Big Data Analytics
The Data Mining process

Roger Bohn
March. 2017

Office hours
RB Tuesday + Thursday 5:10 to 6:15.
Tuesday = office rm 1315; Thursday = Peet's
Sai Kolasani = ? <skolasan@eng.ucsd.edu>

Some material from **Data Mining for Business Analytics**

By Shmueli, et al

Administration

- Office hours
  - TA session = set time
- Web Site = blog + lots of resources
  - For R
  - Available big data sets
  - Supplements to the textbook on data mining issues
- TritonEd site
  - Grading
  - Where to put student/faculty questions and discussion?
- Syllabus: Read it fully
  - Schedule for next 3 weeks is coming.

Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- Data mining—Semi- Automated analysis of massive data sets

Economist: Special report on BD in politics

Technology and politics
The signal and the noise
Mar 26th 2016
Ever easier communications and ever-growing data mountains are transforming politics in unexpected ways, says Ludwig Siegele. What will that do to democracy?

- Technology and politics: The signal and the noise
- Election campaigns: Politics by numbers
- Tracking protest movements: A new kind of weather
- Online collaboration: Connective action
- Local government: How cities score
- Living with technology: The data republic

What can we do with Data Mining?

5

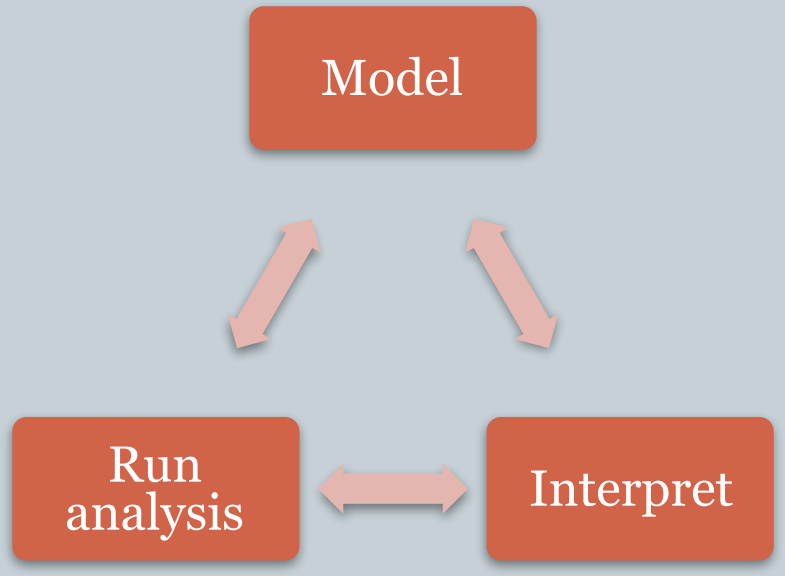
- Exploratory Data Analysis
- Predictive Modeling: Classification and Regression
- Descriptive Modeling
  - Cluster analysis/segmentation
- Discovering Patterns and Rules
  - Association/Dependency rules
  - Sequential patterns
  - Temporal sequences
- Deviation detection

