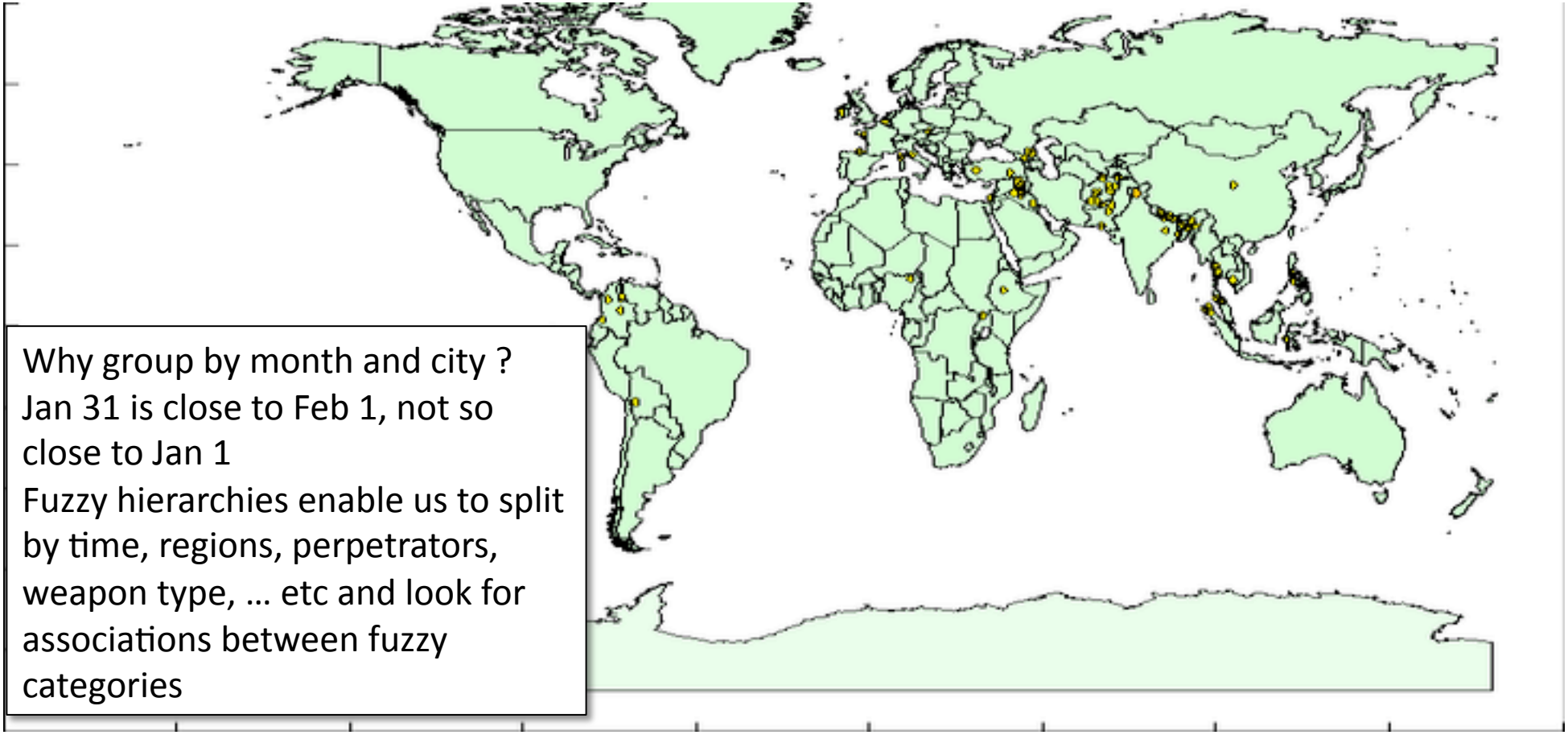
Fuzzy Taxonomies for Creative Knowledge Discovery

Trevor Martin
Zheng Siyao*
Andrei Majidian

1 background - knowledge discovery
2 creativity and the BISON project
3 application to business processes

