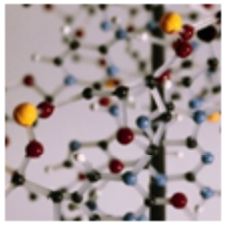
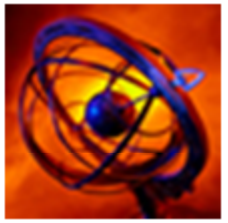


Introduction to Biomedical Ontologies

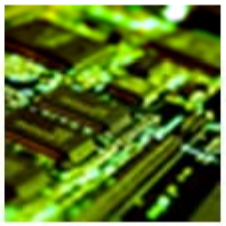
A TWO-DAY INTENSIVE TRAINING COURSE

Barry Smith, University at Buffalo











Background

- Working in ontology since 1975, with bio-ontologists and clinical ontologists since 2002
- Working with Gene Ontology since 2004
- Co-PI of the Protein Ontology (NIH/NIGMS)
- Coordinating Editor of the OBO (Open Biomedical Ontologies) Foundry

NCBO

- Dissemination and Ontology Best Practices of the National Center for Biomedical Ontology (PI Mark Musen, Stanford)
- http://bioontology.org

Example ontologies

Basic Formal Ontology (BFO)

Common Anatomy Reference Ontology (CARO)

Environment Ontology (EnvO)

Foundational Model of Anatomy (FMA)

Infectious Disease Ontology (IDO)

Ontology for Biomedical Investigations (OBI)

Ontology for Clinical Investigations (OCI)

Phenotypic Quality Ontology (PATO)

Relation Ontology (RO)

Ontologies and terminologies examined

SNOMED

Unified Medical Language System

National Cancer Institute Thesaurus

HL7 Reference Information Model

International Classification of Functioning, Disability and Health

Collaborations

Cleveland Clinic Semantic Database for Cardiovascular Surgery Ontology

Duke University Laboratory of Computational Immunology

German Federal Ministry of Heath

European Union Emergency Patient Summary Initiative

University of Pittsburgh Medical Center

University of Texas Southwestern Medical Center

Collaborations (Brain and Behavior)

- UB Task Force for Ontology-Based IT Support for Large-Scale Field Studies in Psychiatry
- Jacobs Neurological Institute, University at Buffalo
- Ontology Task Force (San Diego) of the Biomedical Informatics Research Network (BIRN)
- Neurocommons/Science Commons (MIT)

Agenda • Day 1

- **Introduction: What is an ontology and what is it useful for?**
- Basic Formal Ontology: An upper-level ontology to support scientific research
- Open Biomedical Ontologies (OBO) and the Web Ontology Language (OWL)
- The OBO Relation Ontology

Multiple kinds of data in multiple kinds of silos

Lab / pathology data

Electronic Health Record data

Clinical trial data

Patient histories

Medical imaging

Microarray data

Protein chip data

Flow cytometry

Mass spec

Genotype / SNP data

How to *find* your data?

How to reason with data when you find it?

How to understand the significance of the data *you* collected 3 years earlier?

How to integrate with other people's data?

Part of the solution must involve consensus-based, standardized terminologies and coding schemes

Ontologies facilitate retrieval of data

by allowing grouping of annotations

brain	**20**
hindbrain	**15**
rhombomere	**10**

Query brain without ontology **20**

Query brain with ontology **45**

Making data (re-)usable through standards

- Standards provide
  - common structure and terminology
  - single data source for review (less redundant data)
- Standards allow
  - use of common tools and techniques
  - common training
  - single validation of data

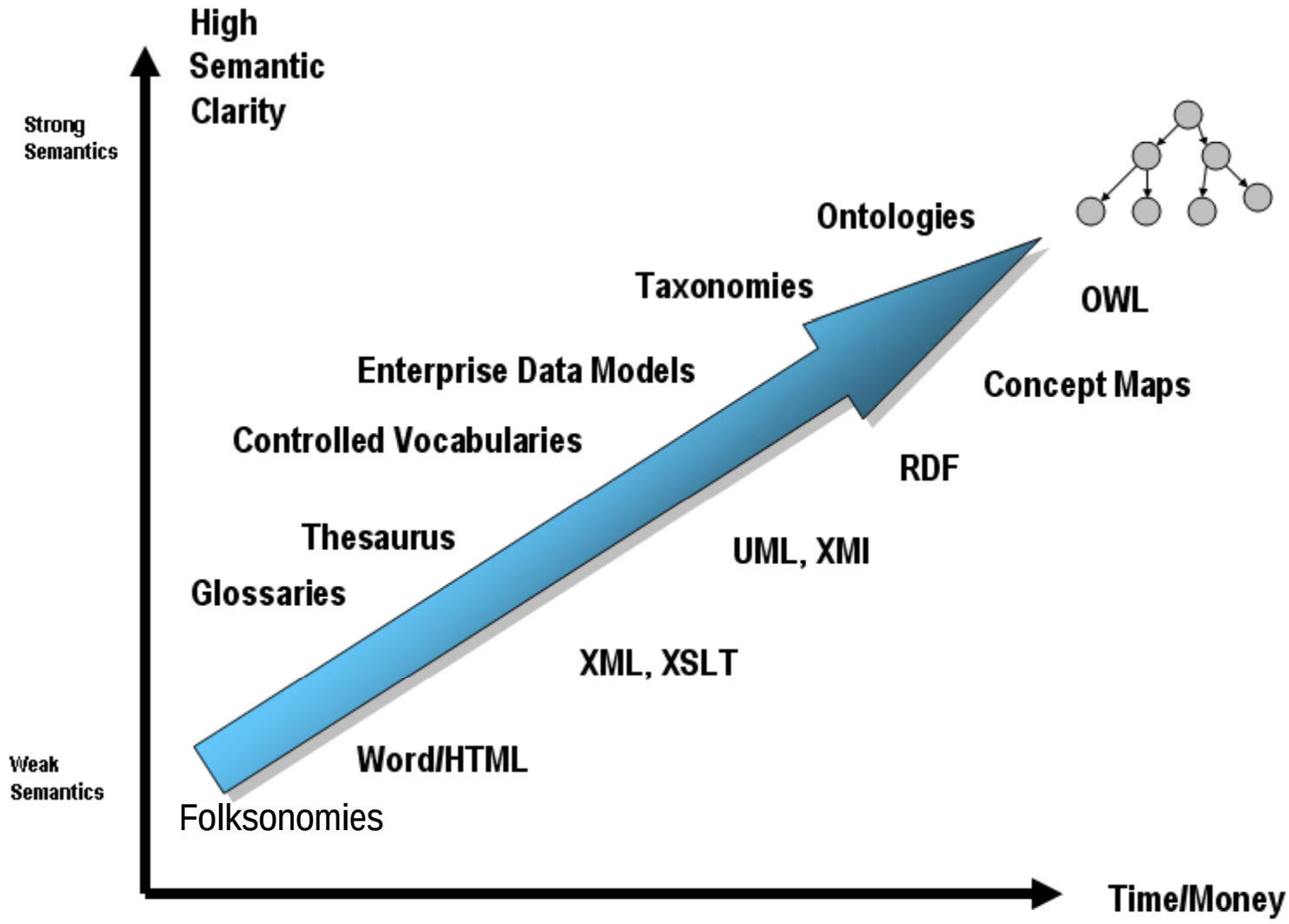
Unifying goal: integration

- within and across domains
- across different species
- across levels of granularity (organ, organism, cell, molecule)
- across different perspectives (physical, biological, clinical)

Problems with standards

- Standards involve considerable costs of re-tooling, maintenance, training, ...
- They pose risks to flexibility
- May break legacy solutions which work locally
- Not all standards are of equal quality
- Bad standards create lasting problems
- ‘Ontology’ = good standards in terminology

Leo Obrst: The Ontology Spectrum



http://www.mkbergman.com/?m=20070516 (slightly modified)

