KPA
Insights through analytics

TECHNION
Israel Institute of Technology

A special topics lecture series on analytics

Professor Ron S. Kenett
ron@kpa-group.com

Samuel Neaman Institute for National Policy Research

Students in Political Sciences (SPO) need to sign up

Professor Ron S. Kenett
ron@kpa-group.com

- Lecture Series in Analytics (Sala Laura)

22/01	10.30-13.30
23/01	9.30-12.30
24/01	10.30-13.30

- Lecture Series in Causality (Sala Laura)

28/01	9.30-12.30
29/01	10.30-13.30

- Seminar on ‘Statistics at a Crossroad’

Via Santa Sofia, 9 - aula M203

30/01 **10.45-11.45**

Task 1: Information quality assessment of a case study (1/3)

Task 2: Trump tweets text analysis

Task 3: German credit data analysis

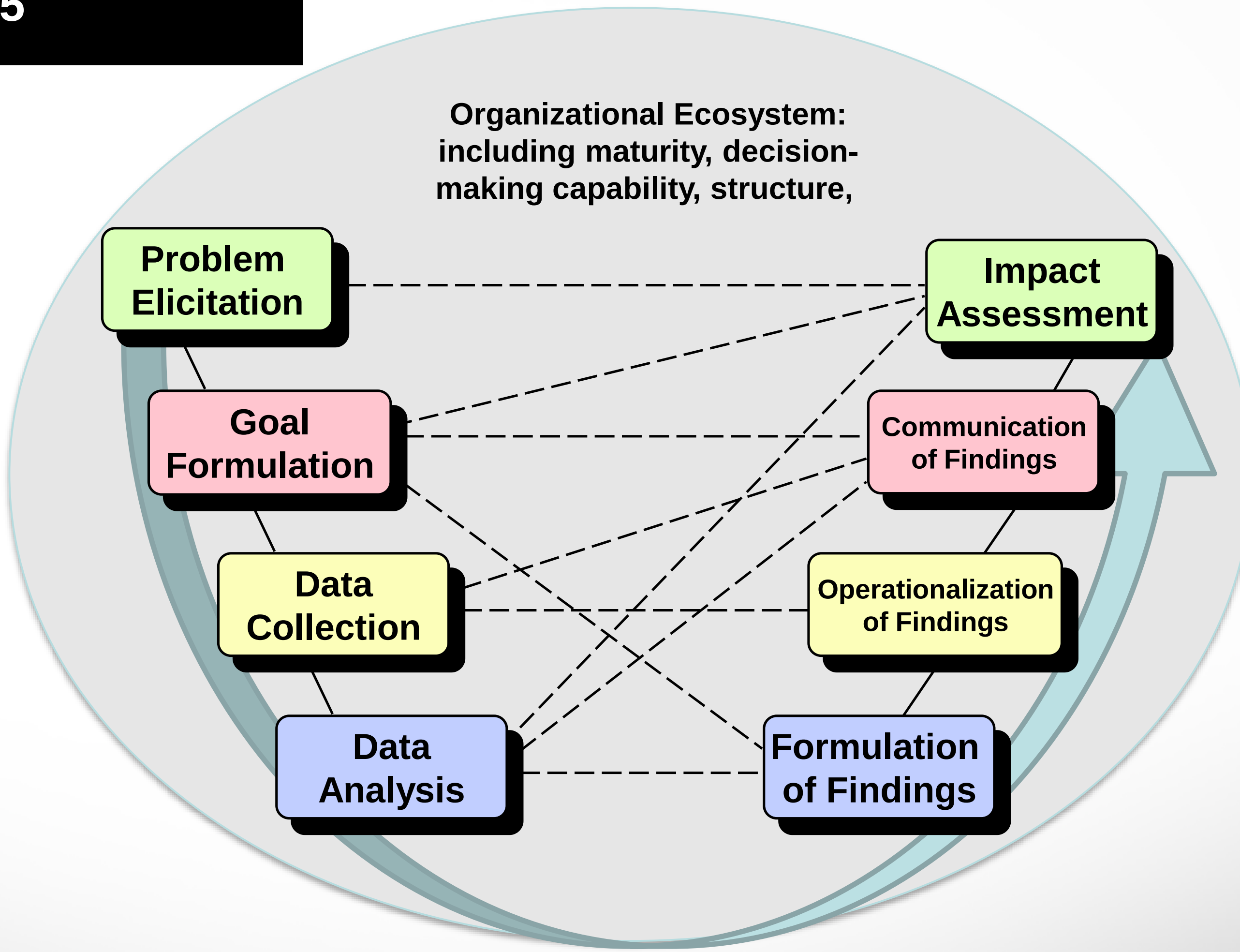
Deadline

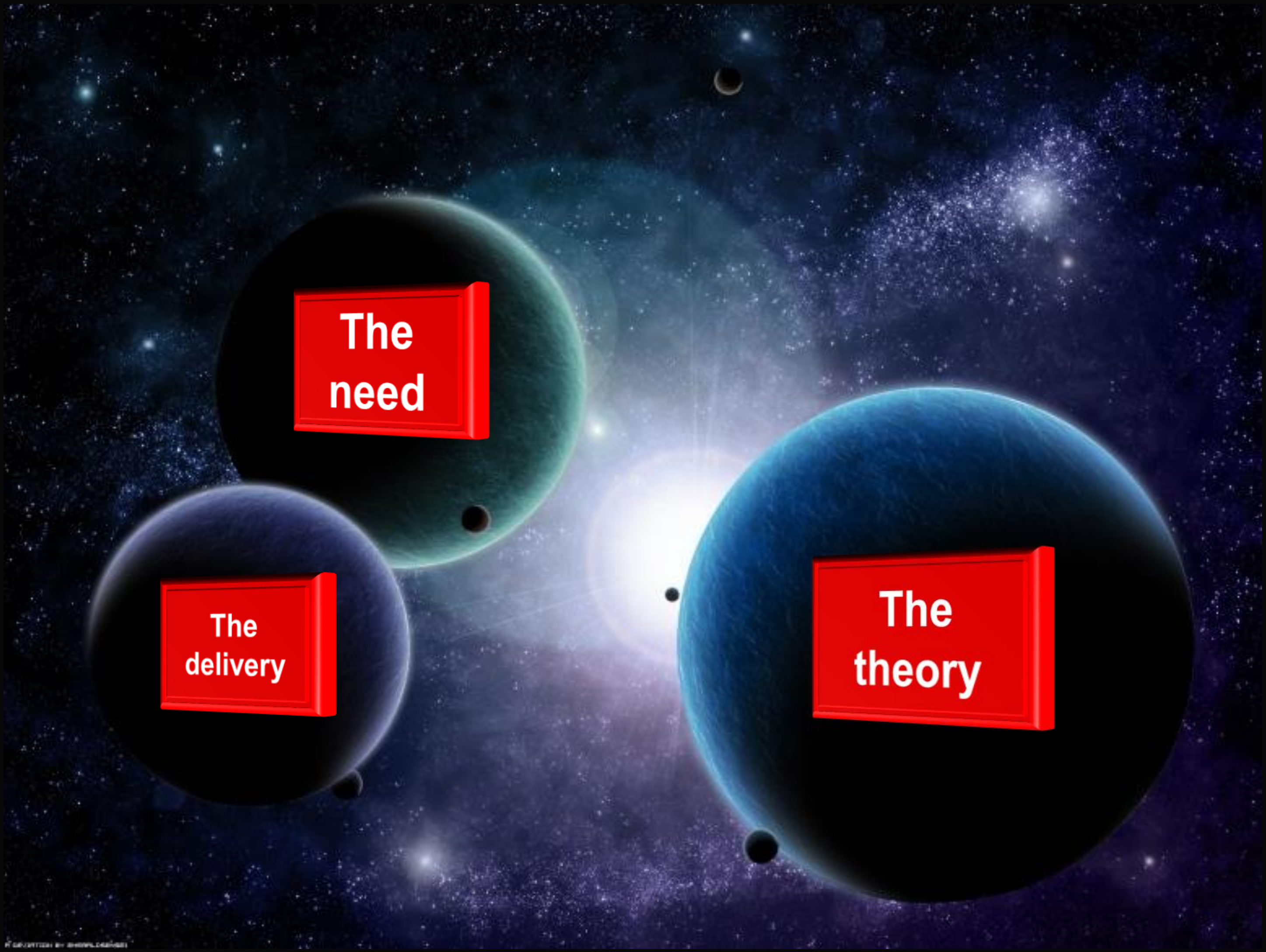
1/3/2020

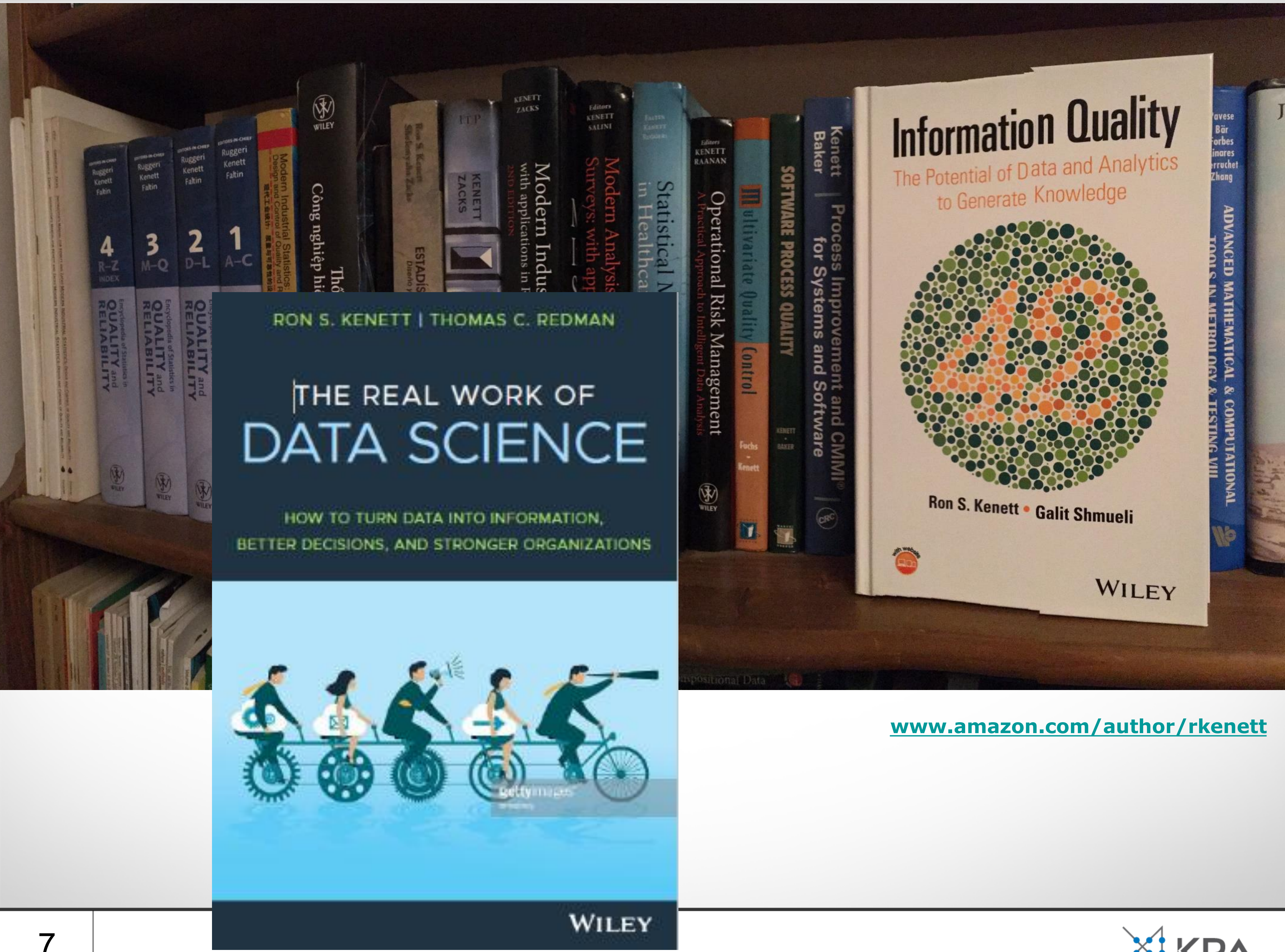
- Background
- Information quality and student group tasks
- The real work of data science
- Decision trees
- Regression trees
- Random forests
- The non performing loans (NPL) case study
- Logistic regression
- Naïve Bayes
- K-means clustering
- Text analytics
- Causality
- Statistics at a crossroad seminar

Applied statistics
is about meeting the challenge
of solving real world problems
with mathematical tools
and statistical thinking

A life cycle view of statistics *Quality Engineering* (with discussion), Vol. 27, No.1, pp. 111-129, 2015




The need
The delivery
The theory



www.amazon.com/author/rkenett

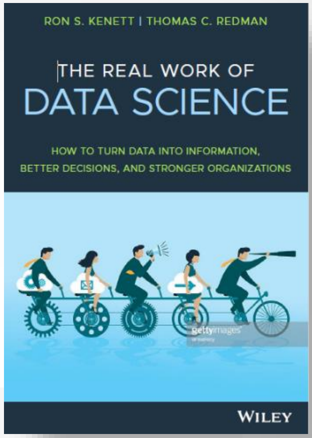
Level 5: Learning and discovery - This is where attention is paid to information quality. Data from different sources is integrated. Chronology of Data and Goal and Generalization is a serious consideration in designing analytic platforms. Leverage causality models.

Level 4: Quality by Design - Experimental thinking is introduced. The data scientist suggests experiments, like A/B testing, to help determine which website is better. Develop causality analysis.

Level 3: Process focus - Probability distributions are part of the game. The idea that changes are statistically significant, or not, is introduced. Some attention is given to model fitting. Introduce causality analysis.

Level 2: Descriptive statistics level – Management asks to see histograms, bar charts and averages. Models are not used, data is analyzed in rather basic ways.

Level 1: Random demand for reports driven by firefighting - New reports address questions such as: How many components of type X did we replace last month or how many people in region Y applied for a loan?



The analytics maturity ladder

Quality Assurance in the Golden

https://www.jmp.com/en_us/events/ondemand/analytically-speaking/quality-assurance-in-the...