Also : *School of Computer Science and Engineering, BeiHang University, China

Previous use case – KDD in multiple sources

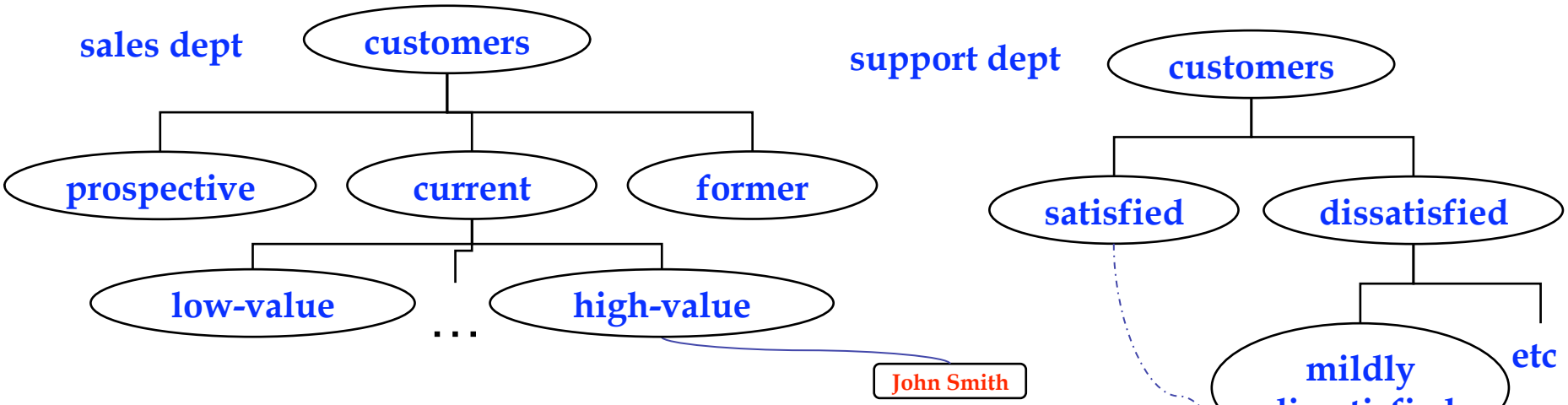


Why group by month and city ?
Jan 31 is close to Feb 1, not so close to Jan 1
Fuzzy hierarchies enable us to split by time, regions, perpetrators, weapon type, … etc and look for associations between fuzzy categories

combines incidents from three sources (up to 80000 incidents of terrorism) into a common representation – when are entities the same, how do we combine and summarise - we can show location of incidents each month (for example)

Knowledge as relations between categories

- it is helpful to know how different hierarchies (views) are related
  - enables reuse of categorised information
  - enables combination of information from different sources

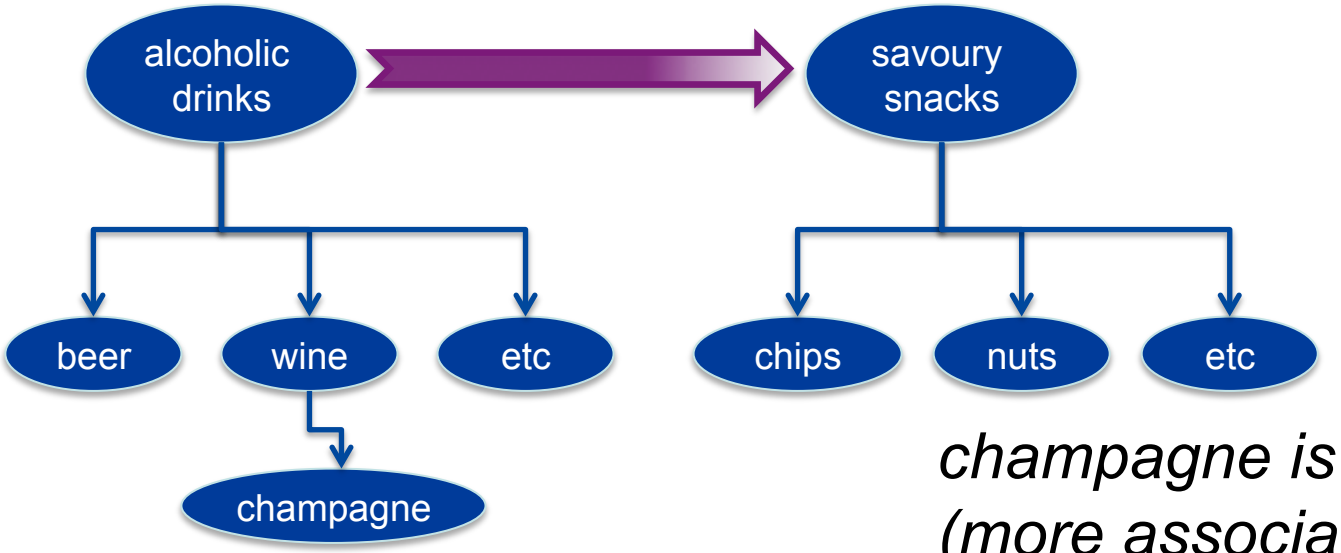


- what can we say with binary logic?
  - a satisfied customer is one who has never complained
  - "all dissatisfied customers are current customers" (false)
  - "at least one current customer is satisfied" (true)

