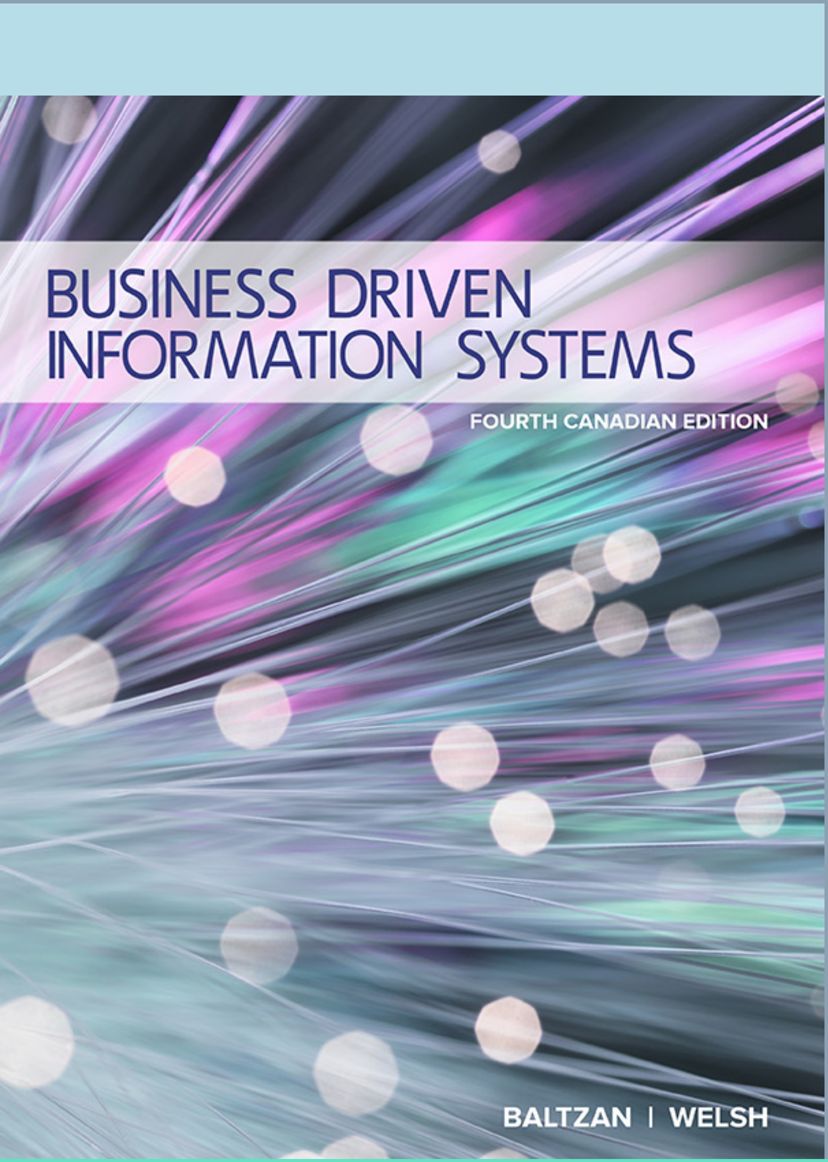
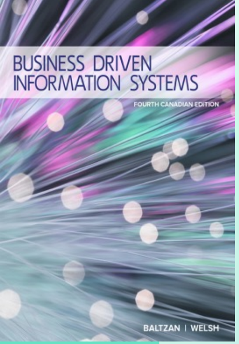


Data Warehouses and Data Mining

(with slides modification and additions by O. Appel)

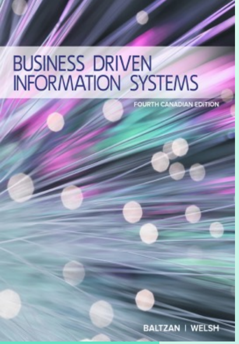
McGraw-Hill-Ryerson



Motivation (O. Appel)

- **What if we have several sources of data?**
  - Accessing Organizational Information
  - Extraction, Transformation & Loading (ETL)
  - Extraction, Loading and Transformation (ELT)
  - Data Warehouse

- **What if we want to uncover intelligence from the data?**
  - Data Mining
  - Data Analytics Techniques
  - Statistical Analysis
  - Simple Prediction



Learning Outcome

4.3

Database Management Systems (DBMS)

Software through which users and application programs interact with a database

Interacting Directly and Indirectly with a Database Through a DBMS

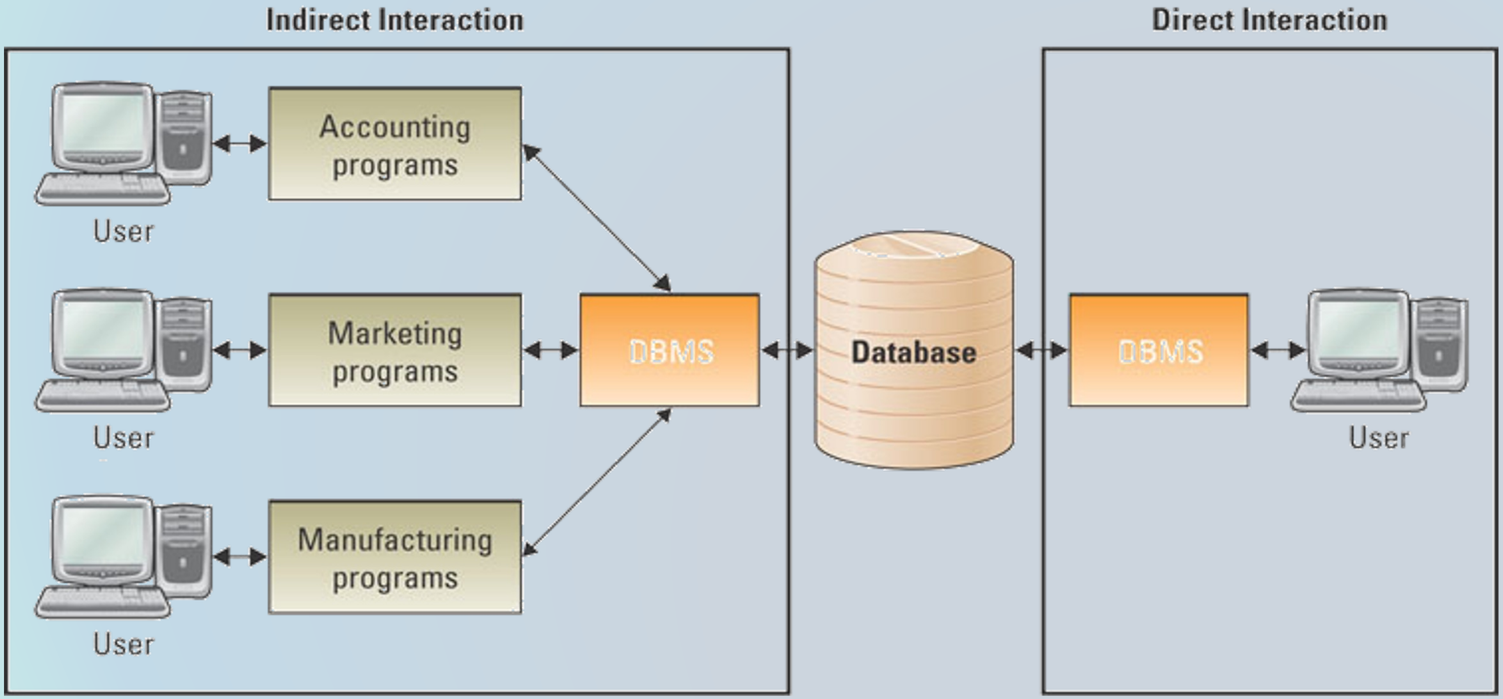
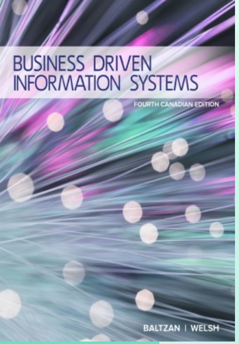


