Data Analysis in Educational Systems

Sebastián Ventura

Department of Computer Sciences and Numerical Analysis
University of Córdoba

Outline

- Introduction
  - Motivation
  - Historical perspective
  - Educational Data Science
- Tasks in Educational Data Science
- Open Challenges
- Conclusions

Introduction

Introduction

- The development of educational systems (web applications, LMSs, MOOCs) has been rising exponentially in the recent years:
  - These systems produce information of high educational value, but usually so abundant that it is impossible to analyze manually.
  - Tools to automatically analyze this kind of data are needed.
- Educational institutions have information systems that store plenty of interesting information:
  - This available information can be used to improve Strategic Planning of these institutions.
  - In this case, tools to analyze that data automatically are needed too.

Introduction
First contributions: EDM

- First references about the automatic discovering of useful knowledge from educational data appeared in the early nineties.
- In the early 2000's several workshops about this topic were organized in conferences like ITS, UM or AIED. The term **Educational Data Mining** was coined then.

Educational Data Mining is a discipline concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings, and using those methods to better understand students, and the settings which they learn in.

- First conference on Educational Data Mining was celebrated in Montreal, 20-21th of June 2008.

Introduction

New Events. More terms for the same discipline?

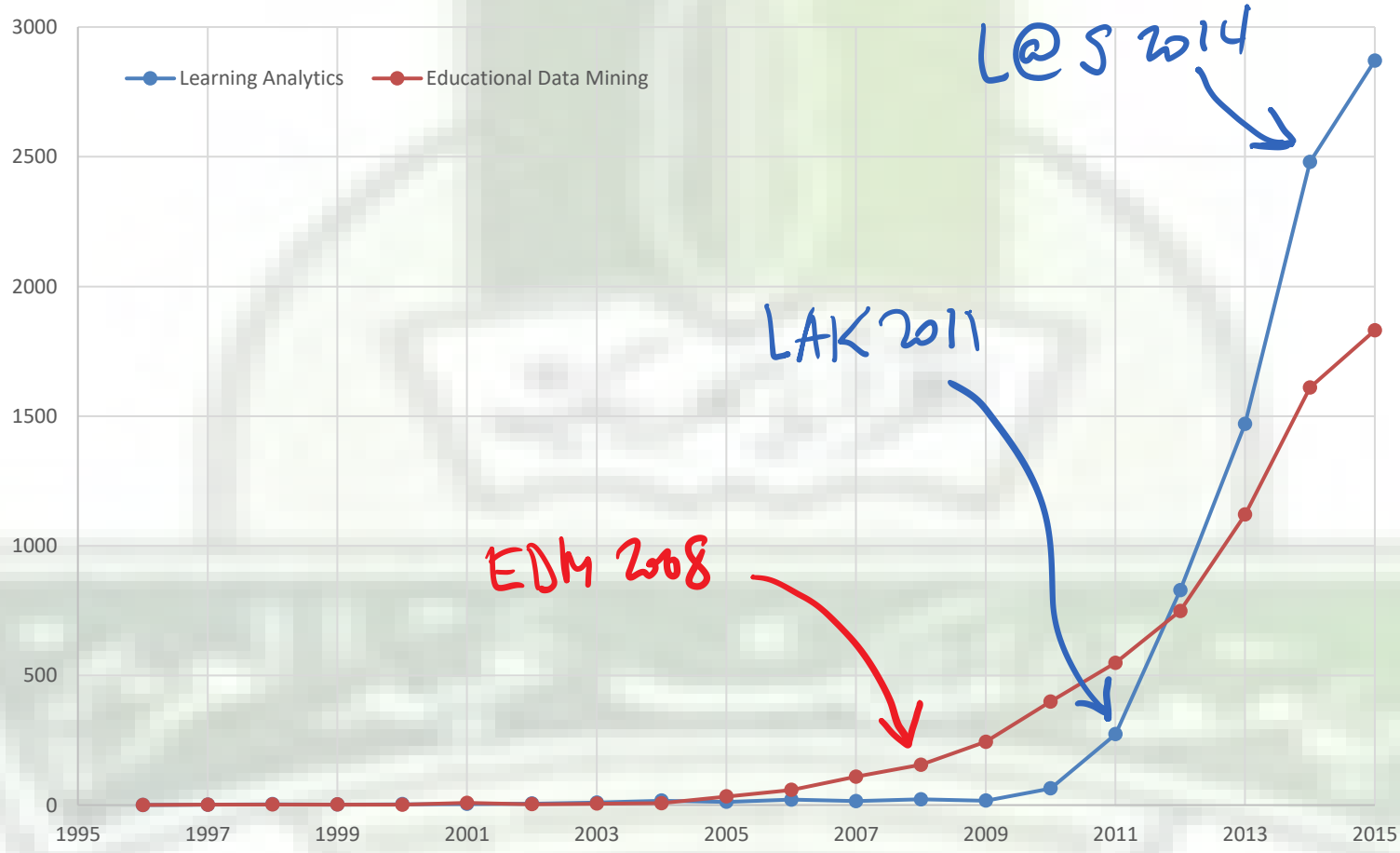
- The number of paper about EDM growth up exponentially.
- The International Educational Data Mining Society was founded in 2011.
- During the same year was celebrated the First International Conference on Learning Analytics and Knowledge (LAK 2011). Its organizers coined the term **Learning Analytics** as

