Quantification of Operational Risk Using Data Mining Methodologies

J. Sebastian Perera, J.D.
National Aeronautics & Space Administration
Lyndon B. Johnson Space Center
Houston, Texas 77058-3696

Outline

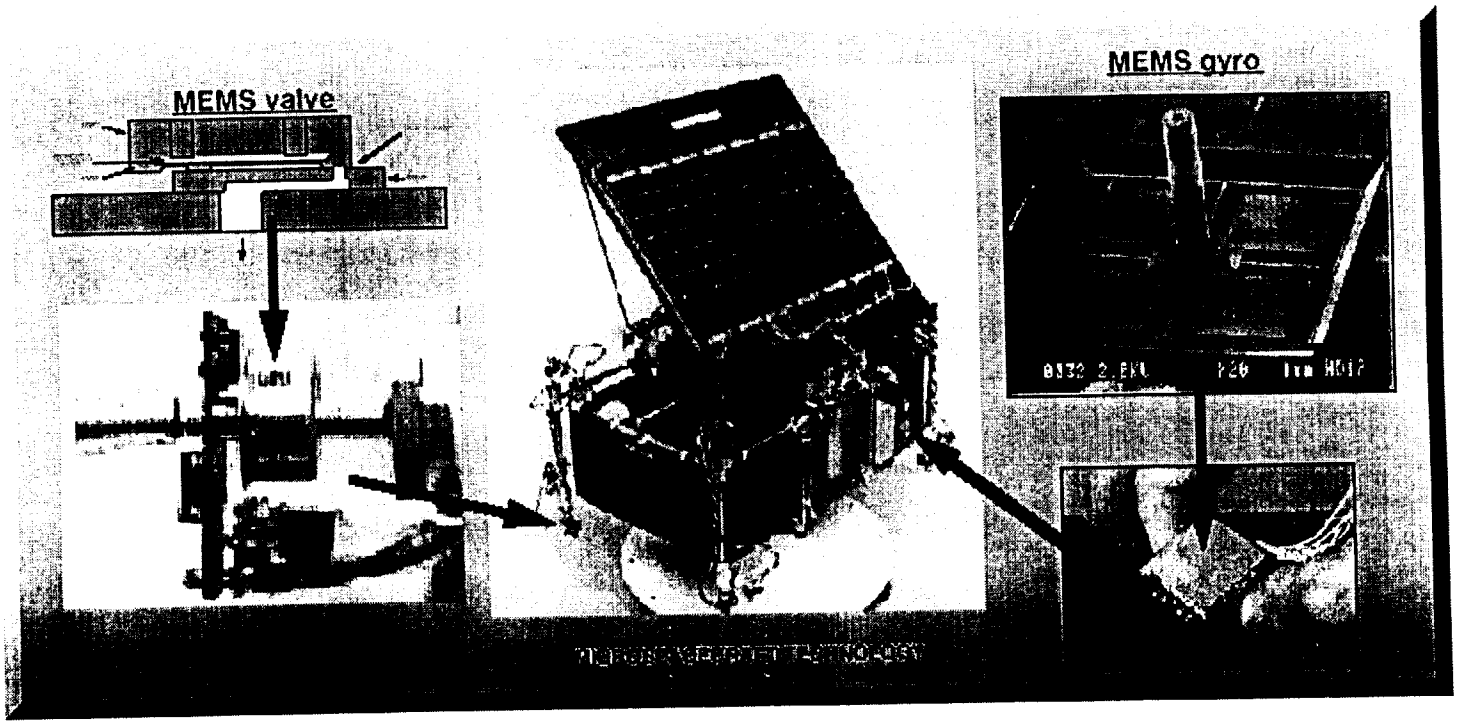
- Introduction
- Prior Applications
- Methodology
- Data Mining Process
- Types of Data Mining Schemes
- A Generic Example
- Practical Considerations
- Summary

Introduction

- What is Data Mining?
  - Data Mining is the process of finding actionable information hidden in raw data.
  - Data Mining helps find hidden patterns, trends, and important relationships often buried in a sea of data
  - Typically, automated software tools based on advanced statistical analysis and data modeling technology can be utilized to automate the data mining process