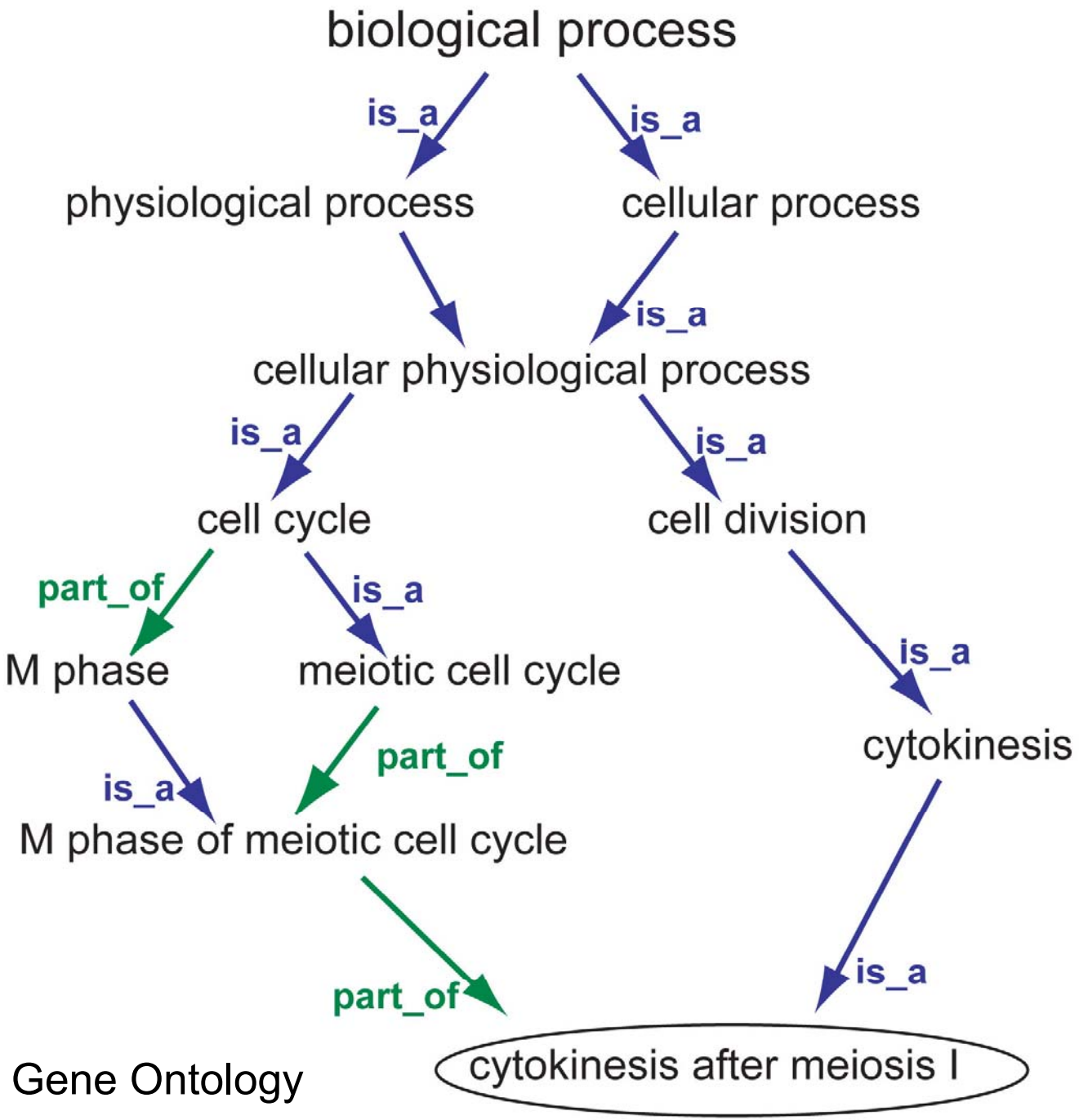
The wisdom of clouds (folksonomies ...)

07 2007 2008 51207 51507 51607 51707 academy accs andy blog centennial ceremony charlottesville coffee collabratory commencement compression concer t conference datlat digital diskutility dtlt dv dvd enc oders engl375mm eve6 fa07 fa08 faculty facultyacad emy ffmpegx foucault fredericksburg fredericksburgnor malandindustrialinstitute freshman fsem100j globalization grad graduation gravatar greenlaw groom header hi storic historical history homecoming img jim jimschair js kenmore learning marker markers mary mashup mashups microsoft nmc nmc2007 ontology patrick p opfly presentation reverend reverendjimtshirt rush sa0 7 sa2007 semantic seminar student studentacademy symposium teaching technologies transfer tshirt umw umwblogs umwead university video virginia virginiahis toricalmarkers visualizing wall

Ontologies are, at least, controlled structured vocabularies

providing definitions and reasoning

including support for automatic validation of ontology structure



from the Gene Ontology

NIH Mandates for Sharing of Research Data

Investigators submitting an NIH application seeking $500,000 or more in any single year are expected to include a *plan for data sharing*

(http://grants.nih.gov/grants/policy/data_sharing)

Program Announcement Number: PAR-07-425

Title: Data Ontologies for Biomedical Research (R01)

NIH Blueprint for Neuroscience Research, (http://neurosciencebluepr int.nih.gov/)
National Cancer Institute (NCI), (http://www.cancer.gov)
National Center for Research Resources (NCRR), (http://www.ncrr.nih.gov/)
National Eye Institute (NEI), (http://www.nei.nih.gov/)
National Heart Lung and Blood Institute (NHLBI), (http://http.nhlbi.nih.gov)
National Human Genome Research Institute (NHGRI), (http://www.genome.gov)
National Institute on Alcohol Abuse and Alcoholism (NIAAA), (http://www.niaaa.nih.gov/)
National Institute of Biomedical Imaging and Bioengineering (NIBIB), (http://www.nibib.nih.gov/)
National Institute of Child Health and Human Development (NICHD), (http://www.nichd.nih.gov/)
National Institute on Drug Abuse (NIDA), (http://www.nida.nih.gov/)
National Institute of Environmental Health Sciences (NIEHS), (http://www.niehs.nih.gov/)
National Institute of General Medical Sciences (NIGMS), (http://www.nigms.nih.gov/)
National Institute of Mental Health (NIMH), (http://www.nimh.nih.gov/)
National Institute of Neurological Disorders and Stroke (NINDS), (http://www.ninds.nih.gov/)
National Institute of Nursing Research (NINR), (http://www.ninr.nih.gov)

PAR-07-425 Purpose

Optimal use of informatics tools and data resources depends upon explicit understandings of concepts related to the data upon which they compute. This is typically accomplished by a tool or resource adopting a formal controlled vocabulary and ontology.

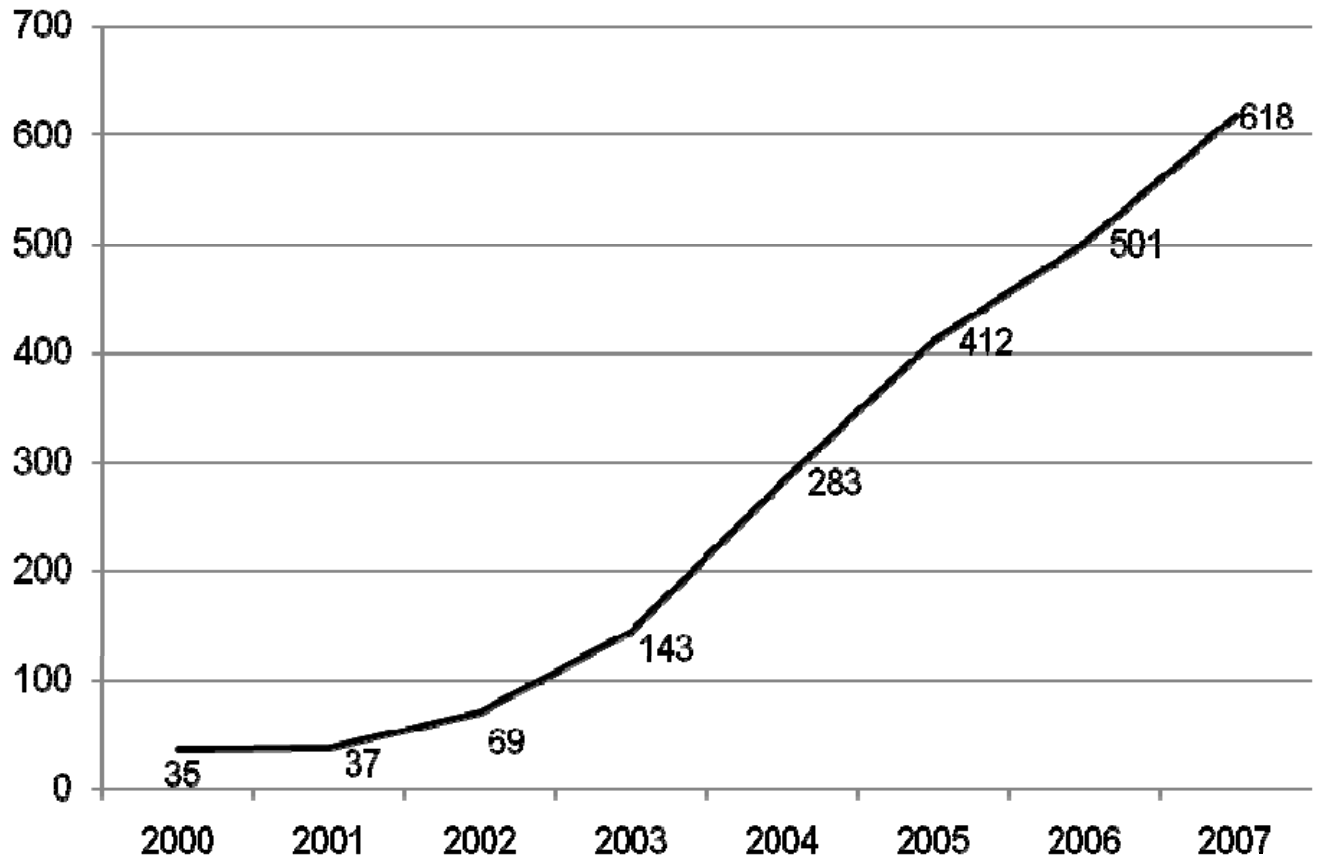
Currently, there is no convenient way to map the knowledge that is contained in one data set to that in another data set, primarily because of differences in language and structure

... in some areas there are emerging standards.

