SPADE♠

A SEMI-SUPERVISED PROBABILISTIC APPROACH FOR DETECTING ERRORS IN TABLES

Minh Pham, Craig A. Knoblock, Muhao Chen, Binh Vu, Jay Pujara

Information Sciences Institute

University of Southern California

Tables are rich sources of structured knowledge

- Millions of tables on the Web
- Providing data for many applications

Beers

	index	beer_name	style	ounces	abv
	1	Pub Beer	American Pale Lager	12.0 oz	0.05
	2	Devil's Cup	American Pale Ale (APA)	12.0 oz.	0.07
	3	Rise of the Phoenix	American IPA	12.0 ounce	0.07
	4	Sinister	American Double / Impe	12.0 oz	0.09%
	5	Sex and Candy	American IPA	12.0 OZ.	0.08
	6	Black Exodus	Oatmeal Stout	12.0 oz	0.08

	Country (or territory)	Capital
1	China (more)	Beijing
2	Japan (more)	Tokyo
3	DR Congo	Kinshasa
4	Russia (more)	Moscow
5	Indonesia (more)	Jakarta
6	South Korea (more)	
7	Egypt (more)	
8	Mexico	

Country

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

Tables can contain errors

- Errors can be detrimental for data applications

index	beer_name	style	ounces	abv
1	Pub Beer	American Pale Lager	12.0 oz	0.05
2	Devil's Cup	American Pale Ale (APA)	12.0 oz.	0.07
3	Rise of the Phoenix	American IPA	12.0 ounce	0.07
4	Sinister	American Double / Impe	12.0 oz	0.09%
5	Sex and Candy	American IPA	12.0 OZ.	0.08
6	Black Exodus	Oatmeal Stout	12.0 oz	0.08

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

Supervised approach: Need of extensive labeling data

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 labeled examples before reaching the error ?

Supervised approach: Imbalanced dataset

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

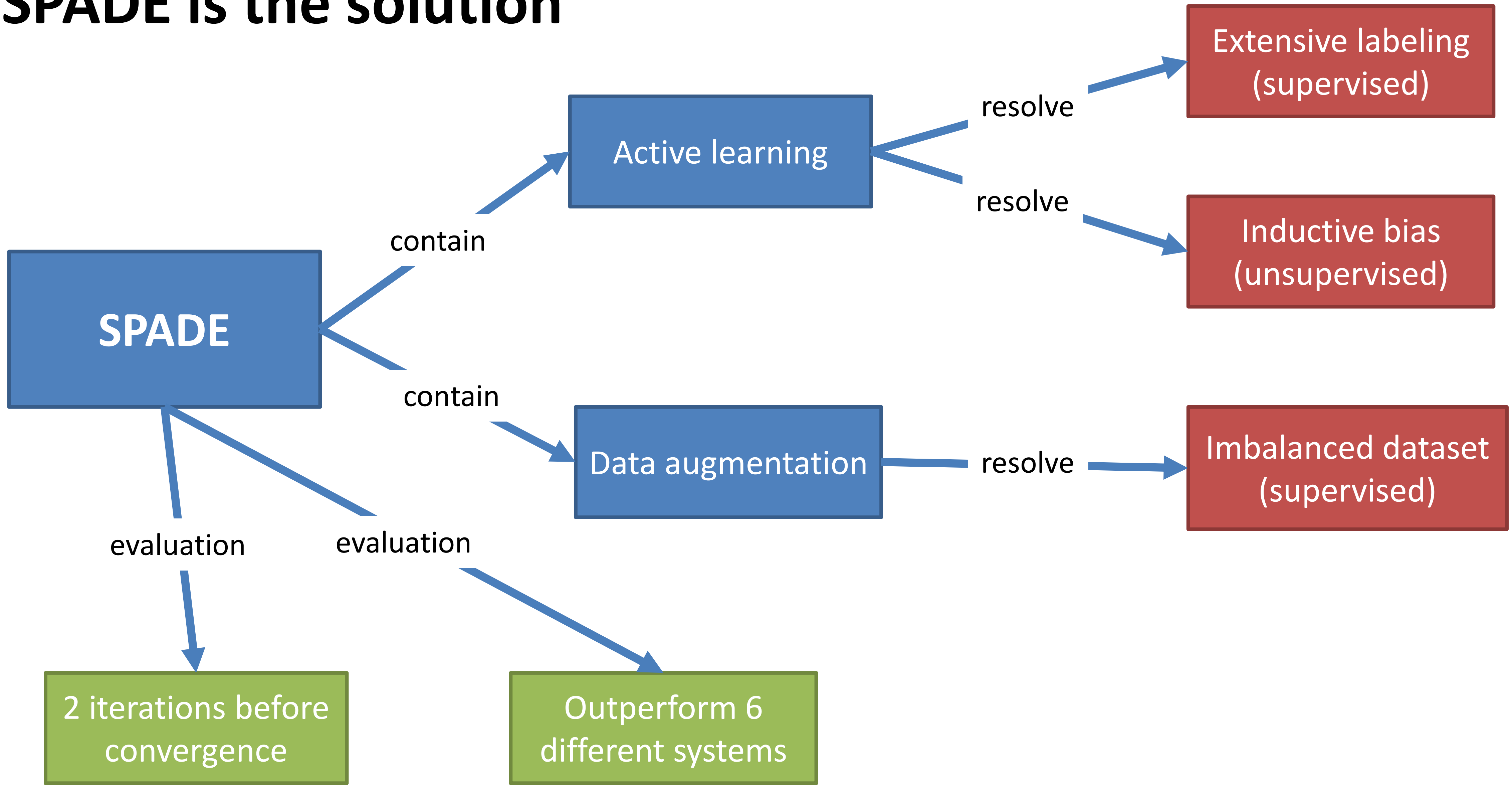
Why only a few errors ?

