Why do we teach Data Science?

Rob Gould
rgould@stat.ucla.edu

Paderborn Colloquium on AI and DS Education at School Level

outline

- The primary motivation for the Mobilize Introduction to Data Science course (IDS)
- Overview of IDS
- Other motivations for school level Data Science
- Conflicts

Motivation for this topic

- We are in a data-literacy crisis, and the crisis must be addressed at school level (and beyond)

- School level data science courses may be designed to meet many constraints and serve many purposes.

- The reasons why students will take DS may conflict with the goal to create a more data-literate society.

The need

Perils

- Ignorance-based decision making
- Privacy weakened, loss of autonomy
- Security comromised
- Increased inequity

Promises

- Quality decision-making in the presence of uncertainty
- Improved control over educational and professional career
- Insight into daily lives
- Improved social equity
- Increased autonomy

Mobilize IDS Motivations

IDS

Introduction to Data Science

introdatascience.org

- To provide students with the intellectual and computational tools needed to pose data-driven questions, analyze data to answer, become constructively critical of data-centered arguments, better understand the role of data in their lives and communities.

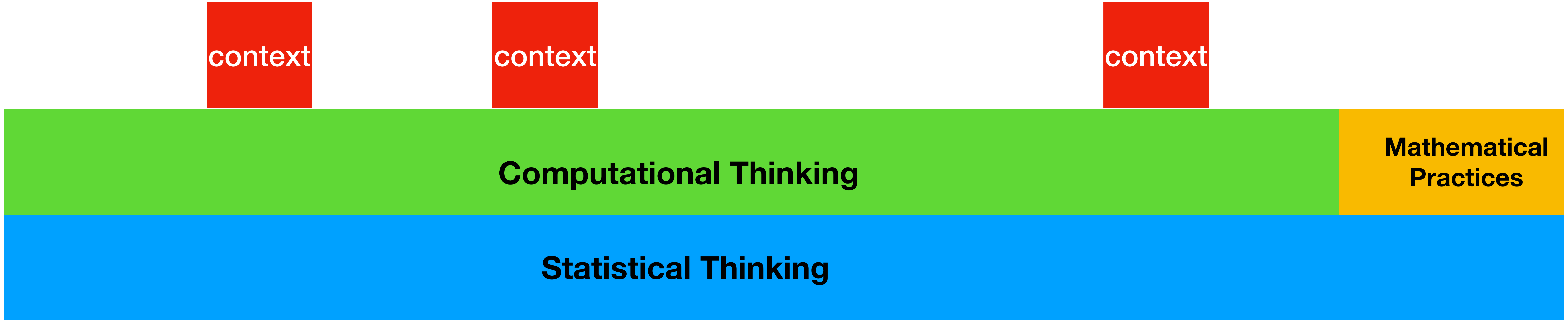
- Optimistic, slightly hidden agenda: engagement with DS will increase interest in STEM education among a diverse group of students, so that more students from under-represented group (women, African-Americans, Latinx) would enter the sciences.

Vision of Data Science

- "Data science is the science and design of (1) actively creating a question to investigate a hypothesis with data, (2) connecting that question with the collection of appropriate data and the application of appropriate methods..., and (3) communicating and making decisions based on....the data and data analysis." - Hicks and Peng (2019)

- A data scientist "must build a bridge between ill-defined questions and unstructured messy data that may (or ma not) be fit to address them, by assessing, cleaning, organizing, integrating and visualizing data, selecting suitable algorithms..., and communicating appropriate inferenes in an ethical manner." --Wise (2020)

IDS Schematic



context

context

context

Computational Thinking

Mathematical Practices

Statistical Thinking

The history of IDS

- Funded by the National Science Foundation to increase numbers of under-represented groups in STEM through engaging in Participatory Sensing

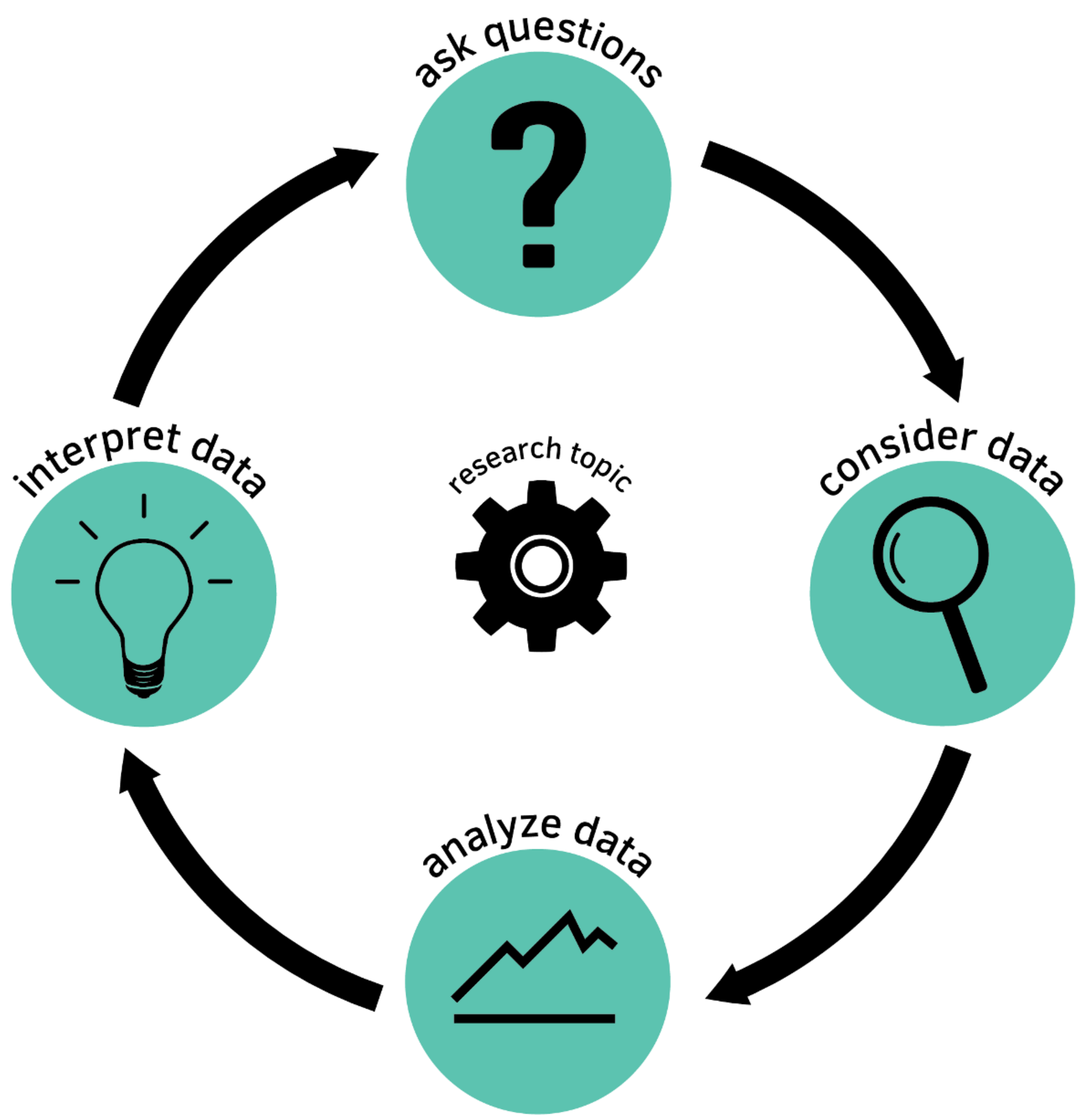
- Partnership between UCLA (Statistics, Computer Science, Education) and Los Angeles Unified School District (LAUSD)

- Based on ASA/NCTM Guidelines for Assessment and Instruction in Statistics Education (GAISE preK-12)

- Key Components: The Data Cycle, Participatory Sensing, R (via Rstudio)

