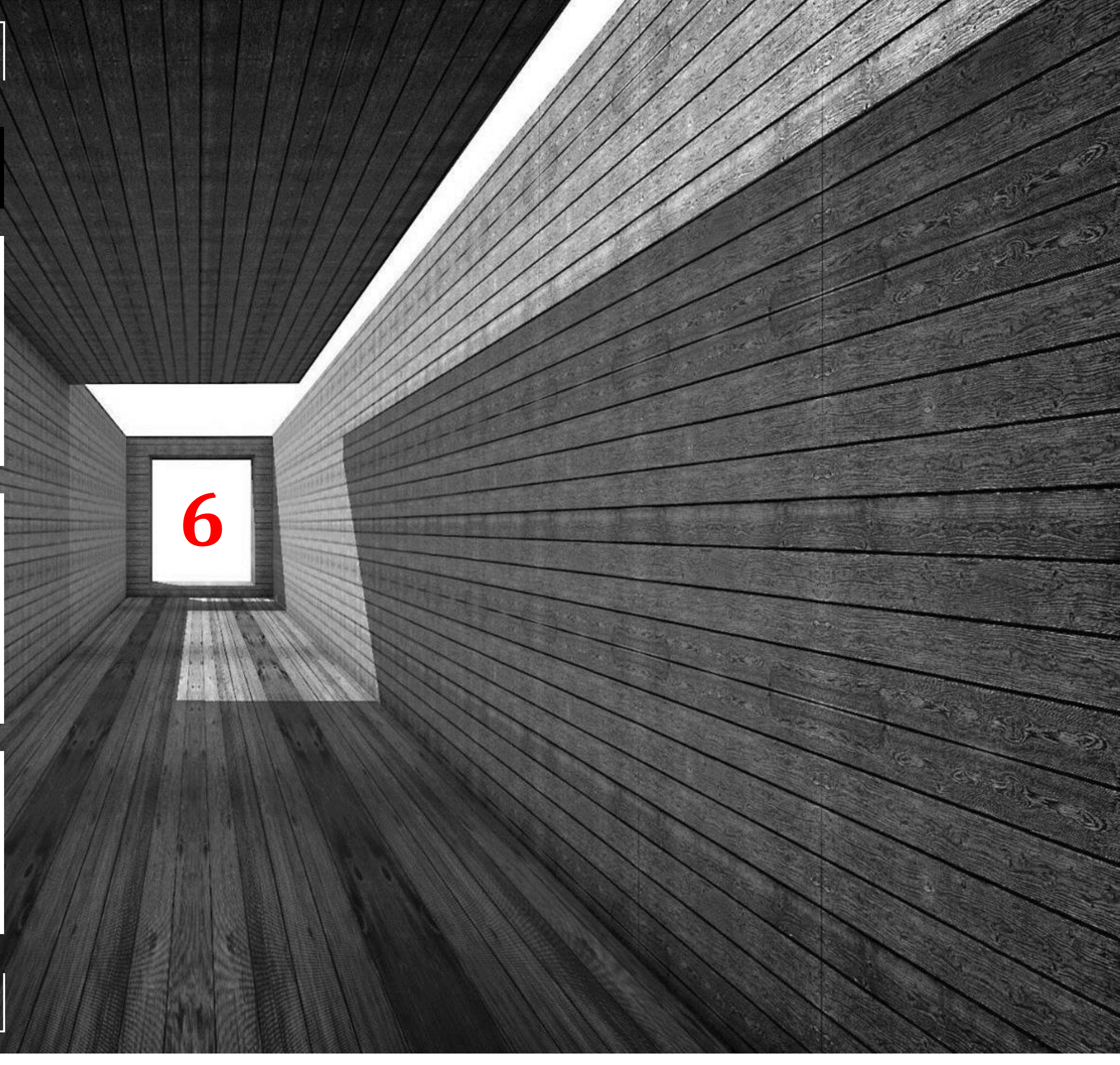
The German University in Cairo

CSEN1095 Data Engineering

Lecture 6
Data Integration



Mervat Abuelkheir
mervat.abuelkheir@guc.edu.eg

Data Integration

Data Integration

- **Merging** data from **multiple** data stores/sources
  - Can be *local*, within same organization perimeters (e.g. across departments)
  - Can be due to mergers/acquisitions of different organizations
  - Can be due to need to use external data sources (e.g. sensors, social feeds)
- Techniques help reduce and avoid *inconsistencies* and *redundancies* in the resulting consolidated dataset
- *Challenges:*
  - Semantic **heterogeneity** → different representations of data, different data scales
  - **Entity identification** problem → join/match keys
  - **Redundancy** → records (numerocity) or attributes (dimensionality)
  - **Structure** of data → functional dependencies and referential constraints

Data Integration – Structure of Data Sources

- **Data formats**
  - Proprietary formats are troublesome
  - Sensor data formats need vendor-specific interpretation
  - XML and JSON are accepted as universal formats, but not all systems abide in production
- **Data modalities**
  - Image data, audio data, video data
  - Medical data
  - No explicit attributes - feature extraction is complex
- **Functional dependencies**
  - Consolidating business rules defined over different database schemas
  - Constraint prioritization is problematic

Data Integration – **Heterogeneity Levels**

- Schema
  - **Schema** mismatch
    - e.g. single student table in one DB, multiple student tables (for different academic years) in another DB
  - **Domain** mismatch
    - e.g. single name attribute in one DB versus first name and last name attributes in another DB
  - **Constraint** mismatch
    - e.g. GPA constraints for student enrollment
- Instance
  - **Entity identification**
    - e.g. same patient in two different hospital databases, with no clear identification value
  - **Format conflict**
    - e.g. DOB for same customer is recorded differently in two databases

Data Integration – **Semantic Heterogeneity**

- **Different definitions**
  - Different views of same entity. Need to agree on meaning or mapping
  - e.g. sales amount means money or # units sold
- **Different representations or encodings**
  - Need to unify
  - e.g. name stored as first-last in one attribute versus name stored as last-first in two attributes
- **Different scales**
  - Need to convert or unify
  - e.g. profits measured per month and profits measured per day, grades maintained differently across educational systems
- **Different timeframes/granularities**
  - Need to timestamp, synchronize, and align
  - e.g. network traffic data and network performance data

Data Integration – Tuple Redundancy and Entity Identification

- Two records within the same DB table representing the **same entity** – **duplicate records**
- Duplicate records can usually be matched using a **name** or **ID** attribute that should be unique
  - But unifying attribute may actually not be identical!

