DATA 8

Summer 2018

Lecture 35

Conclusion

Slides created by Fahad (fhdkmrn@berkeley.edu) and Vinitra (vinitra@berkeley.edu)

Announcements

Final Exam

- **Thursday August 9, 5:00 p.m. to 8:00 p.m.**
- **Le Conte 1, Le Conte 4, and other rooms**
  - Seating assignments to be sent via email
- Bring something to write with and something to erase with; but not food/drink that smells. Water is OK.
- We will provide a couple of reference sheets, with drafts posted on Piazza after lecture
- No calculators or other aids
- Covers the whole course

Next Week

- Monday, Tuesday Wednesday Lectures:
  - TAs will hold review sessions
- No lecture Thursday or Friday
- Monday labs
  - Topical review sessions -- show up to as many as you want
  - Schedule on Piazza after lecture
- Wednesday labs cancelled
- Office hours:
  - All Monday, Tuesday, Wednesday office hours run as normal
  - Thursday, Friday office hours cancelled
- Mock Final: Tuesday night. More information on Piazza!

Final Exam Preparation

- Final exam covers everything
  - List of excluded topics out on Piazza after lecture
- HW 1-11 Solutions released, Labs 1-9 solutions released, Projects 1 and 2 solutions released
- Past exams on the website
  - Fall 2016 is probably the most representative in difficulty
  - Take this one last and time yourself
  - Piazza threads will be available for you to ask questions
  - Answer each others questions!

Overview of the Course

Big Picture of Data 8

1. Python
2. Describing data
3. General concepts of inference and probability
4. Methods of inference
5. Prediction

1. Python

- General features and Table methods: 3.1 - 9.3, 17.3
- `sample_proportions`: 11.1
- `percentile`: 13.1
- `np.average, np.mean, np.std`: 14.1, 14.2
- `minimize`: 15.4

2. Describing Data

- Tables: Chapter 6
- Classifying and cross-classifying: 8.2, 8.3
- Visualizing Distributions: Chapter 7
- Center and spread: 14.1-14.3
- Linear trend and non-linear patterns: 8.1, Chapter 15

3. General Concepts of Inference

- Study, experiment, treatment, control, confounding, randomization, causation, association: Chapter 2
- Distribution, Probability: 7.1, 7.2, 9
- Sampling, probability sample: 10.0
- Probability distribution, empirical distribution, law of averages: Chapter 10
- Population, sample, parameter, statistic: 10.1, 10.3
- Model, null and alternative hypothesis: 16.1

Equally Likely Outcomes

- **If all outcomes are assumed equally likely**, then probabilities are proportions of outcomes:

$$P(A) = \frac{\text{number of outcomes that make A happen}}{\text{total number of outcomes}}$$

$$= \text{proportion of outcomes that make A happen}$$

- 9.5

Probability: Exact Calculations

- Probabilities are between 0 (impossible) and 1 (certain)
- P(event happens) = 1 - P(the event doesn’t happen)
- Chance that two events *A* and *B* both happen

= P(*A* happens) x P(*B* happens given that *A* has happened)

- If event *A* can happen in *exactly one* of two ways, then

  P(*A*) = P(first way) + P(second way)

- 9.5

4. Methods of Inference

- Making conclusions about unknown features of the population or model, based on assumptions of randomness in a sample

Simulation

- Using a computer to mimic a physical experiment
- Uses a for loop
- Examples:
  - Sampling many random samples under a null hypothesis
  - Bootstrapping (sampling with replacement) many times from a random sample
- Oftentimes, aim to create an empirical distribution which approximates the probability distribution

Statistics and Parameters

- If we had population information, we would know all sorts of information from it
    - Models that govern the population
    - If two populations are the same
    - Population parameters
        - Average
        - Median
- All we have is one sample from the population
- Statistic: One number calculated from a sample

Typical Hypothesis Testing

- We try to decide between two models that govern a population
  - One null (chance model), one alternative
- We have one sample of data from a population
  - Is it possible our sample come from the null hypothesis?
- P-Value
  - What’s the chance of seeing our observed data, if the null was true, or further in the direction of the alternative viewpoint?

