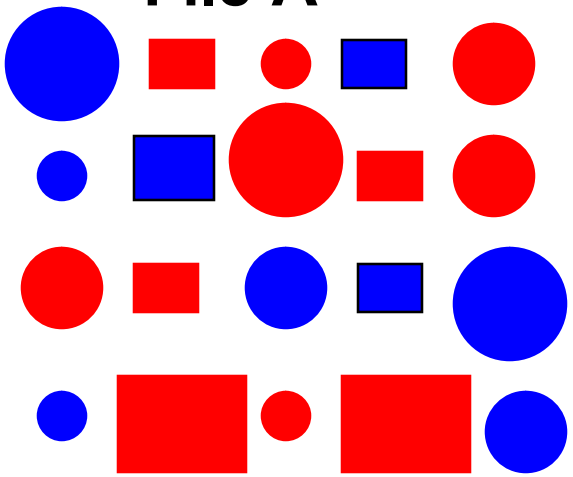


- **Linkage:**

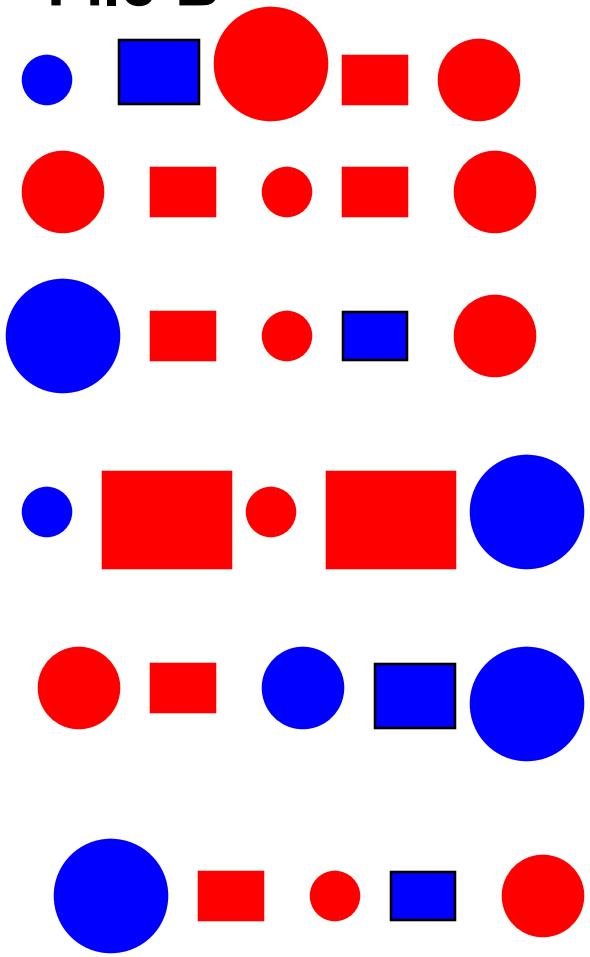


- **Two ‘files’ – records of variables:**

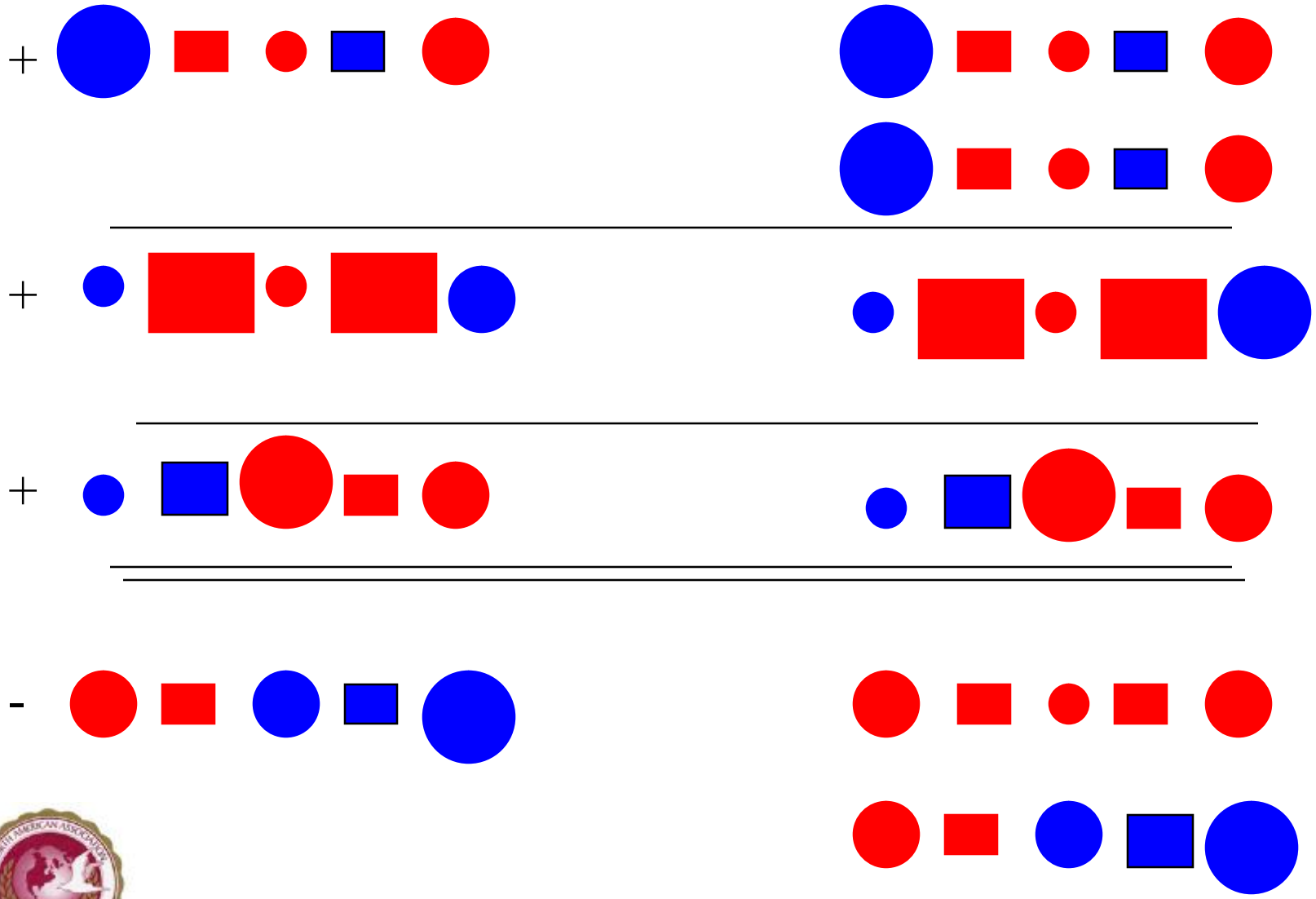
