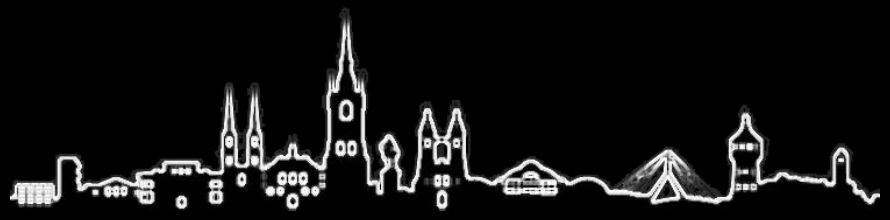


# ODLS 2016

Workshop on Ontologies and Data in Life Sciences,
Sep 29-30, Halle (Saale), Germany

# Ontological interpretation of biomedical database annotations

Filipe Santana da Silva[1], Ludger Jansen[2],
Fred Freitas[1] , Stefan Schulz[3]

[1]Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE), Recife, Brazil
[2]Institut für Philosophie, Universität Rostock, Germany
[3]Institut für Medizinische Informatik, Statistik und Dokumentation, Medizinische Universität Graz, Austria

# Biological Databases and Bio-Ontologies

- Two worlds: Bio-DBs


Bio-ontologies


- How are they related to each other?
- Can their content be expressed by a unified model of meaning?
- Is database content of ontological nature?
- Can OWL be used as a language to express
  - bio-ontology content
  - bio-database structure
  - bio-database content
- Which is the added value?

# Biological Databases and Bio-Ontologies

- Two worlds: Bio-DBs


- Store summarized results of laboratory experiments
- Classical database structure
- Values:
  - Numeric
  - Textual
  - Symbolic (codes from ontologies)

Bio-ontologies


- Provide definitions
- Provide axioms that are universally true
- Obey formal semantics
- Main use case:
  - annotation of biological database entries

# Example – tabular Bio-DB structure

| ID | PR Protein | PR Gene | Organism (NCBI Tax) | GO Biological Process | GO Molecular Function | GO Cellular componente | Ensembl ID | Ensembl Phenotype |
|---|---|---|---|---|---|---|---|---|
| F1MEW4 | CBS | CBS | *Bos tautus* | blood vessel remodelling;... | Cysthationine beta-synthase;... | cytoplasm;... | ENSBTAT00000000184; ... | No phenotype associated |
| Q99707 | MS | MS | *Homo sapiens* | cobalamin metabolic process;... | cobalamin binding;... | cytoplasm;... | ENST000000366577; ENST000000533889 | Neural tube defect.; Megaloblastic anemia;... |
| F1RF82 | MTHFR | MTHFR | *Sus scrofa* | homocysteine metabolic process;... | modified amino acid binding;... | cytosol | ENSSSCT00000003805 | No phenotype associated |
| Q93088 | BHMT | BHMT | *Homo sapiens* | amino acid betaine catabolic process;... | zinc ion binding;... | protein complex;... | ENST00000274353 | Liver tumor; Coronary artery disease;... |

# Example – tabular Bio-DB structure



| ID | PR Protein | PR Gene | Organism (NCBI Tax) | GO Biological Process | GO Molecular Function | GO Cellular componente | Ensembl ID | Ensembl Phenotype |
|---|---|---|---|---|---|---|---|---|
| F1MEW4 | CBS | CBS | *Bos tautus* | blood vessel remodelling;... | Cysthationine beta-synthase;... | cytoplasm;... | ENSBTAT00000000184; ... | No phenotype associated |
| Q99707 | MS | MS | *Homo sapiens* | cobalamin metabolic process;... | cobalamin binding;... | cytoplasm;... | ENST000000366577; ENST000000533889 | Neural tube defect.; Megaloblastic anemia;... |
| F1RF82 | MTHFR | MTHFR | *Sus scrofa* | homocysteine metabolic process;... | modified amino acid binding;... | cytosol | ENSSSCT00000003805 | No phenotype associated |
| Q93088 | BHMT | BHMT | *Homo sapiens* | amino acid betaine catabolic process;... | zinc ion binding;... | protein complex;... | ENST00000274353 | Liver tumor; Coronary artery disease;... |

'is included in'

'is included in'

'is included in '

'is included in'

'is participant in'

