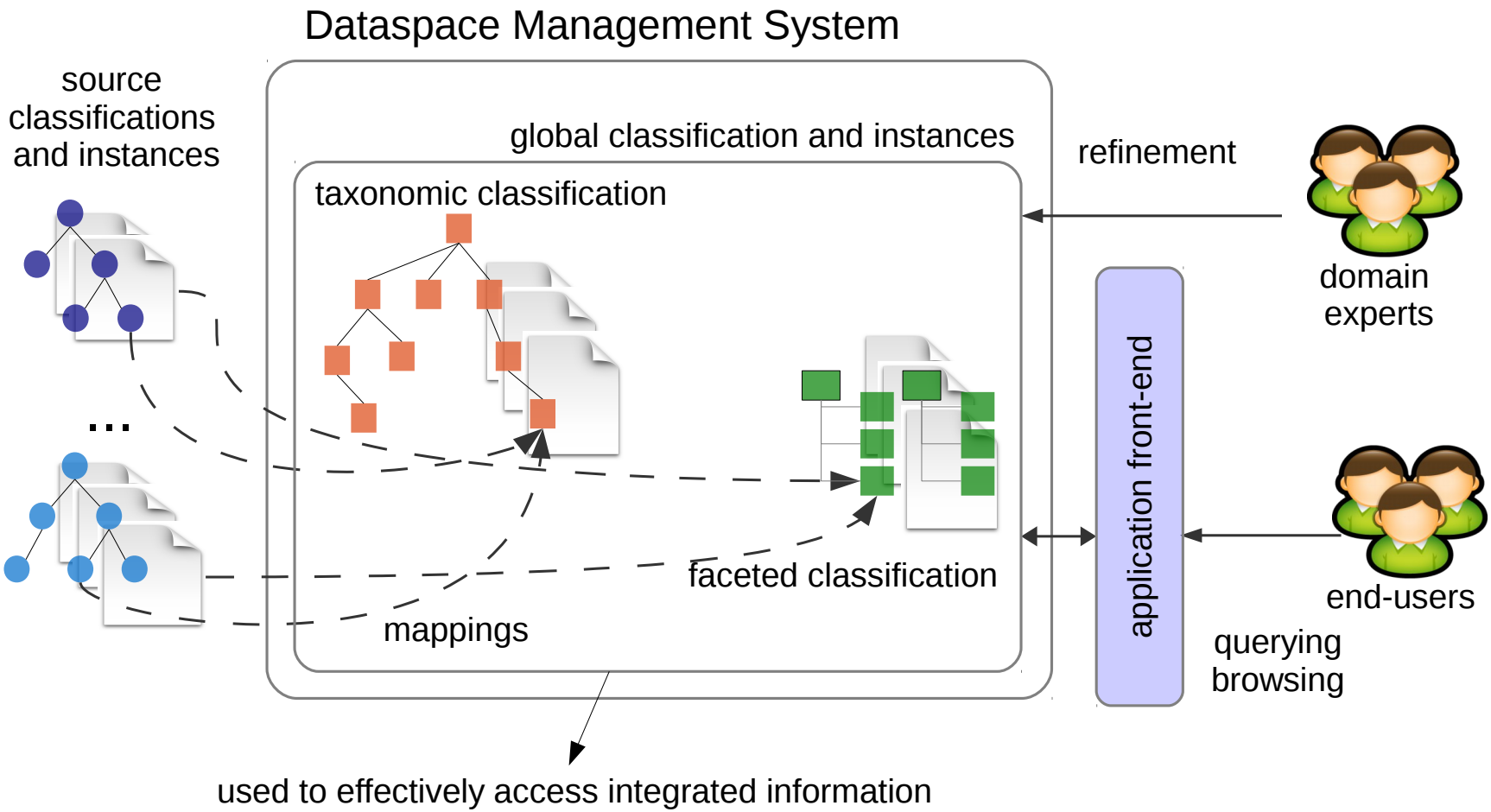
Extracting Facets from Lost Fine-grained Categorizations in Dataspaces

Riccardo Porrini[1,2], Matteo Palmonari[1] and Carlo Batini[1]

[1] **DISCo, University of Milano-Bicocca**
[matteo.palmonari, carlo.batini]@disco.unimib.it

[2] **7Pixel s.r.l.**
riccardo.porrini@trovaprezzi.it

Dataspaces with Multiple Classifications



Taxonomy-based Classification in Price Comparison Engines

Taxonomy: categories organized through a hierarchical structure *(informal)*

Home Page › Giochi e Hobby › Videogiochi › Videogiochi per Console › Videogiochi per PS3

coarse-grained classification helps end-users to recall the "class" of instances

Filtra la ricerca

PREZZO

€ 4,7 - € 1550

MARCA

- Sony (428)
- Electronic Arts (402)
- Activision (262)
- Ubisoft Entertainment (204)
- 2K Games (102)
- Altri(e)...

GENERE

- Azione (1502)
- Sparare (541)
- Sparare in prima persona (FPS) (352)
- Sportivo (332)
- Altro (209)
- Altri(e)...

ETÀ RACCOMANDATA (PEGI)

- 3+ (394)
- 7+ (1...

Risultati: **6025**

Electronic Arts FIFA 14

Scheda tecnica

Videogioco per Sony PlayStation 3, g
ambientazione: realistico. Classificaz

Aggiungi a confronto

Aggiungi a una lista

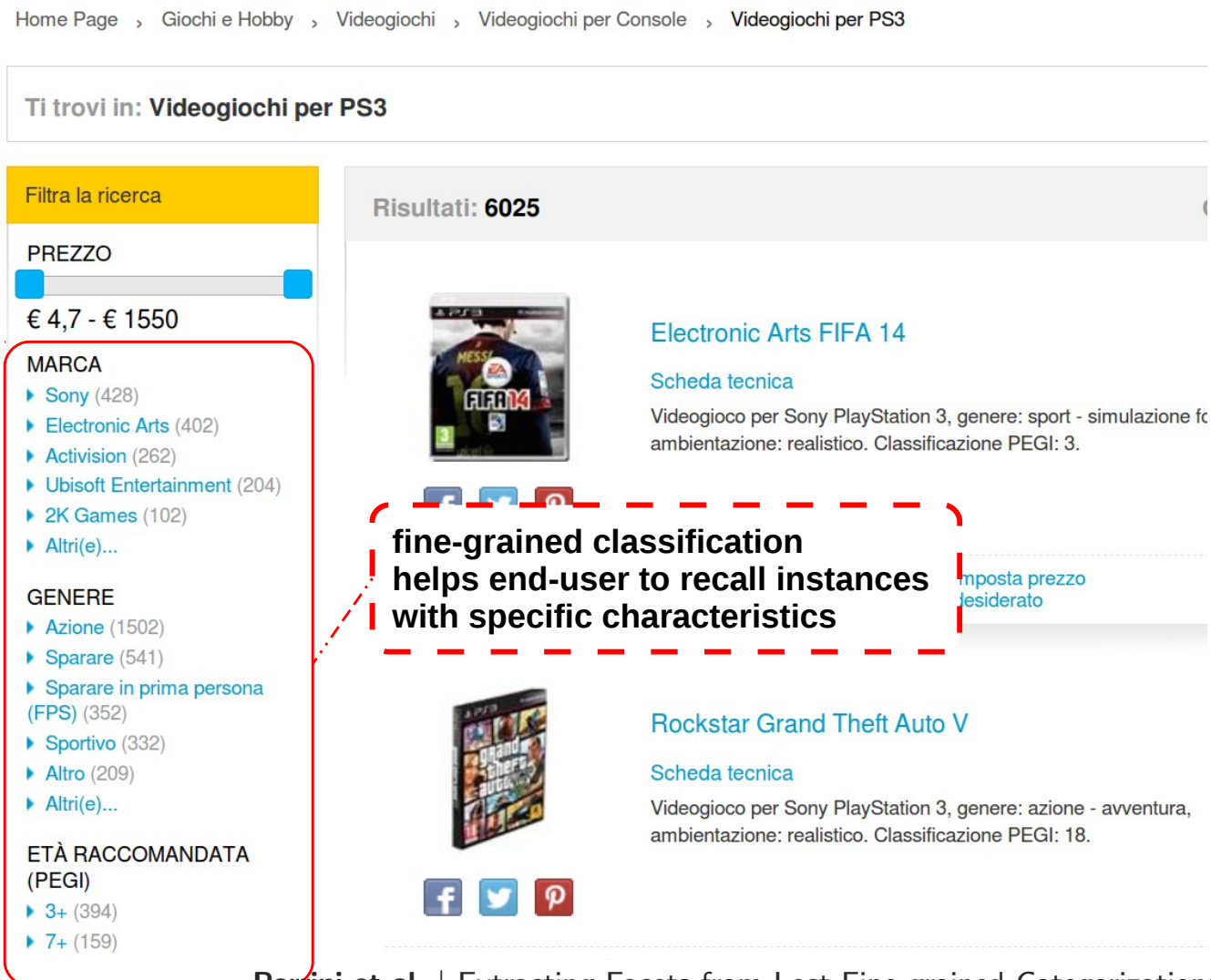
Rockstar Grand Theft Auto V

Scheda tecnica

Videogioco per Sony PlayStation 3, g
ambientazione: realistico. Classificaz

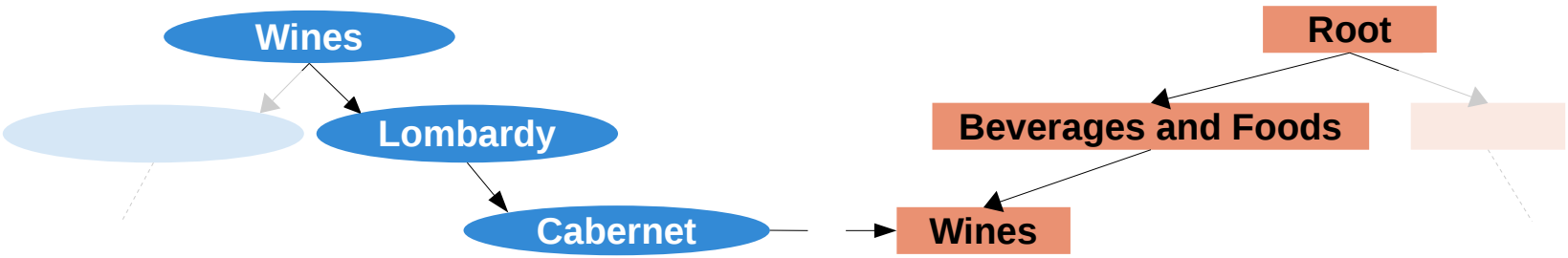
Facet-based Classification in Price Comparison Engines

