



**THE
POWER
TO KNOW®**

Data Mining

Sue Walsh
Higher Education Consulting
SAS

Overview

- Brief Historical Perspective
- Defining Data Mining
- Issues
 - Data Collection and Data Organization
 - Modeling Issues and Data Difficulties
 - Skepticism and Communication
- Applications
- SAS Enterprise Miner Demonstration
- SAS Enterprise Miner versus SAS/STAT
- Another Kind of Data Mining - Text Mining

History

Data Mining, circa 1963

IBM 7090

600 cases

“Machine storage limitations restricted the total number of variables which could be considered at one time to 25.”

IBM 7090

Since 1963

- **Moore's Law:**

The information density on silicon-integrated circuits doubles every 18 to 24 months.

- **Parkinson's Law:**

Work expands to fill the time available for its completion.

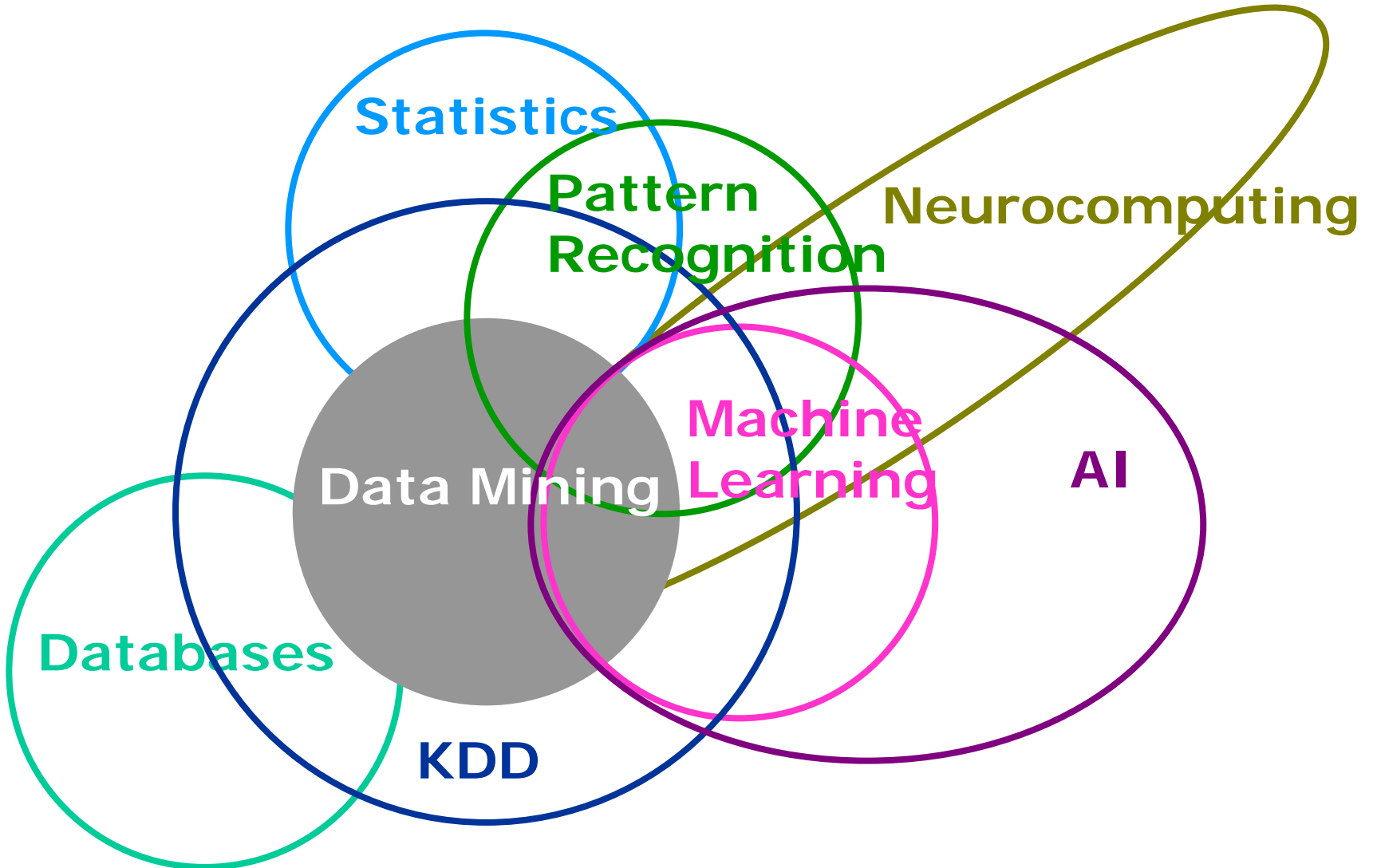
Data Deluge

hospital patient registries
electronic point-of-sale data
stock trades OLTP telephone calls
catalog orders bank transactions
remote sensing images tax returns
airline reservations credit card charges

