Learning on the Web

Fernando Pereira

Thanks to
Koby Crammer, Kuzman Ganchev, João Graça, Yi Liu, Marius Pasça, Franz Och, Joseph Reisinger, Deepak Ravichandran, Stefan Riezler, Partha Talukdar, Ben Taskar, Alexander Vasserman, and other colleagues at Google and University of Pennsylvania.
Penn work funded by NSF

Web Inference and Learning

- Building and using the Web requires algorithms that can draw reliable inferences from the wealth of evidence implicit in Web data
  - How to interpret this term in this context?
  - Does this sentence answer that question?
  - Will this user click on that ad?
- *Learning*: create concise representations to support good inferences

Beyond Supervised Learning

- Explicit annotation is costly and misleading
  - *Expert annotation*: difficult to agree on annotation criteria, repeated failures to achieve self-sustainability:
  - *User annotation*: what's its immediate value to the user, attractive to spammers
- Learning from what users do: we constantly seek and organize information
  - Examples: *machine translation*, *query expansion*
- Lots of (mostly) unlabeled data

Learning from Parallel Data

Statistical Machine Translation

- *Inference*: is this phrase a good translation of that phrase in this context?
- Indirect evidence: translation pairs, monolingual text
- Memory-based (non-parametric) learning:
  - Bilingual: phrase translation table
  - Monolingual: language model

Statistical Machine Translation

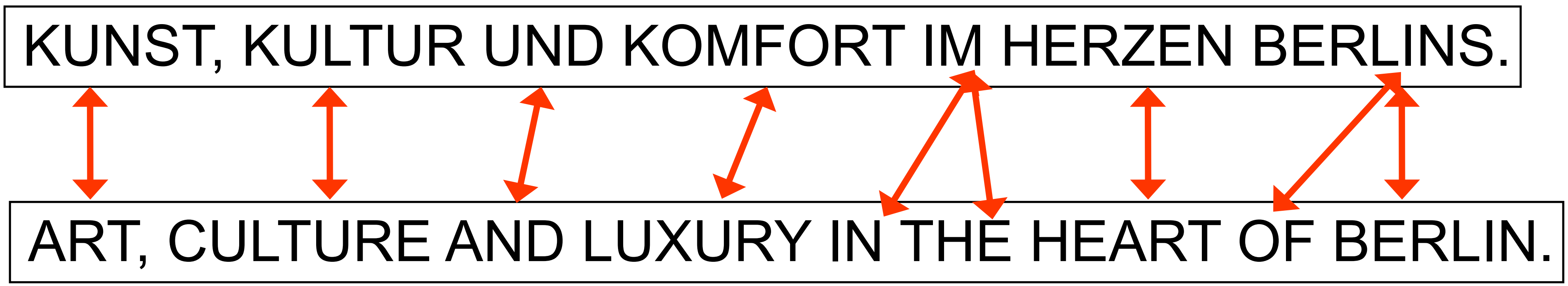
Machine learning from *human* communication

- Parallel texts:

DE	EN
SEHR GEEHRTER GAST! KUNST, KULTUR UND KOMFORT IM HERZEN BERLIN.	DEAR GUESTS, ART, CULTURE AND LUXURY IN THE HEART OF BERLIN.
DIE ÖRTLICHE NETZSPANNUNG BETRÄGT 220/240 VOLT BEI 50 HERTZ.	THE LOCAL VOLTAGE IS 220/240 VOLTS 50 HZ.

Building a Translator

- Align (cf. genomic alignments…)



- Extract phrase pairs from alignments
- Model the statistics of the target language
- To translate a source text:
  - search over phrase-to-phrase translations
  - filter with model of target language

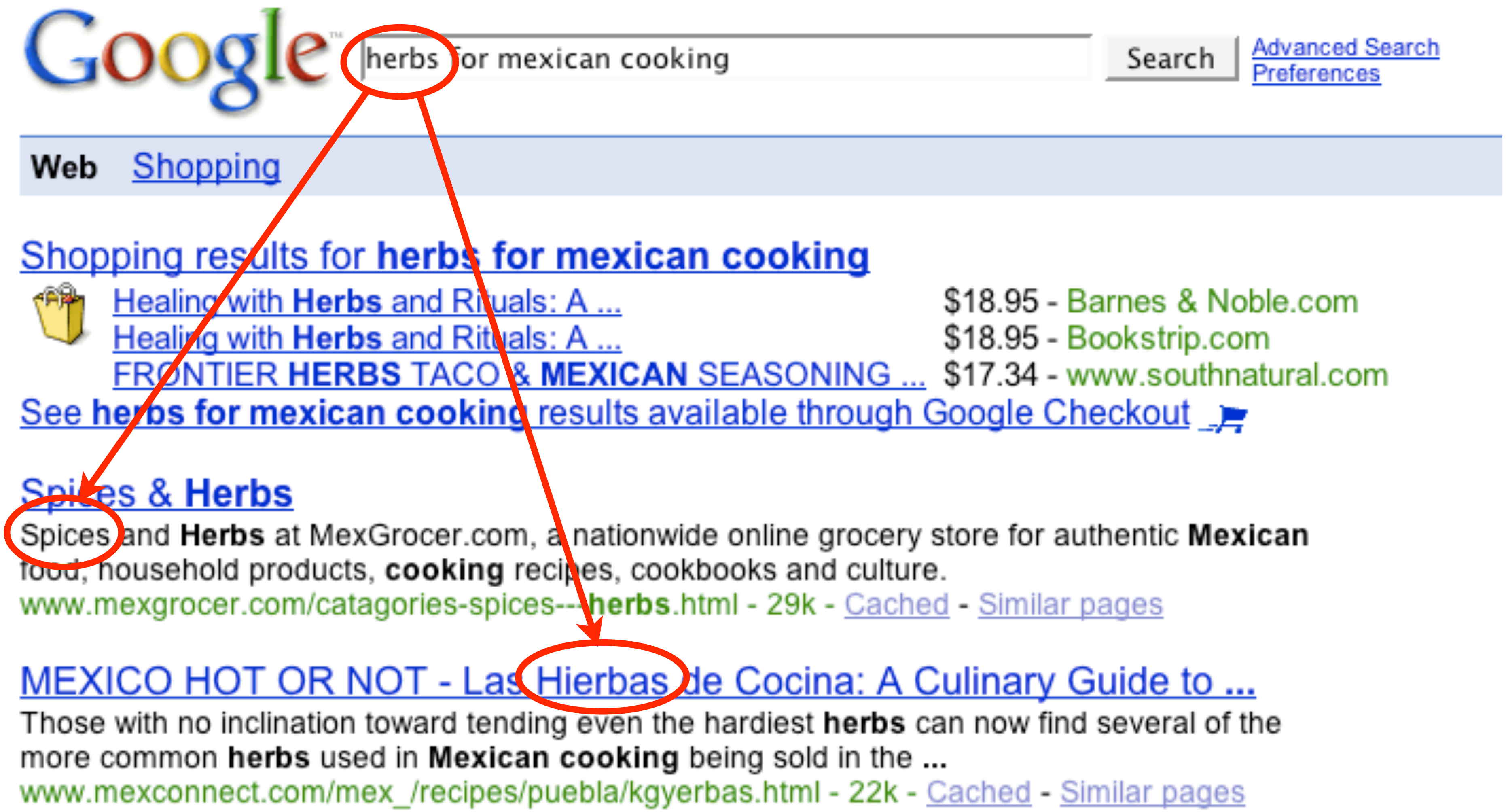
Meaning from Web Search

- *Term mismatches* between queries and documents in information retrieval
  - Query terms and relevant document terms differ
- *Query expansion:* add search terms known to occur in relevant documents
  - Increase recall
  - Decrease query term ambiguity

Query Expansion by Translation

- Training
  - *Align queries* and clicked *snippets*
  - Create translation tables
  - Train *n*-gram language model from query logs
- Use
  - "Translate" many queries
  - Extract and store table of "translated" terms in context

From query to snippets



Ambiguity Resolution

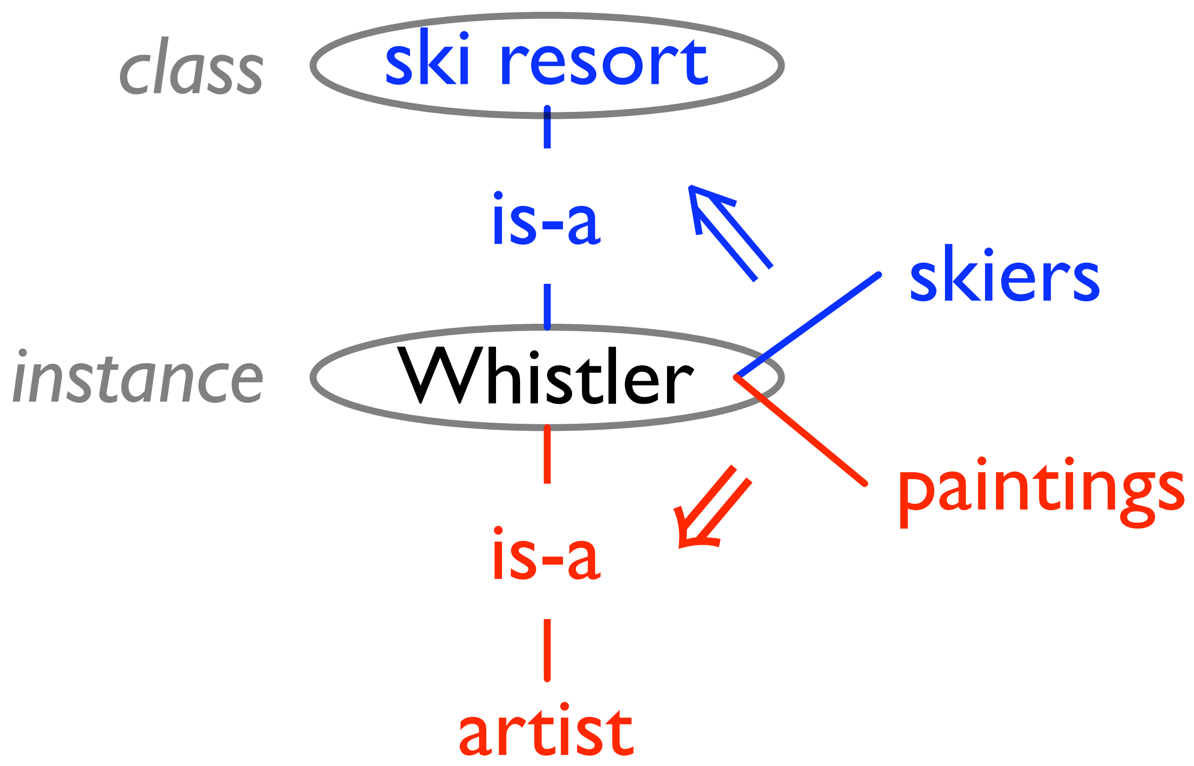
- 5-best phrase-level translations

(herbs,herbs)(for,for)(chronic,chronic)(heartburn, heartburn)
(herbs,herb)(for,for)(chronic,chronic)(heartburn, heartburn)
(herbs,remedies)(for,for)(chronic,chronic)(heartburn, heartburn)
(herbs,medicine)(for,for)(chronic,chronic)(heartburn, heartburn)
(herbs,supplements)(for,for)(chronic,chronic)(heartburn, heartburn)

(herbs,herbs)(for,for)(mexican,mexican)(cooking,cooking)
(herbs,herbs)(for,for))(cooking,cooking)(mexican,mexican
(herbs,herbs)(for,for)(mexican,mexican)(cooking,food)
(mexican,mexican)(herbs,herbs)(for,for)(cooking,cooking)
(herbs,spices)(for,for)(mexican,mexican)(cooking,cooking)

Beyond Parallel Corpora

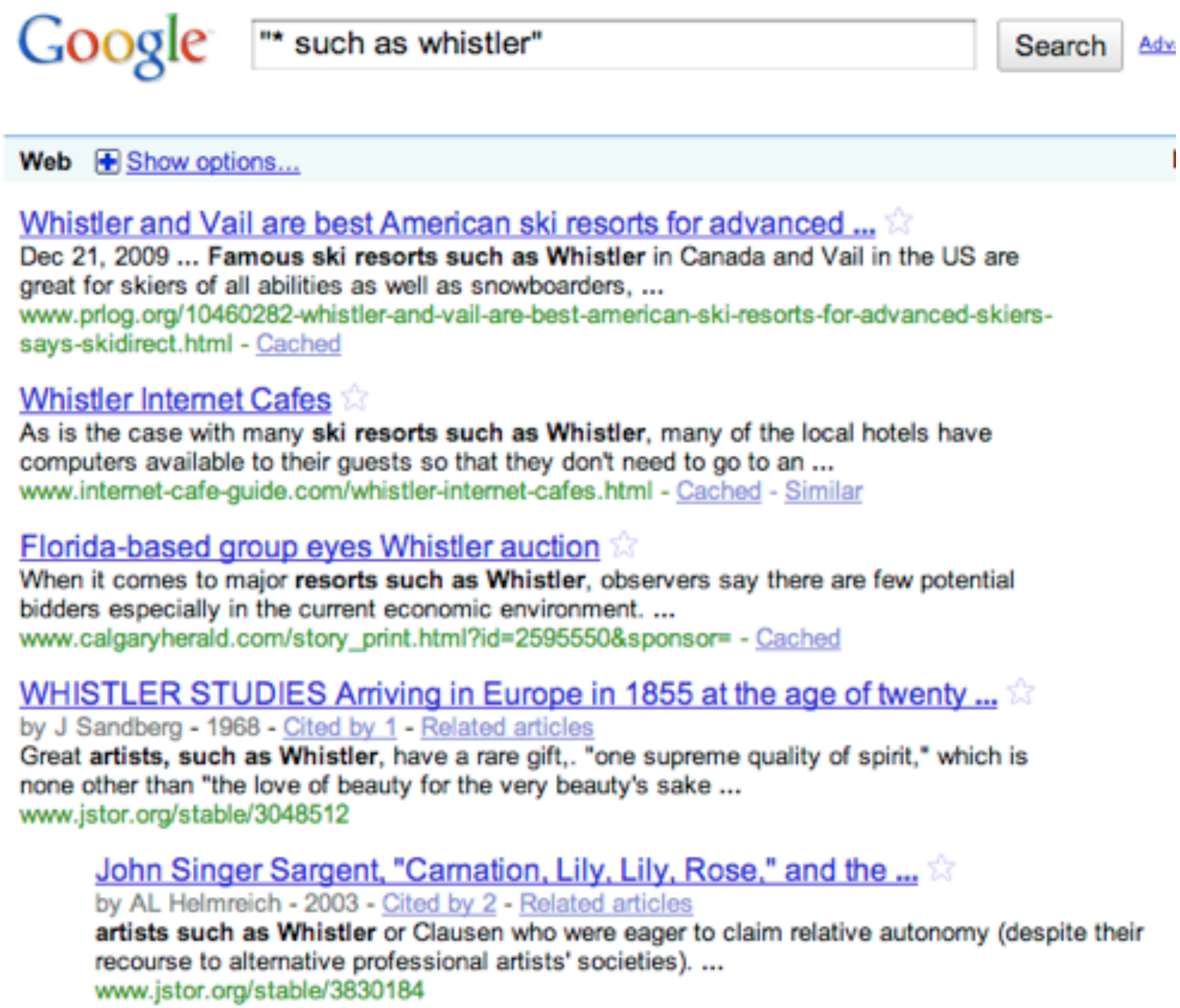
Reading the Web



- Elementary semantic inference
- First step: what are the possible classes for each instance?

