gradescope

Gradescope
AI for Grading

NIPS 2016 Education Workshop

Pieter Abbeel
Professor UC Berkeley
Co-founder Gradescope

http://dsp.rice.edu/ml4ed_nips2016

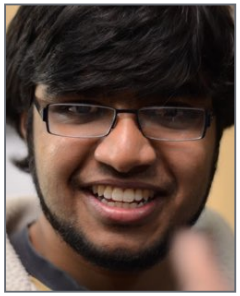
What's in the talk

- How we got started
- How grading works
- ML for analytics: projects we'd love to collaborate on
- AI-assisted grading: what we are working on

gradescope How we got started

Arjun Singh
PhD Robotics

Sergey Karayev
PhD Computer Vision

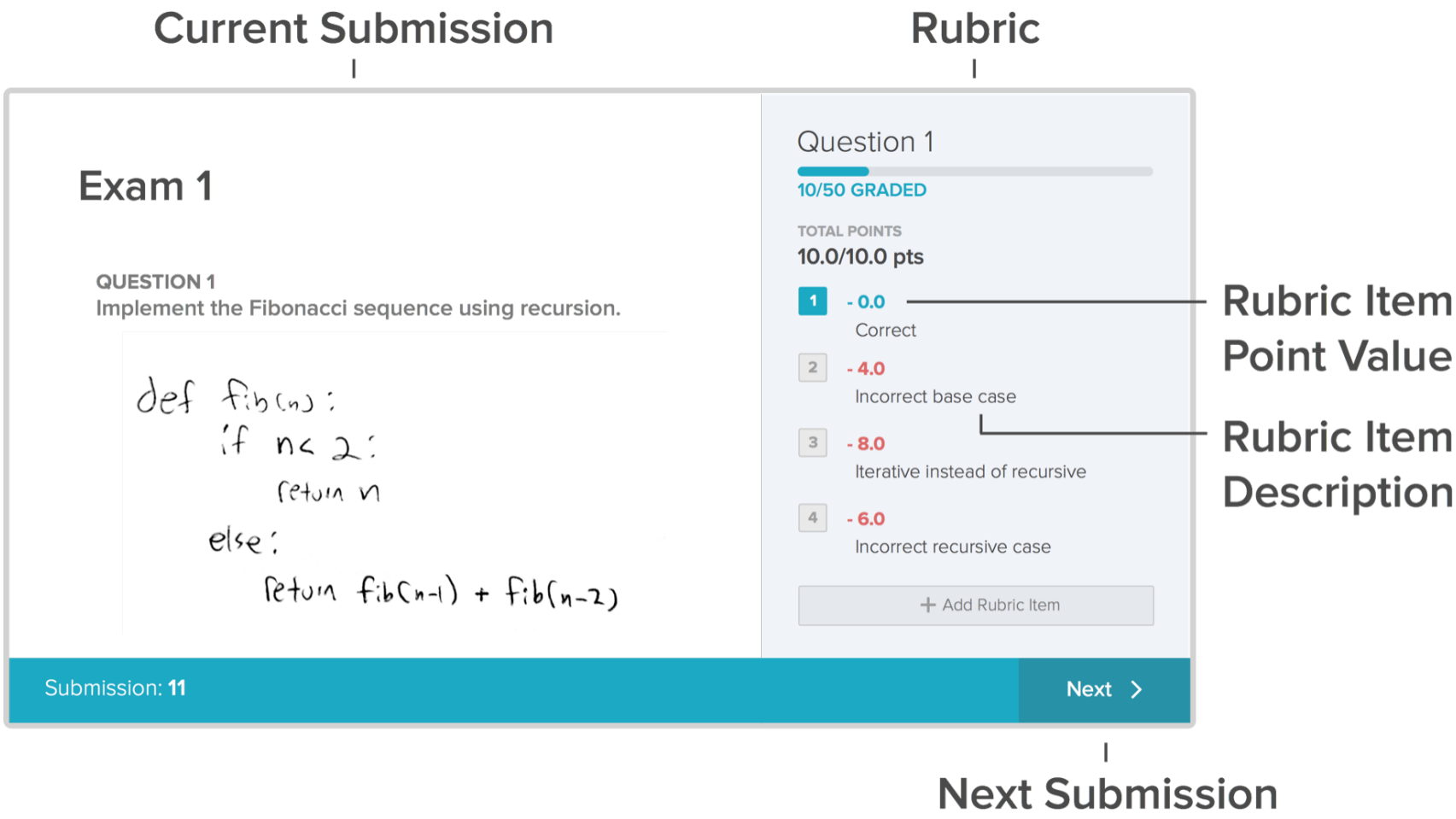


Ibrahim Awwal
MS AI

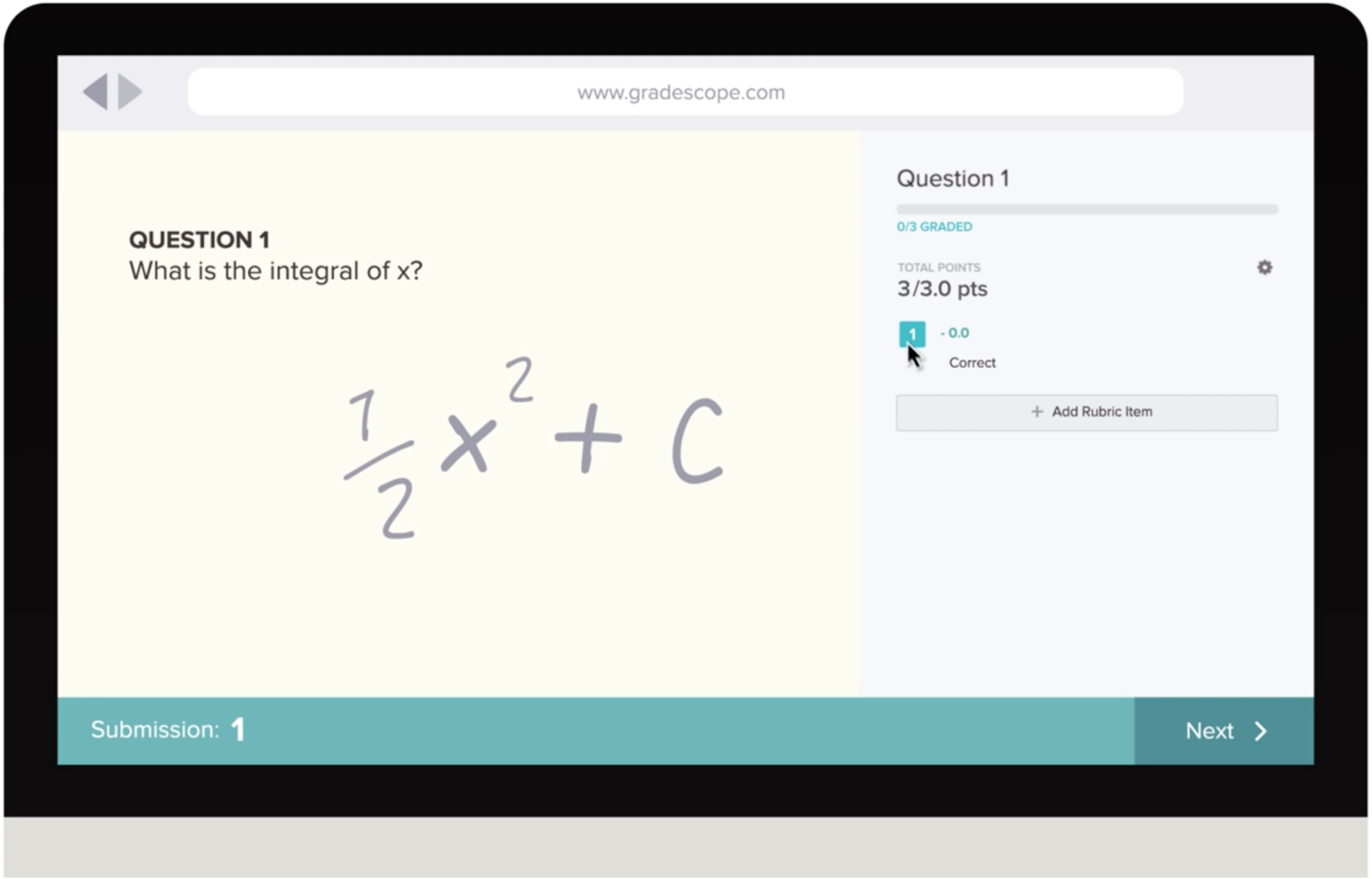
Pieter Abbeel
Professor Robotics

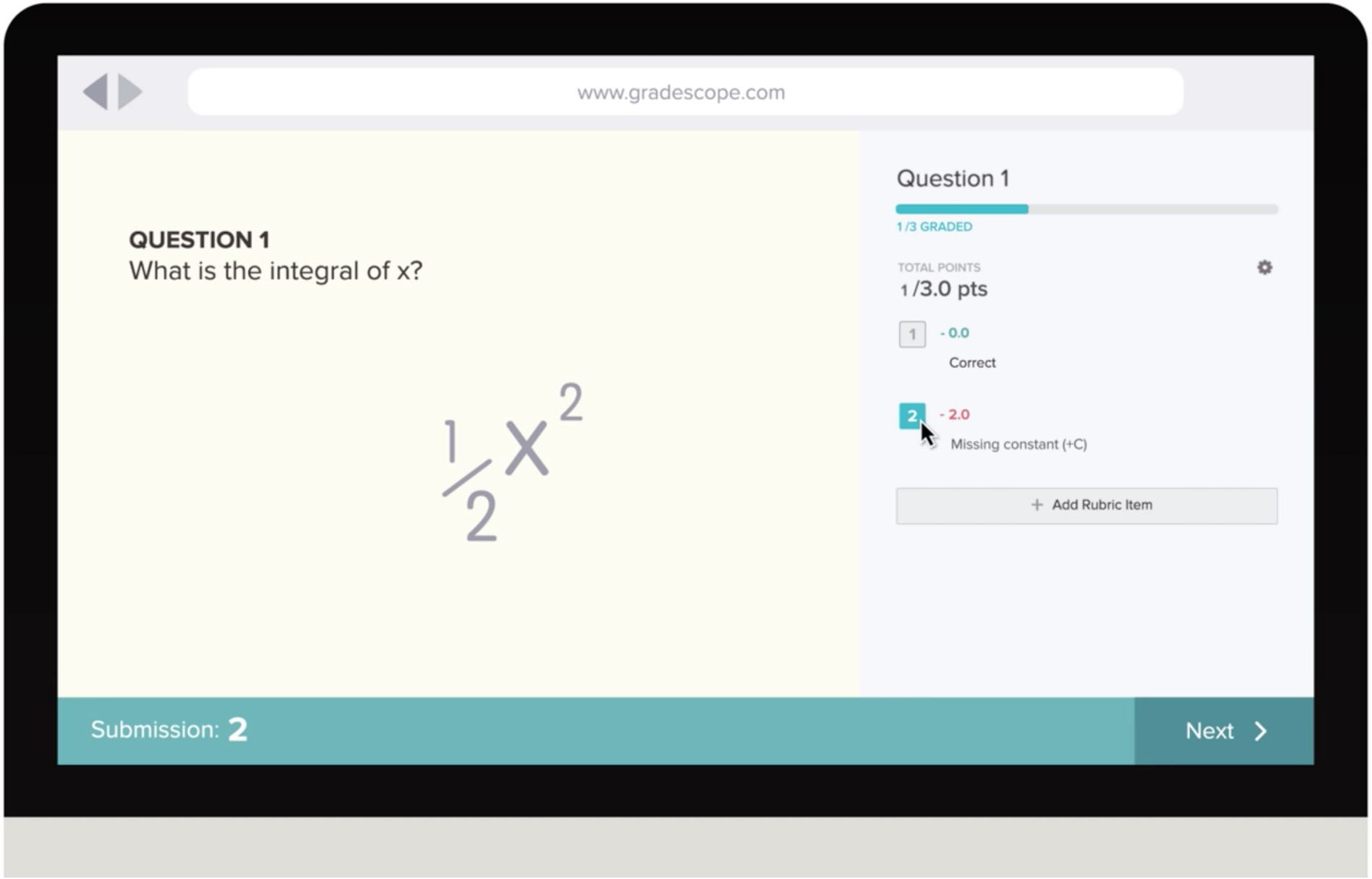
- With TAs, built a tool to grade paper assignments online: faster, more consistent, and prevented cheating.
- Founded company at 50 courses. 2000+ courses today (200+ schools, 15M answers graded).
- Mission is to be the central place for all assessment.

