

On Learning in \mathcal{AL} -log

Francesca A. Lisi
lisi@di.uniba.it



Floriana Esposito
esposito@di.uniba.it

Dipartimento di Informatica
Università degli Studi di Bari
Via Orabona, 4 - 70126 Bari - Italy



CILC 2005, Roma, 21-22 Giugno 2005

Overview

⌘ Background

- ☒ The hybrid system \mathcal{AL} -log

- ☒ Inductive Logic Programming

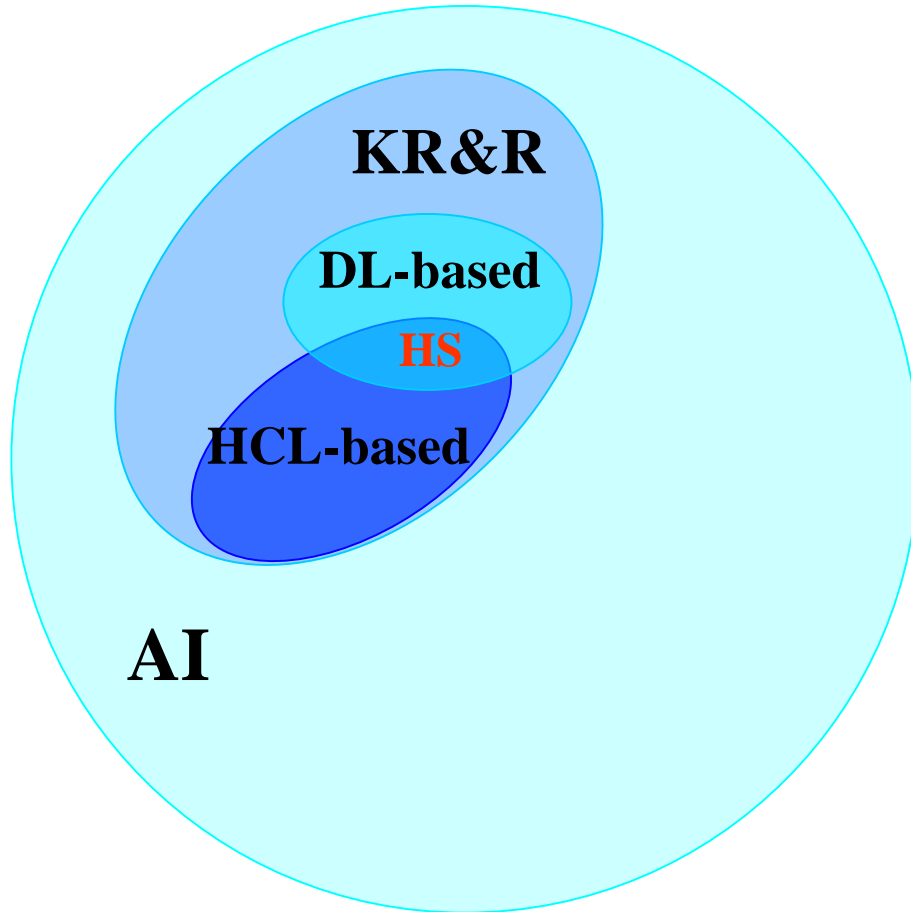
⌘ The general framework

⌘ An instantiation of the framework

⌘ Conclusions



Hybrid systems: combining HCL and DLs



- ⌘ Horn Clausal Logic (HCL) and Description Logics (DLs) are fragments of first-order logic
- ⌘ They can not be compared wrt expressive power but can be combined to obtain more expressive KR&R systems
 - ⊠ $\mathcal{AL}\text{-log} = \mathcal{ALC} + \text{Datalog}$ (Donini et al., 1998)
 - ⊠ CARIN (Levy & Rousset, 1998)



