Discovering Temporal Bisociations for Linking Concepts over Time

Corrado Loglisci, Michelangelo Ceci

Dipartimento di Informatica
University of Bari ”Aldo Moro”

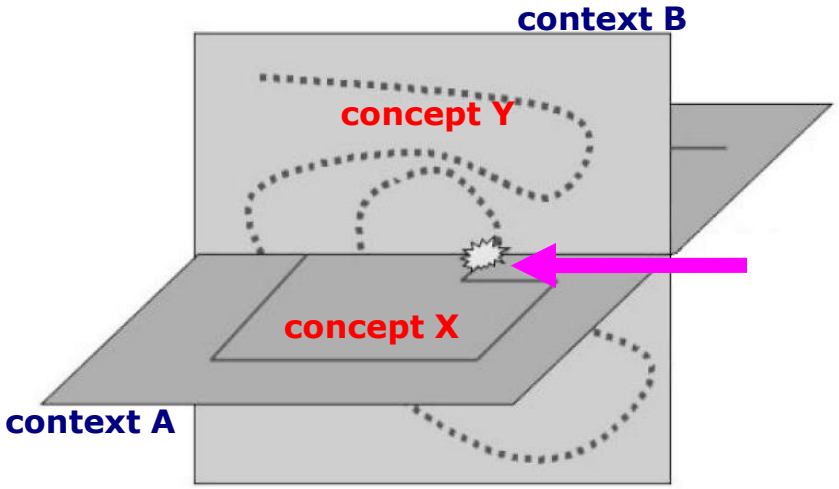
September 5-9th - Athens, Greek

Bisociations

- Bisociations represent interesting relationships between seemingly unconnected concepts from two or more contexts.

- A context can depend on a subjective perspective and can be considered as a domain which collects a set of concepts

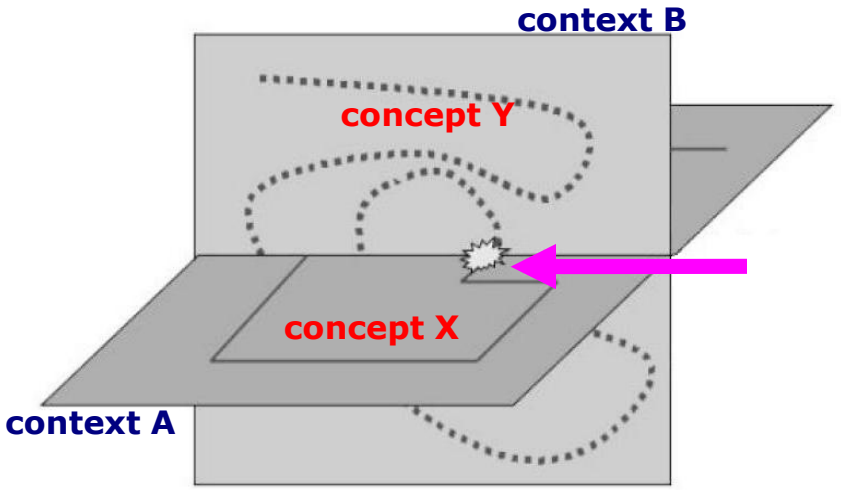
[Berthold et al., PAKDD, 2008
Koestler, Hutchinson Press, 1964]



Bisociations

- Bisociations can be obtained by *comparison*, *abstraction*, *categorization, as well as from analogies and metaphors*

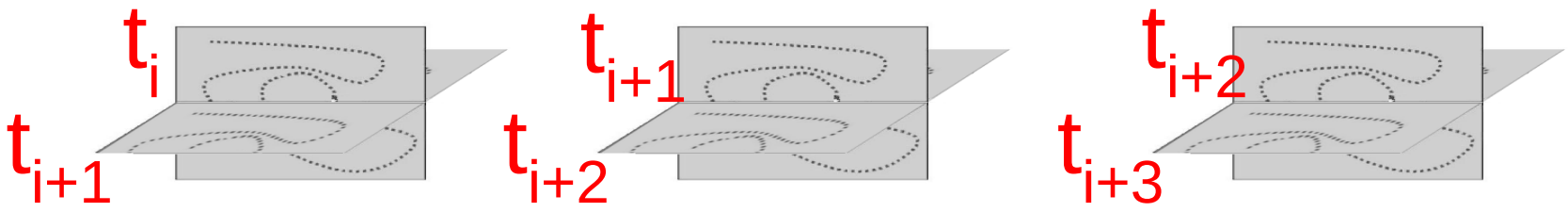
e.g. Scientific discovery needs to create hypotheses worthy of being investigated.
Hypotheses can be generated from bisociations with other contexts.



Computational Challenge

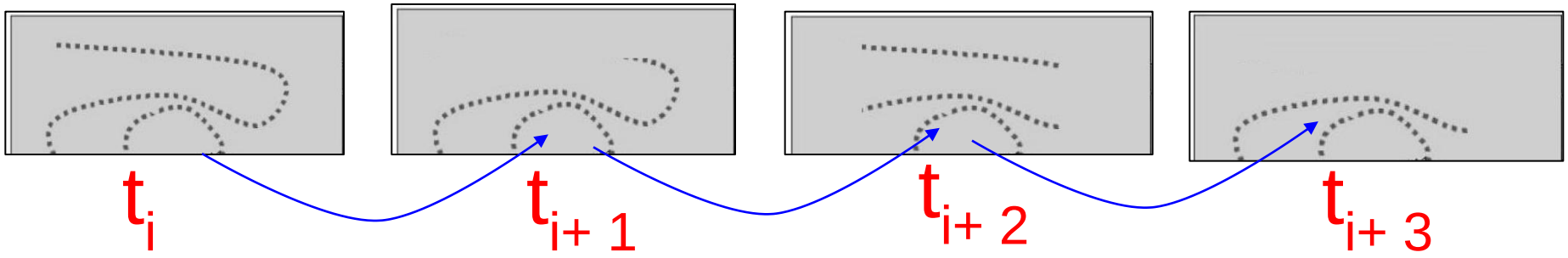
•Contexts are considered static and assimilated to unchangeable domains. However, real-world contexts are in most of cases dynamic in nature → the time-varying component and the dynamic nature have to be taken into account when discovering bisociations

Even only one context can change and can become completely different from what/how it was before



The task: Link Discovery

A dynamic context observed at different time-points can be reasonably seen as a series of distinct static contexts.



