Motivations

Queries Featurization

Queries featurization is crucial for query-driven estimators.

- The query is represented as a collection of four sets:
    - $< Tables >, < Joins >, < Columns >, < Values >$
    - e.g.,
- Query

```
SELECT COUNT(*)
FROM Title t, Company c
WHERE t.t_id = c.t_id
        AND t.Year >= 2000
        AND c.c_id <= 3
        AND c.Zip = 125
```

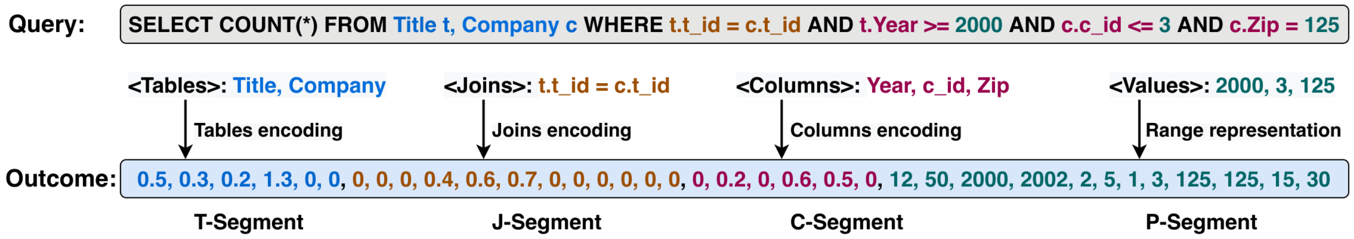
- Query representation
    - `<Tables>: Title, Company`
    - `<Joins>: t.t_id = c.t_id`
    - `<Columns>: Year, c_id, Zip`
    - `<Values>: 2000, 3, 125`
- Existing queries featurization methods cannot capture the fine-grained correlations among (Tables), (Joins), (Columns)
- Existing estimators do not give any quantification of uncertainty of the estimation.

Fauce Overview

- Overview
    - Join Schema -(Basic Information Parsing)->
    - Query Featurization -(Training Data Generation)->
        - Query Encoding
        - Predicates Representation
    - Model Design
- What Fauce includes:
    1. A new query featurization method

2. Uncertainty information of the estimations
3. Uncertainty management module

Query Featurization of Fauce



Query: SELECT COUNT(*) FROM Title t, Company c WHERE t.t_id = c.t_id AND t.Year >= 2000 AND c.c_id <= 3 AND c.Zip = 125

