Refining SQL Queries based on Why-Not Polynomials

Katerina Tzompanaki,

Nicole Bidoit, Melanie Herschel

Theory and Practice of Provenance

MacLean, 9/6/2016

Why-Not Questions / Provenance



Why-Not Questions / Provenance



- Trust assertion.
- Results verification.
- Data cleaning.
- Data integration.
- Data transformations debugging and repairing.

Why-Not Questions / Provenance



- Trust assertion.
- Results verification.
- Data cleaning.
- Data integration.
- ☑ **Data transformations debugging and repairing.**

Problem Description

Problem Setting



Input Query

Test

Missing results

no

Final Query

yes

New Query

Repair

Debug

Problem Setting



Input Query

Test

Missing results

no

Final Query

yes

New Query

Repair

Debug

Problem Setting



Input Query

Test

Missing results

no

Final Query



New Query

yes

Repair

Debug

Problem Setting



Input Query

Test

Missing results

no

Final Query

yes

New Query

Repair

Polynomial Explanations (TaPP14, CIKM15)

✔

Problem Setting



Input Query

Test

Missing results

no

Final Query

yes

New Query

Repair

?

Polynomial Explanations (TaPP14, CIKM15)

✔

Motivating Example

Author

Name	Birthday	Country
Olivier	1990	France
Mark	1985	UK
John	1980	USA
Susan	1975	USA
Frank	1980	Canada

Publication

Author	Page	Citations	Year
Olivier	7	150	2007
John	6	100	2010
John	7	150	2008
John	5	60	2010
Susan	6	190	2011
Susan	8	200	2012
Susan	20	50	2013
Susan	12	90	2015
Frank	12	100	2016



Countries of successful authors?

Motivating Example

Author

Name	Birthday	Country
Olivier	1990	France
Mark	1985	UK
John	1980	USA
Susan	1975	USA
Frank	1980	Canada

Publication

Author	Page	Citations	Year
Olivier	7	150	2007
John	6	100	2010
John	7	150	2008
John	5	60	2010
Susan	6	190	2011
Susan	8	200	2012
Susan	20	50	2013
Susan	12	90	2015
Frank	12	100	2016



Countries of successful authors?

Motivating Example

SQL Query

```
SELECT   A.Country
FROM     Author A, Publication P
WHERE    A.Birthday > 1978  AND
         P.Pages > 10       AND
         P.Citations ≥ 100  AND
         P.Year > 2006      AND
         A.Name = P.Author
```



Motivating Example

Author

Name	Birthday	Country
Olivier	1990	France
Mark	1985	UK
John	1980	USA
Susan	1975	USA
Frank	1980	Canada

Publication

Author	Page	Citations	Year
Olivier	7	150	2007
John	6	100	2010
John	7	150	2008
John	5	60	2010
Susan	6	190	2011
Susan	8	200	2012
Susan	20	50	2013
Susan	12	90	2015
Frank	12	100	2016



Countries of successful authors?

Result

Country
Canada

Why-Not question

Author

Name	Birthday	Country
Olivier	1990	France
Mark	1985	UK
John	1980	USA
Susan	1975	USA
Frank	1980	Canada

Publication

Author	Page	Citations	Year
Olivier	7	150	2007
John	6	100	2010
John	7	150	2008
John	5	60	2010
Susan	6	190	2011
Susan	8	200	2012
Susan	20	50	2013
Susan	12	90	2015
Frank	12	100	2016



Why not USA?

Result

Country
Canada

Why-Not question

Author

Name	Birthday	Country
Olivier	1990	France
Mark	1985	UK
John	1980	USA
Susan	1975	USA
Frank	1980	Canada

Publication

Author	Page	Citations	Year
Olivier	7	150	2007
John	6	100	2010
John	7	150	2008
John	5	60	2010
Susan	6	190	2011
Susan	8	200	2012
Susan	20	50	2013
Susan	12	90	2015
Frank	12	100	2016



Why not USA?

Result

Country
Canada
USA

Query-based explanation

Author

Name	Birthday	Country
Olivier	1990	France
Mark	1985	UK
John	1980	USA
Susan	1975	USA
Frank	1980	Canada

Publication

Author	Page	Citations	Year
Olivier	7	150	2007
John	6	100	2010
John	7	150	2008
John	5	60	2010
Susan	6	190	2011
Susan	8	200	2012
Susan	20	50	2013
Susan	12	90	2015
Frank	12	100	2016



```
SELECT   A.Country
FROM     Author A,
         Publication P
WHERE    A.Birthday > 1978   AND
         P.Pages > 10        AND
         P.Citations ≥ 100   AND
         P.Year > 2006       AND
         A.Name = P.Author
```

Result

Country
Canada
USA

Why-Not Polynomial [TaPP14, CIKM15]

$$\bowtie_{\text{Name=Author}} + 2\sigma_{\text{Pages>10}} + 2\sigma_{\text{Bday>78}}\sigma_{\text{Pages>10}} + 2\sigma_{\text{Bday>78}}\sigma_{\text{Citations}\geq 100} + \ldots$$

$$+\sigma_{\text{Pages>10}}\sigma_{\text{Bday>78}}\sigma_{\text{Citations}\geq 100}\bowtie_{\text{Name=Author}}$$

Why-Not Polynomial [TaPP14, CIKM15]

$$\bowtie_{\text{Name=Author}} + 2\sigma_{\text{Pages>10}} + 2\sigma_{\text{Bday>78}}\sigma_{\text{Pages>10}} + 2\sigma_{\text{Bday>78}}\sigma_{\text{Citations}\geq 100} + \ldots$$

$$+\sigma_{\text{Pages>10}}\sigma_{\text{Bday>78}}\sigma_{\text{Citations}\geq 100}\bowtie_{\text{Name=Author}}$$

All correct query-based explanations

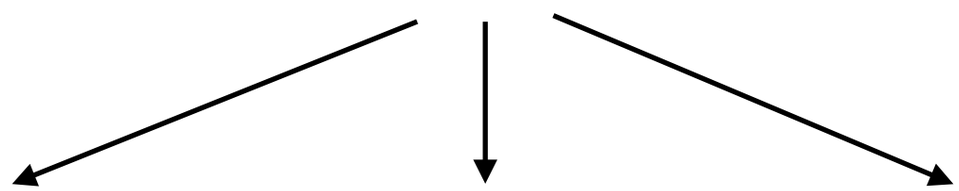


All erroneous query condition combinations

Why-Not Polynomial [TaPP14, CIKM15]

$$\bowtie_{\text{Name=Author}} + 2\sigma_{\text{Pages>10}} + 2\sigma_{\text{Bday>78}}\sigma_{\text{Pages>10}} + 2\sigma_{\text{Bday>78}}\sigma_{\text{Citations}\geq 100} + \ldots$$

$$+\sigma_{\text{Pages>10}}\sigma_{\text{Bday>78}}\sigma_{\text{Citations}\geq 100}\bowtie_{\text{Name=Author}}$$



selections

$\sigma_{\text{Pages>10}}$

$\sigma_{\text{Bday>78}}\sigma_{\text{Pages>10}}$

$\sigma_{\text{Bday>78}}\sigma_{\text{Citations}\geq 100}$

$\sigma_{\text{Pages>10}}\sigma_{\text{Citations}\geq 100}$

joins

$\bowtie_{\text{Name=Author}}$

selections & joins

$\sigma_{\text{Pages>10}}\bowtie_{\text{Name=Author}}$

$\sigma_{\text{Bday>78}}\bowtie_{\text{Name=Author}}$

$\sigma_{\text{Citations}\geq 100}\bowtie_{\text{Name=Author}}$

$\sigma_{\text{Bday>78}}\sigma_{\text{Pages>10}}\bowtie_{\text{Name=Author}}$

$\sigma_{\text{Bday>78}}\sigma_{\text{Pages>10}}\sigma_{\text{Citations}\geq 100}\bowtie_{\text{Name=Author}}$

Query Refinement Problem

Author

Name	Birthday	Country
Olivier	1990	France
Mark	1985	UK
John	1980	USA
Susan	1975	USA
Frank	1980	Canada

Publication

Author	Page	Citations	Year
Olivier	7	150	2007
John	6	100	2010
John	7	150	2008
John	5	60	2010
Susan	6	190	2011
Susan	8	200	2012
Susan	20	50	2013
Susan	12	90	2015
Frank	12	100	2016



```
SELECT   A.Country
FROM     Author A,
         Publication P
WHERE    ?
```

Result

Country
Canada
USA

Query Refinement Problem



- **Vast** number of possible query refinements!
- How do I (efficiently) **prune** this space and still find the ***good*** ones?

Query Refinement Problem



- **Vast** number of possible query refinements!
- How do I (efficiently) **prune** this space and still find the ***good*** ones?



✓ Why-Not polynomial tells us **exactly** which condition combinations to **change**.

✓ Selections: Use values from **chosen** tuples eliminated by the combination.
Efficient tuple selection based on the explanation **schema**.

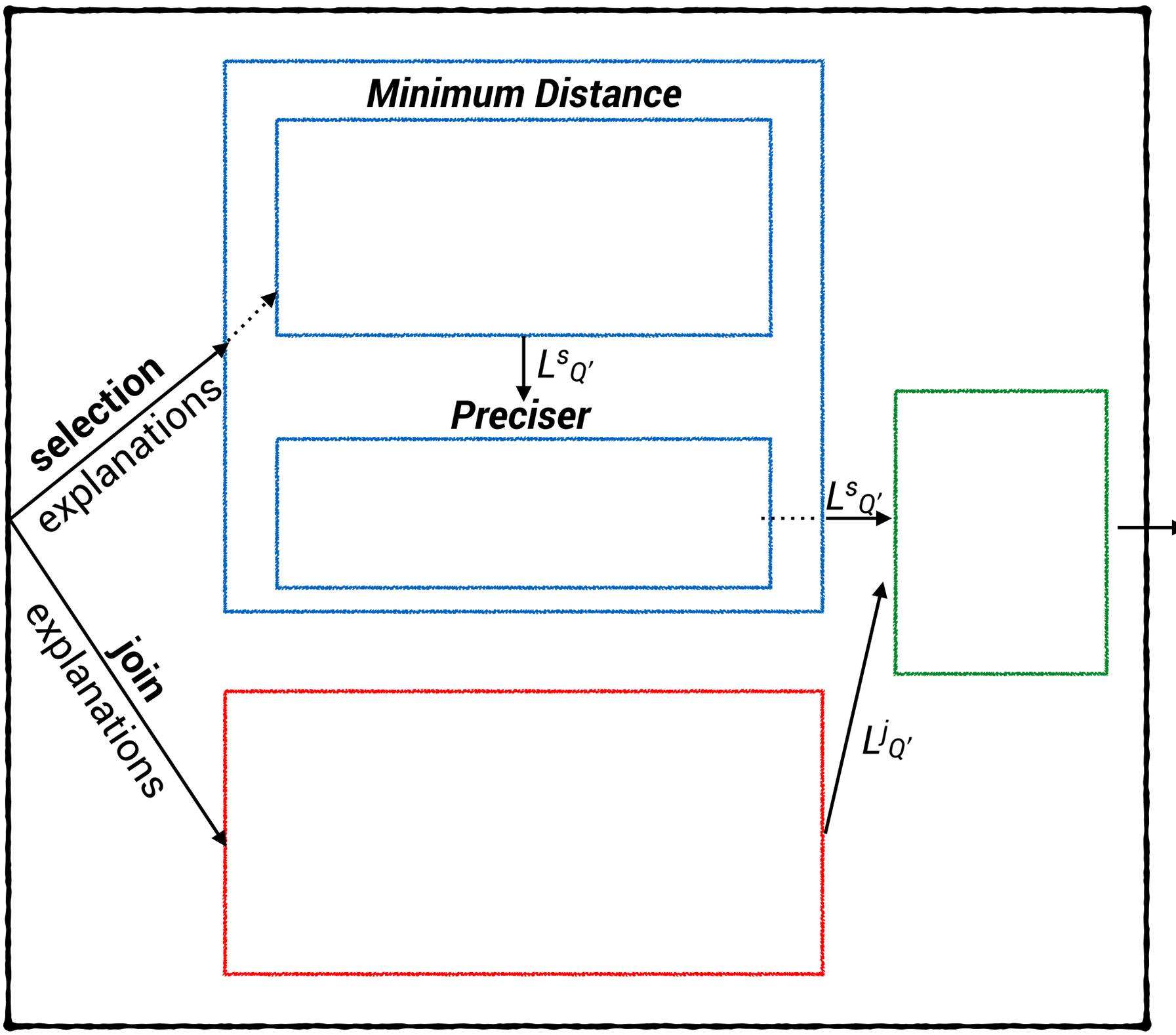
✓ Joins: Use **outer joins**.