- better approach – flexible categories,strong associations
  - "most high-value customers are satisfied customers"
  - NB dynamic data

Relations, hierarchies and exceptions

- association rules allow us to find approximate relations between categories
  - e.g. 72% of people who buy beer also buy chips/nuts
  - fuzzy categories : alcoholic drinks → savoury snacks

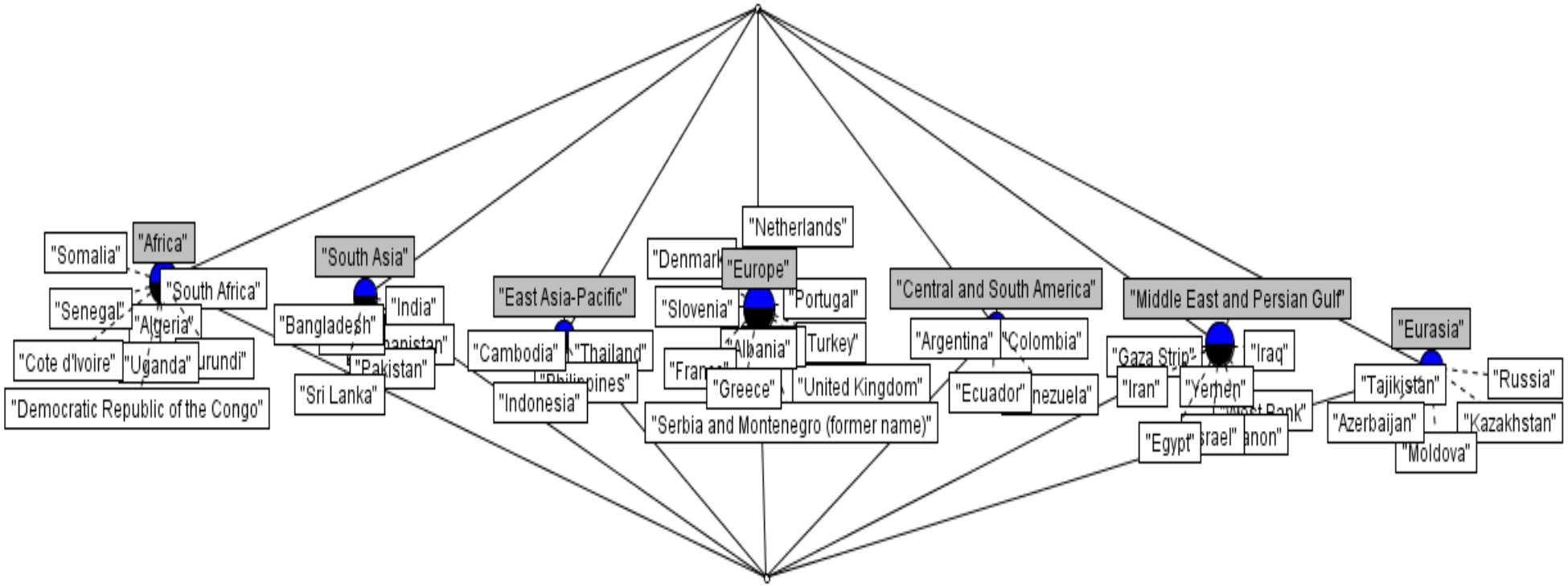


Continual addition of new data means we need to monitor associations over time

champagne is an exception (more associated with chocolates)

Automatic Taxonomy Acquisition

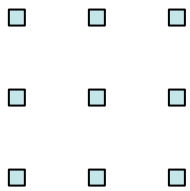
- creating taxonomies is labour intensive
  - often, taxonomic information is embedded in the data
  - can be extracted by formal concept analysis
  - most categories used by humans are not well-defined (fuzzy extensions)



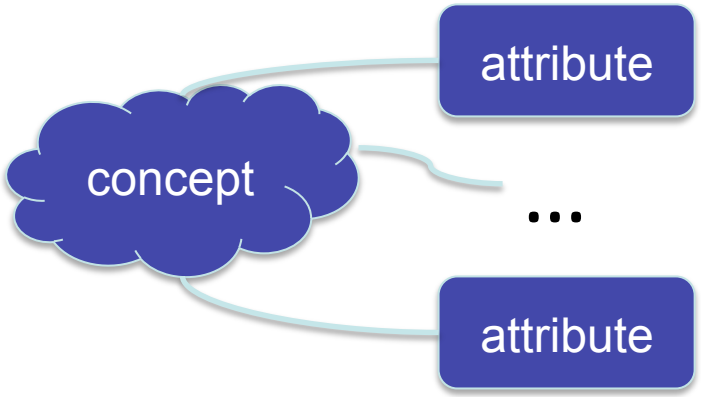
Do concepts constrain creativity?

