Collaborative Research Assistant

2007 Family History Technology Conference

- John Finlay
- Christopher Stolworthy
- Daniel Parker

Introduction

- This presentation will introduce the Research Assistant module for PhpGedView
- It was developed by students from Neumont University
- Tool designed to help genealogy researchers
  - Identify the problems
    - How the Research Assistant help to solve those problems.
  - Artificial Intelligence Techniques
  - Research Workflow
    - How the Research Assistant aids in the workflow

Identify The Problems

- Track research
  - Research is often duplicated due to inaccurate records
  - Research logs are not “nearby” when analyzing data
- Share research
  - How do I know what Uncle Bob in Ohio is researching?
  - What has he already done?
- Determine what to research
  - It can be difficult to analyze records and find the next thing to research
- Losing place
  - It is easy to forget where you were

Identify the Problems

- Enter results
  - There is a **MAJOR GAP** between the research results and the genealogy data
  - Consider the results of a census form and the wealth of data on it
  - Currently requires navigating through many, many different people and entering the same data over and over again

Identify the Problems

- Example 1930 Census

The same source data entered up to 23 times!

6 people in the family
Verify names, relations, gender

Occupations

Parents' Birthplaces

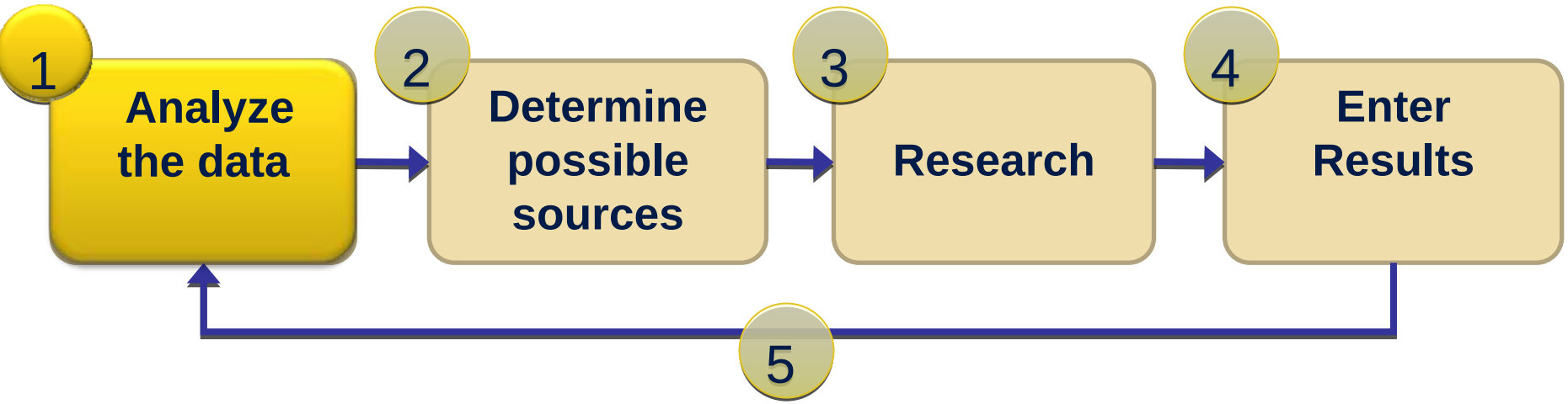
- Requires entering / validating
  - 6 Census facts
  - 6 Birth dates
  - 10 birth places
  - 1 occupation
  - 1 Marriage date
  - Possible notes about previous marriages, deaths of children, etc

Sharing & Tracking Research

- All research is tracked through a *Research Task*
  - Associated with multiple people/families
    - Keeps a log of all research done for a person
  - Associated with a specific source
    - Lookup multiple research tasks at once
  - Assigned to a family member who will complete the task
  - Kept with the genealogy data to simplify lookup and data entry