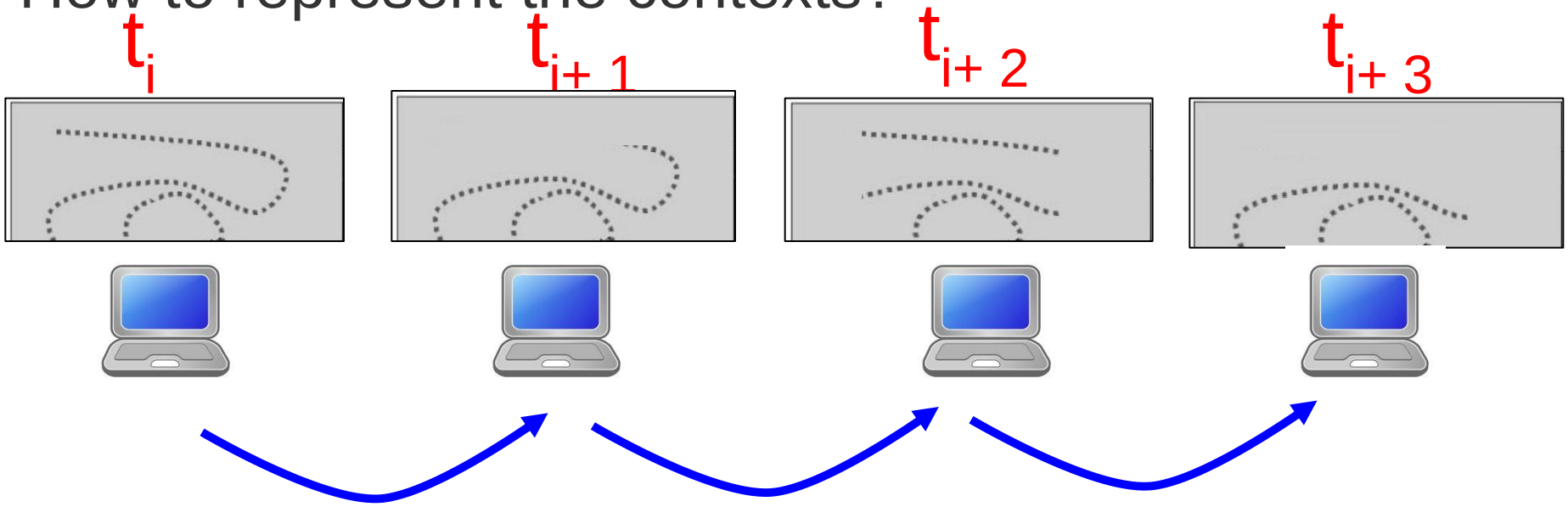
The task:

Linking concepts from distinct static contexts (over time)

→ Temporal Bisociations

The task: Issues

- How to represent the contexts?



- How to perform link discovery?

How to represent the contexts?

Contexts are often represented as networks

(e.g. nodes correspond to named entities annotated with a term frequency vector while the edges represent the co-occurrence of the entities in the documents [Jin et al., 2007])

☺ a single network can aggregate data on the different contexts and the concepts contained within it;

☹ networks generally do not consider the dynamic nature of the data;

☹ computational issues raise when discovering links directly at the level of the data (concepts and contexts)

How to represent the contexts?

Each context is identified by means of a pre-defined temporal discretization and represented as abstract descriptions (e.g., models, patterns).

- dynamic nature is considered through the series of abstract descriptions

- focus on the main characteristics of the contexts
  - lower complexity when discovering links at the level of models or patterns
  - robustness of the discovery process w.r.t. false positive links

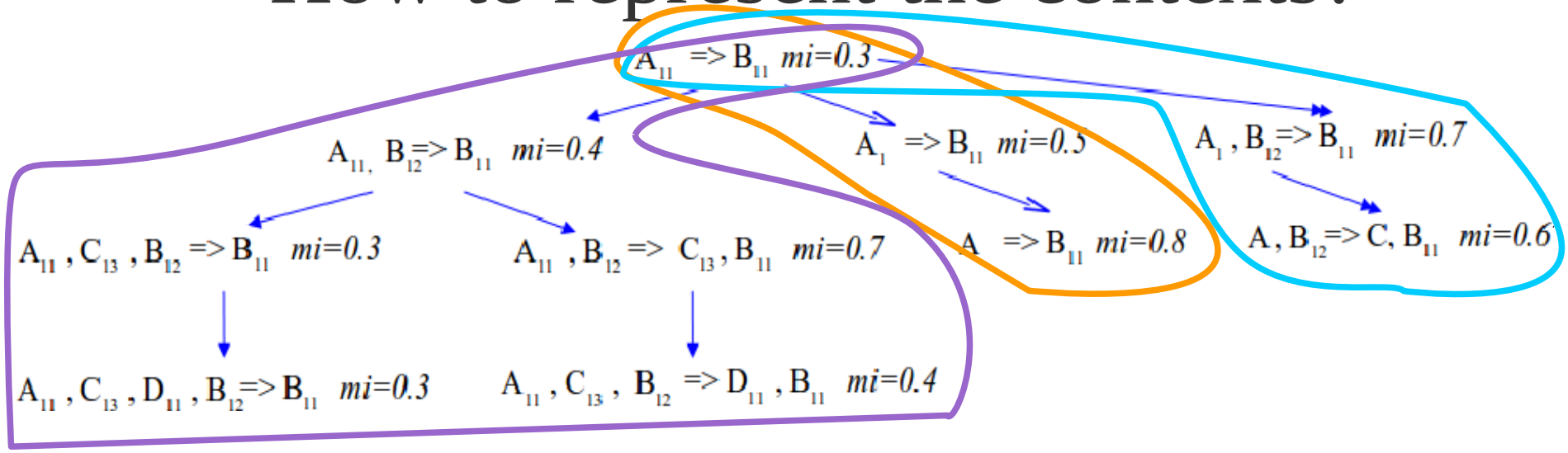
How to represent the contexts?

Abstract descriptions correspond to lattices of association rules(ARs)

- accommodation of is-a background hierarchies to generalize the concepts
- compact lattices by means of *mining non-redundant and minimal ARs from the hierarchies (multiple-level)
- statistical evidence of ARs guarantees robustness

*[Loglisci & Malerba, MLDM 2009]

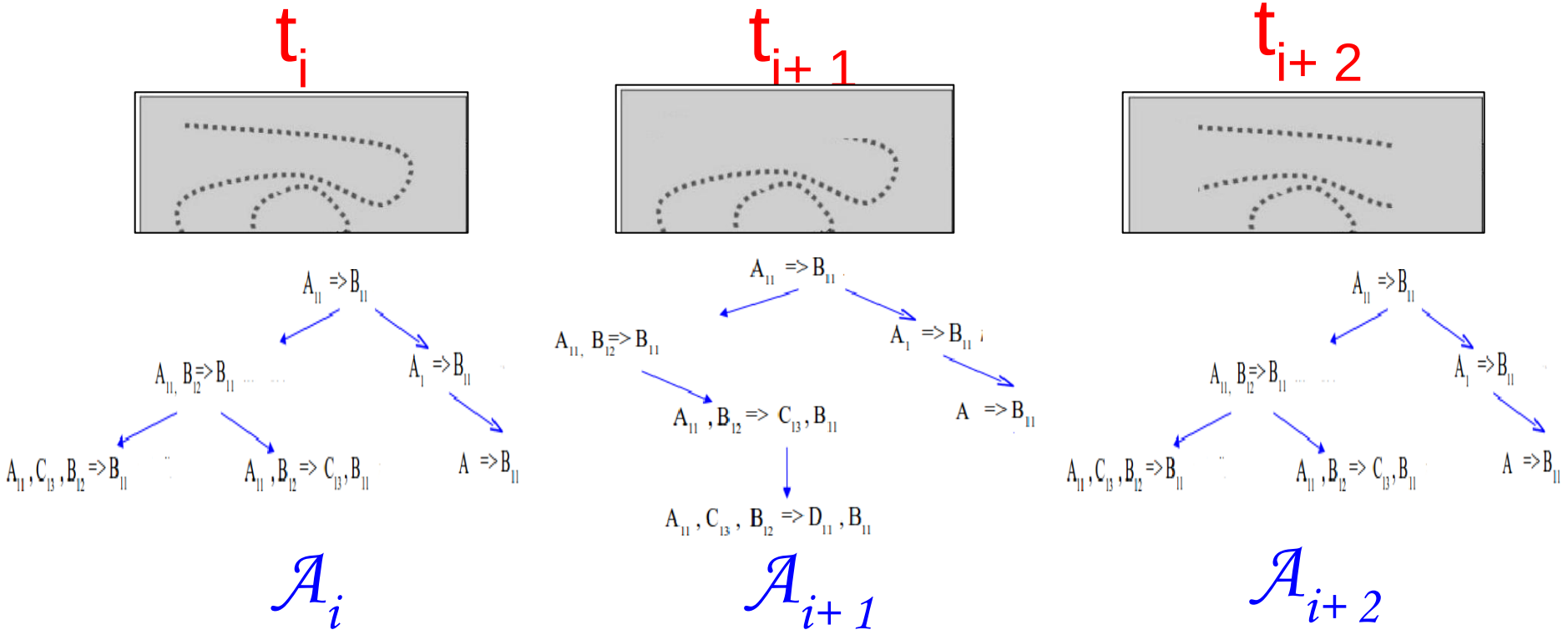
How to represent the contexts?



