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

1. A. C. Bart, R. Whitcomb, E. Tilevich, C. A. Shaffer, D. Kafura, *Computing with CORGIS: Diverse, Real-world Datasets for. Introductory Computing (Best Paper)*, SIGCSE '17, Seattle, Washington. March, 2017.
2. D. Kafura, A. C. Bart, B. Chowdhury, *Design and Preliminary Results From a Computational Thinking Course*. ITiCSE'15, Vilnius, Lithuania. July 6-8, 2015.
3. A. C. Bart, J. Riddle, O. Saleem, B. Chowdhury, E. Tilevich, C. A. Shaffer, D. Kafura, *Motivating Students with Big Data: CORGIS and MUSIC*, Splash-E '14, Portland, Oregon. October 21-23, 2014.
4. A. C. Bart, E. Tilevich, T. Allevato, S. Hall, C. A. Shaffer, *Transforming Introductory Computer Science Projects via Real-Time Web Data*, SIGCSE '14, Atlanta, Georgia. March 5-8, 2014.
5. A. C. Bart, E. Tilevich, C. A. Shaffer, T. Allevato, S. Hall, *Using Real-Time Web Data to Enrich Introductory Computer Science Projects*, Splash-E '13, Indianapolis, Indiana. October 26-31, 2013.

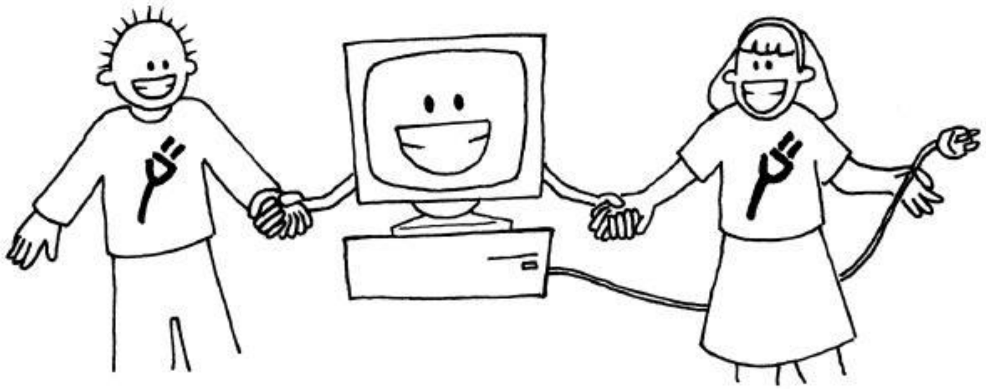
(Related Publications)

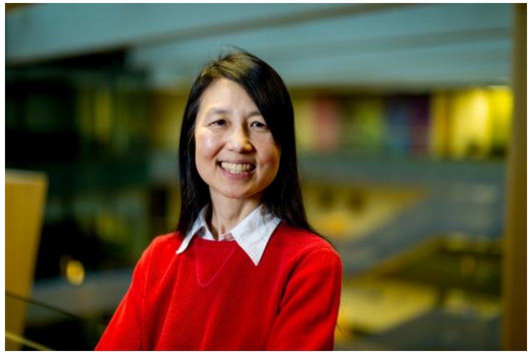
1. A. C. Bart, J. Tibau, E. Tilevich, C. A. Shaffer, D. Kafura, *Design and Evaluation of Open-access, Data Science Programming Environment for Learners*, IEEE Computer '17. May, 2017 (accepted).
2. A. C. Bart, J. Tibau, E. Tilevich, C. A. Shaffer, D. Kafura, *Implementing an Open-access, Data Science Programming Environment for Learners*, COMPSAC '16, Atlanta, Georgia. June 10-15, 2016.
3. A. C. Bart, C. A. Shaffer. *Instructional Design is to Teaching as Software Engineering is to Programming*. SIGCSE '16. Kansas City, MO. March 2-5, 2016.
4. A. C. Bart, E. Tilevich, C. A. Shaffer, D. Kafura, Position Paper: *From Interest to Usefulness with BlockPy, a Block-based, Educational Environment*, Blocks & Beyond '15, Atlanta, Georgia. October 21-23, 2015.

Overview

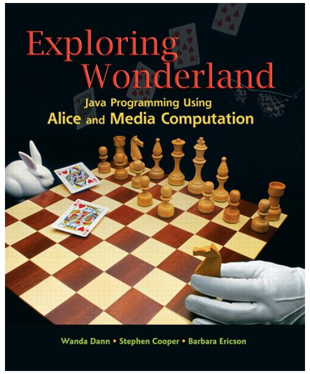
Motivation → Prior Work → Technology → Results

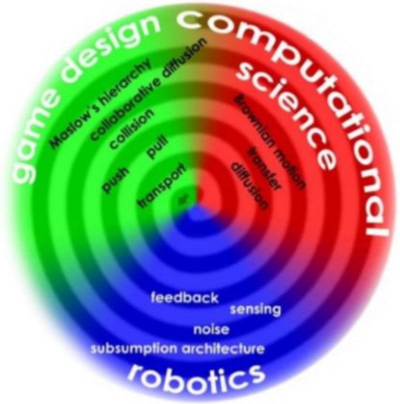
Computer Science For All



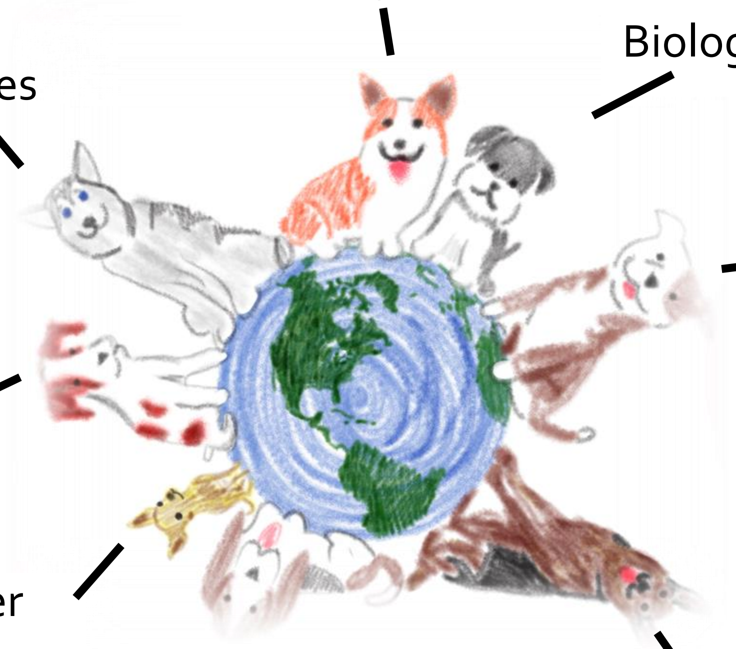


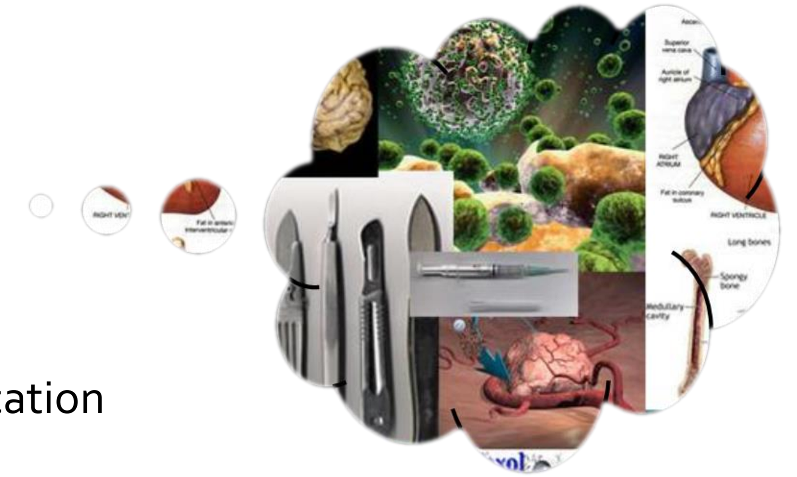
COMPUTATIONAL
THINKING
AT
GOOGLE





Diverse Majors ... with Rich Knowledge





(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?”

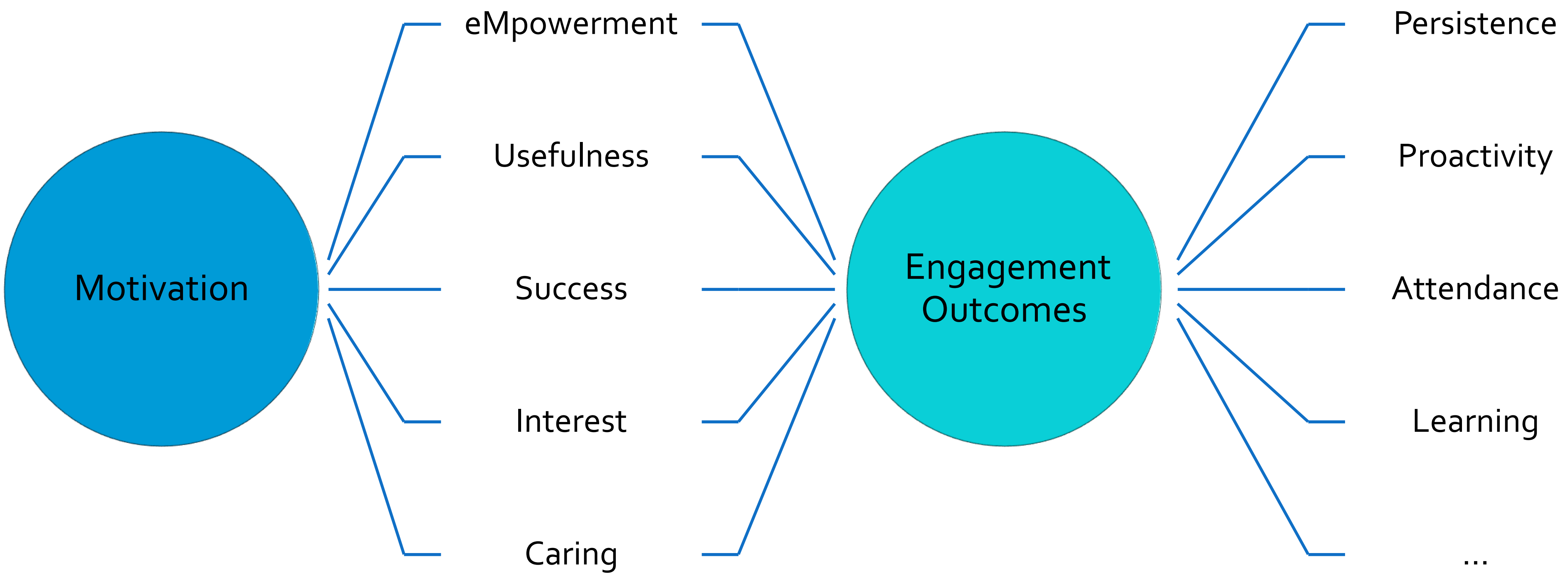
MUSIC Model of Academic Motivation

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

B. D. Jones. Motivating students to engage in learning: The MUSIC model of academic motivation. International Journal of Teaching and Learning in Higher Education, 21(2):272–285, 2009.

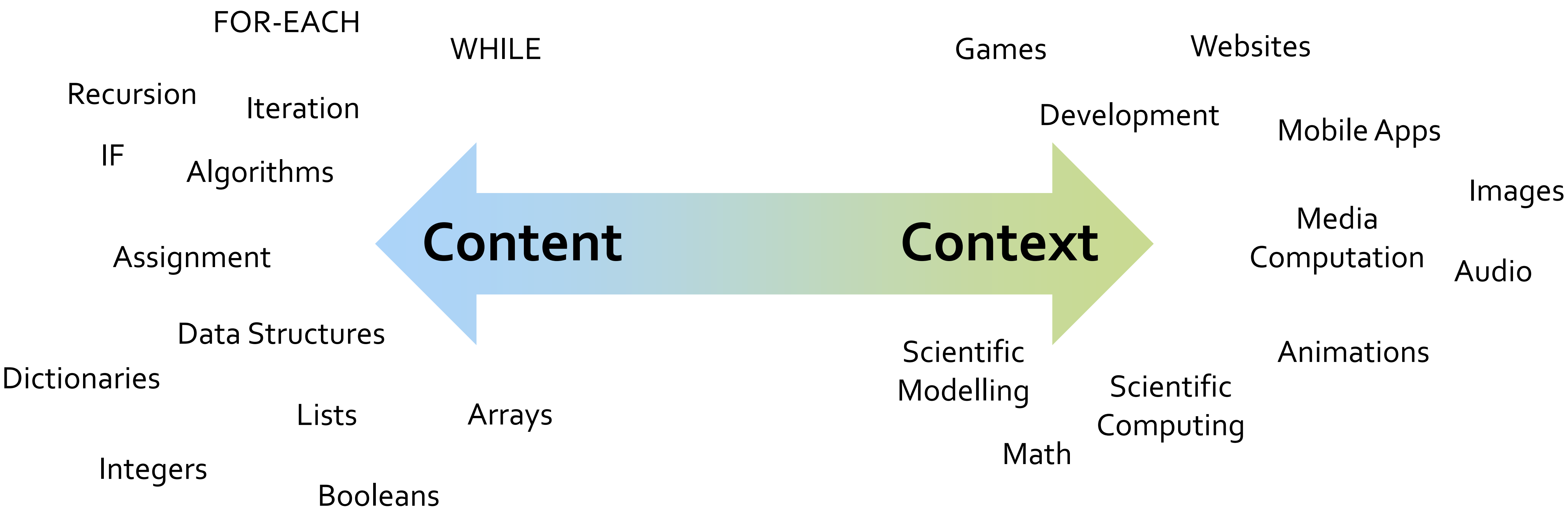
Motivation → Engagement



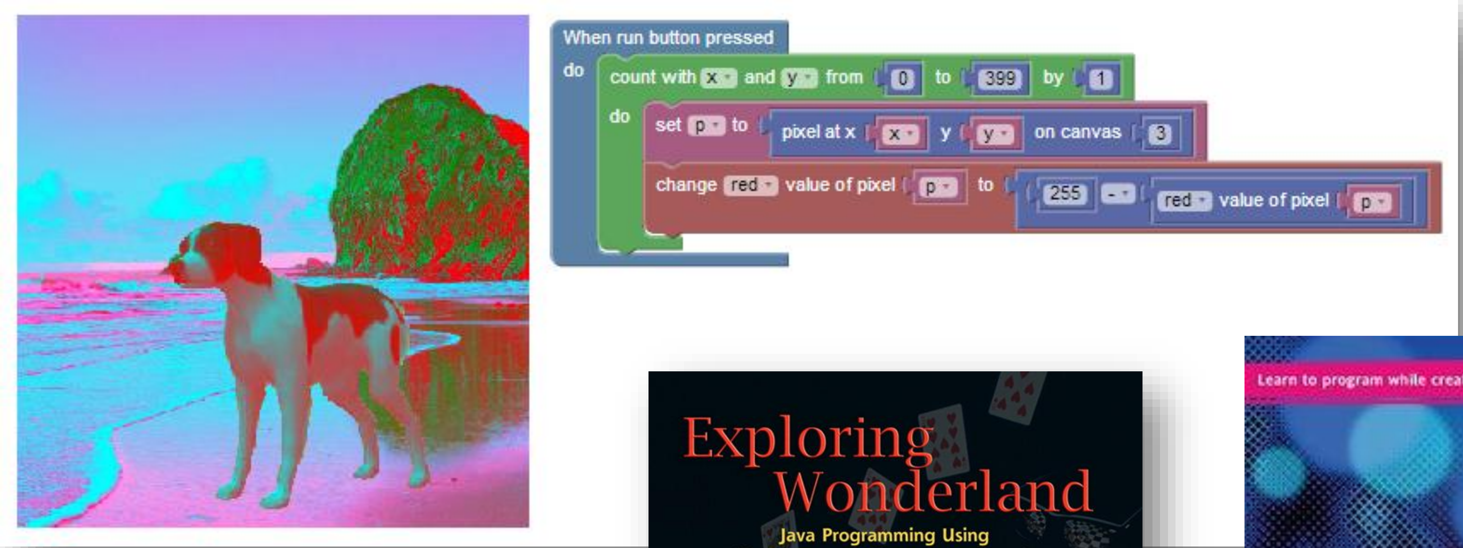
Situated Learning

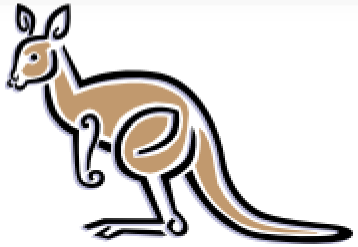
- Lave and Wenger
- “Learning occurs as a function of the activity, context, and culture”

