Knowledge manipulation using OWL and reasoners for drug-discovery

ICBO 2013 Tutorial

July 7th 2013

Samuel Croset

EMBL-EBI ChEMBL UNIVERSITY OF CAMBRIDGE

Errata – July, 8th 2013

Here is a list of points raised during the tutorial and based on feedback from the audience. I will try to address them for a next release of the talk. Send me an email if you need clarification or have more comments croset@ebi.ac.uk - Samuel

Things that can be improved (list not comprehensive):

- Direct semantics versus OWL based semantics → *Could be removed from the talk. The reader can skip that.*
- *is_a relationship as defined by GO corresponds to a rdfs:subClassOf axiom in OWL.*
- *In OWL, is_a is not an object property, it's a built-in primitive construct from the language defining the relashionship between sets of things. Other properties (part-of, regulates, etc…) are defined by OWL object properties.*

Material

- **Files**: http://bit.ly/12flbf8
- **Protégé 4.3**: http://stanford.io/102ZBJO
- **Brain**: http://bit.ly/TYGj4O

Tutorial

- Ask questions!

- What is OWL?
- Why is it particularly interesting for life sciences?
- How to use OWL?
- What is OWL 2EL?
- How to integrate and query biomedical knowledge?

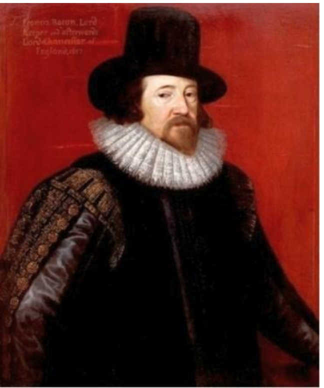
Why learning OWL?

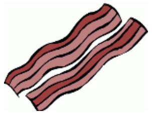
“The scientist is not a person who gives the right answers, he's one who asks the right questions”

— *Claude Lévi-Strauss*

“Half of science is putting forth the right questions”

— *Sir Francis Bacon*

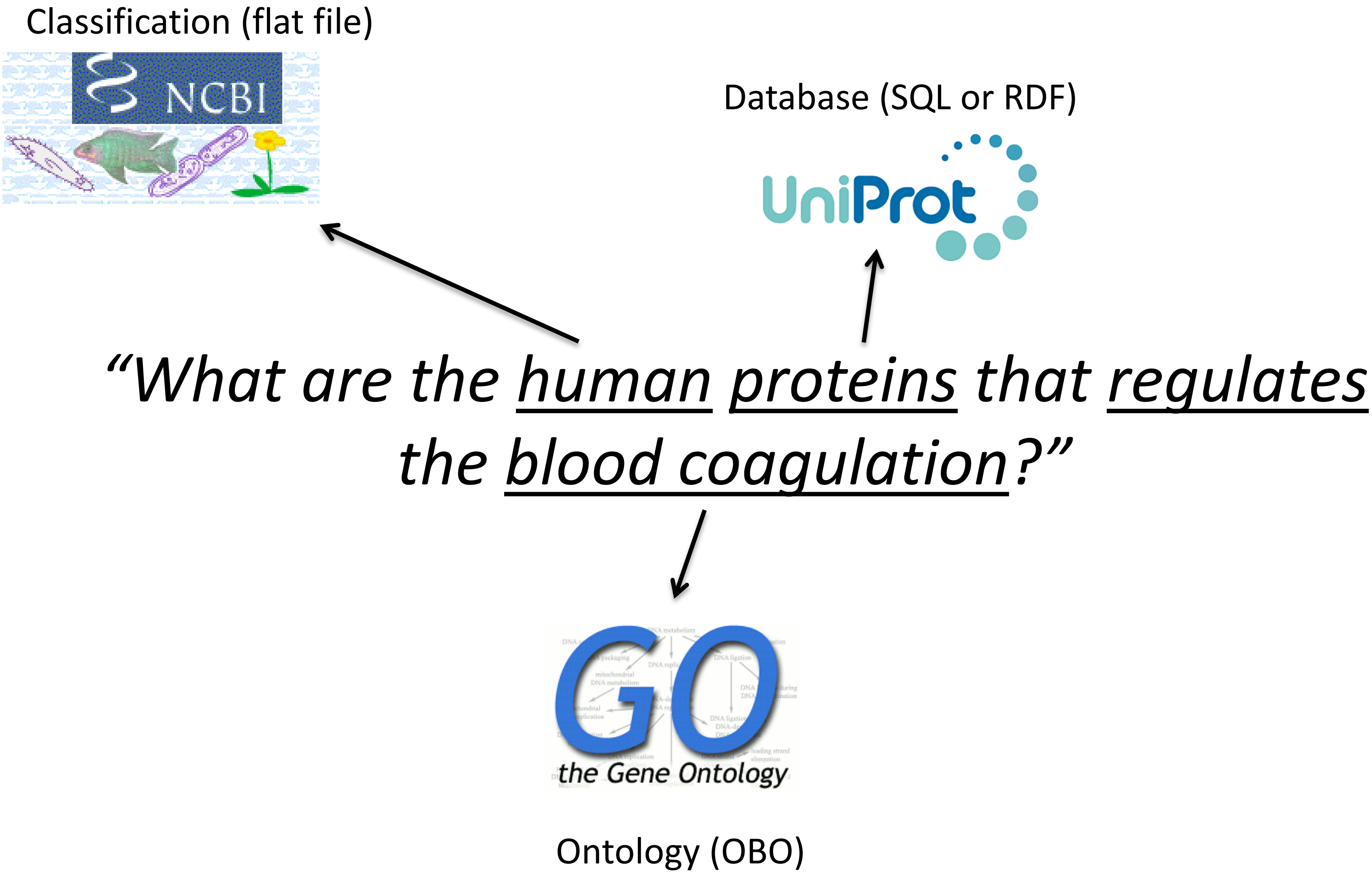




Why learning OWL?

“What are the human proteins that regulates the blood coagulation?”

Why learning OWL?



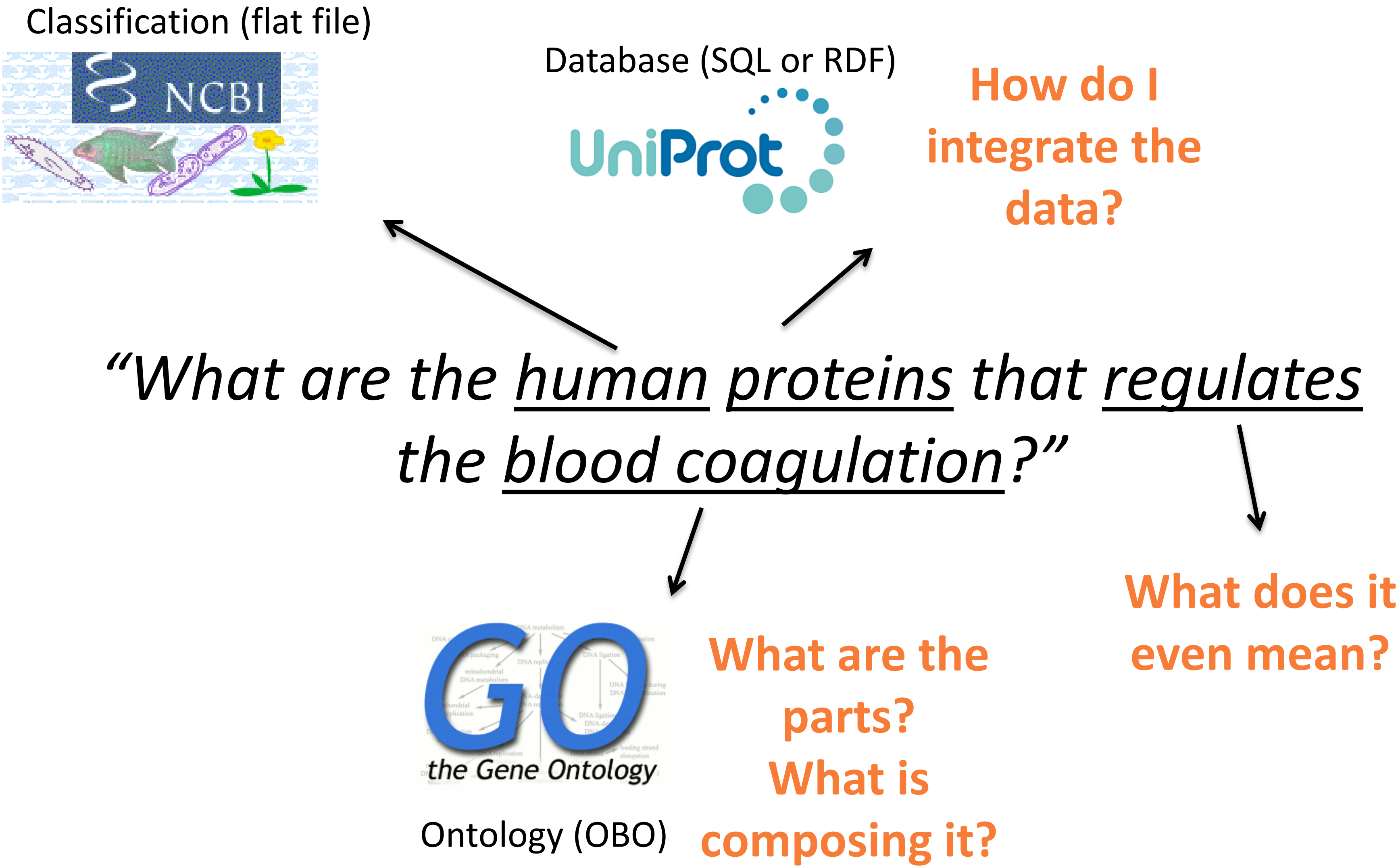
Classification (flat file)

