Principles for Building Biomedical Ontologies

ISMB 2005

April 10, 2007

Ontology (as a branch of philosophy)

- *The science of what is:*

  of the kinds and structures of the objects, and their properties and relations in every area of reality.
- In simple terms, it seeks the classification of entities.
- Defined by a scientific field's vocabulary and by the canonical formulations of its theories.
- Seeks to solve problems which arise in these domains.

In computer science, there is an information handling problem

- Different groups of data-gatherers develop their own idiosyncratic terms and concepts of which they represent information.
- To put this information together, methods must be found to resolve terminological and conceptual incompatibilities.
- Again, and again, and again…

The Solution to this Tower of Babel problem

- A shared, common, backbone taxonomy of relevant entities, and the relationships between them within an application domain
- This is referred to by information scientists as an *'Ontology'*.

Which means…
Instances are not included!

- It is the generalizations that are important
- Please keep this in mind, it is crucial to understanding the tutorial

Principles for Building Biomedical Ontologies

Barry Smith

http://ifomis.de

April 10, 2007

Ontologies as Controlled Vocabularies

- expressing discoveries in the life sciences in a uniform way
- providing a uniform framework for managing annotation data deriving from different sources and with varying types and degrees of evidence

Overview

- Following basic rules helps make better ontologies
- We will work through the principles-based treatment of relations in ontologies, to show how ontologies can become more reliable and more powerful

Why do we need rules for good ontology?

- Ontologies must be intelligible both to humans (for annotation) and to machines (for reasoning and error-checking)
- Unintuitive rules for classification lead to entry errors (problematic links)
- Facilitate training of curators
- Overcome obstacles to alignment with other ontology and terminology systems
- Enhance harvesting of content through automatic reasoning systems

First Rule: Univocity

- Terms (including those describing relations) should have the same meanings on every occasion of use.
- In other words, they should refer to the same kinds of entities in reality

Example of univocity problem in case of *part_of* relation

(Old) Gene Ontology:

- ‘part_of’ = ‘may be part of’
  - flagellum part_of cell
- ‘part_of’ = ‘is at times part of’
  - replication fork part_of the nucleoplasm
- ‘part_of’ = ‘is included as a sub-list in’

Second Rule: Positivity

- Complements of classes are not themselves classes.

- Terms such as ‘non-mammal’ or ‘non-membrane’ do not designate genuine classes.

Third Rule: Objectivity

- Which classes exist is not a function of our biological knowledge.

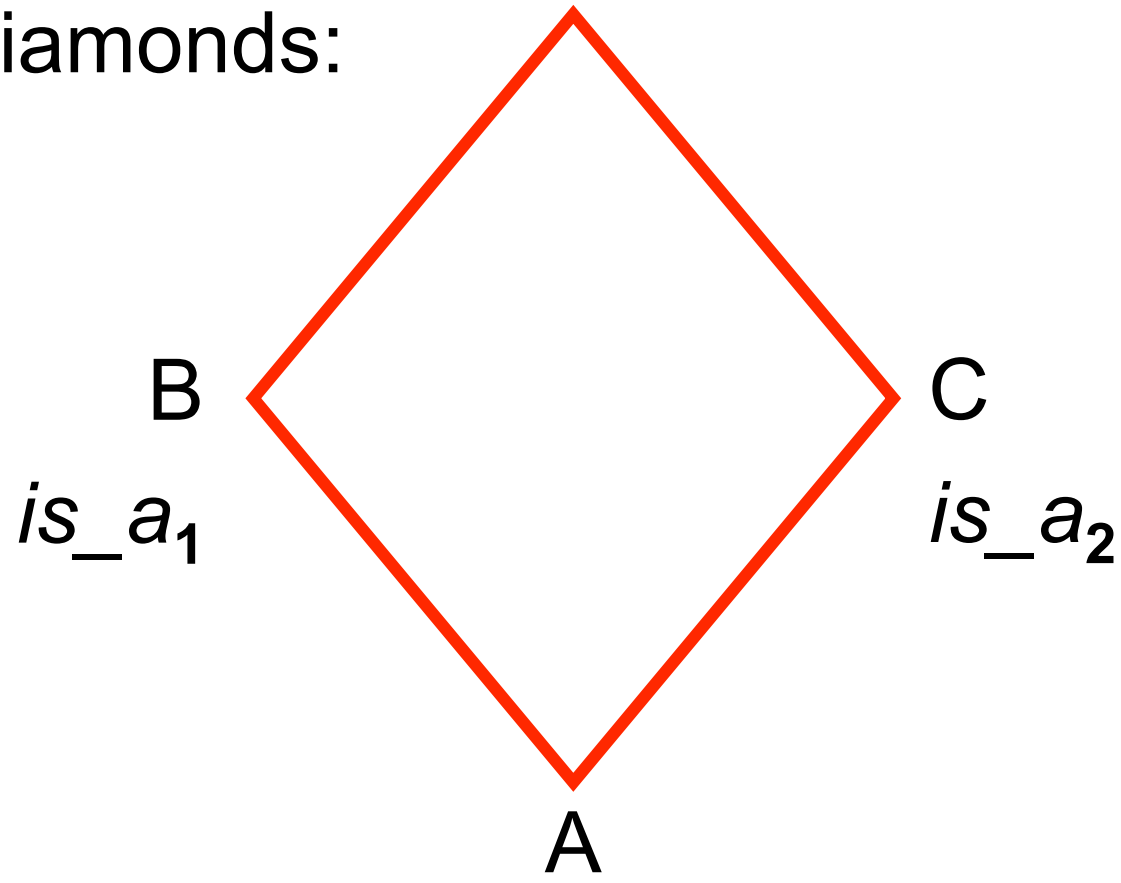
- Terms such as ‘unknown’ or ‘unclassified’ or ‘unlocalized’ do not designate biological natural kinds.

Fourth Rule: Single Inheritance

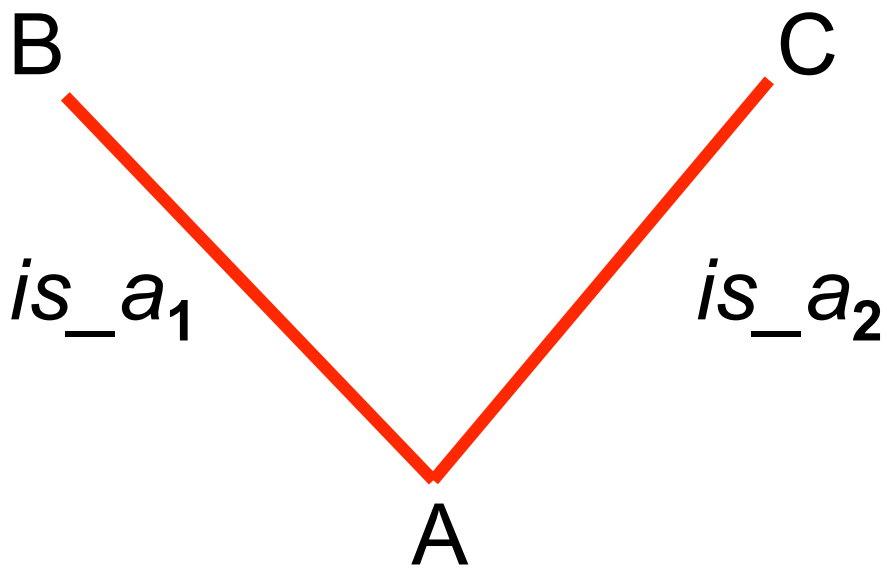
No class in a classificatory hierarchy should have more than one *is_a* parent on the immediate higher level

Rule of Single Inheritance

- no diamonds:



Problems with multiple inheritance



'*is_a*' no longer univocal

'*is_a*' is pressed into service to mean a variety of different things

- shortfalls from single inheritance are often clues to incorrect entry of terms and relations
- the resulting ambiguities make the rules for correct entry difficult to communicate to human curators

is_a Overloading

- serves as obstacle to integration with neighboring ontologies
- The success of ontology alignment depends crucially on the degree to which basic ontological relations such as *is_a* and *part_of* can be relied on as having the same meanings in the different ontologies to be aligned.

Use of multiple inheritance

- The resultant mélange makes coherent integration across ontologies achievable (at best) only under the guidance of human beings with relevant biological knowledge
- How much should reasoning systems be forced to rely on human guidance?

Fifth Rule: Intelligibility of Definitions

- The terms used in a definition should be simpler (more intelligible) than the term to be defined
- otherwise the definition provides no assistance
  - to human understanding
  - for machine processing

To the degree that the above rules are not satisfied, error checking and ontology alignment will be achievable, at best, only with human intervention and via brute force

Some rules are Rules of Thumb

- The world of biomedical research is a world of difficult trade-offs
- The benefits of formal (logical and ontological) rigor need to be balanced
  - Against the constraints of computer tractability,
  - Against the needs of biomedical practitioners.
- BUT alignment and integration of biomedical information resources will be achieved only to the degree that such resources conform to these standard principles of classification and definition

Definitions should be intelligible to both machines and humans

- Machines can cope with the full formal representation
- Humans need to use modularity
- **Plasma membrane**
  - *is a* **cell part** **[immediate parent]**
  - that *surrounds* the **cytoplasm** **[differentia]**

Terms and relations should have clear definitions

- These tell us how the ontology relates to the world of biological instances, meaning the actual particulars in reality:
  - actual cells, actual portions of cytoplasm, and so on…

Sixth Rule: Basis in Reality

- When building or maintaining an ontology, always think carefully at how classes (types, kinds, species) relate to instances in reality

