Educational Data Mining and DataShop

Ryan S.J.d. Baker
Kenneth R. Koedinger

EDM definition

- "the area of scientific inquiry centered around the development of methods for making discoveries within the unique kinds of data that come from educational settings, and using those methods to better understand students and the settings which they learn in." (Baker, in press)

EDM methods

- Often differ from regular DM methods
  - Attempt to exploit multiple levels of meaningful hierarchy
    - Keystroke, answer, session, student, classroom, school
  - Need to account for specific types of non-independence in data

(Some) Classes of method

- Prediction
- Clustering
- Relationship Mining
- Discovery with Models
- Distillation of Data For Human Judgment

Particularly prominent in EDM (as compared to DM in general)

- Prediction
- Clustering
- Relationship Mining
- **Discovery with Models**
- **Distillation of Data For Human Judgment**

- As in other areas of computational science

Prediction

- Develop a model which can infer a single aspect of the data (predicted variable) from some combination of other aspects of the data (predictor variables)
- Classification, regression, knowledge tracing

- Very prominent in EDM
  - EDM2009 best paper
    - Ritter et al
  - EDM2009 best student paper
    - Hershkovitz et al

Clustering

- Find points that naturally group together, splitting full data set into set of clusters

- Related: finding the factors that split the space of data features
  - Example: Principle Component Analysis

- Moderately prominent in EDM

Relationship Mining

- Discover relationships between variables in a data set with many variables
  - Association rule mining
  - Correlation mining
  - Sequential pattern mining
  - Causal data mining

- Prominent in EDM

Discovery with Models

- Pre-existing model (developed with EDM prediction methods… or clustering… or knowledge engineering…)
- Applied to data and used as a component in another analysis
- Moderately prominent in EDM
  - AIED2009 best paper nominee
    - Baker et al

Distillation of Data for Human Judgment

- Making complex data understandable by humans to leverage their judgment
- Moderately prominent in EDM

EDM track schedule (changed)

- Tuesday 10am
  - Bayesian Knowledge Tracing and Discovery with Models
- Tuesday 11am
  - Item Response Theory and Learning Factor Analysis
  - Ken Koedinger
- Tuesday 2:15pm
  - Prediction: Classifiers and Regressors
- Wednesday 11am
  - Principal Component Analysis
  - Geoff Gordon

Where does the data come from?

Educational Software

- Computer tutors and other educational software
  - Fine grained, longitudinal, often across contexts
- Instrumented to give logs of every "transaction" between student and software
  - Transaction = entering an answer, requesting help – some semantic "action"
- Sometimes also instrumented to give logs of mouse movements and keyboard actions (e.g. de Vicente & Pain, 2008)
- Sometimes even physical sensors (e.g. Arroyo et al, 2009)

Other data sources

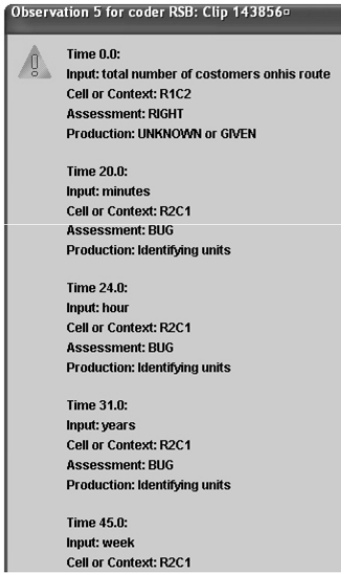
- Other data sources
  - Records of online courses (e.g. WebCT, Moodle) (e.g. Romero et al, 2008; Riley et al, 2009)
  - District or university-level student records
    - Example: www.icpsr.umich.edu/IAED
  - Group collaboration data (e.g. Kay et al, 2008)
  - Records of library borrowing
    - Nice paper on this at EDM2009