Methods

- Schema integration and exact joins over explicit keys
- Metadata → *attribute* name, meaning (semantics), data type, range of values permitted, null rules for handling blank, zero, or null values
  - helps avoid errors in schema integration and data transformation
  - BUT – you don't have control over how and how much metadata are documented if you are not the data collector

Data Integration – **Tuple Redundancy and Entity Identification**

Methods (Cont.)

- If no explicit keys exist to perform exact join, use approximate joins over messy attributes
  - Use most probable attribute (with most unique values in both sources) for join (e.g. Name)
  - May have to use corroborating matches over other attributes! (e.g. Name and Phone #, Address)
  - No standardized representation → needs a lot of manipulation!
- If approximate joins are not possible, maybe infer joins?
  - e.g. do two attributes from two different data sources look like they represent a user's phone number? Use them to join!
  - We can use Regular Expressions for that
  - Or use Approximate Matching (String Matching) – e.g. Levenshtein distance
- Easier method to resolve tuple redundancy is to perform Feature Vector Matching
  - Compute distance (or similarity) between two records incorporating all (or most discriminating) object attributes

Entity Identification Problem – Example

- Two transaction records from two stores:

  Ted Johnson, 3 apples, 09-01-2001

  Theodore Johnson, 2 CDs, 09-02-2001

- No explicit ID to join both records representing the same entity
- We can use approximate matching or similarity algorithms
- But then we will also match Ed Johnson, Eddy Johnson, Todd Johnson to the same entity, which they're not!

Entity Identification – Approximate String Matching

- Measure how far apart two strings are
  - How many edit operations (substitute, insert, delete) are required to change one string into another
- **Levenshtein distance** between two strings a, b of length $|a|$ and $|b|$ respectively is given by
- $$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & if \min(i,j) = 0 \\ min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & otherwise \end{cases}$$
  - $lev_{a,b}(i,j)$ is distance between the first i characters of a and the first j characters of b
  - $1_{(a_i \neq b_j)}$ equals 0 when $a_i = b_j$ and equals 1 otherwise
  - $lev_{a,b}$ is equal to zero if and only if the strings are equal
  - $lev_{a,b}$ is at most the length of the longer string

Entity Identification – Approximate String Matching

Step	How
1	Set $\|a\|$ to be the length of a and set $\|b\|$ to be the length of b. If $\|a\| = 0$, return $\|b\|$ and exit. If $\|b\| = 0$, return $\|a\|$ and exit. Construct a matrix containing $0 \dots \|a\|$ rows and $0 \dots \|b\|$ columns.
2	Initialize the first row to $0 \dots \|a\|$. Initialize the first column to $0 \dots \|b\|$.
3	For each character of a (i from 1 to $\|a\|$). For each character of b (j from 1 to $\|b\|$). If $a[i] = b[j]$, the cost is 0. If $a[i] \neq b[j]$, the cost is 1.
4	Set cell $d[i, j]$ of the matrix to be equal to the minimum of: a. The **cell immediately above** plus 1: $d[i-1, j] + 1$. b. The **cell immediately to the left** plus 1: $d[i, j-1] + 1$. c. The **cell diagonally above and to the left** plus the **cost**: $d[i-1, j-1] + cost$.
5	After the iteration steps (3, 4) are complete, the distance is found in cell $d[\|a\|, \|b\|]$.

Approximate String Matching Example

		G	U	M	B	O
	0	0	1	2	3	4
G	1					
A	2					
M	3					
B	4					
O	5					
L	6					

		G	U	M	B	O
	0	1	2	3	4	5
G	1	0	1	2	3	4
A	2	1	1	2	3	4
M	3	2	2	1	2	3
B	4	3	3	2	1	2
O	5	4	4	3	2	1
L	6	5	5	4	3	**2**

2 edits needed for GUMBO to become GAMBON

Approximate String Matching Example Details

		G	U	M	B	O
	0	1	2	3	4	5
G	1					
A	2					
M	3					
B	4					
O	5					
L	6					

		G	U	M	B	O
	0	1	2	3	4	5
G	1	0	1	1	1	1
A	2					
M	3					
B	4					
O	5					
L	6					

		G	U	M	B	O
	0	1	2	3	4	5
G	1	0	1	2	3	4
A	2					
M	3					
B	4					
O	5					
L	6					

		G	U	M	B	O
	0	1	2	3	4	5
G	1	0	1	2	3	4
A	2	1	1	1	1	1
M	3					
B	4					
O	5					
L	6					

		G	U	M	B	O
	0	1	2	3	4	5
G	1	0	1	2	3	4
A	2	1	1	2	3	4
M	3					
B	4					
O	5					
L	6					

Approximate String Matching Example Details

		G	U	M	B	O
	0	1	2	3	4	5
G	1	0	1	2	3	4
A	2	1	1	2	3	4
M	3	1	1	0	1	1
B	4					
O	5					
L	6					

		G	U	M	B	O
	0	1	2	3	4	5
G	1	0	1	2	3	4
A	2	1	1	2	3	4
M	3	2	2	1	2	3
B	4					
O	5					
L	6					

		G	U	M	B	O
	0	1	2	3	4	5
G	1	0	1	2	3	4
A	2	1	1	2	3	4
M	3	2	2	1	2	3
B	4	1	1	1	0	1
O	5					
L	6					

		G	U	M	B	O
	0	1	2	3	4	5
G	1	0	1	2	3	4
A	2	1	1	2	3	4
M	3	2	2	1	2	3
B	4	3	3	2	1	2
O	5					
L	6					

		G	U	M	B	O
	0	1	2	3	4	5
G	1	0	1	2	3	4
A	2	1	1	2	3	4
M	3	2	2	1	2	3
B	4	3	3	2	1	2
O	5	1	1	1	1	0
L	6					

		G	U	M	B	O
	0	1	2	3	4	5
G	1	0	1	2	3	4
A	2	1	1	2	3	4
M	3	2	2	1	2	3
B	4	3	3	2	1	2
O	5	4	4	3	2	1
L	6					

Data Integration – **Attribute Redundancy** and **Correlation Analysis** (Again)

- *An attribute is redundant* if it can be “derived” from another attribute(s)
- Attribute redundancy is related to **Multicollinearity**
  - Multicollinearity negatively affects some ML algorithms (can exaggerate performance, can mess up parameter estimation)
- Redundancy can be detected by correlation analysis → measure how strongly one attribute *implies* the other, based on the available data
  - *Nominal attributes* → chi-square (χ^2) test
  - *Numeric attributes* → correlation coefficient and covariance

Redundancy and Correlation Analysis

chi-square (χ^2) test for nominal attributes

- *Example*: Are *gender* and *preferred reading* correlated in a dataset with the following observations?

