Introduction to Data Minig in Education

Cristóbal Romero Morales

(cromero@uco.es)

Department of Computer Sciences and Numerical Analysis.
University of Córdoba

KDIS Lab
KNOWLEDGE DISCOVERY AND INTELLIGENT SYSTEMS

What do we call it?

- Statistics
- Machine Learning
- Data mining
- Knowledge Discovery in Data
- Big Data
- Data Analytics
- Data Science
- …?

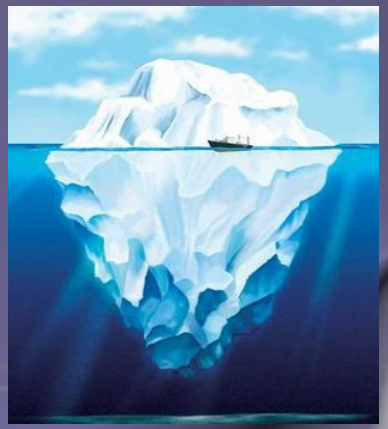
Same Core Idea:
Finding Useful Patterns in Data

Different Emphasis

“In god we trust, all others must bring Data”
William Edwards Deming (1900-1993)

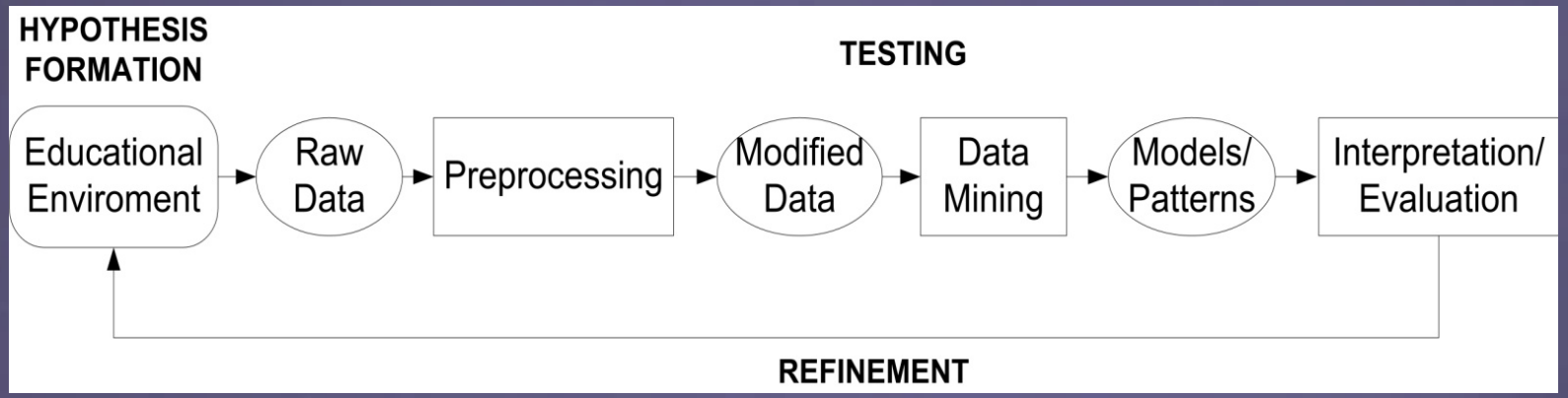
Introduction

- The development of web-based educational systems 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 it 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 also needed.



