

Reconciling the Promise and Pragmatics of Enhancing Computing Pedagogy with Data Science

Austin Cory Bart, Dennis Kafura, Clifford A. Shaffer, Eli Tilevich

Virginia Tech

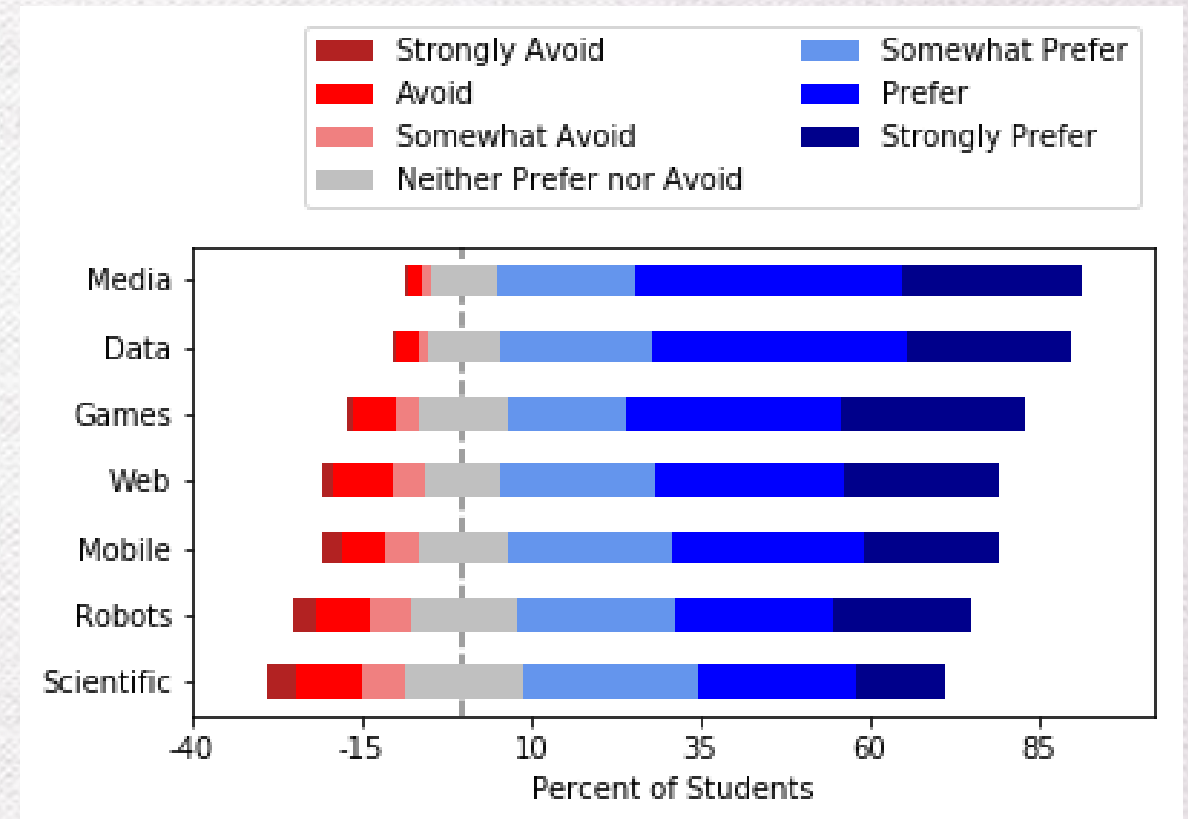
SIGCSE 2018 – Saturday, Mar 24 at 10:45am

Why are we here?

Promise	Problem	Want learners to engage with computing
	Solution	Data Science as a motivating context!*
Pragmatics	Problem	Real-world data is hard
	Solution	Pedagogical datasets!
	Problem	Pedagogical datasets are hard
	(Partial) Solution	A guide for developers to make Pedagogical Datasets