The hybrid system \mathcal{AL} -log: syntax

$$\mathcal{B} = \langle \Sigma, \Pi \rangle$$

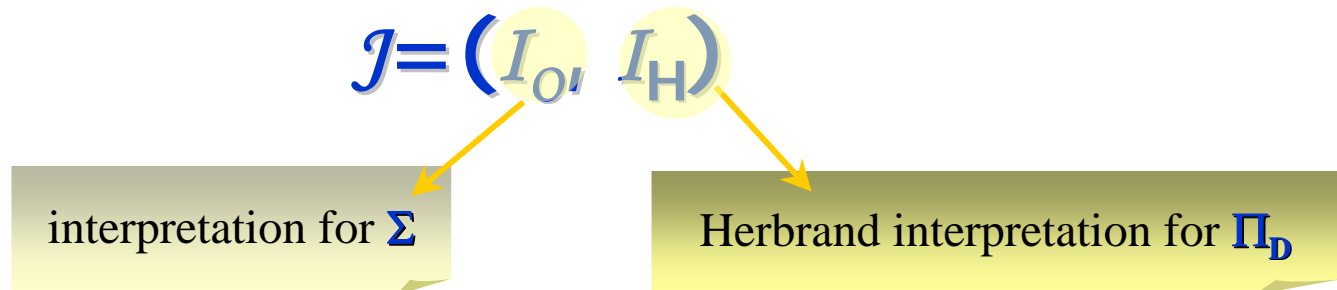
equality axioms $C \equiv D$
inclusion axioms $C \subseteq D$
concept assertions $a:C$
role assertions $\langle a, b \rangle : R$

constrained Datalog clauses
 $\alpha_0 \leftarrow \alpha_1, \dots, \alpha_m \ \& \ \gamma_1, \dots, \gamma_n$
where γ_j must be
a \mathcal{ALC} concept assertion

- ⌘ MiddleEastCountry \equiv AsianCountry $\cap \exists$ Hosts.MiddleEasternEthnicGroup
- ⌘ AsianCountry \subseteq Country
- ⌘ 'IR':AsianCountry, \langle 'IR','Arab' \rangle :Hosts, 'Arab':MiddleEasternEthnicGroup
- ⌘ believes(CountryID, ReligionN) \leftarrow religion(CountryID, ReligionN,Perc)
& CountryID:Country, ReligionN: Religion
- ⌘ religion('IR', 'Shia', 89)



The hybrid system \mathcal{AL} -log: semantics



⌘ \mathcal{J} satisfies \mathcal{B} iff

⊗ it satisfies Σ , and

⊗ for each clause $\alpha_0 \leftarrow \alpha_1, \dots, \alpha_m \ \& \ \gamma_1, \dots, \gamma_n$, for each of its ground instances $\alpha'_0 \leftarrow \alpha'_1, \dots, \alpha'_m \ \& \ \gamma'_1, \dots, \gamma'_n$, either there exists one γ'_i , $1 \leq i \leq n$, that is not satisfied by \mathcal{J} or $\alpha'_0 \leftarrow \alpha'_1, \dots, \alpha'_m$ is satisfied by \mathcal{J}

⌘ OWA of \mathcal{ALC} and CWA of Datalog do not interfere (safe interaction)



The hybrid system \mathcal{AL} -log: reasoning

Consistency of Σ

⌘ Tableau calculus

- ⊞ instance checks ($a:C$ wrt Σ ?)
- ⊞ subsumption checks ($C \sqsubseteq D$ wrt Σ ?)

Query answering wrt Π

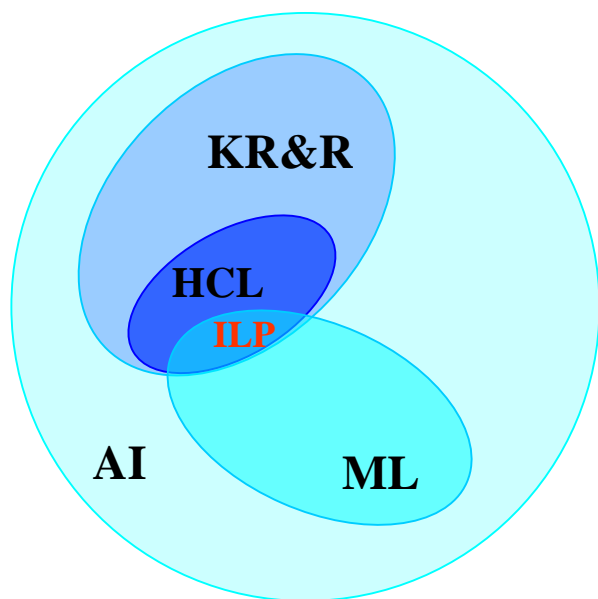
⌘ constrained SLD-resolution (complete by refutation)

- ⊞ Tries to obtain the empty constrained Datalog clause (= only constraints) by applying SLD-resolution
- ⊞ Verifies that constraints in the empty constrained Datalog clause are consistent wrt Σ by applying tableau calculus

⌘ decidable and closed-world reasoning!