ID	Name	Gender	Preferred reading	Last visit	Last book bought
1	Adam	M	Fiction	9/7/2021	Game of Thrones
2	Ali	M	Non-fiction	12/5/2020	Sophie's World
3	Sarah	F	Fiction	13/5/2020	Grapes of Wrath
…	…	…	…	…	…
…	…	…	…	…	…

Redundancy and Correlation Analysis

chi-square (χ^2) test for nominal attributes

- *Example*: Are *gender* and *preferred reading* correlated in a dataset with the following observations?
  - *Contingency table* → summary of **observed values**

		gender		
		male	female	Total
Preferred reading	Fiction	250	200	450
	Non-fiction	50	1000	1050
	Total	300	1200	1500

Redundancy and Correlation Analysis

Hypothesis: the two attributes are <u>independent</u> (not correlated) – **Null hypothesis**

- ***expected (independent) frequencies*** → $e_{ij} = \frac{count\ (A=a_i) \times count\ (B=b_j)}{n}$
- e.g. $e_{11} = \frac{count\ (male) \times count\ (fiction)}{n} = \frac{300 \times 450}{1500} = 90$

		gender		
		male	female	Total
Preferred reading	Fiction	250 (90)	200 (360)	450
	Non-fiction	50 (210)	1000 (840)	1050
	Total	300	1200	1500

Redundancy and Correlation Analysis

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}} = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

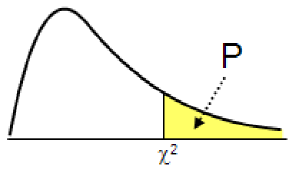
- o_{ij} → observed frequency
- e_{ij} → expected frequency

		gender		
		male	female	Total
Preferred reading	Fiction	250 (90)	200 (360)	450
	Non-fiction	50 (210)	1000 (840)	1050
	Total	300	1200	1500

Redundancy and Correlation Analysis

- For **one degree of freedom at *[p-value](#)* = 0.001 significance level**, the χ^2 value needed to reject the hypothesis is 10.828
  - (source: http://www.medcalc.org/manual/chi-square-table.php)
  - *Degrees of freedom*:
    - If $r > 1$ and $c > 1$, then $df = (r - 1)(c - 1)$
    - If $r = 1$ and $c > 1$, then $df = c - 1$ or if $r > 1$ and $c = 1$, then $df = r - 1$
    - $r = c = 1$ is not allowed

Chi-square distribution table



P

DF	0.995	0.975	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.690	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.180	11.030	13.362	15.507	17.535	18.168	20.090	21.955	24.352	26.124
9	1.735	2.700	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.920	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.300	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697
16	5.142	6.908	20.465	23.542	26.296	28.845	29.633	32.000	34.267	37.146	39.252
17	5.697	7.564	21.615	24.769	27.587	30.191	30.995	33.409	35.718	38.648	40.790
18	6.265	8.231	22.760	25.989	28.869	31.526	32.346	34.805	37.156	40.136	42.312
19	6.844	8.907	23.900	27.204	30.144	32.852	33.687	36.191	38.582	41.610	43.820
20	7.434	9.591	25.038	28.412	31.410	34.170	35.020	37.566	39.997	43.072	45.315
21	8.034	10.283	26.171	29.615	32.671	35.479	36.343	38.932	41.401	44.522	46.797

➢ $507.93 \gg 10.828 \rightarrow$ reject hypothesis that preferred reading and gender are independent!

∴ *Gender* and *preferred reading* are strongly correlated

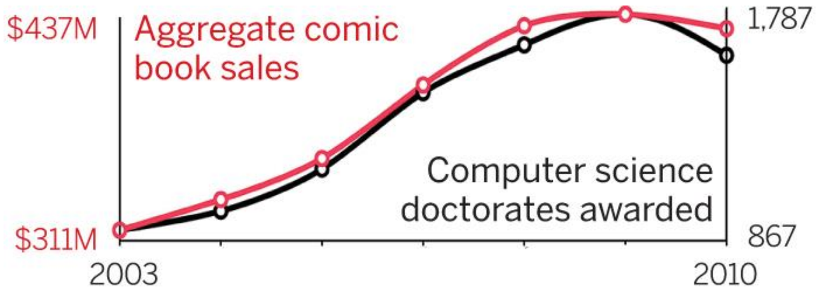
For more on p-values, refer to: https://www.students4bestevidence.net/blog/2016/03/21/p-value-in-plain-english-2/

Redundancy and Correlation Analysis

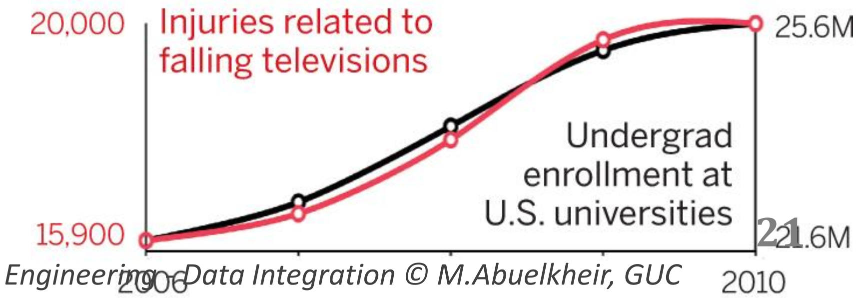
Correlation coefficient for numeric attributes

