ESR 1.2: Traceability in Big Data Processing

Evaluating Trustworthiness of Multiple Overlapping Data Sources

DEDS Winter School, 2022
Athens, Greece

Yeasmin Ara Akter (UPC/AAU)

Supervisors:
Home University (UPC): Alberto Abelló, Petar Jovanovic
Host University (AAU): Katja Hose, Tomer Sagi

OUTLINE

- Introduction
- Motivation
  - Why Truth Discovery?
  - Applications
  - General Principal
- Related Work
- Limitations
- Objectives
- Proposed Architecture
- Existing Prototype
- Running Example
- Reference

Introduction

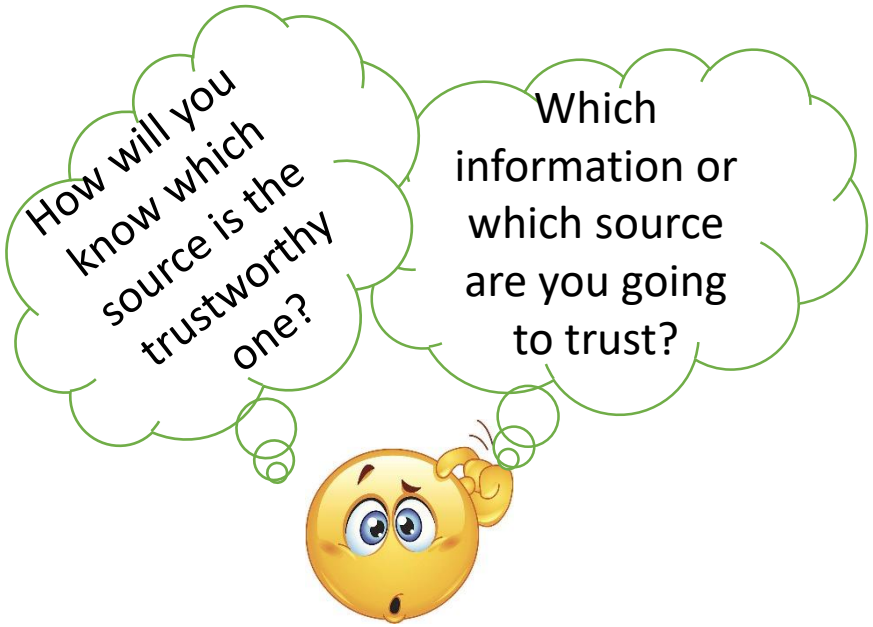
https://www.weather25.com/europe/spain/catalonia/barcelona

https://www.accuweather.com/en/es/barcelona

13^{0} Cloudy

16^{0} C, Sunny

Barcelona Temperature on 6th April



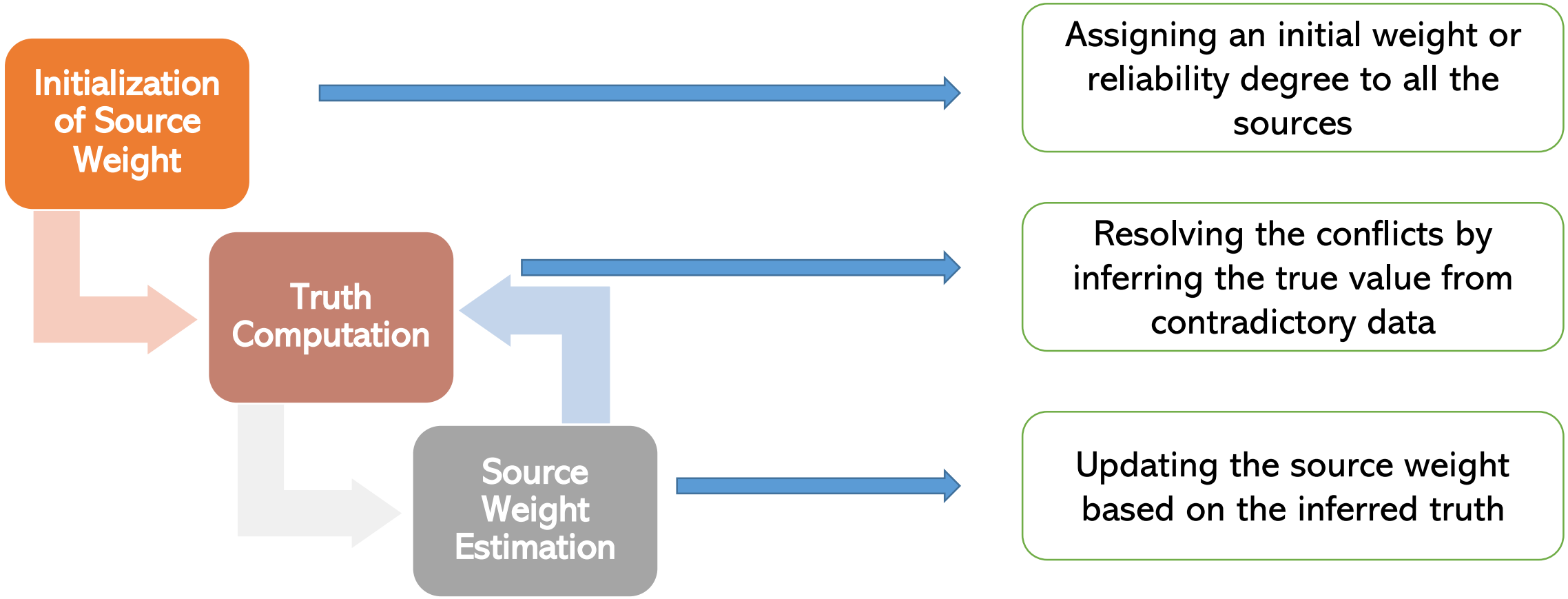
Motivation

- What is truth discovery?
  - Discovering the trusted value from multiple-noisy data sources
- Why it is needed?
  - To resolve the conflicts
  - To integrate true data in a single platform
  - To provide trustable information to the user
  - To reduce the delays of data analytics projects

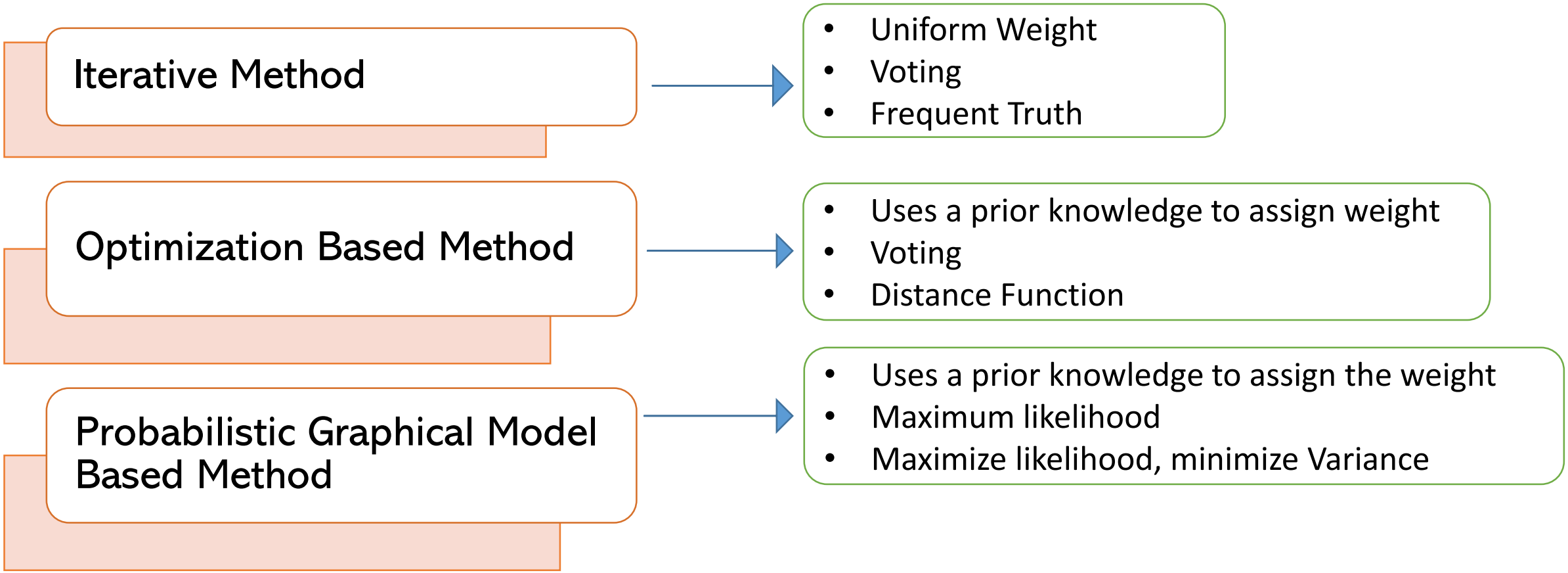
Applications of Evaluating trustworthiness