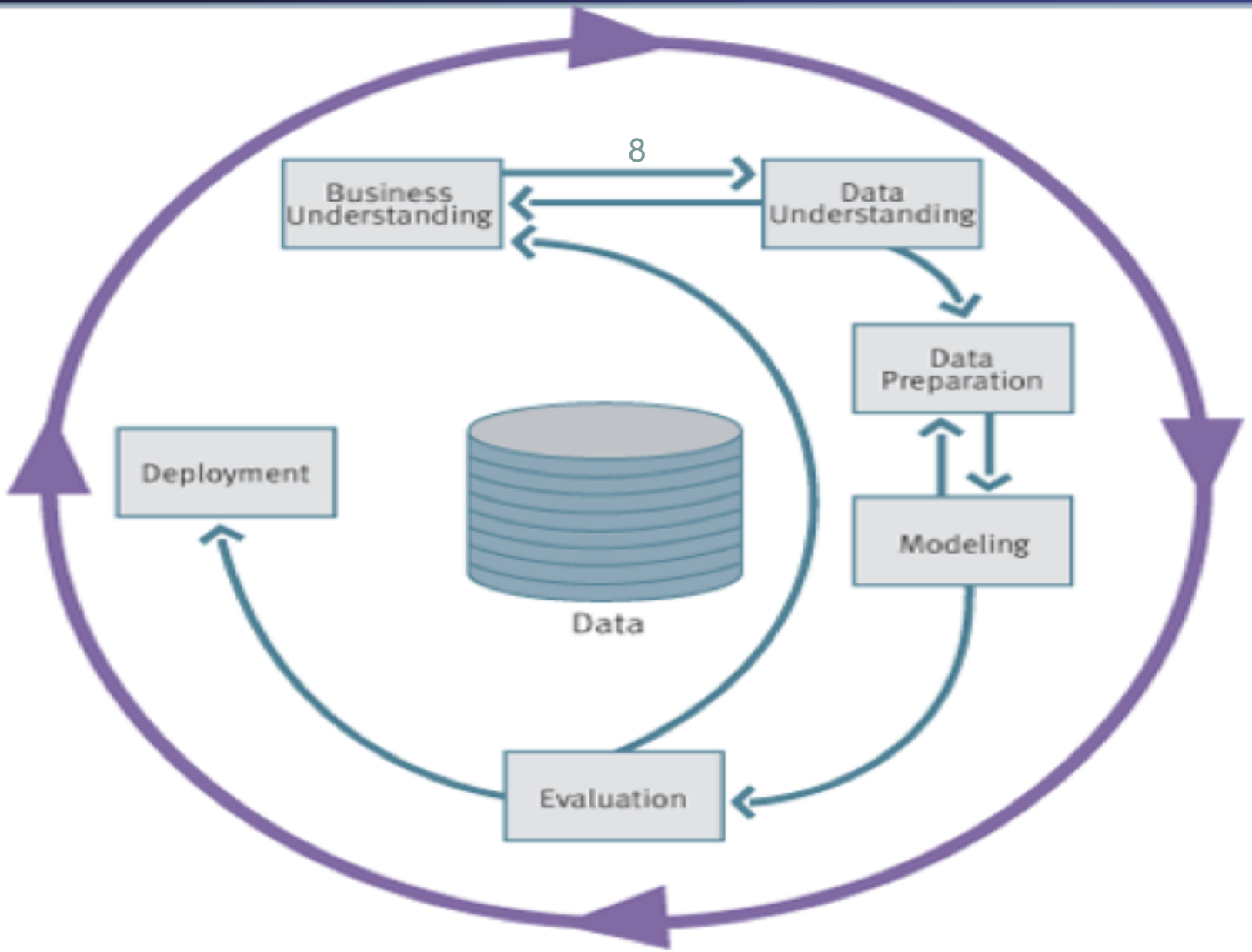
Canonical examples

- Should we approve this transaction? (Is it fraudulent? Likely to fail?)
  - Credit cards
  - Mortgages
- Which financial reports to audit more carefully?
- Which buildings to inspect in NY City?
- What to recommend to a user?

Data Analytics Process

- Choose a good problem
- Gather data
  - Locate, download, examine
  - Clean it e.g missing data, out of range
- Model the problem
  - Create new variables
  - Outcome = f(X1, X2, …. Xn)
  - Select an algorithm
- Run analysis.
- Interpret results
- Iterate until satisfied; DEPLOY
- Enhance, apply broadly, learn



