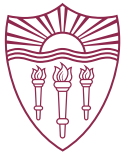


LEARNING DATA TRANSFORMATIONS WITH MINIMAL USER EFFORT

Minh Pham, Craig A. Knoblock, and Jay Pujara

Information Science Institute

University of Southern California

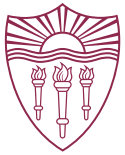
Outline

Problem

Approach

Evaluation

Conclusion and Future Work



PROBLEM

Example: New York City Data

Name	Address	Phone	Website	Latitude	Longitude
Sosa Borella	832 Eighth Ave	2122628282	http://www.sosaborella.com	40.762444	-73.985983
Starbucks	871-879 Eighth Ave	2122467699	http://www.starbucks.com	40.763644	73.985134

Data Integration

Name	Phone	Website	Location
Paramount Hotel	(212) 764-5500	http://www.nycparamount.com/	(40.759132, -73.986348)
Doubletree Guest Suites	2127191600	www.nycdoubletreehotels.com	(40.759055, -73.98471
The Westin New York at Times Square	(212) 868-1900 ext 245	www.westinny.com	(40.757482, -73.988309)

Example: People names

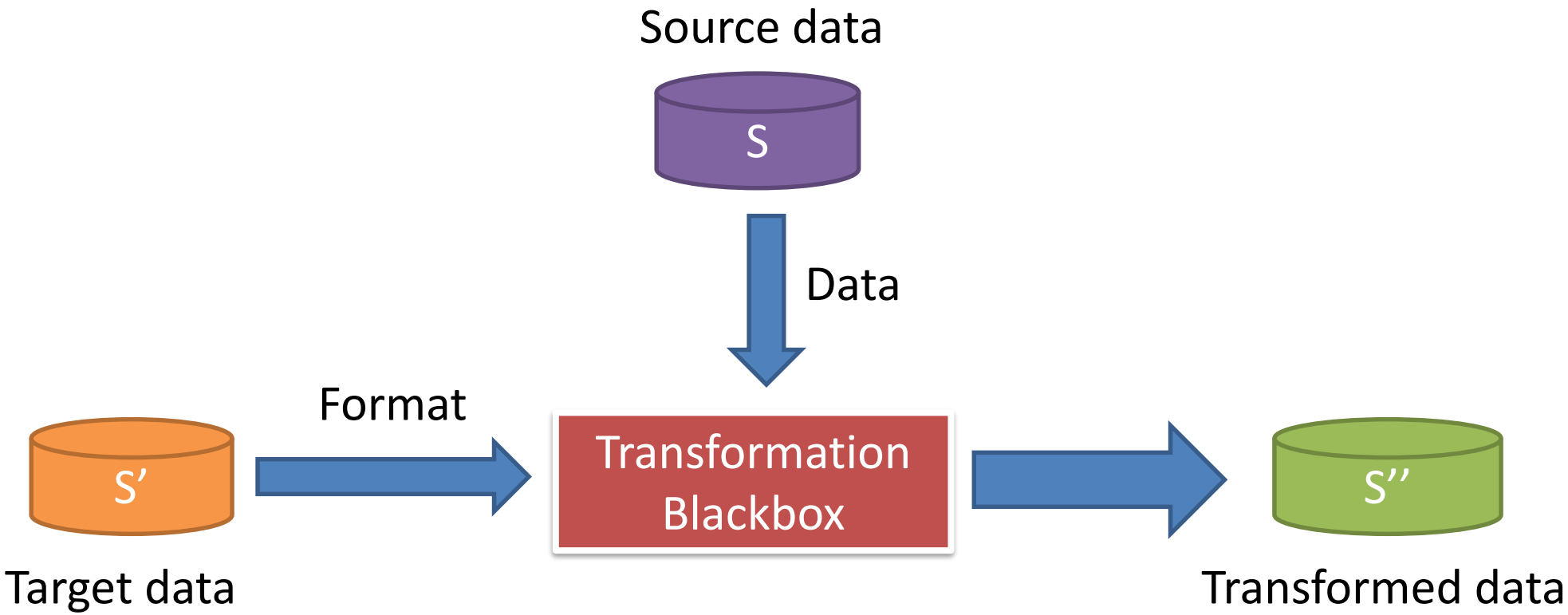
Name
Mark Slipper
Tom A. Clerverley
Cahill, Michael
Edward David
Sergio R. Garcia
Pogba, Paul
…
…

