Exposé: An ontology for machine learning experimentation

Joaquin Vanschoren, K.U.Leuven (Belgium), U. Leiden (The Netherlands)
Larisa Soldatova, University of Aberystwyth (UK)

DM Ontology Jamboree 2010

DTAI

Declaratieve Talen en Artificiële Intelligentie

KATHOLIEKE UNIVERSITEIT LEUVEN

Faculteit Ingenieurswetenschappen

Exposé: An ontology for machine learning experimentation

Joaquin Vanschoren, K.U.Leuven (Belgium), U. Leiden (The Netherlands)
Larisa Soldatova, University of Aberystwyth (UK)

DM Ontology Jamboree 2010

Overview

Ontology lessons
Exposé ontology
Use cases

Ontology lessons

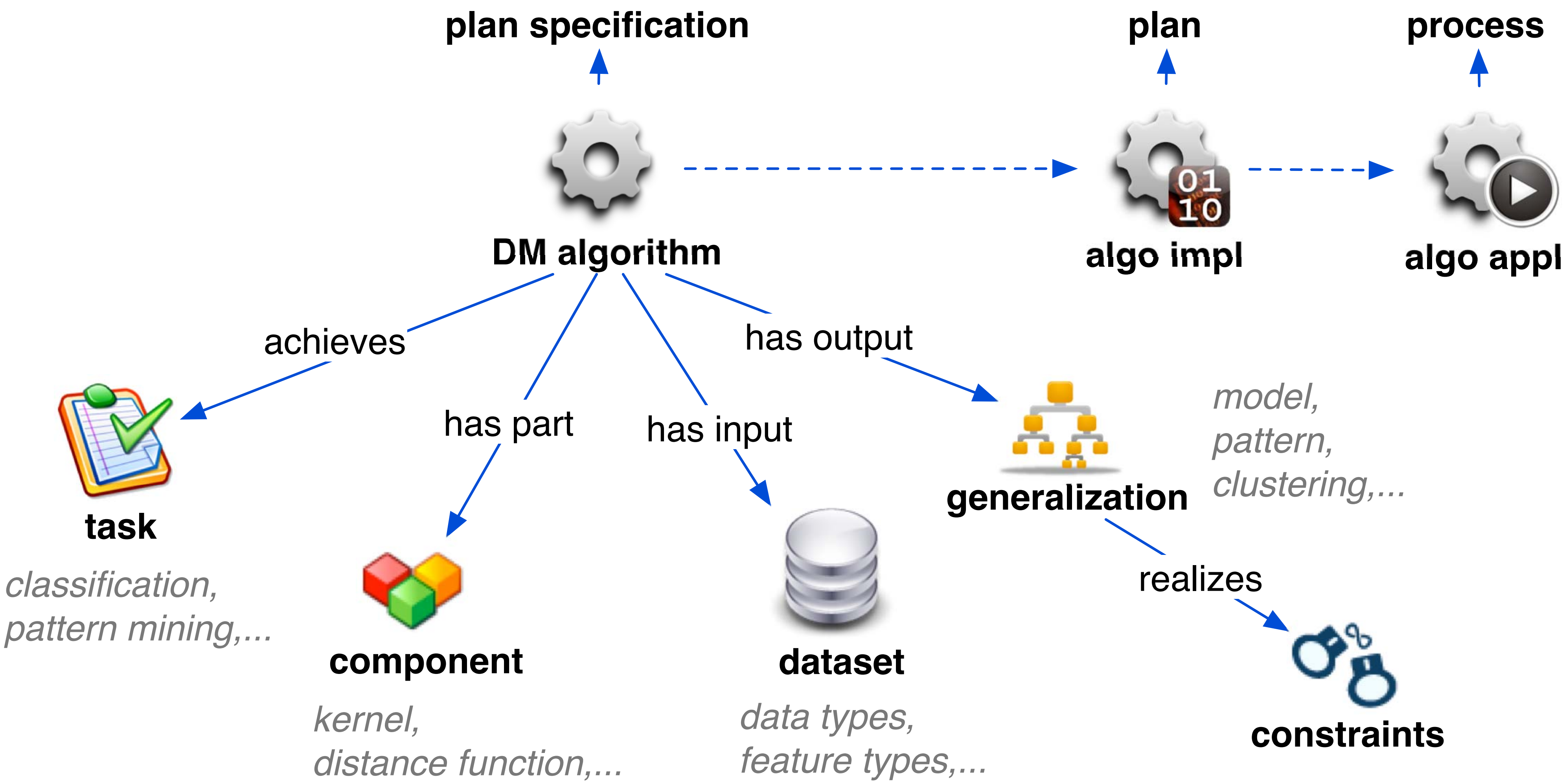
What did we learn from other ontologies

Ontology design

- Start from accepted classes & properties (top-level ontologies, e.g. OBI, RO)
- If possible, reuse prior ontologies to build on their knowledge/consensus
- Use ontology design patterns: reusable patterns for recurrent problems
    - http://ontologydesignpatterns.org
- Check clarity, consistency, extensibility, minimal commitment

Ontology recap: OntoDM (Panov et al., ’09,’10)

- Aim: unified framework for DM research, builds on BFO



plan specification

plan

process

DM algorithm

algo impl

algo appl

achieves

has part

has input

has output

task

classification, pattern mining,...

component

kernel, distance function,...

dataset

data types, feature types,...

generalization

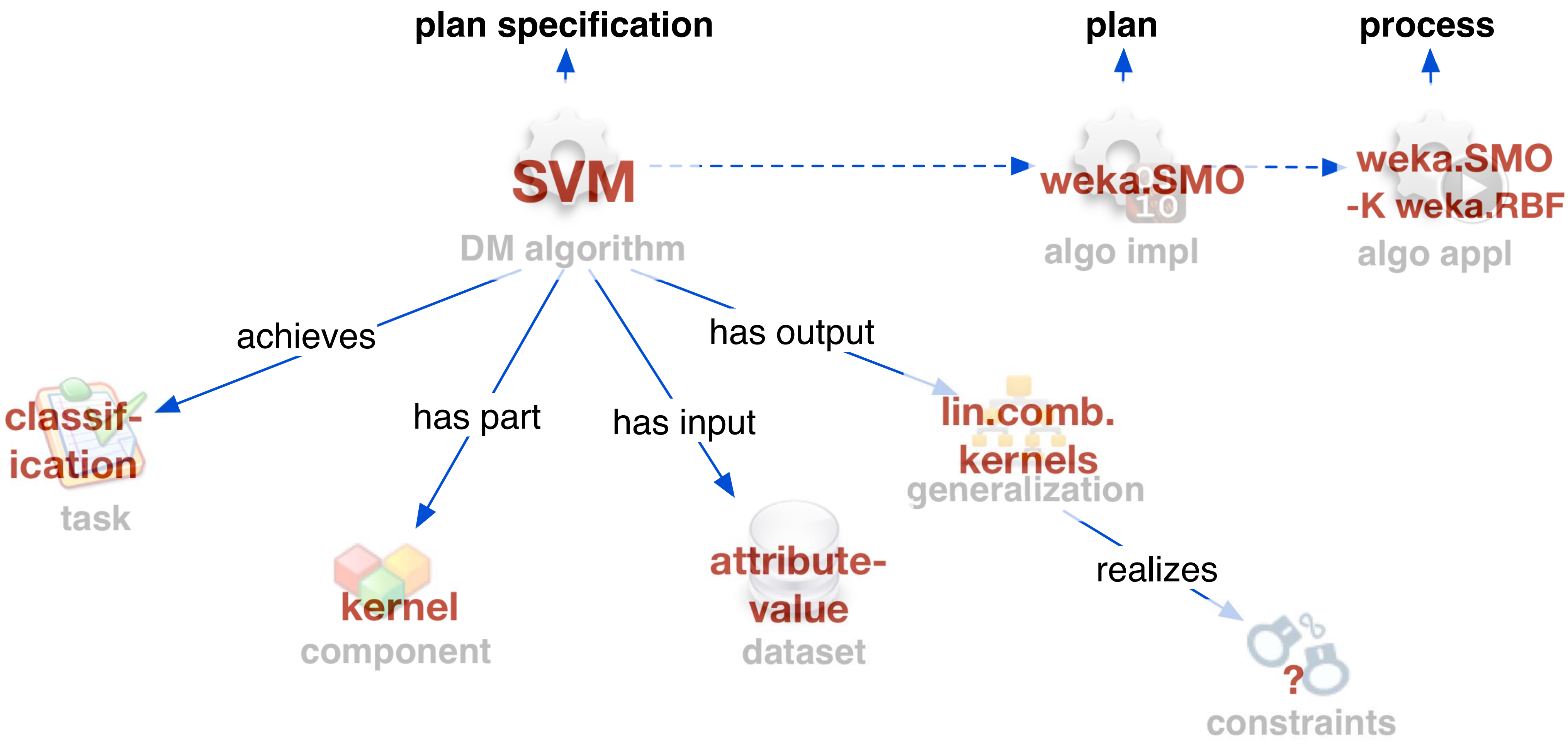
model, pattern, clustering,...

realizes

constraints

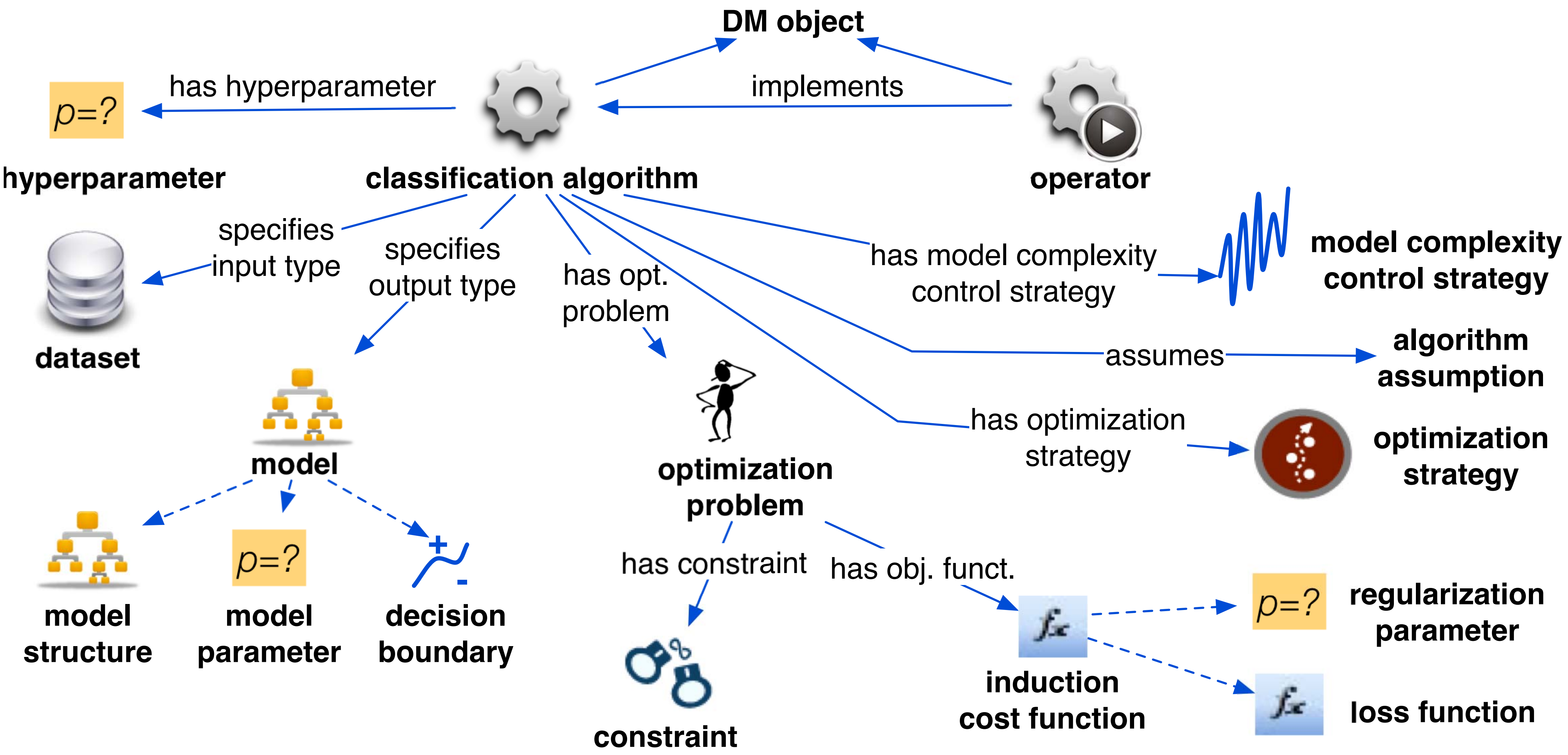
Ontology recap: OntoDM (Panov et al., ’09,’10)

- Aim: unified framework for DM research, builds on BFO



Ontology recap:
DMOP (Hilario et al., ’09)

- Model internal structure of learning algorithms



DM object

has hyperparameter

implements

p=?

hyperparameter

classification algorithm

operator

specifies input type

specifies output type

has opt. problem

has model complexity control strategy

model complexity control strategy

dataset

assumes

algorithm assumption

model

has optimization strategy

optimization strategy

optimization problem

p=?

model structure

model parameter

decision boundary

has constraint

has obj. funct.

p=?

regularization parameter

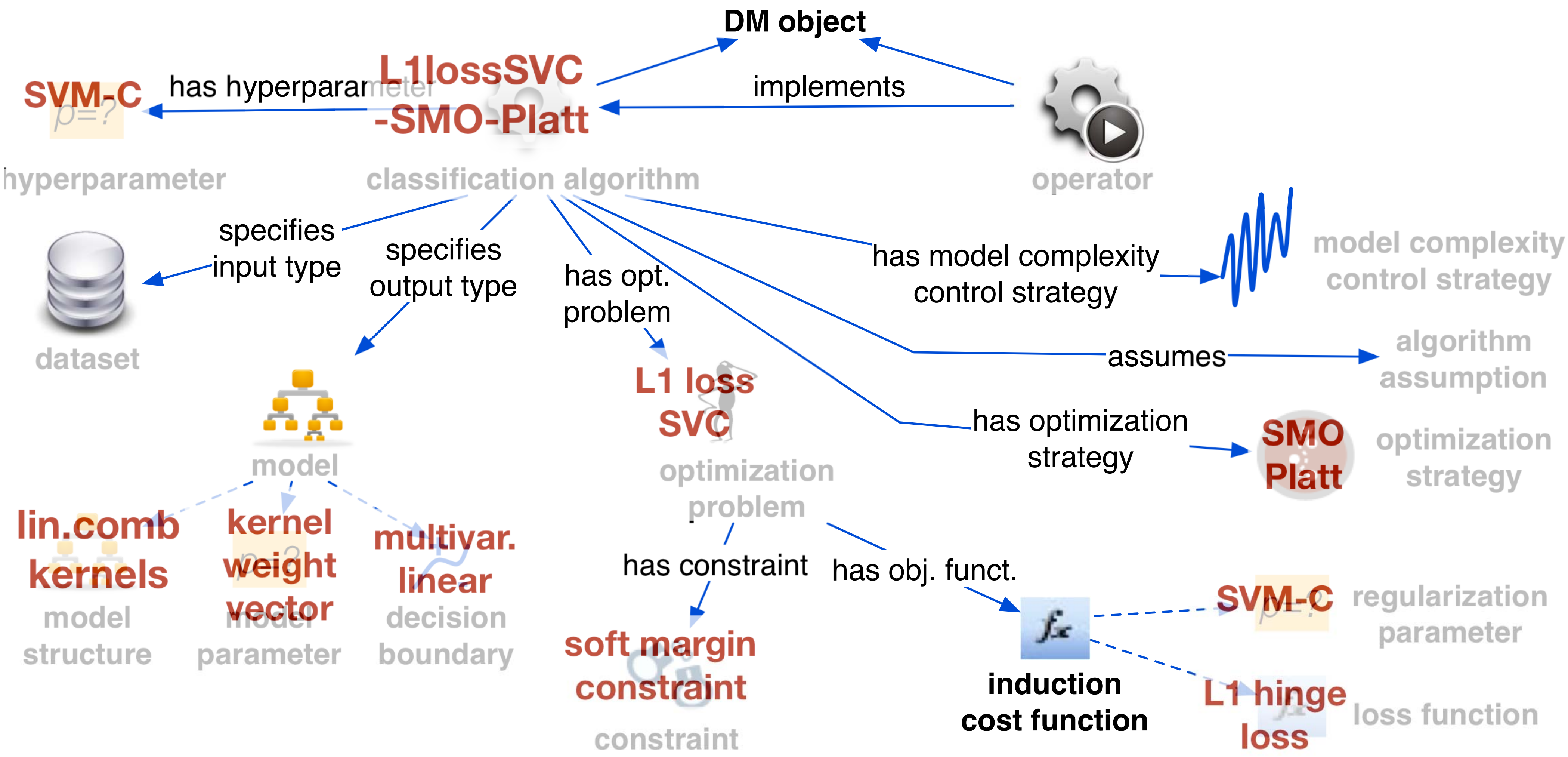
constraint

induction cost function

loss function

Ontology recap: DMOP (Hilario et al., ’09)

- Model internal structure of learning algorithms



DM object

SVM-C $p=?$

has hyperparameter

L1lossSVC -SMO-Platt

implements

hyperparameter

classification algorithm

operator

specifies input type

specifies output type

has opt. problem

has model complexity control strategy

model complexity control strategy

dataset

assumes

algorithm assumption

L1 loss SVC

has optimization strategy

SMO Platt

optimization strategy

model

optimization problem

lin.comb kernels

model structure

kernel weight vector

model parameter

multivar. linear

decision boundary

has constraint

has obj. funct.

SVM-C $p=?$

regularization parameter

soft margin constraint

constraint

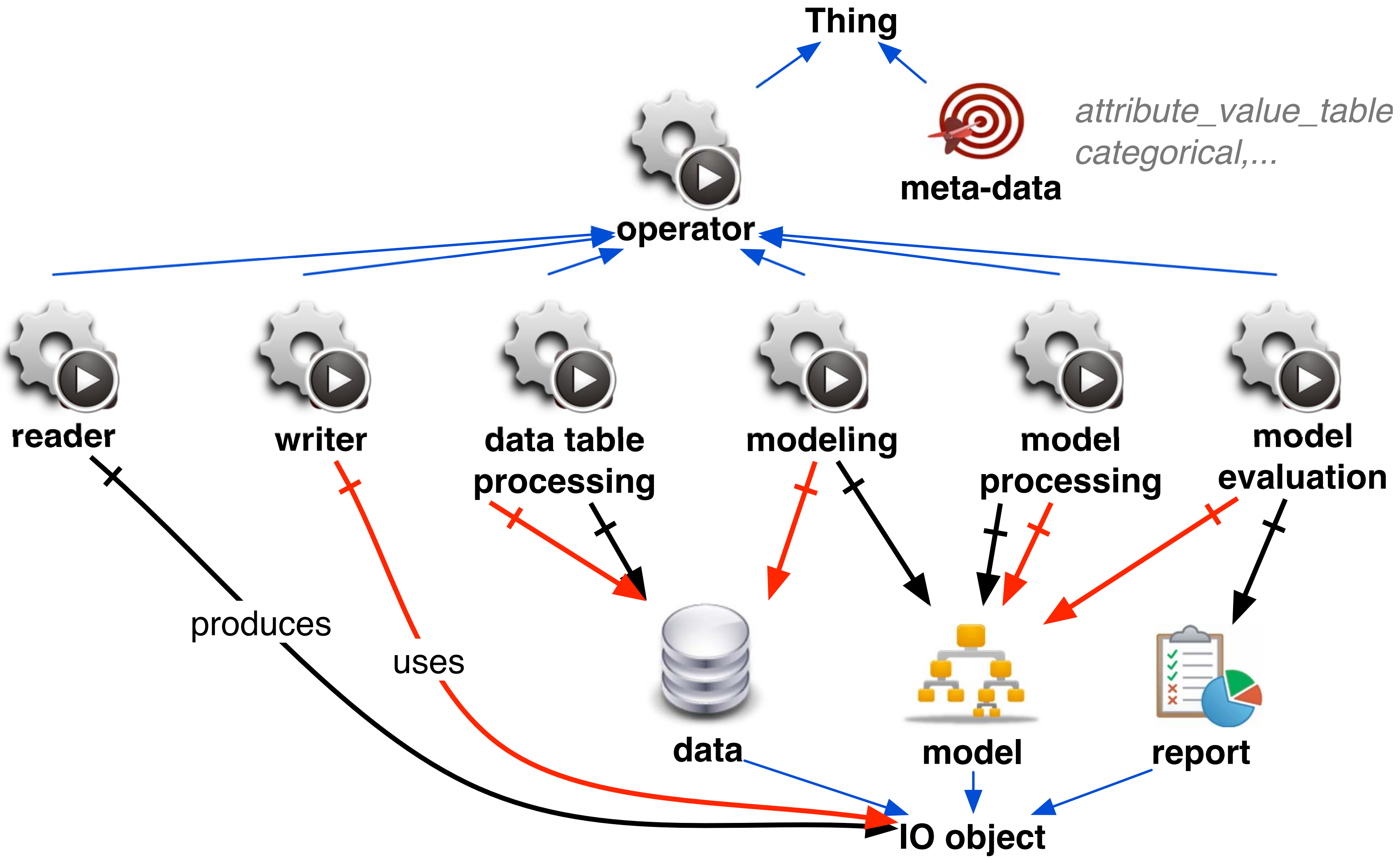
induction cost function

L1 hinge loss

loss function

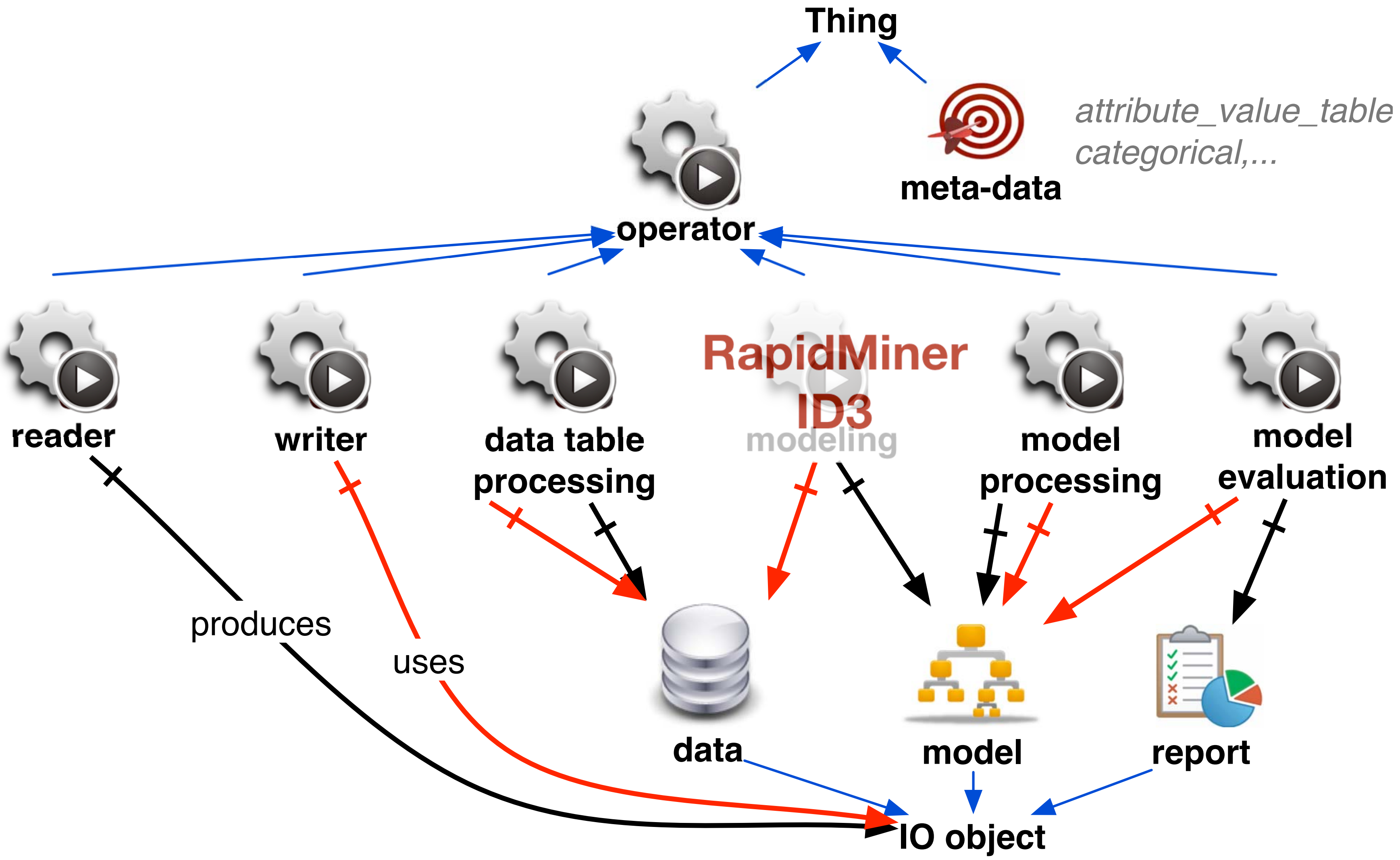
Ontology recap: DMWF (Kietz et al., '09)