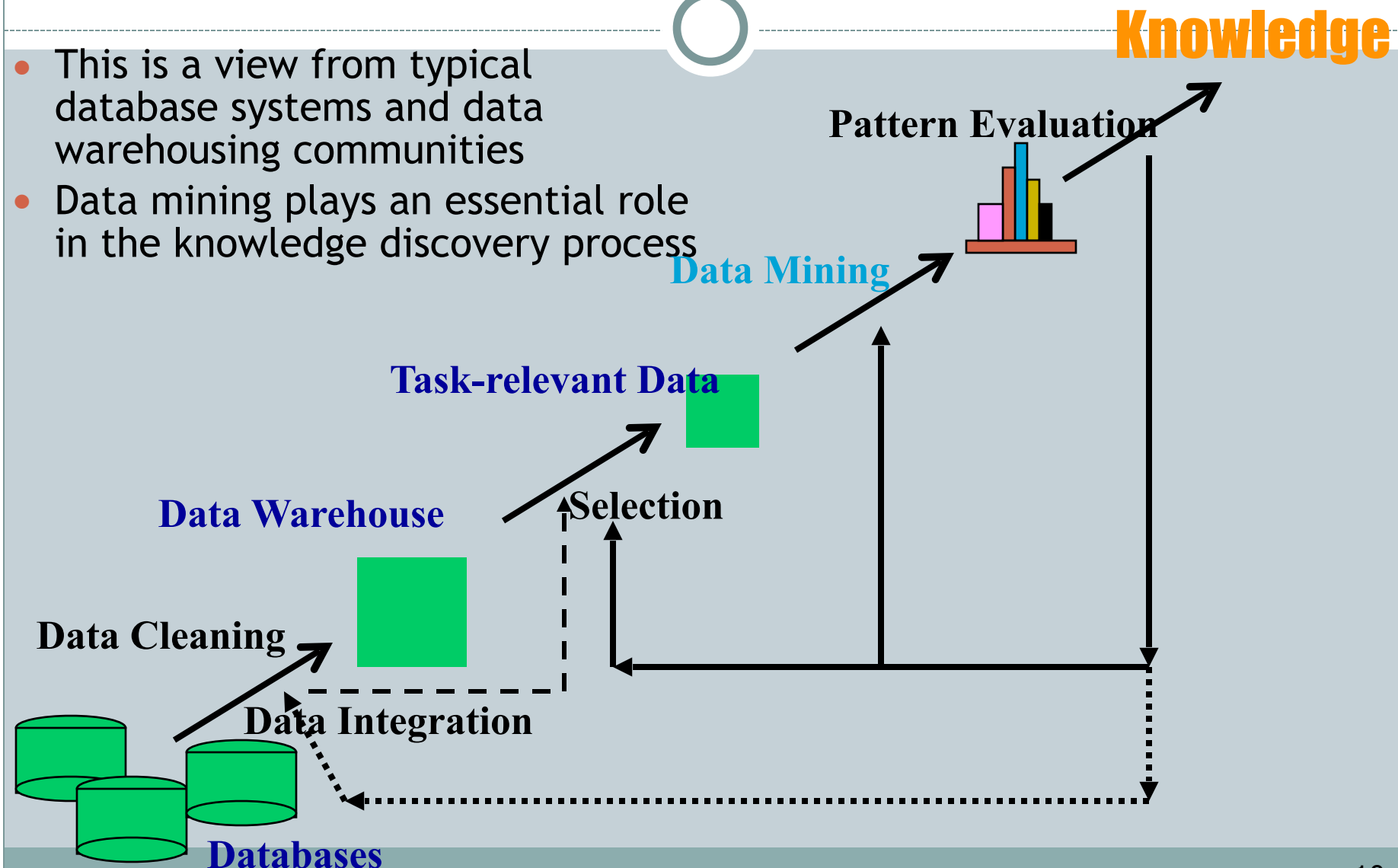

Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Steps in Data Mining

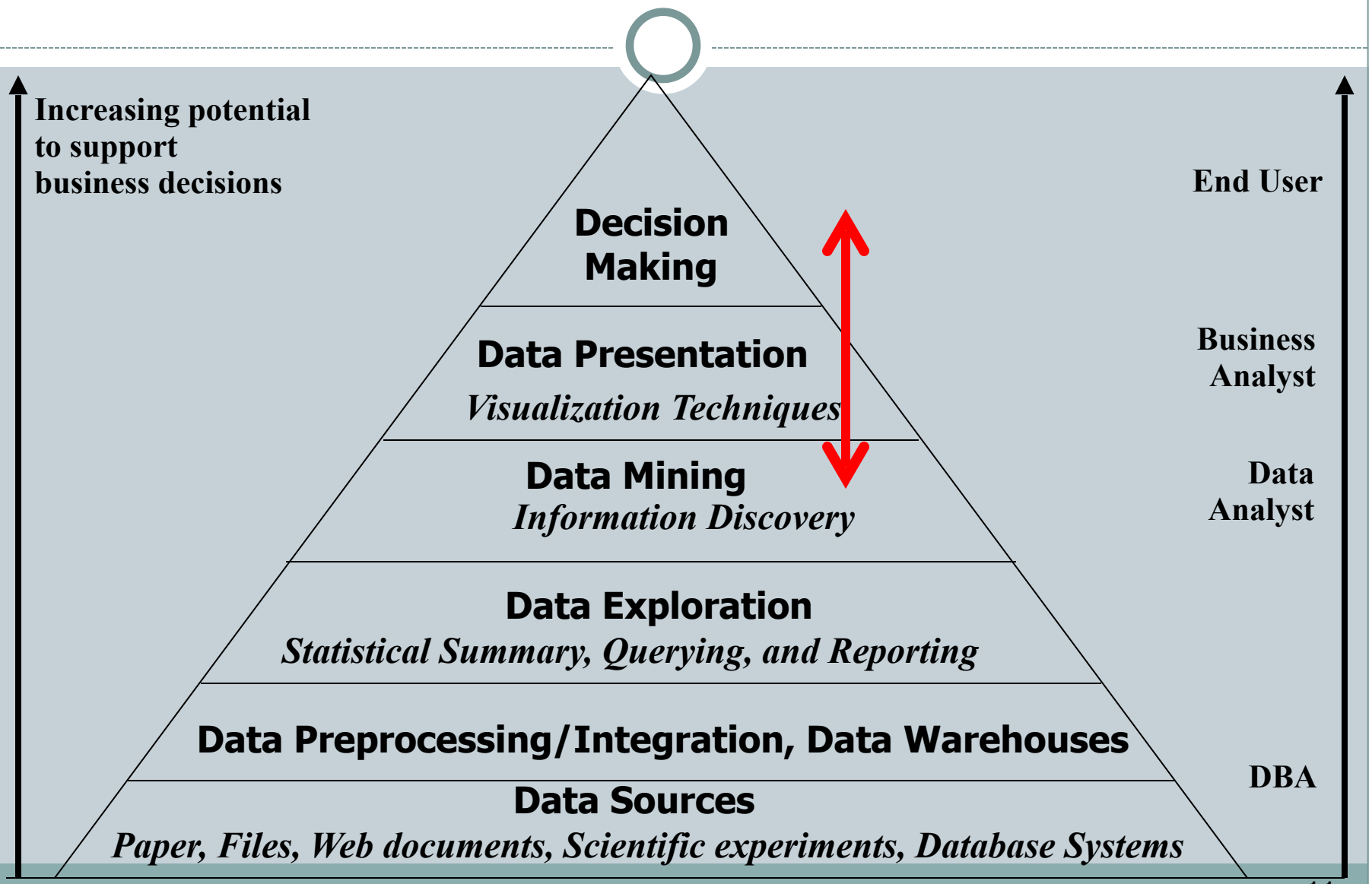
1. Define/understand problem/question/decision
2. Obtain data (may involve random sampling)
3. Explore, clean, pre-process data
4. Specify task (classification, clustering, etc.)
5. Try one or more algorithms (regression, k-Nearest Neighbors, trees, neural networks, etc.)
6. Iterative implementation and “tuning”
7. Assess results – compare models
8. *Deploy* model in production mode. Daily use

Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



Data Mining in Business Intelligence



Supervised Learning = This course

- Goal: Predict a single “target” or “outcome” variable
- Training data, where target value is known
- Methods: Classification and Prediction

Supervised: Classification

- Goal: Predict categorical target (outcome) variable
- Examples: Purchase/no purchase, fraud/no fraud, creditworthy/not creditworthy…
- Each row is a case (customer, tax return, applicant)
- Each column is a variable
- Target variable is often binary (yes/no)

- Deliberately biased classifications: cost of errors

Supervised: Prediction

- Goal: Predict numerical target (outcome) variable
- Examples: sales, revenue, performance
- As in classification:
  - Each row is a case (customer, tax return, applicant)
  - Each column is a variable
- Regression a common tool, but often *not* interested in value of the coefficients *per se*.
