SURF ACADEMY

VOOR EN DOOR PROFESSIONALS IN HET HOGER ONDERWIJS EN ONDERZOEK

TU/e Technische Universiteit Eindhoven University of Technology

# Educational Data Mining & Learning Analytics for All:

## *Potential, Dangers, Challenges*

Mykola Pechenizkiy

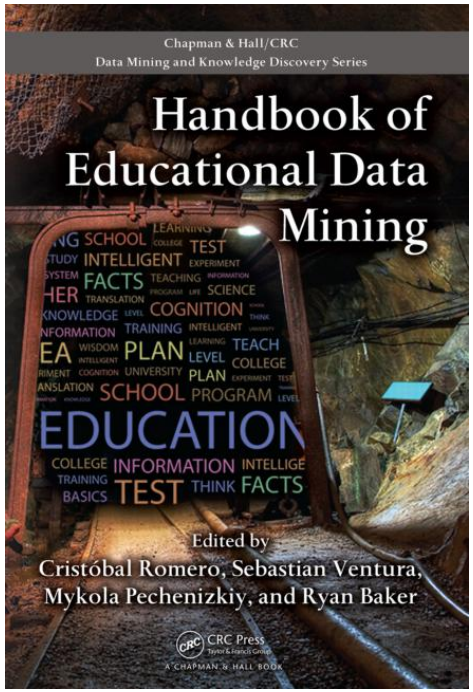
Learning Analytics Seminar
August 30-31, 2011
Utrecht, the Netherlands

http://www.win.tue.nl/~mpechen/talks/las11.pdf

# Short CV of the Presenter

**Mykola Pechenizkiy**
*Assistant Professor* at Dept. of Computer Science, TU/e
*Research interests*: data mining and knowledge discovery; Particularly *predictive analytics* for information systems serving industry, commerse, medicine and education.
http://www.win.tue.nl/~mpechen/ - projects, pubs, talks etc.

Major recent EDM-related activities:





| | |
|---|---|
| Mykola Pechenizkiy & Toon Calders | (Conference Chairs) |
| Cristina Conati & Sebastian Ventura | (PC Chairs) |
| Cristobal Romero & John Stamper | (Posters Chairs) |

# Motivation for This Talk

- Educational Data Mining (EDM)/Learning Analytics (LA) took off – and the question is – what can or should we do about it?
  - more and more educational data becomes available
  - different kinds of data
  - different kinds of data sources
- Unawareness of many stakeholders of what is already available or what is (potentially) possible with EDM/LA technology

# More ICT – More Data Sources

# What Kinds of Data We (Can) Have

- Administrative data
  - Who follows which program, who takes which course, registers for an (interim) exam, reexams
  - Demographics, school grades, etc
- Resource usage data
  - Videocollege, owinfo, studyweb, library resources, …
- LMS (Sakai, Blackboard, Moodle) data
  - More detailed resource usage data
  - Assessment data (online tests)
  - Assignements (text, notes, source code)
  - Forums, collaboration, feedback/help requests
  - Students' evaluation of learning resources
- Educational games, professional learning, e-Health …

# Objectives of This Talk

- Share my vision on and experience in EDM/LA
- Convince you that
  - EDM/LA is a great thing to do
  - There is technology and plenty of concrete techniques already available for developing and integration bits and pieces of EDM/LA into the education at different levels
  - (If there is time left) There are even more challenging research topics that should be studied for having further success

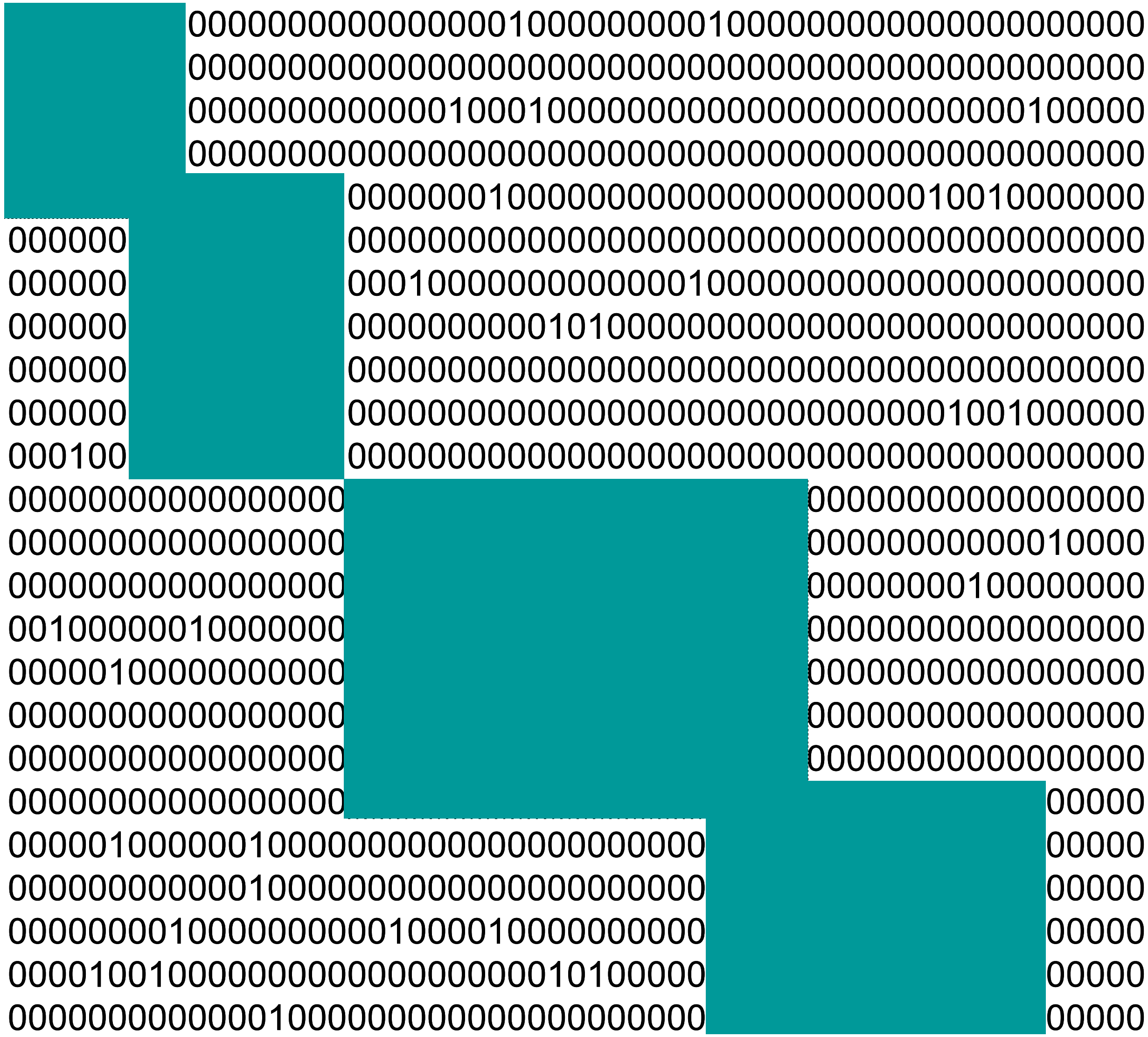
# Outline

- **What EDM/LA is about?**
  - Landscape of tasks & applications
  - Landscape of techniques
  - "Data trumps intuition"
  - Data mining and process mining perspectives
- **Outlook (mostly for the discussion)**:
  - promising directions for immediate development and deployment into the educational practice
  - interesting directions for further research
  - Lots of organizational issues to put EDM/LA into everyday's <u>practice</u>

# Can You See the Pattern?

```
110000000100000001000011001000000011000000000000010000100
111000000100000010000011000000000001010000000000010000000
111000000100000000000011000000000011000000000000010000000
010111000000110111011100100110011100011111100001000000000
000111000000110101011100000110011101011111100001000000000
111000000100000000000011000010000011001000000000010100000
110000000100000000000011000000000011000000000000010000000
111100101100001000000011111000100011000000011000110000000
000111000000110111011100000110011100011111100001000000000
000111011011110111111100000111111000111111100111001100000
000000100000001100000011111000000000000000101100001000000
000000111011111001100000000000110000110000000011100110000
000000111011111001110000000000110000010000000011100110000
000100001011110011100001000001100000100010000011100110000
000000011010110010110110000001100010100000000011100110000
000000100000001000010101111000000000000000001100011000000
000000100000001000000011111000000000000000001100011000000
000000100010001000000011111000000000000000001000011100000
001000100000001000000010111000000000000000001100011000000
000111000000110111011100000110011100011111100001000000000
000111000000110111011100000110011100011111100001000010000
000000100000001000000011111000000000000000001100011000000
```

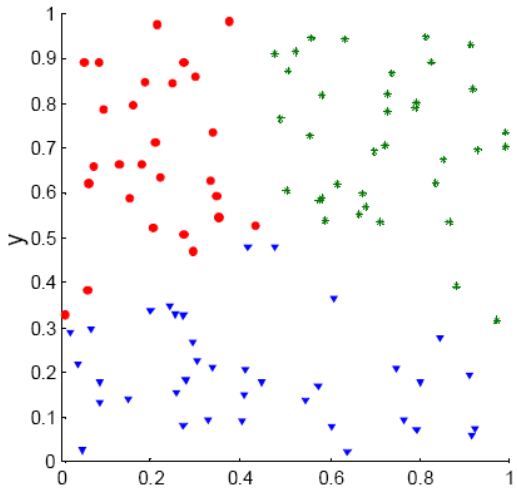
# Can You See the Pattern?

