COMPUTING WITH CORGIS:
DIVERSE, REAL-WORLD DATASETS FOR INTRODUCTORY COMPUTING

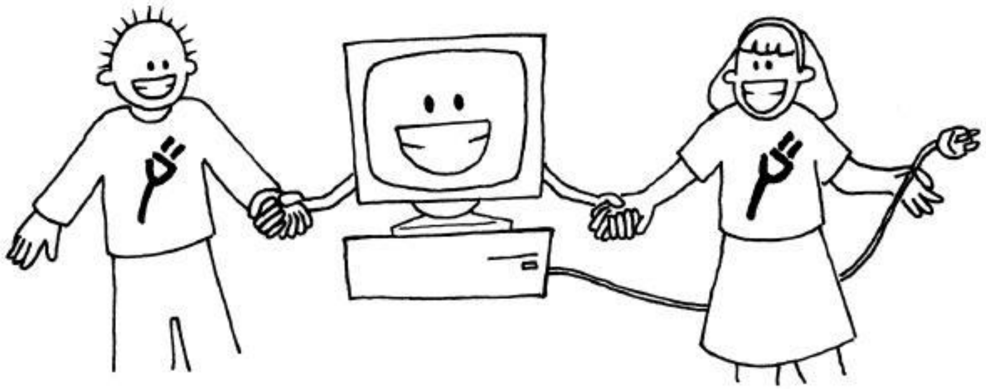
Austin Cory Bart, Ryan Whitcomb,
Dennis Kafura, Clifford A. Shaffer, Eli Tilevich

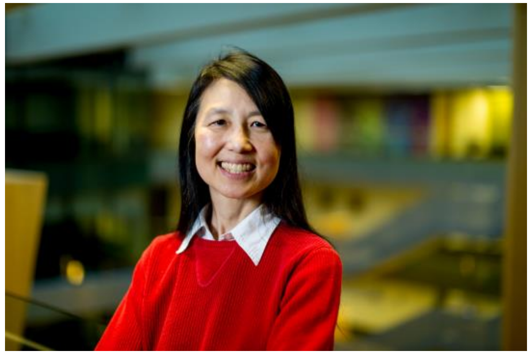
Virginia Tech

Overview

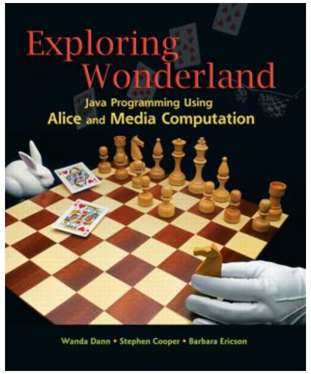
- Bringing real-world data into introductory computing classes
- Via a new system that manages and produces datasets
- In order to motivate non-computing majors

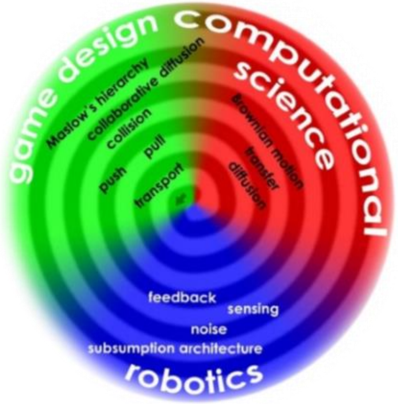
Computer Science For All



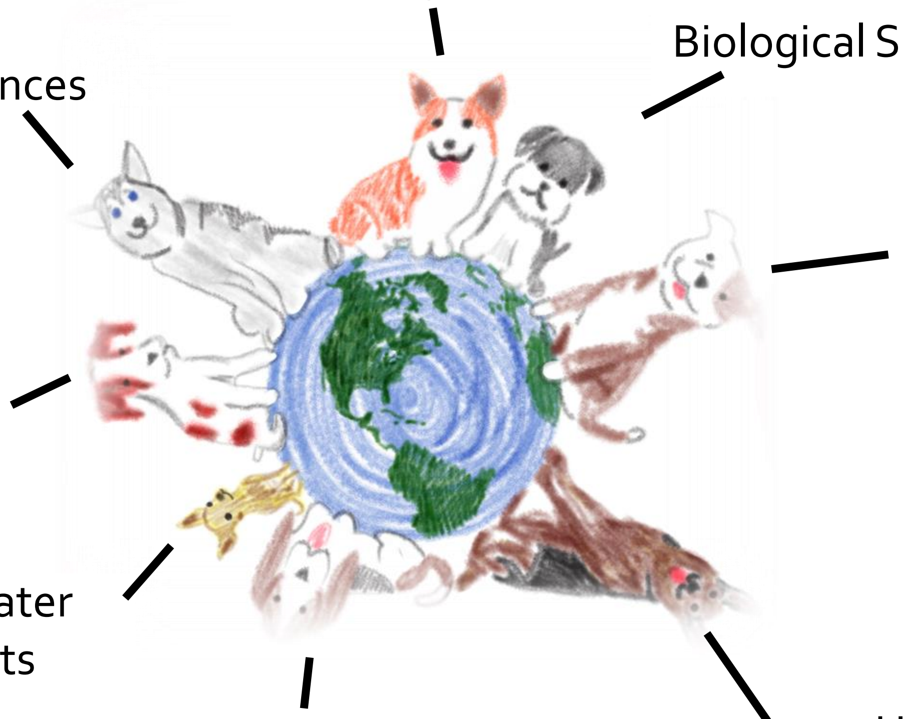


COMPUTATIONAL
THINKING
AT
GOOGLE


Exploring
Wonderland
Java Programming Using
Alice and Media Computation
Wanda Dann • Stephen Cooper • Barbara Ericson



Diverse Majors ... with Rich Knowledge



English

Animal Sciences

Biological Sciences

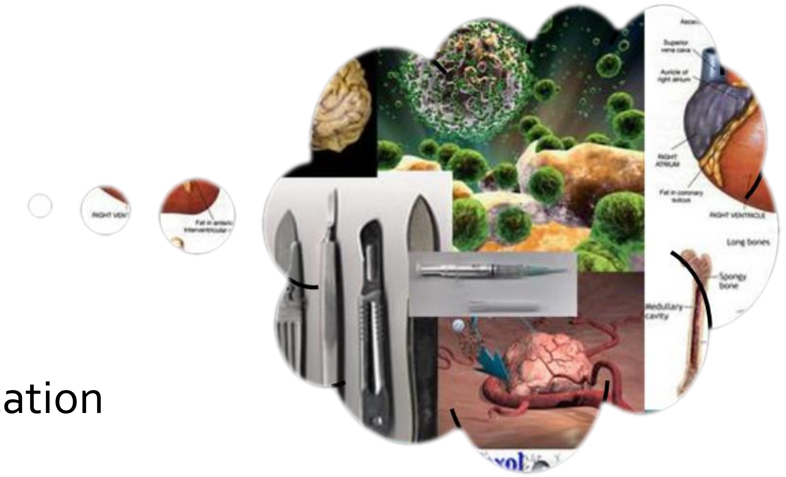
Education

Chemistry

Theater Arts

Building Construction

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?"

