MOTIVATING INTRODUCTORY COMPUTING WITH PEDAGOGICAL DATASETS

Austin Cory Bart
Computer Science Applications, Virginia Tech
March 22, 2017
Thanks!

Clifford A. Shaffer  Eli Tilevich  Dennis Kafura  Brett Jones  Phill Conrad

And many others!
Research Question

“Can a Data Science context motivate introductory computing students, particularly non-Computing majors?”
Contributions

• A model for characterizing student motivation with respect to course components
• New technology to support data science as an introductory computing context
• A large collection of real-world datasets for non-computing majors
• Evidence for value of a data science context as a motivating course component
• Evidence that connects course content with engagement outcomes
Publications


(Related Publications)

Overview

Motivation

Prior Work

Technology

Results
Computer Science For All
Diverse Majors... with Rich Knowledge

- Animal Sciences
- Chemistry
- Theater Arts
- Building Construction
- English
- Biological Sciences
- Education
- History
(1) No Prior Background

“I’ve never done this before.”
(2) Low Self-efficacy

“I have no idea how to do this!”
(3) Unclear on Why

“Why am I doing this?”
MUSIC Model of Academic Motivation

Students are more motivated when they perceive that:

1. they are empowered,
2. the content is useful to their goals,
3. they can be successful,
4. they are interested, and
5. they feel cared for by others in the learning environment

Motivation ➔ Engagement

**Motivation**
- Engagement
- Usefulness
- Success
- Interest
- Caring

**Engagement Outcomes**
- Persistence
- Proactivity
- Attendance
- Learning
- ...

13
Situated Learning

- Lave and Wenger
- “Learning occurs as a function of the activity, context, and culture”
A spectrum

**Context**
- Games
- Websites
- Mobile Apps
- Images
- Audio
- Animations
- Media
- Computation
- Scientific Computing
- Scientific Modelling
- Math
- Development

**Content**
- Recursion
- Iteration
- IF
- Assignment
- Data Structures
- Algorithms
- Lists
- Arrays
- Dictionaries
- Integers
- Booleans
- Algorithms
- Data Structures
- Assignment
- Lists
- Arrays
- Dictionaries
- Integers
- Booleans

FOR-EACH
WHILE
Interesting Contexts
Authenticity

• Situated Learning
• “Relevant”, “Real-world”
• Media Computation as an “Imagineered Authentic Experience”

Why *are* we teaching computing?

“A Tidal Wave of Data”
Highlighted Literature

- DePasquale 2006 – Real-world web APIs in CS2
- Sullivan 2013 – Data Science for non-majors
- Silva 2014 – Big Data in introductory computing
- Hall-Holt 2014 – Statistics in introductory computing
- Anderson 2014 – Real world data in CS1
- Subramanian 2014 – Visualization of data structures with real data (BRIDGES)
Problem – We Need Data

• ICPSR – Tightly controlled datasets
• UCI Machine Learning – Only for machine learning
• Census.gov, Kaggle, etc. – Not ready for beginners
Technology

- RealTimeWeb – real-time data for introductory computing
- CORGIS – real-world data for introductory computing
VT Bus Tracking API

Dr. Eli Tilevich

Dr. Cliff Shaffer
RealTimeWeb – Real-time data
So many Points of Failure!

RealTimeWeb – Secret Sauce

Client Library

Online Web Service

Local Cache File

!.getData()!

[!.searchBusinesses()]
[!.getEarthquakes()]
[!.getBuses()]
[...]
# RealTimeWeb - Deployment

<table>
<thead>
<tr>
<th>Semester</th>
<th>School</th>
<th>Course</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring 2013</td>
<td>Virginia Tech</td>
<td>CS-2</td>
</tr>
<tr>
<td>Fall 2013</td>
<td>University of Delaware</td>
<td>CS-1</td>
</tr>
<tr>
<td></td>
<td>Virginia Tech</td>
<td>CS-2</td>
</tr>
<tr>
<td></td>
<td>Virginia Tech</td>
<td>Data Structures &amp; Algos</td>
</tr>
<tr>
<td>Spring 2014</td>
<td>Virginia Tech</td>
<td>CS-2</td>
</tr>
</tbody>
</table>
RealTimeWeb - Studies

N=370, 14% female
University of Delaware, Virginia Tech
CS1, CS2, and DSA
RealTimeWeb - Hazards

- Limited APIs
- Maintenance was hard
- Impact on CS motivation was minimal
CORGIS

The Collection Of Really Great, Interesting, Situated Datasets
Metrics

44 datasets

267 mB

420,672 rows

9,365,520 values
Datasets
Connecting to Students’ Majors

- Criminal Justice
- Aerospace
- Theater Arts
- Building Construction
- English
- Geological Science
- Education
- History
- Immigration
- Books
- Weather
- Education
- Airlines
- Crime
- Construction
- Theatre
Architecture

Raw Datasets
- CSV: ABC, XYZ, 173, 441, 0, 27, ..., ...
- XML: <?xml version ...>
- TXT: Never gonna give you up, never gonna let you down...
- HTML: <html> <body> <div> ...

