COCOA: COrrelation COefficient-Aware Data Augmentation

Mahdi Esmailoghli[1]

Jorge-Arnulfo Quiané-Ruiz[2]

Ziawasch Abedjan[1, 3]

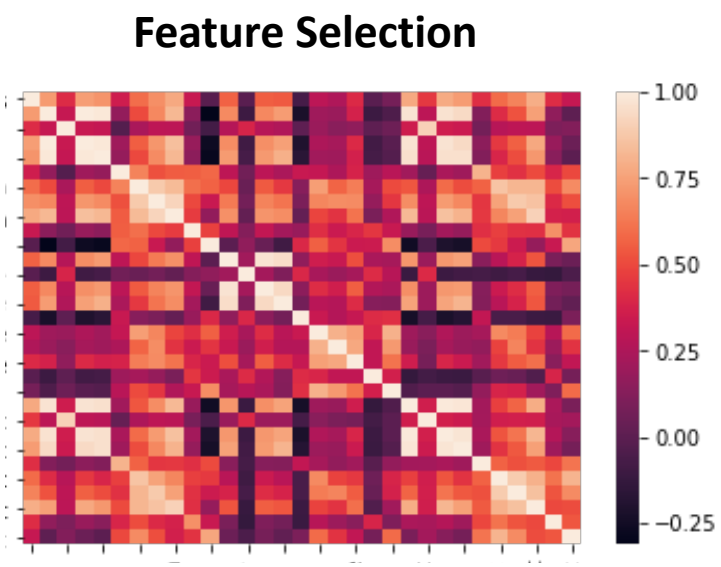
1 Leibniz Universität Hannover

2 TU Berlin

3 L3S Research Center

Correlating Coefficient

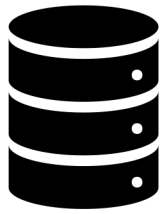
- Dependencies in the data
- Relevance of the feature
- Remove redundancy
- Feature importance
- Data enrichment