Database (SQL or RDF)

“What are the human proteins that regulates the blood coagulation?”

Ontology (OBO)

Why learning OWL?



Classification (flat file)

Database (SQL or RDF)

How do I integrate the data?

"What are the human proteins that regulates the blood coagulation?"

Ontology (OBO)

What are the parts? What is composing it?

What does it even mean?

Why learning OWL?

- Existing resources can already answer the question → But they need to **interact**
- Ontologies are not only labels or annotations for biological concept (“blood coagulation”) → They help to **formalize** problem
- We want to mix traditional ontologies with other **large-scale data**
- We want an **intuitive way** to formulate the query, hiding the implementation

What is OWL?

- The Semantic Web: RDF → URI and triples → Should improve interoperability over the Web
- Need for shared schemas → ontologies
- OWL → **Description logics** and knowledge representation, decidable, attractive and well-understood computational properties.
- (OWL → Direct Semantics or RDF-based semantics)

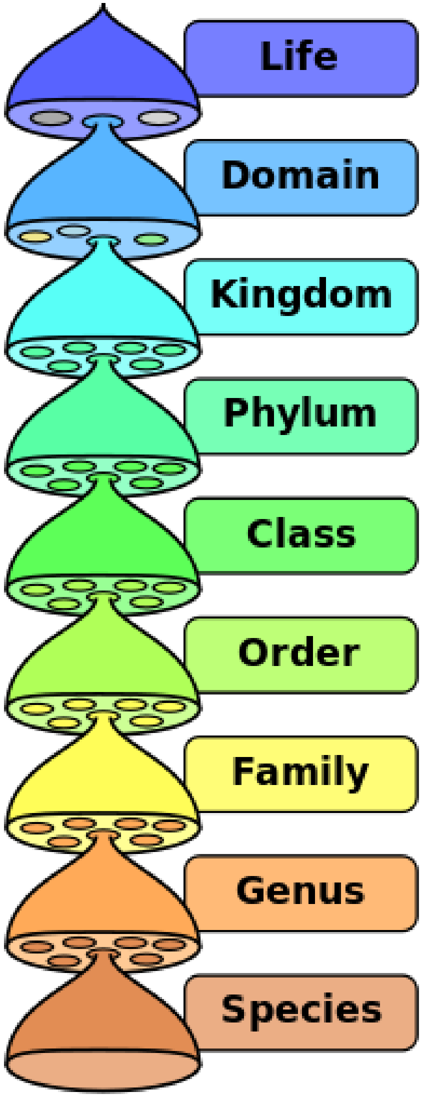
What is OWL?

- **Confusing** relations between OWL, RDF, SPARQL, reasoning, etc…
- Here we deal with the **Direct Semantics** of OWL (no RDF) → It’s easier!
- You get to use the reasoner a lot!
- In OWL you build **knowledge-bases** or **ontologies** (here these terms are synonyms – in the wild people use the two).

OWL and Life Sciences

Advantages versus RDF, SQL and flat files?

- Formal language to represent hierarchical data
- Machine reasoning
- Large-scale (OWL 2EL)
- Knowledge integration
- Composition
- Powerful query mechanism



OWL 2 Terminology

- **It’s all about definitions!**
- **Defining things based on the relations they have**

- **Entities:** elements used to refer to real-world objects
- **Expressions:** combinations of entities to form complex descriptions from basic ones
- **Axioms:** the basic statements that an OWL ontology expresses → Pieces of knowledge

http://www.w3.org/TR/owl2-primer/#Modeling_Knowledge:_Basic_Notions

Entities

- **Classes**: Categories and Terminology
  - *Protein, Human, Drug, Chemical, P53, Binding site, etc…* → **Pretty much everything in life science.**

- **Individuals (objects)**: Instances
  - *Rex the dog, this mouse on the bench, you, etc…*

- **Properties**: Relations between individuals
  - *Part of, regulates, perturbs, etc…*

Axioms

- Statements, pieces of knowledge → express **the truth.**
- How classes and properties relate to each other:
  - All Humans are Mammals → Human is a subclass of Mammal
- **You should always think in terms of individuals.** In biology we don’t really deal much with real individuals, yet classes/properties and axioms are built from relationships between anonymous individuals.
- Our first OWL axiom: **SubClassOf**

Ontology/Knowledge-base

- Set of axioms
- Serialized as “.owl” file – Here using the Manchester syntax (Description logics semantics)
- Example of output (look at the format, don’t try to understand the logic now):

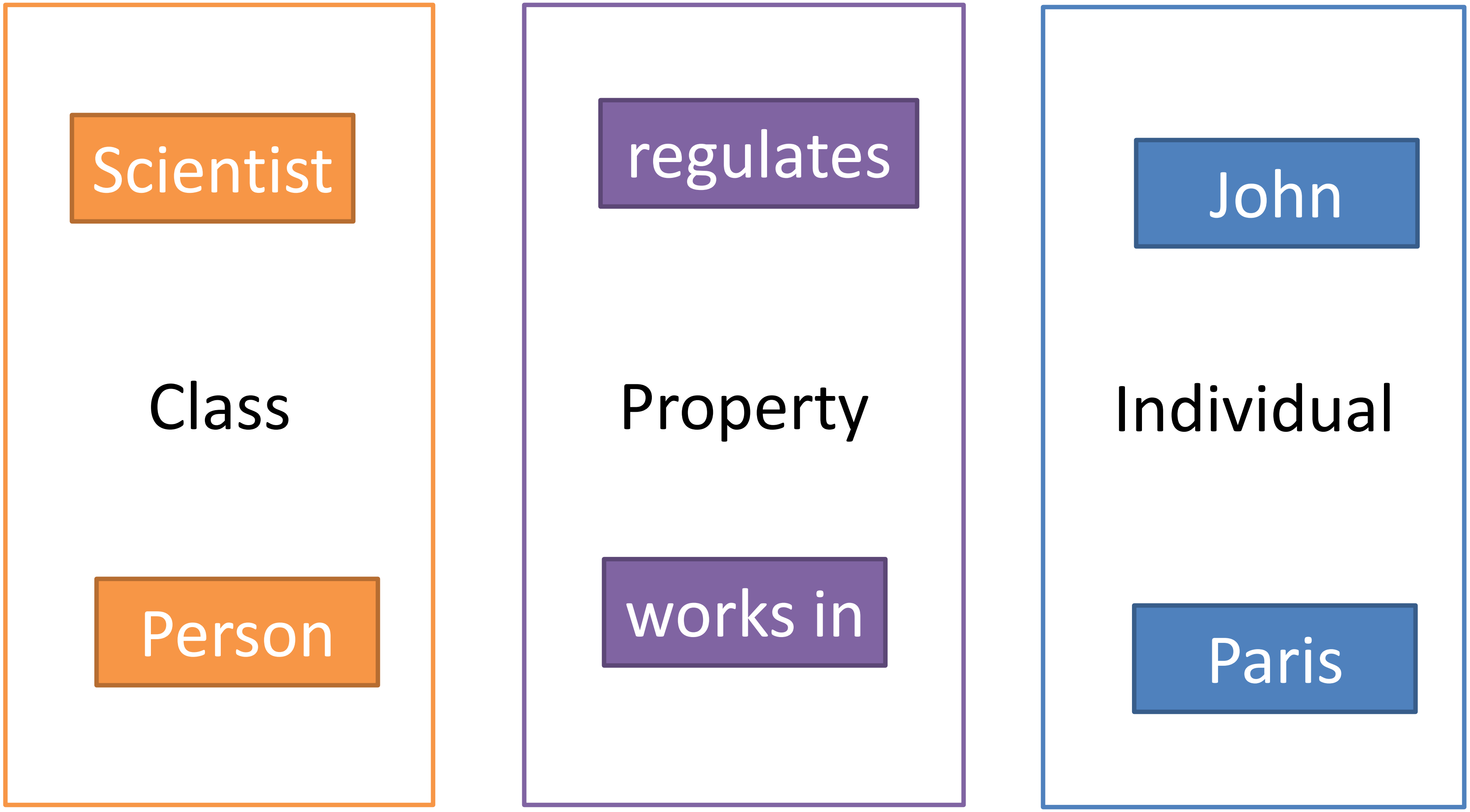
```
ObjectProperty: part-of

Class: owl:Thing

Class: Cell

Class: Nucleus

    SubClassOf:
        part-of some Cell
```