Inductive Logic Programming



⌘ *Originally* Induction of rules from observations and background knowledge within the representation framework of Horn clausal logic (Muggleton, 1990)

⌘ scope of induction: discrimination

⌘ task: prediction

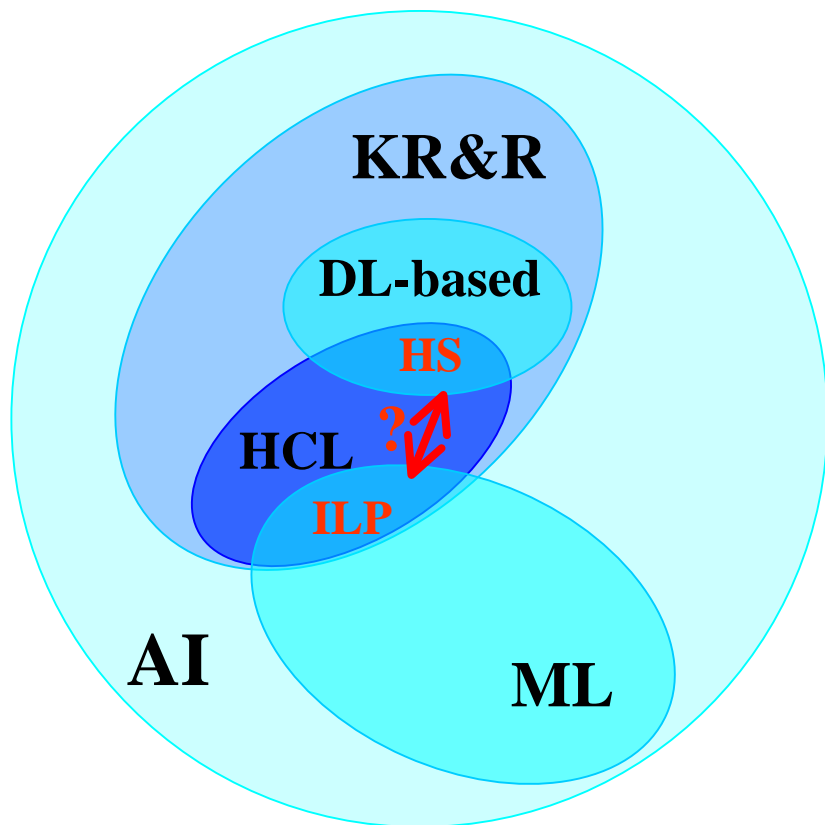
⌘ *Currently* Induction of rules from observations and background knowledge within the representation framework of first-order logic (fragments)

⌘ scope of induction: discrimination/characterization

⌘ task: prediction/description



Inductive Logic Programming



... with DL-HCL hybrid systems

- ⌘ Frisch, *Sorted downward refinement: Building background knowledge into a refinement operator for inductive logic programming*, ILP'99
- ⌘ Rouveirol and Ventos, *Towards learning in CARIN- \mathcal{ALN}* , ILP'00
- ⌘ Kietz, *Learnability of description logic programs*, ILP'02



Learning in \mathcal{AL} -log

Goal

To define a general framework
(i.e. independent from the scope of induction)
for the induction of rules from

- ⌘ a set $O = \{o_i\}$ of observations, and
- ⌘ a background knowledge \mathcal{K} , $\mathcal{K} \cap O = \emptyset$,

within the representation framework of \mathcal{AL} -log



Overview



⌘ Background

⌘ **The general framework**

☑ The language of hypotheses

☑ The generality order

☑ The coverage relations

⌘ An instantiation of the framework

⌘ Conclusions



The language of hypotheses

Hypotheses as **constrained Datalog clauses** compliant with:

⌘ **linkedness**

☑ $h(X) \leftarrow b(X,Y) \ \& \ X:C, Y:D$ **linked**

☑ $h(X) \leftarrow b(X,Y), c(Z) \ \& \ X:C, Y:D$ **not linked**

⌘ **connectedness**

☑ $h(X) \leftarrow b(X,Y) \ \& \ X:C, Y:D$ **connected**

☑ $h(Z) \leftarrow b(X,Y) \ \& \ X:C, Y:D$ **not connected**

⌘ **Object Identity (OI)**

☑ *In a formula terms denoted with different symbols represent different entities of the domain (Semeraro et al., 1998)*

