Exploring the **Purposes** and **Processes** of **Data Wrangling**

Michelle Hoda Wilkerson
Associate Professor
University of California, Berkeley

Berkeley
UNIVERSITY OF CALIFORNIA
School of Education

Setting the Context

Increasing calls for data science education globally

(Bargagliotti et al., 2020; Finzer, 2013; Frischemeier et al., 2021; Kazak, 2021; Ridgway, 2016)

Online illustrations by Storyset

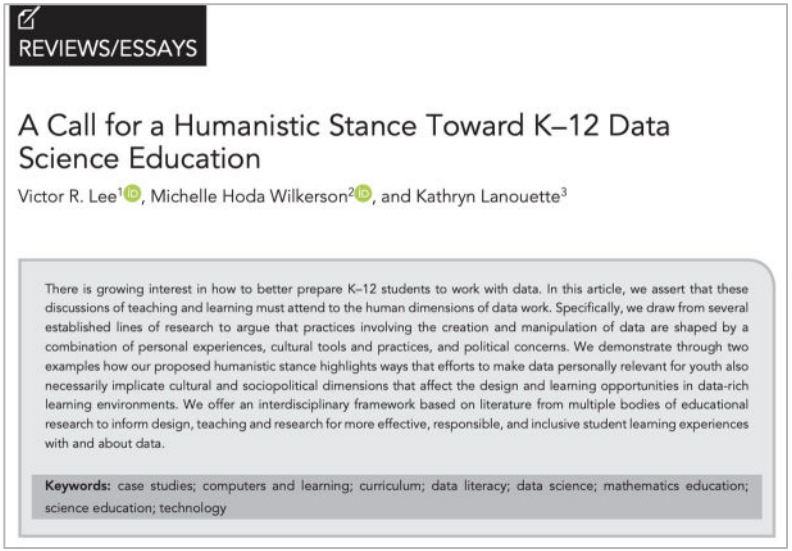
Calls to Broaden the Scope of Data Education

Reasoning / Thinking / Literacy

Data Literacy **Data Acumen** *Data Agency*
Data Fluency Data Literacies
Quantitative Reasoning
Data Science Education *DataViz Literacy*

"...educators, designers, and researchers [should] thoughtfully and systematically consider students' humanistic entanglements with data."

Humanistic Stance



REVIEWS/ESSAYS

A Call for a Humanistic Stance Toward K–12 Data Science Education

Victor R. Lee[1], Michelle Hoda Wilkerson[2], and Kathryn Lanouette[3]

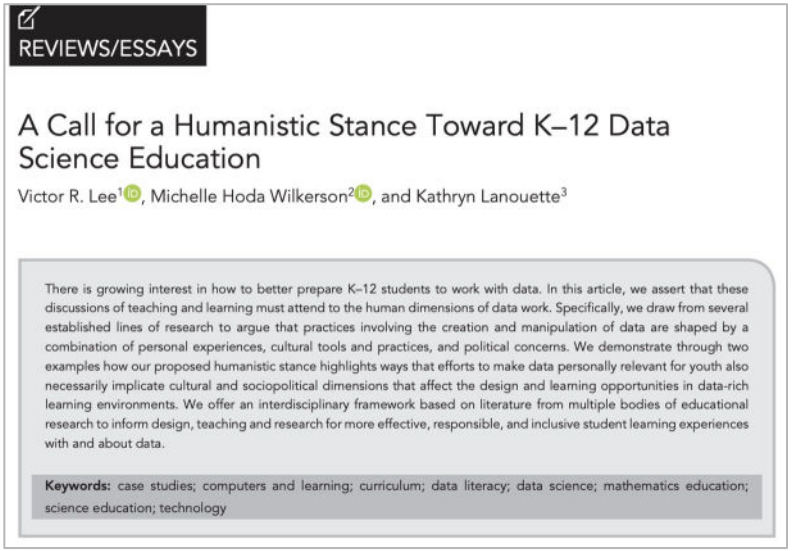
There is growing interest in how to better prepare K–12 students to work with data. In this article, we assert that these discussions of teaching and learning must attend to the human dimensions of data work. Specifically, we draw from several established lines of research to argue that practices involving the creation and manipulation of data are shaped by a combination of personal experiences, cultural tools and practices, and political concerns. We demonstrate through two examples how our proposed humanistic stance highlights ways that efforts to make data personally relevant for youth also necessarily implicate cultural and sociopolitical dimensions that affect the design and learning opportunities in data-rich learning environments. We offer an interdisciplinary framework based on literature from multiple bodies of educational research to inform design, teaching and research for more effective, responsible, and inclusive student learning experiences with and about data.

Keywords: case studies; computers and learning; curriculum; data literacy; data science; mathematics education; science education; technology

"...educators, designers, and researchers [should] thoughtfully and systematically consider students' humanistic entanglements with data."

1. **Students' personal and direct experiences**
2. **Cultural and sociotechnical infrastructures**
3. **Political and social narratives**

Humanistic Stance



REVIEWS/ESSAYS

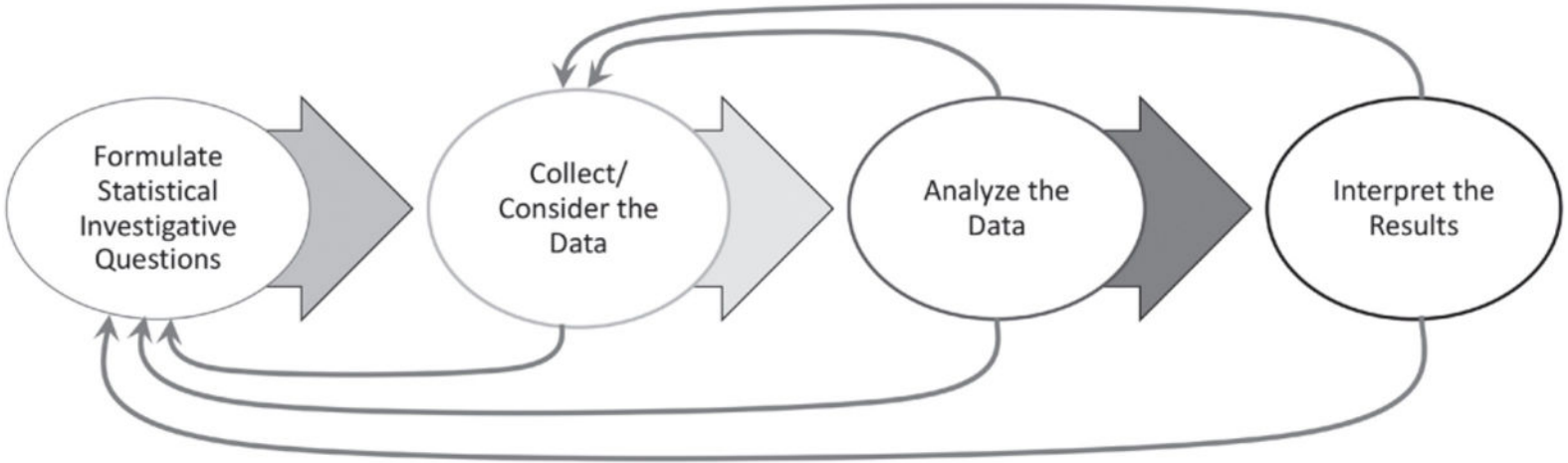
A Call for a Humanistic Stance Toward K–12 Data Science Education

Victor R. Lee[1], Michelle Hoda Wilkerson[2], and Kathryn Lanouette[3]

There is growing interest in how to better prepare K–12 students to work with data. In this article, we assert that these discussions of teaching and learning must attend to the human dimensions of data work. Specifically, we draw from several established lines of research to argue that practices involving the creation and manipulation of data are shaped by a combination of personal experiences, cultural tools and practices, and political concerns. We demonstrate through two examples how our proposed humanistic stance highlights ways that efforts to make data personally relevant for youth also necessarily implicate cultural and sociopolitical dimensions that affect the design and learning opportunities in data-rich learning environments. We offer an interdisciplinary framework based on literature from multiple bodies of educational research to inform design, teaching and research for more effective, responsible, and inclusive student learning experiences with and about data.

