https://ttpoll.com/p/817711

COSC 325: Introduction to Machine Learning

Dr. Hector Santos-Villalobos

THE UNIVERSITY OF TENNESSEE
KNOXVILLE

Lecture 09: Data Splits and Overfitting

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

Class Announcements

Homework:

Homework #3 is out and due 09/29.

Course Project:

Check groups in Canvas.

PRFAQ is due this Friday.

Check Additional Approved Datasets in the Course Project Assignment pane.

Lectures:

Absences: In your email's subject, include the following text "[COSC325 ABSENCE]"

Exams:

Exam #1: Thursday, 10/03

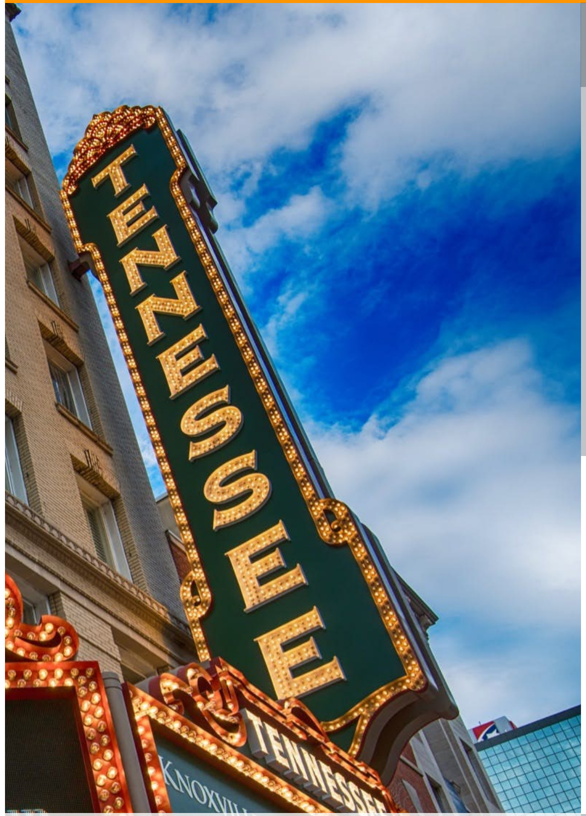
Exam #2: Thursday, 11/21

Online exam window 11 am to 1 pm

://ttpoll.com/p/817711

Tennessee RobUst, Secure, and Trustworthy AI Seminar (TRUST-AI)

TRUST-AI TENNESSEE



Invited Speaker

Dr. Murat Kantarcioglu
Professor
CCI Faculty Fellow
Virginia Tech

Talk Title:

Defending and Defeating AI: Protecting the Good, Attacking the Bad for Privacy, Security and Fairness

THE UNIVERSITY OF TENNESSEE KNOXVILLE

This seminar series is a part of the AI TENNessee Distinguished Seminar Series, sponsored by the AI Tennessee Initiative.

Time: Friday, Sep. 20
12:30 PM - 1:30 PM

Location: MKB 622

MORE INFO

What: UTK Machine Learning Club

Where: **MK 525**

When: **Tuesday** at **5:00**
(including today)

Who: Any experience level

Everyone is welcome to the first meeting of ML club today. Whether you are a beginner looking to learn from our intro to ML lesson series, experienced practitioner who wants to learn from and discuss with other enthusiasts in our reading groups, or you just want to hear from our industry guest speakers and seminars, utkML can help you scratch your machine learning itch!

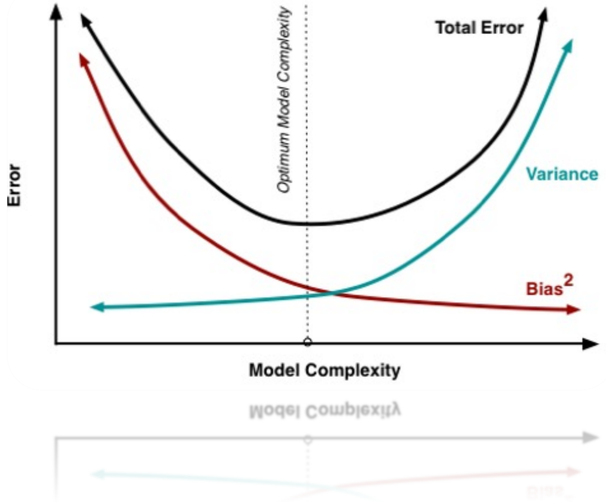
Last Lecture

- Why is linear regression not a good choice for classification problems?
- Logistic Regression for Classification
  - Decision boundary geometry
  - Computational graph
  - Derivatives for GD algorithm
- Binary Cross Entropy Loss
  - Convex for binary problems
  - Derivatives for GD algorithms

https://ttpoll.com/p/817711

Today’s Topics

Overfitting, Variance and Bias



THE UNIVERSITY OF TENNESSEE KNOXVILLE

Pop Quiz

How much time did it take to finish homework #2?

A. Less than 2 hours.

B. 2 to 3 hours.

C. 3 to 5 hours.

D. More than 5 hours.

Pop Quiz

1 | MULTIPLE CHOICE

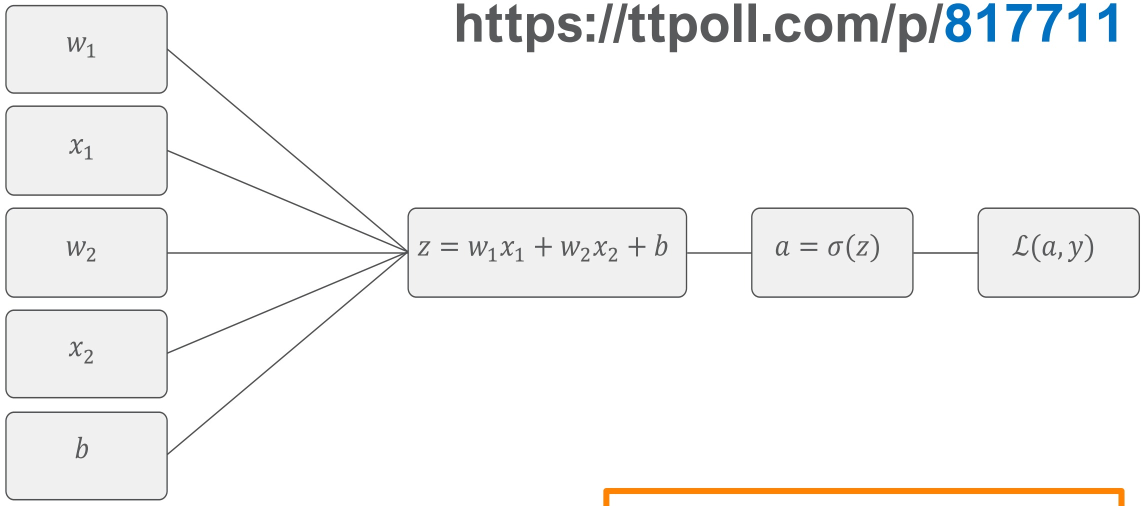
How familiar are you with Overfitting, Variance, and BIas?