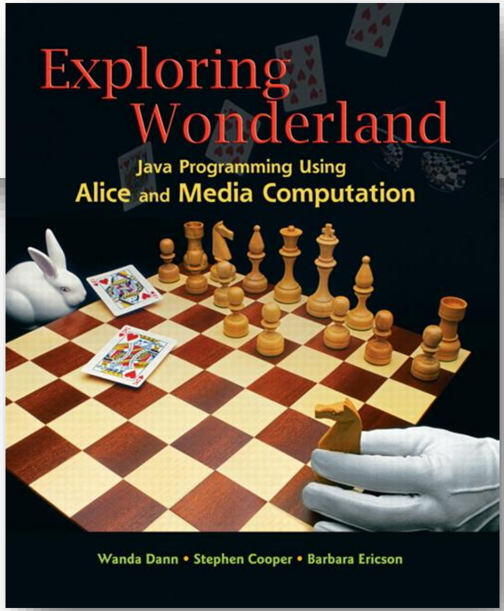
A spectrum

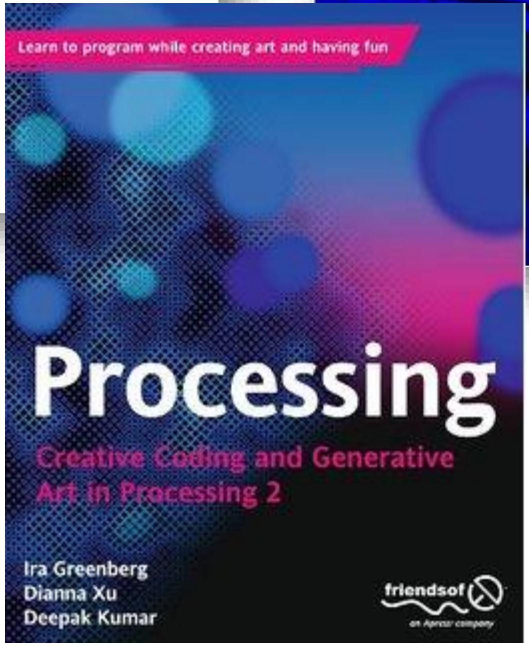

FOR-EACH
WHILE
Recursion
Iteration
IF
Algorithms
Assignment
Data Structures
Dictionaries
Lists
Arrays
Integers
Booleans
Content
Context
Games
Websites
Development
Mobile Apps
Images
Media Computation
Audio
Scientific Modelling
Scientific Computing
Animations
Math

Interesting Contexts







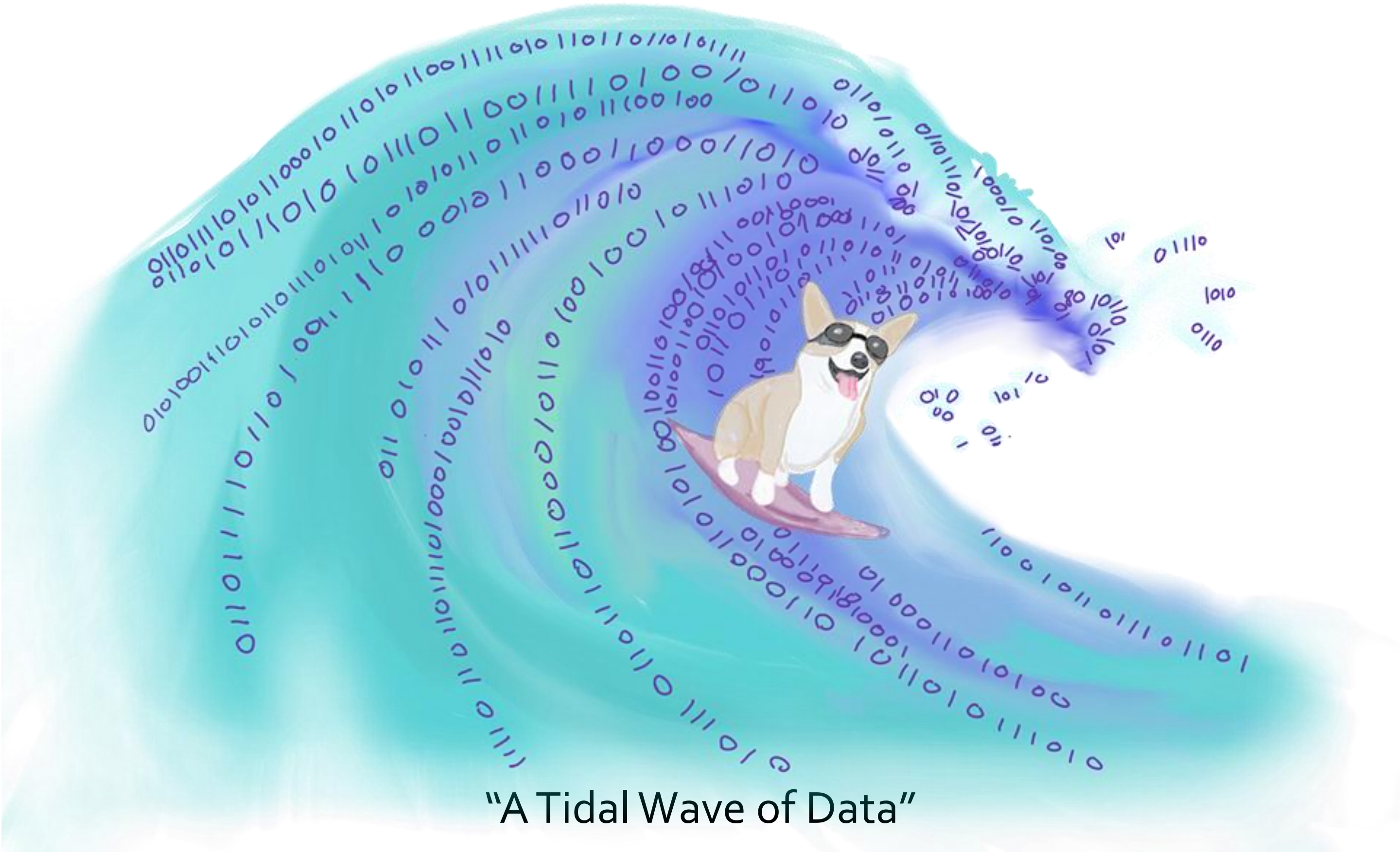


Authenticity

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

**Mark Guzdial and Allison Elliott Tew. 2006. Imagineering inauthentic legitimate peripheral participation: an instructional design approach for motivating computing education. In Proceedings of the second international workshop on Computing education research (ICER '06). New York, NY, USA, 51-58*

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

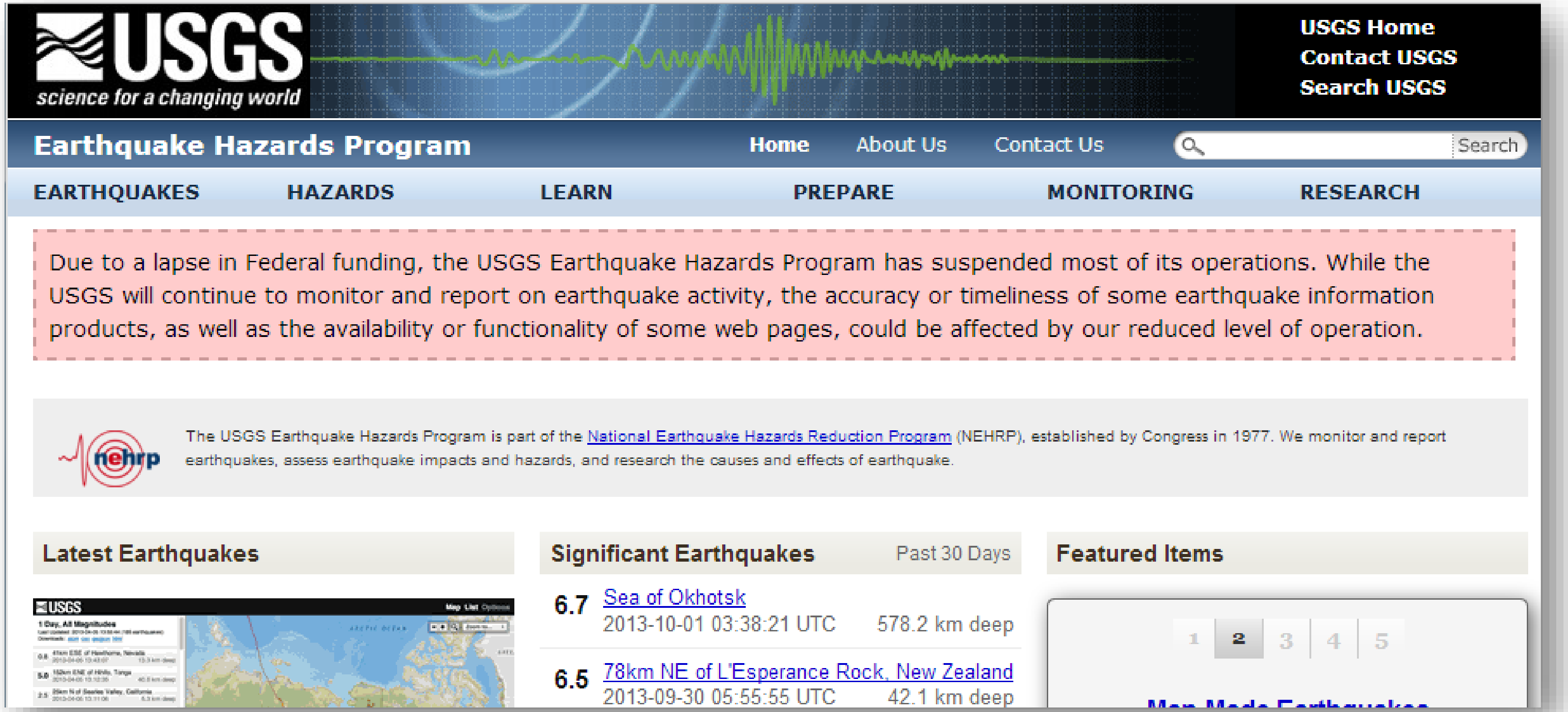
VT Bus Tracking API

Dr. Eli Tilevich

Dr. Cliff Shaffer

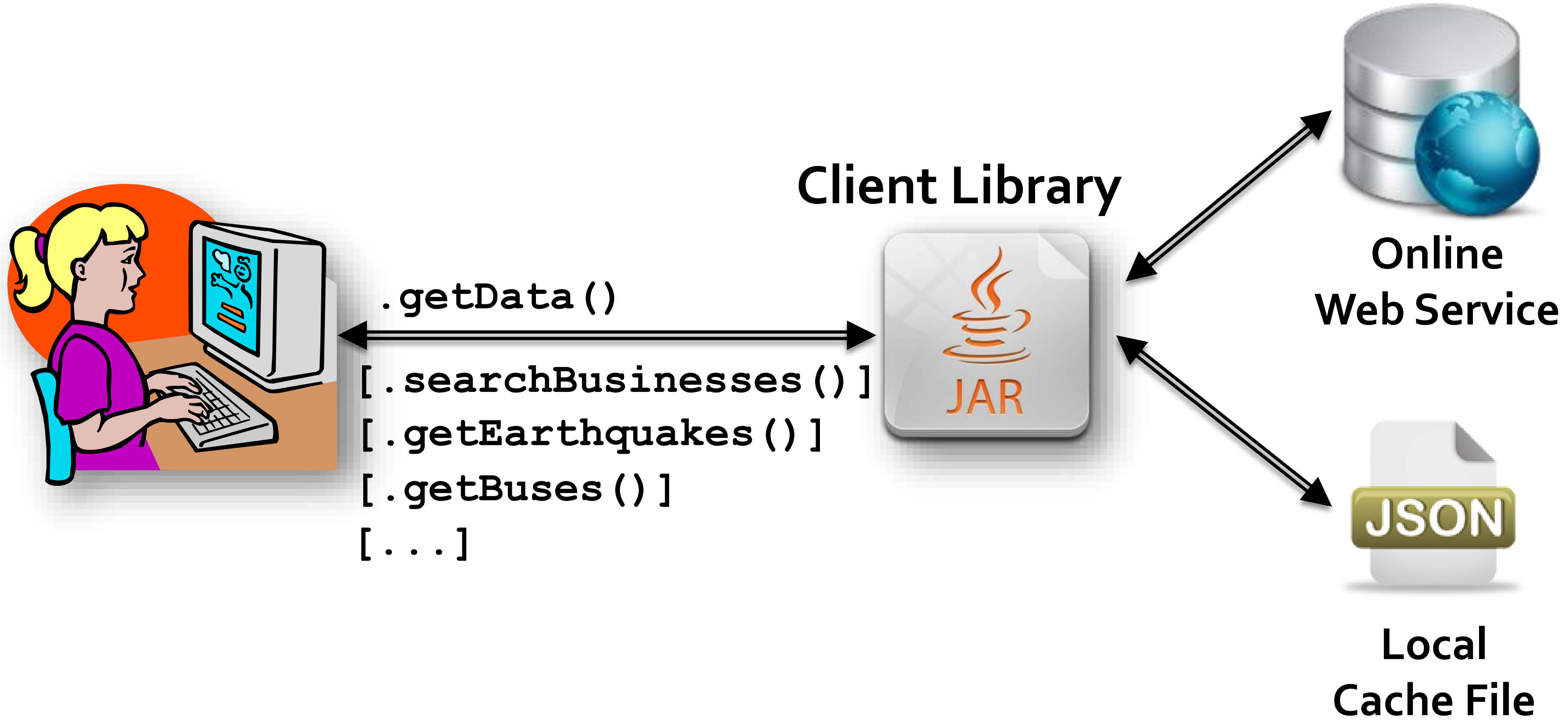
RealTimeWeb – Real-time data

So many Points of Failure!



U.S. Geological Survey, 2013, Earthquakes Hazards Program available on the World Wide Web, accessed [October 7, 2013], at URL [http://earthquake.usgs.gov/].

