INTRODUCTION

FA590 STATISTICAL LEARNING IN FINANCE

Dr. Zonghao Yang

Stevens Institute of Technology

Last Updated: September 1, 2026

Learning Objectives

- Define statistical learning and differentiate between its main paradigms: supervised, unsupervised, parametric, and non-parametric.
- Explain the unique challenges of applying statistical learning to financial data, such as non-stationarity and low signal-to-noise ratios.
- Outline the five stages of a standard statistical learning workflow: sampling, exploration, modification, modeling, and assessment (SEMMA).
- Select appropriate visualization techniques to conduct exploratory data analysis (EDA) and communicate model insights.

Part I

INTRODUCTION

Statistical Learning Problems: Forecasting Sales

- Goal: Future sales figures for a product or service
- Related features:
  - Historical sales data
  - Marketing expenditure
  - Seasonal factors
- Use cases: Inventory/supply chain management, marketing strategy, product development and launch

Zillow

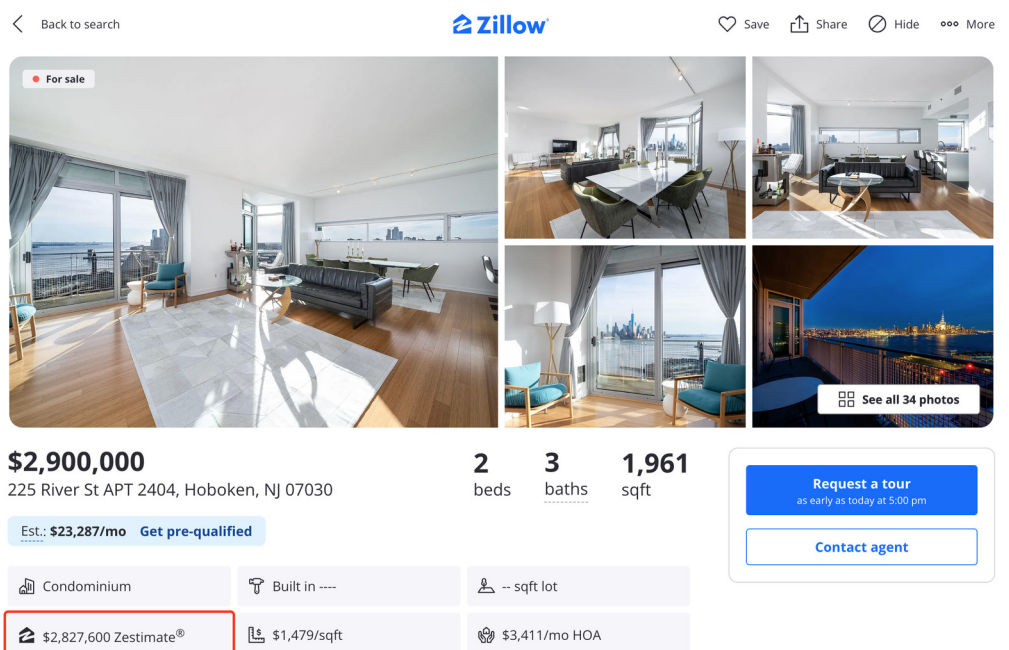


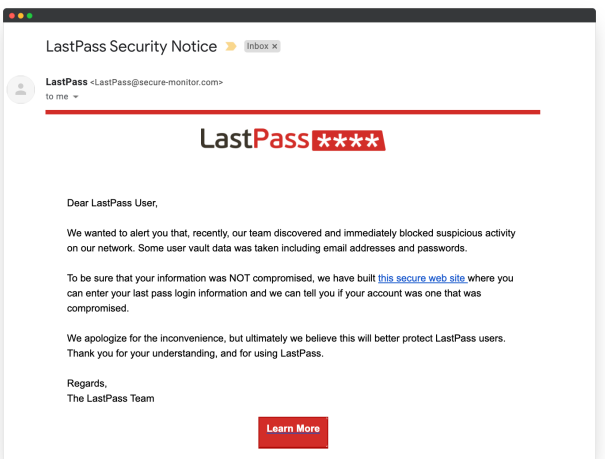
Figure. Listing on Zillow

Predicting Housing Prices

- Goal: The sale price of a house
- Related features:
  - Property characteristics (e.g., square footage, number of bedrooms, lot size)
  - Location factors (e.g., neighborhood, proximity to schools, crime rates)
  - Market conditions (e.g., interest rates, economic indicators)
  - Recent sales of comparable properties
- Use cases: Real estate valuation, mortgage lending, investment analysis, insurance underwriting

Spam Detection

- Goal: Classify emails as either spam or non-spam
- Related features:
  - Sender's email address and domain
  - Presence of specific keywords (e.g., "win", "free", "money", "!")
  - Number of links or URLs within the email
  - Subject line characteristics (e.g., capitalization)
- Use cases: Email filtering and message moderation



Topic Modeling

- Goal: Discover underlying topics in a collection of documents
- Related features:
  - Word frequency counts
  - N-grams
- Use cases: Document summarization, content organization, trend analysis

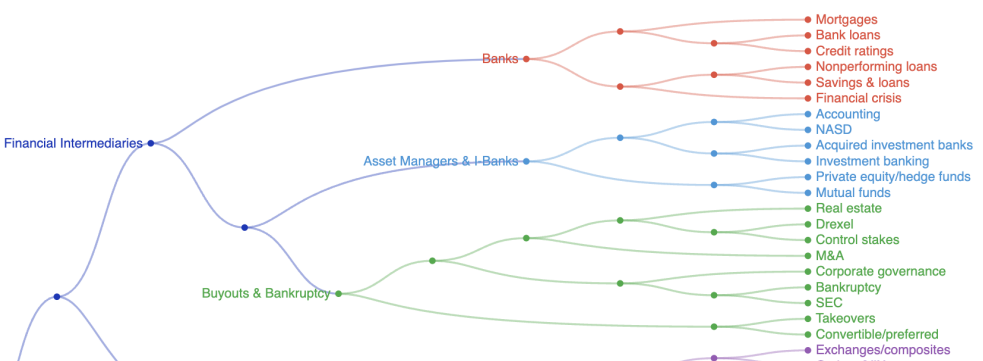


Figure. Topics in *The Wall Street Journal* Articles

Statistical Learning Problems

- Outcome measurement Y
  - Dependent variable, response, target
- Vector of p predictor measurements $X = (X_1, X_2, \ldots, X_p)$
  - Independent variables, inputs, regressors, covariates, features
- For a quantitative response, the relationship between Y and $X = (X_1, X_2, \ldots, X_p)$ can be written as

$$Y = f(X) + \epsilon$$

where $f(\cdot)$ is some fixed but unknown function of $X_1, X_2, \ldots, X_p$, and ϵ is a random error term

Objective of Statistical Learning

$$Y = f(X) + \epsilon$$

- On the basis of observations $(x_1, y_1), \ldots, (x_N, y_N)$, we want to learn the function $f(\cdot)$ so that we can
  - Accurately predict unseen cases
  - Understand which inputs affect the outcomes, and how
  - Assess the quality of our predictions and inferences
- Statistical learning is a fundamental ingredient in the training of a modern data scientist