A. Unfamiliar topics

B. I have heard about these topics

C. I know what these are about.

Scaling to n samples.



- Computing the cost $\boldsymbol{J(w,b)}$

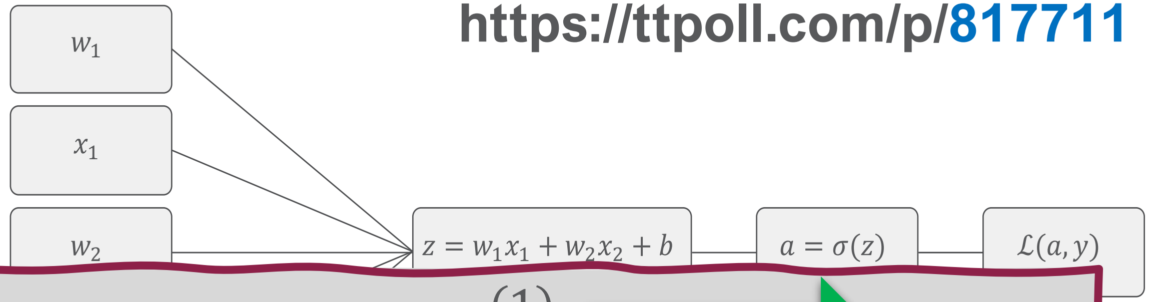
$$J(w,b) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(a^{(i)}, y^{(i)}) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\sigma(x^{(i)}w \quad + b), y^{(i)})$$

Recall:

$$\frac{d\mathcal{L}}{dw_1} = (a - y)x_1$$

- To our benefit, $J(w, b)$ is the average of the measured losses

$$\frac{\partial J(w,b)}{\boldsymbol{\partial w_1}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\mathcal{L}(a^{(i)}, y^{(i)})}{\partial w_1} = \frac{1}{n}\sum_{i=1}^{n}(a^{(i)} - y^{(i)})x_1^{(i)} = \frac{1}{n}\sum_{i=1}^{n}(\sigma(x^{(i)}w \quad + b) - y^{(i)})x_1^{(i)}$$



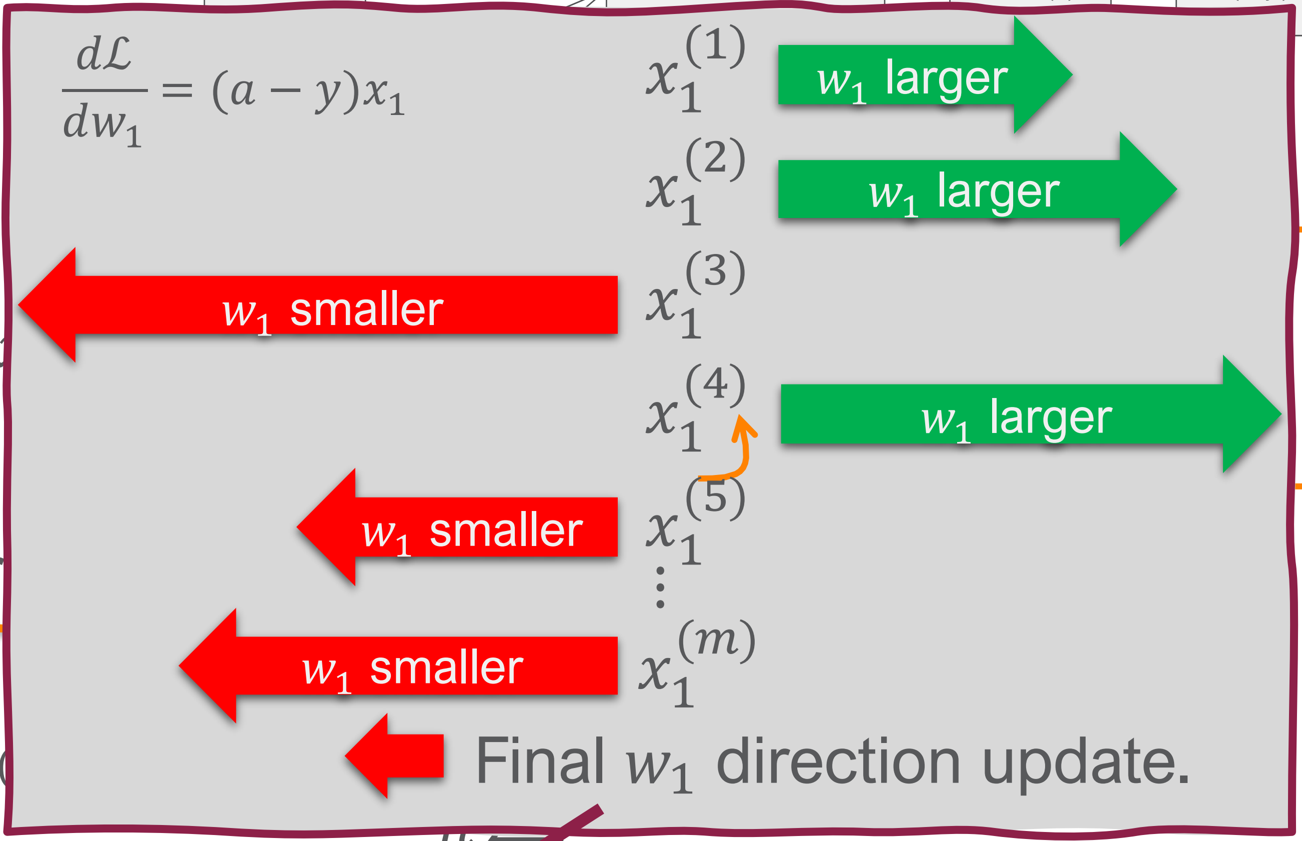
Scaling to n samples.

