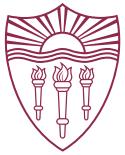


ROBUST AND PROACTIVE ERROR DETECTION AND CORRECTION FOR TABLES

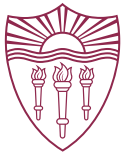
Committee: Craig Knoblock (chair), Cyrus Shahabi, Gerard Hoberg, Muhao Chen, Jay Pujara

Minh Pham

Dissertation Defense

Outline

1. Motivation and Background
2. Syntactic Error Detection
3. Syntactic Error Correction
4. Semantic Error Detection and Correction
5. Related Work
6. Discussion



MOTIVATION AND BACKGROUND

Tables are rich sources of structured knowledge

- Millions of tables on the Web
- Providing data for applications in different domains

Beers

Beer name	Style	Ounces	Abv
Pub Beer	American Pale Ledger	12.0 oz	0.05
Devil's Cup	American Pale Ale (APA)	12.0 oz.	0.07
Rise of the Phoenix	American IPA	12.0 ounce	0.07
Sinister	American Double/Imperial IPA	12.0 oz	0.09%

	Country (or territory)	Capital	Popul
1	China (more)	Beijing	21,
2	Japan (more)	Tokyo	13,
3	DR Congo	Kinshasa	12,
4	Russia (more)	Moscow	12,
5	Indonesia (more)	Jakarta	10,
6	South Korea (more)	Seoul	9,
7	Egypt (more)		
8	Mexico		

Countries

GDP per capita	Voluntary expenditure	Household income	Passenger transport
41 450	2.3	-0.5	138 643
43 746	2.3 |	1.1	132 125
44 720	2.3	0.4	134 954 e

Economics

Syntactic errors

Relate to the format/representation of the data such as misspellings and format inconsistencies

Beer name	Style	Ounces	Abv
Pub Beer	American Pale Ledger	12.0 oz	0.05
Devil's Cup	American Pale Ale (APA)	12.0 oz.	0.07
Rise of the Phoenix	American IPA	12.0 ounce	0.07
Sinister	American Double/Imperial IPA	12.0 oz	0.09%

GDP per capita	Voluntary expenditure	Household income	Passenger transport
41 450	2.3	-0.5	138 643
43 746	2.3 |	1.1	132 125
44 720	2.3	0.4	134 954 e

Semantic errors

Relate to the meaning of the data and usually result in wrong information

Club	Location	Stadium
Al-Ahli	Jeddah	King Abdullah Sports City
Al-Faisaly	Harmah	King Salman Sport City Stadium
Al-Fateh	Al-Hasa	Prince Abdullah bin Jalawi Stadium
Al-Hilal	Riyadh	King Fahd International Stadium Prince Faisal bin Fahd Stadium
Al-Ittihad	Jeddah	King Abdullah Sports City
Al-Khaleej	Saihat	Prince Saud bin Jalawi Stadium
Al Nassr	Riyadh	King Fahd International Stadium Prince Faisal bin Fahd Stadium
Al-Qadisiyah	Khobar	Prince Saud bin Jalawi Stadium
Al-Raed	Buraidah	King Abdullah Sport City Stadium
Al-Shabab	Riyadh	King Fahd International Stadium Prince Faisal bin Fahd Stadium
Al-Taawoun	Buraidah	King Abdullah Sport City Stadium
Al-Wehda	Makkah	King Abdul Aziz Stadium
Hajer	Al-Hasa	Prince Abdullah bin Jalawi Stadium
Najran	Najran	Al Akhdoud Club Stadium

"The Prince Saud bin Jalawi Stadium is …. and it is the home stadium of Al-Qadisiya.

https://en.wikipedia.org/wiki/Prince_Saud_bin_Jalawi_Stadium

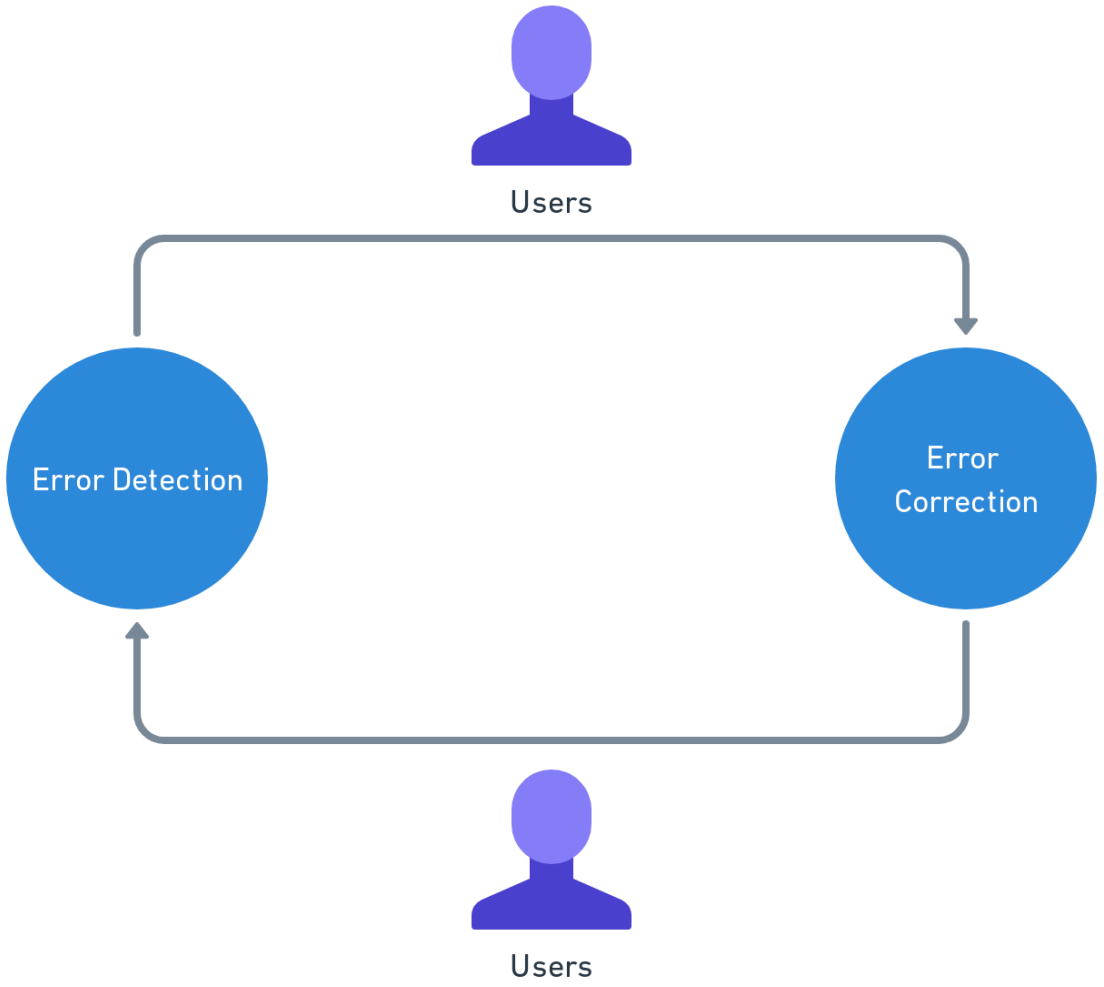
1 October	*The Increasingly Poor Decisions of Todd Margaret*	Channel 4
7 October	*PhoneShop*	
1 November	*Coppers*	
8 November	*Celebrity Coach Trip*	
30 November	*Frankie Boyle's Tramadol Nights*	

PhoneShop is a British sitcom that was first broadcast on Channel 4 as a television pilot on 13 November 2009.

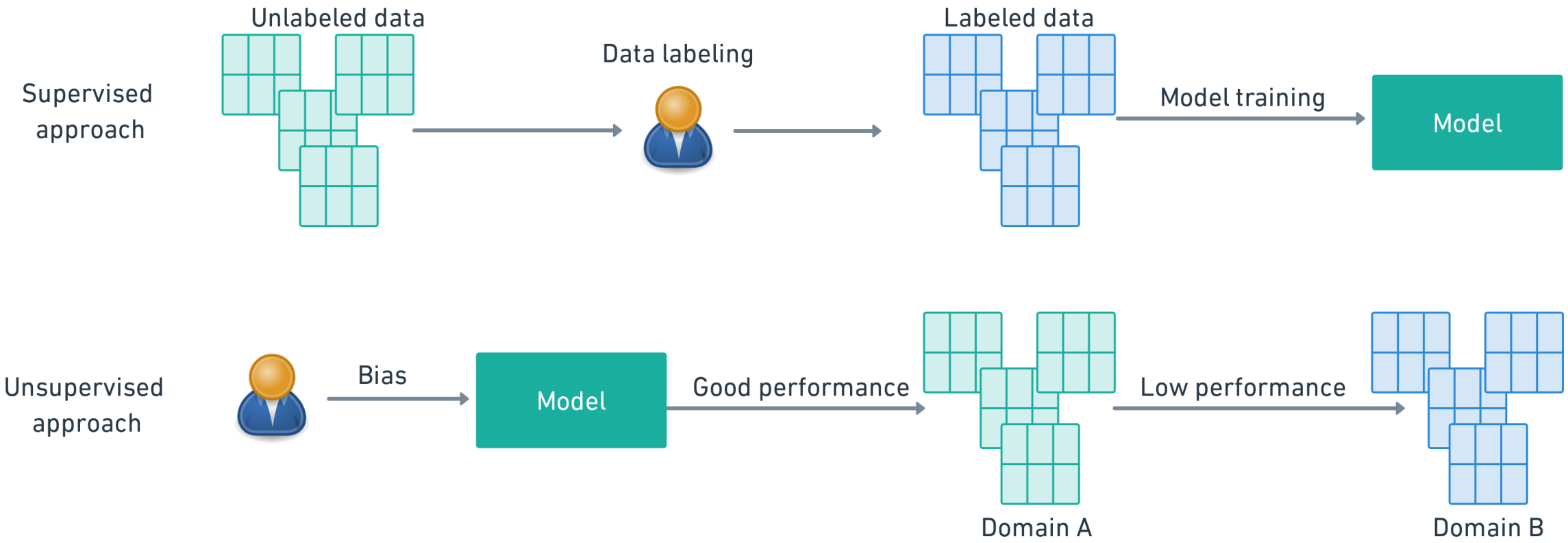
https://en.wikipedia.org/wiki/PhoneShop

Data cleaning

- Includes error detection and error correction
- Usually involves human interaction for data labeling and result verification
- Iterative process



Supervised and unsupervised approaches



Detecting and correcting semantic errors

"The Prince Saud bin Jalawi Stadium is …. and it is the home stadium of Al-Qadisiya."

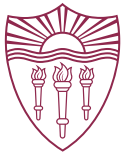
https://en.wikipedia.org/wiki/Prince_Saud_bin_Jalawi_Stadium

Relevant information available on Wikipedia

Club	Location	Stadium
Al-Ahli	Jeddah	King Abdullah Sports City
Al-Faisaly	Harmah	King Salman Sport City Stadium
Al-Fateh	Al-Hasa	Prince Abdullah bin Jalawi Stadium
Al-Hilal	Riyadh	King Fahd International Stadium Prince Faisal bin Fahd Stadium
Al-Ittihad	Jeddah	King Abdullah Sports City
Al-Khaleej	Saihat	Prince Saud bin Jalawi Stadium
Al Nassr	Riyadh	King Fahd International Stadium Prince Faisal bin Fahd Stadium
Al-Qadisiyah	Khobar	Prince Saud bin Jalawi Stadium
Al-Raed	Buraidah	King Abdullah Sport City Stadium
Al-Shabab	Riyadh	King Fahd International Stadium Prince Faisal bin Fahd Stadium
Al-Taawoun	Buraidah	King Abdullah Sport City Stadium
Al-Wehda	Makkah	King Abdul Aziz Stadium
Hajer	Al-Hasa	Prince Abdullah bin Jalawi Stadium
Najran	Najran	Al Akhdoud Club Stadium

Thesis Statement

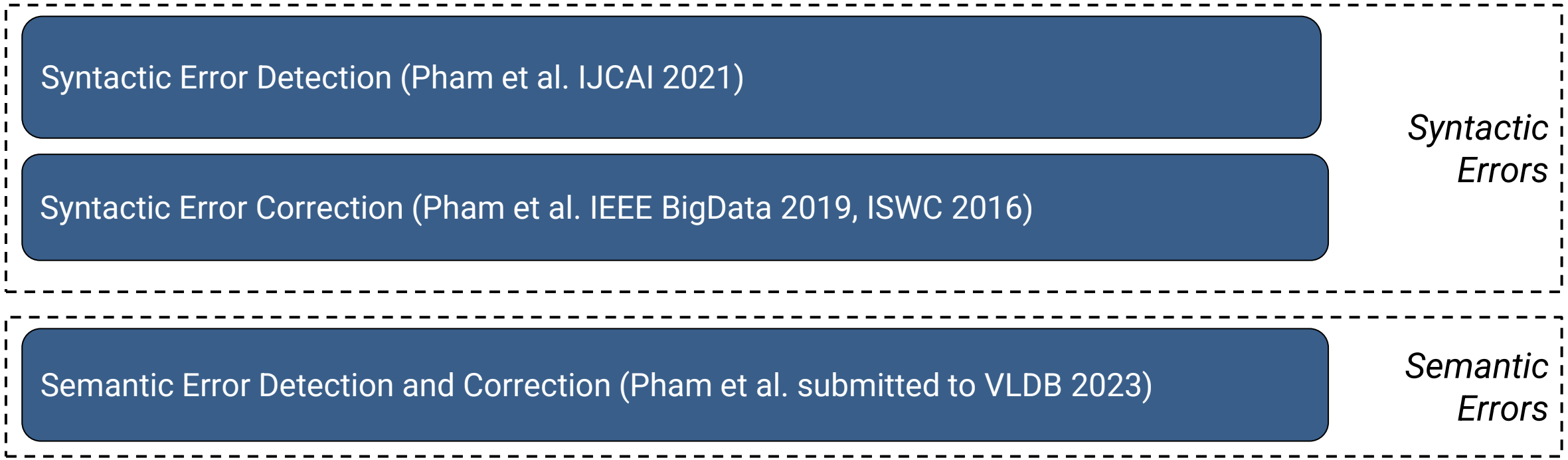
- *Open-domain knowledge and closed-domain weak supervision can be leveraged to reduce human interaction and improve accuracy in syntactic and semantic error detection and correction.*

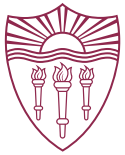


CONTRIBUTIONS

Contributions

- *Open-domain knowledge and closed-domain weak supervision can be leveraged to reduce human interaction and improve accuracy in syntactic and semantic error detection and correction.*



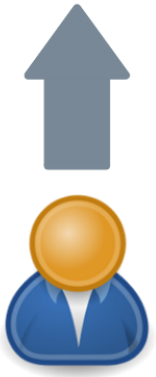


SYNTACTIC ERROR DETECTION

Motivating example: Supervised approach

1000
normal rows

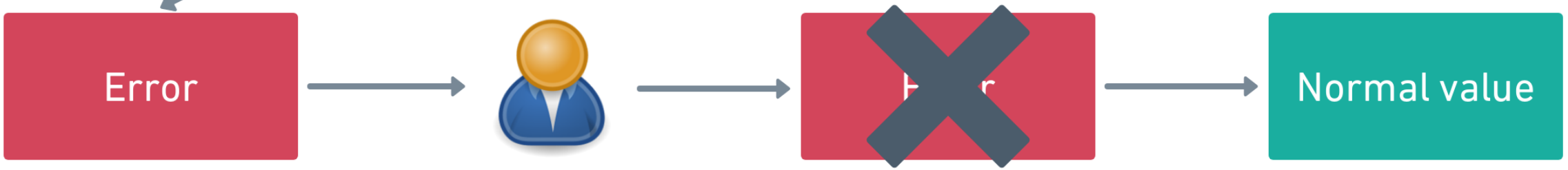
GDP per capita	Voluntary expenditure	Household income	Passenger transport
41 450	2.3	-0.5	138 643
43 746	2.3	1.1	132 125
…	…	…	…
…	…	…	…
44 720	2.3 |	0.4	134 954 e



How many normal rows
need to be labeled before the error?

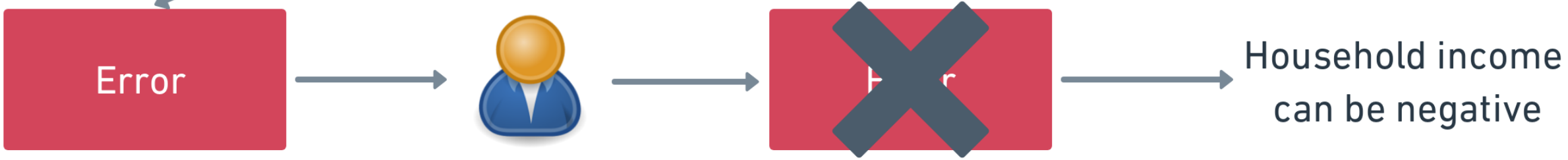
Motivating example: Unsupervised approach

GDP per capita	Voluntary expenditure	Household income	Passenger transport
41 450	2.3	-0.5	138 643
43 746	2.3 |	1.1	132 125
44 720	2.3	0.4	134 954 e

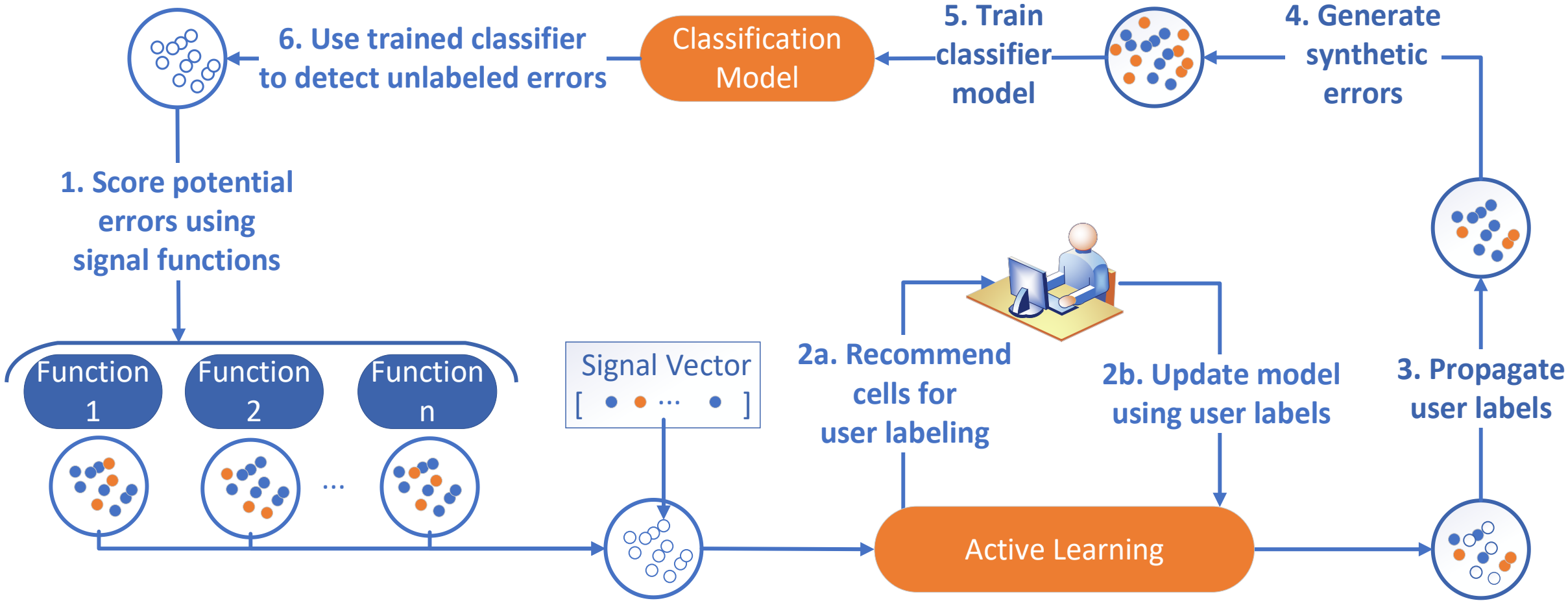


Motivating example: Semi-supervised approach

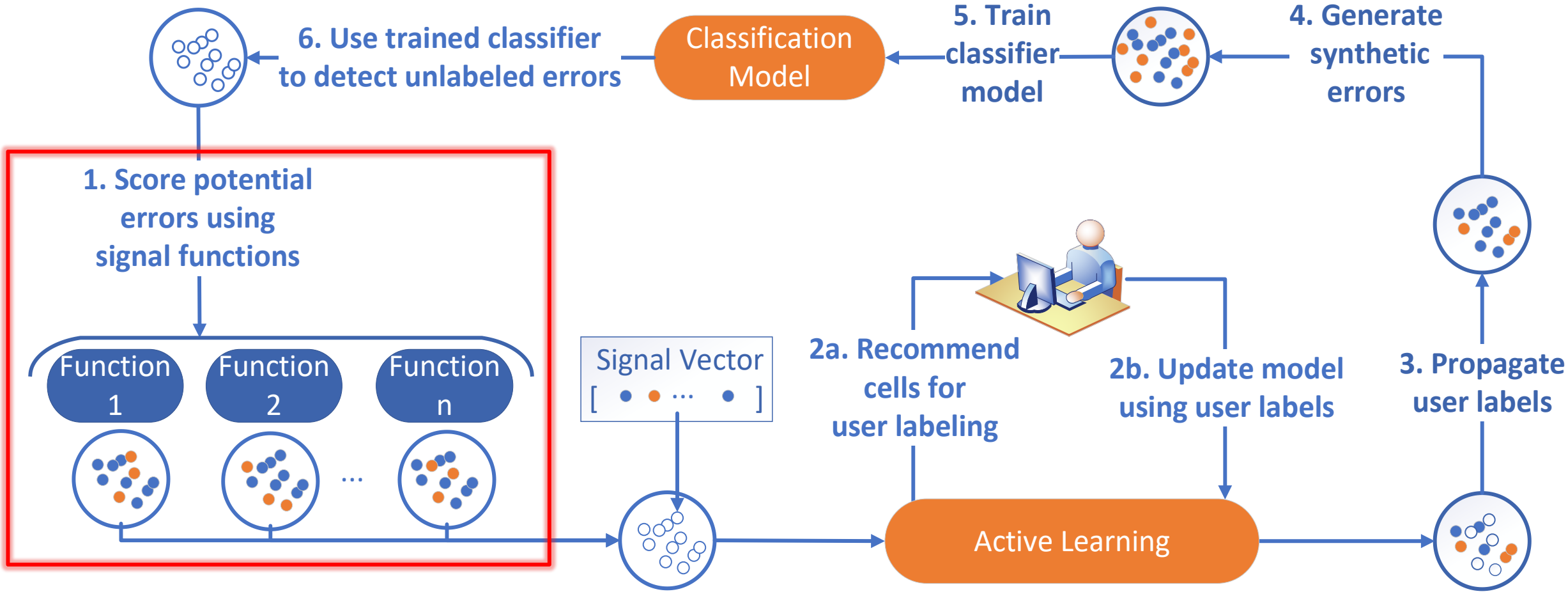
GDP per capita	Voluntary expenditure	Household income	Passenger transport
41 450	2.3	-0.5	138 643
43 746	2.3 |	1.1	132 125
44 720	2.3	0.4	134 954 e



Overall approach



Signal functions



Signal functions

Internal signals

GDP per capita	Voluntary expenditure	Household income	Passenger transport
41 450	2.3	-0.5	138 643
43 746	2.3 |	1.1	132 125
44 720	2.3	0.4	134 954 e

Voluntary expenditure
2.3
2.3 |
2.3

Uncommon format within column → Potential errors

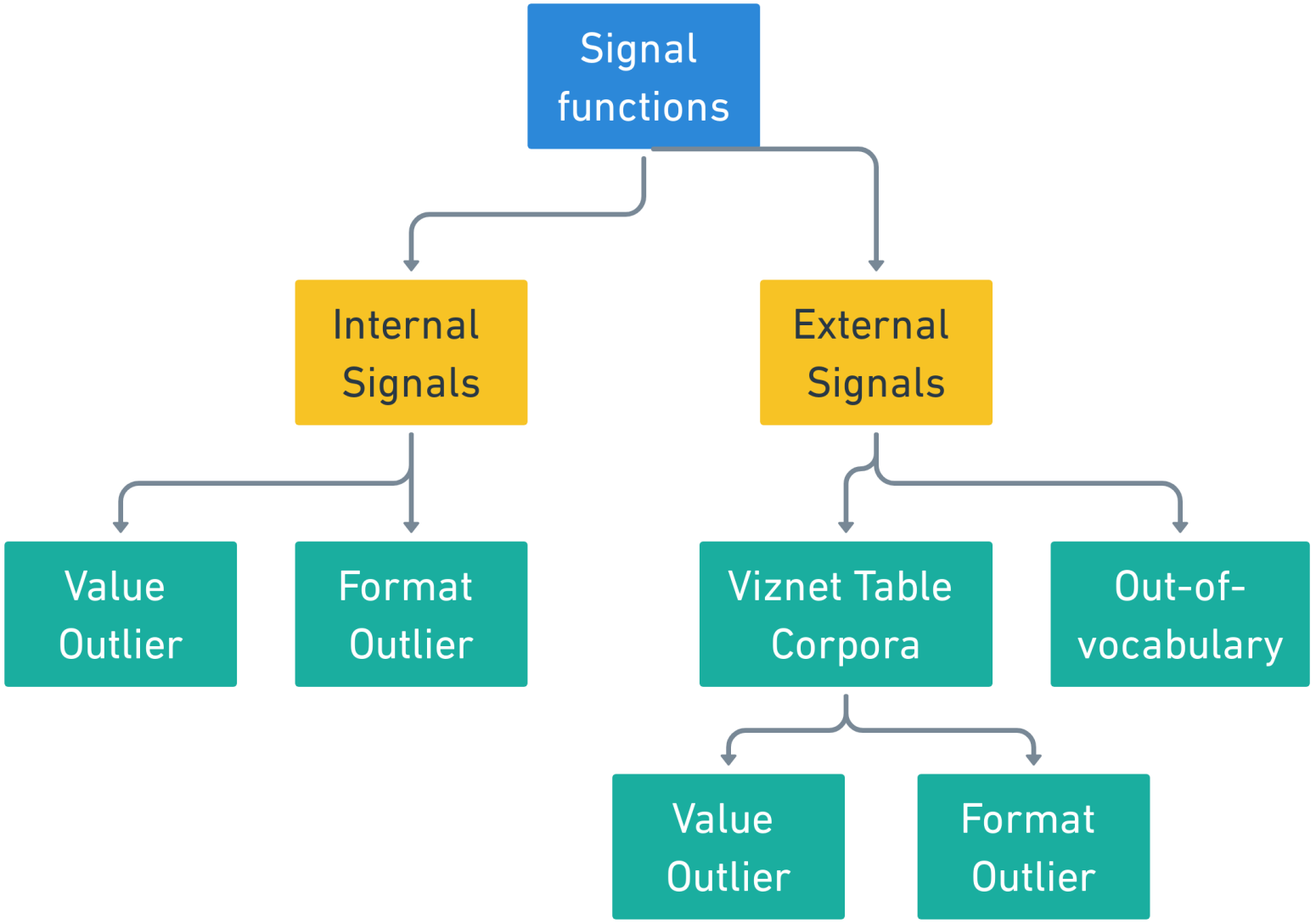
External signals

GDP per capita	Voluntary expenditure	Household income	Passenger transport
41 450	2.3	-0.5	138 643
43 746	2.3 |	1.1	132 125
44 720	2.3	0.4	134 954 e