- Healthcare
- Social Sensing
- Crowdsourcing
- Information Extraction
- Location Based Services
- Sensor network
- Organizational Data

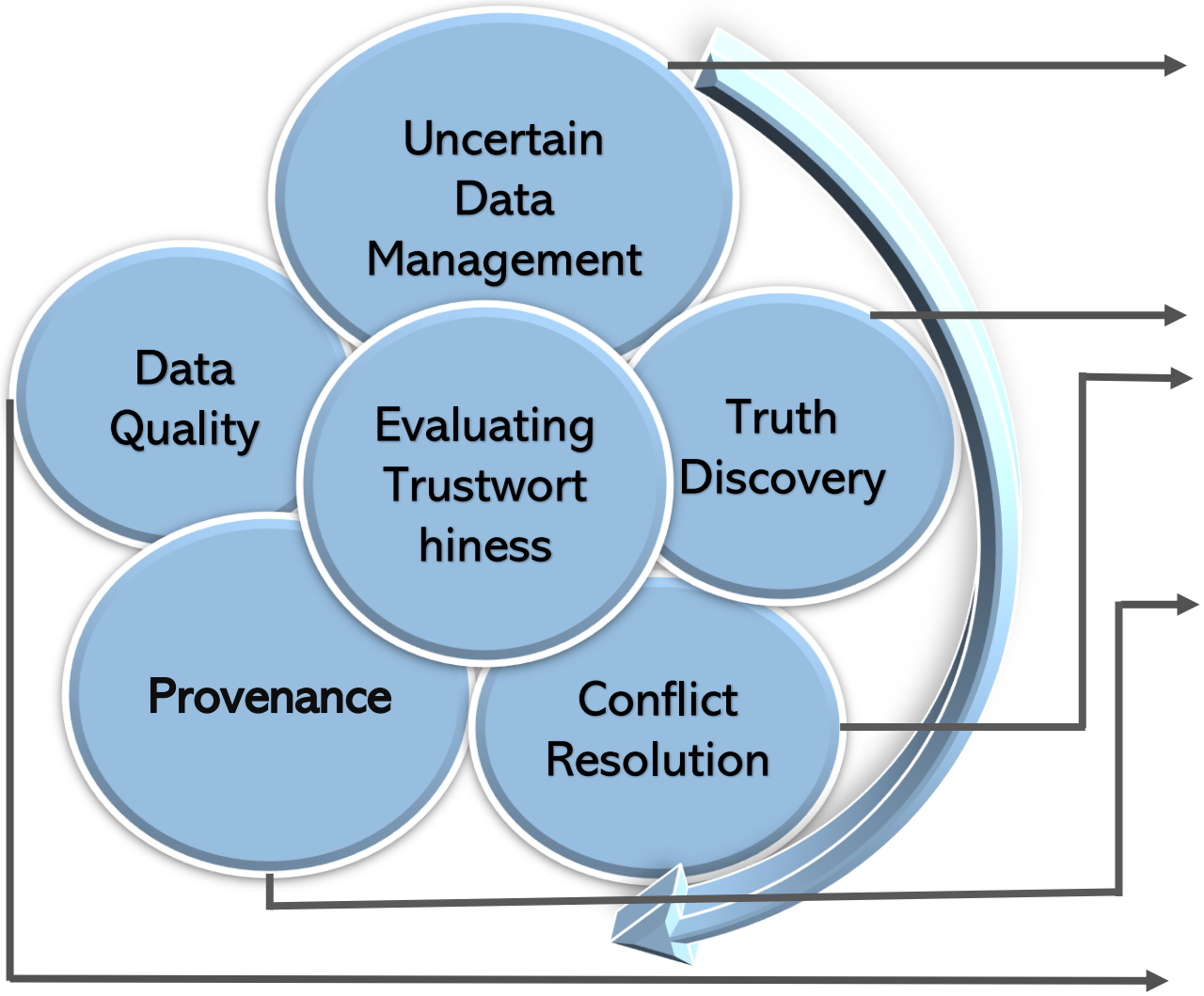
General Principle of Truth Discovery



Truth Discovery Methods



Related Domain



Input Uncertainty of data degrades the data quality and trustworthiness of information

Helps to discover trustworthy information from multiple overlapping data sources and resolves the conflict. Conflict resolution can take place for both categorical and continuous data.

Provenance helps to keep track of the error and improves the traceability and trustability

Improving **data quality** improves the source trustworthiness

Related Work

Systems	Type	Uncertainty Handling	Truth Discovery Method			Evaluation Metric
			Considered Source Dependency	Truth Computation	Ground Truth Evaluation	
Apollo-social [2]	Probabilistic Graphical Model	×	×	Maximum Likelihood	×	Precision, Recall
CATD [3]	Optimization	×	×	Weighted averaging	×	MAE, RMSE
RCHDTD [4]	Optimization	×	×	Weighted Voting Weighted Median	✓	Mean Normalized Absolute Distance (MNAD)
SmartMTD [6]	Probabilistic Graphical Model	×	✓	Majority Voting	✓	Precision, Recall, F1-Score, Execution Time
EPTD [5]	Iterative	×	×	Majority Voting	✓	MAE, RMSE
SRTD [7]	Iterative	✓	×	Majority Voting	✓	Specificity (SPC), Matthews Correlation Coefficient (MCC), Cohen's Kappa (Kappa)
RPPTD [8]	Optimization	×	×	Majority Voting	✓	Execution Time
RTD [9]	Iterative	×	×	Mean Shift Clustering	✓	MAE, MSE, R-Squared

