Predictive Analytics – Session 1

The Predictive Mindset
Predicting with Regression

Associate Professor Ole Maneesoonthorn

Associate Professor in Econometrics and Statistics

Melbourne Business School

O.Maneesoonthorn@mbs.edu

What is your perception of

PREDICTIVE ANALYTICS?

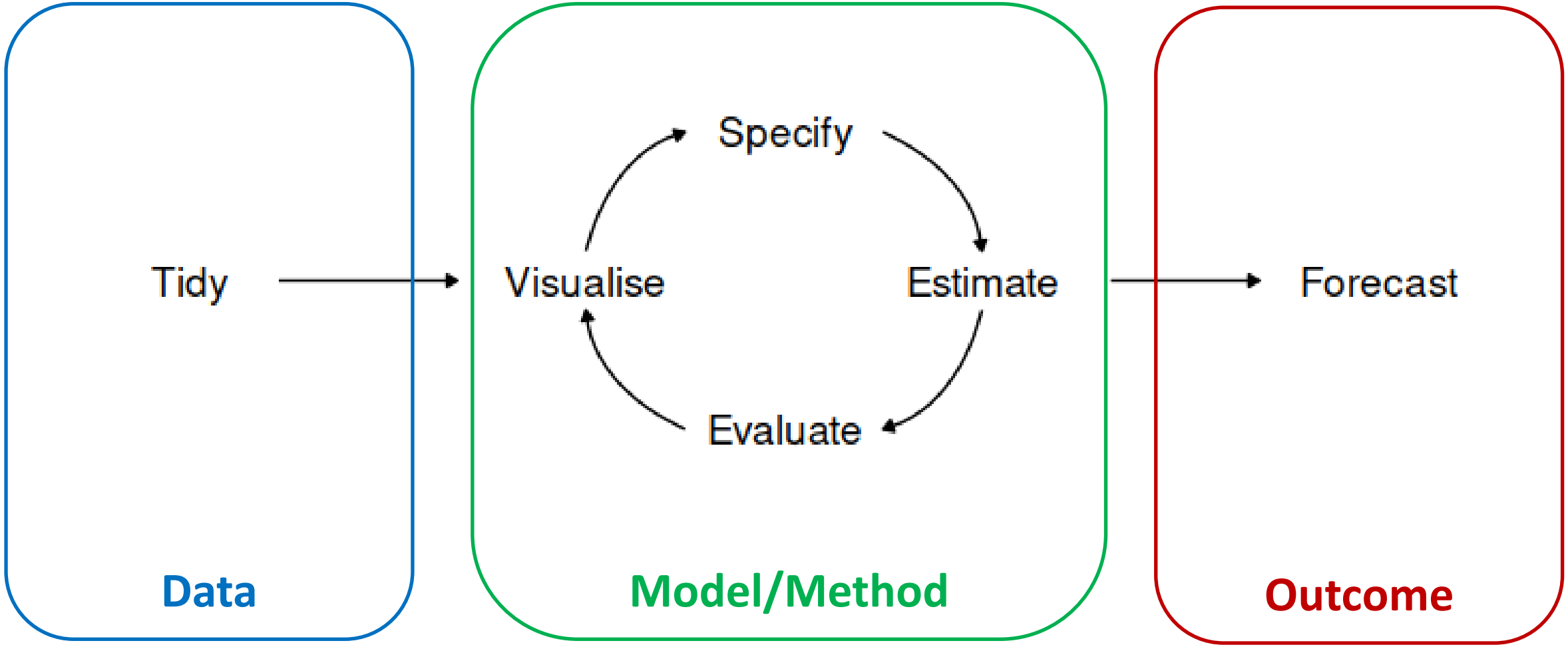
The Predictive Mindset

Predictive Analytics:

- Focus on the future – “Predictive”
- … by learning from the past – “Analytics”
- **For your business context**

