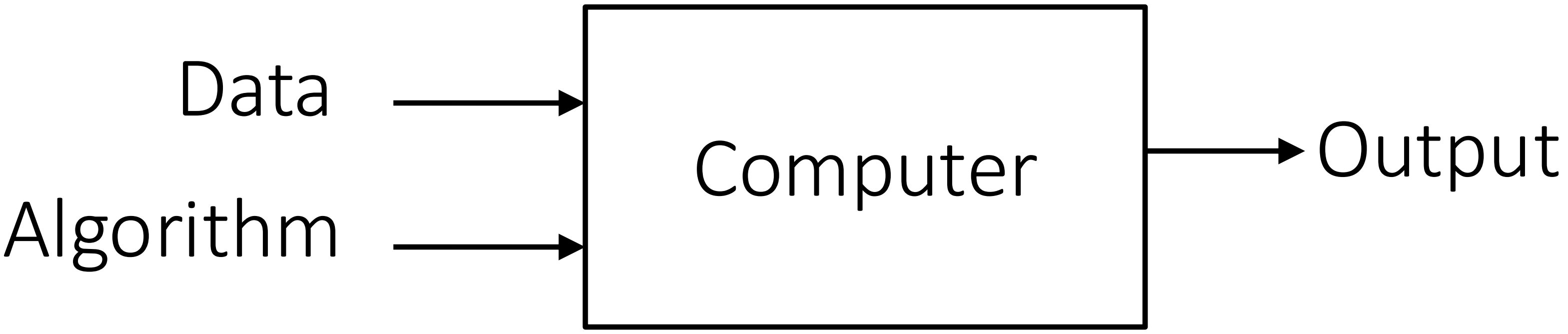
UVA CS 4774: Machine Learning

S0: Lecture 00: Weekly Quiz and QA

Dr. Yanjun Qi

University of Virginia
Department of Computer Science

Traditional Programming



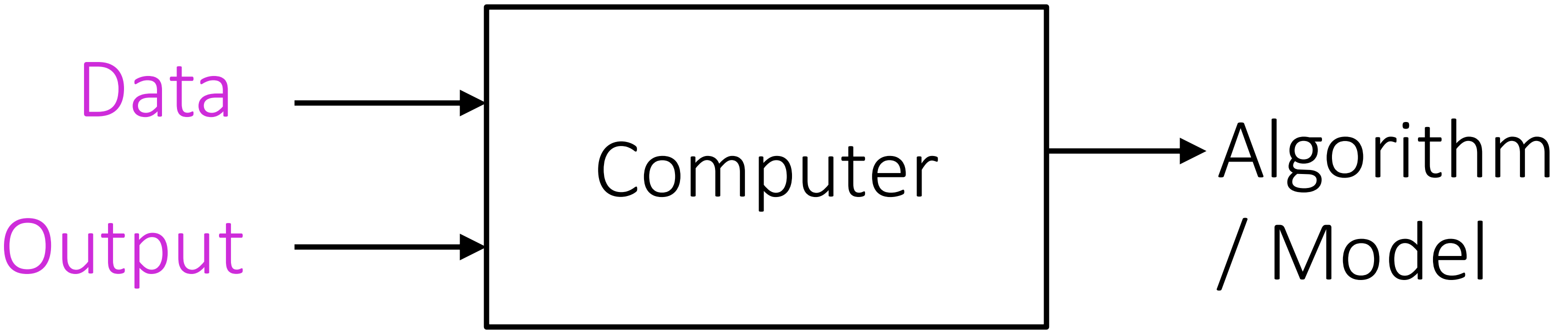
Data

Algorithm

Computer

Output

Machine Learning



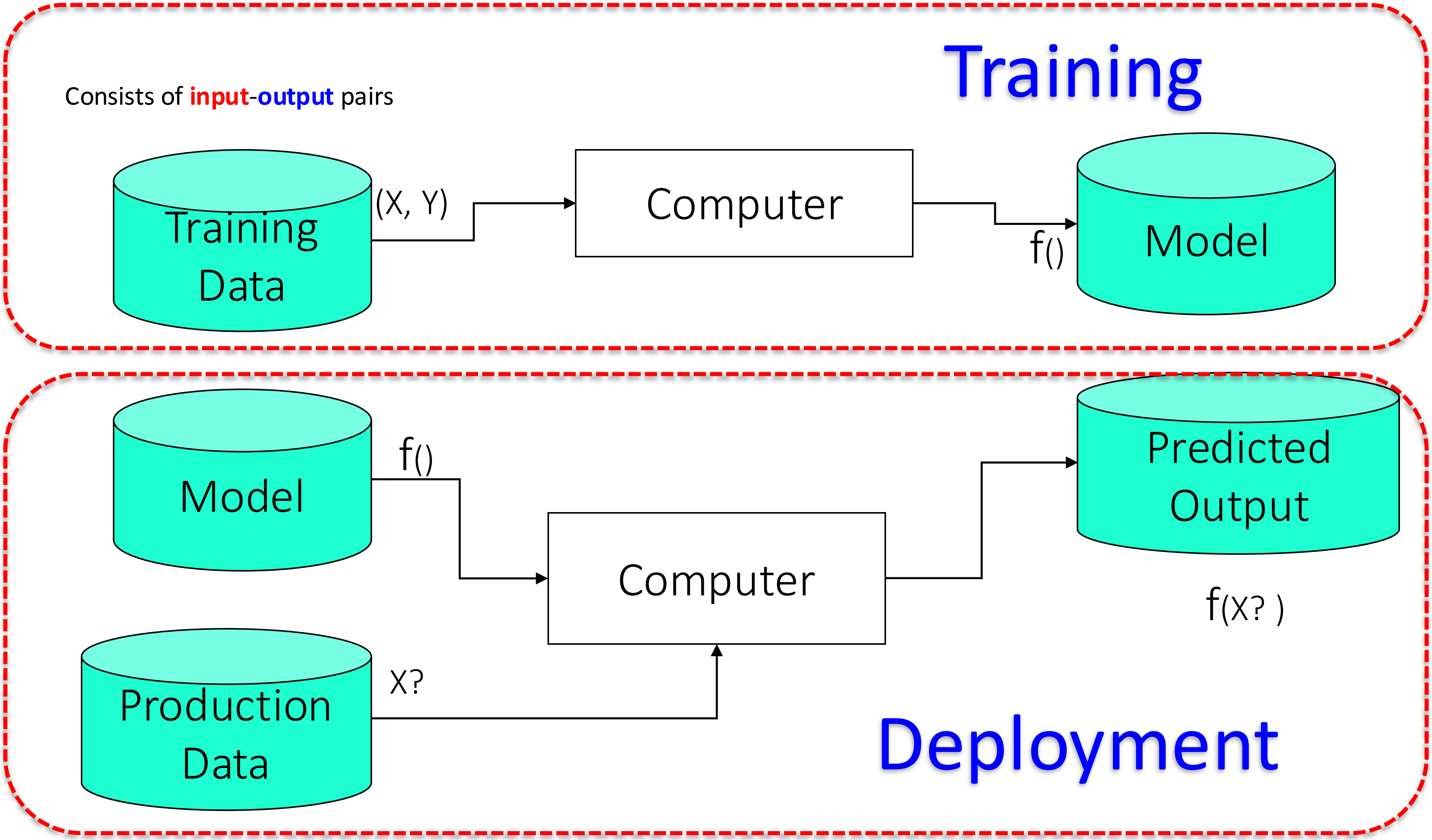
Data

Output

Computer

Algorithm / Model

Two Modes of Machine Learning



Training

Consists of **input**-**output** pairs

Training Data

(X, Y)

Computer

f()

Model

Model

f()

Computer

Predicted Output

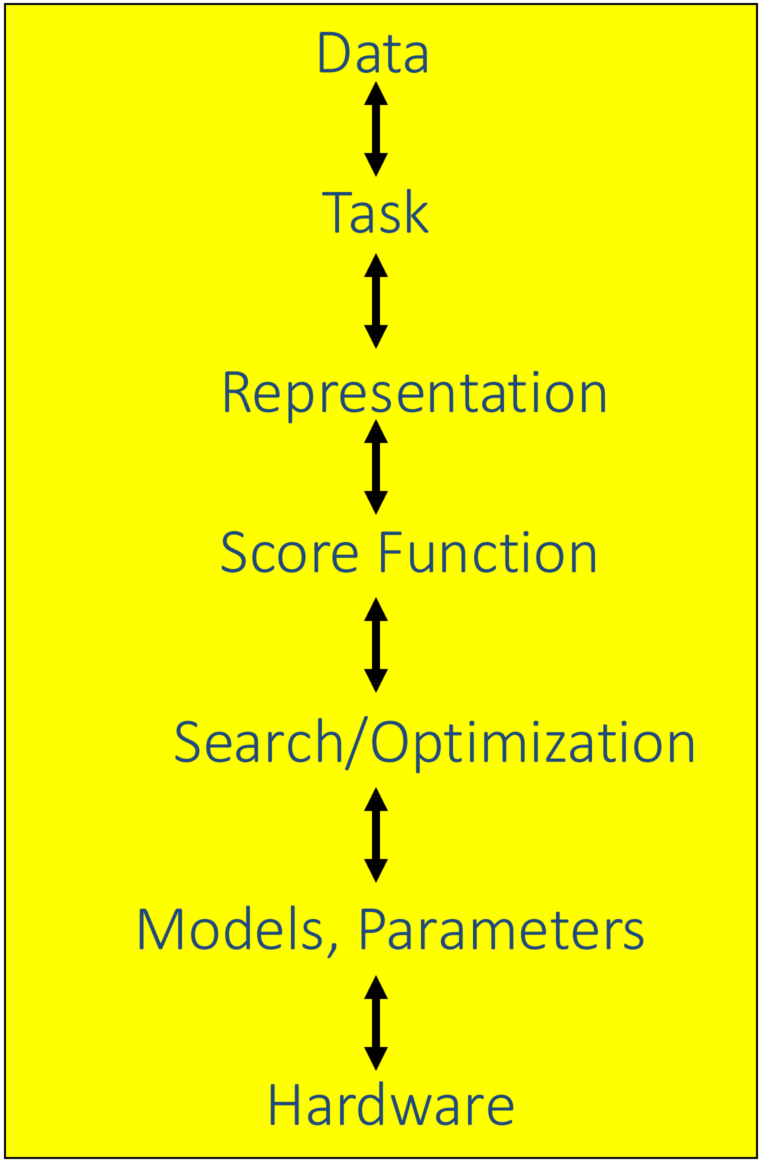
f(X?)

Production Data

X?

Deployment

Machine Learning in a Nutshell



Data

Task

Representation

Score Function

Search/Optimization

Models, Parameters

Hardware

ML grew out of work in AI

Optimize a performance criterion using example data or past experience,

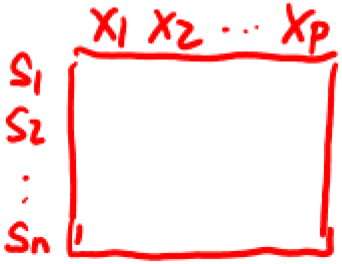
Aiming to generalize to unseen data

Rough Sectioning of this Course

- S1. Basic Supervised Regression + Tabular Data
- S2. Basic Deep Learning + 2D Imaging Data
- S3. Generative and Deep + 1D Sequence Text Data
- S4. Advanced Supervised learning + Tabular Data
- S5. Not Supervised
- S6: Wrap Up + (a few invited tasks, e.g. on AWS)

Course Content Plan → Regarding Data

❑ Tabular / Matrix



❑ 2D Grid Structured: Imaging

x_1	x_2	x_3
x_4	x_5	x_6

❑ 1D Sequential Structured: Text

❑ Graph Structured (Relational)

❑ Set Structured / 3D /

Course Content Plan → Regarding Tasks

- ❑ Regression (supervised)

  Y is a continuous

- ❑ Learning theory

  About f()

- ❑ Classification (supervised)