Someone Told Us

- Text patterns (Hearst 92, Van Durme & Paşca 08)


Google "* such as whistler" Search Adv

Web Show options...

Whistler and Vail are best American ski resorts for advanced ...
Dec 21, 2009 ... **Famous ski resorts such as Whistler** in Canada and Vail in the US are great for skiers of all abilities as well as snowboarders, ...
www.prlog.org/10460282-whistler-and-vail-are-best-american-ski-resorts-for-advanced-skiers-says-skidirect.html - Cached

Whistler Internet Cafes
As is the case with many **ski resorts such as Whistler**, many of the local hotels have computers available to their guests so that they don't need to go to an ...
www.internet-cafe-guide.com/whistler-internet-cafes.html - Cached - Similar

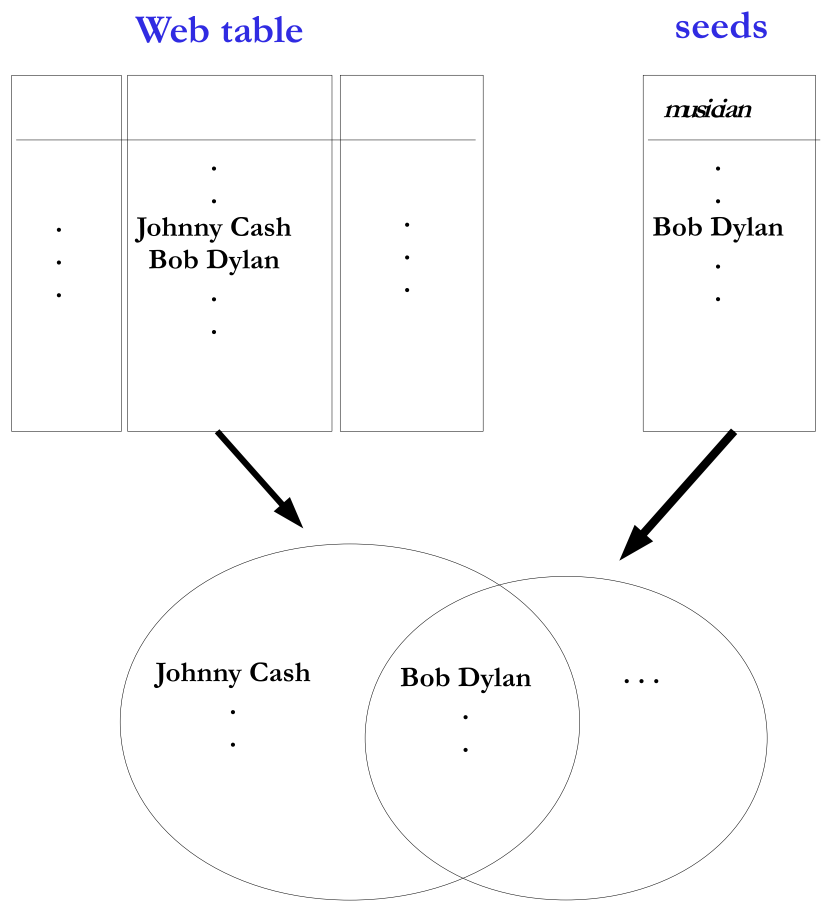
Florida-based group eyes Whistler auction
When it comes to major **resorts such as Whistler**, observers say there are few potential bidders especially in the current economic environment. ...
www.calgaryherald.com/story_print.html?id=2595550&sponsor= - Cached

WHISTLER STUDIES Arriving in Europe in 1855 at the age of twenty ...
by J Sandberg - 1968 - Cited by 1 - Related articles
Great **artists, such as Whistler**, have a rare gift,. "one supreme quality of spirit," which is none other than "the love of beauty for the very beauty's sake ...
www.jstor.org/stable/3048512

John Singer Sargent, "Carnation, Lily, Lily, Rose," and the ...
by AL Helmreich - 2003 - Cited by 2 - Related articles
artists such as Whistler or Clausen who were eager to claim relative autonomy (despite their recourse to alternative professional artists' societies). ...
www.jstor.org/stable/3830184

Informative Co-occurrences

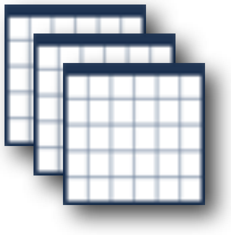
- WebTables (Cafarella *et al.* 08)
  - 154M HTML tables from Web pages
  - Cluster instances in table columns



Combining Information Sources

- *Bootstrapping*:
  - Seed set of (instance, class) pairs
  - Compute instance similarity from additional sources
  - Use similarity to infer new (instance, class) pairs
- *Approach*: label propagation in a graph
  - instance nodes
  - cluster/class nodes
  - bipartite structure

Label Propagation



Label Propagation

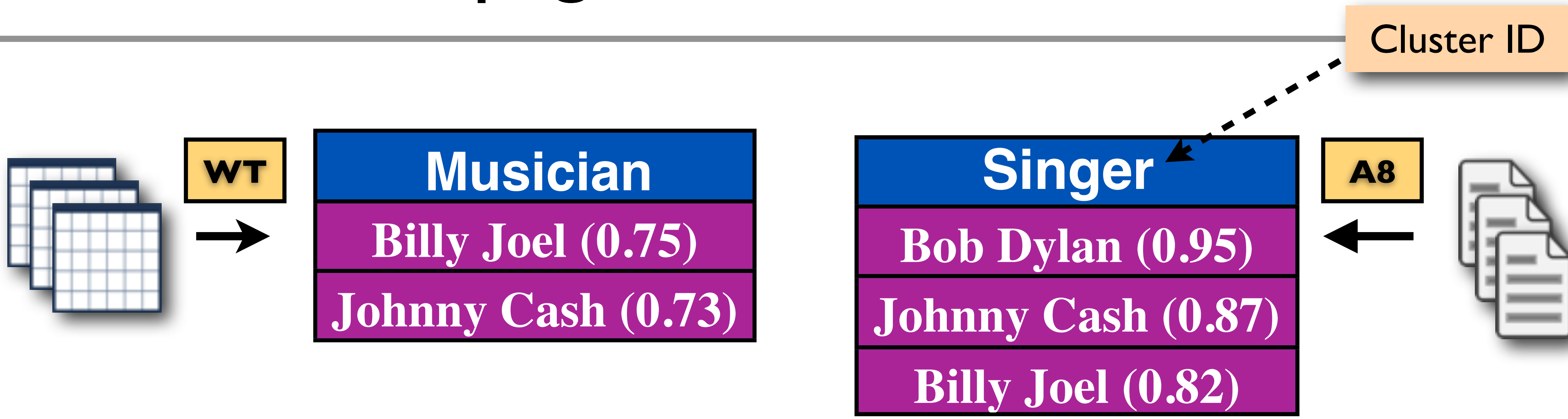
WT →

Musician
Billy Joel (0.75)
Johnny Cash (0.73)

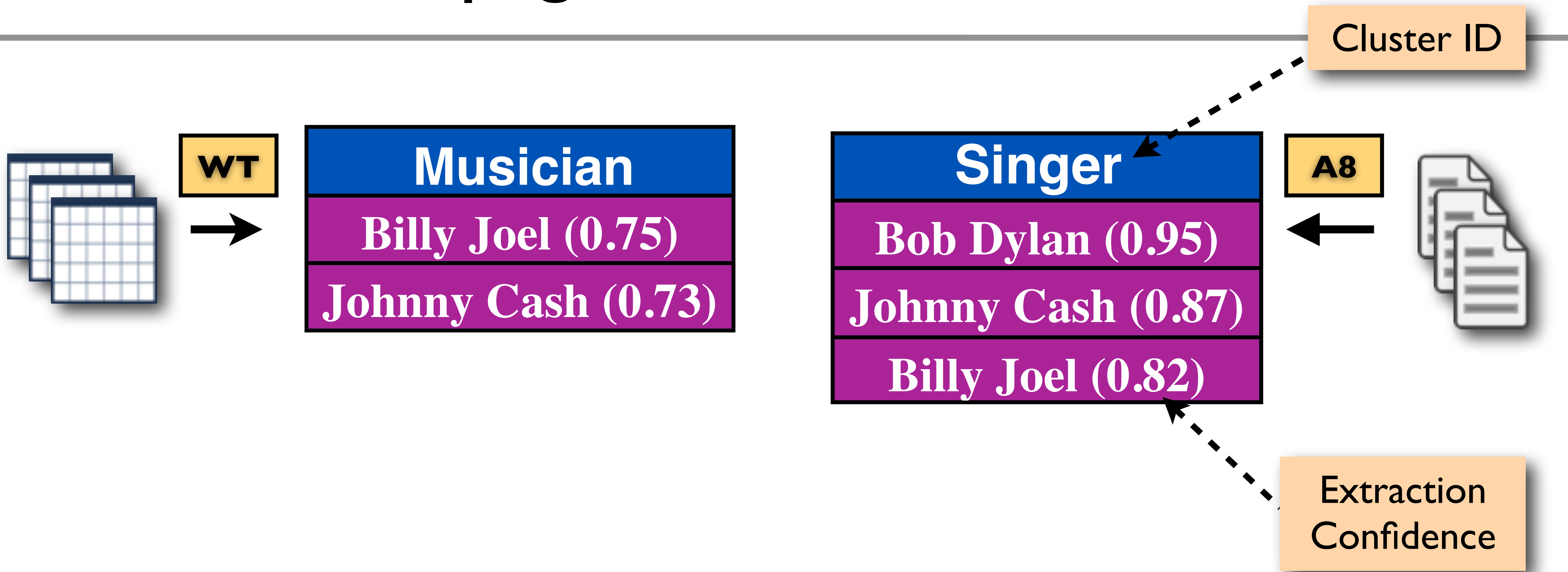
Singer
Bob Dylan (0.95)
Johnny Cash (0.87)
Billy Joel (0.82)