Types of Statistical Learning Problems

- Supervised learning : Labeled outcome variable (Y)
  - Regression : Y is quantitative (e.g., sales)
  - Classification : Y takes values in a finite, unordered set (e.g., spam/ham, digits 0–9)

- Unsupervised learning
  - No outcome variable (Y), just a set of predictors (X)
  - The objective is less well defined: find groups of samples that behave similarly, find features that behave similarly, or find linear combinations of features with the most variation
  - Difficult to know how well you are doing
  - Unsupervised learning can be useful as a pre-processing step for supervised learning

Revisit the Examples: Forecasting Sales

- Response variable (Y): Future sales figures for a product or service
- Input features (X):
  - Historical sales data
  - Marketing expenditure
  - Seasonal factors

Revisit the Examples: Forecasting Sales

- Response variable (Y): Future sales figures for a product or service
- Input features (X):
  - Historical sales data
  - Marketing expenditure
  - Seasonal factors
- **Supervised learning, regression**

Predicting Housing Prices

- Response variable (Y): The sale price of a house
- Input features (X):
  - Property characteristics (e.g., square footage, number of bedrooms, lot size)
  - Location factors (e.g., neighborhood, proximity to schools, crime rates)
  - Market conditions (e.g., interest rates, economic indicators)
  - Recent sales of comparable properties

Predicting Housing Prices

- Response variable (Y): The sale price of a house
- Input features (X):
  - Property characteristics (e.g., square footage, number of bedrooms, lot size)
  - Location factors (e.g., neighborhood, proximity to schools, crime rates)
  - Market conditions (e.g., interest rates, economic indicators)
  - Recent sales of comparable properties
- **Supervised learning, regression**

Spam Detection

- Response variable (Y): Whether the email is spam or non-spam
- Input features (X):
  - Sender's email address and domain
  - Presence of specific keywords (e.g., "win", "free", "money", "!")
  - Number of links or URLs within the email
  - Subject line characteristics (e.g., capitalization)

Spam Detection

- Response variable (Y): Whether the email is spam or non-spam
- Input features (X):
  - Sender's email address and domain
  - Presence of specific keywords (e.g., "win", "free", "money", "!")
  - Number of links or URLs within the email
  - Subject line characteristics (e.g., capitalization)

- **Supervised learning, classification**

Topic Modeling

- Response variable (Y): None
- Input features (X):
  - Word frequency counts
  - N-grams

Topic Modeling

- Response variable (Y): None
- Input features (X):
  - Word frequency counts
  - N-grams

- **Unsupervised learning**

Statistical Learning Problems in Finance: Credit Card

Scenario: You are designing a security system for Visa. You need to flag transactions that look suspicious in real time, such as a card being used in two different countries within an hour.

- Formulate the above scenario as a statistical learning problem

Statistical Learning Problems in Finance: Loan Officer

Scenario: A bank wants to decide whether to approve a mortgage application. They need to know if the borrower is likely to fail to pay back the money.

- Formulate the above scenario as a statistical learning problem

Statistical Learning Problems in Finance: Quant Trader

Scenario: A hedge fund wants to build an algorithm to trade stocks based on technical indicators and news sentiment. They want to predict the return of a stock for the next day.

- Formulate the above scenario as a statistical learning problem

Statistical Learning Algorithms

$$Y = f(X) + \epsilon$$

- Linear regression: Regression

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

- Logistic regression: Classification

$$\hat{Y} = \begin{cases} 1 & \text{if } p(X) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

$$p(X) := Pr(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}}$$

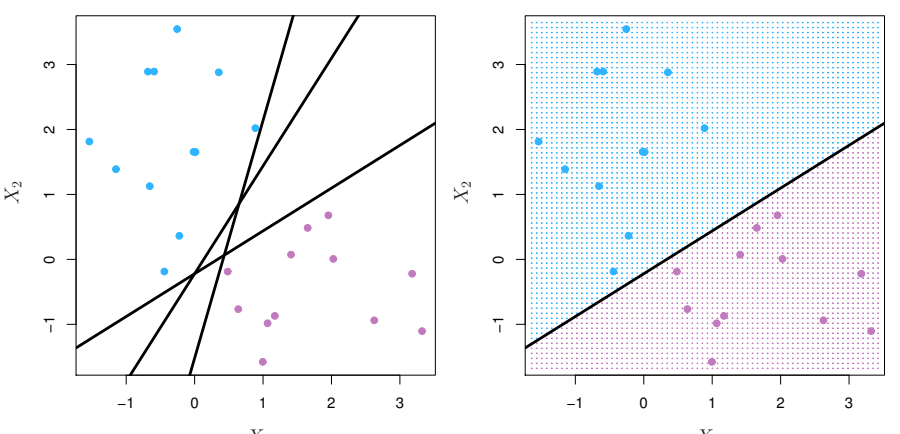
- Naïve Bayesian Classifier: Classification

$$f(X) = \underset{k \in \{1,2,\ldots,K\}}{\arg\max} \; P(Y = k) \prod_{j=1}^{p} P(X_j | Y = k)$$

Statistical Learning Algorithms: Support Vector Machines

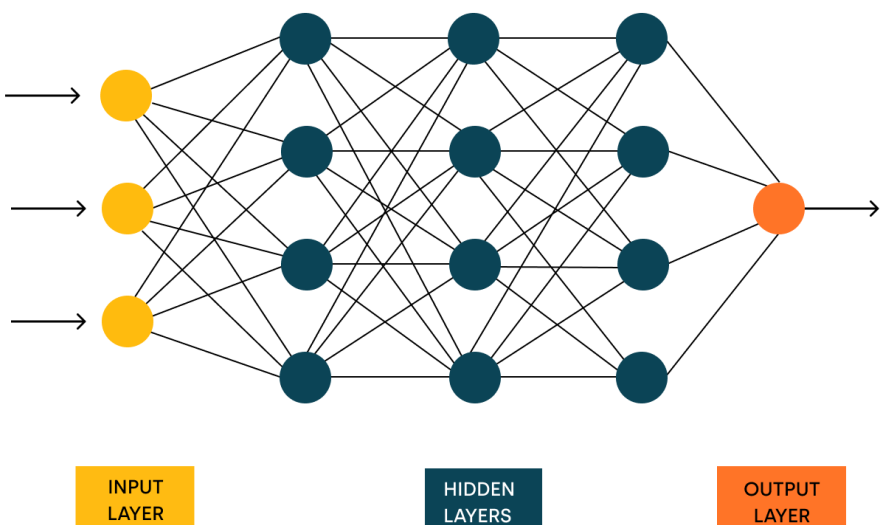
Support vector machines (SVM) find the hyperplane that “best" separates the positive and negative classes:

$$f(X) = \text{sign}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$$



Statistical Learning Algorithms: Neural Networks

Neural networks can be used for both regression and classification

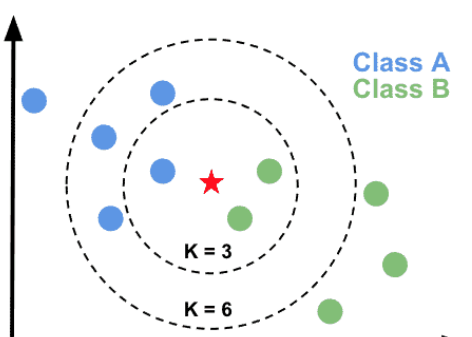


Parametric and Non-Parametric Methods

- Methods can be classified by the purpose of their prediction, **regression** or **classification**
- Methods can also be classified as **parametric** or **non-parametric** depending on whether a specific functional form for $f(X)$ is specified
  - Parametric: Assume a fixed form for the function $f(X)$ that maps inputs to outputs
  - Non-parametric: Do not assume a specific form for $f(X)$; the model structure is determined based on the data
- Examples of non-parametric methods: k-Nearest Neighbors (k-NN), tree-based methods