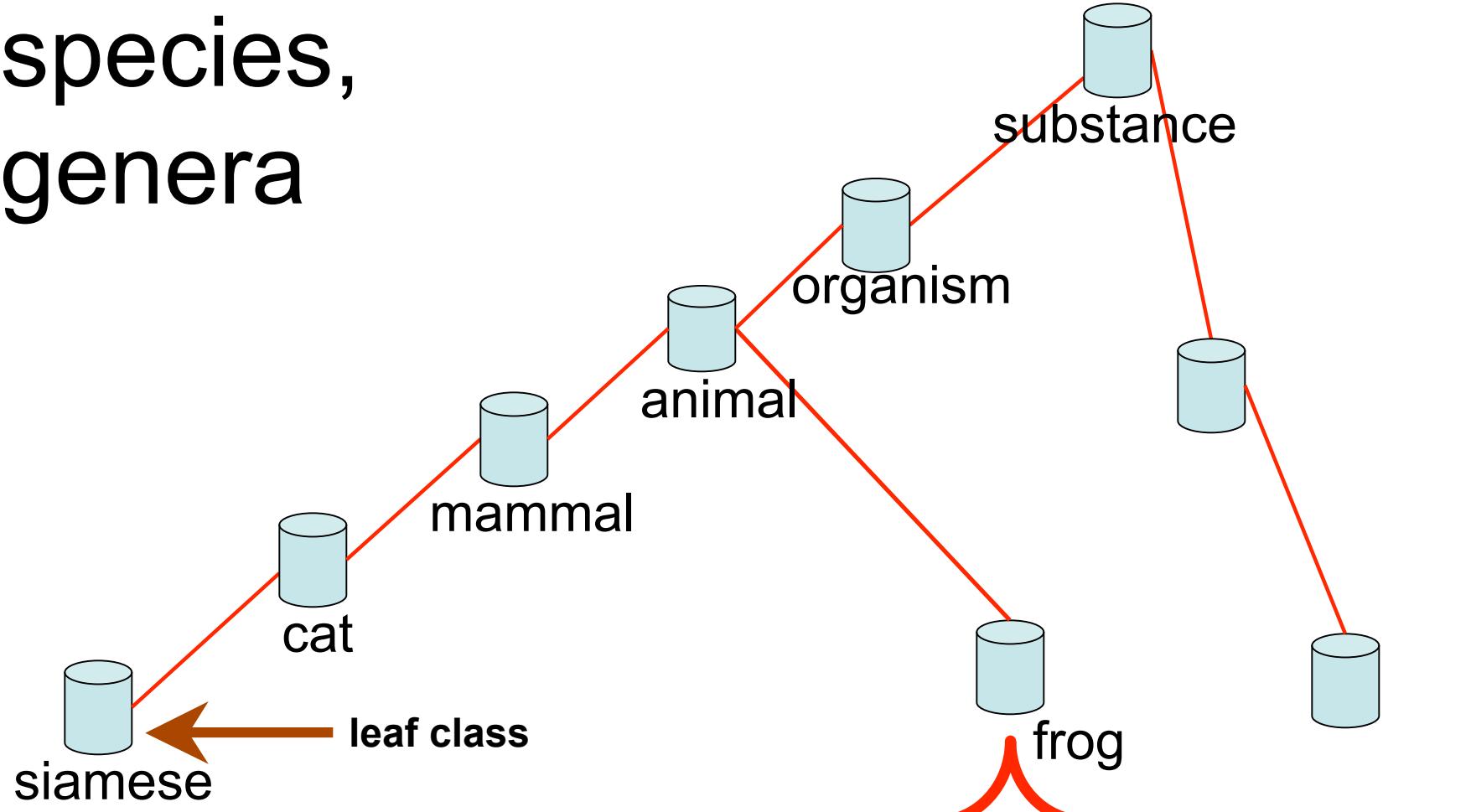
Axioms governing instances

- Every class has at least one instance
- Every genus (parent class) has an instantiated species (differentia + genus)
- Each species (child class) has a smaller class of instances than its genus (parent class)

Axioms governing Instances

- Distinct classes on the same level never share instances
- Distinct leaf classes within a classification never share instances

species, genera



instances

Interoperability

- Ontologies should work together
  - ways should be found to avoid redundancy in ontology building and to support reuse
  - ontologies should be capable of being used by other ontologies (cumulation)

Main obstacle to integration

- Current ontologies do not deal well with
  - Time and
  - Space and
  - Instances (particulars)
- Our definitions should link the terms in the ontology to instances in spatio-temporal reality

Benefits of well-defined relationships

- If the relations in an ontology are well-defined, then reasoning can cascade from one relational assertion ($A\ R_1\ B$) to the next ($B\ R_2\ C$). Relations used in ontologies thus far have not been well defined in this sense.

- *Find all DNA binding proteins* should also find all transcription factor proteins because
  - *Transcription factor is_a DNA binding protein*

How to define *A is_a B*

A is_a B =def.

1. *A* and *B* are names of universals (natural kinds, types) in reality
2. all instances of *A* are as a matter of biological science also instances of B

Biomedical ontology integration / interoperability

- Will never be achieved through integration of meanings or concepts
- The problem is precisely that different user communities use *different concepts*
- ***What's really needed is to have well-defined commonly used relationships***

Idea:

- Move from associative relations between meanings to strictly defined relations between the entities themselves.

- The relations can then be used computationally in the way required

Key idea:
To define ontological relations

- For example: *part_of, develops_from*
- Definitions will enable computation
- It is not enough to look just at classes or types.
  - We need also to take account of *instances* and *time*

Kinds of relations

- Between classes:
  - *is_a*, *part_of*, ...
- Between an instance and a class
  - this explosion **instance_of** the class explosion
- Between instances:
  - Mary's heart **part_of** Mary

Seventh Rule: Distinguish Universals and Instances

- A good ontology must distinguish clearly between
  - **universals (types, kinds, classes)**

  and
  - **instances (tokens, individuals, particulars)**

Don’t forget instances when defining relations

- *part_of* as a relation between classes versus ***part_of*** as a relation between instances
- *nucleus part_of cell*
- your heart ***part_of*** you

Part_of as a relation between classes is more problematic than is standardly supposed

- testis *part_of* human being ?
- heart *part_of* human being ?
- human being *has_part* human testis ?

Why distinguish universals from instances?

- What holds on the level of instances may not hold on the level of universals

- *nucleus adjacent_to cytoplasm*
- **Not:** *cytoplasm adjacent_to nucleus*
- *seminal vesicle adjacent_to urinary bladder*
- **Not:** *urinary bladder adjacent_to seminal vesicle*

part_of

- *part_of* must be time-indexed for spatial universals
- *A part_of B* is defined as:

  Given any instance *a* and any time *t*,

  If *a* is an instance of the universal *A* at *t*,

  then there is some instance *b* of the universal *B*

  such that

  a is an instance-level **part_of** *b* at *t*

Principles for Building Biomedical Ontologies:A GO Perspective

David Hill

Mouse Genome Informatics

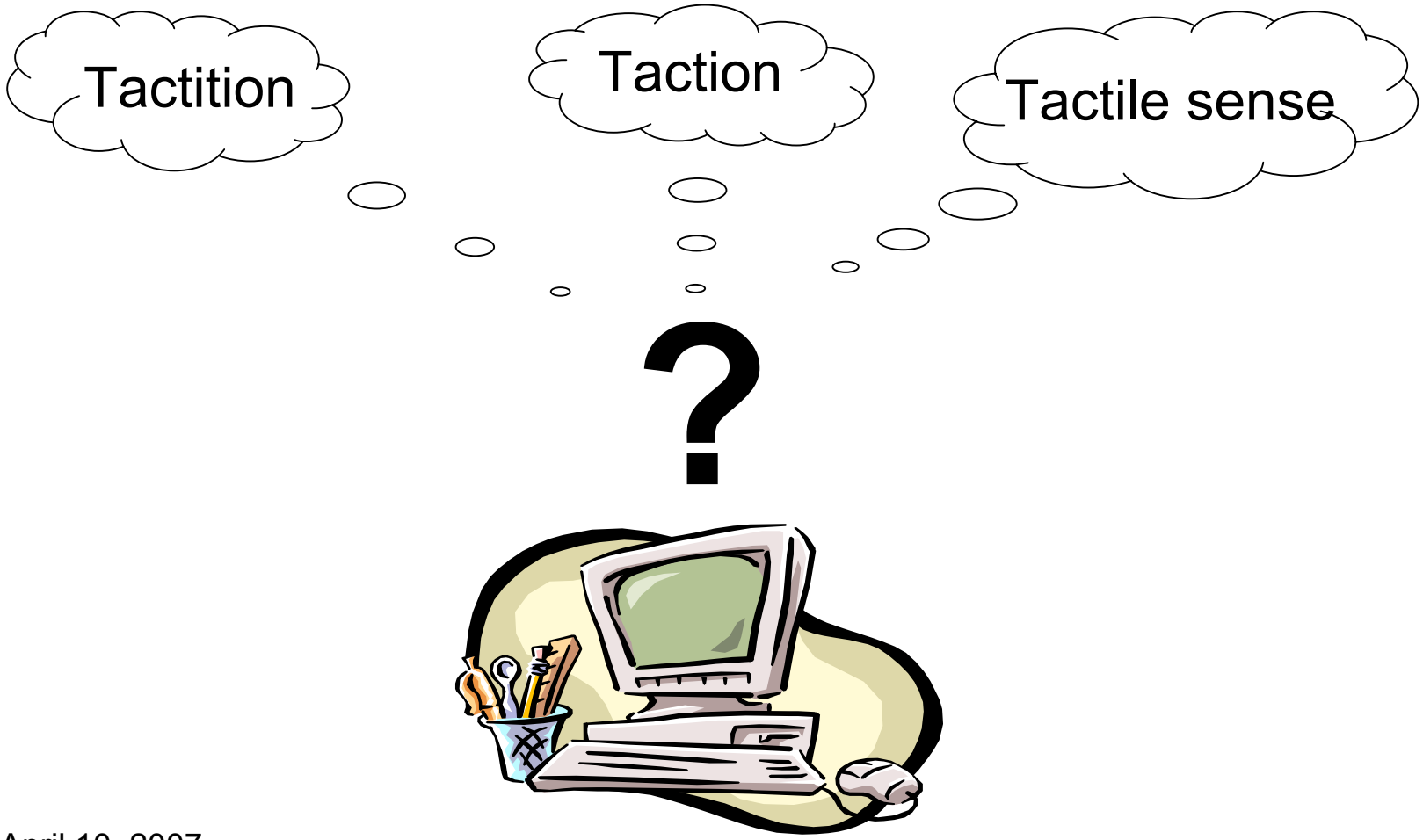
The Jackson Laoratory

April 10, 2007

How has GO dealt with some specific aspects of ontology development?

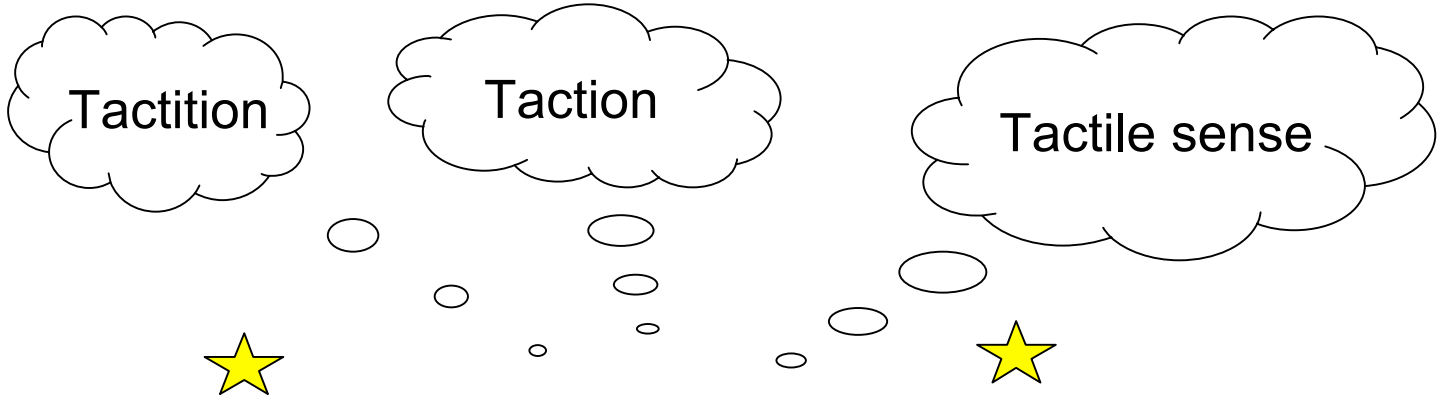
- Univocity
- Positivity
- Objectivity
- Single Inheritance
- Definitions
  - Formal definitions
  - Written definitions
- Basis in Reality
- Universals & Instances
- Ontology Alignment

