Introduction to Forecasting in Data Science

Dr. Khaing S Htun

What is Forecasting?

predict or estimate about the future as accurately as possible with available (historical) data

Some things are easier to forecast than others.

- Exact time of the sun will rise tomorrow morning
- Lottery numbers

The predictability of an event or a quantity depends on several factors including:

- how well we understand the factors that contribute to it
- how much data are available
- whether the forecasts can affect the thing we are trying to forecast.

Is Forecasting Necessary?

Why do we need forecasts?

- To make timely decisions

Who makes forecasts?

- Business, households, government, etc.

What are the ingredients for good forecasting?

- Good judgment

Is Forecasting Necessary?

What is good judgement?

Ask RIGHT questions!!!

a. What are we concerned about?
b. What is the time period to be forecasted?
c. What is the level of aggregation required?
d. How much are we willing to spend to acquire information?
e. Is historical data relevant for the future?
f. What is the penalty for forecasting inaccuracy?

Forecasting problems occur in many fields

- Business and industry
- Economics
- Finance
- Environmental sciences
- Social sciences
- Political sciences

Forecasting Problems

Short-term forecasts

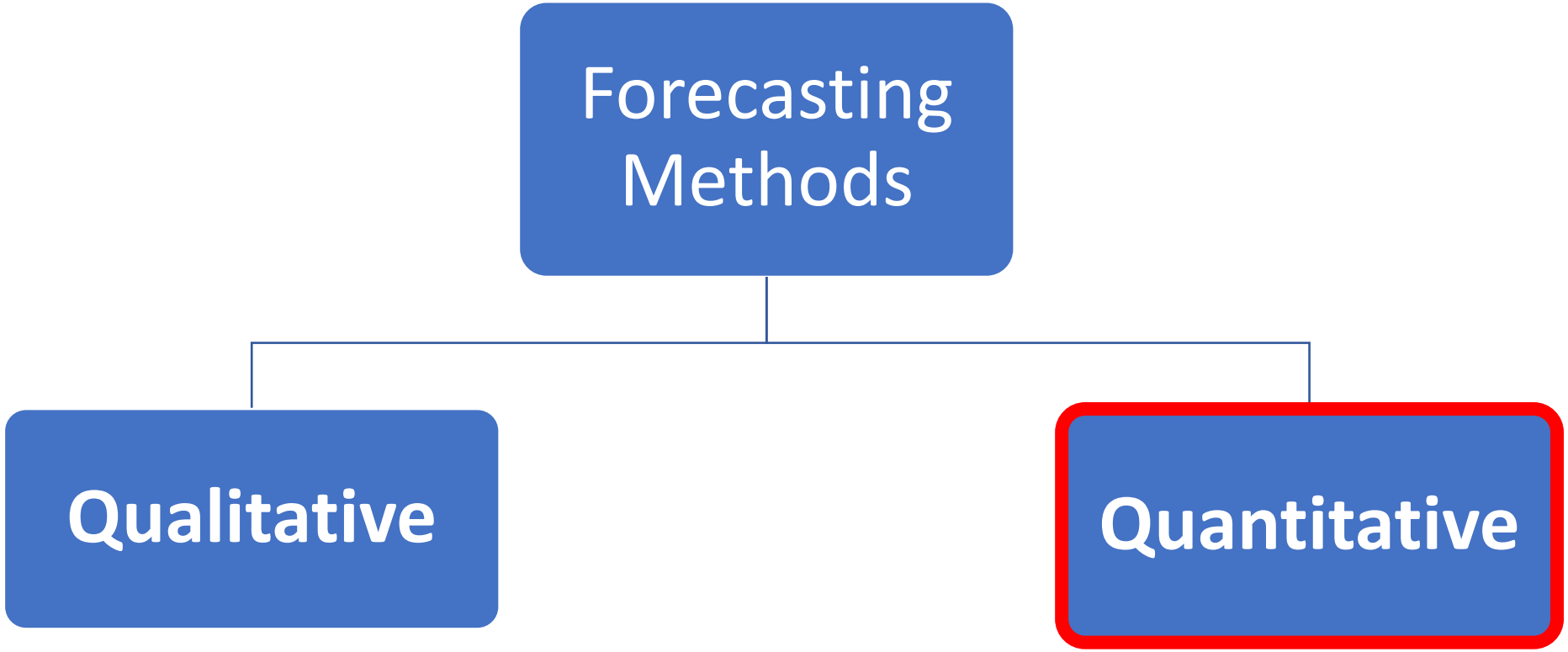
- Predicting only a few periods ahead (hours, days, weeks)
- Typically, bad on modeling and extrapolating patterns in the data