Prior Applications

- Micro-Electromechanical Systems - Small Mechanical/Electrical Devices and Systems
  - Significant positive attributes for space-based use
- A Reliability Tool was Required for Conceptual Design/Planning
  - Develop a flexible, dynamic and robust reliability model for these miniature devices



Methodology

- Create a Flexible Risk Estimation Scheme for Predicting Operational Risk. Must be:
  - Dynamic
  - Robust
  - Relevant
  - Objective
  - Complete
- Model Extrapolates Operational Risk from Previous and Current Risk Attributes and Operational Risk Data:
  - Loss or Damage to Asset
  - Lawsuits
  - Taxation
  - Model Risk
  - Human Error
  - Technology
  - Theft/Fraud
  - Rogue Trading
  - Transactions
  - Customer Satisfaction
  - Regulatory
  - Any Other User-Defined Attributes
- Validate Operational Risk Model

Data Mining Process

- Identify
  - Choose the data sets to be used for training and testing, specify the "input" variables and "output" variables
  - Segment system domain into enterprises (Financial, Manufacturing, etc.)
- Transform
  - Convert original data through mapping, sampling, and feature extraction
  - Compensate for outliers and sparse regions
  - Convert non-numeric data to numeric
- Model
  - Apply one of several algorithms/schemes to learn mathematical relationships among inputs and outputs
- Analyze
  - Define performance expectations and judge model performance

Data Mining Process

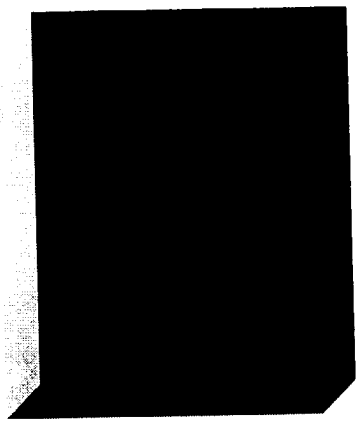
Identify

Audit Reports
Management Reports
External Reports
Management Interviews
Expert Opinion
Loss History
Other Reports/ Documents

Transform *(Numerics)*

Loss or Damage to Asset
Lawsuits
Taxation
Model Risk
Human Error
Technology
Theft/Fraud
Rogue Trading
Transactions
Customer Satisfaction
Regulatory
Any Other User-Defined Attributes

Model



Analyze

Operational Risk Estimate

Output Parameters providing characterization of model performance (i.e. r^2, Absolute Error, Squared Error, etc.)