<Tables>: Title, Company

Tables encoding

<Joins>: t.t_id = c.t_id

Joins encoding

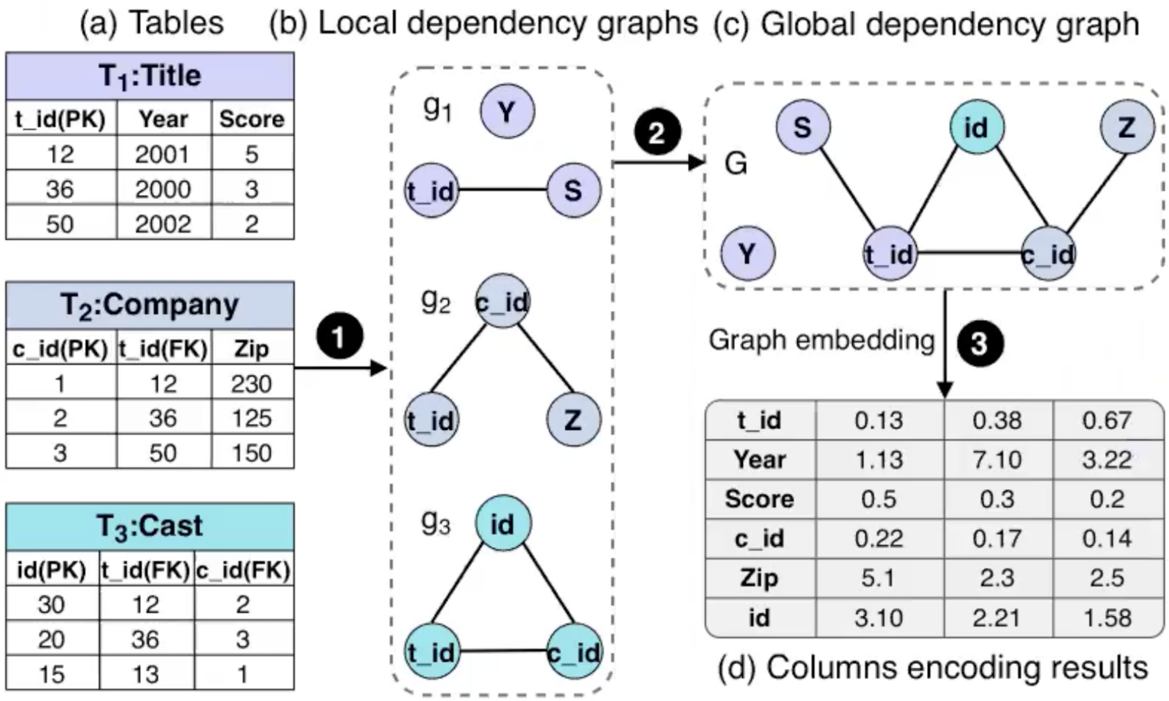
<Columns>: Year, c_id, Zip

Columns encoding

<Values>: 2000, 3, 125

Range representation

Outcome: 0.5, 0.3, 0.2, 1.3, 0, 0, 0, 0, 0, 0.4, 0.6, 0.7, 0, 0, 0, 0, 0, 0, 0, 0.2, 0, 0.6, 0.5, 0, 12, 50, 2000, 2002, 2, 5, 1, 3, 125, 125, 15, 30

- How to encode `<Joins>` of a query into vectors
  - Leverage the semantic information contained in the **Join Schema**.
  - Each sub-graph of Join Schema represent a join relationship among tables
- How to encode `<Columns>` into vectors
  - a method called Columns2Vec has been proposed for the `<Columns>` encoding.
  - Tables -> Local Dependency graphs -> Global dependency graph -> Columns encoding results
  - 

(a) Tables
(b) Local dependency graphs
(c) Global dependency graph
1
2
3
g1
Y
t_id
S
g2
c_id
t_id
Z
g3
id
t_id
c_id
G
S
id
Z
Y
t_id
c_id
Graph embedding
(d) Columns encoding results

T1:Title

t_id(PK)	Year	Score
12	2001	5
36	2000	3
50	2002	2

T2:Company

c_id(PK)	t_id(FK)	Zip
1	12	230
2	36	125
3	50	150

T3:Cast

id(PK)	t_id(FK)	c_id(FK)
30	12	2
20	36	3
15	13	1

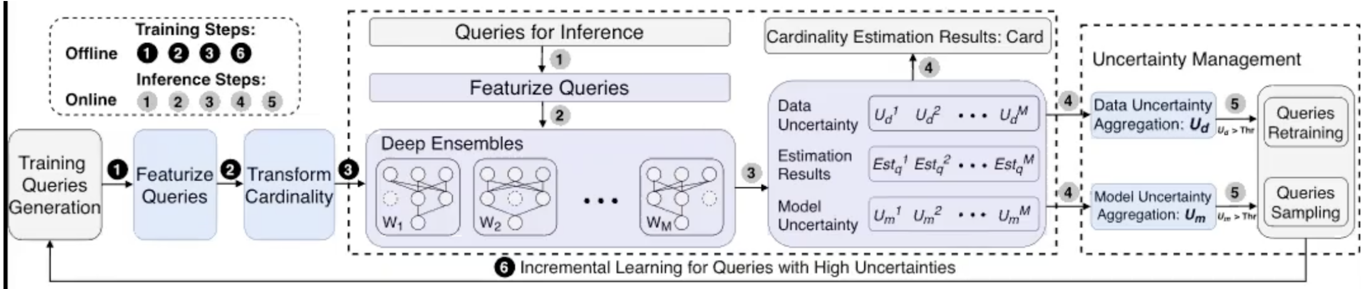
t_id	0.13	0.38	0.67
Year	1.13	7.10	3.22
Score	0.5	0.3	0.2
c_id	0.22	0.17	0.14
Zip	5.1	2.3	2.5
id	3.10	2.21	1.58

Uncertainty Quantification

- The uncertainty consists of **model uncertainty** and **data uncertainty**
  - Model uncertainty: describes how confident the learned model is