File A



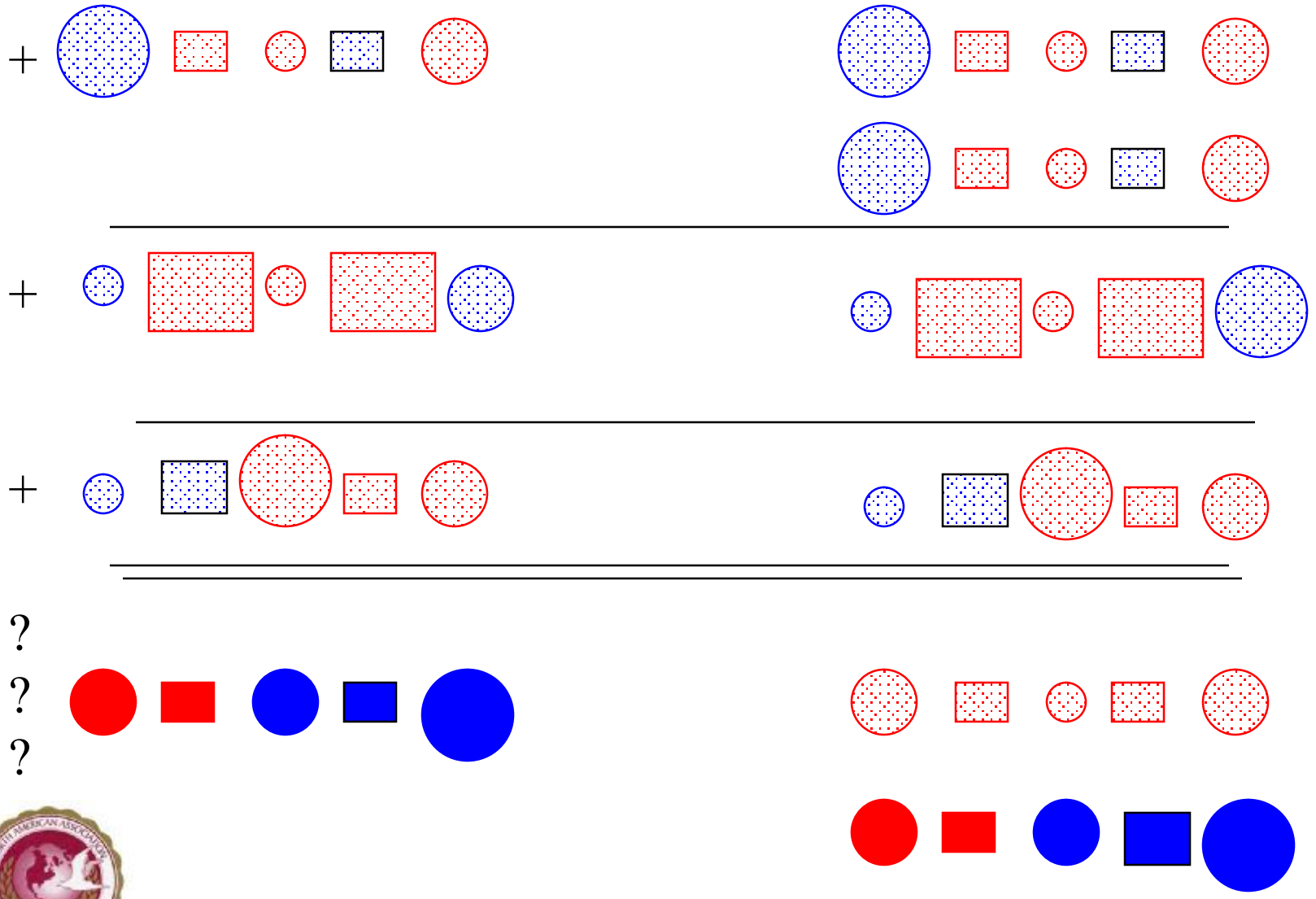
File B



■ Two ‘files’ - Linkage:



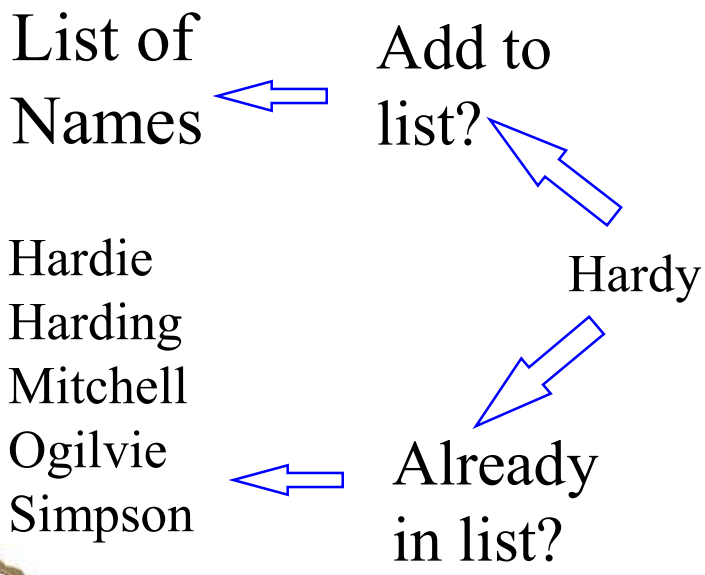
■ Two ‘files’ – Linkage ?:



Why Link Records?