← A8

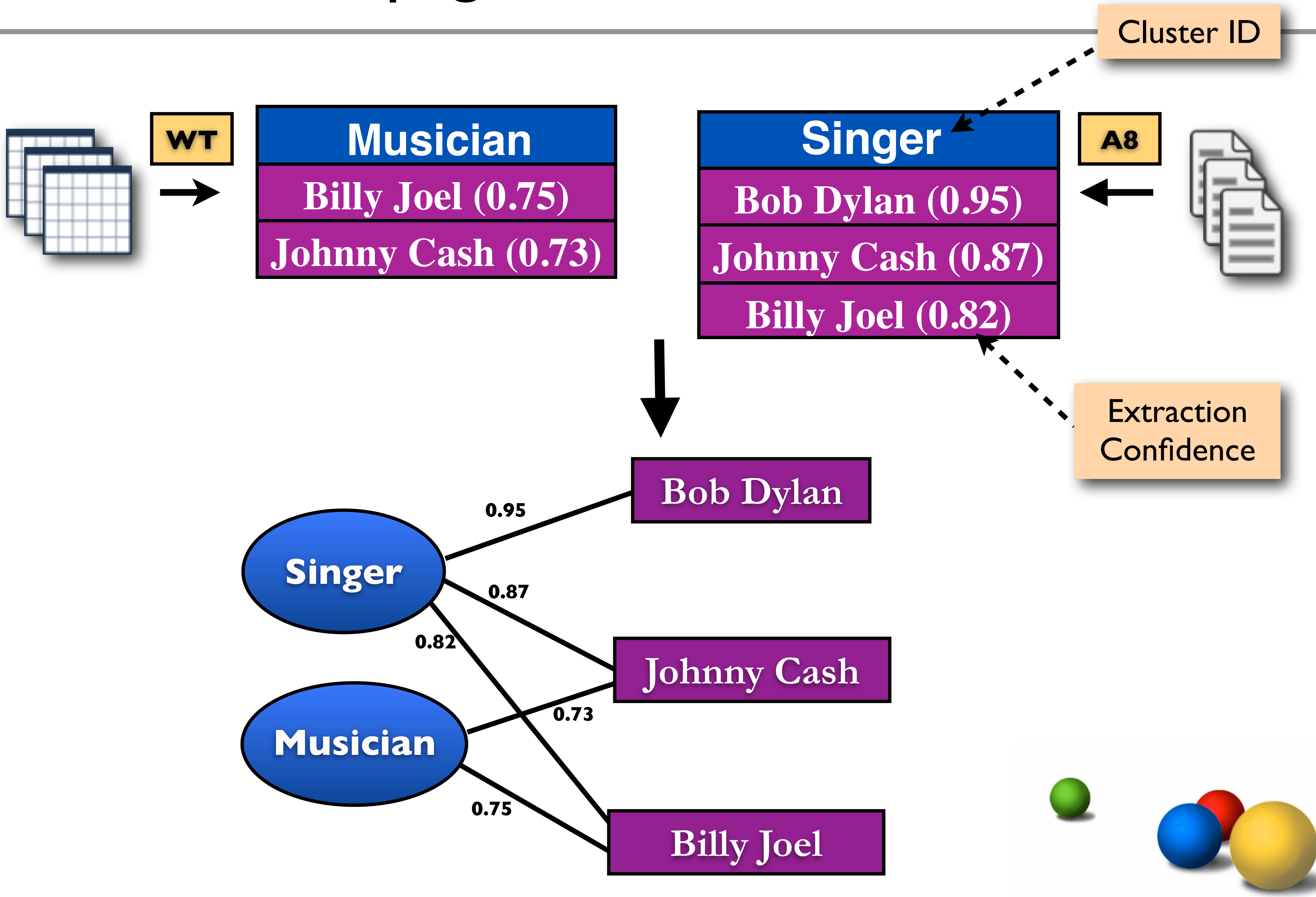
Label Propagation



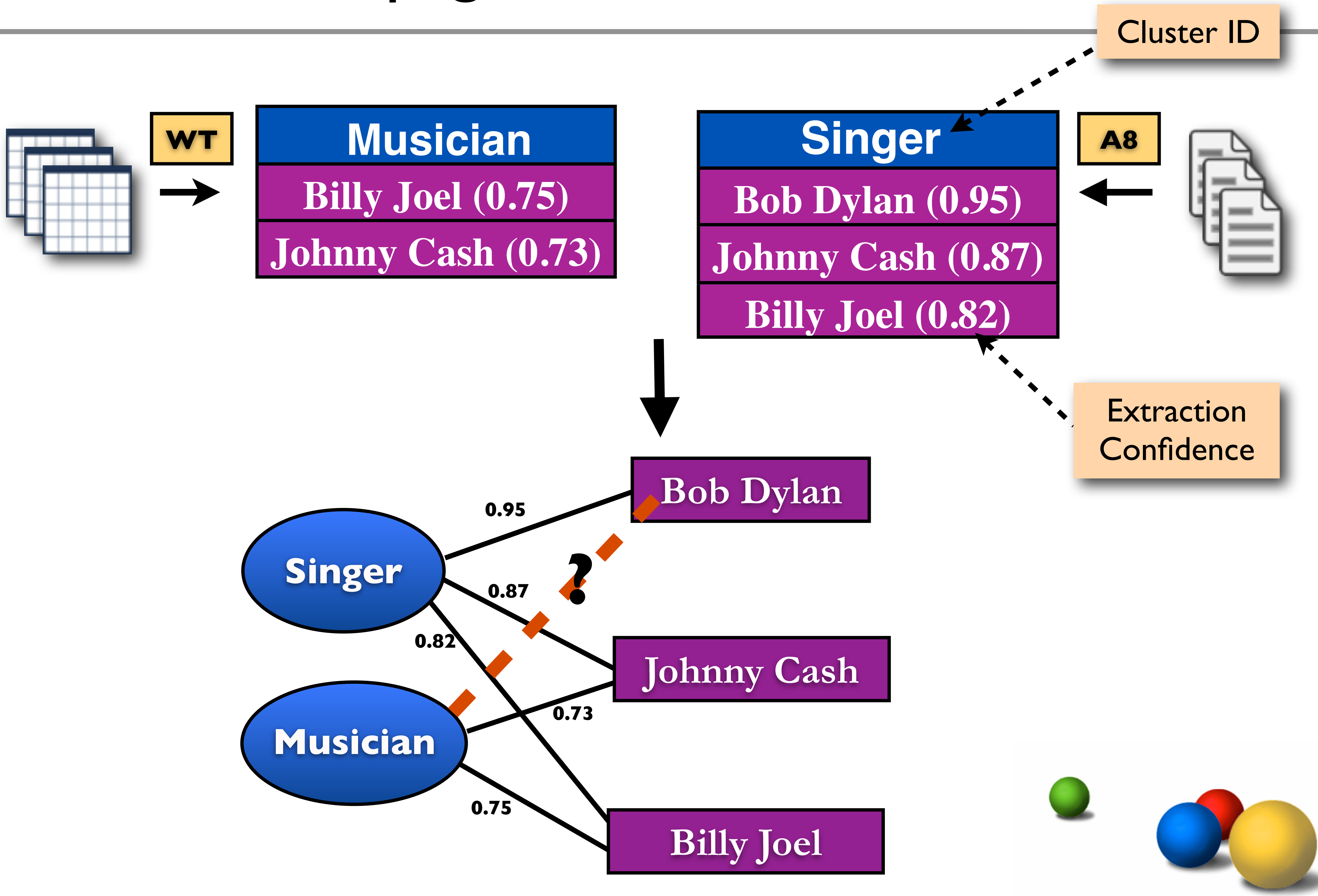
Label Propagation



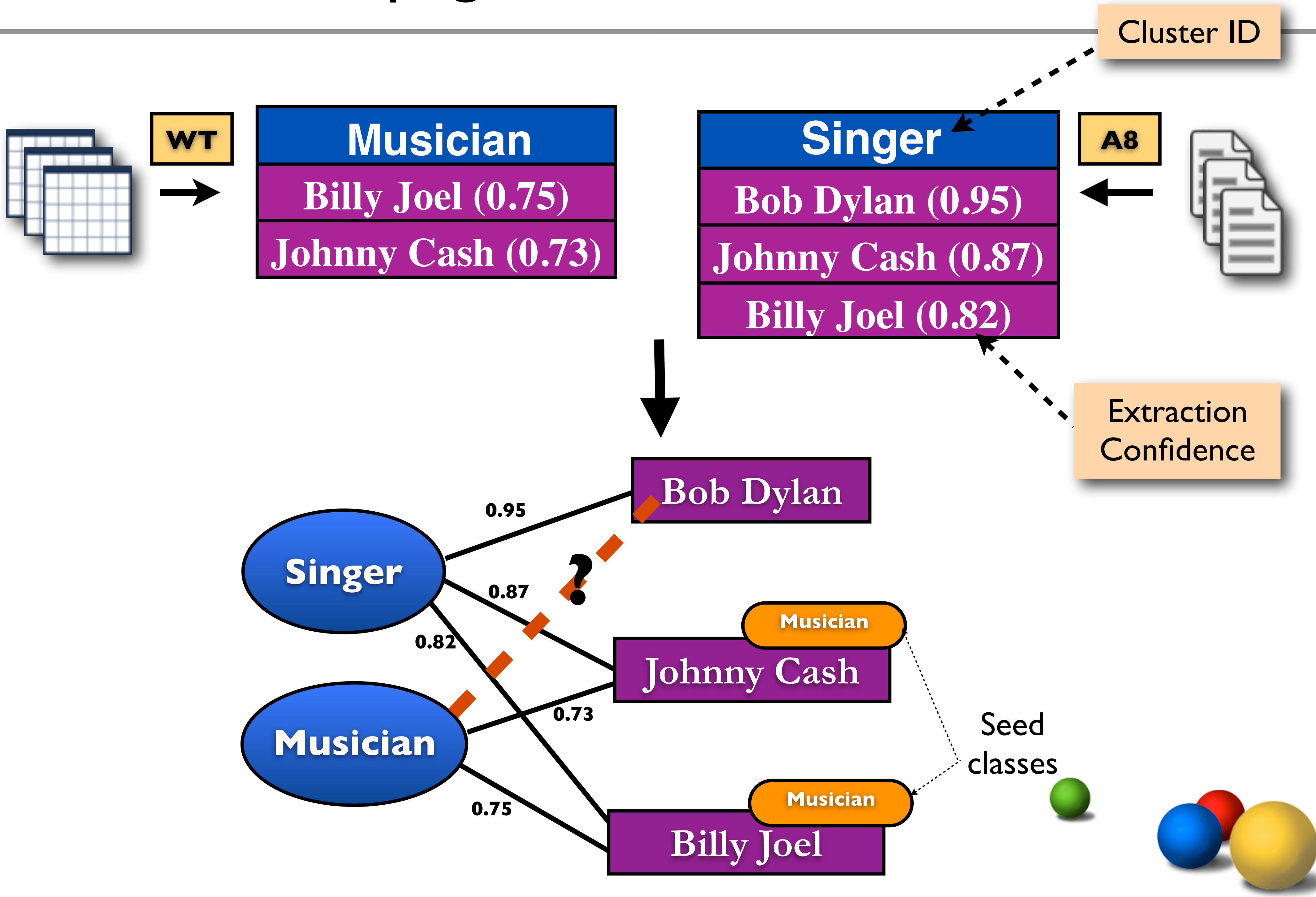
Label Propagation



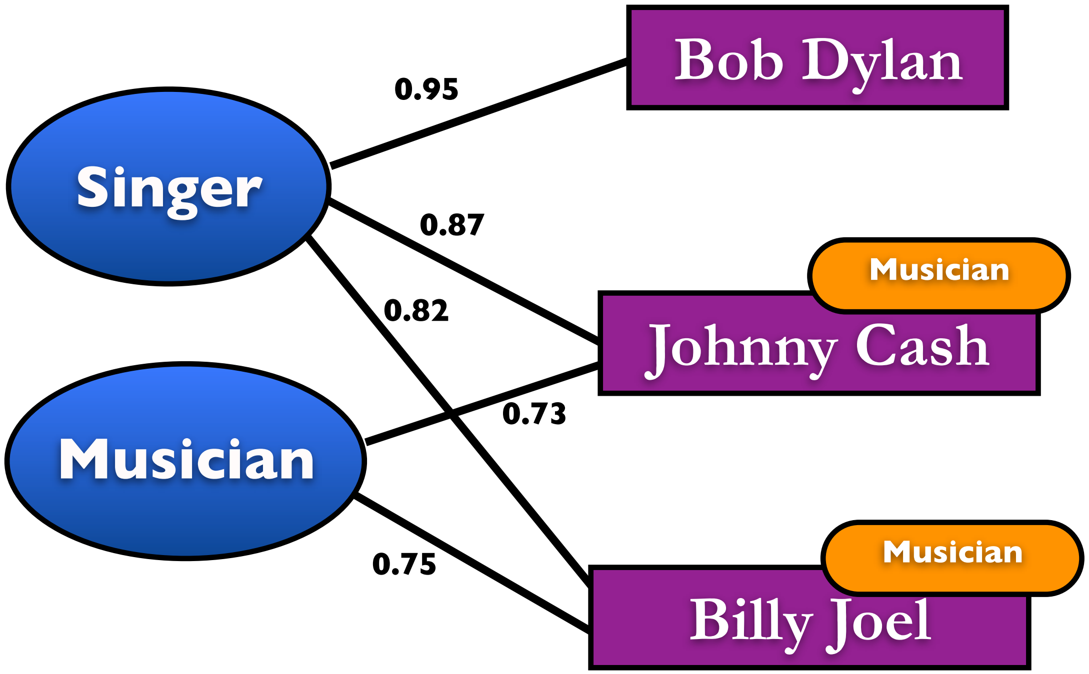
Label Propagation



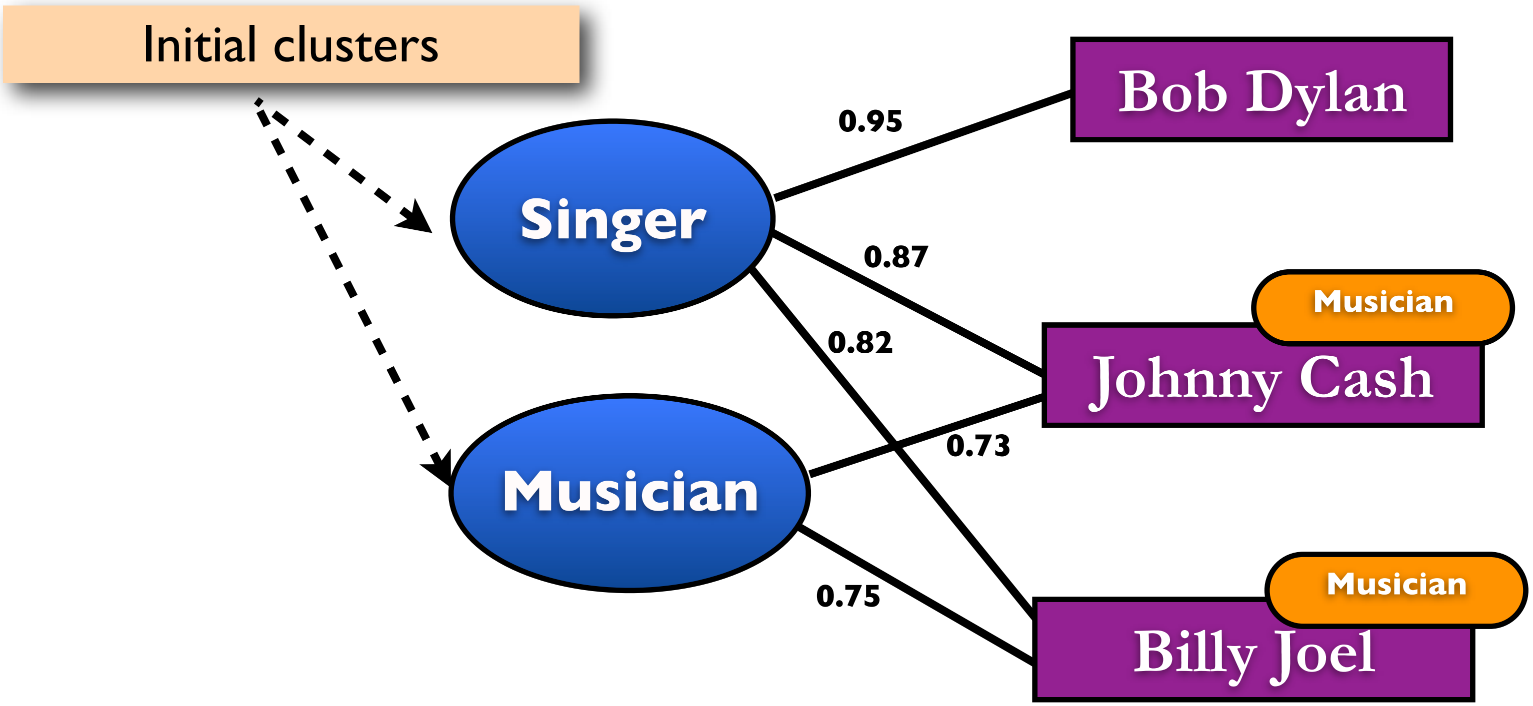
Label Propagation



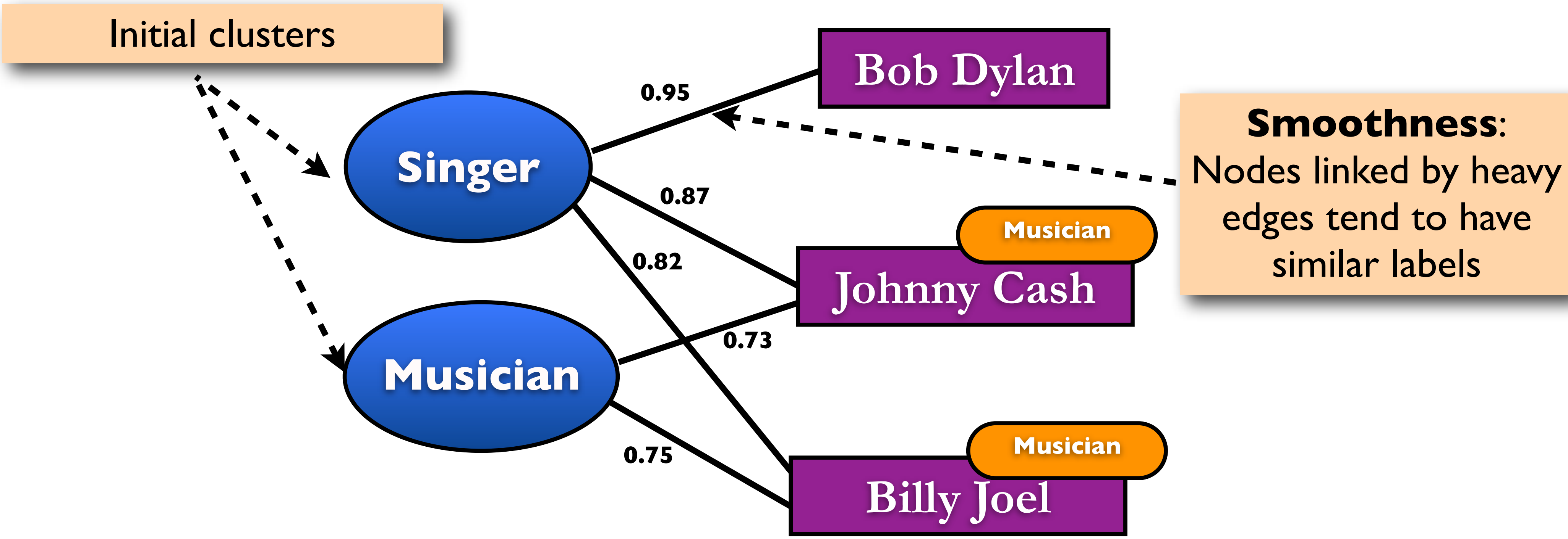
Classes, Clusters, Instances



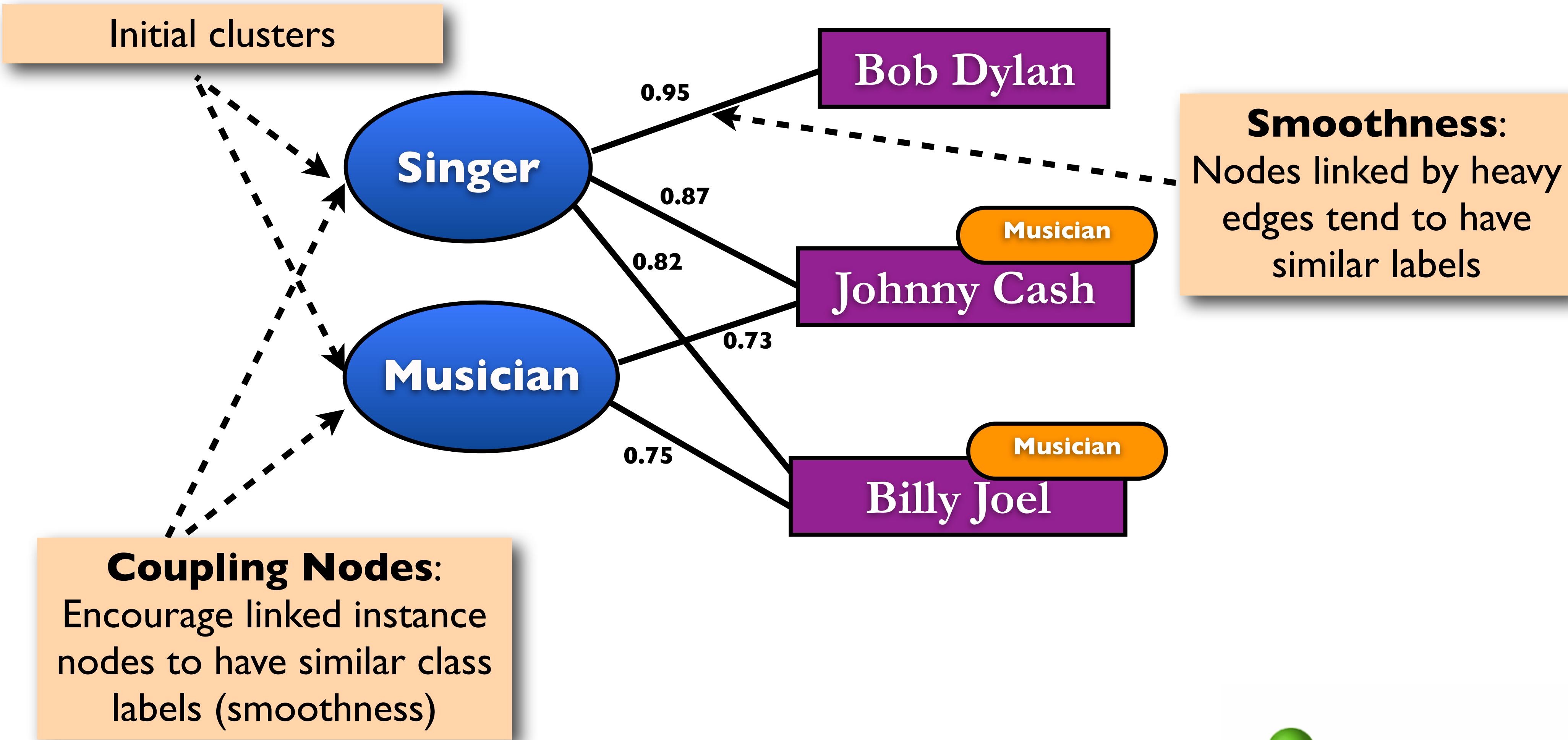
Classes, Clusters, Instances



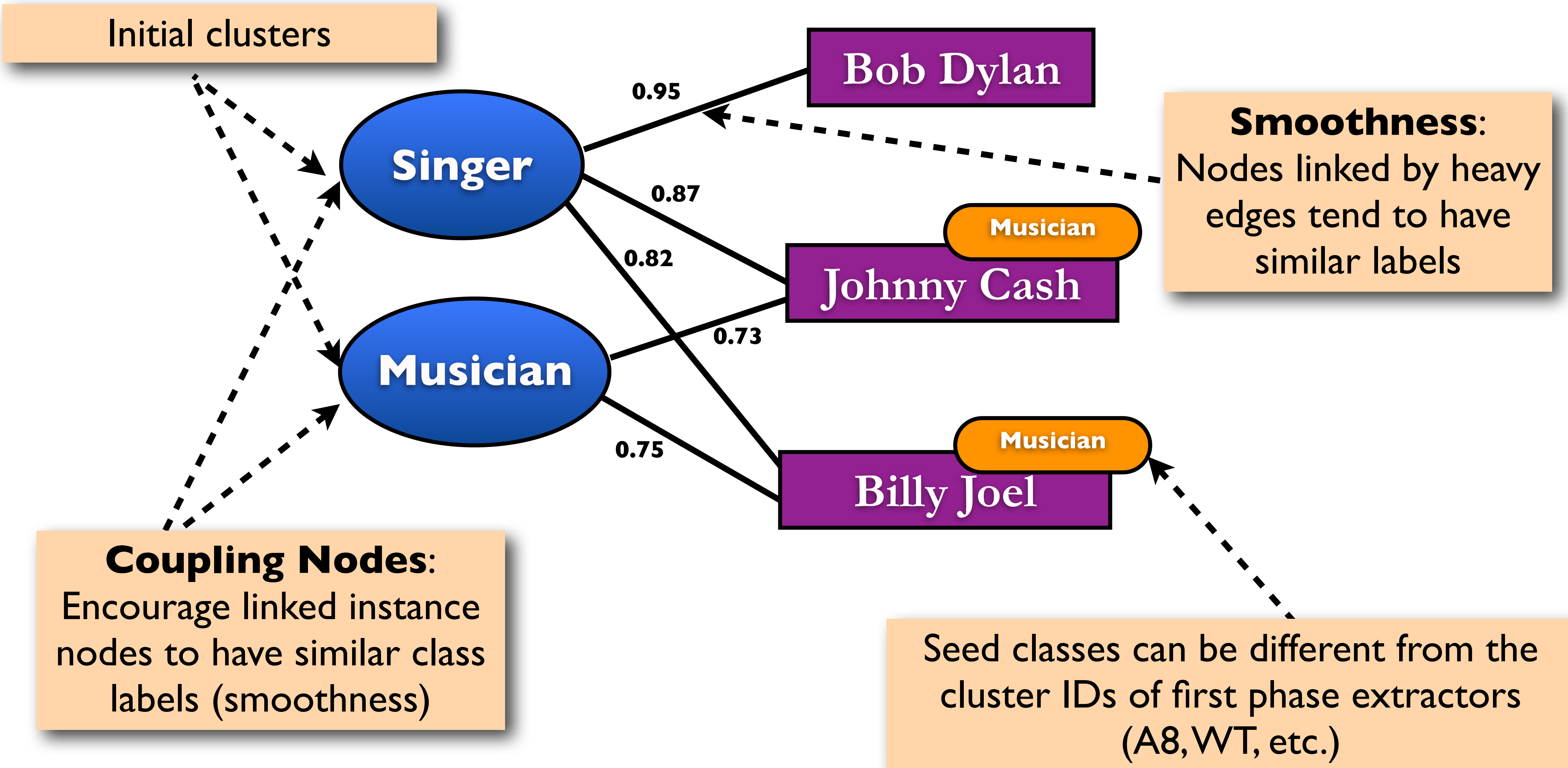