Three kinds of paths (ARs)

- paths for the extension of the rule at the root with longer ARs (a)
- paths for the generalization of the concepts contained in the root (b)
- paths for the generalization of the concepts contained in the root with longer ARs (c)

How to represent the contexts

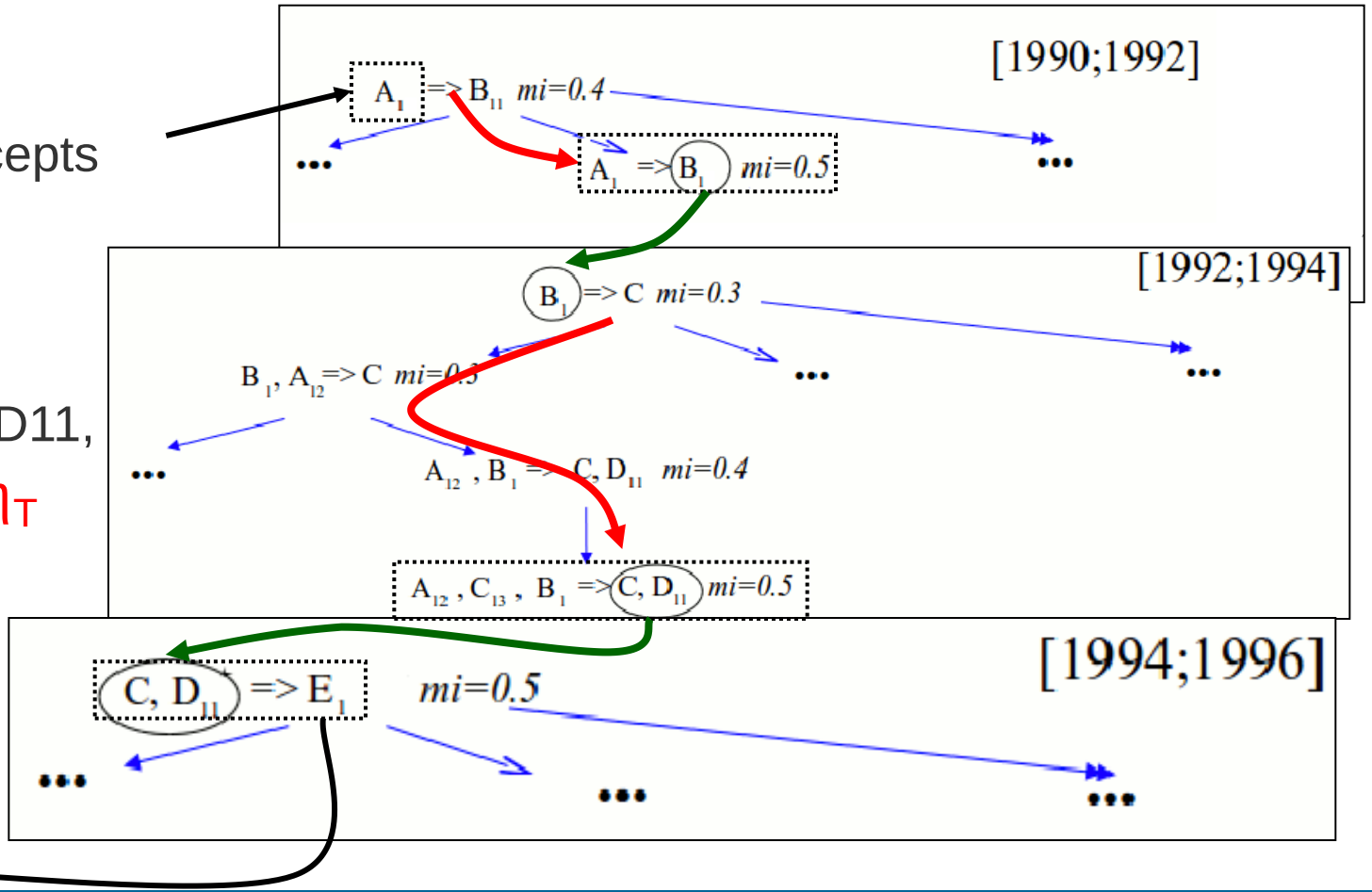


Window size $\geq \Delta_T$ (user-defined threshold - temporal discretization)

How to perform link discovery?

A1, E1 target concepts

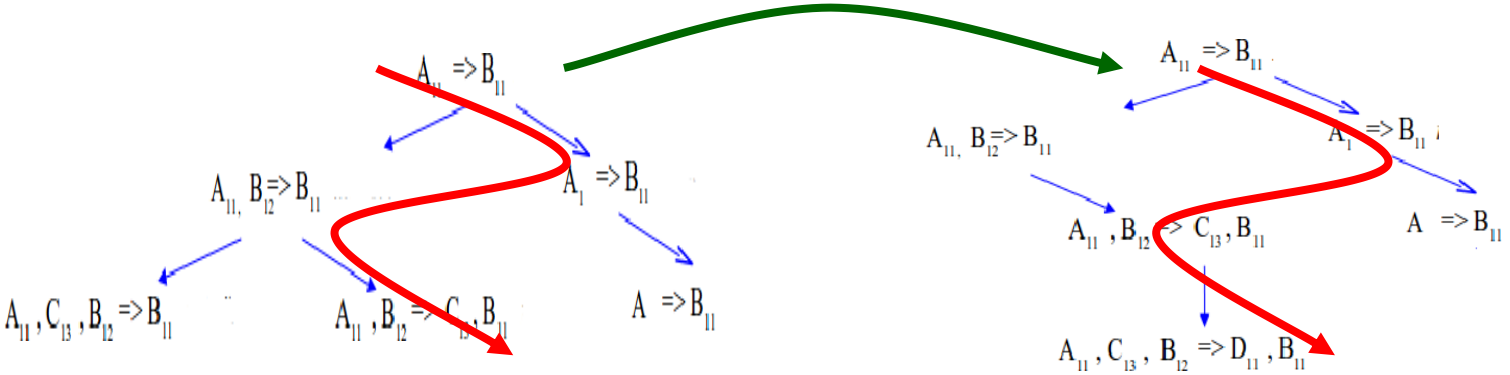
$|\{$**A1** ⇨ B1, A12,C11,B1 ⇨ C,D11, C,D11 ⇨ **E1**$\}| \geq \eta_T$



How to perform link discovery?