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - (n\bar{A}\bar{B})}{n\sigma_A\sigma_B}$$

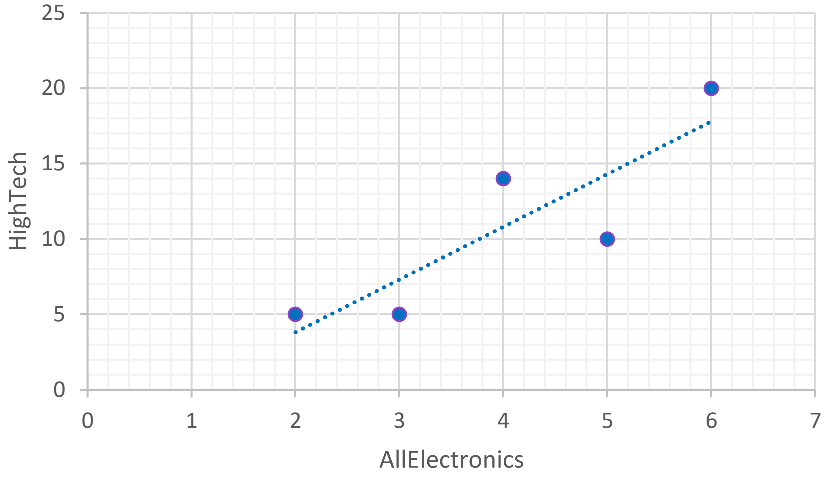
- -1 ≤ $r_{A,B}$ ≤ +1
- If $r_{A,B}$ is *greater* than 0, then *A* and *B* are *positively* correlated
  - The higher the value, the stronger the correlation
- If $r_{A,B}$ = 0, then *A* and *B* are *independent*



- **Correlation does not imply causality!**



Redundancy and Correlation Analysis Example



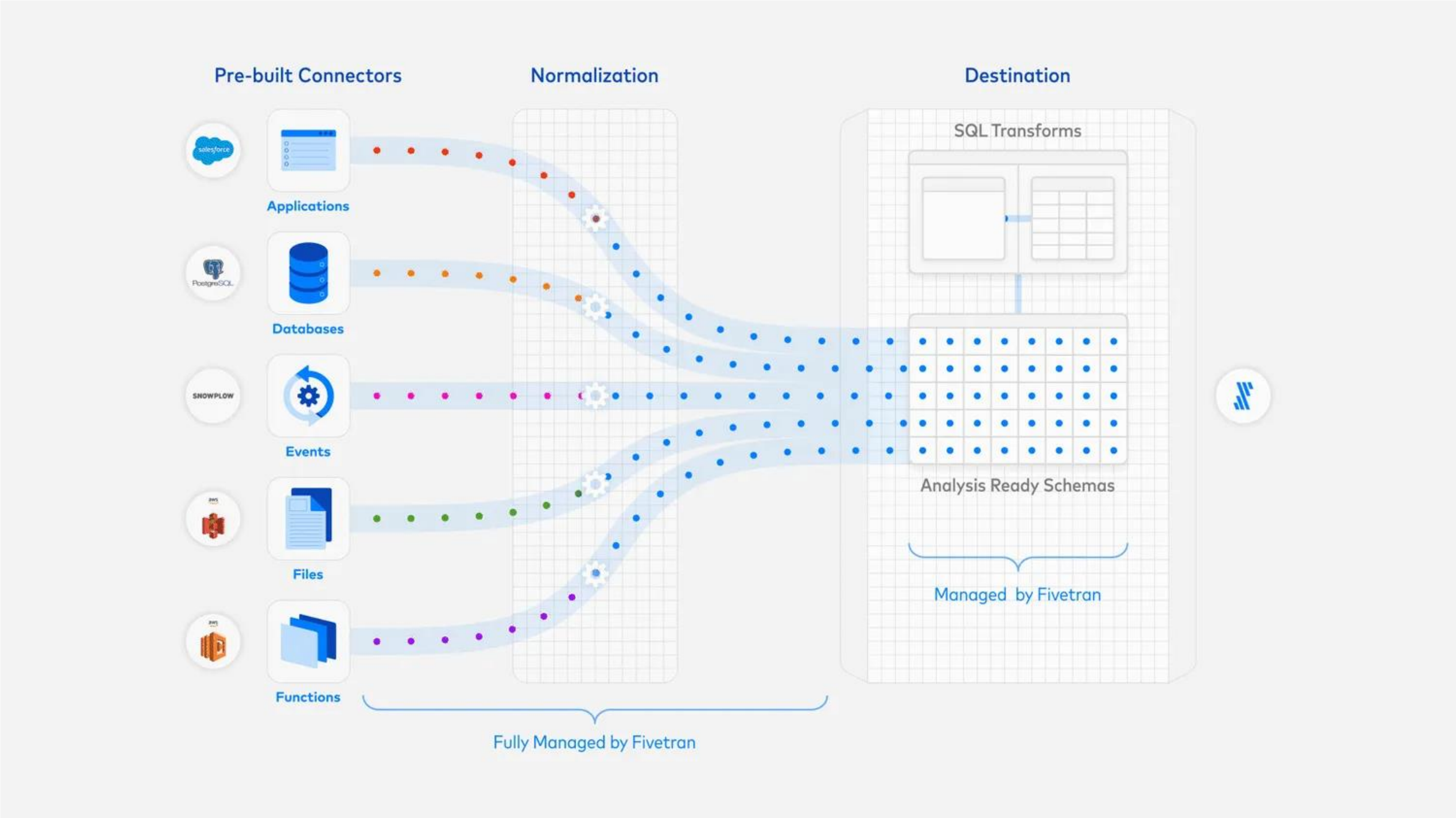
- Stock prices for two companies
  - $\bar{A}$(AllElectronics) = 20/5 = \$4
  - $\bar{B}$(HighTech) = 54/5 = \$10.80

$$r_{A,B} = \frac{(6\times20+5\times10+4\times14+3\times5+2\times5)-(5\times4\times10.80)}{5\times1.4\times5.7} = \frac{251-216}{39.9} \approx 0.88$$