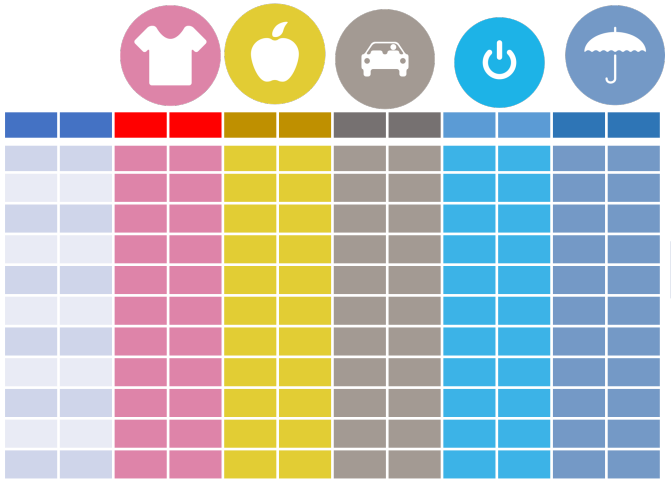


Data Enrichment

Poor Dataset

Country	RO
Germany	.7
U.S.A.	1.2
U.K.	1.0
China	.4





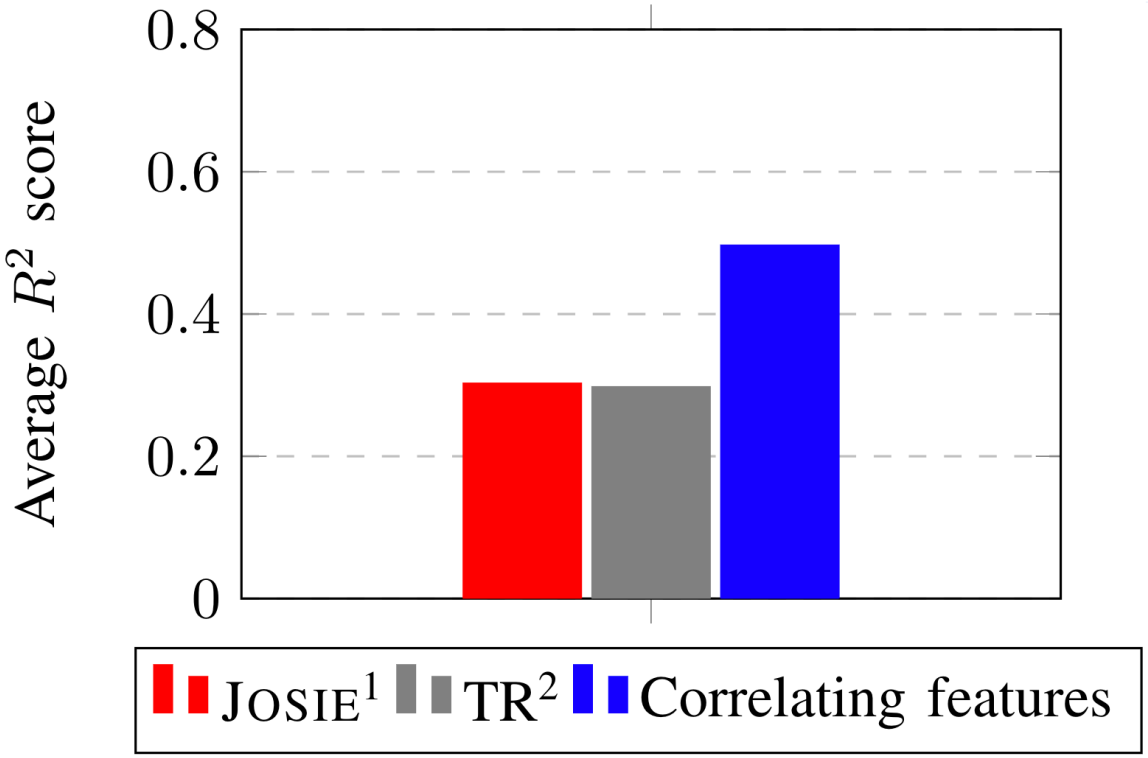
Enriched Dataset

Country	Education	Health	RO
Germany	95	Great	.7
U.S.A.	70	OK	1.2
U.K.	80	OK	1.0
China	90	Great	.4

Correlating Coefficient

Poor Dataset

Country	RO
Germany	.7
U.S.A.	1.2
U.K.	1.0
China	.4



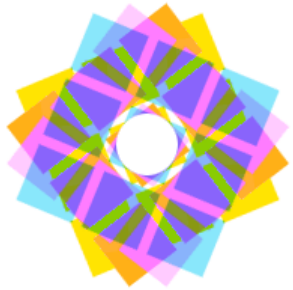
Enriched Dataset

Country	Education	Health	RO
Germany	95	Great	.7
U.S.A.	70	OK	1.2
U.K.	80	OK	1.0
China	90	Great	.4

1- Zhu, Erkang, et al. "Josie: Overlap set similarity search for finding joinable tables in data lakes." Proceedings of the 2019 International Conference on Management of Data. 2019.
2- Kumar, Arun, et al. "To join or not to join? Thinking twice about joins before feature selection." Proceedings of the 2016 International Conference on Management of Data. 2016.

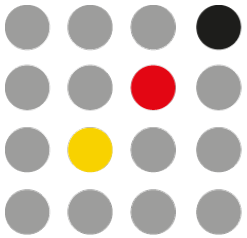
Available Data Lakes

EU Open Data Portal



DWTC
Dresden Web Table Corpus

DATA.GOV



GOVDATA
Das Datenportal für Deutschland

BERLIN OPEN DATA

EUROPEAN DATA PORTAL

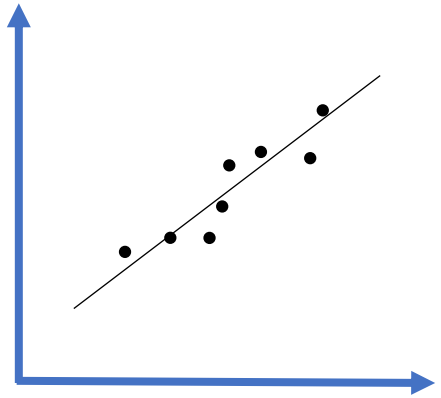
CHICAGO

OpenData

Available Data Lakes

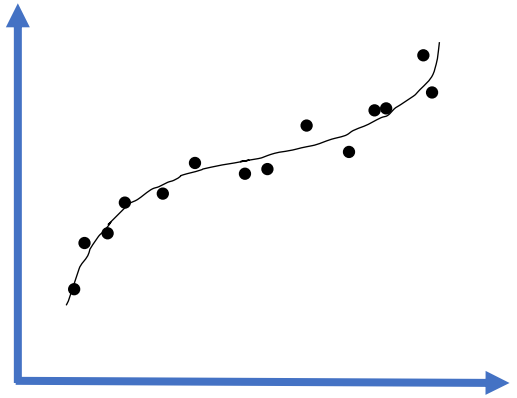
145 Million Tables
870 Million Columns

Time Complexity



Linear Correlation
e.g. Pearson

O(n)