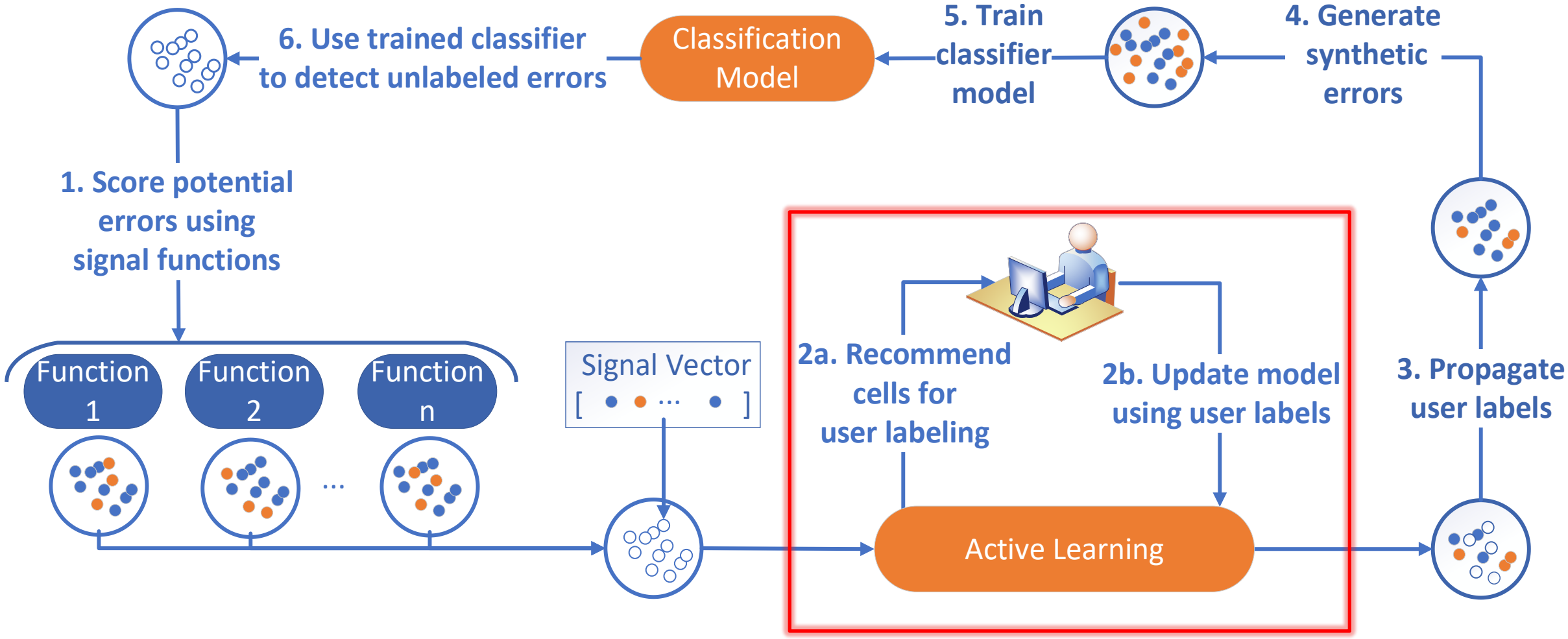
GDP per capita
41 450
43 746
44 720

Uncommon format in Web Table Corpora → Potential errors

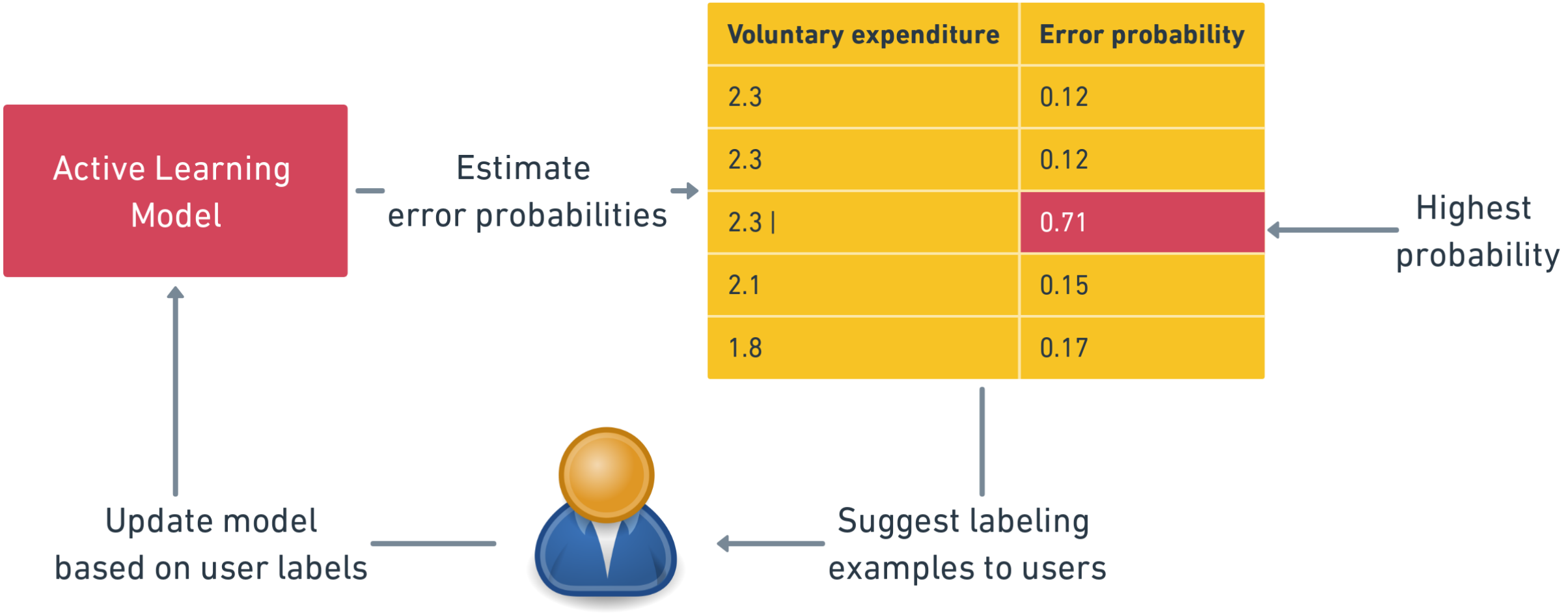
Signal functions



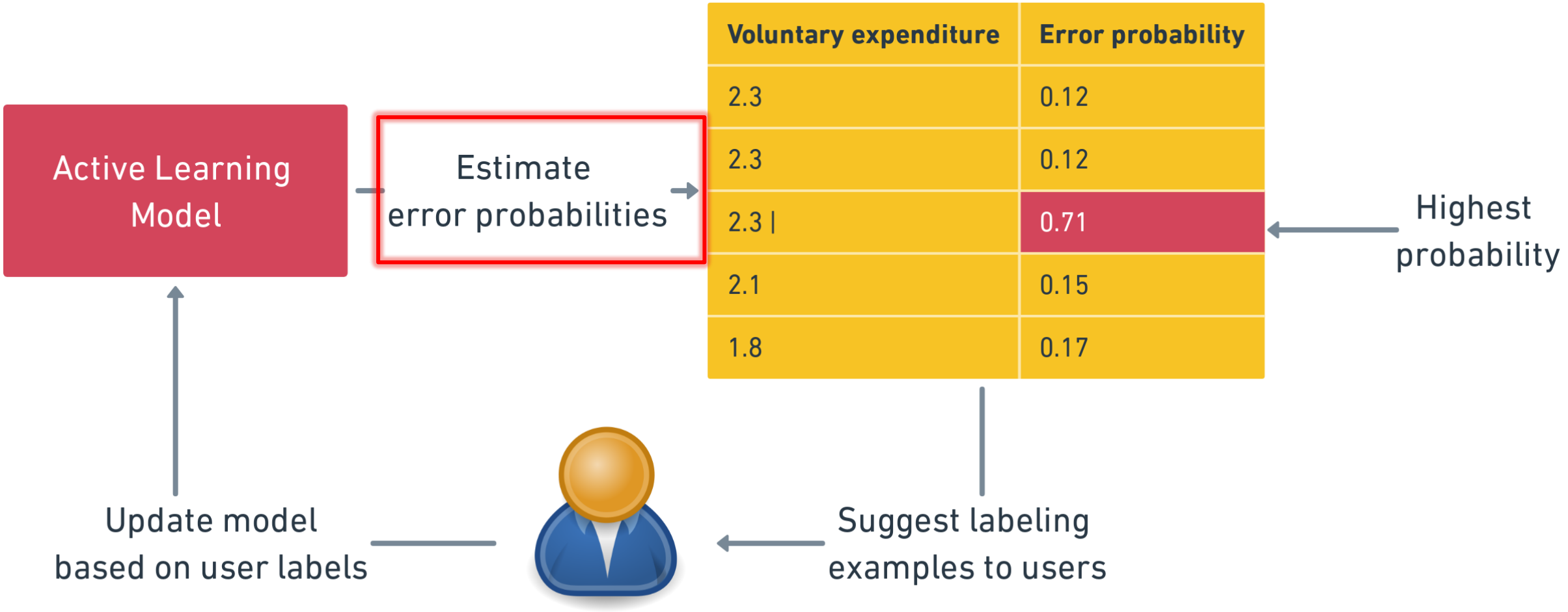
Active learning



Active-learning error detection



Active-learning error detection



Probabilistic Soft Logic (PSL)

- A probabilistic graphical model framework using first-order logic
- Two main elements: predicates and rules
- Predicates can have “soft” values [0,1]

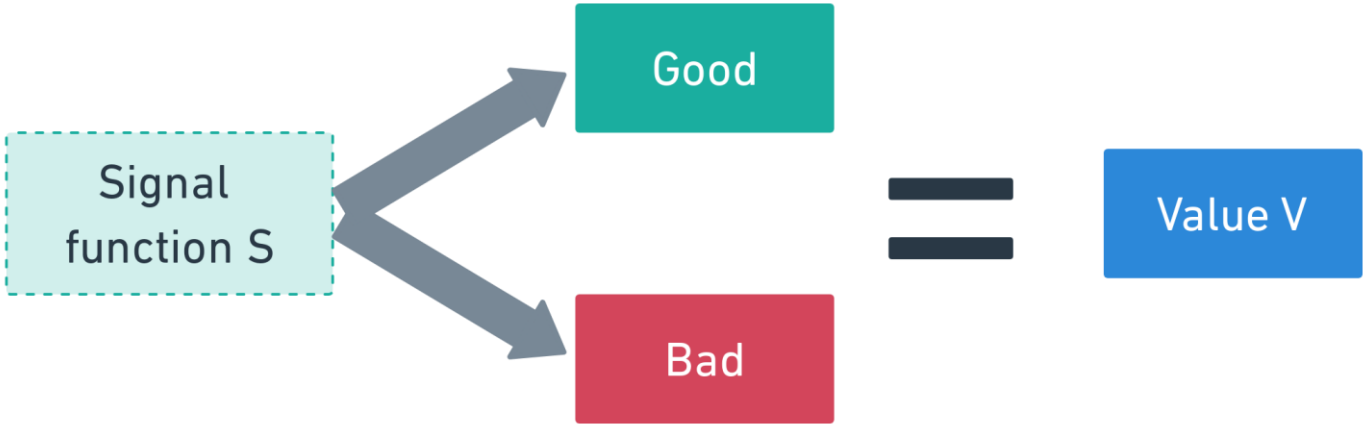
PSL model: Predicates

$HasSignal(c, s) = V$

Indicate value of signal function s when applying on cell c

$BadSignal(s) = V$

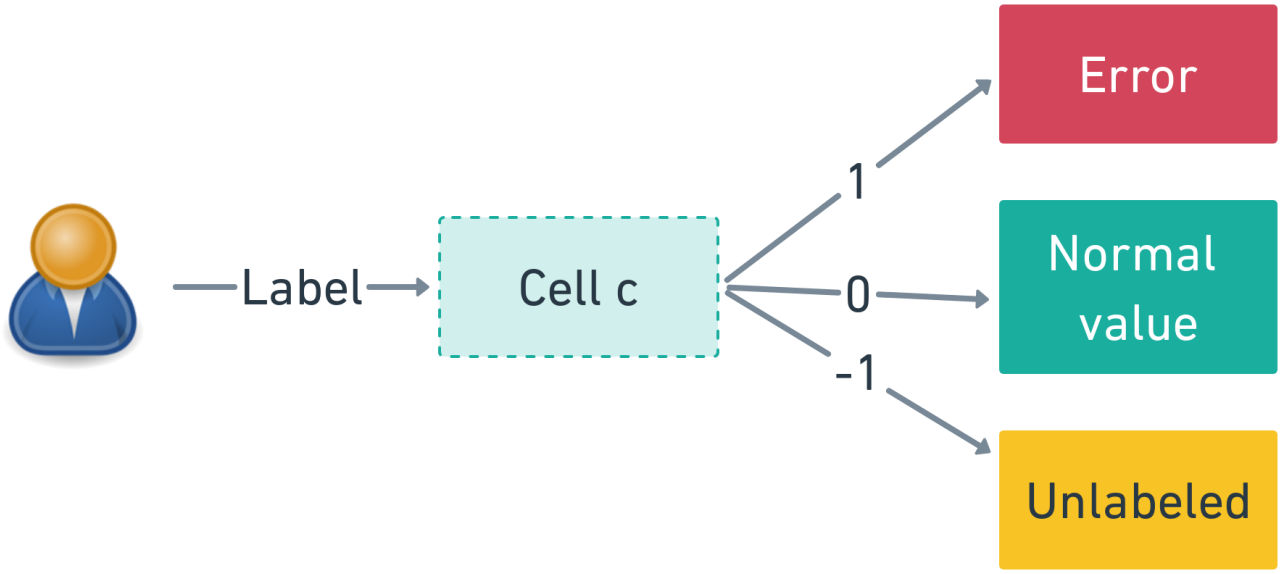
Indicate if a signal is bad or good



PSL model: Predicates

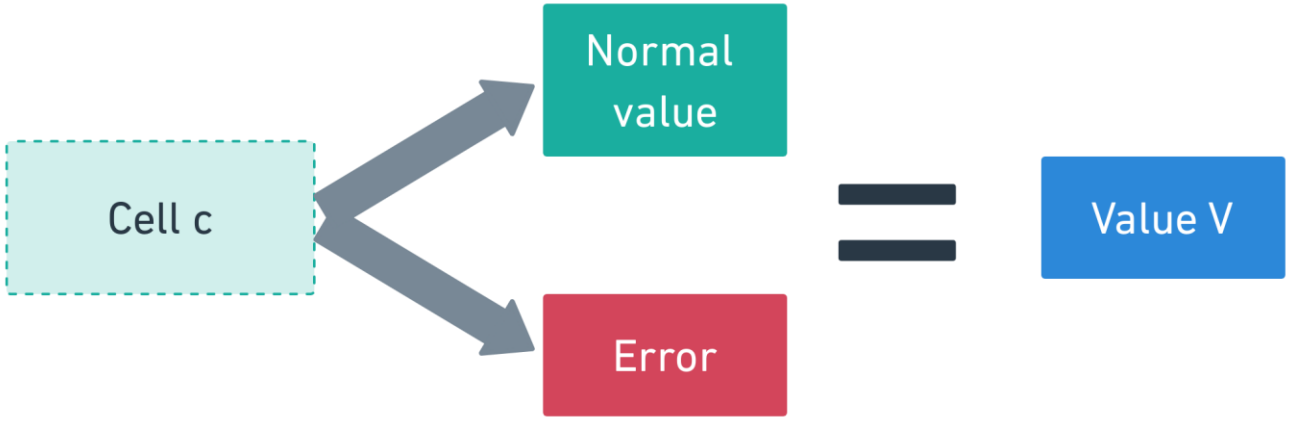
$Label(c, \{-1, 0, 1\}) = \{0, 1\}$

Indicate user label of cell c



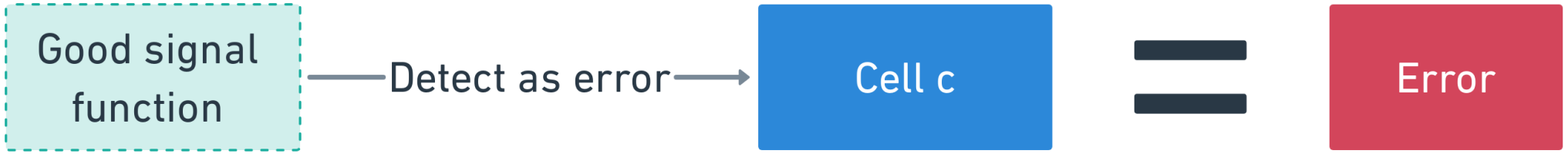
$Error(c) = V$

Indicate error probability of cell c

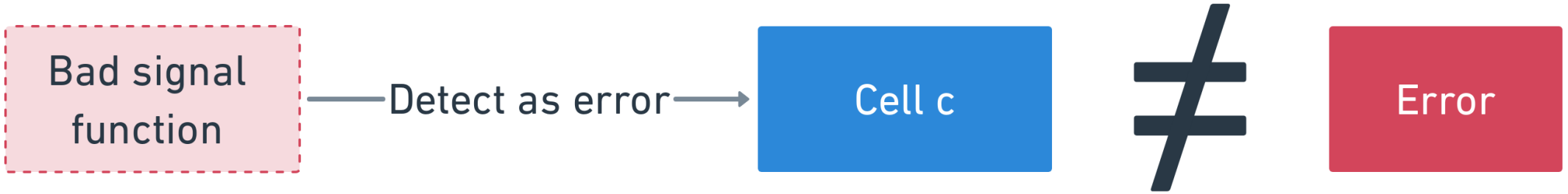


PSL rules: Error probabilities

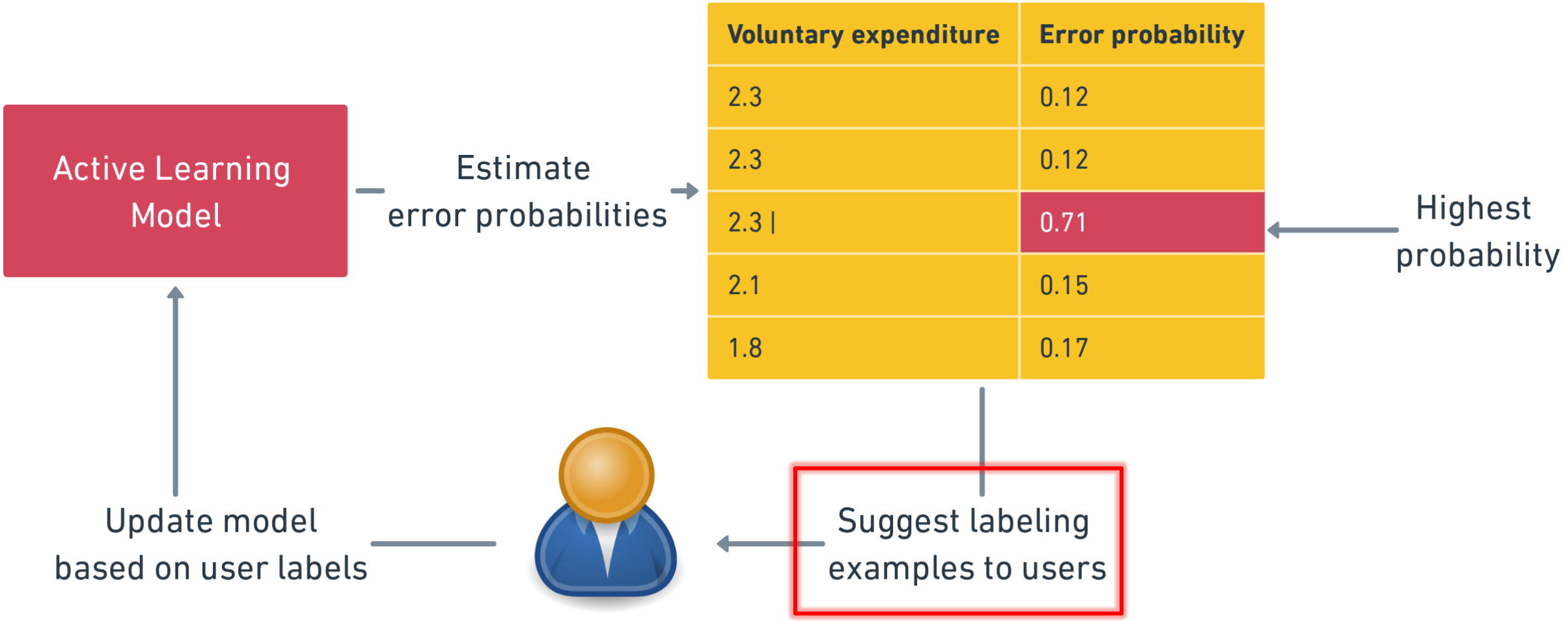
$$\neg BadSignal(s) \wedge HasSignal(c, s) \Rightarrow Error(c)$$



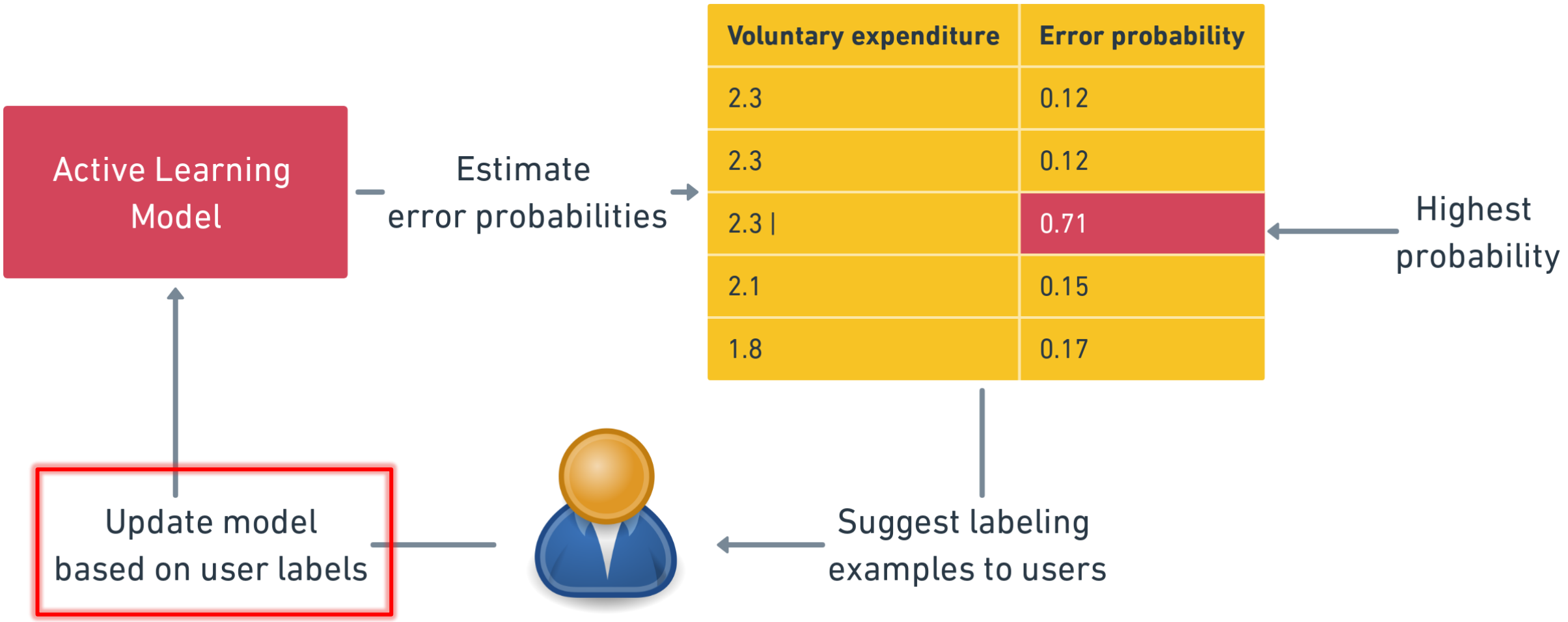
$$BadSignal(s) \wedge HasSignal(c, s) \Rightarrow \neg Error(c)$$



