Current characteristics and historical perspective of CS & IT with/for Biology

C. Maria (Marijke) Keet

KRDB Research Centre
Faculty of Computer Science
Free University of Bozen-Bolzano, Italy
keet@inf.unibz.it

20-10-2005, Bolzano, Italy

Overview

- Introduction
- Historical aspects
- Characteristics
- Brief introduction seminar series programme

Overview

- **Introduction**
- Historical aspects
- Characteristics
- Brief introduction seminar series programme

Introduction

- Disclaimer
    - My background
    - High-level overview (tech details in other seminars)

- CS/IT *for* Biology ⇨ support for biosciences
- CS/IT *with* Biology ⇨‘new’ discipline?
- Biology *for* CS/IT??

Introduction: some terms

Computational Biology

Biomedical Engineering

Biocomputing

(Meta)Genomics,
Proteomics,
Metabolomics

Climate modelling

Bioinformatics

Agricultural Informatics

Theoretical Biology

Geographical Information Systems

Ecoinformatics

Nanotechnology

Computational Chemistry

Environmental engineering

Medical Informatics

In Silico (molecular) Biology

Bio-ontologies

Theoretical Ecology

Mathematics and Biology

Systems Biology

Overview

- Introduction
- **Historical aspects**
- Characteristics
- Brief introduction seminar series programme

Historical aspects

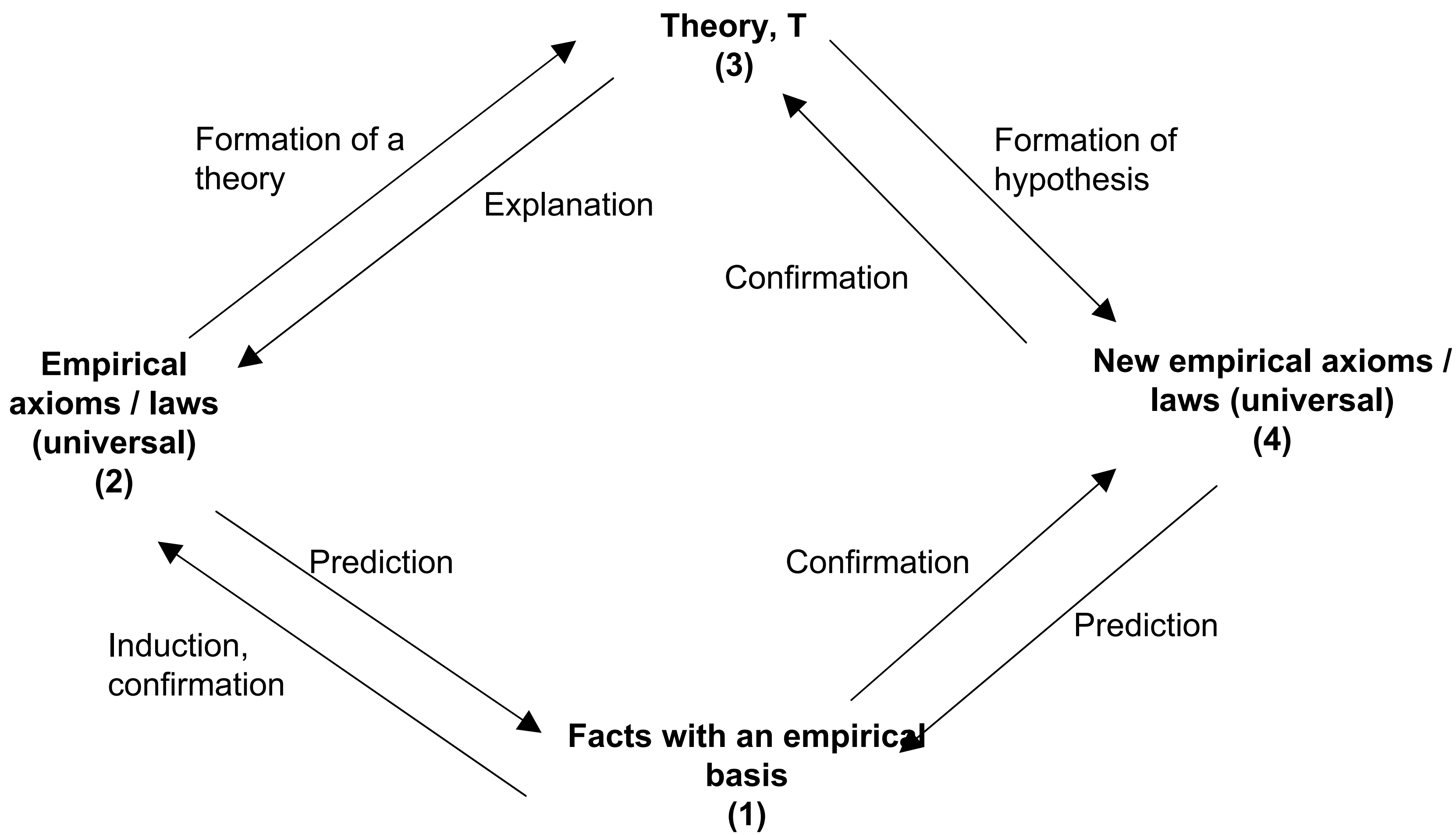
- ± ≥ ‘30s: Systems biology (maths & eco/cellbio)
- ± ≥ ‘50s:
  - Biomechanics/biomedical engineering
  - Analysis machines (hw/sw)
- ± ≥ ‘70s:
  - Climate simulations (supercomputers)
  - Agricultural systems (hw/sw)
  - Medical informatics (DSS)
- ± early ‘90s: start molbio-DBs

