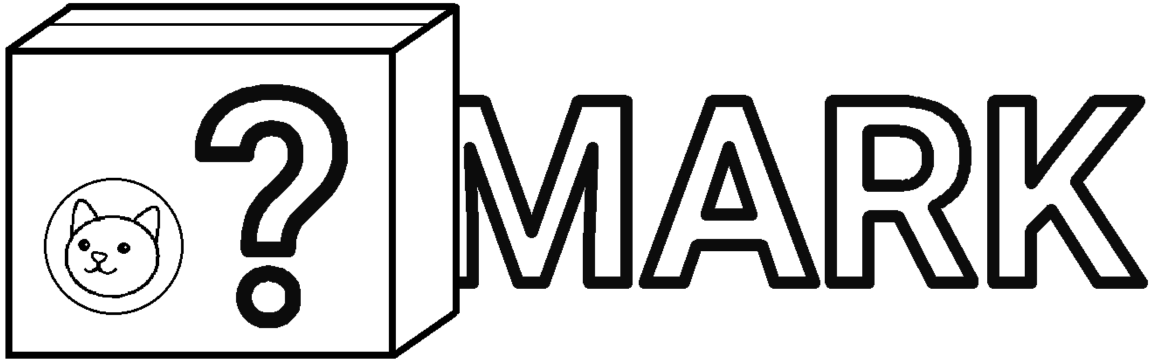
UNIVERSITY OF TWENTE.

QuestionMark

Designing a benchmark for probabilistic databases

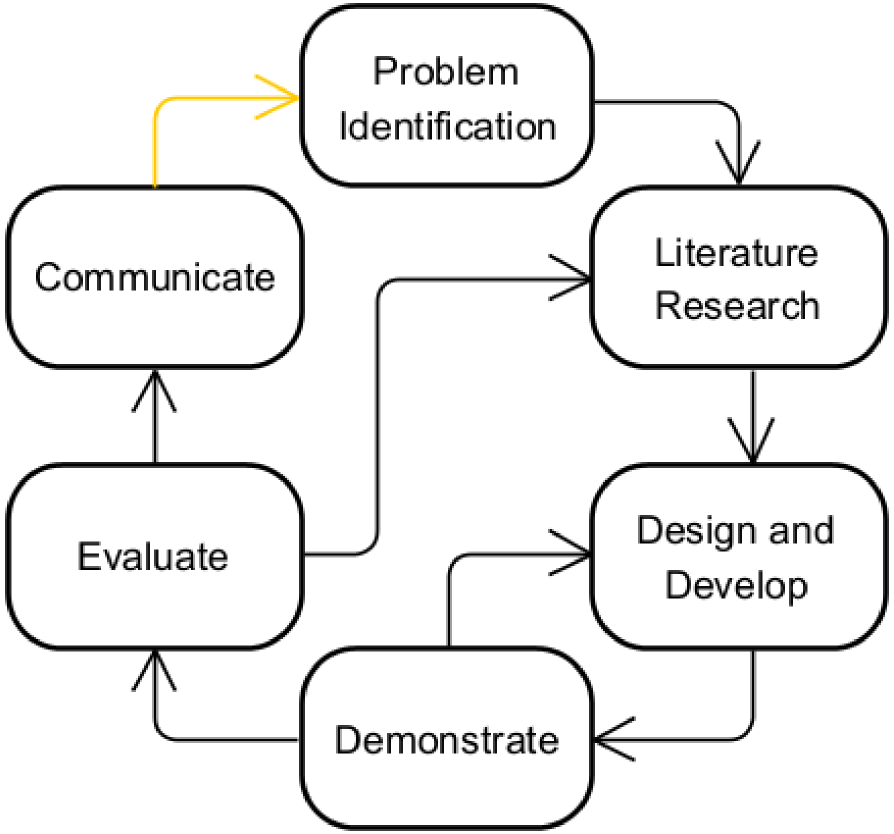
Nikki Zandbergen

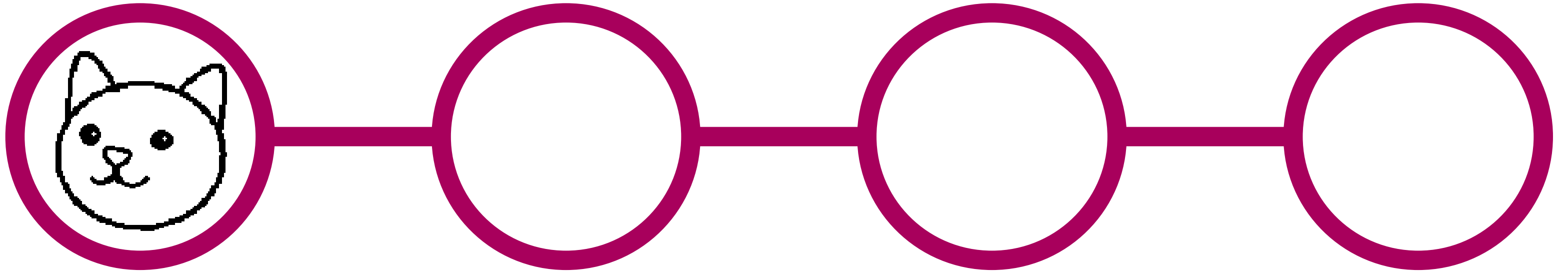


14 JULY 2023

What will be covered?

- Introduction subject
- Background
- QuestionMark
- Conclusion

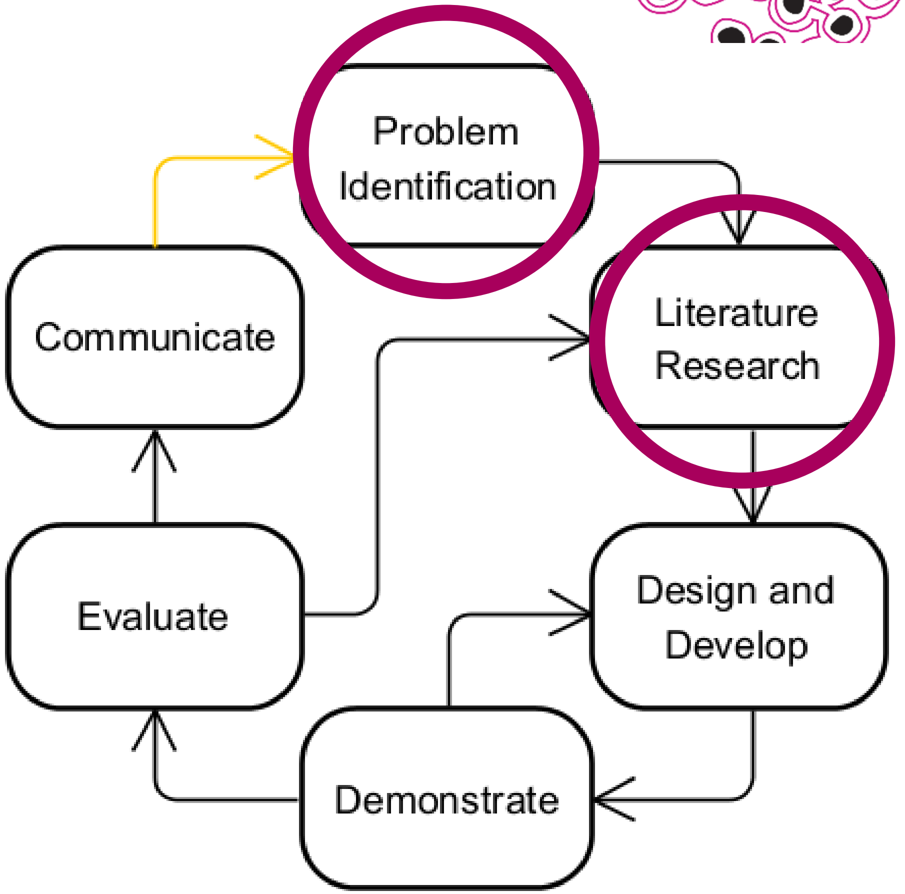


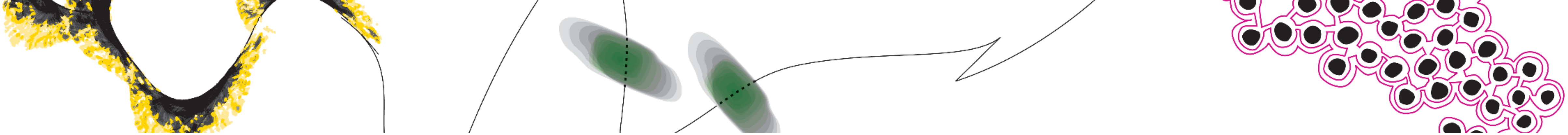


UNIVERSITY OF TWENTE.

What will be covered?

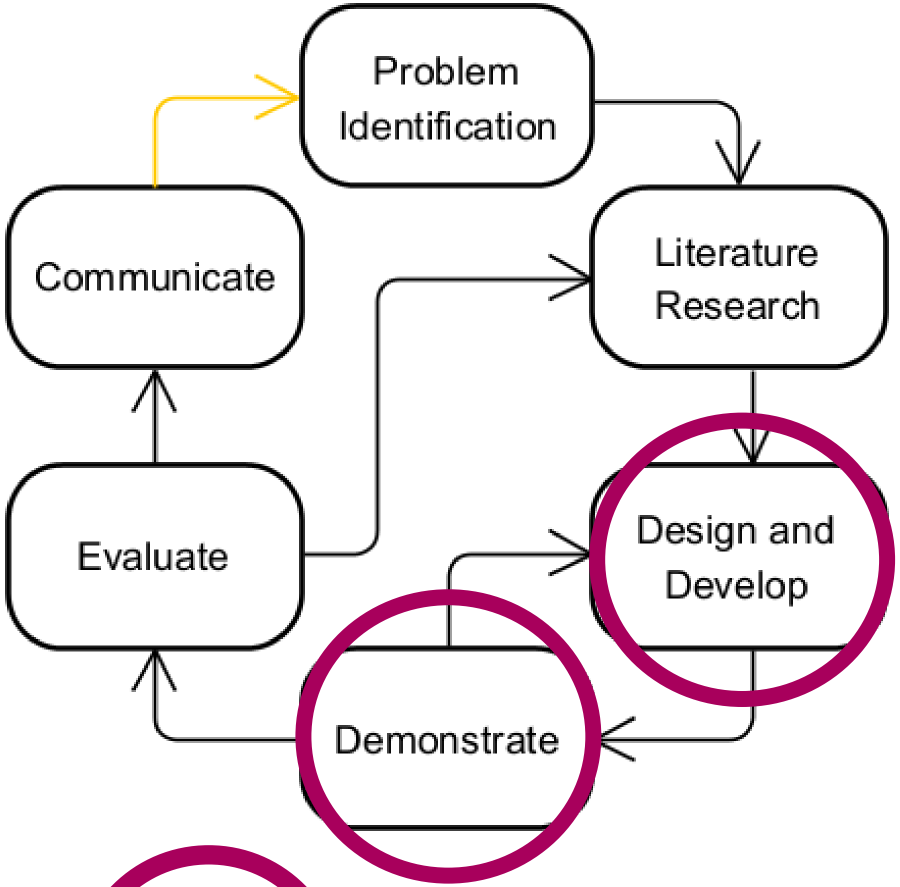
- Introduction subject
- Background
- QuestionMark
- Conclusion

