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**

					Northern Light							
					yes				no			
					Lycos				Lycos			
					yes		no		yes		no	
					HotBot		HotBot		HotBot		HotBot	
					yes	no	yes	no	yes	no	yes	no
AltaVista	yes	Infoseek	yes	Excite	yes	1	0	2	0	0	0	1	0
					no	2	0	3	2	0	0	0	2
			no	Excite	yes	1	0	2	1	0	0	3	4
					no	1	3	0	8	2	0	3	19
	no	Infoseek	yes	Excite	yes	0	0	0	1	0	0	0	0
					no	0	0	1	1	0	0	5	4
			no	Excite	yes	0	0	0	1	0	0	4	22
					no	0	0	7	17	2	3	31	[5] ?

Simple Models Often Work

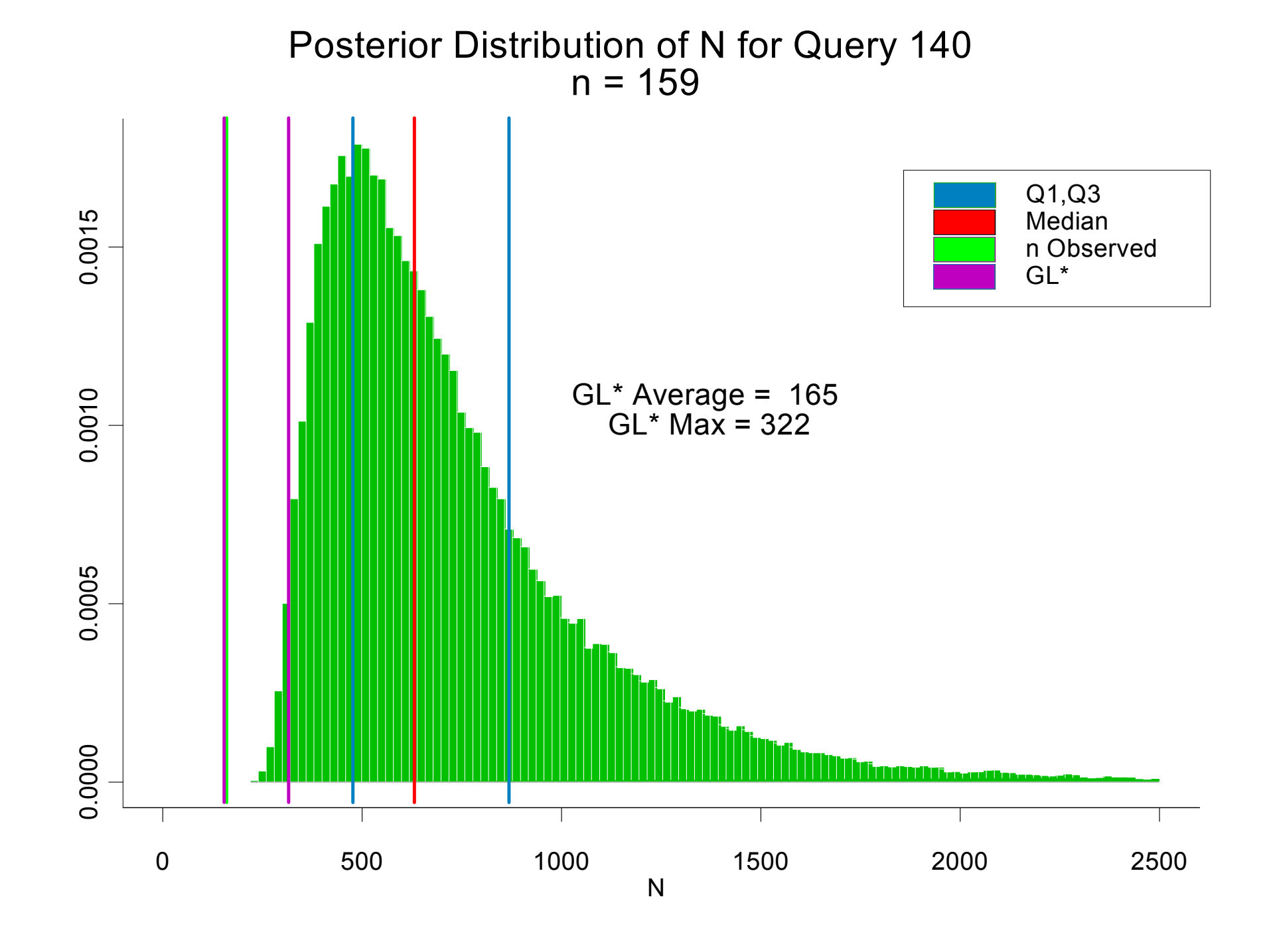
- Let the y_{ij}'s be independent r.v.'s, with

