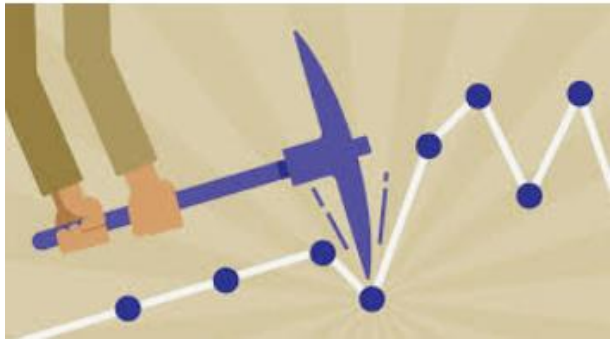
Data Management

Business Intelligence and Data Mining

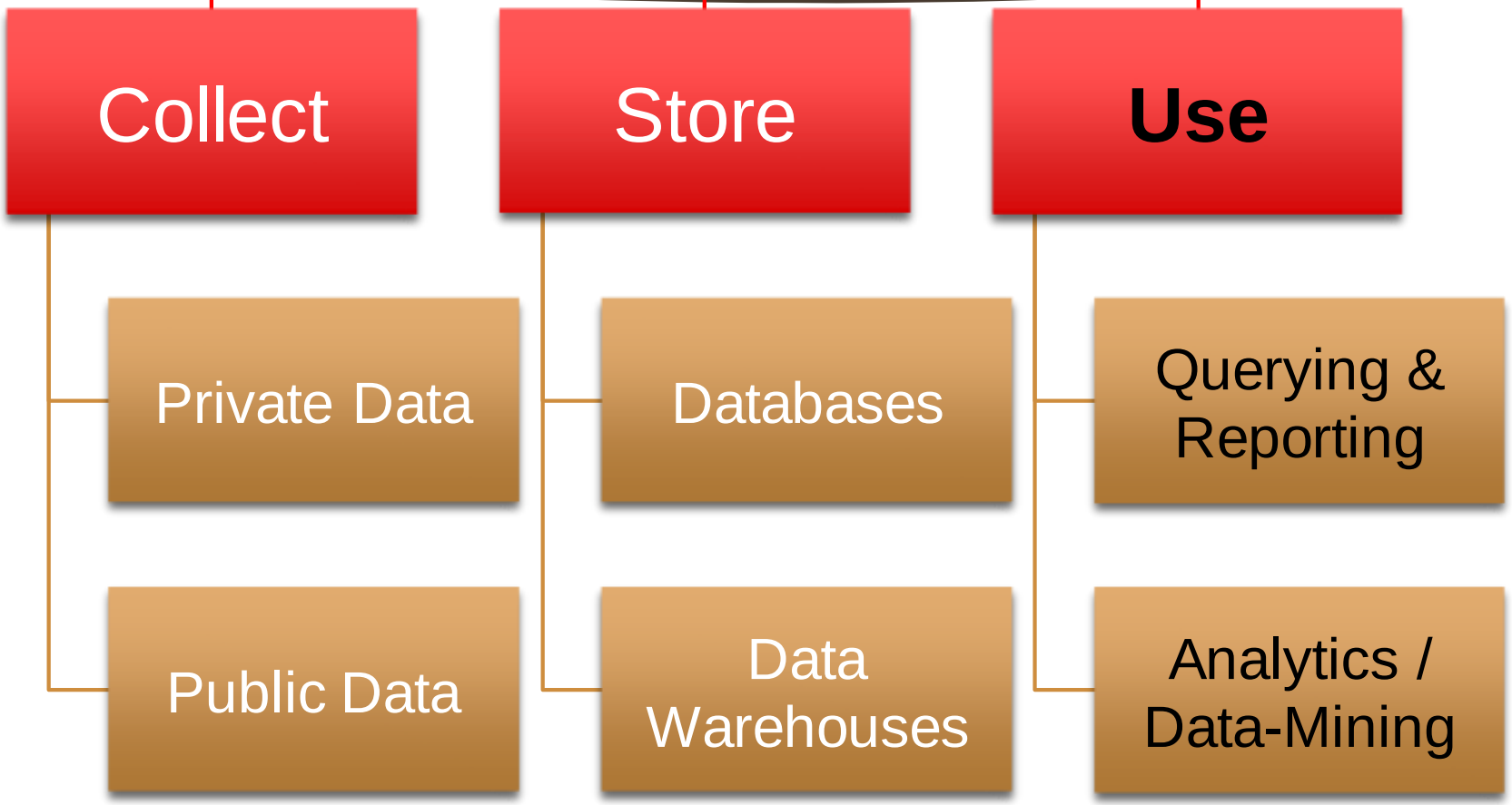


Ref. Ch. 12

Learning Objectives

- Learn the concepts of data warehouses and data mining
- Understand need for analysis data
  - Query and reporting versus data mining
  - Analysis versus predictive
  - Multiple dimensions
- Understand skills and techniques needed for data analytics
- Appreciate role of data analysis in today's society

Data Management



Recall:

Figure 12-1 Storage Capacity Terms

Name	Symbol	Approximate Value for Reference	Actual Value
Byte			8 bits [Store one character]
Kilobyte	KB	About 10^3	2^{10} = 1,024 bytes
Megabyte	MB	About 10^6	2^{20} = 1.024 KB
Gigabyte	GB	About 10^9	2^{30} = 1,024 MB
Terabyte	TB	About 10^{12}	2^{40} = 1,024 GB
Petabyte	PB	About 10^{15}	2^{50} = 1,024 TB
Exabyte	EB	About 10^{18}	2^{60} = 1,024 PB
Zettabyte	ZB	About 10^{21}	2^{70} = 1,024 EB
Yottabyte	YB	About 10^{24}	2^{80} = 1,024 ZB

Implications?
Amount of data being captured and used increases.
Need to manage the data asset.