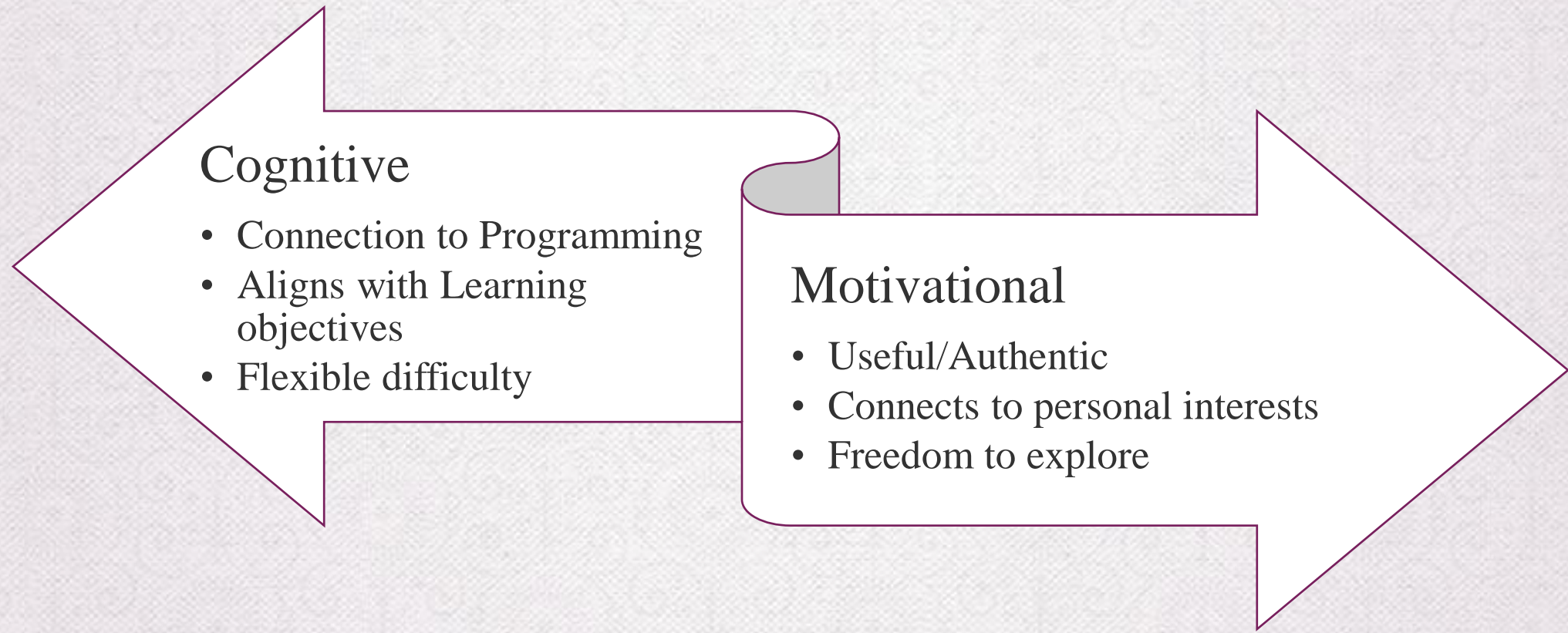
Student Preference for Possible Intro Contexts

- N=346
- Start of semester in non-majors course
 - Arts
 - Humanities
 - Social sciences
- No significant difference* between
 - Media Computation
 - Game Development
 - Data Science
 - But all others are significantly lower!



Mann-Whitney U, $\alpha < .01$

Data Science as an Introductory Context



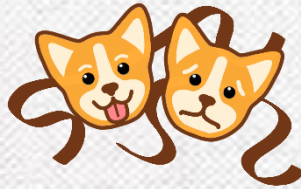
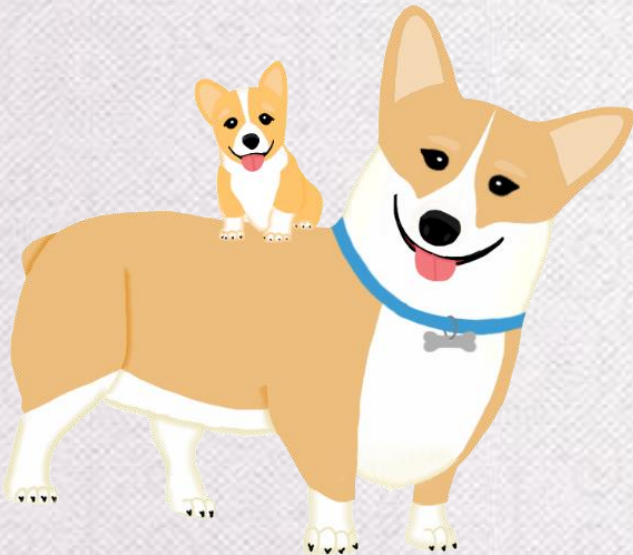
Growing Popularity