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

for page i observed in list j, where

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

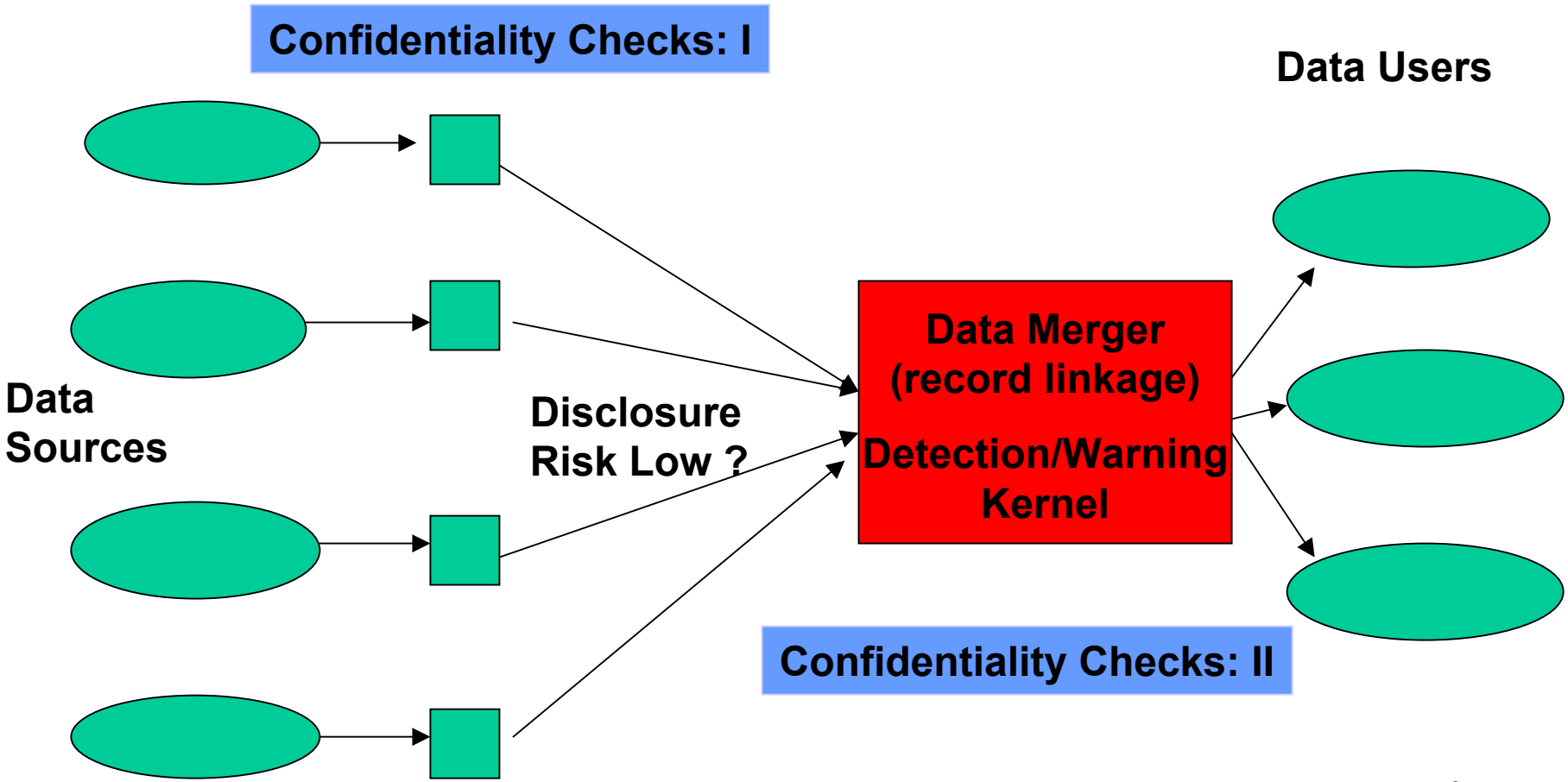
- If we take into account individual heterogeneity represented by $\{\theta_i\}$, samples are "independent."