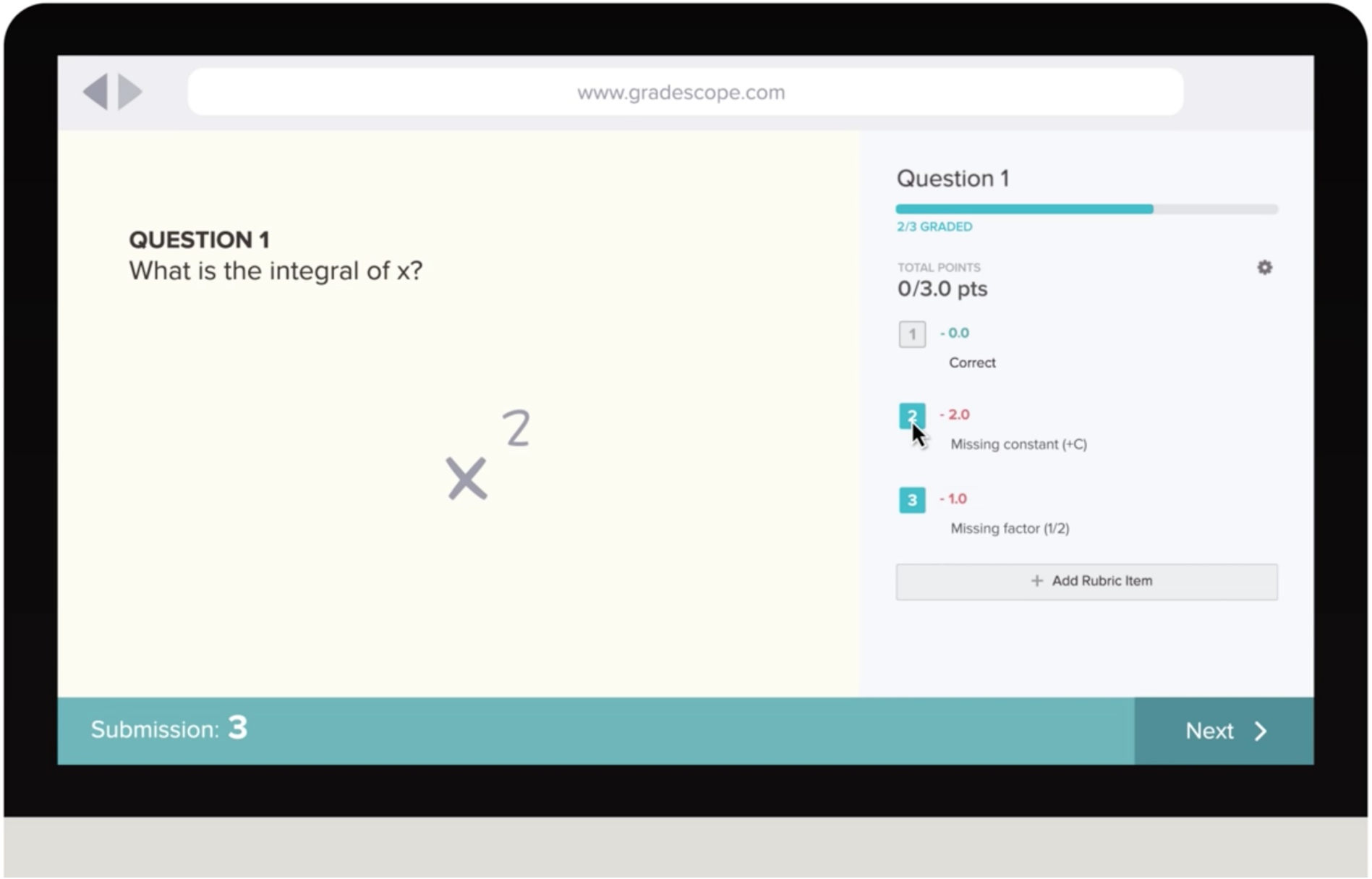
gradescope Grading

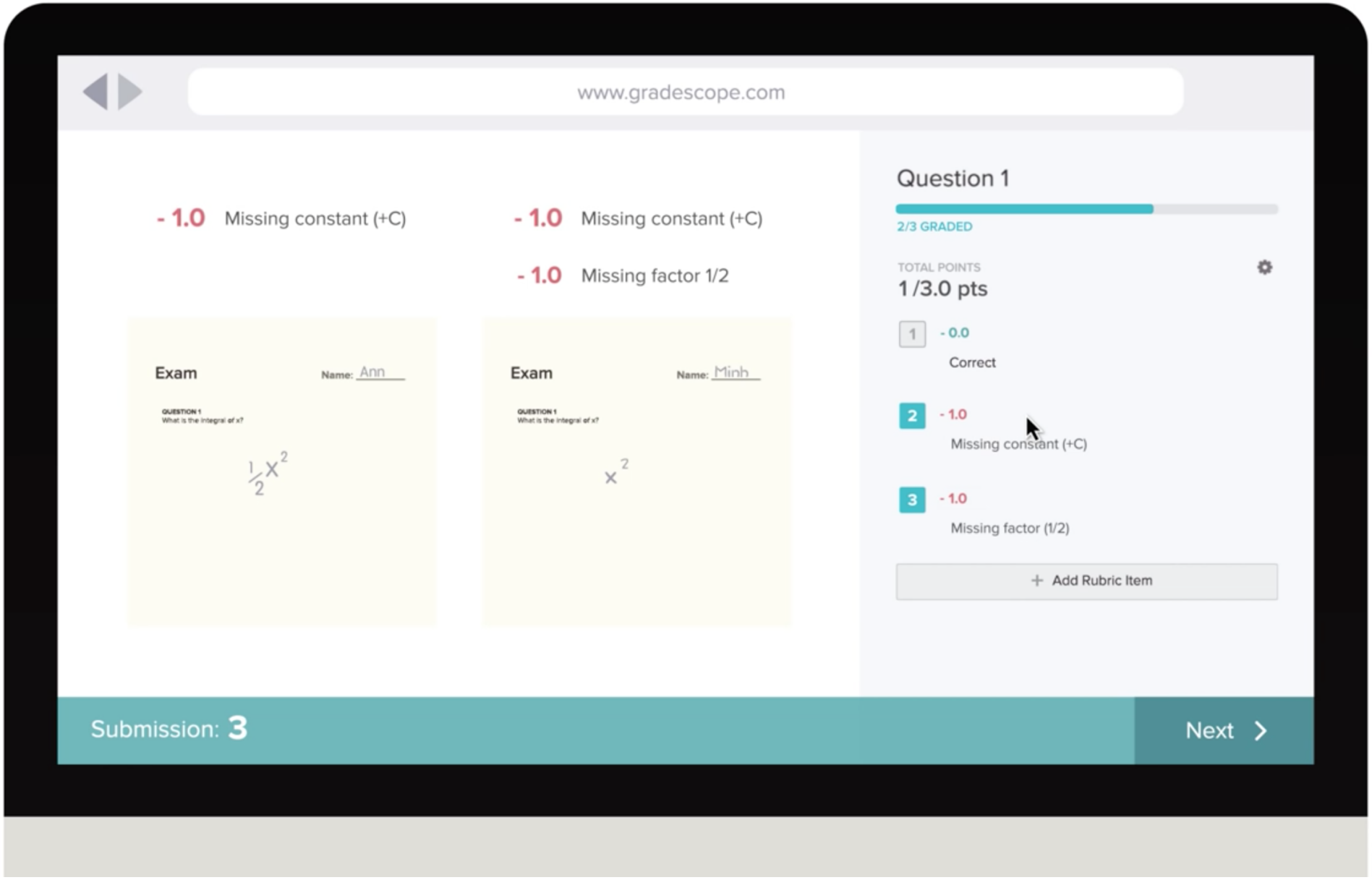


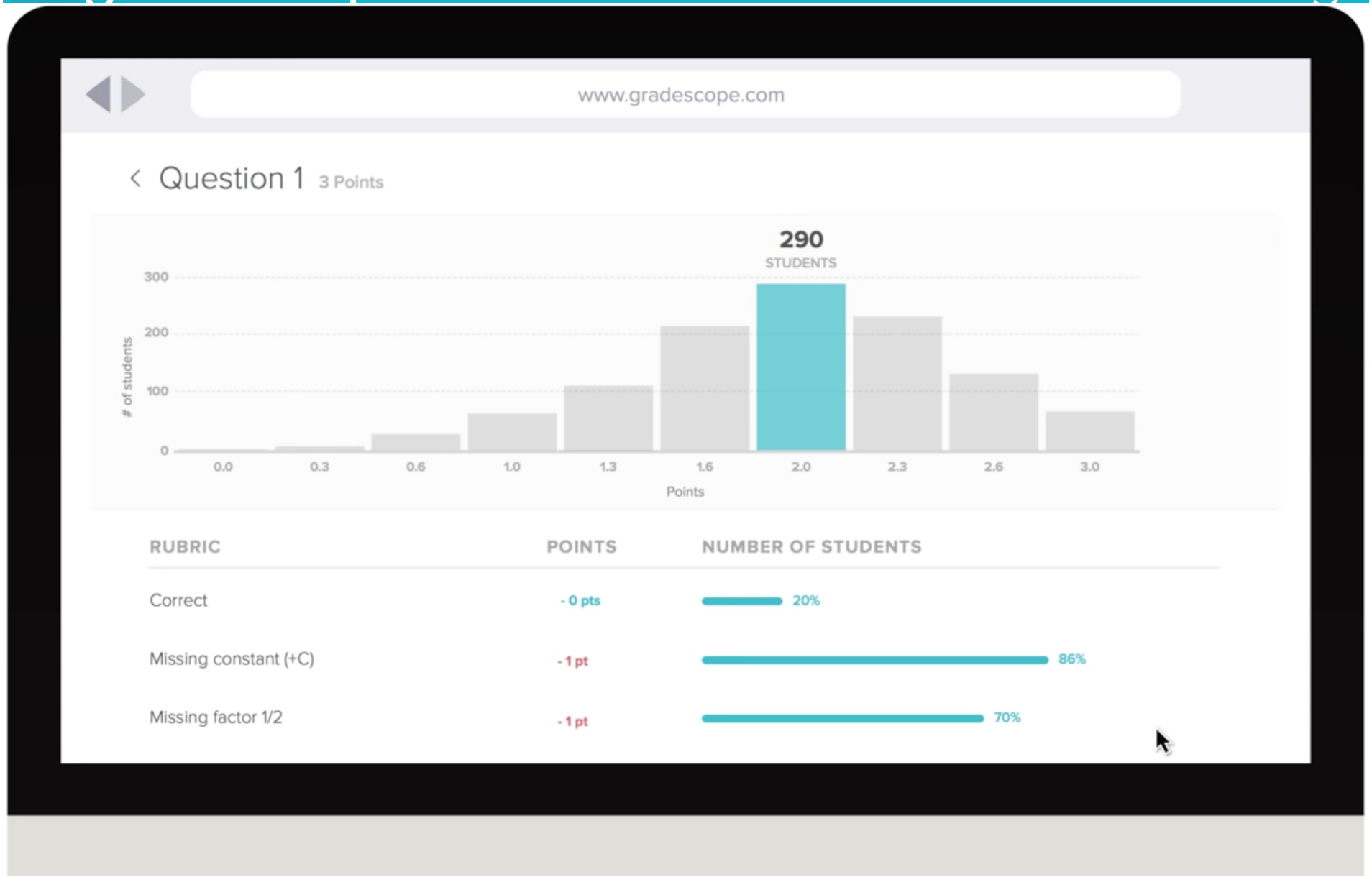
https://www.youtube.com/watch?v=UuIlQiXIjeQ


www.gradescope.com
QUESTION 1
What is the integral of x?
$\frac{1}{2}x^2 + C$
Question 1
0/3 GRADED
TOTAL POINTS
3/3.0 pts
1
- 0.0
Correct
+ Add Rubric Item
Submission: 1
Next >


www.gradescope.com
QUESTION 1
What is the integral of x?
$\frac{1}{2}x^2$
Question 1
1/3 GRADED
TOTAL POINTS
1/3.0 pts
1
- 0.0
Correct
2
- 2.0
Missing constant (+C)
+ Add Rubric Item
Submission: 2
Next >


www.gradescope.com
QUESTION 1
What is the integral of x?
x^2
Question 1
2/3 GRADED
TOTAL POINTS
0/3.0 pts
1 - 0.0
Correct
2 - 2.0
Missing constant (+C)
3 - 1.0
Missing factor (1/2)