  - Instead: forecast outcome for a new case
- Taken together, classification and prediction constitute “predictive analytics”

(Unsupervised) Data Visualization

- Graphs and plots of data
- Histograms, boxplots, bar charts, scatterplots
- Especially useful to examine relationships between pairs of variables

- General concept: Exploratory Data Analysis
  - Where do you *start* with new data?

Pre-processing Data

1. Format conversion e.g. text to numeric
2. Parsing e.g. web data
3. Merging data from multiple sources
4. Dealing with outliers
5. Missing observations (some algorithms don't care)
6. Rare event oversampling
7. Normalizing

- Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets.

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR NY Times AUG. 17, 2014

Convert Variable Types

- Determine the types of pre-processing needed, and algorithms used
- Main distinction: Categorical vs. numeric
- Categorical variables
  - Binary (male/female, student/non-student)
  - Ordered (low, medium, high)
  - Unordered (Ford, Toyota, Honda)

Detecting Outliers

- An outlier is an observation that is “extreme”, being distant from the rest of the data (definition of “distant” is deliberately vague)
- Outliers can have disproportionate influence on models (a problem if it is spurious)
- ~~An important step in data pre-processing is detecting outliers~~
- Once detected, domain knowledge is required to determine if it is an error, or truly extreme.
  - Common example: misplaced decimal point

