Machine learning with biomedical ontologies

Presented by *Sarah Alghamdi, Azza Althagafi, Robert Hoehndorf, Maxat Kulmanov, Sumyyah Toonsi, Fernando Zhapa-Camacho*

Learning Outcomes

- Introduce Ontologies and the description logic

- Discuss unsupervised machine learning methods that can "embed" from one structure to another

- Introduce different methods that use ontologies in machine learning models

- Introduce mOWL, a software library for machine learning with ontologies

- Incorporate mOWL in Biomedical data analysis using different approaches

Preliminaries: What are Ontologies?

- " An ontology is a **logical theory** designed in order to capture the **intended models** corresponding to a certain conceptualization and to **exclude the unintended** ones" … Guarino 2009
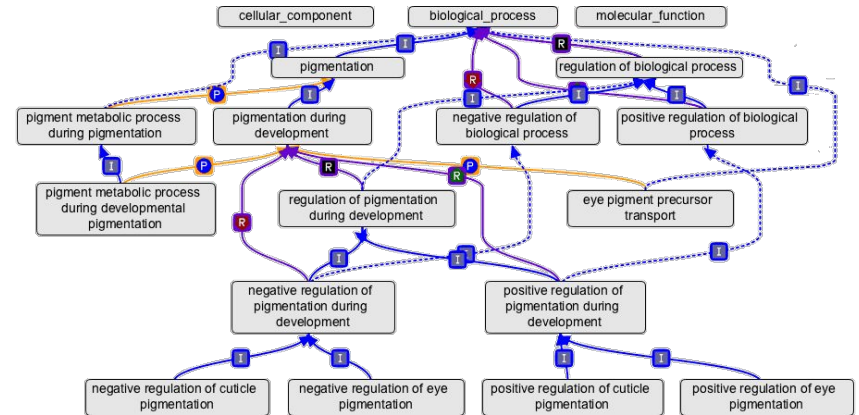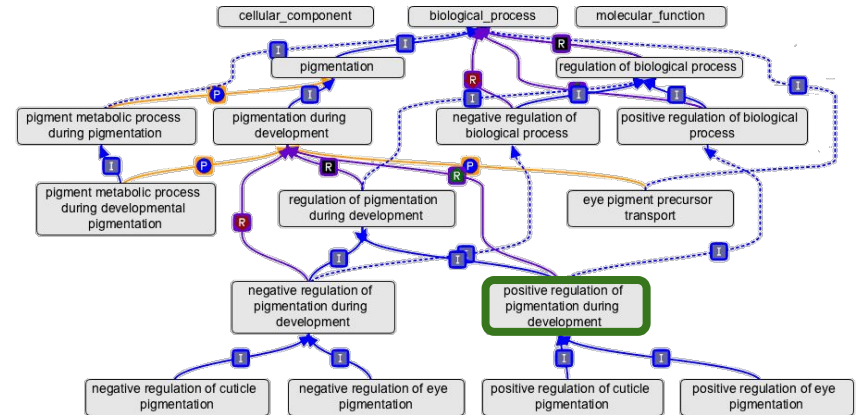
Preliminaries: What are Ontologies?

- " An ontology is a **logical theory** designed in order to capture the **intended models** corresponding to a certain conceptualization and to **exclude the unintended** ones" … Guarino 2009

Preliminaries: What are Ontologies?

- " An ontology is a **logical theory** designed in order to capture the **intended models** corresponding to a certain conceptualization and to **exclude the unintended** ones" … Guarino 2009

Preliminaries: What are Ontologies?

- " An ontology is a **logical theory** designed in order to capture the **intended models** corresponding to a certain conceptualization and to **exclude the unintended** ones" … Guarino 2009

GENEONTOLOGY

$positive regulation of developmental pigmentation \equiv biological process \sqcap \exists positively regulates.developmental pigmentation$

$positive regulation of developmental pigmentation \sqsubseteq \exists positively regulates.developmental pigmentation$

$positive regulation of developmental pigmentation \sqsubseteq \exists regulates.developmental pigmentation$

$positive regulation of developmental pigmentation \sqsubseteq biological process \sqcap \exists regulates.developmental pigmentation$

$positive regulation of developmental pigmentation \sqsubseteq \exists regulates.biological process$