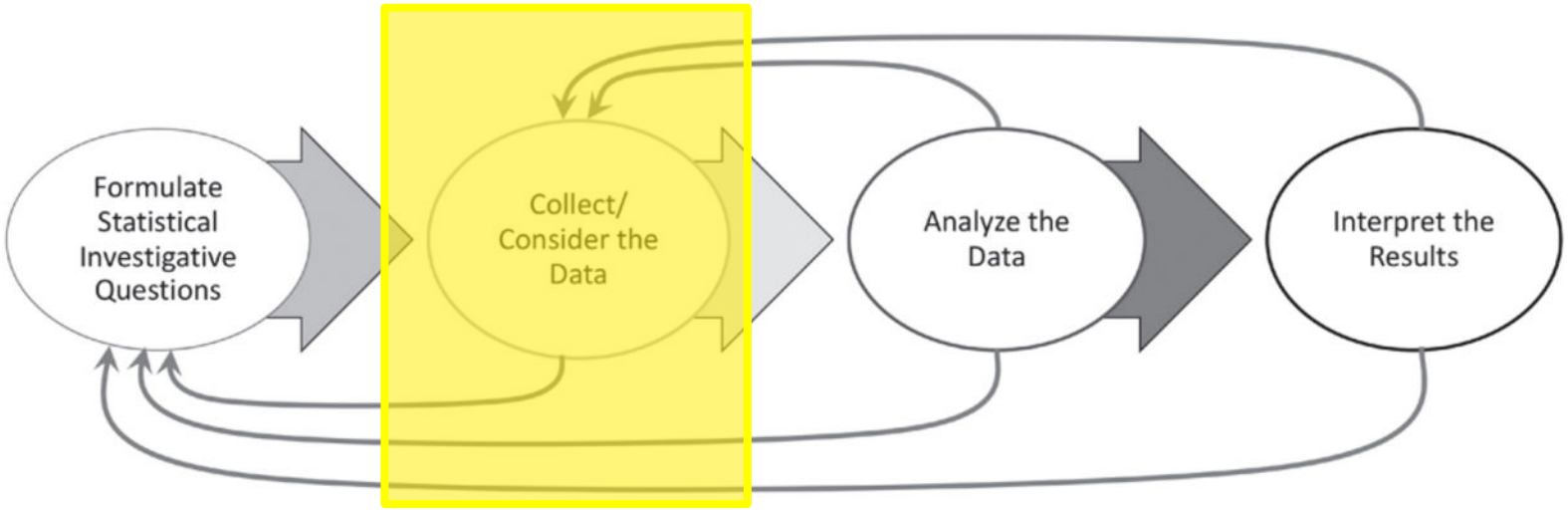
Keywords: case studies; computers and learning; curriculum; data literacy; data science; mathematics education; science education; technology

Conceptual and Computational Skills



GAISE II, 2020

Conceptual and Computational Skills



GAISE II, 2020

Data Preparation

Transnumeration

(Chance, 2006; Wild & Pfannkuch, 1999)

Data Wrangling/ Data Transformation

(Kahn, 2020; Wickham, 2010)

Data Moves

(Erickson et al., 2019)

Guiding Question

It is often taken for granted that focusing on social dimensions of a topic must necessarily short-change skill.

Online illustrations by Storyset

Guiding Question

It is often taken for granted that focusing on social dimensions of a topic must necessarily short-change skill.

How can we **define**, **study**, and **assess** data science education in a way that **balances and integrates** these two important goals?

Online illustrations by Storyset

Plan for Talk

1

Explore

…how goals and data contexts motivate the need for moves

Plan for Talk

1

Explore

...how goals and data contexts motivate the need for moves

2

Analyze

..the process of aligning goal, context, and datasets

Plan for Talk

1
Explore
...how goals and data contexts motivate the need for moves

2
Analyze
..the process of aligning goal, context, and datasets

3
Teach
...using goal and context specific lenses to understand student choices

1

Explore

...how goals and data contexts motivate the need for moves

Activity 1

What do you notice?
What do you wonder?

https://www.nytimes.com/2017/10/09/learning/whats-going-on-in-this-graph-oct-10-2017.html

Context Motivates "Moves"

If I am interested in *a different public's* perceptions of food, I may:

- Reject this dataset.
- Focus only on the foods for which I suspect perceptions are similar across populations (filtering)
- Notice general patterns and consider how those might translate to the next context (grouping)
- Imagine adding, or add, more culturally relevant foods to the dataset (adding/merging records)

Goals Shape Context and "Moves"

If I am interested in….

- … understanding Americans' perceptions of food, this graph is great!
- … evaluating perceptions of some foods you eat, this graph is useful. (filtering)
- … identifying characteristics of unhealthy "health" foods, this graph is useful. (grouping)

- … finding affordable healthy food, this graph is of limited use. (add variables?)
- … understanding global perceptions of food, this graph is of limited use.

2

Analyze

..the process of aligning goal, context, and datasets

Data Wrangling

Goals

Is this dataset useful for my investigative purposes?

Context

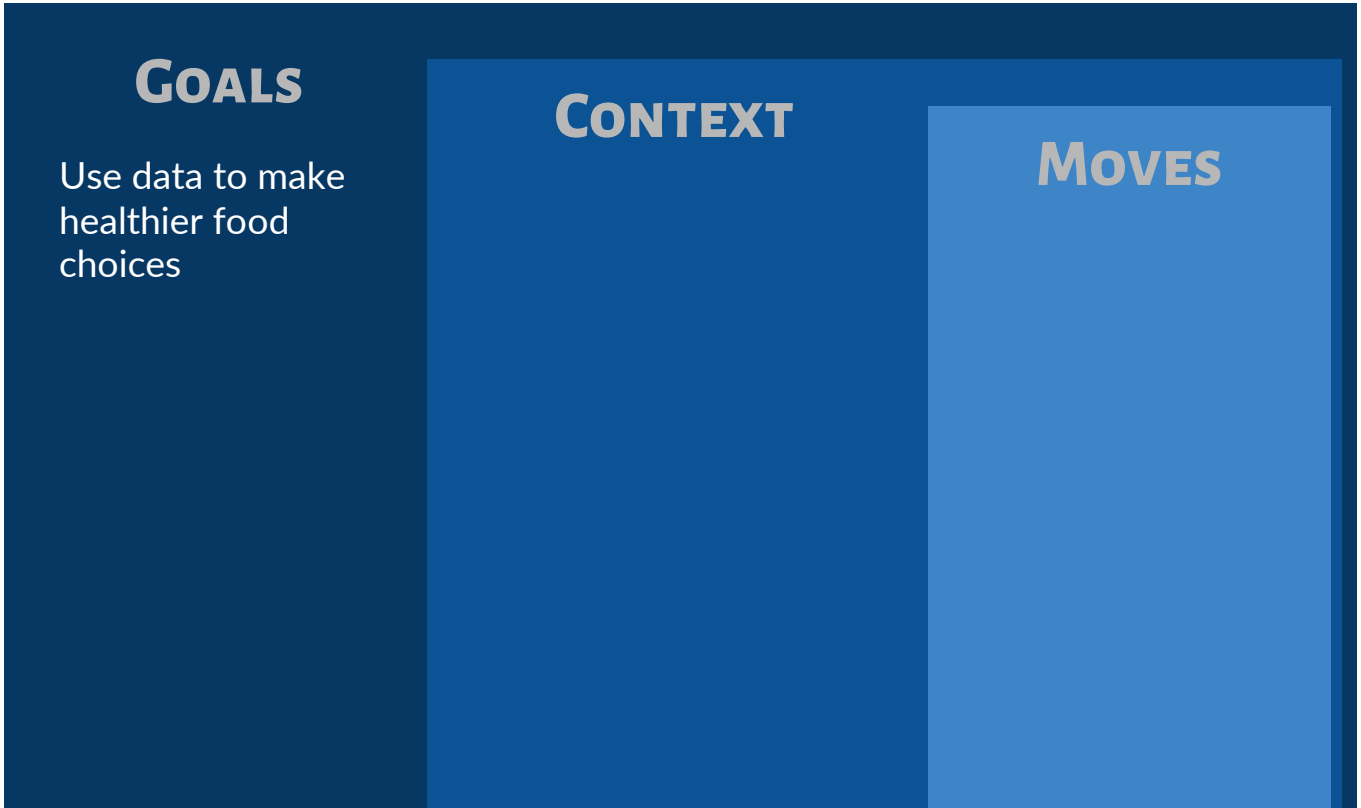
How aligned is this dataset with my goals?