Figure 4.2



Data-Driven Web Sites

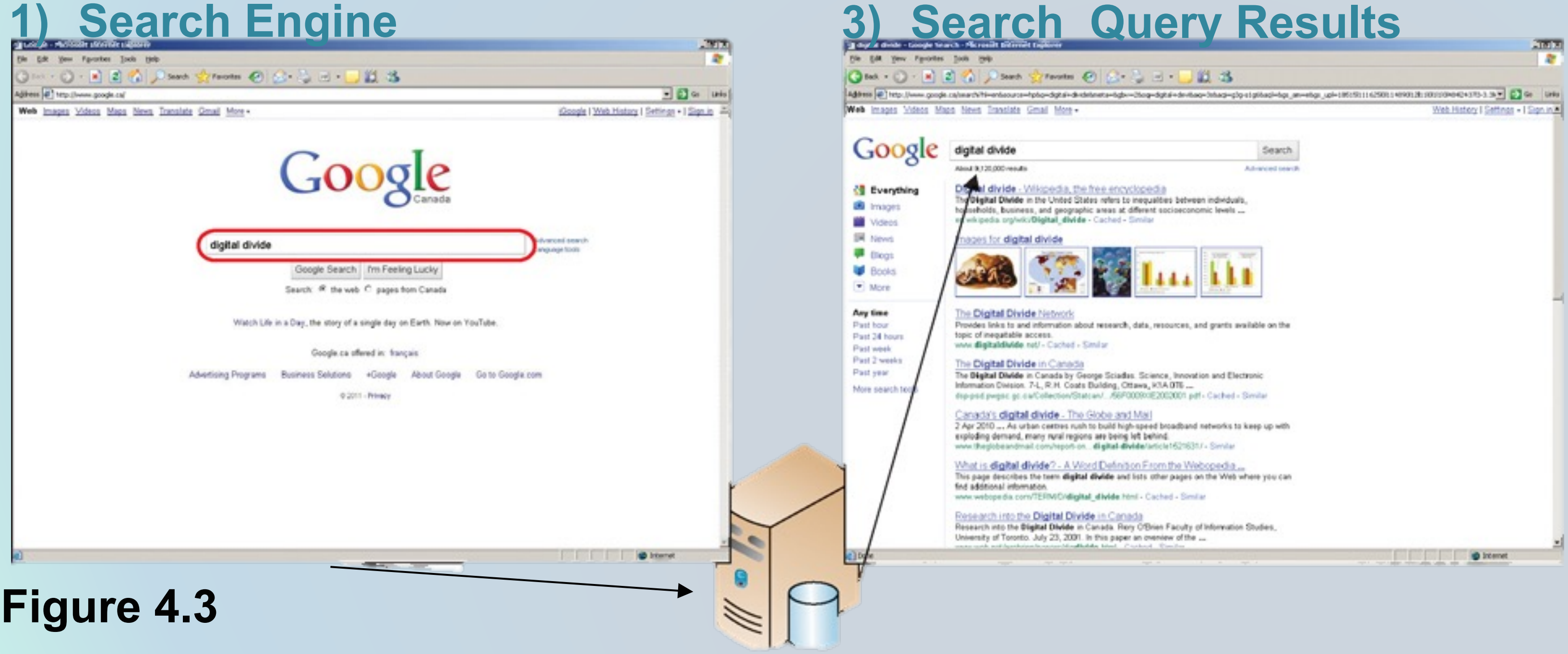
Learning Outcome 4.3

An interactive Web site which uses a database to keep it updated and relevant to the needs of its customers.

A Data-driven Website

1) Search Engine

3) Search Query Results



Digital divide - Wikipedia, the free encyclopedia
The Digital Divide in the United States refers to inequalities between individuals, households, business, and geographic areas at different socioeconomic levels ...
en.wikipedia.org/wiki/Digital_divide - Cached - Similar

Images for digital divide

The Digital Divide Network
Provides links to and information about research, data, resources, and grants available on the topic of inequitable access.
www.digitaldivide.net/ - Cached - Similar

The Digital Divide in Canada
The Digital Divide in Canada by George Sciadas. Science, Innovation and Electronic Information Division. 7-L, R.H. Coats Building, Ottawa, K1A 0T6 ...
dsp-psd.pwgsc.gc.ca/Collection/Statcan/.../56F0009XIE2002001.pdf - Cached - Similar

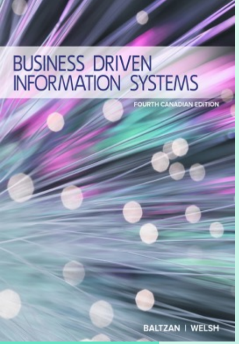
Canada's digital divide - The Globe and Mail
2 Apr 2010 ... As urban centres rush to build high-speed broadband networks to keep up with exploding demand, many rural regions are being left behind.
www.theglobeandmail.com/report-on...digital-divide/article1521631/ - Similar

What is digital divide? - A Word Definition From the Webopedia ...
This page describes the term digital divide and lists other pages on the Web where you can find additional information.
www.webopedia.com/TERM/D/digital_divide.html - Cached - Similar

Research into the Digital Divide in Canada
Research into the Digital Divide in Canada. Rory O'Brien Faculty of Information Studies, University of Toronto. July 23, 2001. In this paper an overview of the ...

