Carnegie Mellon University
Heinz College

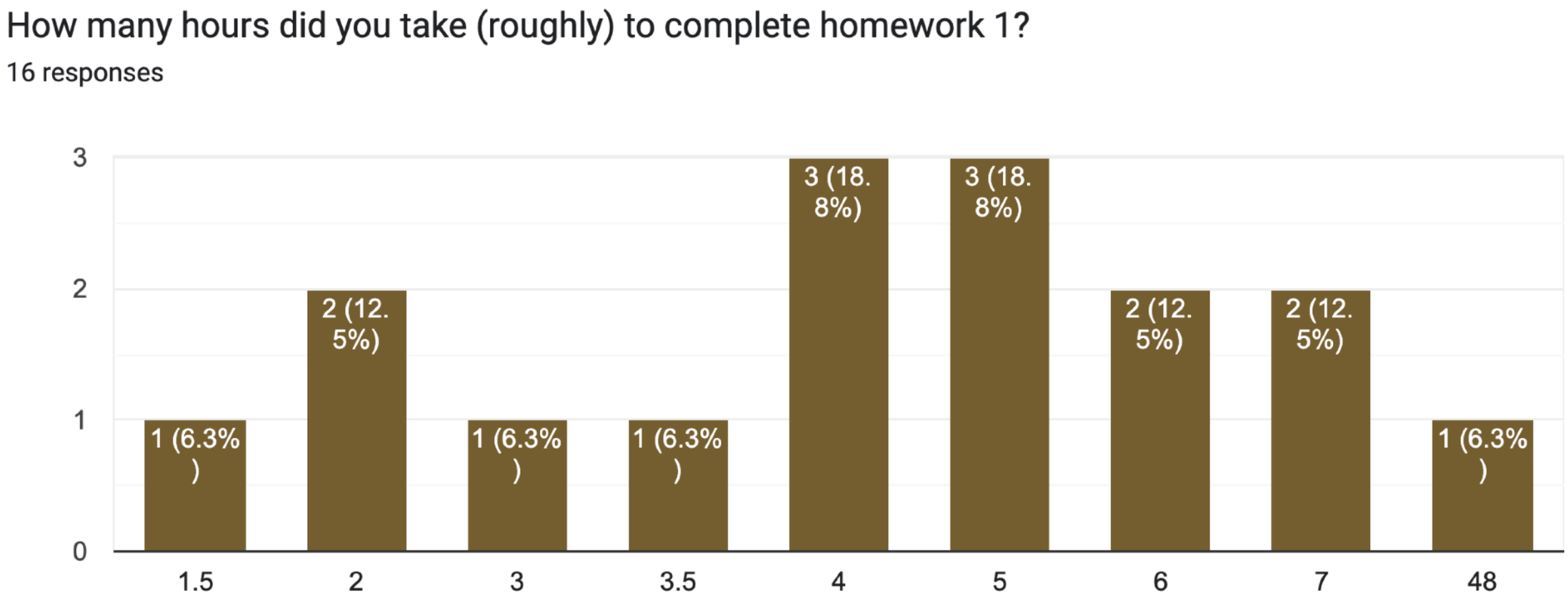
94-775 Unstructured Data Analytics

Lecture 8: Clustering (cont'd)

Slides by George H. Chen

HW1 Questionnaire Results (1/4)