Facet: a clearly defined, mutually exclusive, and collectively exhaustive aspect, property, or characteristic of a class or specific subject [Ta04]



Mappings

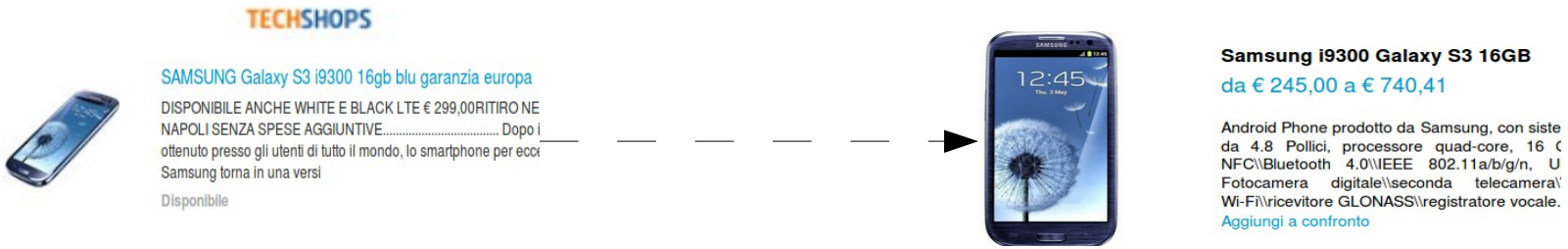
category mapping



source taxonomy

global taxonomy

instance mapping



source instance

global instance

Motivation

Faceted classification bootstrapping

- time and effort consuming
- requires detailed knowledge about dataspace instances

Motivation

Faceted classification bootstrapping

- time and effort consuming
- requires detailed knowledge about dataspace instances

However

- source taxonomies usually provide much more granular classification
- this granular information is lost when mapping specific source categories to generic global ones

Motivation

Faceted classification bootstrapping

- time and effort consuming
- requires detailed knowledge about dataspace instances

However

- source taxonomies usually provide much more granular classification
- this granular information is lost when mapping specific source categories to generic global ones

How about extracting facets from those lost fine-grained source taxonomies?

Problem Statement

Facet F^g

a finite set of values $v_1, \ldots, v_n$ associated with a label used to describe a characteristic of the objects belonging to a global category g

Problem Statement

Facet F^g

a finite set of values $v_1, \ldots, v_n$ associated with a label used to describe a characteristic of the objects belonging to a global category g

Facet Extraction Problem

given a global leaf category g, a set of mappings M from source categories $s_1, \ldots, s_n$ to g in the form $g \leftarrow s_1, \ldots, g \leftarrow s_n$, extract a set $\mathcal{F}^g$ of facets F^g , each one associated with a label

Problem Statement

Facet F^g

a finite set of values $v_1, \ldots, v_n$ associated with a label used to describe a characteristic of the objects belonging to a global category g

Facet Extraction Problem

given a global leaf category g, a set of mappings M from source categories $s_1, \ldots, s_n$ to g in the form $g \leftarrow s_1, \ldots, g \leftarrow s_n$, extract a set $\mathcal{F}^g$ of facets F^g , each one associated with a label

Wines

Problem Statement

Facet F^g

a finite set of values $v_1, \ldots, v_n$ associated with a label used to describe a characteristic of the objects belonging to a global category g

Facet Extraction Problem

given a global leaf category g, a set of mappings M from source categories $s_1, \ldots, s_n$ to g in the form $g \leftarrow s_1, \ldots, g \leftarrow s_n$, extract a set $\mathcal{F}^g$ of facets F^g , each one associated with a label

Wines

Winery Country of Origin

- USA
- China
- Australia
- Italy

Wine Alcohol By Volume

- ☐ Under 10%
- ☐ 10% to 12%
- ☐ 12% to 14%
- ☐ 14% & Up

Grape Variety

- ☐ Blend - White
- ☐ Blend - Other
- ☐ Fruit
- ☐ Muscadine
- ☐ Cabernet Sauvignon
- ☐ Pinot Noir
- ☐ Chardonnay

Wine Bottle Volume

- ☐ 375 mL
- ☐ 500 mL
- ☐ 750 mL

Specialty Wine Type

- ☐ Sustainable
- ☐ Small Lot
- ☐ Kosher
- ☐ Gluten-Free

Wine Vintage

- ☐ 2011
- ☐ 2010
- ☐ 2009
- ☐ 2008
- ☐ 2007

Challenges

Source taxonomies are:

- **many**
  3900 within the TrovaPrezzi italian price comparison engine

Challenges

Source taxonomies are:

- **many**
  3900 within the TrovaPrezzi italian price comparison engine
- **noisy**
  type > white > by vine > chardonnay > producer > firriato

Challenges

Source taxonomies are:

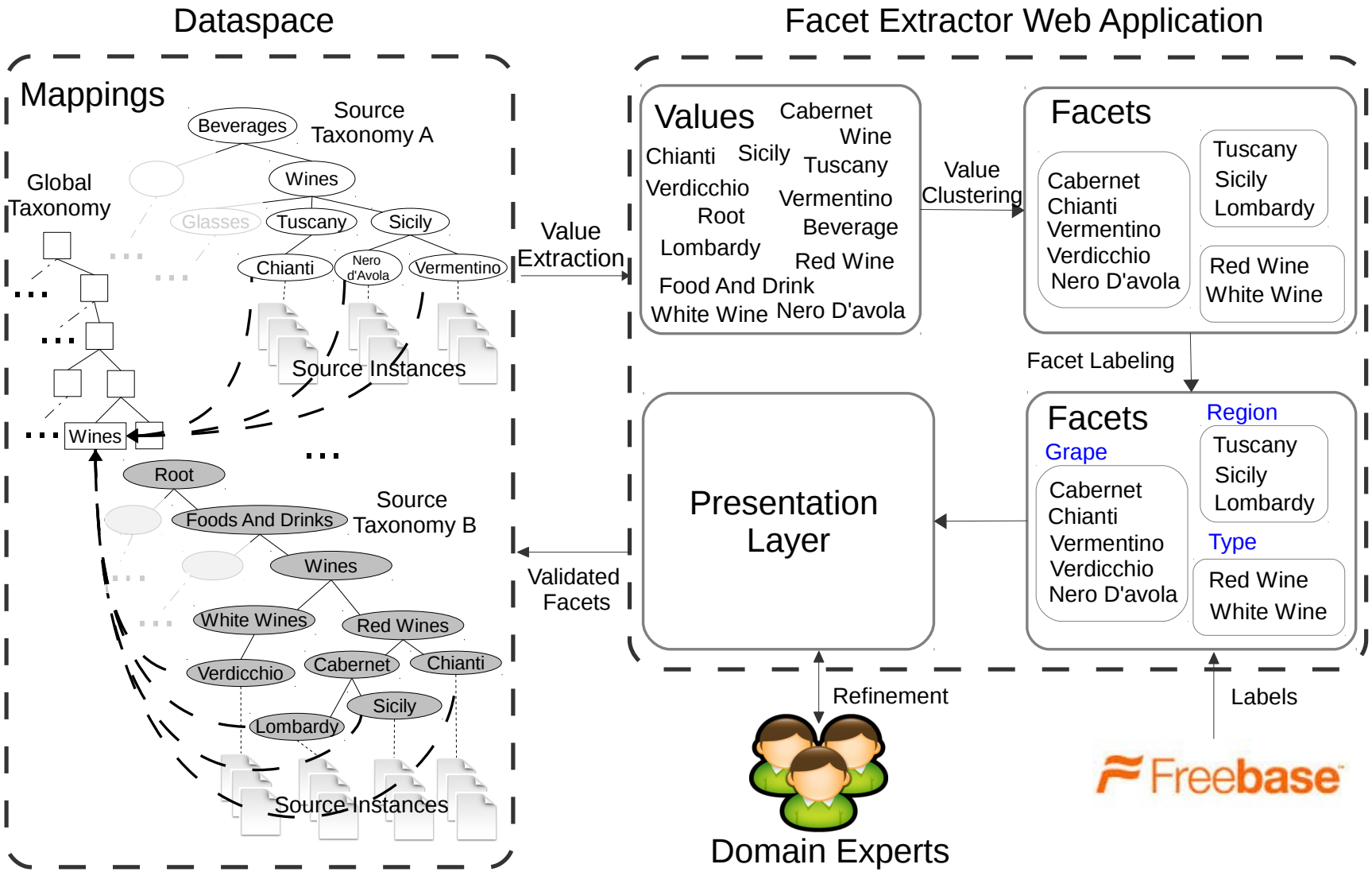
- **many**
  3900 within the TrovaPrezzi italian price comparison engine
- **noisy**
  type > white > by vine > chardonnay > producer > firriato
- **heterogeneous**
  type > white > by vine > chardonnay > producer > firriato
  wines > white wines > greco di tufo