The Data

	<u>Experimental</u>	<u>Opportunistic</u>
Purpose	Research	Operational
Value	Scientific	Commercial
Generation	Actively controlled	Passively observed
Size	Small	Massive
Hygiene	Clean	Dirty
State	Static	Dynamic

The Origins of Data Mining



Solving the Data Puzzle - a Step-by-Step Approach

- Data collection
 - Transactional systems
 - Customer information systems
- Data organization
- Data analysis
- Reporting

The Result → → → Business Decisions

Definition

What Is Data Mining?

- **IT**
 - Complicated database queries
- **ML**
 - Inductive learning from examples
- **Stat**
 - What we were taught not to do

Data Mining – The SAS Definition

Advanced methods for exploring and modeling relationships in large amounts of data.

Solving the Data Puzzle - a Step-by-Step Approach

- Data collection
 - Transactional systems
 - Customer information systems
- Data organization - data warehousing
- Data analysis - data mining
- Reporting
- Action

The SAS Approach to Data Mining SEMMA

- Sample
- Explore
- Modify
- Model
- Assess

Issues

Data Collection and Data Organization

- What data has been collected and where is it?
- How do I combine legacy systems with current data systems?
 - Customer Story
- What is the meaning of some of these data values?

Modeling Issues and Data Difficulties

- Data Preparation
- Rare or Unknown Targets
 - Over Sampling
- Undercoverage
- Dirty Data
 - Errors
 - Missing Values
- Dimension Reduction (Variable Selection)
- Under and Over Fitting
- Temporal Infidelity
- Model Evaluation

Skepticism and Communication

- Skepticism
 - Breaking the Rules (statisticians)
 - Magic (non-analytical individuals)
- Communication

Applications

Health Care

- Drug development – to help uncover less expensive but equally effective drug treatments.
- Medical diagnostics – imaging, real-time monitoring (e.g., predicting women at high risk for emergency C-section).
- Insurance claims analysis – identify customers likely to buy new policies; define behavior patterns of risky customers.

Business and Finance

- Banks - to detect which customers are using which products so they can offer the right mix of products and services to better meet customer needs – cross sell and up sell.
- Credit card companies - to assist in mailing promotional materials to people who are most likely to respond.
- Lenders - to determine which applicants are most likely to default on a loan.

The Absa Group (a South African Bank)

- **Challenge:**
Reduce operating expenses and cut losses by leveraging data to improve security and enhance customer relationships.
- **Solution:**
SAS helped Absa reduce armed robberies by 41 percent over two years, netting a 38 percent reduction in cash loss and an 11 percent increase in customer satisfaction ratings.

Sports and Gambling

- Sports teams – to analyze data to determine favorable player matchups and call the best plays
- Gaming industry - to analyze customer gambling trends at casinos.
- Sports Fanatics – to predict which teams will be chosen for tournament berths as well as to predict game winners.

Education

- Enrollment Management – which students are likely to attend
- Retention/Graduation Analysis – which students will remain enrolled after the first year and/or through graduation
- Donation Prediction – who is likely to donate and how much might they donate
- Faculty Churn – what faculty members are most likely to leave the institution

Other Application Areas

- Insurance – pricing, fraud detection, risk analysis
- Stock Market – market timing, stock selection, risk analysis
- Transportation – performance & network optimization to predict life-cycle costs of road pavement
- Telecommunications – churn reduction
- Retail – market basket analysis to help determine marketing strategies