'is part of'

'includes'

# Example – tabular Bio-DB structure

more abstract:

- a database record informs about experimental evidence that:
  - Proteins of the type *Prot1*
    - participate in Processes of type $BProc_1$… $BProc_k$ within organisms of type $Org_1$
    - are active in cellular components of type $CComp_1$ or $CComp_2$ or … $CComp_x$ within organisms of type $Org_1$
    - participate in Processes that have small molecules of type Mol1, $Mol_2$…$Mol_y$ as outcome
    - within organisms of type $Org_1$ – if dysfunctional – $Org_1$ has dispositions to develop the phenotypes (disorders) $Phen_1$…$Phen_z$

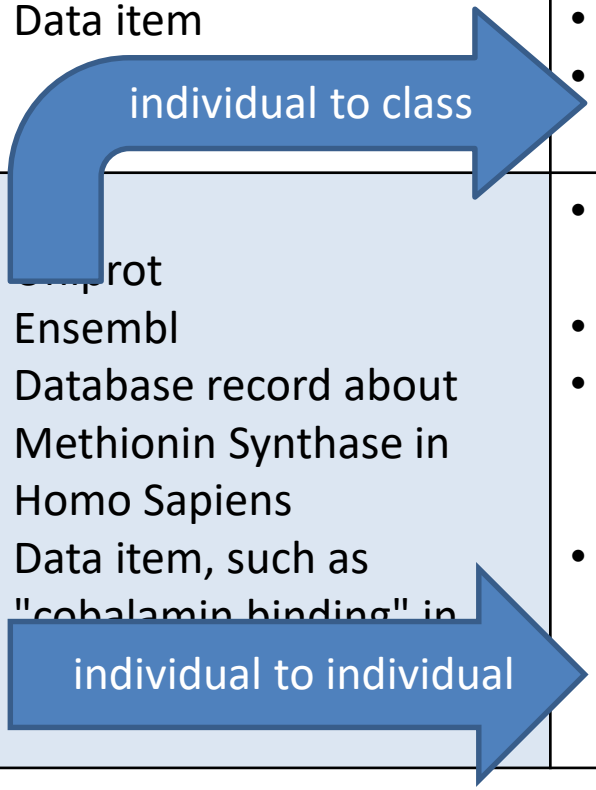
This information is not explicitly contained in the database – it is implicitly shared by database users and curators