Active-learning error detection

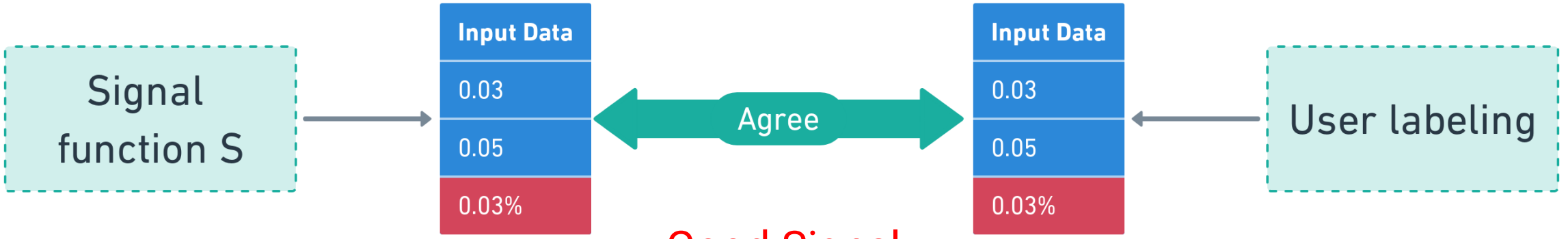


Active-learning error detection



PSL rules: Signal function and user labeling

$$Label(c, 1) \wedge HasSignal(c, s) \Rightarrow \neg BadSignal(s)$$

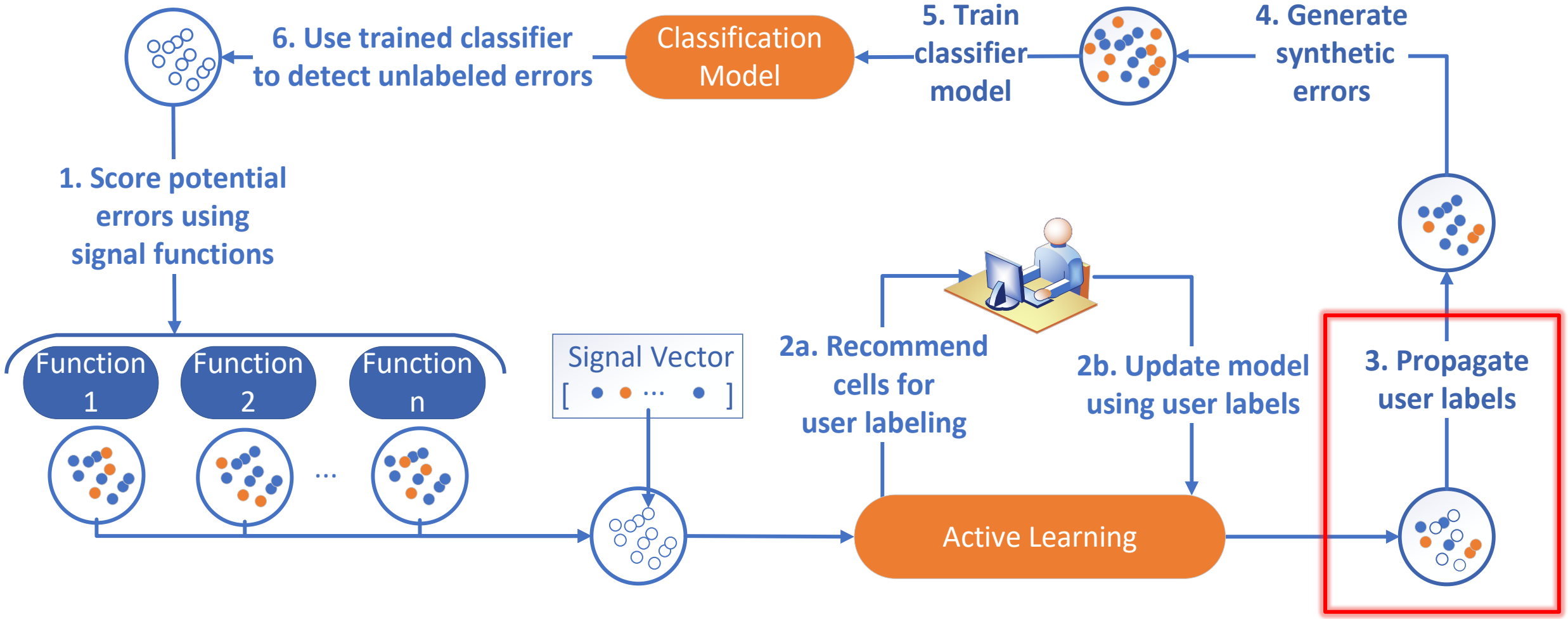


Good Signal

$$Label(c, 0) \wedge HasSignal(c, s) \Rightarrow BadSignal(s)$$

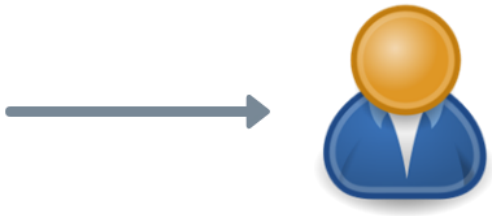
Bad Signal

Label propagation



Label propagation

Data	Score
San Francisco CA	0.7
Los Angeles	0.35
Springdale AR	0.71
Bend	0.11



Data	Score
San Francisco CA	0.7
Los Angeles	0.35
Springdale AR	0.71
Bend	0.11

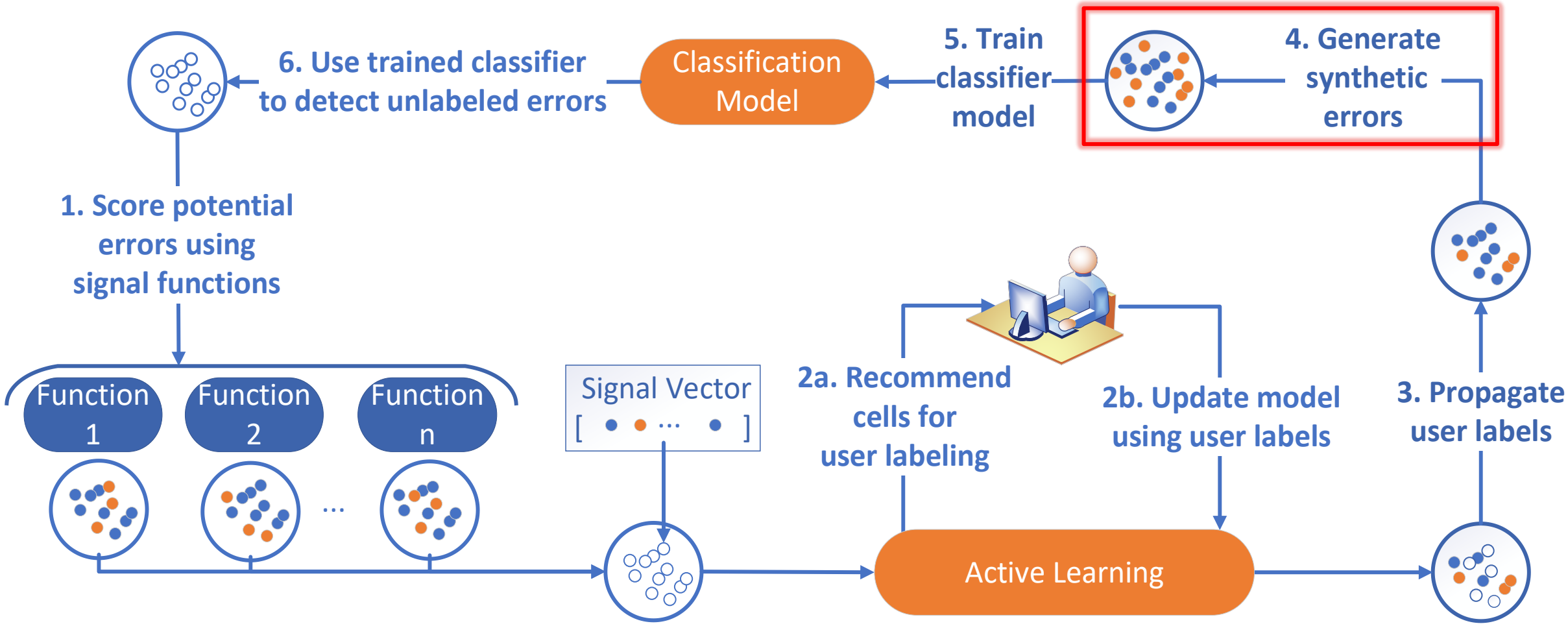
Data	Score
San Francisco CA	0.7
Los Angeles	0.35
Springdale AR	0.71
Bend	0.11

Label propagation

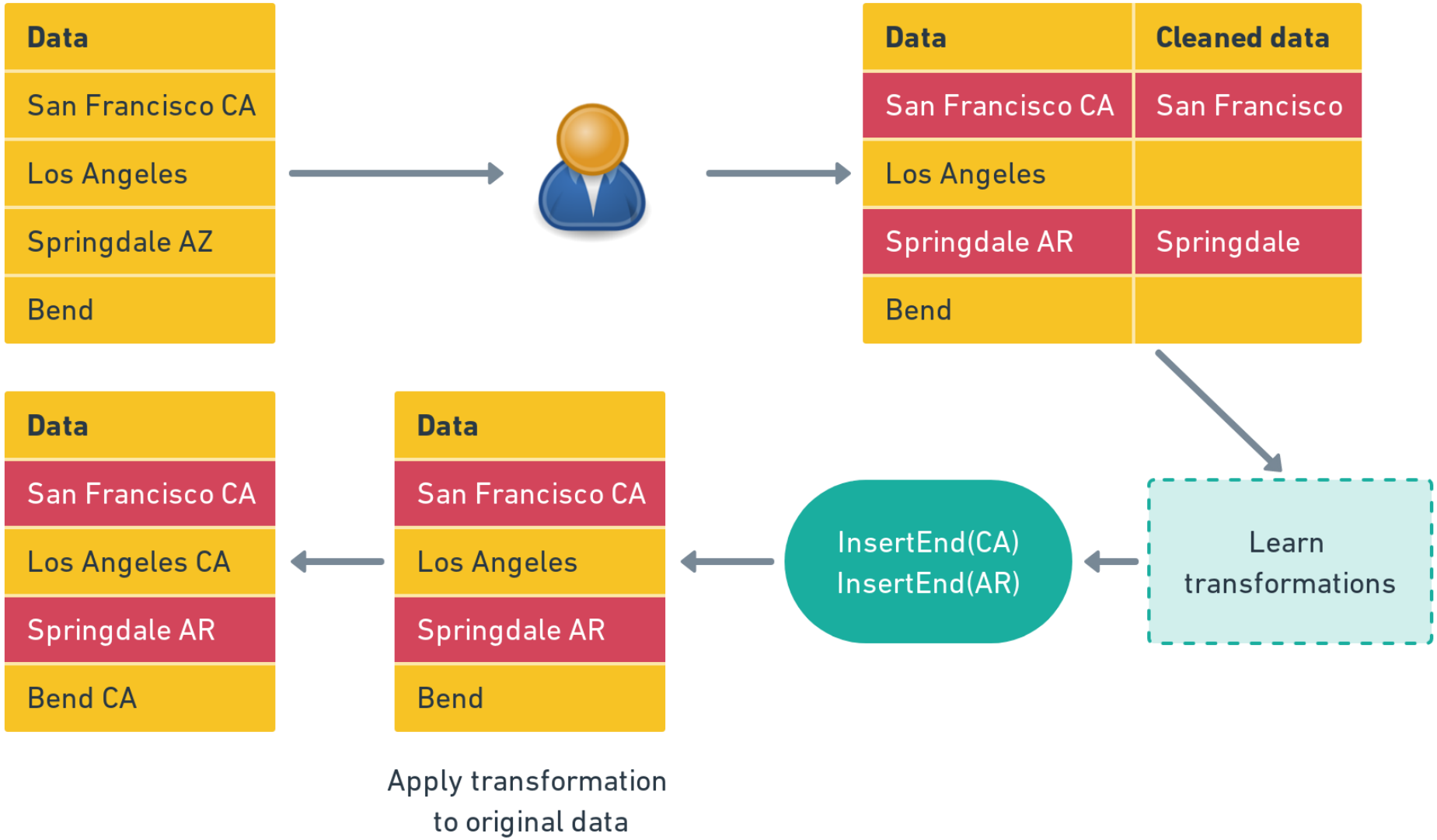
$$d(e) = |e_1 - e_2| \leq \epsilon$$

$$\epsilon = 0.1$$

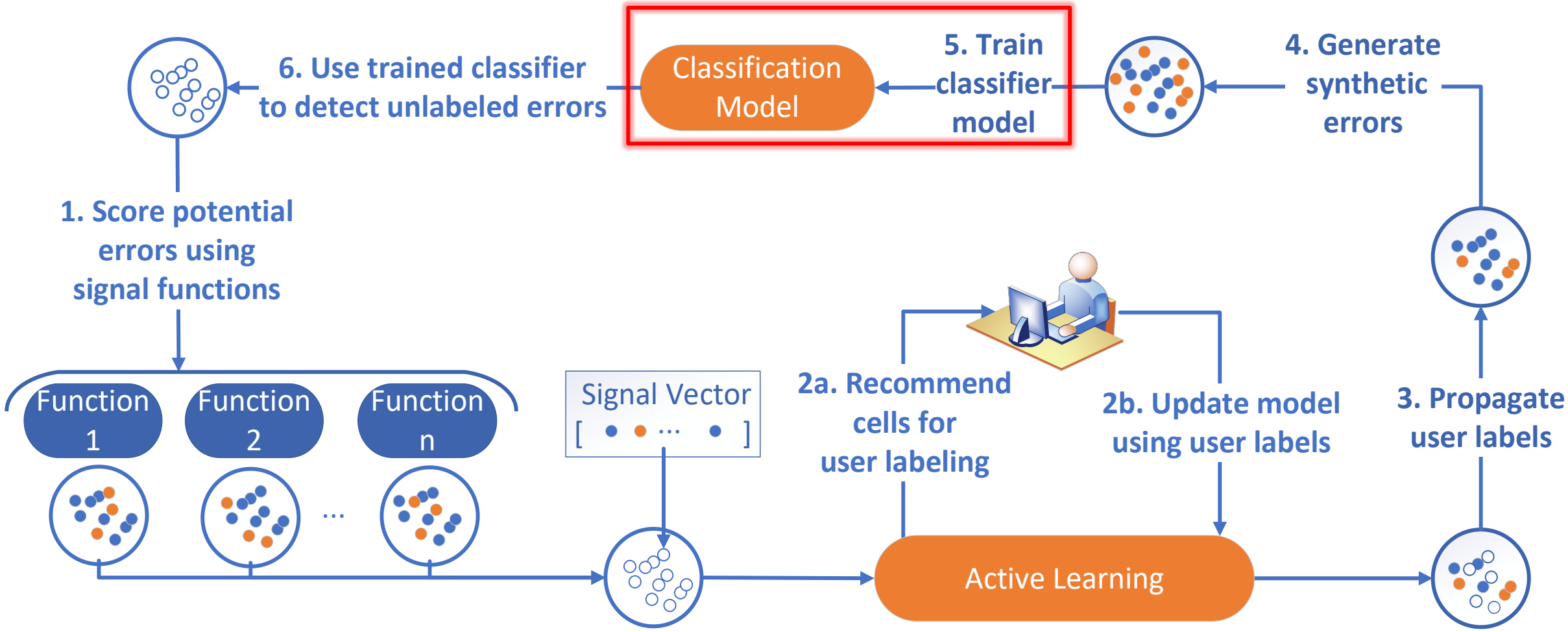
Error generation



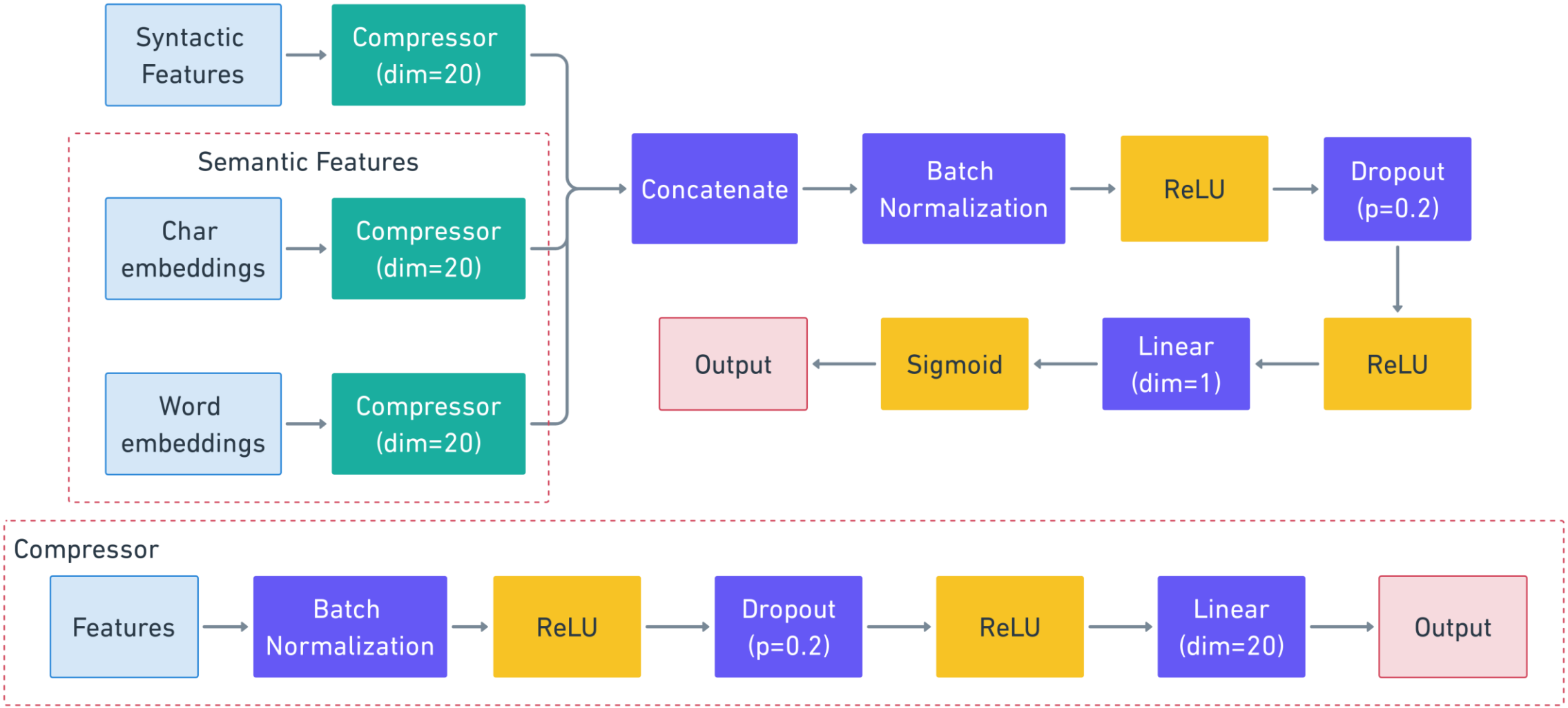
Error generation



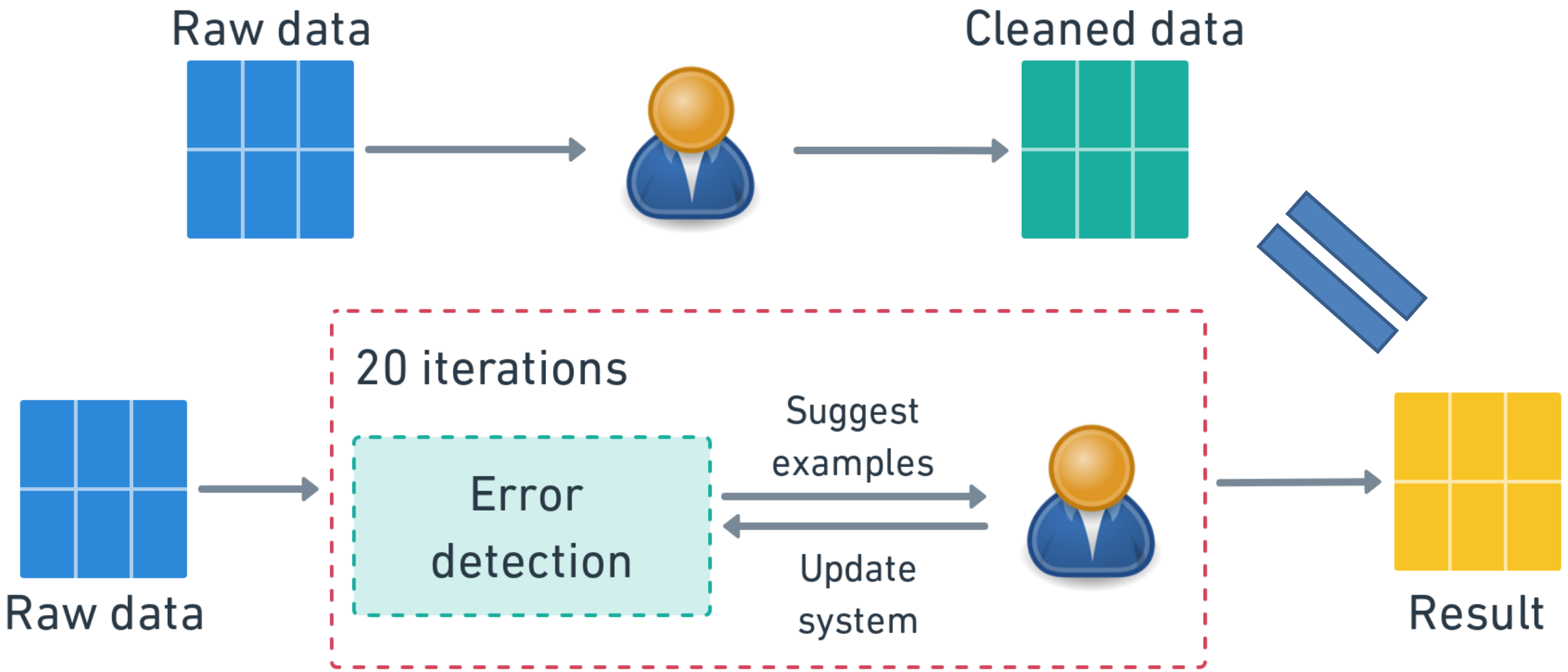
Classifier training



Training classifier model



Evaluation process



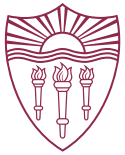
Evaluation result

- SPADE outperforms 6 different systems: Raha [Mahdavi et al., 2019], ED2 [Neutatz et al., 2019], dBoost [Mariet et al., 2016], NADEEF [Dallachiesa et al., 2013], KATARA [Chu et al., 2015], ActiveClean [Krishman et al., 2016]
  - Experiment on 5 datasets from Raha
  - Average of ten runs with $SD = \pm 0.01$, $*$: $SD = \pm 0.02$, $**$: $SD = \pm 0.03$

Approach	Hospital			Beers			Rayyan			Flights			Movies		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
dBoost	0.07	0.37	0.11	0.34	1.00	0.50	0.05	0.18	0.08	0.25	0.34	0.29	0.25	0.79	0.38
NADEEF	0.05	0.37	0.09	0.13	0.06	0.08	0.30	0.85	0.44	0.42	0.93	0.58	**1.00**	0.08	0.16
KATARA	0.44	0.11	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ActiveClean	0.02	0.15	0.04	0.16	1.00	0.28	0.09	**1.00**	0.16	0.30	**0.99**	0.46	0.06	**1.00**	0.12
ED2	0.45	0.29	0.33	**1.00**	0.96	0.98	0.80	0.69	0.74	0.79	0.63	0.68	0.93	0.05	0.13
Raha	**0.94**	0.59	0.72	0.99	0.99	0.99	**0.81**	0.78	0.79	**0.82**	0.81	**0.81**	0.85	0.88	0.86
SPADE	0.93	**1.00**	**0.96**	**1.00**	**1.00**	**1.00**	0.80*	0.92*	**0.85**	0.81**	0.81**	**0.81***	**0.99**	0.83	**0.90**

Summary

- Novel probabilistic active learning model for minimal user labeling
  - capture signals for both internal and external information
  - iteratively update model to recommend the most informative example
- Data augmentation process where we enrich our training datasets with synthetic data
  - propagate labeled data and generates additional errors
  - generalize better to unseen errors
- Semi-supervised approach for error detection with excellent performance



SYNTACTIC ERROR CORRECTION

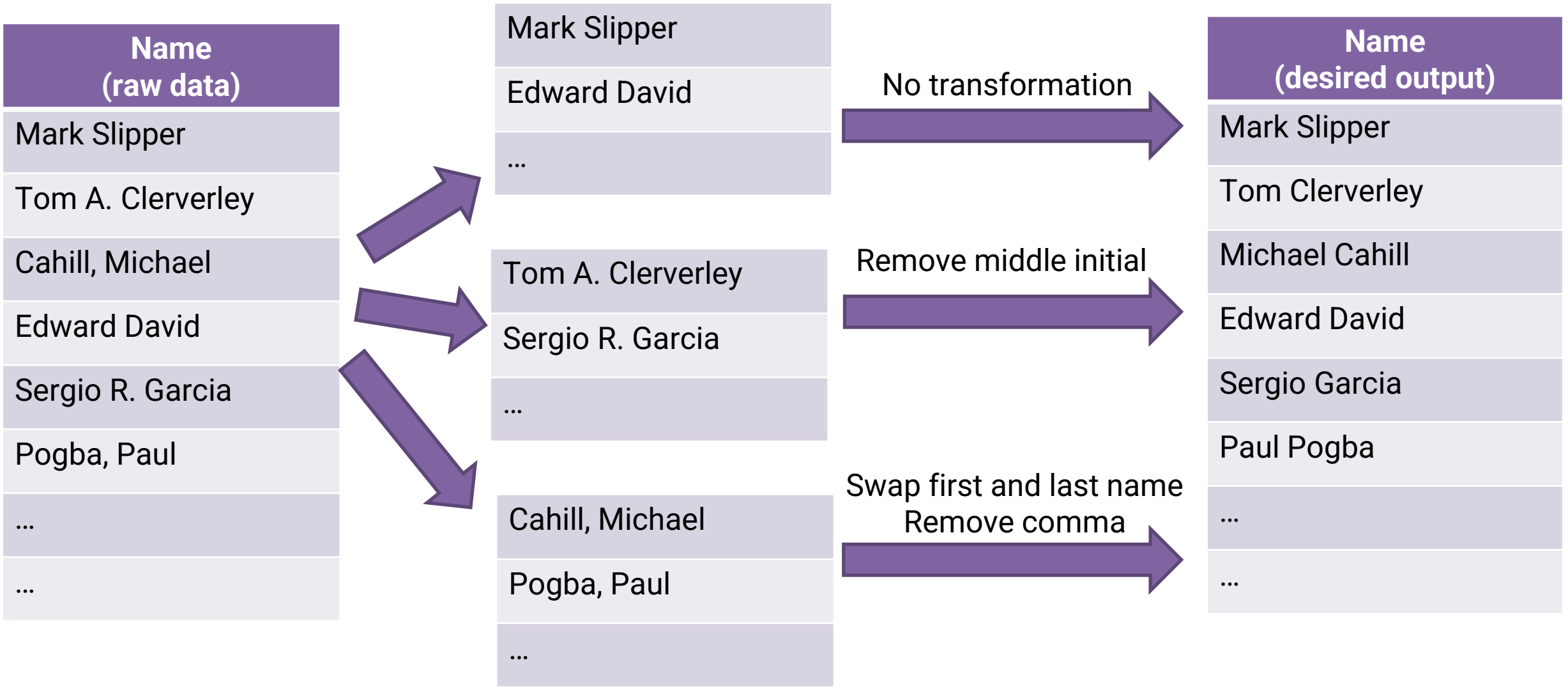
Motivating example: People names

Name (raw data)
Mark Slipper
Tom A. Clerverley
Cahill, Michael
Edward David
Sergio R. Garcia
Pogba, Paul
…
…

Normalize

Name (desired output)
Mark Slipper
Tom Clerverley
Michael Cahill
Edward David
Sergio Garcia
Paul Pogba
…
…

Motivating example: Transformation program



Motivating example: Programming-by-example

Name
Mark Slipper
Tom A. Clerverley
Cahill, Michael
Edward David
Sergio R. Garcia
Pogba, Paul