Cleaned Data
- JSON: 
  
  [
    {...},
    ...
  ]

Final Datasets
- Python
- BlockPy
- SQL
- JSON
- CSV

Specification File

Manual Preprocess

Automatic Build
Java, Python, Racket

// Java
import corgis.crime.StateCrimeLibrary;
import corgis.crime.domain.Report;
import java.util.ArrayList;
public class Main {
    public static void main(String[] args) {
        StateCrimeLibrary scl = new StateCrimeLibrary();
        ArrayList<Report> reports = scl.getAll();
    }
}

; Racket
(require crime)
(define reports (crime-get-all))

# Python
import crime
crime_reports = crime.get_all()
BlockPy
Visualizer Demo
Interventions

• Computational Thinking Course
  ❖ Basic programming
  ❖ Social Impacts
  ❖ Data Science

• 6 semesters taught

• Audience
  ❖ Non-computing majors
  ❖ Freshmen -> Senior
  ❖ Gender balanced
Course Evaluation

- Retention
- More-Computing
- Gender
- Learning

Survey Timeline

1 - Overview
   - Student Demographics
2 - Modelling
   - Motivational Survey #1
   - Keystroke Logs
3 - BlockPy
   - Classwork Completion and Lateness
4 - Spyder
   - Homework Completion and Lateness
   - Motivational Survey #2
5 - Miniproject
   - Reading Quizzes Completion and Lateness
6 - Project
   - Final Project
   - Motivational Survey #3

Attendance
### Motivation × Course Components

#### Motivational Components

<table>
<thead>
<tr>
<th>Motivational Component</th>
<th>eMpowerment</th>
</tr>
</thead>
<tbody>
<tr>
<td>“I believe that I will have freedom to explore my own interests when I…”</td>
<td></td>
</tr>
<tr>
<td>“I believe it will be useful to my long-term career goals to…”</td>
<td>Usefulness</td>
</tr>
<tr>
<td>“I believe I will be successful in this course when I…”</td>
<td>Success</td>
</tr>
<tr>
<td>“I believe it will be interesting to…”</td>
<td>Interest</td>
</tr>
<tr>
<td>“I believe that my instructors and peers will care about me when I…”</td>
<td>Caring</td>
</tr>
</tbody>
</table>

#### Course Component

<table>
<thead>
<tr>
<th>Course Component</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>“... learn to write computer programs”</td>
<td>Programming</td>
</tr>
<tr>
<td>“... learn to work with abstraction”</td>
<td>Abstraction</td>
</tr>
<tr>
<td>“... learn about the social impacts of computing”</td>
<td>Social Ethics</td>
</tr>
<tr>
<td>“... work with real-world data related to my major”</td>
<td>Data Science</td>
</tr>
<tr>
<td>“... work with my cohort”</td>
<td>Collaboration</td>
</tr>
</tbody>
</table>

#### Likert

<table>
<thead>
<tr>
<th>Likert</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly Disagree</td>
</tr>
<tr>
<td>Disagree</td>
</tr>
<tr>
<td>Somewhat Disagree</td>
</tr>
<tr>
<td>Neither Agree nor Disagree</td>
</tr>
<tr>
<td>Somewhat Agree</td>
</tr>
<tr>
<td>Agree</td>
</tr>
<tr>
<td>Strongly Agree</td>
</tr>
</tbody>
</table>
Students’ sense of the usefulness of various course components was highest for the **context**, lowest for the **content**.
V-Shaped Empowerment

N = 85, 62% Female

Students’ sense of agency decreases during the BlockPy and Spyder portions of the course, then increases during the final projects.
V-Shaped Interest

N = 85, 62% Female

Students’ interest decreases during the BlockPy and Spyder portions of the course, then increases during the final projects.
## Preference for Contexts

<table>
<thead>
<tr>
<th>Preference for Contexts</th>
<th>Likert</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Working with data sets related to your major”</td>
<td>Strongly Avoid</td>
</tr>
<tr>
<td>“Working with pictures, sounds, movies”</td>
<td>Avoid</td>
</tr>
<tr>
<td>“Making games and animations”</td>
<td>Somewhat Avoid</td>
</tr>
<tr>
<td>“Making websites”</td>
<td>Neither Prefer nor Avoid</td>
</tr>
<tr>
<td>“Making scientific models of real-world phenomenon”</td>
<td>Somewhat Prefer</td>
</tr>
<tr>
<td>“Controlling robots or drones”</td>
<td>Prefer</td>
</tr>
<tr>
<td>“Making phone apps”</td>
<td>Strongly Prefer</td>
</tr>
</tbody>
</table>

Data | Media | Games | Web | Scientific | Robots | Mobile
Preference for Contexts

N = 85, 62% Female

Students’ preferred a Data Science context over all others at the end, but Media Comp at the beginning. there were a number of V-shaped trends that occurred.