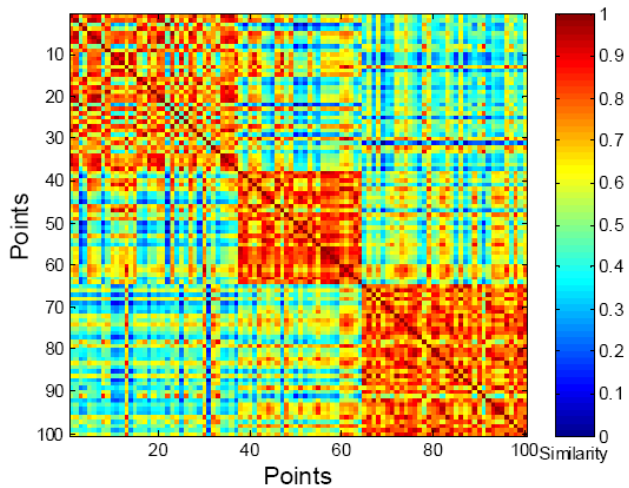


```
         000000000000000010000000001000000000000000000000
         000000000000000000000000000000000000000000000000
         000000000000010001000000000000000000000000100000
         000000000000000000000000000000000000000000000000
                 0000000100000000000000000000010010000000
000000           0000000000000000000000000000000000000000
000000           0001000000000000001000000000000000000000
000000           0000000000101000000000000000000000000000
000000           0000000000000000000000000000000000000000
000000           0000000000000000000000000000001001000000
000100           0000000000000000000000000000000000000000
00000000000000000                       00000000000000000
00000000000000000                       00000000000010000
00000000000000000                       00000000100000000
00100000001000000                       00000000000000000
00000100000000000                       00000000000000000
00000000000000000                       00000000000000000
00000000000000000                       00000000000000000
00000000000000000                                   00000
00000100000010000000000000000000000                 00000
00000000000010000000000000000000000                 00000
00000000100000000001000010000000000                 00000
00001001000000000000000000000101000                 00000
00000000000001000000000000000000000                 00000
```

# This Toy Example Illustrates Two Major Problems

- It is difficult for us to see patterns even when the data is homogeneous and "simple", and the patterns are evidently strong
- It is difficult not to fall into finding/seeing the "patterns" that are really not there
  - Not only our eyes but also tools we use can fool us

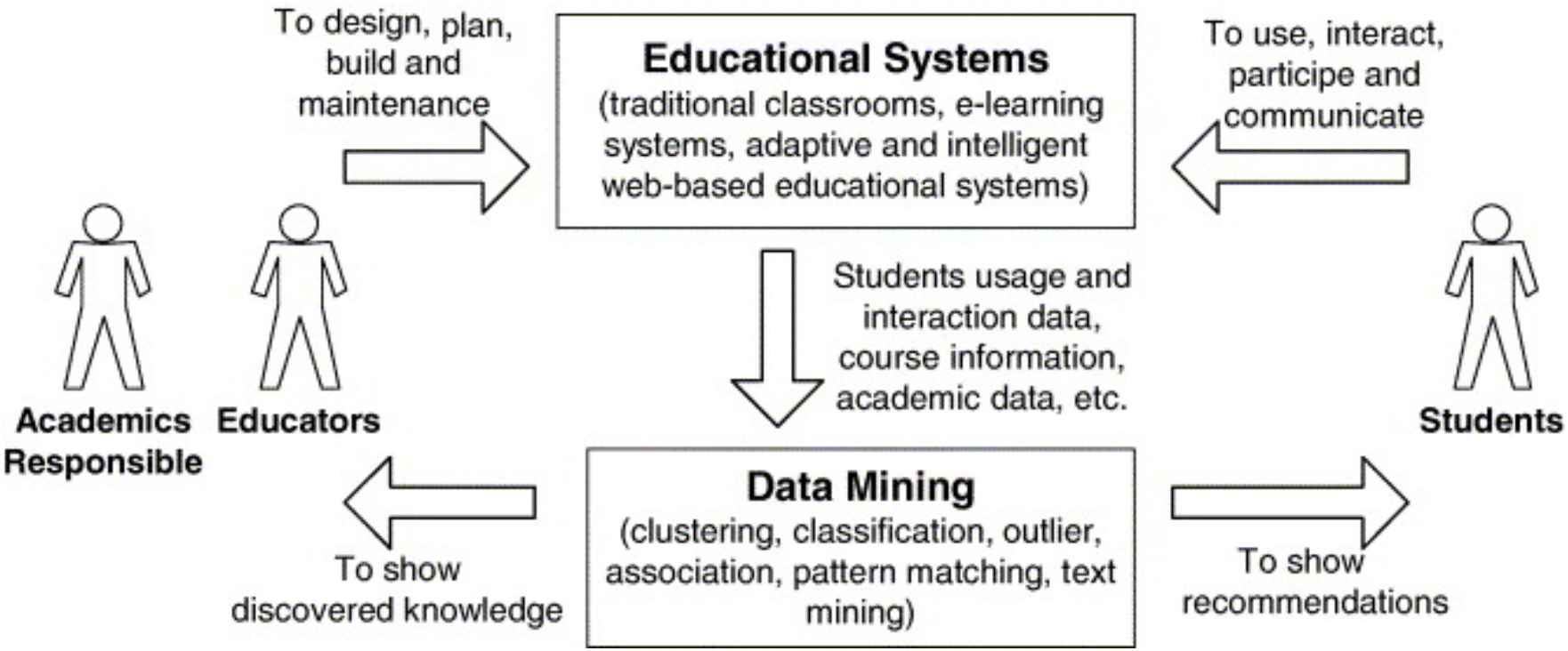


# KDD/DM/EDM/LA – you name it

- **"the use of sophisticated data analysis tools to discover *previously unknown, valid patterns and relationships in large data sets."*** ***(Hand, Mannila, Smyth)***
  - Huge sets of data are being collected and stored
  - Analyzing all data "manually" becomes impossible
- Educational Data Mining can be seen as
  - the process of discovering useful information from the large amount of electronic data collected by educational systems for
    - Learners/students, Teachers, Tutors, Study Advisors, Directors of Education, Educational Researchers, …

# EDM in a Nutshell



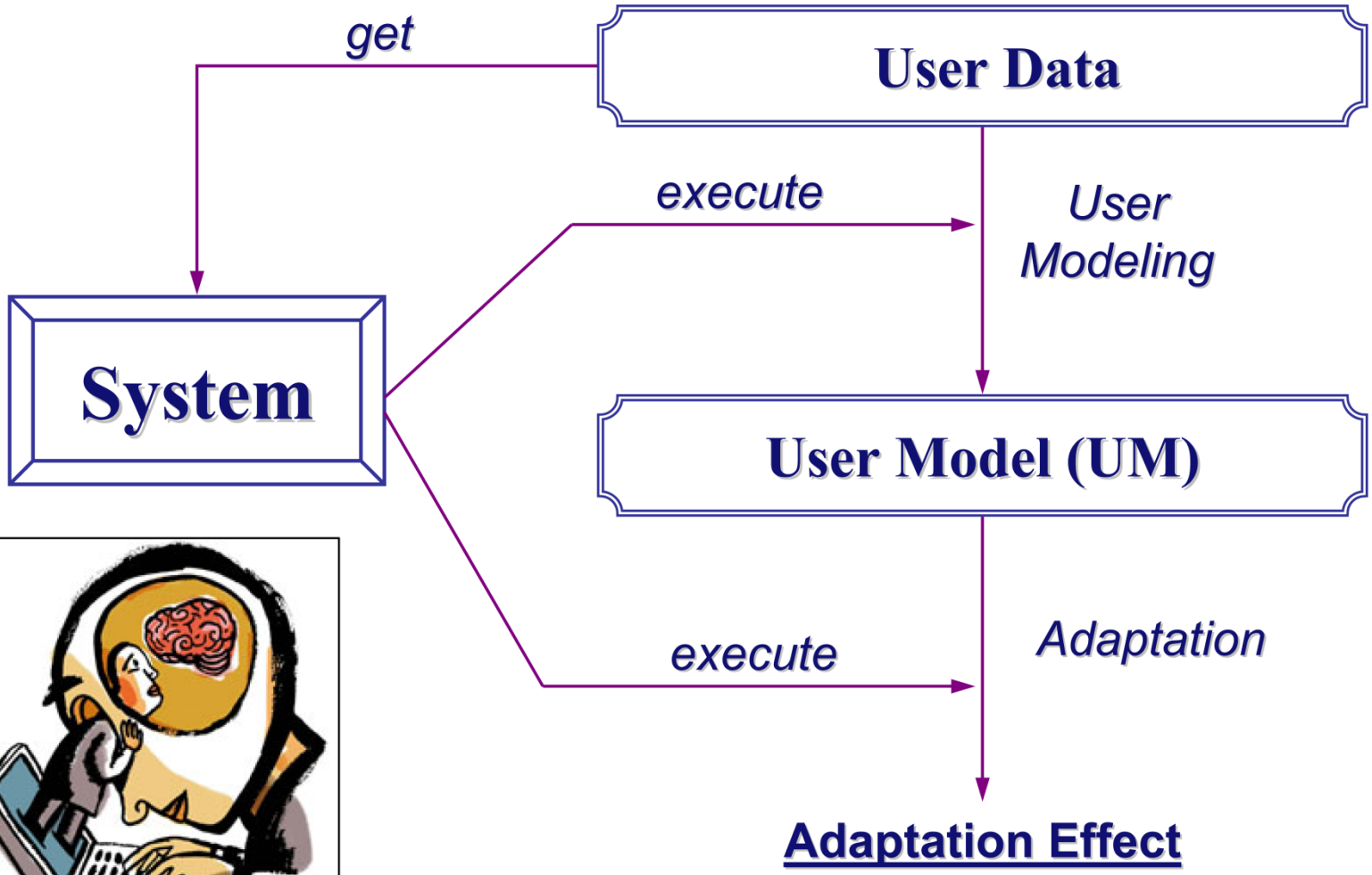

Educational data mining: A survey from 1995 to 2005

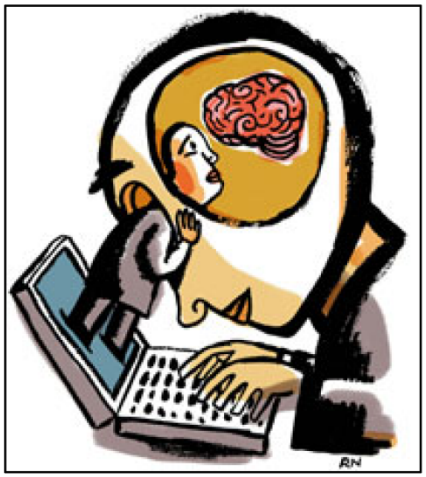
C. Romero & S. Ventura, Expert Systems with Application 33(1)

# What Fields Lay Foundations for EDM/LA

- **Knowledge Discovery from Databases or Data Mining, Process Mining**
- Information Visualization, Visual Analytics
- Recommender Systems, Search and Information Retrieval
- Social Network Analysis (SNA), Text Mining, Sentiment Analysis
- AI in Education (AIED), Intelligent Tutoring Systems (ITS), Adaptive Educational Hypermedia (AH), User Modeling (UM), Technology Enhanced Learning (TEL),
- Psychometrics, Educational Research

# Student Modeling in ITS/AEH



# EDM: Data ⇔ Approach ⇔ Knowledge

**Interactions data**
- Usage logs & contexts

**Administrative data**
- Enrolments
- Results
- Payments
- Graduation
- Employment

**"Feedback" data**
- Opinions
- Preferences
- Needs

**Descriptive data**
- Demographics
- Characteristics

Classification

**Categorizing students**

Clustering

**Grouping similar students**

Association Analysis, Sequence mining

**Find courses taken together or Popular (parts of) study programs**

Visual Analytics

**Facilitate reasoning about the process or results via interactive data/model visualization**

Process mining

**Understanding study curricular**

**Goals**

- Identify high risk students
- Predict new student application rates
- Predict students retention/dropout
- Course planning & scheduling
- Faculty teaching load estimation
- Predict demand for resources (library, cafeteria, housing)
- Predict alumni donation