Challenges

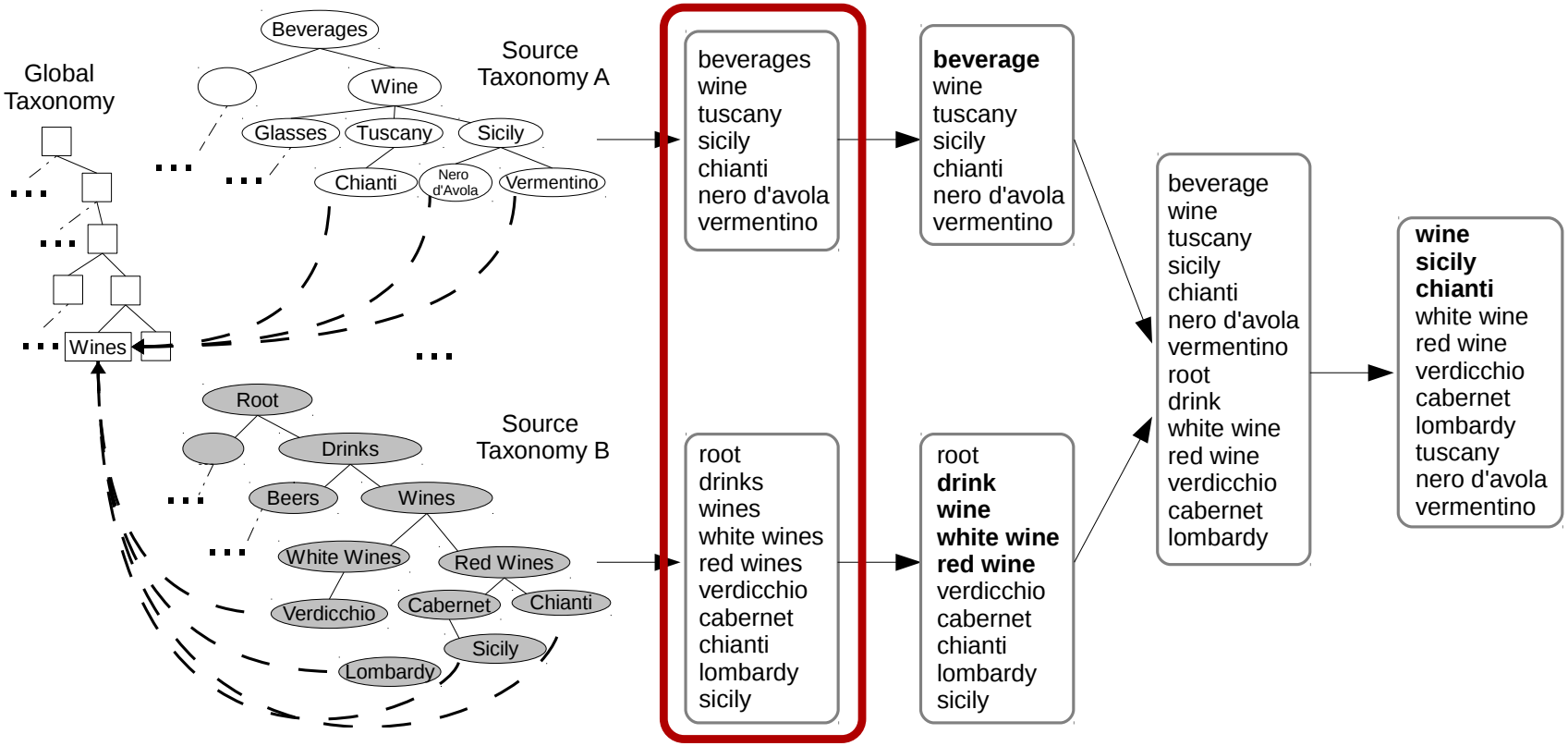
Source taxonomies are:

- **many**
  3900 within the TrovaPrezzi italian price comparison engine
- **noisy**
  type > white > by vine > chardonnay > producer > firriato
- **heterogeneous**
  type > white > by vine > chardonnay > producer > firriato
  wines > white wines > greco di tufo
- **ambiguous** different semantics for different contexts
  red is a wine type for wines
  and a color for shirts

Language independent facet extraction

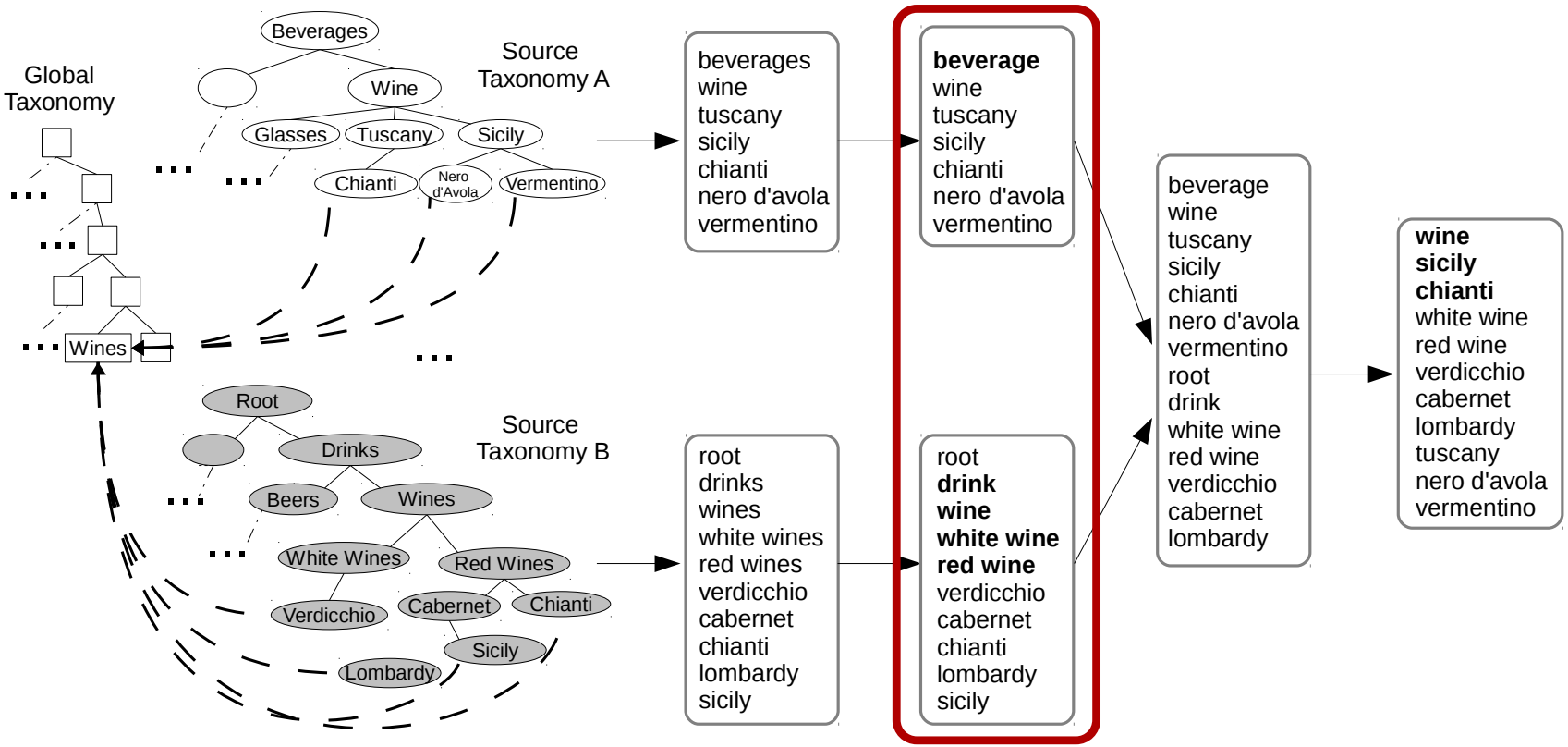


Value Extraction



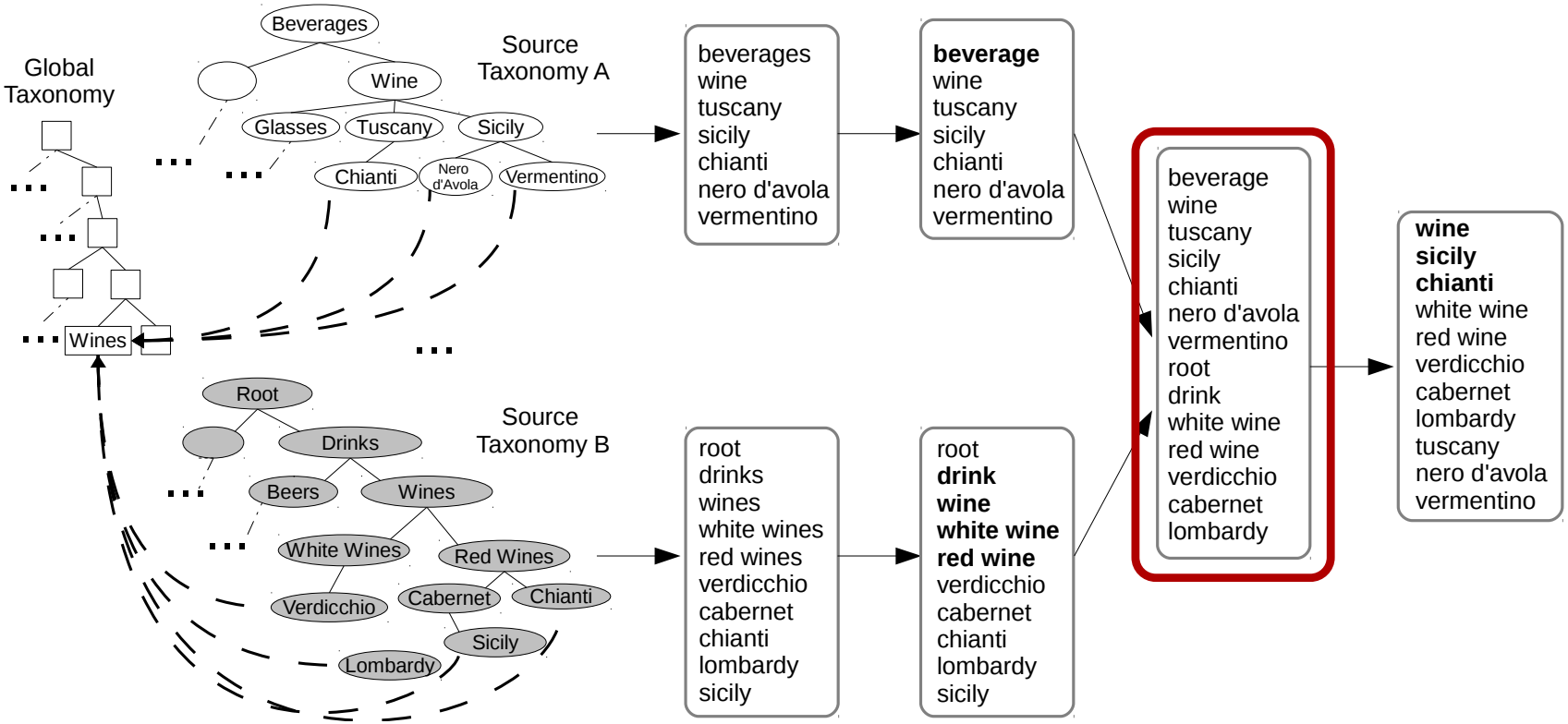
$$V_S^g = \{s \mid \exists\ g \leftarrow s \text{ or } \exists\ g \leftarrow s', \text{ with } s \in S \text{ and } s' \text{ is a descendant of } s\}$$

