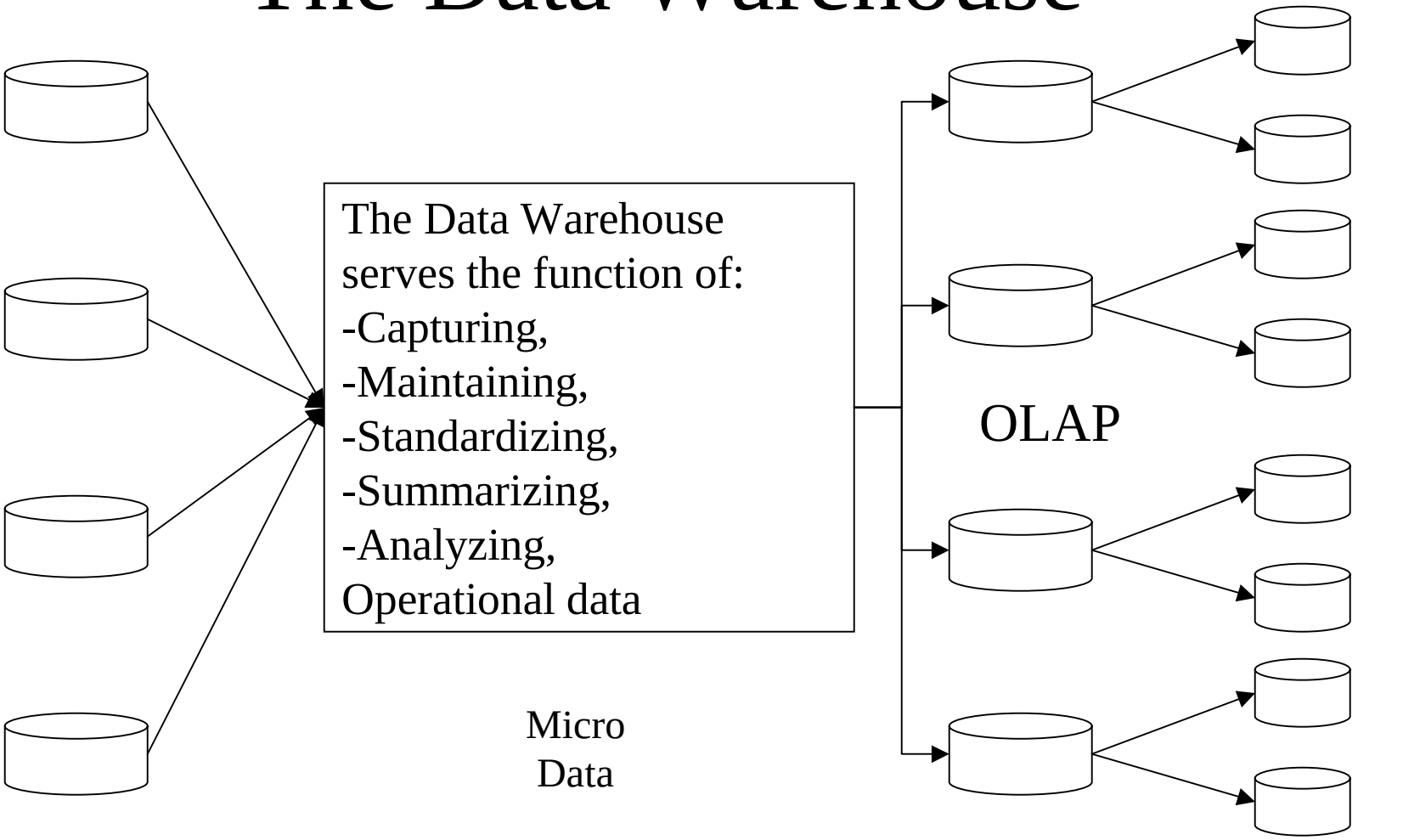
Discussion of “Longitudinal Analysis Using Data Warehouse Techniques”