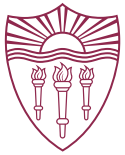
Normalize

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

General Problem

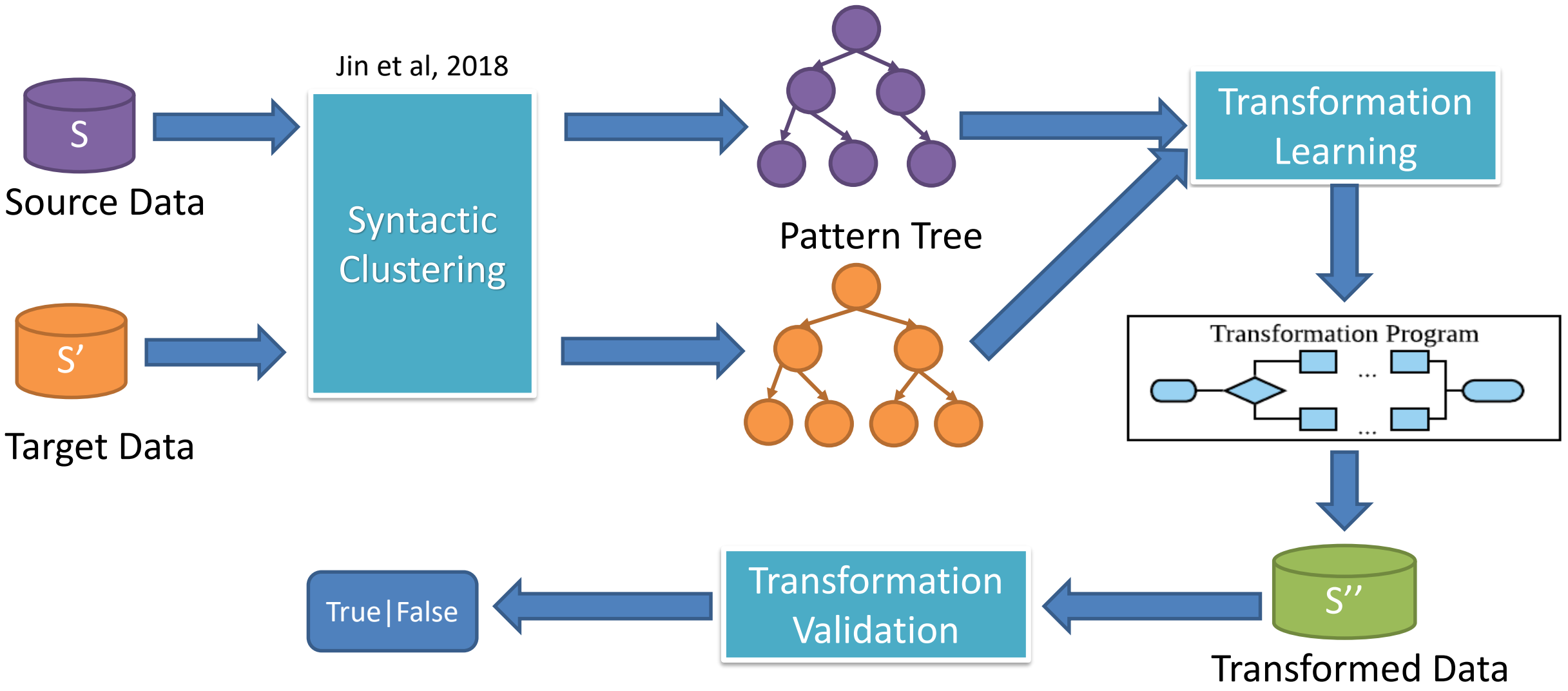
Given two data sources S and S',
learn the transformation program P to transform S to the format of S'