+ Add Rubric Item
Submission: 3
Next >

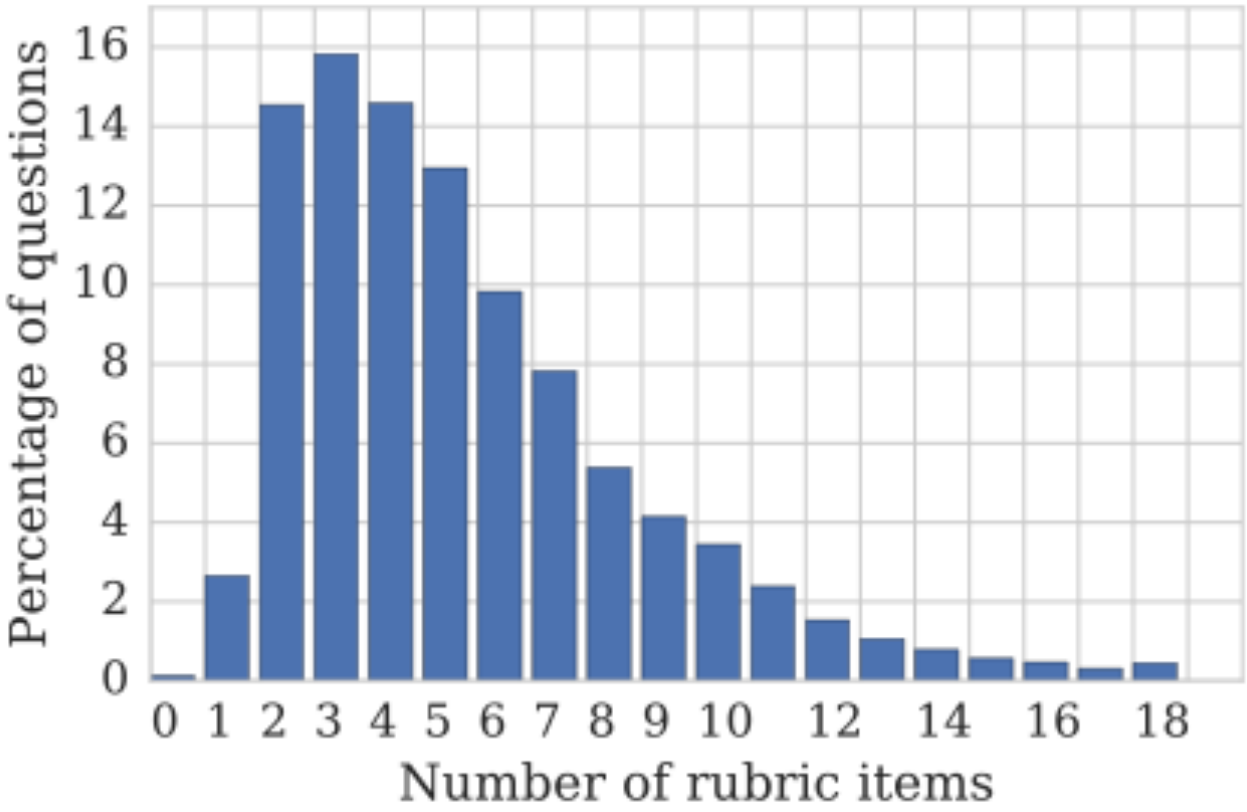

www.gradescope.com
- 1.0 Missing constant (+C)
- 1.0 Missing constant (+C)
- 1.0 Missing factor 1/2
Exam
Name: Ann
QUESTION 1
What is the integral of x?
$\frac{1}{2}x^2$
Exam
Name: Minh
QUESTION 1
What is the integral of x?
x^2
Question 1
2/3 GRADED
TOTAL POINTS
1 /3.0 pts
1 - 0.0
Correct
2 - 1.0
Missing constant (+C)
3 - 1.0
Missing factor (1/2)
+ Add Rubric Item
Submission: 3
Next >


www.gradescope.com
Question 1 3 Points
290
STUDENTS
300
200
100
0
# of students
0.0
0.3
0.6
1.0
1.3
1.6
2.0
2.3
2.6
3.0
Points
RUBRIC
POINTS
NUMBER OF STUDENTS
Correct
- 0 pts
20%
Missing constant (+C)
- 1 pt
86%
Missing factor 1/2
- 1 pt
70%

We would love to collaborate on these!