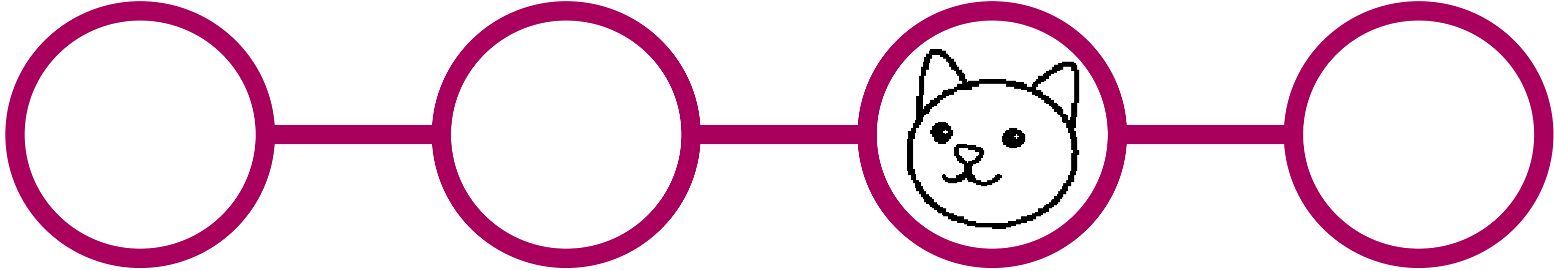




What will be covered?

- Introduction subject
- Background
- QuestionMark
- Conclusion

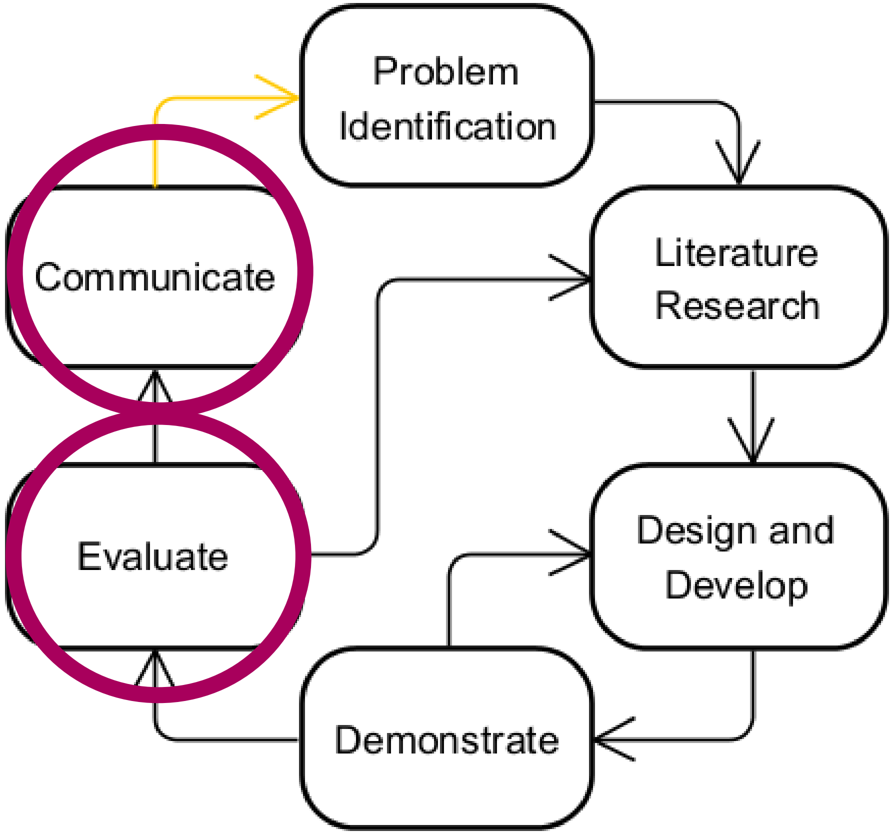




UNIVERSITY OF TWENTE.

What will be covered?

- Introduction subject
- Background
- QuestionMark
- Conclusion



Problem Identification

Literature Research

Design and Develop

Demonstrate

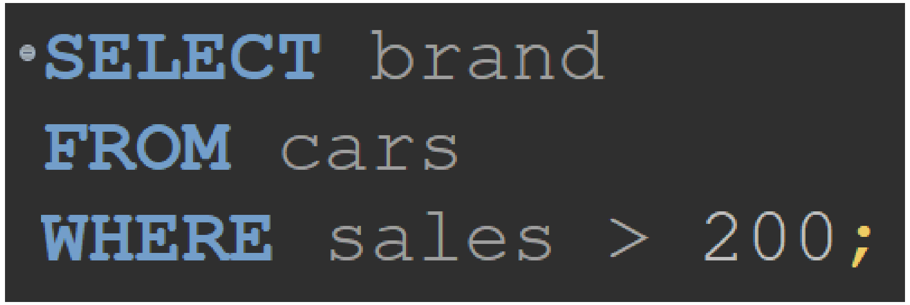
Evaluate

Communicate

Why probabilistic databases?

- Car and cargo company Car&Co
- Lousy data management
- Wants partnerships with top customers

  > 200 sales per month



```
•SELECT brand
 FROM cars
 WHERE sales > 200;
```

Let’s use a probabilistic DBMS!

Let’s use a probabilistic DBMS!

But which to pick?

MAYBMS

MCDB

Trio

DuBio

PossDB

We need a benchmark!

We need a benchmark!

But there is no good one…

Research Questions

- How can a benchmark be designed to test and compare probabilistic database management systems on real-world strain?

- How do the novel probabilistic database DuBio and the state-of-the-art MayBMS perform when benchmarking these technologies with the developed benchmark?

UNIVERSITY
OF TWENTE.

Benchmarking

- Standardised manner to test systems
- Effectiveness. Efficiency. Appeal.
- Dataset and queries.

Probabilistic Databases

- Models a set of possible databases
- Annotated with confidence score
- Possible Worlds
- Probabilistic Database

$$\boxed{\langle R_1^i, \ldots, R_k^i, p^{[i]} \rangle \in W}$$

$$\boxed{\begin{array}{l} W = \{\langle R_1^1, \ldots, R_k^1, p^{[1]} \rangle, \ldots, \langle R_1^n, \ldots, R_k^n, p^{[n]} \rangle\} \\ where \sum_{1 \leq i \leq n} p^{[i]} = 1. \end{array}}$$

DuBio

offers

id	name	sales	_sentence
1	BMW	150	Bdd(a1=1, w1)
2	B.M.W.	127	Bdd(a1=2, w1, a2=1, w2)
3	Audi	194	Bdd(a2=2, w2)

_dict

name	dict
mydict	a1=1:0.3, a1=2:0.7, a2=1:0.4, a2=2:0.6, w1:0.5, w2:0.5

DuBio

offers

id	name	sales	_sentence
1	BMW	150	Bdd(a1=1, w1)
2	B.M.W.	127	Bdd(a1=2, w1, a2=1, w2)
3	Audi	194	Bdd(a2=2, w2)

_dict

name	dict
mydict	a1=1:0.3, a1=2:0.7, a2=1:0.4, a2=2:0.6, w1:0.5, w2:0.5

DuBio

offers

id	name	sales	_sentence
1	BMW	150	Bdd(a1=1, w1)
2	B.M.W.	127	Bdd(a1=2, w1, a2=1, w2)
3	Audi	194	Bdd(a2=2, w2)

_dict

name	dict
mydict	a1=1:0.3, a1=2:0.7, a2=1:0.4, a2=2:0.6, w1:0.5, w2:0.5

MayBMS

offers

id	name	sales	v0	d0	p0	v1	d1	p1
1	BMW	150	1	1	0.3	1	1	0.5
2	B.M.W.	127	1	2	0.7	1	1	0.5
2	B.M.W.	127	2	1	0.4	2	1	0.5
3	Audi	194	2	2	0.6	2	1	0.5

UNIVERSITY
OF TWENTE.

MayBMS

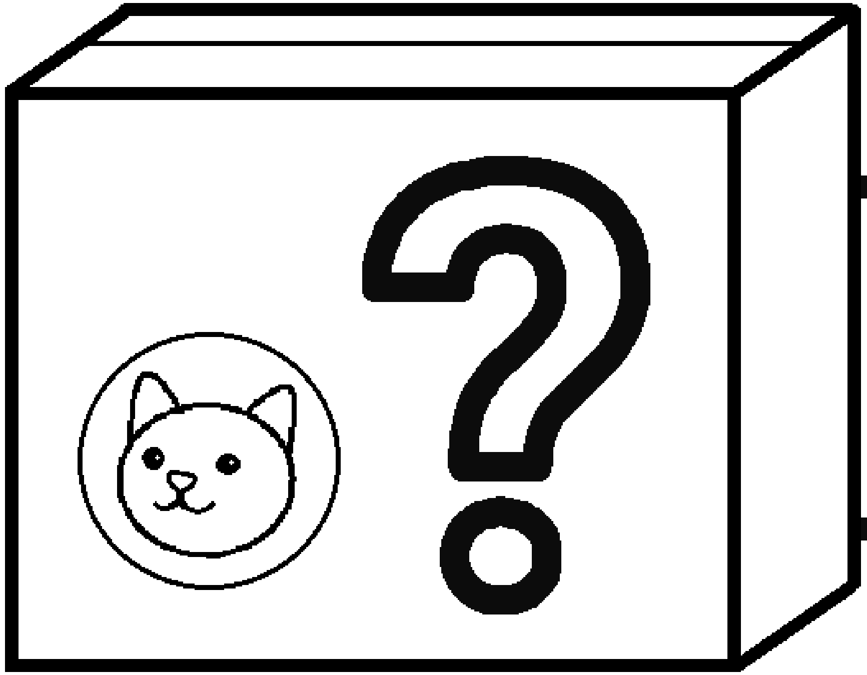
offers

id	name	sales	v0	d0	p0	v1	d1	p1
1	BMW	150	1	1	0.3	1	1	0.5
2	B.M.W.	127	1	2	0.7	1	1	0.5
2	B.M.W.	127	2	1	0.4	2	1	0.5
3	Audi	194	2	2	0.6	2	1	0.5