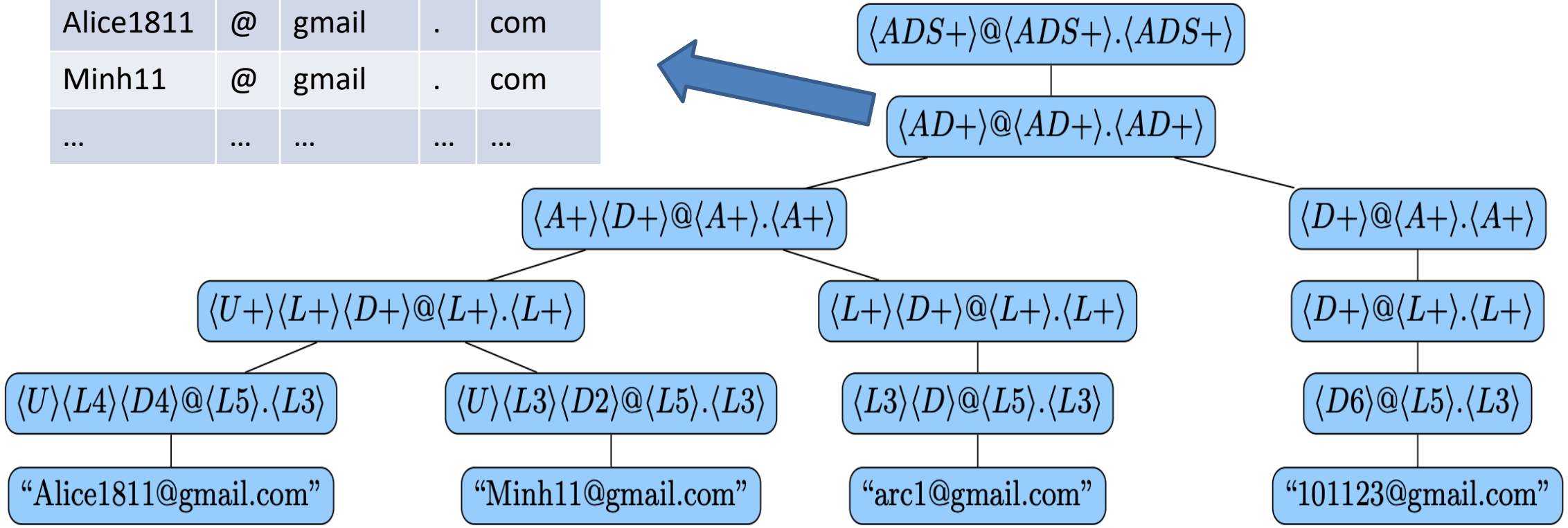
APPROACH

Overall Approach

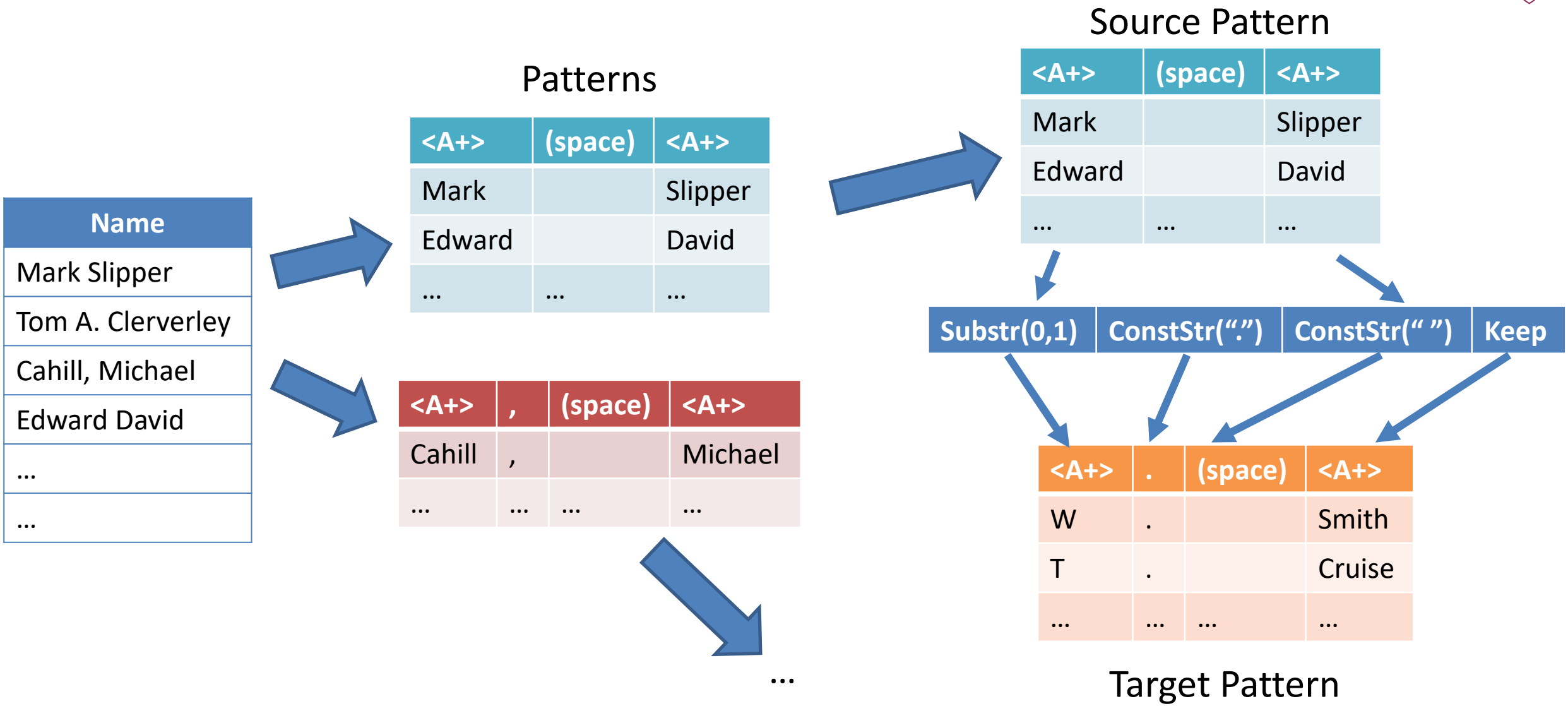


Pattern and Pattern Tree

