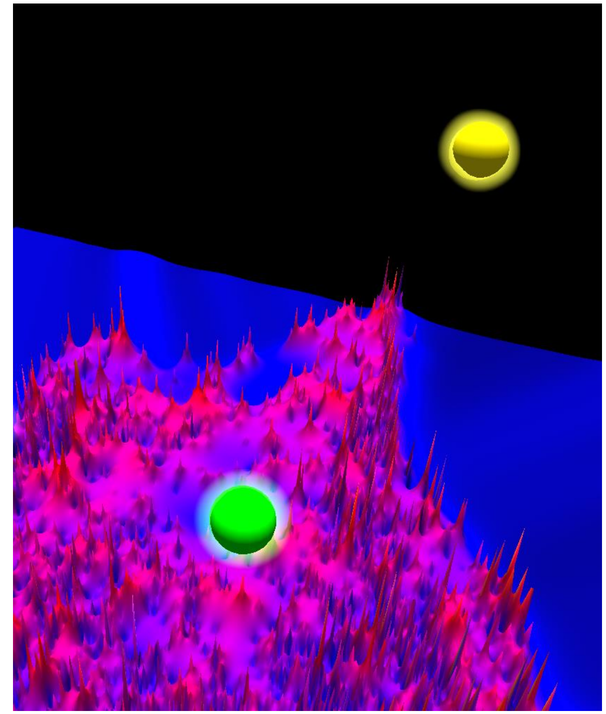
Business Analytics
In Action

Visual Analytics and Data Mining for Business

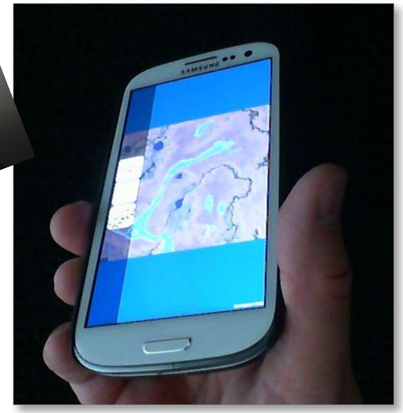
A long story of three cases and three tools

- **Background**
- **About Jacob**
- **Data analytics and model building**
- **Data visualisation for insight**
- **Analytic process and its design**
- **Analytics tools and technology**
- **Professional library**
- **Sensemaking and decision making**
- **Hands on problem solving**
- **Value of information and analytics**
- **Sensemaking framework**
- **Summary**



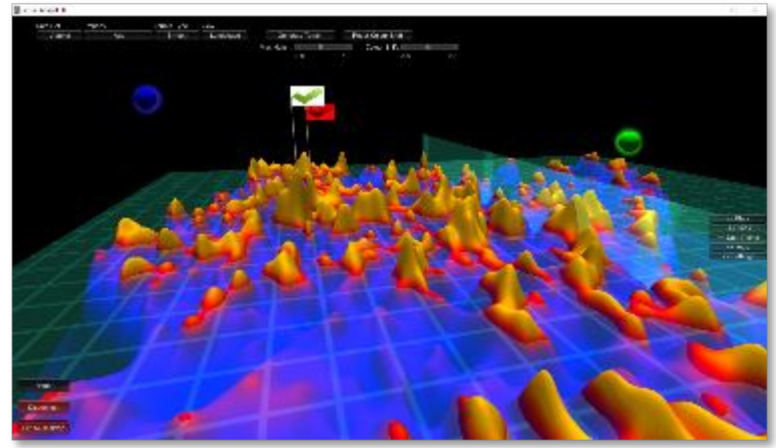
Assoc. Prof. Jacob L. Cybulski
Director of Research
Director of SAS Visual Analytics Collaboratory
Department of IS and Business Analytics
Deakin University
Burwood, Australia

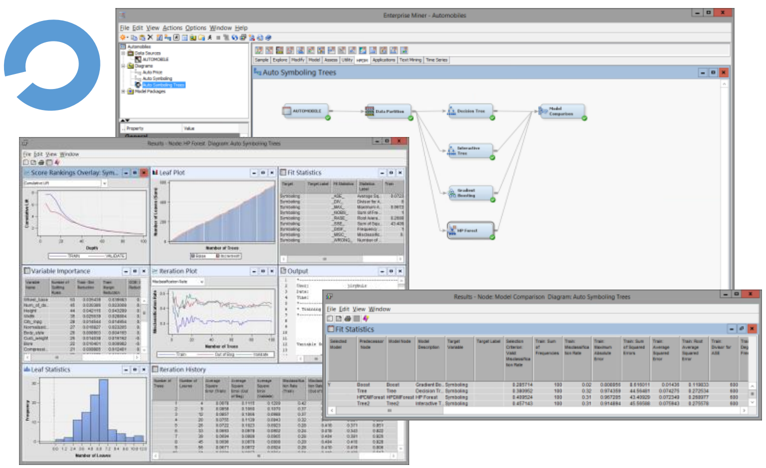
About Jacob's Research



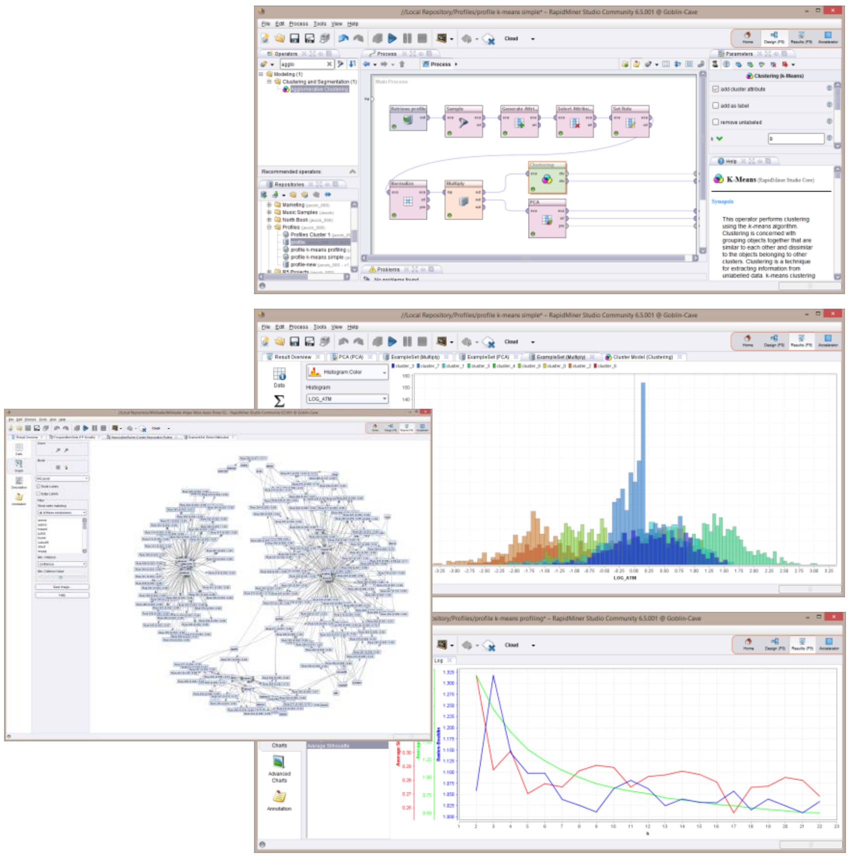
Natural User Interfaces

Collaborative Visual Analytics in 3D





Predictive Analytics

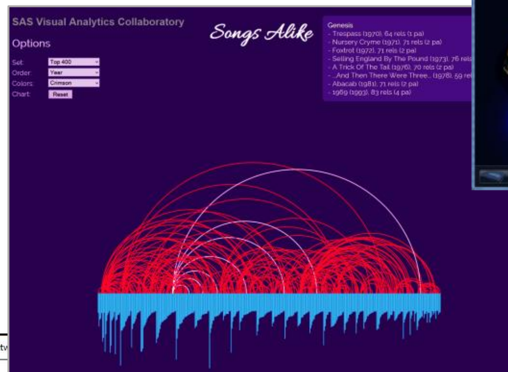


Text and Data Mining

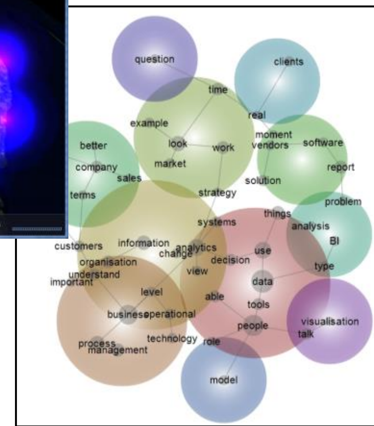
Sample of Jacob's Data Visualisations

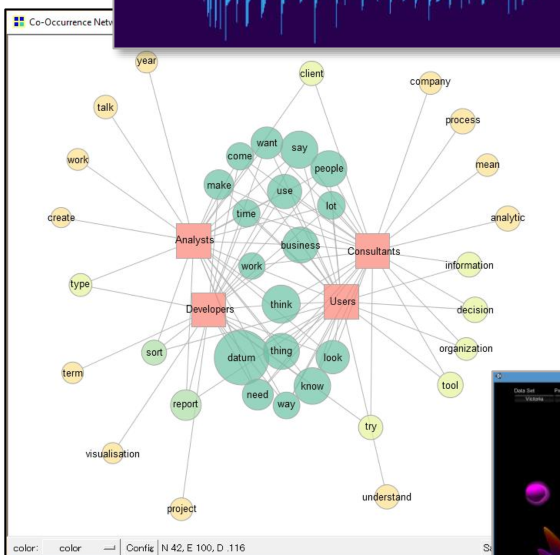
Movie Ticket Sales / WWT Layerscape

Music distribution / JavaScript + D3



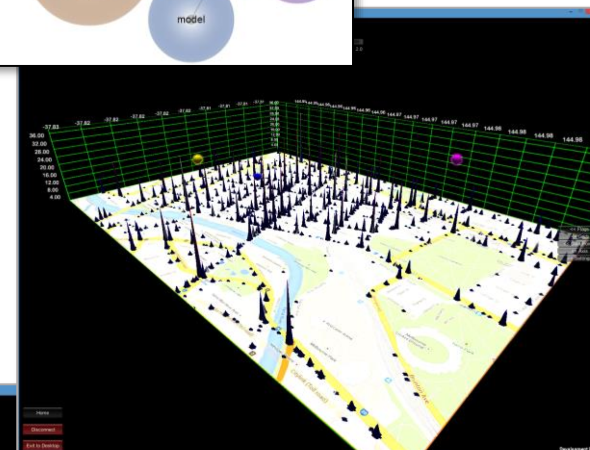
Assignments / Leximancer



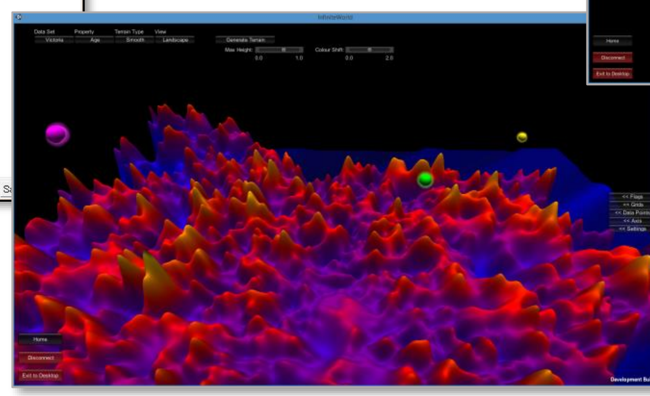


BI Interviews / KH Coder

First Person Art / jMonkey Engine



Traffic Accidents / Visual Analyst 3D in Unity3D



Sensemaking vs Decision Making

Sensemaking Karl Weick (1993, 1995, 2005)	Decision Making Richard Boland (2008)
Sensemaking is a structured process of dealing with ambiguity and uncertainty in organizational settings, aimed at giving meaning to objects and events from the past	Decision-making is a process aiming to evaluate a range of possible actions and to select the best alternative
Sensemaking defines an ongoing socio-cognitive activity that is initiated by organizational actors seeking understanding and control of their environment	Decision making is directed almost completely and without exception to the future impact of decisions, actions and their outcomes
Sensemaking focuses on continuous generation of insights	Decision making focuses on making choices at a specific instance of time
Sensemaking is the prerequisite of informed decision-making (Namvar and Cybulski 2015, 2016)	

Visualisation vs Data Model

Data Science /
Data Analytics

Data Science / Data Analytics is the systematic study of extracting actionable knowledge from data.
(Dhar 2013, CACM V56N12)

Data science relies on methods drawn from many disciplines, e.g.:

- Mathematics
- Statistics
- Operations research
- Information science
- Computer science
- Artificial intelligence
- Data visualisation
- Databases
- Data warehousing
- High performance computing

The main purpose:

- Sensemaking
- Decision making

Typical approaches to data analytics:

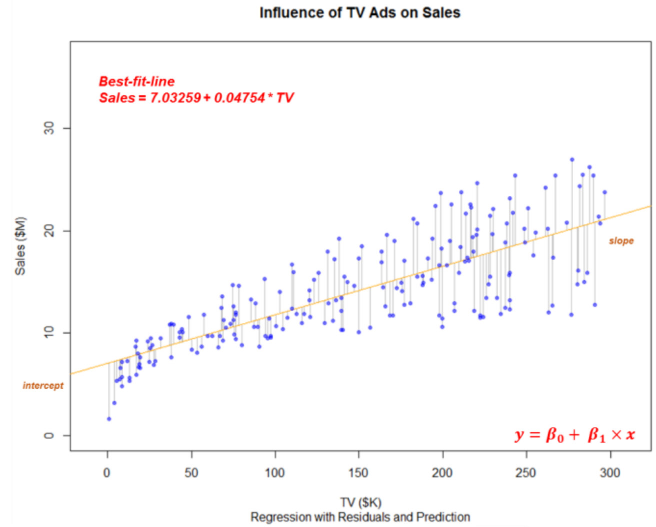
- Statistical methods
  - Linear regression model
  - Logistic regression
  - General linear models
  - Multivariate adaptive regression splines (MARS)
  - Naïve Bayes models
  - Bayesian modelling
  - Association analysis
  - Time series analysis
  - Anomaly analysis
- Machine Learning
  - Decision trees
  - Neural networks
  - Cluster analysis
  - Text mining
  - Support vector machines
  - Genetic algorithms
  - Induction and deduction

Model Building

Data analytics focuses on building and testing of models based on the existing data in order to determine patterns, explain the past and predict future outcomes and trends.

Modern businesses have access to very large data sets, often collected by other organisations and also available in open data repositories. Sometimes the data covers the entire population.

Examples presented set a framework for problem solving by analysing large data sets, leading to more refined outcomes and corrective actions.