Email link RonLinks LinkedIn Facebook News Ynet בנק הפועלים – לקוח Sci-Hub: removing b

jmp Software JMP in Action Events Learn JMP Community Support About Us

ANALYTICALLY SPEAKING

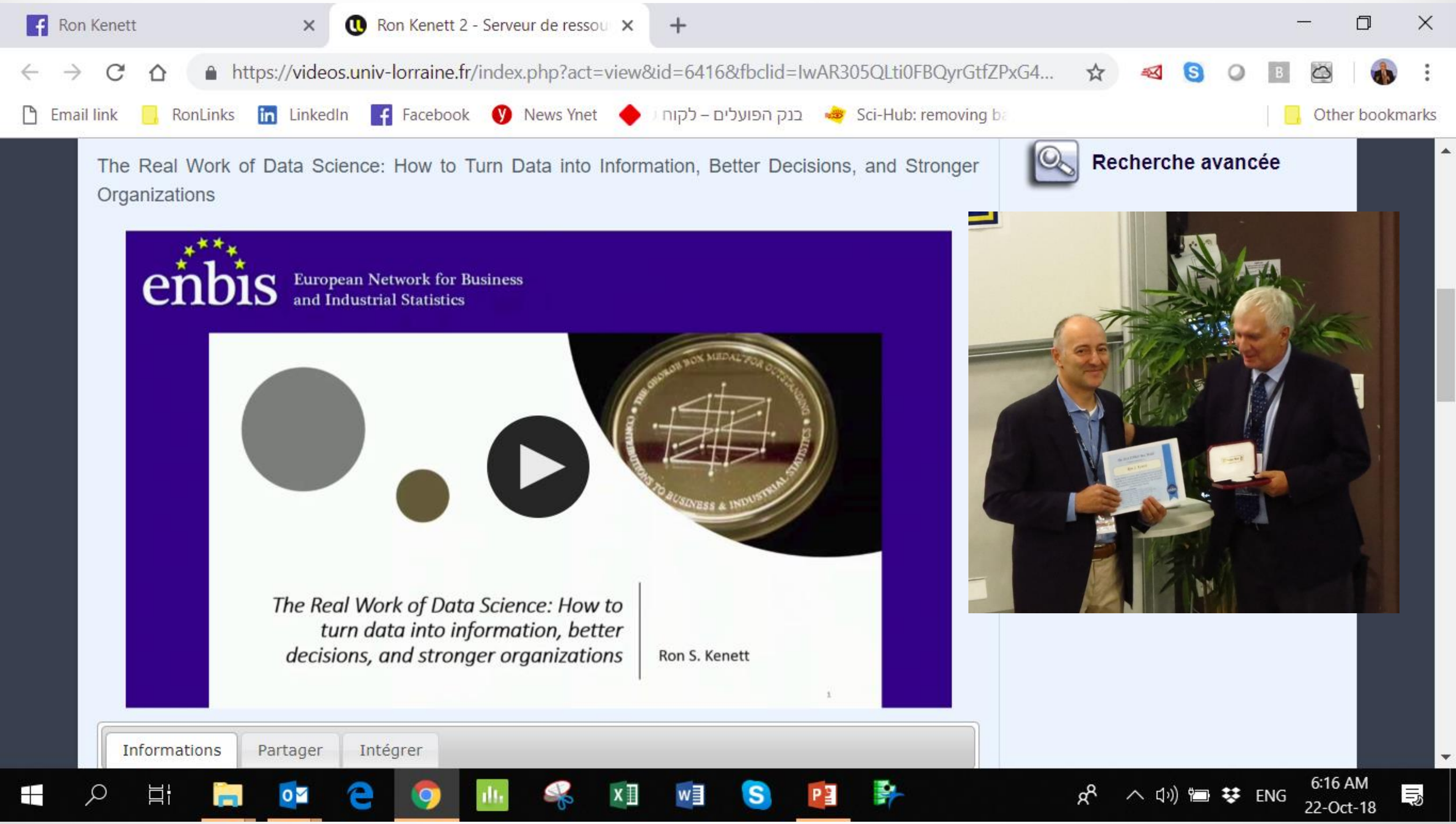
Quality Assurance in the Golden Age of Analyti

With Ron Kenett

With the advent of trans
Industry 4.0, it's clear th
beyond a traditional vie
manufacturing. So how,
in an industrial setting?
says that in the golden a
become the arbiters of
testing architecture, he
engineering challenge a
stakeholders apply the p
quality control and qual
challenges of increased

In addition to discussing
insurance, Kenett also e
his book (co-authored w
of Data Science: How to
Better Decisions and St
be released in 2019.

https://www.jmp.com/en_us/events/ondemand/analytically-speaking/quality-assurance-in-the-golden-age-of-analytics.html



https://www.youtube.com/watch?v=gHoeeuuwcPs&list=PLMCuIG3AKGww8SgP0JQGOXqxu2bFThhIS&index=2&t=235s

WILEY

Contact Us Tech Support

Kenett, Shmueli:
Information Quality: The Potential of Data and Analytics to Generate Knowledge

Home | Resources | More Information

Presentations on InfoQ

requires Adobe Acrobat Reader

Title	Location	Date
Do you want to make an impact with quantitative methods? Make sure you generate high InfoQ	Toulon-Verona Conference, Israel	September 3, 2012
A Workshop on Modern Analysis of Customer Satisfaction Surveys	22nd Colombian Statistics Symposium, The National University of Colombia Bucaramanga, Colombia	July 17, 2012
Quantitative and Qualitative Aspects of Bayesian Networks: A General Approach for Integrating Expert Opinions and Structured Data	Séminaire Parisien de Statistique, Institut Henri Poincare, Paris	April 7, 2014
ENBIS Management Day Round Table Discussion	ENBIS 2011, Coimbra, Portugal	September 7, 2011

Get Help With:

Adobe PDF and Acrobat Reader

* These links will open a new window

https://www.wiley.com//legacy/wileychi/kenett/presentation.html?type=SupplementaryMaterial

KPA


statistics
آمار
статыстыка
statystyka
statisztika
istatistik
statistik
статистика
统计
statistica
परिसंख्यान
tilasto
statistik
istatistika
số liệu thống kê
takwimu
estatística
statistiko
統計
statistikat
tölfræði
סטטיסטיקה
statistic
statistika


numbers
data
statistical analysis
findings
information
insights
Information Quality


statistics
статыстыка
statystyka
statistica
statisztika
istatistik
statistik
статистика
tilasto
số liệu thống kê
takwimu
estatística
statistiko
statistikat
tölfræði
statistic
statistika
統計
统计


Danger Zone
Information Quality

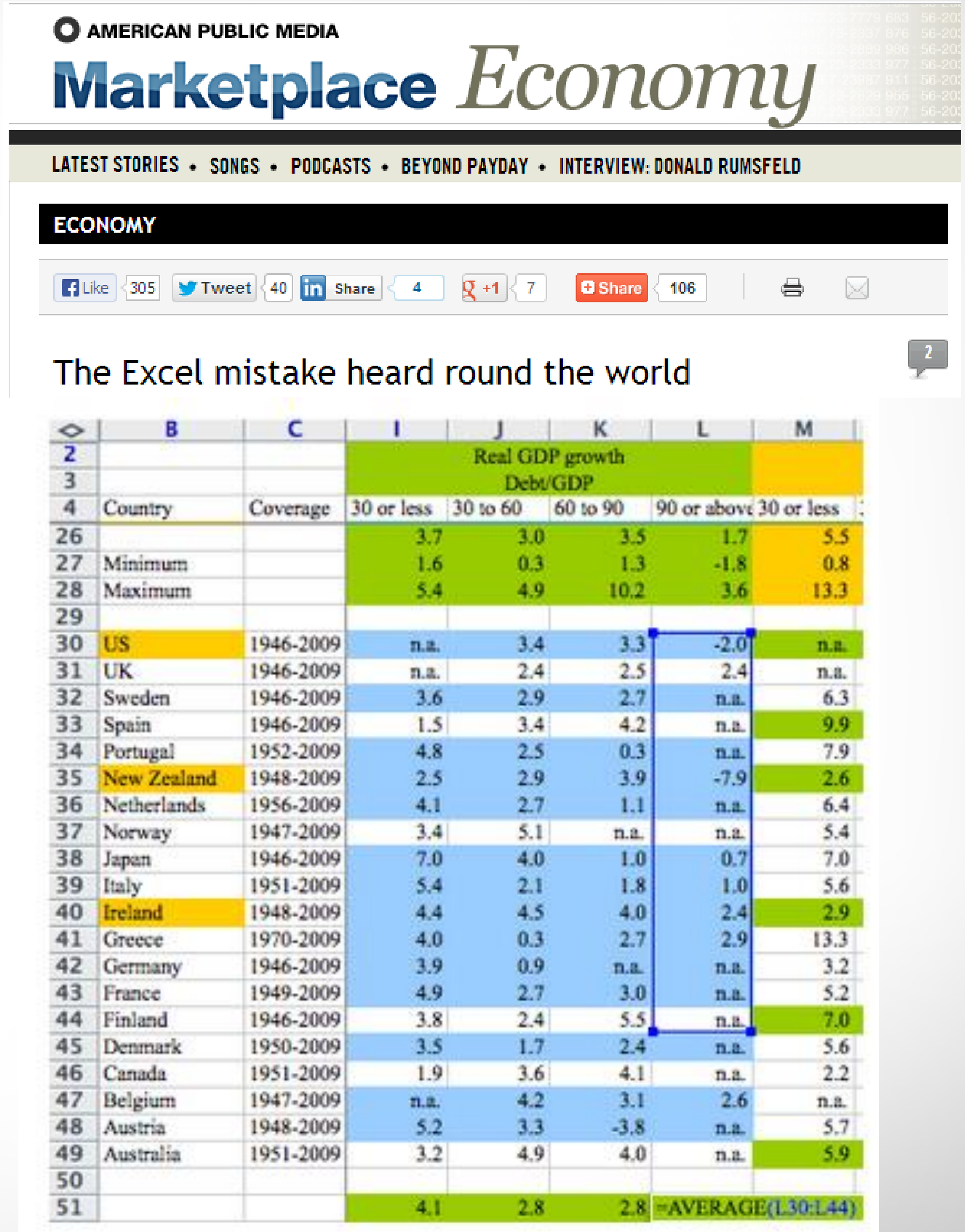
www.marketplace.org/topics/economy/excel-mistake-heard-round-world

"In the last three years, there has been a concerted effort by those in Washington to reduce government spending and reign in the national debt.

One reason for the budget cuts?

Research by two Harvard economists, Ken **Rogoff** and Carmen **Reinhart**. The pair found that when a country owes more than 90 percent of their GDP, it slides into recession."

... Fixing this Excel error transforms high-debt countries from recession to growth


AMERICAN PUBLIC MEDIA

Marketplace Economy

LATEST STORIES • SONGS • PODCASTS • BEYOND PAYDAY • INTERVIEW: DONALD RUMSFELD

ECONOMY

Like 305 Tweet 40 Share 4 +1 7 Share 106

The Excel mistake heard round the world

	B	C	I	J	K	L	M
2			Real GDP growth				
3			Debt/GDP				
4	Country	Coverage	30 or less	30 to 60	60 to 90	90 or above	30 or less
26			3.7	3.0	3.5	1.7	5.5
27	Minimum		1.6	0.3	1.3	-1.8	0.8
28	Maximum		5.4	4.9	10.2	3.6	13.3
29							
30	US	1946-2009	n.a.	3.4	3.3	-2.0	n.a.
31	UK	1946-2009	n.a.	2.4	2.5	2.4	n.a.
32	Sweden	1946-2009	3.6	2.9	2.7	n.a.	6.3
33	Spain	1946-2009	1.5	3.4	4.2	n.a.	9.9
34	Portugal	1952-2009	4.8	2.5	0.3	n.a.	7.9
35	New Zealand	1948-2009	2.5	2.9	3.9	-7.9	2.6
36	Netherlands	1956-2009	4.1	2.7	1.1	n.a.	6.4
37	Norway	1947-2009	3.4	5.1	n.a.	n.a.	5.4
38	Japan	1946-2009	7.0	4.0	1.0	0.7	7.0
39	Italy	1951-2009	5.4	2.1	1.8	1.0	5.6
40	Ireland	1948-2009	4.4	4.5	4.0	2.4	2.9
41	Greece	1970-2009	4.0	0.3	2.7	2.9	13.3
42	Germany	1946-2009	3.9	0.9	n.a.	n.a.	3.2
43	France	1949-2009	4.9	2.7	3.0	n.a.	5.2
44	Finland	1946-2009	3.8	2.4	5.5	n.a.	7.0
45	Denmark	1950-2009	3.5	1.7	2.4	n.a.	5.6
46	Canada	1951-2009	1.9	3.6	4.1	n.a.	2.2
47	Belgium	1947-2009	n.a.	4.2	3.1	2.6	n.a.
48	Austria	1948-2009	5.2	3.3	-3.8	n.a.	5.7
49	Australia	1951-2009	3.2	4.9	4.0	n.a.	5.9
50							
51			4.1	2.8	2.8	=AVERAGE(L30:L44)	

https://www.sciencemag.org/news/2016/08/one-five-genetics-papers-contains-errors-thanks-microsoft-excel

One in five genetics papers contains errors thanks to Microsoft Excel

By Jessica Boddy | Aug. 29, 2016 , 1:45 PM

Autoformatting in Microsoft Excel has caused many a headache—but now, a new study shows that one in five genetics papers in top scientific journals **contains errors from the program**, *The Washington Post* reports. The errors often arose when gene names in a spreadsheet **were automatically changed** to calendar dates or numerical values. For example, one gene called *Septin-2* is commonly shortened to *SEPT2*, but is changed to 2-SEP and stored as the date 2 September 2016 by Excel. The researchers, who published their analysis in *Genome Biology*, say the issue can be fixed by formatting Excel columns as text and remaining vigilant—or switching to Google Sheets, where gene names are stored exactly as they're entered.

Spreadsheets are OK for data entry. But not for calculations.

- Conflates input, code, output, presentation
- UI invites errors, then obscures them
- Debugging extremely hard
- Unit testing hard/impossible
- Replication hard/impossible
- Code review hard
- European Spreadsheet Risk Interest Group horror stories:
  - Reinhart & Rogoff: justification for S. European austerity measures
  - JP Morgan Basel II VAR: risk understated
  - IOC: 10,000 tickets oversold
  - Knox County, TN; W. Baraboo Village, WI; … : errors costing $millions
- According to KPMG and PWC, over 90% of corporate spreadsheets have errors

Bug in the PRNG for many generations of Excel, allegedly fixed in Excel 2010.

Other long-standing bugs in Excel; PRNG still won't accept a seed; etc.

Problems with Excel

https://unimibox.unimi.it/index.php/s/9xWsHEzJamYjZCy

UnimiBox **Short Course Data Science Prof. Kenett** Download all files

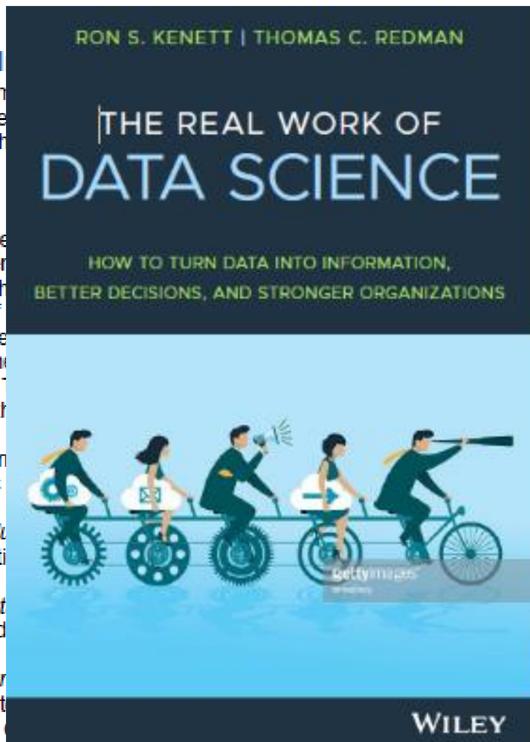
Name	Size	Modified
SAS_JMP_Pro_14	25 KB	39 minutes ago
JMP course files.zip	3 MB	2 hours ago
Kenett Analytics 2020.pdf	11 MB	2 hours ago
Kenett Causality 2020.pdf	11.3 MB	2 hours ago

jmp Statistical Discovery.™ From SAS.

Chapter 13: Evaluating data science outputs more formally

In the last chapter we focused on teaching your colleagues some basics and providing a starte … ecision make … ey gain exper … of inforn … per, facilit … infor … in the use c … orithmic analysis. The information quality framework (InfoQ) addresses outputs from both approaches, in the context of business, academic, services and industrial work.

Class assignment (in teams of ~5)



Assessing I…

The InfoQ fran … analytic work. InfoQ is define … is, f, on a given dataset X, with … ed

As an example … m by launching a customer reter … ners with high potential for ch … nsists of customer usage, lists of … nd problems reported to the … ee, f, which will help him defin … milar churn probabilities. … ign only on customers with

InfoQ, is deten … ally in the context of the specific

(1) *Data resolu…* … ty, and level of data aggregati…

(2) *Data struct…* … ured and unstructured d…

(3) *Data integr…* … ntegrated together? Not … ata definitions, different units …

(4) *Temporal relevance*: Is the time-frame in which the data were collected relevant to the goal?

The Real Work of Data Science: Turning data into information, better decisions, and stronger organizations

(5) *Generalizability*: Are results relevant in a wider context? In particular, is the inference from the sample population to target population appropriate (statistically generalizable, Chapter 8)? Can other considerations be used to generalize the findings?

(6) *Chronology of data and goal*: Are the analyses and needs of the decision-maker synched up in time?

(7) *Operationalization*: Are results presented in terms that can drive action?

(8) *Communication*: Are results presented to decision-makers at the right time and in the right way (as described in Chapter 7)?

See Appendix A3 for a detailed list of questions used in InfoQ assessments.

Importantly, InfoQ helps structure discussions about trade-offs, strengths and weaknesses. Consider the cellular operator noted above and consider a second potential dataset X*. X* includes everything X has, plus data on credit-card churn, but that additional data won't be available for two months. Resolution (the first dimension) goes up, while temporal resolution (the fourth) goes down. Or suppose a new machine-learning analysis, f*, has been conducted in parallel, but results from f and f* don't quite line up. "What to do?" These are the most important discussions for decision-makers, data scientists, and CAOs.

Further, the InfoQ framework can be used in a variety of settings, not just helping decision makers become more sophisticated. It can also be used to assist in the design of a data science project, as a mid-project assessment, and as a post mortem to sort out lessons learned. See Kenett and Shmueli (2016) for a comprehensive discussion of InfoQ and its applications in risk management, healthcare, customer surveys, education and official statistics.

A Hands-On Information Quality Workshop

This workshop uses InfoQ to help an entire team understand the importance of clear goals and what it takes to achieve information quality with respect to those goals. It combines individual work, team discussions, and group presentations, using this information quality framework.