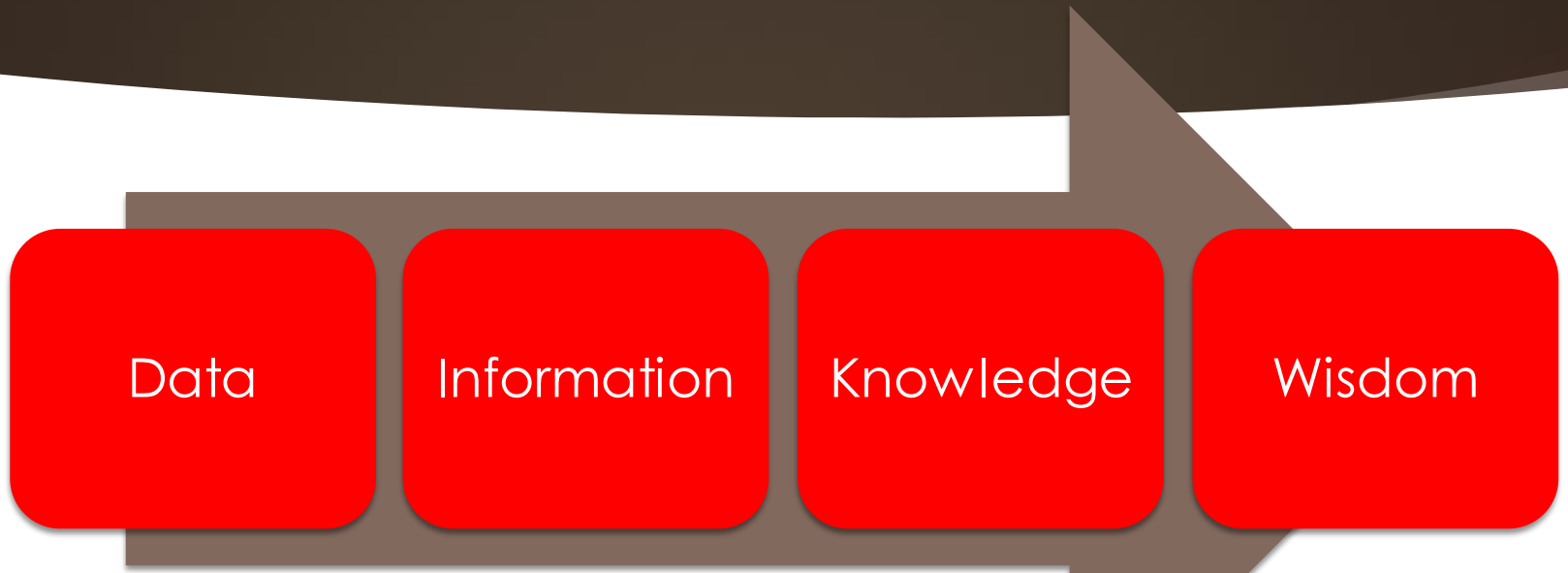
From Storage to Use

Now that we have gathered and organized so much data, what do we do with it?

"The secret of business is to know something that nobody else knows."

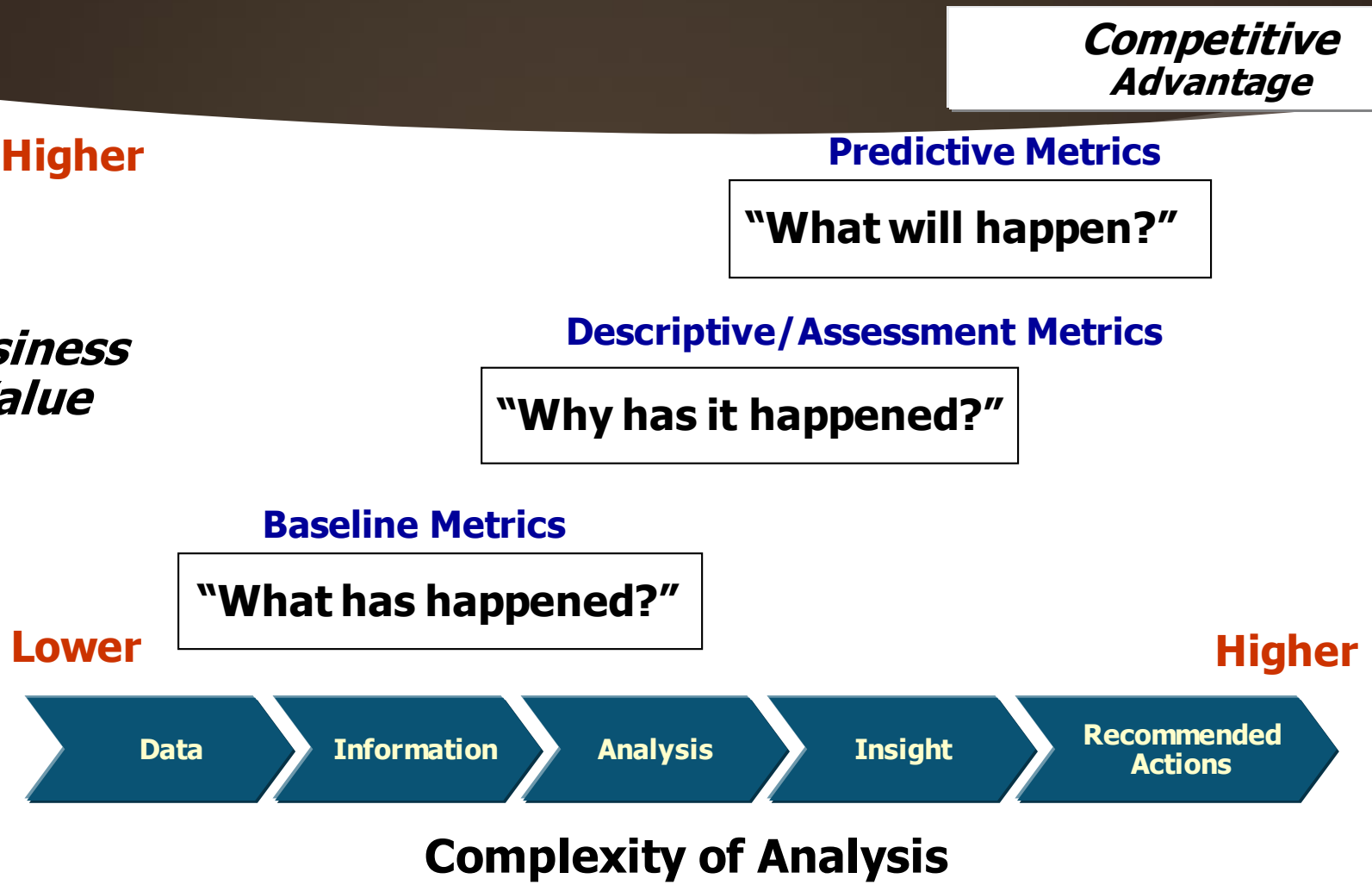
Aristotle Onassis

Recall: This course is about data. What is data(revisited)?



- Data constitute the building blocks of information.
- Information is produced by processing data.
- Information is used to reveal the meaning of data.
- Accurate, relevant, and timely information is the key to good decision making.
- Good decision making is the key to organizational survival in a global environment.

Important: Business Case for Business Intelligence



Source: www.eforceglobal.com

Business Intelligence (BI) Systems [Text]