- Item response theory for rubric-graded questions
- Inferring concepts
- Predicting student outcomes

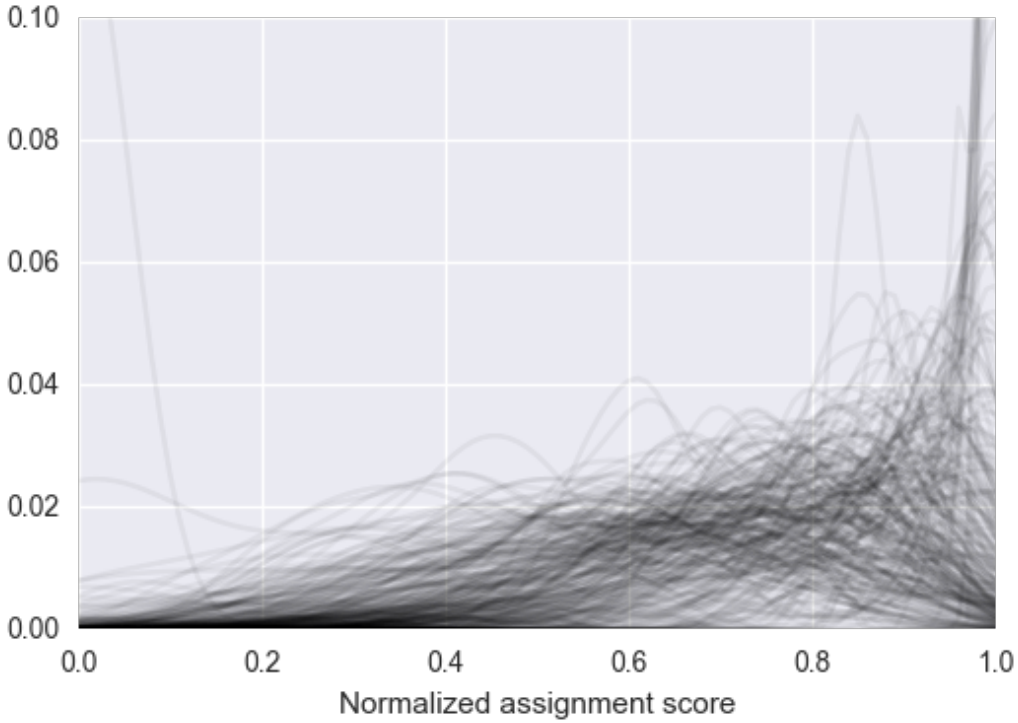
- On Gradescope, mean number of applied rubric items is 5.6. More questions have 8 or more rubric items than 2 or less.
- Approximately equal number of additive and subtractive rubrics.



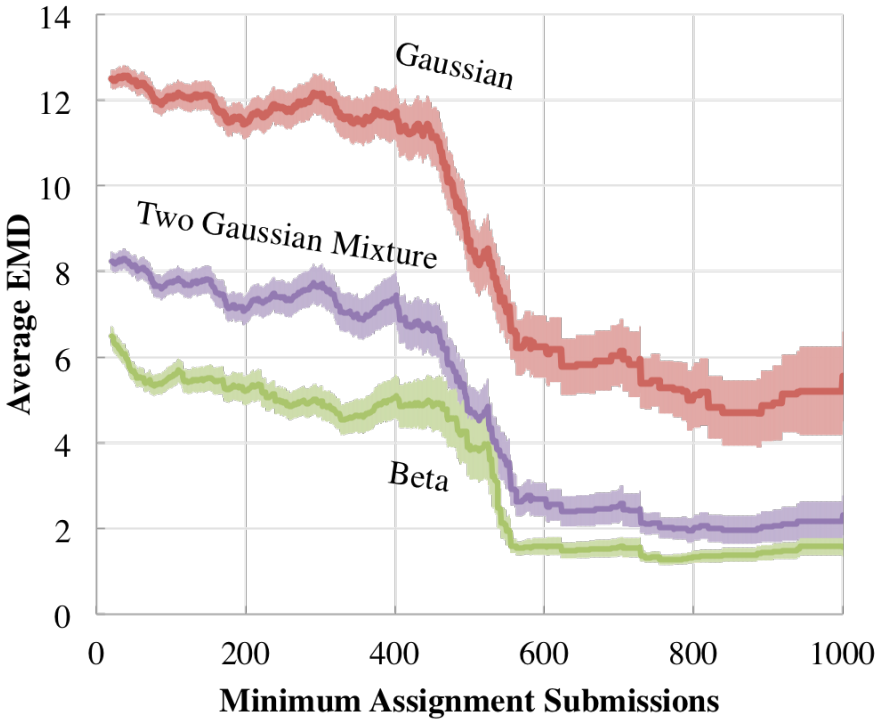
IRT for rubrics

- Assignment and question difficulty do not appear to be normally distributed.

Gradescope assignment score distributions



Gaussian distribution does not fit distributions well



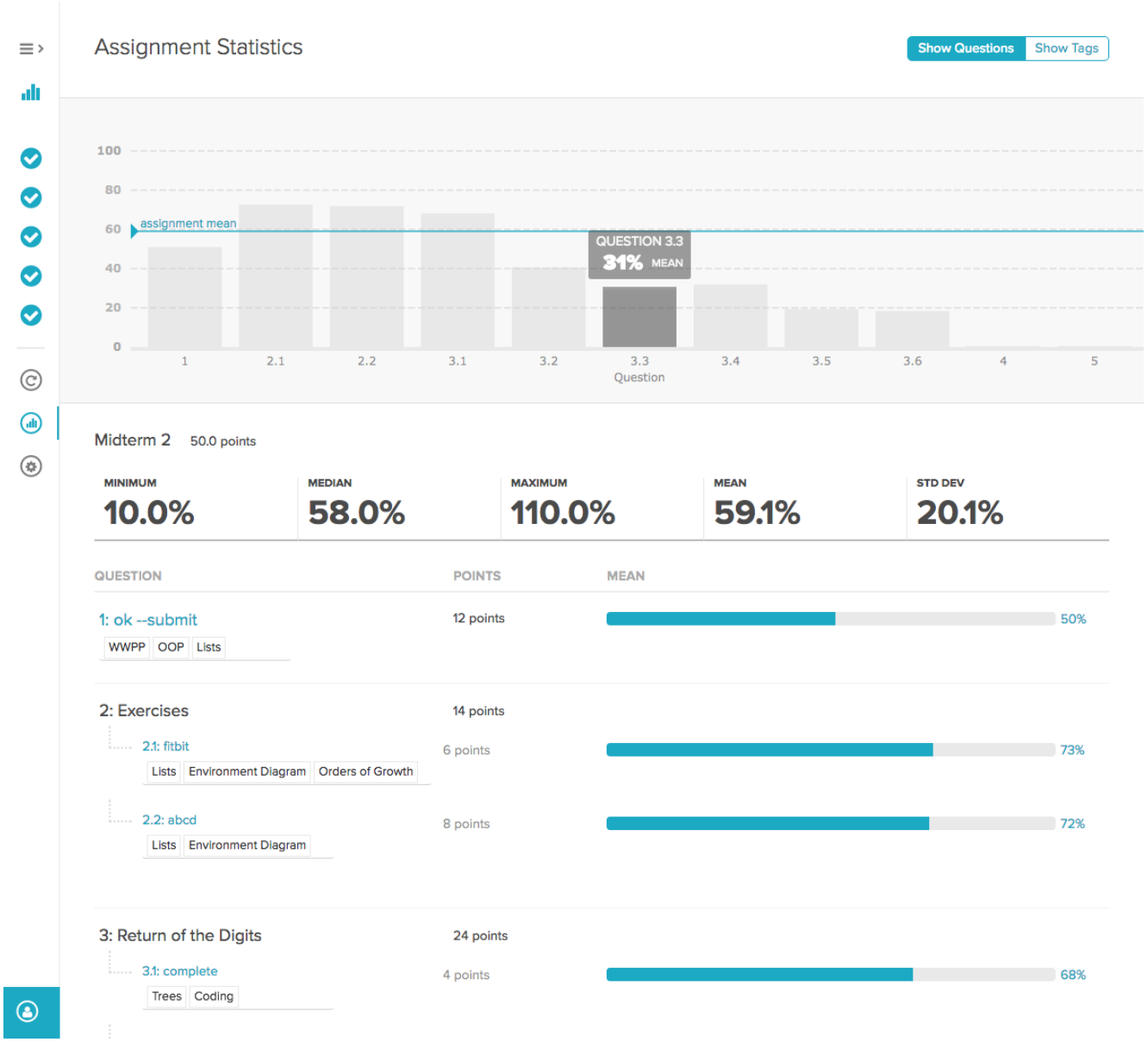
(figure courtesy Chris Piech)

IRT for rubrics

- Would love to collaborate on evaluating polytomous IRT models on our additive and subtractive rubric-graded questions.
- Will probably need to develop new models.

