A Statistical Method for Integrated Data Cleaning and Imputation

Chris Mayfield, Jennifer Neville, Sunil Prabhakar

Real Data is Dirty

And inconsistent, inaccurate, incomplete, outdated, corrupted, ...

- Large majority of data management time spent cleaning up problems
- Decisions based on "dirty data" costs billions of dollars annually

Goal: automated techniques for data quality assessment

- Discover correlations between data attributes
- Localize tuples that violate these dependencies
- Automatically detect/correct subtle errors
- Infer ranges/distributions for missing values

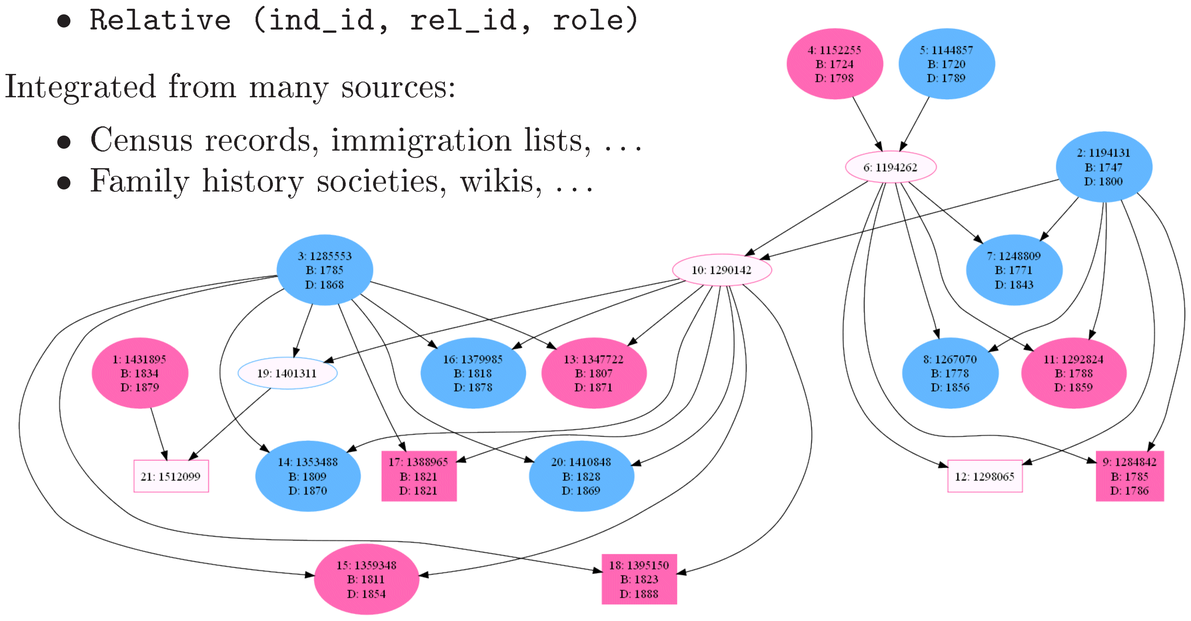
Example: Genealogy Database

Schema:

- `Person (ind_id, birth, death)`
- `Relative (ind_id, rel_id, role)`

Integrated from many sources:

- Census records, immigration lists, ...
- Family history societies, wikis, ...



Baseline Approach

Bayesian network model

- Learn CPDs from data:
  ```
  SELECT birth, death, count(*)
  FROM person
  GROUP BY birth, death
  ```
- Standard algorithms for exact inference (e.g. junction tree)

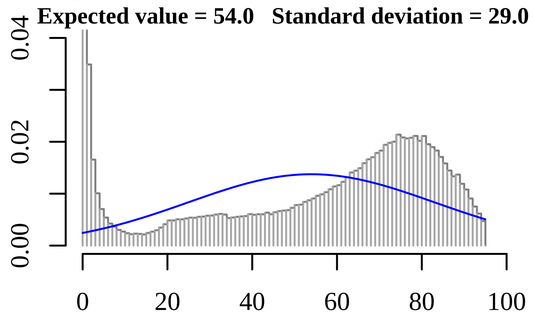
Limitations:

- Too many model parameters (potentially millions)
- Fixed CPD structure (e.g. always two parents)
- Exact inference algorithms are NP complete
- Propagation of errors in underlying data

Convolution Models

Learn the *difference* between related attributes:

```
SELECT model(death - birth) AS age
FROM person
```

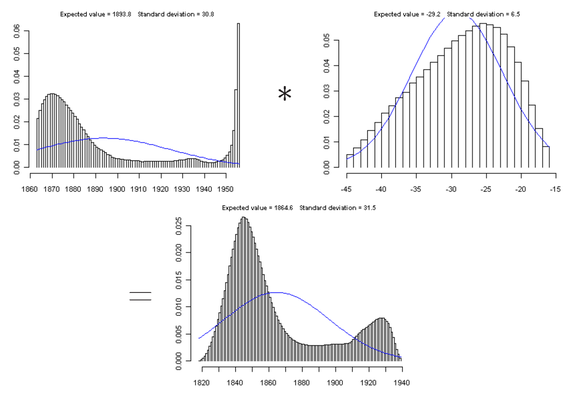


Inference using pdf *addition* (discrete convolution)