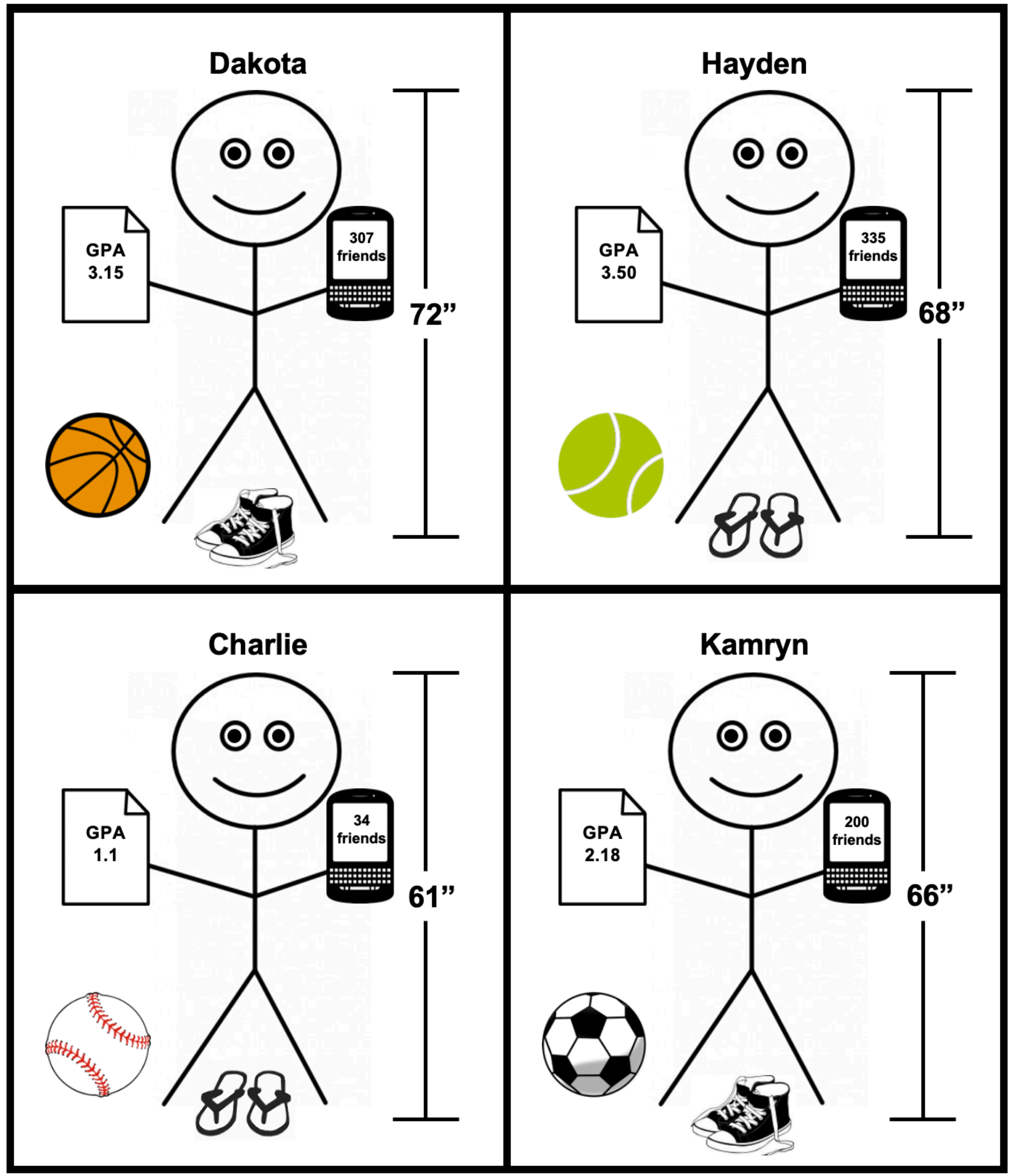
The Statistical Investigation Cycle is at the Foundation of IDS

The Data Cycle



based on the GAISE Statistical Investigation Process and the PPDAC (Wild & Pfannkuch 1999)

Unit 1, Lesson 2: Stick Figures

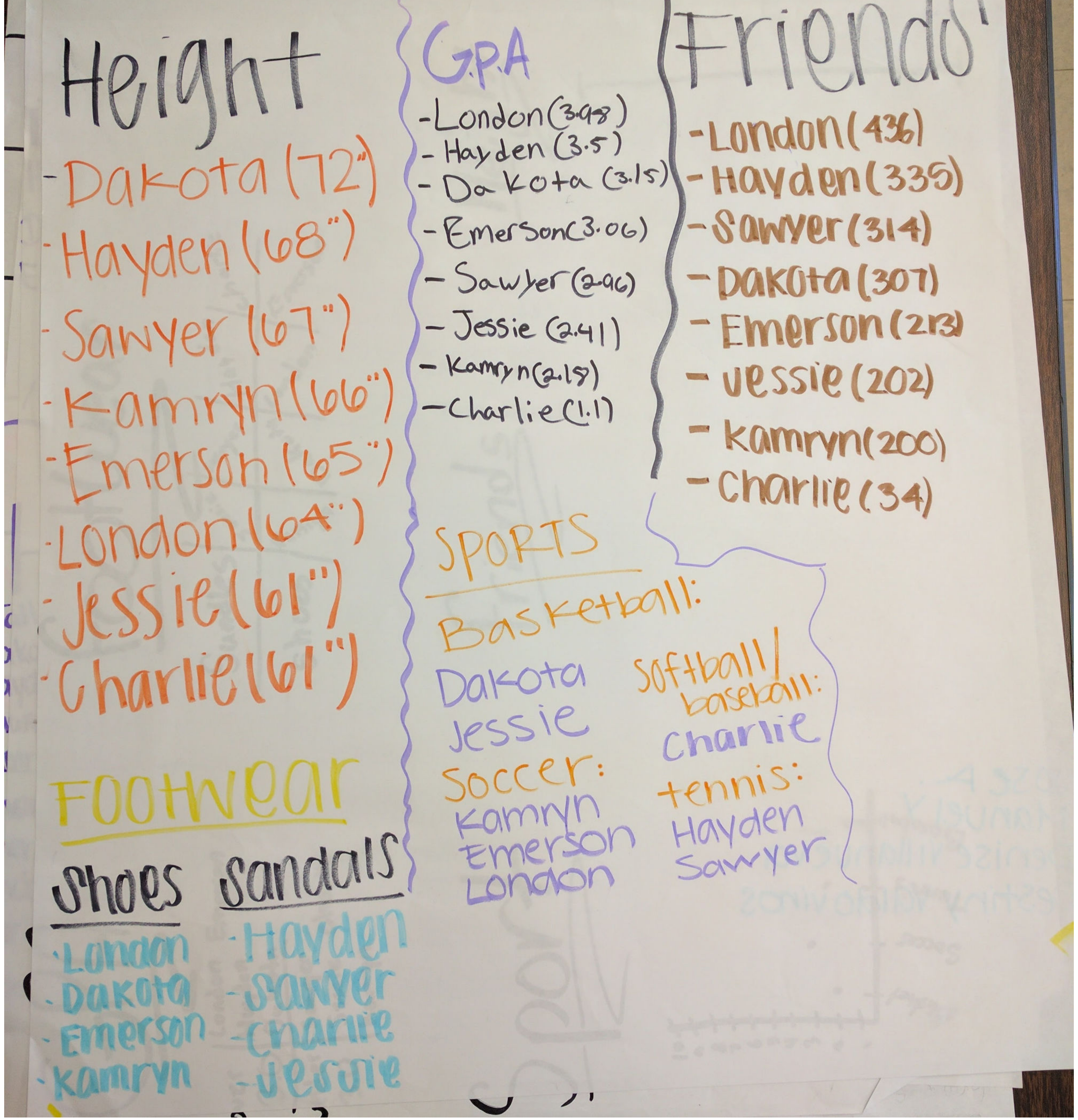


"Collect and record as much information as you can about these people"

"Organize this information on a poster any way that you feel is helpful"

Posters are displayed, and students discuss:

- what are similarities and differences in the ways the data were organized
- what information ('variables') is available?
- which organizations made it easiest to see the variables?