Time point	AllElectronics	HighTech
T1	6	20
T2	5	10
T3	4	14
T4	3	5
T5	2	5

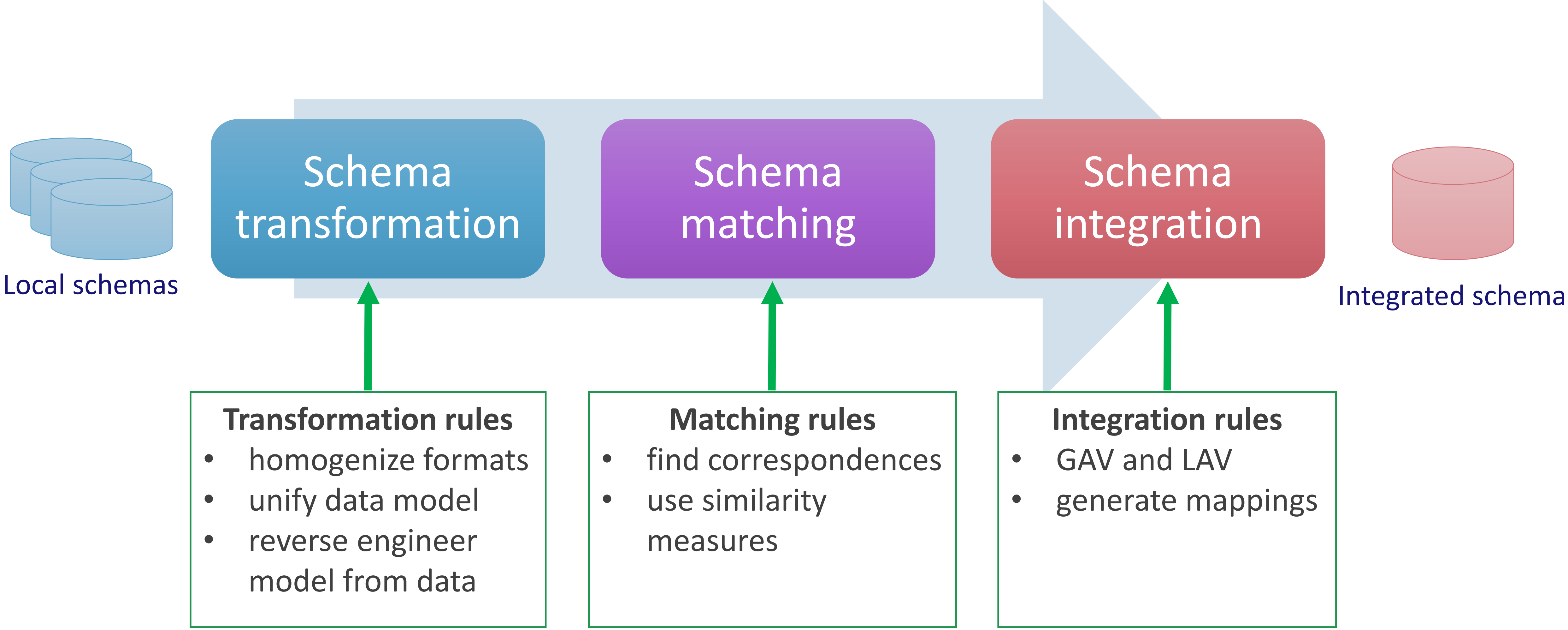
  - +ve correlation→ stock prices of the two companies rise together

Data Integration – **Schema Integration**



https://fivetran.com/

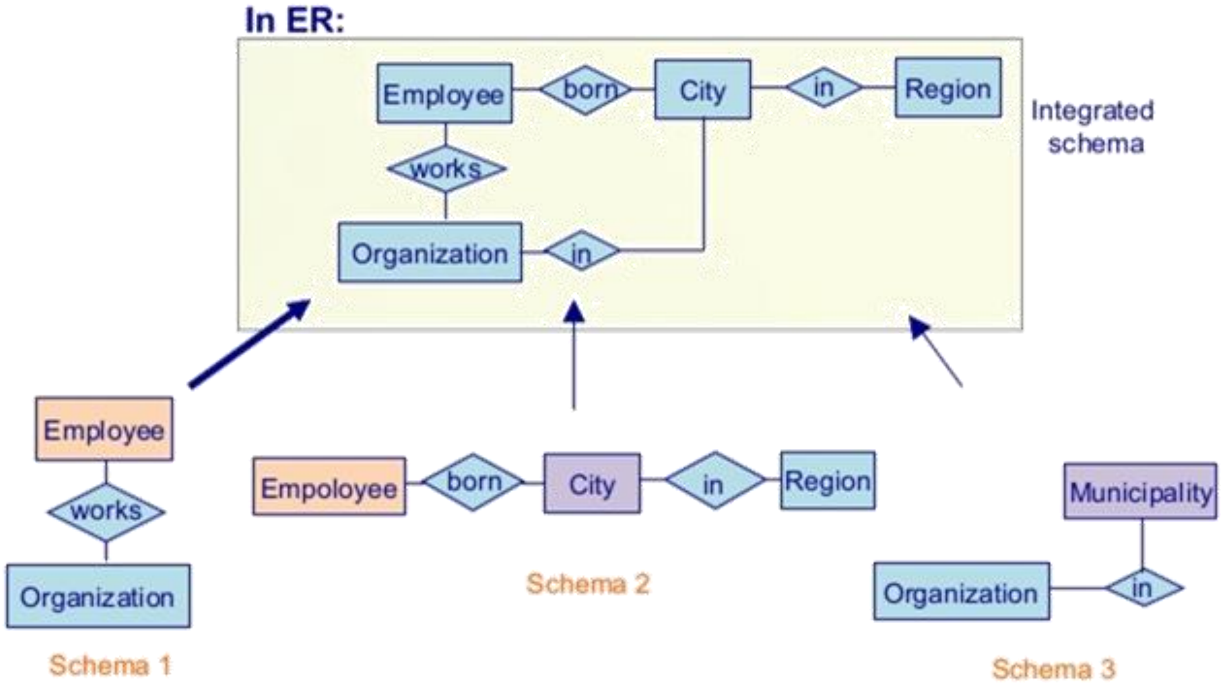
Data Integration – **Schema Integration**