Unsupervised approach: Inductive bias in method design

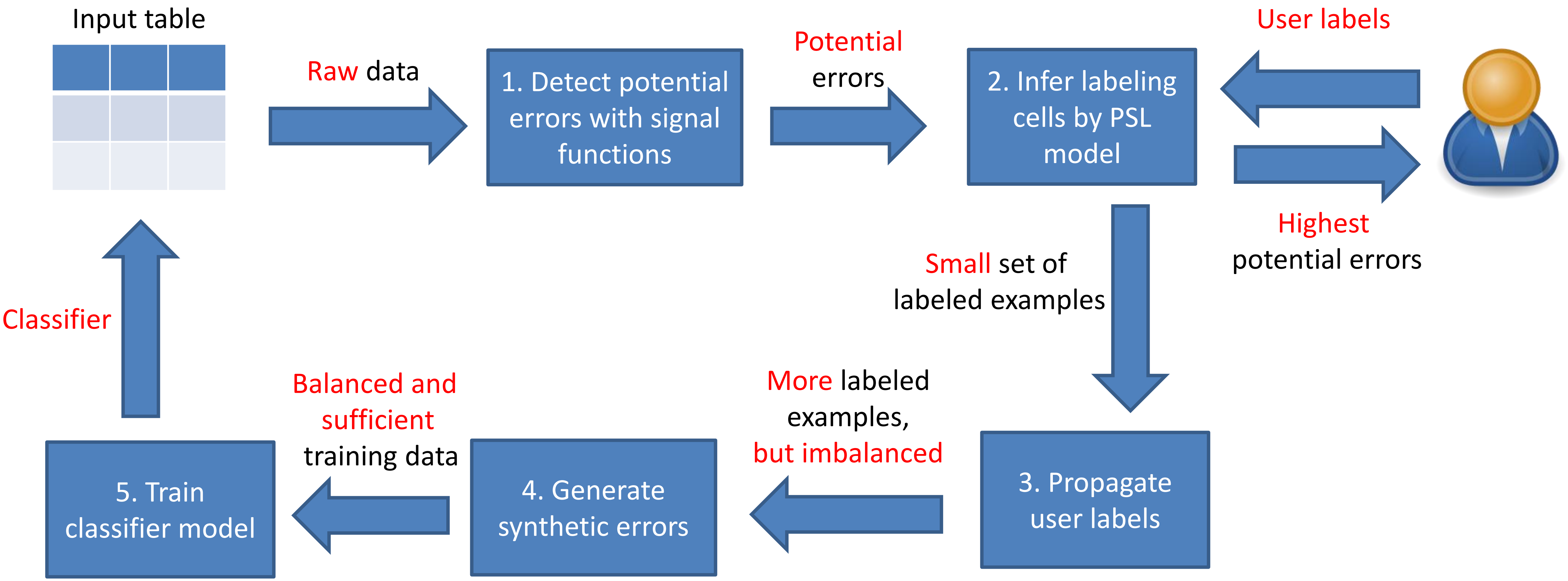
GDP per capita	Voluntary expenditure	Household income	Passenger transport
41 450	2.3	-0.5	138 643
43 746	2.3 |	1.1	132 125
44 720	2.3	0.4	134 954 e

Inductive bias: only negative value in the column

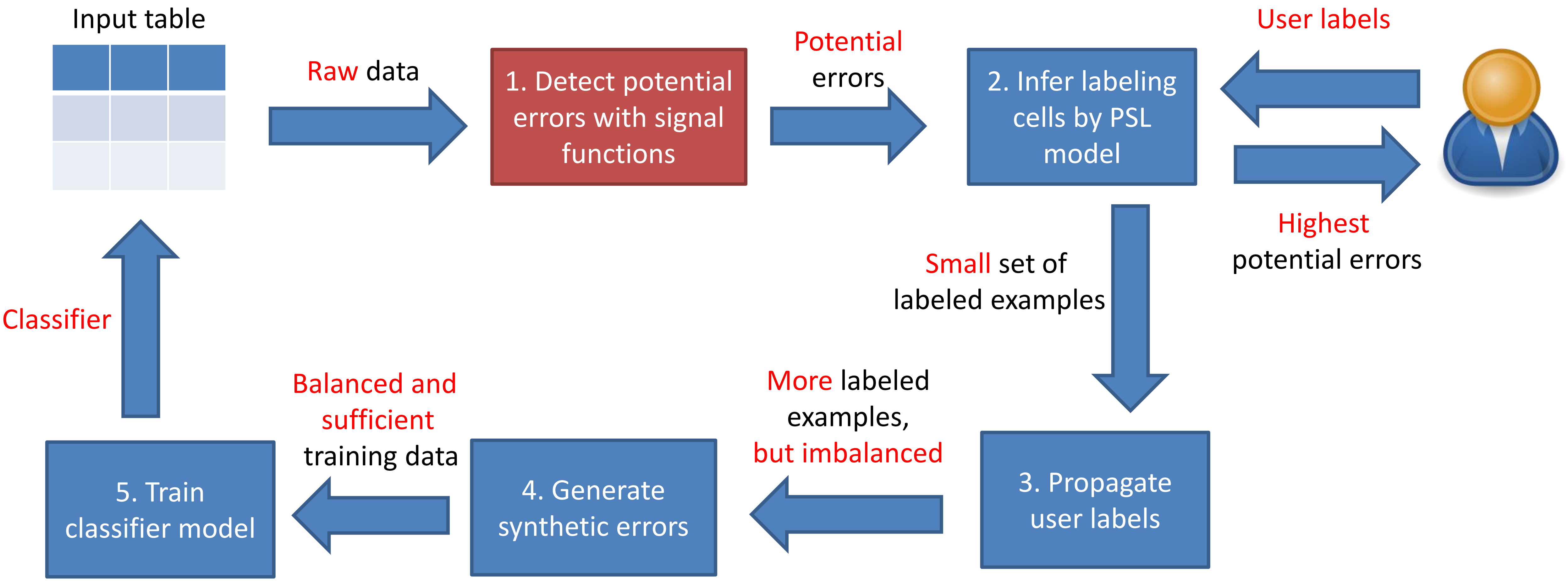
SPADE is the solution



Overall approach



Detect potential errors with signal functions



How to detect potential errors ?

