Learning Efficiently Over Heterogeneous Databases

Jose Picado, Arash Termehchy, Sudhanshu Pathak

Oregon State University

IDEA Lab
Information and Data Management and Analytics

1. Information in a domain is often spread across multiple (heterogeneous) databases

IMDb

movies		
id	title	year
m1	The Hangover (2009)	2009
m2	Star Wars: Episode VII - The Force Awakens (2015)	2015

movies2releasedate		
id	month	year
m1	June	2009
m2	December	2015

...

movies[title] ≈ movies2distributors[title] → movies[title] ⇋ movies2distributors[title]

Matching dependencies (MDs) specify relationships between databases.

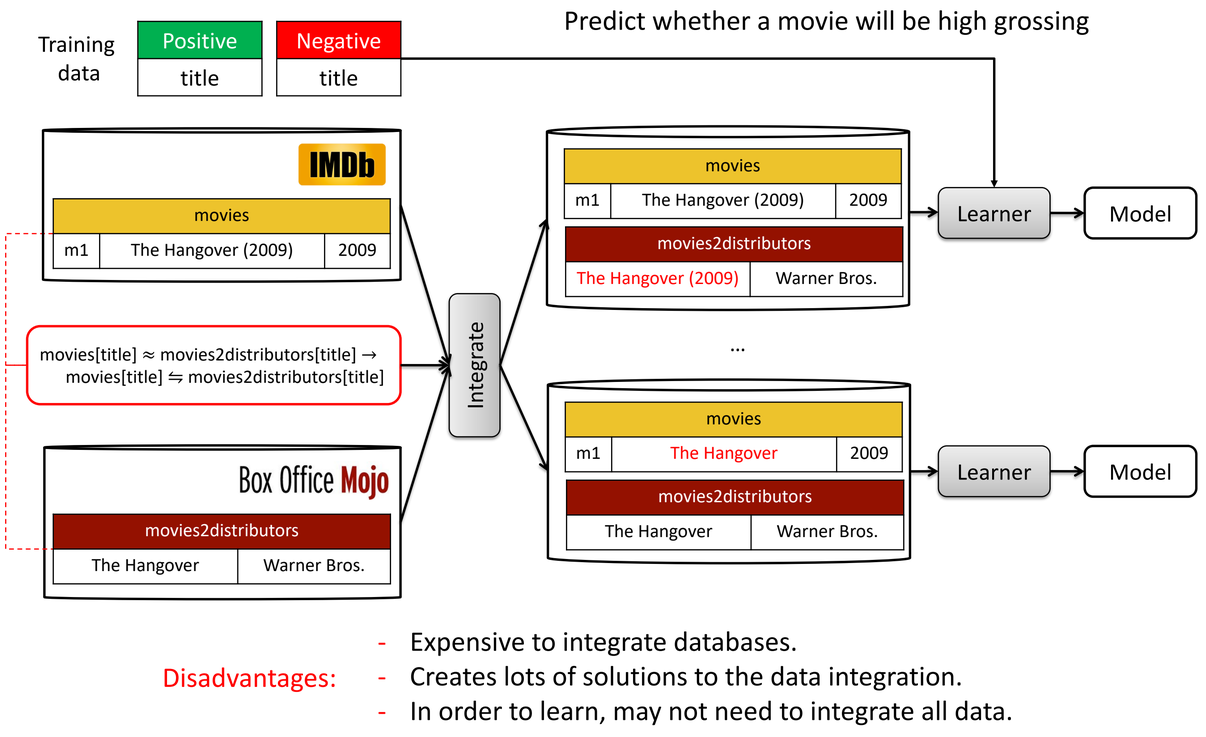
Box Office Mojo

movies2distributors	
title	distributors
The Hangover	Warner Bros.
Star Wars: The Force Awakens	Buena Vista