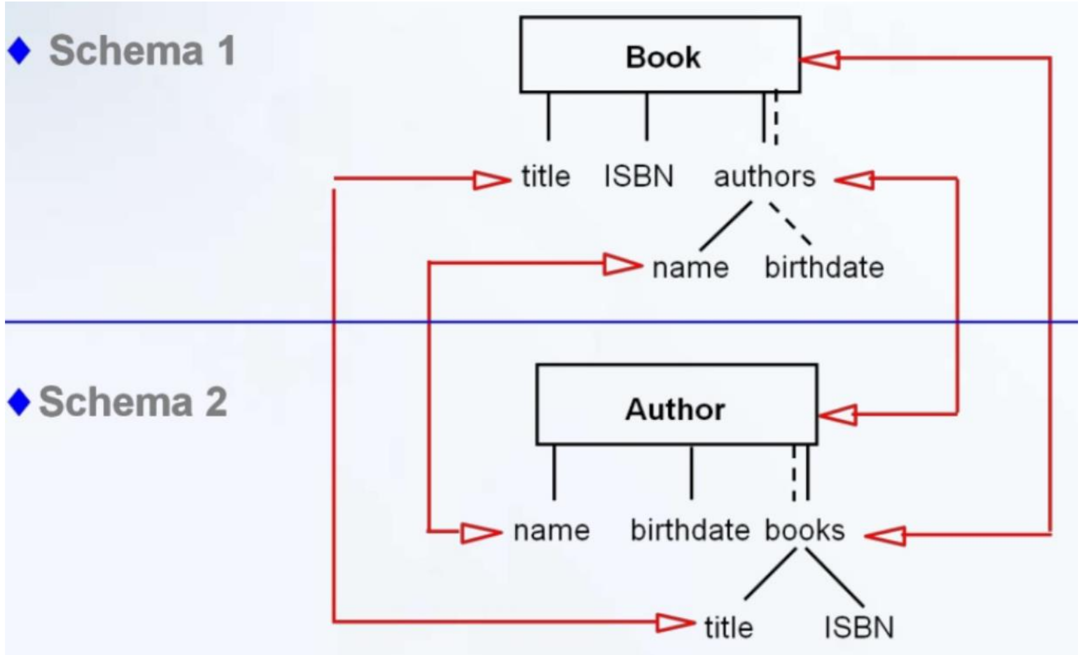
Data Integration – **Schema Integration**

Two main challenges:

- Identify and unify schema elements that relate to the same concept/phenomena
- Identify and resolve conflicts across schemas



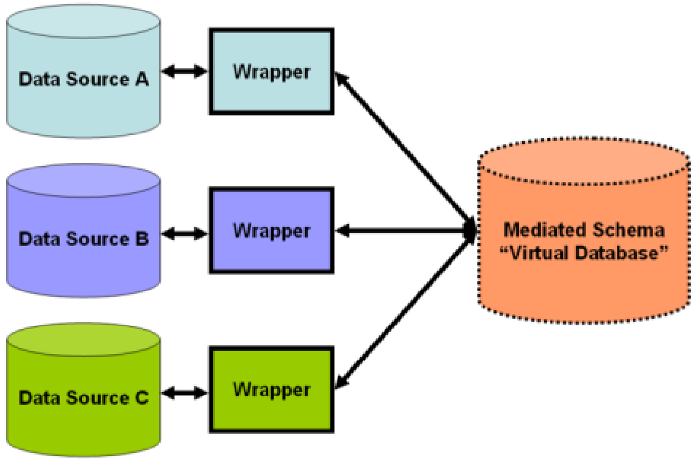
Correspondences relate schema elements that describe same phenomena



Source: Jarrar: Data Schema Integration: https://www.slideshare.net/jarrar02/jarrar-data-schema-integratiov2

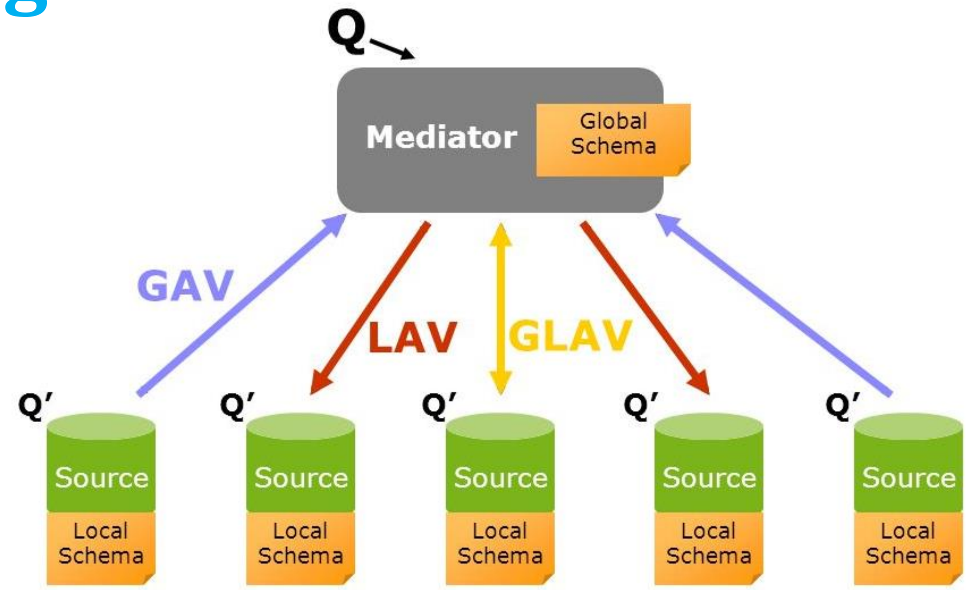
Data Integration – Mediated/View Schema Integration

