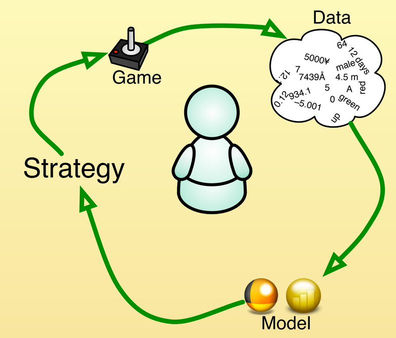
Research on Student Understanding of Data Organization

How do they work with complex and unstructured data?

Fall 2009–Fall 2012

The Core Idea



Students playing computer games generate large quantities of rich, interesting, highly variable data that mostly evaporates into the ether when the game ends. What if in a classroom setting, data from games students played remained accessible to them for analysis? In software and curriculum materials developed by this project, data generated by students playing computer games form the raw material for mathematics classroom activities. Students play a short computer game, analyze the game data, conjecture improved strategies, and test their strategies in another round of the game.

The Essence of a Data Game

The play of a data game consists of brief, suspenseful moves that generate data. The move parameters controlled by a player appear in the data. Examination of the data reveals patterns that are useful in improving a player's game score. At higher levels, the game requires increased sophistication in data modeling and automation of play.

Project Goals

- Improve students' ability to understand and work with data, with special emphasis on large data sets, data visualization, time series, informal inference, and data structures beyond rows and columns.
- Enrich students' understanding of mathematics through learning experiences based on using data generated by playing computer games.
- Expand research in students' understanding of data and chance and in students' ability to learn mathematical content when it is embedded in data-rich contexts.

Foundations—Prior Work

Fathom—Dynamic data analysis software for secondary and college math and science classrooms.

TinkerPlots—Dynamic data exploration software for middle school mathematics classrooms.

TinkerPlots and Fathom have made powerful data visualization tools available to students throughout the US and in a growing number of other countries. Both development teams have been increasingly focused on the need to get data into the hands of students more quickly, for example through scraping data from a web page whose URL is dropped into a document. TinkerPlots is soon to release a new version that includes simulation capabilities.

Published by Key Curriculum Press

Key Staff Members

William Finzer, co-PI
bfinzer@kcptech.com
Vishakha Parvate, research and project manager
Kirk Swenson, software engineer
KCP Technologies
Emeryville, CA

Cliff Konold, co-PI
konold@srri.umass.edu
Craig Miller, software engineer
UMass Amherst
Amherst, MA

Tim Erickson, game and activity developer, tim@eeps.com
Big Time Science
Oakland, CA

Research Objectives

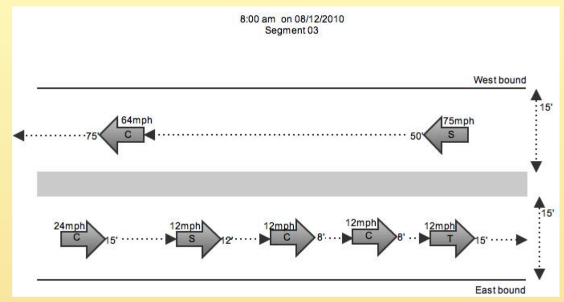
As part of the Data Games project, we are researching how students record and organize multivariate data. This research is informing the design of new software interfaces for Fathom and TinkerPlots that will allow students to explore and understand data that live in other than "flat" data structures — the structures that most software tools currently limit themselves to.

We have designed the Traffic Problem to explore the following questions:

1. What methods do novices and experts use to sytematically record data with multiple attributes?

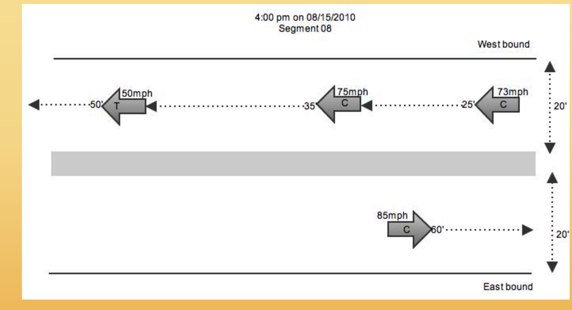
2. In recording data, do students employ a recognizable notion of "case?"