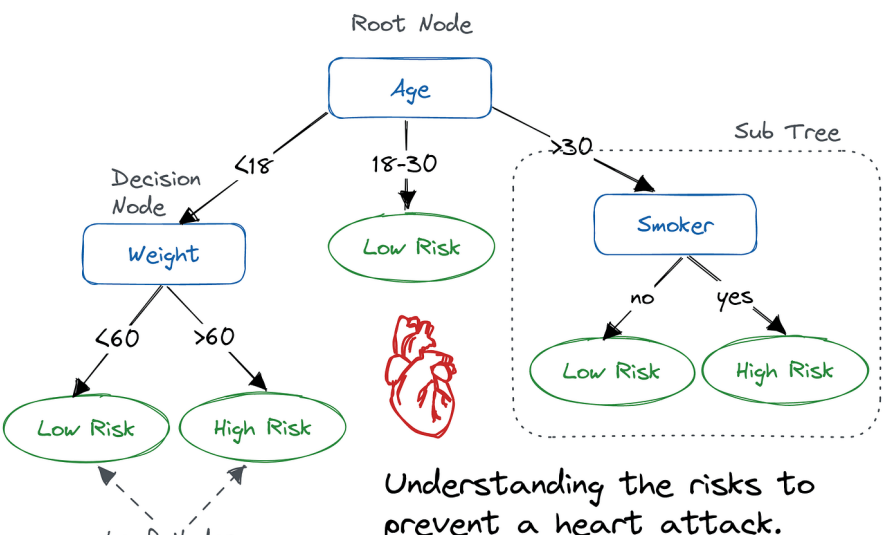
Statistical Learning Algorithms: k-Nearest Neighbors

- Classification: majority label among the k nearest neighbors. Regression: average of the neighbors' values
- k is the number of nearest neighbors
- k-NN algorithm can be used for both regression and classification



Statistical Learning Algorithms: Tree-Based Methods

Tree-based methods can be used for both regression and classification



Parametric vs. Non-Parametric Methods

Method	Advantage	Disadvantage
Parametric	1. Generally more efficient with a smaller sample size 2. Easier to interpret	Assumes a specific form for $f(X)$, which may be wrong
Non-parametric	Does not assume a specific form for $f(X)$	1. Requires more data and computational resources 2. Some methods (e.g., k-NN) must store all the training data

Table. Parametric vs. Non-Parametric Methods

Why Are There So Many Different Methods?

- Each method has advantages and disadvantages
- The usefulness of a method can depend on factors such as
  - the size of the dataset, the types of patterns that exist in the data
  - whether the data meet some underlying assumptions of the method
  - how noisy the data are
  - the particular goal of the analysis
- **Experience is needed to decide which method to use**

Statistical Learning, Machine Learning, and Artificial Intelligence

- **Machine learning** arose as a subfield of **artificial intelligence**
- **Statistical learning** arose as a subfield of **statistics**
- Machine learning has a greater emphasis on large-scale applications and *prediction accuracy*
- Statistical learning emphasizes *interpretability*, *precision*, and *uncertainty*
- The distinction has become more and more blurred, and there is a great deal of "cross-fertilization"

- In this course, we will not distinguish statistical learning, ML, and AI
- These are just terms; the concepts are what matter

Part II

Statistical Learning in Finance

Supervised Learning and Finance

- Asset pricing: Predict expected asset returns, i.e., $\mathbb{E}_t[r_{t+1}]$
- High dimensionality
  - Firm characteristics from accounting data
  - Signals extracted from textual information in corporate disclosures
  - Variables summarizing the history of price and trading volume
- Statistical learning methods are well suited for dealing with high dimensionality, but straight off-the-shelf application of these methods is unlikely to fully exploit their potential

Typical Statistical Learning and Asset Pricing Applications

	Typical ML Application	Asset Pricing
Signal-to-noise	High	Very low
Data dimensions	Many predictors, Many observations	Many predictors Few observations
Aggregation level of interest	Individual outcomes	Portfolio outcomes
Prediction error covariances	Statistical nuisance	Important determinant of portfolio risk
Sparsity	Often sparse	Unclear
Structural change	None	Investors learn from data and adapt

Figure. Typical Statistical Learning vs. Asset Pricing Applications[1]

[1] Nagel, S. (2021). Machine learning in asset pricing (Vol. 1). Princeton University Press.

Signal-to-Noise Ratio

- Compare the following prediction targets
  - Energy consumption of a household, blood pressure of a patient, a dog in an image
  - Daily return of a stock r_{t+1}
- The return has a much lower signal-to-noise ratio
  - The expected return is unobservable, and potentially time-varying
  - The realized return is a noisy signal of the expected return

$$r_{t+1} \neq \mathbb{E}_t[r_{t+1}]$$

- In contrast, you observe the true label in the training set in typical statistical learning applications
- The low signal-to-noise ratio problem is further compounded by the limitations of the data available for training

Data Dimensions

- Usable stock-level return panels cover only several decades
- $\sim 60,000$ trading days on the New York Stock Exchange (NYSE) since May 17, 1792
- $\sim 14,000$ trading days on Nasdaq since February 8, 1971
- The ImageNet dataset contains over 14 million labeled images spread across more than 20,000 categories[2]

- Can we use high-frequency data to increase the number of observations?
  - Merton (1980)[3]: Over a fixed calendar span, higher sampling frequency does not yield more precise estimates of a constant mean return $\mathbb{E}[r]$

[2] ImageNet: https://www.image-net.org/

[3] Merton, R. C. (1980). On estimating the expected return on the market: An exploratory investigation. Journal of Financial Economics, 8(4), 323-361.

Aggregation Level of Interest

- The portfolio perspective
  - Not necessarily interested in obtaining accurate predictions of individual stocks' returns
  - But rather in constructing a portfolio with good risk-return properties

- **Question:** Do methods that deliver the most accurate return forecasts at the individual stock level also give us the best-performing portfolio once we aggregate across stocks?