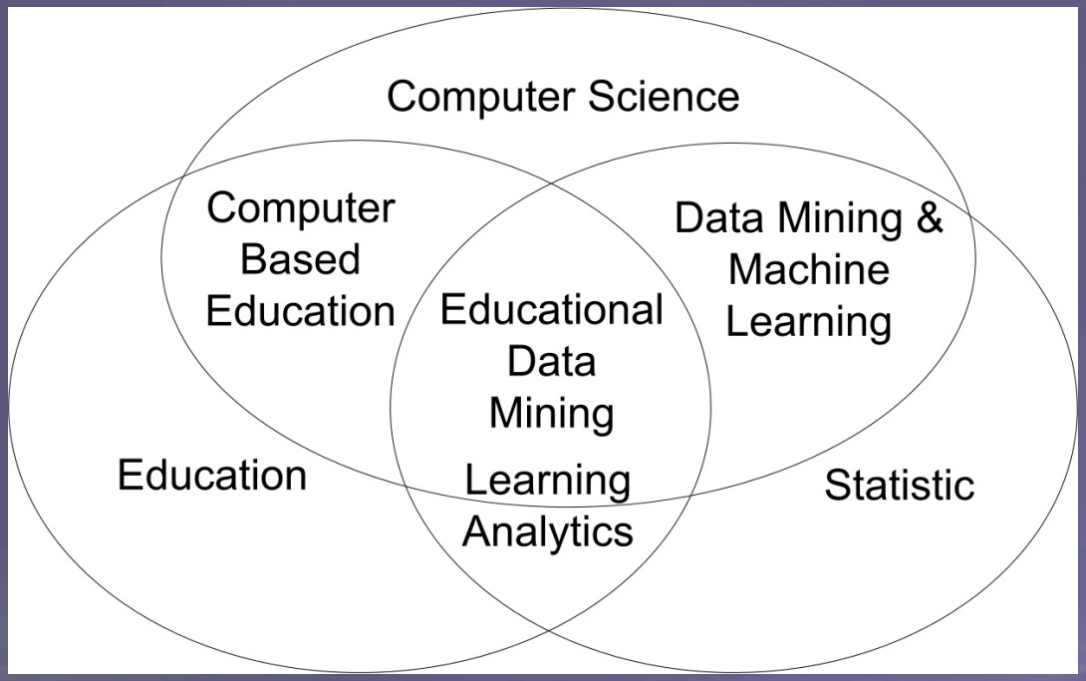
Introduction
What is EDM?

- ***Educational data mining (EDM)*** is the application of data mining techniques to educational environments.



Introduction
Multidisciplinary domain

- ***Educational data mining (EDM)*** is a multidisciplinary domain that is an intersection of 3 domains: computer science, education, statistics.