- Reason about KD operators: in/outputs, conditions/effects (SWRL rules)



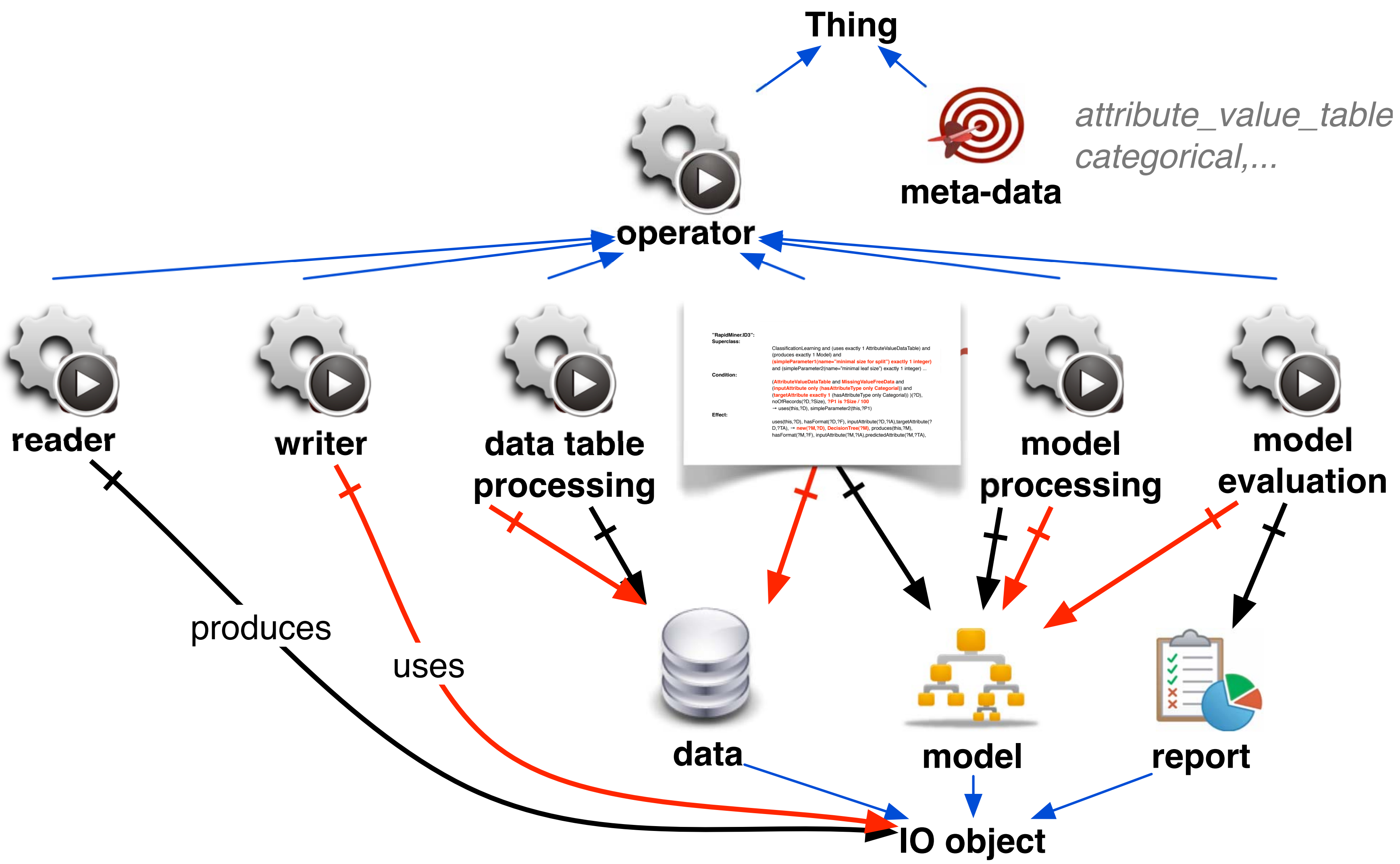
Ontology recap:
DMWF (Kietz et al., ’09)

- Reason about KD operators: in/outputs, conditions/effects (SWRL rules)



Ontology recap:
DMWF (Kietz et al., ’09)

- Reason about KD operators: in/outputs, conditions/effects (SWRL rules)



”RapidMiner.ID3”:

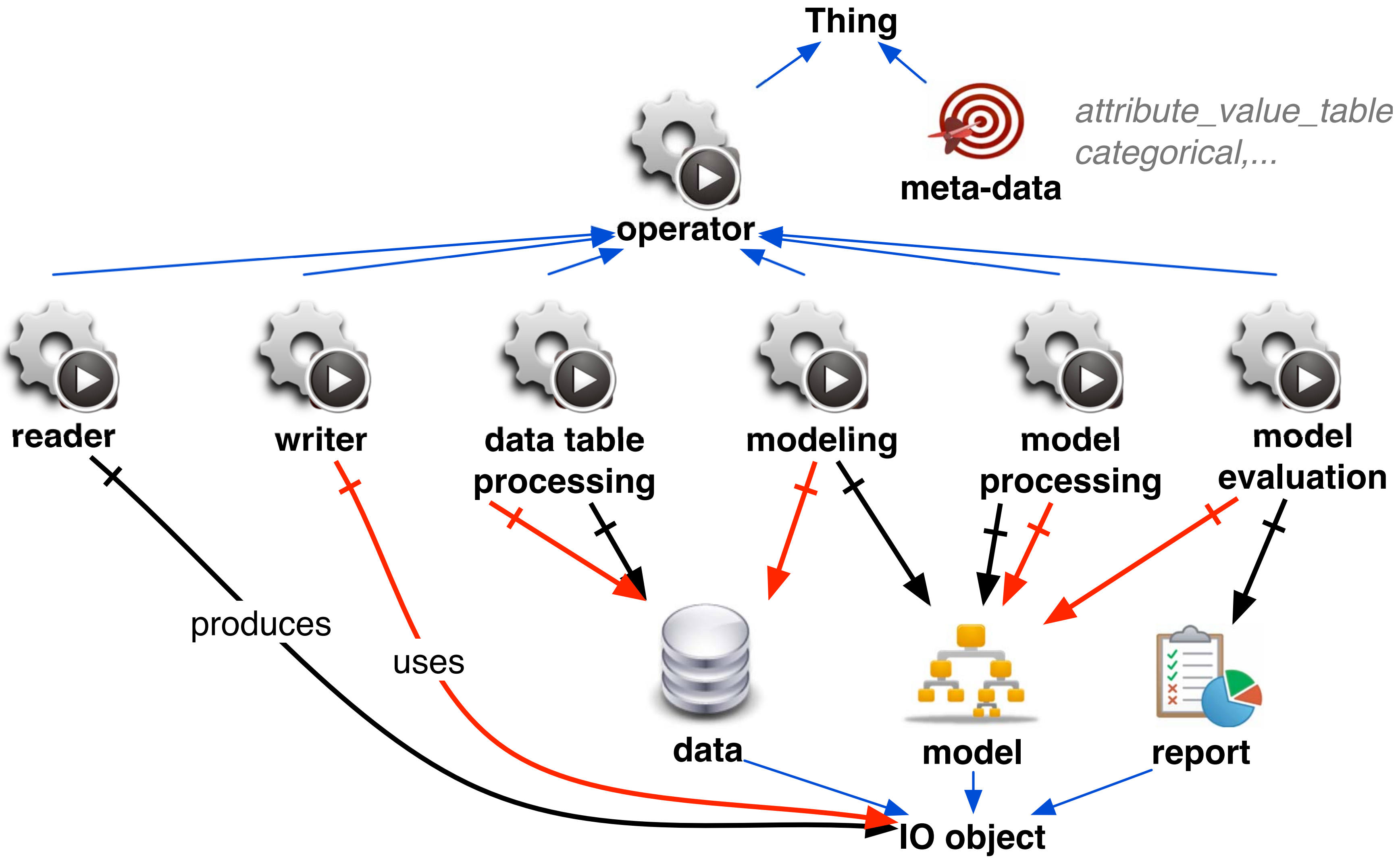
Superclass:
ClassificationLearning and (uses exactly 1 AttributeValueDataTable) and (produces exactly 1 Model) and **(simpleParameter1(name=”minimal size for split”) exactly 1 integer)** and (simpleParameter2(name=”minimal leaf size”) exactly 1 integer) ...

Condition:
(**AttributeValueDataTable** and **MissingValueFreeData** and (**inputAttribute only (hasAttributeType only Categorial**)) and (**targetAttribute exactly 1** (hasAttributeType only Categorial)))(?D), noOfRecords(?D,?Size), **?P1 is ?Size / 100**
→ uses(this,?D), simpleParameter2(this,?P1)

Effect:
uses(this,?D), hasFormat(?D,?F), inputAttribute(?D,?IA),targetAttribute(?D,?TA), → **new(?M,?D)**, **DecisionTree(?M)**, produces(this,?M), hasFormat(?M,?F), inputAttribute(?M,?IA),predictedAttribute(?M,?TA),

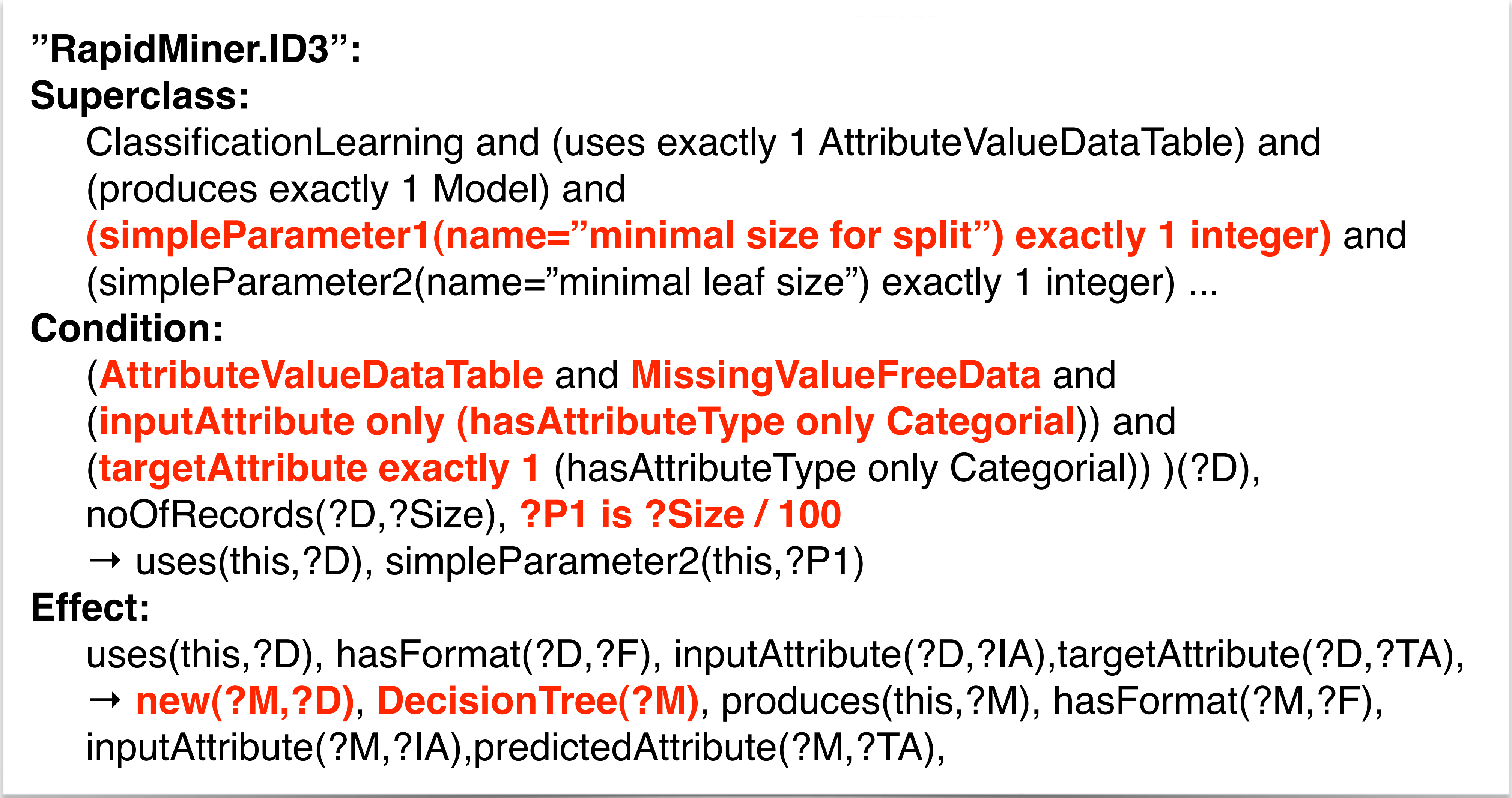
Ontology recap:
DMWF (Kietz et al., ’09)

- Reason about KD operators: in/outputs, conditions/effects (SWRL rules)



Ontology recap:
DMWF (Kietz et al., ’09)

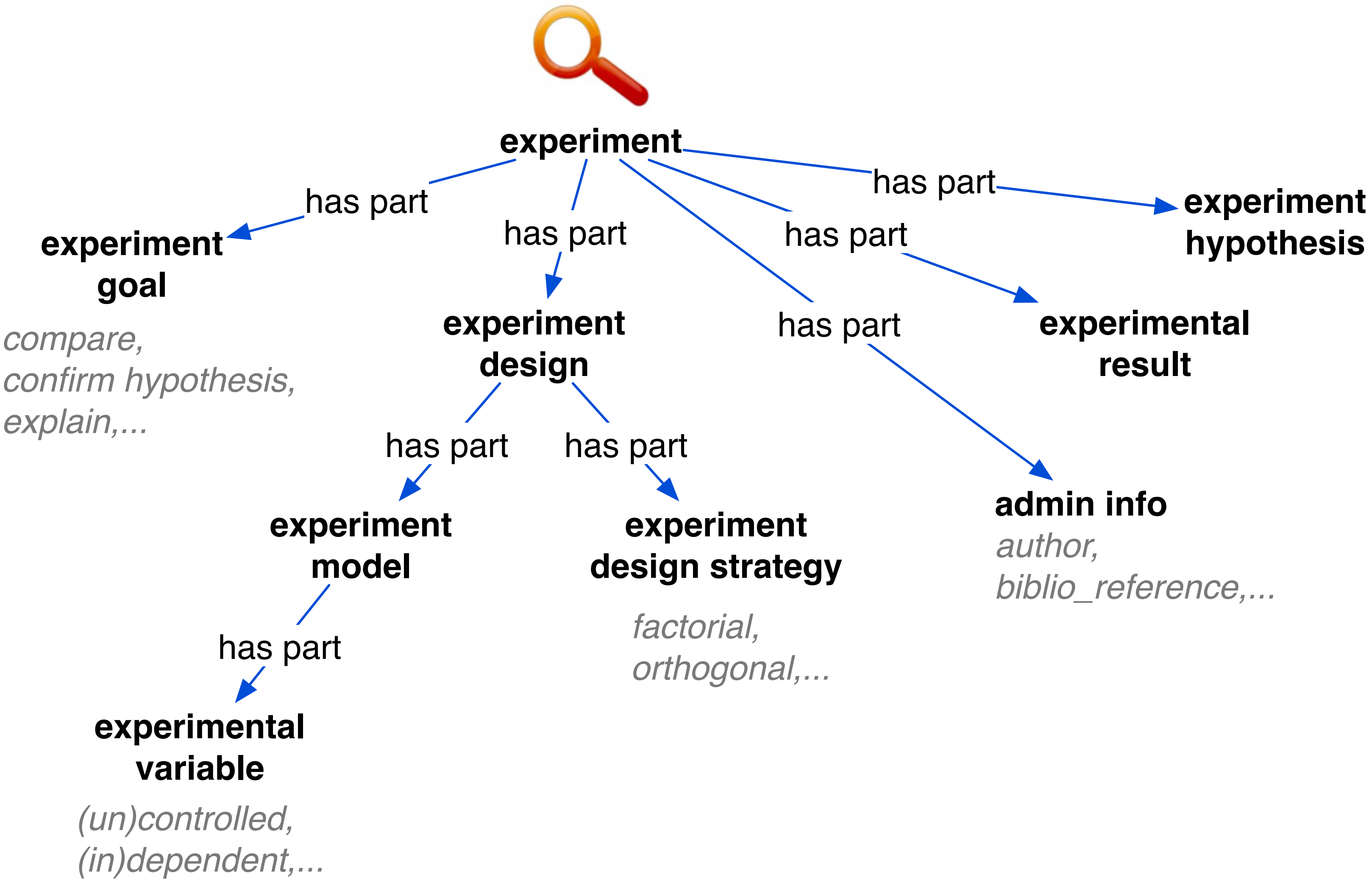
- Reason about KD operators: in/outputs, conditions/effects (SWRL rules)



”RapidMiner.ID3”:
Superclass:
ClassificationLearning and (uses exactly 1 AttributeValueDataTable) and
(produces exactly 1 Model) and
(simpleParameter1(name=”minimal size for split”) exactly 1 integer) and
(simpleParameter2(name=”minimal leaf size”) exactly 1 integer) ...
Condition:
(**AttributeValueDataTable** and **MissingValueFreeData** and
(**inputAttribute only (hasAttributeType only Categorial)**)) and
(**targetAttribute exactly 1** (hasAttributeType only Categorial)))(?D),
noOfRecords(?D,?Size), **?P1 is ?Size / 100**
→ uses(this,?D), simpleParameter2(this,?P1)
Effect:
uses(this,?D), hasFormat(?D,?F), inputAttribute(?D,?IA),targetAttribute(?D,?TA),
→ **new(?M,?D)**, **DecisionTree(?M)**, produces(this,?M), hasFormat(?M,?F),
inputAttribute(?M,?IA),predictedAttribute(?M,?TA),

Ontology recap:
EXPO (Soldatova and King, ’06)

- Make goal and structure of scientific experiments more explicit



experiment

has part

has part

has part

has part

has part

experiment goal

compare,
confirm hypothesis,
explain,...

experiment design

has part

has part

experiment model

has part

experimental variable

(un)controlled,
(in)dependent,...

experiment design strategy

factorial,
orthogonal,...

admin info

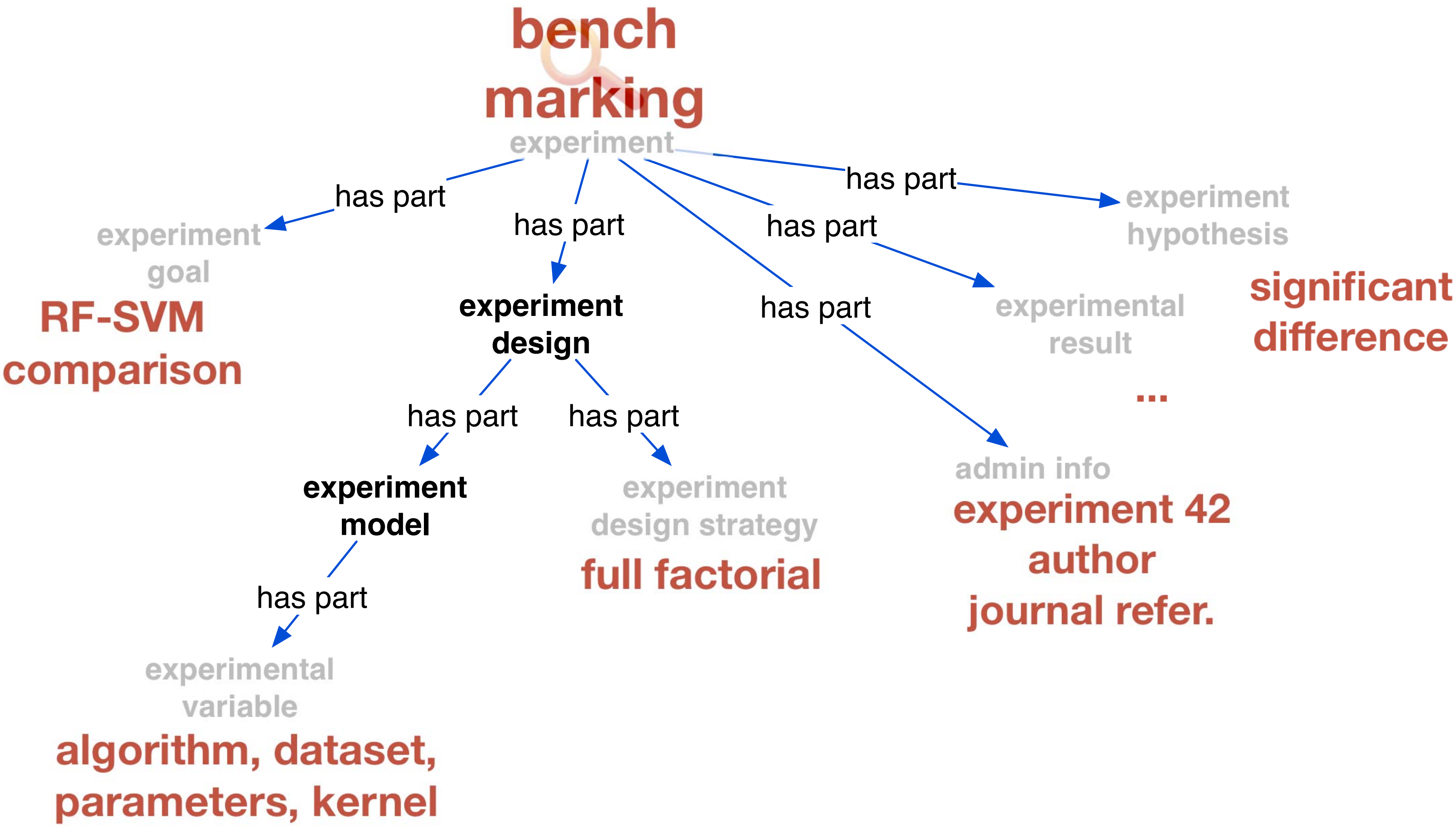
author,
biblio_reference,...

experimental result

experiment hypothesis

Ontology recap:
EXPO (Soldatova and King, ’06)

- Make goal and structure of scientific experiments more explicit



Exposé

an ontology for data mining experimentation

Context

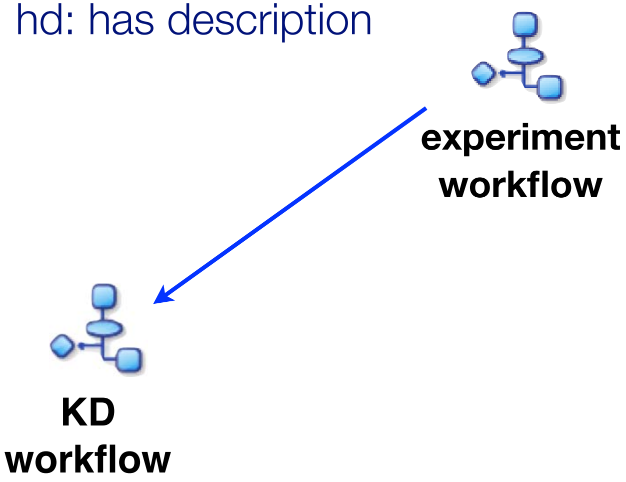
- Giant, public database(s) of data mining experiments
- We need:
    - Common language to share experiments (through DM tools)
    - Intuitive ways to store and query experimental results
- We want:
    - Interoperable ontology: OntoDM for top-level, DMOP for detailed properties of learning algorithms
    - Driven by actual experiments submitted to database
        - New algorithms -> ideally, described by author
        - Instances automatically extracted from database

Problem 1: Experiments

What is a machine learning experiment?
What do we need to know about it?