Historical aspects

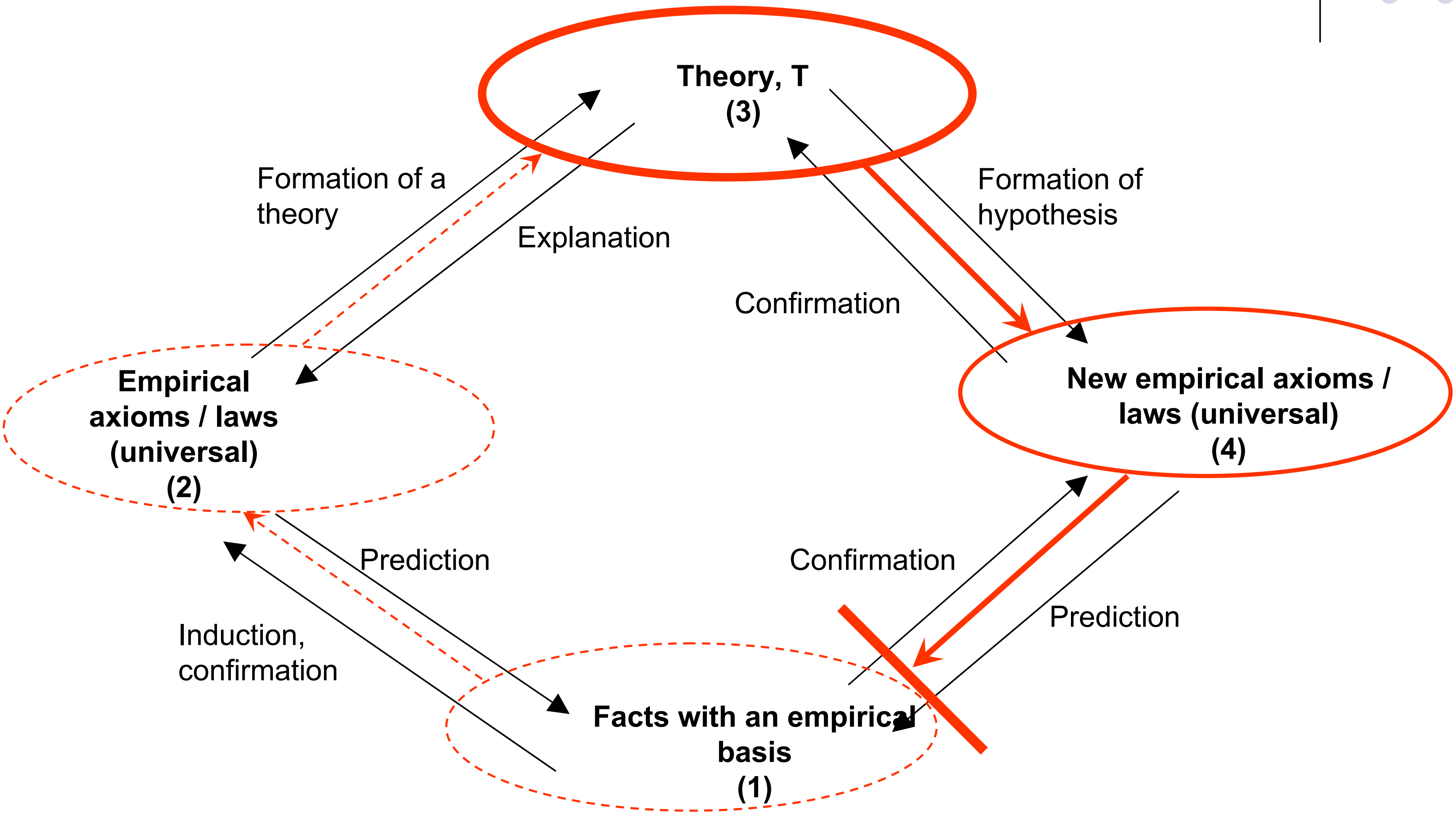
- ± ≥ ‘30s: Systems biology (maths & eco/cellbio)
- ± ≥ ‘50s:
  - Biomechanics/biomedical engineering
  - Analysis machines (hw/sw)
- ± ≥ ‘70s:
  - Climate simulations (supercomputers)
  - Agricultural systems (hw/sw)
  - Medical informatics (DSS)
- ± early ‘90s: start molbio-DBs