Non-Linear Correlation
e.g. Spearman's

O(n.log n)

Categorical data: $O(n^2)$

Robustness

Area (Million sq. miles)	*Calling Code*
0.29	56
0.3	90
3.8	1
0.5	51
600	9800

Pearson = 1.0
Spearman's = 0.1

Goal

Area (Million sq. miles)	*Calling Code*
0.5	51
600	9800

Pearson = 1.0
Spearman's = 0.1

Calculate the **non-linear Spearman's correlation** in **linear time.**

Spearman's Correlation

Area	Code
0.3	90
0.29	56
3.8	1
0.5	51
600	9800

Area	Code
2 0.3	90 4
1 0.29	56 3
4 3.8	1 1
3 0.5	51 2
5 600	9800 5

$$\mathrm{Scc} = \frac{\sum_{i=1}^{m}(R(x_i) - \overline{R(x)})(R(y_i) - \overline{R(y)})}{\sqrt{\sum_{i=1}^{m}(R(x_i) - \overline{R(x)})^2 (R(y_i) - \overline{R(y)})^2}}$$

SCC= 0.1

Spearman's Correlation

Area	Code
0.29	56
0.3	90
3.8	1
0.5	51
600	9800

Area	Code
2 0.3	90 4
1 0.29	56 3
4 3.8	1 1
3 0.5	51 2
5 600	9800 5

$$Scc = \frac{\sum_{i=1}^{m}(R(x_i) - \overline{R(x)})(R(y_i) - \overline{R(y)})}{\sqrt{\sum_{i=1}^{m}(R(x_i) - \overline{R(x)})^2(R(y_i) - \overline{R(y)})^2}}$$

O(n.log n)

$O(n^2)$

Spearman's Correlation

Area	Code
0.29	56
0.5	51
600	9800

	Area	Code	
1	0.29	56	3
3	0.5	51	2
5	600	9800	5

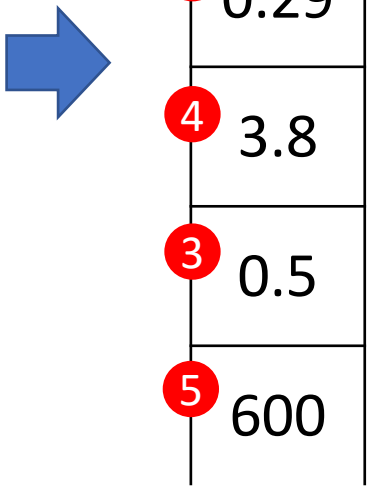
$$Scc = \frac{\sum_{i=1}^{m}(R(x_i) - \overline{R(x)})(R(y_i) - \overline{R(y)})}{\sqrt{\sum_{i=1}^{m}(R(x_i) - \overline{R(x)})^2 (R(y_i) - \overline{R(y)})^2}}$$

$O(n^2)$

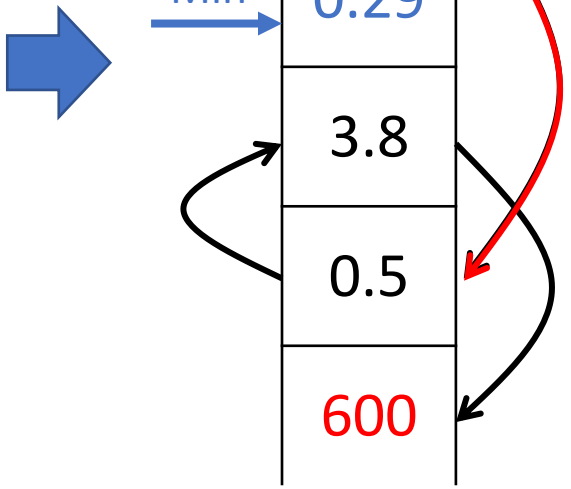
COCOA leverages a **light-weight index** to caculate **non-linear correaltion** in a **linear time**.