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

Different from other values – Internal signal

Potential Errors

How to detect potential errors ?

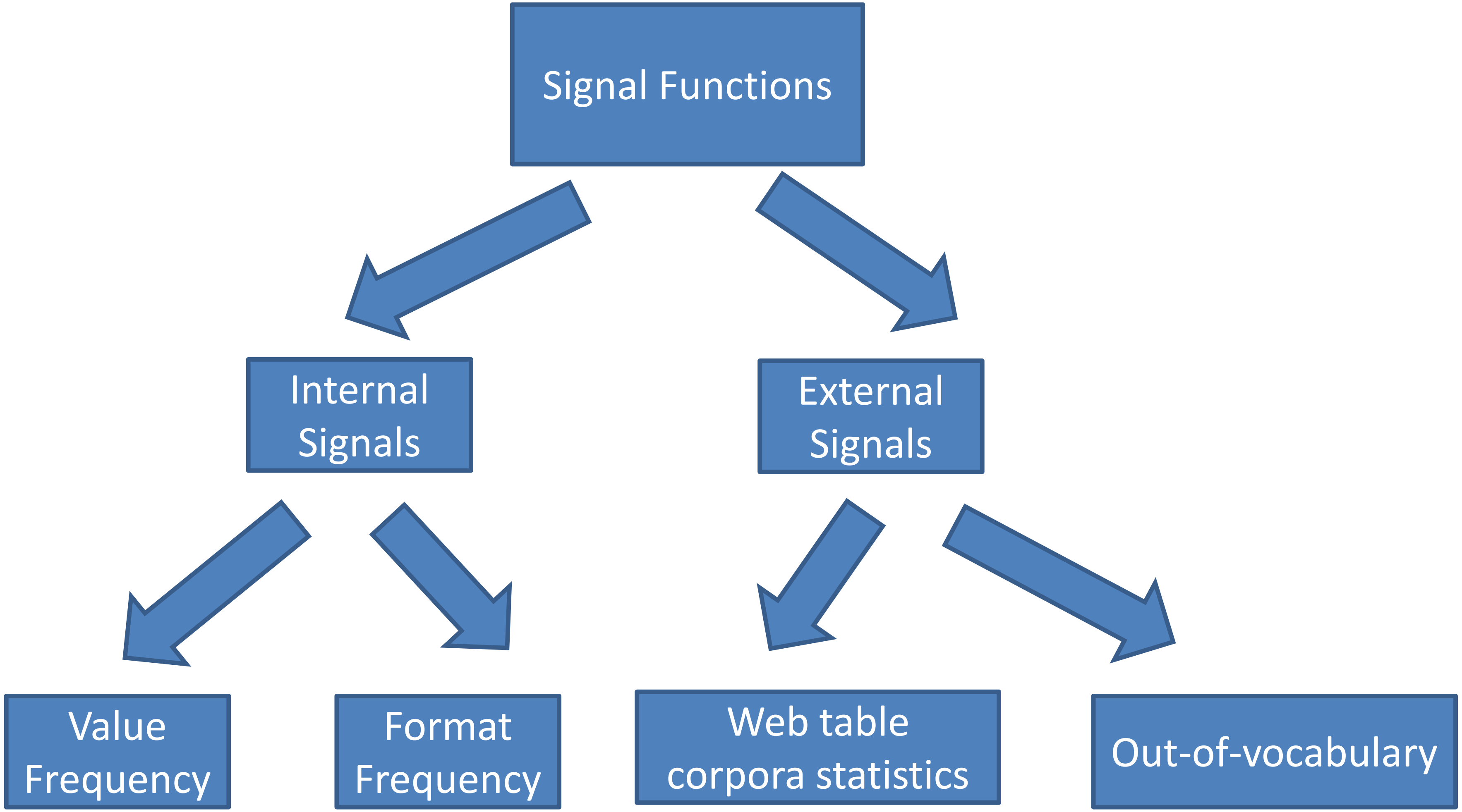
GDP per capita	Voluntary expenditure	Household income	Passenger transport
41 450	2.3	-0.5	138 643
43 746	2.3 |	1.1	132 125
44 720	2.3	0.4	134 954 e