Applications

Marketing Effectiveness

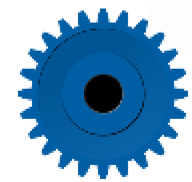
Analysis of Learning Outcomes

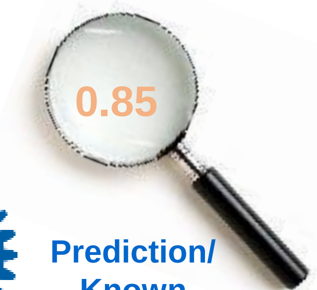
Financial Advice

Fraud Investigation

Disease Diagnosis

Individual Characteristics





Prediction/ Known Outcome

Predictive/Explanatory Model

Mechanics

Creation of analytic models is key to analytics success

Visualisation of data and results generated by the model provides much needed intuition

Aims

Predicting future, acting in the present and explaining the past

Typical Data Mining Process

CRISP-DM

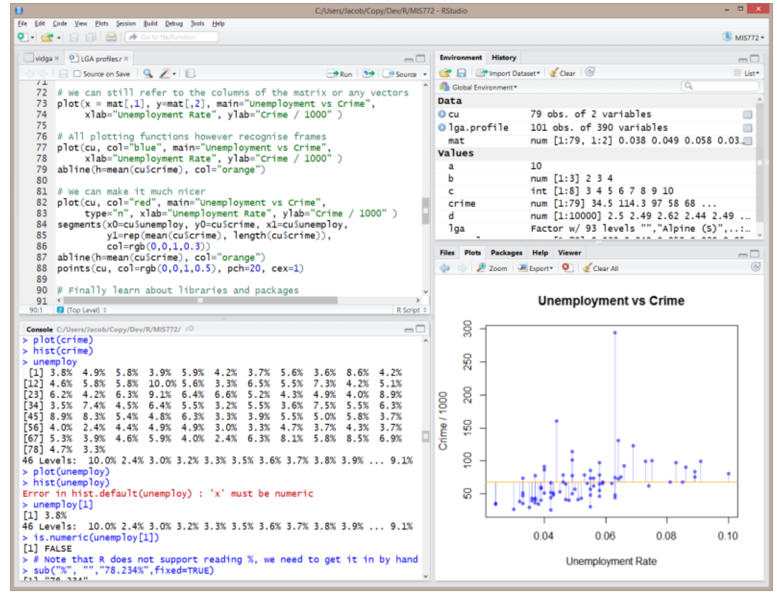
- **Business Understanding:**
  stating project objectives and requirements into a data mining problem.
- **Data Understanding:**
  getting familiar with the data and its interesting features.
- **Data Preparation:**
  getting data ready for modelling,
  to include selection of variables, dealing with errors and omissions, and transforming data to suit the method.
- **Modeling:**
  various techniques are selected
  and applied, and their parameters are optimised.
- **Evaluation:**
  ensuring that the model meets business objectives in terms of its function and the quality of produced results.
- **Deployment:**
  applying the model in practice to solve similar problems using newly collected data.
- **All steps in this process are important, each step in the process is complex, which requires significant effort in its planning and later execution.**

Cross Industry Standard Process for Data Mining

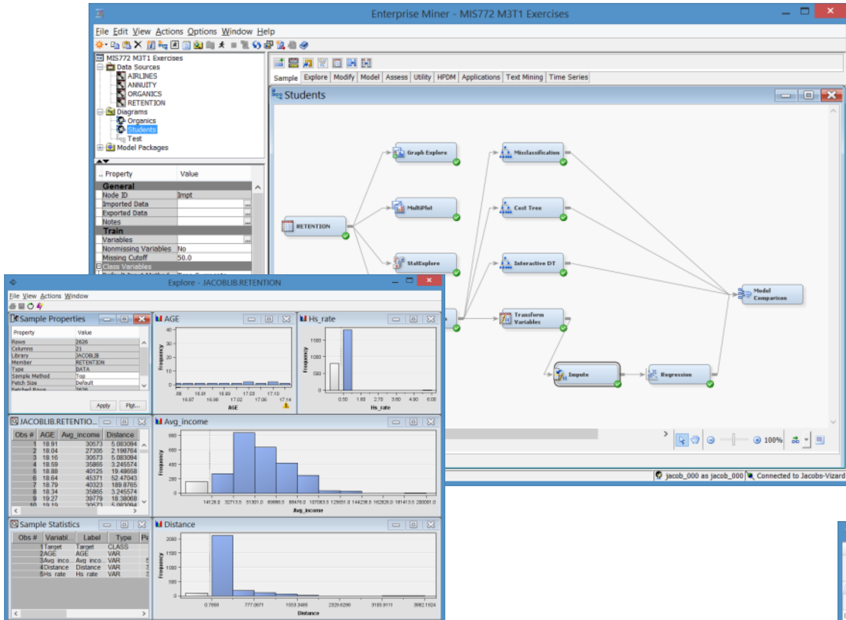
Modern data mining / data analytics tools provide facilities to plan the entire analytic workflow, so that it is reusable and able to produce repeatable results.

Tools of Choice: Model + Visualisation + Process

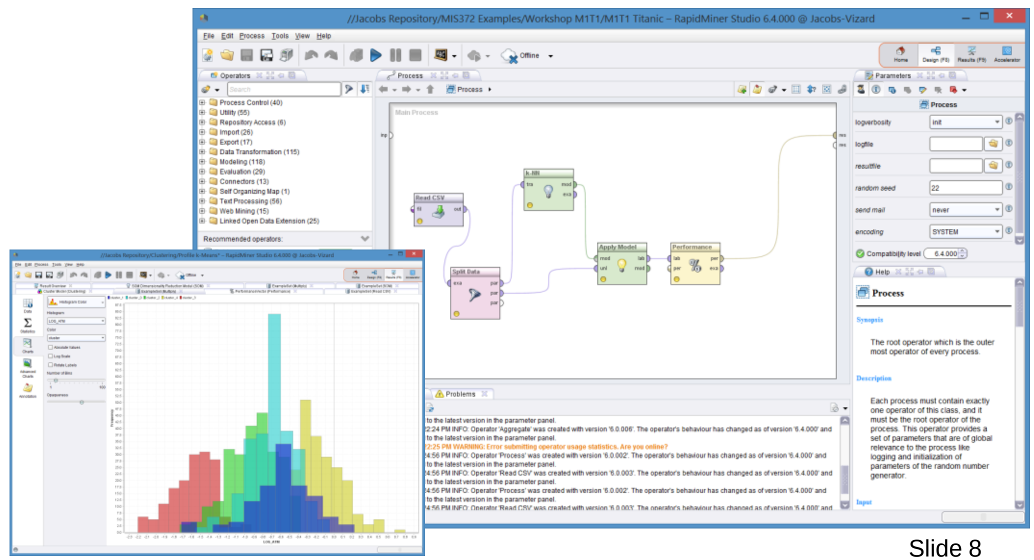
R / MRO / R Studio – Open source statistical software with a programming language and rich libraries



SAS Enterprise Miner / BASE – Commercial defacto industry standard in data mining



RapidMiner Studio – Open source / commercial software with visual analytic process, flexible integration framework and great charts



Other Popular Tools -

- Python + Orange + Anaconda
- KNIME / WEKA
- IBM Watson / SPSS Modeler
- MS Cortana Intelligence / Power BI
- SAP BusinessObjects
- Oracle BI

Sample Project: R with R Studio

Identify predictors of health to reduce the severity of the world's health problems.

The World Bank approached you to assist in the identification of the national-level health quality indicators, which are not directly linked with health care expenditure but rather those hidden in the socio-economic aspects of peoples' living conditions. The World Bank seeks to develop a model of health outcomes, which would be capable of predicting the effects of global social, environmental and economic changes on the lives of people in different countries. They would also like to determine a course of action aimed at improving the situation in the countries most affected by such changes.

You have been asked to identify a number of health quality predictors and subsequently build a k-NN classifier, Regression and Neural Network models in R to predict, evaluate and visualise (on Google Maps) health quality across the world. Suggest a course of action to address the world's health problems.

Select Variables: **Identify several socio-economic predictors of different types of health outcomes.**

Explore Data: **Visualise your data using Google Maps (combined with k-NN insights).**

Analyse Data: **Use correlation in R to establish if there are any interactions between the selected variables. Address the issue of multi-collinearity**

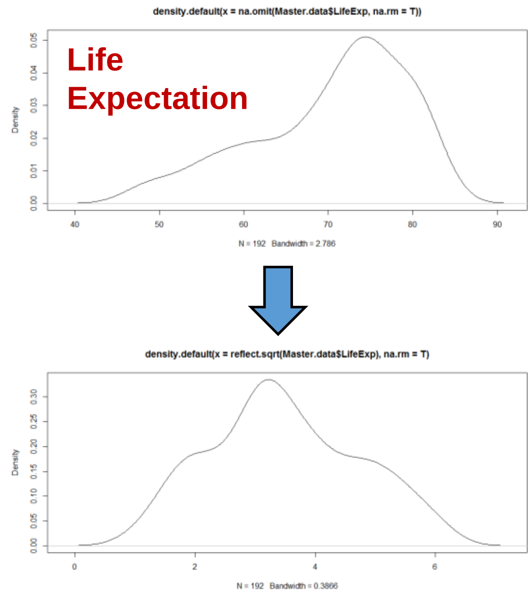
Create Predictive Models: **Create and evaluate predictive models using k-NN and regression methods. Compare all models and their performance.**

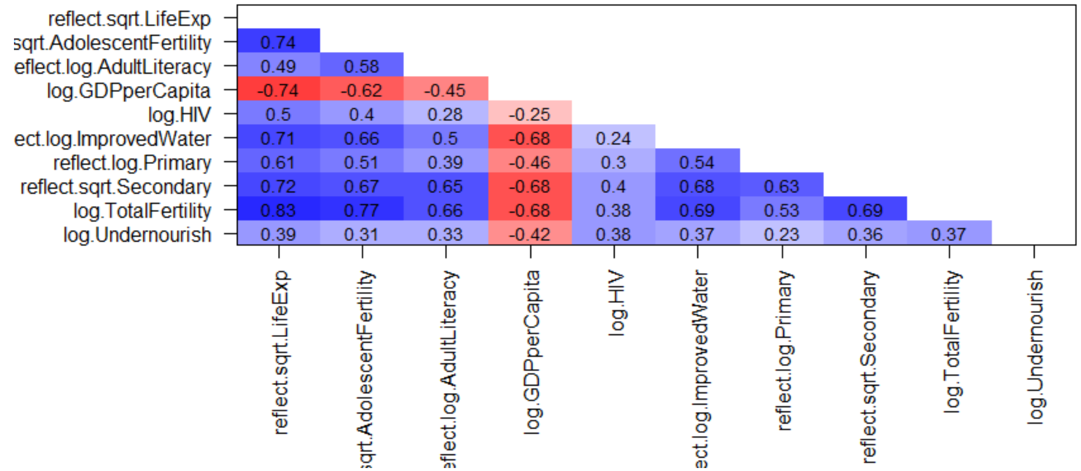
Report: **Report your results and propose a course of action.**

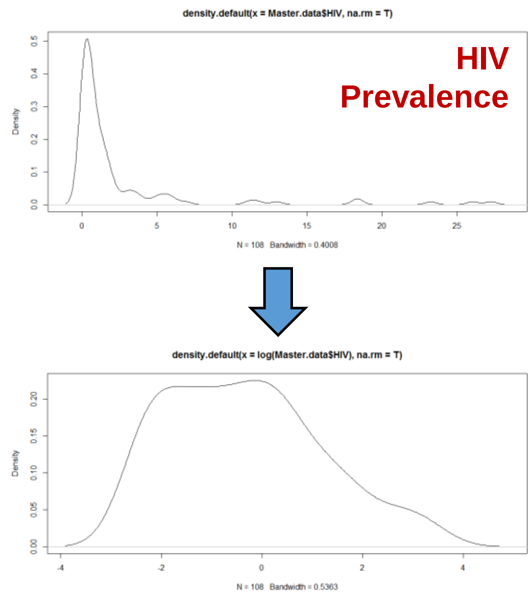
DEAKIN UNIVERSITY

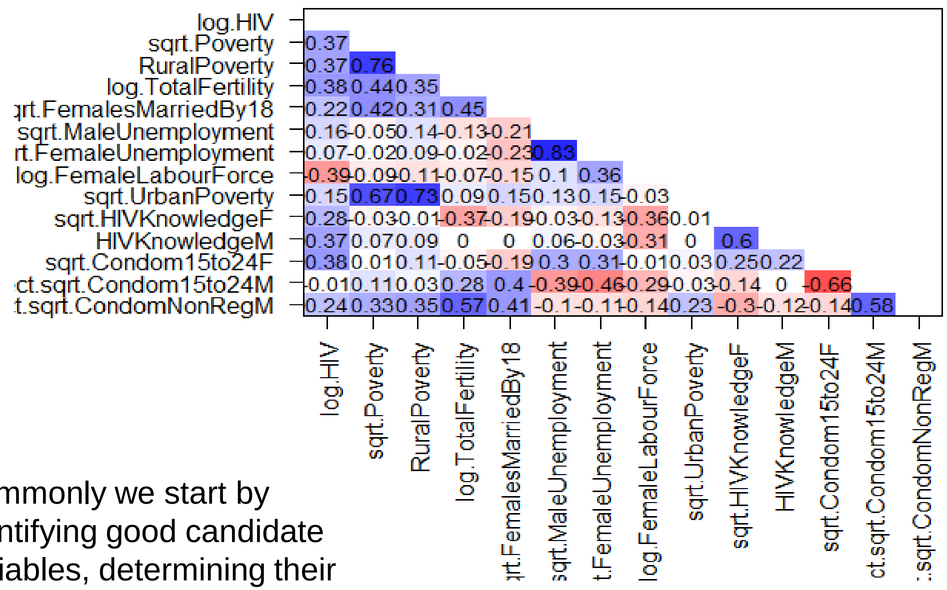
Data Exploration + Data Transformation

(Acknowledgements Daniel Loden 2016)









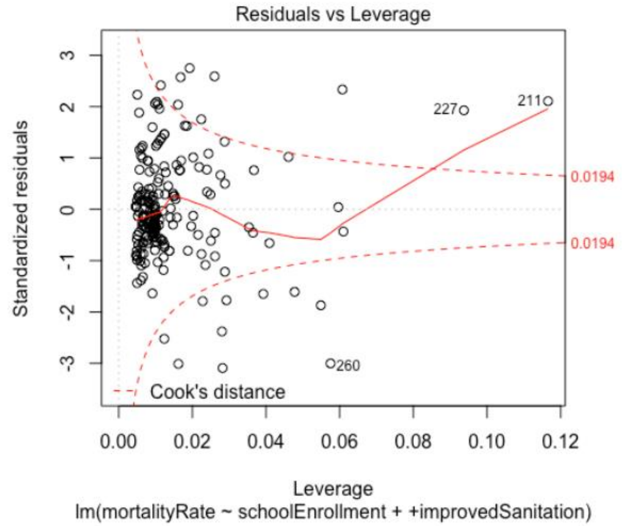
Commonly we start by identifying good candidate variables, determining their relationships and if needed transforming them in this process.

Regression Model Creation
Selection of Predictors

Target: Life expectancy (RF, SQ)			
Predictor	Standardised beta coeff.	p-value	VIF
Adolescent fertility (SQ)	0.129	0.025	2.754
Adult literacy (RF, LN)	-0.188	0.000	1.962
GDP per capita (LN)	-0.252	0.000	1.430
HIV prevalence (LN)	0.179	0.000	1.859
Primary school enrolment (RF, LN)	0.142	0.001	1.214
Total fertility (LN)	0.562	0.000	3.223
LN: natural logarithm; SQ: square-root; RF: reflected			
Adjusted R^2 = 82.7%; F = 117.3; Validation set correlation = 0.895; Training n = 147; Validation n = 63			
Notes on residuals: homoscedastic; approximately normal; no influential outliers			

Total fertility is by far the strongest predictor of life expectancy. Increased fertility is associated with a lower average life expectancy (considering the reflection), holding the other predictors constant.

Cook's distance can also be used here to detect and remove extreme cases from the data set.