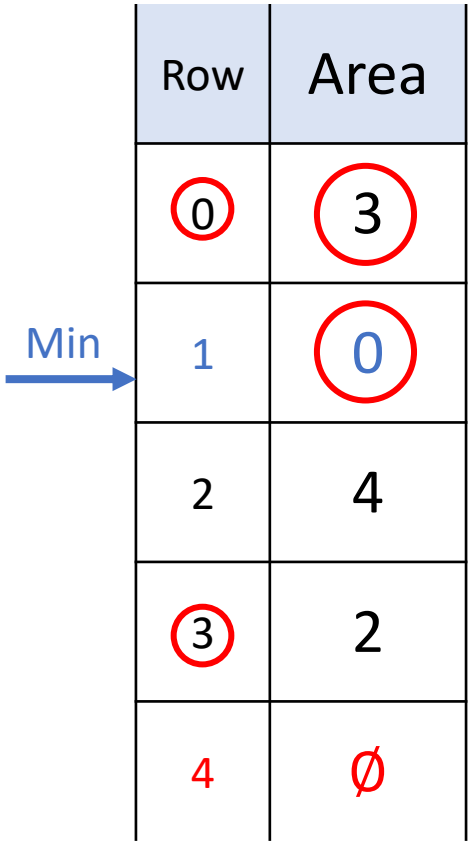
Order Index (Offline)

Country	Area	Code
Turkey	0.3	90
Chile	0.29	56
Canada	3.8	1
Peru	0.5	51
Iran	600	9800



	Area
2	0.3
1	0.29
4	3.8
3	0.5
5	600





COCOA (Online)

1. Retrieving Tables with the highest overlap.
2. Materializing Joinable rows
3. Detecting Joinable rows

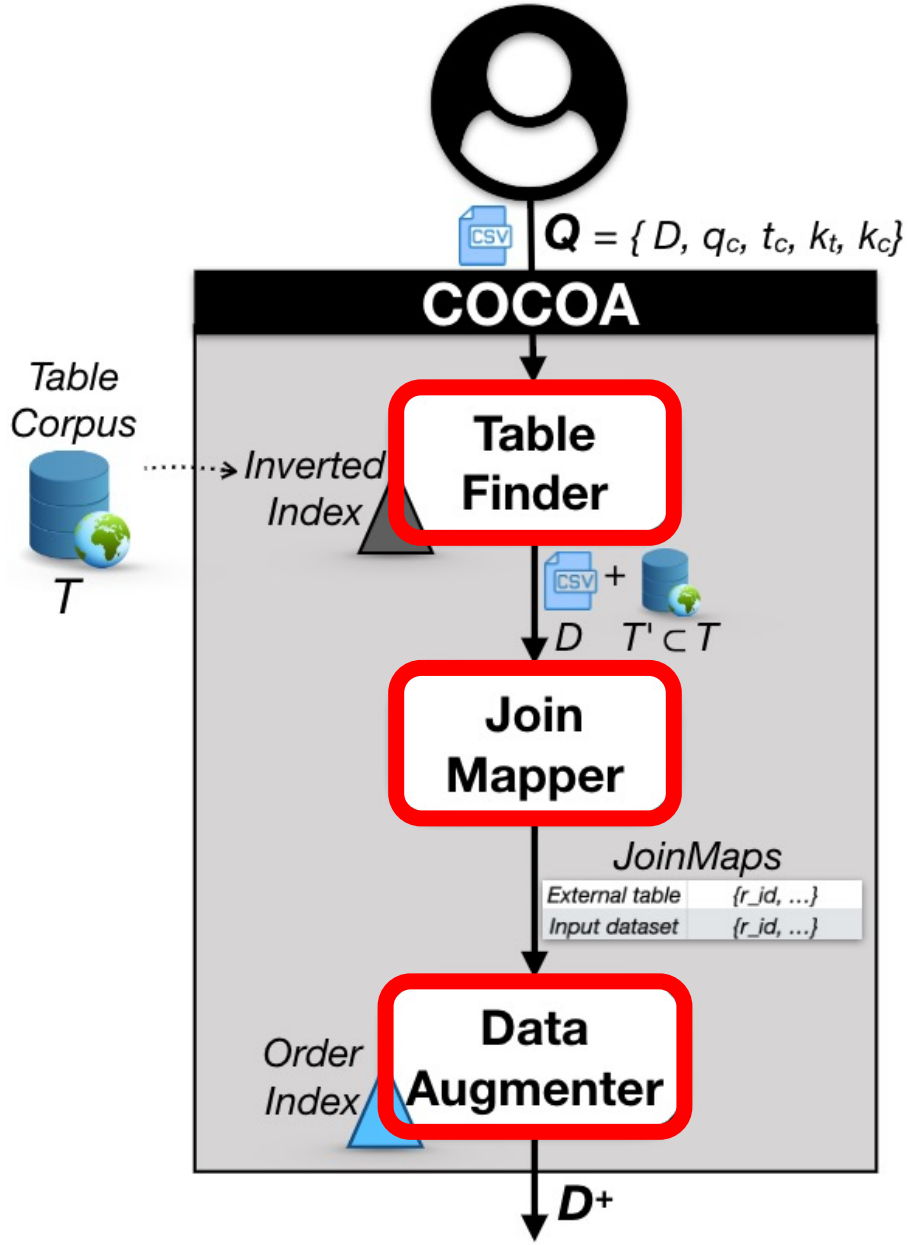
SELECT $T,\ C,\ count(*)$ **AS** $overlap$