Aggregation Level of Interest

- The portfolio perspective
  - Not necessarily interested in obtaining accurate predictions of individual stocks' returns
  - But rather in constructing a portfolio with good risk-return properties

- **Question:** Do methods that deliver the most accurate return forecasts at the individual stock level also give us the best-performing portfolio once we aggregate across stocks?
- Methods that yield better predictions of individual security returns, in the sense of a higher predictive R^2, do not necessarily yield better portfolios in terms of standard metrics such as the portfolio's Sharpe ratio

Portfolio Perspective

Example: Predict returns for two stocks A and B

- Actual returns: r_A and r_B
- Predicted returns: $\hat{r}_A$ and $\hat{r}_B$
- Errors: $\epsilon_A = r_A - \hat{r}_A$ and $\epsilon_B = r_B - \hat{r}_B$

Scenario	**Zero Covariance**	**Positive Covariance**
Covariance	$\text{Cov}(\epsilon_A, \epsilon_B) = 0$	$\text{Cov}(\epsilon_A, \epsilon_B) > 0$
Error Relationship	Errors in predicting r_A and r_B are uncorrelated	If ϵ_A is positive (underestimate of r_A), ϵ_B is likely to be positive too
Implication	Wrong prediction for one stock does not imply wrong prediction for the other	Indicates a common risk factor or systematic bias not accounted for

The portfolio perspective is important in stock return prediction. Ignoring covariance can lead to underestimating risk.

Prediction Error Covariances

- In a setting of many assets, covariances of return prediction errors determine much of the portfolio volatility
- The portfolio volatility plays a big role in determining the risk-return properties of the portfolio

Properties of error covariances could matter at all stages of an SL application in asset pricing, starting with the question of what the optimal algorithm is, how to regularize it, how to evaluate predictive performance, and how to use its output for portfolio construction.

Sparsity

- Sparsity is often high in typical SL applications
  - Image classification: Some regions of an image may be irrelevant for the classification task
  - Genomics: A large number of candidate genes may have no relationship to the disease of interest

- Sparsity is unclear for asset pricing applications
  - Fama-French five-factor model: Market risk, size, value, profitability, and investment
  - The factor zoo: All variables are relevant, and some are more relevant

Structural Change

The data-generating process in financial markets is likely undergoing continuous structural change

- The economy overall is undergoing structural change
- Investors learn from data, and asset prices are the outcome of investment decisions (by humans and machines)
  - Past evidence of return predictability may have induced them to alter their trading strategies in a way that destroyed earlier patterns of return predictability.
  - Predictability that shows up in historical data therefore might not repeat in the same way in future data
- How is this different from other ML applications?
  - Hot dogs are not going to change shape or color in response to the analysis of hot dog image data

Why Statistical Learning in Finance

Many exciting open questions for Statistical Learning in Finance

	Typical ML Application	Asset Pricing
Signal-to-noise	High	Very low
Data dimensions	Many predictors, Many observations	Many predictors Few observations
Aggregation level of interest	Individual outcomes	Portfolio outcomes
Prediction error covariances	Statistical nuisance	Important determinant of portfolio risk
Sparsity	Often sparse	Unclear
Structural change	None	Investors learn from data and adapt

Figure. Typical Statistical Learning vs. Asset Pricing Applications[4]

[4] Nagel, S. (2021). Machine learning in asset pricing (Vol. 1). Princeton University Press.

Part III

Statistical Learning Overview

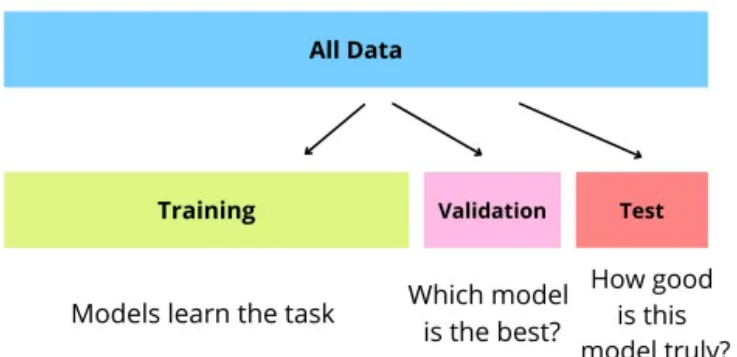
The Steps in Statistical Learning

- Typical steps during statistical learning are abbreviated as **SEMMA**
  1. **S**ample: Take a sample from the dataset; partition into training, validation, and test datasets
  2. **E**xplore: Examine the dataset statistically and graphically
  3. **M**odify: Transform the variables and impute missing values
  4. **M**odel: Fit predictive models (e.g., regression tree, neural network)
  5. **A**ssess: Compare models using a validation dataset

- In this course, we will focus mostly on Steps 4 and 5, since these are general tasks
- In practice, Steps 1–3 are often the most cumbersome, and typically application-specific (they require domain expertise, etc.)

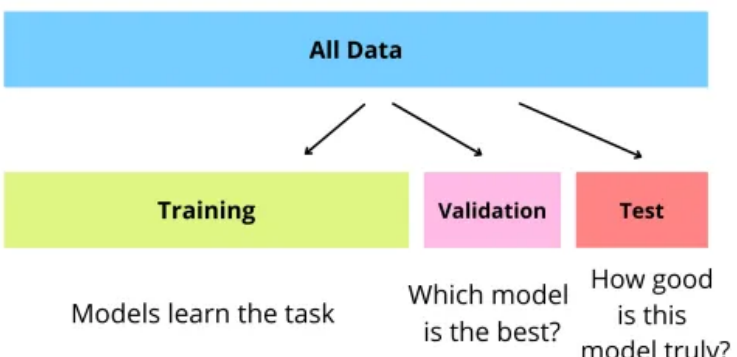
SEMMA: SAMPLE

- When doing statistical learning, we always split the data into a train-test or train-validation-test set
- Training and validation data are in-sample ; test data are out-of-sample
- **Question: Why sample the data?**



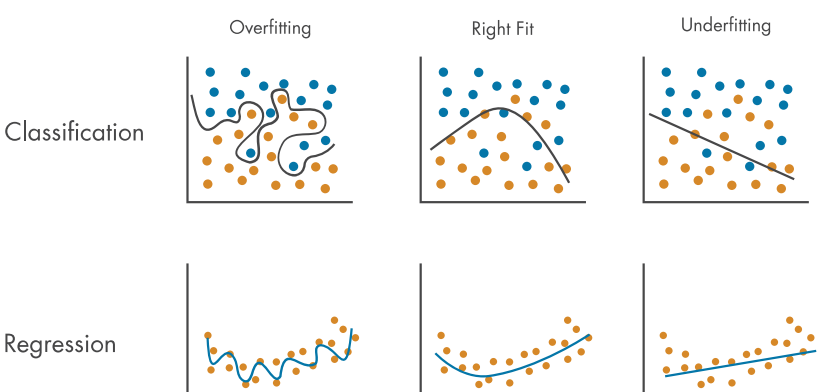
SEMMA: SAMPLE

- When doing statistical learning, we always split the data into a train-test or train-validation-test set
- Training and validation data are in-sample ; test data are out-of-sample
- **Question: Why sample the data?**
- Hold-out samples are crucial for avoiding **overfitting**



