15.561
Information Technology Essentials

Session 15
Business Intelligence: Data Mining and Data Warehousing

Adapted from Chris Dellarocas, U. Md.

Outline

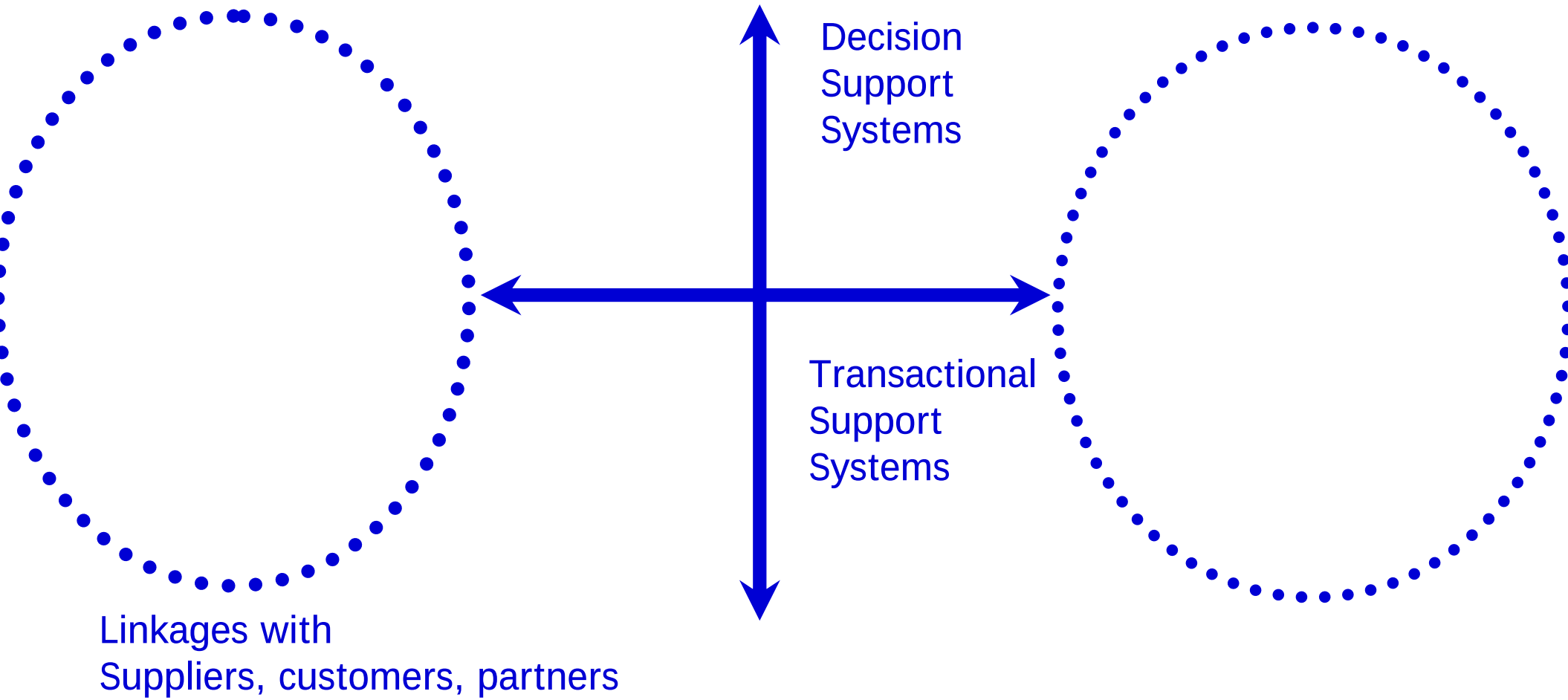
- Operational vs. Decision Support Systems
- What is Business Intelligence?
- Overview of Data Mining
- Case Studies
- Data Warehouses

Major IT applications in business

	Sales and Marketing	Manufacturing	Finance	Accounting	Human resources
Executive Support Systems	5-year sales trend forecasts	5-year operating plan	5-year budget forecasts	Profit planning	Manpower planning
Management Information Systems	Sales management Sales region analysis	Inventory control Production scheduling	Annual budgeting Cost analysis	Capital investment analysis Pricing/profitability analysis	Relocation analysis Contract cost analysis
Knowledge Worker Systems	Engineering workstations	Word processing	Email	Web viewing	Spreadsheets
Transaction Processing Systems	Public web sites Order tracking Order processing	Machine control Plant scheduling Material movement control	Securities trading Cash management	Payroll Accounts payable Accounts receivable	Compensation Training & Development Employee records