Height

- Dakota (72")
- Hayden (68")
- Sawyer (67")
- Kamryn (66")
- Emerson (65")
- London (64")
- Jessie (61")
- Charlie (61")

G.P.A

- London (3.98)
- Hayden (3.5)
- Dakota (3.15)
- Emerson (3.06)
- Sawyer (2.96)
- Jessie (2.41)
- Kamryn (2.18)
- Charlie (1.1)

Friends

- London (436)
- Hayden (335)
- Sawyer (314)
- Dakota (307)
- Emerson (213)
- Jessie (202)
- Kamryn (200)
- Charlie (34)

SPORTS

Basketball:
Dakota
Jessie

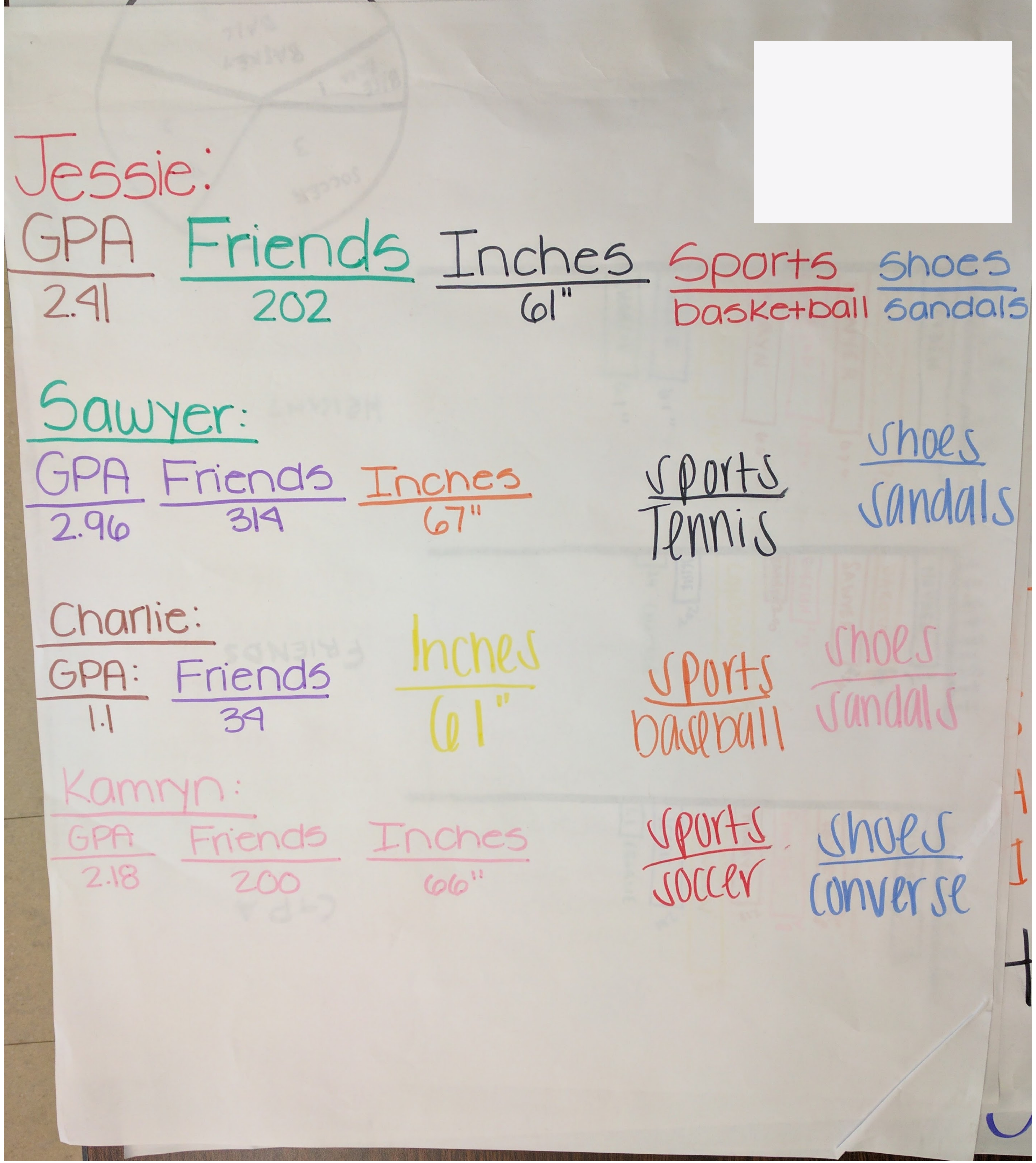
Soccer:
Kamryn
Emerson
London

softball/baseball:
Charlie

tennis:
Hayden
Sawyer

Footwear

Shoes	Sandals
London	Hayden
Dakota	Sawyer
Emerson	Charlie
Kamryn	Jessie



Jessie:

GPA	Friends	Inches	Sports	Shoes
2.41	202	61"	basketball	sandals

Sawyer:

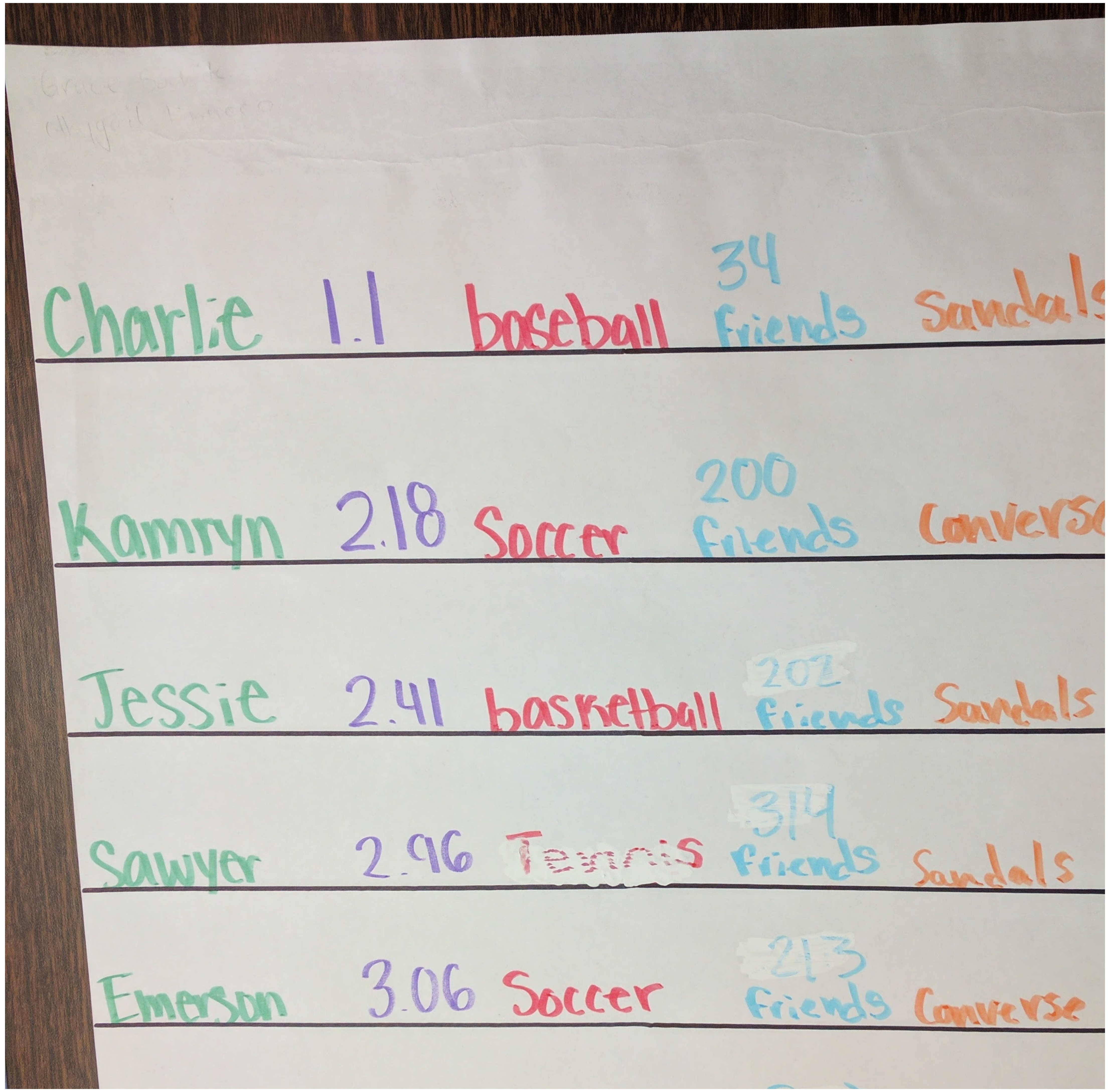
GPA	Friends	Inches	Sports	Shoes
2.96	314	67"	Tennis	Sandals

Charlie:

GPA:	Friends	Inches	Sports	Shoes
1.1	34	61"	baseball	sandals

Kamryn:

GPA	Friends	Inches	Sports	Shoes
2.18	200	66"	soccer	converse



Grace [illegible]
Abigail [illegible]

Charlie	1.1	baseball	34 friends	Sandals
Kamryn	2.18	Soccer	200 friends	Convers
Jessie	2.41	basketball	202 friends	Sandals
Sawyer	2.96	Tennis	314 friends	Sandals
Emerson	3.06	Soccer	213 friends	Converse