MayBMS

offers

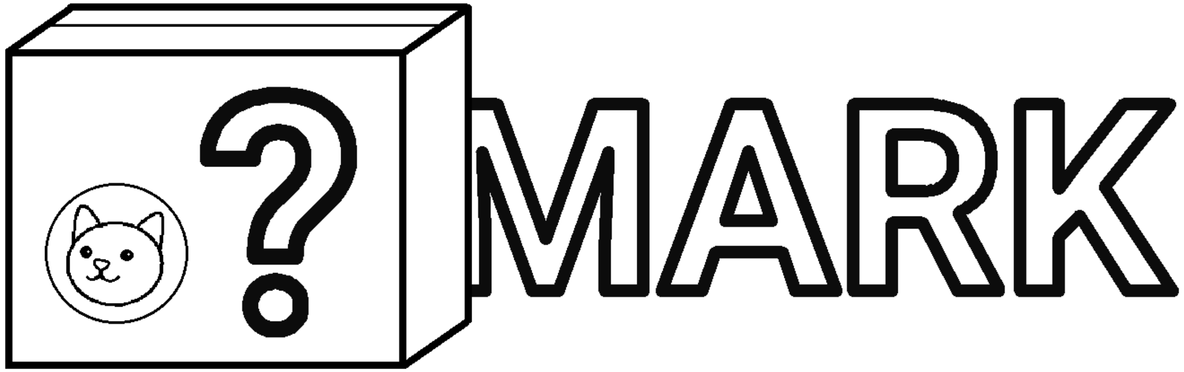
id	name	sales	v0	d0	p0	v1	d1	p1
1	BMW	150	1	1	0.3	1	1	0.5
2	B.M.W.	127	1	2	0.7	1	1	0.5
2	B.M.W.	127	2	1	0.4	2	1	0.5
3	Audi	194	2	2	0.6	2	1	0.5