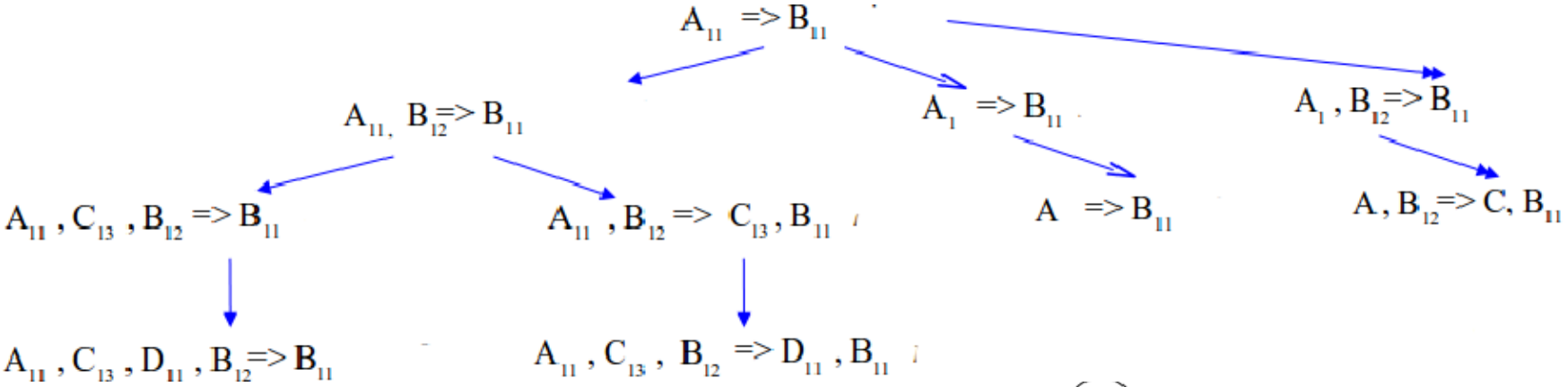
☹ Graph-based analysis techniques are inapplicable to a series of lattices of ARs

- **Exploration** of the lattices and **Chaining** of ARs

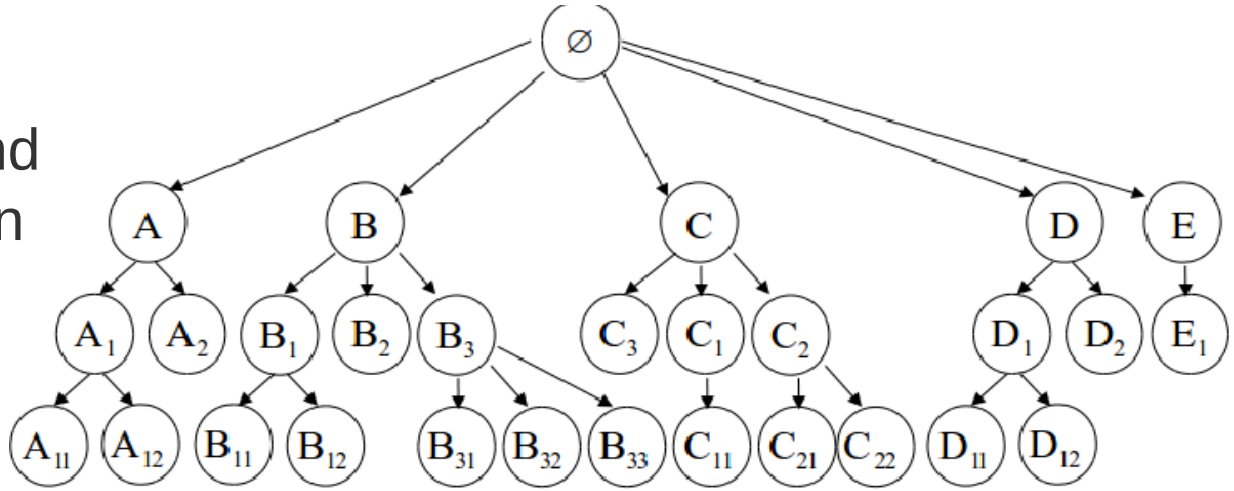


Process guided by the concept of **abstraction** (generalization)
And by an **information theory**–based criterion

How to perform link discovery?



exploitation of background abstraction/generalization relationships



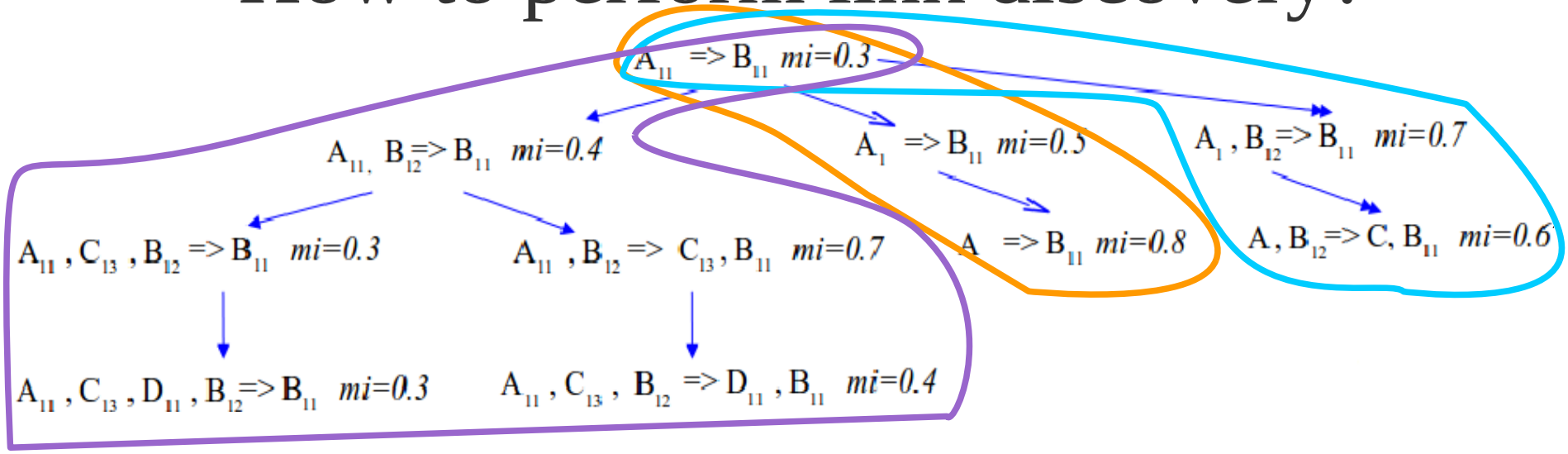
How to perform the link discovery process

- Mutual information *mi* associated to each AR

$$log \frac{supp(Antecedent,\ Consequent)}{supp(Antecedent)*supp(Consequent)}$$

- It expresses the mutual dependence between antecedent and consequent
- Preferred to others for its peculiarity in emphasizing relatively rare concepts that generally occur together and to mitigate the importance of common concepts.

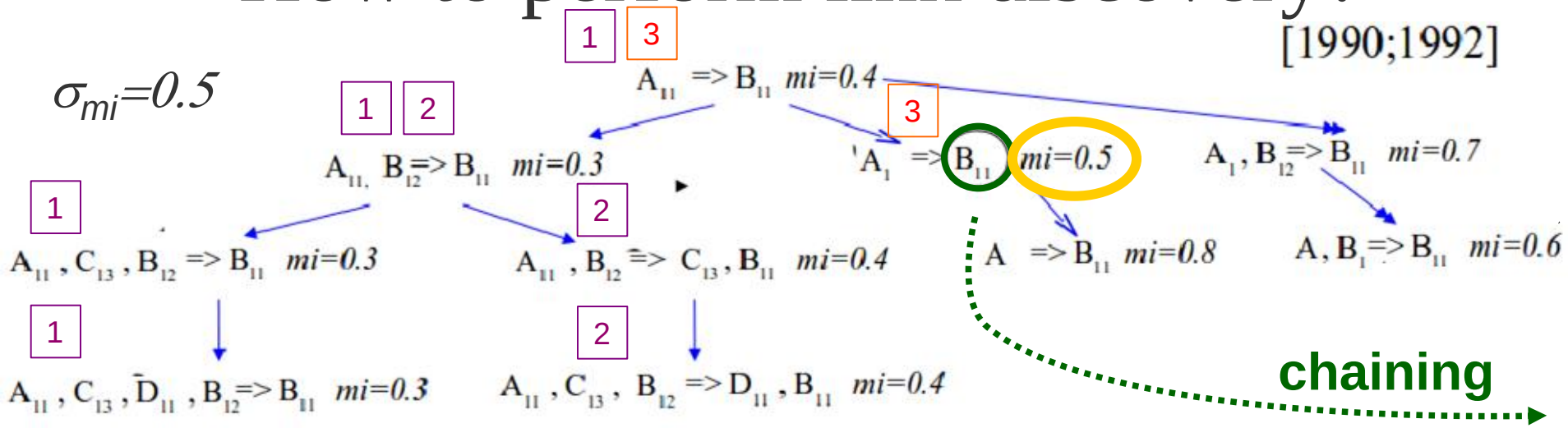
How to perform link discovery?