The Challenge of Univocity:
People call the same thing by different names



April 10, 2007

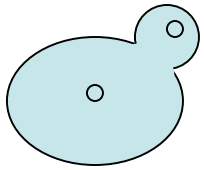
Univocity: GO uses 1 term and many characterized synonyms



perception of touch ; GO:0050975

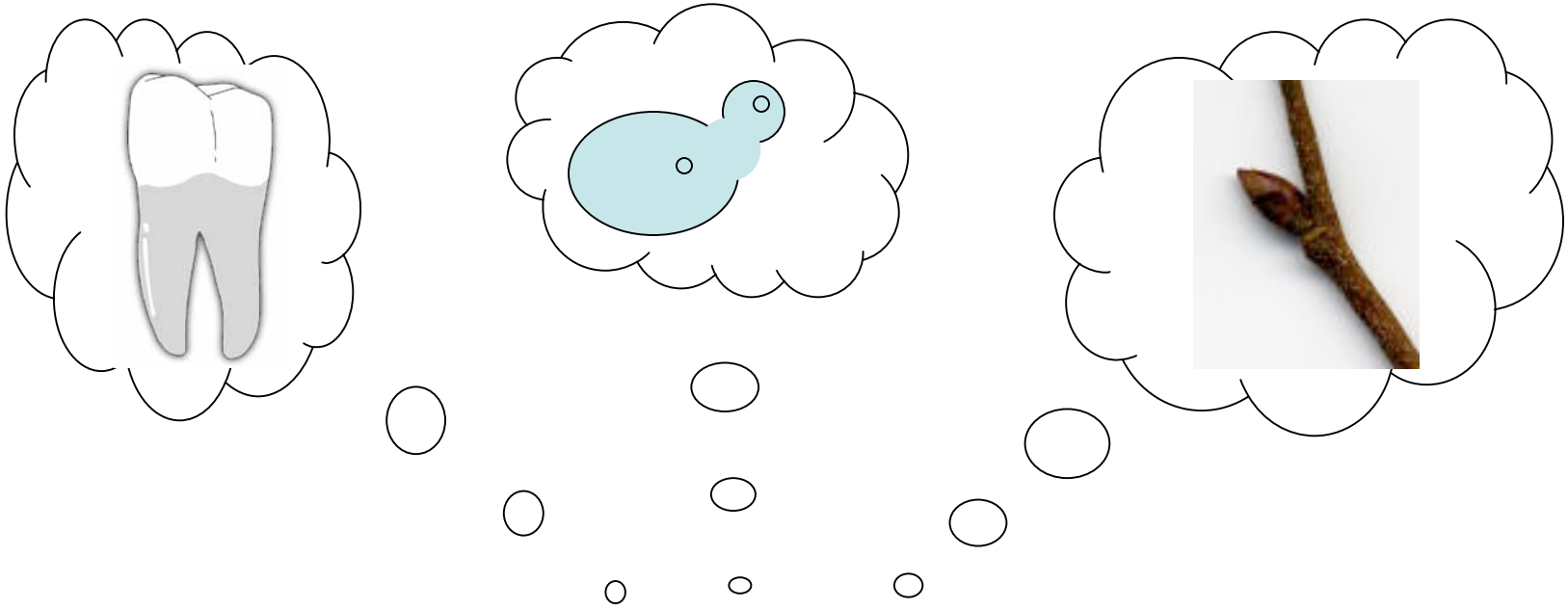
The Challenge of Univocity: People use the same words to describe different things

= bud initiation



= bud initiation

= bud initiation



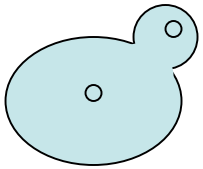
Bud initiation? How is a computer to know?

April 10, 2007

Univocity: GO adds “sensu” descriptors to discriminate among organisms

= bud initiation

sensu *Metazoa*



= bud initiation

sensu *Saccharomyces*

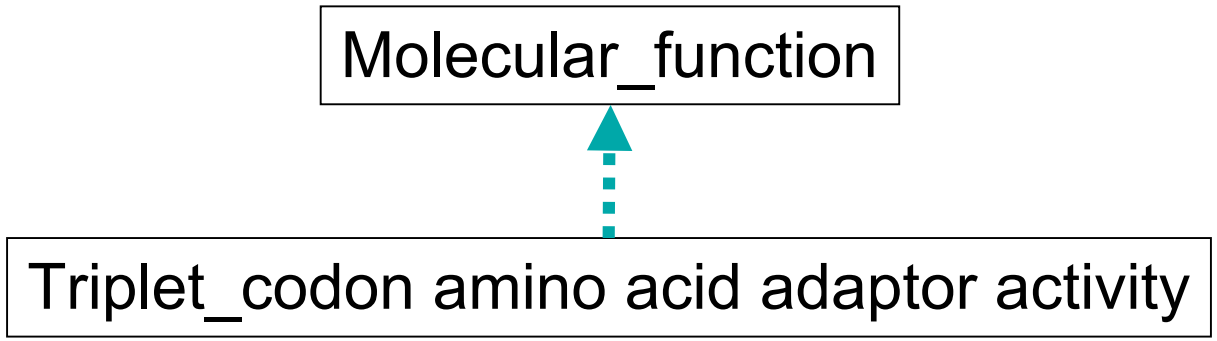
= bud initiation

sensu *Viridiplantae*

The Importance of synonyms for utility: How do we represent the function of tRNA?

Biologically, what does the tRNA do?

Identifies the codon and inserts the amino acid in the growing polypeptide



GO Definition: Mediates the insertion of an amino acid at the correct point in the sequence of a nascent polypeptide chain during protein synthesis.

Synonym: tRNA

But Univocity is also Dependent on a User's Perspective

Development (The biological process whose specific outcome is the progression of an organism over time from an initial condition to a later condition)

--part_of hepatocyte differentiation

----part_of hepatocyte fate commitment

------part_of hepatocyte fate specification

------part_of hepatocyte fate determination

----part_of hepatocyte development

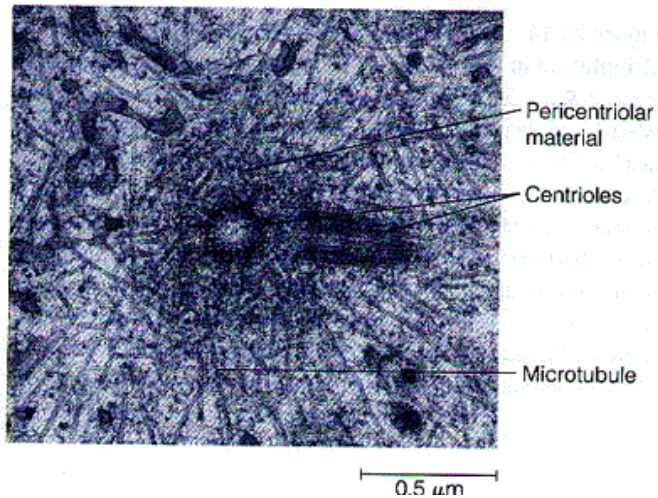
But Univocity is also Dependent on a User's Perspective

So from the perspective of GO a hepatocyte begins development after it is committed to its fate. Its initial condition is after cell fate commitment.

But! A User may ask show me things that have to do with hepatocyte development.

Do they mean show me things that have to do with 'hepatocyte development" or do they mean show me things that have to do with 'development' and a 'hepatocyte'?

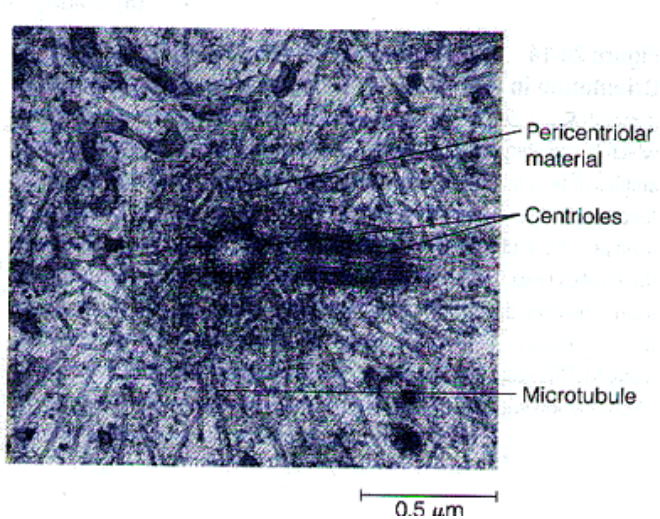
The Challenge of Positivity



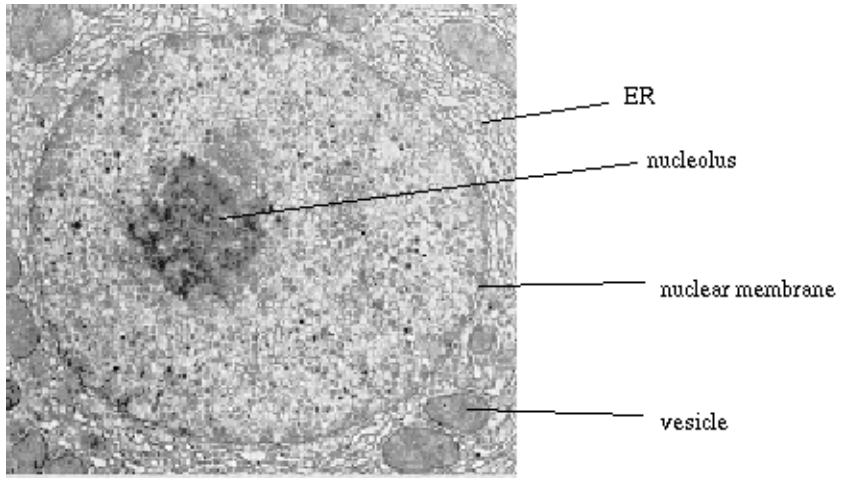
Some organelles are membrane-bound.
A centrosome is not a membrane bound organelle, but it still may be considered an organelle.

April 10, 2007

The Challenge of Positivity: Sometimes absence is a distinction in a Biologist's mind



non-membrane-bound organelle
GO:0043228



membrane-bound organelle
GO:0043227

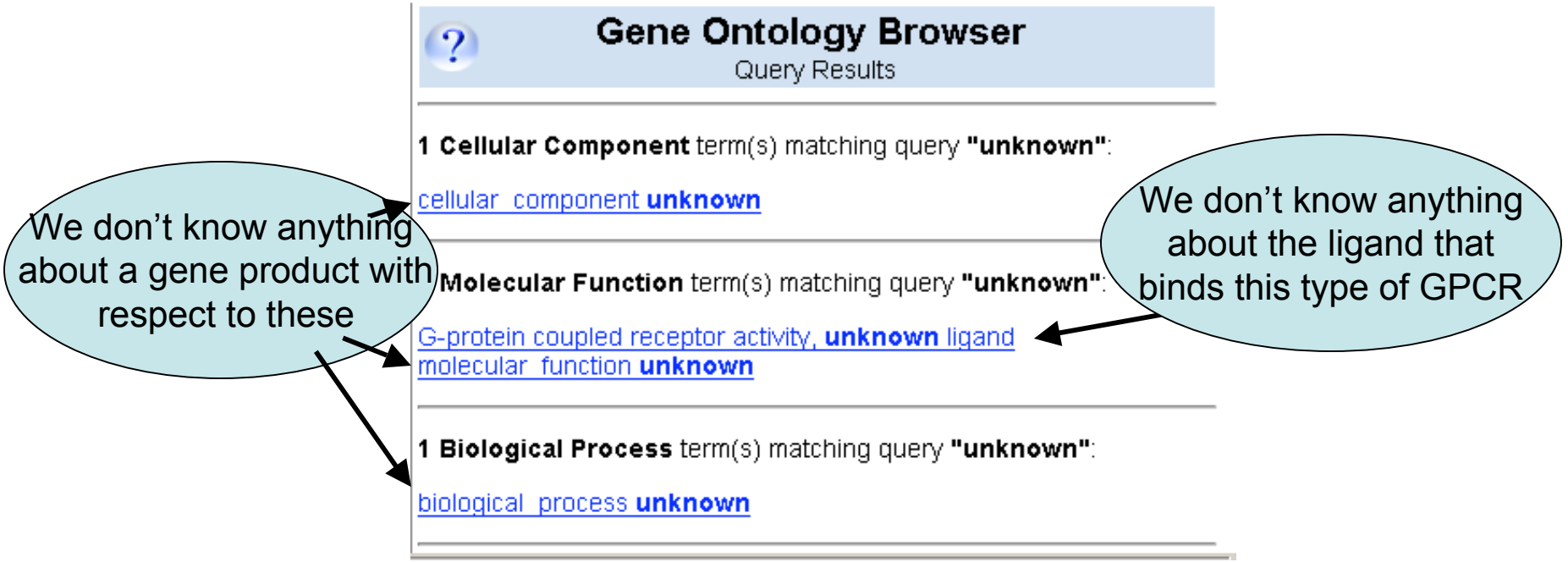
April 10, 2007

Positivity

- Note the logical difference between
  - "*non-membrane-bound organelle*" and
  - "*not a membrane-bound organelle*"

- The latter includes everything that is not a membrane bound organelle!

