Data-Centricity: A Challenge and Opportunity for CS Education

Shriram Krishnamurthi, Brown University
Joint work with Kathi Fisler
What is in (Collegiate) CS1?

- Numbers / Strings
- Variables
- Assignment statements
- Conditionals
- Loops
- Functions
- Arrays
## Data Structure Progress Since the 1970s

<table>
<thead>
<tr>
<th>Year</th>
<th>Language</th>
<th>Data Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1970s</td>
<td>Pascal</td>
<td>arrays</td>
</tr>
<tr>
<td>1980s</td>
<td>C</td>
<td>arrays</td>
</tr>
<tr>
<td>1990s</td>
<td>C++</td>
<td>arrays</td>
</tr>
<tr>
<td>2000-2015</td>
<td>Java</td>
<td>ArrayList</td>
</tr>
<tr>
<td>2015-now</td>
<td>Python</td>
<td>associative arrays</td>
</tr>
</tbody>
</table>
Pressure From Above:
“The Unreasonable Effectiveness of Data”
Pressure From Below:
Code Synthesis from Text

Pressure From Below:
“LowCode/NoCode” Systems
Needs from many more disciplines

Student diversity concerns

Negative societal impacts
Major new curricular trend: Data Science

Courses, bootcamps, even degrees
How Should We View This?

“They don’t teach even the basics!”

“This is a threat to our degree programs!”

“This is actually a (valid) criticism of current CS Ed”
What Are the Valid Criticisms of CS 1?

Curricula do not engage students

Curricula do not connect to society/real-world phenomena

You can’t do anything useful after a course

Computing is used everywhere; students aren’t prepared
Reforming CS for the Data Era
Data Structure Progress Since the 1970s

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Language</th>
<th>Data Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1970s</td>
<td>Pascal</td>
<td>Arrays</td>
</tr>
<tr>
<td>1980s</td>
<td>C</td>
<td>Arrays</td>
</tr>
<tr>
<td>1990s</td>
<td>C++</td>
<td>Arrays</td>
</tr>
<tr>
<td>2000-2015</td>
<td>Java</td>
<td>ArrayList</td>
</tr>
<tr>
<td>2015-now</td>
<td>Python</td>
<td>Associative Arrays</td>
</tr>
</tbody>
</table>
Rich structure

Already parsed

Familiar even to kids
MODELING AS A CORE COMPONENT OF STRUCTURING DATA

CLIFFORD KONOLD
University of Massachusetts Amherst
konold@erra.umass.edu

WILLIAM FINZER
Concord Consortium
wfinzer@concord.org

KOSOOM KREETONG
University of Massachusetts Amherst
kkreetong@edu.edc.umass.edu
<table>
<thead>
<tr>
<th>First Name</th>
<th>Last Name</th>
<th>Email</th>
<th>Tickets</th>
<th>Discount</th>
<th>Birthday</th>
</tr>
</thead>
<tbody>
<tr>
<td>Josie</td>
<td>Zhao</td>
<td><a href="mailto:josie@steele.com">josie@steele.com</a></td>
<td>2</td>
<td>birthday</td>
<td>Jan 1, 1970</td>
</tr>
<tr>
<td>Guy Lewis</td>
<td>Steele Jr</td>
<td><a href="mailto:guy@steele.com">guy@steele.com</a></td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bart</td>
<td>Simple</td>
<td><a href="mailto:bart@simpsons.com">bart@simpsons.com</a></td>
<td>1</td>
<td>student</td>
<td></td>
</tr>
<tr>
<td>Robby</td>
<td></td>
<td><a href="mailto:robbys@simpsons.com">robbys@simpsons.com</a></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shweta</td>
<td>Mysore</td>
<td><a href="mailto:shweta@np.org">shweta@np.org</a></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alvina</td>
<td>Velasquez</td>
<td><a href="mailto:avlina@school.edu">avlina@school.edu</a></td>
<td>3</td>
<td>student</td>
<td></td>
</tr>
<tr>
<td>Sam</td>
<td>Ochibe</td>
<td><a href="mailto:s@sweb.com">s@sweb.com</a></td>
<td>10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data Commons

Data Commons aggregates data from a wide range of sources into a unified database to make it more accessible and useful. More on why we are building Data Commons.
“I chose my dataset because…” Students could choose more than one option. Here, 42.6% said they chose it because it was something they already knew something about, while just as many (40.2%) chose it because “I don’t know much about it, and I was curious”. In addition, 16% said “It affects me personally”. Numerous students used free-response to express specific interests such as representing a favorite hobby or activity, or “I am interested in the medical field”. Still, small numbers contained a potentially negative response: 11.7% said they had no reason, 4.3% said it was their partner’s choice, and 3.2% said they did not have a choice (presumably meaning the dataset was chosen by the teacher).