Exposé: Experiments

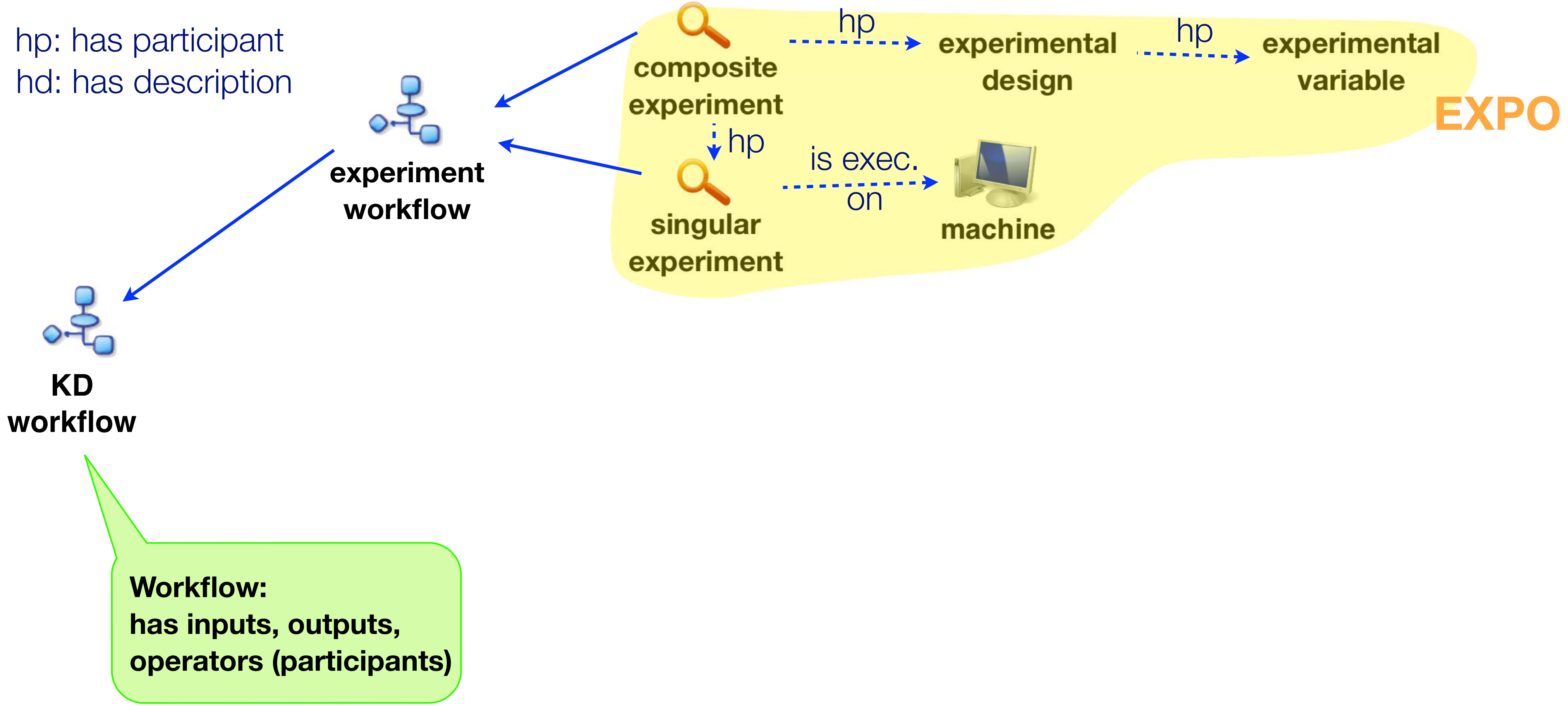
hp: has participant
hd: has description



Workflow:
has inputs, outputs,
operators (participants)

Exposé: Experiments

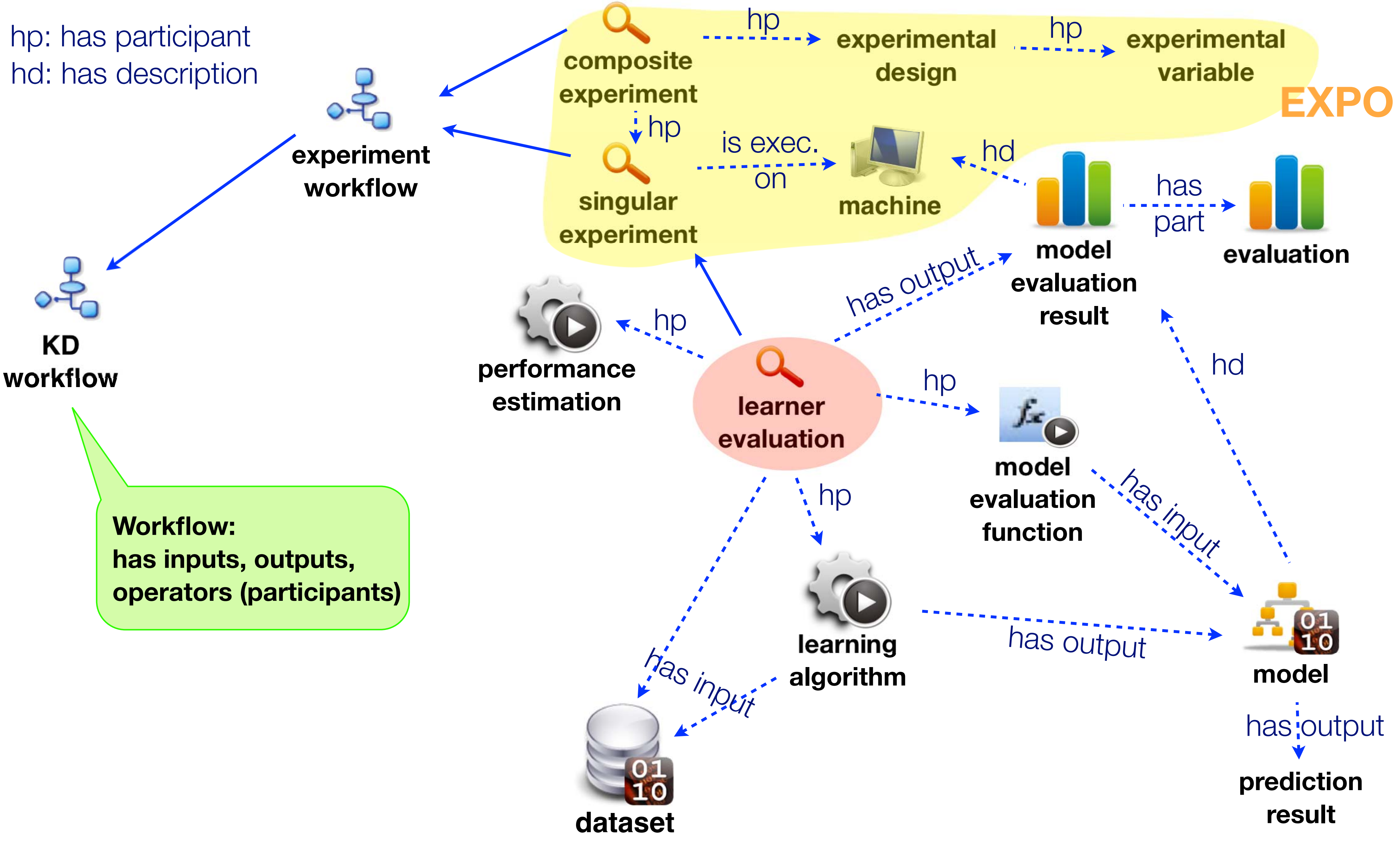
hp: has participant
hd: has description



Workflow:
has inputs, outputs,
operators (participants)

Exposé: Experiments

hp: has participant
hd: has description



Workflow:
has inputs, outputs, operators (participants)

Problem 2: Algorithms

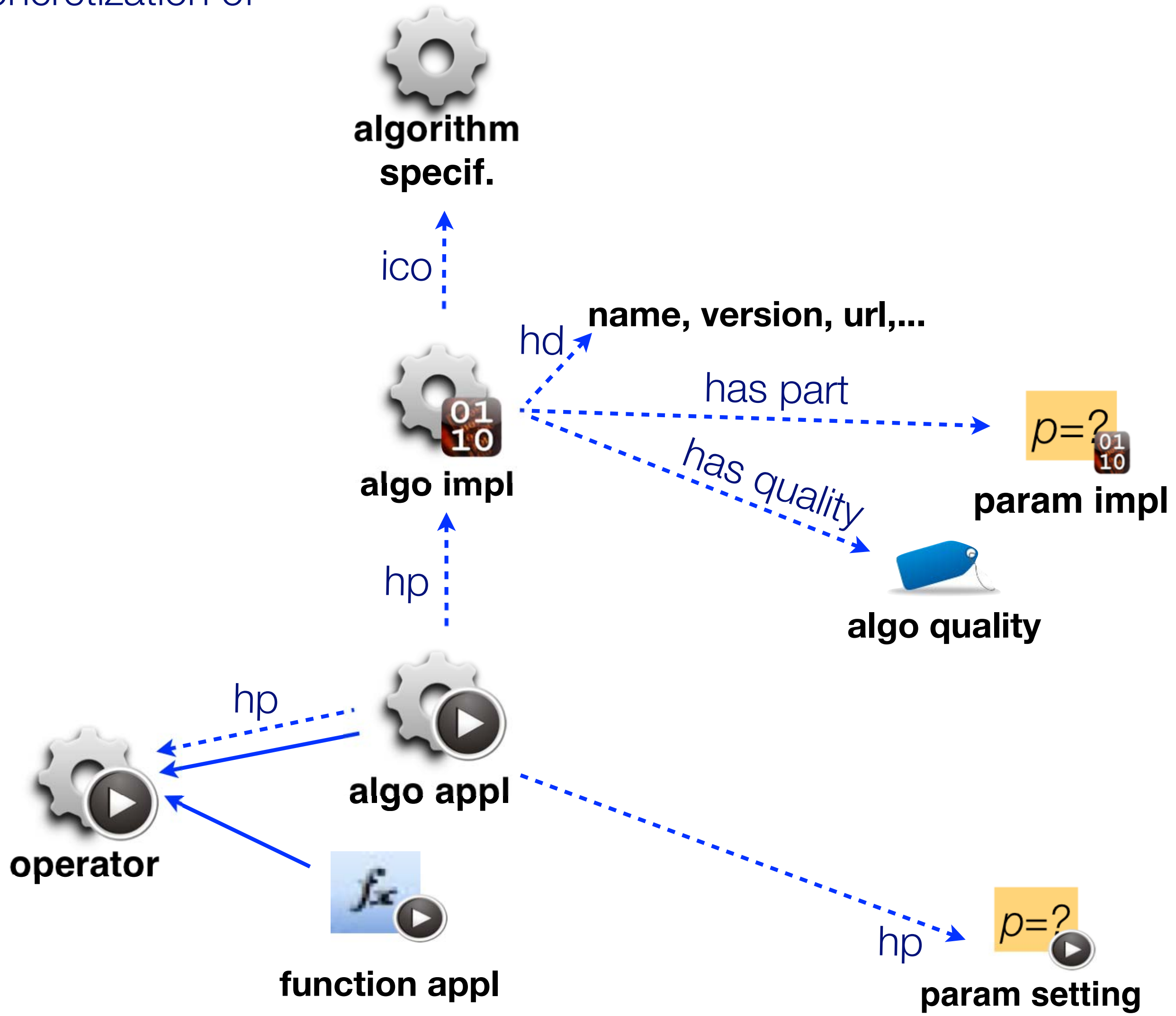
When talking about an algorithm, what is meant?

General algorithm?
Specific implementation? Which version?
When run, which parameters, components?

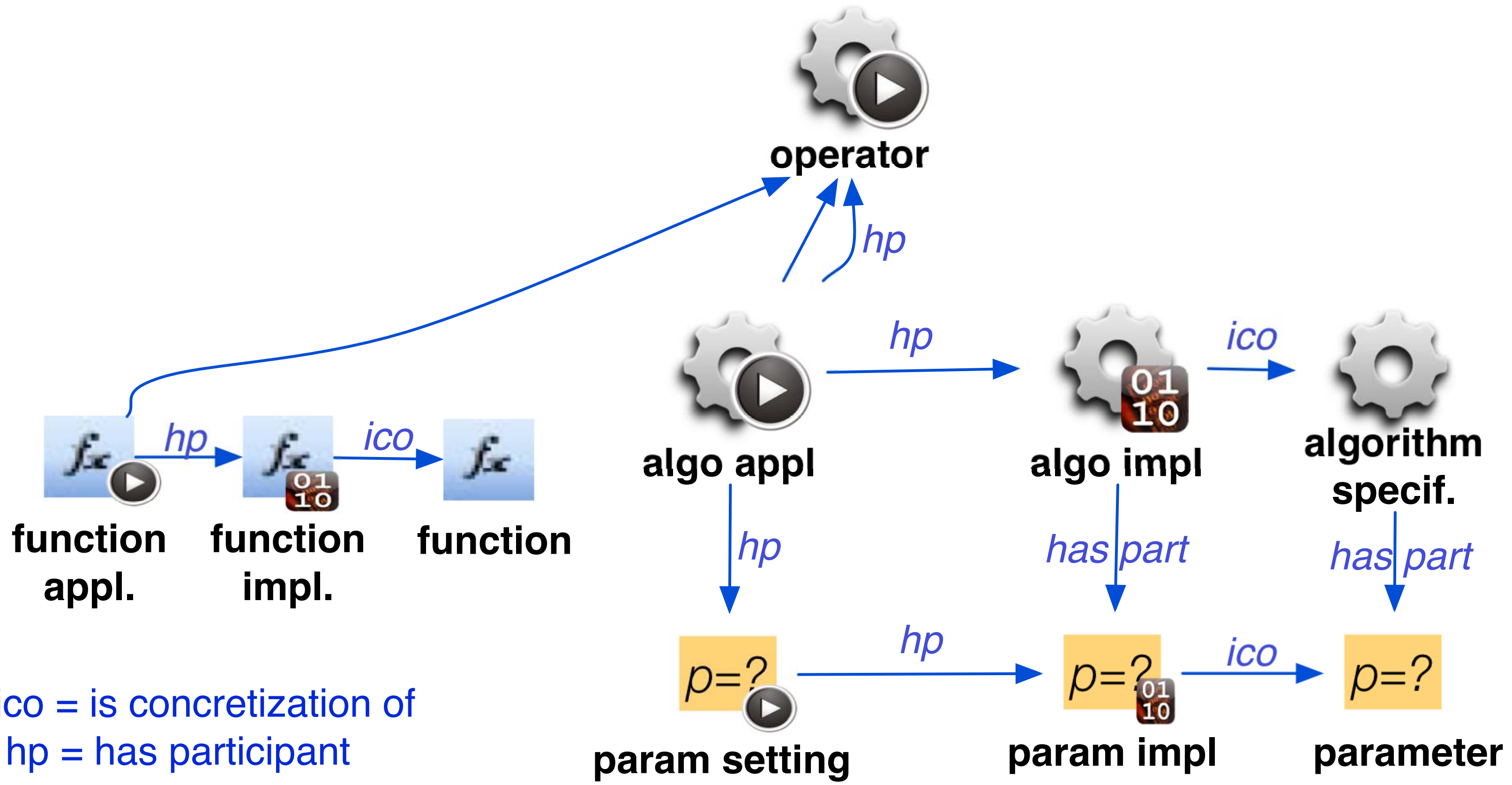
Exposé: Algorithms
Specification, implementation, application

