Chapter 2

Data Structures: Classifying the Various Types of Data Sets

Data Structure Elements

- Data Set:
  - Measurements of items
    - e.g., Yearly sales volume for your 23 salespeople
    - e.g., Cost and number produced, daily, for the past month
- Elementary Units:
  - The items being measured
    - e.g., Salespeople, Days, Companies, Catalogs, …
- A Variable:
  - The type of measurement being done
    - e.g., Sales volume, Cost, Productivity, Number of defects, …

How Many Variables?

- **Univariate data set**: One variable measured for each elementary unit
  - e.g., **Sales** for the top 30 computer companies.
  - Can do: Typical summary, diversity, special features
- **Bivariate data set**: Two variables
  - e.g., **Sales** and **# Employees** for top 30 computer firms
  - Can also do: relationship, prediction
- **Multivariate data set**: Three or more variables
  - e.g., **Sales**, **# Employees**, **Inventories**, **Profits**, …
  - Can also do: predict one from all other variables

Numbers or Categories?

- **Quantitative Variable**: Meaningful numbers
  - e.g., **Sales**, **# Employees**
  - Can add, rank, count
- **Qualitative Variable**: Categories
  - Ordinal Variable: Categories with meaningful ordering
    - e.g., **Bond rating** (AA, A, B, …), **Diamonds** (VSI, SI, …)
    - Can rank, count
  - Nominal Variable: categories without meaningful ordering
    - e.g., **State**, **Type of business**, **Field of study**
    - Can count

Time-Series or Cross-Sectional?

- **Time-Series Data**: Data values recorded in meaningful sequence
  - Elementary units might be **days** or **quarters** or **years**
  - e.g., Daily Dow-Jones stock market average close for the past 90 days
  - e.g., Your firm's quarterly sales over the past 5 years
- **Cross-Sectional Data**: No meaningful sequence
  - e.g., Sales of 30 companies
  - e.g., Productivity of each sales division
  - Easier than time series!

Example

Firm	Sales	Industry Group	S&P Rating
IBM	66,346	Office Equipment	A
Exxon	59,023	Fuel	A-
GE	40,482	Conglomerates	A+
AT&T	34,357	Telecommunications	A-

Cross-Sectional

Example (continued)

Multivariate Data (3 variables)

Firm	Sales	Industry Group	S&P Rating
IBM	66,346	Office Equipment	A
Exxon	59,023	Fuel	A-
GE	40,482	Conglomerates	A+
AT&T	34,357	Telecommunications	A-

Elementary units

Quantitative variable

Nominal Qualitative variable

Ordinal Qualitative variable

Example

Year	Unemployment Rate
2005	4.9%
2006	4.4%
2007	5.0%
2008	7.3%
2009	9.9%
2010	9.3%
2011	8.5%
2012	7.9%
2013	6.7%
2014	5.6%