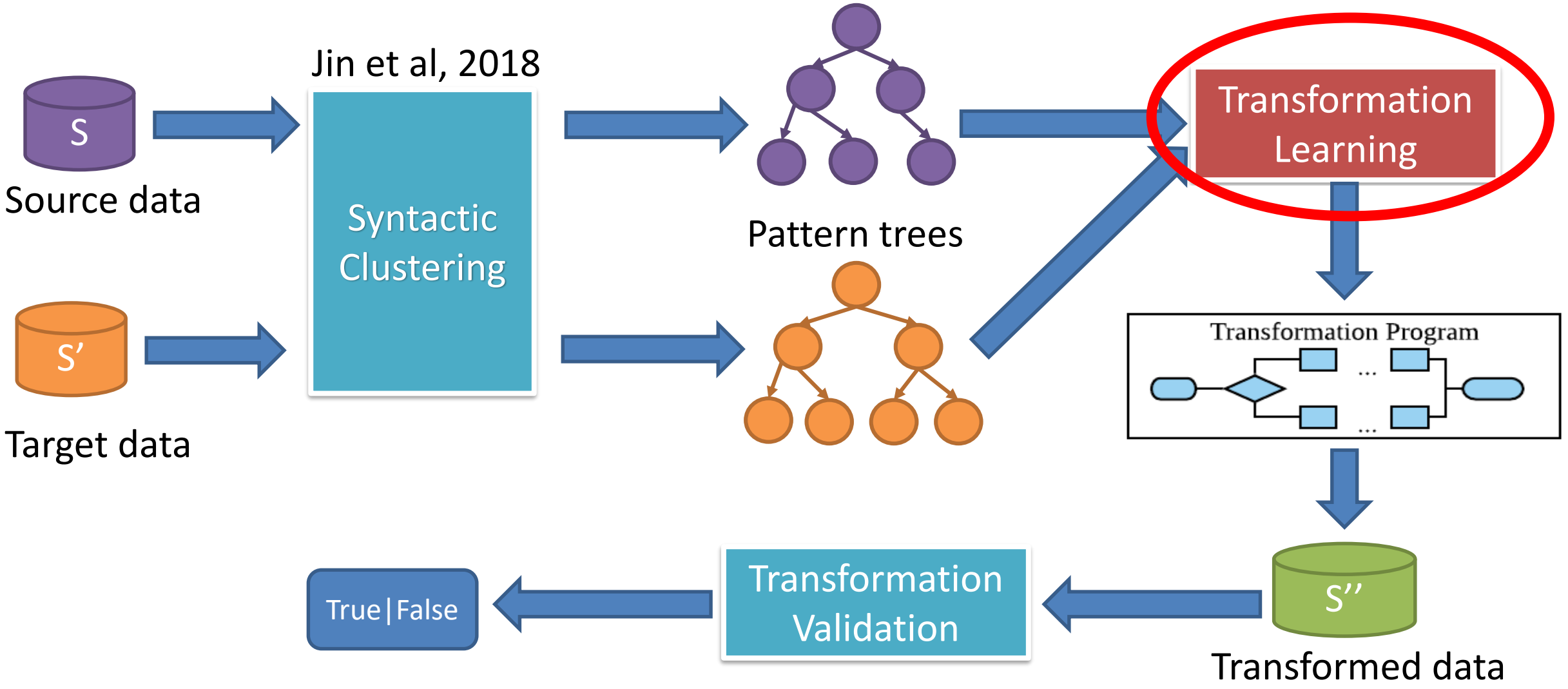
<AD+>	@	<AD+>	.	<AD+>
Alice1811	@	gmail	.	com
Minh11	@	gmail	.	com
…	…	…	…	…