hp: has participant
hd: has description
ico: is concretization of

Similar to OntoDM



Same for functions and parameters



ico = is concretization of
hp = has participant

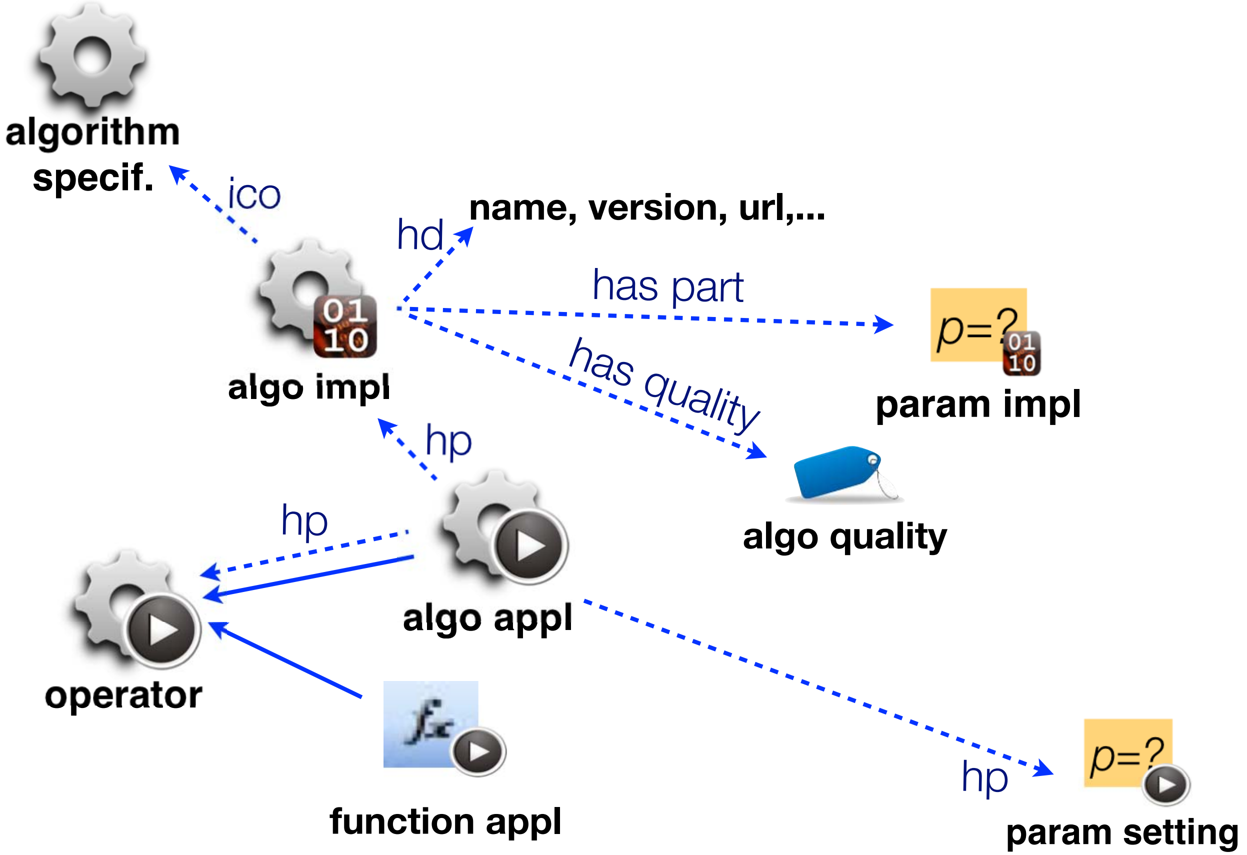
Problem 3: Algorithm composition

plug-in functions, kernels, other algorithms

such components play different roles
-> Agent-role pattern

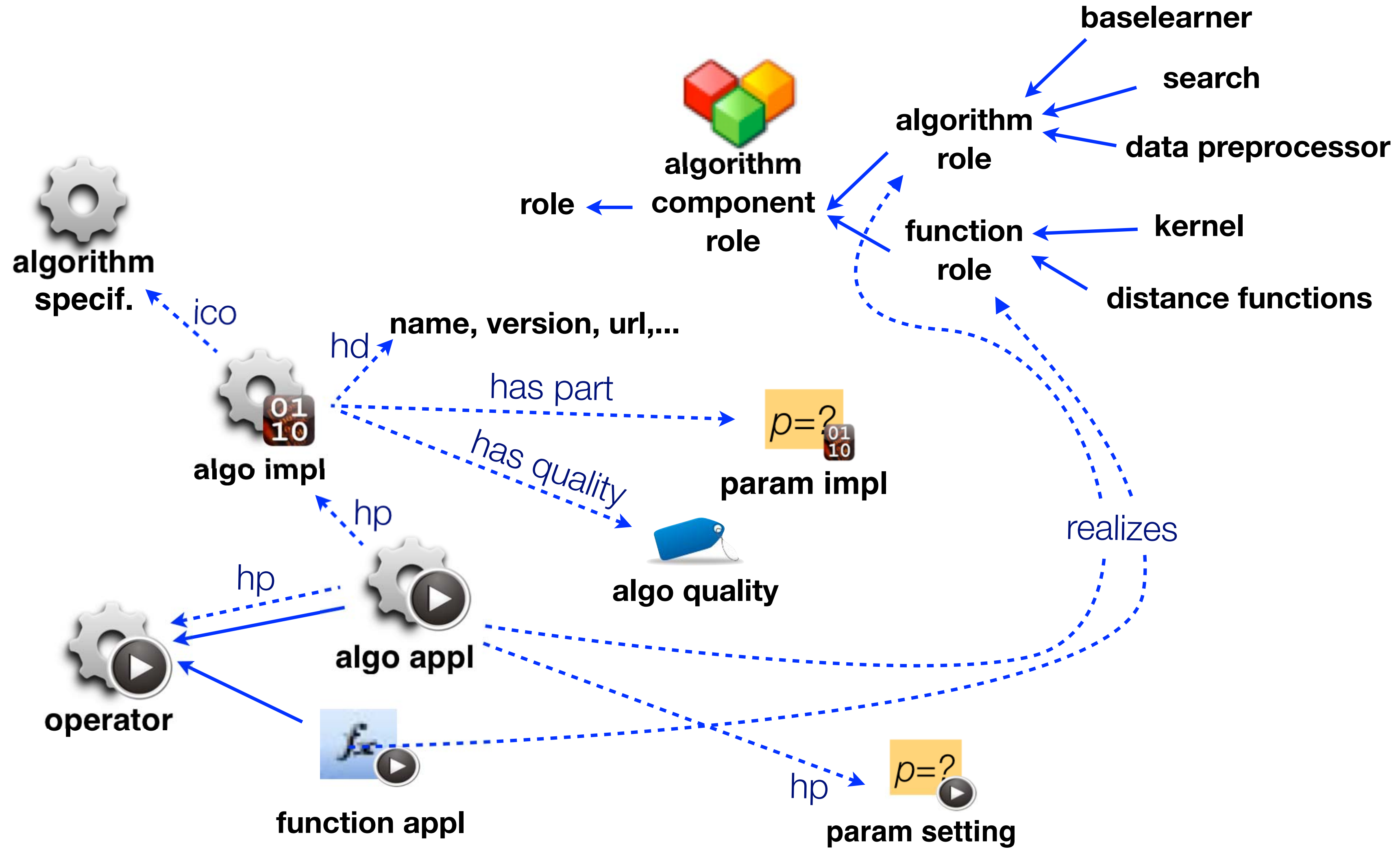
Exposé: Algorithms

hp: has participant
hd: has description



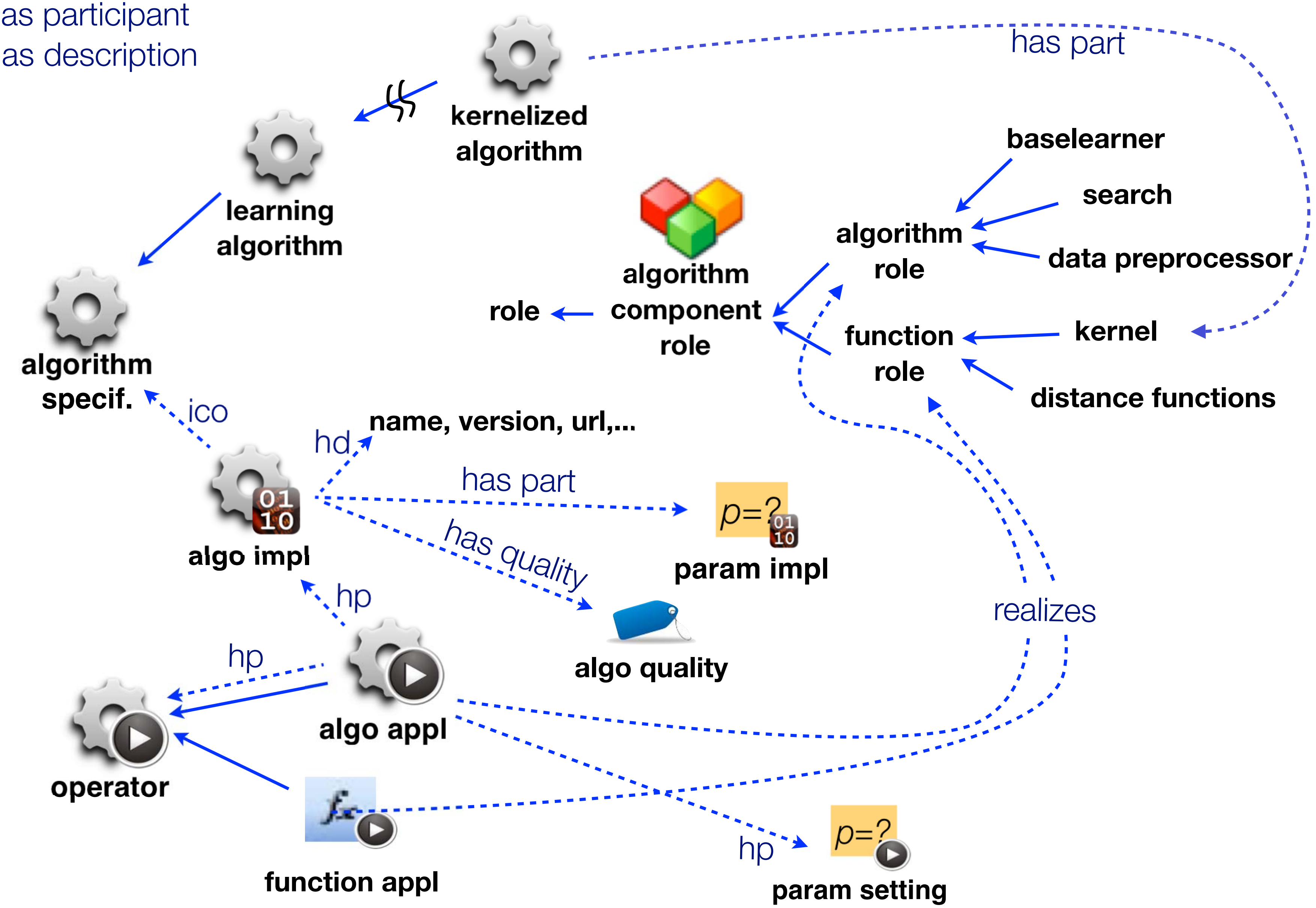
Exposé: Algorithms

hp: has participant
hd: has description



Exposé: Algorithms

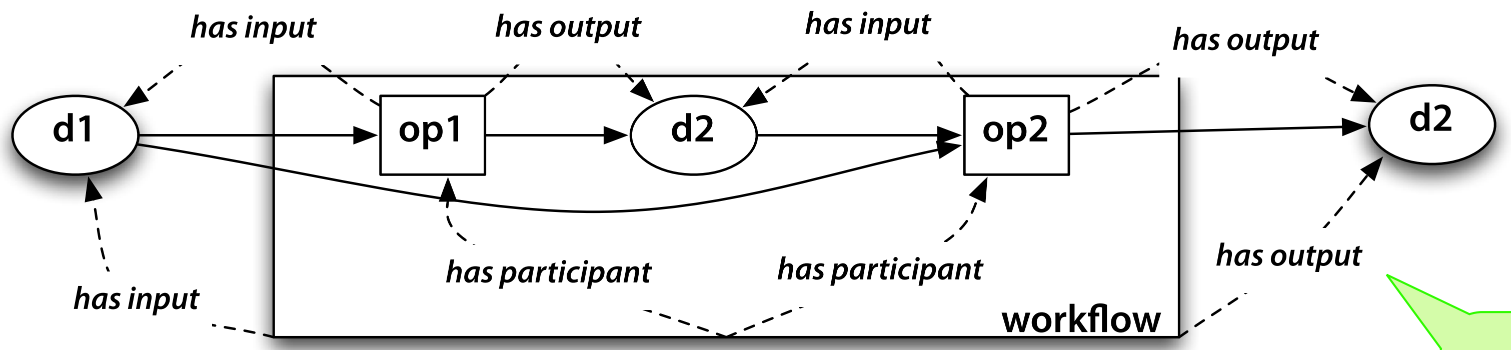
hp: has participant
hd: has description



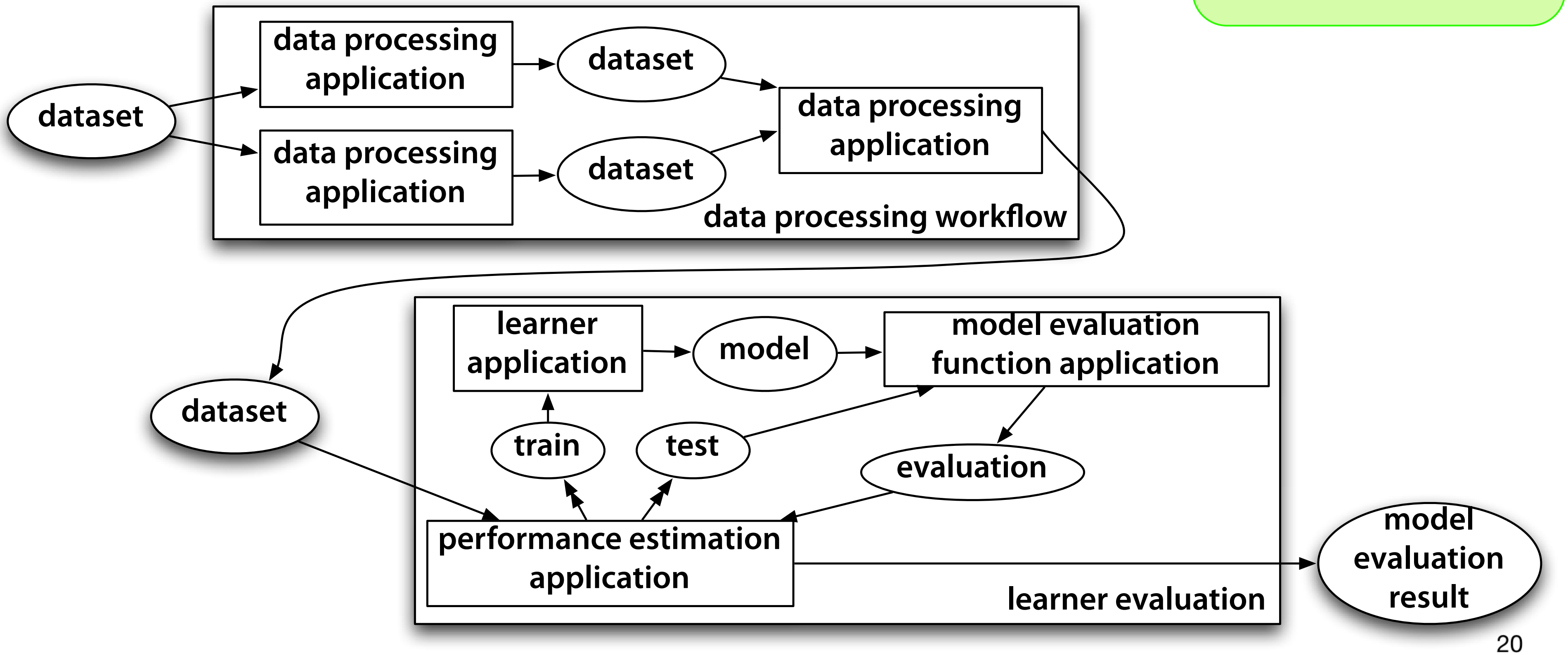
Problem 4: Workflows

Inputs, outputs, operators
Hierarchical: workflows within workflows
Reuse, parameterize common workflows, e.g. k-fold CV

Exposé: workflows



Similar to RapidMiner?

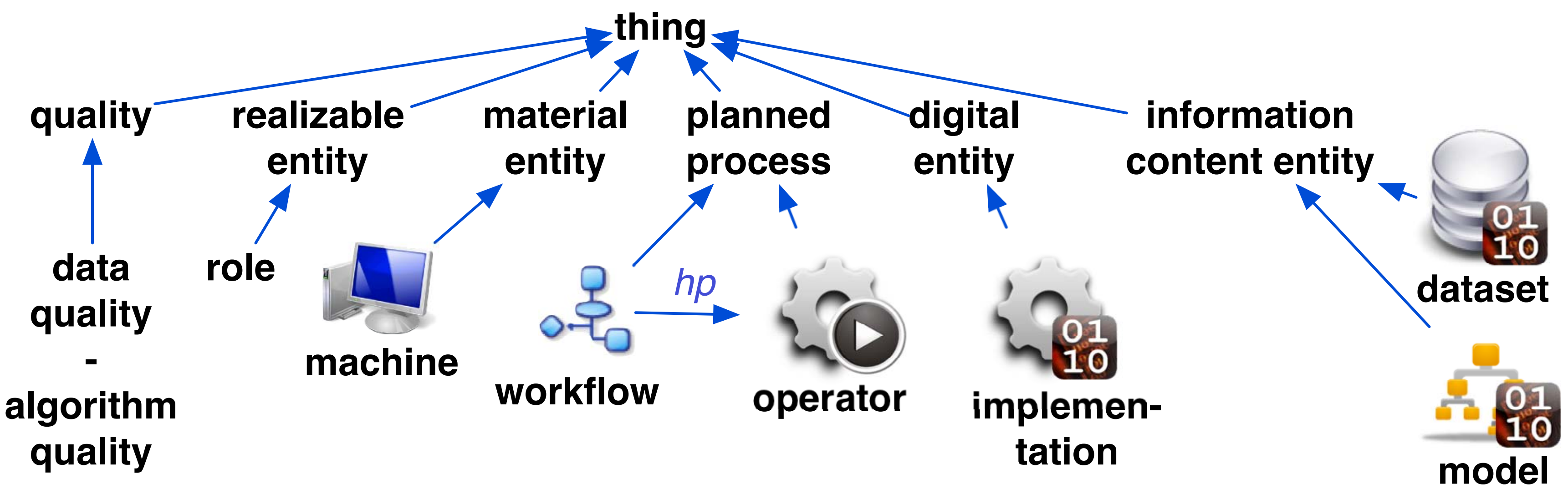

data processing application
dataset
dataset
data processing application
data processing application
dataset
data processing workflow
learner application
model
model evaluation function application
dataset
train
test
evaluation
performance estimation application
learner evaluation
model evaluation result

Problem 5: Reuse

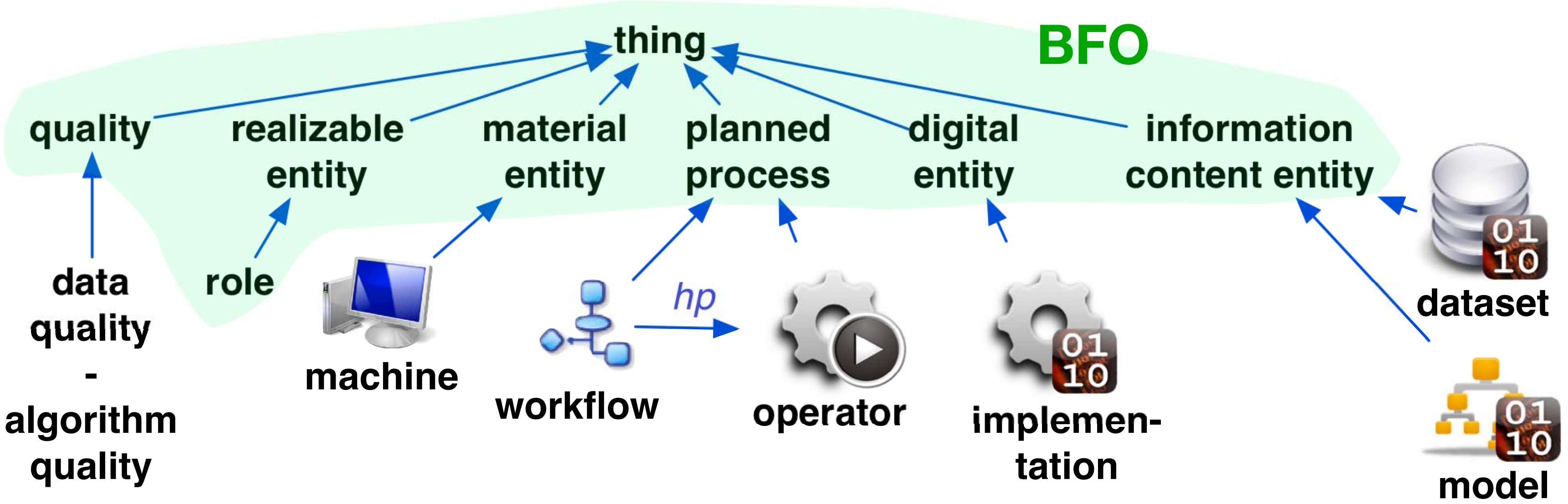
How can we make maximal use of existing ontologies?

OBI: top-level
OntoDM: top-level DM concepts
DMO: operators, learning mechanisms

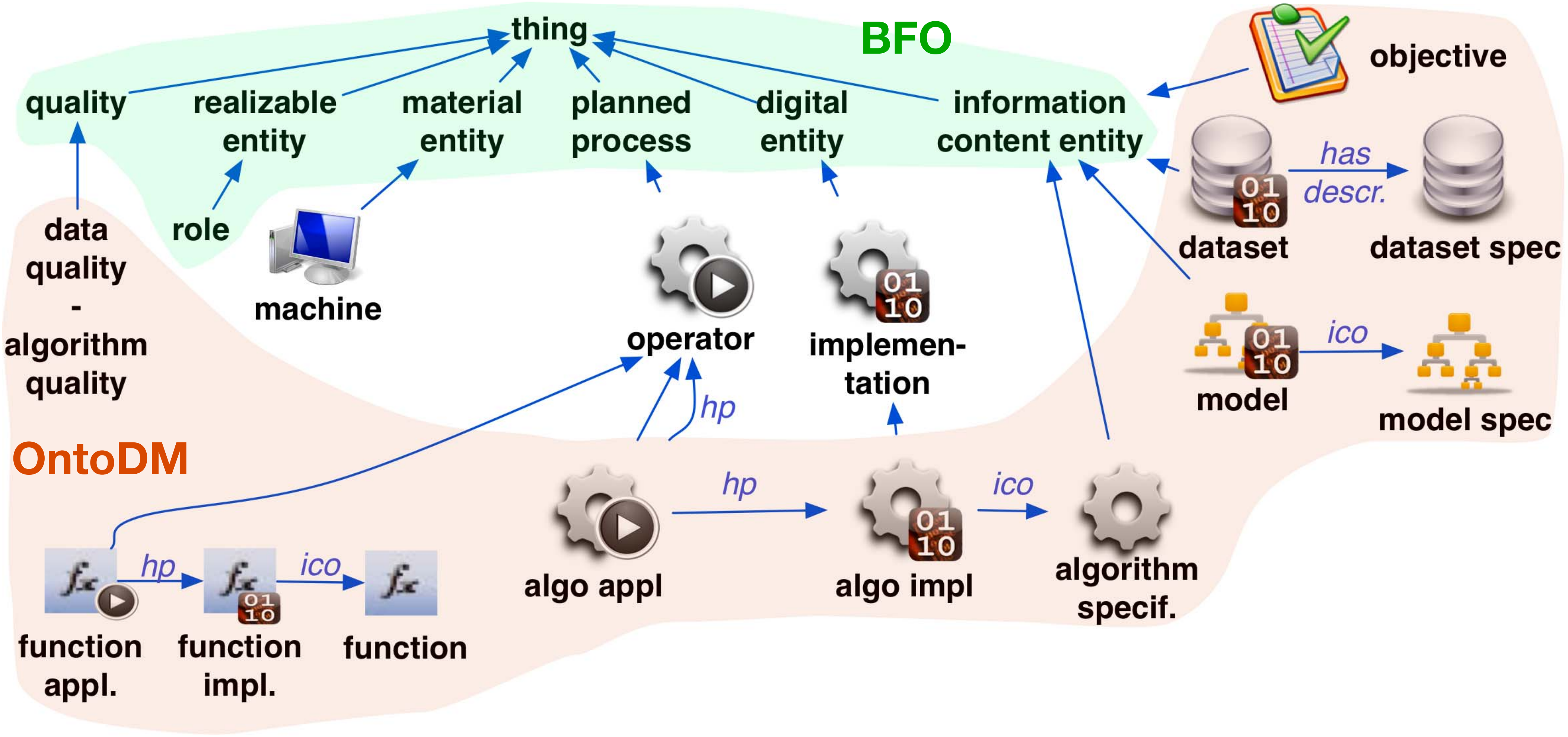
BFO: accepted top-level classes



BFO: accepted top-level classes

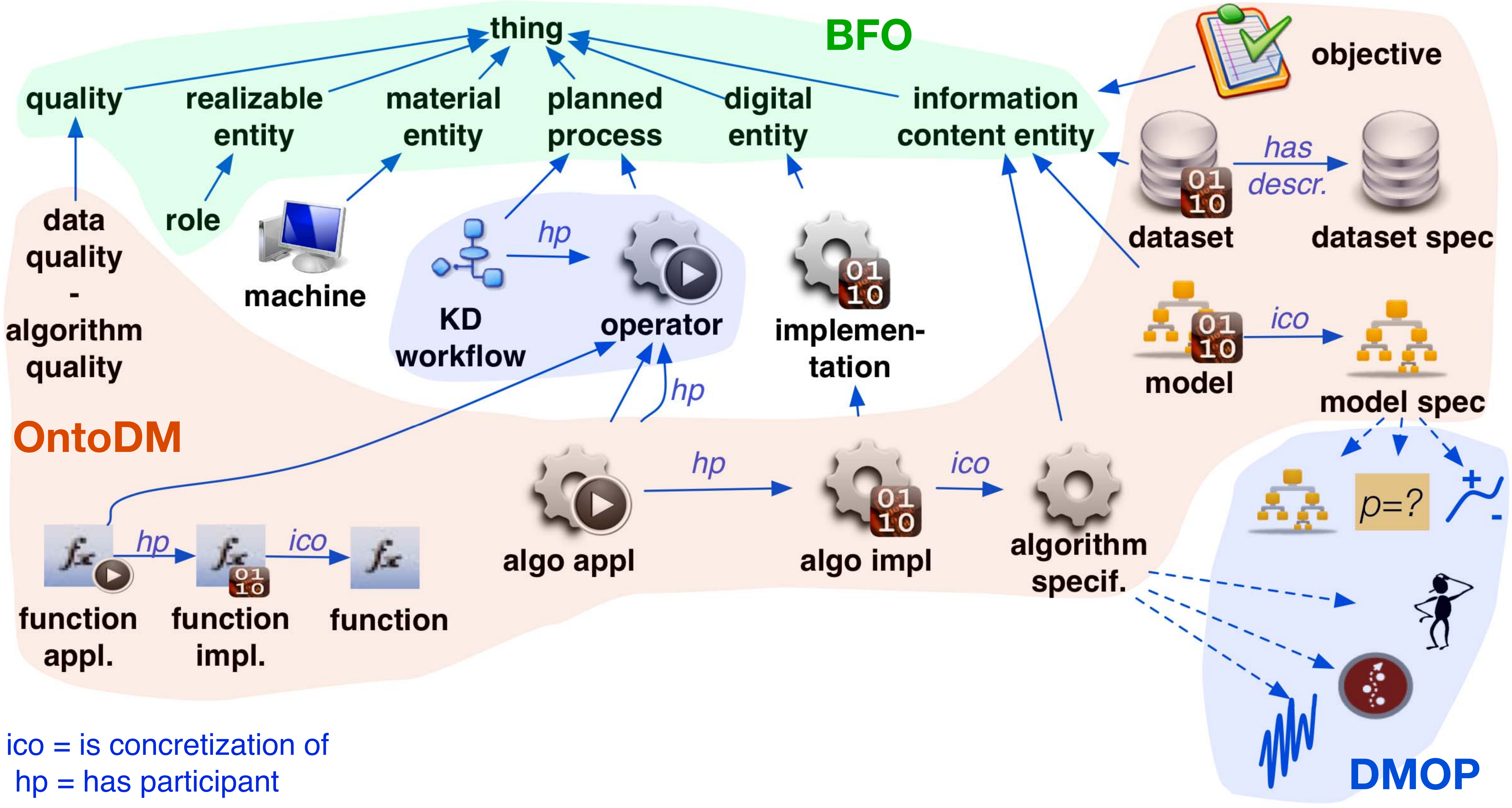


OntoDM: top-level DM concepts



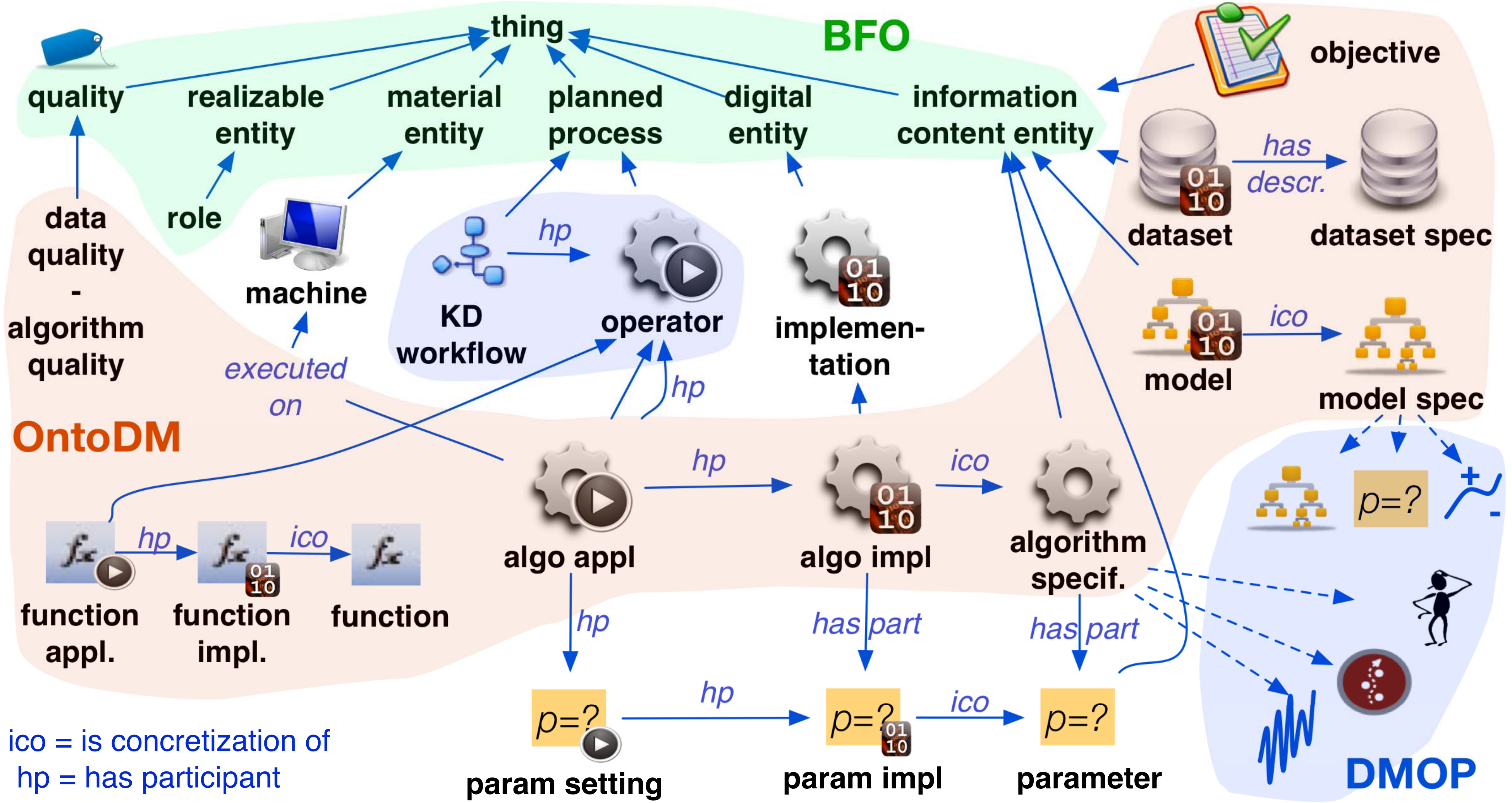
ico = is concretization of
hp = has participant

