Jozef Stefan Institute, Ljubljana, Slovenia
December 2, 2010

Towards an Ontology of Science

Larisa N. Soldatova, PhD

RC UK Fellow
Aberystwyth University, UK

The previous talk:

- Solomonovi seminarji, April 2008
- Formalisation of Science: Ontology Based Projects in Aberystwyth University
- EXPO, EXACT, ART, LABORS, DDI
- 116 views
- http://videolectures.net/larisa_soldatova/

Plan of the talk:

1.Introduction into ontologies.

2.Updates on ontology projects:

- LABORS and DDI projects.
- OBO, OBI
- OntoDM, work led by JSI

4. Research plans:

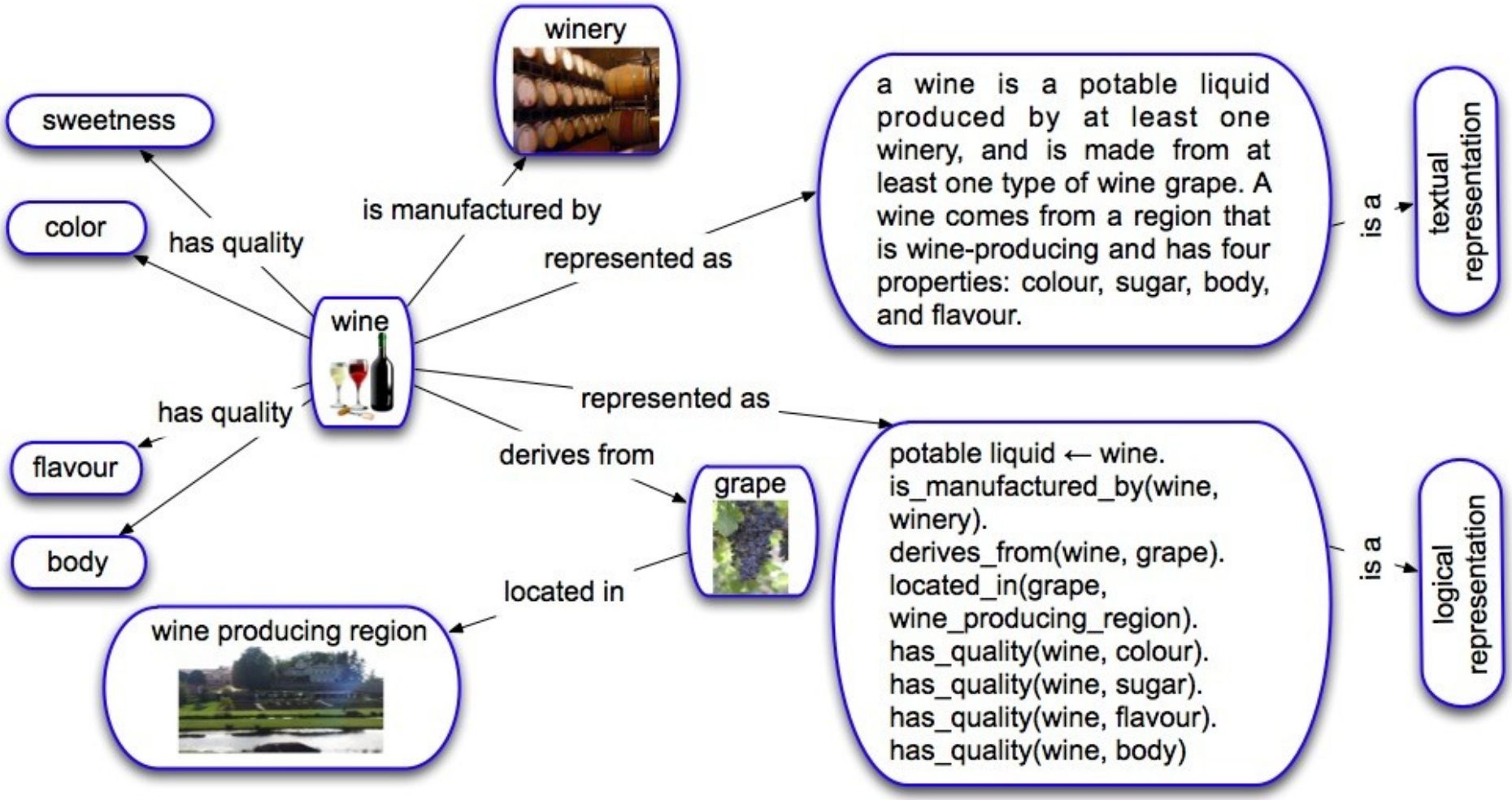
- QSAR modeling
- SBO next generation
- ontological support to automated dynamic modeling

5. Components and applications of an Ontology of Science.

1.Introduction into ontologies

- Ontologies are knowledge representations, theories, logical models
- Ontologies are a key component of Semantic Web.
- Adding semantics to data, services, knowledge.
- Information has machine-processable and machine-understandable semantics.
- Adding logic to RDF (Resource description Framework).

A toy ontology



Example statements

RDF: merlot *has-sweetness* 3

riesling *has-sweetness* 1

riesling *has-sweetness* 5

riesling *grown-in* germany

OWL: riesling *is-a* wine

riesling_grape *is-a* grape

wine *has-sweetness* sweetness

has-sweetness is a functional relation

3 *is-instance-of* sweetness

Inference: what is the best wine to accompany this dish?

(1) area of research interest

A paradigm shift:

- from relational to RDF DB
- from propositional to graph-based data

Companies are investing into data in RDF

A growing demand for new DM algorithms

- how to describe properties of graph-based datasets? (the number of examples, ratio features/numbers(?)...)
- how to select the best algorithms?
- how to evaluate output models?

2. Updates on ontology projects

LABORS for RS

- LABoratory Ontology for Robot Scientists
- The Robot Scientist project:
  www.aber.ac.uk/en/cs/research/cb/projects/robotscienti
- Invited talk at the Third International Workshop on Machine Learning in Systems Biology