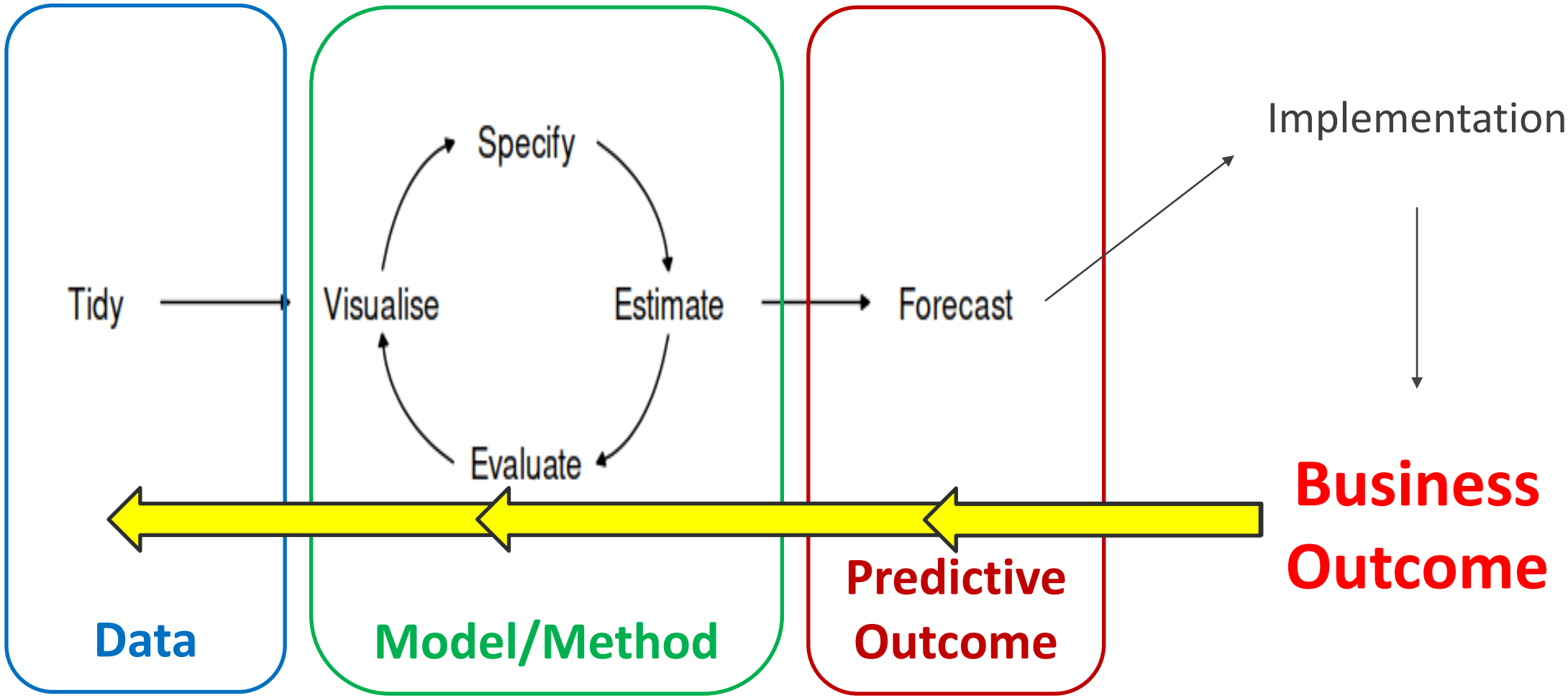
The Predictive Mindset

The predictive workflow



The Predictive Mindset

The predictive workflow



The Predictive Mindset

Example: What do loan lenders want?

		Actual Observation	
		Good loan	Bad loan
Predicted Outcome	Good loan	True Positive	False Positive
	Bad loan	False Negative	True Negative

The Predictive Mindset

Example: What about fraud detection?

		Actual Observation	
		Fraud	Not Fraud
Predicted Outcome	Fraud	True Positive	False Positive
	Not Fraud	False Negative	True Negative

The Predictive Mindset

Discussions:

- Does “good fit” imply “good prediction”?

- Is Prediction = Truth?

The Predictive Mindset

“All models are wrong…

but some are useful…”

George Box (1987)

The Predictive Mindset

Key Questions

- How large is the potential error from predictions?
- How does prediction error impact our business?
- Are there other methods that returns smaller error? How do we compare them?

Predictive Assessments

Quantifying potential error

- We know from the outset that **Prediction ≠ Truth**
- How large is this potential error?