# Categorizing Students

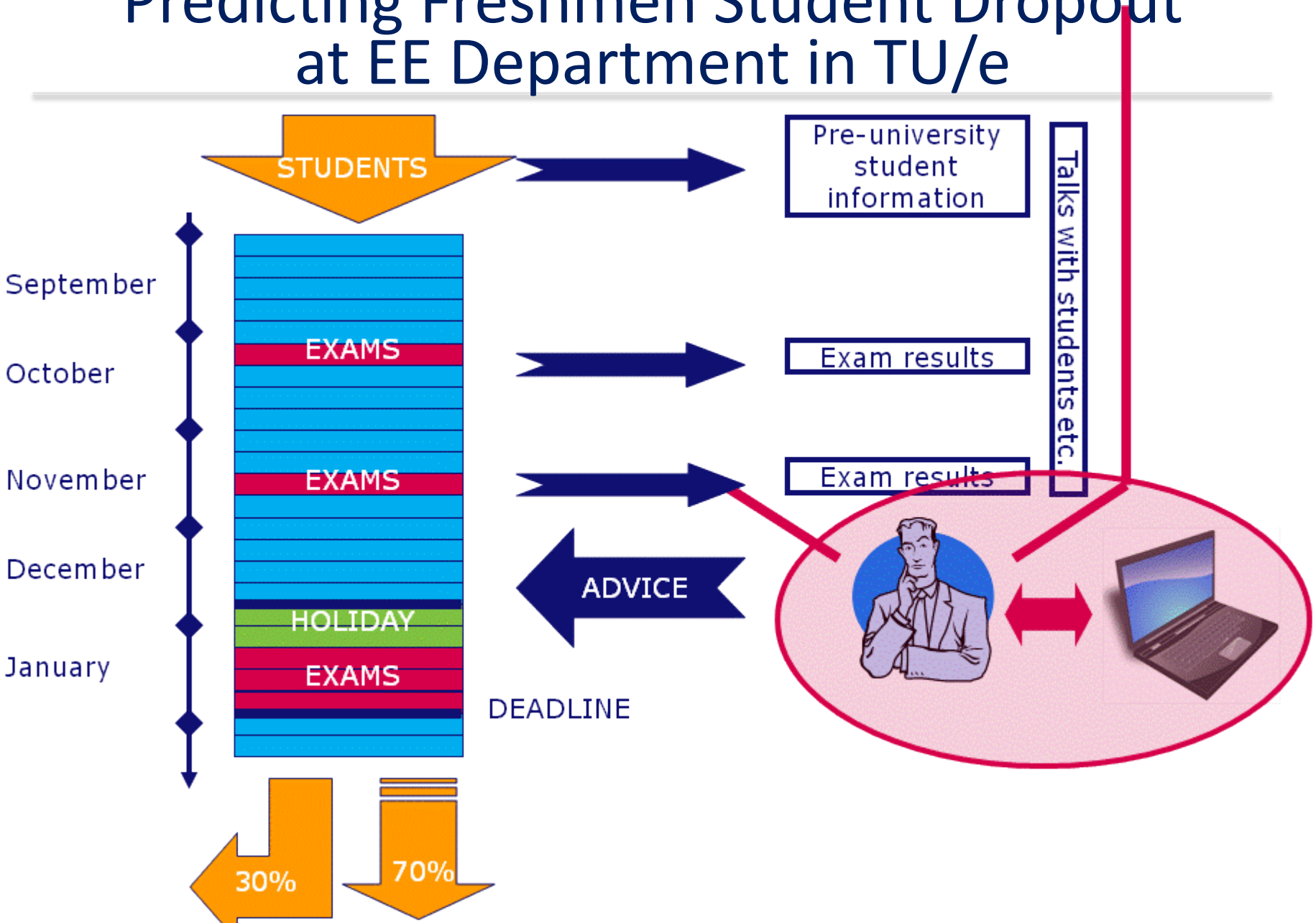
## Student roles in a group project

## Learning style dimensions

# Predicting Student's Success

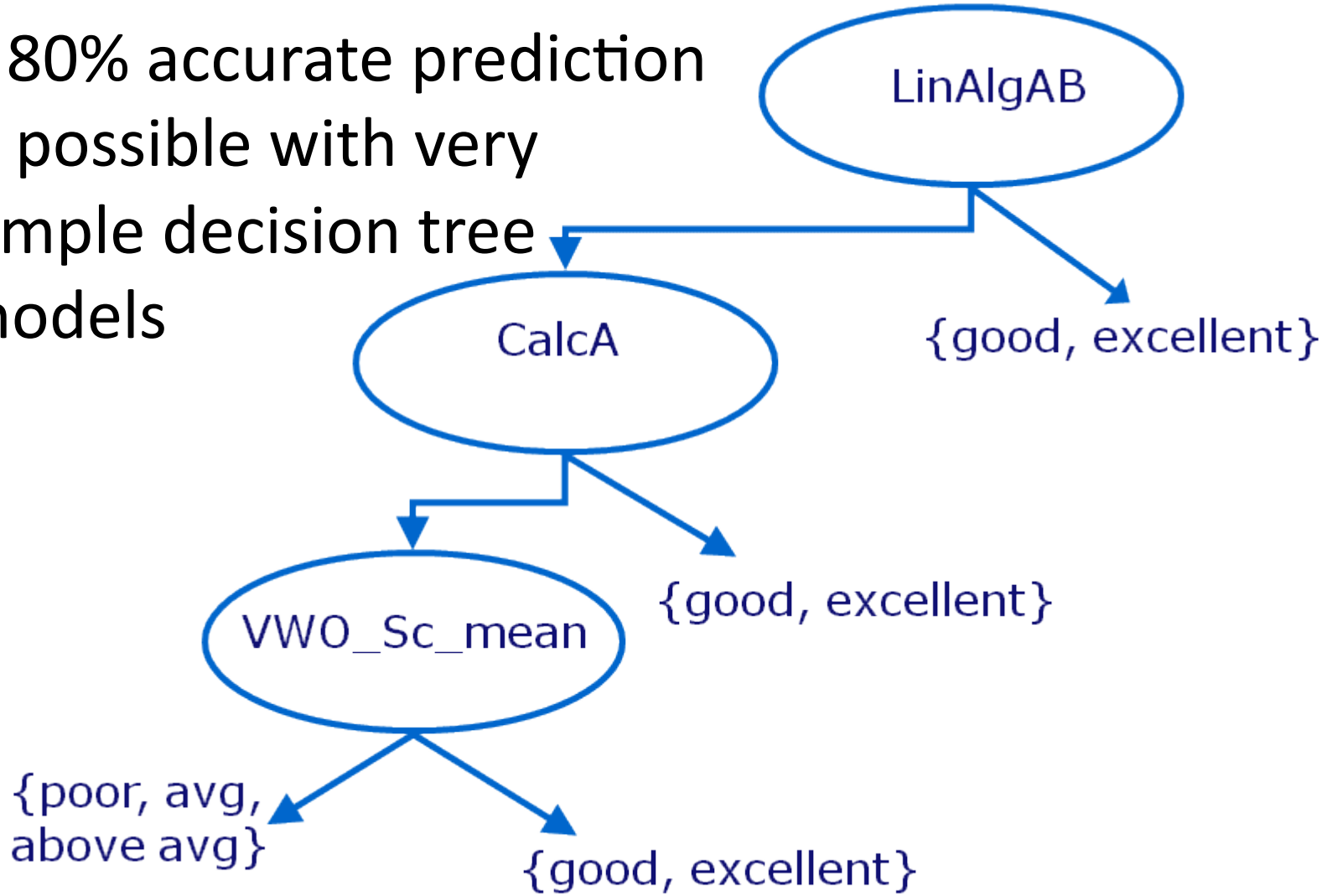
- Predicting the group project outcome
- Predicting whether (and when) a student will eventually graduate
- Predicting student drop out
- Discovering exceptional/atypical behavior

# Predicting Freshmen Student Dropout at EE Department in TU/e

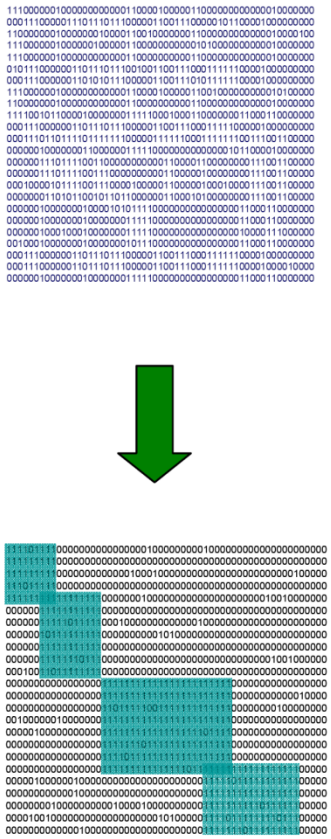


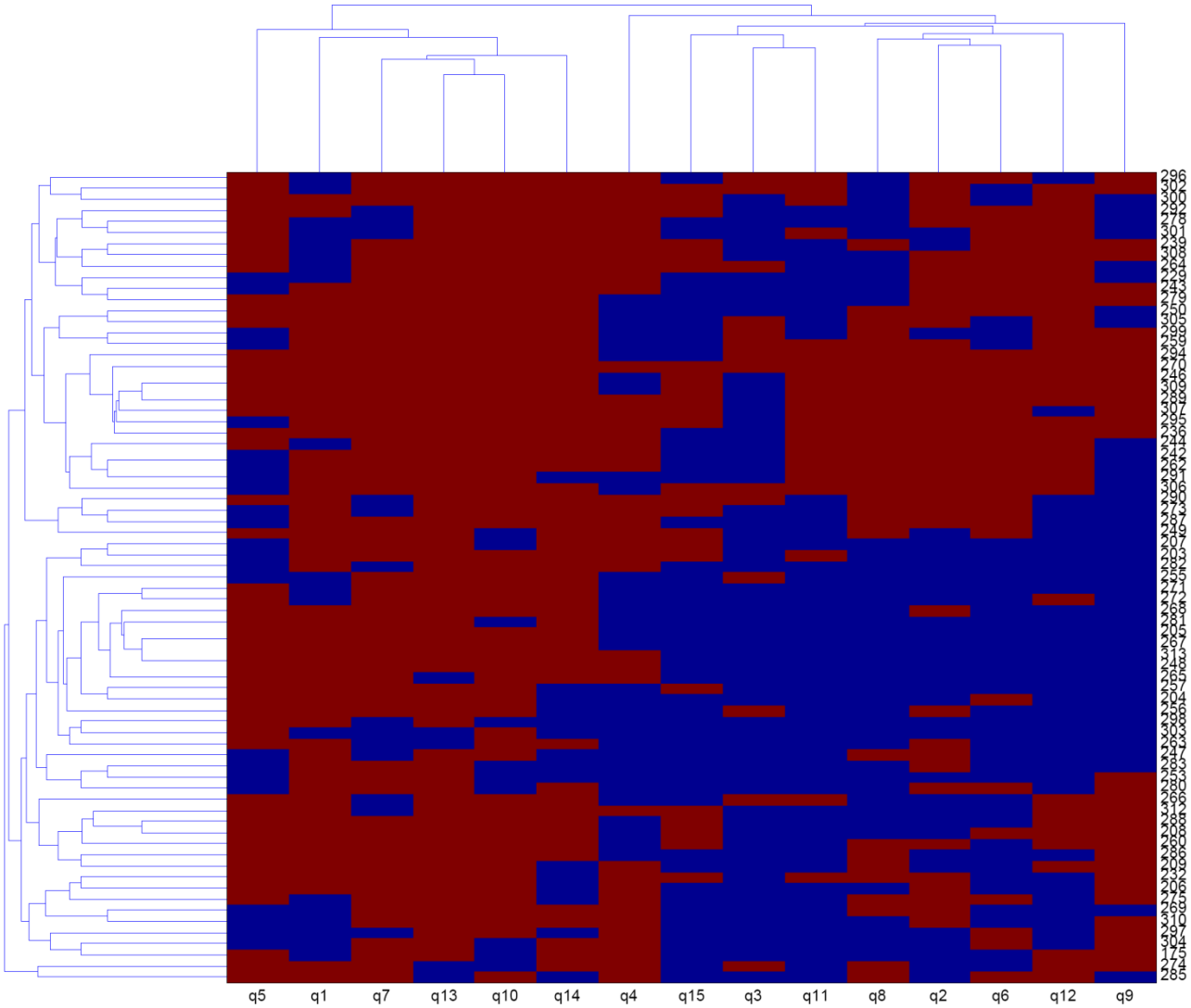
# Results of the Case Study

- ~ 80% accurate prediction is possible with very simple decision tree models



# Which Questions (tasks, courses, etc) Are Difficult/Easy for Which Students?





# (Visual) Data Analytics in EDM

Messages
Groups/courses
Schedule
Tests
Printouts
Questionnaires

Announcements
Study guide
Students
Teachers
Staff
Classrooms

Annual planning

Forms
Notification preferences
Change password

## Class Seniors 07

26 students | Absences | Absences per week | ...uncleared | ...unhandled | Tests

**Alexander Student**

| Date | | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| Thu | 14.4.2011 | | | | | | CORI | | | 1 |
| Wed | 6.4.2011 | CORI | | | | | | | | 1 |
| Thu | 24.3.2011 | | CORI | | | | | | | 1 |
| **Total** | | | | | | | | | | 3 |