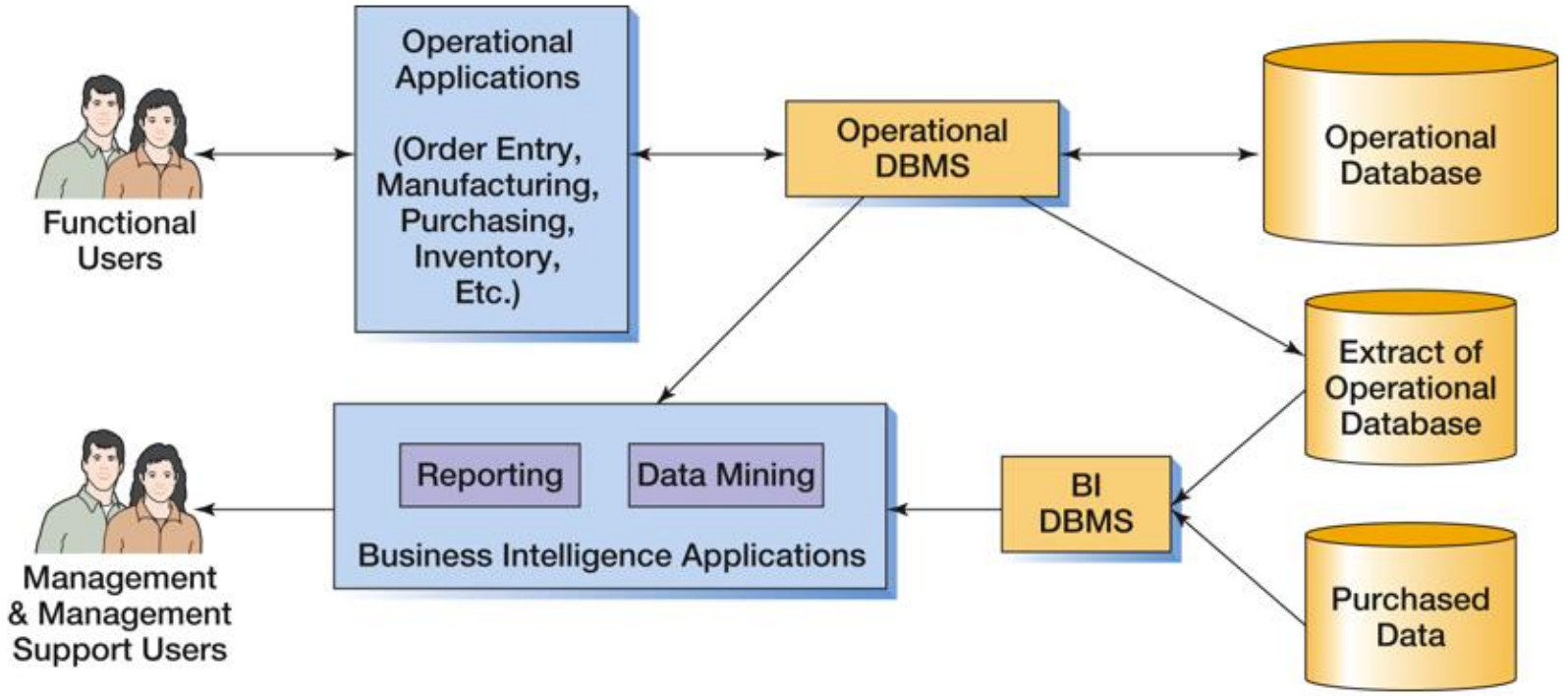
- **Business intelligence (BI)** systems are information systems that assist managers and other professionals:
  - To analyze **current** and **past** activities.
  - To predict future events
- Two broad categories:
  - **Reporting**
  - **Data mining**

Business Intelligence
Source: Wikipedia [Broader view]

- **Business Intelligence** (**BI**) comprises the strategies and technologies used by enterprises for the data analysis of business information. BI technologies provide historical, current and predictive views of business operations. Common functions of business intelligence technologies include reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining, predictive analytics and prescriptive analytics.

- BI technologies can handle large amounts of structured and sometimes unstructured data to help identify, develop and otherwise create new strategic business opportunities. They aim to allow for the easy interpretation of these big data. Identifying new opportunities and implementing an effective strategy based on insights can provide businesses with a competitive market advantage and long-term stability.

Figure 12-2 Relationship Between Operational and BI Systems



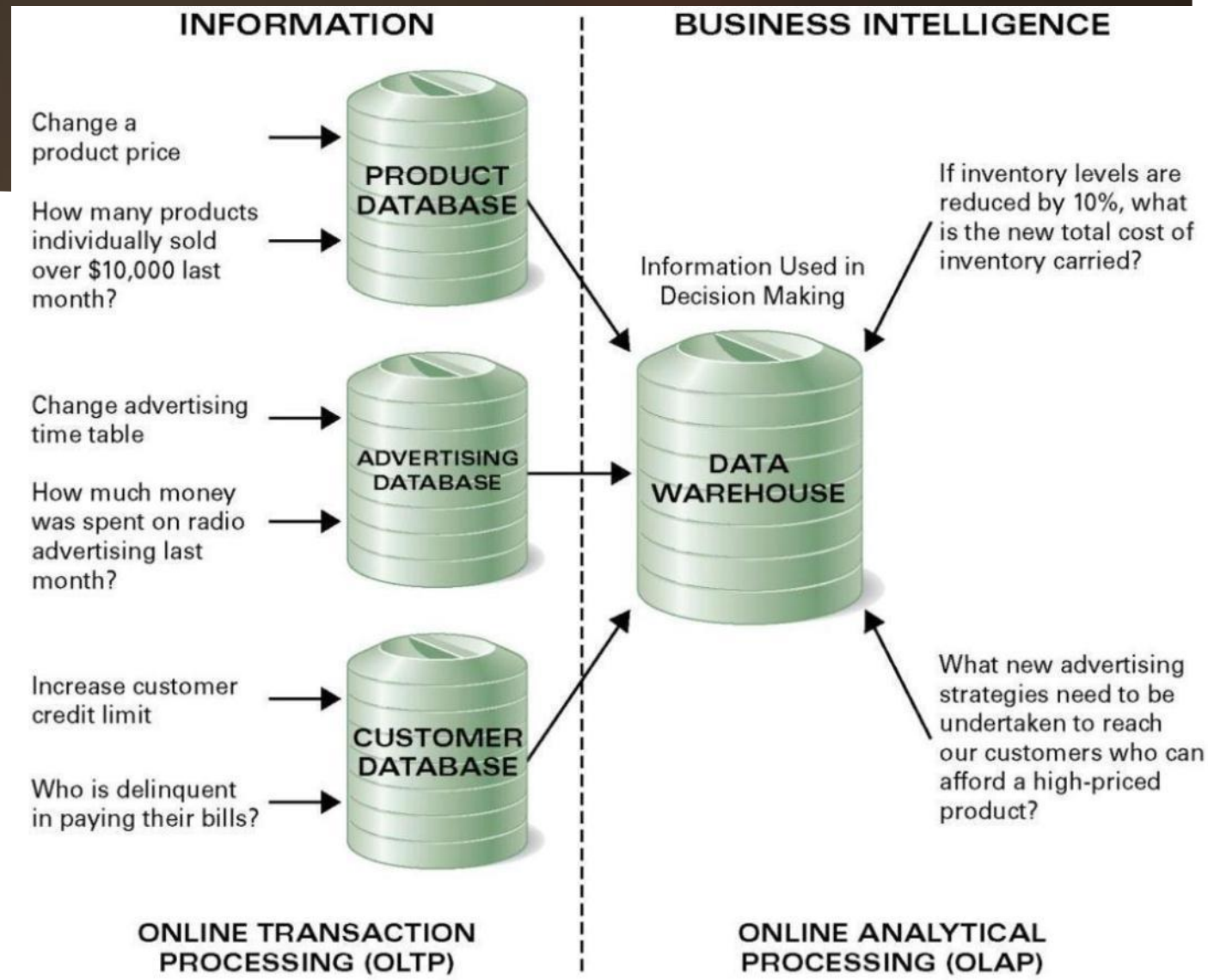
OLTP and OLAP

Online transaction processing (OLTP):

- gather data, process it, and update information
- Operational DBs support OLTP

Online analytical processing (OLAP):