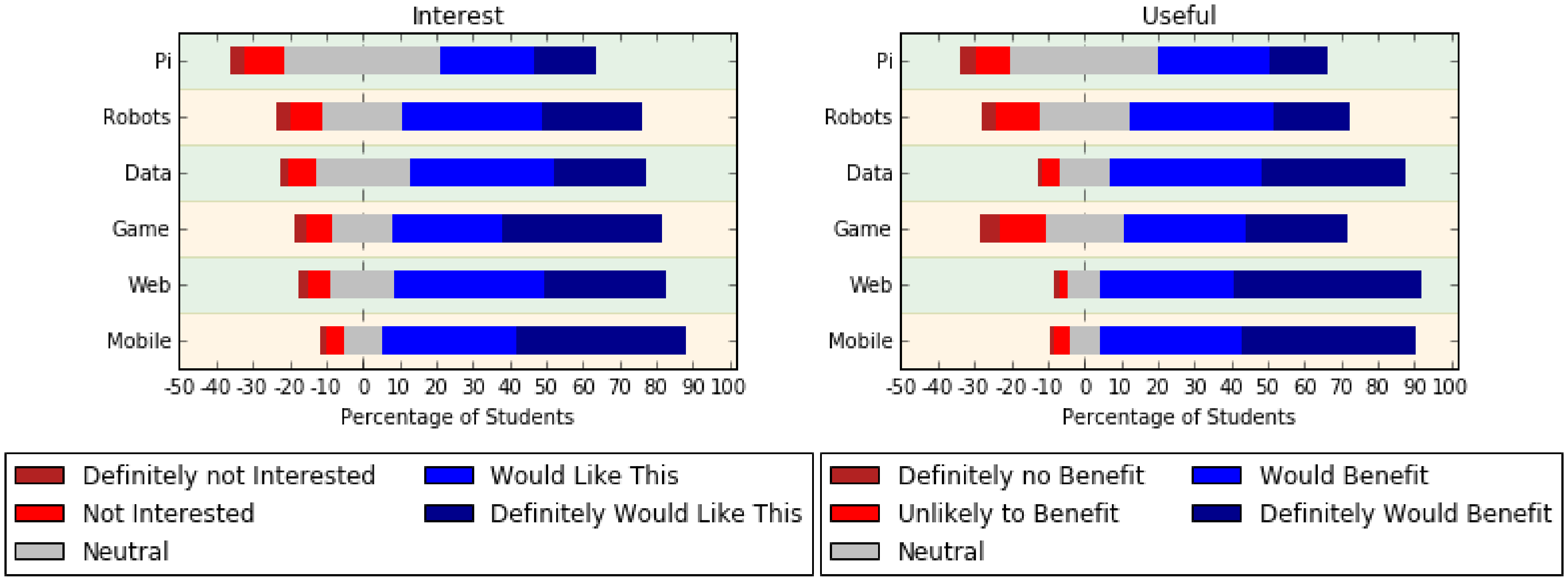
RealTimeWeb – Secret Sauce



RealTimeWeb - Deployment

Semester	School	Course
Spring 2013	Virginia Tech	CS-2
Fall 2013	University of Delaware	CS-1
	Virginia Tech	CS-2
	Virginia Tech	Data Structures & Algos
Spring 2014	Virginia Tech	CS-2

RealTimeWeb - Studies



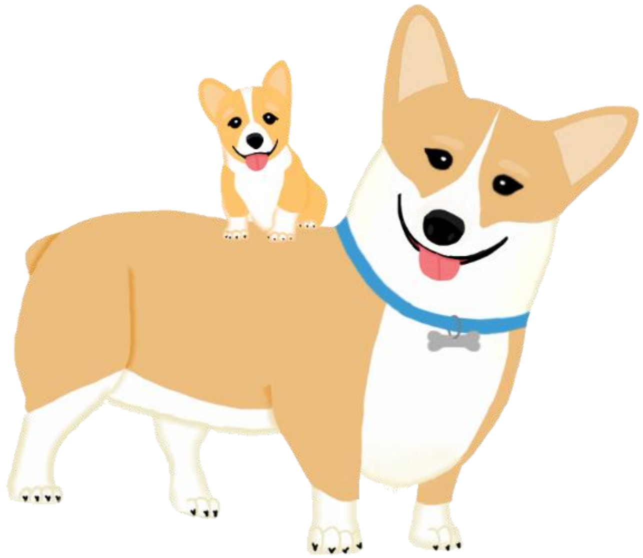
N=370, 14% female
University of Delaware, Virginia Tech
CS1, CS2, and DSA

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


government
health
population
education
buildings
usa
race
america
countries
language
gender
construction
counties
world
death
sales
states
age
food
demographics
national
land area
learning
kindergarten
building
burglary
rural
money
private
households
ethnicity
urban
books
english
job
income
assault
foreign
pain
energy
school
K-12
robbery
hurt
drugs
motor
vehicles
politics
violent
safety
crime
teaching
property
work
art
murder
literature
larceny
unemployment
employment
services
war
legal
nature
wealthy
units
military
legitimacy
institute
renewable
higher education
fossil fuels
warfare
jobs
gun
infection
electricity
religion
global
residential
amazon
development
orphans
longitude
texts
history
majors
kill
disaster
skyscrapers
artwork
court
writing
crude
rate
honor
transportation
shooting
bomb
earthquakes
supreme
chemistry
theatre
effectiveness
waterways
combat
irrigation
votes
outpatient
navy
lakes
labor
election
power
barrister
immigration
medium
musical
drug
nuclear
hiv
latitude
charges
career
damage
marines
substances
sex
supermarket
society
plane
employed
careers
glory
creativity
people
availability
buyer
nutrition
author
magnitude
medicare
grads
international
federal
richter scale
army
security
drug usage
terror
trade
airports
navigability
suffering
media
non-residential
permanants
tower
proteins
style
unemployed
united nations
osha
asian
words
rivers
retail
songs
white
healthcare
publication
biochemistry
aquaculture
artists
broadway
cancer
permits
housing
airplane
law
novel
flights
hazard
terrorism
graduation
sculpture
injury
slavery
judge
height
census
tate
biology
delays
text
fighting
vitamins
tickets
estate
award
economy
slave
attack
civilian
hospitals
person
dams
patient
billionaire
publishers
rich
travel
university
spending
book
finance
composition
aids
industry
lawyer
infrastructure
doctor
expenditure
valuation
genres
candidate
black
diseases
medal
employability
college
minerals
shakespeare
fine art
artist
classic
cars
origin
naturalization
destroy
car
airforce
economics
public
seller
disease
air
tall
injuries
case
access
occupation
agriculture
drugbank
human consumption
slaves
music
hydropower
industrial use
real
fuel
sat
classics

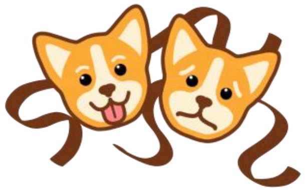
Connecting to Students' Majors

Crime

Criminal Justice

Airlines

Aerospace

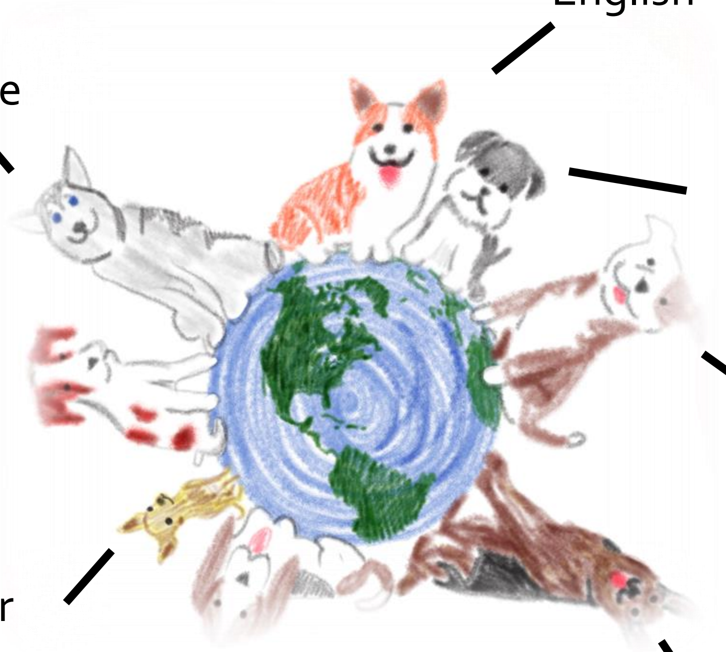


Theater

Theater Arts

Construction

Building Construction



English

Books

Geological Science

Weather

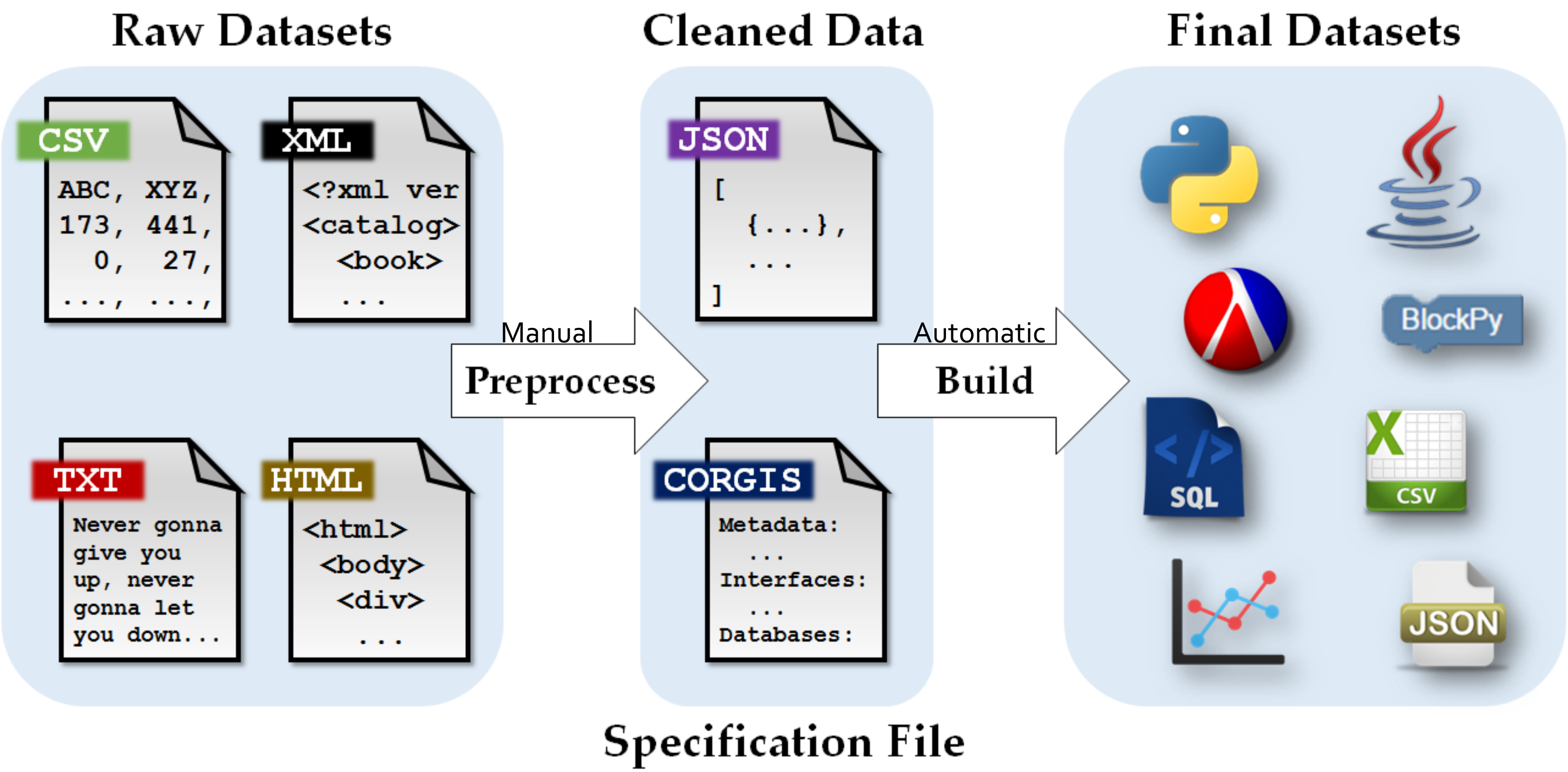
Education

Education

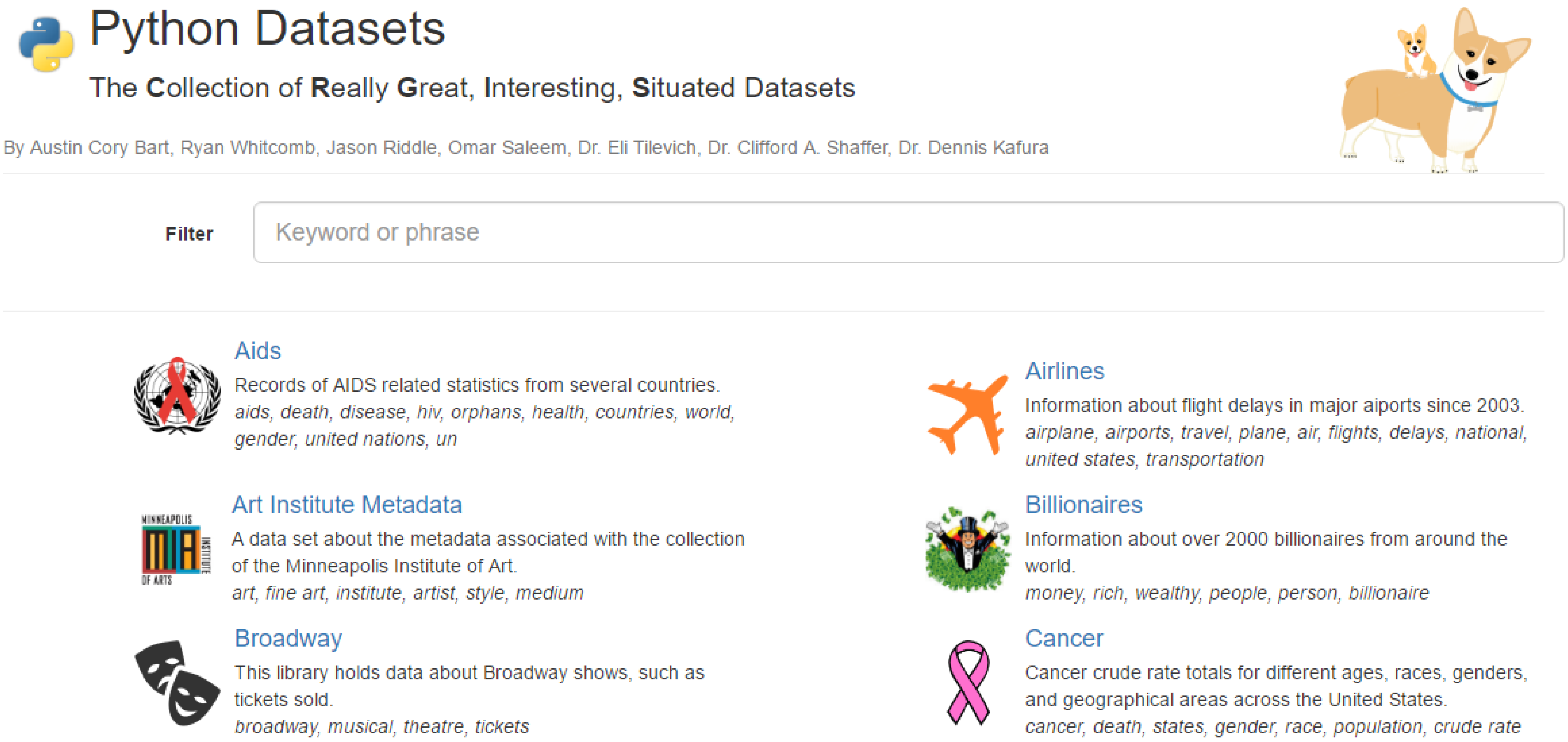
History

Immigration

Architecture



Gallery


Python 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
Keyword or phrase
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 aiports since 2003.
airplane, airports, travel, plane, air, flights, delays, national, united states, transportation
MINNEAPOLIS INSTITUTE OF ARTS
Art Institute Metadata
A data set about the metadata associated with the collection of the Minneapolis Institute of Art.
art, fine art, institute, artist, style, medium
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
Cancer
Cancer crude rate totals for different ages, races, genders, and geographical areas across the United States.
cancer, death, states, gender, race, population, crude rate