Learning Analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs.

- International Society on Learning Analytics Research (SOLAR) was founded in 2013.
- LAK organizers claim that LA and EDM are different disciplines. What do you think?

Introduction

Scientific Production in EDM and LA



Introduction

More related terms…

- There is another discipline closely related to LAK and EDM: **Academic Analytics**

Academic Analytics is the process of evaluating and analyzing organizational data received from university systems for reporting and decision making reasons (Campbell, & Oblinger, 2007).

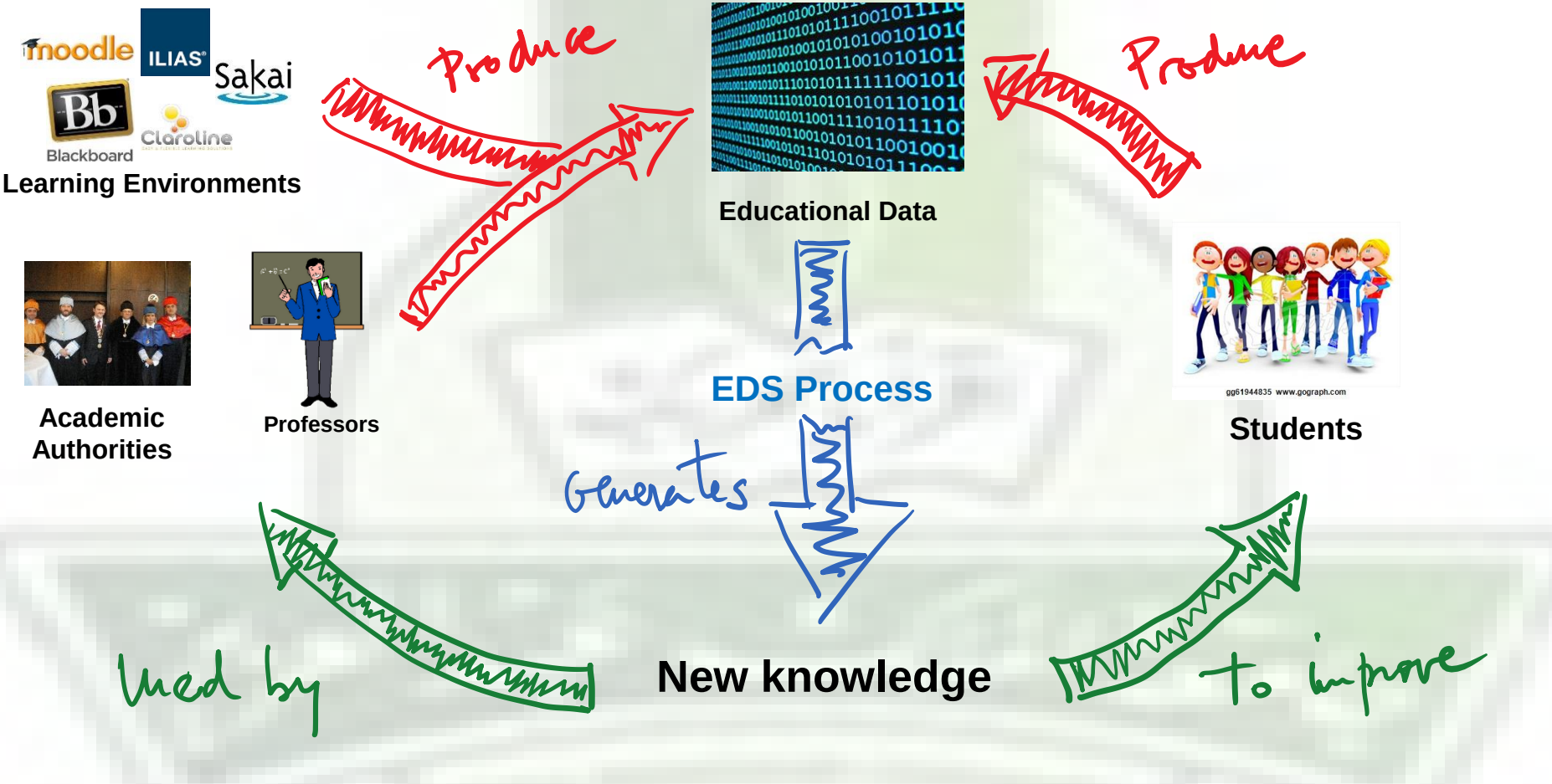
Introduction
Current Picture

- A new term, coined in 2013, is **Educational Data Science**

> **Educational Data Science (EDS)** can be defined as the generalizable extraction of knowledge from educational data.
>
> EDS is an emerging trans-disciplinary field which requires a combination of technical and social skills, an aptitude for engineering and also a profound understanding of the complex world of educational practices and learning in various environments (Piety et al., 2014).

- As can be seen, this definition includes EDM, LA and AA, which may be considered as different aspects of Educational Data Science.

Educational Data Science: Processes and Actors



"The Lifecycle of Educational Data Science"

Tasks in Educational Data Science

Tasks in Educational Data Science

- A **task,** in the EDS context, in a complete analysis or knowledge discovery process oriented to solve a question o problem in the Educational Field.
- The most common *steps* in this task usually are:
  1. Collecting the information to analyze.
  2. Preparing the information
  3. Applying one or more analysis / knowledge discovery algorithms
  4. Evaluating results and generating new useful knowledge
  5. Applying this actionable knowledge, in cases where this is possible

Tasks in EDS (II)
Examples of Tasks

- Predicting student performance
- Automatic recommendation of learning resources to students