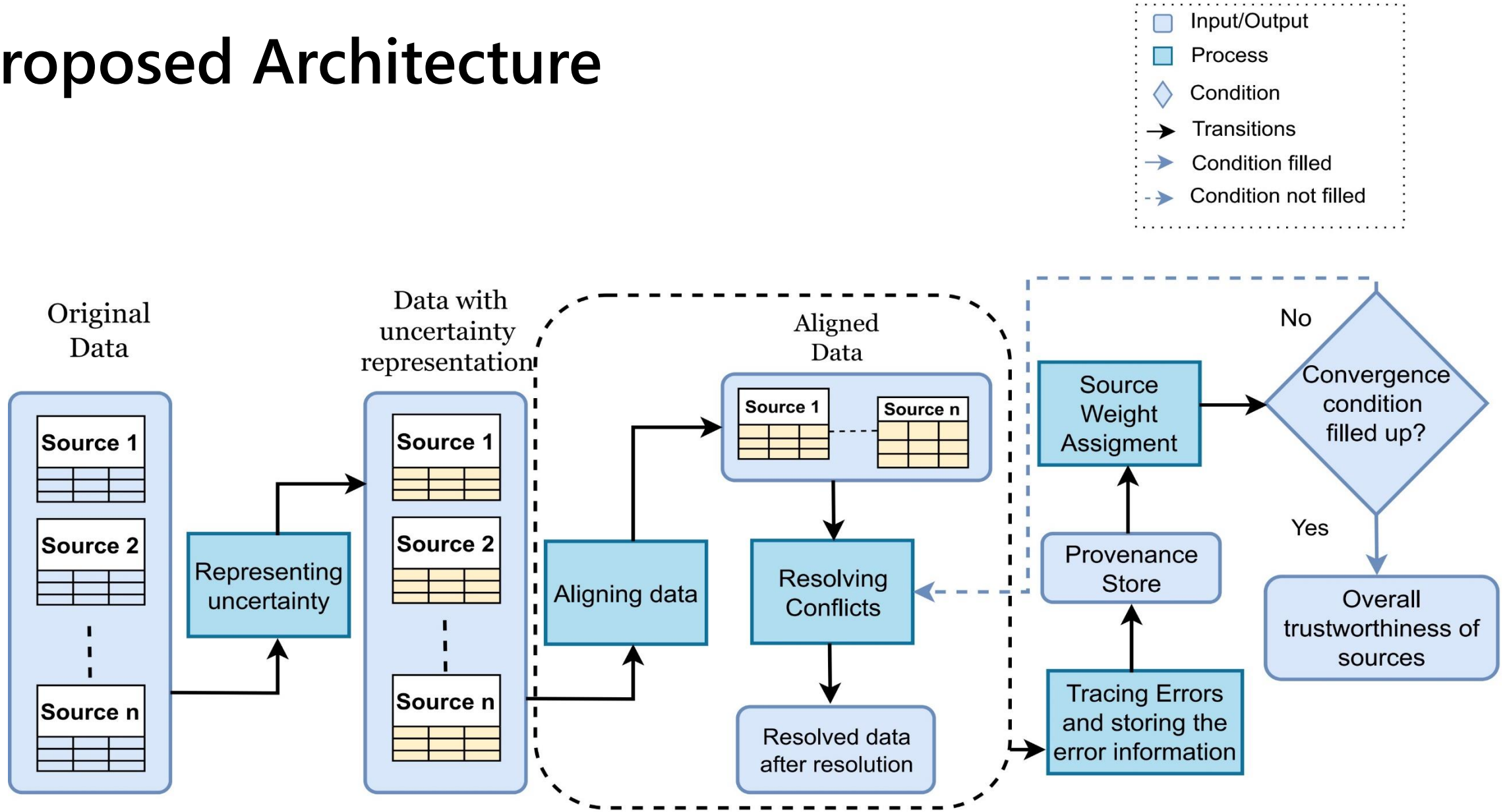
Limitations

- Uncertainty is ignored in most of the trustworthiness evaluation system
- Different data type must be treated differently
- Use of gold standard data
- Error is not traced throughout the workflow
- No specific evaluation metric to provide overall degree of trustworthiness
- Lack of a framework considering all the related domains concurrently

Objective

- Determining a representation method for both uncertain and missing data
- Determining an efficient attribute conflict resolution method that supports aligning data from multiple sources
- Developing an efficient tracing method of data transformations with the help of data provenance techniques to represent the propagation of trust
- Determining a metric to estimate the degree of trustworthiness of sources given multiple overlapping data sources

Proposed Architecture

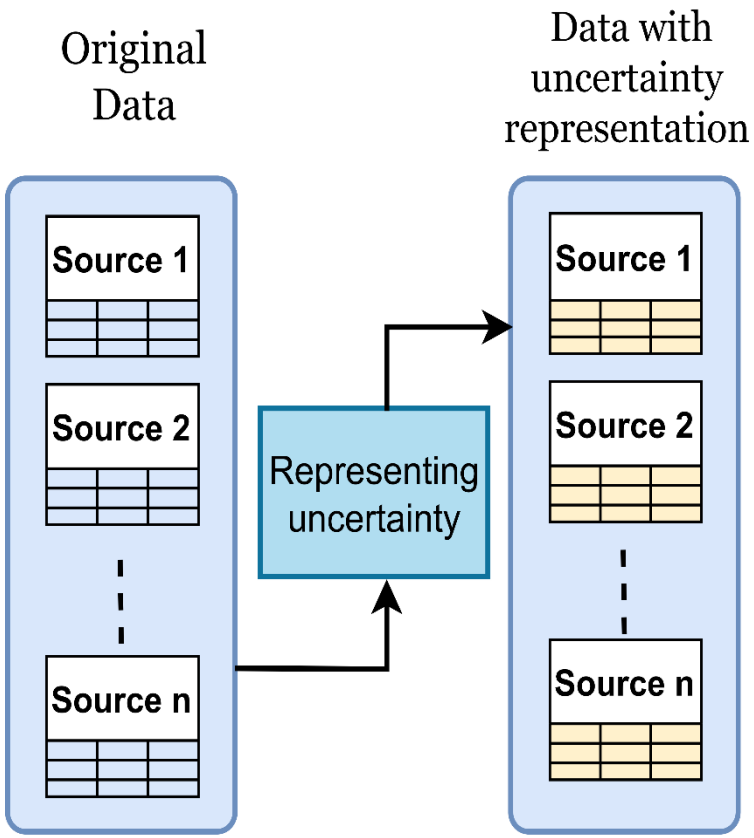


Uncertainty Representation

Uncertainty arises when one source provides a non-null value, but another provides NULL or no information or when two sources provide contradictory data for the same real-world object or when they provide values in a confidence interval.

Challenges:

- How can we differentiate between contradictory or missing data?
- Are the symbolic expressions better than pure NULL to represent the missing data?

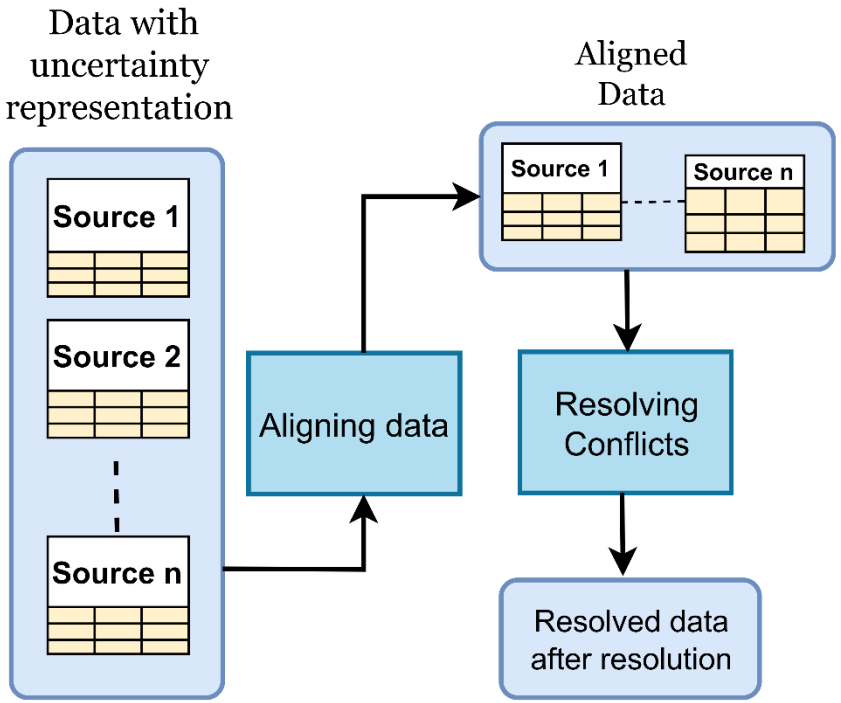


Truth Discovery and Conflict Resolution

Conflict resolution is inferring the true value among multiple contradictory data.

Challenges:

- What would be the best possible way to align the data?
- One source may collect noisy data from another source which affects the truth computation- Source Dependency
- Different Quality of sources in different fields – Proper Domain Subdivision