Introduction

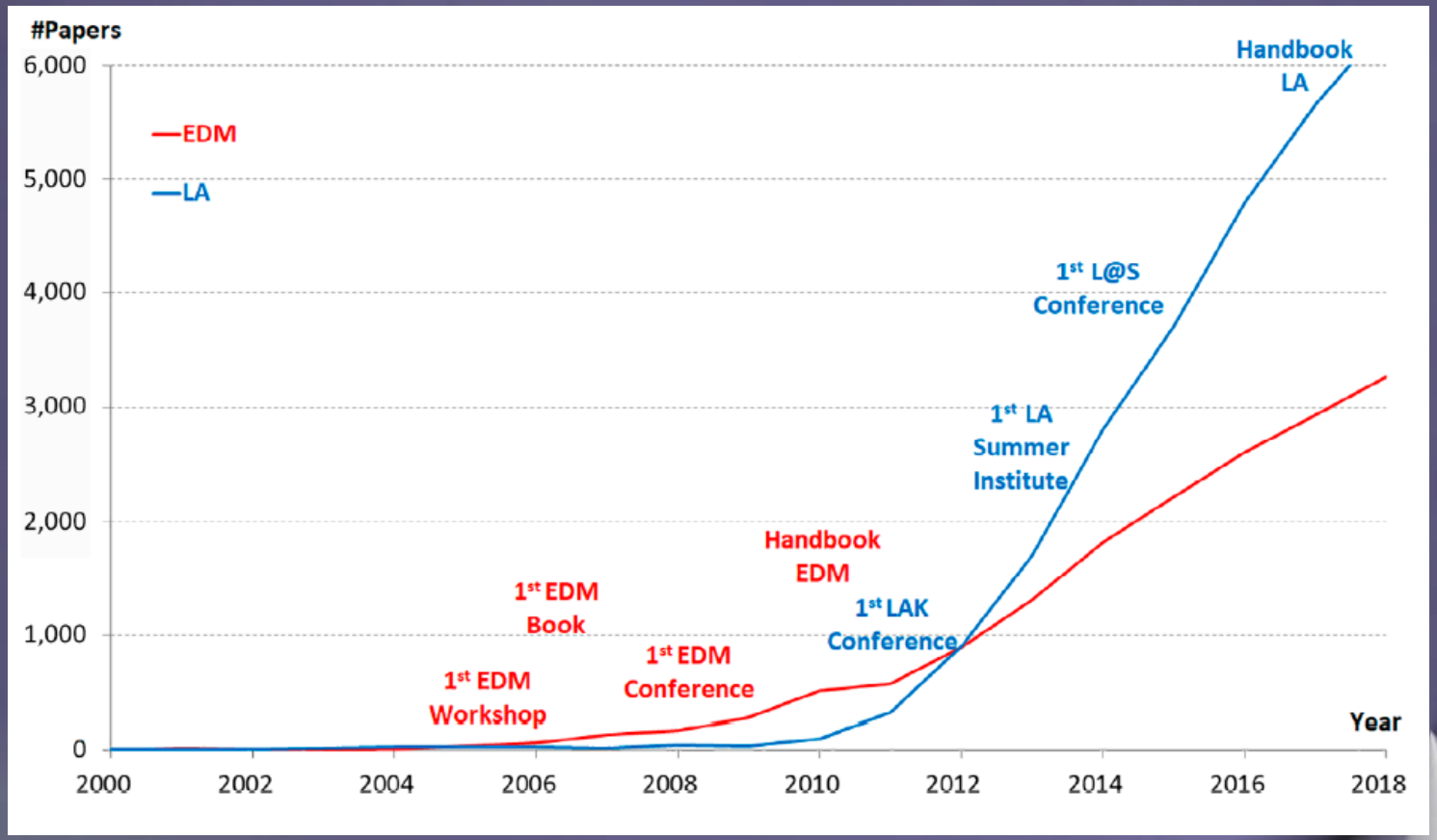
Other areas closely related to EDM

- **Learning analytics**
  - 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.
- **Academic analytics**
  - Bussiness intelligence applied to institutional academic data.
- **Teaching analytics**

