Overview of deliverables

- Form groups (week 3)
- Project proposal (week 5)
- Literature survey (week 7)
- Midterm status report (week 10)
- Project presentations (last week of classes)
- Project report
Data errors

- Prevent
- Detect
- Repair
Data cleaning

- **Cost of data errors:**
  - Poor data quality costs the US economy more than $600 billion per year [Eckerson’02]
  - Erroneous price data in retail databases costs US consumers $2.5 billion each year [Fan et al.’08]
  - Consumes 30-80% of the development time and budget in data warehouses [Fan et al.’08]
Using FDs for cleaning

<table>
<thead>
<tr>
<th>CC</th>
<th>AC</th>
<th>PN</th>
<th>NM</th>
<th>STR</th>
<th>CT</th>
<th>ZIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>908</td>
<td>1111111</td>
<td>Mike</td>
<td>Tree Ave.</td>
<td>MH</td>
<td>07974</td>
</tr>
<tr>
<td>01</td>
<td>908</td>
<td>1111111</td>
<td>Rick</td>
<td>Tree Ave.</td>
<td>MH</td>
<td>07974</td>
</tr>
<tr>
<td>01</td>
<td>212</td>
<td>2222222</td>
<td>Joe</td>
<td>5th Ave</td>
<td>NYC</td>
<td>01202</td>
</tr>
<tr>
<td>01</td>
<td>908</td>
<td>2222222</td>
<td>Jim</td>
<td>Elm Str.</td>
<td>MH</td>
<td>07974</td>
</tr>
<tr>
<td>44</td>
<td>131</td>
<td>3333333</td>
<td>Ben</td>
<td>High St.</td>
<td>EDI</td>
<td>EH4 1DT</td>
</tr>
<tr>
<td>44</td>
<td>131</td>
<td>4444444</td>
<td>Ian</td>
<td>High St.</td>
<td>EDI</td>
<td>EH4 1DT</td>
</tr>
<tr>
<td>44</td>
<td>908</td>
<td>4444444</td>
<td>Ian</td>
<td>Port Pl</td>
<td>MH</td>
<td>W1B 1JH</td>
</tr>
<tr>
<td>01</td>
<td>131</td>
<td>2222222</td>
<td>Sean</td>
<td>3rd Str.</td>
<td>UN</td>
<td></td>
</tr>
</tbody>
</table>

Examples:
FD: STR → ZIP

cFD: [CC, ZIP] → STR, (44, _ | | _)

Challenge 1: discover FDs from data
Challenge 2: use FDs to clean the data

credit: Wenfei Fan
Data fusion

credit: Luna Dong, Divesh Srivastava
Ideas?

- Majority voting?
- What if websites copy information?
- Can you account for correlations?
- What if there are malicious contributors / spammers?
Data disambiguation
Preference-aware Integration of Temporal Data

ABSTRACT

A complete description of an entity of interest cannot always be obtained from a single data source, but rather, it is often necessary to combine data from multiple sources. Applications based on personal data, such as financial services or health care, therefore require effective integration of multiple heterogeneous data sources. To achieve this, we develop preference-aware temporal integration schemes. We motivate these schemes through costs and benefits of temporal integration and propose a set of desirable characteristics for such schemes. Our central assumption is that different users (or applications) need different levels of integration for the same data. In particular, we show how to weight these preferences based on the costs and benefits of data integration.

Bogdan Alexe  
IBM Almaden  
balex@us.ibm.com

Mary Roth  
IBM Almaden and UCSC  
torkroth@us.ibm.com

Wang-Chiew Tan  
UCSC  
tan@cs.ucsc.edu

Exploiting spatial overlap to efficiently compute appearance distances between image windows

ABSTRACT

We present a computationally efficient technique to compute the distance of high-dimensional appearance descriptor vectors between image windows. The method exploits the relation between appearance distance and spatial overlap. We derive an upper bound on appearance distance given the spatial overlap of two windows.

Bogdan Alexe  
ETH Zurich

Viviana Petrescu  
ETH Zurich

Vittorio Ferrari  
ETH Zurich
Ideas?

- Affiliation information?
- Co-authorship graph?
- Paper content?
- Crowdsourcing?
Data diagnosis
Post-factum cleaning with uncertainty

What caused these errors?