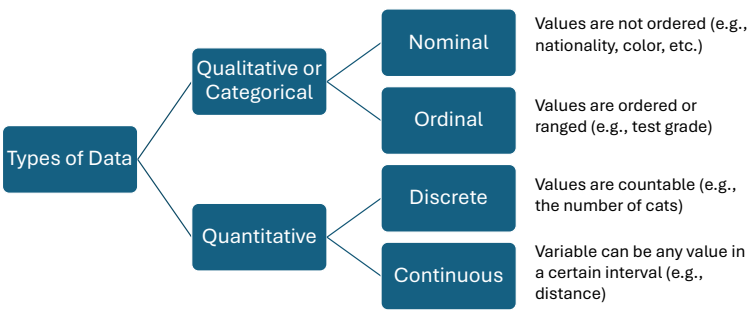
Overfitting and Underfitting

- **Overfitting** The model learns too much from the training data, including noise, making it perform well on the training set but poorly on new data
- **Underfitting** The model is too simple to capture the underlying patterns in the data, resulting in poor performance on both the training and new data
- Question: The accuracy is 99% in training and 60% in validation. Is this scenario overfitting, a good fit, or underfitting?



SEMMA: EXPLORE

- Four types of data

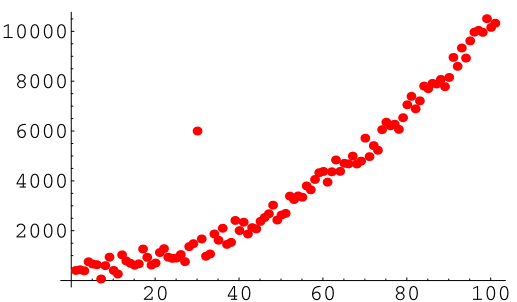


- Another classification: Structured and unstructured data
  - Structured data: Tabular data
  - Unstructured data: Text, images, audio
- **Always visualize data first!**

SEMMA: MODIFY

Outliers Values that lie “far away” from the bulk of the data

- The term *far away* is deliberately left vague because what is or is not called an outlier is problem-specific
- Some outliers are due to erroneous data collection (e.g., measurement error)
- Some outliers are intrinsic to the system and should not be ignored
  - Example: High sales during holiday seasons, traffic spikes on a website, high-value transactions in financial data



SEMMA: MODIFY

Missing Values The absence of data or a lack of recorded information in a dataset

- If the number of data points with missing values is small, those data points might be omitted
- If the number of missing entries is large, you need to consider removing the feature
- Can replace the missing value with an imputed value, based on the other values for that variable across all data points
  - Mean/median
  - Maximum/minimum
  - 0
- Consider the following scenario: The total area of a car garage is missing because the corresponding listing does not have a garage.

SEMMA: MODIFY

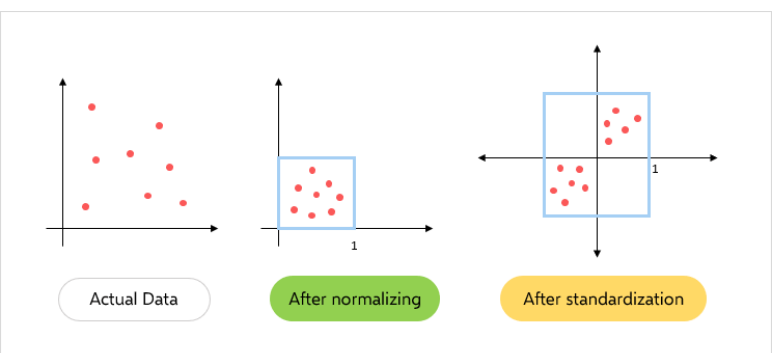
- Standardization

$$\tilde{X} = \frac{X - \mu}{\sigma}$$

- Normalization

$$\tilde{X} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

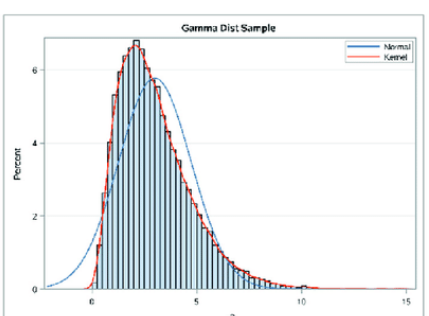
- Both methods bring all values to the same scale



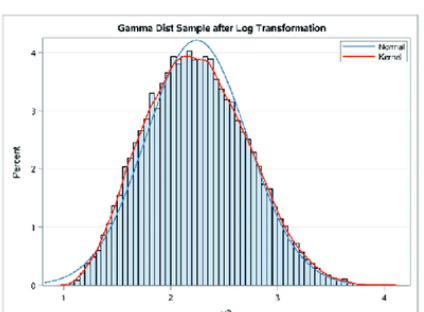
SEMMA: MODIFY

- Some data covers various orders of magnitude (e.g., wealth, salary)
- Standardization or normalization can be misleading and will be skewed toward the larger values
- Often, one first applies a log-transform (and then standardizes)

$$\tilde{X} = \log(X), \quad X > 0$$



(a)



(b)

SEMMA: Model

$$y = f(X) + \epsilon$$

- **Modeling** is choosing the specification for $f(X)$
- Supervised and unsupervised
- Regression and classification
- Parametric and non-parametric
- Learning different model specifications is the main focus of this course

SEMM**A**: Assess

- How well will our prediction model perform when we apply it to new data?
- Regression: For example, mean squared error (MSE)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

- Classification: For example, accuracy, i.e., the percentage of predictions that are correct
- Assessment should consider both in-sample and out-of-sample data
- We will learn various assessment metrics; the selection of metrics depends on the specific context

SEMMA

1. **S**ample: Take a sample from the dataset; partition into training, validation, and test datasets
2. **E**xplore: Examine the dataset statistically and graphically
3. **M**odify: Transform the variables and impute missing values
4. **M**odel: Fit predictive models (e.g., regression tree, neural network)
5. **A**ssess: Compare models using a validation dataset

Part IV

GRAPHICS

Anscombe's Quartet

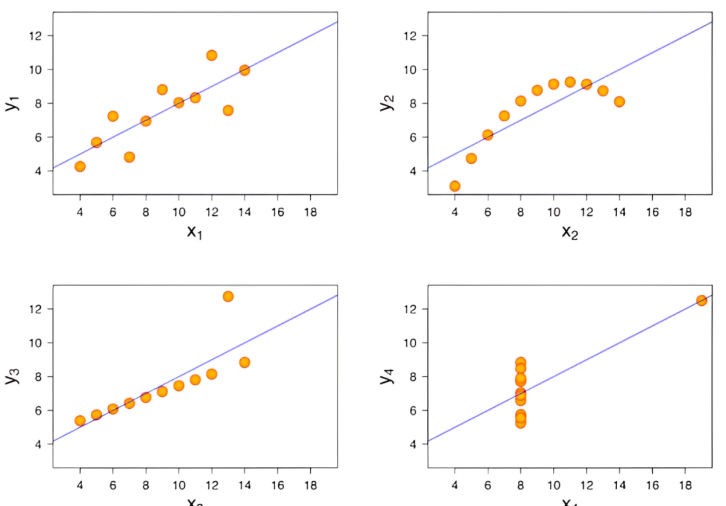
- An illustration of the importance of understanding the distributional properties of data
- Constructed by Francis Anscombe in 1973[5]
- Four sets of data, with very similar descriptive statistics
  - $\bar{x} = 9$, $s_x^2 = 11$
  - $\bar{y} = 7.50$, $s_y^2 = 4.125$
  - $\rho_{x,y} = 0.816$
  - Regression line $y = 3 + 0.5x$

[5] Anscombe, F. J. (1973). Graphs in statistical analysis. The American Statistician, 27(1), 17-21.