- Real World Data used in CS1 courses
 - DePasquele 2006
 - Anderson 2014
 - Hall-Holt 2015
- Teaching Data Science
 - Sullivan 2013
 - P. Anderson 2014
 - Mahadev 2015
 - DePratti 2017
- Technology
 - Subramanian 2014, 2018 – Visualization of data structures with real data (BRIDGES)
 - Hamid 2016, 2017 – Generalized framework for real-time APIs (Sinbad)
 - Bart 2014, 2017 – CORGIS

CORGIS

Collection of Really Great and Interesting dataSets

- Pedagogical Datasets specifically targeted at novices for learning purposes
- 40+ datasets in history, geology, medicine, criminology, etc.
- Free, open-source, available in many formats and interfaces



Theater



Airlines



Books



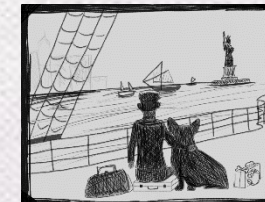
Weather



Construction



Crime



Immigration



Education

Life Is Never Simple

Inadequate datasets

- Schools dataset is broken
- Cars has wrong fields

40+ datasets very hard to maintain

- Datasets change
- Developers come and go

Doesn't cover all disciplines and students

- Missing areas: Fashion, sports

Doesn't cover all courses

- Hope you like lists of dictionaries

Pedagogical Datasets Development is Hard

- Requires Data Science expertise
- Each dataset is its own journey
- Limited data in the world, not always freely available

"The Pragmatics of Pedagogical Dataset Development"

- <https://think.cs.vt.edu/pragmatics>
- Guide to developing datasets
 - Tips
 - Tricks
 - Things to think about
- A living document, please give feedback
 - Overlaps with Data Science guides, but specialized

Pedagogical Concerns

What does
this field
mean?

How do I
get that
field out?

What
about next
semester?

What
about this
major?

What does
this value
mean?

Why
doesn't this
plot look
right?

How do I
give them
the file?

How do I
fix this
issue?



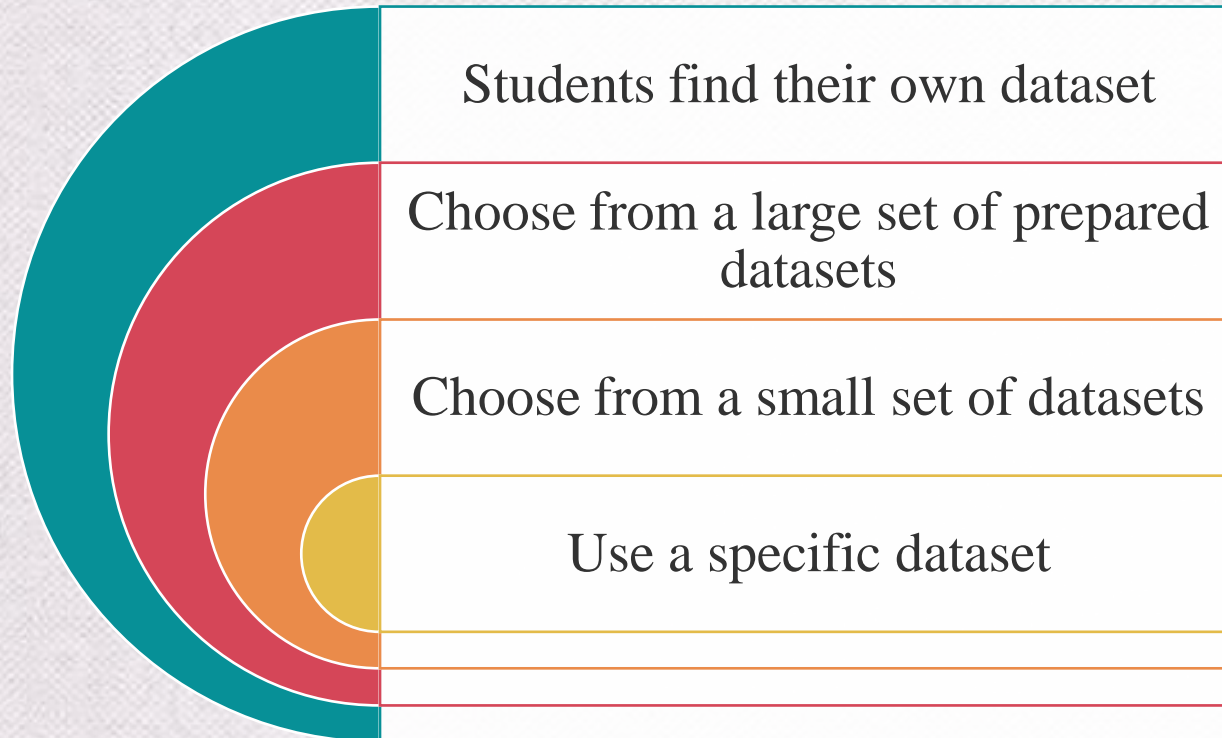
What Are You Teaching?