- Modelling student behavior
- Automatic detection of abnormal student behavior
- Modelling peer-assessment and self-assesment
- Automatic generation of concept maps
- …

Predicting student performance

- Estimating the value of a variable that describes the student's future performance from available information.
  - Historical information (previous evaluations).
  - Other related information (environmental, social, etc.).
- It is a task of great interest, which has multiple uses.
  - Taking corrective actions to improve student achievement, especially when there is the possibility of school failure.
  - Detecting critical factors to improve student performance and / or to prevent its failure.

Predicting student performance (II)

- Has been solved using different methodologies:
  - *Classification*: The variable associated to student performance is categorical (for example “pass” or ”fail”)
  - *Regression*: The variable is numerical (numerical grade, number of failures, etc.)
  - *Nominal regression*: The variable is categorical, but the different labels (grades) follow an strict order, that is A > B > C > D > E.
- Open topics in this field:
  - A better evaluation of prediction models
  - Early prediction

Recommending resources or activities to students

- Generating new knowledge which can be used to make recommendations to students such as the next visit, task or problem to perform.
- This knowledge may also be used to tailor the content, interfaces and learning sequences to each individual student.
- It lets you customize certain aspects of the teaching-learning process
  - Very convenient in distance learning systems

Recommending resources or activities to students (II)

Content-based methods: Analyzing the available items to build a model that informs if a given resource is well suited for a student of group of them

- *Classification methods*. If there is a training set with labeled items.
  - Input: resource features
  - Output: recommended / not recommended
- *Association methods*. If we don't have a class label

