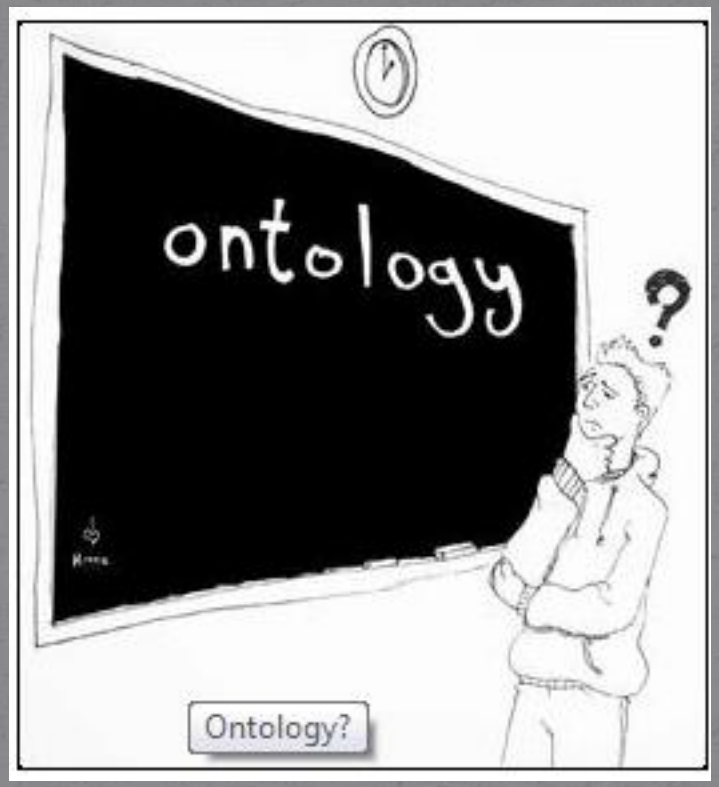
Introduction to ontologies and GO term analysis



- Pantelis Topalis – Bioinformatics Support Group

Metadata

- Data about data.
- Annotation is the process of linking metadata to data.
- Modern databases store and query metadata to facilitate more complicated queries.

Tons of data are available

- High throughput methods yield large amount of data, often annotated in several different but related ways:
  - Annotations related to tissue and developmental stage where a gene is expressed.
  - Geographical distribution of alleles.
  - Resistance to drugs.
  - Ability to transmit a disease.

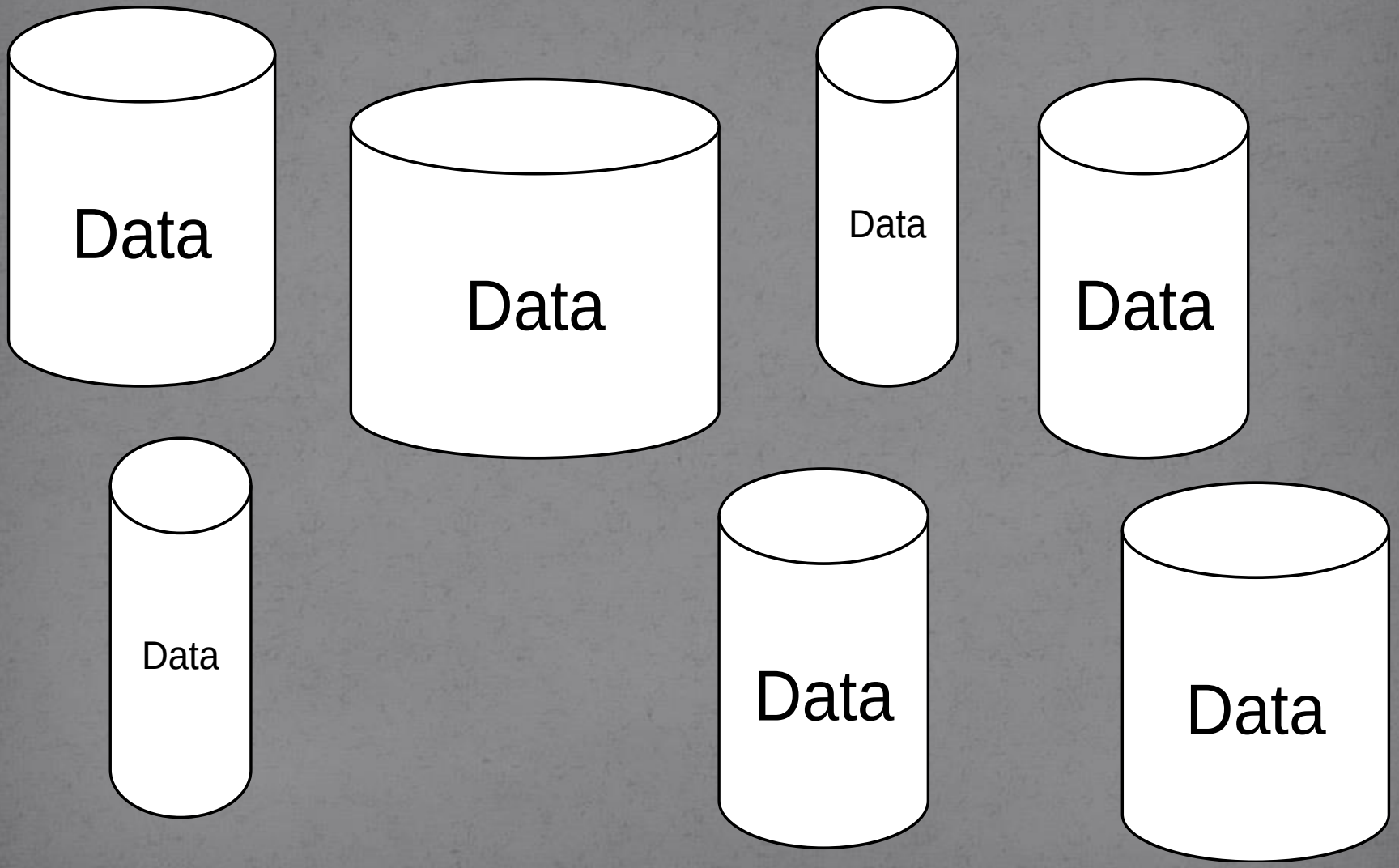
The modern tower of Babel (I)

- Currently data produced by a research project are stored in a homemade database specifically designed to serve the purpose of the project.

- ✓ **Innovative Vector Control Consortium - IVCC**
- ✓ **The WHO/Gates Foundation Vector Biology and Control Project - VBC**
- ✓ **African Network on Vector Research – ANVR**
- ✓ **Malaria Atlas**

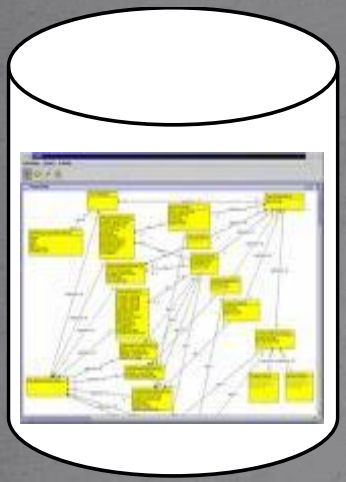
. The modern tower of Babel (II)

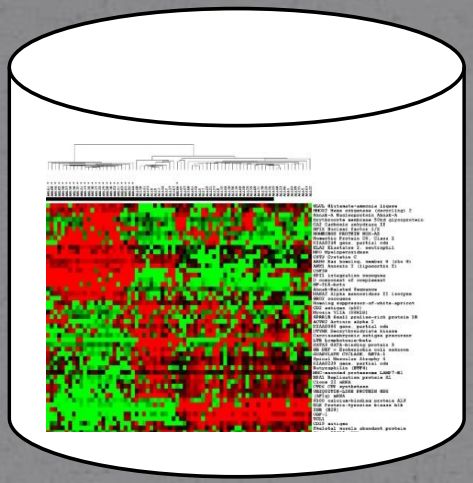
- Not a single format / standard to describe the same datatypes.
  - Genbank vs EMBL
  - Generic Feature Format GFF version 2 vs version 3
  - GVF (Genome Variation Format) vs VCF (Variant Call Format)
  - OBO vs OWL

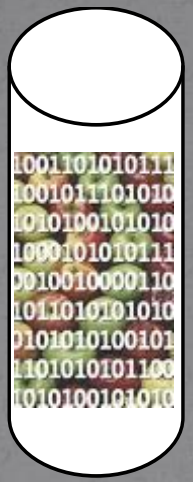

Data
Data
Data
Data
Data
Data
Data
Data

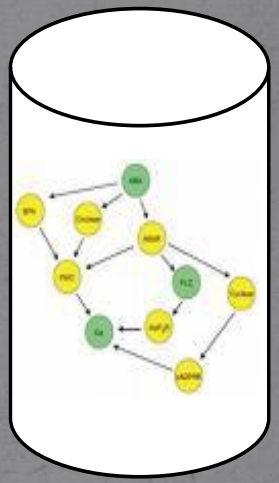
Biological data separated in “silos”

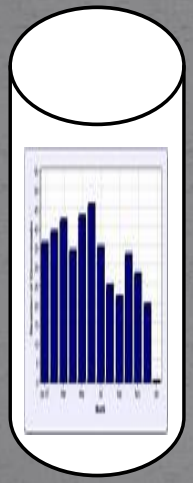
- Lab / pathological data.
- Data related to medical records.
  - Clinical data.
  - Patient history
  - Imaging data
- Gene expression datasets (microarrays, RNAseq).
- Proteomics data (mass spectrometry).
- Genomic Variation data.

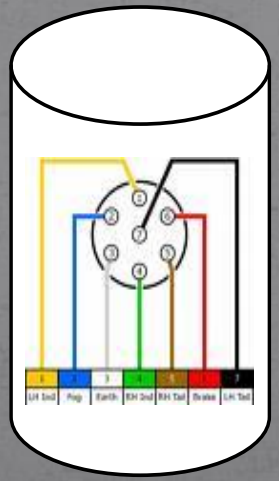


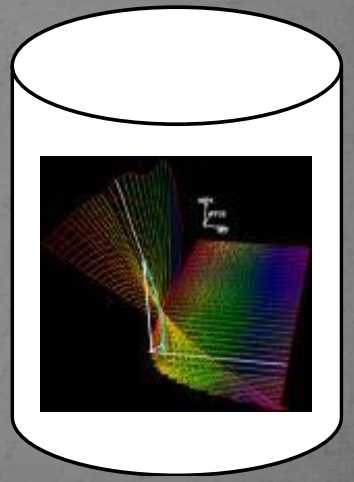










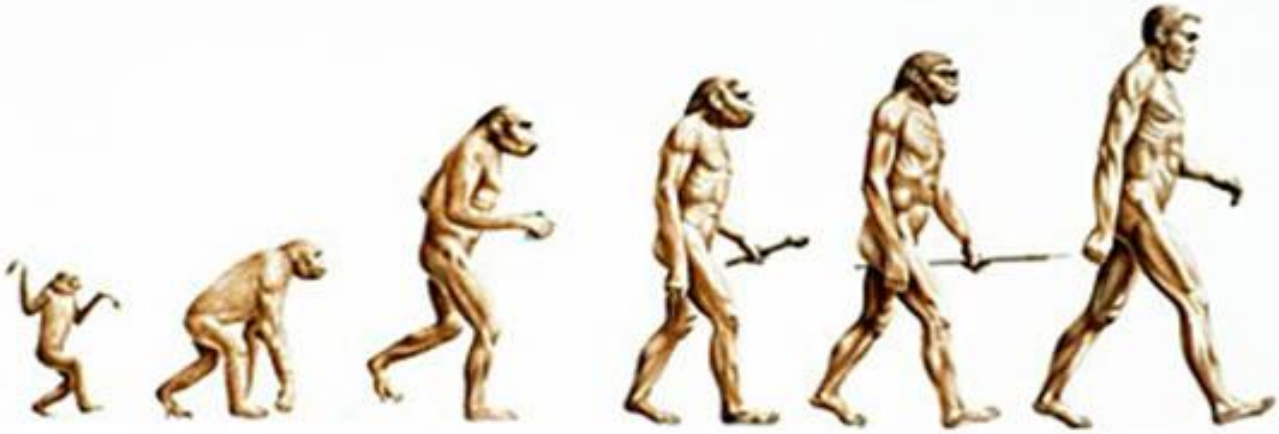


- We need to link all these databases and to make them interoperable.