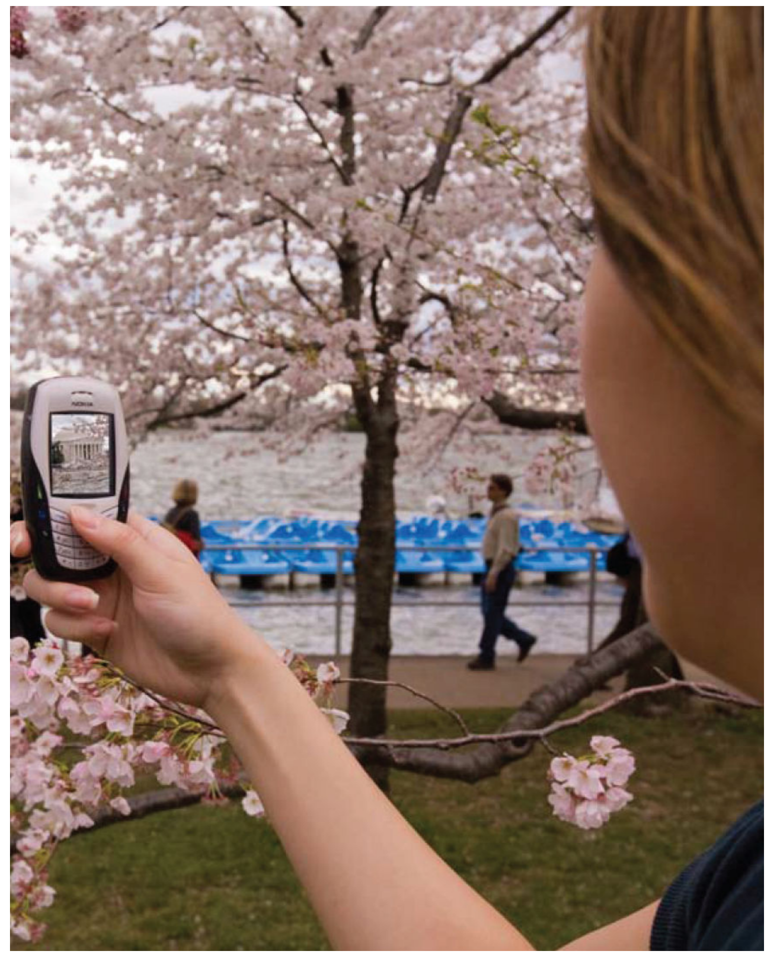
Konold, Finzer, Kreetong (2016)

- **"spreadsheet" format is not natural for many students (and their teachers)**
- **students need to develop the conception of "case"**
- **students have basically sound and solid notions of data**
- **students are comfortable and may even prefer hierarchical representations over spreadsheet representations**

Participatory Sensing

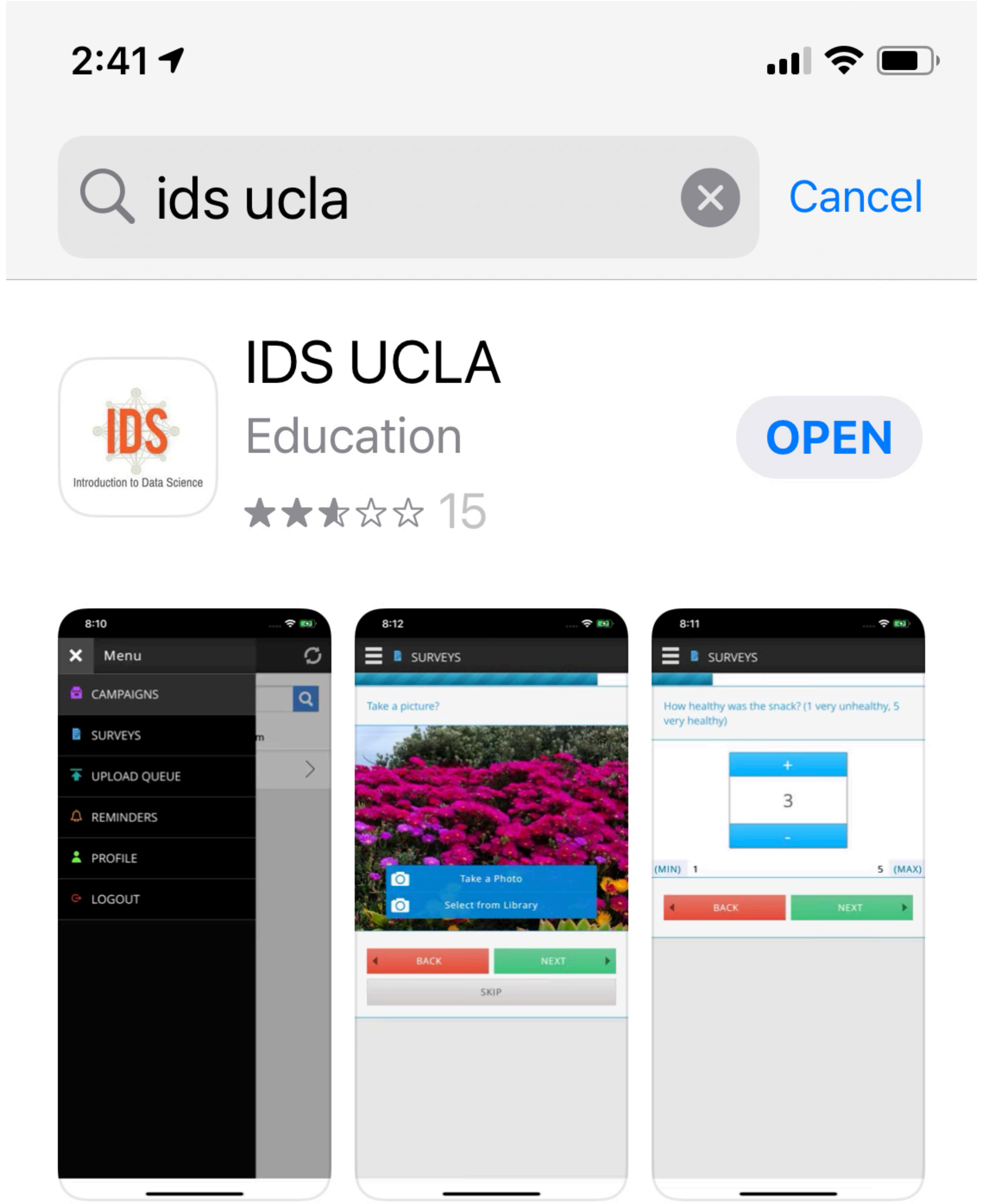
- A data-collection paradigm developed by Deborah Estrin's lab at UCLA (Center for Embedded Network Sensing)
- Students engage in participatory sensing campaigns.
- Mobile devices used to collect data to address various issues: Nutrition, recycling, stress, water conservation
- Students collect numbers, images, words, locations, times, dates.
- They are "human sensors", collecting a stream of data based on triggers, and not random samples.

J. Burke, D. Estin, M. Hansen, A. Parker, N. Ramanathan, M. Reddy, M.B. Srivastava, Participatory Sensing. *Center for Embedded Network Sensing*. (2006).



snack campaign

- Motivating Questions:
    - What is my snacking pattern?
    - How good am I at rating the healthiness of my snack?
    - Do I tend to eat healthy? How does this compare to the rest of my class?
    - Does knowing nutritional value change my habits?
- Data collection: Collect data every time you eat a snack for the next four days.