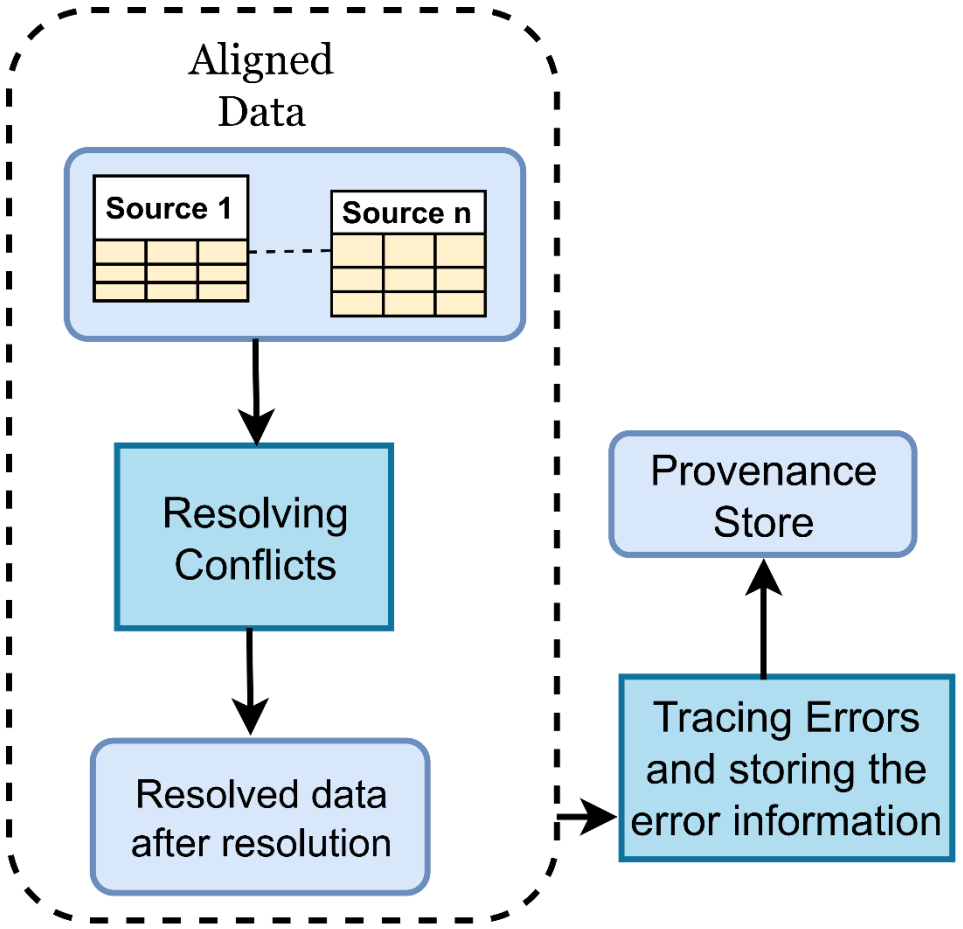


Provenance

Provenance supplies connection between data in the source and the output.

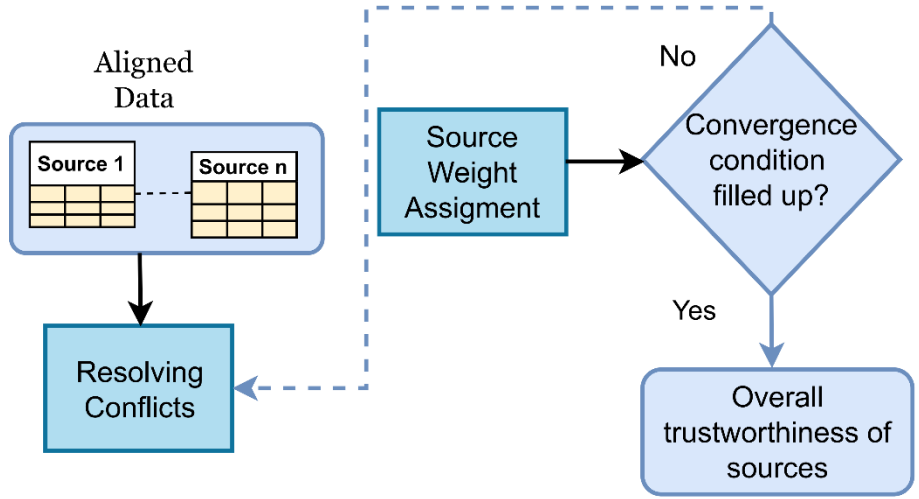
Challenges:

- Difficult to trace the error throughout the data transformation workflow
- To keep the track of error, which provenance technique would be best fit?



Deciding Trustworthiness

Deciding trustworthiness is providing an overall degree of trustworthiness of a source based on which used can make the decision

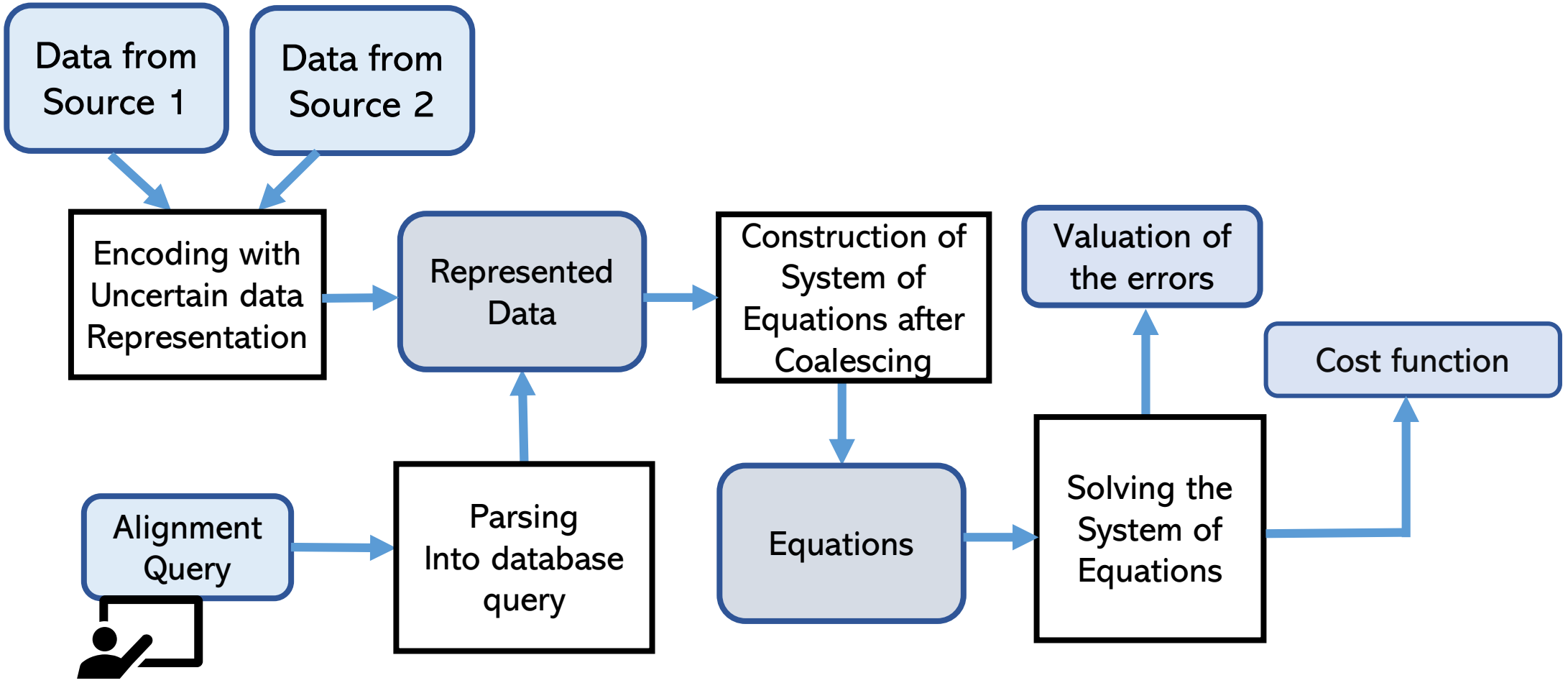


Challenges:

- Difficult to decide the criteria for convergence condition
- Difficult to deciding metric to assign the overall trustworthiness
- Difficult to have the ground truth because it is expensive to have the labelled data while evaluating performance

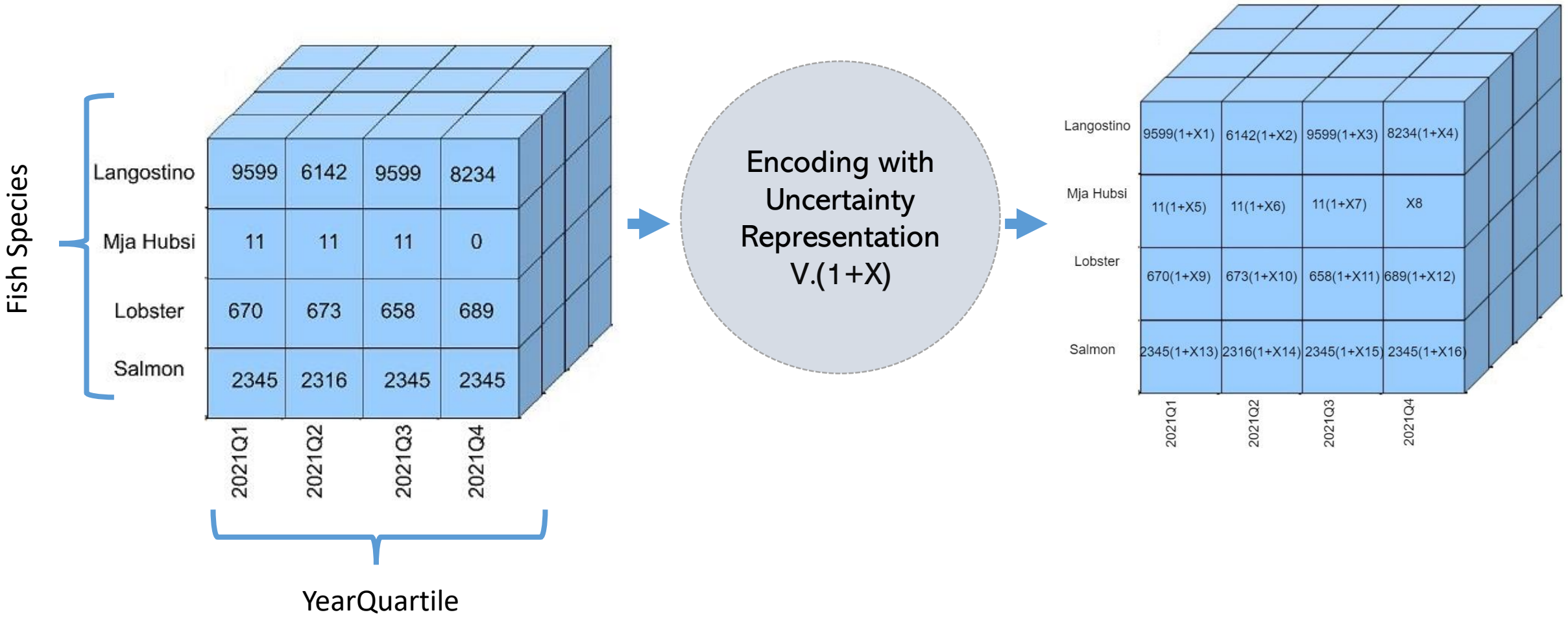
What We Have
ERIS - An existing prototype