- Need to develop IT tools and decision support systems to take advantage and retrieve data from all those sources.

Ontology

- Branch of philosophy focusing to everything that exists.
- (IT) Ontology is a typical network of knowledge representation.
- In biomedical domain , ontologies are ways to represent and organize meta-data.



ΟΝΤΟΛΟΓΙΑ

ΘΗΣΑΥΡΟΣ

ΛΕΞΙΛΟΓΙΟ

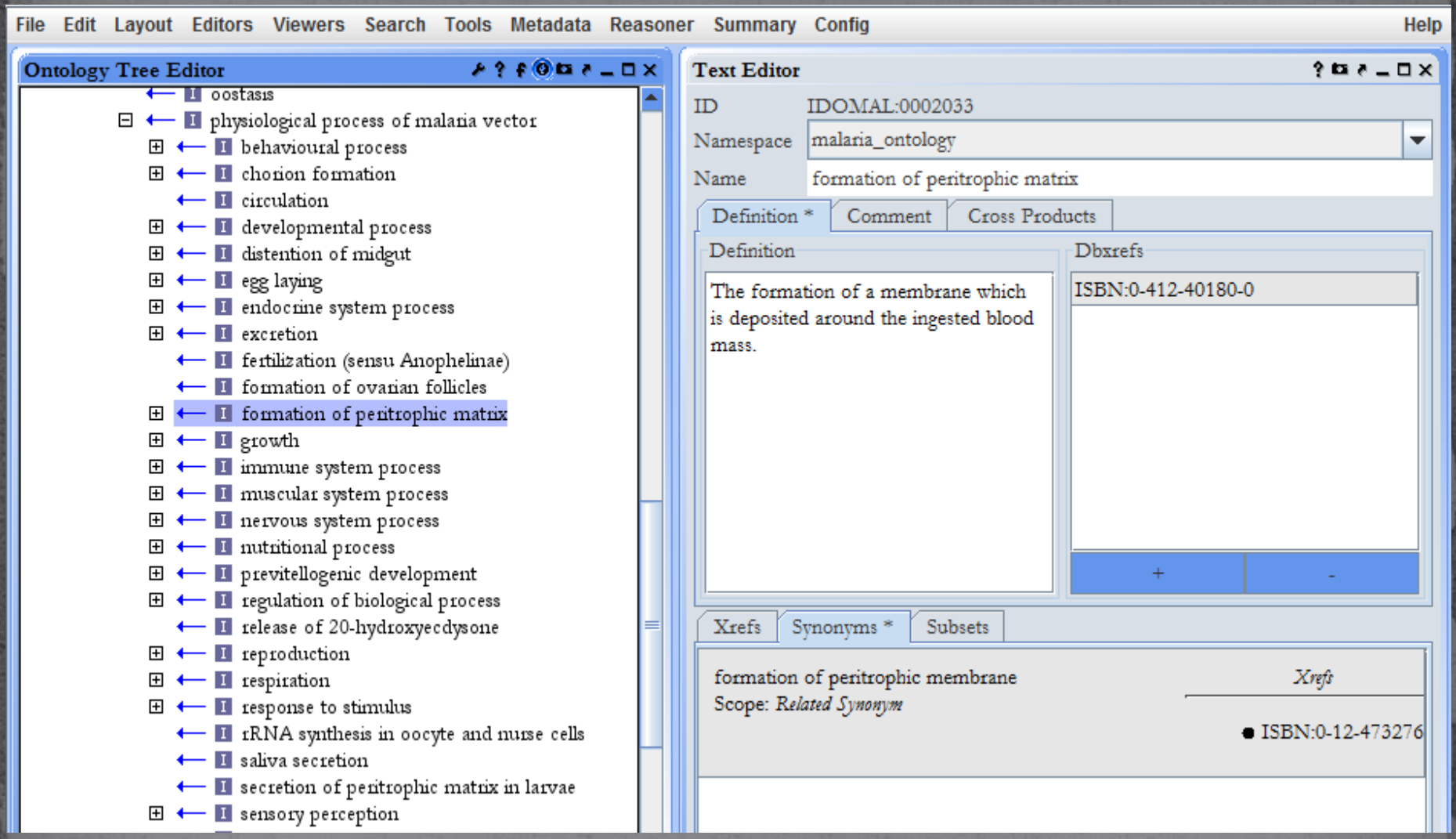
ΤΑΞΙΝΟΜΗΣΗ

ΛΕΞΙΚΟ

ΛΙΣΤΑ ΟΡΩΝ

Basic elements of an ontology

- Terms unambiguously defined.
- Synonyms.
- Connections with the other terms of the ontology via logically defined relations.


File Edit Layout Editors Viewers Search Tools Metadata Reasoner Summary Config Help
Ontology Tree Editor
oostasis
physiological process of malaria vector
behavioural process
chorion formation
circulation
developmental process
distention of midgut
egg laying
endocrine system process
excretion
fertilization (sensu Anophelinae)
formation of ovarian follicles
formation of peritrophic matrix
growth
immune system process
muscular system process
nervous system process
nutritional process
previtellogenic development
regulation of biological process
release of 20-hydroxyecdysone
reproduction
respiration
response to stimulus
rRNA synthesis in oocyte and nurse cells
saliva secretion
secretion of peritrophic matrix in larvae
sensory perception
Text Editor
ID IDOMAL:0002033
Namespace malaria_ontology
Name formation of peritrophic matrix
Definition * Comment Cross Products
Definition
The formation of a membrane which is deposited around the ingested blood mass.
Dbxrefs
ISBN:0-412-40180-0
+ -
Xrefs Synonyms * Subsets
formation of peritrophic membrane
Scope: Related Synonym
Xrefs
ISBN:0-12-473276

Reference and application ontologies

- Reference ontologies describe a scientific domain and provide the basis for the

- Application ontologies which aim to serve the needs of a specific scientific community.

Scientific vs operational ontologies

	Upper level ontologies	Domain specific ontologies
Scientific ontologies	**BFO , Dolce , SUMO**	**GO, FMA, IDO**
Operational ontologies	**FOAF**	**Amazon.com, Library of Congress**

The tower of Babel revisited

- There are many ways to create an ontology but the number of available ontologies is not solving the "data isolation" problem by itself.
- If every community or group creates an ontology using a different jargon interoperability is not promoted.
  - Drug database based on brand names only.
  - Metabolic resistance vs Detoxification process
- Scientific ontologies have to be linked together.

Ways to link ontologies together

- Use common reference ontologies while developing application ontologies for related fields.
  - Disease ontology (DO)
    - Infectious disease ontology (IDO)
      - Malaria ontology (IDOMAL)
      - Dengue Fever ontology (IDODEN)
- Reference ontologies contain generic terms which remain more or less the same, while application ontologies are more flexible to cover the needs as our knowledge is increased.
  - Neither DO or IDO have changed when a fifth serotype of Dengue virus had discovered and added to IDODEN.

Ways to link ontologies (II)

- Common use of relations in all ontologies related to a specific domain.
- OBO Foundry is an international consortium of ontology developers for the biomedical domain has issued a set of rules to facilitate ontology interoperability.

OBO Foundry rules

- Every term has to be related with another term via an *is_a* relation. (*is_a* completeness).
- Every term has only one *is_a* parent.
- There is a predefined set of relations that has to be used in all biomedical ontologies.

Relations available for use in OBO Foundry	
is_a	has_agent
part_of	instance_of
integral_part_of	realizes
proper_part_of	inheres_in
located_in	bearer_of
contained_in	has_quality
adjacent_to	has_function
transformation_of	has_role
derives_from	has_disposition
preceded_by	has_participant

Ontological realism

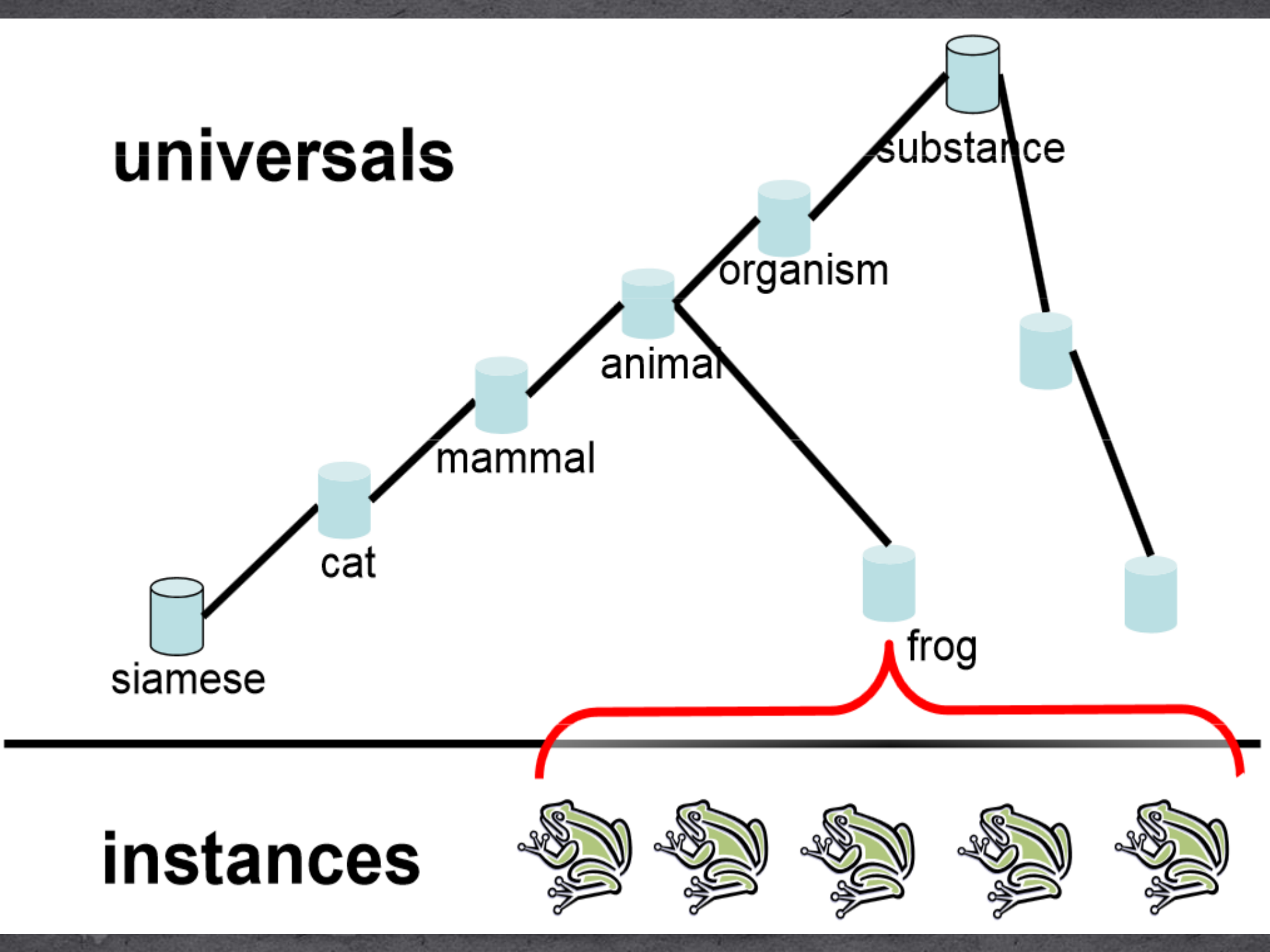
- Describe the scientific domain according to our best knowledge at the present time.
- Add terms and relations in order to describe what is true in reality and not some simplified computer-based model.
- An ontology is the product of collaboration between specialist from several different fields.