$positive regulation of developmental pigmentation \sqsubseteq biological process \sqcap \exists regulates.biological process$
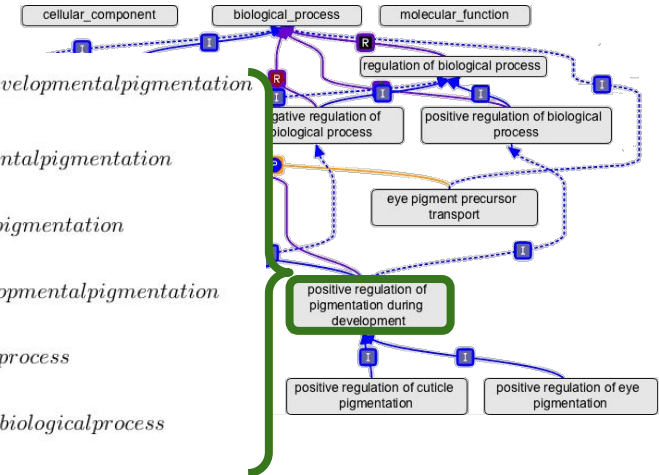
Preliminaries: What are Ontologies?

- " An ontology is a **logical theory** designed in order to capture the **intended models** corresponding to a certain conceptualization and to **exclude the unintended** ones" … Guarino 2009

- Classes and relations

- Standard identifiers

- Axioms and formal definitions

- Metadata:
  - Labels, Synonyms
  - database cross references
 - ….

GO - Gene Ontology

Overview | Browse | DLQuery | SPARQL | Download

Annotation	Value
label	positive regulation of developmental pigmentation
definition	Any process that increases the frequency, rate or extent of the developmental process that results in the deposition of coloring matter in an organism.
class	http://purl.obolibrary.org/obo/GO_0048087
ontology	GO
Equivalent	biological regulation and (positively regulates some developmental pigmentation)
SubClassOf	positively regulates some developmental pigmentation, positive regulation of biological process, regulation of developmental pigmentation
has_obo_namespace	biological_process
synonyms	upregulation of developmental pigmentation, up regulation of developmental pigmentation, stimulation of developmental pigmentation, activation of developmental pigmentation, up-regulation of developmental pigmentation
id	GO:0048087