Value Extraction



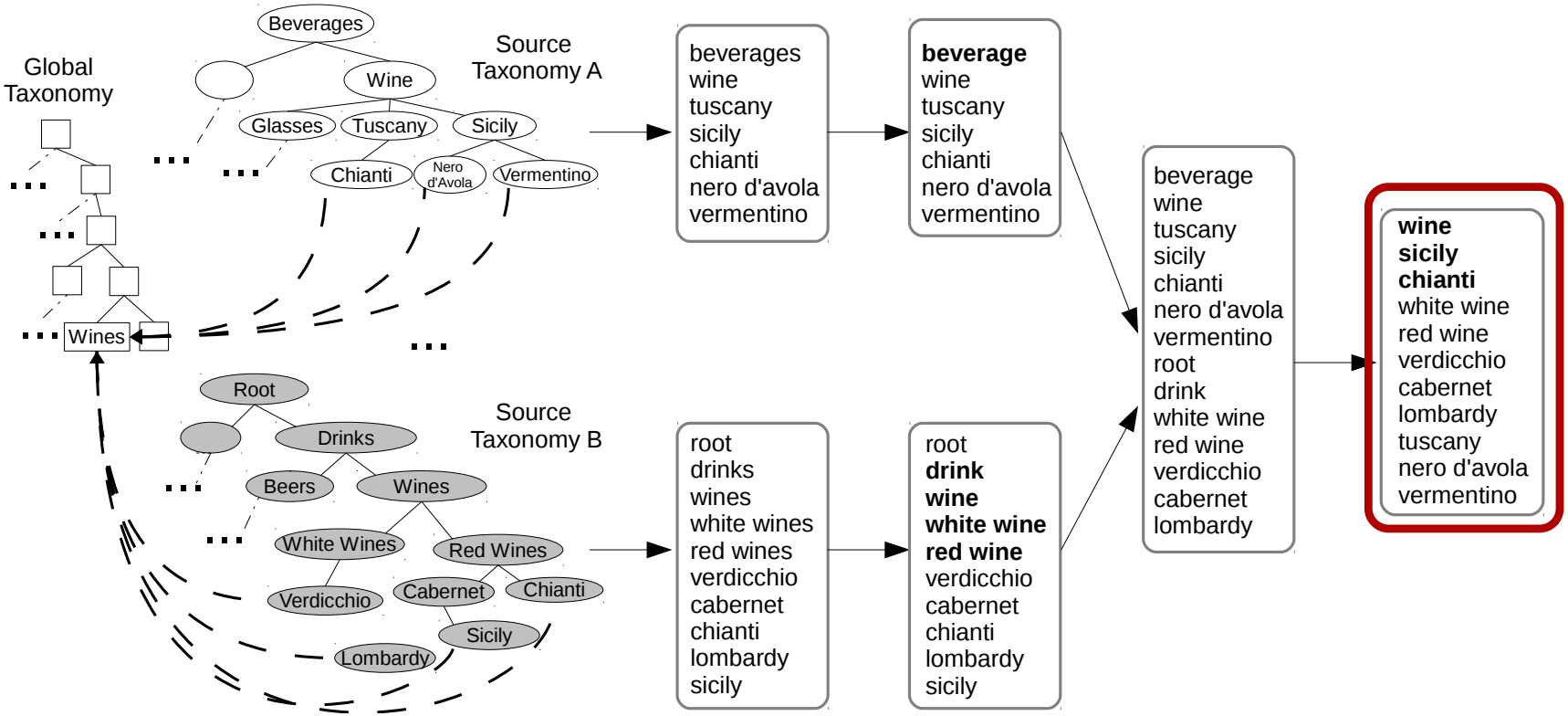
Normalization and stemming

Value Extraction



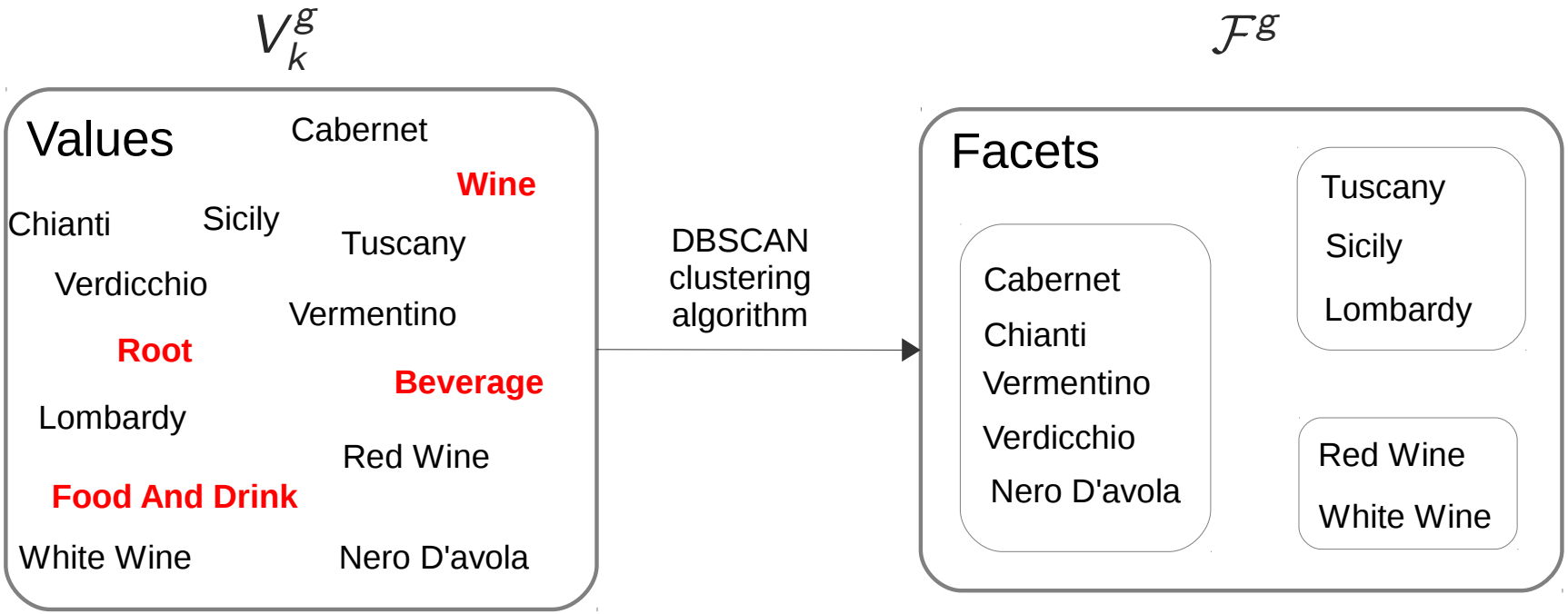
$$V^g = \bigcup_{i=1}^{n} V^g_{S^i}$$

Value Extraction



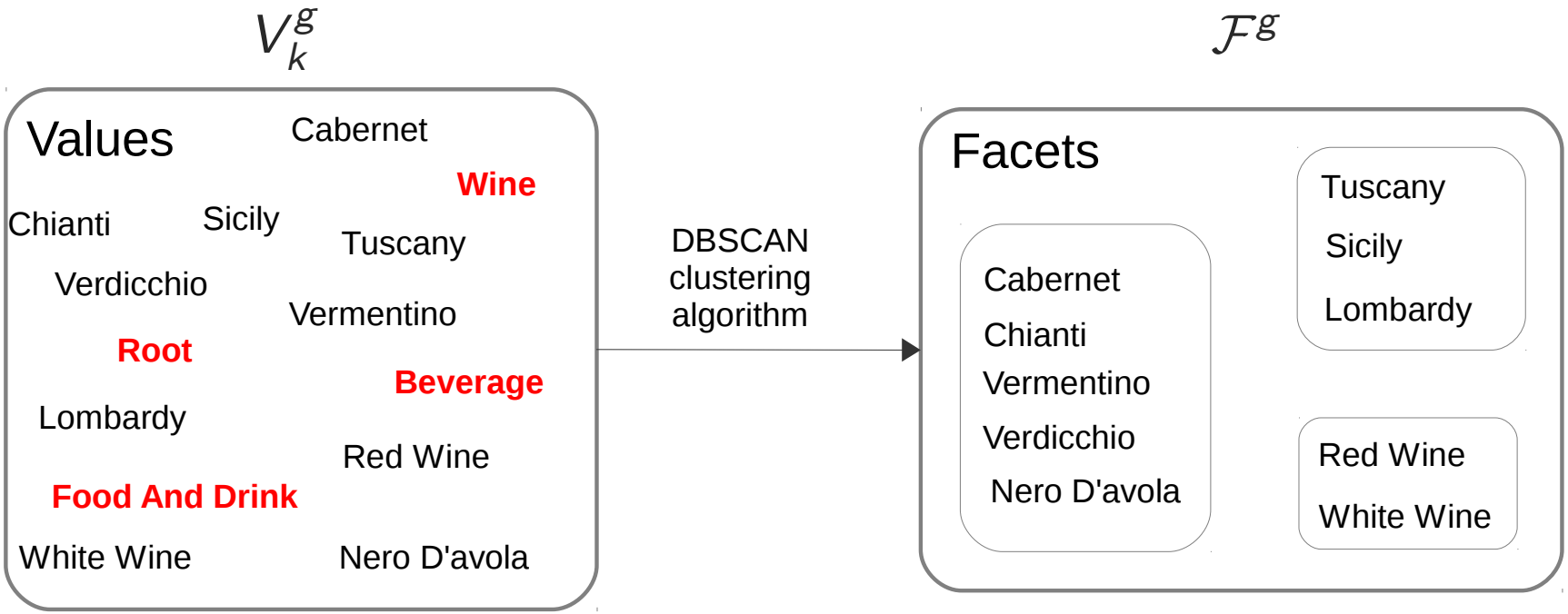
Set V_k^g of the top k frequent values over all $V_{S^i}^g$