A/B Testing

- We have samples from two groups of data
    - Did the two samples come from the same distribution?
    - Is the difference we see just due to random chance?
- Follow normal hypothesis testing
- How do we simulate under the null?
    - If the null was true, no association between group and values
    - Shuffle values randomly, assign them back to original group
- We can conclude if our data shows an association between groups and values

Estimation

- Try to determine a population parameter
- We have one sample
    - Our sample statistic is a decent estimate
- We have a sample of data
    - What if our sample had been different?
- Bootstrap our data and create confidence intervals
    - Quantify our uncertainty about our estimate for the population parameter

Causality

- Tests of hypotheses can help decide that a difference is not due to chance

- But they don't say ***why*** there is a difference …

- Unless the data are from an RCT 12.3
    - In that case a difference that's not due to chance can be ascribed to the treatment

5. Prediction

- Descriptive statistics:
  - One variable (average, SD, etc)
  - Two variables (correlation and regression)
- Classification

Regression Pt. 1

- Use average and standard deviation to describe a distribution
- Use the above to convert data to standard units
- Use this to calculate linear association (correlation) between two variables
- Slope of regression line in standard units turns out to be correlation

Regression Pt. 2

- Create a regression line in original units by finding slope, intercept
- Turns out regression line is the unique line which minimizes root mean squared error
- Analyze residuals of regression predictions to determine if linear regression was a good idea

Regression Inference

- Regression model:
  - Data originally came from a “true line”
  - Take a sample of points, push them off the line randomly (with normal distribution, mean 0)
- We have a sample of points
  - What if our sample had been different?
- Bootstrap our scatter plot
  - Can try and predict the slope, heights at various x-values of the “true line”

Classification

- Binary classification based on attributes 17.1
  - *k*-nearest neighbor classifiers
- Training and test sets 17.2
  - Why these are needed
  - How to generate them
- Implementation: 17.4
  - Distance between two points
  - Class of the majority of the *k* nearest neighbors
- Accuracy: Proportion of test set correctly classified 17.5

Machine Learning

- Supervised Machine Learning
    - Input: Labeled data
    - Output: Prediction for unlabeled example
    - High computational complexity
- Unsupervised Machine Learning
    - Input: Unlabeled data
    - Output: Recognize underlying patterns in the data
    - Low computational complexity

What's Next?

Course Recommendations

Data 100

Data Science Lifecycle

Data 100: Principles and Techniques of Data Science

- **Prepare** students for advanced courses in data-management, machine learning, and statistics
- **Enable** students to start careers as data scientists by working with real-world data, tools, and techniques

NumPy, Pandas, SQL, Spark, Seaborn, SciKitLearn, Plotly

Prerequisites: Data 8, Computing, Math (Linear Algebra)

Prob 140

Probability

Here's the model; what can you say about the sample?

Prob 140: Probability for Data Science (prob140.org)

- Pilot in Spring 2017
- Listed as Statistics 140
- Several members of the course staff recently took it
- The mathematics of chance
- Python and Jupyter are used for computing and for understanding the math better

Programming

- CS 61A: Structure and Interpretation of Computer Programs
    - CS 88: Computational Structures in Data Science
- CS 61B: Data Structures and Algorithms
- STAT 133: Concepts in Computing with Data
- CS 186: Introduction to Databases

Inference

- STAT 135: Concepts of Statistics
- STAT 150: Stochastic Processes
- STAT 151A: Linear Modeling
- STAT 153: Introduction to Time Series
- PB HLTH 142: Intro to Probability and Statistics in Biology

Prediction

- CS 188: Introduction to Artificial Intelligence
- CS 189: Introduction to ML
- IEOR 142: Introduction to ML & Data Analytics
- STAT 154: Modern Statistical Prediction & ML

Data Science Major / Minor

All released information can be found on data.berkeley.edu

Data Science

Why Data Science

- Unprecedented access to data means that we can make new discoveries and more informed decisions
- Computation is a powerful ally in data processing, visualization, prediction, and statistical inference
- People can agree on evidence and measurement

How to Analyze Data

Begin with a question from some domain, make reasonable assumptions about the data and a choice of methods.