Introduction

- Examples of ontologies application in biomedical domain:
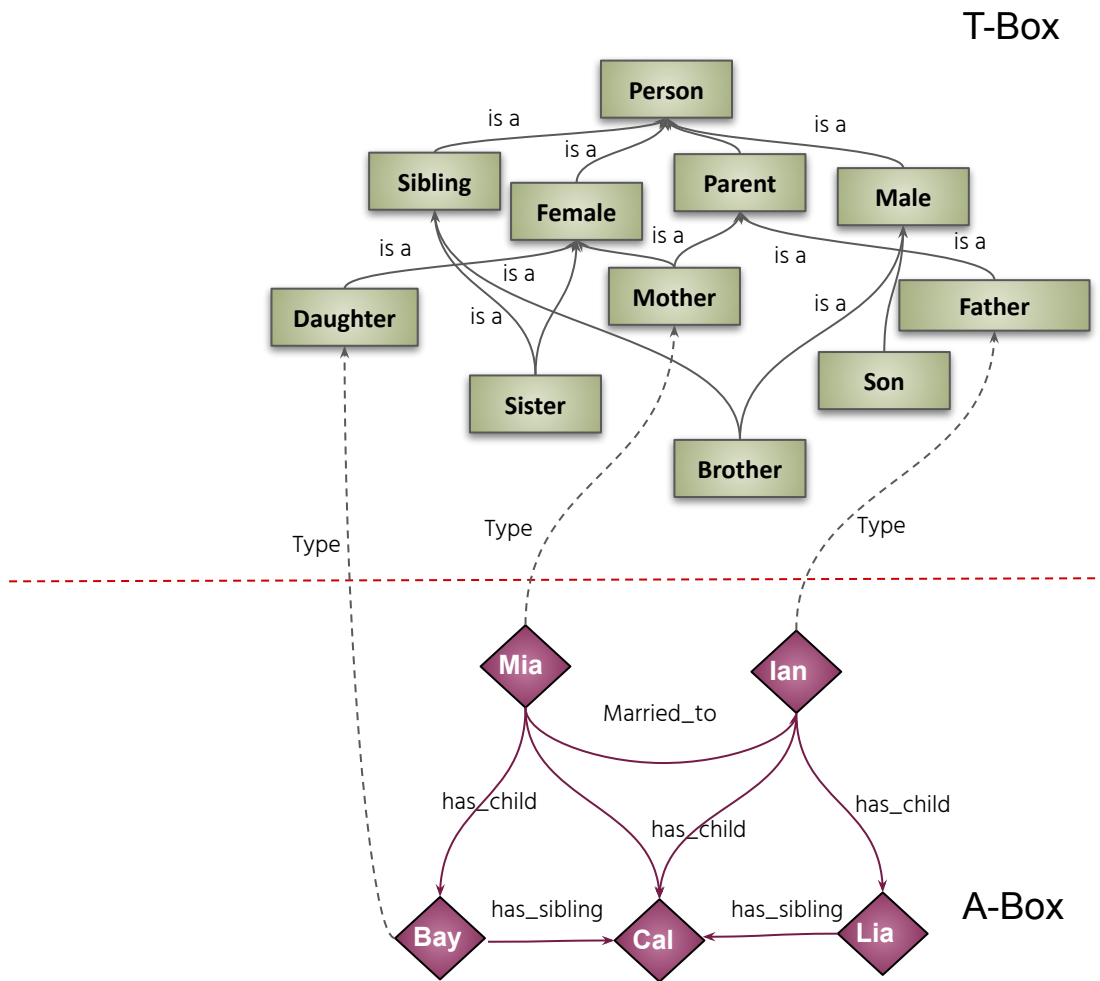
    - Annotation and data integration
    - Ontologies as vocabularies
    - Statistical and predictive data analysis :
        - Enrichment analysis
        - Semantic similarity
    - Regression analysis
    - Relation prediction
    - Classification
        - Supervised
        - Unsupervised
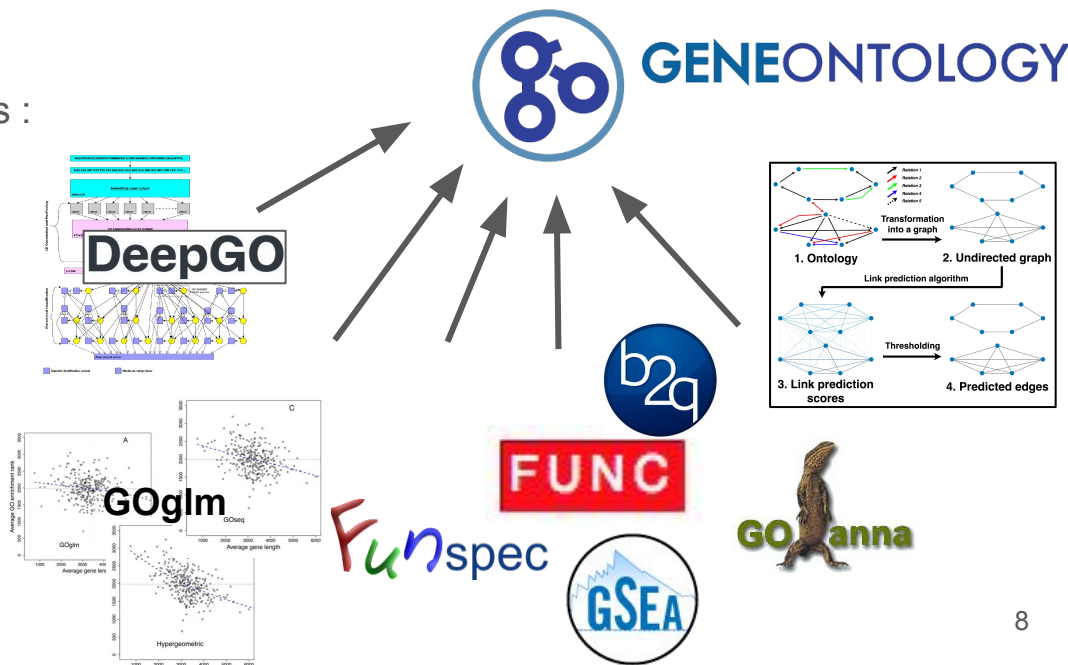
Preliminaries: ontologies

- Ontology consist of :

