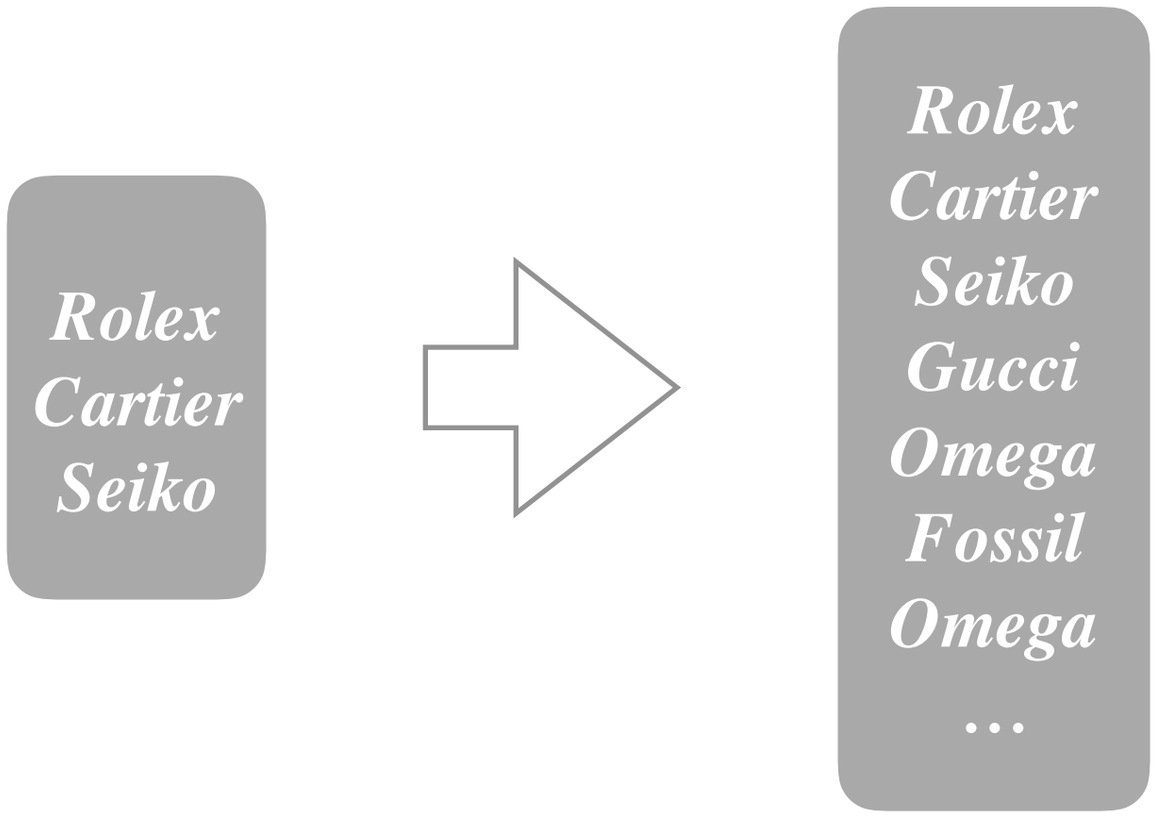
Entity Set Expansion

Set Expansion

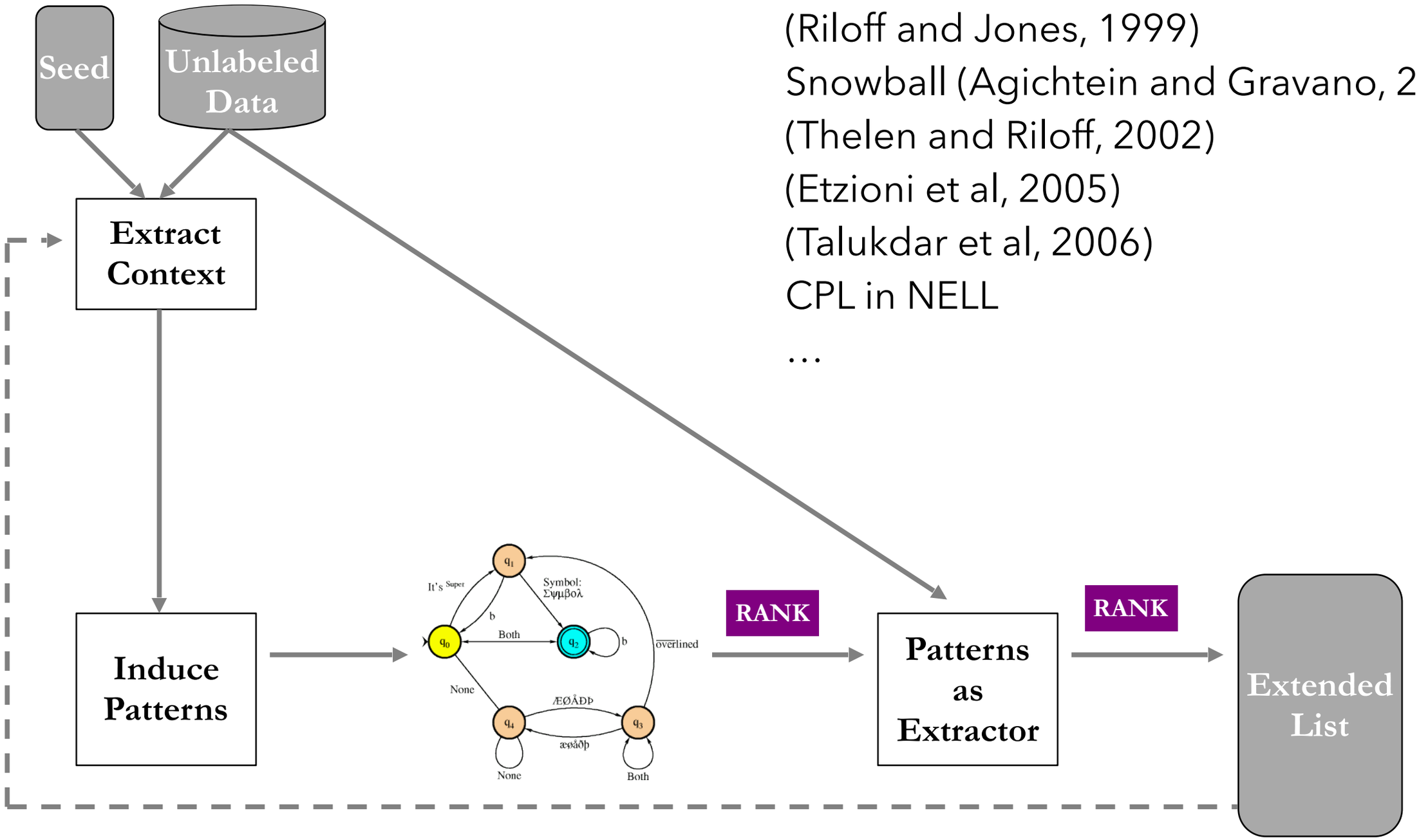
Given seed instances from a class, automatically identify more instances from that class



