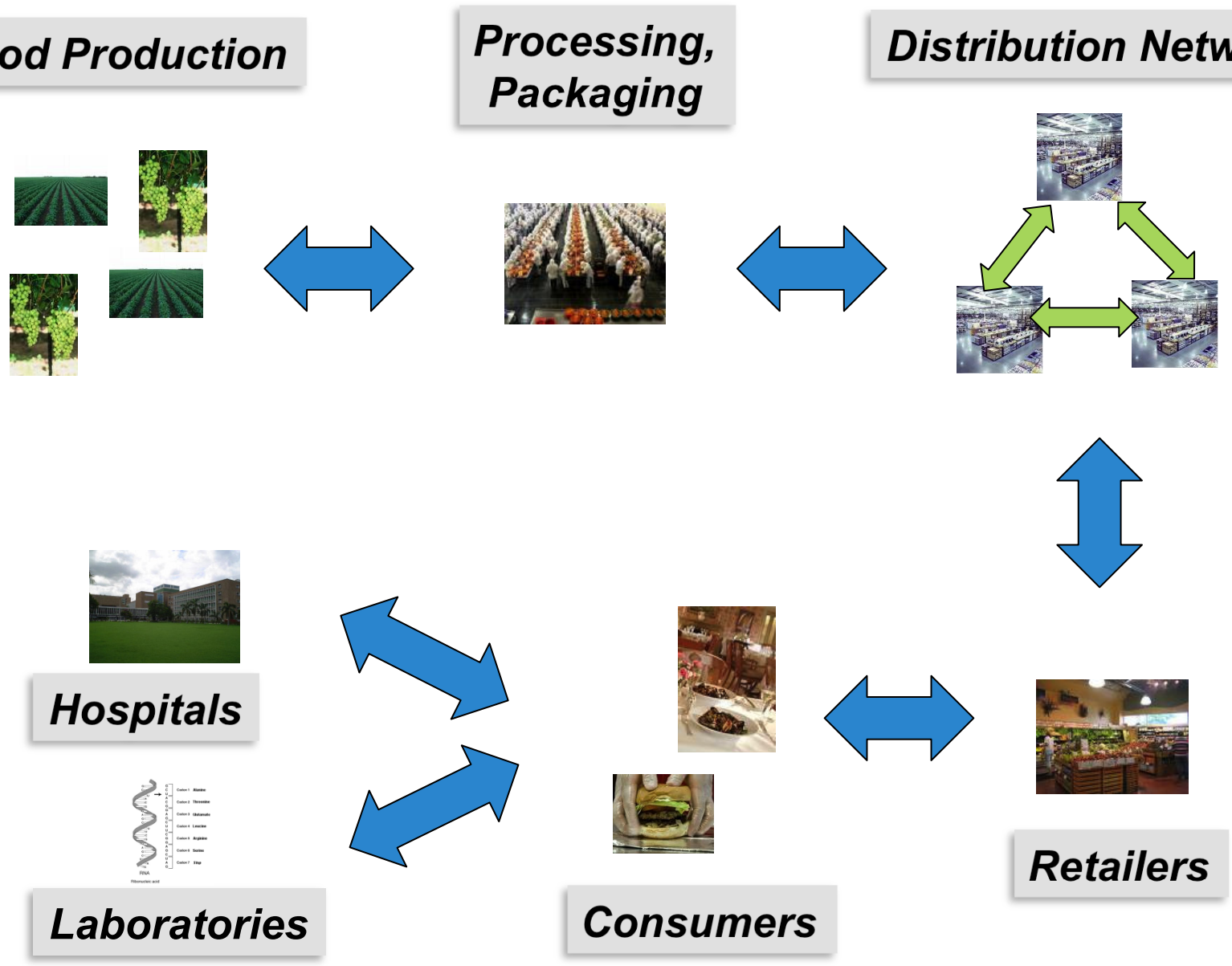
Computer-based Tracking Protocols: Improving Communication between Databases