Uncommon formats – External signal

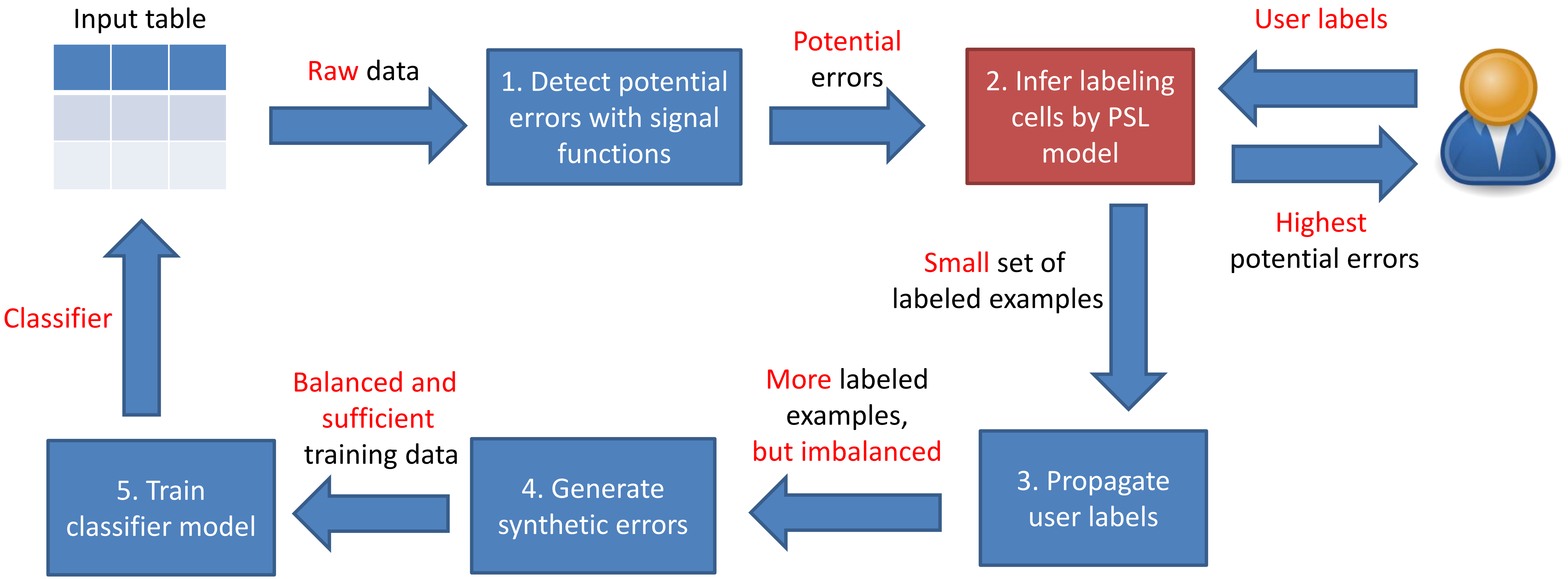
Potential Errors

Huge table corpora

External and internal signals



Overall approach



PSL model: Active learning iteration

PSL Model

Estimate
error probability

Data	Error probabilities
San Francisco CA	0.03
Los Angeles	0.35
Springdale AR	0.71
Bend	0.17
Chicago	0.11

Suggest highest probability
examples to users

Update the good/bad
signals based on user labels

Probabilistic Soft Logic (PSL) model

- A probabilistic graphical model framework using first-order logic

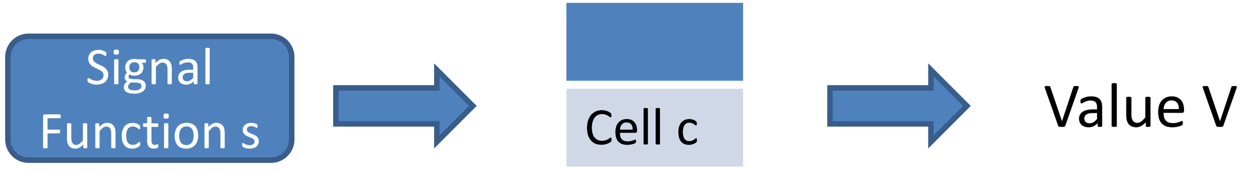
- Two main elements: predicates and rules

- Predicates can have “soft” values [0,1]