Exploration follows this order:

- paths for the extension of the rule at the root with longer ARs (a)
- paths for the generalization of the concepts contained in the root (b)
- paths for the generalization of the concepts contained in the root with longer ARs (c)

How to perform link discovery?

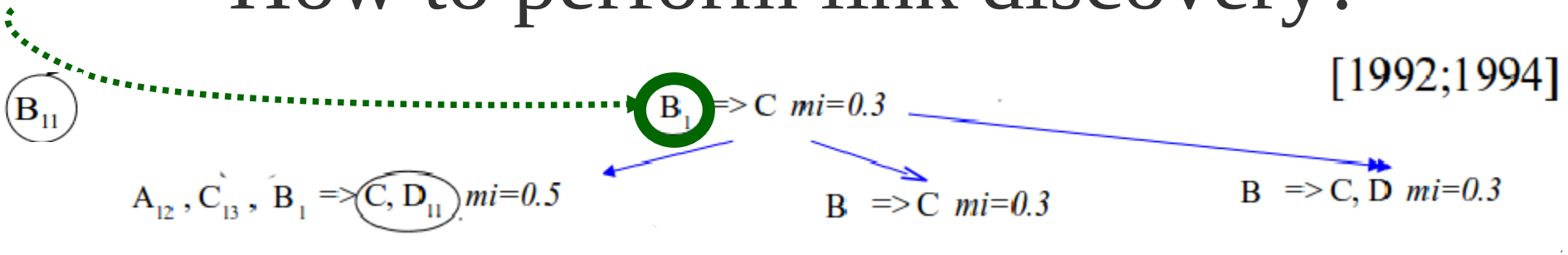


- depth-first search in this order: (a) (b) (c)

- if $mi \geq \sigma_{mi}$ then consider the consequent of the AR for the chain
otherwise, backtrack and continue with paths (a) (b) (c)

-when no AR is identified, a new root is considered

How to perform link discovery?

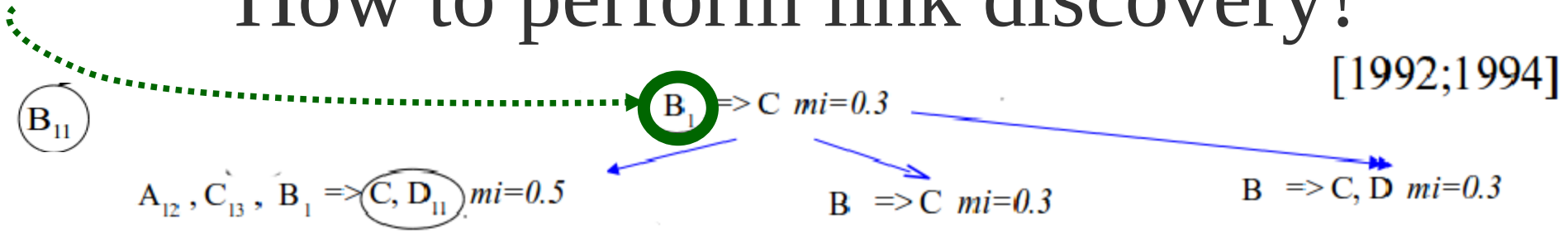


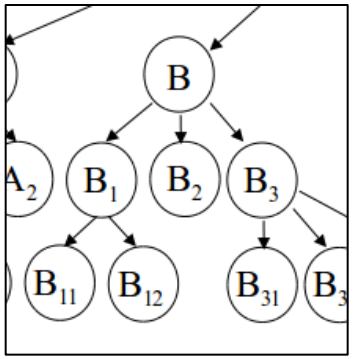
When **chaining** at the next context, we consider ARs *(A1 ⇨B11)*

1. of length two whose antecedent contains only the consequent of the final AR of the previous lattice *(B11 ⇨C1);*

2. of length greater than two whose antecedents contain also the consequent of the final AR of the previous lattice *(B11,D1 ⇨ C1);*

How to perform link discovery?





3. of length two whose antecedents contain only one concept which generalizes the consequent of the final AR of the previous lattice *(B1 ⇨ C1)* ;

4. of length greater than two whose antecedents contain also a concept which generalizes the consequent of the final AR of the previous lattice *(B1,D1 ⇨ C1)* ;

Application

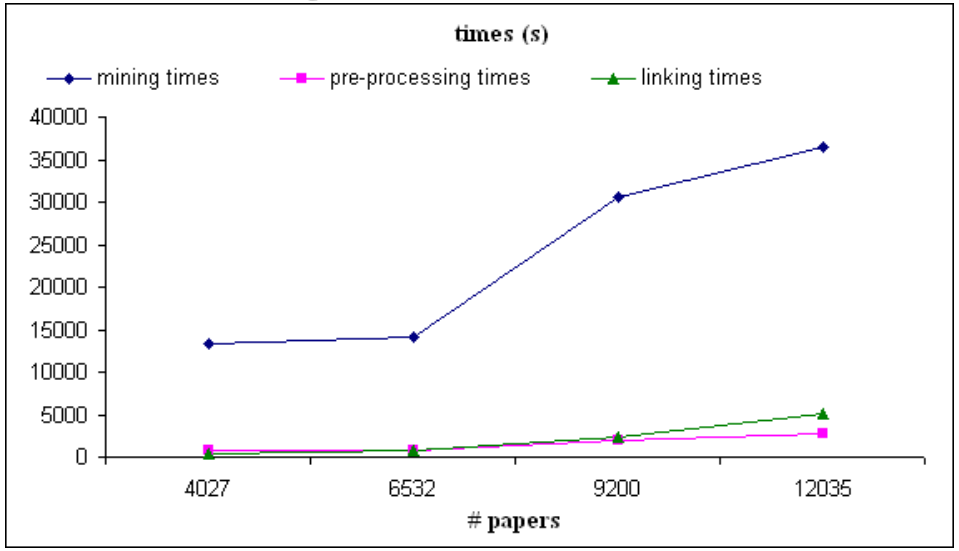
- Scientific discovery based on biomedical literature
- Publications are time-stamped studies on the same topic: scientific literature as a dynamic context.
- Goal: Discovering connections in the biomedical terminology
- Data: Abstracts retrieved from Pubmed search engine

1. *MMD dataset*, queries “migraine” and “magnesium deficiency”
(22223+5311 docs, December 2009)
2. *RDFO dataset*, “raynaud disease” and “fish oils”
(6186+17949 docs, April 2011, *more sparse*)

Application: pre-processing

- Only title, publication date and abstract are considered
- Basic NLP techniques to identify biomedical named entities [Cunningham et al, ACL, 2002]
- Accommodation of Mesh-Terms (http://www.nlm.nih.gov/mesh/) ontologies (abstraction) and thesauri (terms ⇔ concepts)
- Synonyms of the entities replaced with their canonical names [Ferrucci et al, NLE, 2004]
- Time interval-based discretization over years
- Each static context corresponds to the set of abstracts published in the relative time-interval
- Filtering out of Concepts based on TF/IDF (minimum threshold=0.3)