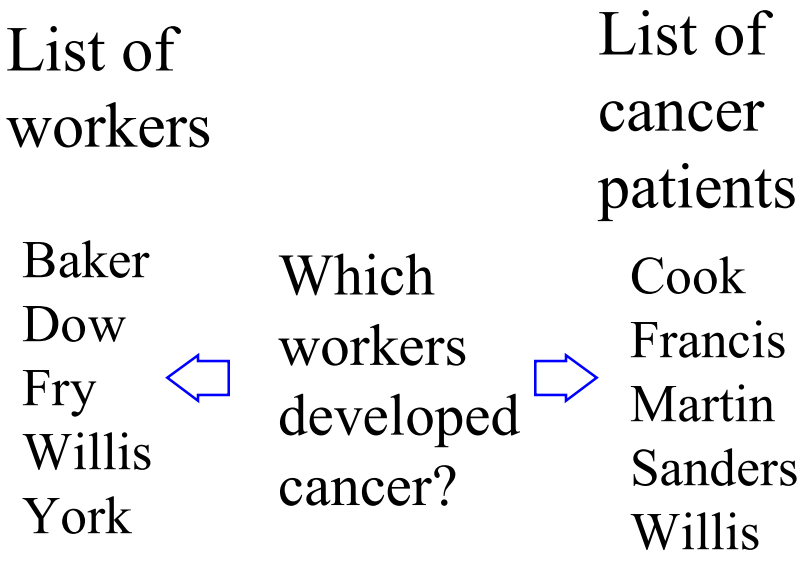
1) Registry Operations

Because you have a Master List and wish to add new names to it.



2) Research Linkages

Because you have two lists and wish to compare them.



Objectives

- Introduce the concept of ‘record linkage’
- Why it is important in the Registry
- Simple examples – the ‘Deterministic’ approach
- Adding some Smarts to the system – the ‘Probabilistic’ approach
- Software example – Link Plus

Why is it important

- Linking new reports, consolidation
- Duplicate detection
- Case Finding
  - Pathology labs
  - Disease index
  - Treatment centers

Linking New Reports

- Often paper lists are provided
- Included on the list are:
  - last name, first name
  - date of birth
  - site, histology
  - date of diagnosis
  - hospital number
- Then the decision is made manually if
  - This is a new person, or
  - New tumor for an existing person, or
  - New report for an existing person/tumor

Duplicate Detection

- **Fundamental requirement for accuracy and validity in our registries**
- **NAACCR yardstick**
  - **Maintain <= 0.1% (1 per 1,000) duplicates**

CaseFinding

- Matching reports from
  - Pathology labs
  - Disease index
  - Treatment centers
- If case is already reported then a new abstract does not need to be completed
- If new treatment is reported for an existing case then treatment alone can be added

Follow Up

Record Linkage used for large scale Follow up efforts:

- **Death Clearance – using State vital statistic file**
- **Hospital discharge data – Statewide utilization file**
- **Medicare – Center for Medicare Services**
- **Department of Motor Vehicles – Drivers information on drivers' licenses and renewals**
- **Medicaid - State's enrollees, targeting low income.**
- **Social Security Death Master File – SSA maintained file of death benefit claims**
- **Voter Registration/Voter History Files - Statewide file of last 6 elections.**
- **National Change of Address. U.S. Postal Service file of individuals reporting change of address in last 3 years.**

Objectives

- **Introduce the concept of 'record linkage'**
- **Why it is important in the Registry**
- **Simple examples – the 'Deterministic' approach**
- **Adding some Smarts to the system – the 'Probabilistic' approach**
- **Software example – Link Plus**

Deterministic approach

- **Computerize the comparison so that:**
  - **Require EVERYTHING to match**

Sndx	Last name	First Name	Site	SSNO	DOB	Sex	DateDx
T220	TEGHISH	MEGGIE	C739	123654789	02011918	2	06152004
T220	TEGHISH	MEGGIE	C739	123456789	02011918	2	06152004

Deterministic approach

- **But WAIT !**
  - **DX Info will often differ on true matches**

Sndx	Last name	First Name	Site	SSNO	DOB	Sex	DateDx
T220	TEGHISH	MEGGIE	C739	123654789	02011918	2	06152004
T220	TEGHISH	MEGGIE	C509	123456789	02011918	2	11152004

Deterministic approach

- **But WAIT !!**
  - **SSN sometimes is missing**

Sndx	Last name	First Name	Site	SSNO	DOB	Sex	DateDx
T220	TEGHISH	MEGGIE	C739	123654789	02011918	2	06152004
T220	TEGHISH	MEGGIE	C509		02011918	2	11152004

Deterministic approach

- **But WAIT !!!**
  - **Birthdate sometimes off**

Sndx	Last name	First Name	Site	SSNO	DOB	Sex	DateDx
T220	TEGHISH	MEGGIE	C739	123654789	02011918	2	06152004
T220	TEGHISH	MEGGIE	C509		02011913	2	11152004