DMO: operators, learning mechanisms



ico = is concretization of
hp = has participant

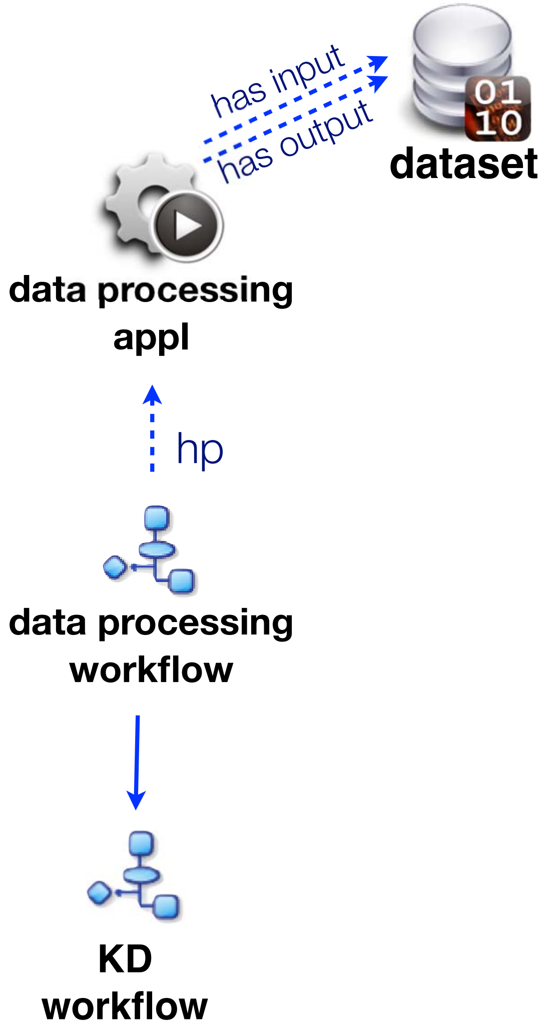
Exposé: top level classes



ico = is concretization of
hp = has participant

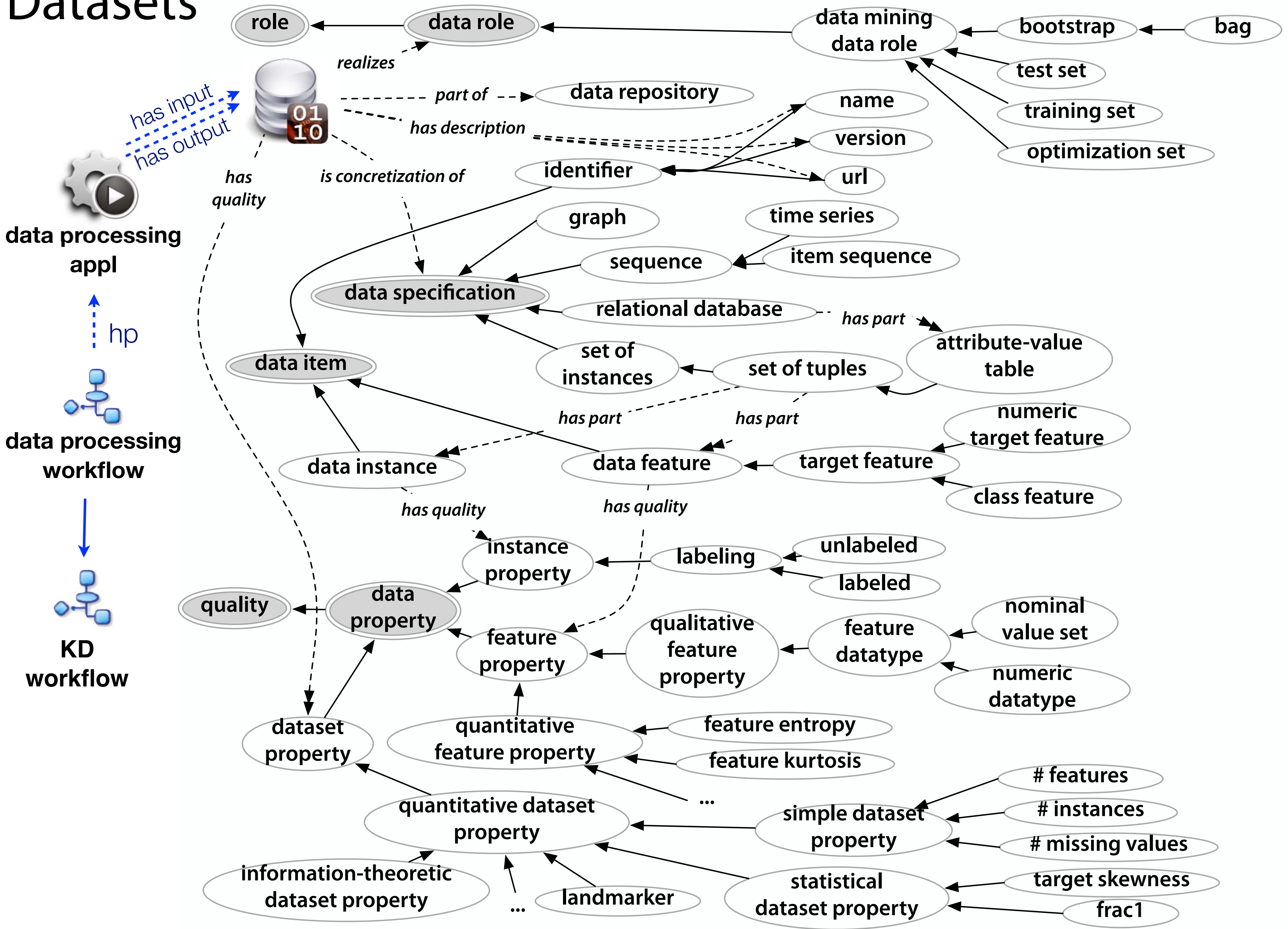
Other aspects

Datasets

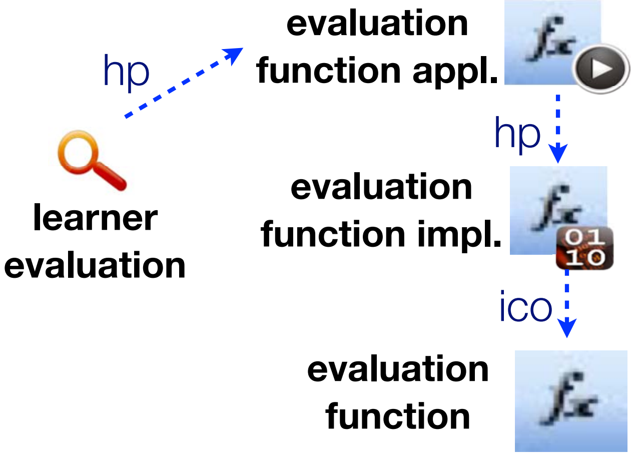


has input
has output
dataset
data processing
appl
hp
data processing
workflow
KD
workflow

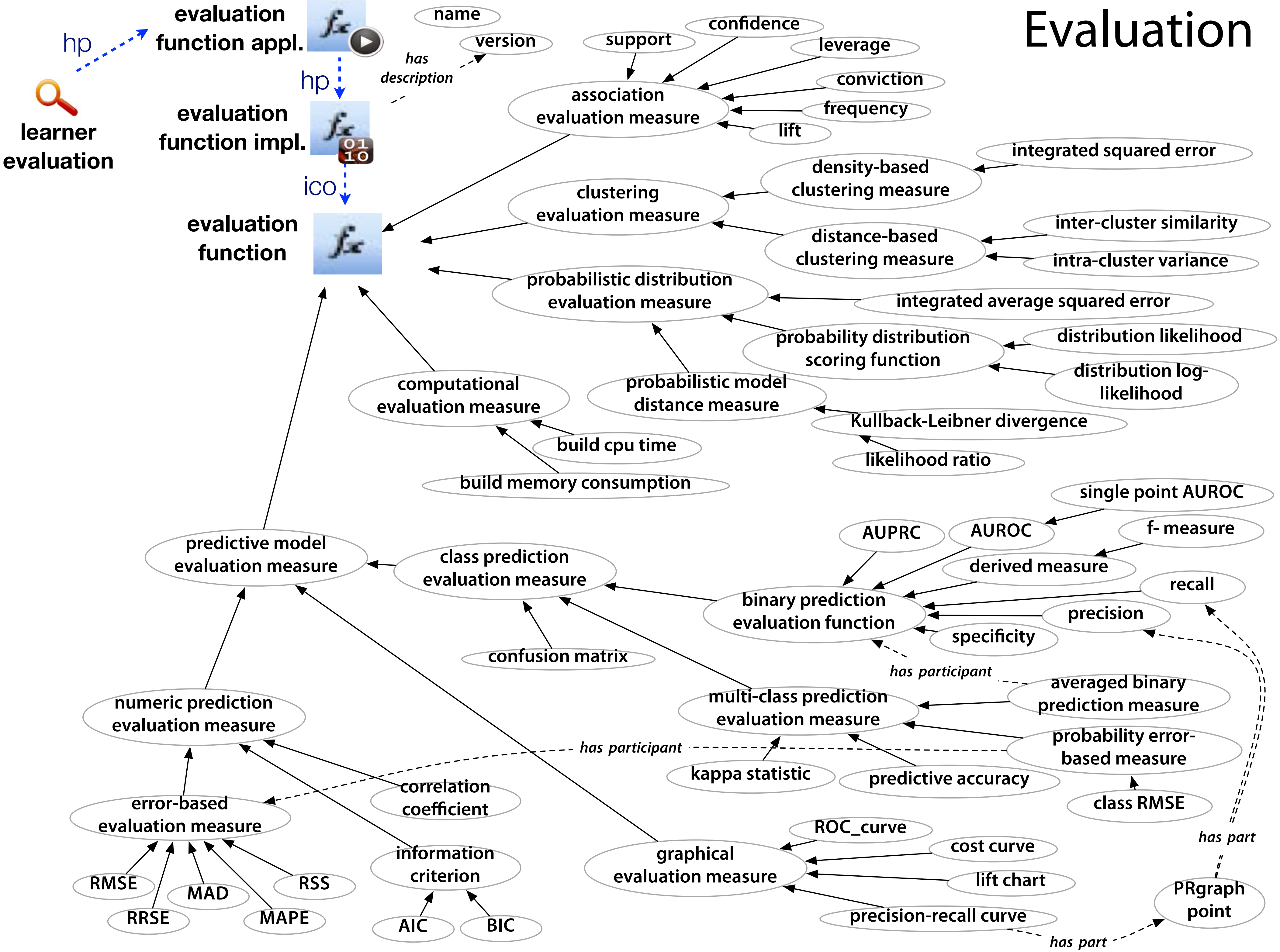
Datasets


role
data role
data mining data role
bootstrap
bag
test set
training set
optimization set
realizes
part of
data repository
has description
name
version
url
identifier
has input
has output
has quality
is concretization of
data processing appl
hp
data processing workflow
KD workflow
graph
time series
sequence
item sequence
data specification
relational database
has part
attribute-value table
data item
set of instances
set of tuples
has part
has part
numeric target feature
data instance
data feature
target feature
class feature
has quality
has quality
instance property
labeling
unlabeled
labeled
quality
data property
feature property
qualitative feature property
feature datatype
nominal value set
numeric datatype
dataset property
quantitative feature property
feature entropy
feature kurtosis
...
# features
# instances
# missing values
simple dataset property
quantitative dataset property
information-theoretic dataset property
...
landmarker
statistical dataset property
target skewness
frac1

Evaluation


evaluation function appl.
hp
learner evaluation
hp
evaluation function impl.
ico
evaluation function

Evaluation



learner evaluation

hp

evaluation function appl.

hp

evaluation function impl.

ico

evaluation function

name

version

has description

association evaluation measure

support

confidence

leverage

conviction

frequency

lift

clustering evaluation measure

density-based clustering measure

integrated squared error

distance-based clustering measure

inter-cluster similarity

intra-cluster variance

probabilistic distribution evaluation measure

integrated average squared error

probability distribution scoring function

distribution likelihood

distribution log-likelihood

probabilistic model distance measure

Kullback-Leibner divergence

likelihood ratio

computational evaluation measure

build cpu time

build memory consumption

predictive model evaluation measure

class prediction evaluation measure

confusion matrix

binary prediction evaluation function

AUPRC

AUROC

single point AUROC

f- measure

derived measure

recall

precision

specificity

has participant

multi-class prediction evaluation measure

averaged binary prediction measure

probability error-based measure

class RMSE

kappa statistic

predictive accuracy

has participant

numeric prediction evaluation measure

error-based evaluation measure

RMSE

RRSE

MAD

MAPE

RSS

correlation coefficient

information criterion

AIC

BIC

graphical evaluation measure

ROC_curve

cost curve

lift chart

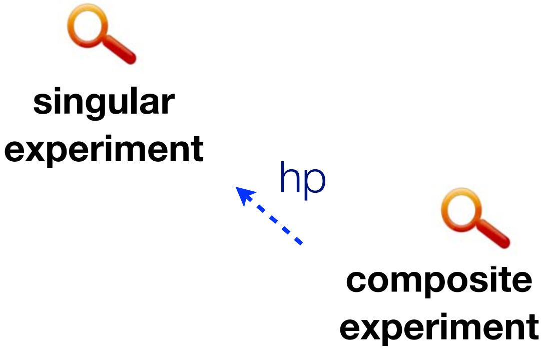
precision-recall curve

has part

has part

PRgraph point

Experiment context

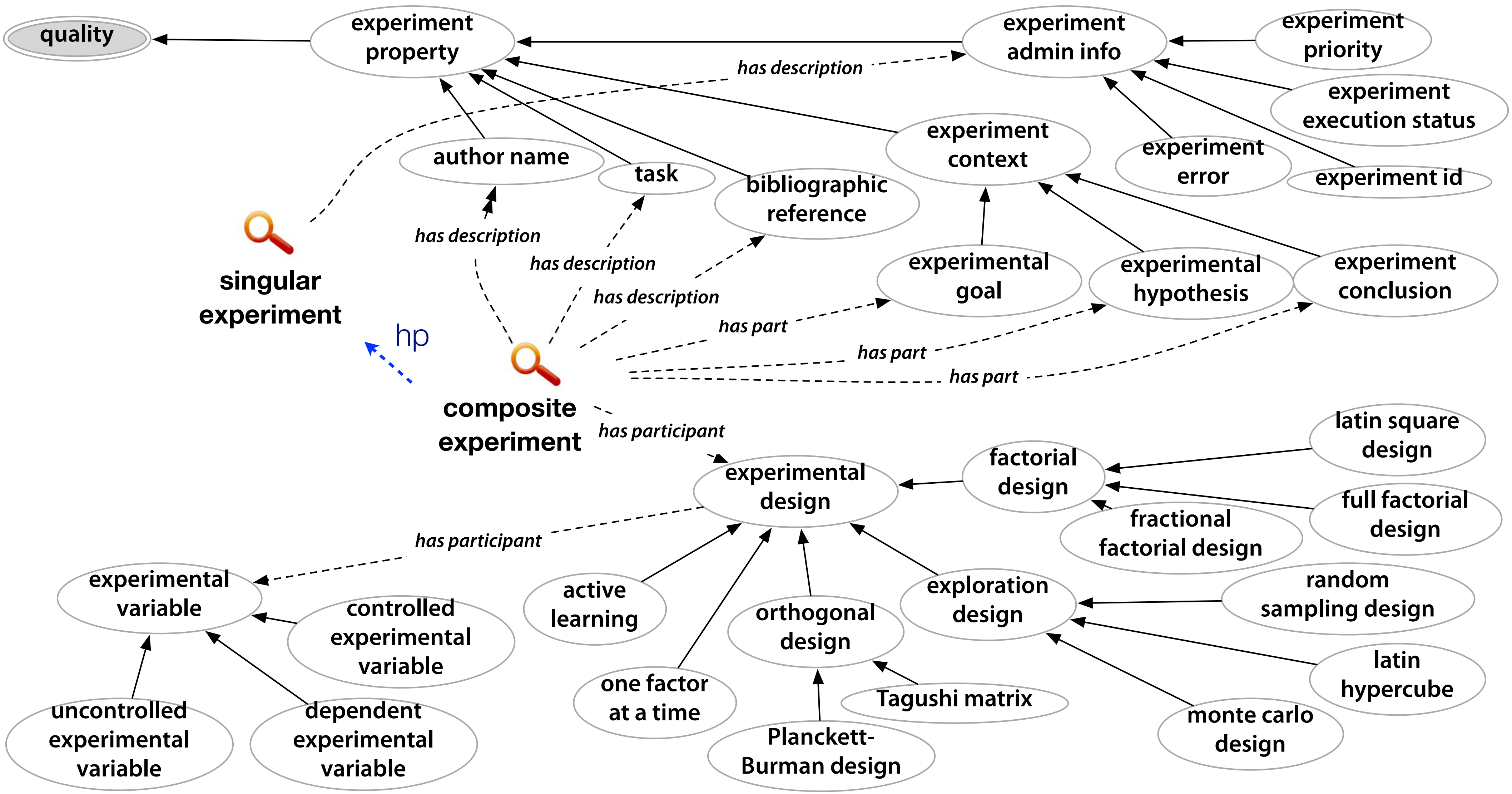


singular experiment

hp

composite experiment

Experiment context



Exposé: final notes

- In total 860 classes, 32 properties (from RO + DMOP)
- Individuals: all algorithms, preprocessors, evaluation from WEKA
  - actually stored in experiment database
  - should be programmatically added (and updated)
- Written in OWL-DL, using Protégé 4.0
- Can be browsed at:
  - http://expdb.cs.kuleuven.be/expdb/expose.owl
  - http://www.e-lico.eu/OWLBrowser2/manage/

Use Cases

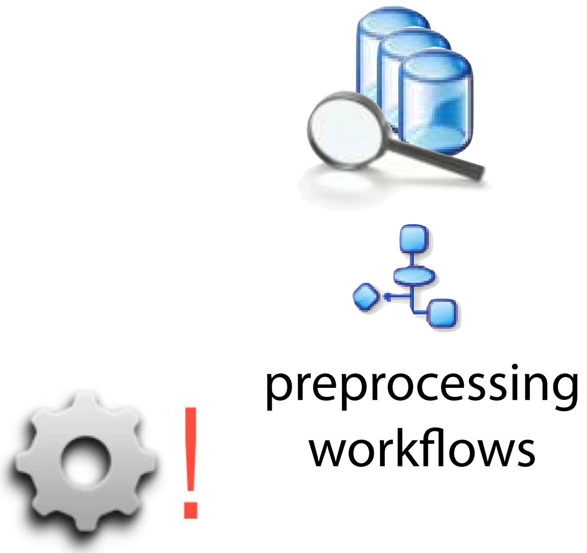
Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable

Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable

new algorithm

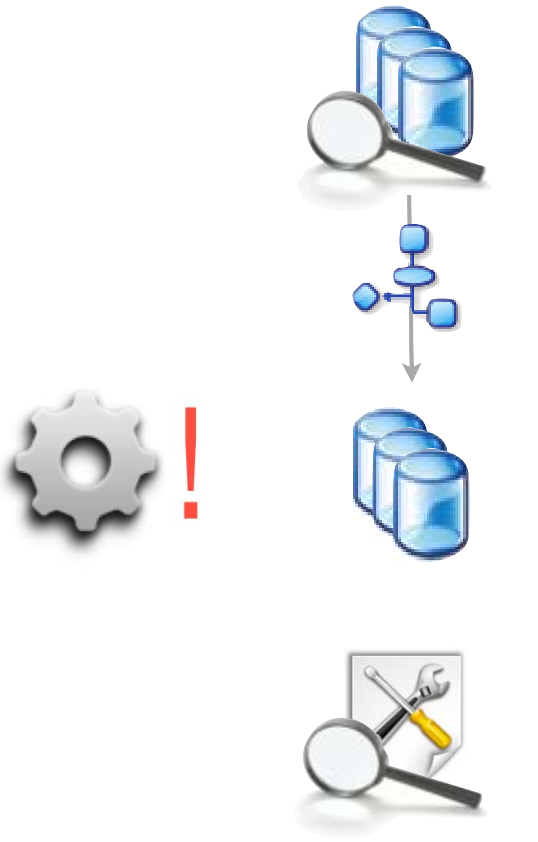
Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable

Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable



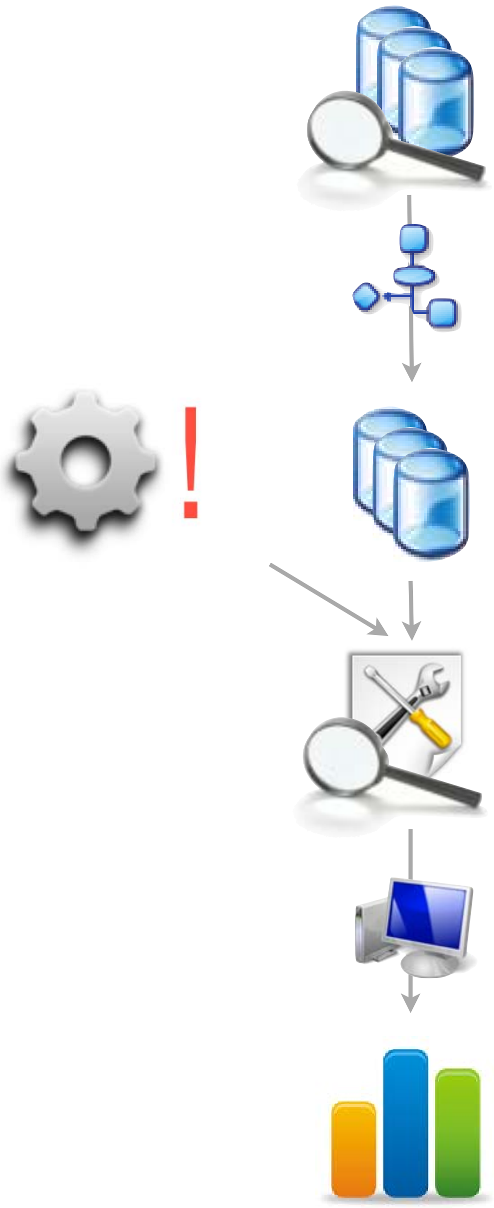
preprocessing workflows

Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable

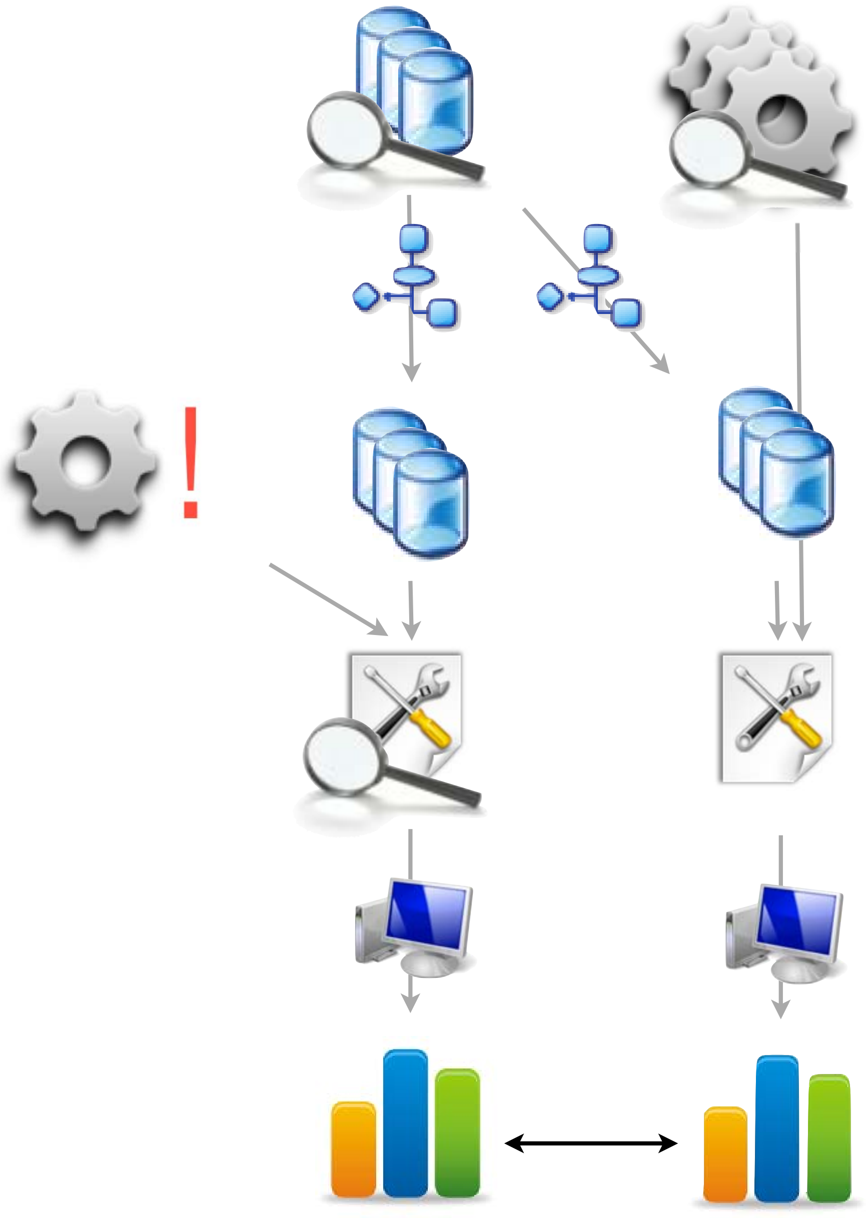


evaluation procedures

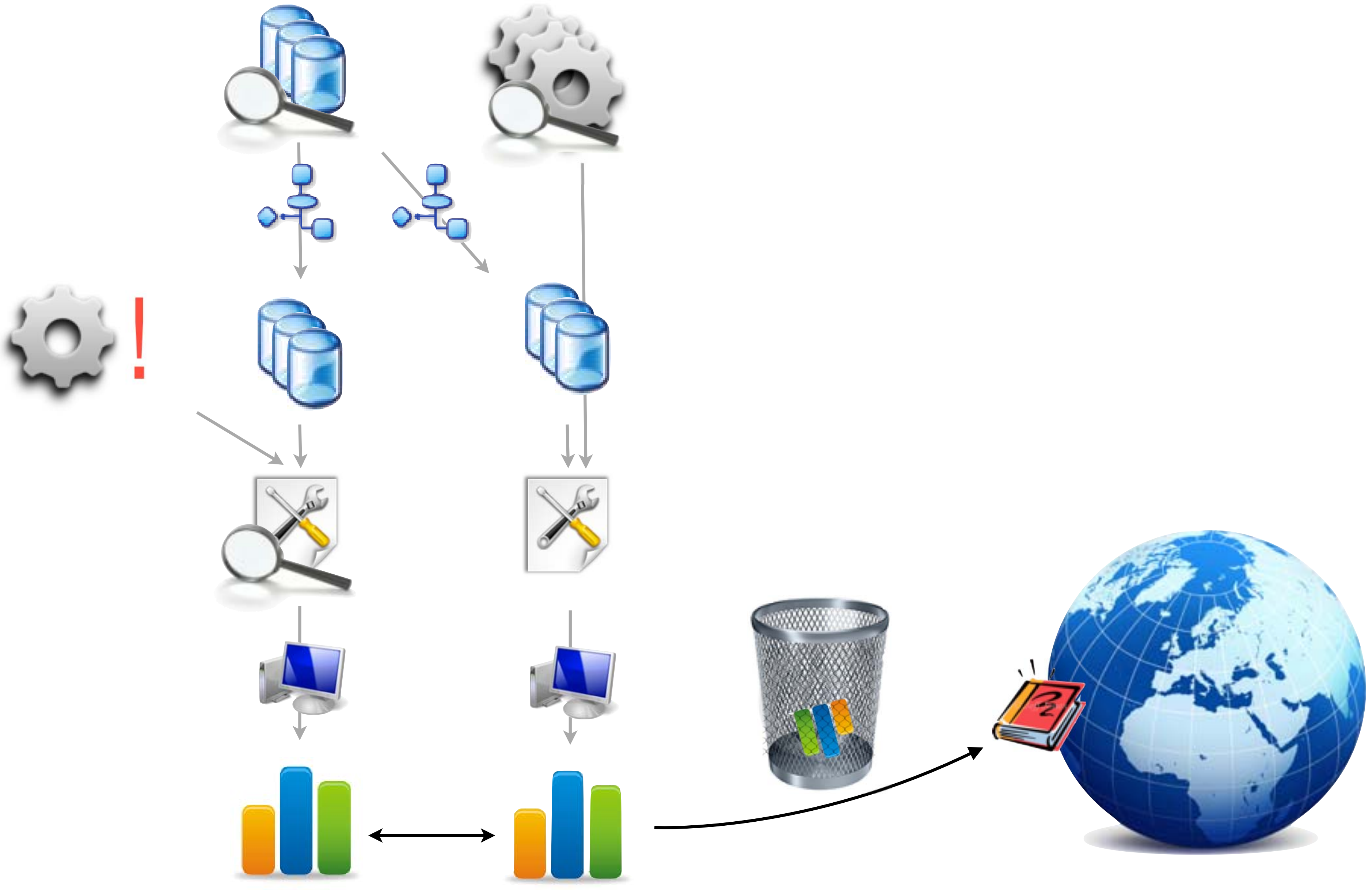
Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable



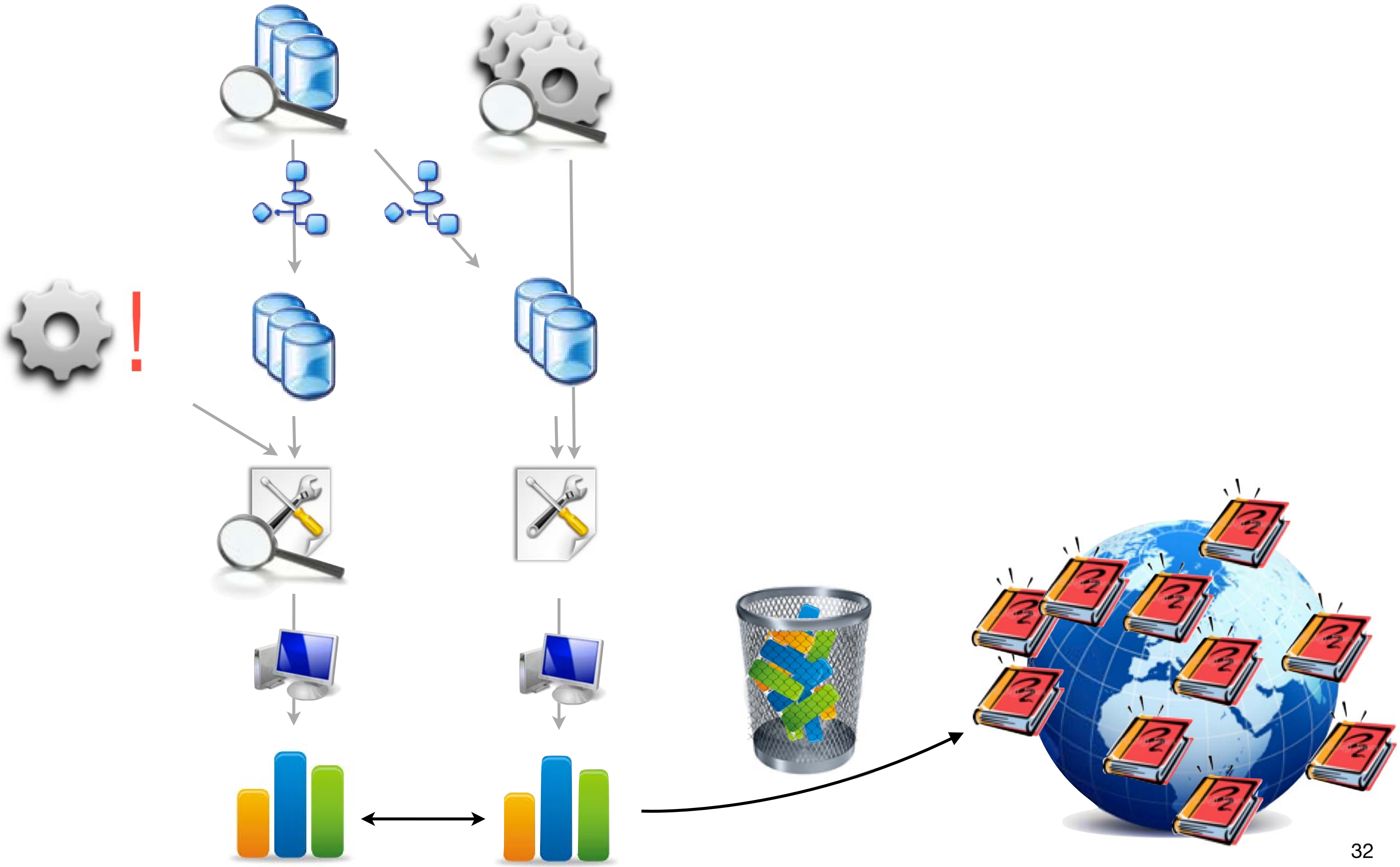
Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable



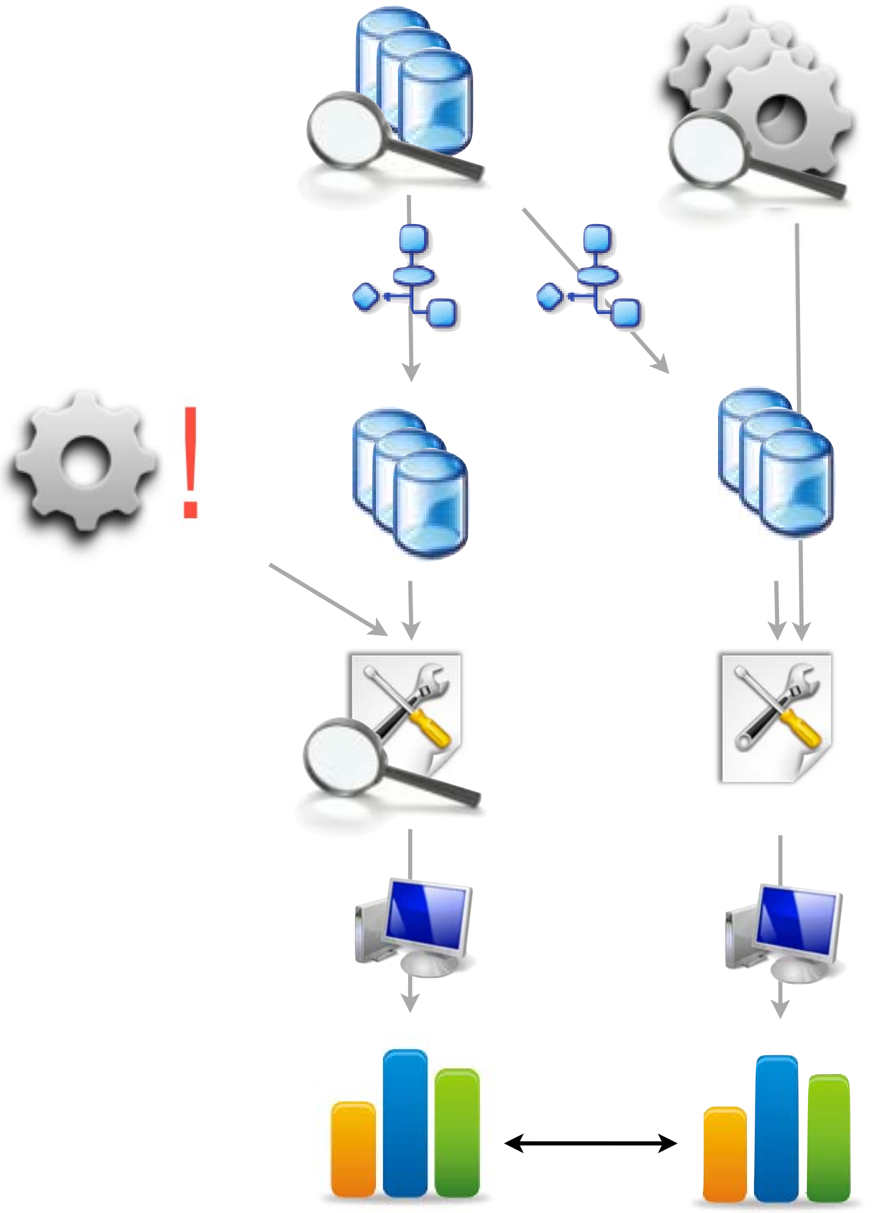
Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable



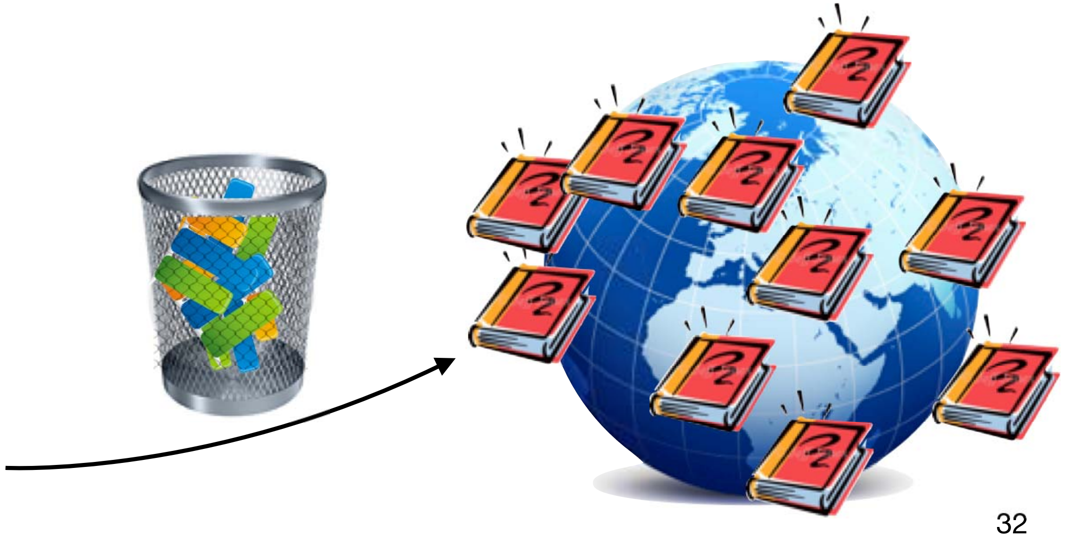
Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable



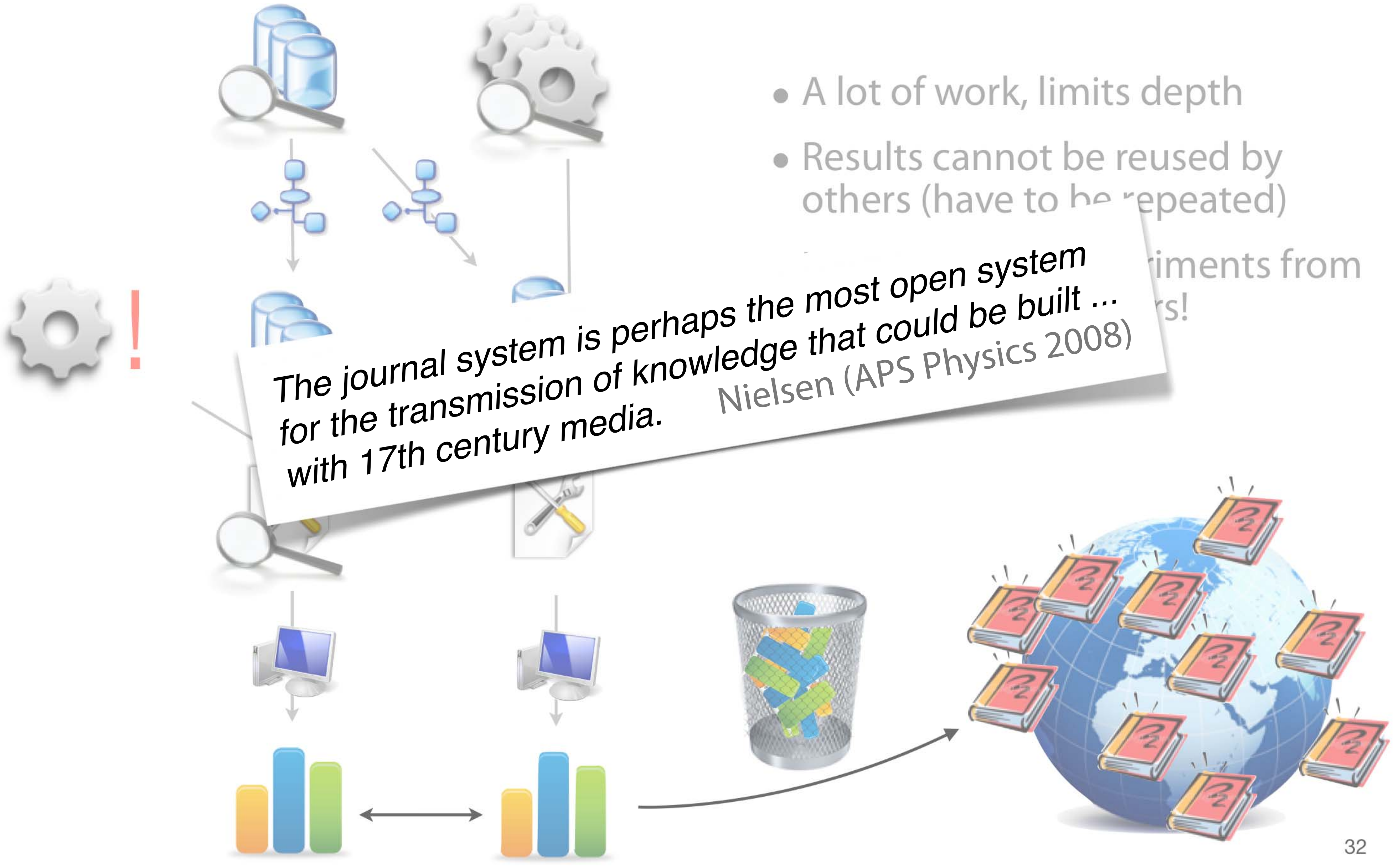
Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable



- A lot of work, limits depth
- Results cannot be reused by others (have to be repeated)
- Hard to repeat experiments from descriptions in papers!



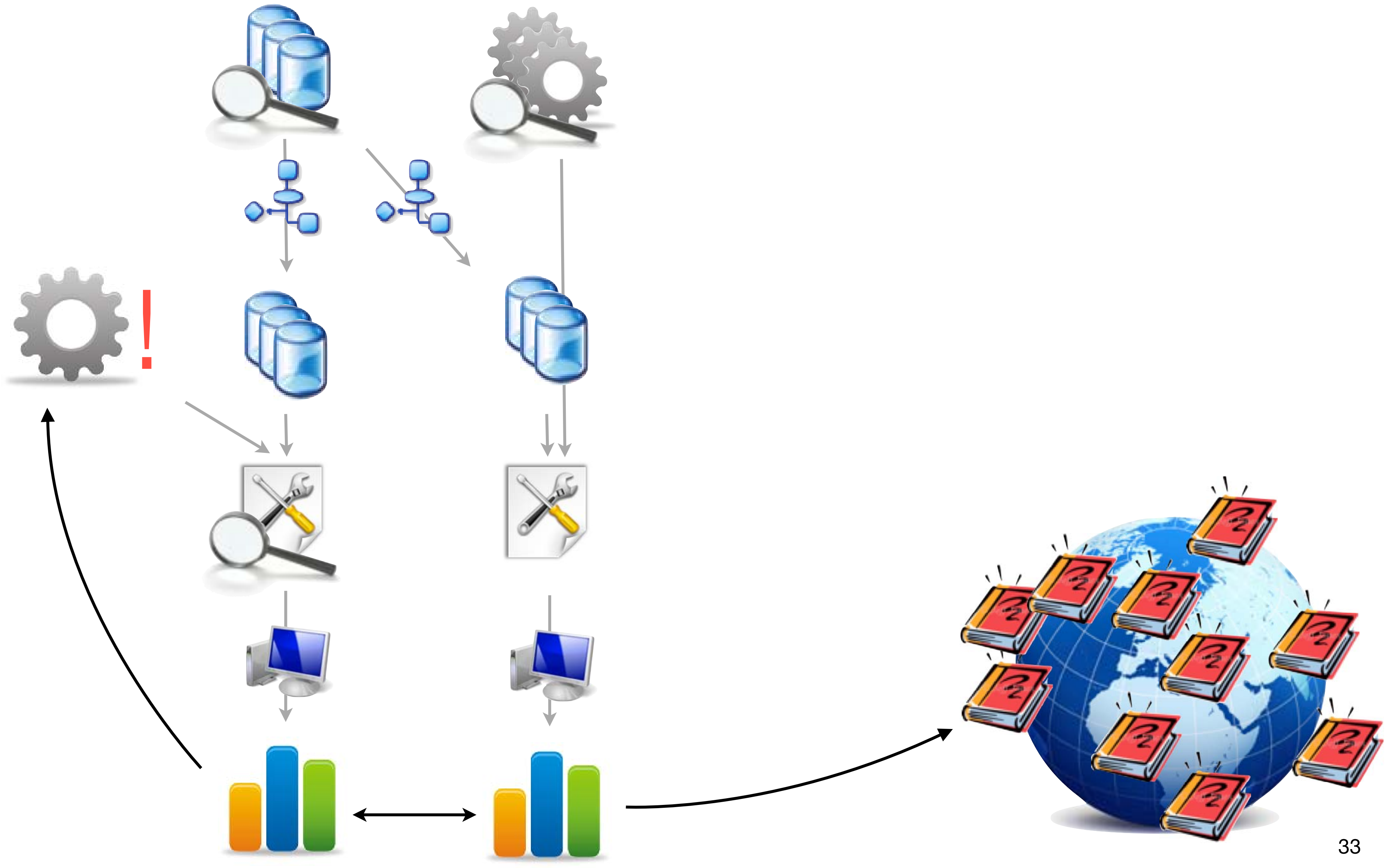
Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable



- A lot of work, limits depth
- Results cannot be reused by others (have to be repeated)
- iments from rs!

> *The journal system is perhaps the most open system for the transmission of knowledge that could be built ... with 17th century media.* Nielsen (APS Physics 2008)

Data mining as an e-science
Ontologies: experiments shared, run automatically

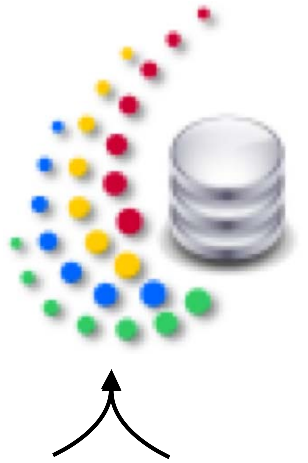


Data mining as an e-science
Ontologies: experiments shared, run automatically

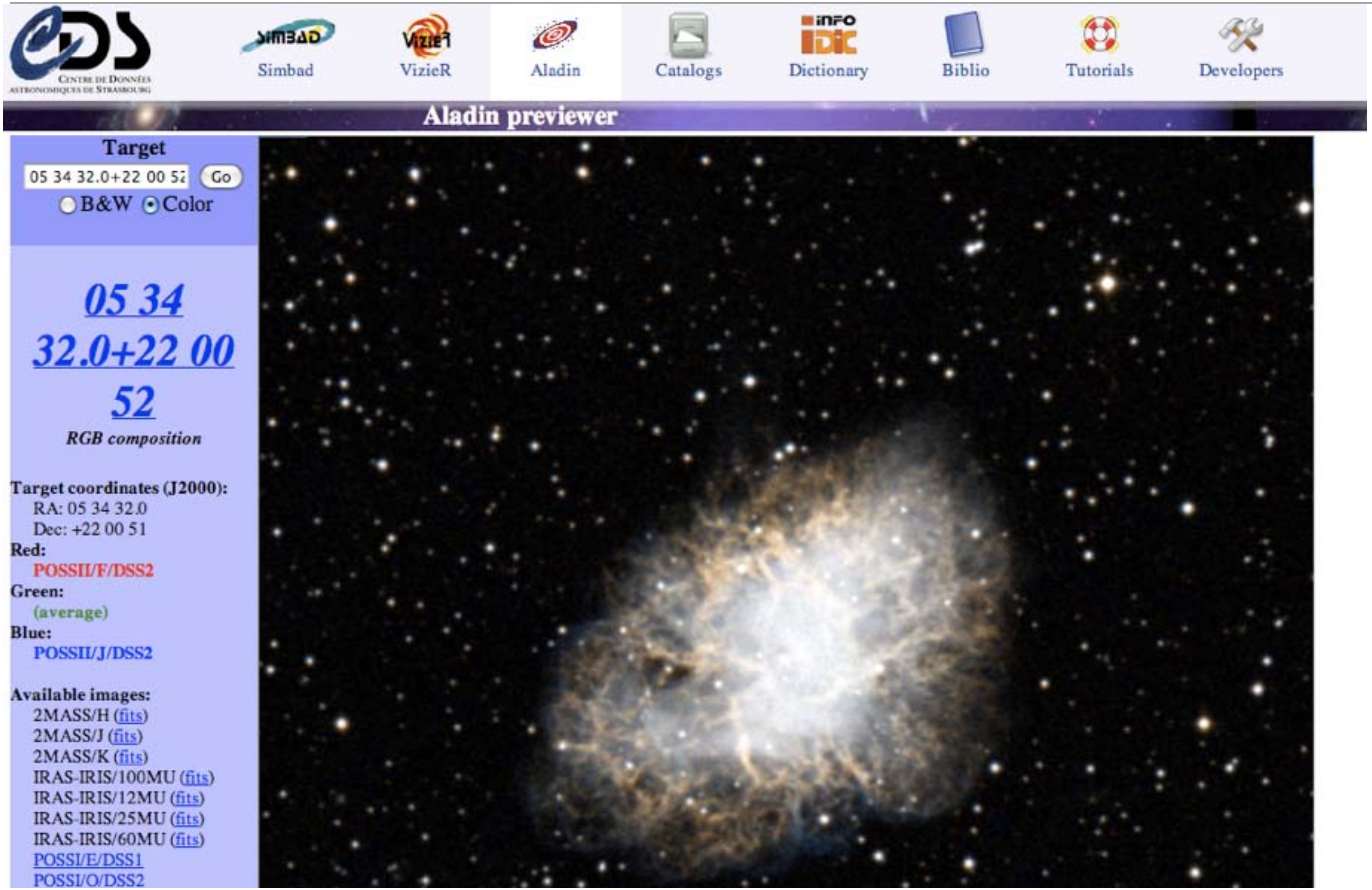
- Share experiments
  - Internet = large, collaborative workspace