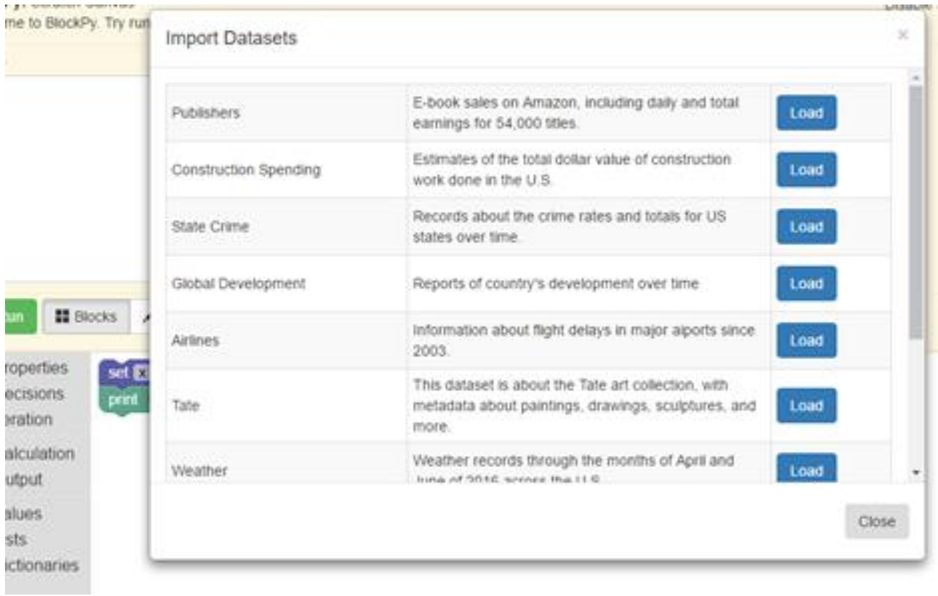
Java, Python, Racket

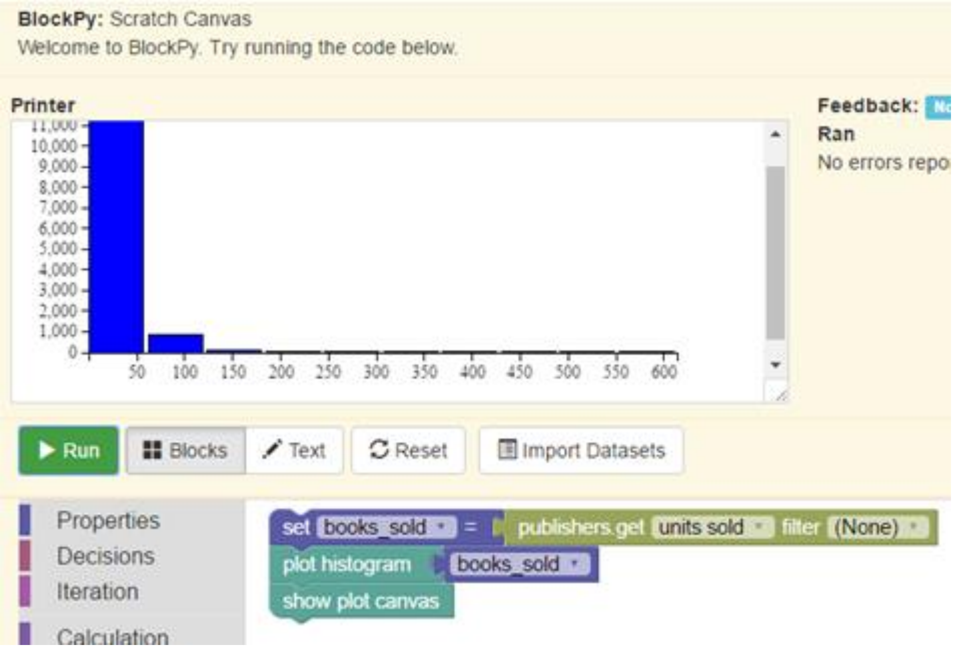
```
// Java
import corgis.crime.StateCrimeLibrary;
import corgis.crime.domain.Report;
import java.util.ArrayList;
public class Main {
        public static void main(String[] args) {
                StateCrimeLibrary scl = new StateCrimeLibrary();
                ArrayList<Report> reports = scl.getAll();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

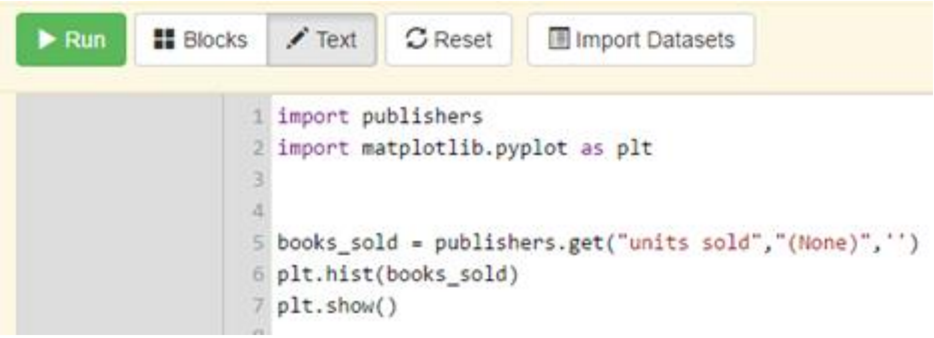
```
; Racket
(require crime)
(define reports (crime-get-all))
```