Deterministic approach

- **But WAIT !!!!**
  - **What about minor misspellings ?**

Sndx	Last name	First Name	Site	SSNO	DOB	Sex	DateDx
T220	TEGISH	MEG	C739	123654789	02011918	2	06152004
T220	TEGHISH	MEGGIE	C509		02011913	2	11152004

Deterministic approach

- These previous slides highlight the use a 'DETERMINISTIC' algorithm for Record Linkage
  - Describes an algorithm in which the correct next step is PRE-defined
  - NOT a bad thing for production environments
  - Easily incorporated into existing data systems

  - Will miss significant numbers of true matchs

    OR

  - Will require HUGE amount of human intervention to REVIEW the results

Deterministic approach logic for Positive Match

- When social security numbers agree:
  - last name/aka matches last name
  - first name/aka matches first name
  - date of birth matches date of birth +/- 4 years
- When social security numbers agree:
  - nysiis(last name) matches
  - first name/aka matches
  - sex matches
  - date of birth +- 4 years matches

Deterministic approach logic for Possible Match

- When Date of Birth agrees:
  - Last Name/AKA matches
  - First Name/AKA matches
- When Date of Birth agrees:
  - Address at diagnosis matches
  - First Name/AKA matches
- When Social Security Number agrees:
  - Date of Birth +/- 4 years matches
  - Soundex (last name) matches

Objectives

- **Introduce the concept of 'record linkage'**
- **Why it is important in the Registry**
- **Simple examples – the 'Deterministic' approach**
- **Adding some Smarts to the system – the 'Probabilistic' approach**
- **Software example – Link Plus**

Probabilistic approach

- **Question:** What do the Humans know…. And how can we 'teach' the computer ??

- **Answer:** Best way is to build in the concept of PROBABILITY.

Probabilistic approach

Probability:

n 1: a measure of how likely it is that some event will occur; "what is the probability of rain?"; "we have a good chance of winning" [syn: chance] 2: the quality of being probable [ant: improbability]

The likelihood that a given event will occur: *little probability of rain tonight.*

Probabilistic approach

- **Recommended over traditional deterministic methods (i.e. exact matching) methods when:**
  - *coding errors*, *reporting variations*, *missing data* **or** *duplicate records* **encountered by registry**
- **Estimate probability / likelihood that two records are from the same person versus not**
- **Frequency Analysis of data values involved (and IMPORTANT)**

Probabilistic approach

- Landmark papers in computerized probabilistic record linkage by several Canadians in 1960s and 1970s (Fellegi & Sunter, Newcombe, Howe)
- Statistics Canada (in collaboration with NCIC) - developed the Generalized Iterative Record Linkage System - GIRLS (based on Fellegi-Sunter model)
  - Details in: Newcombe HB. Handbook of Record Linkage. Oxford University Press, 1988

Probabilistic approach

- **Frequency Analysis – examples:**
  - **How common is the surname ‘Takaharu’ in the Northern Texas Regional Cancer Registry?**
  - **How common is the surname ‘Takaharu’ in the Tokyo Cancer Registry ?**
  - **If you’ve got an ‘iffy’ match – and the Surname is ‘Rumplepinder’ – you likely to take it ?? (say ssn is missing, and mo/day of birth is wrong)**
  - **If you’ve got the same ‘iffy’ match – and the Surname is ‘Jones’ ???**

Probabilistic approach

- **Frequency Analysis – examples:**
  - **You're matching your Cancer file with the Mortality file. What are the impacts of a pair of 'John M Smith' matching with month/yr agreement on birth of 10/23….. Vs the same scenario but an agreement of birth of 10/79**

- **This is a HUGE component of probability**

Probabilistic approach

- **We (ie Registry Folk) intuitively KNOW things about our data**

  Common Names:

Last name	First Name	SSNO	DOB	Sex
TEGHISH	MEGGIE	123654789	02011918	2
TEGHISH	MEGGIE		02101913	2
GOMEZ	MARIA	321123444	04151939	2
GOMEZ	MARIA		04991939	2

Probabilistic approach

- **We (ie Registry Folk) intuitively KNOW things about our data**

The nature of our Cancer Data (C619 in 10 yr old?):

Last name	First Name	Site	SSNO	DOB	Sex
FISHER	JON	C619	331248080	02011918	2
FISHER	JON		331428080	10271995	2

Probabilistic approach

- **Formalization of intuitive concepts regarding outcomes of comparison of personal identifiers**
  - agreement **argues** for linkage **and** ***disagreement against*** **linkage**
  - **partial agreement is less strong than full agreement in supporting linkage**
    - **some types of partial agreements are stronger than others (e.g., truncated rare surname vs residence county code)**