Posterior Distribution of N for Query 140
n = 159
Q1,Q3
Median
n Observed
GL*
GL* Average = 165
GL* Max = 322
0.0000
0.0005
0.0010
0.0015
0
500
1000
1500
2000
2500
N

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**

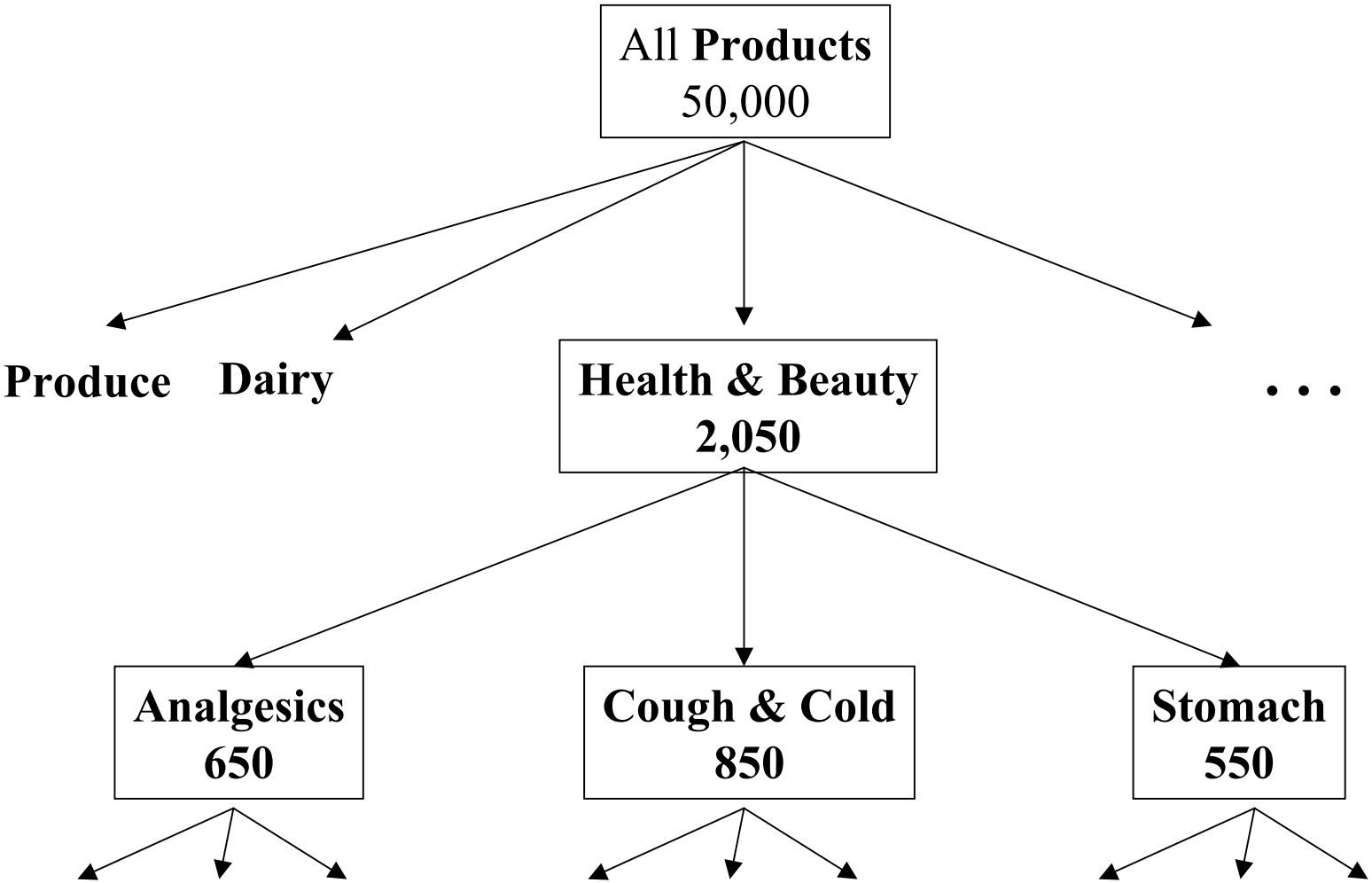