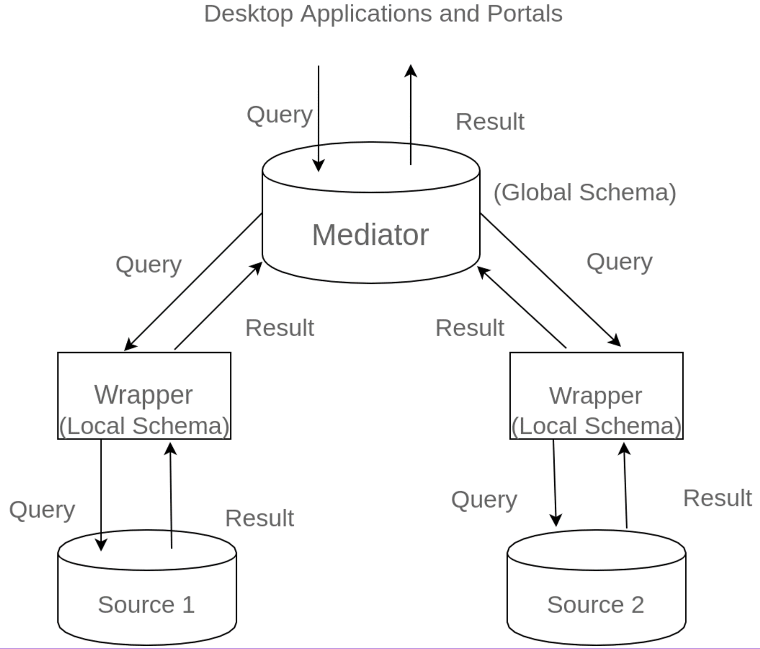
- Service-Oriented Architecture (SOA) Approach
- Provide a unified query-interface to access real time data
  - Allow information to be retrieved directly from original databases
- Mappings between the mediated schema and the schema of original sources
  - Mapping from entities in the mediated schema to entities in the original sources – Global-as-View (GAV) approach
  - Mapping from entities in the original sources to the mediated schema – Local-as-View (LAV) approach
- Translating a query into decomposed queries to match the schema of the original databases



Data Integration – Mediated Schema Integration

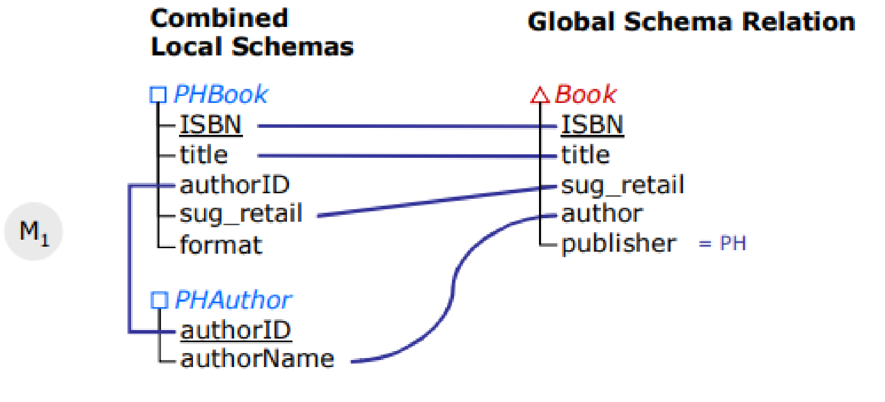
- Global as View
  - Define a **global schema** that acts as a view over existing source schemas
    - Global schema is a function of the local schemas
  - Data is only stored at the sources
  - Given a query over the global schema, mediator will follow the existing rules and templates to convert query into source-specific queries
  - **Wrappers** execute source-specific query on their local schema
  - Results from local sources are merged back together to form final result
  - Addition of new sources is a challenge because schema must be redefined



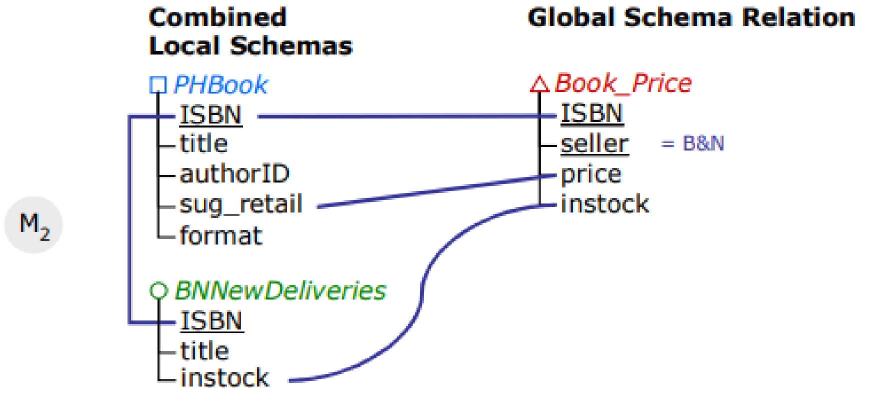


Data Integration – Mediated Schema Integration

- Global as View
  - Define a **global schema** that acts as a view over existing source schemas
    - Global schema is a function of the local schemas
  - Data is only stored at the sources
  - Given a query over the global schema, mediator will follow the existing rules and templates to convert query into source-specific queries
  - **Wrappers** execute source-specific query on their local schema
  - Results from local sources are merged back together to form final result
  - Addition of new sources is a challenge
  - No new information can be modeled if not present in local schemas