Value Clustering



DBSCAN [Es96] Density based clustering

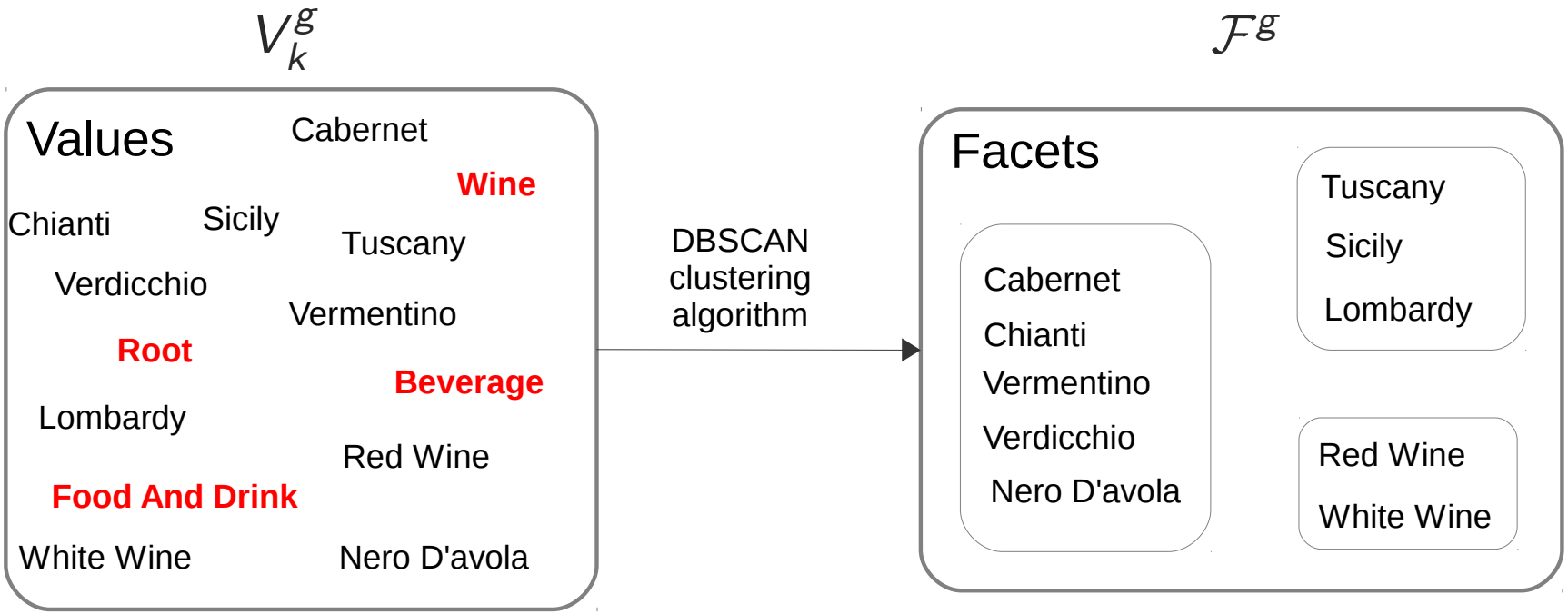
Value Clustering



DBSCAN [Es96] Density based clustering

- incorporates the concept of *noise*

Value Clustering



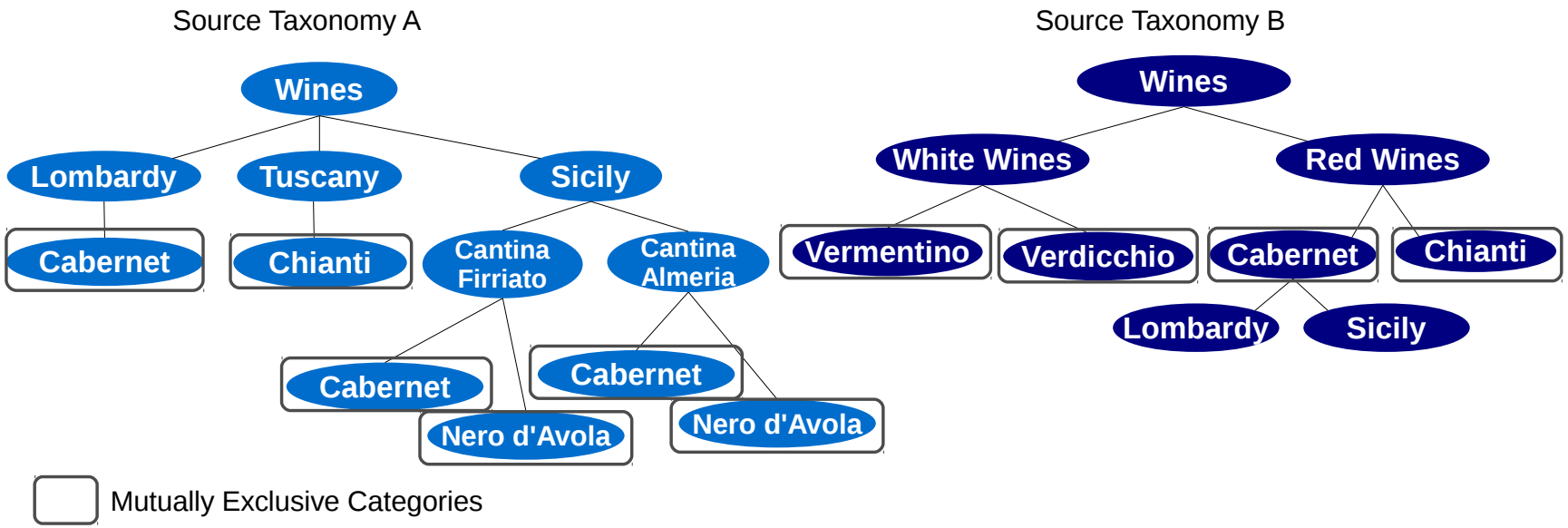
DBSCAN [Es96] Density based clustering

- incorporates the concept of *noise*
- number of clusters (i.e., facets) not known a priori

Source Category Mutual Exclusivity Principle

SCME principle

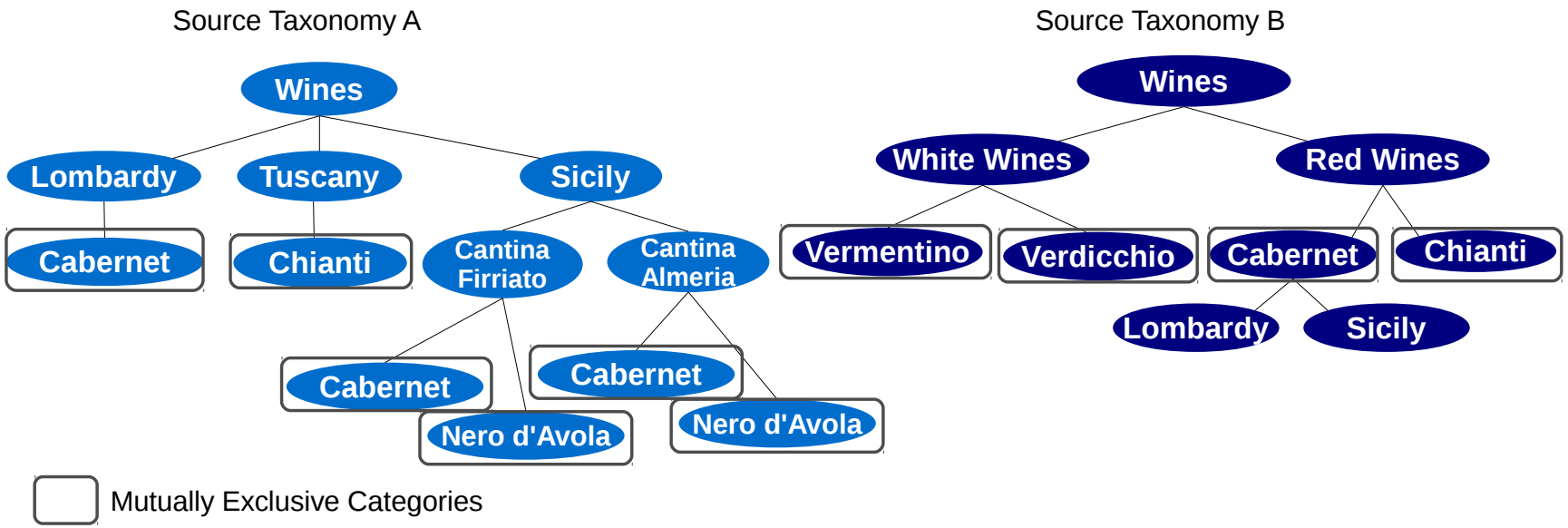
the more two values refer to mutually exclusive categories, the more they should be grouped together into the same facet