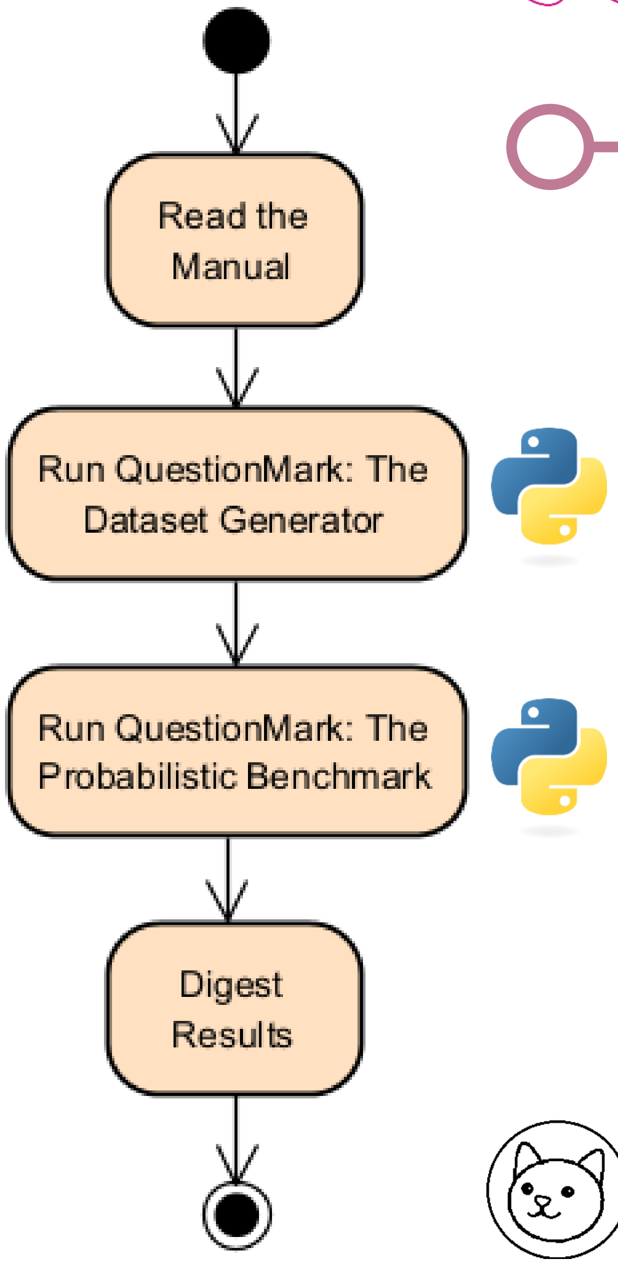


MARK

QuestionMark

- User Manual
- The Dataset Generator
- The Probabilistic Benchmark



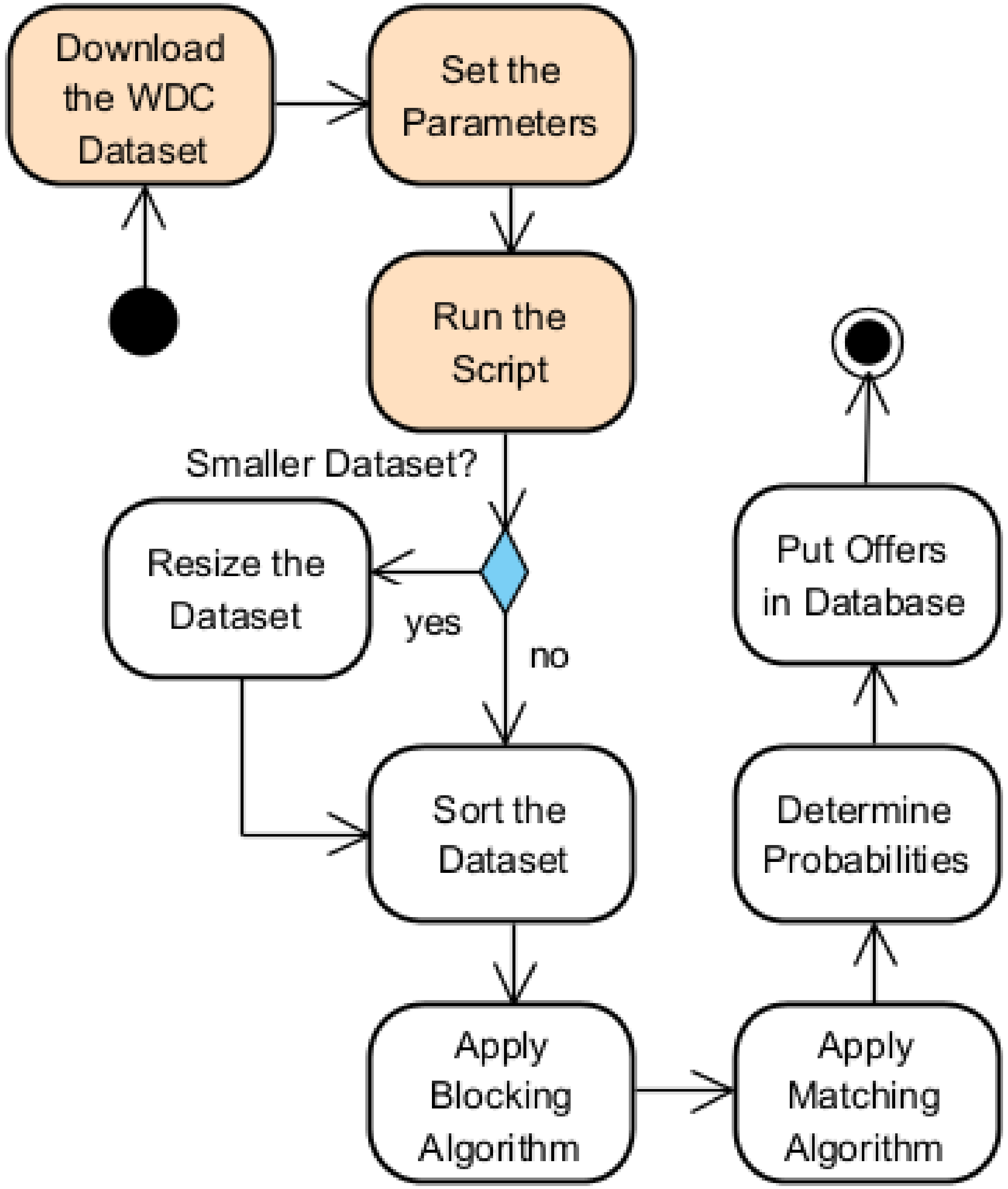

Read the Manual

Run QuestionMark: The Dataset Generator

Run QuestionMark: The Probabilistic Benchmark

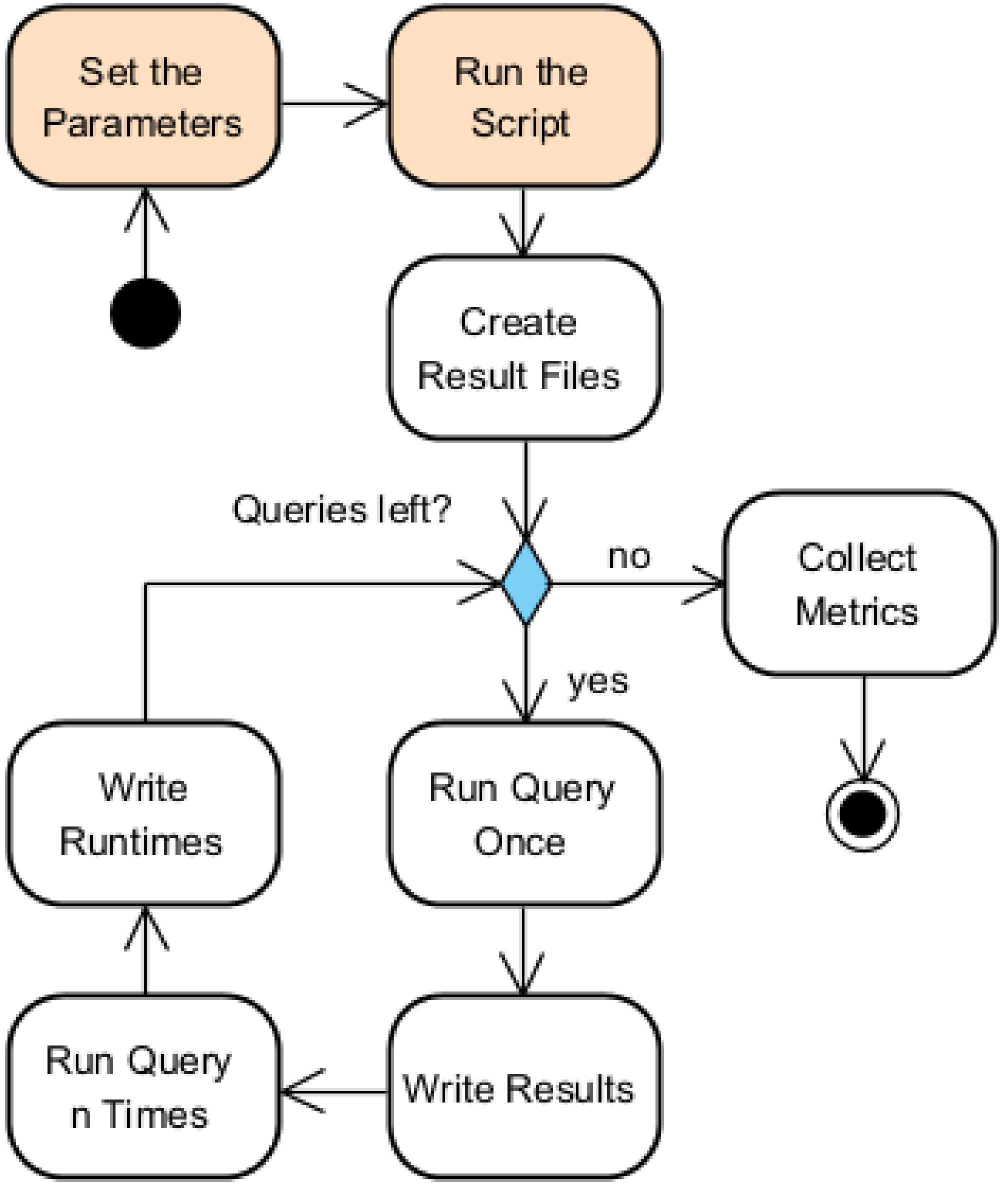
Digest Results

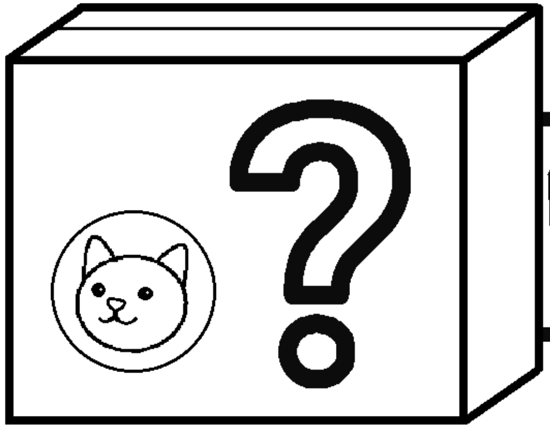
The Dataset Generator



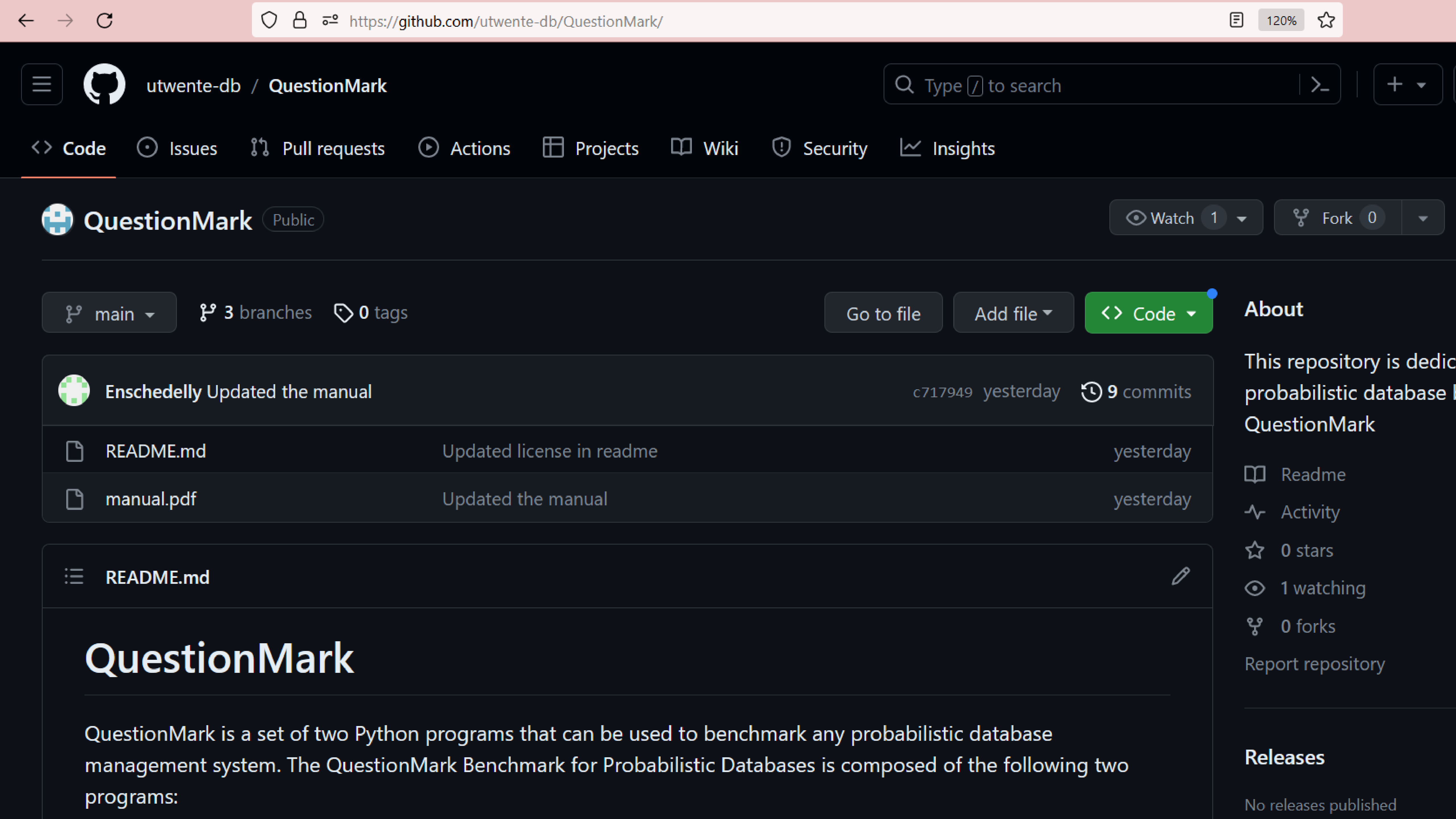
UNIVERSITY OF TWENTE.

The Probabilistic Benchmark





MARK
DEMO



Enschedelly Updated the manual c717949 yesterday 9 commits

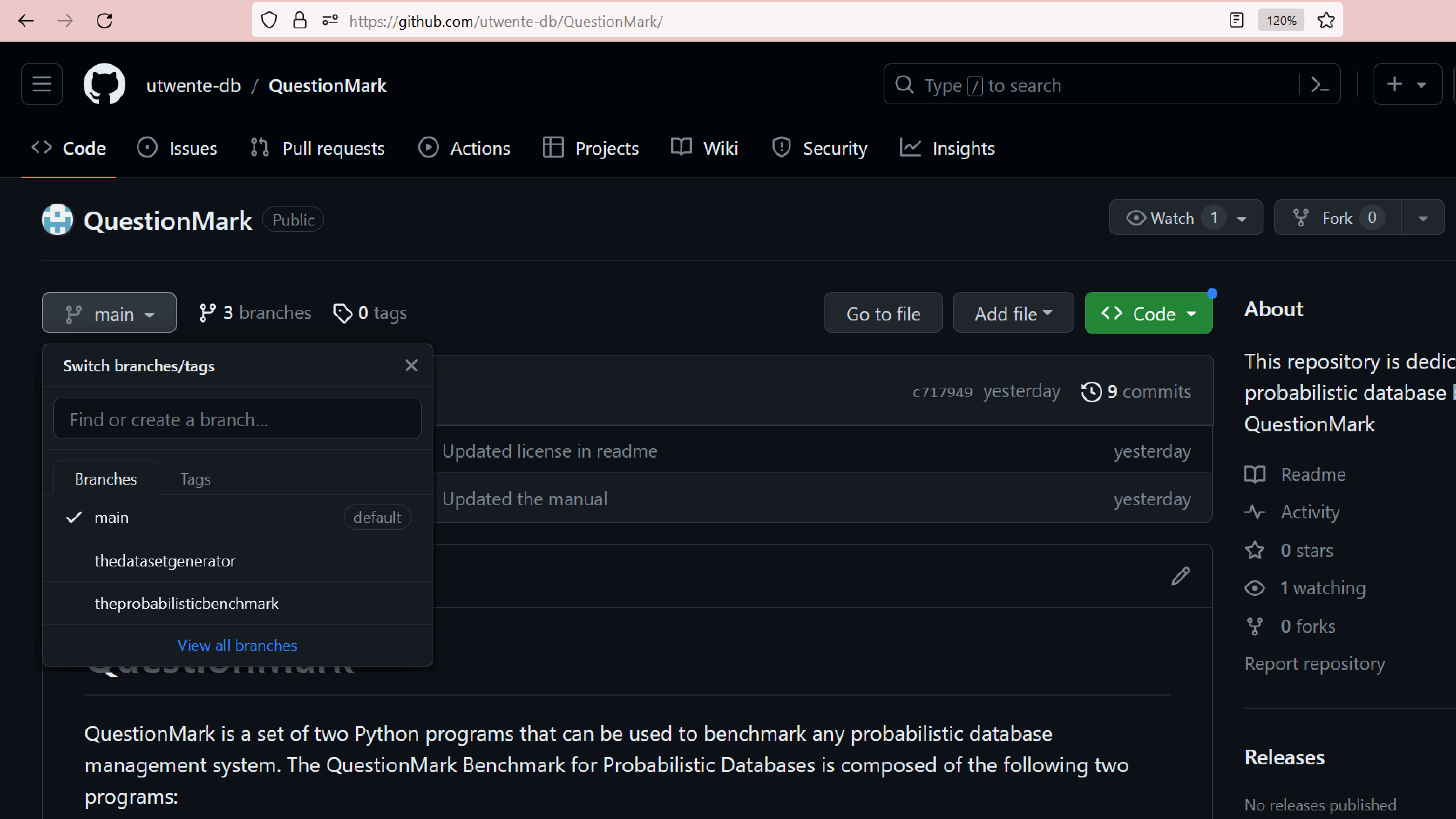
README.md	Updated license in readme	yesterday
manual.pdf	Updated the manual	yesterday