PREDICTION INTERVAL

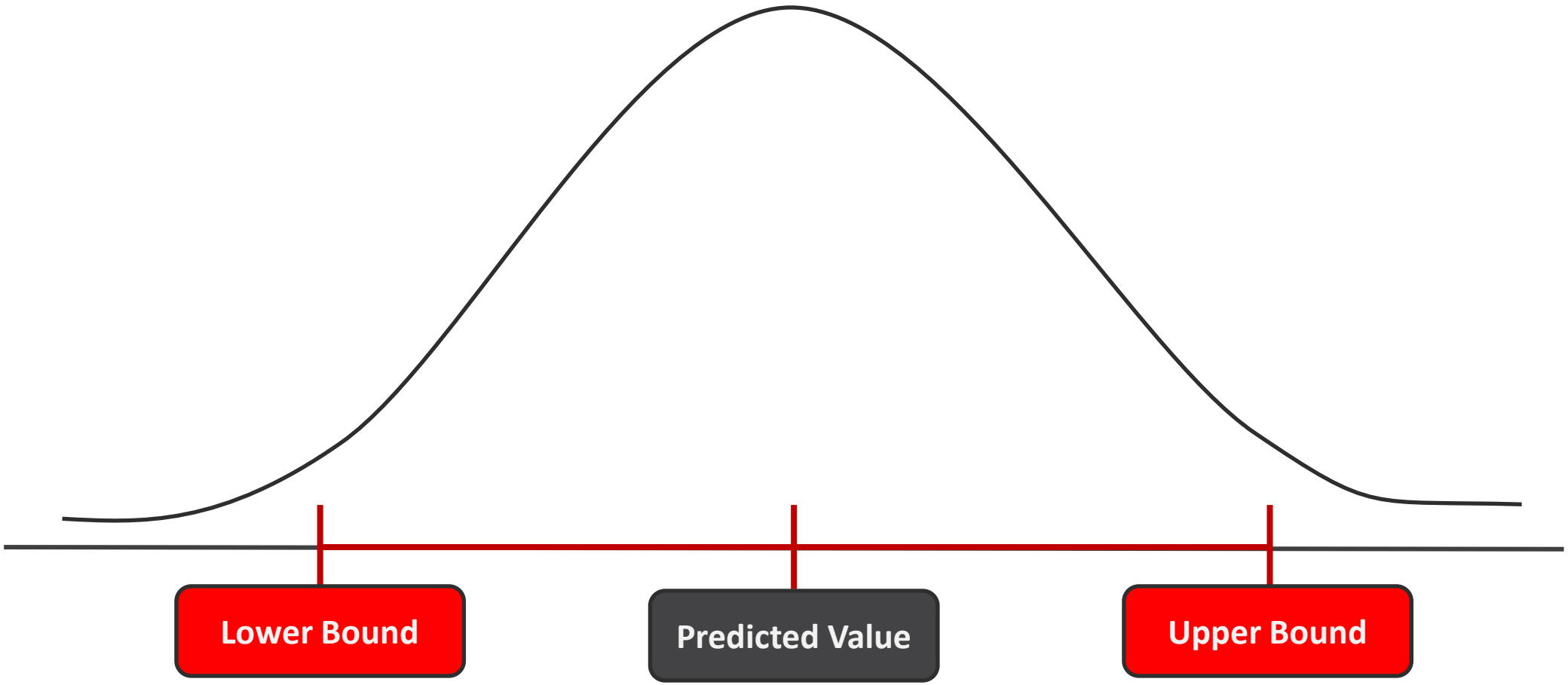
Predictive Assessments

Quantifying potential error

- The **prediction interval**: a range that we expect cover the truth with a given probability
- E.g. 95% prediction interval: we expect that the truth will be within this interval with 95% probability