☑ Extension of the Unique Names Assumption from the semantics of \mathcal{ALC} to the syntax of $\mathcal{AL}\text{-log}$ -> **OI-substitution**



The generality order

\mathcal{B} -subsumption: A model-theoretic definition

⌘ $\mathcal{B} = \mathcal{K} \cup \mathcal{O}$

⌘ $H_1, H_2 \in \mathcal{L}$

H_1 \mathcal{B} -subsumes H_2 if

⌘ for every model \mathcal{J} of \mathcal{B} and

⌘ every ground atom α such that H_2 covers α under \mathcal{J} ,
we have that H_1 covers α under \mathcal{J} .

H_1 is at least as general as H_2 under \mathcal{B} -subsumption, $H_1 \geq_{\mathcal{B}} H_2$,
iff H_1 \mathcal{B} -subsumes H_2



The generality order

\mathcal{B} -subsumption: A proof-theoretic definition

⌘ $\mathcal{B} = \mathcal{K} \cup \mathcal{O}$

⌘ $H_1, H_2 \in \mathcal{L}$

⌘ σ a Skolem substitution for H_2 w.r.t. $\{H_1\} \cup \mathcal{B}$

$H_1 \geq_{\mathcal{B}} H_2$ iff there exists a substitution θ for H_1 such that

⌘ $\text{head}(H_1)\theta = \text{head}(H_2)$

⌘ $\mathcal{B} \cup \text{body}(H_2)\sigma \vdash \text{body}(H_1)\theta\sigma$

⌘ $\text{body}(H_1)\theta\sigma$ is ground.

⌘ Checking $\geq_{\mathcal{B}}$ in \mathcal{AL} -log is decidable.



The coverage relations

Learning from implications

⌘ $H = q(\mathbf{X}) \leftarrow \text{body}(H) \in \mathcal{L}$

⌘ o_i is a ground constrained Datalog clause with head $q(\mathbf{a}_i)$

$H \in \mathcal{L}$ covers $o_i \in O$ under entailment w.r.t. \mathcal{K} iff

⌘ $\mathcal{K} \cup H \models o_i$

⌘ $\mathcal{K} \cup \text{body}(o_i) \cup H \models q(\mathbf{a}_i)$



The coverage relations

Learning from interpretations

⌘ $H = q(\mathbf{X}) \leftarrow \text{body}(H) \in \mathcal{L}$

⌘ $o_i = (q(\mathbf{a}_i), \mathcal{A}_i)$ where \mathcal{A}_i is a set of ground Datalog facts

$H \in \mathcal{L}$ covers $o_i \in \mathcal{O}$ under interpretations w.r.t. \mathcal{K} iff

⌘ $\mathcal{K} \cup \mathcal{A}_i \cup H \models q(\mathbf{a}_i)$

⌘ $\mathcal{K} \cup \mathcal{A}_i \cup H \not\models q(\mathbf{a}_i)$



Overview



⌘ Background

⌘ The general framework

⌘ **An instantiation of the framework**

☑ The setting

☑ The task

☑ The method

⌘ Conclusions



The setting

Characteristic induction from interpretations

- ⌘ \mathcal{L} a language of hypotheses
- ⌘ \mathcal{K} a background knowledge
- ⌘ $M(\mathcal{B})$ model constructed from $\mathcal{B} = \mathcal{K} \cup \mathcal{O}$
- ⌘ $o_i = (q(\mathbf{a}_i), \mathcal{A}_i)$ where \mathcal{A}_i is a set of ground Datalog facts

Finding $\mathcal{H} \subseteq \mathcal{L}$ such that

- ⌘ \mathcal{H} is true in $M(\mathcal{B})$ and
- ⌘ for each $H \in \mathcal{L}$, if H is true in $M(\mathcal{B})$ then $\mathcal{H} \models H$



The task

Frequent Pattern Discovery at l levels of description granularity, $1 \leq l \leq \max G$

