

SYDAG

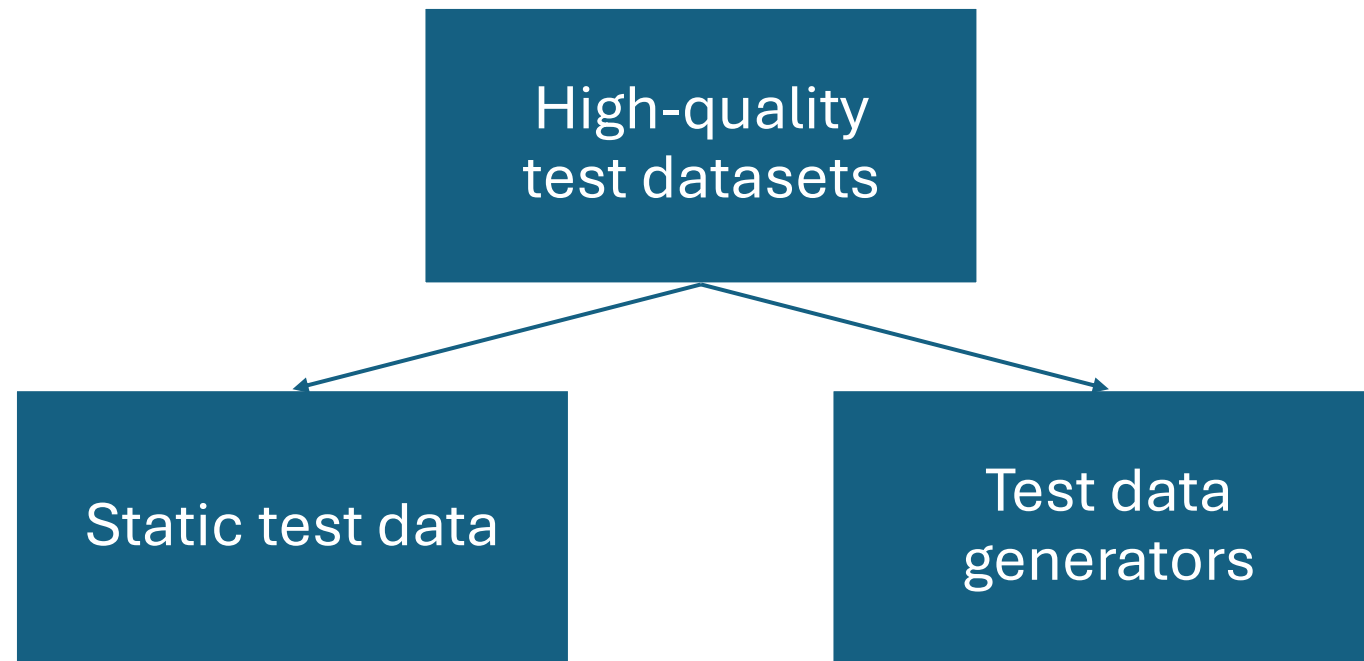
A Synthetic Dataset Generator for Data Integration

Anne Marschner

Supervisor: Prof. Dr. Thorsten Papenbrock

Motivation

Development of data integration tools:



Related Work

Real world datasets

- Thalia ¹⁾
- DIBS ²⁾
- Alaska ³⁾

Dataset generators

- DaPo⁺ ⁴⁾
- EMBench++ ⁵⁾
- eTuner ⁶⁾
- Valentine ⁷⁾

Features

- Duplicates
- Data noise
- Schema noise
- Merging
- Splitting

1) J. Hammer, M. Stonebraker, O. Topsakal, THALIA: Test harness for the assessment of legacy information integration approaches, in: Proceedings of the International Conference on Data Engineering (ICDE), 2005, pp. 485–486.
2) A. M. Cabrera, C. J. Faber, K. Cepeda, R. Derber, C. Epstein, J. Zheng, R. K. Cytron, R. D. Chamberlain, DIBS: A data integration benchmark suite, in: Proceeding of the ACM/SPEC International Conference on Performance Engineering Companion (ICPE), 2018, pp. 25–28.
3) V. Crescenzi, A. D. Angelis, D. Firmani, M. Mazzei, P. Merialdo, F. Piai, D. Srivastava, Alaska: A flexible benchmark for data integration tasks, 2021.
4) F. Panse, W. Wingerath, B. Wollmer, Towards scalable generation of realistic test data for duplicate detection, 2023.
5) E. Ioannou, Y. Velegrakis, EMBench++: Data for a thorough benchmarking of matching-related methods, in: Semantic Web, volume 10, 2019, pp. 435–450.
6) Y. Lee, M. Sayyadian, A. Doan, A. S. Rosenthal, ETuner: Tuning schema matching software using synthetic scenarios, in: VLDB Journal, volume 16, 2007, pp. 97–122.
7) C. Koutras, K. Psarakis, G. Siachamis, A. Ionescu, M. Fragkoulis, A. Bonifati, A. Katsifodimos, Valentine in action: Matching tabular data at scale, in: Proceedings of the VLDB Endowment, volume 14, 2021, pp. 2871–2874.

