# ENSF 611 – Machine Learning for Software Engineers

## Week 2 – Statistical Learning

UNIVERSITY OF CALGARY

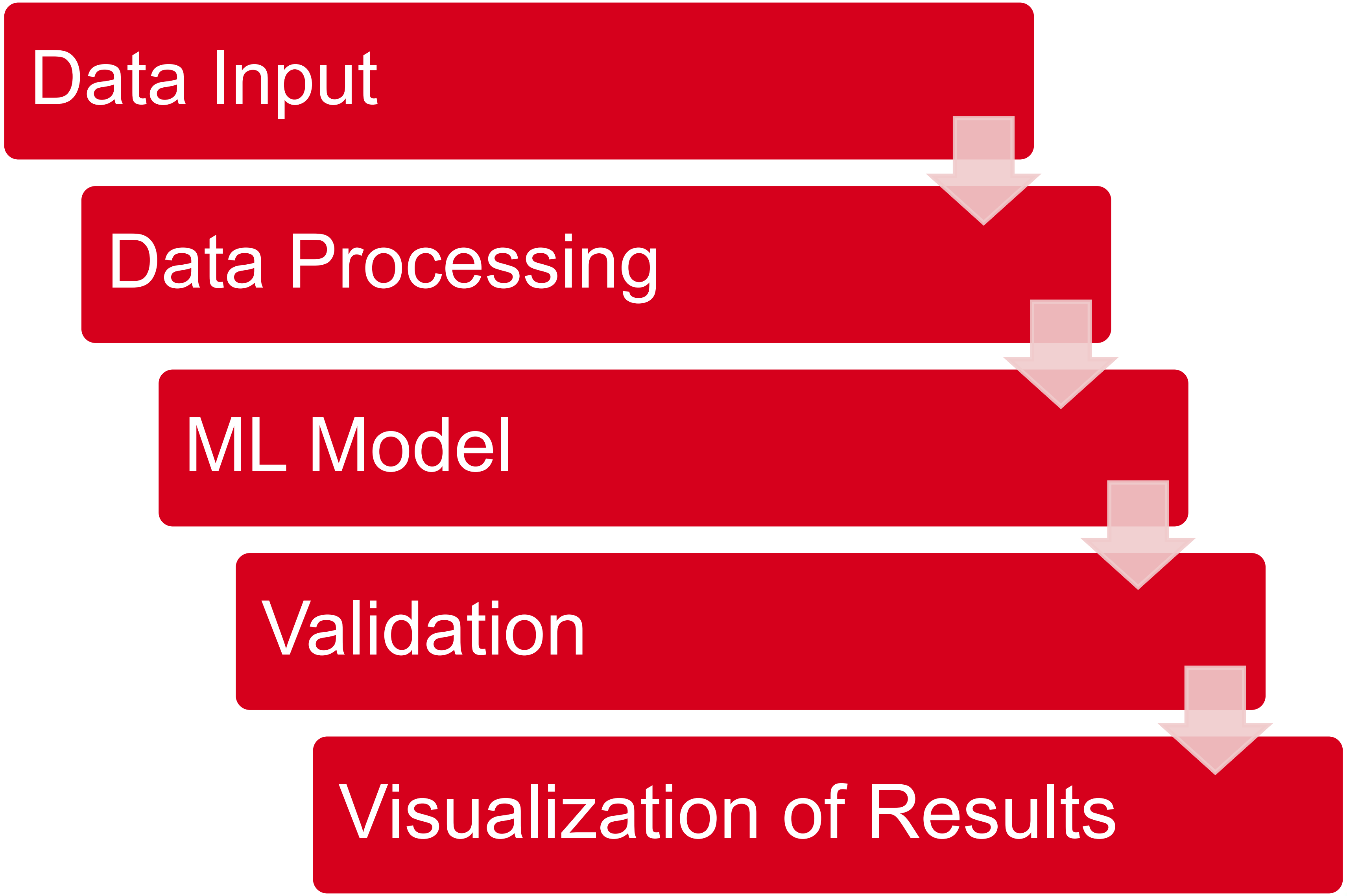
# Lecture Goals

- Introduce the mathematical perspective for machine learning

- Chapters 1 and 2 of *An Introduction to Statistical Learning with Applications in Python*