Predictive Assessments

Quantifying potential error



Predictive Assessments

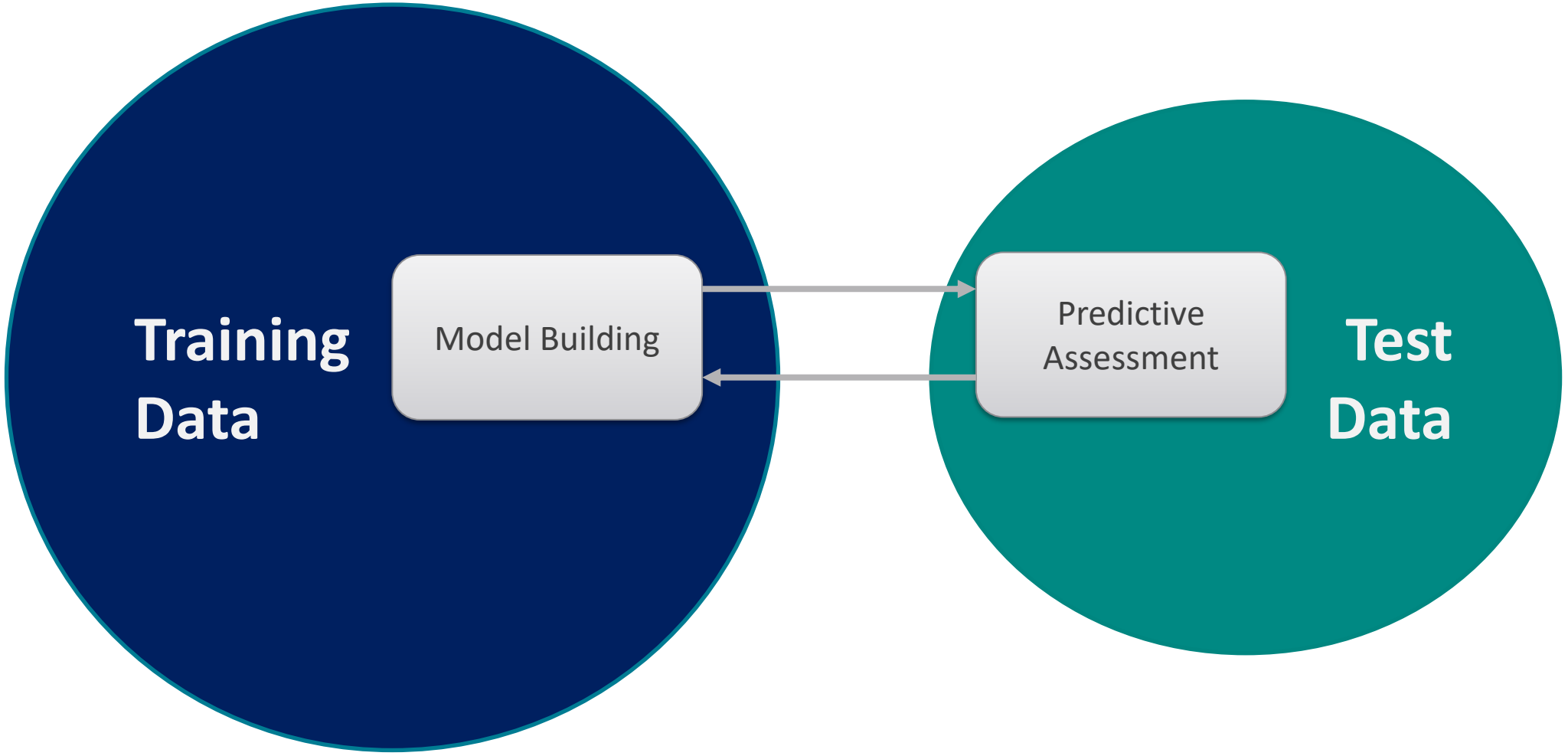
Quantifying potential error

- Small model error → small potential error
- The further away the input is from its centre → larger potential error
- Smaller sample size → larger potential error
- **Narrower** interval for a given probability level is **preferred**

Predictive Assessments

But how do we know if the model/method generate accurate prediction?

Predictive Assessments

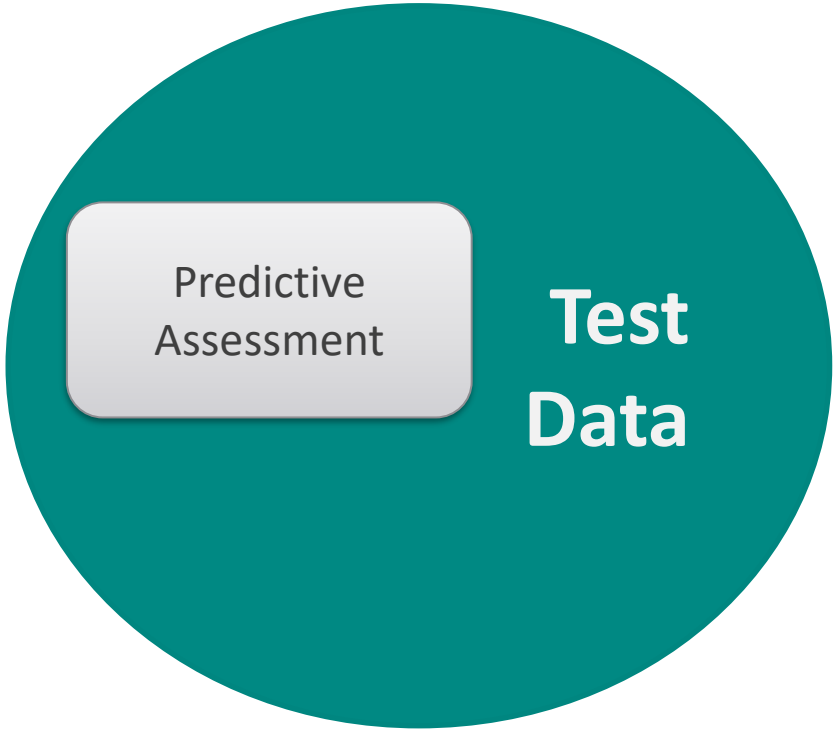


Predictive Assessments

- Split the data into training & test sets
- Both should have **similar characteristics** and **representative** of the population
- Training data – use to build models
- Typically larger in size than the test data