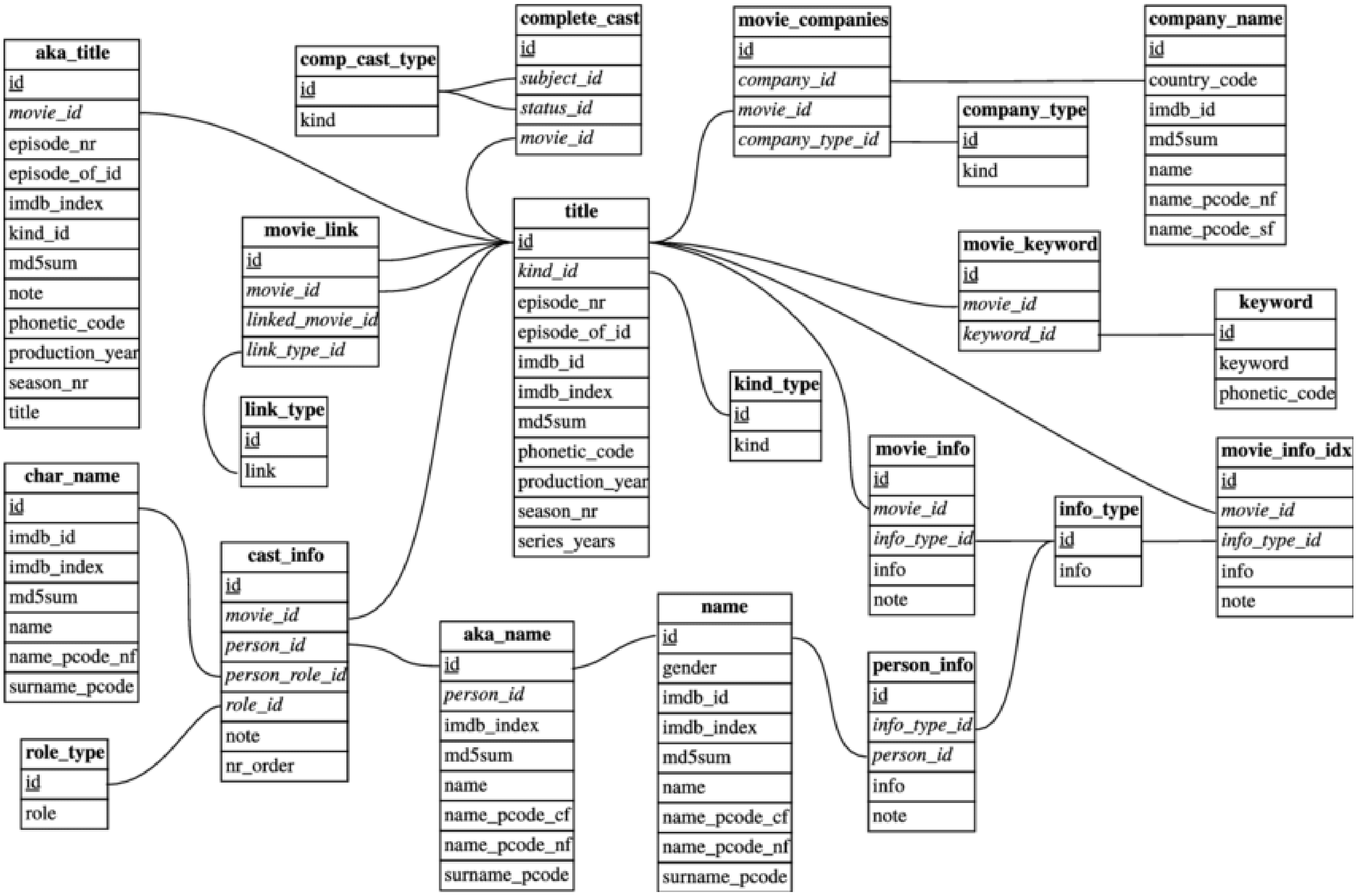
- Data uncertainty: measures how noisy the collected query data are
- $Var(y) = Var(E[y|x]) + E[Var(y|x)]$
- Model uncertainty: $Var(E[y|x])$
- Data uncertainty: $E[Var(y|x)]$
- x denotes the feature vector of a query after featurization. y denotes the query's estimated cardinality

Training and Inference of Fauce



- Data preparation
- Deep ensembles training
- Uncertainty Management

Dataset



t-mi

t-mi_idx

t-mi

t-mc

t-mk