The Challenge of Objectivity: Database users want to know if we don't know anything (Exhaustiveness with respect to knowledge)



Objectivity

- How can we use GO to annotate gene products when we know that we don't have any information about them?
  - Currently GO has terms in each ontology to describe unknown
  - An alternative might be to annotate genes to root nodes and use an evidence code to describe that we have no data.
- Similar strategies could be used for things like receptors where the ligand is unknown.

GPCRs with unknown ligands

Gene Ontology Browser

Term Detail

GO term:	**class A orphan receptor activity**
GO id:	**GO:0001620**
Definition:	**A G-protein coupled receptor that is structurally and functionally related to the rhodopsin receptor, but whose ligand is unknown.**
Number of paths to term:	**2**

ⓘdenotes an 'is-a' relationship
ⓟdenotes a 'part-of' relationship

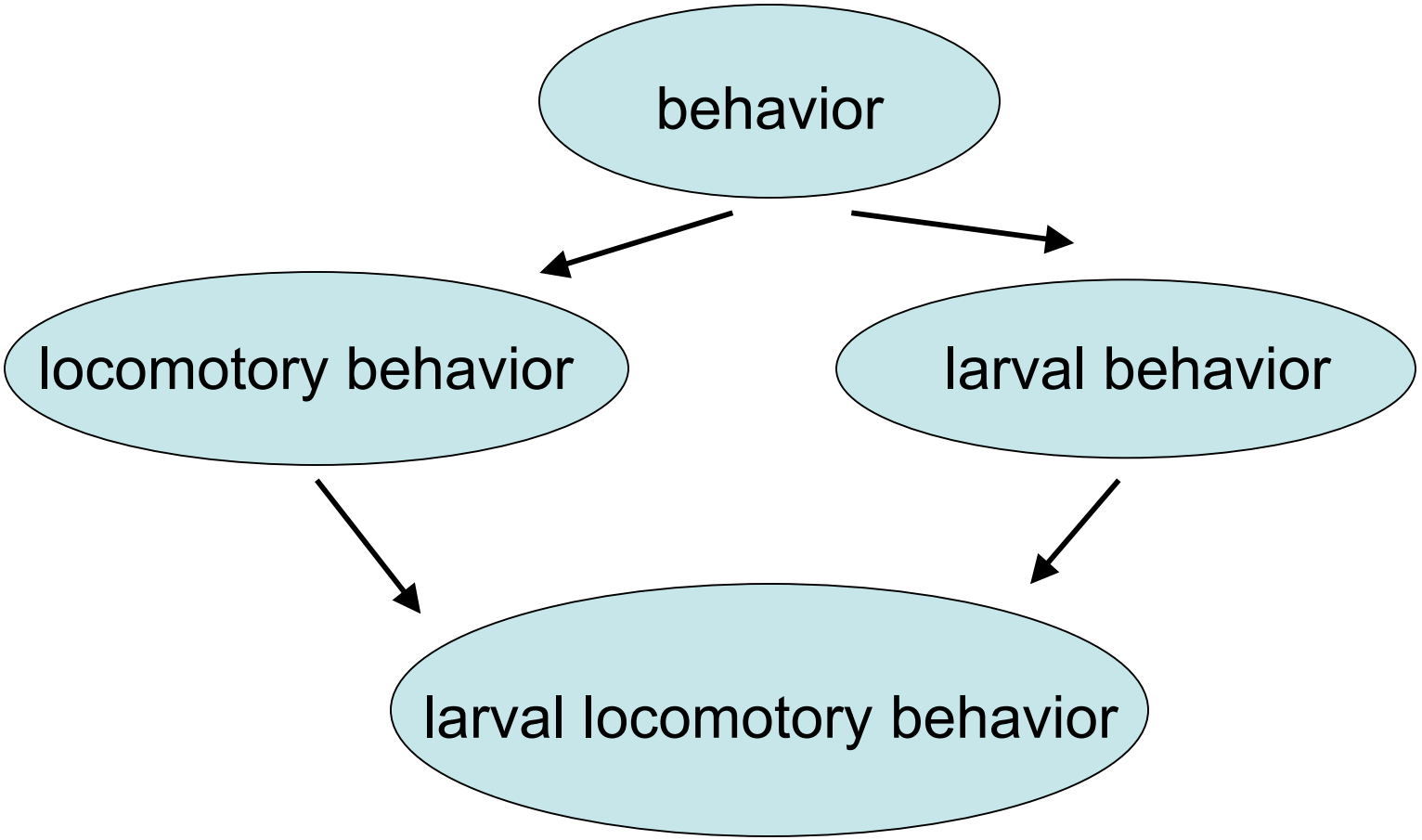
Gene_Ontology
- ⓟmolecular_function
  - ⓘsignal transducer activity
    - ⓘreceptor activity
      - ⓘtransmembrane receptor activity
        - ⓘG-protein coupled receptor activity
          - ⓘG-protein coupled receptor activity, unknown ligand
            - ⓘclass A orphan receptor activity [GO:0001620] *(0 genes, 0 annotations)*
              - ⓘEpstein-Barr Virus-induced receptor activity
              - ⓘG-protein receptor 45-like receptor activity
              - ⓘgastropyloric receptor activity
              - ⓘGP40-like receptor activity
              - ⓘMas proto-oncogene receptor activity
              - ⓘRDC1 receptor activity
              - ⓘsuper conserved receptor expressed in brain receptor activity
            - ⓘclass B orphan receptor activity
            - ⓘclass C orphan receptor activity

We could annotate to this

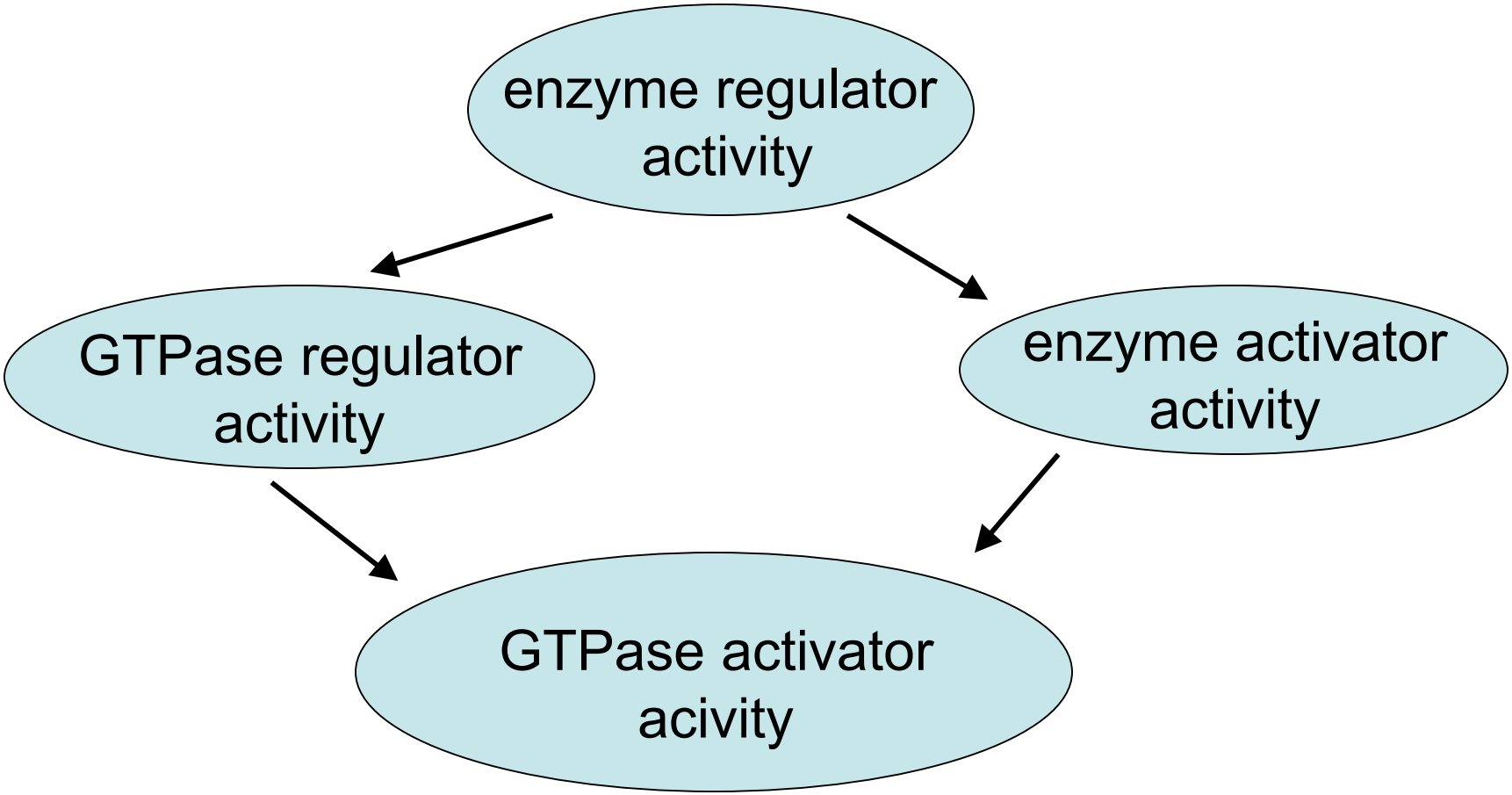
Single Inheritance

- GO has a lot of is_a diamonds
  - Some are due to incompleteness/inaccuracies within the graph
  - Some are due to a mixture of dissimilar entities within the graph at the same level