**Alma Student**

| Date | | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| Wed | 26.1.2011 | | | | | | CORI | | | 1 |
| **Total** | | | | | | | | | | 1 |

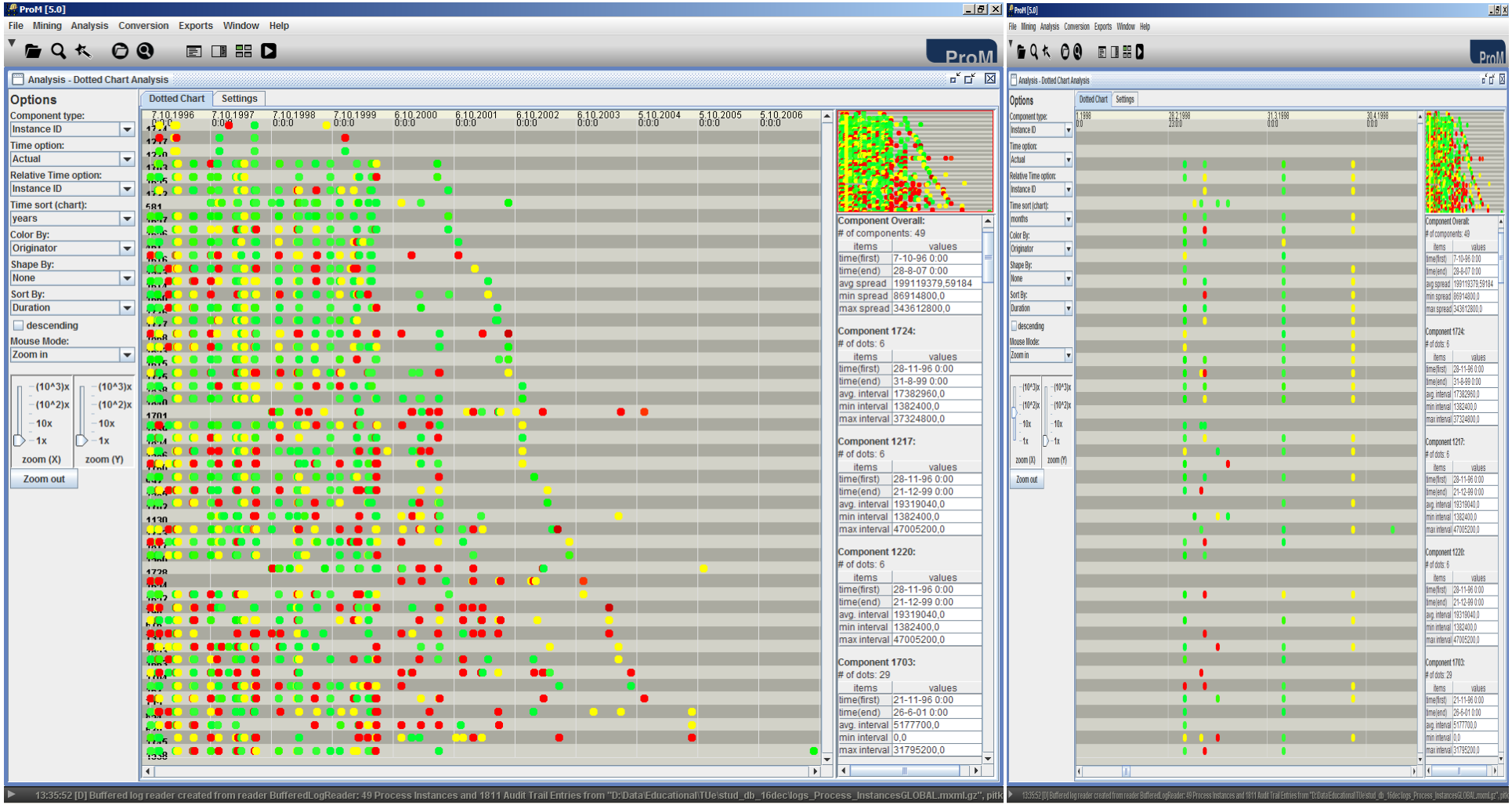
**Beatrice Student**

| Date | | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| Thu | 28.4.2011 | CORI | CORI | CORI | CORI | CORI | | CORI | | 6 |
| **Total** | | | | | | | | | | 6 |

**Beth Student**

| Date | | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| Wed | 11.5.2011 | | | | | CORI | | | | 1 |
| **Total** | | | | | | | | | | 1 |

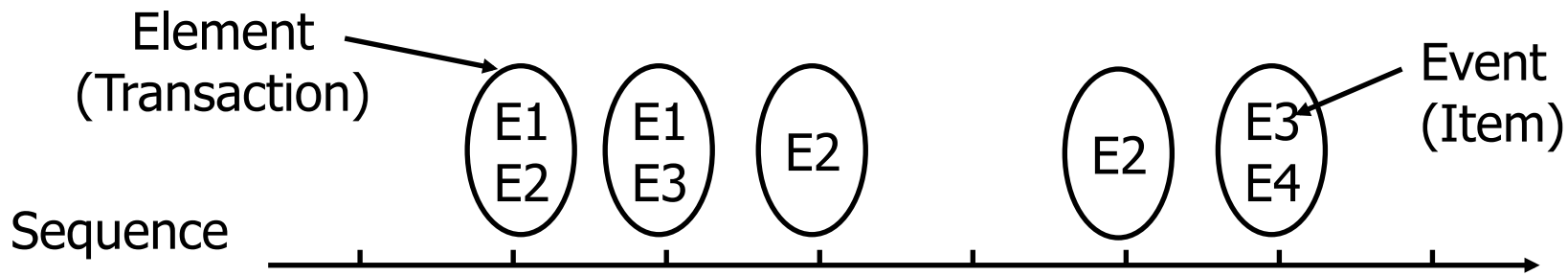
# Students' Performance Started in the Same Year




- Many more examples of Visual Analytics from MagnaView

# Association Analysis/Sequence Mining

- Given: a database of sequences
- Task: Find all subsequences with support ≥ *minsup*
- *E.g. Administrative data as* **Sequence Database**
  - **Sequence** is the history of enrolments into courses of a student
  - **Element** (Transaction) is a set of courses taken in a particular semester (or block)
  - **Event** (Item) – one particular course



# Application Scenarios

- Scenario 1: Find most common types of behavior (and cluster them)
- Scenario 2: Find emerging patterns: such patterns, which capture significant
  - differences in behavior of students who graduated vs. those students who did not
  - changes in behaviour of students from year 2006-07 to 2007-08.
  - in both cases we search for such patters which supports increase significantly from one dataset to another (i.e. in space in the first case and in time in the second case)
- Scenario 3: After finding a bottleneck, find frequent patterns that describe it, i.e. for which students it is the bottleneck and why

| Student | Timestamp | Events |
|---|---|---|
| A | S1 | 2, 3, 5 |
| A | S2 | 6, 1 |
| A | S3 | 1 |
| B | S1 | 4, 5, 6 |
| B | S3 | 2 |
| B | S4 | 7, 8, 1, 2 |
| B | S5 | 1, 6 |
| C | S1 | 1, 8, 7 |

| Student | Graduated |
|---|---|
| A | Yes |
| B | No |
| C | Yes |

# Common Dangers with Mining

- Implication ≠ causality
  - "Diet Coke → Obesity" or "Intensive Care → Death"
- Forgetting about silent evidence and other biases
  - Data driven intelligence is based primarily on the secondary data analysis, i.e. the data was collected for something else rather than particular hypothesis testing
- Data dredging
  - "Torturing the data until they confess"
  - Overfitting – treading noise or some random behavious observed in the data as significant patterns
- Discrimination, False discoveries, Interpretability, Redundancy and volume of output knowledge, …