Given

- ⌘ a data set \mathbf{r}
- ⌘ a taxonomy T where a reference concept C_{ref} and task-relevant concepts are designated
- ⌘ a multi-grained language $\mathcal{L} = \{\mathcal{L}^l\}_{1 \leq l \leq \max G}$ of patterns
- ⌘ a set $\{minsup^l\}_{1 \leq l \leq \max G}$ of support thresholds

Find the set F of all $P \in \mathcal{L}$ that are frequent in \mathbf{r}



The task

Evaluation function

A pattern $P \in \mathcal{L}^l$ with $\text{supp}(P, \mathbf{r}) = s$ is frequent in \mathbf{r} iff

⌘ $s \geq \text{minsup}^l$

⌘ all ancestors of P w.r.t. \mathcal{T} are frequent in \mathbf{r}

A pattern $Q \in \mathcal{L}^h$, $h < l$, is an ancestor of P iff it can be obtained from P by replacing each concept C occurring in P with a concept $D \in \mathcal{T}^h$ such that C is a sub-concept of D



An example of task

Find in:

⌘ the on-line CIA World Fact Book (*data set CIA*)

frequent patterns describing:

⌘ Middle East countries (*reference concept MiddleEastCountry*)

with respect to:

⌘ the religions believed (*task-relevant concepts* from the hierarchy rooted in **Religion**)

☑ e.g., the Muslim religion is a monotheistic religion

⌘ the languages spoken (*task-relevant concepts* from the hierarchy rooted in **Language**)

☑ e.g., the Indo-Iranian language is an Indo-European language

at three levels of description granularity (maxG=3)



The method

Levelwise search (Mannila & Toivonen, 1997)

If a generality order \geq over \mathcal{L} can be defined such that \geq is monotone w.r.t. *supp*, the lattice (\mathcal{L}, \geq) can be searched with a breadth-first strategy level by level of depth

⌘ generation-evaluation cycle

*Cycle on the **depth** ($k > 1$) of (\mathcal{L}, \geq)*

- ⊞ Generate candidate patterns from frequent $(k-1)$ - patterns
- ⊞ Generate frequent patterns from candidate k -patterns



... in our framework

The language \mathcal{L} of patterns

⌘ For each level of description granularity

$$\mathcal{L}' = \langle \mathcal{A}, \gamma, \Gamma' \rangle$$

⌘ Patterns as unary conjunctive queries called **O-queries**

$$Q = q(X) \leftarrow \alpha_1, \dots, \alpha_m \ \& \ X:C_{ref}, \gamma_1, \dots, \gamma_n$$

distinguished
variable



... in our framework

The language \mathcal{L} of patterns: an example

- ⌘ $\mathcal{A} = \{\text{believes}(+,-), \text{speaks}(+,-)\}$
- ⌘ $\gamma = \text{MiddleEastCountry}$
- ⌘ $\Gamma^1 = \{\text{Religion}, \text{Language}\}$
- ⌘ $\Gamma^2 = \{\text{MonotheisticReligion}, \dots, \text{IndoEuropeanLanguage}\}$
- ⌘ $\Gamma^2 = \{\text{MuslimReligion}, \dots, \text{IndoIranianLanguage}\}$

$Q_0 = q(X) \leftarrow \& X: \text{MiddleEastCountry}$ trivial 0-query

$Q_1 = q(X) \leftarrow \text{believes}(X,Y) \& X: \text{MiddleEastCountry}, Y: \text{Religion} \in \mathcal{L}^1$

$Q_2 = q(X) \leftarrow \text{believes}(X,Y), \text{speaks}(X,Z) \&$
 $X: \text{MiddleEastCountry}, Y: \text{MonotheisticReligion}, Z: \text{IndoEuropeanLanguage} \in \mathcal{L}^2$

$Q_3 = q(X) \leftarrow \text{believes}(X,Y), \text{speaks}(X,Z) \&$
 $X: \text{MiddleEastCountry}, Y: \text{MuslimReligion}, Z: \text{IndoIranianLanguage} \in \mathcal{L}^3$



... in our framework

The generality order

⌘ $\geq_{\mathcal{B}}$ is monotone w.r.t support

☐ if $P \geq_{\mathcal{B}} Q$ then $\text{supp}(P, \mathcal{B}) \geq \text{supp}(Q, \mathcal{B})$

Examples

$Q_1 = q(X) \leftarrow \text{believes}(X, Y) \ \& \ X:\text{MiddleEastCountry}, Y:\text{Religion}$

