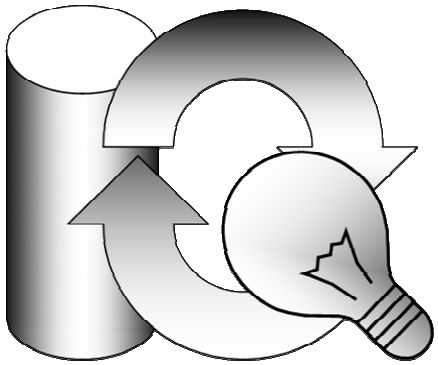
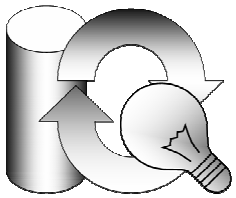
Inductive Databases and Queries for Computational Scientific Discovery

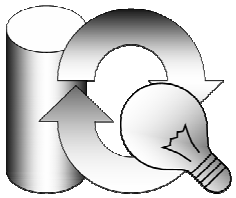


Sašo Džeroski
Jozef Stefan Institute,
Department of Knowledge Technologies
Ljubljana, Slovenia



Outline

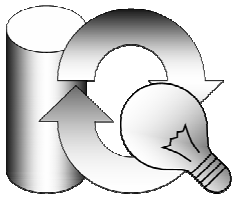
- What is Computational Scientific Discovery
  - Introduction
  - Examples (ecological models, reaction pathways)
- What are Inductive Databases and Queries
  - Introduction
  - Examples (QSAR, integrative genomics)
- How the two can be connected, i.e., how Inductive Databases and Queries can be used for Computational Scientific Discovery



Computational Scientific Discovery

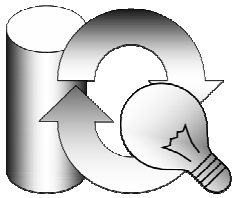
- What is Scientific Discovery:
  The process by which a scientist creates or finds some hitherto unknown knowledge such as class of objects, an empirical law, or an explanatory theory

- Computational Scientific Discovery attempts to provide computational support for this process
  - Early research reconstructed episodes from the history of science
  - Recent efforts in this area have focussed on individual scientific activities (such as formulating quantitative laws) and have led to several new discoveries



Elements of Scientific Behavior

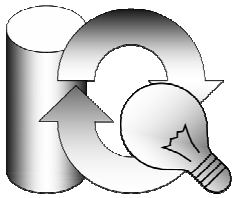
- Scientific knowledge structures
  - Observations
  - *Taxonomies:*
    - Define or describe concepts for a domain, along with specialization relations among them
    - Specify the concepts and terms used to state laws and theories
  - *Laws:* Summarize relations among observed variables, objects or events
  - *Theories:*
    - Statements about the structures or processes that arise in the environment
    - Stated using terms from the domain's taxonomy
    - Interconnect laws into a unified theoretical account
  - Models, Predictions, Explanations (Derived from above)



Elements of Scientific Behavior

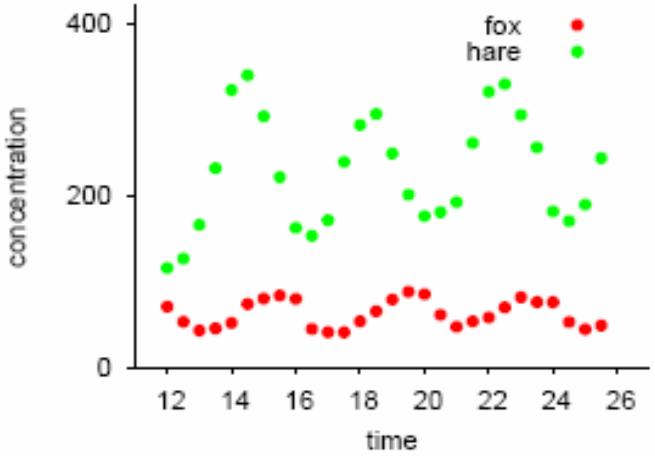
- Scientific processes/activities are concerned with generating and manipulating scientific data and knowledge structures