# Simpson's Paradox

- Success rate of simple and complex operations in two hospitals

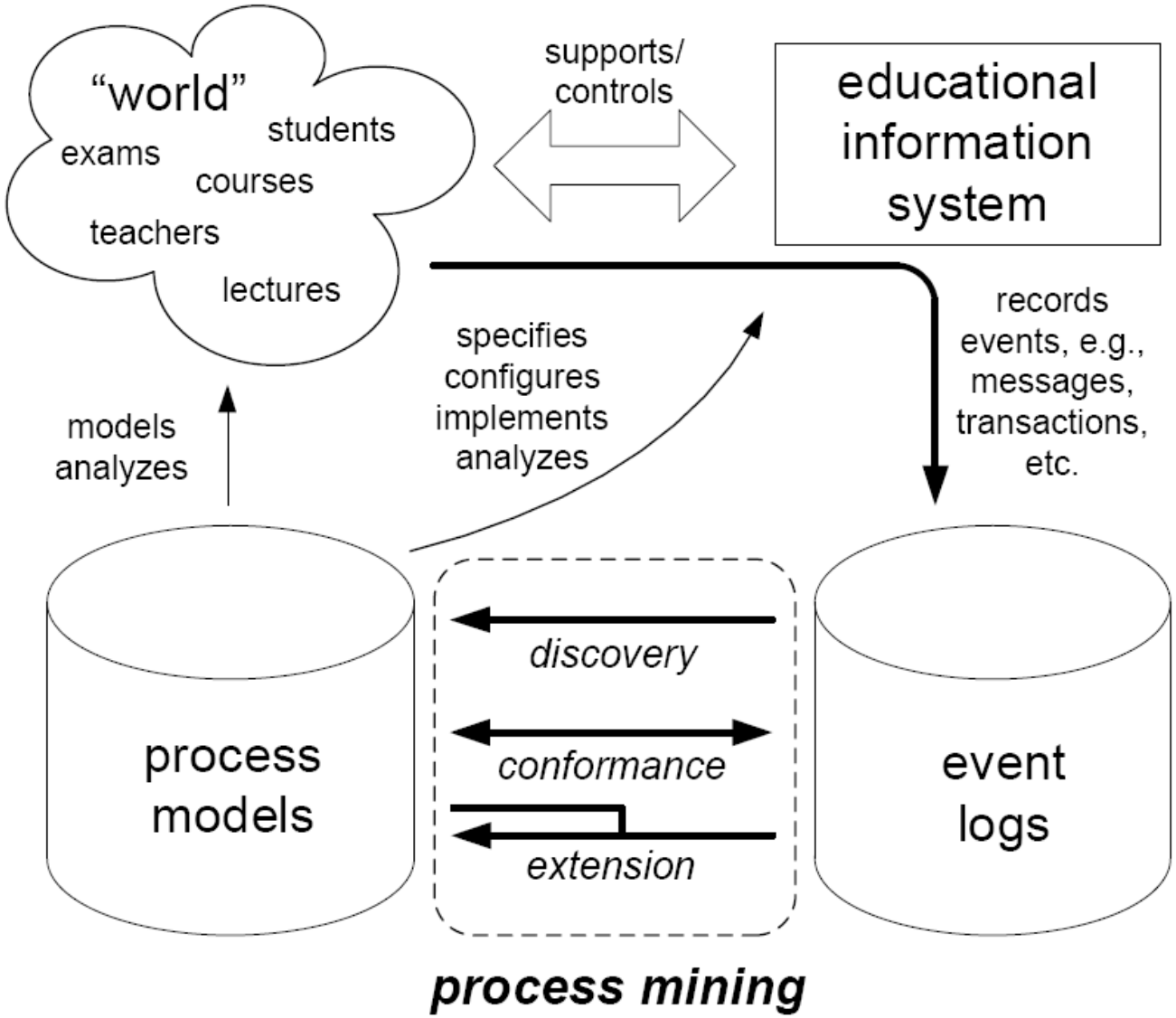
| | Academic | Local |
|---|---|---|
| Simple | 95% | 92% |
| Complex | 75% | 60% |
| Total | 78% | 89% |

| | Academic | Local |
|---|---|---|
| Simple | 190/200 | 920/1000 |
| Complex | 750/1000 | 60/100 |
| Total | 940/1200 | 980/1100 |

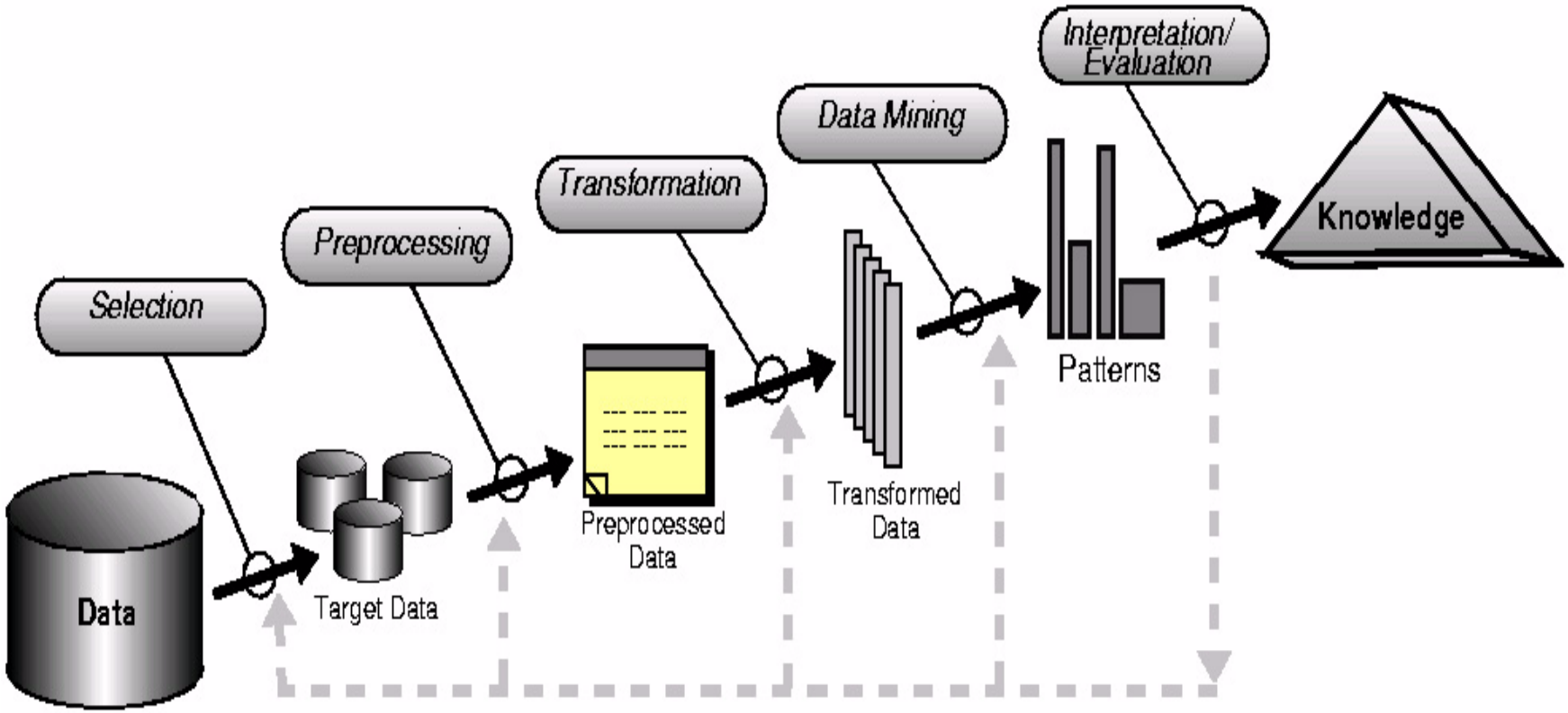
# Many “mining” Research Challenges

- There are some EDM success stories, but also there are plenty of research questions to be solved, like
  - Use of background knowledge in mining
  - Mining for complex patterns, graph mining/ process mining
  - Discovery of emerging and evolving patterns
  - Providing intuitive visualization and explanations
  - …