  Y is a discrete

- ❑ Unsupervised models

  NO Y

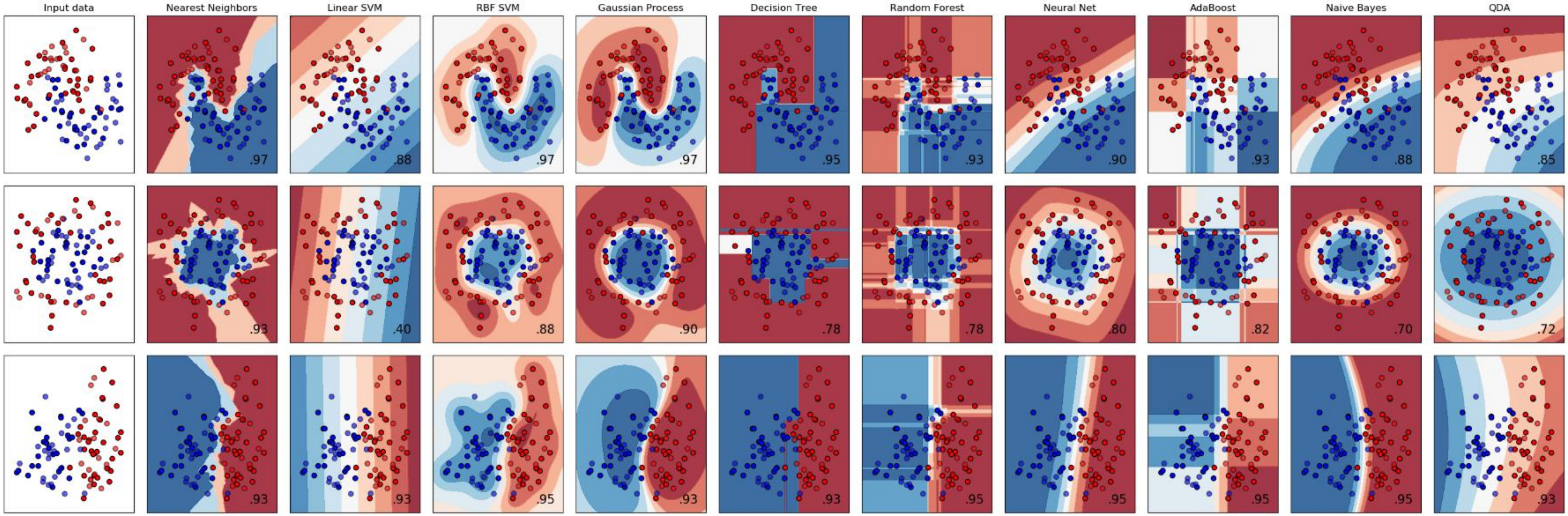
- ❑ Graphical models

  About interactions among Y,X1,. Xp

- ❑ Reinforcement Learning

  Learn to Interact with environment

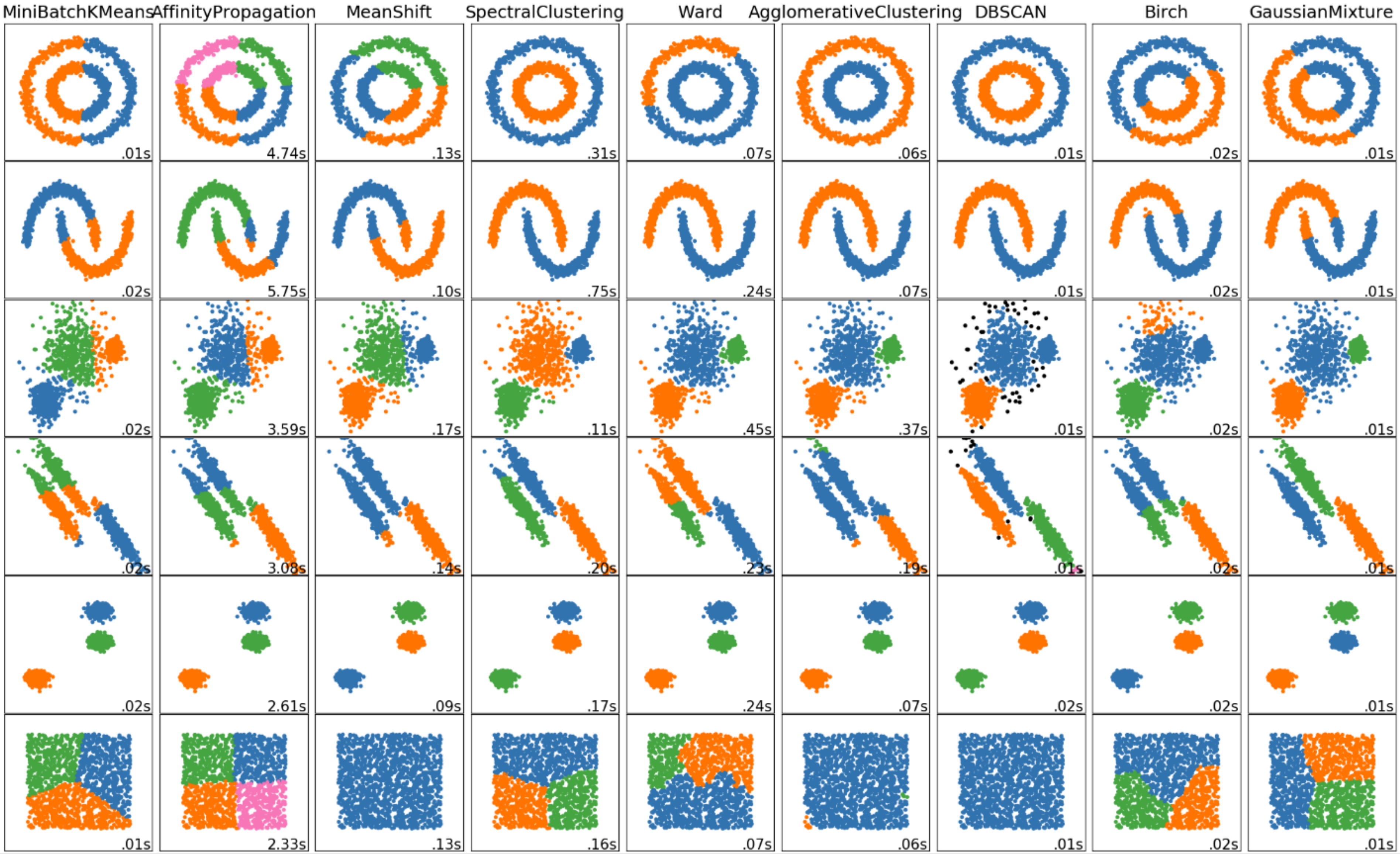
https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html



- ✓different assumptions on data
- ✓different scalability profiles at training time
- ✓different latencies at prediction (test) time
- ✓different model sizes (embedability in mobile devices)
- ✓different level of model interpretability / robustness

Adapt from Olivier Grisel's talk

https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html

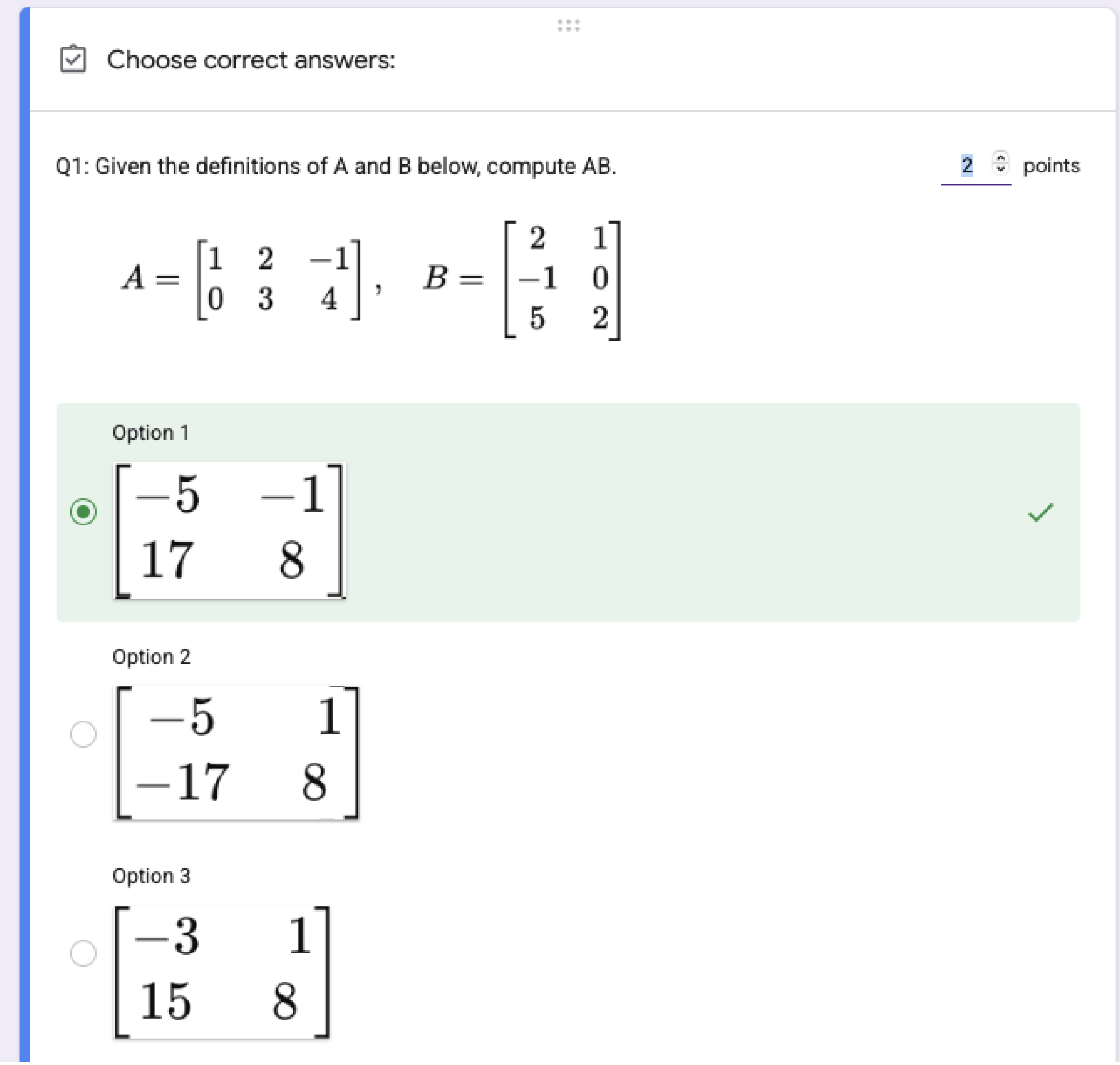


✓different assumptions on data

✓different scalability profiles

✓different model sizes (embedability in mobile devices)

Quiz 1



☑ Choose correct answers:

Q1: Given the definitions of A and B below, compute AB.

2 points

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 0 & 3 & 4 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 1 \\ -1 & 0 \\ 5 & 2 \end{bmatrix}$$

Option 1

● $\begin{bmatrix} -5 & -1 \\ 17 & 8 \end{bmatrix}$ ✓

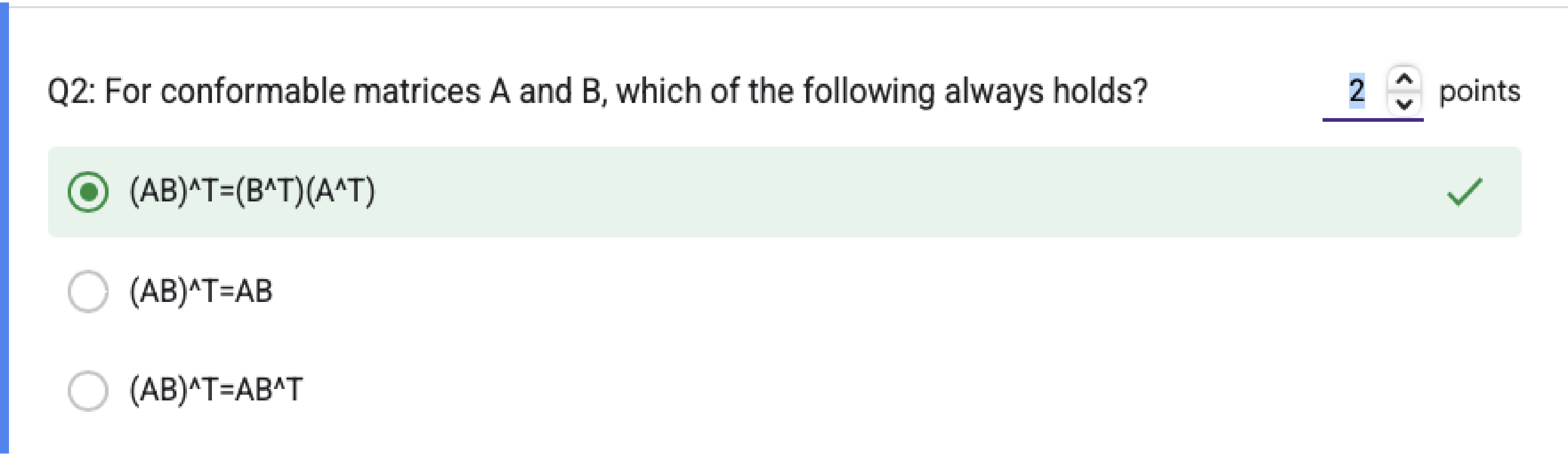
Option 2

○ $\begin{bmatrix} -5 & 1 \\ -17 & 8 \end{bmatrix}$

Option 3

○ $\begin{bmatrix} -3 & 1 \\ 15 & 8 \end{bmatrix}$

Quiz 1



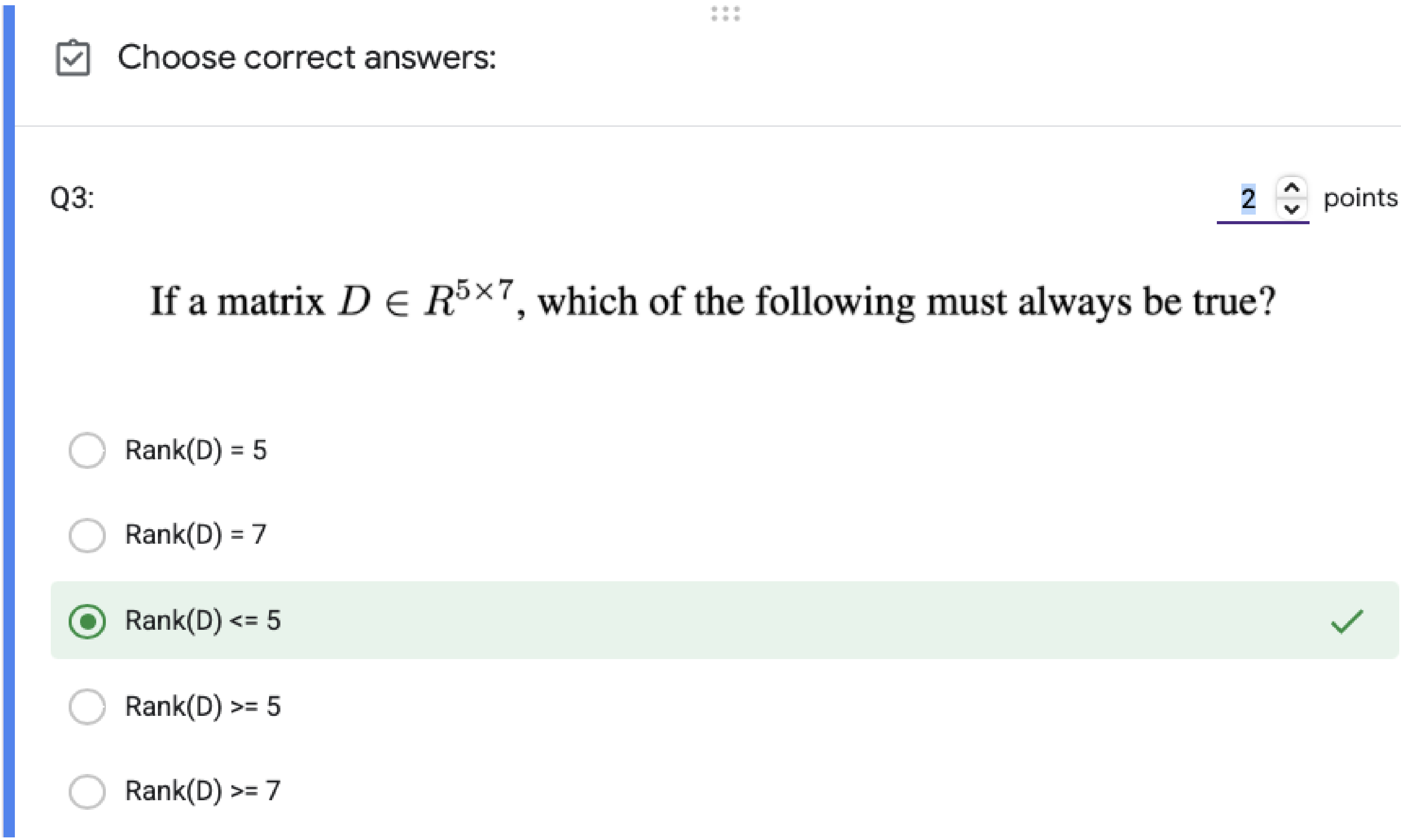
Q2: For conformable matrices A and B, which of the following always holds? 2 points