Refining queries with **FixTed**

FixTed

Input

- Query
- Why-Not Polynomial
- Database



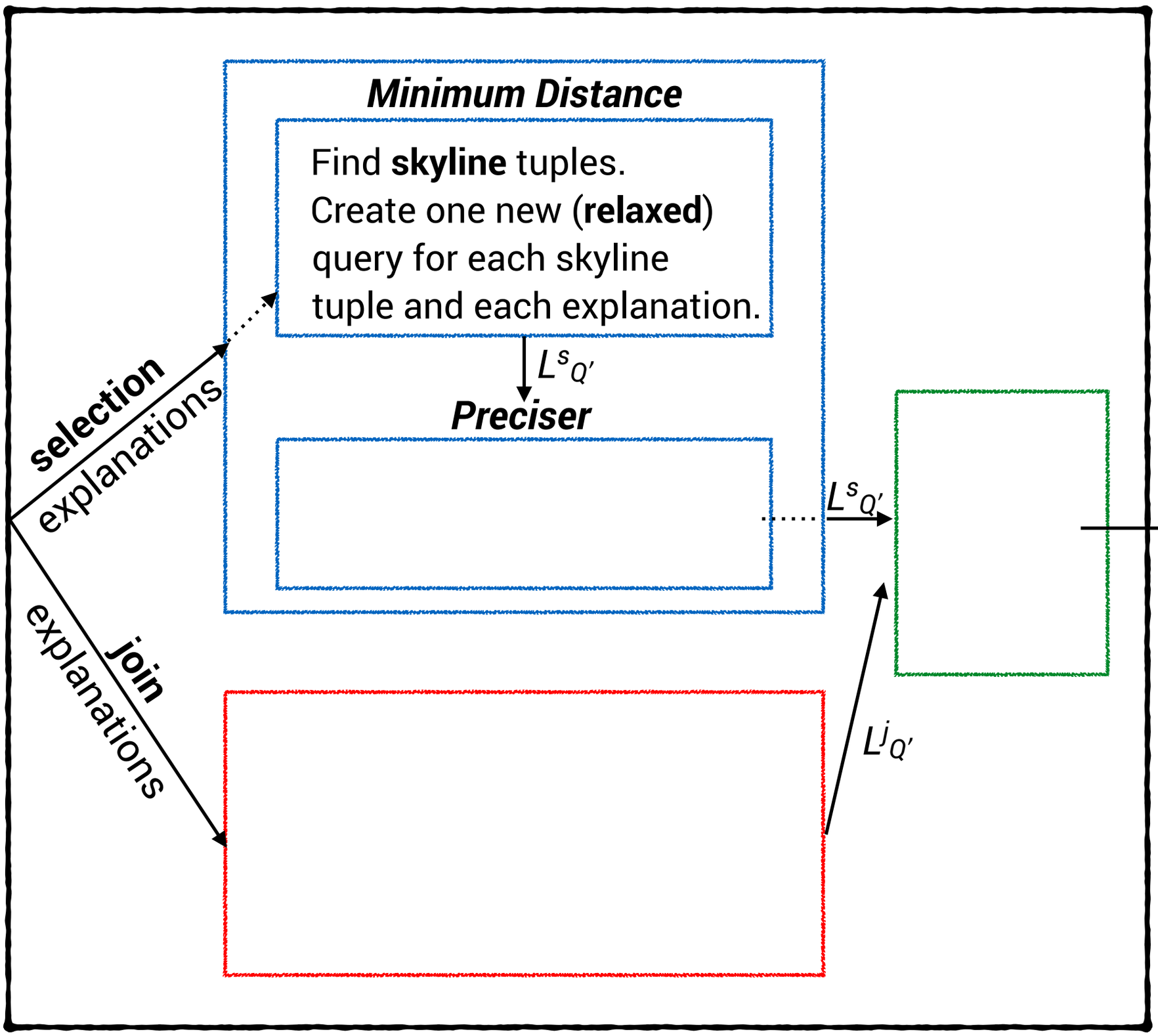
Output

List of ordered refined queries.

FixTed

Input

- Query
- Why-Not Polynomial
- Database



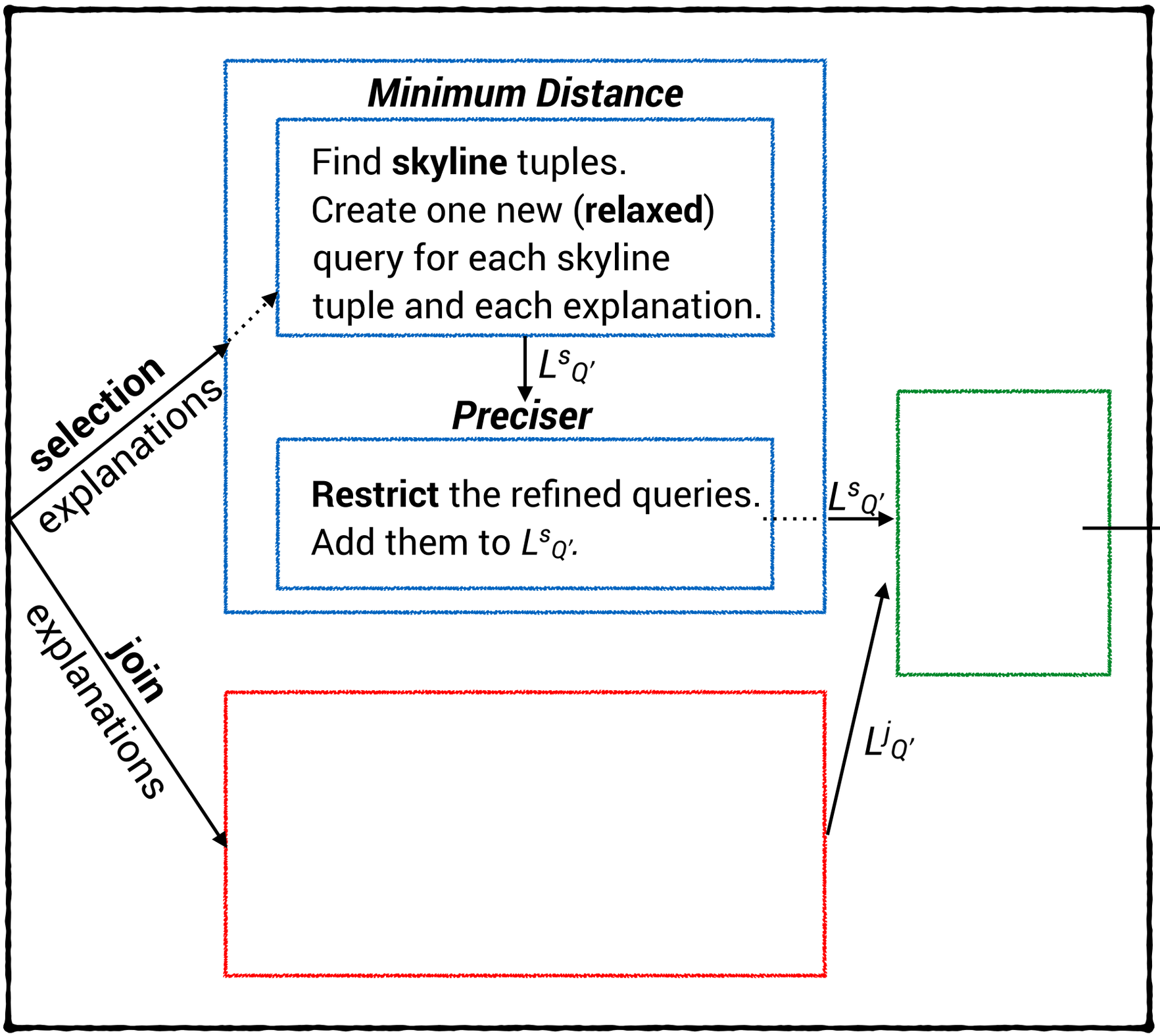
Output

List of ordered refined queries.

FixTed

Input

- Query
- Why-Not Polynomial
- Database



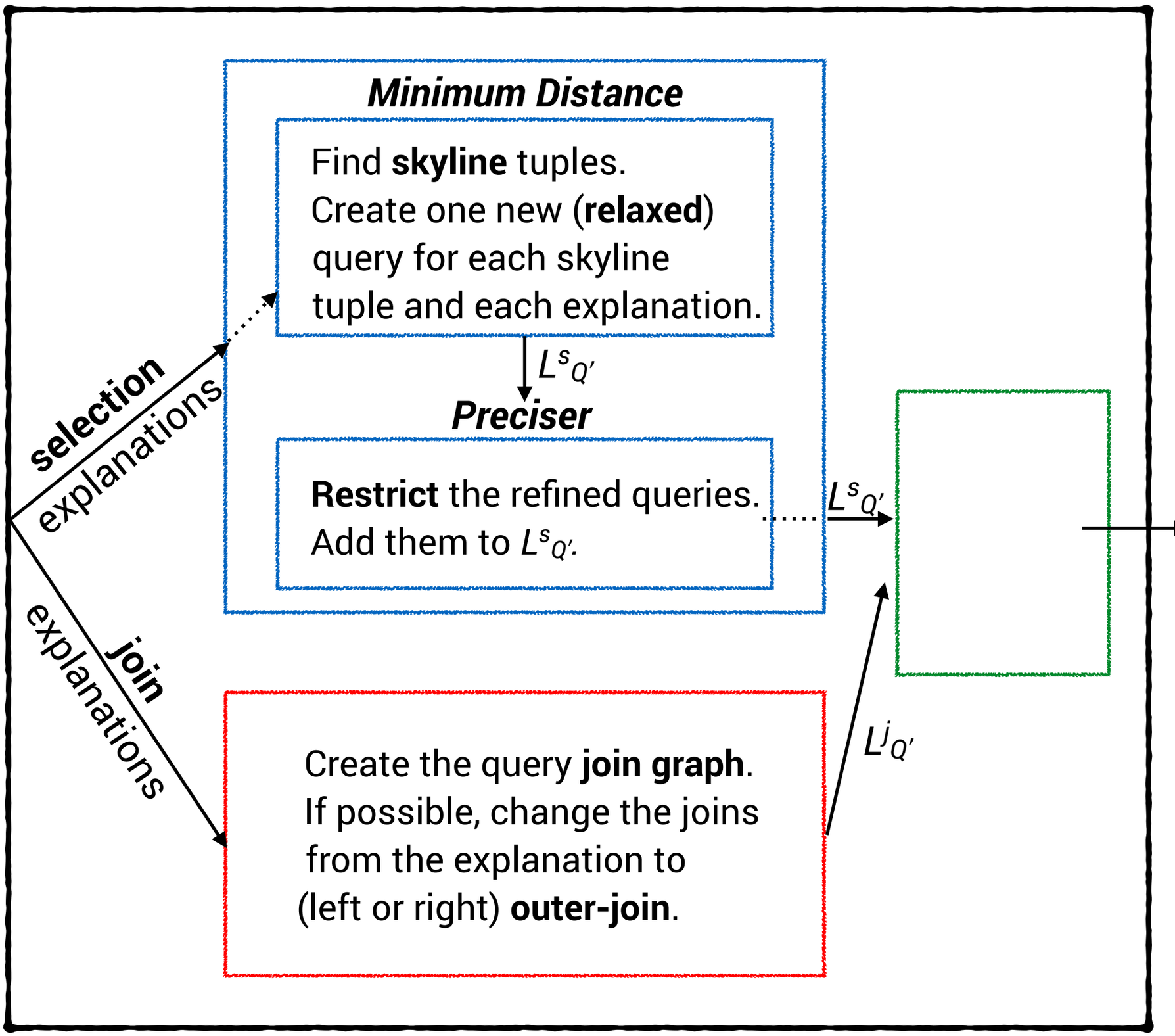
Output

List of ordered refined queries.