movies2budget	
title	budget
The Hangover	35M
Star Wars: The Force Awakens	245M

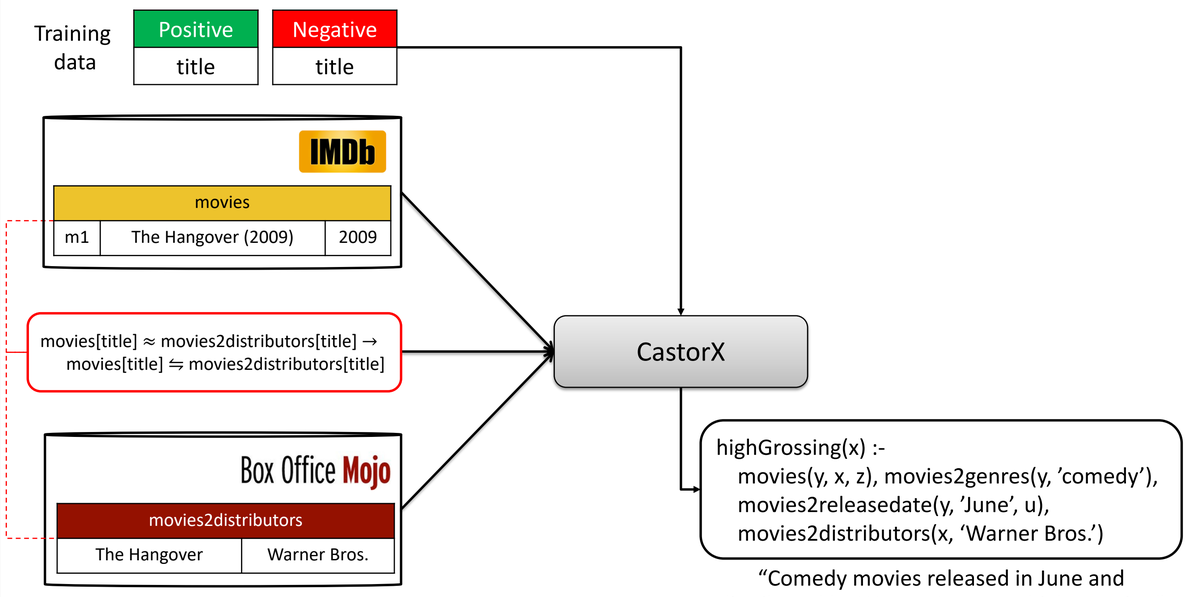
...

2. Learning over multiple databases



Disadvantages:
- Expensive to integrate databases.
- Creates lots of solutions to the data integration.
- In order to learn, may not need to integrate all data.

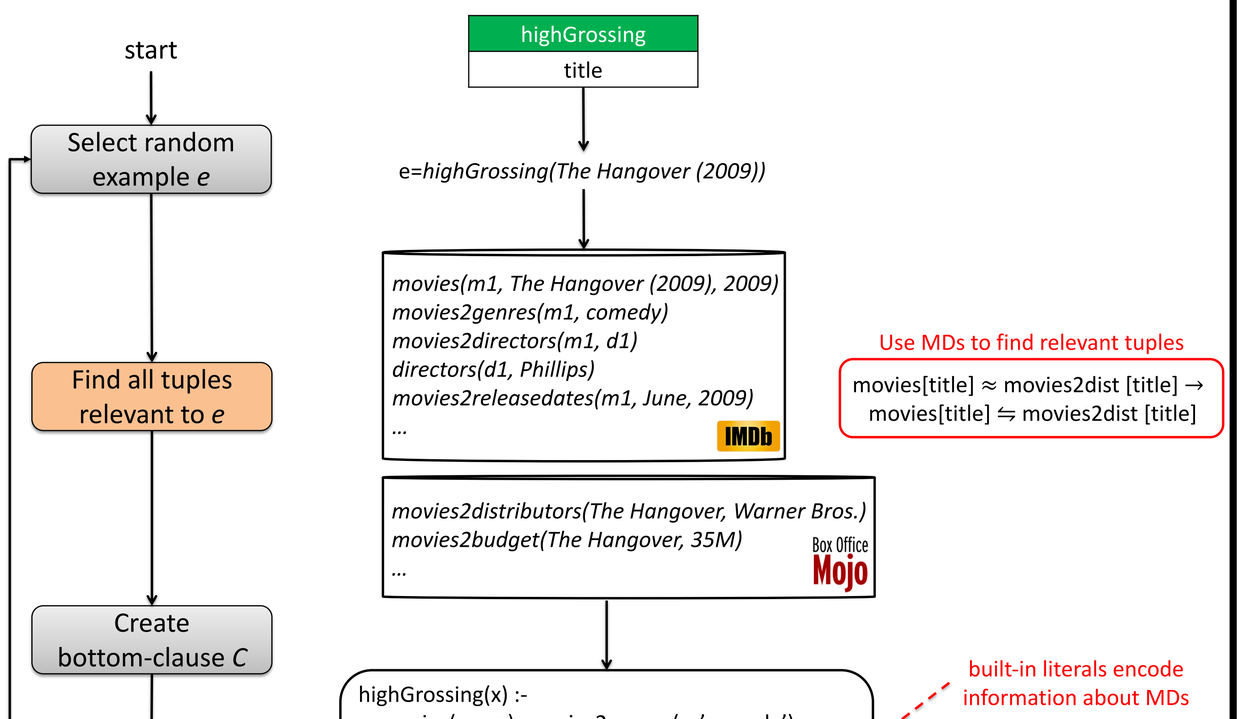
3. CastorX: cross-database relational learning system



Advantages:
+ Use only data relevant to the training data.
+ "Virtual" integration while learning.
+ Learn over non-IID data.
+ Results are interpretable.