● (AB)^T=(B^T)(A^T) ✓

○ (AB)^T=AB

○ (AB)^T=AB^T

Quiz 1



☑ Choose correct answers:

Q3: 2 points

If a matrix $D \in R^{5\times 7}$, which of the following must always be true?

○ Rank(D) = 5

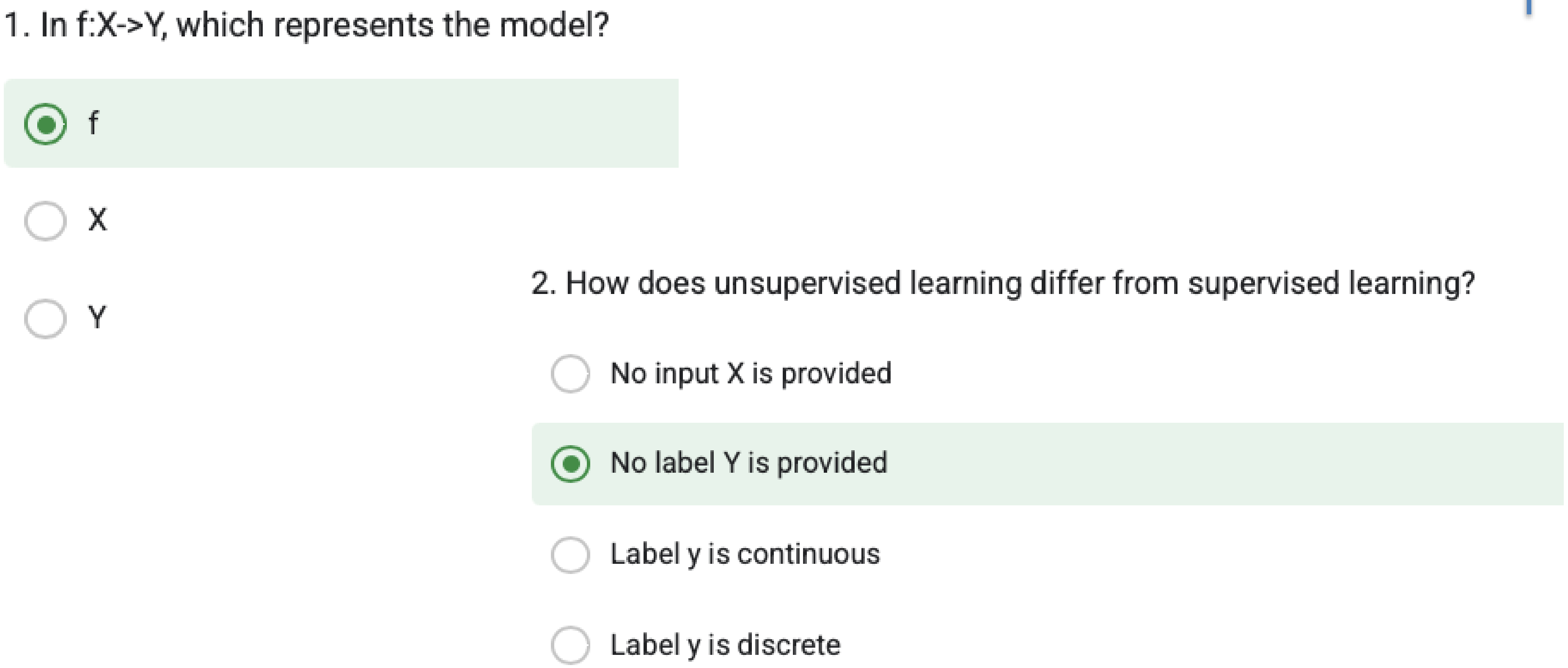
○ Rank(D) = 7

● Rank(D) <= 5 ✓

○ Rank(D) >= 5

○ Rank(D) >= 7

Quiz 2



1. In f:X->Y, which represents the model?

● f

○ X

○ Y

2. How does unsupervised learning differ from supervised learning?

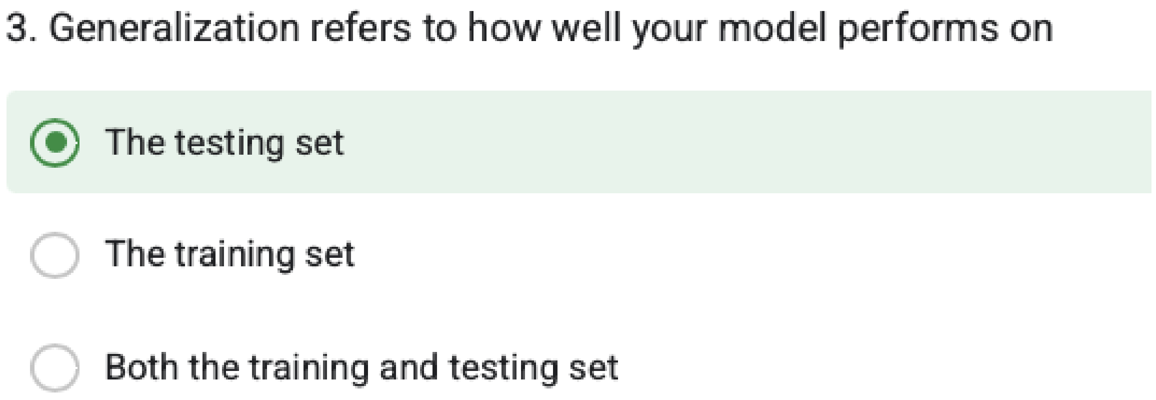
○ No input X is provided

● No label Y is provided

○ Label y is continuous

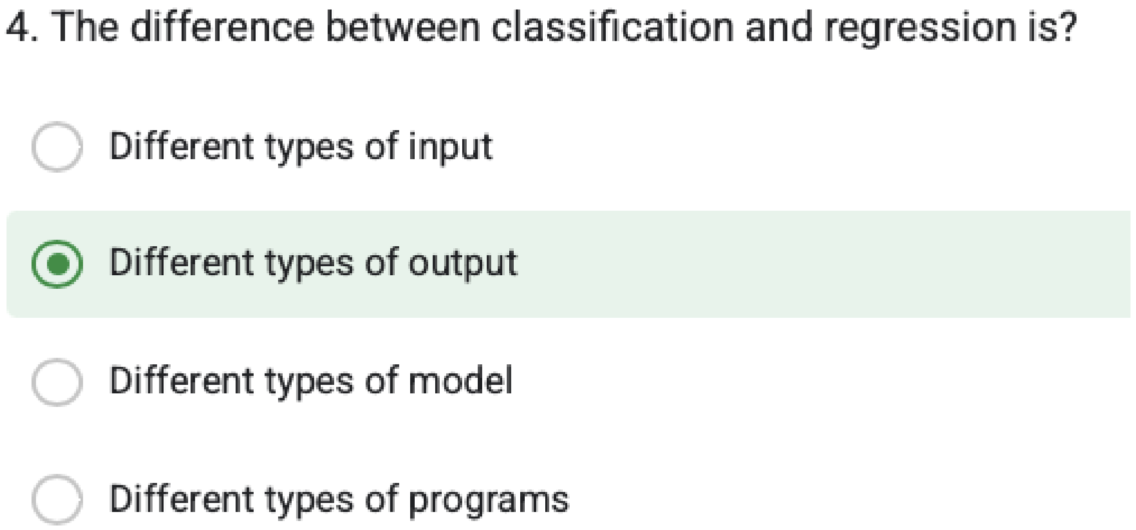
○ Label y is discrete

Quiz 2



3. Generalization refers to how well your model performs on

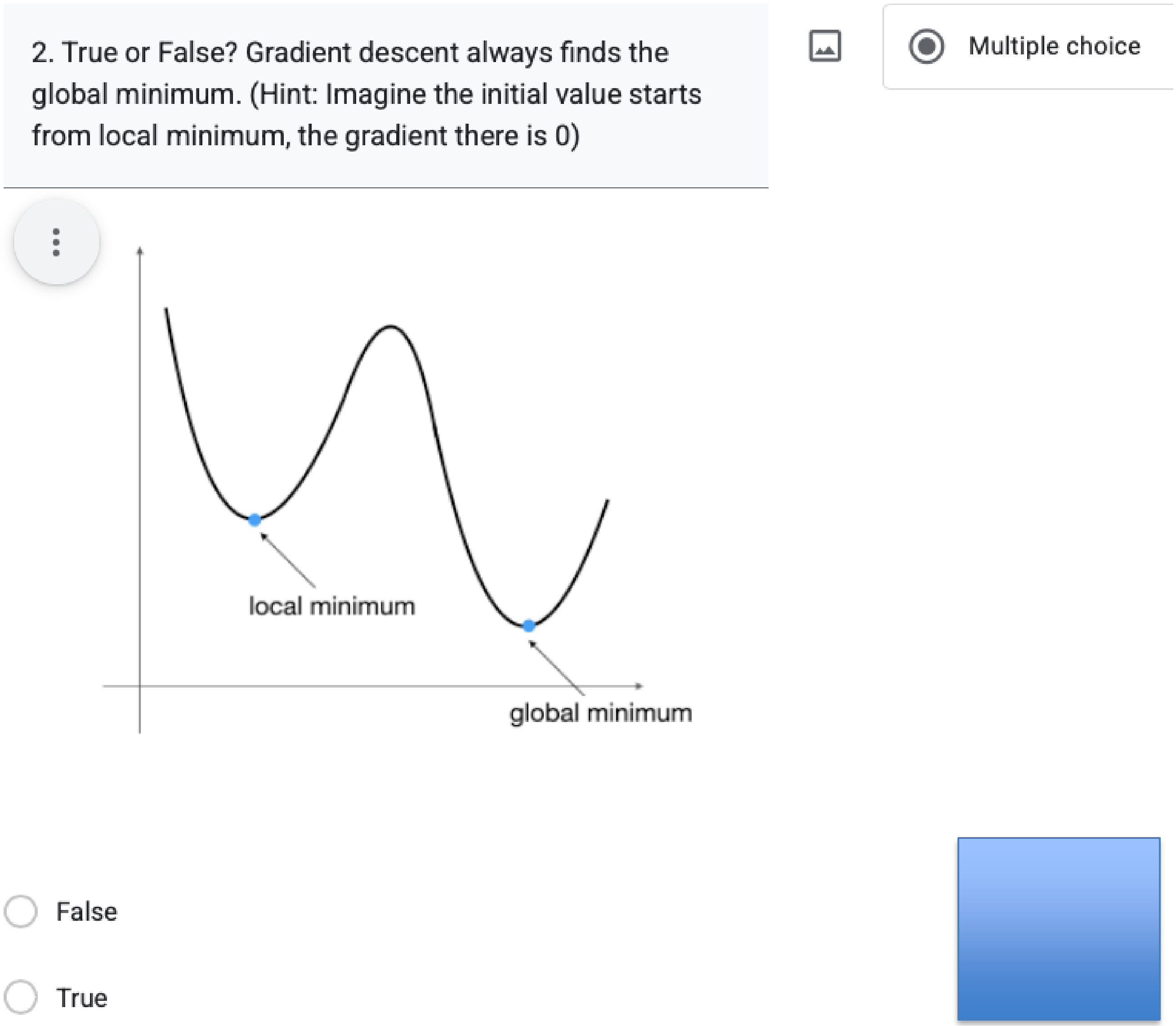
- ● The testing set
- ○ The training set
- ○ Both the training and testing set



4. The difference between classification and regression is?

- ○ Different types of input
- ● Different types of output
- ○ Different types of model
- ○ Different types of programs

Quiz 3 Plus

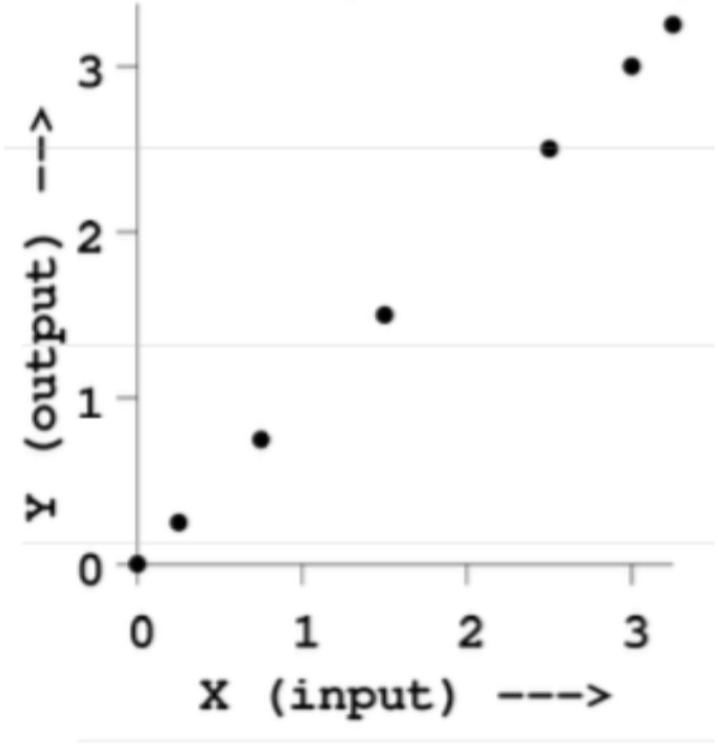


2. True or False? Gradient descent always finds the global minimum. (Hint: Imagine the initial value starts from local minimum, the gradient there is 0)

○ False

○ True

Question 3.1. Linear Regression+ Train-Test Split



Quiz 3 Plus

Figure 1: A reference dataset for regression with one real-valued input (x as horizontal axis) and one real-valued output (y as vertical axis).