Basic elements of an ontology

- Terms unambiguously defined.
- Synonyms.
- Connections with the other terms of the ontology via logically defined relations.

Some definitions

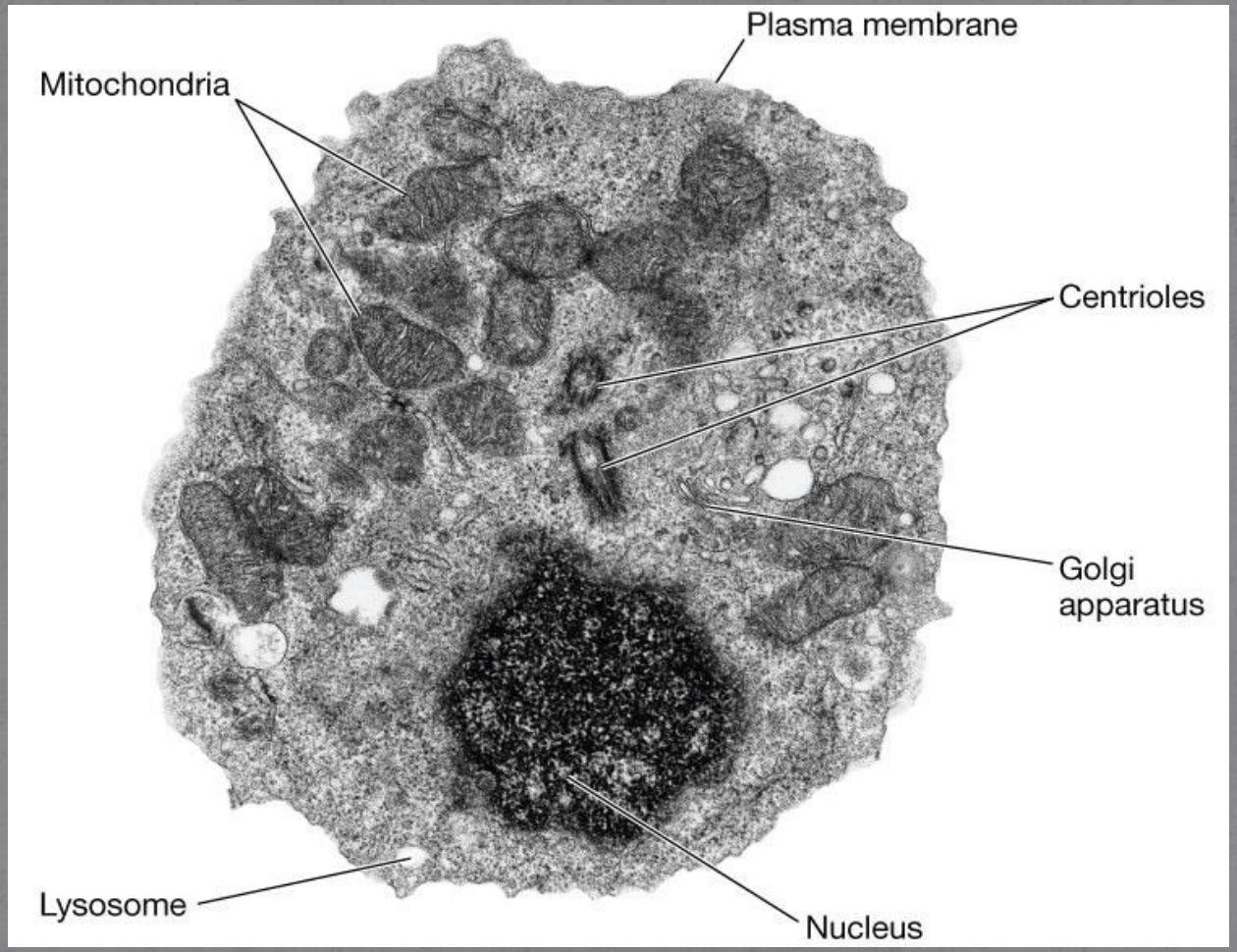
- Entity: Everything that exists: An object, a process, a function, space, text, computer software.
- Universal entities versus instances.
- Human beings versus specific individuals.
- An ontology focuses on universals where as a database focuses on instances.


universals
substance
organism
animal
mammal
cat
siamese
frog
instances

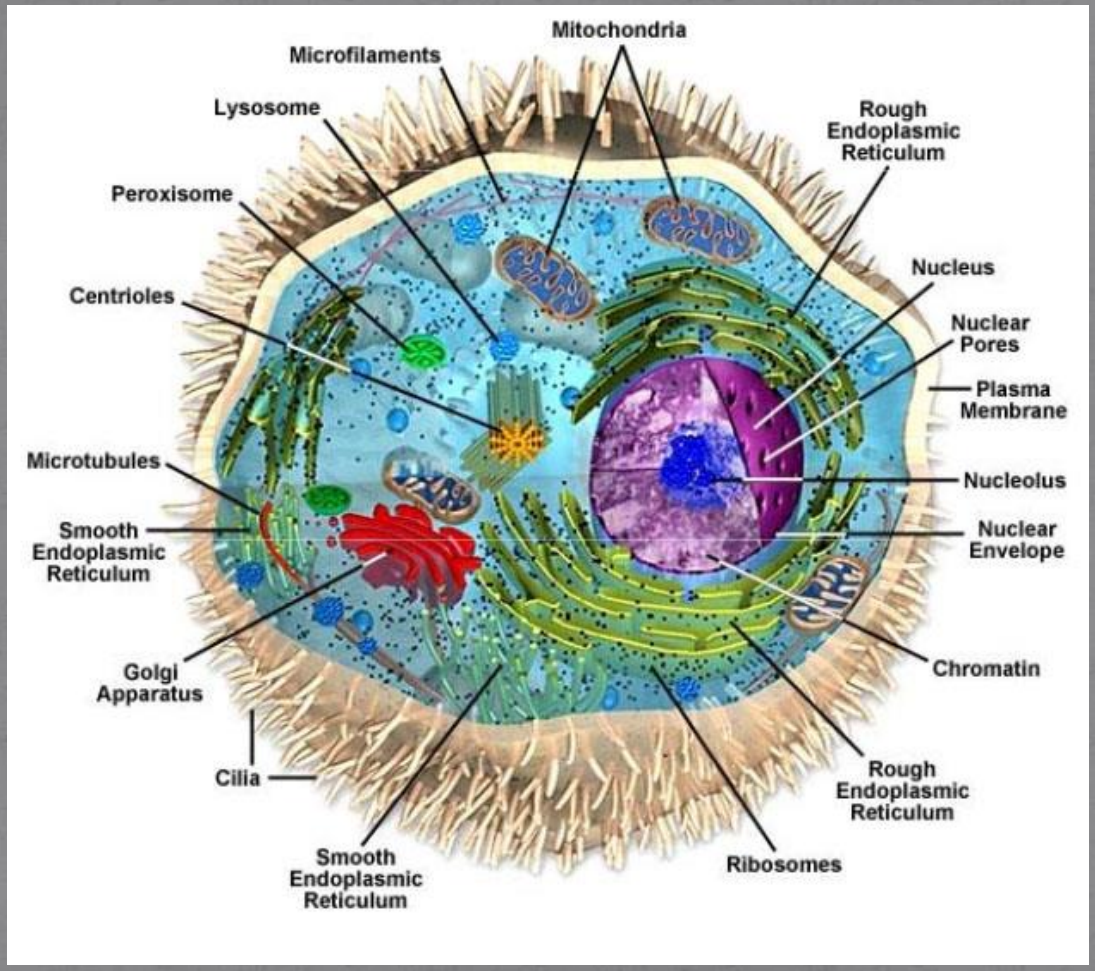
In a scientific ontology

- Every term represents both universal entities and their instances.
- Every term represents only one universal entity.
- Therefore we can define ontology as a representation of universals.

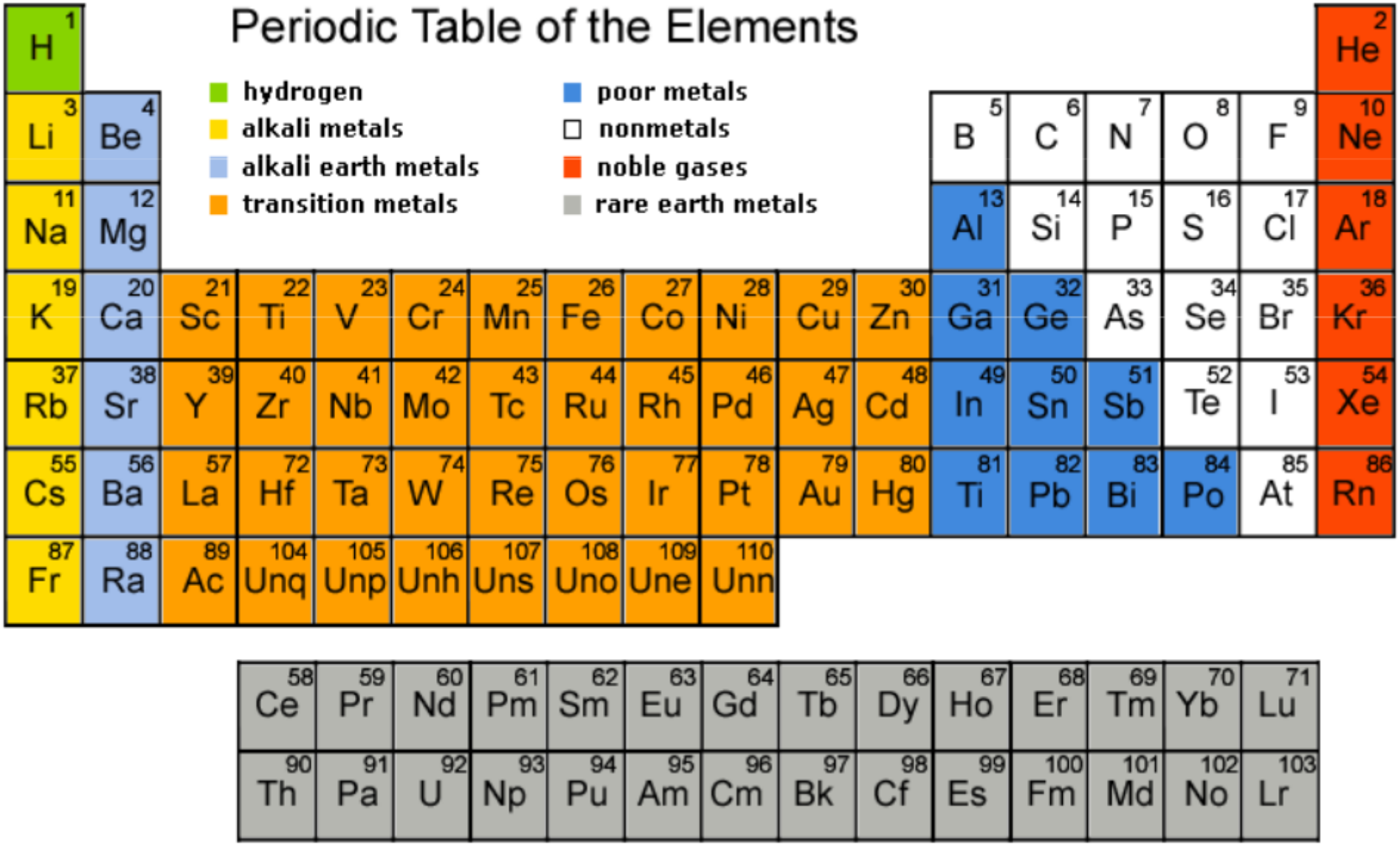
A photograph is a representation of an instance



