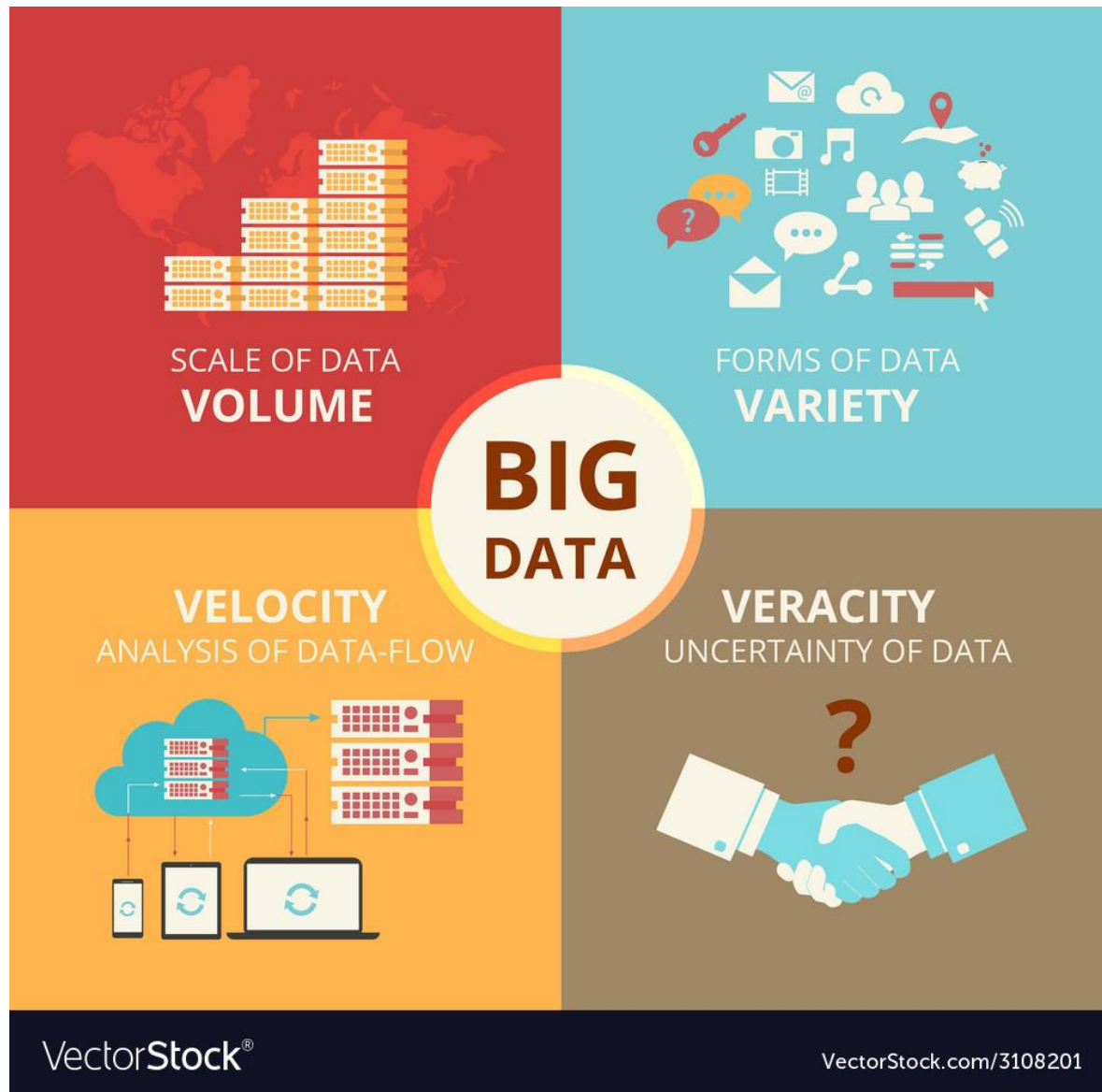


# Enhancing JSON Schema Discovery by Uncovering Hidden Data

COLLEGE OF COMPUTING & INFORMATION SCIENCES  
ROCHESTER INSTITUTE OF TECHNOLOGY  
BY JUSTIN NAMBA



# Big Data

**Unstructured data**

The university has 5600 students.  
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.  
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