  O={C,R,I,⊢}

- T-Box
  - Set of terminological
- A-Box
  - Set of assertions

Preliminaries: ontologies

- Description Logic (DL) is used to formally and explicitly represent ontologies

Name	DL syntax	Semantics
Top concept	$\top$	$\Delta^{\mathcal{I}}$
Bottom concept	$\bot$	$\emptyset$
Concept	C	$C^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$
Concept disjunction	$C_1 \sqcap C_2$	$C_1^{\mathcal{I}} \cap C_2^{\mathcal{I}}$
Concept conjunction	$C_1 \sqcup C_2$	$C_1^{\mathcal{I}} \cup C_2^{\mathcal{I}}$
Concept negation	$\neg C$	$\Delta^{\mathcal{I}} \backslash C^{\mathcal{I}}$
Universal restriction	$\forall R.C$	$\{x \in \Delta^{\mathcal{I}} \mid \forall y \in \Delta^{\mathcal{I}}((x,y) \in R^{\mathcal{I}} \wedge y \in C^{\mathcal{I}})\}$
Existential restriction	$\exists R.C$	$\{x \in \Delta^{\mathcal{I}} \mid \exists y \in \Delta^{\mathcal{I}}((x,y) \in R^{\mathcal{I}} \rightarrow y \in C^{\mathcal{I}})\}$
Subclass of	$C_1 \sqsubseteq C_2$	$C_1^{\mathcal{I}} \subseteq C_2^{\mathcal{I}}$
Subproperty of	$R_1 \sqsubseteq R_2$	$R_1^{\mathcal{I}} \subseteq R_2^{\mathcal{I}}$
Equivalent class	$C_1 \equiv C_2$	$C_1^{\mathcal{I}} = C_2^{\mathcal{I}}$
Equivalent property	$R_1 \equiv R_2$	$R_1^{\mathcal{I}} = R_2^{\mathcal{I}}$

Classes , Relations

$Person \sqsubseteq \top$
$Female \sqcap Male \sqsubseteq \bot$
$Female \sqcup Male \sqsubseteq \top$
$Female \equiv \neg Male$
$Parent \equiv \exists has\_child. Person$
$Son \sqsubseteq Male \sqcap \exists child\_of. Person$
$Mother \sqsubseteq Female \sqcap Parent$
$Sibling \sqsubseteq \exists has\_sibling. Person$
$has\_brother \sqsubseteq has\_sibling$

Preliminaries: ontologies

- Description Logic (DL) is used to formally and explicitly represent ontologies

DL Syntax	Manchester Syntax
$C \sqcap D$	C and D
$C \sqcup D$	C or D
$\neg C$	not C
$\exists R.C$	R some C
$\forall R.C$	R only C
$(\geq nR.C)$	R min n C
$(\leq nR.C)$	R max n C
$(= nR.C)$	R exactly n C
$\{a\} \sqcup \{b\} \sqcup \ldots$	{a b …}

Classes , Relations

$Person \sqsubseteq \top$
$Female \sqcap Male \sqsubseteq \bot$
$Female \sqcup Male \sqsubseteq \top$
$Female \equiv \neg Male$
$Parent \equiv \exists has\_child. Person$
$Son \sqsubseteq Male \sqcap \exists child\_of. Person$
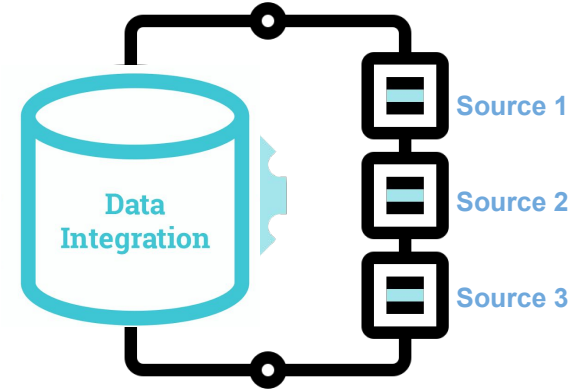$Mother \sqsubseteq Female \sqcap Parent$
$Sibling \sqsubseteq \exists has\_sibling. Person$
$has\_brother \sqsubseteq has\_sibling$

How Ontologies are used in Databases

- Annotations and data integration
  - Ontologies play a crucial role in facilitating data integration across databases due to their usage of standard identifiers for classes and relations

How Ontologies are used in Databases

- Annotations and data integration
  - Ontologies play a crucial role in facilitating data integration across databases due to their usage of standard identifiers for classes and relations