Many applications:

web advertising, knowledge graph population, …

Context Pattern Induction



DIPRE (Brin)
(Riloff and Jones, 1999)
Snowball (Agichtein and Gravano, 2000)
(Thelen and Riloff, 2002)
(Etzioni et al, 2005)
(Talukdar et al, 2006)
CPL in NELL
…

Extractions using Context Patterns

Induced Patterns (containing sequence "*watch*")

gold -<*ENT*>- watch	Richemont , -<*ENT*>- watches	Rolex watches are sold through official -<*ENT*>- and
diamond -<*ENT*>- watch	bought -<*ENT*>- watches	bought a -<*ENT*>- watch
fake -<*ENT*>- watches	fake -<*ENT*>- watch	watchmaker -<*ENT*>- SA
bought -<*ENT*>- watch	diamond -<*ENT*>- watches	Ulysse -<*ENT*>- watches
encrusted -<*ENT*>- watch	stole -<*ENT*>- watches	Rolex watches and -<*ENT*>- watch
stole -<*ENT*>- watch	buy a -<*ENT*>- watch	Rolex , -<*ENT*>- watch
Richemont AG , -<*ENT*>- watches	jewelry , including -<*ENT*>- watch	Rolex and -<*ENT*>- watch
Rolex and -<*ENT*>- watches	watchmaker -<*ENT*>- .	diamond - studded -<*ENT*>- watch
buy -<*ENT*>- watches	jewelry , including -<*ENT*>- watches	diamond - encrusted -<*ENT*>- watch
Cartier and -<*ENT*>- watches	stole a -<*ENT*>- watch	Cartier , and -<*ENT*>- watches
buy -<*ENT*>- watch	Rolex watches and -<*ENT*>- .	buy a -<*ENT*>- watches
gold -<*ENT*>- watches	watchmaker -<*ENT*>- Group	bought a -<*ENT*>- watches

Entities Extracted by Above Patterns (ranked)

Rolex *(most confident)*	Fossil	Swatch
Cartier	Tag Heuer	Super Bowl
Swiss	Chanel	SPOT
Movado	Tiffany	Sekonda
Seiko	TechnoMarine	Rolexes
Gucci	Franck Muller	Harry Winston
Patek Philippe	Versace	Hampton Spirit
Piaget	Raymond Weil	Girard Perregaux
Omega	Guess	Frank Mueller
Citizen	Croton	David Yurman
Armani	Audemars Piguet	Chopard
DVD	DVDs	Chinese
Breitling	Montres Rolex	Armitron
Tourneau	CD	NFL *(least confident)*

Extracted Lists Improve NER Taggers

Training Data (Tokens)	Test-a		
	No List	Seed List	Unsup. List
9229	68.27	70.93	**72.26**
204657	89.52	84.30	**90.48**

SEAL: Set Expansion using the Web

[Wang and Cohen, ICDM 2007]

