Data Integration Overview

Zachary G. Ives
University of Pennsylvania

January 13, 2003

CIS 650 – Data Sharing and the Web

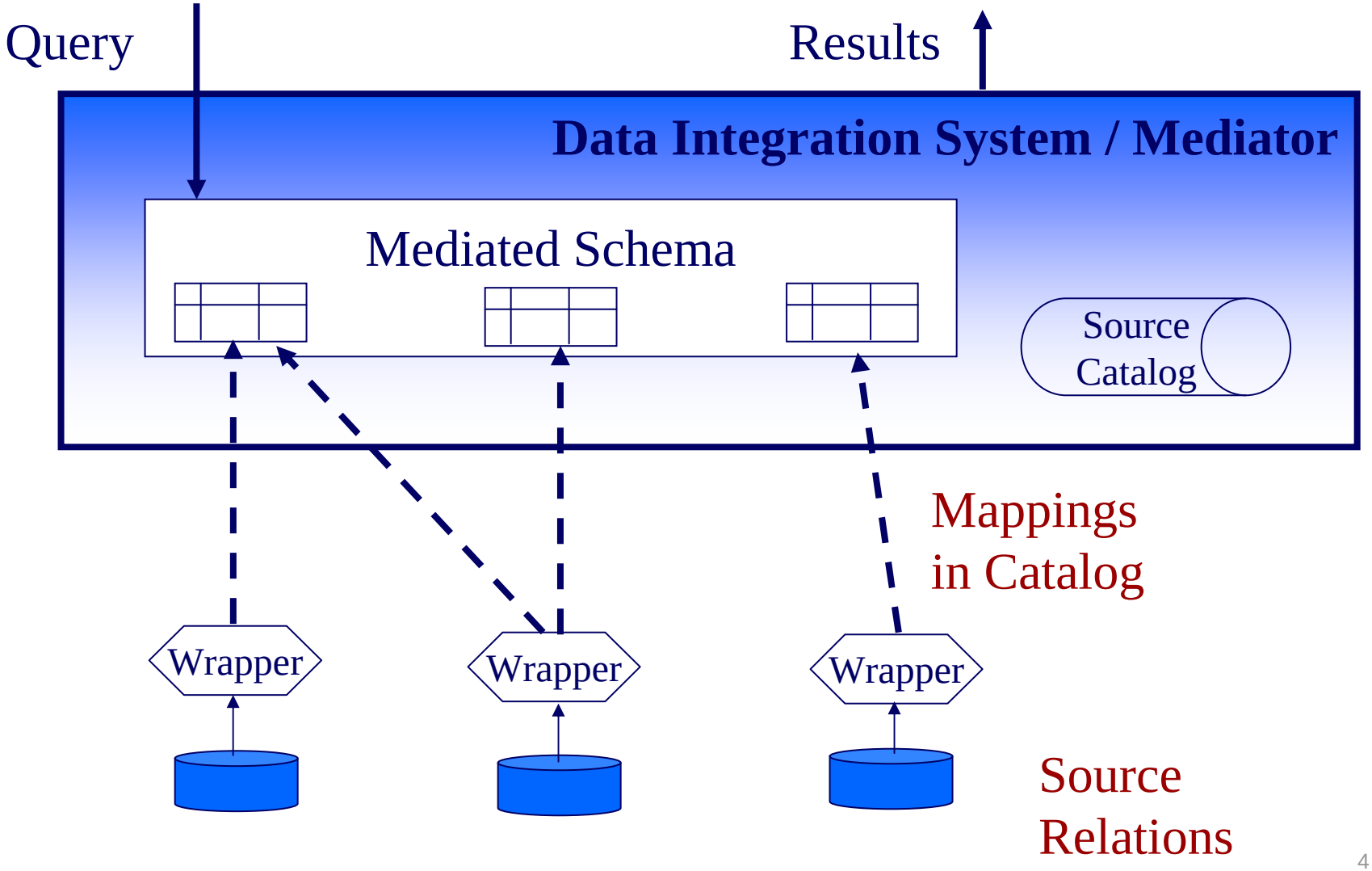
Data Integration / “Mediators”

- Problem: Preponderance of different data sources with overlapping data
  - Different systems within an enterprise
  - Different information brokers on the Web
- Solution: Tie existing data sources with related data into single “mediated” system
  - Benefits: single queryable view (e.g., unified catalog for Amazon shoppers; CRM; etc.)