FixTed

Input

- Query
- Why-Not Polynomial
- Database



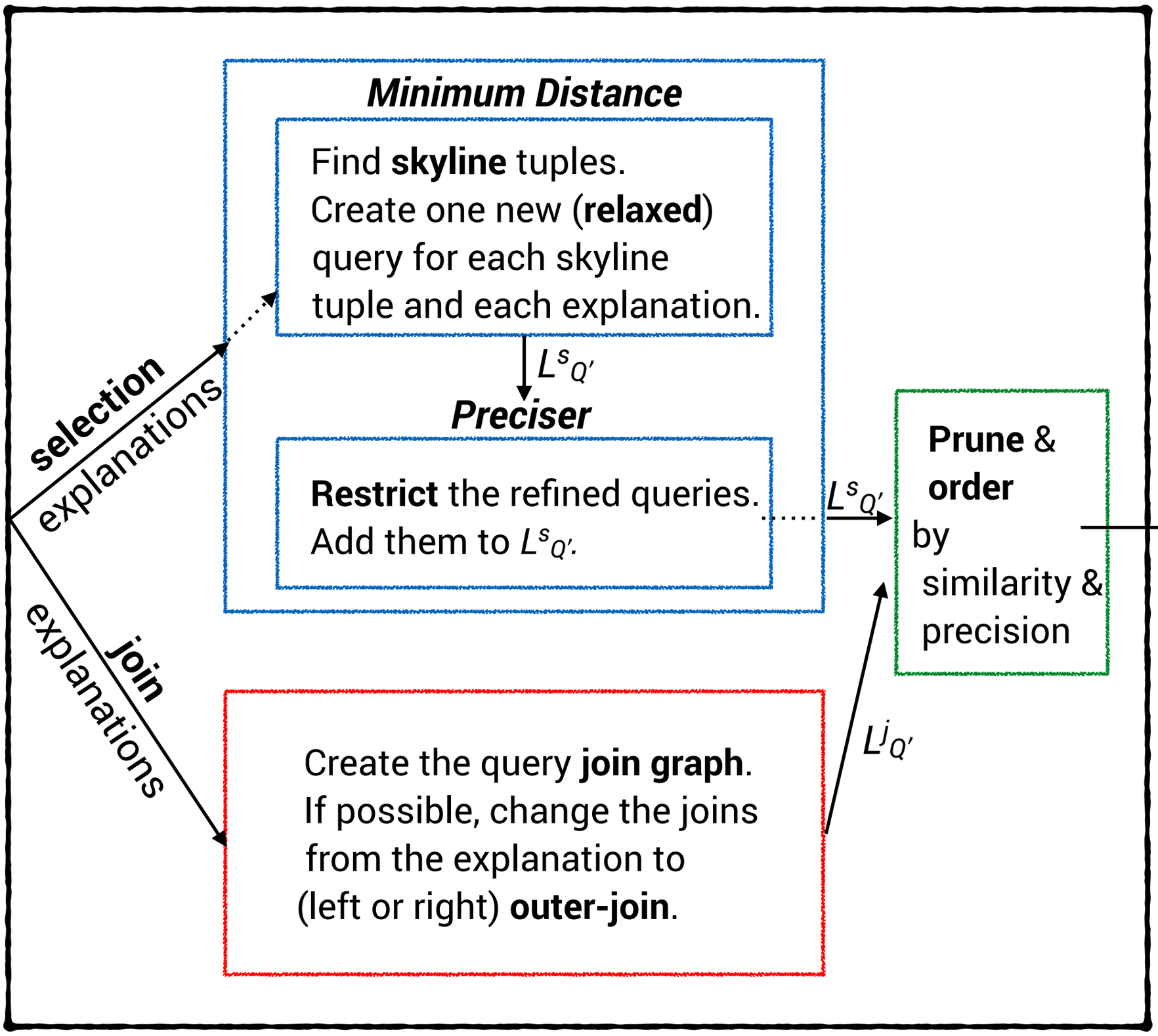
Output

List of ordered refined queries.

FixTed

Input

- Query
- Why-Not Polynomial
- Database



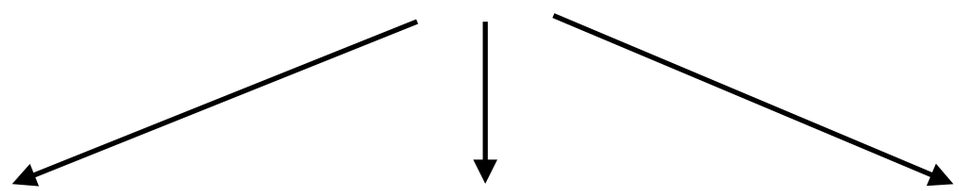
Output

List of ordered refined queries.

Selection Repair

$$\bowtie_{\text{Name=Author}} + 2\sigma_{\text{Pages>10}} + 2\sigma_{\text{Bday>78}}\sigma_{\text{Pages>10}} + 2\sigma_{\text{Bday>78}}\sigma_{\text{Citations}\geq 100} + \ldots$$

$$+\sigma_{\text{Pages>10}}\sigma_{\text{Bday>78}}\sigma_{\text{Citations}\geq 100}\bowtie_{\text{Name=Author}}$$



selections

$\sigma_{\text{Pages>10}}$

$\sigma_{\text{Bday>78}}\sigma_{\text{Pages>10}}$

$\sigma_{\text{Bday>78}}\sigma_{\text{Citations}\geq 100}$

$\sigma_{\text{Pages>10}}\sigma_{\text{Citations}\geq 100}$

joins

$\bowtie_{\text{Name=Author}}$

selections & joins

$\sigma_{\text{Pages>10}}\bowtie_{\text{Name=Author}}$

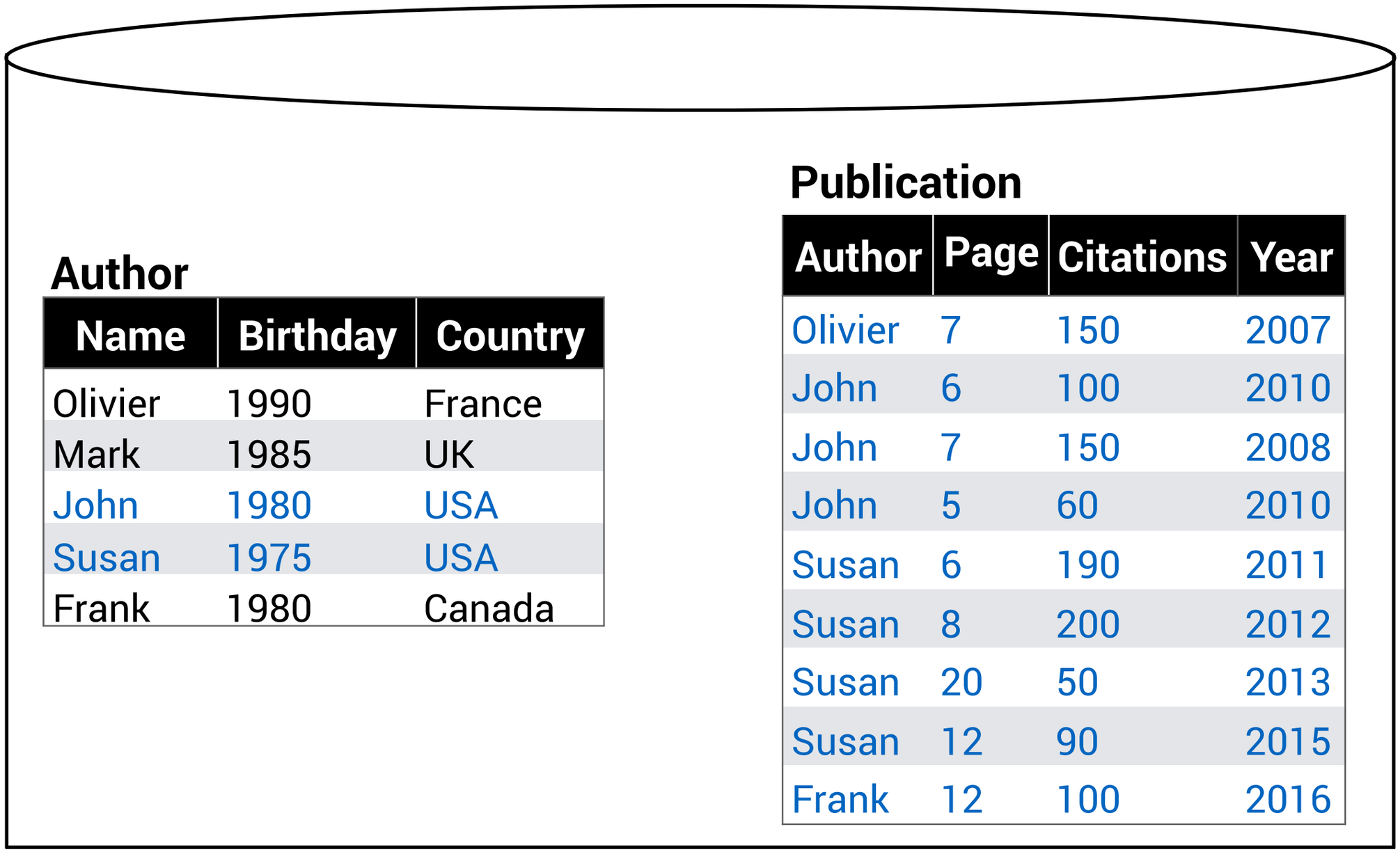
$\sigma_{\text{Bday>78}}\bowtie_{\text{Name=Author}}$

$\sigma_{\text{Citations}\geq 100}\bowtie_{\text{Name=Author}}$

$\sigma_{\text{Bday>78}}\sigma_{\text{Pages>10}}\bowtie_{\text{Name=Author}}$

$\sigma_{\text{Bday>78}}\sigma_{\text{Pages>10}}\sigma_{\text{Citations}\geq 100}\bowtie_{\text{Name=Author}}$

Selection Repair



Author

Name	Birthday	Country
Olivier	1990	France
Mark	1985	UK
John	1980	USA
Susan	1975	USA
Frank	1980	Canada

Publication

Author	Page	Citations	Year
Olivier	7	150	2007
John	6	100	2010
John	7	150	2008
John	5	60	2010
Susan	6	190	2011
Susan	8	200	2012
Susan	20	50	2013
Susan	12	90	2015
Frank	12	100	2016

Compatible tuples

Selection Repair

Author

Name	Birthday	Country
Olivier	1990	France
Mark	1985	UK
John	1980	USA
Susan	1975	USA
Frank	1980	Canada

(*join*)

Publication

Author	Page	Citations	Year
Olivier	7	150	2007
John	6	100	2010
John	7	150	2008
John	5	60	2010
Susan	6	190	2011
Susan	8	200	2012
Susan	20	50	2013
Susan	12	90	2015
Frank	12	100	2016

Selection Repair

Author

Name	Birthday	Country
Olivier	1990	France
Mark	1985	UK
John	1980	USA
Susan	1975	USA
Frank	1980	Canada

(*join*)

