LECTURE 3

SPRING 2021
APPLIED MACHINE LEARNING
CIHANG XIE

SLIDE CREDIT:
NARGES NOROUZI

OFFICE HOURS & SESSIONS

- Monday:
  - **TA Discussion Session** 10 – 11 AM by Minghao Liu
- Tuesday
  - **TA Office Hour** 9 – 10 AM by Minghao Liu
  - **Instructor Office Hour** 12:25 – 1:25 PM
- Wednesday:
  - **Group Tutor Session** 12:30 – 2 PM by Balaram Behera
  - **TA Office Hour** 2 – 3 PM by Molly Zhang
- Friday
  - **TA Discussion Session** 12:30 – 1:30 PM by Molly Zhang
  - **Group Tutor Session** 1:30 – 3 PM by Apala Thakur

GROUPS

Groups (10)

Group 1	1 / 13 students
Group 2	10 / 13 students
Group 3	3 / 13 students
Group 4	Full 13 / 13 students
Group 5	4 / 13 students
Group 6	0 / 13 students
Group 7	Full 13 / 13 students
Group 8	9 / 13 students
Group 9	Full 13 / 13 students
Group 10	Full 13 / 13 students

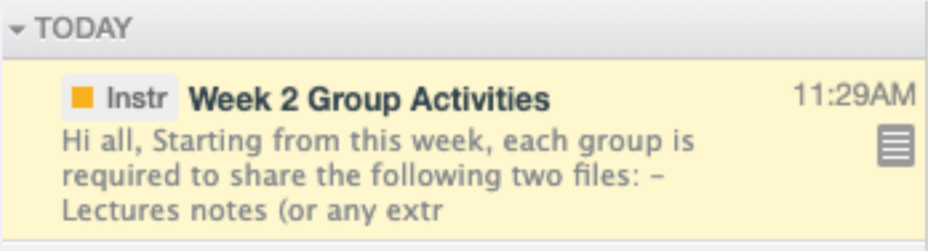
DDL: 11:59PM Wed

https://canvas.ucsc.edu/courses/42540/groups

GROUP ACTIVITIES

- Weekly Piazza Post
  - Lectures notes (or any extra resources on the same topics covered)
  - Exercise questions from weekly topics

DEADLINE is 11:59 pm every Sunday



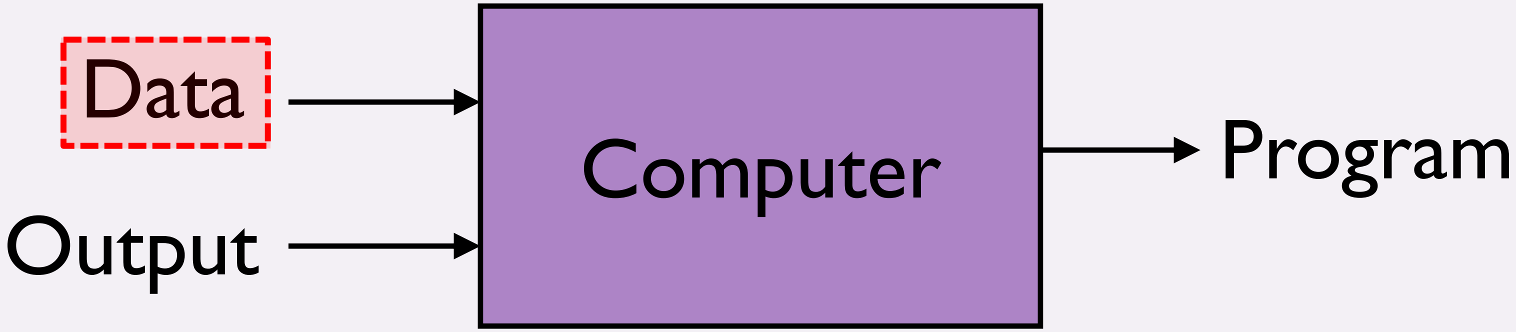
CLASS PARTICIPATION

- Record zoom participation
- In-class quizzes

You will LOSE the **10% class participation and class exercises** if missing class participation for MORE THAN TWO TIMES