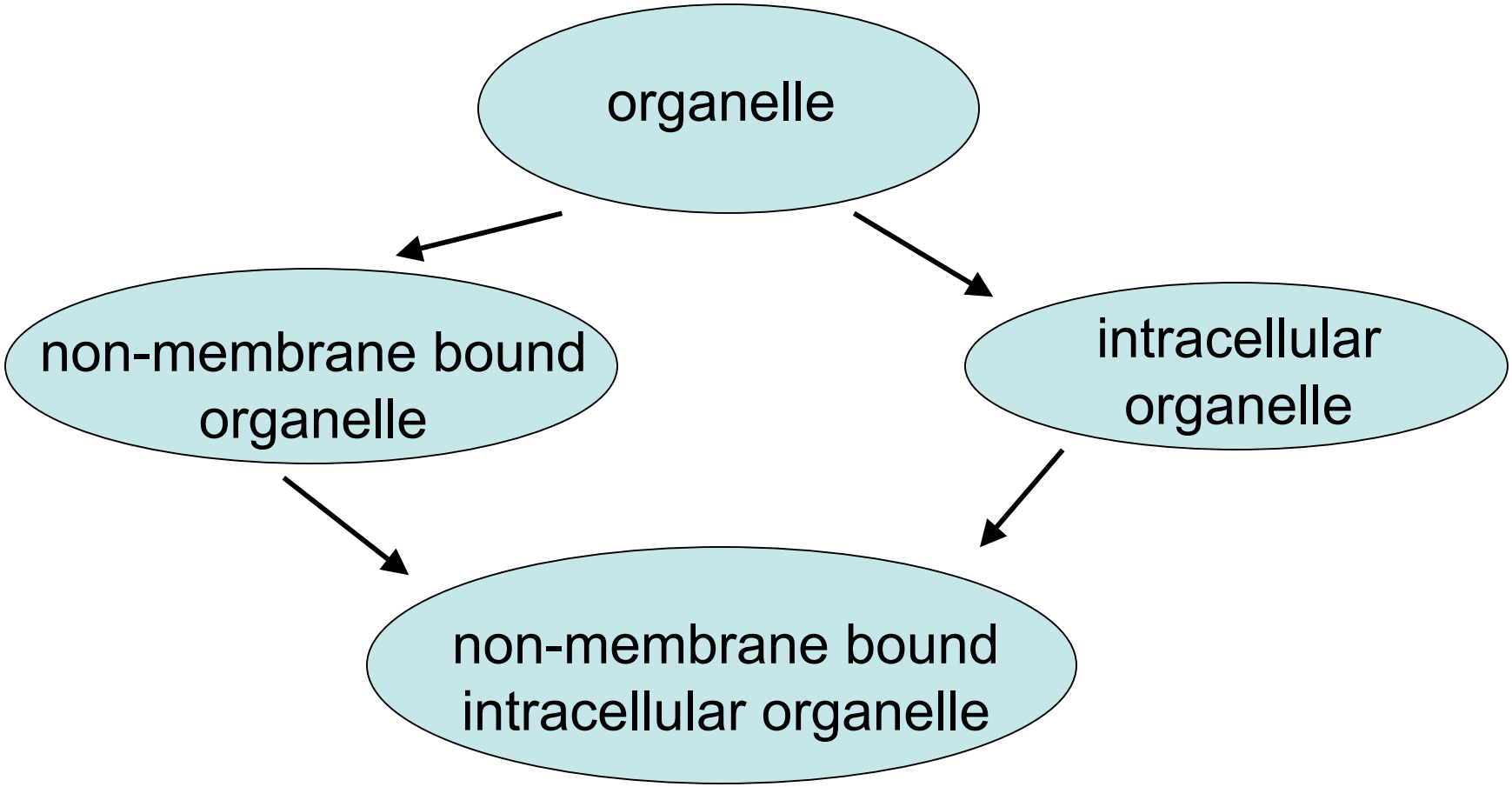
Is_a diamond in GO Process



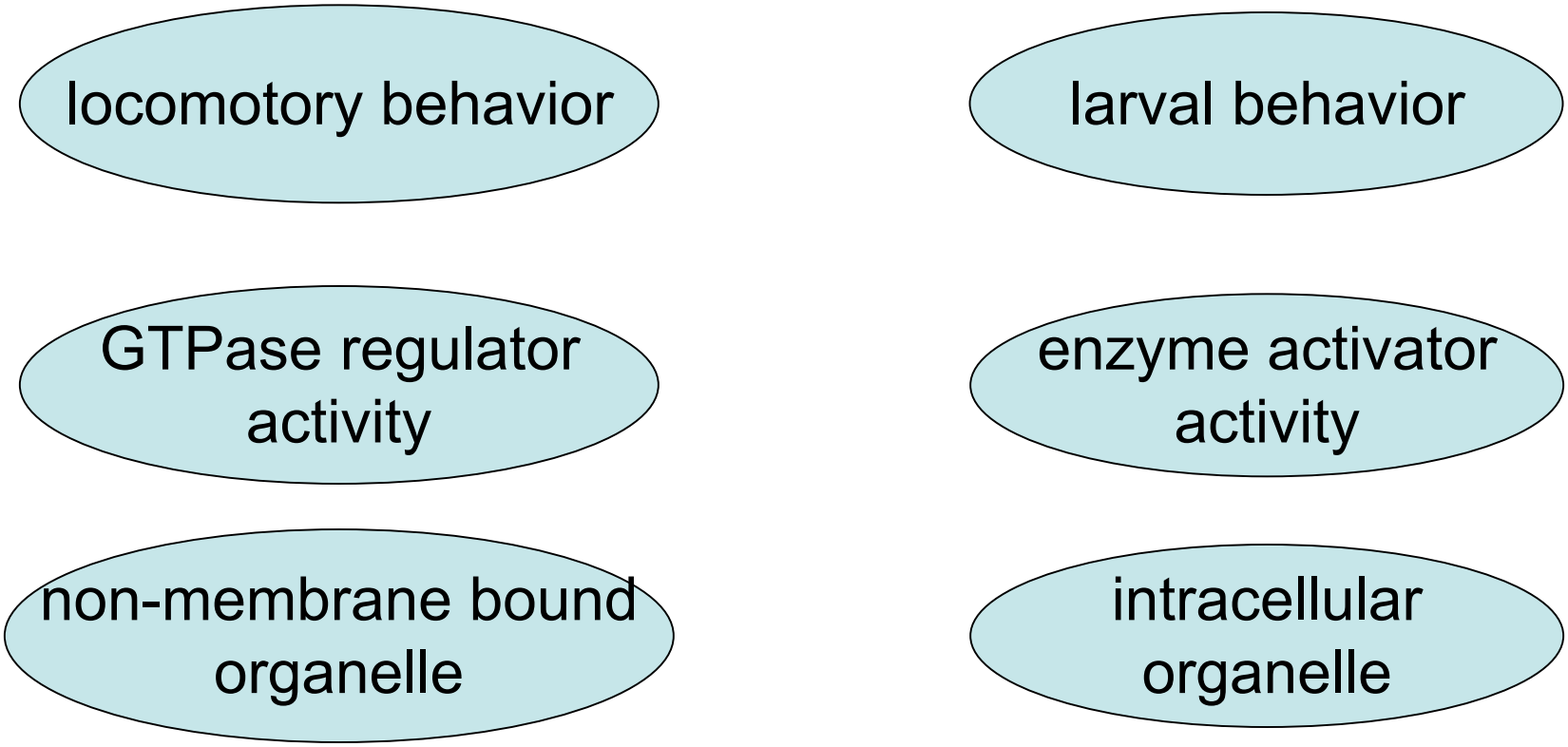
Is_a diamond in GO Function



Is_a diamond in GO Cellular Component

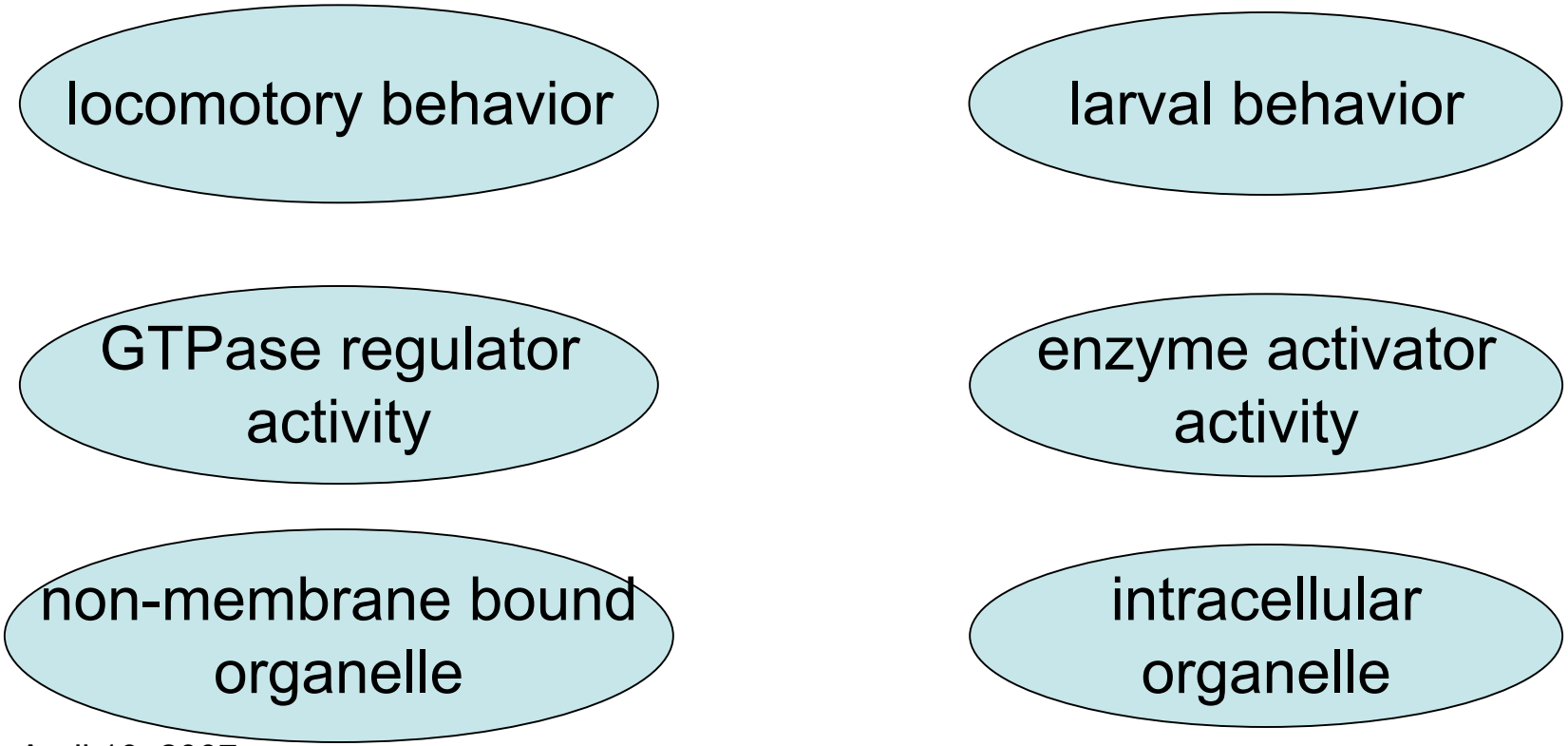


Technically the diamonds are correct, but could be eliminated



What do these pairs have in common?

What do the middle pair of terms all have in common?