- Scientific activities
  - Collecting data/observations
  - *Formation and revision of:*
    - *Taxonomies:* Organize observations into classes and subclasses; define those classes and subclasses
    - *Laws:* Given observed data, find empirical laws
    - *Theories:* Given one or more laws, generate a theory
  - Deriving models, predictions, and explanations



Laws of Dynamic Systems' Behavior

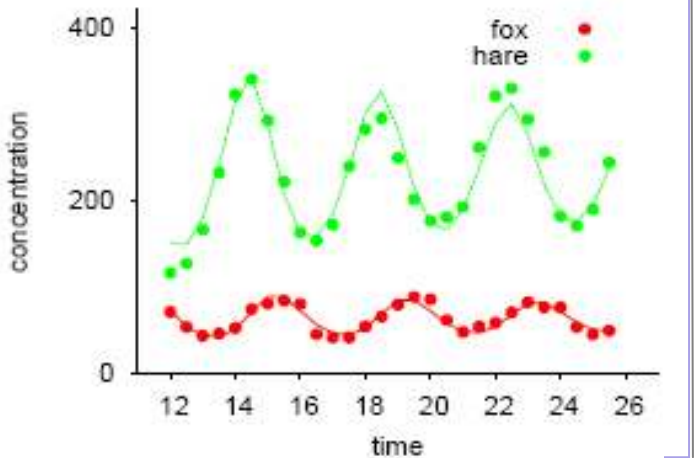
- Input: Observed behavior of dynamics systems

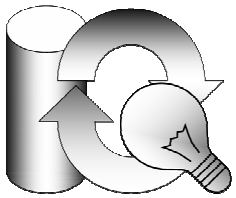


- Output: Set of differential equations

$$\frac{d}{dt} hare = 2.5 \cdot hare - 0.3 \cdot hare \cdot fox$$

$$\frac{d}{dt} fox = 0.1 \cdot 0.3 \cdot hare \cdot fox - 1.2 \cdot fox$$





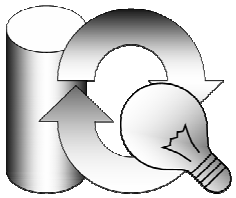
Explanatory Models

- Looking deeper into the model

$$\frac{d}{dt} hare = 2.5 \cdot hare - 0.3 \cdot hare \cdot fox$$
$$\frac{d}{dt} fox = 0.1 \cdot 0.3 \cdot hare \cdot fox - 1.2 \cdot fox$$

- Three processes
  - Exponential growth of hare population
  - Exponential loss of fox population
  - Predator–prey interaction between the two species

- Terms in equations correspond to processes



Domain Knowledge: Generic Processes

- Generic process for predator-prey interaction

  process predator_prey_interaction

  variables $Prey\{species\}, Pred\{species\}$

  parameters $r[0, \text{inf}], e[0, \text{inf}]$

  equations

  $\frac{d}{dt} Prey = -1 \cdot r \cdot Prey \cdot Pred$

  $\frac{d}{dt} Pred = e \cdot r \cdot Prey \cdot Pred$

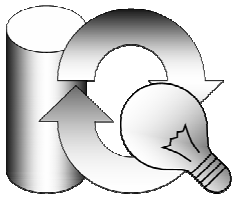
- Instantiation to specific processes

  process predator_prey_interaction

  $\frac{d}{dt} hare = -0.3 \cdot hare \cdot fox$

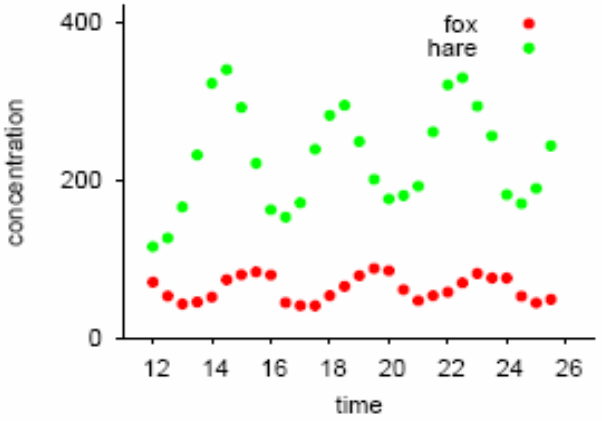
  $\frac{d}{dt} fox = 0.1 \cdot 0.3 \cdot hare \cdot fox$