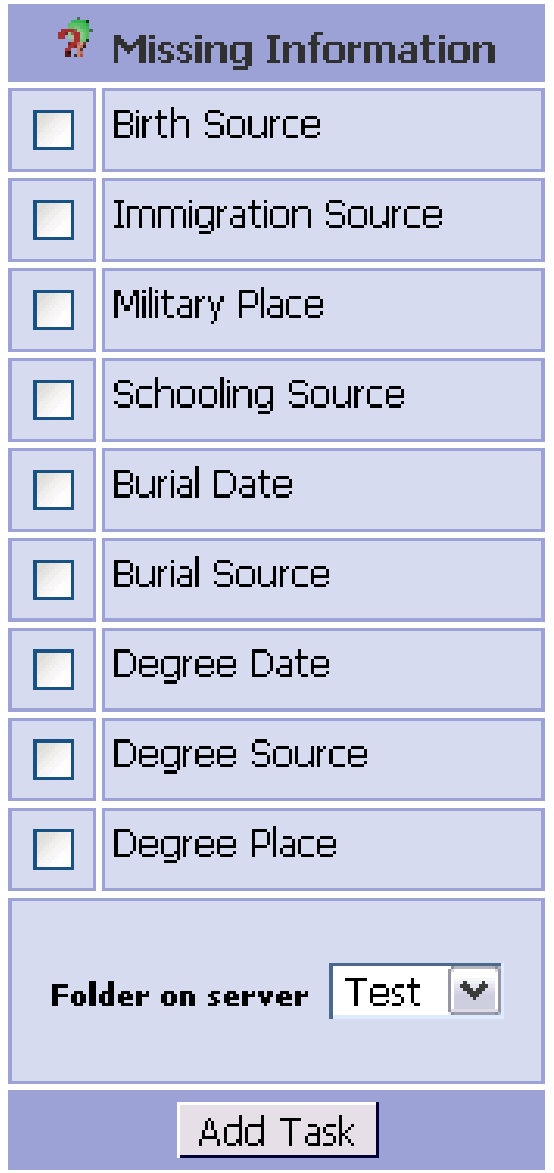
Tracking Research

- Research Workflow



Analyze the Data

- Missing Information
  - Analyze a record and suggest missing information
  - Automatically convert missing information into Research Tasks
- Nice, but how can we provide more?



Analyze the Data

- Bayesian Data Mining
  - Artificial Intelligence technique for predicting trends or highlighting anomalies in large data sets
  - Applied to Genealogy we can use it to help predict events and places for researchers
  - Help researchers narrow and focus their efforts
    - Most likely place
    - Most likely date
    - Most likely source

Analyze the Data

- Create correlation rules of interest
  - How does a child's surname relate to his parents' surnames?
  - How does a child's birth relate to his parents' birth?
  - Use these rules to calculate probabilities
- Each dataset is unique
  - Different cultures have different patronymics
  - Some groups tend to stay where they were born others where they were married
  - Correlation rules need to be uniquely calculated for different datasets