gradescope

Inferring concepts



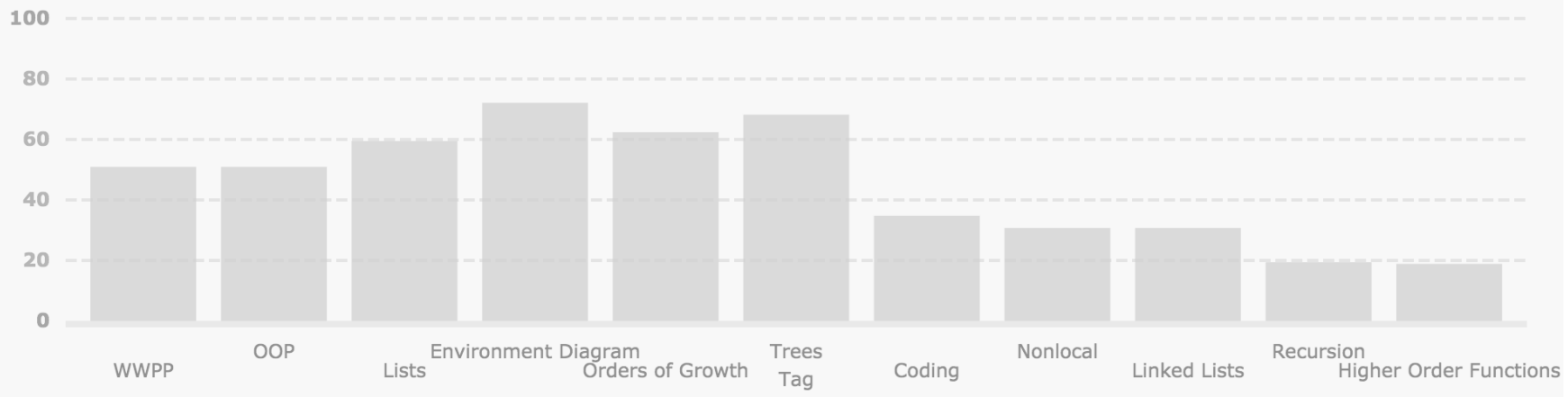
Instructors can tag questions with concepts to see how their assignment breaks down by concept.

gradescope

Inferring concepts

Assignment Statistics

Show Questions | Show Tags



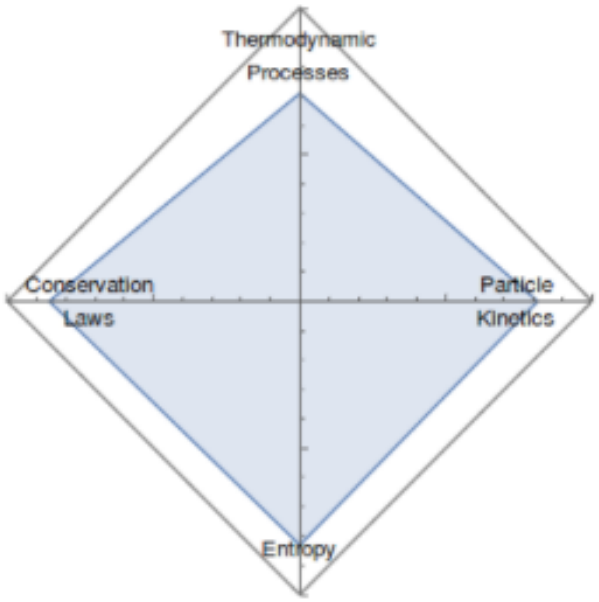
Midterm 2 50.0 points

MINIMUM	MEDIAN	MAXIMUM	MEAN	STD DEV
10.0%	58.0%	110.0%	59.1%	20.1%

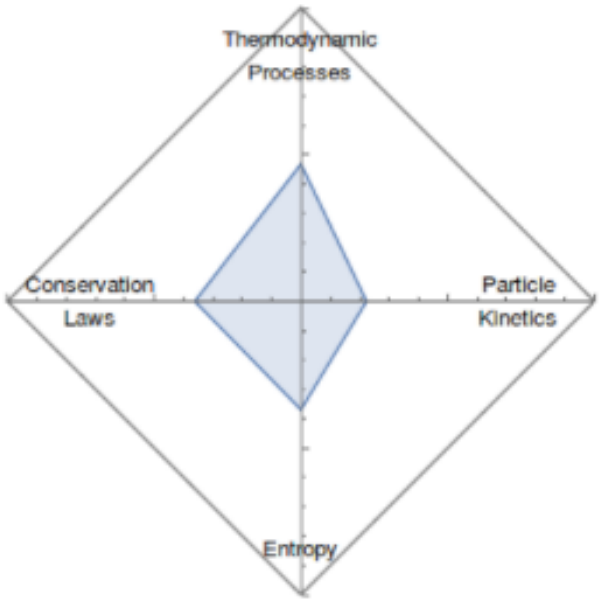
TAG	POINTS	MEAN
WWPP	24 points	50%

Inferring concepts

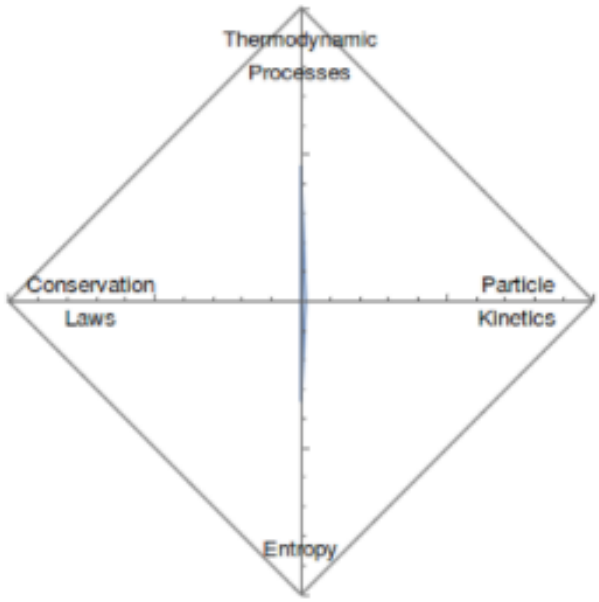
- We would like to infer concepts automatically.
- One way forward: Gradescope user Lenny Evans used topic modeling on rubric item applications.



Best-scoring students



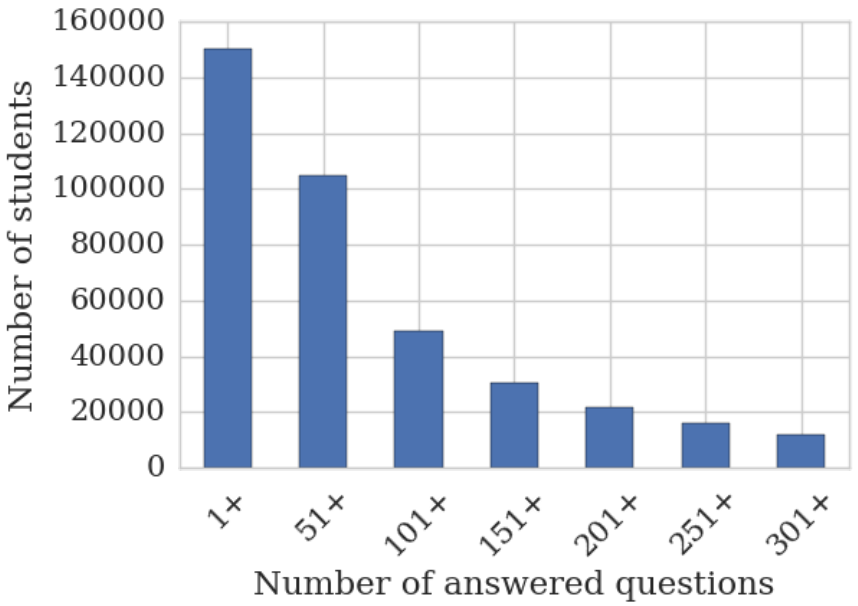
Medium-scoring students

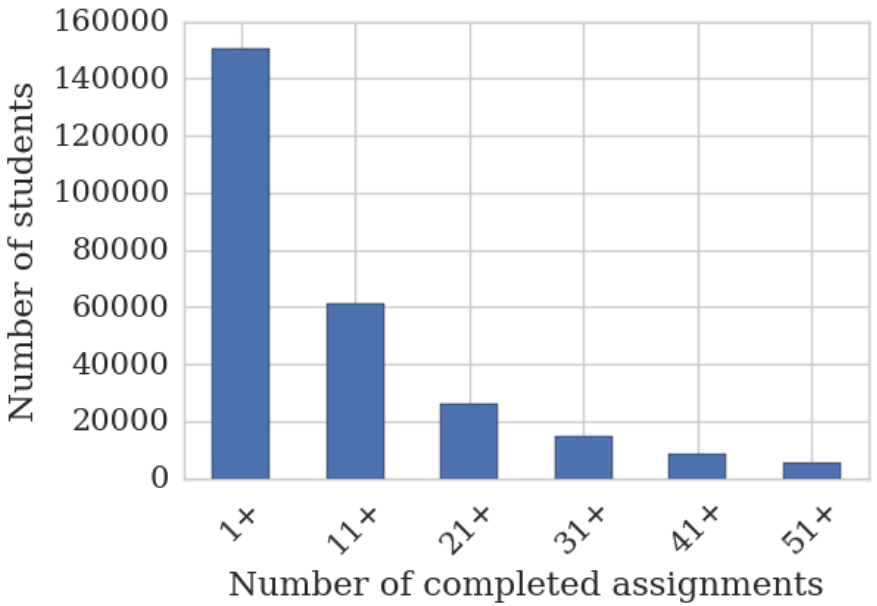


Low-scoring students

http://blog.ultramarineneutrinos.com/topic-modeling-and-gradescope/

- Thousands of students have a significant amount of real work on Gradescope.
- We would like to help institutions with pre-requisite analysis and other tasks.