Approximate Inference

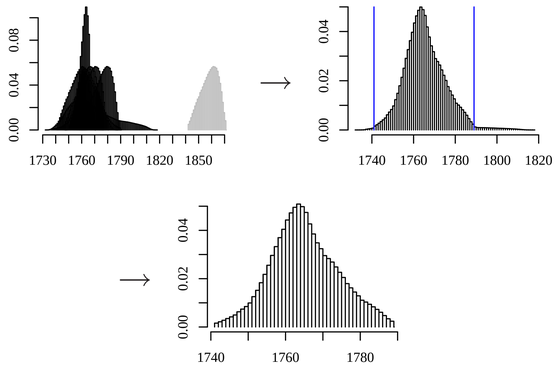
Step 1: Apply

For each value modified in the previous round, construct an output distribution for each model and relative.



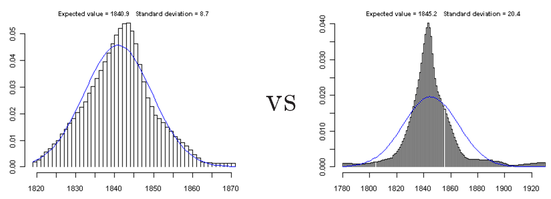
Step 2: Combine

Aggregate and normalize the resulting predictions for each imputation.



Step 3: Evaluate

Accept or reject the resulting distribution by comparing it with the previous version (if any).

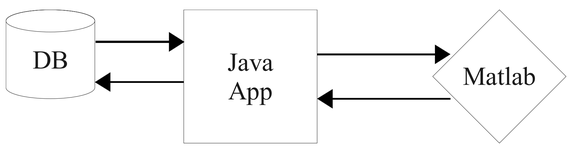


Data cleaning step:

- Is original evidence within expected range?
- Replace outliers with inferences

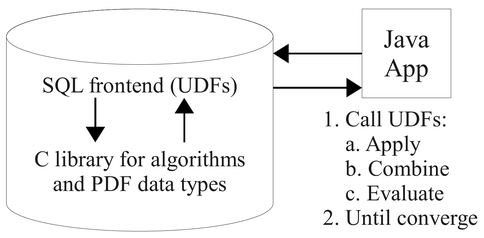
Implementation

Several hours, 2-3 GB RAM



1. For each test blanket:
   a. Query/topsort BN
   b. Query/refine CPDs
   c. Run Matlab (BNT)
   d. Store results in DB

c1. CPD matrices
c2. BN structure
c3. Bind evidence
c4. Run inference
c5. Extract results

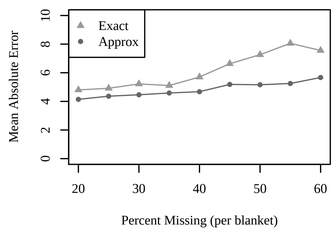


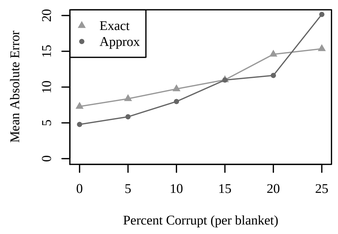
1. Call UDFs:
   a. Apply
   b. Combine
   c. Evaluate
2. Until converge

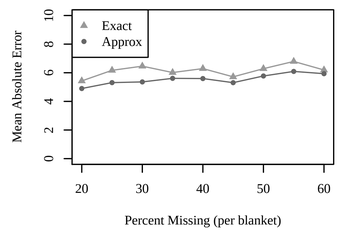
Several minutes, 5-10 MB RAM

Results

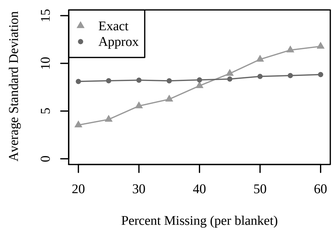
Accuracy:

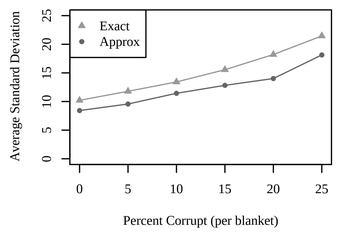


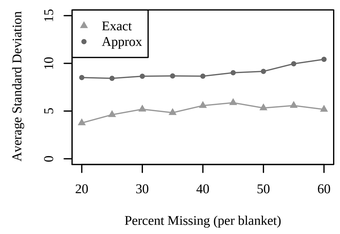




Quality:







Missing data | Corrupted data | Real data (5M people)

Summary

Database-centric approach to approximate inference

Accuracy/quality comparable to Bayesian networks, plus:

- Data cleaning suggestions with high precision
- Significantly more efficient