What is the mean squared training error when running linear regression to fit the data ? (i.e., the model is $y = \beta_0 + \beta_1 x$). Assuming the rightmost three points are in the test set, and the others are in the training set. (you can eyeball the answers.)

L 3,4- Reading Questions

1. Fundamentals of Linear Regression

- What does it mean for a dataset to be a good fit for linear regression?
- Does linear supervised regression only work with data that is already somewhat linear?
- When is it a good time to use linear regression, and under what conditions will it perform best?
- Where is linear regression used in real-world applications today?
- What challenges exist in making linear regression models robust and trustworthy?
- How should we interpret regression coefficients when features are correlated? Does GD handle multicollinearity?
- What does the "bias" term represent conceptually?

L 3,4- Reading Questions

- **2. Loss / Cost Functions**
- What exactly is the meaning of Mean Squared Error (MSE), Mean Absolute Error (MAE), Sum of Squared Errors (SSE)?
- When is MSE preferred over MAE, and what are the trade-offs?
- Why is SSE chosen in linear regression instead of MAE?
- What is the purpose of the ½ factor in quadratic loss?
- How do loss functions differ for convex, concave, and saddle point graphs?
- Where did the SSE loss measurement originate from?
- What is the difference between objective, cost, and loss function terminology?
- How do we choose performance metrics (MSE, MAE, R^2, others) and when should multiple metrics be combined?

L 3,4- Reading Questions

- **3. Gradient Descent & Optimization**
- How does gradient descent (GD) work conceptually?
- What's the difference between GD, stochastic GD (SGD), and mini-batch GD?
- How do we choose learning rate (α) values? Are they fixed or dynamic?
- How do we pick good starting points for GD?
- How does GD behave near local minima, saddle points, or flat regions?
- Are there ways for SGD to escape local minima/saddle points?
- What are good batch sizes, and how do they affect convergence?
- What are the limitations of GD and strategies to overcome them?
- How do we evaluate convergence and know when to stop?
- Could you show a full worked-out example of optimizing with GD step by step?

L 3,4- Reading Questions

- **4. Normal Equation vs Iterative Methods**
- When should we use the Normal Equation versus Gradient Descent or SGD?
- What are the computational trade-offs between closed-form (Normal Eq.) and iterative (GD/SGD) methods?
- What happens if the feature matrix X does not have full rank?
- Why is Strassen’s algorithm for matrix multiplication not always the default, despite being faster in theory?

- **5. Model Selection & Trade-offs**
- How do we know when to choose linear regression vs. more complex models (e.g., Random Forest, SVC)?
- How do we evaluate trade-offs between generalization, efficiency, scalability, and interpretability?
- How does context influence model selection and visualization choices?
- How are these classical regression/optimization topics applied to modern LLMs like ChatGPT or Alexa?

L 3,4- Reading Questions

- **6. Training, Testing, and Generalization**
- How do we decide the split between training and testing data? (e.g., 80/20 rule)
- Is there an optimal ratio of training to test size?
- What does it mean for a model to generalize well? Does it just mean low error?
- Why is performance on training data not a good indicator of generalization?
- What happens if train/test sets have slightly different distributions (distribution shift)?
- When should validation sets be introduced in addition to train/test splits?
- How does generalization relate to overfitting/underfitting (residual patterns, feature poisoning examples)?

- **7. Matrix & Representation Issues**
- Why represent regression in matrix form? How does it help computation and parallelization?
- What's the difference between summation form and matrix form of the loss function?
- How do row vs. column vectors work in NumPy?
- What does it mean for a matrix to be full rank, and why does it matter?

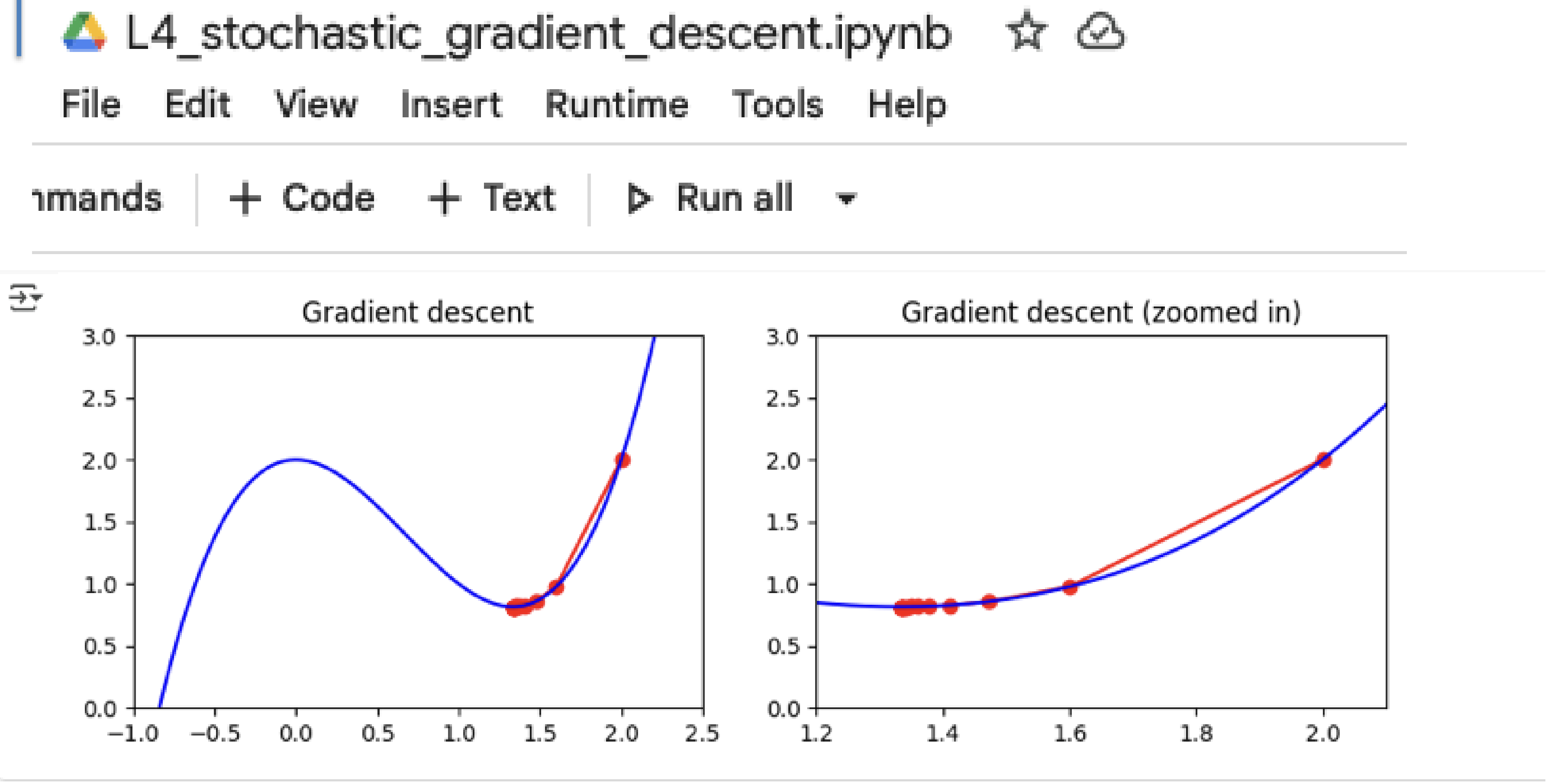
Matrix Representation (p53-)

Lecture 3: Linear Regression Basics

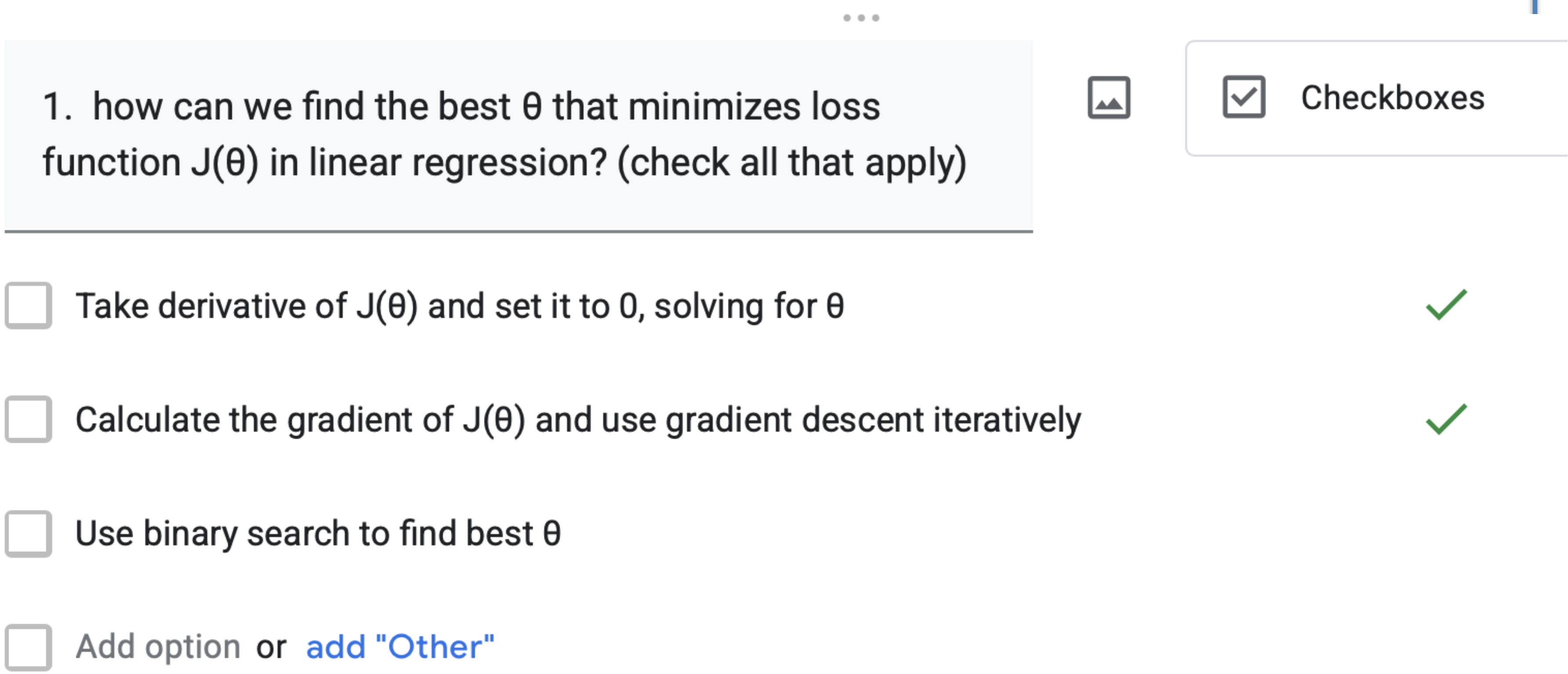
Many architecture details and Algorithm details to consider

- (1): Data parallelization through CPU SIMD / Multithreading/ GPU parallelization / ….
- (2): Memory hierarchical / locality
- (3): Better algorithms, like Strassen’s Matrix Multiply and many others

Learning Rate Code Run



Quiz 3



1. how can we find the best θ that minimizes loss function J(θ) in linear regression? (check all that apply)