We can't photograph a universal but we can create it's representation



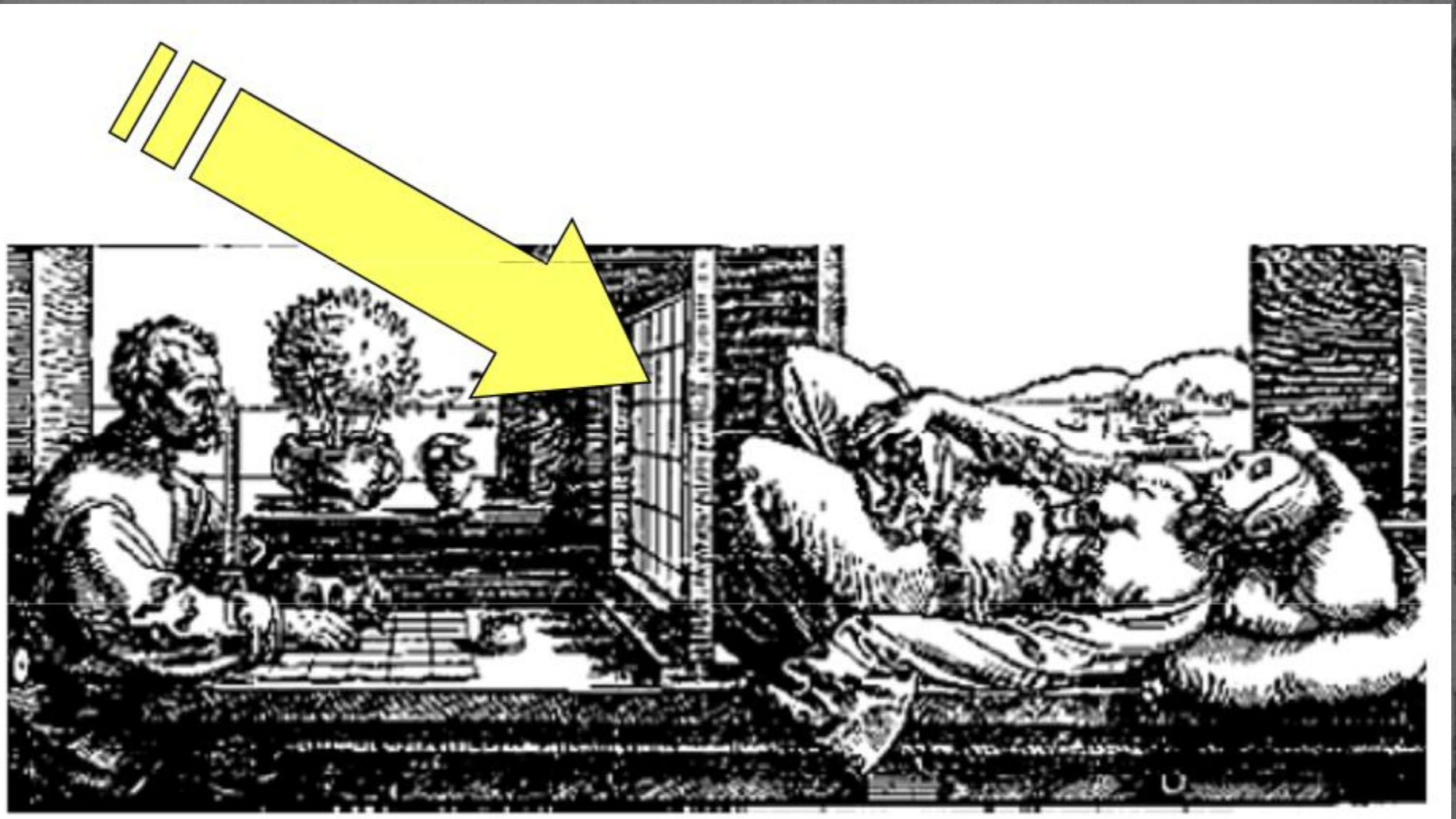
Periodic Table of the Elements

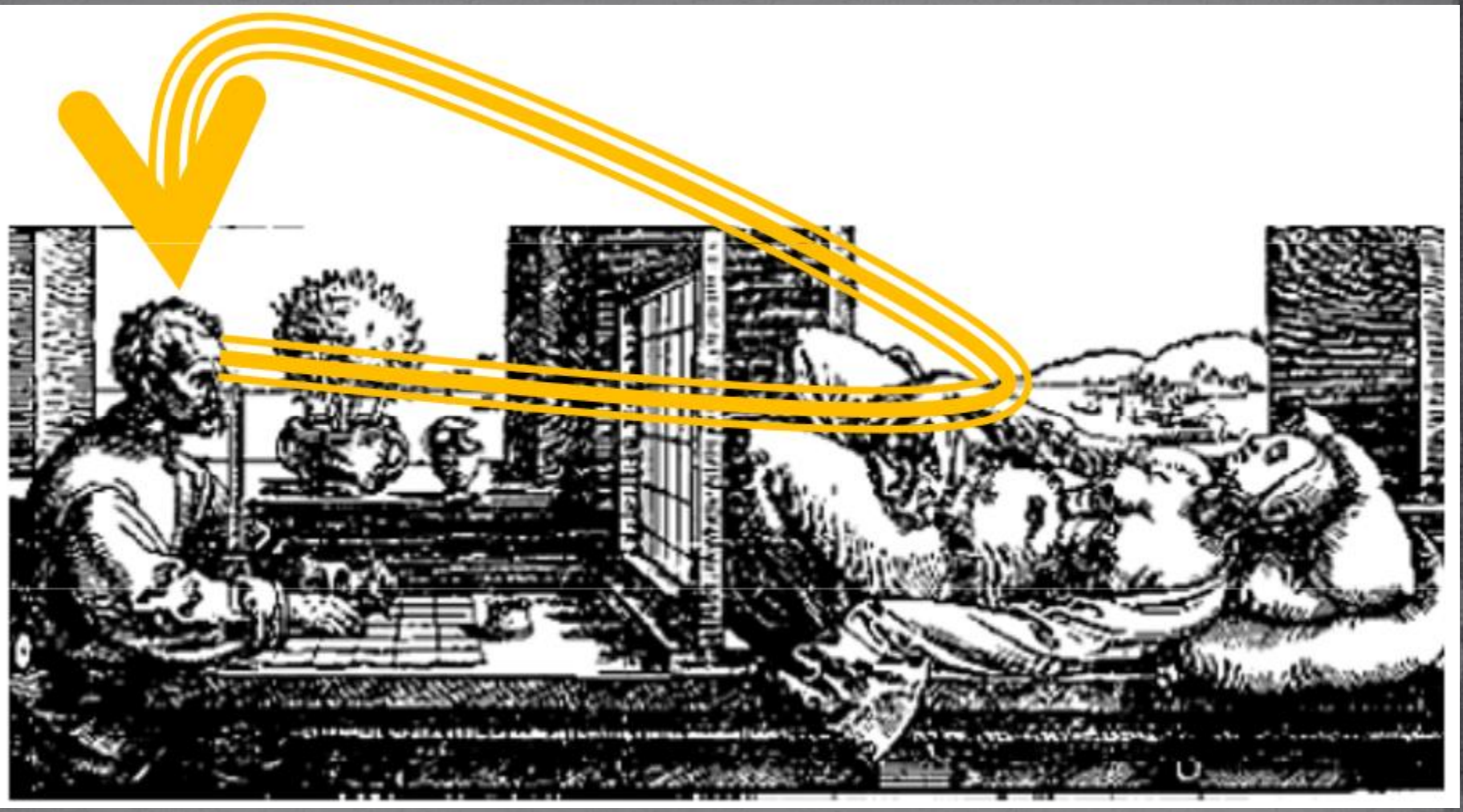


Cognitive representations

Representational artifacts

Reality





Linguistic problems

- Sometimes the same word/term is used for a universal entity and a subset of instances generating confusion.
- Ebola virus can cause death.
- Ebola virus is transmitted uncontrollably in west Africa.