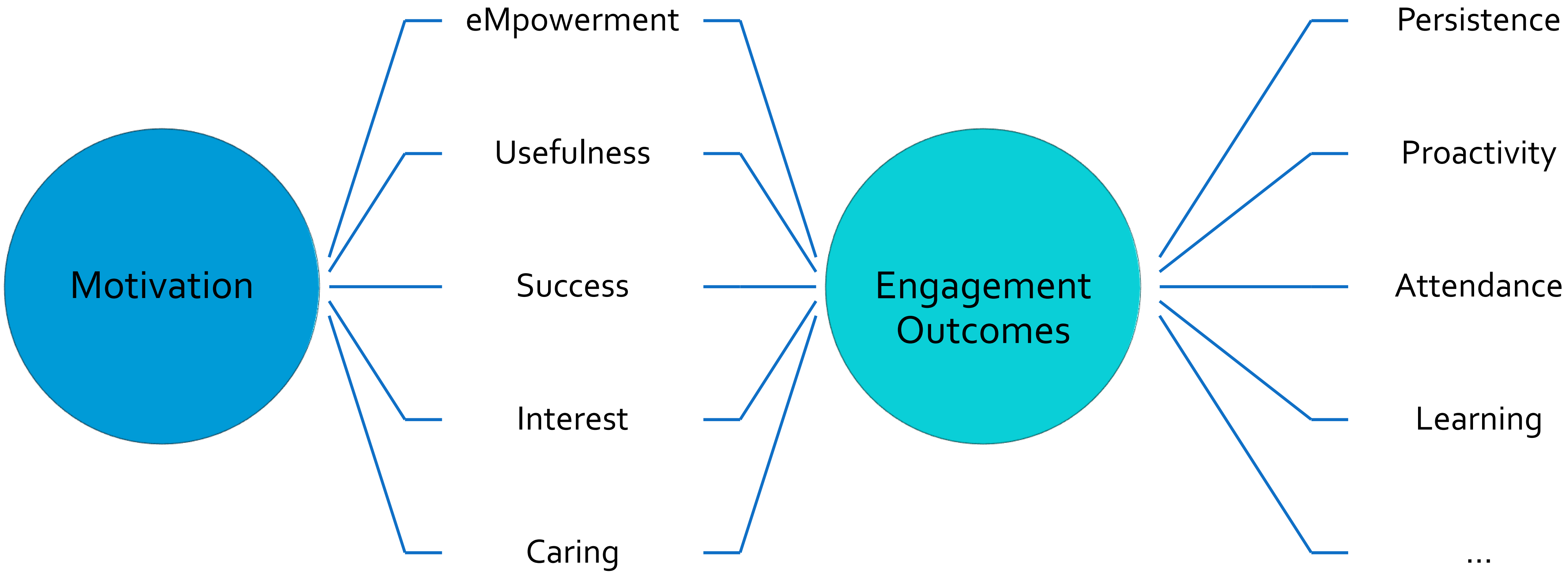
MUSIC Model of Academic Motivation

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

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



A spectrum

FOR-EACH

WHILE

Recursion

Iteration

IF

Algorithms

Assignment

Data Structures

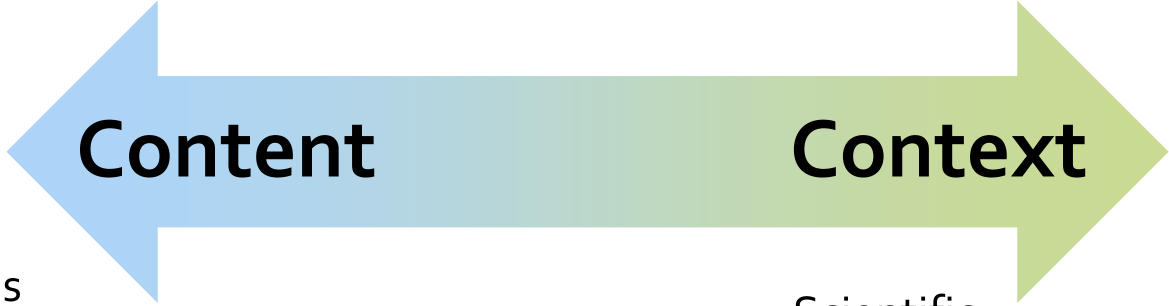
Dictionaries

Lists

Arrays

Integers

Booleans



Content

Context

Games

Websites

Development

Mobile Apps

Images

Media Computation

Audio

Animations

Scientific Modelling

Scientific Computing

Math

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”

State of the Art

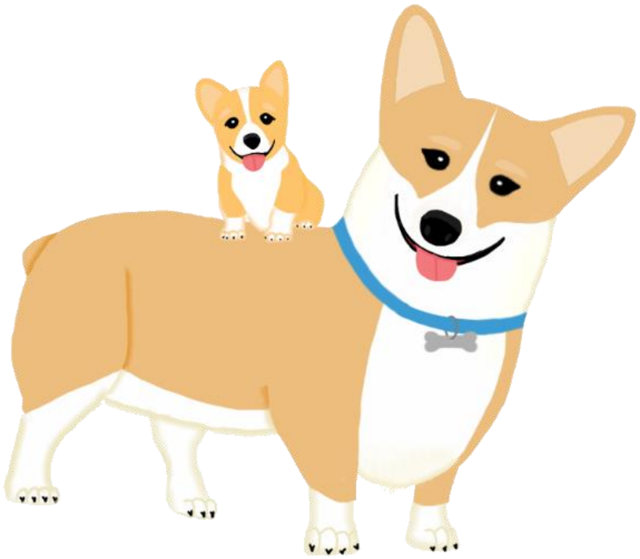
- Bart 2014 – Connecting to real-time APIs (RealTimeWeb)
- Hamid 2016 – More generalized framework for real-time APIs (Sinbad)
- Subramanian 2014 – Visualization of data structures with real data (BRIDGES)
- Anderson 2014 – Real world data in CS1
- Sullivan 2013 – Data Science for non-majors