ids ucla

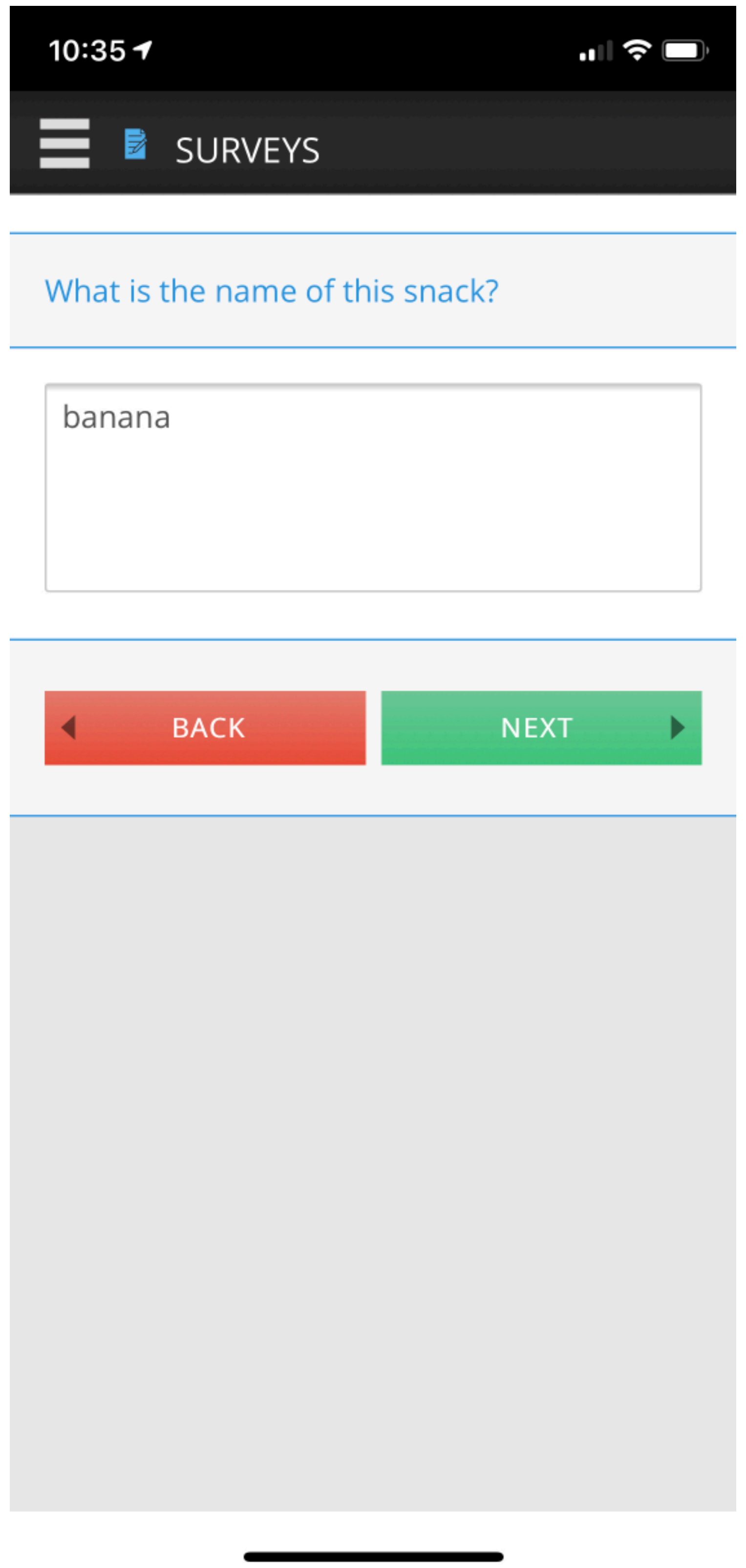
IDS UCLA

Education

★★★☆☆ 15

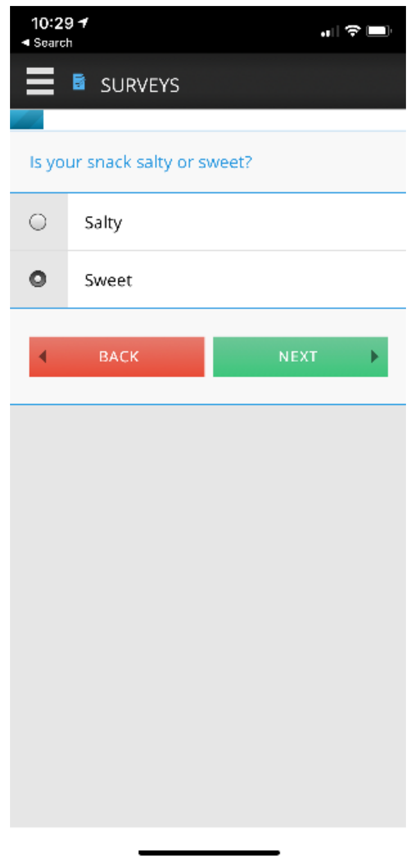
Take a picture?

How healthy was the snack? (1 very unhealthy, 5 very healthy)



What is the name of this snack?

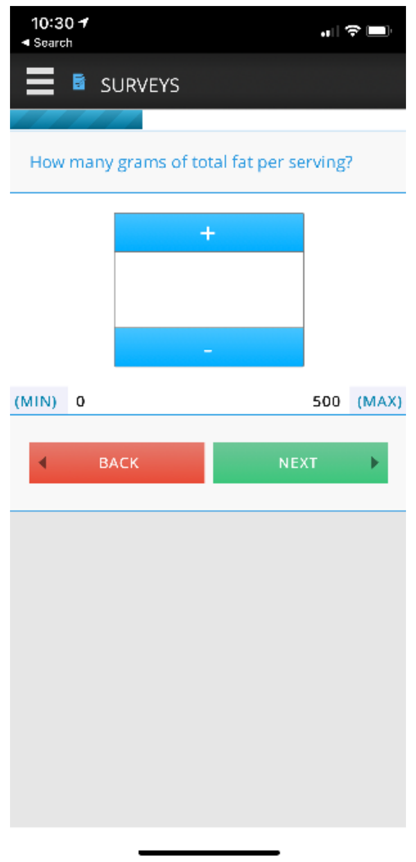
banana



Is your snack salty or sweet?

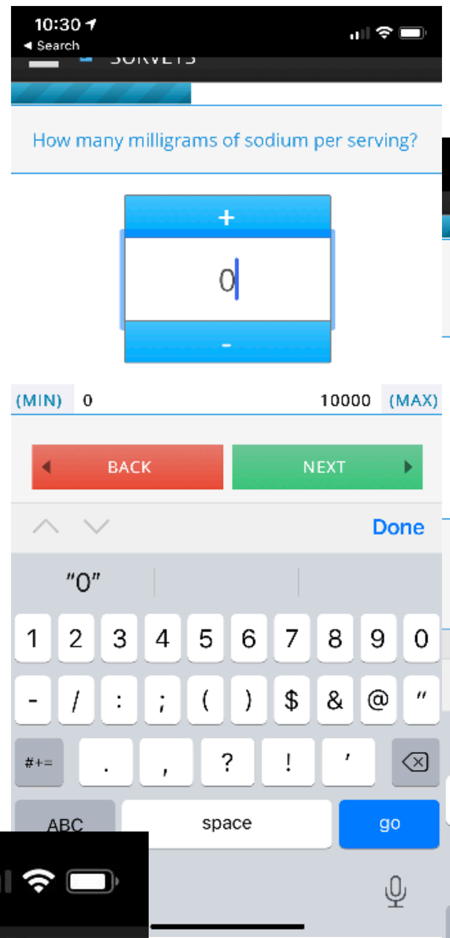
○ Salty

● Sweet



How many grams of total fat per serving?

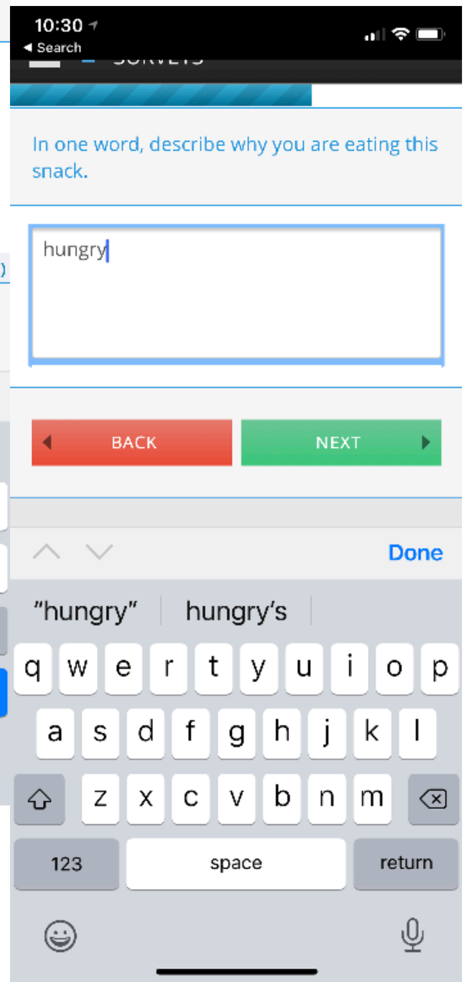
(MIN) 0 500 (MAX)



How many milligrams of sodium per serving?