Predictive Assessments

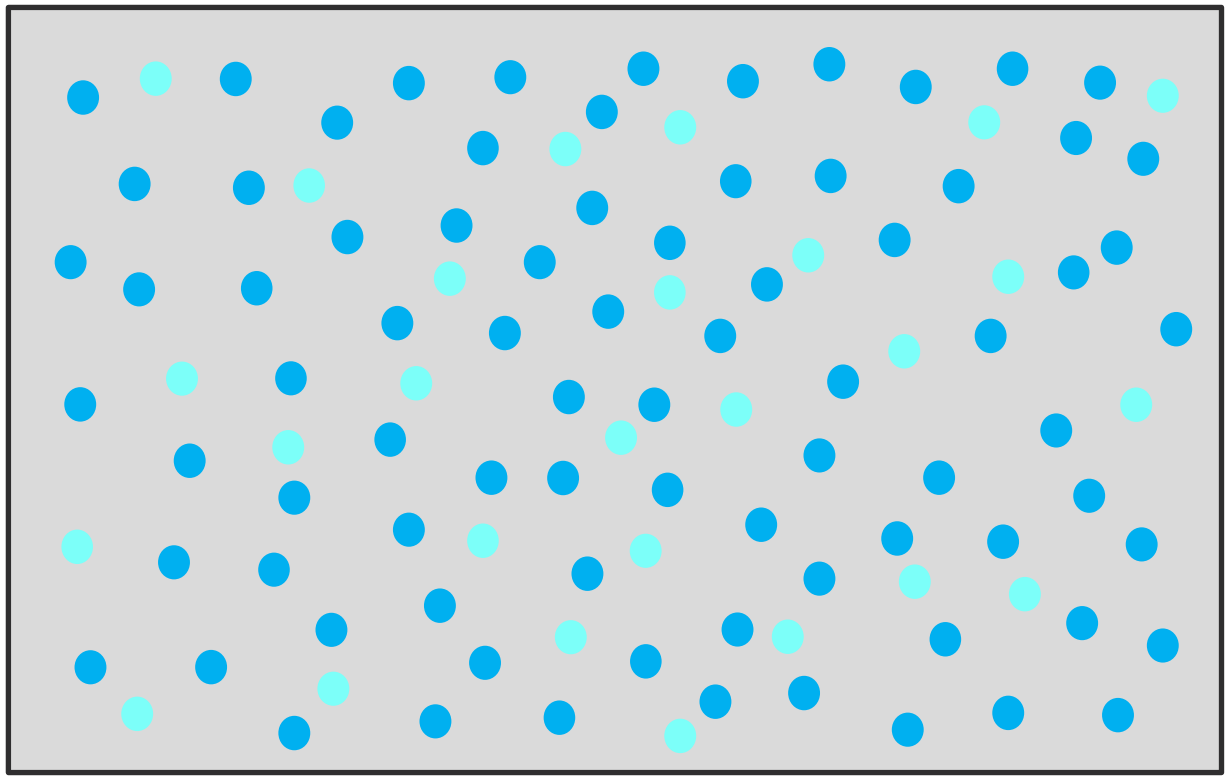
- Use the model to construct from the training data
- Predict the observations using test set characteristics
- Compare the predictions to the actual realizations
- **Error = Actual – Prediction**