Terminology Summary

Output in RDF (turtle – RDF-based semantics):

```
<demo.owl> rdf:type owl:Ontology .

:part-of rdf:type owl:ObjectProperty .

:Cell rdf:type owl:Class .

:Nucleus rdf:type owl:Class ;

  rdfs:subClassOf [ rdf:type owl:Restriction ;
                    owl:onProperty :part-of ;
                    owl:someValuesFrom :Cell
                  ] .

owl:Thing rdf:type owl:Class .
```

Terminology Summary



Scientist

Class

Person

regulates

Property

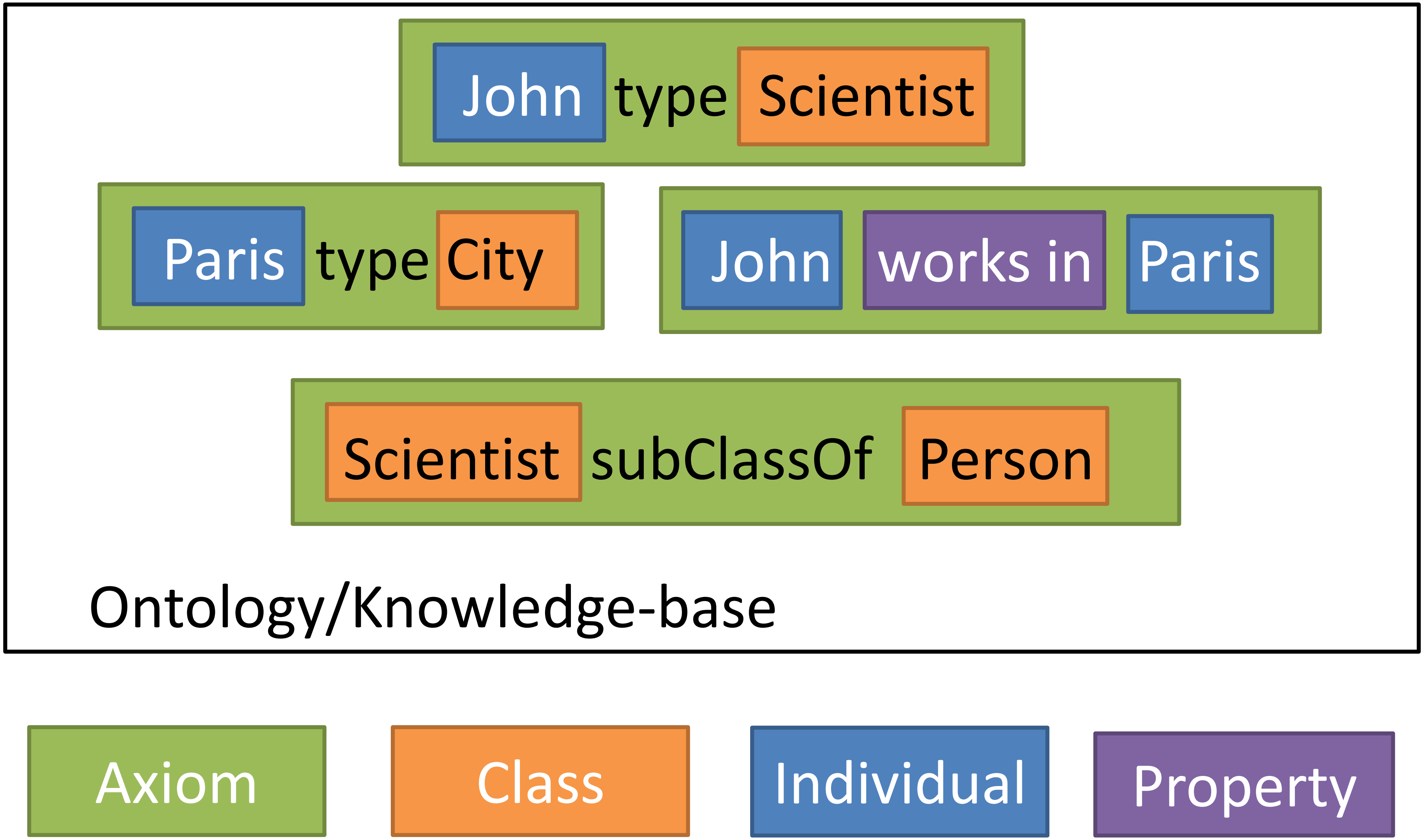
works in

John

Individual

Paris

Terminology Summary



John type Scientist

Paris type City

John works in Paris

Scientist subClassOf Person

Ontology/Knowledge-base

Axiom

Class

Individual

Property

Exercise 1 – Classes and axioms

- Open the file “NCBI-taxonomy-mammals.owl” with a text editor. Can you understand what’s inside?
- Now open the file with Protégé and go under the tab “classes”. You can use the option “render by label” in the “View” menu.
- Can you recognize the classes? What do they describe?
- Can you spot the axioms? What do they capture?

Reasoner

- A program that understand the axioms and can deduce things from it.
- Used to **classify** the ontology.
- **Query engine** for knowledge-bases.
- More or less fast depending on the number and type of axioms.

Exercise 2 - Reasoning

- In Protégé, go under the “DL query” tab and retrieve all descendant classes of the class *Abrothrix* (or *NCBI_156196*).
- What does this query means? What about the results?

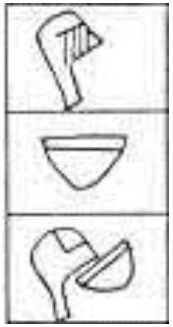
Comparison against mySQL

```
SELECT
  s.*
FROM
  species AS s,
  species AS t
WHERE
  (s.left_value BETWEEN t.left_value AND t.right_value)
AND
  t.common_name='abrothrix';
```

Constructs – Class expressions

- Combining classes and properties to define more things (class expression) → **Composition**
- Intersection: **and**
  - Mammal **and** Omnivore
- Existential Restriction: **some**
  - part-of **some** Cell

Cuneiform script (3000 BC):