  Ross D. King, Automating Science

  117 views, http://videolectures.net/ross_d_king/

LABORS

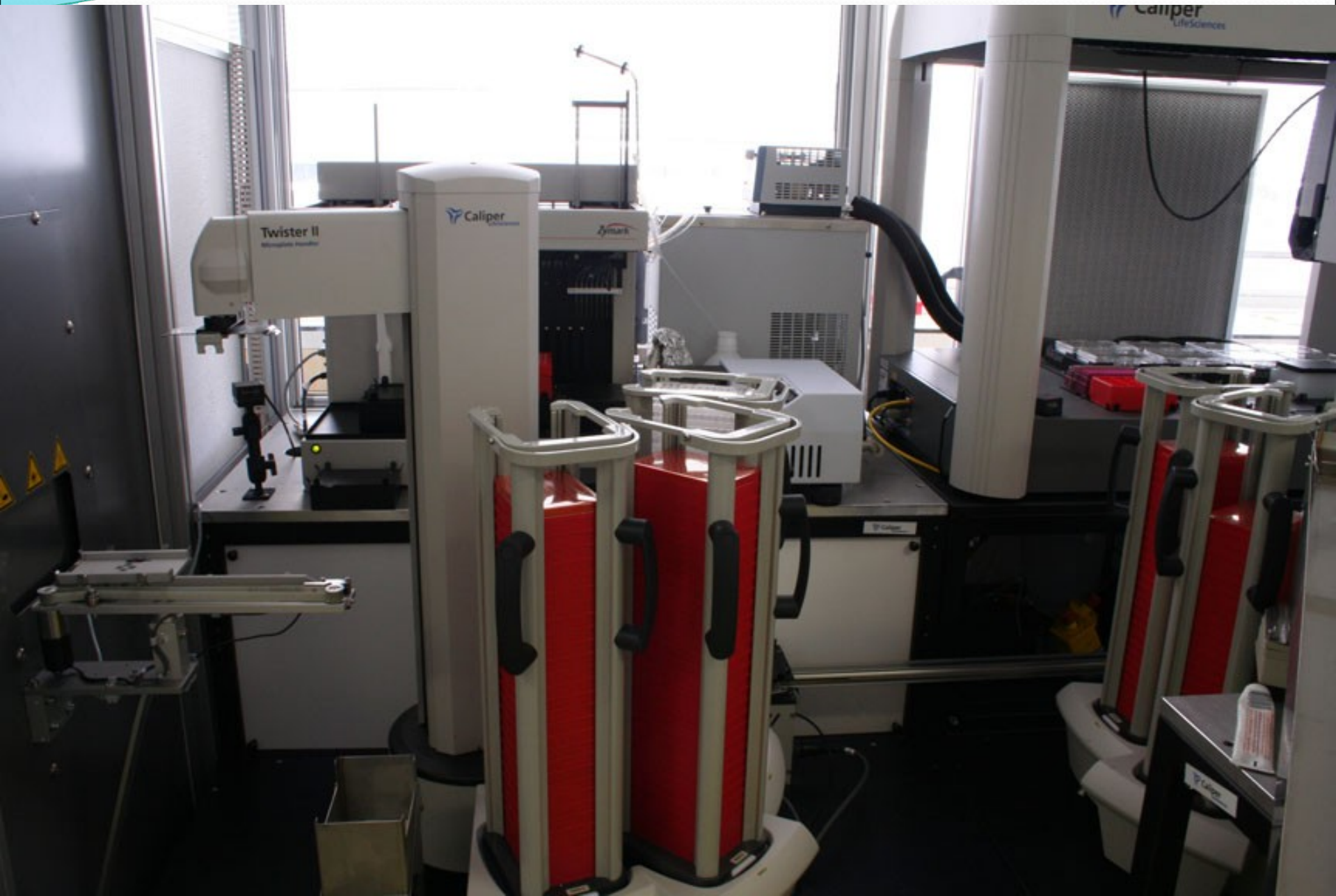
- Formal description of the entities involved in Robot Scientist experiments.
- Identification of metadata essential for the experiment's description and repeatability.
- Modelling a database for the storage of experimental data and track experiment execution.
- Customised version of EXPO.

*King et.al.(2009) The Automation of Science. Science **324**: 85-89.*

*Soldatova L.N., Clare A., Sparkes A. and King, R.D. (2006) An ontology for a Robot Scientist. Bioinformatics (Special issue ISMB) **22/14**: e464-e471.*

*Soldatova, LN & King, RD. (2006) An Ontology of Scientific Experiments. Journal of the Royal Society Interface **3/11**: 795-803.*

A Robot Scientist Adam (A Discovery MAchine)



The RS project

> 500 articles,

TV - BBC, CNN

Radio –

BBC Scotland

BBC Wales,

Radio 1,3,4


TIME

SEARCH TIME.COM

Specials

Special Reports * Best Websites * Worst Cars * Time 100 * Time 100 Roundtable * Best TV Shows * Top 10 * ALL Time 100