Examples include:

- the Unified Medical Language System (**UMLS**),
- the **Gene Ontology**,
- the **caBIG** project,
- Open Biomedical Ontologies (**OBO**)

Uses of ‘ontology’ in PubMed abstracts



Types of ontologies

	Upper-level integrating ontologies	Domain ontologies
Ontologies in support of science		
Administrative ontologies		

Types of ontologies

	Upper-level integrating ontologies	Domain ontologies
Ontologies in support of science	*BFO (Basic Formal Ontology)* *DOLCE, SUMO*	*GO* *FMA* *SNOMED*
Administrative ontologies (e-commerce, etc.)	*FOAF top level:* *person, topic, document, primary topic ...*	*Amazon.com ontology* *Library of Congress Catalog*

Scientific ontologies vs. administrative ontologies

BFO, GO, FMA …

vs.

Library of Congress Catalog, Yahoo ontology, FirstGov Life Events Taxonomy, …

Part of our goal is realized if we can create controlled terminologies

In science we can and must go further than this

Why build scientific ontologies?

- There are many ways to create terminologies
- Multiple terminologies will not solve our data silo problems
- We need to constrain terminologies so that they converge

Evidence-based terminology development

Q: What is to serve as constraint?

A1: Authority?

A2: First in field (Founder effect)?

A3: Best candidate terminology?

A4: Reality, as revealed, incrementally, by experimentally-based science

The standard methodology

- Pragmatics is everything
- It is easier to write useful software if one works with a simplified model
- ("…we can't know what reality is like in any case; we only have our concepts…")
- This looks like a useful model to me
- (One week goes by:) This other thing looks like a useful model to him
- Data in Pittsburgh does not interoperate with data in Vancouver
- Science is siloed

The methodology of ontological realism

- Find out what the world is like by doing science, talking to other scientists and working continuously with them to ensure that you don't go wrong
- Build representations adequate to this world, not to some simplified model in your laptop
- Ontology is ineluctably a multi-disciplinary enterprise – need to work hard to overcome the resultant terminological confusions

Our first job is in to create a common understanding of terms such as:

- universal, type, kind, class
- instance
- model
- representation
- data

Entity =def

anything which exists, including things and processes, functions and qualities, beliefs and actions, documents and software

Scientific ontologies have special features

Every term must be such that the developers of the ontology believe it to refer to some entity on the basis of the best current scientific evidence

(Important role of instances that we can observe in the laboratory)

Administrative ontologies

- Entities may be brought into existence by the ontology itself. (Convention ...)
- Highly task-dependent – reusability and compatibility not (always) important
- Can be secret
- Are comparable to software artifacts

For scientific ontologies

openness, reusability and

compatibility with neighboring scientific ontologies are crucial

- Scientific ontologies must evolve gracefully
- Scientific ontologies must be evidence-based
- Scientific ontologies are comparable to scientific theories

The central distinction

***universal* vs. instance**

(*catalog* vs. inventory)

(*science text* vs. diary)

(*human being* vs. Arnold Schwarzenegger)

Science texts are representations of universals in reality

= representations of what is *general* in reality

Clinical guidelines are representations of universals in reality

diseases, therapies, diagnostic procedures (measurements) are generals, with particular instances in particular patients

Ontologies are representations of universals in reality

aka kinds, types, categories, species, genera, ...

instances

A	515287	**DC3300 Dust Collector Fan**
B	521683	**Gilmer Belt**
C	521682	**Motor Drive Belt**

universals

Catalog vs. inventory

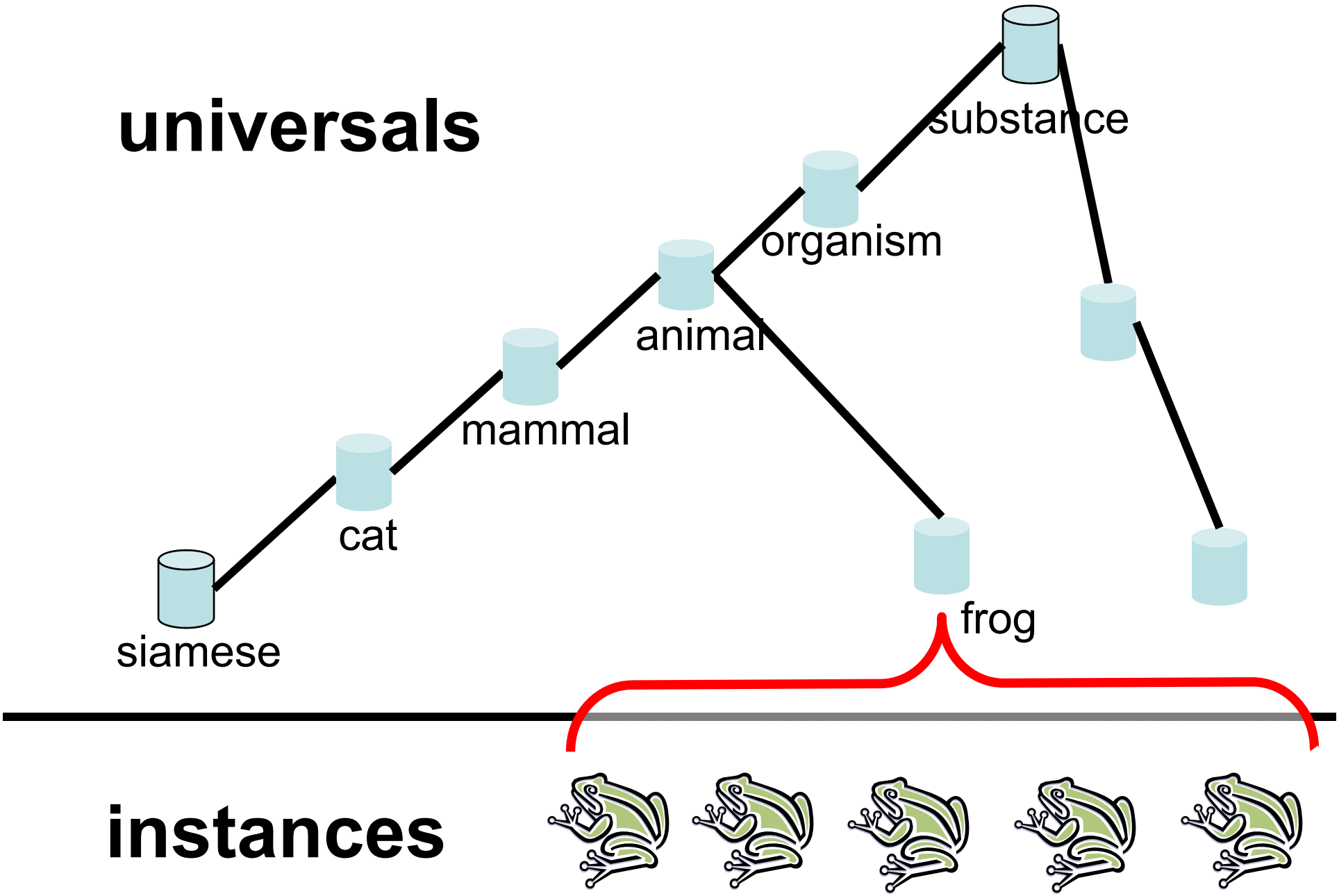
For scientific ontologies

it is *generalizations* (universals) that are important

For databases it is (normally) instances that are important

= particulars in reality:

- patient #0000000001
- headache #000000004
- MRI image #23300014, etc.