- Computing the cost $\boldsymbol{J(w,b)}$

$$J(w,b) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(a^{(i)}, y^{(i)}) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\sigma($$

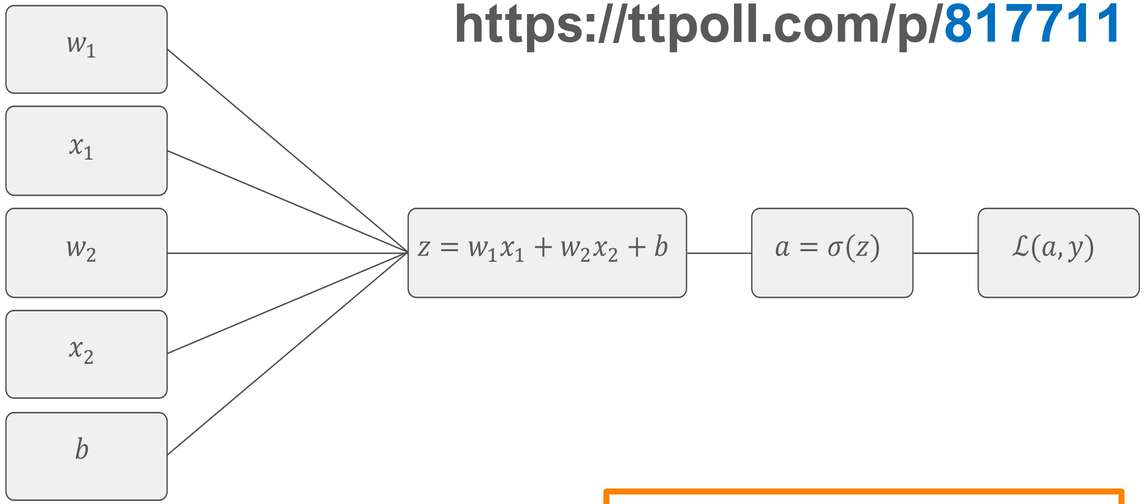
- To our benefit, $J(w,b)$ is the aver

$$\frac{\partial J(w,b)}{\boldsymbol{\partial w_1}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\mathcal{L}(a^{(i)}, y^{(i)})}{\partial w_1} = \frac{1}{n}\sum_{i=1}^{n}(a$$



https://ttpoll.com/p/817711

Scaling to n samples.



- Computing the cost $\boldsymbol{J(w,b)}$

$$J(w,b) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(a^{(i)}, y^{(i)}) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\sigma(x^{(i)}w + b), y^{(i)})$$

Recall:

$$\frac{d\mathcal{L}}{db} = (a - y)$$

- To our benefit, $J(w,b)$ is the average of the measured losses

$$\frac{\partial J(w,b)}{\boldsymbol{\partial b}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\mathcal{L}(a^{(i)}, y^{(i)})}{\partial b} = \frac{1}{n}\sum_{i=1}^{n}(a^{(i)} - y^{(i)}) = \frac{1}{n}\sum_{i=1}^{n}(\sigma(x^{(i)}w + b) - y^{(i)})$$

https://ttpoll.com/p/817711

Gradient Descent Algorithm

Repeat from k=1 to num_iterations:

$J_k = 0, dw_1 = 0, dw_2 = 0, \text{and } db = 0$

Initialization of aggregating variables

Repeat from $i = 1{:}n$ *(Batch GD)*

$z = w_1 x_1^{(i)} + w_2 x_2^{(i)} + b$

$a = \sigma(z)$

$J_k := J - \left[y^{(i)} \log(a) + \left(1 - y^{(i)}\right) \log(1 - a)\right]$

$dz = \left(a^{(i)} - y^{(i)}\right)$

$$\left.\begin{aligned} dw_1 &:= dw_1 + (dz) x_1^{(i)} \\ dw_2 &:= dw_2 + (dz) x_2^{(i)} \\ db &:= db + (dz) \end{aligned}\right\}$$

Cumulative parameter derivative for each sample.

End of Sample Loop

$J_k := J_k/n, dw_1 := dw_1/n, dw_2 := dw_2/n, \text{and } db := db/n$

Expected Values

$w_1 := w_1 - \alpha dw_1, w_2 := w_2 - \alpha dw_2, b := b - \alpha db$

Parameter Update

End of GD Loop

Vectorized Gradient Descent Algorithm

$J = 0, \qquad X := [1, X]$

Adding columns of ones to account for intercept parameter.

Repeat from k=1 to num_iterations:

$Z = XW$

(n,m+1)(m+1,1)=(n,1)

$\mathrm{A} = \sigma(Z)$

$J_k = \left(\frac{-1}{n}\right)\left[Y^T \log(A) + (1 - Y^T)\log(1 - A)\right]$

(1,n)(n,1)+(1,n)(n,1) = scalar

$dZ = (A - Y)$

$\mathrm{d}W = \left(\frac{1}{n}\right) X^T dZ$

(m+1,n)(n,1)=(m+1,1)

$W := W - \alpha dW$

Parameter Update

End of GD Loop

Notebook Time

Pop Quiz

2 | MULTIPLE CHOICE

Vectorize the following computation of C. Assume A and B are matrices of shape (n,1).

```
C=0
for k=1 in range(len(A))
    C = C + A[k]*B[k]
```

A. C = A*B

B. C=A@B

C. C=np.dot(A,B)

D. C=np.dot(A.T,B)

Pop Quiz

2 | MULTIPLE CHOICE

Vectorize the following computation of C. Assume A and B are matrices of shape (n,1).

```
C=0
for k=1 in range(len(A))
    C = C + A[k]*B[k]
```

A. C = A*B

B. C=A@B

C. C=np.dot(A,B)

D. C=np.dot(A.T,B)

Model Evaluation

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

Why do we want to evaluate our model?

To ensure it is generalizing well.

Generalization Performance

- Want the model to "generalize" well to __________ data.

Generalization Performance

- We want the model to "generalize" well to ***unseen*** data.