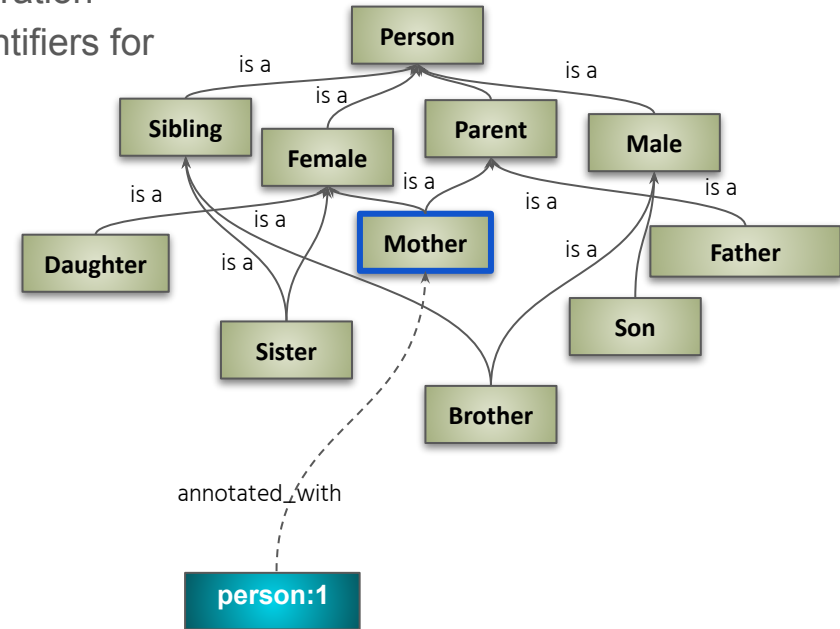# How Ontologies are used in Databases

- Annotations and data integration

GAF fields

The annotation flat file format is comprised of 17 tab-delimited fields.

Column	Content	Required?	Cardinality	Example	
1	DB	required	1	UniProtKB	
2	DB Object ID	required	1	P12345	
3	DB Object Symbol	required	1	PHO3	
4	Qualifier	required	1 or 2	NOTIinvolved_in	
5	GO ID	required	1	GO:0003993	
6	DB:Reference (IDB:Reference)	required	1 or greater	PMID:2676709	
7	Evidence Code	required	1	IMP	
8	With (or) From	optional	0 or greater	GO:0000346	
9	Aspect	required	1	F	
10	DB Object Name	optional	0 or 1	Toll-like receptor 4	
11	DB Object Synonym (ISynonym)	optional	0 or greater	hToll	Tollbooth
12	DB Object Type	required	1	protein	
13	Taxon(Itaxon)	required	1 or 2	taxon:9606	
14	Date	required	1	20090118	
15	Assigned By	required	1	SGD	
16	Annotation Extension	optional	0 or greater	part_of(CL:0000576)	
17	Gene Product Form ID	optional	0 or 1	UniProtKB:P12345-2	

How Ontologies are used in Databases

- Annotations and data integration

1. UniProtKB
2. A0A024RBG1
3. NUDT4B
4. enables
5. GO:0003723
6. GO_REF:0000043
7. IEA
8. UniProtKB-KW:KW-0694
9. F
10. Diphosphoinositol polyphosphate phosphohydrolase
11. NUDT4B
12. NUDT4B
13. Protein
14. taxon:9606   20221109
15. UniProt
16.
17.

1. MGI
2. MGI:1913300
3. 0610009B22Rik
4. enables
5. GO:0001222
6. MGI:MGI:4834177|GO_REF:0000096
7. ISO
8. UniProtKB:P0DI82
9. F
10. RIKEN cDNA 0610009B22 gene
11.
12. protein_coding_gene
13. taxon:10090
14. 20210709
15. MGI
16.
17.

GAF fields

The annotation flat file format is comprised of 17 tab-delimited fields.

Column	Content	Required?	Cardinality	Example	
1	DB	required	1	UniProtKB	
2	DB Object ID	required	1	P12345	
3	DB Object Symbol	required	1	PHO3	
4	Qualifier	required	1 or 2	NOTIinvolved_in	
5	GO ID	required	1	GO:0003993	
6	DB:Reference (IDB:Reference)	required	1 or greater	PMID:2676709	
7	Evidence Code	required	1	IMP	
8	With (or) From	optional	0 or greater	GO:0000346	
9	Aspect	required	1	F	
10	DB Object Name	optional	0 or 1	Toll-like receptor 4	
11	DB Object Synonym (ISynonym)	optional	0 or greater	hToll	Tollbooth
12	DB Object Type	required	1	protein	
13	Taxon(Itaxon)	required	1 or 2	taxon:9606	
14	Date	required	1	20090118	
15	Assigned By	required	1	SGD	
16	Annotation Extension	optional	0 or greater	part_of(CL:0000576)	
17	Gene Product Form ID	optional	0 or 1	UniProtKB:P12345-2	