Handling Missing Data

- Many algorithms will not process records with missing values. Default is to drop those records.
- Solution 1: Omission
  - If a small number of records have missing values, can omit them
  - If many records are missing values on a small set of variables, can drop those variables (or use proxies)
  - If many records have missing values, omission is not practical
- Solution 2: Imputation
  - Replace missing values with reasonable substitutes
  - Lets you keep the record and use the rest of its (non-missing) information
- Solution 3: Use an algorithm that handles missing data (Classification Trees)

Rare event oversampling

- Often the event of interest is rare
- Examples: response to mailing, fraud, …
  - Only a few percent of total sample.
- Sampling may yield too few “interesting” cases to effectively train a model
- A popular solution: oversample the rare cases to obtain a more balanced training set
- Later, need to adjust results for the oversampling

Normalizing (Standardizing) Data

- Needed when variables with the largest scales would dominate and skew results
  - Needed for some algorithms (eg kNN); not for others (regression)
- Puts all variables on same scale
- Is weight in g or kg? Meters or feet or km?
- Normalizing function: Subtract mean and divide by standard deviation
- Alternative: scale to 0-1 by subtracting minimum and dividing by the range
  - Useful when the data contain dummies and numeric
  - Sometimes best *not* to normalize. More insight from coefficient values.

Overfitting (cont.)

Causes:

- Too many predictors
- A model with too many parameters
- Trying many different models

Consequence: Deployed model will not work as well as expected with completely new data.

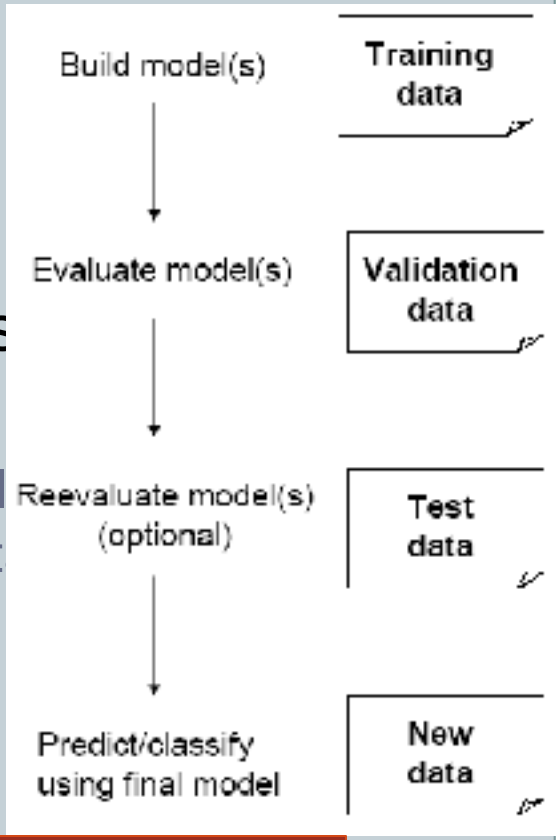
Partitioning the Data

Problem: How well will our model perform with new data?

Solution: Separate data into two parts

- Training partition to develop the model
- Validation partition to implement the model and evaluate its performance on "new" dat

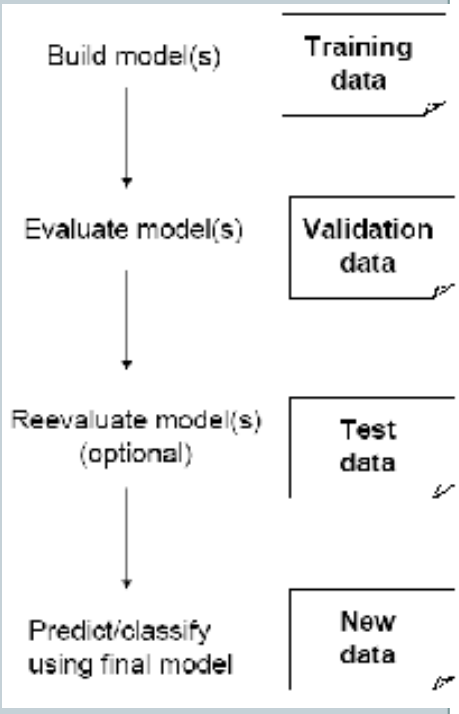
Addresses the issue of overfitting



A Big Idea in Data Analytics

Multiple Partitions

- When a model is developed on **training data**, it can overfit the training data (hence need to assess on validation)
- Assessing multiple models on same **validation data** can *overfit* validation data ("p hacking")
- Some methods use the validation data to choose a parameter. This too can lead to overfitting the validation data
- Solution: final selected model is applied to a third ***test*** **partition.** Realistic estimate of its performance on new data