Analyze the Data

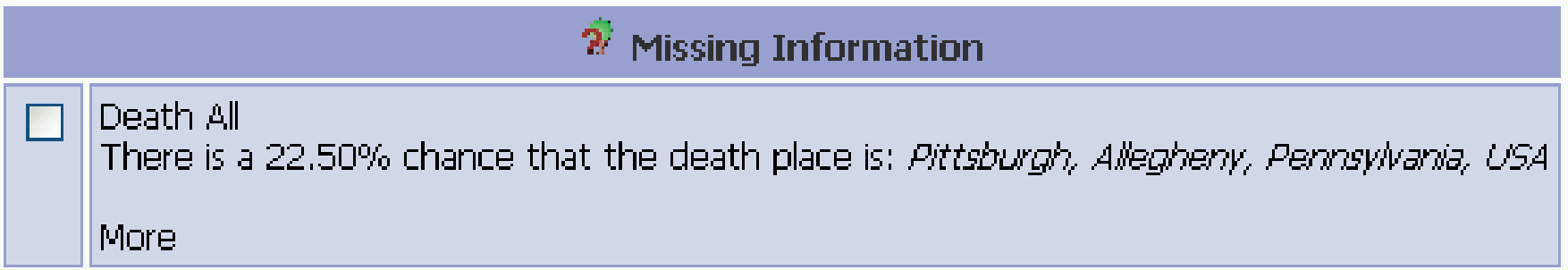
Data Correlations			
Local Data	Related Record	Related Data	Local Percentage
Surname	Father	Surname	86.82%
Death Place	Spouse	Death Place	46.59%
Burial Place	Self	Death Place	40.88%
Birth Place	Spouse	Birth Place	31.11%
Christening Place	Self	Birth Place	29.44%
Birth Place	Father	Birth Place	22.85%
Death Place	Marriage Place	Marriage Place	22.50%
Birth Place	Marriage Place	Marriage Place	22.42%
Birth Place	Mother	Birth Place	22.39%
Given Names	Paternal Grandfather's Given Name	Given Names	22.26%
Given Names	Father	Given Names	22.10%
Birth Place	Parents' Marriage Place	Marriage Place	21.02%
Given Names	Mother	Given Names	20.86%
Given Names	Maternal Grandmother's Given Name	Given Names	20.54%
Death Place	Self	Birth Place	19.12%

Analyze the Data

- Local Correlations
  - Calculate the rules with a smaller dataset
  - Localize the dataset around a person and their close relatives
  - Average the probabilities to get a more localized correlation

Analyze the Data

- We can now apply these correlations to our missing information
  - Suggest the most likely places for events to occur

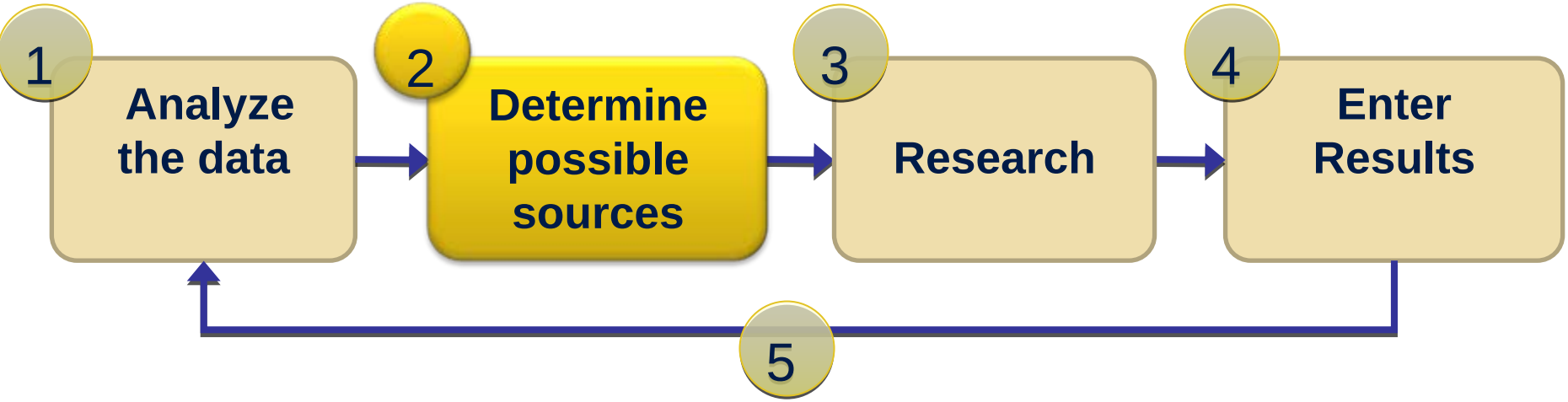


Analyze the Data

- Future work to do:
  - Possibility for AI to infer its own rules as it analyzes the data
  - Combine probabilities for rules that have matching data
    - What is the probability that the death place is Indiana given that the birth and marriage place are Indiana
    - More Bayes law
  - Broaden place localities
    - Currently only match on exact place match
    - Broaden to match on county and perhaps state