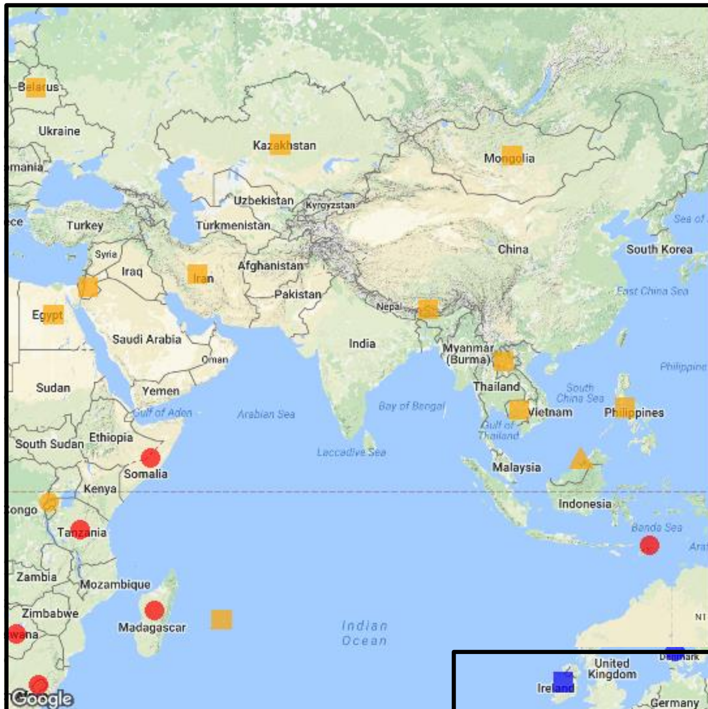


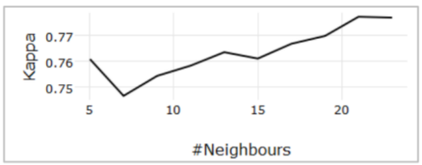
Target: HIV prevalence (LN)			
Predictor	Standardised beta coeff.	p-value	VIF
Total fertility (LN)	0.613	0.000	1.383
Female unemployment (SQ)	0.293	0.000	1.295
Female labour force (RF, LN)	-0.352	0.000	1.584
Knowledge of HIV amongst females (SQ)	0.365	0.000	1.657
Condom use amongst 15 to 24 year-old females (SQ)	0.225	0.002	1.263
LN: natural logarithm; SQ: square-root; RF: reflected			
Adjusted R^2 = 70.7%; F = 36.7; Validation set correlation = 0.786; Training n = 75; Validation n = 33			
Notes on residuals: homoscedastic; approximately normal; no influential outliers			

Total fertility is by far the strongest predictor of HIV prevalence, with this outcome increasing in line with fertility, when holding the other predictors constant.

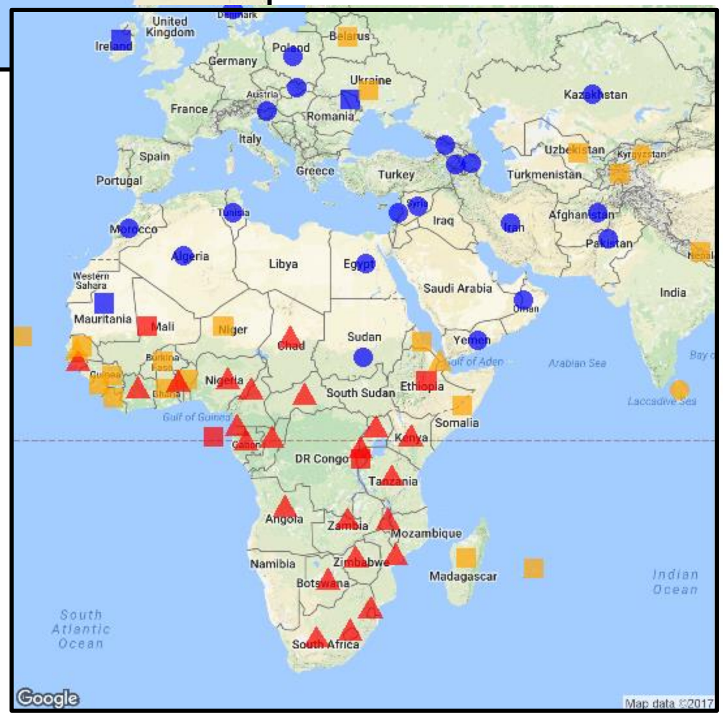
If you check these variables, you'd think twice if indeed they are good "predictors", or something different!

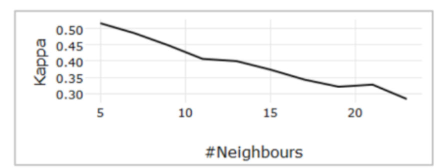
Validation Visualised





Classifier	Accuracy	Kappa
k-NN	69.8%	0.517
Naïve Bayes	66.3%	0.420
Gradient Boosting Machines	71.2%	0.532
Random Forest	75.5%	0.601





An interesting way of deploying visualization is for model validation. Not only we can see if the model predicts well but also where it fails to predict accurately.

Demo Project: Prediction

Predict Litigation for Compensation Recovery.

A significant portion of a company's loss-expense ratio goes to defending disputed claims. A major insurance company was concerned about the rising cost of bodily injury claims. They want to reduce the cost of litigation by analysing its transactional data and creating a predictive model that could forecast which customers are more likely to engage lawyers. Such capability is likely to result in lower claims settlements and reduced loss ratios.

Create a predictive model in SAS Enterprise Miner using both structured and unstructured data of the past worker's compensation claims to determine the likelihood of claim litigation and the consequent subrogation. Use several different modelling approaches and select the most effective one or use all of them simultaneously in an ensemble.

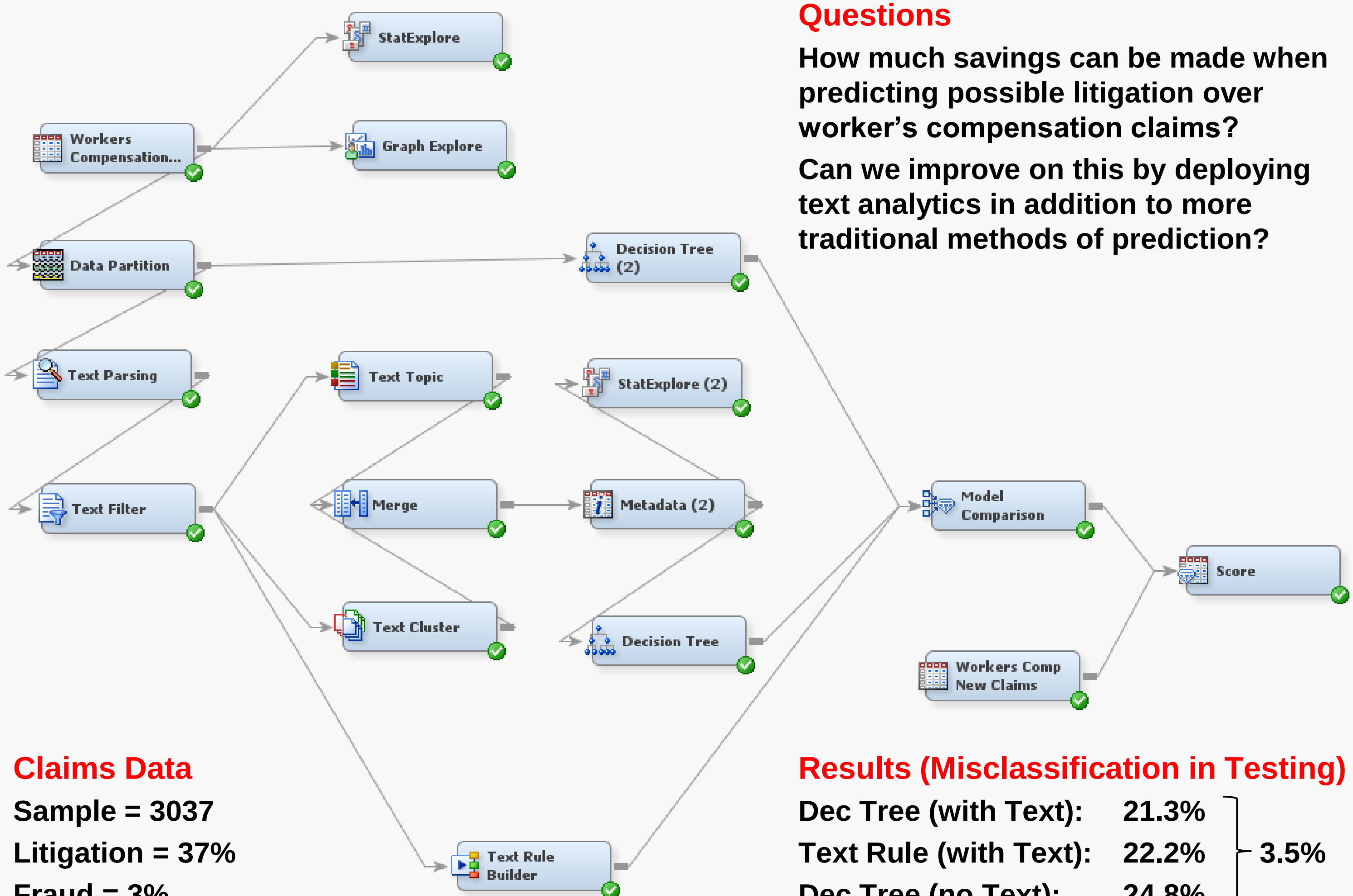
Structured Models: **Create a number of predictive models (e.g. Neural Nets, Regression and Decision Trees) based on the structured data, evaluate and optimise their performance**

Text Analysis Models: **Perform cluster and topic analysis of the provided text. Evaluate the model performance**

Model Integration: **Create an ensemble model integrating recommendation of all models**

Predicting Litigation for Compensation Recovery

This SAS Enterprise Miner model combines structured and unstructured data to predict possible litigation to recover worker's compensation claims, which could add over $200k to the cost of a claim (valid or invalid).



Questions

How much savings can be made when predicting possible litigation over worker's compensation claims?

Can we improve on this by deploying text analytics in addition to more traditional methods of prediction?

Claims Data

Sample = 3037

Litigation = 37%

Fraud = 3%

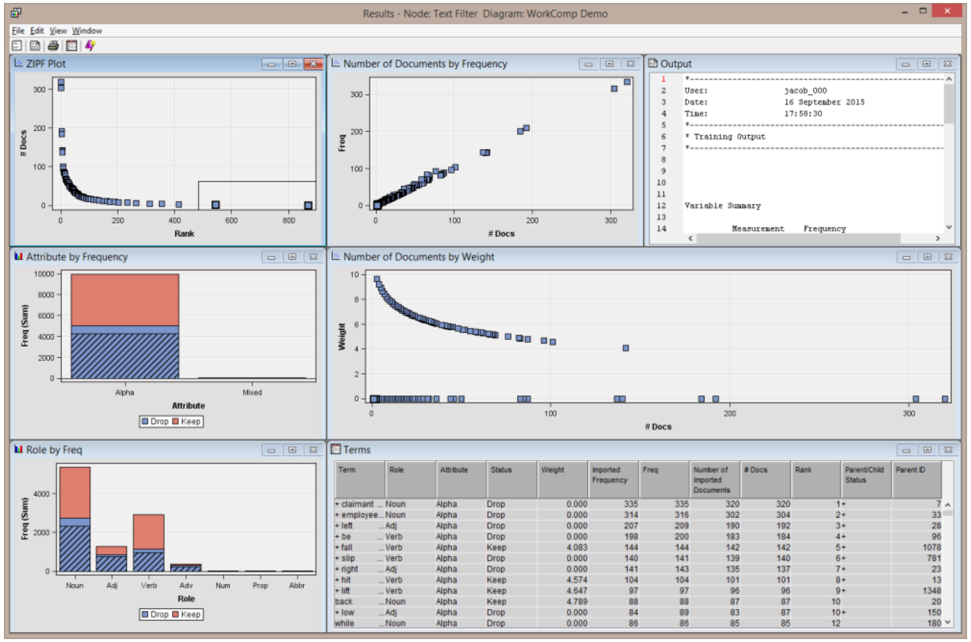
Results (Misclassification in Testing)

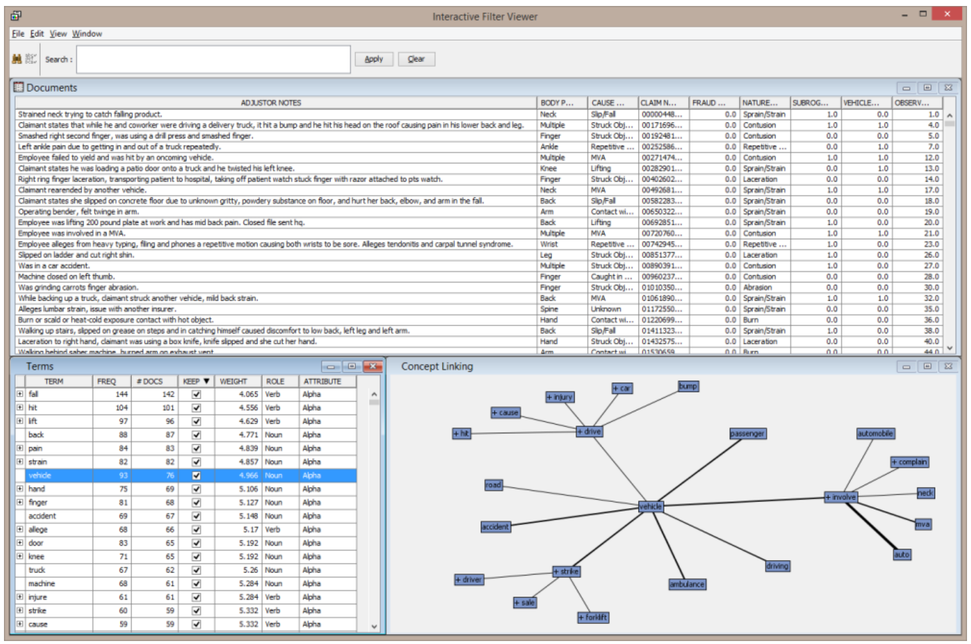
Dec Tree (with Text):	21.3%
Text Rule (with Text):	22.2%
Dec Tree (no Text):	24.8%