Probabilistic approach

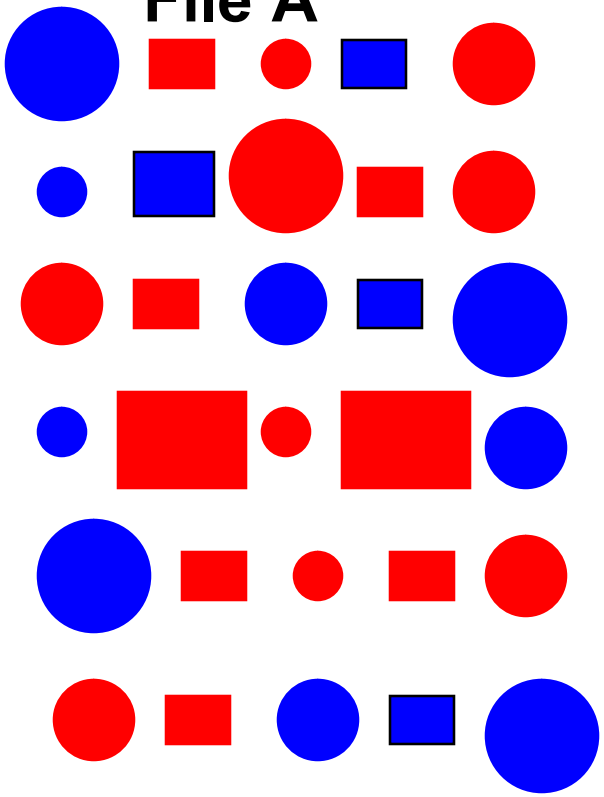
- **Agreement on an uncommon value argues more *strongly* for linkage than a common value (e.g., surname Drazinsky vs Smith)**
- **Agreement on a more specific attribute argues more strongly for linkage than agreement on a less specific one (e.g, SSN # vs sex variable)**
- **Agreement on more attributes, disagreement on few, supports linkage**

Probabilistic approach

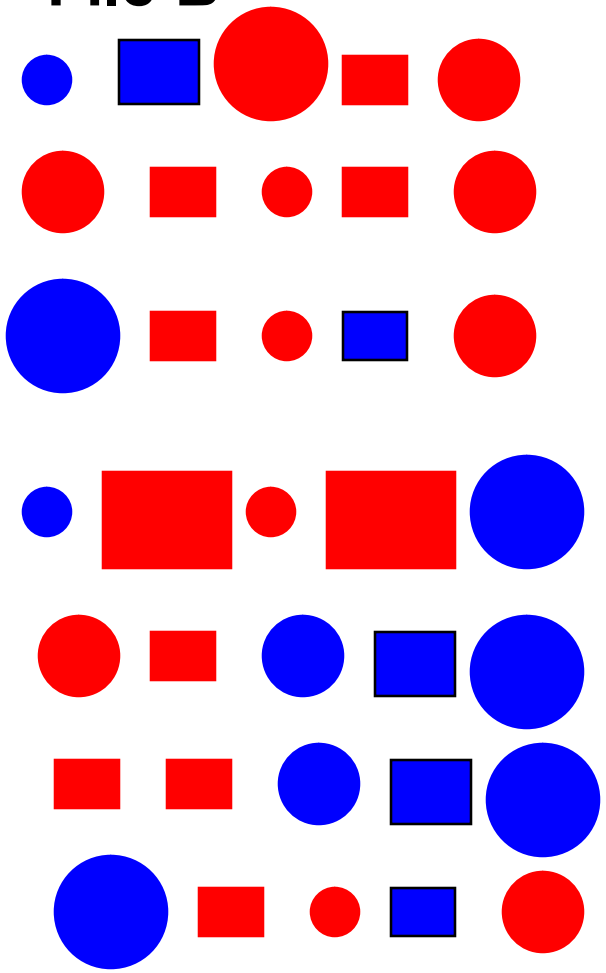
- **Blocking:**
  - **probabilistic linkage step that reduces the number of record comparisons between files**
  - **records for the two files / single file to be linked partitioned into mutually exclusive and exhaustive blocks**
  - **comparisons subsequently made *within* blocks**
  - **implemented by "sorting" the two files by one or more identifying variables**