Figure 4.3

2) Database



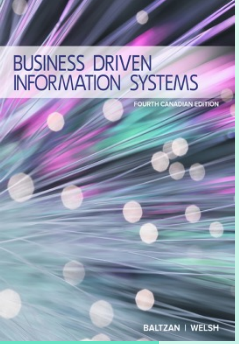
Data Integration

Learning Outcome

4.3

Allows separate systems to communicate directly with each other.

- **Forward integration** takes information entered into a given system and sends it automatically to all downstream systems and processes.
- **Backward integration** takes information entered into a given system and sends it automatically to all upstream systems and processes.



Forward and Backward Integration

Learning Outcome

4.3

Forward and Backward Customer Data Integration

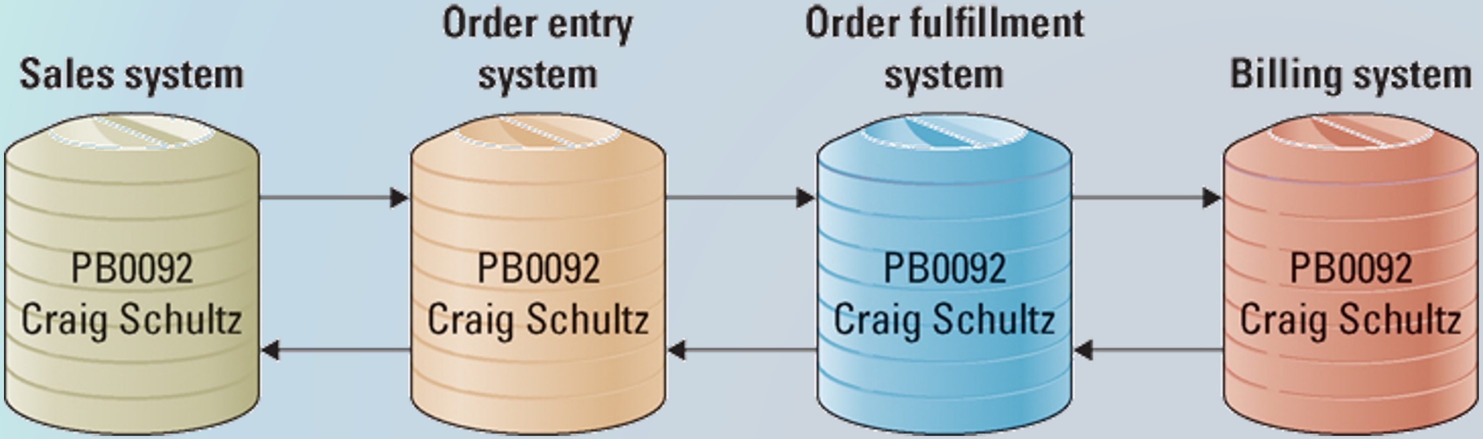
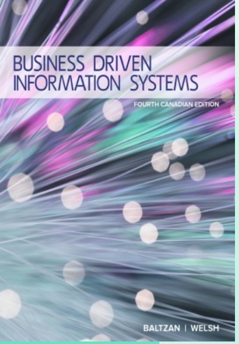