$\left.\right\}$ 3.5%

DEAKIN UNIVERSITY

Adding Text to a Model

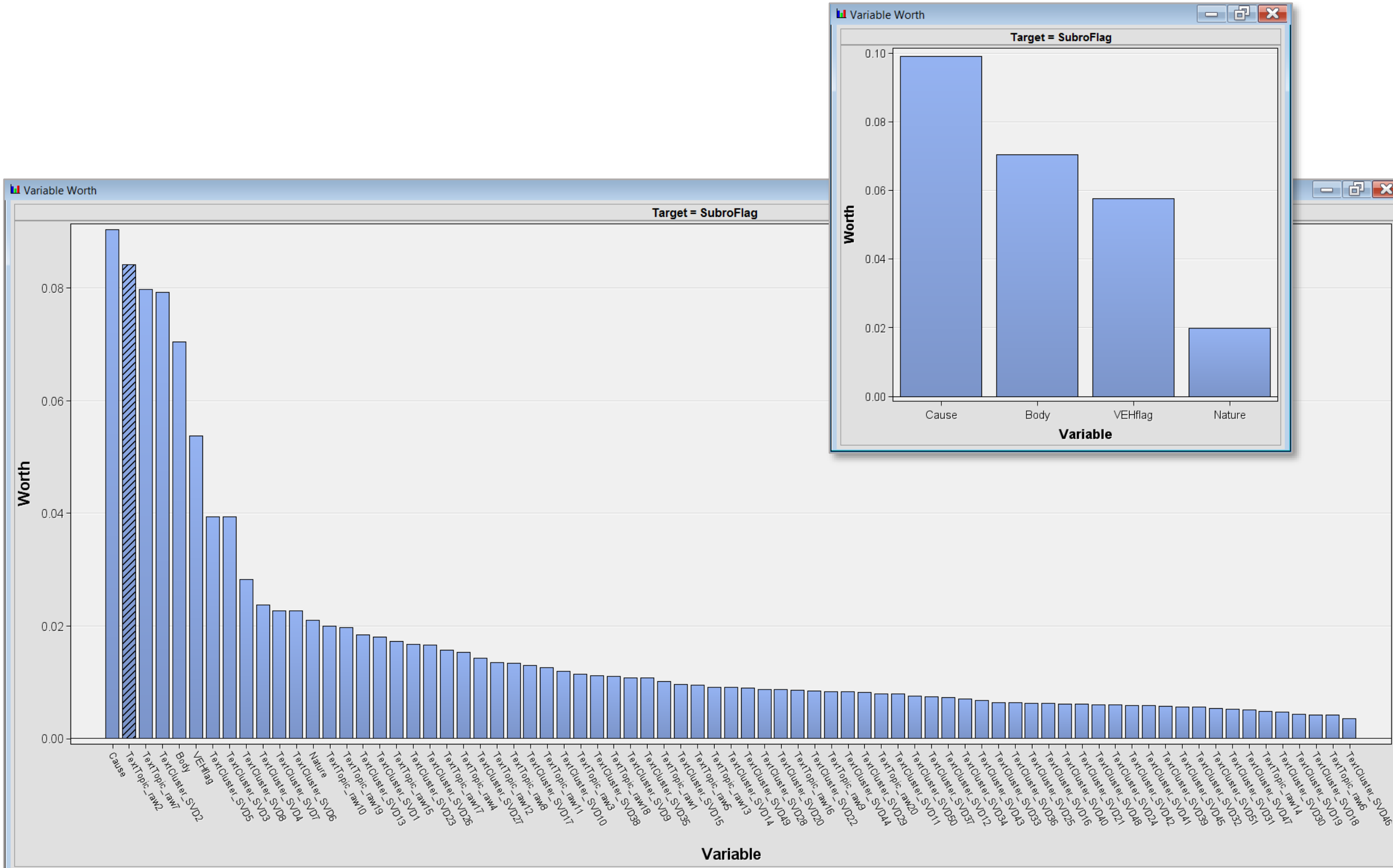




The main aim of text analytics is to convert text variables into a collection of structured variables that could be used in prediction. This process involves:

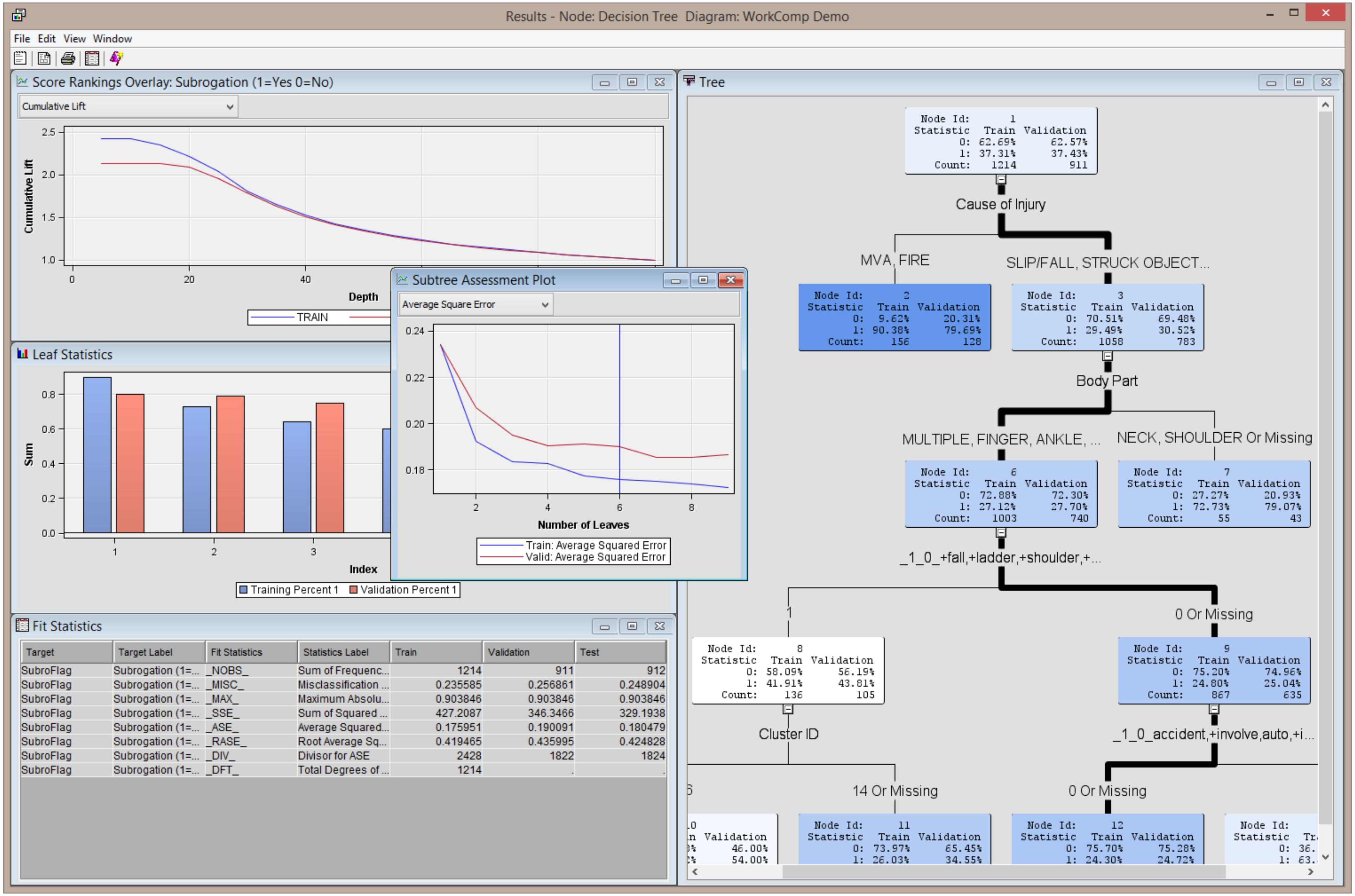
- Preparing data
- Parsing text variables to identify significant terms
- Filtering terms to create vector representations of text where the terms act as document variables
- Clustering term variables to reduce dimensionality
- Creation of topic variables which represent co-occurring terms
- Use of structured variables, cluster and topic variables to create a predictive model
- Model validation, testing and scoring

Explore and Prepare Data



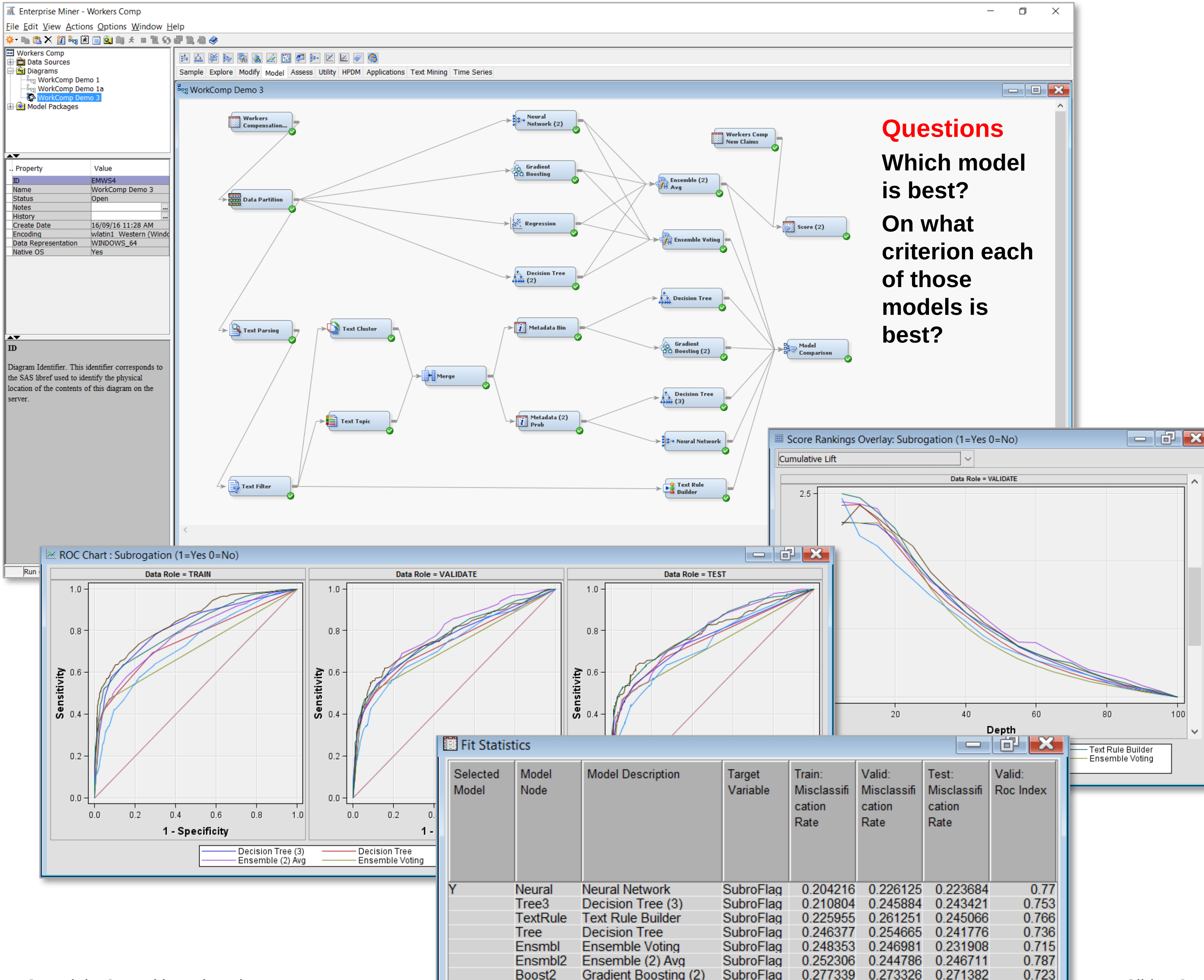
The initial analysis of Workers Compensation data shows the importance of structured variables via their logworth for predicting subrogation, e.g. "Cause" and "Body" (injury). However, as soon as text variables are added two topic and one cluster variables are now considered of higher importance than "Body", which is clearly captured within the processed text.

Developing and optimization of predictive models



A model such as a Decision Tree can be developed and tested to assess its performance. We can also analyse the model structure to determine the impact of text vs structured variables on the produced results.

Integration of Several Models into One Process



Selected Model	Model Node	Model Description	Target Variable	Train: Misclassification Rate	Valid: Misclassification Rate	Test: Misclassification Rate	Valid: Roc Index
Y	Neural	Neural Network	SubroFlag	0.204216	0.226125	0.223684	0.77
	Tree3	Decision Tree (3)	SubroFlag	0.210804	0.245884	0.243421	0.753
	TextRule	Text Rule Builder	SubroFlag	0.225955	0.261251	0.245066	0.766
	Tree	Decision Tree	SubroFlag	0.246377	0.254665	0.241776	0.736
	Ensmbl	Ensemble Voting	SubroFlag	0.248353	0.246981	0.231908	0.715
	Ensmbl2	Ensemble (2) Avg	SubroFlag	0.252306	0.244786	0.246711	0.787
	Boost2	Gradient Boosting (2)	SubroFlag	0.277339	0.273326	0.271382	0.723

Questions

Which model is best?

On what criterion each of those models is best?

Demo Project: Sensemaking

Understand the characteristics of customers for marketing purposes.

It is a common practice to survey customers visiting a store to identify their characteristics, which could subsequently be used for marketing purposes, e.g. to target groups of customers with offers specifically tailored to their needs. This dataset contains a survey of 6,876 customers visiting a shopping mall in San Francisco Bay area.

Create an exploratory model in RapidMiner Studio using the survey data to segment the customers based on 13 demographics attributes, which can also be used to estimate income.

An alternative to a survey, customers can also be studied based on their past shopping behaviour, their use of loyalty schemes, online navigation and click throughs, etc.

Cluster Models: **Use k-mean clustering of data with a view to create a marketing campaign targeting specific segments of customers.**

Model Evaluation: **Evaluate the cluster model and determine the optimum number of clusters for the purpose.**

Predictive Models: **Use data clusters as new variables useful in predicting customer income.**

Plan of Action

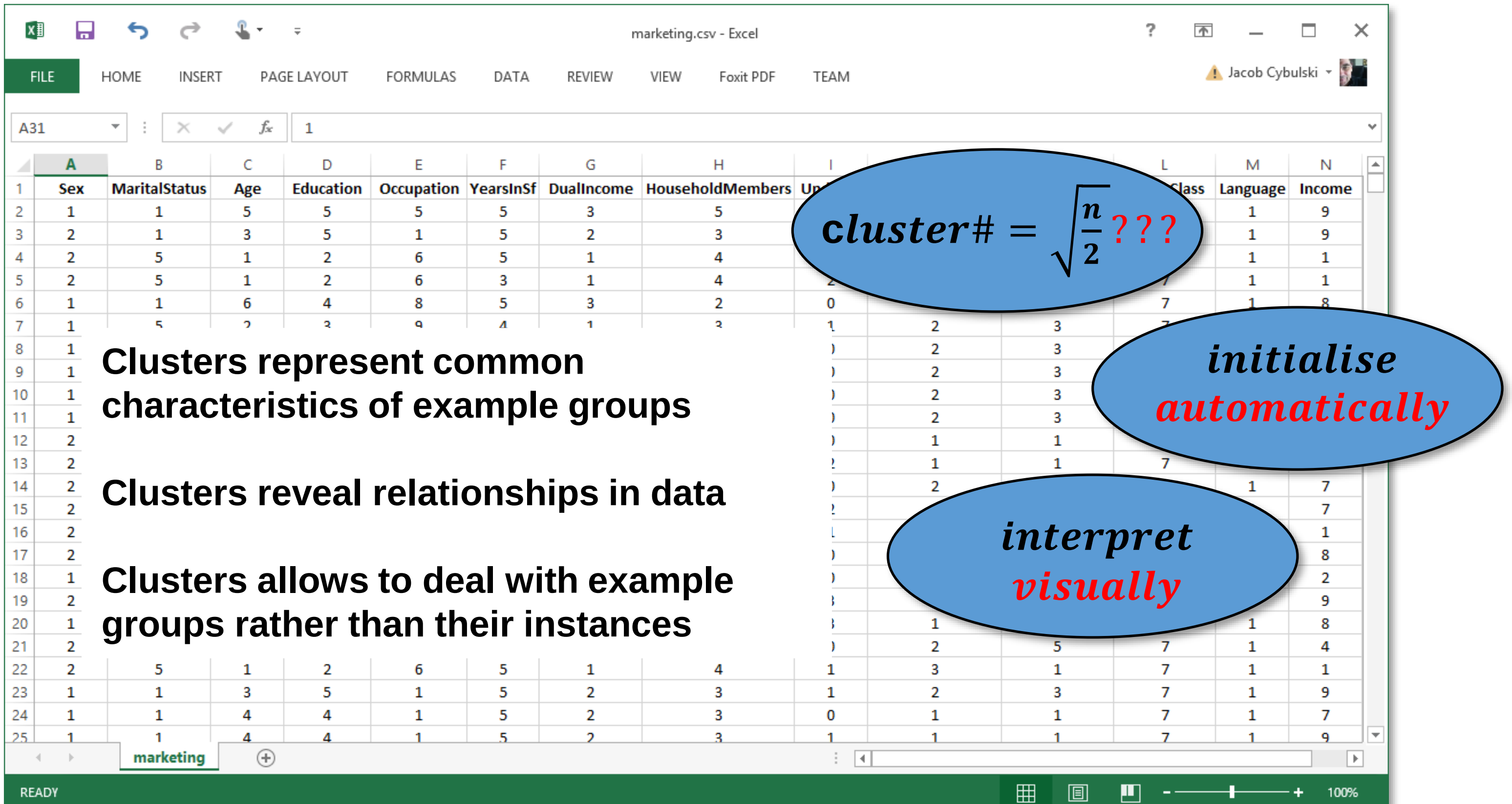
This dataset contains data from a survey of customers in a shopping mall in the San Francisco Bay area.

