# Understanding Self-Directed Learning with Sequential Pattern Mining

Sungeun An[1], Spencer Rugaber[1] Jennifer Hammock[2], and Ashok K. Goel[1]

[1] School of Interactive Computing, Georgia Institute of Technology, Atlanta GA 30308, USA
[2] National Museum of Natural History, Smithsonian Institution, Washington, D.C. 20002, USA
sungeun.an@gatech.edu

**Abstract.** We describe a study on the use of an online laboratory for self-directed learning through the construction and simulation of conceptual models of ecological systems. We analyzed the modeling behaviors of 315 learners and 822 instances of learner-generated models using a sequential pattern mining technique. We found three types of learner behaviors: observation, construction, and exploration. We found that while the observation behavior was most common, exploration led to models of higher quality.

**Keywords:** Self-directed learning · Modeling and simulation · Online laboratory · Learning analytics.

## 1 Introduction

Self-directed online learning is becoming increasingly prevalent [5][9]. Self-directed learning here refers to non-formal inquiry-based learning outside classroom settings. One challenge in using online laboratories for self-directed learning outside K-12 pedagogical contexts is measurement of learning outcomes as there will be a large variance in the phenomena being modeled as well as in the goals and behaviors of the learners. Many studies on the use of online laboratories for learning focus on pedagogical contexts in K-12 education with well-defined problems and well-defined learning goals, assessments, and outcomes [2][7][4][6]. At present there is a lack of understanding of the processes and outcomes of self-directed learning in online laboratories. As online laboratories become increasingly widespread, it is important to not only formulate appropriate measures of learning but also to validate learning theories and findings from the literature.

To explore this research goal, we used VERA, a publicly available online laboratory for modeling ecological systems [1]. VERA is a web application that enables users to construct conceptual models of ecological systems and run agent-based simulations of these models. This allows users to explore multiple hypotheses about ecological phenomena and perform “what if” experiments to either explain an ecological phenomenon or predict the outcomes of changes to an ecological system. We investigate two research questions. *(1) What kinds of learning* *behaviors emerge in self-directed learning using VERA? (2) How do the learning behaviors relate to model quality?* In this study, the learning goals, as well as the demographics of the learners or even their precise geographical location are unknown; only the modeling behaviors and outcomes are observable.

## 2 Data Analysis

We analyzed the behaviors of 315 learners and the outcomes of 822 models generated by the learners over three years (2018-21). This section describes four analysis tasks: defining activities, creating activity sequences, segmenting activity sequences, and clustering similar sequences.

### 2.1 Learning Behaviors

Learners' log data within the VERA system creates timestamped records of actions such as adding a component, removing a component, or connecting two components with a relationship. These individual actions were categorized into three activity classes: *model construction, parameterization, and simulation* [7]. A *Model Construction* activity is defined as an insertion of a component or a relationship into a model or removal of a portion of the model. A *Parameterization* activity is defined as modification of a component's or relationship's parameter value. A *Simulation* activity is defined as the execution of a simulation.

We extracted activity sequences for every model created by a learner. For instance, if a learner performed a series of actions–adding a component, adding another component, and running a simulation–the activity sequence is 'ccs' (construction, construction, simulation). Given that an activity has no time duration in our data, we focus on the transition from one activity to another. This makes for 822 activity sequences, one for each model created by the 315 learners.

The activity sequences were divided into three groups of similar lengths (short, medium, long) based on two local minima in density using a segmentation optimization method (Kernel Density Estimation). Too short or too long sequences that are above a threshold of mean + 2*SD and below the threshold of mean - 2*SD were eliminated (N=33). Then the Levenshtein Distance was applied within each length group [8]. An Agglomerative Hierarchical method, the most common type of hierarchical clustering to group objects in clusters based on their similarity, is used to aggregate the most similar sequences based on the Levenshtein distance matrix [3].

### 2.2 Model Outcomes

We used two proxies to measure model quality. *Model complexity* is defined as the total number of model components and relationships (referred as *depth* in [9]). *Model variety* is defined as the number of unique components and relationships used in the model (commonly referred as *breadth* [9]).

## 3 Results and Discussions

A total of seven clusters from three length groups were derived based on hierarchical structure of the dendrogram and visually compared and merged into three clusters. Figure 1 illustrates the resulting three clusters in VERA with 16 randomly selected example sequences for each cluster using the visualization technique in [3]. Each horizontal line in the figure shows a sequence of activities in a model, the length of an activity in a sequence corresponds to the frequency of the activity. The sequence clusters have the following characteristics:

1. **Type 1** (N=382): *Observation.* The learners engage in experimenting with different simulation parameters with very little or no evidence of construction of conceptual models.
2. **Type 2** (N=338): *Construction.* The learners engage in short sessions of model construction with little or no simulation of the conceptual models.
3. **Type 3** (N=69): *Exploration (or Full Cycle).* The learners engage in a full cycle of model construction, parameterization, and simulation.