------------------MID ‘90s EXPLOSION-----------------

Historical aspects

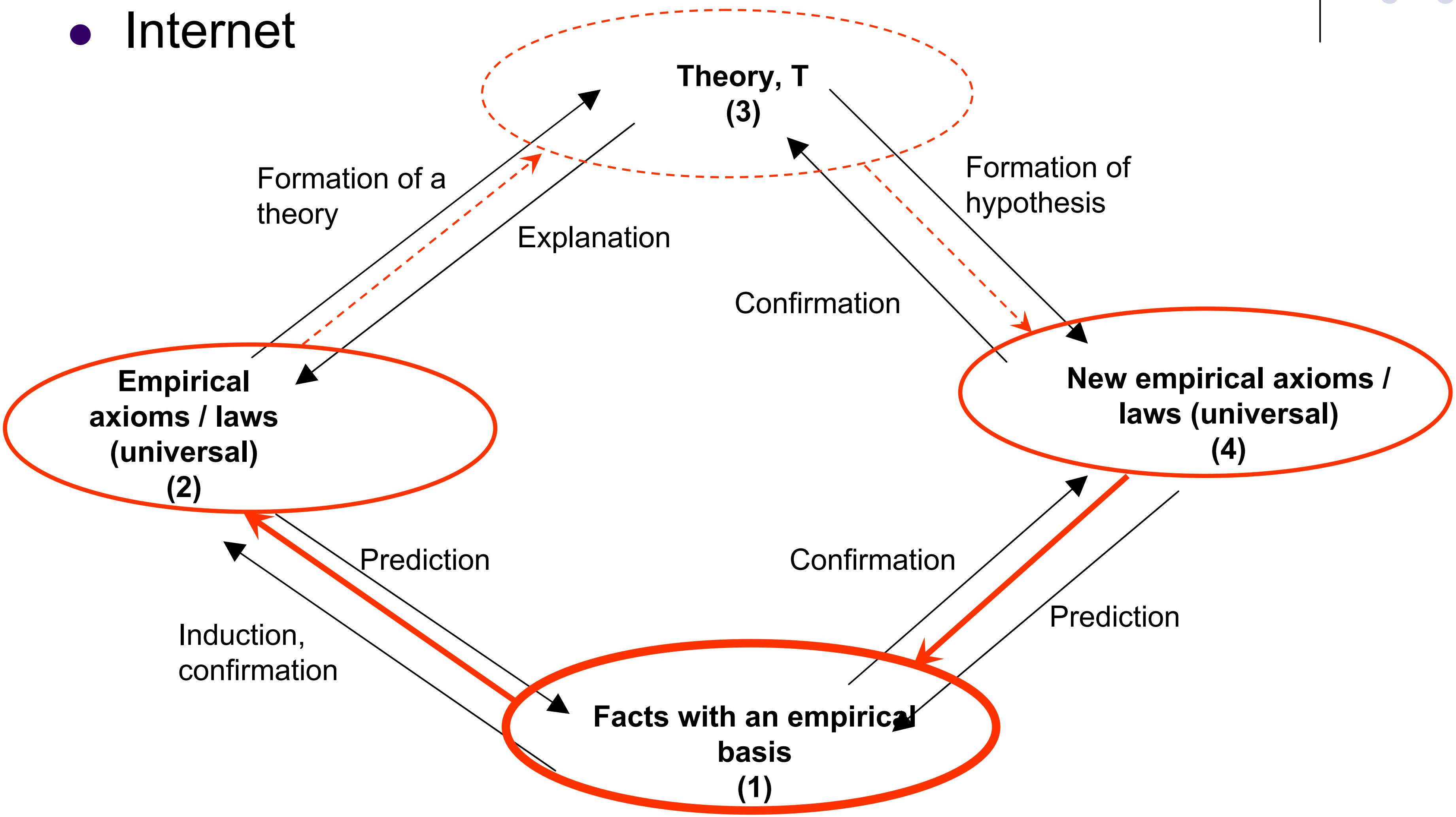


Historical aspects: < mid '90s



Historical aspects: ≥ mid ‘90s

- Cheaper computers
- Internet



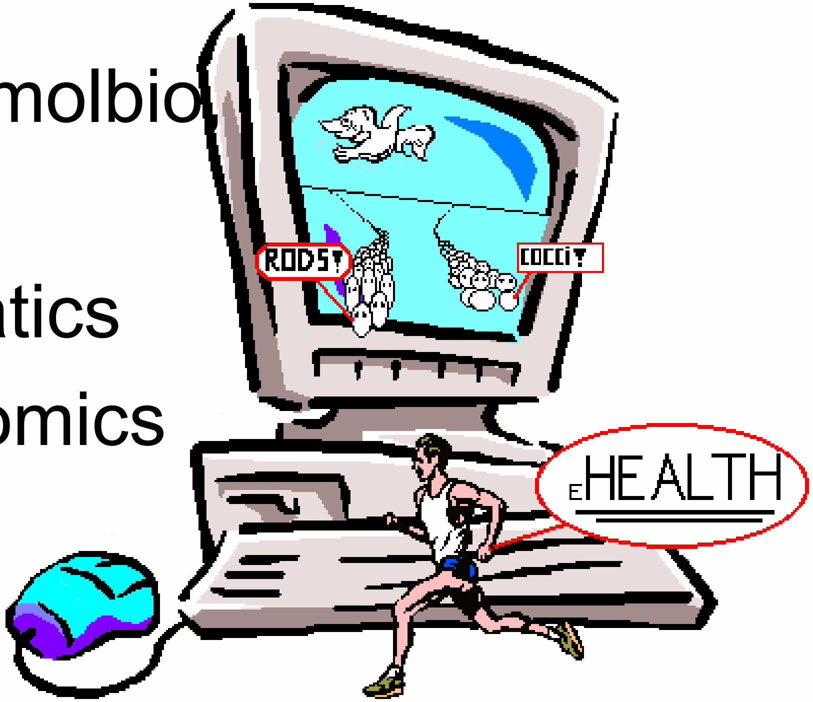
New possibilities for biology, in particular *molecular* biology

Historical aspects

- ± ≥ mid/late ‘90s: sequence comparison algorithms, more databases, database integration (DWH, FD), software tools
- 1998: start Gene Ontology Consortium
- ± ≥ ‘00s: bio-ontologies, computational linguistics (both NLP and NLG), data mining and pattern finding, Grids, workflow, ‘hypothesis generation’

Historical aspects

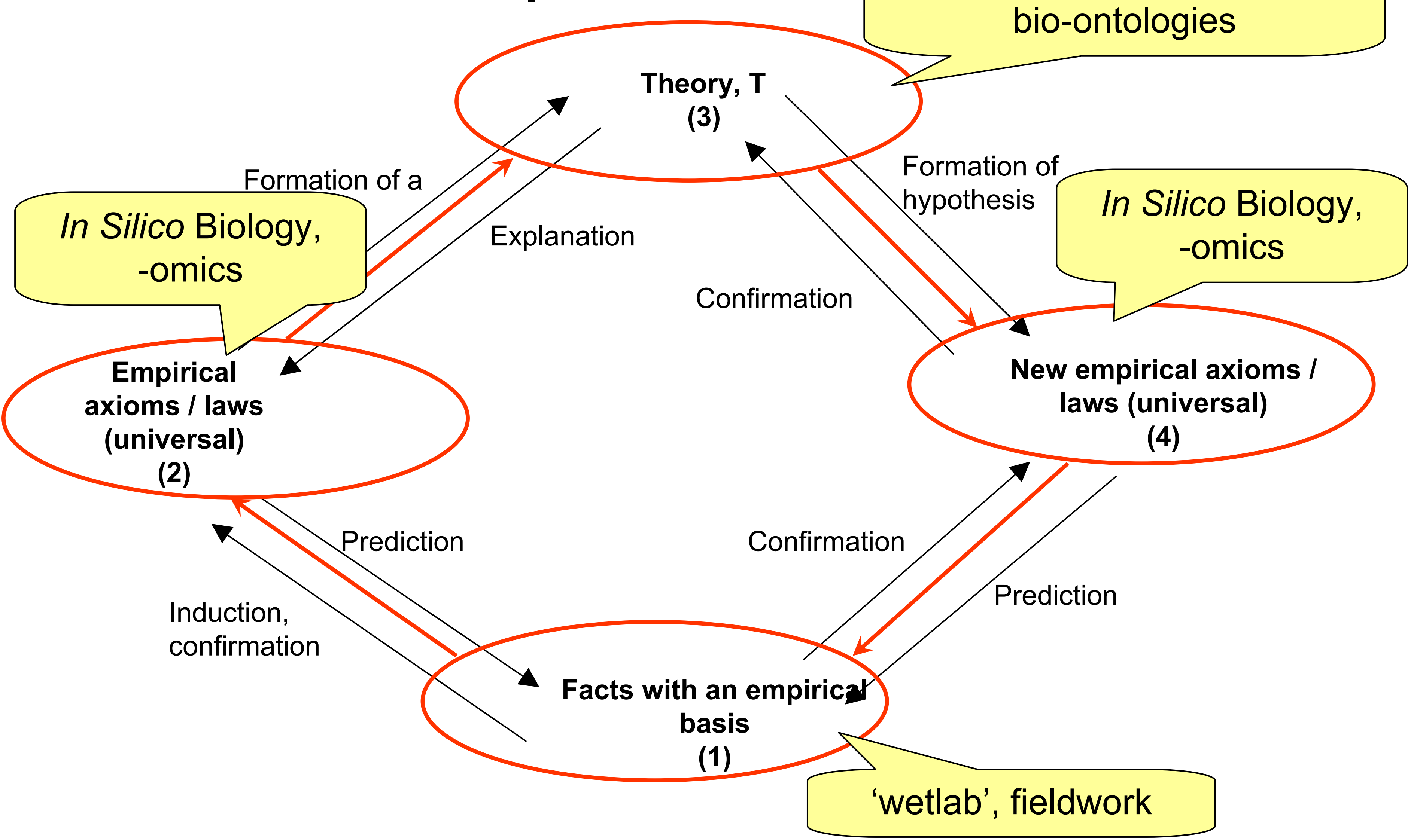
- ± ≥ mid/late '90s: sequence comparison algorithms, more databases, database integration (DWH, FD), software tools
- 1998: start Gene Ontology Consortium
- ± ≥ '00s: bio-ontologies, computational linguistics (both NLP and NLG), data mining and pattern finding, Grids, workflow, 'hypothesis generation'
- bioinformatics: from bio to molbio to biomed (OBO vs. OBO). Counter-action: ecoinformatics
- systems biology, metagenomics