# (Educational) Process Mining Framework

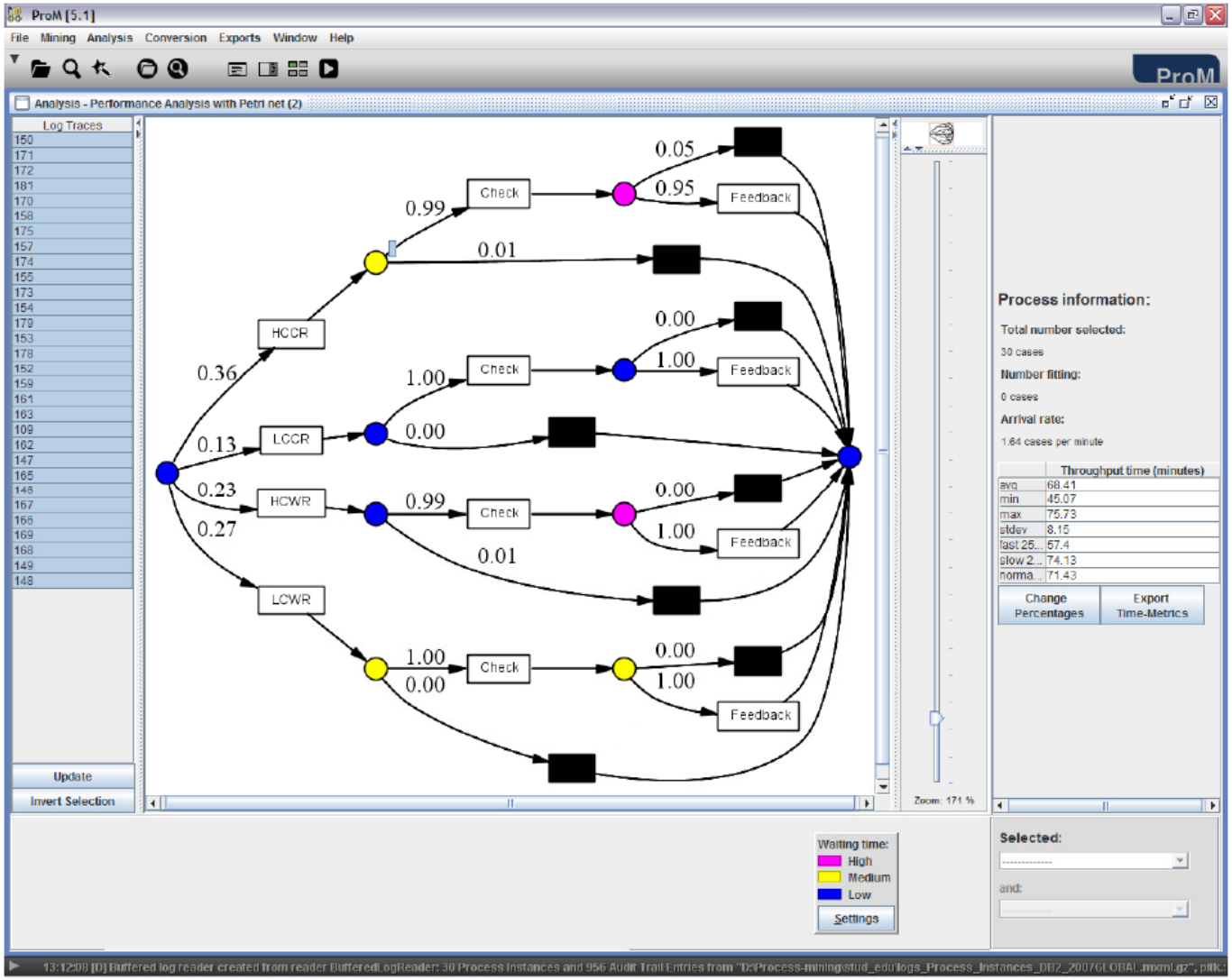


# Not to be Mixed with KDD as a Process

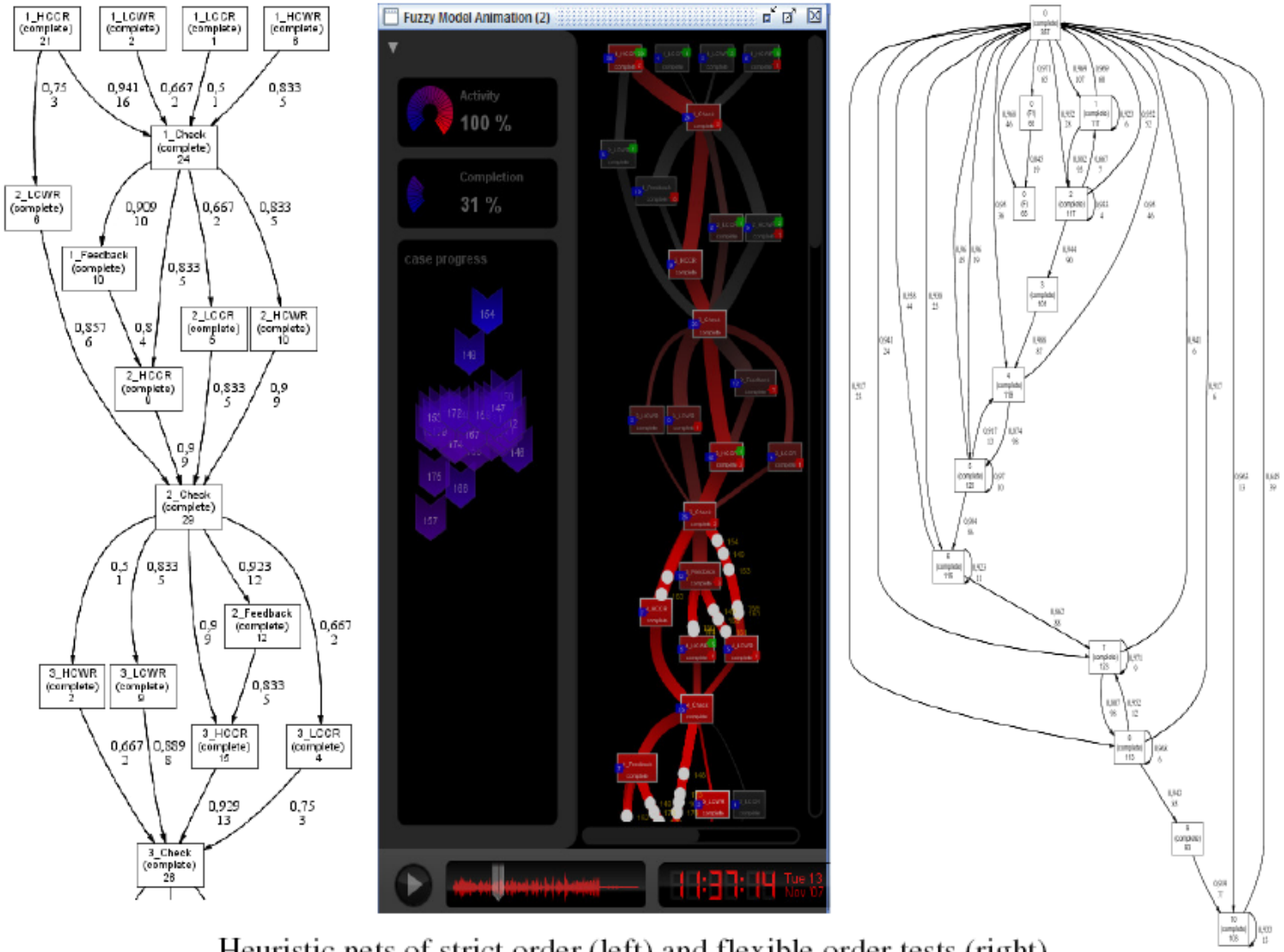



Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R.,
*Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1997.

# Process Analysis/Conformance Checking ex.



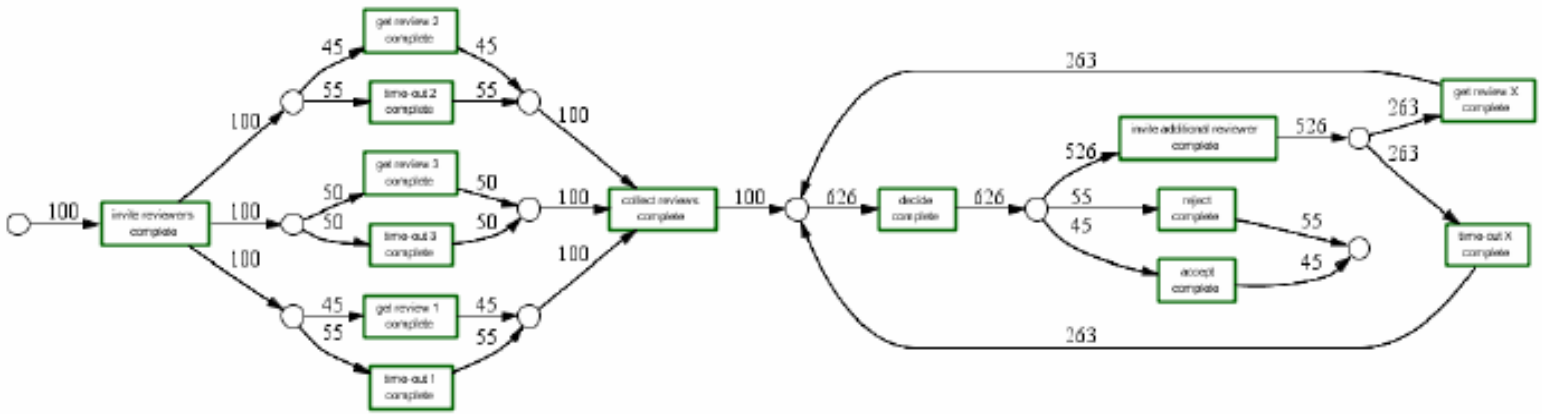
# Process Discovery Example




Heuristic nets of strict order (left) and flexible order tests (right)