Visualize, then quantify!

Perhaps the most important part: Interpretation of the results in the language of the domain, without statistical jargon.

How *Not* to Analyze Data

Begin with a question from some domain, make reasonable assumptions about the data and a choice of methods.

Visualize, then quantify!

Perhaps the most important part: Interpretation of the results in the language of the domain, without statistical jargon.

How to Analyze Data in 2018

Begin with a question from some domain, make reasonable assumptions about the data and a choice of methods.

Visualize, then quantify! Do both using computation.

Perhaps the most important part: Interpretation of the results in the language of the domain, without statistical jargon.

The Design of Data 8

- Table manipulation using Python
- Working with whole distributions, not just means
- Decisions based on sampling: assessing models
- Estimation based on resampling
- Understanding sampling variability
- Prediction

Data Science in the Future

Our Journeys

A Request

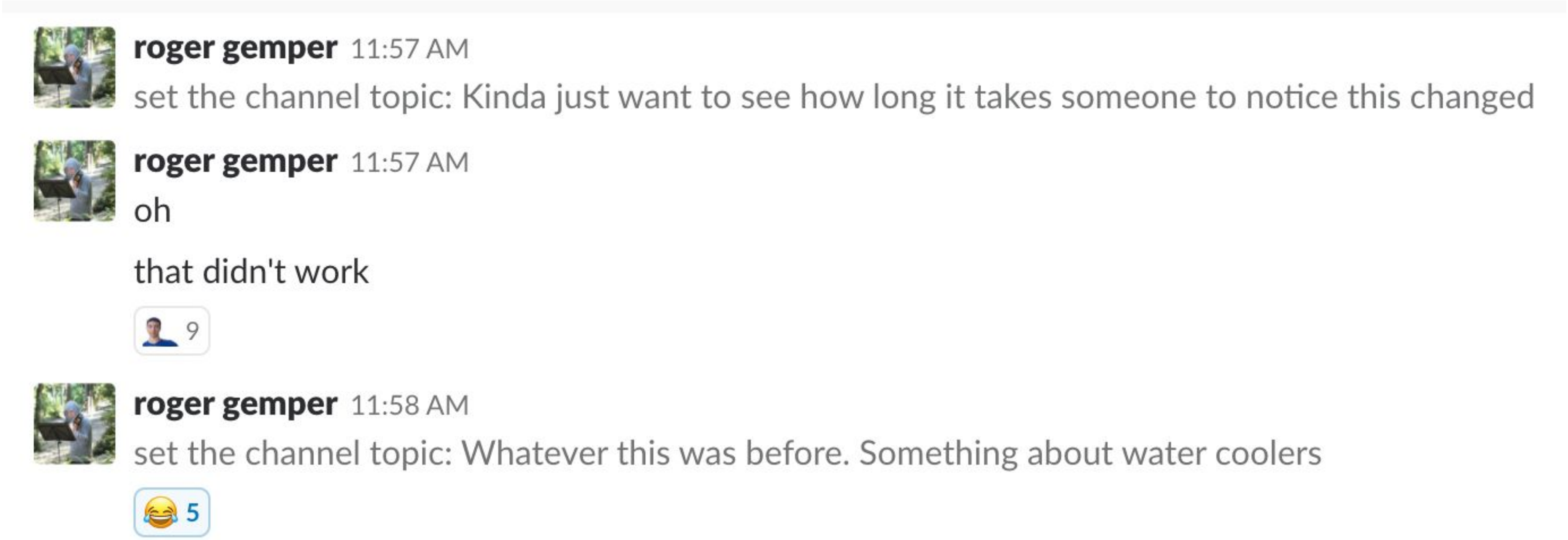
Please fill out the course evaluations.