Phase I: Individual work
Please consider the four steps below and document each for further discussion.

Step 1: The background

a. Pick an organization to focus on. It should be one that you know reasonably well, such as your current or previous place of employment, a school, hospital, or restaurant.

b. Answer the following: Who are this organization's most important customers and suppliers? What are its most important products and services?

Install InfoQ.jmpaddin

Jmp.com/infoqscore

https://community.jmp.com/kvoqx44227/attachments/kvoqx44227/add-ins/338/1/InfoQ.jmpaddin

Class assignment (in teams of ~5)

1. Select one of the three supplied case studies
2. Review the report and presentation.
3. Evaluate its information quality using JMP add in.
4. Prepare a ppt report and assign a spokesperson

Task 1/3 to get a pass/fail grade

Step 2: The data
List various data sources that are available to support help decision makers pursue that goal. In evaluating data sources, focus on data quality and data clarity. Data quality reflects to what extent the data can be trusted and data clarity represents the way data elements are defined and collected by various parts of the organization. This step specifies the X component of InfoQ.

Step 3: The analysis
Identify several approaches for analyzing the data in order to help the organization achieve its goal. In this step alternatives methods of analysis, f_1, f_2, …, f_p, are identified and listed.

Step 4: Assessment:
Assess the data and the potential analysis on eight info Q dimensions "dimensions" using a 1-5 score where 1 means "very poorly" and 5" very well."

1. *Data resolution*. When the data are on the right level of granularity, the scale of measurement scale is appropriate, and the level of aggregation appropriate, score a "5."
2. *Data structure*. When there are important gaps in the data coverage, score a "1."
3. *Data integration*. A "5" corresponds to integration into a seamless whole.
4. *Temporal relevance*. When the data is timely with respect to the goal, score a "5.".
5. *Generalizability*. When what we learn can be generalized to many other circumstances, score a "5."
6. *Chronology of data and goal*. When the analysis and recommendations can be completed in a timely fashion from a decision-making perspective, score a "5.".
7. *Operationalization*. If the analyses are unlikely to lead to concrete actions that provide business benefit, score a "1."
8. *Communication*. If the "who," (needs the information), "what," "when," "why," and "how" are clear, score a "5."

Note: An application for recording InfoQ scores, which also allows for a range of values reflecting uncertainty in the score, is available for download from the Wiley website of Kenett and Shmueli (2016). The application requires installation of the JMP software and provides an overall InfoQ score based on the geometric mean of the individual dimension scores.

Three case studies (1)

1. Predicting Changes in Quarterly Corporate Earnings Using Economic Indicators

http://www.galitshmueli.com/data-mining-project/predicting-changes-quarterly-corporate-earnings-using-economic-indicators

This study looks at corporate earnings in relation to an existing theory of business forecasting developed by Joseph H. Ellis (former research analyst at Goldman Sachs).

Three case studies (2)

2. Predicting ZILLOW.com's Zestimate accuracy

http://www.galitshmueli.com/data-mining-project/predicting-zillowcom-s-zestimate-accuracy

Zillow.com is a free real estate service that calculates an estimated home valuation ("Zestimate") as a starting point for anyone to see for most homes in the U.S. The study looks at the accuracy of Zestimates.

Three case studies (3)

3. Predicting First Day Returns for Japanese IPOs

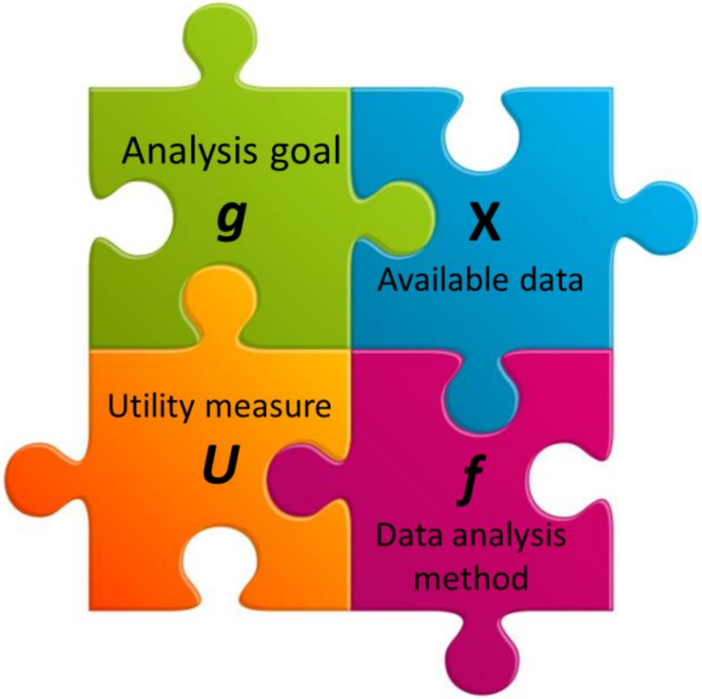
http://www.galitshmueli.com/data-mining-project/predicting-first-day-returns-japanese-ipos

An Initial Public Offering (IPO) is the first sale of stock by a company to the public. The study looks at the first-day returns on IPOs of Japanese companies.

Information Quality

The potential of a particular dataset to achieve a particular goal using a given empirical analysis method

g	**A specific analysis goal**
X	**The available dataset**
f	**An empirical analysis method**
U	**A utility measure**



$$InfoQ(f,X,g) = U(\, f(X|g)\,)$$

Depends on quality of *g, X, f, U* and relationship between them

Kenett, R.S. and Shmueli , G. (2013) On Information Quality, http://ssrn.com/abstract=1464444
Journal of the Royal Statistical Society, Series A (with discussion), 176(4).

Explain, predict, describe
enumerative, analytic,
exploratory, confirmatory

Goal Specification

- “error of the third kind” - giving the right answer to the wrong question – A. Kimball
- “Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise” – John Tukey

Typical Goals of Customer Surveys

Goal 1. **Decide** where to launch improvement initiatives

Goal 2. **Highlight** drivers of overall satisfaction

Goal 3. **Detect** positive or negative trends in customer satisfaction

Goal 4. **Identify** best practices by comparing products

Goal 5. **Determine** strengths and weaknesses

Goal 6. **Set up** improvement goals

Goal 7. **Design** a balanced scorecard with customer inputs

Goal 8. **Communicate** the results using graphics

Goal 9. **Assess** the reliability of the questionnaire

Goal 10. **Improve** the questionnaire for future use

Data Source

- Primary, secondary
- Observational, experiment
- Single, multiple sources
- Collection instrument, protocol

Data Size and Dimension

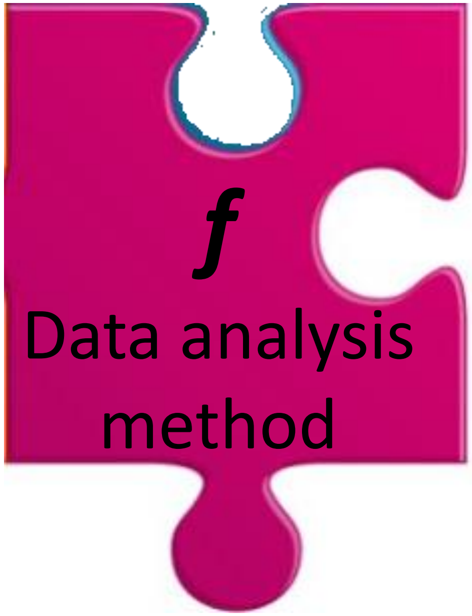
- # observations
- # variables

Data Type

- Continuous, categorical, semantic
- Structured, un-, semi-structured
- Cross-sectional, time series, panel, network, geographical

Data Quality

- "Zeroth Problem - How do the data relate to the problem, and what other data might be relevant?" – C. Mallows
- *Quality of Statistical Data* (IMF, OECD) - usefulness of summary statistics for a particular goal (7 dimensions)



Statistical models and methods

- Parametric, semi-, non-parametric
- Classic, Bayesian

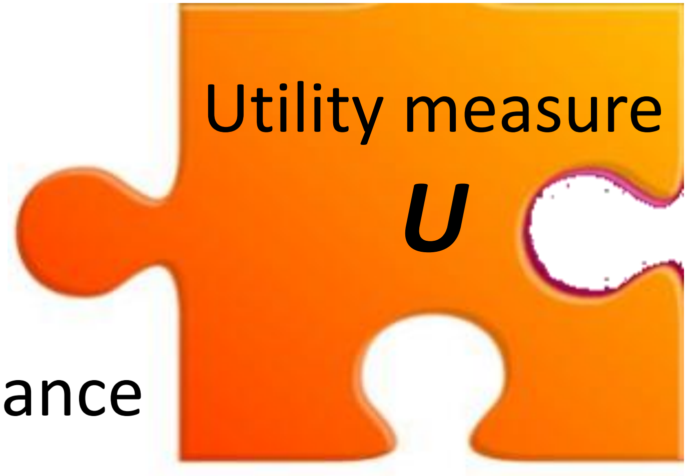
Data mining algorithms

Graphical methods

Operations research methods

Analysis Quality

- "poor models and poor analysis techniques, or even analyzing the data in a totally incorrect way." - B. Godfrey
- Analyst expertise
- Software availability
- The focus of statistics education



- Predictive accuracy, lift
- Goodness-of-fit
- Statistical power, statistical significance
- Strength-of-fit
- Expected costs, gains
- Bias reduction, bias-variance tradeoff

Utility Measure

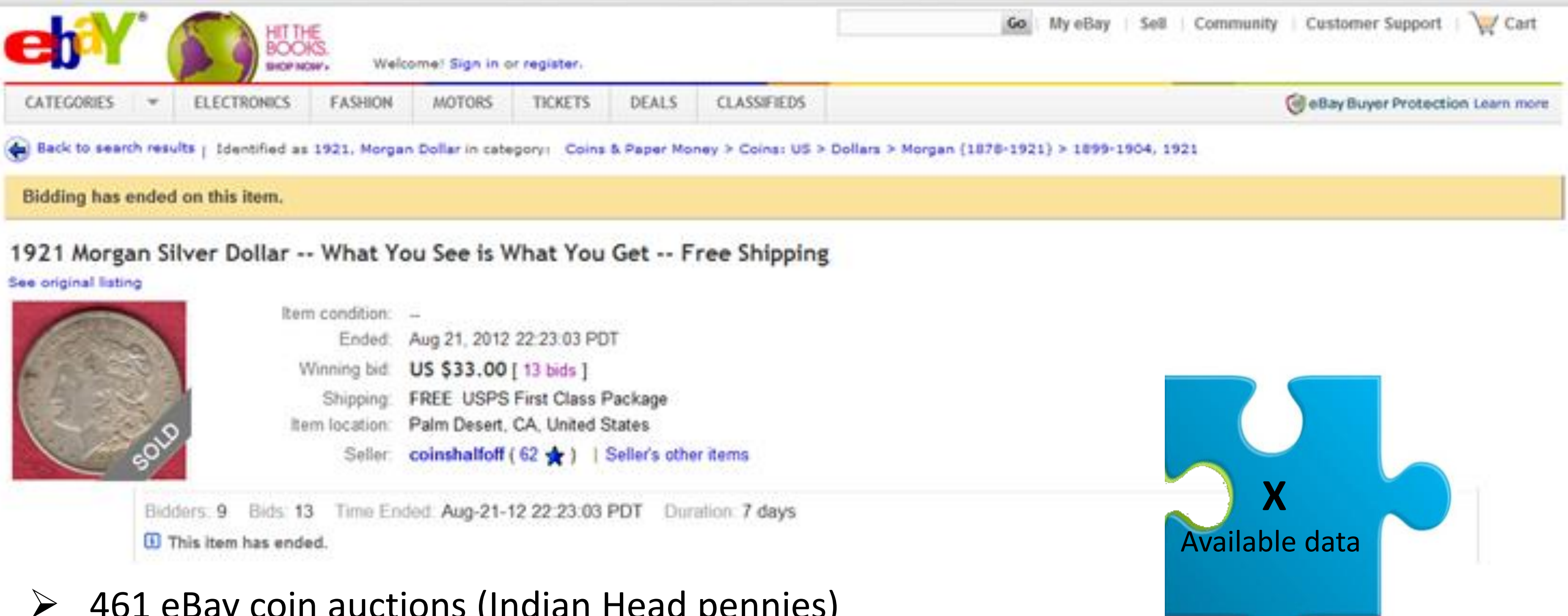
- Adequate metric from analysis standpoint (R^2, holdout data)
- AUC, ROC, confusion matrix
- MAPE, RMSE, AIC, BIC, generalizability
- Adequate metric from domain standpoint

An example....

Goal of study:

1. Predict the final price of an Ebay auction at start of auction
2. Predict price during ongoing auction
3. Predict the auctions with the highest prices (ranking)
4. Identify factors that determine the final price of an eBay auction?

"Pennies from ebay: The determinants of price in online auctions" Lucking-Reiley D., Bryan D., Prasad N. & Reeves D. *Journal of Indust. Econ.*, 2007



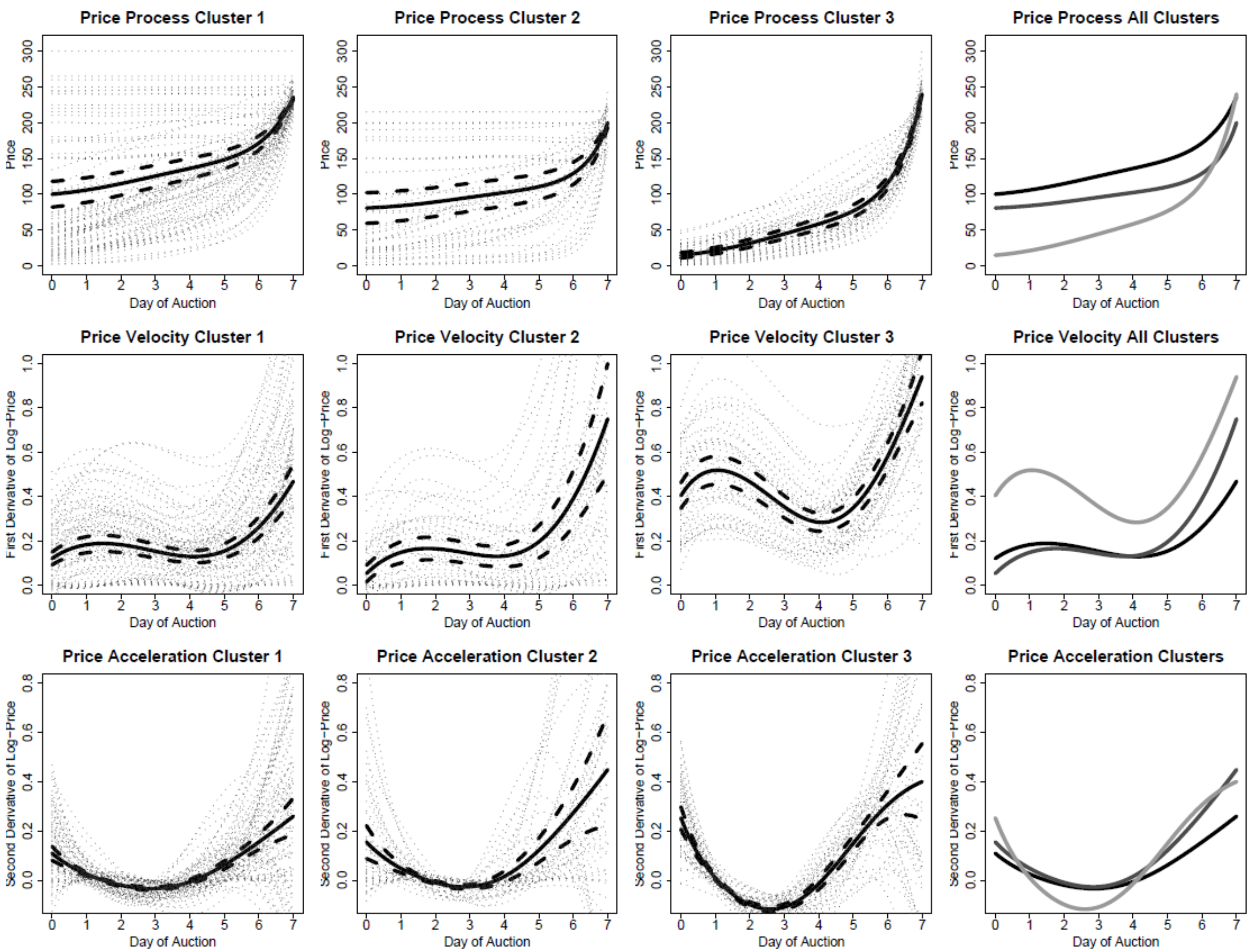
- 461 eBay coin auctions (Indian Head pennies)
- Auction characteristics
  - Duration
  - Open and close prices
  - Number of bids and bidders
  - Secret reserve price
  - Weekday/weekend ending
- Seller characteristics
  - Seller rating
- Item characteristics
  - Year and grade of coin