- Either we want…
  - ***High*** generalization ***accuracy***
  - ***Low*** generalization ***error***

Assumptions

- i.i.d. assumption: inputs samples are independent, and training and test examples are identically distributed and drawn from the same probability distribution _________.
- For some random model that has not been fit to the training set, we expect both the training and test error to be _____________.
- For some model fit to the training set, we expect the training error be _________ than the test error.
- The training error or accuracy provides an ________________ __________ estimate of the generalization performance.

Assumptions

- i.i.d. assumption: inputs samples are independent, and training and test examples are identically distributed and drawn from the same probability distribution $\underline{f(x, y)}$.
- For some random model that has not been fit to the training set, we expect both the training and test error to be ***similar***.
- For some model fit to the training set, we expect the training error be ***lower*** than the test error.
- The training error or accuracy provides an ***optimistically biased*** estimate of the generalization performance.

Training, Validation, and Test Sets

- **Training set:** samples drawn from $f(x, y)$ used to train/adjust the *parameters* in model $h(x)$.

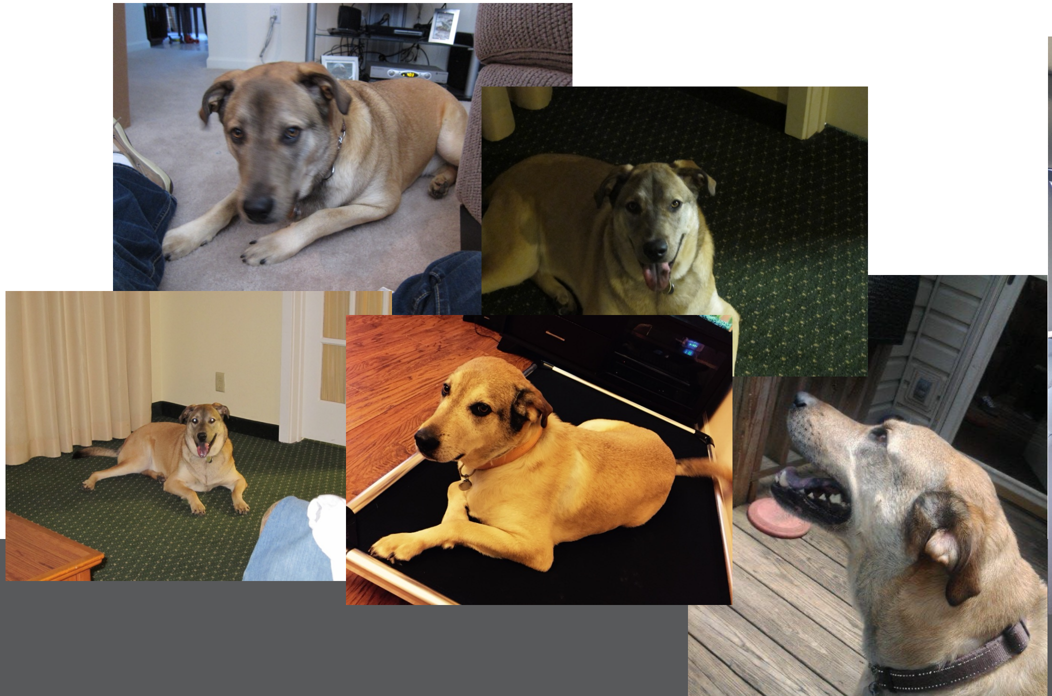
- **Validation set:** samples drawn from $f(x, y)$ used to evaluate model performance and adjust the *hyperparameters* in model $h(x)$.

- **Test set:** samples drawn from $f(x, y)$ used to evaluate the final model with unseen data.

Practical Advice on Data Splits

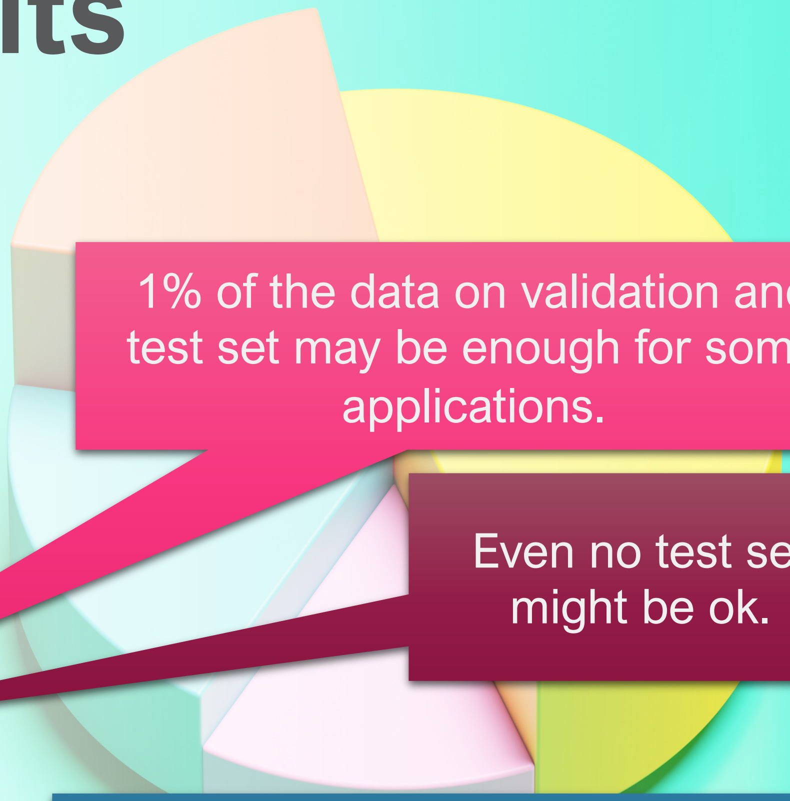
- Most times, random sampling works fine unless…
  - Unbalanced classes – Stratified split
  - Differences in the data (e.g., quality)

Defects, anomalies, disease.



Practical Advice on Data Splits

- Most times, random sampling works fine unless…
  - Unbalanced classes – Stratified split
  - Differences in the data (e.g., quality)
- Typical splits {Training, Validation, Testing}
  - {60, 20, 20}, {70,15,15}, {80,10,10}
  - Validation and testing set splits are about adequate data representation



1% of the data on validation and test set may be enough for some applications.

Even no test set might be ok.