Figure 4.5



Integrated Customer Data

Learning Outcome

4.3

Integrated Customer Data

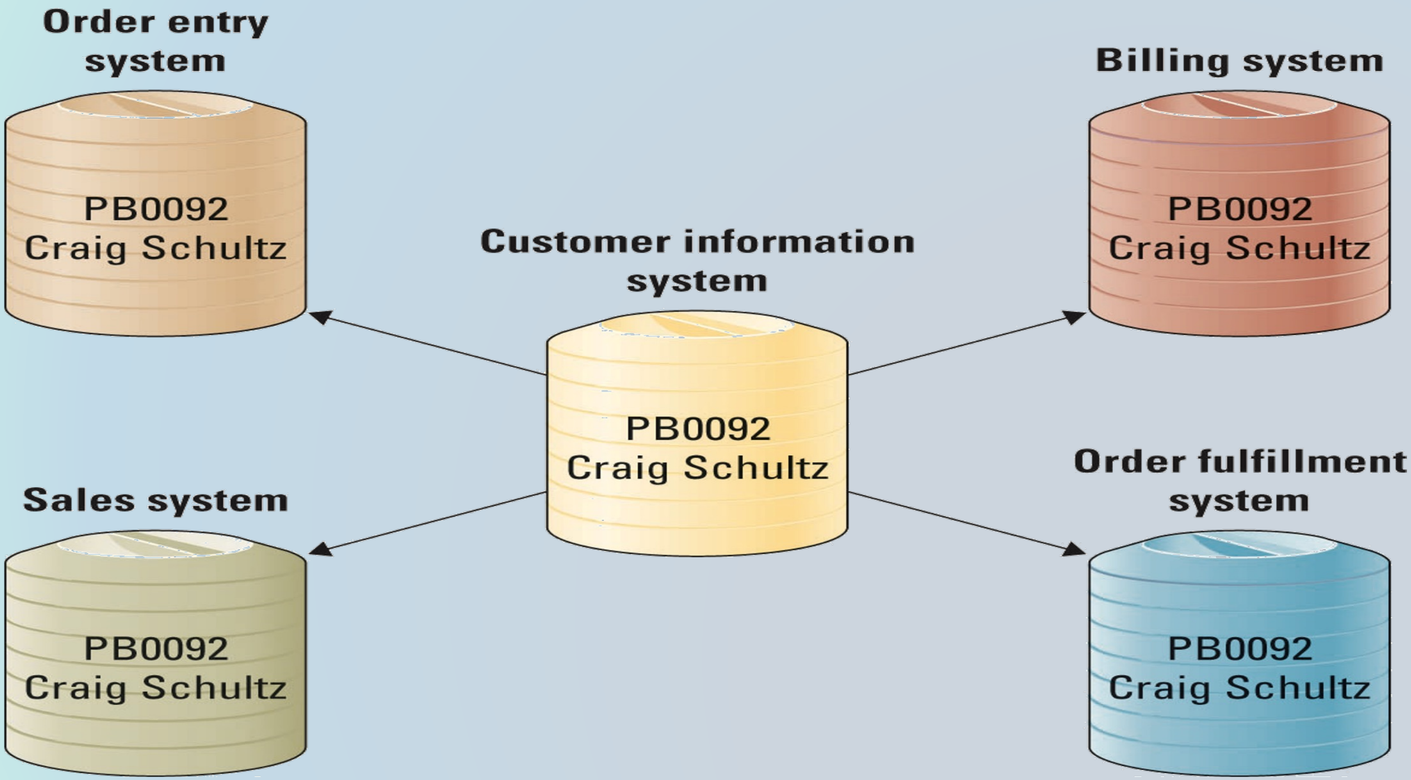
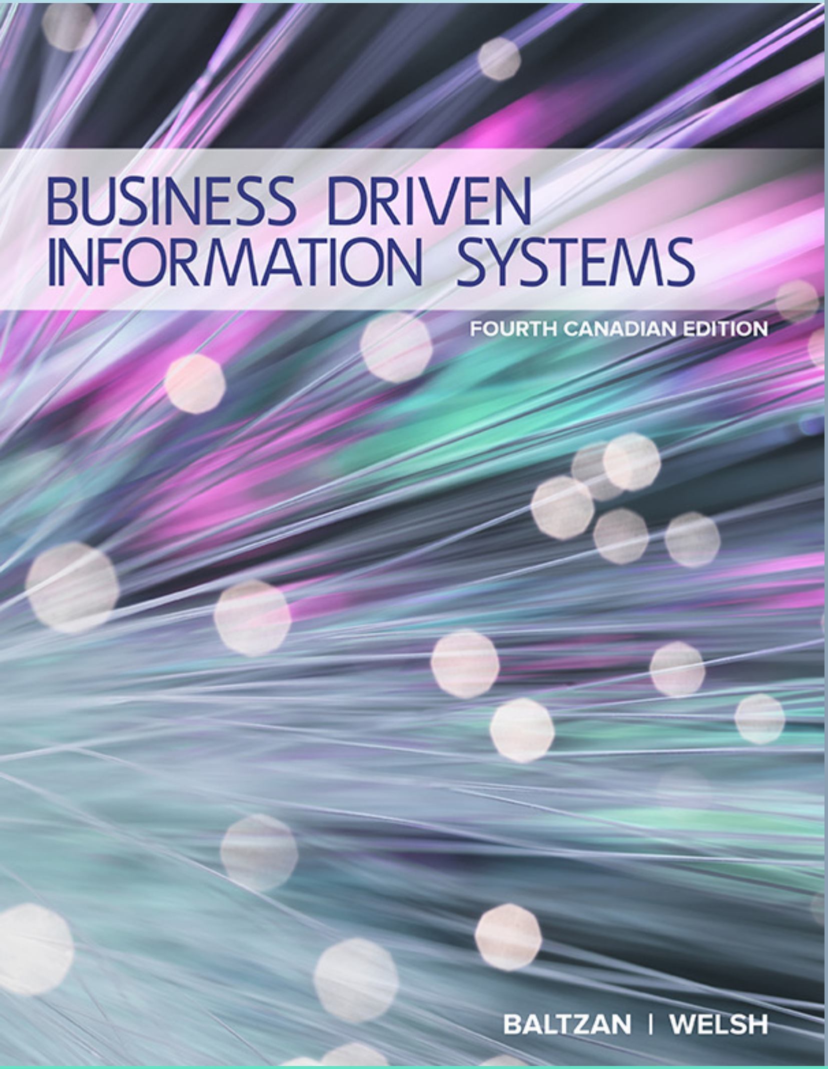
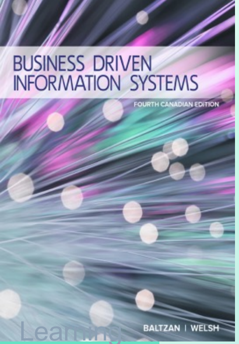


Figure 4.6



DATA WAREHOUSING

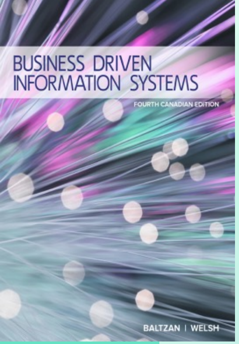


Learning Outcome

4.4

History of Data Warehousing

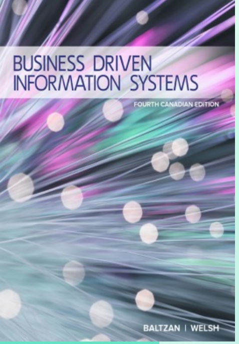
- In the 1990's, Functional systems were too cumbersome & inefficient
  - Operations systems and data were not integrated.
  - Little historic data, little trend information
  - Quality issues
  - Good for transactions processing, not analysis
- Turn of the millennium
  - Data scattered over too many platforms
  - Complex analysis was not timely



Data Warehouse Fundamentals

Learning Outcome
4.4

- **Data warehouse**
  - – A logical collection of information
  - – Gathered from many different operational databases
  - – Supports strategic business analysis activities and decision-making tasks.
- Primary Purpose
  - – To aggregate information throughout an organization
  - – Not a location for ALL data, **only data of interest**.

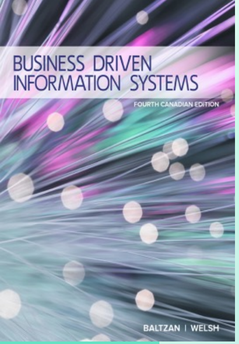


Characteristics of Data Warehouses

Learning Outcome

4.4

- Subject oriented
  - Information is organized around a major organizational subject area, e.g.. Customers
- Integrated
  - Sourced from a variety of internal operational systems and external databases into a coherent whole
- Time-variant
  - Time-stamped according to its cycle (daily, yearly etc.)
- Non-volatile
  - Once loaded, data does not change



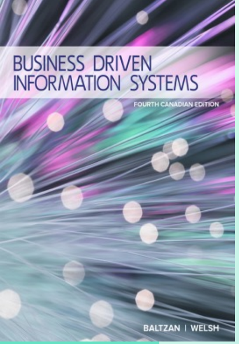
Learning Outcome

4.4

Data Warehouse Fundamentals

- **Extraction, transformation, and loading (ETL)**
  - A process that extracts information from internal and external databases,
  - Transforms the information using a common set of enterprise definitions
  - Loads the information into a data warehouse.