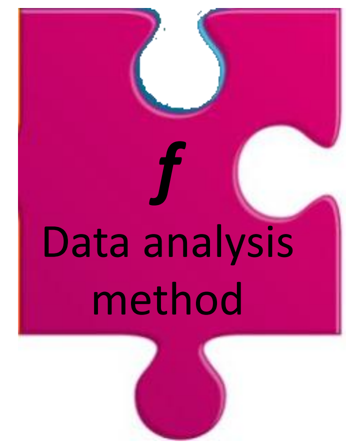
"Pennies from ebay: The determinants of price in online auctions"
Lucking-Reiley D., Bryan D., Prasad N. & Reeves D. *Journal of Indust. Econ.*, 2007

Abhijit Banerjee and Esther Duflo: The Nobel couple fighting poverty
The team pioneered "randomized controlled trials", or RCTs, in economics. https://www.bbc.com/news/world-asia-india-50048519

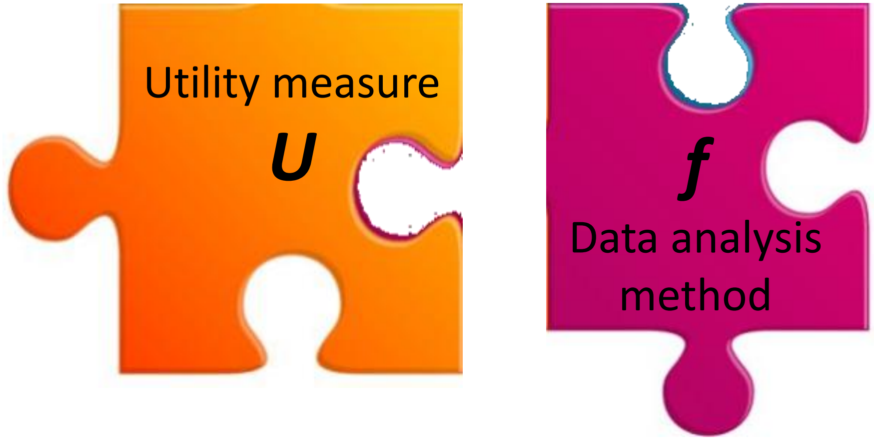
An example....

Dimension Reduction



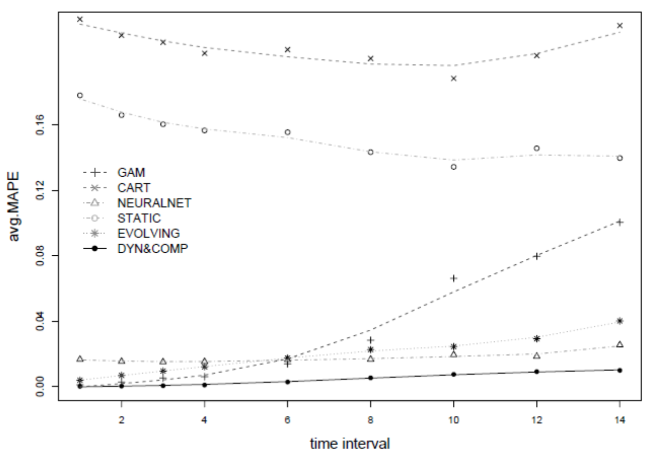


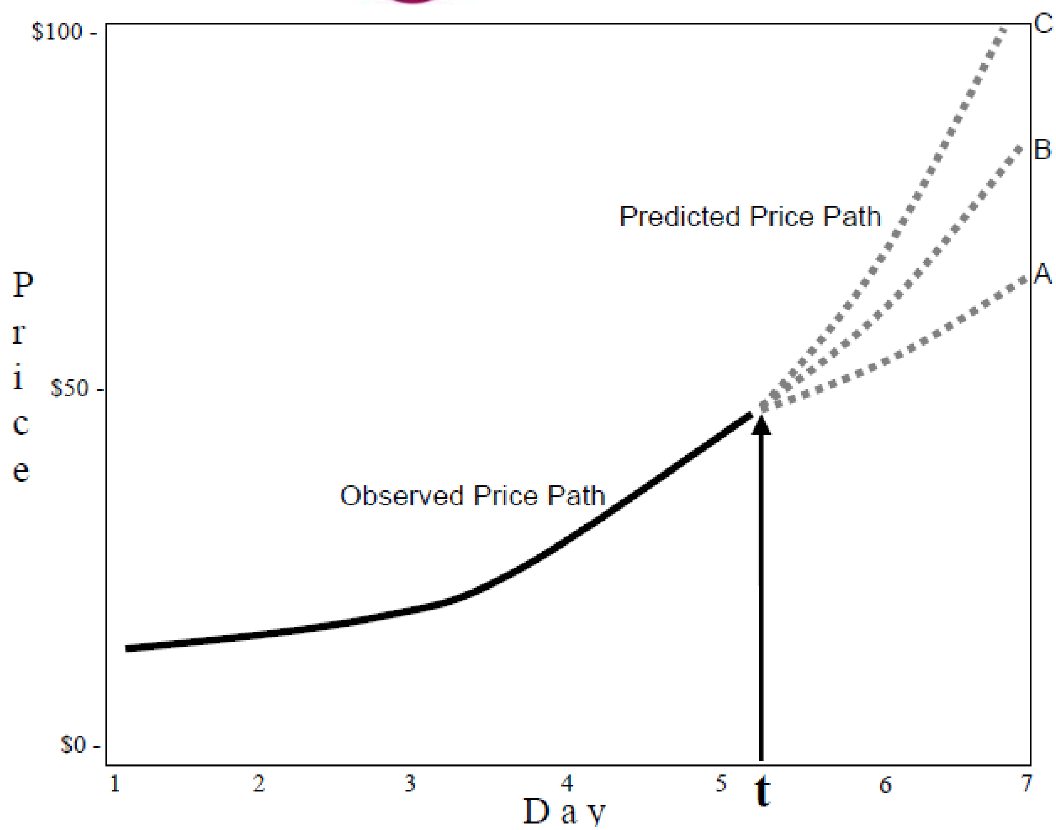
An example....



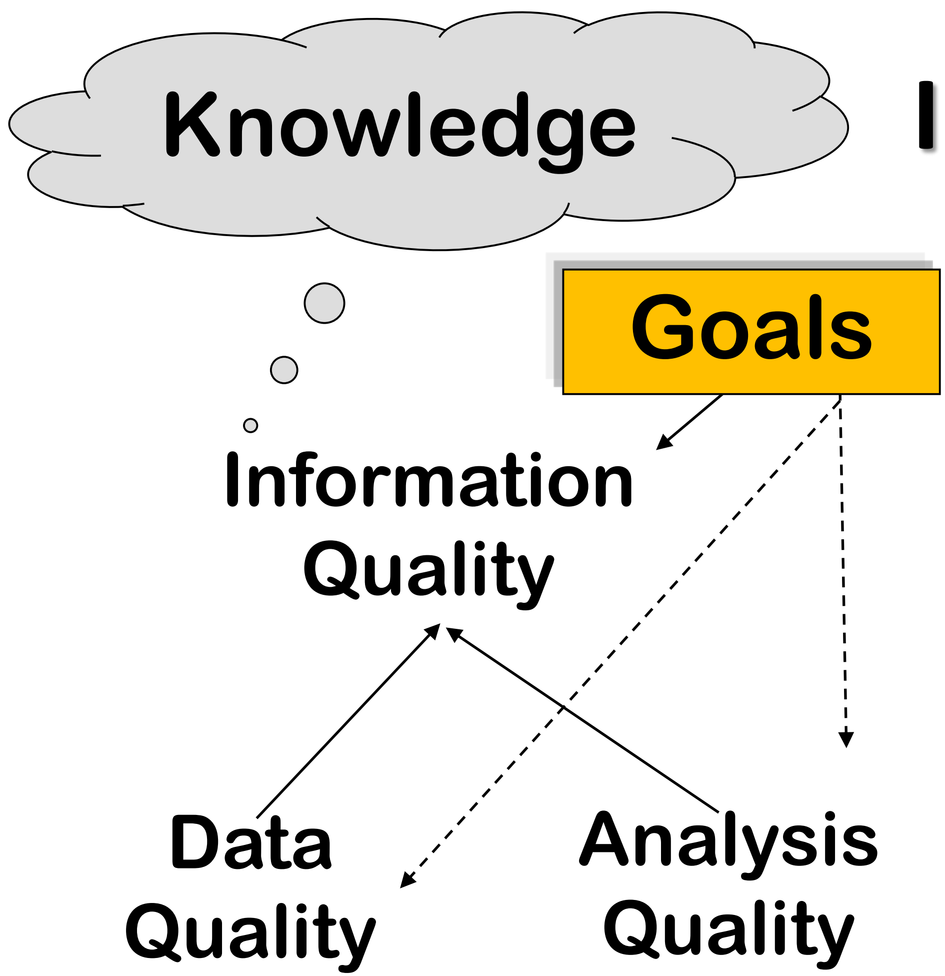
Prediction error:

- Holdout data
- Metrics such as *MAPE* and *RMSE*





Information Quality



$$InfoQ(f,X,g) = U(f(X|g))$$

g	**A specific analysis goal**
X	**The available dataset**
f	**An empirical analysis method**
U	**A utility measure**

What

How

1. Data resolution
2. Data structure
3. Data integration
4. Temporal relevance
5. Chronology of data and goal
6. Generalizability
7. Operationalization
8. Communication

Primary Data
- Experimental
- Observational

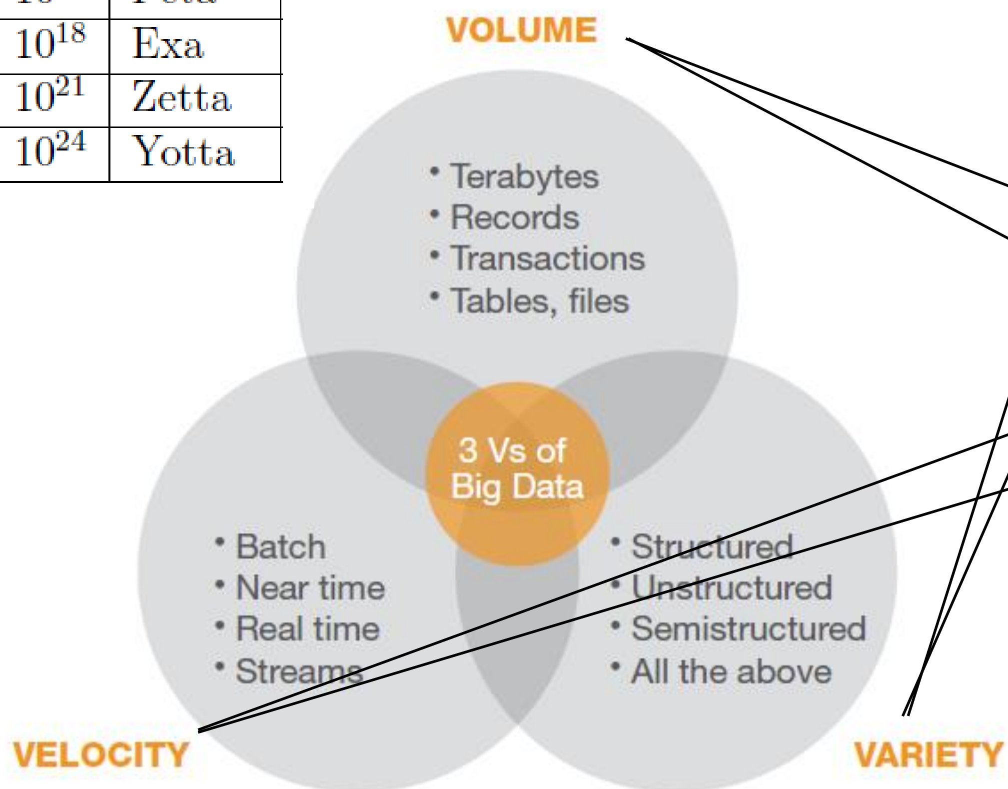
Secondary Data
- Experimental
- Observational

Massive data sets

Big data Analytics

Power	Prefix
10^9	Giga
10^{12}	Tera
10^{15}	Peta
10^{18}	Exa
10^{21}	Zetta
10^{24}	Yotta

V V V

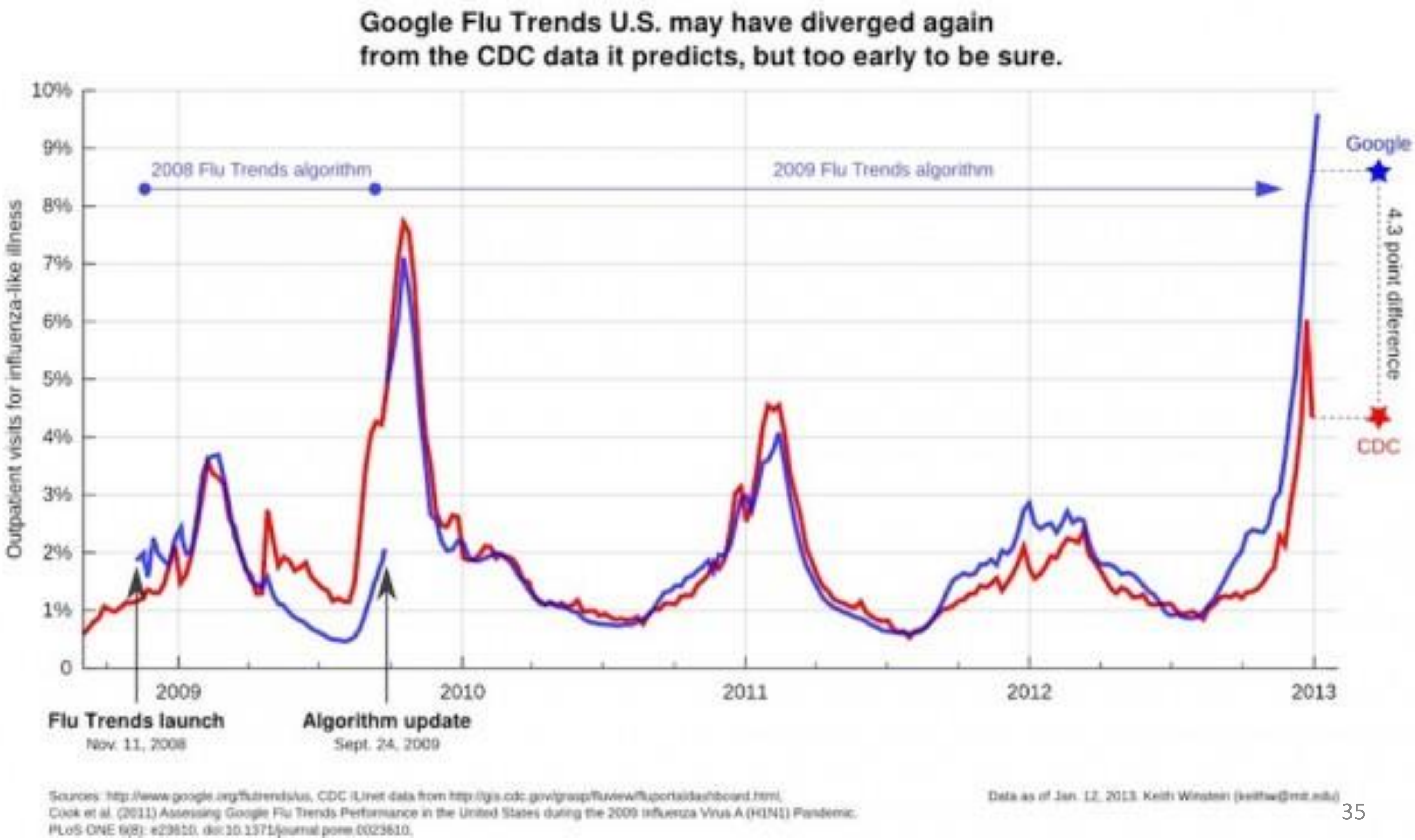


1. Data resolution
2. Data structure
3. Data integration
4. Temporal relevance
5. Chronology of data and goal
6. Generalizability
7. Operationalization
8. Communication