PSL model: Predicates

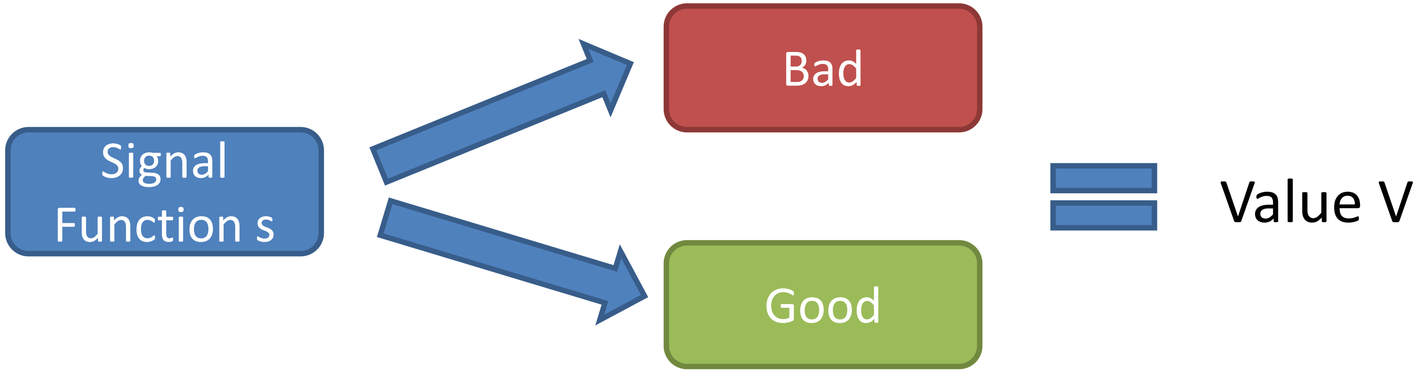
$HasSignal(c, s) = V$

Indicate value of signal function s when applying on cell c



$BadSignal(s) = V$

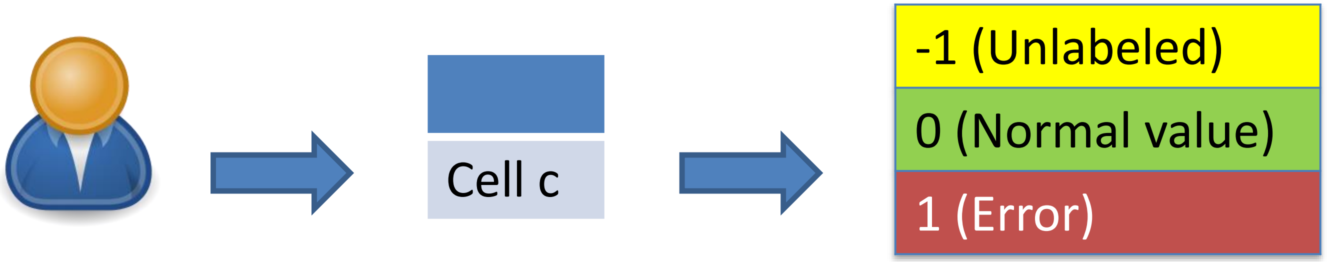
Indicate if a signal is bad or good



PSL model: Predicates

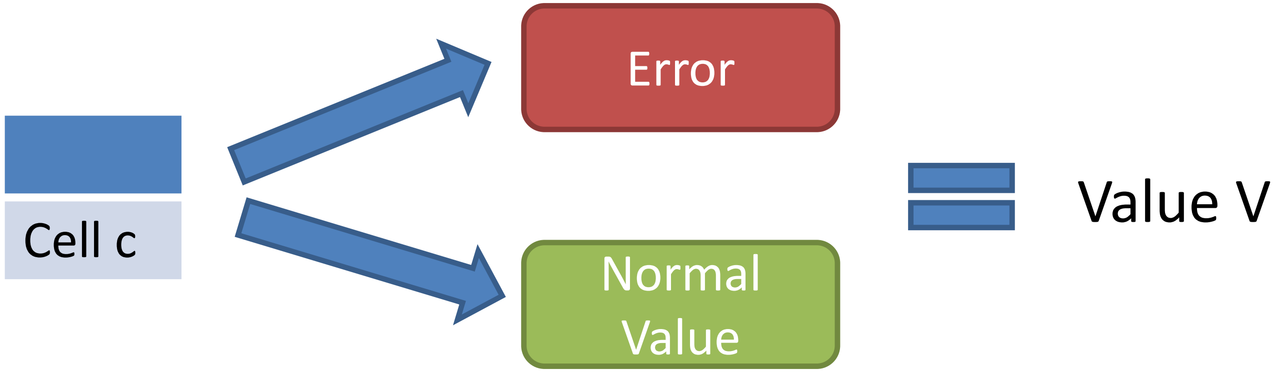
$Label(c, \{-1, 0, 1\}) = \{0, 1\}$

Indicate user label of cell c



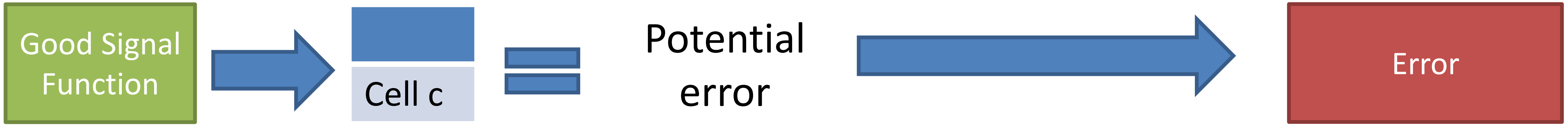
$Error(c) = V$

Indicate error probability of cell c

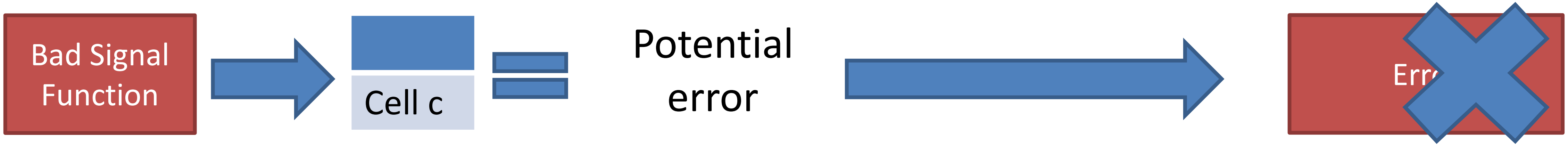


PSL rules: Error probabilities

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



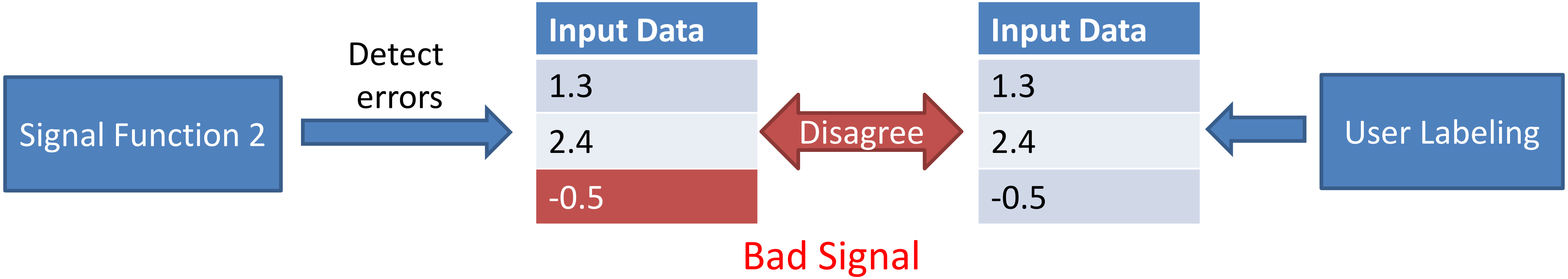
$$BadSignal(s) \wedge HasSignal(c,s) \Rightarrow \neg Error(c)$$