$Q_5 = q(X) \leftarrow \text{believes}(X, Y) \ \& \ X:\text{MiddleEastCountry}, Y:\text{MonotheisticReligion}$

$Q_6 = q(X) \leftarrow \text{believes}(X, Y), \text{believes}(X, Z) \ \& \ X:\text{MiddleEastCountry}, Y:\text{Religion}$

$Q_1 \geq_{\mathcal{B}} Q_5$ (but not viceversa due to **constrained SLD-resolution**)

$Q_1 \geq_{\mathcal{B}} Q_6$ (but not viceversa due to **the OI bias**)



The Semantic Web: layers and languages

ORL (OWL Rules Language)

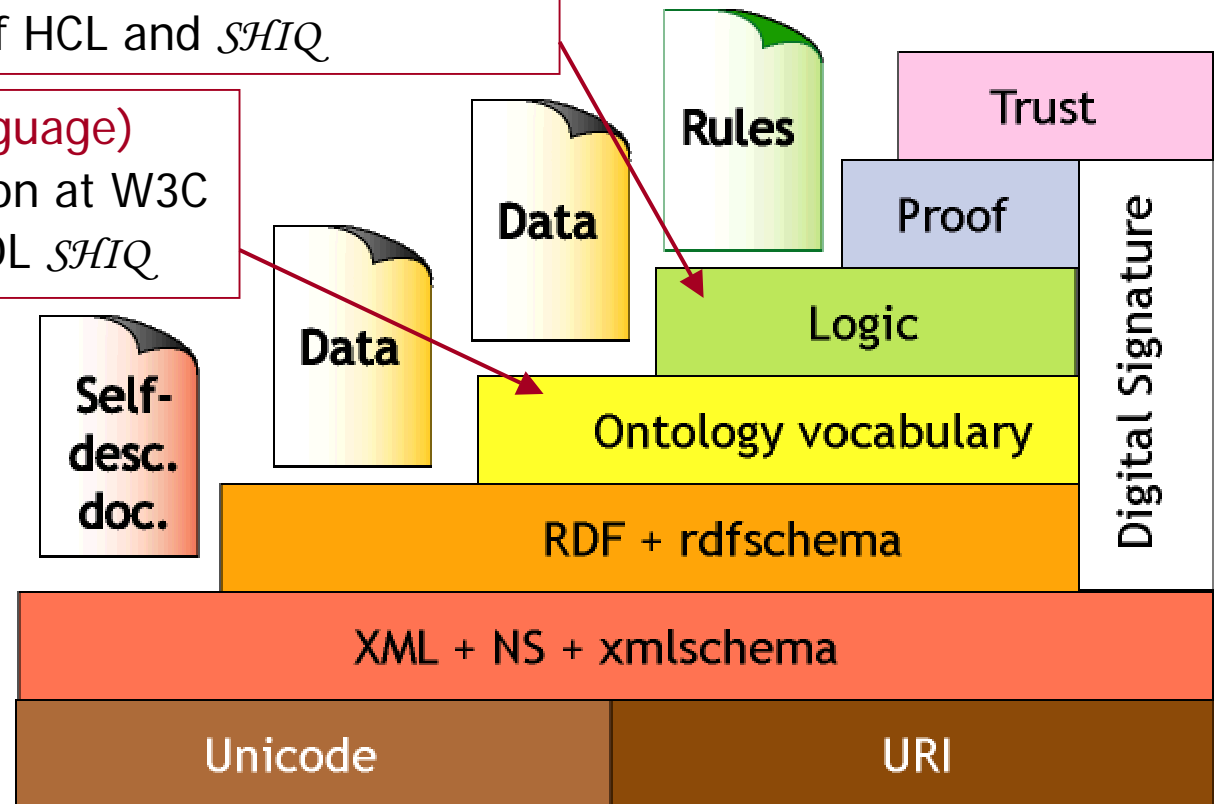
- ⌘ submitted to W3C for standardization
- ⌘ integration of HCL and *SHIQ*

OWL (Ontology Web Language)

- ⌘ ongoing standardization at W3C
- ⌘ as expressive as the DL *SHIQ*

Our view

- ⌘ $\mathcal{ALC} \subset \mathcal{SHIQ}$
- ⌘ Datalog \subset HCL
- ⌘ \mathcal{AL} -log is a good approximation of ORL



