ESR 1.2: Traceability in Big Data Processing

Evaluating Trustworthiness of Multiple Overlapping Data Sources

DEDS Summer School, 2023
Barcelona, Spain

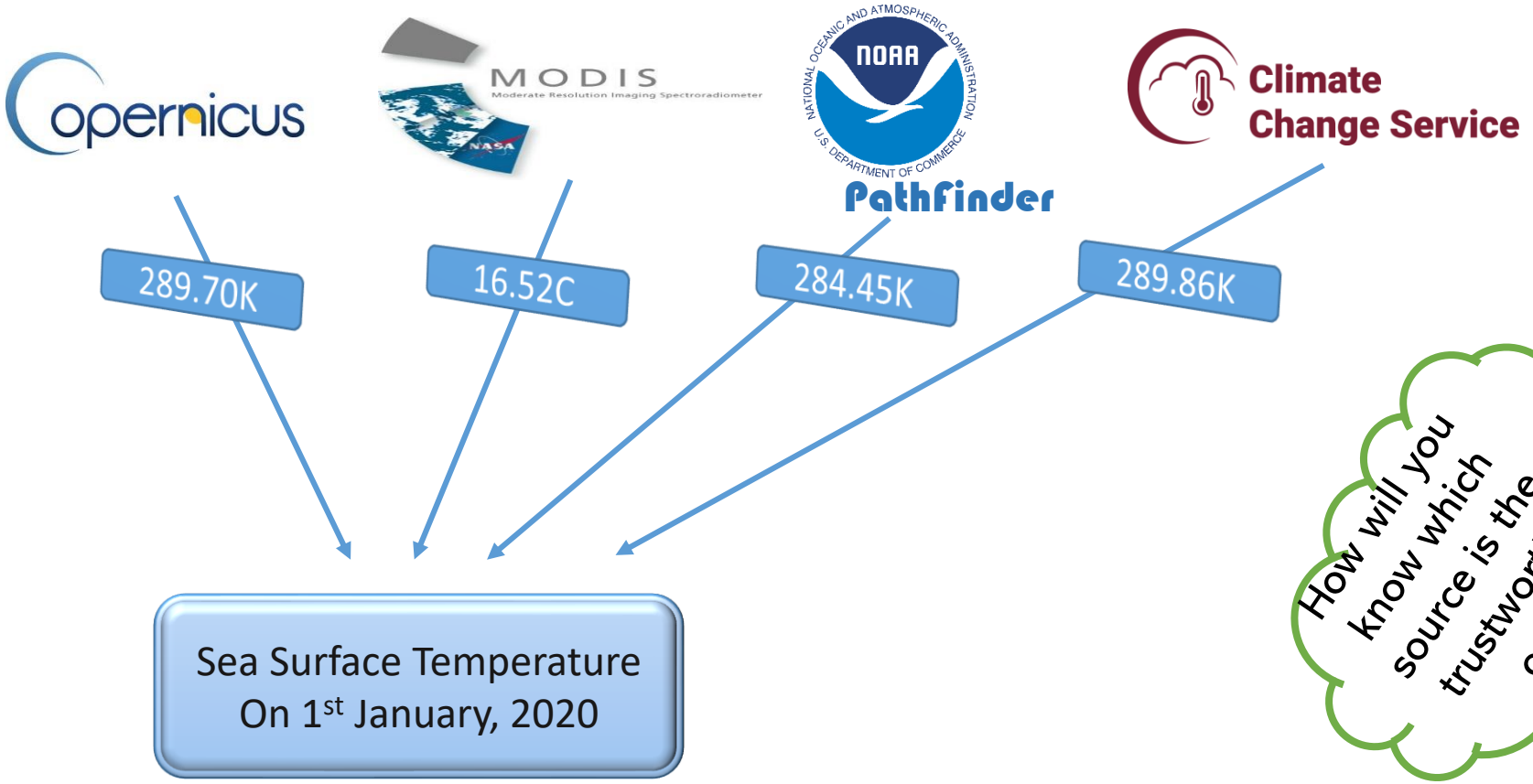
Yeasmin Ara Akter (UPC/AAU)

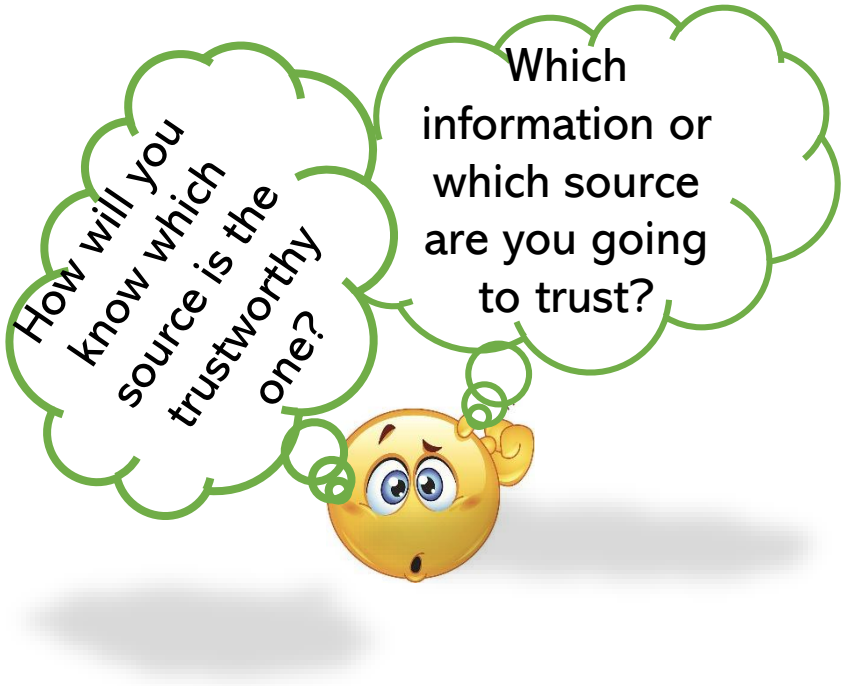
Supervisors:
Home University (UPC): Alberto Abelló, Petar Jovanovic
Host University (AAU): Katja Hose, Tomer Sagi

OUTLINE

- Introduction
- Motivation
  - Information Fusion
  - Truth Discovery
  - General Principal
- Objectives
- Proposed Architecture
- Existing Prototype
- Use Case
- Graphical User Interface
- Conclusion

Introduction





Motivation

- Information Fusion:
  - Process of integrating multiple data sources to produce more consistent, accurate, and useful information than that provided by any individual data source.
  - To get fused information, quality data is needed to get a consistent data source
  - One of the main task of information fusion is discovering the truth from multiple overlapping sources

Motivation

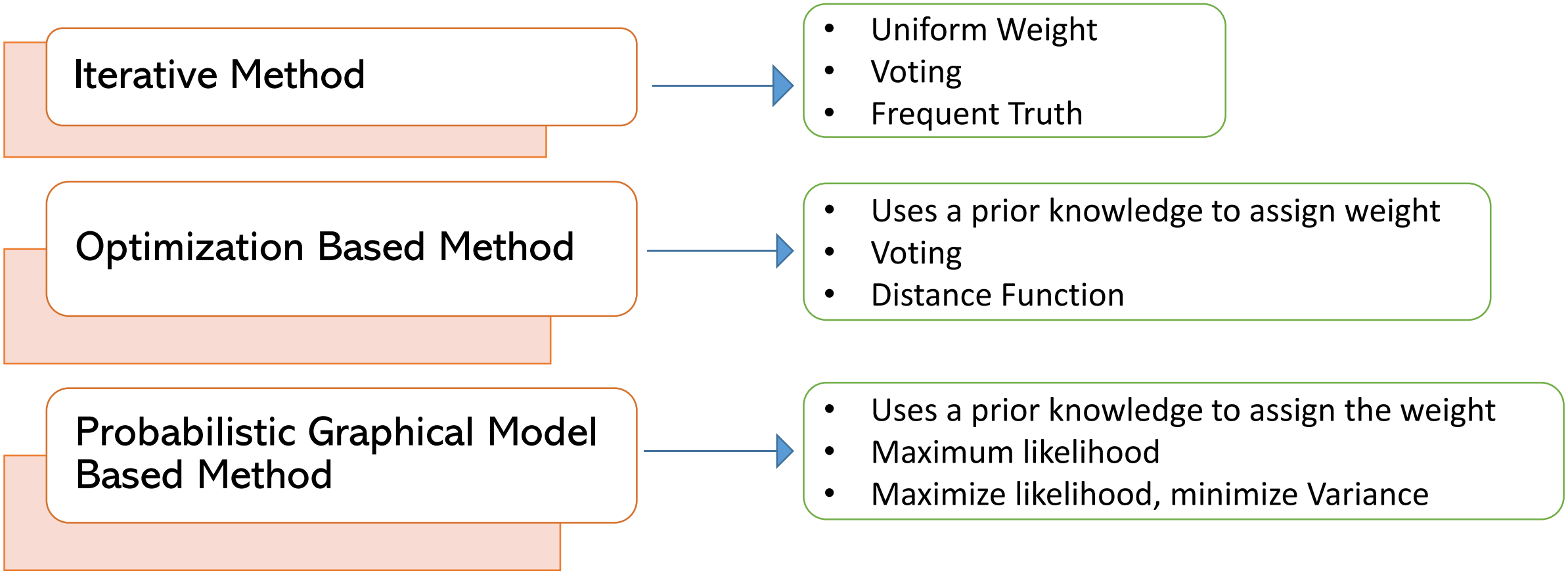
- Truth discovery- Discovering the trusted value from multiple-noisy data sources

- Necessity -
  - 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

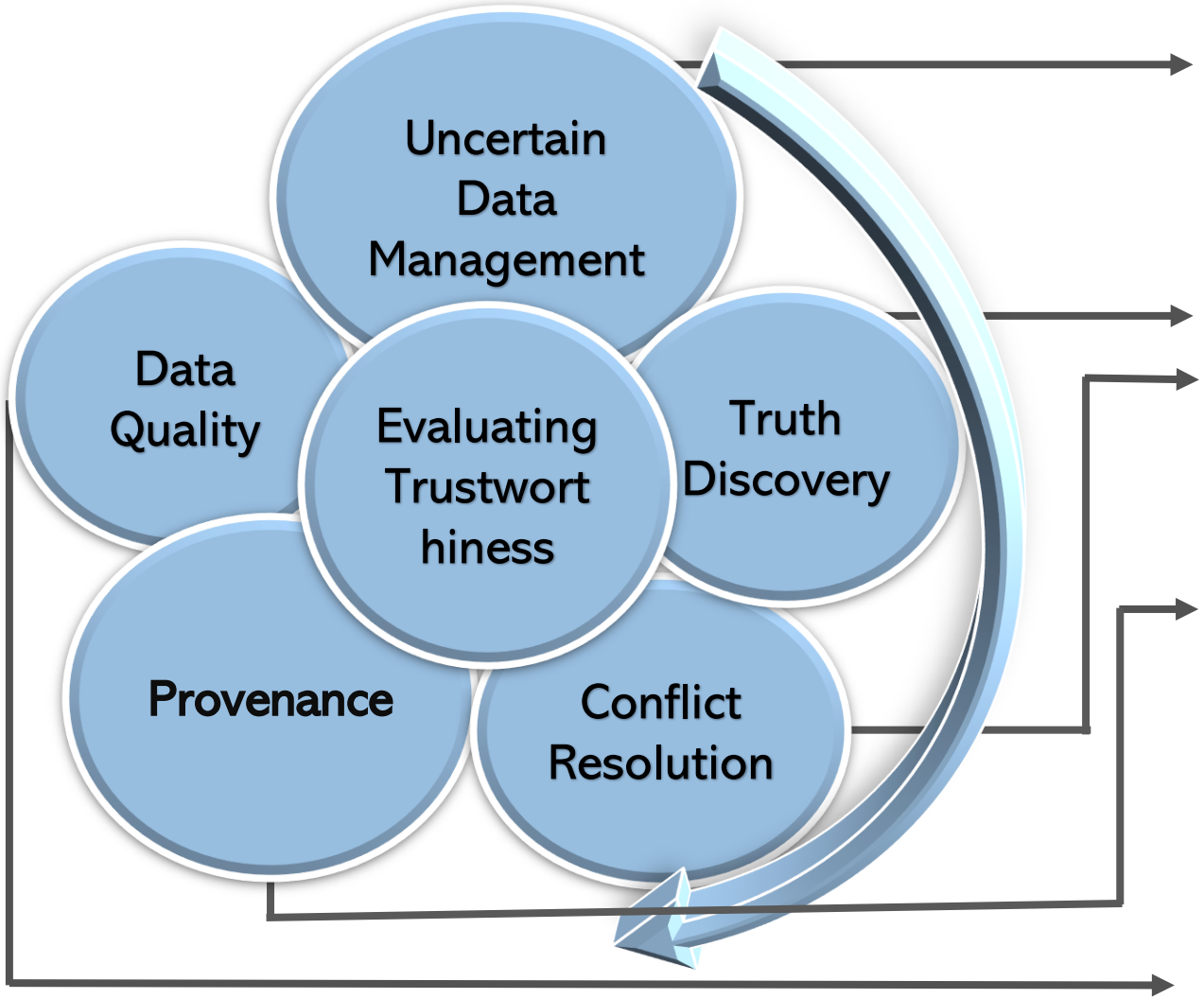
Applications of Evaluating trustworthiness