The Team

Staff

- GSIs
- Tutors
- Lab Assistants

Joining the Team



roger gemper 11:57 AM

set the channel topic: Kinda just want to see how long it takes someone to notice this changed

roger gemper 11:57 AM

oh

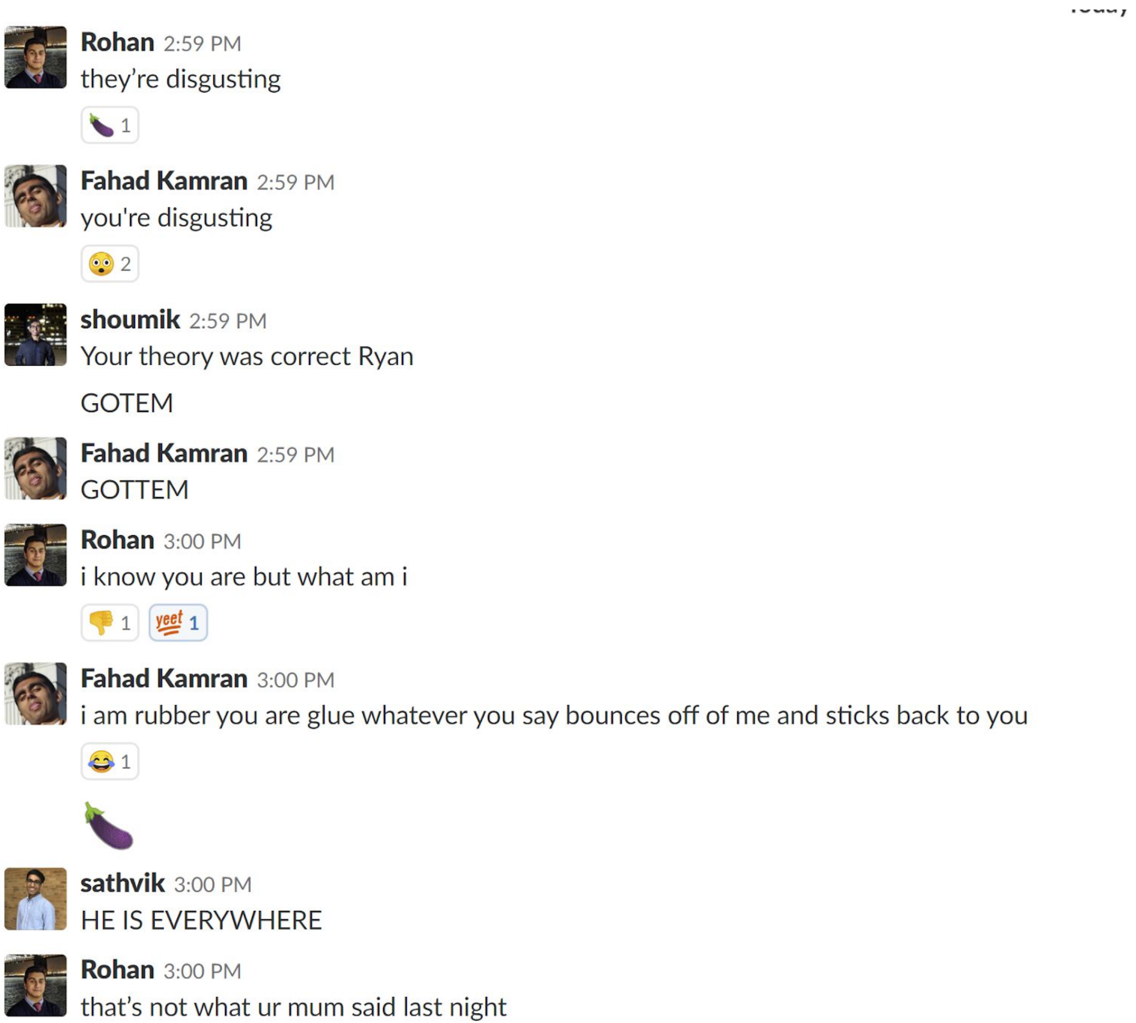
that didn't work

👤 9

roger gemper 11:58 AM

set the channel topic: Whatever this was before. Something about water coolers

😂 5



Today

Rohan 2:59 PM
they’re disgusting

🍆 1

Fahad Kamran 2:59 PM
you're disgusting

😳 2

shoumik 2:59 PM
Your theory was correct Ryan

GOTEM

Fahad Kamran 2:59 PM
GOTTEM

Rohan 3:00 PM
i know you are but what am i

👎 1 yeet 1

Fahad Kamran 3:00 PM
i am rubber you are glue whatever you say bounces off of me and sticks back to you