Overall, then, we see that the personalization is accompanied in general by a high degree of engagement and, most importantly, low degrees of dis-engagement. This suggests that the technique of providing a (limited) range of curated datasets still gives many students room for personal expression.
<table>
<thead>
<tr>
<th>First Name</th>
<th>Last Name</th>
<th>Email</th>
<th>Tickets</th>
<th>Discount</th>
<th>Birthday</th>
</tr>
</thead>
<tbody>
<tr>
<td>Josie</td>
<td>Zhao</td>
<td><a href="mailto:josie@zoho.com">josie@zoho.com</a></td>
<td>2</td>
<td>birthday</td>
<td>Jan 1, 1970</td>
</tr>
<tr>
<td>Guy Lewis</td>
<td>Steele Jr</td>
<td><a href="mailto:guy@steele.com">guy@steele.com</a></td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bart</td>
<td>Simple</td>
<td><a href="mailto:bart@simpsons.com">bart@simpsons.com</a></td>
<td>1</td>
<td>student</td>
<td></td>
</tr>
<tr>
<td>Robby</td>
<td></td>
<td><a href="mailto:robby@roby.com">robby@roby.com</a></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shweta</td>
<td>Mysore</td>
<td><a href="mailto:shweta@np.org">shweta@np.org</a></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alvina</td>
<td>Velasquez</td>
<td><a href="mailto:avlina@school.edu">avlina@school.edu</a></td>
<td>three</td>
<td>student</td>
<td></td>
</tr>
<tr>
<td>Sam</td>
<td>Ochibe</td>
<td><a href="mailto:s@sweb.com">s@sweb.com</a></td>
<td>10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

“How many tickets sold with a student discount?”

Interesting problems

Concrete, easy to envision and manipulate

Lend themselves to decomposition and planning
After 30 years, why hasn’t somebody beaten the Rainfall Problem? Why can’t someone teach a course with the explicit goal of their students doing much better on the Rainfall Problem — then publish how they did it? We ought to make measurable progress.

I don’t think that this is an impossible goal. In fact, I bet that some of the existing research projects in computing education could “beat” (generate published reports with better results) these current studies.

- The TeachScheme approach focuses on design based on data. I bet that their students could beat the Rainfall Problem or the McCracken working group problem.
Variety of programming forms (functions, methods, loops, ...)

```py
# Sieve tickets using discount:
t1 = sieve tickets using discount:
    discount == "student"
end

sum(extract tickets from t1 end)
```
Data quality, normalization, cleansing

Consequences of bad data
<table>
<thead>
<tr>
<th>First Name</th>
<th>Last Name</th>
<th>Email</th>
<th>Tickets</th>
<th>Discount</th>
<th>Birthday</th>
</tr>
</thead>
<tbody>
<tr>
<td>Josie</td>
<td>Zhao</td>
<td><a href="mailto:josie@steele.com">josie@steele.com</a></td>
<td>2</td>
<td>birthday</td>
<td>Jan 1, 1970</td>
</tr>
<tr>
<td>Guy Lewis</td>
<td>Steele Jr</td>
<td><a href="mailto:guy@steele.com">guy@steele.com</a></td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bart</td>
<td>Simple</td>
<td><a href="mailto:bart@simpsons.com">bart@simpsons.com</a></td>
<td>1</td>
<td>student</td>
<td></td>
</tr>
<tr>
<td>Robby</td>
<td></td>
<td>robbby</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shweta</td>
<td>Mysore</td>
<td><a href="mailto:shweta@np.org">shweta@np.org</a></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alvina</td>
<td>Velasquez</td>
<td><a href="mailto:avlina@school.edu">avlina@school.edu</a></td>
<td>three</td>
<td>student</td>
<td></td>
</tr>
<tr>
<td>Sam</td>
<td>Ochibe</td>
<td><a href="mailto:s@sweb.com">s@sweb.com</a></td>
<td>10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What gets collected about you?