Moves

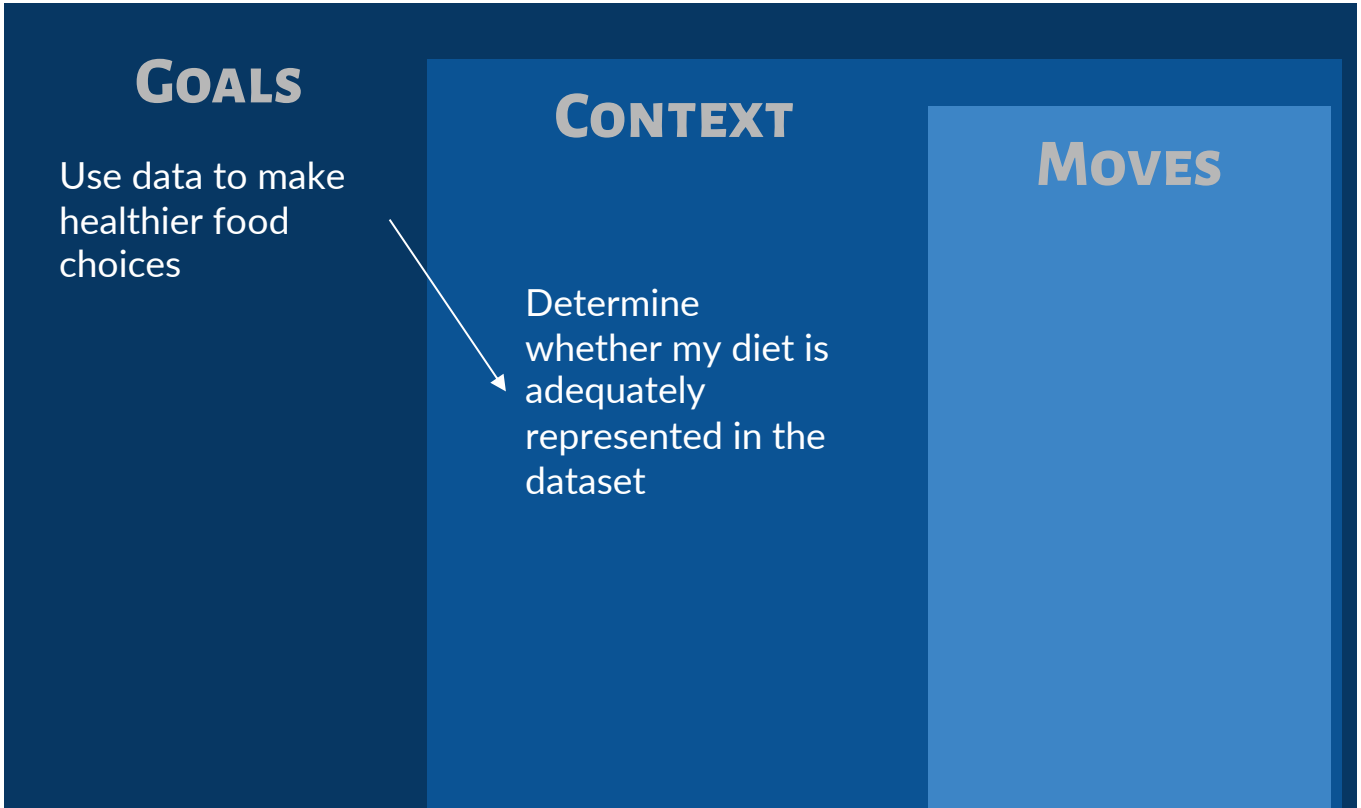
What can I do to make this dataset more aligned with my goals?