LECTURE 2

- Dataset Splitting
- Feature Engineering
- Data Cleaning
- Feature Crossing



EXERCISE

HTTPS://BIT.LY/2OQUMZ2

CROSSING ONE-HOT VECTORS

- Example: applying feature crossing to **latitude** and **longitude**

binned_latitude = [0, 0, 1]

binned_longitude = [0, 1, 0]

binned_latitude(lat) = [0 < lat <= 10,
10 < lat <= 20,
20 < lat <= 30]

binned_longitude(lon) = [0 < lon <= 15,
15 < lon <= 30,
30 < lon <= 45]

Consider ALL possibilities here

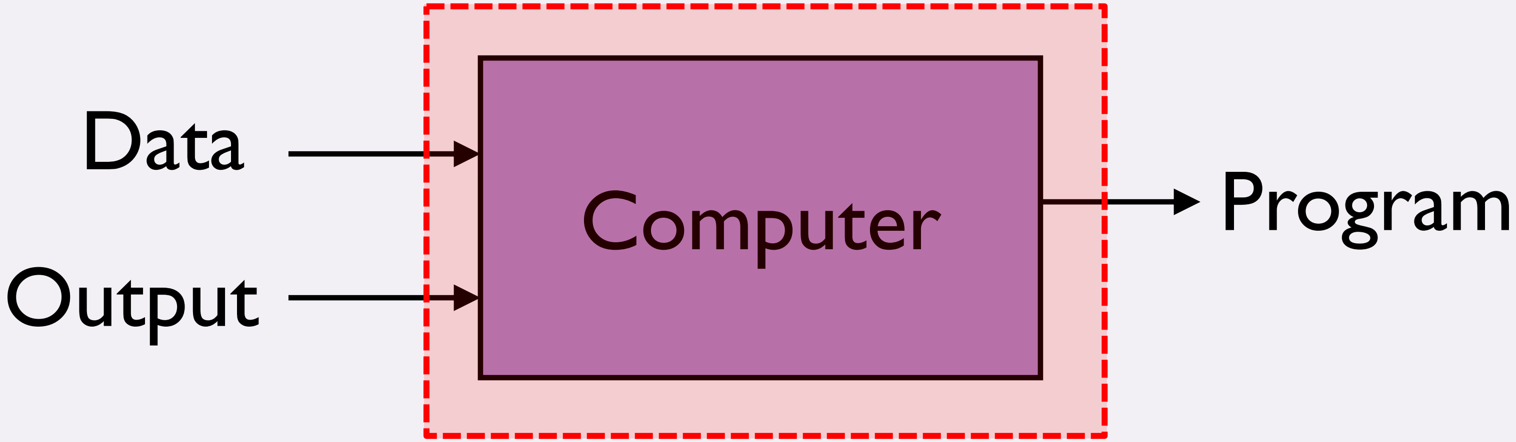
CROSSING ONE-HOT VECTORS

9 possible situations

binned_latitude_X_longitude(lat, lon) = [0 < lat <= 10 AND 0 < lon <= 15,
0 < lat <= 10 AND 15 < lon <= 30,
0 < lat <= 10 AND 30 < lon <= 45,
10 < lat <= 20 AND 0 < lon <= 15,
10 < lat <= 20 AND 15 < lon <= 30,
10 < lat <= 20 AND 30 < lon <= 45,
20 < lat <= 30 AND 0 < lon <= 15,
20 < lat <= 30 AND 15 < lon <= 30,
20 < lat <= 30 AND 30 < lon <= 45]

