[redacted]

[redacted]

[redacted]

INFORMATION EXTRACTED FROM DATA

People generate significant amounts of digital data daily. Some always-on devices are collecting geographic location data constantly, while social media sites are collecting premium data based on your usage.

People can use computer programs to process information as well as to gain insight and knowledge. Information is the collection of facts and patterns extracted from data.

Gaining insight from this valuable data involves a combination of statistics, mathematics, programming, and problem solving. Large data sets may be analyzed computationally to reveal patterns, trends, and associations. These trends are powerful predictors of future behaviors. Investors are constantly reviewing trends in past pricing to influence their future investment decisions. However, sometimes trends can be misinterpreted and result in business disasters.

Digitally processed data may show correlation between variables. A correlation found in data does not necessarily indicate that a causal relationship exists. Additional research is needed to understand the exact nature of the relationship.

Often, the size of the data set affects the amount of information that can be extracted from it. A single source often does not contain the data needed to draw a conclusion. Combining data from variety of sources may be necessary to formulate a conclusion.

Depending on how the data were collected, the information may not be uniform. For example, if users entered data into an open field, the way they chose to abbreviate, spell, or capitalize something may vary from user to user. Data sets pose challenges regardless of size, such as:

- The need to clean data
- Incomplete data
- Invalid data
- The need to combine data sources

Cleaning data is a process that makes the data uniform without changing their meaning. One example is replacing all equivalent abbreviations with the same word. This can also be done with various spellings and with different capitalizations.

Data can get too large for traditional data-processing applications. The ability to process data depends on the capabilities of the users and their tools. Social media activity generates an enormous amount of data. In the absence of a data-processing application, much of this data will go unexamined. All of the information in the data is too large to examine by hand in real time. Some data sets are difficult to process using a single computer and may require parallel systems. Parallel systems are fully covered in Chapter 5, “Big Idea 4: Computer Systems and Networks.”

Problems of bias are often created by the types and sources of data being collected. Bias is not eliminated by simply collecting more data. A large amount of data is generated by humans. Algorithms that use this data will reflect this bias.

Despite the advantages of big data, a large sample size can magnify the bias associated with the data being used. Data can have little value if the sample is not representative of the population to which the results will be generalized. Computing bias is covered more completely in Chapter 6, “Big Idea 5: Impact of Computing.”

Predicting Algorithms

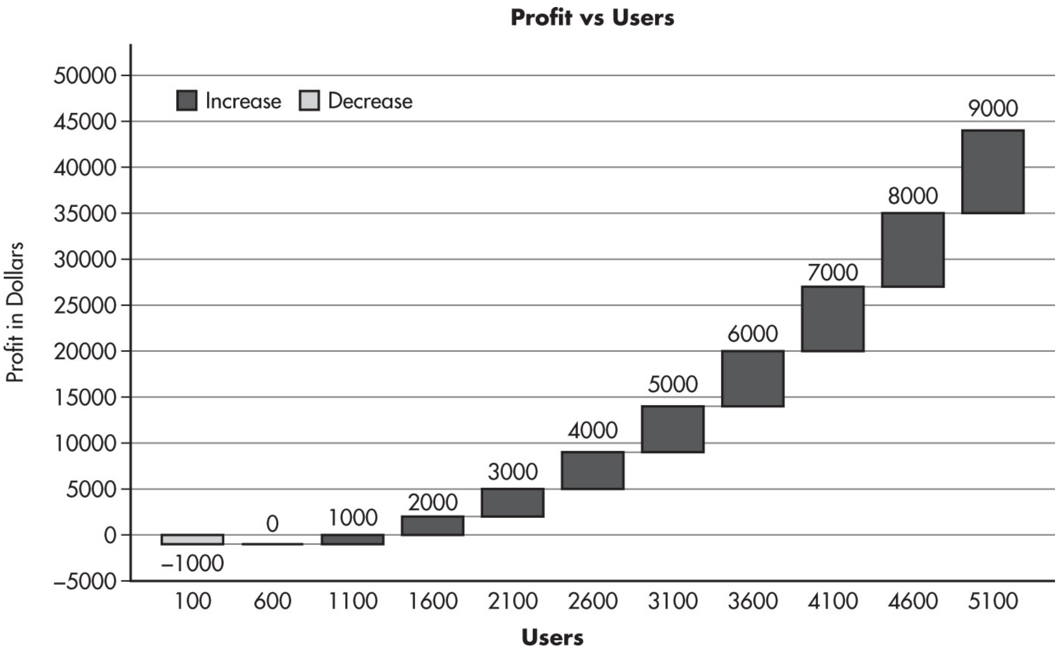
Predicting algorithms use information collected from big data to influence our daily lives. For example:

- A credit card company can use purchasing patterns to identify when to extend credit or flag a purchase for possible fraud.
- Social media sites can use patterns to target advertising based on viewing habits.
- An online store analyzing customers' past purchases can suggest new products the customer may be interested in buying.
- An entertainment application may recommend an additional movie to watch based on the viewer's interests.
- Algorithms can be used to prevent crimes by identifying crime "hot spots." The police can then step up patrols in those areas.

Visualization of Data

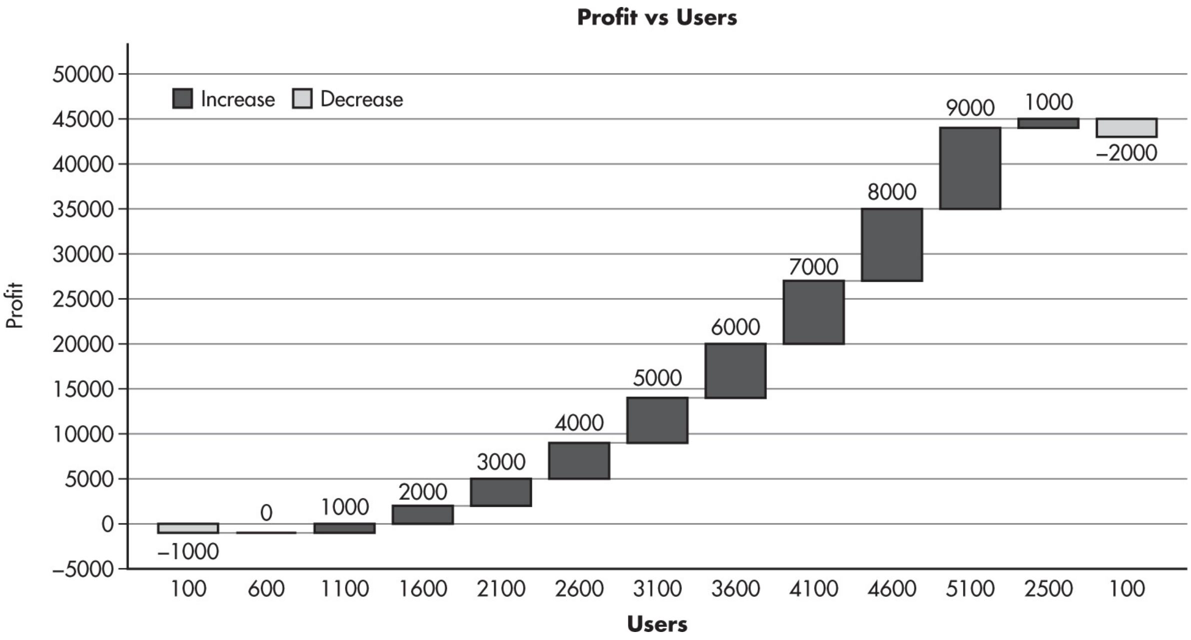
Using appropriate visualizations when presenting digitally processed data can help one gain insight and knowledge. Although big data is a powerful tool, the data will lose their value if they cannot be presented in a way that can be interpreted. Visualization tools can communicate information about data. Column charts, line graphs, pie charts, bar charts, XY charts, radar charts, histograms, and waterfall charts can make complex data easier to interpret.

For example, the graph below plots users versus profit. When looking at the trends from this graph, it looks like a direct relationship exists between the number of users and profit. The company might want to invest in drawing more members or spending on advertisers to draw in new members.