😂 1

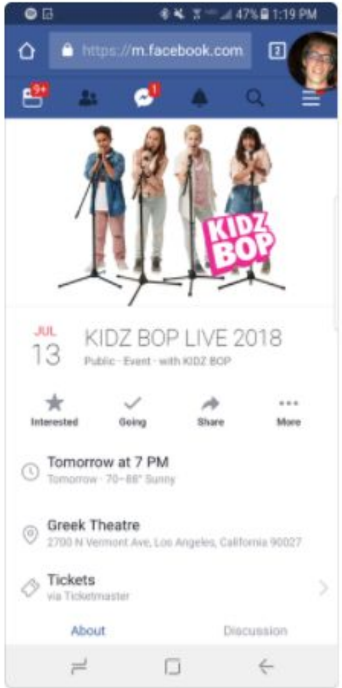
🍆

sathvik 3:00 PM
HE IS EVERYWHERE

Rohan 3:00 PM
that’s not what ur mum said last night

Fahad Kamran 1:20 PM

Screenshot_20180712-131957.png ▾



JUL
13

KIDZ BOP LIVE 2018

Public · Event · with KIDZ BOP

Tomorrow at 7 PM

Tomorrow · 70–88° Sunny

Greek Theatre

2700 N Vermont Ave, Los Angeles, California 90027

Tickets

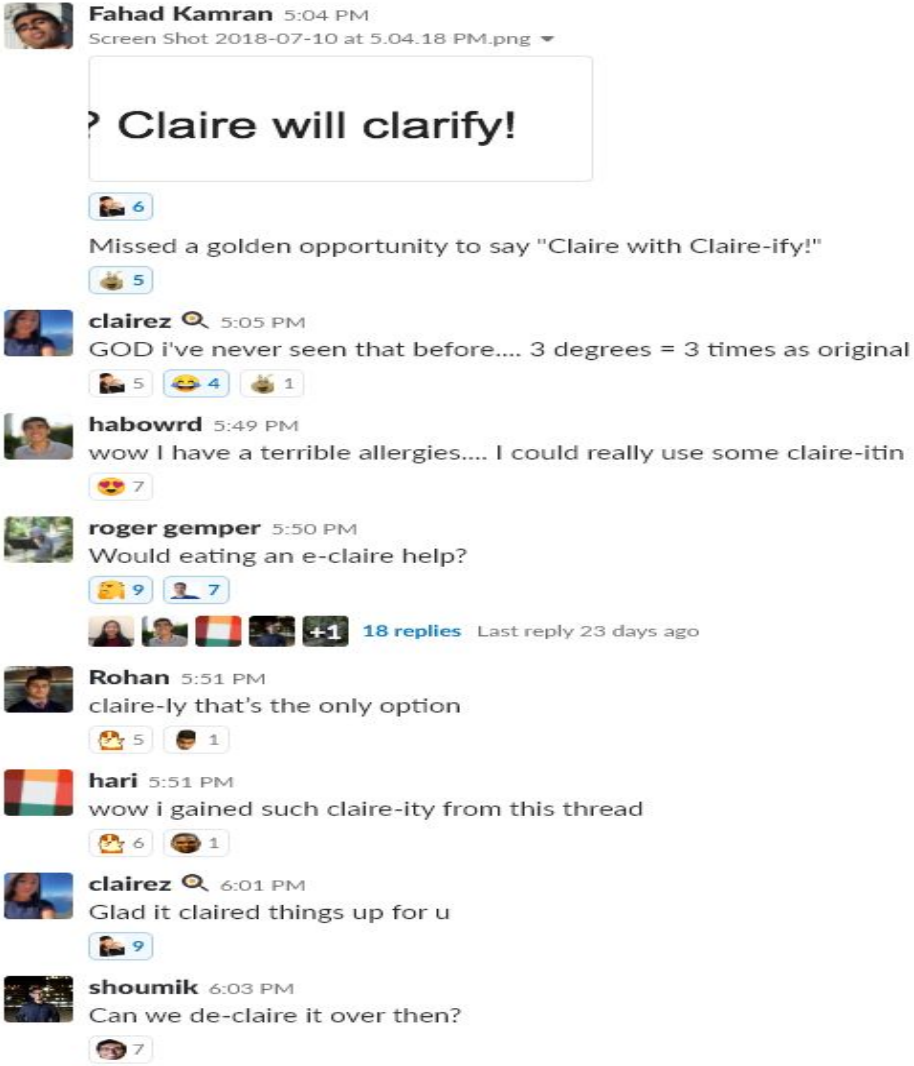
via Ticketmaster

We were looking for something to do tomorrow night right???

❤️ 3

Rohan 1:21 PM

sick let's drive down to LA



Fahad Kamran 5:04 PM
Screen Shot 2018-07-10 at 5.04.18 PM.png

? Claire will clarify!

6

Missed a golden opportunity to say "Claire with Claire-ify!"

5

clairez 5:05 PM
GOD i've never seen that before.... 3 degrees = 3 times as original

5 😂 4 1

habowrd 5:49 PM
wow I have a terrible allergies.... I could really use some claire-itin

😍 7

roger gemper 5:50 PM
Would eating an e-claire help?

