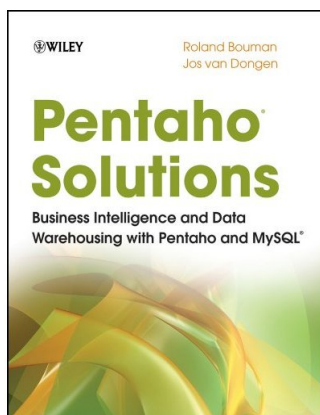
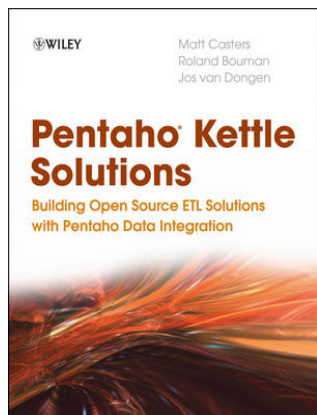




# Jos van Dongen

> 18 jr BI  
Principal Consultant  
Author/speaker/analyst

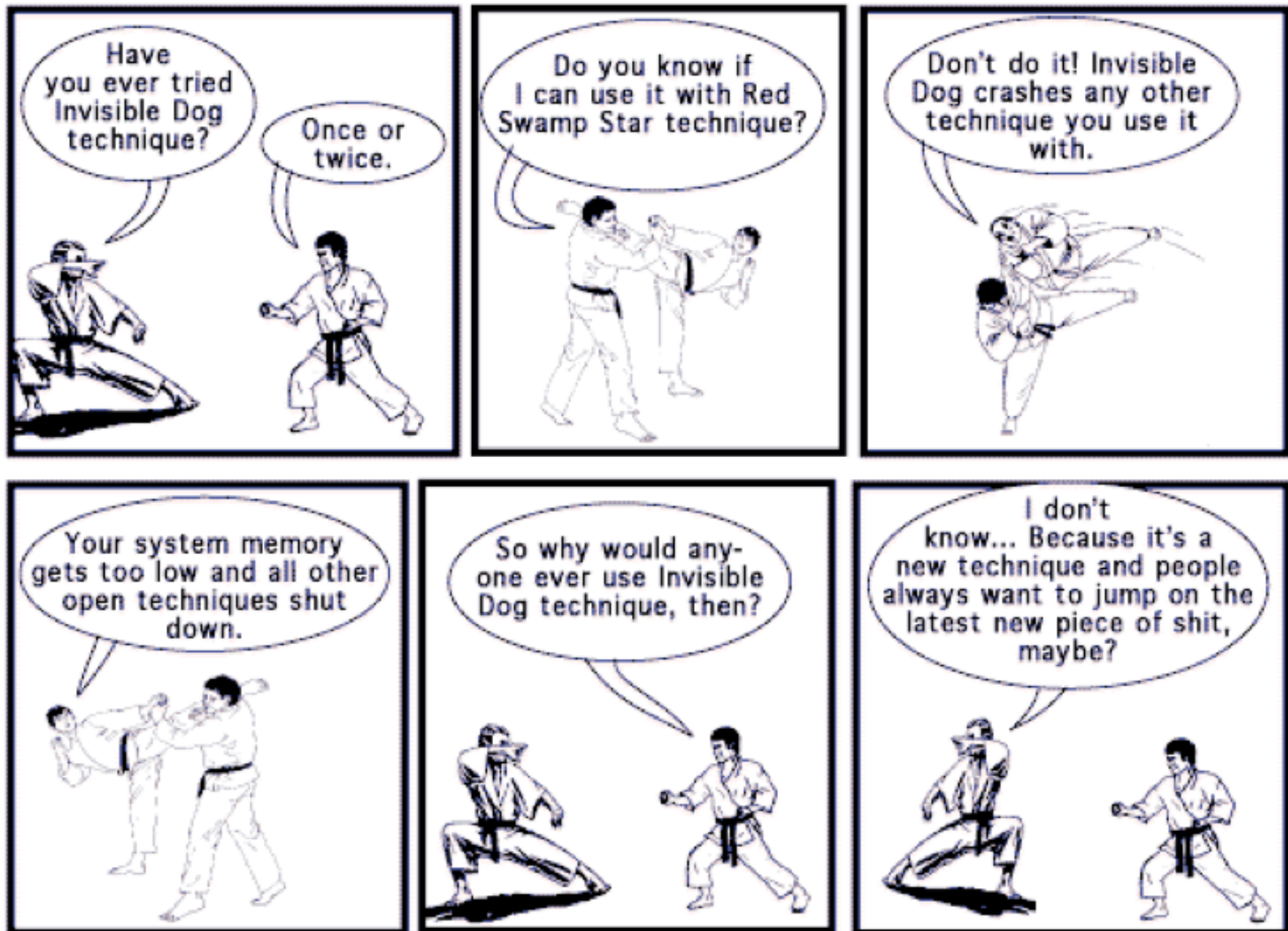


Web: [www.tholis.com](http://www.tholis.com)  
Email: [jos@tholis.com](mailto:jos@tholis.com)  
Phone: +31-(0)50-2304510  
Skype: tholis.jos  
LinkedIn: jvdongen  
Twitter: josvandongen





# About Data Mining Techniques...



# Just remember....

Every model is wrong...