Prototype Workflow



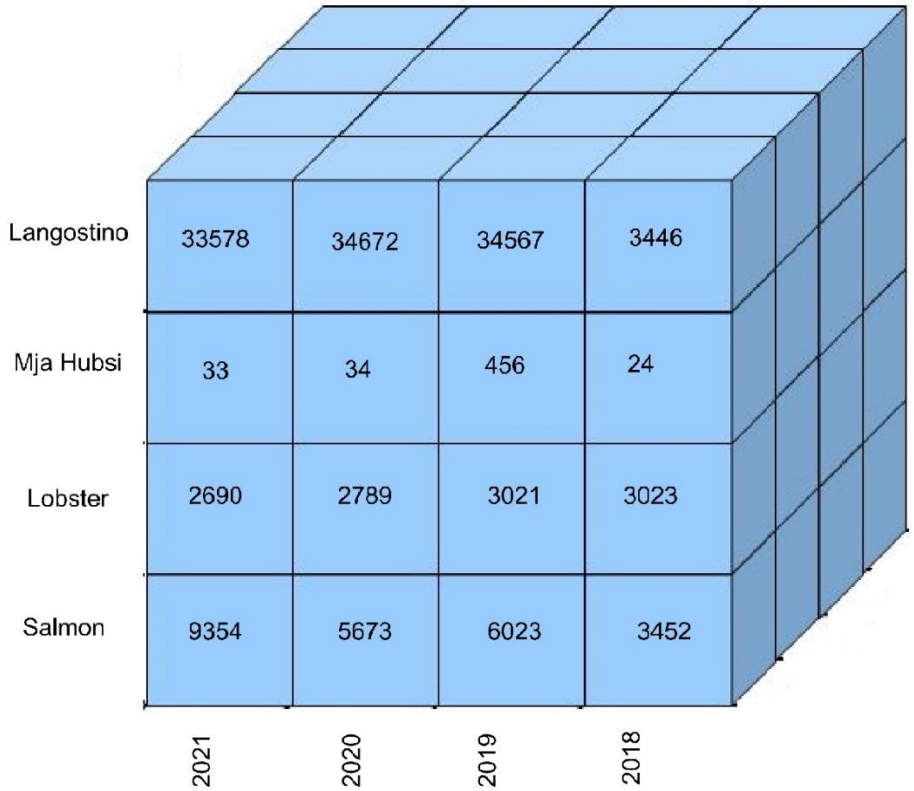
Running Example

Table ESP from Source 1

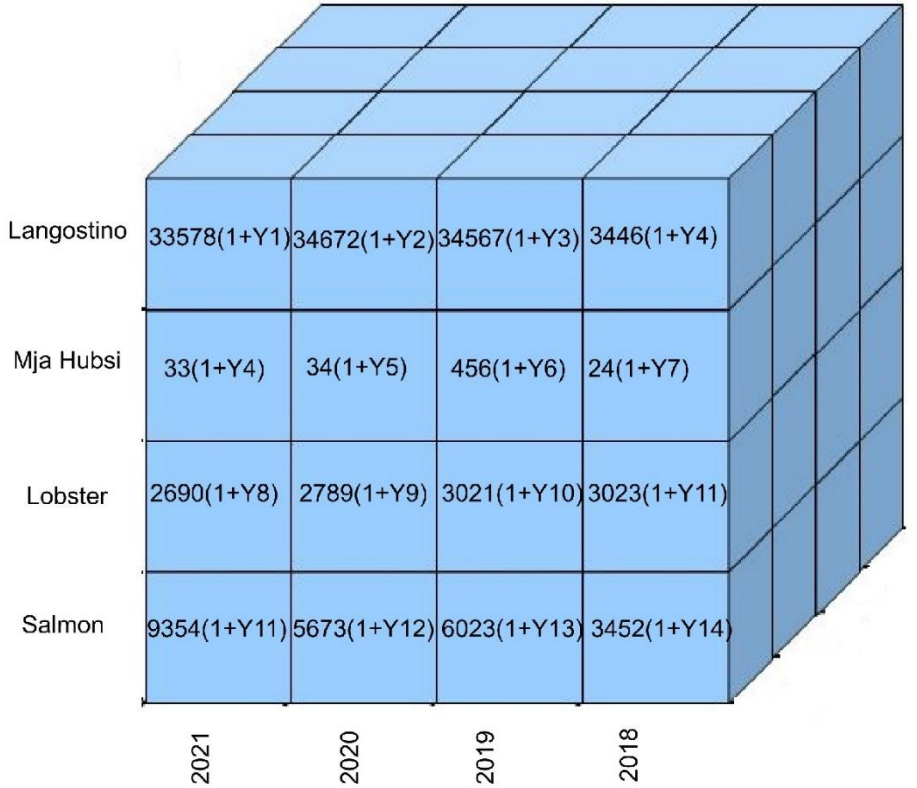


Running Example

Table ALLESP from Source 1

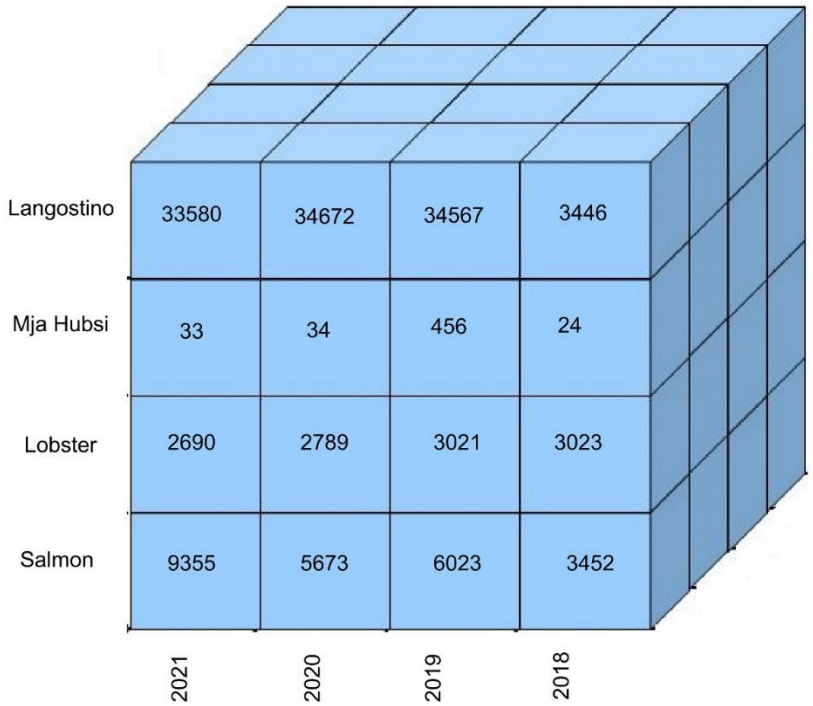


Encoding with Uncertainty Representation V.(1+Y)

