DataShop is an open data repository and set of associated visualization and analysis tools for fine-grained, longitudinal learner data.

DataShop enables you to:
- Import and export datasets
- Visualize student performance
- Perform learning curve analysis
- Discover cognitive models of student knowledge
- Use data to improve instruction
- Programatically retrieve data via web services

What data is in DataShop?
DataShop contains vast amounts of free data.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Datasets</th>
<th>Student Actions</th>
<th>Students</th>
<th>StudentHours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>60</td>
<td>2,838,843</td>
<td>3,263</td>
<td>5,661</td>
</tr>
<tr>
<td>Math</td>
<td>73</td>
<td>32,201,021</td>
<td>8,887</td>
<td>99,084</td>
</tr>
<tr>
<td>Science</td>
<td>30</td>
<td>3,675,770</td>
<td>3,890</td>
<td>19,930</td>
</tr>
<tr>
<td>Other</td>
<td>30</td>
<td>4,789,816</td>
<td>11,058</td>
<td>28,698</td>
</tr>
<tr>
<td>Total</td>
<td>193</td>
<td>43,950,460</td>
<td>27,098</td>
<td>153,273</td>
</tr>
</tbody>
</table>

Over 150,000 student hours!

What are some analyses that have used DataShop data?
- Adapting to when students game an intelligent tutoring system (Baker et. al.)
- Self-explanatory in the classroom: Learning curve evidence (Hausmann and VanLehn)
- Optimizing knowledge component learning of Chinese using a dynamic structural model of practice (Pavlik et. al.)
- Extending learning factors analysis to model reading transfer (Leszczenski and Beck)
- Exploring alternative methods for error attribution in learning curves analyses (Nwaigwe et. al.)

Current features

- Performance Profiler tool for exploring the data
- Knowledge component model analysis with learning curves
- Learning curve point decomposition
- Easy knowledge component model creation

New features

- Web services architecture

Coming in future releases

- Easy-to-use API for using web services to add fields to an existing dataset
- Ability to upload models to DataShop for dissemination, making it easier to build on each other’s work
- Improved scalability and speed: accommodate a planned 100x increase in quantity of data in DataShop
- Support for timestamps with millisecond precision
- Error bars on learning curves

KDD Cup 2010
Educational Data Mining Challenge
› http://pslcdatashop.org/KDDCup

Awarded to the PSLC and DataShop, this year’s challenge asked participants to predict student performance on mathematical problems from logs of student interaction with Intelligent Tutoring Systems.

The KDD Cup is the premier data mining competition that attracts top data mining and machine learning experts from industry and academia.

The competition ended on June 8. There were:
- 655 registered participants
- 130 participants who submitted predictions
- 3,400 submissions

The datasets used for the challenge were:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Students</th>
<th>Steps</th>
<th>File size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algebra I 2008-2009</td>
<td>3,310</td>
<td>9,426,966</td>
<td>3 GB</td>
</tr>
<tr>
<td>Bridge to Algebra 2008-2009</td>
<td>6,043</td>
<td>20,768,884</td>
<td>5.43 GB</td>
</tr>
</tbody>
</table>

The competition addressed questions of both scientific and practical importance. Improved models could be saving millions of hours of students’ time and effort in learning algebra. These models should both increase achievement levels and reduce time needed to learn.