Synthetic Dataset Generator

SYDAG

input dataset &
configuration



determine keys of
relation



split relation into
new datasets



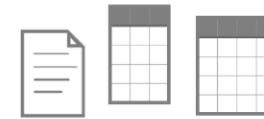
normalize datasets



insert schema noise



output datasets,
mapping & metadata



generate metadata
and mapping



shuffle columns or
rows

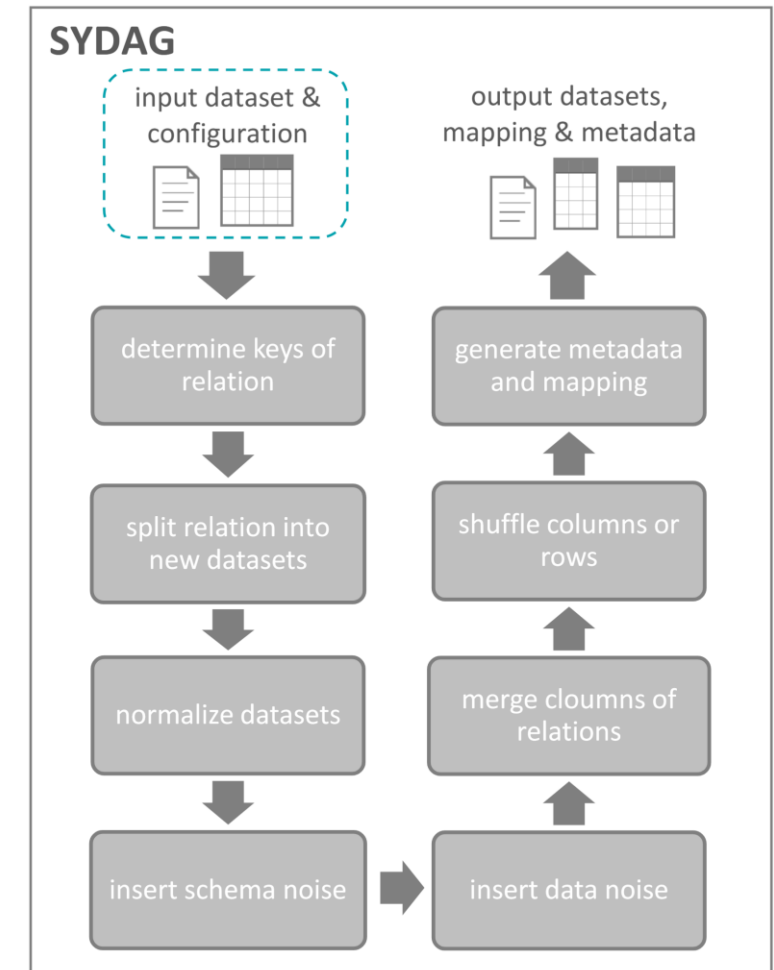
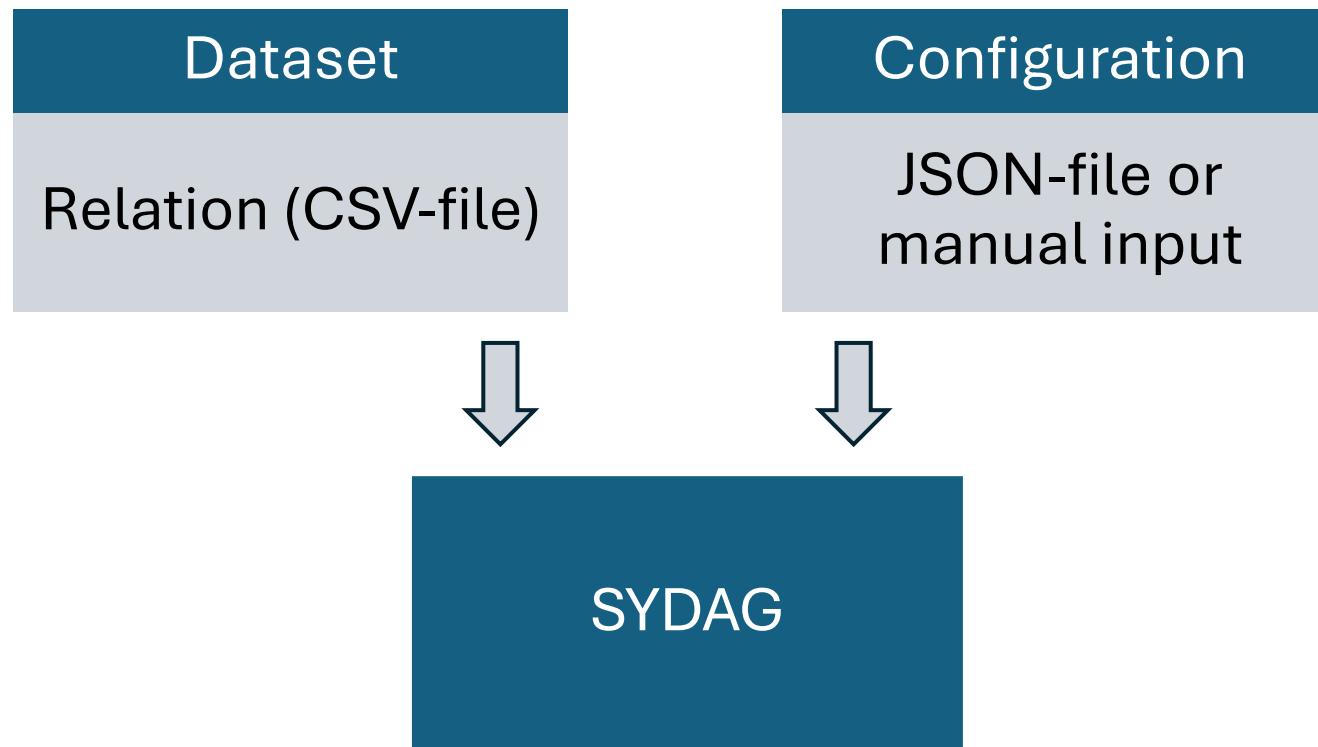