Adapted from Laudon & Laudon, *Management Information Systems: Organization and Technology*, Prentice Hall, 1998, p. 39

Operational vs. Decision Support Systems



Operational vs. Decision Support Systems

- **Operational Systems**
  - Support day to day transactions
  - Contain current, "up to date" data
  - Examples: customer orders, inventory levels, bank account balances

- **Decision Support Systems**
  - Support strategic decision making
  - Contain historical, "summarized" data
  - Examples: performance summary, customer profitability, market segmentation

Example of an Operational Application: Order Entry

Orders

Bill To: Rattlesnake Canyon Grocery
2817 Milton Dr.
Albuquerque NM 87110
USA

Ship To: Rattlesnake Canyon Grocery
2817 Milton Dr.
Albuquerque NM 87110
USA

Salesperson: Peacock, Margaret

Ship Via: ☑ Speedy ☐ United ☐ Federal

Order ID: 10024 Order Date: 19-Jun-91 Required Date: 17-Jul-91 Shipped Date: 21-Jun-91

Prod ID:	Product:	Unit Price:	Quantity:	Discount:	Extended Price:
43	Ipoh Coffee	$32.00	10	0%	$320.00
53	Perth Pasties	$22.90	15	0%	$343.50
56	Gnocchi di nonna Alice	$26.00	60	0%	$1,560.00

Record: 1 of 3

NORTHWIND TRADERS

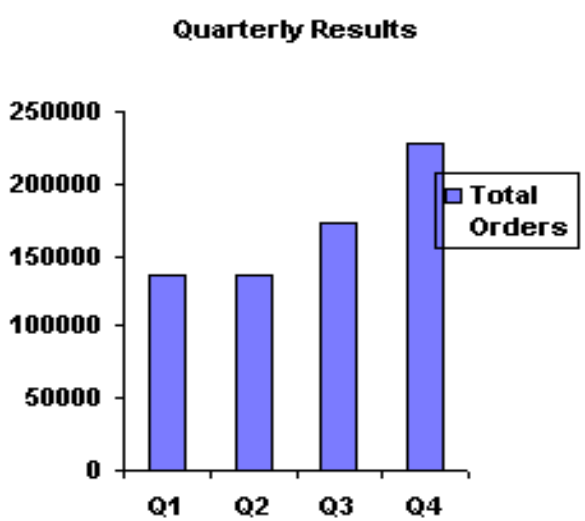
Subtotal:	$2,223.50
Freight:	$5.19
Total:	$2,228.69

Record: 26 of 1080

Example of a DSS Application: Annual performance summary

NORTHWIND TRADERS

Performance Summary for year: 1993



Most valuable products:

Product Name:	Order Total:
Côte de Blaye	63463.97
Raclette Courdavault	39160
Thüringer Rostbratwurst	34546.14
Gnocchi di nonna Alice	31304.4
Maniimun Dried Apples	25312.8

Record: 1 of 77

Best customers:

Company Name:	Order Total:
QUICK-Stop	79495.3
Save-a-lot Markets	62685.49
Ernst Handel	47300.55
Königlich Essen	21584.87
Rattlesnake Canyon Grocery	19639.15

Record: 1 of 87

Most important countries:

Ship Country:	Order Total:
Germany	151046.38
USA	136108.2
Austria	56690.13
France	46935.04
Brazil	45463.73

Record: 1 of 21

What is Business Intelligence?

- Collecting and refining information from many sources
- Analyzing and presenting the information in useful ways
- So people can make better business decisions

What is Data Mining?

- Using a combination of artificial intelligence and statistical analysis to analyze data
- and discover useful patterns that are "hidden" there

Sample Data Mining Applications

- Direct Marketing
  - identify which prospects should be included in a mailing list
- Market segmentation
  - identify common characteristics of customers who buy same products
- Customer churn
  - Predict which customers are likely to leave your company for a competitor
- Market Basket Analysis
  - Identify what products are likely to be bought together
- Insurance Claims Analysis
  - discover patterns of fraudulent transactions
  - compare current transactions against those patterns

Business uses of data mining

Essentially five tasks…

- Classification
  - Classify credit applicants as low, medium, high risk
  - Classify insurance claims as normal, suspicious
- Estimation
  - Estimate the probability of a direct mailing response
  - Estimate the lifetime value of a customer
- Prediction
  - Predict which customers will leave within six months
  - Predict the size of the balance that will be transferred by a credit card prospect