By Marton Vucsan and Jan Kardaun,
Statistics Netherlands

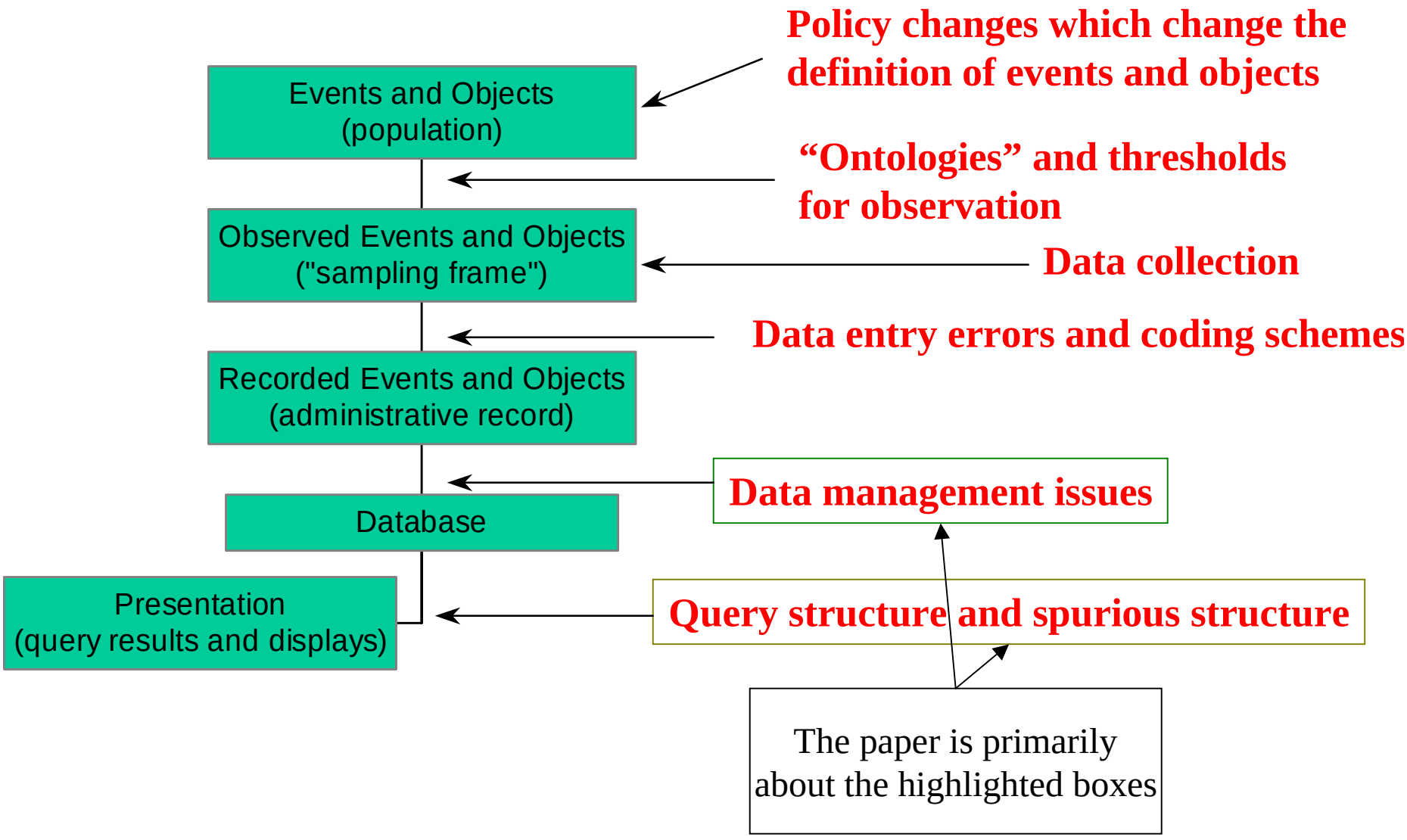
The Data Warehouse

- William Inmon, about 1990
- Basic problem:
  - Operational data are changing all the time;
  - Different analyses (at different times, using different queries) give different results;
  - History cannot be reconstructed.
- Solution:
  - Maintain “operational data” in a repository
  - Perform analyses on the repository (warehouse)
  - Standardize data
  - Perform historical analyses at will

The Data Warehouse



How Administrative Records Are Created and Used



Challenges for the Data Warehouse

- Data Quality and Database Ontologies
  - A delivery address suitable for receiving a payment check may not suffice for putting individuals at a street address
  - Data coding differs across different databases
    - 2 sexes (M, F) vs. multiple (XX,XY,XXY)
    - 4 races x 2 Hispanic origin vs. 5 races (Hispanic treated as a race)
  - Transaction data ≠ person data
  - How many names does a person have (and in what order)
- Example: Proxy Addresses