Amol Deshpande
Database Group
Department of Computer Science
University of Maryland

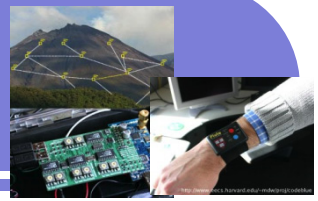
Overview

- Food tracking and traceability crucial in today's world
  - Public health
  - Responding to outbreaks
  - Accountability
  - Operational efficiency
- Tremendous amounts of useful information out there – and growing daily
  - A small fraction in use today
  - Typically ad hoc, labor-intensive to use it
- How do we exploit this information ?
- What are the challenges that we will face ?
  - Many form active research areas in databases today

Data Overload



Data Overload

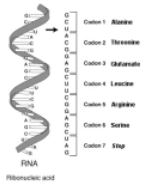


Wireless sensor networks

- Unprecedented and rapidly increasing instrumentation of our every-day world
  - Data from suppliers, warehouses, supermarkets
  - Patient monitoring data (EHR)
  - Sensor networks that monitor: container temperatures during transit, patients' vitals signs, water/air contaminants
  - RFID devices that can track at the pallet level
  - More efficient and faster genome sequencers
  - Social networks
- Potentially very useful data
  - Proactive alert and anomaly monitoring
  - More informed decision-making; better response to outbreaks
  - Faster information dissemination to interested parties

Distributed measurement networks (e.g. GPS)

RFID



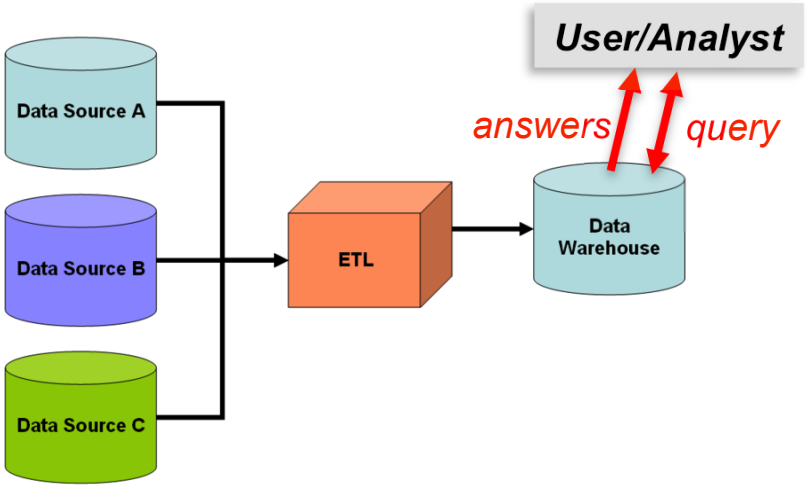
Genome data

Challenges

- Highly distributed data sources
  - Continuously generating data at high volumes
- Entities that don’t talk to each other
  - For legal or technical reasons
- Privacy
- Noisy, erroneous, often incomplete data
- Data provenance (“tracking”)
- Often don’t know what to do with the data
  - Need to support data analytics, data mining
- …

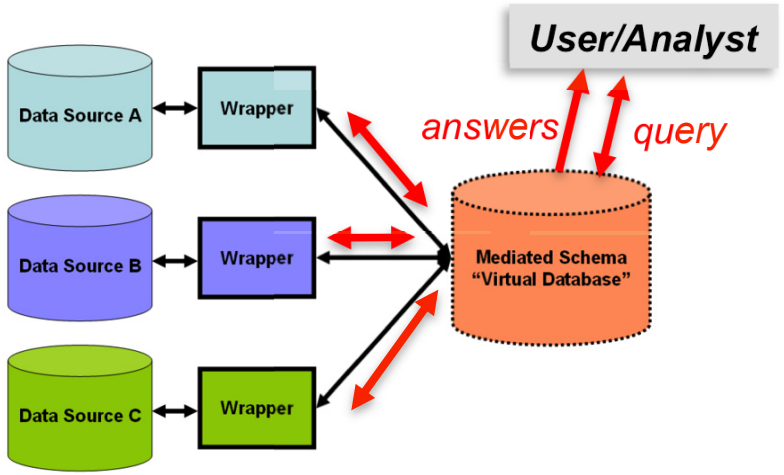
Data Integration

- Combining information from distributed, heterogeneous data sources
  - Different administrative domains
  - Typically different "schemas"