InfoQ

Russom, P., Big Data Analytics, TDWI Best Practices Report, Q4 2011

#1 Data Resolution



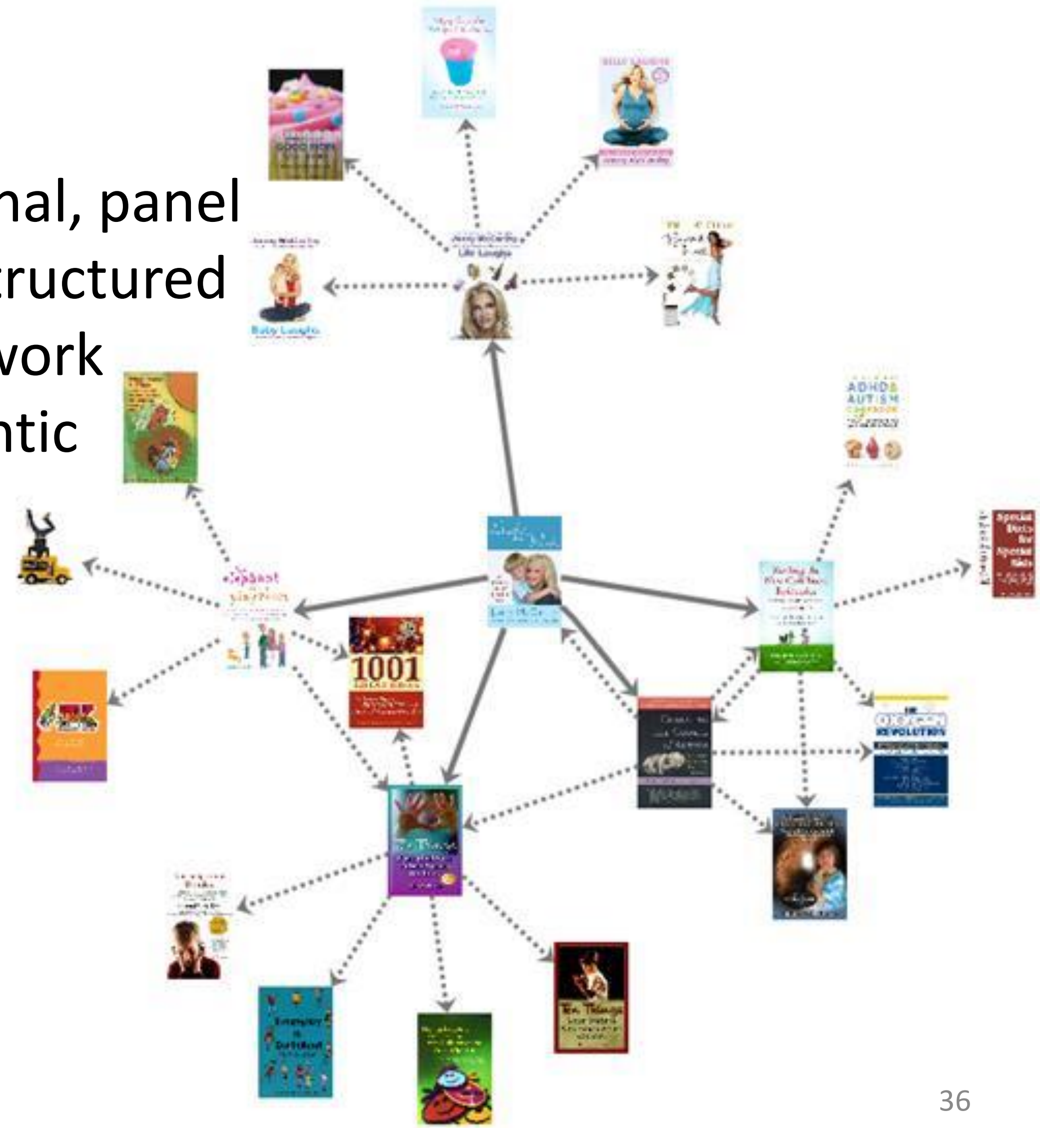
#2 Data Structure

Data Types

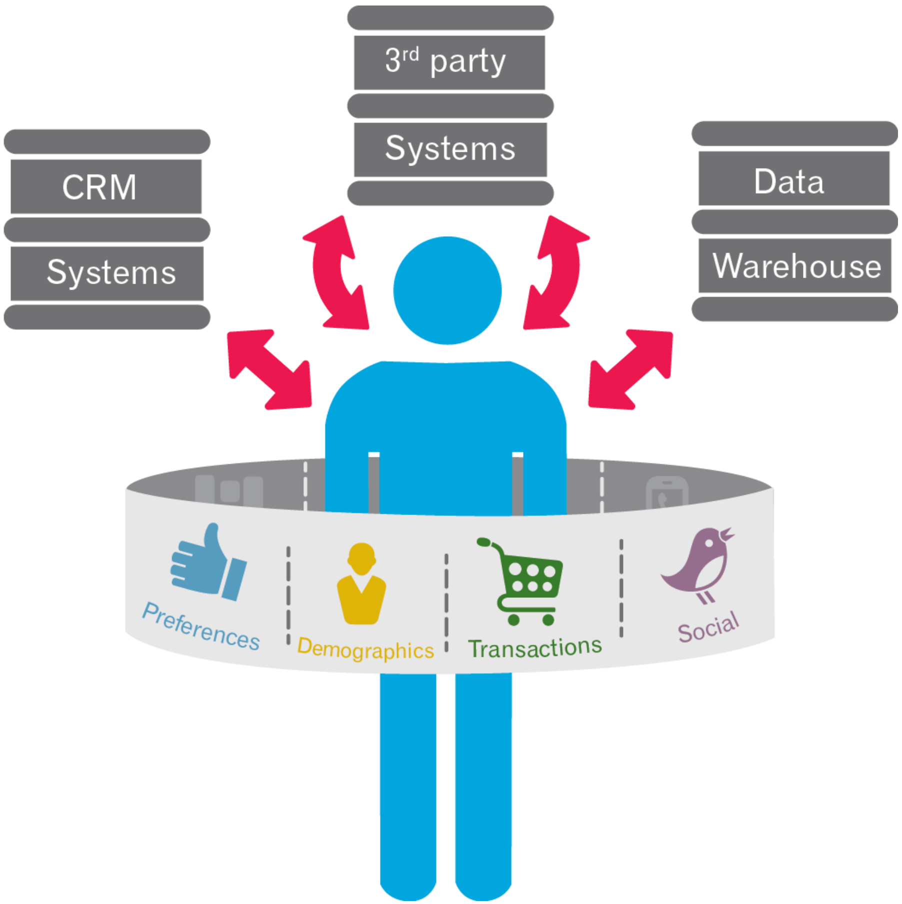
- Time series, cross-sectional, panel
- Structured, semi-, non-structured
- Geographic, spatial, network
- Text, audio, video, semantic
- Discrete, continuous

Data Characteristics

Corrupted and missing values due to study design or data collection mechanism

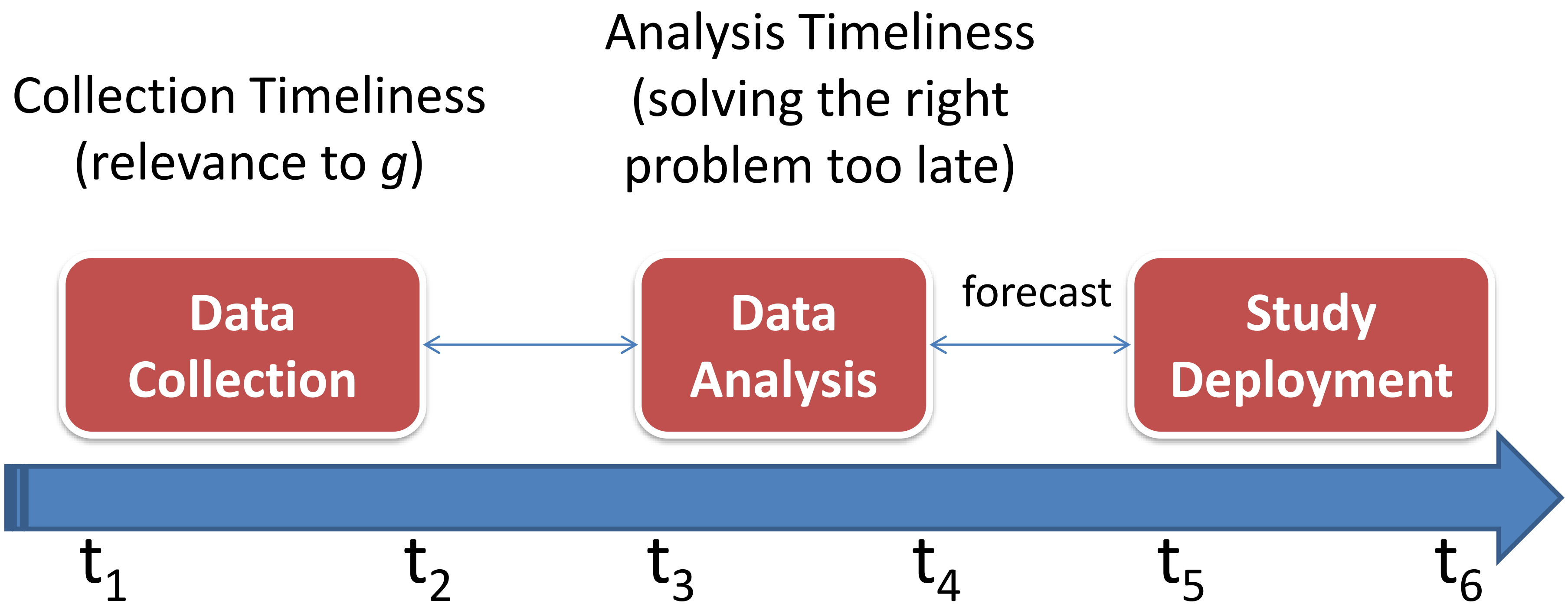


#3 Data Integration



Linkage, privacy-preserving methods: Increase or decrease InfoQ?

#4 Temporal Relevance



g: Prospective vs. retrospective; longitudinal vs. snapshot

Nature of X, complexity of *f*

#5 Chronology of Data & Goal

Air Quality Index (AQI) Values	Levels of Health Concern
0 to 50	Good
51-100	Moderate
101-150	Unhealthy for Sensitive Groups
151-200	Unhealthy
201-300	Very Unhealthy
301 to 500	Hazardous

http://www.airnow.gov/?action=aqibasics.aqi

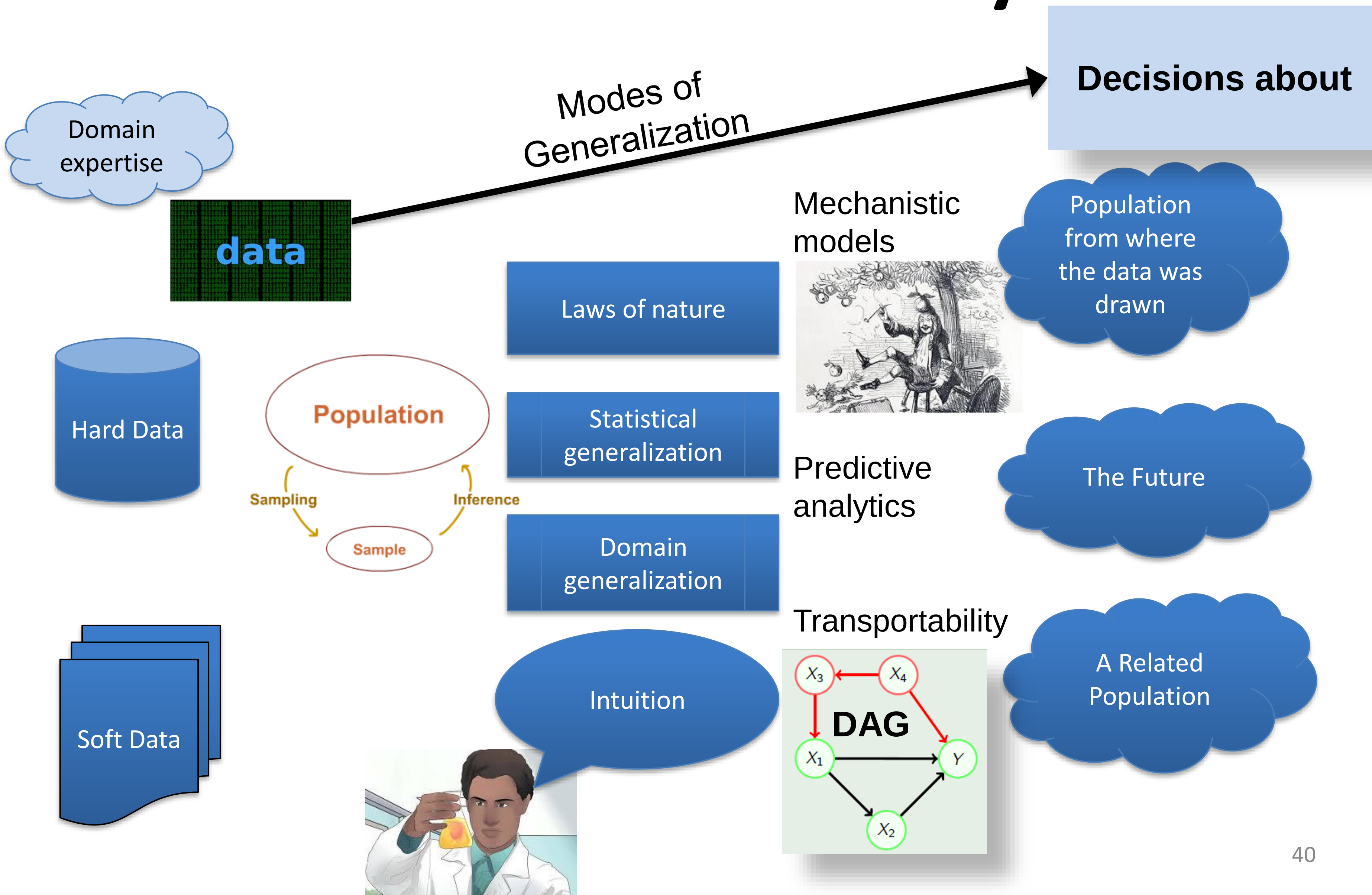
Data: Daily AQI in a city

g_1: Reverse-engineer AQI

g_2: Forecast AQI

Retrospective/prospective
Ex-post availability
Endogeneity

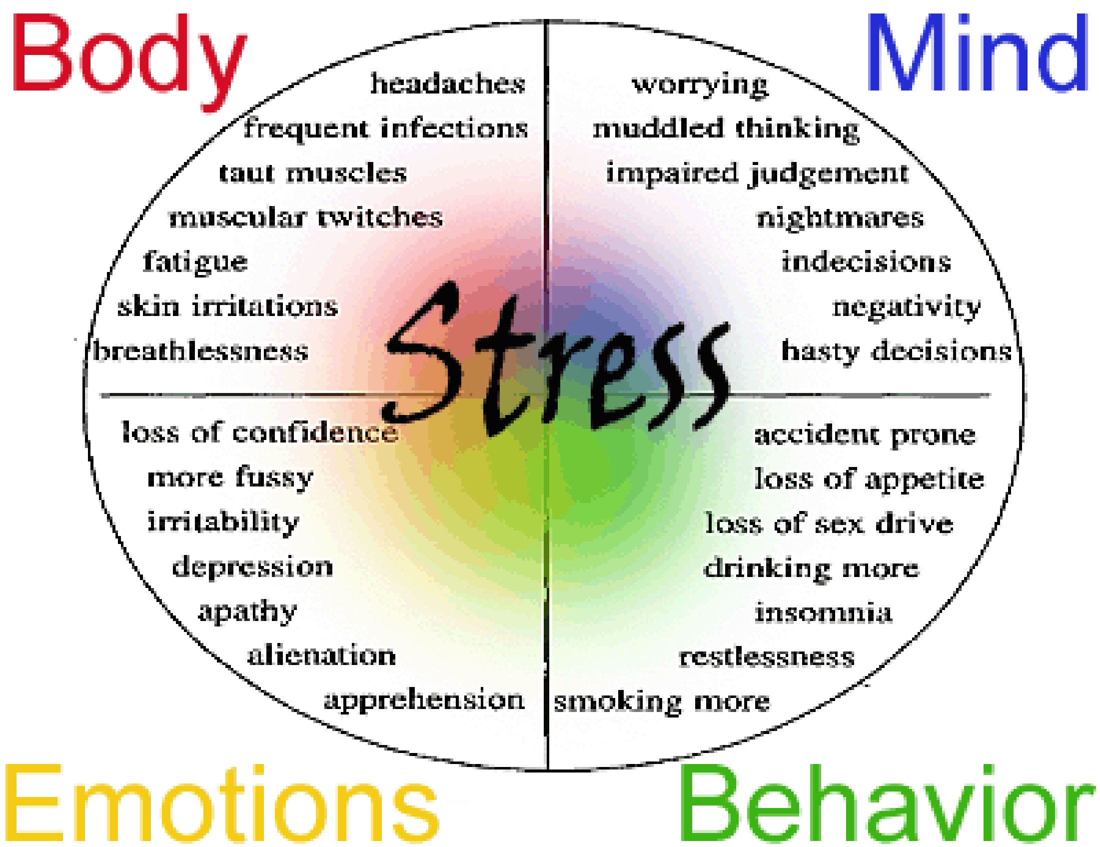
#6 Generalizability



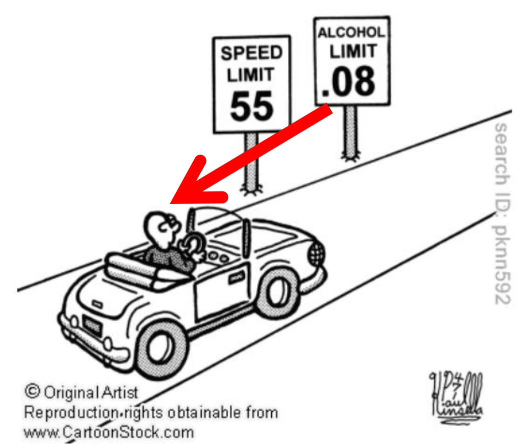
#7 (Construct) Operationalization

X: construct

X = θ(χ) operationalization (measurable)