Concept *starting* to be used in classic regression analysis

- Instead of trying to estimate the forecast errors, just measure them!
  - Standard error of residuals
- Statistical estimates of errors have long list of assumptions:
  - Homoskedastic errors
  - No autocorrelation
  - No important omitted variables (hah)
- Cross-validation concept:
  - Train (60%) Validate (20%) Test(20%) sample
  - Then shuffle the 3 samples, and repeat

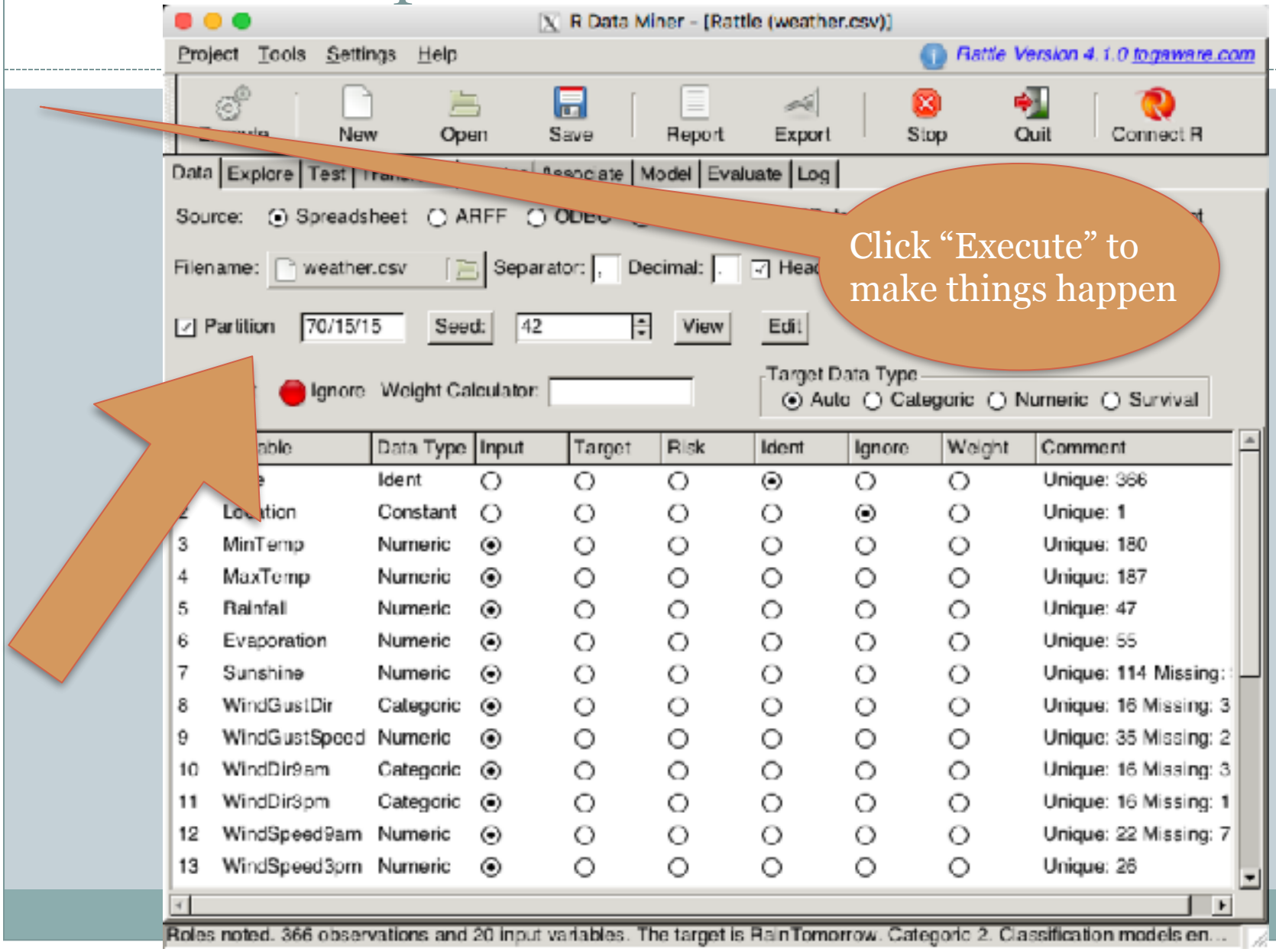
Summary

- Data Mining includes many supervised methods (Classification & Prediction) + some unsupervised methods (Association Rules, Data Reduction, Data Exploration & Visualization)
- Before algorithms can be applied, data must be characterized and pre-processed. *This takes work! And thought! And creativity!*
- To evaluate performance and to avoid overfitting, partition the data
- Data mining methods are applied to a part of a large dataset, and then the best model is used to analyze the rest of dataset