Using a Data Warehouse

+ Efficient
+ Low-latency query processing
+ Very well-developed technology
- Stale data
- Expensive; not always feasible



Using a Mediated Schema

+ "Real-time" data
+ Cheaper hardware costs
- Conceptually harder to set up
- High latencies

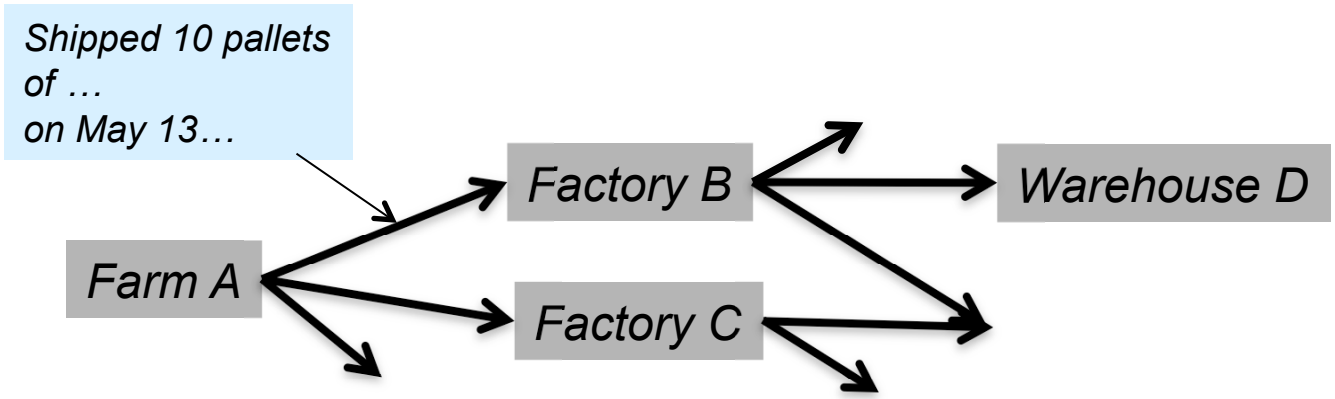
Data Integration: Challenges

- Semantic heterogeneity
  - How to design translation routines or wrappers ?
  - Must figure out the correspondences between different data sources
    - Commonly called “schema mapping” or “schema matching”
  - Standardization (e.g. using XML) ideal, but often not feasible
- Setting up the data integration
  - E.g. constructing the queries to convert data back and forth
- Typically semi-automatic process
  - Find possible correspondences and mappings
  - Refine using user feedback
- Active research area in the database community

Semantic Integration Research in the Database Community: A Brief Survey;
A. Doan and A. Halevy; AI Magazine, Special Issue on Semantic Integration, Spring 2005.

Data Analysis and Processing

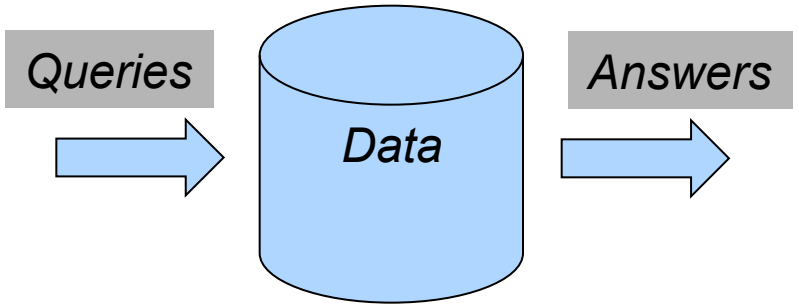
- What data model and query language ?
  - Relational databases ?
    - Tabular representation may not naturally capture the environment
    - Is SQL the right query language ?
      - SQL doesn't handle temporal data very well
  - No real alternative for large datasets
    - Recent development of "cloud computing" may be a solution – enables renting compute cycles as needed