What gets inferred about you?

What decisions are made on those inferences?
## Limitations

Ancestry = table: name, birthyear, eyecolor, fempar, malepar

<table>
<thead>
<tr>
<th>row</th>
<th>name</th>
<th>birthyear</th>
<th>eyecolor</th>
<th>fempar</th>
<th>malepar</th>
</tr>
</thead>
<tbody>
<tr>
<td>row 1</td>
<td>&quot;Anna&quot;</td>
<td>1997</td>
<td>&quot;blue&quot;</td>
<td>&quot;Susan&quot;</td>
<td>&quot;Charlie&quot;</td>
</tr>
<tr>
<td>row 2</td>
<td>&quot;Susan&quot;</td>
<td>1971</td>
<td>&quot;blue&quot;</td>
<td>&quot;Ellen&quot;</td>
<td>&quot;Bill&quot;</td>
</tr>
<tr>
<td>row 3</td>
<td>&quot;Charlie&quot;</td>
<td>1972</td>
<td>&quot;green&quot;</td>
<td>&quot;NoInfo&quot;</td>
<td>&quot;NoInfo&quot;</td>
</tr>
<tr>
<td>row 4</td>
<td>&quot;Ellen&quot;</td>
<td>1945</td>
<td>&quot;brown&quot;</td>
<td>&quot;Laura&quot;</td>
<td>&quot;John&quot;</td>
</tr>
</tbody>
</table>

...  

Motivate new data structures (and rest of CS)
A Data-Centric Introduction to Computing

Kathi Fisler  Shriram Krishnamurthi  Benjamin S. Lerner  Joe Gibbs Politz
Data-Centric Intro to Computing (DCIC)

Structure of code follows structure of image
Each topic motivated by something students want to do with a dataset

Information and code have structure

Computations transform and summarize data(sets)

Sometimes, we aggregate information across datapoints

Attributes within dataset might have structure

Datapoints might have relationships

Programs sometimes need to update datasets

Improve efficiency of working with associative data
Research-Driven Curriculum!

Using Design Alternatives to Learn About Data Organizations
Xingjian Gu  Max A. Heller  Stella Li  Yanyan Ren  Kathi Fisler  Shriram Krishnamurthi
Contact: slik@cs.brown.edu
Computer Science Department
Brown University
Providence, RI, USA

Developing Behavioral Concepts of Higher-Order Functions
Shriram Krishnamurthi
Brown University
Providence, RI, USA
Kathi Fisler
Brown University
Providence, RI, USA

What Help Do Students Seek in TA Office Hours?
Yanyan Ren
Computer Science Department
Brown University
Providence, Rhode Island, USA
Shriram Krishnamurthi
Computer Science Department
Brown University
Providence, Rhode Island, USA
Kathi Fisler
Computer Science Department
Brown University
Providence, Rhode Island, USA

Adapting Student IDEs for Blind Programmers
Emmanuel Schanzer
schanzer@bootstrapworld.org
Brown University / Bootstrap
Providence, RI, USA
Sina Bahram
sina@sina bahram.com
Prime Access Consulting
Cary, NC, USA
Shriram Krishnamurthi
schanzer@bootstrapworld.org
Brown University / Bootstrap
Providence, RI, USA

Early Post-Secondary Student Performance of Adversarial Thinking
Nick Young
Brown University
Providence, RI, USA
Shriram Krishnamurthi
Brown University
Providence, RI, USA
Data-centricity: a challenge and opportunity for computing education

Authors: Shiram Krishnamurthi, Kathi Fisler

Publication: Communications of the ACM • july 2020 • https://doi.org/10.1145/3408056

All materials are free and on-line!

CACM July 2020
dcic-world.org

CSCI 0111
Computing Foundations: Data

(at Brown & others)
Stepping Back Out
Many CS1 courses are here. Frequently overlooked. Needs nontrivial CS. Increasing calls for Social Responsibility. Should have data & stats. Appealing to students across campus.
What About Primary/Secondary?

Ports all these ideas to students as young as 10yo
State of CS 1

The Rise of Data Science

Redesigning CS 1 with Data

dcic-world.org