Data Wrangling

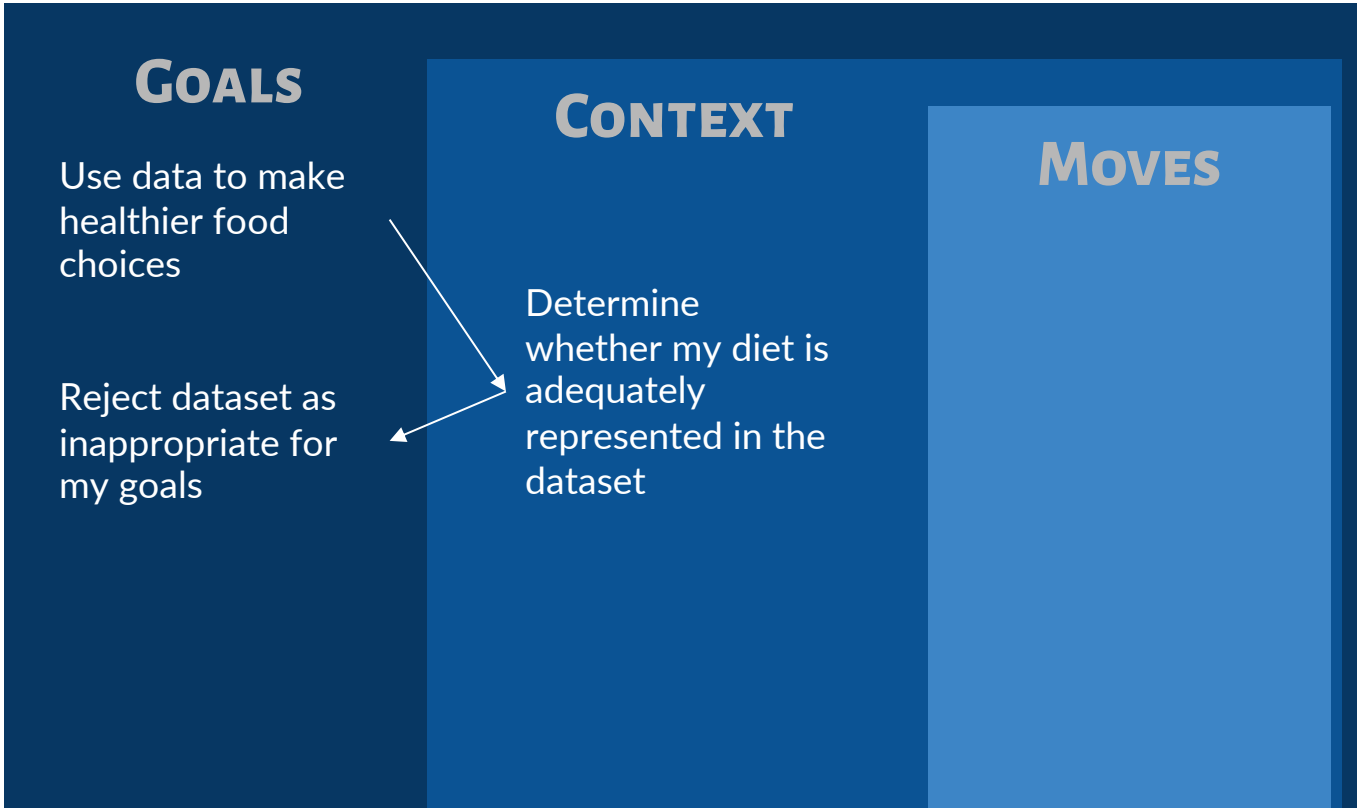
Data Wrangling



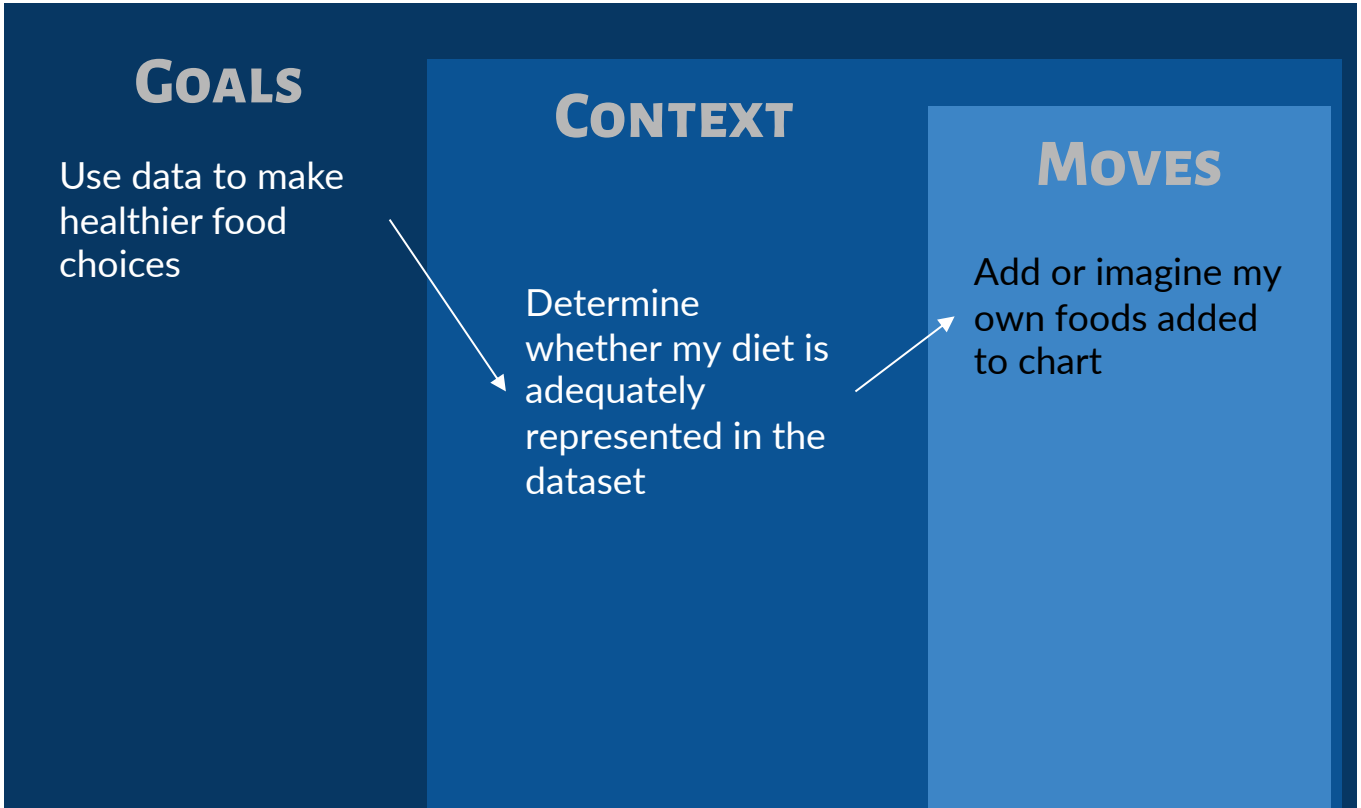
Data Wrangling



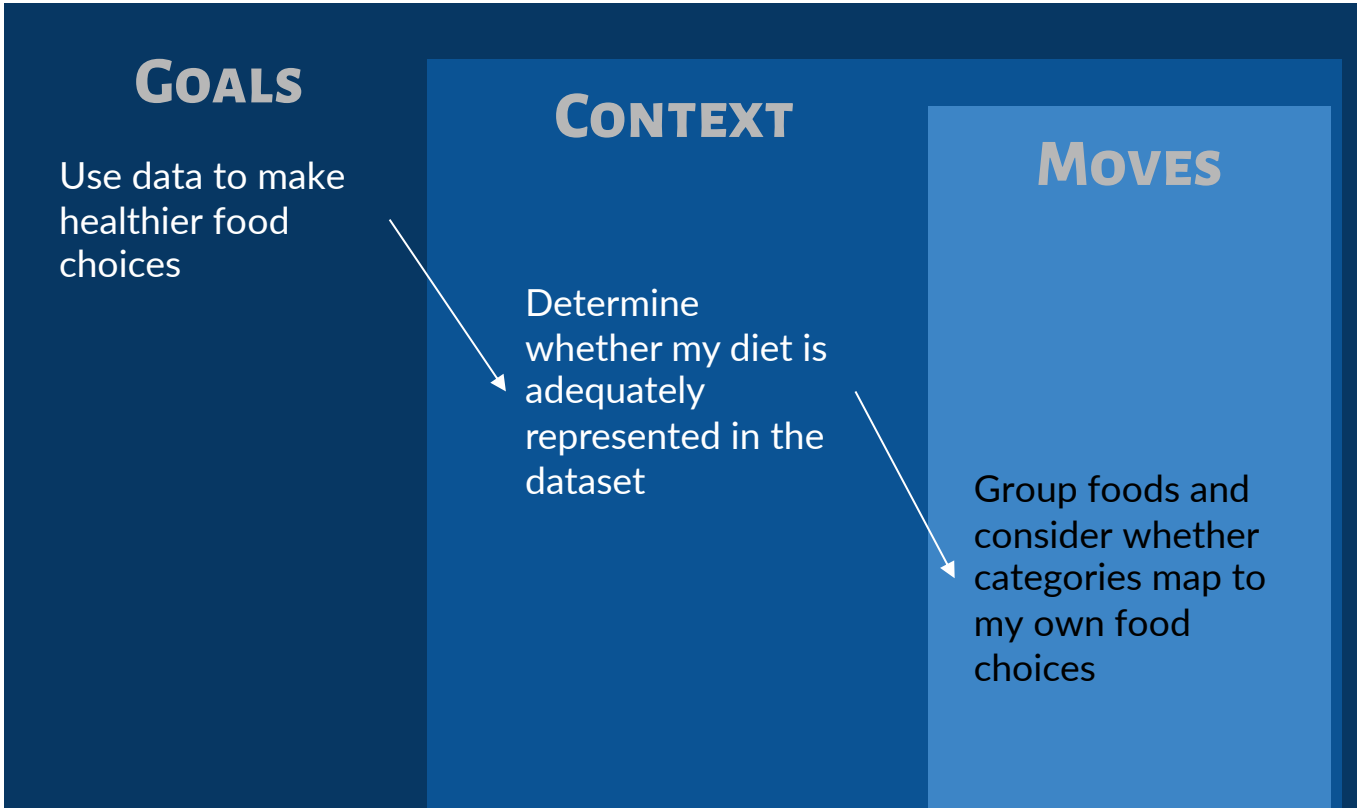
Data Wrangling



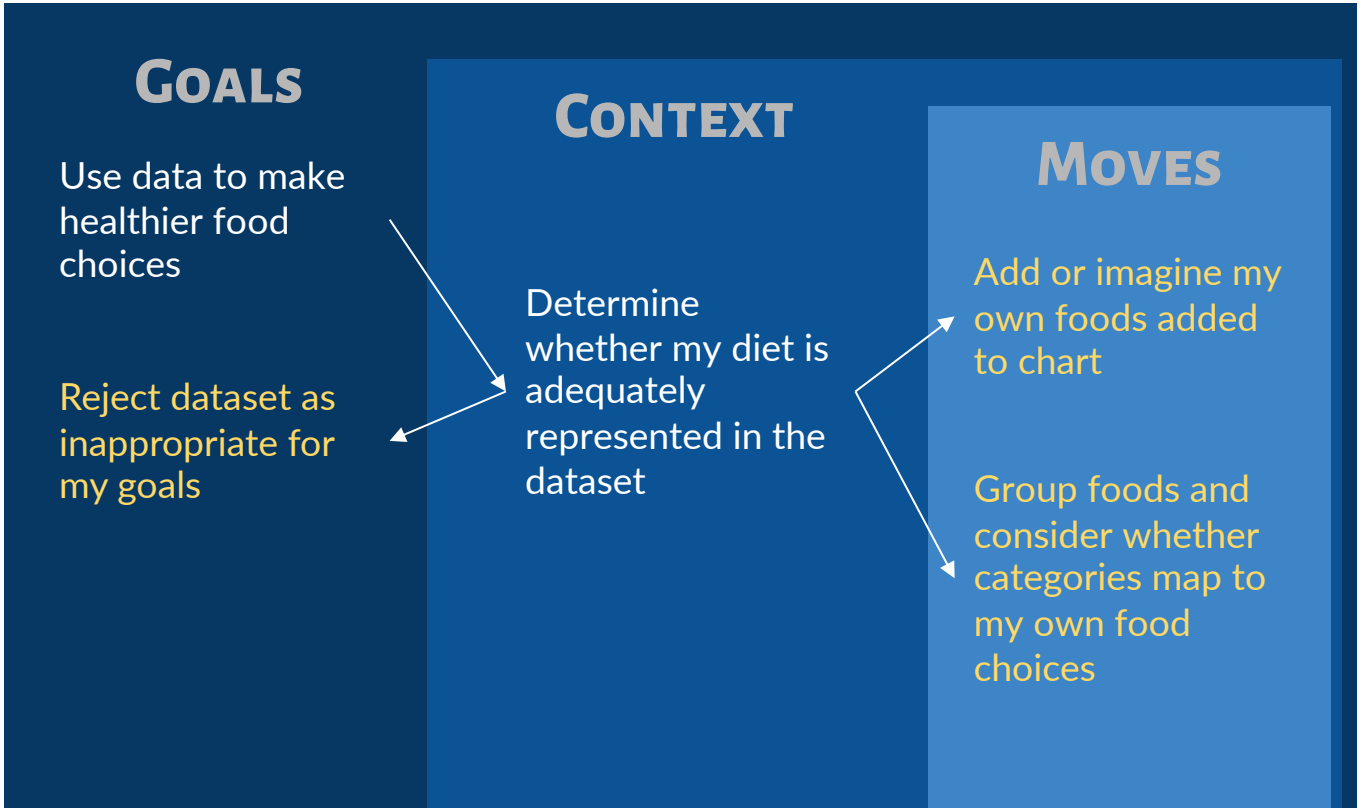
Data Wrangling



Data Wrangling



Data Wrangling



Study Context

Stebbins is a “Data Science Game” developed by Concord Consortium. It links ideas related to evolution, natural selection, data analysis, and gameplay.

https://learn.concord.org/resources/1245/stebbins

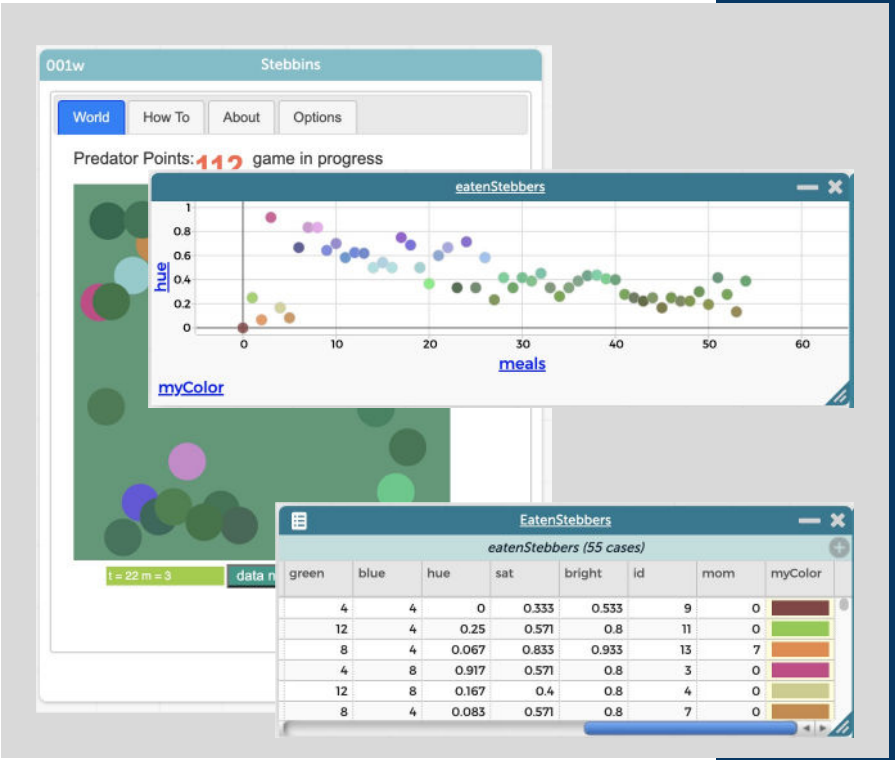
Study Context

Players are a **predator** who must eat to survive. They eat “Stebbins,” creatures whose color is inherited, but with slight mutations. Stebbins move in an environment with poisonous “crud” and a solid colored background.