Medium-term forecasts

- One to two years into the future, typically

Long-term forecasts

- Several years into the future



- No (relevant) data availability
- Structured approach without using historical data

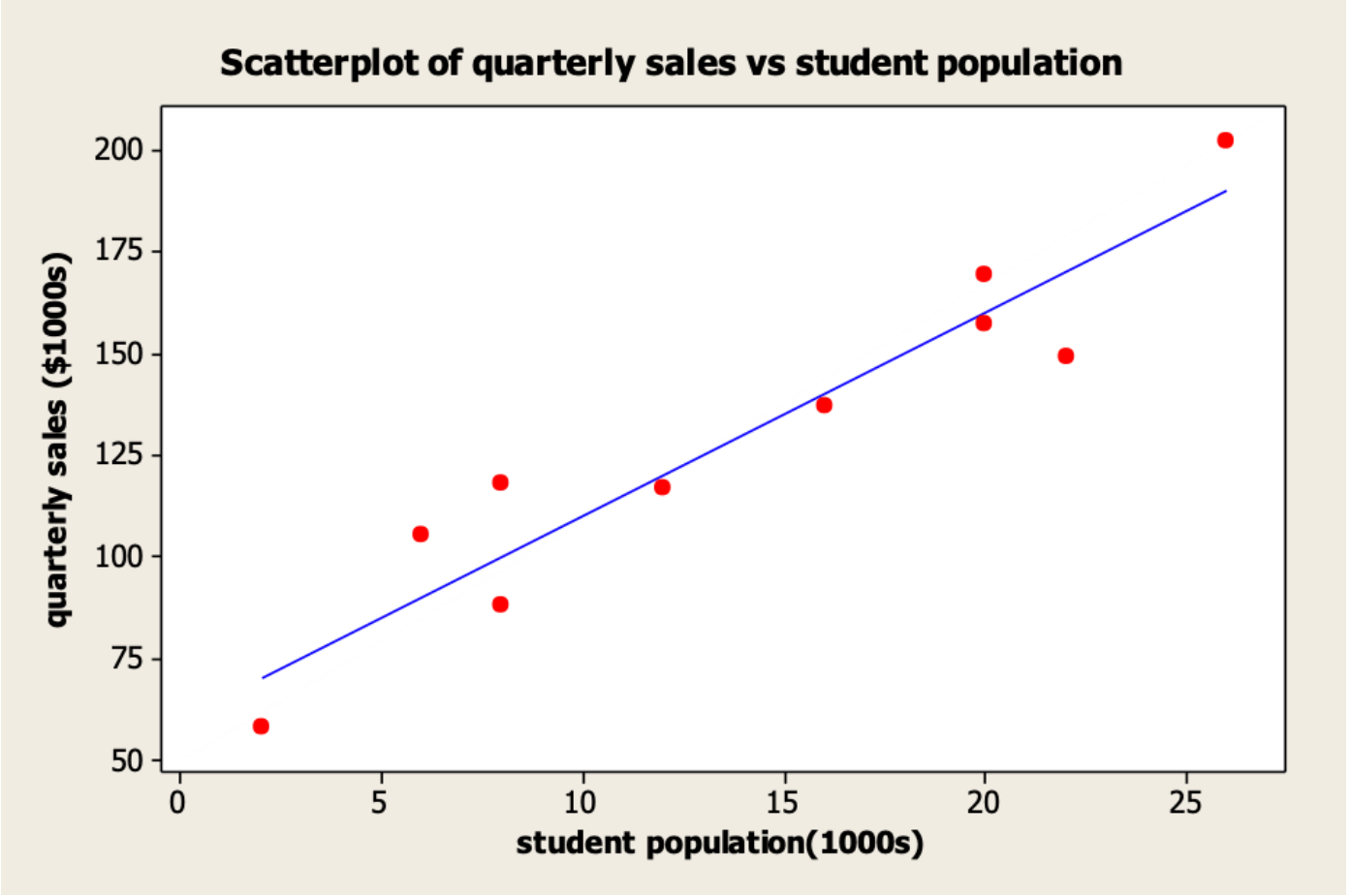
- Past data availability
- Reasonable to assume from past patterns will continue in the future