Business uses of data mining

- **Affinity Grouping**
  - Find out items customers are likely to buy together
  - Find out what books to recommend to Amazon.com users
- **Description**
  - Help understand large volumes of data by uncovering interesting patterns

Overview of Data Mining Techniques

- Market Basket Analysis
- Automatic Clustering
- Decision Trees and Rule Induction
- Neural Networks

Market Basket Analysis

- Association and sequence discovery
- Principal concepts
  - Support or Prevalence: frequency that a particular association appears in the database
  - Confidence: conditional predictability of B, given A
- Example:
  - Total daily transactions: 1,000
  - Number which include "soda": 500
  - Number which include "orange juice": 800
  - Number which include "soda" and "orange juice": 450
  - SUPPORT for "soda and orange juice" = 45% (450/1,000)
  - CONFIDENCE of "soda → orange juice" = 90% (450/500)
  - CONFIDENCE of "orange juice → soda" = 56% (450/800)

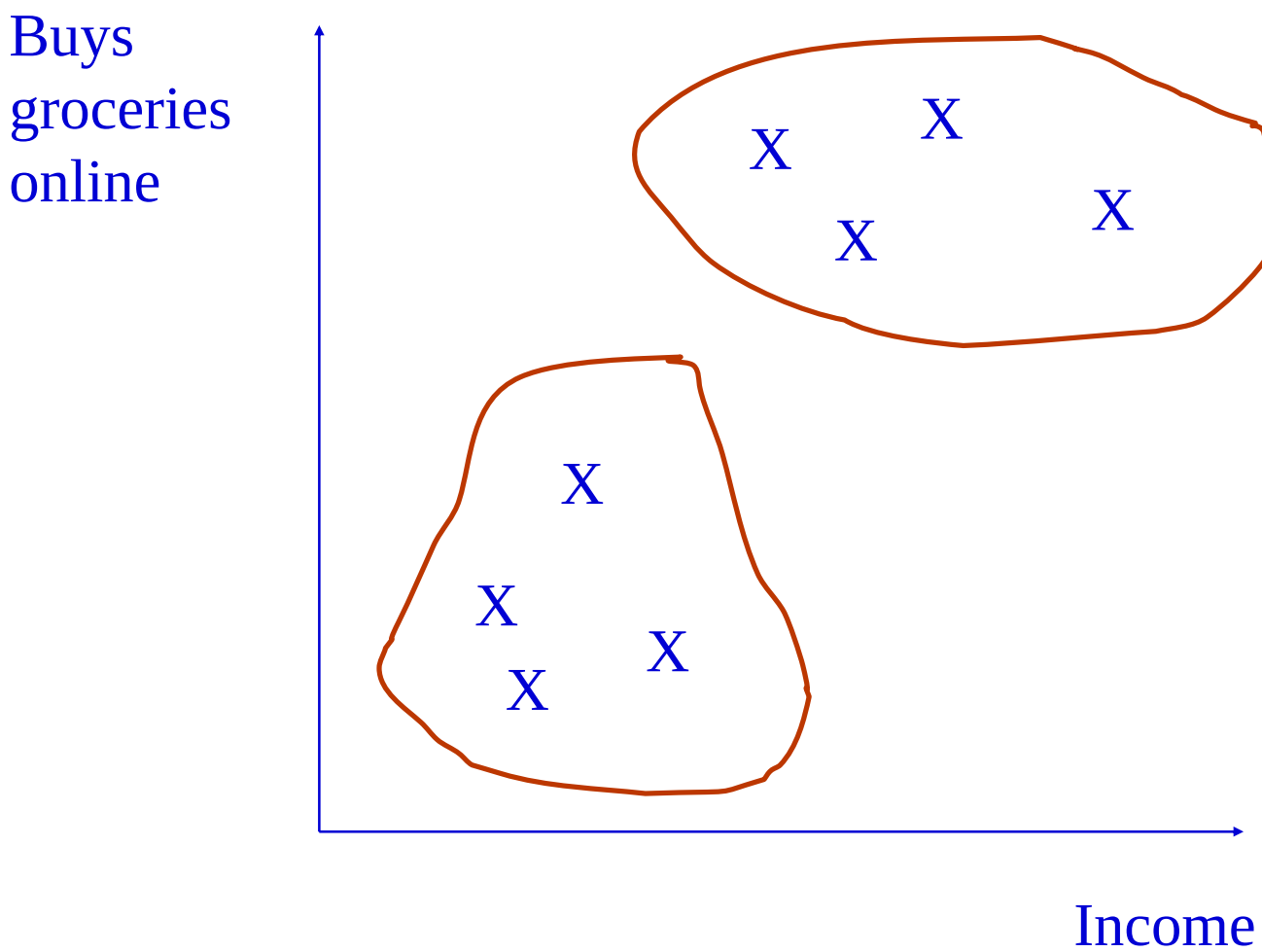
Applying Market Basket Analysis

- Create co-occurrence matrix
  - What is the right set of items???
- Generate useful rules