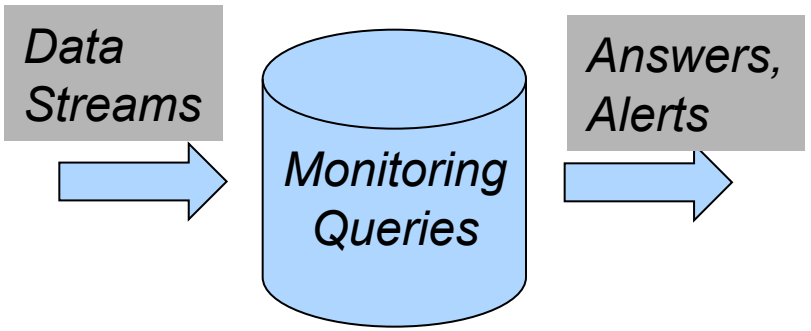
Graph representation of the supply chain data

Data Analysis and Processing

- Must be able to do real-time monitoring over the data
  - By using “continuous queries” over “data streams”
    - Over both relational data and XML data
    - Need to be able to specify: “only consider data generated in last day”
  - Complex event processing
    - Tell me if “A happens followed by B within 1 hour”
  - Anomaly detection
- Active research area in recent years
  - Several commercial systems (Streambase, Truviso, IBM System S…)



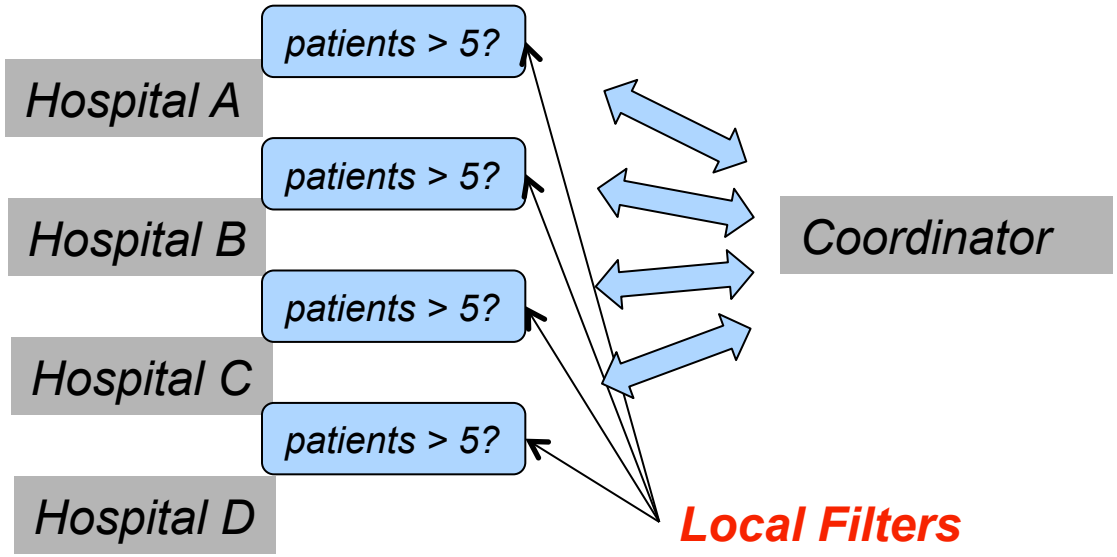
Traditional Model



Streaming Model

Data Analysis and Processing

- Data is naturally distributed
  - Bringing it together at one location not feasible
  - Data must be processed where it is
- Need distributed aggregation and threshold monitoring
  - *Alert if "number of patients with flue is more than 100 in the entire state"*