- Causal explanation vs. prediction, description
- Theory vs. data
- Data: Questionnaire, physio measurement



#7 (Action) Operationalization

In the pre-publication drafts of *Quality, Productivity, and Competitive Position* Dr. Deming wrote:

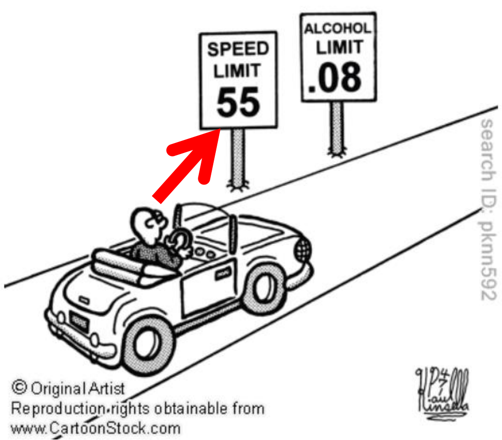
> "An operational definition consists of (1) a criterion to be applied to an object or a group of objects, (2) a test of compliance for the object or group, and (3) a decision rule for interpreting the test results as to whether the object or group is, or is not, in compliance."

In Dr. Deming's own conversations, when individuals would start telling him about what they or their organization were planning to do, he would invariably have one of two responses for them: "By what method?" or "How will you know?" Either one of these questions would generally end the conversation since the individual would have no answer. After discerning this pattern to Dr. Deming's responses, it finally occurred to me that these two questions corresponded to the last two parts of an operational definition. This realization, in turn, resulted in a generalization of an operational definition to become:

(1) What do you want to accomplish?

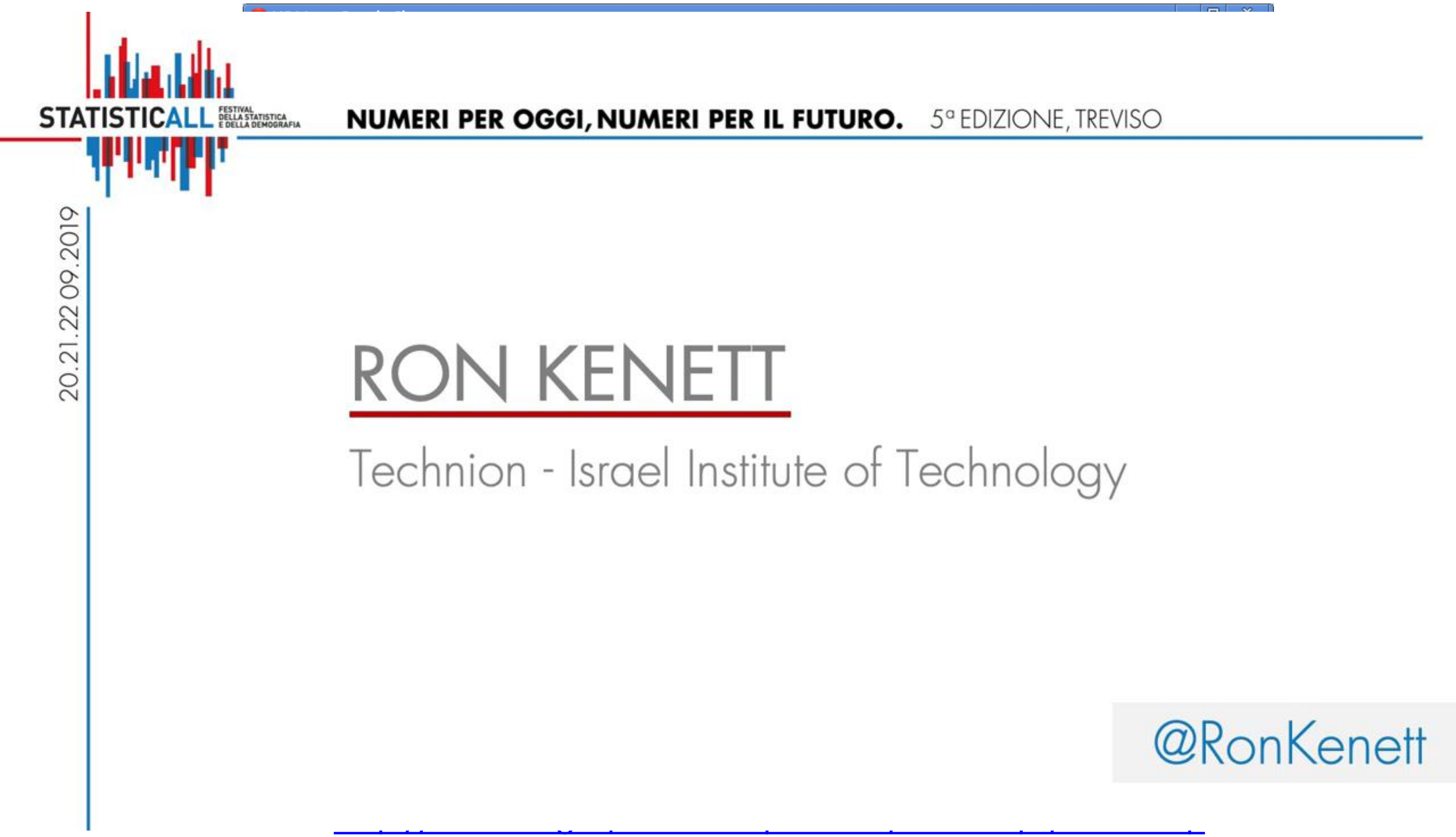
(2) By what method will you accomplish it?

(3) How will you know when you have accomplished it?



http://www.spcpress.com/pdf/DJW187.pdf

#8 Communication

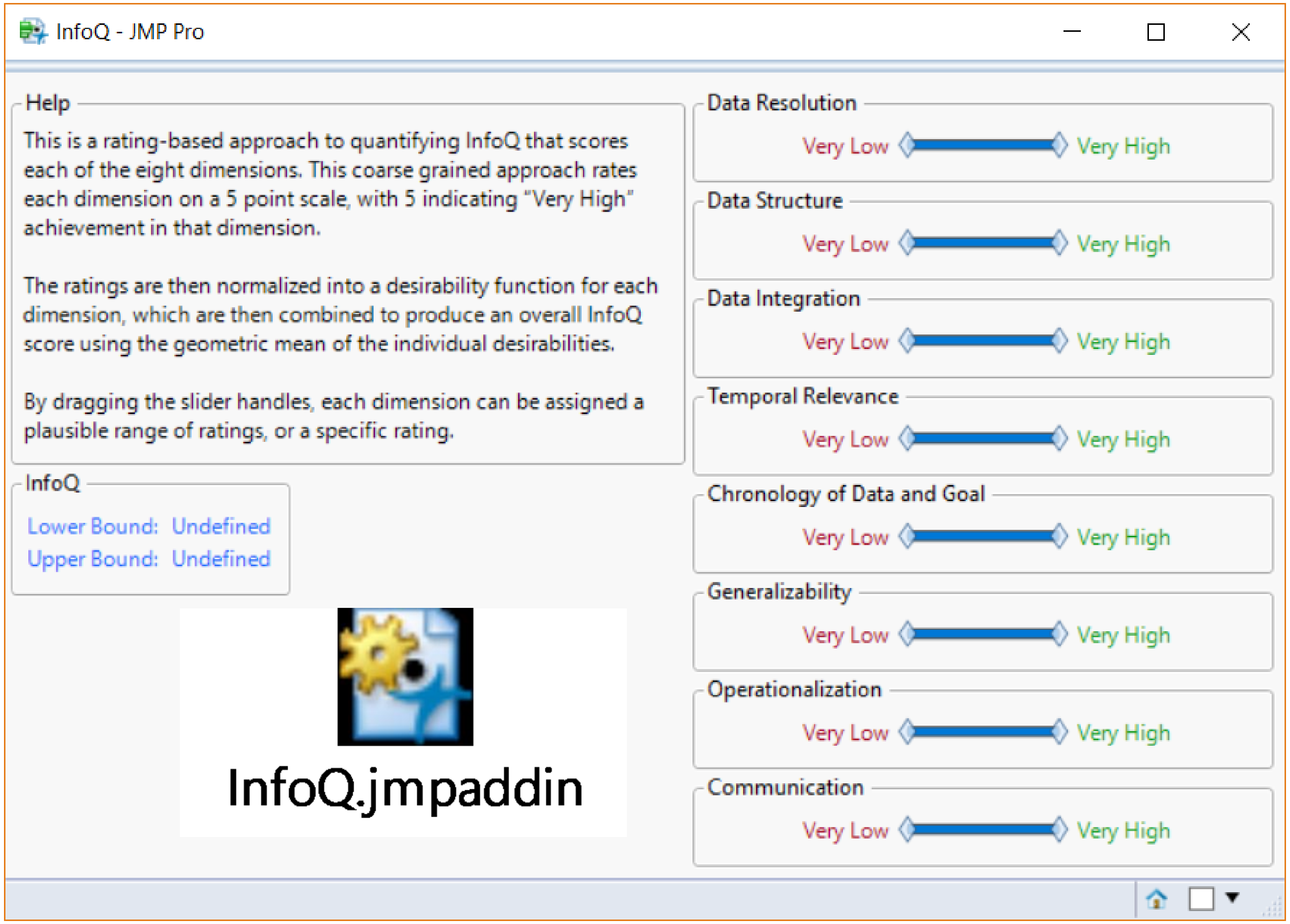


Open JMP, double click on InfoQ.jmpaddin

Assessing InfoQ

Rating-based assessment (1-5 scale on each dimension)

$$InfoQ\ Score = [d_1(Y_1)\ \ d_2(Y_2)\ \ \ldots\ \ d_8(Y_8)]^{1/8}$$



Three case studies (1)

1. Predicting Changes in Quarterly Corporate Earnings Using Economic Indicators

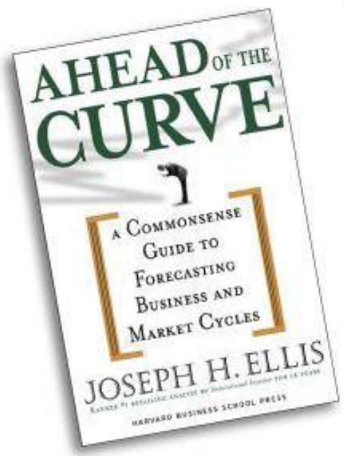
Stages in economic downturn: 1) the peak, 2) modest slowing, 3) intensifying worrying by investors (a lot of panic selling occurs in this stage), and 4) the advent of recession. **Can we predict the economic slowdown in corporate earnings (S&P 500 EPS) well in advance?**

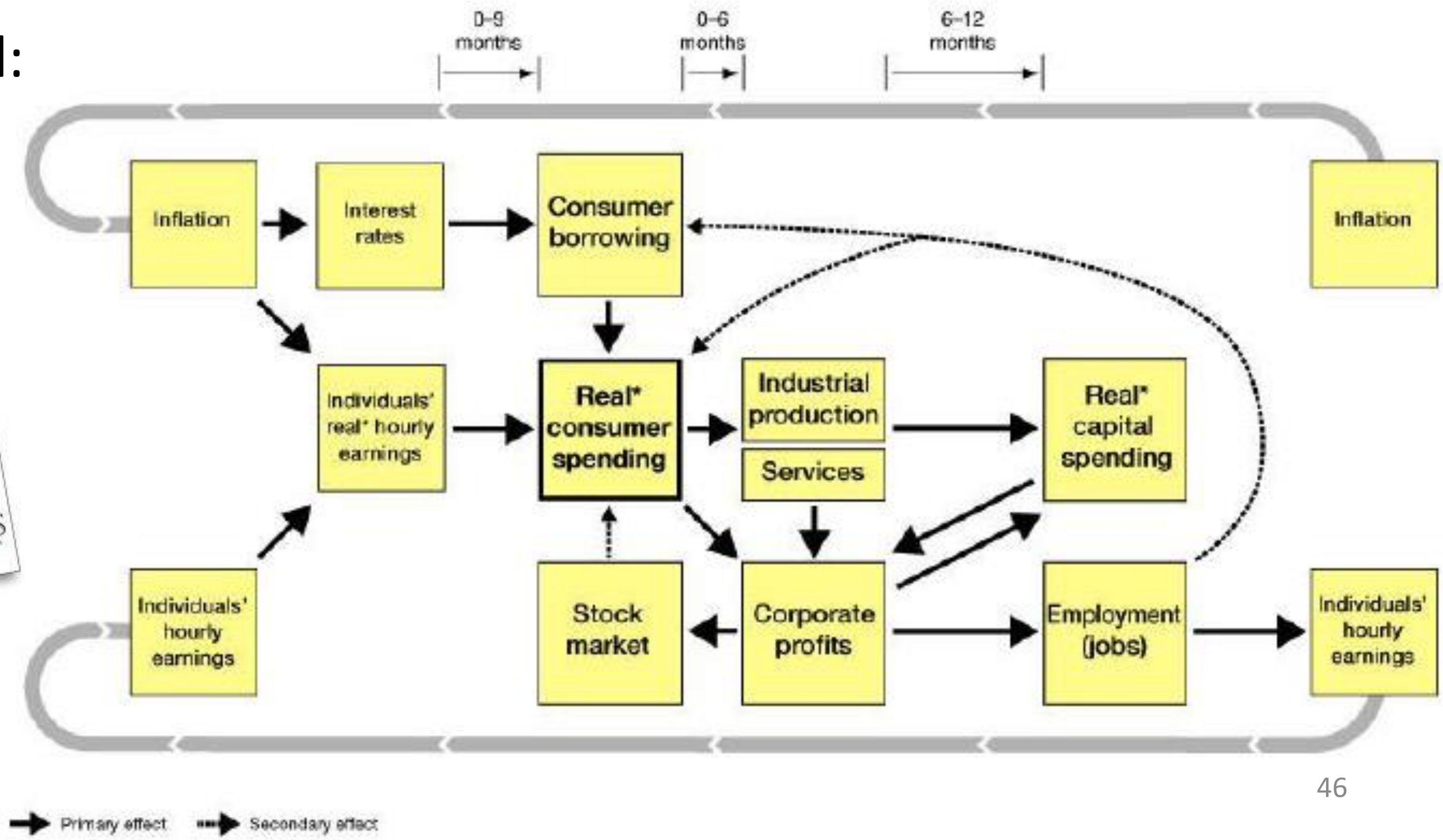
Ellis claims (based on observations) there is a 0-9 month lag between wages and its effect on consumer spending. 0-6 months until changes in consumer spending affects changes in industrial production. Another 6-12 months between industrial production and capital spending. And finally, another 6-12 between capital spending and its effects on Corporate Profits.