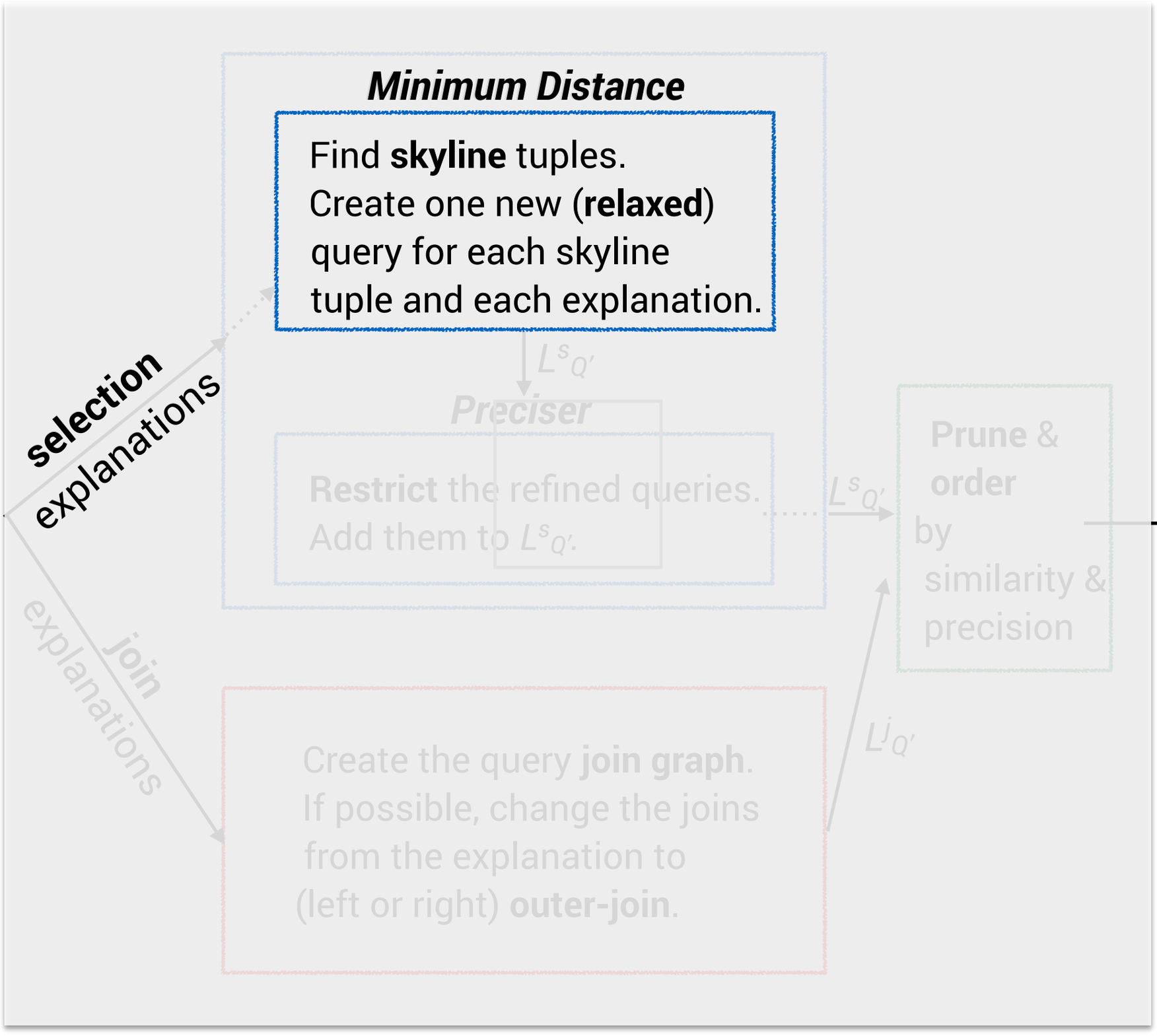
Publication

Author	Page	Citations	Year	*Explanations*
Olivier	7	150	2007	
John	6	100	2010	$\sigma_{Pages>10}$
John	7	150	2008	$\sigma_{Pages>10}$
John	5	60	2010	$\sigma_{Pages>10}\sigma_{Citations\geq 100}$
Susan	6	190	2011	$\sigma_{Bday>78}\sigma_{Pages>10}$
Susan	8	200	2012	$\sigma_{Bday>78}\sigma_{Pages>10}$
Susan	20	50	2013	$\sigma_{Bday>78}\sigma_{Citations\geq 100}$
Susan	12	90	2015	$\sigma_{Bday>78}\sigma_{Citations\geq 100}$
Frank	12	100	2016	

FixTed

Input

- Query
- Why-Not Polynomial
- Database

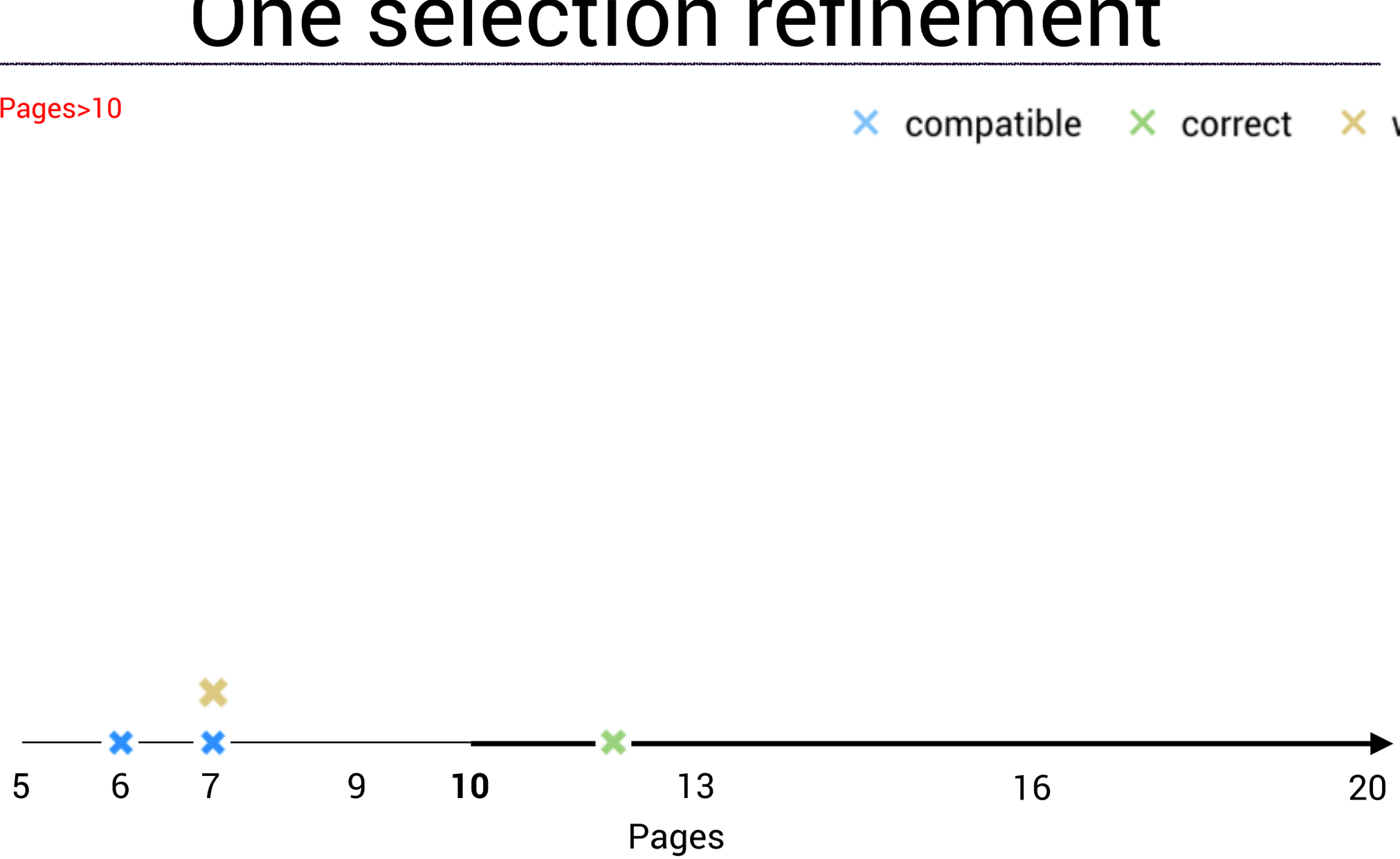


Output

List of ordered refined queries.