The goal is to identify segments of customers based on 13 demographics attributes, which can be used to estimate income.

First: What kind of problem would clustering of this data solve?

Method: k-Means, which searches for centers of clusters

Initial question: How many clusters?



Clusters represent common characteristics of example groups

Clusters reveal relationships in data

Clusters allows to deal with example groups rather than their instances

Maximum Number of Clusters to Consider?

1. HOUSEHOLD INCOME PA
1. Less than $10,000
2. $10,000 to $14,999
3. $15,000 to $19,999
4. $20,000 to $24,999
5. $25,000 to $29,999
6. $30,000 to $39,999
7. $40,000 to $49,999
8. $50,000 to $74,999
9. $75,000 or more

2. SEX
1. Male
2. Female

3. MARITAL STATUS
1. Married
2. Living together, not married
3. Divorced or separated
4. Widowed
5. Single, never married

4. AGE
1. 14 thru 17
2. 18 thru 24
3. 25 thru 34
4. 35 thru 44
5. 45 thru 54
6. 55 thru 64
7. 65 and Over

5. EDUCATION
1. Grade 8 or less
2. Grades 9 to 11
3. Graduated high school
4. 1 to 3 years of college
5. College graduate
6. Grad Study

6. OCCUPATION
1. Professional/Managerial
2. Sales Worker
3. Laborer/Driver
4. Clerical/Service Worker
5. Homemaker
6. Student, HS or College
7. Military
8. Retired
9. Unemployed

7. HOW LONG LIVED IN SF AREA?
1. Less than one year
2. One to three years
3. Four to six years
4. Seven to ten years
5. More than ten years

8. DUAL INCOMES (IF MARRIED)
1. Not Married
2. Yes
3. No

9. PERSONS IN YOUR HOUSEHOLD
1. One… 9. Nine or more

10. PERSONS IN HOUSEHOLD UNDER 18
0. None… 9. Nine or more

11. HOUSEHOLDER STATUS
1. Own
2. Rent
3. Live with Parents/Family

12. TYPE OF HOME
1. House
2. Condominium
3. Apartment
4. Mobile Home
5. Other

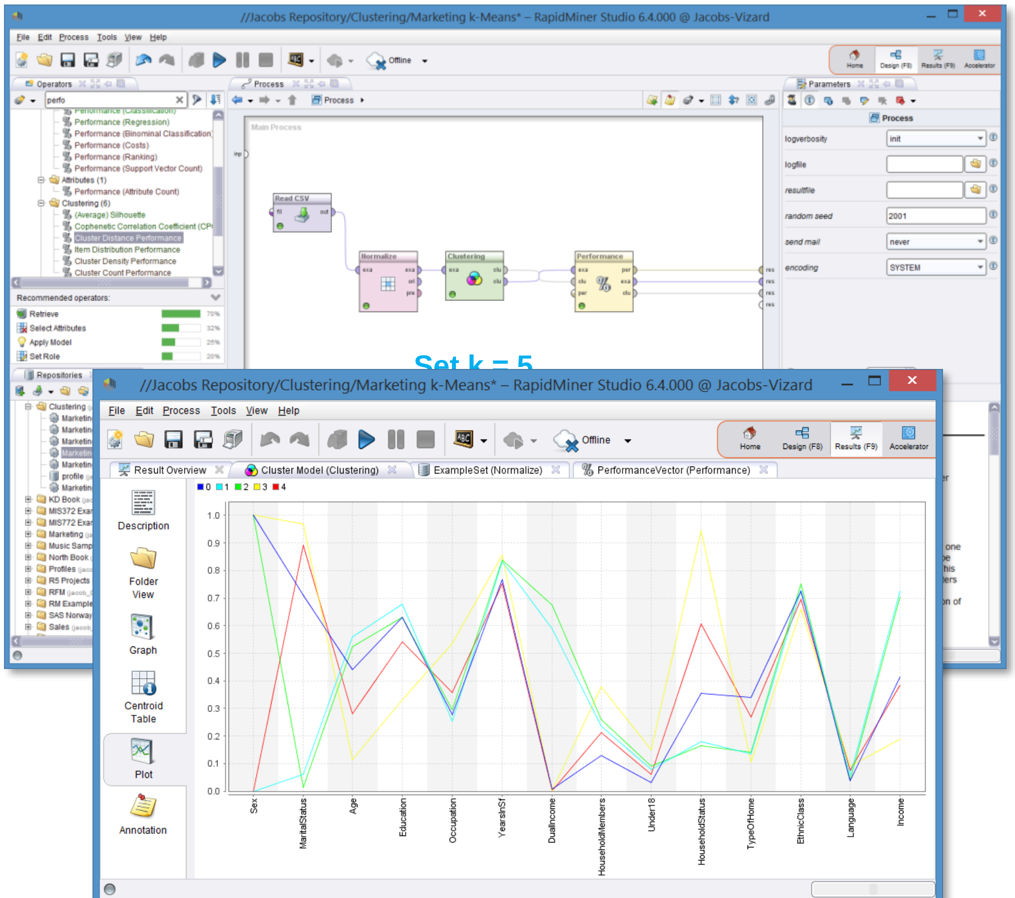
$$for\ n = 6876$$
$$cluster\# = \sqrt{\frac{n}{2}} = 59$$

13. ETHNIC CLASSIFICATION
1. American Indian
2. Asian
3. Black
4. East Indian
5. Hispanic
6. Pacific Islander
7. White
8. Other

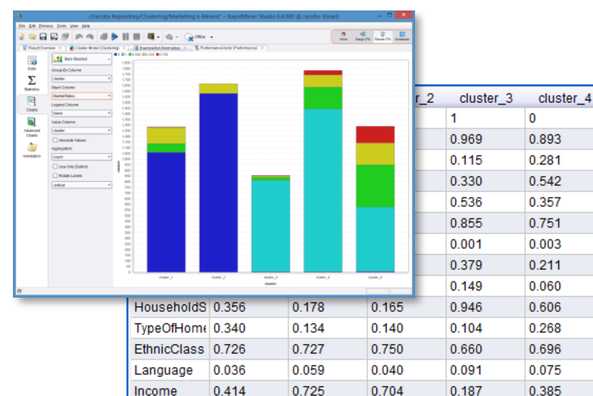
14. LANGUAGE SPOKEN AT HOME?
1. English
2. Spanish
3. Other

Marketing Example: K-Means Simple Process

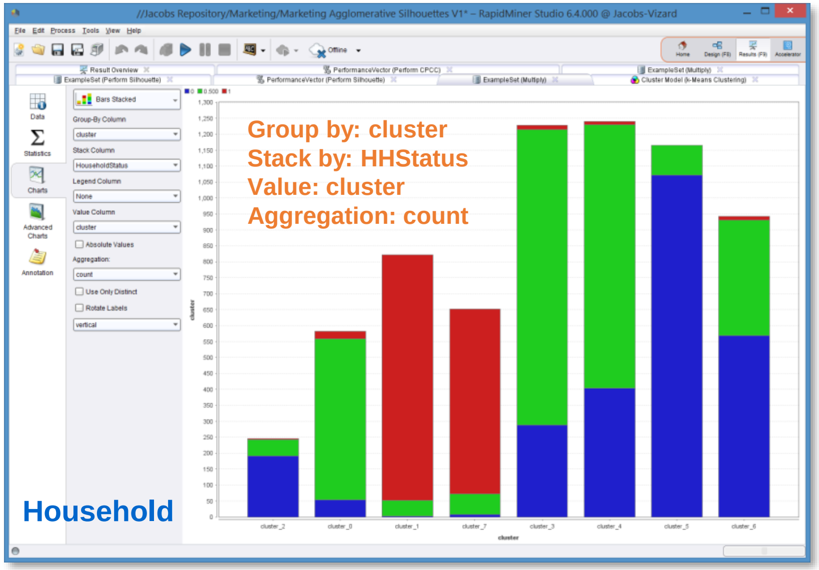
- Select variables for clustering – to the best of your knowledge they must be important in defining clusters / segments
- Reduce dimensionality – high dimensional clusters are hard to find
- Use only numeric variables

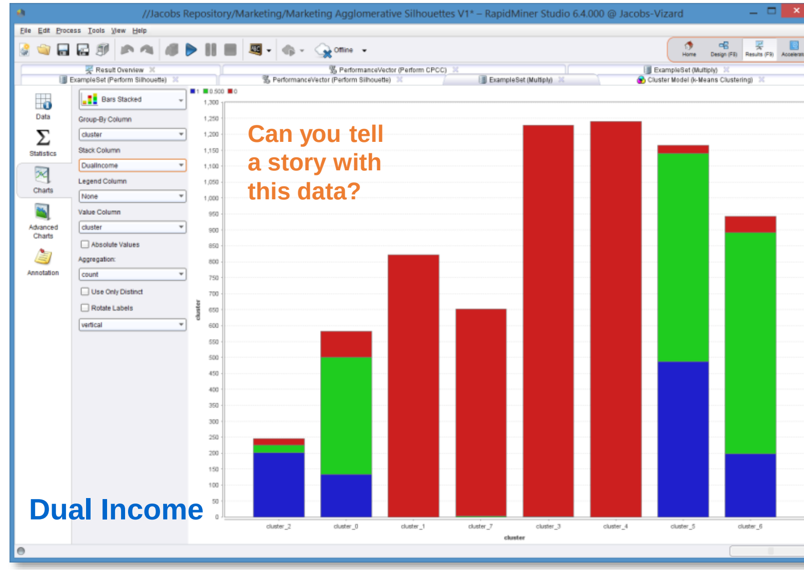


- Every dimension should be of equal importance
- Variables selected for clustering should not be highly related – related attribute increase their weight in clustering
- Optimise clustering to suit its purpose
- Use your domain knowledge in the optimisation process
- Consider different clustering algorithms
- Visualise results for interpretation

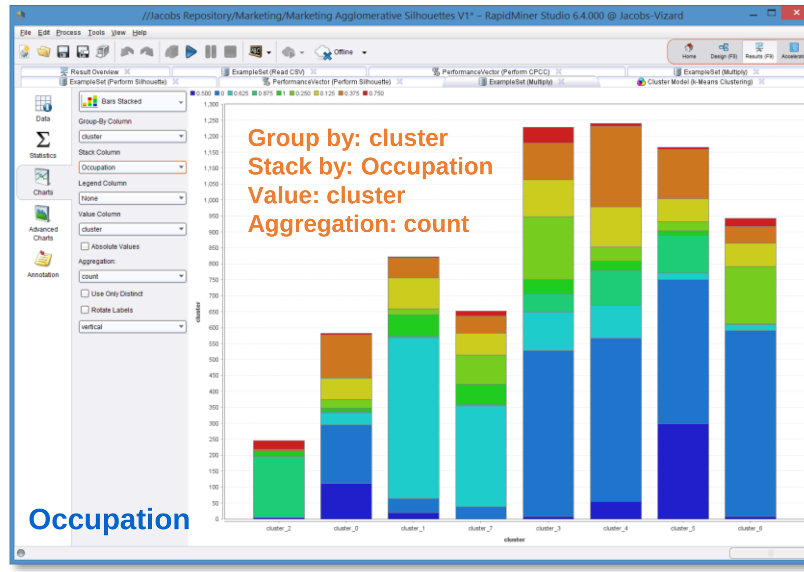


How to Analyse Clusters?
Stacked Bars / Clusters





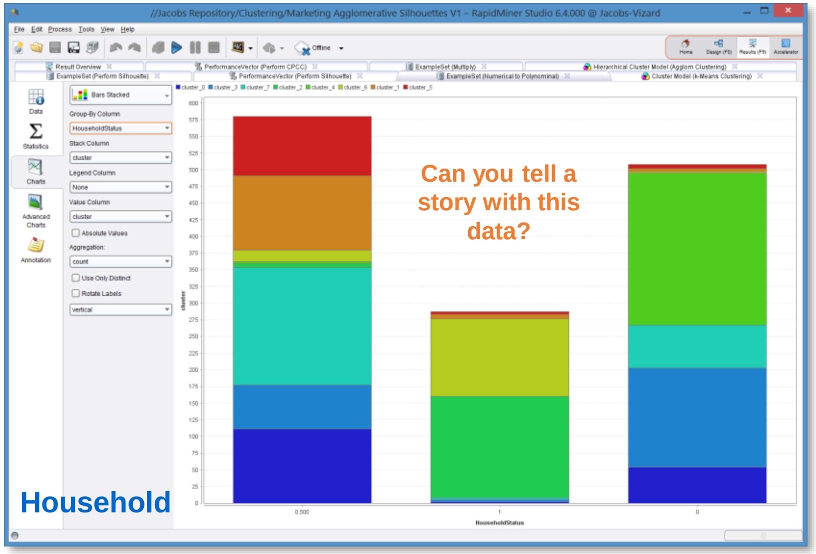
Variables can be numerical but must be binned!