Study Context

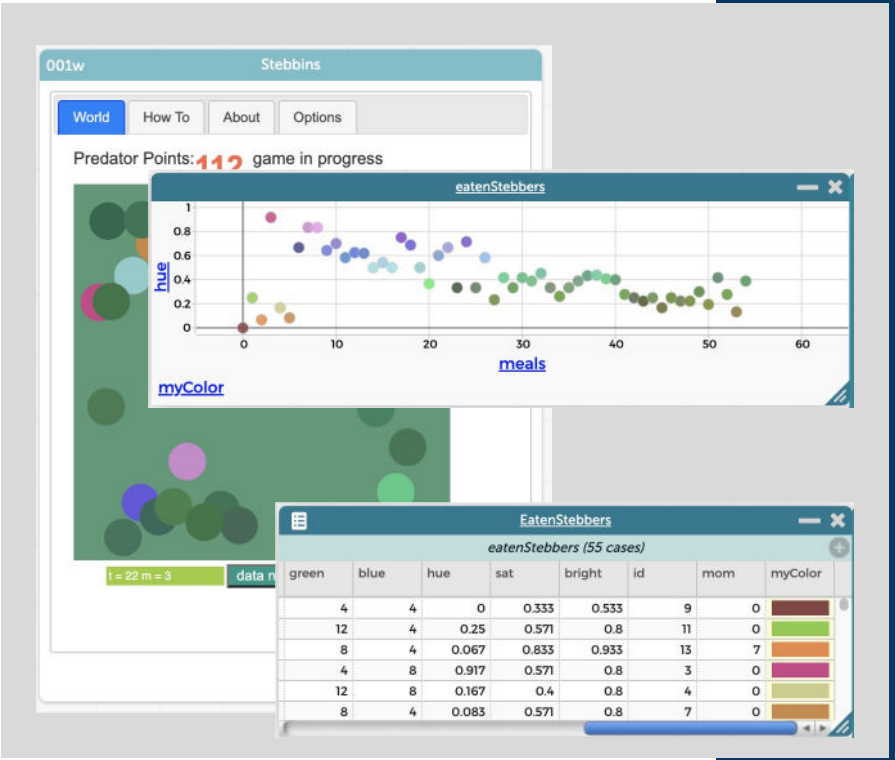
Players are a **predator** who must eat to survive. They eat “Stebbins,” creatures whose color is inherited, but with slight mutations. Stebbins move in an environment with poisonous “crud” and a solid colored background.

They also have access to data, which we hoped would help them analyze underlying scientific principles to optimize gameplay.



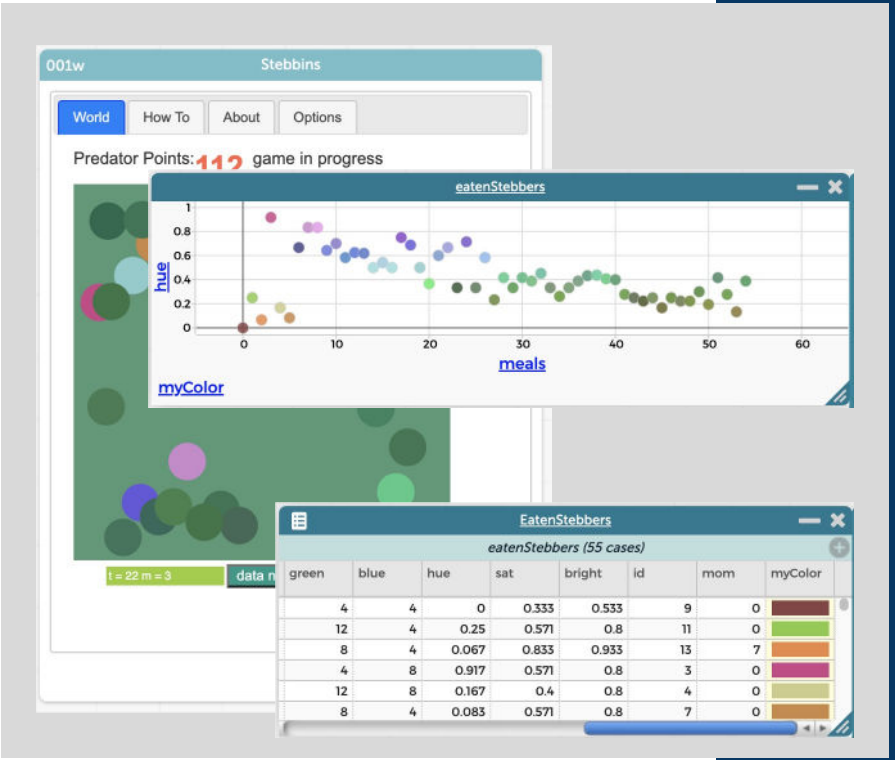
Study Context

We interviewed high school and early university students as they played Stebbins and other similar data-enriched games, and asked them to think aloud as they used data to reason about what they observed.



Study Context

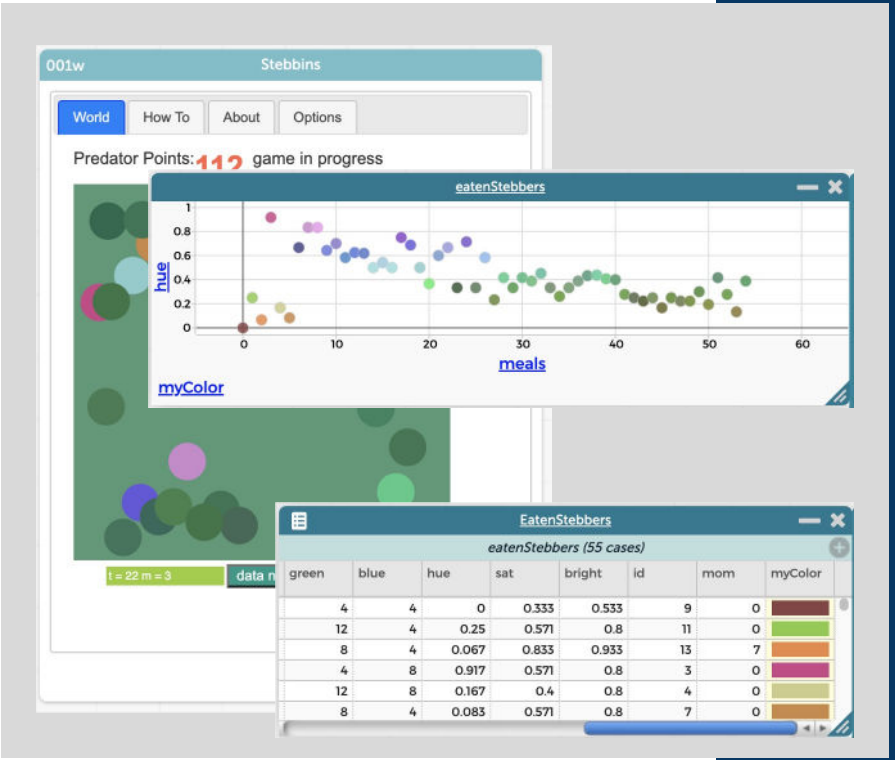
Students engaged with data in unexpected ways, often informed by goals other than bettering gameplay through scientific understanding.



Study Context

Students engaged with data in unexpected ways, often informed by goals other than bettering gameplay through scientific understanding.

To better understand this, we focused on moments during interviews during which students executed or articulated a need data moves.