They are all differentiated from the parent term by a different factor

locomotory behavior

larval behavior

Type of behavior vs. what is behaving

GTPase regulator activity

enzyme activator activity

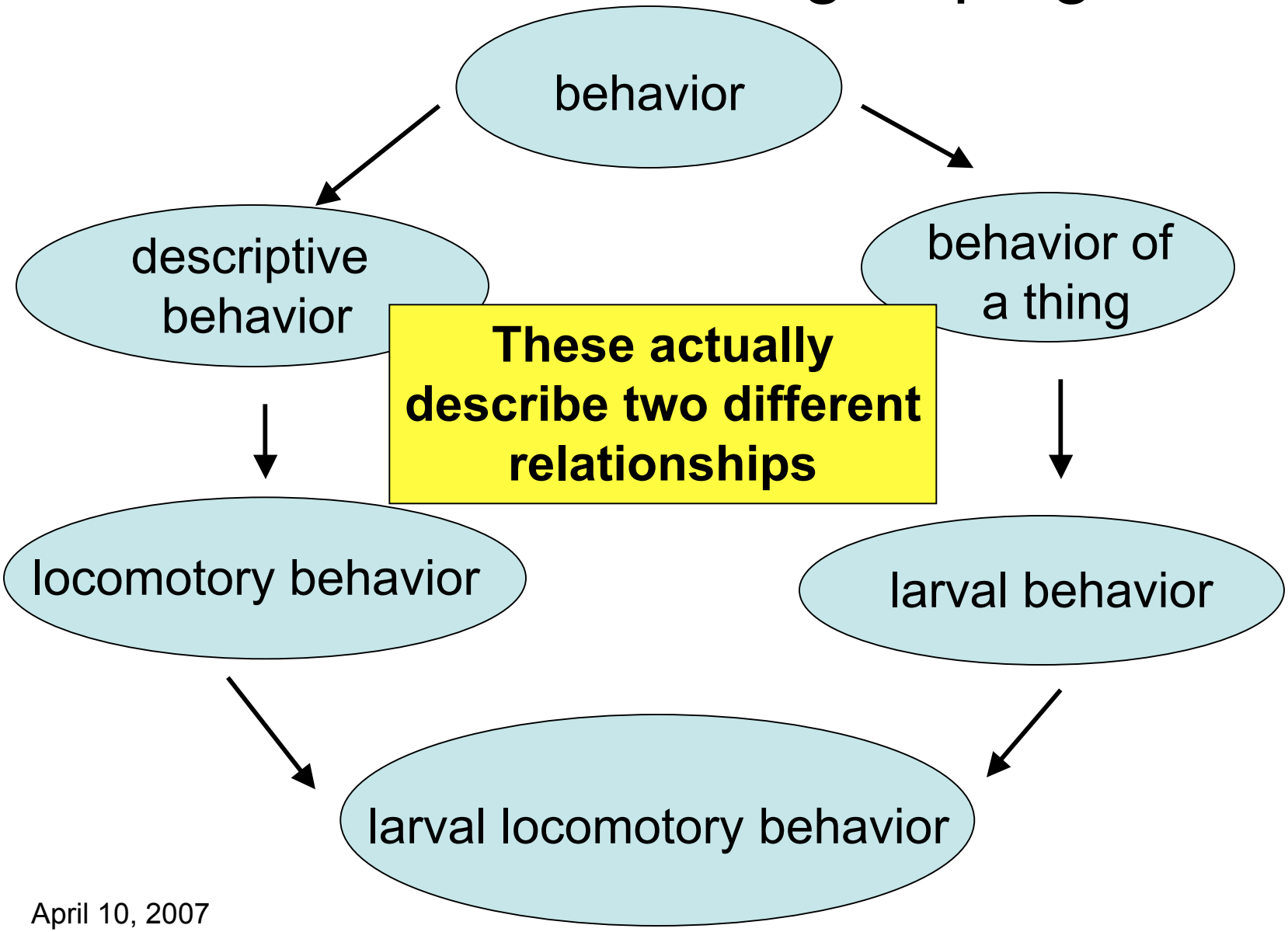
What is regulated vs. type of regulator

non-membrane bound organelle

intracellular organelle

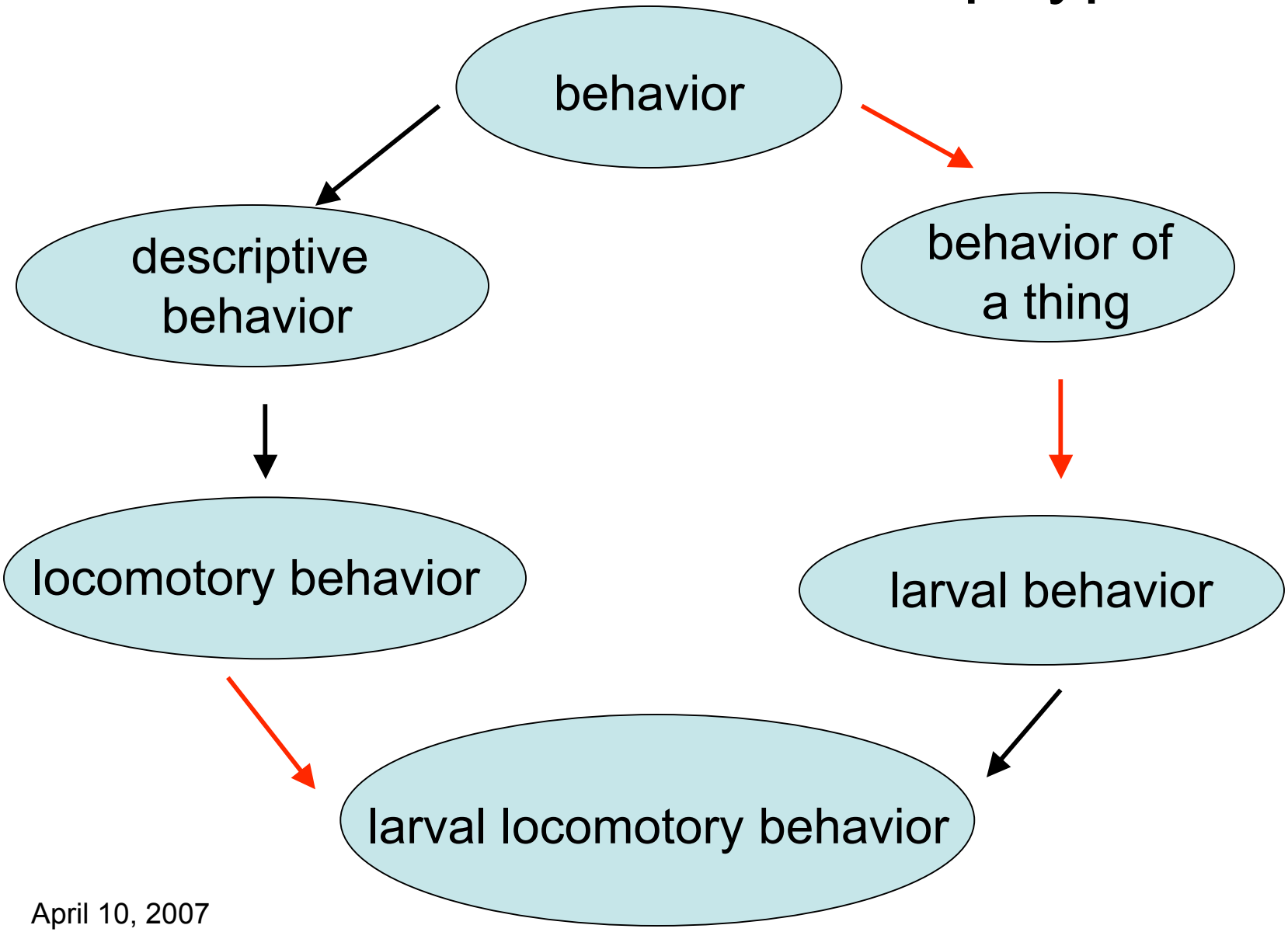
Type of organelle vs. location of organelle

Insert an intermediate grouping term

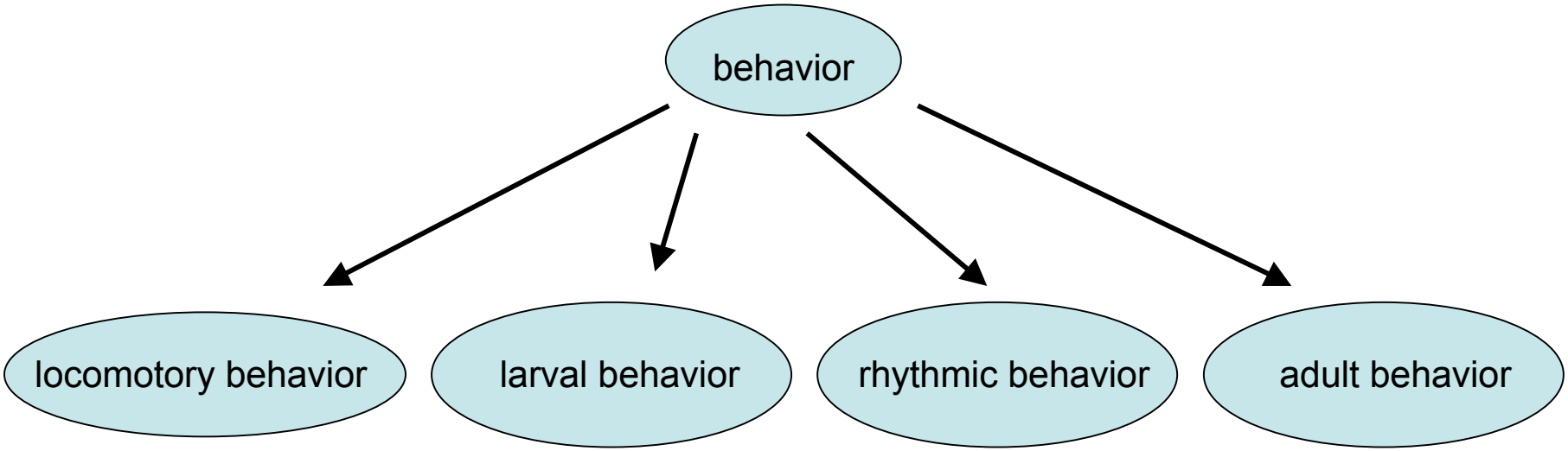


April 10, 2007

Create a new relationship type



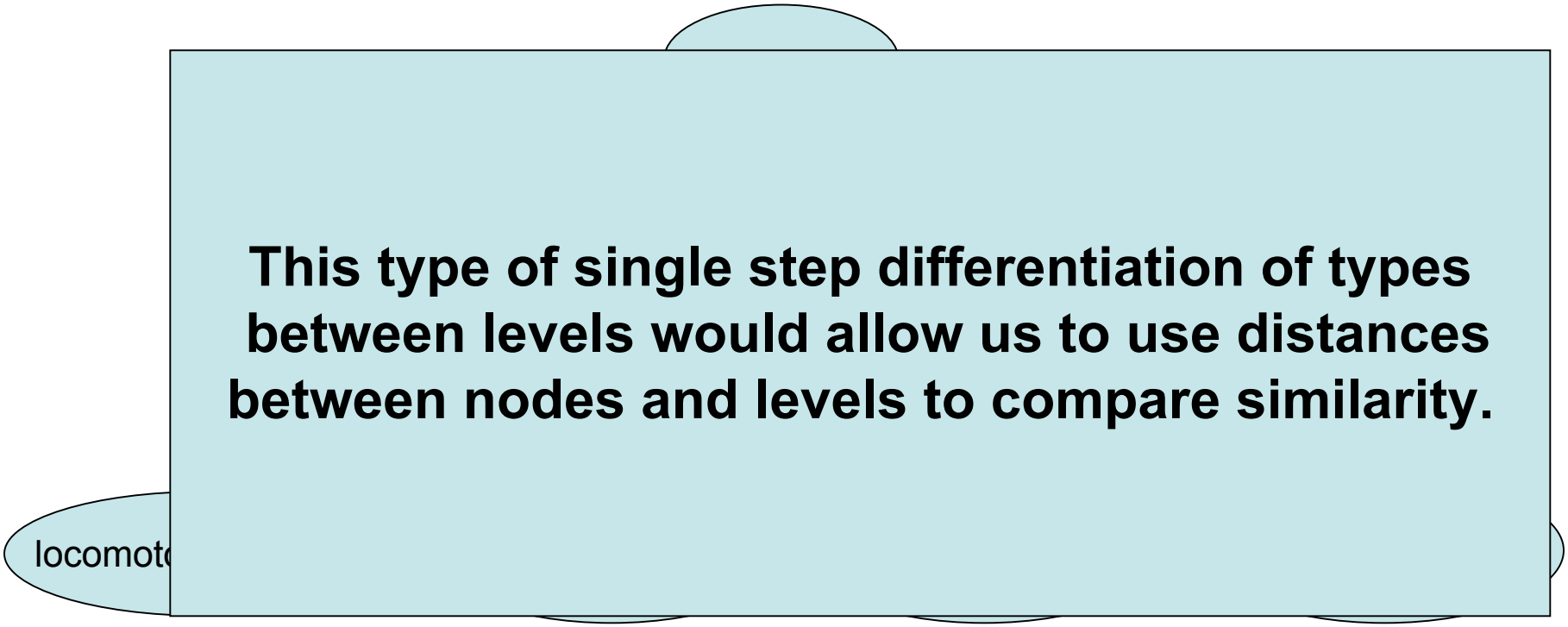
Why insert terms that no one would use?