Historical aspects

A few terms

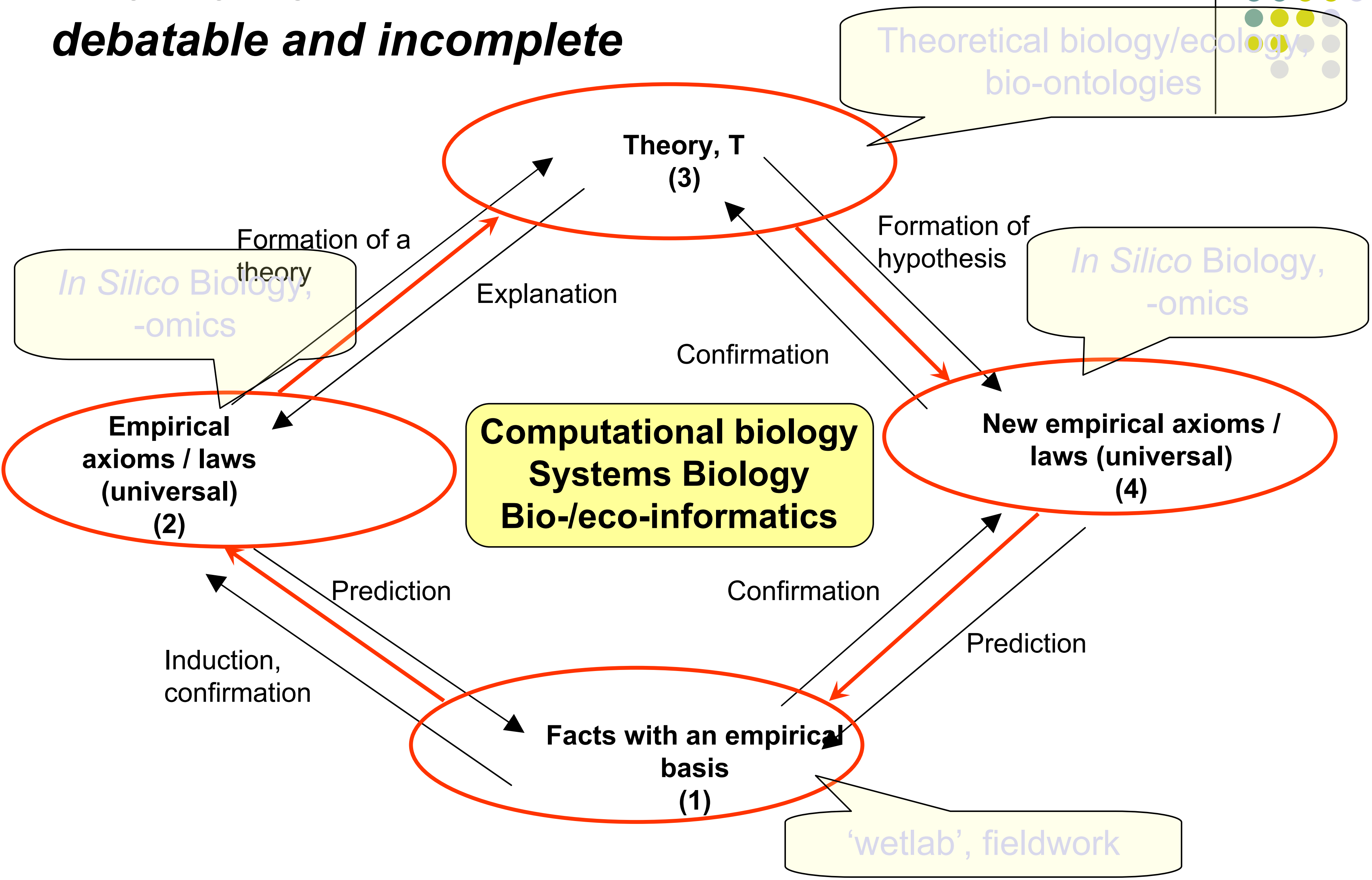
debatable and incomplete



Historical aspects

A few terms

debatable and incomplete



Overview

- Introduction
- Historical aspects
- **Characteristics**
- Brief introduction seminar series programme

Characteristics

- (some of the) CS/IT disciplines:

  Hardware/Firmware: robotics, grid computing & supercomputing, analyzers

  Software: Software engineering (devel., programming), Workflows, Databases (CM, DB devel., integration, temporal DB), distributed processing, Graphics & visualisation, HCI, comp. linguistics, neural networks, KBS, data mining

Characteristics

- (some of the) CS/IT disciplines:

  Hardware/Firmware: robotics, grid computing & supercomputing, analyzers

  Software: Software engineering (devel., programming), Workflows, Databases (CM, DB devel., integration, temporal DB), distributed processing, Graphics & visualisation, HCI, comp. linguistics, neural networks, KBS, data mining

- (some of the) Life sciences disciplines:

  Core sciences:

  Biology: virology, microbiology (bacteriology, fungi), plant sciences, animal sciences (entomology, nematology, ornithology, anatomy, ethology…), taxonomy

  Molecular Biology & Chemistry: biochemistry, enzymology, cell physiology, genomics, proteomics, metabolomics, system biology

  Ecology & environmental sciences: theoretical and experimental ecology (trophic levels, nutrient cycles, the niche), climatology, system biology

  Applied Sciences: biomedicine, agriculture, food science

Characteristics

Combine any you like from the two boxes ☺

robotics, grid computing & supercomputing, analyzers, software engineering (devel., programming), Workflows, Databases (CM, DB devel., integration, temporal DB), distributed processing, Graphics & visualisation, HCI, comp. linguistics, neural networks, KBS, data mining

virology, microbiology (bacteriology, fungi), plant sciences, animal sciences (entomology, nematology, ornithology, anatomy, ethology, …), taxonomy, biochemistry, enzymology, cell physiology, genomics, proteomics, metabolomics, system biology, theoretical and experimental ecology (trophic levels, nutrient cycles, the niche), climatology, system biology, biomedicine, agriculture, food science