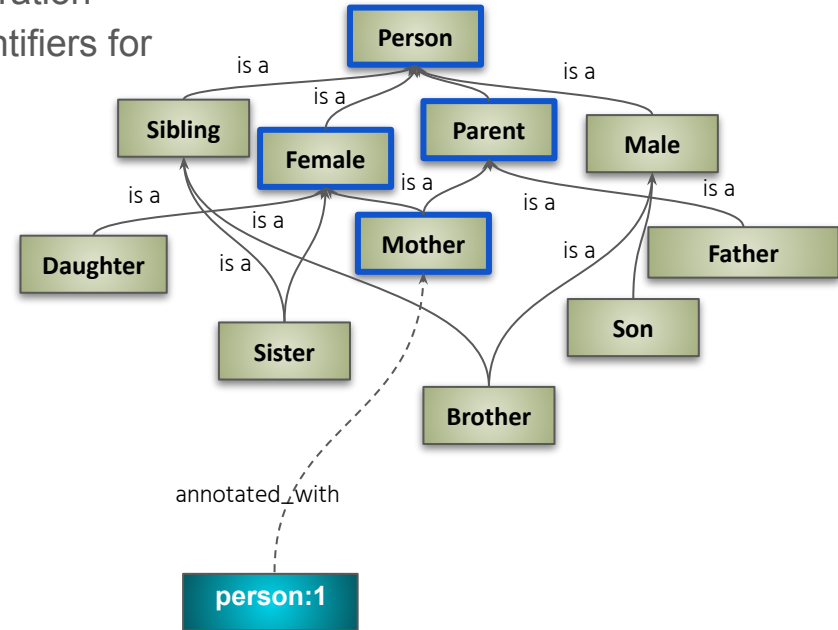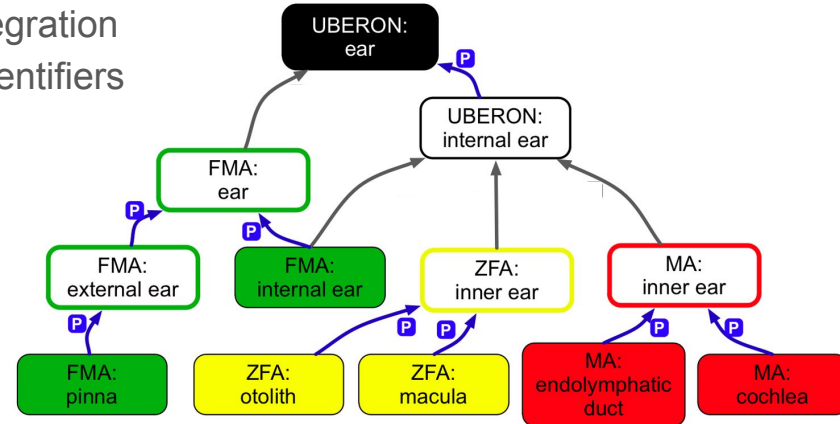# How Ontologies are used in Databases

- Annotations and data integration
  - Ontologies play a crucial role in facilitating data integration across databases due to their usage of standard identifiers for classes and relations

- True path rule:
  - Annotation for a class is passed to its ancestors

How Ontologies are used in Databases

- Annotations and data integration
  - Ontologies play a crucial role in facilitating data integration across databases due to their usage of standard identifiers for classes and relations

- True path rule:
  - Annotation for a class is passed to its ancestors

How Ontologies are used in Databases

- Annotations and data integration
  - Ontologies play a crucial role in facilitating data integration across databases due to their usage of standard identifiers for classes and relations

- True path rule:
  - Annotation for a class is passed to its ancestors

[1]Washington, N. L., Haendel, et al. (2009). Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS biology*, 7(11), e1000247.
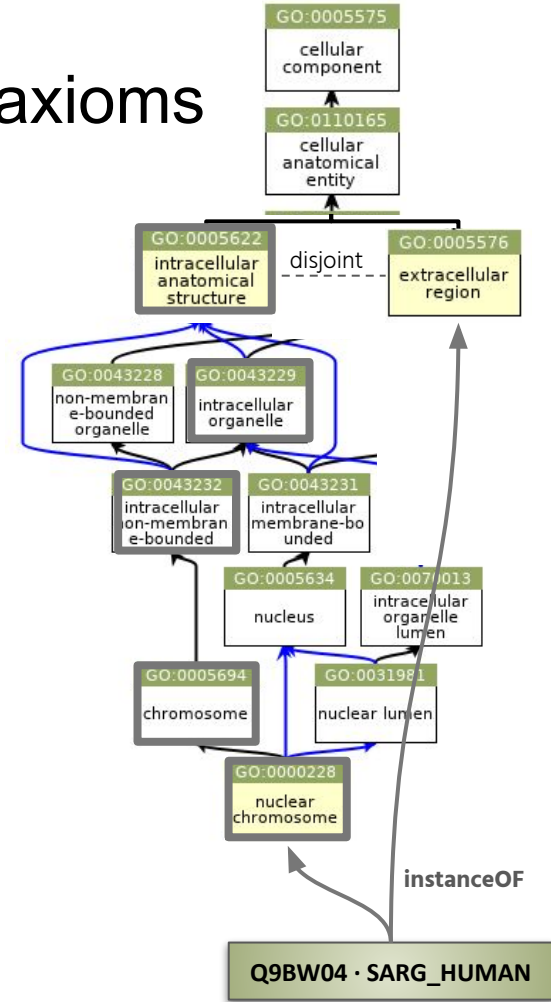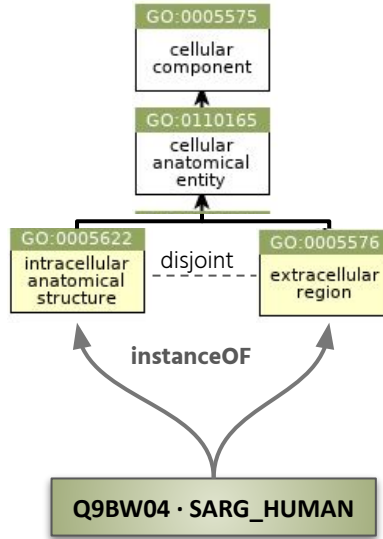
Transforming GO annotations to ontology axioms