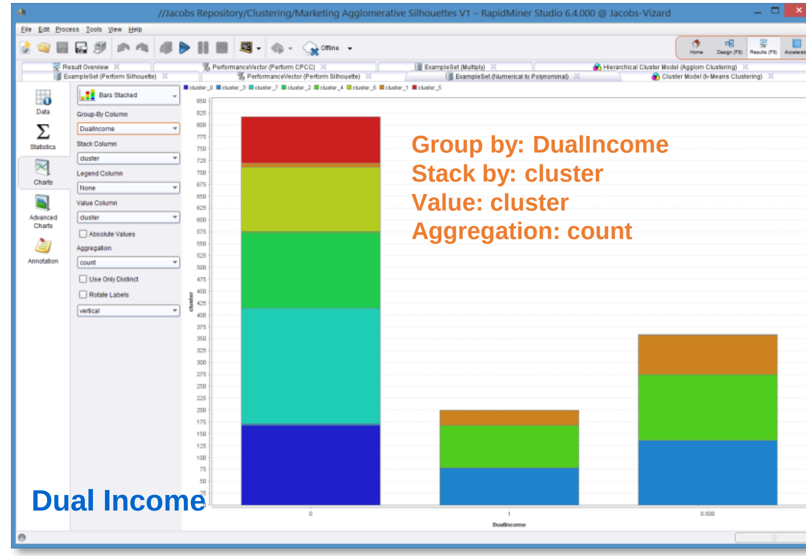


Frequency of attribute values in clusters

- **Householder status: blue (Own), green (Rent), red (With Family)**
  Clusters 1 and 7 - people living with the family
- **Dual income: red (Not Married), green (Yes), blue (No)**
  Clusters 1, 7, 3 and 4 - singles
- **Occupation: grey blue (Professn), yellow (Sales), l. green (Labor), orange (Clerical), d. blue (Home), l. blue (Student), red (Military), bright green (Retired & Unempl)**
  Clusters 1 and 7 – mainly students
- **Conclusion: students are single and live with their family (cluster 1 and 7)**

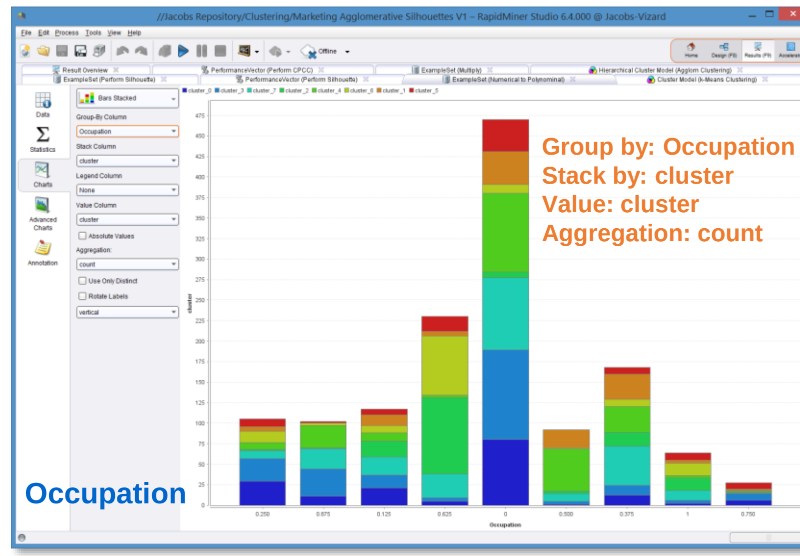
How to Analyse Clusters? Stacked Bars / Variables





- **Householder status:**
  0 (Own), 0.5 (Rent), 1 (With Family)
  Clusters 6 and 4 - people living with the family (1)
- **Dual income:**
  0 (Not Married), 0.5 (Yes), 1 (No)
  Clusters 0, 4, 6 and 5 – singles (0)
- **Occupation: 0 (Professn), 0.125 (Sales), 0.250 (Labor), 0.375 (Clerical), 0.500 (Home), 0.625 (Student), 0.750 (Military), 0.875 (Retired), 1 (Unempl)**
  Clusters 6 and 4 – students (0.625)
- **Conclusion: students are single and live with their family (cluster 4 and 6)**

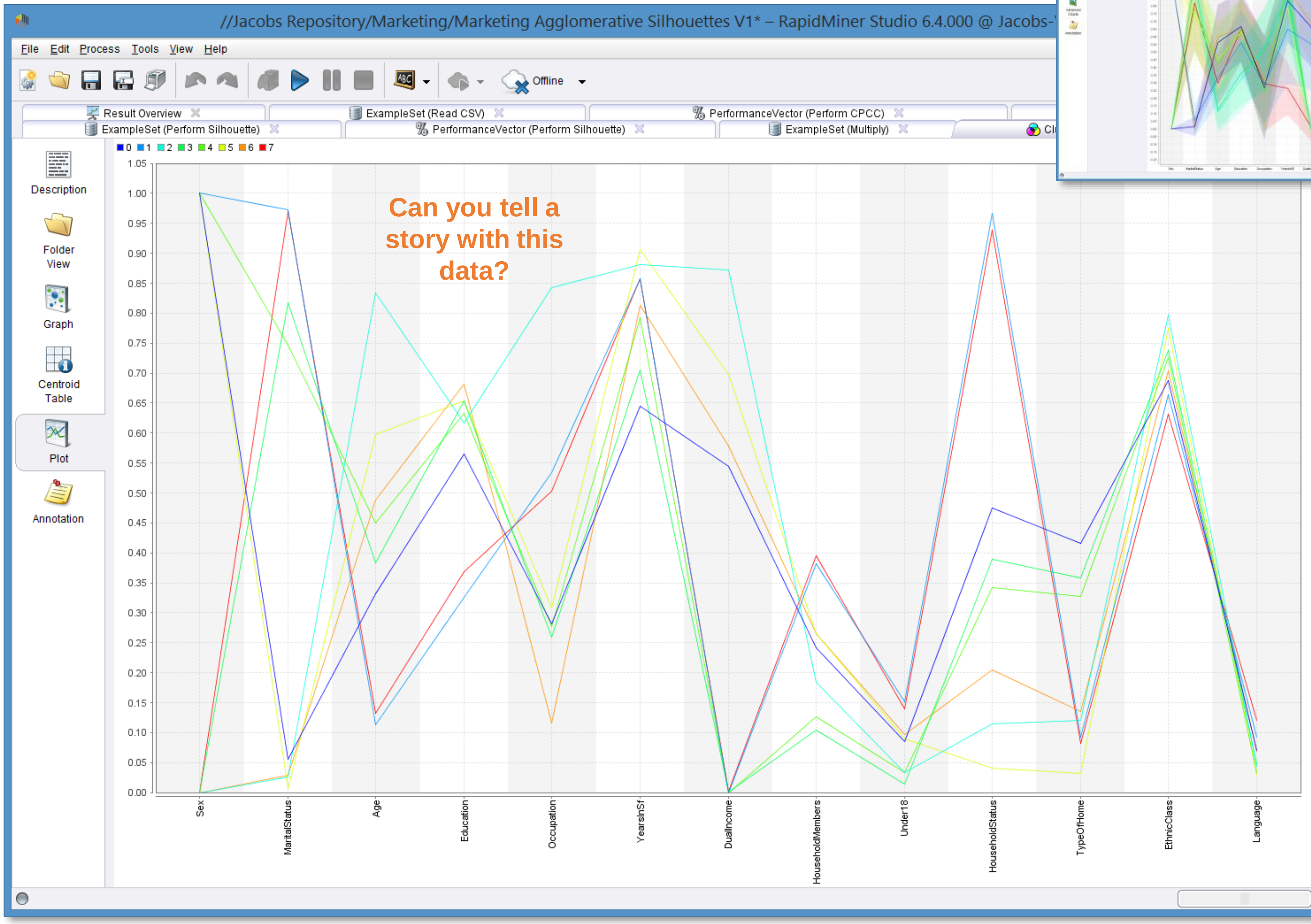
All variables need to be nominal or need to be binned

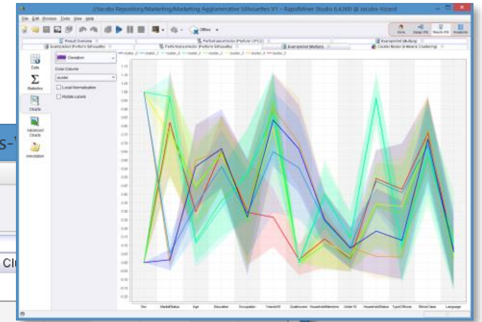


Frequency of clusters in attribute values

How to Analyse Clusters? Parallel Charts

Many cluster models directly provide visualization of their properties, e.g. k-means and its centroids





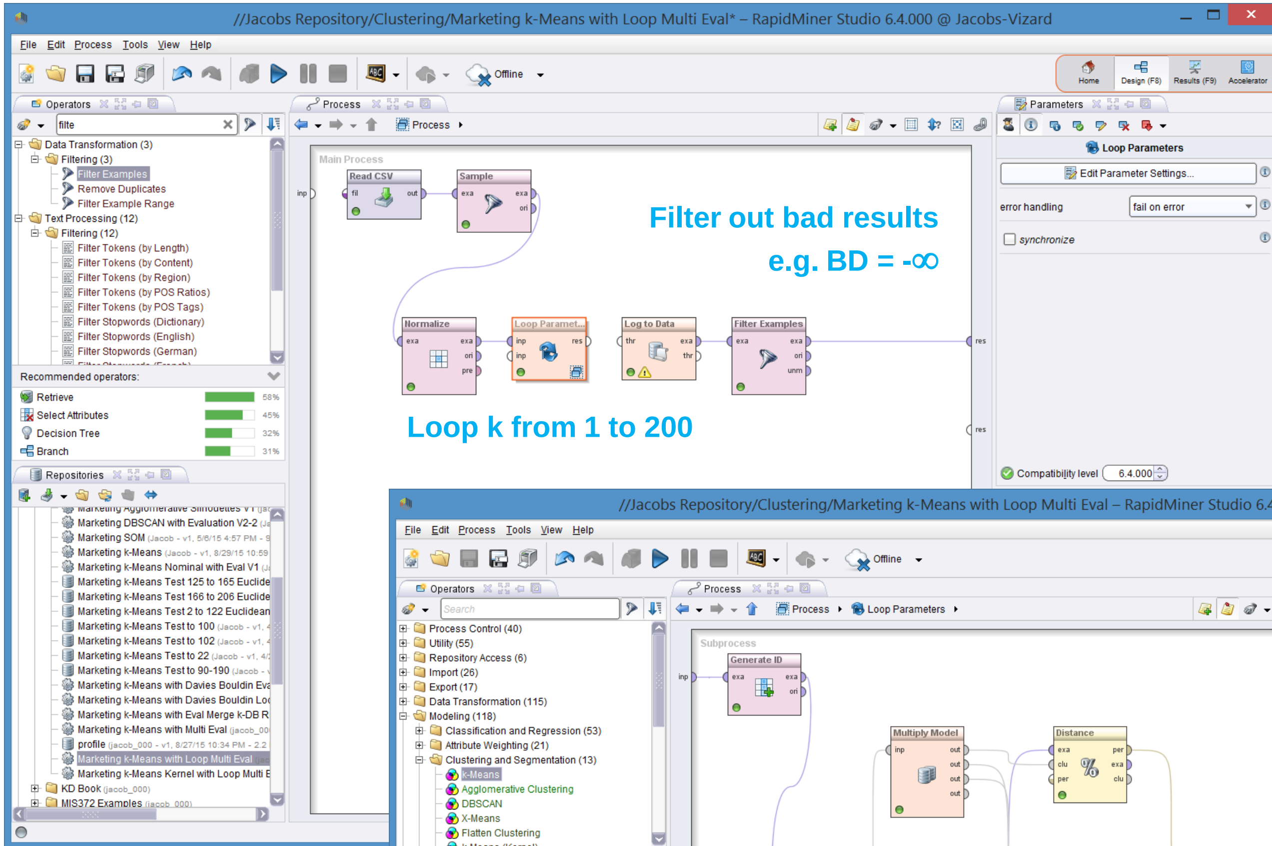
Cluster 7 (red line): young single men, high-school education, living with parents in a house, mainly students. A similar chart can be produced separately to indicate standard deviation in each band of cluster values.

DEAKIN UNIVERSITY

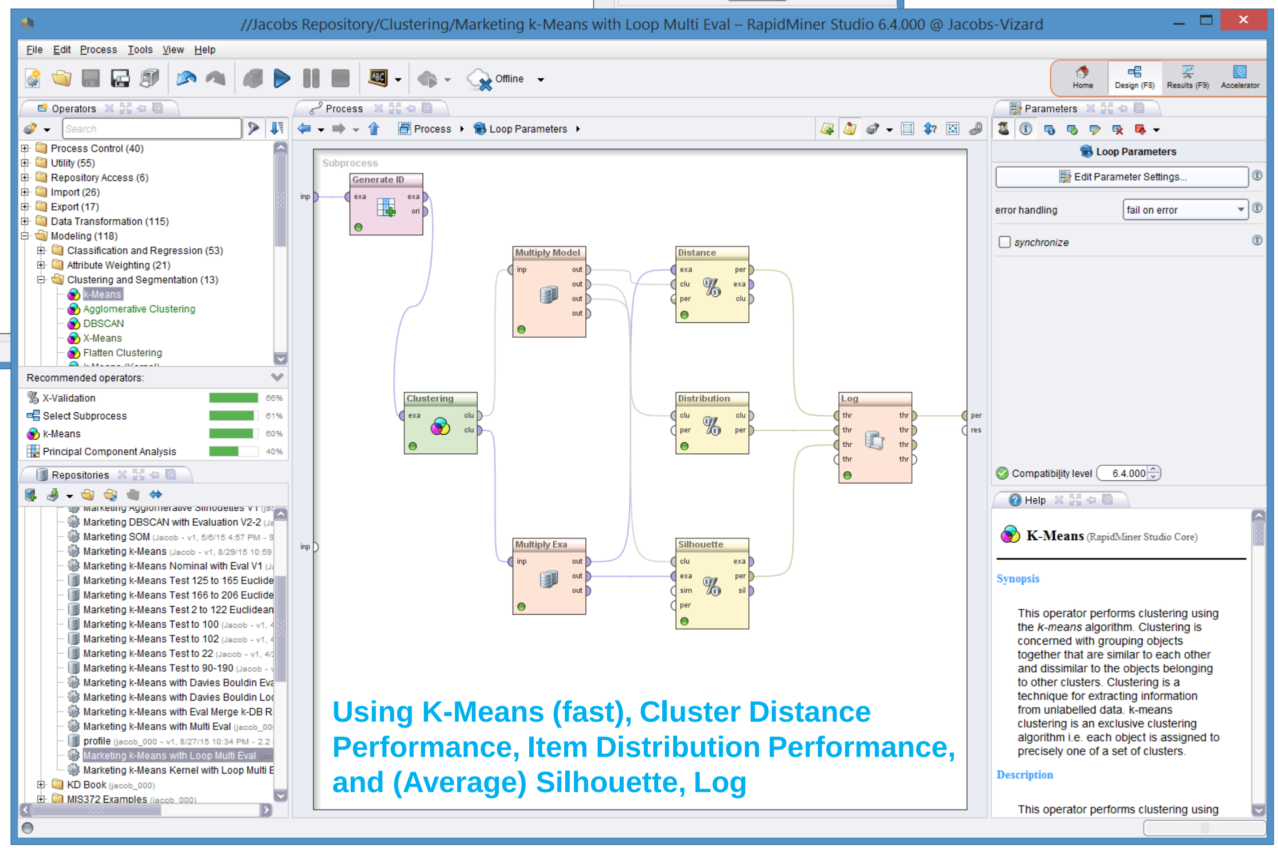
Cluster Optimisation
What should be the value K

- Clusters should consist of data points that have high degree of similarity (small average distance between cluster members and centroid).
- Clusters themselves (or their centroids) should be relatively dissimilar (large average distance between centroids).
- For many applications clusters should have a similar number of members (but not always).
- There should be a minimum unclustered data points.
- There are several approaches to measure the “goodness” of data clustering. RapidMiner provides several performance metrics for flat clusters, e.g.
  - Distance measures
  - Density measures
  - Distribution measures
- Such measures can be taken iteratively while varying a number of model parameters, e.g. k (the number of clusters).
- By plotting the performance measures against clustering parameters, it is possible to detect their best combination, e.g.
  - We can select the best value of k by finding the smallest value of clustering performance metric, e.g. Davies-Bouldin
- Some data mining software, such as R and Python (RapidMiner via a plugin), support calculation of cluster silhouettes, which is based on the ratio between the average dissimilarity of cluster members to each other vs. the from members of other clusters. The measure of dissimilarity can be based on many different metrics.

