Chapter 15 – Handling Time Series
处理时间序列

Instructor: Zach Zhizhong ZHOU,
Shanghai Jiao Tong University
主讲教师：周志中，上海交通大学

Data Mining for Business Intelligence

Shmueli, Patel & Bruce

Main ideas 主要思想

- Forecast future values of a time series 预测一个时间序列的未来值
- Distinction between forecasting (main focus) and describing/explaining 预测（主要目的）和描述/解释的区别。
- Four components of time series: 时间序列的四个要素
  - Level 水平
  - Trend 趋势
  - Seasonality 季节性
  - noise 噪音

Explain vrs. Predict 解释和预测

Explanation is the goal of “time series analysis” 时间序列分析的目标是解释

Models are based on causal argument 模型建立于因果论断基础上。

Models are not “black-box” 模型不是“黑箱”

Forecasting (our focus) seeks to predict future values 预测目标是预测未来值

Time Series Components 时间序列的要素

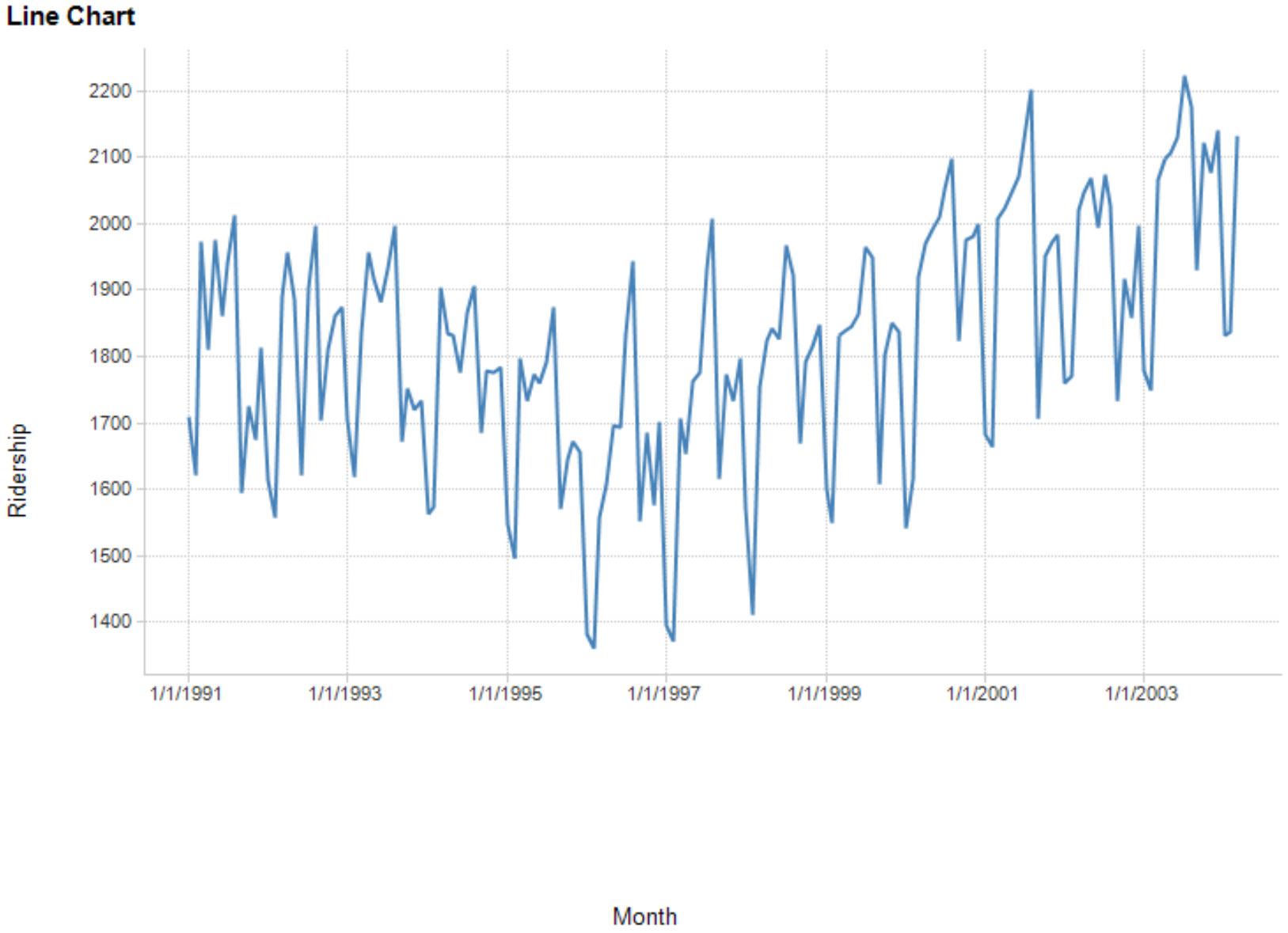
- Level 水平
- Trend 趋势
- Seasonality 季节性
- Noise 噪音

Amtrak Ridership (monthly)

Level - about 1,800,000 passengers per month

Appears to have U-shaped trend

Amtrak Ridership



Zoom to 3 years (1997-1999)

Seasonality* appears: 季节性出现

Summer peaks 夏天乘客达到顶峰

Noise: 噪音

Departure from the general level that is neither trend nor seasonality 偏离通常的水平，但这种偏离不是趋势或者季节性引起的。

*Seasonality is any cyclical pattern. Here it is seasons of the year, but could be any cyclical pattern (daily, weekly, monthly, etc.) 季节性可以是任何循环往复的模式，它可以是一年的季节，也可以是其他任何循环往复模式（如按日、星期、或者月份的循环往复）。

Amtrak Ridership – zoom to 3-years

Partitioning 数据分块

Divide data into training portion and validation portion 将数据分割成训练数据集和验证数据集

Test model on the validation portion 在验证数据集上测试模型

Partitioning is not random 数据分块不是随机的

Random partitioning would leave holes in the data, which causes problems 随机数据分块会导致数据出现空洞，会产生问题。

- Forecasting methods assume regular sequential data 预测方法假设正常的序列数据。

Instead of random selection, divide data into two parts 不对数据进行随机分块，而是将数据分成2部分。

- Train on early data 在早期数据上训练模型
- Validate on later data 在后期数据上验证模型

Summary 总结

Focus is to predict (not describe/explain) 聚焦于预测而不是描述或者解释

Four components 四个要素

- Level 水平
- Trend 趋势
- Seasonality 季节性
- Noise 噪音

Partition data by dividing into early/late 将数据分块成早期数据和晚期数据。