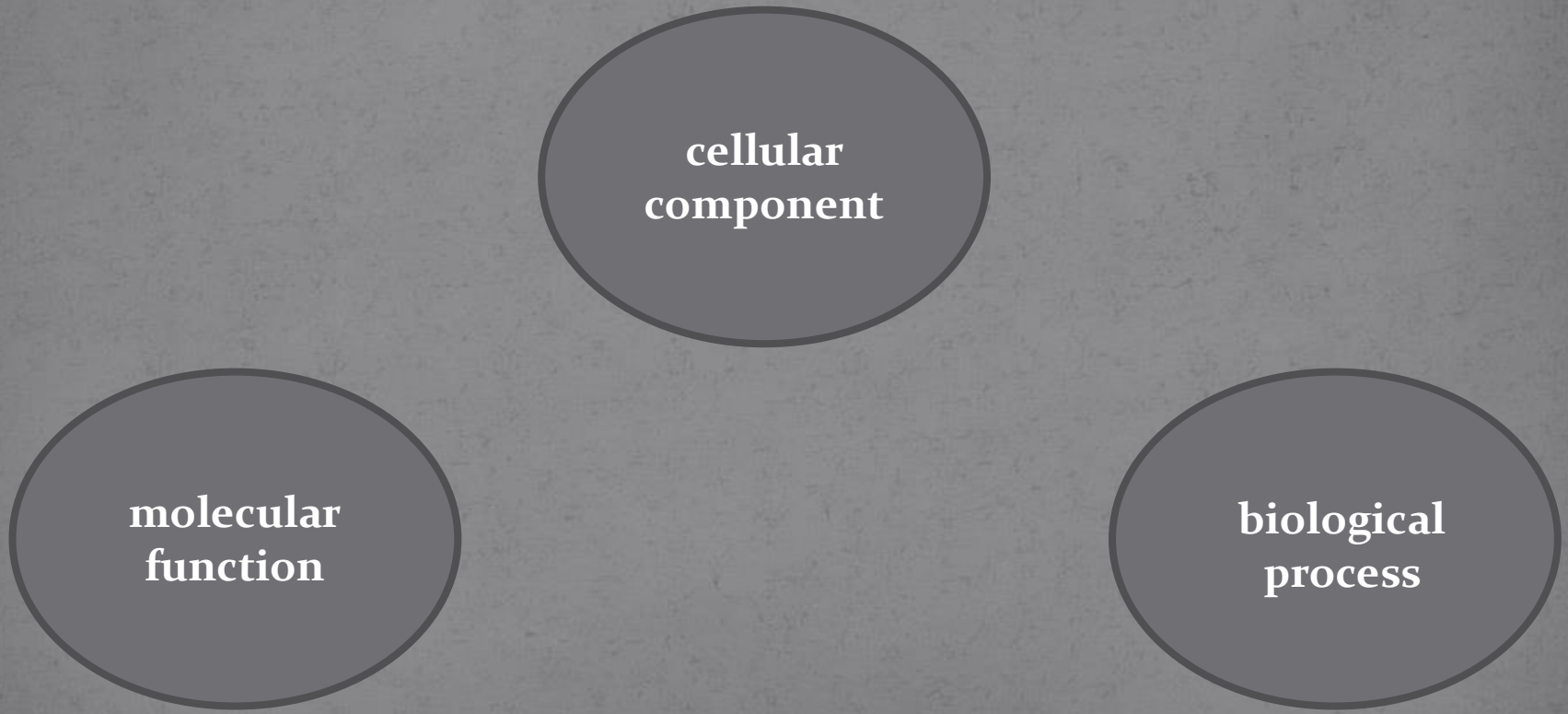
Scientific ontology

- An artifact to represent universals and the relations that relate them

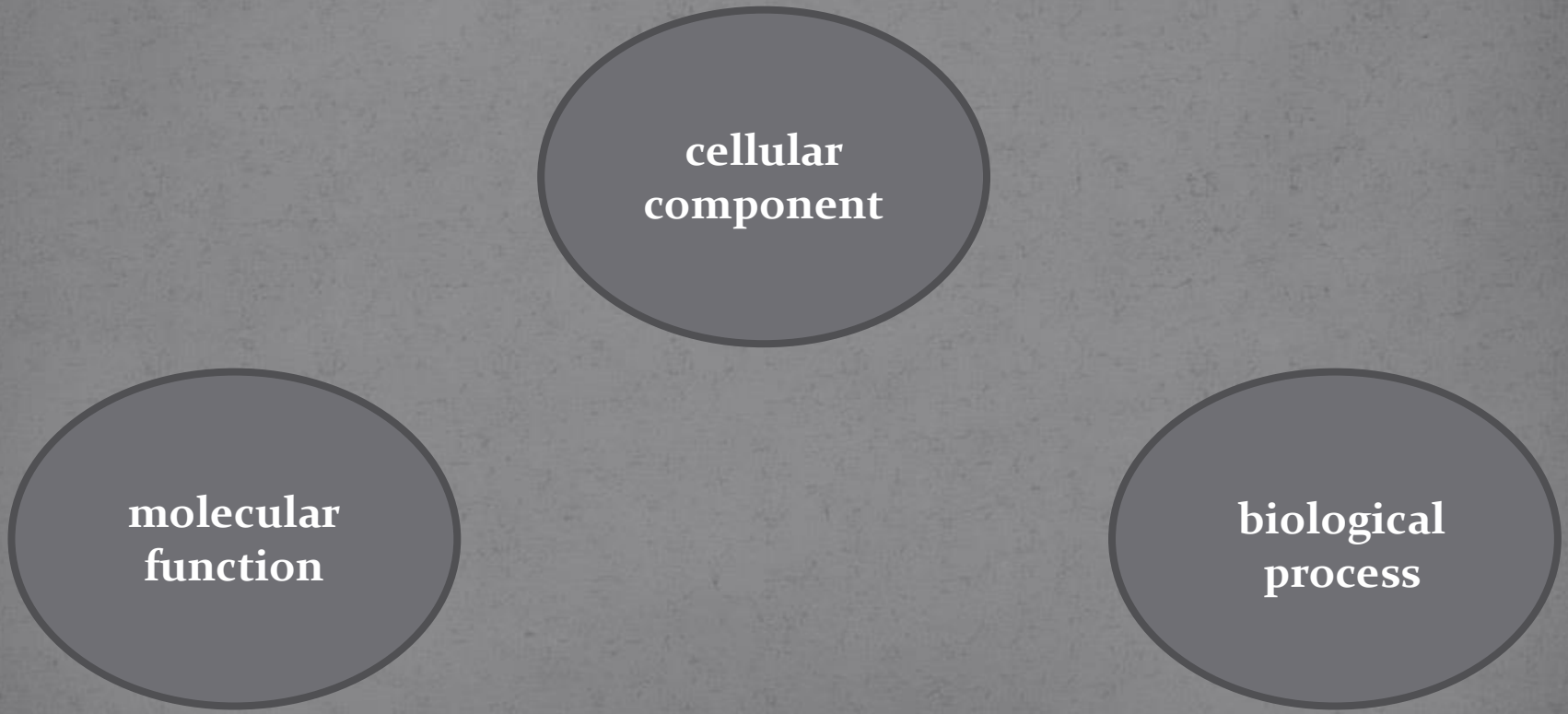
Gene Ontology (GO)

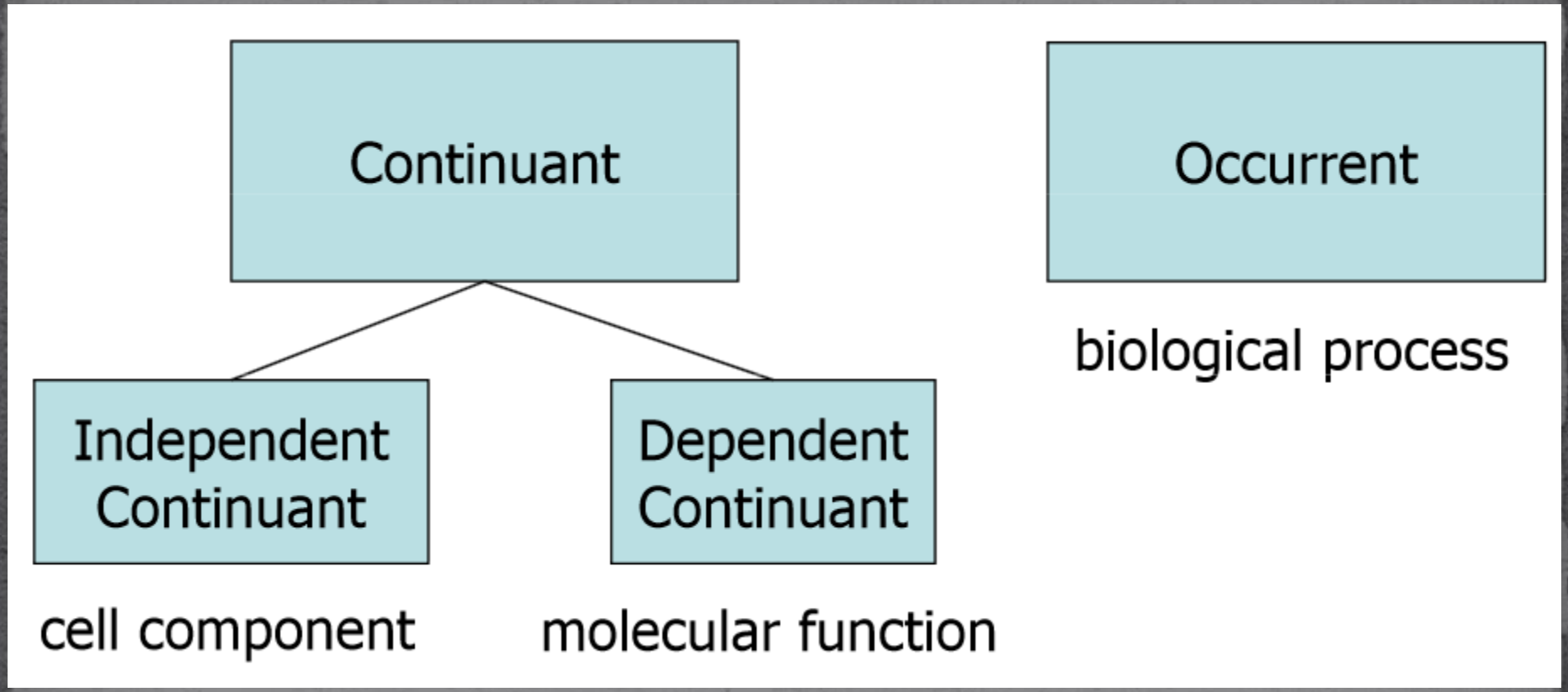
- A structured representation of the qualities / properties of genes available to all interested to the universal biological reality.

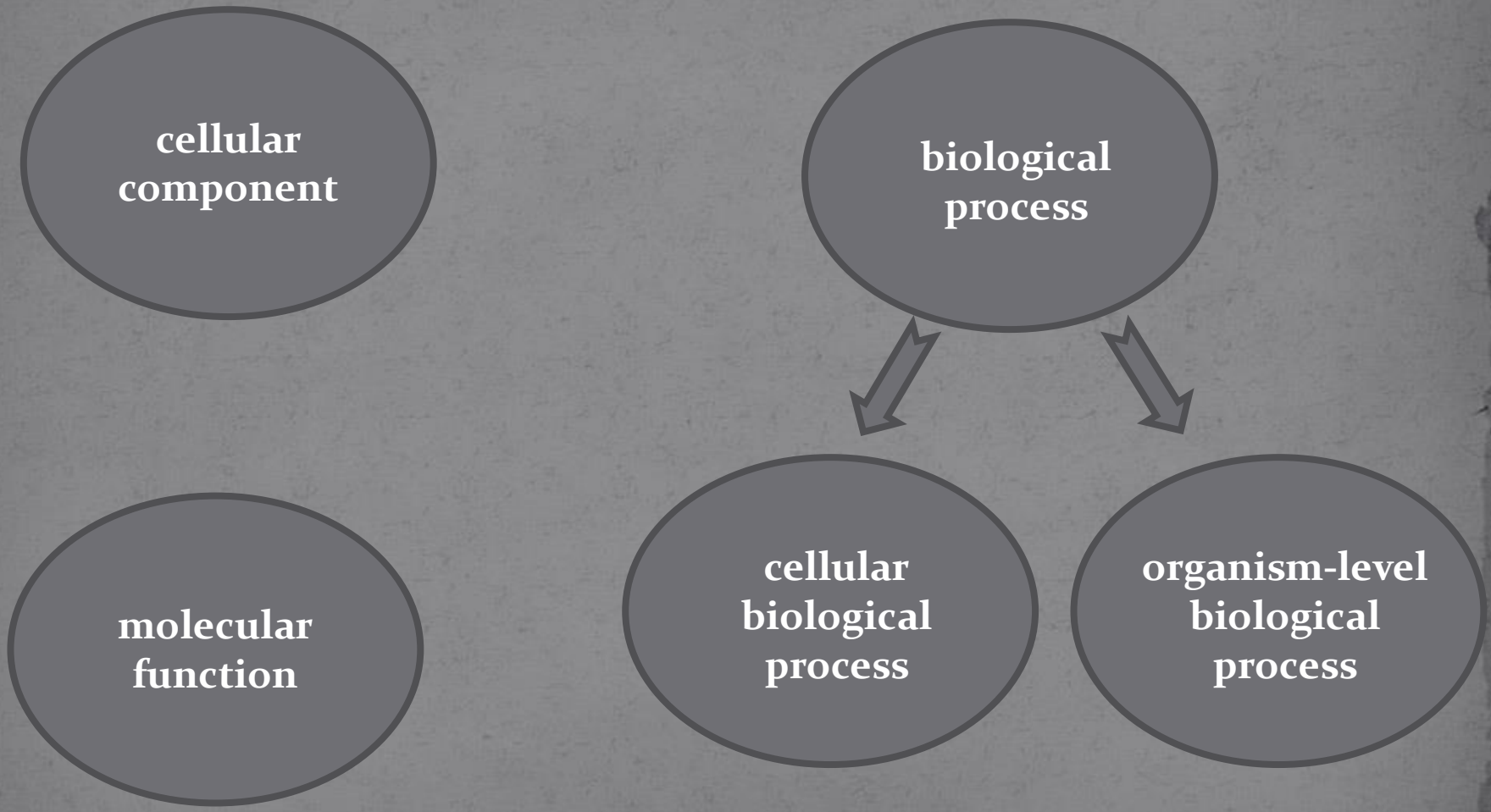
GO is not an ontology, but 3 independent ontologies together