The Top 10 Everything of 2009

TIME charts the highs and lows of the past year in 50 wide-ranging lists

Select a Section | Story | All Best and Worst Lists

Top 10 Scientific Discoveries

BACK NEXT

4. A Robot Performs Science

374 of 500 | View All

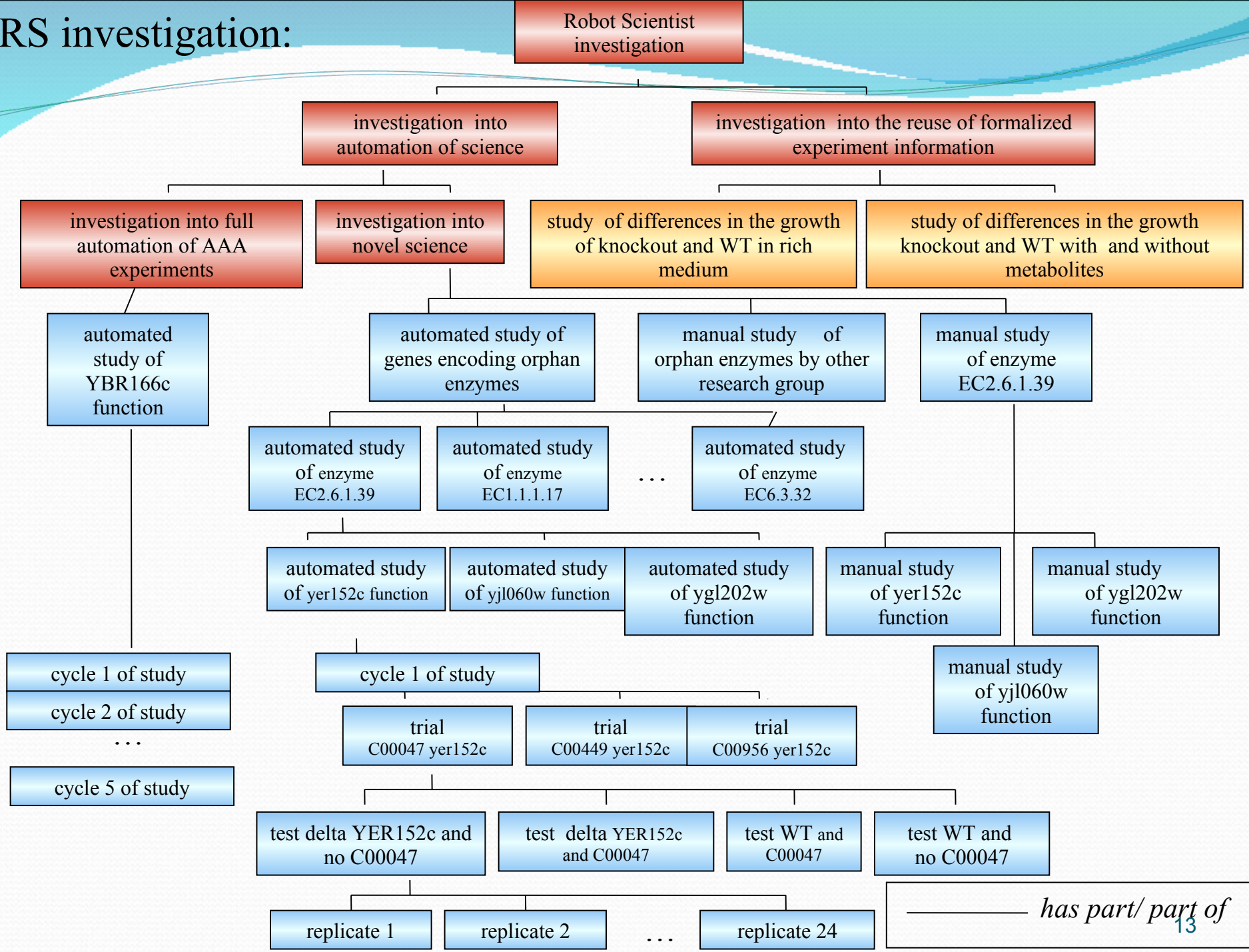
By EBEN HARRELL Tuesday, Dec. 08, 2009

By any standard, it was an elementary discovery — the identification of the role of about a dozen genes in a yeast cell. But what made this finding a major breakthrough was the unlikely form of th scientist: a robot. In April, "Adam," a machine designed at Aberystwyth University in Wales, became the first robotic system to make a novel scientific discovery with virtually no human intellectual input. Robots have long been used in experiments — their vast computational power assisted in the sequencing of the human genome, for example — but Adam was the first to complete the cycle from hypothesis to experiment to reformulated hypothesis without human intervention. Interviewed after Adam's experiment appeared in *Science*, inventor Ross Kin argued that artificial intelligence had almost limitless scientific potential — and that a computer would one day make a discovery akin to Einstein's special theory of relativity. "There isn't any intrinsic reason why that wouldn't happen," he said. "A computer can mak beautiful chess moves, but it's not doing anything special. In my view, that's what's going t

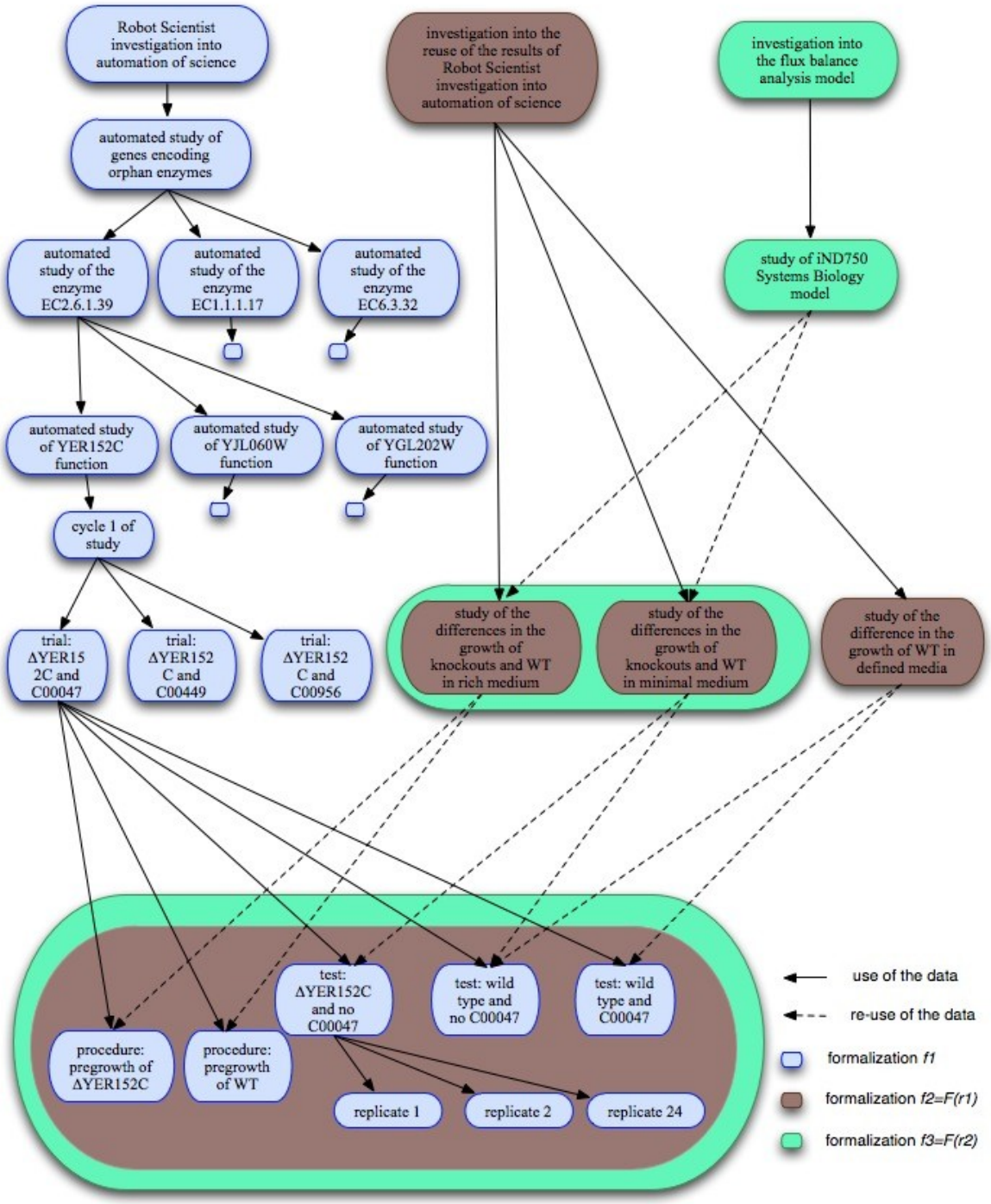
King *et.al.(*2009) The Automation of Science. *Science* ***324****: 85-89.*

King, *et al* (2009) Make Way for Robot Scientists. *Science* ***325****: 945*

RS investigation:



Data re-use


Robot Scientist investigation into automation of science
investigation into the reuse of the results of Robot Scientist investigation into automation of science
investigation into the flux balance analysis model
automated study of genes encoding orphan enzymes
automated study of the enzyme EC2.6.1.39
automated study of the enzyme EC1.1.1.17
automated study of the enzyme EC6.3.32
study of iND750 Systems Biology model
automated study of YER152C function
automated study of YJL060W function
automated study of YGL202W function
cycle 1 of study
trial: ΔYER15 2C and C00047
trial: ΔYER152 C and C00449
trial: ΔYER152 C and C00956
study of the differences in the growth of knockouts and WT in rich medium
study of the differences in the growth of knockouts and WT in minimal medium
study of the difference in the growth of WT in defined media
test: ΔYER152C and no C00047
test: wild type and no C00047
test: wild type and C00047
procedure: pregrowth of ΔYER152C
procedure: pregrowth of WT
replicate 1
replicate 2
replicate 24
use of the data
re-use of the data
formalization f1
formalization f2=F(r1)
formalization f3=F(r2)

Representation of research hypotheses (H)

- The heart of scientific research.
- H as textual entities:

  SWAN, swan.mindinformatics.org

  OBI, obi-ontology.org
- H in logic:

  HyBrow, www.hybrow.org / HyQue

  The Large Scale Discovery of Scientific Hypotheses

  http://arrafunding.uchicago.edu/investigators/rzhetsky_a.shtml
- A need to express precise semantic meaning of H to

  generate H

  test H

  reason with H

Computed research hypotheses

- automated investigations.
- machine learning programs (based on induction) are used to help design drugs;
- the annotation of a genome is essentially a large set of (based on abduction) hypotheses generated by sequence similarity programs;

Automatic generation of hypotheses

1)Machine processable representation of the domain knowledge.