* No significant difference with Media Computation in S3, according to matched-pairs T-test
## Engagement (Intent to Continue)

<table>
<thead>
<tr>
<th>Intent to Continue</th>
<th>Likert</th>
</tr>
</thead>
<tbody>
<tr>
<td>“I will try to learn more about computing, either through a course or on my own.”</td>
<td>Strongly Disagree</td>
</tr>
<tr>
<td></td>
<td>Disagree</td>
</tr>
<tr>
<td>“I will recommend this class to others.”</td>
<td>Somewhat Disagree</td>
</tr>
<tr>
<td></td>
<td>Neither Agree nor Disagree</td>
</tr>
<tr>
<td>“I will directly apply what I have learned in my career.”</td>
<td>Somewhat Agree</td>
</tr>
<tr>
<td></td>
<td>Agree</td>
</tr>
<tr>
<td></td>
<td>Strongly Agree</td>
</tr>
</tbody>
</table>
Engagement (Intent to Continue)

N = 85, 62% Female

Although students would recommend the course, many did not intend to continue learning more computing or applying what they learned. The trend was negative from S1 to S2, and polarizing in S2 to S3.
Engagement vs. Components

Pearson correlation of “Student’s intent to continue learning computing” with students’ perception of each course and motivational component

<table>
<thead>
<tr>
<th></th>
<th>eMpowerment</th>
<th>Usefulness</th>
<th>Success</th>
<th>Interest</th>
<th>Caring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstraction</td>
<td>.087</td>
<td>.276</td>
<td>.184</td>
<td>.124</td>
<td>.288</td>
</tr>
<tr>
<td>Cohort</td>
<td>-.011</td>
<td>.064</td>
<td>.046</td>
<td>.001</td>
<td>.152</td>
</tr>
<tr>
<td>Data</td>
<td>-.046</td>
<td>.088</td>
<td>.019</td>
<td>.115</td>
<td>.134</td>
</tr>
<tr>
<td>Ethics</td>
<td>.025</td>
<td>.203</td>
<td>.196</td>
<td>.082</td>
<td>.255</td>
</tr>
<tr>
<td>Programming</td>
<td>.166</td>
<td><strong>.406</strong></td>
<td><strong>.354</strong></td>
<td><strong>.341</strong></td>
<td>.257</td>
</tr>
</tbody>
</table>

N = 85, 62% Female

Intent to continue seems to be correlated with the **content**, not the **context**.
Limitations

• Only included students who...
  ❖ Completed all three surveys
  ❖ Gave consent
  ❖ Self-enrolled in the course

• Self-report data

• N=85, relatively small sample

• Might not generalize to other institutions

• Anonymized, not anonymous
Take-aways

• Data Science seems to be a preferable context for students, across genders, by the end of the course

• The format of the final project was an important motivating factor

• Context, and in particular Data Science, can seem to provide motivation in ways that content cannot

• But some engagement outcomes might be more connected to content than context
Future Work

• Expand CORGIS
  ❖ More Datasets
  ❖ Better Datasets
  ❖ More Tools
  ❖ More Domains

• Expand Studies
  ❖ Confirm results
  ❖ Connect motivation to learning outcomes
  ❖ Determine causality of content’s relationship with intent to continue
Questions?

https://think.cs.vt.edu/corgis

Artwork by Eleonor Bart
Trends in Motivation
<table>
<thead>
<tr>
<th>Spring 2016</th>
<th>eMpowerment</th>
<th>Usefulness</th>
<th>Success</th>
<th>Interest</th>
<th>Caring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstraction</td>
<td>.458</td>
<td>.699</td>
<td>.614</td>
<td>.488</td>
<td></td>
</tr>
<tr>
<td>Cohort</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethics</td>
<td>.485</td>
<td>.418</td>
<td>.323</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Programming</td>
<td>.437</td>
<td><strong>.823</strong></td>
<td>.600</td>
<td>.638</td>
<td></td>
</tr>
</tbody>
</table>

Continue Learning, Applying, and/or Recommend Course
N =36
50% female
Students’ Perception of Caring