- **Data mart**
  - Contains a subset of data warehouse information
  - Extracted to be analyzed for specific objectives.



Model of a Typical Data Warehouse

Learning Outcome

4.4

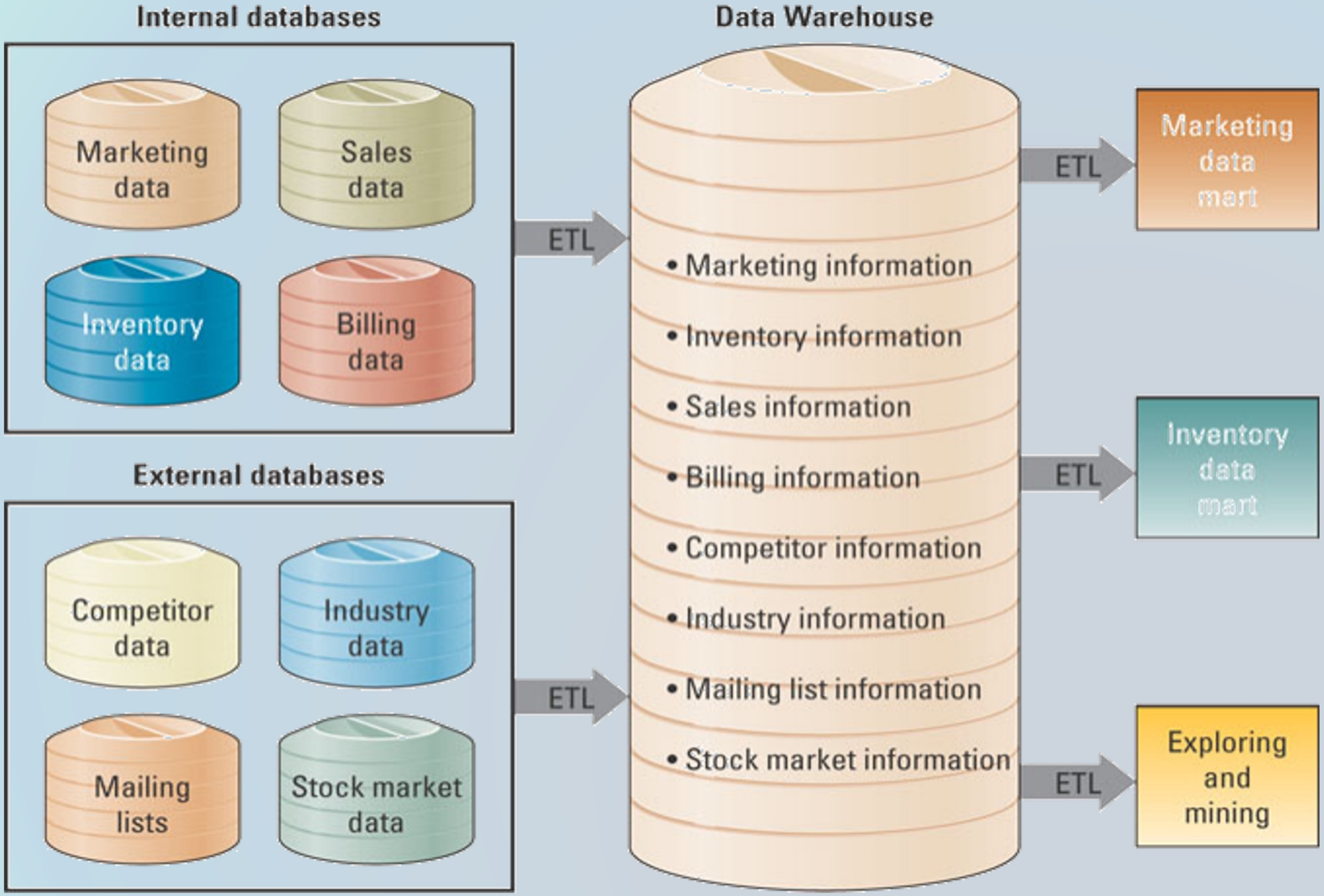
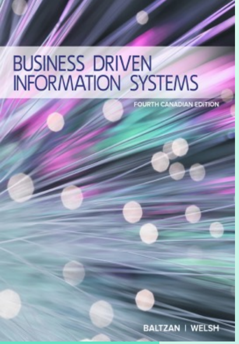


Figure 4.7

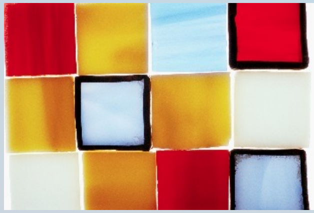


Multi-dimensional Analysis

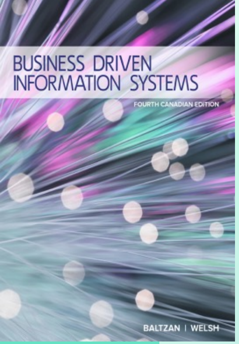
Learning Outcome

4.4

- Databases contain information in two-dimensional tables…rows and columns



- Data warehouse information is three-dimensional…layers of rows and columns
  - Each **Dimension** is a particular characteristic of the information; an attribute.
  - **Cube** is acommon term for the representation of multi-dimensional information.



Multi-dimensional Analysis

Learning Outcome

4.4

A Cube of Information for Performing Multi-Dimensional Analysis on Three Stores for Five Products and Four Promotions.