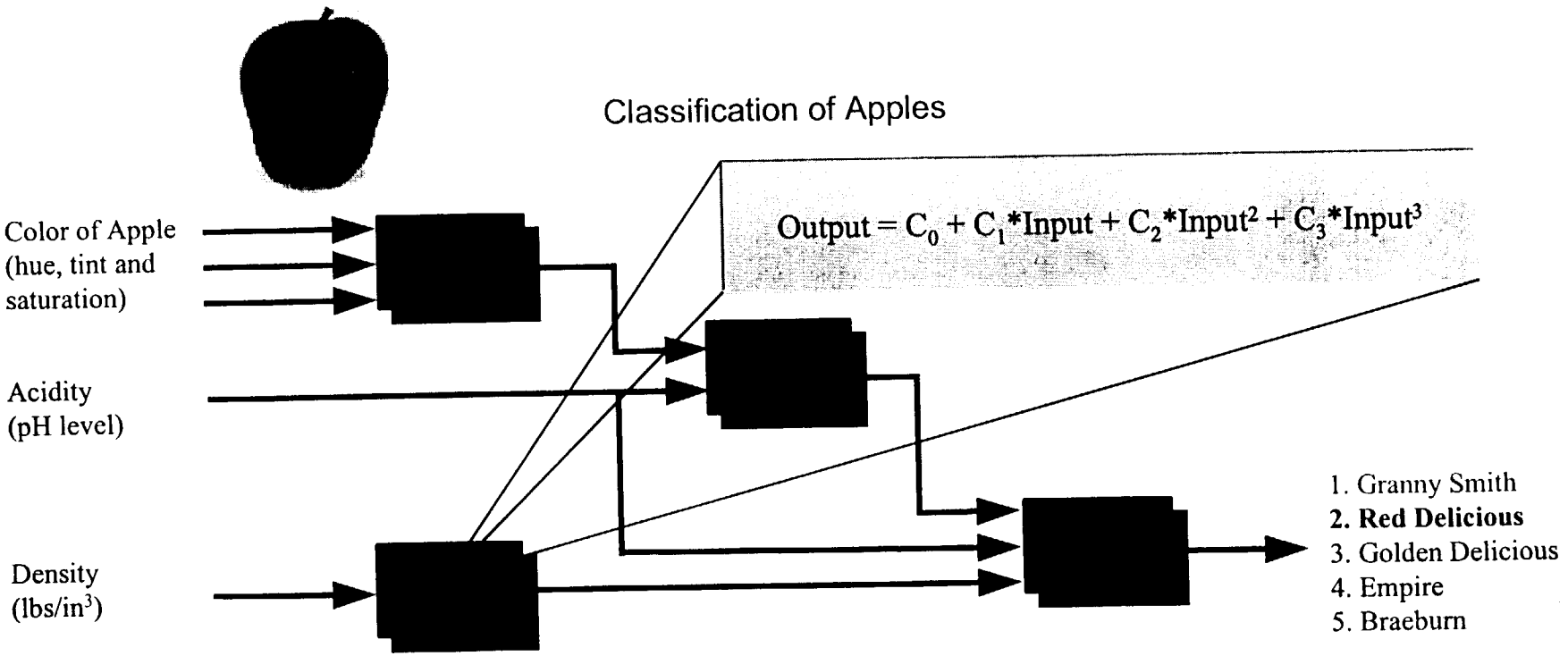
Data Mining Schemes

- Statistical Networks
- K-Nearest Neighbors
- Logistic Regression
- Decision Tree
- Hybrids
- Many others

Data Mining Schemes

Statistical Networks

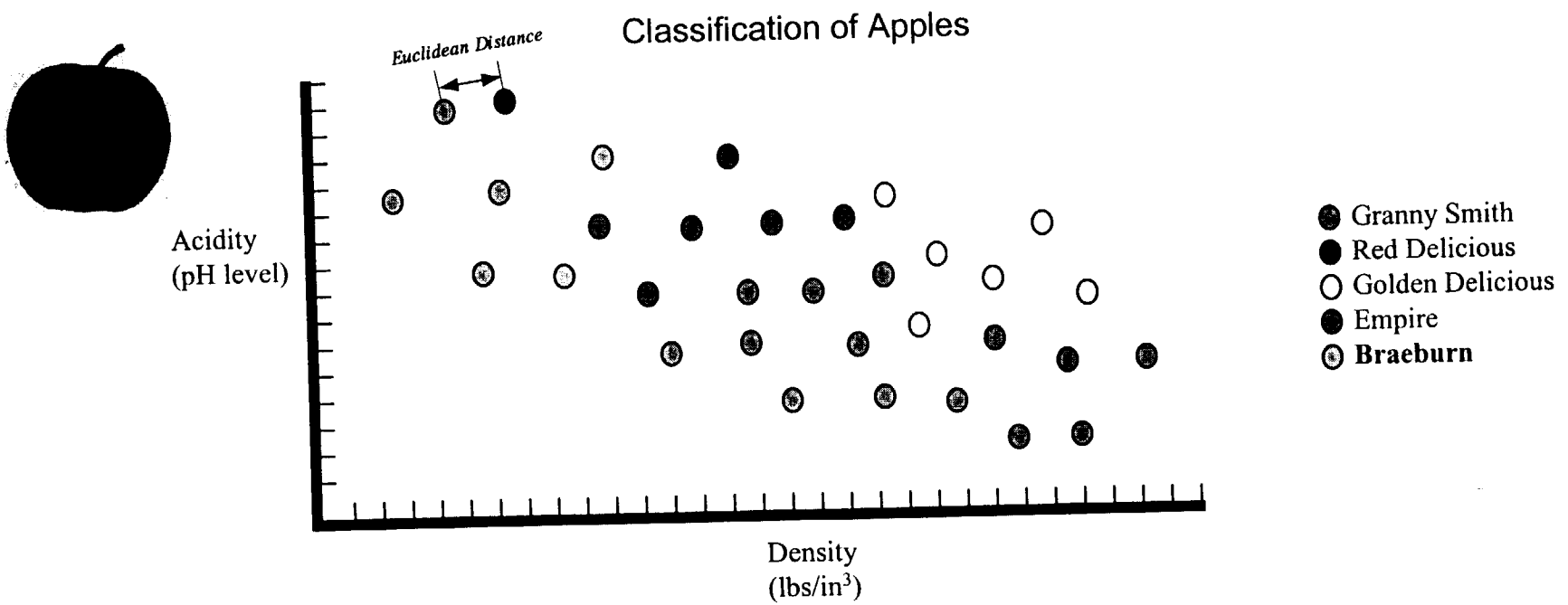
- Type of neural network with proven statistical learning functions
- Based on statistical learning algorithm, tries many different types of networks to converge on optimal
- Capture a large number of complex relationships in a very compact form