1. Canon
2. Nikon
3. Olympus

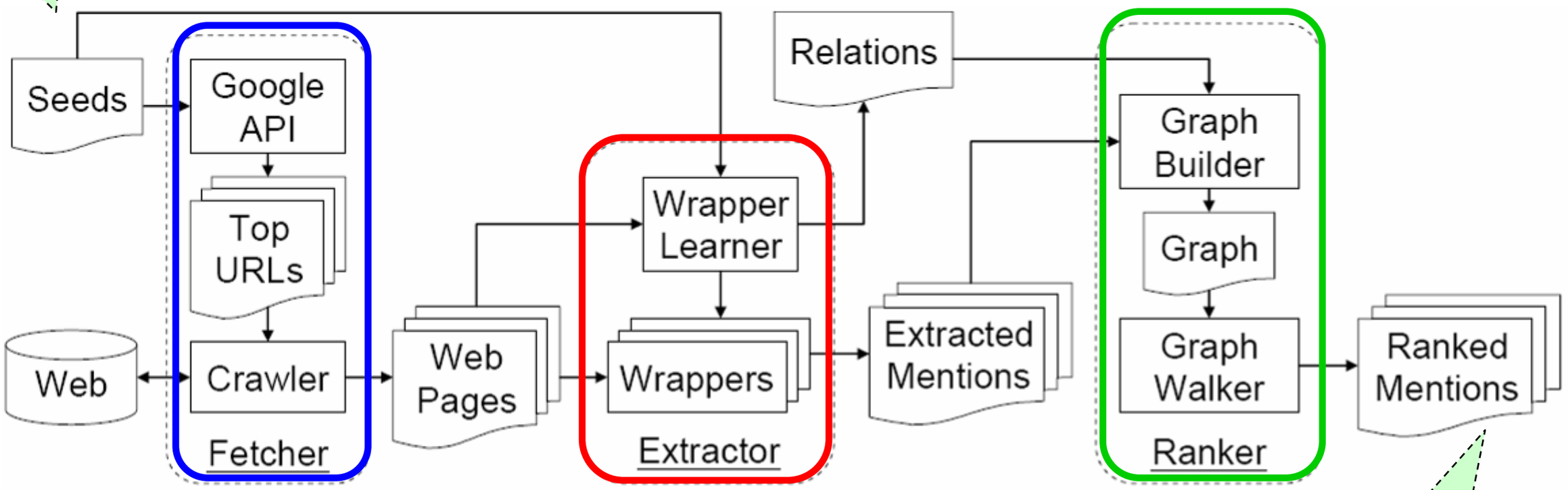


Figure 1. Flow chart of the SEAL system

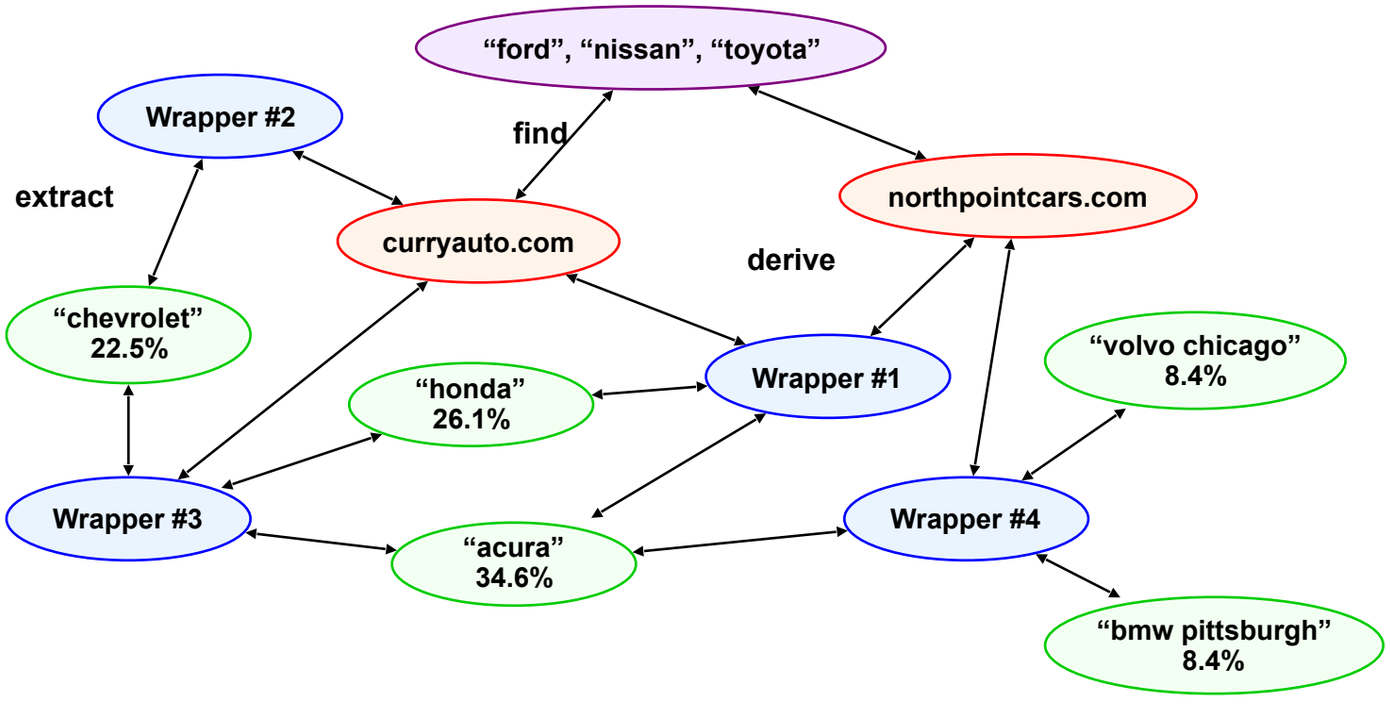
- **Fetcher**: download web pages from the Web
- **Extractor**: learn wrappers from web pages
- **Ranker**: rank entities extracted by wrappers

4. Pentax
5. Sony
6. Kodak
7. Minolta
8. Panasonic
9. Casio
10. Leica
11. Fuji
12. Samsung
13. …

```
<li class="ford"><a href="http://www.curryauto.com/">
<img src="/common/logos/ford/logo-horiz-rgb-lg-dkbg.gif" alt="3"></a>
    <ul><li class="last"><a href="http://www.curryauto.com/">
        <span class="dName">Curry Ford</span>...</li></ul>
</li>

<li class="nissan"><a href="http://www.curryauto.com/">
<img src="/common/logos/nissan/logo-horiz-rgb-lg-dkbg.gif" alt="6"></a>
    <ul><li class="last"><a href="http://www.geisauto.com/">
        <span class="dName">Curry Nissan</span>...</li></ul>
</li>
```

Top three are the seeds

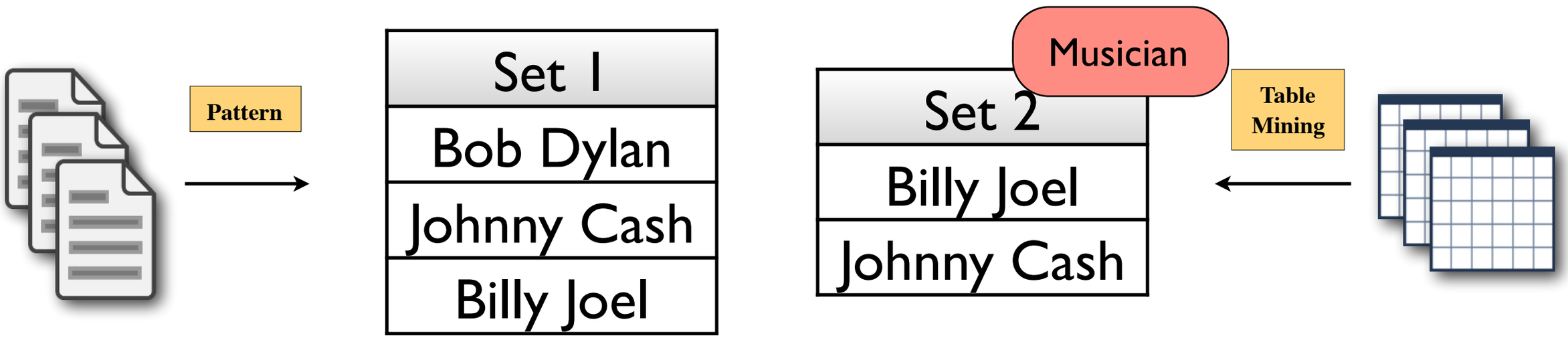


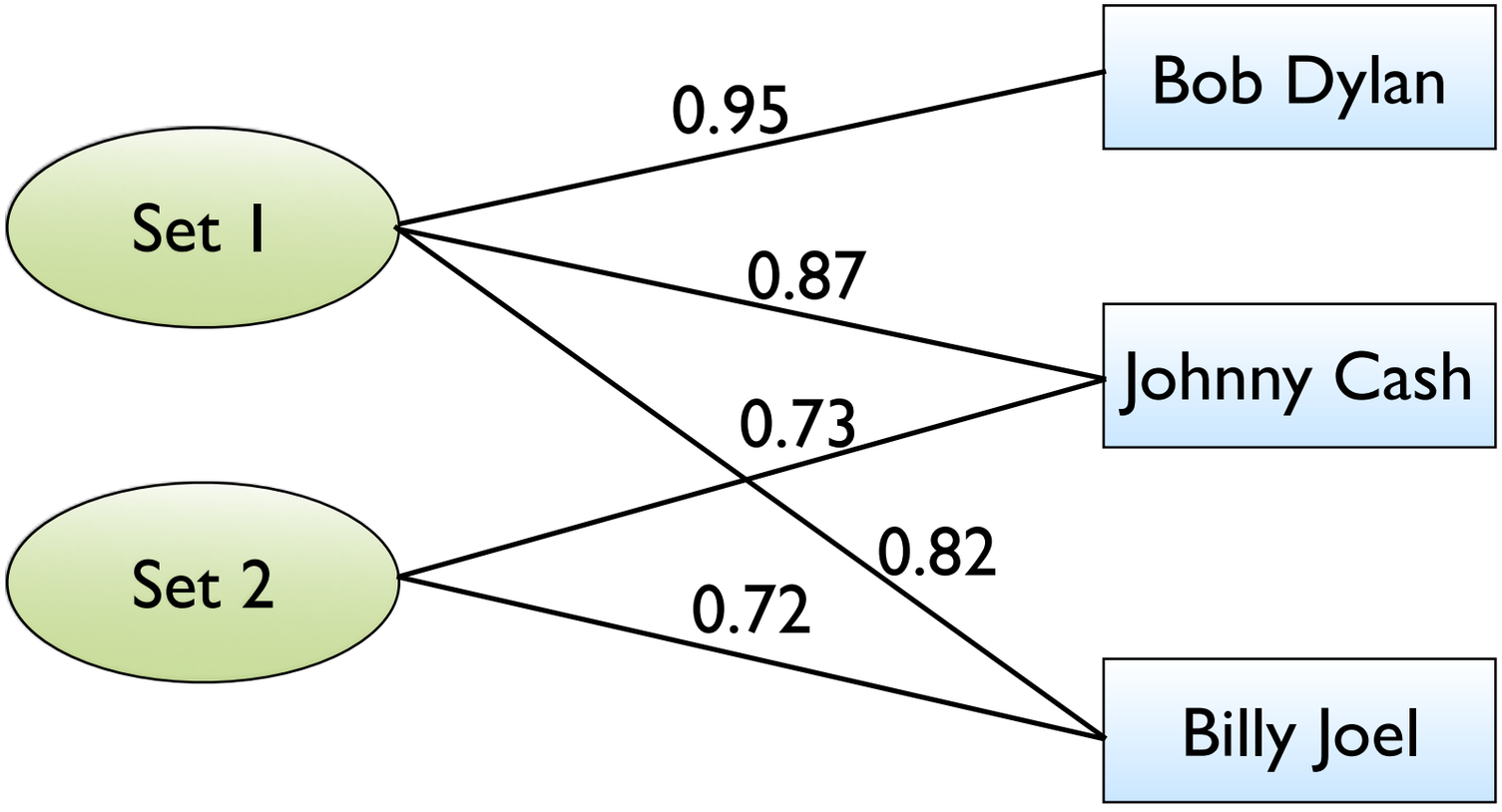
#	Entity
1	andrew mccallum
2	michael collins
3	john lafferty
4	naftali tishby
5	fernando pereira
6	zoubin ghahramani
7	daphne koller
8	thomas hofmann
9	thorsten joachims
10	david heckerman
11	nir friedman
12	tom mitchell
13	dan roth
14	william w. cohen
15	mark craven
16	roni rosenfeld
17	david mcallester
18	yoram singer
19	michael i. jordan
20	eugene charniak
21	amir globerson
22	yiming yang
23	yoshua bengio
24	sridhar mahadevan