- In this case: Pred=fox, Prey=hare, r=0.3, e=0.1



Process-based Models of Dyn Sys

- Input: Observed behavior + Set of generic processes



process predator_prey_interaction

variables $Prey\{species\}, Pred\{species\}$

parameters $r[0, \text{inf}], e[0, \text{inf}]$

equations

$$\frac{d}{dt} Prey = -1 \cdot r \cdot Prey \cdot Pred$$
$$\frac{d}{dt} Pred = e \cdot r \cdot Prey \cdot Pred$$

- Output: Set of instantiated processes + ODEs

process exponential_growth

$$\frac{d}{dt} hare = 2.5 \cdot hare$$

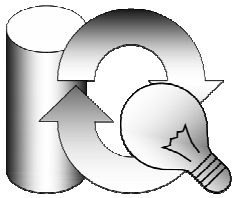
process exponential_loss

$$\frac{d}{dt} fox = -1.2 \cdot fox$$

process predator_prey_interaction

$$\frac{d}{dt} hare = -0.3 \cdot hare \cdot fox$$
$$\frac{d}{dt} fox = 0.1 \cdot 0.3 \cdot hare \cdot fox$$

$$\frac{d}{dt} hare = 2.5 \cdot hare - 0.3 \cdot hare \cdot fox$$
$$\frac{d}{dt} fox = 0.1 \cdot 0.3 \cdot hare \cdot fox - 1.2 \cdot fox$$



Integrating Data and Knowledge

- Using different types of domain knowledge
  - Background knowledge on basic processes
  - Using existing models and revising them

$$
\begin{aligned}
NPPc &= \max(0, E \cdot IPAR) \\
E &= 0.389 \cdot T1 \cdot T2 \cdot W \\
T1 &= 0.8 + 0.02 \cdot topt - 0.0005 \cdot topt^2 \\
T2 &= 1.1814/((1+\exp(0.2 \cdot (TDIFF - 10))) \cdot (1+\exp(0.3 \cdot (-TDIF \\
TDIFF &= topt - tempc \\
W &= 0.5 + 0.5 \cdot eet/PET \\
PET &= 1.6 \cdot (10 \cdot \max(tempc, 0)/ahi)^A \cdot pet\_tw\_m \\
A &= 0.000000675 \cdot ahi^3 - 0.0000771 \cdot ahi^2 + 0.01792 \cdot ahi + 0.49 \\
IPAR &= FPAR\_FAS \cdot monthly\_solar \cdot SOL\_CONV \cdot 0.5 \\
FPAR\_FAS &= \min((SR\_FAS - 1.08)/srdiff, 0.95) \\
SR\_FAS &= (1 + fas\_ndvi/1000)/(1 - fas\_ndvi/1000) \\
SOL\_CONV &= 0.0864 \cdot days\_per\_month
\end{aligned}
$$

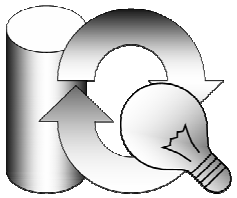
$$
\begin{aligned}
NPPc &= \max(0, E \cdot IPAR) \\
E &= 0.402 \cdot T1^{0.624} \cdot T2^{0.215} \cdot W^0 \\
T1 &= 0.680 + 0.270 \cdot topt - 0 \cdot topt^2 \\
T2 &= 1.1814/((1+\exp(0.2 \cdot (TDIFF - 10))) \cdot (1+\exp(0.3 \cdot (-TDIFF - 10)))) \\
TDIFF &= topt - tempc \\
TDIFF &= topt - tempc \\
W &= 0.5 + 0.5 \cdot eet/PET \\
PET &= 1.6 \cdot (10 \cdot \max(tempc, 0)/ahi)^A \cdot pet\_tw\_m \\
A &= 0.000000675 \cdot ahi^3 - 0.0000771 \cdot ahi^2 + 0.01792 \cdot ahi + 0.49239 \\
IPAR &= FPAR\_FAS \cdot monthly\_solar \cdot SOL\_CONV \cdot 0.5 \\
FPAR\_FAS &= \min((SR\_FAS - 1.08)/srdiff, 0.95) \\
SR\_FAS &= (1 + fas\_ndvi/750)/(1 - fas\_ndvi/750) \\
SOL\_CONV &= 0.0864 \cdot days\_per\_month
\end{aligned}
$$