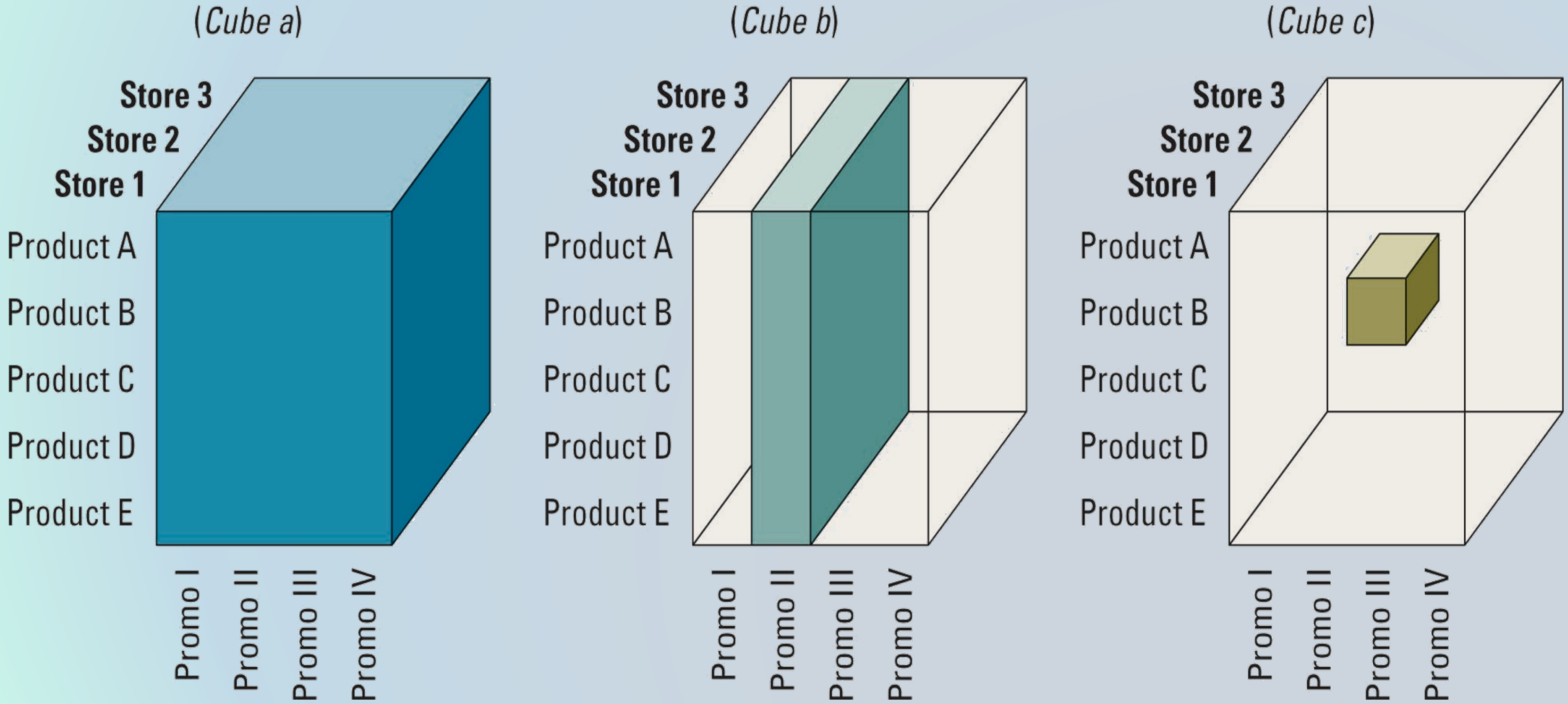
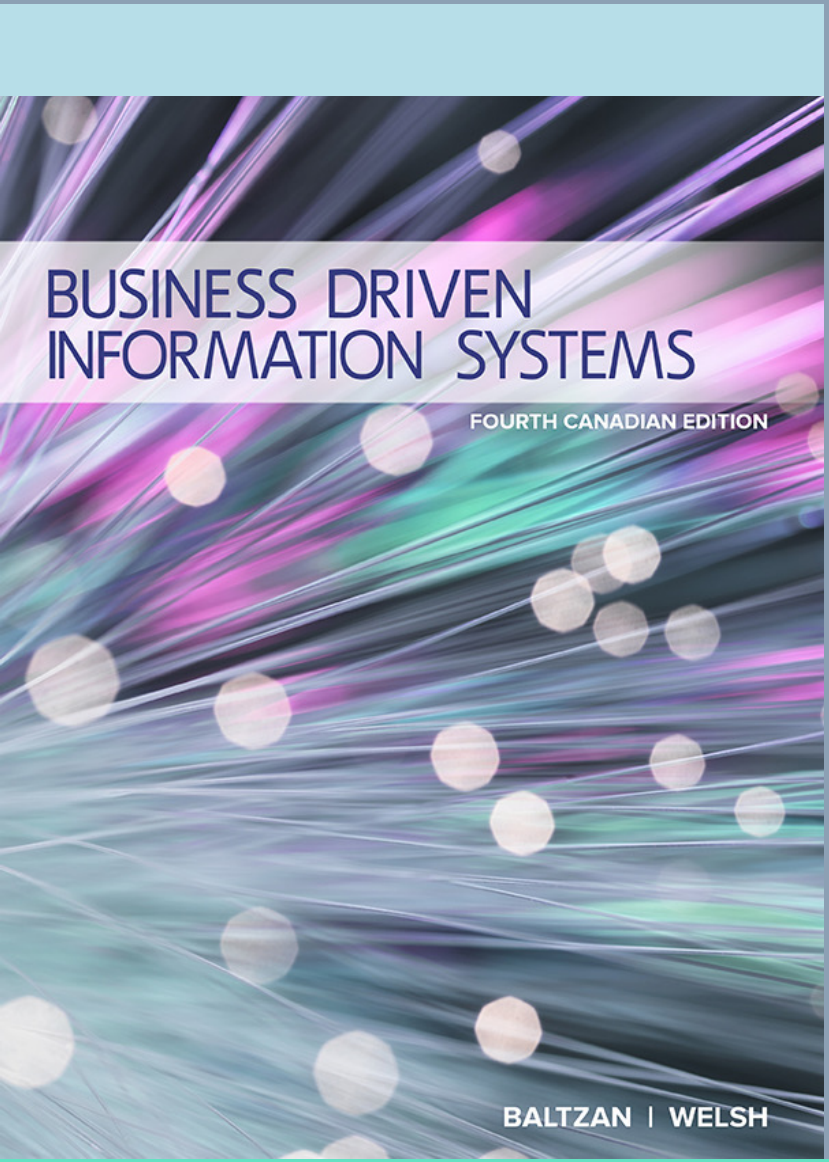
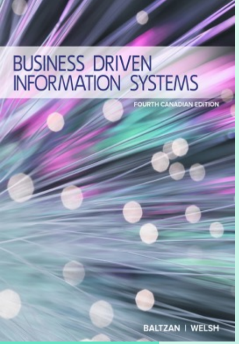


Figure 4.8


BUSINESS DRIVEN
INFORMATION SYSTEMS
FOURTH CANADIAN EDITION
BALTZAN | WELSH

DATA MINING

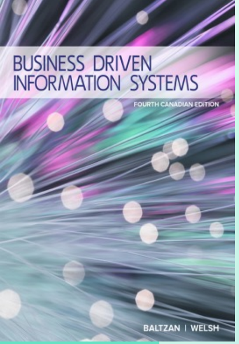


Learning Outcome

4.5

Data Mining

- The process of analyzing data to extract information.
  - **Drilling Down** progresses through increasing levels of detail.
  - **Drilling Up** works through increasing levels of summarization.
- **Data Mining Tools**
  - Variety of techniques that find patterns and relationships in large volumes of information.
  - Specialized technologies and functionalities including Query tools, reporting tools, statistical tools and intelligence agents.