Marketing Example: In Search of K



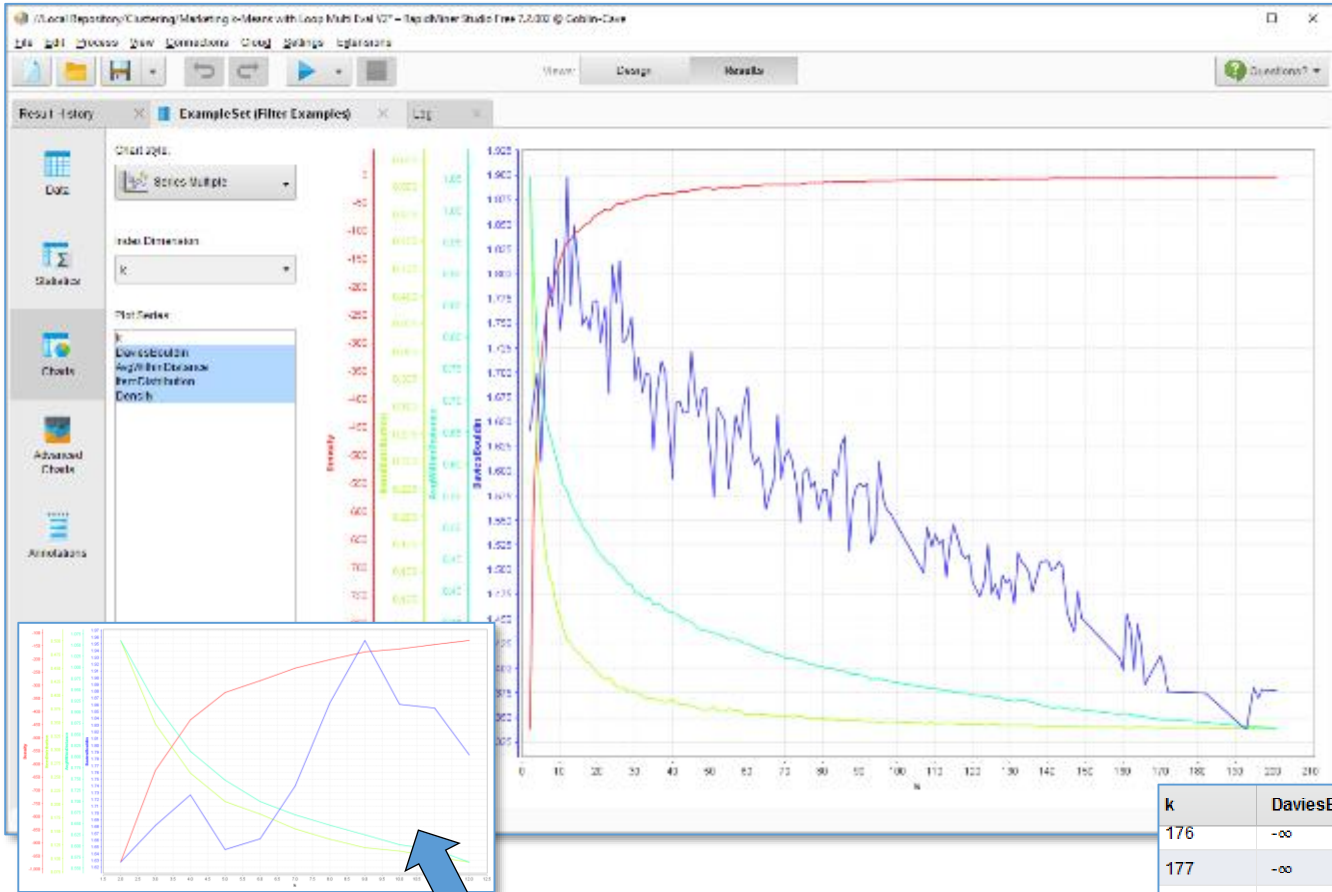
- Sample 10% of your examples
- Create a loop
- Measure and log several types of performance indices



- We can run the process for different values of *k*, e.g. from 2 to 102
- Then investigate the results
- Identify best k - may be difficult

DEAKIN UNIVERSITY

Let us Test Performance Indicators for k = 1..300



Steps of 5 between 5 and 800 showed no significant change of performance: ***DB*** is dropping

Sometimes it is best to use a small selection of variables to get clustering, try using only: Age, Income and Occupation.

k	DaviesB...	AvgWit...	ItemDis...	Density
176	-∞	0.196	0.008	-6.034
177	-∞	0.195	0.008	-6.059
178	-∞	0.195	0.008	-5.954
179	-∞	0.195	0.008	-5.930
180	-∞	0.194	0.008	-5.894
181	-∞	0.193	0.008	-5.861
182	1.374	0.193	0.008	-5.692
183	-∞	0.193	0.008	-5.660
184	-∞	0.193	0.008	-5.758
185	-∞	0.192	0.008	-5.715
186	-∞	0.192	0.008	-5.701
187	-∞	0.191	0.008	-5.778
[illegible]	-∞	0.191	0.008	-5.772

Measure cluster performance based on the selected metrics, such as:

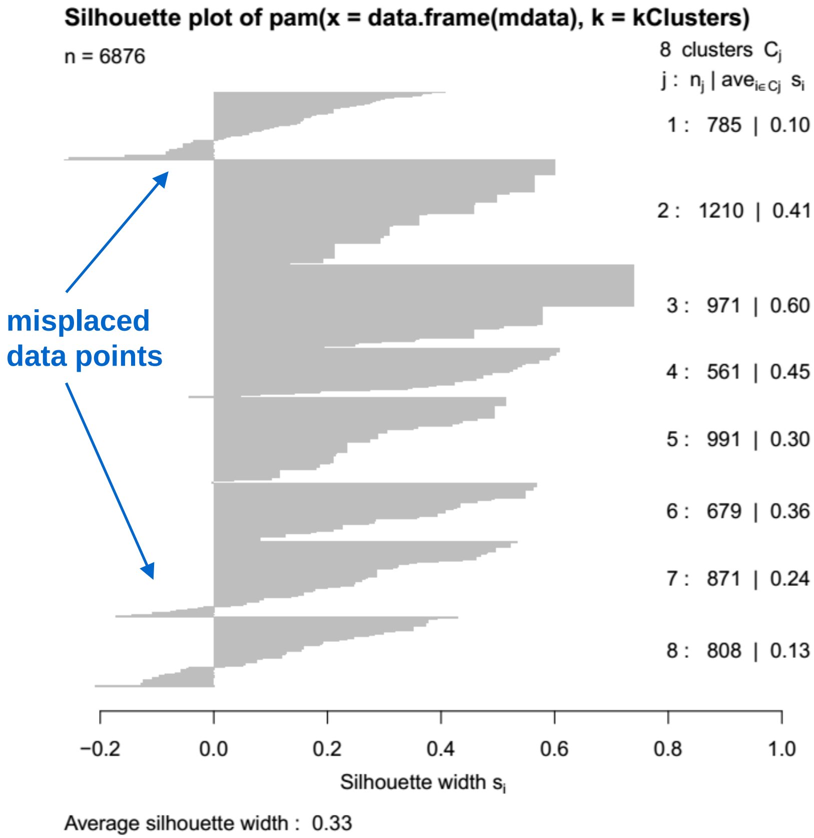
- **Davies-Bouldin**
  blue - find minimum
- **Silhouette**
  May need to try Silhouettes

We hope the best 2 < k < 12 is here but you cannot always rely on the performance indices!

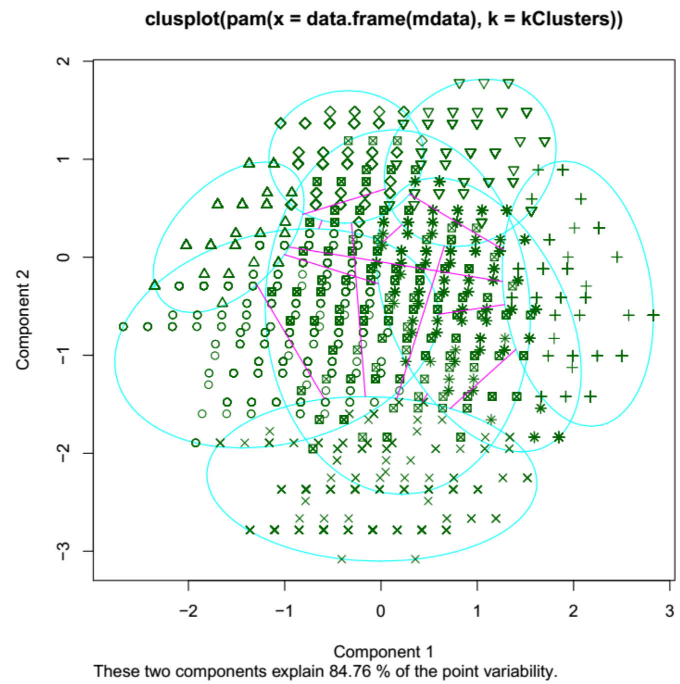
Protect yourself against empty clusters which produce -∞ values in Davies-Bouldin index

DEAKIN UNIVERSITY

Medoid Cluster Analysis Silhouettes



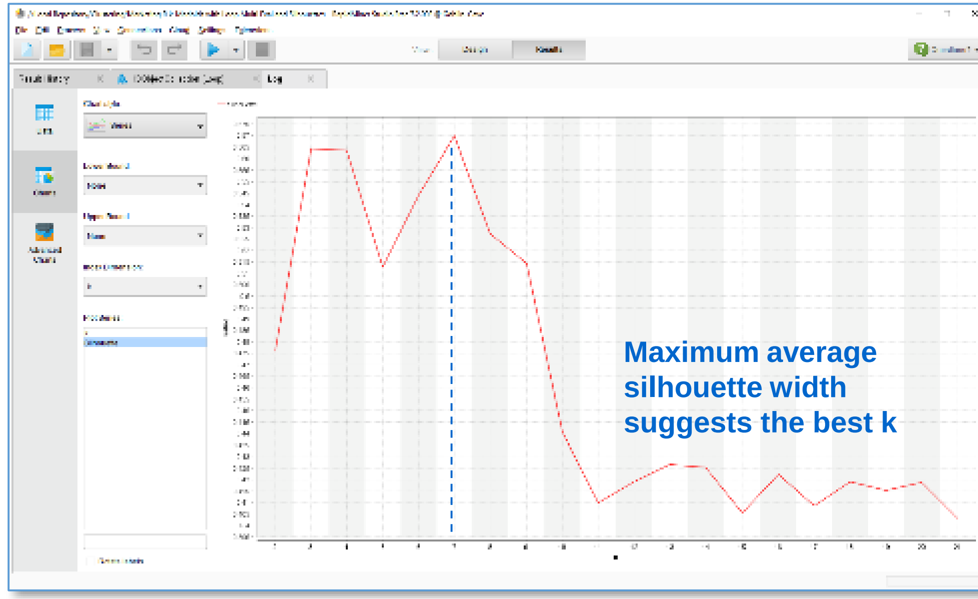
- Flat clusters can also be visualized using silhouettes (left)
- Silhouettes show distribution of dissimilarities between data pairs, i.e. those inside and those outside clusters (widths)
- Silhouette widths are in range -1..1, where the width close to 1 indicates a point near its medoid, -1 indicates it should belong to another cluster
- Average silhouette width is a good indicator of the overall clustering

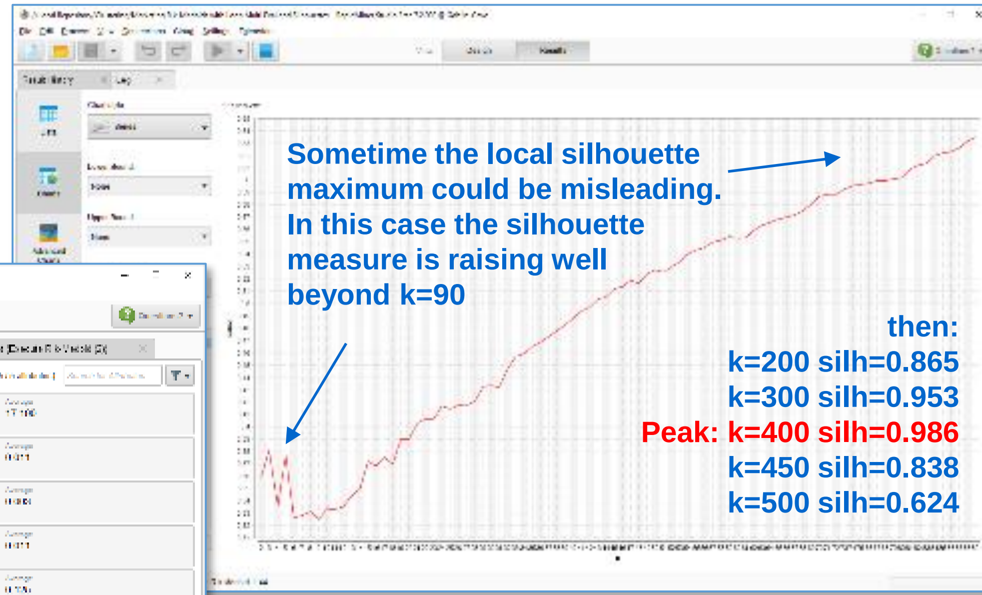


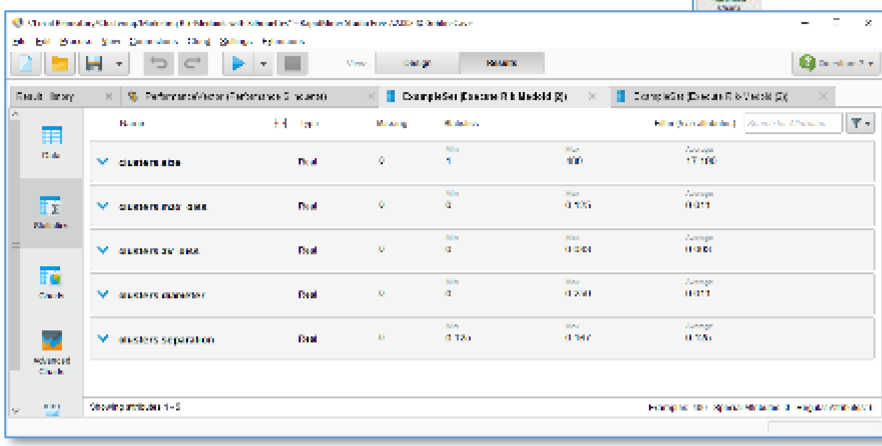
- Clusters can be visualised by plotting their data points in 2D space (right)
- The plot preserves proximity of data points and shows cluster boundaries and their distances
- The method relies on using two principal components of clustered multidimensional data
- RapidMiner can only access silhouettes via R or Python scripting or plugins

Medoid Cluster Analysis Silhouettes

- Silhouette measure identifies best spaced clusters, especially for automatic processing
- Similarly to other optimisation methods silhouette measures could be used to look for the best cluster size
- When we find the maximum average silhouette measure, we adopt its k as optimum
- Beware that local maximum could be misleading, so experiment with a range of reasonable cluster sizes