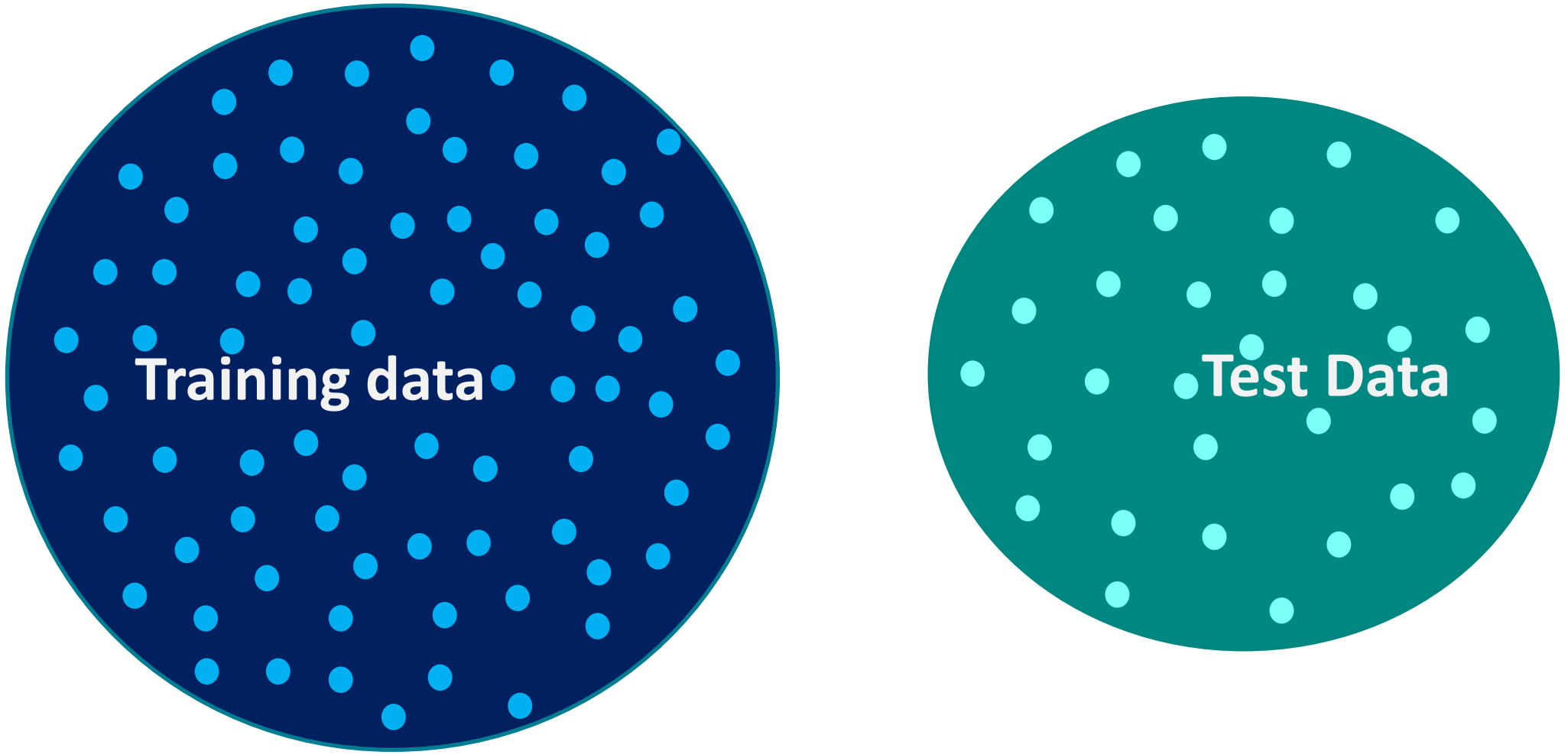
Predictive Assessments

Training vs test split

- For cross-sectional data, we need to make sure that the two samples are representative of population



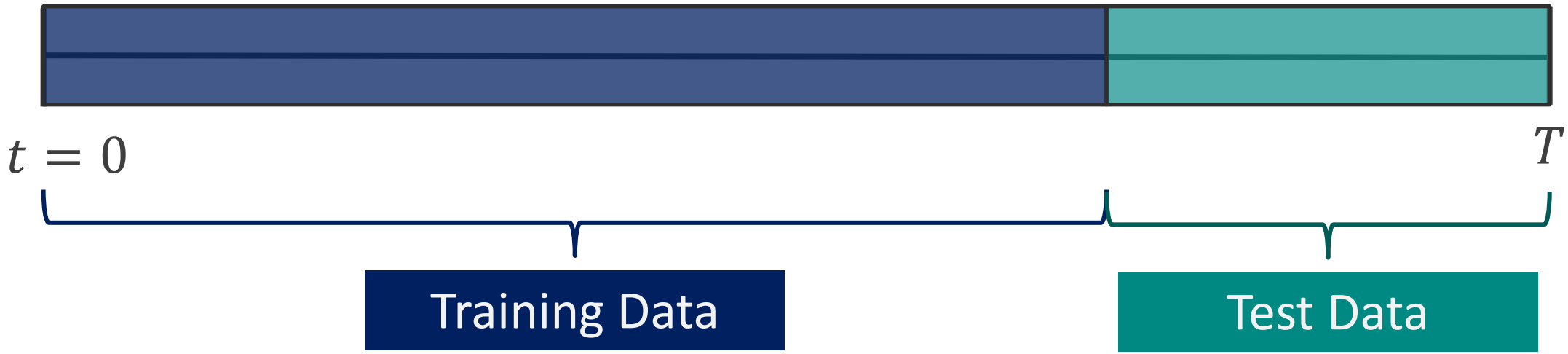
Predictive Assessments



Predictive Assessments

Training vs test split

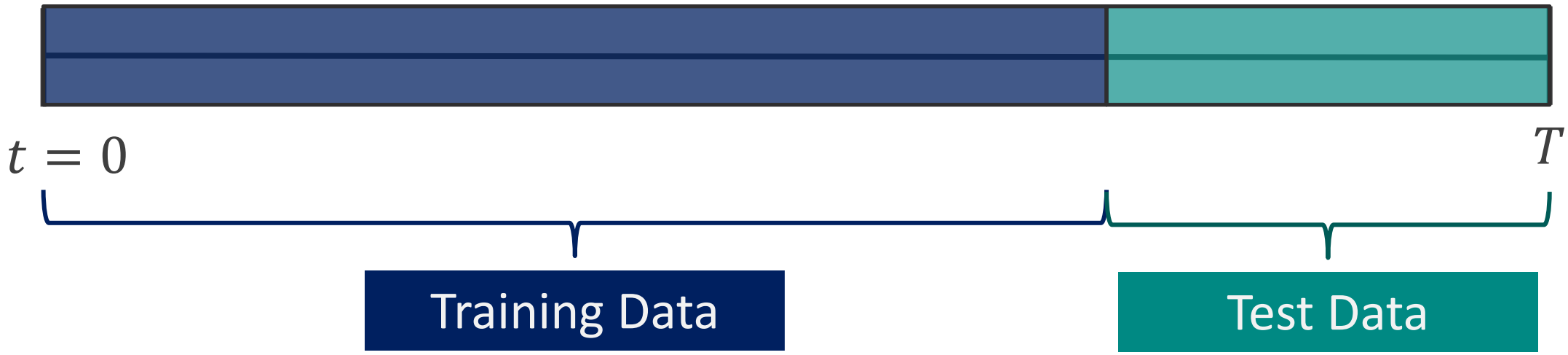
- For time series data, there is a natural sequence
- Data split by time



Predictive Assessments

Time series predictions

- **Long-horizon** prediction
- Short-horizon **real time** prediction



Predictive Assessments

Evaluating predictive accuracy

- Various statistics available to measure predictive accuracy
- Computed over the **test data**
- **E.g. Mean squared error**

$$MSE = average(error^2)$$

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(e_i)^2$$

Where $e_i = Actual_i - Predicted_i$

Predictive Assessments

Commonly reported measures of predictive accuracy:

- Root mean square error (RMSE)
- Mean absolute error (MAE)
- Mean absolute percentage error (MAPE)
- Mean absolute scaled error (MASE)

Smaller values are preferred!