Test and validation sets from the same sample distribution.

Practical Advice on Data Splits

- Most times, random sampling works fine unless…
  - Unbalanced classes – Stratified split
  - Differences in the data (e.g., quality)
- Typical splits {Training, Validation, Testing}
  - {60, 20, 20}, {70,15,15}, {80,10,10}
  - Validation and testing set splits are about adequate data representation
- Avoid data leakage

THE UNIVERSITY OF TENNESSEE KNOXVILLE

Random Sampling

Data Leakage

Dataset

Dog 3
Dog 1b
Dog 1
Dog 2

Cat 1
Cat 1b
Cat 2b
Cat 2

Horse 1
Horse 2
Horse 4
Horse 3

Training

Dog 3
Dog 1
Dog 2
Cat 1
Cat 1b
Cat 2
Horse 1
Horse 2
Horse 4

Validation

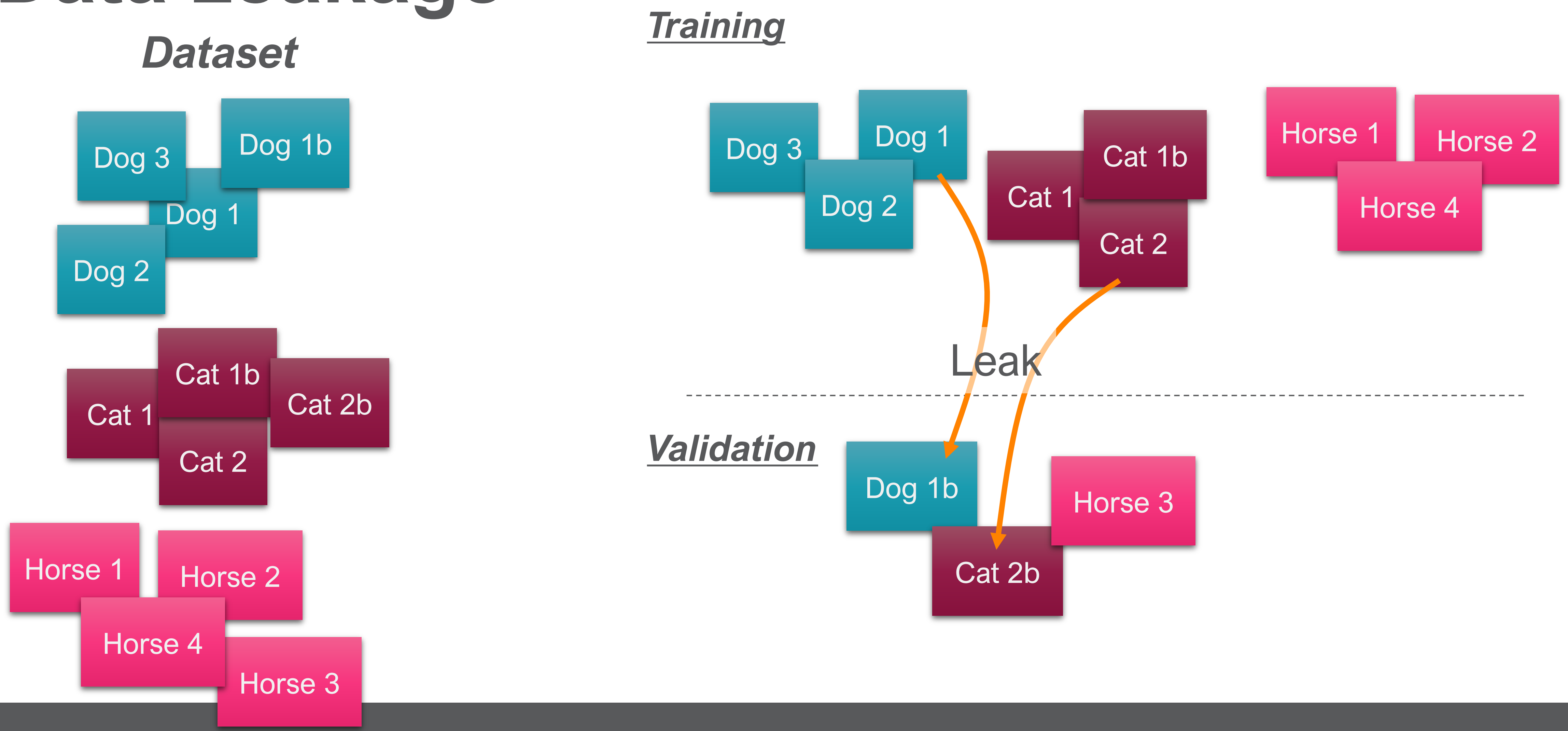
Dog 1b
Horse 3
Cat 2b

https://ttpoll.com/p/817711

Data Leakage

Dataset

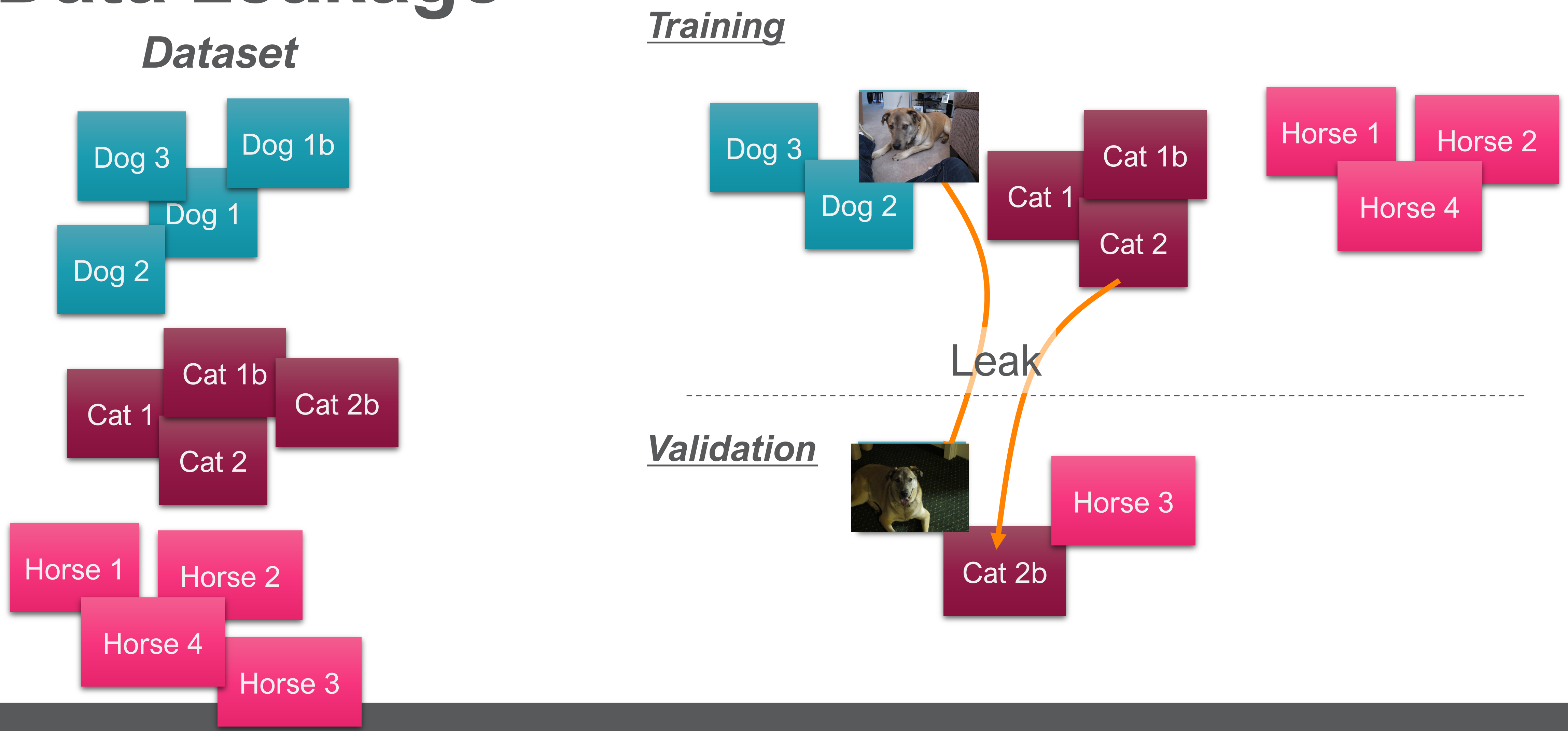
Random Sampling



Data Leakage

Random Sampling

Dataset



https://ttpoll.com/p/817711

Data Leakage

Dataset

Dog 3
Dog 1b
Dog 1
Dog 2

Cat 1b
Cat 1
Cat 2b
Cat 2

Horse 1
Horse 2
Horse 4
Horse 3

Correct Sampling

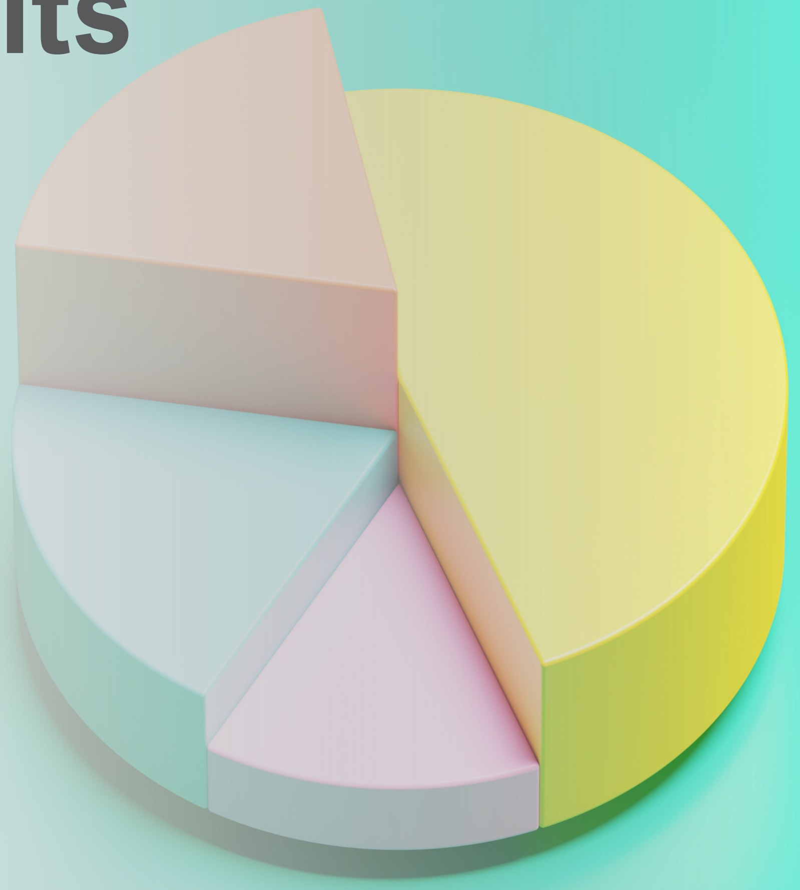
Training

Dog 1
Dog 1b
Dog 2
Cat 1b
Cat 1
Horse 1
Horse 2
Horse 4

Validation

Dog 3
Horse 3
Cat 2
Cat 2b

Practical Advice on Data Splits



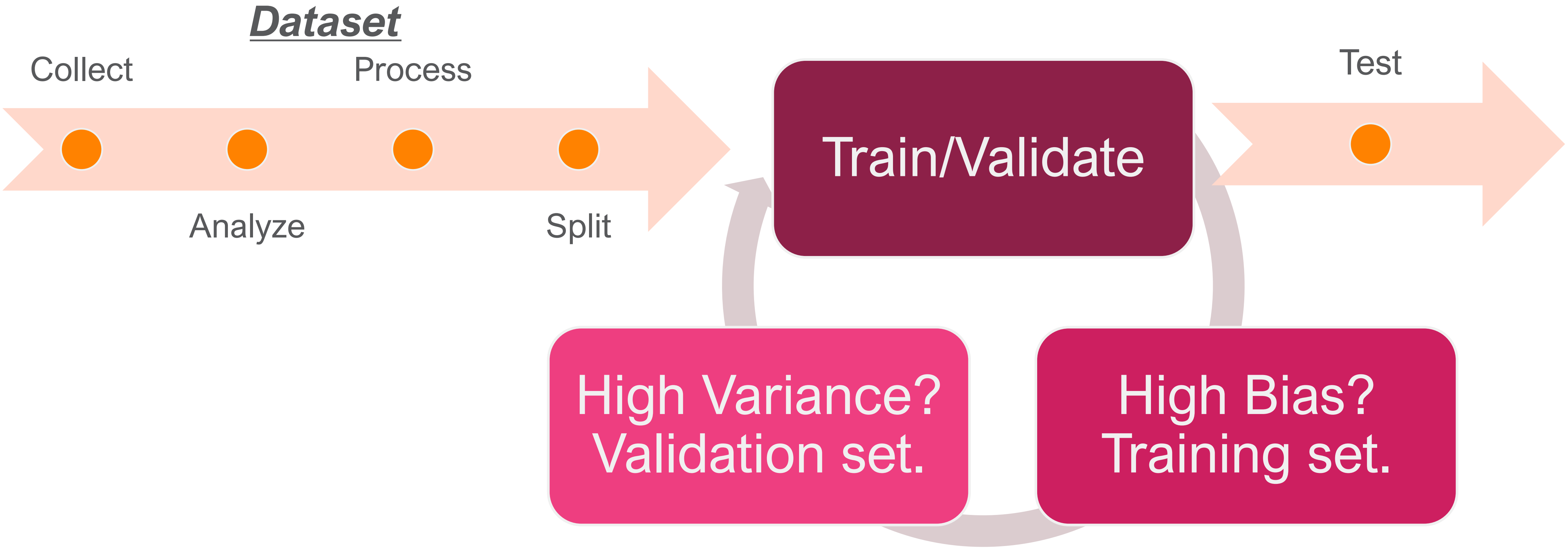
- Most times, random sampling works fine unless…