UNIVERSITY OF CALGARY

# ML Workflow



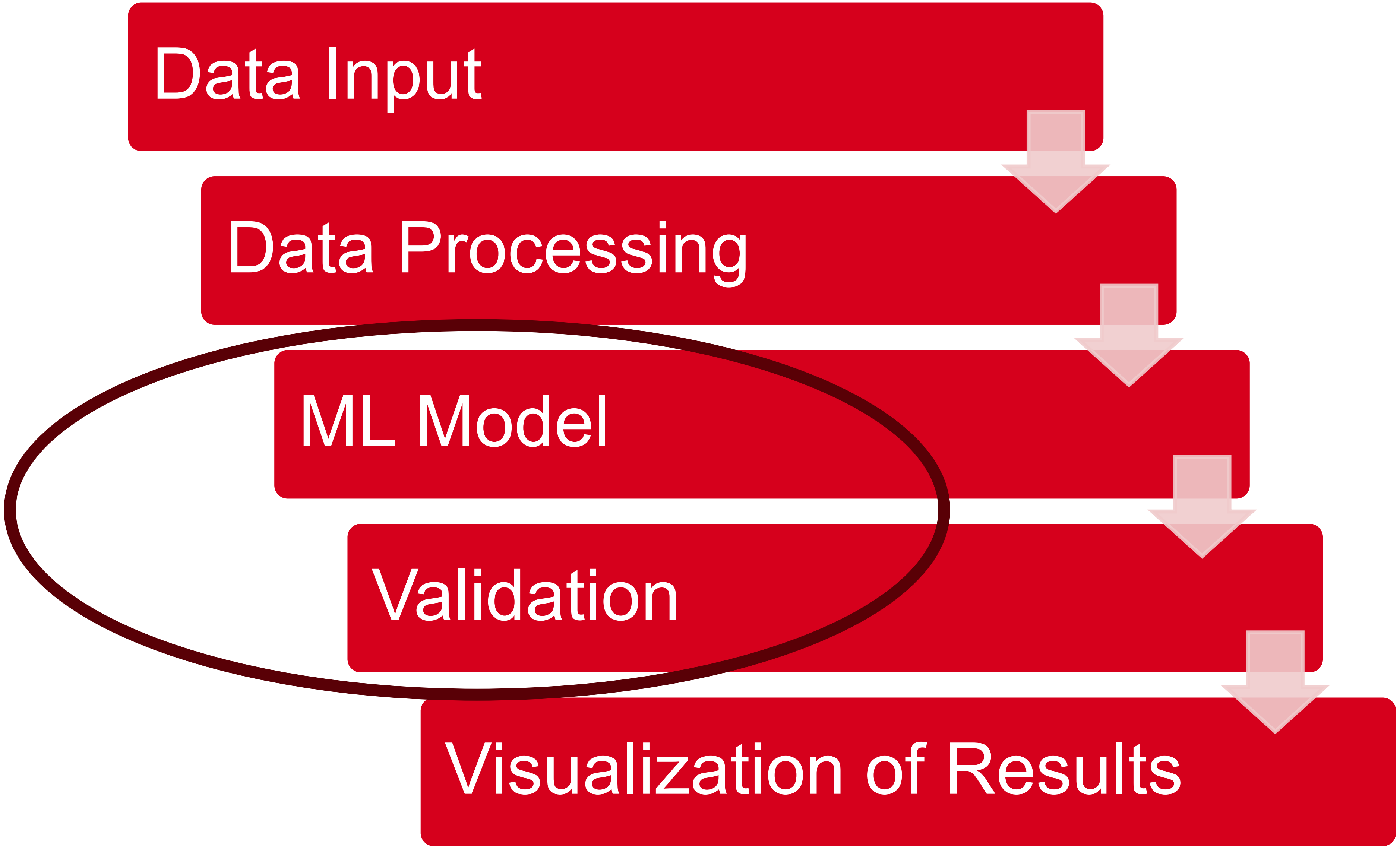
Data Input

Data Processing

ML Model

Validation

Visualization of Results



Data Input

Data Processing

ML Model

Validation

Visualization of Results

# What is Statistical Learning?

Do not edit

How to change the design

# What is a synonym of statistical learning?

# What is statistical (machine) learning?

- Statistical learning refers to a vast set of tools for **understanding data**
- These tools can be classified as **supervised** or **unsupervised**
- Supervised statistical learning involves building a statistical model for **predicting**, or **estimating**, an **output** based on one or more **inputs**

# Output-input relationship

- In mathematics, we represent the relationship between a set of inputs and an output as:

$$y = f(x)$$

- Where *y* represents the output, *x* represents the input and *f(…)* represents the relationship between the two

# Wage Example

- Imagine that you have been given income survey data, which includes the wage, education level, age and year
- You would like to use this data to predict what the wages will be for other samples
- To do this, you need to figure out what is the relationship between wage and the other measured factors