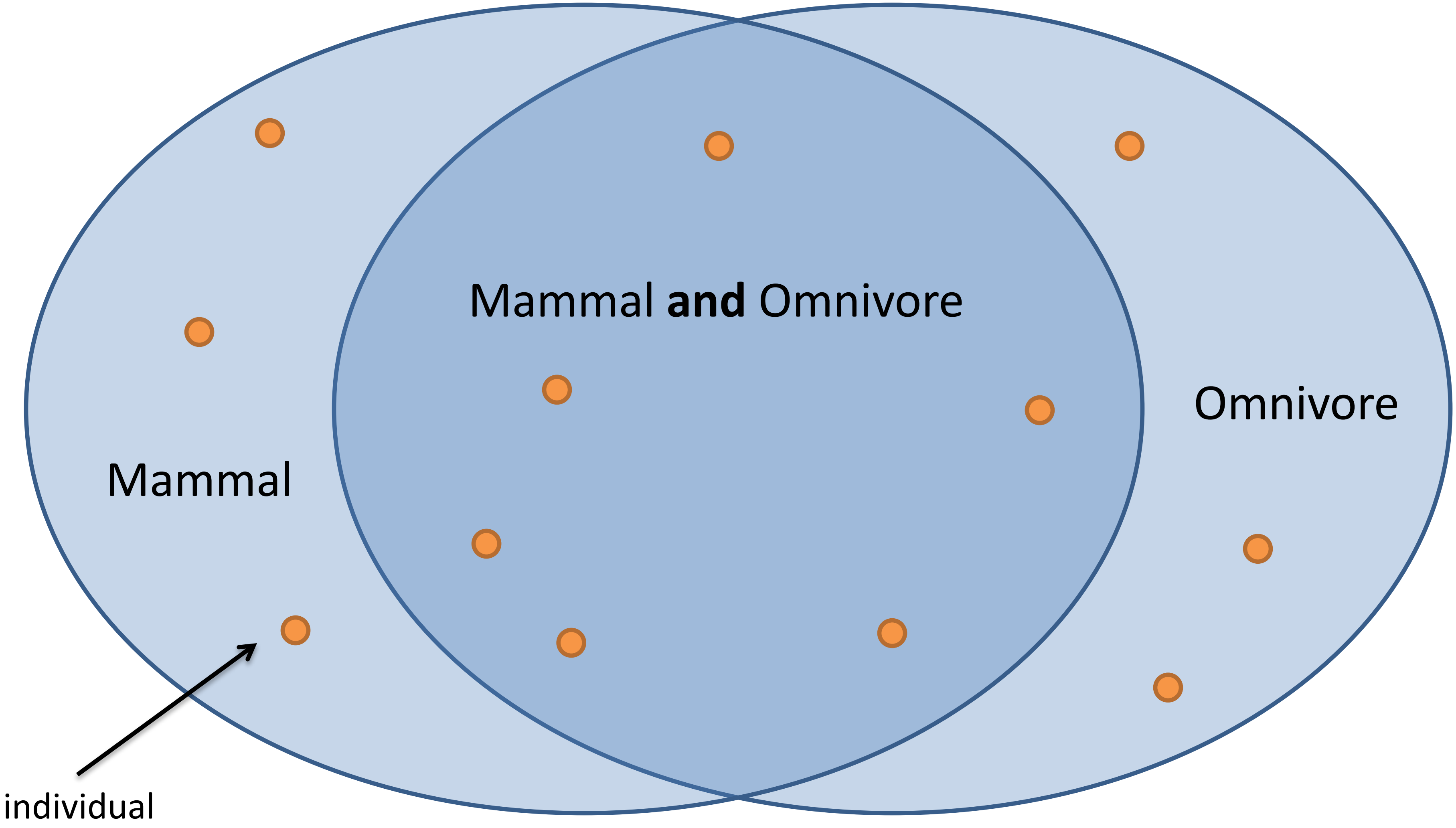
Head

Food

Eat

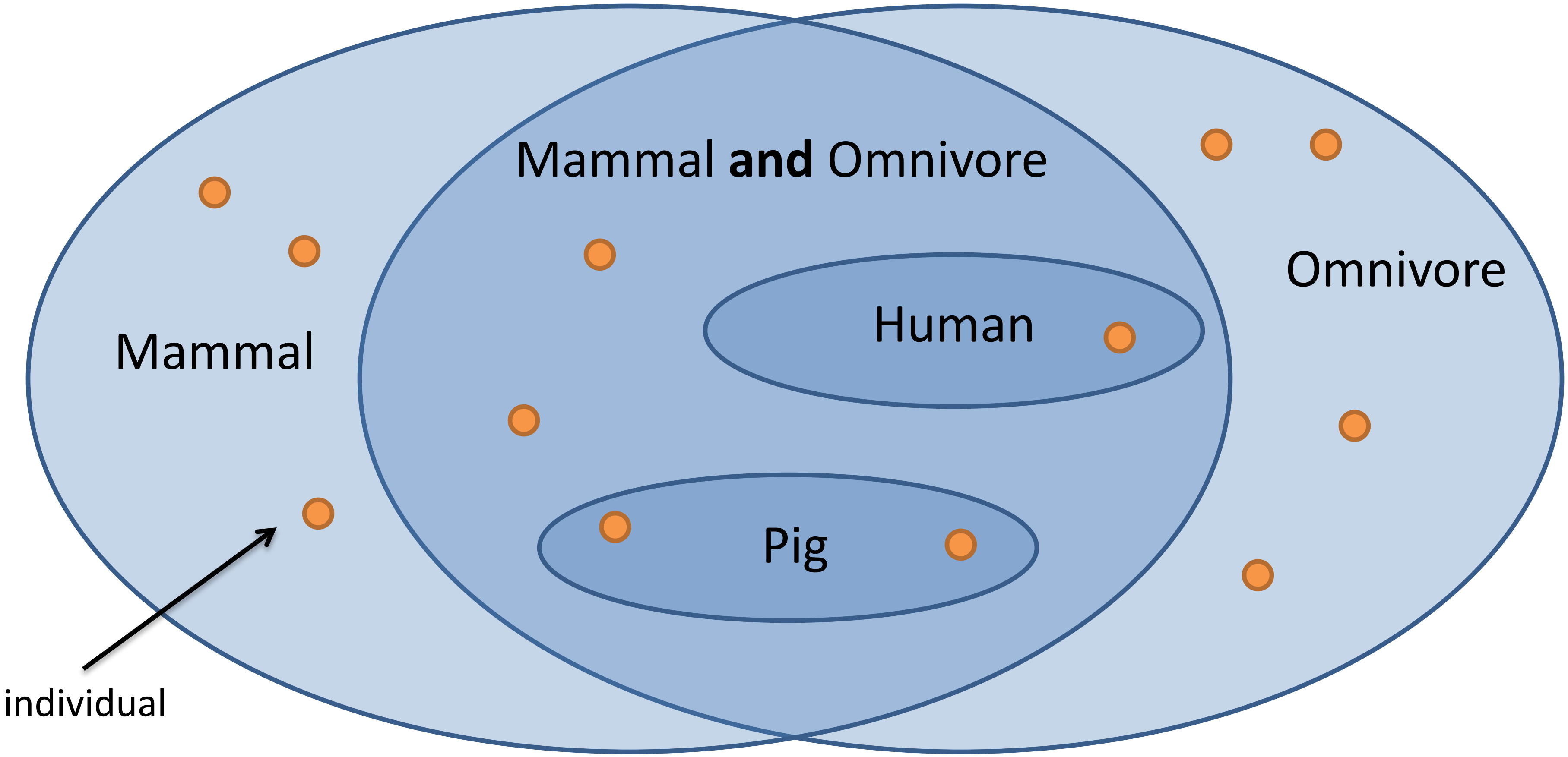
http://en.wikipedia.org/wiki/Cuneiform

Construct: **and**

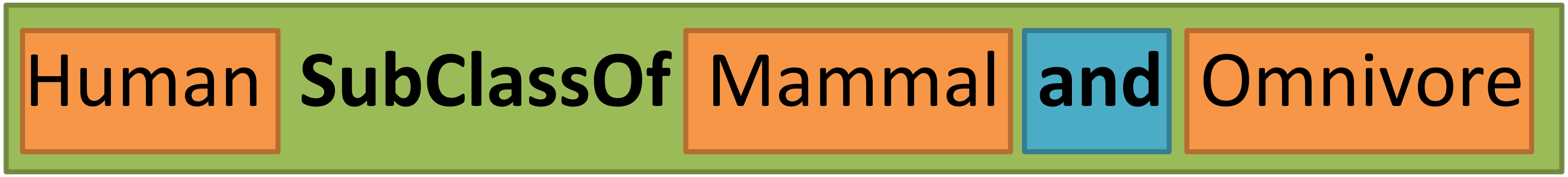


Constructs & axioms

Human **SubClassOf** Mammal **and** Omnivore



Constructs & axioms



Human **SubClassOf** Mammal **and** Omnivore

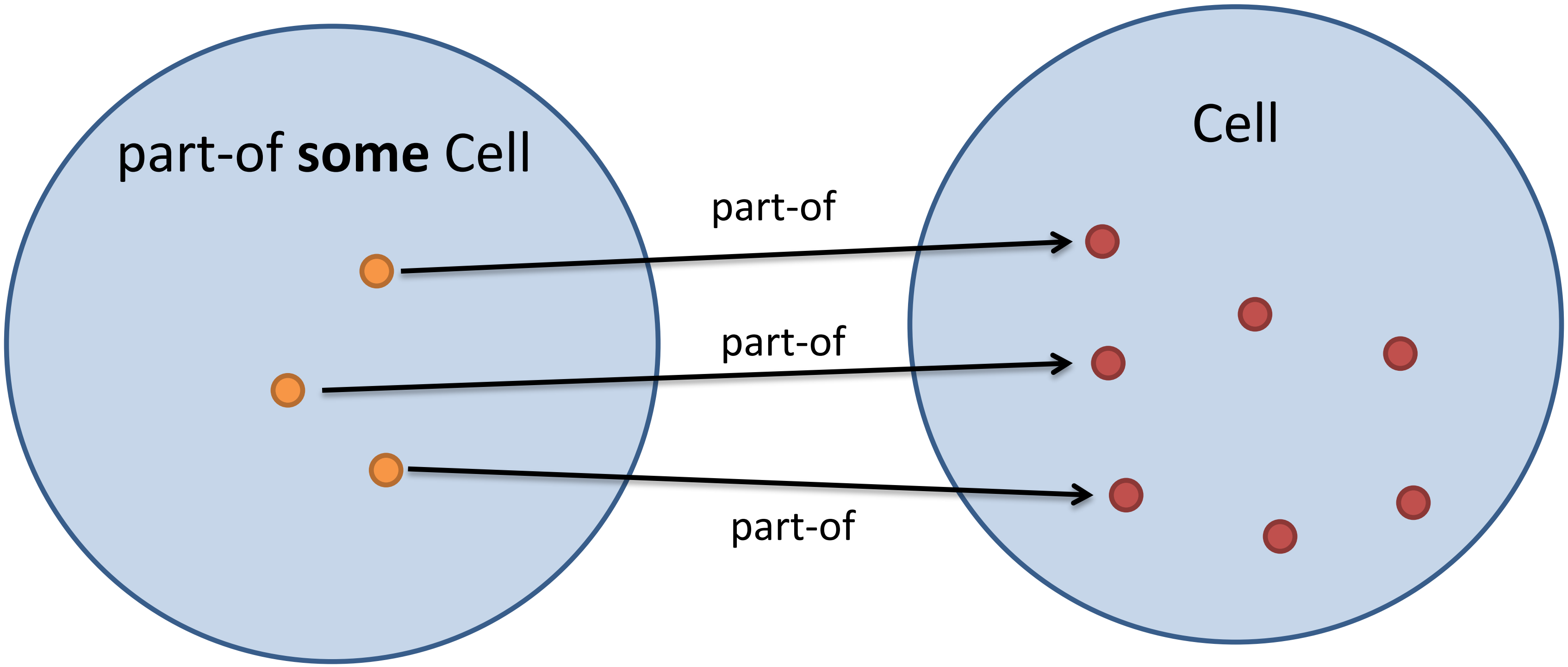
This definition (Mammal and Omnivore) of the concept “Human ” is **partial.**

- Every human must be at least a mammal and an omnivore according to our definition.
- *But it’s not because you are a mammal and an omnivore that you are necessary human!!*