  JOHN WILSON
  C/O MARY SMITH
  1004 LAUREL LANE
  ROCKMONT, MD 22345

  The address is (presumably) for Mary Smith. John Wilson may or may not live there.
- Example: Naming Conventions

  Dean Harold Judson
  vs. Jane Marie Barker-Jones
  Vs. Irene Marie Zimmerman y Diaz
  Vs. Myoung Kim (Kim Myoung)

  "Judson" vs. "Barker-Jones" vs. "Zimmerman y Diaz" vs. "Myoung"

Challenges for the Data Warehouse, continued

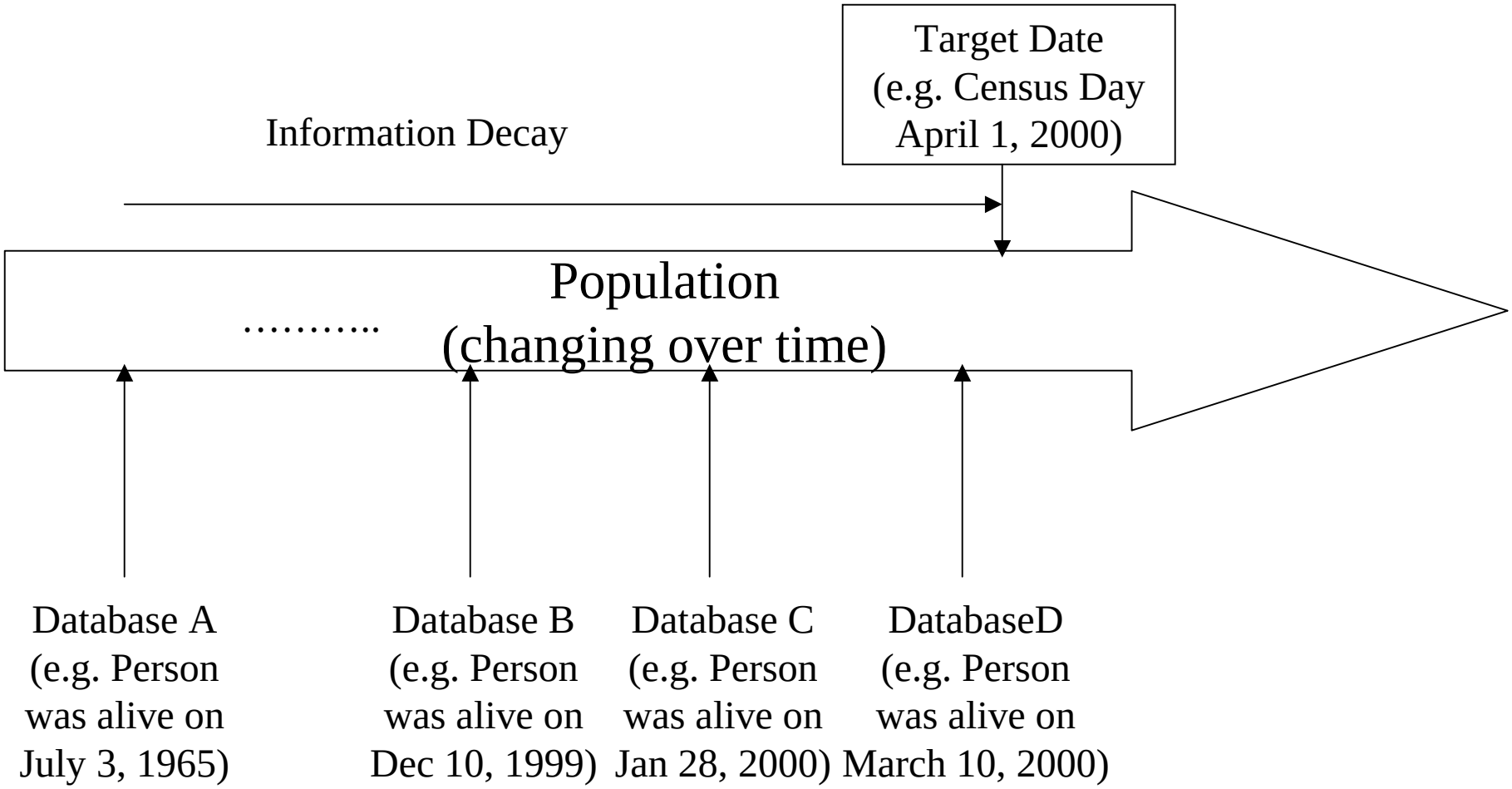
- Changing information states
  - Distinct problem from "point in time" data collection
  - Information states change over time/over databases
    - Address information ages over time and varies over databases