Next week: Run some analyses w Rattle

- Update your software e.g. MacOS 10.12
- Start RStudio
- install.packages("Rattle") # only once
- # It will ask to install other packages. "Yes" to all
- library(rattle)
- rattle()
- DMRR textbook Chapter 3
- In RStudio:
  - dim(weather) #Comes with the Rattle package

Import *weather* into Rattle

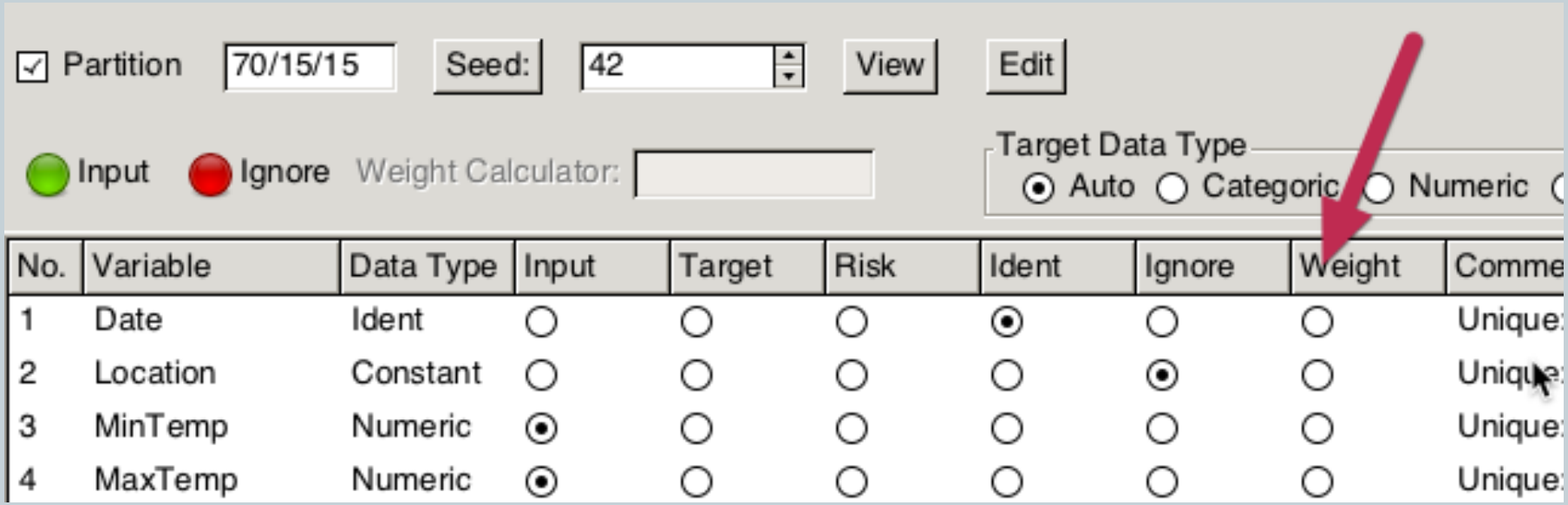


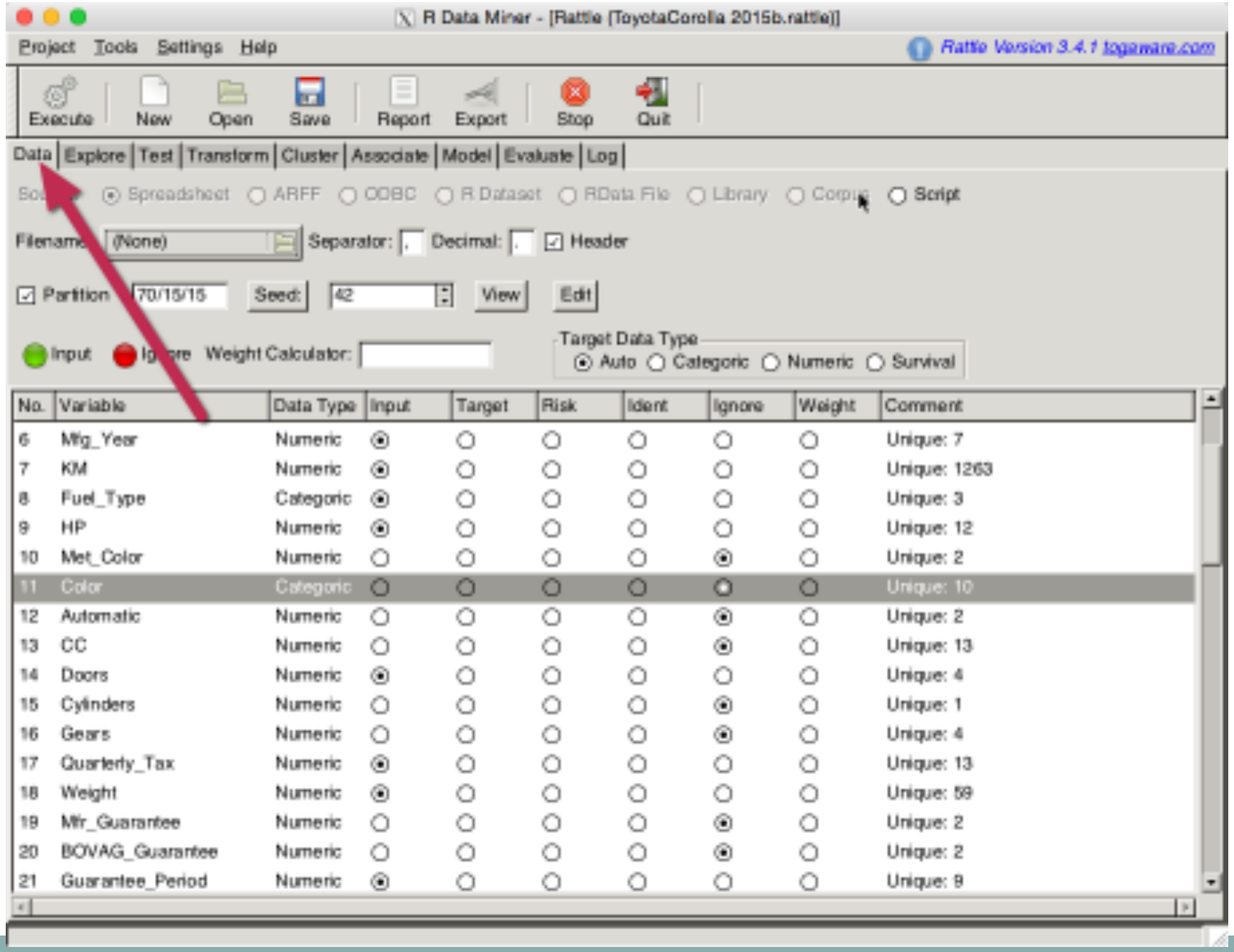
Rattle Text DMRR by Williams

- Chapter 2 tutorial; Chapter 4 on loading data
  - Data frame *weather* is built into Rattle
  - Remember to click on "Execute" button
- Our first model: Decision Trees in DMRR Chapter 11
- Do decision tree tutorial, Section 11.4

- Next Thursday: Use more interesting data, more interesting analysis. Toyota price data
- Get started by loading + look at Toyota price data.

Partitioning the data




R Data Miner - [Rattle (ToyotaCorolla 2015b.rattle)]
Project Tools Settings Help
Rattle Version 3.4.1 togaware.com
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Spreadsheet ARFF ODBC R Dataset RData File Library Corpus Script
Filename (None) Separator: , Decimal: . Header
Partition 70/15/15 Seed: 42 View Edit
Input Ignore Weight Calculator:
Target Data Type
Auto Categoric Numeric Survival
No. Variable Data Type Input Target Risk Ident Ignore Weight Comment
6 Mfg_Year Numeric Unique: 7
7 KM Numeric Unique: 1263
8 Fuel_Type Categoric Unique: 3
9 HP Numeric Unique: 12
10 Met_Color Numeric Unique: 2
11 Color Categoric Unique: 10
12 Automatic Numeric Unique: 2
13 CC Numeric Unique: 13
14 Doors Numeric Unique: 4
15 Cylinders Numeric Unique: 1
16 Gears Numeric Unique: 4
17 Quarterly_Tax Numeric Unique: 13
18 Weight Numeric Unique: 59
19 Mfr_Guarantee Numeric Unique: 2
20 BOVAG_Guarantee Numeric Unique: 2
21 Guarantee_Period Numeric Unique: 9