Anscombe's Quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Anscombe's Quartet



Takeaway: Look at your data before doing any analysis!

WHY GRAPHICS?

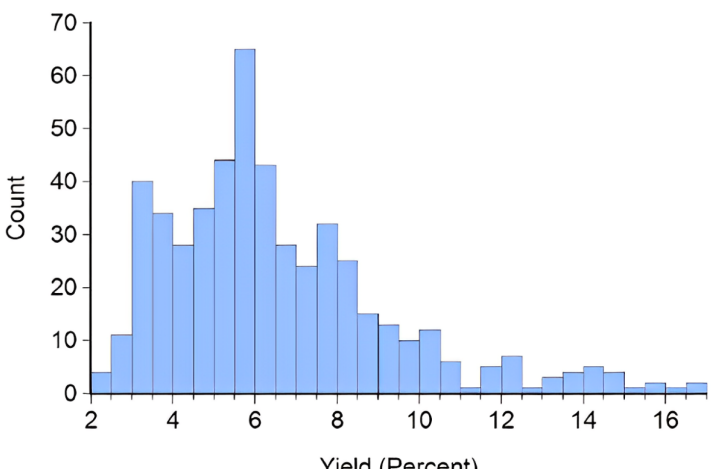
- Graphical analysis helps us summarize and reveal patterns in data
  - Identify anomalies (or outliers) in data
  - Facilitate and encourage comparison of different pieces of data
  - Enable us to present a large amount of data in a small space
  - Make a huge data set coherent
- Simple graphical techniques
  - Univariate, multivariate
  - Time series vs. distributional shape
  - Relational graphics

GRAPHICS

- Univariate data
  - Histograms and density estimates can help us learn about distributional shape: symmetric, skewed, fat-tailed, ...
  - Time series plots reveal dynamics such as trend, seasonality, cycle, outliers, ...
- Multivariate data
  - Scatterplots for relations: Does a relation exist? Is it linear or non-linear? Are there outliers?

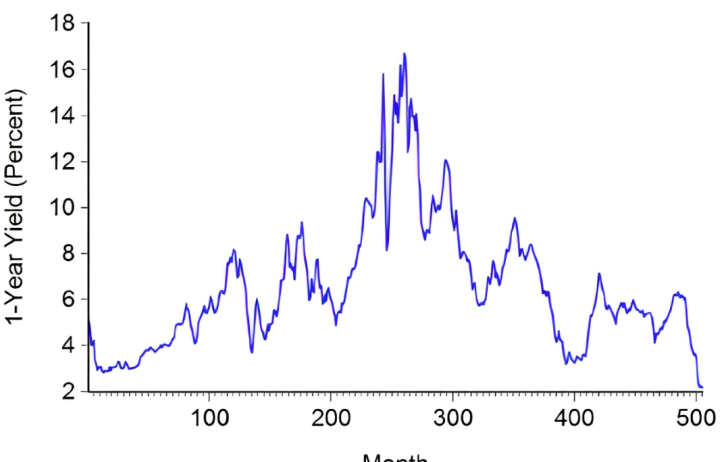
HISTOGRAM

- Histogram reveals distributional shape
- Example: One-year government bond yield



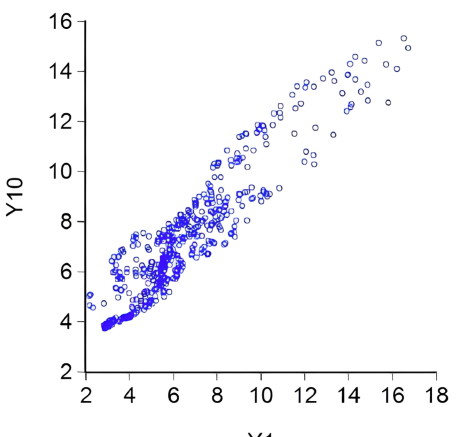
Time Series Plot

- Time series plot reveals dynamics
- Example: One-year government bond yield



SCATTERPLOT

- Scatterplot reveals the relation between two variables
- Example: One-year and 10-year government bond yields