# Ontological framework

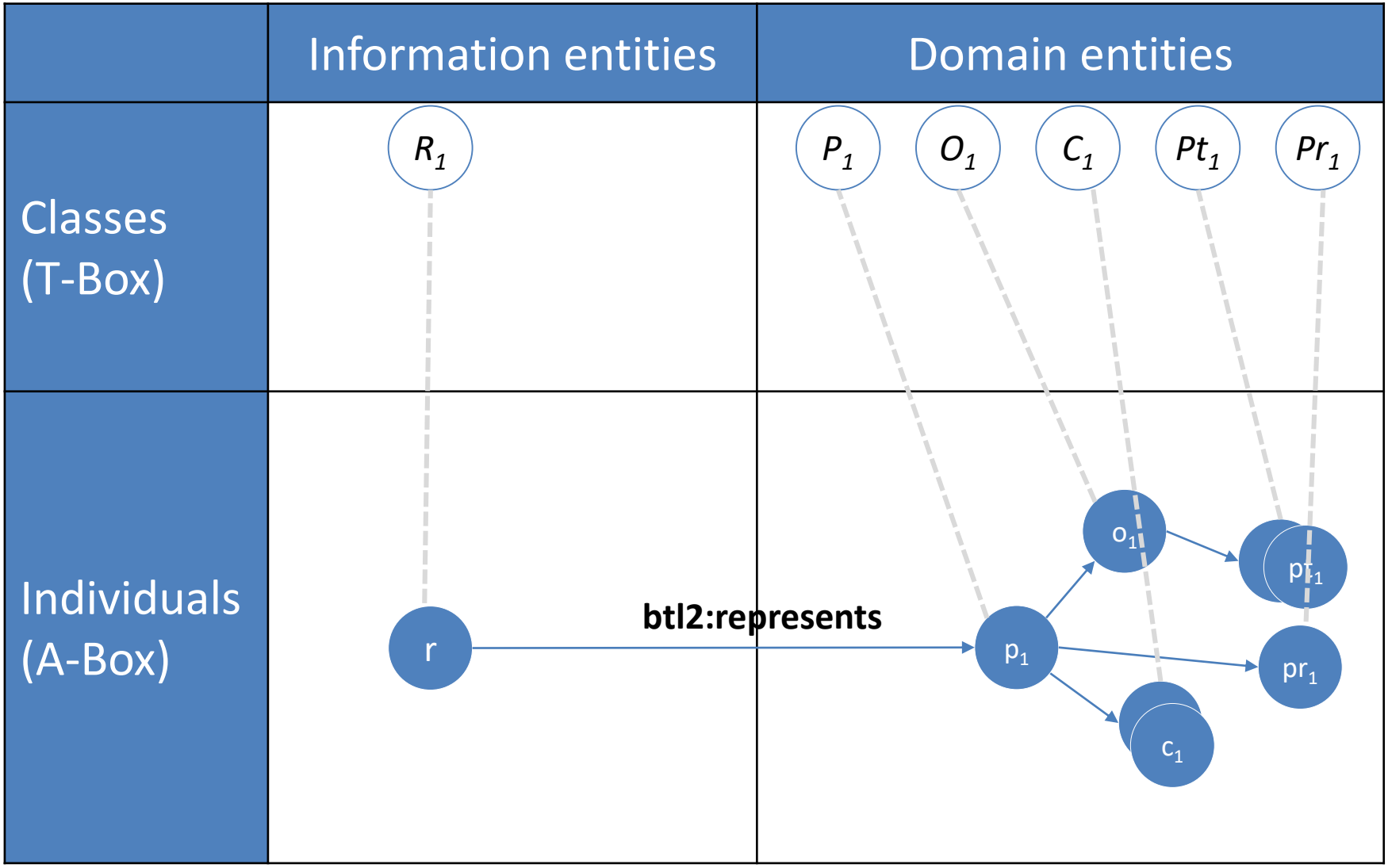
| | Information entities | Domain entities |
|---|---|---|
| Classes (T-Box) | • Bio-DB<br>• Database record<br>• Data item | • *Homo sapiens*<br>• *Megaloblastic anemia*<br>• *Cobalamin binding*<br>• *Methionin synthase*<br>• *(…)* |
| Individuals (A-Box) | • Uniprot<br>• Ensembl<br>• Database record about Methionin Synthase in Homo Sapiens<br>• Data item, such as "cobalamin binding" in this record | • John Doe, of which tissue is stored in a biobank and analysed in a lab<br>• John's megaloblatic anemia<br>• A cobalamin binding process observed in the lab within a tissue sample from John<br>• a dysfunctional Methionin synthase protein molecule in John's tissue |

# Denotation

| | Information entities | Domain entities |
|---|---|---|
| Classes (T-Box) | • Bio-DB<br>• Database record<br>• Data item | • *Homo sapiens*<br>• *Megaloblastic anemia*<br>• *Cobalamin binding*<br>• *Methionin synthase*<br>• *(…)* |
| Individuals (A-Box) | • UniProt<br>• Ensembl<br>• Database record about Methionin Synthase in Homo Sapiens<br>• Data item, such as "cobalamin binding" in | • John Doe, of which tissue is stored in a biobank and analysed in a lab<br>• John's megaloblatic anemia<br>• A cobalamin binding process observed in the lab within a tissue sample from John<br>• a dysfunctional Methionin synthase protein molecule in John's tissue |