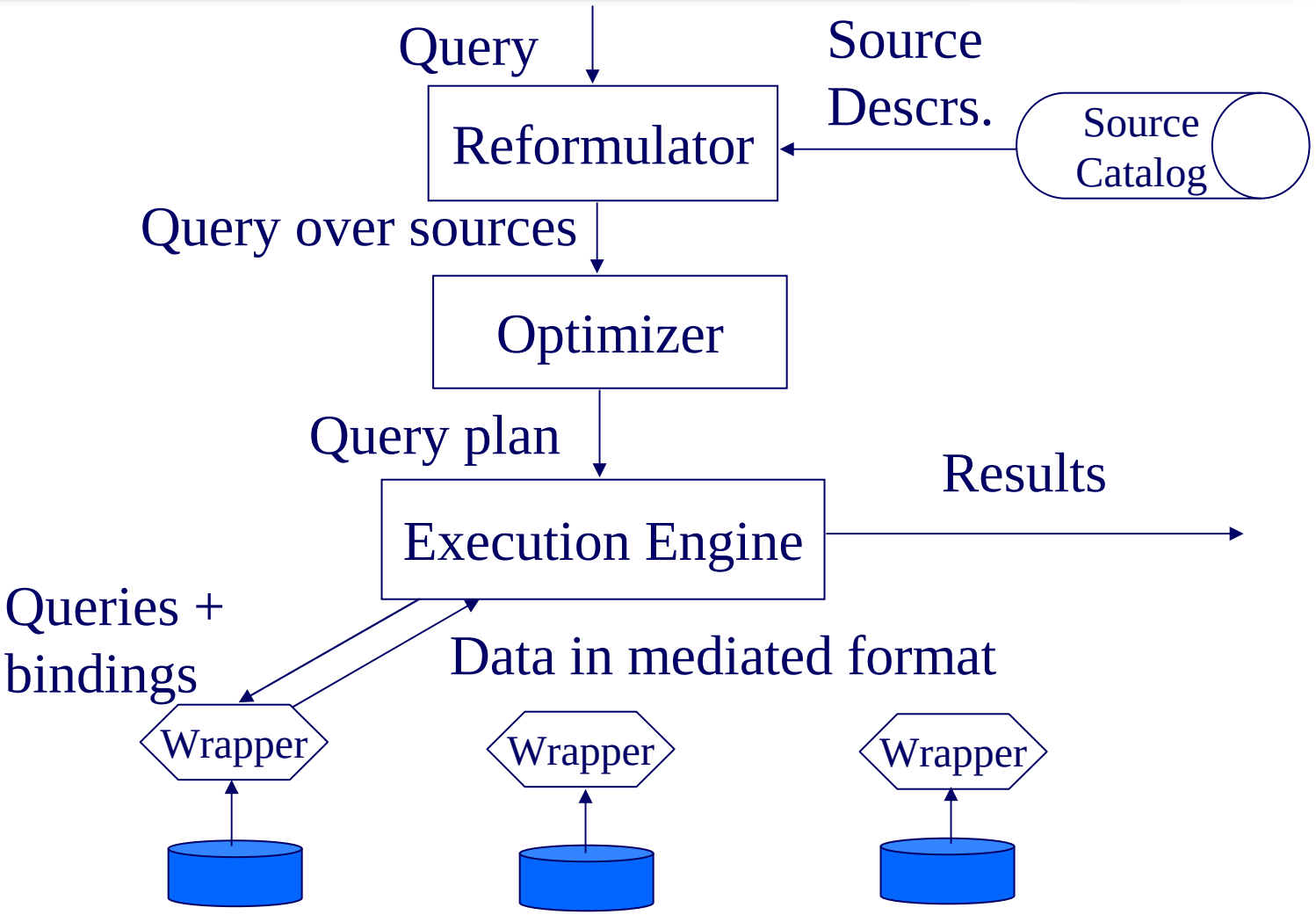
Building a Data Integration System

- Create a middleware “mediator” or “data integration system” over the sources
  - Can be warehoused (a data warehouse) or *virtual*
  - Presents a uniform query interface and schema
  - Abstracts away multitude of sources; consults them for relevant data
    - Unifies different source data formats (and possibly schemas)
    - Sources are generally *autonomous*, not designed to be integrated
  - Sources may be local DBs or remote web sources/services
  - Sources may require certain input to return output (e.g., web forms): binding patterns

Typical Data Integration Components



Typical Data Integration Architecture



Some Important Design Points

- Garlic [Haas+97] – IBM Almaden (now in DB2)
  - Focus: intranet, SQL, few well-profiled source types
  - No mediated schema
- TSIMMIS [Garcia-Molina+97] – Stanford
  - Focus: semistructured data (OEM), OQL-based language (Lorel)
  - Mediated schema defined in terms of sources
- Information Manifold [Levy+96] – AT&T Research
  - Focus: local-as-view mappings, relational model
  - Sources defined in terms of mediated schema