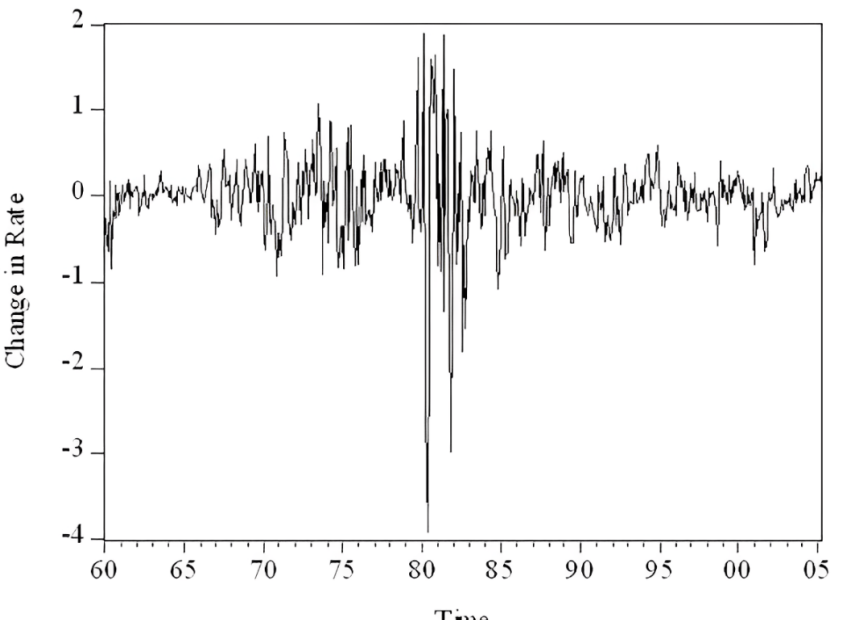


Principles of Graphical Style

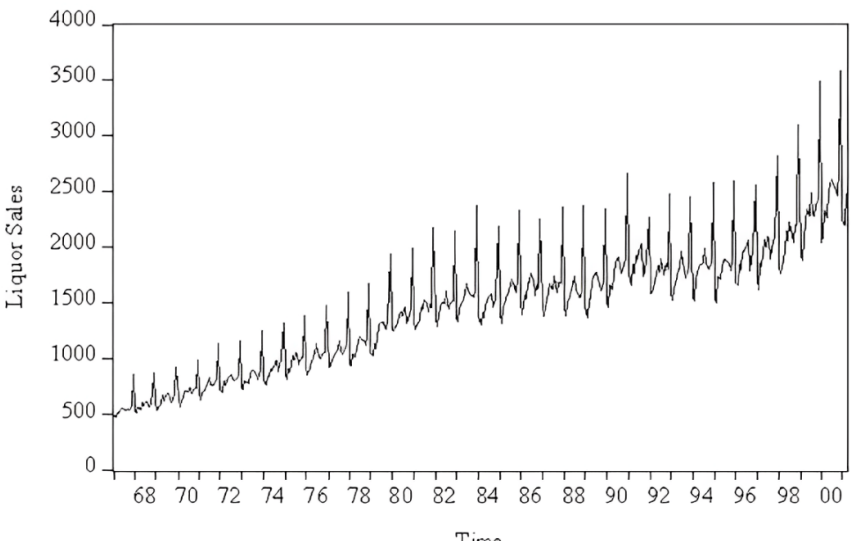
Primary purpose: Summarize and reveal patterns in data

- Know your goals
- Know your audience; make the presentation appealing to the audience
- Show the data within the bounds of reason
  - Use common scales and avoid distortion
  - Minimize non-data text and labels to make graphics look cleaner
  - Maximize graphical data density
- Revise and edit, again and again; graphics using software defaults are seldom satisfactory

More Graphics



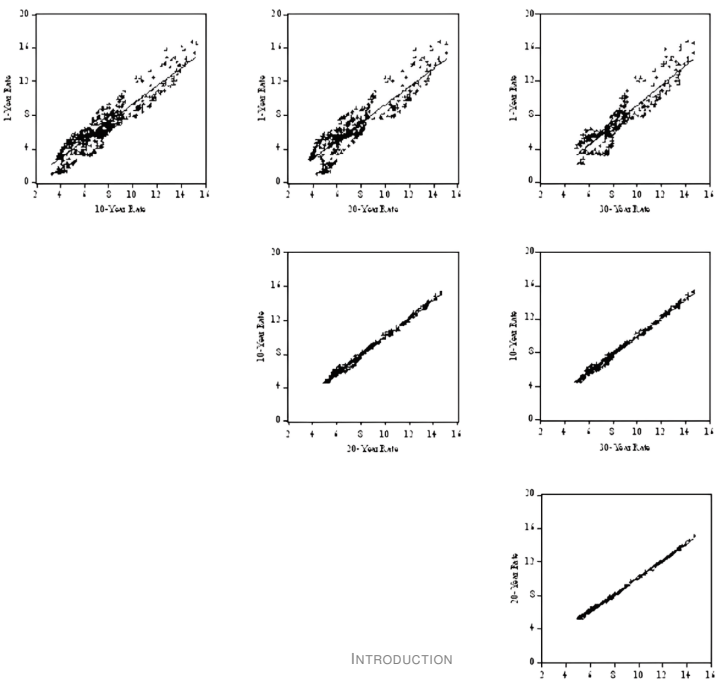
More Graphics



What are the spikes?

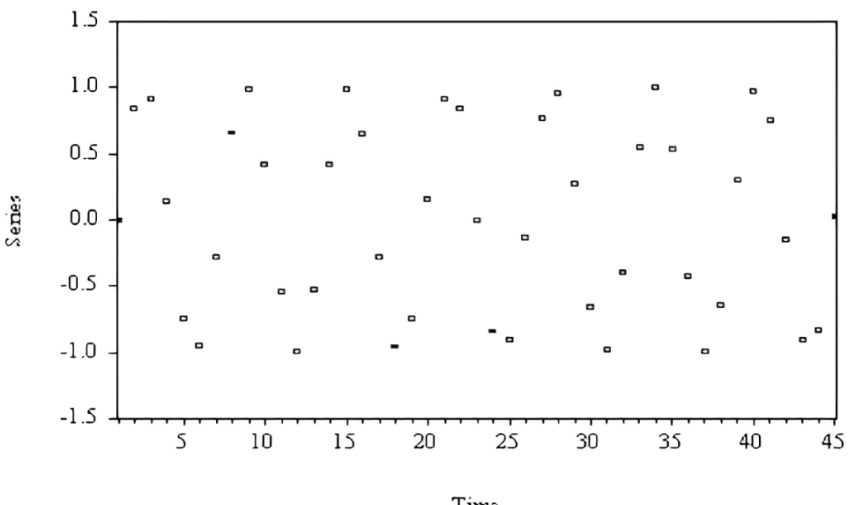
More Graphics

Scatterplot matrix: 1-, 10-, 20-, and 30-year Treasury Bond rates



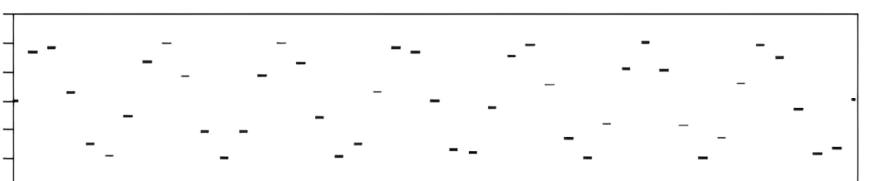
More Graphics

What's the pattern in this plot? 1:1.6 aspect ratio



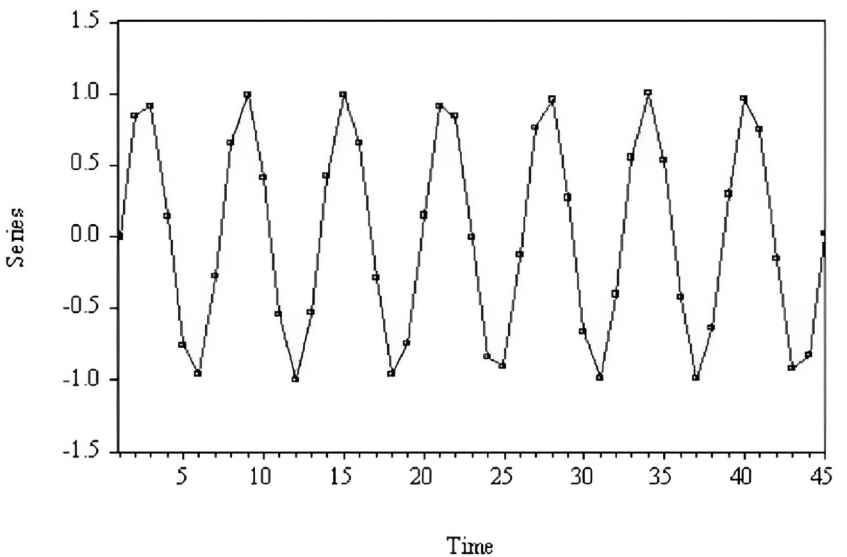
More Graphics

What about now? (Same data, banked to 45 degrees)