Source Category Mutual Exclusivity Principle

SCME principle

the more two values refer to mutually exclusive categories, the more they should be grouped together into the same facet

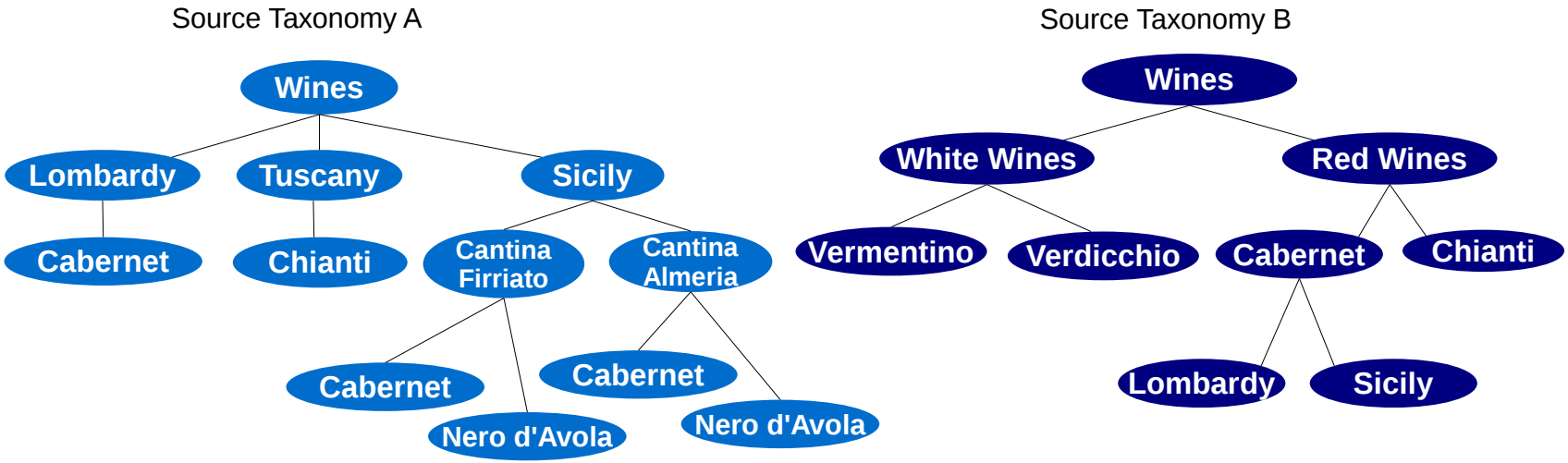


Hint

given two source categories s_1 and s_2, their occurrence as siblings indicates that s_1 and s_2 are mutually exclusive

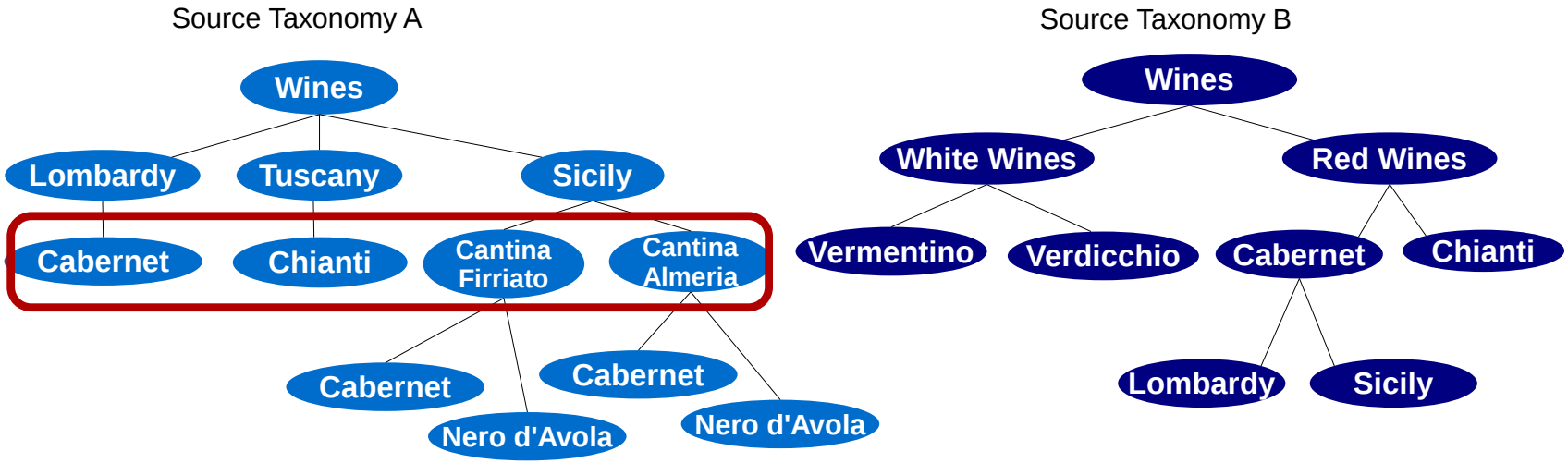
Taxonomy Layer Distance

captures the **SCME** principle by considering the co-occurrence of categories on a same taxonomy layer



Taxonomy Layer Distance

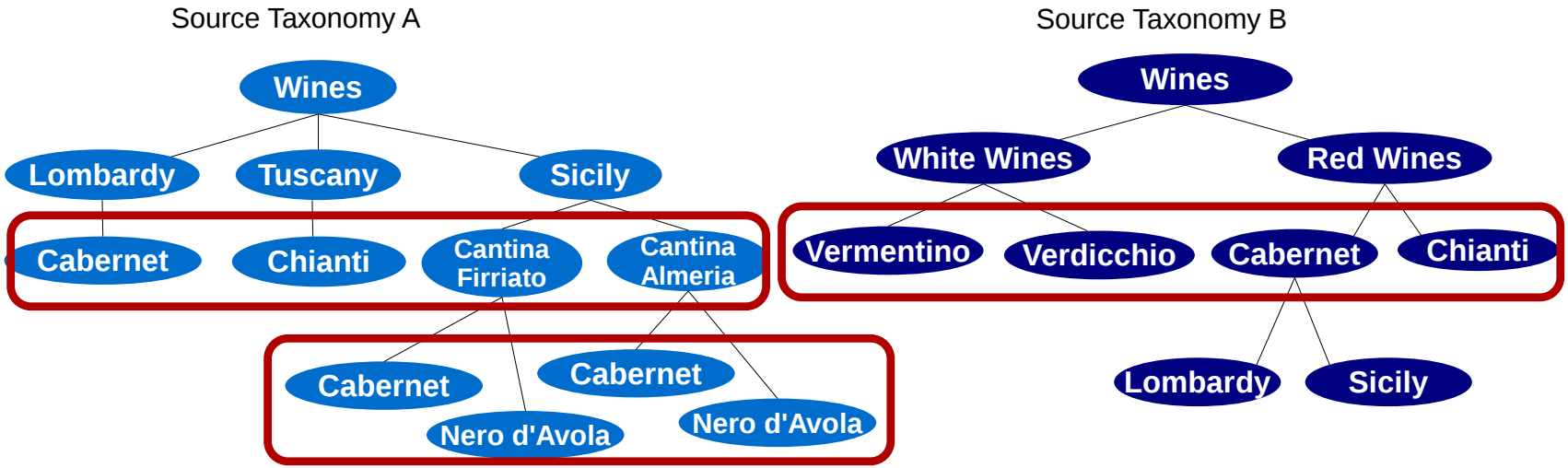
captures the **SCME** principle by considering the co-occurrence of categories on a same taxonomy layer



A taxonomy *layer* l^S of S is the set of all categories that are at the same distance from the taxonomy root

Taxonomy Layer Distance

captures the **SCME** principle by considering the co-occurrence of categories on a same taxonomy layer



A value v is represented by the set $L_v = \bigcup_{i=1}^{n} L_v^{S^i}$ of layers containing v for each source taxonomy S, where $L_v^S = \{I^S \mid v \in I^S\}$ is the set of layers containing v in the source taxonomy S

Taxonomy Layer Distance

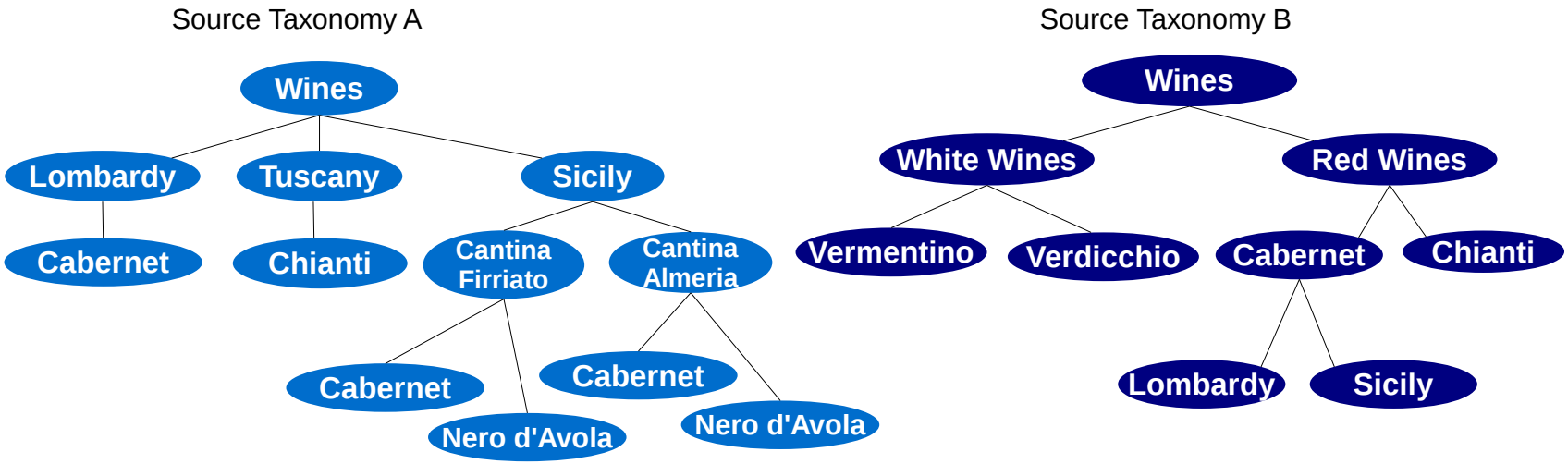
$$\mathrm{TLD}(v_1, v_2) = 1 - \frac{|L_{v_1} \cap L_{v_2}|}{|L_{v_1} \cup L_{v_2}|}$$