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

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

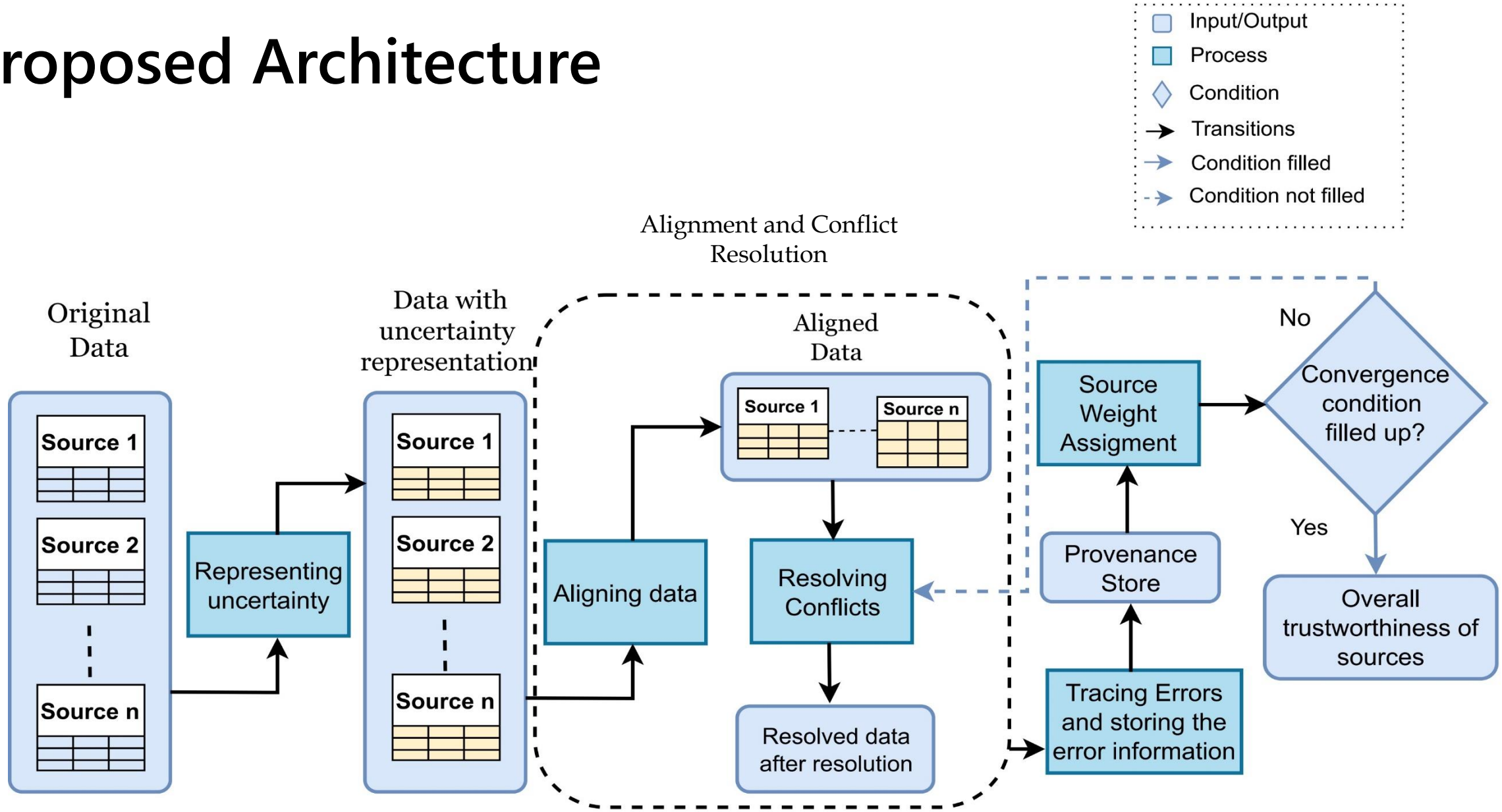
Limitations

- 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

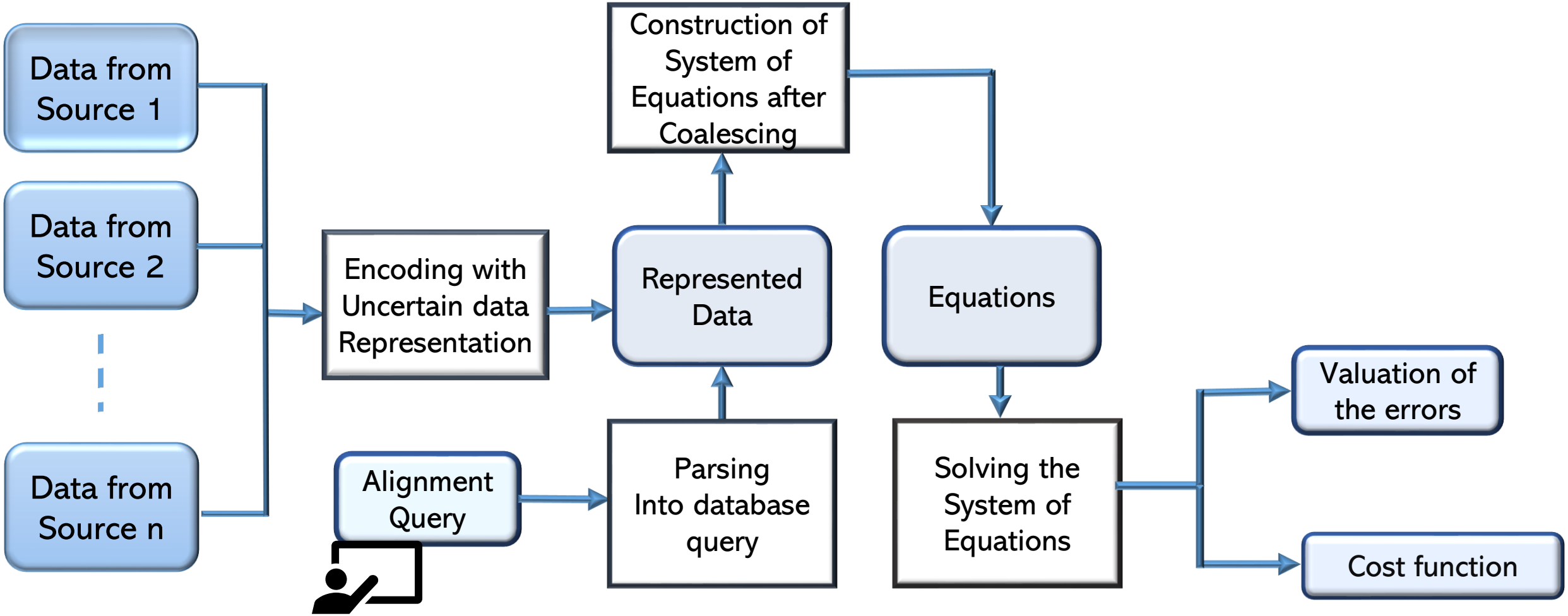
Proposed Architecture



ERIS

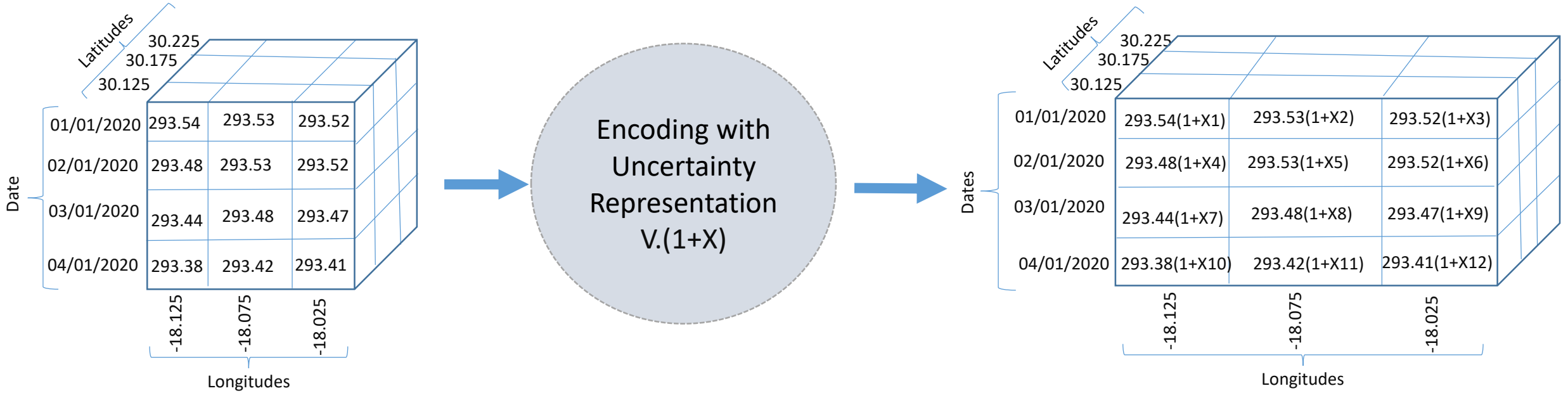
(A Prototype System)