This repository is dedic
probabilistic database
QuestionMark

README.md

QuestionMark

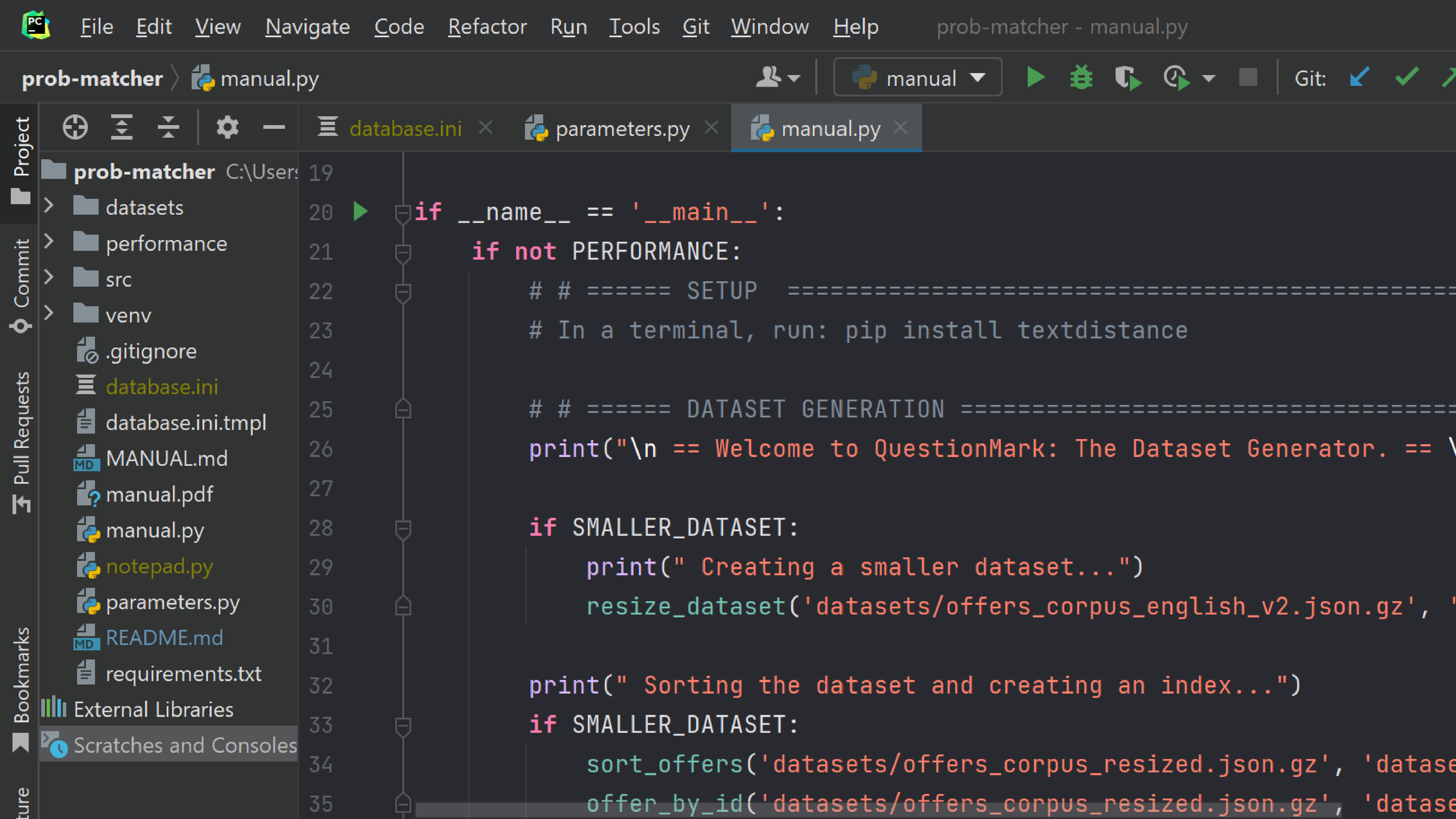
QuestionMark is a set of two Python programs that can be used to benchmark any probabilistic database management system. The QuestionMark Benchmark for Probabilistic Databases is composed of the following two programs:



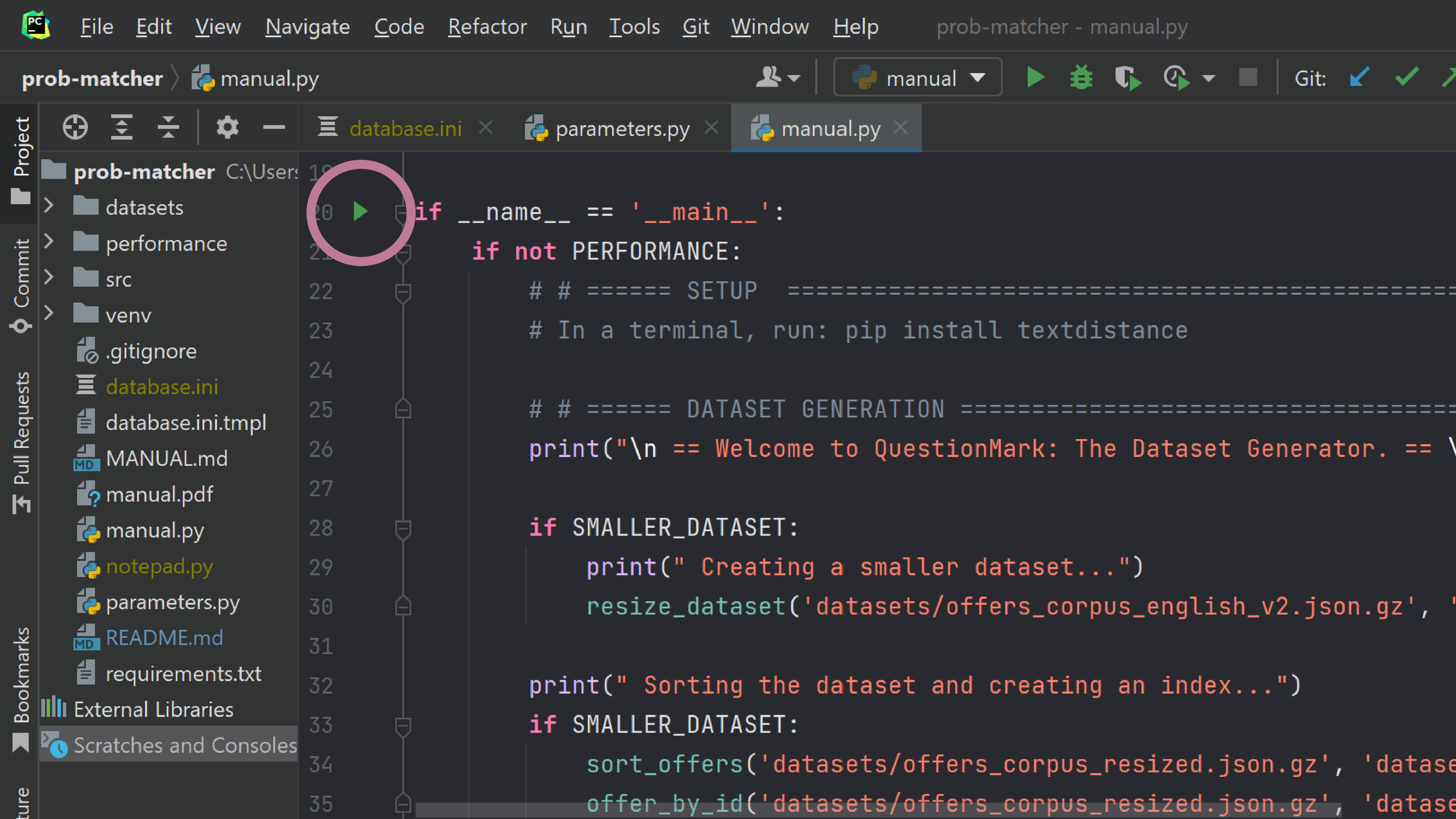
This repository is dedic
probabilistic database
QuestionMark

QuestionMark

QuestionMark is a set of two Python programs that can be used to benchmark any probabilistic database management system. The QuestionMark Benchmark for Probabilistic Databases is composed of the following two programs:



```

if __name__ == '__main__':
    if not PERFORMANCE:
        # # ====== SETUP  ==========================================
        # In a terminal, run: pip install textdistance

        # # ====== DATASET GENERATION ================================
        print("\n == Welcome to QuestionMark: The Dataset Generator. == \

        if SMALLER_DATASET:
            print(" Creating a smaller dataset...")
            resize_dataset('datasets/offers_corpus_english_v2.json.gz', '

        print(" Sorting the dataset and creating an index...")
        if SMALLER_DATASET:
            sort_offers('datasets/offers_corpus_resized.json.gz', 'datase
            offer_by_id('datasets/offers_corpus_resized.json.gz', 'datase
```