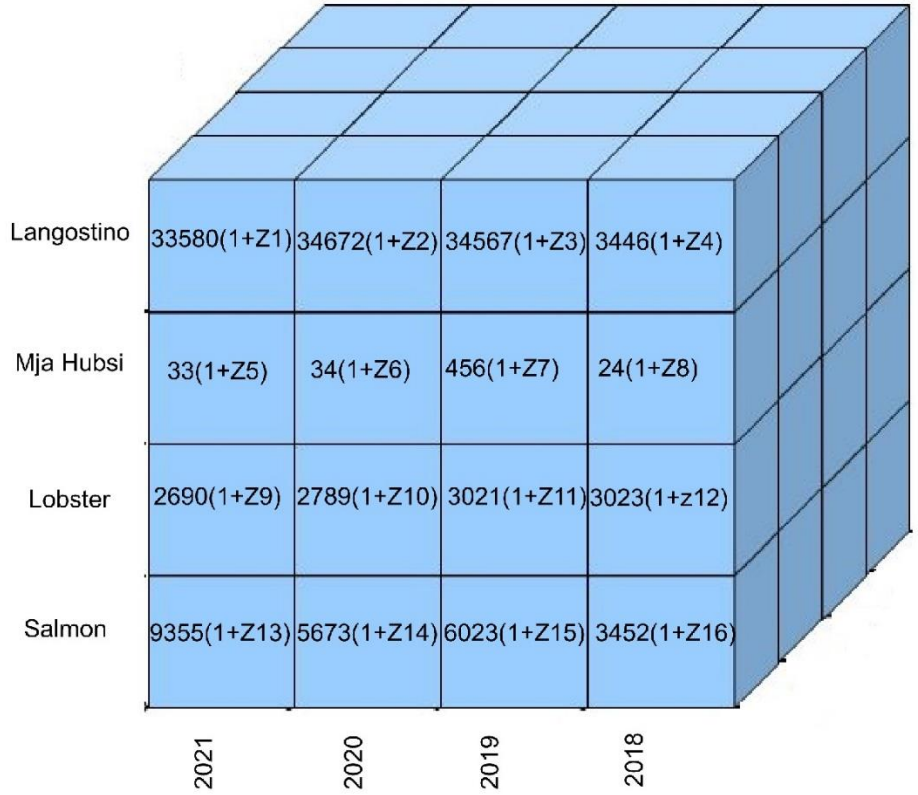


Running Example

Table AESP from Source 2



Encoding with Uncertainty Representation V.(1+Z)



Running Example

Alignment Query

$$\left(\sigma_{FishSpecies,YearQuartile}\left(\ {}_{YearQuartile}\gamma_{SUM(NumberfoFish)}\right)\right) \cup \left(\sigma_{FishSpecies,\,Year,Tnof}(ALLESP)\right)$$

Parsing
Into database query

Select (σ),
Project (π),
ProjectAway ($\hat{\pi}$),
Join ($\bowtie$),
Renaming (ρ),
Difference (\ ),
Aggregation (γ),
UNION (∪),
DUNION (⊎),
Coalescing (κ)

```
t1:= SELECT FishSpecies, YearQuartile, SUM(NumberofFish) AS Tnof FROM ESP GROUP BY YearQuartile
t2:= SELECT FishSpecies, Year, Tnof FROM ALLESP
t3:= t1 UNION ALL t2
```

Running Example

Table ESP from Source 1

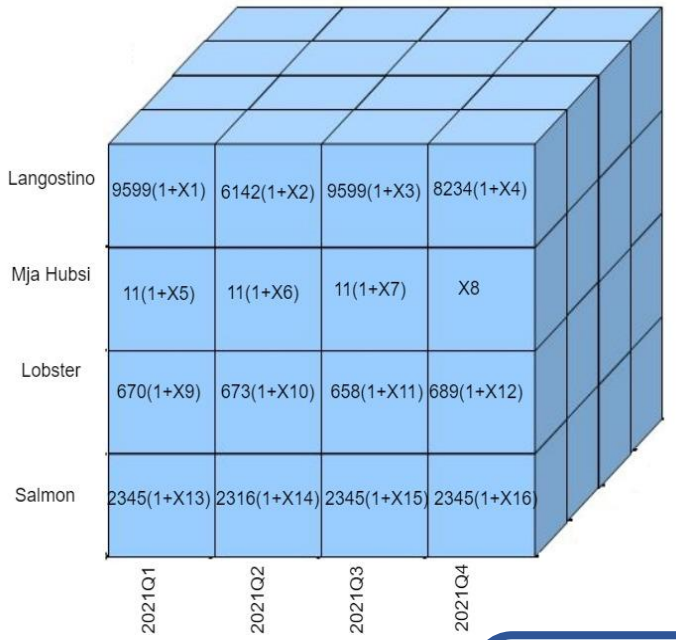
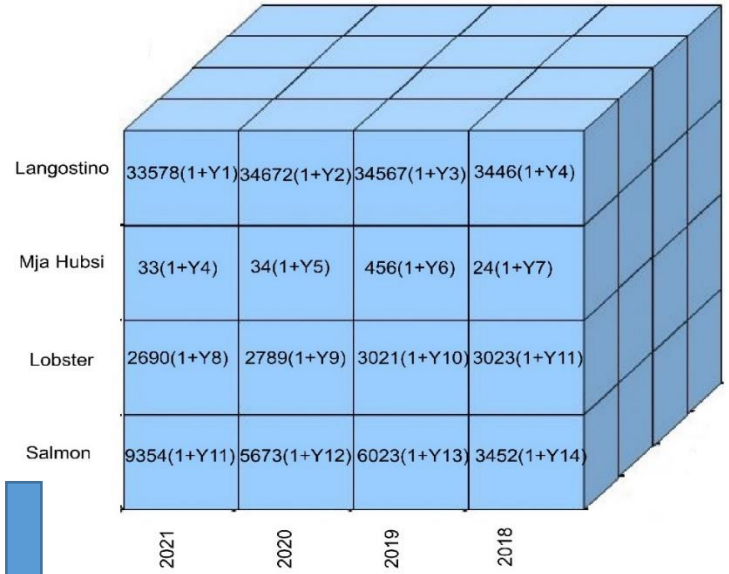


Table ALLESP from Source 1



Construction of System of Equations After Coalescing

Fish Species	Aggregation from ESP 2021	From ALLESP 2021
Langostino	33574	33578
Mja Hubsi	33	33
Lobster	2690	2690
Salmon	9351	9354

Integrity Constraint

Equation will be generated only when there is disagreement to maintain the functional dependency

Running Example

Table ESP from Source 1

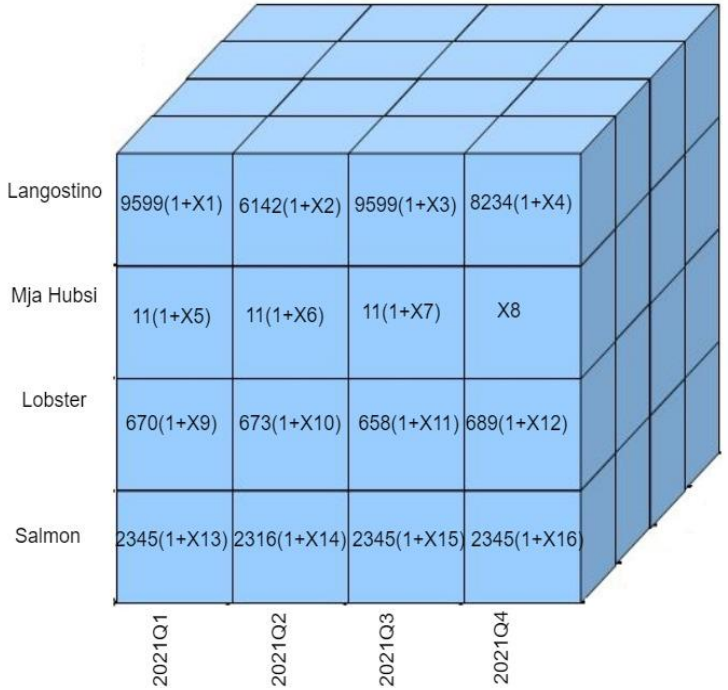
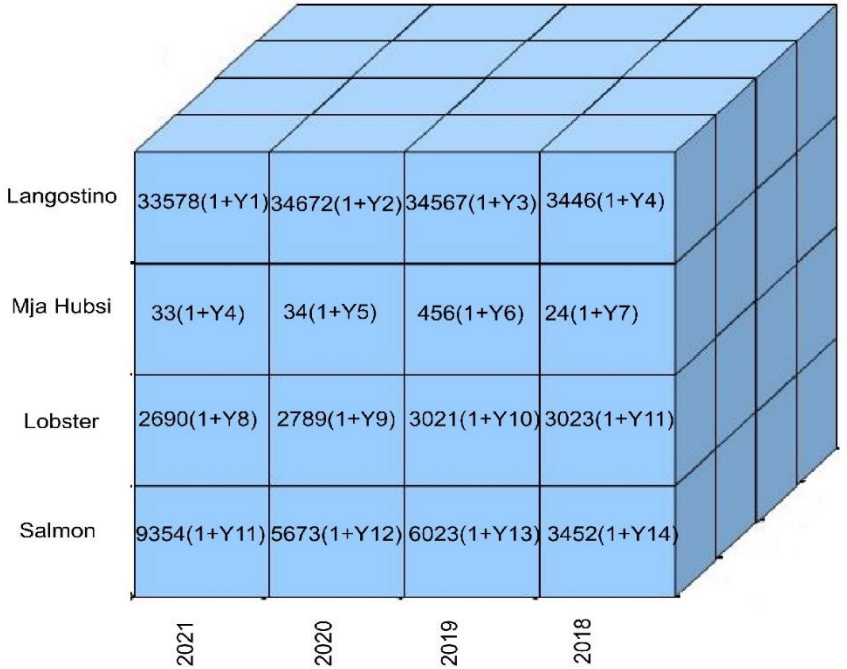