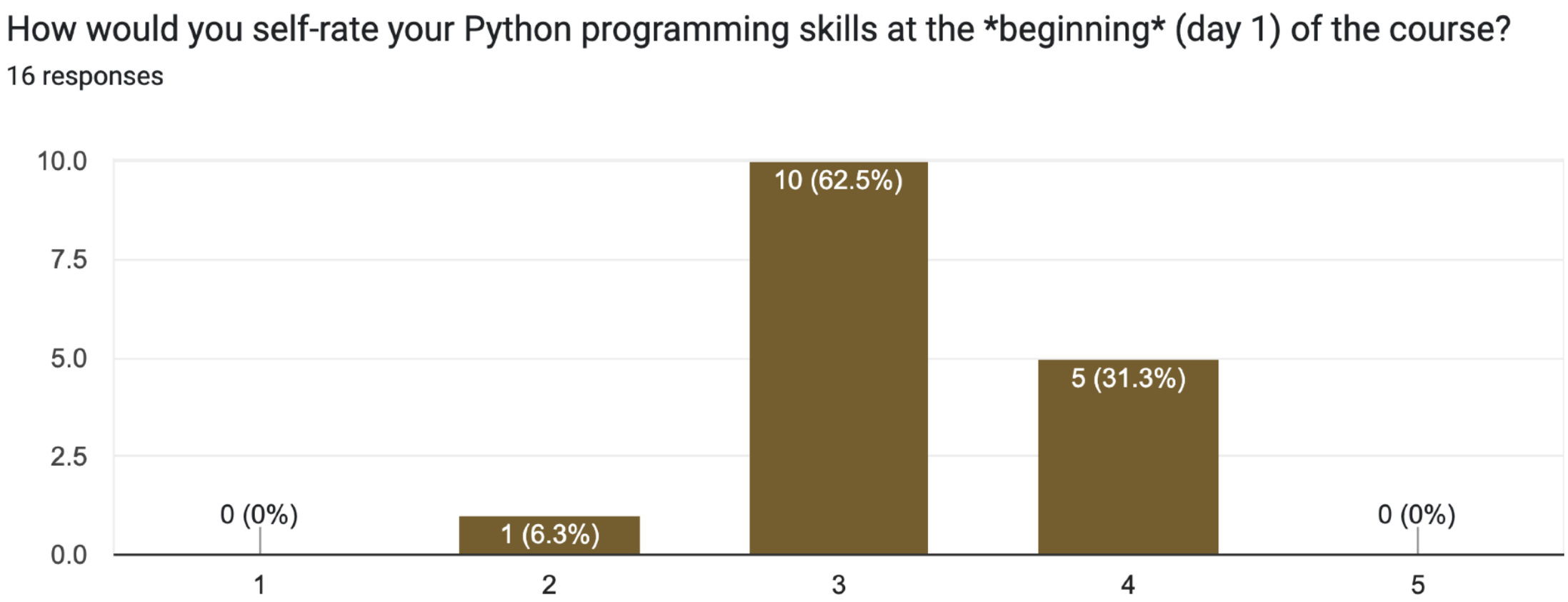
Overall: we target the homework to be doable within 15 hours
(so for the most part, students are doing well!)



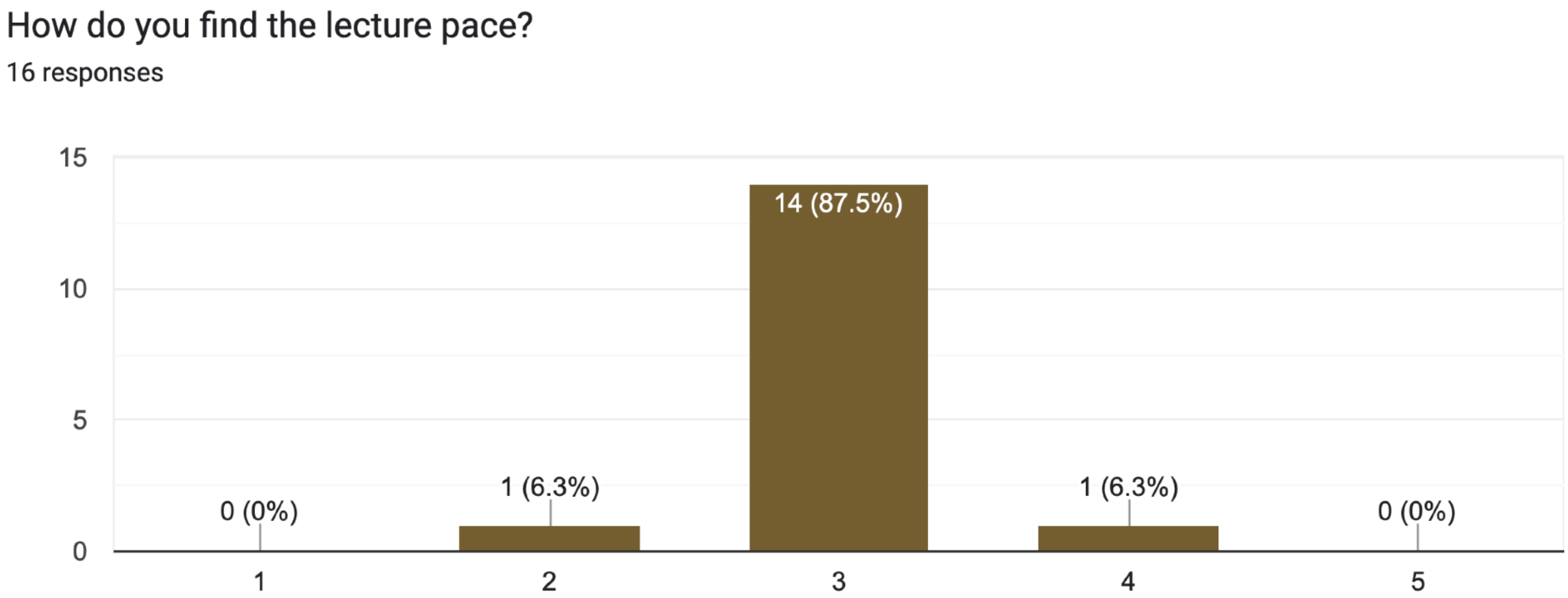
Whoa! So fast!
(It takes time to even just read the questions!)

Is this a typo?
48 is unusually large!