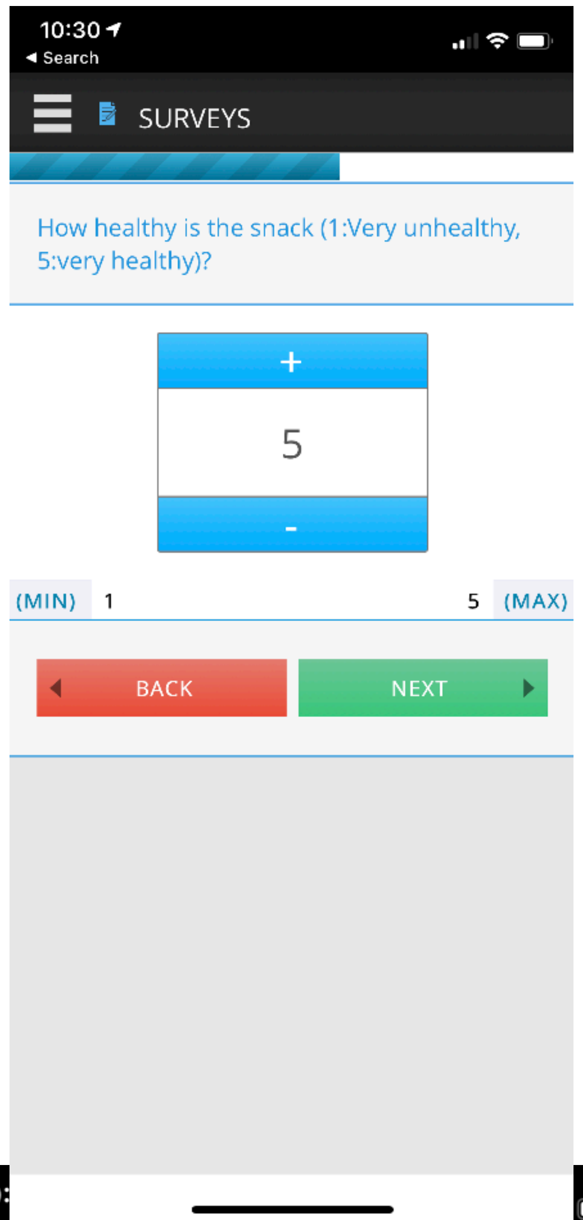
0

(MIN) 0 10000 (MAX)



In one word, describe why you are eating this snack.

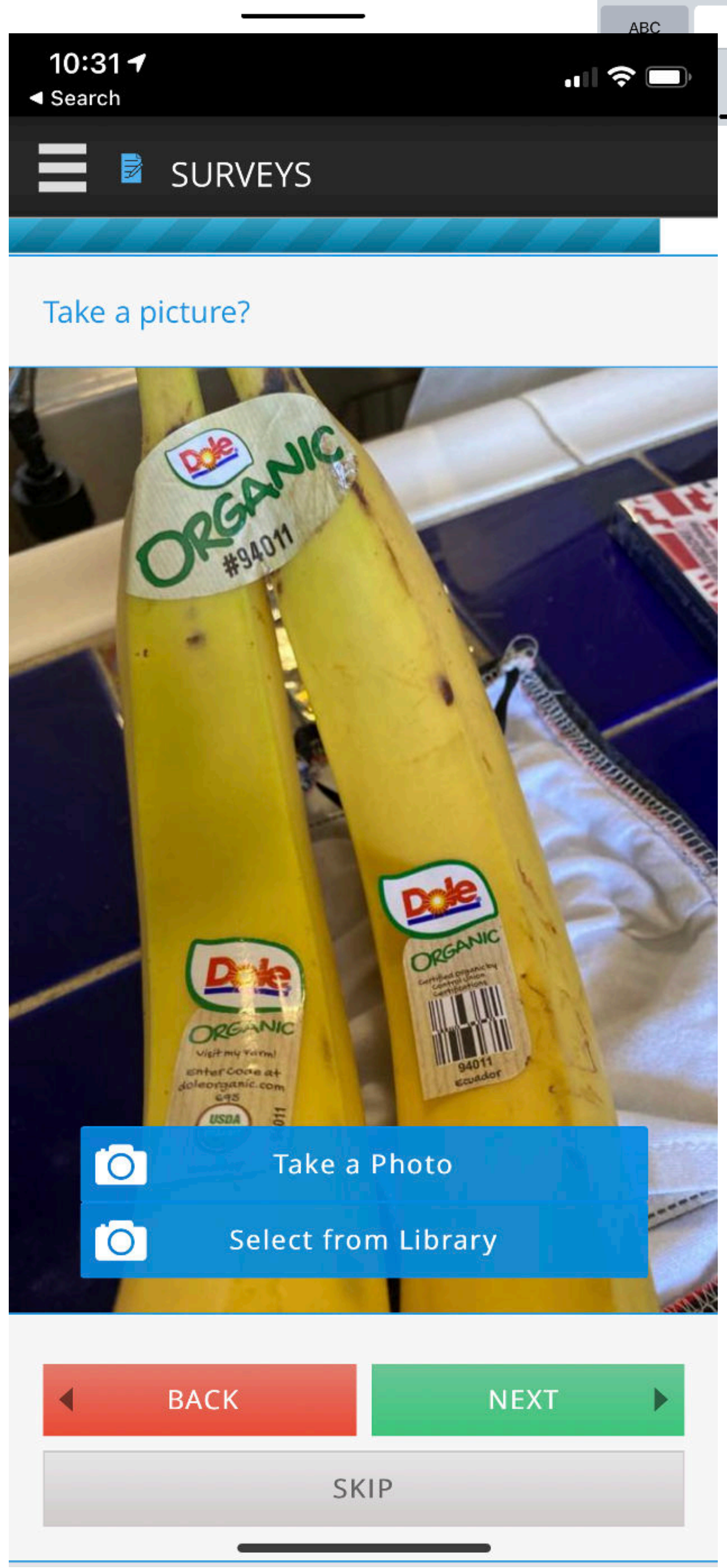
hungry



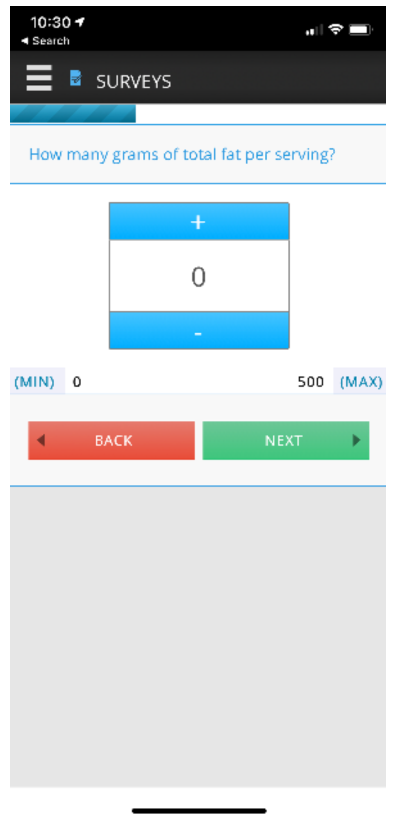
How healthy is the snack (1:Very unhealthy, 5:very healthy)?

5

(MIN) 1 5 (MAX)



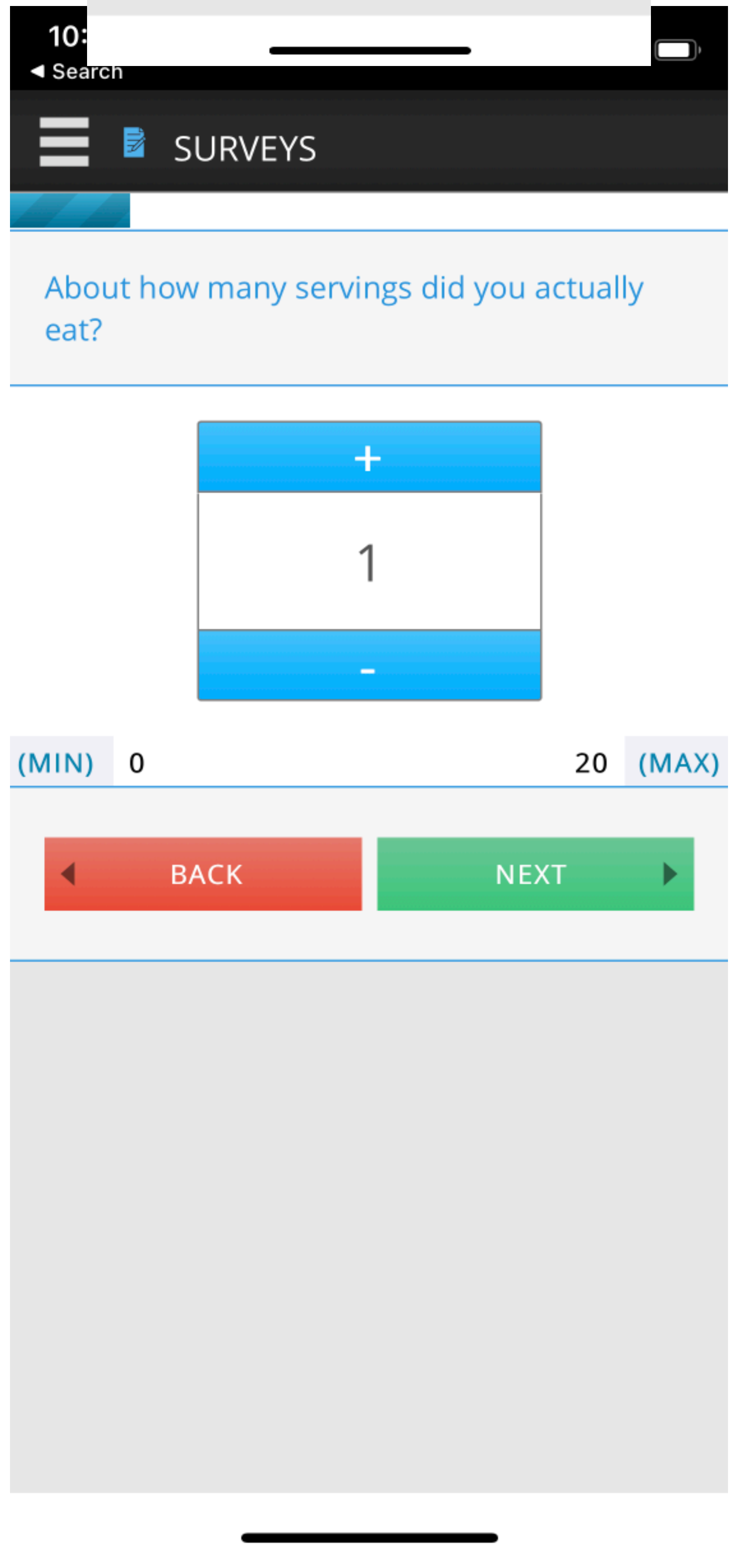
Take a picture?