Problem – We Need Data

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

CORGIS

The Collection Of Really Great, Interesting, Situated Datasets



Metrics

42 datasets

267 mB

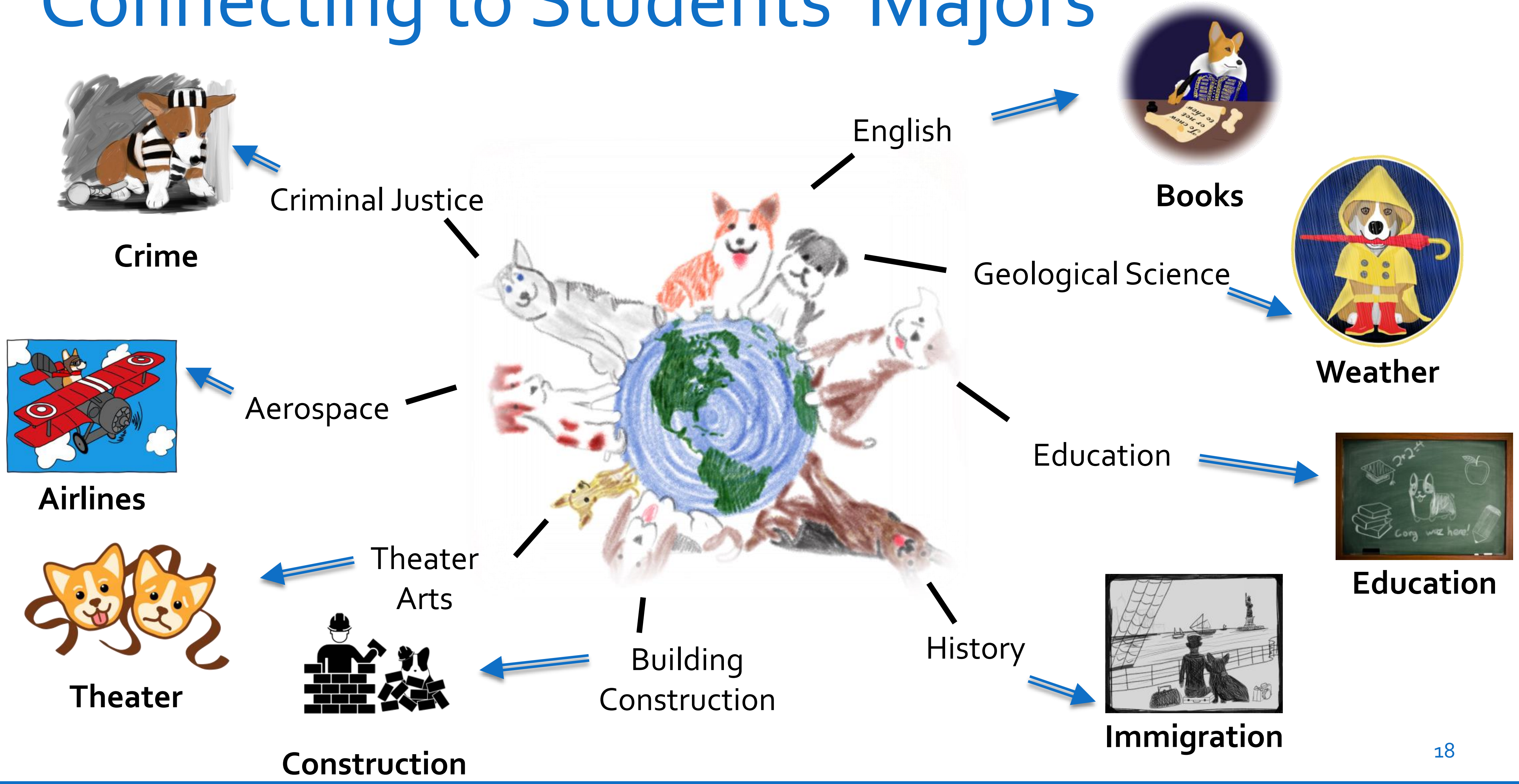
420,672 rows

9,365,520 values

Datasets

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

Connecting to Students’ Majors



Criminal Justice

Crime

Aerospace

Airlines

Theater Arts

Theater

Building Construction

Construction

English

Books

Geological Science

Weather

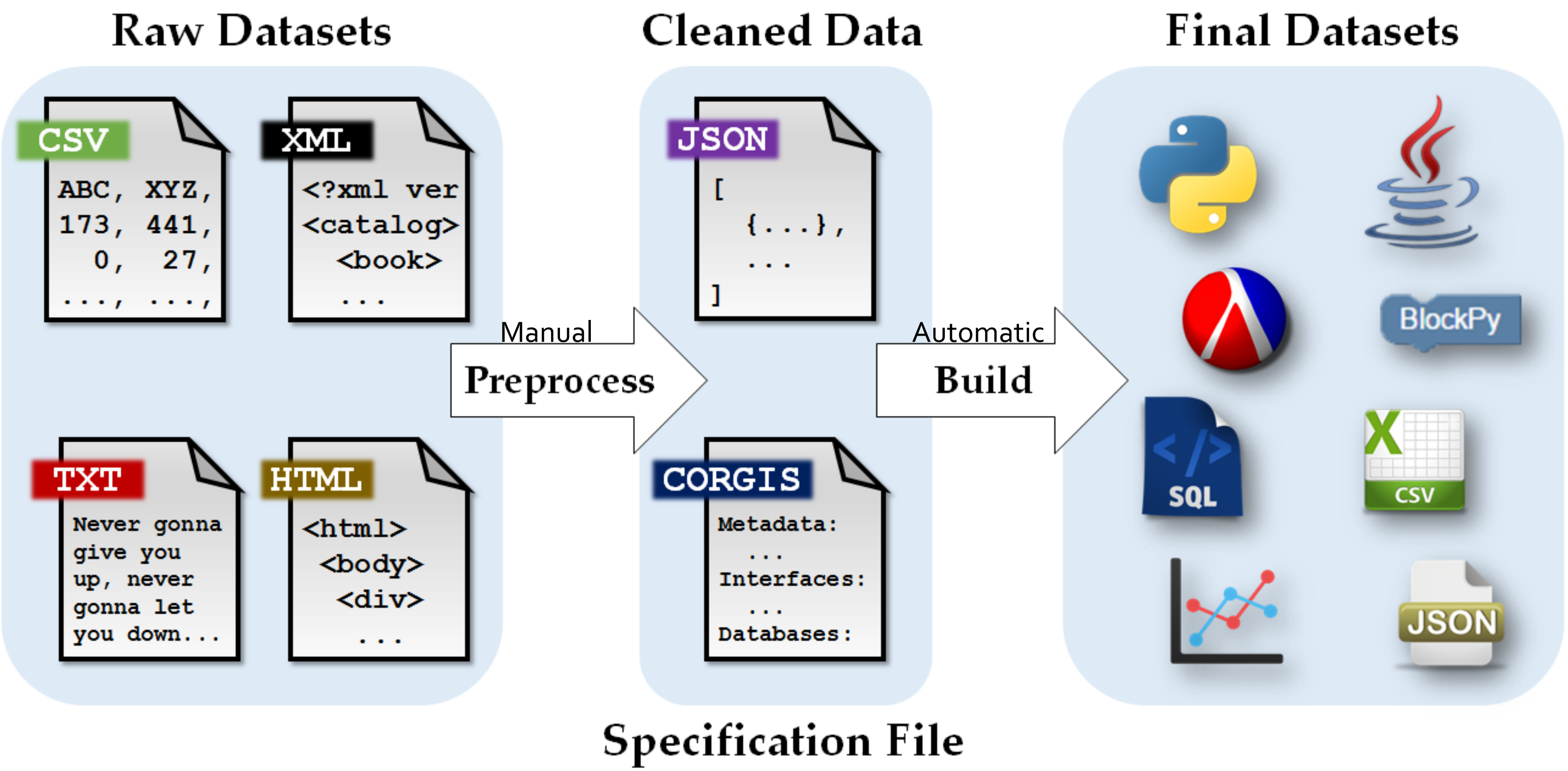
Education

Education

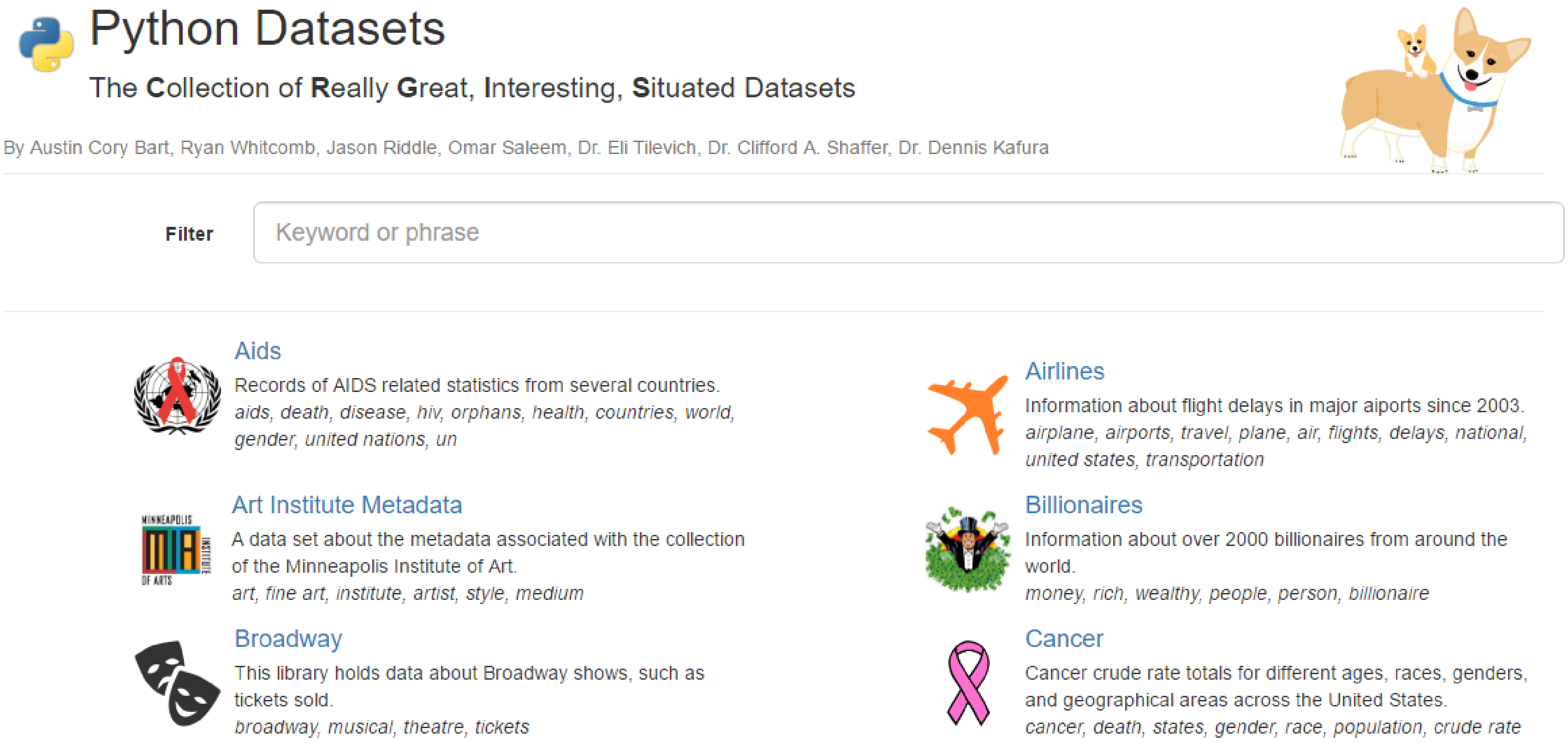
History

Immigration

Architecture



Gallery



Python Datasets

The **C**ollection of **R**eally **G**reat, **I**nteresting, **S**ituated 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