Predictive Assessments

Commonly reported measures of predictive accuracy:

- Root mean square error (RMSE)

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(e_i)^2}$$

- In the same unit as the data
- Akin to the out-of-sample standard error of the regression
- **Commonly used** and reported metric
- Highly **influenced by extreme values** due to the power function

Predictive Assessments

Commonly reported measures of predictive accuracy:

- Mean absolute error (MAE)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|e_i|$$

- In the same unit as the data
- **Overcome the issue of extreme value** domination in the MSE and RMSE

Predictive Assessments

Commonly reported measures of predictive accuracy:

- Mean absolute percentage error (MAPE)

$$MAPE = 100 \times \frac{1}{N} \sum_{i=1}^{N} \left| \frac{e_i}{Actual_i} \right|$$

- In percentage unit – unit insensitive
- Measures how large the predictive error is **relative** to actual value
- Con: Undefined if “Actual” is zero, and gets “explosive” when “Actual” is close to zero
- **Do not use if the data can the value of zero!**

Predictive Assessments

Commonly reported measures of predictive accuracy:

- Mean absolute scaled error (MASE)

$$MASE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{e_i}{BE}\right|$$

where $BE = \frac{1}{N}\sum_{i=1}^{N}|Actual_i - Naive|$

- Measures how well a method predicts **relative to the naïve** prediction
- For cross sectional data, the naïve method is the sample mean

Predictive Assessments

Commonly reported measures of predictive accuracy:

- Mean absolute scaled error (MASE)

$$MASE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{e_i}{BE}\right|$$

- Benchmark value for MASE is 1.
- **MASE = 1:** your method is doing as good as the naïve
- **MASE > 1:** your method is doing worse than the naïve
- **MASE < 1:** your method is doing better than the naïve ← **PREFERRED!**

Predictive Assessments

Evaluating predictive accuracy

- Note that these are **statistical** metrics
- That is, they are simply measure of **distances** (or relative distances)
- In many cases, there will be **conflicts** across these measures
- The impact of prediction errors on the **business** may not be captured accurately
- Consider defining **user-specific loss** functions of prediction errors

Predictive Assessments

Evaluating predictive accuracy

- Example: predicting demand impacts has certain implications on inventory management.
- Would overpredicting have the same consequence as underpredicting demand?

Predictive Assessments

Evaluating predictive accuracy

- Example: predicting real estate prices.
- What would be the consequence of errors for a banker using the predicted price to assess loan collateral?
- What would be the consequence of errors for a real estate agent using the predicted price to propose an advertise price?