Types of Quantitative Forecasting Methods

1. **Regression Analysis**
2. **Time Series Analysis**

Regression Analysis

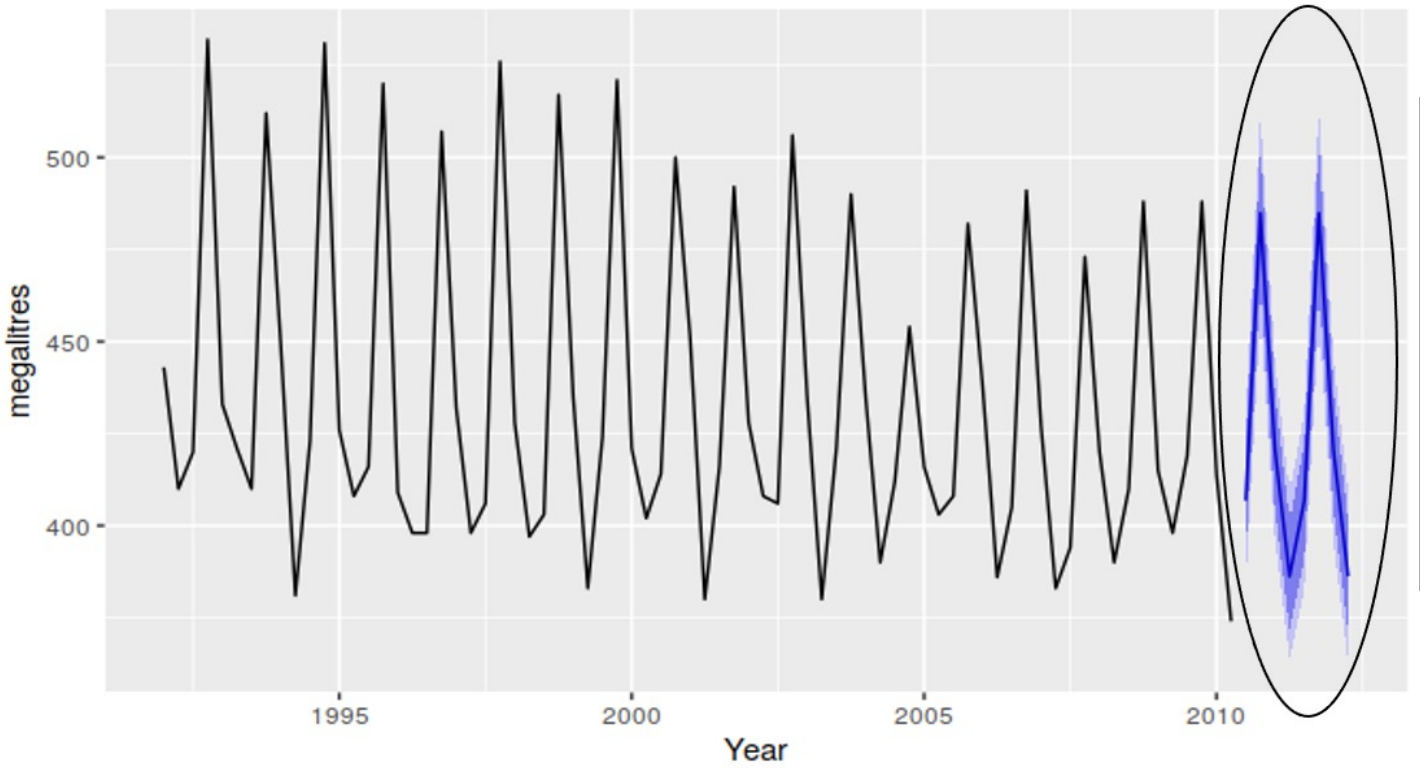
- Set of statistical methods to estimate the relationships between dependent and independent(s) variables
- Strength of the relationships
- To build regression model
- Variations
  1. **Linear regression**
  2. **Multiple regression**
  3. Non-linear regression



the scatterplot between the quarterly sales and the student population of Armand's Pizza restaurants