  - Weed out the trivial and the inexplicable from the useful
- Figure out how to act on them
- Similar techniques can be applied to time series for mining useful *sequences* of actions

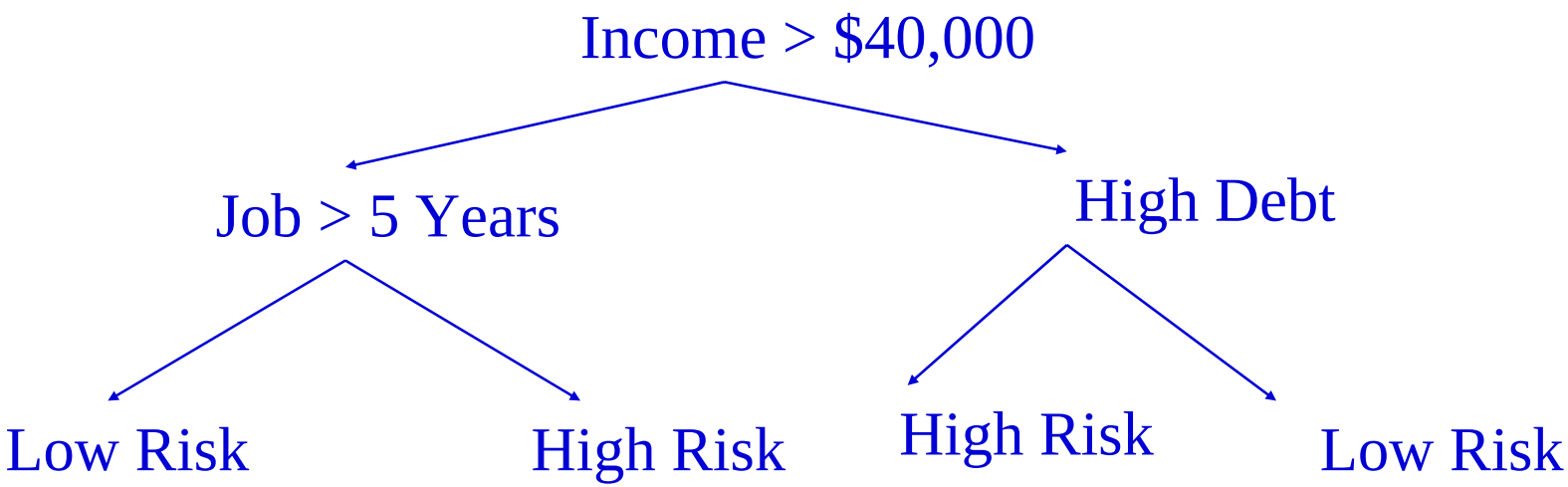
Clustering

- **Divide a database into groups ("clusters")**
- **Goal: Find groups that are very different from each other, and whose members are similar to each other**
- **Number and attributes of these groups are *not known in advance***

Clustering (example)



Decision Trees



- **Data mining is used to construct the tree**
- **Example algorithm: CART (Classification and Regression Trees)**

Decision tree construction algorithms

- Start with a training set (i.e. preclassified records of loan customers)
  - Each customer record contains
    - » Independent variables: income, time with employer, debt
    - » Dependent variable: outcome of past loan
- Find the independent variable that best splits the records into groups where one single class (low risk, high risk) predominates
  - Measure used: entropy of information (diversity)
  - Objective:
    - » max[diversity before – (diversity left + diversity right)]
- Repeat recursively to generate lower levels of tree