universals
substance
organism
animal
mammal
cat
siamese
frog
instances

In a scientific ontology

every node in the ontology should represent both universals *and* the corresponding instances in reality

every term should reflect instances – it is instances which form the objects of our experiments

to do this is hard work …

Each term in an ontology represents exactly one universal

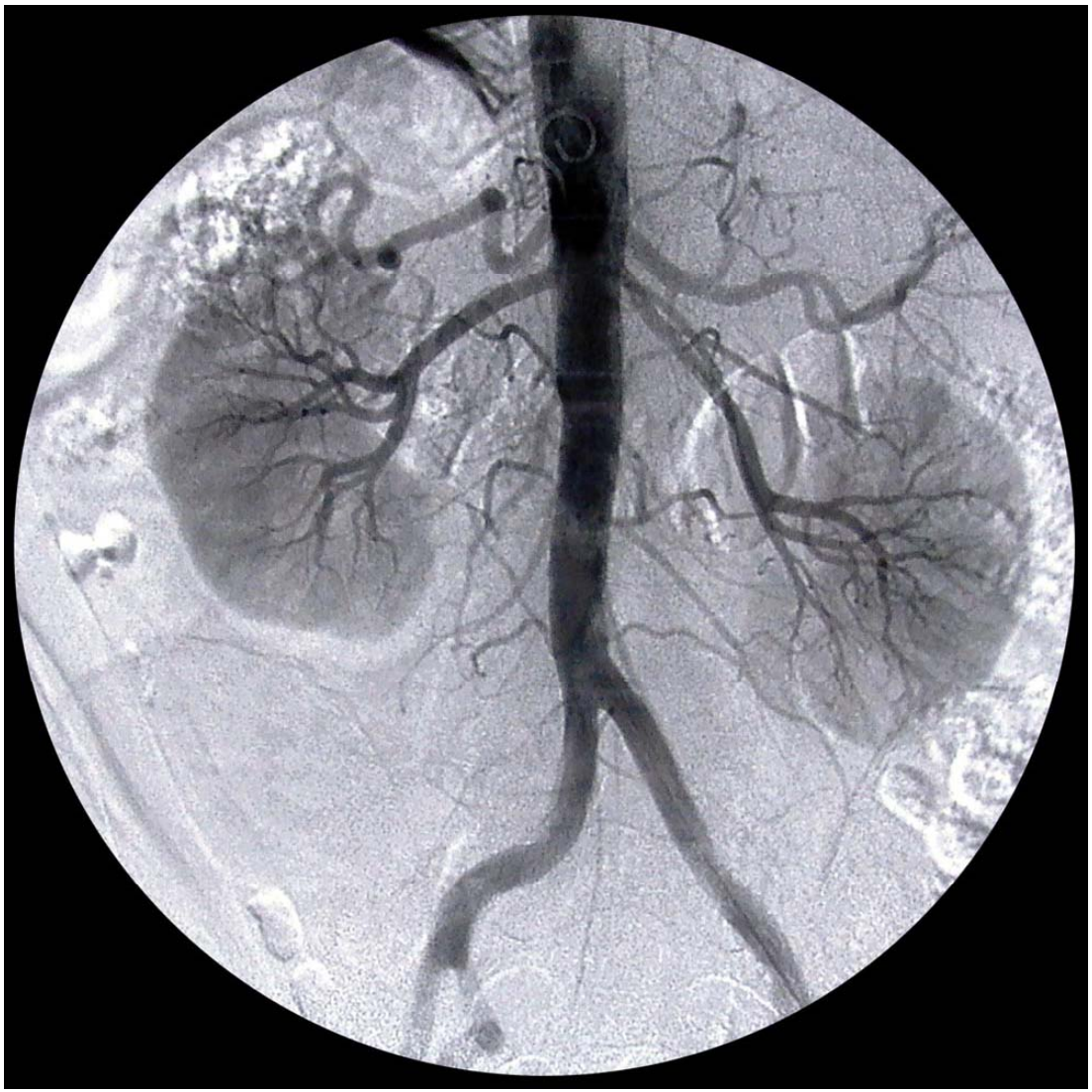
For this reason ontology terms should be **singular nouns**

- headache
- human being
- drug administration

An ontology is a representation of universals

We learn about universals in reality from looking at the results of scientific experiments as expressed in the form of scientific theories – which describe, not what is particular in reality, but what is general

A photographic image is a representation of an instance



Three Levels to Keep Straight

- Level 1: the entities in reality, both instances and universals
- Level 2: cognitive representations of this reality on the part of scientists ...
- Level 3: publicly accessible concretizations of these cognitive representations in textual and graphical artifacts

Ontology development

starts with: Level 2 = the cognitive representations of practitioners or researchers in the relevant domain

results in: Level 3 representational artifacts (comparable to maps, science texts, dictionaries)

Domain =def.

a portion of reality that forms the subject-matter of a single science or technology or mode of study;

proteomics

HIV

demographics

...

Representation =def.

an image, idea, map, picture, name or description ... of some entity or entities

two kinds of representation:

- analogue (photographs)
- digital/composite/syntactically structured

Representational units =def

terms, icons, alphanumeric identifiers ... which refer, or are intended to refer, to entities

and which are minimal (‘atoms’)