  - Unbalanced classes – Stratified split
  - Differences in the data (e.g., quality)
- Typical splits {Training, Validation, Testing}
  - {60, 20, 20}, {70,15,15}, {80,10,10}
  - Validation and testing set splits are about adequate data representation
- Avoid data leakage
  - E.g., time series data split chronologically
  - E.g., instances of the same sample assign to same set.

Practical Advice on Data Splits

- Most times, random sampling works fine unless…
  - Unbalanced classes – Stratified split
  - Differences in the data (e.g., quality)
- Typical splits {Training, Validation, Testing}
  - {60, 20, 20}, {70,15,15}, {80,10,10}
  - Validation and testing set splits are about adequate data representation
- Avoid data leakage
  - E.g., time series data split chronologically
  - E.g., instances of the same sample assign to same set.

THE UNIVERSITY OF TENNESSEE KNOXVILLE

So far


Dataset
Collect
Process
Analyze
Split
Train/Validate
Test
High Variance?
Validation set.
High Bias?
Training set.

Pop Quiz

3 | MULTIPLE CHOICE POINTS: 1 | Edit

We have a dataset of handwritten digits with 1 million samples independently and identically distributed drawn from f(x,y). What is a suitable sample allocation for training, validation, and test sets?

A. Trainig 80%, Validation 10%, Test 10%

B. Trainig 60%, Validation 20%, Test 20%