SAM SMITH
BOX 2 RURAL ROUTE 37
WESTPORT, VA 32784
(Dated 10/14/98 from Medicare)
(Not Geocodable)

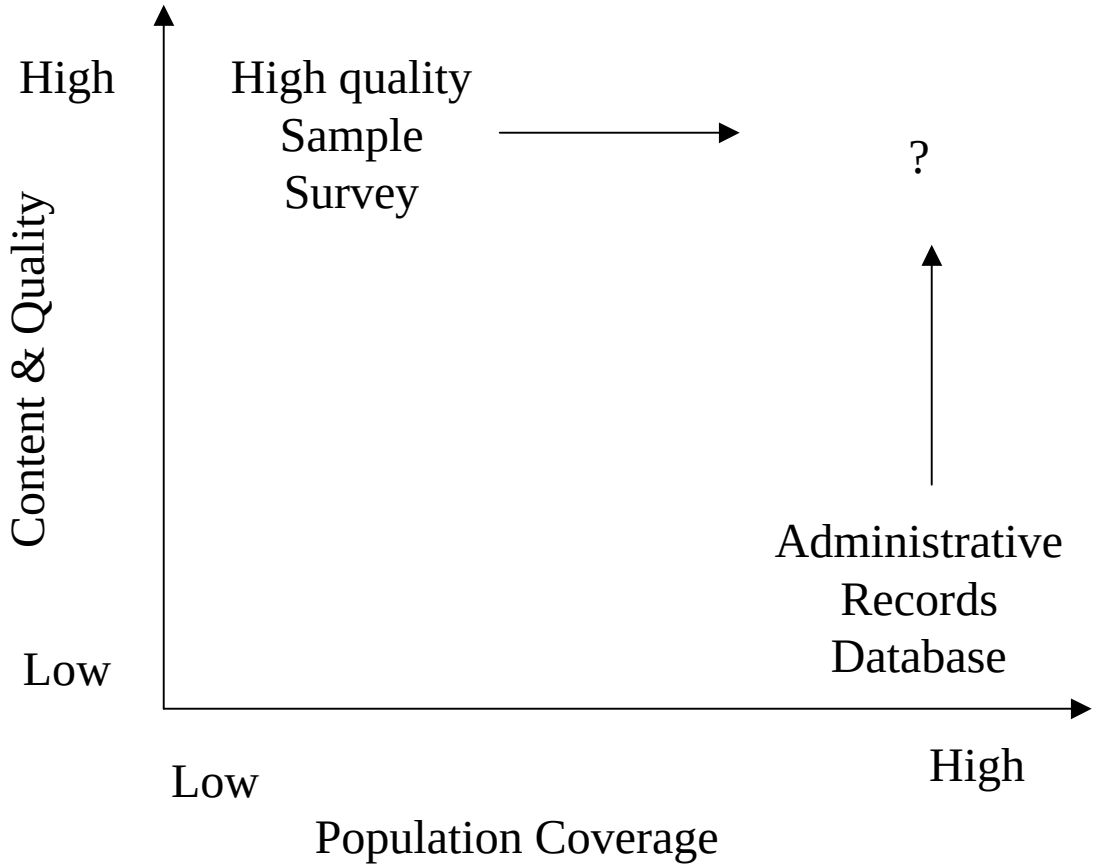
SAM SMITH
486 MAIN STREET
FAIRFIELD, VA 33412
(From Tax Year 97 IRS file, filed sometime in 1998)
(Geocodable)

  - Mortality information ages over time and varies over databases
  - One database provides information about the other, provided that matching can be performed
  - Data processing requires complex, and substantively important, decision logic at each step