Data Mining Schemes

K-Nearest Neighbor

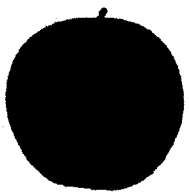
- Computes distances between sample and prior data
- The data set is the model
- Very effective for lower dimensional problems



Data Mining Schemes

Decision Tree

- Uses induction to learn the form of a "decision tree" for a given data set
- After the tree is established for a given data set, the tree is traversed answering questions at each decision node to classify an input example
- The classification result is reached when the tree is traversed to a leaf node



Color of Apple (hue, tint and saturation)

Acidity (pH level)

Density (lbs/in^3)

Classification of Apples

```
If Hue > 155 AND Tint < 16 Then
        If Acidity > 8.9 Then
                If Density < 0.12 Then Empire
        Else
                If Density < 0.18 Then Golden Delicious
                Else
                        If Saturation > 123 Then Empire
                        Else Red Delicious
Else If Hue < 131 AND Tint > 136
        If Acidity > 7.4 Then
                If Density > 0.11 Then Granny Smith
                Else
```

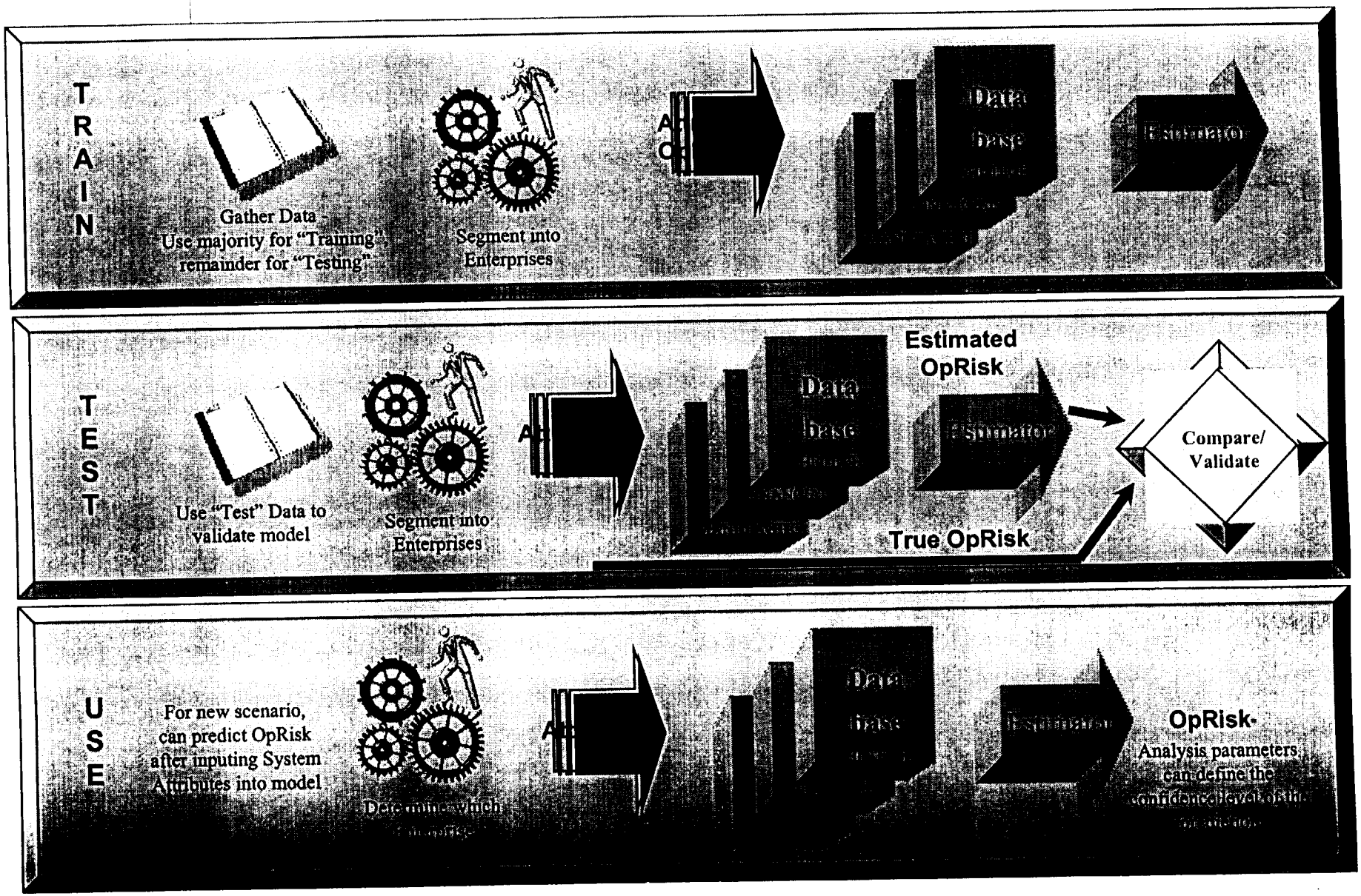
1. Granny Smith
2. Red Delicious
3. Golden Delicious
4. Empire
5. Braeburn

Generic Example

Use a Model (e.g. Neural Network, Interpolating Polynomials, etc.) to Predict Operational Risk

- Select Attributes that Have Correlation to Operational Risk
- Gather Data on Attributes and Operational Risk
- Segment Data into Enterprises (Financial, Manufacturing, High Tech, etc.)
- Data Transform
- Train Each Model with Inputs (Attributes) and Outputs (Operational Risk)
- Test Model: Supply Inputs to Model and Compare Outputs to Actual Known Operational Risk Values
- Now Use Trained Model to Predict Operational Risk of New Scenario

Illustration



Practical Considerations

- Not a "silver bullet"
- Don't need to understand algorithms in detail
  - Many software vendors have easy-to-use modeling software
- Automated software will "fit" data to many different types of modeling schemes, and let user select best model
- Should segment database for each type of enterprise to improve effectivity of the models
- Data manipulations is key to success
- Should have panel review process to validate the results
- Should "train" network with only recent data
- Bad Correlations?:
  - Not enough data (quantity, sparse regions, etc.)
  - Not correct data manipulations/transformations
  - Not right inputs

Summary

- Powerful & flexible modeling scheme
- Many different schemes available for data mining
- Commercial software vendors have software that do all the work for you
- User defines the inputs and outputs of the system
- User "trains" the system and evaluates the effectivity of the different models
- Data transformations and segmentation can be critical to the process

Contact Info

J. Sebastian Perera, J.D.

Phone: (281) 483- 5814 (Voice)

(281) 512- 7393 (Pager)

E-mail: jeevan.s.perera1@jsc.nasa.gov