Prototype Workflow



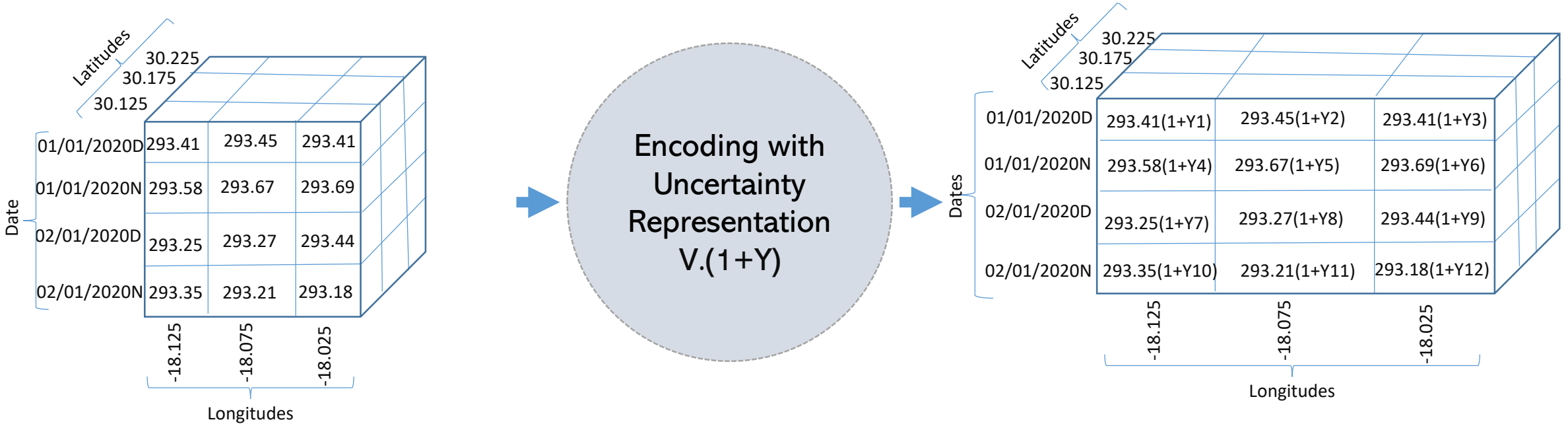
Running Example

Table MA from Source Copernicus

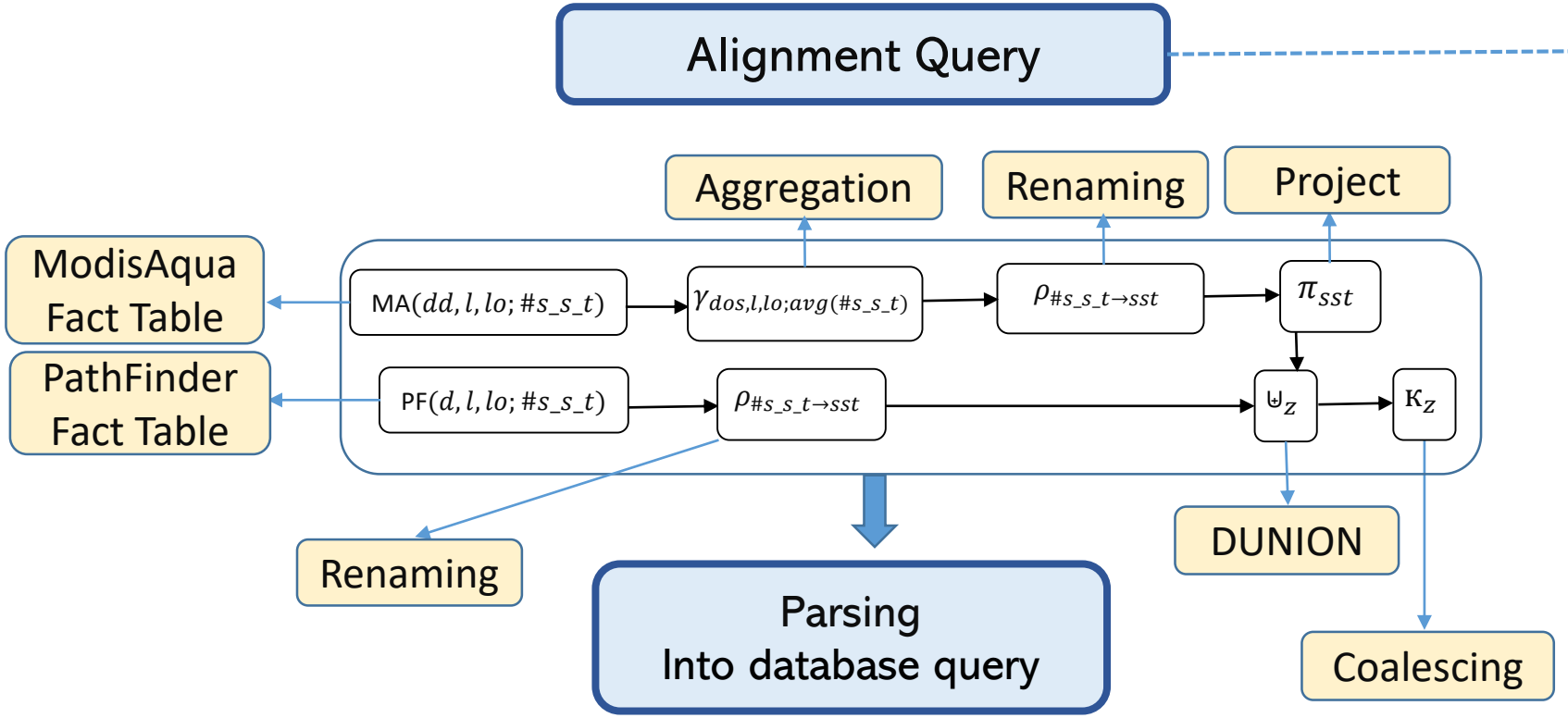


Running Example

Table PF from Source Climate



Running Example



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

t1:= SELECT dos, avg(s_s_t) as sst FROM CT JOIN semiday on climate_temperature.date = semiday.id WHERE date='20200101D' GROUP BY dos ORDER BY dos;
t2:= SELECT s_s_t as sst FROM COT WHERE date='20200101'
t3:= t1 UNION ALL t2

Running Example

Table MA from Source 1

Latitudes / Longitudes	-18.125	-18.075	-18.025
01/01/2020	293.54(1+X1)	293.53(1+X2)	293.52(1+X3)
02/01/2020	293.46(1+X4)	293.53(1+X5)	293.52(1+X6)
03/01/2020	293.44(1+X7)	293.48(1+X8)	293.47(1+X9)
04/01/2020	293.38(1+X10)	293.42(1+X11)	293.41(1+X12)

Table PF from Source 2

Dates / Latitudes	30.225	30.175	30.125
01/01/2020D	293.41(1+Y1)	293.45(1+Y2)	293.41(1+Y3)
01/01/2020N	293.58(1+Y4)	293.67(1+Y5)	293.69(1+Y6)
02/01/2020D	293.51(1+Y7)	293.47(1+Y8)	293.44(1+Y9)
02/01/2020N	293.39(1+Y10)	293.21(1+Y11)	293.18(1+Y12)

Longitudes: -18.125, -18.075, -18.025

Construction of System of Equations After Coalescing

Integrity Constraint

Sea_Surface Temperature	Aggregated SST from CT	SST from COT
01/01/2020	293.49	293.54
02/01/2020	293.45	293.46

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

Running Example

Table MA from Source 1

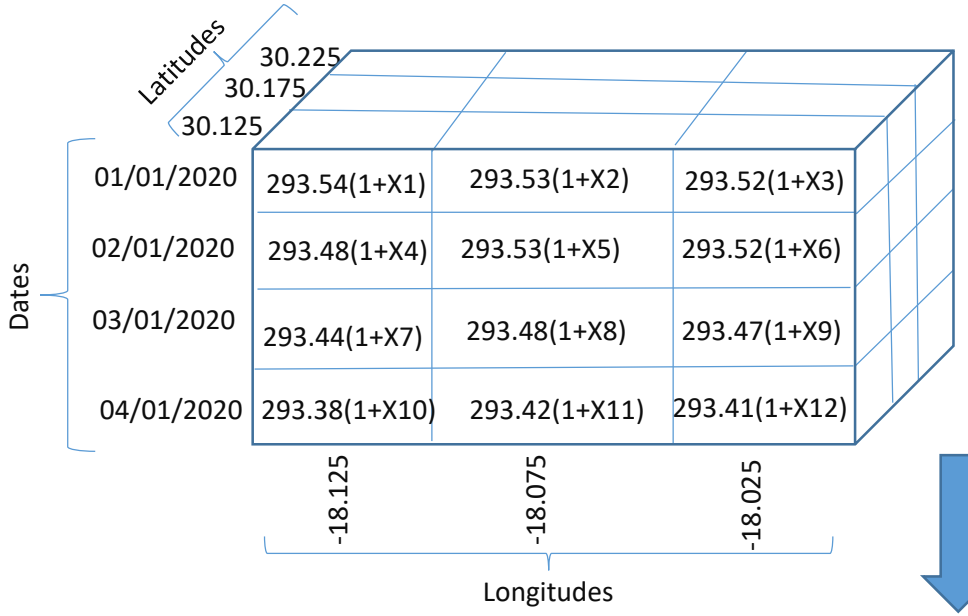
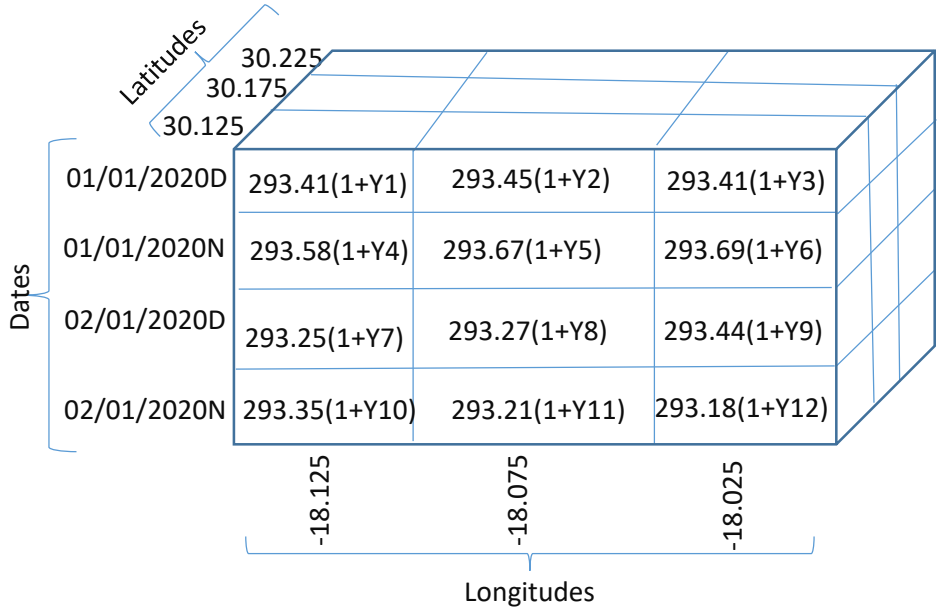


Table PF from Source 2

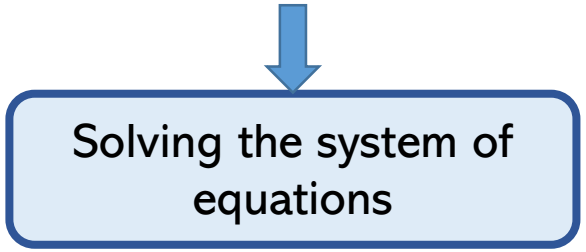


Construction of System of Equations after Coalescing

Date	Latitude = 30.125, Longitude = -18.125
01/01/2020	(293.41(1+Y1) + 293.58(1+Y4))/2 = 293.54 (1+X1)

Running Example

Date	Different Source
01/01/2020	(293.41(1+Y1) + 293.58(1+Y4))/2 = 293.54 (1+X1)



Variables	Y1	Y4	X1
Valuation	-1	1	0.000136
$\frac{\sum X_i^2}{n}$	0.667		

Use Case (Environmental Data)

Copernicus

- Environmental Variables – SST
- Daily Data
- Data Collection based on multiple sensors
- Spatial Resolution - 0.05
- Area Coverage- Mediterranean Sea

Moderate Resolution Imaging Spectroradiometer

- Environmental Variables – SST
- Data per Day and Night
- Data Collection based on multiple sensors
- Spatial Resolution- 0.04
- Area Coverage- Whole world

Climate Change Service

- Environmental Variables – SST
- Data per Day and Night
- Data Collection based on multiple sensors
- Resolution 0.05
- Area Coverage - Whole world

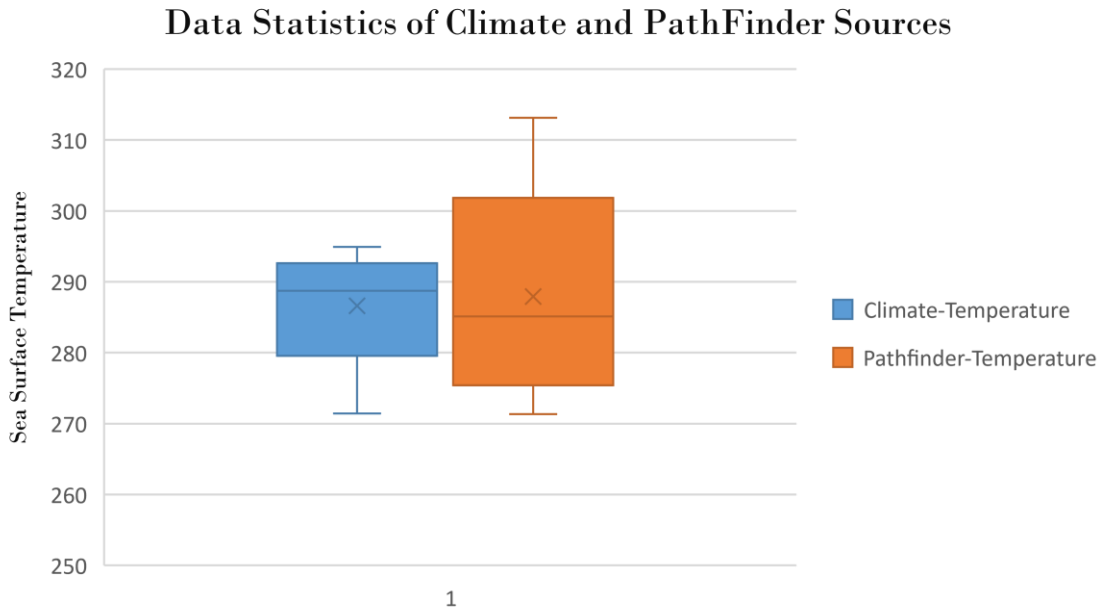
National Oceanic and Atmospheric Administration-PathFinder