Construct: **some**

Existential restriction: Weird construct at first, but useful while dealing with incomplete knowledge

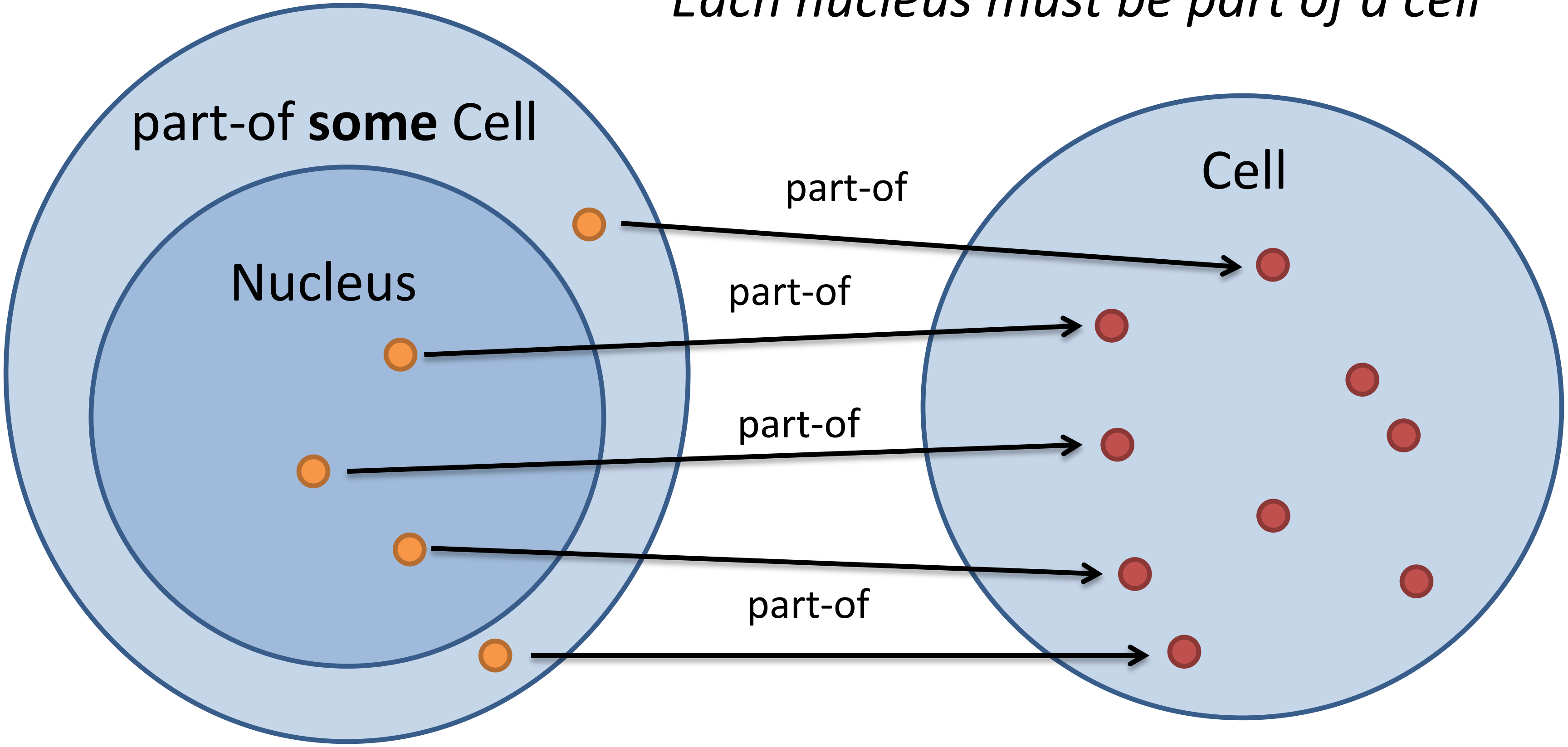
P some C: if it exists then a least one instance of C linked by P



Constructs & axioms

Nucleus **SubClassOf** part-of **some** Cell

“Each nucleus must be part of a cell”



Exercise 3 – Implementing the axiom

- Create a new project inside Protégé.
- Implement “Human SubClassOf Mammal and Omnivore”
- Run the reasoner and look at the hierarchy of classes. Does it make sense?
- That’s the main role of the reasoner → classifying things based on their definiti

- “Conceptual Lego”