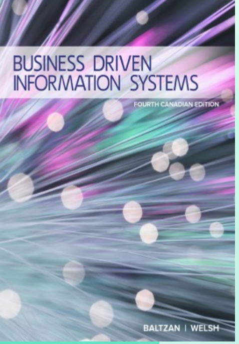
Data Mining Activities

Learning Outcome

4.5

Apply algorithms to information sets to uncover inherent trends and patterns which are used to develop new business strategies.

- **Classification**
  - Assigning records to one of a pre-defined set of classes
- **Estimation**
  - Determining the values for an unknown continuous variable behavior
- **Affinity grouping**
  - Which things go together
- **Clustering**
  - Breaks up a heterogeneous population of records into a number of more homogenous subgroups.



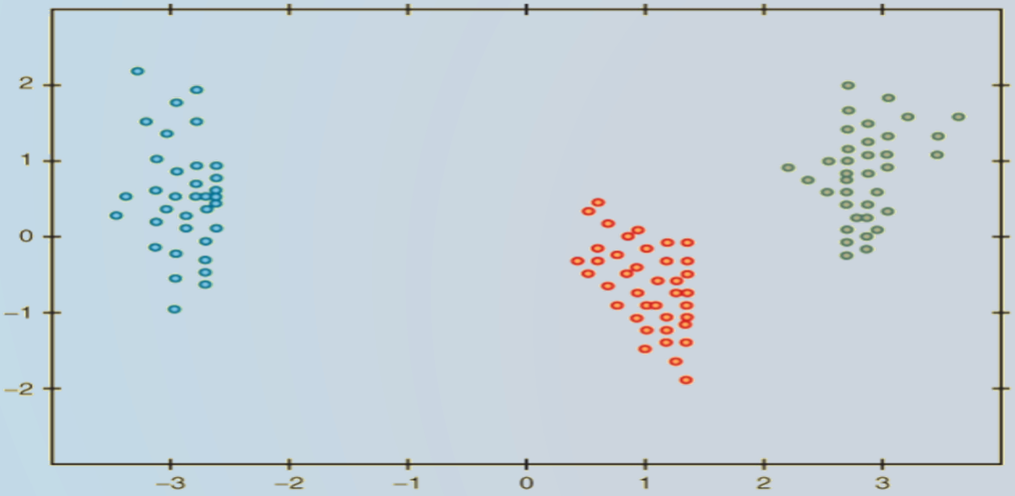
Data Mining Techniques

Learning Outcome

4.5

Cluster analysis

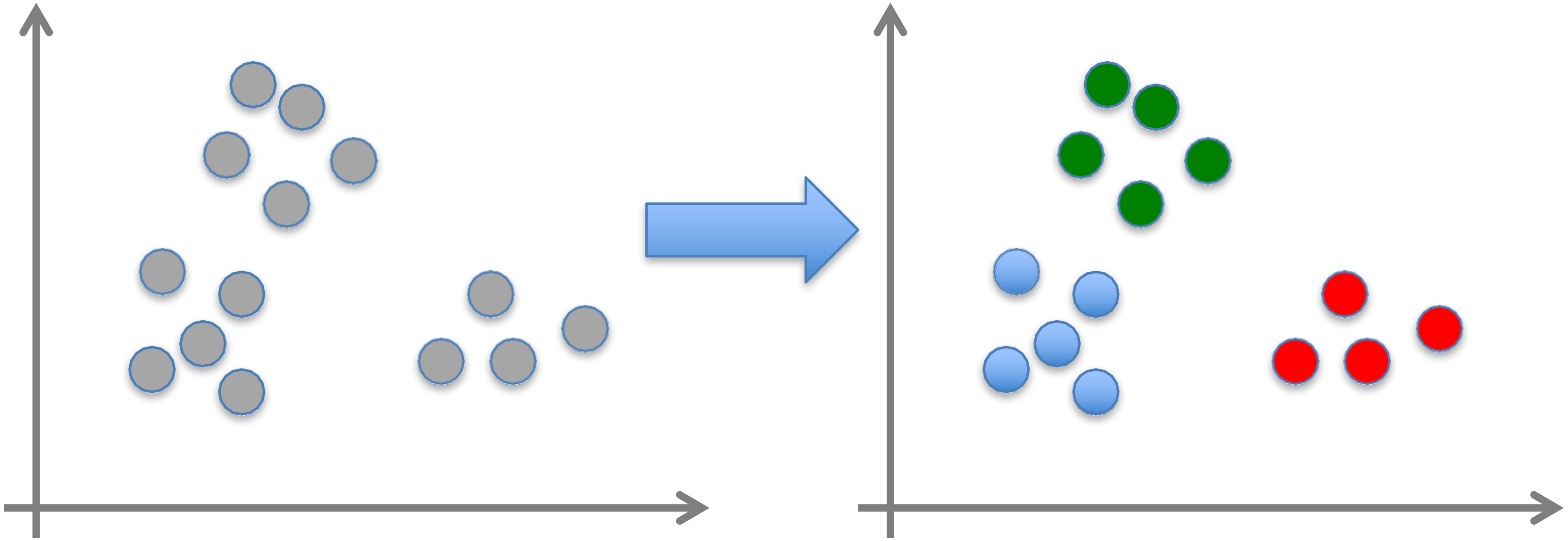
- A statistical technique used to divide an information set into mutually exclusive groups such that the members of each group are as close together as possible to one another and the different groups are as far apart as possible.