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

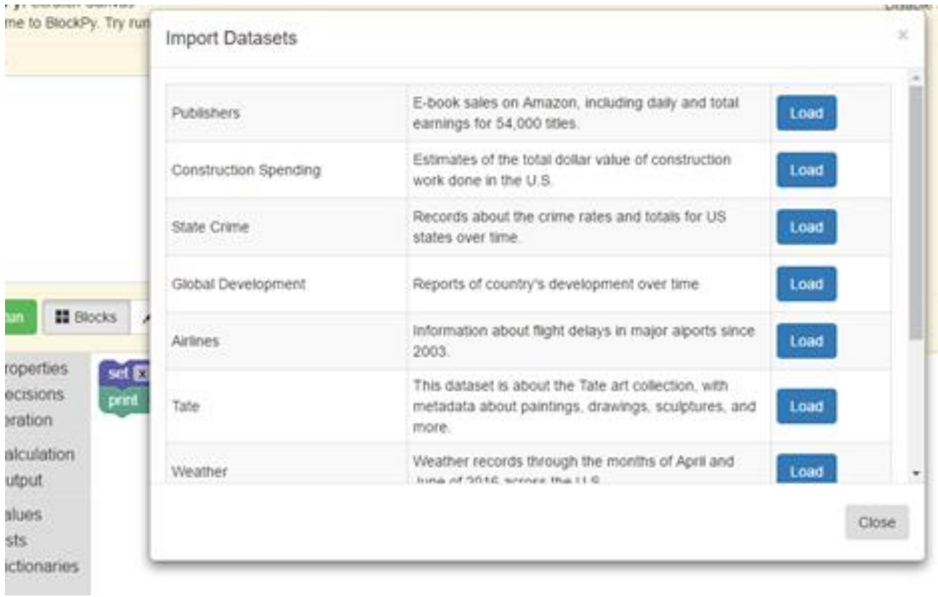
Java, Python, Racket

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

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

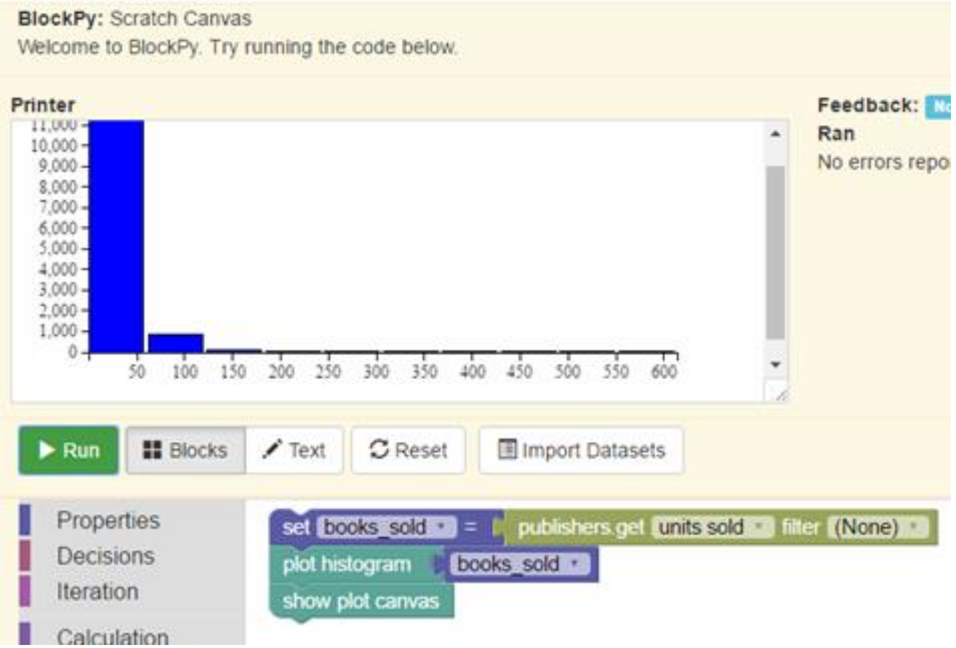
BlockPy



me to BlockPy. Try run

Import Datasets

Publishers	E-book sales on Amazon, including daily and total earnings for 54,000 titles.	Load
Construction Spending	Estimates of the total dollar value of construction work done in the U.S.	Load
State Crime	Records about the crime rates and totals for US states over time.	Load
Global Development	Reports of country's development over time	Load
Airlines	Information about flight delays in major aiports since 2003.	Load
Tate	This dataset is about the Tate art collection, with metadata about paintings, drawings, sculptures, and more.	Load
Weather	Weather records through the months of April and June of 2016 across the U.S.	Load



BlockPy: Scratch Canvas

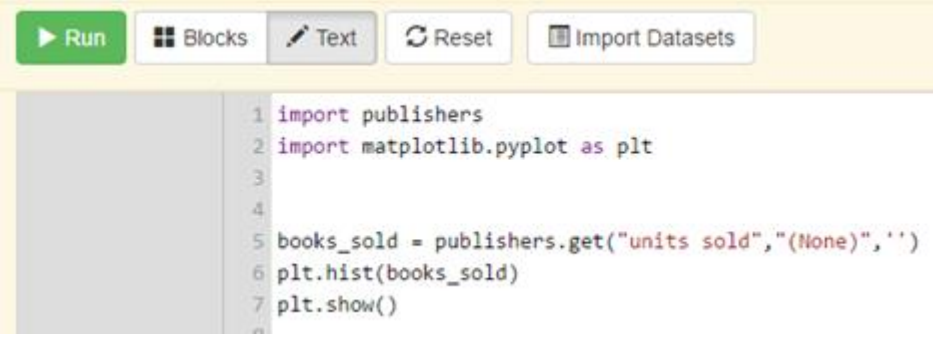
Welcome to BlockPy. Try running the code below.

Printer

Feedback: No

Ran

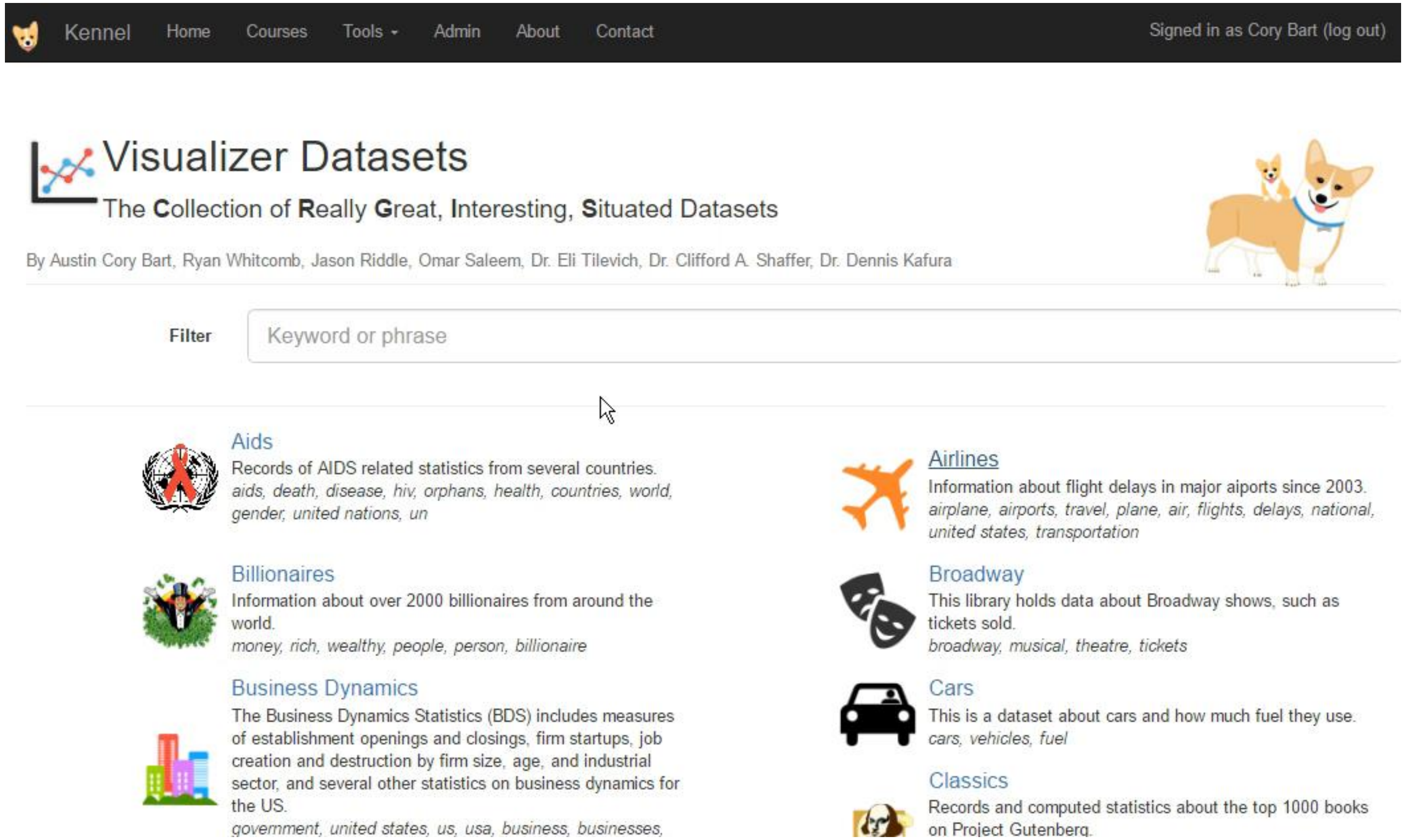
No errors repo



```
import publishers
import matplotlib.pyplot as plt


books_sold = publishers.get("units sold","(None)",'')
plt.hist(books_sold)
plt.show()
```

Visualizer Demo



Visualizer Datasets

The **C**ollection of **R**eally **G**reat, **I**nteresting, **S**ituated 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

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

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,

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

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

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

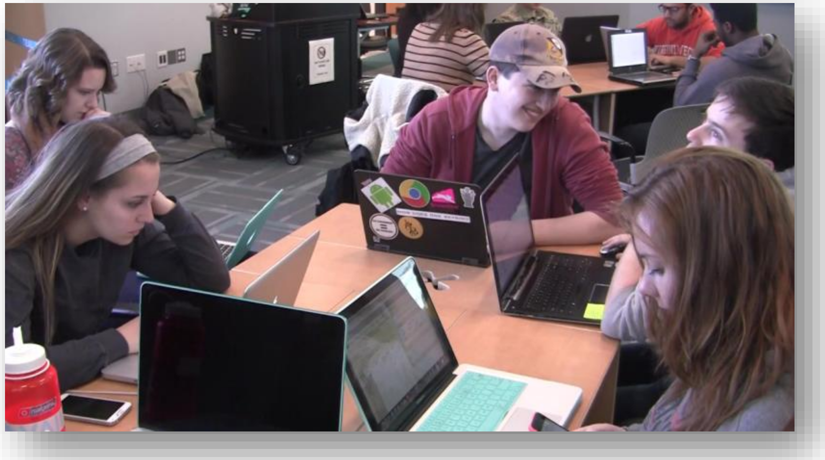
Classics
Records and computed statistics about the top 1000 books on Project Gutenberg.