Classes, Clusters, Instances



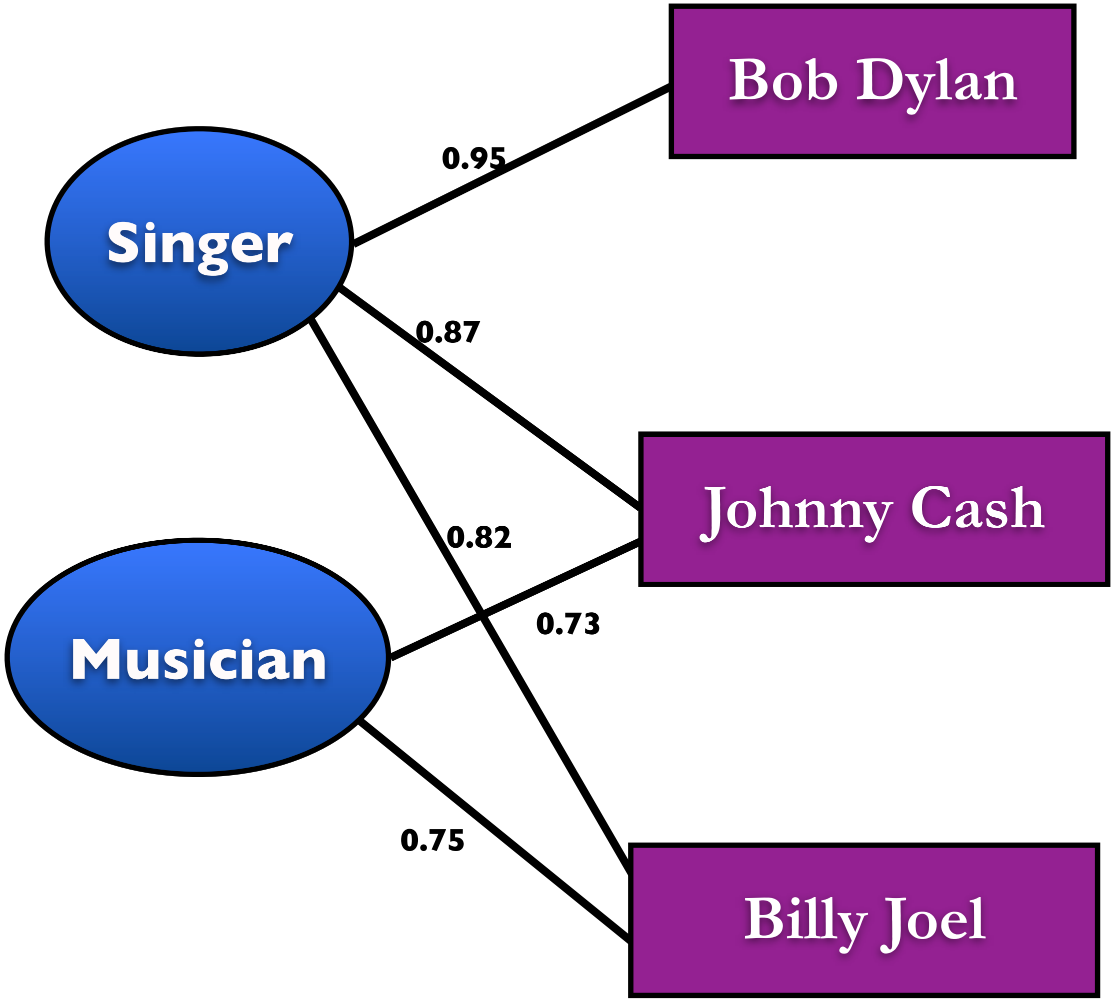
Classes, Clusters, Instances



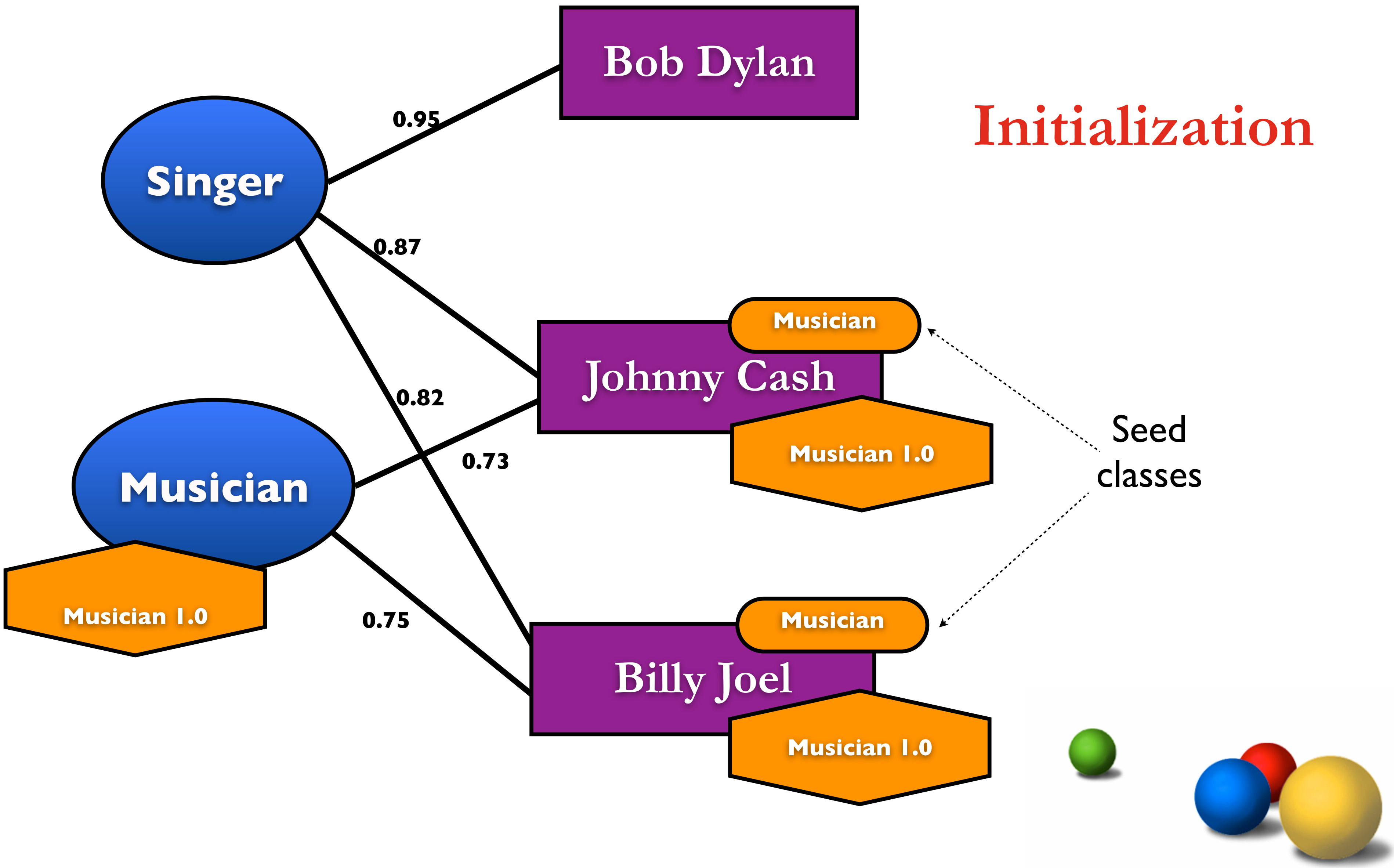
Classes, Clusters, Instances



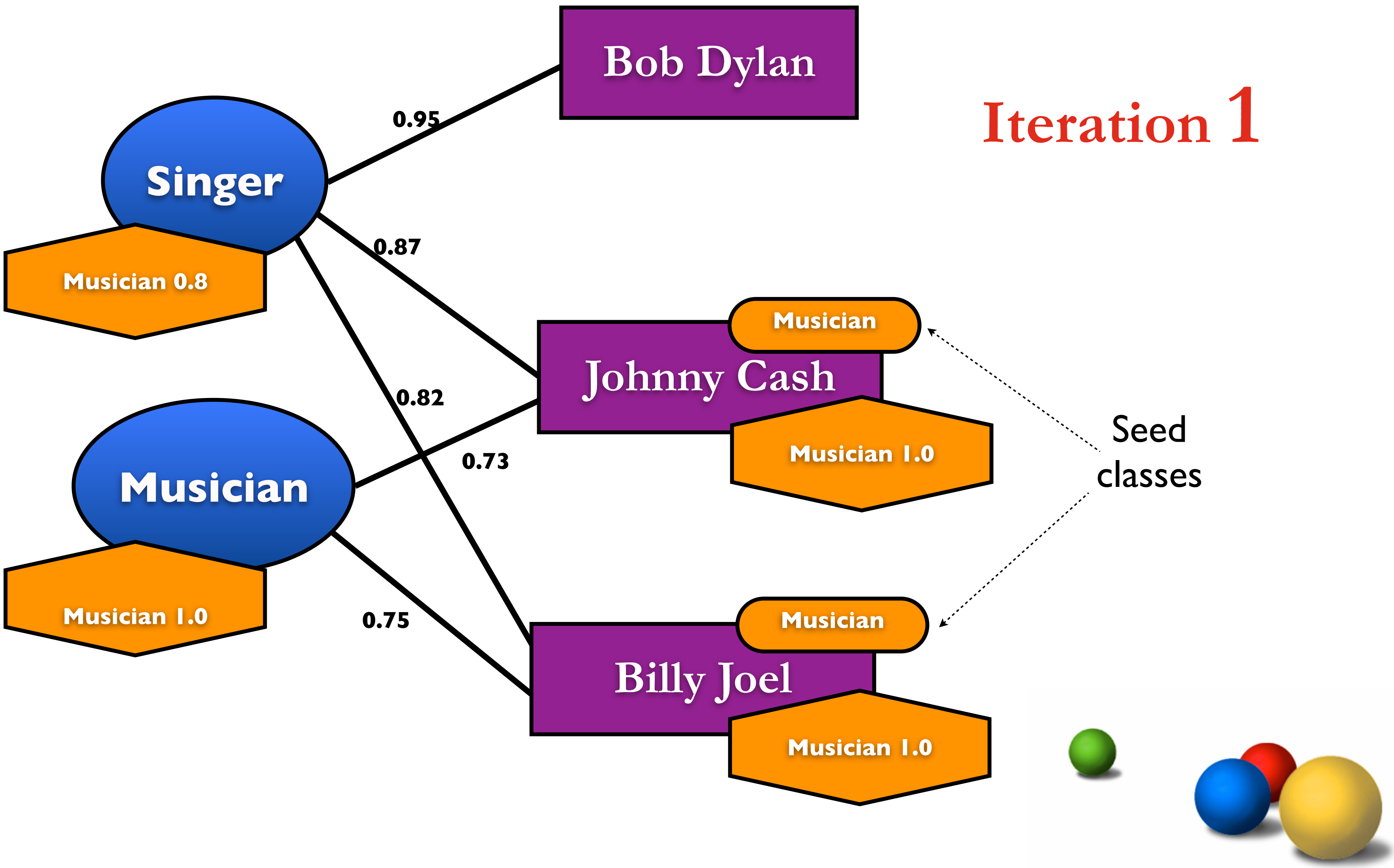
Label Propagation



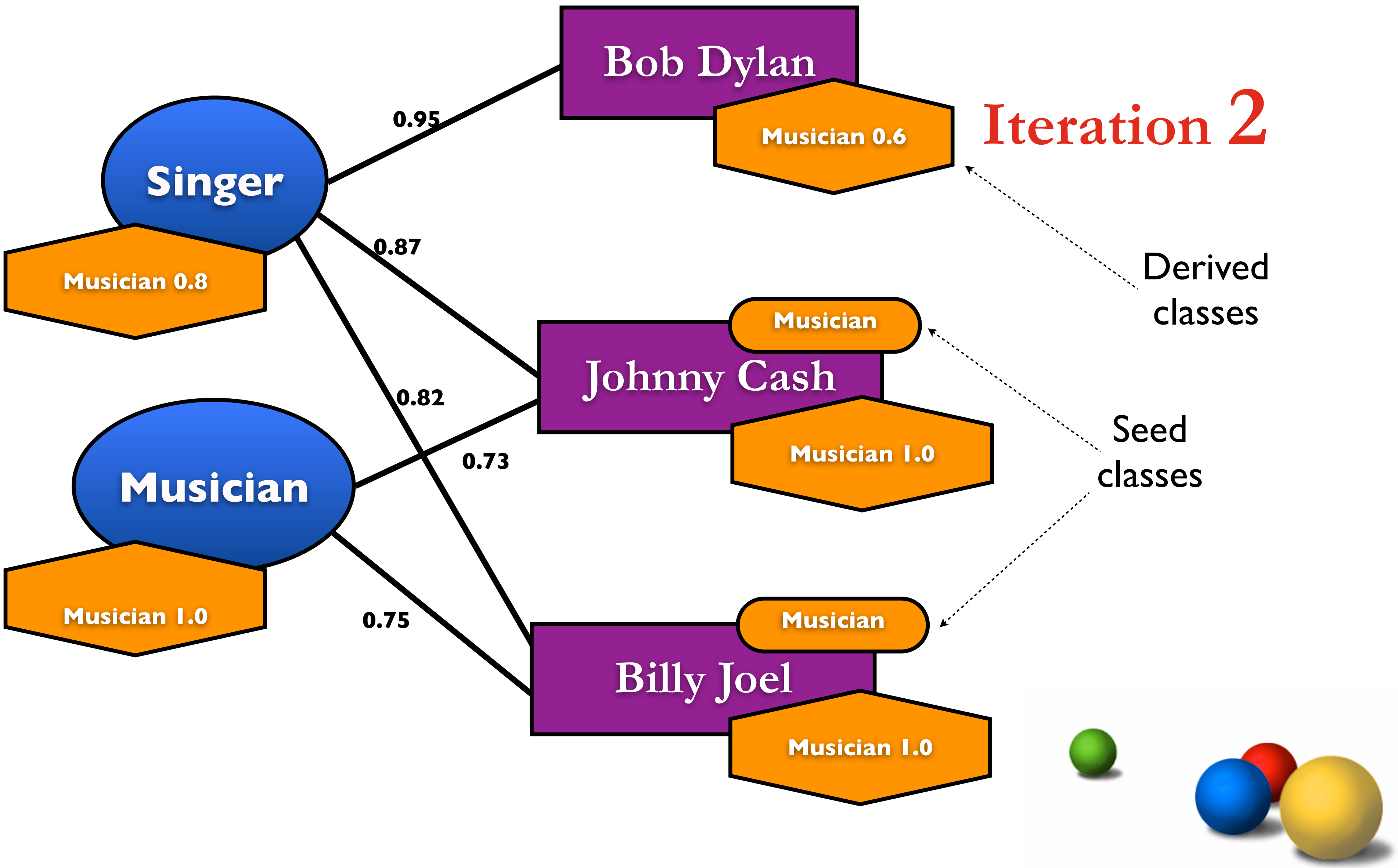
Label Propagation



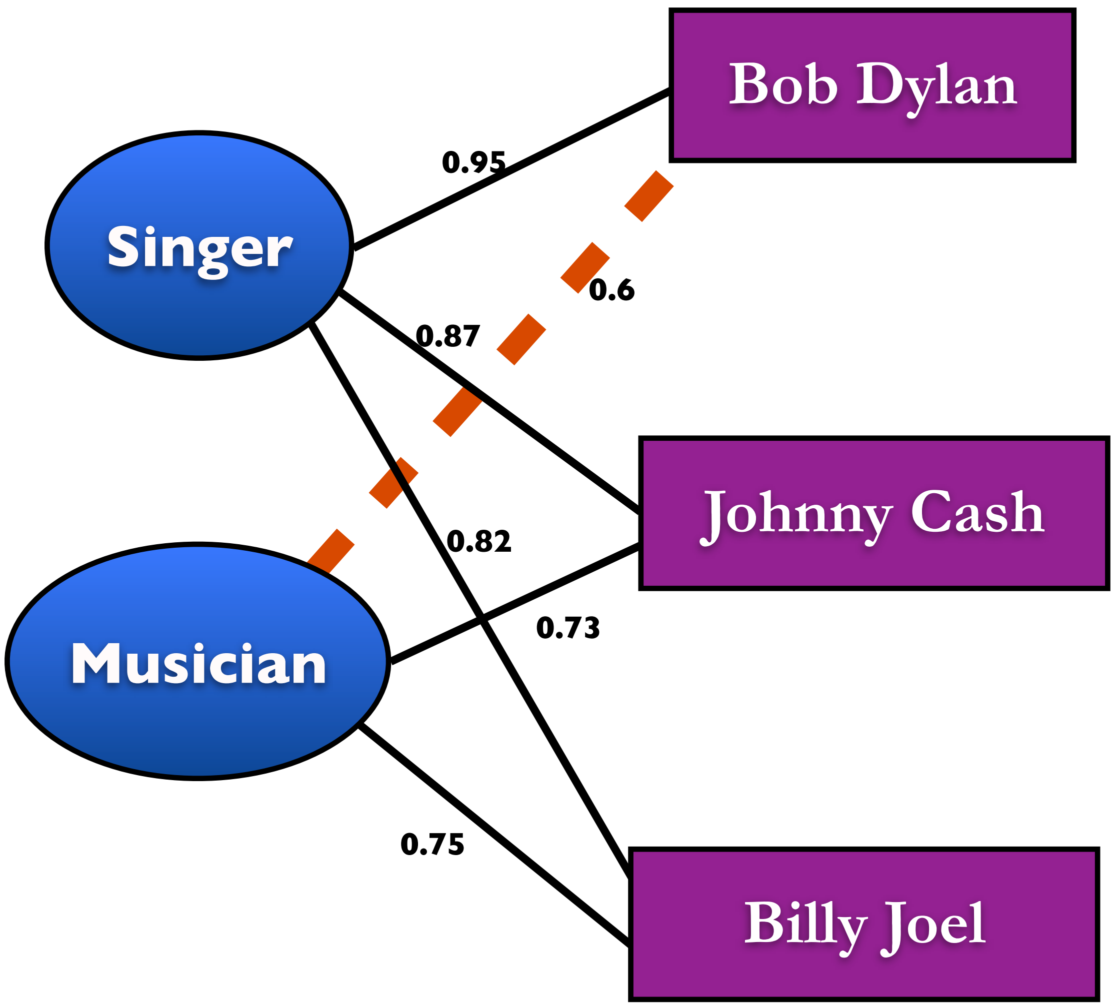
Label Propagation



Label Propagation



Label Propagation



Iteration 2

Class Assignment for Given Instances

Class Assignment for Given Instances

- A8
- Propagation
- WebTables

Class Assignment for Given Instances

924k (class, instance) pairs extracted from 100M web documents

A8
Propagation
WebTables

74M (class, instance) pairs extracted from WebTables

Class Assignment for Given Instances

A8
Propagation
WebTables

Graph with 1.4M nodes, 75M edges

Class Assignment for Given Instances

$$\text{MRR} = \frac{1}{|\text{test-set}|} \sum_{v \in \text{test-set}} \frac{1}{\text{rank}_v(\text{class}(v))}$$