Conceptual Confidentiality Kernel



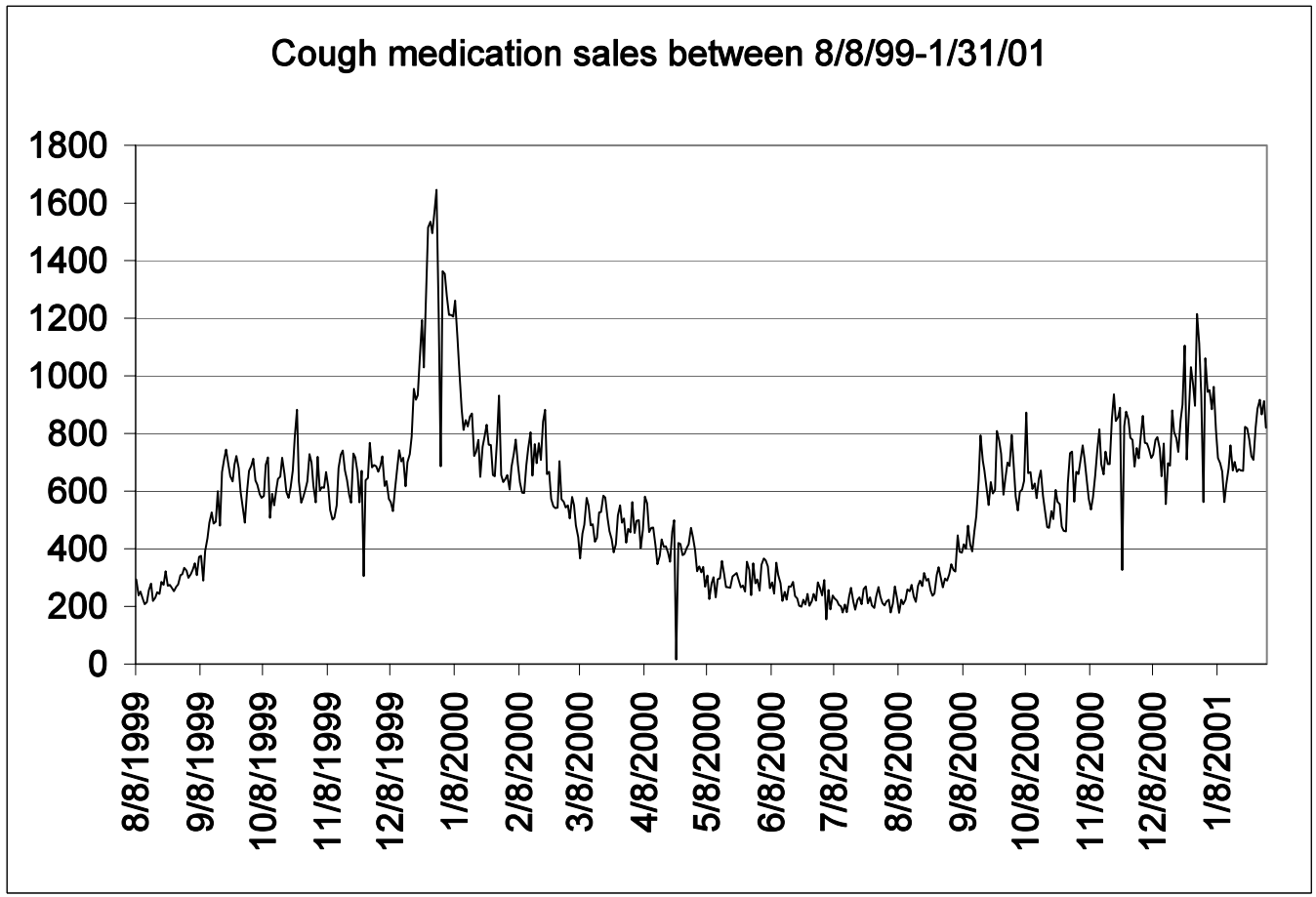
Time and Space

- **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

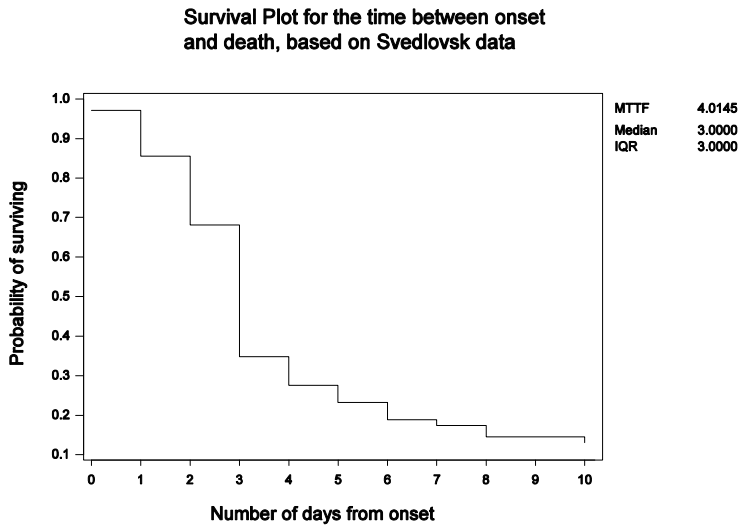


Confounding Natural Periodicities