concepts = convenient groupings

- concepts are central to (conscious) human thought and communication
- logic (*logos* = *word, thought, idea)*
- creativity = finding new concepts and re-interpreting /recombining concepts in novel ways
- Koestler : "*The creative act is not an act of creation in the sense of the Old Testament. It does not create something out of nothing: it uncovers, selects, re-shuffles, combines, synthesizes already existing facts, idea, faculties, skills. The more familiar the parts, the more striking the new whole*"
- e.g. what is the minimum number of straight lines needed to join these dots ?
  Could a computer produce the answer?



Automating Creativity?



choose an attribute (feature)
change it
what are the consequences ?
what is the hardest part of this approach ?

The Gramophone Co. trademark gramophone, 1898

music player
- LP
- stylus
- cabinet
- loudspeaker

FP7-211898 BISON

Bisociation Networks for Creative Information Discovery

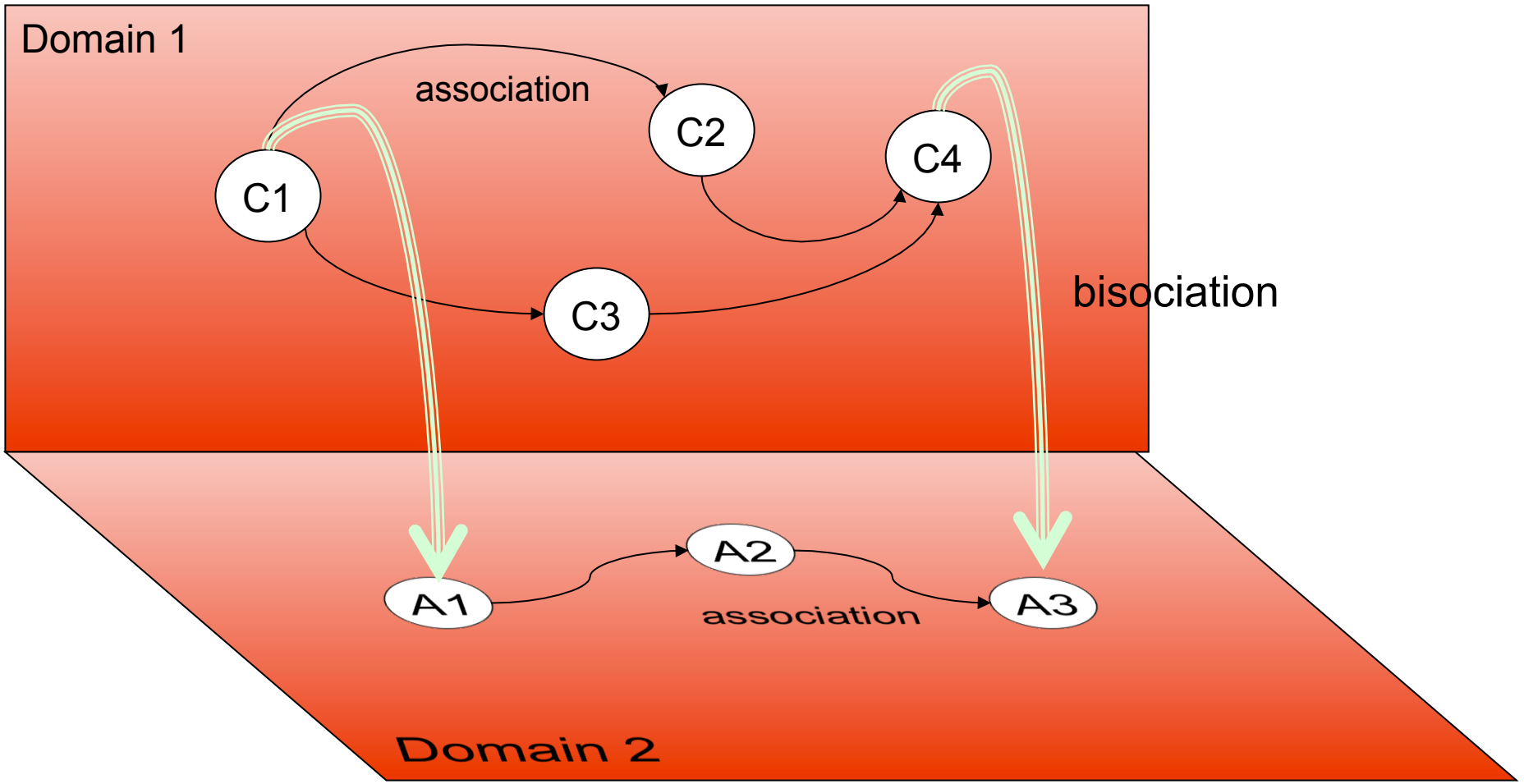
"Develop a bisociative information discovery framework and implement an open-source BISON platform for interactive and scalable processing of massive distributed collections of heterogeneous information content."