Jaccard Distance between the two sets of taxonomy layers where two values v_1 and v_2 occur

Taxonomy Layer Distance

$$\mathrm{TLD}(v_1, v_2) = 1 - \frac{|L_{v_1} \cap L_{v_2}|}{|L_{v_1} \cup L_{v_2}|}$$

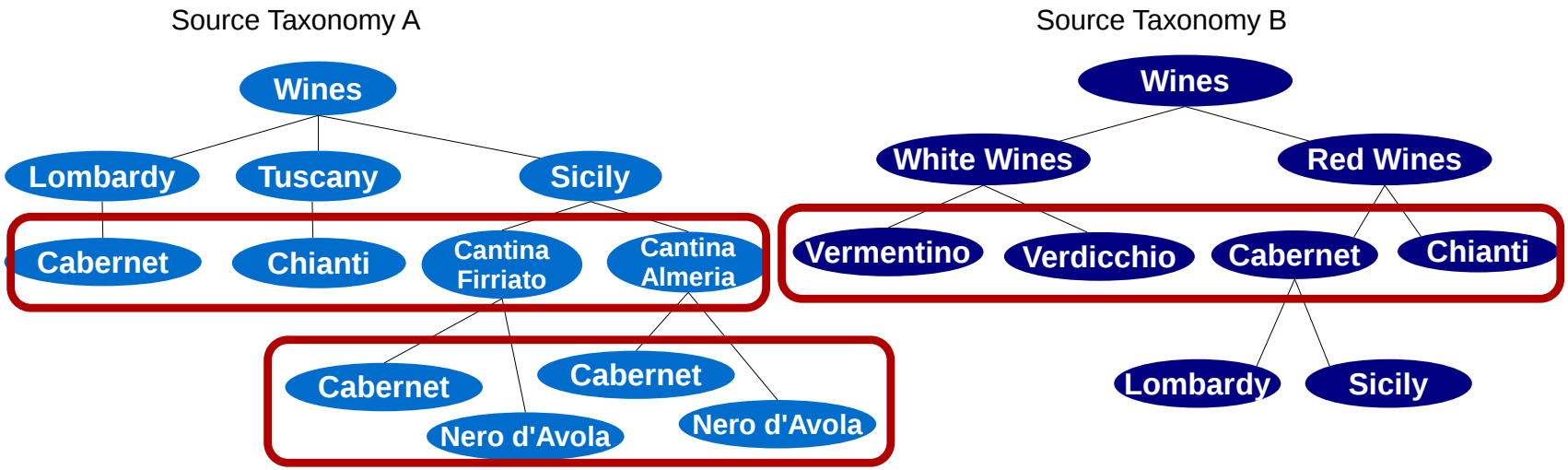
Jaccard Distance between the two sets of taxonomy layers where two values v_1 and v_2 occur



Taxonomy Layer Distance

$$\mathrm{TLD}(v_1, v_2) = 1 - \frac{|L_{v_1} \cap L_{v_2}|}{|L_{v_1} \cup L_{v_2}|}$$

Jaccard Distance between the two sets of taxonomy layers where two values v_1 and v_2 occur

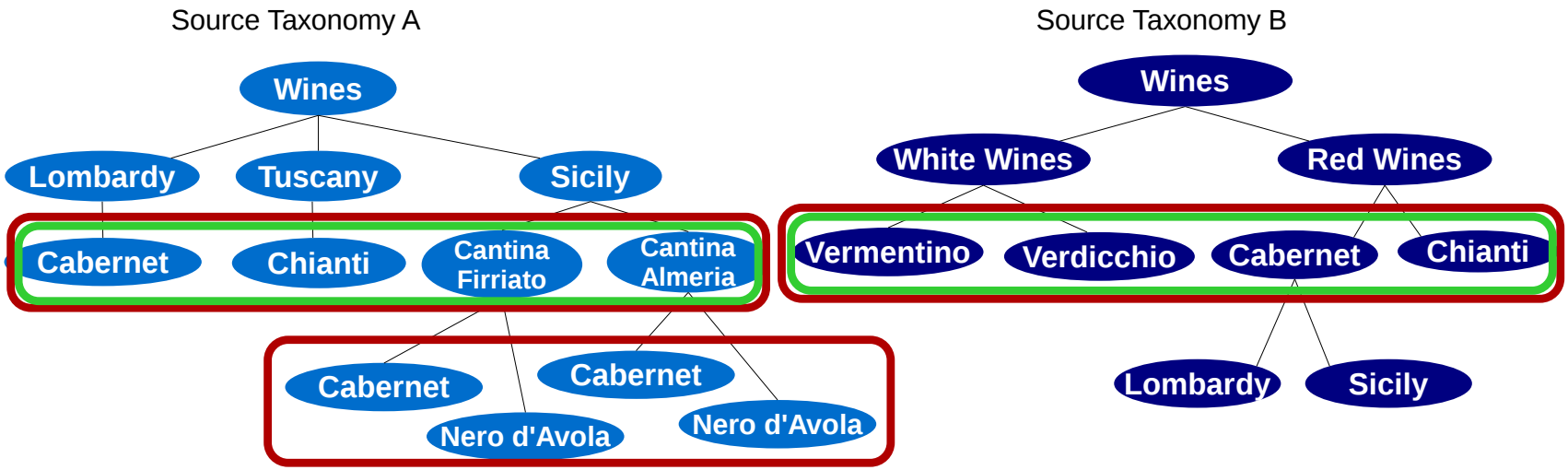


cabernet

Taxonomy Layer Distance

$$\mathrm{TLD}(v_1, v_2) = 1 - \frac{|L_{v_1} \cap L_{v_2}|}{|L_{v_1} \cup L_{v_2}|}$$

Jaccard Distance between the two sets of taxonomy layers where two values v_1 and v_2 occur

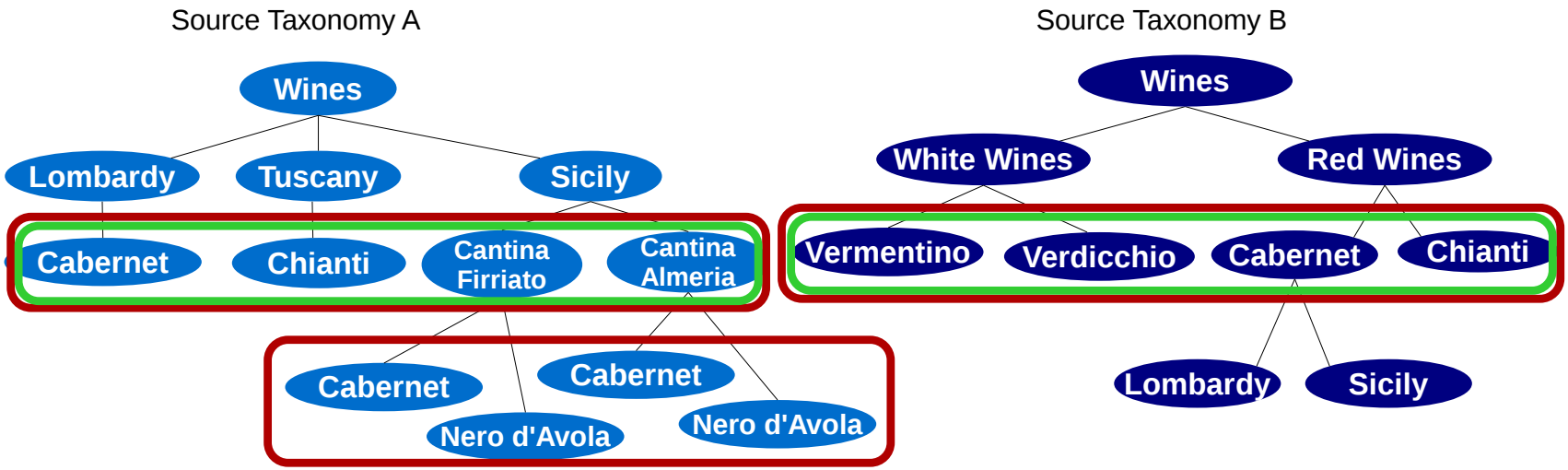


cabernet *chianti*

Taxonomy Layer Distance

$$\mathrm{TLD}(v_1, v_2) = 1 - \frac{|L_{v_1} \cap L_{v_2}|}{|L_{v_1} \cup L_{v_2}|}$$

Jaccard Distance between the two sets of taxonomy layers where two values v_1 and v_2 occur



cabernet *chianti*

$$\mathrm{TLD}(\mathit{cabernet}, \mathit{chianti}) = 1 - \frac{|L_{\mathit{cabernet}} \cap L_{\mathit{chianti}}|}{|L_{\mathit{cabernet}} \cup L_{\mathit{chianti}}|} = 1 - \frac{2}{3} = \frac{1}{3}$$