# Case 1: database entry represents individuals

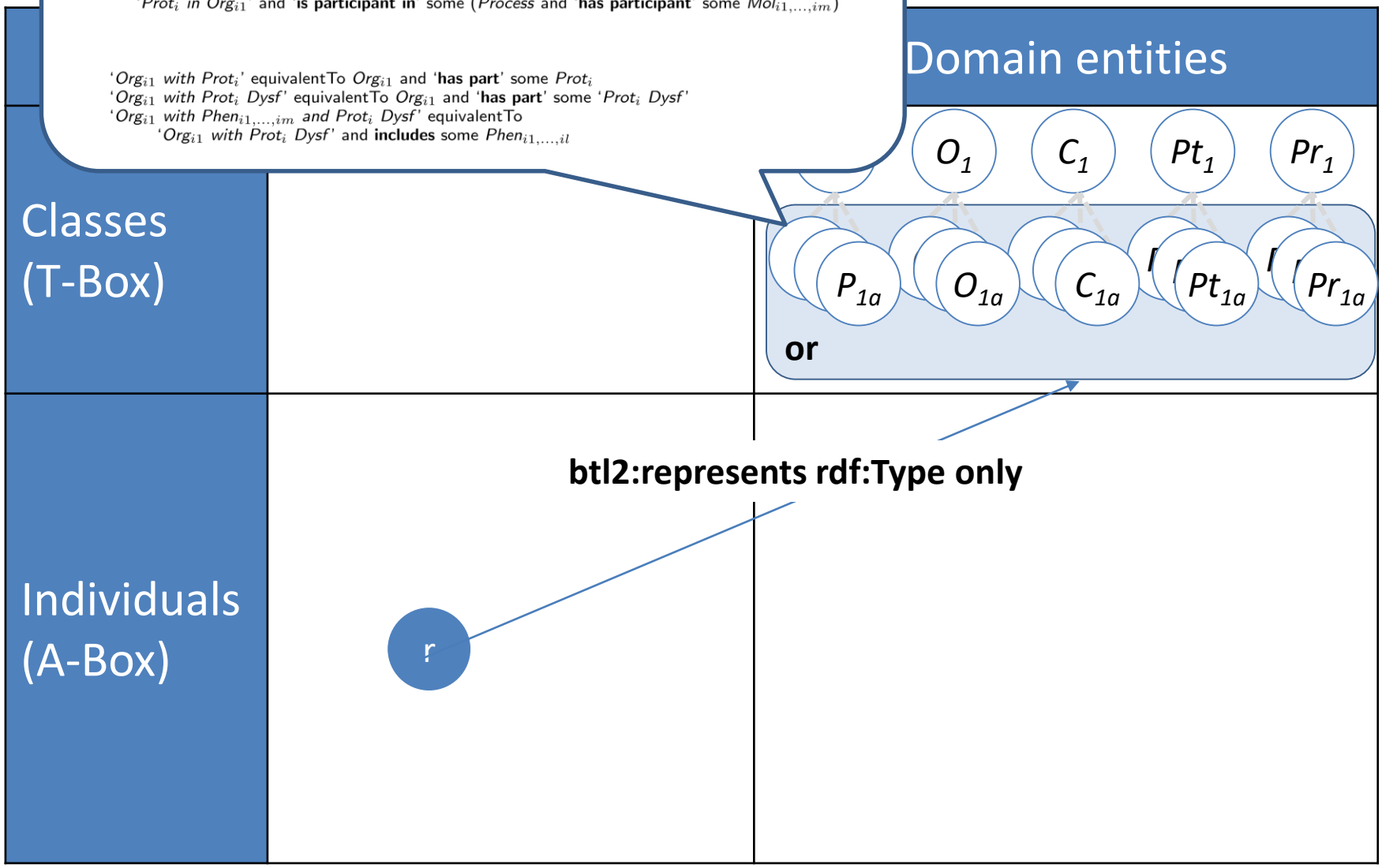


# resents classes

## Multiple defined subclasses

'$Prot_i$ *Dysf in* $Org_{i1}$ *with Phen* $_{i1,\ldots,il}$' equivalentTo
'$Prot_i$ *Dysf in* $Org_{i1}$ and '**is part of**' some ($Org_{i1}$ and **includes** some $Phen_{i1,\ldots,im}$)
'$Prot_i$ *in* $Org_{i1}$ *in BProc* $_{i1,\ldots,il}$' equivalentTo '$Prot_i$ *in* $Org_{i1}$' and '**is participant in**' some $BProc_{i1,\ldots,im}$
'$Prot_i$ *in* $Org_{i1}$ *in CComp* $_{i1,\ldots,il}$' equivalentTo '$Prot_i$ *in* $Org_{i1}$ and '**is included in**' some $CComp_{i1,\ldots,im}$
'$Prot_i$ *in* $Org_{i1}$ *with Mol* $_{i1,\ldots,il}$' equivalentTo
'$Prot_i$ *in* $Org_{i1}$' and '**is participant in**' some (*Process* and '**has participant**' some $Mol_{i1,\ldots,im}$)