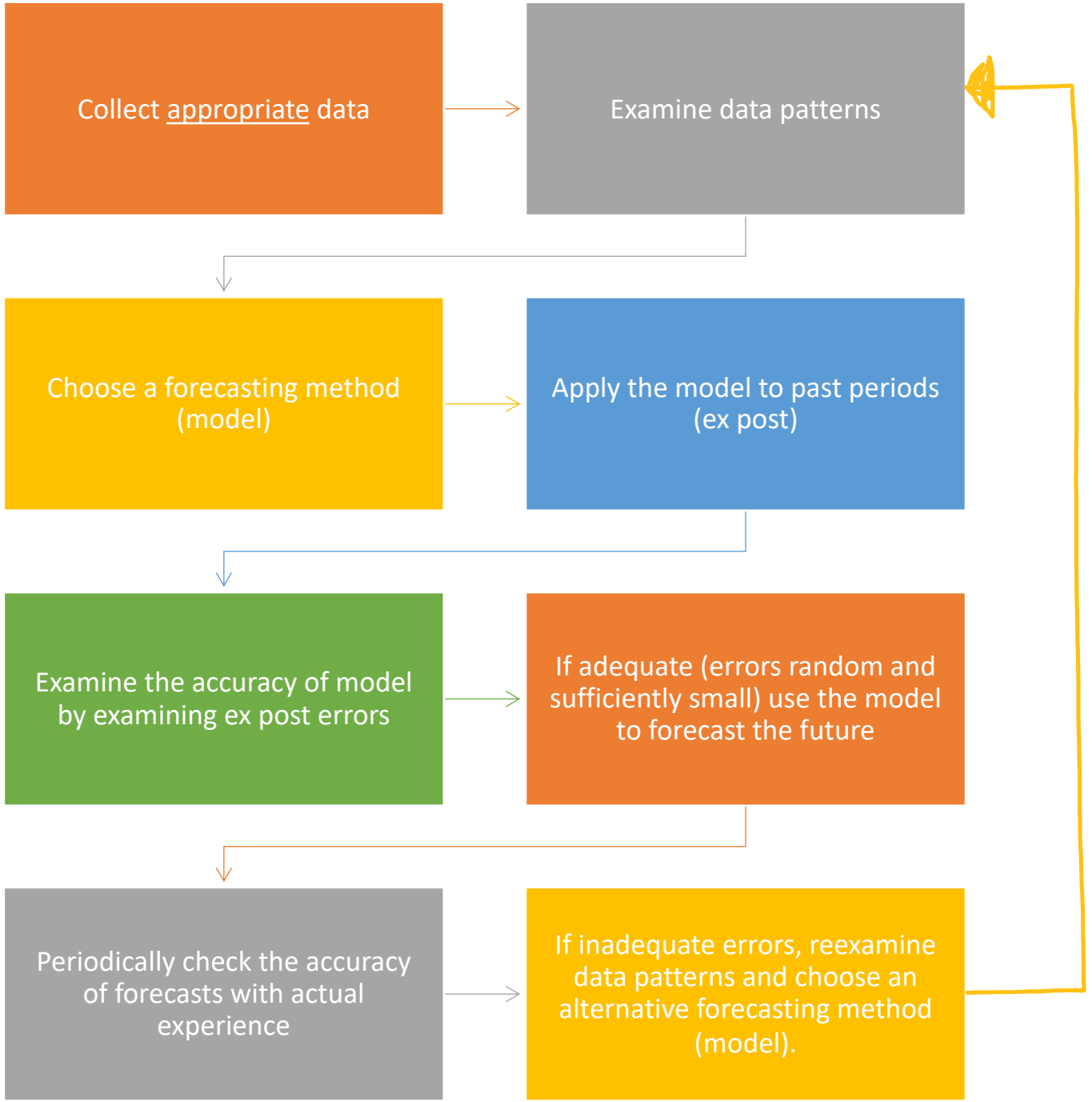
Time Series Analysis

- Observed sequentially over time **– at regular intervals of time** (hourly, daily, weekly, monthly, quarterly, annually)
- *Possible occurrence of irregularly spaced time series*
- Estimate how the sequence of observations will continue into the future
- Time series models
  1. Decomposition models
  2. Exponential smoothing models
  3. ARIMA models



Australian quarterly beer production: 1992Q1 – 2010Q2, with two years of forecasts

Forecasting Process



Types of Data

1. **Cross-sectional data**
   - collected at a single point in time

Company	Exchange	Annual Sales ($ millions)	Earnings / Share ($)
Advanced Comm. Systems	OTC	75.10	**0.32**
Ag-Chem Equipment Co.	OTC	321.10	**0.48**
Aztec Manufacturing Co.	NYSE	79.70	**1.18**
Cal-Maine Foods, Inc.	OTC	314.10	**0.38**
Chesapeake Utilities	NYSE	174.50	**1.13**
Dataram Corporation	AMEX	73.10	**0.86**
Energy South, Inc.	OTC	74.00	**1.67**
Gencor Industries, Inc.	AMEX	263.30	**1.96**
Industrial Scientific	OTC	43.50	**2.03**
Keystone Consolodated	**NYSE**	**365.70**	**0.86**

2. **A time series data**
   - collected at regular intervals over time

Month	Number of VCRs sold
January	**123**
February	**130**
March	**125**
April	**138**
May	**145**
June	**142**
July	**141**
August	**146**
September	**147**
October	**157**
November	**150**
December	**160**