Facet Labelling

Reconcile the values of each facet F^g to the Freebase knowledge base

Facet Labelling

Reconcile the values of each facet F^g to the Freebase knowledge base

- submit each facet value as a keyword query

Facet Labelling

Reconcile the values of each facet F^g to the Freebase knowledge base

- submit each facet value as a keyword query
- obtain a list of Freebase entities

Facet Labelling

Reconcile the values of each facet F^g to the Freebase knowledge base

- submit each facet value as a keyword query
- obtain a list of Freebase entities
- select the type of each entity

Facet Labelling

Reconcile the values of each facet F^g to the Freebase knowledge base

- submit each facet value as a keyword query
- obtain a list of Freebase entities
- select the type of each entity
- pick the most frequent type

Evaluation

Goal

show that TLD effectively captures the SCME principle and supports domain experts in facets definition

Evaluation

Goal

show that TLD effectively captures the SCME principle and supports domain experts in facets definition

Comparison with:

- Leacock and Chodorow (LC) similarity [Le98]
  shortest path scaled by the depth of the taxonomy
- Wu and Palmer (WP) similarity [Wu94]
  distance from nearest common ancestor and distance of the nearest common ancestor from the taxonomy root

Evaluation

Goal

show that TLD effectively captures the SCME principle and supports domain experts in facets definition

Comparison with:

- Leacock and Chodorow (LC) similarity [Le98]
  shortest path scaled by the depth of the taxonomy
- Wu and Palmer (WP) similarity [Wu94]
  distance from nearest common ancestor and distance of the nearest common ancestor from the taxonomy root

Evaluation using real world data from the italian PCE TrovaPrezzi

- 10 global categories
- 688 source taxonomies
- 22594 leaf mappings
- ran the extraction phase
- values manually grouped in facets by domain experts

Evaluation - Metrics

State-of-the-art evaluation campaign [Do11, Ka12]

Evaluation - Metrics

State-of-the-art evaluation campaign [Do11, Ka12]

- **Value Effectiveness:** Precision(P), Recall(R), FMeasure(F_1) evaluate the ability to filter noisy values out

Evaluation - Metrics

State-of-the-art evaluation campaign [Do11, Ka12]

- **Value Effectiveness:** Precision(P), Recall(R), FMeasure(F_1) evaluate the ability to filter noisy values out
- **Value Clustering Effectiveness:** Purity (P^*), Normalized Mutual Information (NMI^*), Entropy (E^*), FMeasure (F^*)

Evaluation - Metrics

State-of-the-art evaluation campaign [Do11, Ka12]

- **Value Effectiveness:** Precision(P), Recall(R), FMeasure(F_1) evaluate the ability to filter noisy values out
- **Value Clustering Effectiveness:** Purity (P^*), Normalized Mutual Information (NMI^*), Entropy (E^*), FMeasure (F^*)
- **Overall quality:** aggregates facet value precision P, facet value recall R and clustering F-measure F^*

$$PRF^* = \frac{3 * P * R * F^*}{R * P + P * F + P * R}$$

Evaluation - Results

	Value Effectiveness			*Clustering Effectiveness*				*Quality*
	P	R	F_1	F^*	NMI^*	Purity	E^*	PRF^*
LC	0.394	0.953	0.537	0.666	0.709	0.220	0.685	0.531
WP	0.377	**0.984**	0.525	0.682	0.714	0.210	0.744	0.520
TLD	**0.416**	0.901	**0.541**	**0.719**	**0.746**	**0.286**	**0.416**	**0.558**

Evaluation - Results

	Value Effectiveness			*Clustering Effectiveness*				*Quality*
	P	R	F_1	F^*	NMI^*	Purity	E^*	PRF^*
LC	0.394	0.953	0.537	0.666	0.709	0.220	0.685	0.531
WP	0.377	**0.984**	0.525	0.682	0.714	0.210	0.744	0.520
TLD	**0.416**	0.901	**0.541**	**0.719**	**0.746**	**0.286**	**0.416**	**0.558**

- TLD more effective in finding relevant facet values and discarding noisy ones (high F_1)

Evaluation - Results

	Value Effectiveness			*Clustering Effectiveness*				*Quality*
	P	R	F_1	F^*	NMI^*	Purity	E^*	PRF^*
LC	0.394	0.953	0.537	0.666	0.709	0.220	0.685	0.531
WP	0.377	**0.984**	0.525	0.682	0.714	0.210	0.744	0.520
TLD	**0.416**	0.901	**0.541**	**0.719**	**0.746**	**0.286**	**0.416**	**0.558**

- TLD more effective in finding relevant facet values and discarding noisy ones (high F_1)
- TLD more effective in clustering homogeneous values (high *clustering effectiveness*)

Evaluation - Results

LC	F_1^g = {Wine, Red Wine, White Wine, . . . , Piedmont, Lombardy, . . . , Sicily, Donnafugata, Cusumano, . . . , Alessandro di Camporeale, . . . , France} (98)
WP	F_1^g = {Wine, Red Wine, White Wine, . . . , Piedmont, Lombardy, . . . , Sicily, Donnafugata, Cusumano, . . . , France} (100)
TLD	F_1^g = { Piedmont, Tuscany, Sicily, , . . . , France } (14) F_2^g = { Red, White, Rosé } (3) F_3^g = { Red Wine, White Wine, Rosé Wine } (3) F_4^g = { Moscato, Chardonnay, . . . , Merlot } (13) F_5^g = { Tuscany Wine, Sicily Wine} (2) F_6^g = { Donnafugata, Cusumano, . . . , Principi di Butera } (27)
Gold Standard	F_1^g = { Piedmont, Lombardy, . . . , Sicily } (21) F_2^g = { Red Wine, White Wine, . . . , Rosé Wine } (14) F_3^g = { Donnafugata, Cusumano, . . . , Alessandro di Camporeale} (12)

Conclusions

- semi-automatic, language independent approach to facets extraction from heterogeneous taxonomies within dataspaces

Future work

Conclusions

- semi-automatic, language independent approach to facets extraction from heterogeneous taxonomies within dataspaces
- TLD, a novel metric that captures the source categories mutual exclusivity

Future work

Conclusions

- semi-automatic, language independent approach to facets extraction from heterogeneous taxonomies within dataspaces
- TLD, a novel metric that captures the source categories mutual exclusivity
- evaluation shows that TLD outperforms state-of-the-art metrics

Future work

Conclusions

- semi-automatic, language independent approach to facets extraction from heterogeneous taxonomies within dataspaces
- TLD, a novel metric that captures the source categories mutual exclusivity
- evaluation shows that TLD outperforms state-of-the-art metrics

Future work

- improvement of the labelling phase (e.g., reconciliation with known Semantic Web ontologies)

Conclusions

- semi-automatic, language independent approach to facets extraction from heterogeneous taxonomies within dataspaces
- TLD, a novel metric that captures the source categories mutual exclusivity
- evaluation shows that TLD outperforms state-of-the-art metrics

Future work

- improvement of the labelling phase (e.g., reconciliation with known Semantic Web ontologies)
- integration of evidence coming from additional input (e.g., user queries)

Thank You

Questions?

riccardo.porrini@disco.unimib.it
http://rporrini.info

Backup

Related Work

Facet extraction

- *document corpora* [St07, Da08, We13, Me13]
  focus on faceted hierarchies - specific for unstructured data
- search engines' *query logs* and *documents* [Li09, Pa09, Po11]
  user search queries as a primary source of information
- search engines' *query results* [Ya10, Do11, Ka12, Ko13]
  integrate and rank facets already present in web documents

Similarity-Relatedness between taxonomy categories

- Leacock and Chodorow similarity [Le98]
- Wu and Palmer similarity [Wu94]
- ...
  not designed for heterogeneous taxonomies

Evaluation - Number of discovered facets

	$\lvert\mathcal{F}_*^g\rvert$	LC	WP	TLD
Dogs and Cats Food	3	1	1	7
Grappe, Liquors, Aperitives	1	1	1	6
Wines	3	1	1	6
Beers	2	6	3	14
DVD Movies	2	2	1	3
Rings	4	1	2	7
Blu-Ray Movies	2	2	2	5
Musical Instruments	6	1	1	5
Ski and Snowboards	1	1	1	7
Necklaces	8	2	3	11

References

[Fr05] Franklin et al. From Databases to Dataspaces: A New Abstraction for Information Management. *ACM SIGMOD Record*, 2005

[Ta04] Taylor et al. Wynars introduction to cataloging and classification. *Libraries Unlimited*, 2004

[Da08] Dakka et al. Automatic extraction of useful facet hierarchies from text databases. *ICDE*, 2008

[Me13] Medelyan et al. Constructing a focused taxonomy from a document collection. *ESWC*, 2013

[St07] Stoica et al. Automating creation of hierarchical faceted metadata structures. *HLT-NAACL*, 2007

[We13] Wei et al. Dft-extractor: a system to extract domain-specific faceted taxonomies from wikipedia. *WWW*, 2013

[Li09] Li et al. Extracting structured information from user queries with semi-supervised conditional random fields. *SIGIR*, 2009

[Pa09] Pasca et al. Web-derived resources for web information retrieval: from conceptual hierarchies to attribute hierarchies. *SIGIR*, 2009

[Po11] Pound et al. Facet discovery for structured web search: a query-log mining approach. *SIGMOD*, 2011

[Do11] Dou et al. Finding dimensions for queries. *CIKM*, 2011

[Ka12] Kawano et al. On-the-fly generation of facets as navigation signs for web objects. *DASFAA*, 2012

[Ko13] Kong et al. Extracting query facets from search results. *SIGIR*, 2013

[Ya10] Yan et al. Facetedpedia: enabling query-dependent faceted search for wikipedia. *CIKM*, 2010

[Le98] Leacock et al. Combining local context and wordnet similarity for word sense identification. *MIT Press*, 1998

[Wu94] Wu et al. Verb semantics and lexical selection. *ACL*, 1994

[Es96] Ester et al. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD*, 1996