Composite representation =def

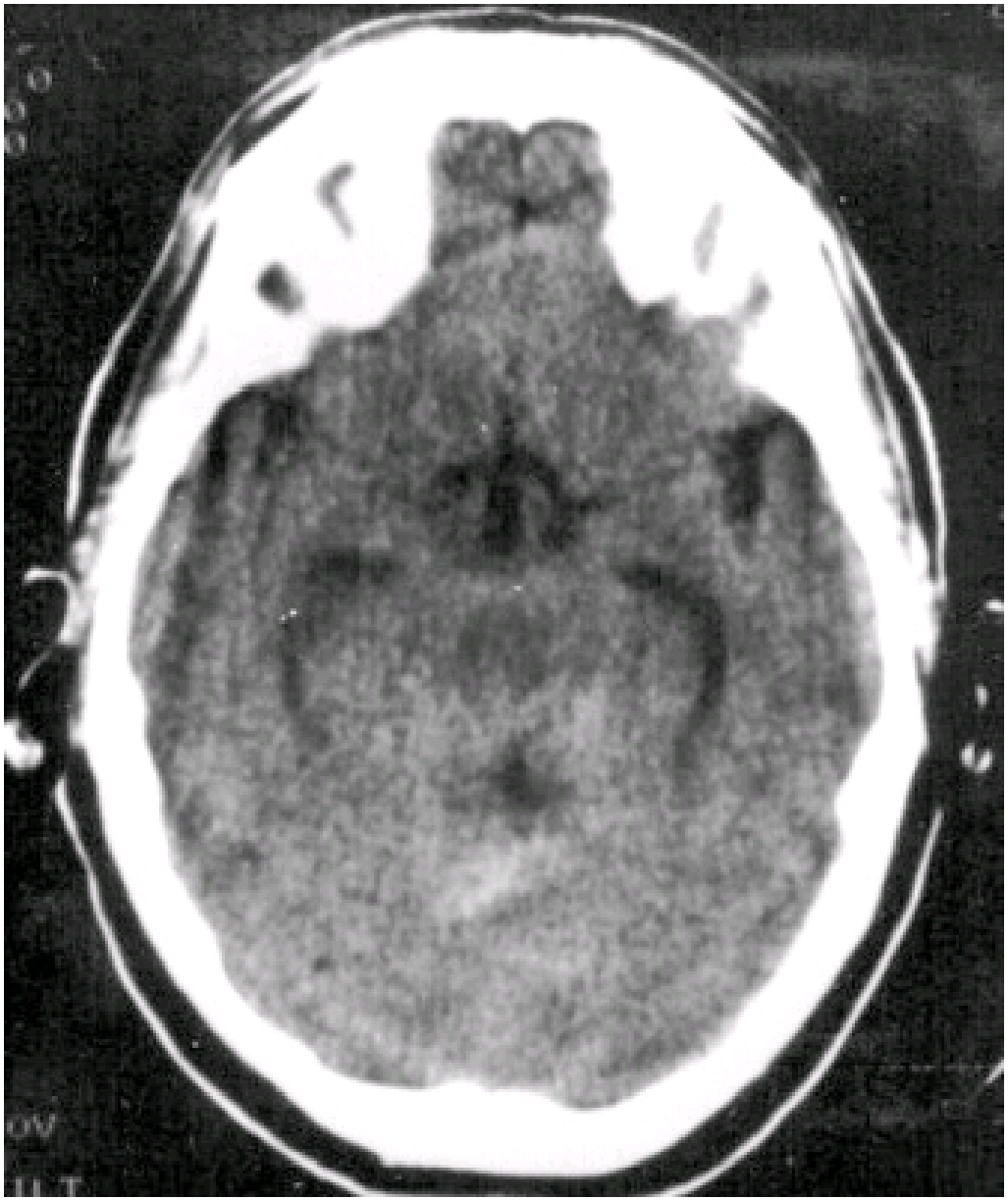
a representation

(1) built out of representational units

which

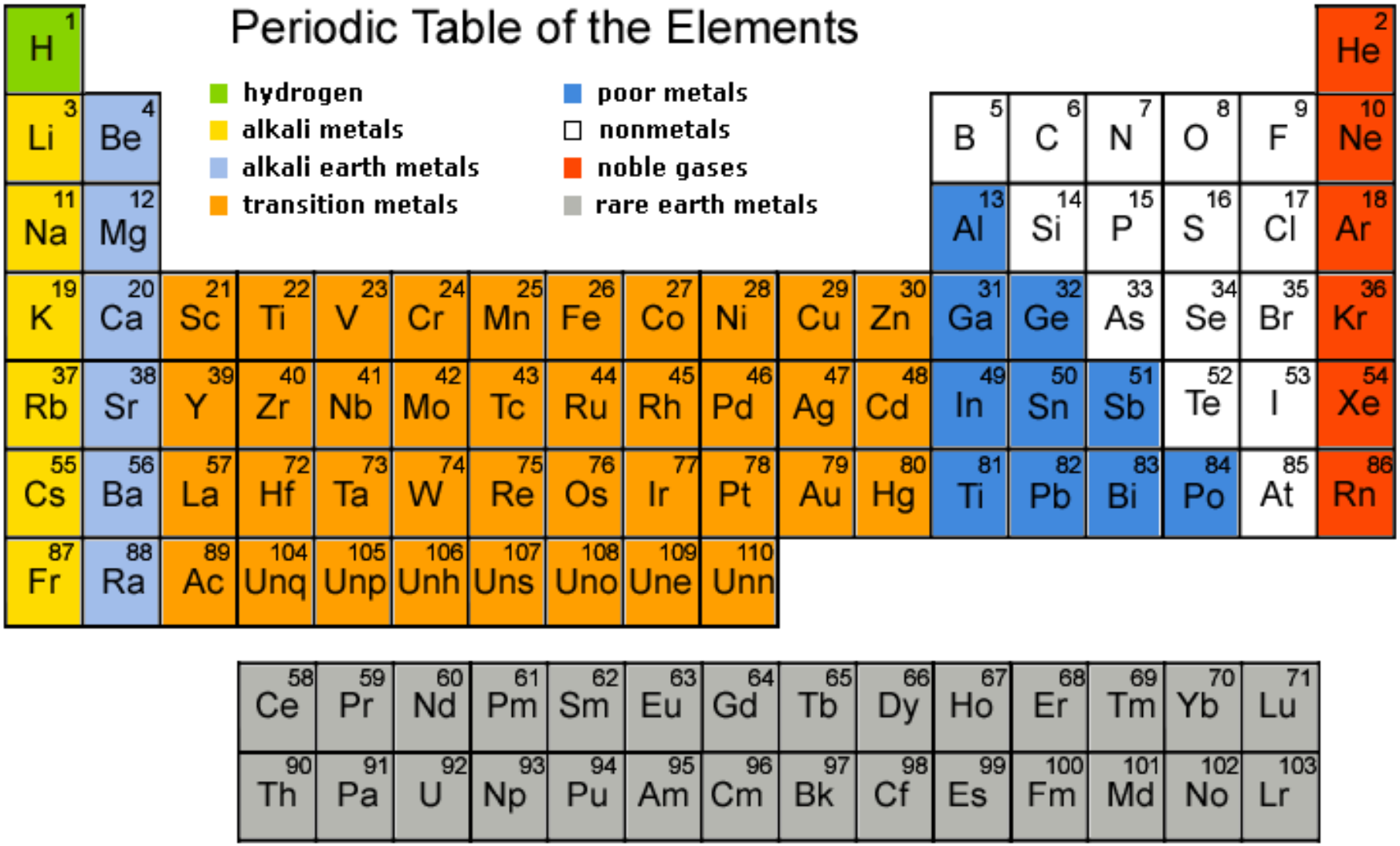
(2) form a structure that mirrors, or is intended to mirror, the entities in some domain

Analogue representations



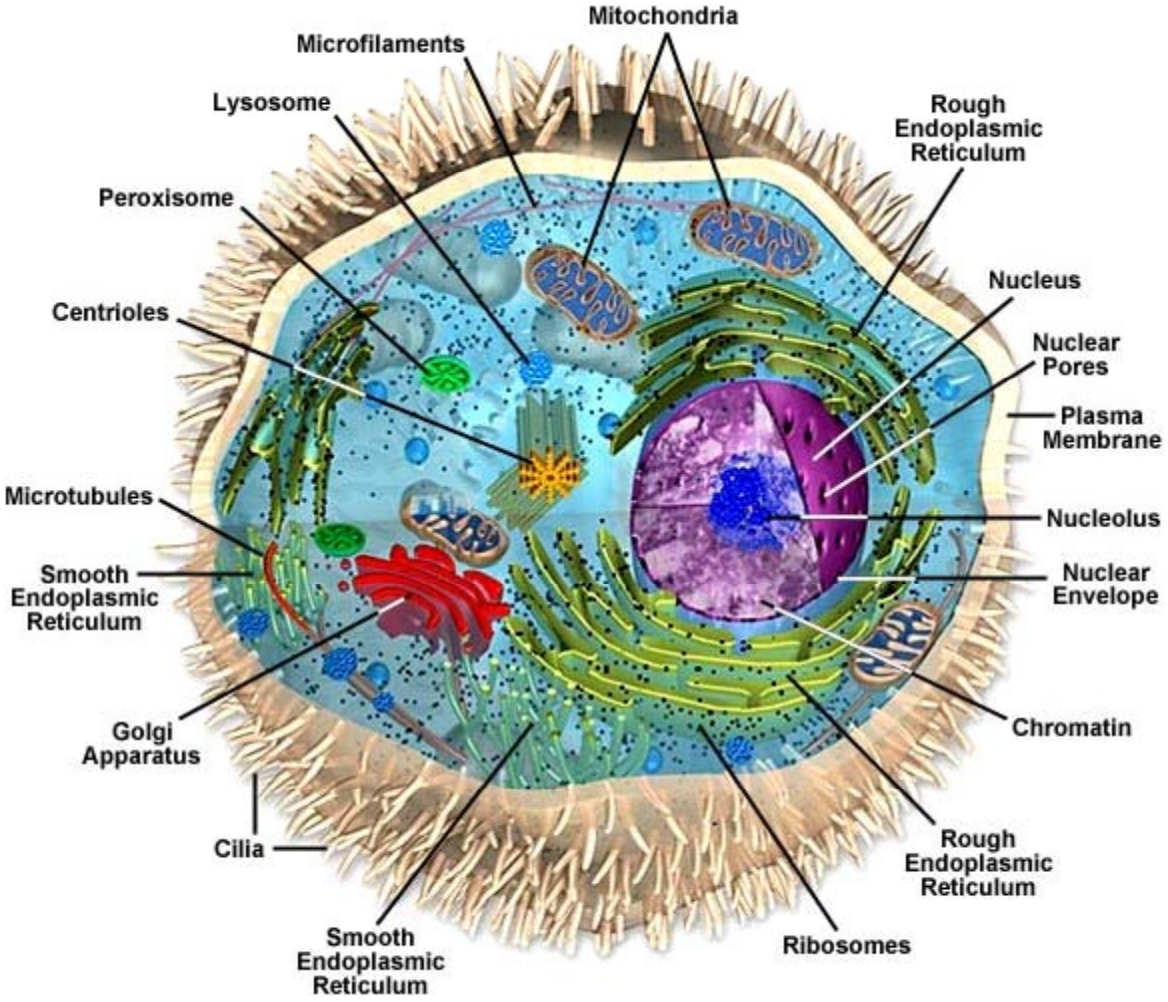
The Periodic Table

H 1	Periodic Table																He 2
Li 3	Be 4											B 5	C 6	N 7	O 8	F 9	Ne 10
Na 11	Mg 12											Al 13	Si 14	P 15	S 16	Cl 17	Ar 18
K 19	Ca 20	Sc 21	Ti 22	V 23	Cr 24	Mn 25	Fe 26	Co 27	Ni 28	Cu 29	Zn 30	Ga 31	Ge 32	As 33	Se 34	Br 35	Kr 36
Rb 37	Sr 38	Y 39	Zr 40	Nb 41	Mo 42	Tc 43	Ru 44	Rh 45	Pd 46	Ag 47	Cd 48	In 49	Sn 50	Sb 51	Te 52	I 53	Xe 54
Cs 55	Ba 56	La 57	Hf 72	Ta 73	W 74	Re 75	Os 76	Ir 77	Pt 78	**Au** 79	Hg 80	Tl 81	Pb 82	Bi 83	Po 84	At 85	Rn 86
Fr 87	Ra 88	Ac 89	Rf 104	Ha 105	?? 106												
Lanthinide Series		Ce 58	Pr 59	Nd 60	Pm 61	Sm 62	Eu 63	Gd 64	Tb 65	Dy 66	Ho 67	Er 68	Tm 69	Yb 70	Lu 71		
Actinide Series		Th 90	Pa 91	U 92	Np 93	Pu 94	Am 95	Cm 96	Bk 97	Cf 98	Es 99	Fm 100	Md 101	No 102	Lr 103		