Data mining as an e-science
Ontologies: experiments shared, run automatically

- Share experiments
  - Internet = large, collaborative workspace
- Store them in ***experiment databases***
  - Ensure reproducibility
  - Reuse millions of prior experiments
  - Use all info on algorithms, datasets
  - Results universally accessible and useful



e-Sciences
Astrophysics: Virtual Observatories



Target

05 34 32.0+22 00 52

○ B&W ● Color

05 34 32.0+22 00 52

RGB composition

Target coordinates (J2000):
RA: 05 34 32.0
Dec: +22 00 51

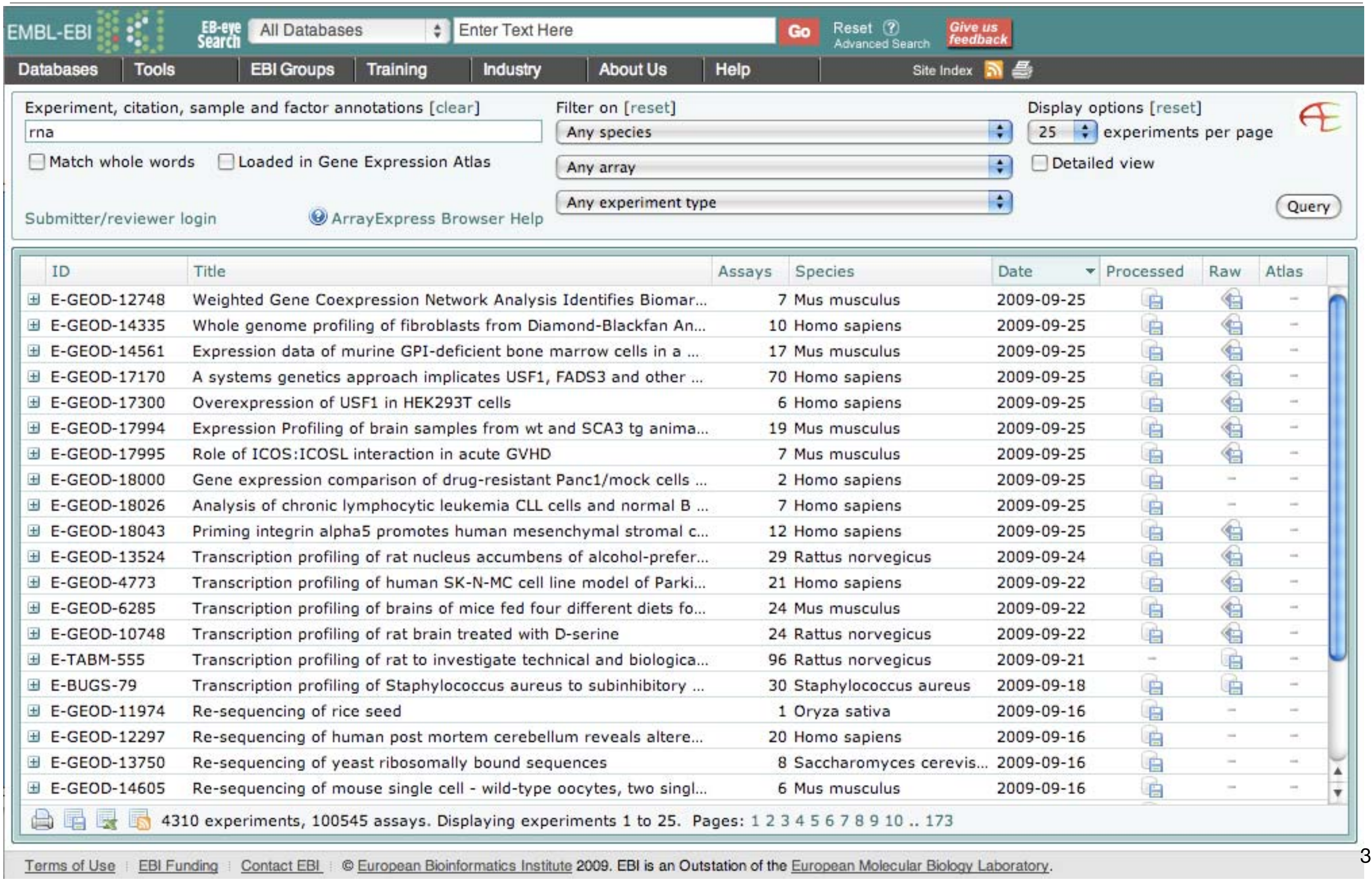
Red:
POSSII/F/DSS2

Green:
(average)

Blue:
POSSII/J/DSS2

Available images:
2MASS/H (fits)
2MASS/J (fits)
2MASS/K (fits)
IRAS-IRIS/100MU (fits)
IRAS-IRIS/12MU (fits)
IRAS-IRIS/25MU (fits)
IRAS-IRIS/60MU (fits)
POSSI/E/DSS1
POSSI/O/DSS2

e-Sciences
Bio-informatics: Micro-array Databases



Experiment, citation, sample and factor annotations [clear]

rna

□ Match whole words □ Loaded in Gene Expression Atlas

Submitter/reviewer login

ArrayExpress Browser Help

Filter on [reset]

Any species

Any array

Any experiment type

Display options [reset]

25 experiments per page

□ Detailed view

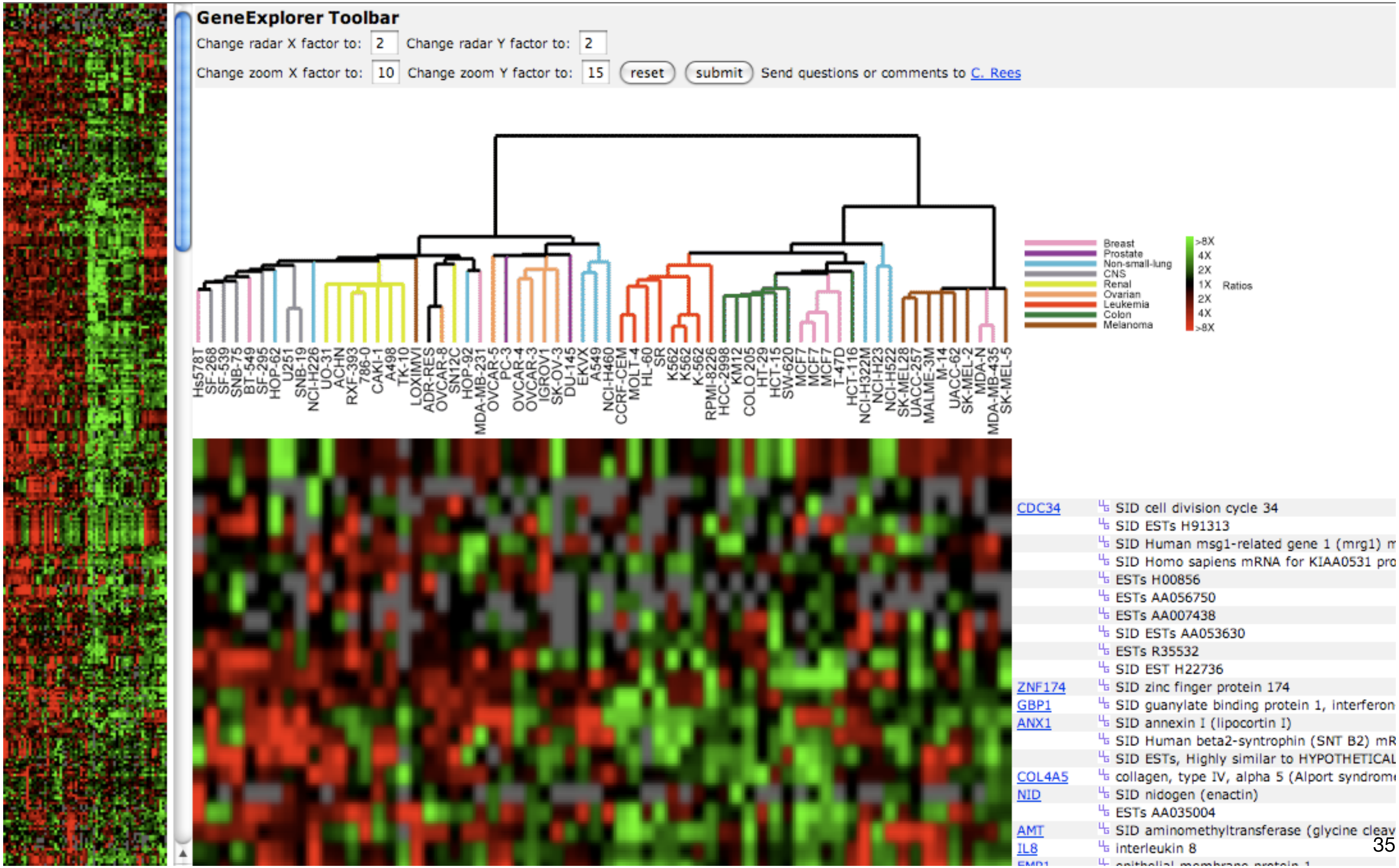
Query

	ID	Title	Assays	Species	Date	Processed	Raw	Atlas
⊞	E-GEOD-12748	Weighted Gene Coexpression Network Analysis Identifies Biomar...	7	Mus musculus	2009-09-25	▤	▤	-
⊞	E-GEOD-14335	Whole genome profiling of fibroblasts from Diamond-Blackfan An...	10	Homo sapiens	2009-09-25	▤	▤	-
⊞	E-GEOD-14561	Expression data of murine GPI-deficient bone marrow cells in a ...	17	Mus musculus	2009-09-25	▤	▤	-
⊞	E-GEOD-17170	A systems genetics approach implicates USF1, FADS3 and other ...	70	Homo sapiens	2009-09-25	▤	▤	-
⊞	E-GEOD-17300	Overexpression of USF1 in HEK293T cells	6	Homo sapiens	2009-09-25	▤	▤	-
⊞	E-GEOD-17994	Expression Profiling of brain samples from wt and SCA3 tg anima...	19	Mus musculus	2009-09-25	▤	▤	-
⊞	E-GEOD-17995	Role of ICOS:ICOSL interaction in acute GVHD	7	Mus musculus	2009-09-25	▤	▤	-
⊞	E-GEOD-18000	Gene expression comparison of drug-resistant Panc1/mock cells ...	2	Homo sapiens	2009-09-25	▤	-	-
⊞	E-GEOD-18026	Analysis of chronic lymphocytic leukemia CLL cells and normal B ...	7	Homo sapiens	2009-09-25	▤	-	-
⊞	E-GEOD-18043	Priming integrin alpha5 promotes human mesenchymal stromal c...	12	Homo sapiens	2009-09-25	▤	▤	-
⊞	E-GEOD-13524	Transcription profiling of rat nucleus accumbens of alcohol-prefer...	29	Rattus norvegicus	2009-09-24	▤	▤	-
⊞	E-GEOD-4773	Transcription profiling of human SK-N-MC cell line model of Parki...	21	Homo sapiens	2009-09-22	▤	▤	-
⊞	E-GEOD-6285	Transcription profiling of brains of mice fed four different diets fo...	24	Mus musculus	2009-09-22	▤	▤	-
⊞	E-GEOD-10748	Transcription profiling of rat brain treated with D-serine	24	Rattus norvegicus	2009-09-22	▤	▤	-
⊞	E-TABM-555	Transcription profiling of rat to investigate technical and biologica...	96	Rattus norvegicus	2009-09-21	-	▤	-
⊞	E-BUGS-79	Transcription profiling of Staphylococcus aureus to subinhibitory ...	30	Staphylococcus aureus	2009-09-18	▤	▤	-
⊞	E-GEOD-11974	Re-sequencing of rice seed	1	Oryza sativa	2009-09-16	▤	-	-
⊞	E-GEOD-12297	Re-sequencing of human post mortem cerebellum reveals altere...	20	Homo sapiens	2009-09-16	▤	-	-
⊞	E-GEOD-13750	Re-sequencing of yeast ribosomally bound sequences	8	Saccharomyces cerevis...	2009-09-16	▤	-	-
⊞	E-GEOD-14605	Re-sequencing of mouse single cell - wild-type oocytes, two singl...	6	Mus musculus	2009-09-16	▤	-	-

4310 experiments, 100545 assays. Displaying experiments 1 to 25. Pages: 1 2 3 4 5 6 7 8 9 10 .. 173

Terms of Use | EBI Funding | Contact EBI | © European Bioinformatics Institute 2009. EBI is an Outstation of the European Molecular Biology Laboratory.

e-Sciences
Bio-informatics: Micro-array Databases



GeneExplorer Toolbar

Change radar X factor to: 2 Change radar Y factor to: 2

Change zoom X factor to: 10 Change zoom Y factor to: 15 reset submit Send questions or comments to C. Rees

CDC34	SID cell division cycle 34
	SID ESTs H91313
	SID Human msg1-related gene 1 (mrg1) m
	SID Homo sapiens mRNA for KIAA0531 pro
	ESTs H00856
	ESTs AA056750
	ESTs AA007438
	SID ESTs AA053630
	ESTs R35532
	SID EST H22736
ZNF174	SID zinc finger protein 174
GBP1	SID guanylate binding protein 1, interferon
ANX1	SID annexin I (lipocortin I)
	SID Human beta2-syntrophin (SNT B2) mR
	SID ESTs, Highly similar to HYPOTHETICAL
COL4A5	collagen, type IV, alpha 5 (Alport syndrom
NID	SID nidogen (enactin)
	ESTs AA035004
AMT	SID aminomethyltransferase (glycine cleav
IL8	interleukin 8
EMP1	epithelial membrane protein 1

Collaborative Experimentation
Why?

Collaborative Experimentation
Why?

Reproducibility

Good science

Collaborative Experimentation
Why?

Reproducibility

Good science

Quick, easy analysis

Querying: Answer questions

Test hypotheses

Collaborative Experimentation
Why?

Reproducibility

Good science

Quick, easy analysis

Querying: Answer questions

Test hypotheses

Reuse

♻ Save time & energy
(e.g. benchmarking)

Collaborative Experimentation
Why?

Reproducibility

Good science

Quick, easy analysis

Querying: Answer questions

Test hypotheses

Reuse

Save time & energy

(e.g. benchmarking)

Generalizability:

Plug into prior results: larger studies

Collaborative Experimentation
Why?

Reproducibility

Good science

Quick, easy analysis

Querying: Answer questions
Test hypotheses

Integration

Data mining tools
import/export

Reuse

Save time & energy
(e.g. benchmarking)

Generalizability:

Plug into prior results: larger studies

Collaborative Experimentation
Why?

Reproducibility

Good science

Quick, easy analysis

Querying: Answer questions

Test hypotheses

Integration

Data mining tools

import/export

Reuse

♻ Save time & energy

(e.g. benchmarking)

Reference

'Map' of known approaches

Compare to state-of-the-art

Includes negative results

Generalizability:

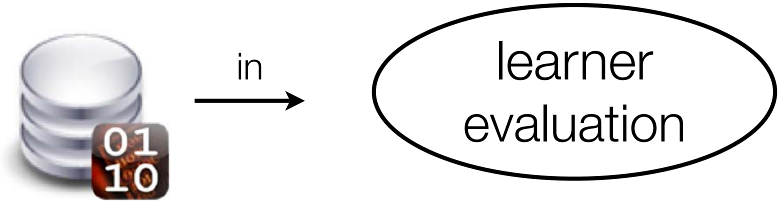
Plug into prior results: larger studies

Use Case 1

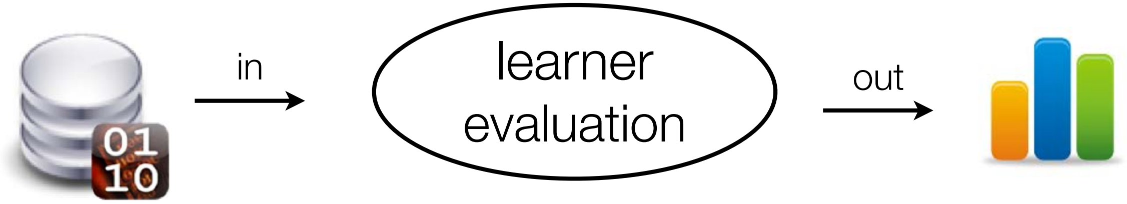
Describe experiments in a common language
-> sharing or running experiments on grid

Use Exposé to define common language: ExpML

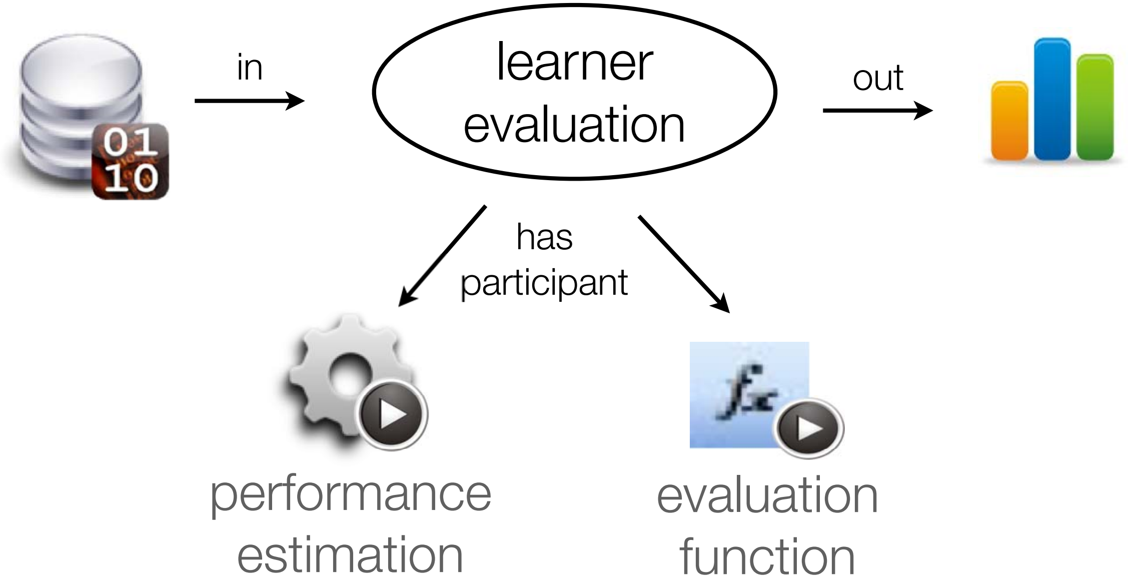
Use Exposé to define common language: ExpML



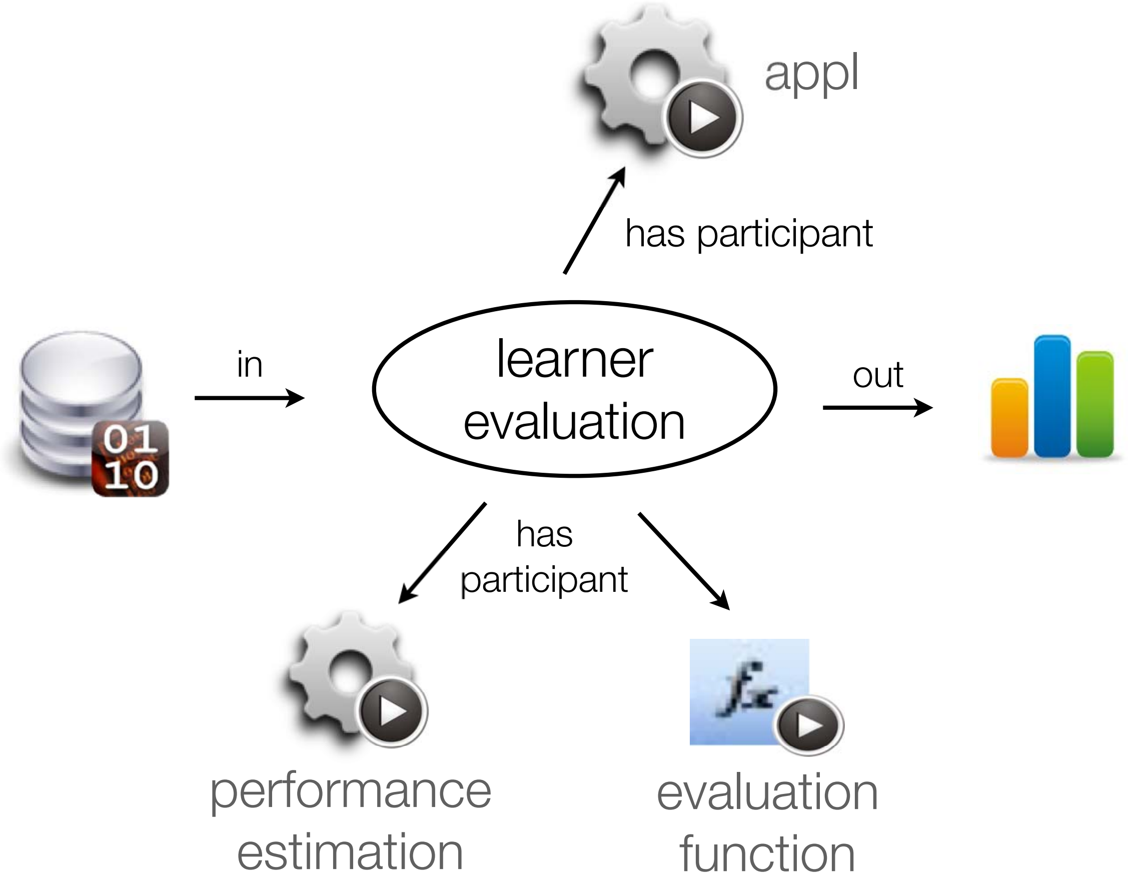
Use Exposé to define common language: ExpML