C. Trainig 70%, Validation 20%, Test 10%

D. Trainig 90%, Validation 5%, Test 5%

Pop Quiz

3 | MULTIPLE CHOICE

POINTS: 1 | Edit

We have a dataset of handwritten digits with 1 million samples independently and identically distributed drawn from f(x,y). What is a suitable sample allocation for training, validation, and test sets?

A. Trainig 80%, Validation 10%, Test 10%

B. Trainig 60%, Validation 20%, Test 20%

C. Trainig 70%, Validation 20%, Test 10%

D. Trainig 90%, Validation 5%, Test 5%

ImageNet-k: ~1.3M samples, 1,000 classes

- Training 85%
- Validation 3.8%, 50 samples per class
- Testing 7.7%, 100 samples, per class

Overfitting and Underfitting

Model Capacity

Capacity: the ability of a model to represent a wide variety of functions that map input data to output predictions. Also known as model complexity.

$$\mathcal{H} = \{h(X): X \rightarrow y\},$$

where $\mathcal{H}$ is the hypothesis space, which consists of all possible functions that the model $h(X)$ can learn on its architecture and parameters

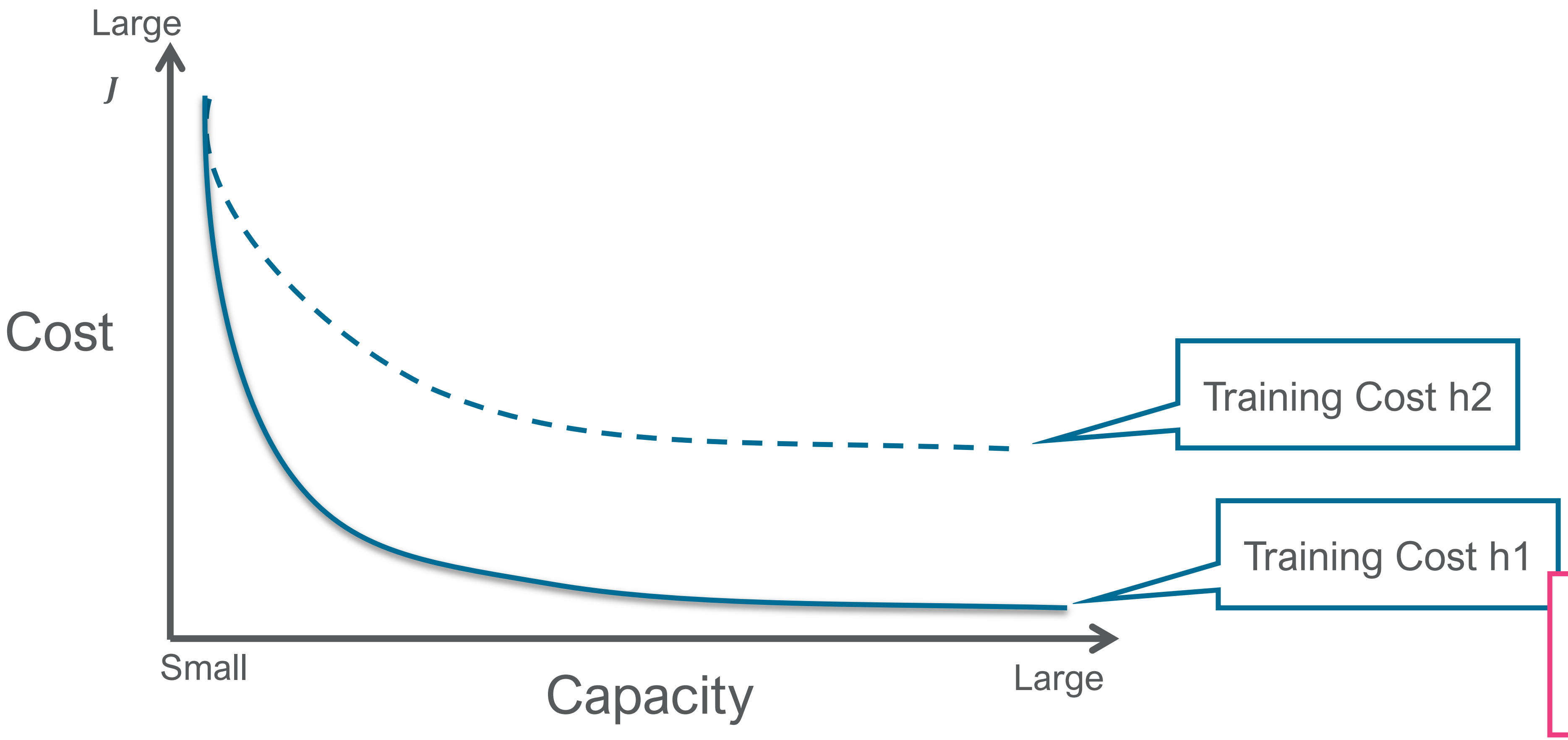
- Low capacity models (e.g., linear models) have smaller hypothesis space and can only represent simpler functions.
- High capacity models (e.g., deep neural networks) have a larger hypothesis space, enabling them to approximate more complex functions.