Task and Materials



City planners are studying the traffic along roads that lead into and out of the city. As part of the study, they are collecting data at various times of the day along short road segments. Here are two such road snapshots.

The snapshots include all of the information that the planners want collected: the time and date .... So on the 8:00 a.m. snapshot at the top left is a car (C) going 64 mph following 75 feet behind the vehicle in front. Following behind it is an SUV.



Research Protocol Summary

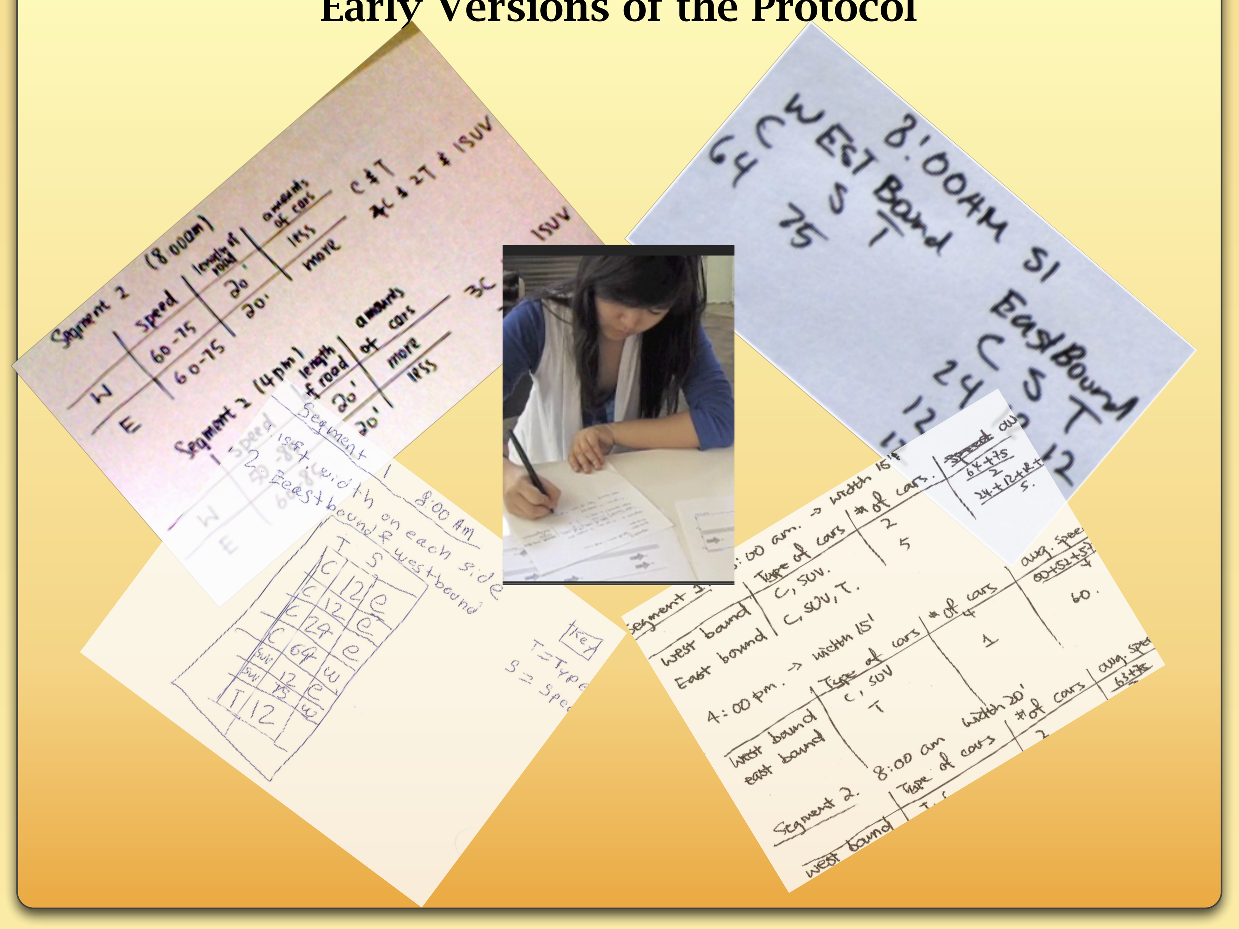
Actual protocol is available to browse on the table below