By the structure of this graph, locomotory behavior has the same relationship to larval behavior as to rhythmic behavior

April 10, 2007

Why insert terms (types) that no one would use?



But actually, locomotory behavior/rhythmic behavior and larval behavior/adult behavior group naturally

GO Definitions

Gene Ontology Browser

Term Detail

GO term:	**cell differentiation**
GO id:	**GO:0030154**
Definition:	**The process whereby relatively unspecialized cells, e.g. embryonic or regenerative cells, acquire specialized structural and/or functional features that characterize the cells, tissues, or organs of the mature organism or some other relatively stable phase of the organism's life history.**

A definition written by a biologist:
necessary & sufficient conditions
written definition
(not computable)

Gene_Ontology
- Ⓟbiological_process
  - ①cellular process
    - ①cell communication +
    - ①cell differentiation [GO:0030154] *(493 genes, 649 annotations)*
      - ①adipocyte differentiation +
      - ①antipodal cell differentiation +
      - ①cardiac cell differentiation +

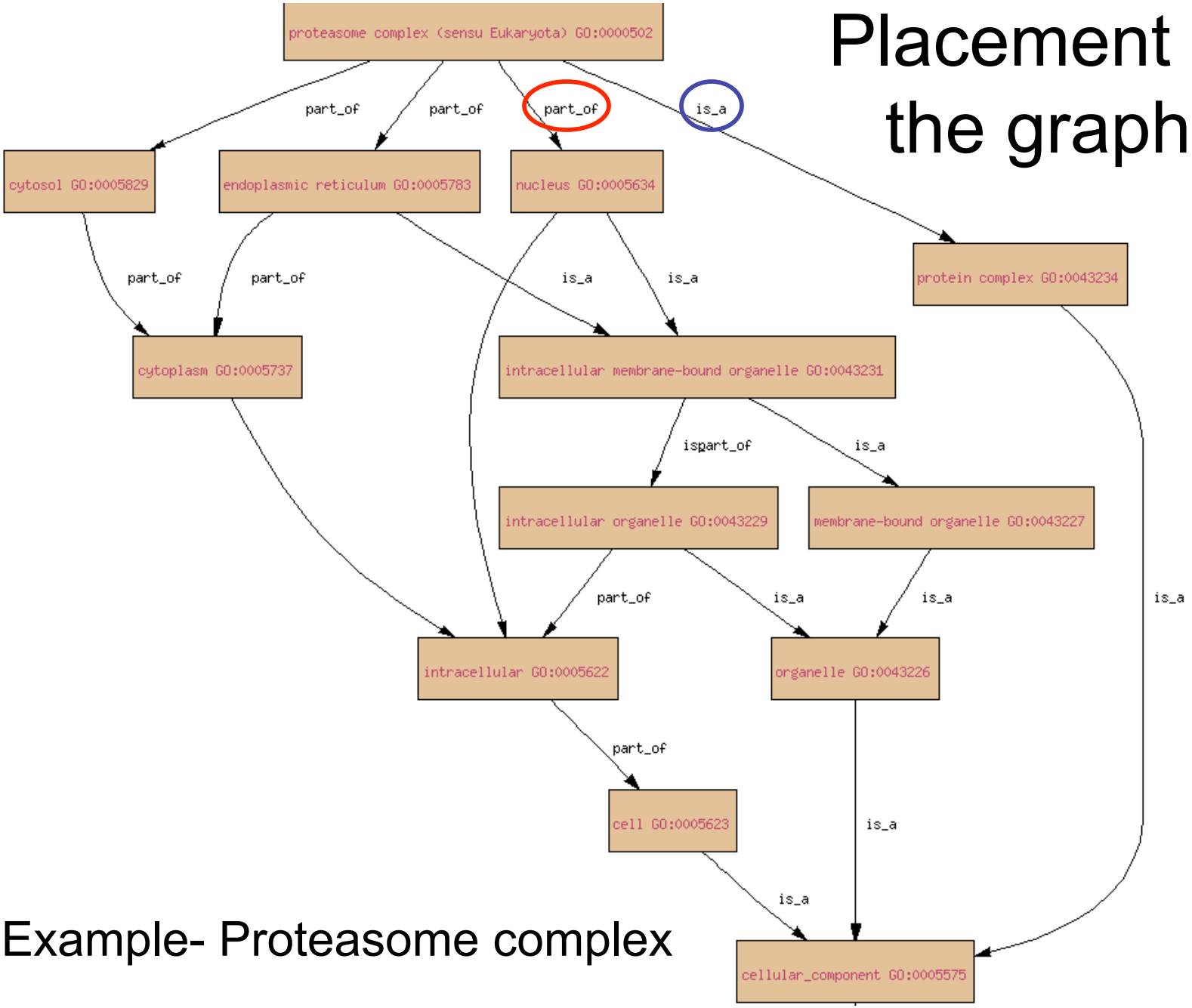
Gene_Ontology
- Ⓟbiological_process
  - ①development
    - ①abscission +
    - ①aging +
    - ①blastocyst development +
    - ①blastocyst hatching
    - ①cell development +
    - Ⓟcell differentiation [GO:0030154] *(493 genes, 649 annotations)*
      - ①adipocyte differentiation +

Graph structure:
necessary conditions
formal
(computable)

Relationships and definitions

- The set of *necessary conditions* is determined by the graph
  - This can be considered a *partial* definition
- Important considerations:
  - Placement in the graph- selecting parents
  - Appropriate relationships to different parents
  - True path violation

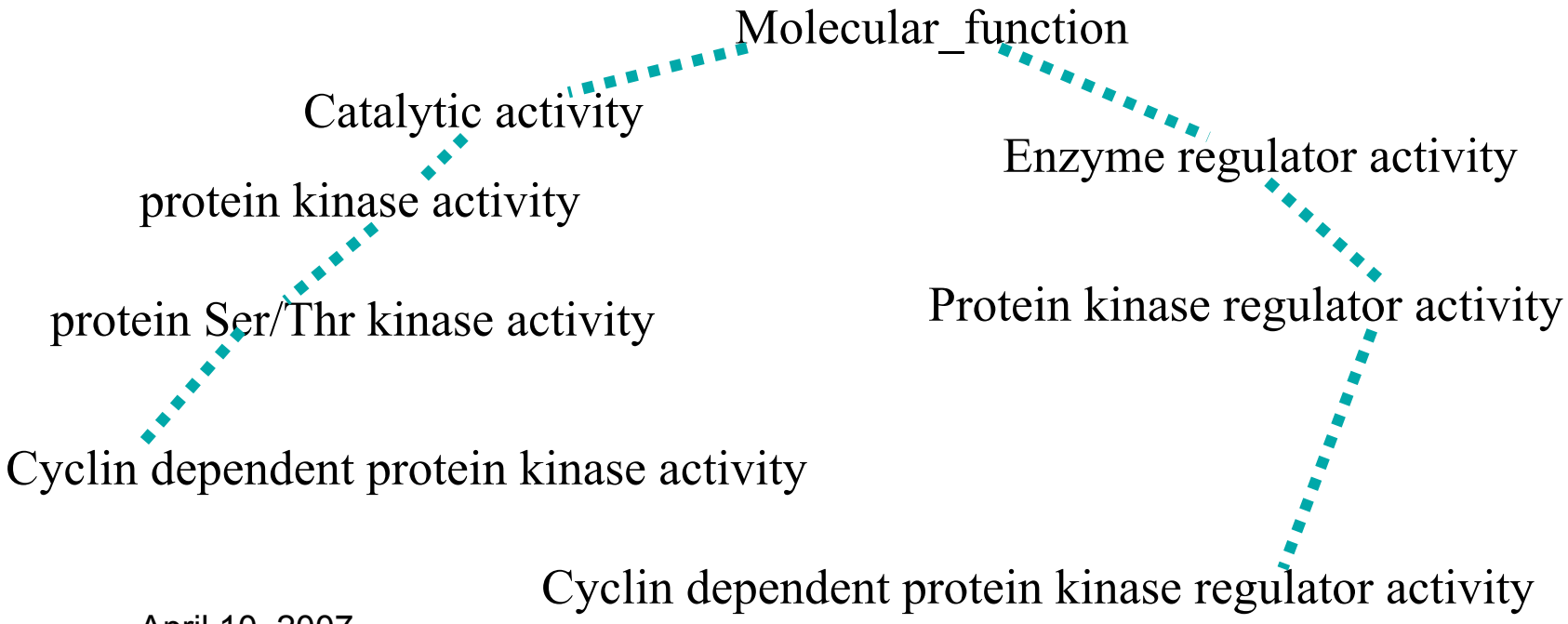
Placement in the graph



- Example- Proteasome complex

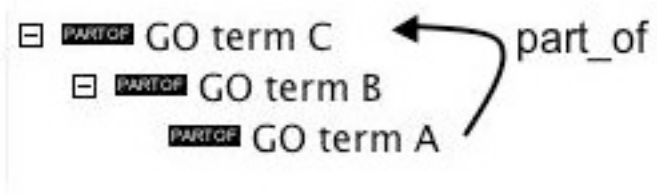
The importance of relationships

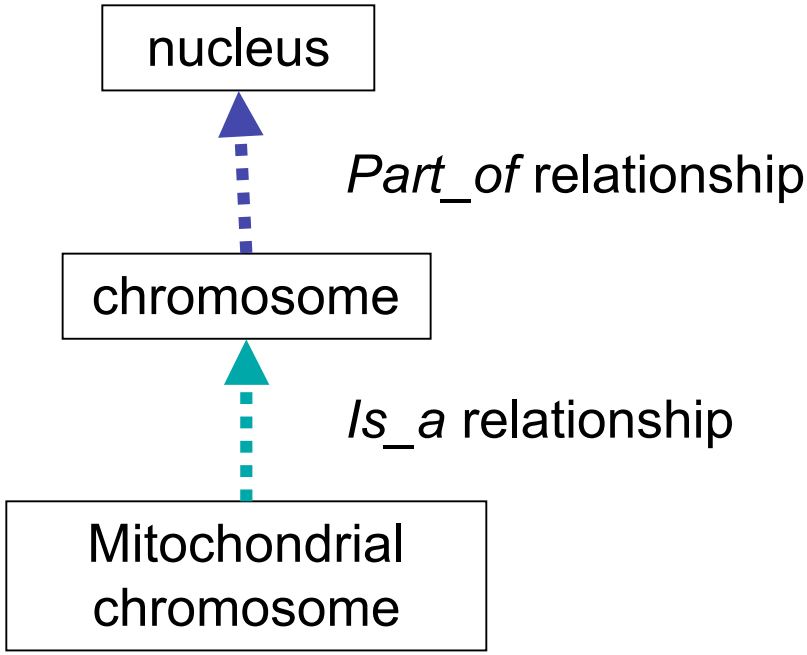
- Cyclin dependent protein kinase
  - Complex has a catalytic and a regulatory subunit
  - How do we represent these activities (function) in the ontology?
  - Do we need a new relationship type (regulates)?



We must avoid true path violations

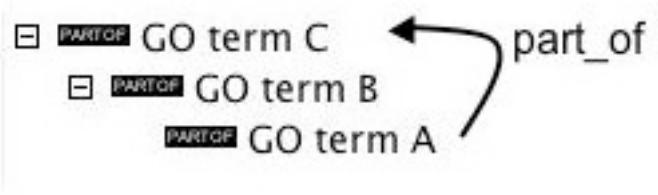
.."the pathway from a child term all the way up to its top-level parent(s) must always be true".