9 7

+1 18 replies Last reply 23 days ago

Rohan 5:51 PM
claire-ly that's the only option

5 1

hari 5:51 PM
wow i gained such claire-ity from this thread

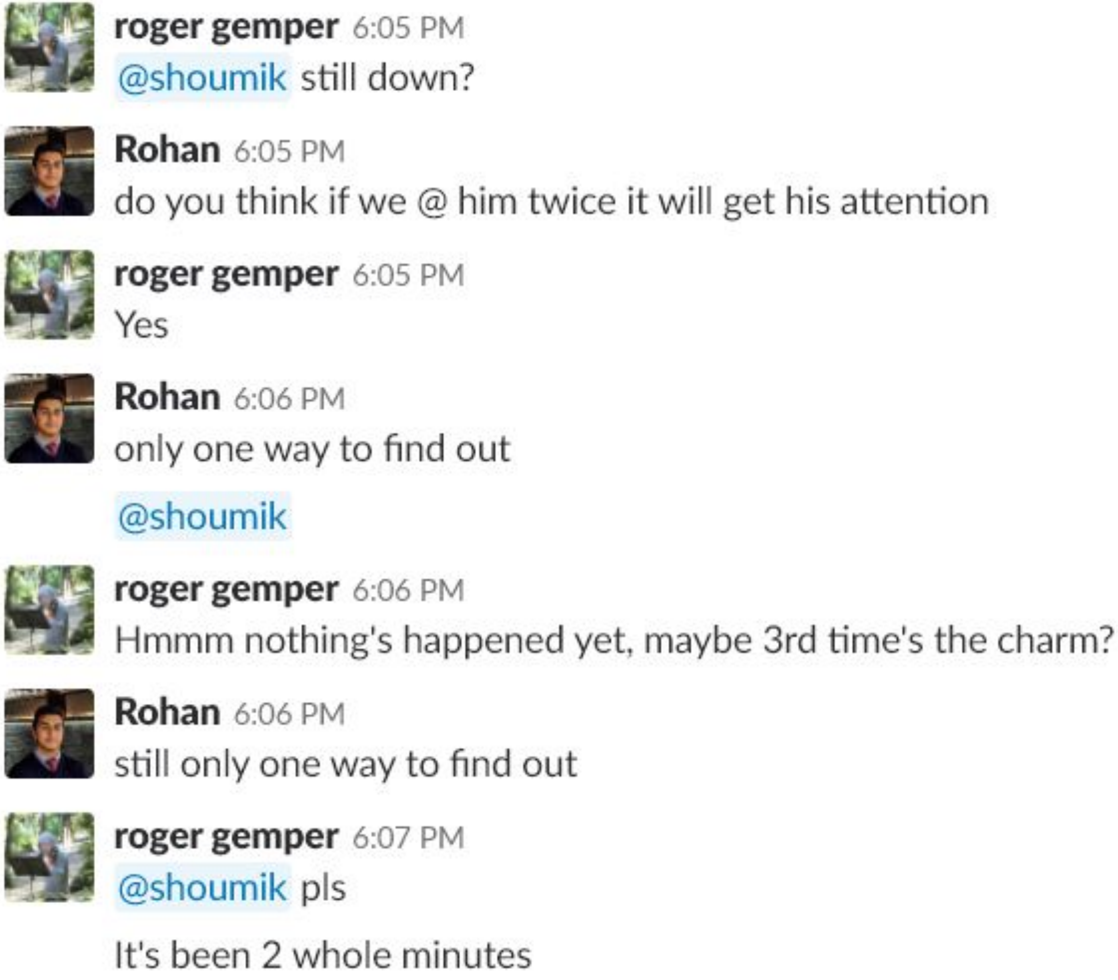
6 1

clairez 6:01 PM
Glad it claired things up for u

9

shoumik 6:03 PM
Can we de-claire it over then?

7



roger gemper 6:05 PM
@shoumik still down?

Rohan 6:05 PM
do you think if we @ him twice it will get his attention

roger gemper 6:05 PM
Yes

Rohan 6:06 PM
only one way to find out