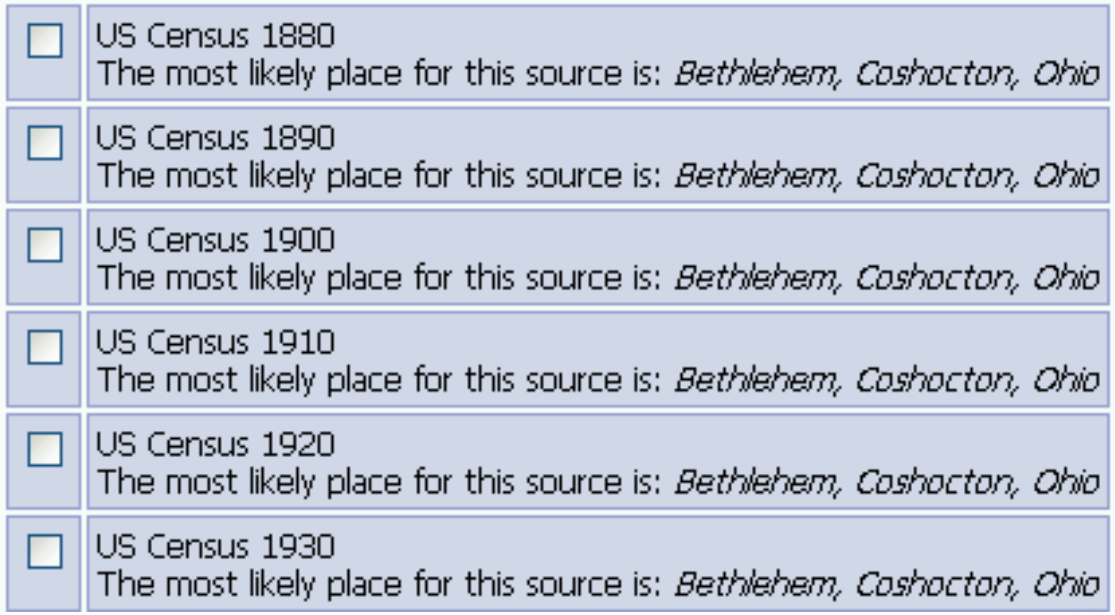
Tracking Research

- Research Workflow



Determining Possible Sources

- Help the researcher determine possible sources of their information
- Requires a database of source information to look in
- Example to the right shows supplementing missing information with US census sources