Use Exposé to define common language: ExpML



Use Exposé to define common language: ExpML



appl

has participant

in

learner evaluation

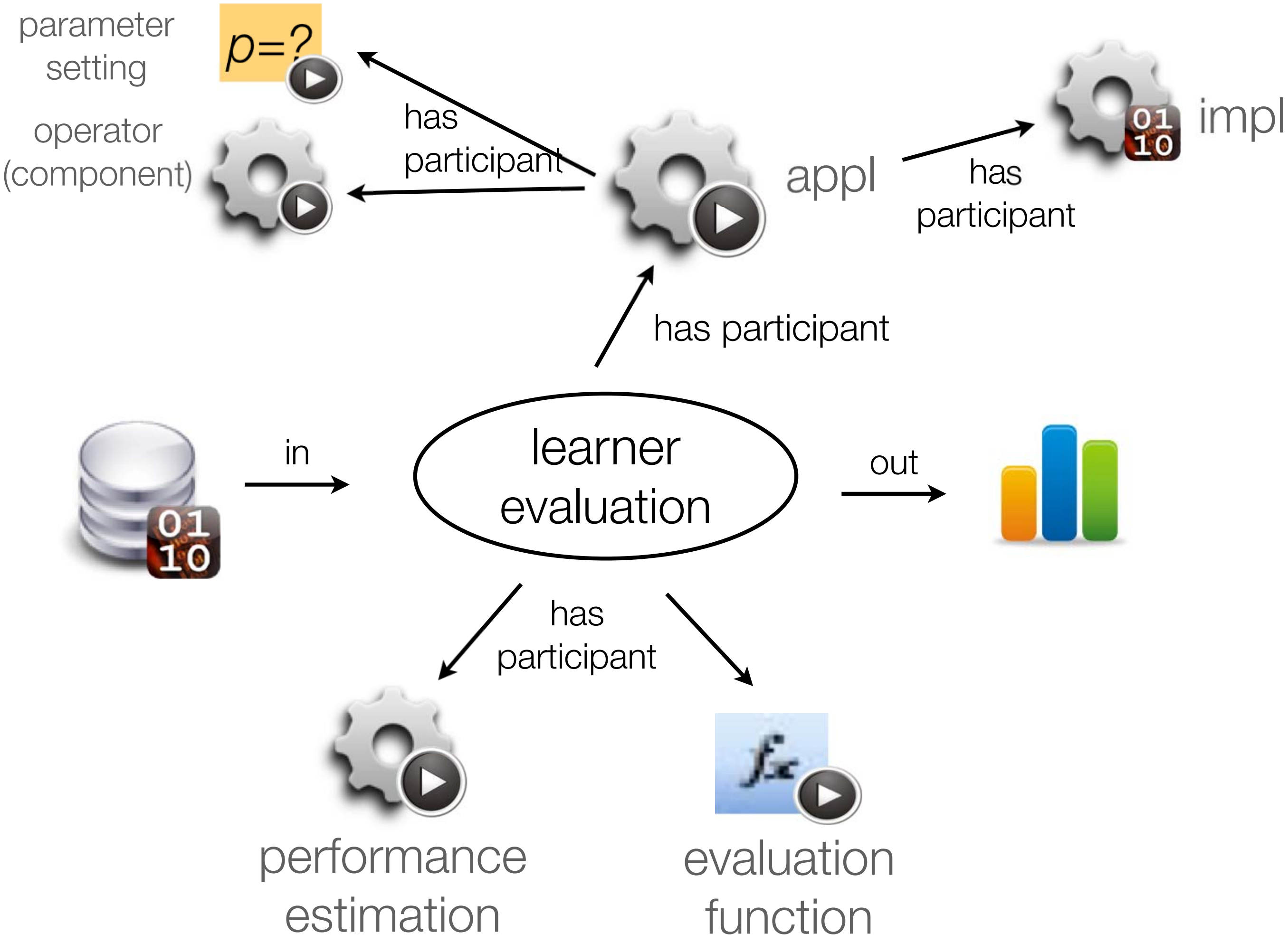
out

has participant

performance estimation

evaluation function

Use Exposé to define common language: ExpML



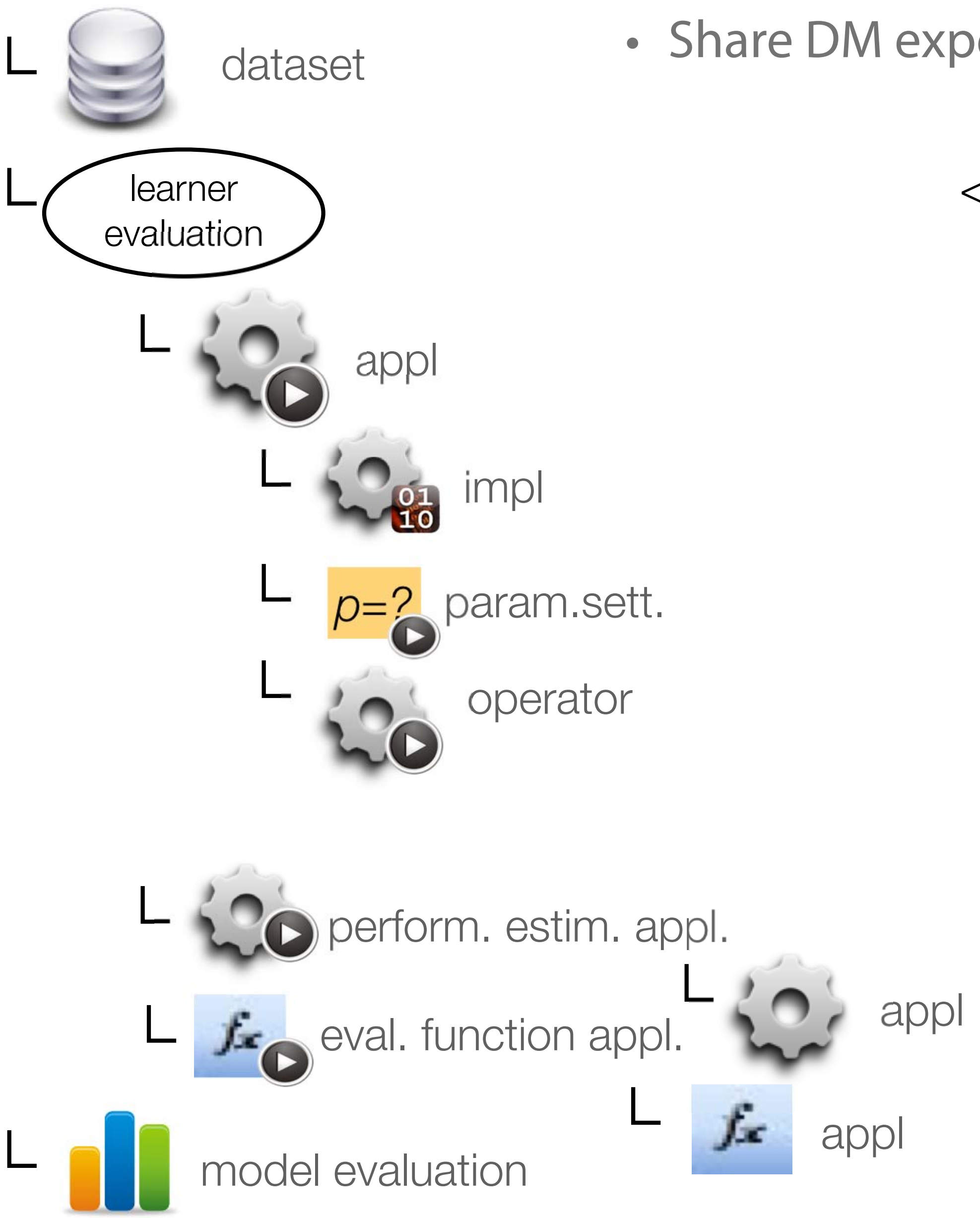
ExpML: a markup language for DM experiments

- Share DM experiments, XML-based

- └ dataset
- └ learner evaluation
    - └ appl
        - └ impl
        - └ param.sett.
        - └ operator
    - └ perform. estim. appl.
    - └ eval. function appl.
        - └ appl
        - └ appl
- └ model evaluation

ExpML: a markup language for DM experiments

- Share DM experiments, XML-based



└ dataset

└ learner evaluation

└ appl

└ impl

└ param.sett.

└ operator

└ perform. estim. appl.

└ eval. function appl.

└ appl

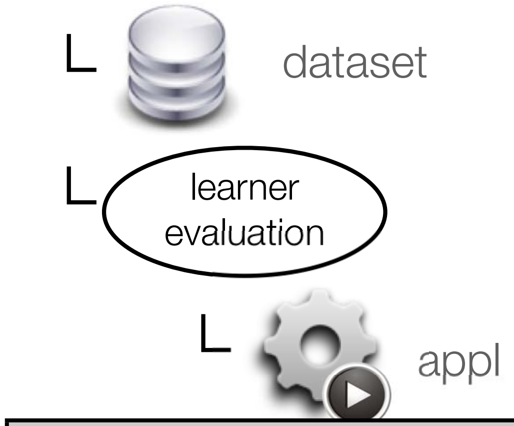
└ appl

└ model evaluation

```
<expml>
    <dataset id='d1'>
    <learner evaluation id='e1' input_data='d1'>
        <learner_appl>
            <learner_impl name=... version=...>
            <parameter_setting name='P' value='100'/>
            <learner_appl role= 'base-learner'>
                ...
        </learner_appl>
        <performance_estimation_appl>
        ...
        <model_evaluation_function_appl>
        ...
    </learner_evaluation>
    <model_evaluation_result output_of='e1'>
        <evaluation name='accuracy' value= '0.99'>
        ...
```

ExpML: a markup language for DM experiments

- Share DM experiments, XML-based



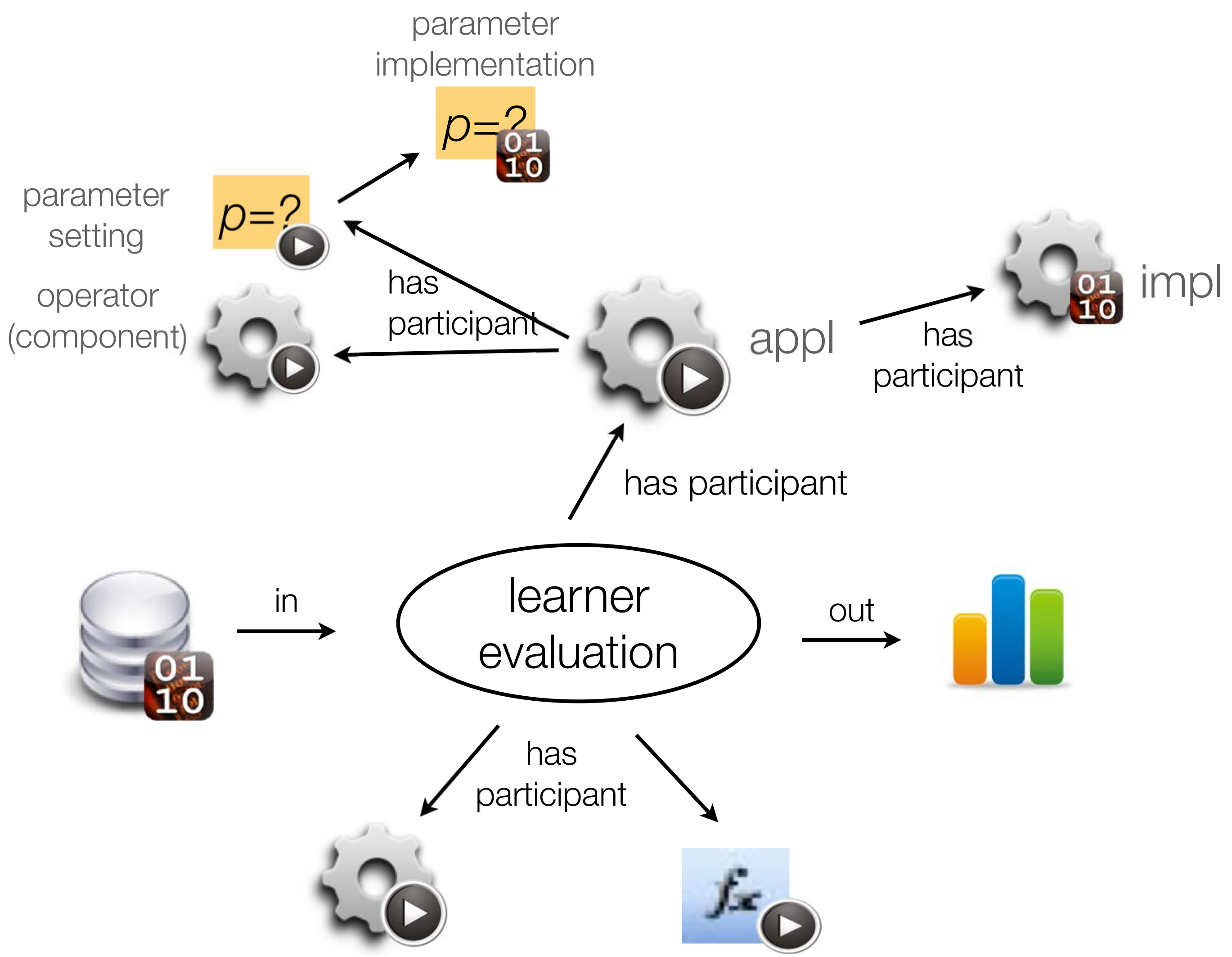
ontology	XML
has-part,has-participant	XML subelement
	(with role attribute)
has-description	(required) attribute
has-quality	`property’ subelement
is-concretization-of	implementation_of attr.
part-of	attributes
has-specific-input	input_data attribute
has-specified-output	output_of attribute

```
<expml>
    <dataset id=‘d1’>
    <learner evaluation id=‘e1’ input_data=‘d1’>
        <learner_appl>
            <learner_impl name=... version=...>
            <parameter_setting name=‘P’ value=‘100’/>
            <learner_appl role= ‘base-learner’>
                ...
        </learner_appl>
        <performance_estimation_appl>
        ...
        <model_evaluation_function_appl>
        ...
    </learner_evaluation>
    <model_evaluation_result output_of=‘e1’>
        <evaluation name=‘accuracy’ value= ‘0.99’>
        ...
```

Use Case 2

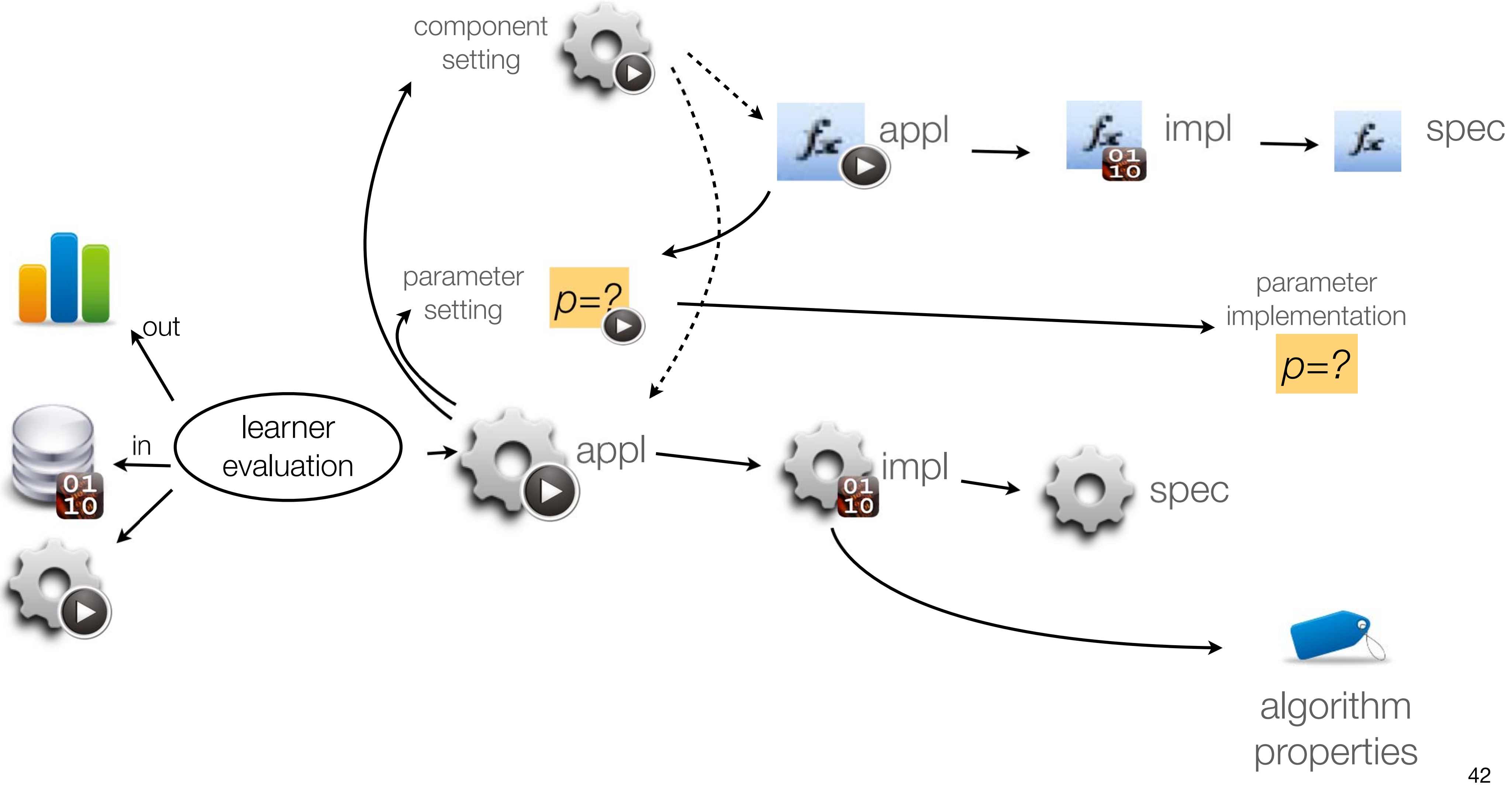
Collect experiments in a database
to query all empirical results

ExpDB: a database to share experiments



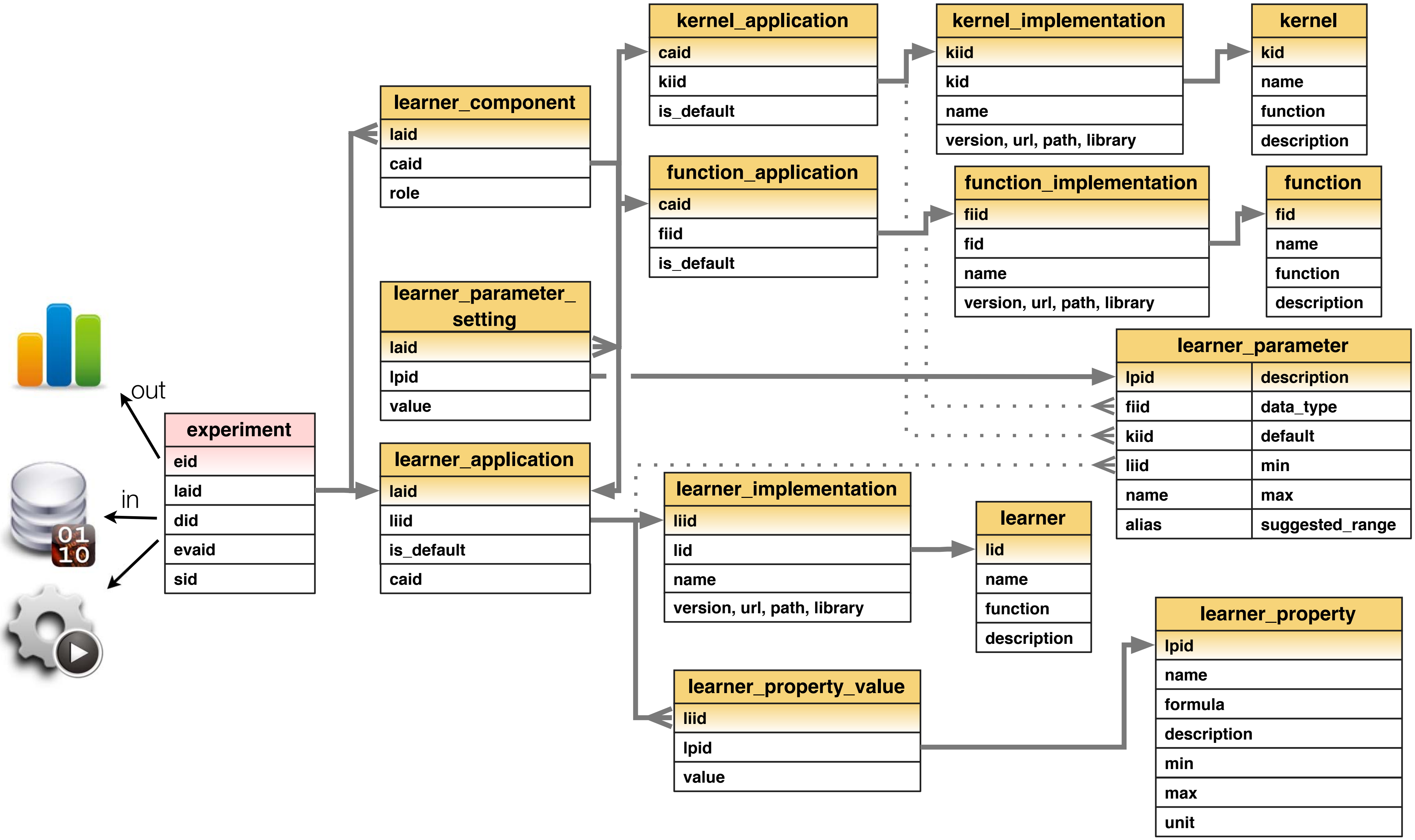
Experiment Database

>650,000 experiments, 54 algorithms, >87 datasets, 45 evaluation measures, 2 data processors, bias-variance analysis



Experiment Database

>650,000 experiments, 54 algorithms,
>87 datasets, 45 evaluation measures,
2 data processors, bias-variance analysis



experiment
eid
laid
did
evaid
sid

learner_component
laid
caid
role

learner_parameter_setting
laid
lpid
value

learner_application
laid
liid
is_default
caid

kernel_application
caid
kiid
is_default

function_application
caid
fiid
is_default

kernel_implementation
kiid
kid
name
version, url, path, library

function_implementation
fiid
fid
name
version, url, path, library

kernel
kid
name
function
description

function
fid
name
function
description

learner_parameter	
lpid	description
fiid	data_type
kiid	default
liid	min
name	max
alias	suggested_range

learner_implementation
liid
lid
name
version, url, path, library

learner
lid
name
function
description

learner_property_value
liid
lpid
value

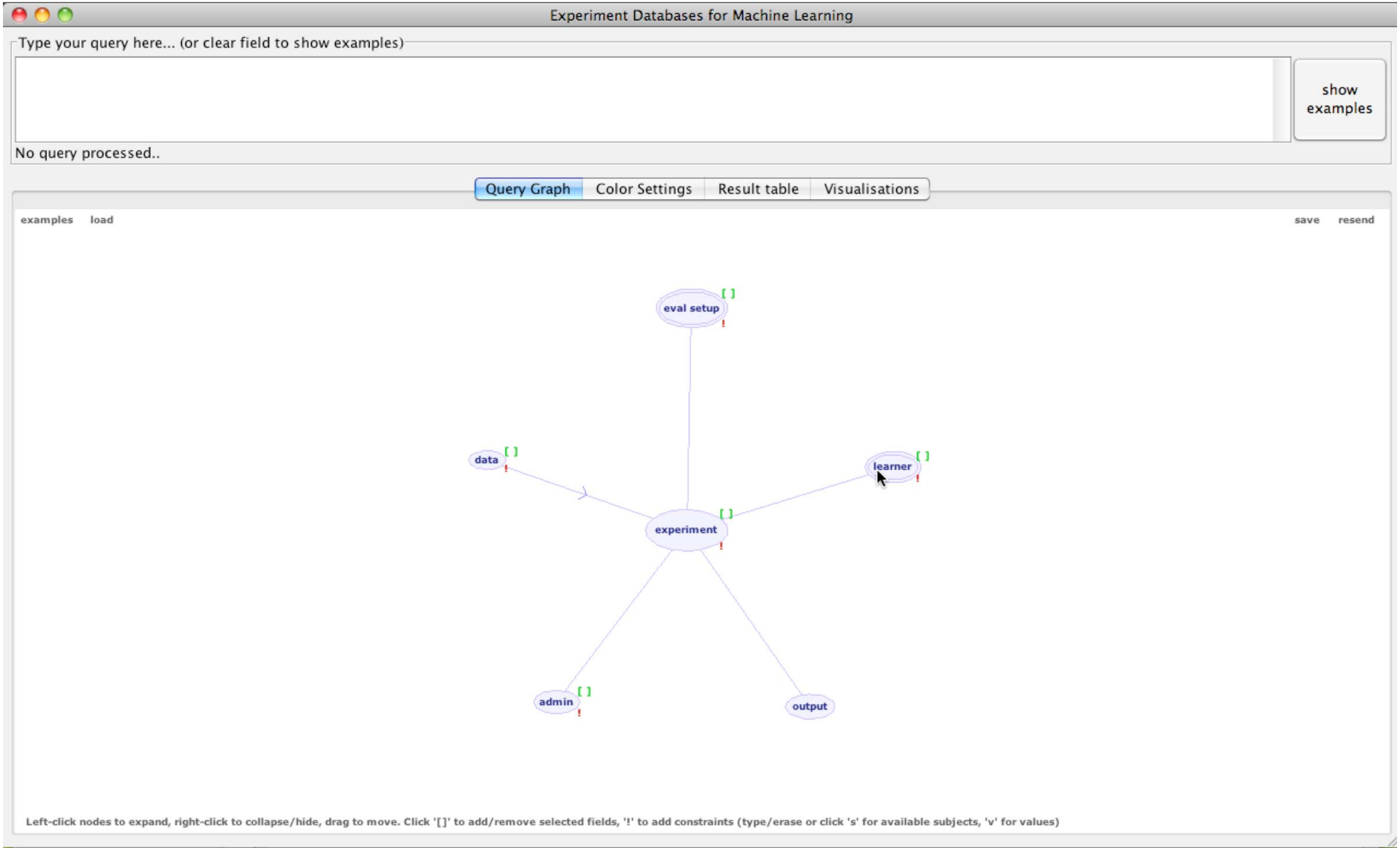
learner_property
lpid
name
formula
description
min
max
unit

Use Case 3

Intuitive querying

Query Interface (YouTube “experiment database”)

http://expdb.cs.kuleuven.be



Experiment Databases for Machine Learning

Type your query here... (or clear field to show examples)

No query processed..

Left-click nodes to expand, right-click to collapse/hide, drag to move. Click '[]' to add/remove selected fields, '!' to add constraints (type/erase or click 's' for available subjects, 'v' for values)

The way ahead

- 3rd generation of tools could make data mining into e-science
    - Experiments shared, reused, run worldwide
    - Repeatable, generalizable, reusable
- Cooperation on a standardized ontology for data mining?
- Automatic ontology extraction: DM paper -> ontology extension
- RDF experiment databases?
- Open problems:
    - Queriable models, auto-population (active meta-learning), quality control

Hvala

Danke

Thanks

Xie Xie

Diolch

Toda

Merci

Grazie

Spasiba

Efharisto

Obrigado

Arigato

Köszönöm

Tesekkurler

Dank U

Dhanyavaad

Gracias

http://expdb.cs.kuleuven.be