  - Completing partially specified models

$$\dot{h} = \frac{f(a)}{A}(h_{sea} - h + h_0) + \frac{Q_f}{A} + g(W_{vel}, W_{dir})$$

$$f(a) = 5 + 5 \cdot a + 5 \cdot a^2 + 5 \cdot a^3 - 1.01 \cdot a^4$$

$$
\begin{aligned}
g(W_{\text{vel}}, W_{\text{dir}}) = & -0.00137 - 0.0106 \cdot \cos W_{\text{dir}} \\
& +0.218 \cdot \cos W_{\text{dir}} \cdot \sin W_{\text{dir}} \\
& +0.0106 \cdot W_{\text{vel}} \cdot \cos W_{\text{dir}} \cdot \sin W_{\text{dir}} \\
& -0.0128 \cdot W_{\text{vel}}^2 \cdot \cos W_{\text{dir}} \cdot \sin W_{\text{dir}} \\
& -0.000428 \cdot W_{\text{vel}}^3 \cdot \cos W_{\text{dir}} \cdot \sin W_{\text{dir}}.
\end{aligned}
$$



Example Applications: Ecology

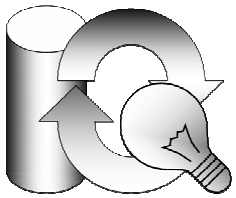
- Modelling aquatic ecosystems
  - Venice lagoon

$$\dot{\texttt{biomass}} = 4.79 \cdot 10^{-5} \cdot \texttt{biomass} \cdot (1 - \frac{\texttt{biomass}}{0.844}) + 0.406 \cdot \texttt{biomass} \cdot (1 - e^{-0.216 \cdot \texttt{temp}}) \cdot (1 - e^{-0.413 \cdot \texttt{DO}}) \cdot \frac{\texttt{NH3}}{\texttt{NH3} + 1} - 0.0343 \cdot \texttt{biomass}$$

  - Lake Glumsoe, Denmark

$$\dot{\texttt{phyto}} = 0.553 \cdot \texttt{temp} \cdot \frac{\texttt{phosp}}{0.0264 + \texttt{phosp}} - 4.35 \cdot \texttt{phyto} - 8.67 \cdot \texttt{phyto} \cdot \texttt{zoo}$$

  - Many other: Lake Bled (Slovenia), Lake Kasumigaura (Japan), Lake Greifensee (Switzerland), Lake Kinnereth (Israel), Lake Ohrid (Macedonia)



Example Apps: Metabolic Networks

model glycolysis_kinetics;

process flux_combination_G3P;

equations $d[G3P,t,1] = 2.0828 * G3P_{+flux} + 0.0002 * G3P_{-flux}$;

process flux_combination_3PG;

equations $d[3PG,t,1] = 1.2251 * 3PG_{+flux} + 4.3892 * 3PG_{-flux}$;

process flux_combination_F16BP;

equations $d[F16BP,t,1] = 3.2353 * F16BP_{+flux} + 1.2893 * F16BP_{-flux}$;

process flux_combination_F6P;

equations $d[F6P,t,1] = 9.8457 * F6P_{+flux} + 7.9592 * F6P_{-flux}$;

process flux_combination_DHAP;

equations $d[DHAP,t,1] = 1.5514 * DHAP_{+flux} + 0.2402 * DHAP_{-flux}$;

process flux_combination_G6P;

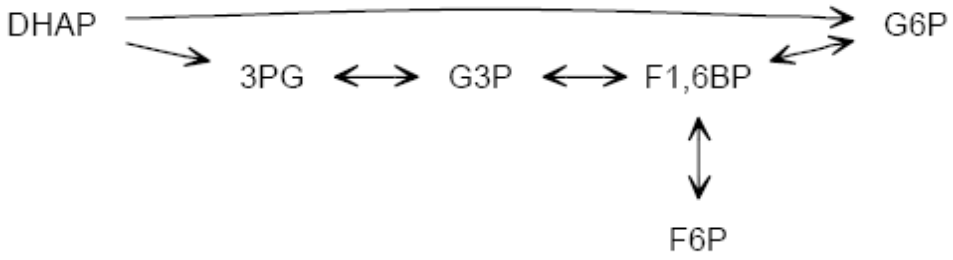
equations $d[G6P,t,1] = 0.1119 * G6P_{+flux} + 0.1557 * G6P_{-flux}$;