FROM $tbl\_inverted\_index$

row in C_1	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
row in q	∅	3	∅	∅	∅	5	1	4	2

ORDER BY $overlap$ **DESC**

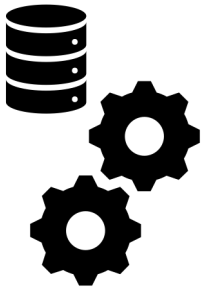
LIMIT k_t



Correlation Calculation (Online)

Input Dataset

Row	Name (q)	Pop. (t)	Pop. (Rank)
1	Russia	144	5
2	Turkey	81	4
3	Switzerland	9	1
4	Sweden	10	2
5	Canada	37	3



Joinable Table

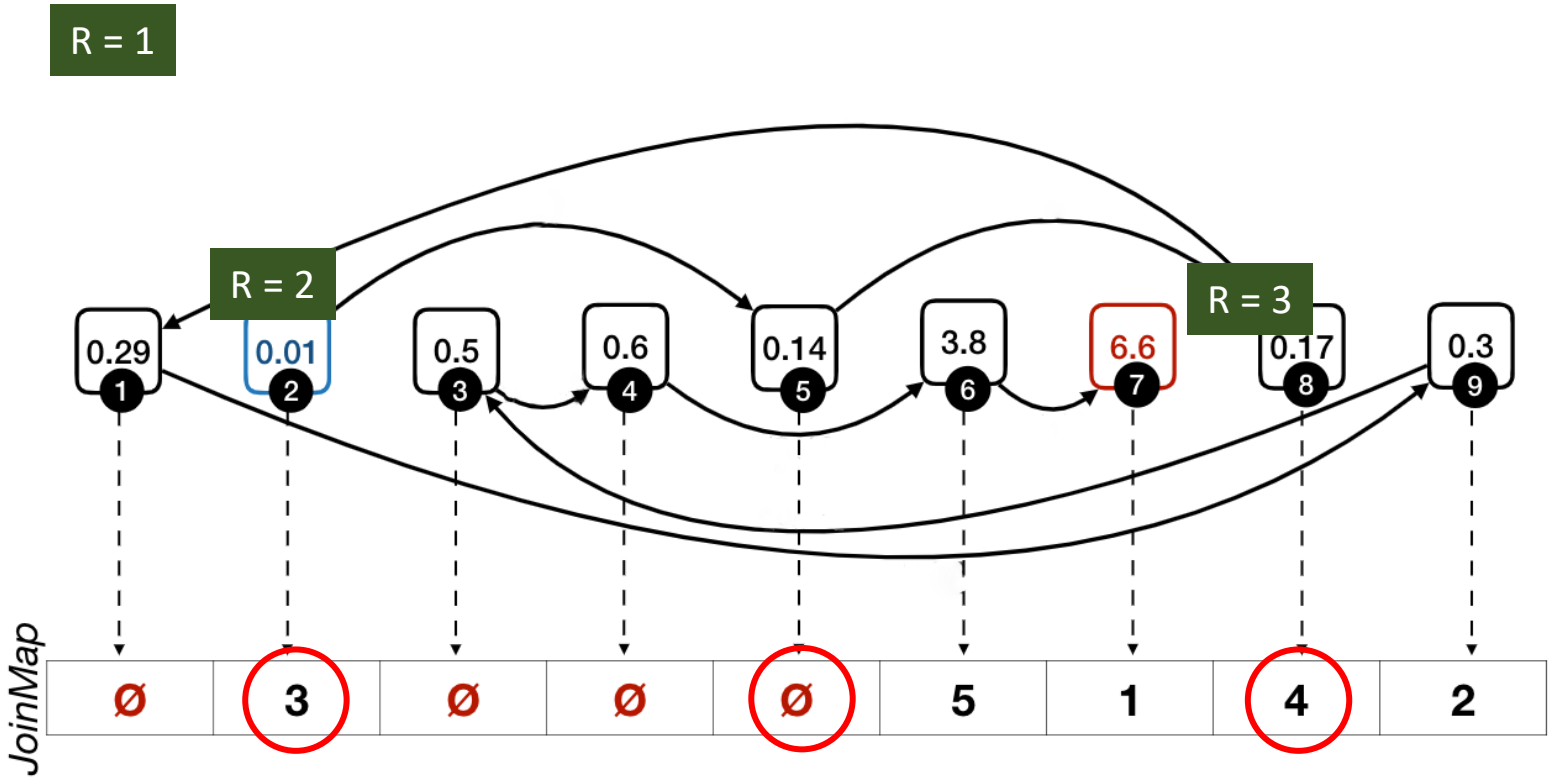
Row	Country	Area
1	Chile	0.29
2	Switzerland	0.01
3	Peru	0.5
4	Iran	0.6
5	Germany	0.14
6	Canada	3.8
7	Russia	6.6
8	Sweden	0.17
9	Turkey	0.3