TODAY

- Linear Regression
- Least Squares Method



REGRESSION

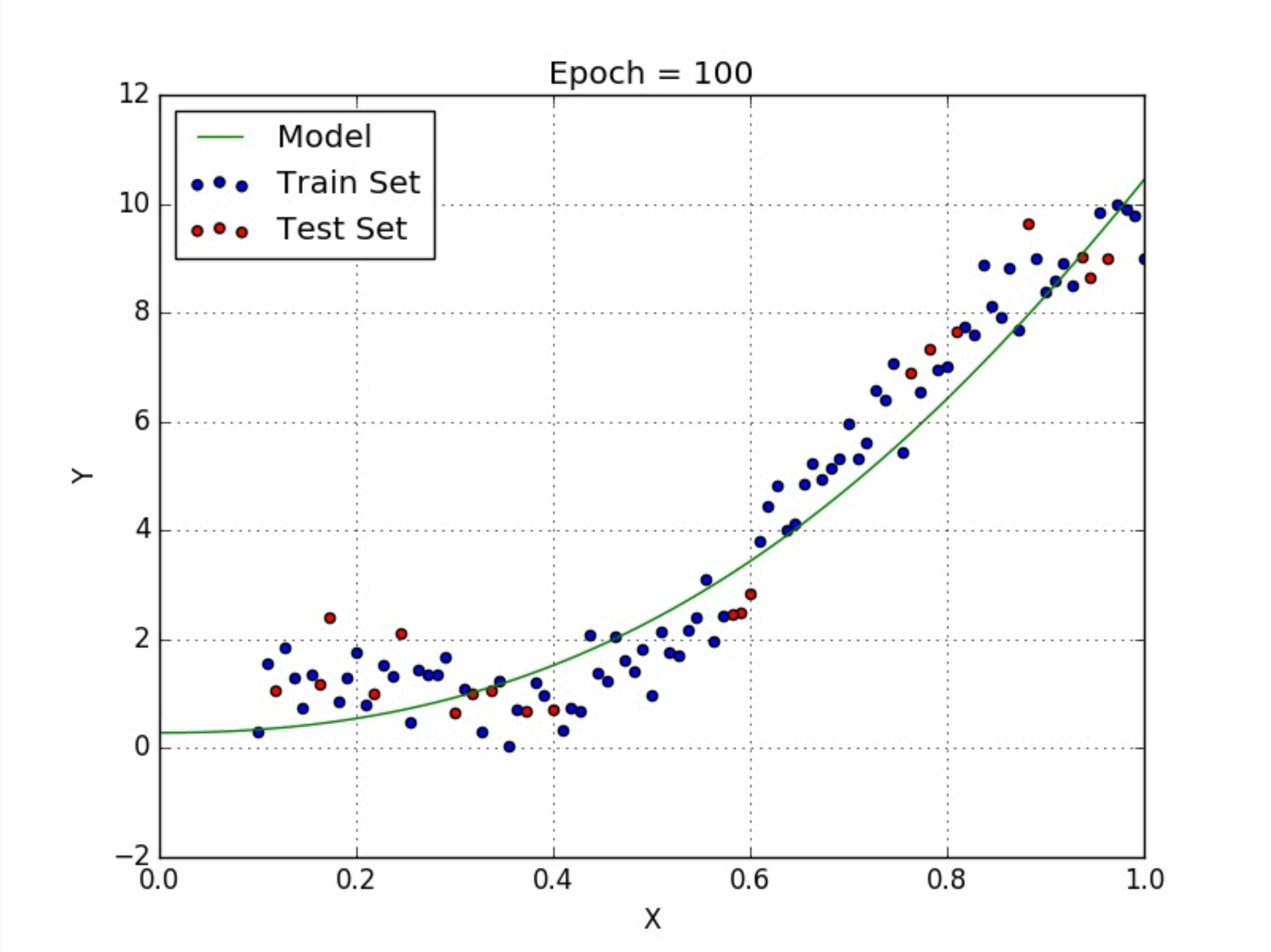


Image credit: https://www.cs.toronto.edu/~frossard/post/multiple_linear_regression/

REGRESSION

- A statistical measure that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables)