Hypotheses

- Context provides motivation
- Students have some preference for Data Science
- The usefulness of the context connects to engagement outcomes as strongly as the content

Interventions

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



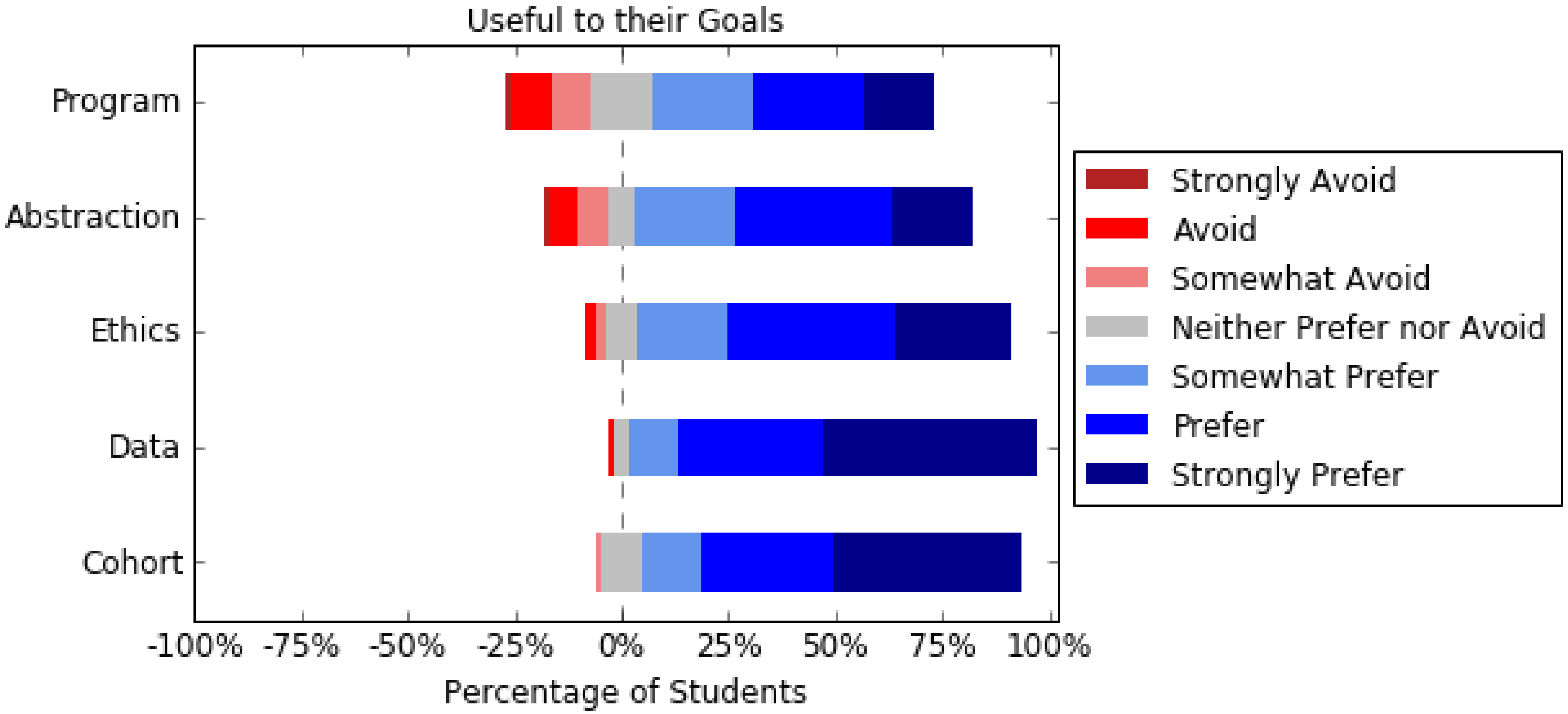
Motivation × Course Components

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

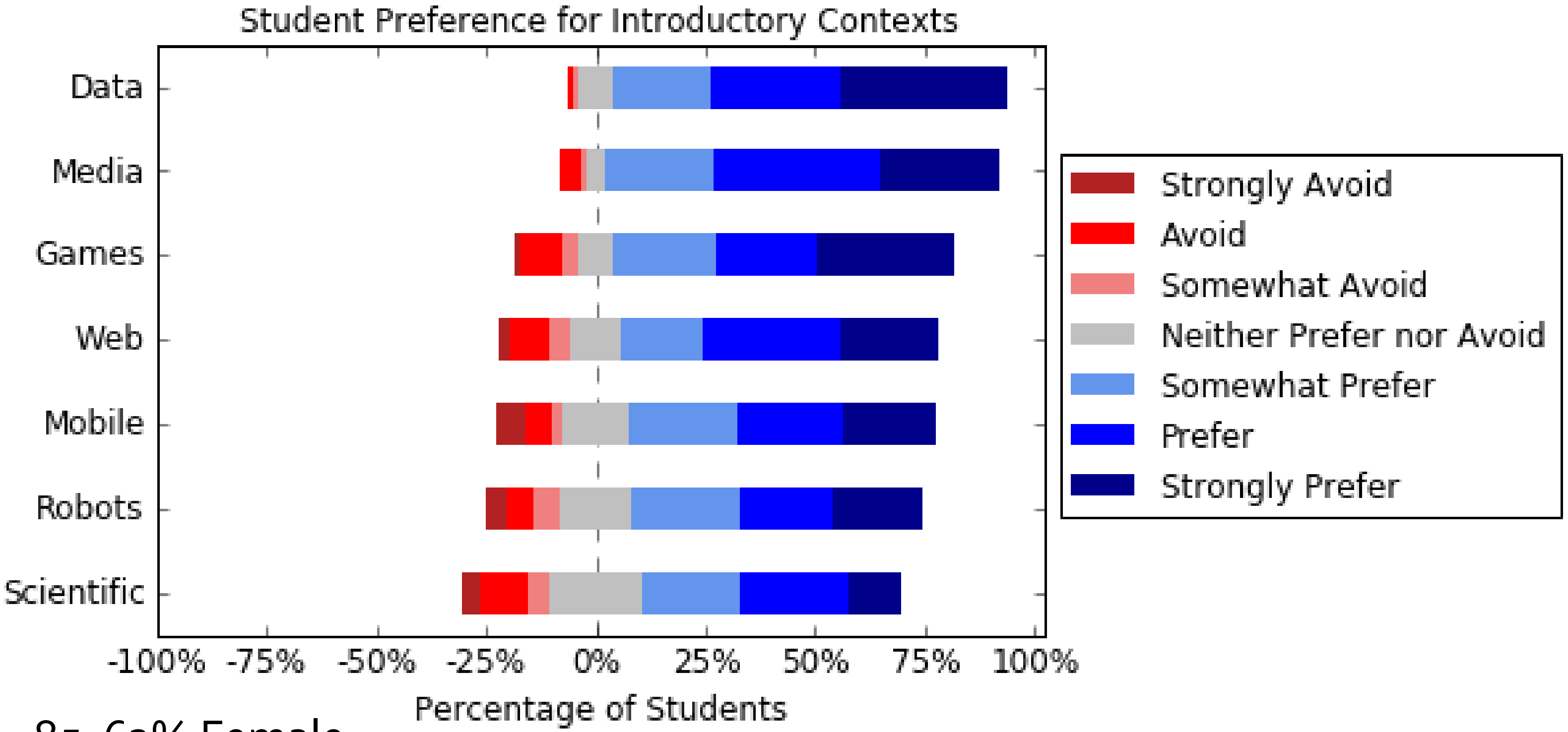
Students’ sense of the usefulness of various course components was highest for the **context**, lowest for the **content**.

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

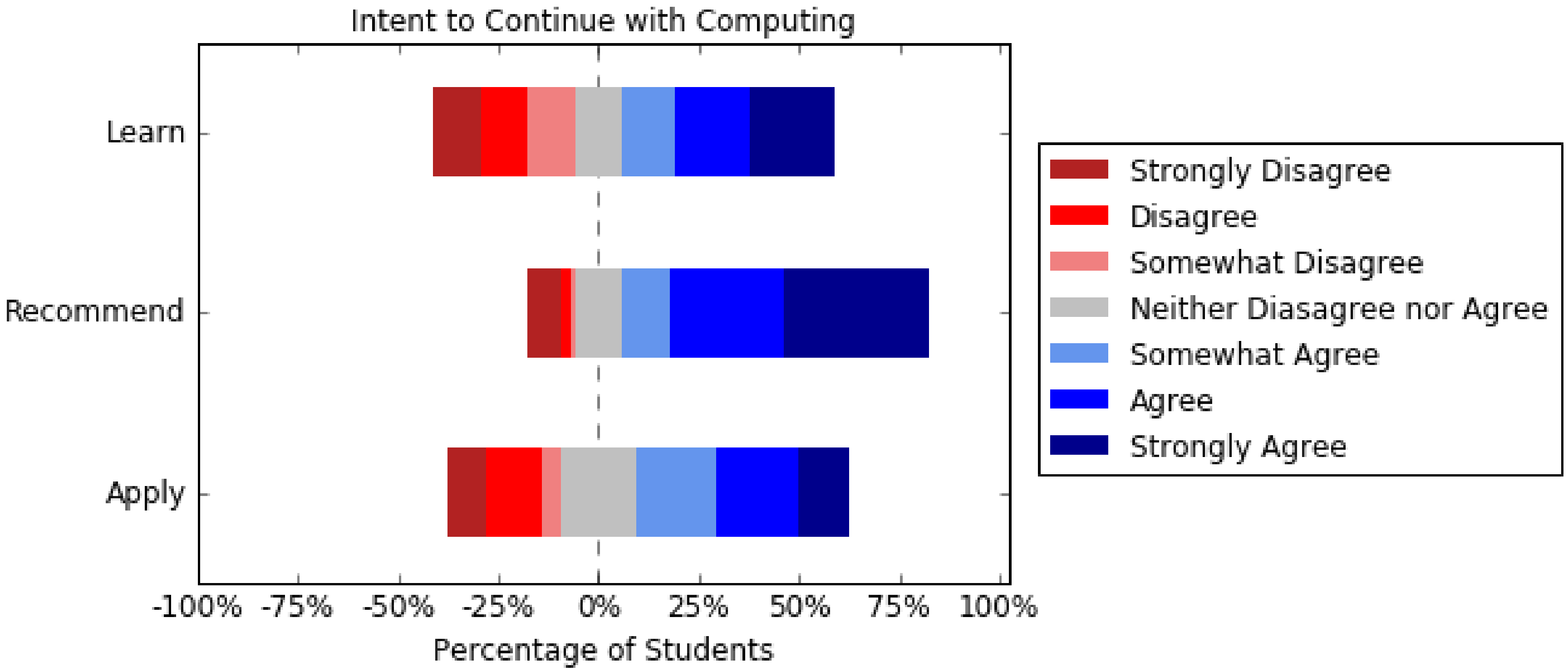
** No significant difference with Media Computation according to KW test*

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.