Periodic Table of the Elements
hydrogen
alkali metals
alkali earth metals
transition metals
poor metals
nonmetals
noble gases
rare earth metals
H 1
He 2
Li 3
Be 4
B 5
C 6
N 7
O 8
F 9
Ne 10
Na 11
Mg 12
Al 13
Si 14
P 15
S 16
Cl 17
Ar 18
K 19
Ca 20
Sc 21
Ti 22
V 23
Cr 24
Mn 25
Fe 26
Co 27
Ni 28
Cu 29
Zn 30
Ga 31
Ge 32
As 33
Se 34
Br 35
Kr 36
Rb 37
Sr 38
Y 39
Zr 40
Nb 41
Mo 42
Tc 43
Ru 44
Rh 45
Pd 46
Ag 47
Cd 48
In 49
Sn 50
Sb 51
Te 52
I 53
Xe 54
Cs 55
Ba 56
La 57
Hf 72
Ta 73
W 74
Re 75
Os 76
Ir 77
Pt 78
Au 79
Hg 80
Ti 81
Pb 82
Bi 83
Po 84
At 85
Rn 86
Fr 87
Ra 88
Ac 89
Unq 104
Unp 105
Unh 106
Uns 107
Uno 108
Une 109
Unn 110
Ce 58
Pr 59
Nd 60
Pm 61
Sm 62
Eu 63
Gd 64
Tb 65
Dy 66
Ho 67
Er 68
Tm 69
Yb 70
Lu 71
Th 90
Pa 91
U 92
Np 93
Pu 94
Am 95
Cm 96
Bk 97
Cf 98
Es 99
Fm 100
Md 101
No 102
Lr 103