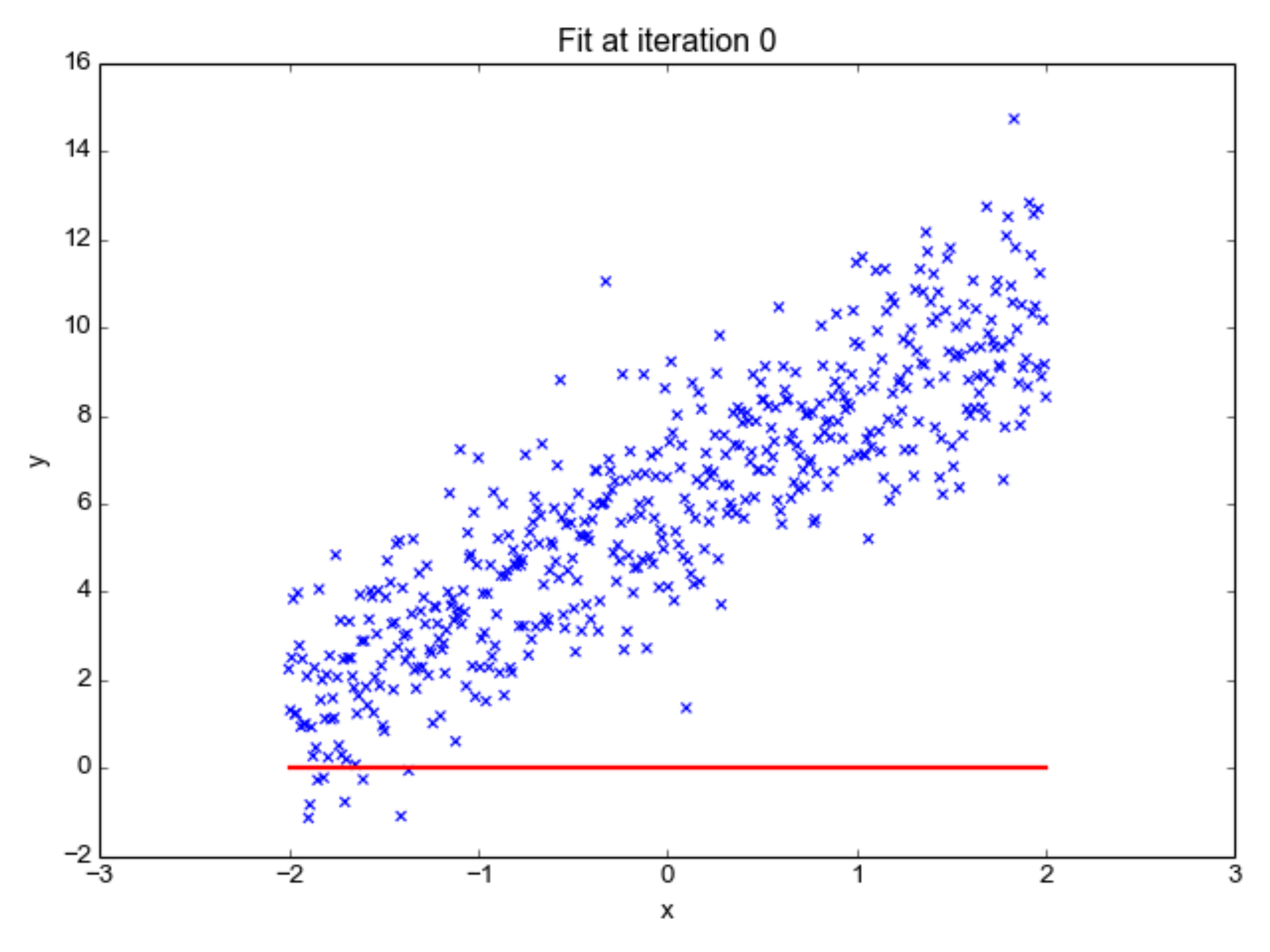
- Forecast value of a dependent variable (Y) from the value of independent variables ($X_1, X_2, X_3, \ldots$)

- It is widely used for prediction, estimation, hypothesis testing, and modeling causal relationships

DEPENDENT & INDEPENDENT VARIABLES

- Independent variables are regarded as inputs to a system and may take on different values freely.
- Dependent variables are those values that change as a consequence of changes in other values in the system.
- Independent variable is also called as predictor or explanatory variable and it is denoted by X.
- Dependent variable is also called as response variable and it is denoted by Y.