Study Context

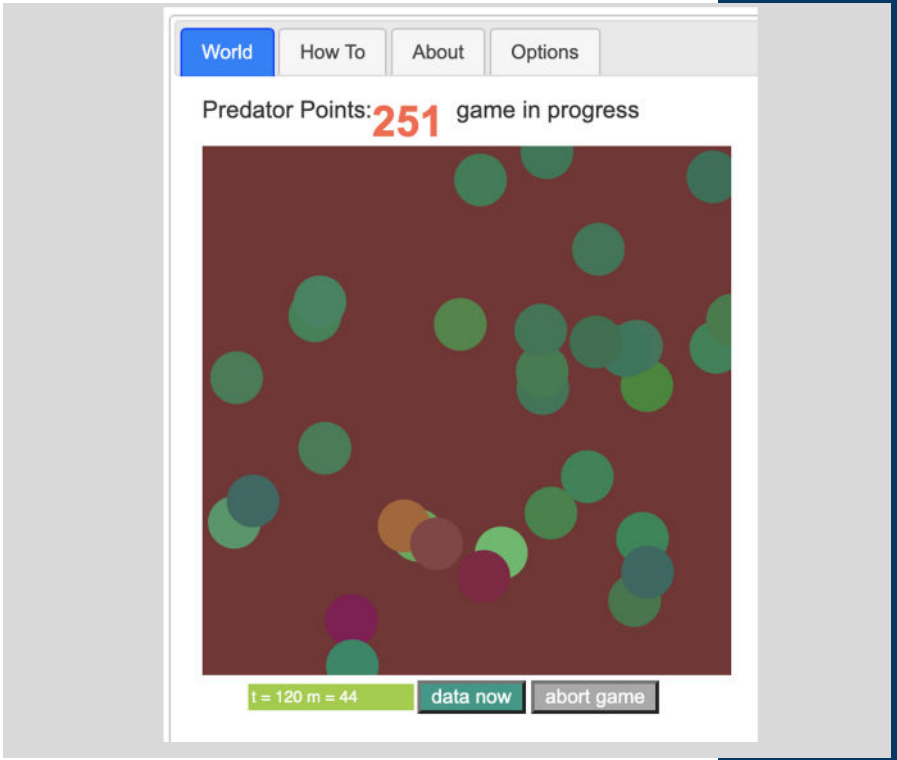
In the clips you are about to see, students have played Stebbins for 2 rounds. In the first round, “Stebbers” converge to the color of the poisonous “crud,” as game players work to avoid getting poisoned by a meal.

This simulates **mimicry**.

Study Context

In the second round, we adjust the background color settings so that Stebbers are equally likely to converge to *either* the color of the poisonous crud, or the color of the background thus rendering them harder to see and catch.

This simulates **mimicry** and **camouflage** simultaneously.



Example 1:
Data Rejected as Irrelevant

Example 1

GOAL

Use data to identify explanatory school biology topic

CONTEXT

DATA MOVE

Example 1

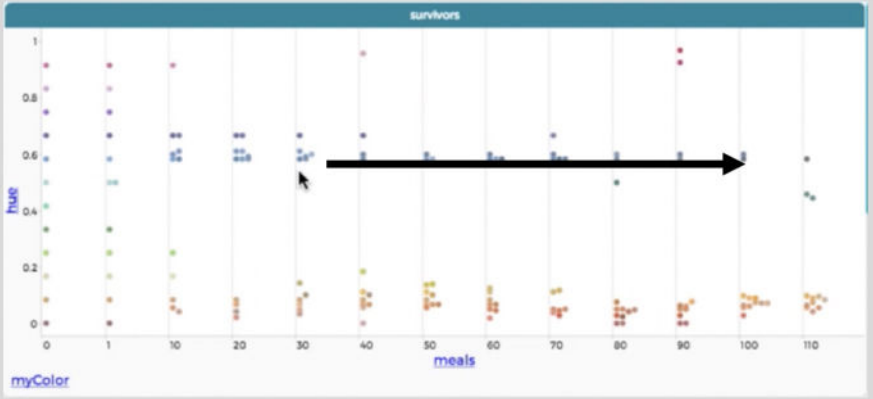
GOAL

Use data to identify explanatory school biology topic

CONTEXT

Convergence of blue color aligns with natural selection and mimicry

DATA MOVE



b) K points to blue cluster of points, then moves mouse quickly to the right.

since these guys who didn't [*look like poisonous crud*] were eaten, and didn't immediately die out, that, this isn't—

Example 1

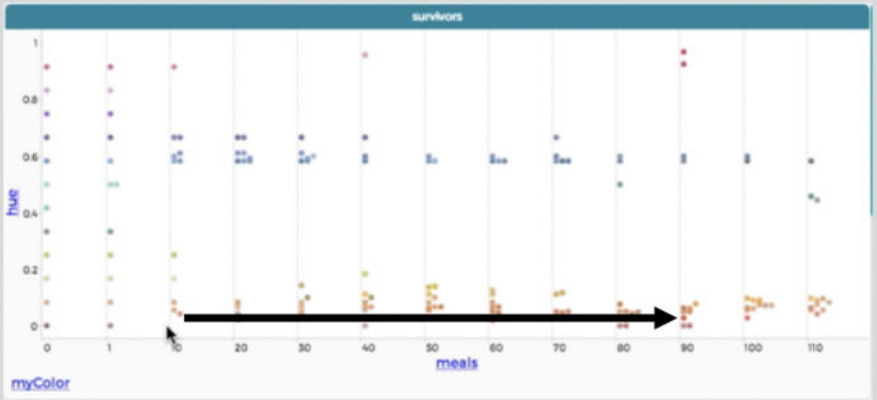
GOAL

Use data to identify explanatory school biology topic

CONTEXT

Simultaneous convergence of orange color did not evidently reflect school topics

DATA MOVE



a) L moves mouse slowly to the right, over orange/yellow data points.

these would have died out completely. So, this is, **this isn't relevant data to an actual population**.

Example 1

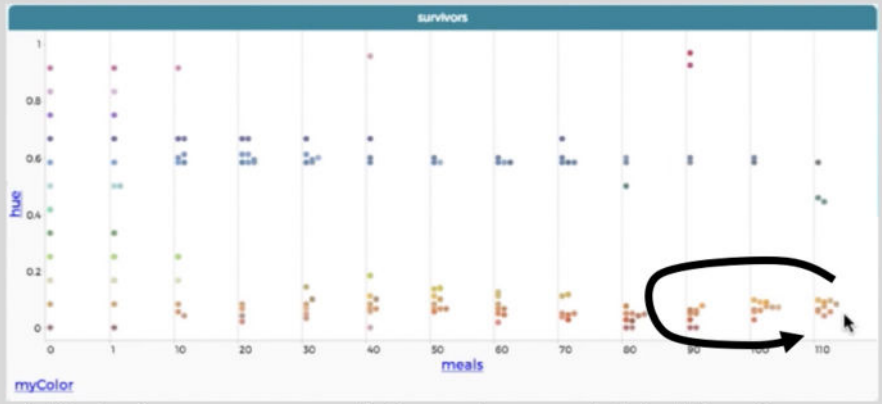
GOAL

Use data to identify explanatory school biology topic

CONTEXT

Simultaneous convergence of orange color did not evidently reflect school topics

DATA MOVE



c) K circles orange points on lower right side of graph.

Example 1

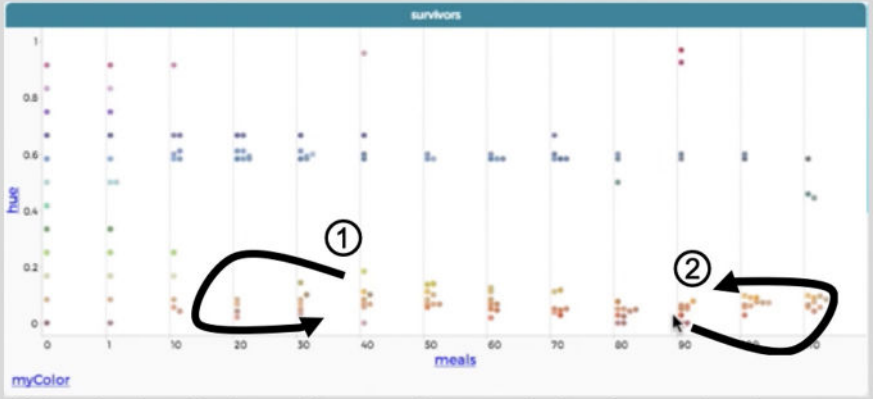
GOAL

Reject data as invalid connection to school topics

CONTEXT

Simultaneous convergence of orange color did not evidently reflect school topics

DATA MOVE



d) L circles light yellow points to left of graph, then darker red points on lower right side of graph.