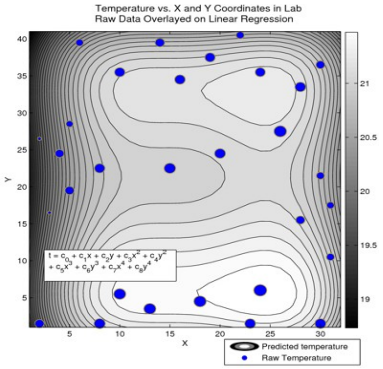


M. Franklin et al.; Design Considerations for High Fan-in Systems: The HiFi Approach; CIDR 2005
M. Garofalakis; http://www.softnet.tuc.gr/~minos/Papers/eds09dstreams.pdf; 2009

Statistical Modeling

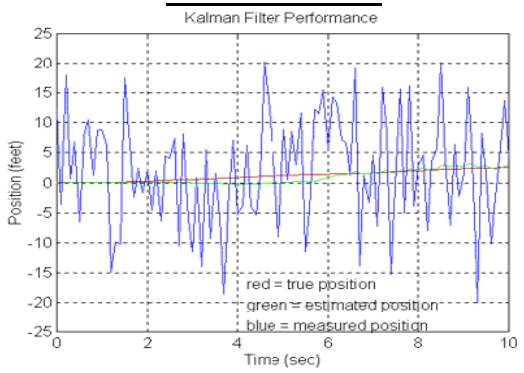
- Need for real-time *statistical modeling* of data
  - Regression, interpolation, classification, anomaly detection, forecasting and prediction models often used
  - Recently Bayesian approaches have become popular
- Very little work on integrating these within a data management framework
  - MATLAB not sufficiently easy to use or scalable in a distributed setting

Temperature monitoring



Regression/interpolation models

GPS Data



Kalman Filters …

A. Deshpande, S. Madden; MauveDB: Supporting Model-based User Views in Database Systems; SIGMOD 2006

B. Kanagal, A. Deshpande; Online Filtering, Smoothing and Probabilistic Modeling of Streaming data; ICDE 2008

Data Uncertainty

- Real-world data exhibits many uncertainties
  - Measurement noise, errors (sensor, RFID data)
  - Trust, reputation or quality issues (data integration)
  - Probabilities (classification, inference)
  - Imprecise queries (not sure what we are looking for)
  - Erroneous, but faster, detection devices

- Current database systems typically ignore them
  - Can lead to inaccuracies
  - May prevent early detection because of low confidences
    - *If many analysts individually suspect an outbreak with low confidence, the overall likelihood should be high*

Data Uncertainty

- Instead, represent the uncertainty explicitly in the data and reason about it during processing
  - Each event associated with a “probability of being true”
  - An attribute associated with a set of values with a probability distribution, instead of just one value
- Use probability theory principles to operate upon them
  - May choose to use something different (e.g. for reputation)
- Active research area in databases
  - With several projects here at UMD

Amol Deshpande et al.; Graphical Models for Uncertain Data; 2009.
Nilesh N. Dalvi, Dan Suciu: Management of probabilistic data: foundations and challenges; PODS 2007.

Other Challenges

- Privacy
  - Many entities don't want to share data
    - For good reasons
  - In many cases, we only need aggregated data
    - Hide the sensitive information somehow
  - Many mechanisms developed in recent years
    - *k-diversity, l-anonymity;* Data perturbations
  - But still hard to guarantee data privacy
- Data provenance
  - Similar notion to "traceability"
  - How to do this over distributed data streams, complex aggregation and fusion…

Other Challenges

- Data acquisition
  - Can we automate the data acquisition process ?
    - Which food to test, and what tests ?
    - Who to interview ?
  - The cost of acquisition is high
  - Can use statistical prediction models to decide the most important information to acquire

Amol Deshpande et al.; Model-driven Data Acquisition; VLDB 2004
Jure Leskovec et al.; Cost-effective Outbreak Detection in Networks; SIGKDD 2007.

Thanks !

- Questions ?