"I have coined the term 'bisociation' in order to make a distinction between the routine skills of thinking on a single 'plane', as it were, and the creative act, which, as I shall try to show, always operates on more than one plane."

Arthur Koestler, The Act of Creation

#1: University of Konstanz, Germany (Coordinator): Michael Berthold
#2: University of Ulster, United Kingdom Werner Dubitzky
#3: Josef Stefan Institute, Slovenia Nada Lavrač Igor Mozetic
#4: Katholieke Universiteit Leuven, Belgium Luc de Raedt
#5: Otto-von-Guericke University Magdeburg, Germany Andreas Nurnberger
#6: University of Helsinki, Finland Hannu Toivonen
#7: University of Bristol, United Kingdom Trevor Martin
#8: European Centre for Soft Computing, Spain Christian Borgelt

Bisociation vs Association

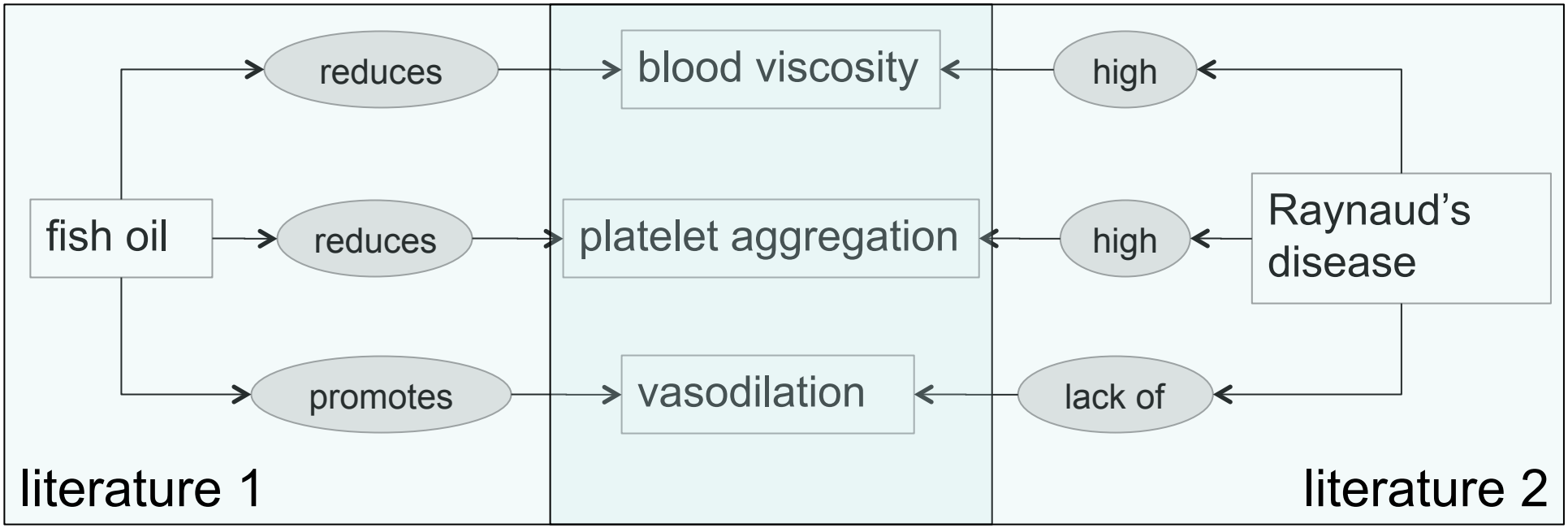


Need – mappings between domains (vocabularies, ontologies)
Arguable – apply standard methods to a “super-domain” ?