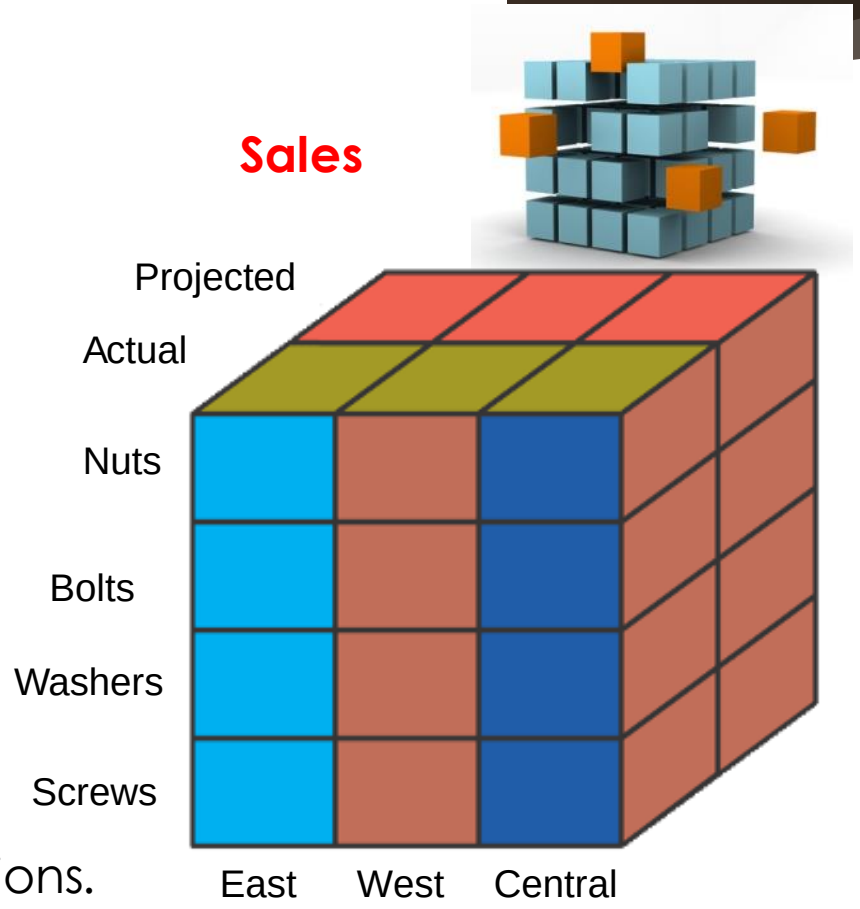
- manipulation of data to support decision making
- Data Warehouses support OLAP



Source: McGraw Hill

Online Analytical Processing (OLAP)

- **Functions**
  - Sum, count, average, etc.
- **OLAP report**
  - **Measure:** data item of interest
    - Total sales, average sales, average cost
  - **Dimension:** characteristic of a measure
    - Purchase date, customer type
    - Customer location

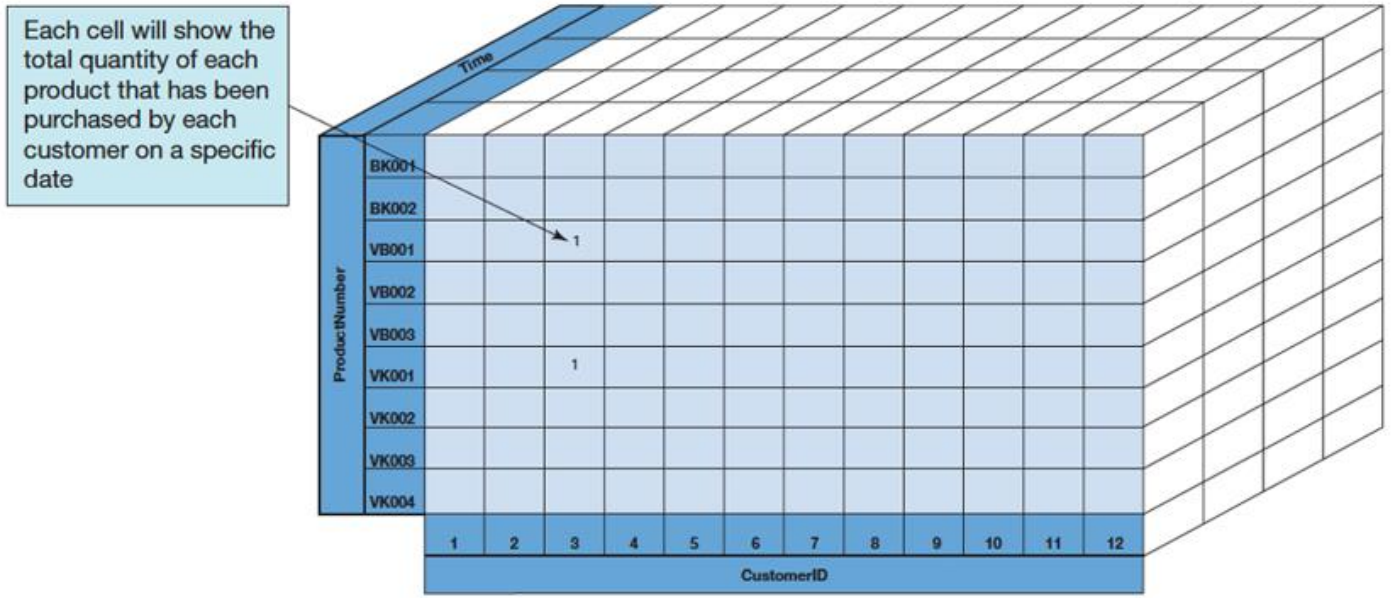


OLAP Cube

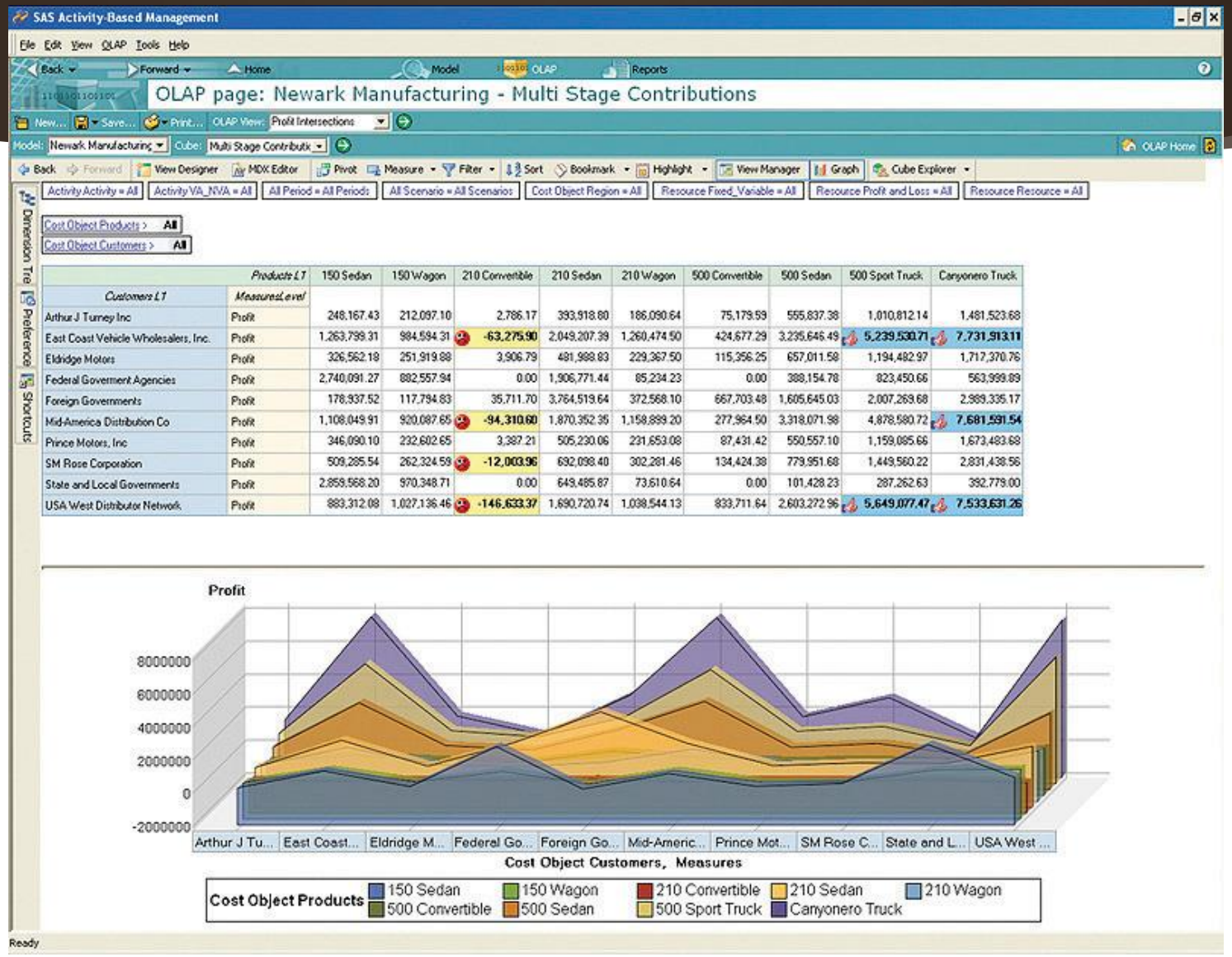
A presentation of a measure with associated dimensions.