- Problem when annotations are added as instances of GO classes
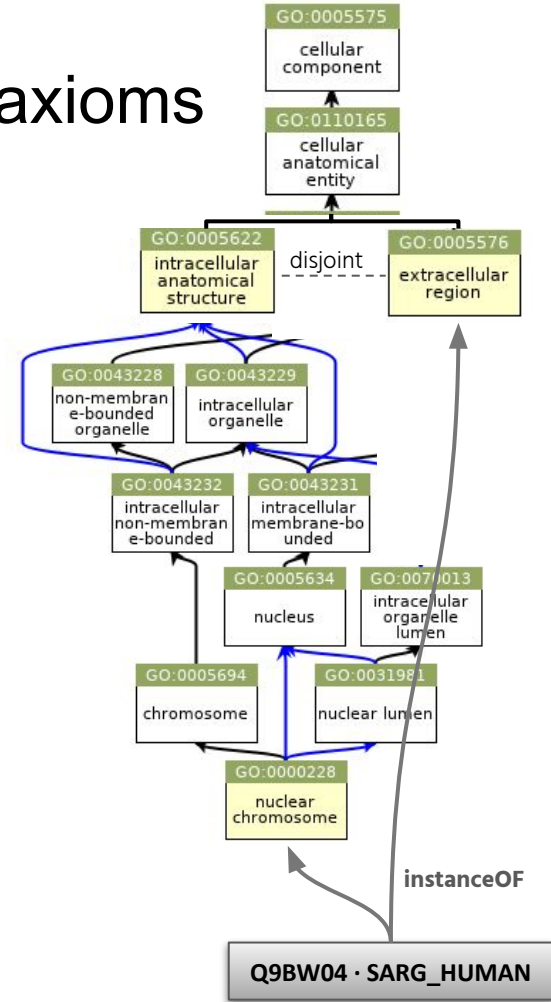
Transforming GO annotations to ontology axioms

- Problem when annotations are added as instances of GO classes

Transforming GO annotations to ontology axioms

- Problem when annotations are added as instances of GO classes
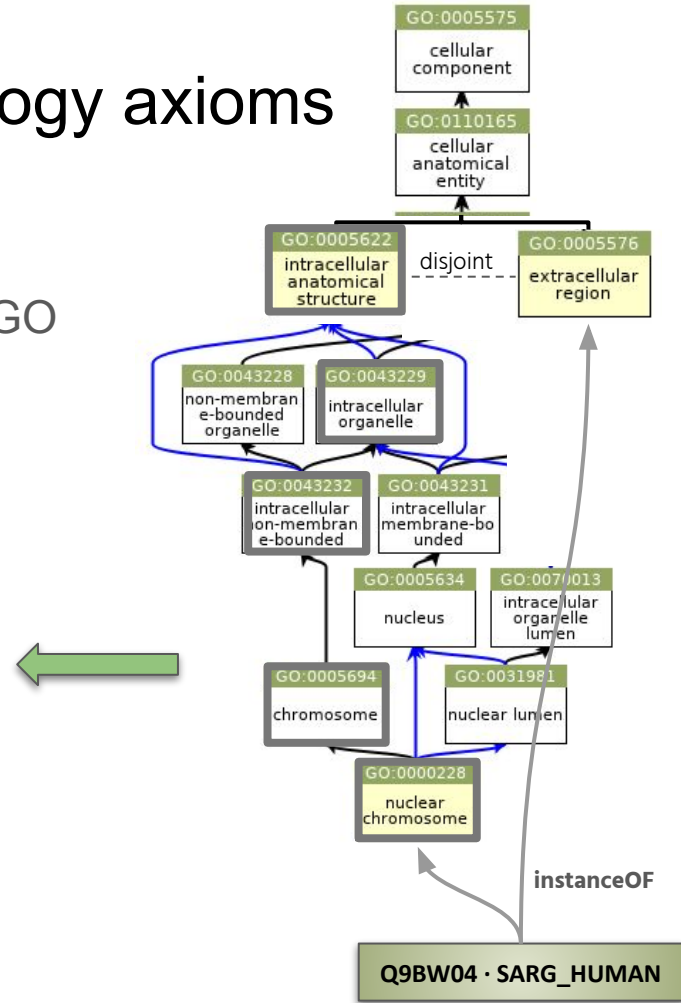
Transforming GO annotations to ontology axioms