...but some are useful

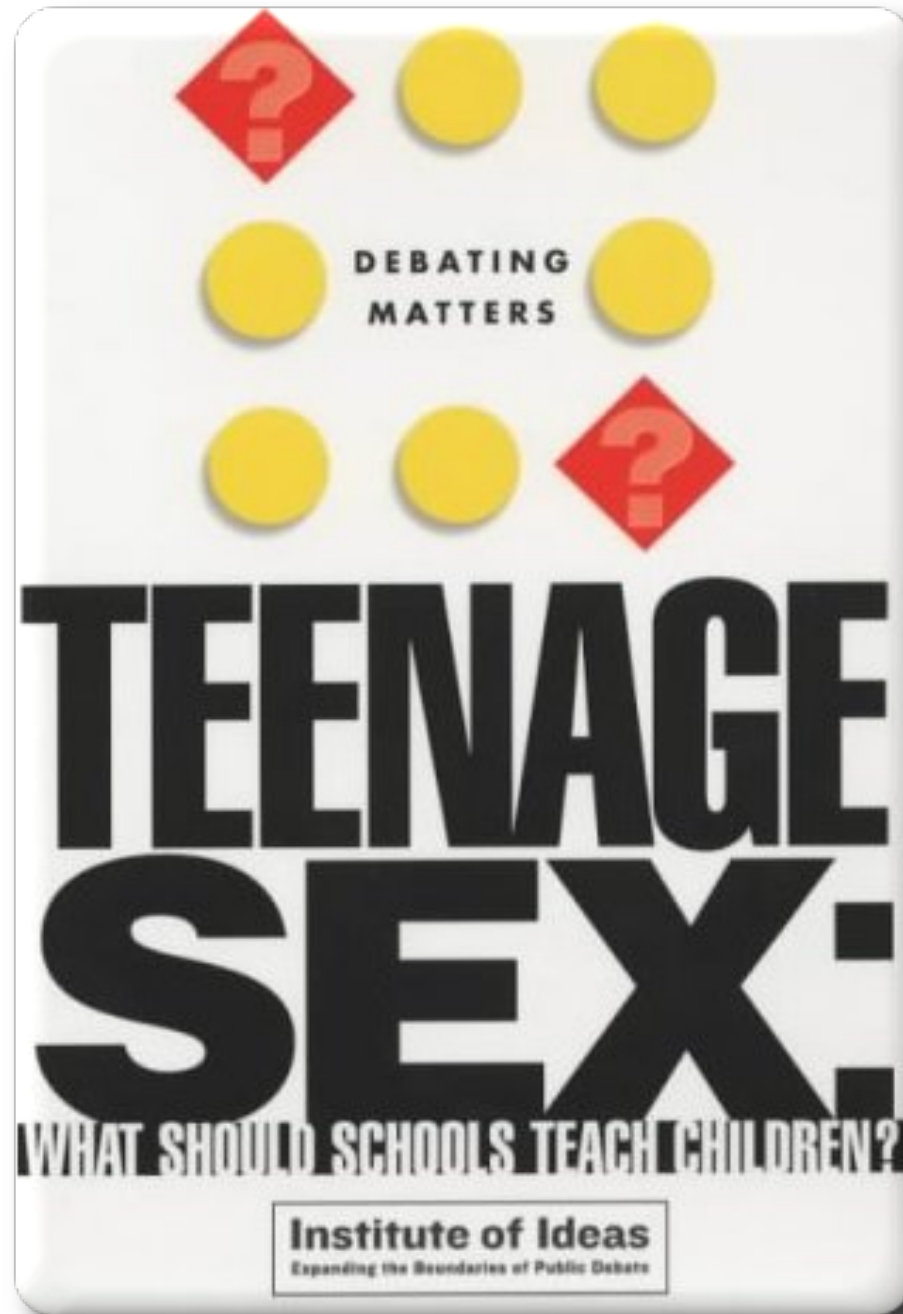


<b>A</b> RSA MEX URU FRA	<b>B</b> ARG NGA KOR GRE	<b>C</b> ENG USA ALG SLO	<b>D</b> GER AUS SCG GHA	<b>E</b> NED DEN JPN CMR	<b>F</b> ITA PAR NZL SVK	<b>G</b> BRA PRK CIV POR	<b>H</b> SPA SUI HON CHI
<b>1/8</b> ? ? ? ? 1-A 2-B 1-C 2-D	? ? ? ? 1-E 2-F 1-G 2-H	? ? ? ? 1-B 2-A 1-D 2-C	? ? ? ? 1-F 2-E 1-H 2-G	<b>1/4</b> ? ? ? ? ? ? ? ?	<b>S</b> ? ? ? ?	<b>F</b> ? ? ? ?	

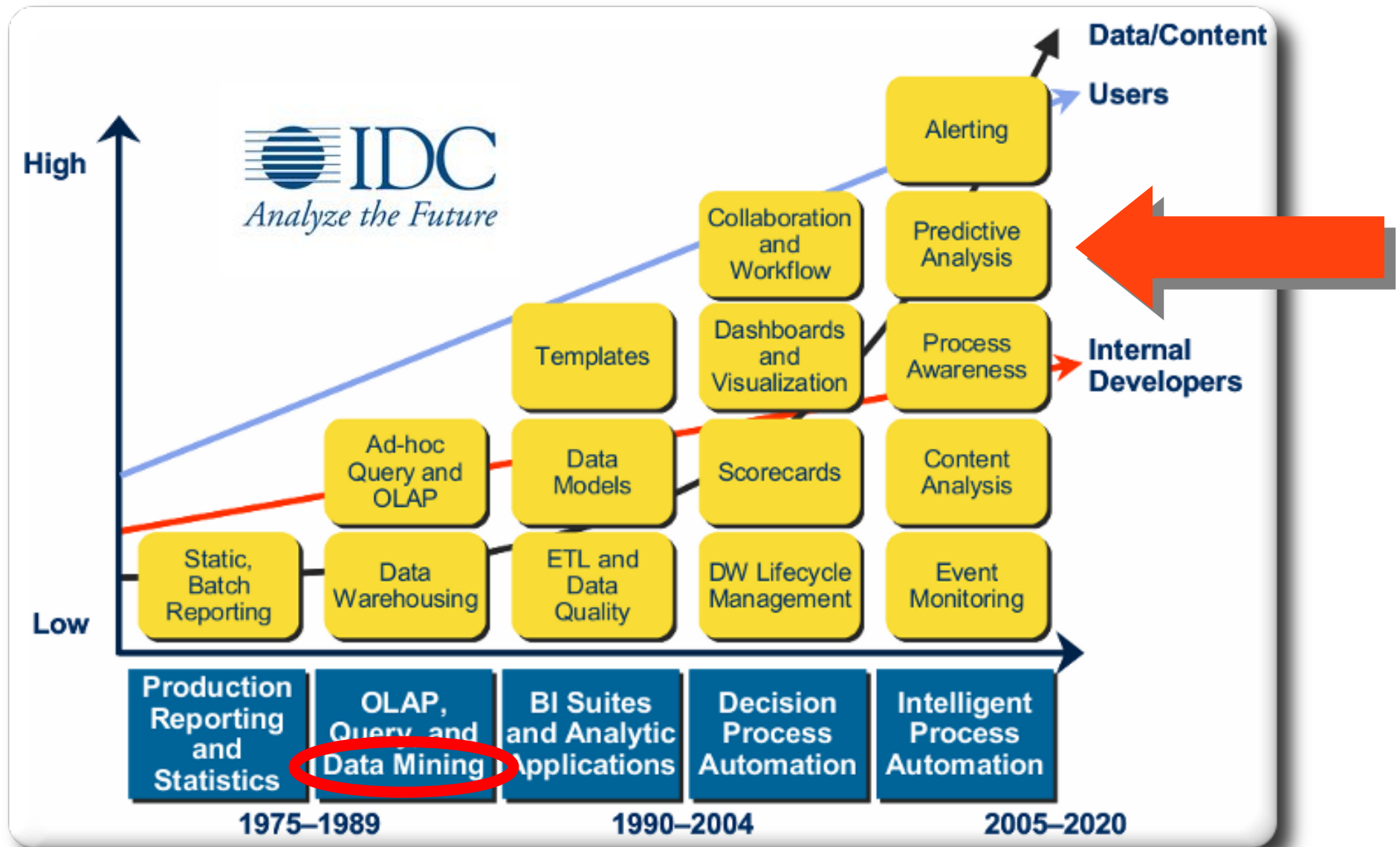
ROUND of 16	QuarterFinals	SEMIFINALS	FINAL	SEMIFINALS	QuarterFinals	ROUND of 16
<p>26 JUNIO</p> <p>Uruguay Korea Republic 69.0% - 31.0%</p> <p>26 JUNIO</p> <p>USA Ghana 60.4% - 39.6%</p> <p>27 JUNIO</p> <p>Netherlands Italy 60.8% - 39.2%</p> <p>27 JUNIO</p> <p>Brazil Chile 74.9% - 25.1%</p>	<p>2 JULIO</p> <p>Uruguay USA 56.8% - 43.2%</p> <p>3 JULIO</p> <p>Netherlands Brazil 42.0% - 58.0%</p>	<p>6 JULIO</p> <p>Uruguay Brazil 26.8% - 73.2%</p>	<p>11 JULIO</p> <p>Brazil Spain 57.2% - 42.8%</p> <p>10 JULIO</p> <p>Uruguay Argentina 34.8% - 65.2%</p>	<p>7 JULIO</p> <p>Argentina Spain 48.1% - 51.9%</p>	<p>2 JULIO</p> <p>Germany Argentina 46.9% - 53.1%</p> <p>3 JULIO</p> <p>Paraguay Spain 35.9% - 64.1%</p>	<p>28 JUNIO</p> <p>Germany England 51.7% - 48.3%</p> <p>28 JUNIO</p> <p>Argentina Mexico 61.7% - 38.3%</p> <p>29 JUNIO</p> <p>Paraguay Denmark 50.3% - 49.7%</p> <p>29 JUNIO</p> <p>Spain Portugal 57.3% - 42.7%</p>