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					
Cohort					
Data					
Ethics					
Programming		**.406**	**.354**	**.341**	

Not significantly Correlated!

Significant

N = 85, 62% Female

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

Take-aways

- Data Science seems to be a preferable context for students, across genders.
- 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

- More Datasets
- Maintenance
- Connecting motivation to learning outcomes

Thanks!

Clifford A. Shaffer

Dennis Kafura

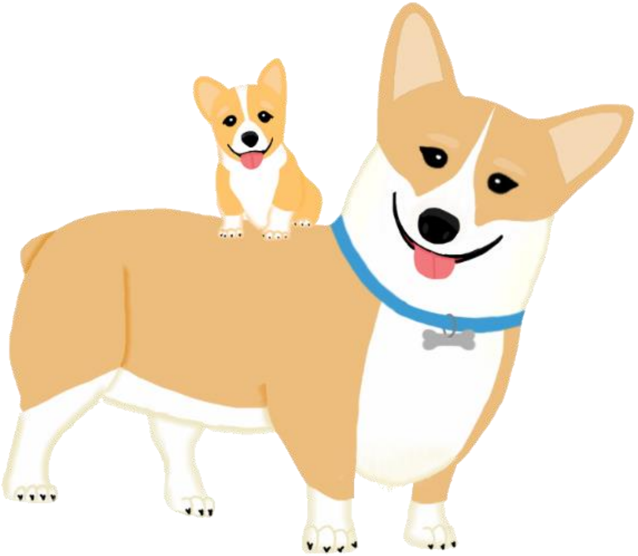
Eli Tilevich

Ryan Whitcomb

We gratefully acknowledge the support of Virginia Tech and the National Science Foundation under Grants NSF DGE 0822220, NSF IUSE 1624320, and NSF IUSE 479632.

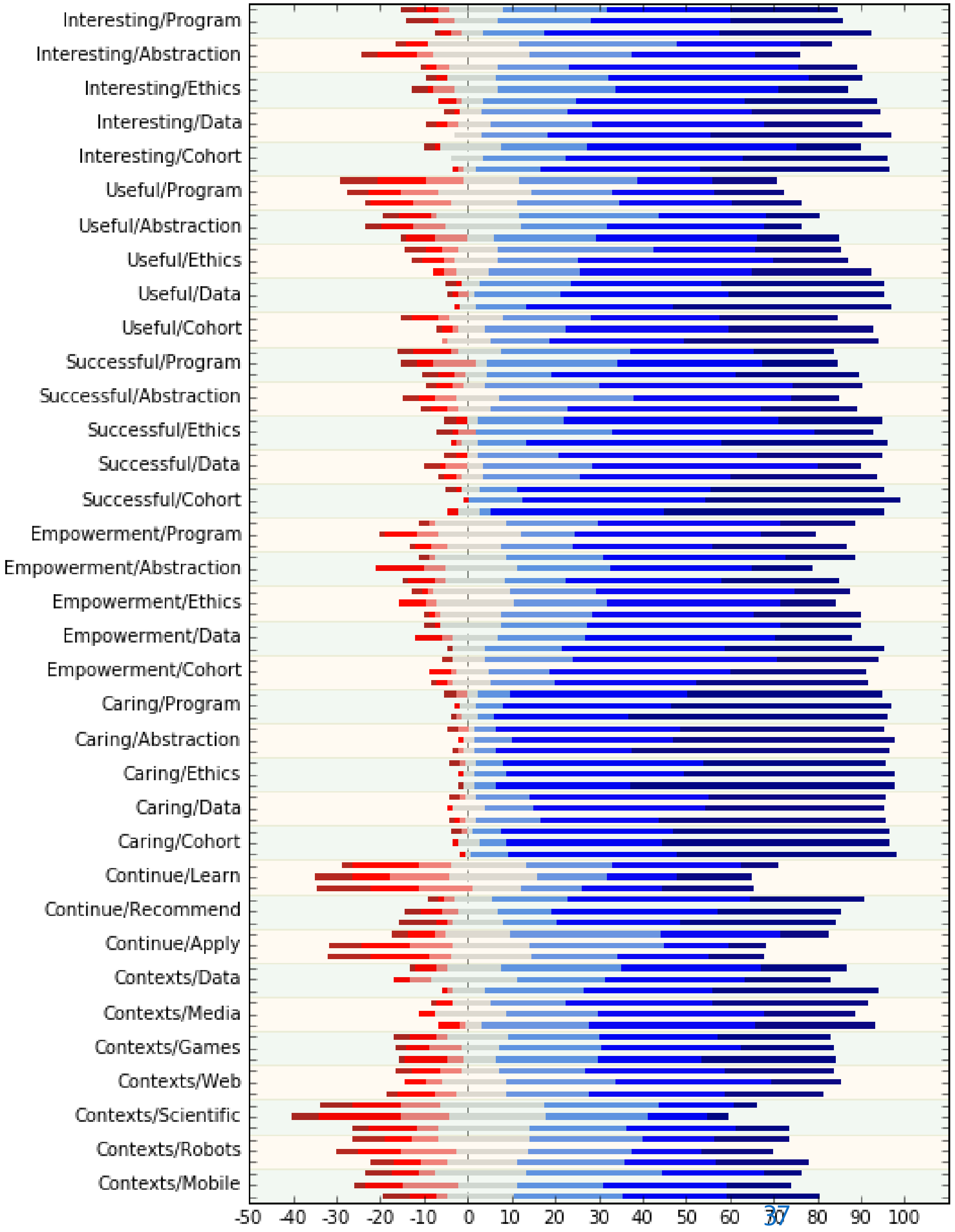
Questions?