HW1 Questionnaire Results (2/4)



HW1 Questionnaire Results (3/4)



👍

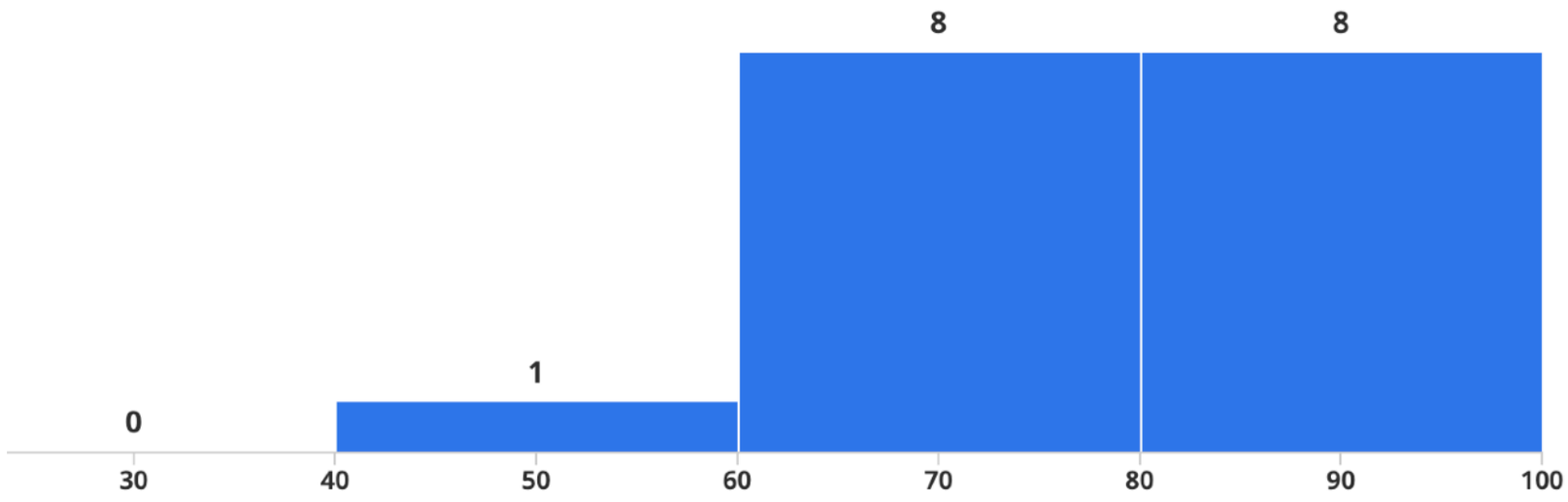
HW1 Questionnaire Results (4/4)

- ~52% of students said that they have not taken machine learning/ deep learning coursework (aside from 90-803)

- Most confusing topic mentioned in questionnaire has been dimensionality reduction (PCA, manifold learning)
    - Hopefully as you see it some more in upcoming lectures, these concepts will sink in a bit more
    - Manifold learning results are often indeed not straightforward to interpret/explain… don't worry about this too much…

- A number of students asked for more interactivity in lecture
    - I'll try to ask more questions to check students' understanding in lecture
    - Earlier on I was a bit more worried about pacing but we'll try to ease pacing once we get to the prediction part of the course

Quiz 1



Median	Maximum	Mean	Std Dev
78.0	**99.0**	**76.76**	**11.36**

These stats are typical of my quizzes (don't panic—no one's at risk of failing)

Remember: letter grades are assigned based on a curve

Solutions are in Canvas -> Files -> "Quiz 1 solutions.pdf"

Regrade requests (use Gradescope's regrade request feature)
are due **Monday April 7, 11:59pm**
(for if you think there's a genuine grading error)

HW1 has also been graded: regrade requests also due Monday April 7, 11:59pm (use Gradescope's regrade request feature)

(Flashback) Final Project Rubric

- **Policy question (15%):** what public policy question are you addressing? Please be clear and concise.
- **Data analysis (30%):** clearly state what part of your data are unstructured (some but not all of the data you are analyzing must be unstructured), and carefully justify every step of your analysis with supporting visualizations/intermediate outputs as needed
- **Code (30%):** your code should actually run!
- **Conclusions (15%):** come up with insights that are based on your quantitative data analysis and that address your original policy question
- **Presentation (10%):** how polished is your final project presentation? — this is based on the live presentation your group makes (changes made to the slides after the presentation don't affect this score)

(Flashback) Final Project Rubric

- **Code (30%):** your code should actually run!

Please have 1 notebook that does preprocessing & saves preprocessed data
(we will **not** attempt to re-run this preprocessing notebook)

Load in your saved preprocessed data into another notebook or notebooks to conduct subsequent analyses
(we will try to run this notebook; please provide saved preprocessed data in a compressed format)

Clustering on Text

Resuming the demo from last time…