# Data Mining is like...

©Tom Breur, 2009

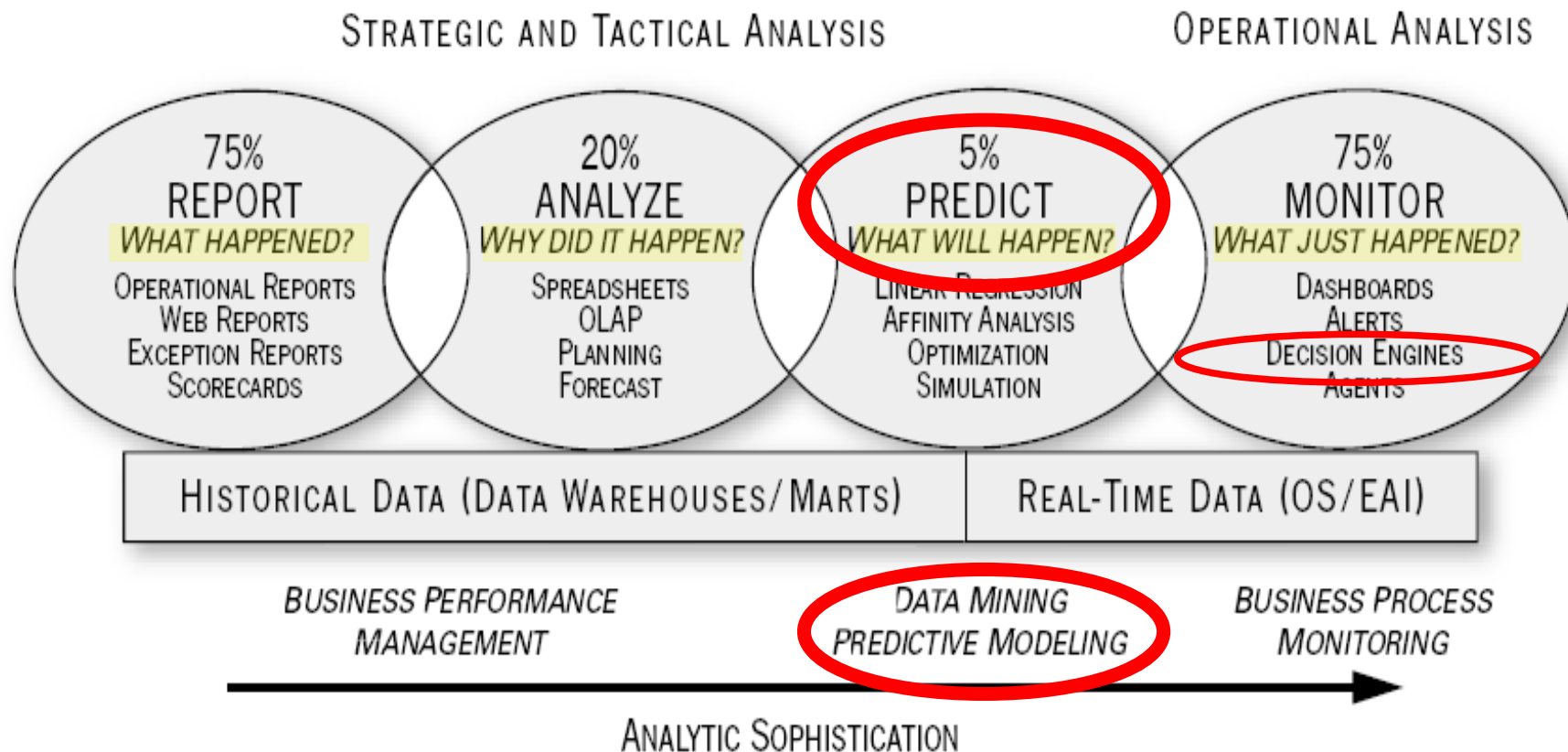


# Predictive Analytics is next



# Data Mining & BI

## The Landscape for Analytical Tools



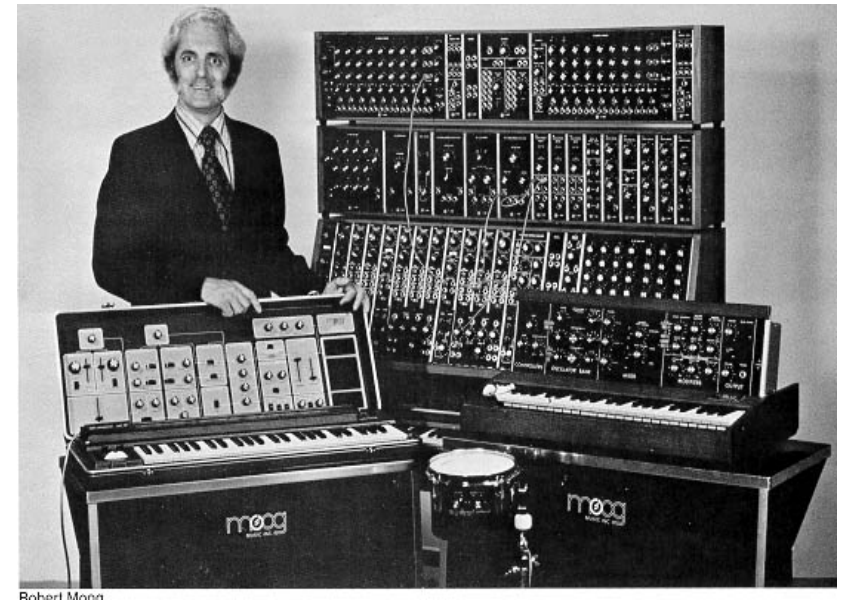
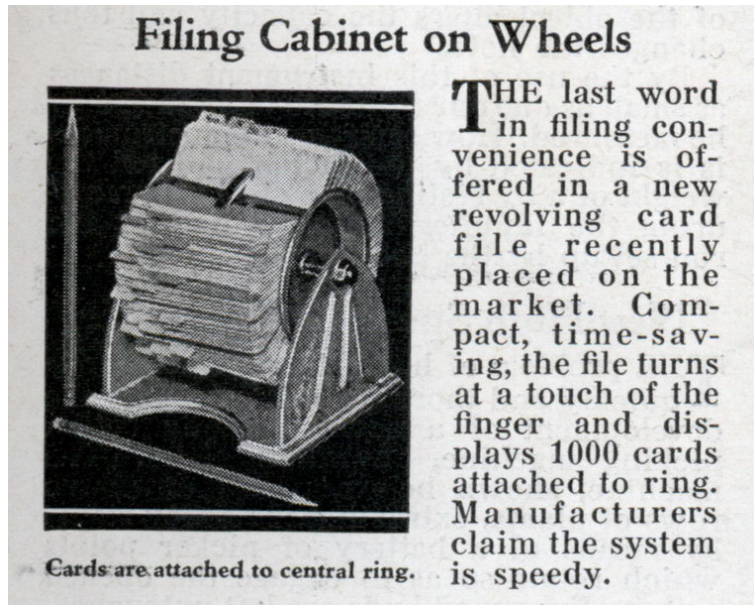
Source: TDWI® ([www.tdwi.org](http://www.tdwi.org))



