Progsnap: Sharing Programming Snapshots for Research

David Hovemeyer, York College of Pennsylvania
Arto Hellas, University of Helsinki
Andrew Petersen, University of Toronto Mississauga
Jaime Spacco, Knox College

tl;dr

- Our goal: to support easier, automated analysis of programming snapshot data
- Progsnap is a data interchange format for programming snapshot data that includes ...
  - A Python library for reading Progsnap data sets
  - Support for data from multiple institutions
  - Export utilities for CloudCoder [2], and PCRS [4]
- `https://cloudcoderdotorg.github.io/progsnap-spec`

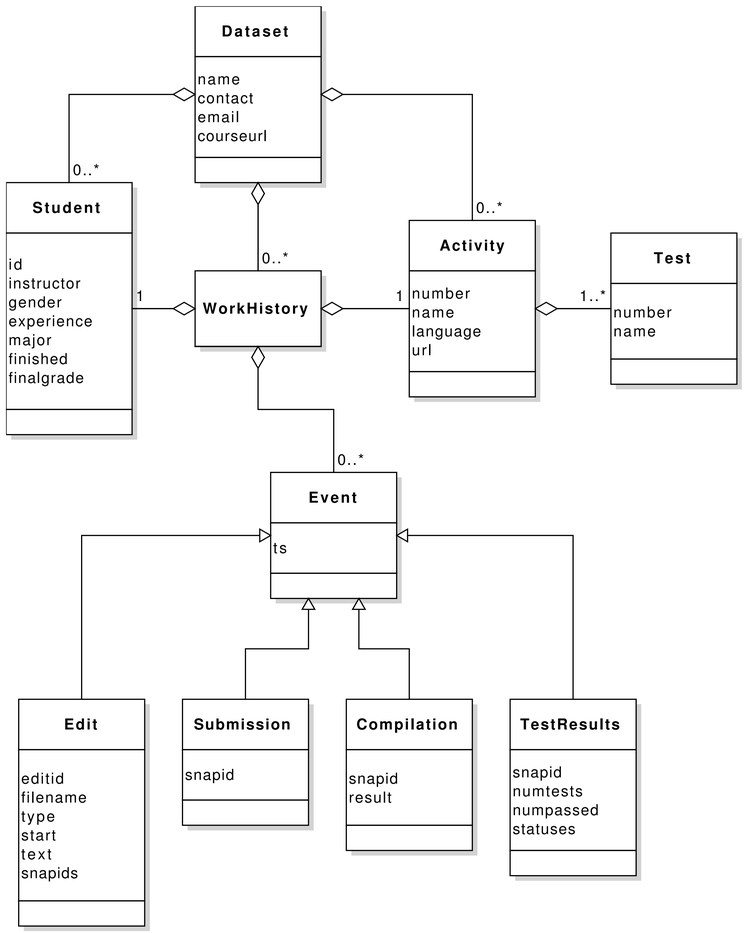
The problem

- Programming exercise systems are collecting lots of data:
  - Edits
  - Submissions
  - Compilation results
  - Test results
- The wealth of data creates opportunities to study how students learn to program
- Problem: different systems use different data models, data is stored in different formats, and different amounts and types of data are collected
- For these reasons, automating data analysis is difficult
  - Especially for multi-institution studies where different systems are being used

Our solution

- **Progsnap** [3] is a data model and data representation for programming snapshot data
- Data model is intended to be generic and easily exportable from programming exercise systems (see UML on right)
- A Python library makes it easy to read and analyze data in Progsnap datasets (see code example on right)
- Data representation is JSON objects stored in text files (possibly in zip files)
- Extensibility: custom tags and fields may be used for information beyond scope of standard data types

Data model (v0.1)



Isn't this the same as Blackbox [1]?

- Progsnap data represents work in a *course*: more information about context is known
- Progsnap is an interchange format, not a repository

Future directions (v0.2+)

- Richer data model:
  - Assignments (collection of related Activities)
  - Compiler diagnostics for Compilation events
  - More accurate diagnostics for runtime exceptions
- Data export from more systems
- Publish/access data via network

Code example

```
# Find the activity with the highest average number of
# submissions per student (omitting instructors)

import sys, progsnap

filename = sys.argv[1]
dataset = progsnap.Dataset(filename)

activity = None
highest_avg = 0

for a in dataset.activities():
  total = 0
  n = 0
  for wh in dataset.work_histories_for_activity(a):
    student = dataset.student_for_id(wh.student_id())
    if not student.instructor():
      subs = [e for e in wh.events()
              if type(e) is progsnap.Submission]
      total += len(subs)
      n += 1
  avg = total / n
  if avg > highest_avg:
    activity = a
    highest_avg = avg

print("Activity {}, {} submissions/student"
      .format(activity.number(), highest_avg))
```

References

[1] Neil Christopher Charles Brown, Michael Kölling, Davin McCall, and Ian Utting. Blackbox: A large scale repository of novice programmers' activity. In *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, SIGCSE '14, pages 223–228, New York, NY, USA, 2014. ACM.

[2] Andrei Papancea, Jaime Spacco, and David Hovemeyer. An open platform for managing short programming exercises. In *Proceedings of the Ninth Annual International ACM Conference on International Computing Education Research*, ICER '13, pages 47–52, New York, NY, USA, 2013. ACM.

[3] Progsnap: Home page. `https://cloudcoderdotorg.github.io/progsnap-spec`, 2017.

[4] Daniel Zingaro, Yuliya Cherenkova, Olessia Karpova, and Andrew Petersen. Facilitating code-writing in pi classes. In *Proceedings of the 44th ACM Technical Symposium on Computer Science Education*, SIGCSE '13, pages 585–590, New York, NY, USA, 2013. ACM.