```
# Python
import crime
crime_reports = crime.get_all()
```

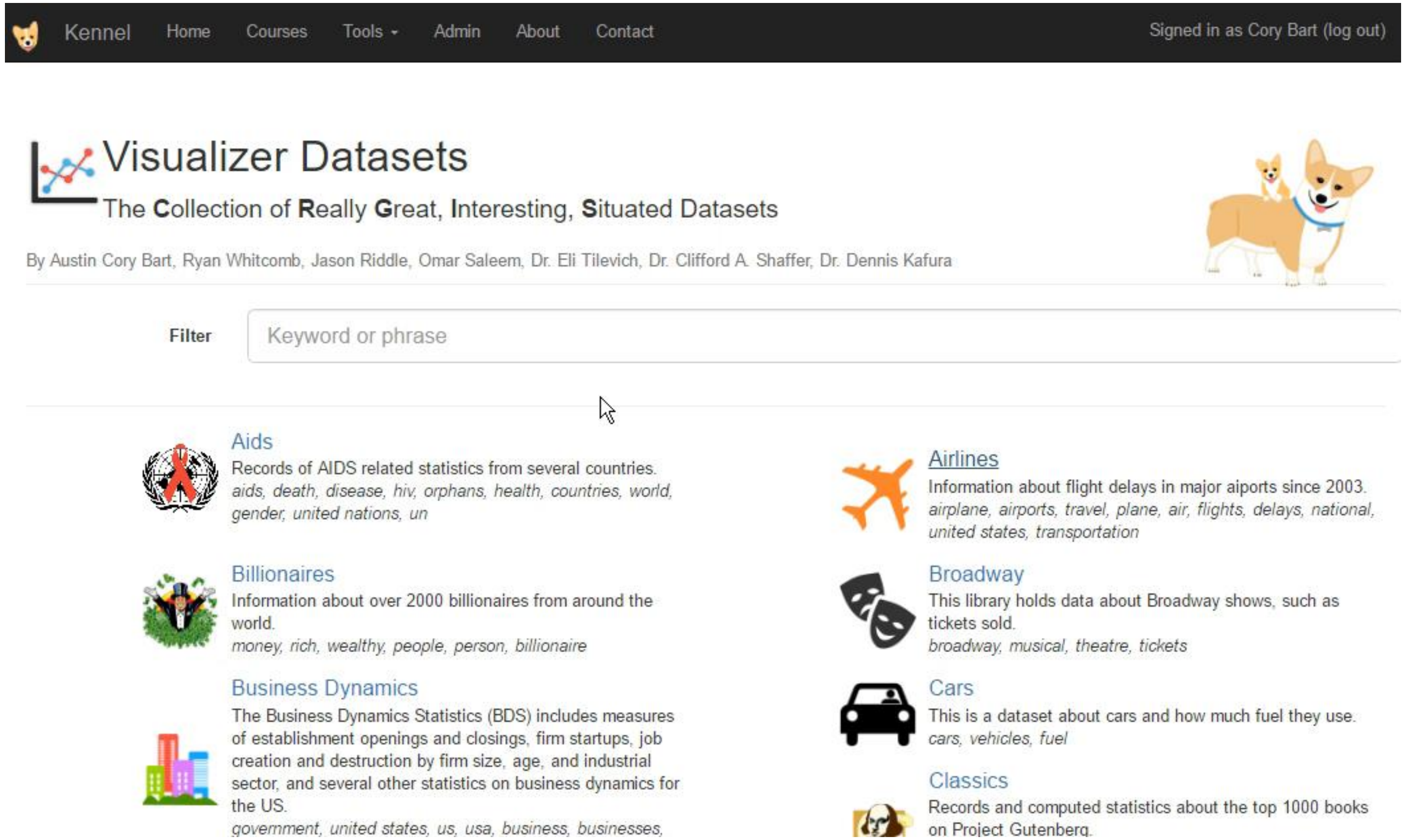
BlockPy


Import Datasets
Publishers
E-book sales on Amazon, including daily and total earnings for 54,000 titles.
Construction Spending
Estimates of the total dollar value of construction work done in the U.S.
State Crime
Records about the crime rates and totals for US states over time.
Global Development
Reports of country's development over time
Airlines
Information about flight delays in major aiports since 2003.
Tate
This dataset is about the Tate art collection, with metadata about paintings, drawings, sculptures, and more.
Weather
Weather records through the months of April and
Load
Close
Blocks


BlockPy: Scratch Canvas
Welcome to BlockPy. Try running the code below.
Printer
Feedback:
Ran
Run
Blocks
Text
Reset
Import Datasets
Properties
Decisions
Iteration
Calculation
set books_sold = publishers.get units sold filter (None)
plot histogram books_sold
show plot canvas


Run
Blocks
Text
Reset
Import Datasets
import publishers
import matplotlib.pyplot as plt
books_sold = publishers.get("units sold","(None)",'')
plt.hist(books_sold)
plt.show()

Visualizer Demo

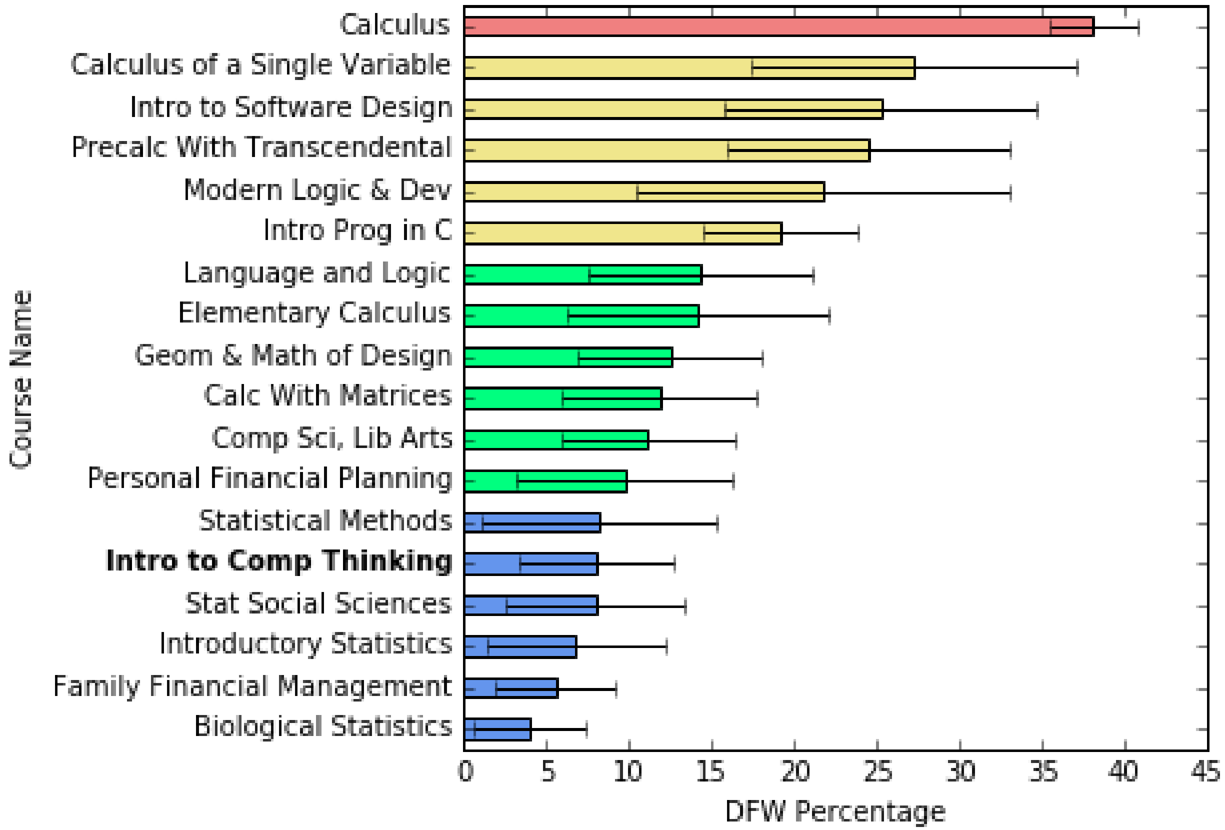


Interventions

- Computational Thinking Course
  - ❖Basic programming
  - ❖Social Impacts
  - ❖Data Science
- 6 semesters taught
- Audience
  - ❖Non-computing majors
  - ❖Freshmen -> Senior
  - ❖Gender balanced

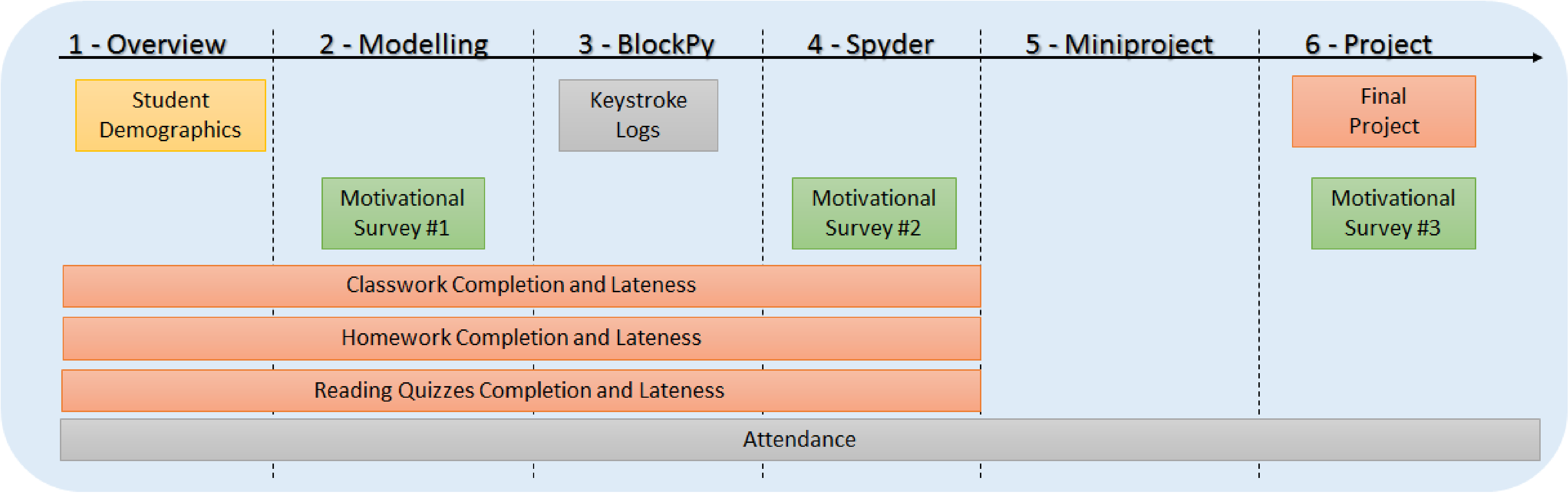
Course Evaluation

- **Retention**
- More-Computing
- Gender
- Learning



Mark Guzdial. 2013. Exploring hypotheses about media computation. In Proceedings of the ninth annual international ACM conference on International computing education research (ICER '13).

Survey Timeline



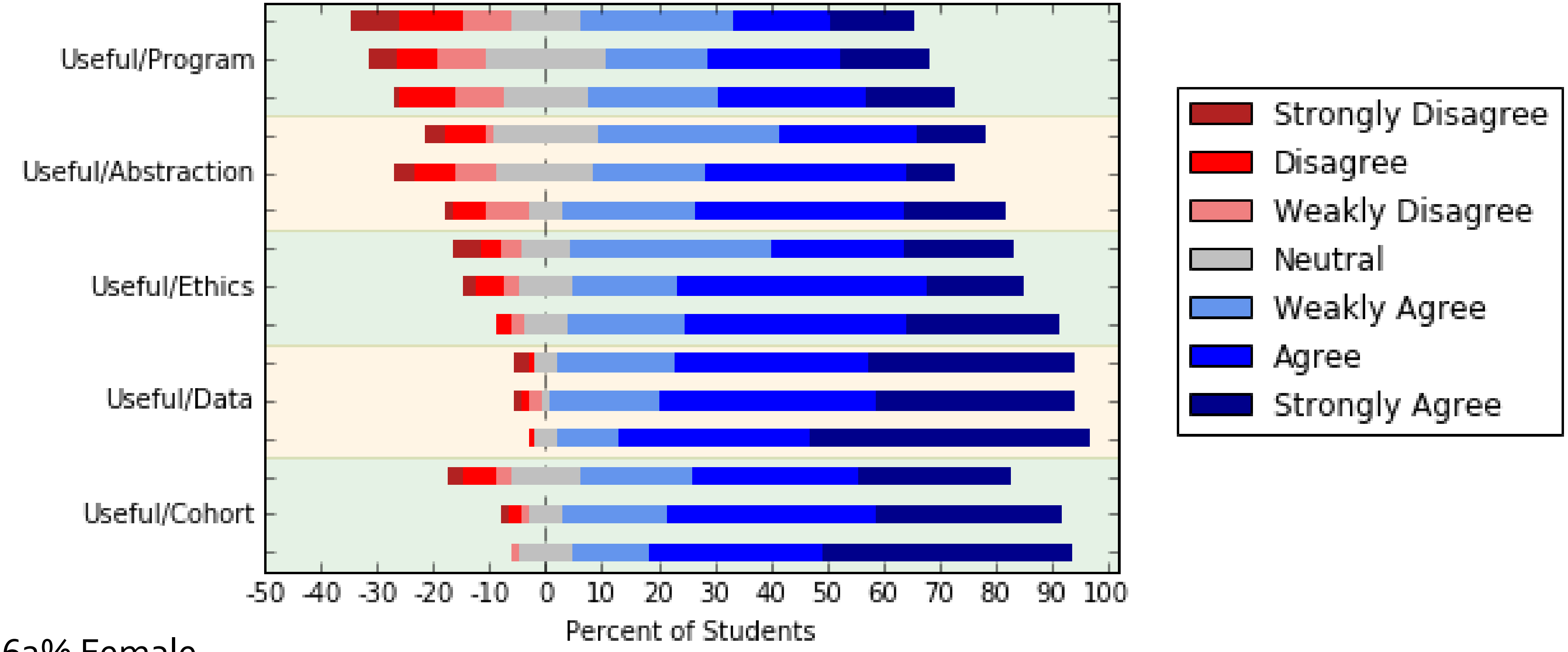
Motivation × Course Components

Motivational Components	
"I believe that I will have freedom to explore my own interests when I…"	eMpowerment
"I believe it will be useful to my long-term career goals to…"	Usefulness
"I believe I will be successful in this course when I…"	Success
"I believe it will be interesting to…"	Interest
"I believe that my instuctors and peers will care about me when I…"	Caring

Course Component	
"… learn to write computer programs"	Programming Content
"… learn to work with abstraction"	Abstraction Content