# Why Now?

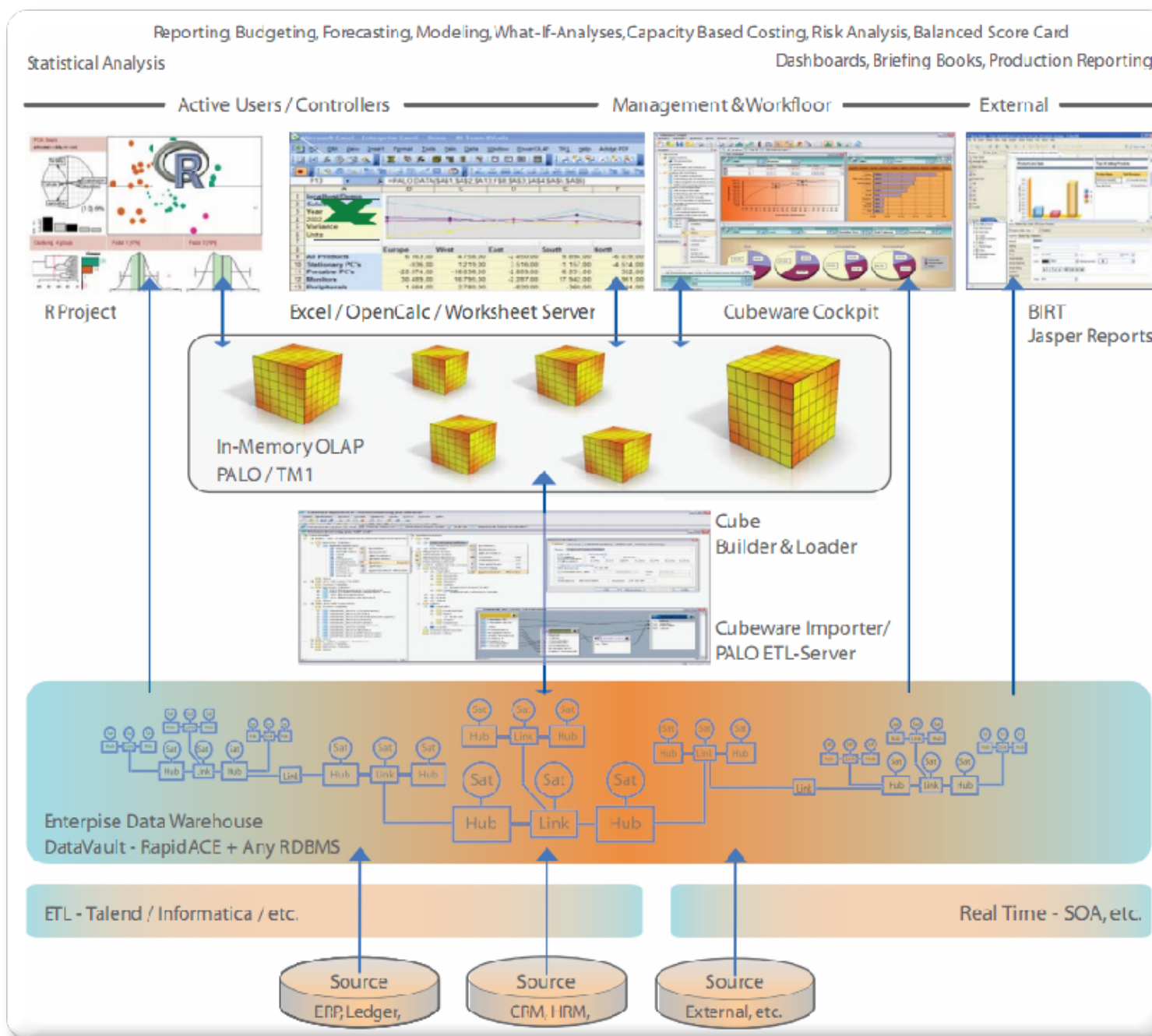
## The Perfect Storm:

- Data collection, particularly of people's behavior, makes new things possible.
- The data volume and complexity require methods other than basic query and reporting.
- Commodity computing makes complex work possible.

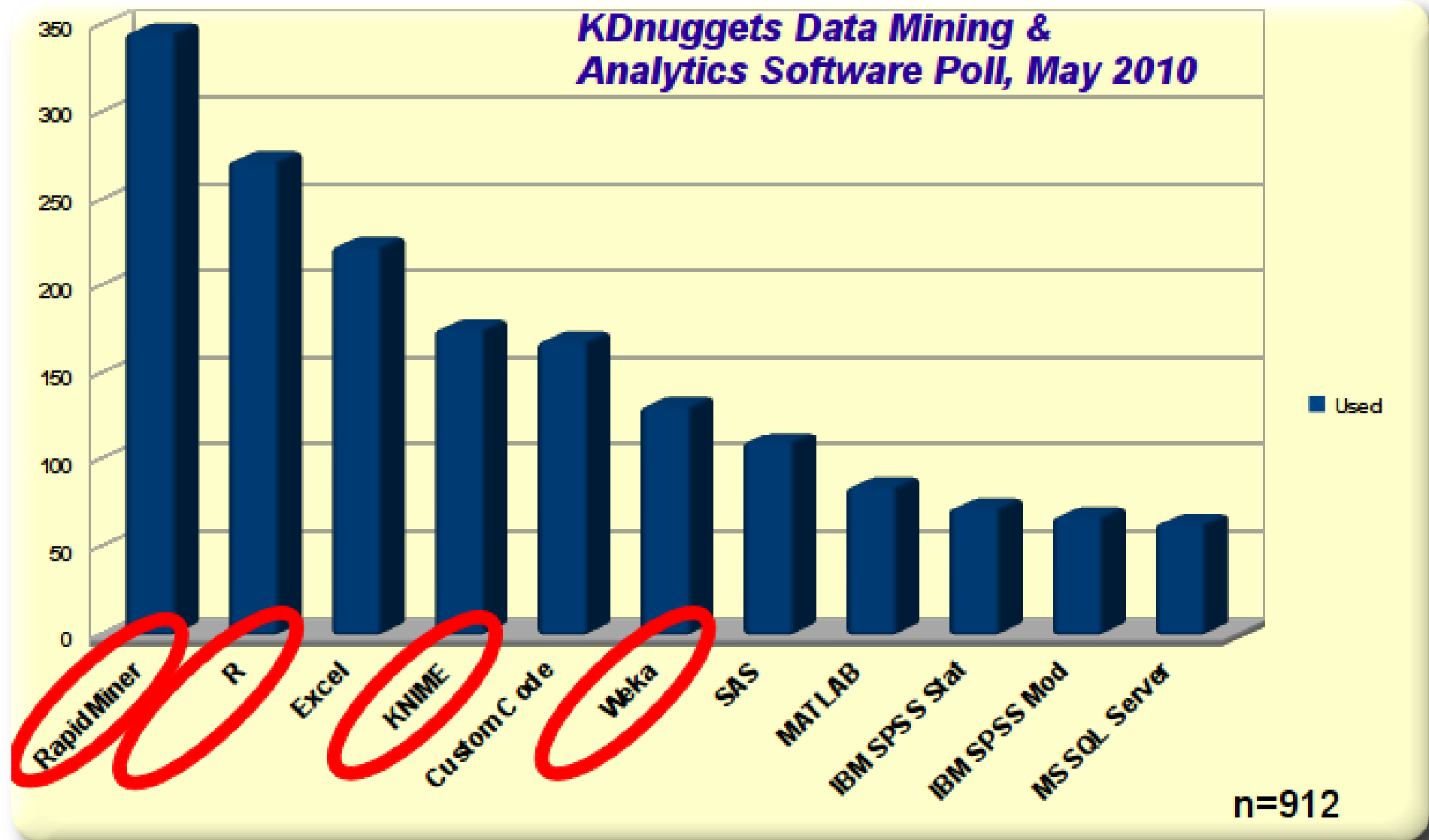


Robert Moog

# We're building on existing data infrastructure



# The tools people use...





# What You Need for Data Mining

## ◇ Data

- ◇ DWs have readily available data and ETL tools to get missing data.
- ◇ BI tools help you identify and isolate the data you need.

## ◇ Clean data

- ◇ Your models are as good as the data that goes in. Most data mining techniques are sensitive to noise, inaccuracy, and missing values.

## ◇ Pre-processed clean data

- ◇ Most of the time you can't run against your existing data. You need to prepare the data:
  - ◇ Group continuous variables, symbolic to numeric transformations, coding categorical or discrete values

## ◇ Model building, testing, validation, and operational data storage

- ◇ The process of building, testing, validating is iterative
- ◇ You need data management infrastructure to run your models in production.

***Like a DW, ~70% of analytics project effort is spent preparing data.***

# What is Data Mining?

*“The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”*

Fayyad et al.

*“The extraction of hidden predictive information from large databases”*

Kurt Thearling

✓ But also:

- Knowledge Discovery
- Machine Learning
- “Anything the computer does that we don’t understand”(MM)
- “Statistics + Marketing”(frustrated statisticians)

# Different Data Perspectives

## ✓ BI

- Result is a report displaying known data
- Data is read once and reported on
- Looks at complete data set
- Missing data prohibited
- Data can be pre-aggregated
- 100 % correctness required

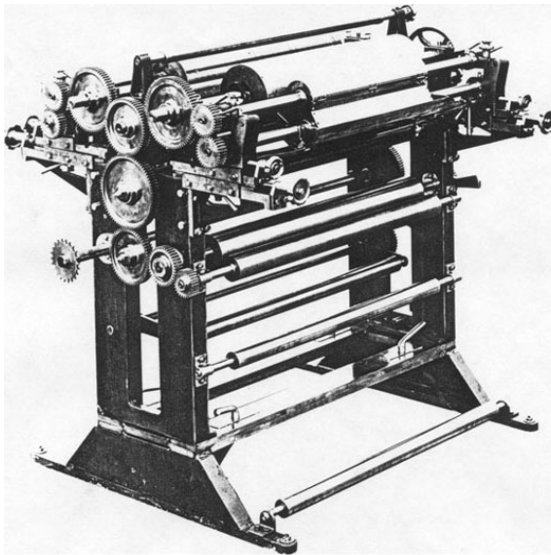
## ✓ Data Mining

- Result is a previously unknown prediction or segmentation
- Data is processed multiple times to yield better results
- Splits data into training & test sets
- Expects missing data
- Detailed data required
- Getting an 80% accurate prediction is Gold