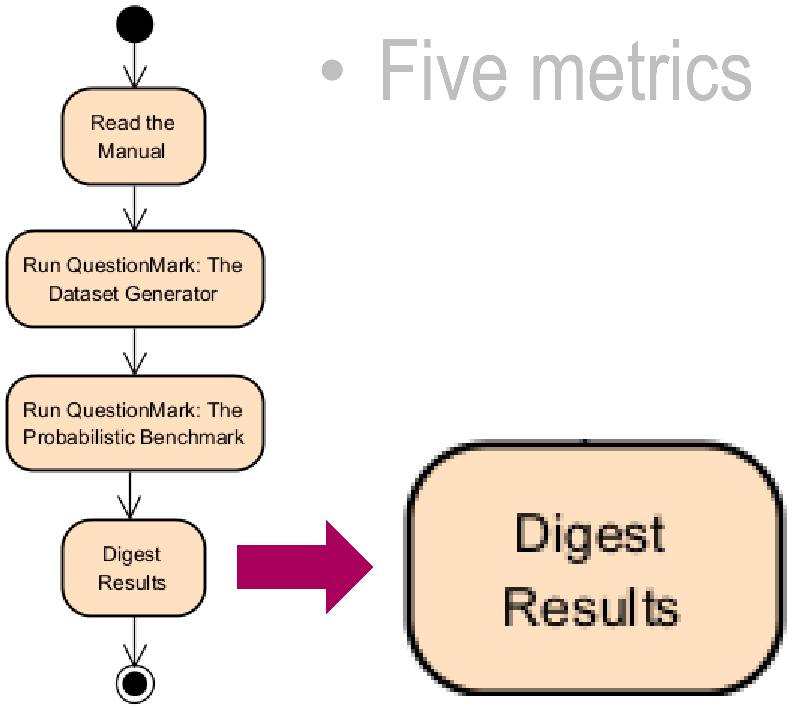


```
if __name__ == '__main__':
    if not PERFORMANCE:
        # # ====== SETUP  ===========================================
        # In a terminal, run: pip install textdistance

        # # ====== DATASET GENERATION ==================================
        print("\n == Welcome to QuestionMark: The Dataset Generator. == \

        if SMALLER_DATASET:
            print(" Creating a smaller dataset...")
            resize_dataset('datasets/offers_corpus_english_v2.json.gz', '

        print(" Sorting the dataset and creating an index...")
        if SMALLER_DATASET:
            sort_offers('datasets/offers_corpus_resized.json.gz', 'datase
            offer_by_id('datasets/offers_corpus_resized.json.gz', 'datase
```

Digesting the Results

- Three aspects
- Five metrics

- Effectiveness
- Efficiency
- Appeal


Read the Manual

Run QuestionMark: The Dataset Generator

Run QuestionMark: The Probabilistic Benchmark

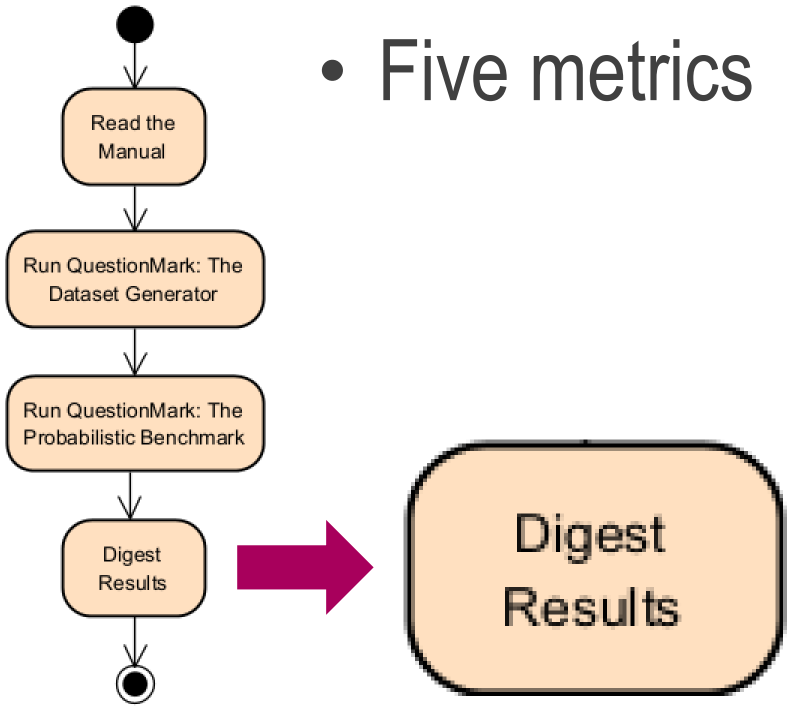
Digest Results

Digest Results

UNIVERSITY OF TWENTE.

Digesting the Results

- Three aspects
- Five metrics



Read the Manual

Run QuestionMark: The Dataset Generator

Run QuestionMark: The Probabilistic Benchmark

Digest Results

Digest Results

- Query Functionality Coverage
- Brevity of the Query Dialect
- Runtime of Queries
- Probabilistic Data Overhead
- User Friendliness

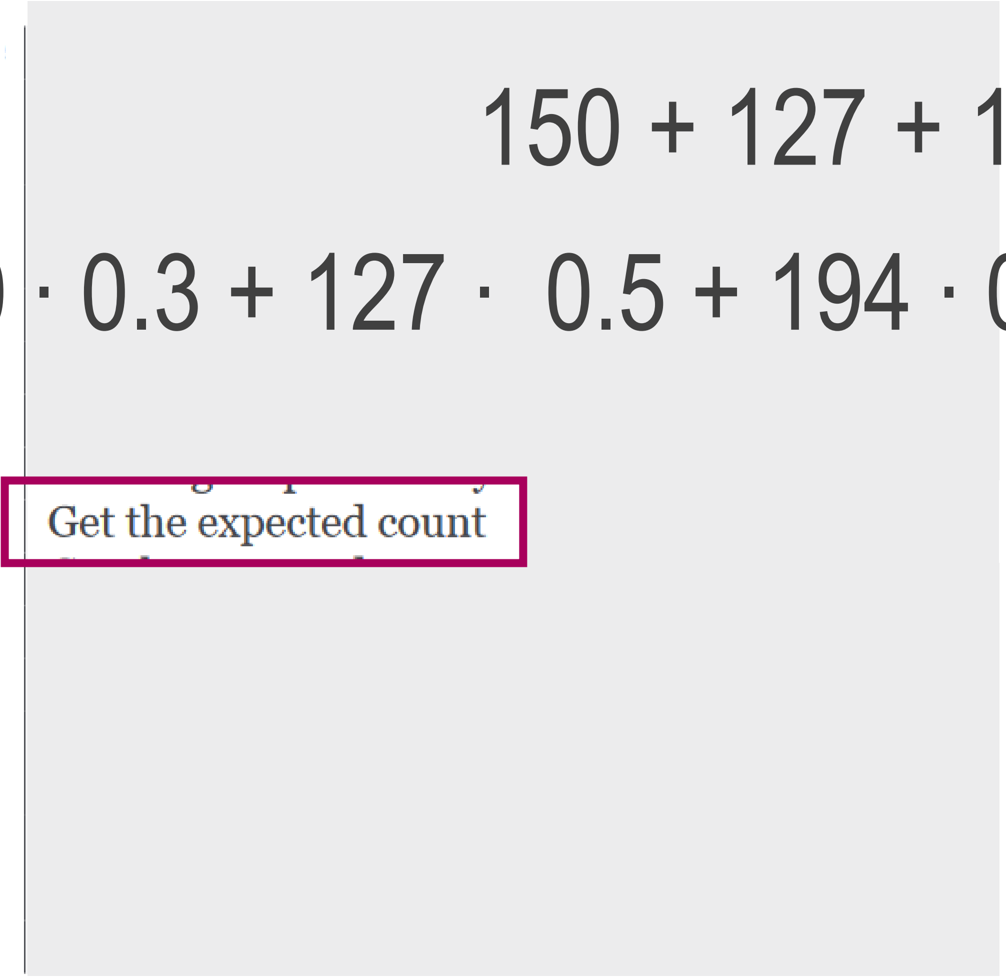
Putting QuestionMark to the test

- DuBio
- MayBMS

1 – Query Functionality Coverage

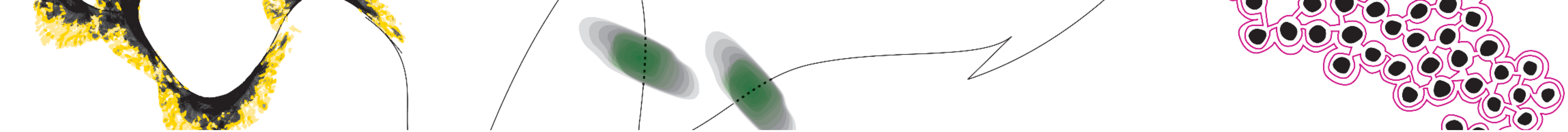
$$150 + 127 + 194 = 471$$

$$150 \cdot 0.3 + 127 \cdot 0.5 + 194 \cdot 0.2 = 147.3$$

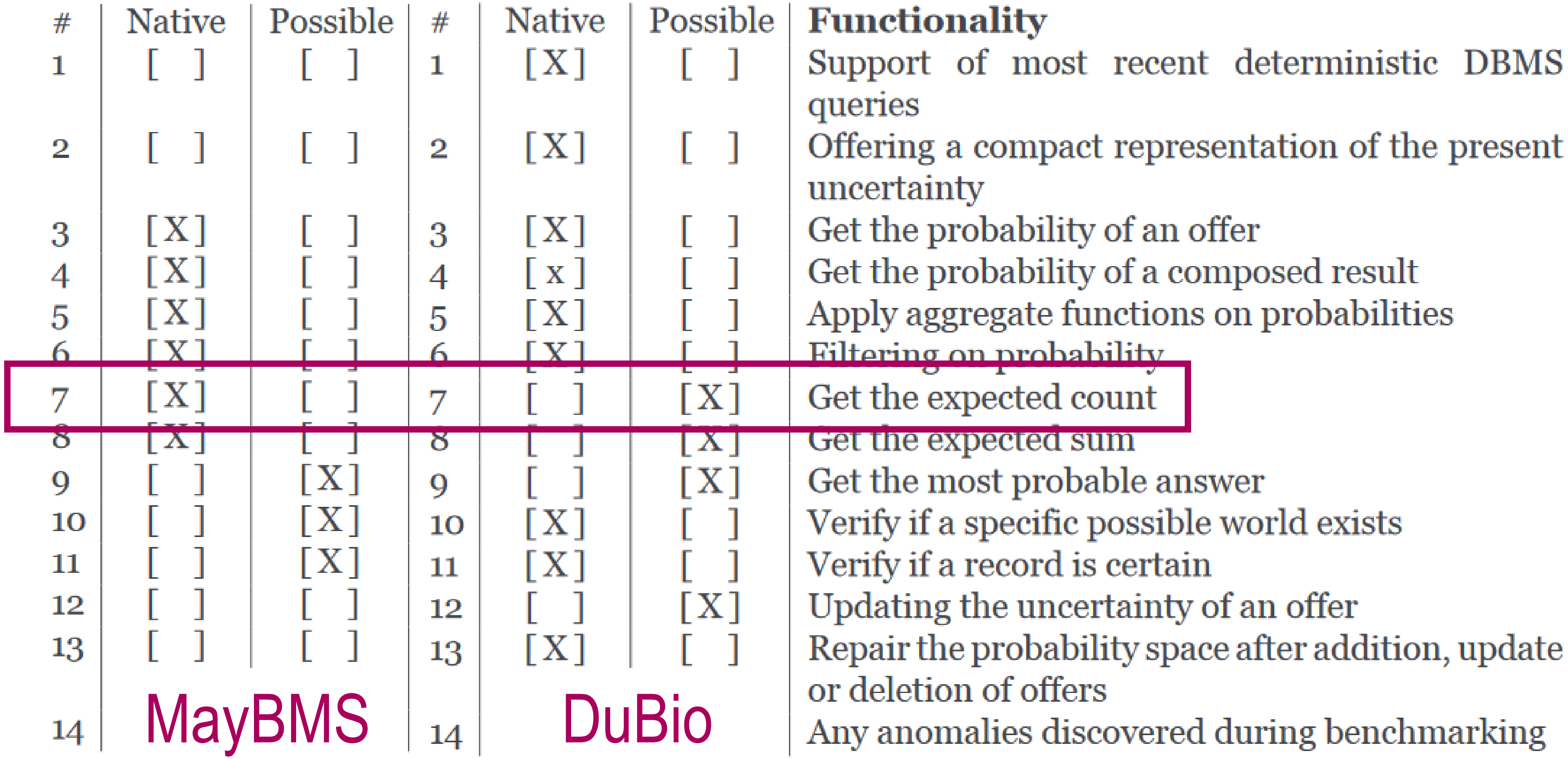


Get the expected count

UNIVERSITY OF TWENTE.