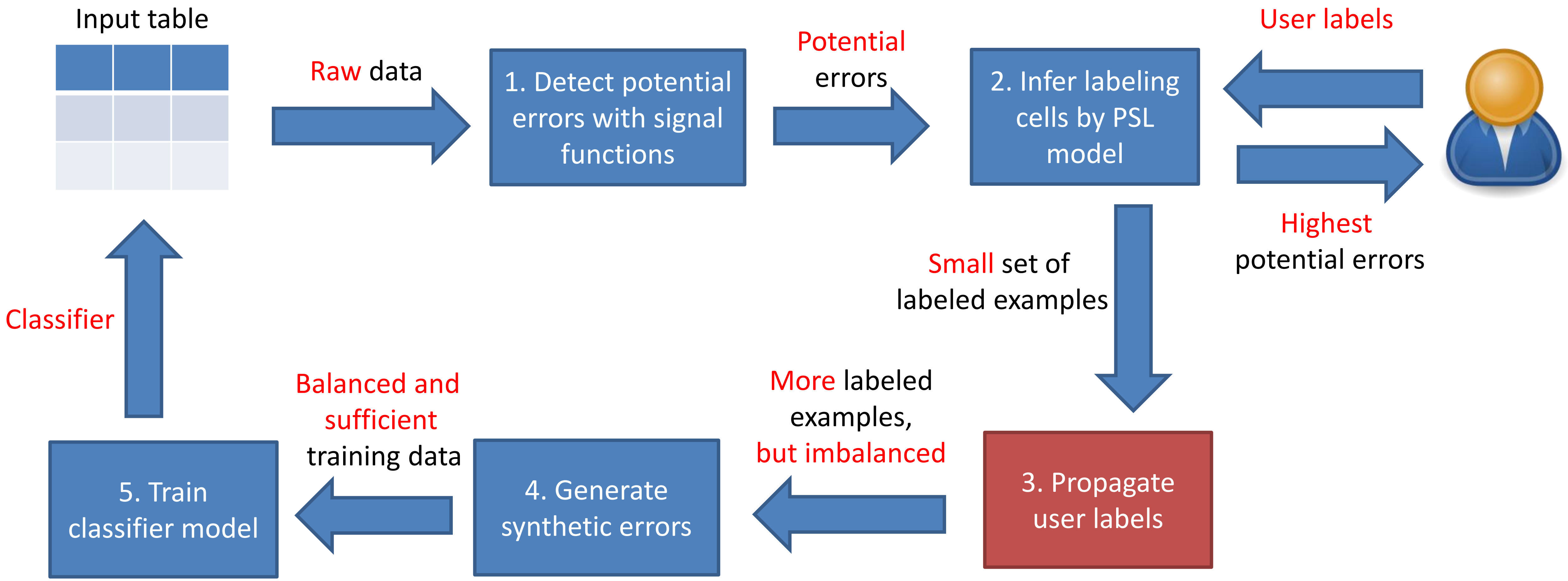
PSL rules: Signal function and user labeling

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

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



Overall approach



Label propagation

Data	Error probability
San Francisco CA	0.7
Los Angeles	0.35
Springdale AR	0.71
Bend	0.17
Chicago	0.11

User labeling

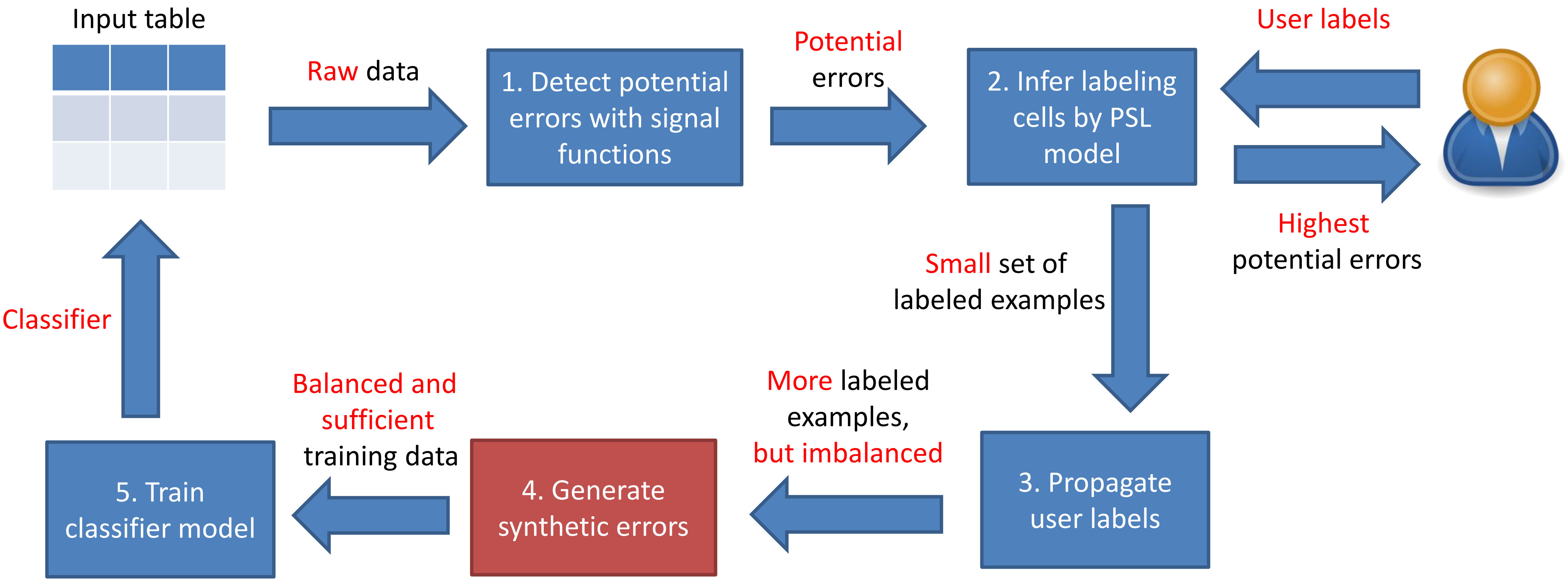
Data	Error probability
San Francisco CA	0.7
Los Angeles	0.35
Springdale AR	0.71
Bend	0.17
Chicago	0.11