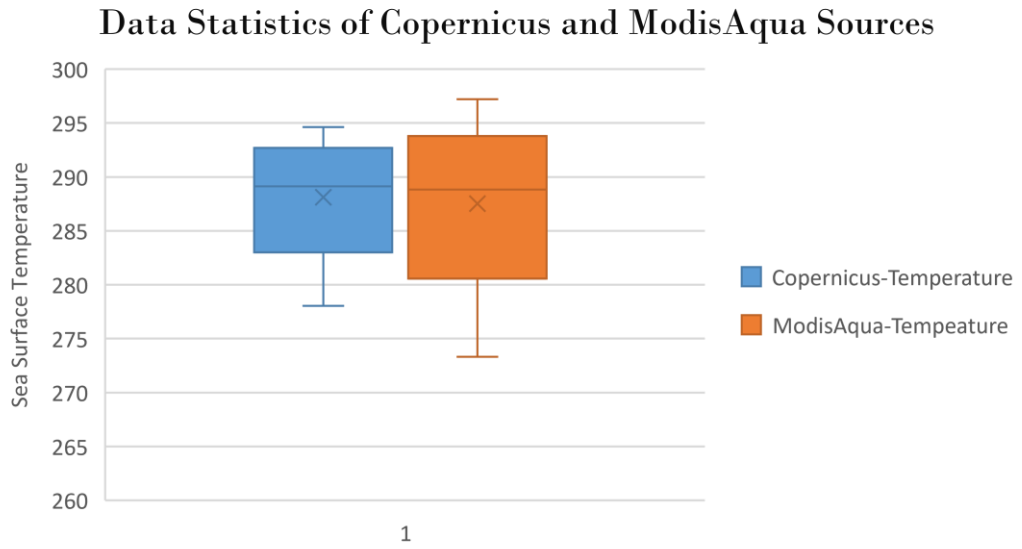
- Environmental Variables – SST
- Daily, Monthly data
- Data Collection based on multiple sensors
- Resolution 0.04
- Area Coverage- Whole world

Use Case (Data Statistics)

Sources	Values	Total Data	Number of Nulls (%)	Minimum	Maximum	Units
Copernicus	Temperature (Daily)	85,354,584	93.7%	278.029	294.630	Kelvin
Climate Data	Temperature (day and Night)	20,124,895	99.25%	271.440	294.910	Kelvin
Modis-Aqua Data	Temperature (Daily)	10,735,362	78.8%	0.179	24.064	Degree_C
Pathfinder Data	Temperature (Day and Night)	21,470,724	3.4%	271.349	313.14	Kelvin

Use Case (Data Statistics)





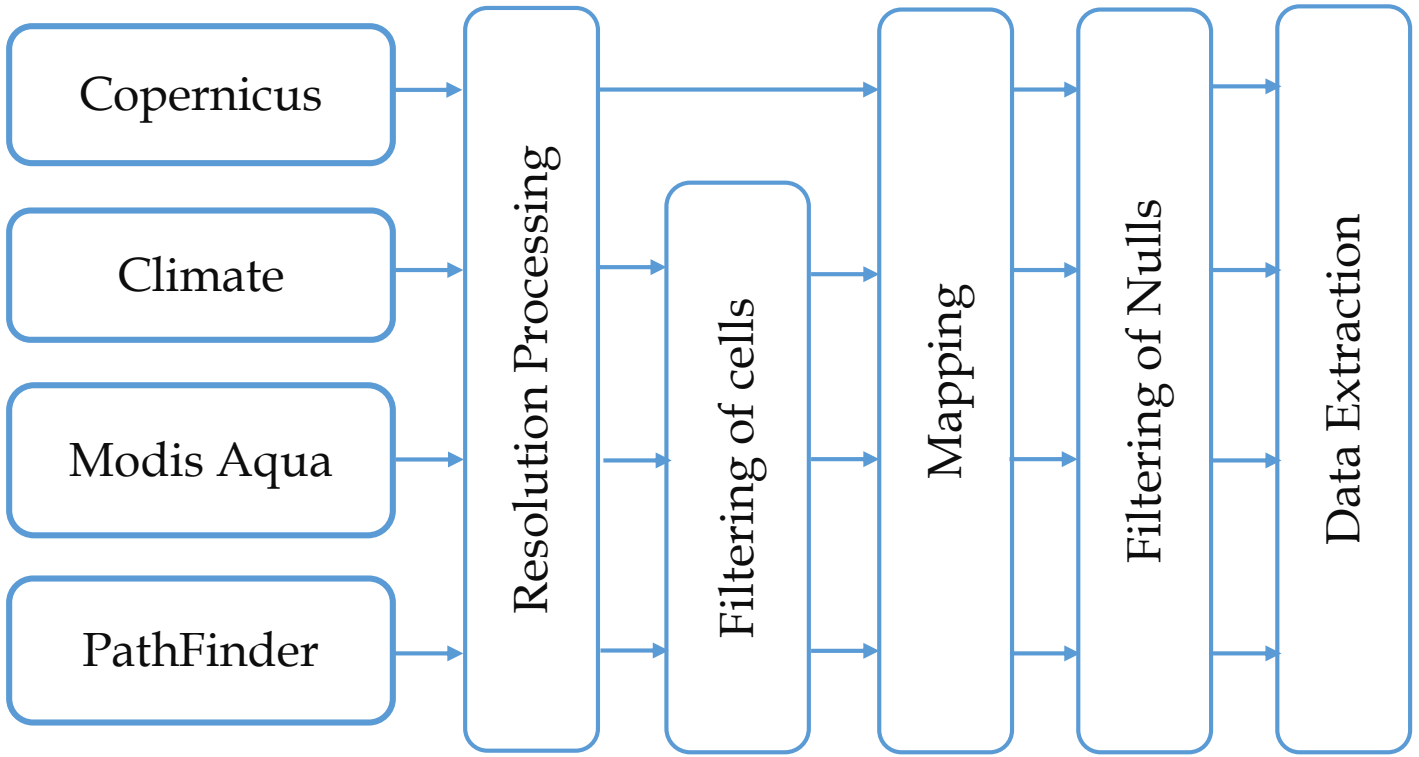
Climate- Min -271.44, Max -294.91 , Average – 288.668551177

PathFinder- Min – 271.35, Max – 313.14, Average – 284.88329

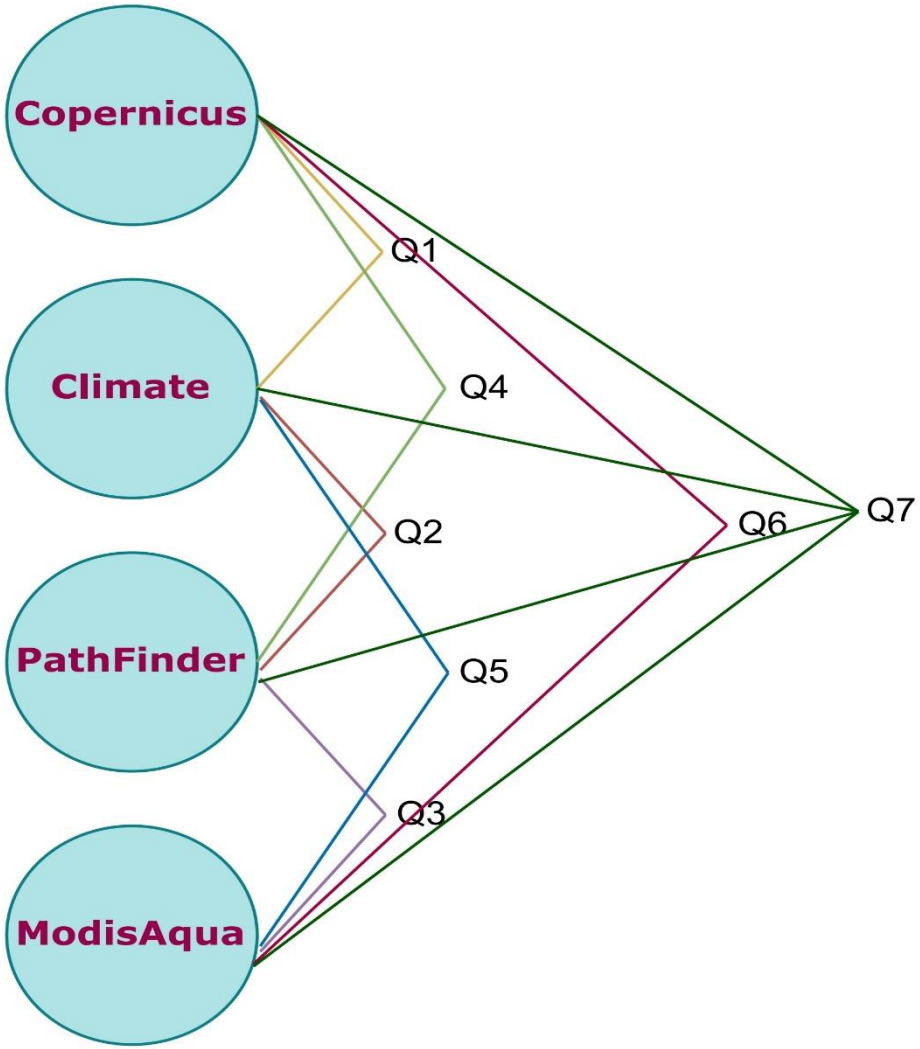
Copernicus- Min -278.03, Max – 294.63 , Average – 289.1963

Modis Aqua- Min – 273.33, Max – 297.215 , Average – 288.7300

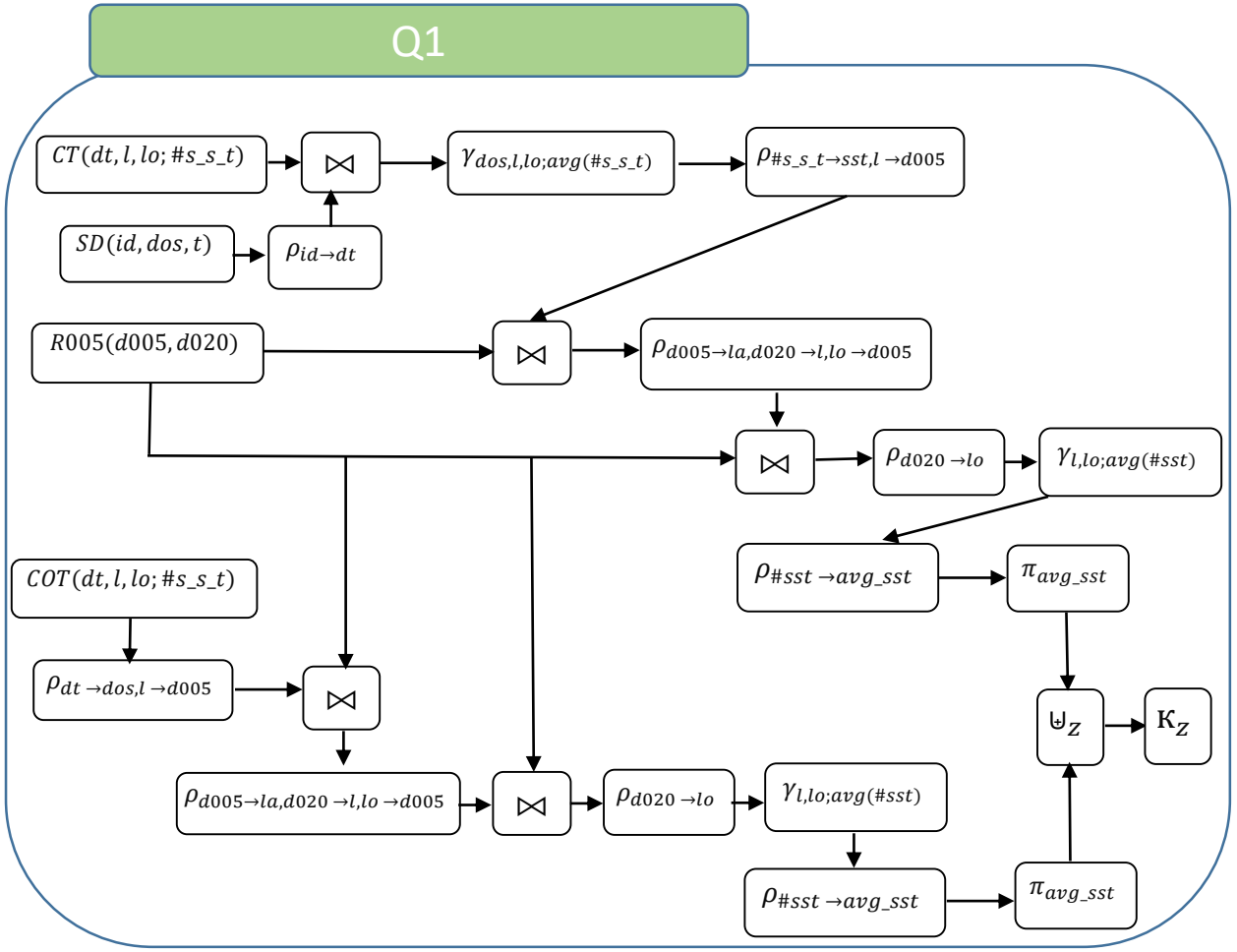
Use Case (Data Preparation)