#	Entity
1	睡美人
2	灰姑娘
3	白雪公主
4	小紅帽
5	美人魚
6	小美人魚
7	美女與野獸
8	花木蘭
9	青蛙王子
10	貝兒
11	木偶奇遇記
12	糖果屋
13	三隻小豬
14	茉莉公主
15	茉莉
16	愛麗絲夢遊仙境
17	寶嘉康蒂
18	長髮姑娘
19	人魚公主
20	紅舞鞋
21	唐老鴨
22	長靴貓
23	拇指神童
24	小熊維尼

Combining Patterns and Tables using Graph-based SSL

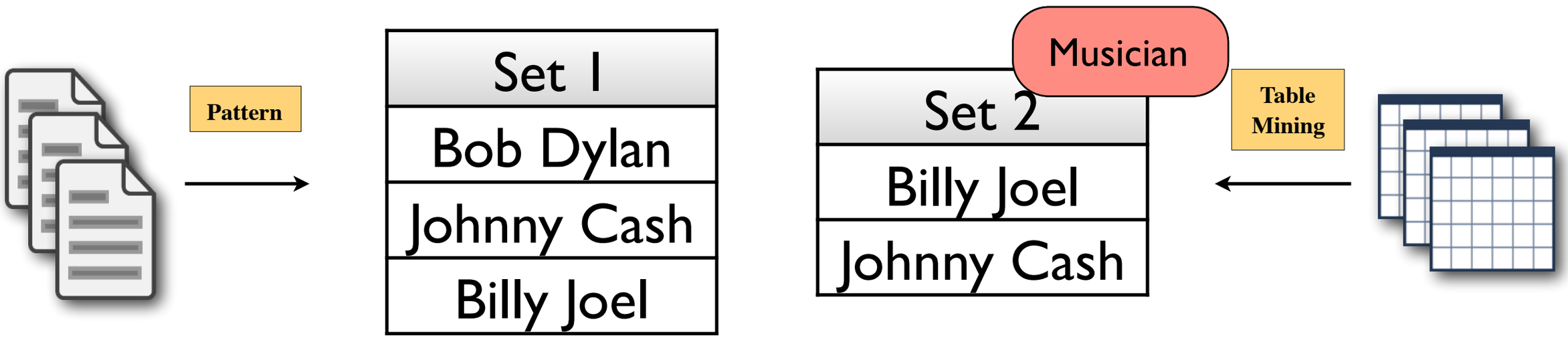
[Talukdar et al., EMNLP 2008, 2010]