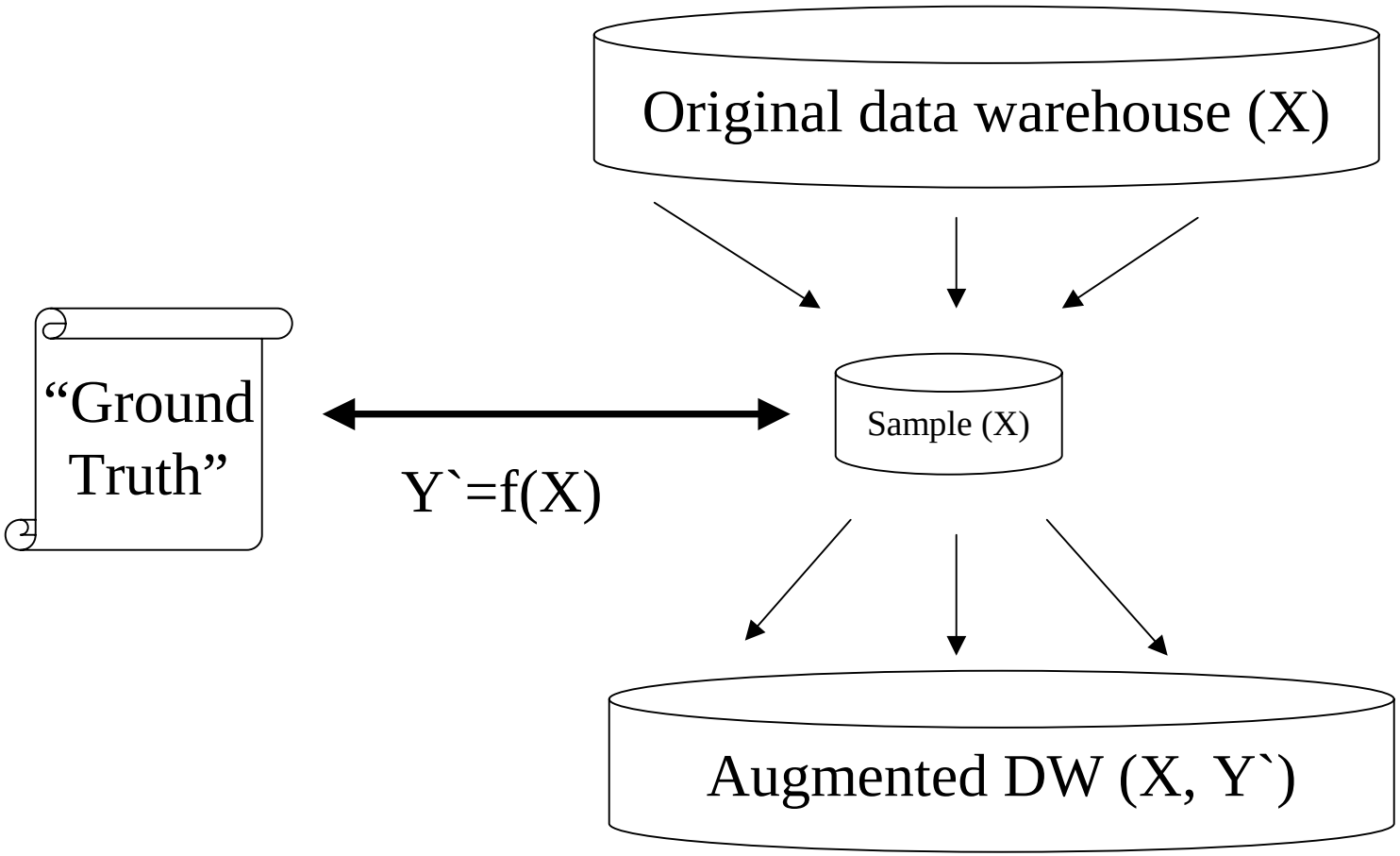
Thinking Dynamically about Information Decay



Trade-off: Coverage vs. Quality



A model for “augmenting” the data warehouse (DW)



Specific Thoughts

- The Good News:
  - Data Warehouse is the right approach
  - OLAP is very useful for “ad hoc” queries
  - Relational database is right framework
  - Recognizing the time dimension of data
  - More complex analyses
- The Bad News:
  - Combining data with different ontologies is very difficult
  - Changing population and time reference is a challenge
  - Instantaneous Access?
  - We need to augment our data warehouse

Specific Questions

- Does a relational database (as opposed to a “flat file”) really have no “theoretical impact”?
- Can you do a logistic regression with current OLAP tools?
- Pre-defined queries and subsets to create analysis databases?
- How important is instantaneous access, really?
- What about the fact that the population is constantly changing?
- What about data quality concerns?