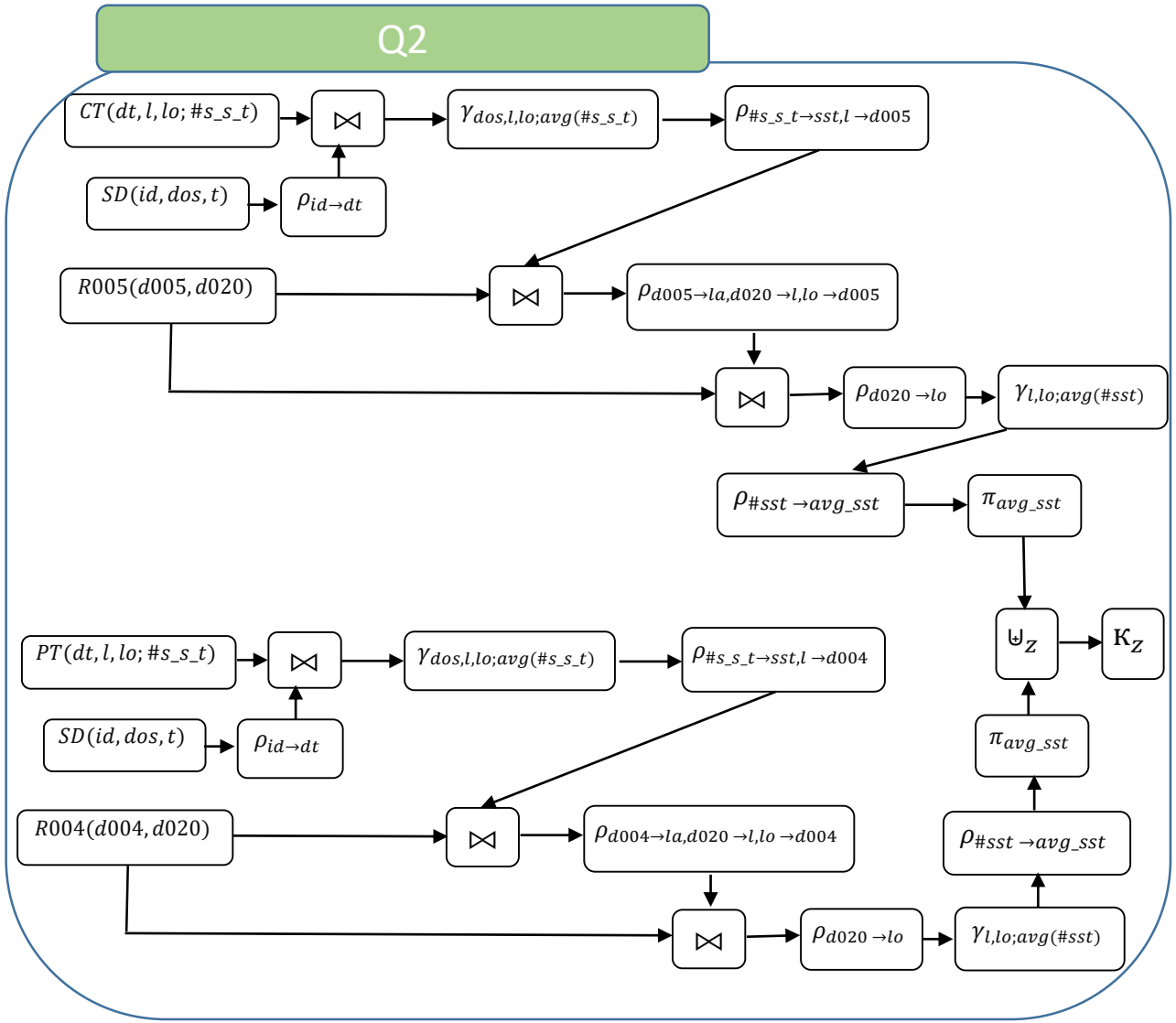


Use Case (Combination of Queries)

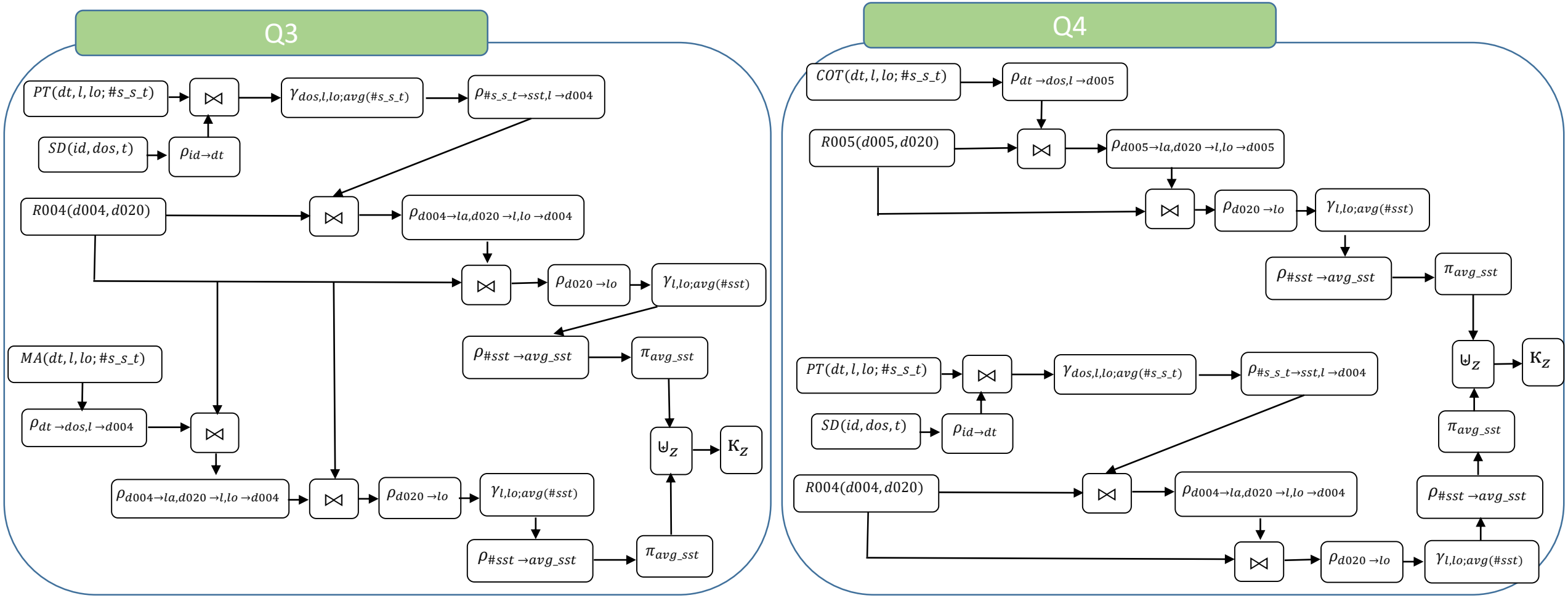


Use Case (Queries)

