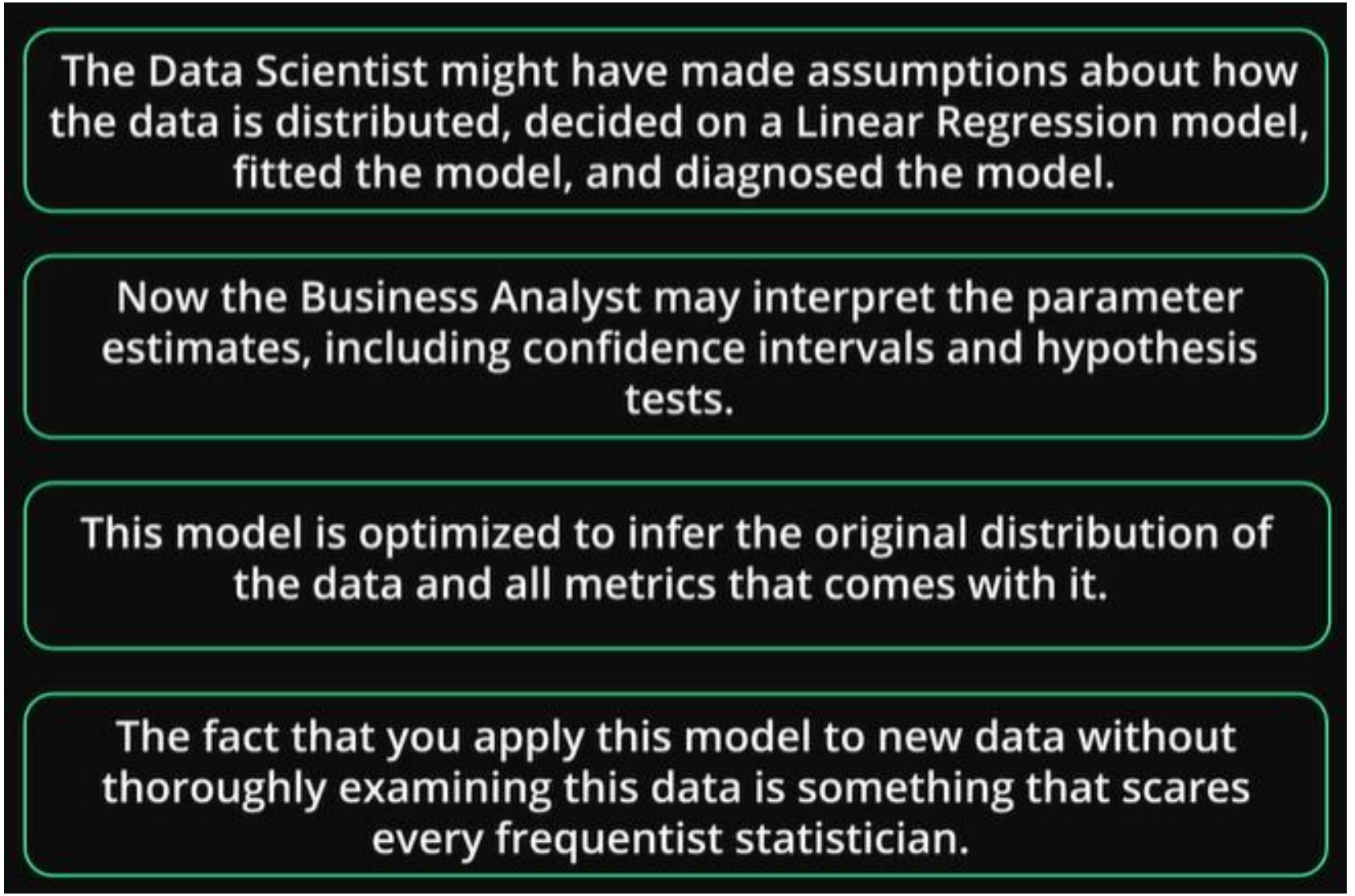
Recap – ML Fundamentals

Rina BUOY

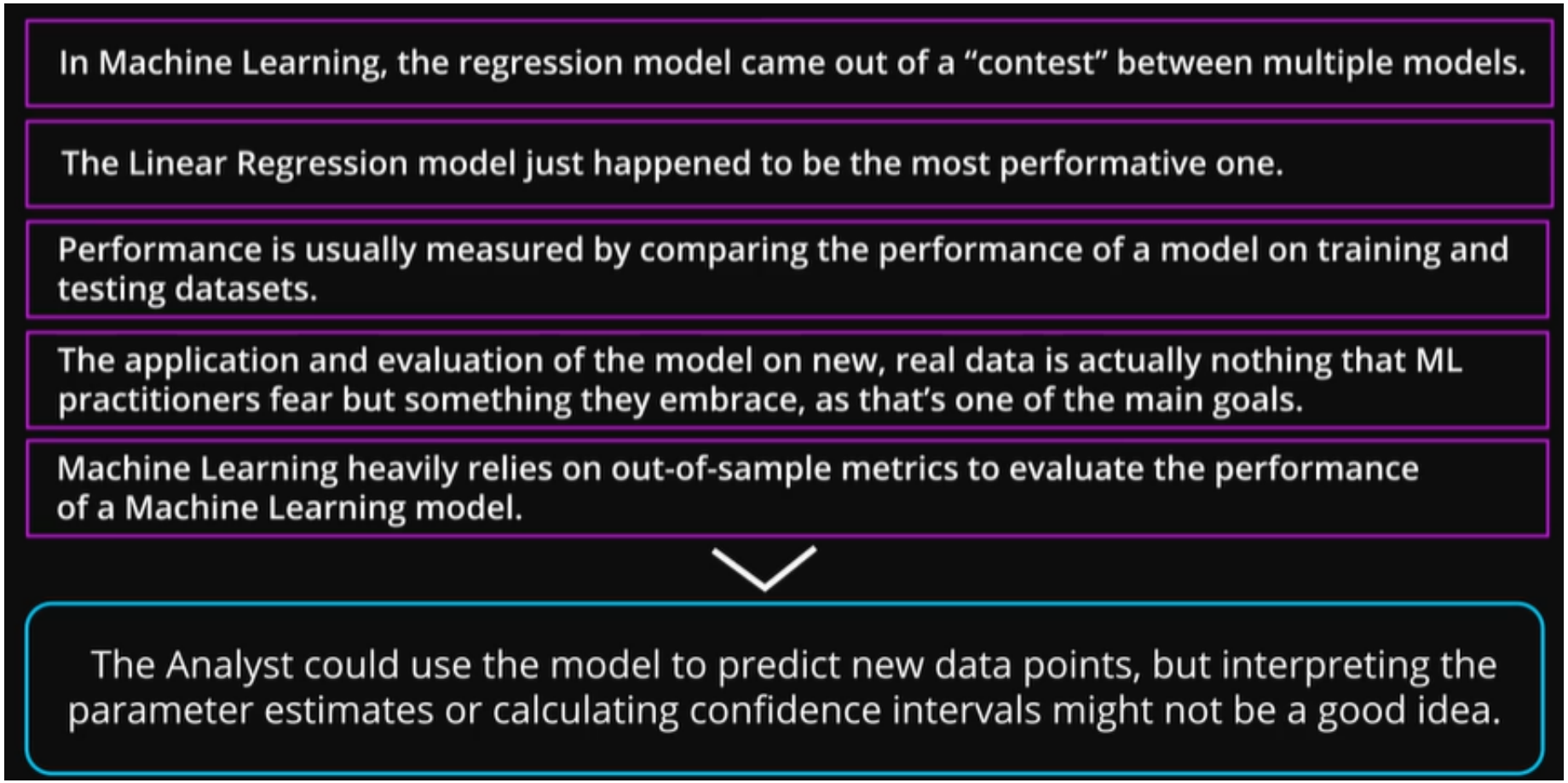
AMERICAN UNIVERSITY OF PHNOM PENH

STUDY LOCALLY. LIVE GLOBALLY.

Statistics Mindset



Machine Learning Mindsets



Statical Evaluation

Training Data

- Used to fit models (estimate parameters).
- Typically, 60-80% of the data set.

Validation Data

- Used to measure the **error of candidate models**.
- Training and validation are performed iteratively until model achieves the desired performance.