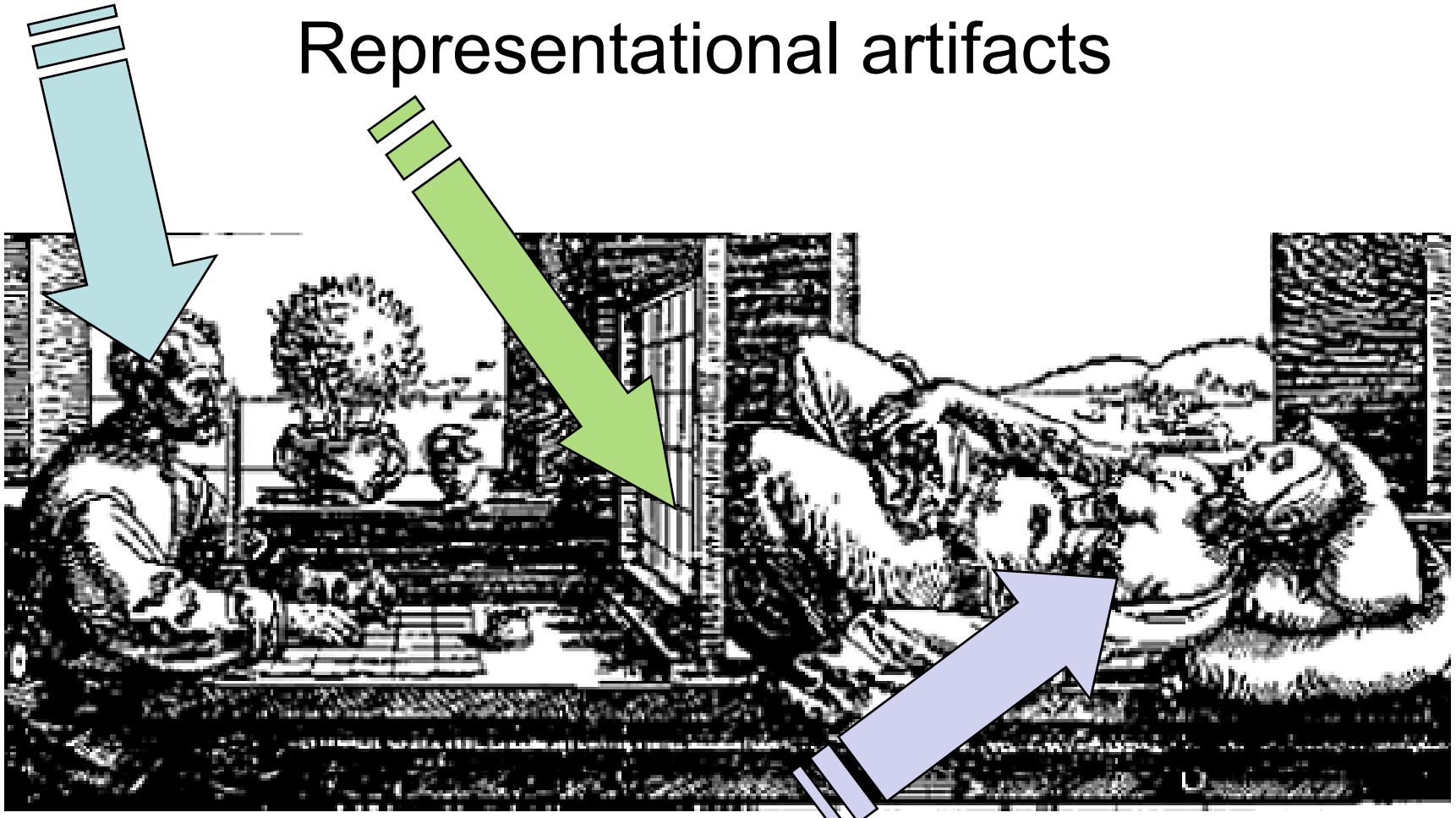
We can’t take photographs of universals

But we can create cartoons and diagrams



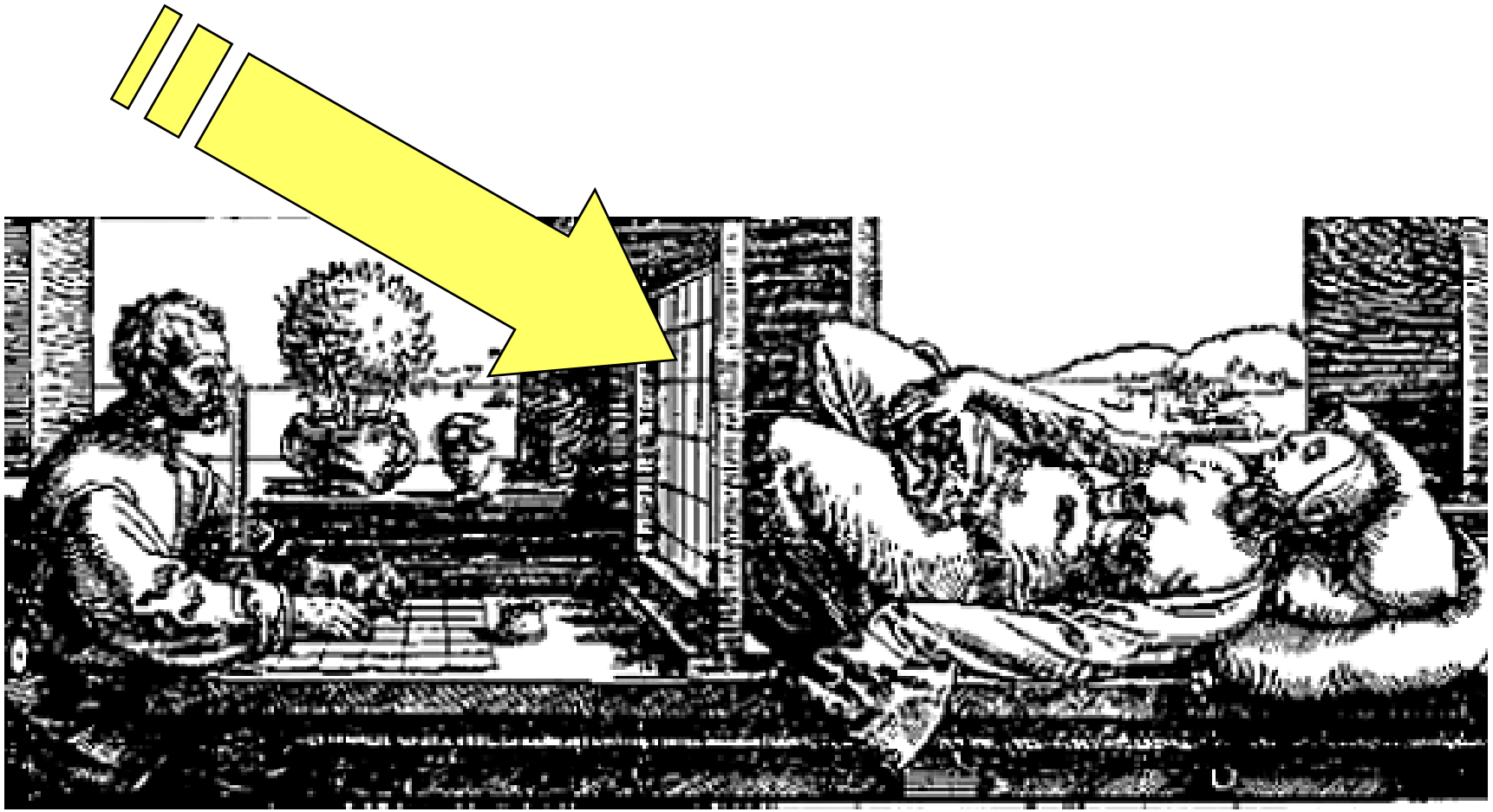
Cognitive representations

Representational artifacts



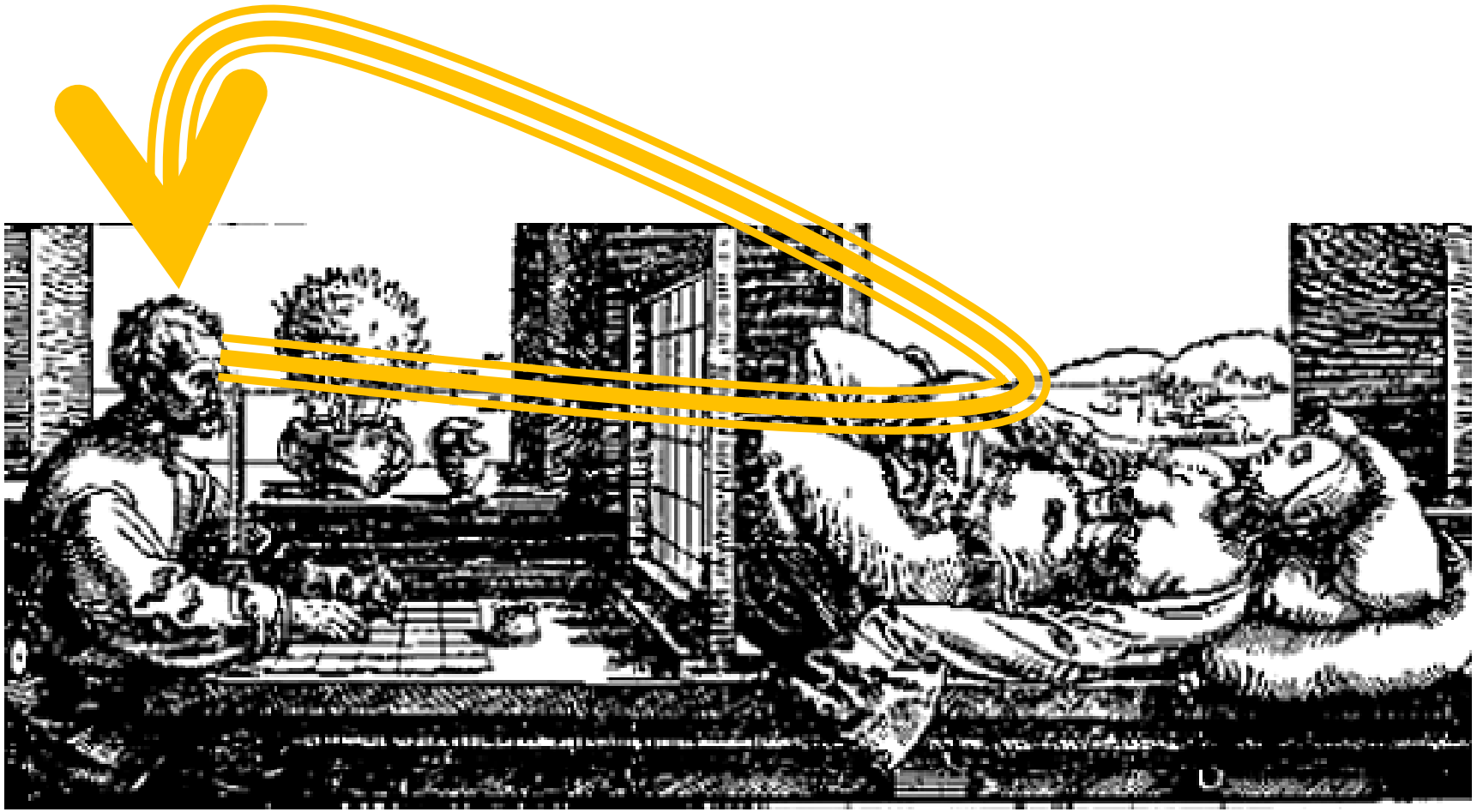
Reality

Ontologies are here



or here

Ontologies do *not* represent concepts in people's heads



Like the scientific theories from which they derive, they represent universals in reality

e.g. *leg*

Compare the typical relations used in medical ontologies

part_of

connected_to

adjacent_to

causes

treats ...

“leg” is not the name of a concept

concepts do not stand in

part_of

connected_to

adjacent_to

causes

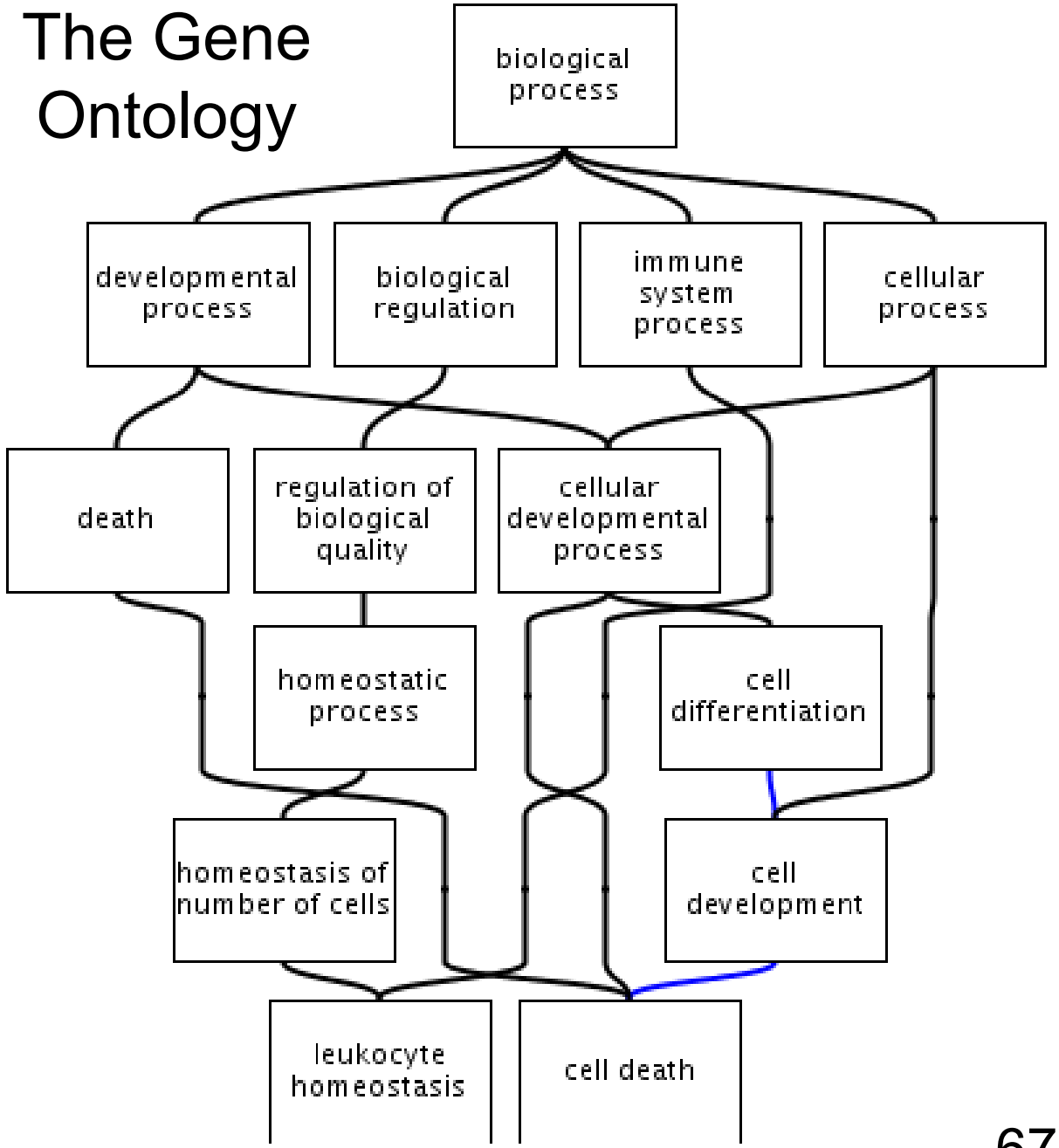
treats ...

relations to each other

The Gene Ontology

is_a —

part_of —



How do we know which general terms designate universals?

Roughly: terms used in a plurality of sciences to designate entities about which we have a plurality of different kinds of testable propositions / laws

(compare: cell, electron, membrane ...)

