# Data Curation Means …

# by

# Michael Stonebraker

DBg Database Group
*MIT Computer Science and Artificial Intelligence Lab*

# For K Data Sources:

- Identify the data sources!
- Ingest the data
- Clean the data
- Transform the data
- Perform schema integration
- Perform entity consolidation

# Simple Example --- 2 Data Sources

- Employee (name, salary, hobbies, age, city, state)
- Person (p-id, wages, address, birthday, year_born, likes)

# And 2 Records

- Sam Madden, $4000, {bike, dogs}, 36, Cambridge, Mass.

- Samuel E. Madden, $5000, Newton Ma., October 4, 1975, bicycling

# Data Curation (1)

- Ingest
  - Read the 2 records and store in a common place
- Clean
  - $4000 and $5000: both wrong? One right? Both right? (May have to ask an expert)
- Tranform
  - October 4, 1975 → 39
  - Now clean 39 and 36

# Data Curation (2)

- Schema Integration
  - hobbies same as likes?
  - Person same as Employee?
- Entity consolidation
  - 2 Sams or 1 Sam?

# Data Curation (3)

- Making use of “trusted” data sources, if available
  - Dictionary of hobbies, …

# Papers in this Session

- Consider various aspects of data curation
  - Web Tables
  - “finding” issue
  - Sandbox for experimentation
  - In 18 minutes or less
  - Leaving 15 minutes for discussion

# Advertisement

- This problem is killing most enterprises!!!!
  - Customer integration for cross selling
  - Purchasing integration to get “most favored nation” terms
  - Medical data records
  - …