**Semi-structured data**

```
<University>  
  <Student ID="1">  
    <Name>John</Name>  
    <Age>18</Age>  
    <Degree>B.Sc.</Degree>  
  </Student>  
  <Student ID="2">  
    <Name>David</Name>  
    <Age>31</Age>  
    <Degree>Ph.D. </Degree>  
  </Student>  
  ....  
</University>
```

**Structured data**

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

```
{ "asin": "0309069963", "categories": [ ["Books"] ],  
  "salesRank": { "Books": 2174268 },  
  "related": { "also_bought": [ "0465022227" ], "buy_after_viewing": [ "0465022227" ],  
              "also_viewed": [ "0465022227", "0309069963" ], "bought_together": [ "0309069963" ] } }  
  
{ "asin": "B007M6IMQ0", "title": "Adrienne Vittadini Footwear Women's Vida Flat...",  
  "salesRank": { "Shoes": 139961, "Clothing": 596278 },  
  "related": { "also_bought": [ "B006WVESEK", "B007VMCFLC" ], "buy_after_viewing": [ "B006WVESEK" ],  
              "also_viewed": [ "B006WVESEK", "B00880CLHE" ], "bought_together": [ "B006WVESEK" ] } }
```

## JSON & Nested JSON Documents

# JSON disadvantage

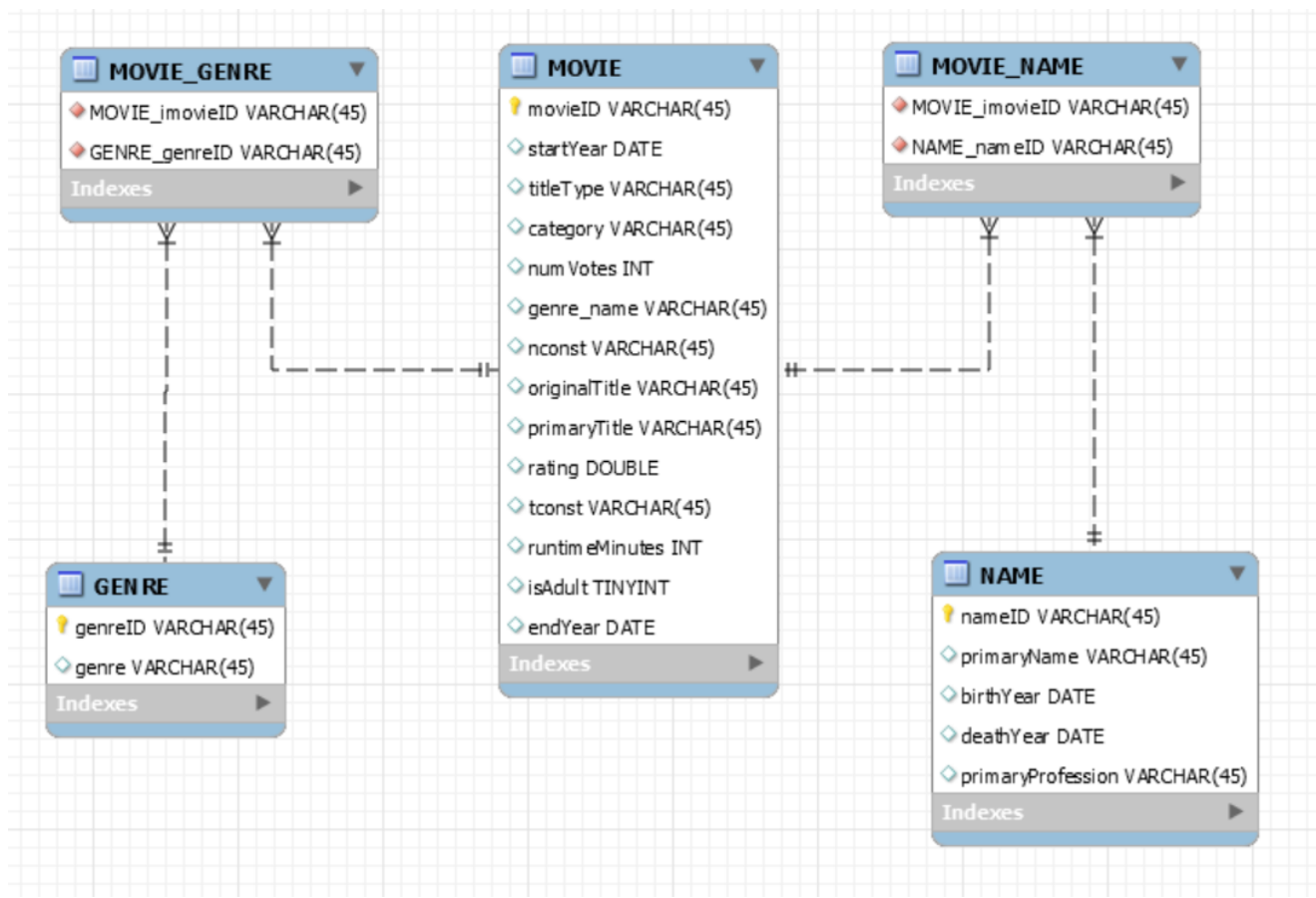
No  
predefined  
schema

Complicated  
analysis

# Goal of research

- ▶ Enhance the quality of discovered JSON schemas to make analysis easier

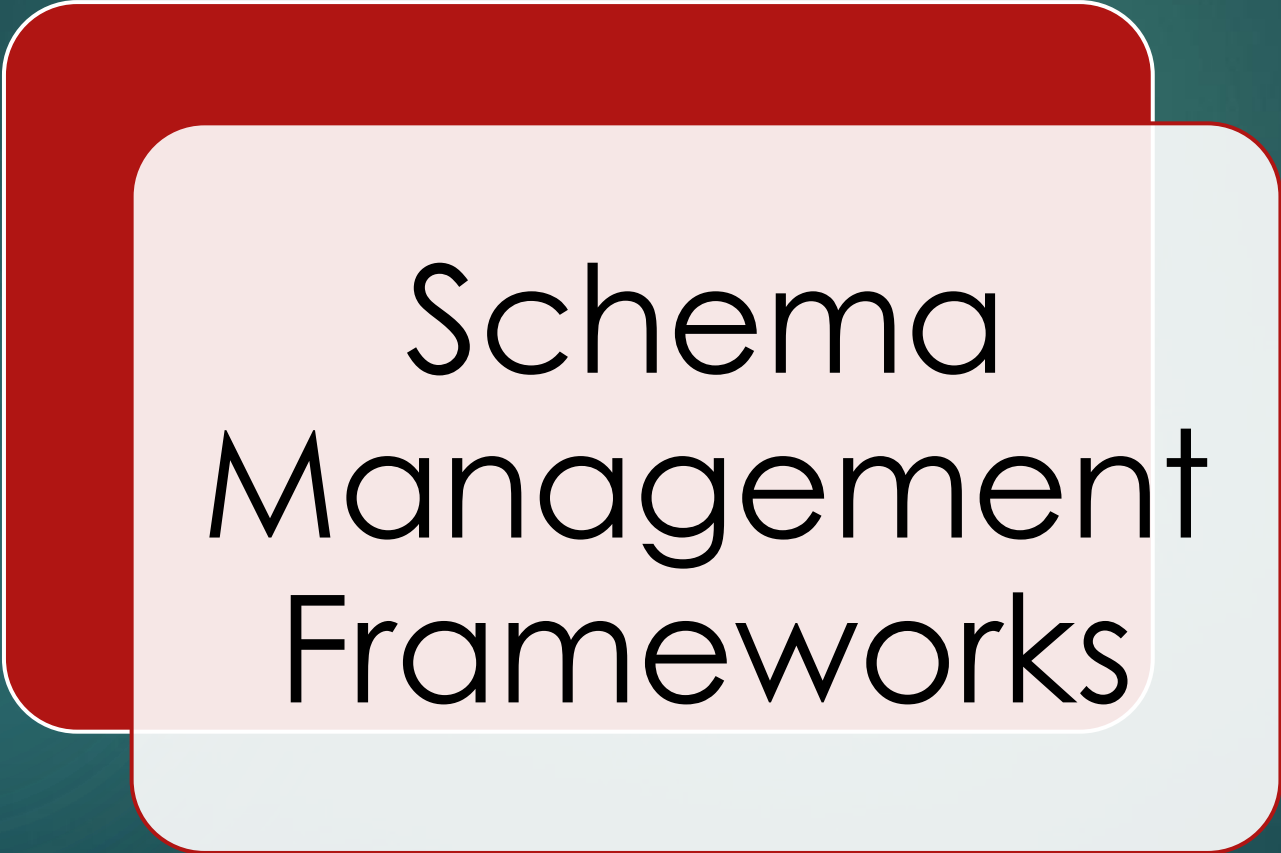




# Schema Discovery

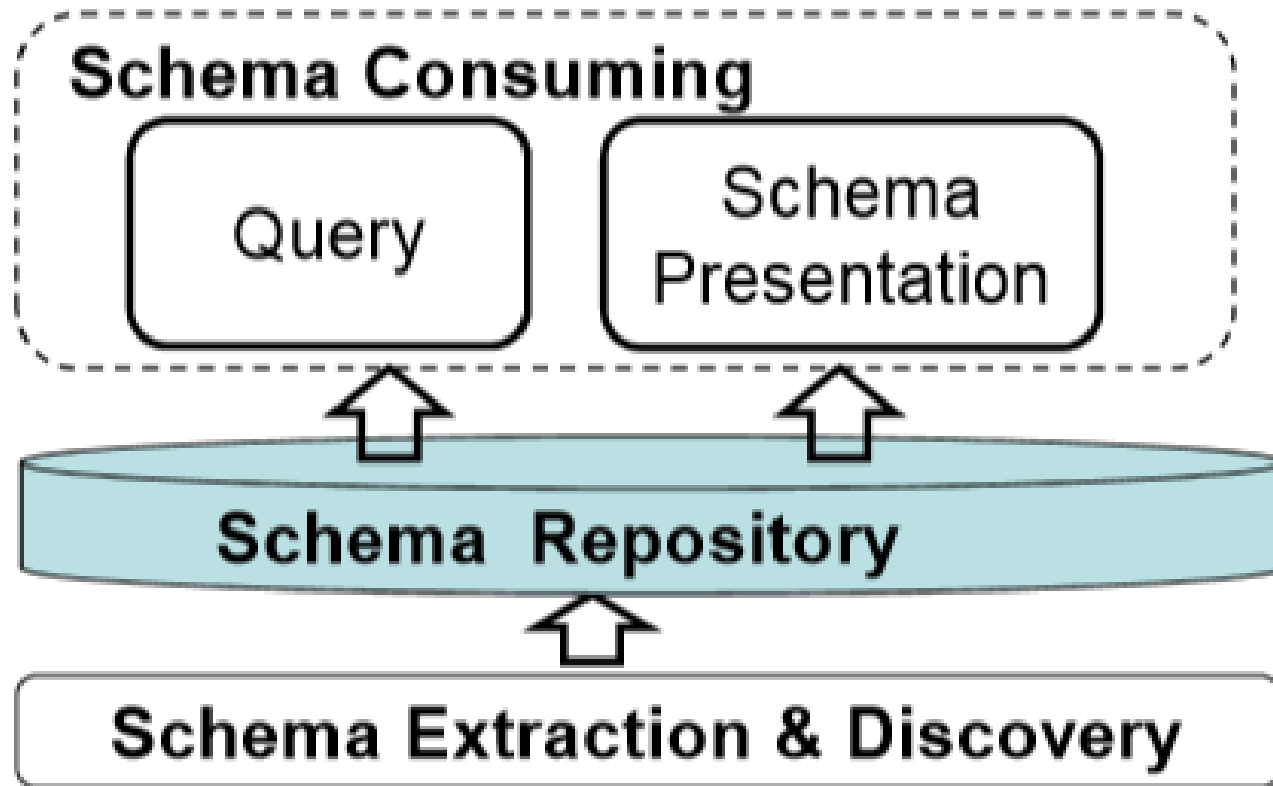
# Introduction of related works

8

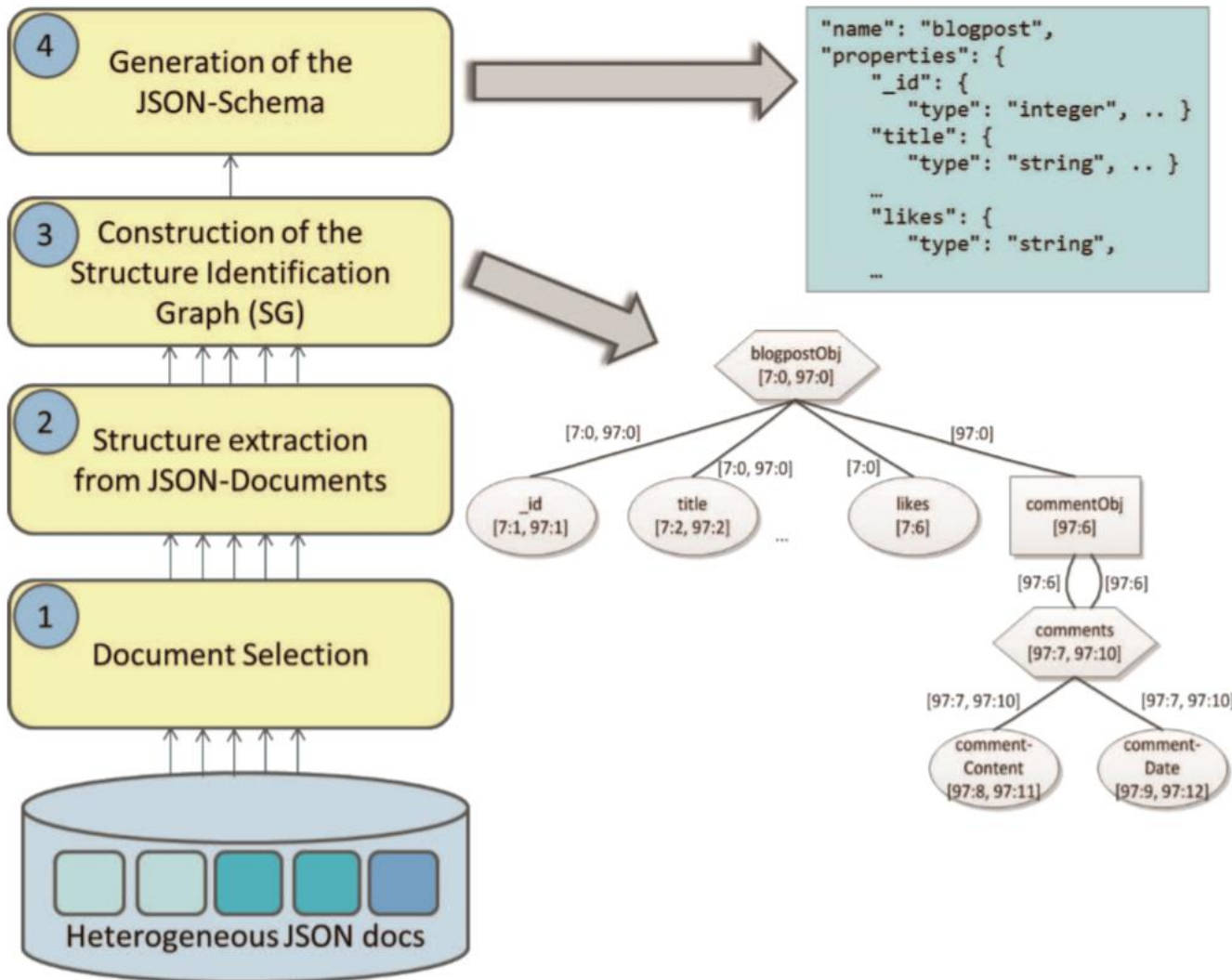


Schema  
Management  
Frameworks

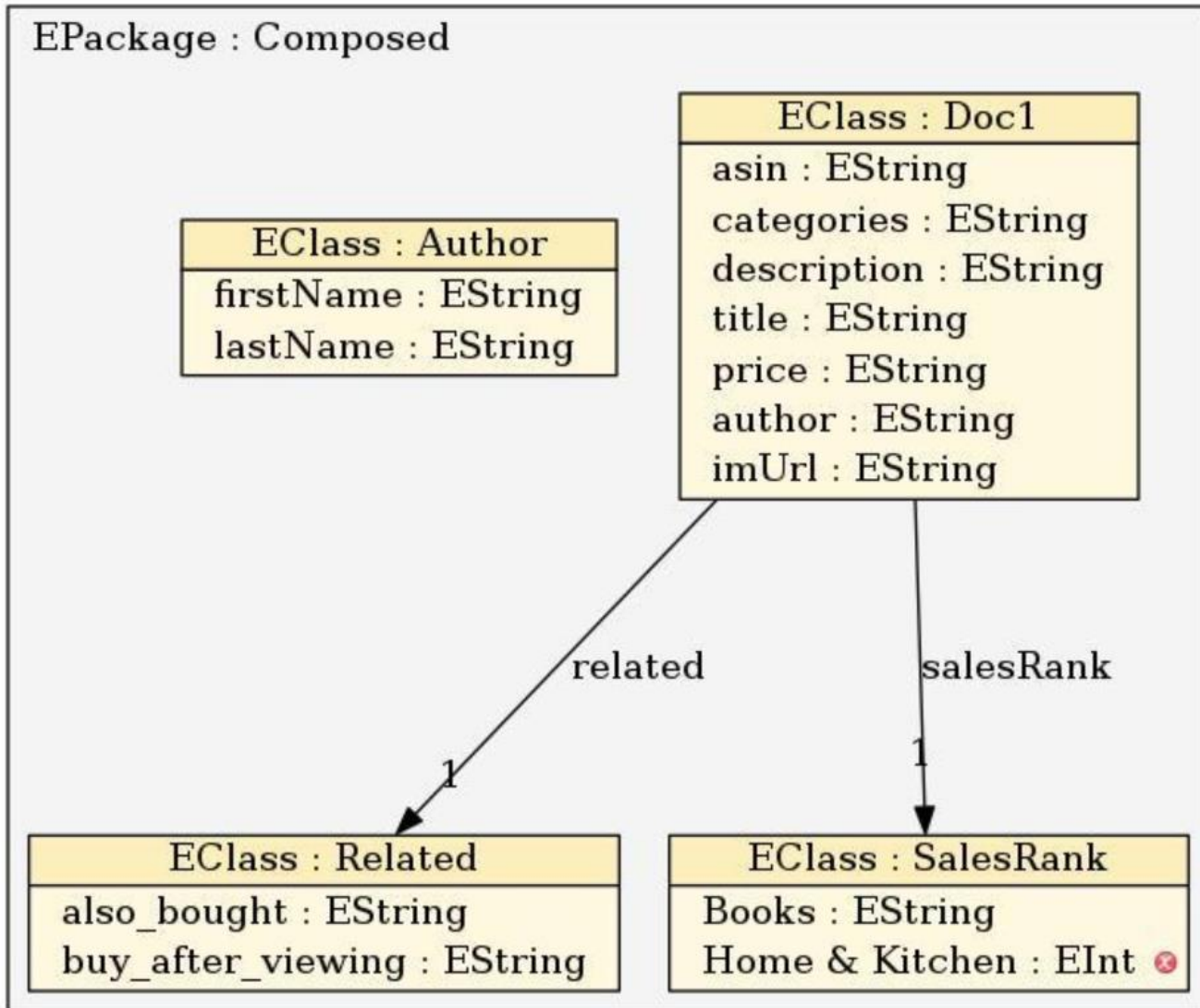