merge cloumns of
relations



insert data noise

Input

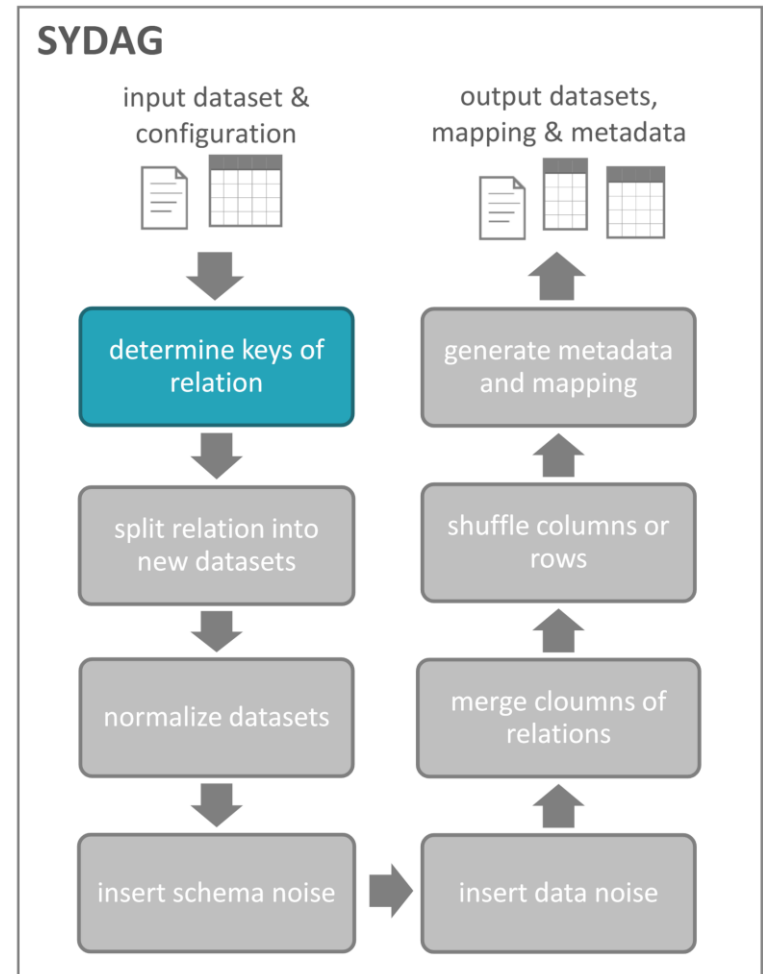
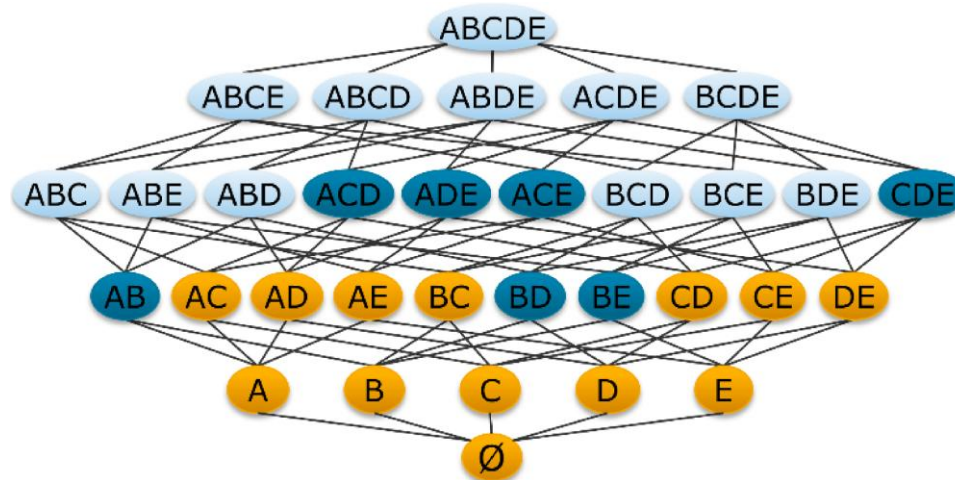


Key Determination

HyUCC ¹⁾

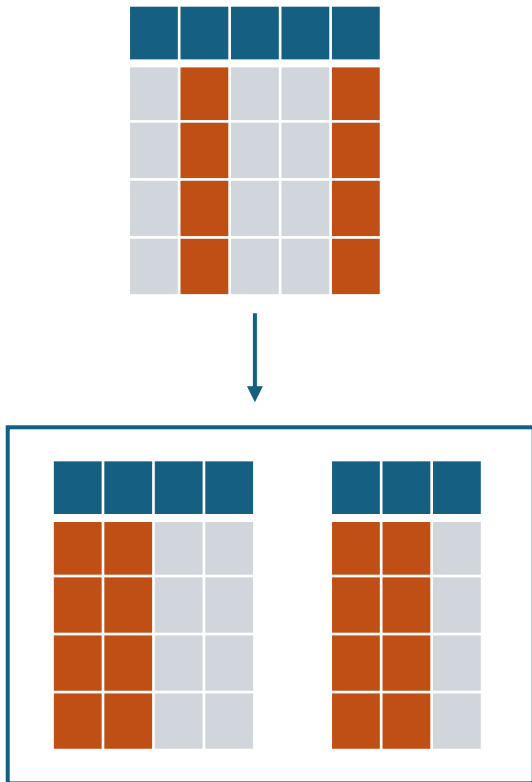
– Key profiling algorithm

UCC discovery in a powerset lattice:



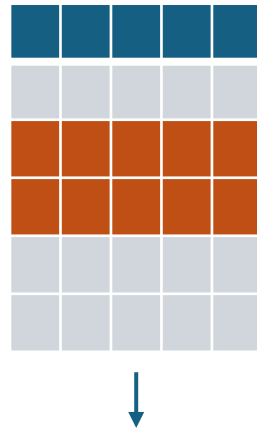
Splitting

Vertical split:

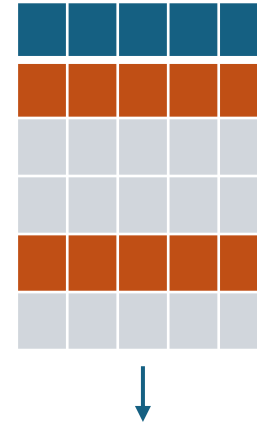


Horizontal split:

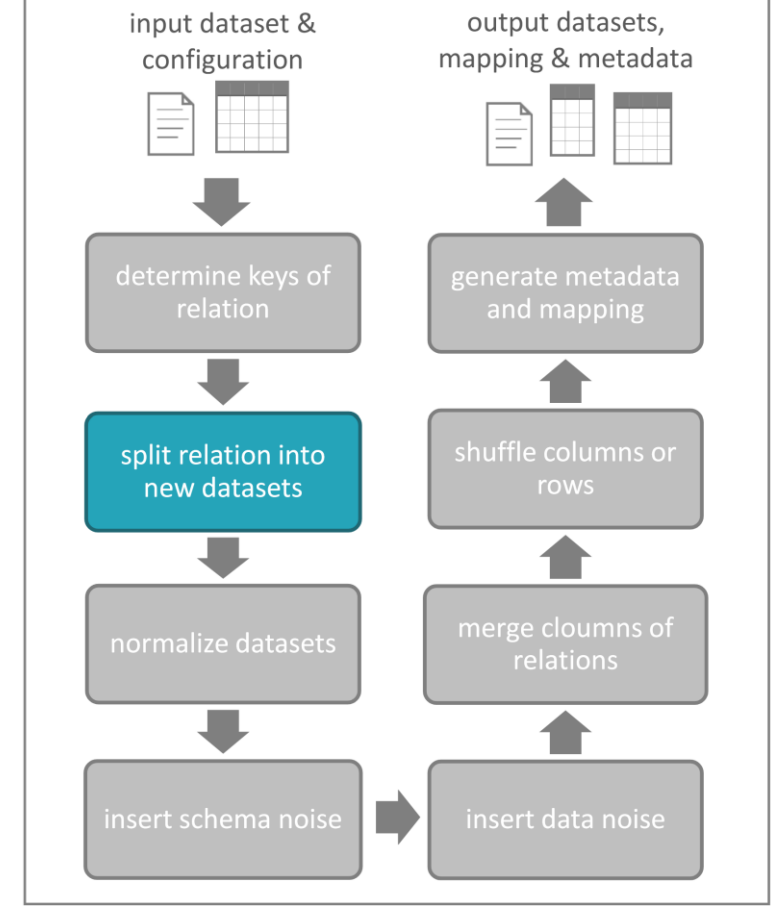
Block overlap:



Random overlap:



SYDAG

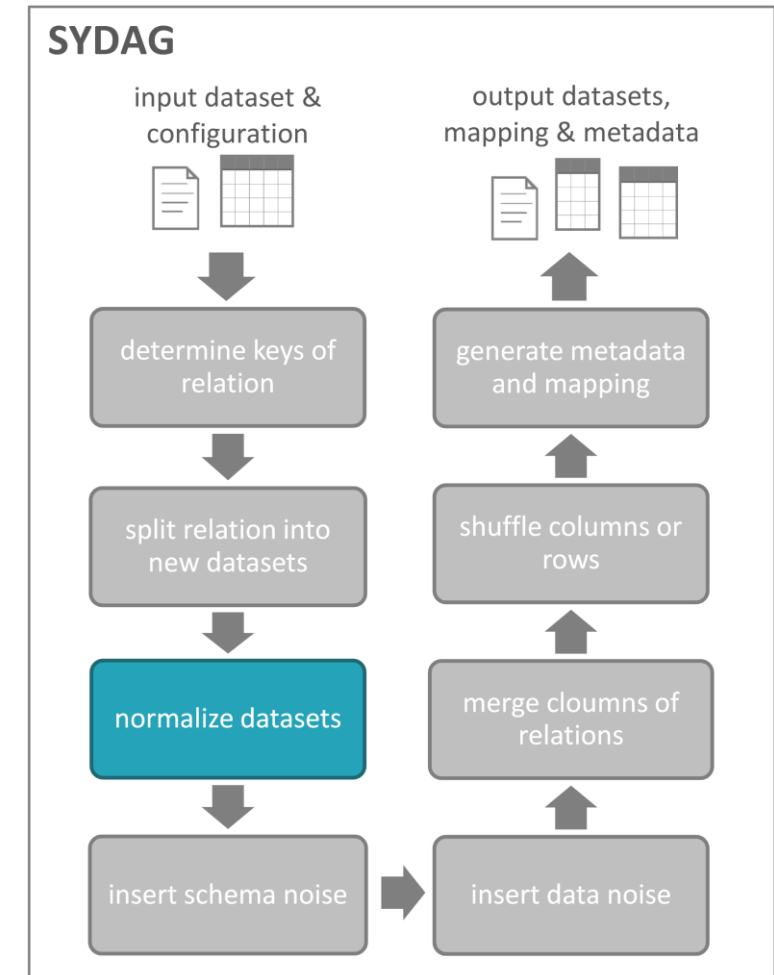
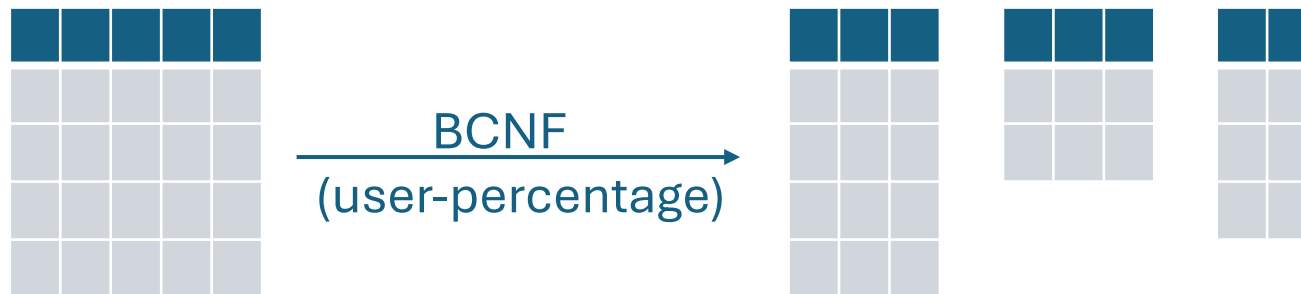


Normalization

Normalize ¹⁾

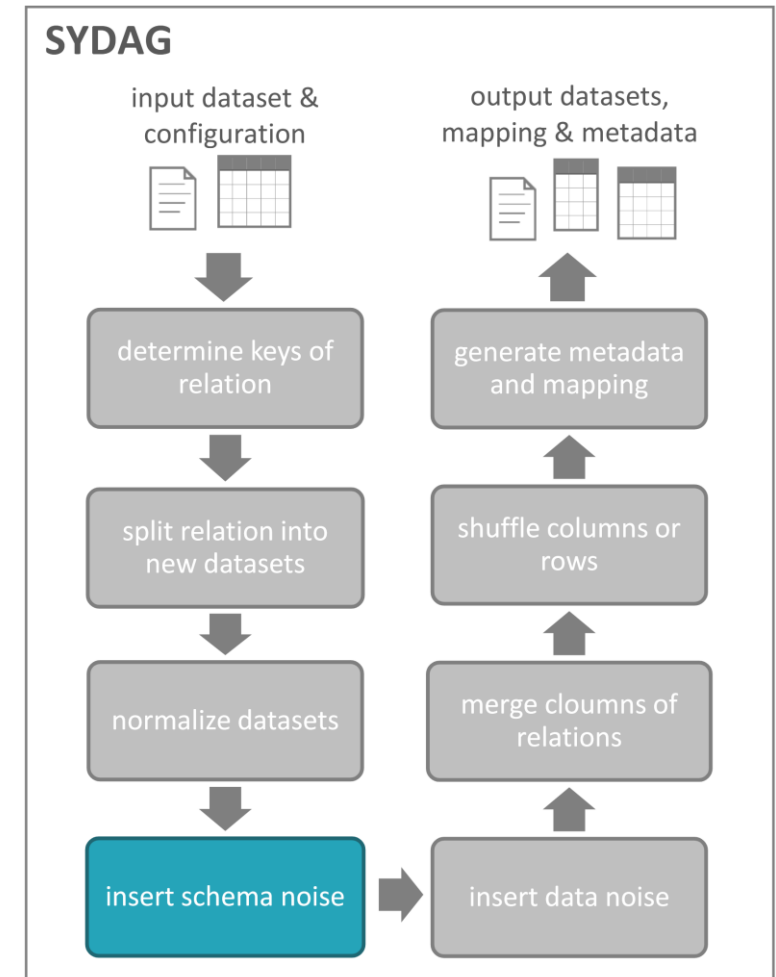
– schema normalization algorithm

For each dataset:



Schema Noise

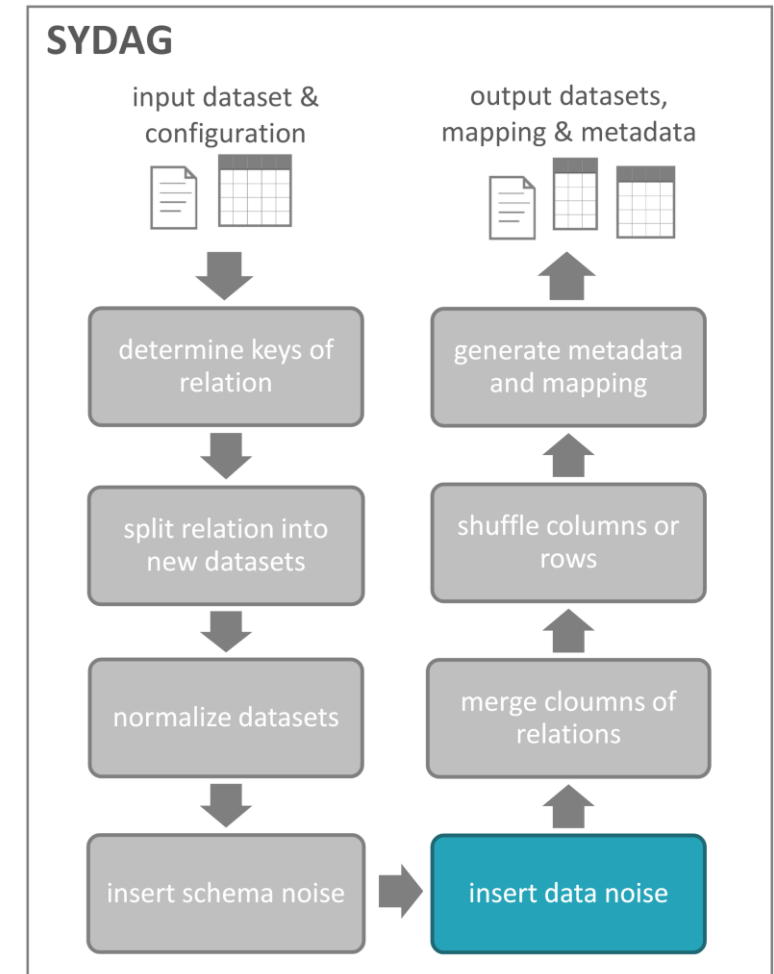
Method	„last name“
Remove vowels	„lst nm“
Abbreviate first letters	„L.N.“
Shorten to random length	„LaNam“
Shuffle letters	„matlensa“
Shuffle words	„name last“
Random string	„ko7gik“
Random prefix	„n3_last name“
Synonyms	„surname“
Translation	„Nachname“



Data Noise

Method	„Mike Brown“
Empty value	„“
Phonetic error	„Myke Brown“
OCR error	„Mike 8rown“
Typing error	„Hike Browj“
Reformatting	„Mike_Brown“
Abbreviate random length	„M.Br.“
Map string to numeric value	„1“