# Value of Data Mining

- Automated discovery of patterns in data
- Predictive capabilities
- (by) Making better use of data that a business already collects in the normal course of business operations

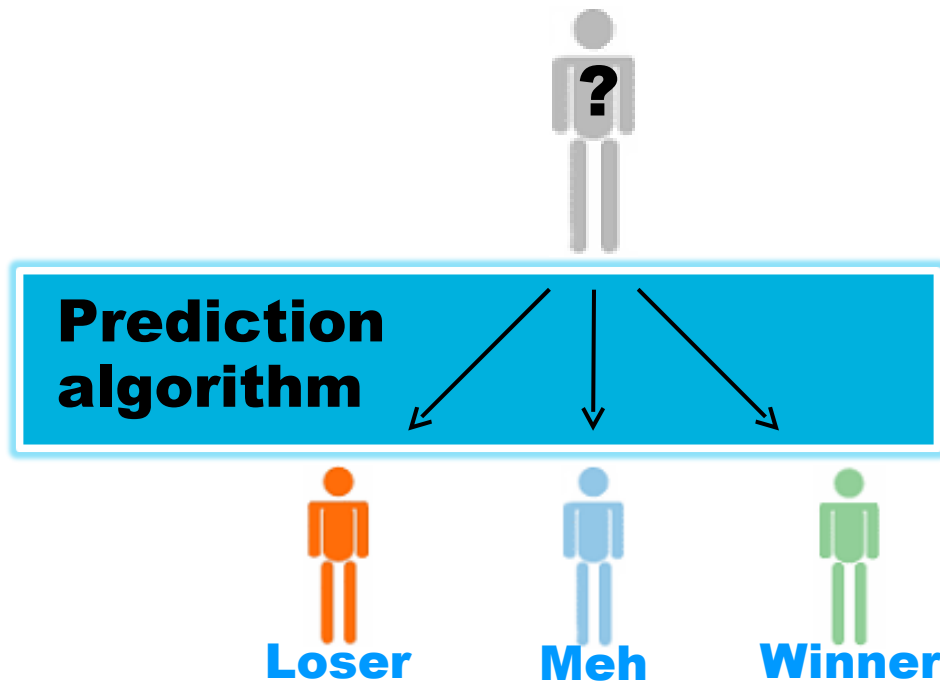


5288



# A Useful Way to Look at Techniques

## Predictive models



**Predictive models can determine an output variable or rank-order based on the input data.**

**What:** Use known variables to predict unknown or future values of other variables.

**For example:** Determine the odds of an event. Will a customer respond to an offer? Is this a fraudulent transaction?

### **Techniques:**

- (Linear) regression
- Neural networks\*
- Decision, regression trees

# A Useful Way to Look at Techniques

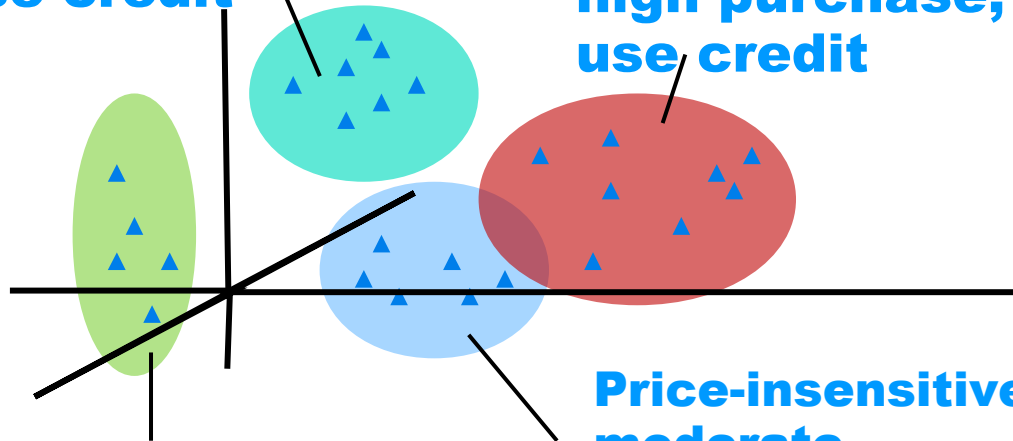
## Descriptive Models

Price-sensitive,  
low purchase,  
use credit

Price-sensitive,  
high purchase,  
use credit

Price-insensitive,  
low purchase, no  
credit

Price-insensitive,  
moderate  
purchase, no  
credit



*In theory* you could use these clusters of the existing base to predict the behavior of new customers based on the same attributes.

**What:** Find relationships and patterns, or define models that describe the data in question.

**For example:** Define customer segments for optimal email marketing response.

## Techniques:

- Classification and clustering
- Association rules
- Sequential pattern discovery
- Logistic regression



# Groups of Techniques

**Most techniques fit into two groups based on how you work with them:**

**Supervised learning**: a person has to define the correct output for some portion of the data. Data is divided into training sets used for model building and test sets for validating the results.

✦ **What constitutes a good vs. bad credit risk?**

**Unsupervised learning**: the algorithm functions without an agent specifying the action to take on a given set of input. The algorithm derives the pattern.

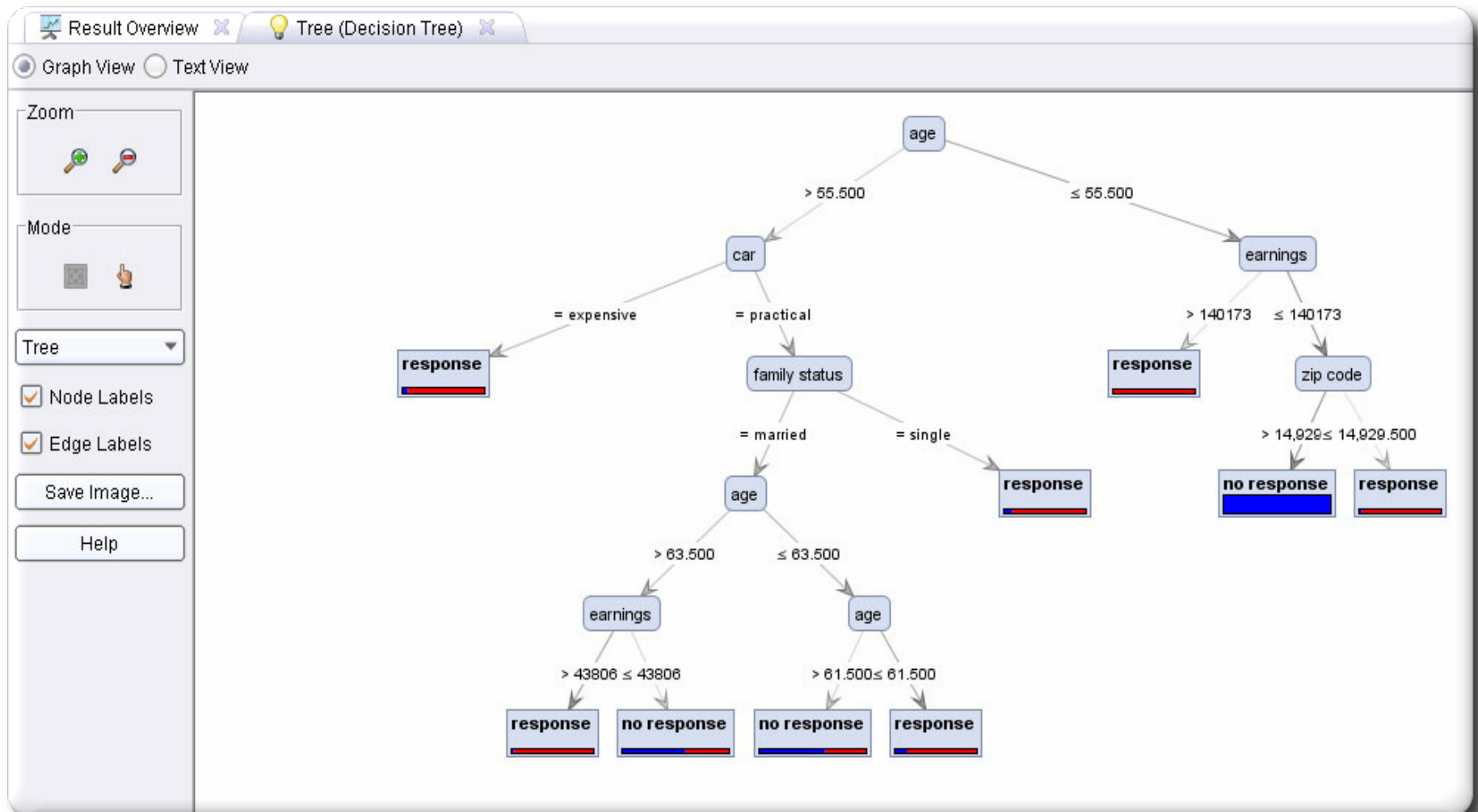
✦ **What is the best way to segment customers for marketing?**