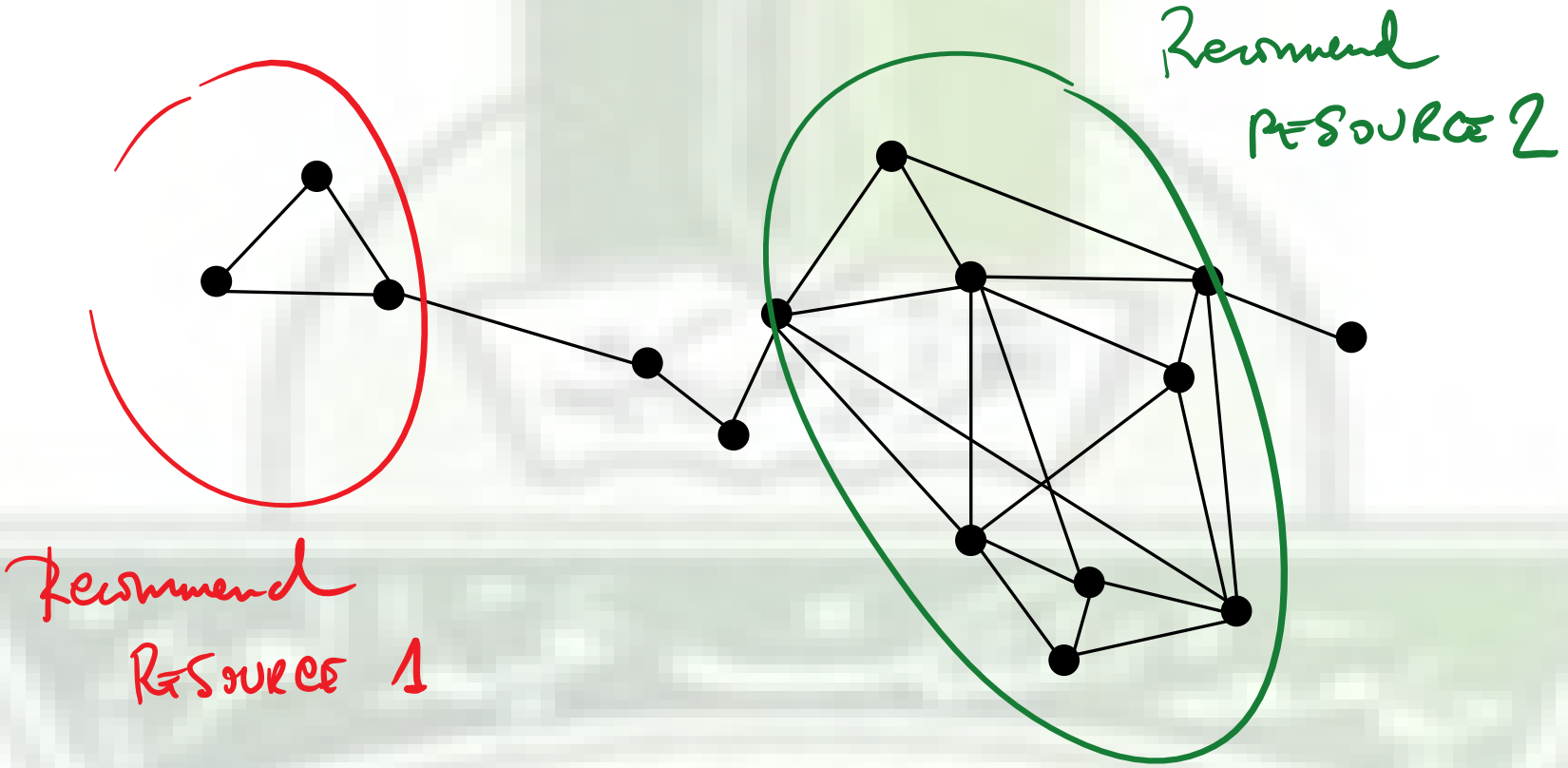
Both methods presents the *cold start* problem. "At the beginning we don't have enough information to build the model".

Recommending resources or activities to students (III)

Collaborative filtering: Recommend to a user the same resources that have worked well with other users similar to him.

- *Clustering methods*. Once we have obtained the groups or similar users, we can find what resources have been used by them and recommend to new users belonging to these clusters
- Recently has been applied the *analysis of social networks*. Instead of creating the clusters we recommend resources that have been successful to nearest neighbors in the social network.

Recommending resources or activities to students (IV)



Detection of undesired student behavior

Unwanted student behavior is a very broad concept, including:

- Performing wrong actions
- Misuse of facilities
- Attempts to cheat the system
- Other issues: detection of low motivation, school failure or student dropout.