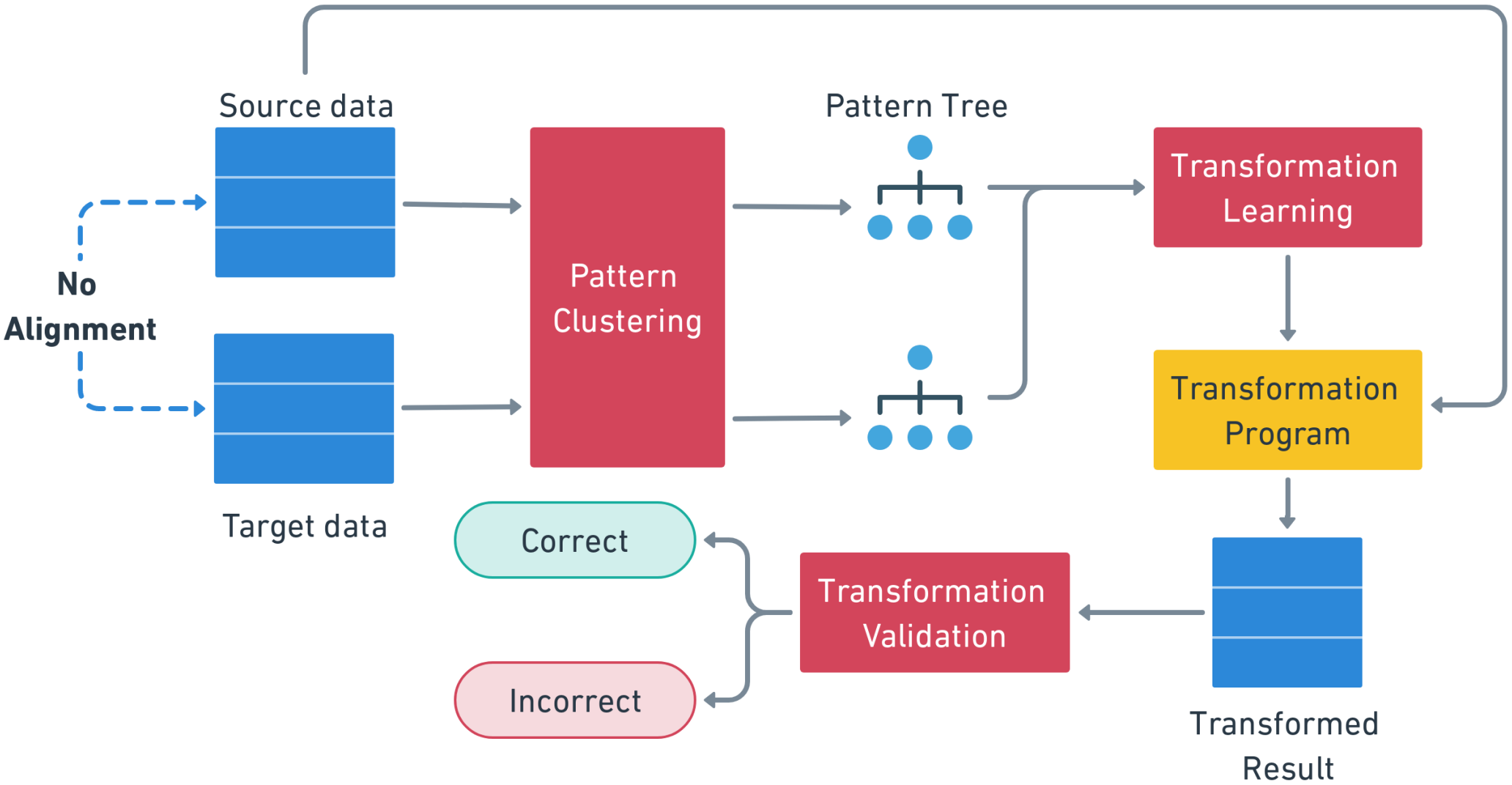
User label

Name
Mark Slipper
Tom Clerverley

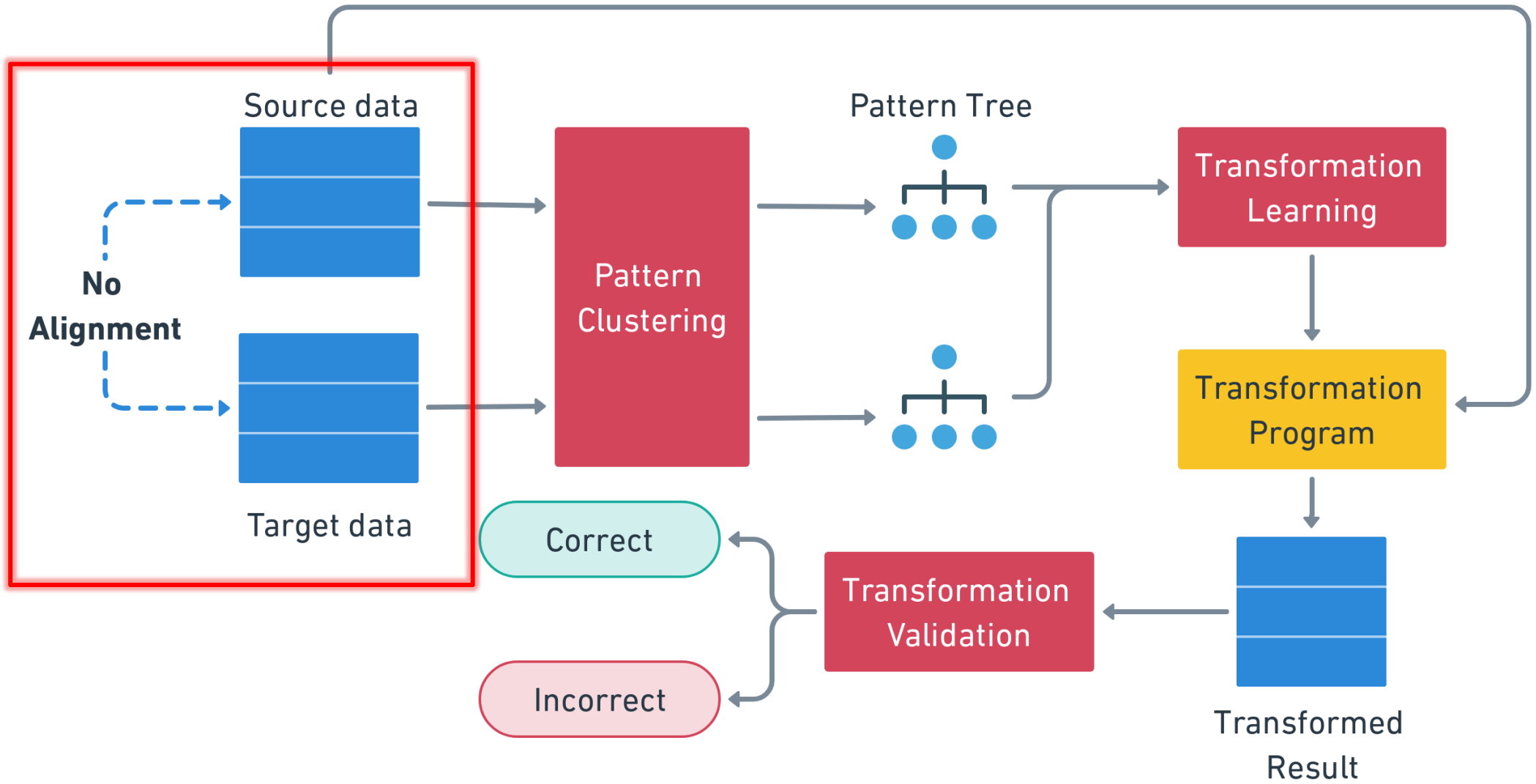
PBE Transform

Name
Mark Slipper
Tom Clerverley
Cahill Michael (X)
Edward David
Sergio Garcia
Pogba Paul (X)

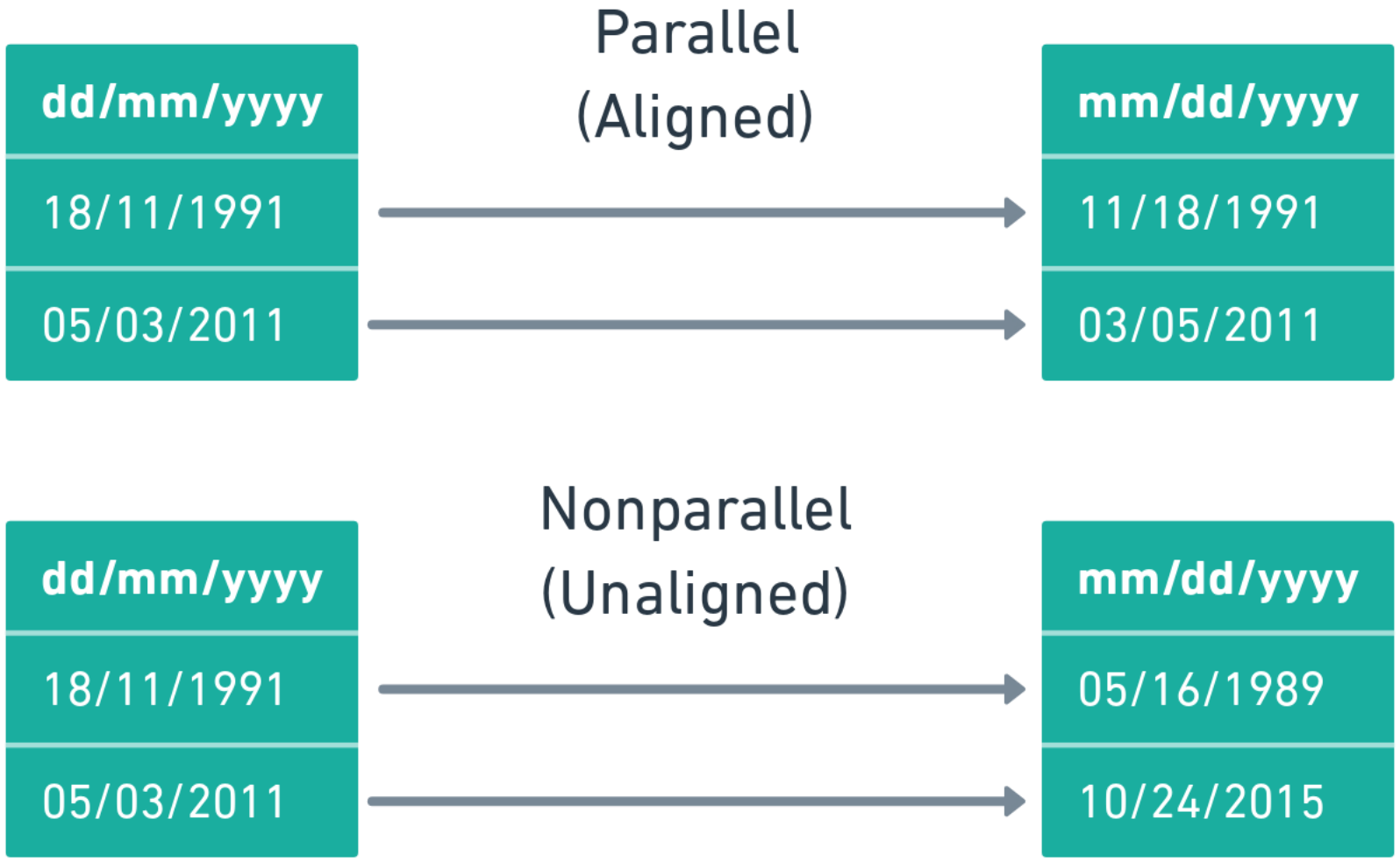
Overall approach



Overall approach



Parallel and nonparallel input-output



Input: Programming-by-example

Name
Mark Slipper
Tom A. Clerverley
Cahill, Michael
Edward David
Sergio R. Garcia
Pogba, Paul

User label

Name
Mark Slipper
Tom Clerverley

PBE Transform

Name
Mark Slipper
Tom Clerverley
Cahill Michael (X)
Edward David
Sergio Garcia
Pogba Paul (X)

Input: Our method

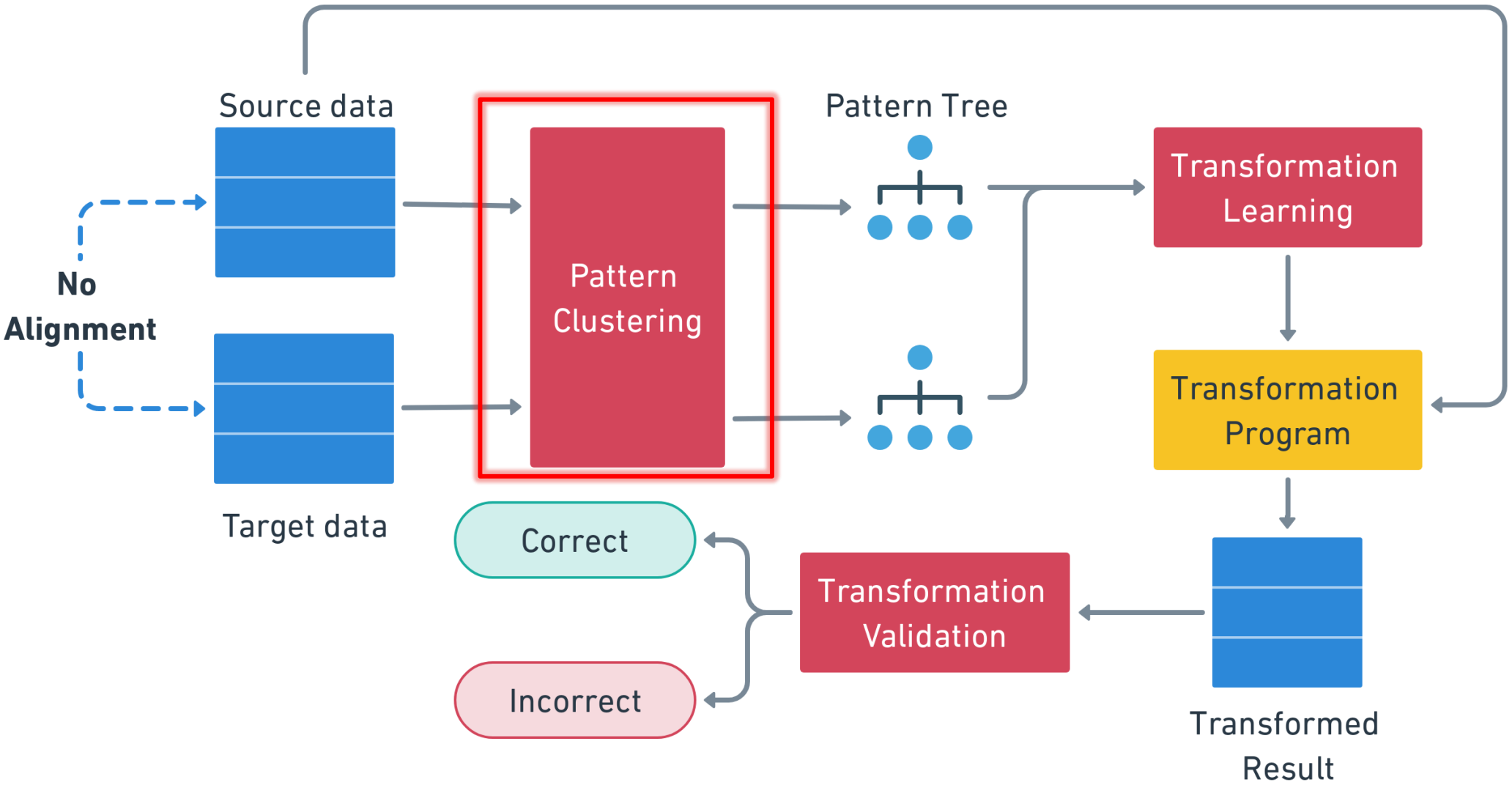
Name
Mark Slipper
Tom A. Clerverley
Cahill, Michael
Edward David
Sergio R. Garcia
Pogba, Paul

Example of desired format

Name
Lionel Messi
Cristiano Ronaldo

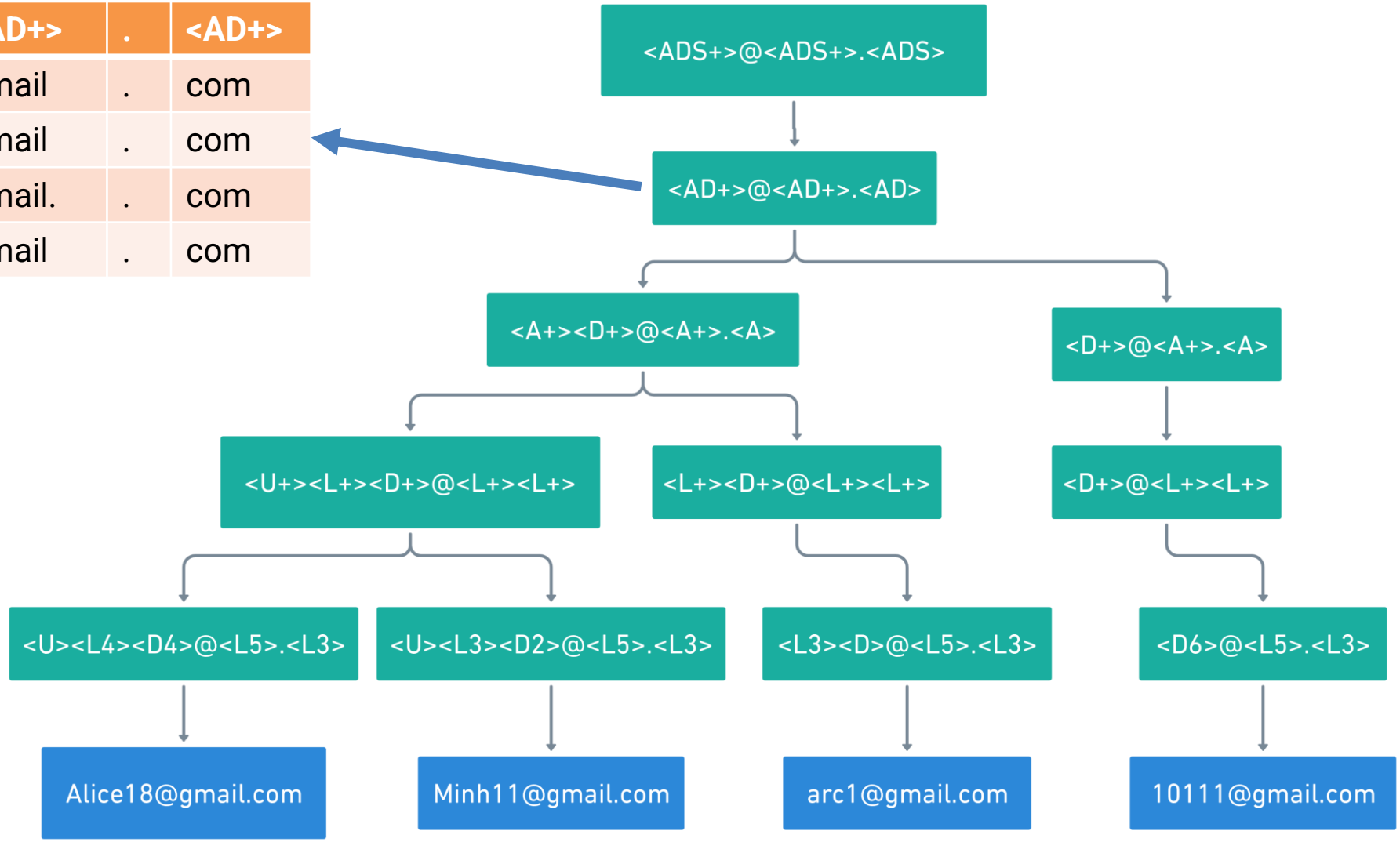
Name
Mark Slipper
Tom Clerverley
Michael Cahill
Edward David
Sergio Garcia
Paul Pogba

Pattern clustering

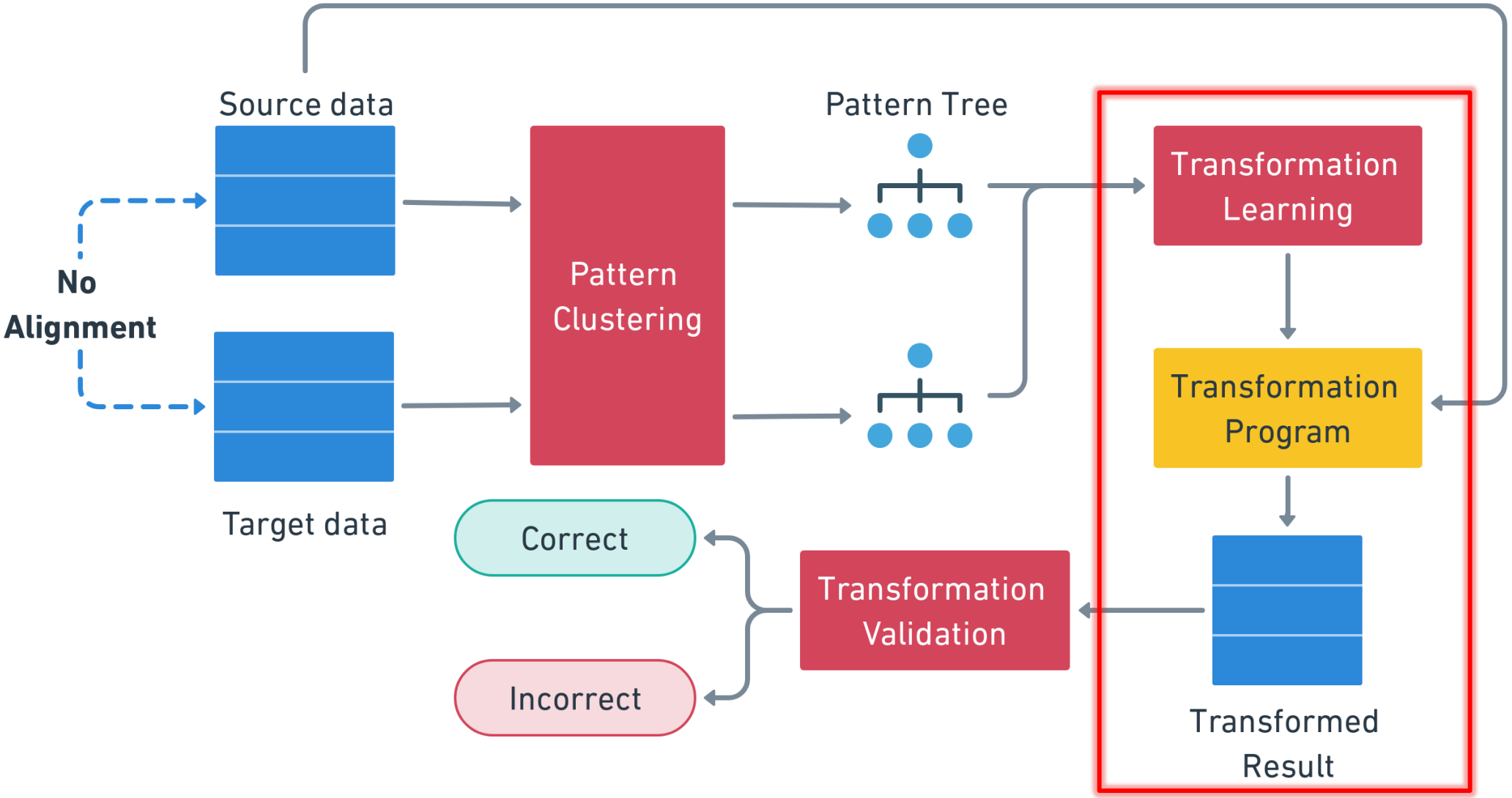


Pattern tree

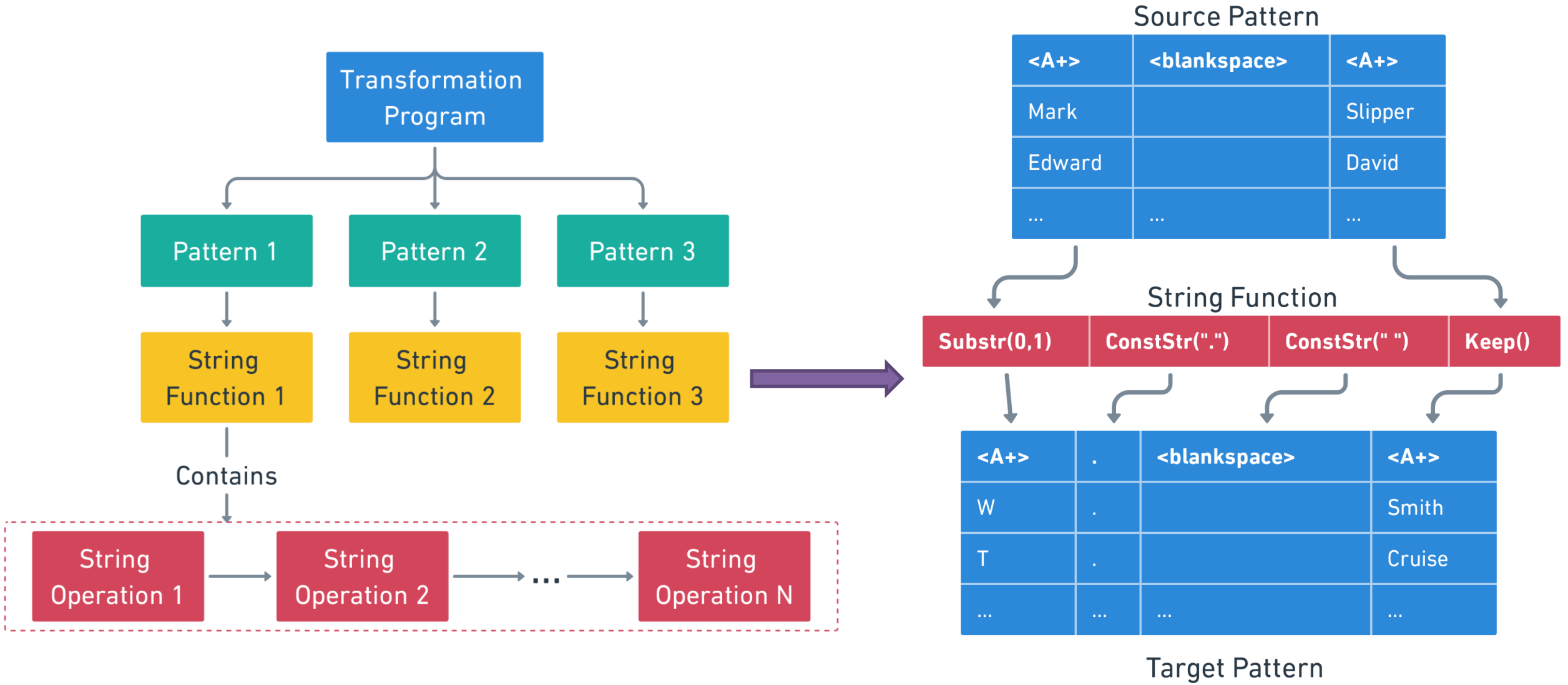
<AD+>	@	<AD+>	.	<AD+>
Alice1811	@	gmail	.	com
Minh11	@	gmail	.	com
arc1	@	gmail.	.	com
10111	@	gmail	.	com