Graph of increasing profits

Predicting trends is not a guarantee of future usage. For example, the above graph cannot predict an innovation that could make this current innovation obsolete. It can be dangerous to draw conclusions based on good data and assume that those conclusions apply across the board or that past patterns will remain consistent. Often, a single source does not contain the data needed to draw a conclusion (see graph below). It may be necessary to combine data from a variety of sources to formulate a conclusion.



Graph of increasing profits with decrease at end of graph.

Predicting algorithms use historical data to predict future events. This data are used to build a mathematical model that encompasses trends. That predictive model is then used on current data to predict what will happen next.

Example Twenty-Four

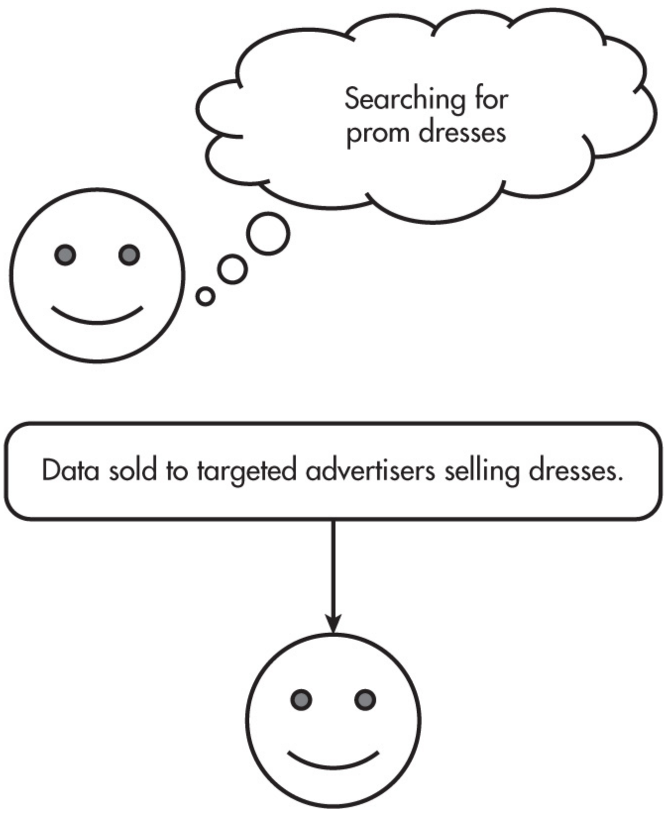
What can be learned from the following data table kept in a pet store?

Date	Pet Food	City	Number of Times Purchased
7/2018	Kibbles	Orlando	10
7/2018	Bill-Jackson	Orlando	2
7/2018	Science Food	Altamonte Springs	23
7/2018	Bill-Jackson	Maitland	37
7/2018	Kibbles	Altamonte Springs	1

Answers:

- The date when a certain dog food was purchased the greatest number of times
- The total number of cities in which a certain food was purchased
- The total number of foods purchased in a certain city during a month

E-commerce sites use data to determine how much inventory to hold and how to price products. Additionally, data about product views and purchases power the recommendation engine, which drives a large portion of sales. Data allow for personalized and effective advertisement. Sometimes an e-commerce site knows what you want to buy before you do.


Searching for prom dresses
Data sold to targeted advertisers selling dresses.

Targeted advertisers

Example Twenty-Five

A high school principal is interested in predicting the number of students passing a state-level exam. She created a computer model that uses data from third-party software showing an increasing student pass rate for the exam. The model provided by the software company predicts a 90% student pass rate. The actual percentage of students passing the state exam was 74%. When creating a model, all real-world variables cannot be represented. In the case of a model not accurately predicting outcomes, addition information can be added to make a more accurate prediction. What are some possible additions to the model to make it more reliable in predicting the student pass rate?

Answer:

- Refine the model to include data from more sources other than the third-party software due to the financial interest in the software being used.
- Refine the model to include student data from other schools.

- Refine the model to include information about the community, such as redistricting.

[redacted]

[redacted]

[redacted]

[redacted]