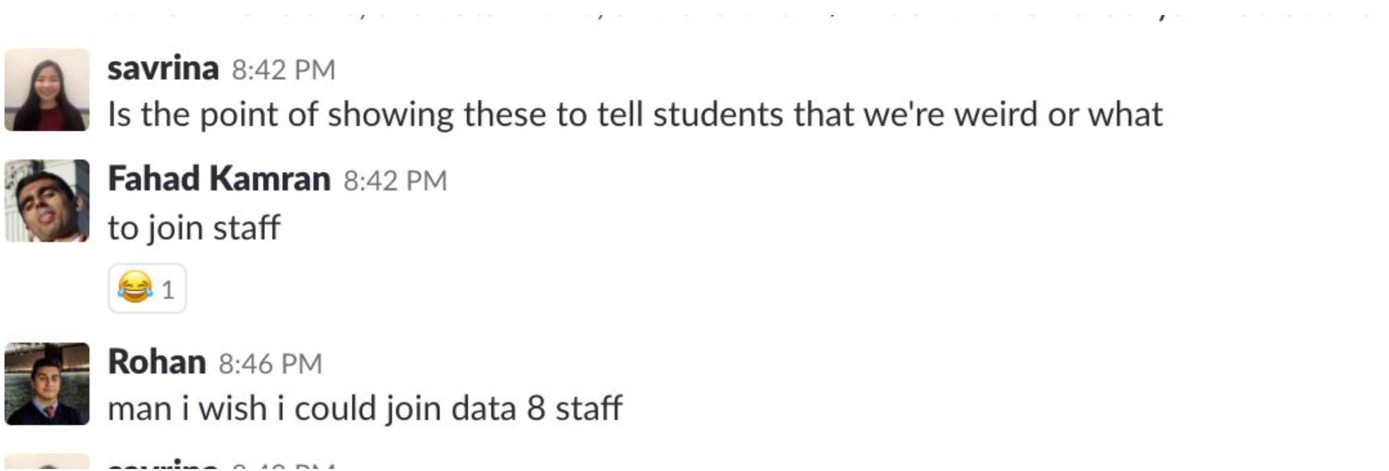
@shoumik

roger gemper 6:06 PM
Hmmm nothing's happened yet, maybe 3rd time's the charm?

Rohan 6:06 PM
still only one way to find out

roger gemper 6:07 PM
@shoumik pls

It's been 2 whole minutes



savrina 8:42 PM

Is the point of showing these to tell students that we're weird or what

Fahad Kamran 8:42 PM

to join staff

😂 1

Rohan 8:46 PM

man i wish i could join data 8 staff

Thank you!

Come get boba with us
(drinks not included)