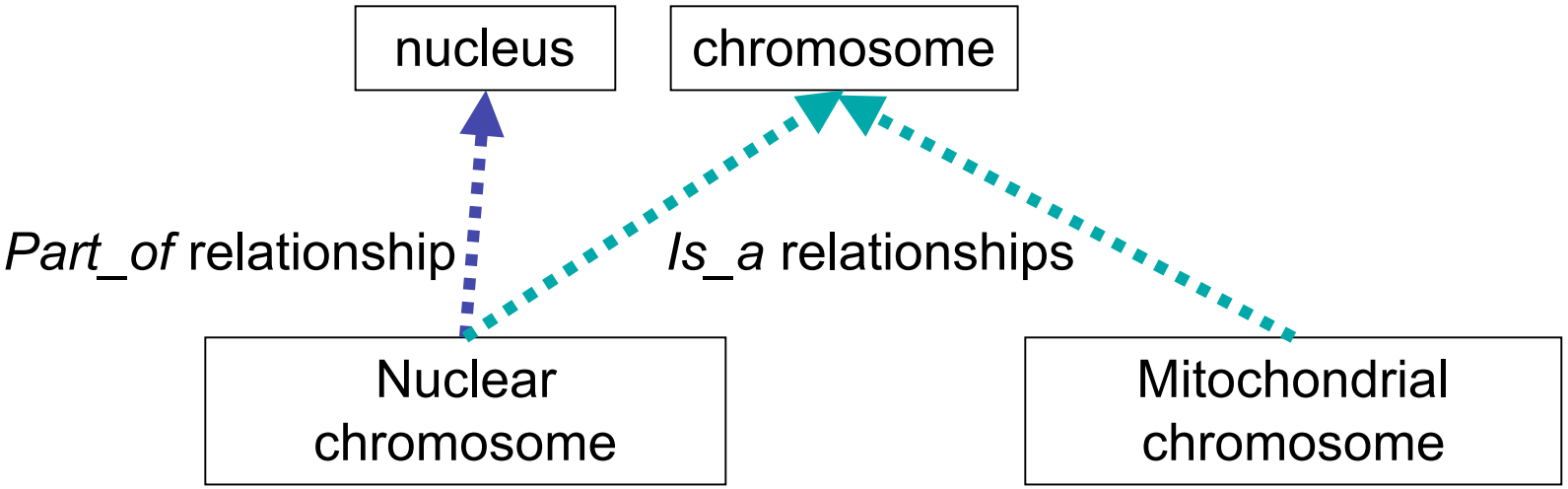




We must avoid true path violations

..”the pathway from a child term all the way up to its top-level parent(s) must always be true".





GO textual definitions: Related GO terms have similarly structured (normalized) definitions

GO term: **neuron cell differentiation**
GO id: **GO:0030182**
Definition: **Processes whereby a relatively unspecialized cell acquires specialized features of a neuron.**

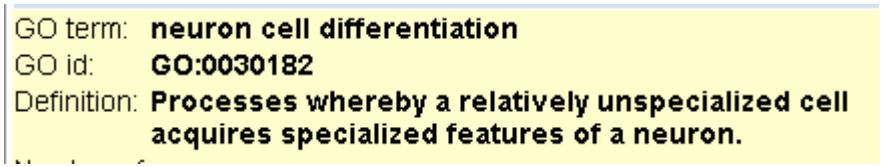
GO term: **heterocyst cell differentiation**
GO id: **GO:0043158**
Definition: **Processes whereby a relatively unspecialized cell acquires specialized features of a heterocyst, a differentiated cell in certain cyanobacteria whose purpose is to fix nitrogen.**

GO term: **cardiac cell differentiation**
GO id: **GO:0035051**
Definition: **The processes whereby a relatively unspecialized cell acquires the specialized structural and/or functional features of a cell that will form part of the cardiac organ of an individual.**

GO term: **muscle cell differentiation**
GO id: **GO:0042692**
Definition: **The process whereby a relatively unspecialized cell acquires specialized features of a muscle cell.**

GO term: **glial cell differentiation**
Synonym: **glia cell differentiation**
GO id: **GO:0010001**
Definition: **Processes whereby a relatively unspecialized cell acquires the specialized features of a glial cell.**

Structured definitions contain both **genus** and **differentiae**


GO term: **neuron cell differentiation**
GO id: **GO:0030182**
Definition: **Processes whereby a relatively unspecialized cell acquires specialized features of a neuron.**

Essence = Genus + Differentiae

neuron cell differentiation =
Genus: **differentiation** (processes whereby a relatively unspecialized cell acquires the specialized features of..)
Differentiae: *acquires features of* a **neuron**

Basis in Reality

But, since GO is representing a science, GO actually represents paradigms.
Therefore, it is essential that GO is able to change!

content

- Annotators are experts in their fields
- Annotators constantly read the scientific literature

Types and Instances

For the sake of GO, types are the terms and instances are the gene product attributes that are annotated to them.

Types and Instances

- When should we create a new type as opposed to multiple annotations?
- When the the biology represents a universal principal. Receptor signaling protein tyrosine kinase activity does not represent receptor signaling protein activity and tyrosine kinase activity independently.

Ontology alignment

One of the current goals of GO is to align:

Cell Types in GO with **Cell Types in the Cell Ontology**

cone cell fate commitment	⟷	retinal_cone_cell
keratinocyte differentiation	⟷	keratinocyte
adipocyte differentiation	⟷	fat_cell
dendritic cell activation	⟷	dendritic_cell
lymphocyte proliferation	⟷	lymphocyte
T-cell homeostasis	⟷	T_lymphocyte
garland cell differentiation	⟷	garland_cell
heterocyst cell differentiation	⟷	heterocyst

Alignment of the Two Ontologies will permit the generation of consistent and complete definitions

GO term:	**osteoblast differentiation**
Synonym:	**osteoblast cell differentiation**
GO id:	**GO:0001649**
Definition:	**Processes whereby a relatively unspecialized cell acquires the specialized features of an osteoblast, the mesodermal cell that gives rise to bone.**

GO

+

id: CL:0000062
name: osteoblast
def: "A bone-forming cell which secretes an extracellular matrix. Hydroxyapatite crystals are then deposited into the matrix to form bone." [MESH:A.11.329.629]
is_a: CL:0000055
relationship: develops_from CL:0000008
relationship: develops_from CL:0000375

Cell type

=

Osteoblast differentiation: Processes whereby an osteoprogenitor cell or a cranial neural crest cell acquires the specialized features of an osteoblast, a bone-forming cell which secretes extracellular matrix.

New Definition

Alignment of the Two Ontologies will permit the generation of consistent and complete definitions

id: GO:0001649

name: osteoblast differentiation

synonym: osteoblast cell differentiation
genus: differentiation GO:0030154 (differentiation)
differentium: *acquires_features_of* CL:0000062 (osteoblast)
definition (text): Processes whereby a relatively unspecialized cell acquires the specialized features of an osteoblast, the mesodermal cell that gives rise to bone

Formal definitions with necessary and sufficient conditions, in both human readable and computer readable forms

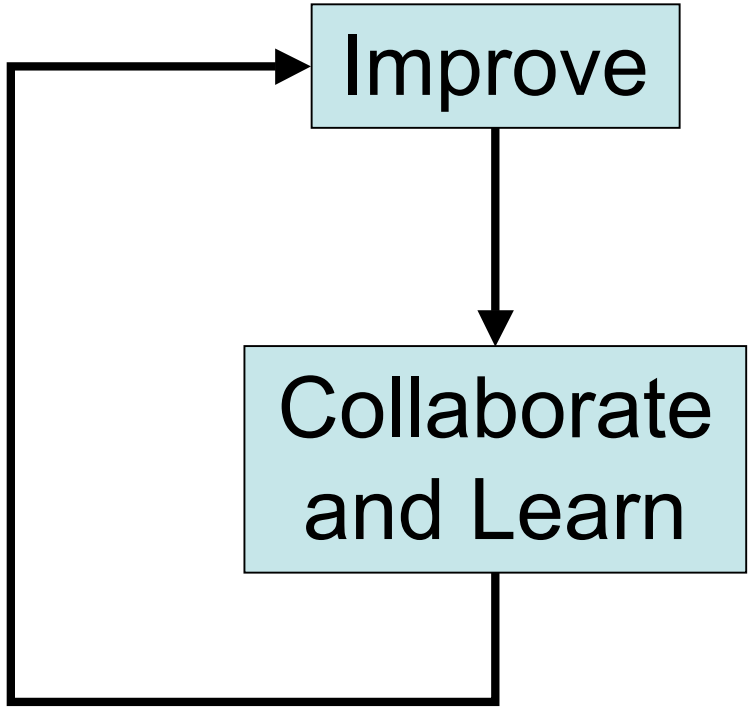
Other Ontologies that can be aligned with GO

- Chemical ontologies
  - 3,4-dihydroxy-2-butanone-4-phosphate synthase activity
- Anatomy ontologies
  - metanephros development
- GO itself
  - mitochondrial inner membrane peptidase activity

But Eventually…

Molecular function	GO	gene ontology.obo	yes
Biological process	GO	gene ontology.obo	yes
Cellular component	GO	gene ontology.obo	yes
Human developmental anatomy, timed version	EHDA	human dev anat staged.ontology	yes
Human developmental anatomy, abstract version	EHDAA	human dev anat abstract.ontology	yes
Human disease	DOID	DO 08 18 03.txt	no
Biological imaging methods	FBbi	image.ontology	no
Protein domain	IPR	entry.list	yes
Multiple alignment	RO	mao.obo	no
Medaka fish anatomy and development	MFO	medaka anatomy.ontology and medaka anatomy.definitions	yes
MESH	MESH	MESH to GO and MESH definitions	no
Mus gross anatomy and development	EMAP	EMAP.ontology	yes
Mus adult gross anatomy	MA	MA.ontology	yes
Mouse pathology	MPATH	mouse pathology.ontology	yes
Mammalian phenotype	MP	MPheno.ontology and MP.defs	no
NCI Thesaurus	NCIt	EVS ftp site	no
SwissProt organismal classification	[none]	[none]	yes
OBO relationship types	OBO_REL	relationship.obo	yes
Context	PM	context.ontology and context.definition	no
Plant anatomy	PO	anatomy.ontology and anatomy.definition	yes
Plant environmental conditions	EO	environment ontology.obo	no
Plasmodium development	PLO	PLO ontology.txt and PLO defs.shtml	yes
PATO	PATO	attribute and value.obo	yes
Physico-chemical process	REX	rex.obo	no
Sequence types and features	SO	so.ontology and so.definition	yes
NCBI organismal classification	taxon	taxonomy.dat	no
Caenorhabditis gross anatomy	[none]	[none]	no
C. elegans development	WBls	worm development.ontology and worm development.definitions	yes
Zebrafish anatomy and development	ZDB	zebrafish anatomy.ontology	yes

Building Ontology



A tribute to Lewis Carroll

Once master the machinery of Symbolic Logic, and you have a mental occupation always at hand, of absorbing interest, and one that will be of real use to you in any subject you may take up. It will give you clearness of thought - the ability to see your way through a puzzle - the habit of arranging your ideas in an orderly and get-at-able form - and, more valuable than all, the power to detect fallacies, and to tear to pieces the flimsy illogical arguments, which you will so continually encounter in books, in newspapers, in speeches, and even in sermons, and which so easily delude those who have never taken the trouble to master this fascinating Art.

Lewis Carroll

(a) All babies are illogical.
(b) Nobody is despised who can manage a crocodile.
(c) Illogical persons are despised

Can a baby can manage a crocodile?

April 10, 2007