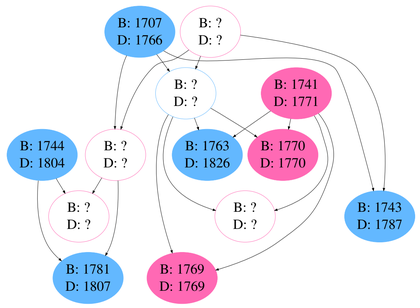
ERACER: A Database Approach for Statistical Inference and Data Cleaning

Chris Mayfield, Jennifer Neville, Sunil Prabhakar
Department of Computer Science, Purdue University

Example Databases

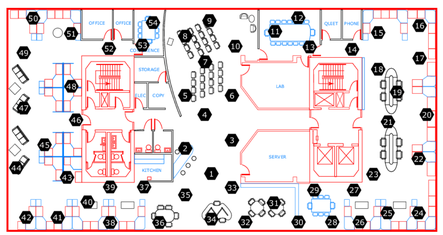
Genealogy Data

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



Sensor Networks

- `Sensor (epoch, mote_id, temp, humid, light, volt)`
- `Neighbor (id1, id2, distance)`



ERACER Framework

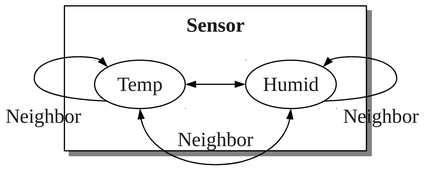
Learning (one time, offline)

1. Extract graph structure
2. RDNs learn parameters

Inference (multiple iterations)

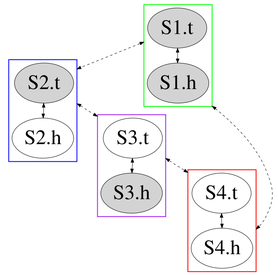
3. Apply component models
4. Combine inputs/outputs
5. Evaluate resulting pdfs
6. Repeat until converges

Model *template*



- Node for each attribute
- Edge b/t each attribute (both *within* and *across*)

Unrolled instance



(by tuple)

Key Design Choices

- Relational dependency networks
- Approximate inference algorithm
- Group by tuples (not attributes)
- Database implementation (UDFs)

Integrated data cleaning

- Run inference for *all* attributes
- Compare known values with pdfs
- Replace outliers with predictions
- Do not propagate suspected errors

SQL Interface (steps 3–5)

```
SELECT erace(i, n)
FROM node AS i
  LEFT JOIN link AS l ON i.nid = l.id1
  LEFT JOIN node AS n ON l.id2 = n.nid
GROUP BY i;
```

t_i	h_i
?	40%

t_n	h_n
21°	38%
35°	23%
24°	?

Query Output

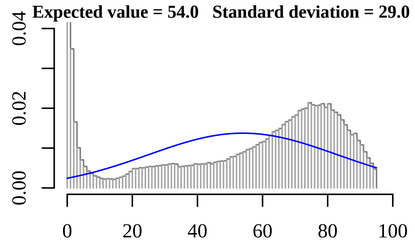
- PDFs for missing data
- Flags for dirty data

Component Models

Convolution

- *death age*: $M_{DA} = P(I.d - I.b)$

```
SELECT hist(death - birth)
FROM person;
```



Regression

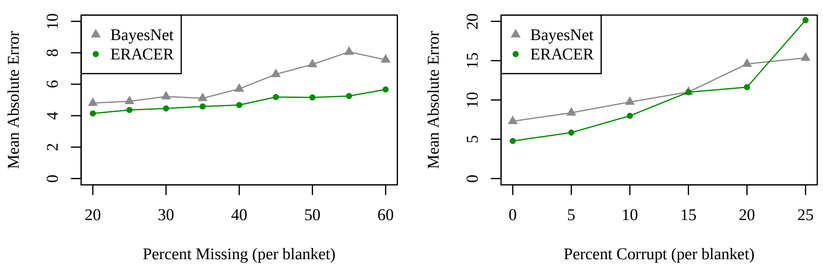
- *mean temperature*:
  $S.t \sim \beta_0$
  $+ \beta_1 \cdot S.h$
  $+ \beta_2 \cdot avg(N.t)$
  $+ \beta_3 \cdot avg(N.h)$

In General

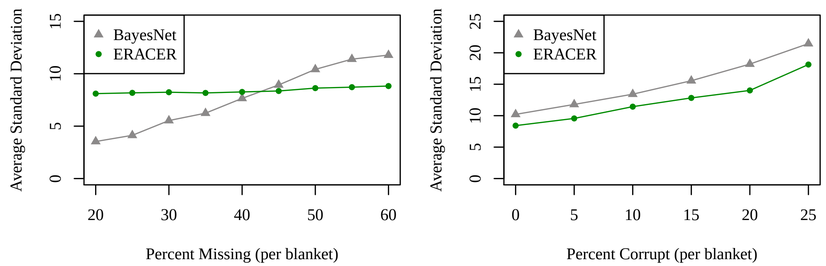
- Predict each attribute from others
- Use aggregation for heterogeneity

Selected Results

Accuracy of pdfs



Variance of pdfs



See the Paper!

Details for the `erace` aggregate

- Applying component models
- Data cleaning algorithms
- Inference feedback control

Other highlights

- Comparison to Bayesian networks
- Many more experiments & results