Method	„5“
Random numeric value	„4“
Random numeric outlier	„103“



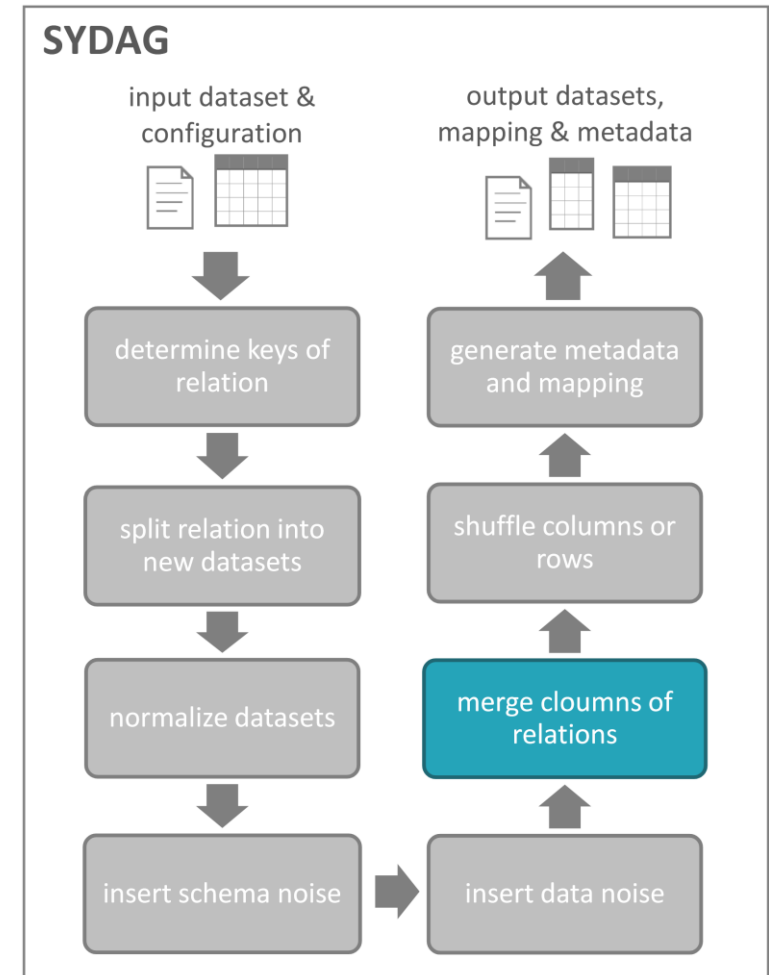
Merging

- Merge columns within relation

Name	Surname	Age
Jane	Smith	22
Tim	Miller	45
Mike	Brown	14

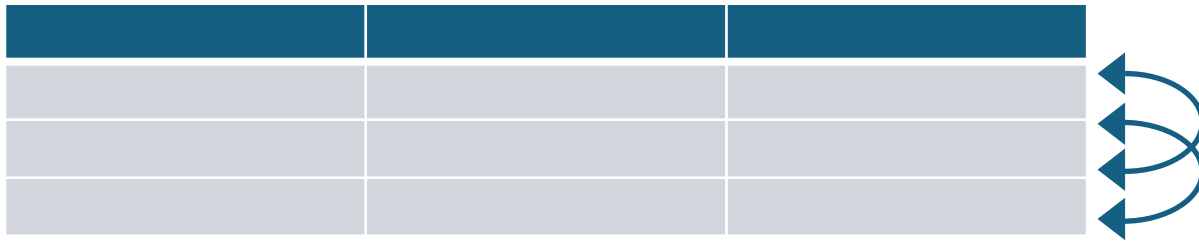
Merge

Name, Surname	Age
Jane, Smith	22
Tim, Miller	45
Mike, Brown	14

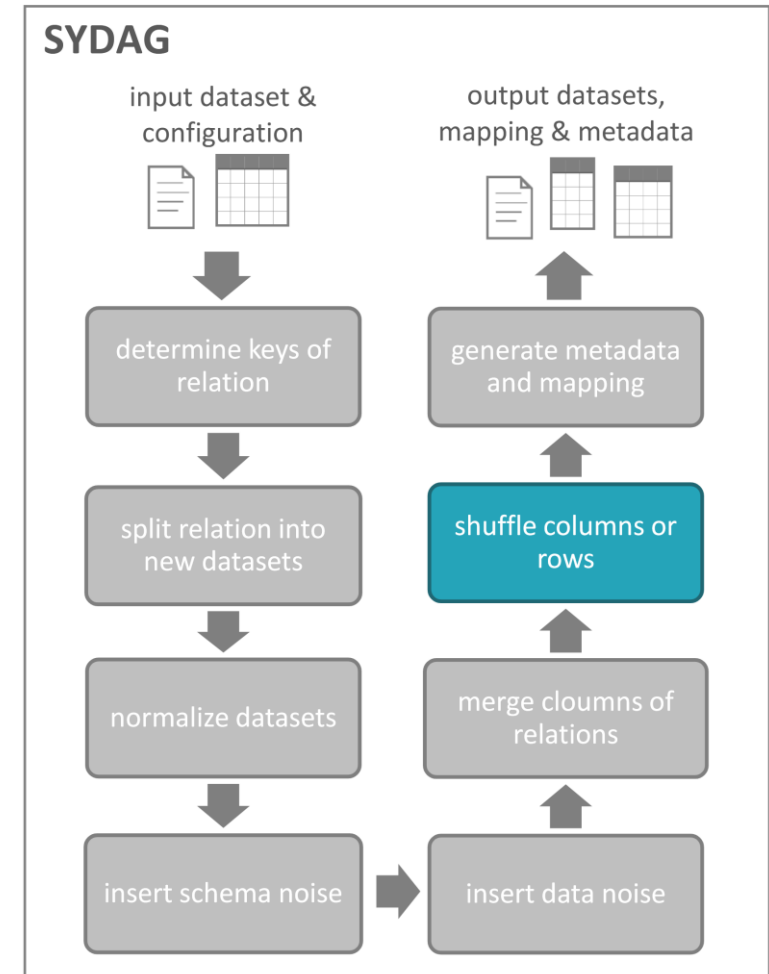


Shuffling

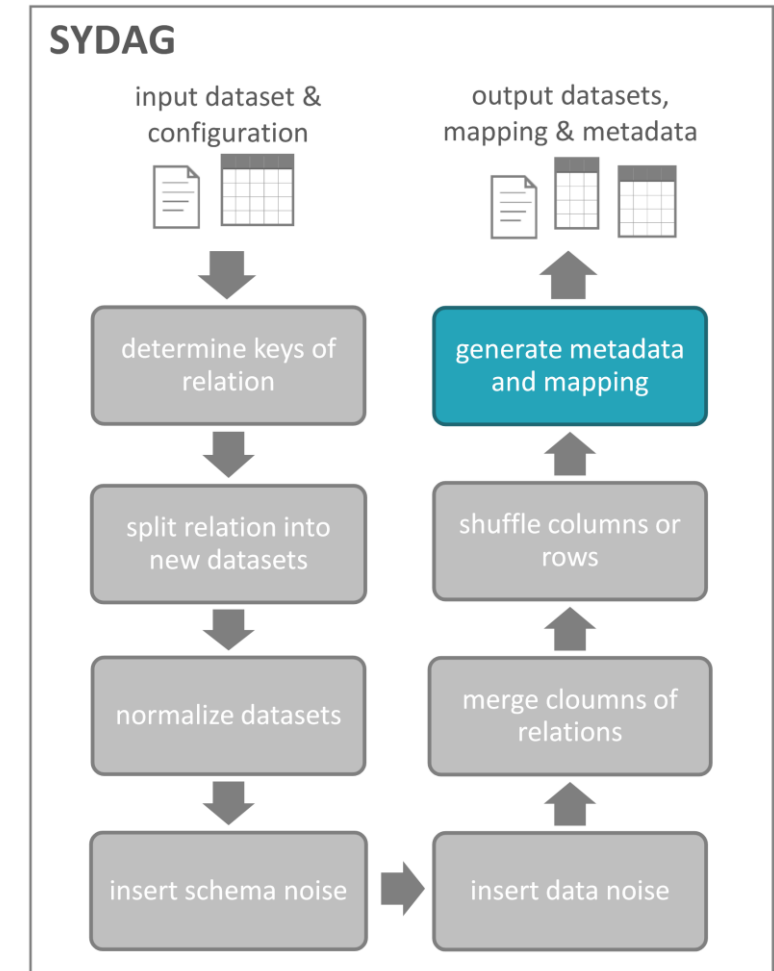
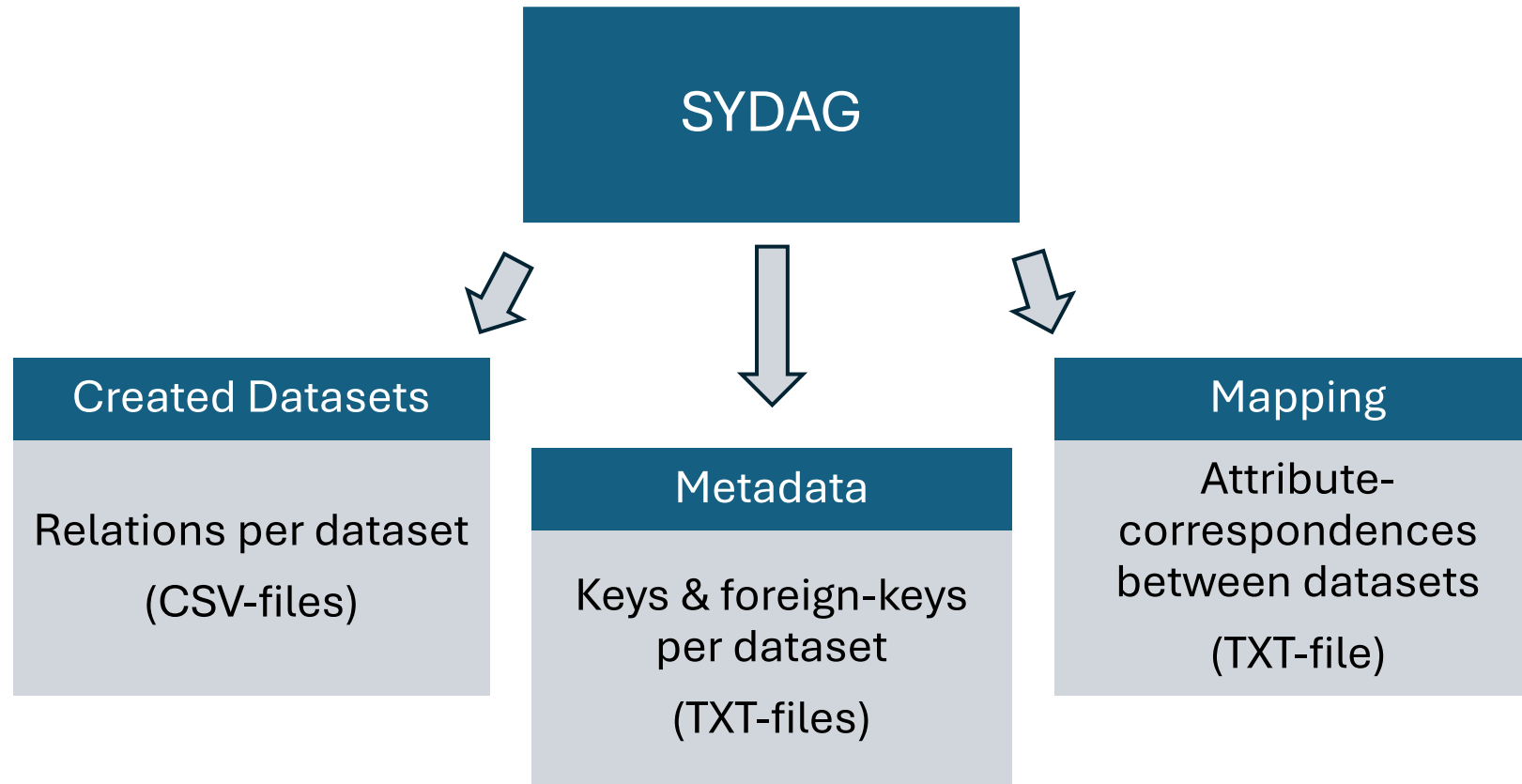
- Shuffle rows:



- Shuffle columns:



Output Generation



Usage

STEP 1

[Datasource](#)

STEP 2

Select Mode

STEP 3

Select Split

STEP 4

Select Structure

STEP 5

Select Schema Noise

STEP 6

Select Data Noise

STEP 7

Select Shuffle

STEP 8

Summary

Upload your Dataset

Please provide a CSV file.

CSV File

Browse... No file selected.

Does the CSV file have headers?

No

Separator

e.g., ',' or ';'

Quote Character

e.g., '"'

Escape Character

e.g., '\'

Next Step

STEP 1

Datasource

STEP 2

Select Mode

STEP 3

Select Split

STEP 4

Select Structure

STEP 5

Select Schema Noise

STEP 6

[Select Data Noise](#)

STEP 7

Select Shuffle

STEP 8

Summary

Select Noise Options for Data

For each dataset, choose whether to add noise to the data. If enabled, specify the error methods.

Dataset A

☒ Enable Noise

☐ Break Key Constraints

Dataset A Error Methods

Select methods for Dataset A

Selected:

Percentage of noisy Columns/ Rows

Percentage: 26 %

Percentage of noisy entries in Column/ Row

Percentage: 12 %

Go Back

Dataset B

☒ Enable Noise

☐ Break Key Constraints

Dataset B Error Methods

addRandomPrefix shuffleWords

Search...

☐ (Select All)