Class =def.

a maximal collection of particulars referred to by a general term

the class *A* =def. the collection of all particular *A*'s

where '*A*' is a general term (e.g. 'brother of Elvis fan', 'cell')

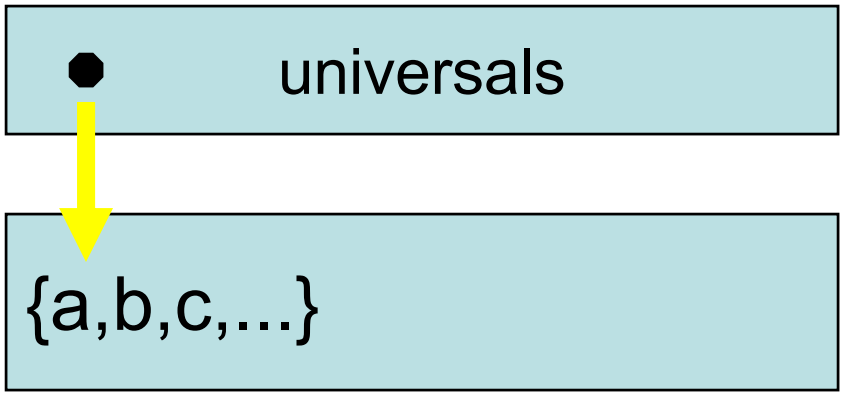
Classes are on the same level as the instances which they contain

Extension =def

the collection of all particular *A*'s, where '*A*' is the name of a universal

universals vs. their extensions

The extension of the *universal A* is the class of *A's* instances



collections of particulars

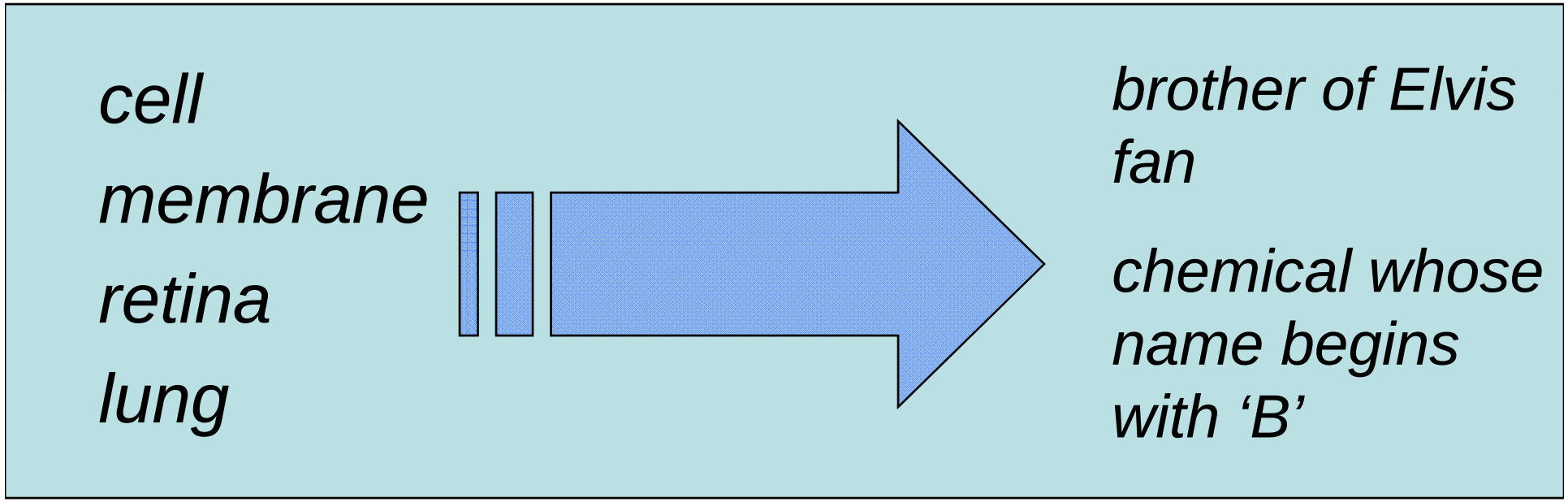
Problem

The same general term can be used to refer both to universals and to collections of particulars.

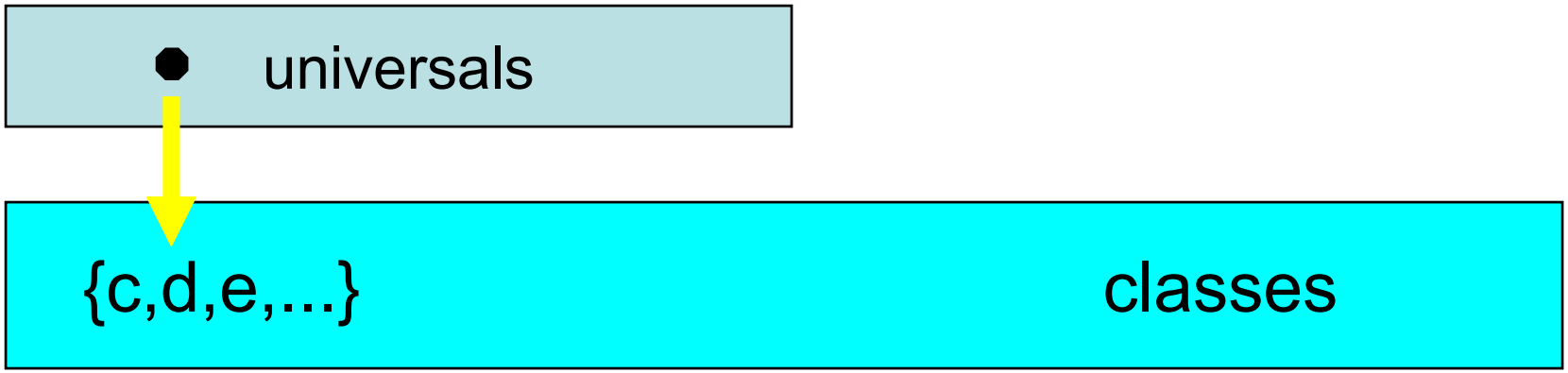
HIV is an infectious retrovirus

HIV is spreading very rapidly through Asia

a spectrum of cases



Not all classes correspond to universals



Administrative ontologies often go beyond universals

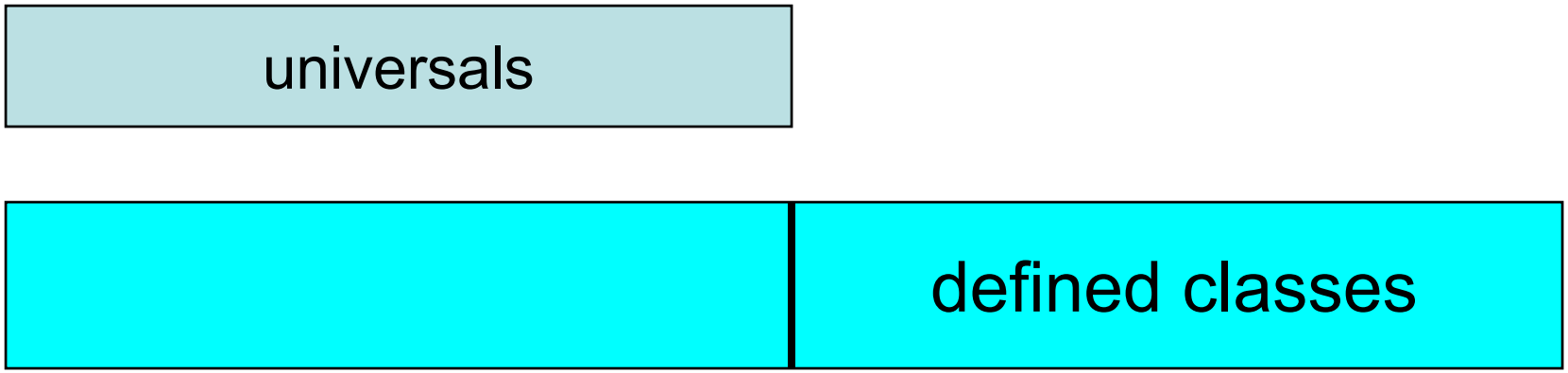
Fall on stairs or ladders in water transport injuring occupant of small boat, unpowered

Railway accident involving collision with rolling stock and injuring pedal cyclist

Non-traffic accident involving motor-driven snow vehicle injuring pedestrian

ICD (WHO International Classification of Diseases)

universals vs. classes



Defined class =def

a class defined by a general term which does not designate a universal

person called ‘Chris’

person with diabetes in Maryland on 4 June 1952

OWL (Ontology Web Language) is a good representation of defined classes

sibling of Finnish spy

member of Abba aged > 50 years

property-owning farm employee

such set-theoretic combinations are at the heart of many administrative ontologies

(Scientific) Ontology =def.

a representational artifact whose representational units (which may be drawn from a natural or from some formalized language) are intended to represent

1. universals in reality

2. those relations between these universals which obtain universally (= for all instances)

*lung **is_a** anatomical structure*

*lobe of lung **part_of** lung*