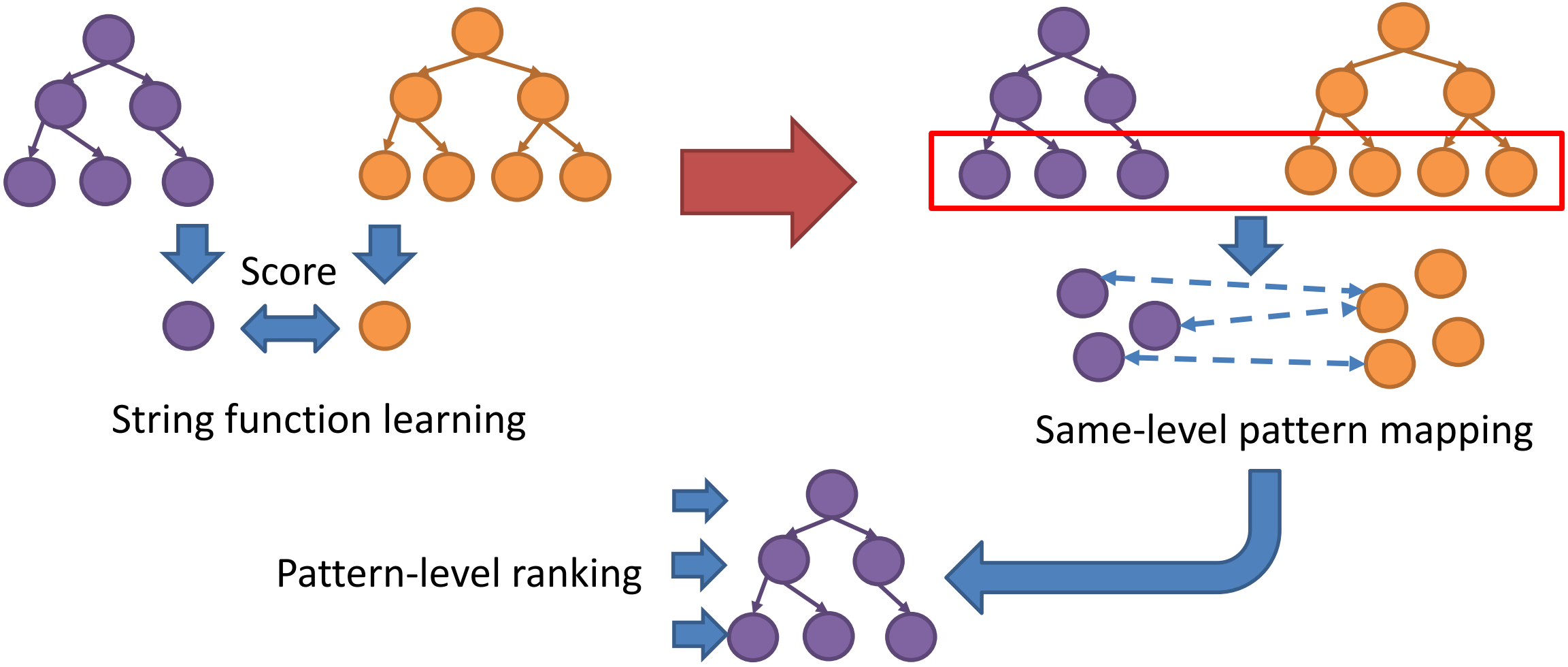
Transformation Program



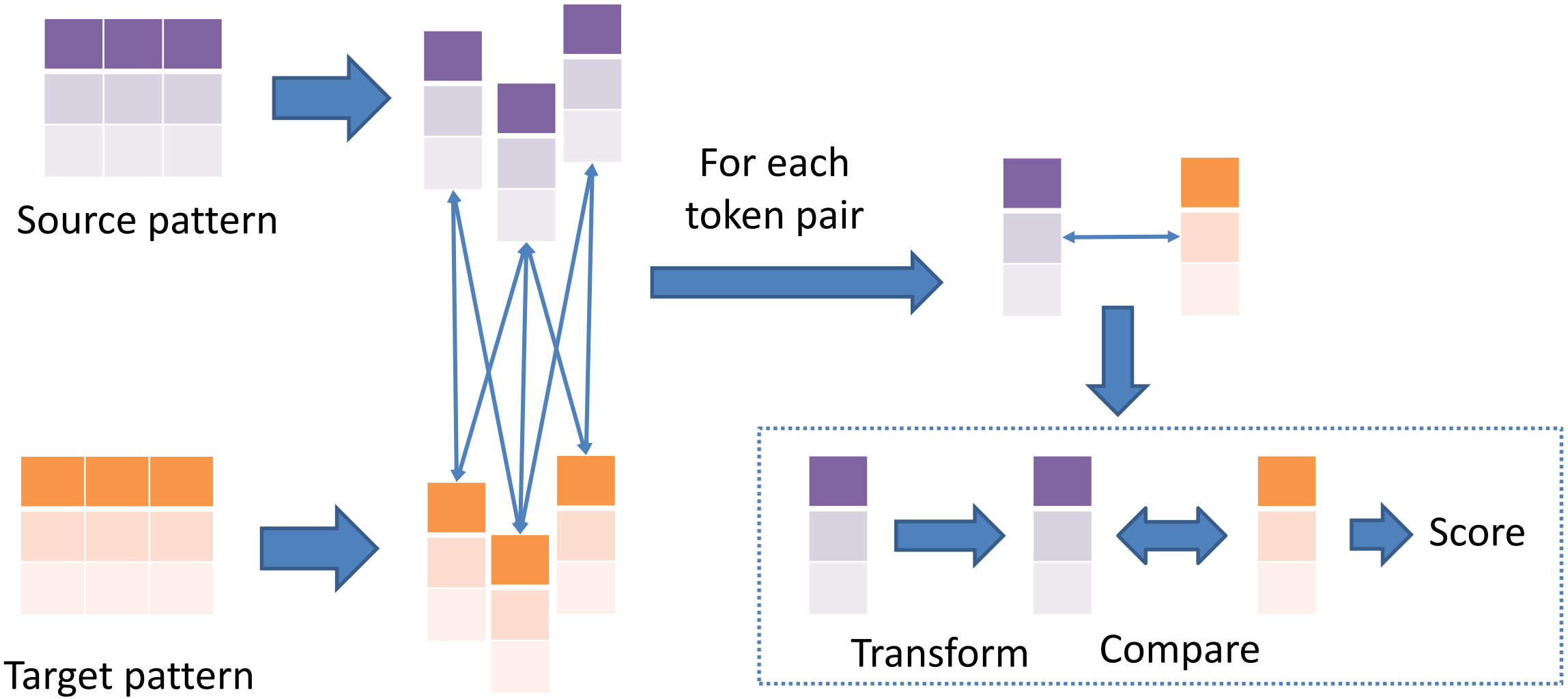
Transformation Learning



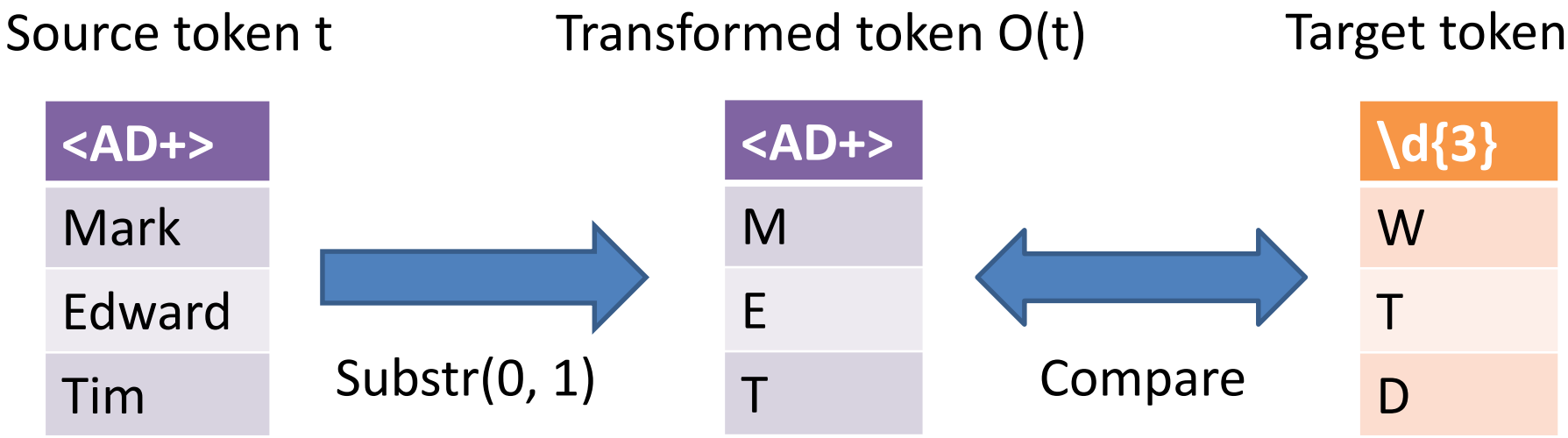
Transformation Learning – Bottom Up



String Function Learning



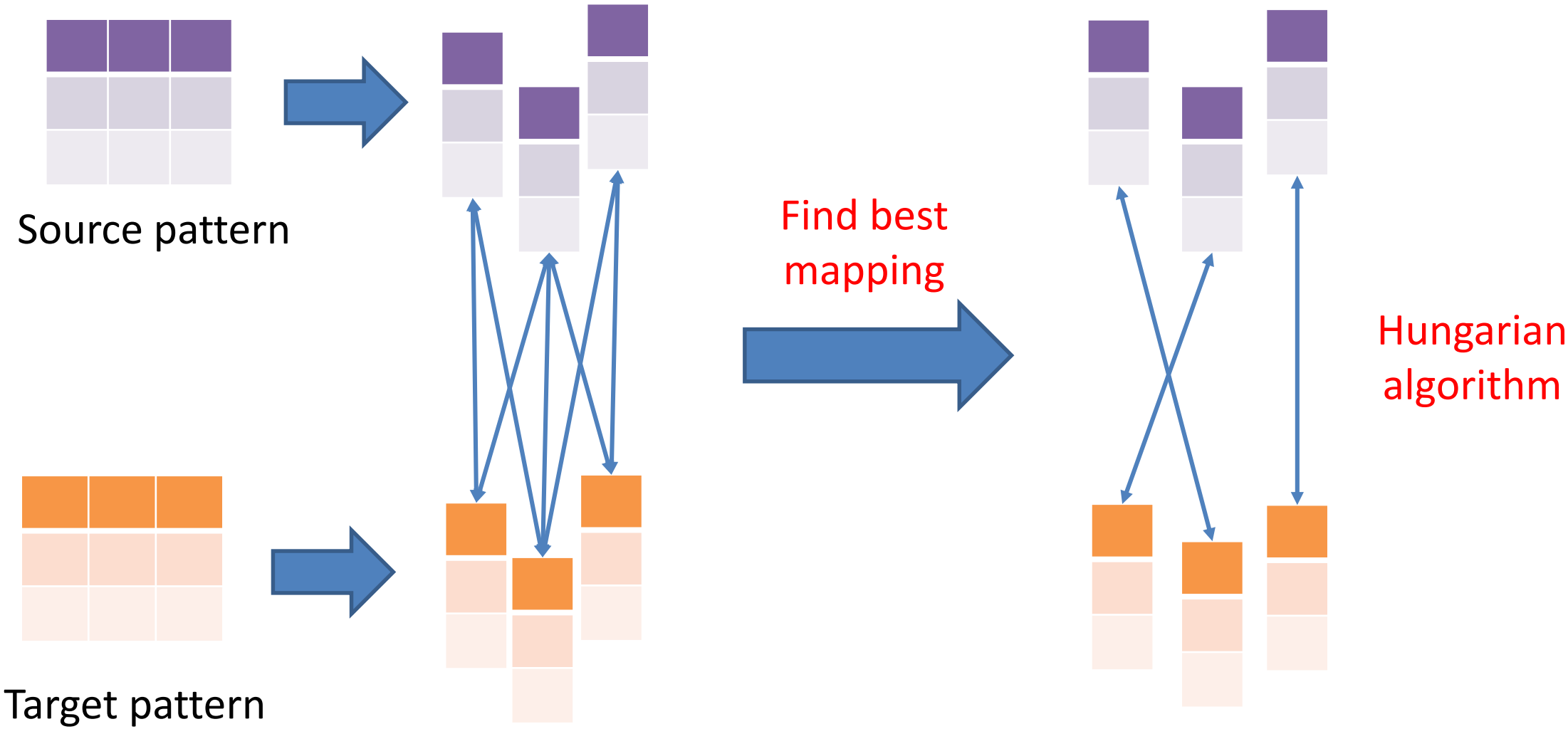
Scoring Model



$$Score(O, t, t') = sim(O(t), t)$$

$$Score(t, t') = \max_{O} Score(O, t, t')$$

String Function Learning



Same-level Pattern Mapping

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

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

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

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

Source patterns

Target patterns

Find the best mapping => Learn the correct transformation

Pattern-level ranking

Source patterns

<U+>	<L+>	(space)	<U+>	<L+>
M	ark		S	lipper
E	daward		D	avid
…		…	…	

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

<ADS+>
Mark Slipper
Edward Daid
…

What is the best level to learn the transformation ?