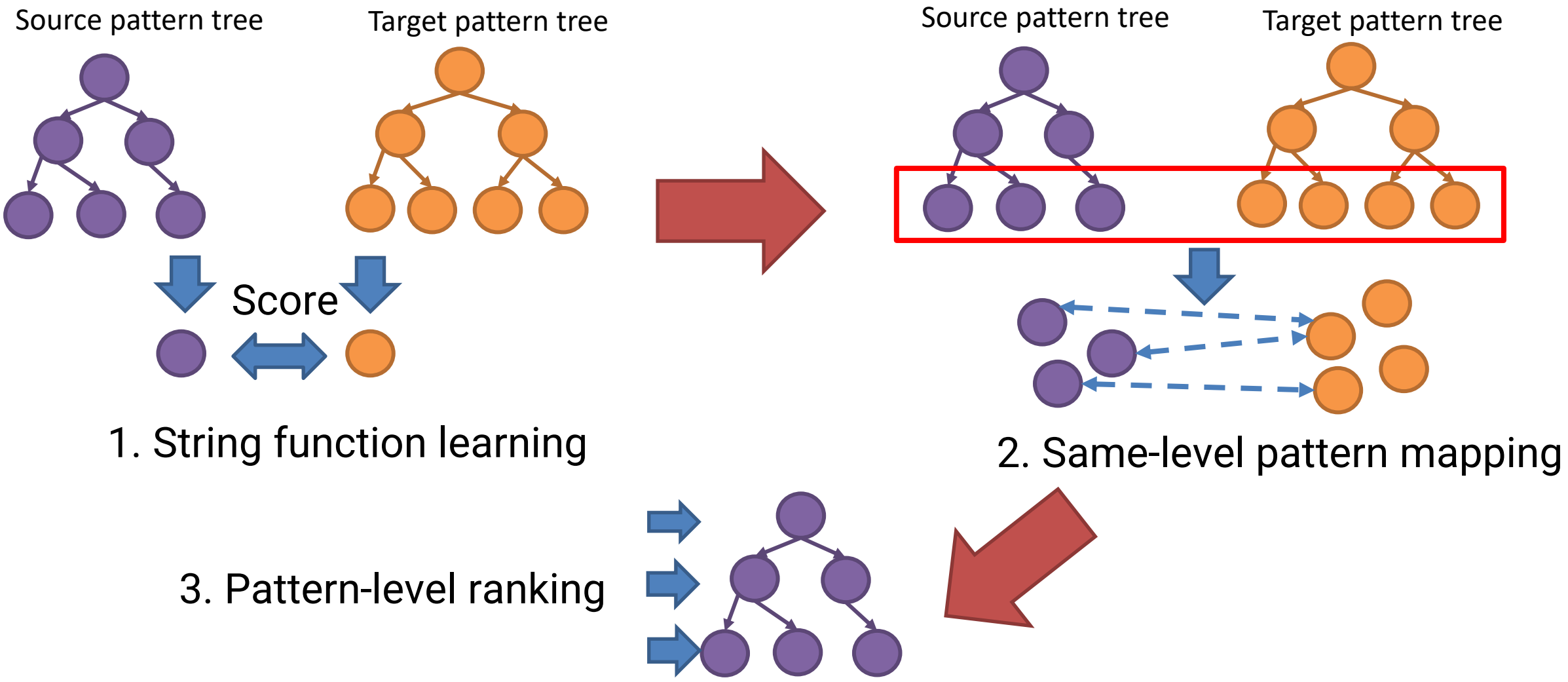
Transformation learning



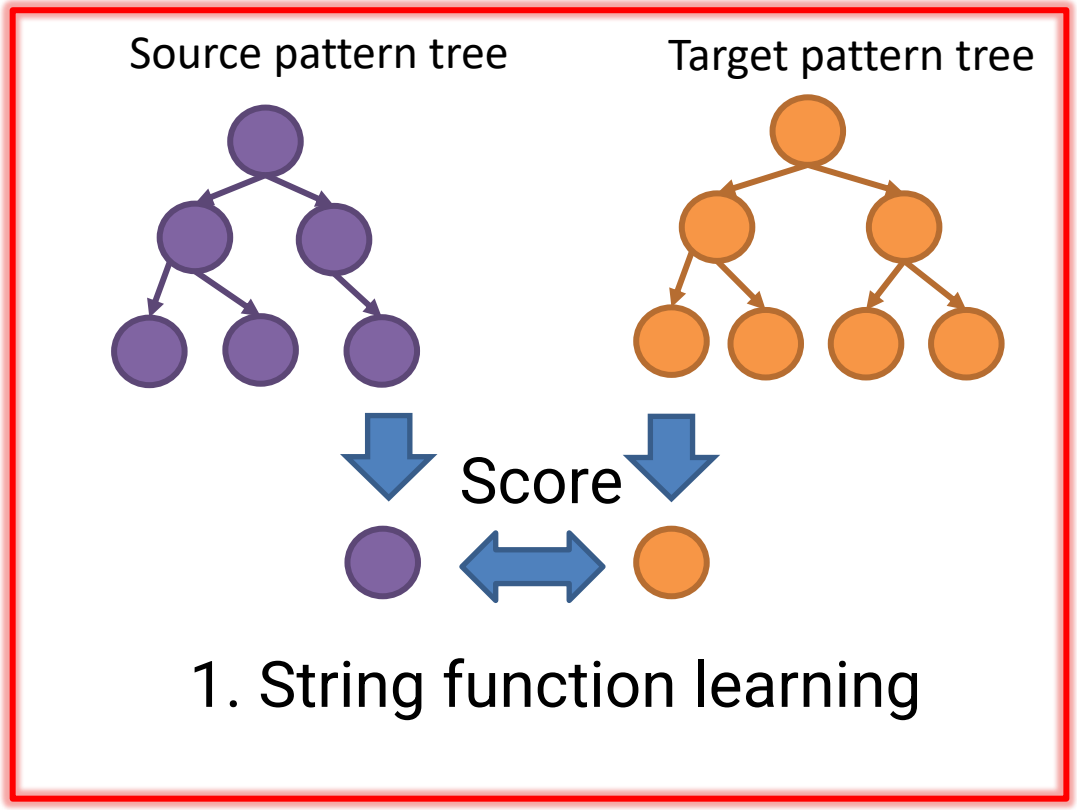
Transformation program

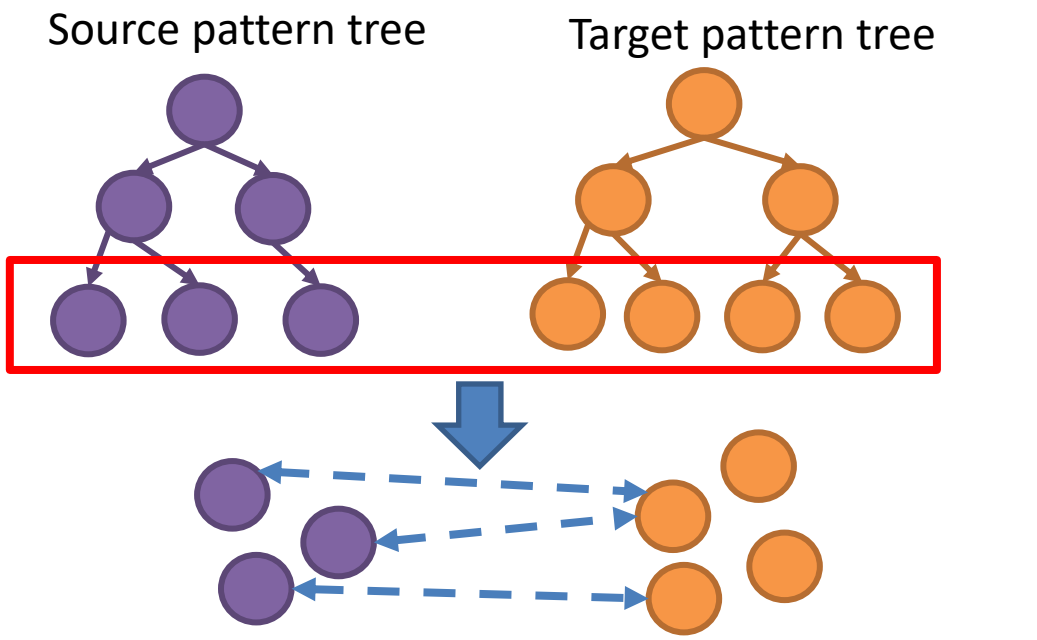


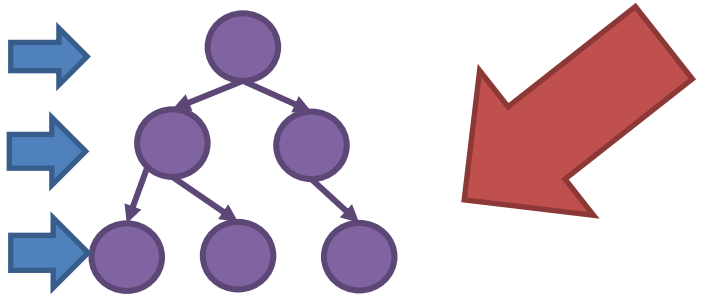
Transformation learning



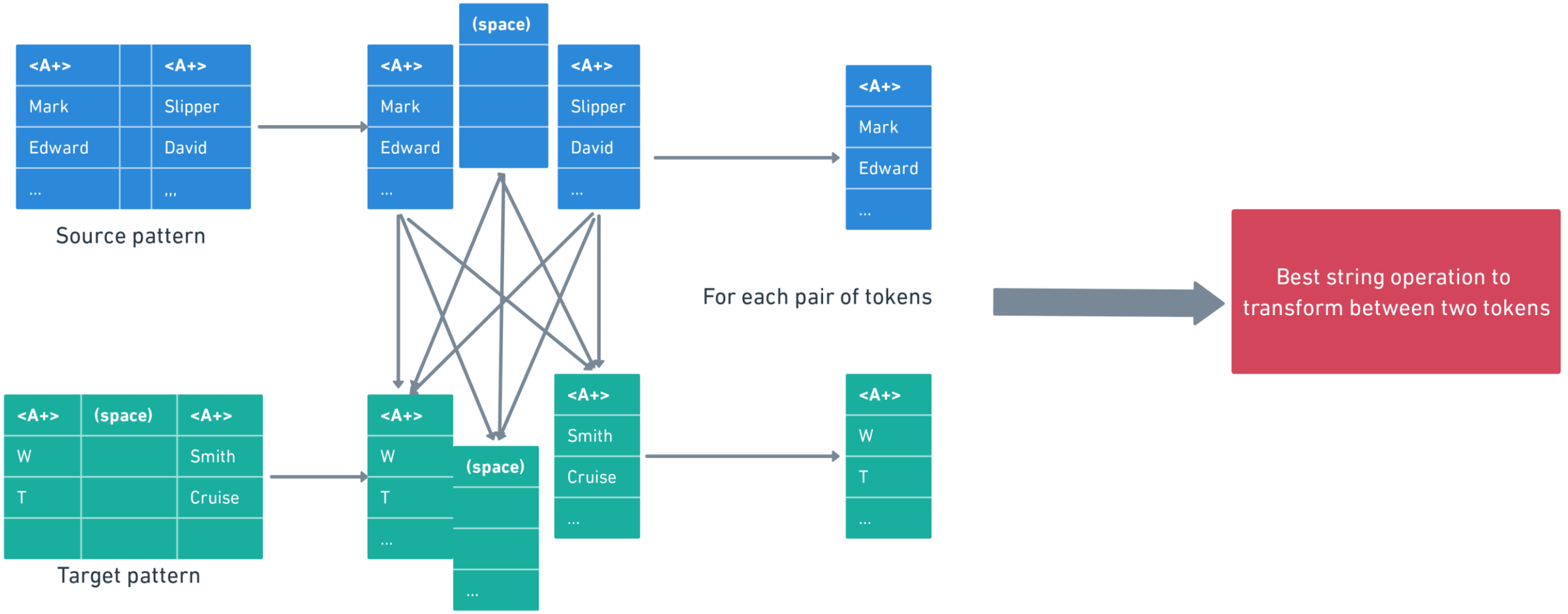
String function learning



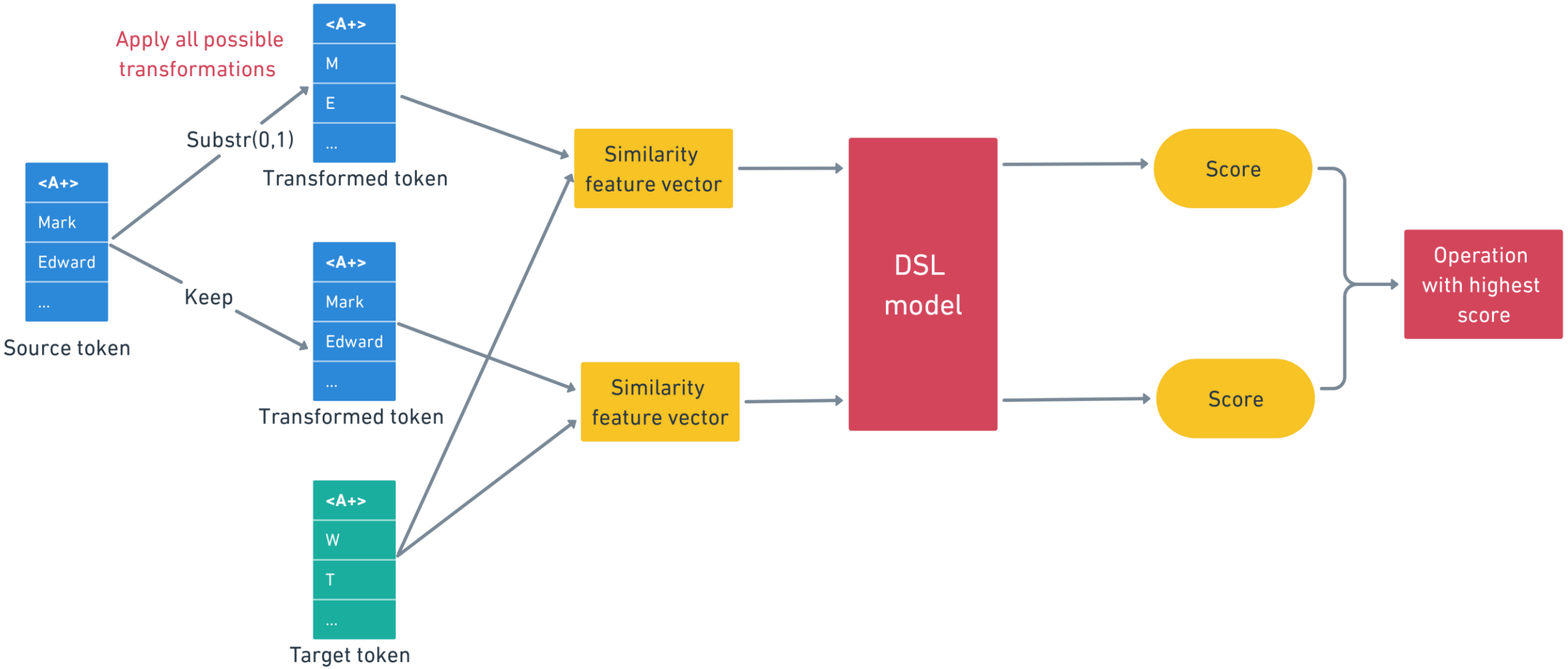




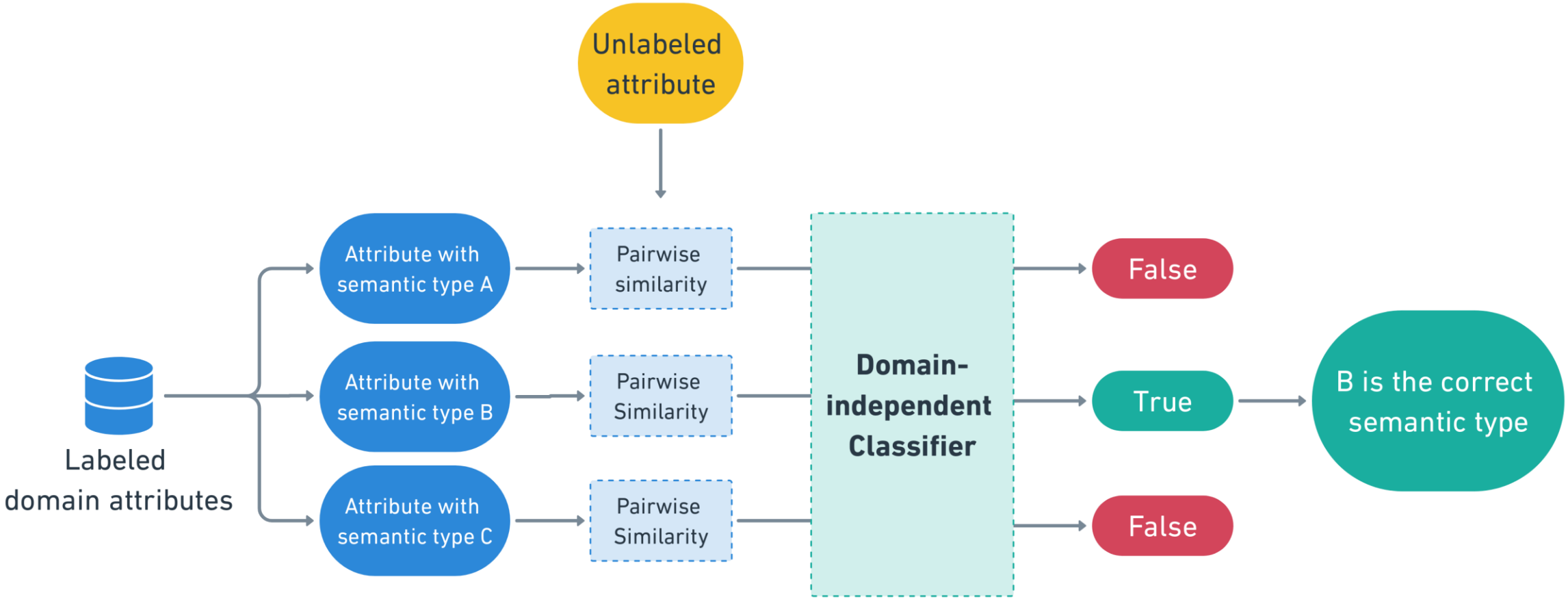
String function learning



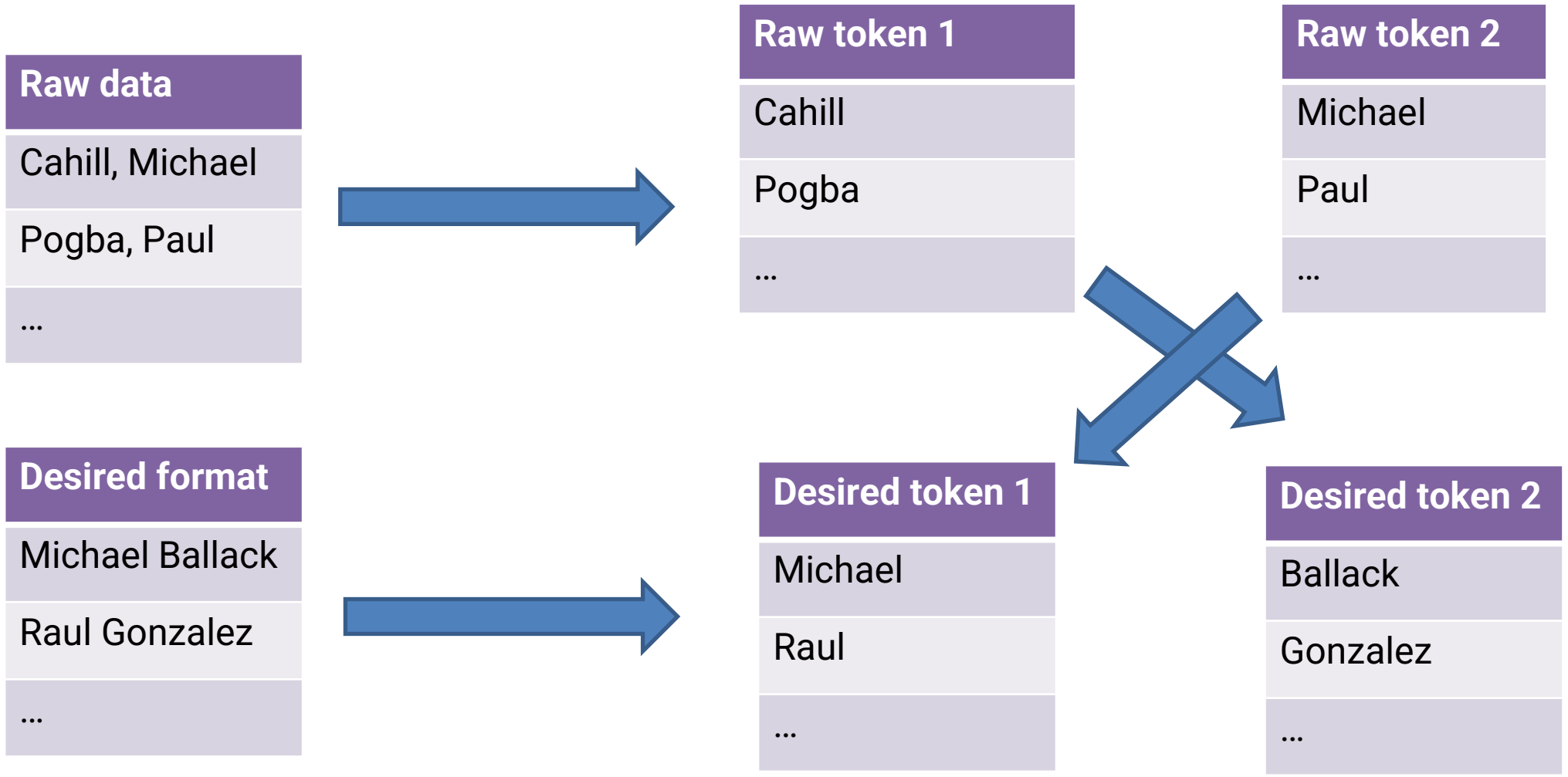
String function learning



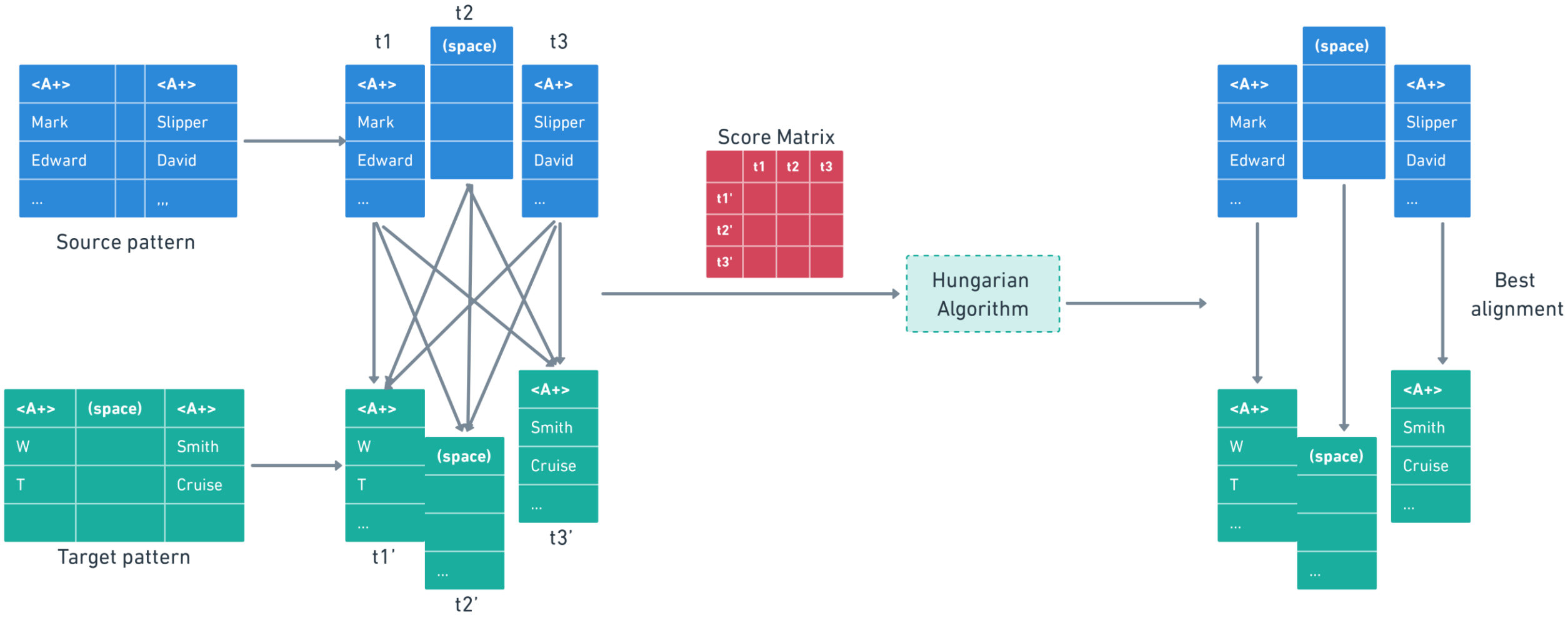
DSL model (ISWC 2016)



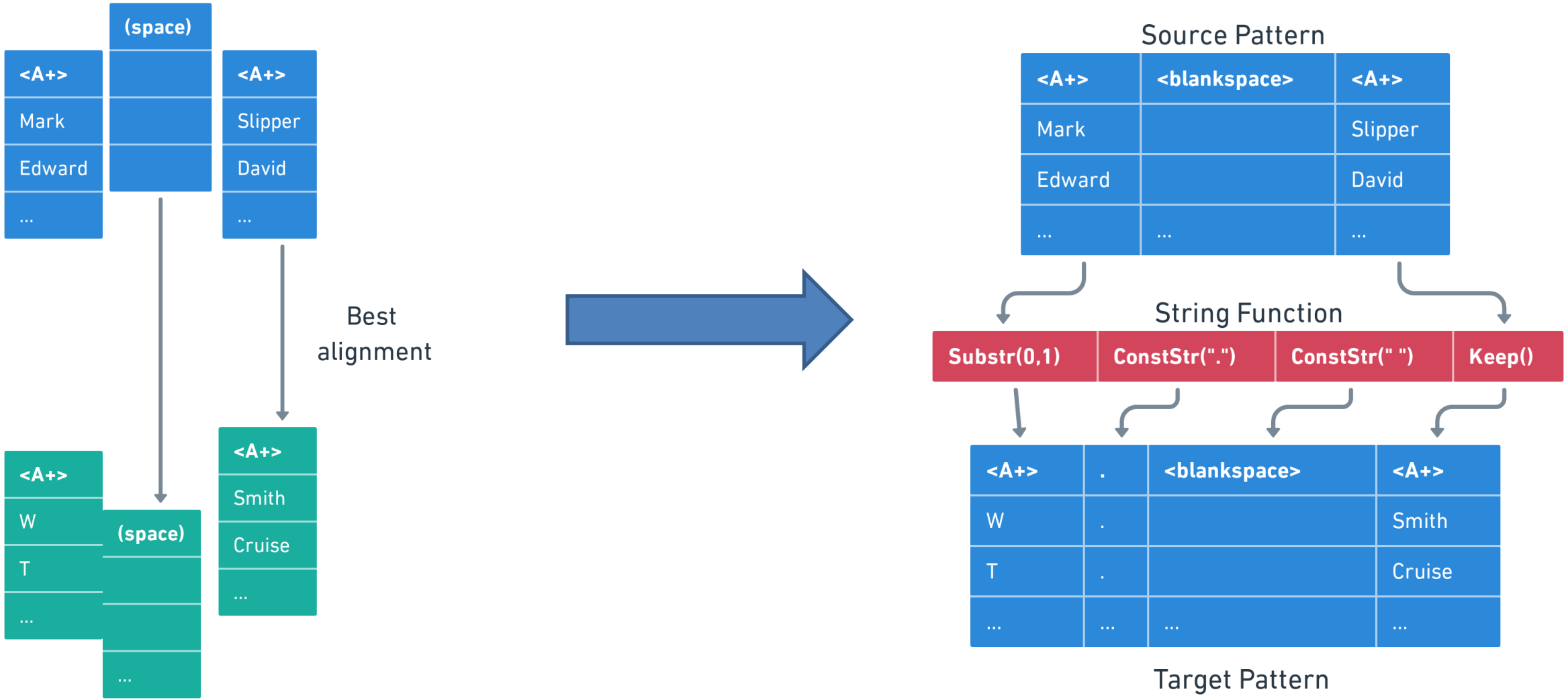
DSL model


Raw data
Cahill, Michael
Pogba, Paul
…
Raw token 1
Cahill
Pogba
…
Raw token 2
Michael
Paul
…
Desired format
Michael Ballack
Raul Gonzalez
…
Desired token 1
Michael
Raul
…
Desired token 2
Ballack
Gonzalez
…