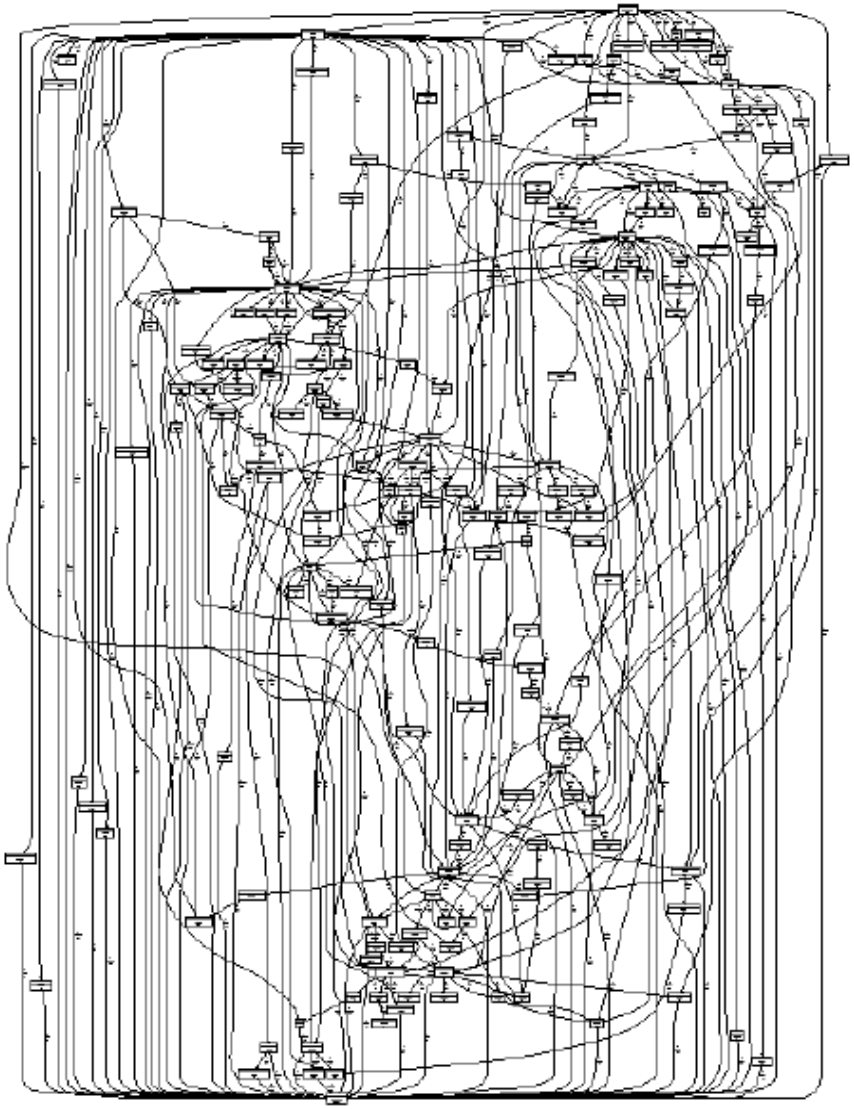
# The Objective is to Mine …

- structured, easy to understand process models
  - just like this one

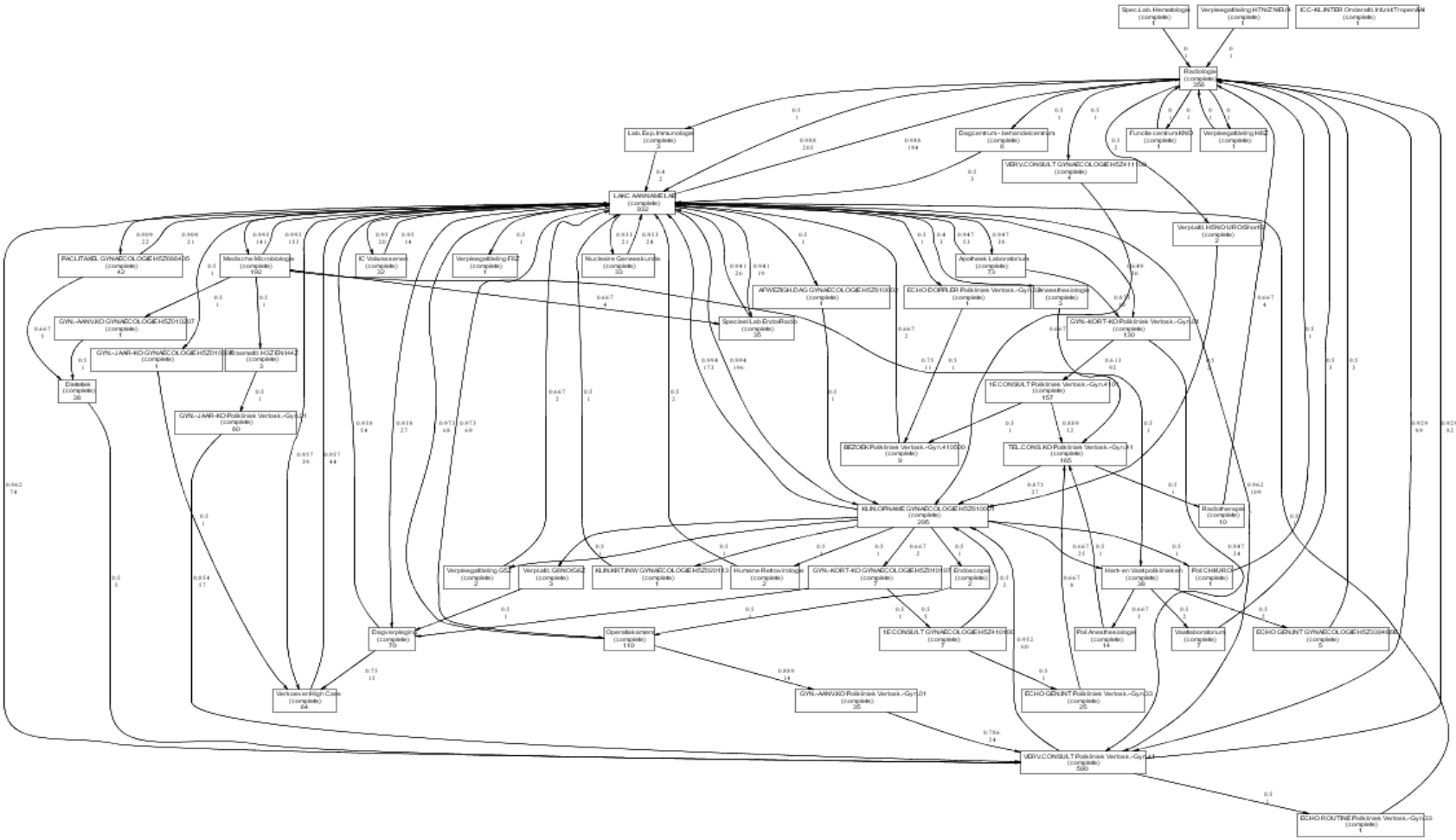



- but …

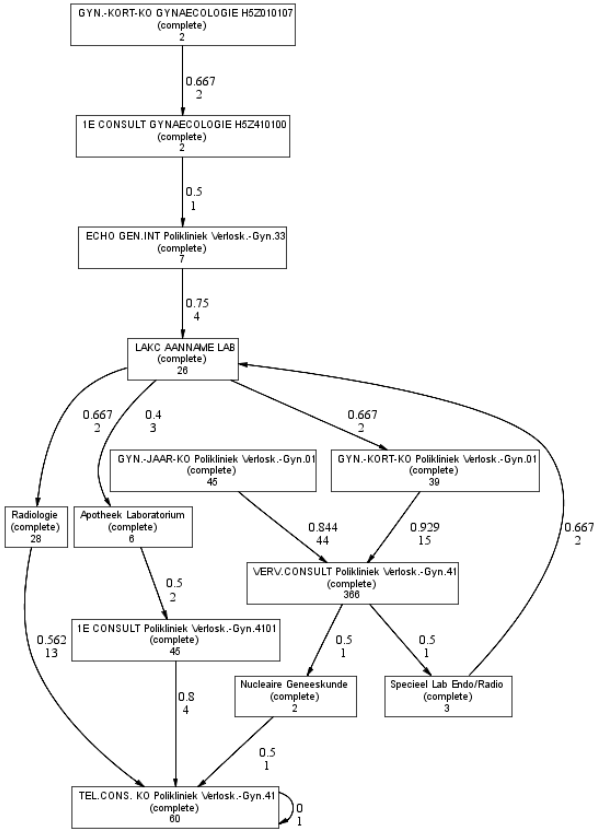
# … in Reality They May Look Like Spaghetti



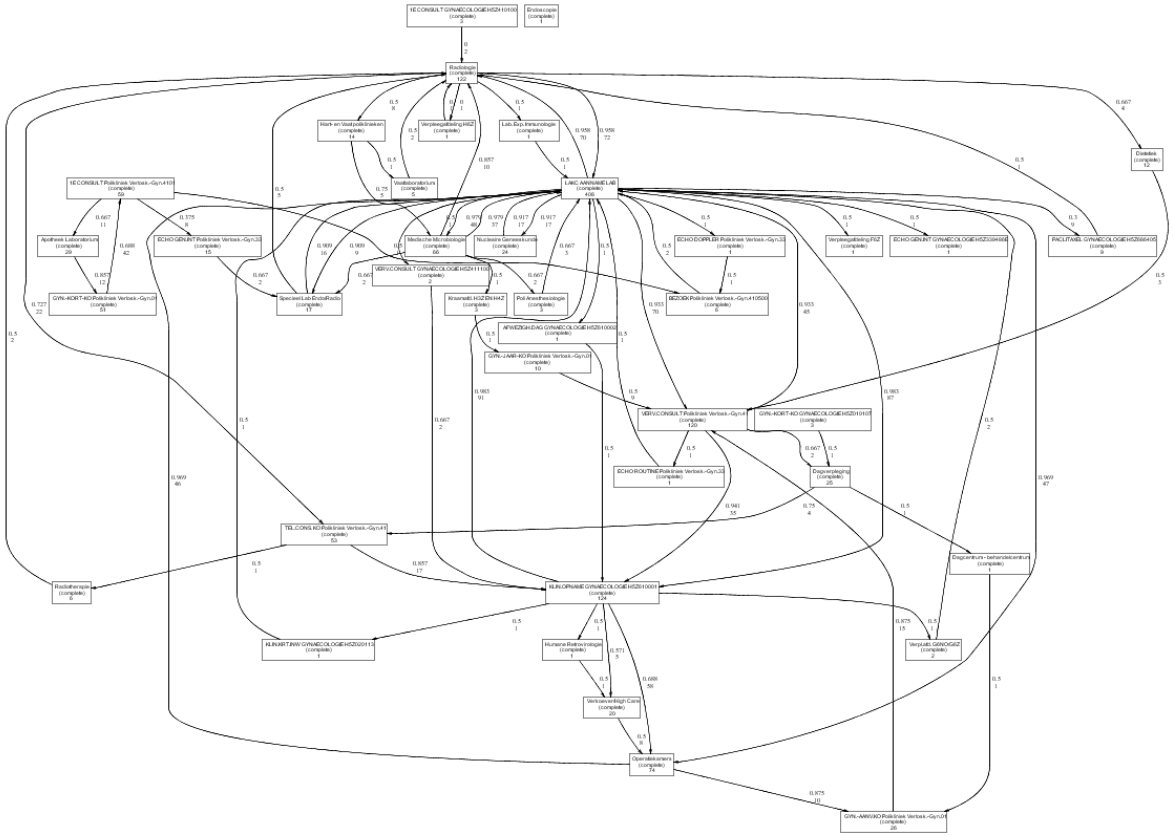
# Process Model from the Original Log



# Process Models from the Clusters




**Diagnosis process**

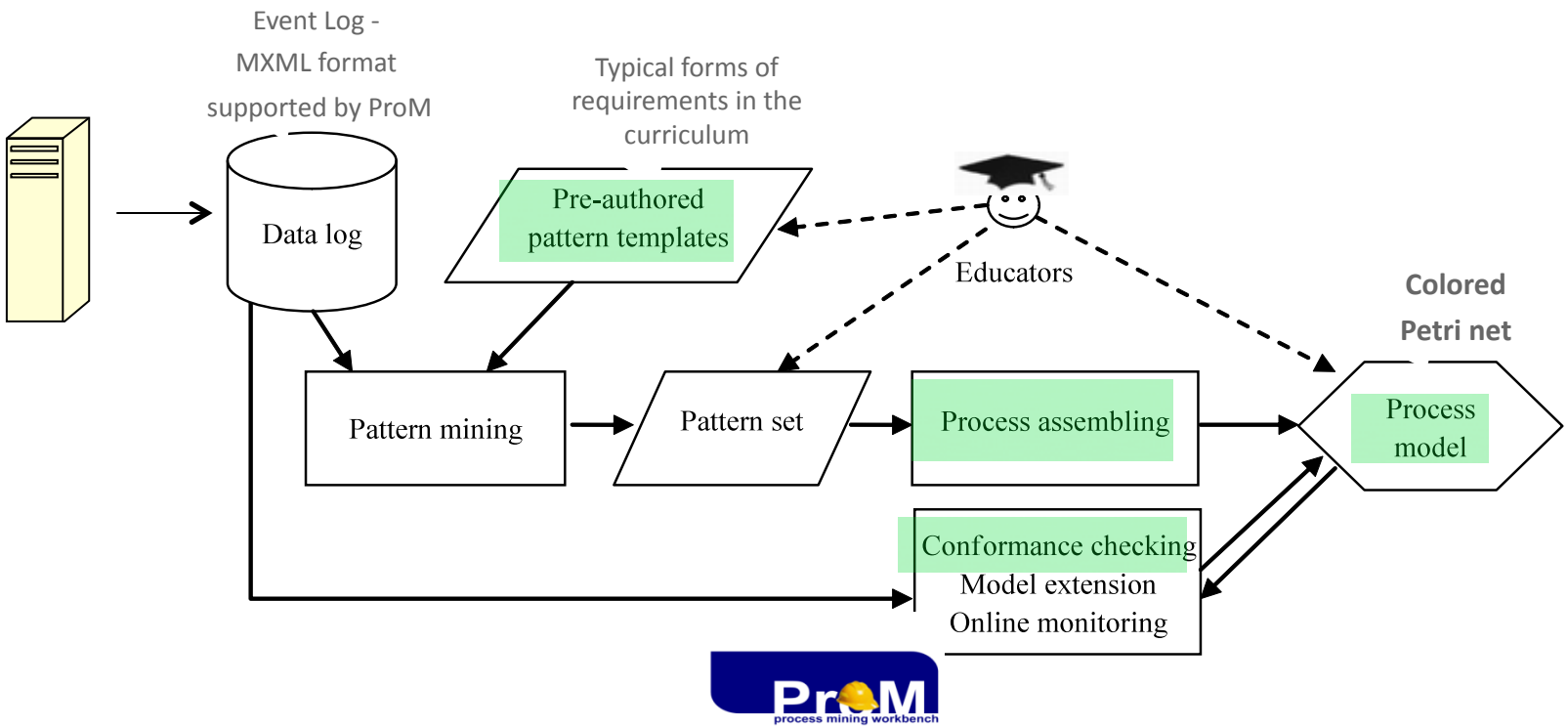


**Treatment process**

# A More Sound Approach

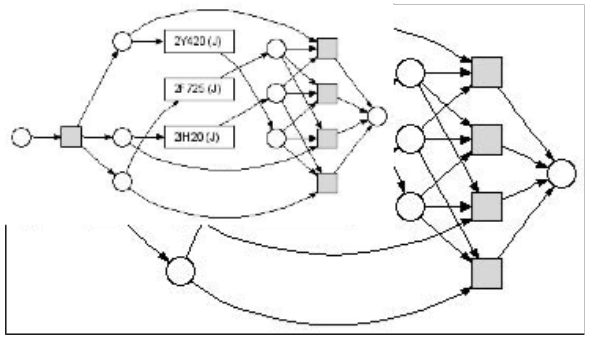
Isolate a set of standard curriculum patterns and based on this patterns

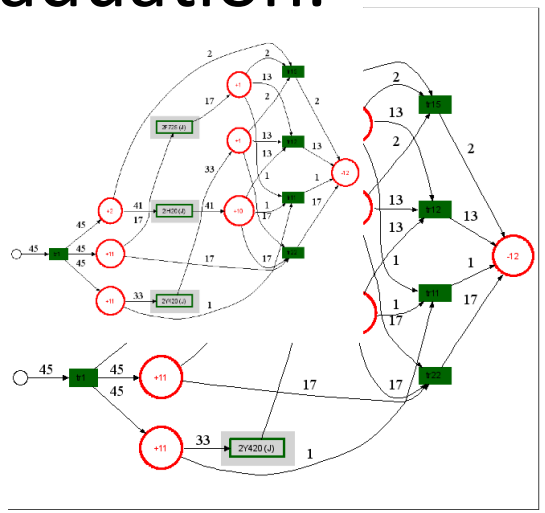
1. mine the curriculum as an executable quantified formal model and analyze it, or
2. (first) manually devise a formal model of the assumed curriculum and test it against the data.



# Example 2-out-of-3 Pattern Check

- At least 2 courses from { 2Y420,2F725,2IH20 } must be taken before graduation.