Introduction

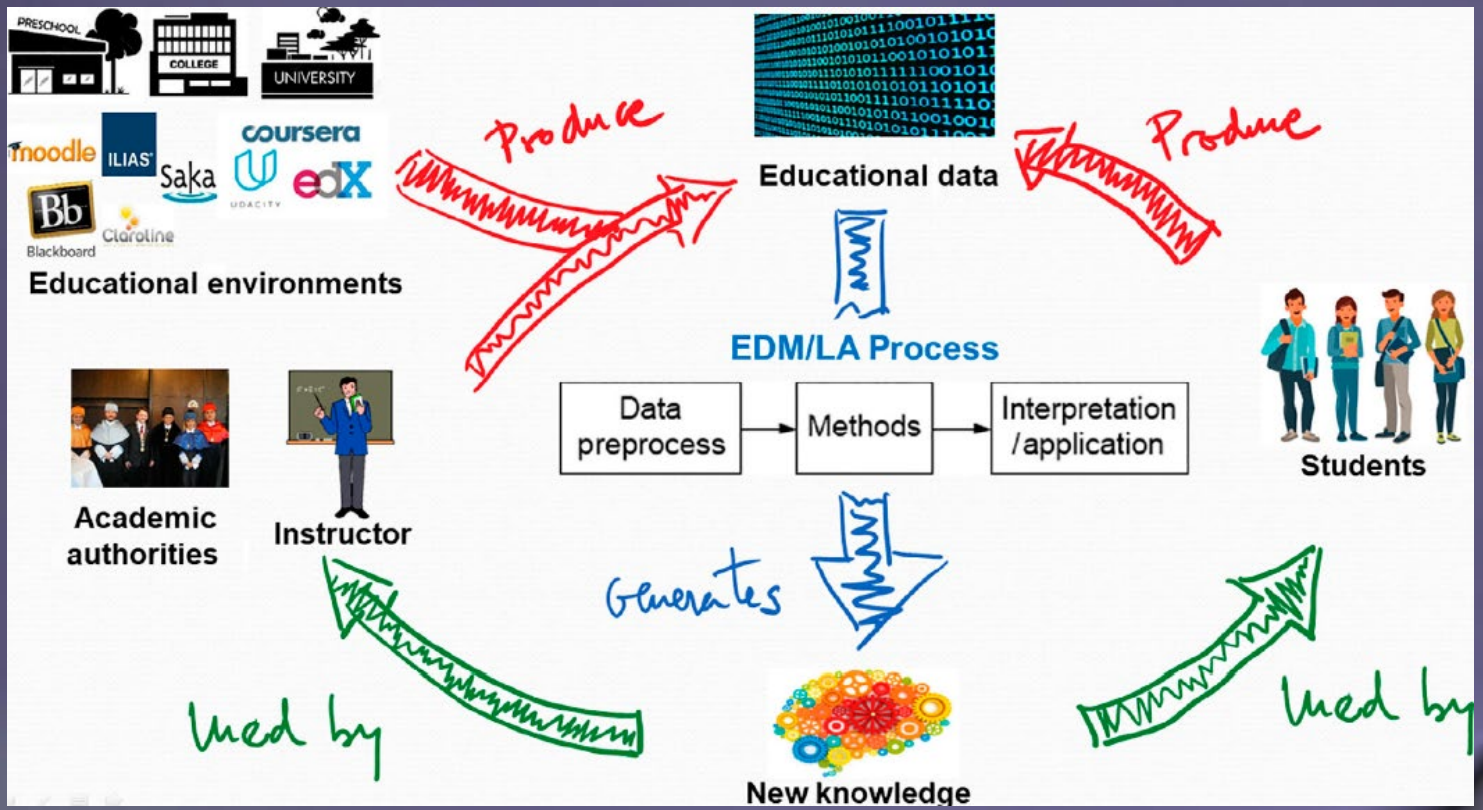
EDM and Learning Analytics progression

Evolution of EDM and LA references in Google Scholar



Process and actors

The Lifecycle of Educational Data Science:

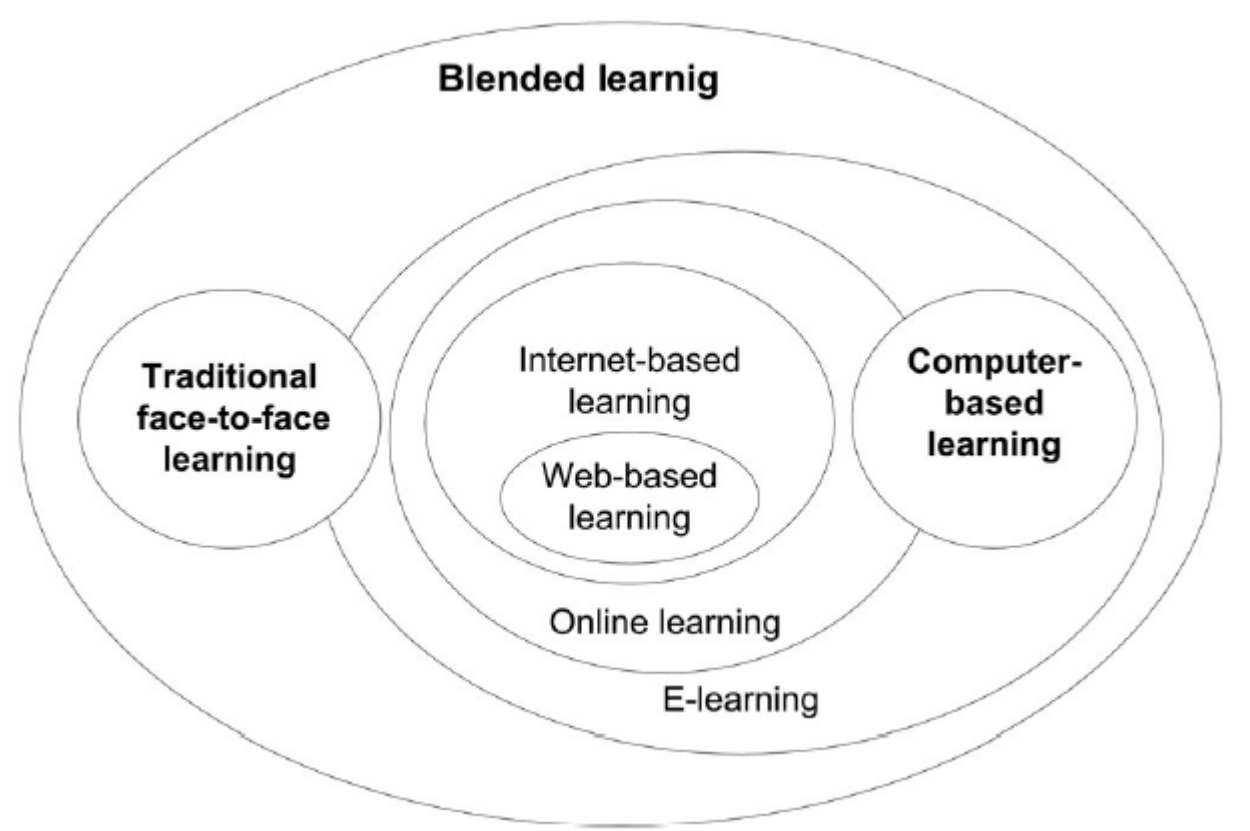


Goals

- The knowledge that can be discovered from educational data is very diverse.
- Our main objective when we applied EDS techniques depends on:
  - Who is addressed the knowledge we will extract
    - Students
    - Teachers
    - Academic authorities
  - What kind of information is available?
    - A priori
    - A posteriori
  - What is our environmment of interest?
    - Traditional learning
    - Distance learning
  - ...

Data

Types of Educational Environments



Data
Characteristics

- The information come from different sources of data.
- There are a lot of incomplete and loss data because not all students carry out all the activities.
- User/Students are clearly identified.
- There is a great number of available instances and attributes that may required tasks of filtering for selecting the most important.
- Educational data have different level of granularity.
- Some transformation such as discretization of number are normally used for improving the comprehensibility of data and the obtained models.

Tasks

- **Low level tasks**. Similar to ML/DM, but the knowledge we want to discover is extracted from educational data.

GENERAL (DM)

- **High level tasks**. Try to solve a problem in the educational context. Involves one or more low level tasks, as well as the interpretation of results.

SPECIFIC (EDM)

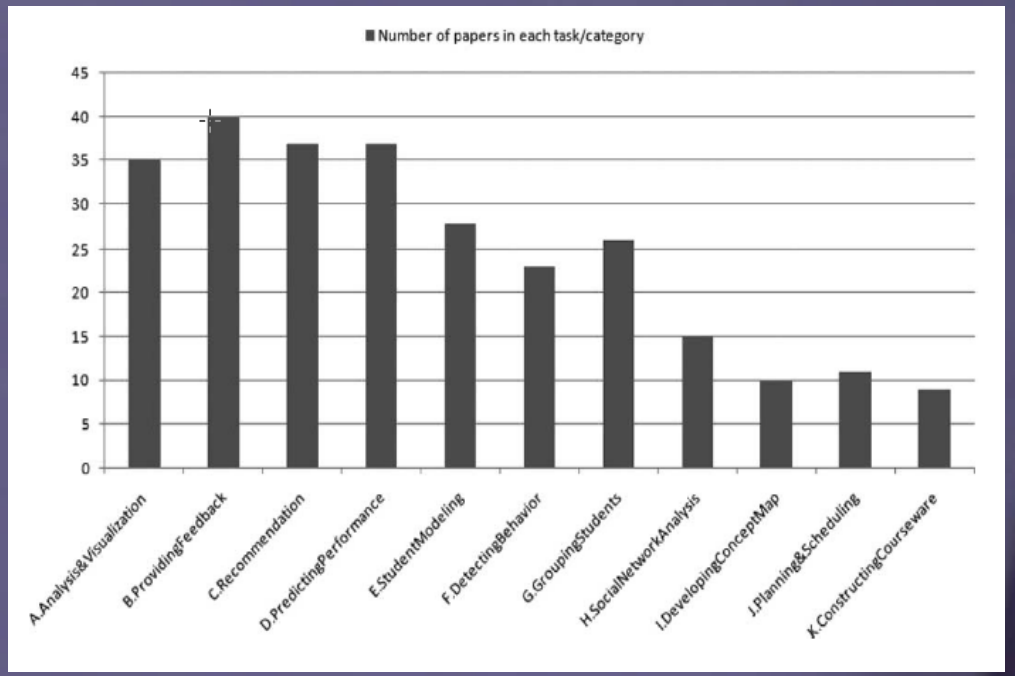
Tasks
Low level (ML/DM) tasks

- **Predictive tasks**
  - Supervised. Output information is available
  - Examples
    - Classification
    - Regression
- **Descriptive tasks**
  - Unsupervised. Output information is **not** available
  - Examples
    - Association
    - Clustering