Detection of undesired student behavior (II)

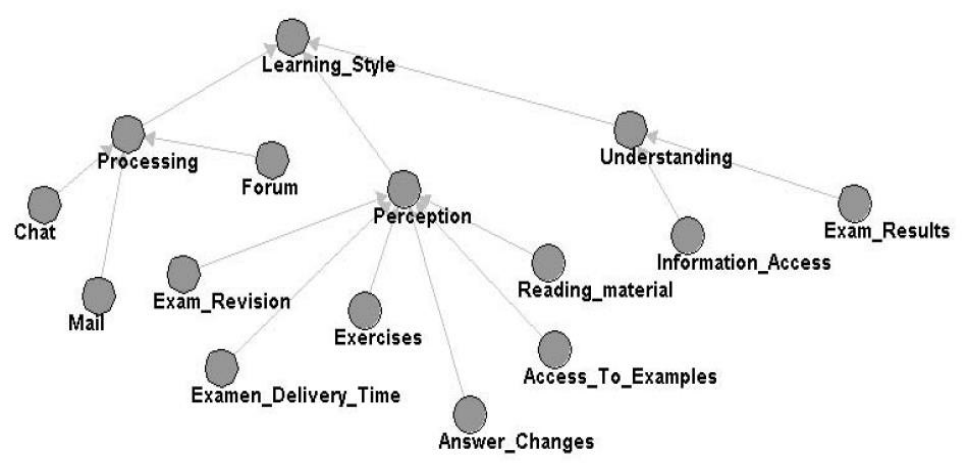
- *Classification*: Build a model that distinguish wanted and unwanted behavior.
- *Anomaly detection methods*: Apply clustering methods and detects data that cannot be included in any group.
- *Association rule mining and/or subgroup discovery*: Find rules that explain the anomalous behavior of a group of students.

Modelling student behavior

- Developing cognitive models of student users of an educational system, including a modeling skills and declarative knowledge.
- The interest of this work is manifold:
  - Allows the construction of intelligent tutoring systems using this model for teaching and custom-tailored to the characteristics student.
  - The information embodied may shed light on understanding the psychological mechanisms that influence learning.

Modelling student behavior (II)

- One of the most popular models to represent student behavior are bayesian networks



- Association rule mining has also been used to model student behavior in adaptive hypermedia systems

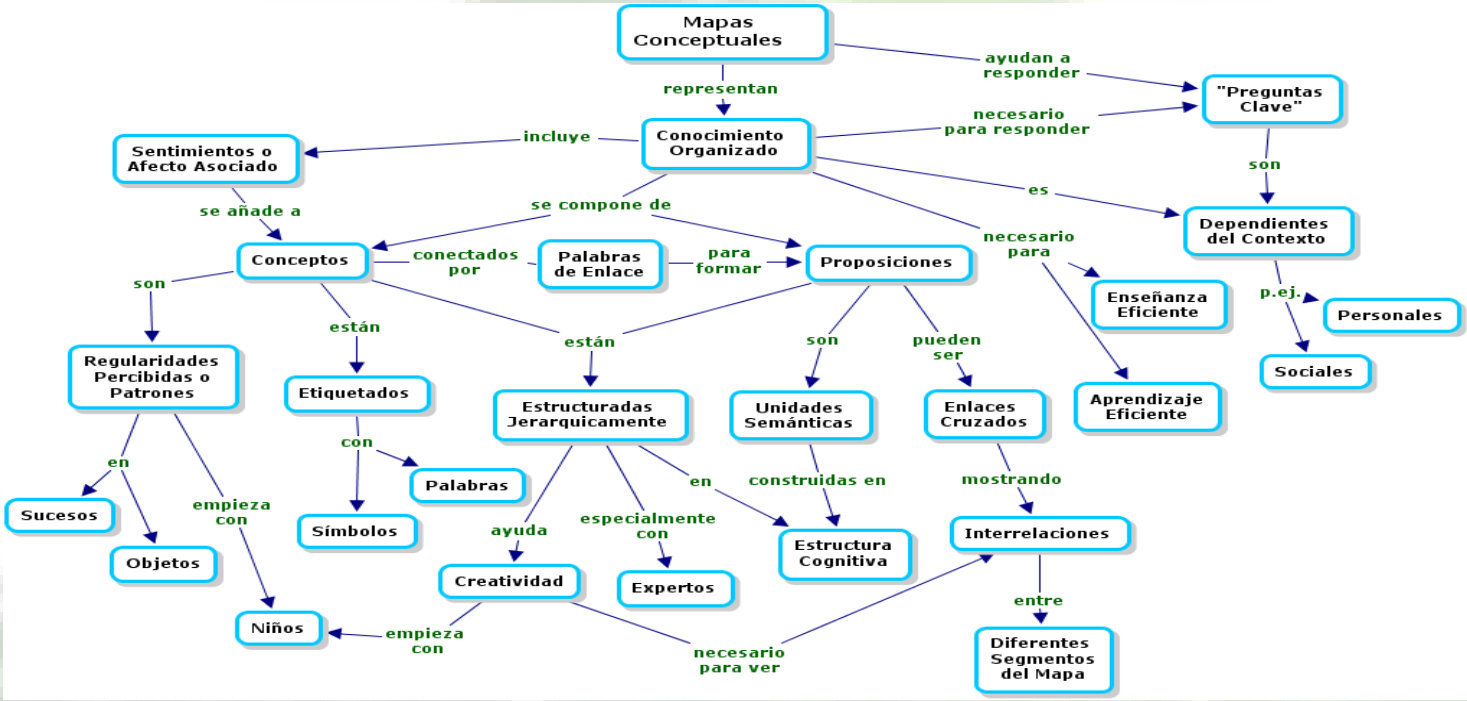
Modelling Self-Assessment and Peer-Assessment