☐ replaceWithSynonyms

☒ addRandomPrefix

☐ replaceWithTranslation

☒ shuffleWords

☐ generateMissingValue

☐ generatePhoneticError

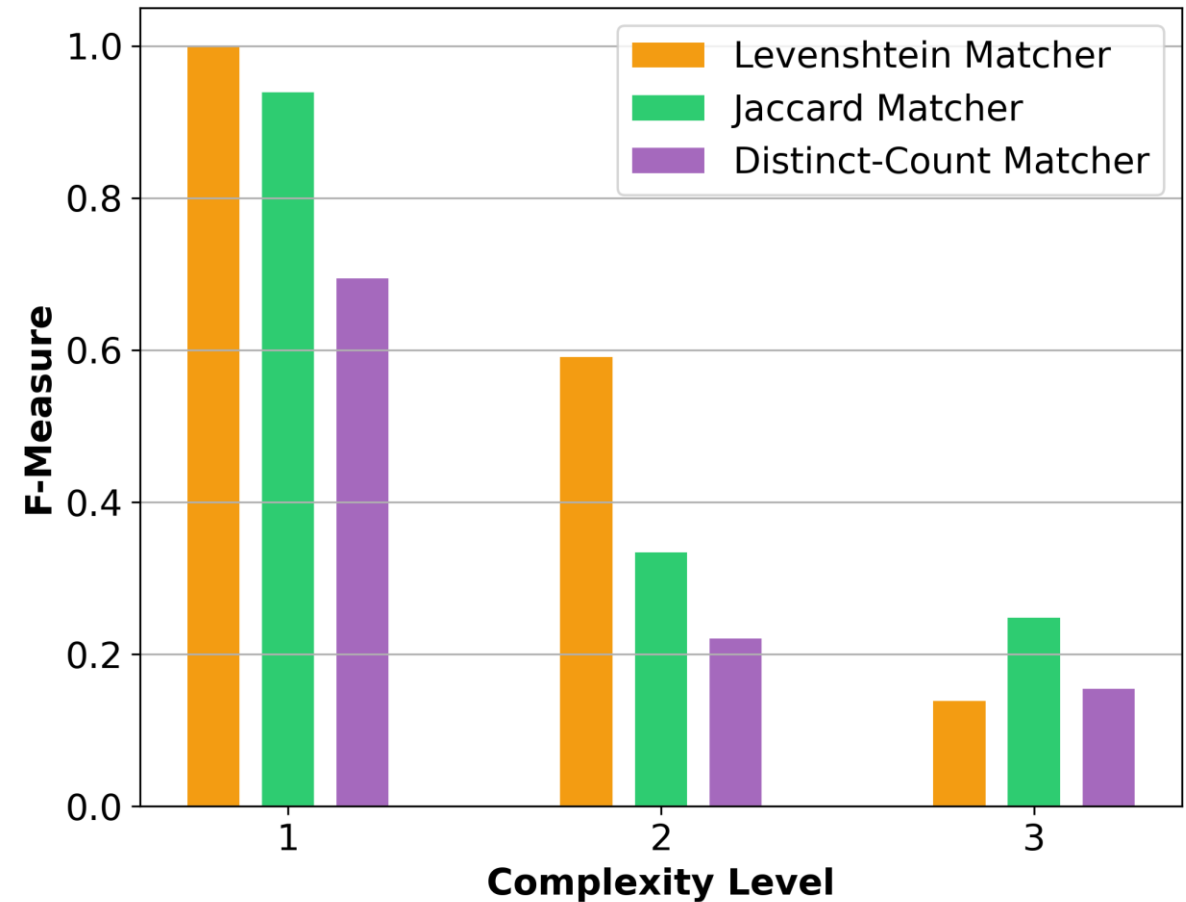
☐ generateOCRError

☐ changeFormat

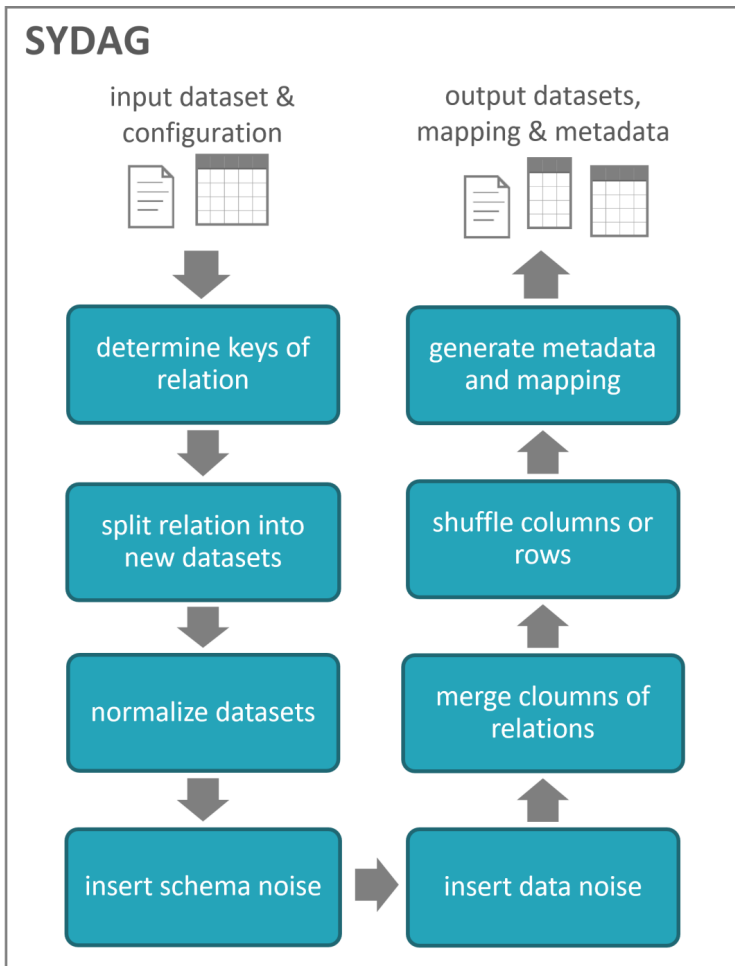
Next Step

Evaluation

- 4 datasets
- 3 complexity levels
→ 12 integration scenarios
- F-Measure:
 - Averaged across datasets



Summary



Future directions:

- Extend evaluation
- Different input-formats
- Performance

Upload your Dataset

Please provide a CSV file.

CSV File
Browse... No file selected.

Does the CSV file have headers?
☐ No

Separator
e.g., ',' or '\t'

Quote Character
e.g., '"'

Escape Character
e.g., '\'

Select Noise Options for Data

For each dataset, choose whether to add noise to the data. If enabled, specify the error methods.

Dataset A

☒ Enable Noise
☐ Break Key Constraints

Dataset A Error Methods
Select methods for Dataset A

Selected:

Percentage of noisy Columns/ Rows
Percentage: 55 %

Percentage of noisy entries in Column/ Row
Percentage: 23 %

Dataset B

☒ Enable Noise
☐ Break Key Constraints

Dataset B Error Methods
addRandomPrefix shuffleWords

Search...

- ☐ (Select All)
- ☐ replaceWithSynonyms
- ☒ addRandomPrefix
- ☐ replaceWithTranslation
- ☒ shuffleWords
- ☐ generateMissingValue
- ☐ generatePhoneticError
- ☐ generateOCRError
- ☐ changeFormat