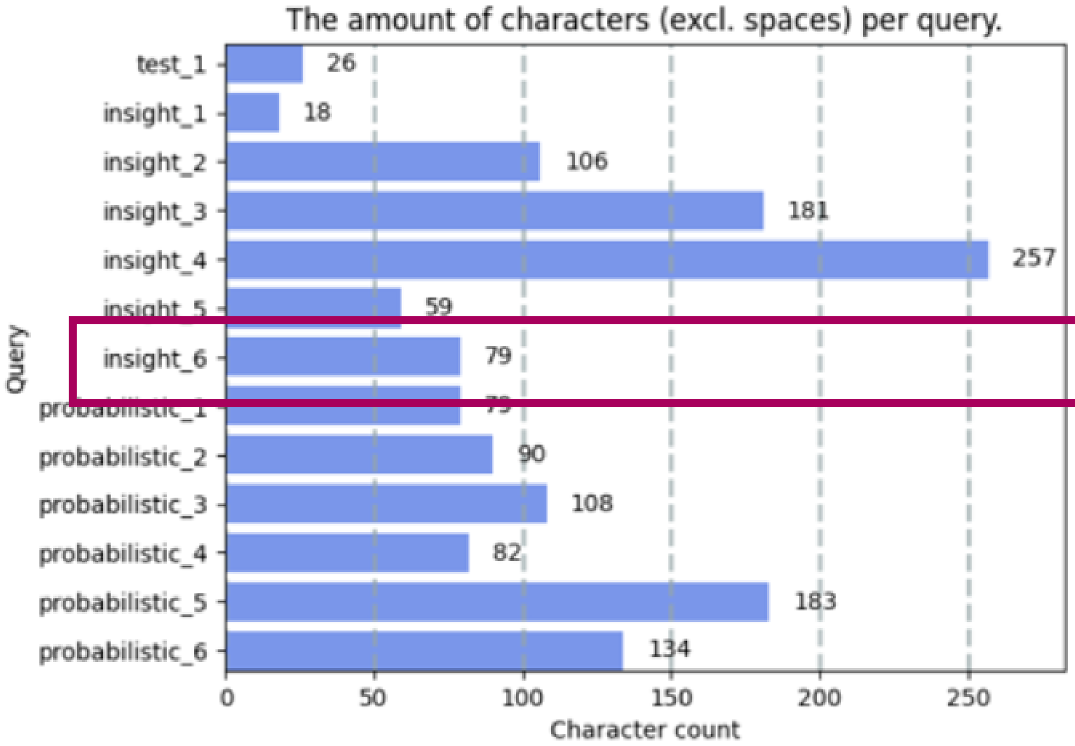


1 – Query Functionality Coverage

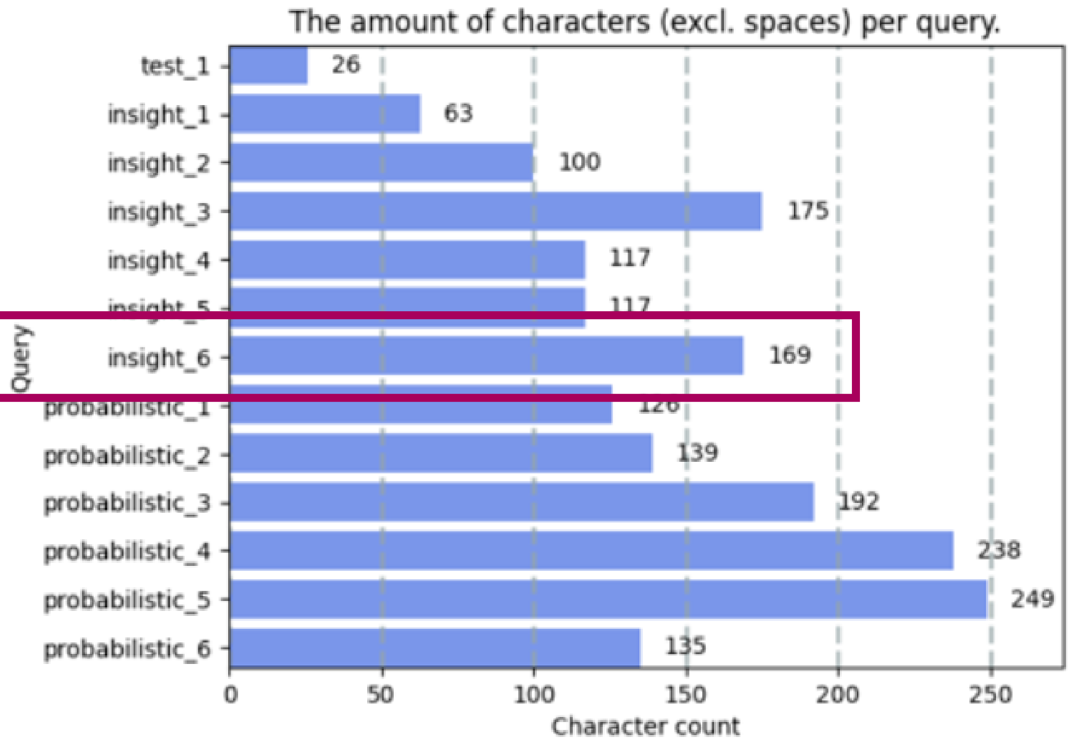


#	Native	Possible	#	Native	Possible	**Functionality**
1	[ ]	[ ]	1	[X]	[ ]	Support of most recent deterministic DBMS queries
2	[ ]	[ ]	2	[X]	[ ]	Offering a compact representation of the present uncertainty
3	[X]	[ ]	3	[X]	[ ]	Get the probability of an offer
4	[X]	[ ]	4	[x]	[ ]	Get the probability of a composed result
5	[X]	[ ]	5	[X]	[ ]	Apply aggregate functions on probabilities
6	[X]	[ ]	6	[X]	[ ]	Filtering on probability
7	[X]	[ ]	7	[ ]	[X]	Get the expected count
8	[X]	[ ]	8	[ ]	[X]	Get the expected sum
9	[ ]	[X]	9	[ ]	[X]	Get the most probable answer
10	[ ]	[X]	10	[X]	[ ]	Verify if a specific possible world exists
11	[ ]	[X]	11	[X]	[ ]	Verify if a record is certain
12	[ ]	[ ]	12	[ ]	[X]	Updating the uncertainty of an offer
13	[ ]	[ ]	13	[X]	[ ]	Repair the probability space after addition, update or deletion of offers
14	MayBMS		14	DuBio		Any anomalies discovered during benchmarking

2 – Brevity of the Query Dialect



MayBMS

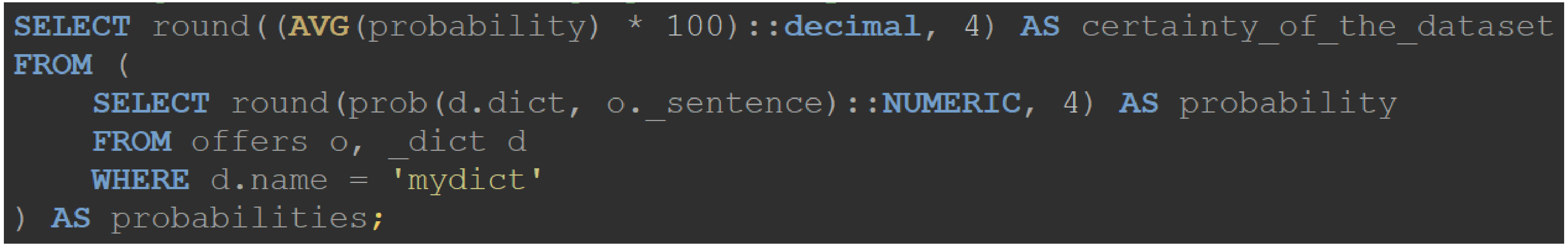


DuBio

UNIVERSITY OF TWENTE.