"… learn about the social impacts of computing"	Social Ethics Content
"… work with real-world data related to my major"	Data Science Context
"… work with my cohort"	Collaboration Facilitation

Likert
Strongly Disagree
Disagree
Somewhat Disagree
Neither Agree nor Disagree
Somewhat Agree
Agree
Strongly Agree

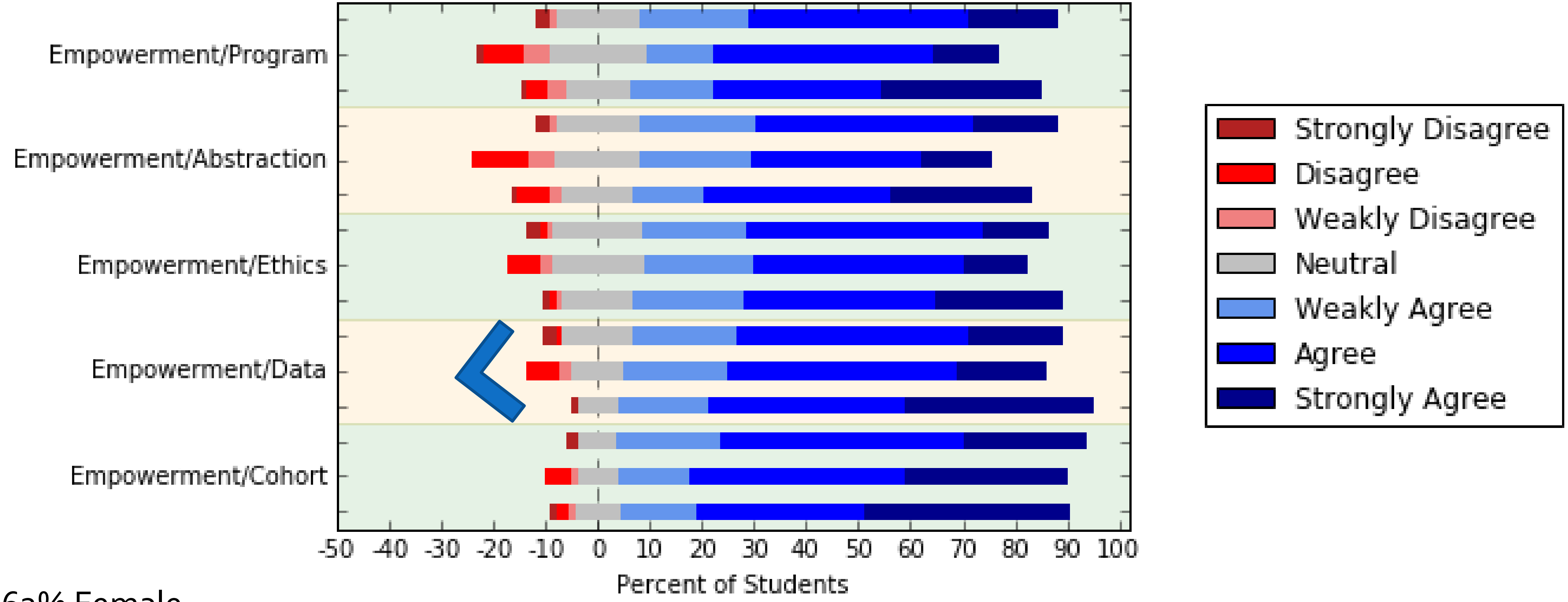
Context is Useful



N = 85, 62% Female

Students' sense of the usefulness of various course components was highest for the **context**, lowest for the **content**.

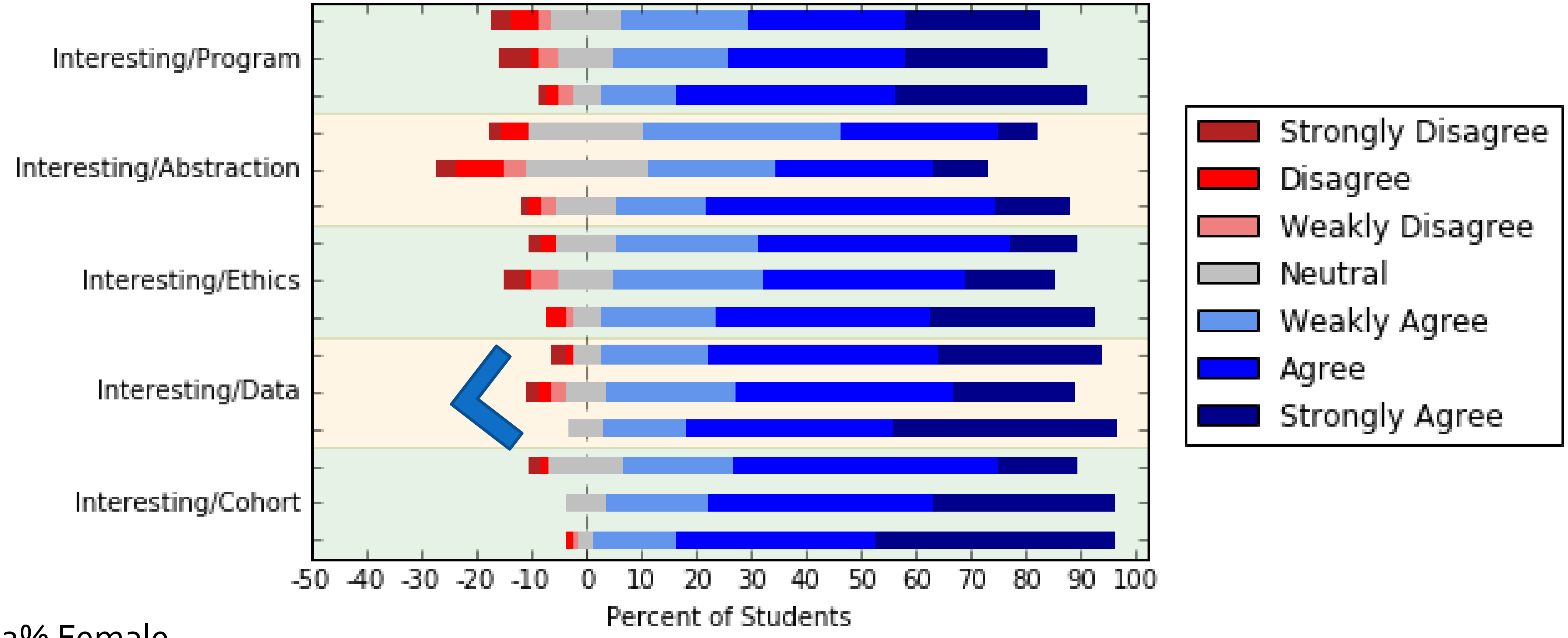
V-Shaped Empowerment



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

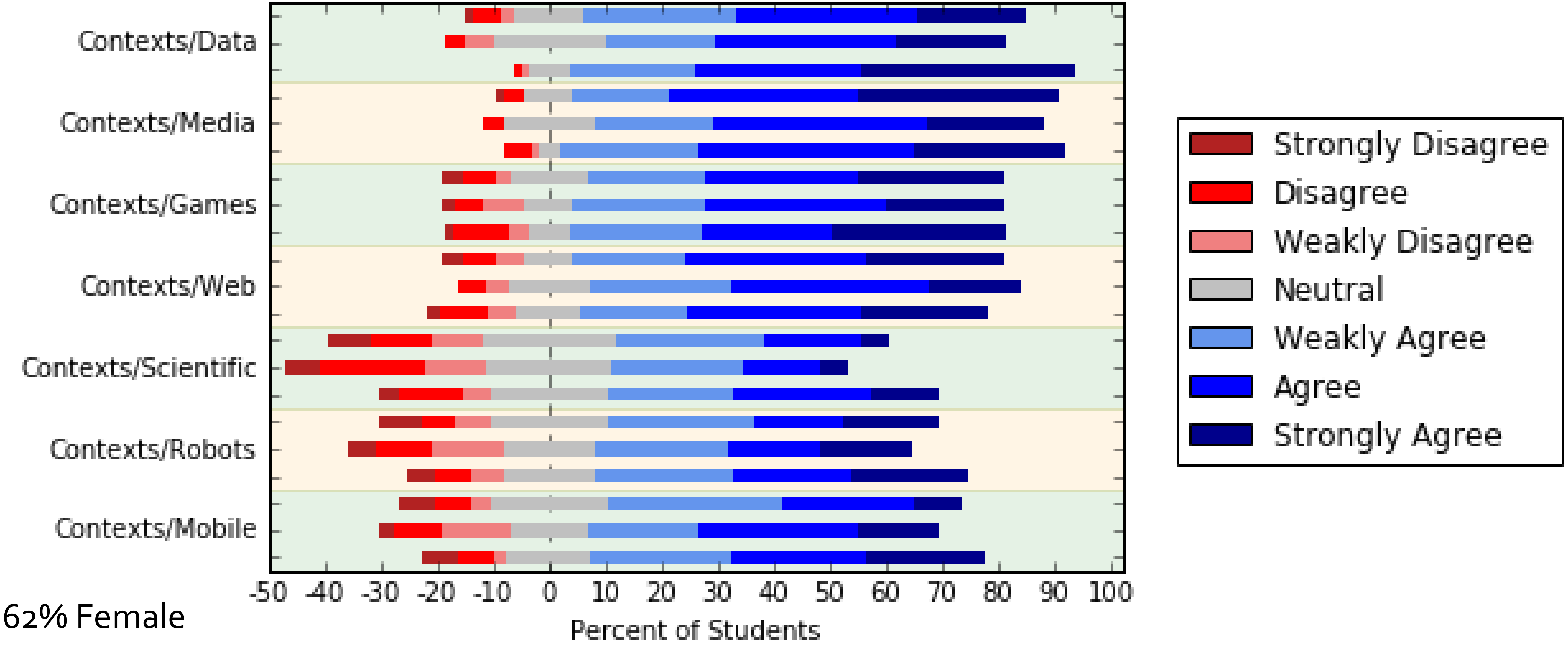
Students' interest decreases during the BlockPy and Spyder portions of the course, then increases during the final projects.

Preference for Contexts

Preference for Contexts	
"Working with data sets related to your major"	Data
"Working with pictures, sounds, movies"	Media
"Making games and animations"	Games
"Making websites"	Web
"Making scientific models of real-world phenomenon"	Scientific
"Controlling robots or drones"	Robots
"Making phone apps"	Mobile

Likert
Strongly Avoid
Avoid
Somewhat Avoid
Neither Prefer nor Avoid
Somewhat Prefer
Prefer
Strongly Prefer

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.

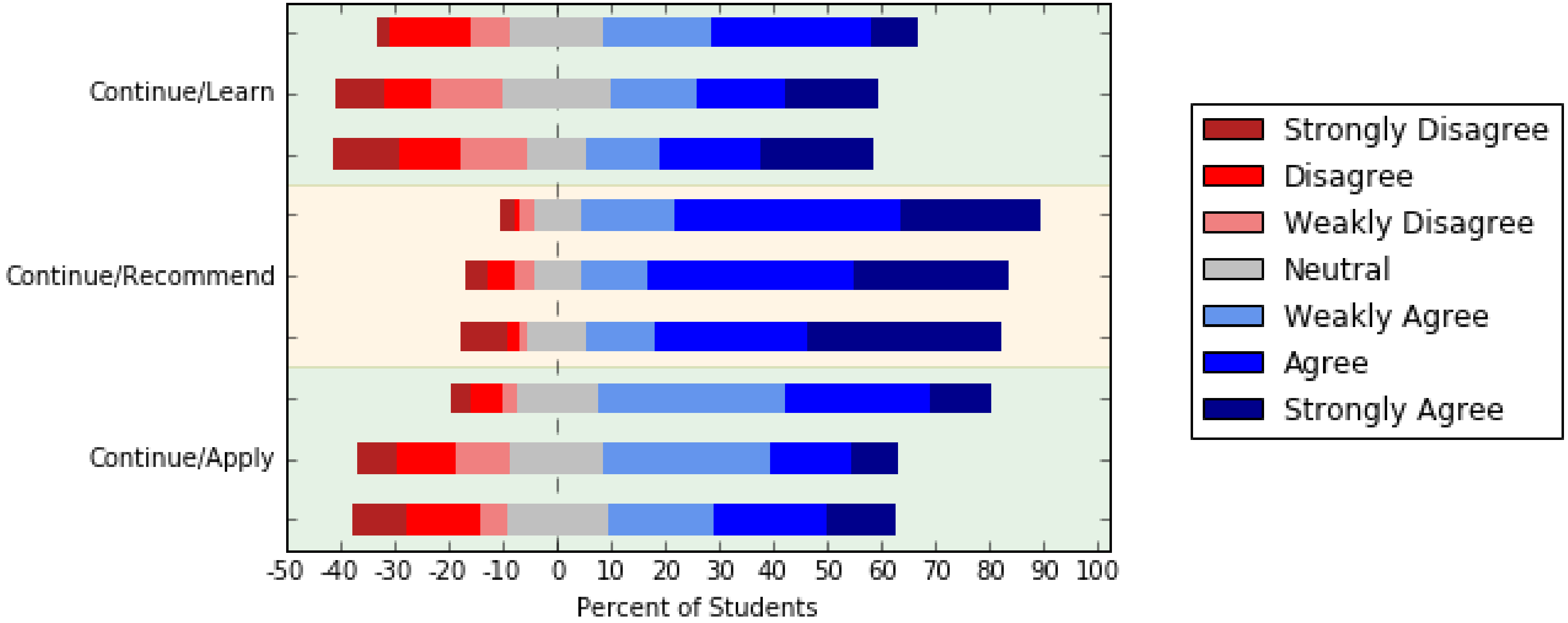
** No significant difference with Media Computation in S3, according to matched-pairs T-test*

Engagement (Intent to Continue)

Intent to Continue	
"I will try to learn more about computing, either through a course or on my own."	Learn
"I will recommend this class to others."	Recommend
"I will directly apply what I have learned in my career."	Apply

Likert
Strongly Disagree
Disagree
Somewhat Disagree
Neither Agree nor Disagree
Somewhat Agree
Agree
Strongly Agree

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

Fall 2016	eMpowerment	Usefulness	Success	Interest	Caring
Abstraction	.087	.276	.184	.124	.288
Cohort	-.011	.064	.046	.001	.152
Data	-.046	.088	.019	.115	.134
Ethics	.025	.203	.196	.082	.255
Programming	.166	**.406**	**.354**	**.341**	.257

Not significantly Correlated!

Significant

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

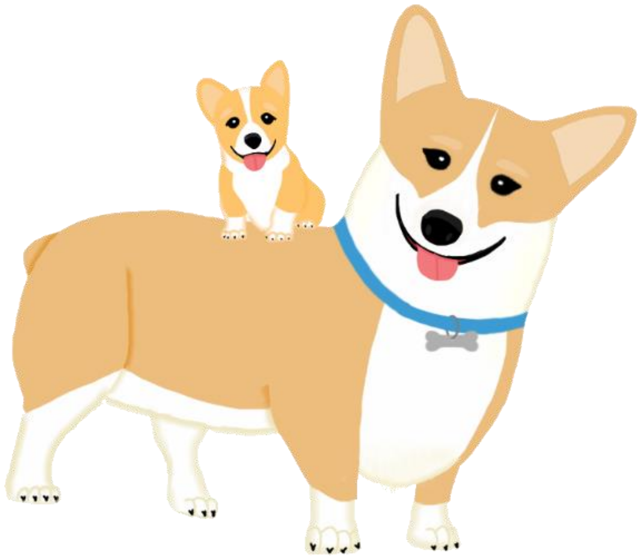
- Data Science seems to be a preferable context for students, across genders, by the end of the course
- The format of the final project was an important motivating factor