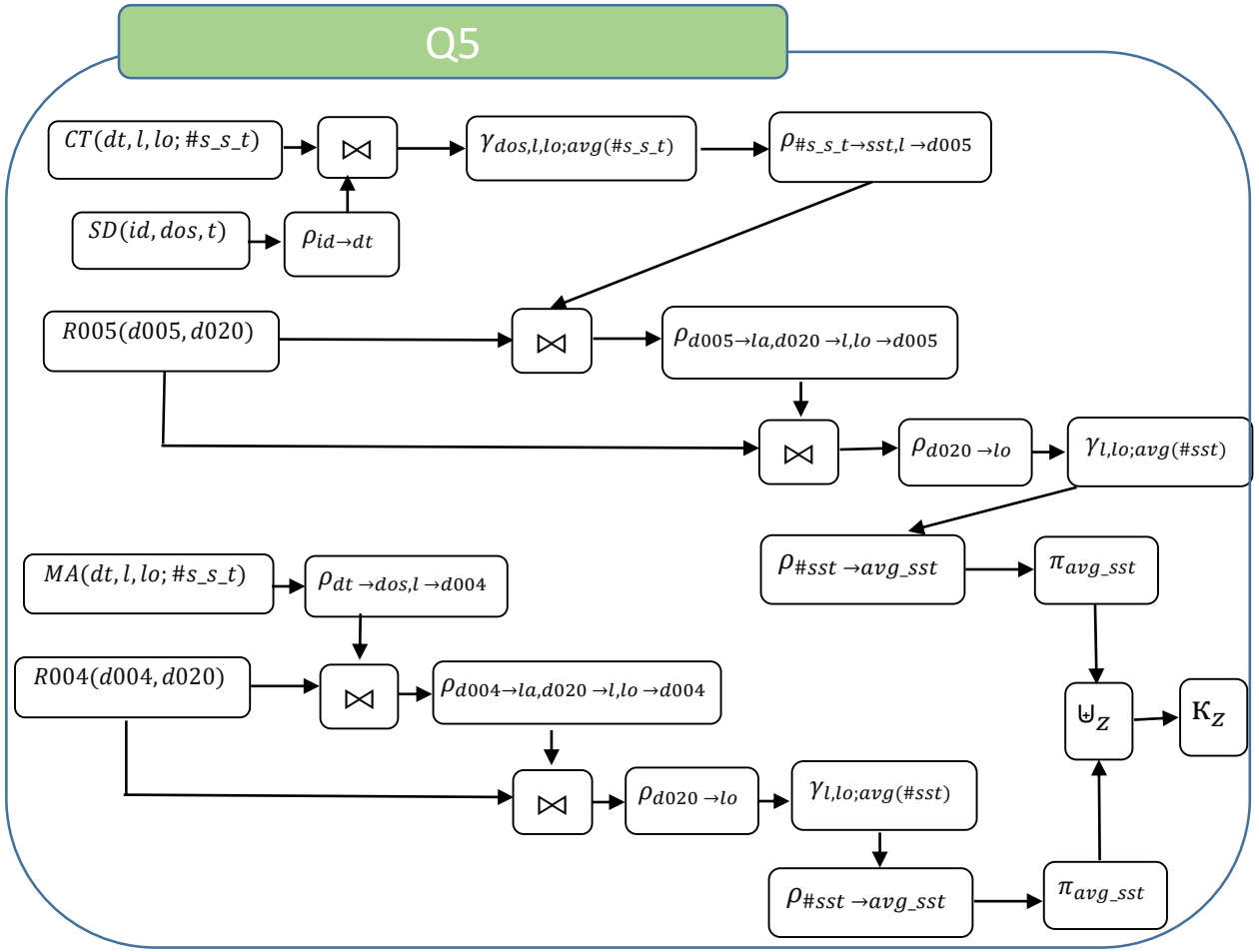


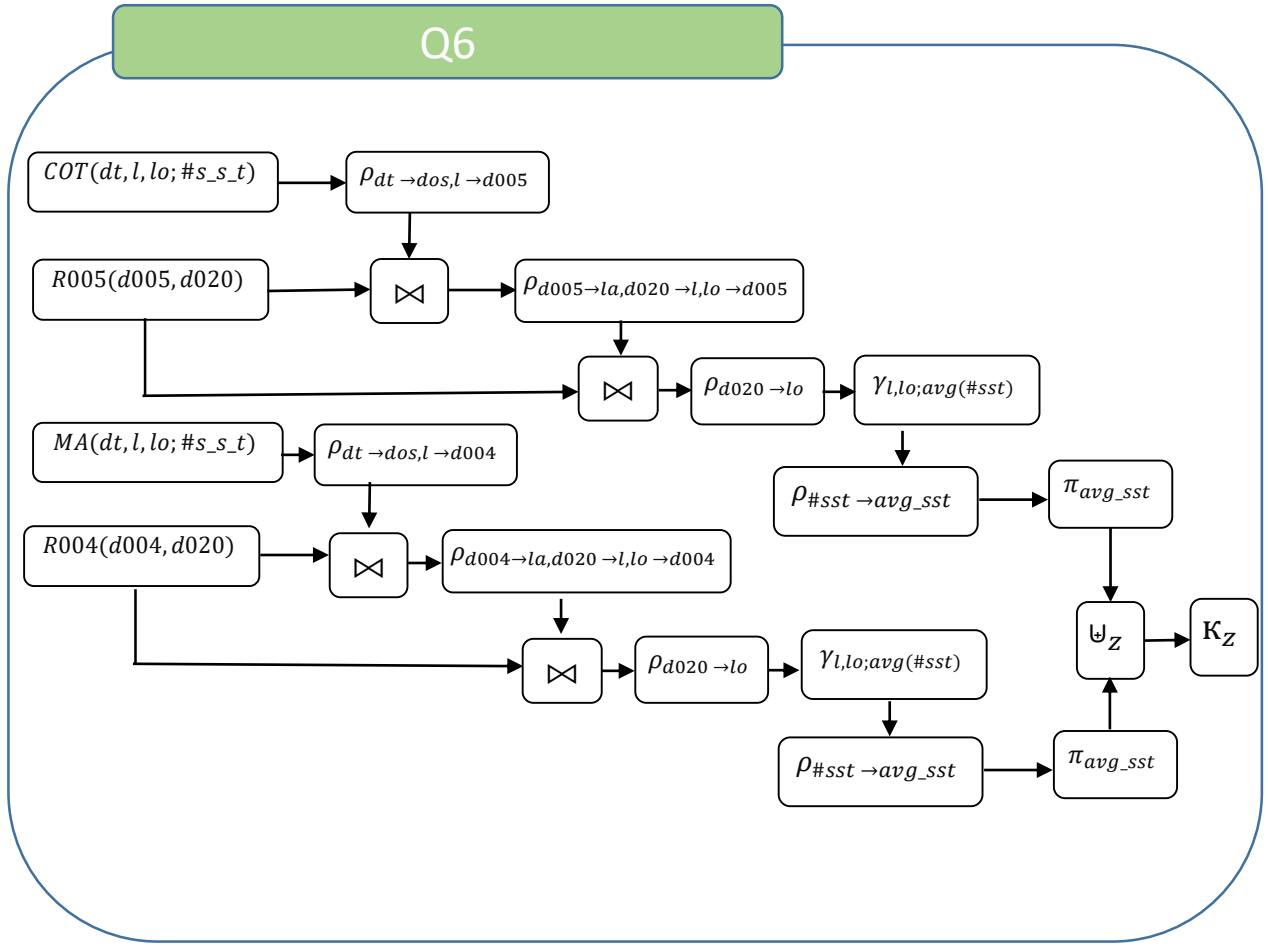


Use Case (Queries)

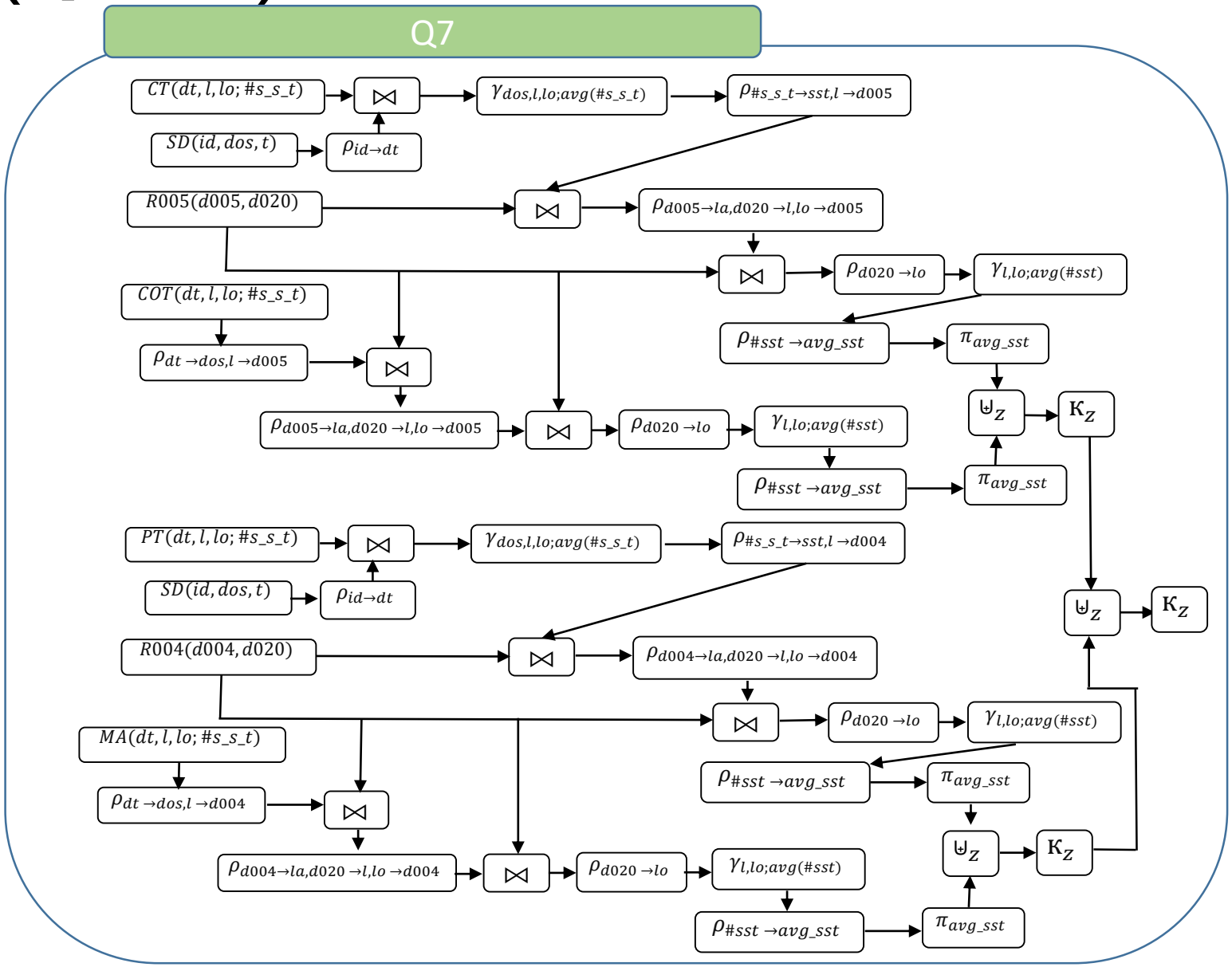


Use Case (Queries)



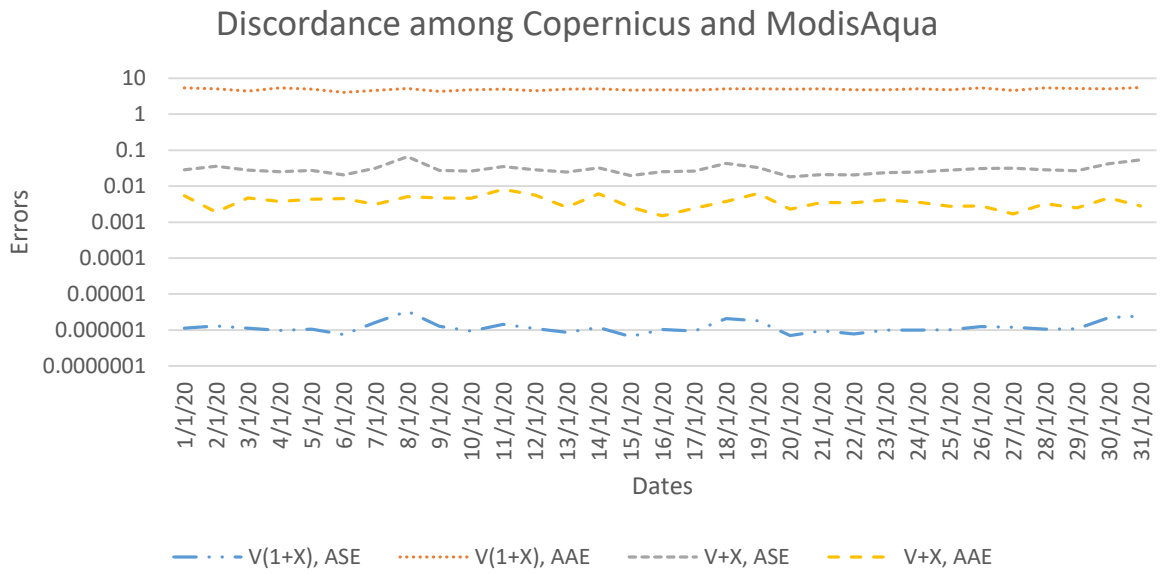


Use Case (Queries)



Results (Copernicus - ModisAqua)

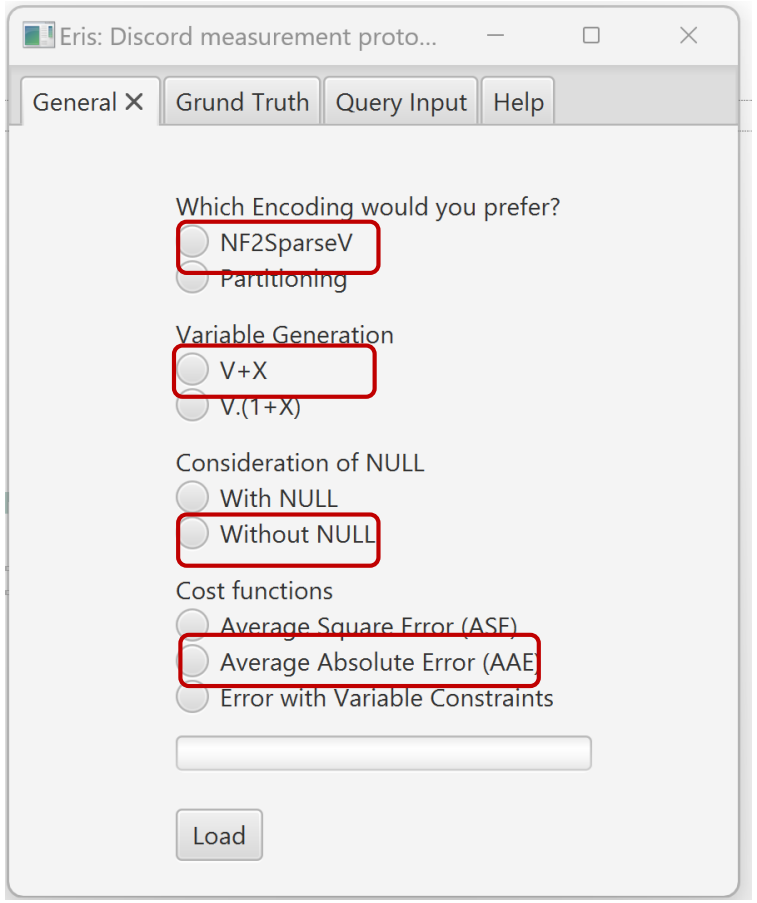
- Data Encoding Techniques
  - Nf2_sparsev
  - Partitioning
- Representation of Uncertainty
  - Variable Generation
    - V+X
    - V(1+X)
- Consideration of NULL
  - With NULL or
  - Without NULL
- Different Options to evaluate cost
  - Average Square Error (ASE)
  - Average Absolute Error (AAE)