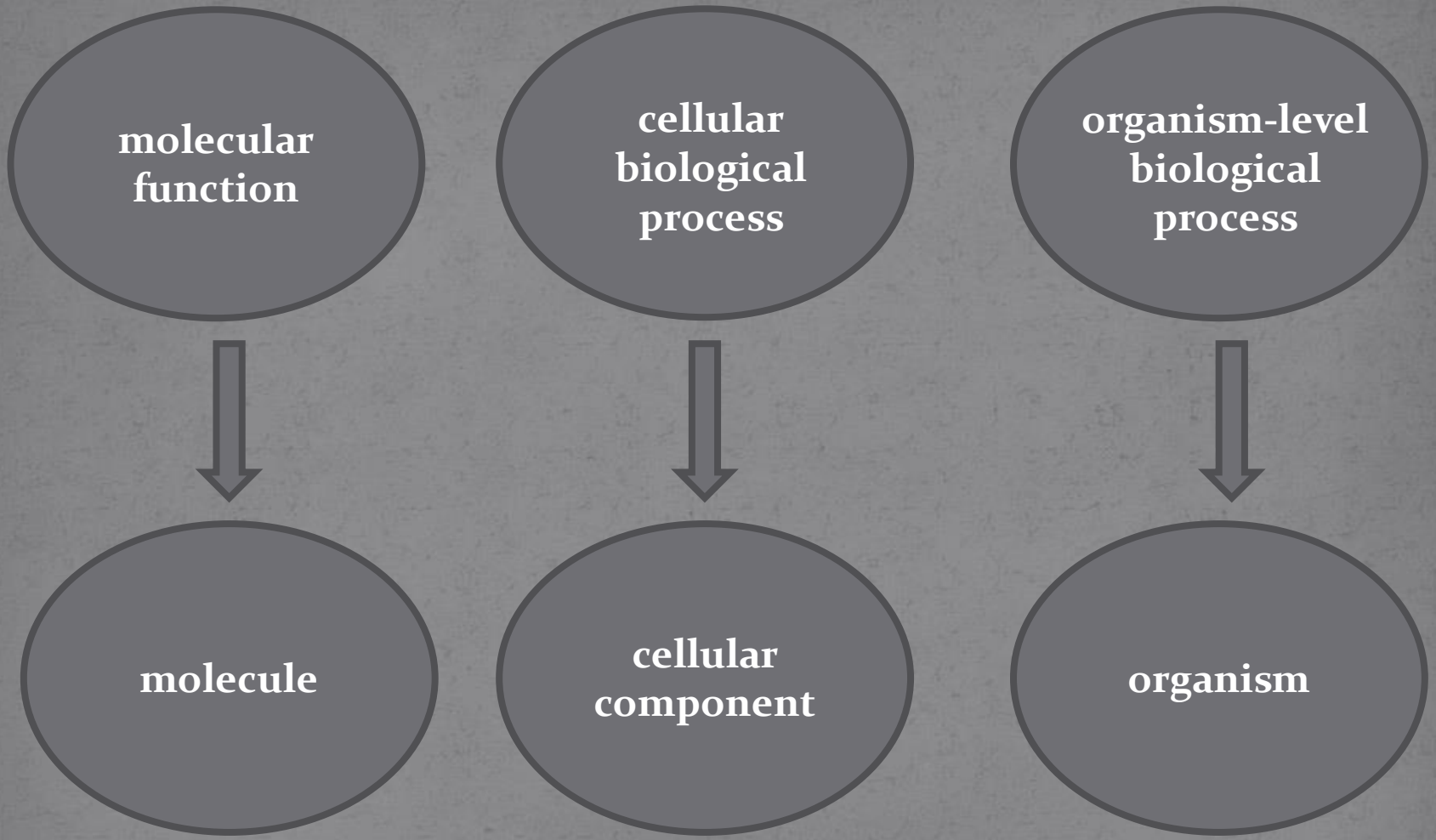


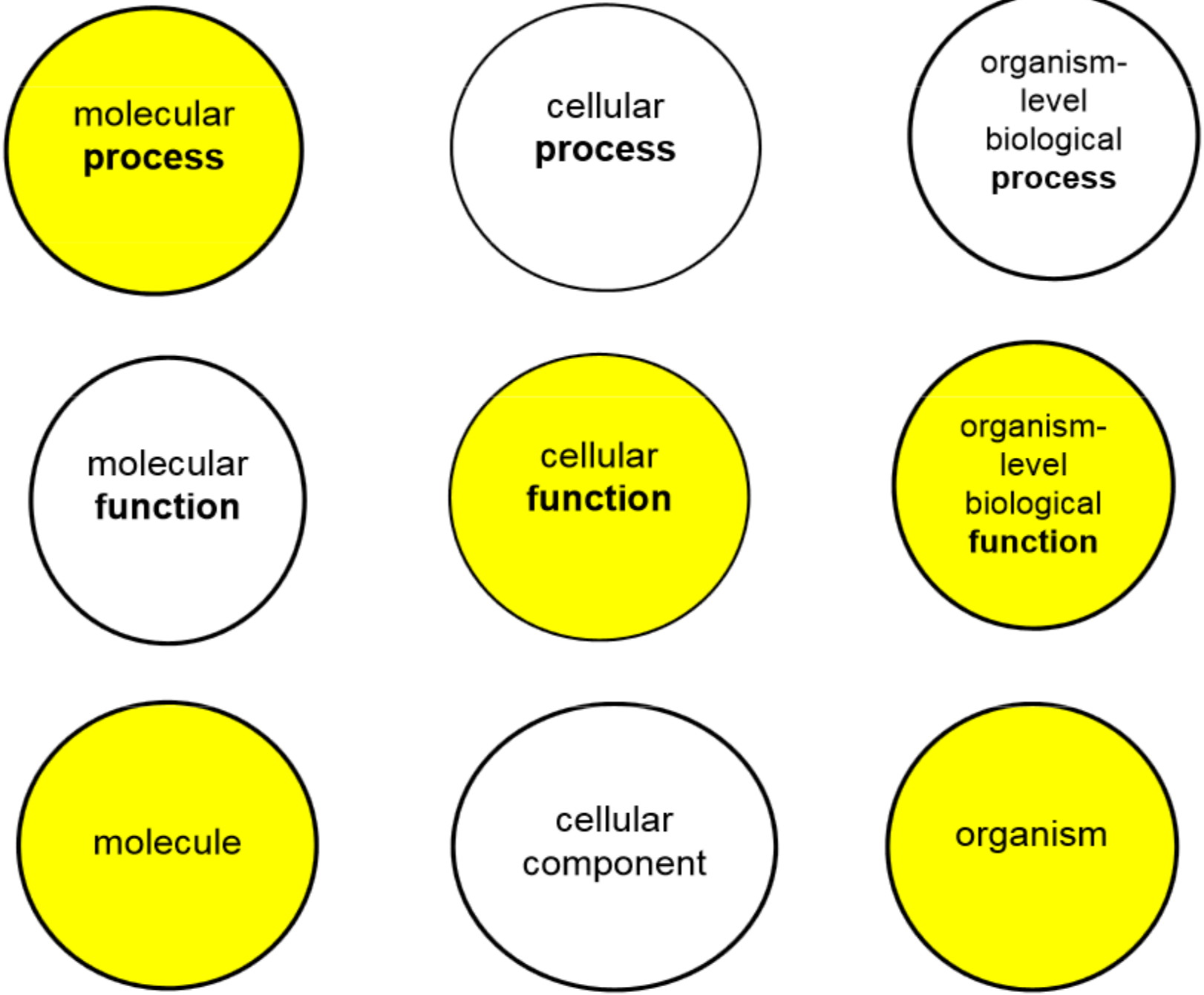
No connections between the 3 main components of GO

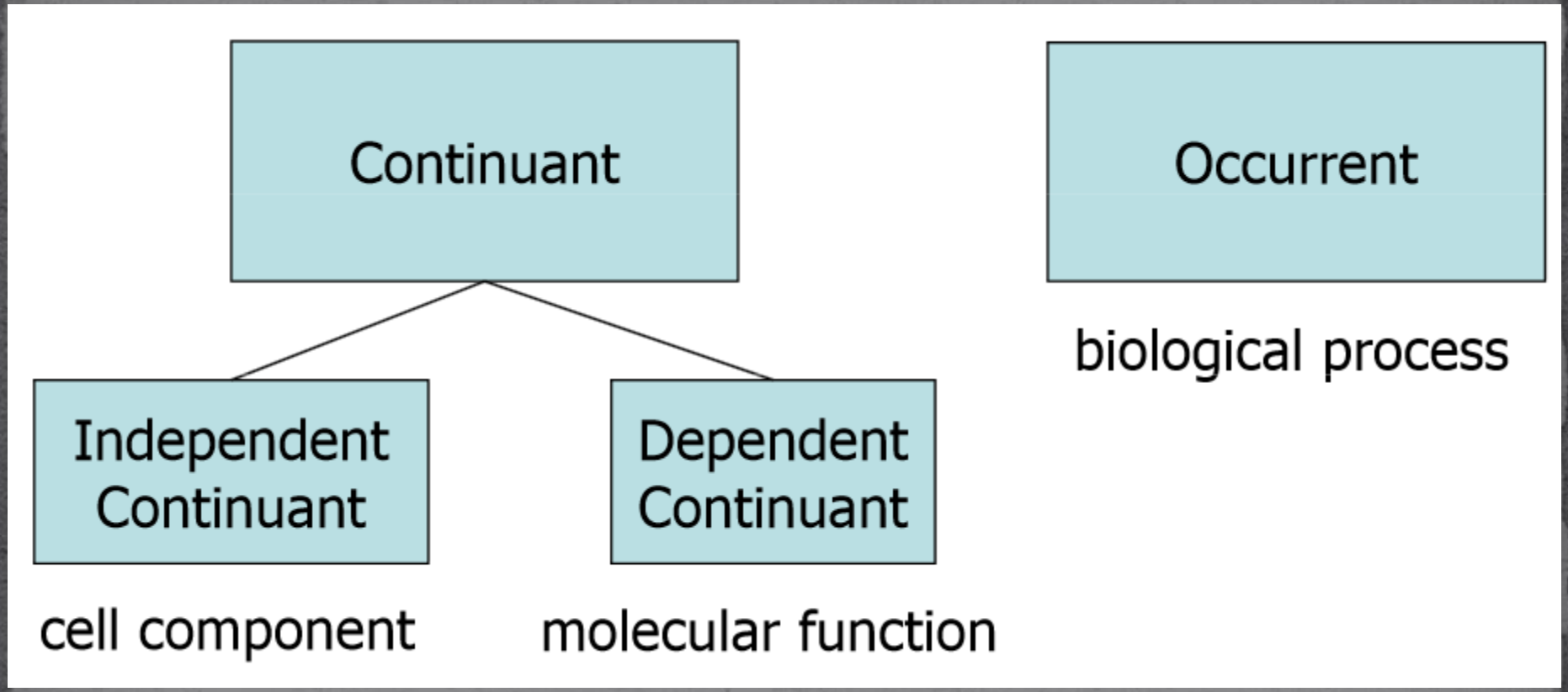



Continuant
Occurrent
biological process
Independent Continuant
Dependent Continuant
cell component
molecular function


cellular component
molecular function
biological process
cellular biological process
organism-level biological process


molecular function
cellular biological process
organism-level biological process
molecule
cellular component
organism