Join Map

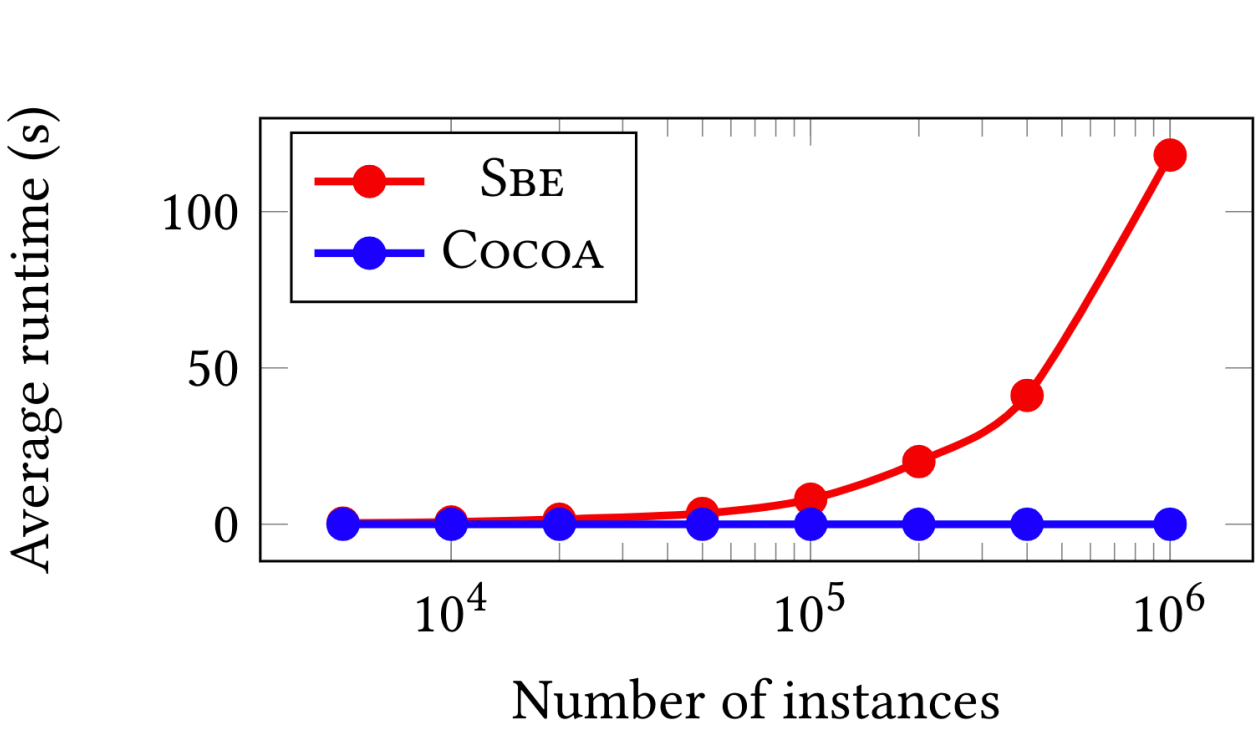
Row in **Country**	1	2	3	4	5	6	7	8	9
Row in **q**	Ø	3	Ø	Ø	Ø	5	1	4	2