Lists	<ul style="list-style-type: none">• Iteration• Index lookup
Dictionaries	<ul style="list-style-type: none">• Dictionary lookup• Struct access
Dictionaries of dictionaries	<ul style="list-style-type: none">• Traversal of heterogenous data structures• Chained lookup
Lists of dictionaries	<ul style="list-style-type: none">• Iteration + Dictionary Lookup
Lists of lists (2D Array)	<ul style="list-style-type: none">• Nested loops

Suggestions

- Consider LOs
- Target structure

Assignment Scoping



More Freedom
Opportunities for creativity

More quality control
Easier to grade

What About Next Semester? Year? Decade?

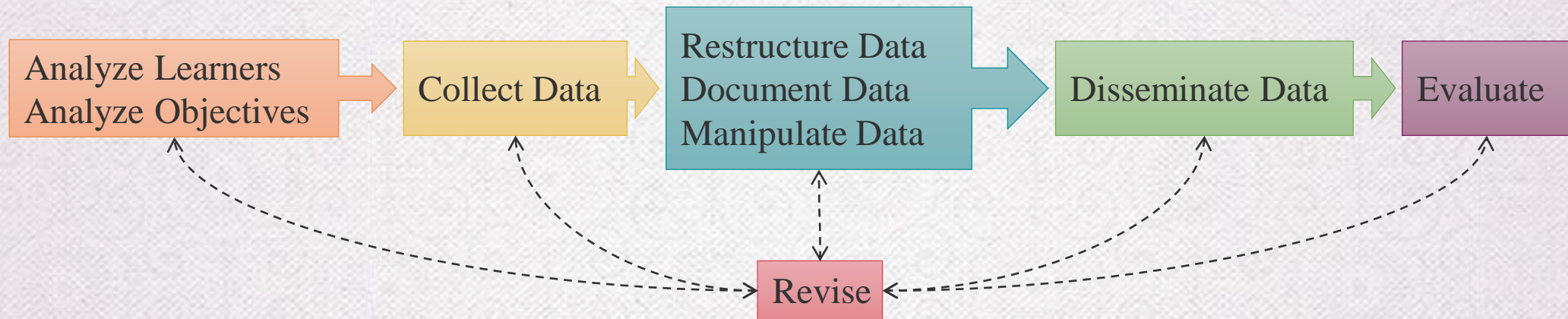
- Data decays, changes
- Datasets update, but not automatically
- Fixes, revising



Suggestions

- Formal processes
- Document

How Will You Approach This?



Suggestions

- Think systematically

Is the Size Right?

- Students have limited processing speed
- Students have limited space
- Students might not be able to download



Suggestions

- Pivot/Bin/Aggregate
- Stack/Unstack
- Sample down

Broadway Dataset

[illegible]

"I opened the file and it's broken!"

How Do I Disseminate?

- Hosting?
- Instructions?
- Digital Literacy?

Suggestions

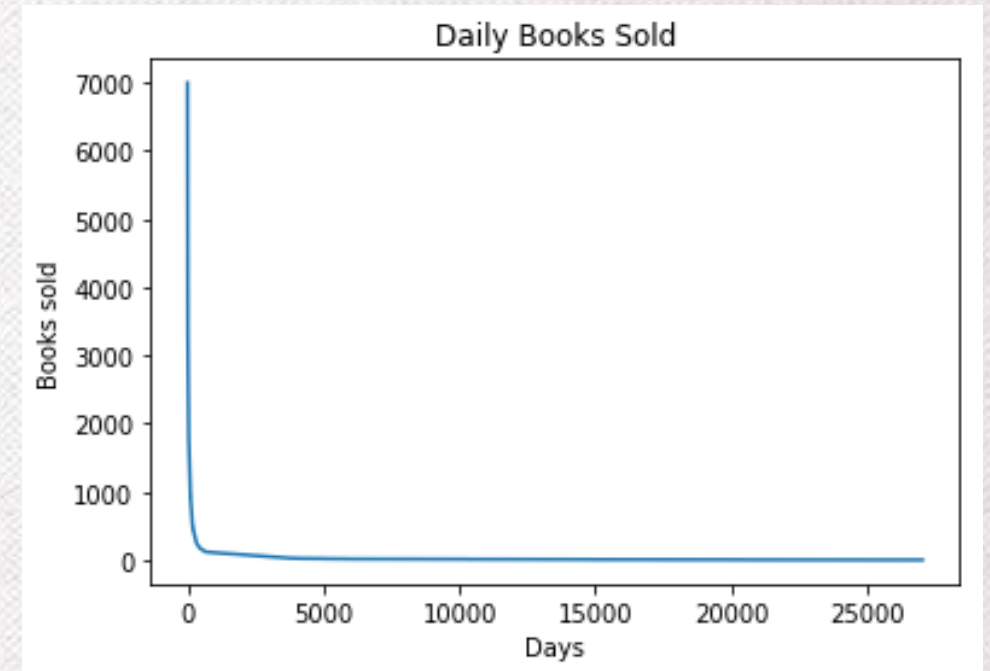
- Write instructions carefully
- Prepare "Common Issues"
- Have support available