Signal Detection

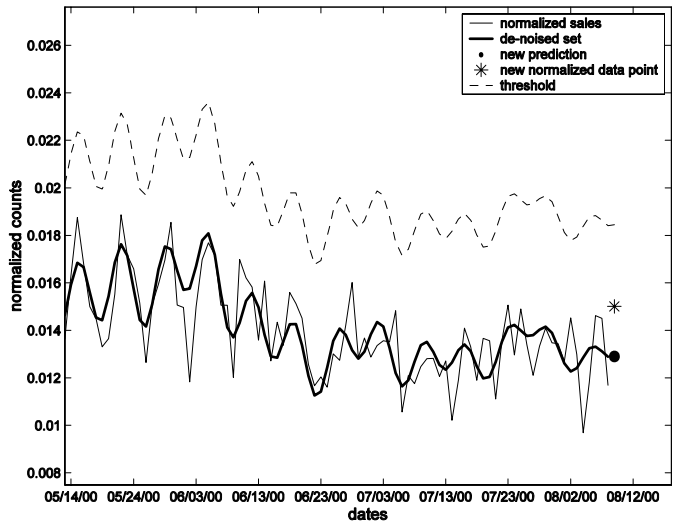
- **Adverse events ➔ Discovery of cause**
  - **e.g., detecting signature of outbreak in response to anthrax attack**



  - **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**