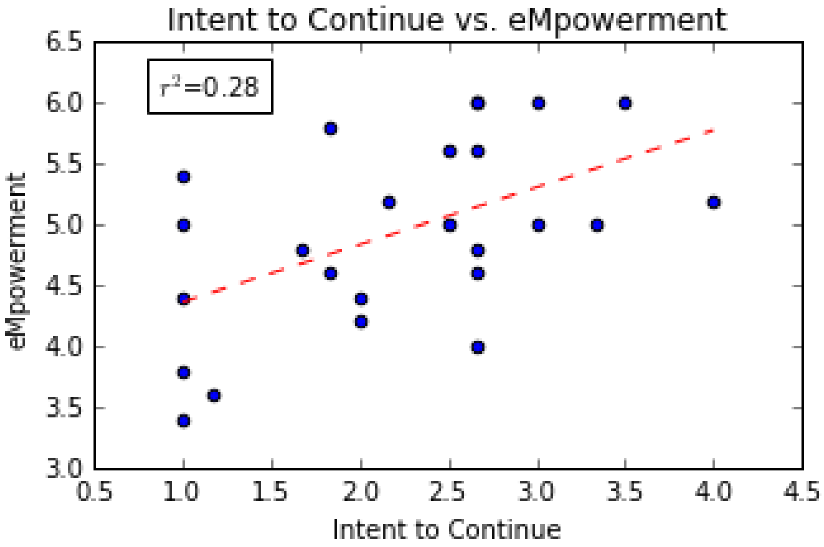
https://think.cs.vt.edu/corgis

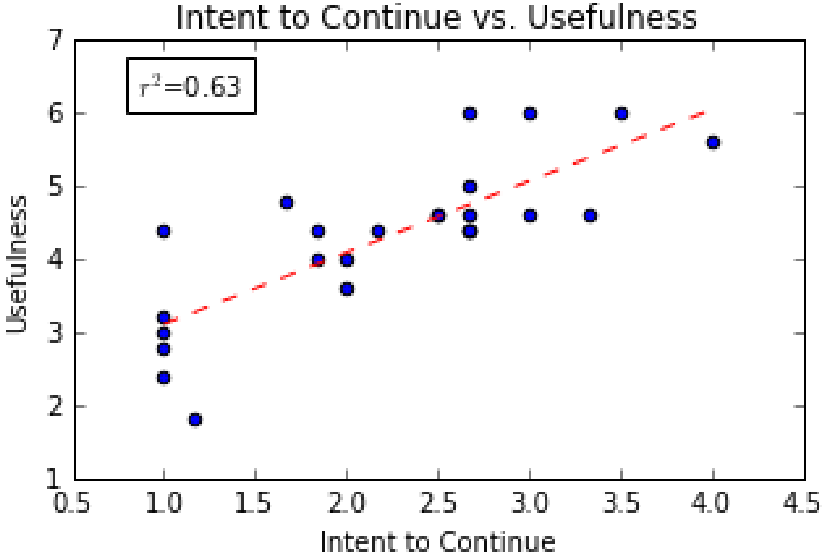


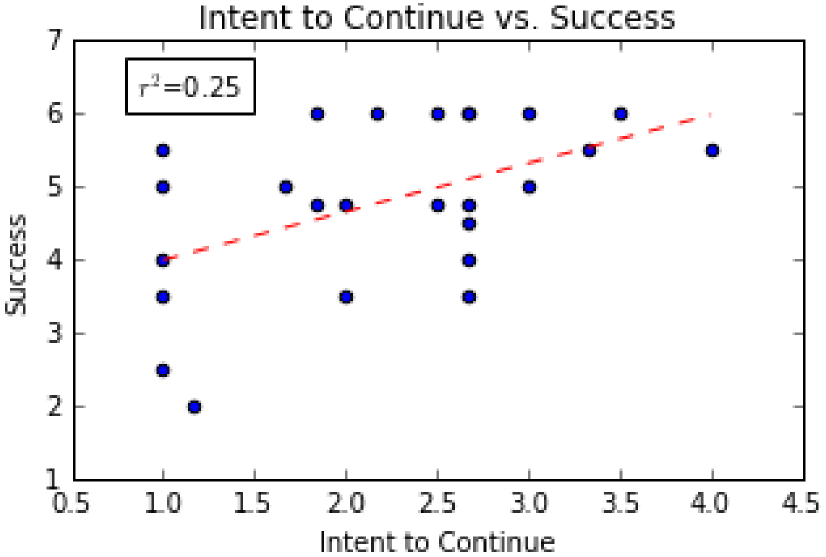
Artwork by Eleonor Bart

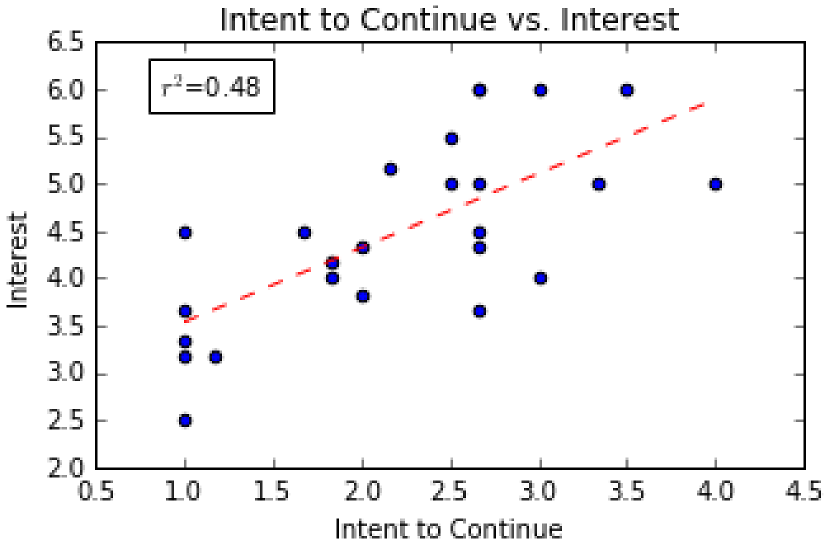
Trends in Motivation

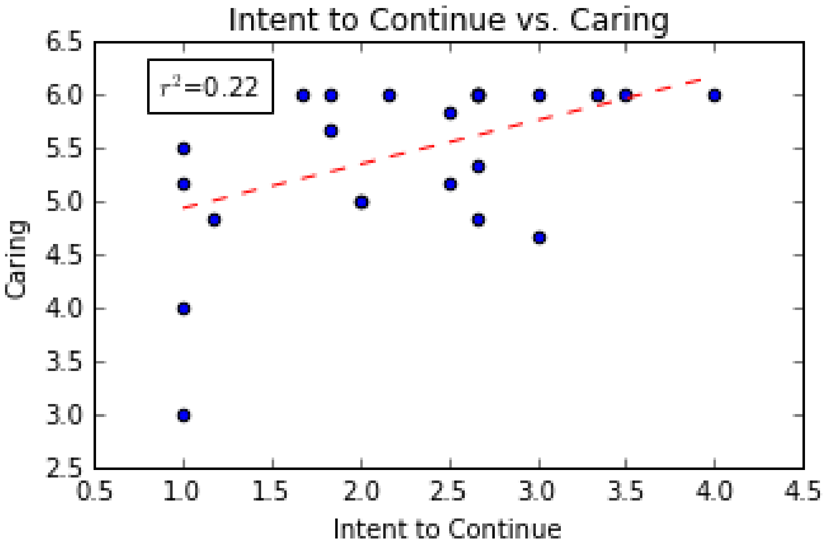


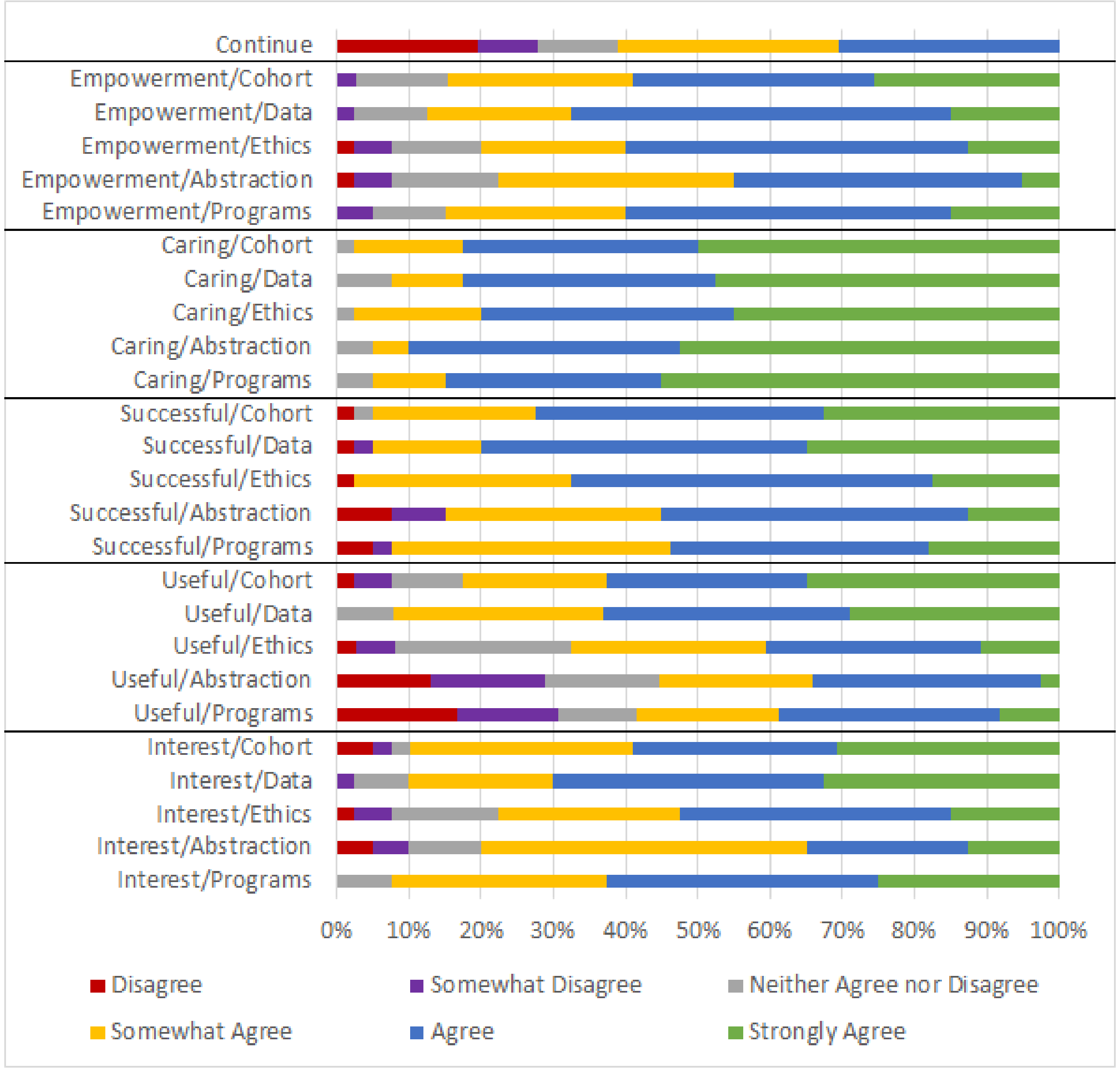

Intent to Continue vs. eMpowerment
r^2=0.28
eMpowerment
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Intent to Continue


Intent to Continue vs. Usefulness
r^2=0.63
Usefulness
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Intent to Continue