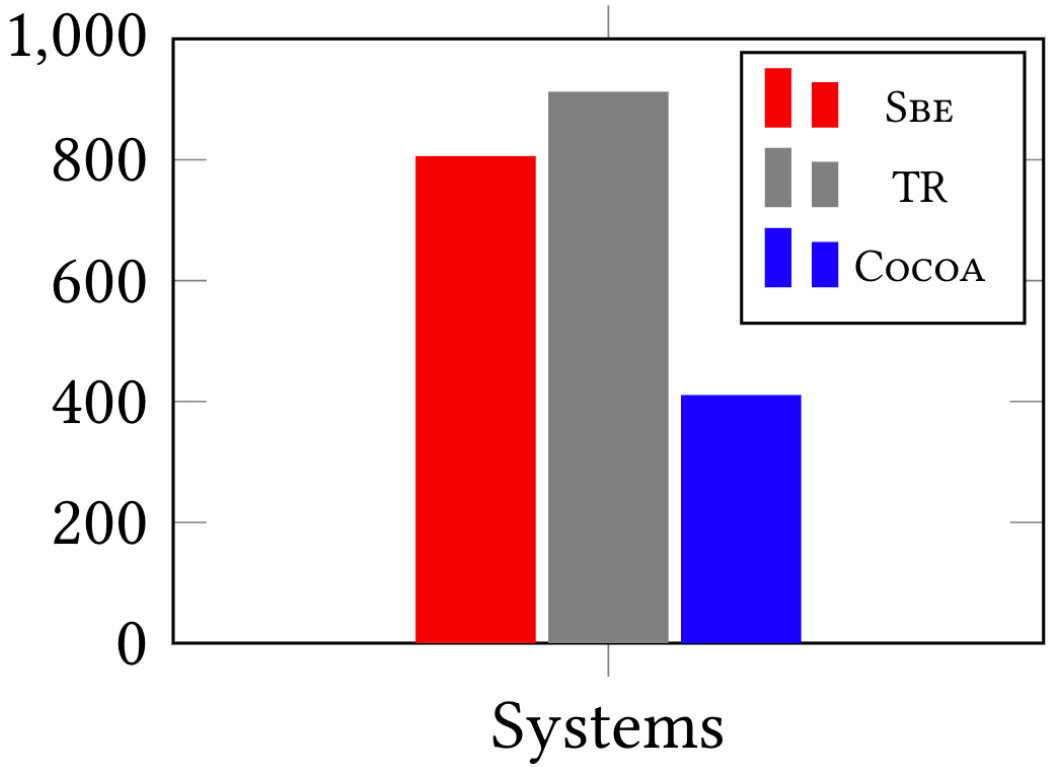
Correlation Calculation (Online)

Row	Country	Pop.	Pop. (Rank)	Area (Rank)
1	Russia	144	5	5
2	Turkey	81	4	3
3	Switzerland	9	1	1
4	Sweden	10	2	2
5	Canada	37	3	4



Results


Average runtime (s)
100
50
0
SBE
COCOA
10^4
10^5
10^6
Number of instances


1,000
800
600
400
200
0
SBE
TR
COCOA
Systems

COCOA: COrrelation COefficient-Aware Data Augmentation

Mahdi Esmailoghli[1]

Jorge-Arnulfo Quiané-Ruiz[2]

Ziawasch Abedjan[1, 3]

1 Leibniz Universität Hannover

2 TU Berlin

3 L3S Research Center