OWL concepts

Class : Basic block

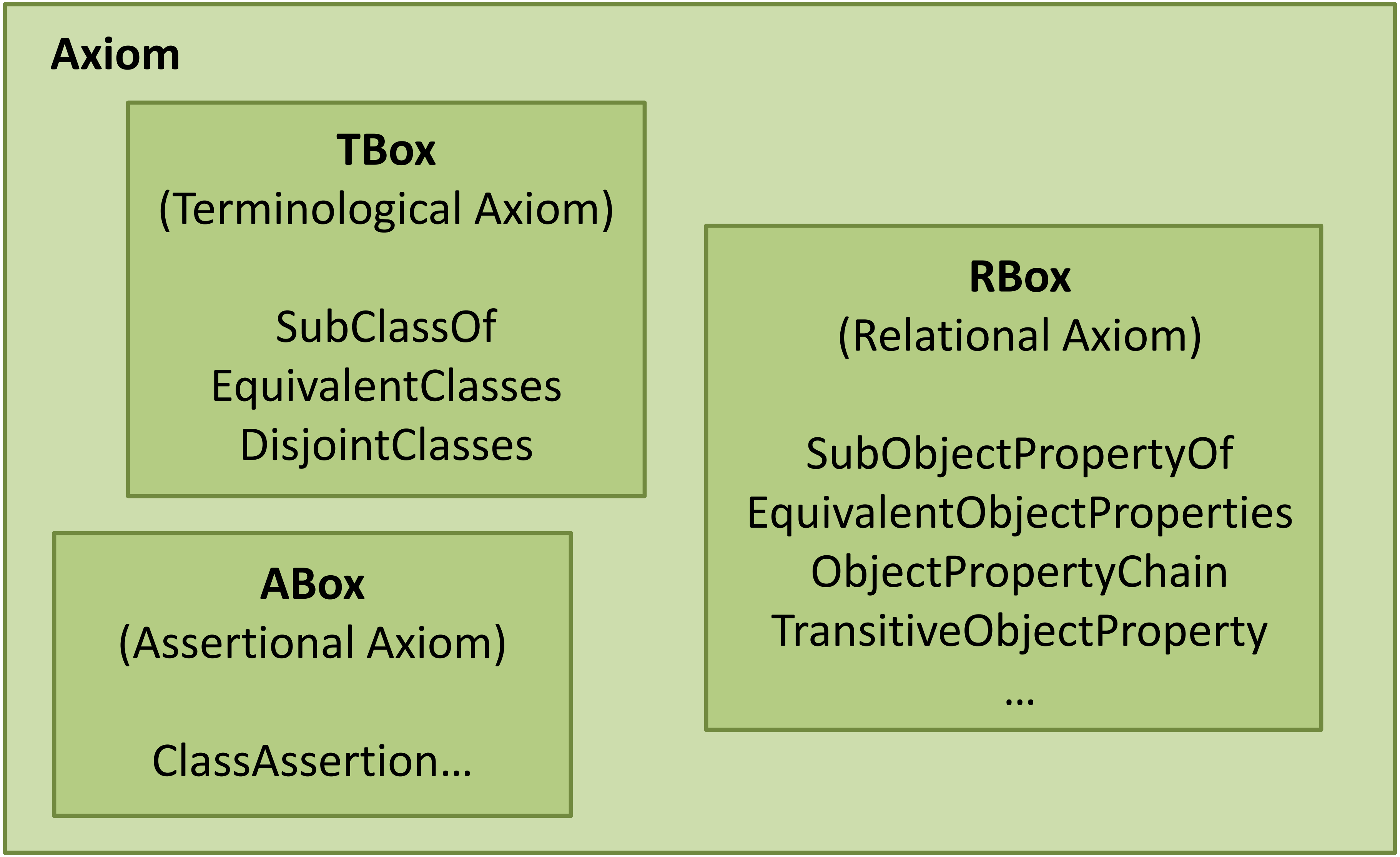
Property : Basic block

Constructor : Used in class expressions

Class Expression : Class , Property , Constructor

Axiom : Relations between these entities.

OWL Concepts



Axiom

TBox
(Terminological Axiom)

SubClassOf
EquivalentClasses
DisjointClasses

ABox
(Assertional Axiom)

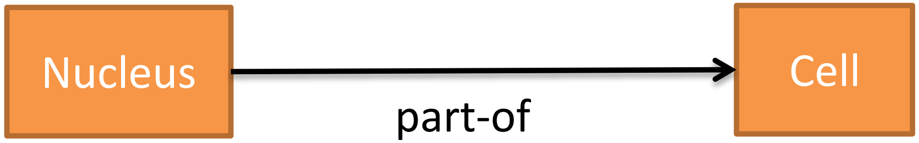
ClassAssertion…

RBox
(Relational Axiom)

SubObjectPropertyOf
EquivalentObjectProperties
ObjectPropertyChain
TransitiveObjectProperty
…

Real-life example: The Gene Ontology

- Open Biomedical Ontology (OBO) format originally.



- Moved to OWL → Stronger semantics

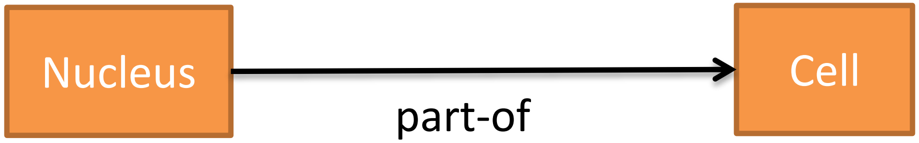
http://www.geneontology.org/GO.ontology-ext.relations.shtml

GO constructs

- Central pattern:

A **SubClassOf** P **some** B

Nucleus **SubClassOf** part-of **some** Cell

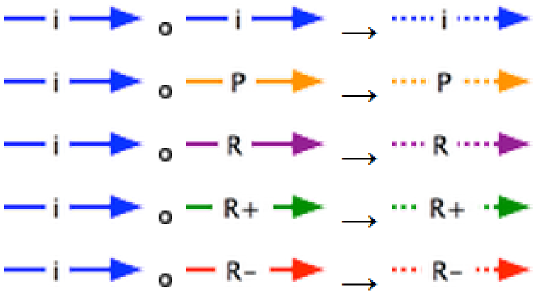
(


)

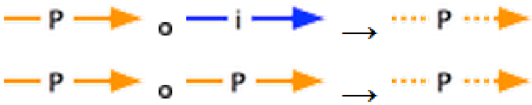
http://www.geneontology.org/GO.ontology-ext.relations.shtml