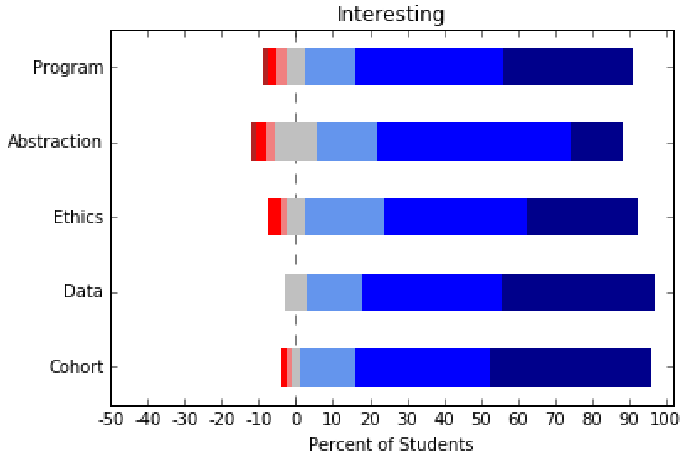

Intent to Continue vs. Success
r^2=0.25
Success
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Intent to Continue

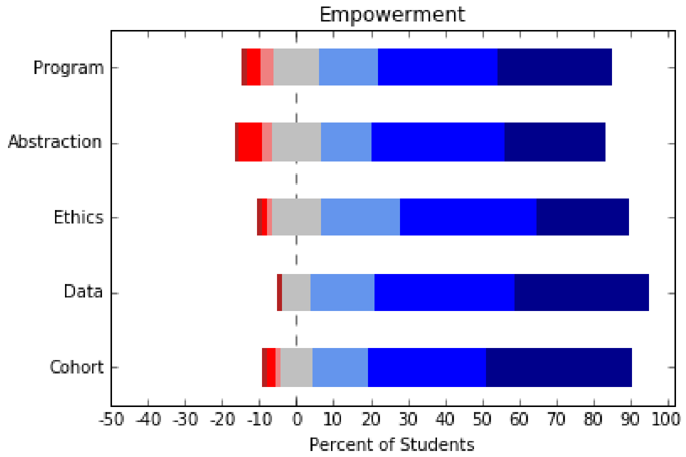

Intent to Continue vs. Interest
r^2=0.48
Interest
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Intent to Continue

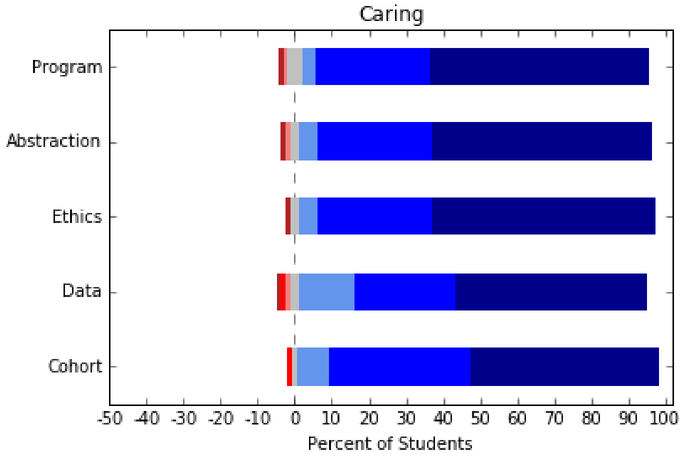

Intent to Continue vs. Caring
r^2=0.22
Caring
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Intent to Continue

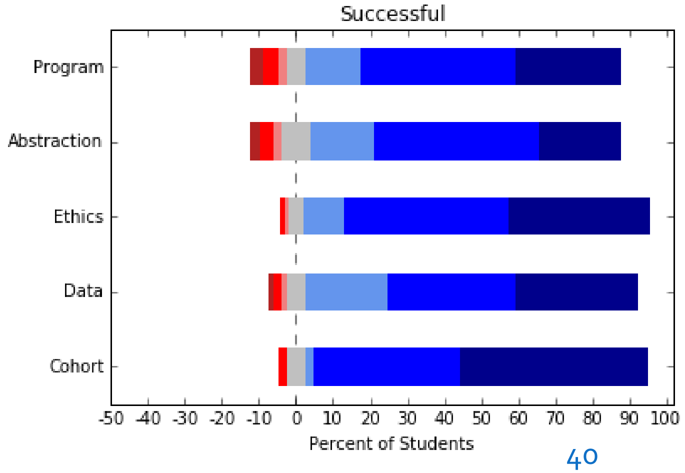

Continue
Empowerment/Cohort
Empowerment/Data
Empowerment/Ethics
Empowerment/Abstraction
Empowerment/Programs
Caring/Cohort
Caring/Data
Caring/Ethics
Caring/Abstraction
Caring/Programs
Successful/Cohort
Successful/Data
Successful/Ethics
Successful/Abstraction
Successful/Programs
Useful/Cohort
Useful/Data
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

Other Components









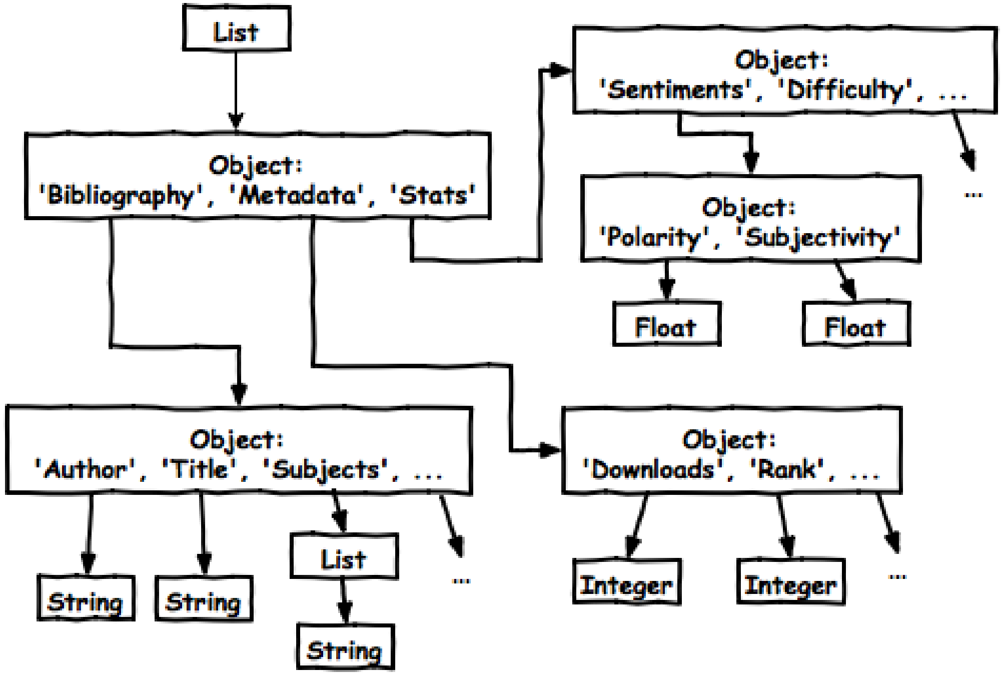
Spring 2016	eMpowerment	Usefulness	Success	Interest	Caring
Abstraction	.458	.699	.614	.488	
Cohort					
Data					
Ethics		.485	.418	.323	
Programming	.437	**.823**	.600	.638	

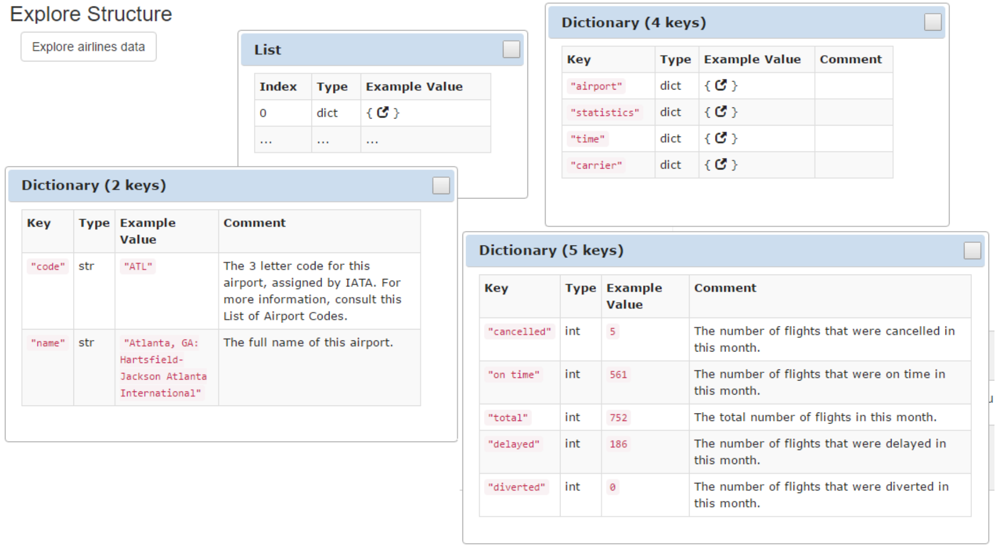
Continue Learning, Applying, and/or Recommend Course

N =36

50% female

Structure


List
Object:
'Bibliography', 'Metadata', 'Stats'
Object:
'Sentiments', 'Difficulty', ...
...
Object:
'Polarity', 'Subjectivity'
Float
Float
Object:
'Author', 'Title', 'Subjects', ...
String
String
List
String
...
Object:
'Downloads', 'Rank', ...
Integer
Integer
...


Explore Structure

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.