☑ Checkboxes

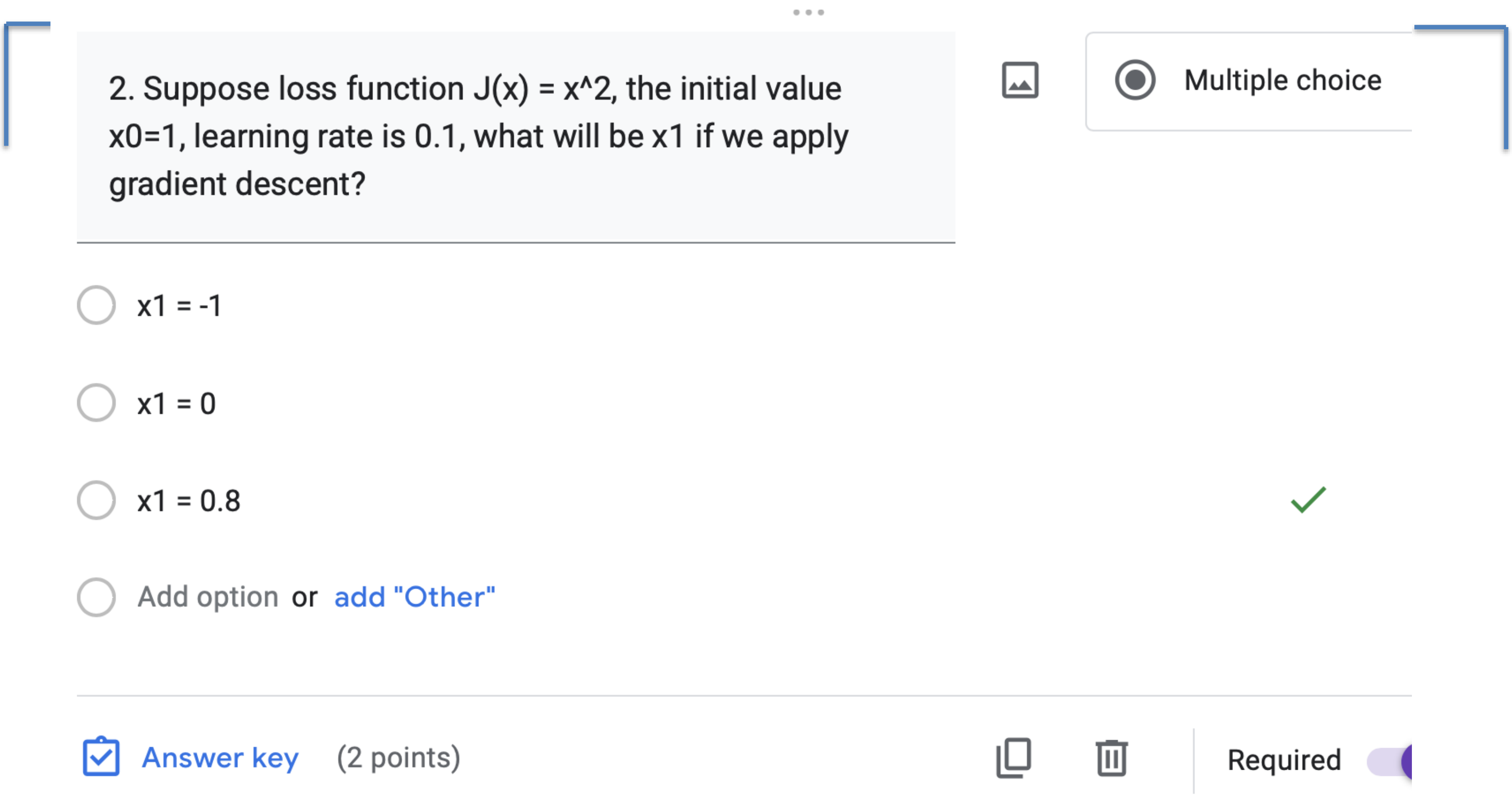
□ Take derivative of J(θ) and set it to 0, solving for θ ✓

□ Calculate the gradient of J(θ) and use gradient descent iteratively ✓

□ Use binary search to find best θ

□ Add option or add "Other"

Quiz 3



● Multiple choice

2. Suppose loss function J(x) = x^2, the initial value x0=1, learning rate is 0.1, what will be x1 if we apply gradient descent?

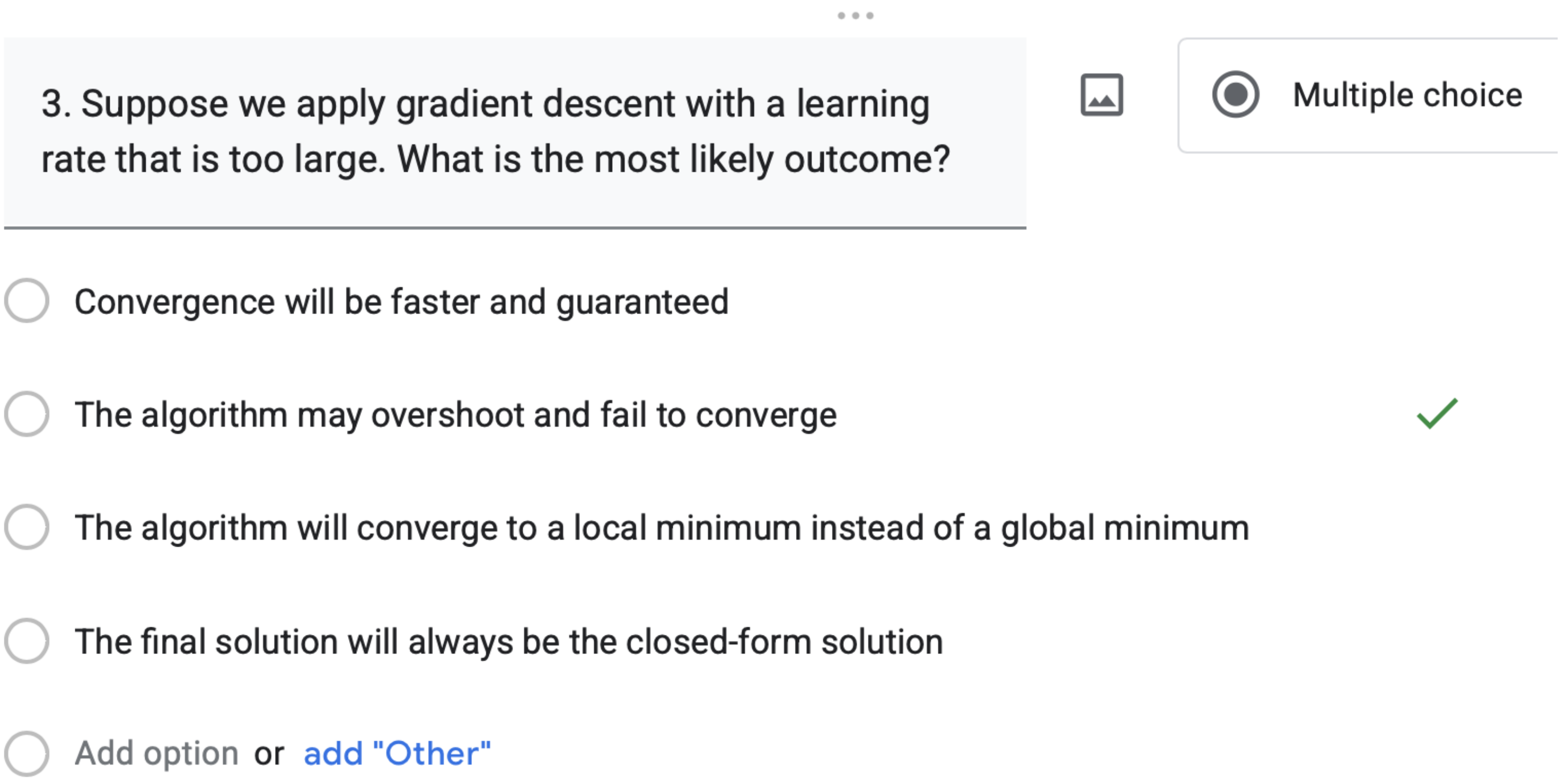
○ x1 = -1

○ x1 = 0

○ x1 = 0.8 ✓

○ Add option or add "Other"

Quiz 3



3. Suppose we apply gradient descent with a learning rate that is too large. What is the most likely outcome?

● Multiple choice

○ Convergence will be faster and guaranteed

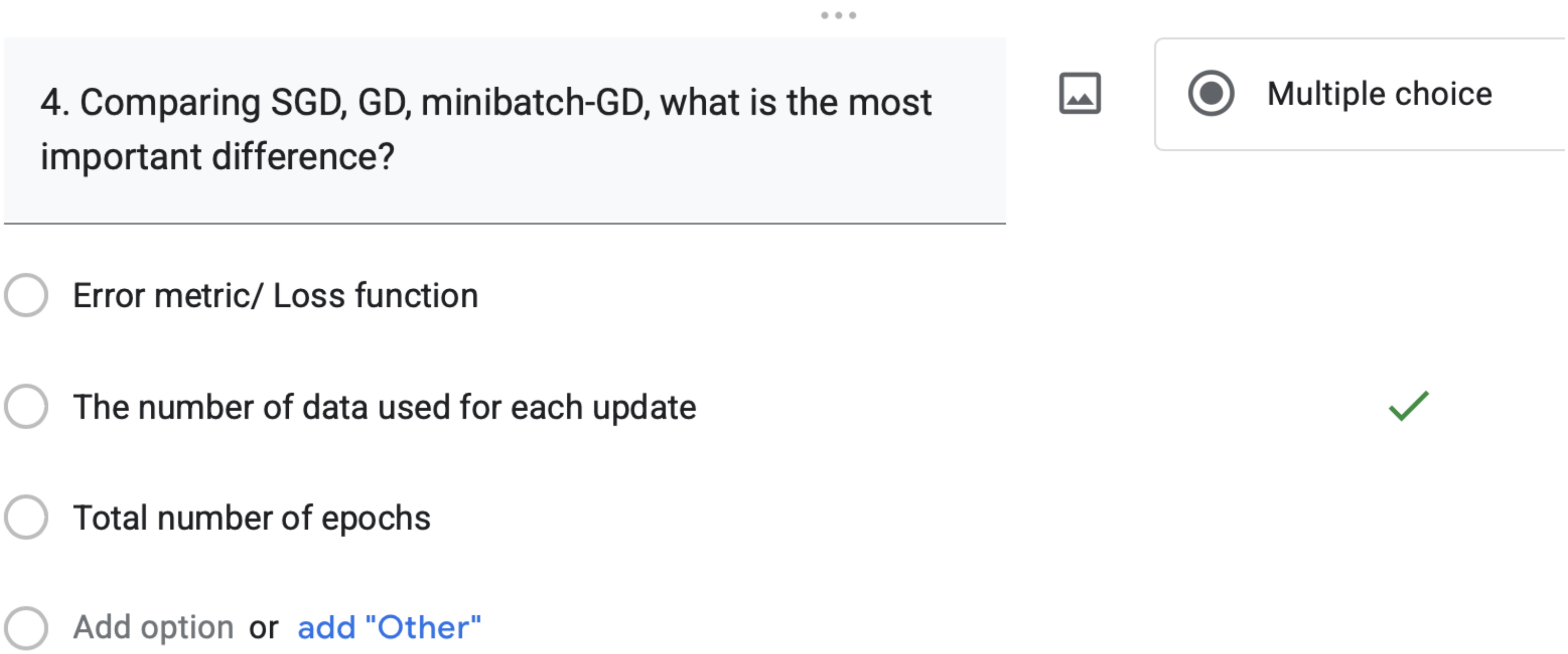
○ The algorithm may overshoot and fail to converge ✓

○ The algorithm will converge to a local minimum instead of a global minimum

○ The final solution will always be the closed-form solution

○ Add option or add "Other"

Quiz 3



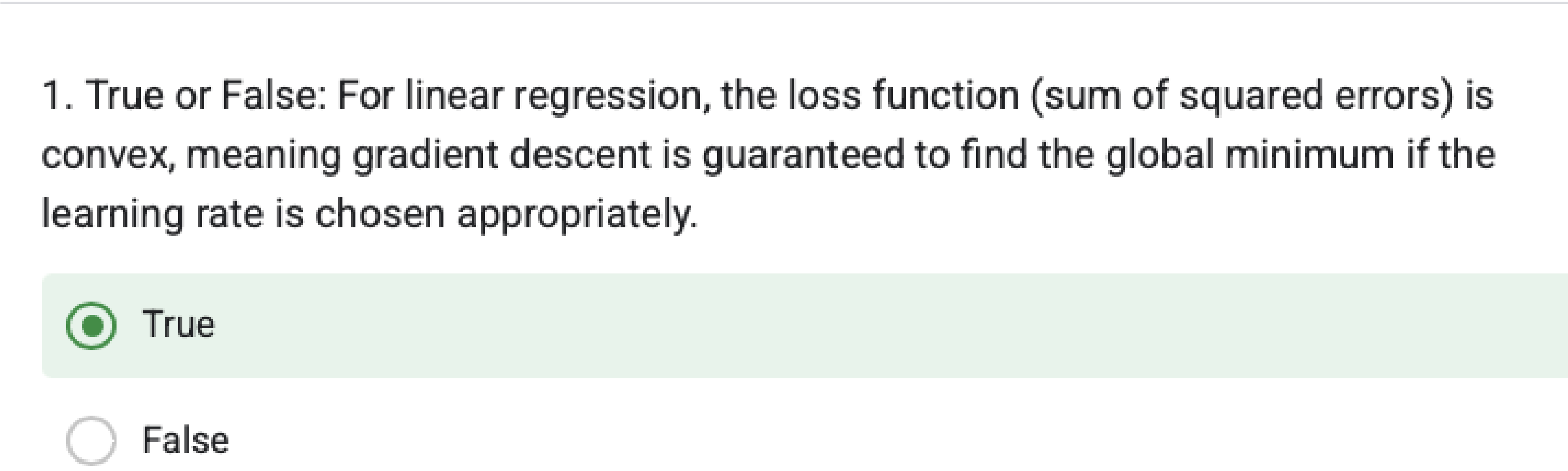
4. Comparing SGD, GD, minibatch-GD, what is the most important difference?