Publishers Dataset

```
import publishers
import matplotlib.pyplot as plt

books = publishers.get_books()
daily_units_sold= []
for book in books:
    units_sold = book['daily average']['units sold']
    daily_units_sold.append(units_sold)

plt.plot(daily_units_sold)
plt.xlabel("Days")
plt.ylabel("Books sold")
plt.title("Daily Books Sold")
plt.show()
```



"... This data is NOT time-oriented; it is a collection of a bunch of different books at a single point in time."

"Principle of Least Effort"

- Students take the cognitive path of least resistance
 - Even if it makes no sense
- Work with the grain, not against it



Suggestions

- Natural key orderings
- Adjust expectations
- Field test

What's that Field?


- Abbreviations
 - "Temp" => "Temperature"? "# of Flights" => "Number of Flights"
- Capitalization
 - "Temperature" or "temperature"? "numberOfDogs" or "NumberOfDogs"?
- Allowing special symbols?
 - "High & Low Temperature"?
- Specificity?
 - "Crime Rate" => "Number of Crimes per 10k People"
- Units
 - Celsius or Fahrenheit?


Suggestions

- Choose names carefully
- Document
- Field-test

Interactive Documentation


Data Visualizer

 Kennel Home Courses Tools Admin About Contact Signed in as Cory Bart (log out)

 **Visualizer Datasets**
The Collection of Really Great, Interesting, Situated Datasets


By Austin Cory Bart, Ryan Whitcomb, Jason Riddle, Omar Saleem, Dr. Eli Tilevich, Dr. Clifford A. Shaffer, Dr. Dennis Kafura

Filter




Aids

Records of AIDS related statistics from several countries.
aids, death, disease, hiv, orphans, health, countries, world, gender, united nations, un




Airlines

Information about flight delays in major airports since 2003.
airplane, airports, travel, plane, air, flights, delays, national, united states, transportation




Billionaires

Information about over 2000 billionaires from around the world.
money, rich, wealthy, people, person, billionaire




Broadway

This library holds data about Broadway shows, such as tickets sold.
broadway, musical, theatre, tickets




Business Dynamics

The Business Dynamics Statistics (BDS) includes measures of establishment openings and closings, firm startups, job creation and destruction by firm size, age, and industrial sector, and several other statistics on business dynamics for the US.
government, united states, us, usa, business, businesses



Cars

This is a dataset about cars and how much fuel they use.
cars, vehicles, fuel



Classics

Records and computed statistics about the top 1000 books on Project Gutenberg.

Data Explorer

Overview

This library holds data about over Broadway shows, grouped over weeklong periods. Only shows that reported capacity were included. This dataset is made available by the Broadway League (the national trade association for the Broadway industry), and you can view the data.

Explore Structure

Explore Broadway data

Downloads

Download all of the following files.

1. [broadway.py](#) 
2. [broadway.db](#) 

Could Students Use It?

- Questions
 - Did students use it?
 - Did students enjoy it?
 - Did they understand it?
 - Did they know what to do when confused?
 - Did they have enough options?
- Mechanics
 - Surveys
 - Interviews
- Revise!

Suggestions

- Plan for Data Collection
- Plan up-front!

More Information

- Read the complete pragmatics
 - <https://think.cs.vt.edu/pragmatics>
 - Please contribute lessons, ideas, common issues through our github
- Try making a dataset
- Contribute to CORGIS
 - Need maintainers, documenters, developers
 - Good project for motivated students!
- Gratefully acknowledge the support of Virginia Tech and the National Science Foundation under Grants NSF DGE 0822220, NSF IUSE 1624320, and NSF IUSE 479632.