Data

- Periodicity $p$
- Has Signal? $h$
- Speed $s$
- Rate of Change $r$
- Average Strength $a$
- Zero crossing rate $z$
- Spectral roll-off $c$
- Average Intensity $i$

Transformations

- Is Walking?
  \[ M(p > P_w, R_s < r < R_w, -h \lor (s < S_w)) \]
- Is Driving?
  \[ M(p < P_d, r > R_d, h, s > S_d) \]
- Alone?
  \[ (A_2 \geq a > A_1) \lor ((a > A_2) \land (z > Z)) \lor ((a > A_3) \land (z < Z) \land (c > C)) \]
- Is Indoor?
  \[ M(\neg h, i < I_i) \]
- Is Meeting?
  \[ M(\neg h, i < I_m, a > A_m, z > Z_m) \]

Outputs

- true
- false
- false
- false

Sensors may be faulty or inhibited

It is not straightforward to spot such errors in the provenance
Post-factum cleaning with uncertainty

- What if we don’t know for sure that there is an error?
- What if we know for sure that there is an error, but we don’t know where?
Example: find all actors with Bacon number 2.

```
SELECT a3.fname, a3.lname
FROM Actor a0, Casts c0, Casts c1,
    Casts c2, Casts c3, Actor a3
WHERE a0.fname = 'Kevin' AND a0.lname = 'Bacon' AND
    c0.pid = a0.id AND c0.mid = c1.mid AND
    c1.pid = c2.pid AND c2.mid = c3.mid AND
    c3.pid = a3.id AND
    NOT (a3.fname = 'Kevin' and a3.lname = 'Bacon') AND
    NOT EXISTS (SELECT xc1.pid
                FROM Actor xa0, Casts xc0, Casts xc1
                WHERE xa0.fname = 'Kevin' AND
                      xa0.lname = 'Bacon' AND
                      xa0.id = xc0.pid AND
                      xc0.mid = xc1.mid AND xc1.pid = a3.id)
GROUP BY a3.id, a3.fname, a3.lname;
```
Ideas?

- Data sampling?
- Visualizations?
- Different query languages?
- Query by example?
  - Instead of writing a query, give examples of the results
- Can you query with natural language?
Query containment/equivalence

- **Equivalence:** \( q_1 \equiv q_2 \iff q_1(D) = q_2(D), \forall D \)

- **Containment:** \( q_1 \subseteq q_2 \iff q_1(D) \subseteq q_2(D), \forall D \)

- Query rewrite with views.

- Idea: use SAT solvers

**NP-hard**
Databases for applications

- A declarative language for animation
  - Create characters by “joining” existing components
  - Find rigs (skeletons) for a given character
Databases for applications

- Navigating and querying large image data

credit: Sloan digital sky survey
Probabilistic databases

- What are the probabilities in the join results?
- Can you find query classes for which inference is tractable?

credit: Dan Suciu
Provenance-based content authenticity

- Provenance: sources and process that derives new data.

  - Openly authored
  - Mistaken contributions
  - Malicious contributions

- Idea:
  - Use history to assess trustworthiness
  - Analyze edit behavior
  - Analyze citations
  - Propagate trust
Data auditing

- Example: medical records
  - Who should have access?

- Query language to model security violations
  - Query execution engine for this language

- Trust propagation
  - What is data gets tainted?
  - What other data is affected?
  - Which data is safe?
Data understanding

Support to understand surprising results

What if the results are not surprising? Can DBs add context?

Explain by example
Fairness in Big Data

Data-driven decisions
Where to offer services
Product recommendations
Who gets a loan
Crime sentencing
Self-driving cars
Work authorization (e.g., e-Verify)

Accuracy is better for majorities
Other projects

- ACM SIGMOD programming contest
- PVLDB experiment and analysis track
  - thorough evaluation of existing approaches
Brainstorming session

- Form groups of 6-7.
- Take 5 min to think individually and fill as much as you can the questions on the first page of the printout.
- Present your ideas in a circle as follows:
  - Each person describes their idea at high level. No more than 1 minute.
  - The person on his/her left repeats the idea in their own words. Then proceeds to describe his/her idea.
- Discuss and fill the second page of the printout