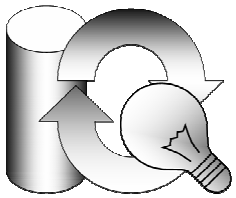
process reversible_G3P_F16BP;

equations $G3P_{+flux} = G3P\hat{}0.0824 * F16BP\hat{}0.1451$;
$G3P_{-flux} = G3P\hat{}0.7173 * F16BP\hat{}1$;
$F16BP_{+flux} = G3P\hat{}0.1678 * F16BP\hat{}0.4607$;
$F16BP_{-flux} = G3P\hat{}0 * F16BP\hat{}0.0010$;

process reversible_3PG_G3P;

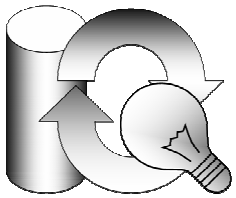
equations $3PG_{+flux} = 3PG\hat{}0.2755 * G3P\hat{}0.2959$;
$3PG_{-flux} = 3PG\hat{}0.3810 * G3P\hat{}0.6193$;
$G3P_{+flux} = 3PG\hat{}0.2166 * G3P\hat{}0.2742$;
$G3P_{-flux} = 3PG\hat{}0.5907 * G3P\hat{}0.3825$;





CSD Focusses

- On standard scientific formalisms (e.g., equations, pathways) introduced and routinely used by scientists
- The results should be communicable with domain scientists and publishable in relevant scientific literature
- Integration of domain knowledge is of primary importance (e.g., concepts from the relevant scientific domain, existing laws/models)
- Interaction with domain scientist and incremental approach also crucial
- Many of these concerns ill met by data mining, some addressed by inductive databases/queries



Inductive Databases and Queries

- A database perspective on knowledge discovery:

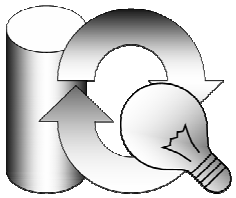
Knowledge discovery processes are query processes

- ”There is no discovery in KDD, it’s all a matter of the expressive power of the query language”

- Inductive database = Database + Patterns/Models

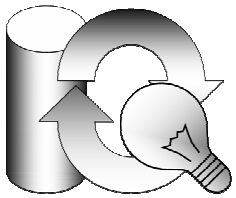
- Sets of patterns can be materialized or views
- Data mining operations = Inductive queries

- IQ: Inductive Queries for Mining Patterns and Models (EU funded project, Future and Emerging Technol.)



Inductive Queries

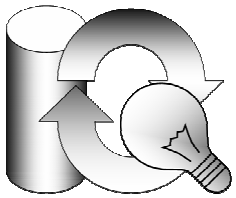
- Inductive query = Set of constraints that a pattern/model has to satisfy
  - Language constraints (only on the pattern/model)
  - Evaluation constraints (concern the validity of the pattern/model with respect to a database)
- Given IDB = D + B + P, we have diff types of queries
  - **Data retrieval (D + B -> D)**: "classical" database query
  - **Cross over (D + B + P -> D)**: uses patterns and data to obtain new data
  - **Processing patterns (P + B -> P)**: patterns queried without access to the data (post-processing)
  - **Data mining (D + B + P -> P)**: new patterns generated on the basis of the data and the existing patterns



Inductive Databases for QSAR

QSAR = Quantitative Structure Activity Relationships

- Basic data structure: Molecule
  - Represented as labeled graph, or
  - relationally through atom/bond facts
- Patterns: Molecular fragments/substructures
- Models: Equations (linear) or other predictive models (e.g., regression trees) based on bulk features and molecular fragments as indicator variables
- Domain knowledge: Functional groups