- Self Assessment and Peer Assessment are two interesting evaluation techniques that have gain relevance with the appearance of MOOCs and other distance learning Systems.
- In Peer-Assessment, students evaluate the work of their peers.
- Each work is evaluated by several students and the final grade is an average of these assessments.
- Usually there is a bias between grades provided by students and the one provide by teachers.
- The problem consist on finding a good correction to convert this student average in a right grade.

Automatic generation of conceptual maps

- Conceptual maps are used to graphically represent concepts or ideas that have a hierarchical relationship.
- A conceptual map is a network in which the concepts are the nodes of the network, and there are a number of edges that serve to relate some concepts with others.
- Is a structured way to visualize the most relevant information on a topic.

Automatic generation of conceptual maps (II)



The development of concept maps can be very laborious, especially when we want to represent the domain is complicated.

Automatic generation of conceptual maps (III)

Two techniques have been used to generate conceptual maps automatically:

- Association mining: Rules mined represent relationships between concepts to include in the map.
- Text mining: These techniques have been used to extract the keywords that represent the concepts to include in the map

New Challenges

New Challenges

- Development of good tools for EDS
  - Personalized tasks
  - Post processing of models
  - Use by non-experts in Data Science
- Data Mining in MOOCs:
  - Big Number of students
  - Student Retention:
    - Dropout detection
    - Personalization
  - Self- and Peer-Assessment
- Evaluation from multiple perspectives
- Mining Institutional Data (Big Data Mining)
- …

Conclusions

Conclusions

- Educational Information hides knowledge useful to improve Learning and to get a better insight about it.
- Educational Data Science applies Data Analysis to perform this task.
- Since the nineties a lot of interesting applications have been described in this field
- There are still a lot of open problems
- A main problem. Availability of good quality educational data.

References

- C. Romero & S. Ventura. Educational Data Mining: A Survey from 1995 to 2005. *Expert Systems with Applications*, 33(**1**), 135-146, 2007.
- C. Romero & S. Ventura (eds.). *Data Mining in e-learning*. Advances in Management Information, Vol. 4. WIT Press. Wessex (UK), 2006.
- C. Romero, S. Ventura & E. García. Data Mining in Course Management Systems: MOODLE Case Study and Tutorial. *Computers and Education*, 51(**1**), 368-384, 2008.
- C. Romero & S. Ventura. Educational Data Mining: A Review of the State-of-the-Art. *IEEE Tansactions on Systems, Man and Cybernetics. Part C: Applications and Reviews*, 40(**6**), 601-618, 2010.
- C. Romero, S. Ventura, M. Pechenizkiy & R. S. de J. Baker (eds.). *Handbook of Educational Data Mining*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. CRC Press, 2010.
- C. Romero & S. Ventura: Data mining in education. *Wiley Interdisc. Rev.: Data Mining and Knowledge Discovery* 3(1): 12-27 (2013).
- C. Romero & S. Ventura: Data Science in MOOCs. *Wiley Interdisc. Rev.: Data Mining and Knowledge Discovery*. To appear (2016).


شكرا
The Mosque of Cordoba (169-633 AH)