Tasks

High level tasks

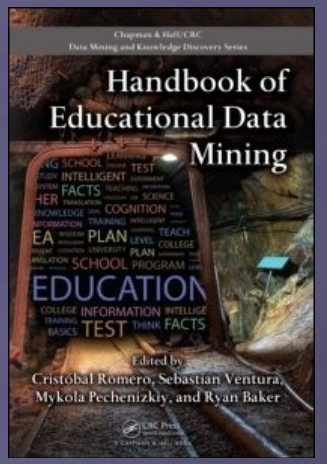
- Analysis & Visiualization.
- Providing Feedback.
- Recomendation.
- Predicting Performance.
- Student Modeling.
- Detecting Behavior.
- Grouping Students.
- Social Network Analysis.
- Developing Concep Map.
- Planing & Scheduling.
- Constructing Courseware.



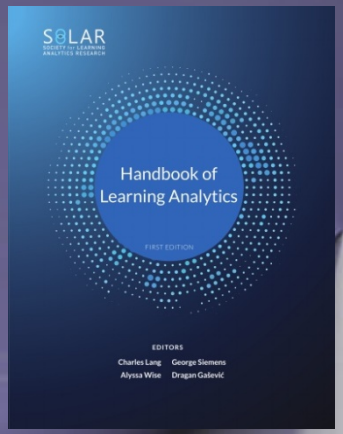
Publications

Books

- *Data Mining in E-Learning*.
  C. Romero & S. Ventura (Eds).
  Editorial WIT Press, 2006.



- Handbook of Educational Data Mining.
  C.Romero, S. Ventura,
  M. Pechenizky, R. Baker. (Eds).
  Editorial CRC Press, Taylor & Francis Group. 2010.



- *Handbook of Learning Analytics*.
  C. Lang, G. Siemens, A. Wise, D. Gašević
  SOLAR, 2017.

Publications

- JOURNALS:
  - Journal of Educational Data Mining
  - Journal of Learning Analytics
  - Journal of Artificial Intelligence in Education
- CONFERENCES:
  - *International Conference on Educational Data Mining (EDM)*
  - *Learning Analytics & Knowledge (LA)*
  - *International Conference on Artificial Intelligence in Education (AIED)*
  - ACM Conference on Learning at Scale (l@s)

Publications
Surveys/Reviews

- Baker, R., Yacef, K. The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Educational Data Mining, 1, 1, 3-17. 2009.
- C. Romero, S. Ventura. Educational Data Mining: A Review of the State-of-the-Art. IEEE Transactions on Systems, Man, and Cybernetics--Part C: Applications and Reviews. 40:6, pp. 601 – 618. 2010.
- Karen Cator. Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics. Report of the U.S. Office of Educational Technology. 2012.
- Romero, C., & Ventura, S. Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1355. 2020.

DM Software

Weka is one of the most popular software packages for Data Mining

http://www.cs.waikato.ac.nz/~ml/weka/

This is a very popular DM tool, developed in Java

http://rapidminer.com

R is a programming language that was initially created to perform statistics, but it has also used in DM

https://www.r-project.org/

Free EDM datasets

Datasets	URL	Descripction
Canvas Network dataset	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/1XORAL	De-identified data from Canvas Network open courses (running January 2014 - September 2015), along with related documentation.
DataShop	https://pslcdatashop.web.cmu.edu/index.jsp?datasets=public	LearnSphere's DataShop provides a central repository to secure and store research ITS data and set of analysis and reporting tools.
HarvardX-MITx dataset	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26147	De-identified data from the first year of MITx and HarvardX MOOC courses on the edX platform along with related documentation.
MOOC-Ed Dataset	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ZZH3UB	Communications taking place between learners in two offerings of the Massively Open Online Course for Educators (MOOC-Eds).
Open University Learning Analytics Dataset	https://analyse.kmi.open.ac.uk/open_dataset	It contains data about courses, students and their interactions with Moodle for seven selected courses.
Student Performance Dataset	https://archive.ics.uci.edu/ml/datasets/Student+Performance	This data approach student achievement in secondary education of two Portuguese schools.

Thanks.

Questions?