# Decision Trees

- **Sets of decisions represented as a tree-structure**
- **Decisions generate rules for the classification of a dataset.**
- **Classification methods:**
  - **CART/C&RT (Classification and Regression Trees: 2 way splits)**
  - **CHAID (Chi Square Automatic Interaction Detection)**
  - **ID3/C4.5/C5.0 (C5.0 only commercially available)**
  - **J48 (adapted C4.5: Weka Tree algorithm)**
- **Tree algorithms are recursive**
  - **Top level is attribute with the highest information gain**
  - **Next, recurse**


# Decision Tree Example

➤ **Classification/Prediction with Decision Trees  $\approx$  40% DM**



# Decision Trees in Depth (1)

## 1. Data Set: Attributes plus Classifier (outcome)

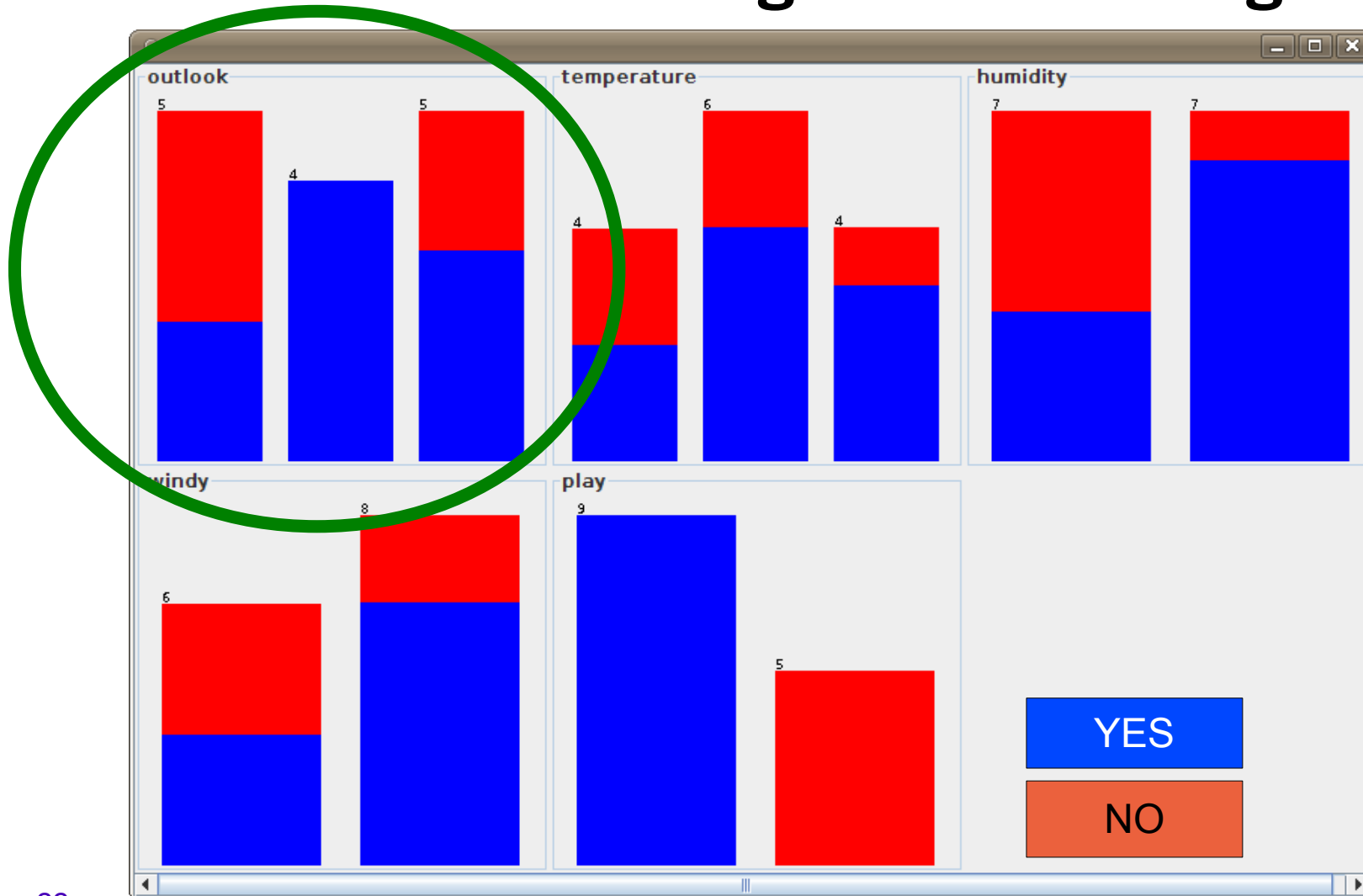


outlook	temperature	humidity	windy	play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no



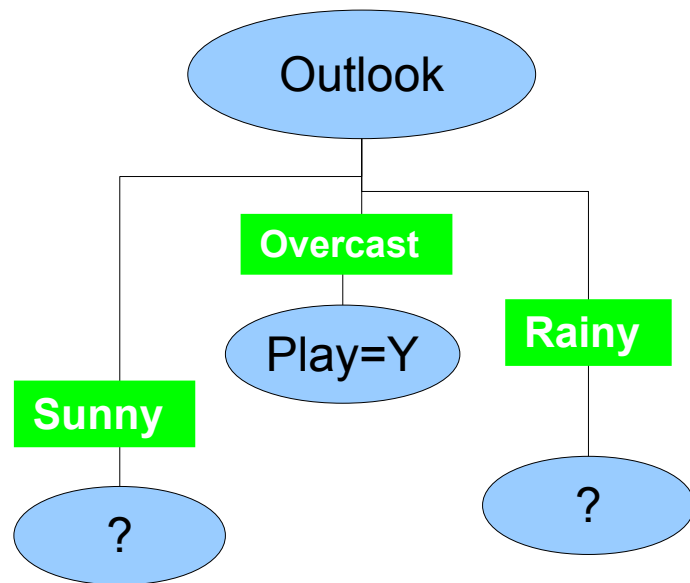
# Decision Trees in Depth (2)