– An OLAP cube can have *any* number of axes.

Figure 12-17: Three-Dimensional: Time-ProductNumber-CustomerID Cube



OLAP report compares multiple dimensions. Company is along the vertical axis; product is along the horizontal axis. Many OLAP tools can present graphs of multidimensional data.



Data Mining

- Includes:
  - Identifying valid, novel, potentially useful, and ultimately understandable patterns in data
  - Searching for relationships, patterns, and trends not known to exist or not visible
  - Providing answers to questions decision maker not thought to ask
- Requires:
  - Information technology
  - Statistics
  - Business knowledge

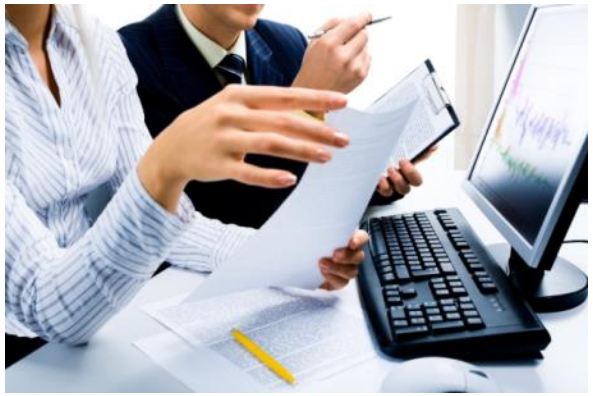


Figure 12-25 Convergence of Disciplines for Data Mining

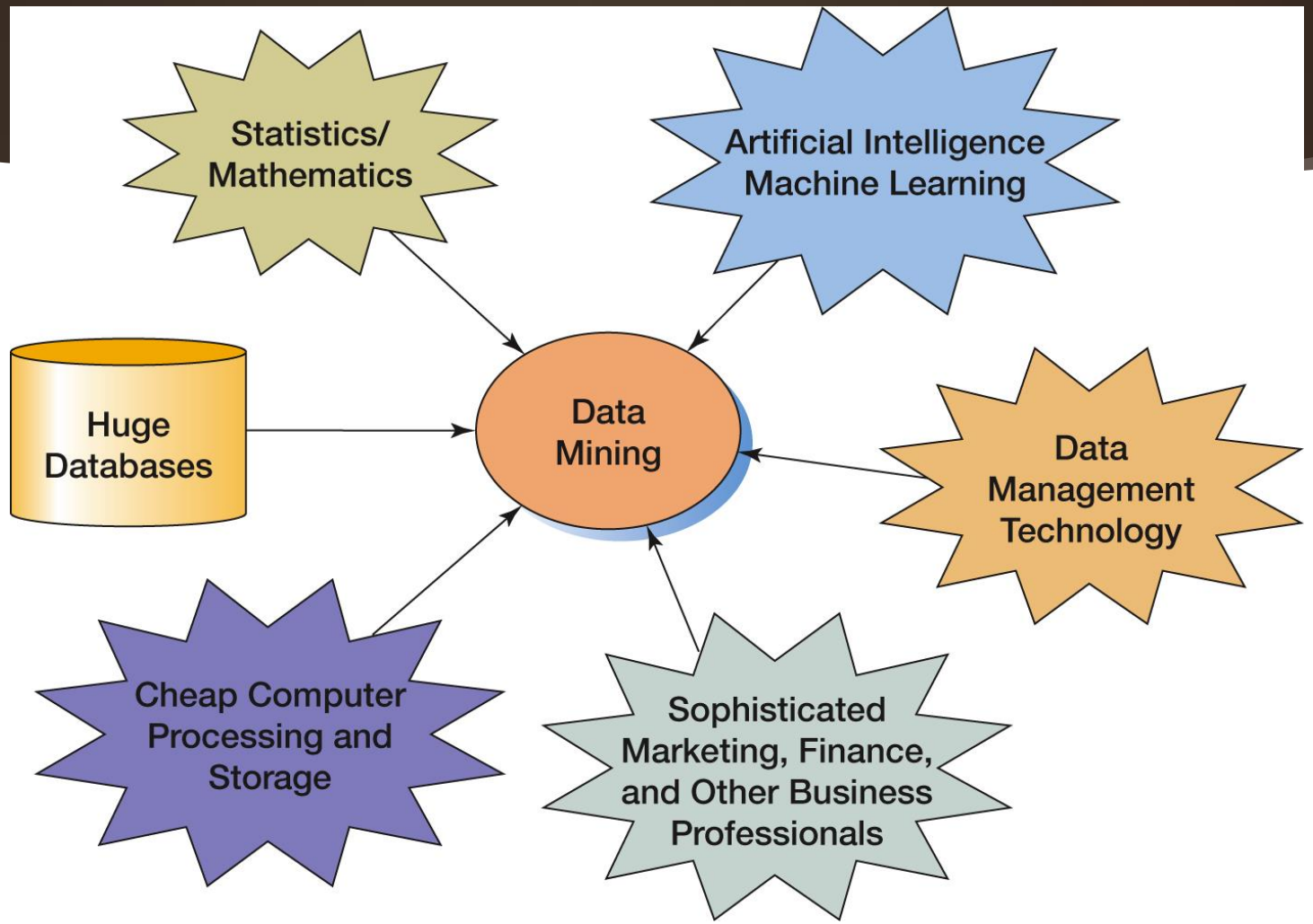
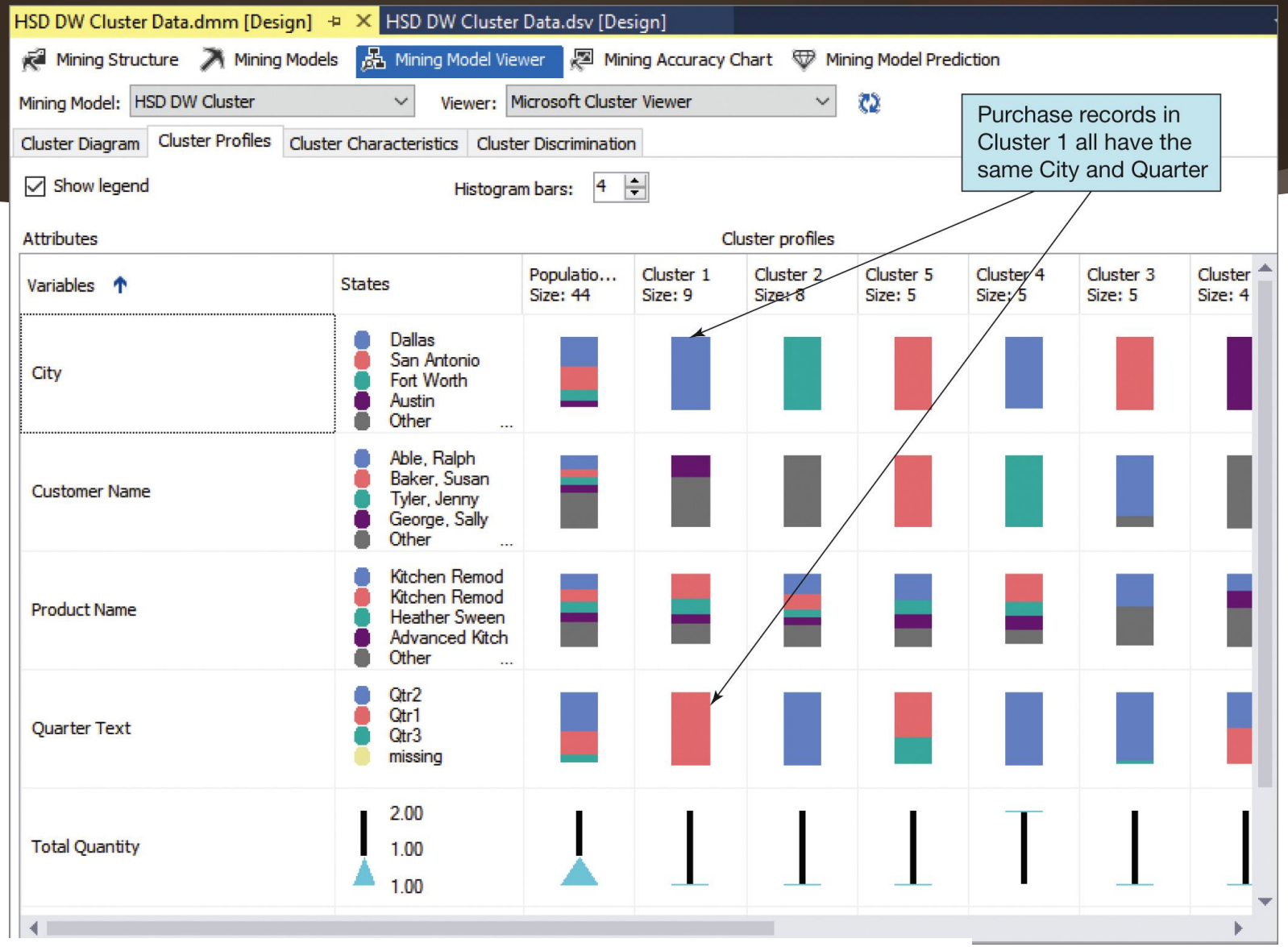