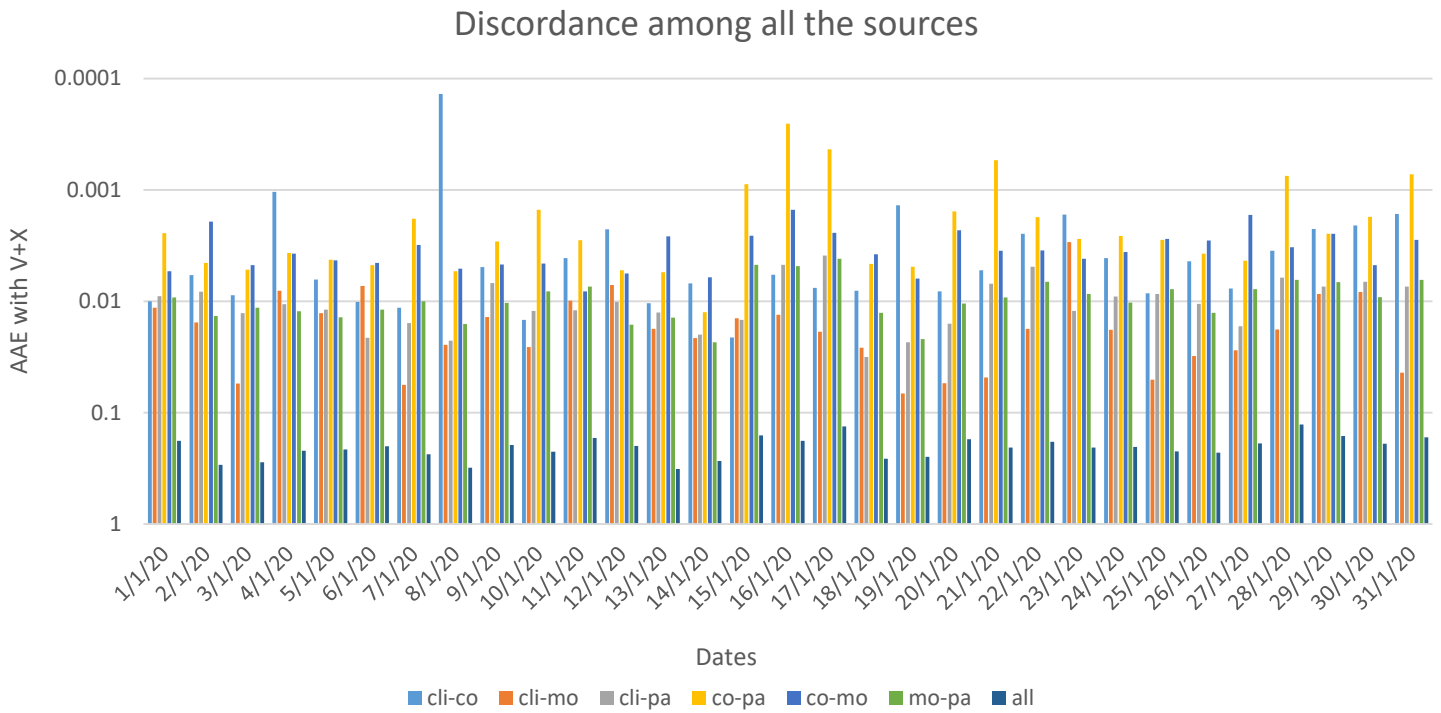


Absolute error is higher than the relative errors while the errors between the estimated values and the true values are relatively small when squared

Errors follow almost similar trend for same variable generation

Results (All sources)





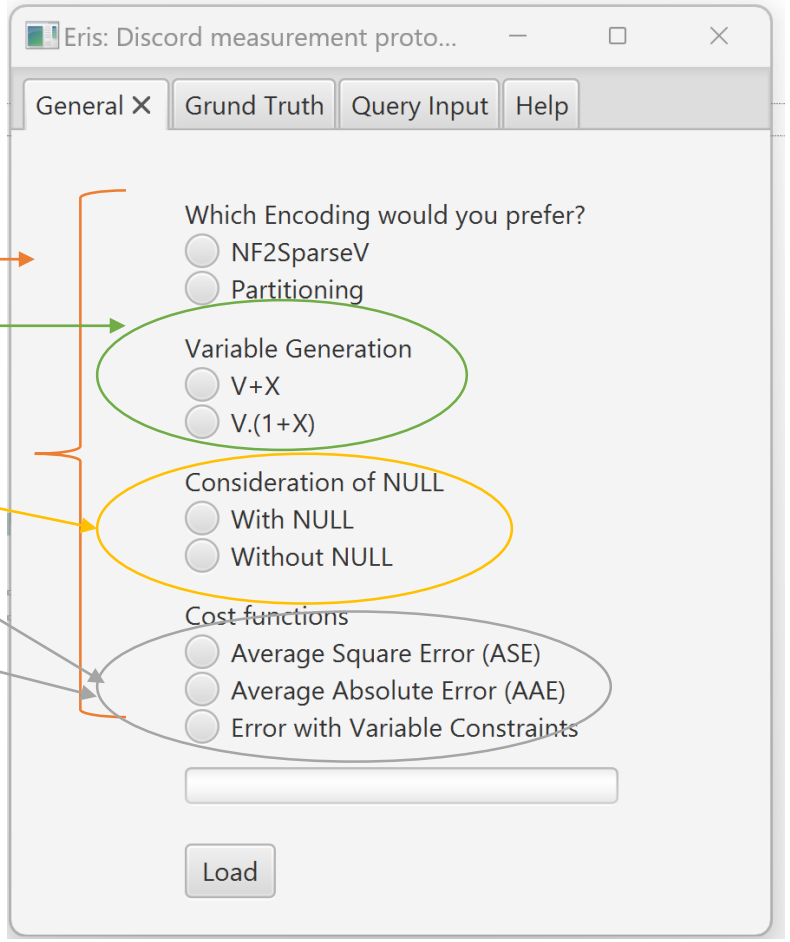
- Copernicus and pathfinder has lowest errors as they provide highest amount of data according to their resolution
- Combination of all the sources has highest error as they have provide data in different resolution and dates

Tasks to be Completed

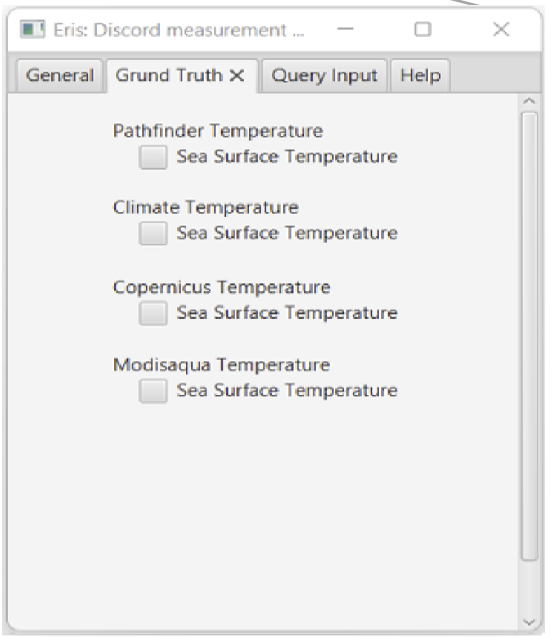
- Prototype Implementation
  - Creation of GUI
  - Implementing parsing of sequence of operations
  - Implementing Optionality of NULLs in Cost Function
  - Implementing Different Options to Generate Variables
  - Implementing Different Options to evaluate cost
  - Implementing Constrains in variables
- Prototype use:
  - Identifying a use case
  - Applying the prototype to the use case
- Publication:
  - Writing Demo Paper
  - Finding Potential Venue for submission

Prototype Implementation

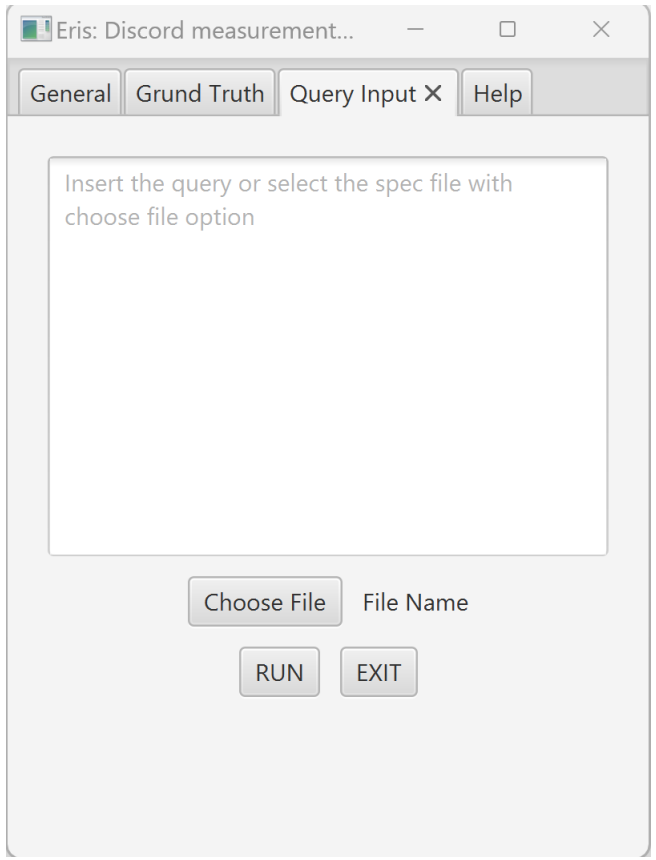
- Creation of GUI
- Implementing parsing of sequence of operations
- Implementing Different Options to Generate Variables
- Implementing Optionality of NULLs
- Implementing Different Options to evaluate cost
- Implementing Constrains in variables