2)Inference of hypotheses.

3) Selection of hypotheses.

4) Deduction of experimental consequences.

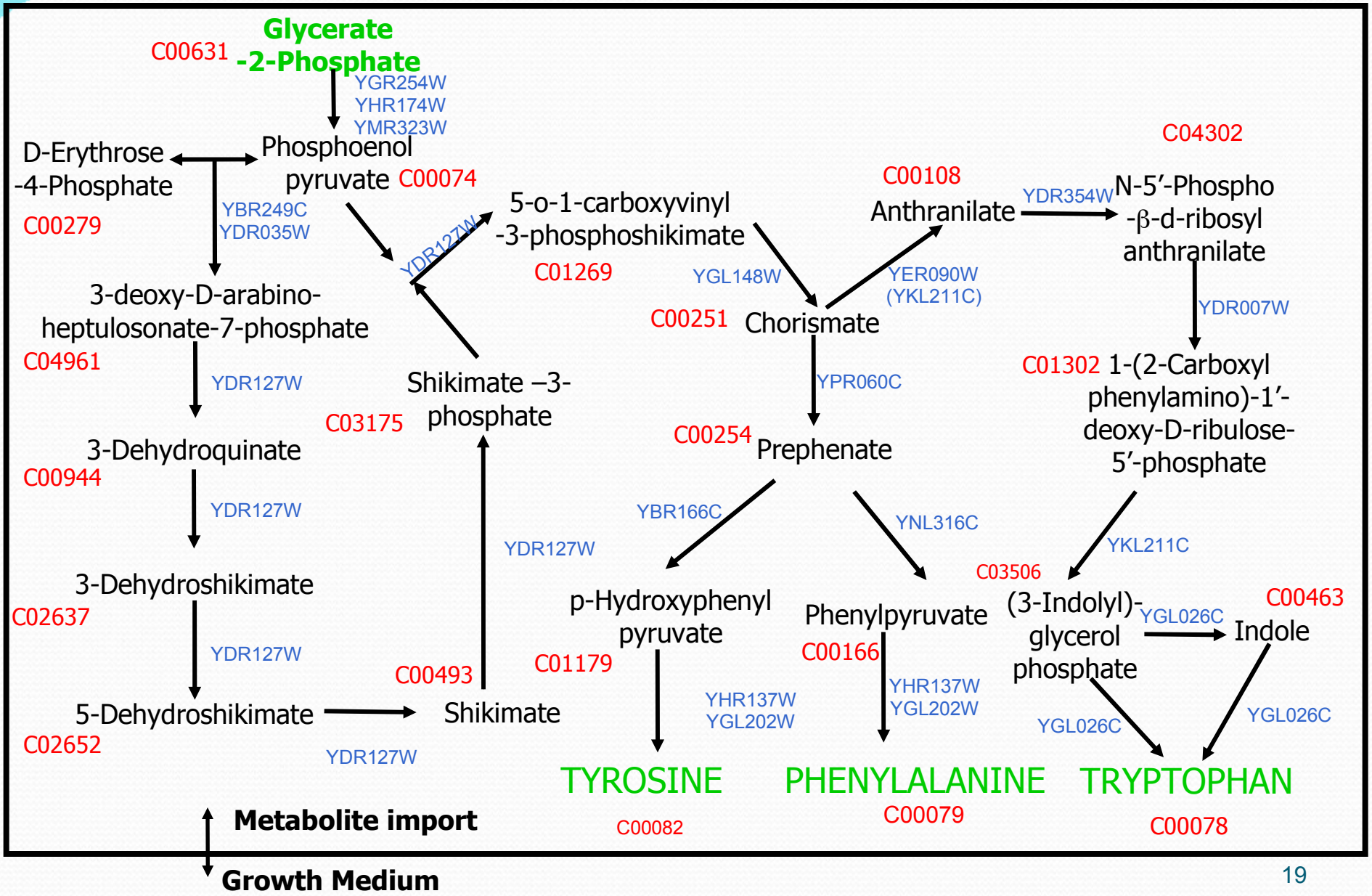
(1) Machine processable representation of the domain knowledge: Yeast metabolic model

- Our group has developed a logical formalism for modelling metabolic pathways (encoded in Prolog)*. This is essentially a directed graph: with metabolites as nodes and enzymes as arcs.

- If a path can be found from cell inputs (metabolites in the growth medium) to all the cell outputs (essential compounds), then the cell can grow.

** Whelan, K. E. and King, R. D. (2008) Using a logical model to predict the growth of yeast. BMC Bioinformatics 2008, 9:97*

Phenyalanine, Tyrosine, and Tryptophan Pathways for S. cerivisae



(2) Inference of hypotheses

- Abductive Logic Programming for the inference of missing arcs/labels in our metabolic graph.

- It is not possible to find a path from the inputs (growth medium) to all the end-point metabolites using only reactions encoded by known genes.