Characteristics

- Is it *inter*disciplinary?

interdisciplinary teams vs. interdisciplinary **people**

"An interdisciplinary team is a committee in which members identify themselves as an expert in something else besides the actual scientific problem at hand, and abdicate responsibility for the majority of the work because it's not their field." [Eddy05]

Characteristics

…. Then, what you get:

- ***Many*** **bio-databases**: topical (one or two granularity levels, GOLD, HGVBase), "species" specific (FlyBase, AceDB), context (Bad Bug Book), primary source vs. boutique DBs (TIGR), ... see also [Galperin05]
- ***Many single-issue software tools***, mainly for data analysis of DB content, visualisation, some simulation
- High degree of ***autonomy*** of development and maintenance of IT tools, most tools poorly maintained, some databases better maintained than others
- ***Sub-optimal data management***: data is ill structured, reliability of data becoming an issue, data redundancy, data incompatibility (between DBs and between DB and analysis tools)
- Results reported in many different kinds of journals and conferences: ***where & how to find the right info?***
- Education did not keep up (biologist who can program a bit, computer scientist who googles for bio info)
- Differences in paradigms, research approaches, cultures, between science – engineering:
    - CS/IT-based (the motivating/toy example) vs. bio-based (fixes/addressed *this* issue/little problem)
    - Theory - Experiment
    - Mainly technology push
- Difference in use of knowledge (e.g. hierarchical, object-oriented etc vs. associative knowledge) and 'nature of the knowledge' (certainty vs. good-enough, conjectures, change)
- Everybody (claims to) develop(s) the ultimate best solution. There is a lot of stuff, but a) one can't see the forest for the trees on what is there and b) ***there is an awakening realisation that the bioinformatics tools at present don't do quite what biologists had in mind the tools would do in helping them to do their research more effectively***. In addition: there are ***moving targets***.
- The "jumping on the bandwagon" feeling is very present, there's a lot of money going around in bioinformatics, lots of over-optimistic promises, ethical side largely ignored.

Note: some tools are really useful!

Characteristics

- Is it *inter*disciplinary?

interdisciplinary teams vs. interdisciplinary people

"An interdisciplinary team is a committee in which members identify themselves as an expert in something else besides the actual scientific problem at hand, and abdicate responsibility for the majority of the work because it's not their field." [Eddy05]

a) 'bridge' function, multilingual

b) apply approach/methodology *a* from discipline *x* to problem *b* from discipline *y* , and v.v. and scale up and merge

Then: "…inventing new ways to look at the world." [Eddy05]
e.g.: fixing the biologist's radio [Lazebnik02]

Characteristics

- Is it *inter*disciplinary?
- Or maybe ***ante***disciplinary?

Watson & Crick were ornithologist and physicist by training, but did molecular biology, even though the term & discipline didn't exist yet.

⇨ *Will/does/did the combination of CS, IT, and Bio result in one, or more, new disciplines?*

⇨ *Does a named set of a few characteristic activities make a (new) (sub)discipline?*

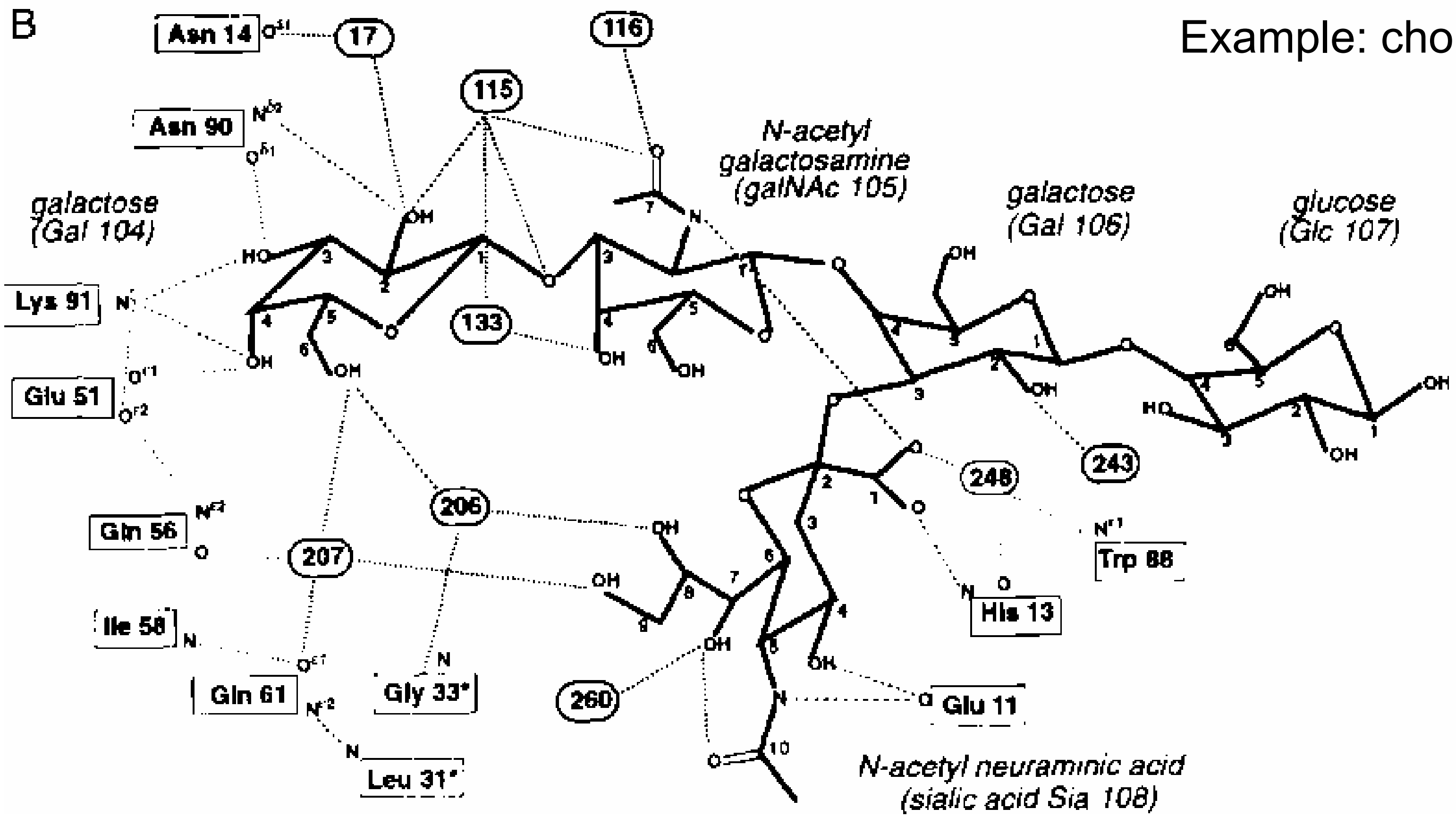
Characteristics

- Some examples of integrative approaches
  - Virtual cell [ecell] : cell physiologists, (bio)chemists, geneticist, modellers, programmers, visualisation, mathematicians