## 2. Find Attribute with highest information gain



# Decision Trees in Depth (2)

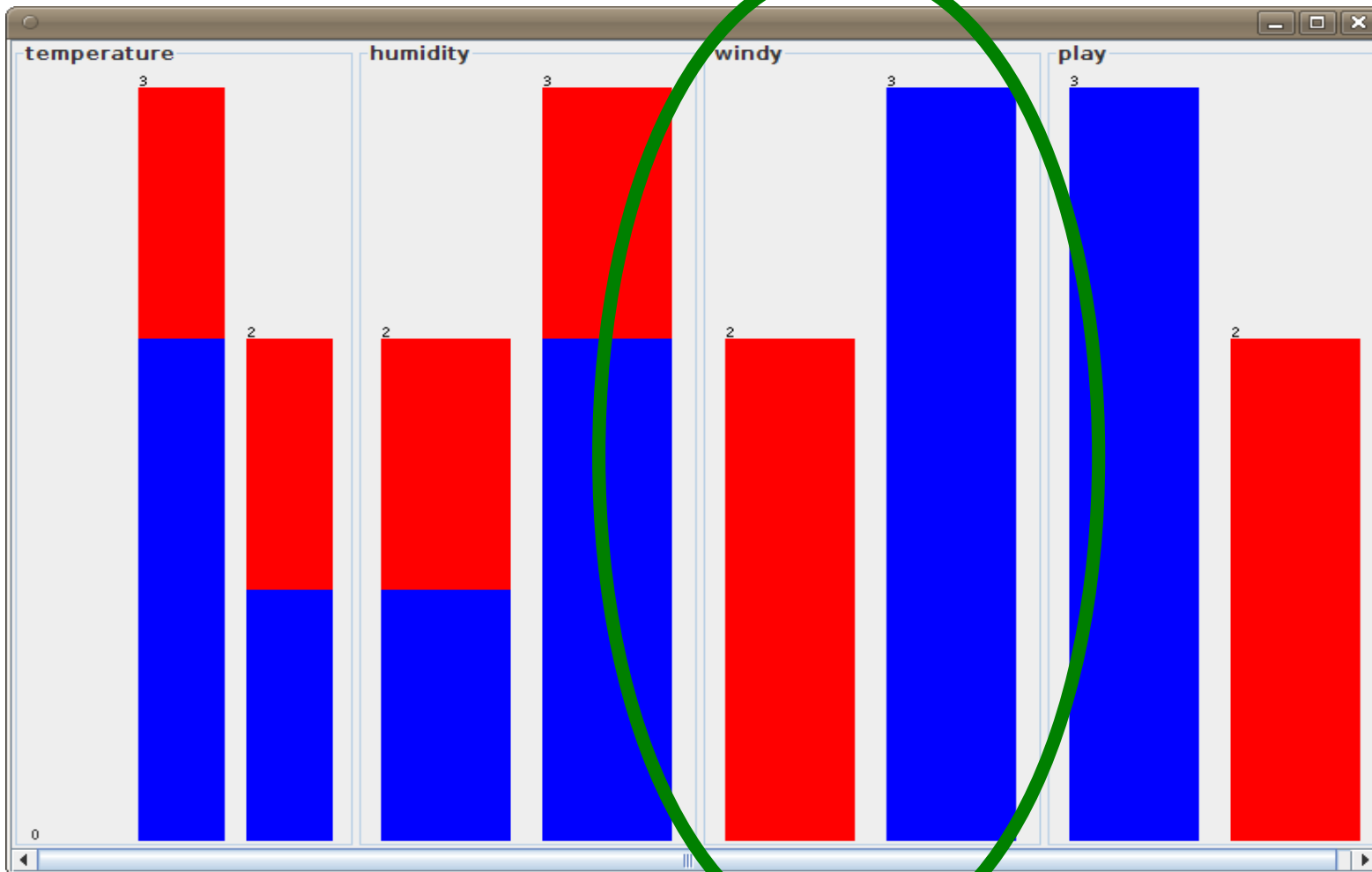
➤ Next, recurse



outlook	temperature	humidity	windy	play
overcast	hot	high	FALSE	yes
overcast	cool	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
rainy	mild	normal	FALSE	yes
rainy	mild	high	TRUE	no
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

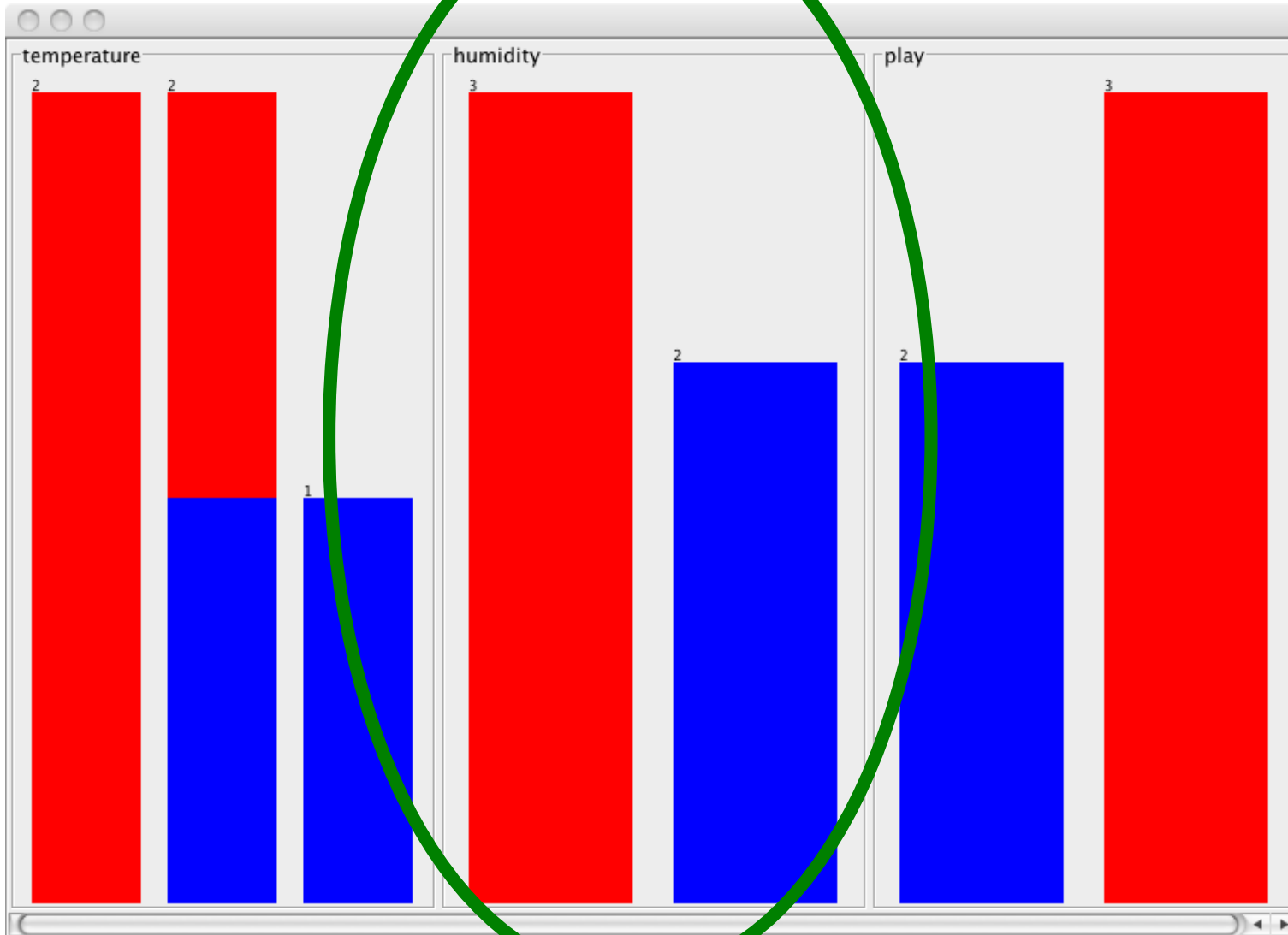
# Decision Trees in Depth (4)

➤ Evaluate 'Rainy'