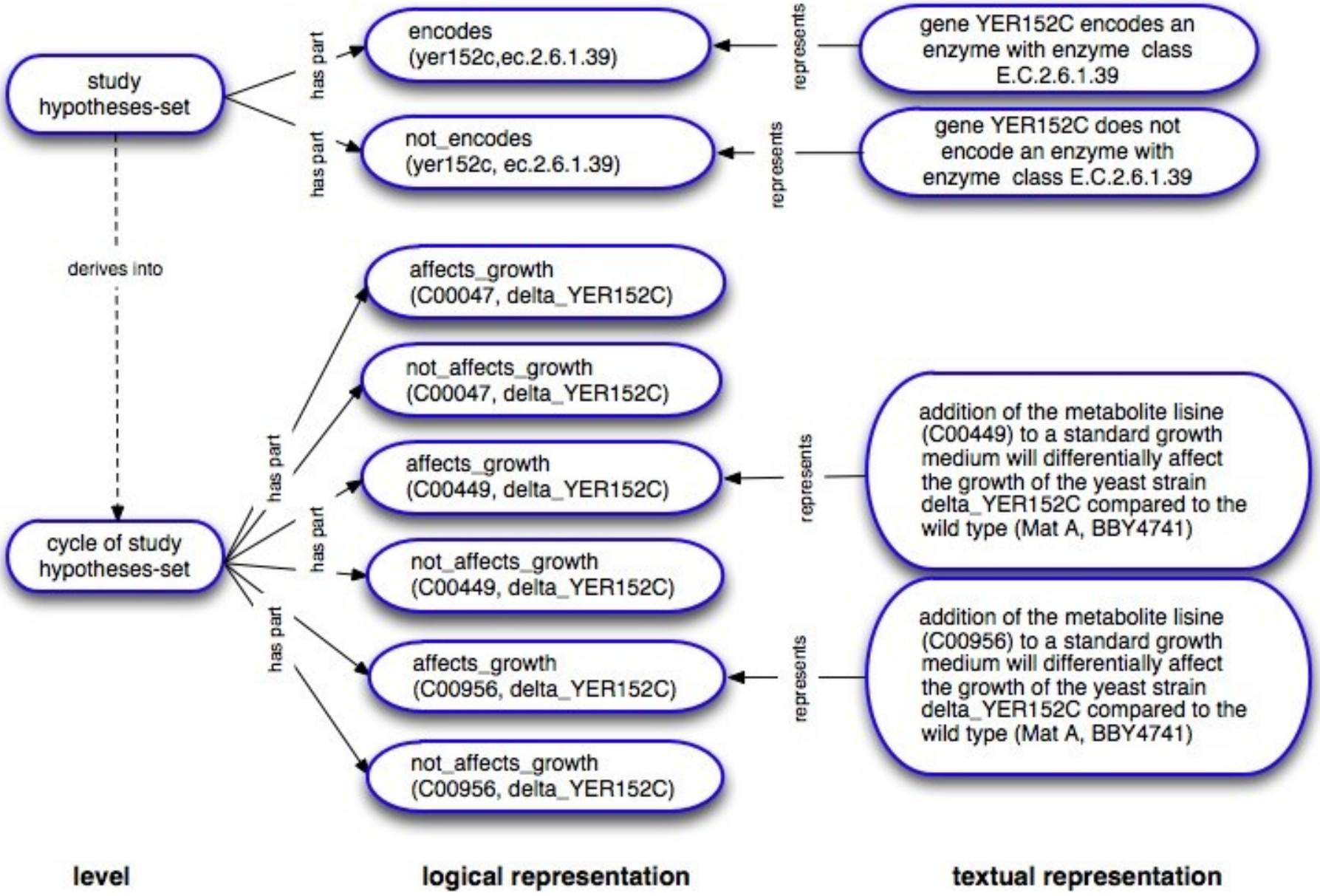
- Automated strategy, based on using EC enzyme class of missing reactions, is to identify genes that code for this EC class in other organism, then find homologous genes in yeast.

(3) Selection of a set of the hypotheses

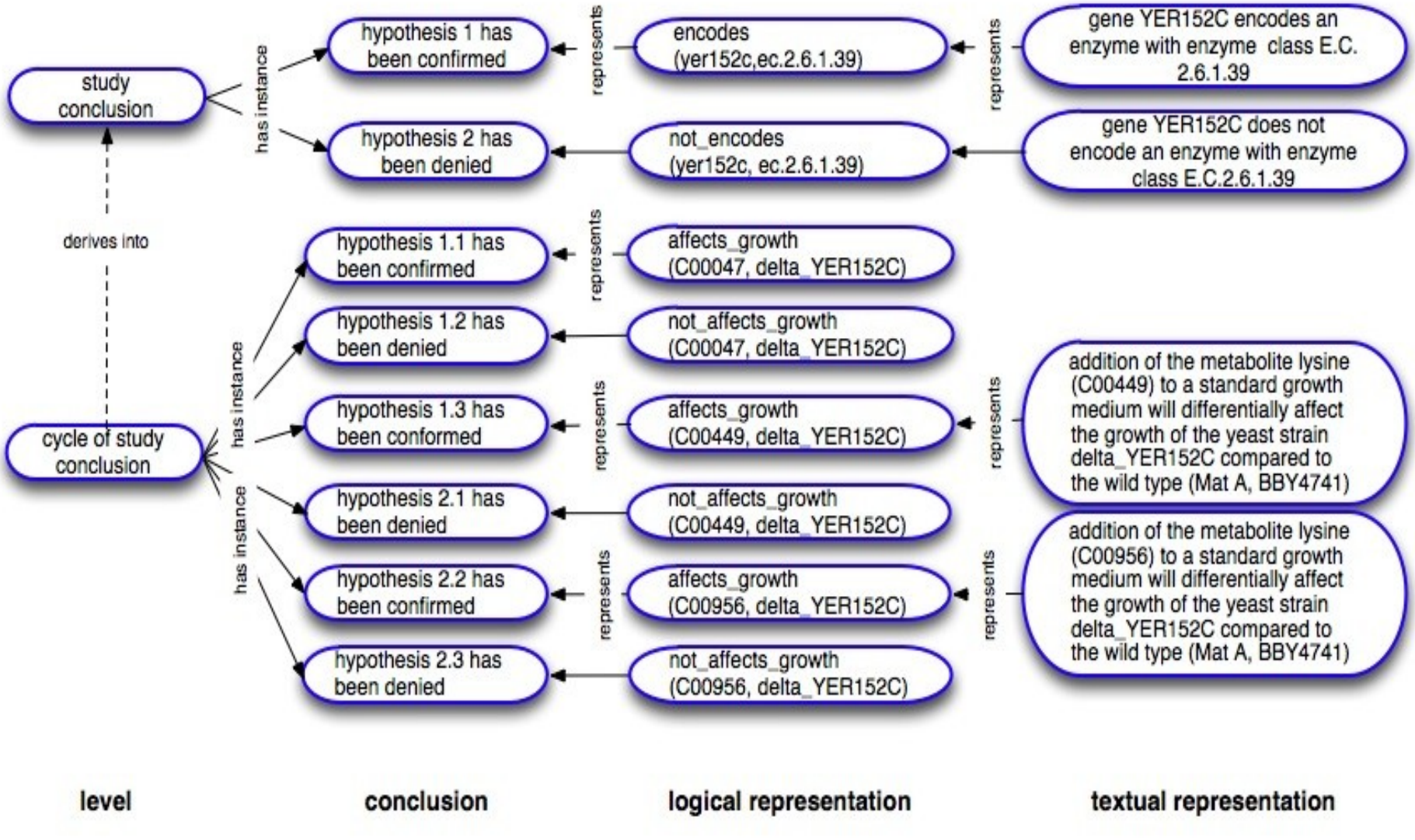
- max probability;
- max information;
- min time;
- min cost.

(4) Deduction of experimental consequences

Specification of hypotheses:



Representation of conclusions



Cyclic investigations

- Each cycle of investigation has a specified input *hypotheses set*.
- RS designs and run experiments to test each hypothesis from the set.
- RS analyses the results of the experiments and makes conclusions about whether a particular hypothesis has been confirmed or rejected.
- The rejected hypotheses are eliminated from the input *hypotheses set* and the remaining set of hypotheses are considered as a specified output of the current cycle of the study.
- RS updates its current domain model and generates a new set of hypotheses, where the rejected on previous cycles hypotheses are excluded.
- This set is considered as a specified input *hypotheses set* for the next cycle of the study.
- RS continues to run cycles of studies until the *hypotheses set* is empty or the robot runs out of specified resources.