Time series

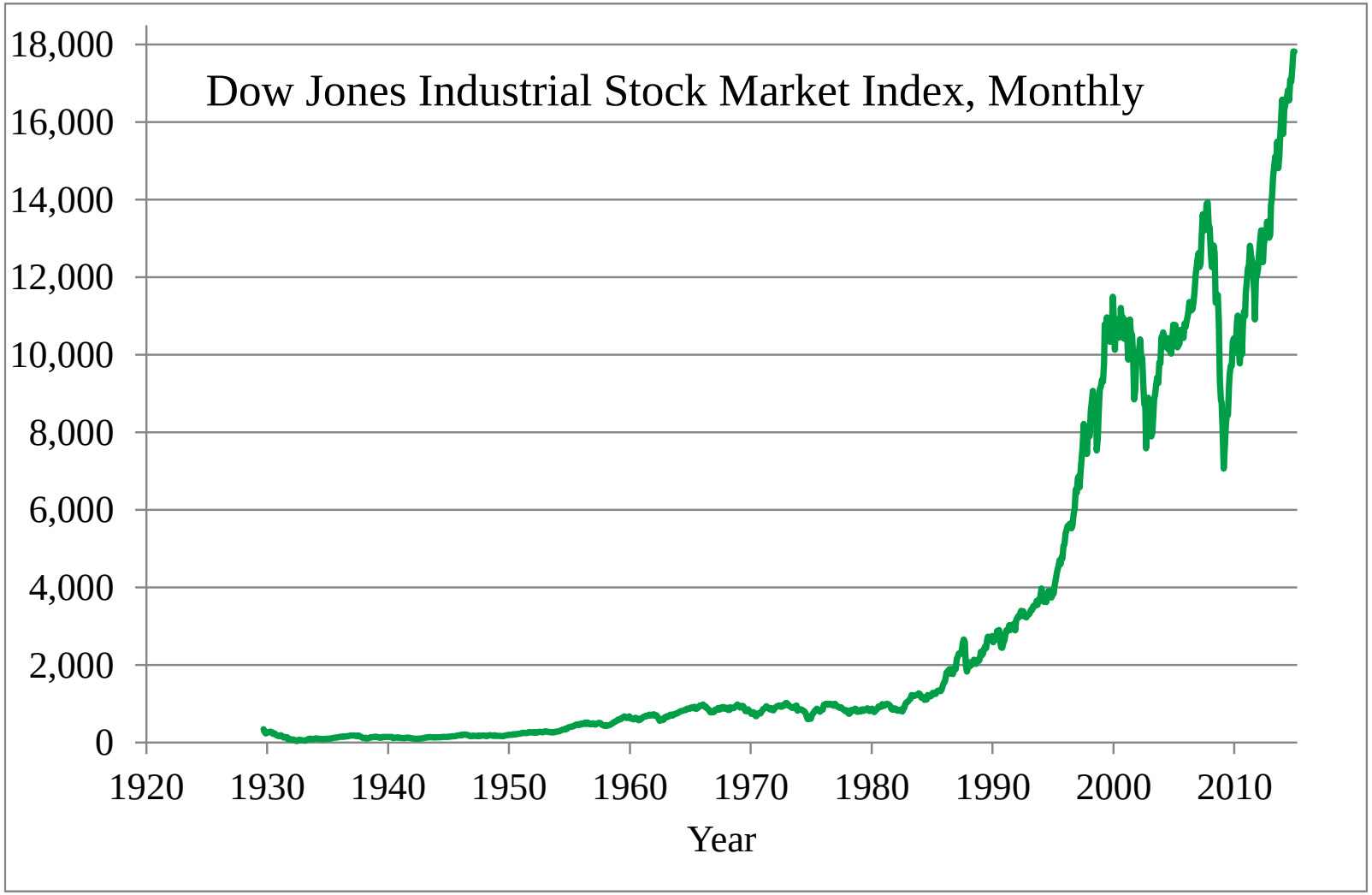
Example

Year	Unemployment Rate
2005	4.9%
2006	4.4%
2007	5.0%
2008	7.3%
2009	9.9%
2010	9.3%
2011	8.5%
2012	7.9%
2013	6.7%
2014	5.6%

Elementary unit defined by "year"

Quantitative data

Stock Market – Time Series



Sources of Data

- **Primary Data**
  - When you control the design and data collection
    - Production data from your factory
    - Your firm's marketing studies
- **Secondary Data**
  - When you use data previously collected by others for their own purposes
    - Government data: economics and demographics
    - Media reports – TV, newspapers, Internet
    - Companies that specialize in gathering data

Observational Study and Experiment

- **Observational Data**
  - Measurements are observed as they occur naturally as part of a system
    - Production data from your factory under ordinary circumstances
    - Characteristics of your customers
- **Experiment**
  - Involves manipulation to control some characteristic(s) of the system so that we can assess **causation**.
    - A/B Testing: Show one of two ads chosen at random, then observe the result
    - Two random groups for testing medicine effectiveness: Treatment group and control group