○ Error metric/ Loss function

○ The number of data used for each update ✓

○ Total number of epochs

Quiz 3 Plus

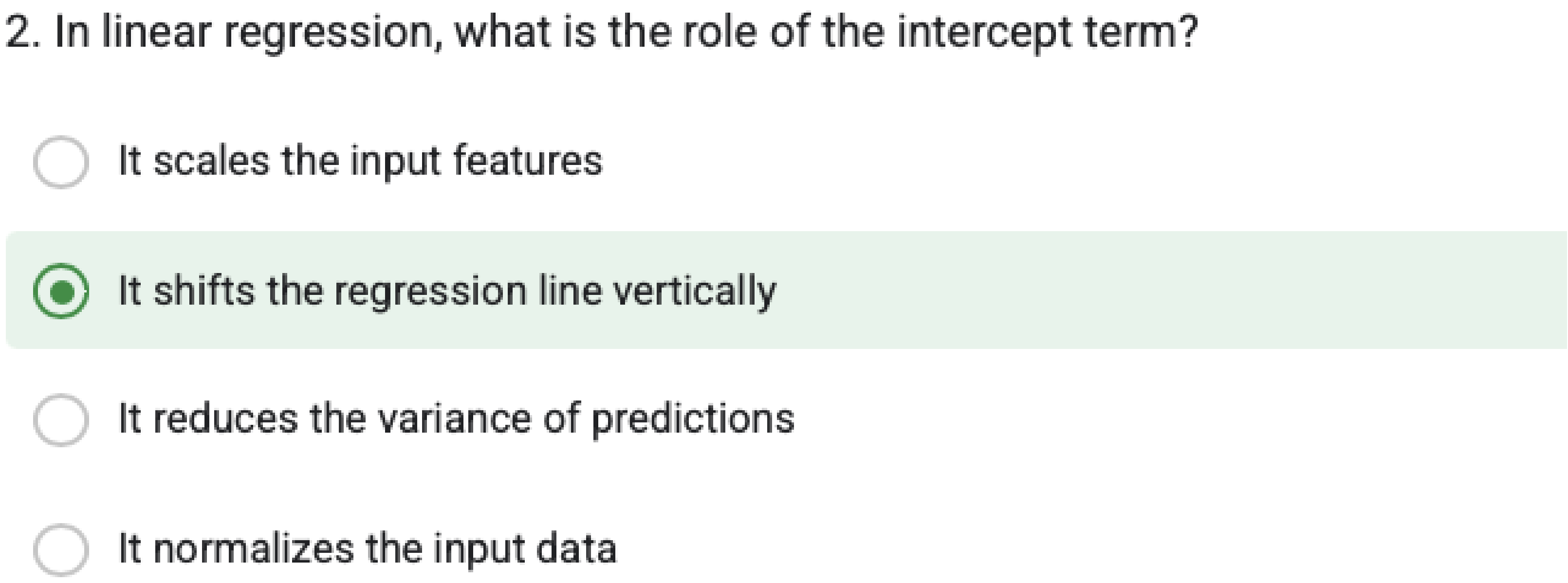


1. True or False: For linear regression, the loss function (sum of squared errors) is convex, meaning gradient descent is guaranteed to find the global minimum if the learning rate is chosen appropriately.

● True

○ False

Quiz 3 Plus



2. In linear regression, what is the role of the intercept term?

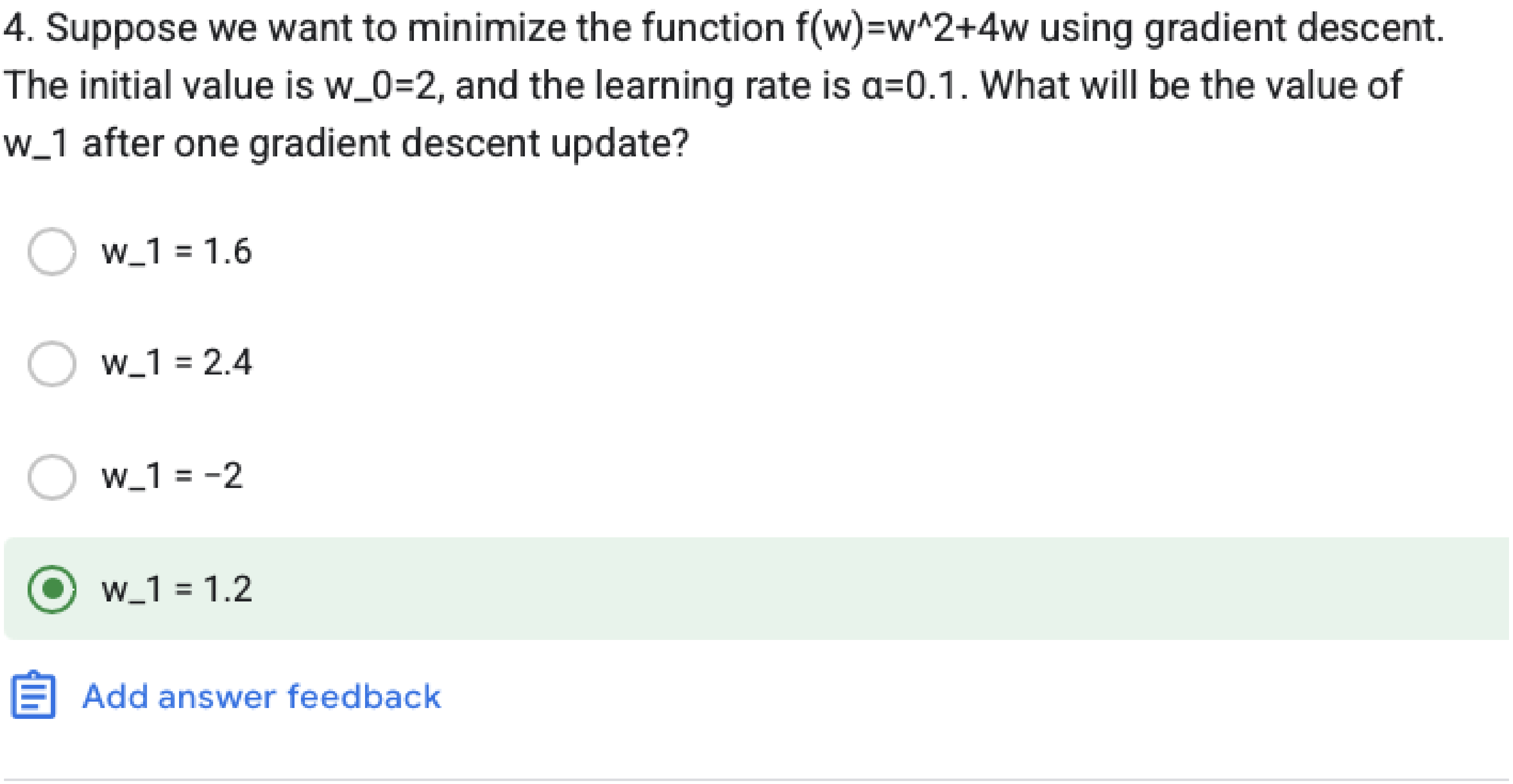
○ It scales the input features

● It shifts the regression line vertically

○ It reduces the variance of predictions

○ It normalizes the input data

Quiz 3 Plus



4. Suppose we want to minimize the function f(w)=w^2+4w using gradient descent. The initial value is w_0=2, and the learning rate is α=0.1. What will be the value of w_1 after one gradient descent update?

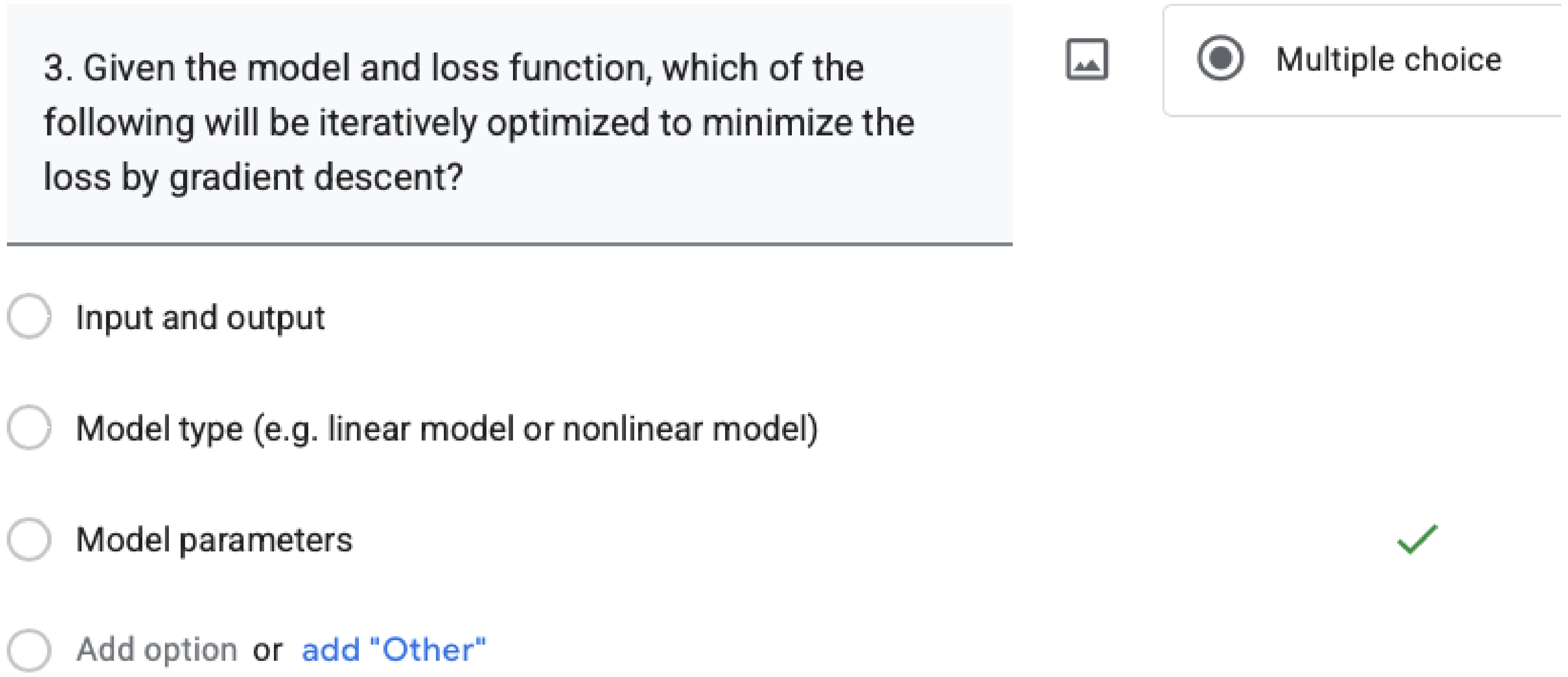
○ w_1 = 1.6

○ w_1 = 2.4

○ w_1 = −2

● w_1 = 1.2

Quiz 3 Plus



3. Given the model and loss function, which of the following will be iteratively optimized to minimize the loss by gradient descent?

◉ Multiple choice

○ Input and output

○ Model type (e.g. linear model or nonlinear model)

○ Model parameters ✓

○ Add option or add "Other"

L5 – e.g. LR with radial-basis functions

- E.g.: LR with RBF regression:

$$\underset{\boxed{?}}{\hat{y}} = \theta_0 + \square_{j=1}^{m} \theta_j \varphi_j(x) = \varphi(x)^T \theta$$

$$\underset{\boxed{?}}{\varphi}(x) := \left[1, K_{\lambda_1}(x, r_1), K_{\lambda 2}(x, r_2), K_{\lambda_3}(x, r_3), K_{\lambda 4}(x, r_4) \right]^T$$

$$\theta^* = \left(\varphi^T \varphi\right)^{-1} \varphi^T \vec{y}$$

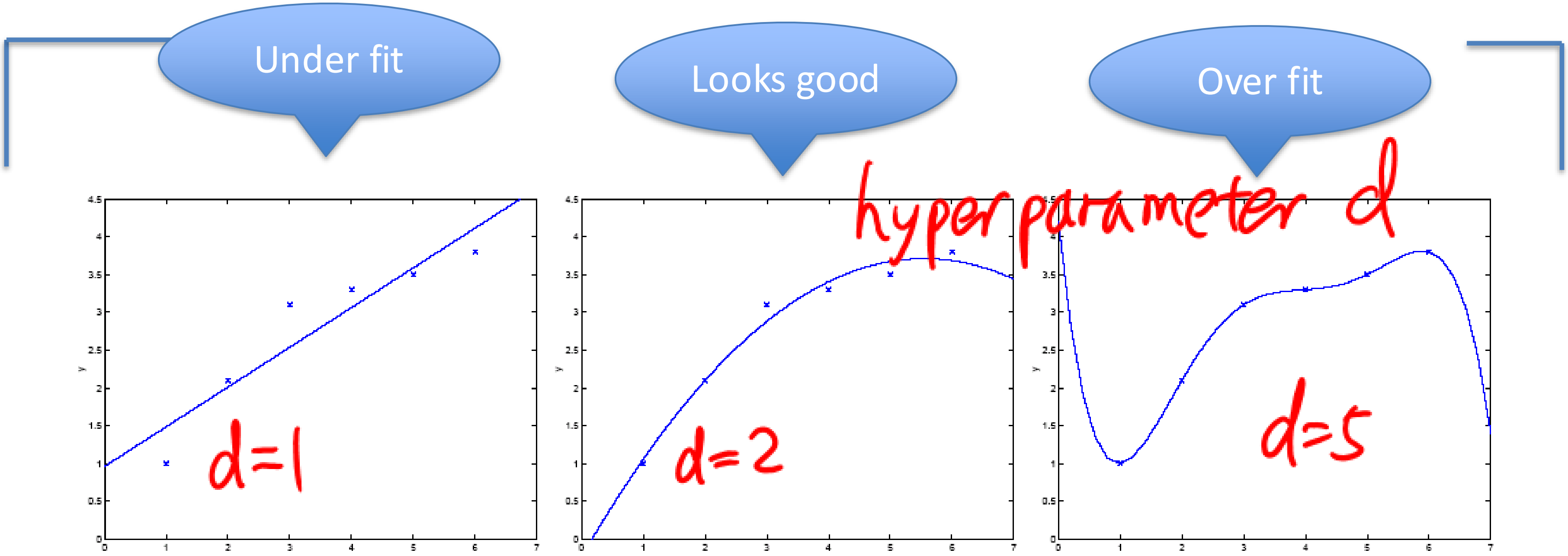
$$\vec{\Theta} = [\Theta_0, \Theta_1, \Theta_2, \Theta_3, \Theta_4]^T$$

$$\begin{array}{cccc} r_1 & r_2 & \lambda_3 & r_4 \\ \lambda_1 & \lambda_2 & r_3 & \lambda 4 \end{array} \Bigg\} \text{hyper para}$$

L6: Main issues: Model Selection

- How to select the right model type? How to select hyperparameter for a model type?
  - E.g. what polynomial degree d for polynomial regression
  - E.g., where to put the centers for the RBF kernels? How wide?
  - E.g. which basis type? Polynomial or RBF?