(a) Mapping for global relation Book

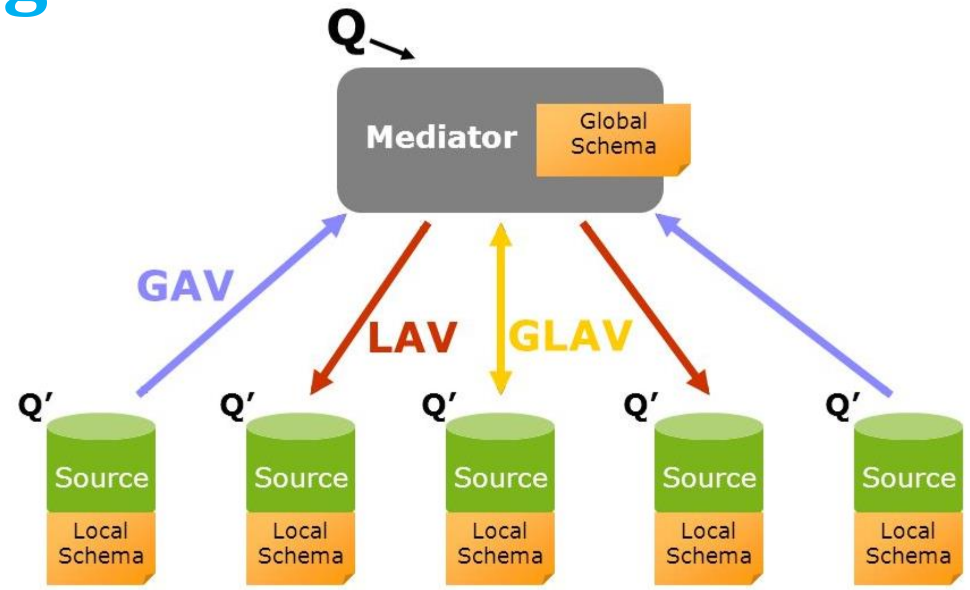


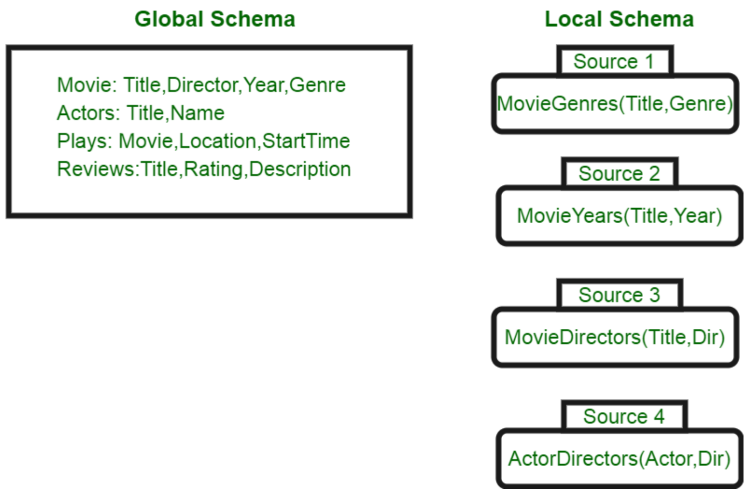
(b) Mapping for global relation Book_Price

```
V1(ISBN, title, sug_retail, authorName, “PH”)
V2(ISBN, “B&N”, sug_retail, instock)
```

Data Integration – Mediated Schema Integration

- Local as View
  - Each local schema is described as a **view** over a global schema (a complete vision of what is needed)
  - View – which data in schema is present in source?
  - Data is still stored at sources
  - Global schema is not altered as new sources join/leave – only mappings change
  - Addition of new sources is flexible

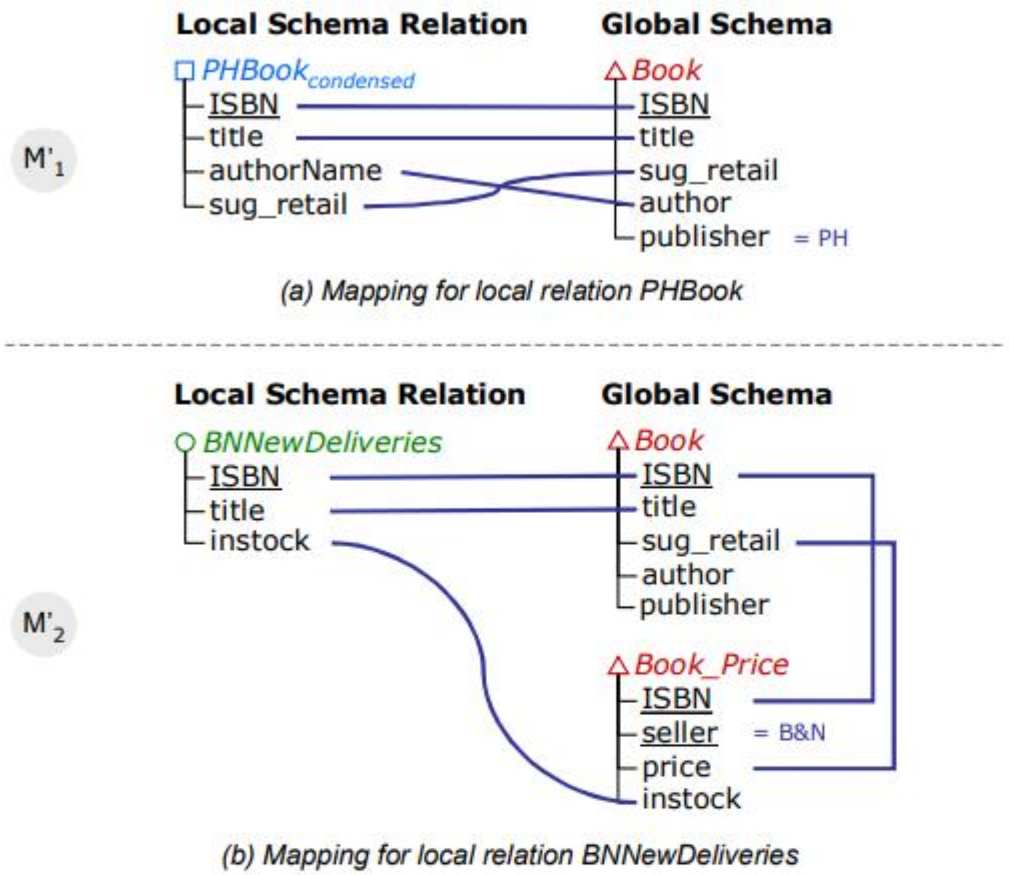




- Think of the example local schemas on the right – what would they look like if GAV was used?

Data Integration – Mediated Schema Integration

- Local as View
  - Each local schema is described as a **view** over a global schema (a complete vision of what is needed)
  - View – which data in schema is present in source?
  - Data is still stored at sources
  - Global schema is not altered as new sources join/leave – only mappings change
  - Addition of new sources is flexible
  - Information in sources not easily handled in global schema
  - No unique global database is possible because of the suggested mapping (virtual mediation)



https://dbucsd.github.io/paperpdfs/2009_7.pdf

Thank You