THE FIRST ORDER LINEAR MODEL



Assume such relationship is **linear**

Image credit: https://towardsdatascience.com/linear-regression-the-easier-way-6f941aa471ea

THE FIRST ORDER LINEAR MODEL

$$Y = \theta_0 + \theta_1 X$$

- Y = dependent/outcome/response variable
- X = independent/predictor/explanatory variable
- θ_0 = Y-intercept
- θ_1 = slope of the line

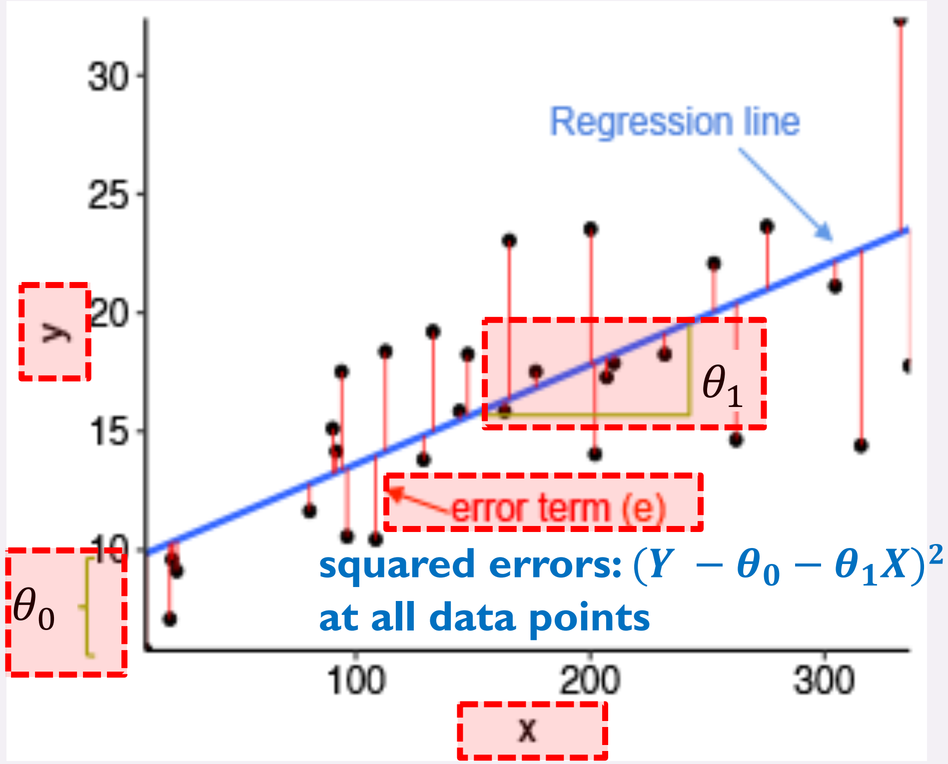


Image credit: https://nextjournal.com/intelrefinery/simple-linear-regression

REGRESSION HYPOTHESIS

- Hypothesis:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ \ldots + \theta_d x_d = \sum_{j=0}^{d} \theta_j x_j$$

Our data has d-dimension

e.g., {house size, house location, …, year built}

REGRESSION DATA

- Given
  - Data:

    $$X = \{x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(n)}\} \text{ where } x^{(i)} \in \mathcal{R}^d$$

  - Corresponding labels:

    $$Y = \{y^{(1)}, y^{(2)}, y^{(3)}, \dots, y^{(n)}\} \text{ where } y^{(i)} \in \mathcal{R}$$

LEAST SQUARES LINEAR REGRESSION

- Cost Function

Summation over the whole dataset

$$Cost(\theta) = \frac{1}{2\times n} \sum_{i=0}^{n} \underbrace{\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2}_{\text{The squared error on a single data point}}$$