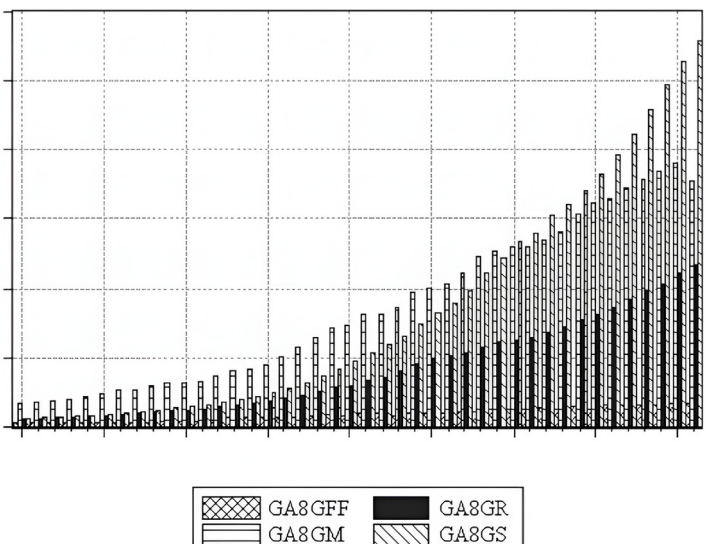
More Graphics

Back to 1:1.6 aspect ratio. Now what is the pattern?



GRAPHICS: A CASE STUDY

What's wrong with this picture?



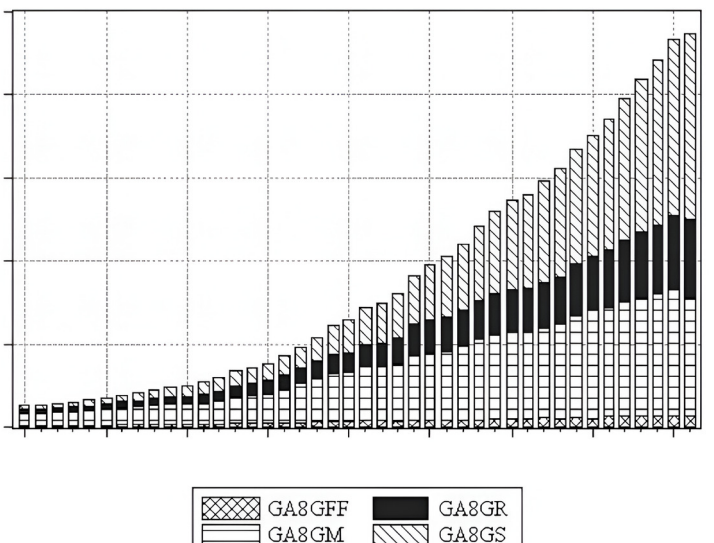
GRAPHICS: A CASE STUDY

What do you learn?

- Data $\neq$ knowledge
- Data $\neq$ insight
- The plot contains lots of data, but no insight
- You can have data without knowledge

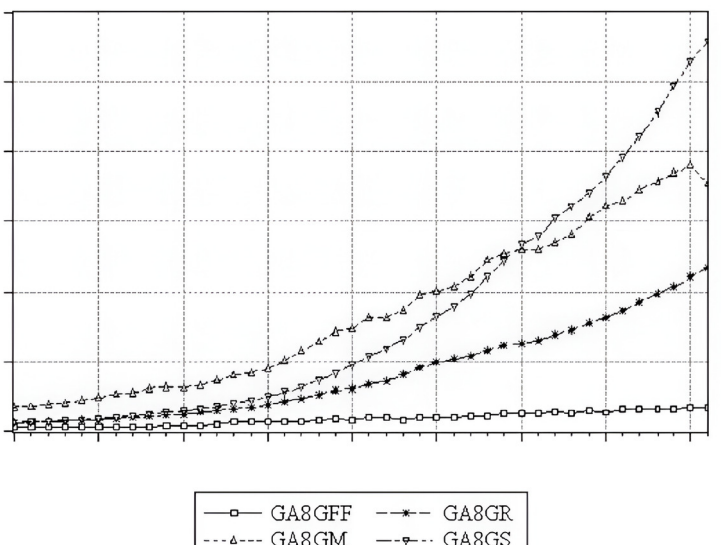
GRAPHICS: A CASE STUDY

Is this better or worse than before?

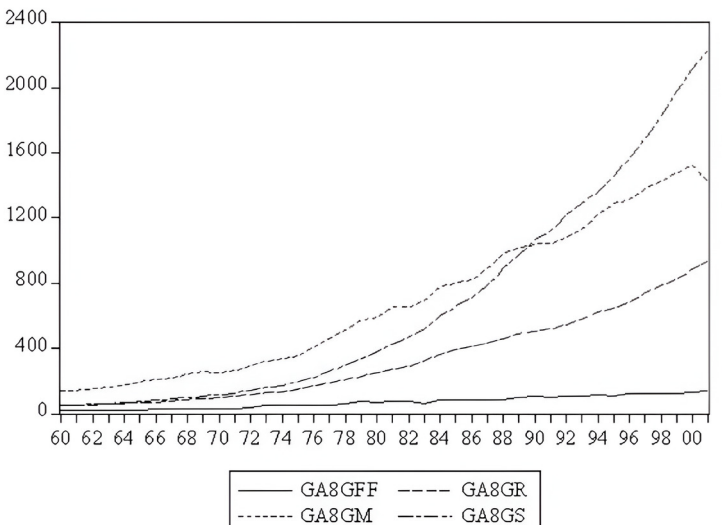


GRAPHICS: A CASE STUDY

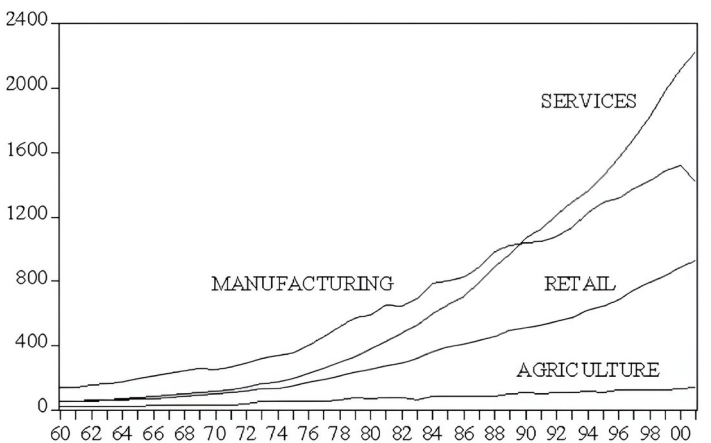
How about now?



GRAPHICS: A CASE STUDY



GRAPHICS: A CASE STUDY



GRAPHICS: A CASE STUDY

Figure. Components of Nominal GDP (Billions of Current Dollars, Annual)