GO - RBox

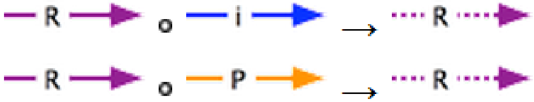
is a ∘ …



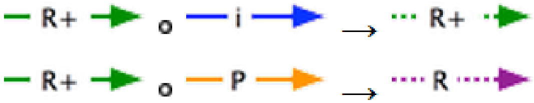
part of ∘ …



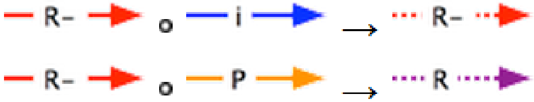
regulates ∘ …

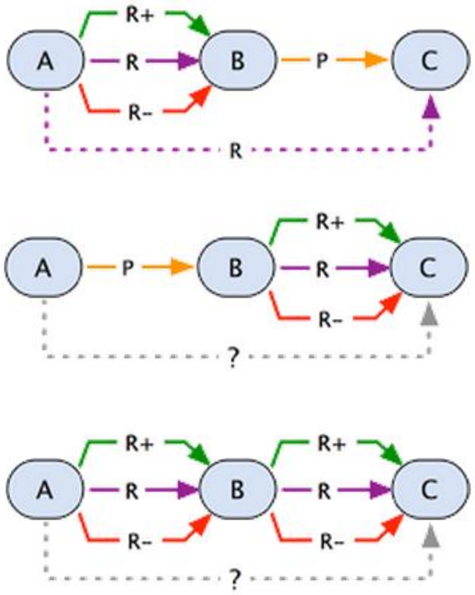


positively regulates ∘ …



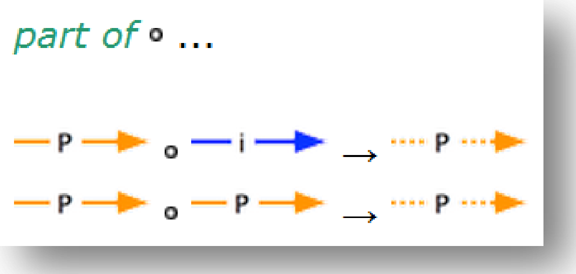
negatively regulates ∘ …



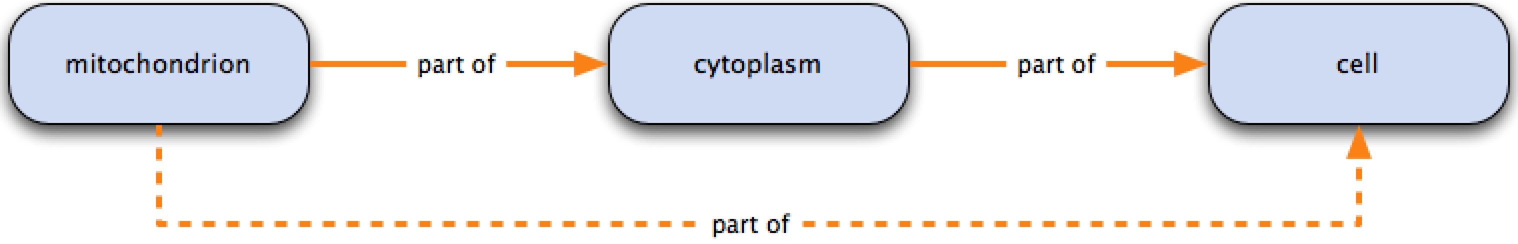


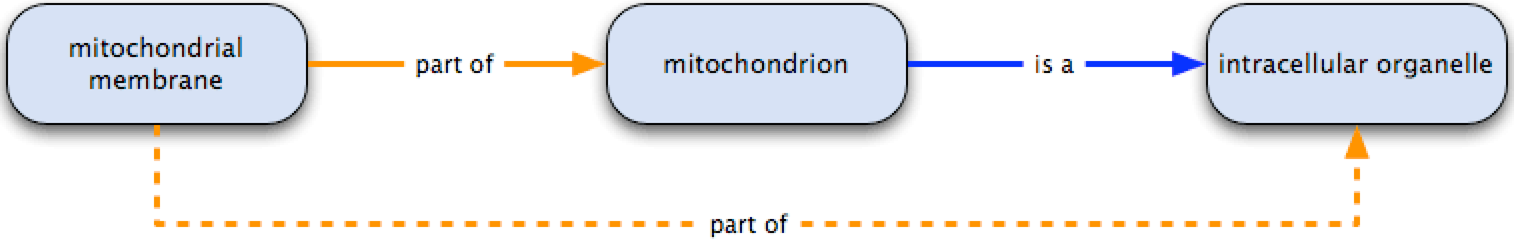
http://www.geneontology.org/GO.ontology-ext.relations.shtml

GO – Rbox: part-of



Transitivity

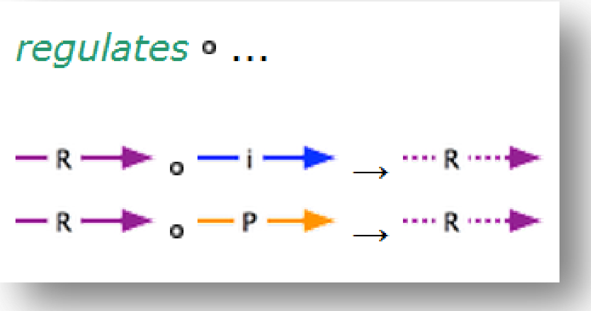




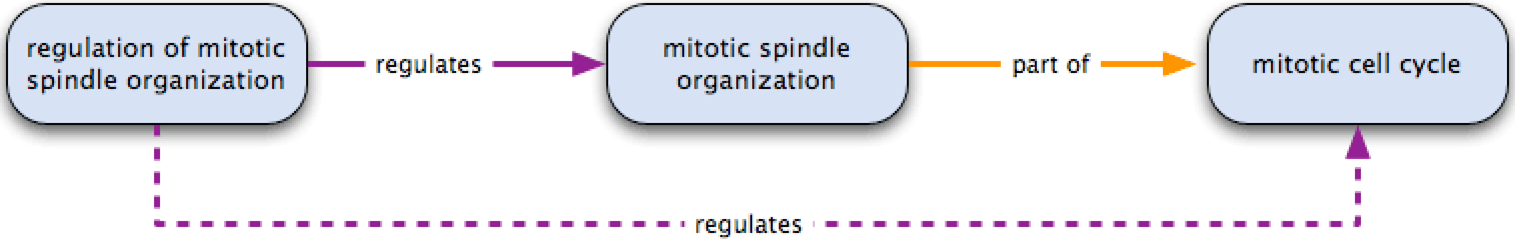
Exercise 4 – Transitive property

- Open the “gene_ontology.owl” file.
- What are the things that are a *biological_process and part_of some 'wound healing‘* ?
- Look at the class “*blood coagulation, common pathway*”. Is it obvious for this class to be in the results?

GO – Rbox: regulates



Chain


regulation of mitotic spindle organization
regulates
mitotic spindle organization
part of
mitotic cell cycle
regulates

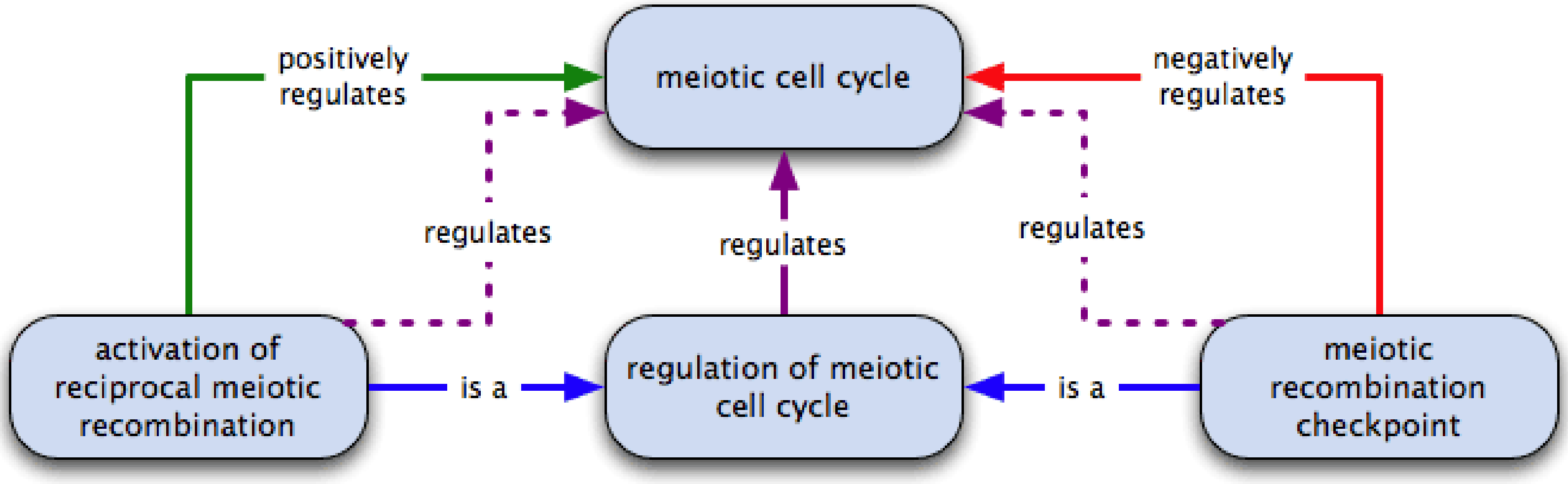
Exercise 5 – Chained properties

- Look at the “regulates” property inside Protégé.

- What are the things that are a *biological_process and regulates some 'mitotic cell cycle’* ?
- Look at the class “*positive regulation of syncytial blastoderm mitotic cell cycle*”
- Is it obvious for this class to be in the results?

GO – Rbox: positively/negatively regulates

SubProperty

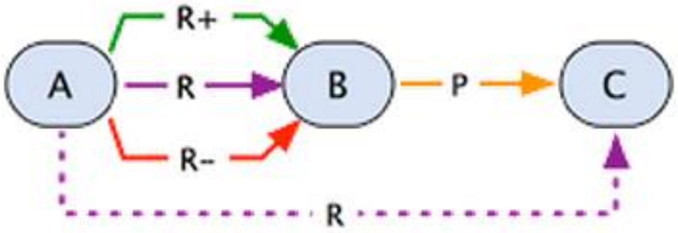

positively regulates
meiotic cell cycle
negatively regulates
regulates
regulates
regulates
activation of reciprocal meiotic recombination
is a
regulation of meiotic cell cycle
is a
meiotic recombination checkpoint

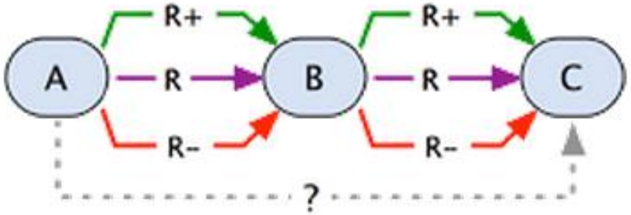
Exercise 6 – Sub Properties

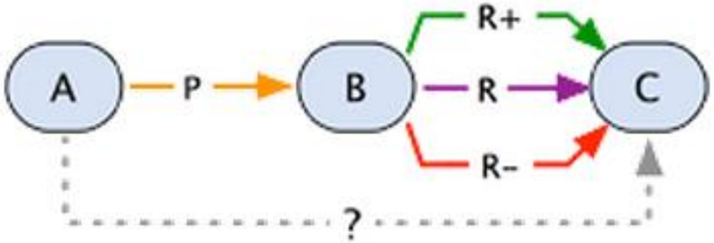
- Look at the “positively-regulates” property inside Protégé.
- What are the things that are a *biological_process and positively_regulates some 'mitotic cell cycle'* ?
- Are they different from the things that are *biological_process and regulates some 'mitotic cell cycle'*?