Three case studies (1)

1. Predicting Changes in Quarterly Corporate Earnings Using Economic Indicators

Ellis model:



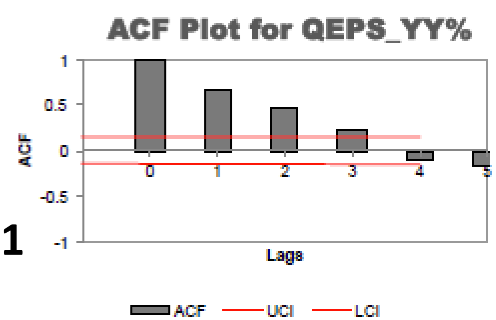


Three case studies (1)

1. Predicting Changes in Quarterly Corporate Earnings Using Economic Indicators

The data: i) 180 quarters. 6 [Economic] x variables. Ii) Change in S&P EPS = y variable, iii) All variables transformed to year vs year % change, iv(All data used is publicly available via websites of US agencies: BEA, BLS, FED, and S&P.

The analysis: XLMiner on these different versions of datasets. Partitioned it. Ran predictor applications: ACF Plots, MLR, Regression Tree – full and pruned.

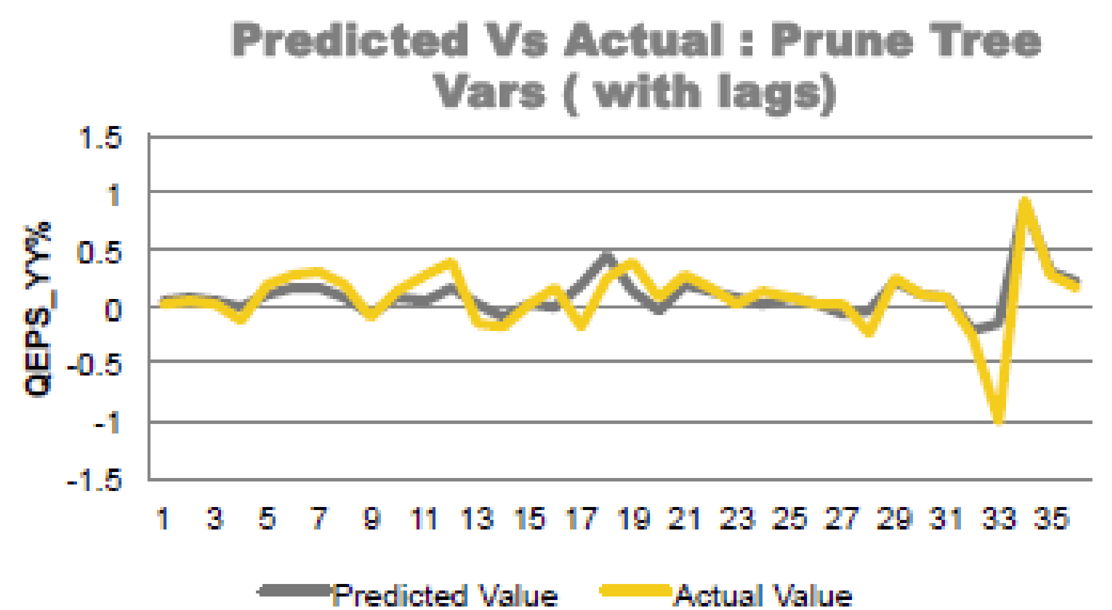


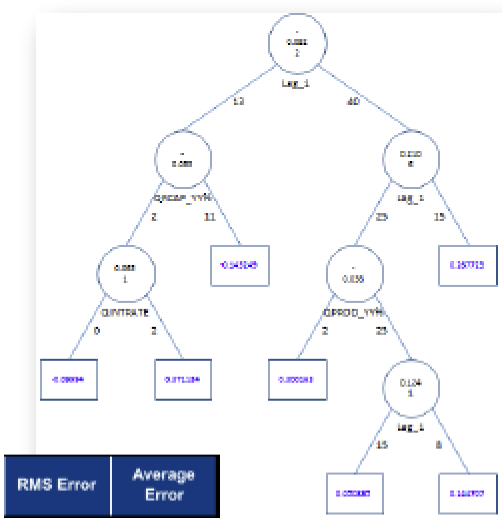
Auto Correlation Chart. Based on this, took Lag_1 as one of the predictors. Lag_1 = QEPS_YY(Q-1)

Three case studies (1)

1. Predicting Changes in Quarterly Corporate Earnings Using Economic Indicators

*QEPS_YY%(t) = 0.0486 + 0.747*QEPS_YY%(t-1) -0.517*QRCAP_YY%(t-2)*





RMS Error	Average Error
0.1896416	-0.0131939

Three case studies (1)

3

Data Resolution: 3
After estimation, measures regarding the goodness of fit such as The R-squared measure are not high

4-5

Data Structure: 5
No problem of missing data.
Moreover all data collections start from the same data (1964)

5

Data Integration: 5
We have a good integration of data.
During the research, all data went through all process of normalization

4

Temporal Relevance: 4
We started from 1964 since previous data were missing. With more data the anlaysis would be more accurate

2-3

Generalizability: 2
The analysis regards only the S&P index. In order to generalize the results of the project we should use data also from other source that are not always available

5

Chronology of data and goal: 5
Prediction is the aim of the project.
As a result the chronology of data is very important

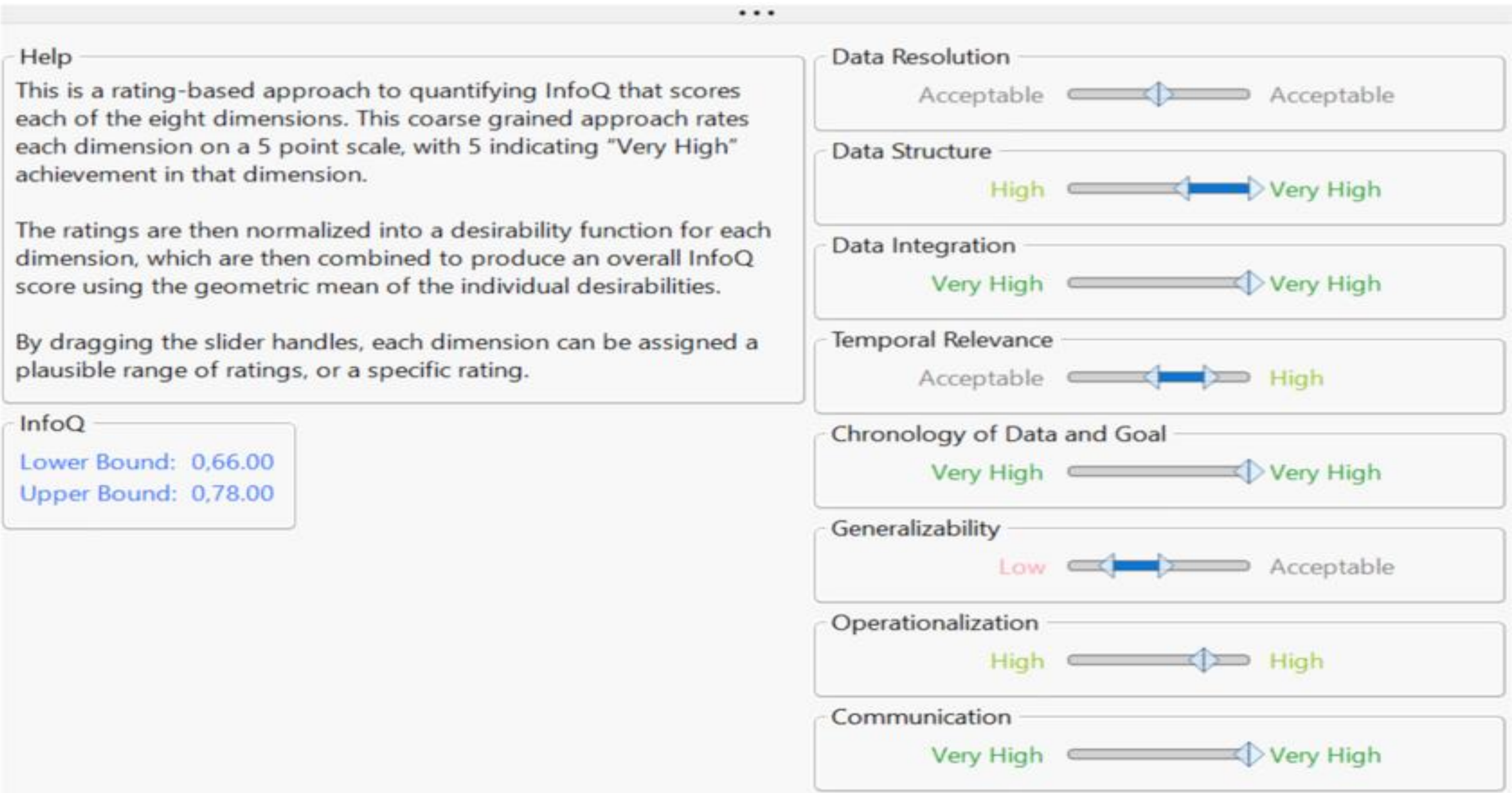
4

Operationalization: 4
The project can be applied in real life context. It would be interesting to show the result for other kind of index

5

Communication: 5
The analysis is clearly explained step by step from data processing to conlusion

Three case studies (1)



Three case studies (2)

2. Predicting ZILLOW.com's Zestimate accuracy

- "Zillow.com" is a real estate service launched in 2006
- It calculates a Zestimate-home valuation for most homes in the U.S
- For MD and VA it gets only about 26% of predictions within the +/-5% range only.

1. Home Type (Single Family, Condo , etc)
2. No of Bed Rooms
3. No of Bath Rooms
4. Total Area –Sqft
5. Lot size –Sqft
6. No of Stories
7. Total Rooms
8. Distance from Metro
9. Primary School Rank
10. Middle School Rank
11. High School Rank
12. Age of house at Sale
13. Sale Season (Fall , Winter , etc)
14. Recession Period (Y/N)
15. Sales Volume

The data

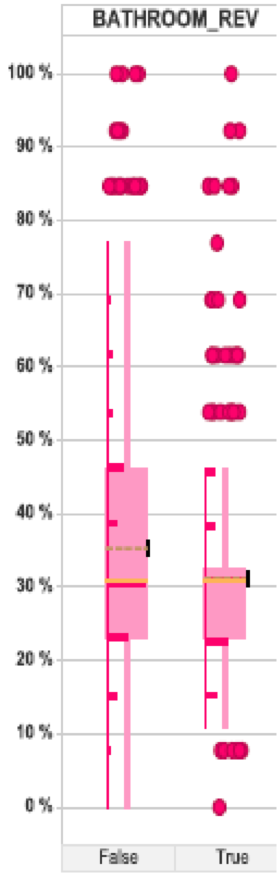
Three case studies (2)

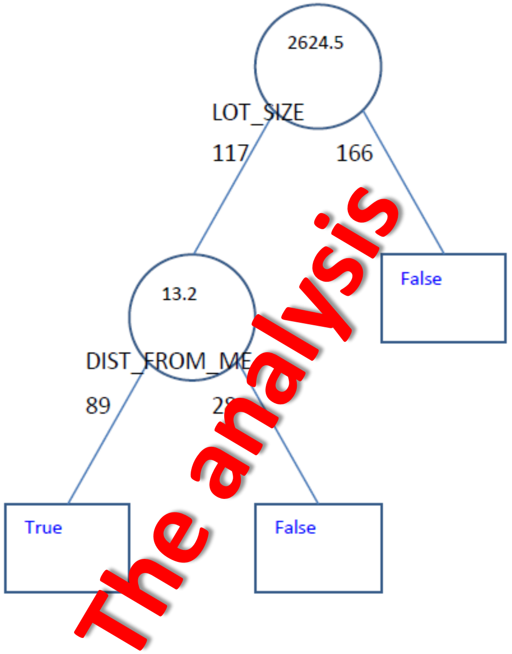
2. Predicting ZILLOW.com's Zestimate accuracy

- Data collected, cleansed and merged from 4 sources –Zillow , Redfin, School Digger and Google Maps
- 17 counties (29 Zip codes) in Northern VA

House sales data

- Before Data Clean up: **3500+**
- After Data Clean up: **1416**
- Y –***Is Zestimate correct*** (Y/N) 37.6%/62.43%
- X –15 variables (5+ variables where discarded from initial set)





Three case studies (2)

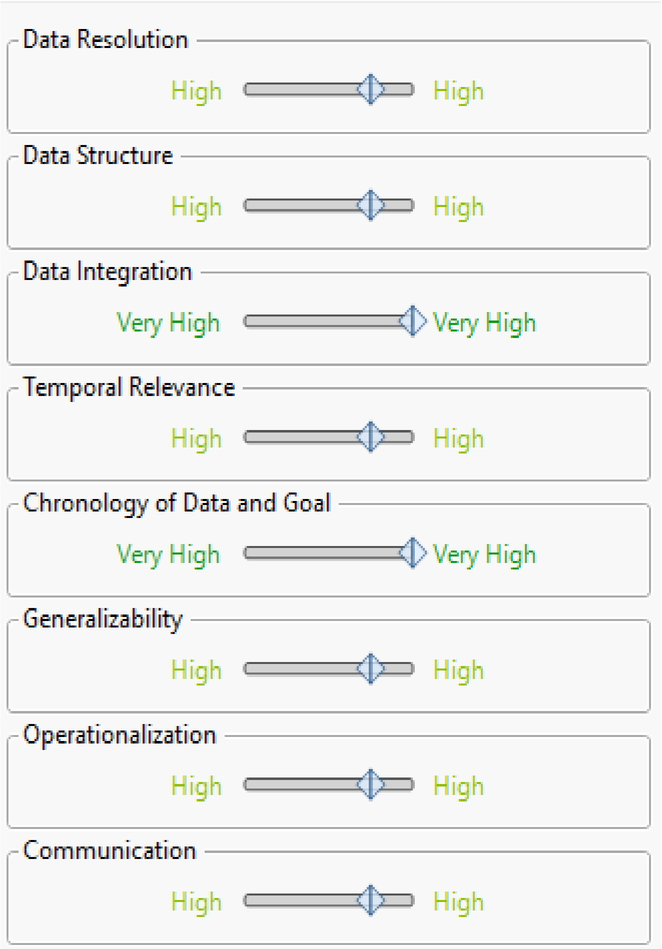
2. Predicting ZILLOW.com's Zestimate accuracy

Logistic Regression

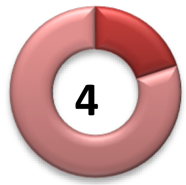
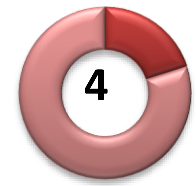
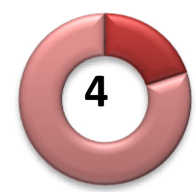
Input variables	Coefficient
Constant term	-4.65478611
BATHROOM_REV	0.38922957
LOG(SQFT)	0.2396526
log(LOT_SIZE)	0.38037464
TOTALROOMS_REV	-0.19049983
Age_of_house_at_Sale	0.01936915
Binned_PrimarySchoolRank	0.0735151
Binned_MiddleSchoolRank	-0.09299159
Binned_HighSchoolRank	0.04271848

Class	# Cases	# Errors	% Error
FALSE	184	22	11.96
TRUE	99	71	71.72
Overall	283	93	32.86

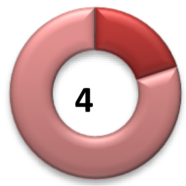
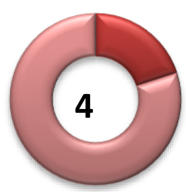
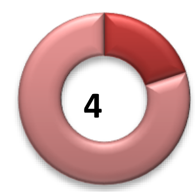
InfoQ=81%



Three case studies (2)

Data Resolution  Appropriate scale used	**Data Structure**  No significant gaps in the data coverage
Data integration  Data from different sources and formats were merged to get a more robust and complete data set	**Temporal Relevance**  Data used span the boom and bust periods of the housing market, but may not reflect truly the normal market scenario

Three case studies (2)

Generalizability  It can be generalized to other Northern Virginia States but probably not to other parts of the US or the world at large	**Chronology of data and goal**  Analysis and recommendations are available now and those interest in buying or selling a house can use them
Operationalization  Buyers and sellers can rely on the assessment but data used in the model needs to be updated periodically	**Communication**  Results duly published online but some advertisement about it will inform more prospective users of its availability

Three case studies (3)

3. Predicting First Day Returns for Japanese IPOs

Goal: To predict the First Day returns on Japanese IPOs (based on first day closing price), using public information available prior to the offer

The data: i) Japanese IPO data from 1997-2009*, ii) 1561 IPOs, iii) Industry(categorical) : 35 industries - 3 were spelling errors, corrected

Remove Air Trans (1), Fishery & Forestry (2) industries

–Removed first 128 entries (1997-1999) as they had no data for 2 columns : Underwriter's fees & Allocation to BRLM