Often annotated…

- Quantitative Field Observations

Often annotated…

- Text Replay Labelings



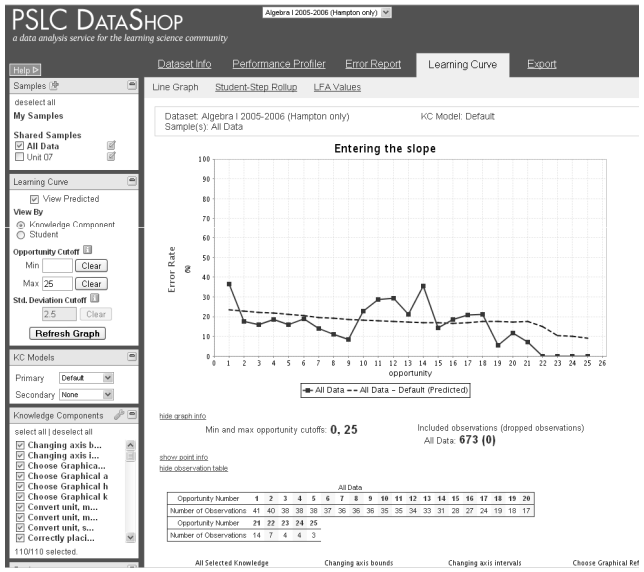
One excellent source of data

- Led to
  - 7 of 17 full papers at EDM2008
  - 8 of 20 full papers at EDM2009
- Used in exploratory data analysis as well as EDM

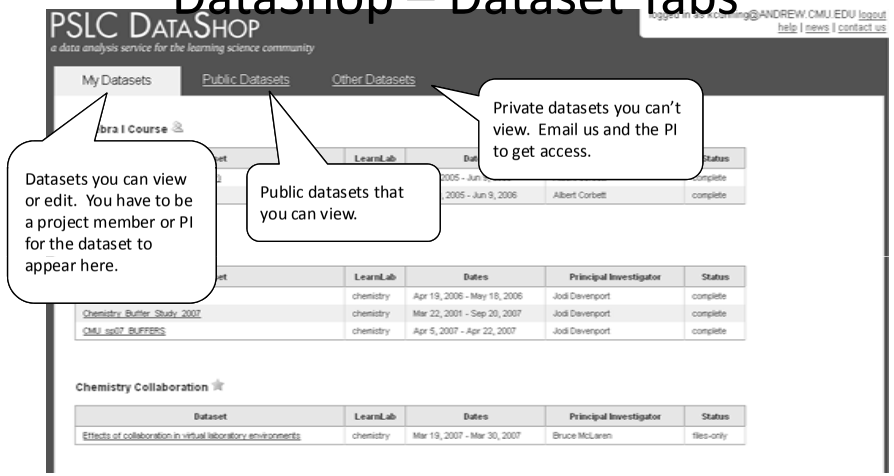
PSLC DataShop

- Free access, free to use these data sets in any analysis or publication you want
- Mostly collected in real classrooms by students using educational software for weeks or months or whole years
  - Intelligent tutors, virtual labs, online courseware, optimized drill, …
  - Math, physics, chemistry, Chinese, ESL, French, domain-general invention & meta-cognition skills …