Demonstration

Data Mining with SAS Enterprise Miner versus with SAS/STAT

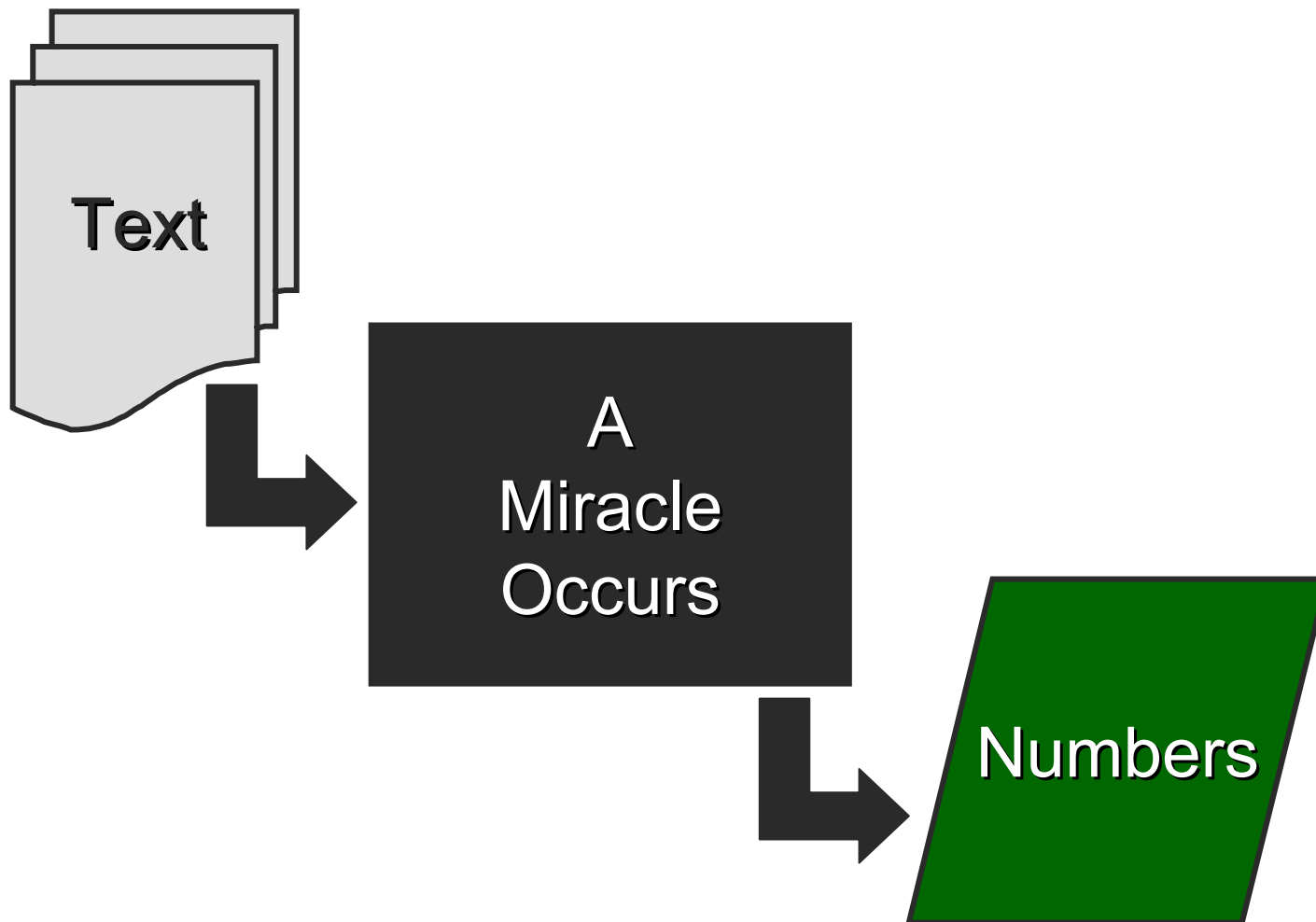
- Features in SAS Enterprise Miner not in SAS/STAT
 - Decision trees
 - Neural networks
 - Automatic data splitting
 - Automatic score code
 - Model comparison tool
- Features in SAS/STAT not in SAS Enterprise Miner
 - Diagnostic statistics
- The products offer different model evaluation statistics because of the difference in purpose.

Another Kind of Data Mining

Text Mining – What is it?

- Text mining is a process that employs a set of algorithms for converting unstructured text into structured data objects and the quantitative methods used to analyze these data objects.
- “SAS defines text mining as the process of investigating a large collection of free-form documents in order to discover and use the knowledge that exists in the collection as a whole.” (*SAS[®] Text Miner: Distilling Textual Data for Competitive Business Advantage*)

Another View of Text Mining



Text Mining Applications

- Automotive Early Warning System
 - Wallace and Cermack (2004) describe the use of text mining for warranty analysis related to the TREAD act.
- Medical Information Management
 - TextWise Labs uses sophisticated text mining methodology to extract medical information from disparate data sources on the Internet.
 - Computer Science Innovations Inc. is developing an application for the National Cancer Institute that automatically converts medical records into XML data.

Text Mining Applications

- Insurance Claim Fraud
 - Insurance companies employ Special Investigative Units (SIU) to investigate claims for fraud. Data mining methods can be employed to automate the process of referral. Text mining methods are applied to claims examiner notes, physician reports, and other textual data to enhance predictive accuracy.

- Technical Support
 - Sanders and DeVault (2004) describe a process that employs text mining to improve efficiency in a technical support environment.