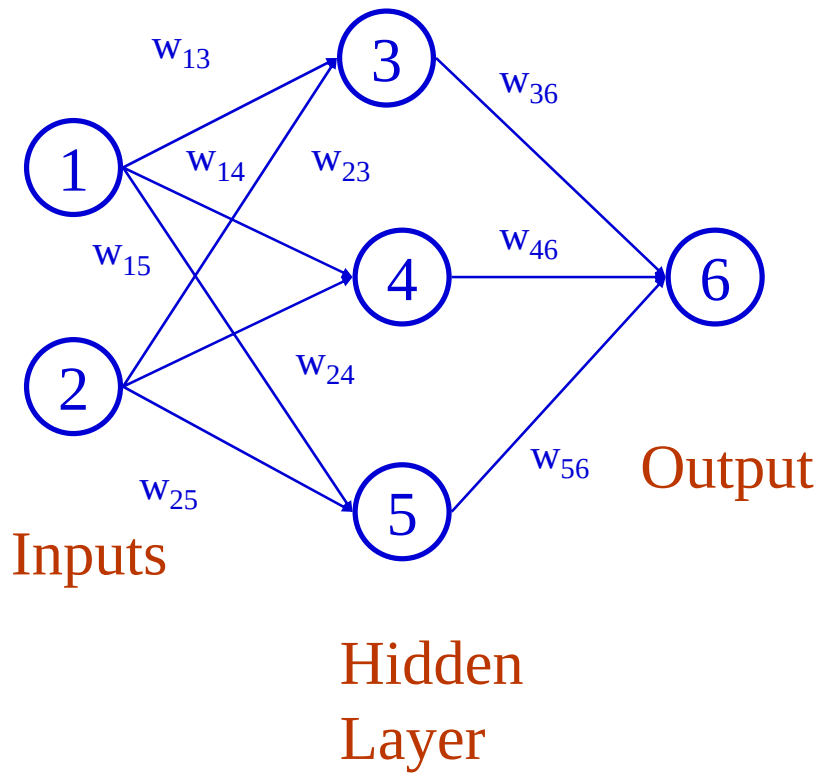
Decision Tree pros and cons

- **Pros**
  - One of the most intuitive techniques, people really like decision trees
  - Really helps get some intuition as to what is going on
  - Can lead to direct actions/decision procedures

- **Cons**
  - Independent variables are not always the best separators
  - Maybe some of them are correlated/redundant
  - Maybe the best splitter is a linear combination of those variables (remember factor analysis)

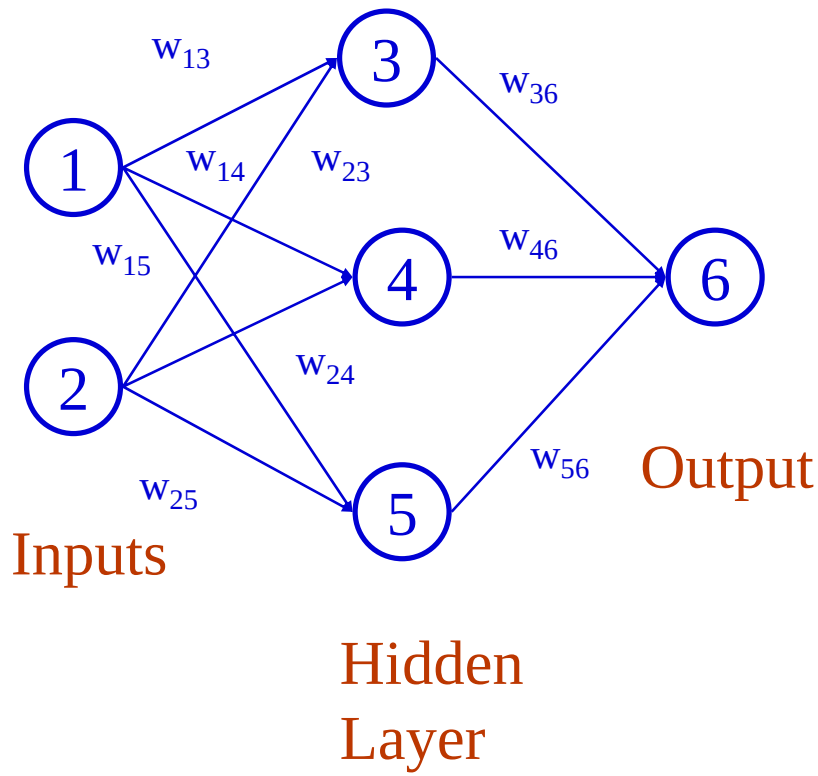
Neural Networks

- Powerful method for constructing predictive models
- Each node applies an activation function to its input
- Activation function results are multiplied by w_{ij} and passed on to output