Results: running times

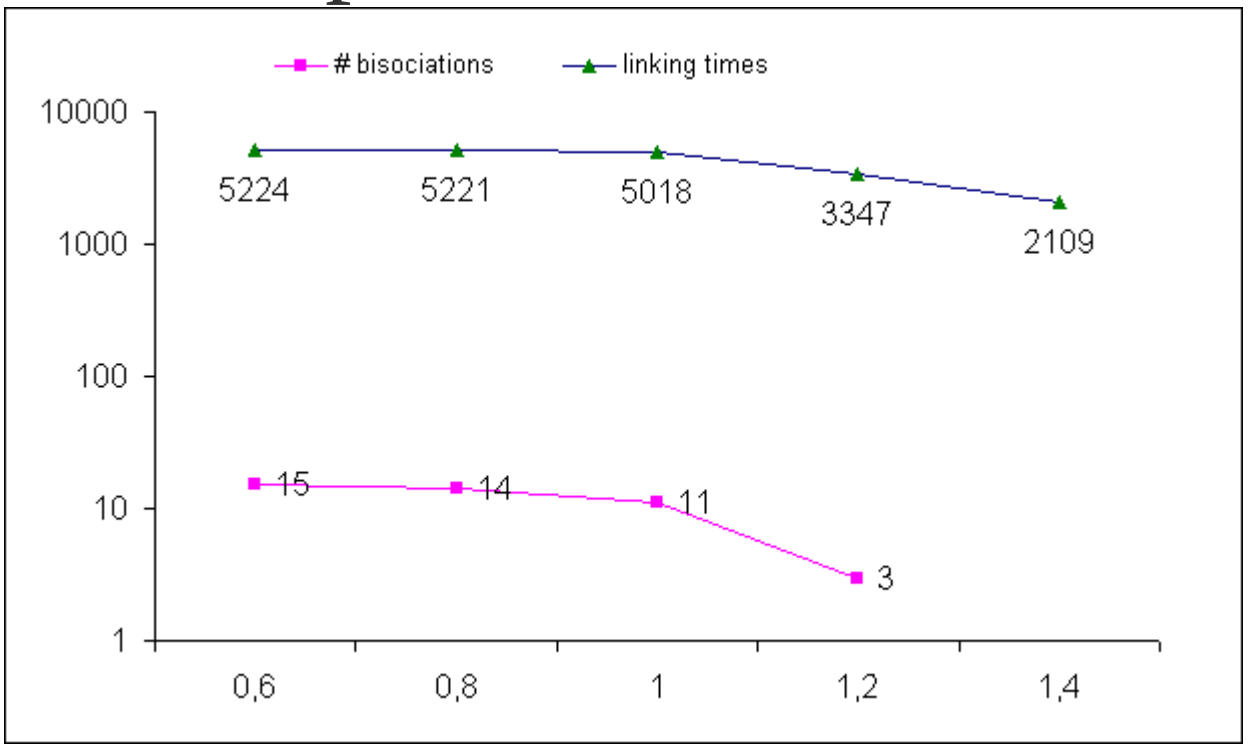


- Δ_T = 1 year
- minSup=0.3
- minConf=0.7
- σ_{mi}=1
- [2000;2009] of MMD

$[\eta_{\mathcal{T}}]$	$[\tau_1;\tau_m]$	avg papers	# ARs	# papers
3	[2000;2003]	1342	15302	4027
5	[2000;2005]	2177	16143	6532
7	[2000;2007]	3066	22880	9200
9	[2000;2009]	4011	27908	12035

- computational cost mainly due to the ARs mining (integration of hierarchies)
- linear growth for the link discovery process (heuristics into the lattices)

Results: parameter σ_{mi}

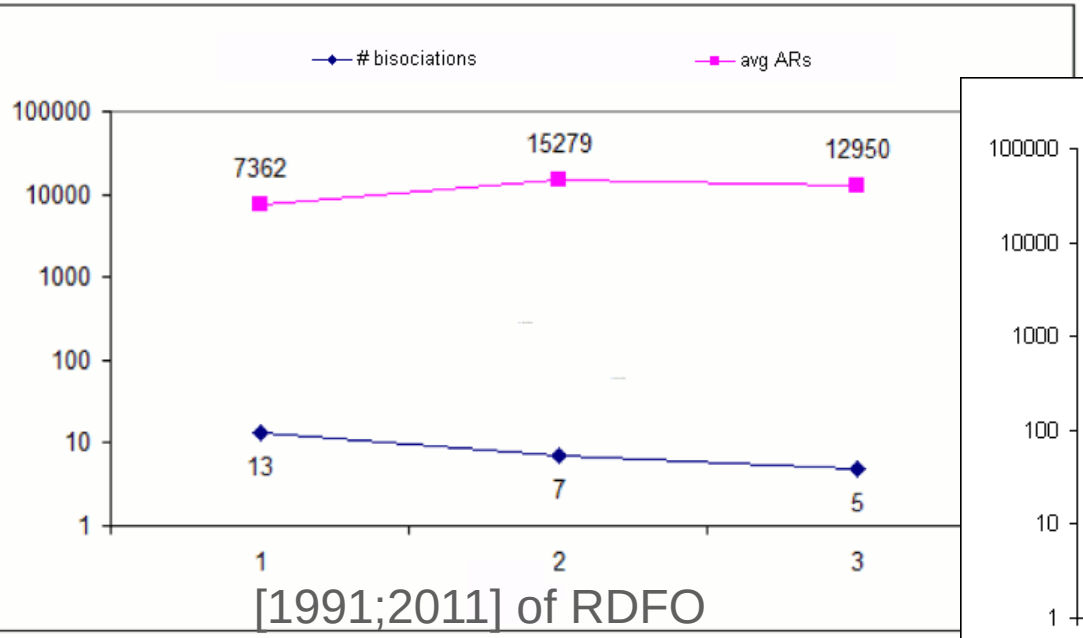


- Δ_T = 3 years
- minSup=0.3
- minConf=0.7
- η_T=9
- [1990;2008] of MMD

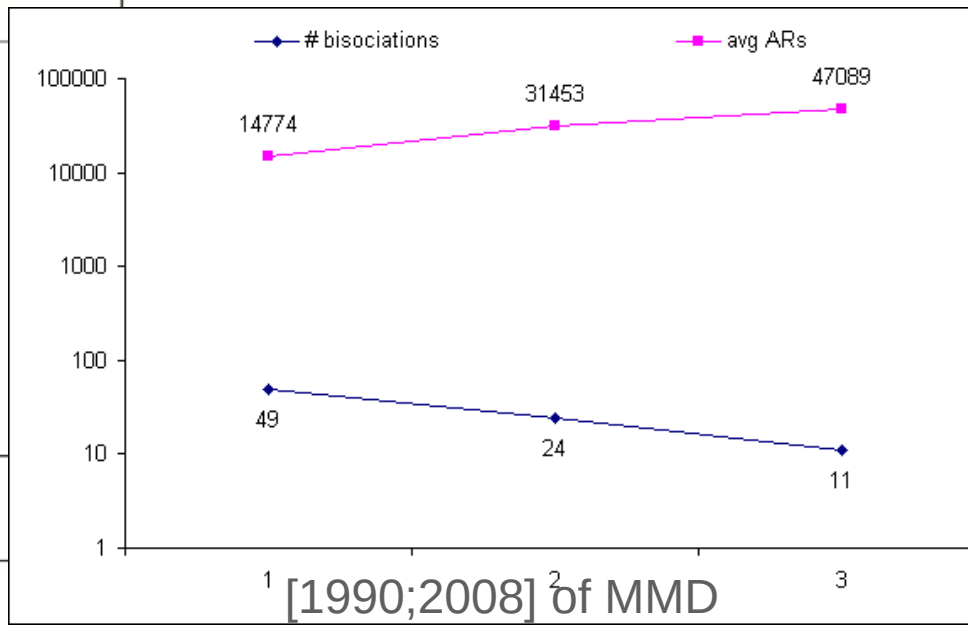
- maintainability of the set of bisociations (links over 27 years, $\Delta_T * \eta_T$)
- dependence from high values of σ_{mi} (strong correlation due to insertion of canonical names)