4. One representation to encode all possible solutions to the data integration

- Encode information about matching dependencies inside the model.
- Save computations by learning one model that is valid over all possible solutions of the data integration.

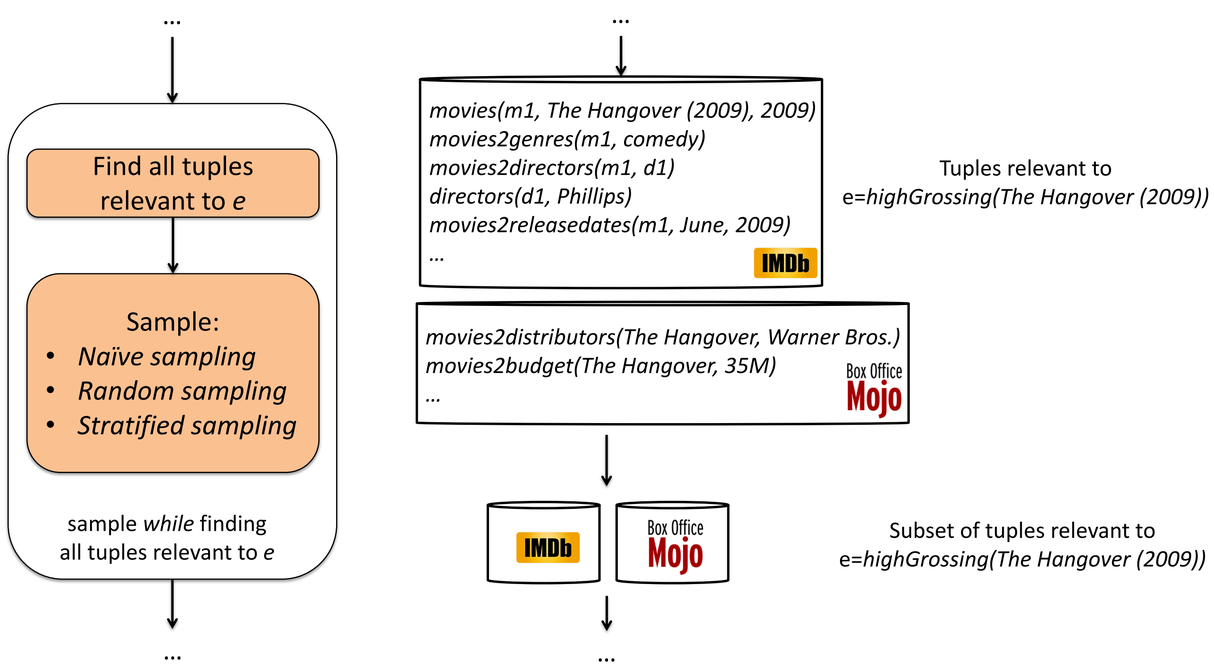


Use MDs to find relevant tuples

movies[title] ≈ movies2dist [title] → movies[title] ⇋ movies2dist [title]

5. Efficient learning using sampling

- CastorX learns efficiently by sampling the tuples relevant to a positive example *e*.

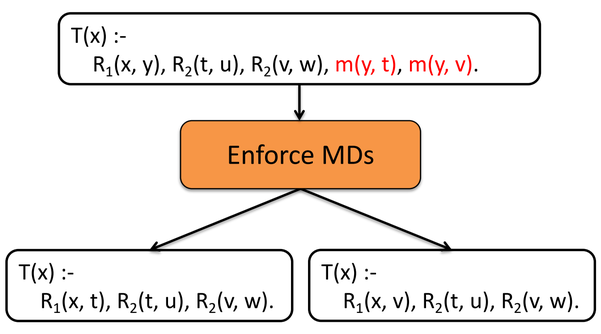


6. Enforcing matching dependencies

- After enforcing MDs, learned definitions may be:
  - Exactly the same
  - Equivalent (through homomorphism)
  - Different (can still get approximate answers)

Schema:
$R_1(A, B)$
$R_2(C, D)$

Matching dependency:
$R_1[B] \approx R_2[C] \rightarrow R_1[B] \leftrightharpoons R_2[C]$



7. Experiments

HIV DB: chemical compounds
- Target: anti-HIV(compound)
- 7.8M tuples
- 2K positive, 4K negative examples

IMDb + Box Office Mojo
- Target: highGrossing(title)
- 9M and 100K tuples
- 1K positive, 2K negative examples

Sampling method	Precision	Recall	Time (minutes)
Naïve	0.84	0.87	27.99
Random	0.79	0.81	12.57
Stratified	0.84	0.90	24.97
Naïve	0.86	0.78	59.9
Stratified	0.95	0.78	95