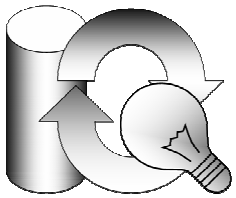


Inductive Databases for QSAR

Inductive queries

- Find frequent patterns (molecular fragments)
- Check for occurrence of fragments in molecules to obtain features
- Build predictive models from bulk features and molecular fragments/functional groups as indicator variables

Underlying application: Drug design



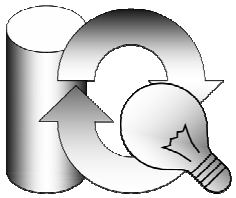
Example Inductive Queries for QSAR

Let us be given datasets $D1$ and $D2$ of molecules

$Q1$: In the context of dataset $D1$, find all molecular fragments that
- appear in the compound AZT (which is a drug for AIDS)
- occur frequently in the active compounds ($\geq$ 15% of them) and
- occur infrequently in the inactive ones ($\leq$ 5% of them)

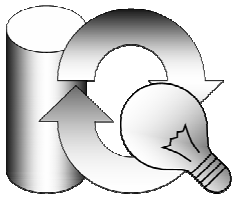
$Q2$: Use the fragments resulting from $Q1$
as features to describe the molecules in $D2$

$Q3$: Use the data resulting from $Q2$
to find a decision tree for predicting activity that
- is of size at most 7 (leaves)
- is as accurate as possible



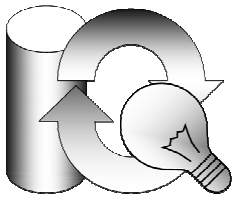
IDBs for Integrative Genomics

- Basic data structure: A microarray
  - In the dataset, rows are patients (with diagnoses),
  - columns are probes/genes,
  - entries are gene expression levels
- Patterns: Rankings of genes (wrt differental expression in the light of diagnosis)
- Models: Relational regression trees/rules predicting the rank of a gene in terms of DK
- Domain knowledge: gene ontology, gene interactions, pathways



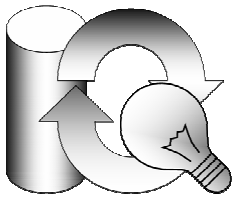
IDBs for Integrative Genomics

- Take microarray data from three neuroblastoma studies (M1, M2, M3), where for each patient we have the status (relapse or 'no event')
- On each of these datasets, rank the genes wrt differential expression in relapse vs. 'no event' producing rankings R1, R2, R3
- From R1, R2, and R3, produce an aggregate ranking R
- Build a model for predicting the rank R of a gene from the domain knowledge, i.e., characterize highly ranked genes in terms of GO/int./pathways



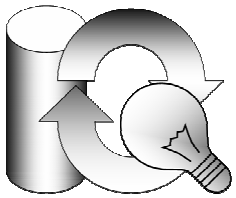
IDBs for Integrative Genomics

- Take microarray data from neuroblastoma patients (N) and Wilm's tumor (W), as well as controls (C)
- Rank the genes wrt differential expression in N vs. C and W vs. C, producing rankings R1, and R2
- Find the pathways with the highest number of highly ranked genes (according to R1 and R2 separately)
- Find the pathways common for R1 and R2
- Underlying application: identify genes/pathways to be targeted with new drugs



IDBs and IQs for CSD

- IDBs and IQs address some of the central concerns of Computational Scientific Discovery
  - The explicit storage of patterns/models and background knowledge allows for the (re)use of domain knowledge together with data
  - The process of (inductive) querying is interactive and allows for significant user involvement
  - The use of constraint-based data mining approaches allows for additional influence of the user on the discovery process



Outlook

- Scientific task: Construct a model of a new lake ecosystem, for which some measurements are available
- First, find a model from the existing literature that has been constructed for a similar ecosystem [query on patterns/models]
- Apply this model to the dataset at hand [cross-over query]
- If the fit of the model to the data is bad, revise the model by using the data or construct a new model by using data and domain knowledge [IQ]
- For this, both scientific data and models need to be stored in (distributed) scientific IDBs!