Discussion: how to record research hypotheses

- Explicitly;
- Constructively;
- Systematically;
- Statistically significantly;
- Record negative hypotheses.

Soldatova,*et al* (2011) Representation of research hypotheses. *J. of Biomedical Semantics* (invited).

An ontology for the description of Drug Discovery Investigations*

- DDI is an application ontology for OBI (Ontology for Biomedical Investigations.
- DDI complies with Open Biomedical Ontologies (OBO) Foundry principle.
- Relations in DDI are from OBO Relation Ontology (RO).
- DDI complies with other OBO ontologies, such as ChEBI (Chemical entities of biological interest).
- DDI complies with Basic Formal Ontology (BFO).

**Qi, D., Ross D. King, R.D., Hopkins, A., Bickerton, R., Soldatova, L.N.(2010) An Ontology of Description of Drug Design Investigations. Journal of Integrative Bioinformatics*

OBO Foundry

- ~ 80 open bio-medical ontologies (OBO)

http://www.obofoundry.org/

- OBO F is a legal entity
- Principles:

- Using BFO top classes, RO relations
- Orthogonal
- In OWL or OBO

.....

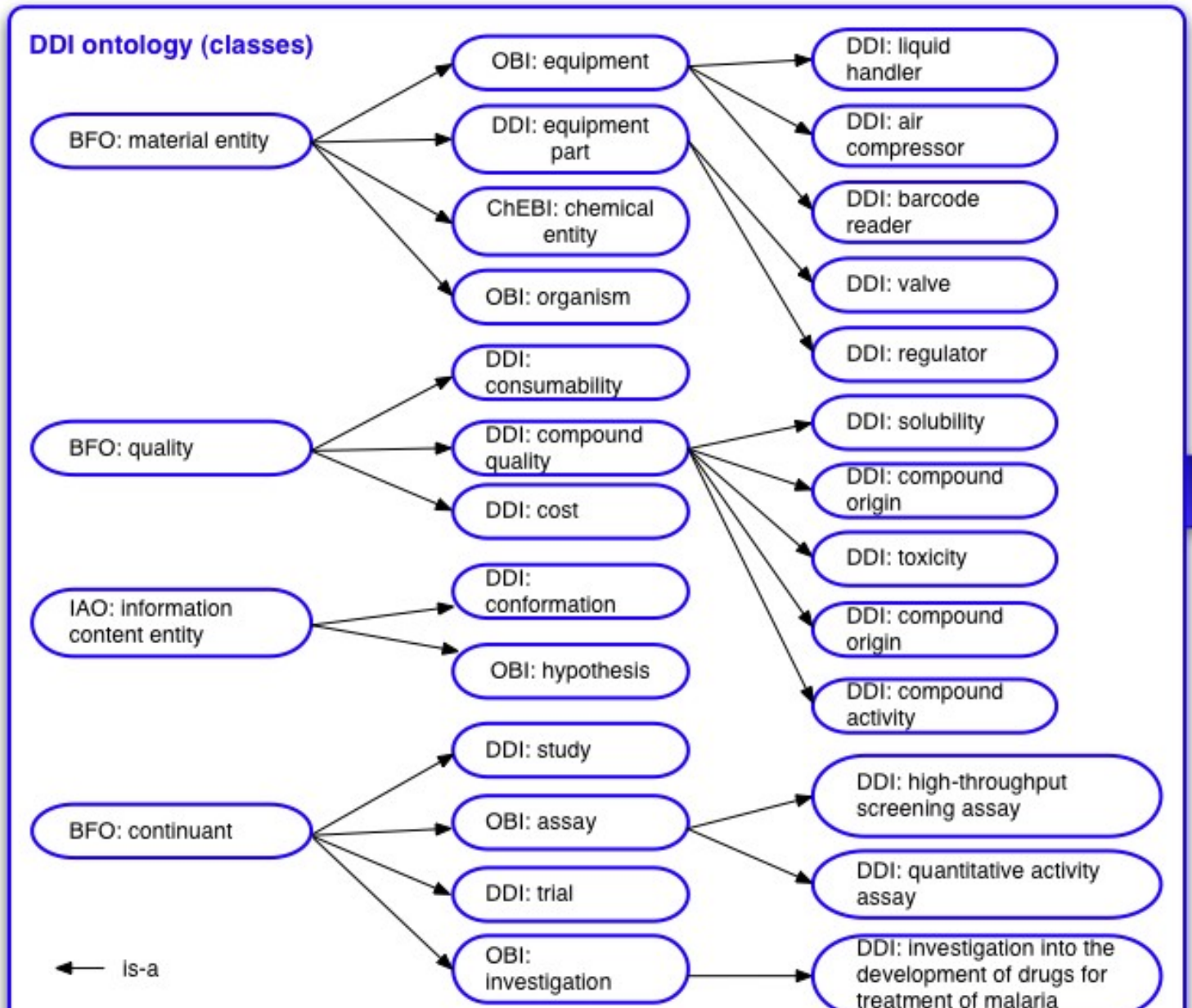
- Candidate ontologies and 8 [6] foundry ontologies

A Robot Scientist for drug screening and design Eve

What Eve offers to DD

- Advanced Artificial Intelligence.
- Innovative data mining (active learning, relational representation).
- More informative guesses about compound activity.

DDI


DDI ontology (classes)
BFO: material entity
BFO: quality
IAO: information content entity
BFO: continuant
OBI: equipment
DDI: equipment part
ChEBI: chemical entity
OBI: organism
DDI: consumability
DDI: compound quality
DDI: cost
DDI: conformation
OBI: hypothesis
DDI: study
OBI: assay
DDI: trial
OBI: investigation
DDI: liquid handler
DDI: air compressor
DDI: barcode reader
DDI: valve
DDI: regulator
DDI: solubility
DDI: compound origin
DDI: toxicity
DDI: compound origin
DDI: compound activity
DDI: high-throughput screening assay
DDI: quantitative activity assay
DDI: investigation into the development of drugs for treatment of malaria
is-a


Database (instances of DDI classes)

Problems in text mining

- Only 70% of NE, and 30% of relations
- Lack of semantic definitions in DD terms
- OBO failed to provide clear semantic definitions for:

<u>enzymes:</u>

OBI -a material entity, SBO – functional entity, TMO – a protein

do not define links to other terms, part from *has-function* catalysis

<u>kinases:</u> no term

DDI-based text mining: enzyme

enzyme *is-a* material entity

enzyme *is-a* protein

enzyme *has-function* catalytic activity [GO:0003824]

enzyme *participates-in* chemical reaction

enzyme *participates-in* some biological process

metabolomic process *is-a* biological process

signalling processes *is-a* biological process

enzyme *bearer-of* efficiency

enzyme *has-role* some drug target

enzyme *participates-in* enzyme assay

……

DDI-based text mining: kinase

kinase *is-a* protein

kinase *is-a* enzyme

kinase *participates-in* kinase assay

kinase assay *realises* assay design

kinase assay *has-specified-output* IC50

kinase assay *has-specified-output* EC50

kinase assay *has-specified-output* kinase activity value

kinase assay *has-specified-output* Ki

kinase has-function *kinase* activity

kinase *is-realised-by* phosphorylation

phosphate group *participates-in* phosphorilation

phosphate group *located-in* binding site

kinase *has-quality* selectivity ….