Example 1

GOAL

Reject data as invalid connection to school topics

CONTEXT

Simultaneous convergence of orange color did not evidently reflect school topics

DATA MOVE

[*no data moves executed*]

Example 2:
Data for Inquiry about Gameplay

Example 2

GOAL

Use Living Stebbers dataset to inform gameplay

CONTEXT

DATA MOVE

Example 2

GOAL

Use Living Stebbers dataset to inform gameplay

CONTEXT

Dataset includes records of score and Stebbers' color; useful for strategy

DATA MOVE

Example 2

GOAL

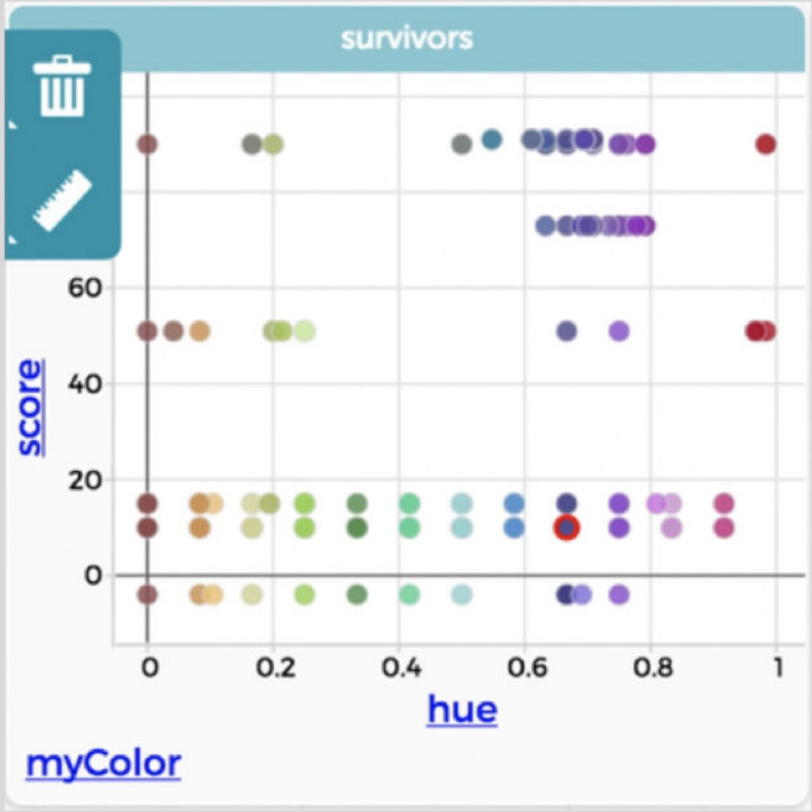
Use Living Stebbers dataset to inform gameplay

CONTEXT

Dataset includes records of score and Stebbers' color; useful for strategy

DATA MOVE

Creates graph of score by color



a) Graph after N* colors points to indicate the color of each surviving Stebber.

Example 2

GOAL

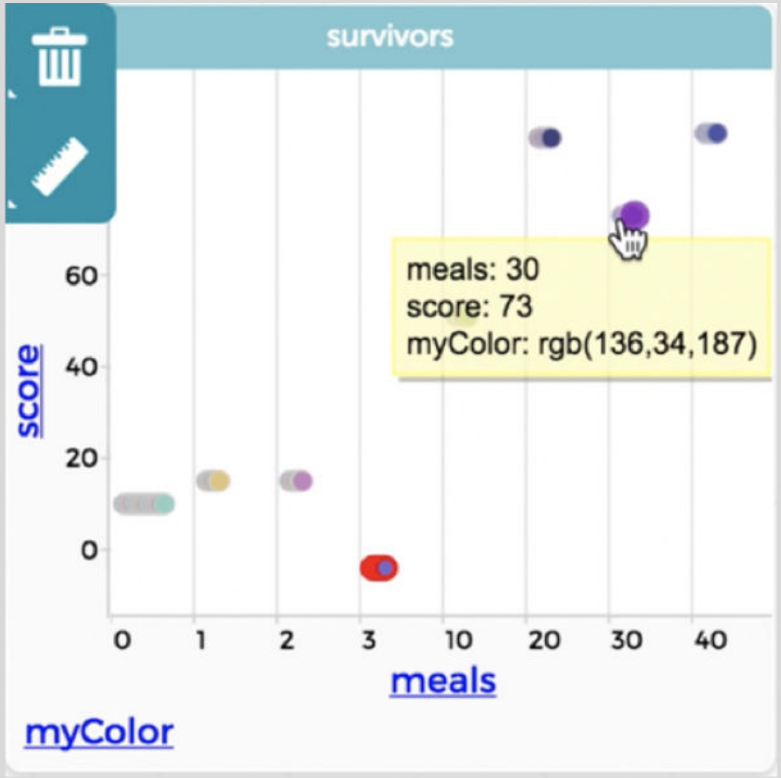
Use Living Stebbers dataset to inform gameplay

CONTEXT

Dataset includes records of score and Stebbers' color; useful for strategy

DATA MOVE

Creates graph of score by # eaten



c) Pam reconfigures graph to explore the relationship between score, color, and number of meals.

Example 2

GOAL

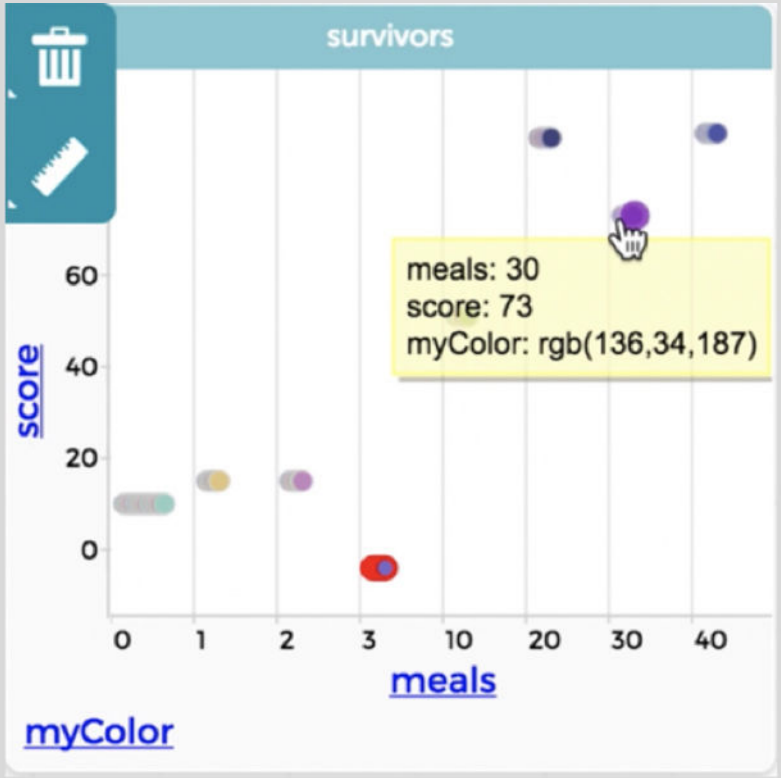
Use Living Stebbers dataset to inform gameplay

CONTEXT

Graph does not match experience; P infers data are not eaten Stebbers

DATA MOVE

Creates graph of score by # eaten



c) Pam reconfigures graph to explore the relationship between score, color, and number of meals.

Example 2

GOAL

Use Eaten Stebbers dataset to inform gameplay

CONTEXT

Graph does not match experience; P infers data are not eaten Stebbers

DATA MOVE

Creates graph of score by # eaten

Example 2

GOAL

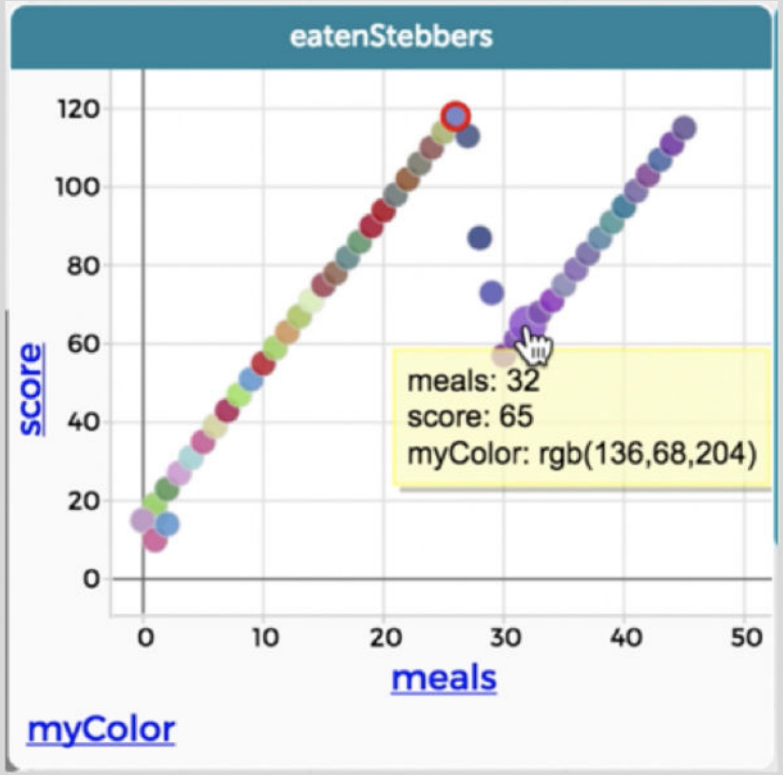
Use Eaten Stebbers dataset to inform gameplay

CONTEXT

Dataset includes records of score and *eaten* Stebbers' color; useful for strategy

DATA MOVE

Creates graph of score, # eaten, color



d) Pam creates eatenStebbers graph to explore relationship between score, color of eaten Stebbers, and meals.