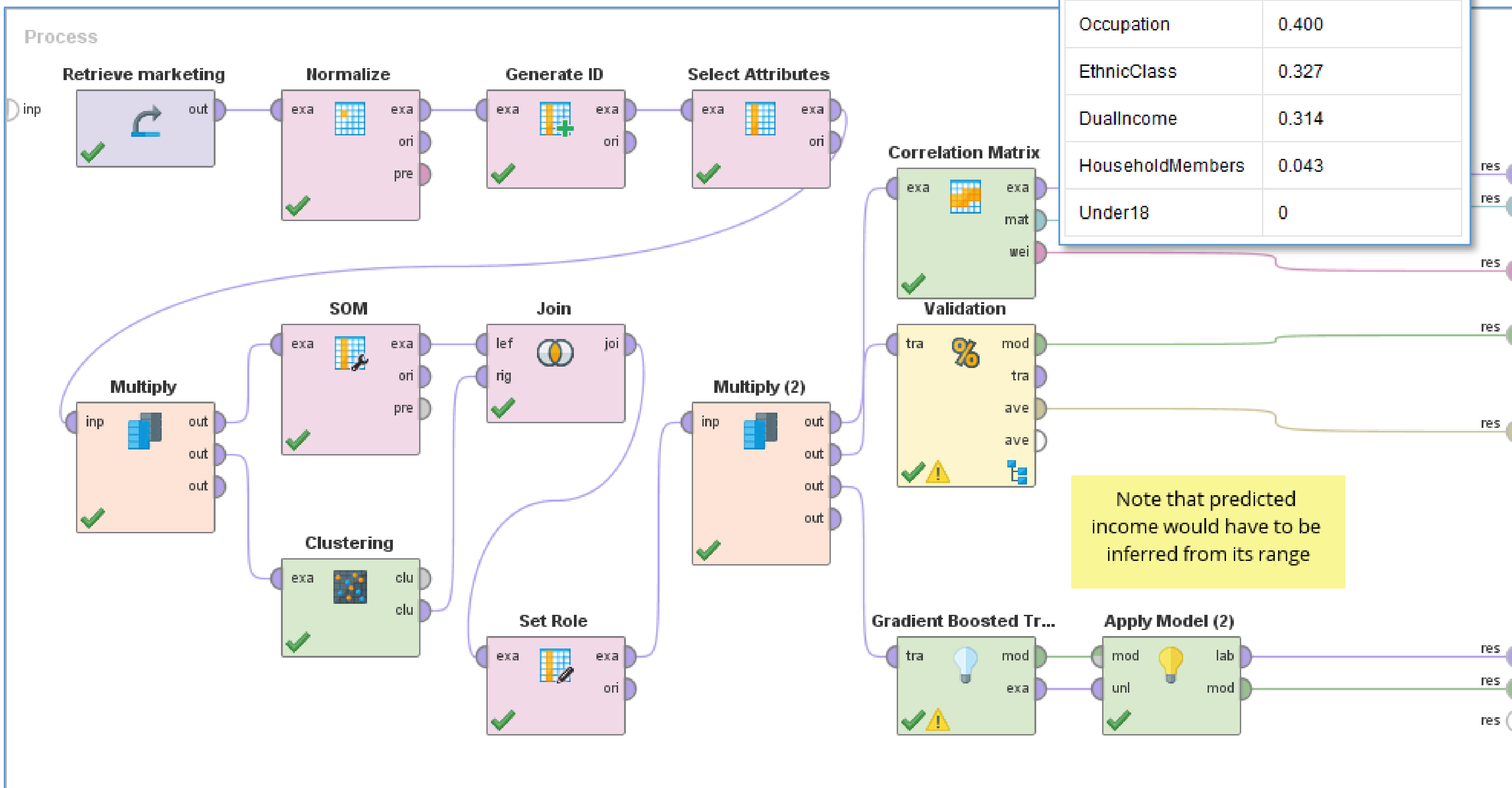




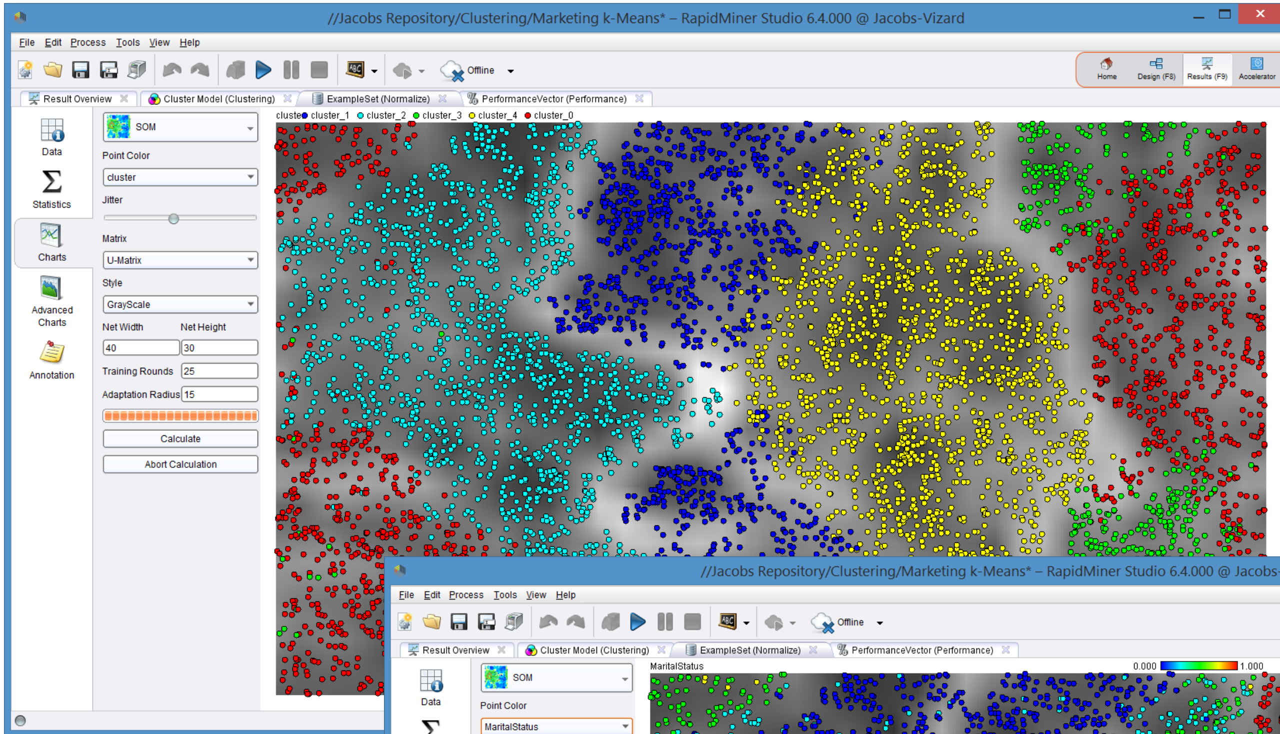
Clustering and Prediction

- There are many uses of clustering:
  - Data exploration
  - Reduction of variables or observations
  - Improvement of prediction
- As an example, we use clustering and SOM to generate extra variables that could be used to improve model prediction.
- Note importance of variables after clustering.

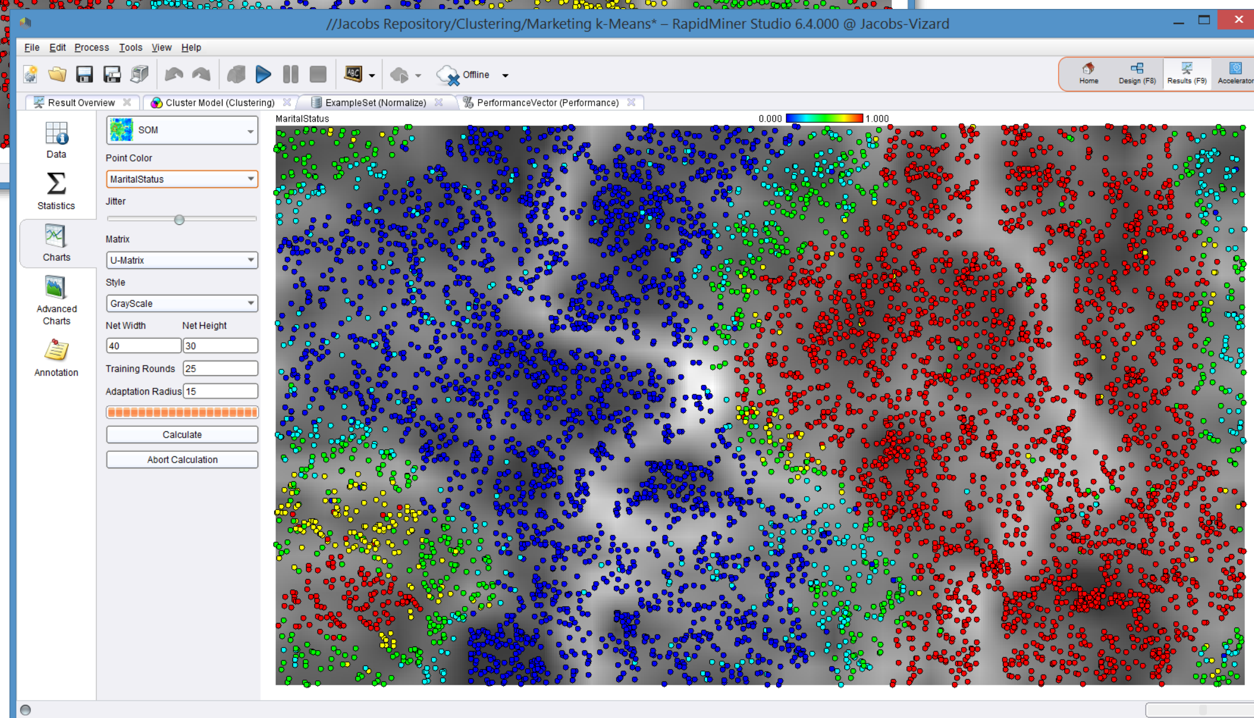
attribute	weight ↓
cluster	1
SOM_0	0.743
TypeOfHome	0.700
YearsInSf	0.586
Education	0.562
SOM_1	0.557
Age	0.466
Language	0.421
Occupation	0.400
EthnicClass	0.327
DualIncome	0.314
HouseholdMembers	0.043
Under18	0



Play with SOM Visualisation



Cluster as point colour



Marital status as point colour

- Select SOM as a chart
- Press Calculate
- Experiment with options
- What can you say?

DEAKIN UNIVERSITY

More "Template" Projects:

R, RM and EMiner

Identify predictors of crime to assist community response planning.

A global agency offering services to local governments across the globe approached you to create a tool capable of predicting crime in communities. As a pilot they provided you with the FBI population and crime data collected in the USA over the five year period. Your job is to select a number of socio-economic predictors of crime and construct a predictive model to be used for the capacity planning by the law enforcement agencies.

You have been asked to identify a number of socio-economic predictors of several types of crime. Establish any interactions between the predictors and targets.

Clean and explore data, use R / R Studio to build the k-NN and Naïve Bayes classifiers, as well as Regression models, evaluate their performance, report the results.

Determine manufacturing problems in vehicles to initiate their recall.

The client is US National Highway Traffic Safety Administration (NHTSA, pronounced "NITS-uh"). They are responsible for reducing deaths, injuries and economic losses resulting from motor vehicle crashes. They require an Early Warning System for potential safety issues associated with automotive vehicles due to manufacturing problems. They require an analytic model to be developed, capable of predicting the likelihood of a vehicle crash, based on the vehicle safety complaints. When the likelihood of crashes is high, NHTSA will initiate a recall of vehicles likely to be affected.

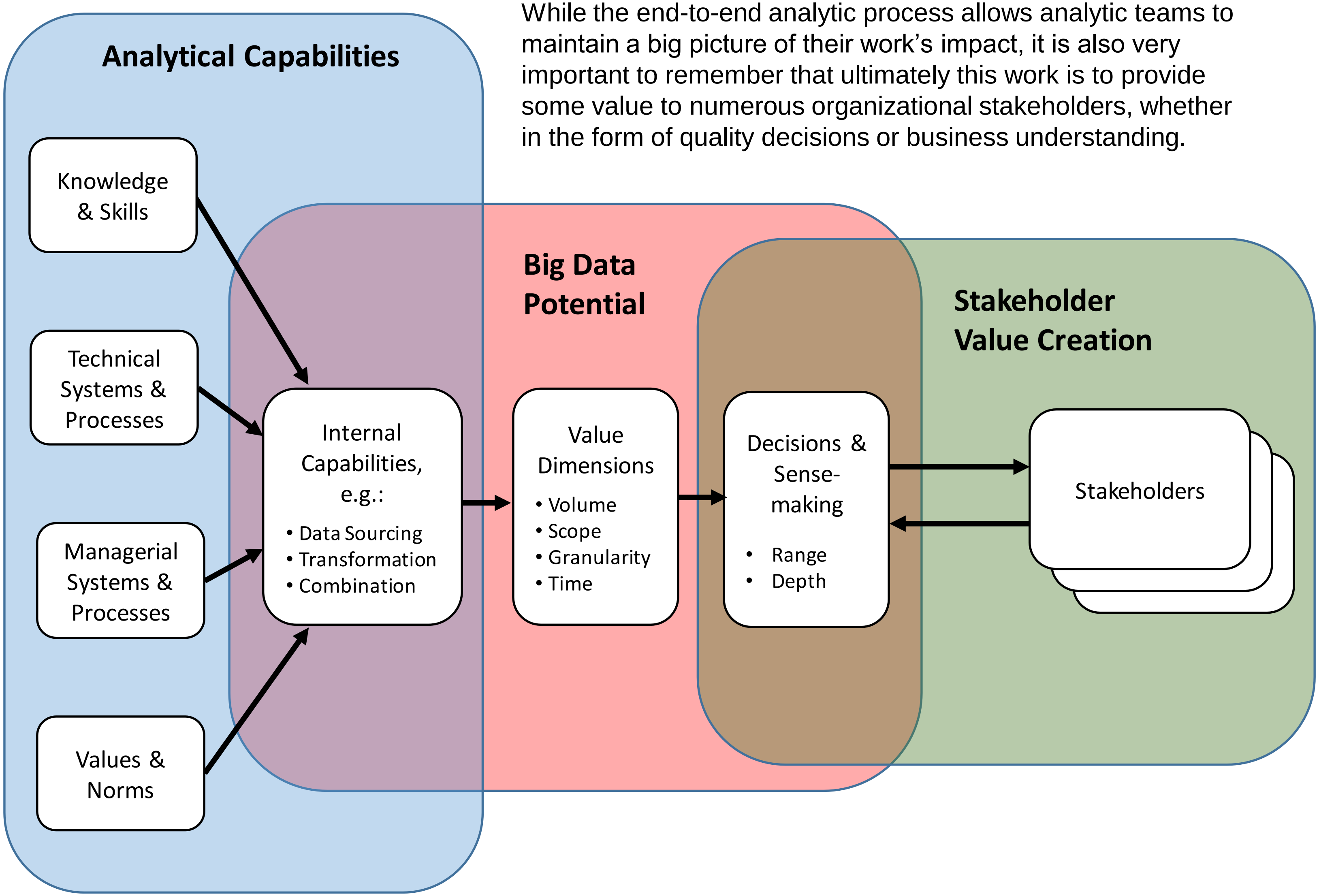
You have been asked to create a number of predictive models using both structured and text data, evaluate and compare their performance with SAS Enterprise Miner.

Select the best predictive model and use it to suggest what vehicles should be recalled from the roads.

DEAKIN UNIVERSITY

Value of Information and Data Analytics

While the end-to-end analytic process allows analytic teams to maintain a big picture of their work's impact, it is also very important to remember that ultimately this work is to provide some value to numerous organizational stakeholders, whether in the form of quality decisions or business understanding.



(Adapted from Rens Scheepers 2016)

Conclusions

- Sensemaking is a prerequisite to decision making
- The key to data analytics is data modelling
- Some of the models are predictive and some explanatory
- Data visualisation provides intuition but supports analytics
- Analytic process assures reusability of models
- There are many analytic tools, in a wide range of features and prices, some provide very high productivity
- While R and Python are the most popular analytic tools their productivity value is relatively low
- The key to high analytic productivity is the process support
- Never exclude text from the analytic process
- All models need to be optimised
- Many measurements used in model optimisation have “preconditions”, which need to be checked
- Tools such as SAS EMiner and RapidMiner provide extensions to enrich their feature set (e.g. R and Python)
- Never lose sight of business value in data analytics!

Some R(L)ight Reading