(2) area of research interest

ontological support for text mining:

- Full power of logical representations
- Innovative DM algorithms

ontology learning from texts, data

- Innovative DM algorithms

OntoDM

- A proposal for a general ontology of data mining (DM)*.
- Based on a general DM framework.**
- Defines the key entities of DM, i.e. DM task, DM algorithm, generalisation, data, dataset, evaluation.

* Panov, P., Soldatova, L., Dzeroski, S., (2010) Representing Entities in the OntoDM Data Mining Ontology. In Inductive Databases and Constraint-Based Data-Mining. (Eds) S. Dzeroski, B. Goethals, P. Panov. Springer.

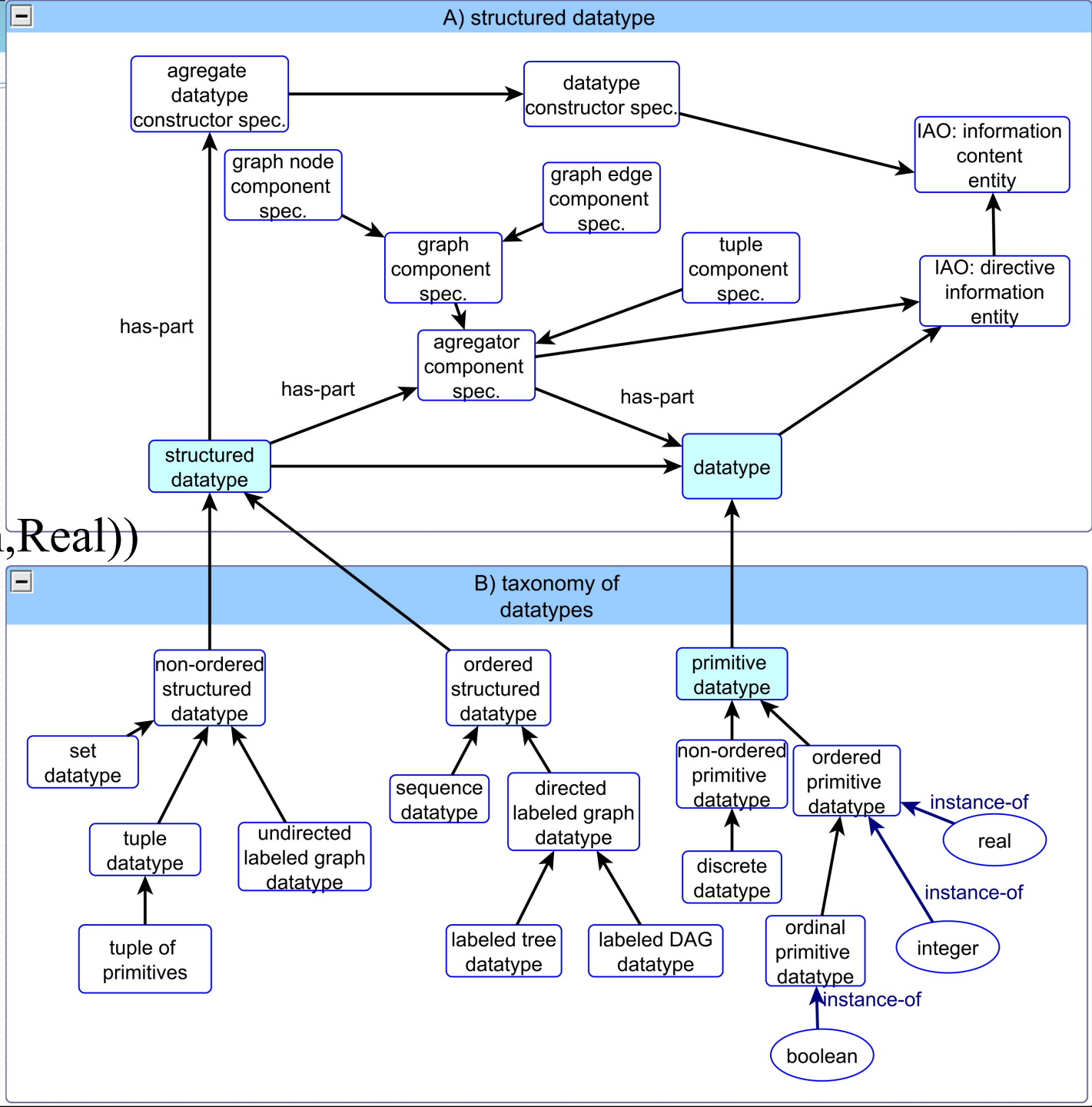
**S. Dzeroski. Towards a general framework for data mining. In S. Dzeroski and J. Struyf, (Eds) KDID, v. 4747 of Lecture Notes in Computer Science, 259–300.

OntoDM motivation:

- Minimum information for the description of DM investigations;
- Support representation and annotation of DM algorithms;
- Support representation and annotation of DM scenarios;
- Selection of suitable algorithms for datasets;
- Search for different implementations of algorithms;
- Support composition of algorithms, etc.

Taxonomy of datatypes

example:
Set(Tuple(Boolean,Real))



Data Mining Task

- task of data mining is to produce a generalization from a given set of data
- OntoDM defines four fundamental tasks:

- estimating the joint probability distribution
- learning a predictive model
- clustering
- pattern discovery

Data Mining Algorithm

data mining algorithm *is-a* algorithm implemented as computer program and designed to solve a data mining task.