Semantic Web Mining: A scenario in ORL

⌘ **Input:** OWL expressions, e.g.

```
<owl:Class rdf:ID="MiddleEastCountry" />
<owl:Class rdf:ID="MuslimReligion">
  <rdfs:subClassOf rdf:resource="#MonotheisticReligion" />
</owl:Class>
<owl:Class rdf:ID="IndoIranianLanguage">
  <rdfs:subClassOf rdf:resource="#IndoEuropeanLanguage" />
</owl:Class>
<MiddleEastCountry rdf:ID="Iran" />
<MuslimReligion rdf:ID="Sunni" />
<IndoIranianLanguage rdf:ID="Persian" />
```

⌘ **Output:** ORL expressions, e.g. (see next two slides)



Semantic Web Mining: A scenario in ORL (2)

<owlx:Rule>

<owlx:antecedent>

<owlx:classAtom> <owlx:class="&MiddleEastCountry" /> <owlx:Variable owlx:name="_X" /> </owlx:classAtom>

<owlx:classAtom> <owlx:class="&MonotheisticReligion" /> <owlx:Variable owlx:name="_Y" /> </owlx:classAtom>

<owlx:classAtom> <owlx:class="&IndoEuropeanLanguage" /> <owlx:Variable owlx:name="_Z"/> </owlx:classAtom>

<owlx:individualPropertyAtom owlx:property="&believes">

<owlx:Variable owlx:name="_X" /> <owlx:Variable owlx:name="_Y" />

</owlx:individualPropertyAtom>

<owlx:individualPropertyAtom owlx:property="&speaks">

<owlx:Variable owlx:name="_X" /> <owlx:Variable owlx:name="_Z" />

</owlx:individualPropertyAtom>

</owlx:antecedent>

<owlx:consequent>

<owlx:individualPropertyAtom owlx:property="&q">

<owlx:Variable owlx:name="_X" />

</owlx:individualPropertyAtom>

</owlx:consequent>

</owlx:Rule>



- ⌘ MonotheisticReligion and IndoEuropeanLanguage belong to the 2nd layer of the input taxonomy T
- ⌘ This is a description of (a subset of) the individuals of MiddleEastCountry at the level l=2 of granularity!

Semantic Web Mining: A scenario in ORL (3)

<owlx:Rule>

<owlx:antecedent>

<owlx:classAtom> <owlx:class="&MiddleEastCountry" /> <owlx:Variable owlx:name="_X" /> </owlx:classAtom>

<owlx:classAtom> <owlx:class="&MuslimReligion" /> <owlx:Variable owlx:name="_Y" /> </owlx:classAtom>

<owlx:classAtom> <owlx:class="&IndolranianLanguage" /> <owlx:Variable owlx:name="_Z"/> </owlx:classAtom>

<owlx:individualPropertyAtom owlx:property="&believes">

<owlx:Variable owlx:name="_X" /> <owlx:Variable owlx:name="_Y" />

</owlx:individualPropertyAtom>

<owlx:individualPropertyAtom owlx:property="&speaks">

<owlx:Variable owlx:name="_X" /> <owlx:Variable owlx:name="_Z" />

</owlx:individualPropertyAtom>

</owlx:antecedent>

<owlx:consequent>

<owlx:individualPropertyAtom owlx:property="&q">

<owlx:Variable owlx:name="_X" />

</owlx:individualPropertyAtom>

</owlx:consequent>

</owlx:Rule>



- ⌘ MuslimReligion and IndolranianLanguage belong to the 3rd layer of the input taxonomy T
- ⌘ This is a description of (a subset of) the individuals of MiddleEastCountry at the level I=3 of granularity!

Conclusions

- ⌘ Hybrid KR&R systems integrating DLs and HCL are very interesting because they combine the expressive and deductive power of both
- ⌘ \mathcal{AL} -log is to be preferred to CARIN because it provides
 - ☑ Safe interaction between DL and HCL part
 - ☑ decidable and closed-world reasoning mechanisms
 - ☑ expressive and deductive power *enough* for the *actual* needs of application areas, e.g. the Semantic Web, that (can) use hybrid KR&R systems
- ⌘ Semantic Web Mining seems to be a killer application for learning in \mathcal{AL} -log