Exercise 7 – Verifying properties

- Are we respecting the GO specifications?







Summary GO

- Concepts are defined using one construct only (A SubClassOf P some B).
- Rich RBox
- OWL is helpful to represent these relations, helps to abstract away.

Knowledge integration

- We would like to answer questions over all different source of knowledge.
- *"Thrombosis is a widespread condition and a leading cause of death in the UK."*
- We would like to find a new protein target in order to treat thrombosis.
- Here we would like to know "*what are the human proteins that regulates the blood coagulation*".

Knowledge-bases

- Species: NCBI taxonomy
- Biological Process: Gene Ontology
- Proteins: Uniprot

Exercise 8 – Integrating knowledge

- Open the file uniprot.owl
- Do you understand its content? Look for the class "Protein"
- Now open the file "integrated.owl"
- How would you formulate the question "*what are the human proteins that regulates the blood coagulation*" in OWL?
- *involved_in some (regulates some 'blood coagulation') and expressed_in some 'Homo sapiens'*

Implementation using Brain

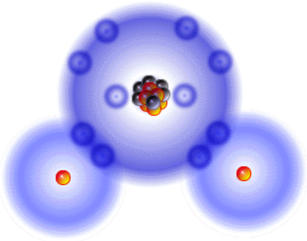
```
Brain brain = new Brain();

brain.learn("data/gene_ontology.owl");
brain.learn("data/NCBI-taxonomy-mammals.owl");
brain.learn("data/uniprot.owl");

String query = "involved_in some (regulates some GO_0007596) and
expressed_in some NCBI_9606“;
List<String> subClasses = brain.getSubClasses(query,false);

brain.sleep();
```

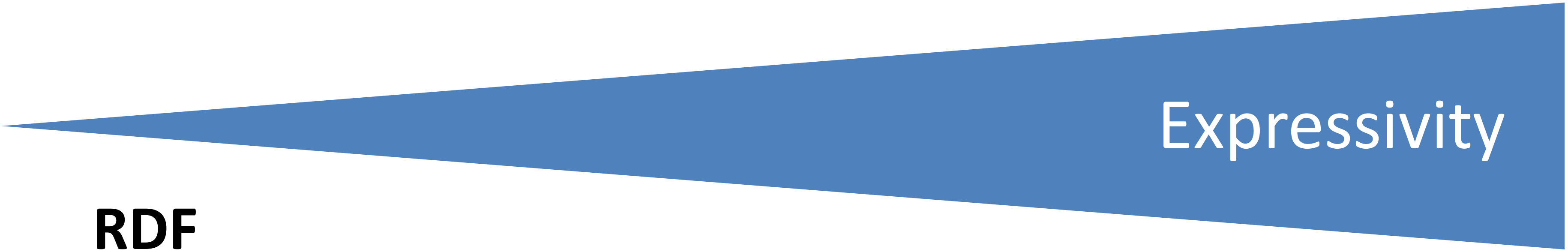
Large-scale implementation

- OWL is computing intensive → **OWL 2EL**
- Less axioms and constructs → easier for you to remember and easier for the reasoner to compute
- Suited for life sciences → lots of classes, few instances

H_2O

H — O — H





RDF
SPARQL

PSPACE
(all constructs)

NP
(AND, FILTER, UNION)

LOGSPACE
(AND, FILTER)

RDFS

PTIME

OWL2 EL

PTIME

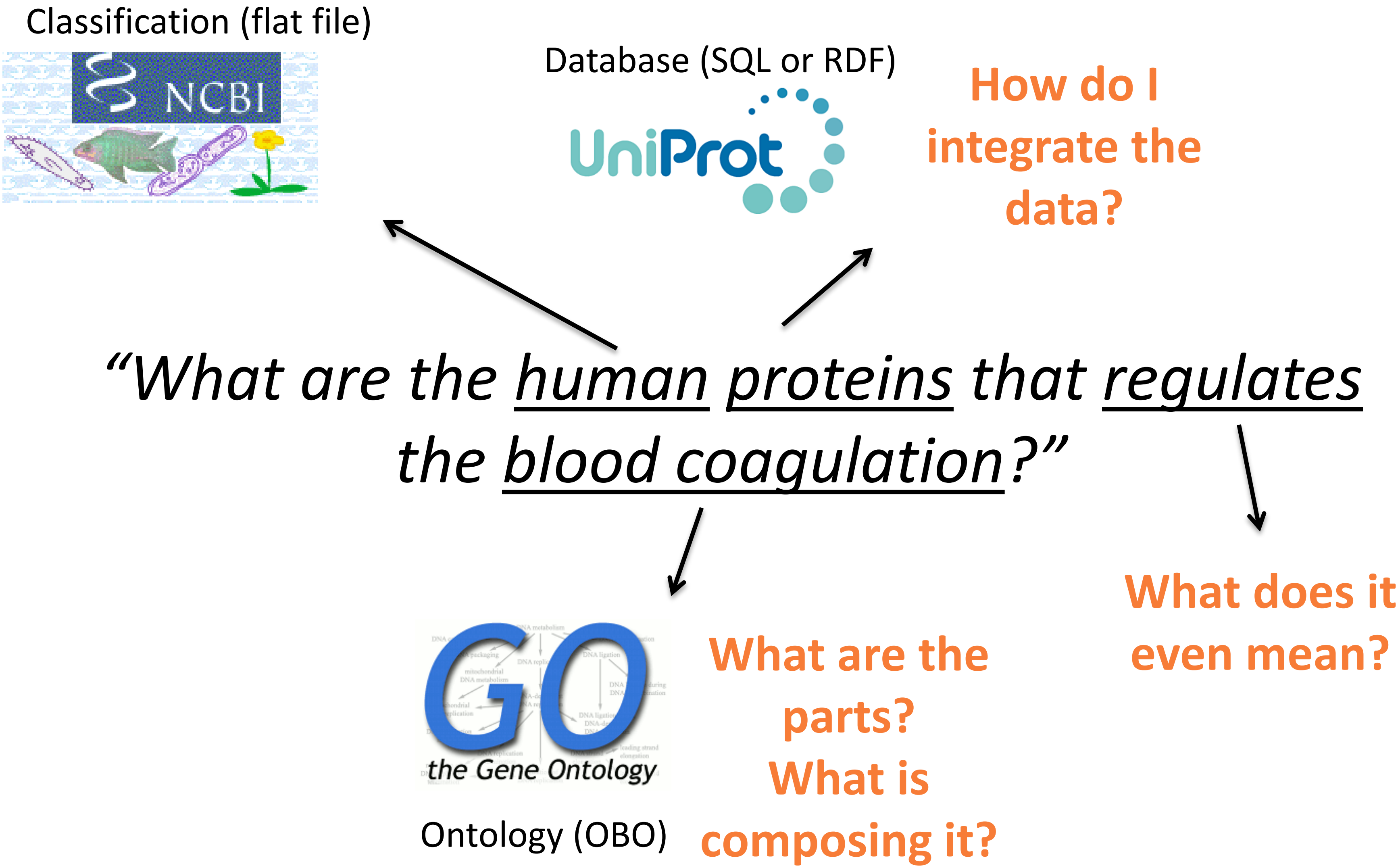
Tractable Parallelism

OWL2

N2EXPTIME-complete

http://www.w3.org/TR/owl2-profiles/

Why learning OWL?



Classification (flat file)

Database (SQL or RDF)

How do I integrate the data?

“What are the human proteins that regulates the blood coagulation?”

What does it even mean?

What are the parts? What is composing it?

Ontology (OBO)

Conclusion

- Ask questions!

- What is OWL?
- Why is it particularly interesting for life sciences?
- How to use OWL?
- What is OWL 2EL?
- How to integrate and query biomedical knowledge?

Thank you!

- croset@ebi.ac.uk
- More questions: StackOverflow (tag “OWL”)
- If you think things could be improved please send feedback, fork or contribute

EMBL-EBI