N = 85, 62% Female
We seem to be good instructors
N = 85, 62% Female
V-shaped in some cases, but otherwise increasing
Most students (85%) received a Good or Excellent on each element.
## Situated Learning vs. Motivation

<table>
<thead>
<tr>
<th>Situated Learning Component:</th>
<th>Context</th>
<th>Content</th>
<th>Facilitations</th>
<th>Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
<td>&quot;Game Design&quot;</td>
<td>&quot;For Loops&quot;</td>
<td>Blocks-based environment, teaching assistants, etc.</td>
<td>Exams, performance review, code review</td>
</tr>
<tr>
<td>eMpowerment</td>
<td>Am I restricted by the context to explore what I want?</td>
<td>Do I have control over the depth/breadth/direction of what I am learning?</td>
<td>Do these scaffolds let me accomplish things I couldn't?</td>
<td>Can I explore my limitations and successes in this assessment?</td>
</tr>
<tr>
<td>Usefulness</td>
<td>Is this situated in a topic that's worth learning?</td>
<td>Is the content itself worth learning?</td>
<td>Do these scaffolds let me learn enough to still be useful?</td>
<td>Do I feel that performing well on the assessment is important?</td>
</tr>
<tr>
<td>Success</td>
<td>Do I believe I can understand this context?</td>
<td>Do I believe I can understand this material?</td>
<td>Do these scaffolds hinder me or help me?</td>
<td>Can I succeed at this assessment?</td>
</tr>
<tr>
<td>Interest</td>
<td>Is this situated in something I find boring/interesting?</td>
<td>Is the material inherently interesting?</td>
<td>Do the scaffolds support my interest in the activity or detract from the experience?</td>
<td>Am I interested in the assessment experience?</td>
</tr>
<tr>
<td>Caring</td>
<td>Does the context give opportunities for the instructor and peers to show they care?</td>
<td>Does the content give opportunities for the instructor and peers to show they care?</td>
<td>Do the scaffolds give opportunities for the instructor and peers to provide support?</td>
<td>Does the assessment give opportunities for the instructor and peers to show they care?</td>
</tr>
</tbody>
</table>
Big Idea: Real-World Data
Complete Picture

1. Designs
2. Generates
3. Caches
4. Submits
5. Assigns
6. Uses
7a. Access
7b. Alt Access

Developer

Client Library Specification

Real-time Data Source

Client Library Data Cache

RealTimeWeb Gallery

Instructor

Students

Client Libraries
Situated Learning Framework
Choi & Hannafin

Context

Content

Assessment

Assessment

Assessment

Scaffolds

Scaffolds

Scaffolds

Scaffolds
Cache Files = Sophisticated Snapshots

getEarthquakes() => [ <raw usgs data>, <raw usgs data>, ...]

june_18_2013.json

<table>
<thead>
<tr>
<th>Call</th>
<th>Returns</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>5 earthquakes</td>
</tr>
<tr>
<td>#2</td>
<td>2 earthquakes</td>
</tr>
<tr>
<td>#3</td>
<td>7 earthquakes</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Three Components

Client Libraries

Curated Gallery

Library Generator
Gallery - Initial Offering

- Earthquakes
- Weather
- Stocks
- Reddit
- Magic the Gathering
Client Library Building

- API Spec
- Jinja2 Templates
- Java
Pedagogical Dataset Design

1. General Advice
   1. Have a plan
   2. Build for your audience
   3. Iterate
   4. Standardize your process
   5. Keep a clean workspace
   6. Manage dataset health
   7. Beware breaking convention
   8. Work in phases
   9. Understand the context

2. Collecting data
   1. Hunting sources
   2. Working with file formats
   3. Scraping web data
   4. Mining real-time data
   5. Legality of your data
   6. Synthesizing datasets

3. Restructuring data
   1. Choose your target structure
   2. Layering columnar data
   3. Converting XML to JSON
   4. Working with indexes
   5. Collapsing fields
   6. Stacking data
   7. Redundant total field

4. Manipulating the data
   1. Standardize fields
   2. Names are important
   3. Working with bad data
   4. Cleaning up by hand
   5. Reshaping data
   6. Extending a dataset with divined data

5. Working with Data Types
   1. Numbers
   2. Textual
   3. Dates and times
   4. Measurements
   5. Locations
   6. URLs
   7. Enumerated data

6. Knowing the data
   1. Nobody reads the documentation
   2. Learning the structure
   3. Learning the distribution
   4. Disseminating materials
   5. Monitor usage
4. CONCLUSION

In this exposition, I showed how to infuse some algorithmic and mathematical aspects to guide the programming experience. The main theme is Fibonacci (and the golden ratio), which is a pleasant topic for many students. The typical paradigm that I support here is to first start with a warm up question (one that is not too trivial), then to...