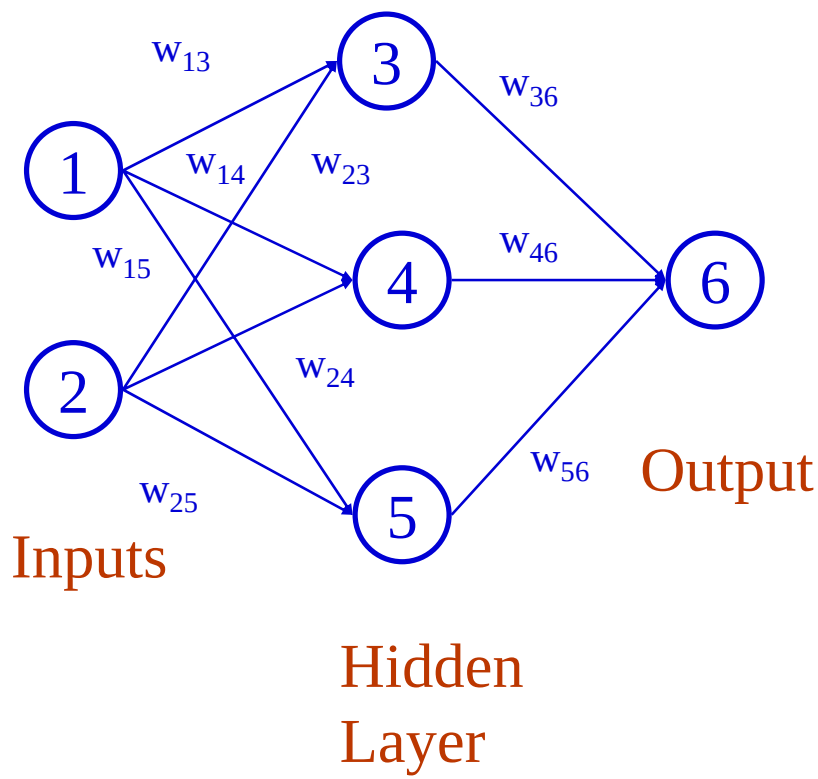
Neural Networks

- **Weights are determined using a "training set", I.e. a number of test cases where both the inputs and the outputs are known**



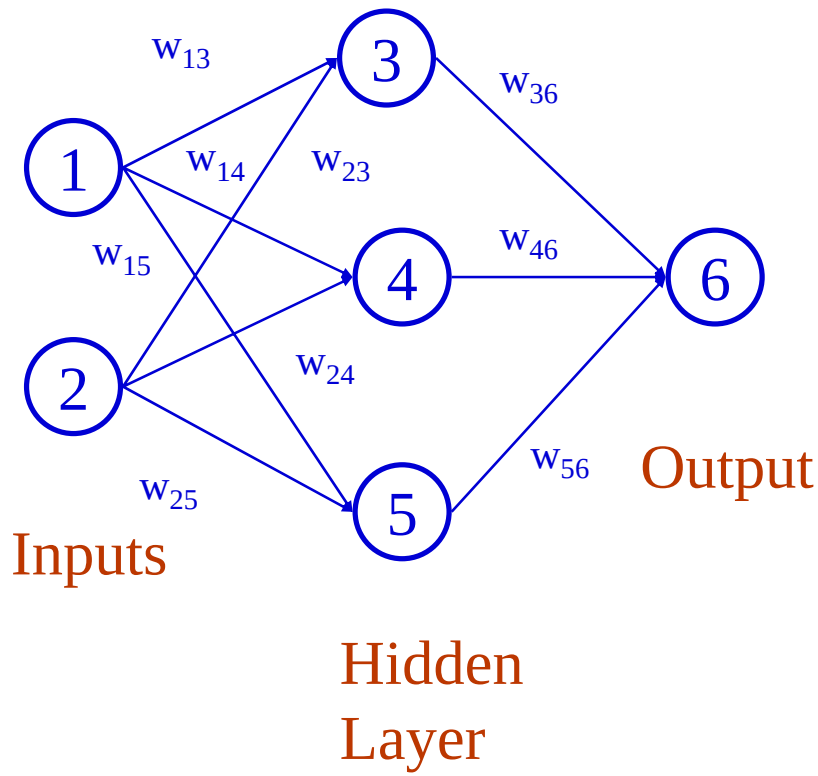
Neural Networks

- **Example: Build a neural net to calculate credit risk for loan applicants**
- **Inputs: annual income, loan amount, loan duration**
- **Outputs: probability of default [0,1]**
- **Training set: data from past customers with known outcomes**