Example 3:
Data for Inquiry about Phenomenon

Example 3

GOAL

Use dataset to understand mechanism for converging to two colors

CONTEXT

DATA MOVE

Example 3

GOAL

Use dataset to understand mechanism for converging to two colors

CONTEXT

Dataset illustrates shifts in hue due to mutation and predation

DATA MOVE

Example 3

GOAL

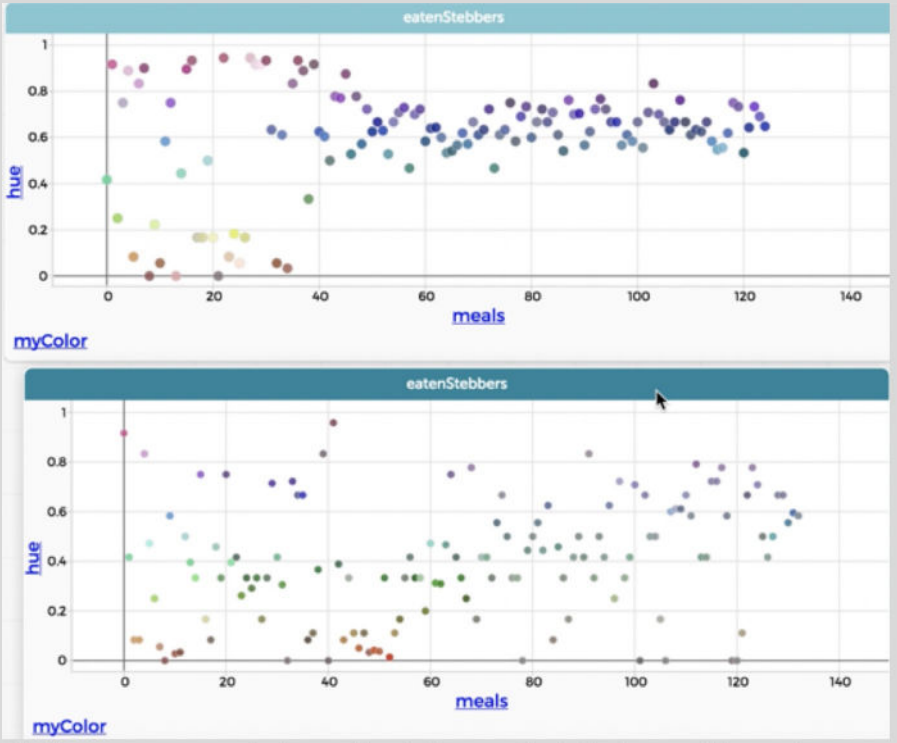
Use dataset to understand mechanism for converging to two colors

CONTEXT

Dataset illustrates shifts in hue due to mutation and predation

DATA MOVE

Reorganize graphs to facilitate comparison



a) J rearranges plots so that Game 1 data is positioned directly above Game 2 data, with “meals” axis aligned.

Example 3

GOAL

Use dataset to understand mechanism for converging to two colors

CONTEXT

Dataset illustrates shifts in hue due to mutation and predation

DATA MOVE

Reorganize graphs to facilitate comparison

b) J gestures to highlight change from scattered points toward convergence, then traces converging points back to their corresponding location on the y axis at approximately 0.6.

Example 3

GOAL

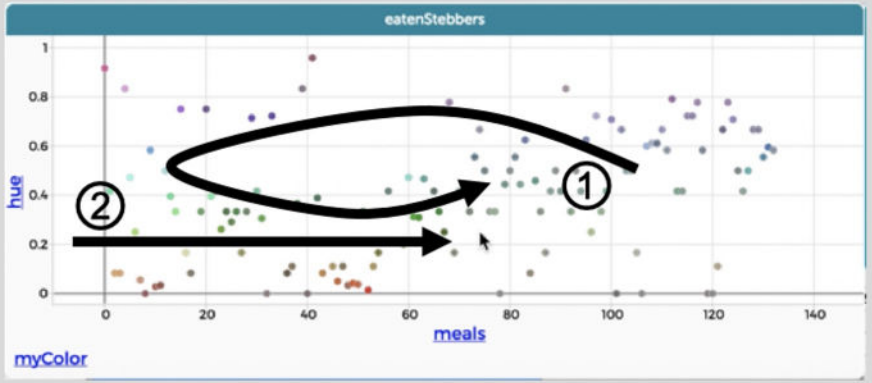
Use dataset to understand mechanism for converging to two colors

CONTEXT

Dataset highlights similarities between converging colors and crud/background

DATA MOVE

Reorganize graphs to facilitate comparison



d) J sweeps mouse across scattering of points, then points to hue value 0.2 on the y axis and extends mouse toward data along that value.

Example 3

GOAL

Use dataset to understand mechanism for converging to two colors

CONTEXT

Dataset highlights similarities between converging colors and crud/background

DATA MOVE

Offline, students construct a ‘difference’ measure, (bg hue) - (Stebbins hue)

Data Wrangling

3

Teach

...using goal and context specific lenses to understand student choices

Assessment

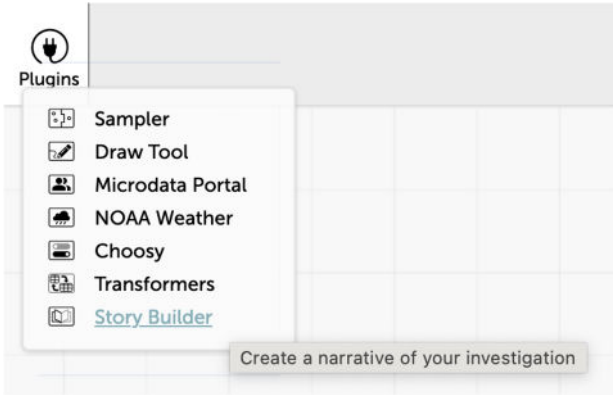
Targeted Support

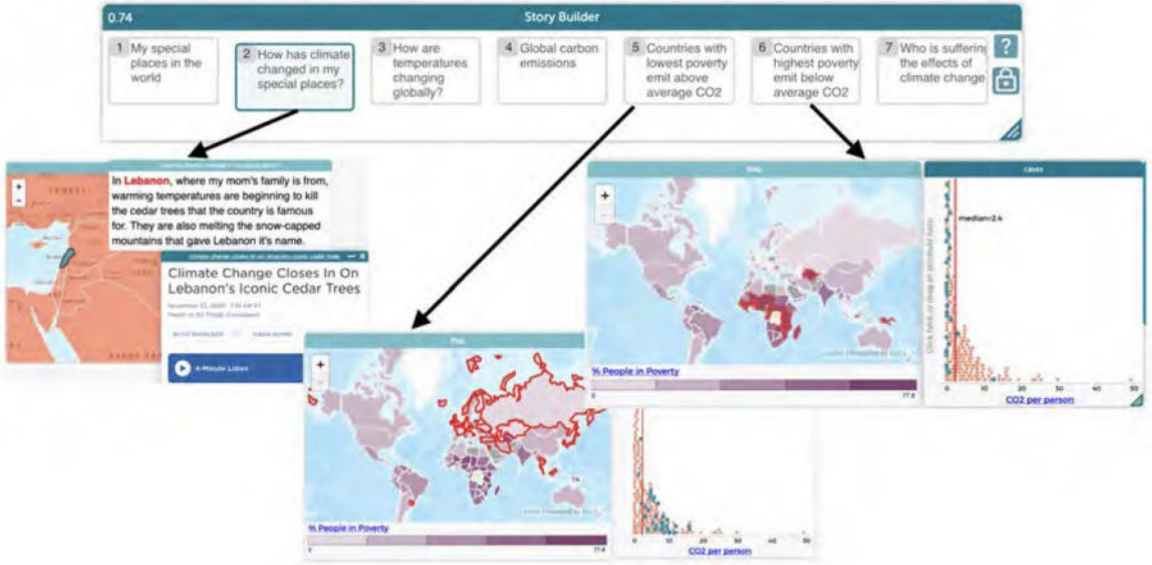
Data WRITING Stories
Choosing Breakfast Food using Data that Reflects What We Care About

Data WRITING Stories
Exploring Climate Through Data and a Special Place

Design

Writing Data Stories





Thank you!

mwilkers@berkeley.edu

Resource Links

https://tinyurl.com/mhwpaderborn
tinyurl.com/wds-units
tinyurl.com/wds-databytes
codap.concord.org/app

Data Science Games Team

Bill Finzer, Tim Erickson, Frieda Reischmann, Natalya St Clair, Nicole Bulalacao, Kathryn Lanouette, Becca Shareff

Writing Data Stories Team

Kris Gutiérrez, Hollylynne Lee, Bill Finzer, Tim Erickson, Lina Haldar, Anthony Petrosino, Cherise McBride, Emily Reigh, Kathryn Lanouette, Lisette Lopez, David Stokes, Krista Cortes, Eddie Rivero, Colette Roberto, Nicole Bulalacao, Jacob Barton, Meg Escudé, Celia Wei, Michael Bakal

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, infographics & images by Freepik

IIS-1900606
IIS-1530578

slidesgo