Literature-based discovery

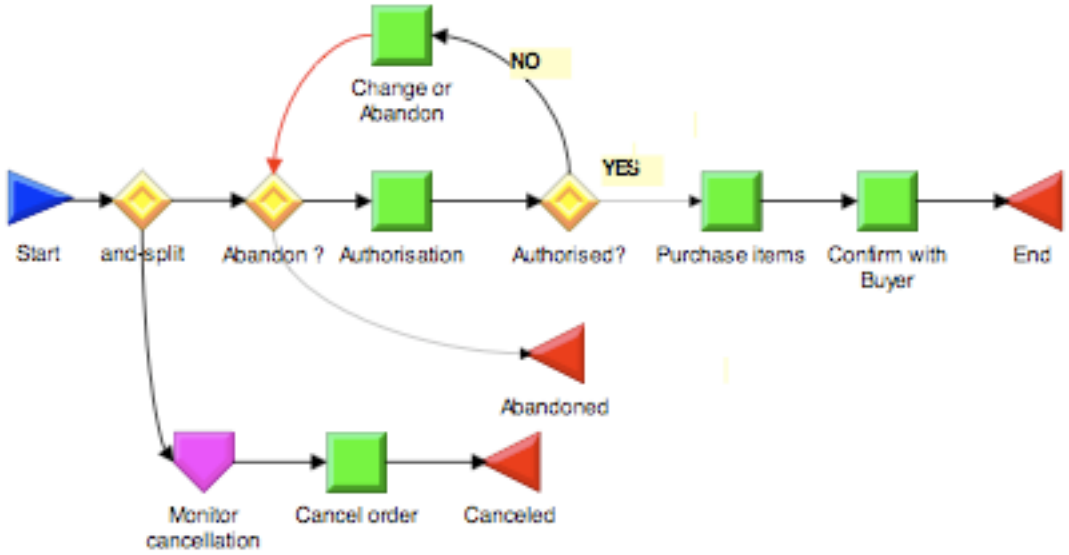
Swanson (1986) from exploration of MEDLINE

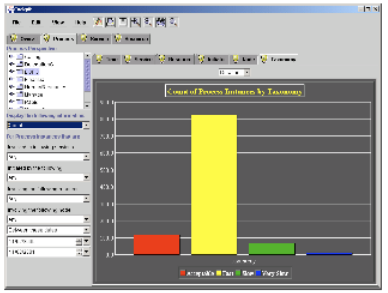
– hypothesis : fish oil to treat Raynuad's disease



1988 : suggested link between magnesium deficiency and migraine
Process can be assisted by MeSH ontology / Unified Medical Language System

Business Process Intelligence





- business processes modelled by sequence of tasks
  - e.g. customer order, fault report, sales enquiry, …
  - monitored at key points (time to respond, number of visits, …)
  - linked by transitions, may have sub-tasks, internal states, …
  - typically specified in XML
  - improve performance by monitoring indicators
  - more radical improvement – “process re-engineering”
  - aim : mining to find abstract process models, apply bisociation

Objects, Attributes and Values

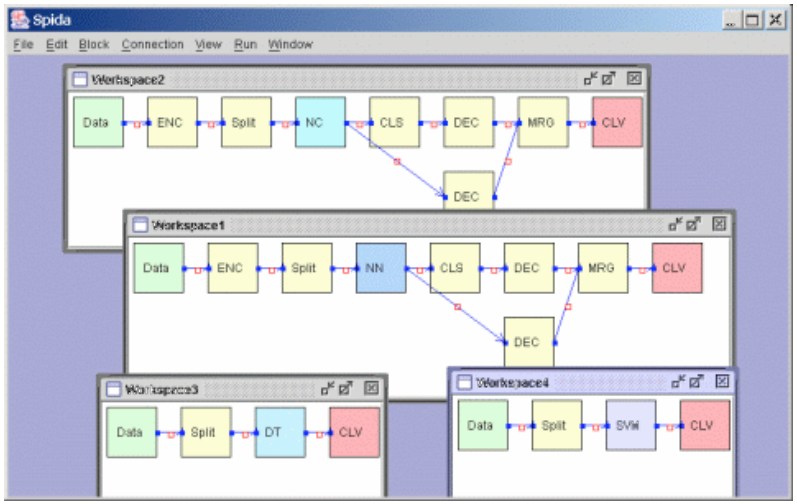
```
<WorkflowLog>
    <Process id="XYZ">
      <ProcessInstance id="1492491">
        <Data>
          <Attribute name="CLEARING_MU">NRERECY1</Attribute>
          <Attribute name="DATA_DATE">2008-10-28T00:00:00.000</Attribute>
          <Attribute name="FAULT_NUMBER">CL0TVQ10</Attribute>
          <Attribute name="FIRST_HANDLE_TYPE">UNKN</Attribute>
          <Attribute name="FIRST_MU">BRDCLIDS</Attribute>
        </Data>
        <AuditTrailEntry>
          <WorkflowModelElement>start</WorkflowModelElement>
          <EventType>start</EventType>
          <Timestamp>2008-10-28T12:14:31.000</Timestamp>
        </AuditTrailEntry>
      …
…
```