Label propagation

$$d(e) = |e_1 - e_2| \leq \epsilon$$
$$\epsilon = 0.1$$

Data	Error probability
San Francisco CA	0.7
Los Angeles	0.35
Springdale AR	0.71
Bend	0.17
Chicago	0.11

Overall approach



Synthetic error generation

Data	Error Score
San Francisco CA	0.55
Los Angeles	0.35
Springdale AR	0.59
Bend	0.17
Chicago	0.11

Data	Cleaned data
San Francisco CA	San Francisco
Los Angeles	
Springdale AR	Springdale
Bend	
Chicago	

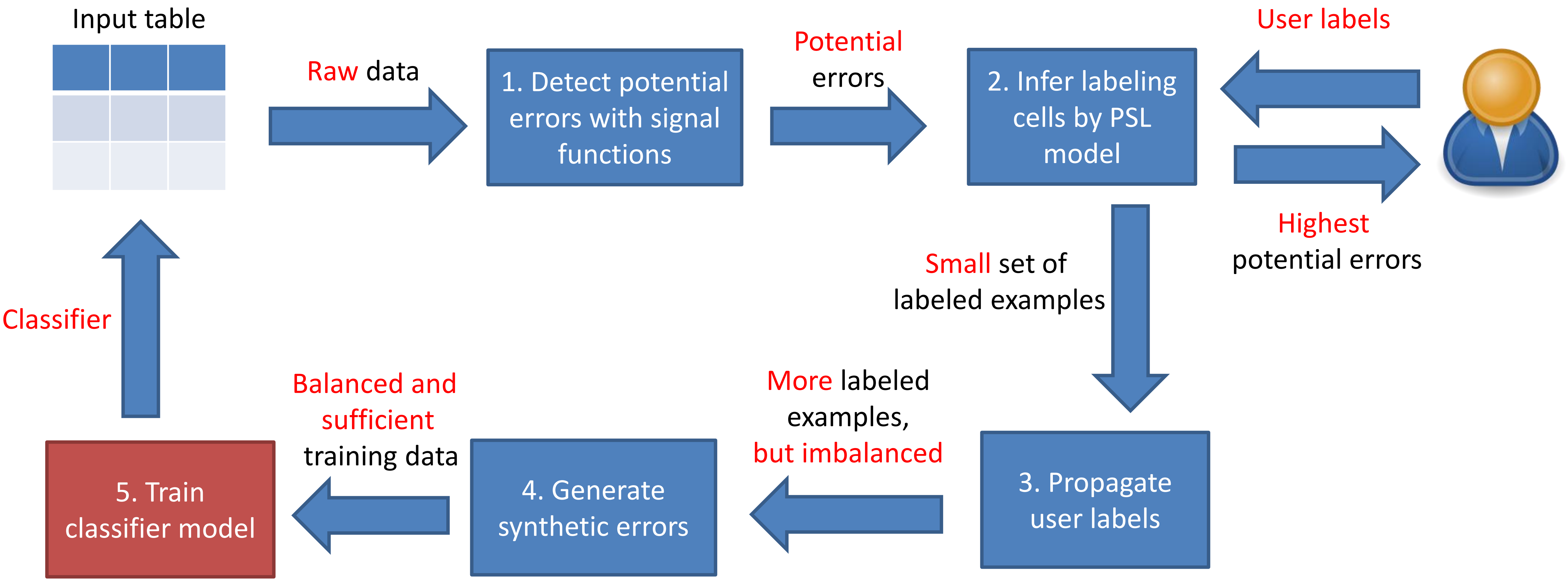
Learn transformations to convert cleaned data to erroneous data

INSERT_END(“CA”)
INSERT_END(“AR”)

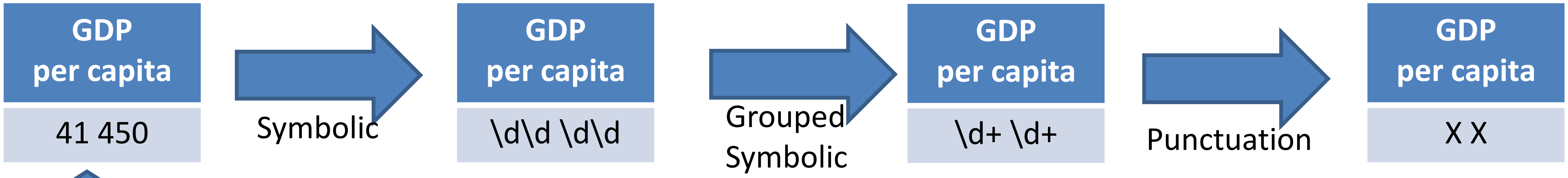
Data
San Francisco CA
Los Angeles
Springdale AR
Bend
Chicago