"On a blank sheet of paper, record the data from these two snapshots. Your data sheet should not be a drawing of these snapshots. It should be an organized record of the data values on the snapshots. ... Later, we may give your completed data sheet to other students who will use it to answer several questions about the flow of traffic along this road. These students will not see the snapshots you saw. Therefore, your data sheet needs to include all of the information these student will need to answer the questions we give them."

Revision Cycles

In the process of refining the Traffic Problem, we have given various versions of it to 31 novices and 6 experts. Most of the novices were high school students.

Student Work from Early Versions of the Protocol



Preliminary Analyses

Our initial assumption was that novices would not record and organize data in a way that would allow them later to answer questions about relations among attributes. Surprisingly, however, the majority of them created forms that housed all the information in the snap shots and organized the data in a way that maintained all the important relations. Specifically:

- About half organized data in "flat" tables.
- A majority used an organizational method that kept the information about individual vehicles together in such a way that it is possible to determine, e.g., the correlation between distance and speed.
- Many employed an organizational method consistent with a hierarchical data model, in that they partitioned information spatially to reflect different case levels (date/time, lane direction, vehicle information). This general method is consistent with our tentative design of "hierarchical" tables for hosting data.

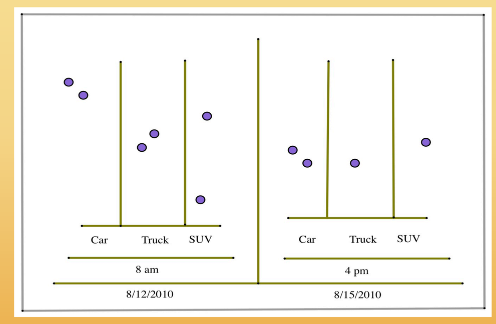
We are developing coding schemes that attempt to capture the building blocks of these various data organization methods.

Software Implications

Date	Time	Lane Information: Direction	Width	Vehicle Information: Type	Speed	Distance
8/12/2010	8am	West	15	Car	64	75
				SUV	75	50
		East	15	Car	24	15
				SUV	12	12
				Car	12	8
				Car	12	8
				Truck	12	15
8/15/2010	4pm	West	20	Truck	50	50
				Car	75	35
				Car	73	25
		East	20	Car	85	60

A prototype of nested data tables. The nesting of attributes within levels addresses the desire of many novices to record data efficiently so as to avoid the seemingly needless repetition of, e.g., date and time for every vehicle. But it also allows analysis at different levels of case without actively restructuring the unit of analysis.

Partitioned plot spaces is an early graphic interface designed to provide a visual way to represent data at different levels by creating partitions of graph space named by attribute values rather than by attribute names.



Acknowledgments

This material is based upon work supported by the National Science Foundation under:
KCP Technologies Award ID: 0918735
UMass Amherst Award ID: 0918653
Grant period: September 1, 2009 through August 31, 2012
Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

The Data Games project wishes to thank Doug Page and his students at Galileo Academy of Science and Technology, Josh Zucker, Lick-Wilmerding High school, Lina Huang and project staff at KCP Technologies and UMass Amherst, for their valuable support of this research work.

"Flat" data table

Vehicle data clustered by Date/Time

Vehicle data clustered by 1) Date/Time and 2) Direction