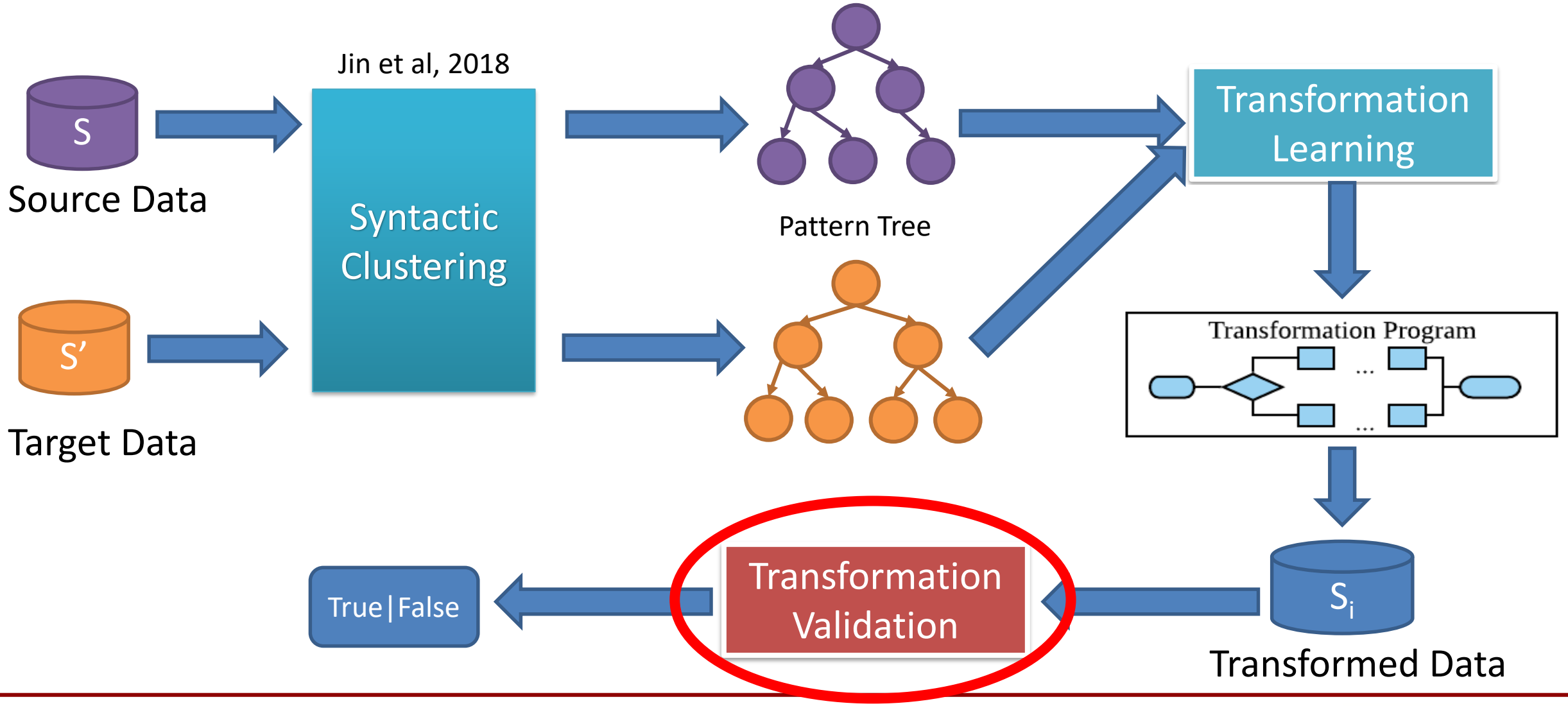
<U+>	.	(space)	<U+>	<L+>
W	.		S	mith
T	.		C	ruise
…		…	…	

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

<ADS+>
W. Smith
T. Cruise
…

Target patterns

Transformation Learning



Transformation Validation

Transformed data
Desiree S.
Alaina P.

Transformed data
R. Mcgaughey
C. Latimore

Target data
Andrew C.
Bradford L.

Have the same pattern?

Source Pattern

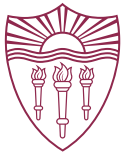
<AD+>		<AD+>
Desiree		Seamons
Chong		Aylward