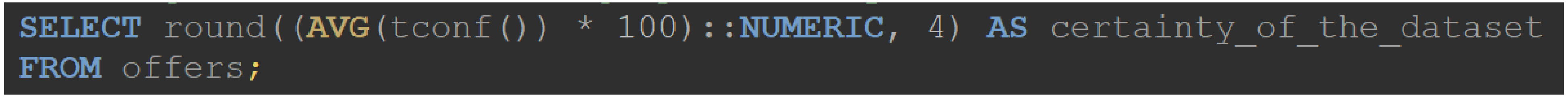
2 – Brevity of the Query Dialect

DuBio



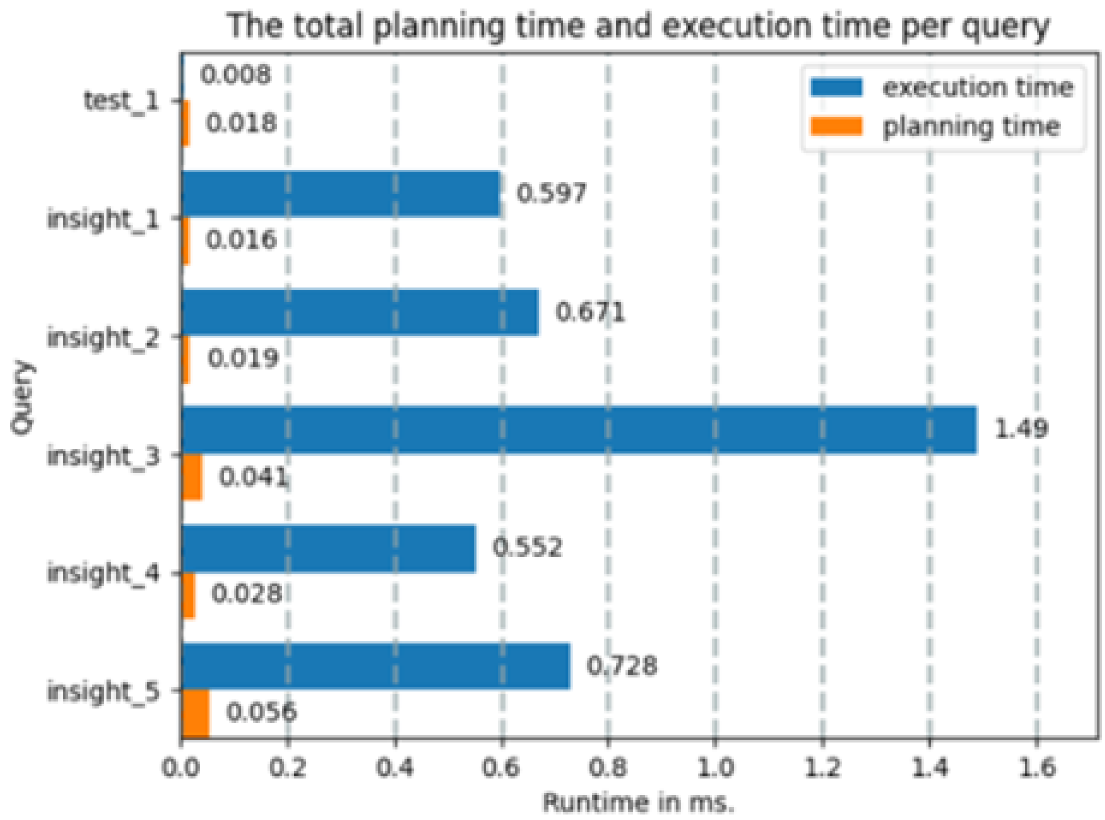
```
SELECT round((AVG(probability) * 100)::decimal, 4) AS certainty_of_the_dataset
FROM (
    SELECT round(prob(d.dict, o._sentence)::NUMERIC, 4) AS probability
    FROM offers o, _dict d
    WHERE d.name = 'mydict'
) AS probabilities;
```

MayBMS

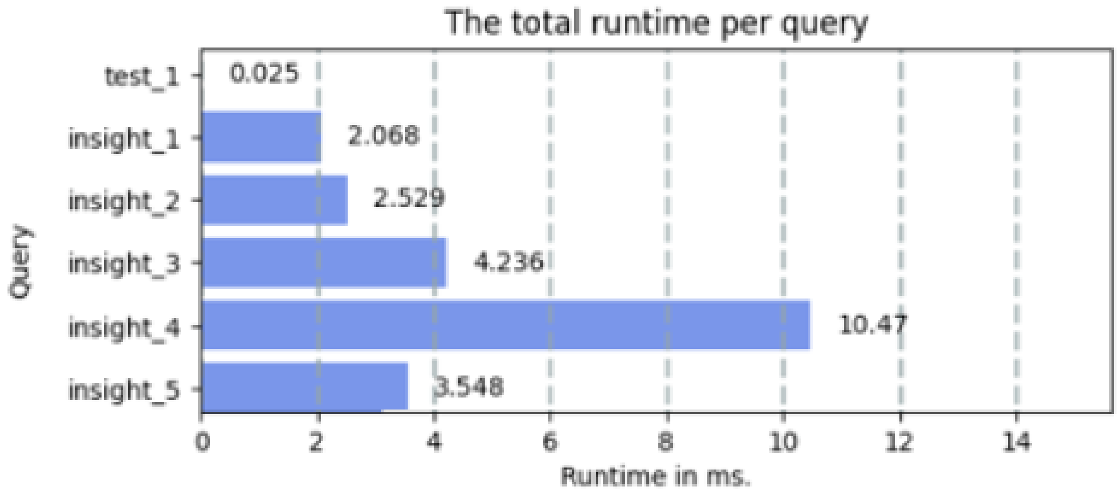


```
SELECT round((AVG(tconf()) * 100)::NUMERIC, 4) AS certainty_of_the_dataset
FROM offers;
```

3 – Runtime of Queries



DuBio



MayBMS

4 – Probabilistic Data Overhead

offers

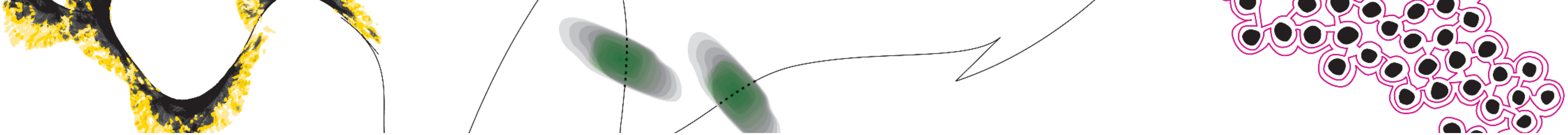
id	name	sales	_sentence
1	BMW	150	Bdd(a1=1, w1)
2	B.M.W.	127	Bdd(a1=2, w1, a2=1, w2)
3	Audi	194	Bdd(a2=2, w2)

_dict

name	dict
mydict	a1=1:0.3, a1=2:0.7, a2=1:0.4, a2=2:0.6, w1:0.5, w2:0.5

DuBio

UNIVERSITY OF TWENTE.



4 – Probabilistic Data Overhead

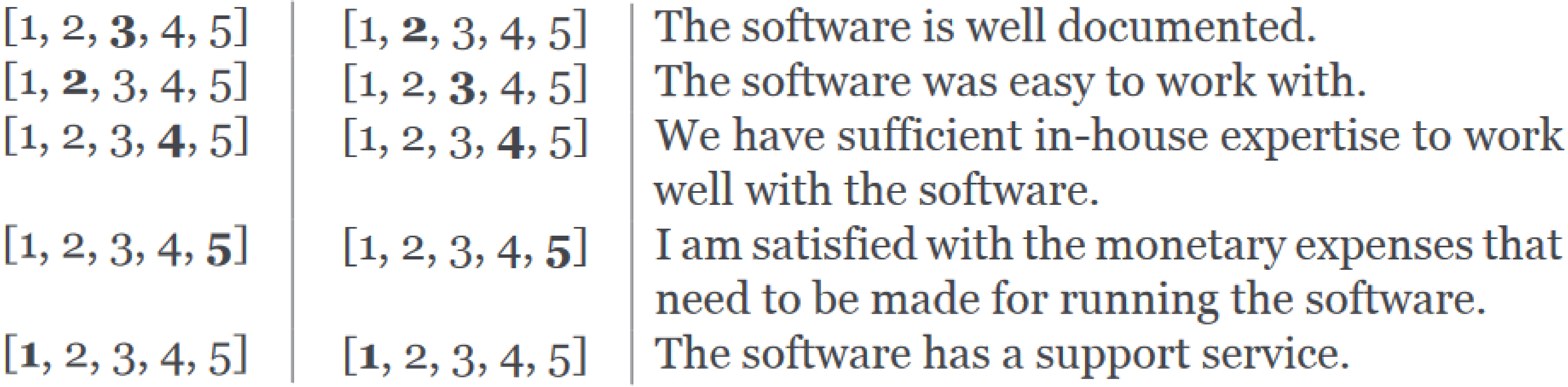
offers

id	name	sales	v0	d0	p0	v1	d1	p1
1	BMW	150	1	1	0.3	1	1	0.5
2	B.M.W.	127	1	2	0.7	1	1	0.5
2	B.M.W.	127	2	1	0.4	2	1	0.5
3	Audi	194	2	2	0.6	2	1	0.5

MayBMS

UNIVERSITY OF TWENTE.

5 – User Friendliness



[1, 2, **3**, 4, 5]	[1, **2**, 3, 4, 5]	The software is well documented.
[1, **2**, 3, 4, 5]	[1, 2, **3**, 4, 5]	The software was easy to work with.
[1, 2, 3, **4**, 5]	[1, 2, 3, **4**, 5]	We have sufficient in-house expertise to work well with the software.
[1, 2, 3, 4, **5**]	[1, 2, 3, 4, **5**]	I am satisfied with the monetary expenses that need to be made for running the software.
[**1**, 2, 3, 4, 5]	[**1**, 2, 3, 4, 5]	The software has a support service.

MayBMS DuBio

Conclusion

- Limitations identified
- Fulfils purpose

QuestionMark is ready to guide the future of databases!

Thank you!

Questions.

?MARK

UNIVERSITY
OF TWENTE.

Thank you!

Tea and cake time :D

Appendix: Dataset Selection

- The dataset is a good representation of the real world, both in the type of data and in size.
- The dataset contains enough uncertainty to be suitable for data integration purposes.
- The dataset should be freely available.
- The dataset should be versioned. Experiments conducted on the dataset should be reproducible.
- The dataset is suitable to be inserted in a relational database management system.

Appendix: The WDC Dataset

- Web Data Commons Product Data Corpus and Gold Standard for Large-Scale Product Matching (LSPM) version 2.0
- English subset
- 43 thousand websites
- 16 million product offers
- 10 million clusters
- Cluster sizes from 1 to 80 offers per cluster
- id, cluster_id, title, brand, category, description, price, identifiers, +2
- 2.8 GB compressed

UNIVERSITY
OF MANNHEIM

Appendix: Product Matching

- Data Preparation
- Search Space Reduction – using a Rule-Based blocking algorithm
    - Incrementally-Adaptive Sorted Neighborhood Blocking
    - Improved Suffix Array Blocking
- Attribute Value Matching – using a matching algorithm
    - Attribute-Based Entity Resolution
- Classification
    - Probabilistic clustering
    - Removing Inconsistent world graphs.
- Verification

UNIVERSITY OF TWENTE.

Appendix: The Dataset Generator

- 16 Python files
- 2013 lines of code

Appendix: The Probabilistic Benchmark

- 11 Python files
- 822 lines of code