- **Blocking – the concept:**

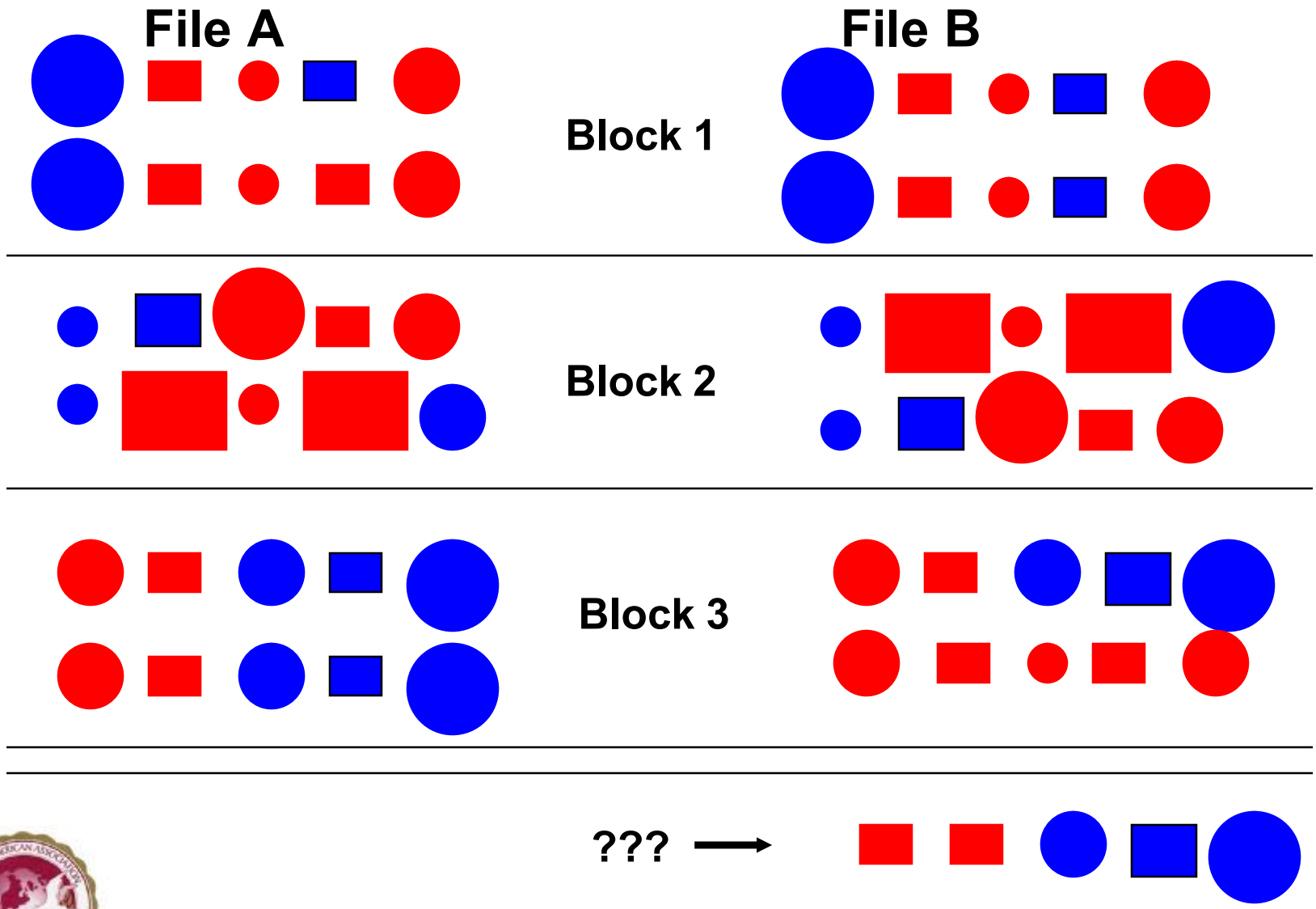
File A



File B



■ Blocking – the concept:



Probabilistic approach

- **Once comparisons within blocks are made:**
  - ***weight* calculated for each field comparison, and total weight derived by summing these separate field comparisons across all fields that have identifying value**
    - » **e.g., surname, given names, birth date**
- **Define thresholds for automatically accepting and rejecting a link**
  - **gray area / marginal links reviewed manually**

Objectives

- **Introduce the concept of 'record linkage'**
- **Why it is important in the Registry**
- **Simple examples – the 'Deterministic' approach**
- **Adding some Smarts to the system – the 'Probabilistic' approach**
- **Software example – Link Plus**

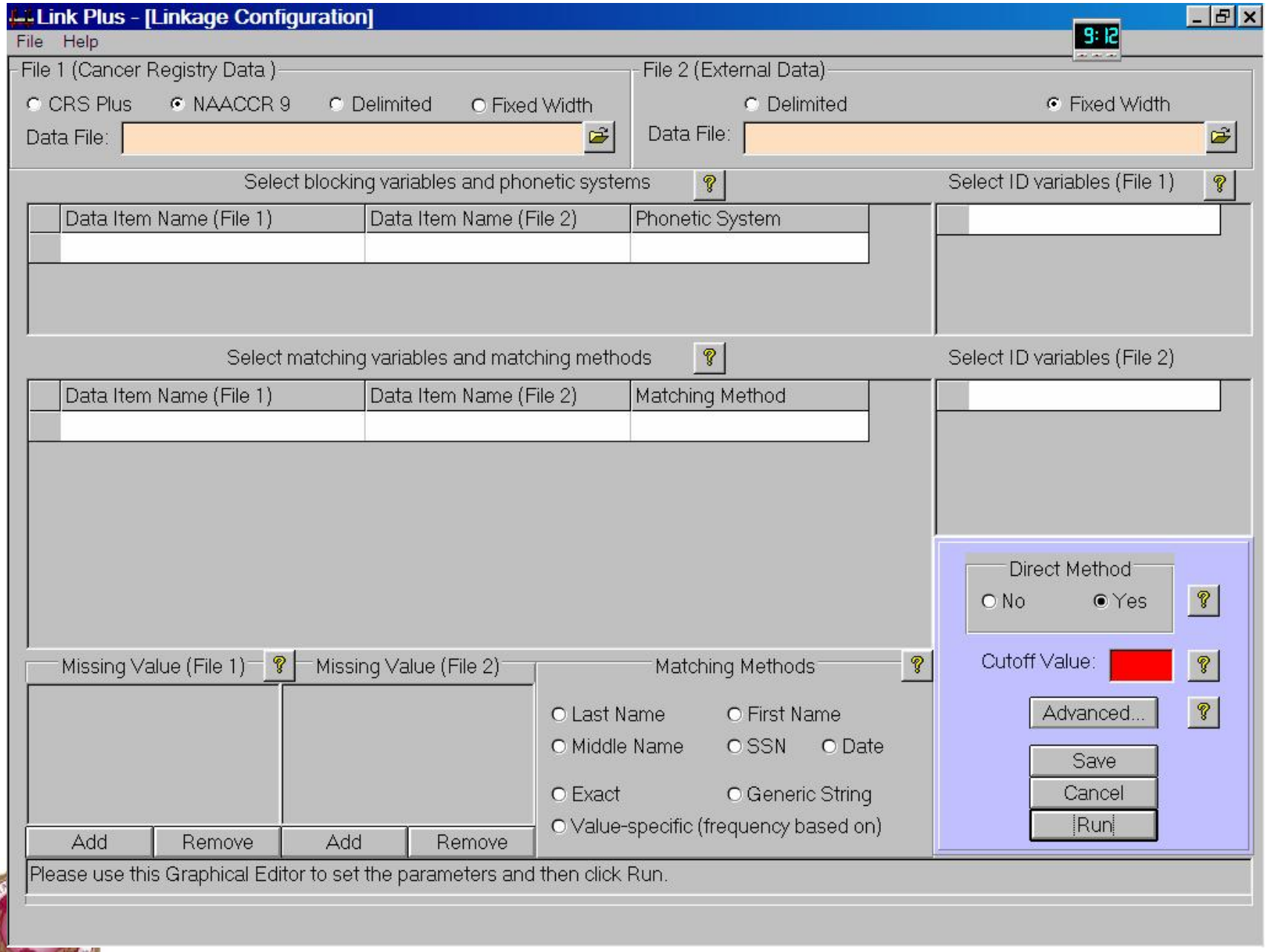
Record Linkage Software:

