DIMACS Workshop
Opening-Closing Comments

Stephen E. Fienberg
Department of Statistics &
Center for Automated Learning and Discovery
Carnegie Mellon University
Pittsburgh, PA, U.S.A.
Some Integrative Themes

• Integrating diverse data sources
• Privacy/confidentiality
• Data across time and space
• Signal detection and setting cutoffs
• Datamining to the rescue?
• Models and methods of inference
Integrating Diverse Data Sources

• Public health data/non-traditional data
  – Grocery store sales
  – Pharmacy sales
  – School attendance records

• Matching records/identifiers?
  – Fellegi–Sunter and modern Bayesian embellishments
  – Capture-recapture methods for estimating population totals of exposure and infection
What Do Following Populations Have in Common?

- Fish
- Penguins
- Homeless
- Prostitutes in Glasgow
- Italians with diabetes
- Atrocities in Kosovo
- People in the U.S.
- People infected with HIV virus
- Adolescent injuries in Pittsburgh, PA
- WWW
# Multiple List Data for Query 140

\[ n=159 \]

<table>
<thead>
<tr>
<th>AltaVista</th>
<th>Northern Light</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes Excite</td>
<td>HotBot</td>
</tr>
<tr>
<td>yes Infoseek</td>
<td>yes Lycos</td>
</tr>
<tr>
<td>yes Infoseek</td>
<td>no Lycos</td>
</tr>
<tr>
<td>yes Infoseek</td>
<td>HotBot</td>
</tr>
<tr>
<td>no Infoseek</td>
<td>yes Lycos</td>
</tr>
<tr>
<td>no Infoseek</td>
<td>no Lycos</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>yes</th>
<th>no</th>
<th>yes</th>
<th>no</th>
<th>yes</th>
<th>no</th>
<th>yes</th>
<th>no</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes Infoseek</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>yes Infoseek</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>yes Infoseek</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>no Infoseek</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>8</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>yes Infoseek</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>yes Infoseek</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>
| no Infoseek     | 0   | 0  | 7   | 17 | 2   | 3  | 31  | \?


Let the $y_{ij}$’s be independent r.v.’s, with

$$p_{ij} = \Pr \{y_{ij} = 1\}$$

for page $i$ observed in list $j$, where

$$\log \left\{ \frac{p_{ij}}{(1-p_{ij})} \right\} = \theta_i + \beta_j \quad i = 1, 2, \ldots, N; \quad j = 1, 2, \ldots, k.$$ 

If we take into account individual heterogeneity represented by $\{\theta_i\}$, samples are “independent.”
Posterior Distribution of N for Query 140

n = 159

GL* Average = 165
GL* Max = 322
Privacy/Confidentiality

• Matching records raises major issues of privacy and confidentiality
  – Can we integrate sources without identifiers?
  – Role of intermediaries for linkage and then application of disclosure limitation methods
Conceptual Confidentiality Kernel

Confidentiality Checks: I

Data Sources

Disclosure Risk Low?

Confidentiality Checks: II

Data Merger (record linkage) Detection/Warning Kernel

Data Users
Time and Space

• Recording timing of occurrence of events is crucial component of data

• Data result in multivariate time series or point processes for events/purchases/reports
  – Multiple products purchased
  – Doctors visits
  – School absences

• Spatial information makes data sparser

• Crude counts versus individual records
Supermarket Sales Records

All Products
50,000

Produce

Dairy

Health & Beauty
2,050

Analgesics
650

Cough & Cold
850

Stomach
550

...
Confounding Natural Periodicities

Cough medication sales between 8/8/99-1/31/01
Signal Detection

- Adverse events \(\Rightarrow\) Discovery of cause
  - e.g., detecting signature of outbreak in response to anthrax attack

  - What about alternative explanations?
Setting Detection Cutoffs

- **Fixed thresholds?**

- **Tradeoff between false positives and false negatives**

- **Nature of followup?**
  - Back to privacy issues again
What Are We Looking For?

• Anticipating specific problems, e.g., in response to smallpox vaccination campaign

• Surveillance systems to measure everything
Datamining to the Rescue?

• **Bad News:**
  – For broad based screening and surveillance, \( p >> n \) and we encounter curse of dimensionality
  – Model selection on large numbers of features has major problems

• **Good News:**
  – For prediction we may be willing to settle for black box (or at least gray box) predictions
  – Datamining methods may turn out to be useful here but jury is out
Models and Inference Methods

• Black box approaches (including simple “robust” methods) versus models for underlying phenomena

• Frequentist vs. Bayesian methods
  – Specifying likelihood is hard
  – Picking priors based on real information or for smoothing is relatively easy

• First get statistical tools that work, and then figure out how to move them into the field or to approximate