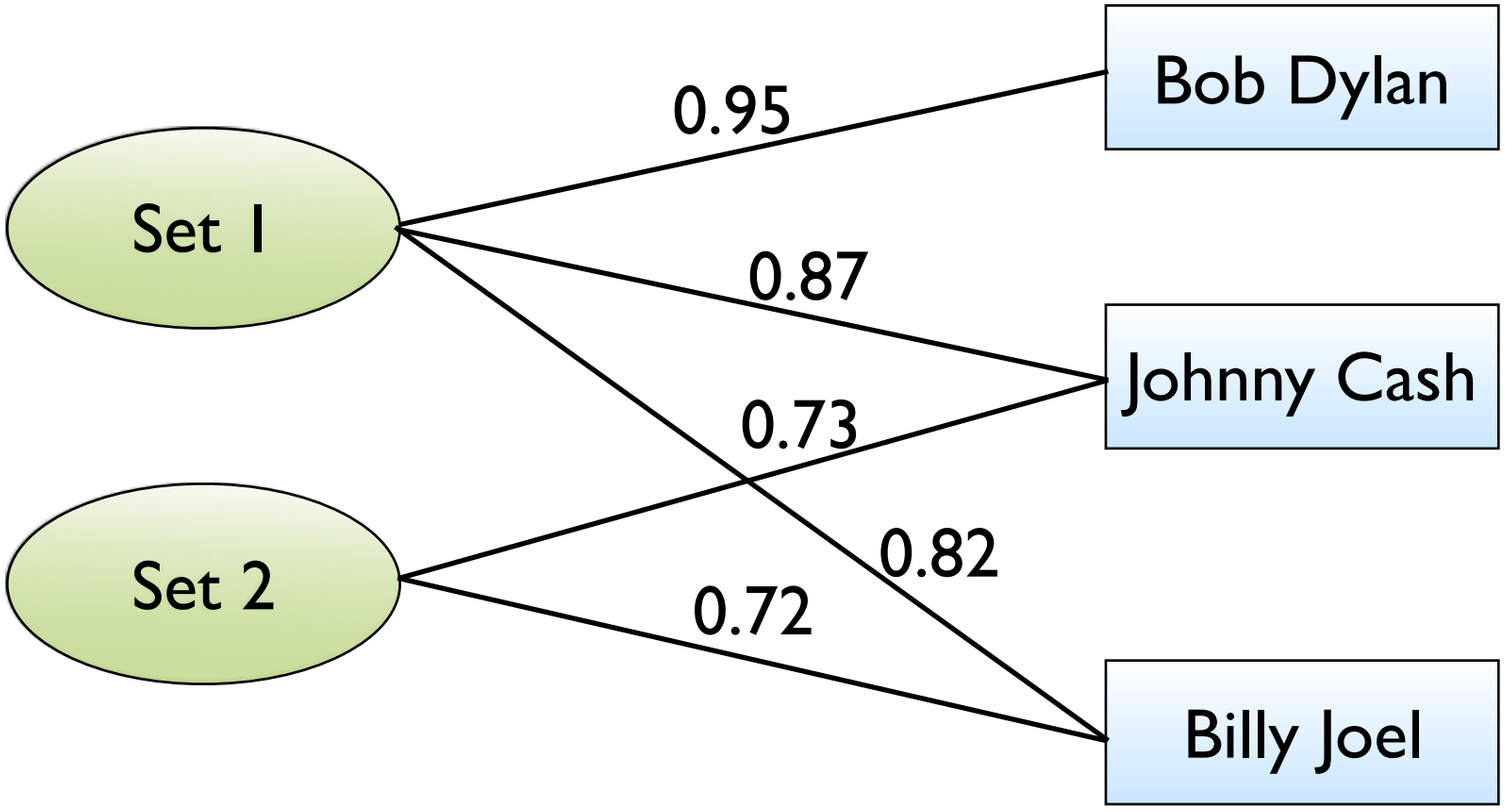




Combining Patterns and Tables using Graph-based SSL

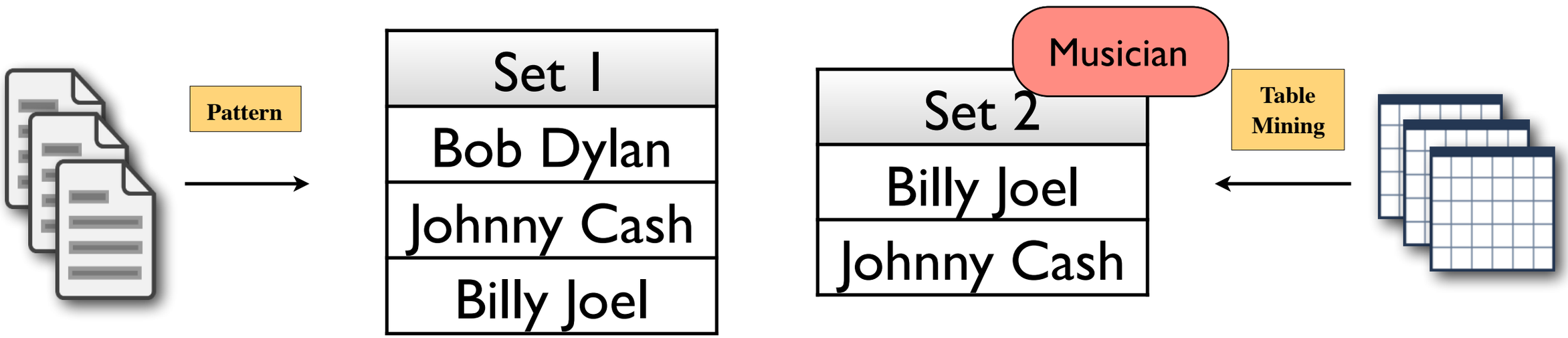
[Talukdar et al., EMNLP 2008, 2010]