Figure 12-26 Clustering in SQL Server Analysis Services



Why Mine Data?

- **Great deal of data collected and warehoused**
  - Web data
  - Purchases (retail, grocery)
  - Bank/Credit Card transactions
  - Exhaust data
- **Powerful processing**
- **Competitive pressure**
  - Better, customized services for potential advantage
  - Customer Relationship Management

"Data is the fuel that marketers run on."

Northwestern University Professor

Source: Tan et al.

Data Mining: Applications
[And many more]

- **Marketing and Promotion Targeting**
  - Prospects for e-mailing list
- **Customer Segmentation**
  - Common characteristics of customers who buy same products
- **Market Basket Analysis**
  - Which products likely to be bought together
- **Customer Churn**
  - Which customers likely to leave
- **Fraud Detection**
  - Patterns of fraudulent transactions; compare current transactions
- **Collaborative Filtering**
  - Personalization based on similar customers
- **Financial Modeling**
  - Trading systems based upon historical data
- **Hiring and Promotion**
  - Based upon employee characteristics

Source: T. Melone, Gallaugher

Data Mining: Techniques and Tasks

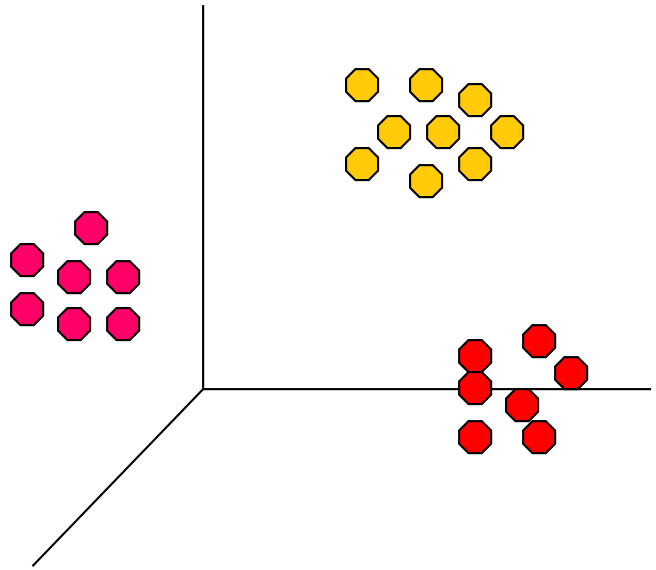
- **Descriptive/Assessment** Techniques [**patterns**]
  - Clustering: data segmentation
  - Association Rule Discovery: market-basket analysis
  - Sequential Pattern Discovery (A then B)
  - Characterization: generalization / summarization
- **Predictive** Techniques [**predict**]
  - Classification: categories
  - Regression
  - Deviation Detection

Source: Tan et al.

Clustering

- Data points in common cluster are similar
- Data points in different clusters are dissimilar

Intra-cluster distances are minimized



Inter-cluster distances are maximized

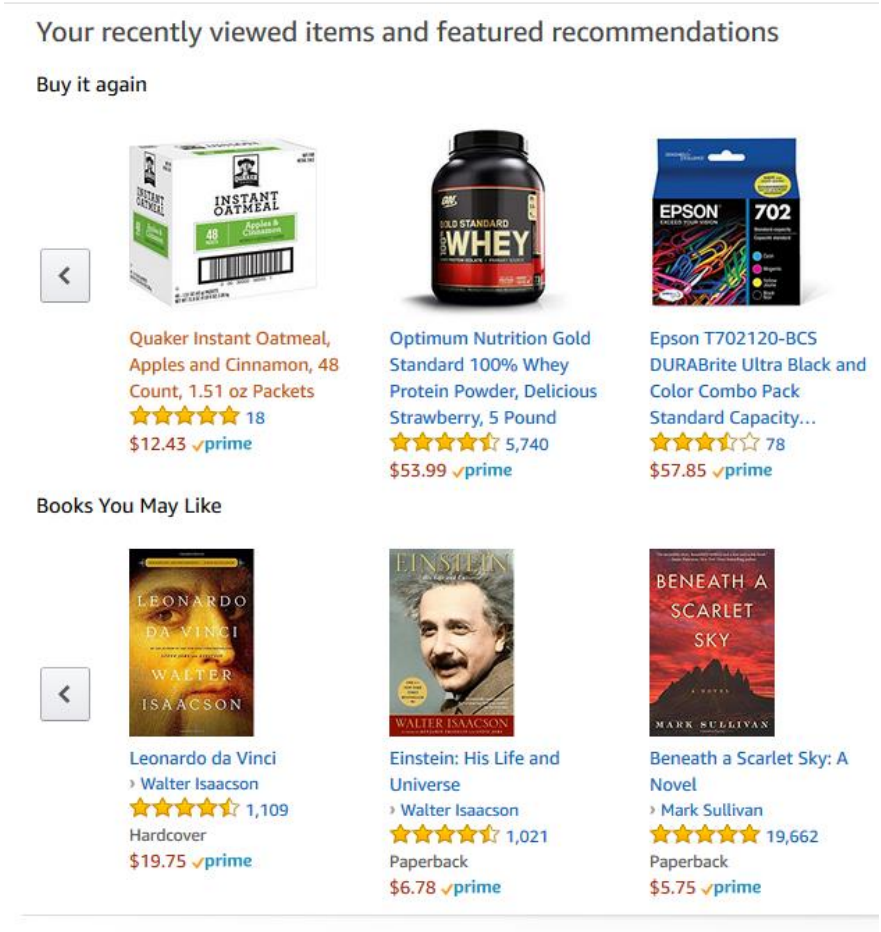
Source: Tan et al.