has-specified-input dataset (examples of a given type)

has-specified-output generalization (from a given type on a given datatype)

has-part algorithm component

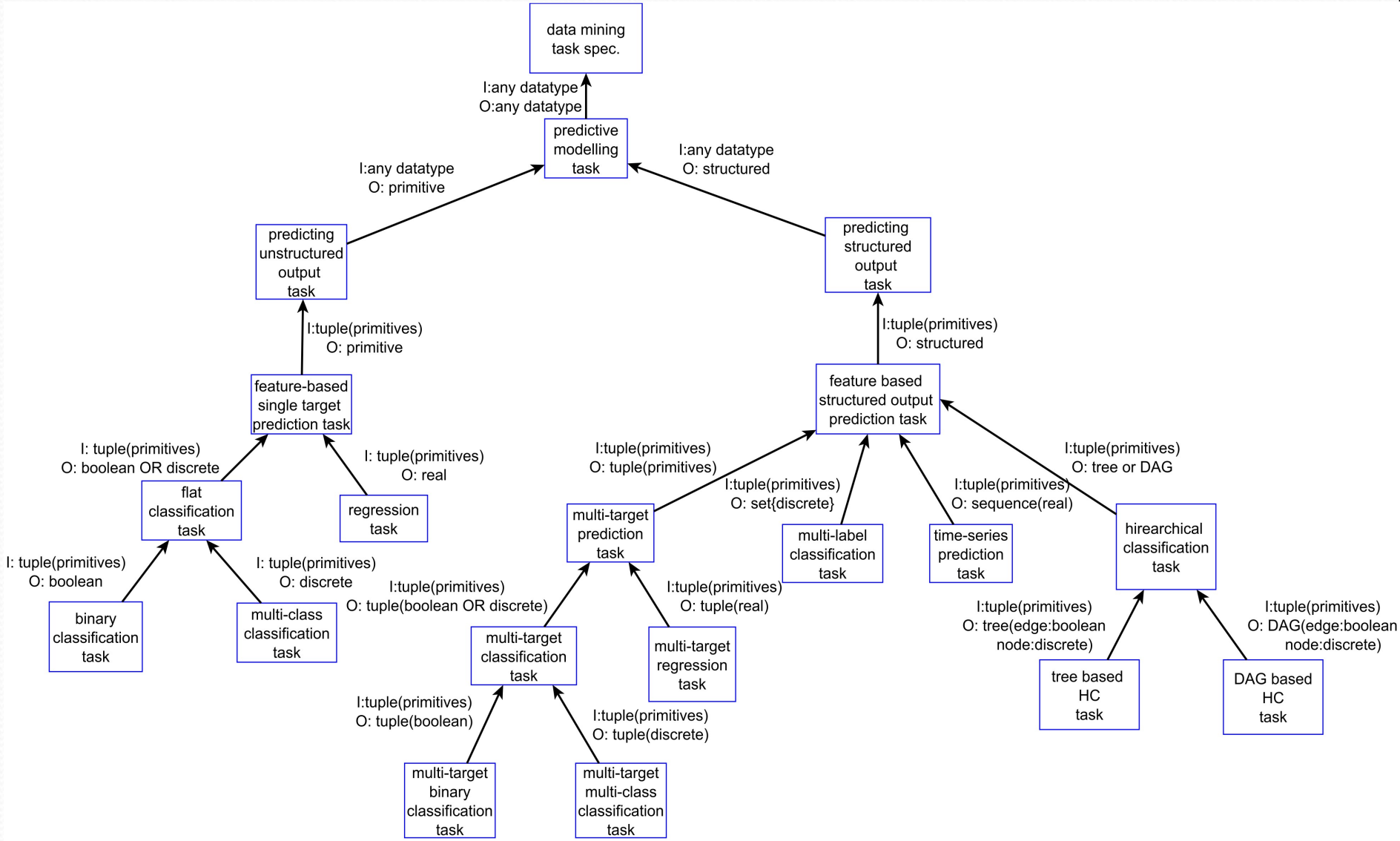
distance *is-a* algorithm component

feature *is-a* algorithm component

kernel *is-a* algorithm component

generality operator *is-a* algorithm component

Predictive modeling algorithms



OntoDM & DDI for QSAR modelling

- Lack of standards in the area of drug discovery (DD).
- QSARs modelling based on propositional datasets.
- A need to support evaluation of QSAR models (different models can be evaluated differently).
- A need to define and record properties of datasets.

(3) area of research interest

OntoDM can offer to DD semantic support of a QSAR modelling workflow:

- Selection of the best DM algorithms
- Meta-learning
- Active learning

OntoDM and DDI – based interfaces to support of a QSAR modelling workflow.

(4) area of research interest

OntoDM, Expose (Leuven, Belgium), and DMOP (e-Lico EU projects) have agreed to form DMO Foundry.

- the DMO Core
- set up principles
- a Portal
- the standard
- the MI

(5) area of research interest

SBO (systems biology ontology) next generation

- to fix major problems

- SBO is used for SBML, a graphic language, interfaces, et.

Prof. Sašo Džeroski, Jožef Stefan Institute, Slovenia

Dr Ljupco Todorovski, University of Ljubljana, Slovenia

SBO problems

SBO is a member of OBO, but it does not follow the principles:

- top classes

BFO : object, processes, quality, function, role,...

SBO: functional entity, material entity, interaction, math expression, participant role,...

- relations

RO: *has-participant, has-quality,...*

SBO: the class 'relationship' *is-a* 'interaction'

- Duplication of classes, instead of import, i.e. chemical entities
- Did not use any logic, i.e. the class 'equivalence'

SBO examples

- Enzyme can not be a material entity
- Material entity can not be an interaction outcome
- Empty set, observable *is-a* material entity is
- Class 'material entity of unspecified nature'
- Logical operators (AND, OR) are interactions
- Biological activity *is-a* interaction
- Process *is-a* interaction

Analysis of ontologies

- Soldatova, LN & King, RD. (2005) Are the Current Ontologies used in Biology Good Ontologies? Nature Biotechnology 9/23: 1096-1098.
- Soldatova, LN & King, RD. (2006) Reply to Wrestling with SUMO and bio-ontologies. Nature Biotechnology 24: 23.
- Schierz, A.C., Soldatova, L.N. and King, R.D. (2007) Overhauling the PDB. Nature Biotechnology 25/4: 437-442.
- Schierz, A.C., Soldatova, L. N. and King, R.D. (2007) Reply to Overhauling the PDB. Nature Biotechnology 25/8: 846.

(6) area of research interest

Prof. Sašo Džeroski, Jožef Stefan Institute, Slovenia

Dr Ljupco Todorovski, University of Ljubljana, Slovenia

automated modelling of dynamic systems*

- SW for modelling domain knowledge
- Ontology for dynamic systems

*Džeroski, S. & Todorovski, L. Equation discovery for systems biology. Current Opinion in Biotechnology, 2008.

(7) area of research interest

Prof. Sašo Džeroski, Jožef Stefan Institute, Slovenia
Dr Ljupco Todorovski, University of Ljubljana, Slovenia

automated modelling of ecosystems*

- SW for modelling domain knowledge
- Ontology for ecosystems
- Ecoinformatics, http://www.ecoinformatics.org/

*Atanasova,et al Constructing a library of domain knowledge for automated modelling of aquatic systems. Ecological modelling, 2006.

(8) area of research interest

Prof. Sašo Džeroski, Jožef Stefan Institute, Slovenia

- Biological datatypes (i.e. microarray data, sequencing data)
- Properties of biological datasets (i.e. GEO dataset)

An Ontology of Science

- Based on an automated discovery
- Based on the work done in Knowledge Discovery
- Beyond BFO: models, theories, hypotheses, paradigm, paradoxes, believes, opinions.
- Components: taxonomy of datatypes, tasks, methods
- Applications: research workflow, research interactions, projects

Thank you