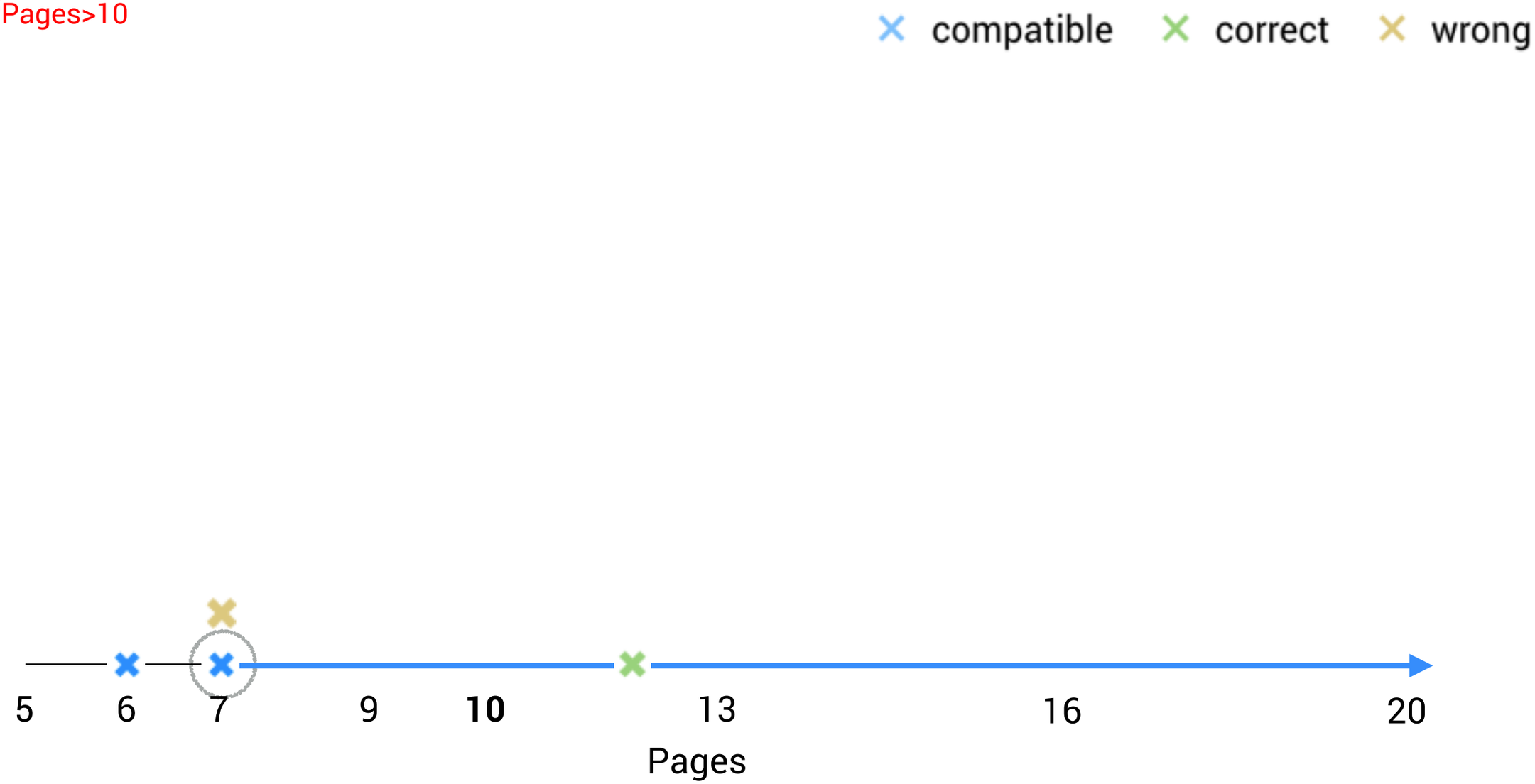
One selection refinement

$\sigma_{Pages>10}$



One selection refinement

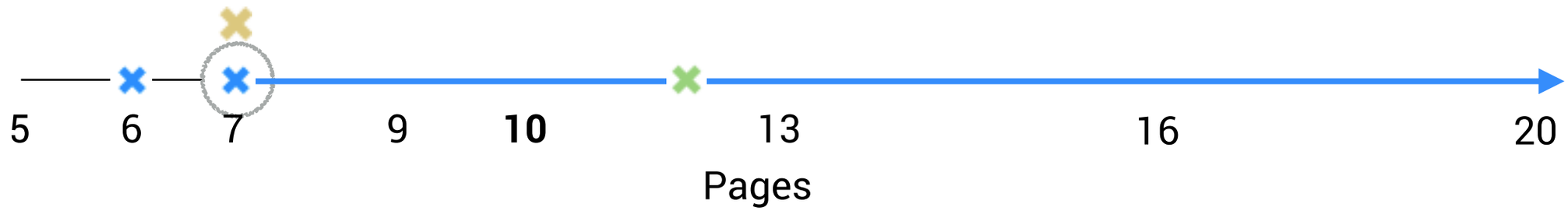
$\sigma_{Pages>10}$



One selection refinement

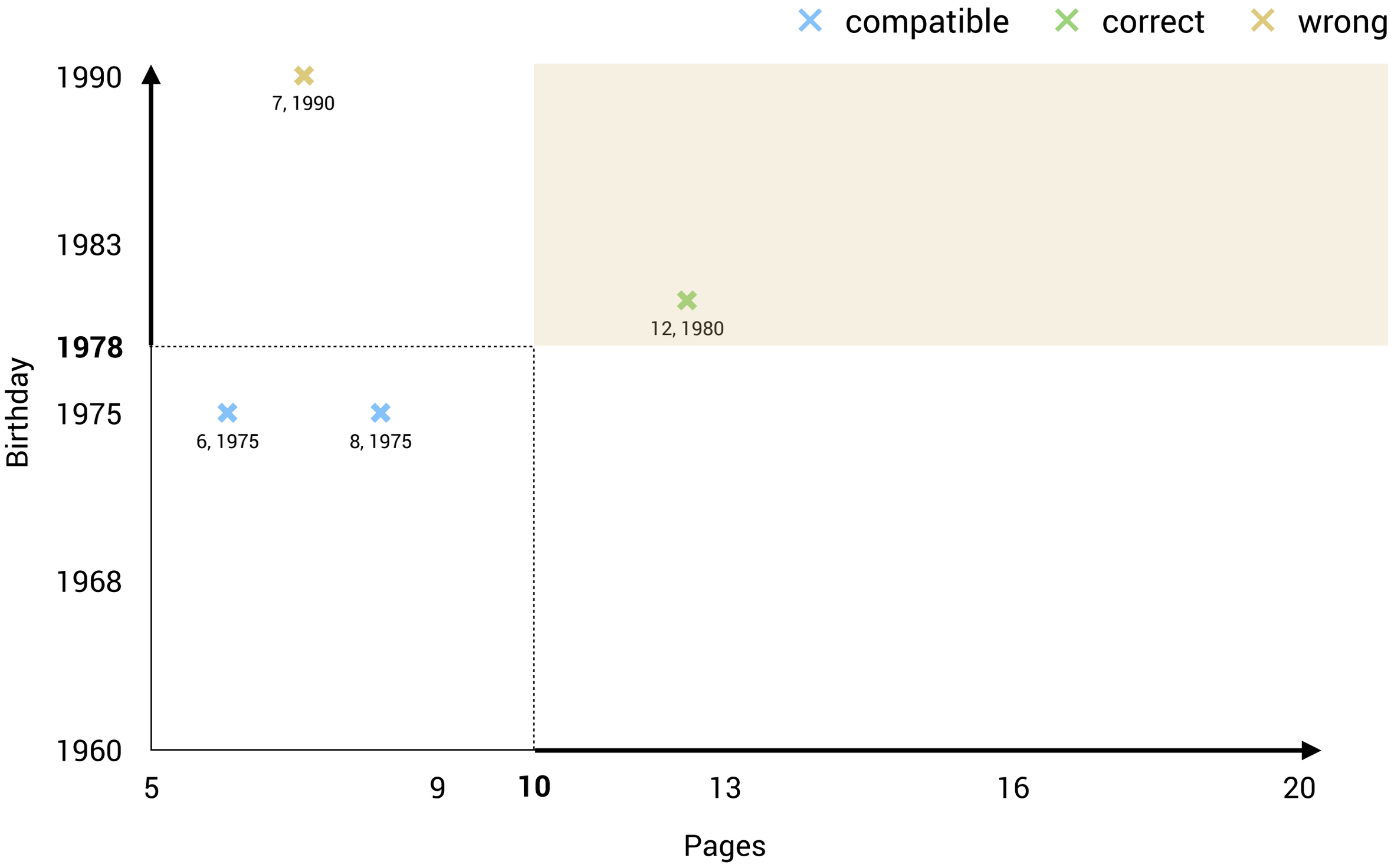
Refine with 7

$\sigma_{Pages>10}$ ⟹ $\sigma_{Pages\geq 7}$



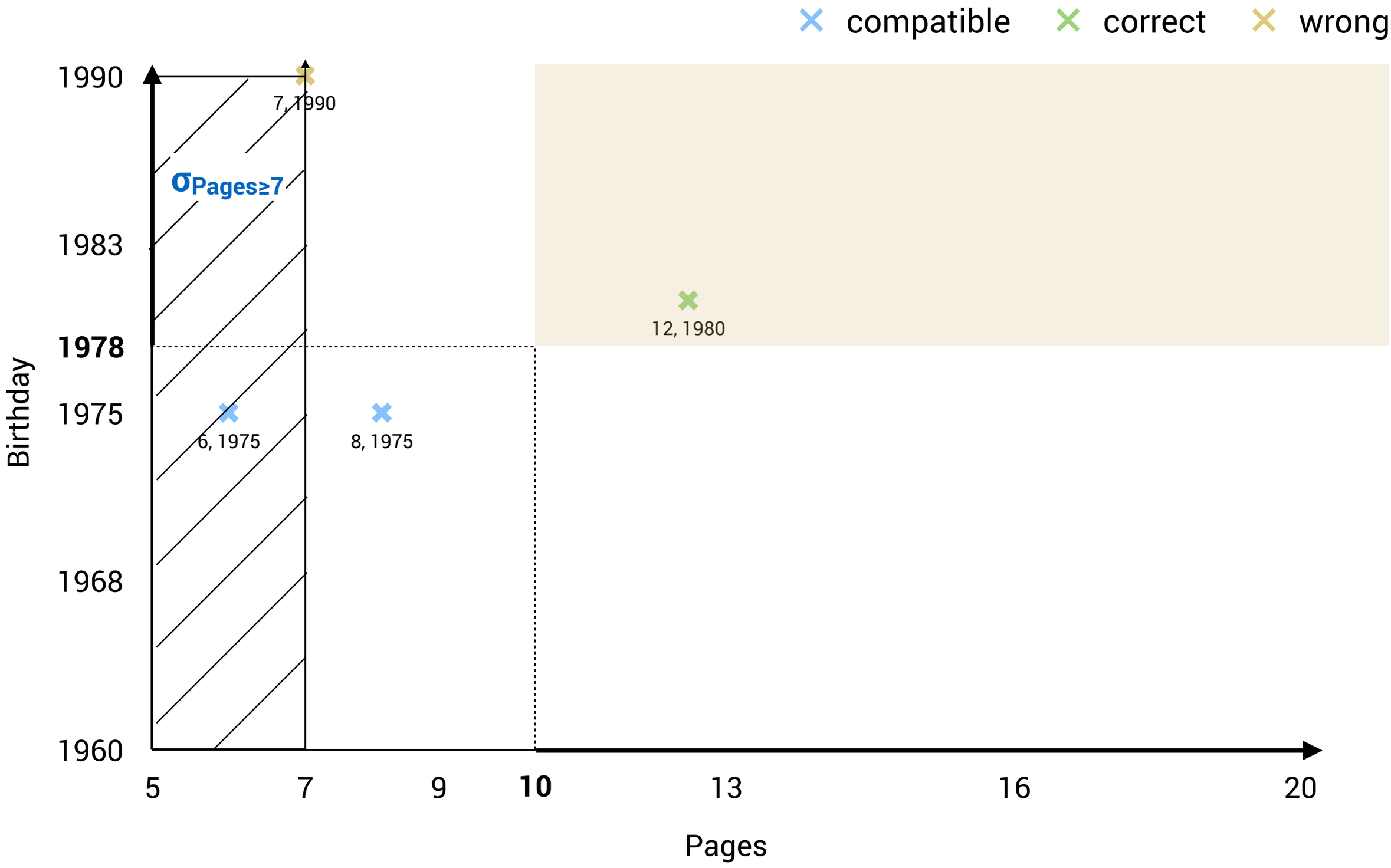
Two selections refinement

$\sigma_{Bday>78}\sigma_{Pages>10}$



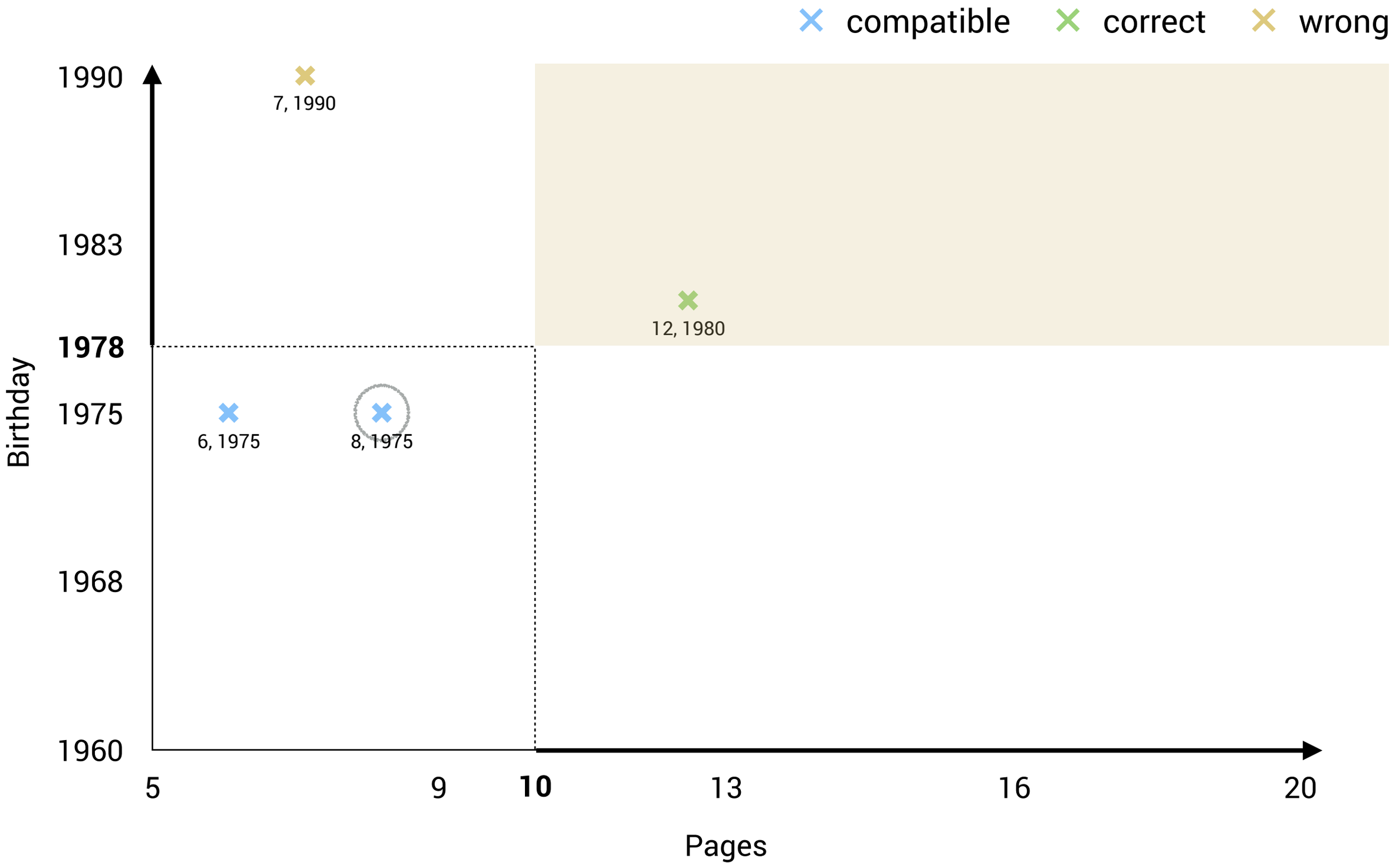
Two selections refinement

$\sigma_{Bday>78}\sigma_{Pages>10}$