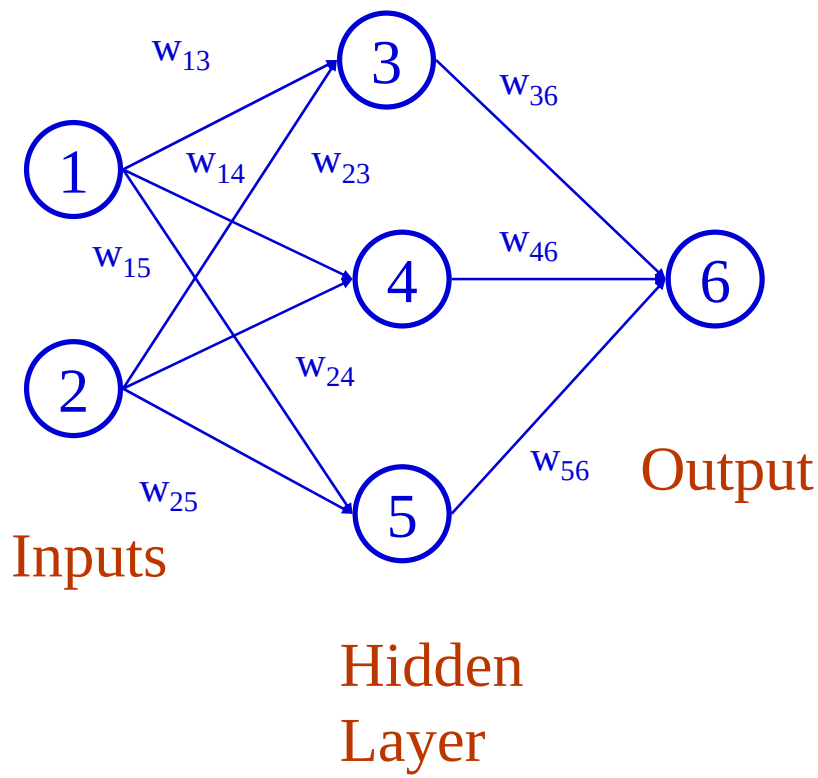
Neural Networks

- Start from an initial estimate for the weights
- Feed the independent variables for the first record into inputs 1 and 2
- Compare with output and calculate error
- Update estimates of weights by back-propagating error



Neural Networks

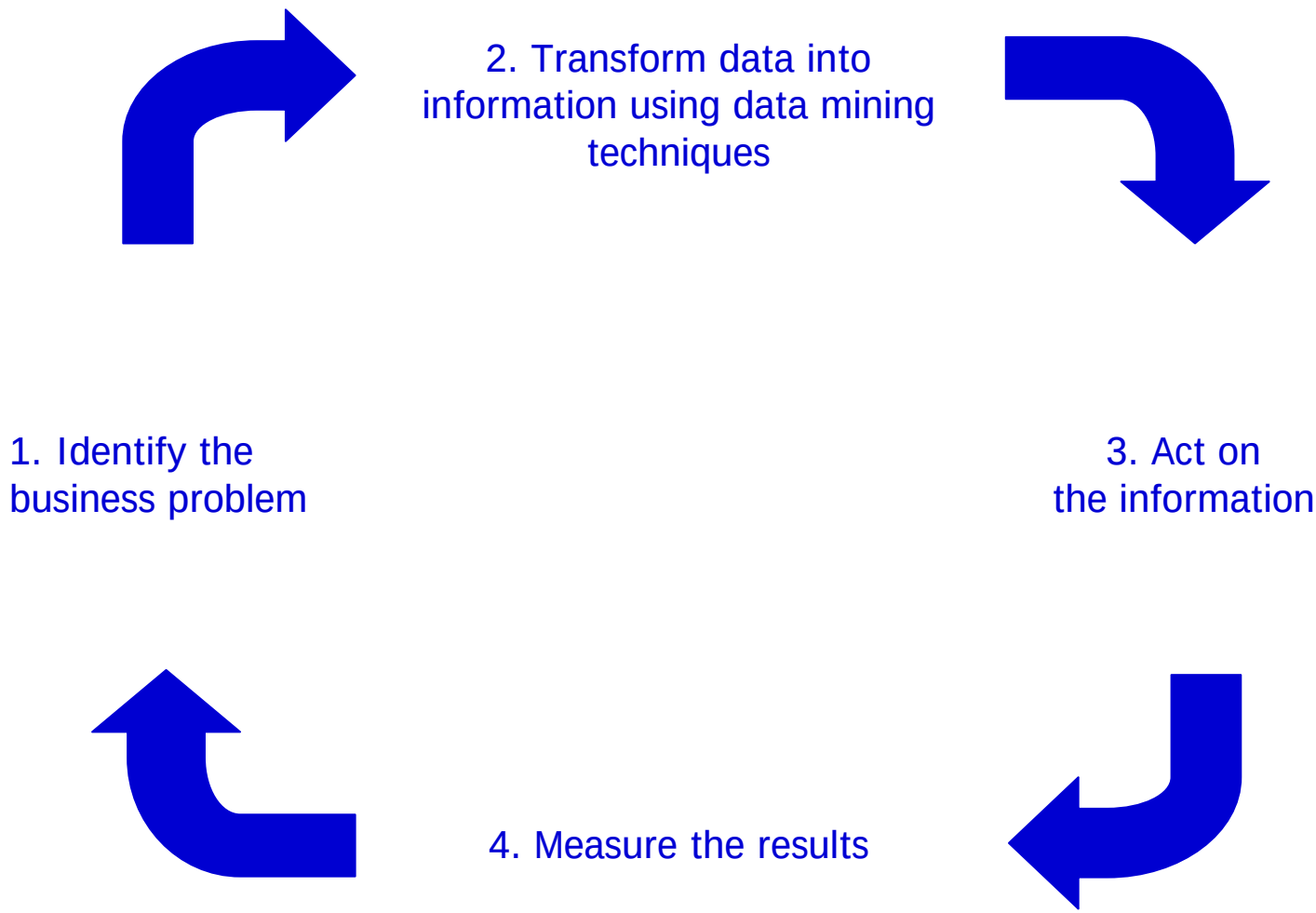
- **Repeat with next training set record until model converges**