# Schema Management Framework



# JSON Schema Extraction



# JSON discoverer

# Limitations of Related Works

Only  
provide  
structure

Insufficient  
semantic  
information

# Keys Distinction: Static vs. Dynamic

```
{ "asin": "0309069963", "categories": [ ["Books"] ],  
  "salesRank": { "Books": 2174268 },  
  "related": { "also_bought": [ "0465022227" ], "buy_after_viewing": [ "0465022227" ],  
              "also_viewed": [ "0465022227", "0309069963" ], "bought_together": [ "0309069963" ] } }  
  
{ "asin": "B007M6IMQ0", "title": "Adrienne Vittadini Footwear Women's Vida Flat...",  
  "salesRank": { "Shoes": 139961, "Clothing": 596278 },  
  "related": { "also_bought": [ "B006WVESEK", "B007VMCFLC" ], "buy_after_viewing": [ "B006WVESEK" ],  
              "also_viewed": [ "B006WVESEK", "B00880CLHE" ], "bought_together": [ "B006WVESEK" ] } }
```

# Jxplain

19

- ▶ Entropy-based model
  - ▶ Datatype Entropy
  - ▶ Key Entropy
- ▶ Evaluation problem
  - ▶ Doesn't evaluate if a key is static or dynamic

# Solution: Feature-Based Classifier

20

1. Extract JSON keys
2. Choose useful features
3. Apply binary classification
4. Calculate Metrics

# Feature Domains

21

1. Intrinsic Characteristics
  1. Percentage & Nesting Level
2. Central Tendency
  1. Mean
3. Statistical Dispersion
  1. Range, Standard Deviation, Entropy
4. Distribution Shape
  1. Skewness & Kurtosis
5. Semantic & Contextual Similarity
  1. Distinct Subkeys, Distinct Sub-keys Datatypes, Average Sub-key Contextual Similarity
6. Structural Similarity
  1. Grouping