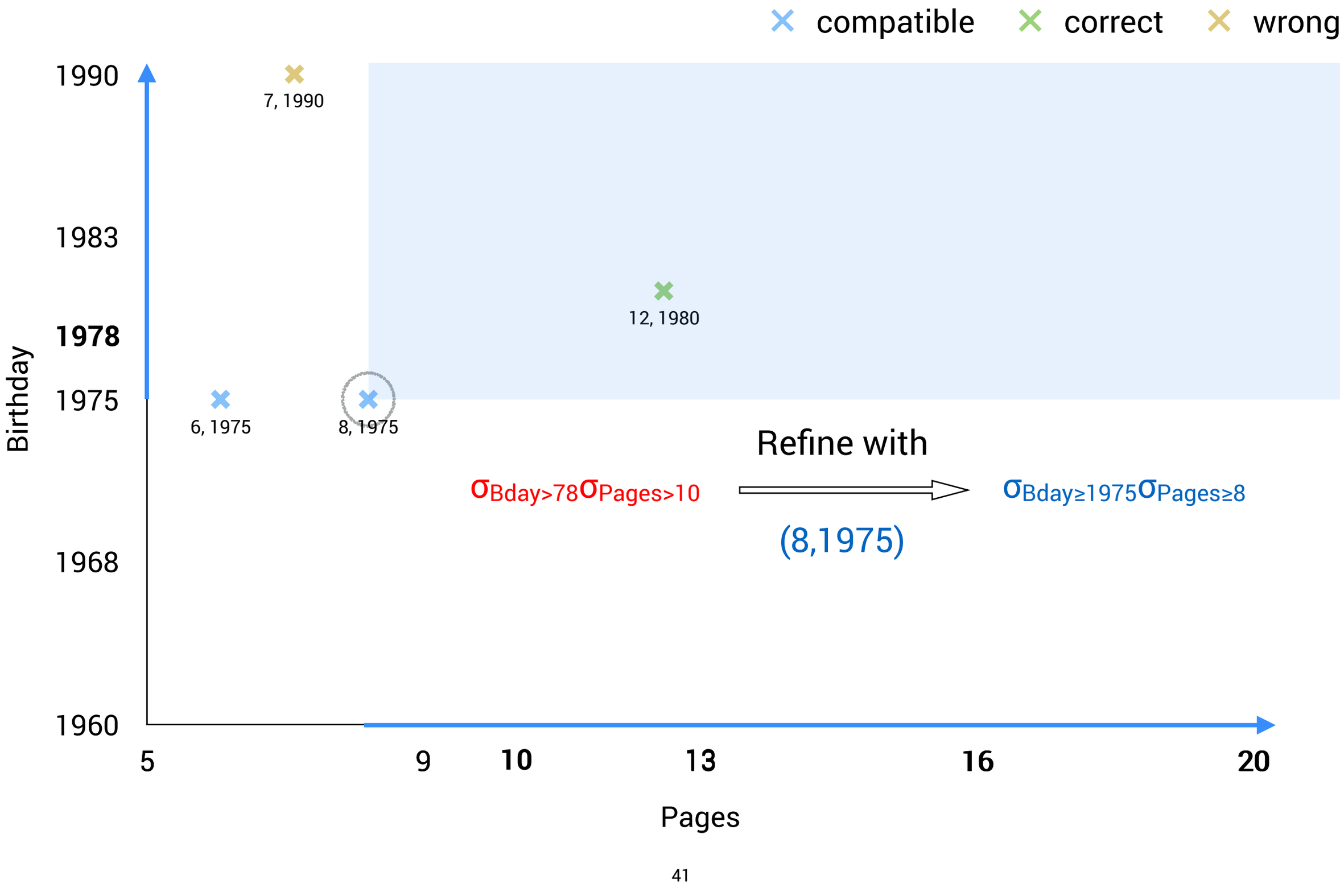


Two selections refinement

$\sigma_{Bday>78}\sigma_{Pages>10}$



Two selections refinement



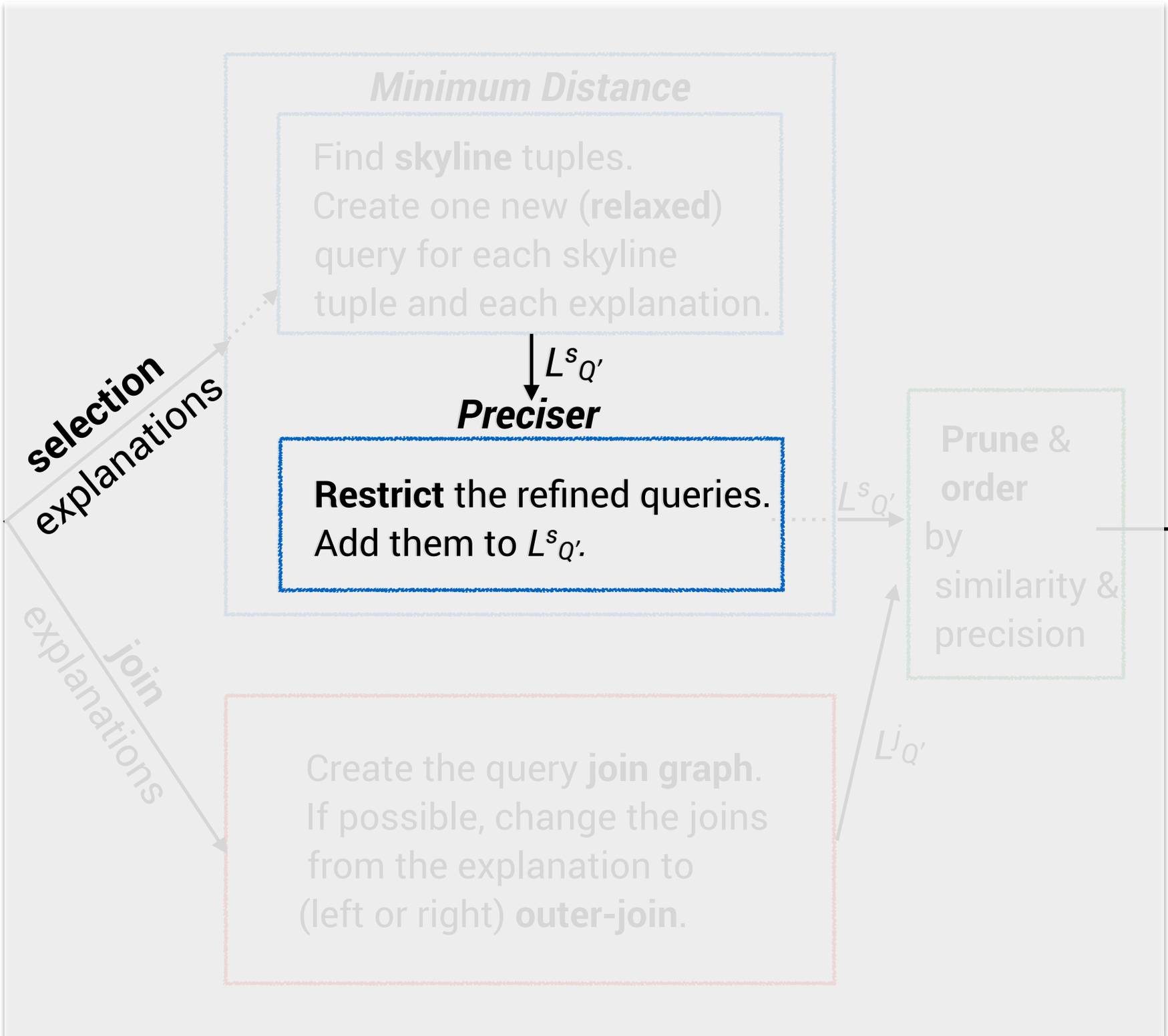
Minimum Distance Refinements

Refinement (L^s_Q)	New conditions
Q_1	$\sigma_{Pages \geq 7}$
Q_2	$\sigma_{Bday \geq 1975}\sigma_{Pages \geq 8}$
Q_3	$\sigma_{Bday \geq 1975}\sigma_{Citations \geq 90}$

FixTed

Input

- Query
- Why-Not Polynomial
- Database



Output

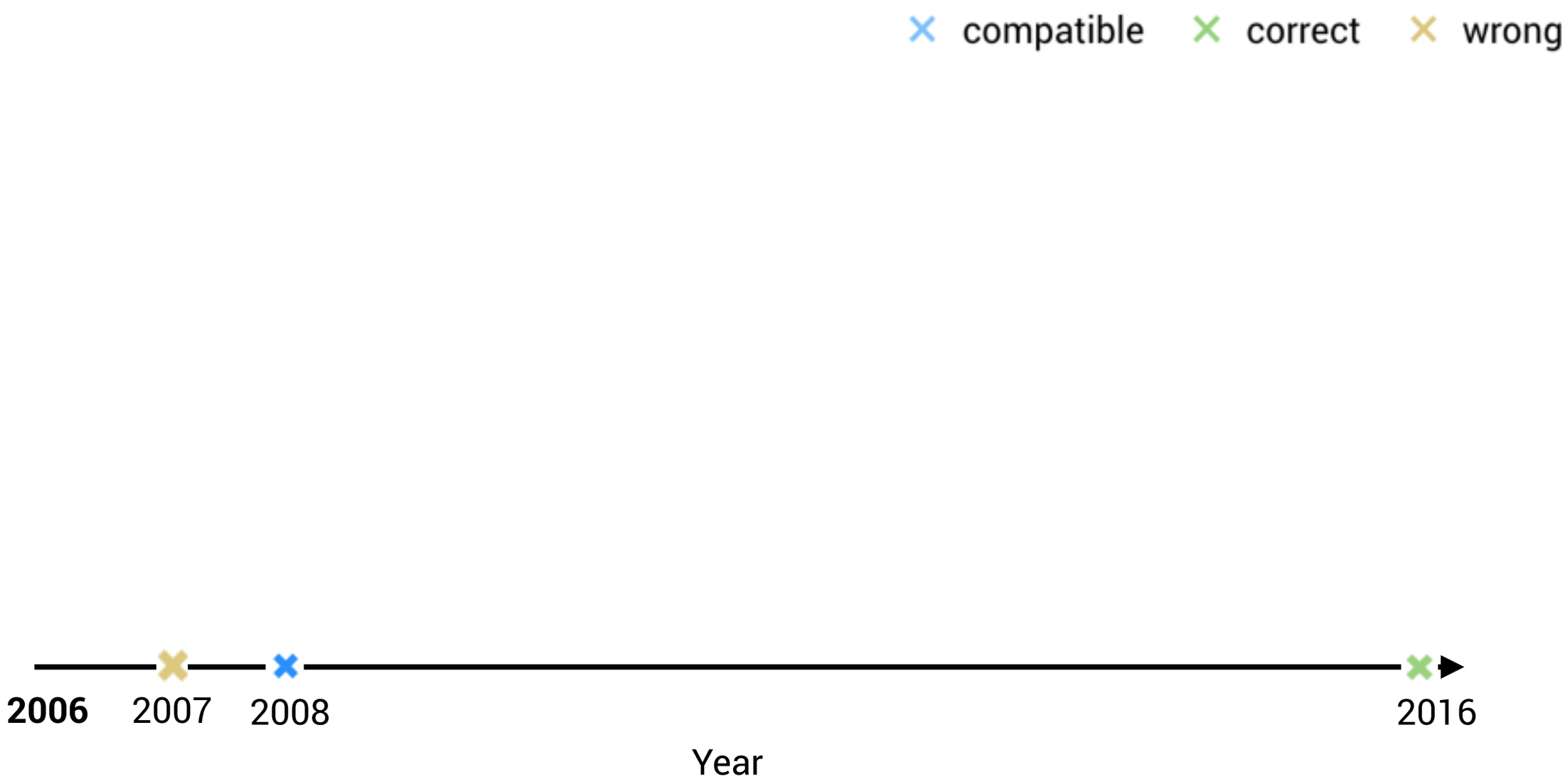
List of ordered refined queries.