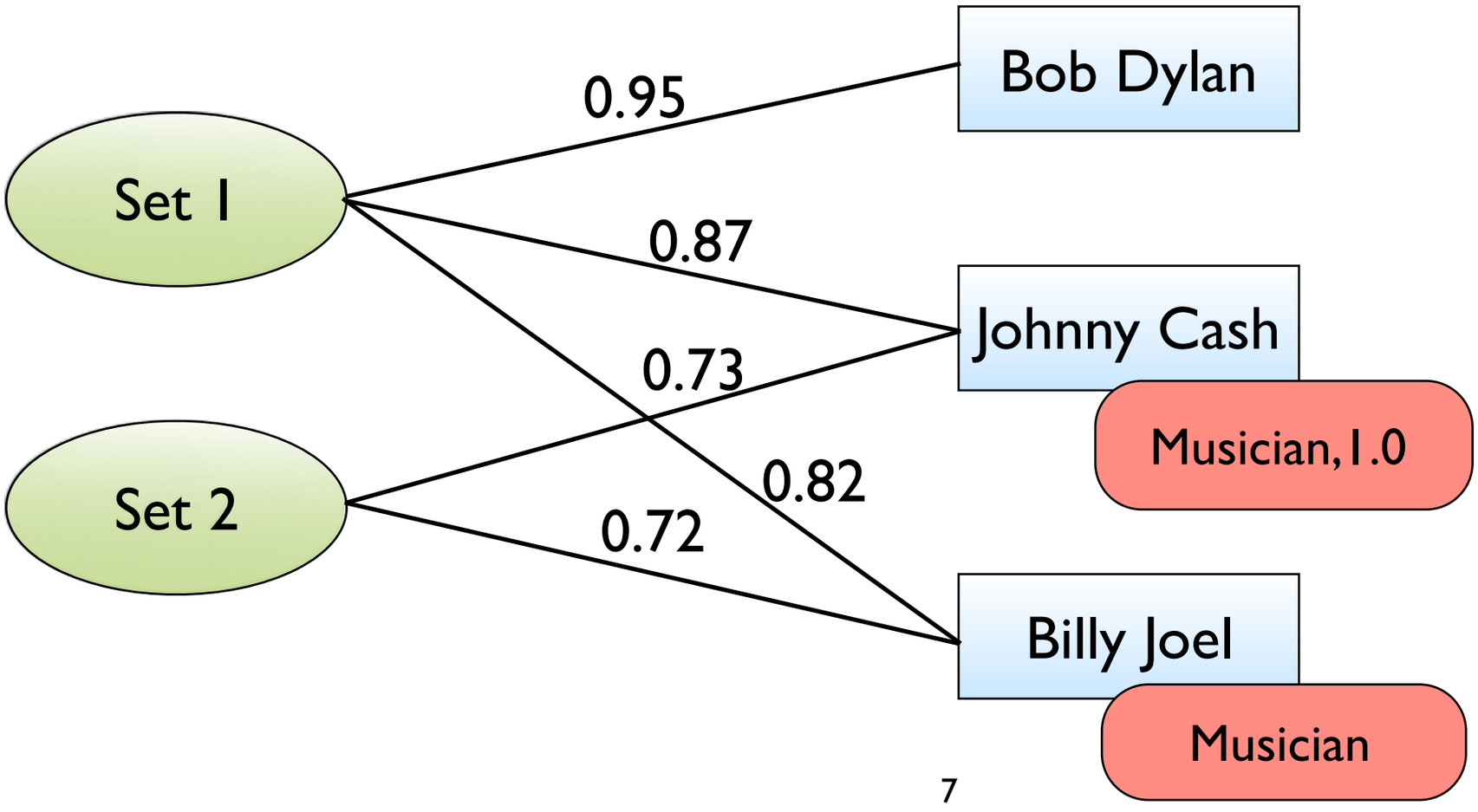




Combining Patterns and Tables using Graph-based SSL

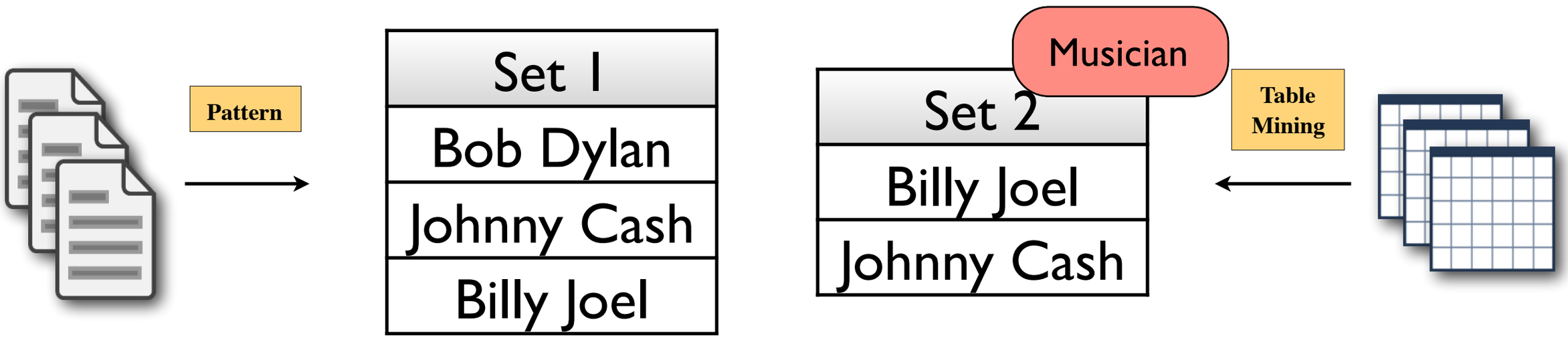
[Talukdar et al., EMNLP 2008, 2010]

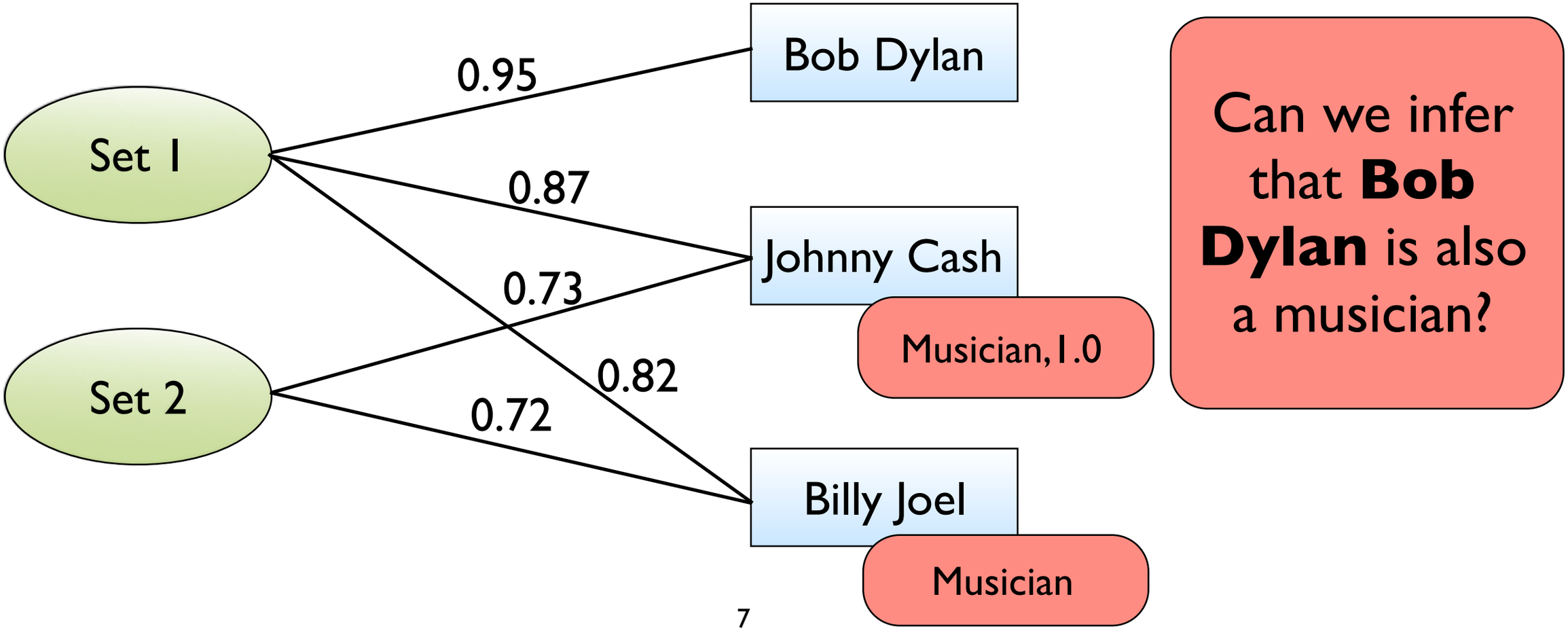




Combining Patterns and Tables using Graph-based SSL

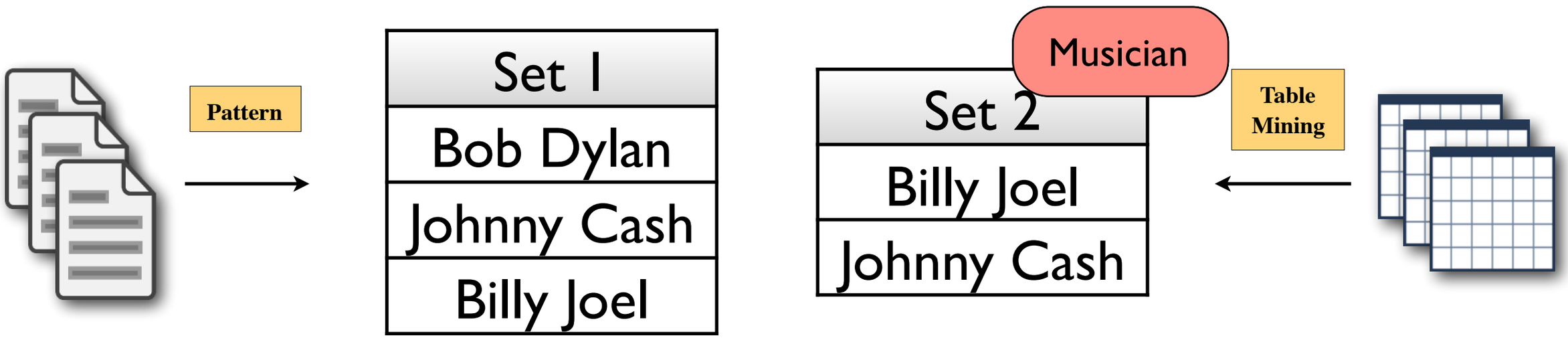
[Talukdar et al., EMNLP 2008, 2010]