# Complex Features

- ▶ Semantic & Contextual Similarity
  - ▶ Distinct Subkeys, Distinct Sub-keys Datatypes, Average Sub-key Contextual Similarity
- ▶ Structural Similarity
  - ▶ Grouping

```
{ "pegi": {  
  "pegi_url": "https://store.cloudflare",  
  "pegi_tags": ["Blood", "and", "Gore"]},  
  "requirements": {  
    "minimum": {  
      "windows": {  
        "processor": "1 GHz Intel...",  
        "memory": "1024 MB RAM",  
      },  
      "linux": {  
        "processor": "1 GHz Intel...",  
        "memory": "1024 MB RAM",  
      },  
      "macOS": {  
        "processor": "SSE2 inst...",  
        "memory": ""}}}}}
```

# Data Pre-processing

23

1. Data Source
2. Feature Extraction
3. Normalization
4. Data labelization
5. Oversample minority class

# Evaluation

24

- ▶ Classifiers
  - ▶ Logistic Regression
  - ▶ Random Forest
  - ▶ Support Vector Machines (SVMs)
- ▶ Evaluation
  - ▶ Average F1-Score of dynamic keys

# Results

25

	Intrinsic Feat.	Central Tend. Feat.	Dispersion Feat.	Dist. Shape Feat.	Add. Feat.	Grouping
Classifier	F1-score	F1-score	F1-score	F1-score	F1-score	F1-score
Logistic regression	0.0897	0.0906	0.0921	0.0826	0.4875	0.4875
Random forest	0.1106	0.1198	0.1447	0.1272	0.5616	0.5016
SVMs	0.1110	0.1129	0.1029	0.0880	0.4218	0.4218

# Conclusion

## Summary

- Insufficient information from discovered JSON Schemas
- Feature-Based Classifier
- Test & Evaluate algorithms

## Future Works

- Identify which features are detrimental to the classifiers
- Analyze mis-classified keys



## Questions & Answers