Preciser

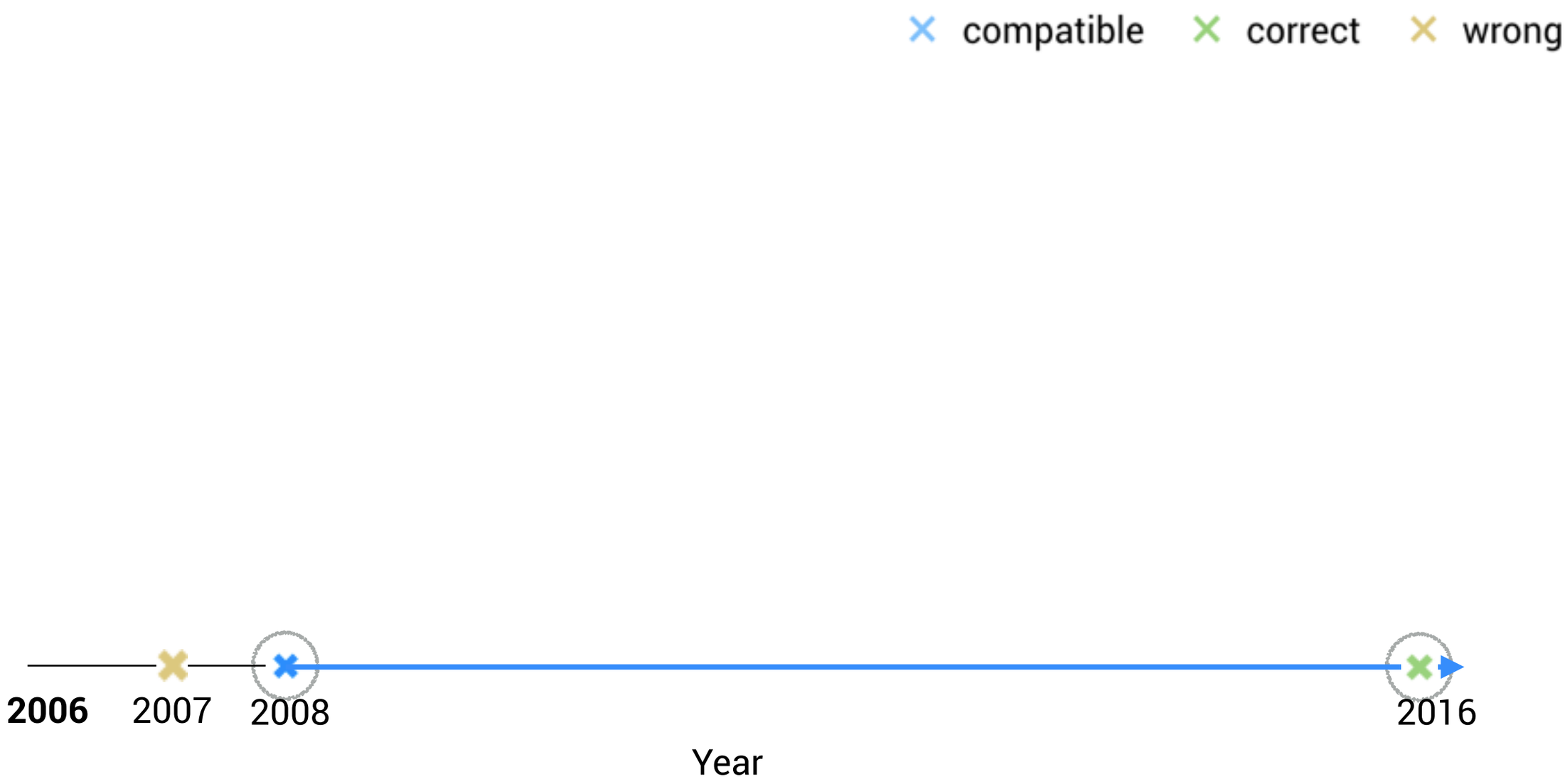


Refinement (L^s_Q)	New conditions	False positive tuples
Q_1	$\sigma_{Pages \geq 7}$	1
Q_2	$\sigma_{Bday \geq 1975}\sigma_{Pages \geq 8}$	0
Q_3	$\sigma_{Bday \geq 1975}\sigma_{Citations \geq 90}$	0

Preciser



Preciser



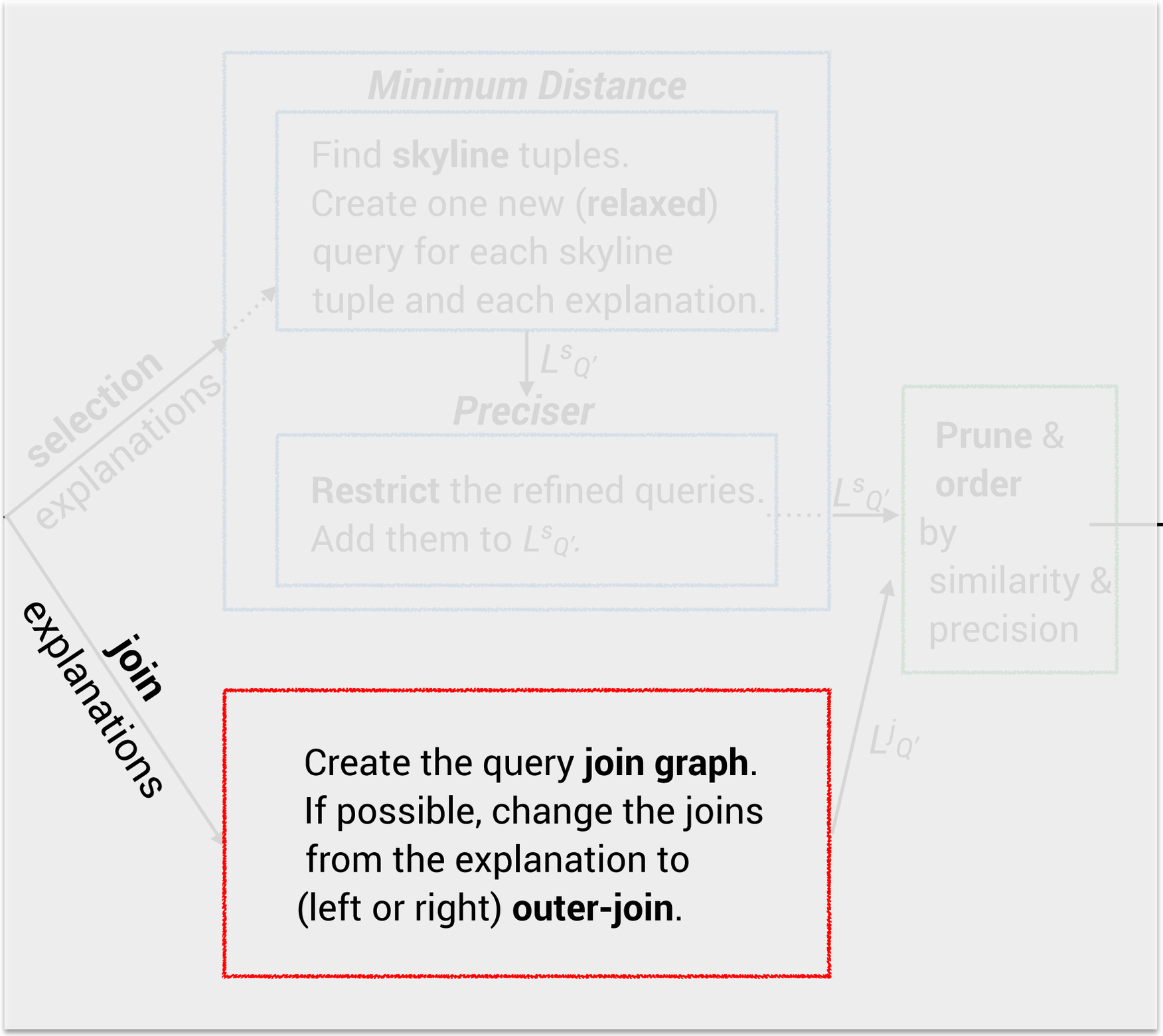
Preciser

Refinement (L^s_Q)	New conditions	False positive tuples
Q_1	$\sigma_{Pages \geq 7}$	1
Q_2	$\sigma_{Bday \geq 1975}\sigma_{Pages \geq 8}$	0
Q_3	$\sigma_{Bday \geq 1975}\sigma_{Citations \geq 90}$	0
Q_4	**$\sigma_{Pages \geq 7}\sigma_{Year \geq 2008}$**	**0**
Q_5	**$\sigma_{Pages \geq 7}\sigma_{Bday \geq 1980}$**	**0**

FixTed

Input

- Query
- Why-Not Polynomial
- Database



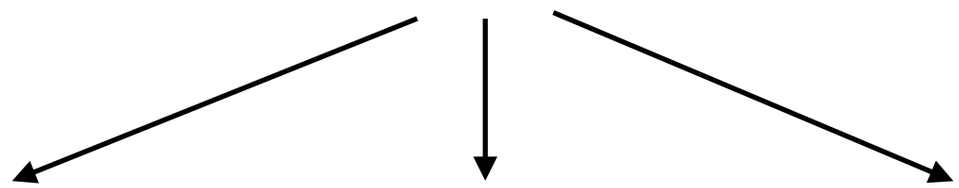
Output

List of ordered refined queries.