Assisted grading

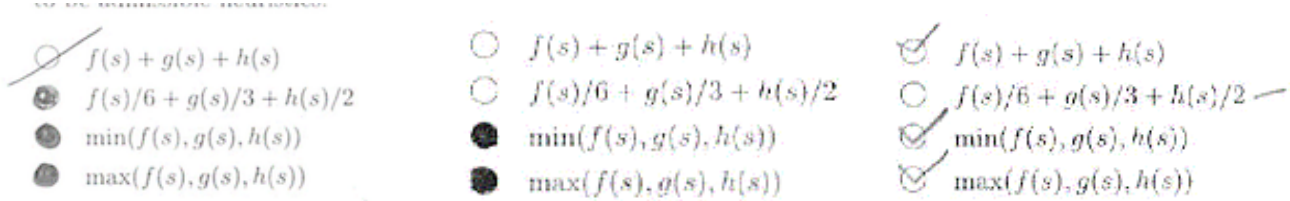
- Our path to auto-grading: grouping answers for the instructors to grade
- Spectrum of question complexity
- Group review interface

Our approach

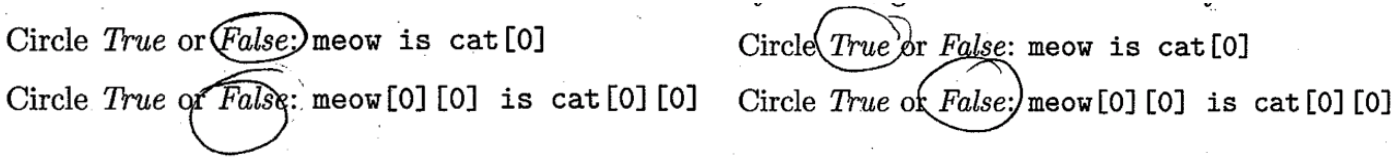
- We value rubric-based grading: partial credit and useful feedback to the student.
  Goal is to scale, not replace, instructor grading.

gradescope Question types: first

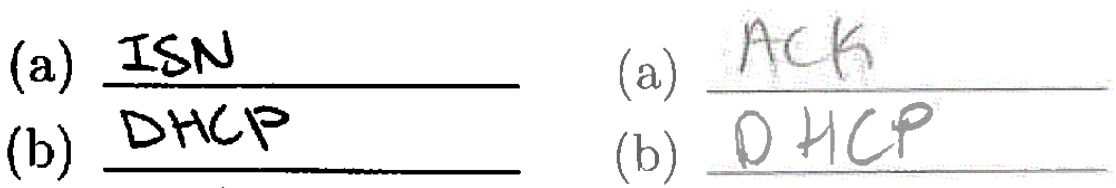
Scantron-like



Freeform M.C.



Constrained text



gradescope Question types: next

Math

Q1.1 [3pt] What is the integral of x?

$\frac{1}{2}x^2 + C$

Q1.1 [3pt] What is the integral of x?

$\frac{1}{2}x^2 + C$

Code

(a) (2 pt) Select the titles of all movies that have a rating greater than 7 in alphabetical order.

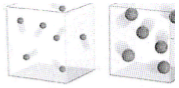
```
select title from ratings
where rating > 7
order by title;
```

(a) (2 pt) Select the titles of all movies that have a rating greater

```
select title from ratings
where rating > 7
order by title
```

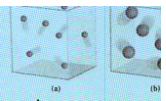
Short answer

2. (5 points) The figure to the right represents three ideal gas samples (a, b, c). Assume that the mass of each particle is proportional to its size, and that all the gas samples are at the same temperature, in which sample is the pressure highest? Explain your reasoning.



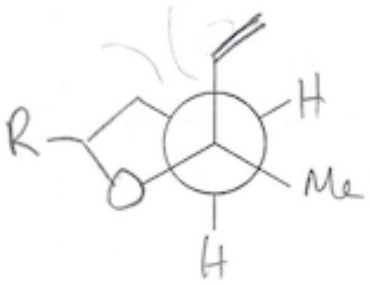
Sample A, because more open space allows to move more freely & b/c particles are moving faster.

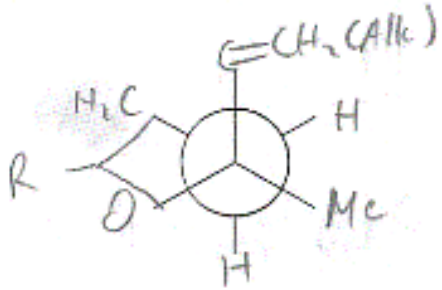
2. (5 points) The figure to the right represents three ideal gas samples (a, b, c). Assume that the mass of each particle is proportional to its size, and that all the gas samples are at the same temperature, in which sample is the pressure highest? Explain your reasoning.

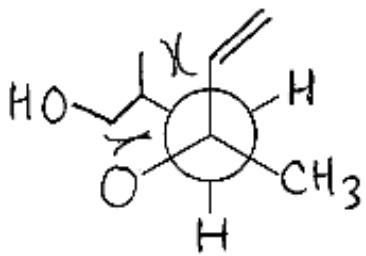


C has the highest pressure. For ideal gases, $P=\frac{nRT}{V}$, since R, T, and V are the same, the gas with larger n has higher pressure. By counting, c has 11 molecules while a and b only have 7.

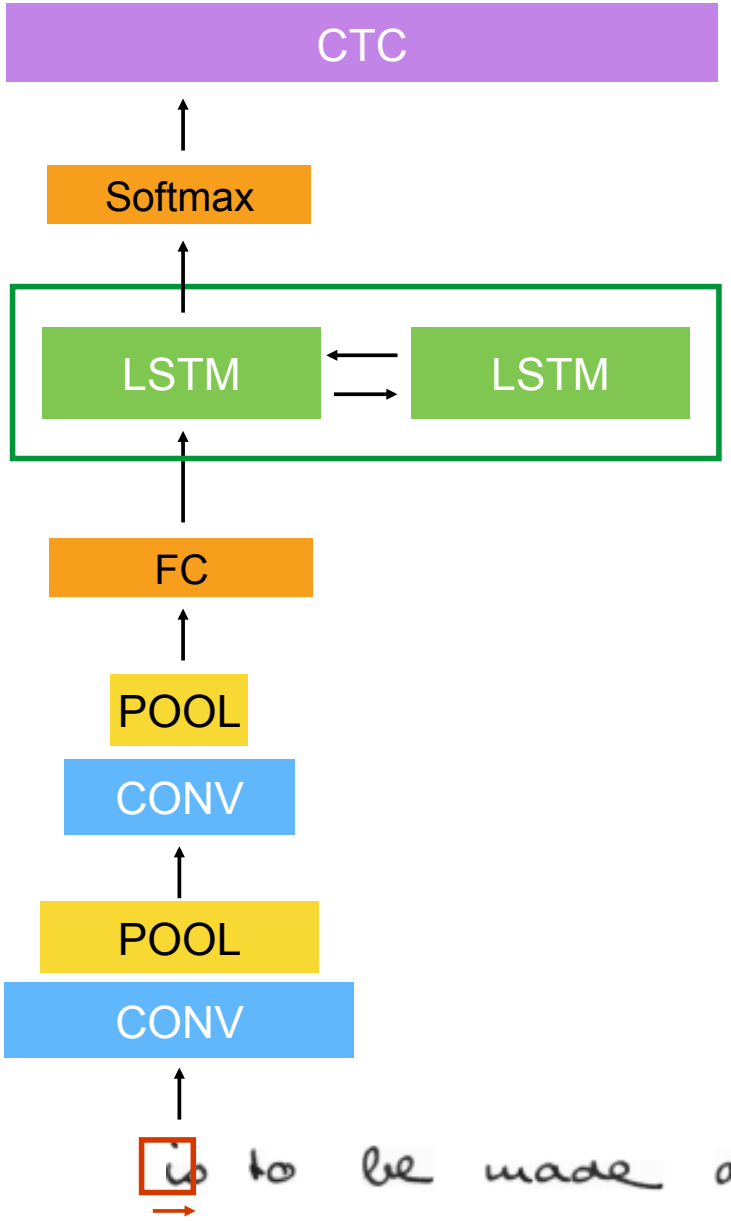
Diagrams







gradescope Handwriting recognition



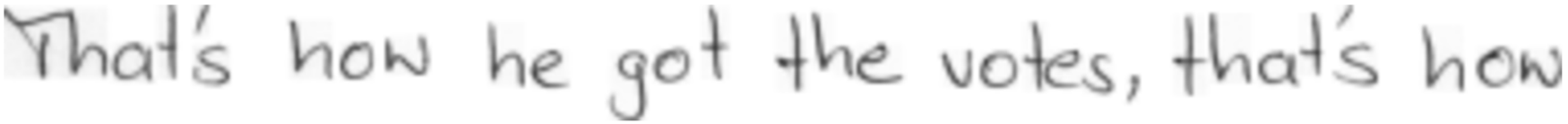
→ “is to be made at a meeting at Labour”

Model architecture based on a combination of LeNet and Alex Graves PAMI 2009.