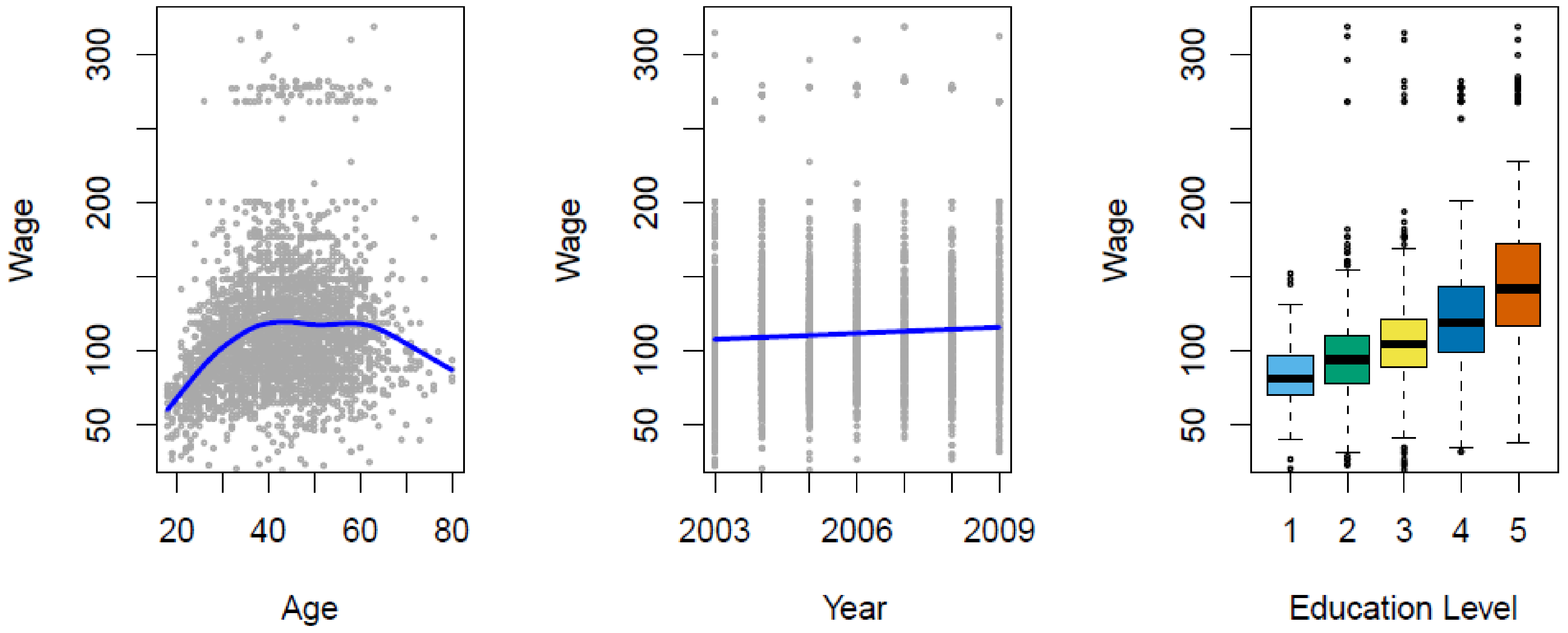


**FIGURE 1.1.** Wage *data, which contains income survey information for men from the central Atlantic region of the United States.* Left: wage *as a function of* age. *On average,* wage *increases with* age *until about* 60 *years of age, at which point it begins to decline.* Center: wage *as a function of* year. *There is a slow but steady increase of approximately* $10,000 *in the average* wage *between* 2003 *and* 2009. Right: *Boxplots displaying* wage *as a function of* education, *with* 1 *indicating the lowest level (no high school diploma) and* 5 *the highest level (an advanced graduate degree). On average,* wage *increases with the level of education.*

# Wage Example

- Since the wage seems to be impacted by all three factors (age, year and education), the most accurate prediction would be given by combining all three into one model
- Assuming that:

$$wage = f(age, year, education)$$

- How do we find the best function to give us the most accurate prediction?
- Knowing that we need to account for the non-linear relationship between wage and age

Do not edit

How to change the design

# What are potential sources of model inaccuracy ?

# Model Accuracy

- Can we be 100% accurate with our prediction?
- A more general form of the input-output expression:

$$y = f(x) + \in$$

Where ∈ is a random error term, which is independent of X and has a mean of zero

# Model Accuracy

- We can predict the y values using the given data:

$$\hat{y} = \hat{f}(x)$$

Where $\hat{y}$ are the predicted values based on the estimated relationship, represented by $\hat{f}$, and the given data, $x$

# Model Accuracy

- Considering a given estimated function and a given input, assume that both are fixed, and that the only variability comes from the error term:

$$E(y-\hat{y})^2 = E(f(x)+\in -\hat{f}(x))^2$$

$$E(y-\hat{y})^2 = \underbrace{(f(x)-\hat{f}(x))^2}_{\text{Reducible}} + \underbrace{Var(\in)}_{\text{Irreducible}}$$

Where $E(y-\hat{y})^2$ represents the average (**expected value**) of the squared expected difference between the predicted and actual values, and $Var(\in)$ represents the variance associated with the error term

# Model Accuracy

- Since the random error term (irreducible error) is independent of the input, there will always be some error in our prediction
  - This error is usually caused by missing information
  - You can’t account for every factor that influences the output
- The goal of machine learning is to minimize the **reducible error** by selecting the best fitting model, based on the available data

# What does this mean?

- The training data is used to estimate the relationship ($\hat{f}$) between the given input and output

- The testing data is used to:
  - Estimate the predicted values ($\hat{y}$) using ($\hat{f}$)
  - Compare the predicted values ($\hat{y}$) to the given values ($y$) to calculate the accuracy (classification) or error (regression)

# What’s next?

- Discuss what metrics to use to compare model performance

- How do we decide if a model works well or not?

- What is the relationship between model performance and model complexity?