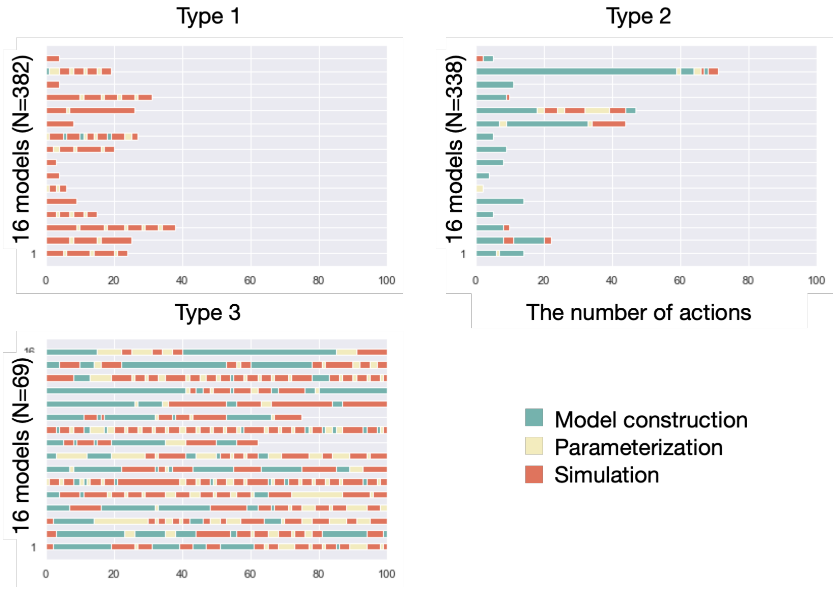

**Fig. 1.** Three Behavior Clusters of Similar Activity Sequences.

There was a statistically significant difference in model quality among the types as determined by one-way ANOVA test (complexity: $p$<.001, $f$=75.36; model variety: $p$<.001, $f$=26.80) and t-tests for pairwise comparisons. The conceptual models that manifested Type 3 behavior had the most complex models

($M$=12.5) followed by Type 1 ($M$=8.52) and Type 2 ($M$=6.22). (Type 1 & 2: $p$<.005, $t$=2.9835, Type 1 & 3: $p$<.001, $t$=−7.6527, Type 2 & 3: $p$<.001, $t$=−11.2651). The conceptual models that manifested Type 3 behavior had the most variety models ($M$=3.5) followed by Type 1 ($M$=2.9) and Type 2 ($M$=2.3) (Type 1 & 2: $p$<.01, $t$=2.6965, Type 1 & 3: $p$<.001, $t$=−5.8629, Type 2 & 3: $p$<.001, $t$=−6.5342).

## 4 Conclusion

We derive two main conclusions from the results. First, learners manifest three types of modeling behaviors in self-directed learning using VERA: observation (simulation focused), construction (construction focused), and full exploration (model construction, evaluation and revision). Second, learners who explored the full cycle of model construction, evaluation and revision generated models of higher quality.

**Acknowledgements** This research was supported by US NSF grant #1636848. We thank members of the VERA project, especially Luke Eglington and Stephen Buckley. This research was conducted in accordance with IRB protocol #H18258.

## References

1. An, S., Bates, R., Hammock, J., Rugaber, S., Weigel, E., Goel, A.: Scientific modeling using large scale knowledge. In: International Conference on Artificial Intelligence in Education. pp. 20–24. Springer (2020)
2. Basu, S., Dickes, A., Kinnebrew, J.S., Sengupta, P., Biswas, G.: Ctsim: A computational thinking environment for learning science through simulation and modeling. In: CSEDU. pp. 369–378. Aachen, Germany (2013)
3. Desmarais, M., Lemieux, F.: Clustering and visualizing study state sequences. In: Educational Data Mining 2013 (2013)
4. Gobert, J.D., Sao Pedro, M., Raziuddin, J., Baker, R.S.: From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. Journal of the Learning Sciences **22**(4), 521–563 (2013)
5. Haythornthwaite, C., Kumar, P., Gruzd, A., Gilbert, S., Esteve del Valle, M., Paulin, D.: Learning in the wild: coding for learning and practice on reddit. Learning, media and technology **43**(3), 219–235 (2018)
6. van Joolingen, W.R., de Jong, T., Lazonder, A.W., Savelsbergh, E.R., Manlove, S.: Co-lab: research and development of an online learning environment for collaborative scientific discovery learning. Computers in human behavior **21**(4), 671–688 (2005)
7. Joyner, D.A., Goel, A.K., Papin, N.M.: Mila–s: generation of agent-based simulations from conceptual models of complex systems. In: Proceedings of the 19th international conference on intelligent user interfaces. pp. 289–298 (2014)
8. Levenshtein, V.I., et al.: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady. vol. 10, pp. 707–710. Soviet Union (1966)
9. Scaffidi, C., Chambers, C.: Skill progression demonstrated by users in the scratch animation environment. International Journal of Human-Computer Interaction **28**(6), 383–398 (2012)