How many grams of total fat per serving?

0

(MIN) 0 500 (MAX)

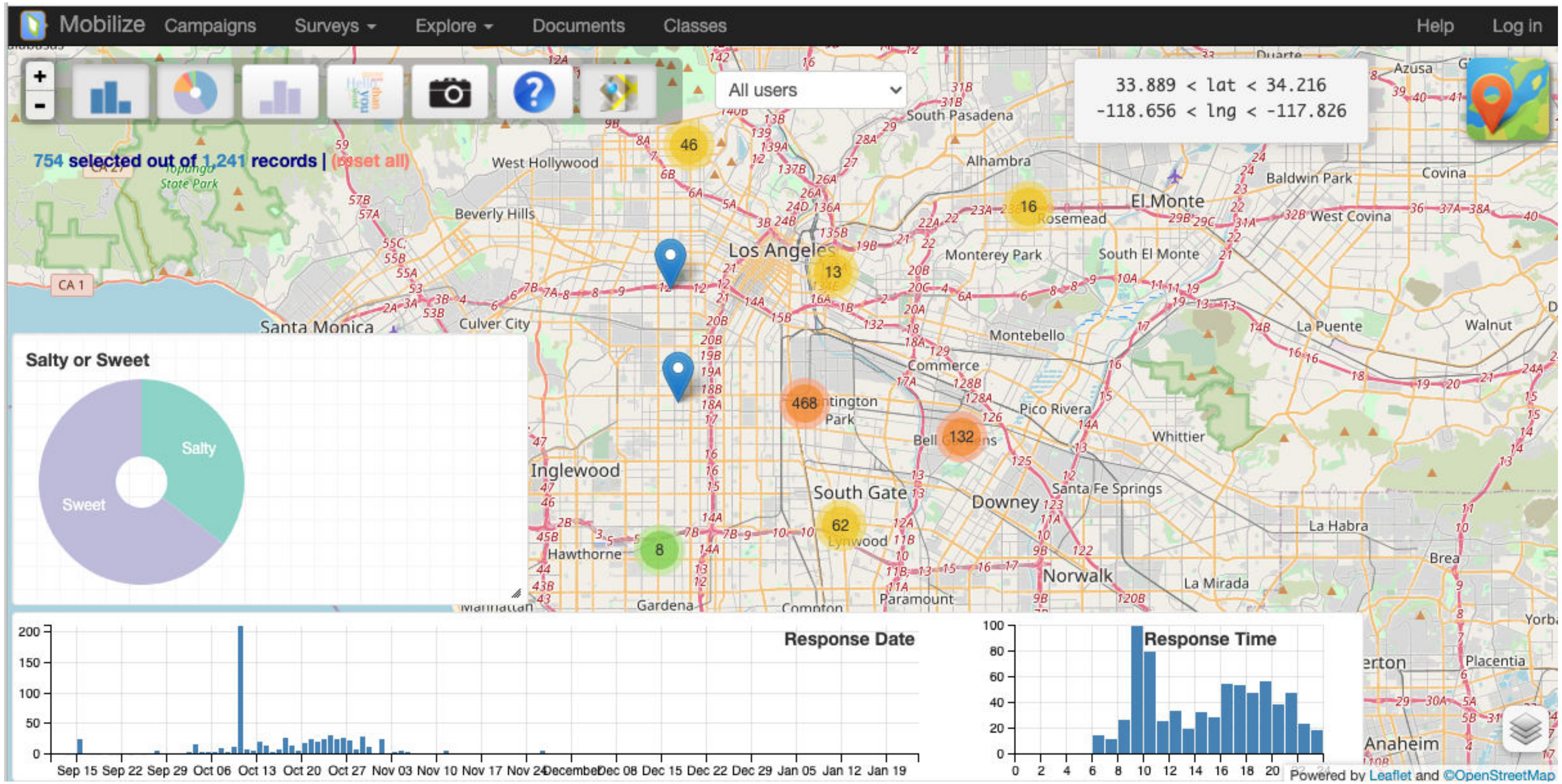


About how many servings did you actually eat?

1

(MIN) 0 20 (MAX)

https://sandbox.mobilizingcs.org/#dashboard/#urn:public:nutrition



754 selected out of 1,241 records | (reset all)

All users

33.889 < lat < 34.216
-118.656 < lng < -117.826

3. Markers

Practicum
The Data Cycle & My Food Habits

Instructions:

With a partner, you will engage in the Data Cycle to address the Research Topic:

How good are we at identifying healthy and unhealthy snacks?

Task:

1. Create a Data Cycle poster.
2. The poster should illustrate how the Data Cycle is used to address the Research Topic.
3. Use RStudio to create at least one statistical graphic. The graphic MUST be included on the poster.
4. You and your partner will present your findings with appropriate evidence from the data.

Awards:

Your teacher will select the top posters in the following categories:

- Best Statistical Question
- Most Interesting Statistical Graphic
- Best Illustration of the Data Cycle

Calories

[redacted]

1. I wonder if we eat more calories than we should?

2. I wonder how many calories are in the snacks we eat?

3. I wonder when we consume the most calories?

Prompts

1. What did you eat?

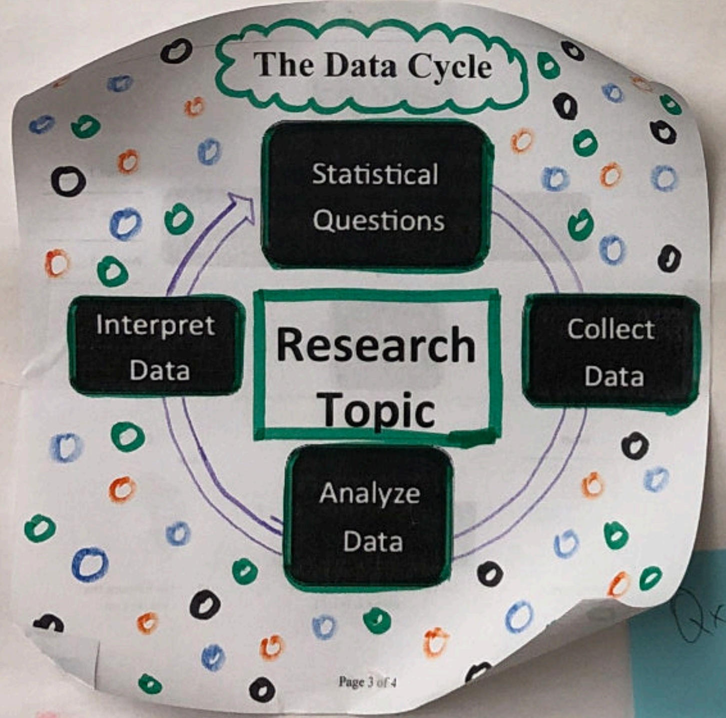
2. how many calories do you think were in the snack?