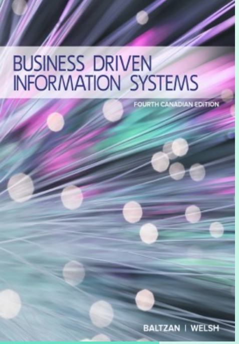


Copyright © 2015 McGraw-Hill Ryerson Limited

Unsupervised Learning - Clustering

- Given x_1, x_2, ..., x_n (without labels)
- Output hidden structure behind the x's





Association Detection

Learning Outcome 4.5

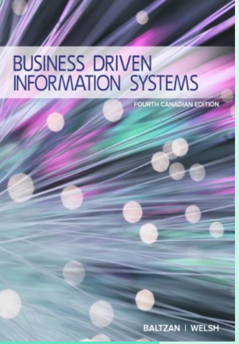
Association detection

Reveals the relationship between variables along with the nature and frequency of the relationships

- **Rule Generators**
  - Form business rules from the data mining applications
  - Predict business events and their probability of occurrence
- **Market basket analysis**
  - Analyzes websites & checkout scanners
  - Predict future buyer behaviour

Data Collection for Market Basket Analysis

Figure 4.14



Statistical Analysis

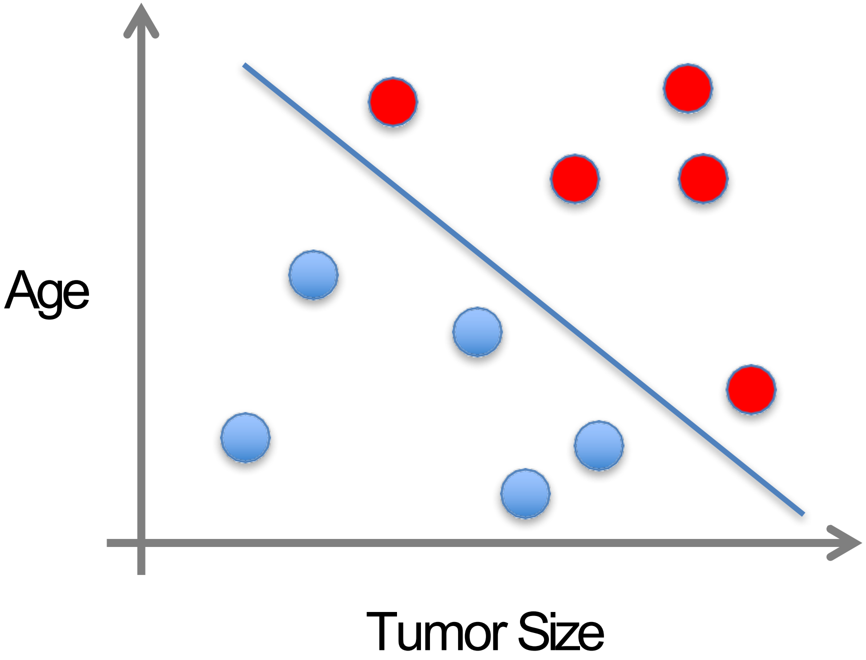
Learning Outcome 4.5

Performs such functions as information correlations, distributions, calculations, and variance analysis

- **Definition** of qualitative variables and assigns them numerical scales. Then, builds models, forecasts and trends based on consumer testing.
- **Forecast** – Predictions made on the basis of time-series information
- **Time-series information** – Data collected at regular, equal-spaced, periods. Used for trend analysis.
- Many large vendors provide end-to-end data mining decision tools with predictive analytical capabilities.
- Many Data Scientists write their own scripts

Prediction

- x can be multi-dimensional
  - – Each dimension corresponds to an attribute



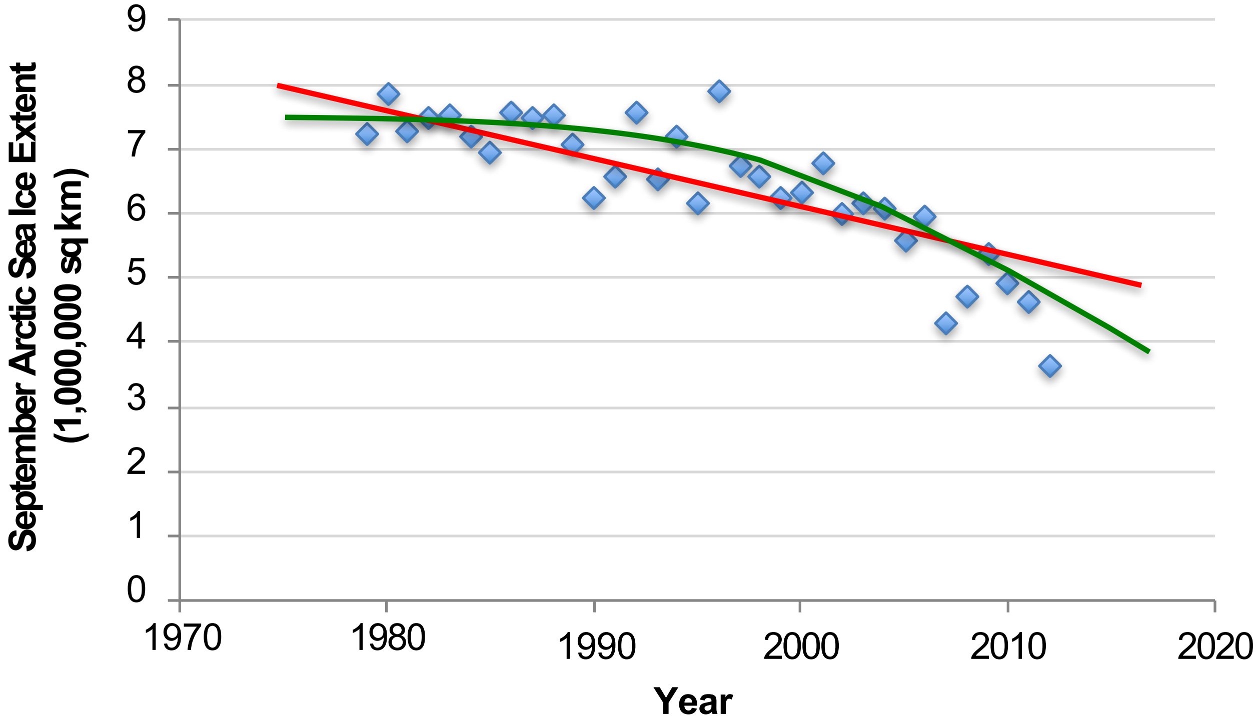
- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape

…

Based on example by Andrew Ng

Linear Regression

- Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
  - – y is real-valued == regression



Data from G. Witt. Journal of Statistics Education, Volume 21, Number 1 (2013)