Predicting with Regression

- The **multiple linear regression – how do we predict?**

$$Y = c + b_1X_1 + b_2X_2 + \cdots + b_kX_k + error$$

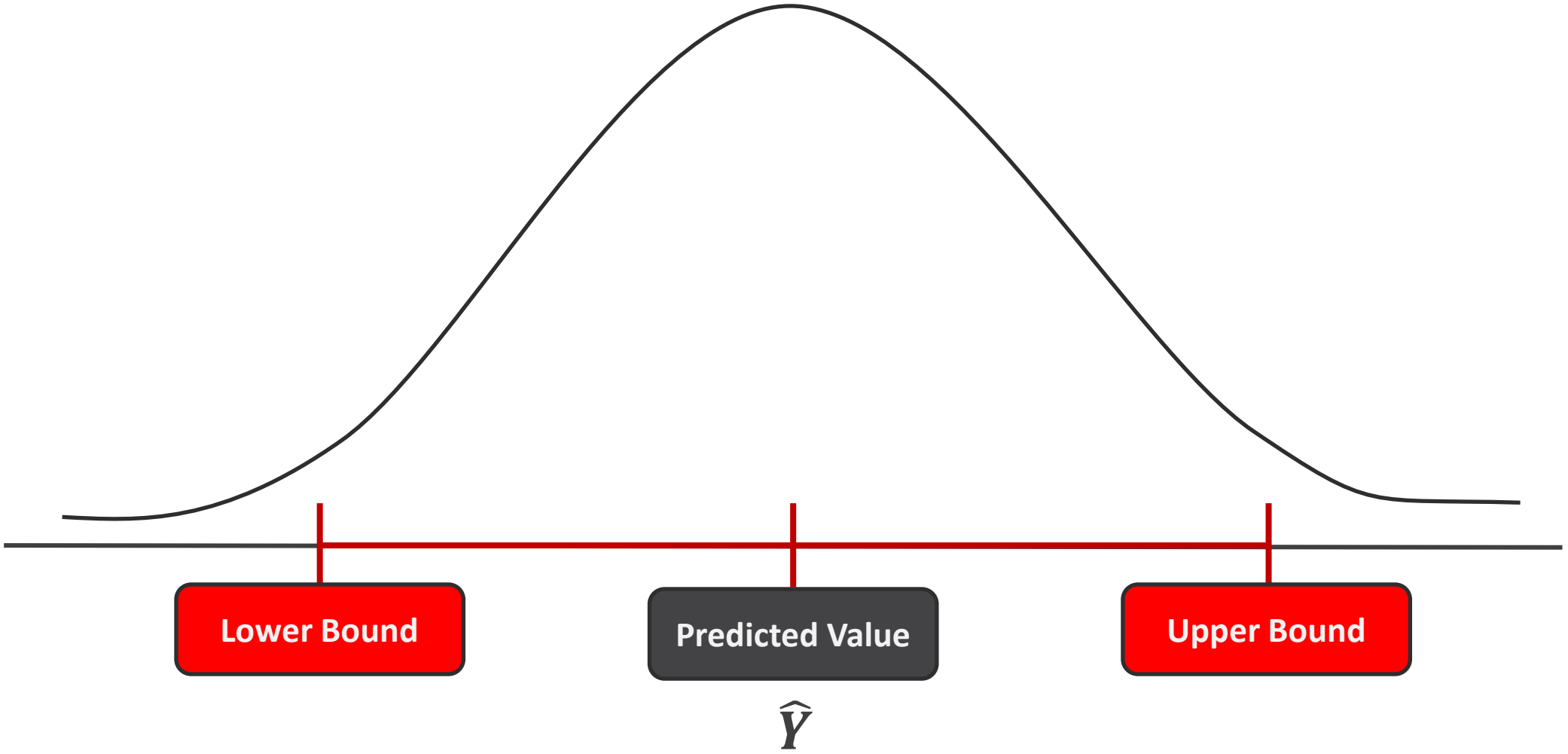
- **Form an expectation!**

$$\hat{Y} = E(Y|X_1, X_2, \ldots, X_k) = c + b_1X_1 + b_2X_2 + \cdots + b_kX_k$$

- Because the error is always present, $actual \neq prediction$

Predicting with Regression

Quantifying potential error



Predicting with Regression

- The **multiple linear regression – the prediction**

$$\hat{Y} = E(Y|X_1, X_2, \dots, X_k) = c + b_1X_1 + b_2X_2 + \cdots + b_kX_k$$

- **Quantifying the uncertainty** – the 95% prediction interval

$$LowerBound = \hat{Y} - 1.96(Prediction\ Error)$$

$$UpperBound = \hat{Y} + 1.96(Prediction\ Error)$$

- The model indicates that there is a 95% **probability** that the actual value will be between the lower bound and the upper bound

Predictive Assessments Work Through

Regression example

- Goal: Construct a **predictive model** for property prices (Real Estate data)
- Let's compare
  - Multiple linear regression model
  - Stepwise regression
  - Nonlinear regression with quadratic and interactions
- **Real Estate Prices with Regression page**

Predictive Assessments Work Through

Regression example

- Goal: Construct a predictive model for property prices (Real Estate data)
- What do the statistical metrics say?

	RMSE	MAE	MAPE	MASE
Linear	296.337	185.768	13.885	0.592
Stepwise	295.628	186.304	13.940	0.593
Nonlinear	**280.542**	**175.747**	**13.387**	**0.560**

Green is best

Red is worst

Predictive Assessments Work Through

Regression example

- What if you are a banker? How do the models compare if we value a conservative prediction more than an overestimate? (as lenders!)
- →Give double the weight to errors from overestimation

	Banker
Linear	259.730
Stepwise	260.297
Nonlinear	**245.350**

Green is best

Red is worst

Predictive Assessments Work Through

Regression example

- What if you are a real estate agent?
  - Low: error is within 10% of actual price.
  - Medium: error is within between 10% and 30% of actual price.
  - High: error is greater than 30% of actual price.
- The medium level errors are penalized 5 times greater than the low level error.
- The high level errors are penalized 10 times greater.

Predictive Assessments Work Through

Regression example

- What if you are a real estate agent?
- Loss function a bit more complex…

	RE Agent
Linear	75.437
Stepwise	77.586
Nonlinear	80.461

Green is best

Red is worst

Predictive Assessments Work Through

which ever loss function would need to be theoritically best.

Regression example

- Goal: Construct a predictive model for property prices (Real Estate data)
- Collating all statistics

	RMSE	MAE	MAPE	MASE	Banker	RE Agent
Linear	296.337	185.768	13.885	0.592	259.730	**75.437**
Stepwise	295.628	186.304	13.940	0.593	260.297	77.586
Nonlinear	**280.542**	**175.747**	**13.387**	**0.560**	**245.350**	80.461

Green is best

Red is worst