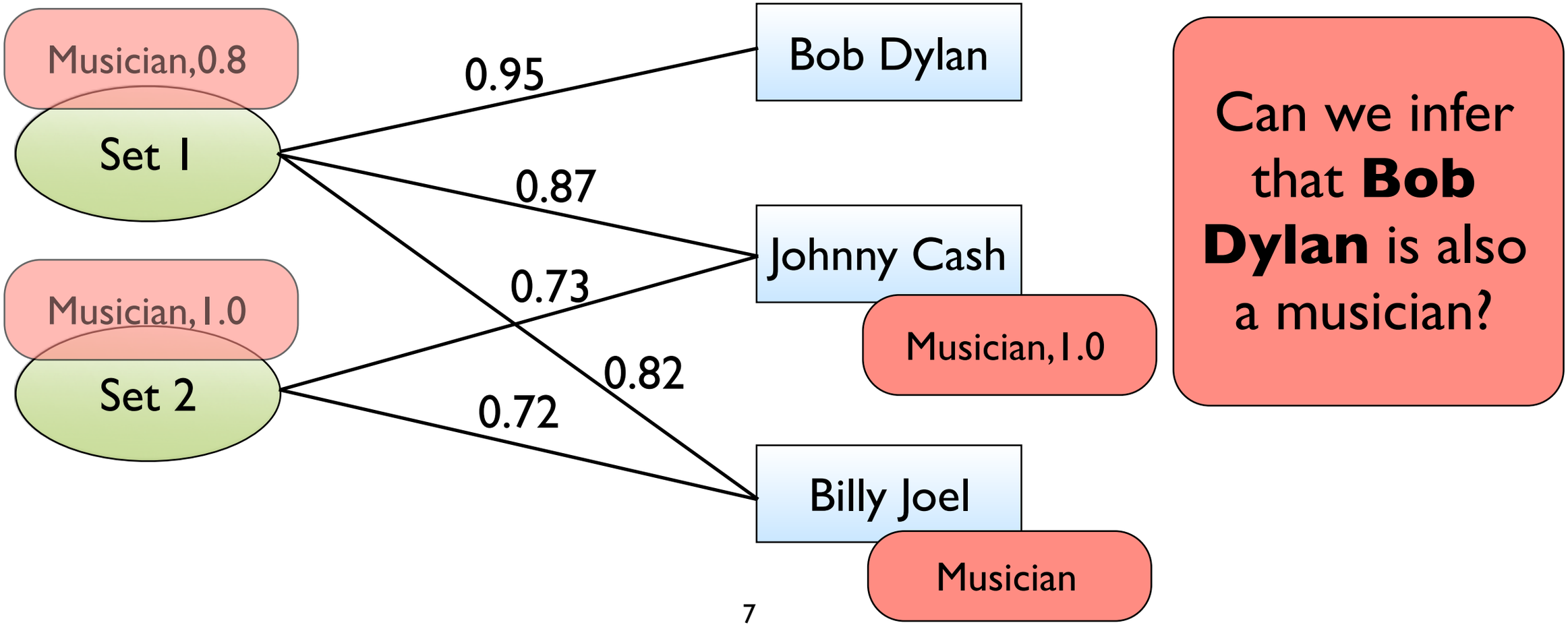




Combining Patterns and Tables using Graph-based SSL

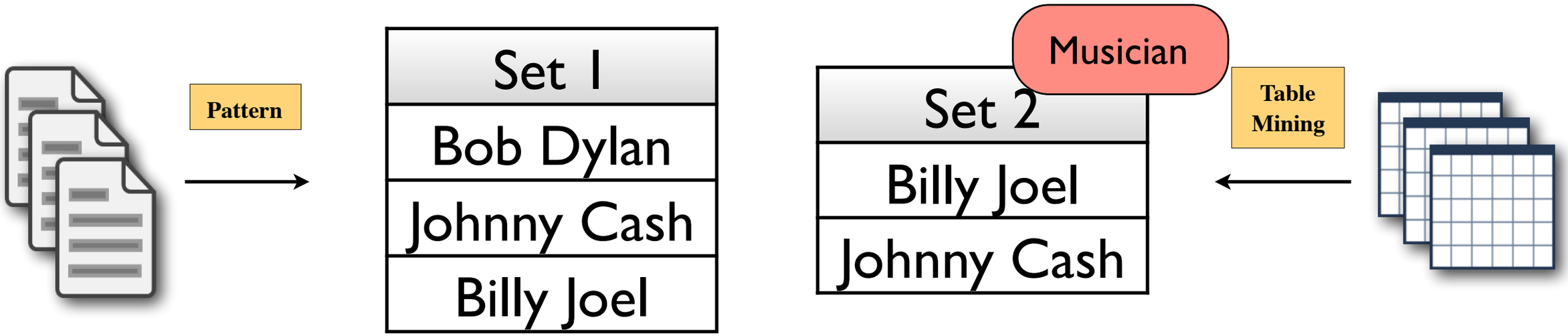
[Talukdar et al., EMNLP 2008, 2010]