Which one is first names

<AD+>
Samuel
Maryann

Target Pattern

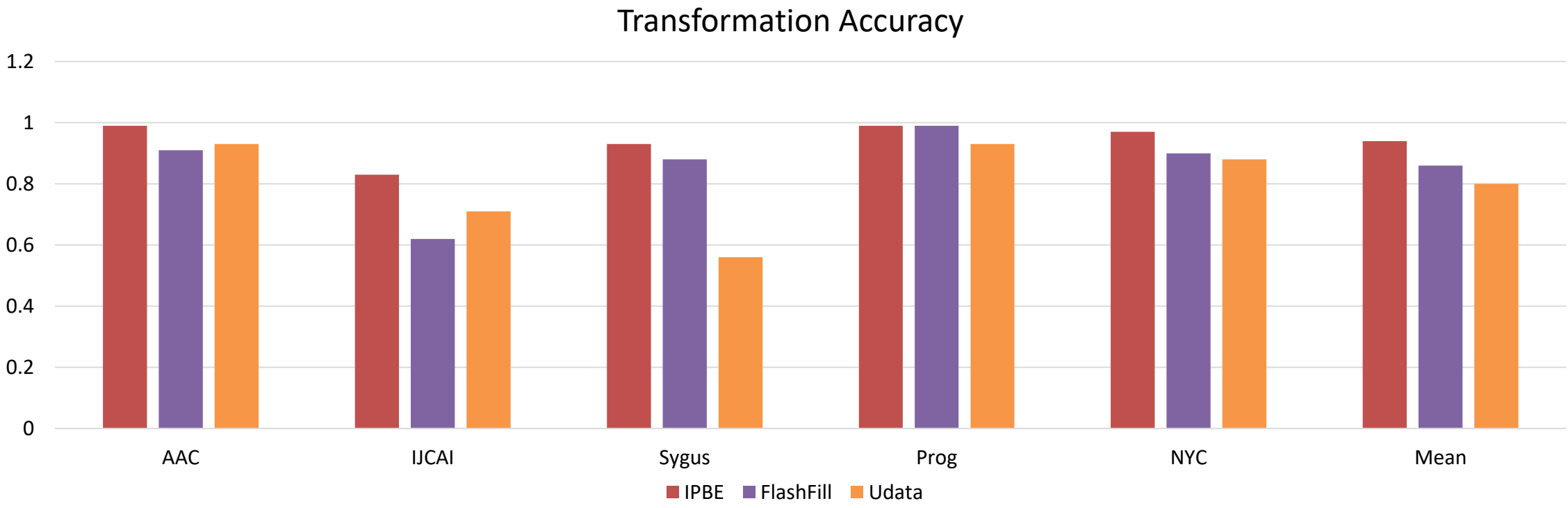
Ambigous in matching tokens



EVALUATION

Evaluation

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



Validation Evaluation

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

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

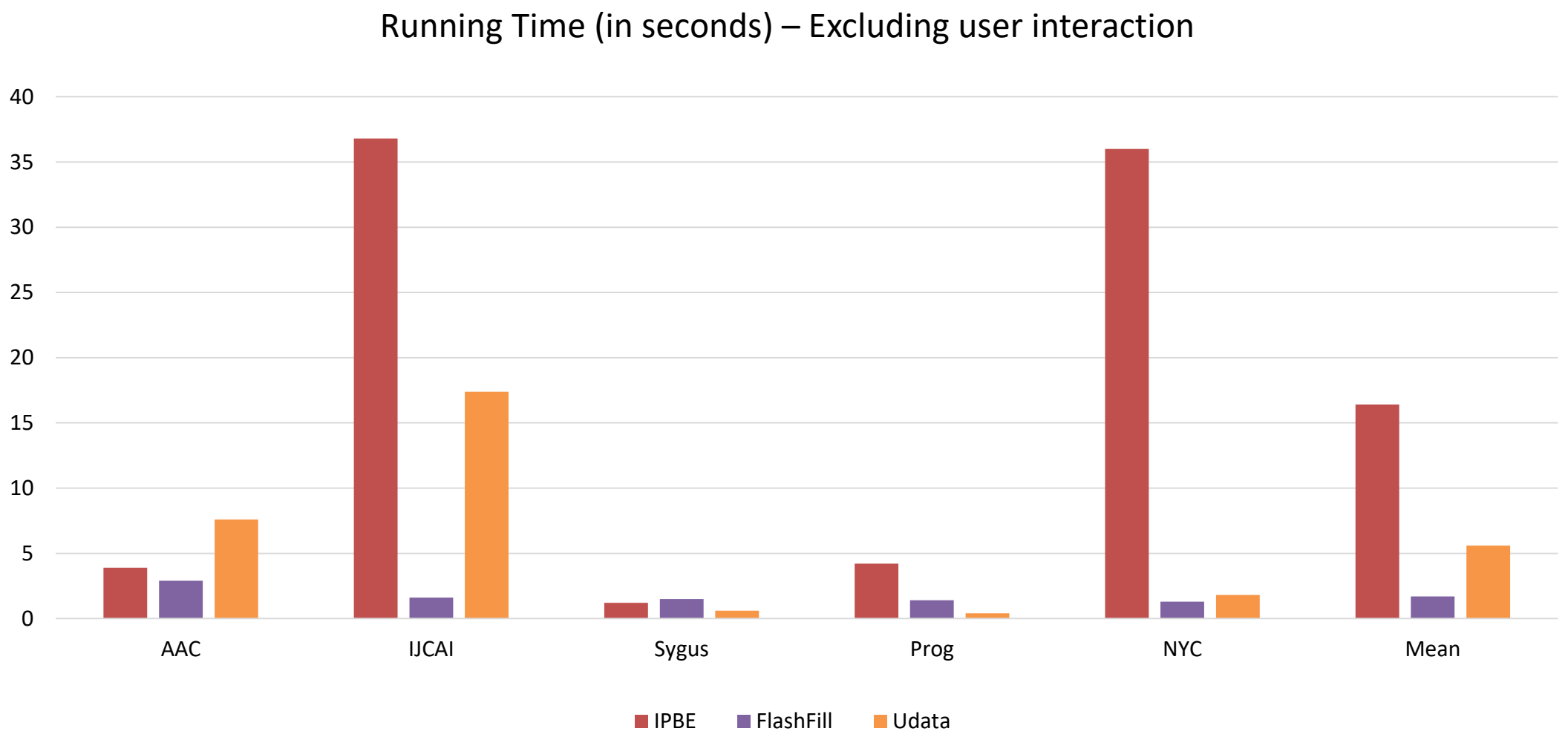
Validation Result		**Groundtruth**	
		Incorrect Transform	Correct Transform
Validation Prediction	Incorrect Transform	**20.0%**	**11.7%**
	Correct Transform	**0.2%**	**68.1%**

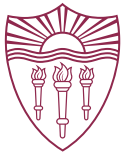
31% scenarios require human-interaction

vs

100% scenarios require human-interaction without UData

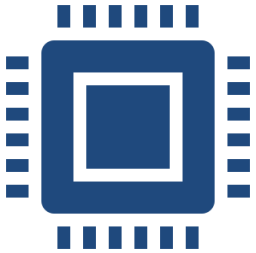
Running Time





CONCLUSION

Conclusion and Future Work



Conclusion

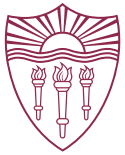
An unsupervised data transformation system with high accuracy

A validation module which can detect “almost” any error made by the system

Future Work

Improve scalability and reduce running time

Include semantic transformation



THANK YOU