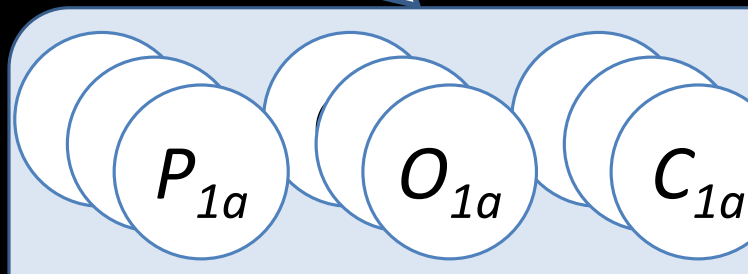
'$Org_{i1}$ *with* $Prot_i$' equivalentTo $Org_{i1}$ and '**has part**' some $Prot_i$
'$Org_{i1}$ *with* $Prot_i$ *Dysf*' equivalentTo $Org_{i1}$ and '**has part**' some '$Prot_i$ *Dysf*'
'$Org_{i1}$ *with* $Phen_{i1,\ldots,im}$ *and* $Prot_i$ *Dysf*' equivalentTo
'$Org_{i1}$ *with* $Prot_i$ *Dysf*' and **includes** some $Phen_{i1,\ldots,il}$



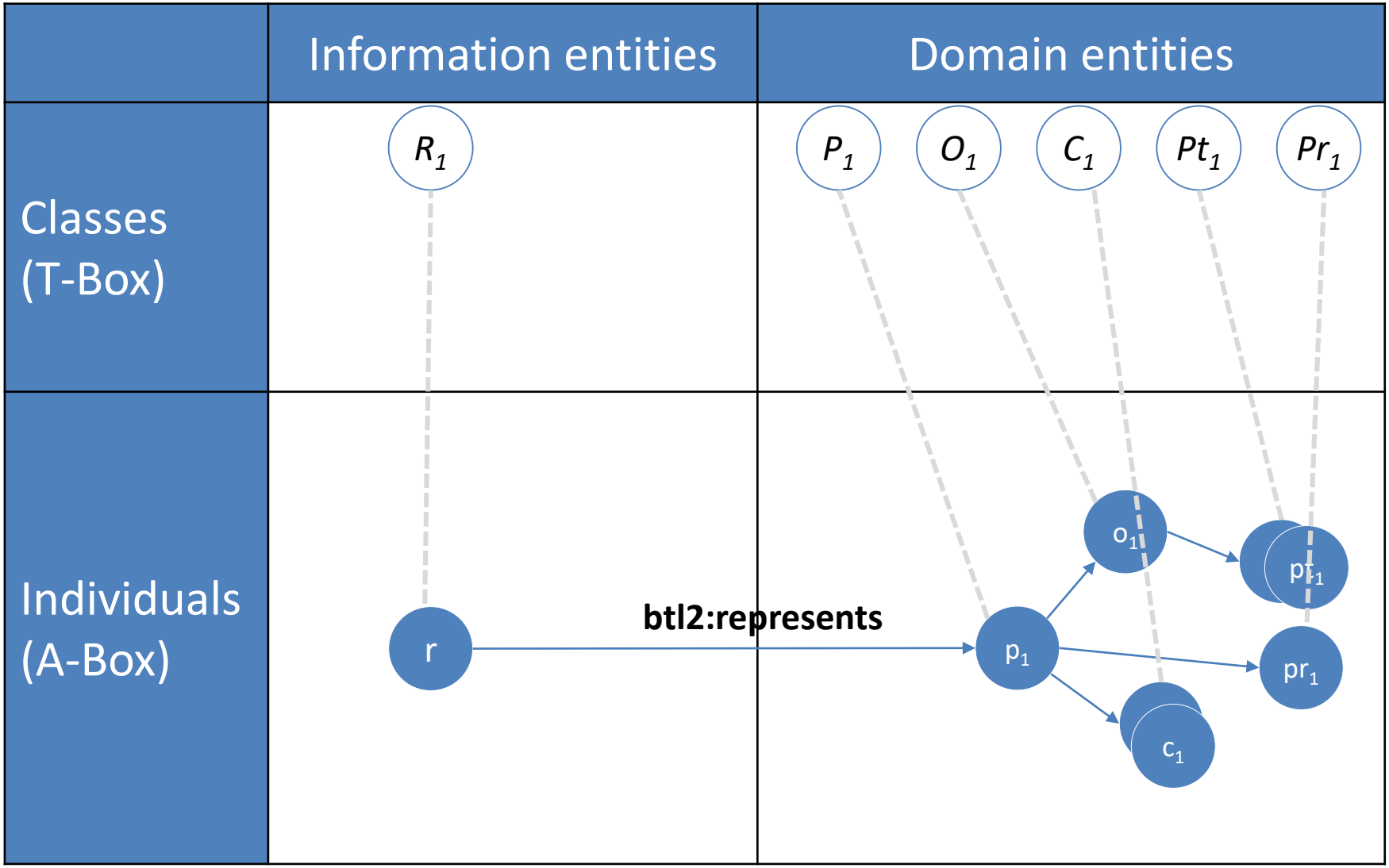
# Multiple defined subclasses

'*Prot*$_i$ *Dysf in Org*$_{i1}$ *with Phen* $_{i1,\ldots,il}$' equivalentTo
'*Prot*$_i$ *Dysf in Org*$_{i1}$ and '**is part of**' some (*Org*$_{i1}$ and **includes** some *Phen*$_{i1,\ldots,im}$)
'*Prot*$_i$ *in Org*$_{i1}$ *in BProc* $_{i1,\ldots,il}$' equivalentTo '*Prot*$_i$ *in Org*$_{i1}$' and '**is participant in**' some *BProc*$_{i1,\ldots,im}$
'*Prot*$_i$ *in Org*$_{i1}$ *in CComp* $_{i1,\ldots,il}$' equivalentTo '*Prot*$_i$ *in Org*$_{i1}$ and '**is included in**' some *CComp*$_{i1,\ldots,im}$
'*Prot*$_i$ *in Org*$_{i1}$ *with Mol* $_{i1,\ldots,il}$' equivalentTo
'*Prot*$_i$ *in Org*$_{i1}$' and '**is participant in**' some (*Process* and '**has participant**' some *Mol*$_{i1,\ldots,im}$)