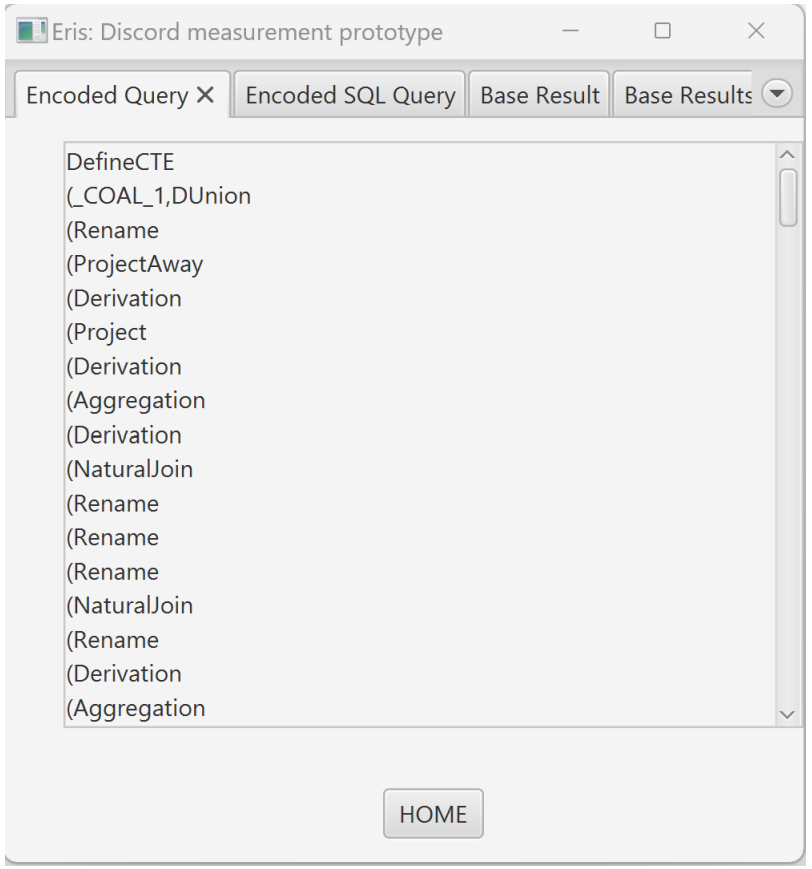


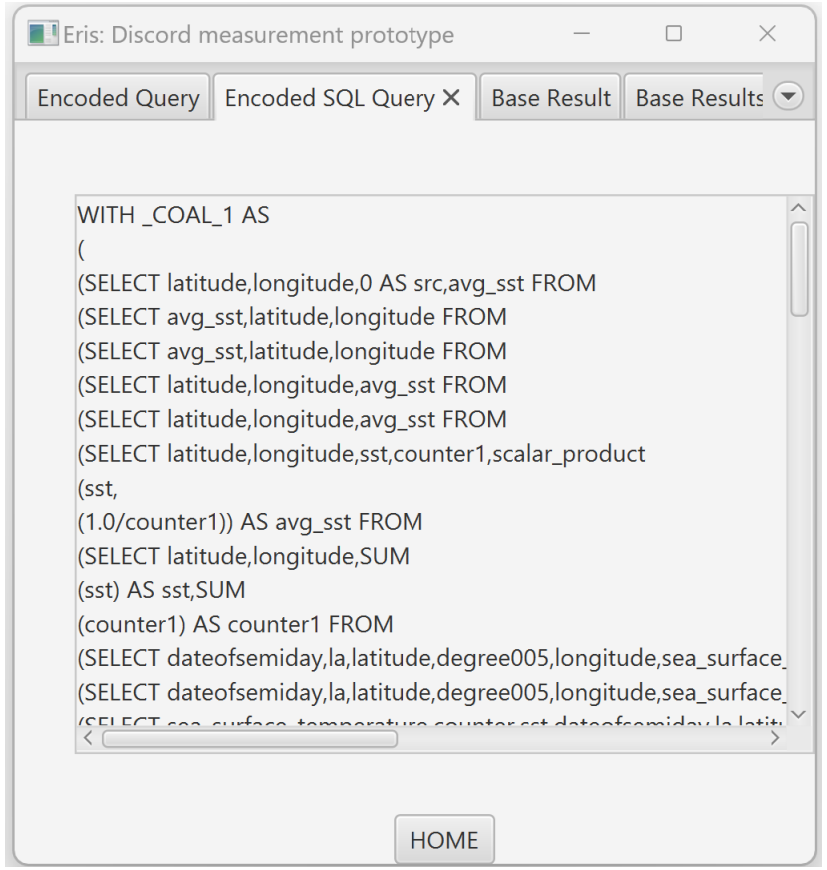
- Ground truth options



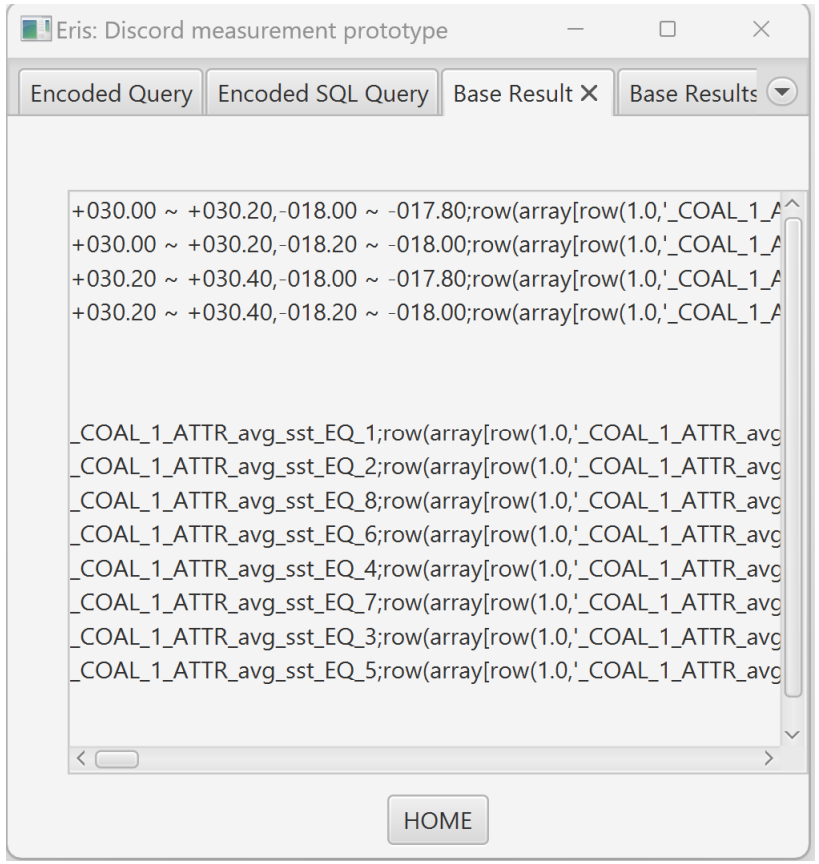
GUI outputs

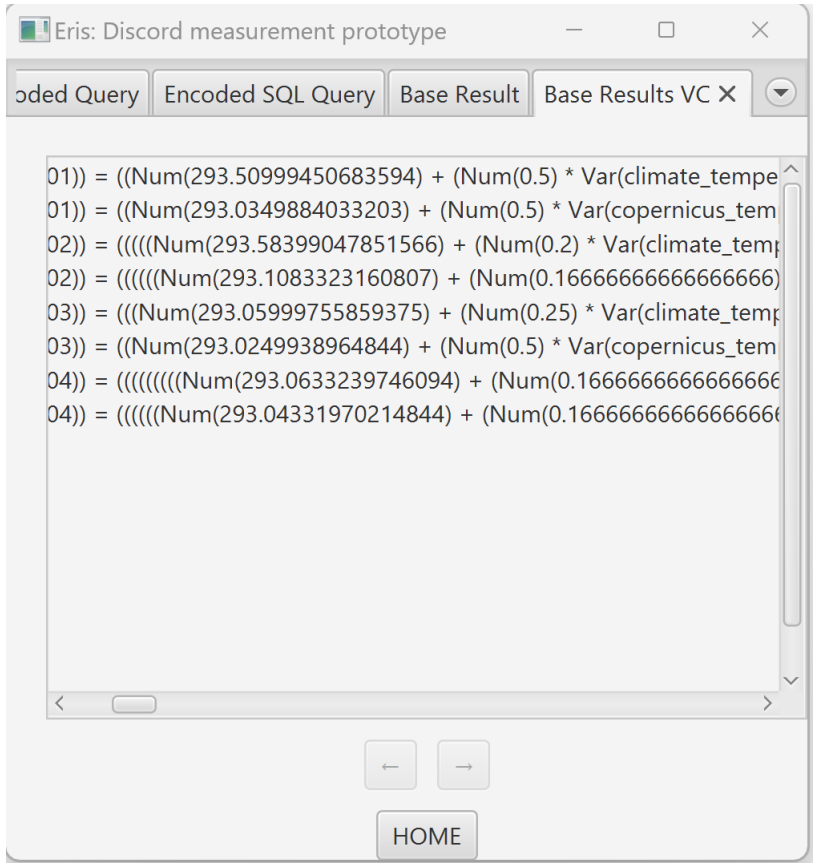

Eris: Discord measurement...
General
Grund Truth
Query Input
Help
Insert the query or select the spec file with choose file option
Choose File
File Name
RUN
EXIT

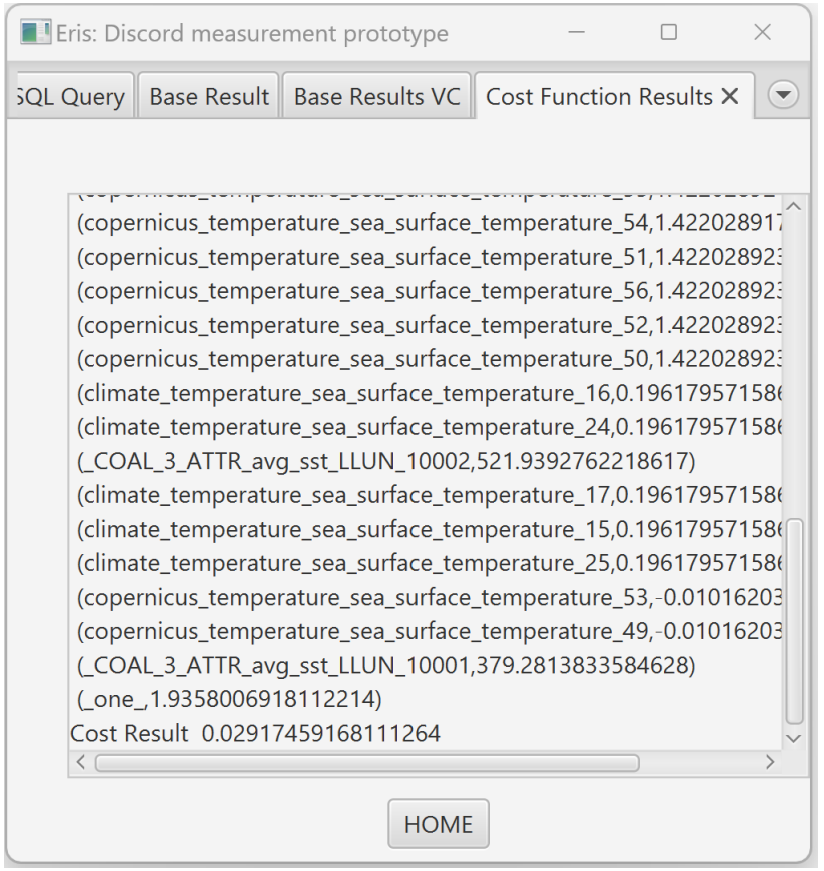

Eris: Discord measurement prototype
Encoded Query
Encoded SQL Query
Base Result
Base Results
DefineCTE
(_COAL_1,DUnion
(Rename
(ProjectAway
(Derivation
(Project
(Derivation
(Aggregation
(Derivation
(NaturalJoin
(Rename
(Rename
(Rename
(NaturalJoin
(Rename
(Derivation
(Aggregation
HOME


Eris: Discord measurement prototype
Encoded Query
Encoded SQL Query
Base Result
Base Results
WITH _COAL_1 AS
(
(SELECT latitude,longitude,0 AS src,avg_sst FROM
(SELECT avg_sst,latitude,longitude FROM
(SELECT avg_sst,latitude,longitude FROM
(SELECT latitude,longitude,avg_sst FROM
(SELECT latitude,longitude,avg_sst FROM
(SELECT latitude,longitude,sst,counter1,scalar_product
(sst,
(1.0/counter1)) AS avg_sst FROM
(SELECT latitude,longitude,SUM
(sst) AS sst,SUM
(counter1) AS counter1 FROM
(SELECT dateofsemiday,la,latitude,degree005,longitude,sea_surface_
(SELECT dateofsemiday,la,latitude,degree005,longitude,sea_surface_
HOME

GUI outputs


Eris: Discord measurement prototype
Encoded Query
Encoded SQL Query
Base Result
Base Results
+030.00 ~ +030.20,-018.00 ~ -017.80;row(array[row(1.0,'_COAL_1_A
+030.00 ~ +030.20,-018.20 ~ -018.00;row(array[row(1.0,'_COAL_1_A
+030.20 ~ +030.40,-018.00 ~ -017.80;row(array[row(1.0,'_COAL_1_A
+030.20 ~ +030.40,-018.20 ~ -018.00;row(array[row(1.0,'_COAL_1_A
_COAL_1_ATTR_avg_sst_EQ_1;row(array[row(1.0,'_COAL_1_ATTR_avg
_COAL_1_ATTR_avg_sst_EQ_2;row(array[row(1.0,'_COAL_1_ATTR_avg
_COAL_1_ATTR_avg_sst_EQ_8;row(array[row(1.0,'_COAL_1_ATTR_avg
_COAL_1_ATTR_avg_sst_EQ_6;row(array[row(1.0,'_COAL_1_ATTR_avg
_COAL_1_ATTR_avg_sst_EQ_4;row(array[row(1.0,'_COAL_1_ATTR_avg
_COAL_1_ATTR_avg_sst_EQ_7;row(array[row(1.0,'_COAL_1_ATTR_avg
_COAL_1_ATTR_avg_sst_EQ_3;row(array[row(1.0,'_COAL_1_ATTR_avg
_COAL_1_ATTR_avg_sst_EQ_5;row(array[row(1.0,'_COAL_1_ATTR_avg
HOME


Eris: Discord measurement prototype
oded Query
Encoded SQL Query
Base Result
Base Results VC
01)) = ((Num(293.50999450683594) + (Num(0.5) * Var(climate_tempe
01)) = ((Num(293.0349884033203) + (Num(0.5) * Var(copernicus_tem
02)) = (((((Num(293.58399047851566) + (Num(0.2) * Var(climate_temp
02)) = ((((((Num(293.1083323160807) + (Num(0.16666666666666666)
03)) = (((Num(293.05999755859375) + (Num(0.25) * Var(climate_temp
03)) = ((Num(293.0249938964844) + (Num(0.5) * Var(copernicus_tem
04)) = (((((((((Num(293.0633239746094) + (Num(0.1666666666666666
04)) = ((((((Num(293.04331970214844) + (Num(0.1666666666666666
←
→
HOME


Eris: Discord measurement prototype
SQL Query
Base Result
Base Results VC
Cost Function Results
(copernicus_temperature_sea_surface_temperature_54,1.422028917
(copernicus_temperature_sea_surface_temperature_51,1.422028923
(copernicus_temperature_sea_surface_temperature_56,1.422028923
(copernicus_temperature_sea_surface_temperature_52,1.422028923
(copernicus_temperature_sea_surface_temperature_50,1.422028923
(climate_temperature_sea_surface_temperature_16,0.196179571586
(climate_temperature_sea_surface_temperature_24,0.196179571586
(_COAL_3_ATTR_avg_sst_LLUN_10002,521.9392762218617)
(climate_temperature_sea_surface_temperature_17,0.196179571586
(climate_temperature_sea_surface_temperature_15,0.196179571586
(climate_temperature_sea_surface_temperature_25,0.196179571586
(copernicus_temperature_sea_surface_temperature_53,-0.01016203
(copernicus_temperature_sea_surface_temperature_49,-0.01016203
(_COAL_3_ATTR_avg_sst_LLUN_10001,379.2813833584628)
(_one_,1.9358006918112214)
Cost Result 0.02917459168111264
HOME

Thank you for attention

Any Questions?