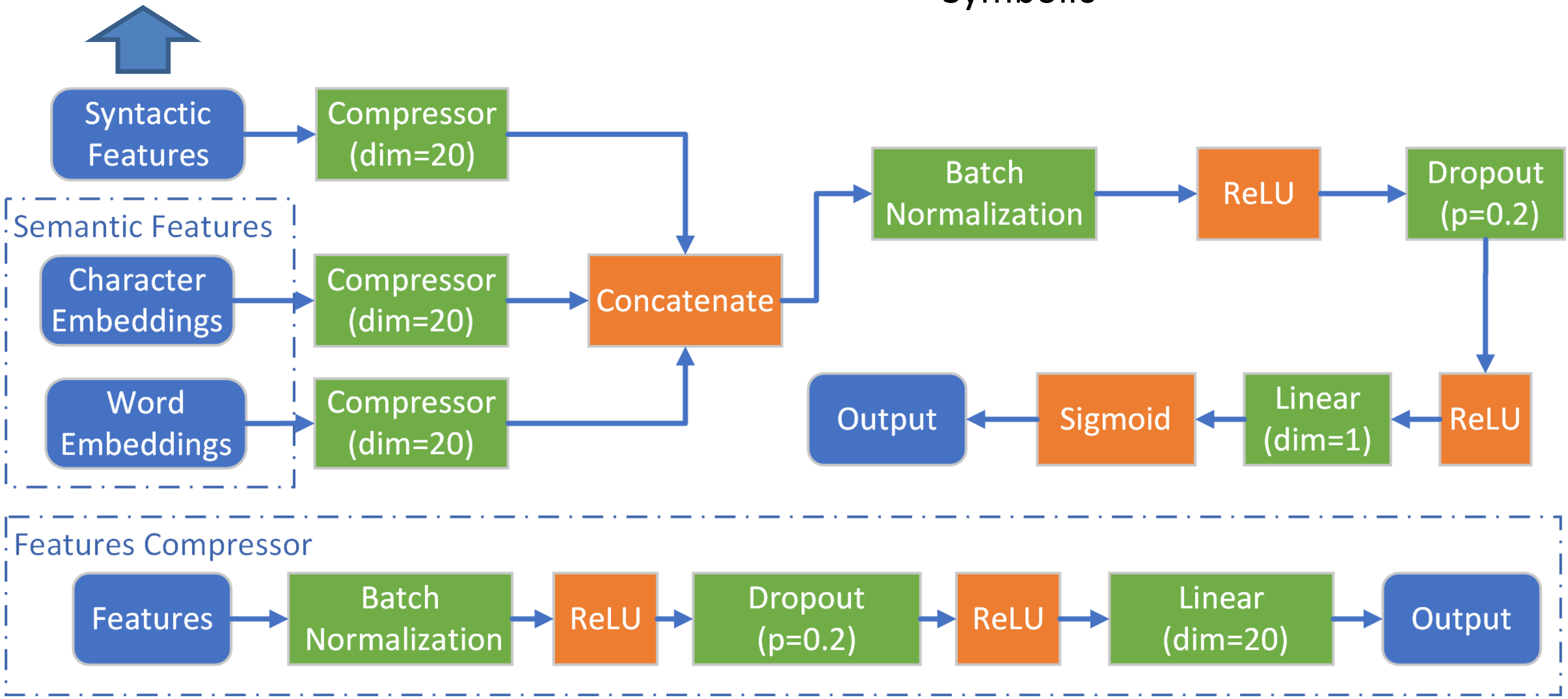
Generated Errors
San Francisco CA CA
Los Angeles CA
Springdale AR AR
Bend CA
Chicago AR

Overall approach



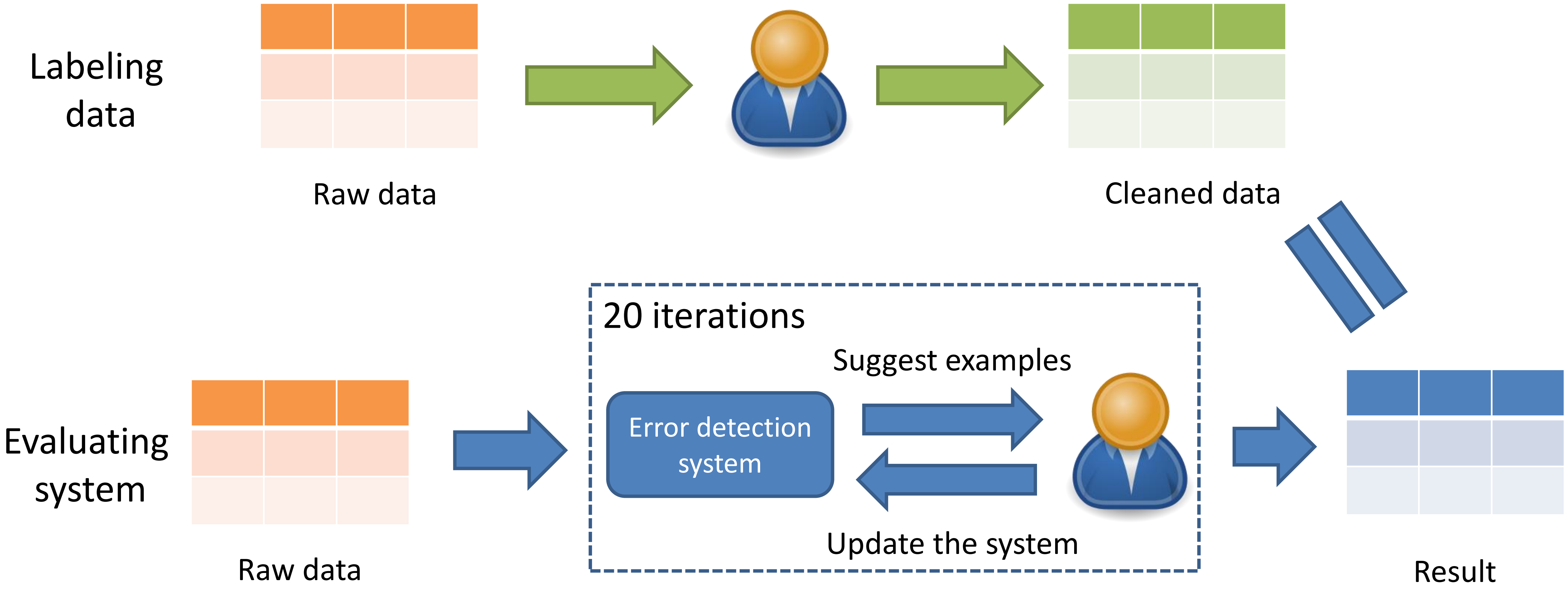
Training classifier model





- Optimizer: AdamW
- Batch size: 128
- Learning rate: 0.0005
- Loss function: Cross Entropy
- Epochs: 500 * 128 / size

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**

Conclusion

- 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