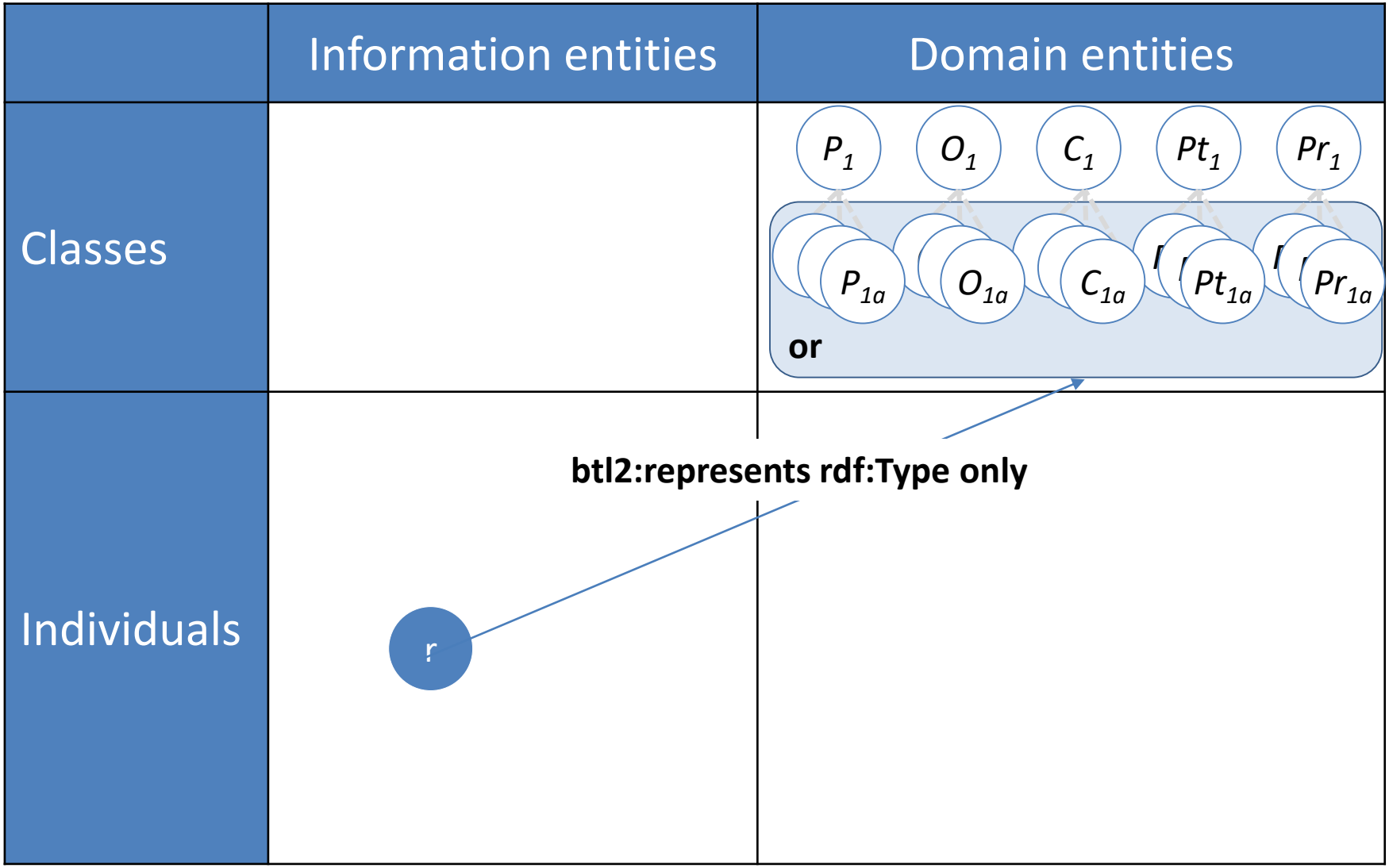
'*Org*$_{i1}$ *with Prot*$_i$' equivalentTo *Org*$_{i1}$ and '**has part**' some *Prot*$_i$
'*Org*$_{i1}$ *with Prot*$_i$ *Dysf*' equivalentTo *Org*$_{i1}$ and '**has part**' some '*Prot*$_i$ *Dysf*'
'*Org*$_{i1}$ *with Phen*$_{i1,\ldots,im}$ *and Prot*$_i$ *Dysf*' equivalentTo
'*Org*$_{i1}$ *with Prot*$_i$ *Dysf*' and **includes** some *Phen*$_{i1,\ldots,il}$