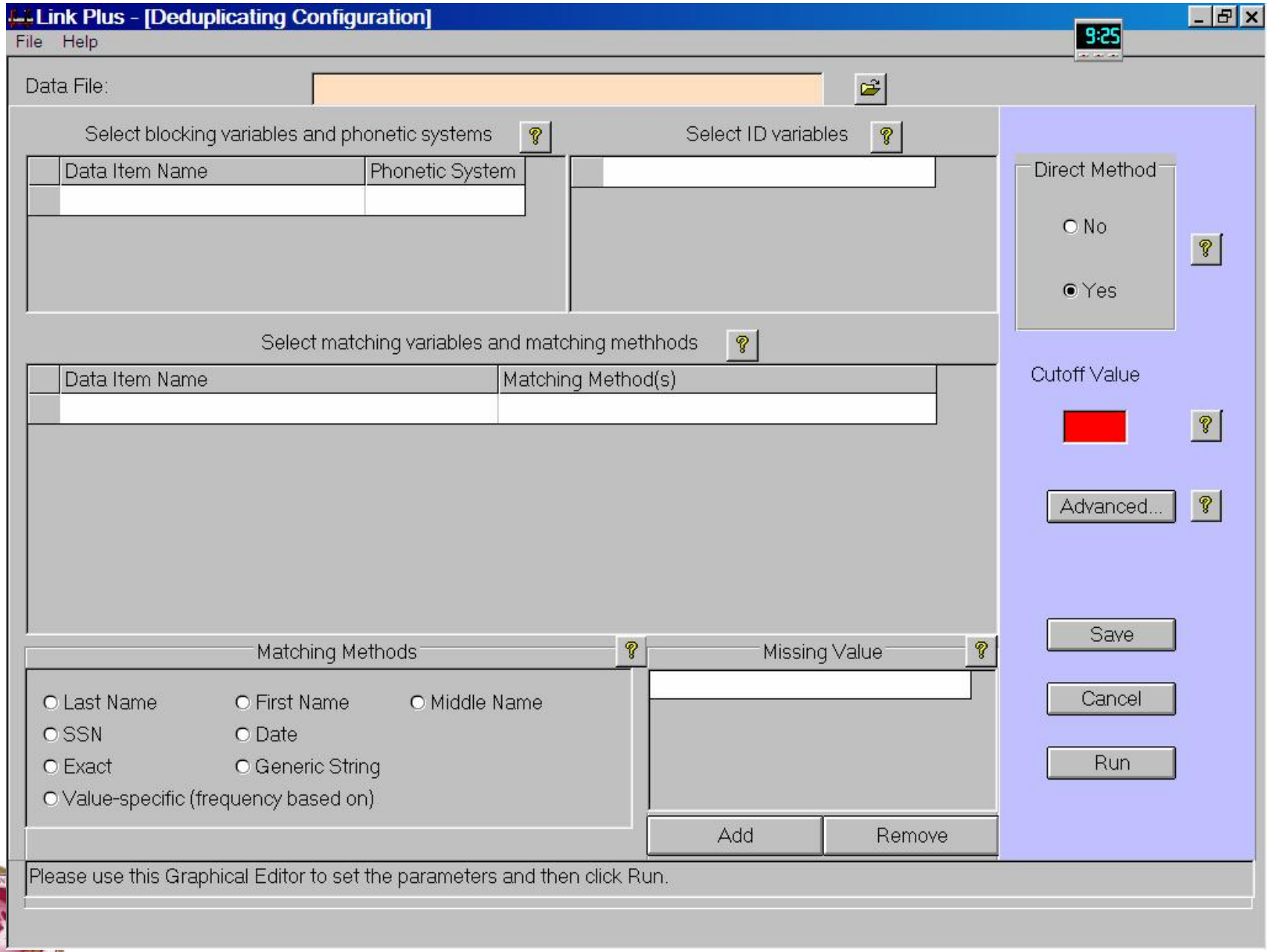
Link Plus
CDC & NPCR

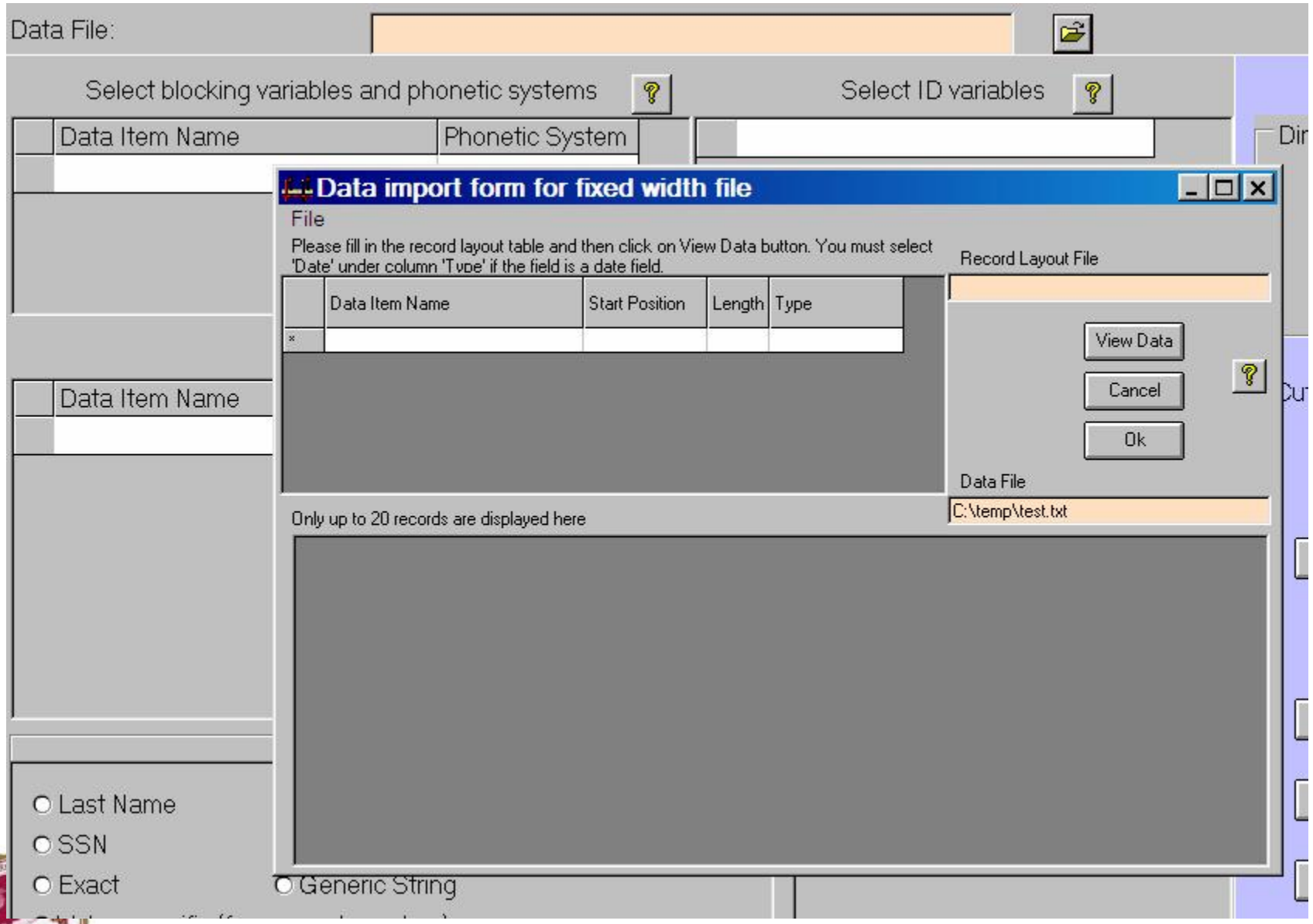
Joe Rogers, David Gu, Scott Van Heest

http://www.cdc.gov/cancer/registryplus/lp.htm

Boston, June 2005


Link Plus - [Linkage Configuration]
File Help
9:12
File 1 (Cancer Registry Data)
CRS Plus
NAACCR 9
Delimited
Fixed Width
Data File:
File 2 (External Data)
Delimited
Fixed Width
Data File:
Select blocking variables and phonetic systems
Data Item Name (File 1)
Data Item Name (File 2)
Phonetic System
Select ID variables (File 1)
Select matching variables and matching methods
Data Item Name (File 1)
Data Item Name (File 2)
Matching Method
Select ID variables (File 2)
Direct Method
No
Yes
Cutoff Value:
Advanced...
Save
Cancel
Run
Missing Value (File 1)
Missing Value (File 2)
Add
Remove
Add
Remove
Matching Methods
Last Name
First Name
Middle Name
SSN
Date
Exact
Generic String
Value-specific (frequency based on)
Please use this Graphical Editor to set the parameters and then click Run.


Link Plus - [Deduplicating Configuration]
File Help
9:25
Data File:
Select blocking variables and phonetic systems
Data Item Name
Phonetic System
Select ID variables
Direct Method
No
Yes
Select matching variables and matching methhods
Data Item Name
Matching Method(s)
Cutoff Value
Advanced...
Matching Methods
Last Name
First Name
Middle Name
SSN
Date
Exact
Generic String
Value-specific (frequency based on)
Missing Value
Add
Remove
Save
Cancel
Run
Please use this Graphical Editor to set the parameters and then click Run.


Data File:
Select blocking variables and phonetic systems
Select ID variables
Data Item Name
Phonetic System