Table ALLESP from Source 1



Construction of System of Equations after Coalescing

Fish Species	2021
Langostino	9599 (1+X1) + 6142 (1+X2) + 9599 (1+X3) + 8234 (1+X4) = 33578 (1+Y1)
Salmon	2345 (1+X13) + 2316 (1+X14) + 2345 (1+X15) + 2345 (1+X16) = 9354 (1+Y11)

Running Example

Table ALLESP from Source 1

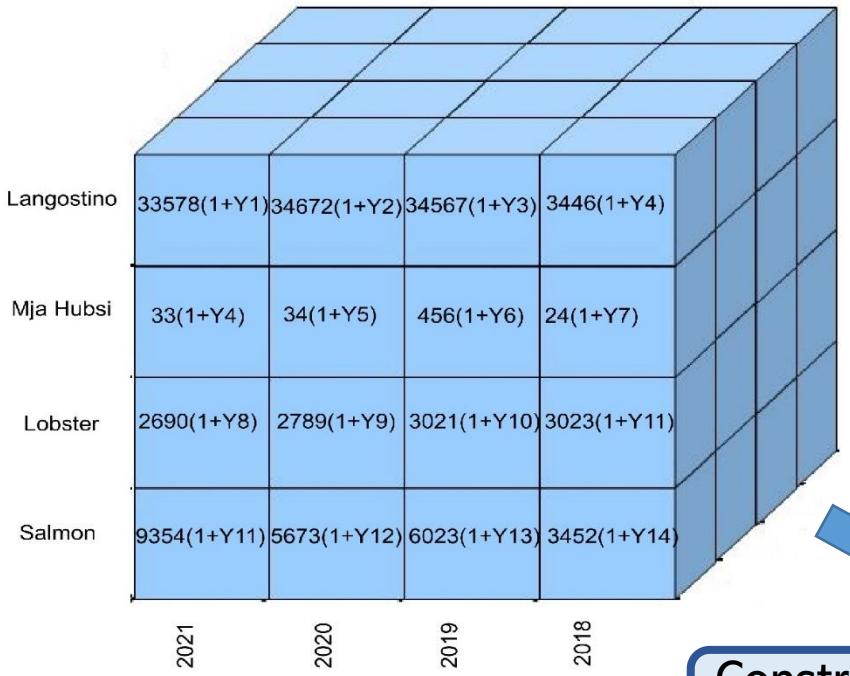
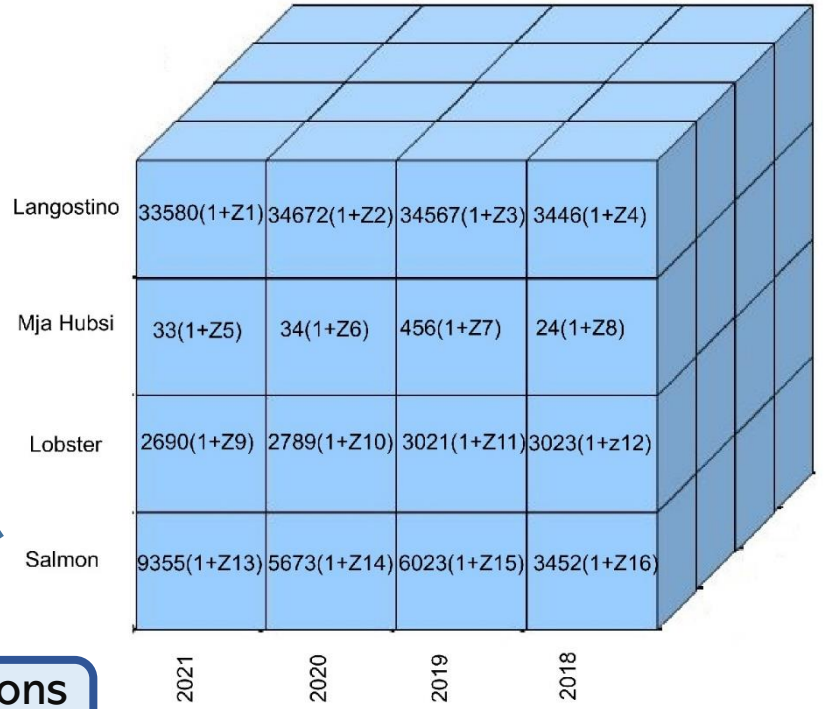


Table AESP from Source 2

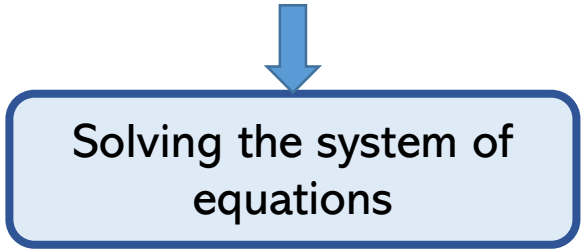


Construction of System of Equations between multiple sources

Fish Species	2021
Langostino	33578 (1+Y1) = 33580 (1+Z1)
Salmon	9354 (1+Y11) = 9355 (1+Z13)

Running Example

Fish Species	Within same source	Within different source
Langostino	9599 (1+X1) + 6142 (1+X2) + 9599 (1+X3) + 8234 (1+X4) = 33578 (1+Y1)	33578 (1+Y1) = 33580 (1+Z1)