Neural networks pros and cons

- **Pros**
  - Versatile, give good results in complicated domains
- **Cons**
  - Neural nets cannot explain the data
  - Inputs and outputs usually need to be massaged into fixed intervals (e.g., between -1 and +1)

The Virtuous Cycle of Data Mining



Case study 1: Bank is losing customers…

- Attrition rate greater than acquisition rate
- More profitable customers seem to be the ones to go
- What can the bank do?

Bank is losing customers…

- **Step 1: Identify the opportunity for data analysis**
  - Reducing attrition is a profitable opportunity

- **Step 2: Decide what data to use**
  - Traditional approach: surveys
  - New approach: Data Mining

Bank is losing customers…

- Clustering analysis on call-center detail
- Interesting clusters that contain many people who are no longer customers
- Cluster X: People considerably older than average customer and less likely to have mortgage or credit card
- Cluster Y: People who have several accounts, tend to call after hours and have to wait when they call. Almost never visit a branch and often use foreign ATMs

Bank is losing customers…

- Step 3: Turn results of data mining into action

Case study 2: Bank of America

- BoA wants to expand its portfolio of home equity loans
- Direct mail campaigns have been disappointing
- Current common-sense models of likely prospects
  - People with college-age children
  - People with high but variable incomes

Enter data mining…

- BoA maintains a large historical DB of its retail customers
- Used past customers who had (had not) obtained the product to build a decision tree that classified a customer as likely (not likely) to respond to a home equity loan
- Performed clustering of customers
- An interesting cluster came up:
  - 39% of people in cluster had both personal and business accounts with the bank
  - This cluster accounted for 27% of the 11% of customers who had been classified by the DT as likely respondents to a home equity offer

Completing the "cycle"

3. The resulting Actions (*Act*)

- Develop a campaign strategy based on the new understanding of the market
- The acceptance rate for the home equity offers more than doubled

4. Completing the Cycle (*Measure*)

- Transformation of the retail side of Bank of America from a mass-marketing institution to a targeted-marketing institution (learning institution)
- Product mix best for each customer => "Market basket analysis" came to exist

What is a data warehouse?

- A collection of data from multiple sources
  - » within the company
  - » outside the company
- Usually includes data relevant to the entire enterprise
- Usually includes summary data and historical data as well as current operational data
- Usually requires “cleaning” and other integration before use
- Therefore, usually stored in separate databases from current operational data

What is a data mart?

- A subset of a data warehouse focused on a particular subject or department

Data Warehousing considerations

- What data to include?
- How to reconcile inconsistencies?
- How often to update?

To delve deeper

- **Recommended books**

  Data Mining Techniques: Michael J. A. Berry and Gordon Linoff

- **Useful collections of links**
  - http://databases.about.com/cs/datamining/