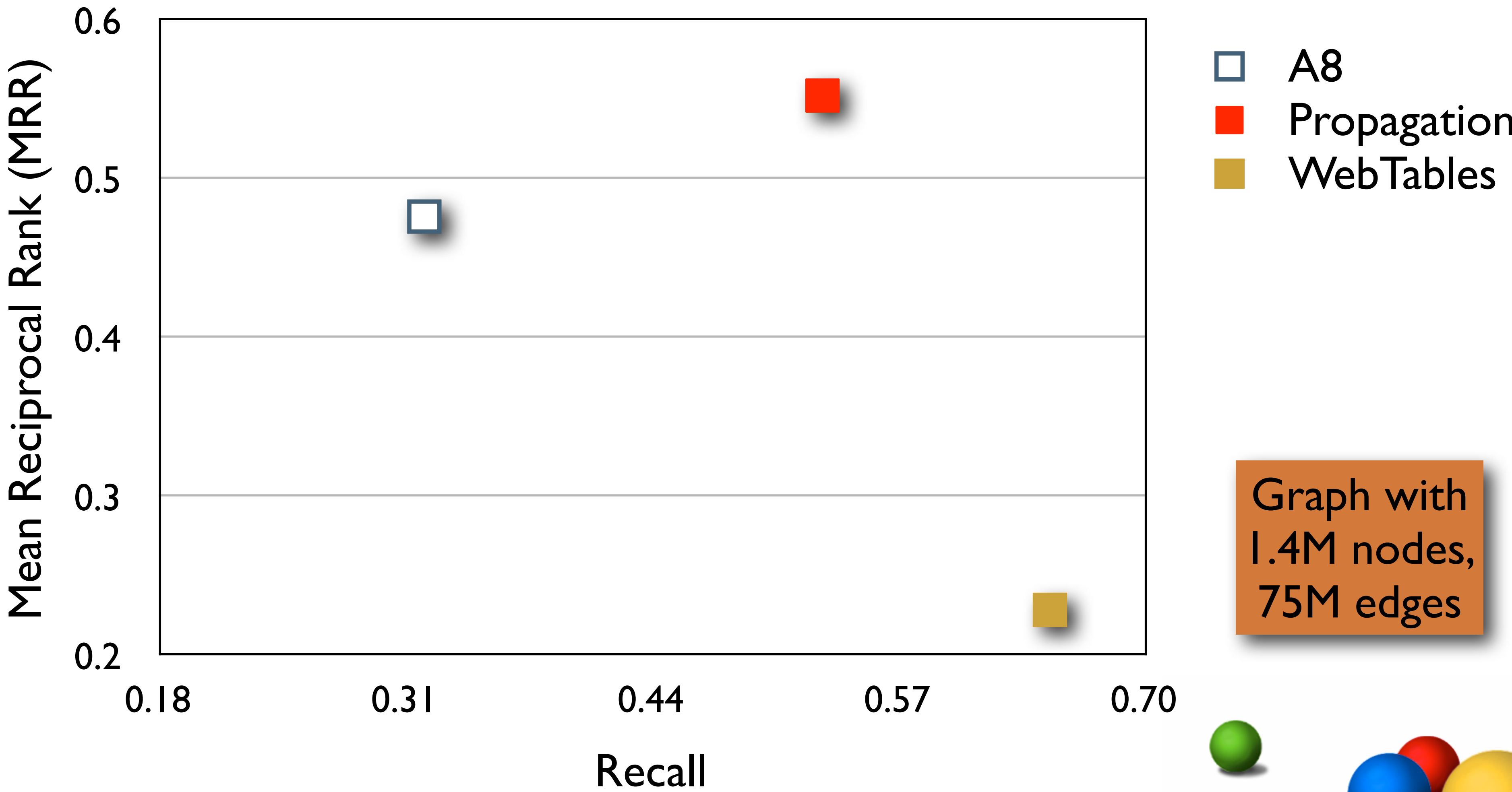
- A8
- Propagation
- WebTables

Graph with 1.4M nodes, 75M edges

Class Assignment for Given Instances

$$\mathrm{MRR} = \frac{1}{|\text{test-set}|} \sum_{v \in \text{test-set}} \frac{1}{\mathrm{rank}_v(\mathrm{class}(v))}$$

Evaluation against Wordnet excerpt (38 classes, 8910 instances)



Graph with 1.4M nodes, 75M edges

Finding More Good Instances

- Instances found solely by label propagation:

Class	Precision at 100 (non-A8 extractions)
Book Publishers	87.36
Federal Agencies	29.89
NFL Players	94.95
Scientific Journals	90.82
Mammal Species	84.27

Scientific Journals	Journal of Physics, Nature, Structural and Molecular Biology, Sciences Sociales et santé, Kidney and Blood Pressure Research, American Journal of Physiology–Cell Physiology
NFL Players	Tony Gonzales, Thabiti Davis, Taylor Stubblefield, Ron Dixon, Rodney Hannah
Book Publishers	Small Night Shade Books, House of Anansi Press, Highwater Books, Distributed Art Publishers, Copper Canyon Press

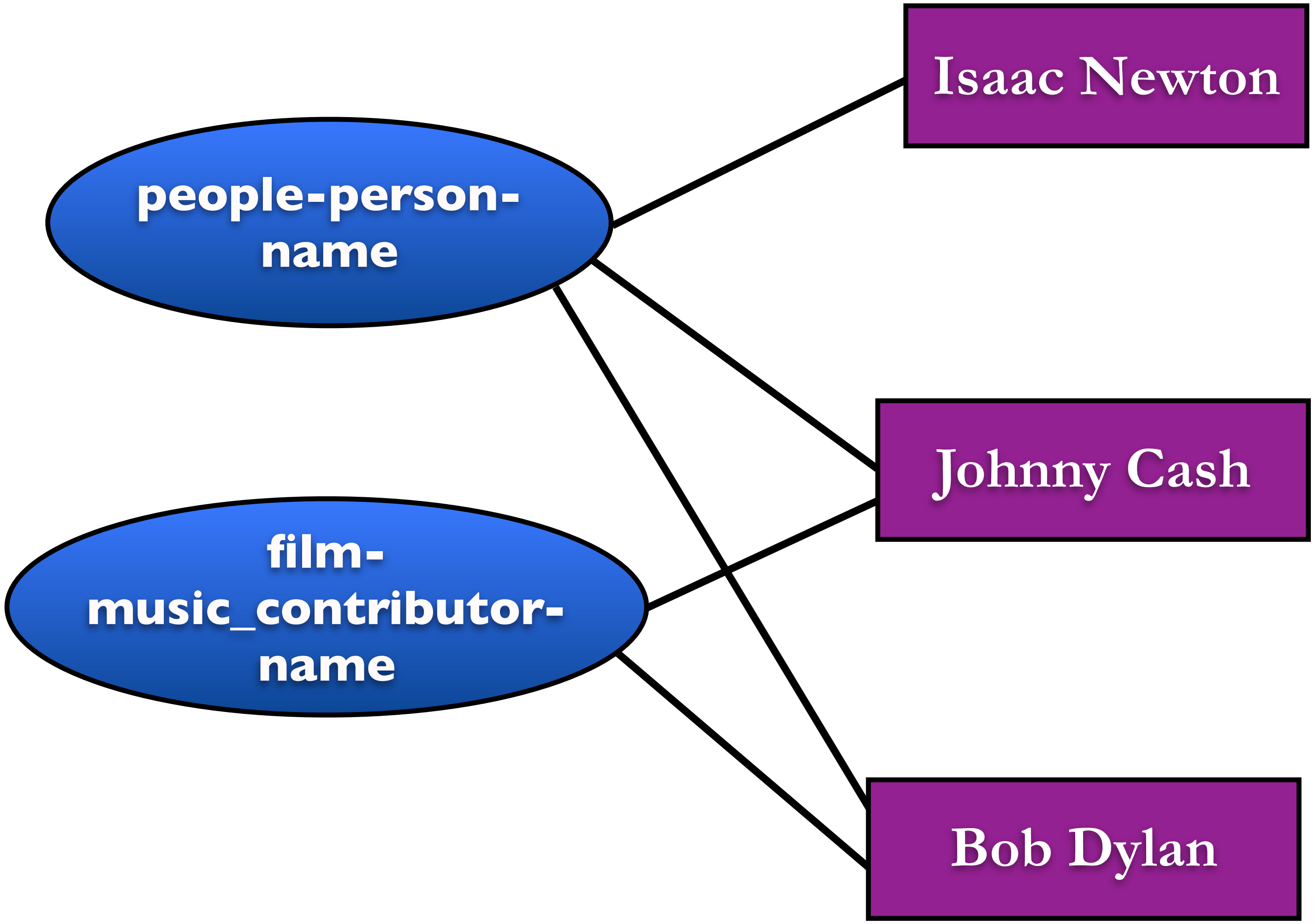
Discovering Similar Classes

- Label propagation by-product:

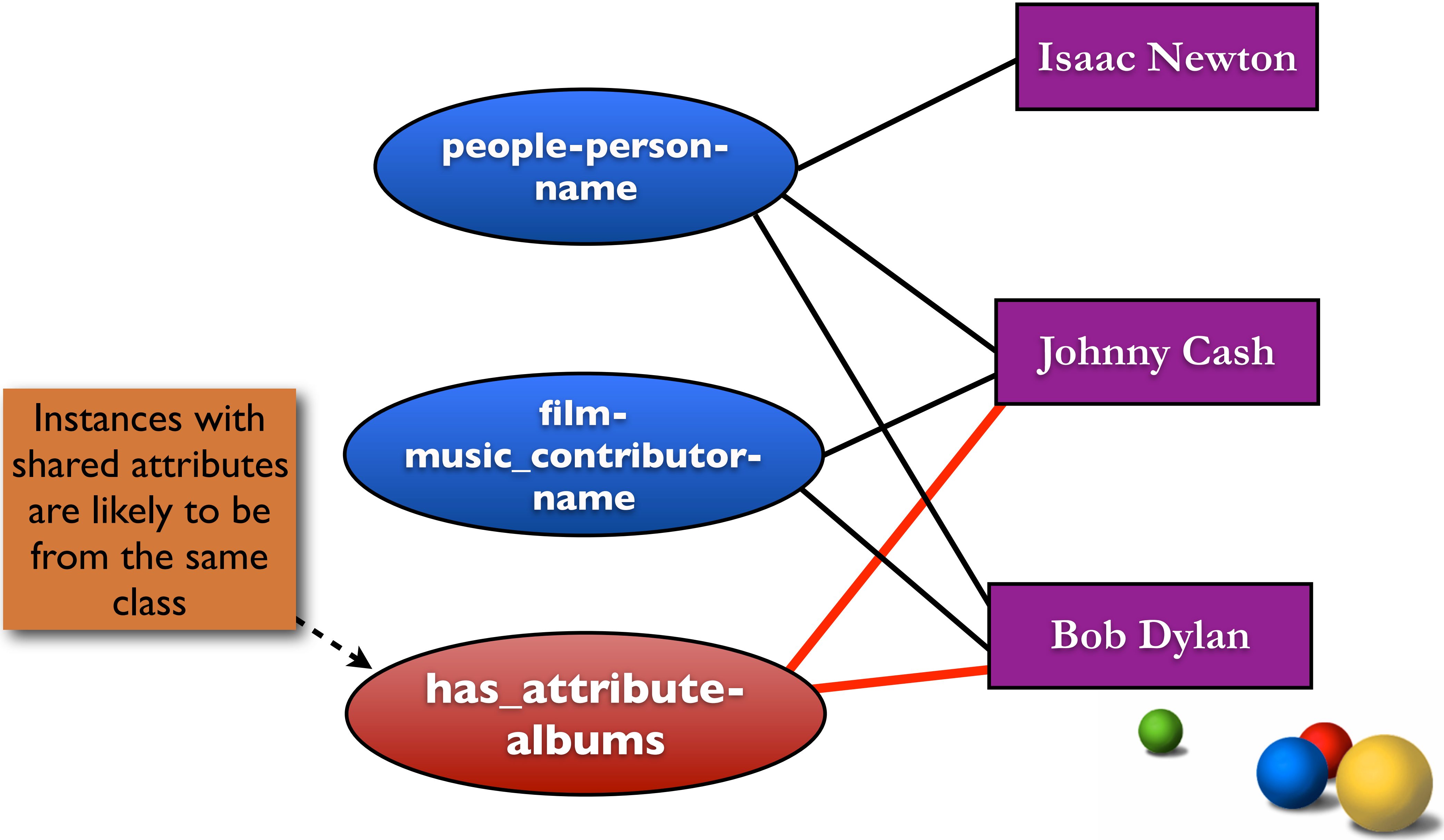
Seed Class	Non-Seed Class Labels Discovered
Book Publishers	small presses, journal publishers, educational publishers, academic publishers, commercial publishers
NFL Players	sports figures, football greats, football players, backs, quarterbacks
Scientific Journals	prestigious journals, peer-reviewed journals, refereed journals, scholarly journals, academic journals