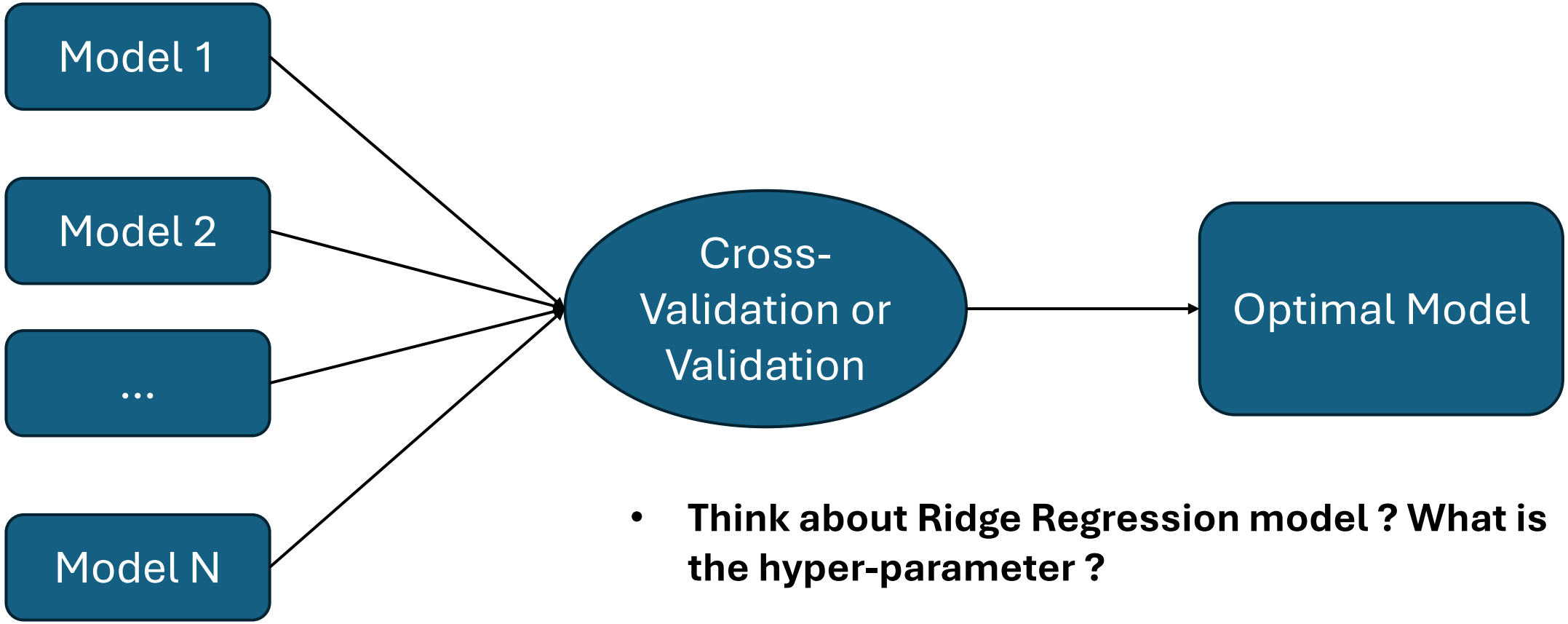
Test Data

- Used to measure the **performance of the final selected model**.
- True blind test as validation data was used multiple times during training.

Dealing with Imbalanced Dataset - Target Variable Strategies

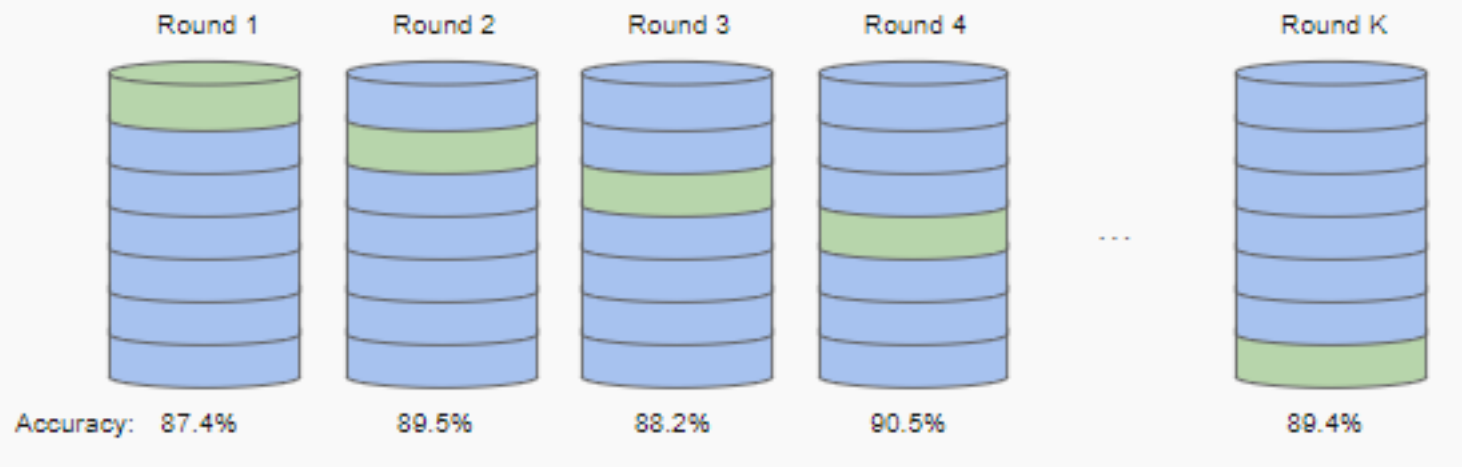
Oversampling the minority class	• Randomly duplicate samples from the minority class until the classes are balanced out.
Undersampling the majority class	• Remove samples from the majority class randomly until both classes contain an equal number of observations.
Generate synthetic data	• Create new, synthetic samples for the minority class by interpolating existing data points.
Penalize algorithms	• Modify learning algorithm to consider class imbalance.
Try a different algorithm	• Algorithms that usually work well on imbalanced data: random forests, boosted trees, naive bayes classifiers, K nearest neighbors. • Algorithms that usually do not work well on imbalanced data: linear regression, logistic regression, neural networks.

Model Selection



- **Think about Ridge Regression model ? What is the hyper-parameter ?**
- **Think about Polynomial Regression model ? What is the hyper-parameter ?**

Cross Validation – For Model Selection/Hyper-parameter Tuning



- Split the data into K folds.
- K-1 folds are used for training, 1-fold for validation.
- Shuffle the validation fold k times.

Alternatively