Computational Discovery of Scientific Knowledge

Advances in technology have enabled the collection of data from scientific observations, simulations, and experiments at an ever-increasing pace. For the scientist and engineer to benefit from these enhanced data collecting capabilities, it is becoming clear that semi-automated data analysis techniques must be applied to find the useful information in the data. Computational scientific discovery methods can be used to this end: they focus on applying computational methods to automate scientific activities, such as finding laws from observational data. In contrast to mining scientific data, which focuses on building highly predictive models, computational scientific discovery puts a strong emphasis on discovering knowledge represented in formalisms used by scientists and engineers, such as numeric equations and reaction pathways.

This state-of-the-art survey provides an introduction to computational approaches to the discovery of scientific knowledge and gives an overview of recent advances in this area, including techniques and applications in environmental and life sciences. The 15 articles presented are partly inspired by the contributions of the International Symposium on Computational Discovery of Communicable Knowledge, held in Stanford, CA, USA in March 2001. More representative coverage of recent research in computational scientific discovery is achieved by a significant number of additional invited contributions.

In parallel to the printed book, each new volume is published electronically in LNCS Online.

Detailed information on LNCS can be found at
www.springer.com/lncs

Proposals for publication should be sent to
LNCS Editorial, Tiergartenstr. 17, 69121 Heidelberg, Germany
E-mail: lncs@springer.com

ISSN 0302-9743

ISBN 978-3-540-73919-7

springer.com

Lecture Notes in Artificial Intelligence

Lecture Notes in Computer Science

Džeroski
Todorovski (Eds.)

LNAI 4660

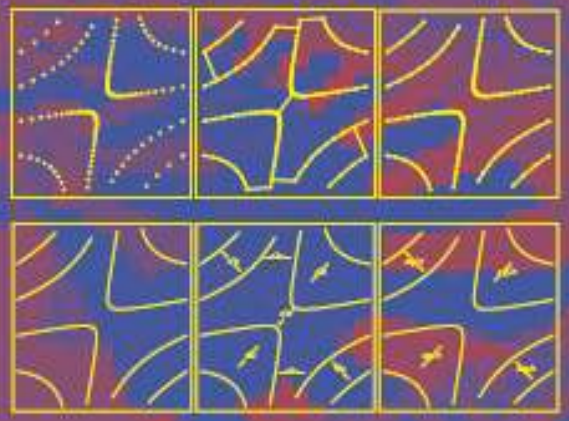
Computational Discovery of Scientific Knowledge

State-of-the-Art Survey

Sašo Džeroski
Ljupčo Todorovski (Eds.)

Computational Discovery of Scientific Knowledge

Introduction, Techniques, and Applications in Environmental and Life Sciences



Springer

Lecture Notes in Computer Science

The LNCS series reports state-of-the-art results in computer science research, development, and education, at a high level and in both printed and electronic form. Enjoying tight cooperation with the R&D community, with numerous individuals, as well as with prestigious organizations and societies, LNCS has grown into the most comprehensive computer science research forum available.

The scope of LNCS, including its subseries LNAI and LNBI, spans the whole range of computer science and information technology including interdisciplinary topics in a variety of application fields. The type of material published traditionally includes

- proceedings (published in time for the respective conference)
- post-proceedings (consisting of thoroughly revised final full papers)
- research monographs (which may be based on outstanding PhD work, research projects, technical reports, etc.)

More recently, several color-cover sublines have been added featuring, beyond a collection of papers, various added-value components; these sublines include

- tutorials (textbook-like monographs or collections of lectures given at advanced courses)
- state-of-the-art surveys (offering complete and mediated coverage of a topic)
- hot topics (introducing emergent topics to the broader community)

In parallel to the printed book, each new volume is published electronically in LNCS Online.

Detailed information on LNCS can be found at www.springer.com/lncs

Proposals for publication should be sent to

LNCS Editorial, Tiergartenstr. 17, 69121 Heidelberg, Germany

E-mail: lncs@springer.com

ISSN 0302-9743

ISBN 978-3-540-75548-7

Lecture Notes in Computer Science

LNCS LNAI LNBI

springer.com

Džeroski • Struyf (Eds.)

LNCS 4747

Knowledge Discovery in Inductive Databases

LNCS 4747

KDID 2006

Sašo Džeroski Jan Struyf (Eds.)

Knowledge Discovery in Inductive Databases

5th International Workshop, KDID 2006
Berlin, Germany, September 2006
Revised Selected and Invited Papers

Springer