Results: parameter Δ_T

- $\eta_T = 9$, $\sigma_{mi}=1$



[1991;2011] of RDFO



[1990;2008] of MMD

- maintainability of the set of bisociations (RDFO more sparse)
- the shorter Δ_T the higher #bisociations (the number of static contexts and lattices is higher)

Results: browsing bisociations with X=magnesium deficiency, Y=migraine

- $\Delta_T = 2$
- minSup=0.3
- minConf=0.7
- σ_{mi} =0.8
- η_T=2
- [1980;2009]
- length= 7

[1983;1985] **Magnesium** ⇒ Pain AND Nervous System Diseases [support=0.304, confidence=0.97, mi=1.559]

[1985;1987] Elements AND Anatomy AND Diseases ⇒ Metals [support=0.306, confidence=1.0, mi=1.58]

[1987;1989] Metals AND Hemic and Immune Systems ⇒ Elements [support=0.306, confidence=1.0, mi=1.58]

[1989;1991] Chemicals and Drugs AND Blood ⇒ Plasma [support=0.306, confidence=1.0, mi=1.577348321089182

[1991;1993] Immunologic and Biological Factors AND Elements AND Anatomy ⇒ Metals [support=0.317, confidence=1.0, mi=1.57]

[1993;1995] Chemicals and Drugs AND Biological Sciences AND Neurologic Manifestations ⇒ Neurologic Manifestations [support=0.302, confidence=1.0, mi=1.59]

[1995;1997] Cerebrovascular Disorders AND Pathological Conditions, Signs and Symptoms ⇒ **Migraine** [support=0.304 confidence=1.0, mi=1.604]

Results: browsing bisociations with X=migraine, Y= magnesium deficiency

- $\Delta_T = 2$
- minSup=0.3
- minConf=0.7
- σ_{mi} =0.8
- η_T=2
- [1980;2009]
- length= 6

[1983;1985] **Migraine** ⇒ Pathological Conditions, Signs and Symptoms [support=0.301, confidence=1.0, mi=1.67]

[1985;1987] Metals, Light AND Diseases ⇒ Metals, Alkaline Earth AND Metals, Light [support=0.32, confidence=1.0, mi=0.94]

[1987;1989] Metals, Alkaline Earth AND Metals, Light AND Metals, Light ⇒ Biological Sciences [support=0.31, confidence=0.41, mi=1.35]

[1989;1991] Biological Sciences AND Metals, Alkaline Earth ⇒ Metals [support=0.30, confidence=1.0, mi=0.94]

[1991;1993] Metals, Alkaline Earth AND Metals AND Diseases ⇒ Magnesium [support=0.31, confidence=0.96, mi=1.17]

[1993;1995] Biological Sciences AND Magnesium ⇒ **Magnesium** [support=0.30, confidence=1.0, mi=1.21]

Conclusions & Future Work

The approach extends the task of discovering bisociations to the temporal dimension and considers the dynamic nature of the domain:

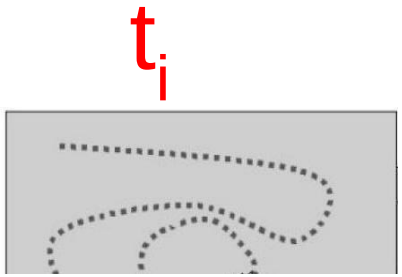
- Process which explores the contexts over time (sequentially)
- Representation of the data in the form of models or patterns
- Capture links which have not been discovered when observing the domain as static, but which may have developed over time, when considering the dynamic nature

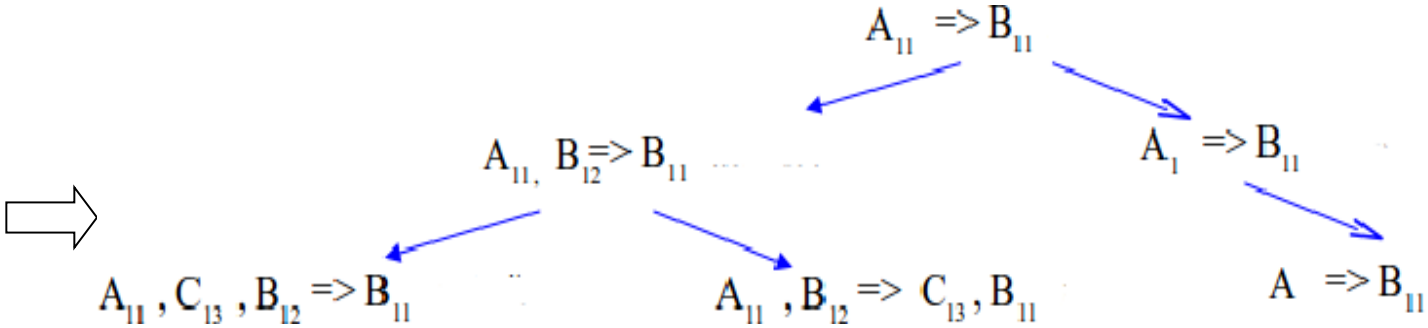
✓ Determination of contexts with Online Discretization techniques

✓ Discovering Bisociations with gaps

Questions?

How to represent the contexts





*Mining multiple-level non-redundant and minimal ARs:

- generation of multiple-level closed *concept*-sets

  Y is closed iff no supersets of Y is supported by the same set of data of Y

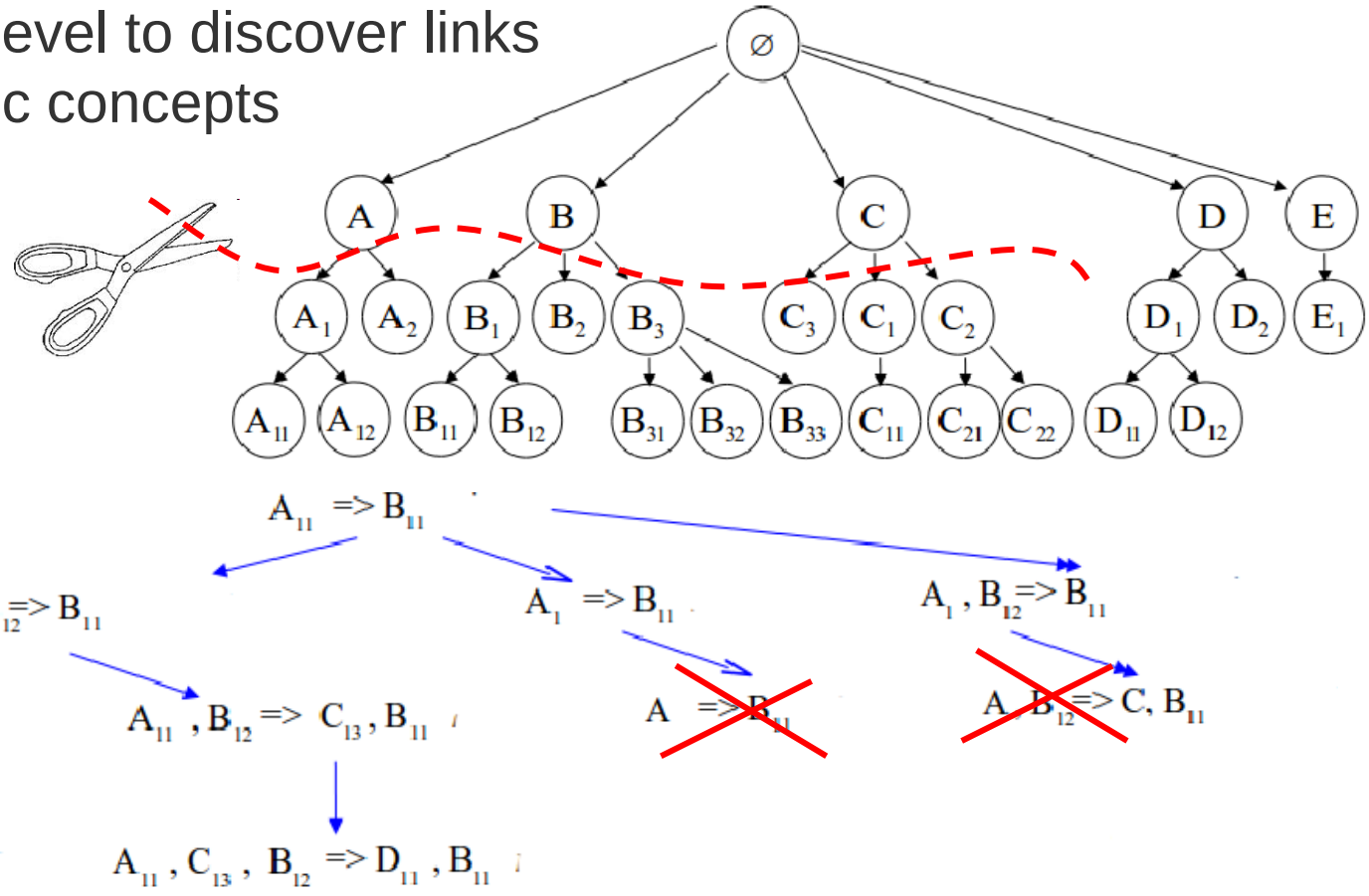
- discovery of multiple-level minimal ARs

R1:X1⇨Y1 is minimal iff not ∃R2 : X2 ⇨ Y2 with identical support and confidence of R1, for which X2 ⊆ X1, Y1 ⊆ Y2 .

*[Loglisci & Malerba, MLDM 2009]

How to represent the contexts

Pruning highest level to discover links with more specific concepts



How to perform the link discovery process

- A temporal bisociation B is a sequence of abstract descriptions

$A_1, \ldots, A_i, \ldots, A_{m-1}, A_1 \in \mathcal{A}_1, \ldots, A_i \in \mathcal{A}_i, \ldots, A_{m-1} \in \mathcal{A}_{m-1}$,

- Given X, Y target concepts, Find temporal bisociations $\mathcal{B}$

  - $\forall B \in \mathcal{B}: X \in A_1, Y \in A_{m-1}$
  - $|\{A_1, A_2, \ldots A_{m-1}\}| \geq \eta_T$ user-defined threshold

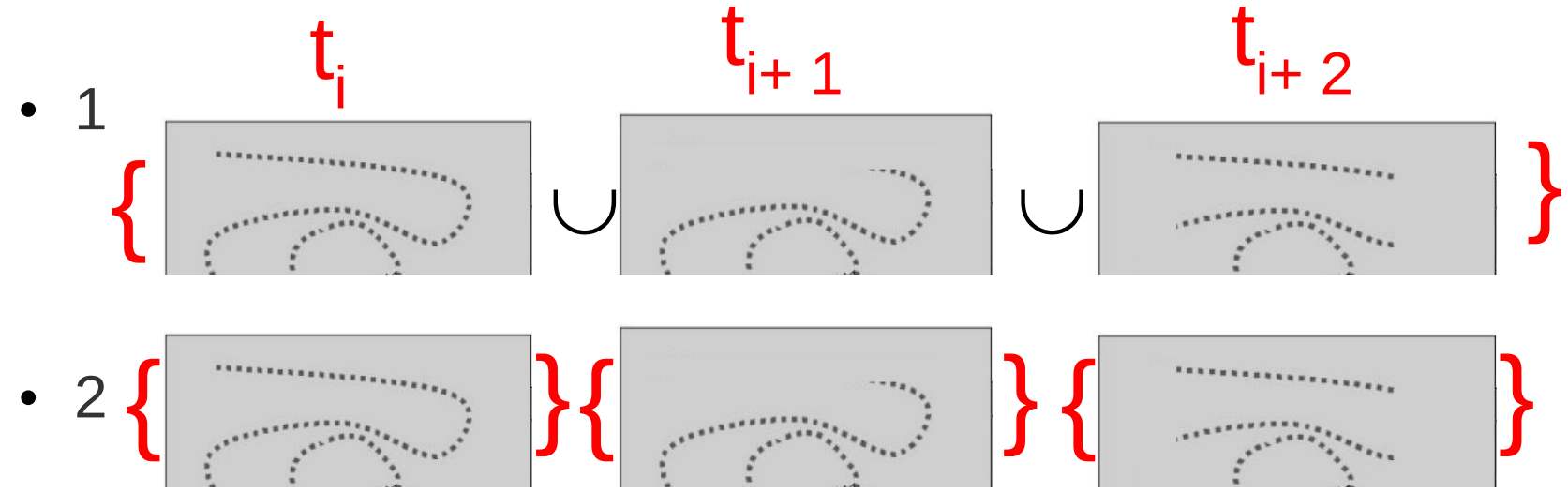
For instance, X ⇨ W, W ⇨ J, J ⇨ Z, Z ⇨ Y

A_1 A_2 A_3 A_4

How to perform the link discovery process

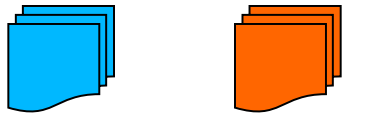
- Excluding direct connections between target concepts 1. at the level of distinct static contexts and 2. at the level of overall dynamic context

- Association Rules Mining on



Results: simulating *Swanson's discoveries

- Manual comparison with systems discovering links based on ABC model (X→Z→Y)
- Statistical evidence



*[Swanson, Persp. in Biol. Med., 1988]

Results: X= Migraine, Y=Magnesium Deficiency

- Bitola system [Hristovski, IJ Med.Inf, 2005]:

Minimum support=0

- # links: 2620
- 2 links with support in [0.1;0,209]

- Proposed approach:

- $\Delta_T = 1$
- minSup=0.3
- minConf=0.4
- σ_{mi}=1
- η_T=2
- [1989;1997]

- # bisociations: 1
- avg support=0.31
- avg *mi*=1.157
- 2/2 intermediate concepts of BITOLA covered

Results: X= Migraine, Y=Magnesium Deficiency

Arrowsmith system [Swanson et al, KDD, 1996]:

Minimum relevance=0

- # links: 598
- 15 links with relevance in [0.89;1)

Proposed approach:

- $\Delta_T = 1$
- minSup=0.3
- minConf=0.4
- σ_{mi} in [0.5;1]
- η_T=2
- [1989;1997]

- # bisociations: 1
- avg support=0.31
- avg *mi*=1.157
- 9/15 intermediate concepts of ARROWSMITH covered
- 2/15 not covered
- 4/15 not present in the used terminology