IAM Offline Handwriting Recognition DB (2002)

	Trainval	Test
# lines	7081	2781
# writers	339	161

gradescope Handwriting recognition



Thal's how he got the votes, that's how

with bleek, watted hair and sfriking

But all three were to the casual

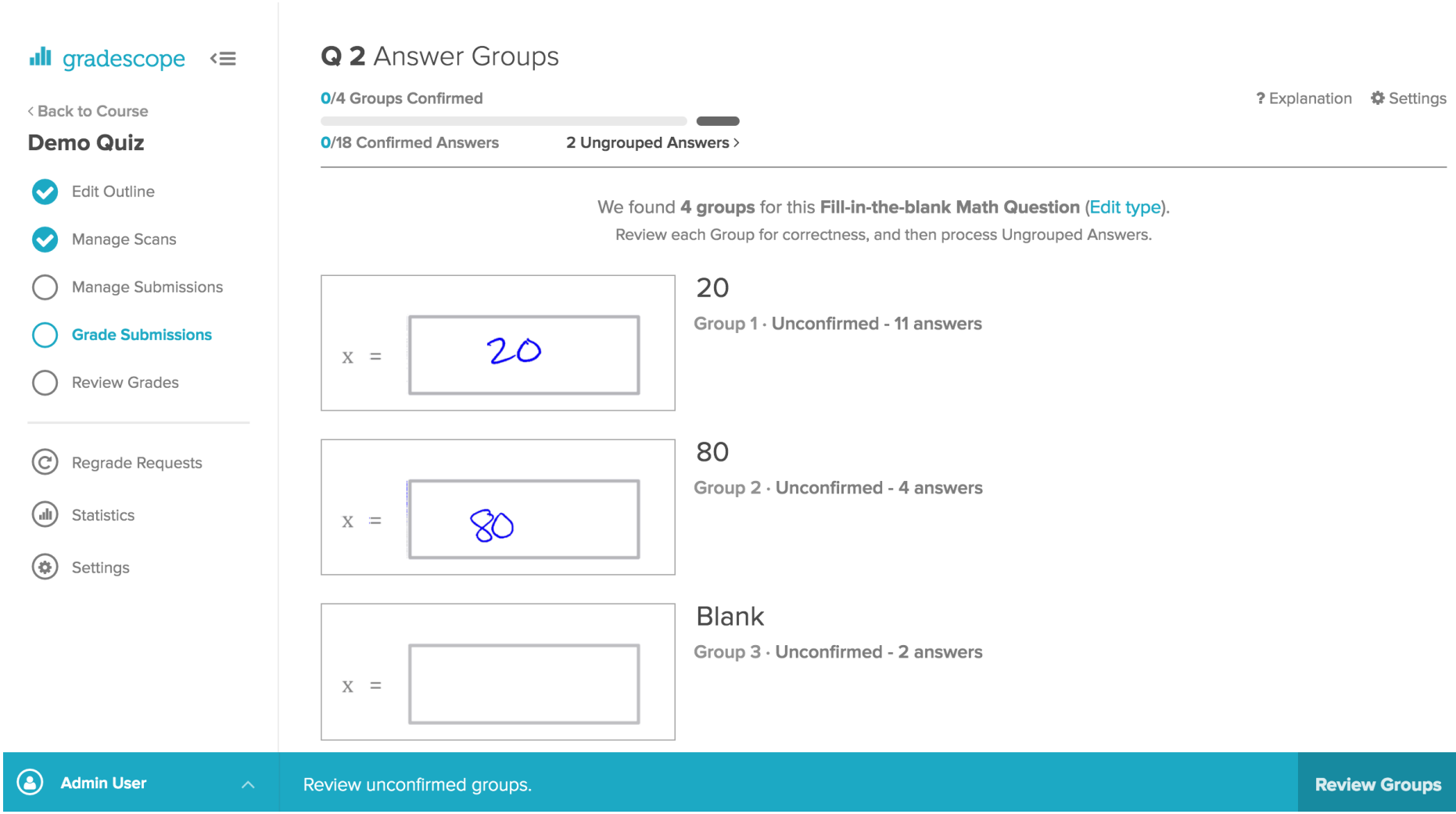
He love frou his breakfait -nook becch

wars impatient far it. Teis affroon, then, wechinbed

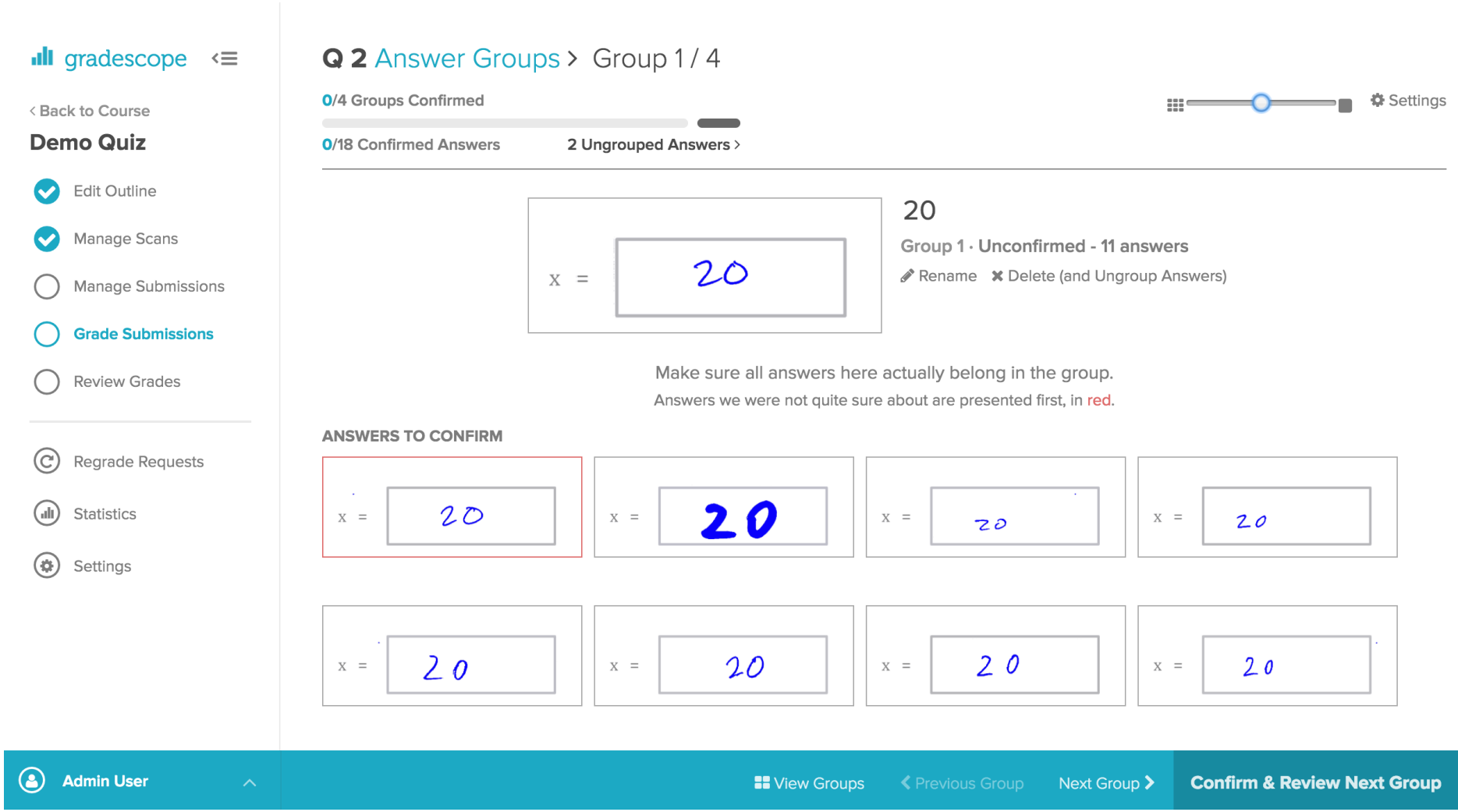
83.5% character accuracy
(no language model)

$$acc = 100 * \left(1 - \frac{\text{insertions} + \text{substitutions} + \text{deletions}}{\text{total length of test set transcriptions}}\right)$$