- Context, and in particular Data Science, can seem to provide motivation in ways that content cannot
- But some engagement outcomes 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

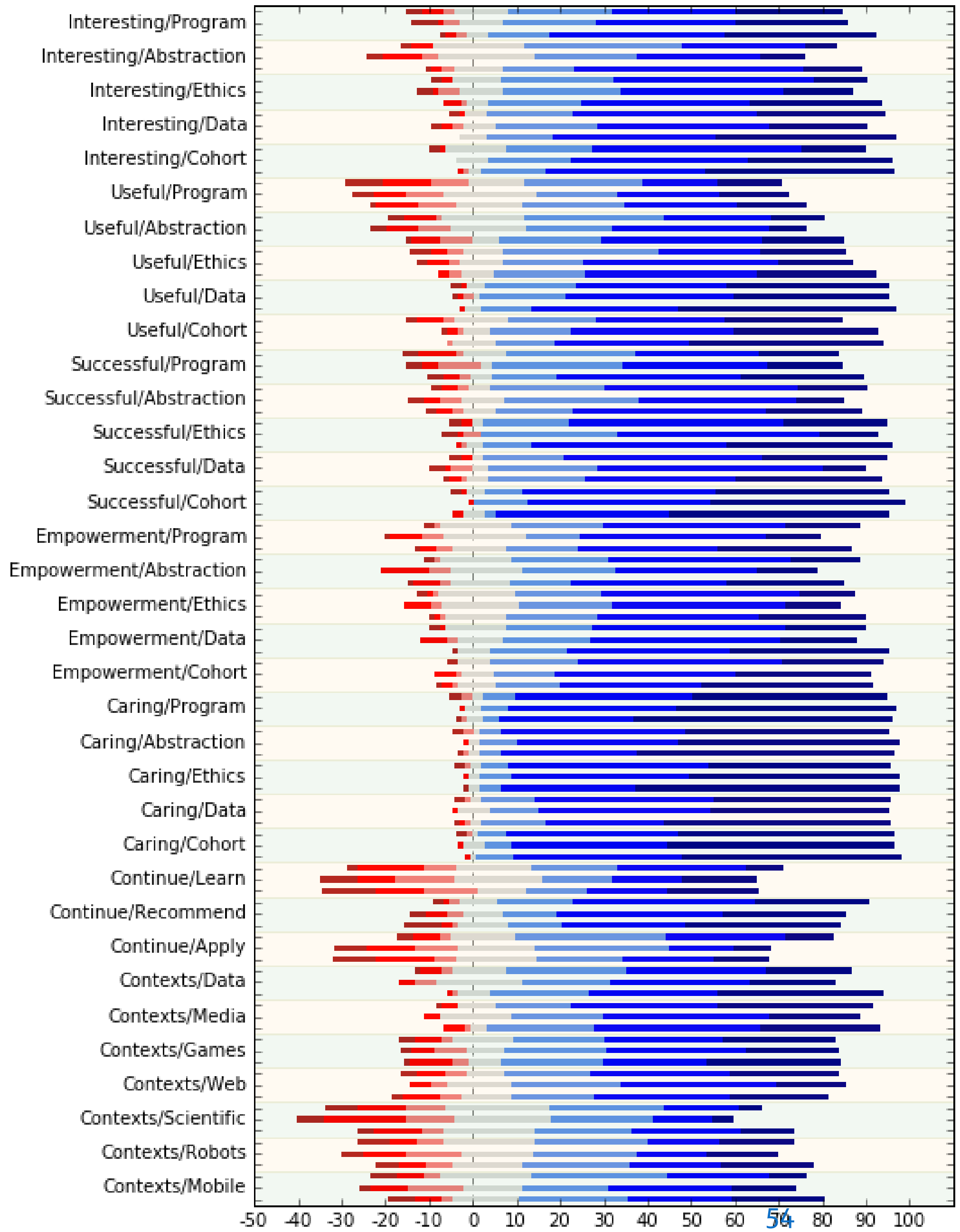
Questions?

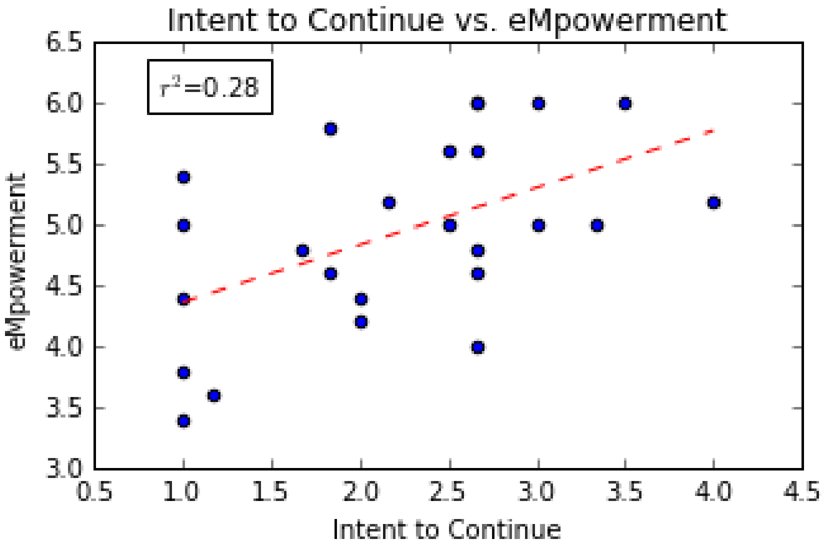
https://think.cs.vt.edu/corgis

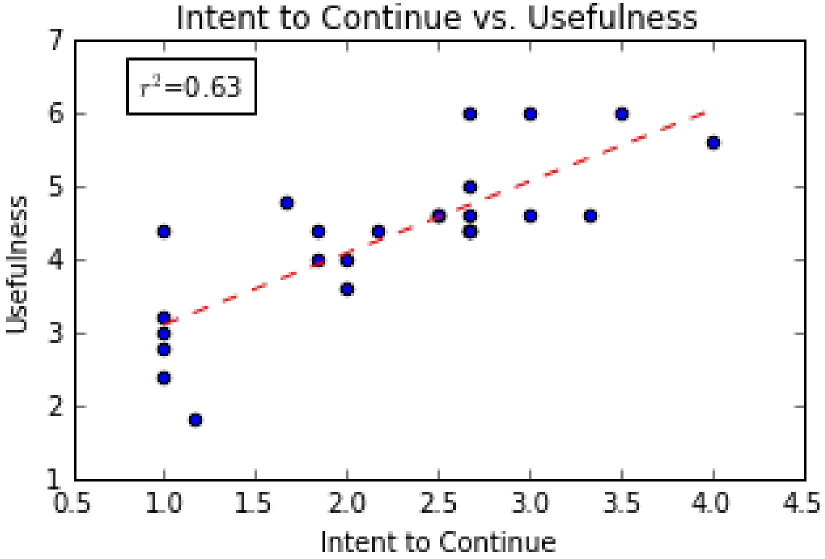


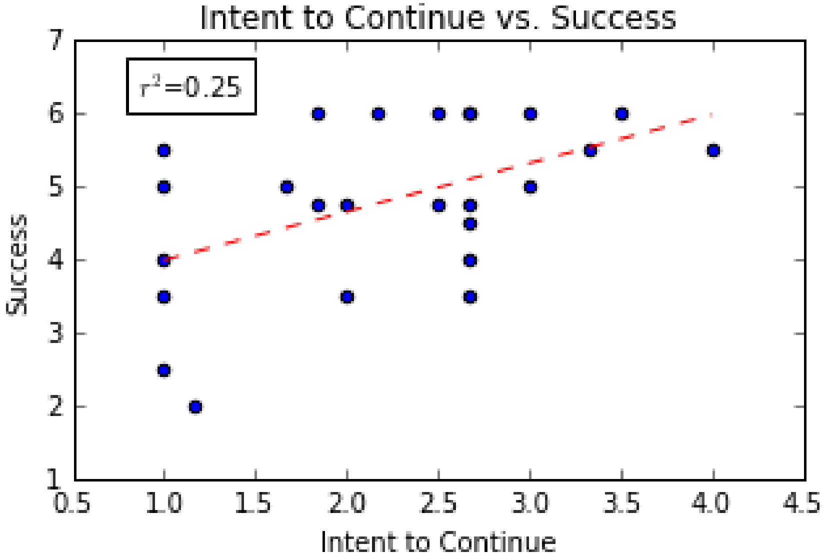
Artwork by Eleonor Bart

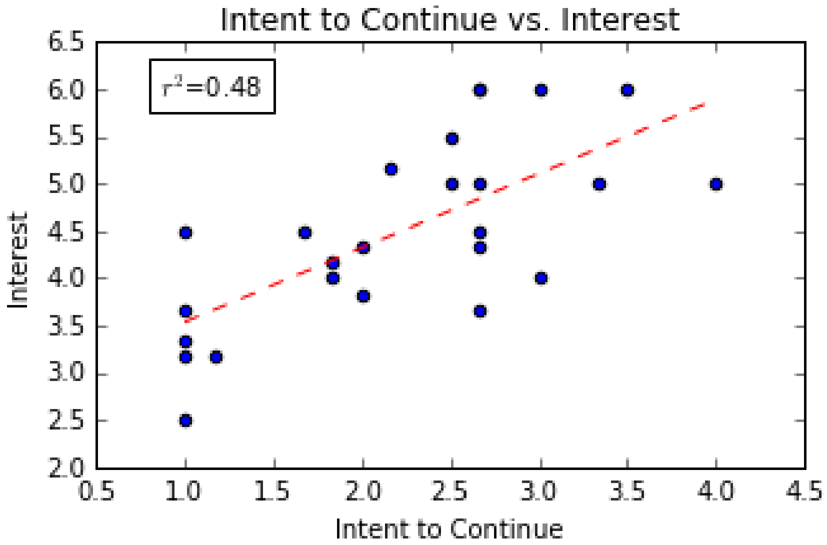
Trends in Motivation

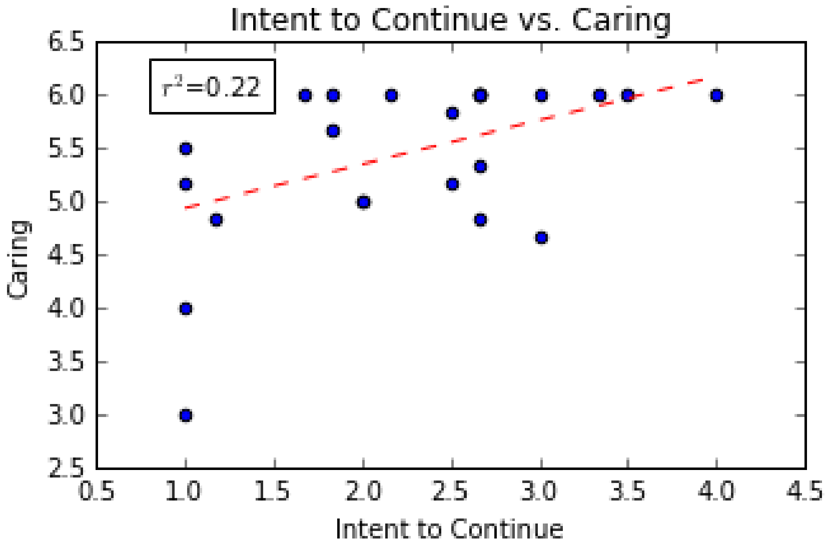


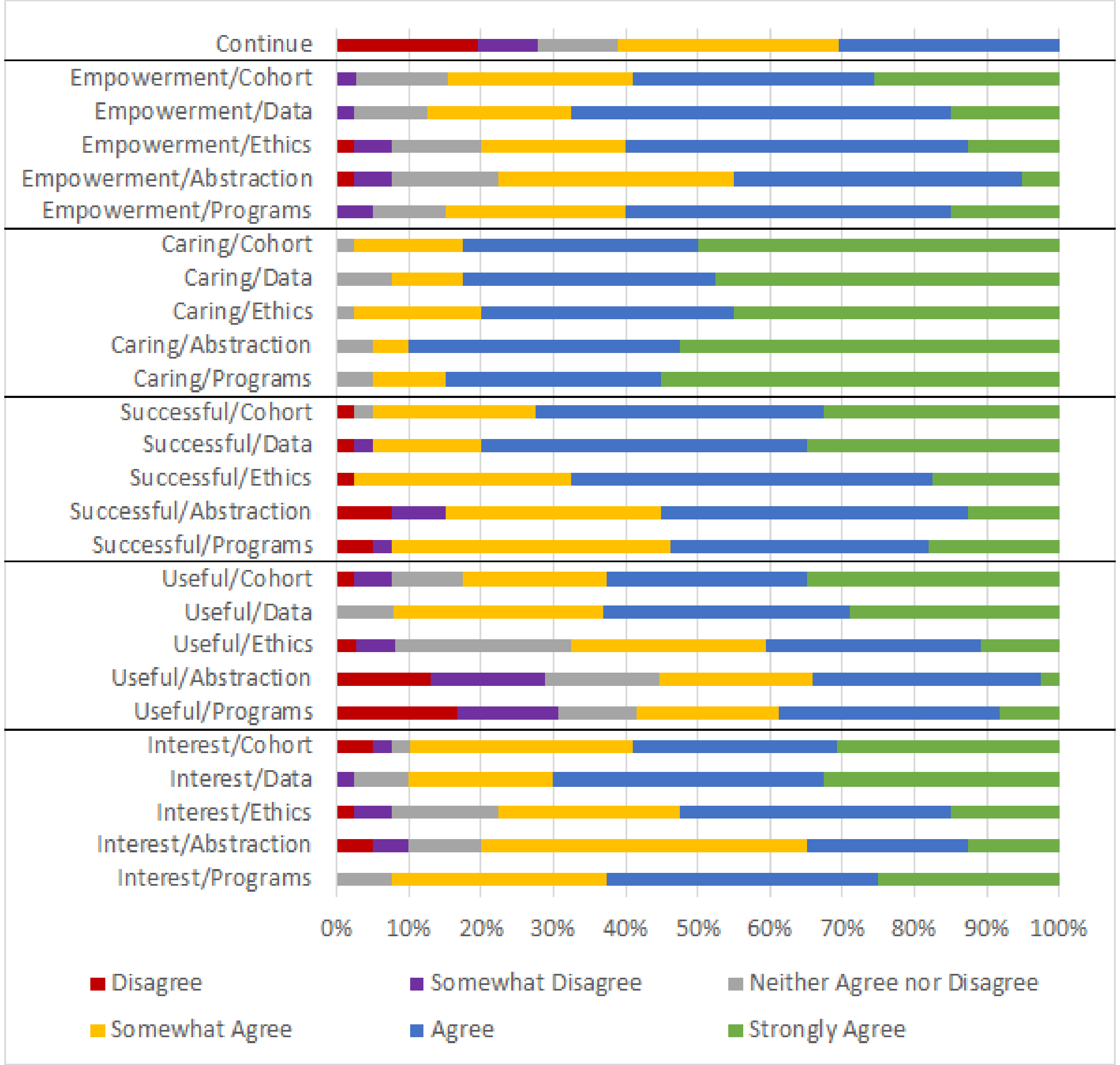

Intent to Continue vs. eMpowerment
r²=0.28
eMpowerment
Intent to Continue


Intent to Continue vs. Usefulness
r²=0.63
Usefulness
Intent to Continue


Intent to Continue vs. Success
r²=0.25
Success
Intent to Continue


Intent to Continue vs. Interest
r²=0.48
Interest
Intent to Continue


Intent to Continue vs. Caring
r²=0.22
Caring
Intent to Continue


Continue
Empowerment/Cohort
Empowerment/Data
Empowerment/Ethics
Empowerment/Abstraction
Empowerment/Programs
Caring/Cohort
Caring/Data
Caring/Ethics
Caring/Abstraction
Caring/Programs
Successful/Cohort
Successful/Data
Successful/Ethics
Successful/Abstraction
Successful/Programs
Useful/Cohort
Useful/Data
Useful/Ethics
Useful/Abstraction
Useful/Programs
Interest/Cohort
Interest/Data
Interest/Ethics
Interest/Abstraction
Interest/Programs
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Disagree
Somewhat Disagree
Neither Agree nor Disagree
Somewhat Agree
Agree
Strongly Agree

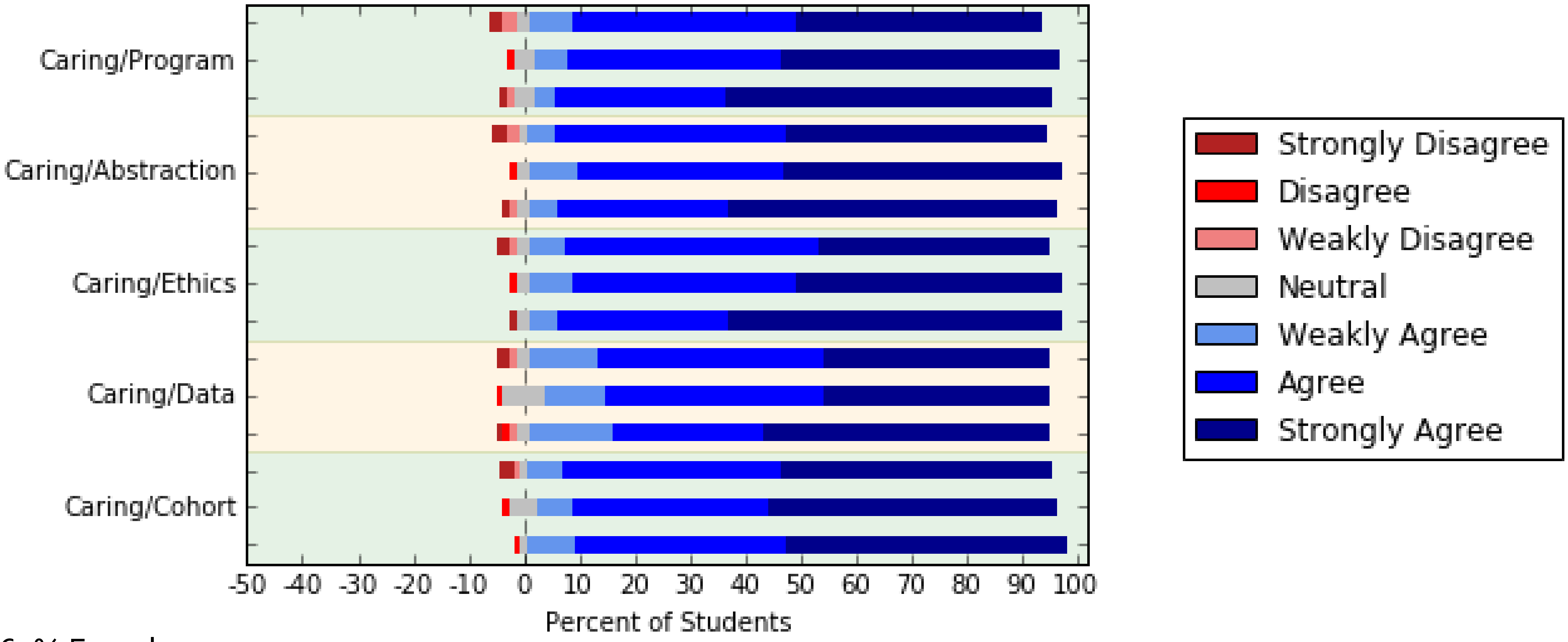
Spring 2016	eMpowerment	Usefulness	Success	Interest	Caring
Abstraction	.458	.699	.614	.488	
Cohort					
Data					
Ethics		.485	.418	.323	
Programming	.437	**.823**	.600	.638	

Continue Learning, Applying, and/or Recommend Course

N =36

50% female

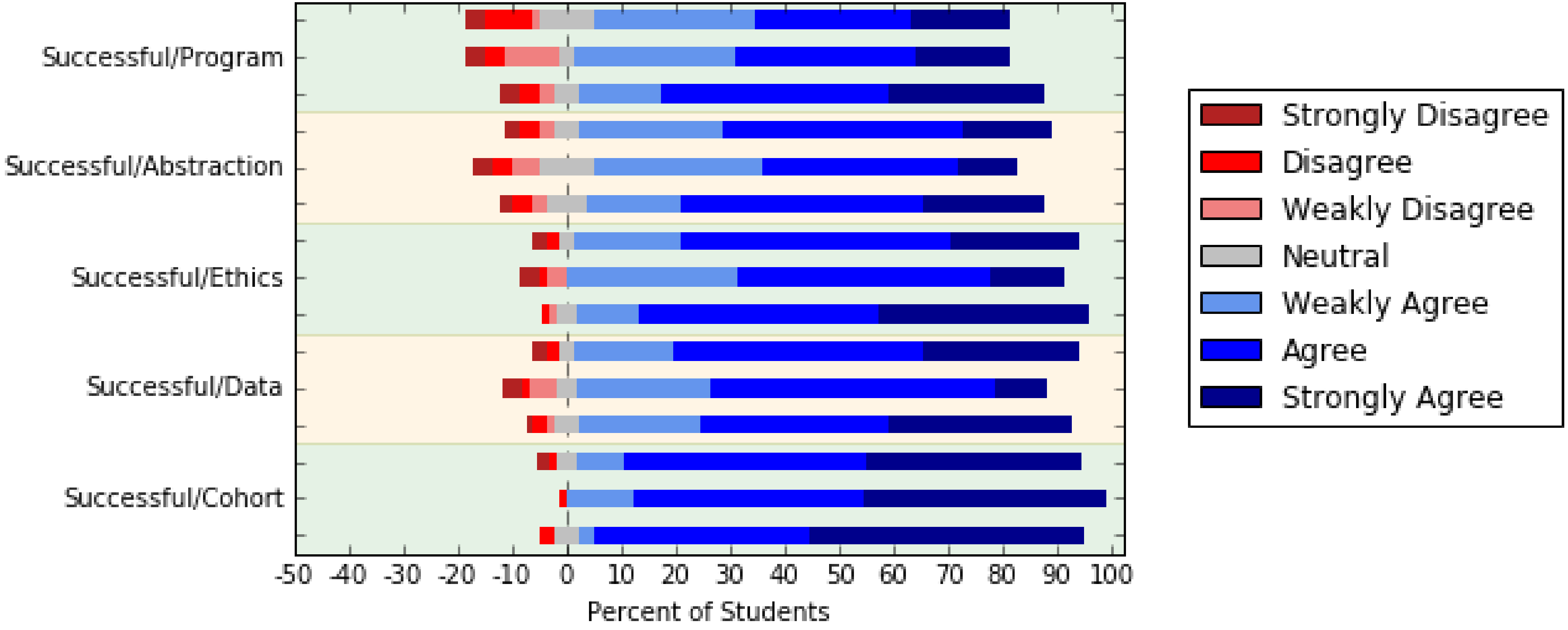
Students’ Perception of Caring



N = 85, 62% Female

We seem to be good instructors

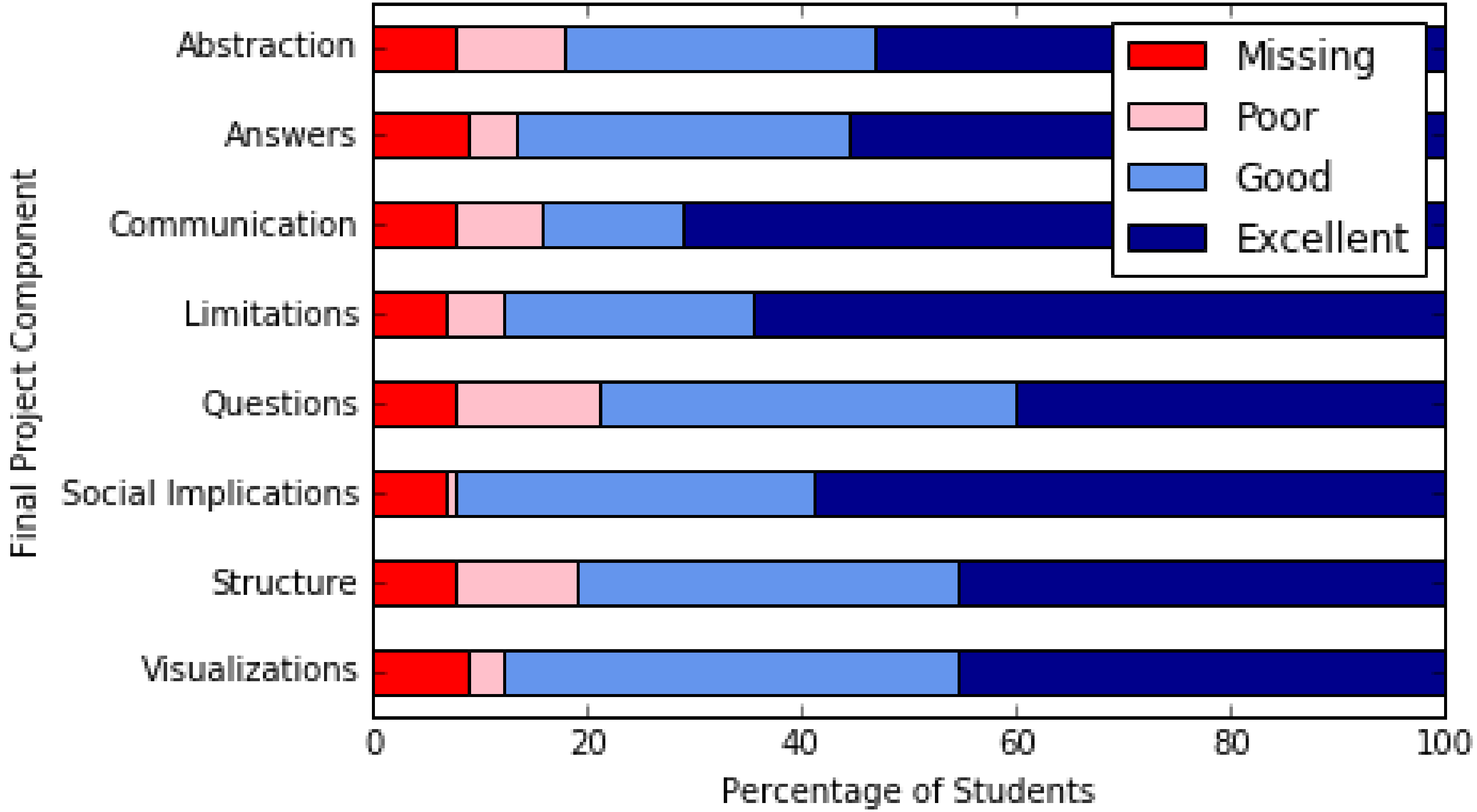
Students' Self-Efficacy



N = 85, 62% Female

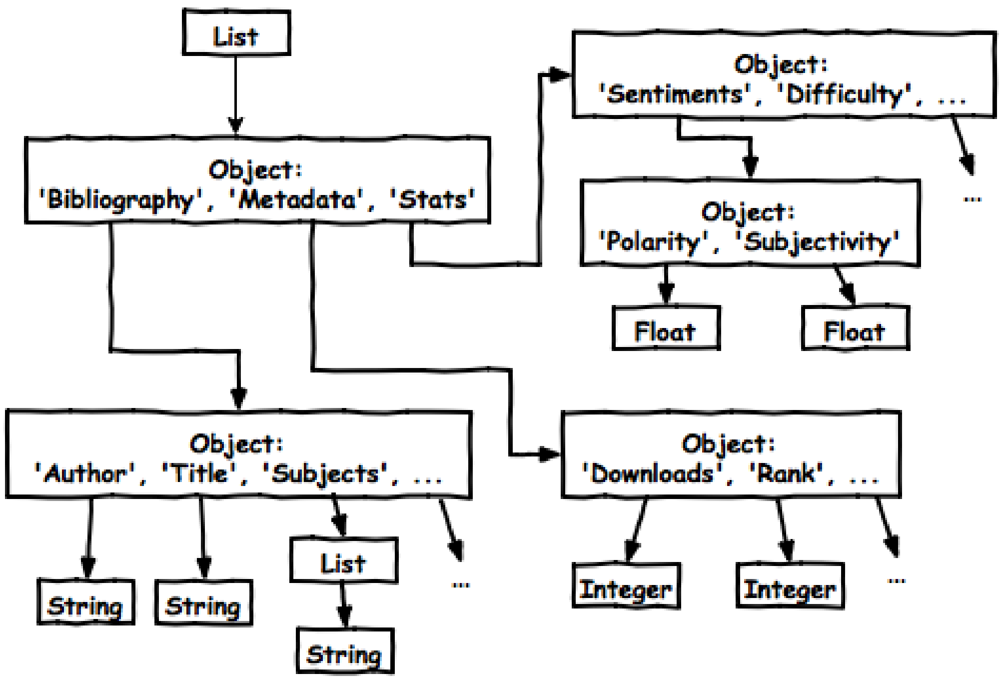
V-shaped in some cases, but otherwise increasing

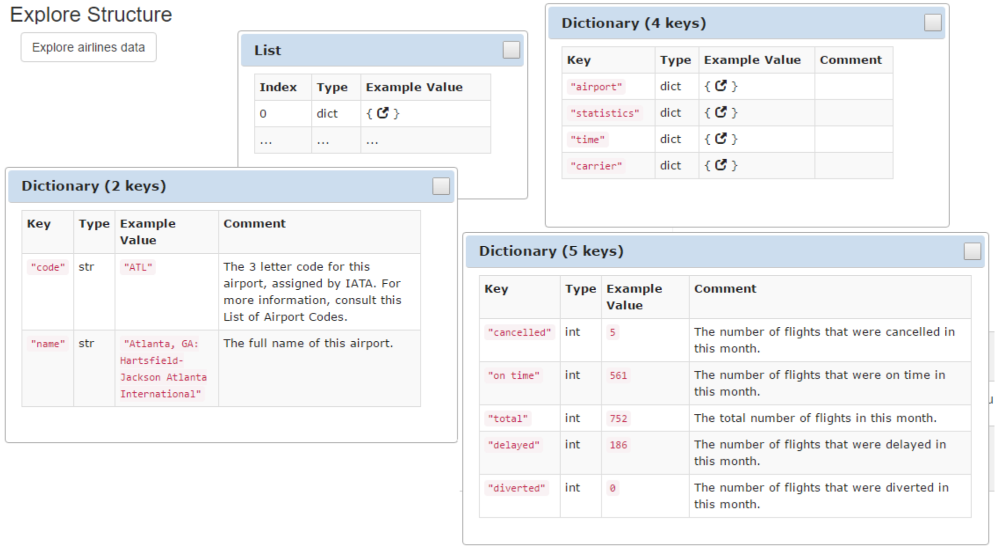
Final Project Scores



Most students (85%) received a Good or Excellent on each element

Structure





Explore Structure

Explore airlines data

List

Index	Type	Example Value
0	dict	{ }
...	...	...