Averaging

- Fit by solving

$$\min_{\theta} Cost(\theta)$$

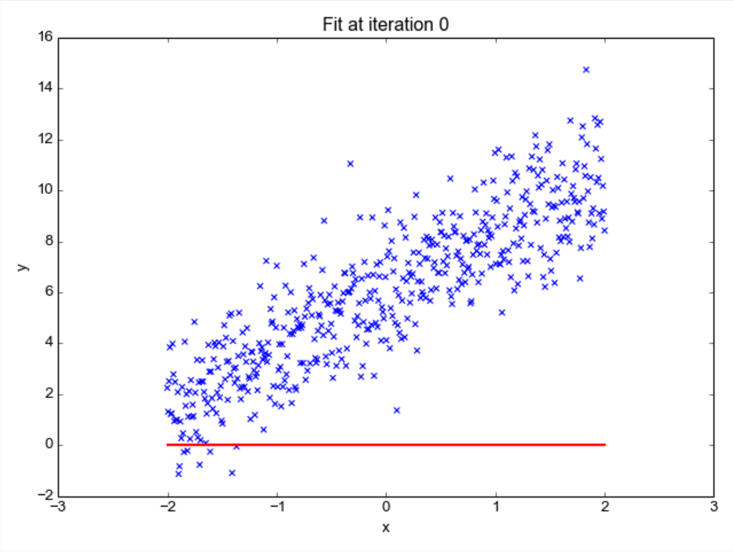


Image credit: https://towardsdatascience.com/linear-regression-the-easier-way-6f941aa471ea

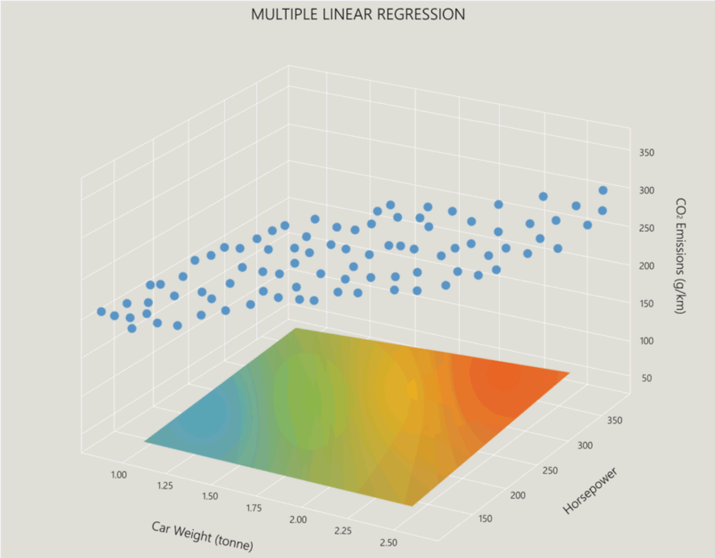
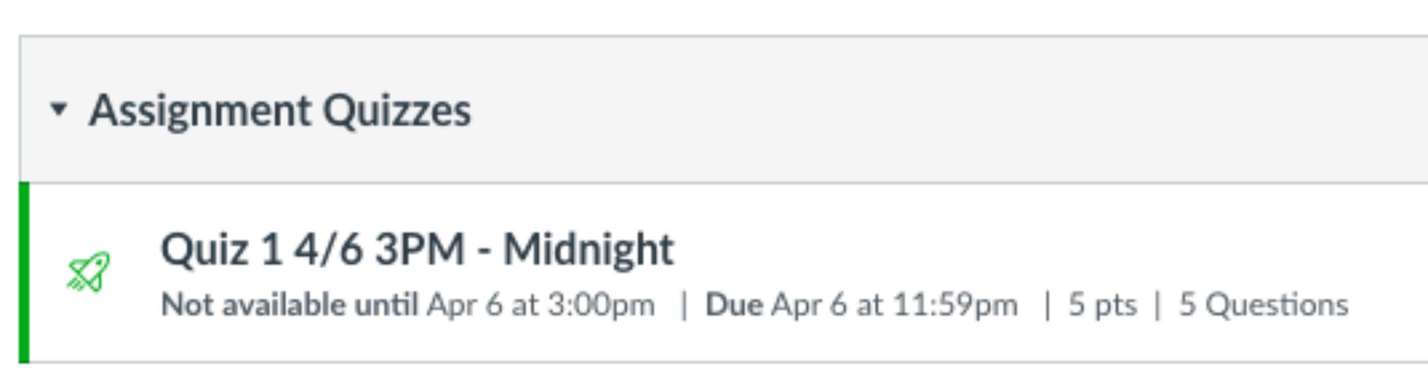


Image credit: https://jaggiz.wordpress.com/2019/10/26/data-and-ai-regression/

QUESTIONS?

QUIZ 1


Assignment Quizzes
Quiz 1 4/6 3PM - Midnight
Not available until Apr 6 at 3:00pm | Due Apr 6 at 11:59pm | 5 pts | 5 Questions