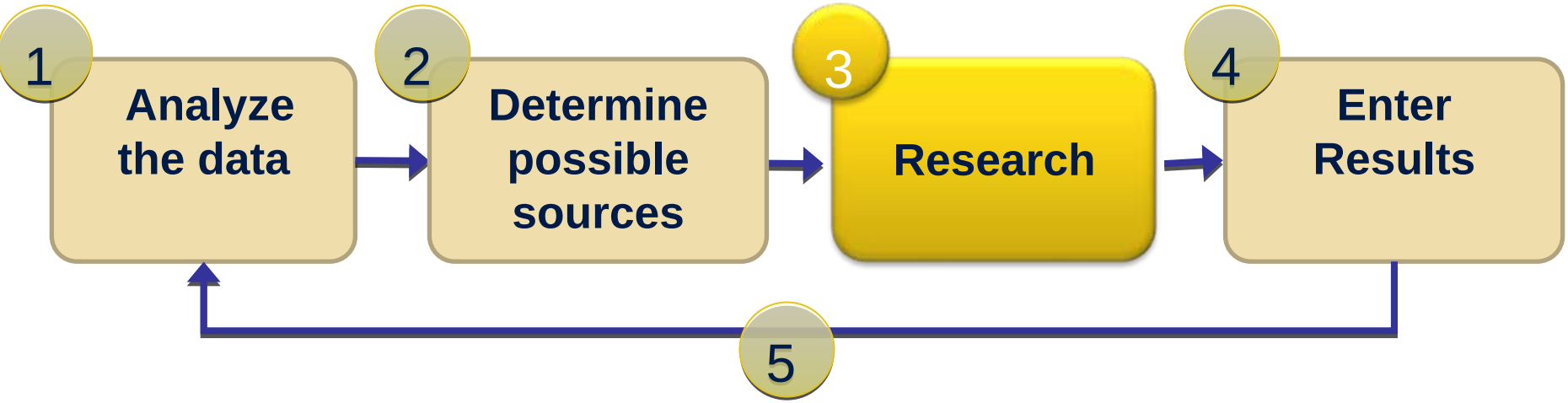


Determining Possible Sources

- Future Work
  - Improved locality search. Again to broaden the search to match on county and state.
  - Tie it into the FHL Catalogue
  - Common global repository for sources with a Web Service API we can query

Tracking Research

- Research Workflow

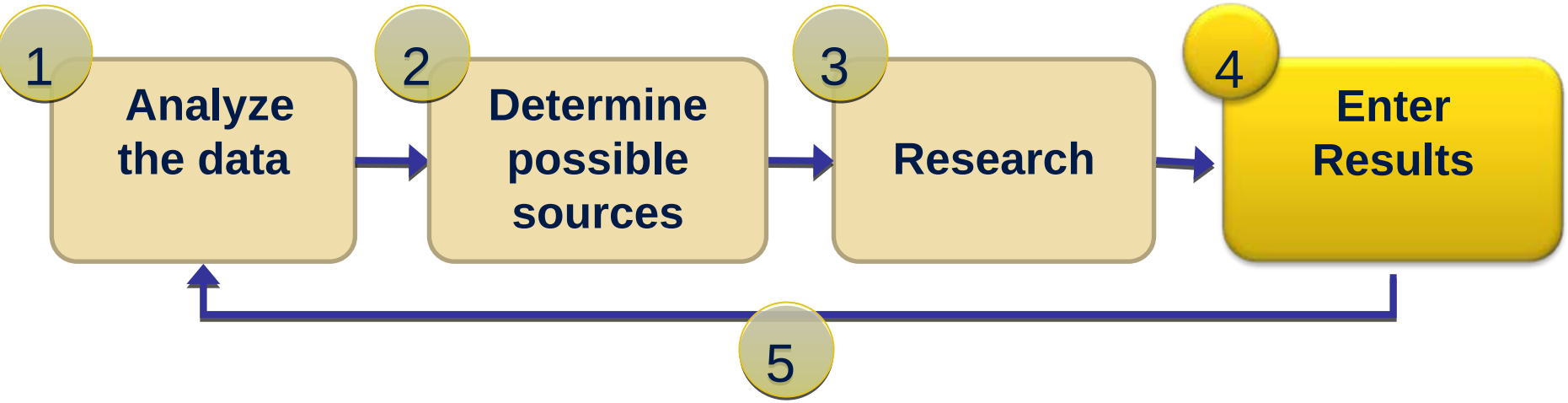


Research

- Auto-Search Assistant
  - Automatically pull data from a person's record so that it can be searched more easily
- Pluggable Architecture
  - Easy to add new sites to search
- Demonstration:
  - http://localhost/pgv-nu/individual.php?pid=I6541&ged=test.ged&tab=5

Tracking Research

- Research Workflow



Entering Results

- Unique Source citation forms
  - Enter in data the way it appears in the source record
  - Enter data only once!
  - Structured forms allow us to automatically infer facts
  - Pluggable architecture allows us to easily add new forms

- Remember the 23 things to enter from the census record?
  - Demonstration
  - http://localhost/pgv-nu/individual.php?pid=I716&tab=5

Conclusion

- PhpGedView Research Assistant Module simplifies technology for genealogy researchers
  - Aids in analyzing data through artificial intelligence techniques
  - Helps researchers find possible sources
  - Brings research tools closer to the data
  - Simplifies data entry
  - Distributed, Collaborative