What Model Order to Select?



$y = \theta_0 + \theta_1 x$

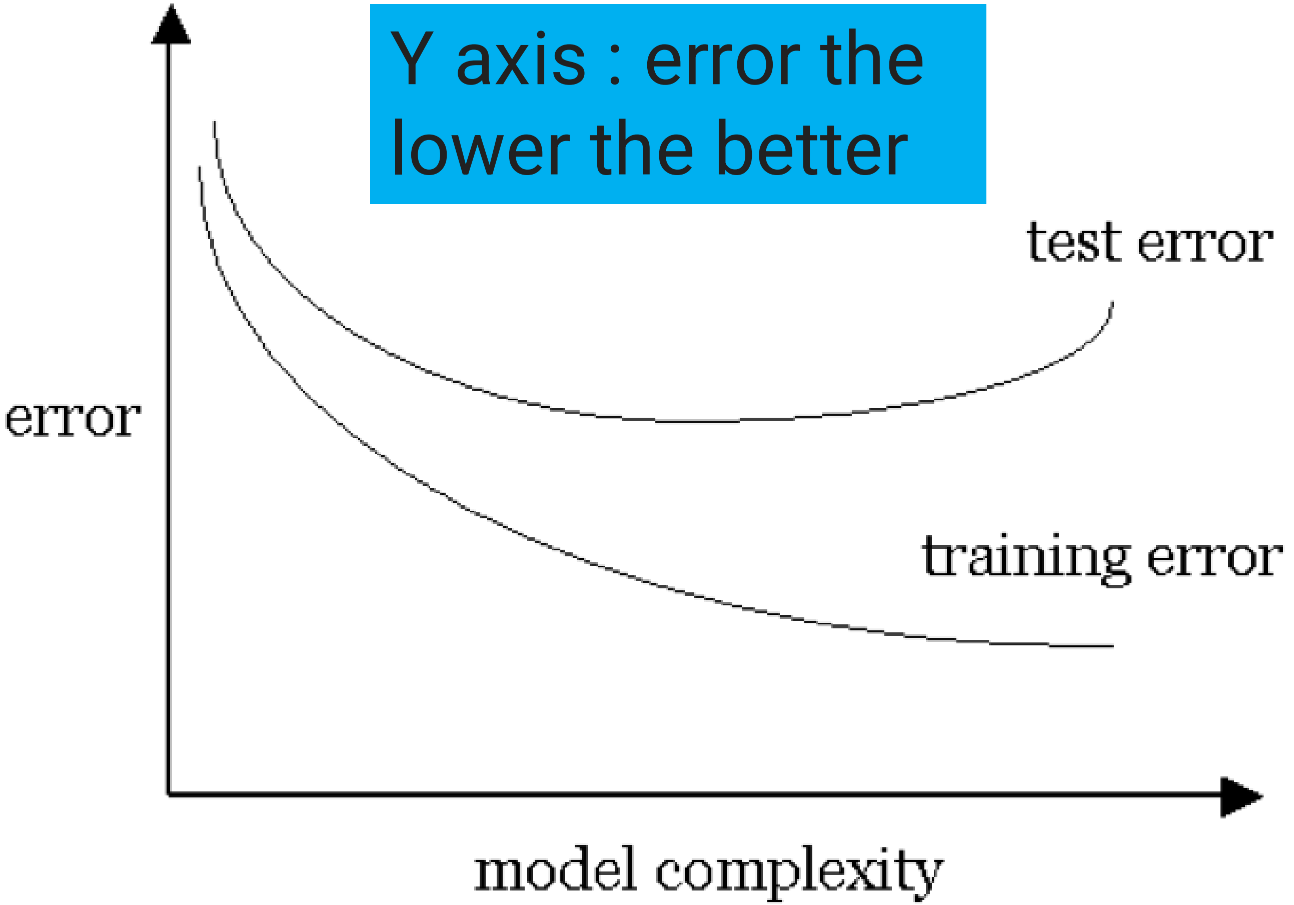
$y = \theta_0 + \theta_1 x + \theta_2 x^2$

$y = \sum_{j=0}^{5} \theta_j x^j$

Generalisation: learn function / hypothesis from past data in order to "explain", "predict", "model" or "control" new data examples

(a) Train-validation /
(b) K-fold Cross
Validation /

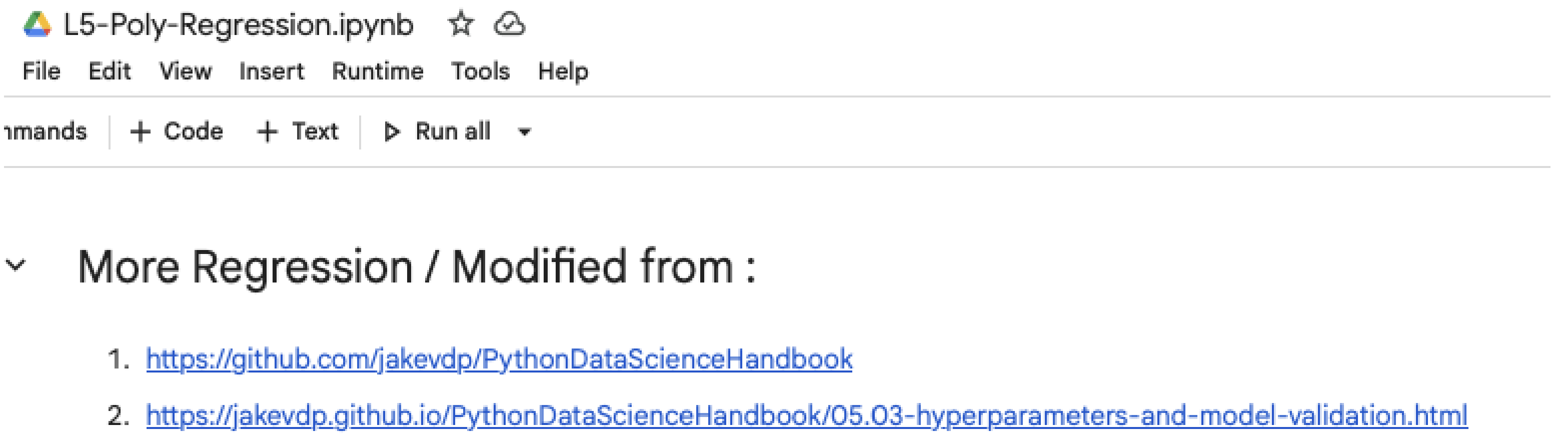
A Plot for Model Selection



Y axis : error the lower the better

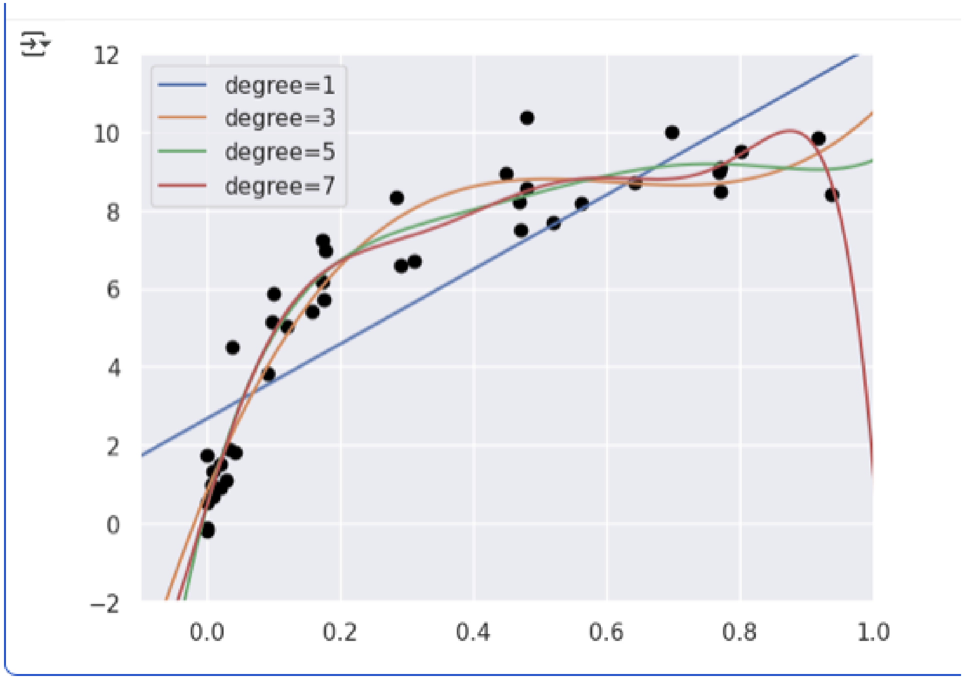
k-CV on train to choose model and hyperparameter / then a separate test set to assess future performance

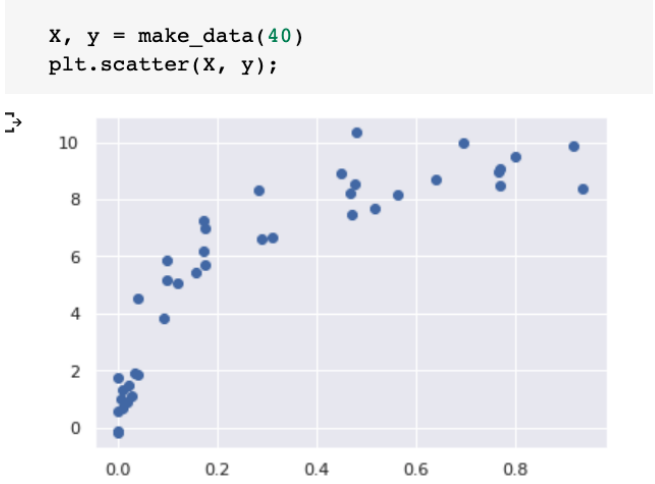
Polynomial Regression Code Run



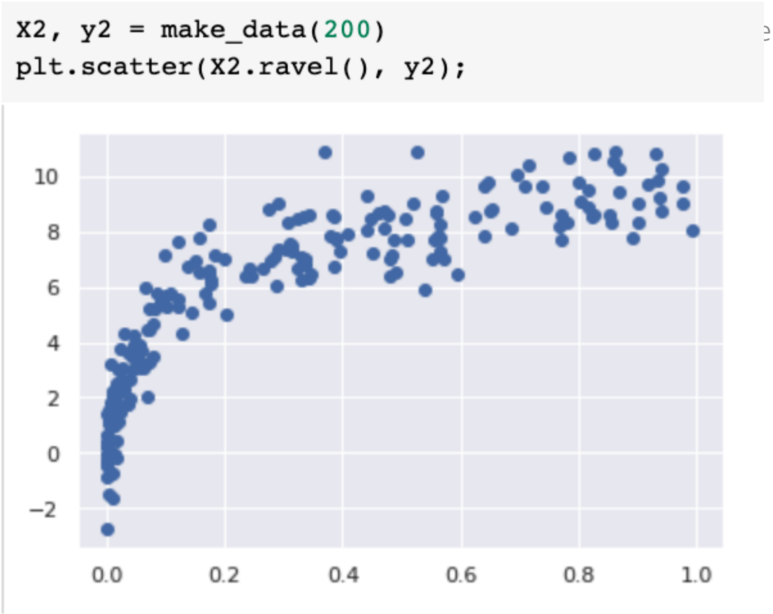
More Regression / Modified from :

1. https://github.com/jakevdp/PythonDataScienceHandbook
2. https://jakevdp.github.io/PythonDataScienceHandbook/05.03-hyperparameters-and-model-validation.html