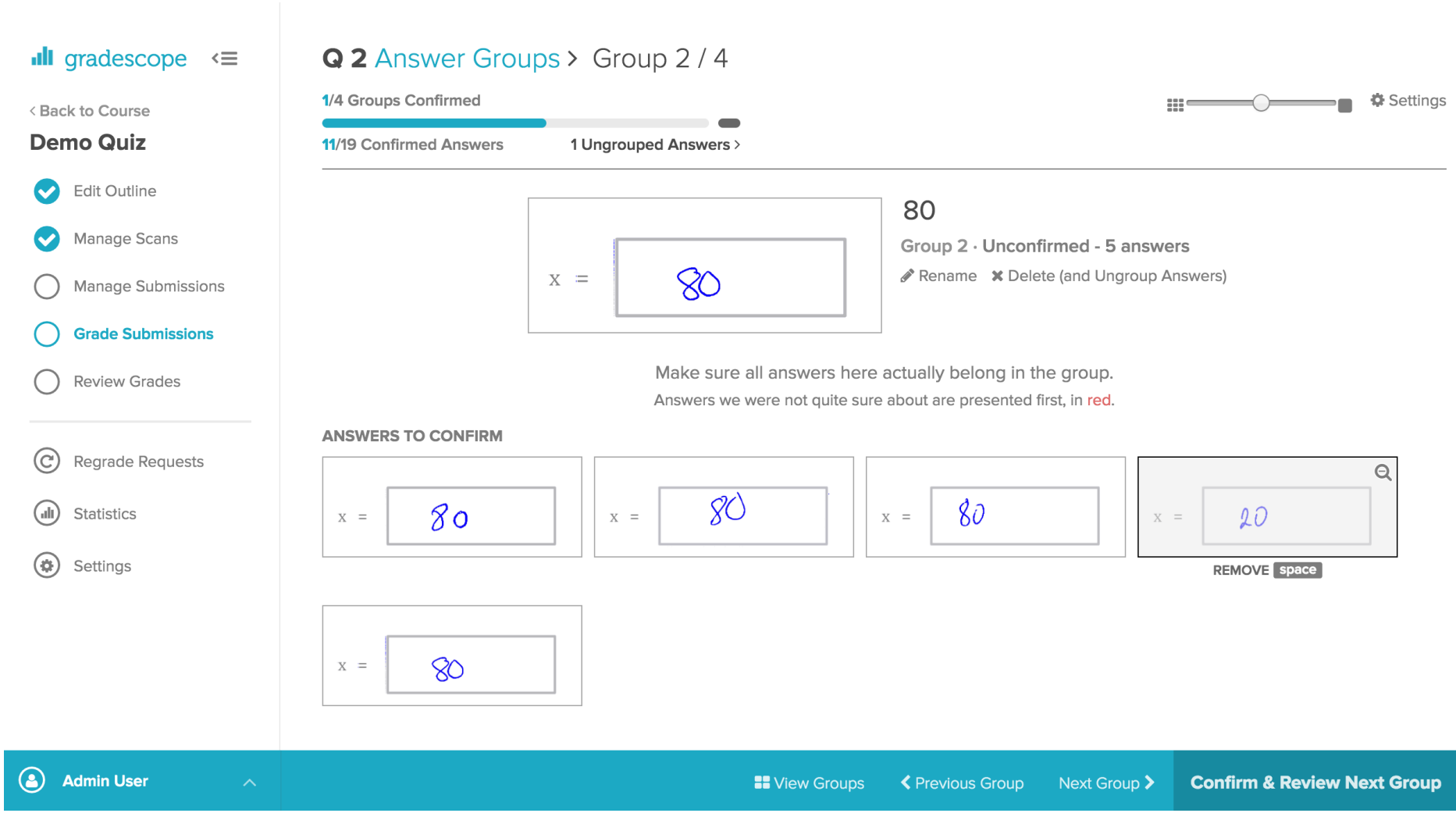
gradescope Group review interface



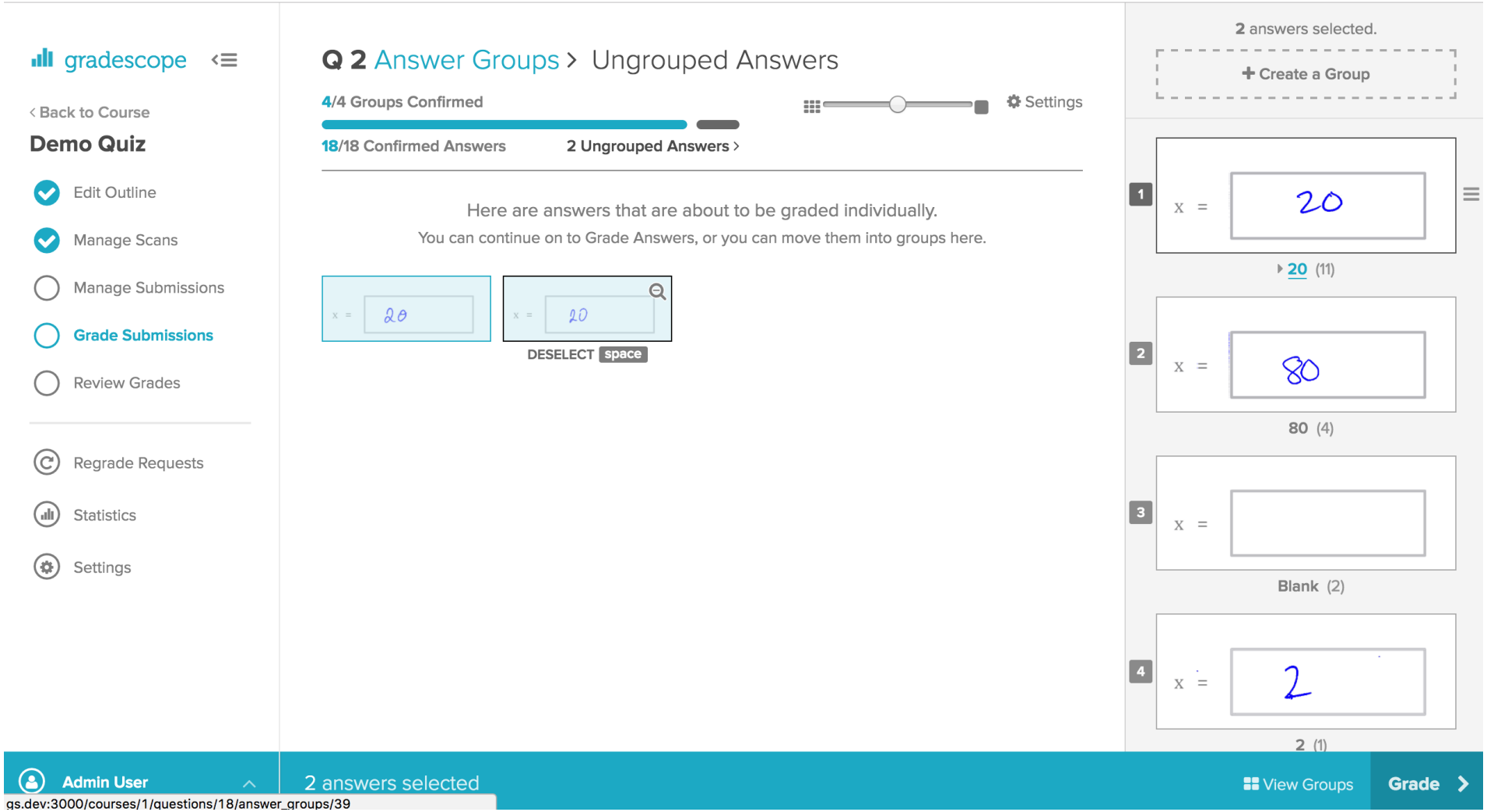
gradescope — Group review interface



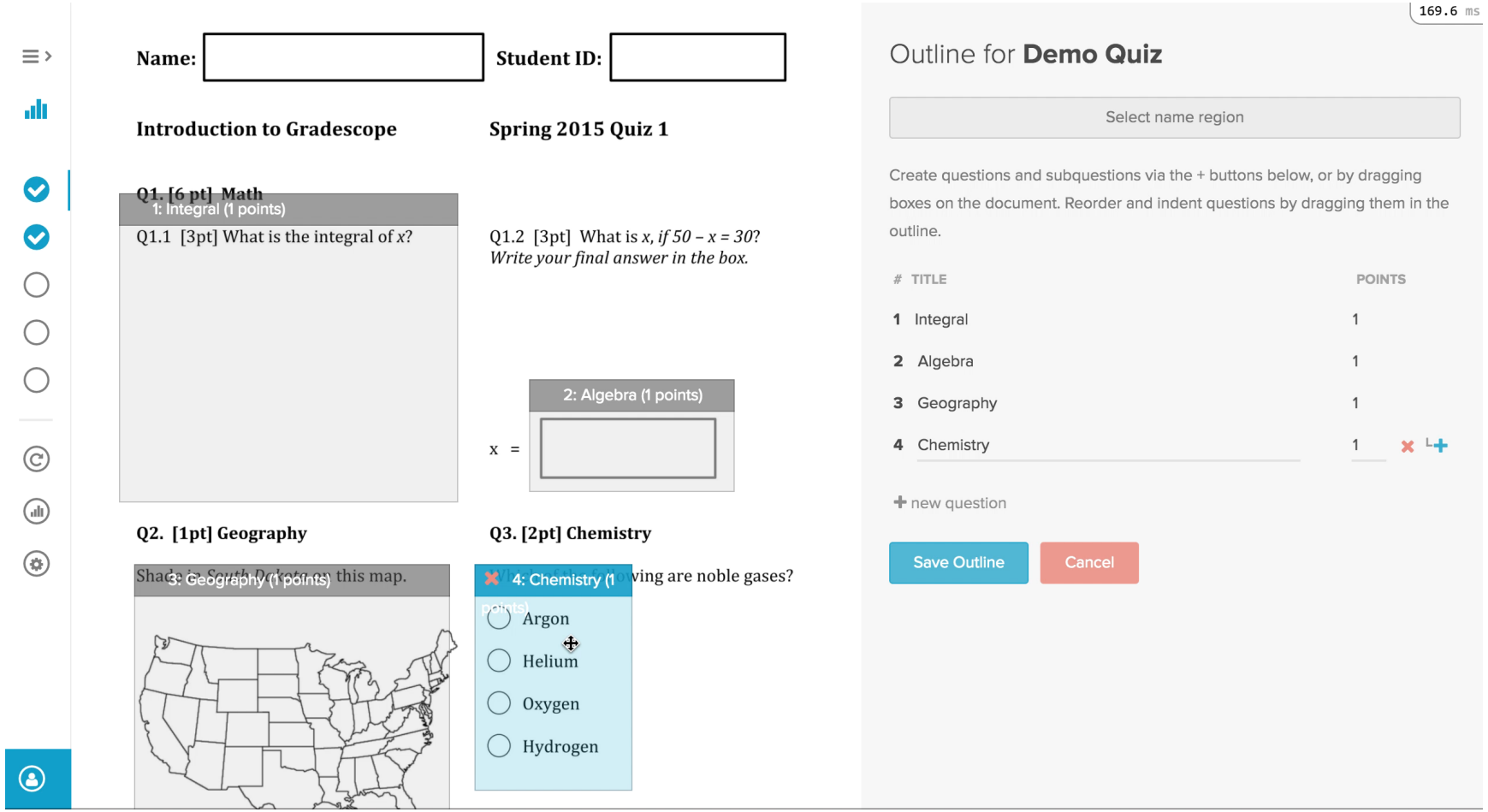
gradescope — Group review interface



gradescope Group review interface



gradescope Group review interface



gradescope

Thank you!

NIPS 2016 Education Workshop

Pieter Abbeel
pabbeel@cs.berkeley.edu

Professor UC Berkeley
Co-founder Gradescope

http://dsp.rice.edu/ml4ed_nips2016