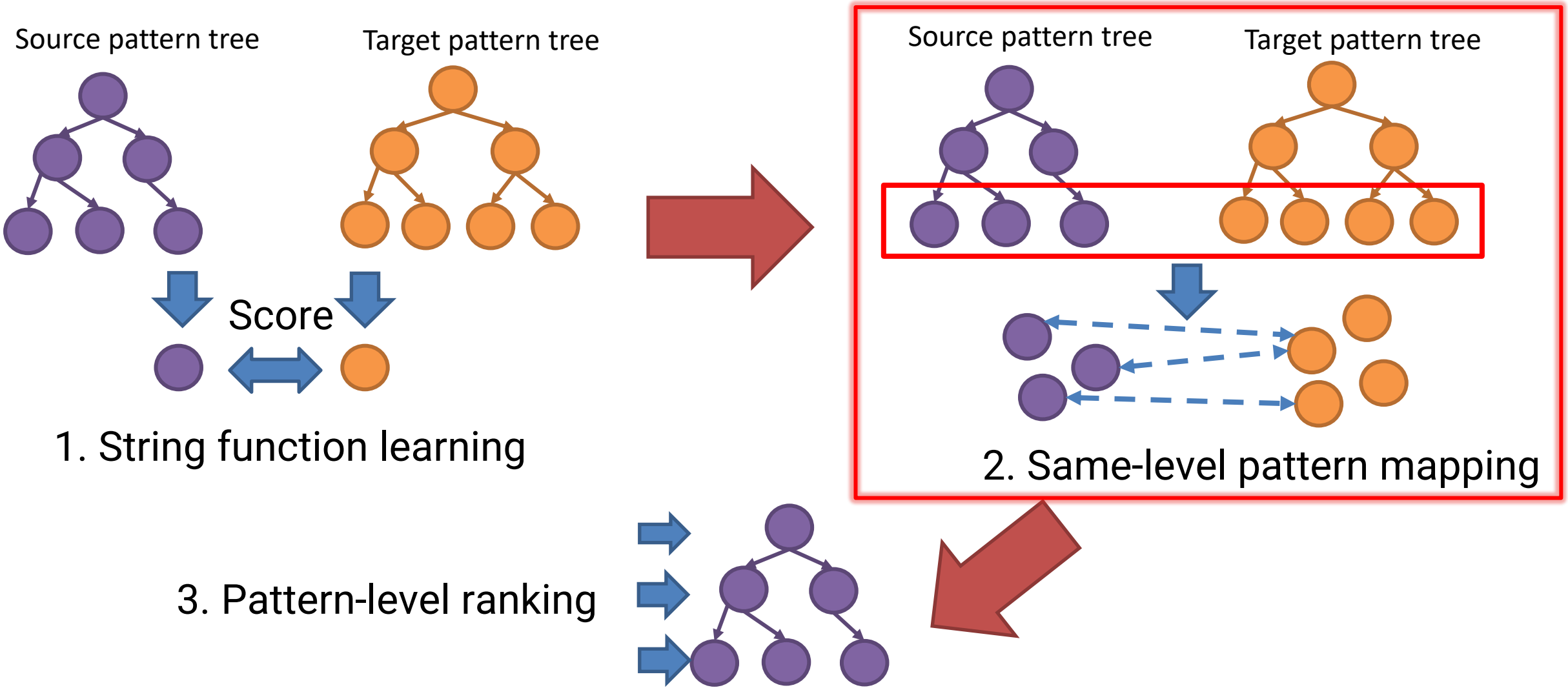
String function learning



String function learning - Output



Same-level pattern mapping

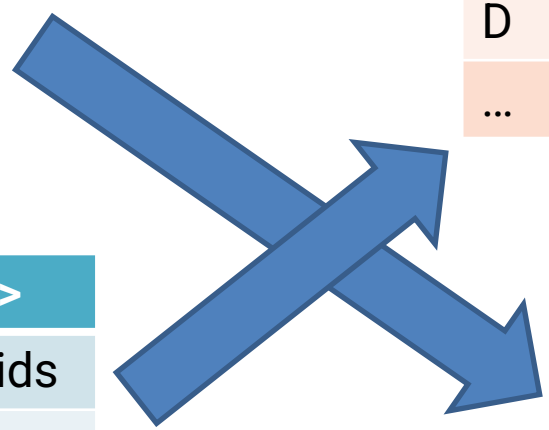


Same-level pattern mapping

<A+>	(space)	<A+>
Mark		Slipper
Edward		David
…	…	…

<A+>	(space)	<A+>		<A+>
Edgar		Steven		Davids
Jose		Luis		Garcia
…	…	…	…	

Source patterns



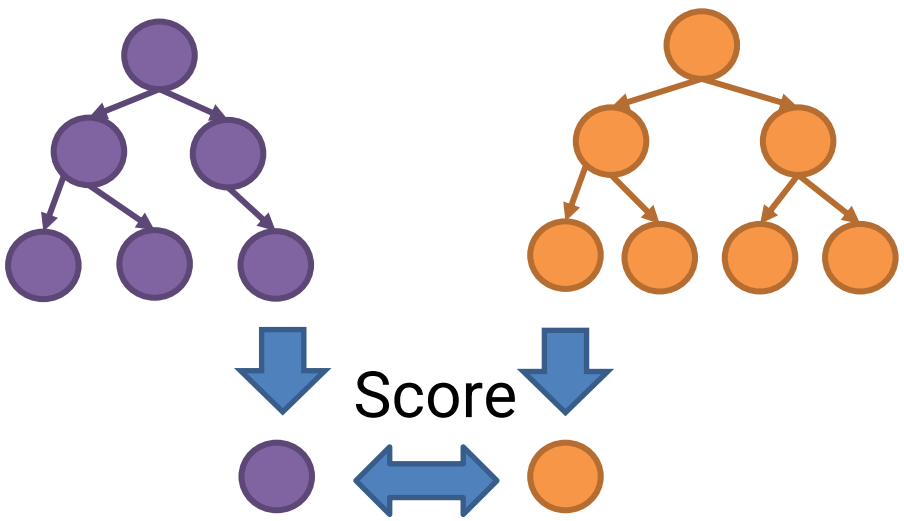
<A+>	.	<A+>	.	(space)	<A+>
J	.	P	.		Marquess
D	.	C	.		Leary
…	…			…	…

<A+>	.	(space)	<A+>
W	.		Smith
T	.		Cruise
…	…	…	…

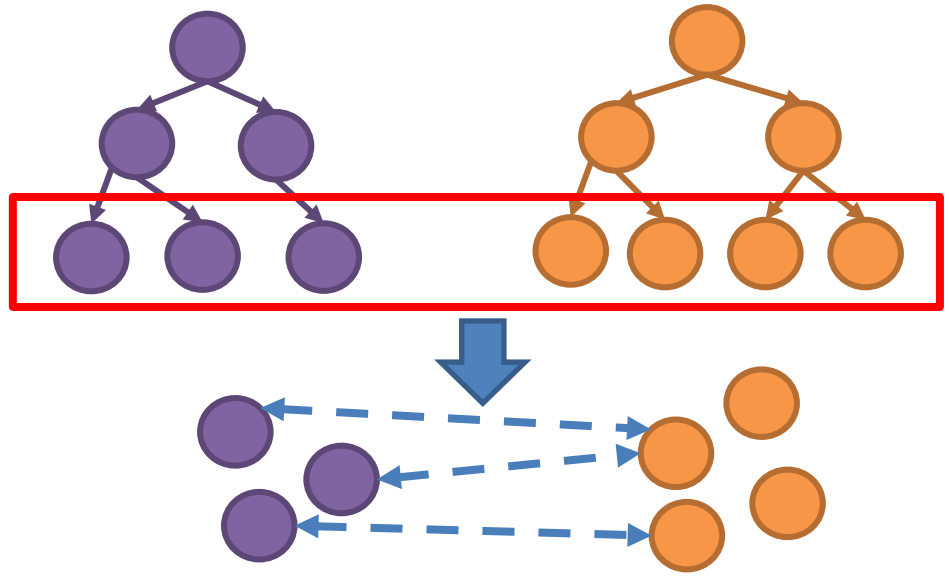
Target patterns

Find the best mapping based on sum of DSL scores

Pattern-level ranking

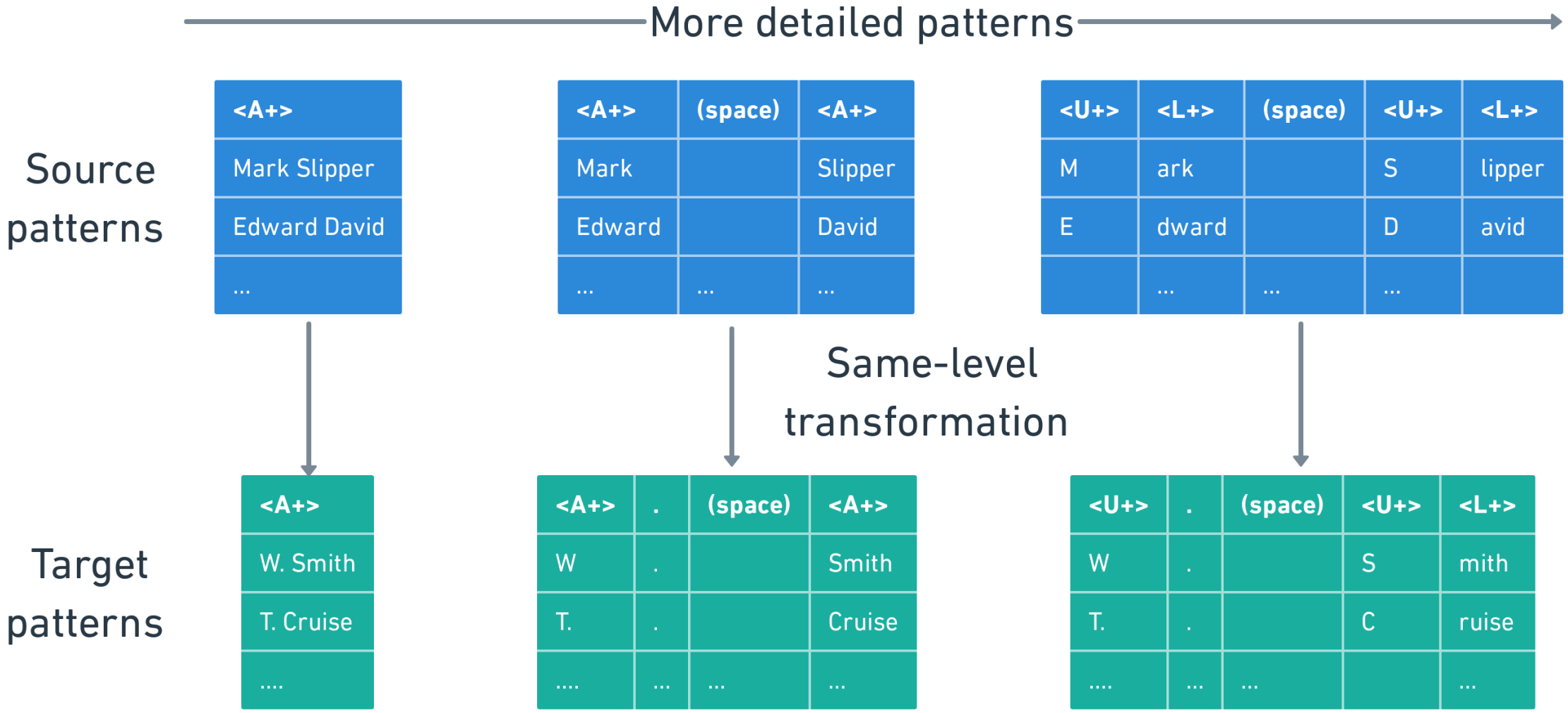


1. String function learning



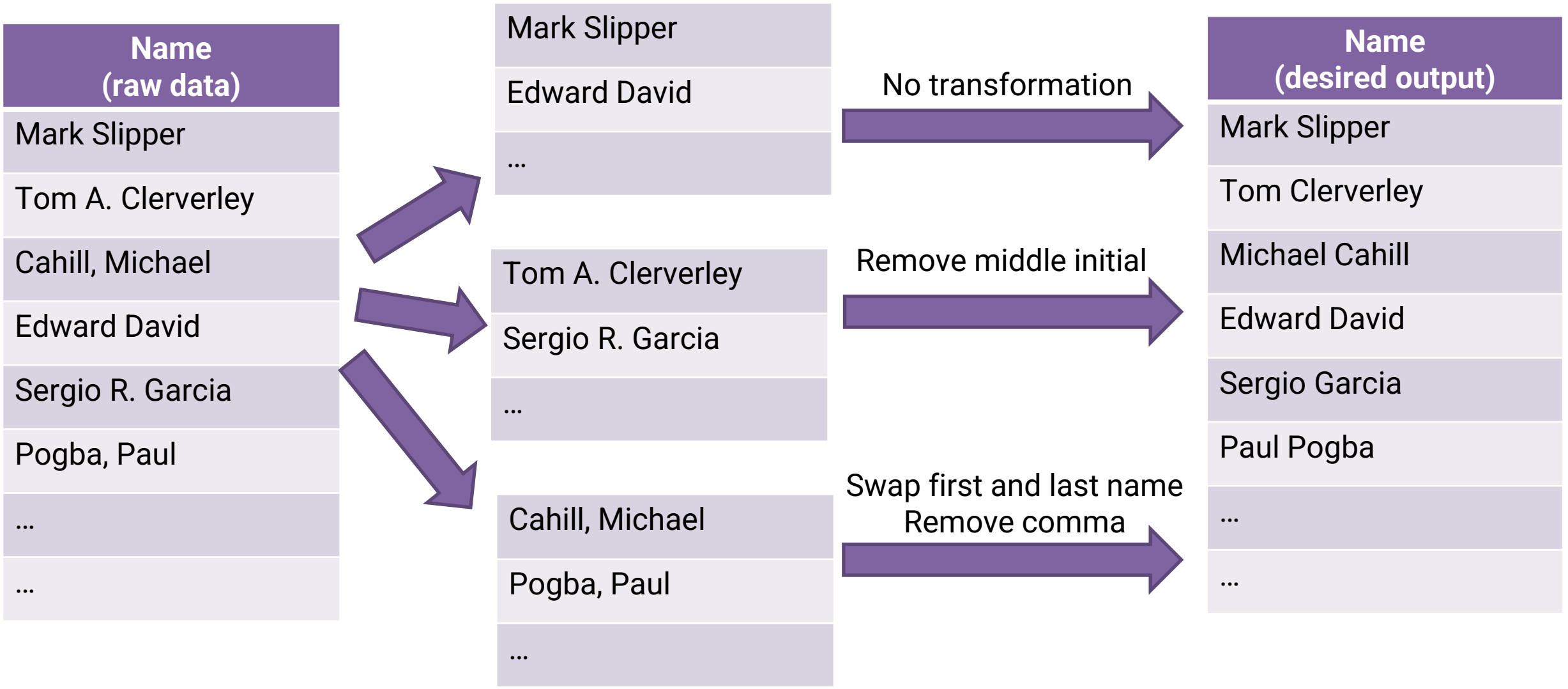
2. Same-level pattern mapping

Pattern-level ranking

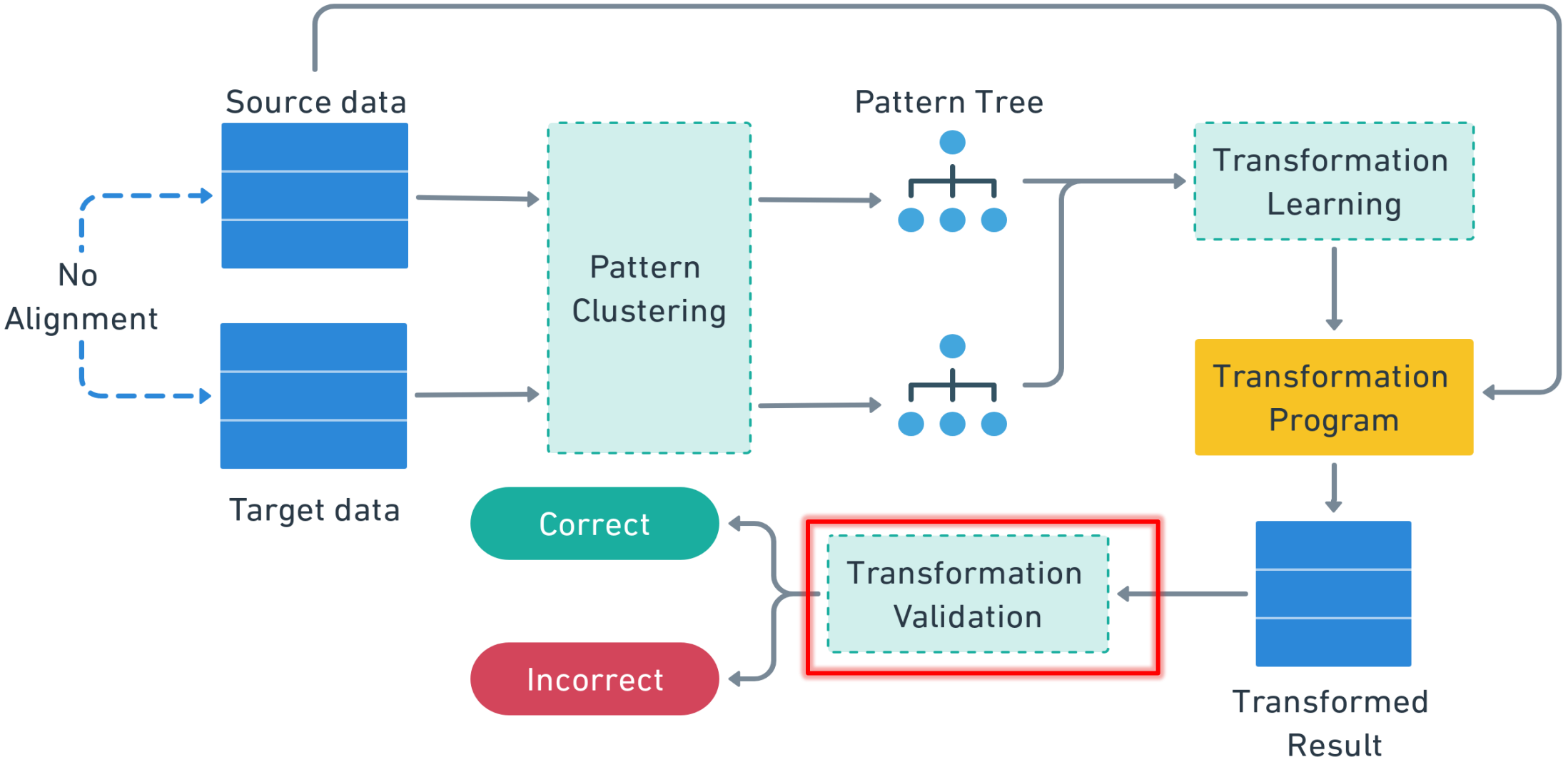


Choose the best pattern level to learn transformations based on DSL scores

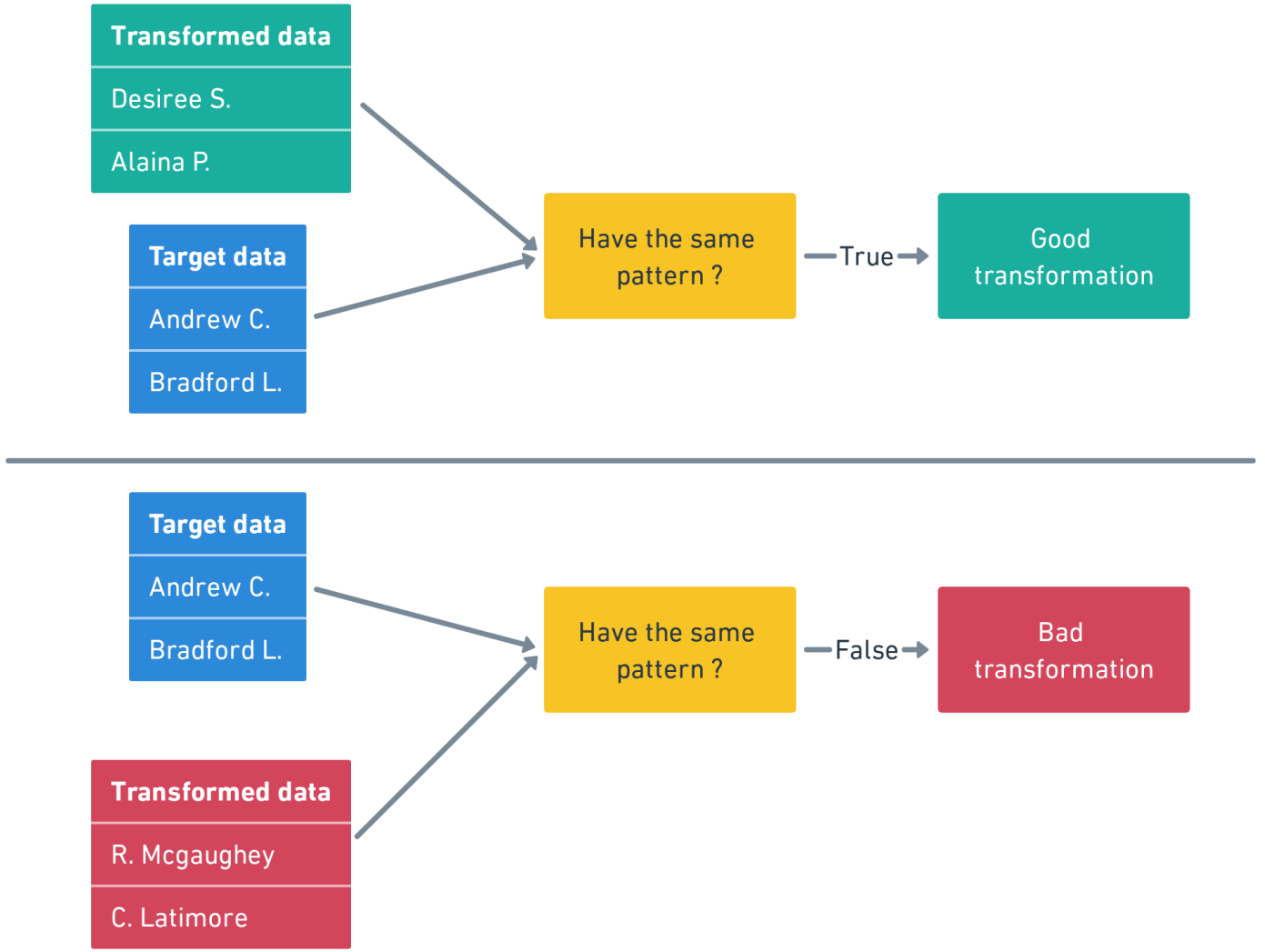
Output: Transformation program



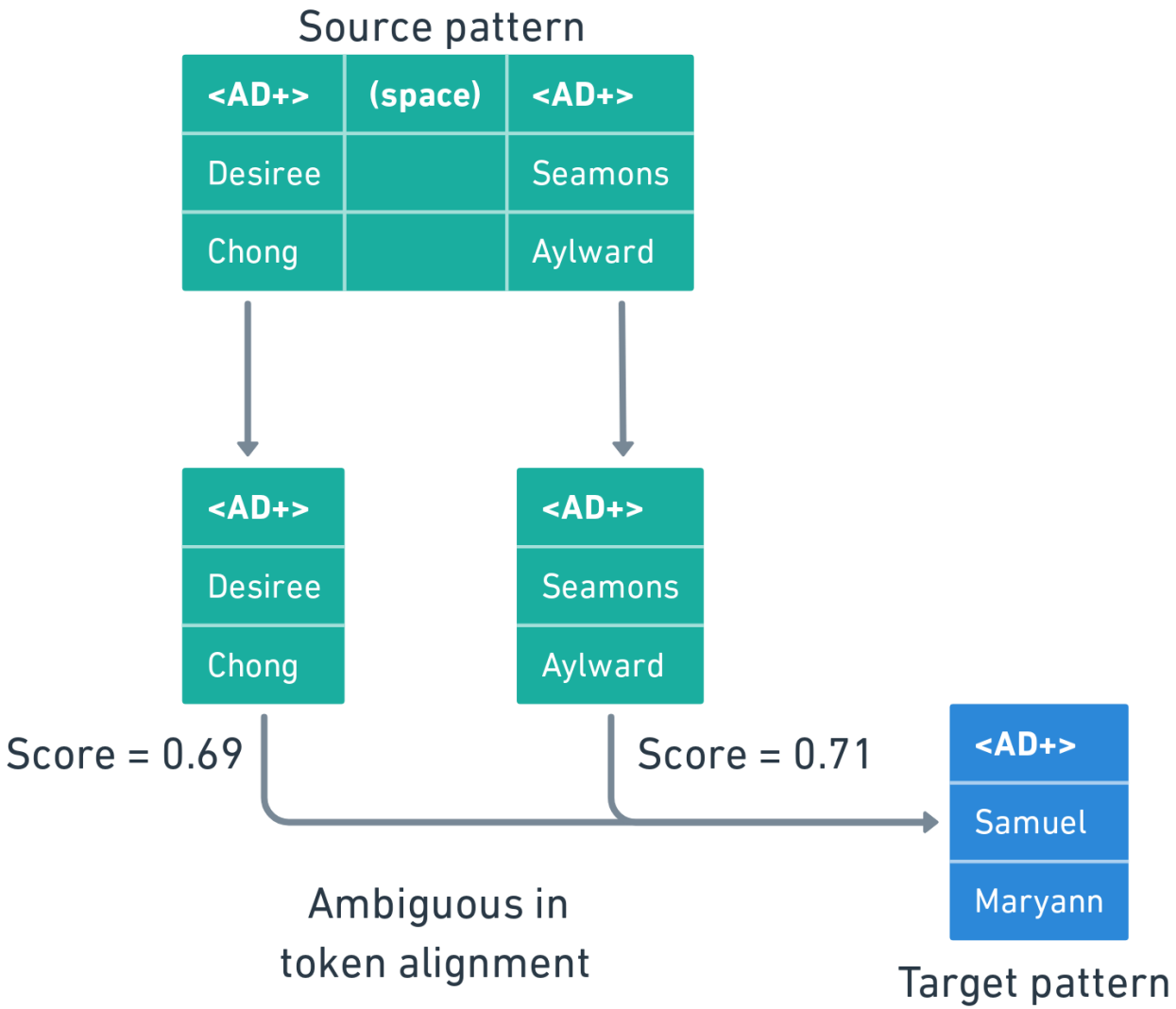
Transformation validation



Transformation validation

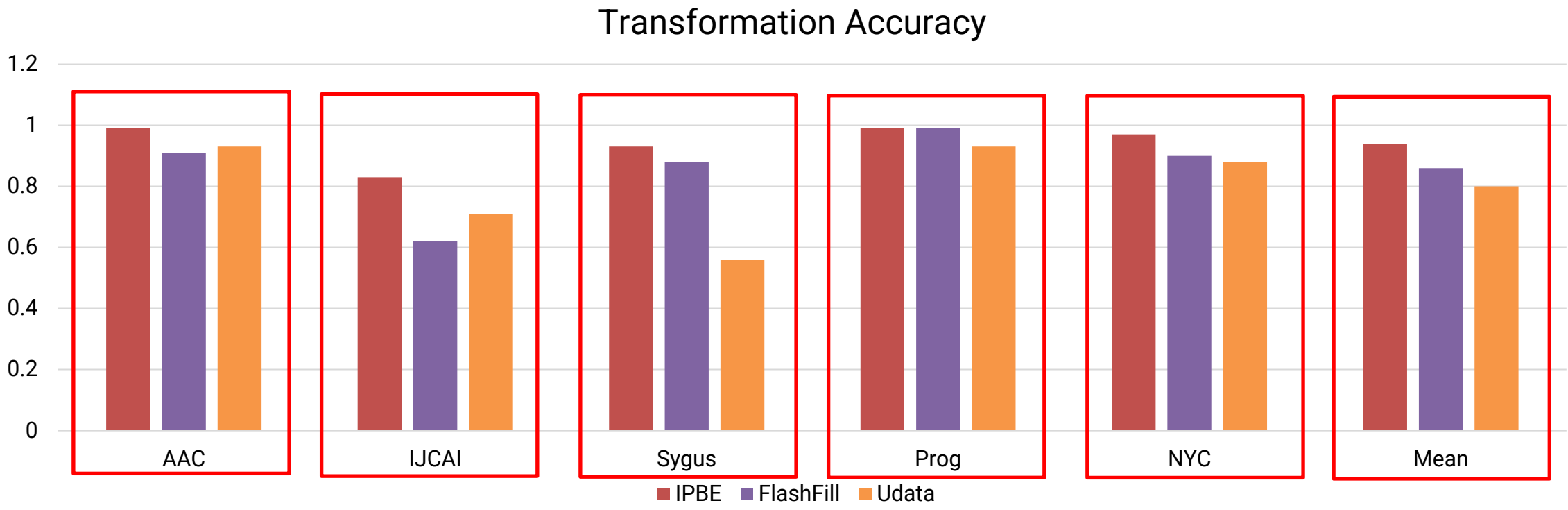


Transformation validation



Evaluation

- Our system: UDATA
- Two baseline systems:
  - IPBE (Wu et al, 2015)
  - FlashFill (Gulwani et al, 2012)



Transformation validation evaluation

Goal of validation: find all incorrect transformations in the systems = high recall

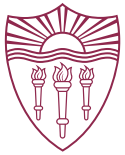
	Precision	Recall	F-measure
Validation Result	**0.63**	**0.99**	**0.73**

Validation Result		**Groundtruth**	
		Incorrect Transform	Correct Transform
Validation Prediction	Incorrect Transform	**20.0%**	**11.7%**
	Correct Transform	**0.2%**	**68.1%**

Summary

- Novel unsupervised approach for data transformation and error correction
  - construct syntactic patterns of string values
  - learn transformations programs by aligning semantically similar tokens

- Validation method with a near-perfect recall to capture the wrong transformation



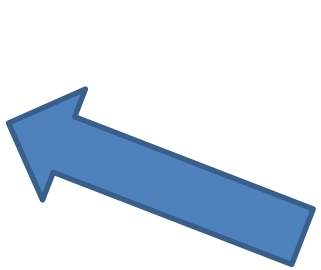
SEMANTIC ERROR DETECTION AND CORRECTION

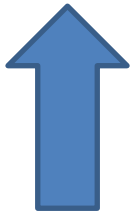
Motivating example: Wikipedia

Club	Location	Stadium
Al-Ahli	Jeddah	King Abdullah Sports City
Al-Faisaly	Harmah	King Salman Sport City Stadium
Al-Fateh	Al-Hasa	Prince Abdullah bin Jalawi Stadium
Al-Hilal	Riyadh	King Fahd International Stadium Prince Faisal bin Fahd Stadium
Al-Ittihad	Jeddah	King Abdullah Sports City
Al-Khaleej	Saihat	Prince Saud bin Jalawi Stadium
Al Nassr	Riyadh	King Fahd International Stadium Prince Faisal bin Fahd Stadium
Al-Qadisiyah	Khobar	Prince Saud bin Jalawi Stadium
Al-Raed	Buraidah	King Abdullah Sport City Stadium
Al-Shabab	Riyadh	King Fahd International Stadium Prince Faisal bin Fahd Stadium
Al-Taawoun	Buraidah	King Abdullah Sport City Stadium
Al-Wehda	Makkah	King Abdul Aziz Stadium
Hajer	Al-Hasa	Prince Abdullah bin Jalawi Stadium
Najran	Najran	Al Akhdoud Club Stadium

"The Prince Saud bin Jalawi Stadium is …. and it is the home stadium of Al-Qadisiya.