Join Repair

$$\bowtie_{Name=Author} + 2\sigma_{Pages>10} + 2\sigma_{Bday>78}\sigma_{Pages>10} + 2\sigma_{Bday>78}\sigma_{Citations\geq 100} + \ldots$$

$$+\sigma_{Pages>10}\sigma_{Bday>78}\sigma_{Citations\geq 100}\bowtie_{Name=Author}$$



selections

$\sigma_{Pages>10}$

$\sigma_{Bday>78}\sigma_{Pages>10}$

$\sigma_{Bday>78}\sigma_{Citations\geq 100}$

$\sigma_{Pages>10}\sigma_{Citations\geq 100}$

joins

$\bowtie_{Name=Author}$

selections & joins

$\sigma_{Pages>10}\bowtie_{Name=Author}$

$\sigma_{Bday>78}\bowtie_{Name=Author}$

$\sigma_{Citations\geq 100}\bowtie_{Name=Author}$

$\sigma_{Bday>78}\sigma_{Pages>10}\bowtie_{Name=Author}$

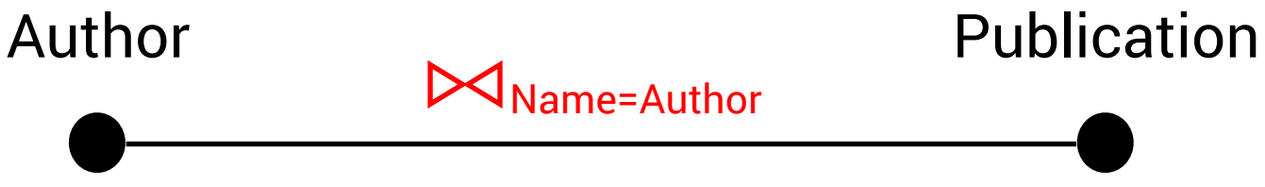
$\sigma_{Bday>78}\sigma_{Pages>10}\sigma_{Citations\geq 100}\bowtie_{Name=Author}$

Join Repair

Query join graph



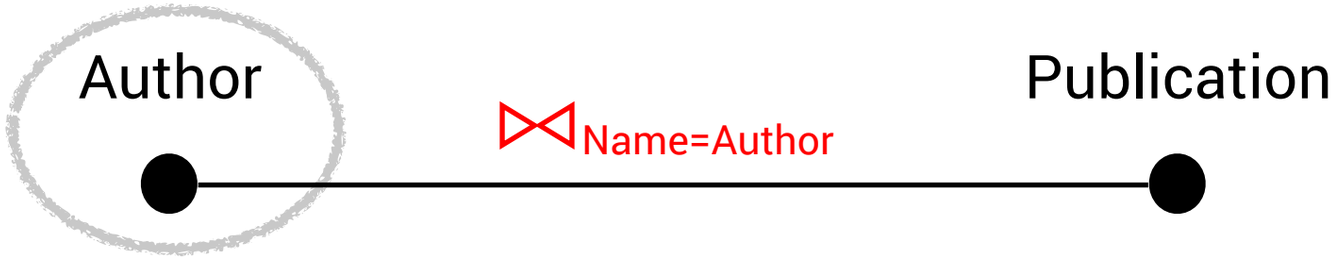
Join Repair





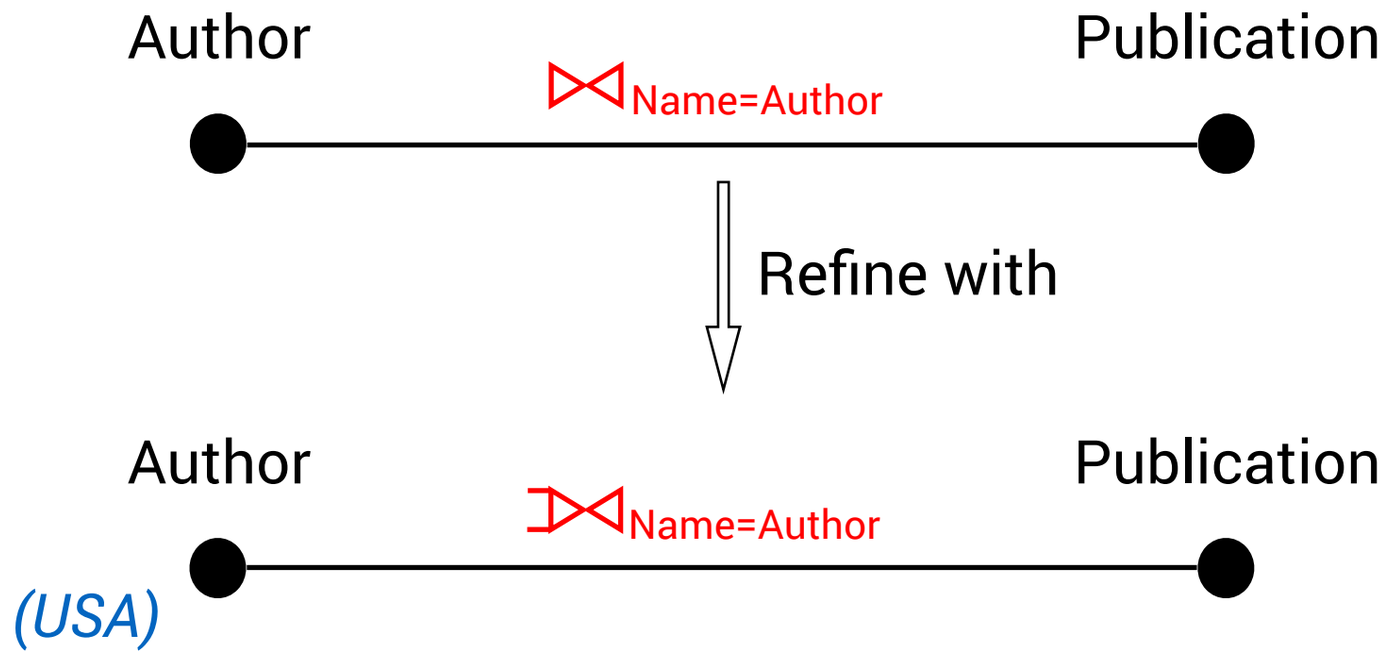
Join Repair

Direct relation





Join Repair


Author
Name=Author
Publication
Refine with
Author
Name=Author
Publication
(USA)

Join Repair

Original query

```
SELECT   A.Country
FROM     Author A, Publication P
WHERE    A.Birthday > 1978  AND
         P.Pages > 10       AND
         P.Citations ≥ 100  AND
         P.Year > 2006      AND
         A.Name = P.Author
```

Refined query

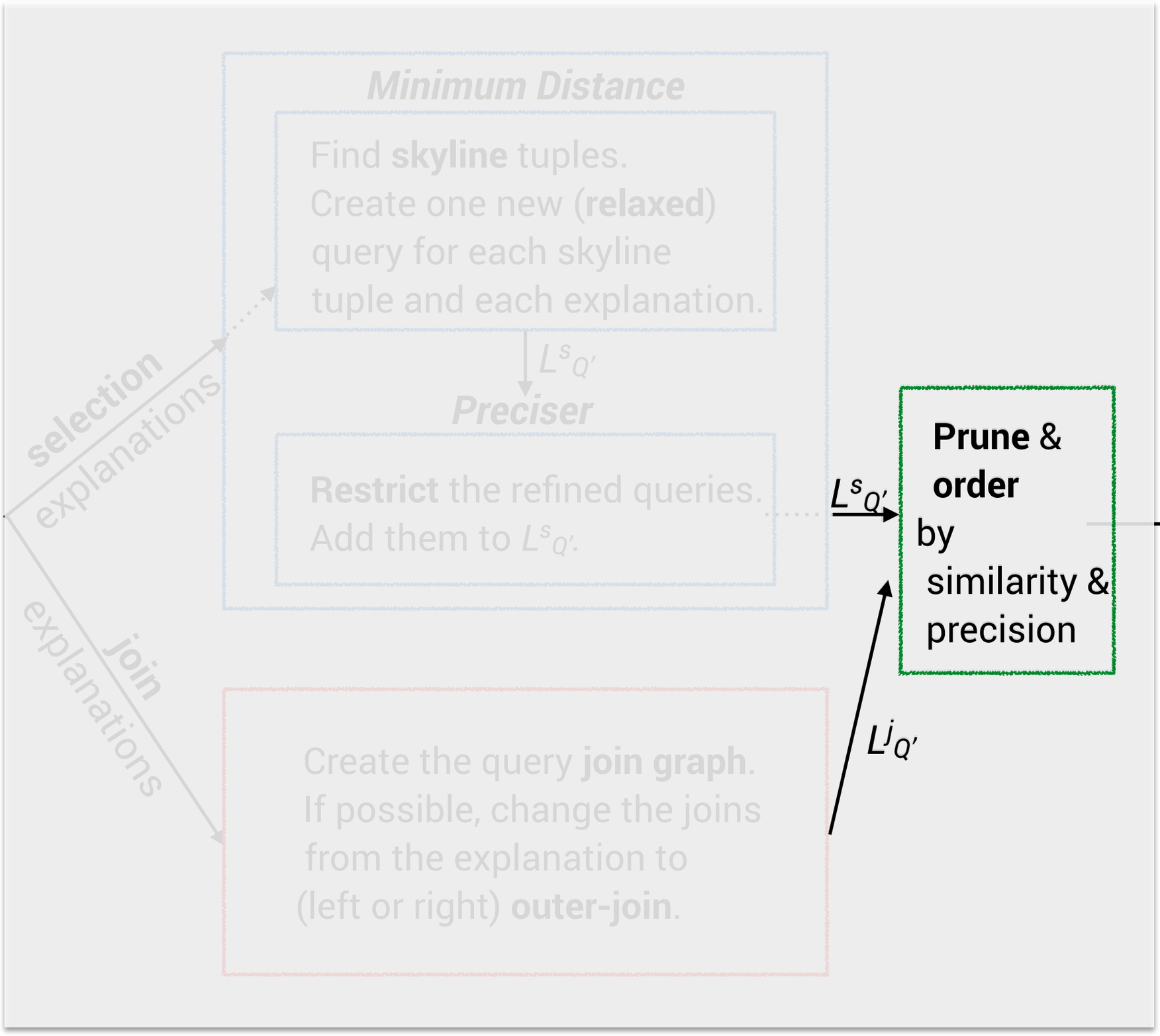
Q_6

```
SELECT   A2.Country
FROM     ( SELECT A.Name, A.Country
           FROM Author A
           WHERE A.Birthday > 1978 ) as A2
 LEFT OUTER JOIN ( SELECT P.Author
                   FROM Publication P
                   WHERE P.Pages > 10 AND P.Citations ≥ 100)
                   as P2
ON   A2.Name = P2.Author
```

FixTed

Input

- Query
- Why-Not Polynomial
- Database



Output

List of ordered refined queries.

Prune and Order

Metrics

Similarity:

Number of changed conditions

Distance between the original and changed condition values

Number of added conditions

Precision: $\frac{|Q'[D]| - FP_{Q'}}{|Q'[D]|}$

Skyline of refined queries based on the metrics.

Prune and Order

All Refinements	New conditions
Q_1	$\sigma_{Pages \geq 7}$
Q_2	$\sigma_{Bday \geq 1975}\sigma_{Pages \geq 8}$
Q_3	$\sigma_{Bday \geq 1975}\sigma_{Citations \geq 90}$
Q_4	$\sigma_{Pages \geq 7}\sigma_{Year \geq 2008}$
Q_5	$\sigma_{Pages \geq 7}\sigma_{Bday \geq 1980}$
Q_6	$\bowtie_{Name=Author}$

Most similar

Pruned & Ordered Refinements	New conditions
Q_2	$\sigma_{Bday \geq 1975}\sigma_{Citations \geq 90}$
Q_1	$\sigma_{Pages \geq 7}$
Q_6	$\bowtie_{Name=Author}$

Contributions and Future Work

Contributions of *FixTed*:

- Query refinement involving **join** and **selection** reparation.
- Leverage of **explanations** given by the **Why-Not polynomial**, to **optimize** the computation.
- Computation of the **most similar** refinement.
- Efficiency **improvement** of the **skyline** tuples computation by:
  - Reducing the number of considered attributes.
  - Reducing the number of considered input tuples.
- Join reparation using **outer joins**.

Contributions and Future Work

Future Work:

- Thorough experimental evaluation.
- Further investigation for join reparation (e.g., by using integrity constraints).

Thank you! Questions?

Contributions of *FixTed*:

Query refinement involving **join** and **selection** reparation.

Leverage of **explanations** given by the **Why-Not polynomial**, to **optimize** the computation.

Computation of the **most similar** refinement.

Efficiency **improvement** of the **skyline** tuples computation by:

Reducing the number of considered attributes.

Reducing the number of considered input tuples.

Join reparation using **outer joins**.

Future Work:

Thorough experimental evaluation.

Further investigation for join reparation.