Semantic Constraints for Better Classes

Semantic Constraints for Better Classes



Semantic Constraints for Better Classes

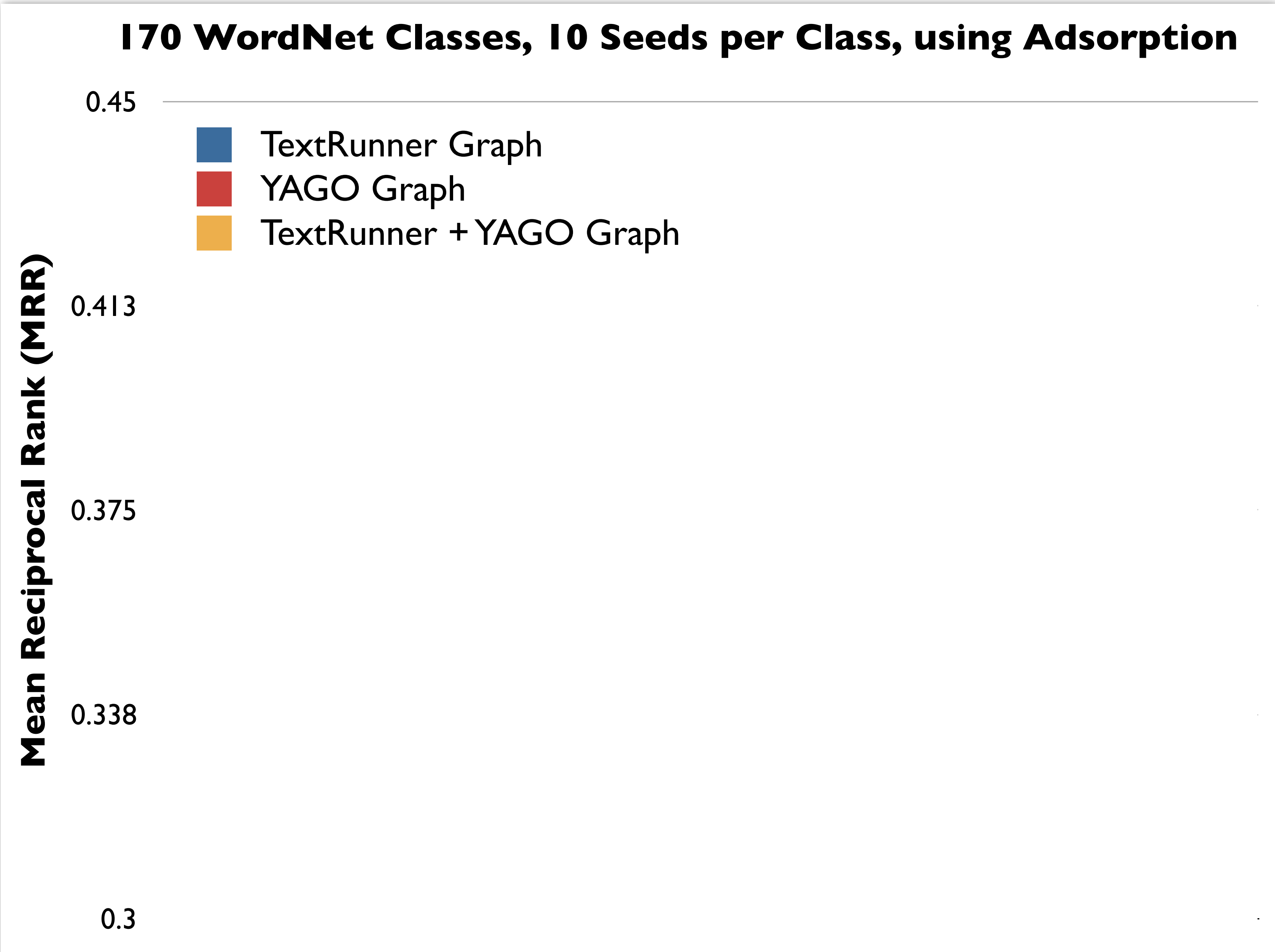


Experiments with Public Sources

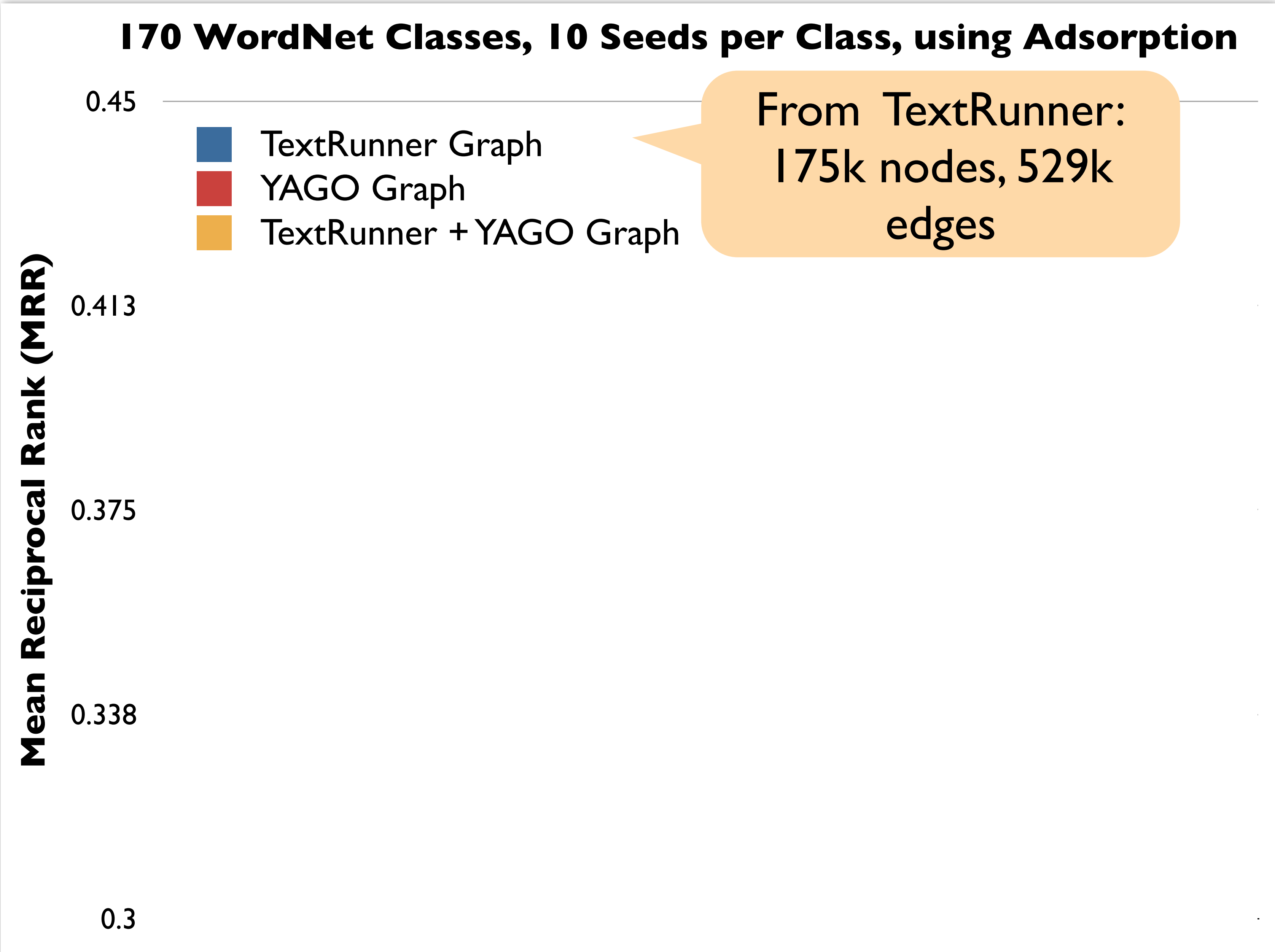
- Make Talukdar *et al.* results easy to reproduce and extend
  - Freebase: multiple-sourced relational tables
  - Pantel *et al.* 09 gold-standard hypernyms
  - TextRunner (U. of Washington): hypernyms from open-domain extraction
  - YAGO (Suchanek *et al.*, 07): entity-attribute knowledge base curated from Wikipedia and Wordnet