Scale – hundreds of processes, tens of thousands of process instances

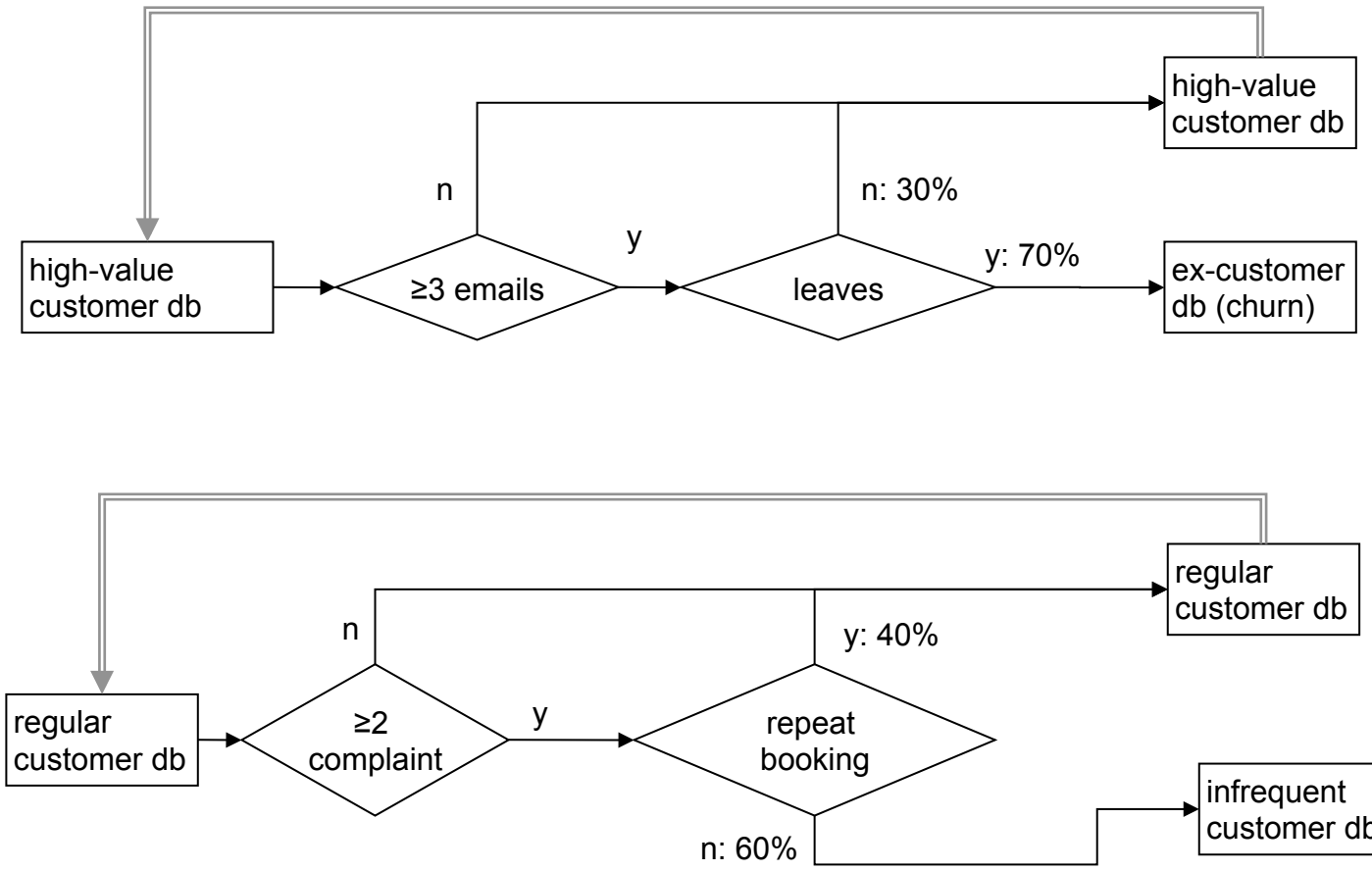
Where does Bison fit ?