Data import form for fixed width file
File
Please fill in the record layout table and then click on View Data button. You must select 'Date' under column 'Type' if the field is a date field.
Data Item Name
Start Position
Length
Type
Record Layout File
View Data
Cancel
Ok
Data File
C:\temp\test.txt
Only up to 20 records are displayed here
Data Item Name
Last Name
SSN
Exact
Generic String

Frequently Asked Questions (FAQ)

- **Q. Can Link Plus perform the data linkage on two data files if neither one is that cancer registry data file?**
  A. Yes.
- **Q. What data linkages can Link Plus perform?**
  A. Although designed for use in cancer registries, Link Plus can be used for general linkage purposes.
- **Q. If I want to link a file in NAACCR 10 format with another file, should I select file type 'Fixed Width' instead of 'NAACCR'?**
  A. Yes. At this time, Link Plus is set up to work with NAACCR layout version 9 only. (next version will contain NAACCR 10 – Fall 2005)

- **Q. What is the difference between using the matching method 'last name' and the 'value specific frequency' for the last name variable?**
  A. Last name matching uses not only value specific frequency matching but also adds approximate matching. It accounts for hyphenated names and possible interchange of last name and first name as well.
- **Q. How does one indicate null or empty fields as unknown values?**
  A. Null or empty fields are treated as unknown fields automatically.
- **Q. What is the maximum file size Link Plus can work with?**
  A. There is no maximum file size.

- **Integraged Clerical Review ??**

NEXT VERSION !!! Fall 2005