Better Classes with YAGO Attributes

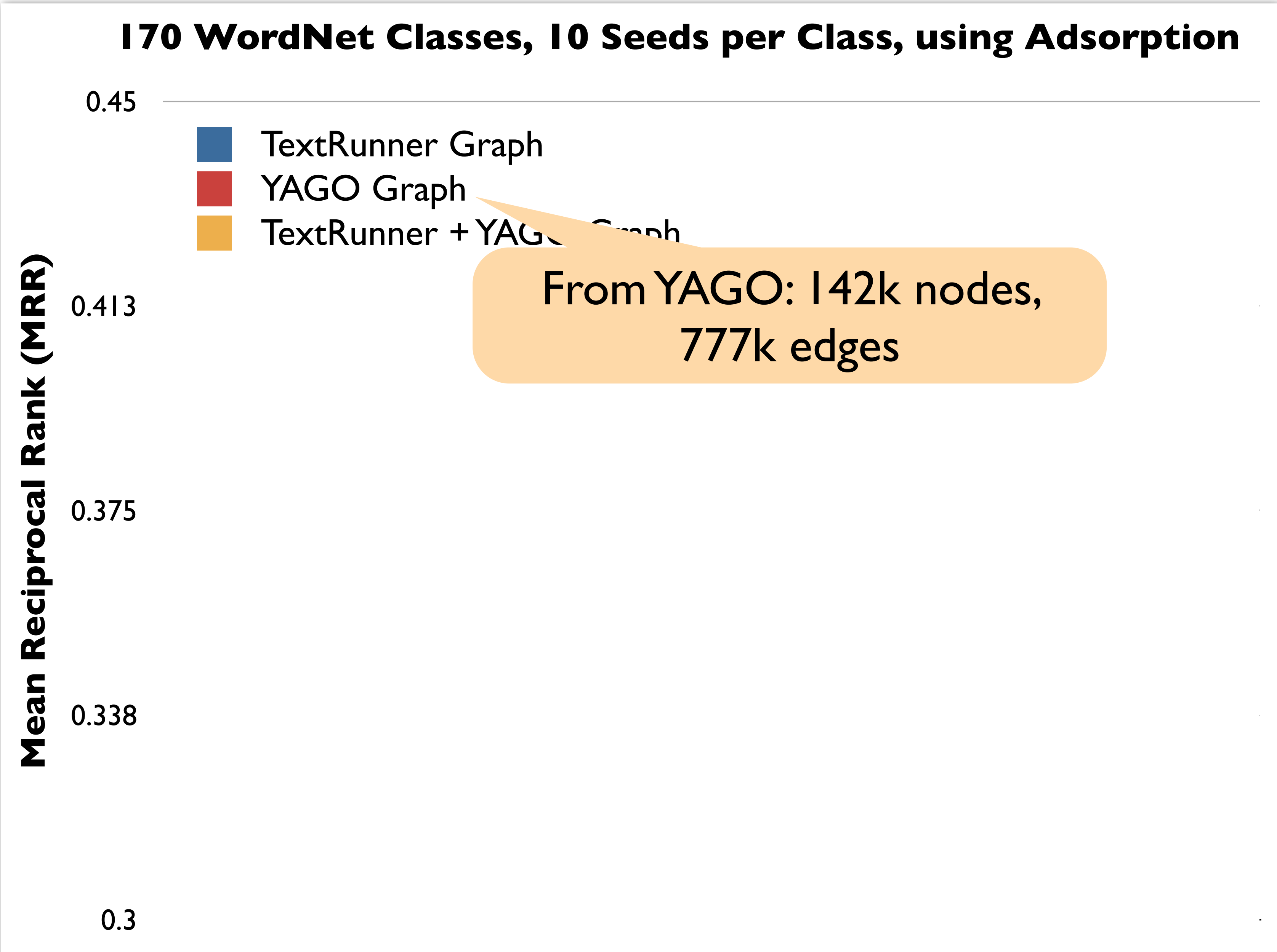
Better Classes with YAGO Attributes



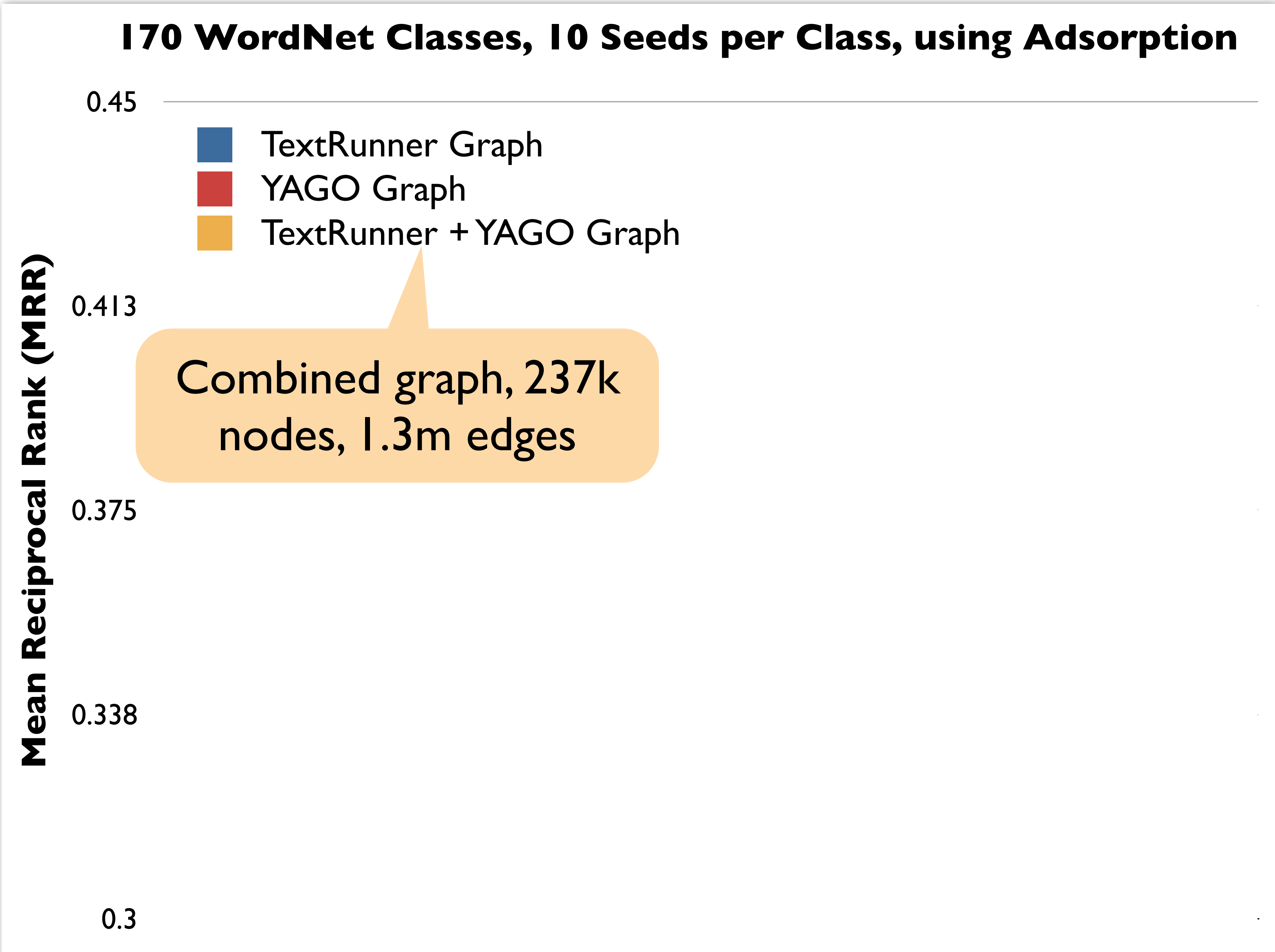
Better Classes with YAGO Attributes



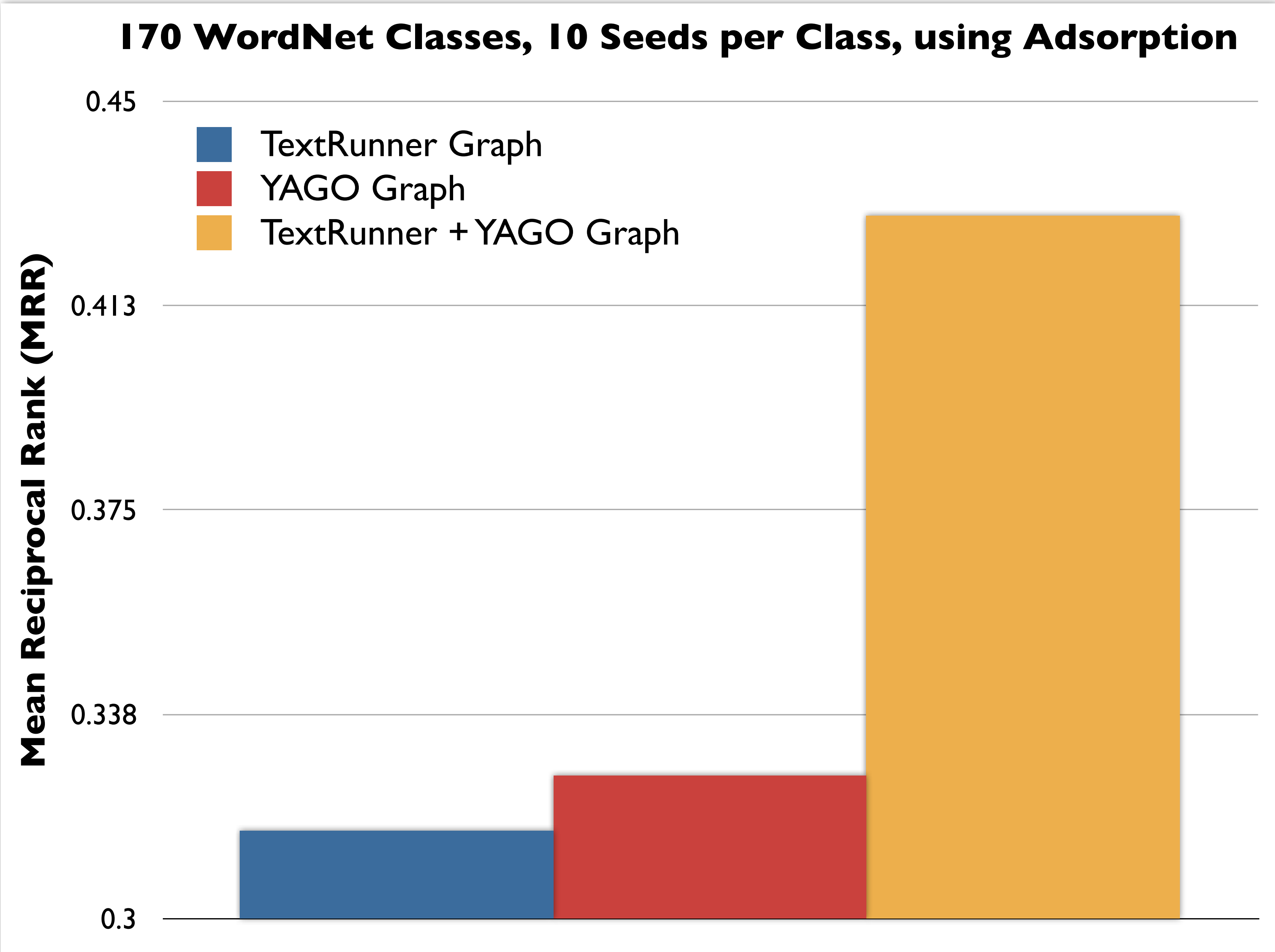
Better Classes with YAGO Attributes



Better Classes with YAGO Attributes



Better Classes with YAGO Attributes



Classes for Attributes

- Qualitative evidence that attribute nodes propagate the right information

YAGO Attribute	**Top-2 WordNet Classes Assigned by MAD** (example instances for each class are shown in brackets)
has_currency	**wordnet_country_108544813 (Burma, Afghanistan)** wordnet_region_108630039 (Aosta Valley, Southern Flinders Ranges)
works_at	**wordnet_scientist_110560637 (Aage Niels Bohr, Adi Shamir)** wordnet_person_100007846 (Catherine Cornelius, Jamie White)
has_capital	**wordnet_state_108654360 (Agusan del Norte, Bali)** wordnet_region_108630039 (Aosta Valley, Southern Flinders Ranges)
born_in	**wordnet_boxer_109870208 (George Chuvalo, Fernando Montiel)** wordnet_chancellor_109906986 (Godon Brown, Bill Bryson)
has_isbn	**wordnet_book_106410904 (Past Imperfect, Berlin Diary)** wordnet_magazine_106595351 (Railway Age, Investors Chronicle)

Propagation Objective

- MAD [Talukdar & Crammer 09] (simplified)

$$\arg\min_{\hat{\boldsymbol{Y}}} \sum_{l=1}^{m+1} \left[\|\boldsymbol{S}\hat{\boldsymbol{Y}}_l - \boldsymbol{S}\boldsymbol{Y}_l\|^2 + \mu_1 \sum_{u,v} \boldsymbol{M}_{uv} (\hat{\boldsymbol{Y}}_{ul} - \hat{\boldsymbol{Y}}_{vl})^2 + \mu_2 \|\hat{\boldsymbol{Y}}_l - \boldsymbol{R}_l\|^2 \right]$$

- m labels, $+1$ dummy label
- $\boldsymbol{M} = \boldsymbol{W}^\top + \boldsymbol{W}$ is the symmetrized weight matrix
- $\hat{\boldsymbol{Y}}_{vl}$: weight of label l on node v
- $\boldsymbol{Y}_{vl}$: seed weight for label l on node v
- $\boldsymbol{S}$: diagonal matrix, nonzero for seed nodes
- $\boldsymbol{R}_{vl}$: regularization target for label l on node v

A Propagation Algorithm

Inputs $\boldsymbol{Y}, \boldsymbol{R} : |V| \times (|L| + 1)$, $\boldsymbol{W} : |V| \times |V|$, $\boldsymbol{S} : |V| \times |V|$ diagonal

$\hat{\boldsymbol{Y}} \leftarrow \boldsymbol{Y}$

$\boldsymbol{M} = \boldsymbol{W} + \boldsymbol{W}^\top$

$Z_v \leftarrow \boldsymbol{S}_{vv} + \mu_1 \sum_{u \neq v} \boldsymbol{M}_{vu} + \mu_2 \quad \forall v \in V$

repeat

    for all $v \in V$ **do**

        $\hat{\boldsymbol{Y}}_v \leftarrow \frac{1}{Z_v}\left((\boldsymbol{S}\boldsymbol{Y})_v + \mu_1 \boldsymbol{M}_{v\cdot}\hat{\boldsymbol{Y}} + \mu_2 \boldsymbol{R}_v\right)$

    end for

until convergence

- Some details of the construction of the input matrices omitted for simplicity
- Converges under reasonable assumptions
- Many variants, alternative objectives (Subramanya and Bilmes 2008)

Good News

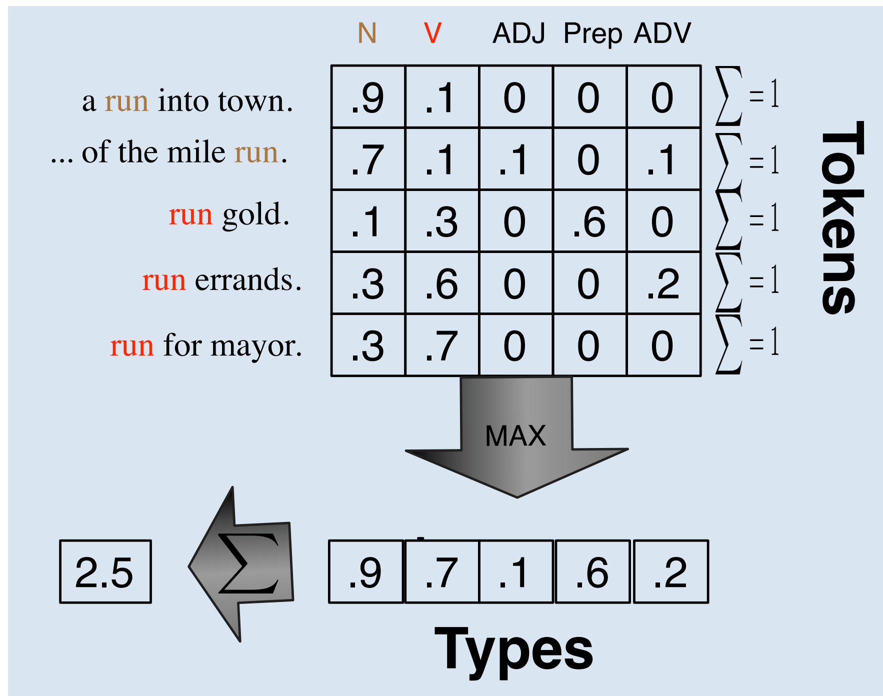
- Label propagation can combine multiple information sources effectively
- Useful coverage of class-instance relations, much bigger than in previous work
- Embarrassingly parallel algorithms
- Graph representation can encode a variety of useful constraints

Limitations

- Can't express "few classes per instance"
- How to classify instances in context?
  - Whistler paintings *vs* Whistler skiers
- Propagation is additive, averaging
  - Algorithmically nice (convexity, convergence)
  - But it can't "push back" to express incompatibilities between classes
    - artist $\supset$ painter $\not\subset$ ski-resort

Few Classes Per Instance

- If a classification is to be informative, it must have limited ambiguity
- *Posterior regularization*: constrain the ambiguity of final labeling
- POS tagging pilot (Graça *et al.* 09): motivated by this work, but easier to test
- Penalize the ℓ_1/ℓ_∞ norm of the posterior distribution of classes given instances



Contextual Classification

- Graph-based approach:
  - One node per mention (lots of nodes!)
  - Link mention nodes that have similar contexts
- Language modeling approach:
  - Class labels are also terms
  - Model probability of class given context
  - Find most likely class for each mention given its context

Back to Supervised Learning

- Where do edge weights come from?
- These experiments: heuristic scoring functions drawn from language modeling and information retrieval
- Can we learn edge scoring functions?
  - Minkov & Cohen: learning from random walks
  - McCallum's group: learning factor potentials by M-H sampling
  - Learn from user feedback (Talukdar *et al.*, SIGMOD 2010)

Summary

- First steps in inferring broad-coverage semantic relationships from the actions of Web users:
  - What they write
  - How they interact with search results
- Multiple correlated sources provide a wealth of indirect supervision
- (Some) graph-based algorithms scale up
- Related work:
  - Wang & Cohen's SEAL and follow-ups
  - Weikum lab's SOFIE

Current Work

- Web-scale contextual semantic annotation
  - is-a, co-reference
- Combining multiple relationships
  - The distinct senses of “Whistler“ belong to disjoint co-reference classes
- Probabilistic interpretations
  - Relational factor graphs
  - Non-parametric Bayesian models