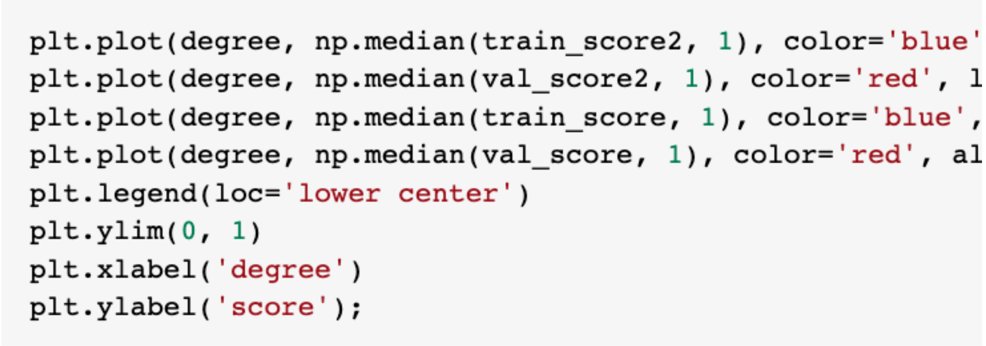




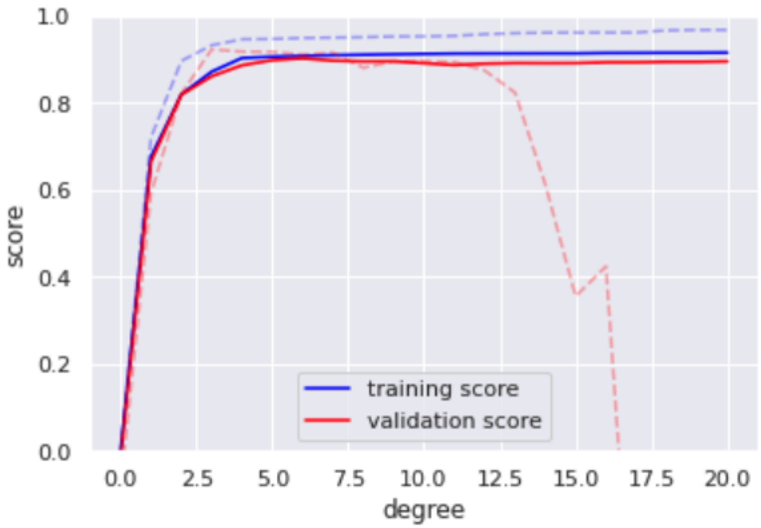
```
X, y = make_data(40)
plt.scatter(X, y);
```



```
X2, y2 = make_data(200)
plt.scatter(X2.ravel(), y2);
```

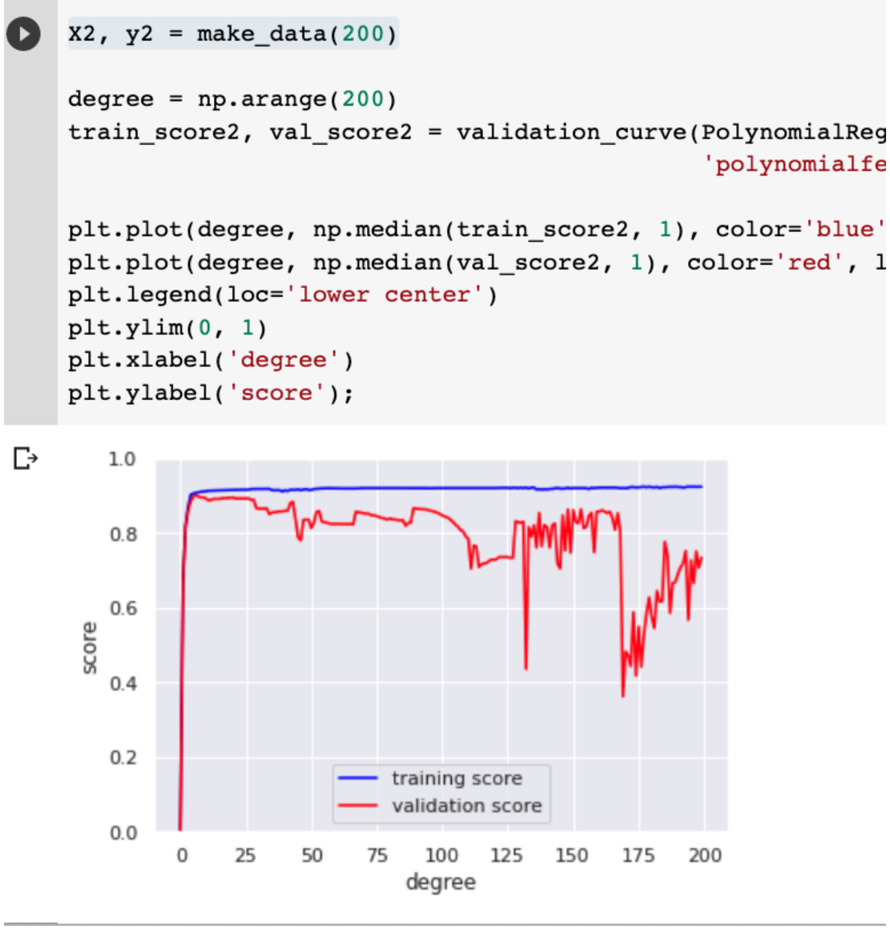


```
plt.plot(degree, np.median(train_score2, 1), color='blue'
plt.plot(degree, np.median(val_score2, 1), color='red', l
plt.plot(degree, np.median(train_score, 1), color='blue',
plt.plot(degree, np.median(val_score, 1), color='red', al
plt.legend(loc='lower center')
plt.ylim(0, 1)
plt.xlabel('degree')
plt.ylabel('score');
```



Behavior of the validation curve:

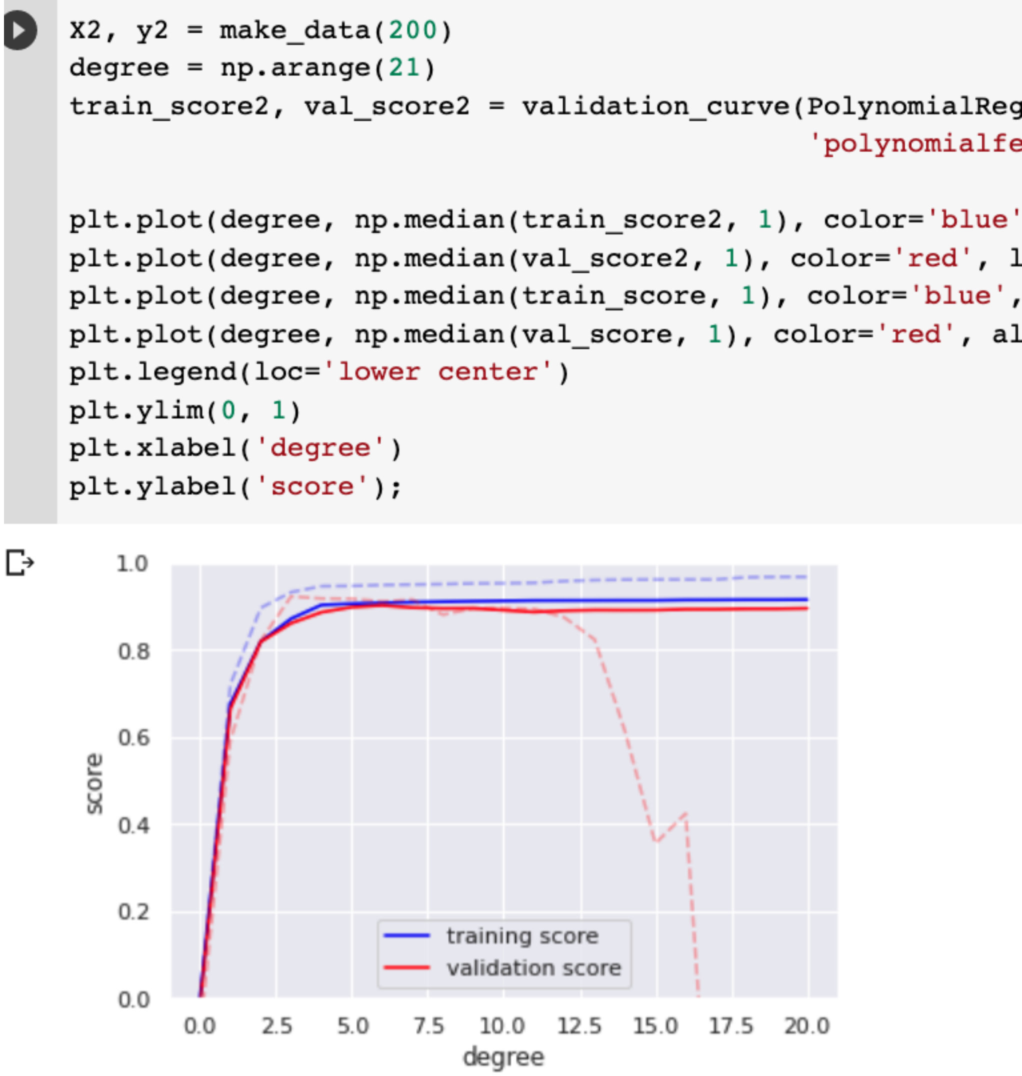
- the model complexity
- the number of training points



```
X2, y2 = make_data(200)

degree = np.arange(200)
train_score2, val_score2 = validation_curve(PolynomialReg
                                             'polynomialfe

plt.plot(degree, np.median(train_score2, 1), color='blue'
plt.plot(degree, np.median(val_score2, 1), color='red', l
plt.legend(loc='lower center')
plt.ylim(0, 1)
plt.xlabel('degree')
plt.ylabel('score');
```



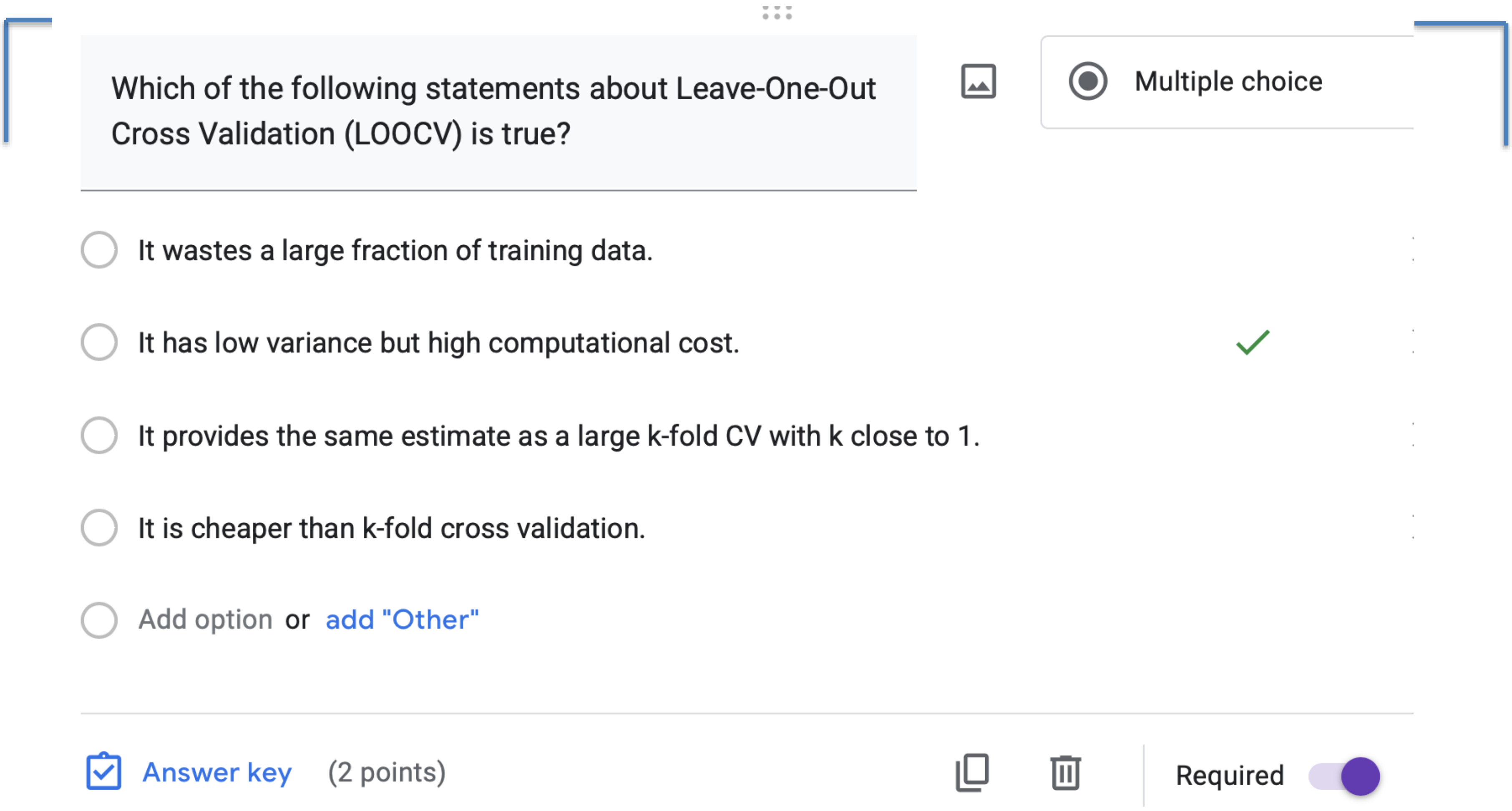
```
X2, y2 = make_data(200)
degree = np.arange(21)
train_score2, val_score2 = validation_curve(PolynomialReg
                                             'polynomialfe

plt.plot(degree, np.median(train_score2, 1), color='blue'
plt.plot(degree, np.median(val_score2, 1), color='red', l
plt.plot(degree, np.median(train_score, 1), color='blue',
plt.plot(degree, np.median(val_score, 1), color='red', al
plt.legend(loc='lower center')
plt.ylim(0, 1)
plt.xlabel('degree')
plt.ylabel('score');
```

Interesting Relation between

- the right range of model complexity
- the number of training points

Quiz 4 plus



Which of the following statements about Leave-One-Out Cross Validation (LOOCV) is true?

● Multiple choice

○ It wastes a large fraction of training data.

○ It has low variance but high computational cost. ✓

○ It provides the same estimate as a large k-fold CV with k close to 1.

○ It is cheaper than k-fold cross validation.

○ Add option or add "Other"

Answer key (2 points)

Required

need to make assumptions that are able to generalize

- **Underfitting:** model is too “simple” to represent all the relevant characteristics
  - High bias and low variance
  - High training error and high test error

- **Overfitting:** model is too “complex” and fits irrelevant characteristics (noise) in the data
  - Low bias and high variance
  - Low training error and high test error

A Gentle Touch of Bias - Variance Tradeoff

(More details … Later)

Slide credit: L. Lazebnik

L 5/ L6 Readings