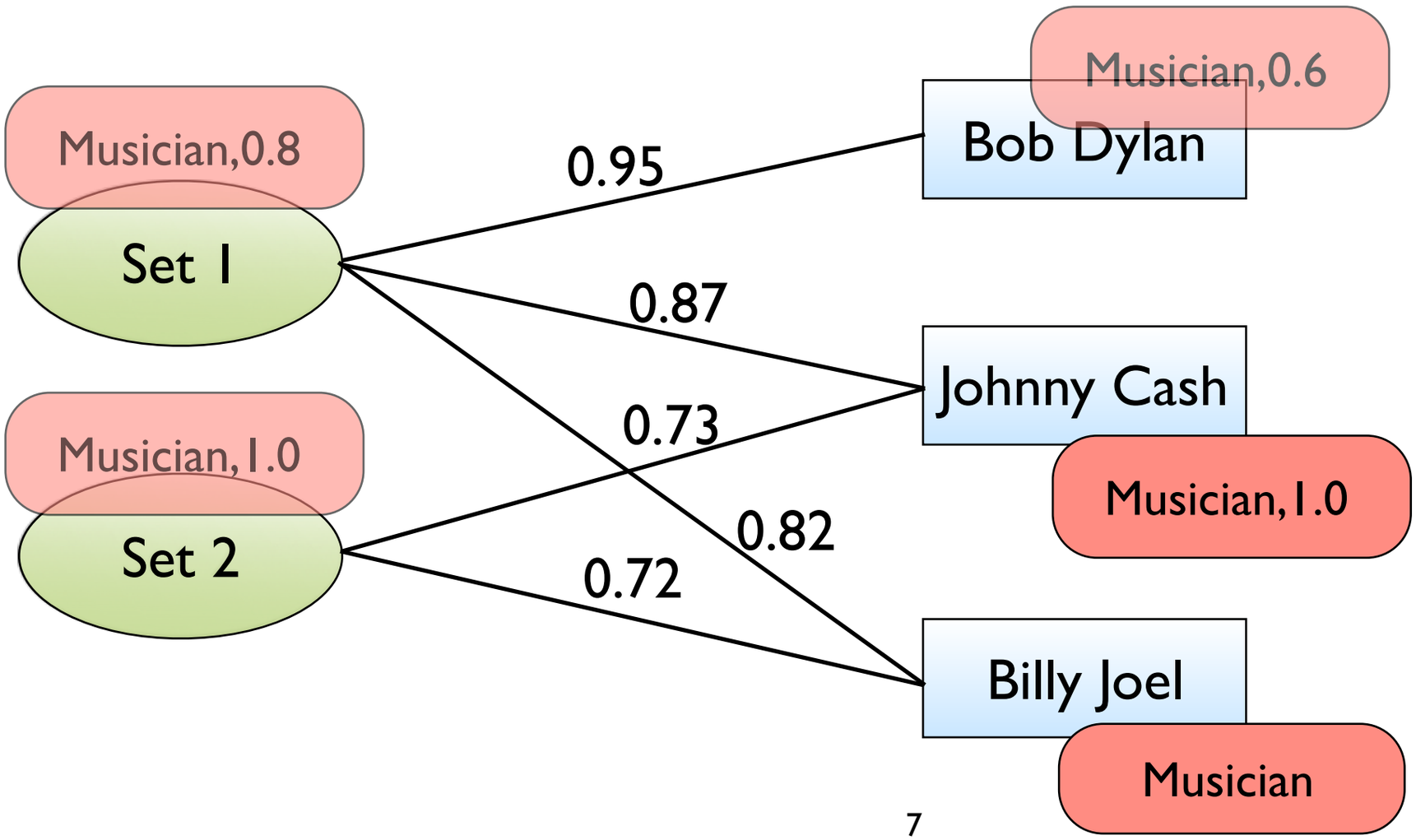




Combining Patterns and Tables using Graph-based SSL

[Talukdar et al., EMNLP 2008, 2010]

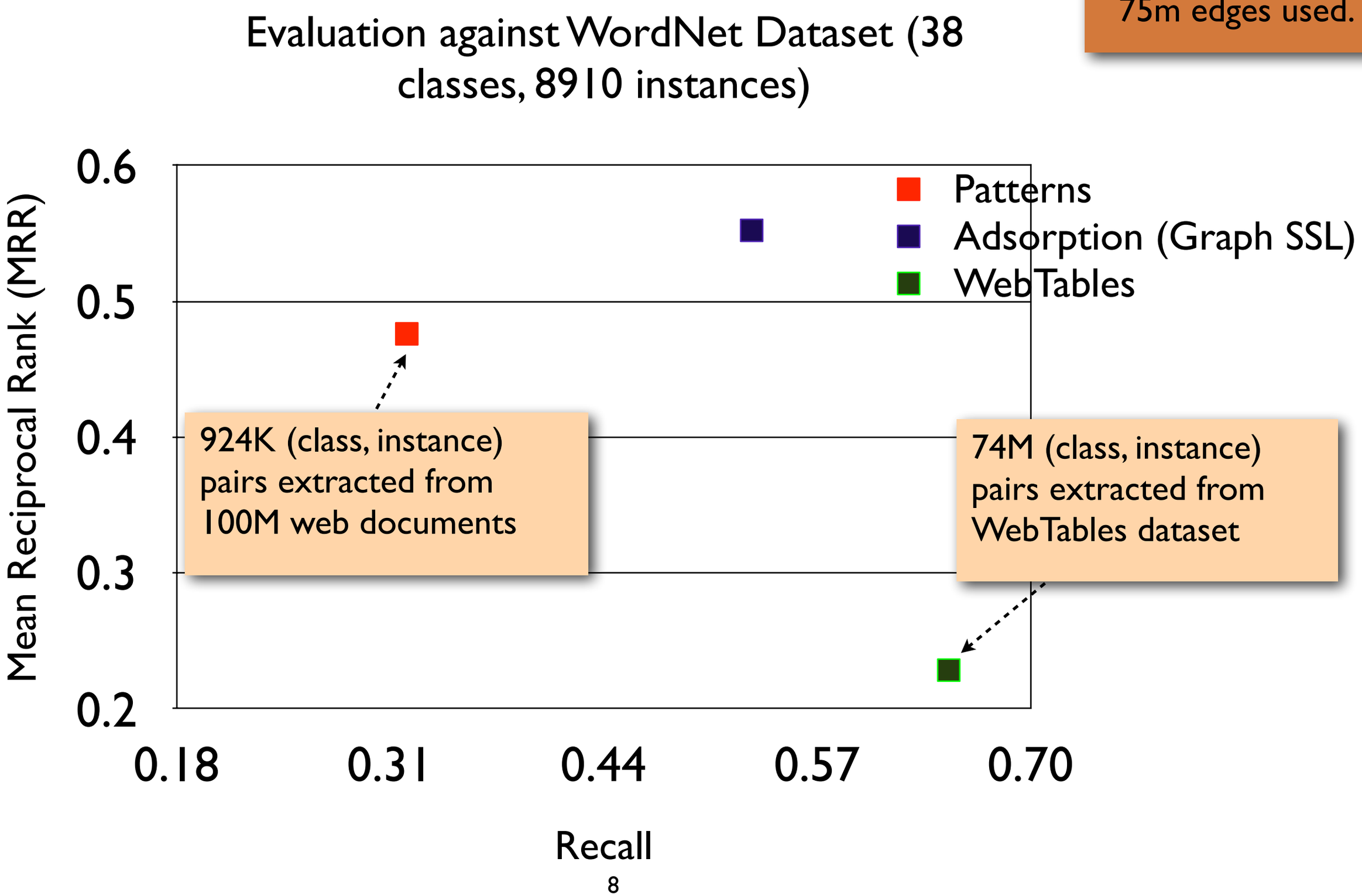




Can we infer that **Bob Dylan** is also a musician?

Extraction for Known Instances

Graph with
1.4m nodes,
75m edges used.



Extracted Pairs

Total classes: 9081

Class	Some non-seed Instances found
Scientific Journals	Journal of Physics, Nature, Structural and Molecular Biology, Sciences Sociales et sante, Kidney and Blood Pressure Research, American Journal of Physiology-Cell Physiology, …
NFL Players	Tony Gonzales, Thabiti Davis, Taylor Stubblefield, Ron Dixon, Rodney Hannan, …
Book Publishers	Small Night Shade Books, House of Ansari Press, Highwater Books, Distributed Art Publishers, Cooper Canyon Press, …

EgoSet [Rong et al., WSDM 2016]