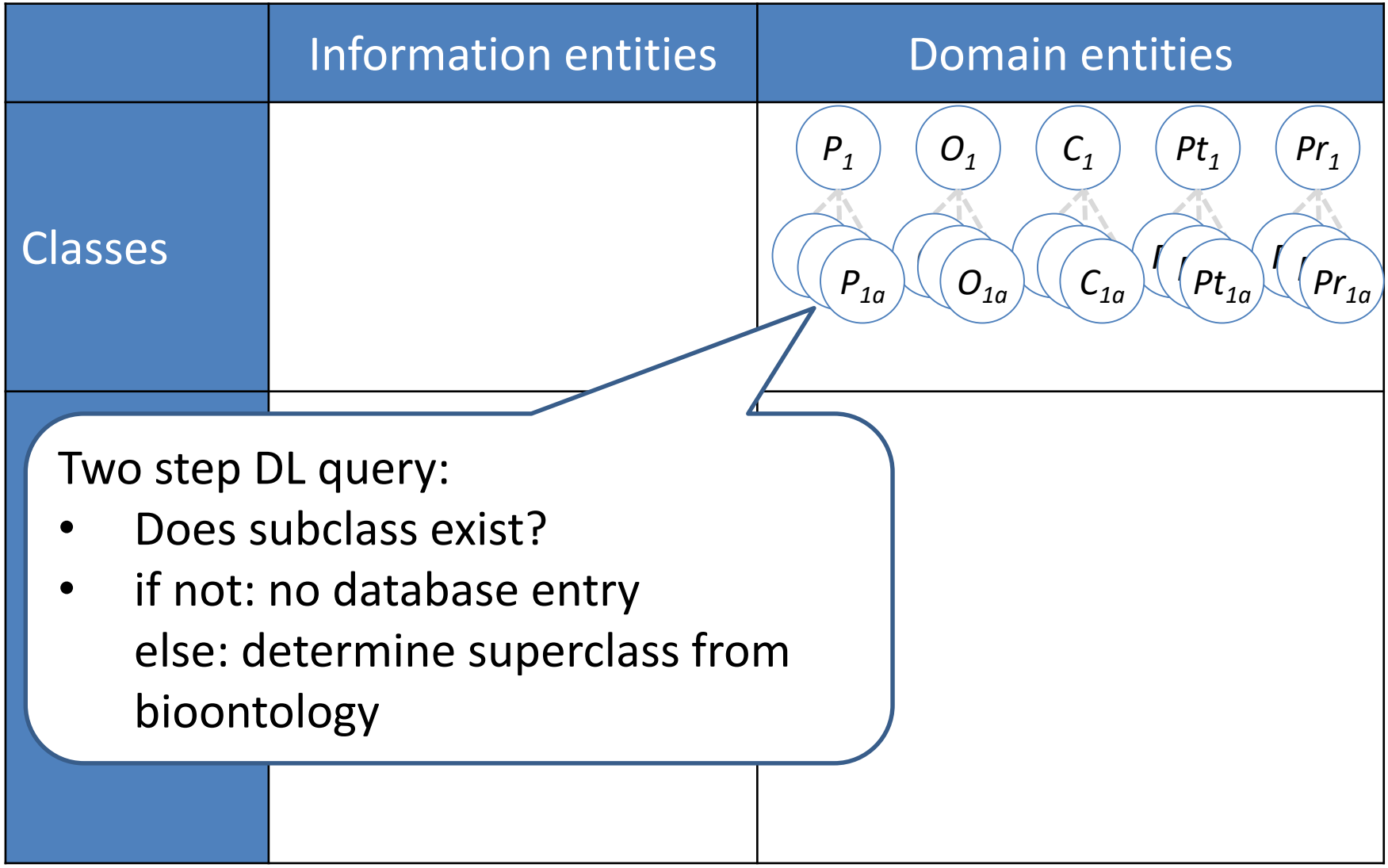
# Querying: A-box query for individuals



# Querying: T-box query for subclasses



# Querying: T-box query for subclasses



# Competency questions

(Q1) Which biological processes have proteins of the kind $Prot_i$ as participants?

*BProc*$_1$ and (‘**has participant**’ some *Prot*$_i$)

(Q2) In which cellular locations is $Prot_i$ active in organisms of the type $Org_1$?

‘*Cellular component*’ and (‘**is included in**’ some *Org*$_1$)
and (**includes** some *Prot*$_i$)

(Q3) Which proteins are involved in processes of the type BProc in organisms of the type $Org_1$?

*Protein* and (‘ **is participant in**’ some *BProc*$_1$ )
and (‘**is included in**’ some *Org*$_1$)

# Evaluation (one database record)

| Model | Q1 | Q2 | Q3 | Classes | Individuals | Axioms/ Assertions |
|---|---|---|---|---|---|---|
| A-Box | bp1001, bp2001, bp3001 | cc1001, cc2001, cc3001 | p1004 | 24 | 51 | 207 |
| T-Box | $BProc_1$ | $CComp_1$ | $Prot_i$ | 68 | 0 | 149 |

# Discussion (I)

- Both modelling solutions:
  - highly productive
  - scaling problems to be expected
- A-Box solution (prototypical individuals):
  - A-box reasoning more costly
  - Makes existential assumptions
- T-Box solution (multiple subclasses)
  - Theoretically allows non-referential entries
  - Simplified model: EL++

# Discussion (II)

- Do biological database refer to ontological content?
  - No "real" universal statements on biological entities
  - Even no existential assumption
  - Dispositional statements to be discussed (see paper)
- Exercise best described as ontological representation of referring individuals
- Possible use case: non-disruptive querying of Bio-DBs where axioms of the annotation ontologies need to be explored

# Conclusion

- Four ontological approaches - IND, SUBC, DISP and HYB
  - Structure and content of BIO-DBs
- Solution:
  - Expressiveness, DB retrieval and retrieval based on DL queries
  - Interpretation:
    - Denoted entities as prototypical individuals
    - Creation of defined subclasses
    - Database content as reporting dispositions

Funding: This work was funded by *Conselho Nacional de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) 3914/2014-03 and *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq) 140698/2012-4.