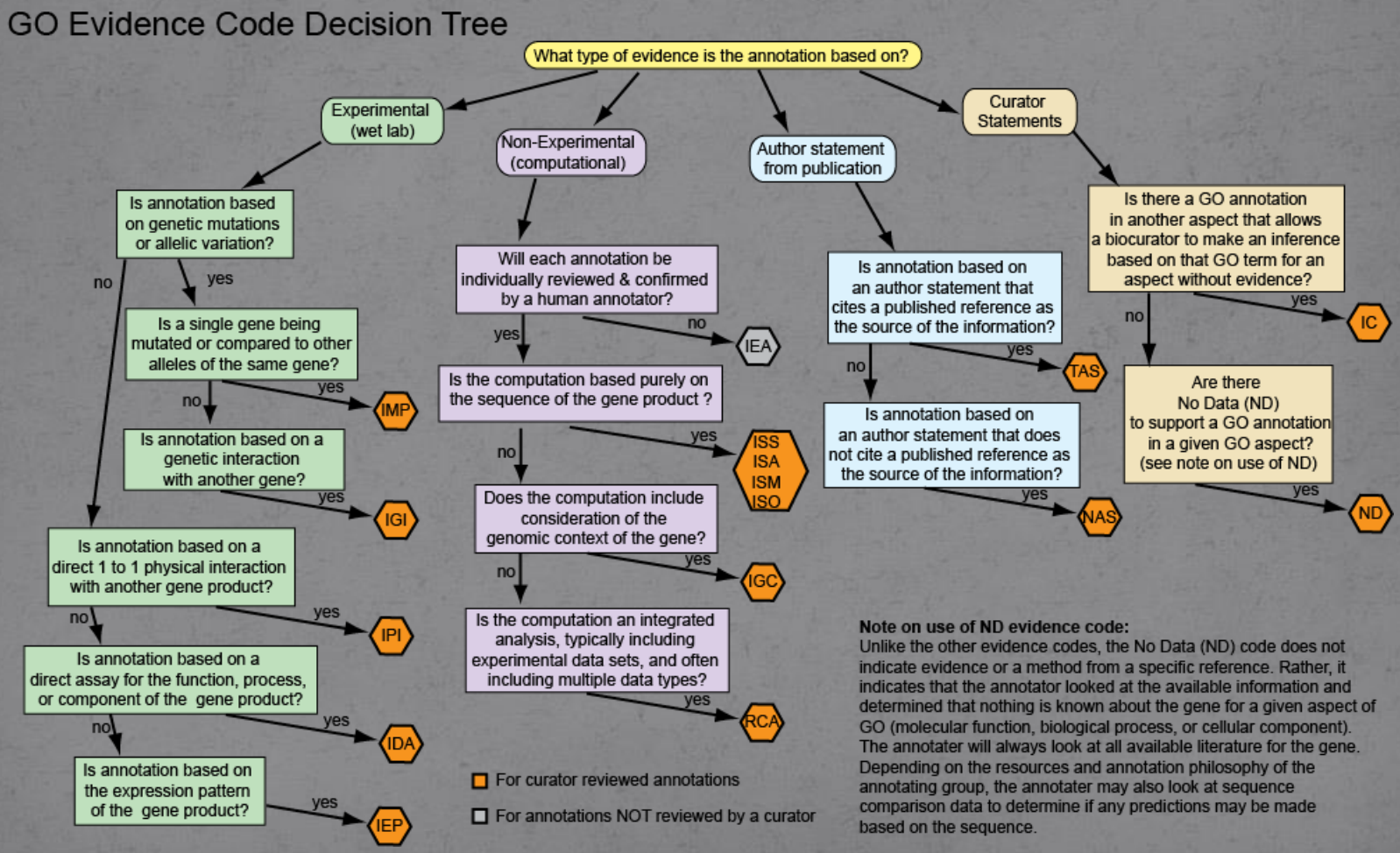

molecular process
cellular process
organism-level biological process
molecular function
cellular function
organism-level biological function
molecule
cellular component
organism


Continuant
Occurrent
biological process
Independent Continuant
Dependent Continuant
cell component
molecular function

GO curation

- GO curators develop the ontology and also provide annotations on behalf of the model organism databases.

Evidence codes



GO usage

- The Gene Ontology (GO) provides core biological knowledge representation for modern biologists, whether computationally or experimentally based.
- GO annotations provide largely species-neutral, comprehensive statements about what gene products do.
- The general user of GO resources often misses fundamental distinctions about GO structures, GO annotations, and what can and can not be extrapolated from GO resources.

Know the Source of the GO Annotations You Use

- The GO site has the most comprehensive and current sets of annotations.
- Annotations are also contained within external tools and applications, although these may not be updated often.

Understand the Scope of GO Annotations

- The GO annotation stream focuses on the capture of the knowledge about the functional activities of specific proteins, the larger biological process (such as photosynthesis) as part of which these specific functions collectively act, and the cellular locality where all this occurs.

Consider Differences in Evidence Codes

- There are 21 evidence codes currently in use to document how data are obtained, and GO may adopt an evidence code classification in the near future. Some evidence codes indicate different classes of experiments such as “inferred by direct assay.” Some indicate different approaches to prediction from comparative analysis such as “inferred from sequence orthology.”

Probe Completeness of GO Annotations

- While the GO annotation corpus generally has excellent broad coverage of available knowledge about gene products via structurally based annotations from sources such as InterPro, the completeness of annotations derived from biomedical literature is uneven because there are not enough GO curators to keep them current on all fronts.

- The lack of a GO annotation does not mean that a gene product does not perform a particular function or act in a particular role.

Understand the Complexity of the GO Structure

- The GO structure, relations, and terminology are modified every day by GO ontology editors. New relations between terms are added as the GO refines the representation of biological knowledge.

- There is a need to know the version of the ontology which is being used by our analysis tool.

Choose Analysis Tools Carefully

- Hundreds of GO-focused applications and tools are available.
- Term enrichment analysis, a common use of GO resources, is incorporated into many different applications and analysis tool sites. Different implementations/algorithms may give different results.

AmiGO 2

More information on quick search

Quick search

Search

Get Started with Grebe

Use the Grebe Search Wizard to **get started** in exploring the Gene Ontology data.

Go »

Advanced Search

Interactively **search** the Gene Ontology data for annotations, gene products, and terms using a powerful search syntax and filters.

Search

GOOSE

Use GOOSE to query a legacy GO database with **SQL** or edit one of the templates.

Go »

Term Enrichment Service

Your genes here...

biological process

H. sapiens

Submit

Powered by PANTHER

Advanced »

Statistics

View the most recent **statistics** about the Gene Ontology data on the main site.

Go »

And Much More...

Many **more tools** are available from the software list, such as alternate searching modes, Visualize, non-JavaScript pages.

Go »

http://www.amigo.org

Term annotations

What are the direct and indirect annotations to term [ ] ? Go »

What are the direct and indirect annotations for organism (scientific name) [ ] to term [ ] ? Go »

Gene product annotations

What are all the annotations for gene product [ ] Go »

Gene products associated with terms

What are the gene products annotated to term [ ] but *not* term [ ] ? Go »

What are the gene products annotated to term [ ] *and* term [ ] ? Go »

Protein family

What are the annotations associated with the protein family [ ] ? Go »

AmiGO 2

More information on quick search

Quick search

Search

Get Started with Grebe

Use the Grebe Search Wizard to **get started** in exploring the Gene Ontology data.

Go »

Advanced Search

Interactively **search** the Gene Ontology data for annotations, gene products, and terms using a powerful search syntax and filters.

Search ▾

GOOSE

Use GOOSE to query a legacy GO database with **SQL** or edit one of the templates.

Go »

Term Enrichment Service

Your genes here...

biological process ▾

H. sapiens ▾ Submit

Powered by PA ER

Advanced »

Statistics

View the most recent **statistics** about the Gene Ontology data on the main site.

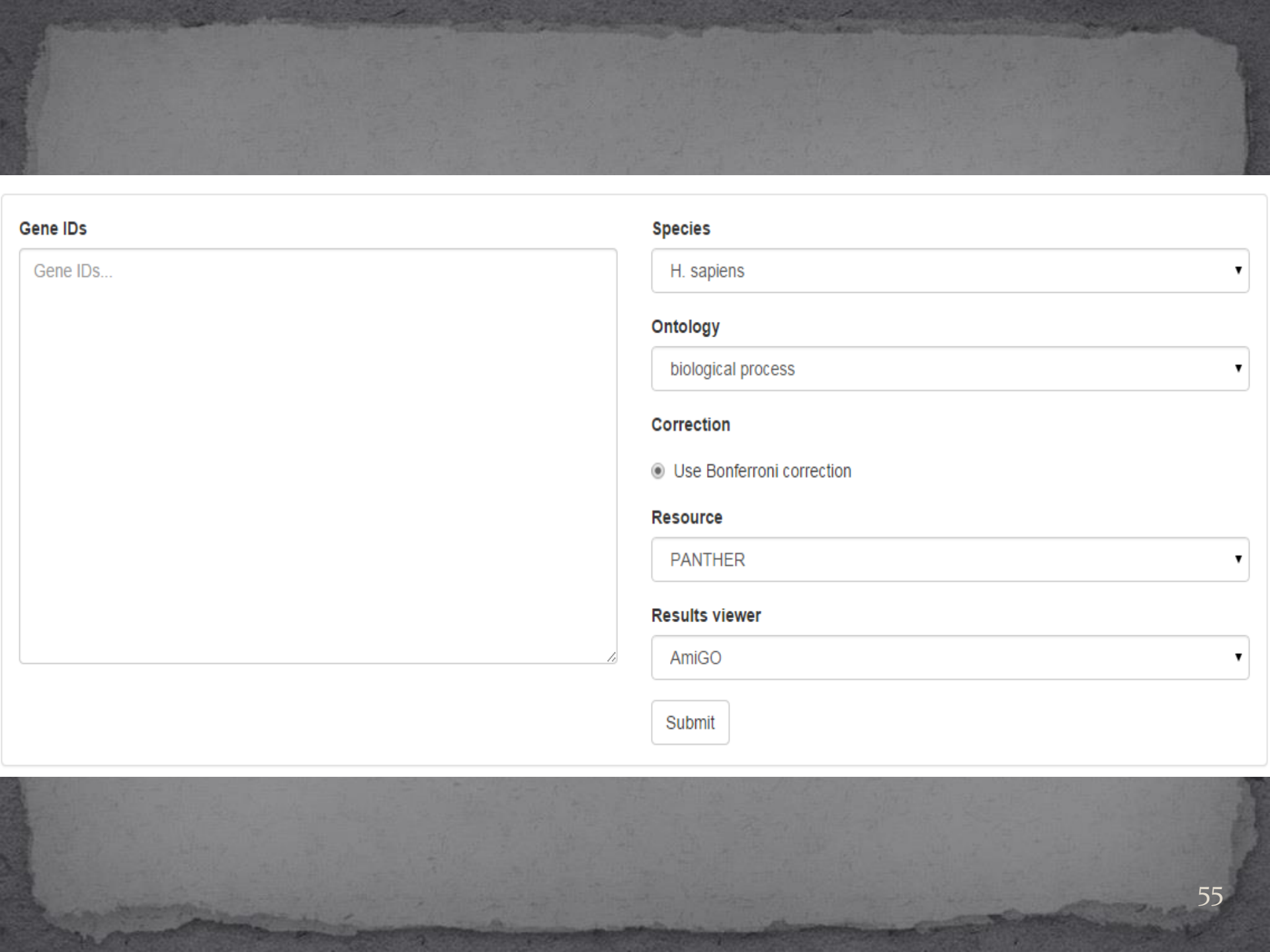
Go »

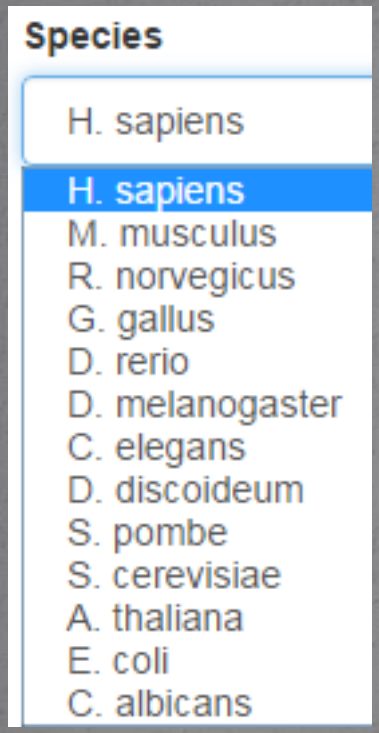
And Much More...