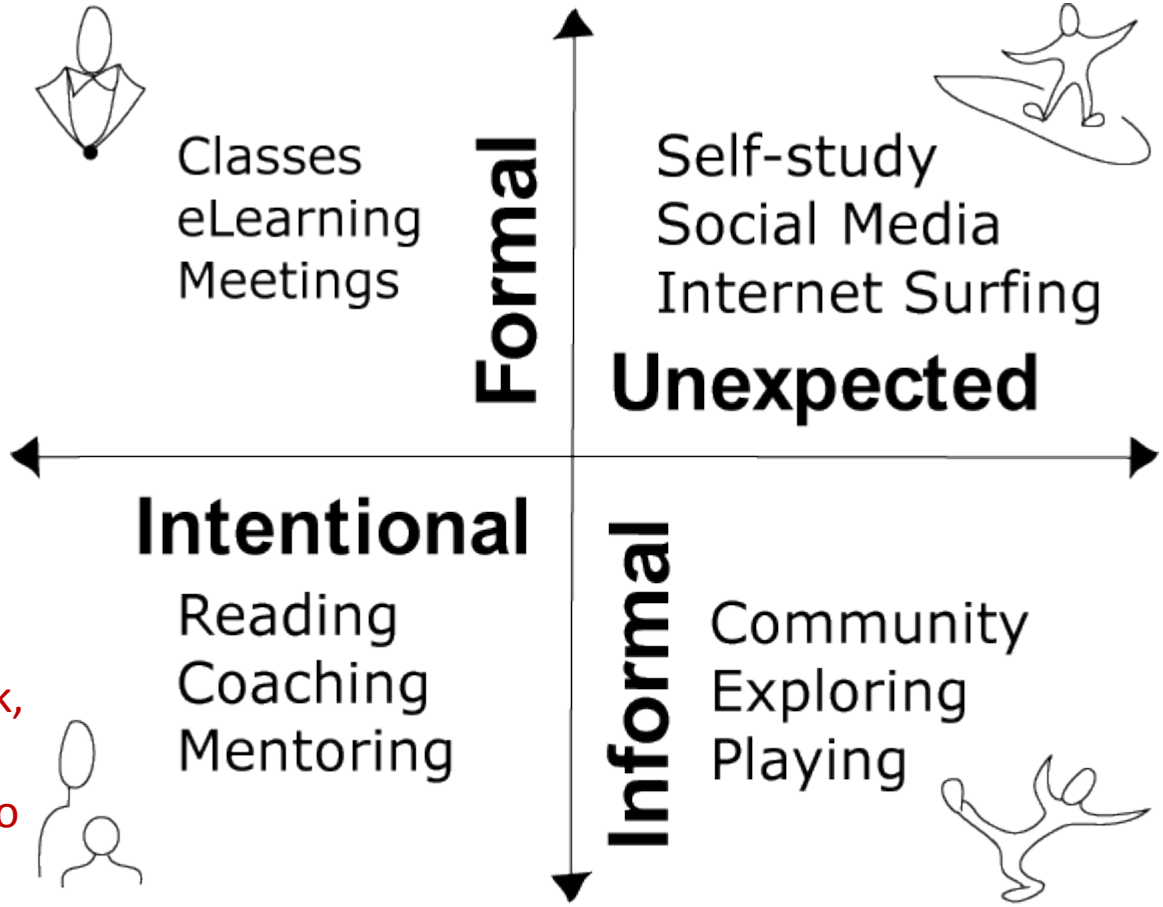


# Four Major Types of Learning & Types of Questions EDM Can Assist with

How to (re) organize the classes, or assessment, or placement of materials based on usage and performance data

How to help learners in (re-) finding useful material, done whether individually or collaboratively with peers




How to identify those who would benefit from provided feedback, study advice or other help; How to decide which kind of help would be most effective?

How to help learners in (re-) finding useful material, done whether individually or collaboratively with peers

# A Wide Scope of "learning"

- Traditional education at primary, secondary and high-school (algebra tutors), and University levels
- eLearning and Blended learning (LMS like Sakai, Moodle, Blackboard; SQL tutor)
- Professional education – (pilot, military simulators)
- Rehabilitation, elementary skills like reading and performing arithmetic operations ("Neure", "Ekapeli")
- eHealth, Patient education (Philips "Motiva", RPM)
- Learning becomes more informal, mobile, social and ICT/data/information intense in all areas
  - EDM is just making the first steps to address this

# EDM Evolution

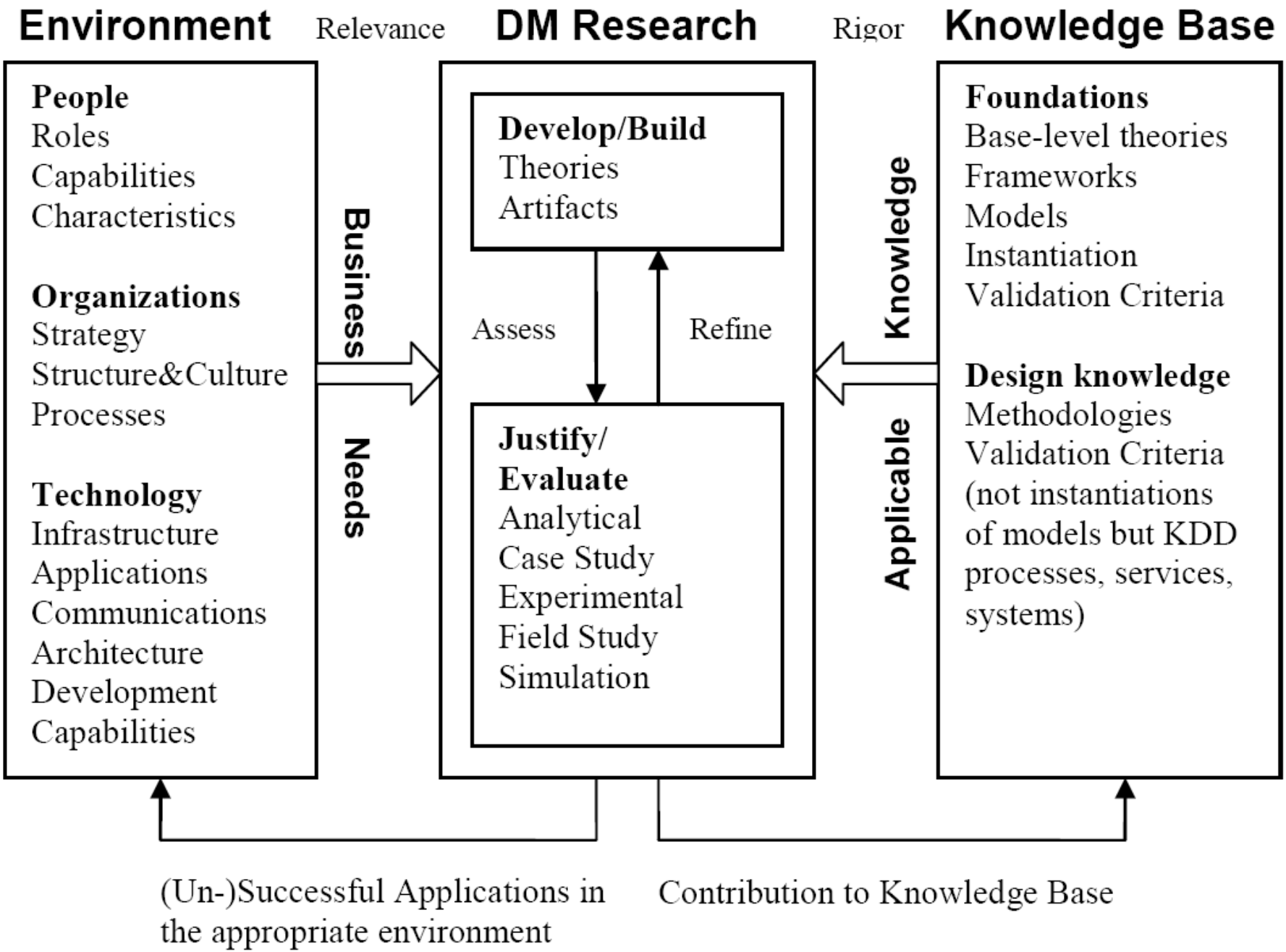
- Phase I: Educational research curiosity driven
  - Come up with a hypothesis, collect data, test and publish the results
- Phase II: DM research curiosity driven
  - data is there, collected because it can be collected
  - Do EDM, generate interesting hypothesis, test, and publish the results
- Phase III: Educational needs driven
  - Technology is there, knowledge is there,
  - Time to start the valorization - Carnegie Learning Lab Success
- Phase IV: Synergy of R&D
  - Understanding that not all know-how is there
  - Understanding (by researchers) that assessment in non-lab settings is crucial
  - Understanding (by stakeholders) that R&D forms a natural cycle

# From EDM's past and presence to EDM of tomorrow

- Lyrics by Jack Mostow, CMU
  - *All my data from my tutor,*
    *I put on a USB*
    *That I accidentally swallowed*
    *Now my study's history.*
  - *All my data,*
    *Educational data that I mine,*
    *Now is lost and gone forever --*
    *Dreadful sorry, data mine*
  - ... full text can be found at EDM 2011 website

Reflects on not easy early days of EDMers

- Indeed, this relates primarily to Phases I & II, while we need to think more about Phases III & IV
  - how to organize data collection to turn data instances into "first-class citizens" rather than some byproduct
  - how to address ethical, privacy and scrutability issues
  - how to enable and organize EDM activities at different levels, individual, institutional, national levels


Environment
Relevance
DM Research
Rigor
Knowledge Base
People
Roles
Capabilities
Characteristics
Organizations
Strategy
Structure&Culture
Processes
Technology
Infrastructure
Applications
Communications
Architecture
Development
Capabilities
Business
Needs
Develop/Build
Theories
Artifacts
Assess
Refine
Justify/
Evaluate
Analytical
Case Study
Experimental
Field Study
Simulation
Knowledge
Applicable
Foundations
Base-level theories
Frameworks
Models
Instantiation
Validation Criteria
Design knowledge
Methodologies
Validation Criteria
(not instantiations
of models but KDD
processes, services,
systems)
(Un-)Successful Applications in
the appropriate environment
Contribution to Knowledge Base

# Take Home (& for the discussion) Messages

- Many stakeholders: students, lecturers, tutors, study advisors, directors of education, educational institution and national level
- Lot's of potential benefits for each category of stakeholders
- Popular data mining problem formulations fit well to the educational domain and there are state-of-the-art techniques to address them
- Promising directions for further research from the applications perspective

# Still Open (and often not technology related) Questions

- Privacy and ethics
  - What is EDMer's phi
  - Is EDM always ethica
  - Is EDM a threat to p
    - Students, educato


- Who are EDM stakeholders?
- Organization of data collection processes
- Organization of EDM activities

# Acknowledgements


- Thanks to many colleagues at TU/e:
  - Wil van der Aalst
  - Nikola Trcka
  - Ekaterina Vasilyeva
  - Toon Calders
  - Paul De Bra
  - Jan Vleeshouwers
- SURFfoundation
  - John Doove
- …
- I wish I could acknowledge funding agencies here

# Thank you!

- *Questions*
- *Suggestions*
- *Collaboration*

- all warmly welcome