3. Why did you eat that snack?

4. Do you think you ate the right amount of calories?

Qx 3

Qx 2



The Data Cycle

Statistical Questions

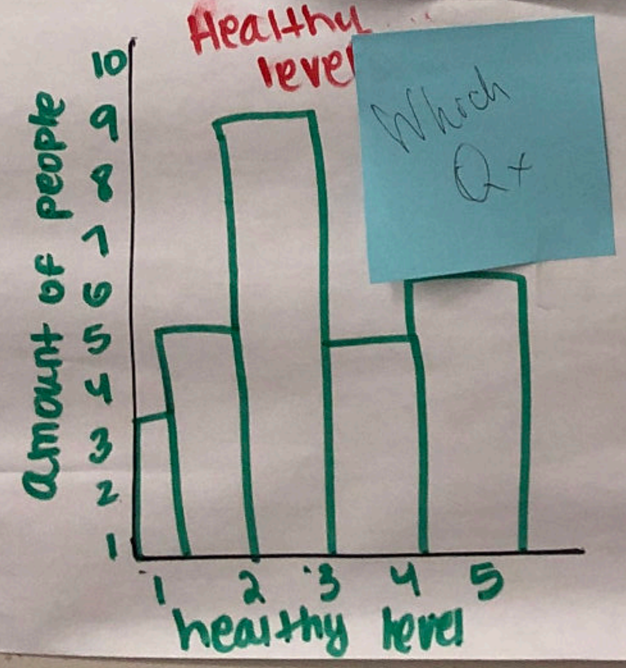
Collect Data

Analyze Data

Interpret Data

Research Topic

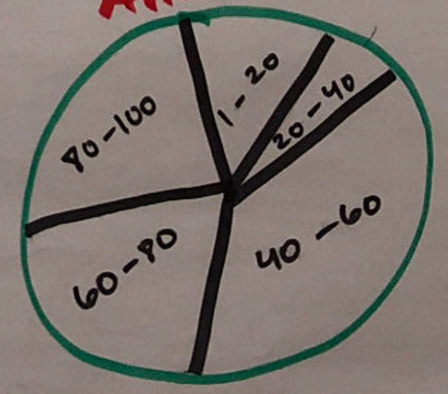
Page 3 of 4



why

hungry because tired sad angry Sport bored wanted someone gave it to me to I dont know

Amount of calories



Why teach coding?

- Learning to "code" using R has many advantages:
  - Students use code to communicate models and ideas
  - Students more easily understand code than mathematical notation
  - Teaches reproducible research habits and communication
  - Some coding is needed for students to access data.
- Heinzman (2020):
  - Students find that coding is "helpful" and "productive" for solving problems. (Heinzman, 2020)
  - Students find it "efficient" and "empowering" (Heinzman, 2020)

IDS Today

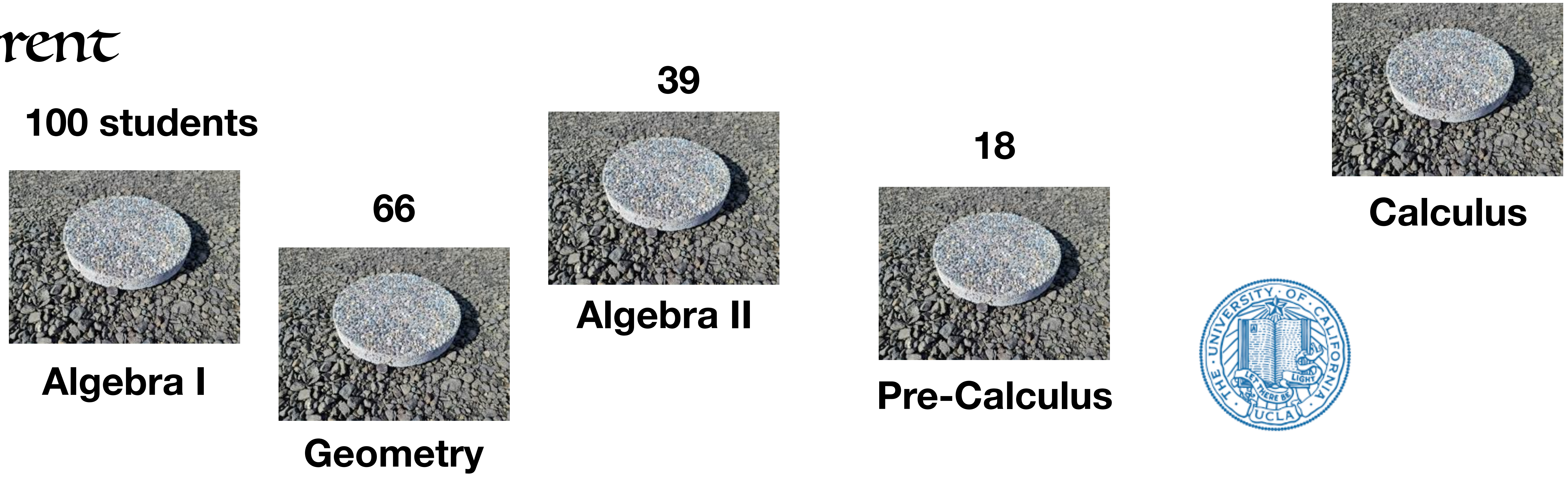
- 48 districts
- 135 High Schools
- To date: 28,000 students
- We make contract with school districts to provide two years of professional development (9 days for new teachers, 3-5 days for more advanced teachers.)
- We provide software technology

Other purposes

- Data Science to improve citizenship: Universal data literacy
- Data science to improve mathematics
- Data science to improve access to university study
- Data science to improve computer science

Math "Pathways" to University

Current



100 students

Algebra I

66

Geometry

39

Algebra II

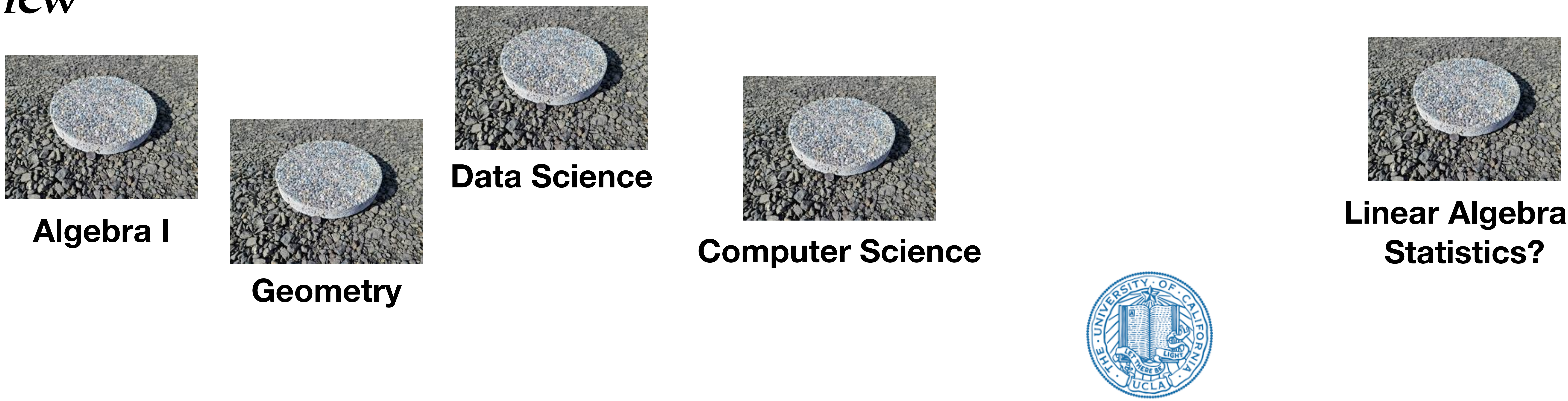
18

Pre-Calculus

Calculus

Gao, N. and Johnson H., "Improving College Pathways in California", Public Policy Institute of California, 2017

New



Algebra I

Geometry

Data Science

Computer Science

Linear Algebra?
Statistics?

Challenges to Data Science

- Students forced to choose between Data Science and the Calculus Pathway.
- Statistics, Computer Science, Engineering, Mathematics, Physics, Chemistry still require university applicants to follow the calculus pathway
- Some mathematicians in California are pushing back vigorously to prevent resources going "away" from calculus and towards data science.
- Perception: Data Science is "less rigorous"

Universal Data Literacy

- Science-oriented students, middle-class and up, ambitious students can skip data literacy
- And so universal data literacy is undermined

Proposed solution

- Emphasize that data science is worthy and necessary in itself
- Develop data science pathways in schools
- Emphasize K-12 data science curriculum
- Build bridges from school data science to university study, including sciences