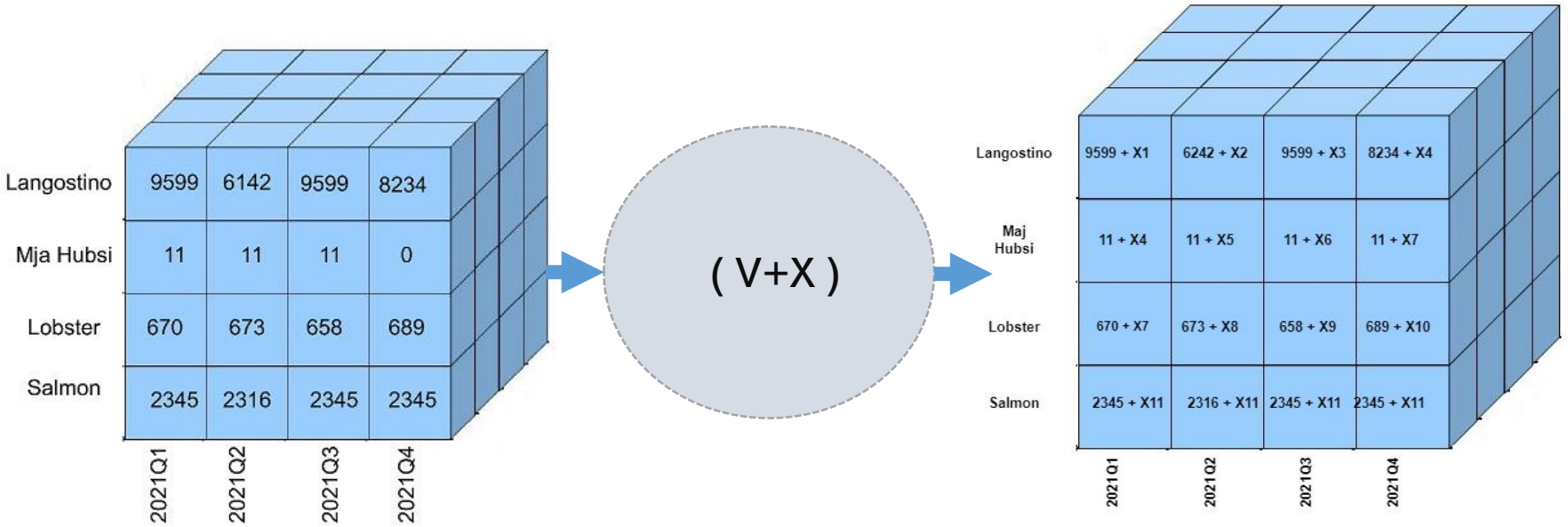


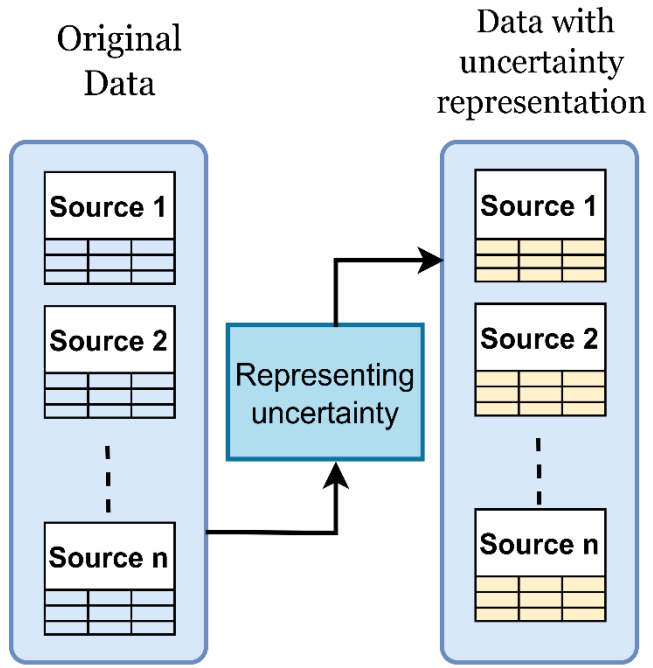
Variables	X1	X2	X3	X4	Y1	Z1
Valuation	-1	1	0	0	0.468	0.468
$\sum X_i^2$	0.406					
Valuation	0	0	0	1	0.245	0.245
$\sum X_i^2$	0.1866					

Ongoing Work

Applying different linear expression (e.g., V+X) to represent the uncertain data

- To observe how the unknown, null or uncertain values react according to their different way of symbolic representation





Ongoing Work

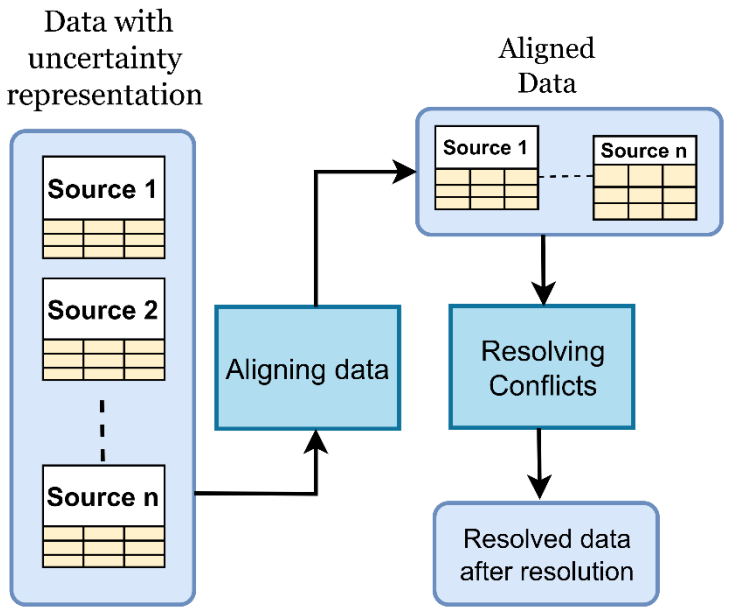
Adding positive value constraints or Interval constraints

- Solving the system of equations helps to resolve the data conflicts among the sources

  - Positive **valuations** of variables or valuation under a given interval have an impact on cost function hence overall trustworthiness

Variables	X1	X2	X3	X4	Y1	Z1
Valuation	-1	1	0	0	0.468	0.468
$\sum \lvert X_i \rvert$	0.406					
Valuation	0	0	0	1	0.245	0.245
$\sum \lvert X_i \rvert$	0.1866					

positive or within a given interval



Ongoing Work

- ✓ Applying **Average Absolute Error** as cost function

  To determine a metric which can give the best result as degree of trustworthiness

- ✓ Evaluation with **ground truth or without ground truth**

  If ground truth is available, considering both the cases may increase the trustworthiness of the sources

Variables	X1	X2	X3	X4	Y1	Z1
Valuation	1	1	0	0	0.468	0.468
$\sum \lvert X_i \rvert$	0.489					
Valuation	0	0	0	1	0.245	0.245
$\sum \lvert X_i \rvert$	0.248					

Next Tasks.....

- ✓ Applying the Average Absolute Error
- ✓ Adding the value constraints in the valuation
- ✓ Generalizing the system with or without ground truth
- ✓ Proposing a provenance approach to trace the error from the source to whole workflow
- ✓ Determining an evaluation metric
- ✓ Determining the Truth computation convergence criteria

References

SL	Paper
1	Li, Yaliang, et al. "A survey on truth discovery." *ACM Sigkdd Explorations Newsletter* 17.2 (2016): 1-16.
2	Wang, Dong, et al. "Using humans as sensors: an estimation-theoretic perspective." *IPSN-14 proceedings of the 13th international symposium on information processing in sensor networks*. IEEE, 2014.
3	Li, Qi, et al. "A confidence-aware approach for truth discovery on long-tail data." *Proceedings of the VLDB Endowment* 8.4 (2014): 425-436.
4	Li, Yaliang, et al. "Conflicts to harmony: A framework for resolving conflicts in heterogeneous data by truth discovery." *IEEE Transactions on Knowledge and Data Engineering* 28.8 (2016): 1986-1999.
5	Xu, Guowen, et al. "Efficient and privacy-preserving truth discovery in mobile crowd sensing systems." *IEEE Transactions on Vehicular Technology* 68.4 (2019): 3854-3865.
6	Fang, Xiu Susie, et al. "Smartmtd: A graph-based approach for effective multi-truth discovery." *arXiv preprint arXiv:1708.02018* (2017).
7	Zhang, Daniel, et al. "On scalable and robust truth discovery in big data social media sensing applications." *IEEE transactions on big data* 5.2 (2018): 195-208.

References

SL	Paper
8	Chen, Jingxue, et al. "RPPTD: robust privacy-preserving truth discovery scheme." *IEEE systems journal* (2021).
9	Chen, Jingxue, et al. "Robust Truth Discovery Scheme Based on Mean Shift Clustering Algorithm." *Journal of Internet Technology* 22.4 (2021): 835-842.
10	Parag Agrawal, Omar Benjelloun, Anish Das Sarma, Chris Hayworth, Shubha Nabar, Tomoe Sugihara, and Jennifer Widom. Trio: A system for data, uncertainty, and lineage. Proc. of VLDB 2006 (demonstration description), 2006.
11	Dai, Chenyun, et al. "An approach to evaluate data trustworthiness based on data provenance." *Workshop on Secure Data Management*. Springer, Berlin, Heidelberg, 2008.
12	Ikbal Taleb, Mohamed Adel Serhani, and Rachida Dssouli. Big data quality: A survey. In 2018 IEEE International Congress on Big Data (BigData Congress), pages 166–173. IEEE, 2018
13	Suraj Juddoo. Overview of data quality challenges in the context of big data. In 2015 International Conference on Computing, Communication and Security (ICCCS), pages 1–9. IEEE, 2015

Thank you for attention

Any Questions?