Association: Market-Basket Analysis

Customer buying behavior

- which products customers tend to buy together
- probability of customer purchase
- cross-selling opportunities
  - "Customers who bought X also bought Y"
  - Recent purchases
  - What you may like

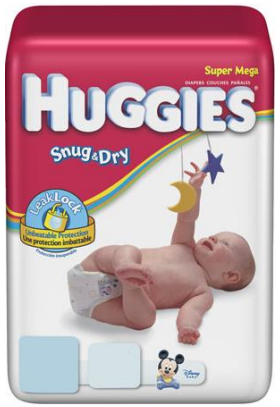
[



Association Rule: Discovery

Supermarket Shelf Management

- Goal
  - Identify items bought together by sufficient number of customers
- Approach
  - Process point-of-sale data from barcode scanners
  - Identify dependencies among items (patterns)
- Classic rule
  - If customer buys diapers and milk, then that customer very likely to buy beer



Source: Tan et al.

Predictive: Customer Churn

Predict whether customer likely to be lost to competitor [Classification]

- Example (retail): Find model for loyalty
  - Find customer attributes from past/present transaction data
    - E.g., how often customers purchase, where made purchases from, time of the day purchases made, etc.
  - Label customers loyal or disloyal

Source: Tan et al.

Predictive Technique: Regression

Predict value of a given variable based on values of other variables

- **Examples**
  - **Sales of new product based on advertising, location, etc.**
  - **Wind velocities as a function of temperature, humidity, air pressure, etc.**

[Regression Equation y = a + bx]

X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0).

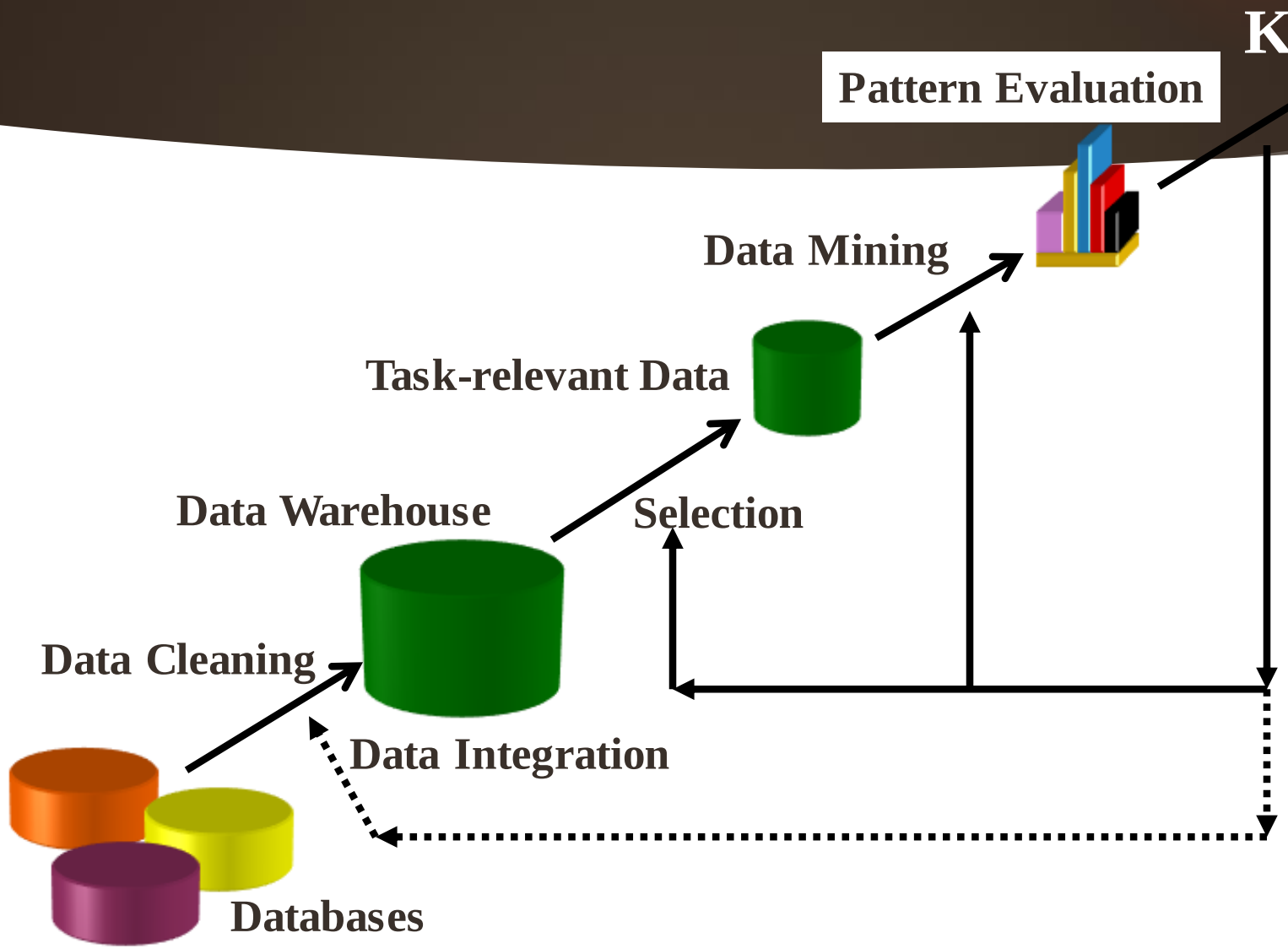
Source: Tan et al.