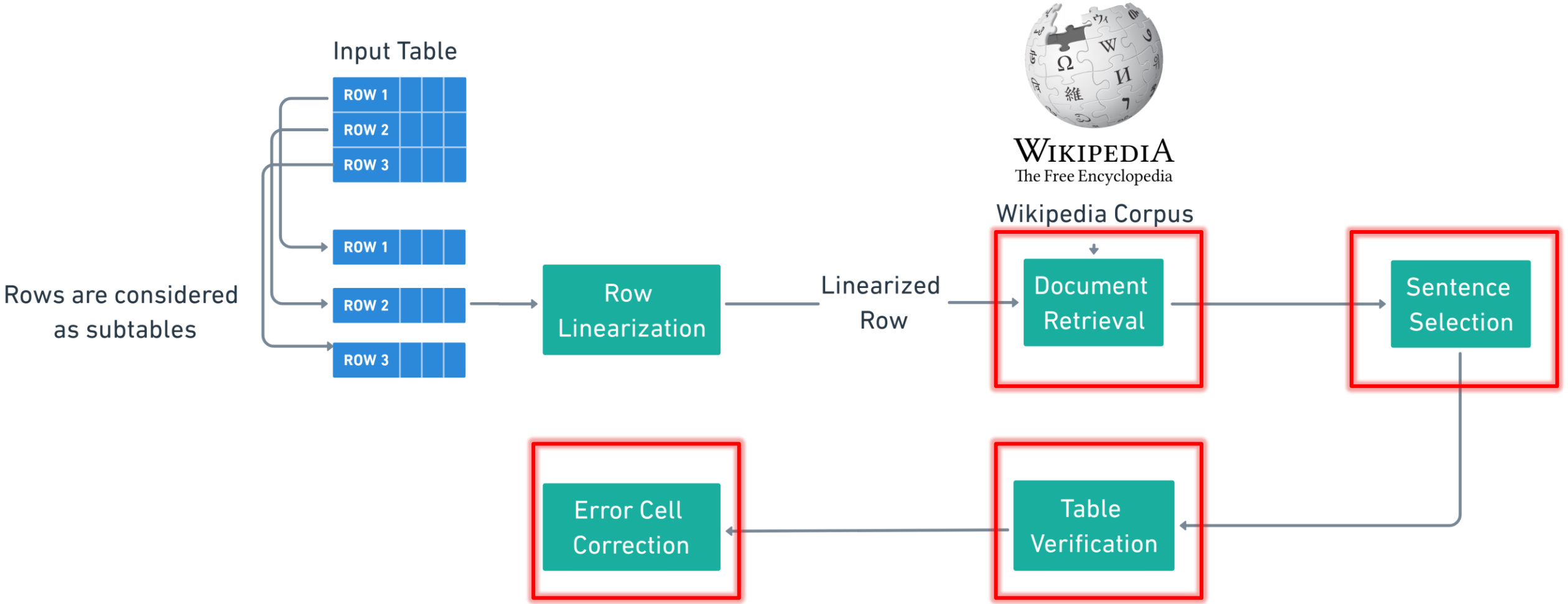
https://en.wikipedia.org/wiki/Prince_Saud_bin_Jalawi_Stadium





Locating this text in the corpus helps resolving the error

Overall approach



Examples

<table>
<tr><th>Phase</th><th>Output</th></tr>
<tr><td>Data input</td><td>
<table>
<tr><th>Opponent</th><th>Event</th><th>Date</th></tr>
<tr><td>Daniel Sarafian</td><td>UFC 174</td><td>June 14, 2014</td></tr>
</table>
</td></tr>
<tr><td>Document retrieval</td><td>“Kiichi Kunimoto’
“Daniel Sarafian”
“Neil Magny”</td></tr>
<tr><td>Sentence selection</td><td>“Kunimoto [...] UFC 174 on June 14, 2014, facing Daniel Sarafian.”
“Sarafian faced Kiichi Kunimoto [...] on June 14, 2014, at UFC 174.”</td></tr>
<tr><td>Table verification</td><td>“Kunimoto [...] UFC 174 on June 14, 2014, facing Daniel Sarafian.” ⇒ Entailed
“Sarafian faced Kiichi Kunimoto [...] on June 14, 2014, at UFC 174.” ⇒ Entailed</td></tr>
</table>

Example of a correct table

Examples

<table>
<tr><th>Phase</th><th>Output</th></tr>
<tr><td>Data input</td><td>
<table>
<tr><th>Years</th><th>Title</th><th>Role</th><th>Location</th></tr>
<tr><td>2010</td><td>Equus</td><td>Alan Strang</td><td>HERE Arts Center</td></tr>
</table>
</td></tr>
<tr><td>Document retrieval</td><td>“Sam Underwood’
“James Rado”</td></tr>
<tr><td>Sentence selection</td><td>“[...], Underwood was asked to play the part of Alan Strang in a production of "Equus" at the John Drew Theatre [...]”</td></tr>
<tr><td>Table verification</td><td>“[...], Underwood was asked to play the part of Alan Strang in a production of ‘Equus’ at the John Drew Theatre [...]” ⇒ Not entailed</td></tr>
<tr><td>Error cell correction</td><td>Location ⇒ John Drew Theatre</td></tr>
</table>

Example of an incorrect table

ToTTo dataset

- A dataset for controlled table-to-text generation

Tee Grizzley

Section Title: Awards and nominations
Table Section Text: *None*

Year	Award	Category	Work	Result
2017	2017 BET Hip Hop Awards	Best New Hip Hop Artist	Himself	Nominated
	2017 BET Hip Hop Awards	Best Mixtape	My Moment	Nominated
2018	MTV Video Music Awards	Push Artist of the Year	Himself	Nominated

Original sentence(s)

He received two 2017 BET Hip Hop Award nominations for Best New Hip-Hop Artist and Best Mixtape for My Moment.

Final sentence(s)

Tee Grizzley received two 2017 BET Hip Hop Award nominations for Best New Hip-Hop Artist, and for Best Mixtape for My Moment.

Overall approach

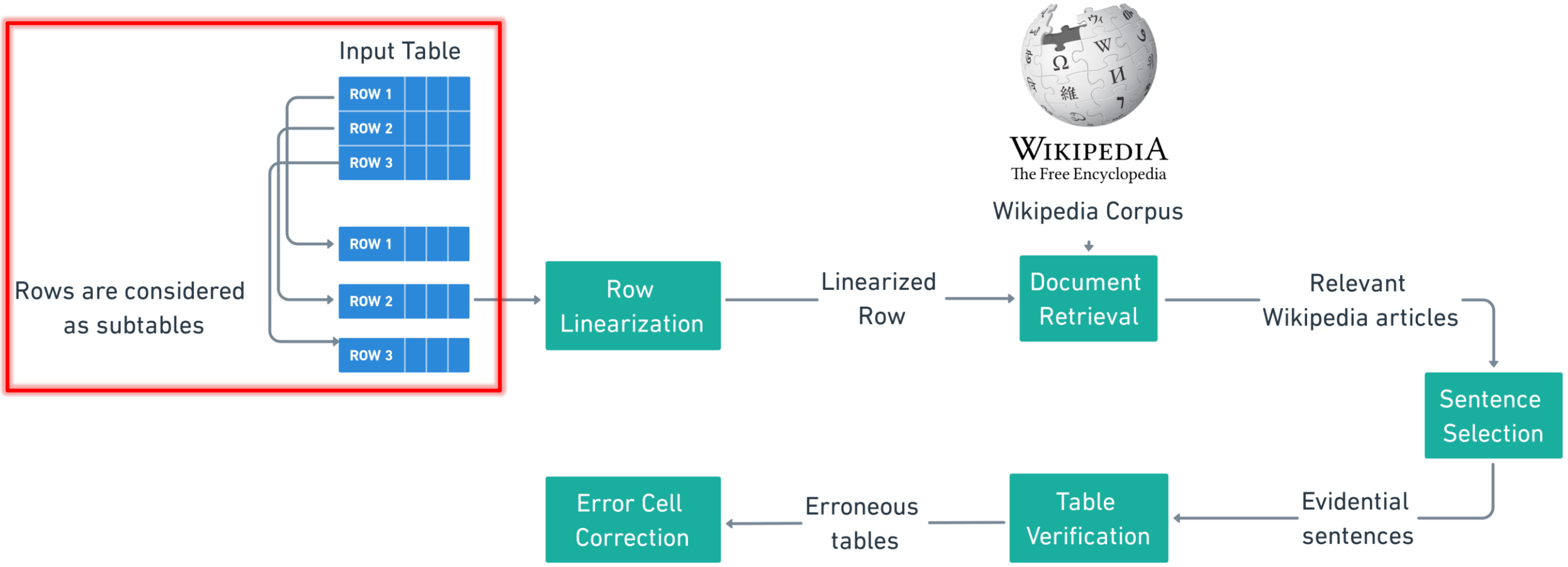


Table linearization

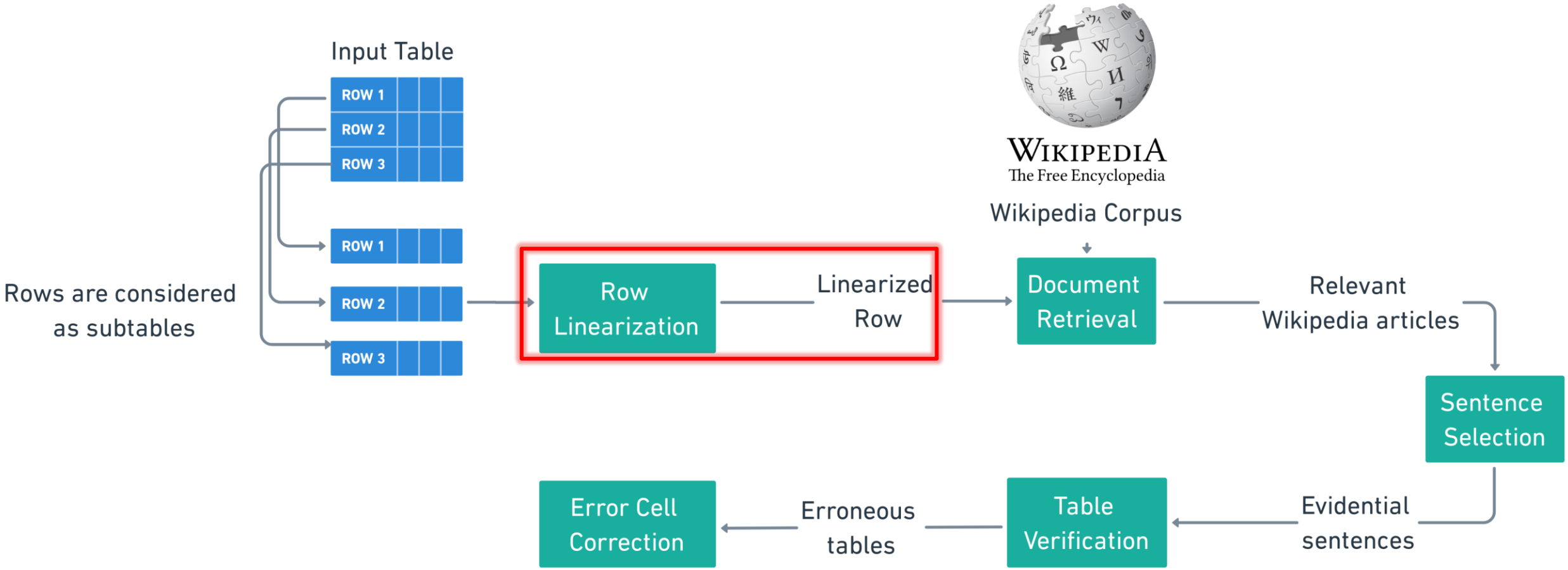


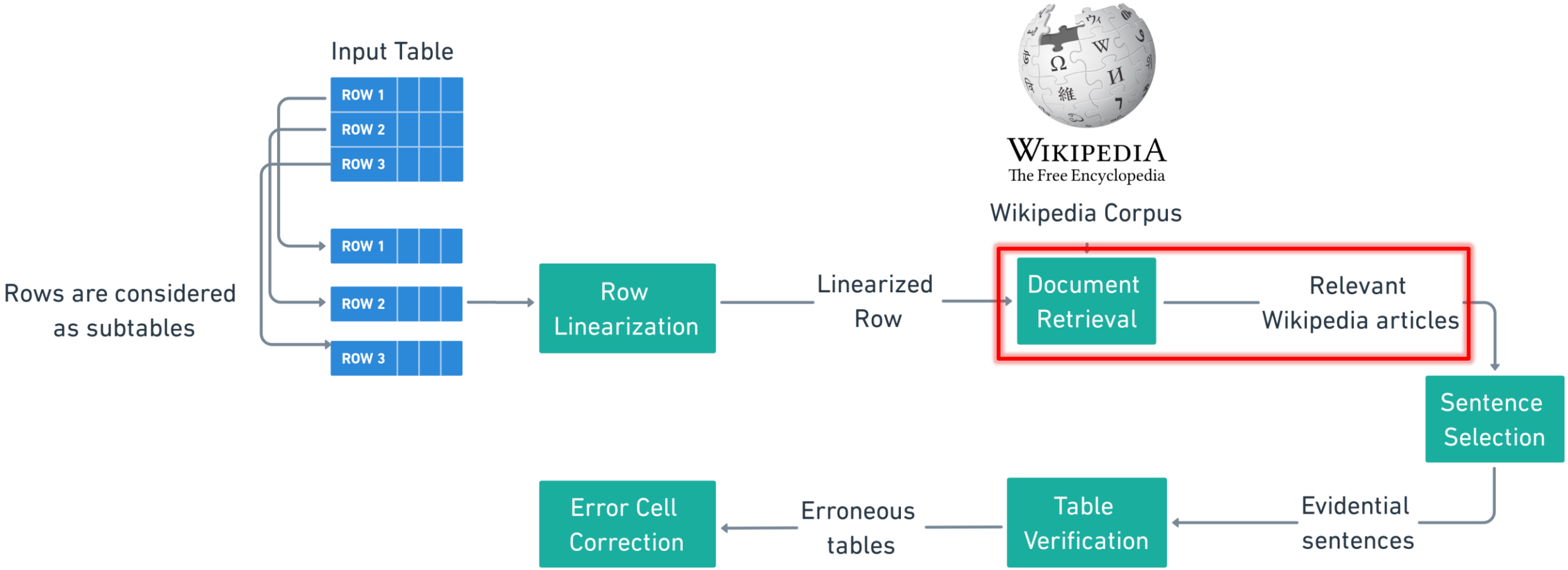
Table linearization

Page Title: Kiichi Kunimoto

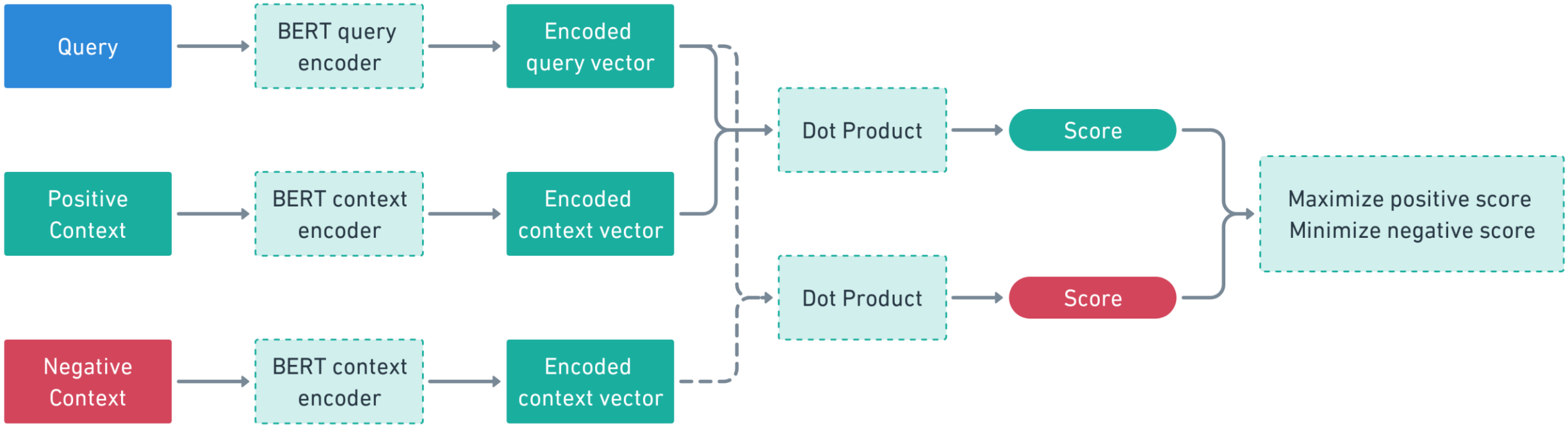
Opponent	Event	Date
Daniel Sarafian	UFC 1974	June 14, 2014

```
<table><page_title> Kiichi Kunimoto</page_title>
    <row>
        <cell> <col_header> Opponent </col_header> Daniel Sarafian </cell>
        <cell> <col_header> Event </col_header> UFC 1974</cell>
        <cell> <col_header> Date </col_header> June 14, 2014 </cell>
    </row>
</table>
```

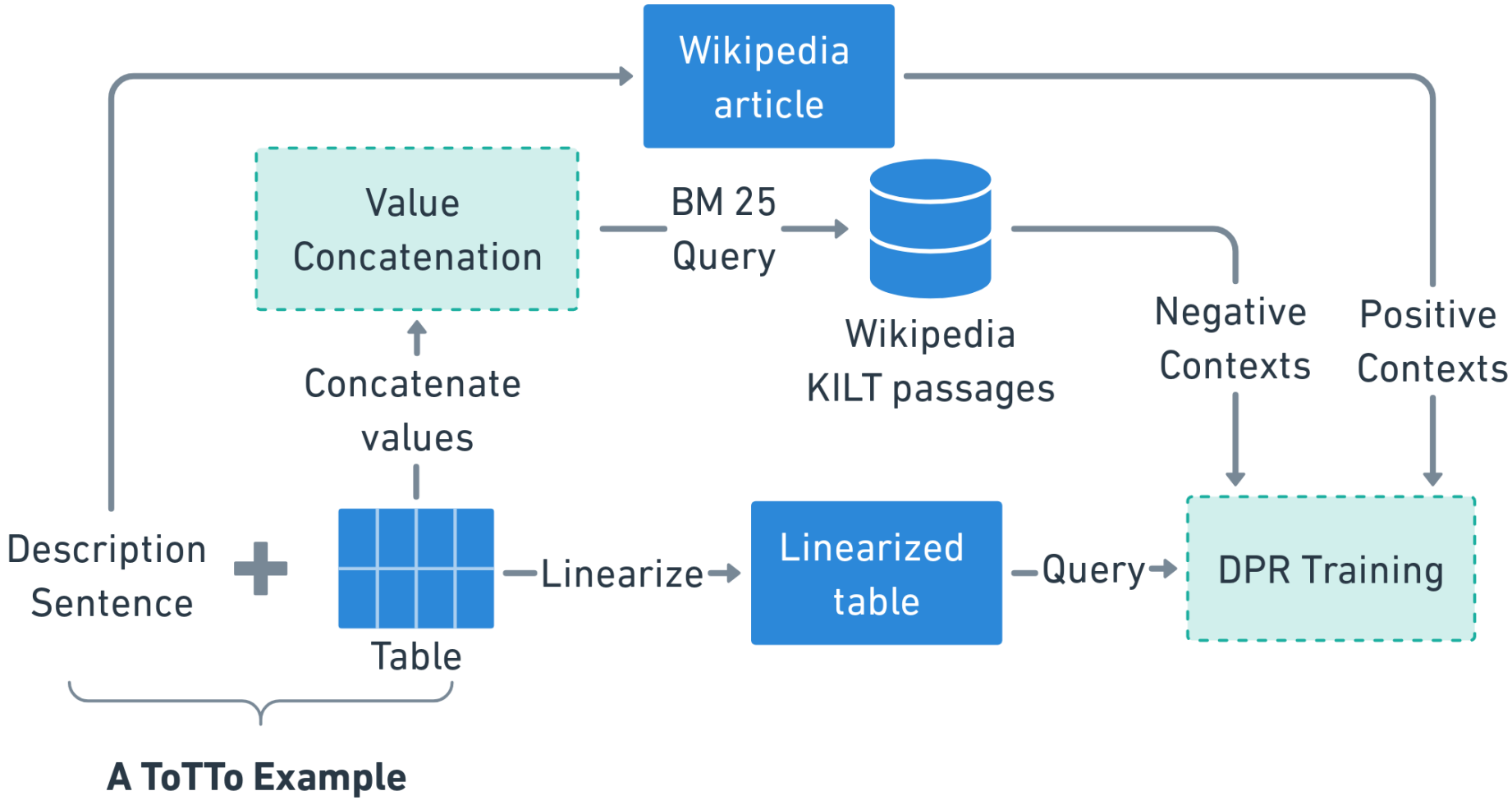
Document retrieval



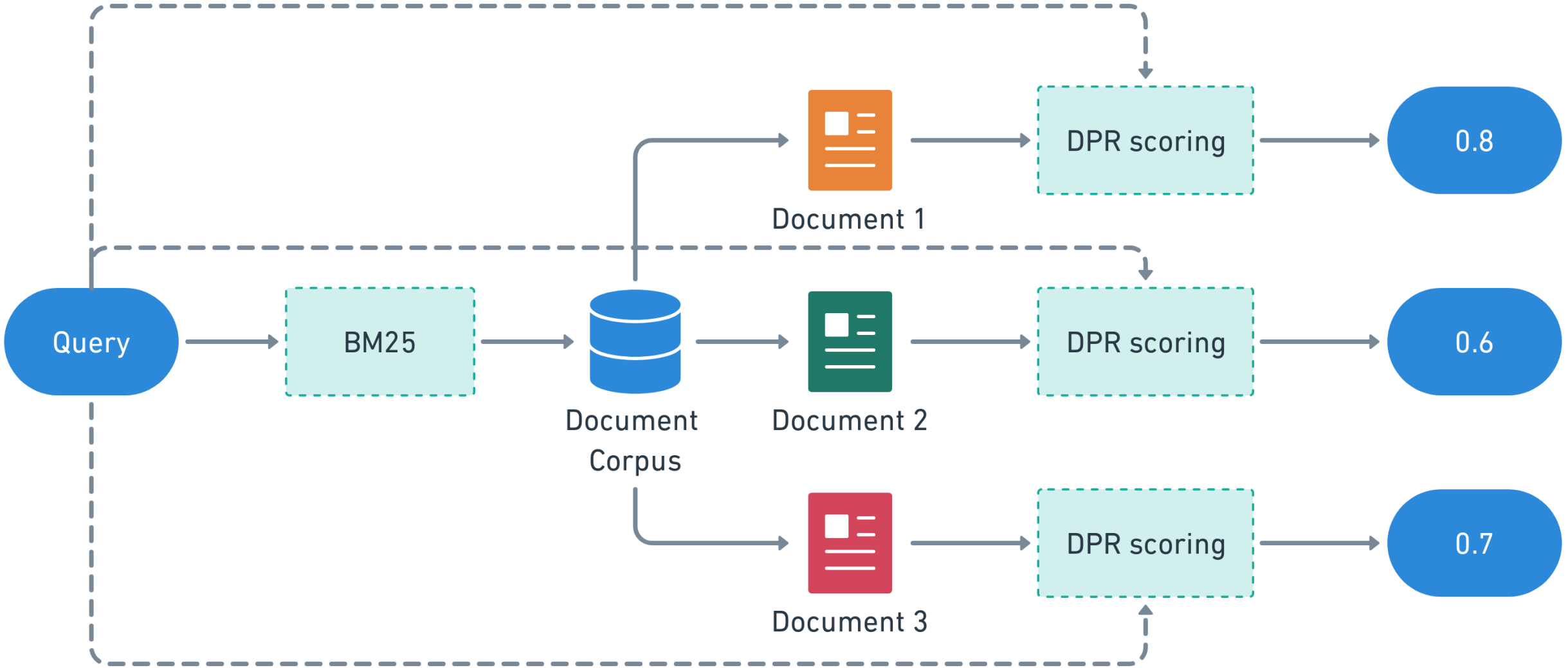
Dense Passage Retrieval (DPR)



