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)

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
- Expected value = 54.0   Standard deviation = 29.0

ERACER Framework

SQL Interface (steps 3–5)
- Use aggregation for heterogeneity
- Predict each attribute from others
- Apply component models
- Inference feedback control

Component Models
Convolution
- \( M_{DA} = P(I.d - I.b) \)

Regression
- \( S.t \sim \beta_0 + \beta_1 \cdot S.h + \beta_2 \cdot \text{avg}(N.t) + \beta_3 \cdot \text{avg}(N.h) \)

In General
- Predict each attribute from others
- Use aggregation for heterogeneity

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

Selected Results

Accuracy of pdfs
- BayesNet
- ERACER

Variance of pdfs
- BayesNet
- ERACER

See the Paper!
Details for the \( \text{erace} \) aggregate
- Applying component models
- Data cleaning algorithms
- Inference feedback control

Other highlights
- Comparison to Bayesian networks
- Many more experiments & results