Data Mining Pitfalls

- Wrong estimates from bad data
  - overexposed to risk
- Models not effective when market does not behave as in past
  - Data mining: regularities from history
- Pattern uncovered;
  - best choice for response less clear
  - association does not dictate trend nor causality
- Over-engineering
  - Model has so many variables, solution might only work on subset of data used to create it

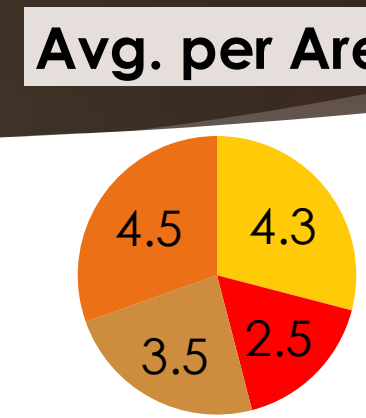
Source: Gallaugher

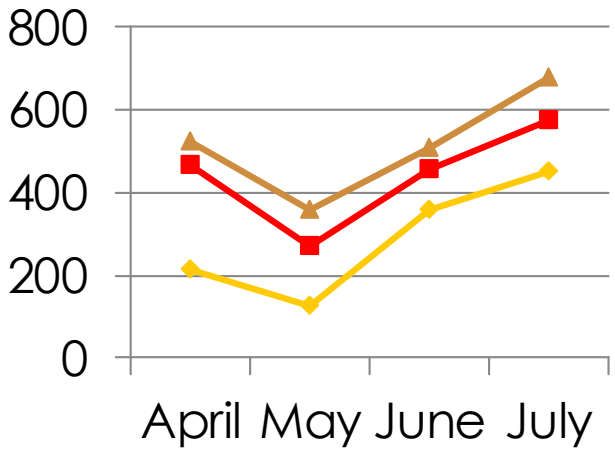
Important: Knowledge Production

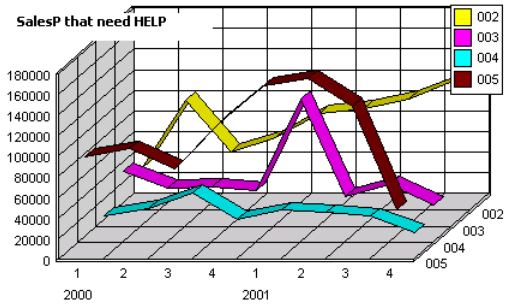


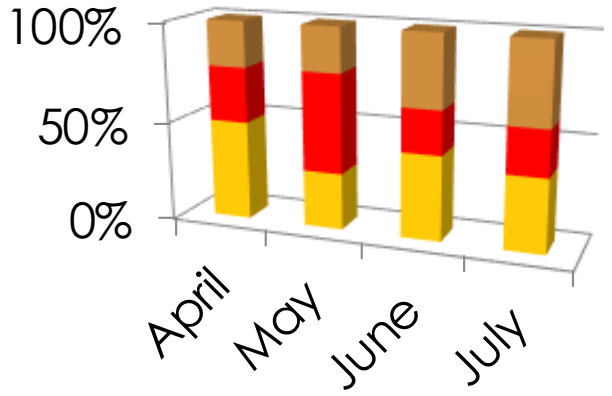
Representation: Visualization

	Cars	Trucks	Buses
April	24,024	12,408	1,201
May	21,585	8,502	1,842
June	19,684	10,582	2,022
July	27,254	15,206	2,145





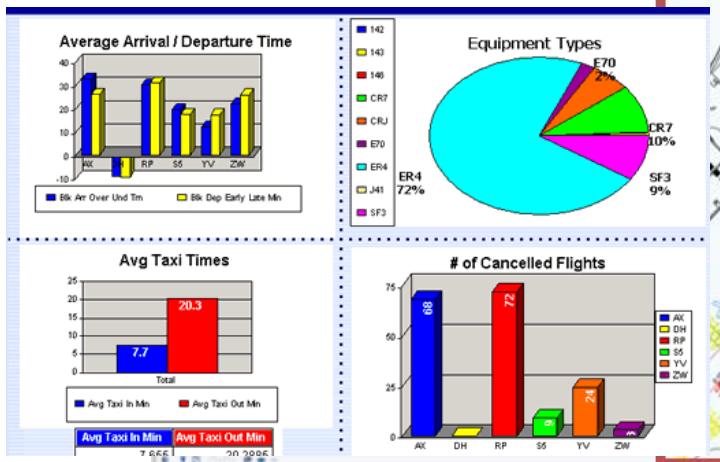


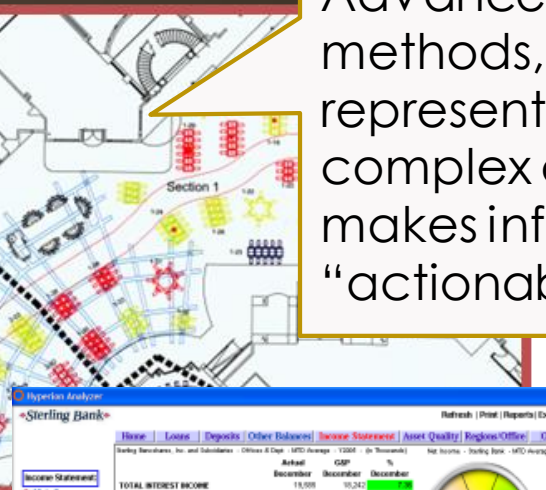


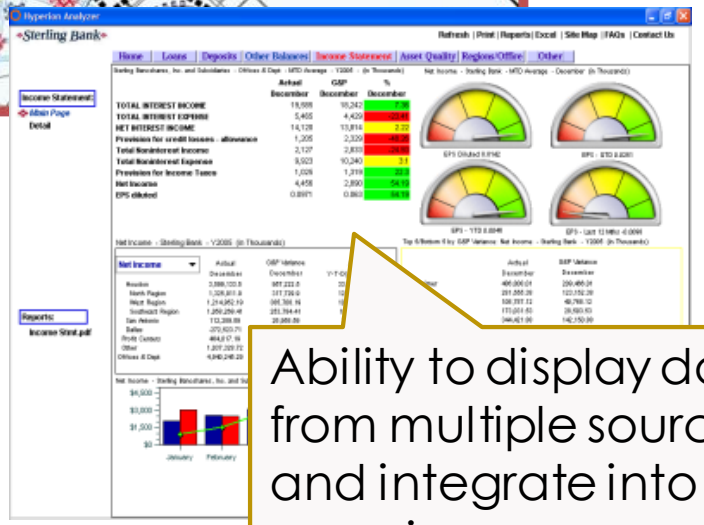
Business Dashboards: full-fledged performance management tools

Display of critical indicators

Managers: graphical glance at key performance metrics







Advanced visualization methods, with ability to represent large, complex data sets, makes information "actionable"

Pre-defined navigation paths to facilitate analysis and improve user experience

Ability to display data from multiple sources and integrate into one view

Source: Watson

Management Issues

- Data asset
  - Manage quantity and quality
  - Protect and secure
- Appropriate use of data
  - Privacy
  - Ethics
- Business impact
  - Exploit to provide value

Application: Inventory Management

The right item, at the right store, at the right time and at the right price

Walmart
Save money. Live better.

"Our business strategy depends on detailed data at every level - every cost, every line item is carefully analyzed enabling better merchandising decisions to be made on a daily basis. It is the foundation for maintaining Wal-mart's competitive edge and its continuing success in providing every day low prices and superior customer satisfactions."

Randy Mott, Wal-Mart

http://www.autonews.com/article/20171002/OEM06/171009988/randy-mott-gm-it-architect

Hired IT graduates and outsourced them within US for GM.

Data-driven decision making

- Increasingly standardized corporate data and access to rich, third-party datasets; all leveraged by cheap, fast computing and easy-to-use software; enabling age of data-driven, fact-based decision making
- Big data: General term used to describe massive amount of data available to today's managers.
- Business intelligence (BI): Term combining aspects of reporting, data exploration and ad hoc queries, and sophisticated data modeling and analysis
- Analytics: Term describing the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions

Source: Gallaugher

Conclusion

- *Data Management*
  - Crucial part of business
- *Collect data*
  - Internal and external sources
- *Represent and Store data*
  - Database management systems
  - Data warehouses
- *Use data*
  - Business intelligence
  - Data mining and marketing applications
  - Operational efficiencies

More coming . . .

Georgia State University
J. Mack Robinson
COLLEGE of BUSINESS