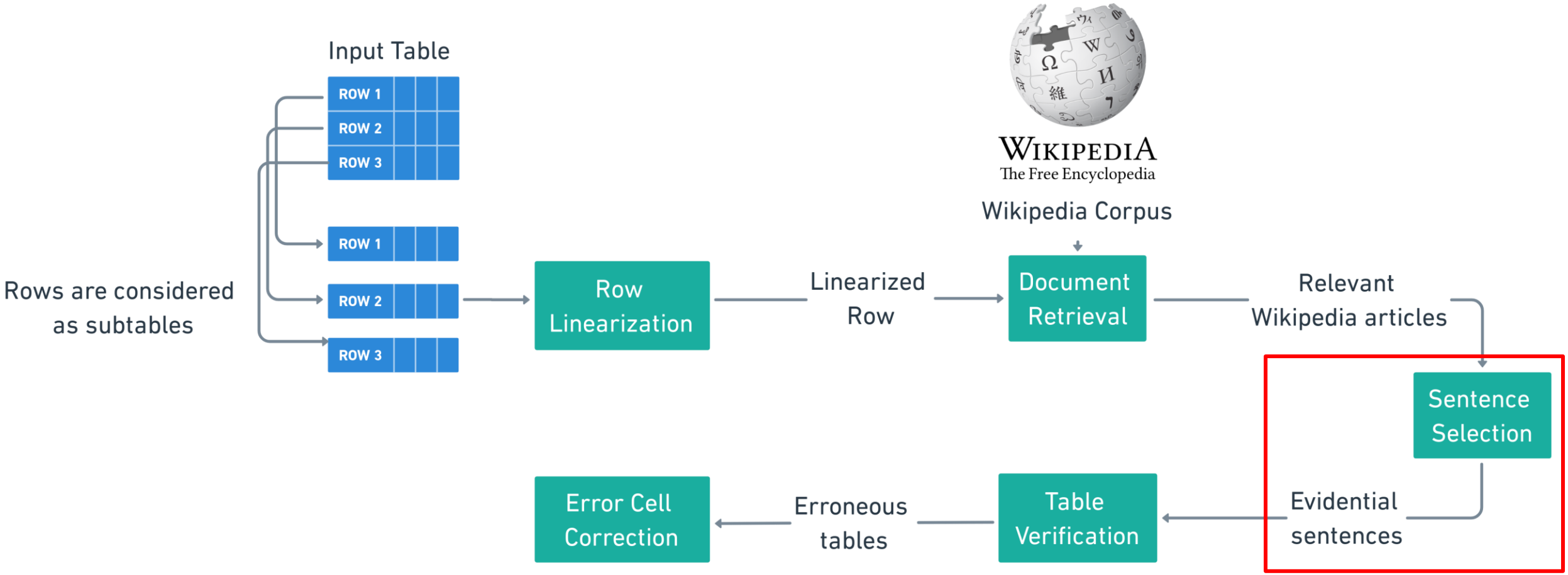
DPR training



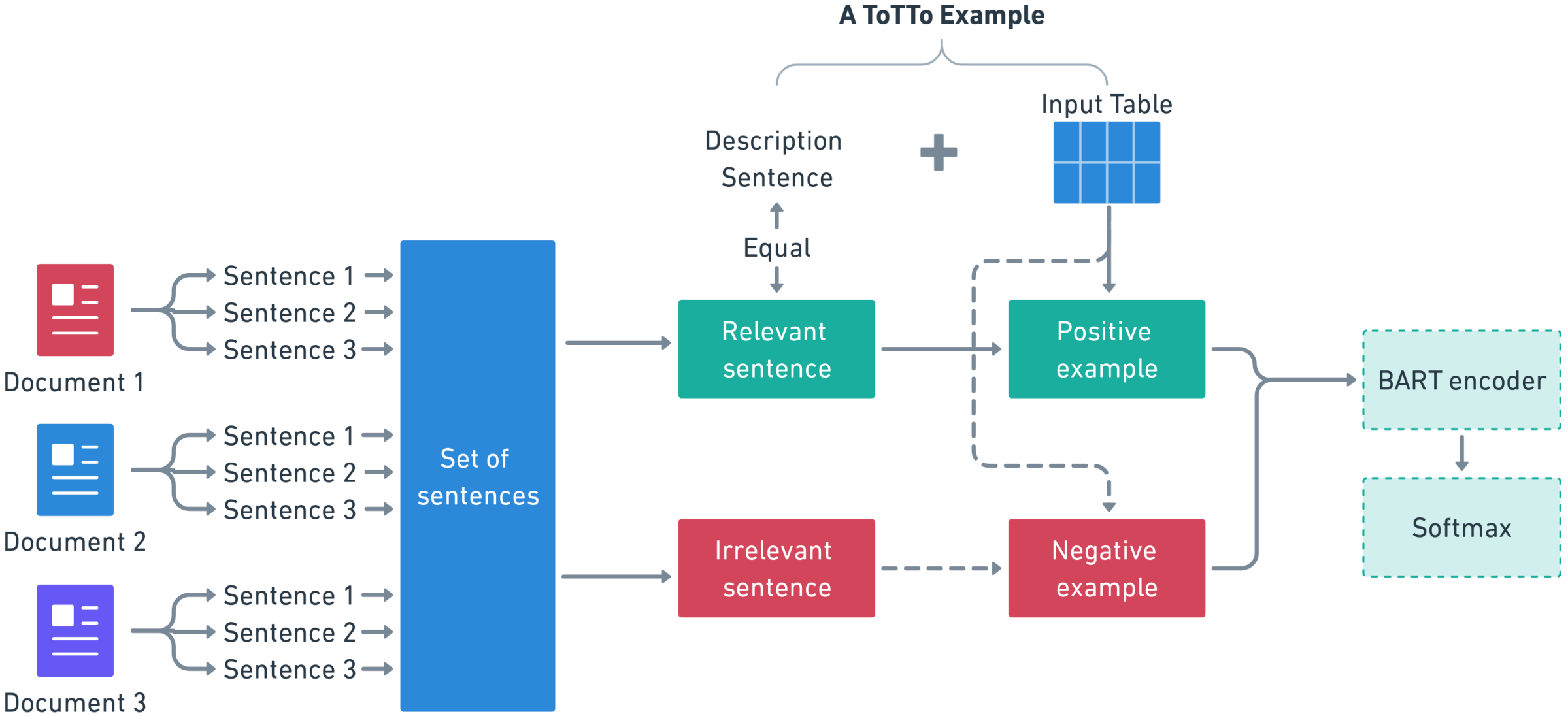
Document retrieval



Sentence selection



Sentence selection - Training



Sentence selection - Inference

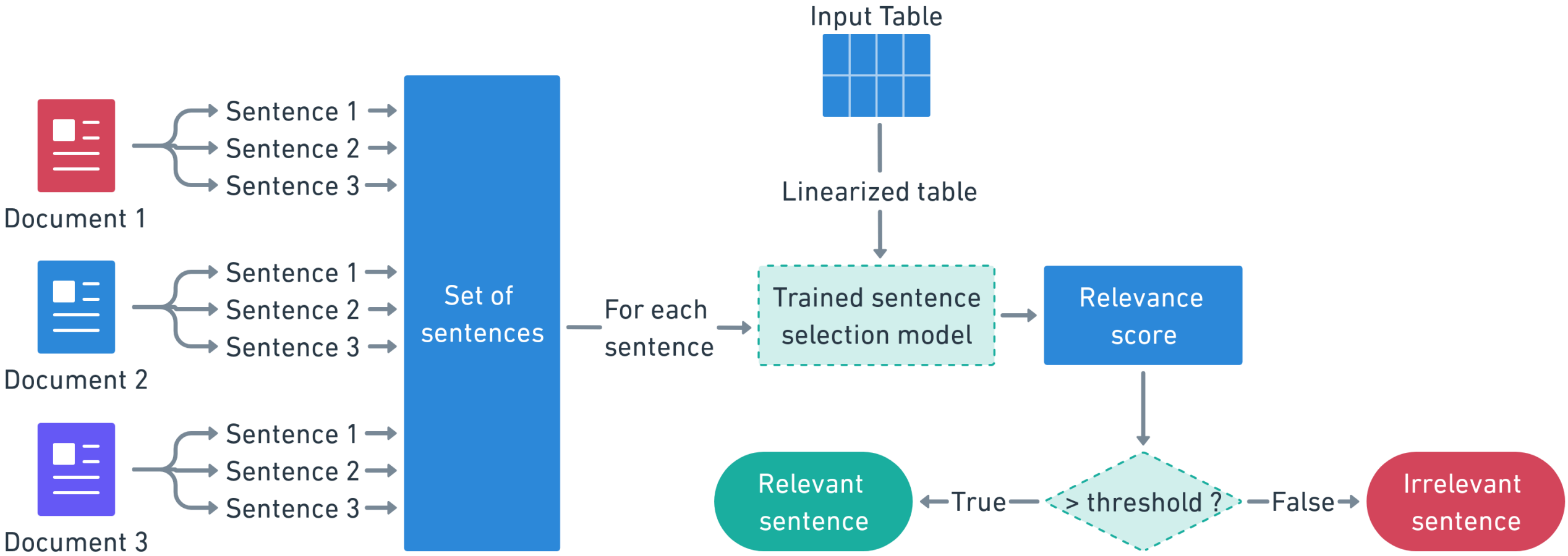


Table verification

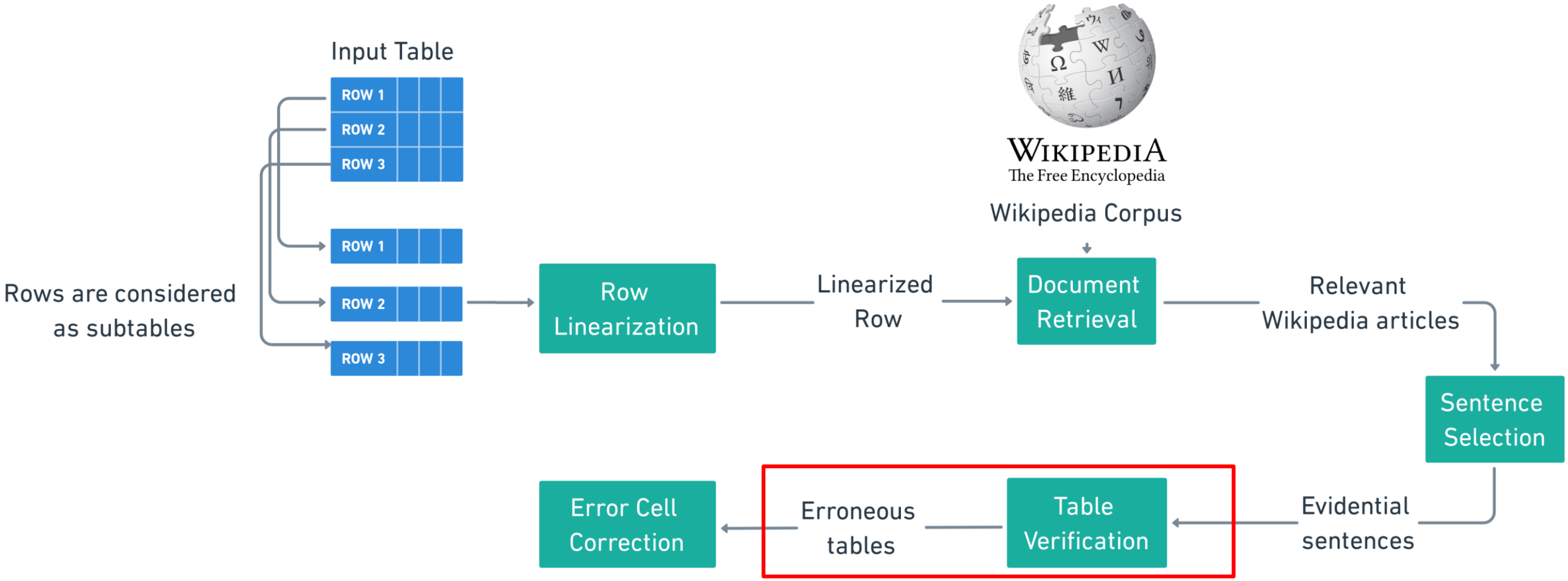
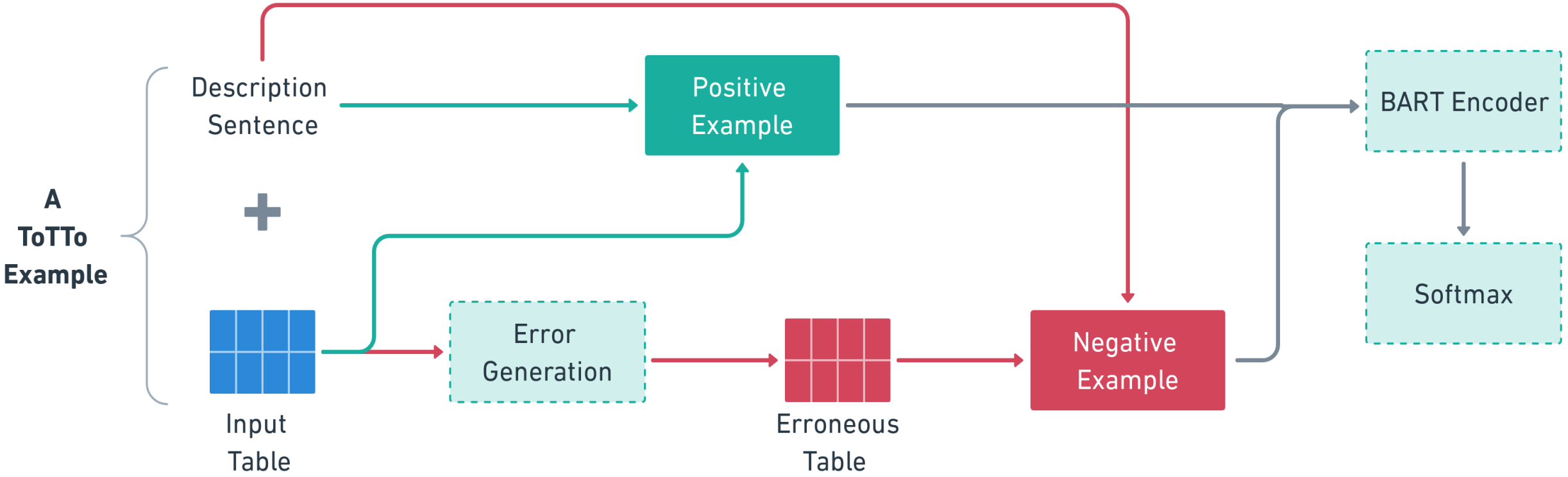


Table verification - Training



Error generation

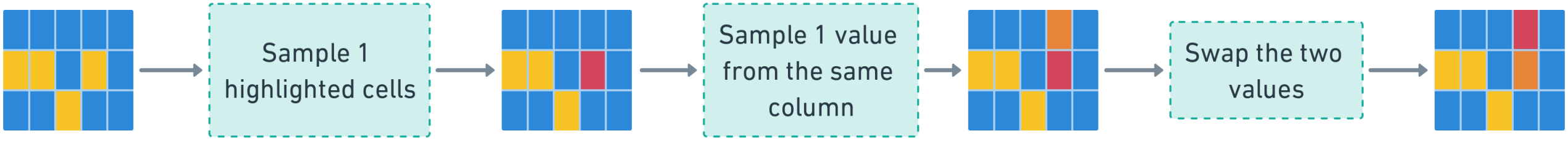
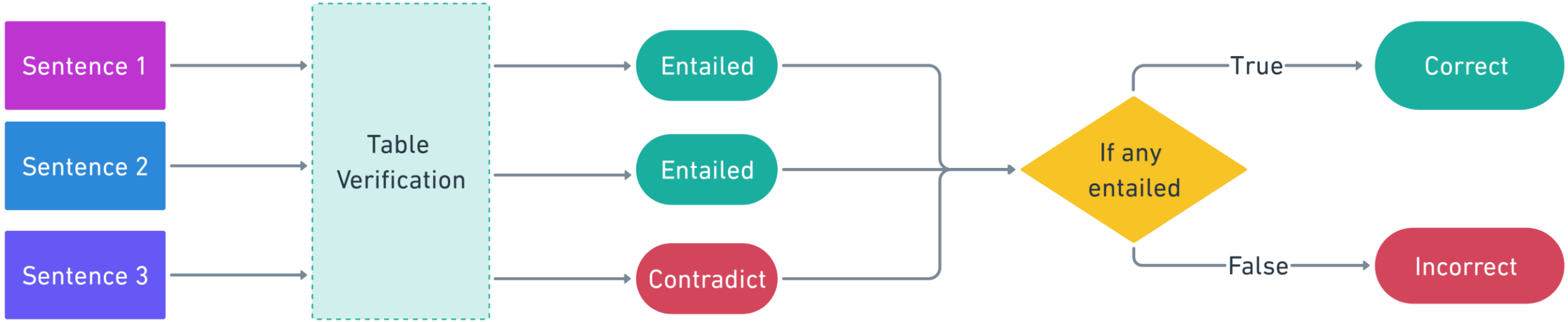
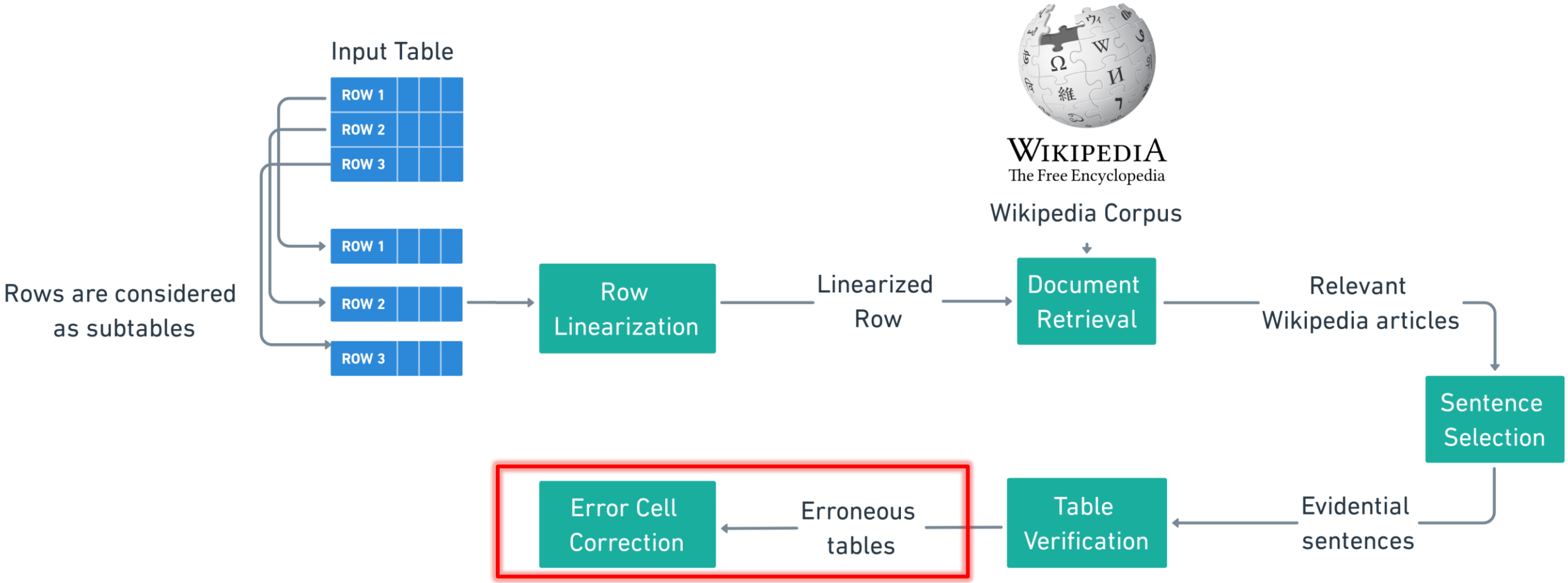


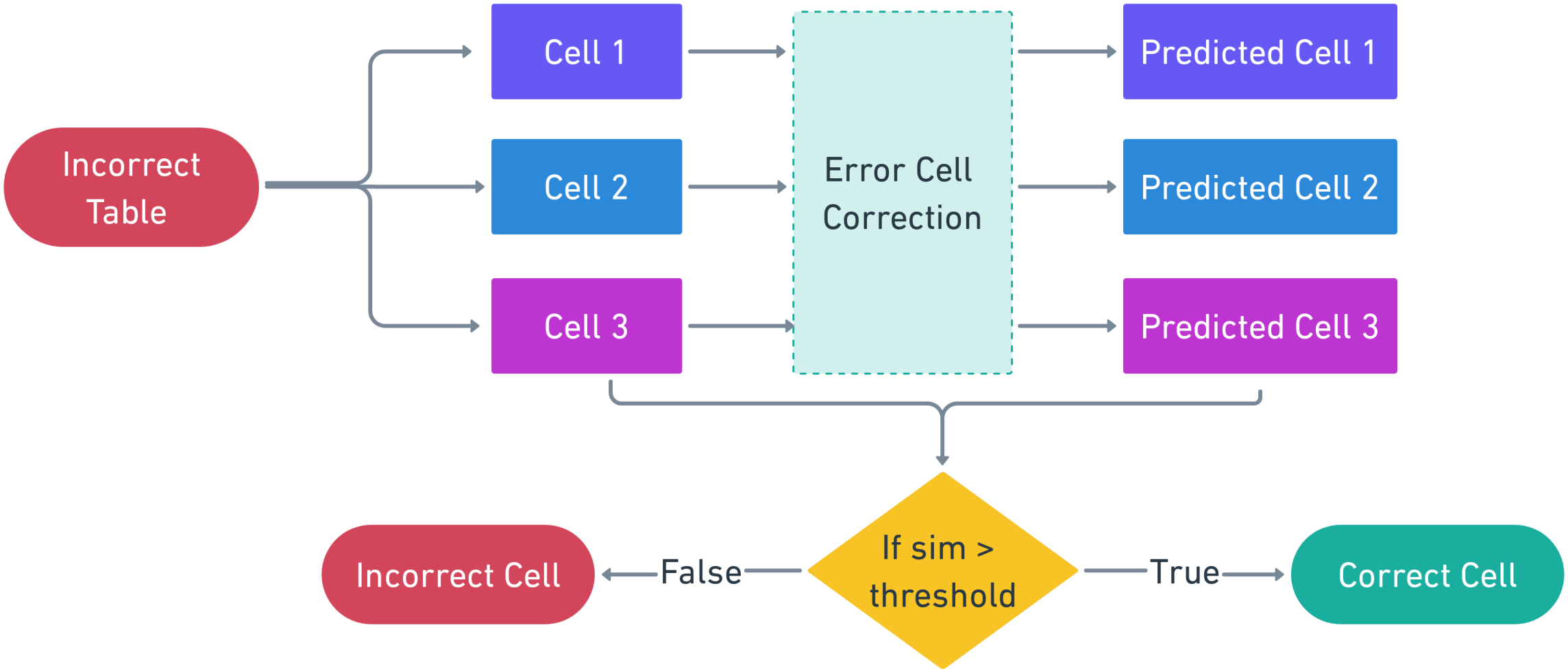
Table verification - Inference



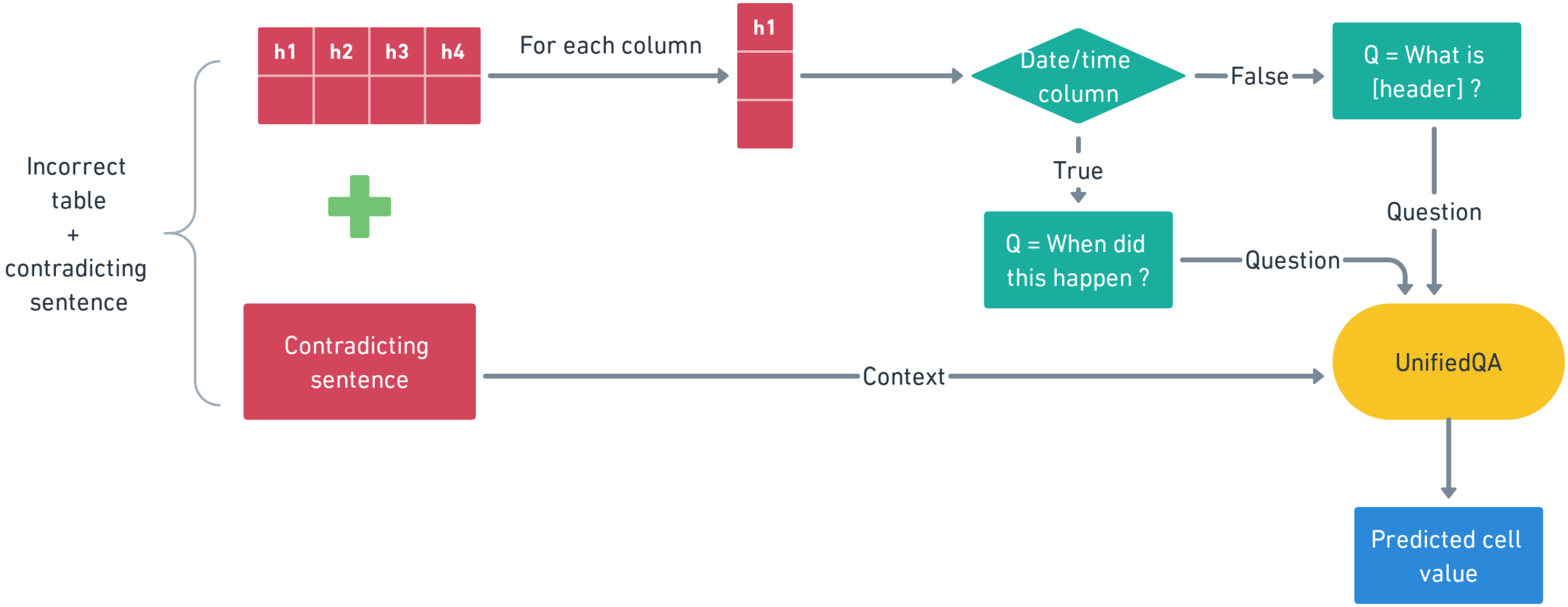
Error cell correction



Error cell correction



Error cell correction



Evaluation

- Document retrieval

	Recall@1	Recall@5	Recall@10
BM25	0.36	0.65	0.80
DPR	0.18	0.56	0.78
DPR reranking (SEED)	**0.54**	**0.93**	**0.95**

- Sentence selection:
  - Infotab + Tapas: Table-based Fact Verification systems

	Precision	Recall	F1
InfoTab	0.844	0.931	0.886
TAPAS	0.973	0.951	0.962
SEED	**0.976**	**0.985**	**0.981**

Evaluation

- Table verification

	Precision	Recall	F1
InfoTab	N/A	N/A	N/A
TAPAS	0.926	0.916	0.921
SEED	**0.985**	**0.957**	**0.961**

- Error cell correction

Recall	MRR
0.65	0.46

Evaluation

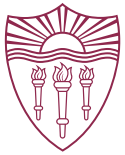
- End-to-end evaluation
  - 2 table-based fact verification baselines: InfoTab and TAPAS
  - InfoTab does not converge during table verification training
  - InfoTab and TAPAS use SEED's document retrieval result

Method	Document Retrieval			Sentence selection			Table Verification		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
InfoTab	0.25	0.88	0.39	0.05	0.56	0.09	N/A	N/A	N/A
TAPAS	0.25	0.88	0.39	0.23	0.27	0.25	0.51	0.43	0.47
SEED	0.25	0.88	0.39	**0.67**	**0.82**	**0.74**	**0.88**	**0.83**	**0.86**

Summary

- Novel approach to detect semantic errors in tables
  - hybrid approach (BM25 + DPR) for document retrieval
  - present multiple strategies to augment data for model training

- Question-answering-based approach for semantic error correction in tables



RELATED WORK

Syntactic error detection

- Supervised error detection
  - HoloDetect [Heidari et al. ‘17], Auto-Detect [Huang et al.’18], Uni-Detect [Wang et al.’19]
- Interactive error detection
  - NADEEF [Dallachiesa et al. ‘13], KATARA [Chu et al. ’15]; dBoost [Mariet et al. ’16]
- Semi-supervised error detection
  - ED2 [Neutatz et al. ‘19], Raha [Mahdavi et al. ’19]

Our approach uses both active learning and data augmentation techniques to minimize the amount of user labeling while maintaining sufficient training data to train a high accuracy model.

Syntactic error correction

- Supervised data transformation
  - *Programming-by-example*: FlashFill [Gulwani et al. ‘12], IPBE [Wu et al. ‘15, Wu et al. ‘16], Blinkfill [Singh et al. ’16], Unifacta [Jin et al. ’18]
  - *Machine learning*: [Wang et al. ’14], [Shu et al. ’17], [Devlin et al. ’18]
- Interactive cleaning
  - Potter [Rama et al. ‘01], Wrangler [Kandel et al. ‘11]
- Complex transformation:
  - Extracted from web data: DataXFormer [Abedjan et al. ’16]

Our approach requires no aligned input/output from user labeling and thus minimizes the amount of user interaction.

Semantic error detection and correction

- Document Retrieval
  - DPR [Karpukhin et al. ‘20], RAG [Lewis et al. ‘20], DensePhrases [Lee et al. ‘21], KGI [Glass et al. ‘21], Re2G [Glass et al. ‘22]

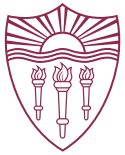
Our document retrieval approach combines DPR and keyword search to improve recall of numeric and named entity data, which are popular in tables.

- Table-based Fact Verification
  - [Eisenschlos et al. ‘20], [Wang et al. ‘21]’, Infotab [Gupta ‘20]

Our approach addresses the opposite problem (use text to verify table)

- Table Pretraining
  - TAPAS [Herzig et al. ‘20], TAPEX [Liu et al. ‘21]

Our approach leverages pretrained language models → better for NLI tasks



DISCUSSION

Conclusion

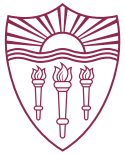
- Novel approaches to syntactic and semantic error detection and correction

- Leverage open-domain knowledge and closed-domain weak supervision to reduce human interaction
  - Open-domain knowledge: mining web text for semantic error detection and correction
  - Closed-domain weak supervision:
    - Syntactic error detection: signals and active learning
    - Syntactic error correction: user verification

Future Work

- Approach:
  - Optimize models to decrease training time and inference time
  - Support complex syntactic error correction/detection by mining web context (Web tables, web forms, public code repositories)
  - Support active learning for semantic error correction/detection
  - Joint training and joint inference for all approaches

- System:
  - Interactive system for unified syntactic and semantic data cleaning



THANK YOU