- Annotations to T-Box

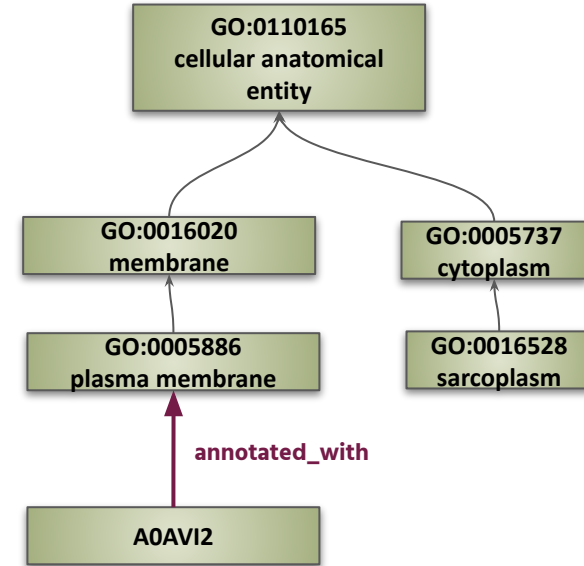 The annotated entity **C** is added as a class to the ontology

  The annotation is added a relation R as follows

  $$\mathbf{C} \sqsubseteq \exists R. \mathbf{D}$$

Example:

Annotating protein **A0AVI2** To Gene Ontology

- ○ **A0AVI2** $\sqsubseteq \exists$ *annotated_with.* ***Plasma membrane***
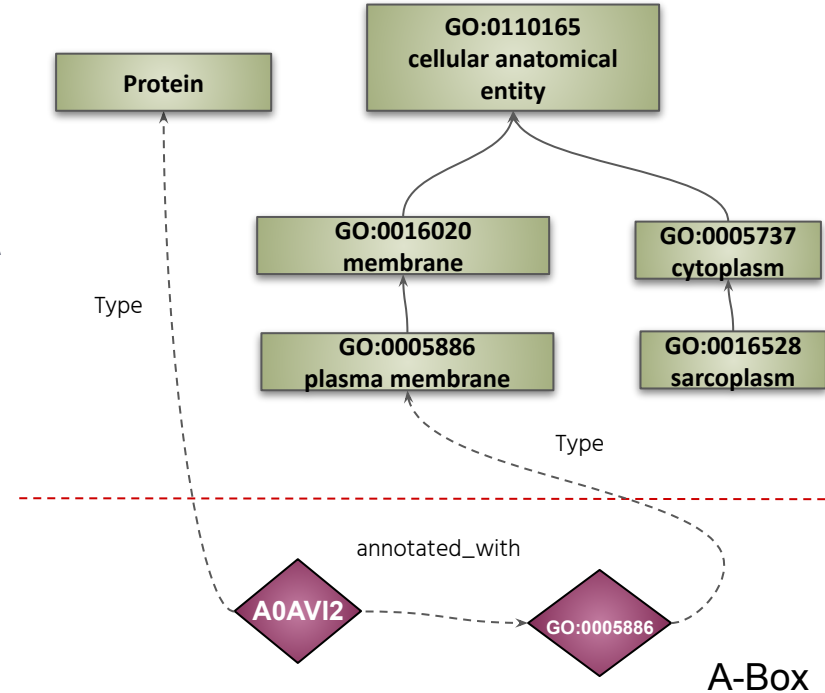
Following this

- Ontologies and text mining → utilizing class labels and descriptions

- Graph based embedding → utilizing axioms and class labels and descriptions

- Syntactic embedding → utilizing axioms and class labels and descriptions

- Semantic embedding → utilizing axioms