Overfitting and Underfitting

- ***Underfitting:*** both the training and validation errors are large.
  - Usually, the result of a low-capacity model
- ***Overfitting:*** gap between training and validation error
  - Validation error >> Training Error
- For a large hypothesis space being searched by a learning algorithm, there is a high tendency to ______fit

$$\mathcal{H} = \{h(X): X \rightarrow y\}, \text{ where } \mathcal{H} \text{ is very large}$$

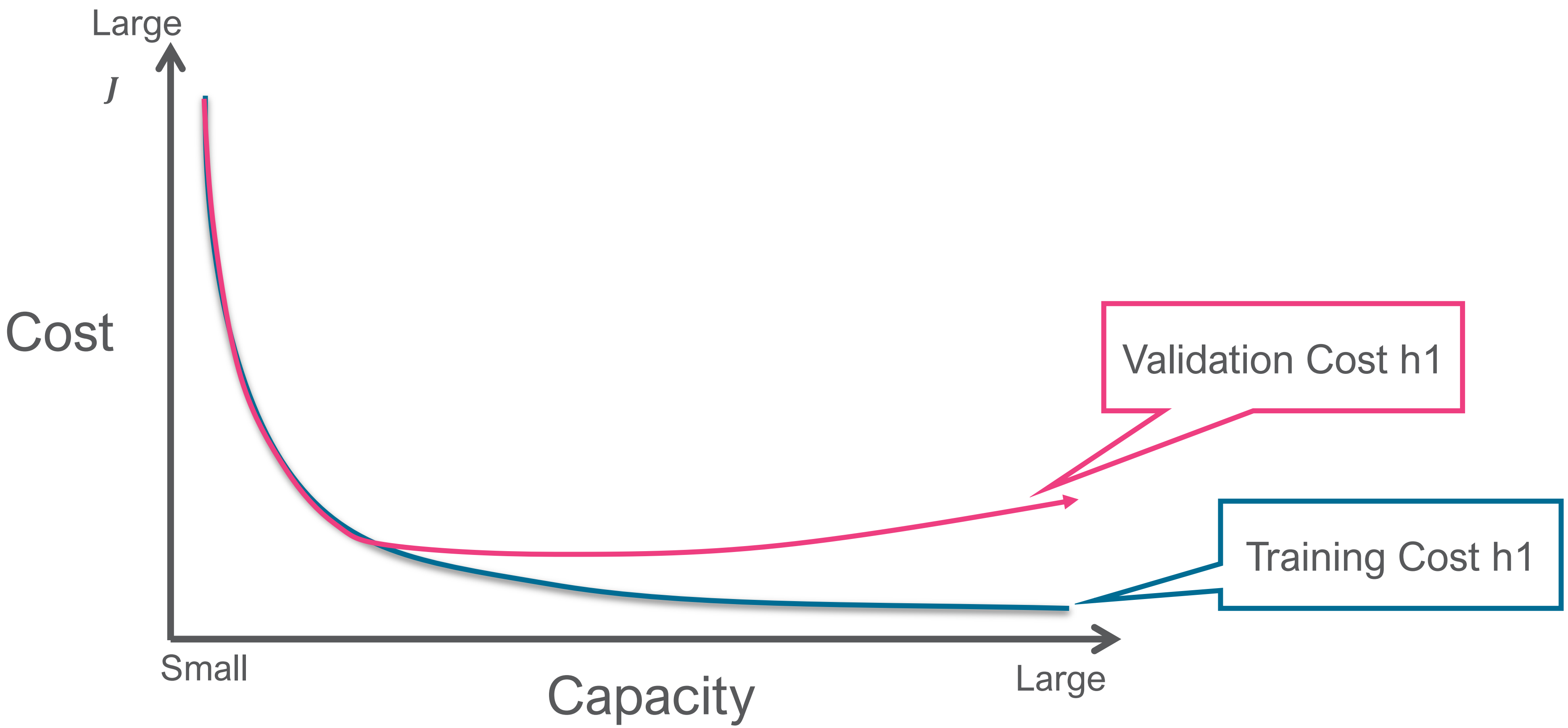
Overfitting and Underfitting



h_2 is _____fitting

Can we say anything about h_1?

Overfitting and Underfitting



h_1 is overfitting

Review

- Vectorized GD for logistic regression classification.
- Model evaluation
  - Dataset split
    - Training, validation, and testing
    - Random sampling while avoiding data leaks
  - Capacity
  - Overfitting