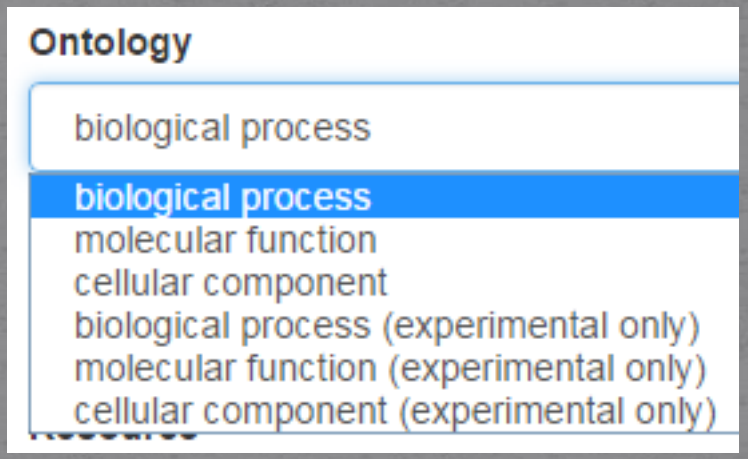
Many **more tools** are available from the software list, such as alternate searching modes, Visualize, non-JavaScript pages.

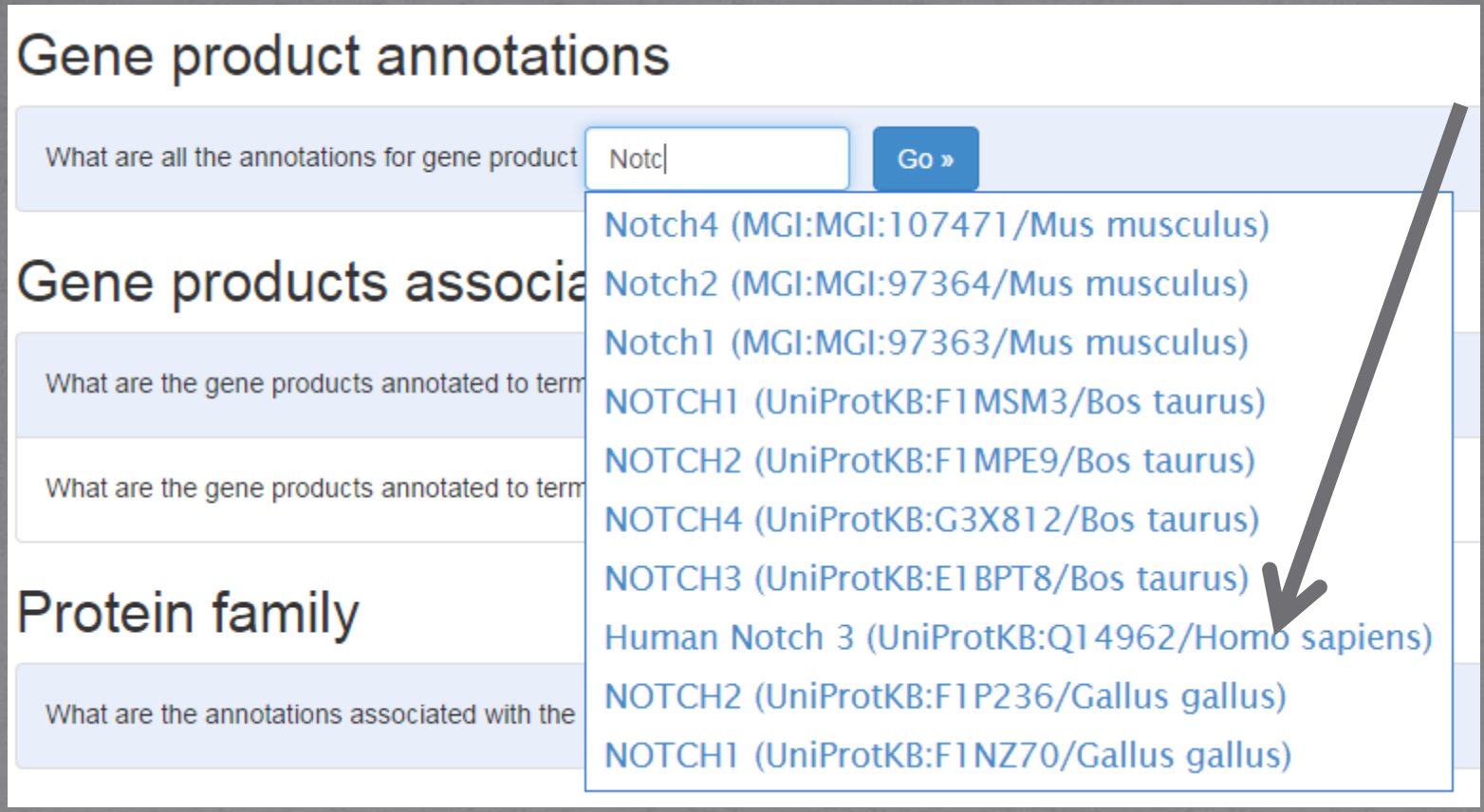
Go »

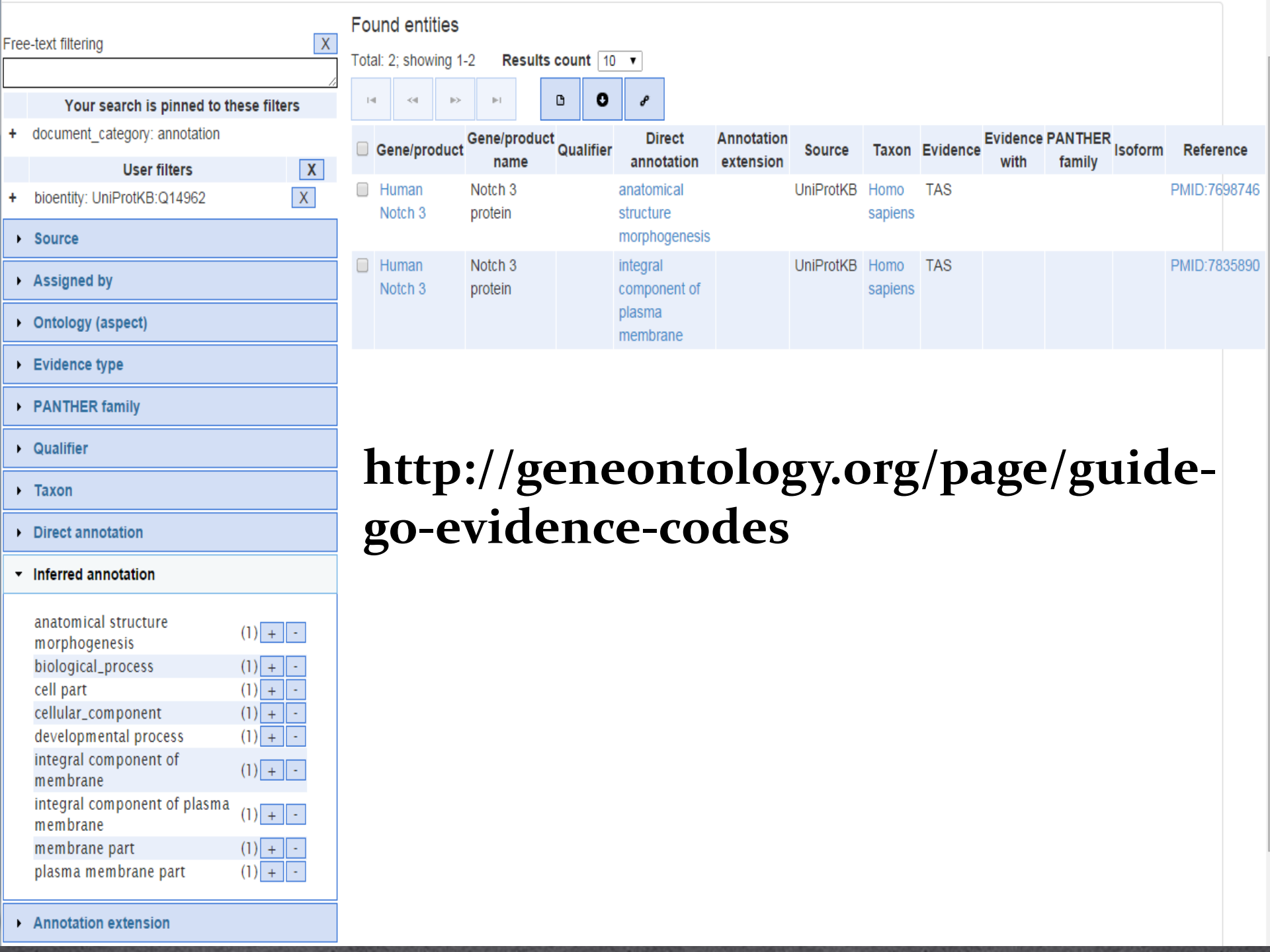
http://www.amigo.org


Gene IDs
Gene IDs...
Species
H. sapiens
Ontology
biological process
Correction
Use Bonferroni correction
Resource
PANTHER
Results viewer
AmiGO
Submit


Species
H. sapiens
H. sapiens
M. musculus
R. norvegicus
G. gallus
D. rerio
D. melanogaster
C. elegans
D. discoideum
S. pombe
S. cerevisiae
A. thaliana
E. coli
C. albicans


Ontology
biological process
biological process
molecular function
cellular component
biological process (experimental only)
molecular function (experimental only)
cellular component (experimental only)


Gene product annotations
What are all the annotations for gene product
Notc
Go »
Notch4 (MGI:MGI:107471/Mus musculus)
Notch2 (MGI:MGI:97364/Mus musculus)
Notch1 (MGI:MGI:97363/Mus musculus)
NOTCH1 (UniProtKB:F1MSM3/Bos taurus)
NOTCH2 (UniProtKB:F1MPE9/Bos taurus)
NOTCH4 (UniProtKB:G3X812/Bos taurus)
NOTCH3 (UniProtKB:E1BPT8/Bos taurus)
Human Notch 3 (UniProtKB:Q14962/Homo sapiens)
NOTCH2 (UniProtKB:F1P236/Gallus gallus)
NOTCH1 (UniProtKB:F1NZ70/Gallus gallus)
Gene products associa
What are the gene products annotated to term
What are the gene products annotated to term
Protein family
What are the annotations associated with the



http://geneontology.org/page/guide-go-evidence-codes

Thank you!