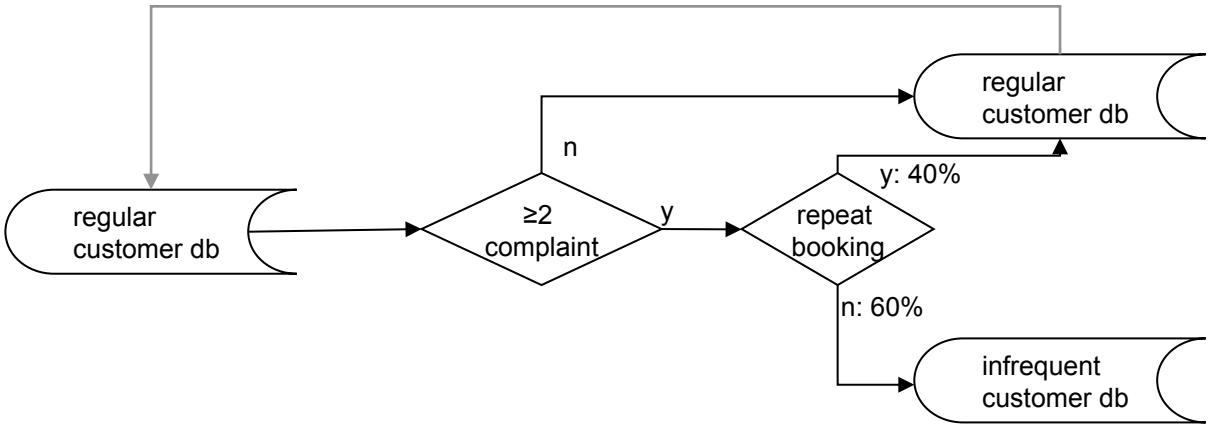
- xml → process graph is relatively straightforward (but underlying taxonomies may need work)
- Bison tasks – identify process similarities
  (intra- or inter- process)

- use similarity metrics to suggest process "transplants"
- *bisociation– take the components apart, change them, put them together in different ways, recognise when we have a good solution*
- benchmark by
  (i) run process simulation package,
  check performance indicators
  (ii) human evaluation
  (if available)

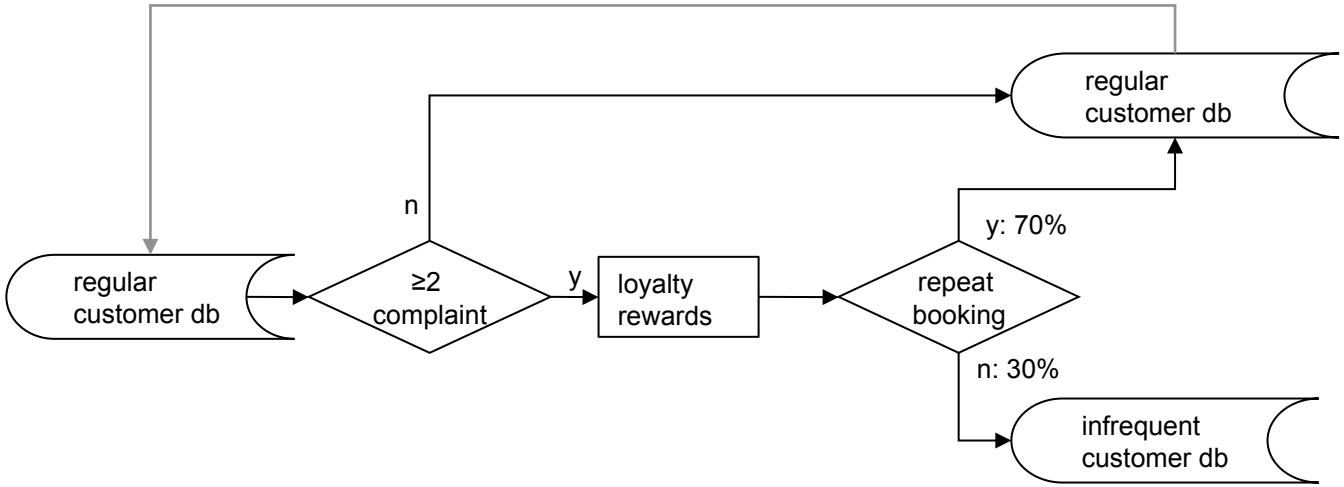


Simplified example





before



after

obvious lesson for the ISP!

many (less obvious) parallels in processes identified in the demonstrator dataset

Finally …

- other (text-based) Bison demonstrators
  - bio- / pharma- literature mining with “semantic” annotations
  - matching research demonstrators to corporate customer “needs” and interests
  - information-finding behaviour in web forum

  (all involve free text plus (hierarchical) keywords)

– Don’t re-invent analogical / case-based reasoning

– Early stage of work – comments welcome
  - semantic markup, uncertain hierarchies, uncertain match between domains → URSW