https://pslcdatashop.web.cmu.edu/index.jsp



DataShop – Dataset Tabs

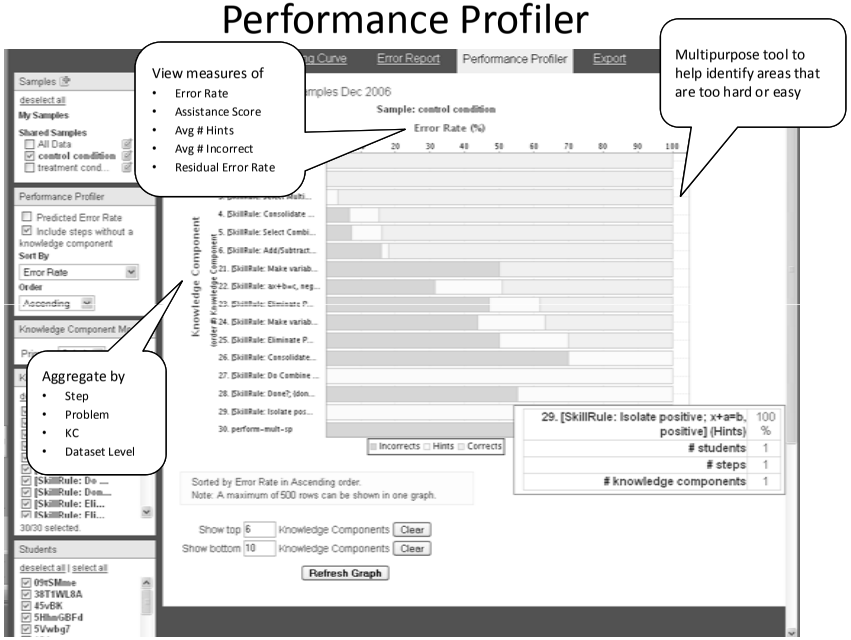


Dataset Info

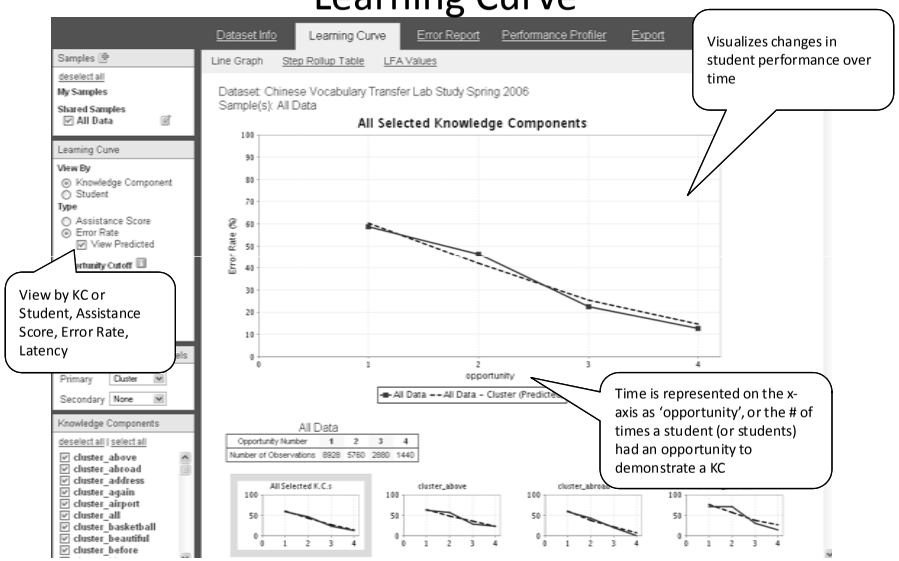
Wide range of Exploratory Data Analysis functionality

- A few quick examples

Performance Profiler

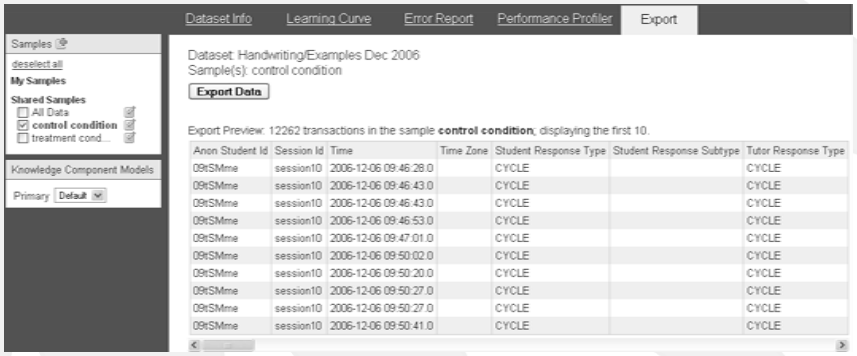


Learning Curve



Export

- Two types of export available
  - By Transaction
  - By Step
- Anonymized, tab-delimited file
- Easy to import into Excel, Weka, R
- Very fast even for large data sets



Plus

- We also host non-PSLC data
  - You have control over who gets access to your data
  - Eliminates storage and access-control problems
  - Makes it easy for EDM researchers looking for data to find your data and use it (and write papers with you or cite your papers)
- If you're interested in storing your data with us, we can chat about this while you're here