Dictionary (4 keys)

Key	Type	Example Value	Comment
"airport"	dict	{ }	
"statistics"	dict	{ }	
"time"	dict	{ }	
"carrier"	dict	{ }	

Dictionary (2 keys)

Key	Type	Example Value	Comment
"code"	str	"ATL"	The 3 letter code for this airport, assigned by IATA. For more information, consult this List of Airport Codes.
"name"	str	"Atlanta, GA: Hartsfield-Jackson Atlanta International"	The full name of this airport.

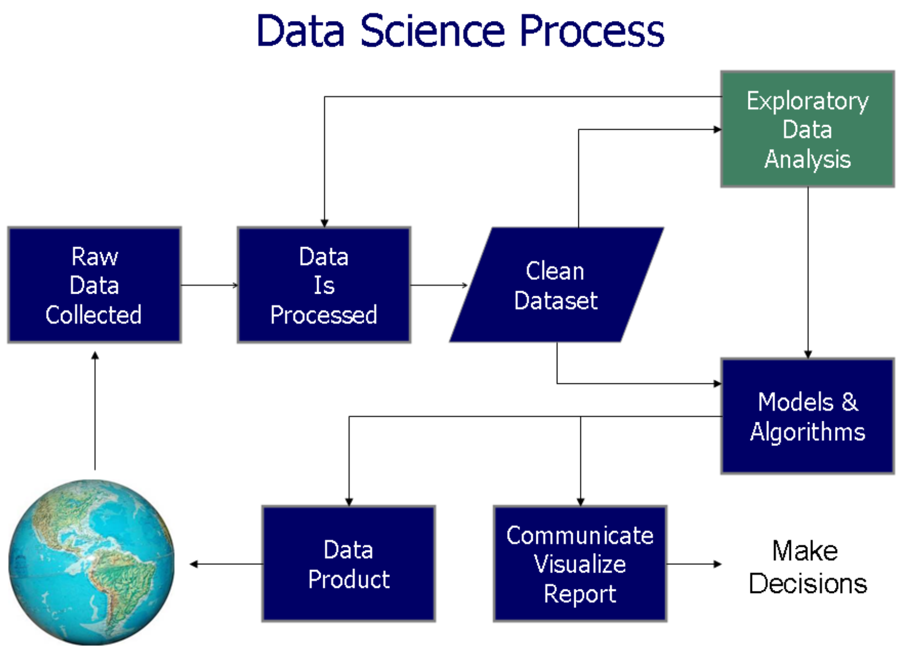
Dictionary (5 keys)

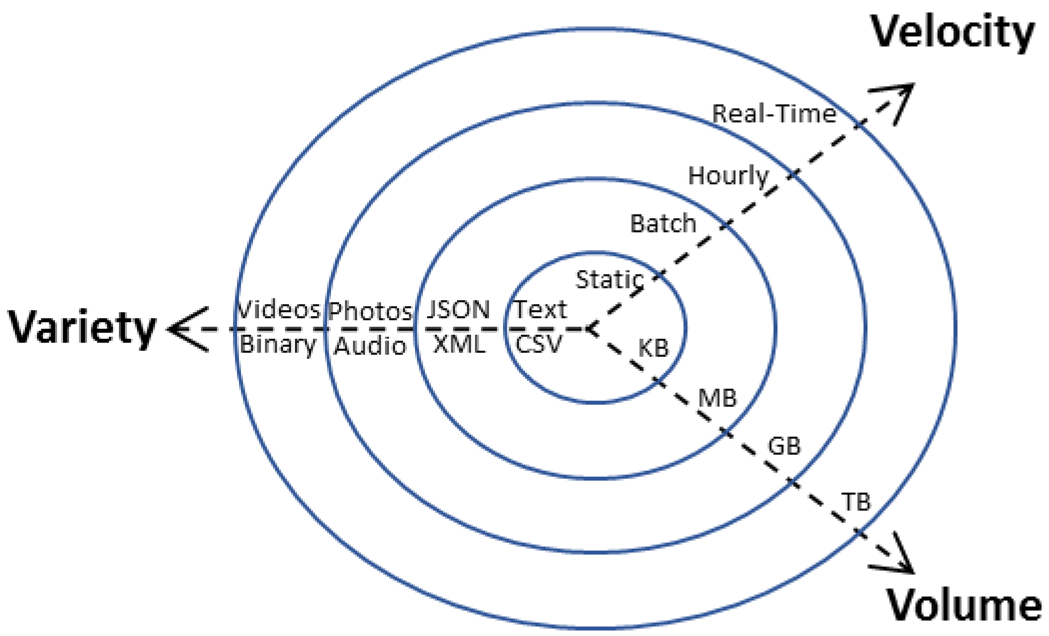
Key	Type	Example Value	Comment
"cancelled"	int	5	The number of flights that were cancelled in this month.
"on time"	int	561	The number of flights that were on time in this month.
"total"	int	752	The total number of flights in this month.
"delayed"	int	186	The number of flights that were delayed in this month.
"diverted"	int	0	The number of flights that were diverted in this month.

Situated Learning vs. Motivation

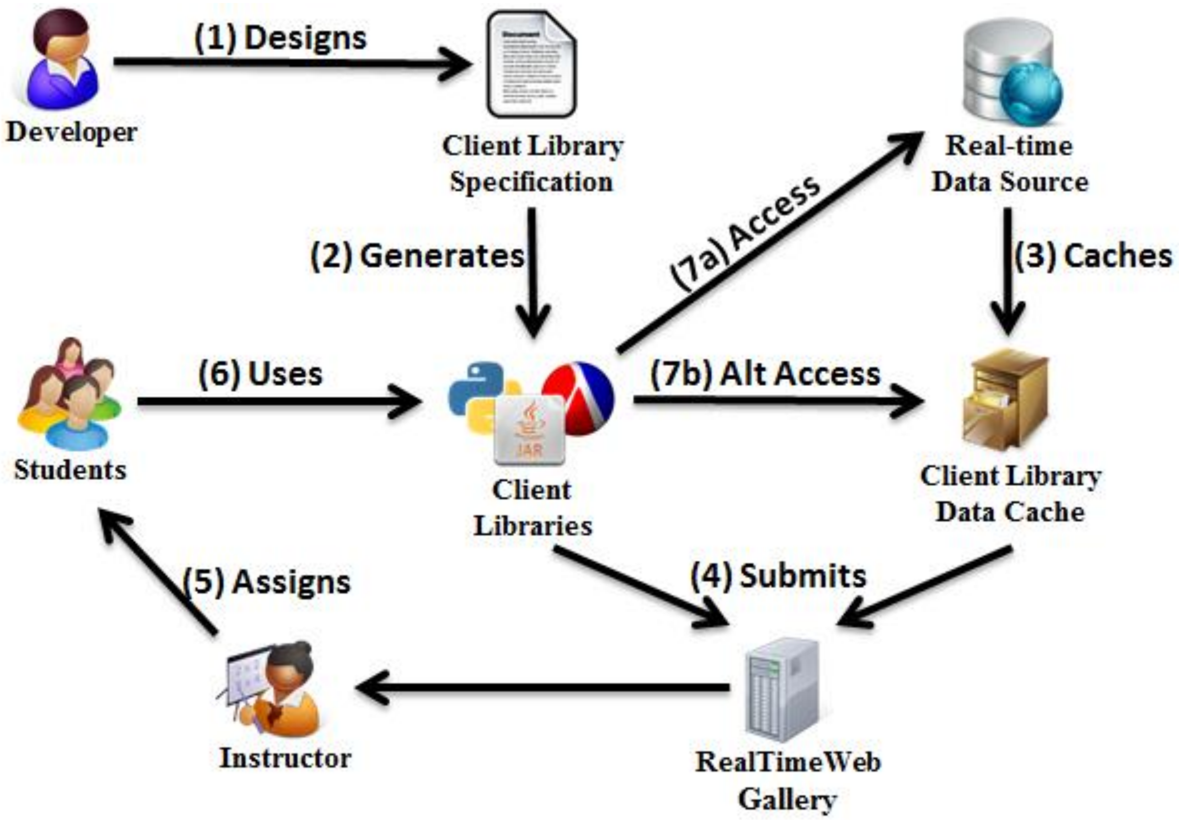
Situated Learning Component:	Context	Content	Facilitations	Assessment
Example	"Game Design"	"For Loops"	Blocks-based environment, teaching assistants, etc.	Exams, performance review, code review
eMpowerment	Am I restricted by the context to explore what I want?	Do I have control over the depth/breadth/direction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

Big Idea: Real-World Data

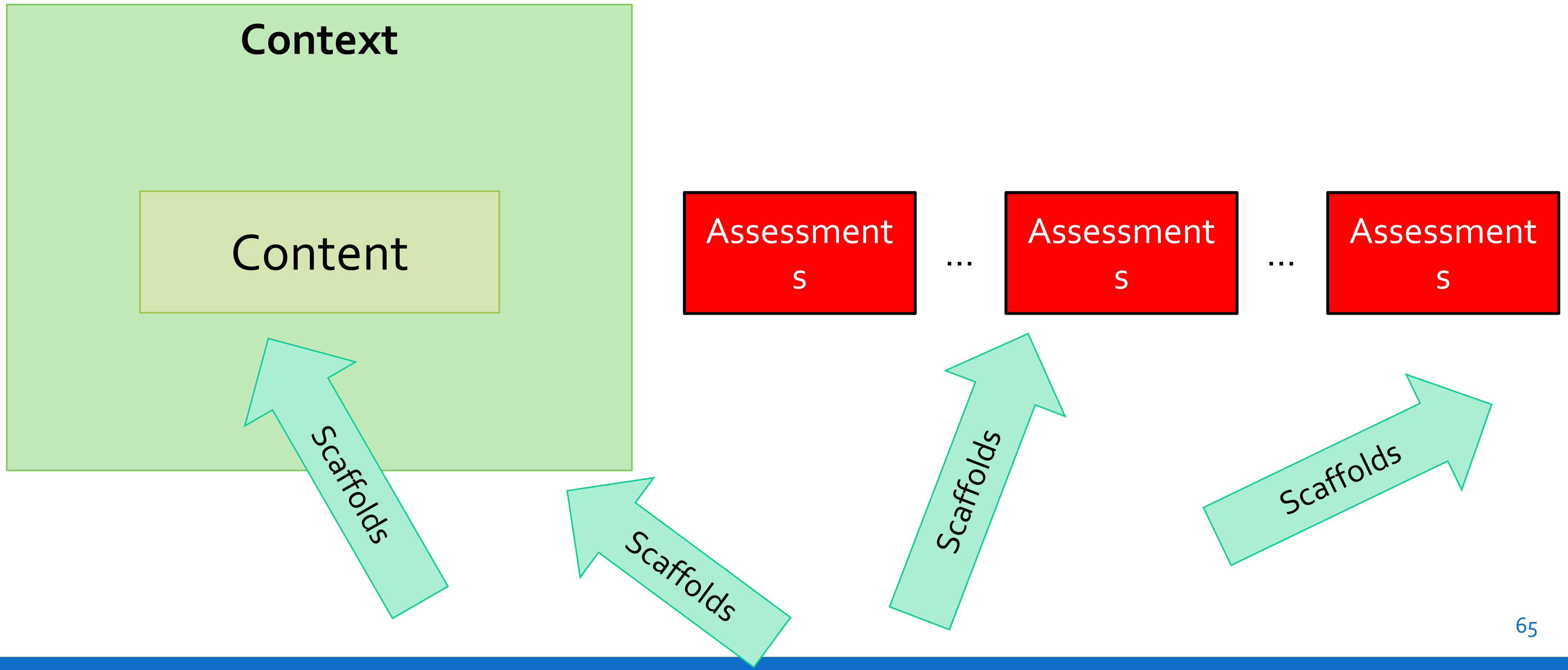




Complete Picture



Situated Learning Framework
Choi & Hannafin



Cache Files = Sophisticated Snapshots

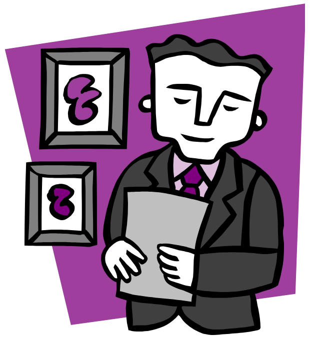
june_18_2013.json

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

Call	Returns
#1	5 earthquakes
#2	2 earthquakes
#3	7 earthquakes
…	…

Three Components

Client Libraries



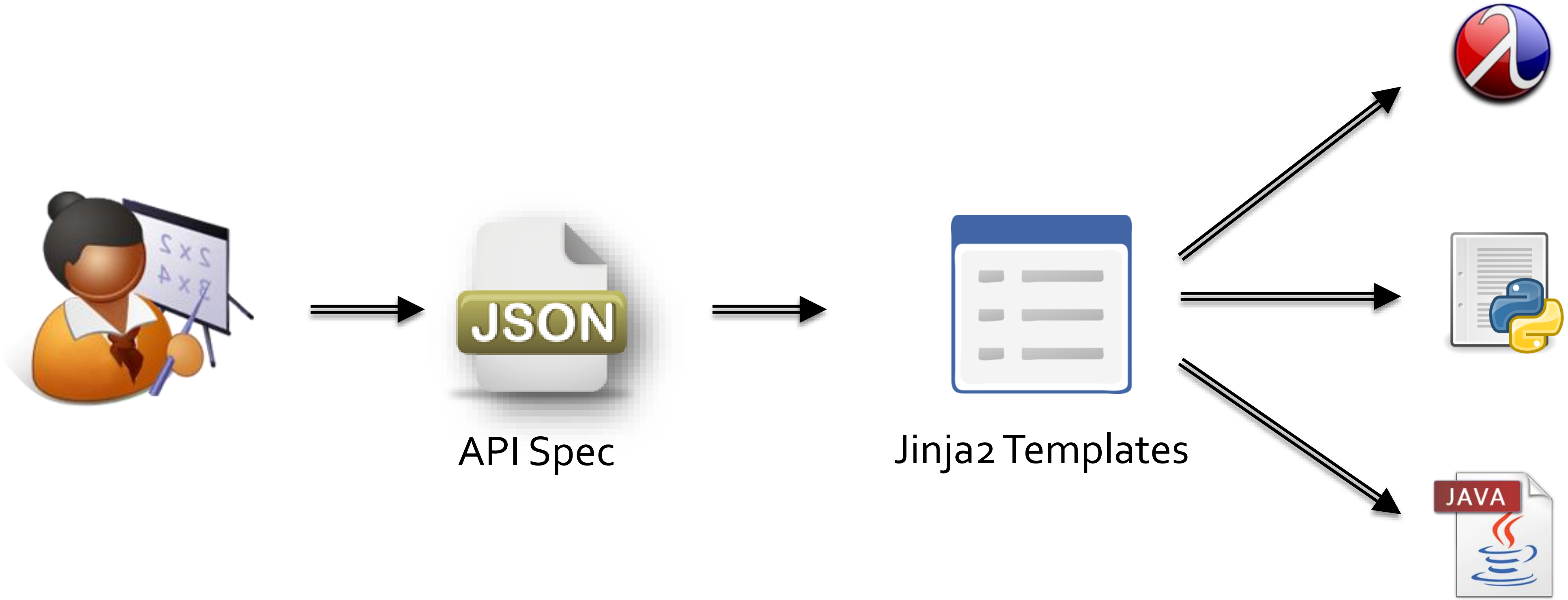
Curated Gallery

Library Generator

Gallery - Initial Offering

- Earthquakes
- Weather
- Stocks
- Reddit
- Magic the Gathering

Client Library Building



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

Contexts: Math and Business

Pure Math (e.g., Fibonacci)

$$1,\ 1,\ 2,\ 3,\ 5,\ 8,\ 13,\ 21,\ 34,\ 55,\ 89,\ 144,\ \ldots$$

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

Saad Mneimneh. 2015. Fibonacci in The Curriculum: Not Just a Bad Recurrence. In Proceedings of the 46th ACM Technical Symposium on Computer Science Education (SIGCSE '15). ACM, New York, NY, USA, 253-258.