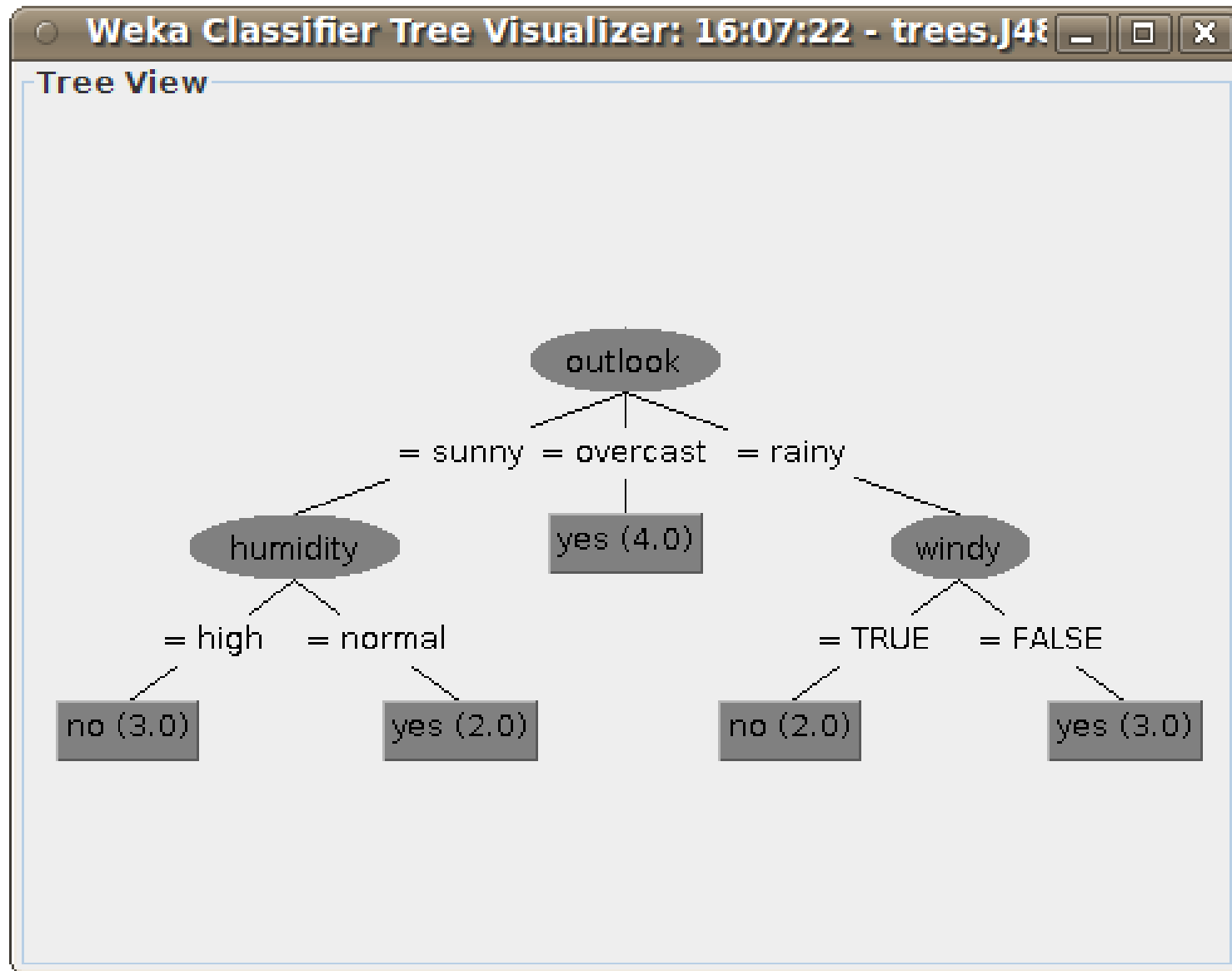
# Decision Trees in Depth (5)

➤ Evaluate 'Sunny'





# Decision Tree Result

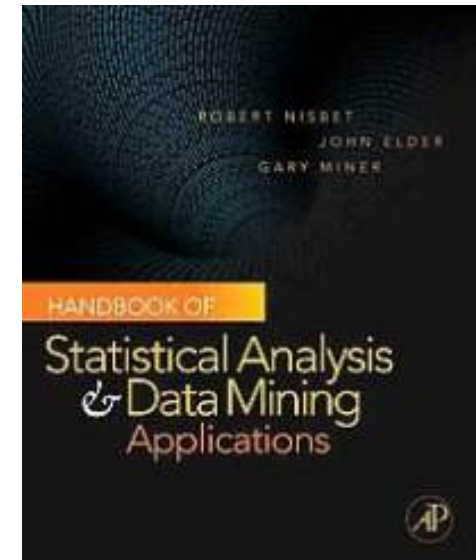
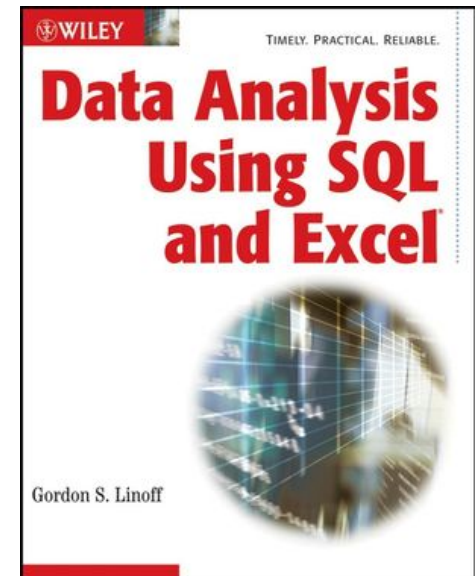


# **Demo:**

# **Weka scoring with Kettle**

# Where to go from here?

- Read 'Competing on Analytics'
- Move on to 'Data Analysis Using SQL and Excel'
- Then buy 'Handbook of Statistical Analysis & Data Mining Applications'
- Statistics for business:
  - <http://home.ubalt.edu/ntsbarsh/Business-stat/opre504.htm>
- Data Mining:
  - [www.rapid-i.com](http://www.rapid-i.com) (RapidMiner)
  - <http://www.thearling.com>
  - <http://www.autonlab.org/tutorials/>
- For free text books, search [www.scribd.com](http://www.scribd.com)



# More Resources to Get You Started

## Books:

- ✧ **Data Mining Techniques: For Marketing, Sales and Customer Support, Michael J. Barry and Gordon Linoff**
- ✧ **Data Preparation for Data Mining, Dorian Pyle**
- ✧ **Data Mining Algorithms, Elbe Frank, Ian Witten, Jim Gray**
- ✧ **An Introduction to Information Retrieval, Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze**
- ✧ **Information Retrieval, C. J. van Rijsbergen**
- ✧ **The Visual Display of Quantitative Information, Edward R. Tufte**

## Journals, Newsletters, Web Sites:

- ✧ **SIG KDD Explorations, Newsletter of the ACM SIG on Knowledge Discovery and Data Mining**
- ✧ **IEEE Transactions on Pattern Analysis and Machine Intelligence**
- ✧ **KDNuggets data mining resources: [www.kdnuggets.com](http://www.kdnuggets.com)**
- ✧ **Flowing Data, visualization resources: <http://flowingdata.com/>**
- ✧ **Infoaesthetics, visual design resources: <http://infosthetics.com/>**
- ✧ **Visual Complexity, visualization resources: [www.visualcomplexity.com/vc/index.cfm](http://www.visualcomplexity.com/vc/index.cfm)**
- ✧ **Recommendation systems resources:  
<http://www.deitel.com/ResourceCenters/Web20/RecommenderSystems/tabid/1229/Default.aspx>**
- ✧ **The Impoverished Social Scientist's Guide to Free Statistical Software and Resources:  
<http://maltman.hmdc.harvard.edu/socsci.shtml>**



# Free Stuff So You Can Work Cheaply

- ✧ **WEKA** <http://www.cs.waikato.ac.nz/ml/weka/>
- ✧ **IND decision tree software** <http://opensource.arc.nasa.gov/software/ind/>
- ✧ **Clustering** <http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/>
- ✧ **Parallel Sets** <http://eagereyes.org/parallel-sets#download>
- ✧ **RapidMiner** <http://rapid-i.com/content/blogcategory/38/69/>
- ✧ **Knime** <http://www.knime.org/>
- ✧ **Orange** <http://www.ailab.si/Orange/>
- ✧ **R statistics software** <http://www.r-project.org/>
- ✧ **ARC statistics software** <http://www.stat.umn.edu/arc/software.html>
- ✧ **Octave numerical and matrix computation** <http://www.gnu.org/software/octave/>
- ✧ **Processing** <http://www.processing.org/>
- ✧ **Circos** <http://mkweb.bcgsc.ca/circos/>
- ✧ **Treemap** <http://www.cs.umd.edu/hcil/treemap/>
- ✧ **Many Eyes** <http://manyeyes.alphaworks.ibm.com/manyeyes/>