(Virtual) Cells (& cell components) example

- Basic options:- Schematic & measurements

Example: cholera toxin 1994



B: Schematic representation of hydrogen bonding interactions involving the G_{MI} pentasaccharide in subunit B#5 of the pentamer. The peptide residues shown belong to a single monomer, except for the involvement of Gly **33** from an adjacent monomer. Solvent molecules are depicted as ovals. Electron density at the other **4** copies of the pentamer is insufficient to model the full pentasaccharide. In subunit B#5, however, the conformation of the glucose end of the pentasaccharide is stabilized by hydrogen bonds (not shown) to a separate molecule related by the crystallographic symmetry operation ***(x** + 1/2, *y* + 1/2, ***E)* .** [meritt94]

(Virtual) Cells

(& cell components)

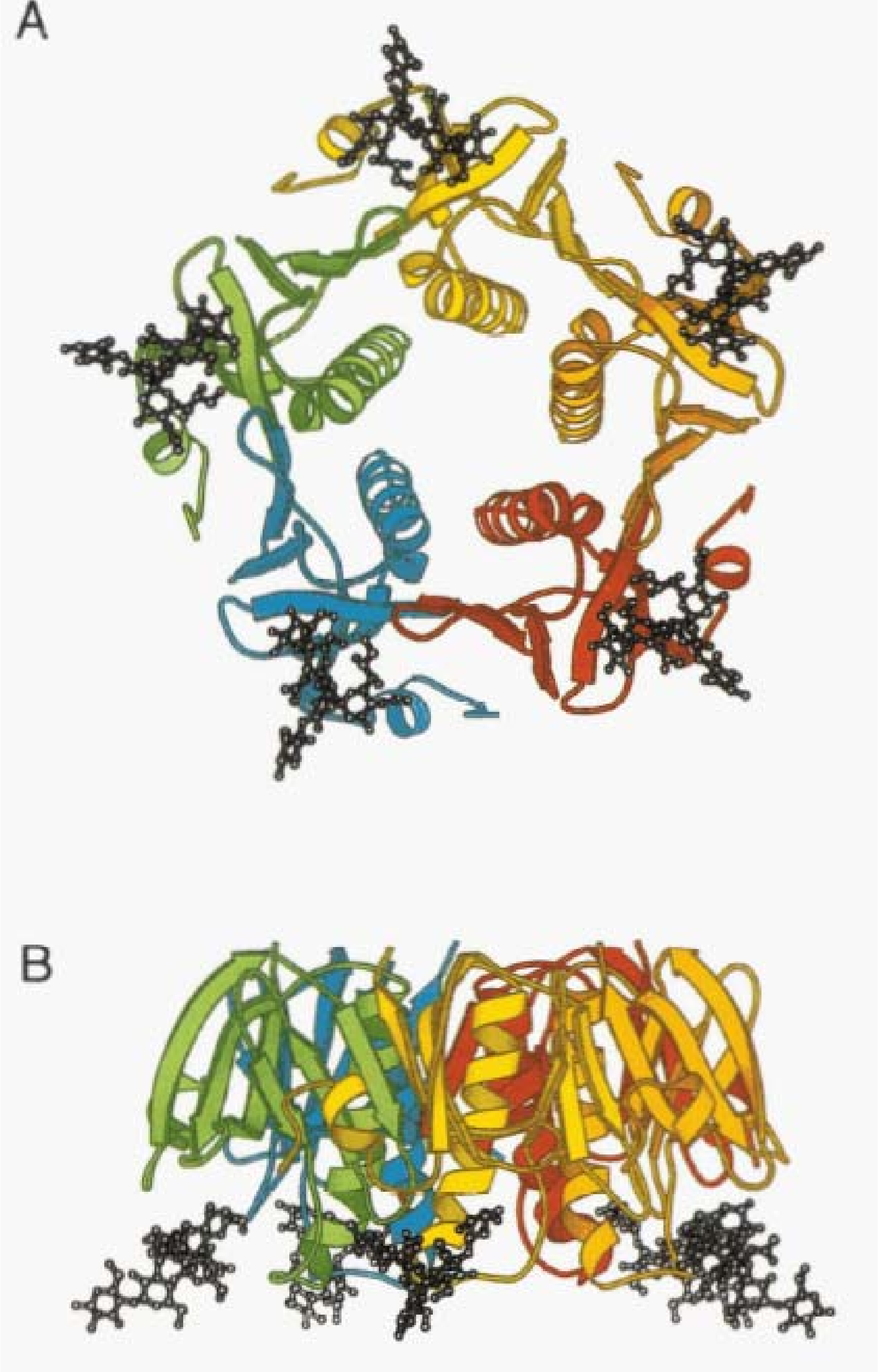
example

- Basic options

more cholera toxin:

(3D) structure of the protein

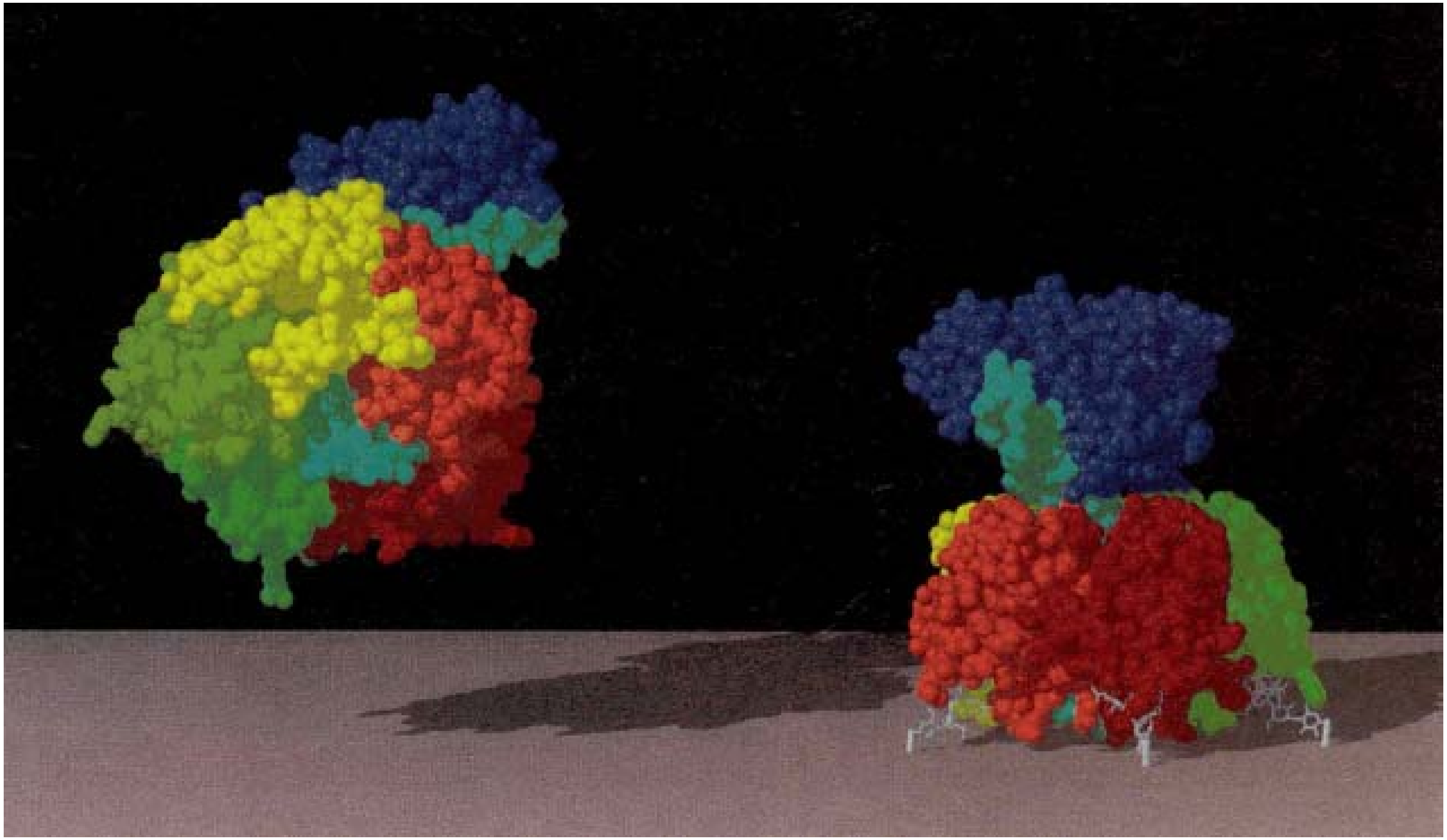
[merritt94]



(Virtual) Cells (& cell components) example

- Basic options: even more cholera toxin, now "in action"

the B-pentamer of the toxin binds specifically to the branched pentasaccharide moiety of ganglioside $\mathbf{G_{MI}}$, on the surface of target human intestinal epithelial cells. [meritt94]



(Virtual) Cells example

- Basic options:

  - Schematic e.g. choleratoxin 1994 [meritt94]

- Progress over the past decade, with *new*, *more*, and *better*:

  - Simulations for education

    "fig 9."http://cellix.imolbio.oeaw.ac.at/Videotour/video_tour_5.html

    "Translation: the movie" http://vcell.ndsu.nodak.edu/~christjo/vcell/animationSite/index.htm

  - Simulations with data

    schematic & measurements

  - Videos of real things at (sub)cellular level

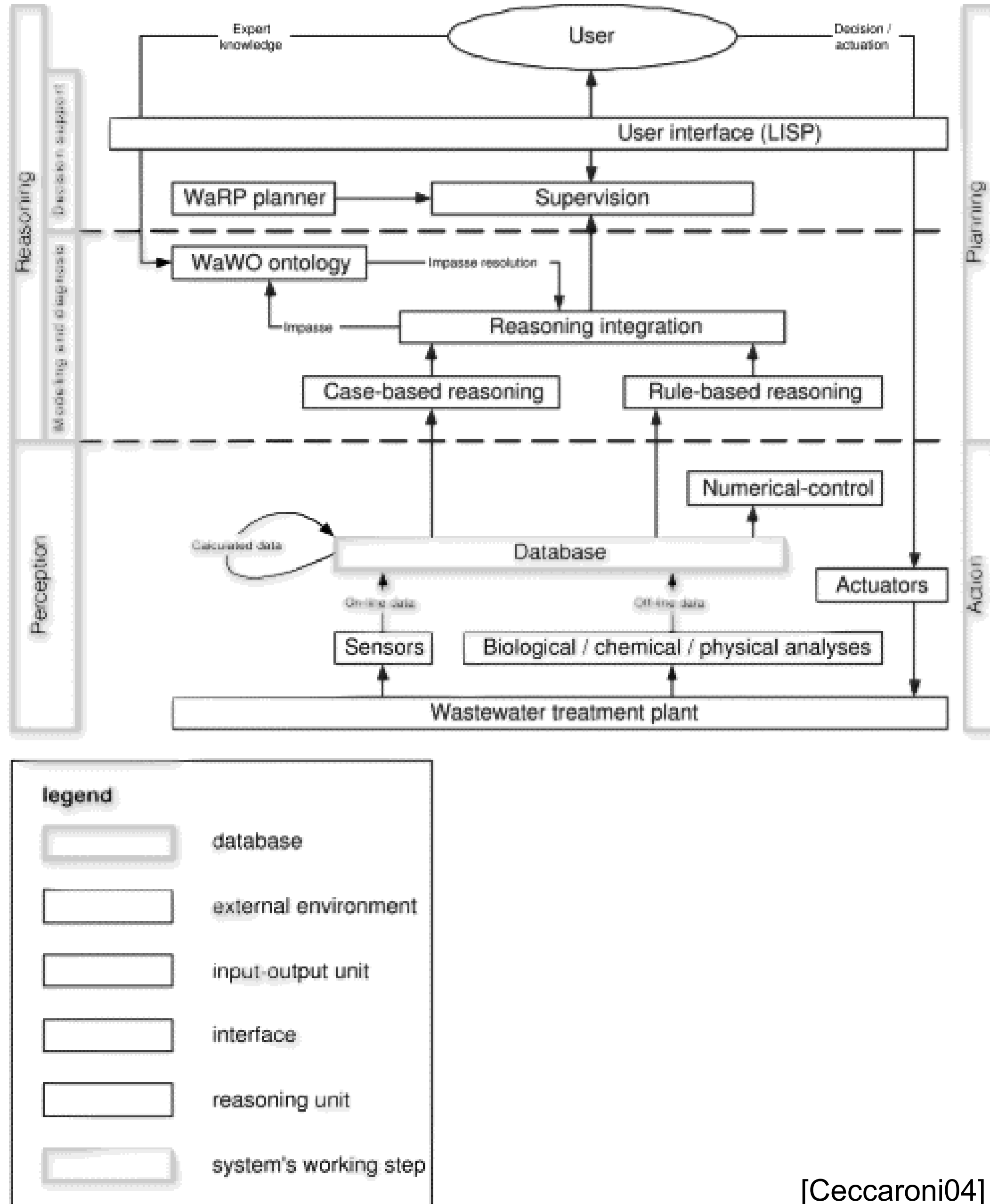
    see yourself: http://cellix.imolbio.oeaw.ac.at/Videotour/video_tour.html

Characteristics

- Some examples of integrative approaches
  - Virtual cell [ecell] : cell physiologists, (bio)chemists, geneticist, modellers, programmers, visualisation, mathematicians
  - Sequence comparisons: geneticists, data miners, mathematicians/statisticians, visualisation, DB, Web-access
  - Applied sciences: e.g. medical / agriculture / environmental engineering. example

Environmental Engineering

- OntoWEDSS: Environmental decision support system for managing active sludge in wastewater treatment plants.



[Ceccaroni04]

Characteristics

- Some examples of integrative approaches
  - Virtual cell [ecell] : cell physiologists, (bio)chemists, geneticist, modellers, programmers, visualisation, mathematicians
  - Sequence comparisons: geneticists, data miners, mathematicians/statisticians, visualisation, DB, Web-access
  - Applied sciences
  - Ecology: example

Ecology - example

- Knowledge Network for Biocomplexity

"…network intended to facilitate ecological and environmental research on biocomplexity. For scientists, the KNB is an efficient way to discover, access, interpret, integrate and analyze complex ecological data from a highly-distributed set of field stations, laboratories, research sites, and individual researchers… focuses on research into informatics and biocomplexity, through the development of software products…" http://knb.ecoinformatics.org/index.jsp

⇨ metadata, storage resource broker, distributed data management tool, data integration, quality assurance, hypothesis modelling (Bayesian network), visualisation tools

Characteristics

- Some examples of integrative approaches
  - Virtual cell [ecell] : cell physiologists, (bio)chemists, geneticist, modellers, programmers, visualisation, mathematicians
  - Sequence comparisons: geneticists, data miners, mathematicians/statisticians, …
  - Applied sciences
  - Ecology
  - Bio-Linux http://www.biolinux.org/ , 60 software packages
    BioBrew http://bioinformatics.org/project/?group_id=273, 223 groups
  - Exercise 'Conservation efforts'

Conservation effort exercise

1. Let *X* be an endangered species and you want to conserve it.
2. You are in the fortunate position to have enough resources (researchers, technicians, money, etc) to set up a project to fulfill (1).
3. Goal:

   how to come up with the "recovery plan"?

Conservation effort exercise

- Known variables:

  Habitat of species *X*, the site(s) where the organisms live at present, population census

  Its diet and predator

  Dispersal, migration characteristics

- What we need to know:

  - What did change at the sites, and how?

  - Are there any viable other sites to (re)introduce *X*, if yes, where?

  - Is the gene pool diverse enough?

Characteristics

- Some examples of integrative approaches
  - Virtual cell [ecell]: cell physiologists, (bio)chemists, geneticist, modellers, programmers, visualisation, mathematicians
  - Sequence comparisons: geneticists, data miners, mathematicians/statisticians, visualisation, DB, Web-access
  - Medical info system/agriculture
  - Ecology
  - Bio-Linux http://www.biolinux.org/ , 60 software packages

    BioBrew http://bioinformatics.org/project/?group_id=273, 223 groups
  - Exercise 'Conservation efforts'
- **Don't forget *intra*disciplinary science**

Overview

- Introduction
- Historical aspects
- Characteristics
- Brief introduction seminar series programme

Titles & Topics

- Ethics
- Computational Systems Biology
- Virtual laboratory for integrative bioinformatics research
- A standardized and dynamic approach for immunogenetics and immunoinformatics
- Querying biological data sources
- Requirements for natural language understanding in referent-tracking based electronic health records
- Ontology design patterns and practical expertise: roles, tasks and techniques in the agricultural domain
- Ecological niche

References & related websites

- Some **Journals & Conferences**:
  - lists: http://www.lri.fr/~cohen/confSem/confSem.html , http://www.conferencealerts.com/bioinform.htm
- Some **websites & organisations**:
  - Gene Ontology Consortium: http://www.geneontology.org
  - Society for Mathematical Biology: http://www.smb.org/
  - International Society for Computational Biology: http://www.iscb.org/
  - European Molecular Biology Laboratory: http://www.embl.org/
  - European Bioinformatics Institute: http://www.ebi.ac.uk
  - European Federation for Information Technology in Agriculture, Food and the Environment: http://www.efita.net/
  - International Medical Informatics Association: http://www.imia.org
- Some **Projects**:
  - Marine genomics: http://www.marine-genomics-europe.org
  - Science Environment for Ecological Knowledge: http://seek.ecoinformatics.org
  - Semantic Mining in Biomedicine: http://www.semanticmining.org
  - Biopattern: http://www.biopattern.org/index.html
  - InfoBioMed: http//www.infobiomed.org
  - EurEthNet: http://www.eureth.net

- Referenced literature:

  - [Ceccaroni04] Ceccaroni, L., Cortés, U. and Sànchez-Marrè, M. OntoWEDSS: augmenting environmental decision-support systems with ontologies. *Environmental Modelling & Software*, 2004, 19 (9): 785-797.
  - [ecell] E-cell project. http://www.ndsu.nodak.edu/instruct/mcclean/vc/wwwic-vc5.htm (there are many more projects)
  - [Eddy05] Eddy, S.R. “Antedisciplinary” Science. *PLoS Computational Biology*, 2005, 1(1): e6. http://compbiol.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pcbi.0010006
  - [Galperin05] Galperin, M.Y. The Molecular Biology Database Collection: 2005 update. *Nucleic Acids Research*, 2005, 33: D5-D24. http://nar.oxfordjournals.org/cgi/content/full/33/suppl_1/D5
  - [Lazebnik05] Lazebnik, Y. Can a Biologist Fix a Radio? – or, What I Learned while Studying Apoptosis. *Cancer Cell*, 2002, 2: 179-182. http://www.protein.bio.msu.ru/biokhimiya/contents/v69/pdf/bcm_1403.pdf
  - [Merritt94] Merritt, E.A., Sarfaty, S., Akker, F. van den, L‘Hoir, C., Martial, J.A., Hol, W.G.J. Crystal structure of cholera toxin B-pentamer bound to receptor G_{M1}pentasaccharide. *Protein Science*, 1994, 3: 166-175.