–New Columns

Minimum bid size

Secondary Offering %age

–Creation of Dummy Variables

BRLMs – 3, on the basis of Gross proceeds of IPO

Industry – 4, binned by average return

Market – whether the IPO was OTC or not

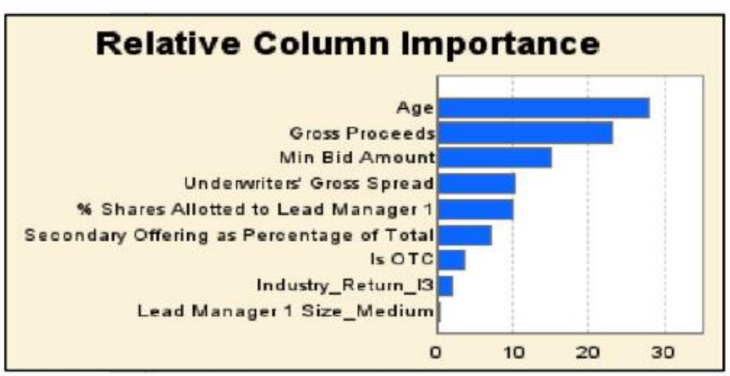
*Kaneko and Pettway's Japanese IPO Database (KP-JIPO)
http://www.fbc.keio.ac.jp/~kaneko/KP-JIPO/top.htm

Three case studies (3)

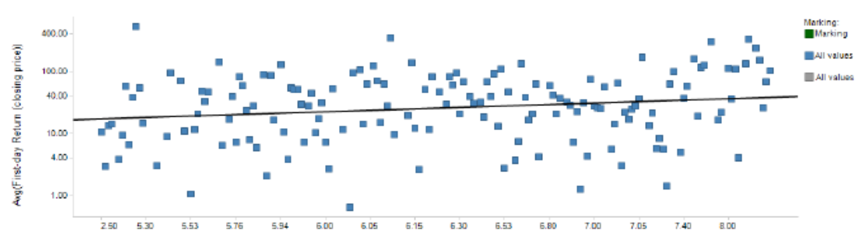
3. Predicting First Day Returns for Japanese IPOs

1) Age of company at time of IPO
2) Gross Proceeds (size of IPO)
3) Minimum Bid Amount
4) IS_OTC listing
5) Secondary offering as %age of total
5) Percentage shares allocated to Lead Manager 1
7) Underwriter's Gross Spread (fees as %age of size of IPO)
8) Industry_Type (binned categorical variable – 4 categories)
9) Lead_Manager (binned categorical variable – 3 categories)

InfoQ=51%



Prediction algorithms do not give a reasonable prediction of IPO returns from public information. (High RMSE: 90%)



Three case studies (3)

5

Data Resolution

data are suitable for the report goal and furthermore they decided to aggregate where it was possible to do it, i.e. industry, of the 33 industries in the raw data, they binned them 4 categories of industries representing the 4 types of patterns in the first day returns observed as a function of industry.

4

Data Structure

there were some missing data in the percentage of allocation to Lead Manager which was considered as an important predictor. In the data cleaning procedure it has been decided to remove them but, since they were a small percentage (128/1561≈8%), it does not affect in a critical way the all dataset.

3

Data integration

Obviously, the practice of integrating multiple sources usually creates new knowledge. The consequence of doing this it's the inflow of InfoQ. Anyway, here in this report there was no need to search for other data sources in order to solve any kind of integration problem. So data were taken from one single source. However, visiting the source of the database it is possible to understand that all the data were taken from different sources.

2

Temporal Relevance

the all data set could have been divided into two sub-sets: one data set collected during the two period after the financial crises of the 1997 and 2008, the other one collected during the economical growth right before the Great Recession.

Three case studies (3)

3

Chronology of data and goal

Since ours is a predictive model, we have to consider the temporal relation that links the input variables; we have high values of this parameter when the variables are available at the time of prediction. However also the endogeneity can occur when some variables are omitted from the dataset. We have to consider its effect on a predictive model that is different from an explanatory study, infact it can be increase the infoQ.

2

Generalizability

this report is based on a dataset corresponding to a determinate temporal range of observations, then the analysis method that we need to use does not bring any new theory that could be used to generalize. In this case the available sample represents the complete population to analyse without the addition of new data.

3,5

Operationalization

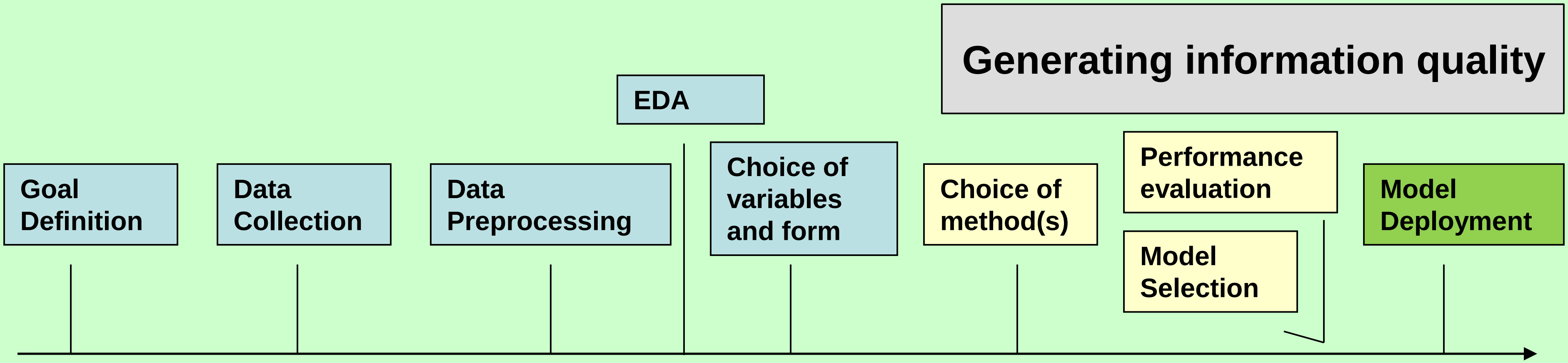
I gave 4 for the construct operationalization because in a predictive task the InfoQ relies on the quality of the data and here we have it. Besides, data are stable in the sense that further studies can make use of them for other purposes. However, since that with the action operationalization we want to assess if a report leads to clear follow-up actions from the information provided, I gave 3 to this sub-dimension because personally I am not lead to any follow-up actions but in any case I think that it could be possible to improve the entire analysis by focusing on a subgroup of data which is shorter in period of time and closer to the present.

4

Communication

if we read this report with the sufficient level of attention, we will understand that it gives the exact information we need in order to understand the conclusion it leads to, without unnecessary details. The data are represented in schematic way and the subdivision categories of variables are explained clearly. Anyway, I have not given 5 because there are some points in the report in which there is a need to study in deep what it is saying.

The Roadmap to Predictive Models



Predictive task

Action: Evaluate predictability; compare to existing models

Risks: Over-fitting; costs of prediction error

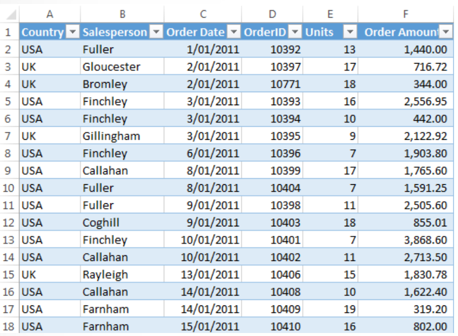
Supervised vs. Unsupervised Learning

- **Supervised learning:** discover patterns in the data that relate data attributes with a target (class) attribute.
  - These patterns are then utilized to predict the values of the target attribute in future data instances.
- **Unsupervised learning:** The data has no target attribute.
  - We want to explore the data to find some intrinsic structures in it.

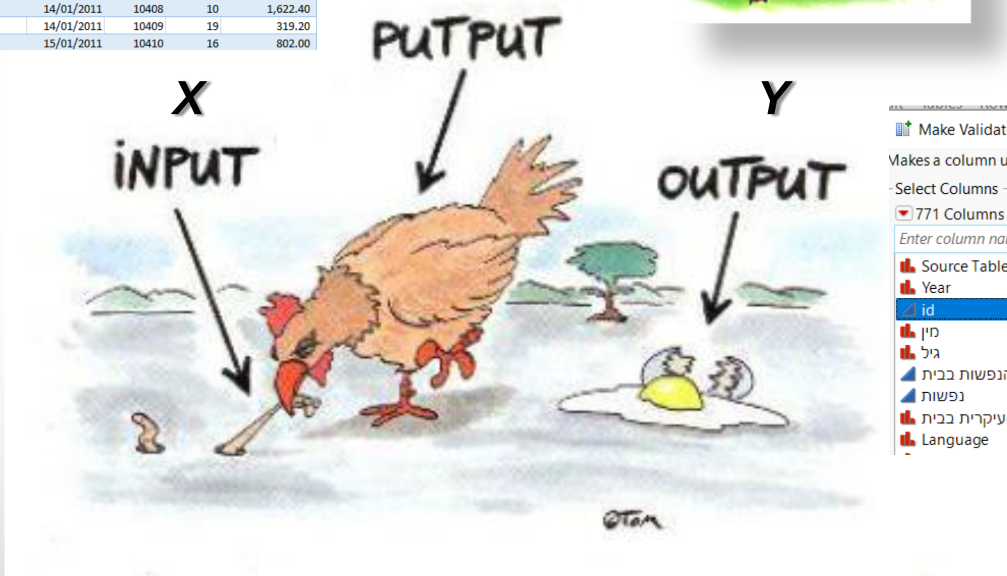
Supervised Learning

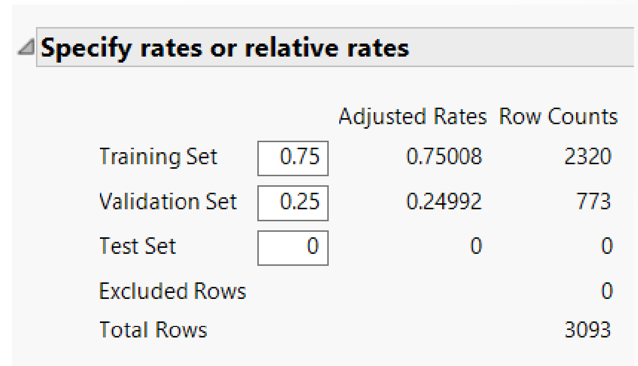
“0” Training data
“1” Validation data
“2” Testing data

Holdout set



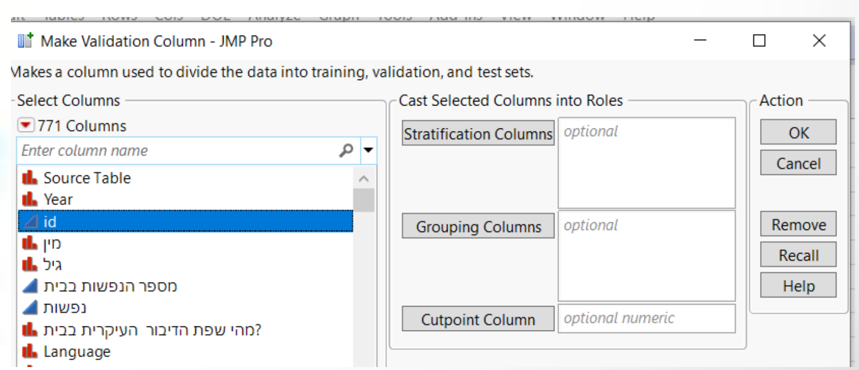
	A	B	C	D	E	F
1	Country	Salesperson	Order Date	OrderID	Units	Order Amoun
2	USA	Fuller	1/01/2011	10392	13	1,440.00
3	UK	Gloucester	2/01/2011	10397	17	716.72
4	UK	Bromley	2/01/2011	10771	18	344.00
5	USA	Finchley	3/01/2011	10393	16	2,556.95
6	USA	Finchley	3/01/2011	10394	10	442.00
7	UK	Gillingham	3/01/2011	10395	9	2,122.92
8	USA	Finchley	6/01/2011	10396	7	1,903.80
9	USA	Callahan	8/01/2011	10399	17	1,765.60
10	USA	Fuller	8/01/2011	10404	7	1,591.25
11	USA	Fuller	9/01/2011	10398	11	2,505.60
12	USA	Coghill	9/01/2011	10403	18	855.01
13	USA	Finchley	10/01/2011	10401	7	3,868.60
14	USA	Callahan	10/01/2011	10402	11	2,713.50
15	UK	Rayleigh	13/01/2011	10406	15	1,830.78
16	USA	Callahan	14/01/2011	10408	10	1,622.40
17	USA	Farnham	14/01/2011	10409	19	319.20
18	USA	Farnham	15/01/2011	10410	16	802.00





Specify rates or relative rates

		Adjusted Rates	Row Counts
Training Set	0.75	0.75008	2320
Validation Set	0.25	0.24992	773
Test Set	0	0	0
Excluded Rows			0
Total Rows			3093

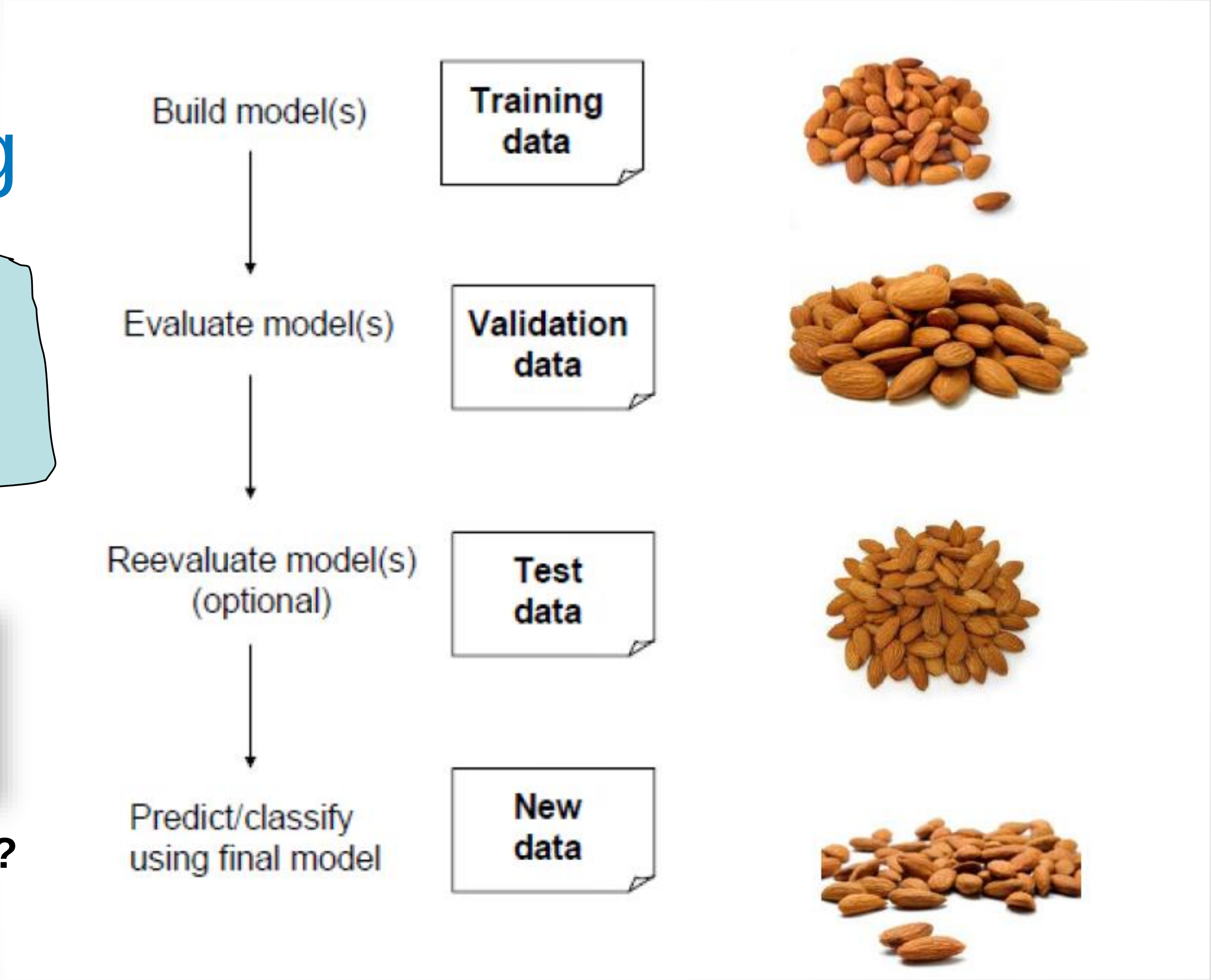


KPA

Data Partitioning

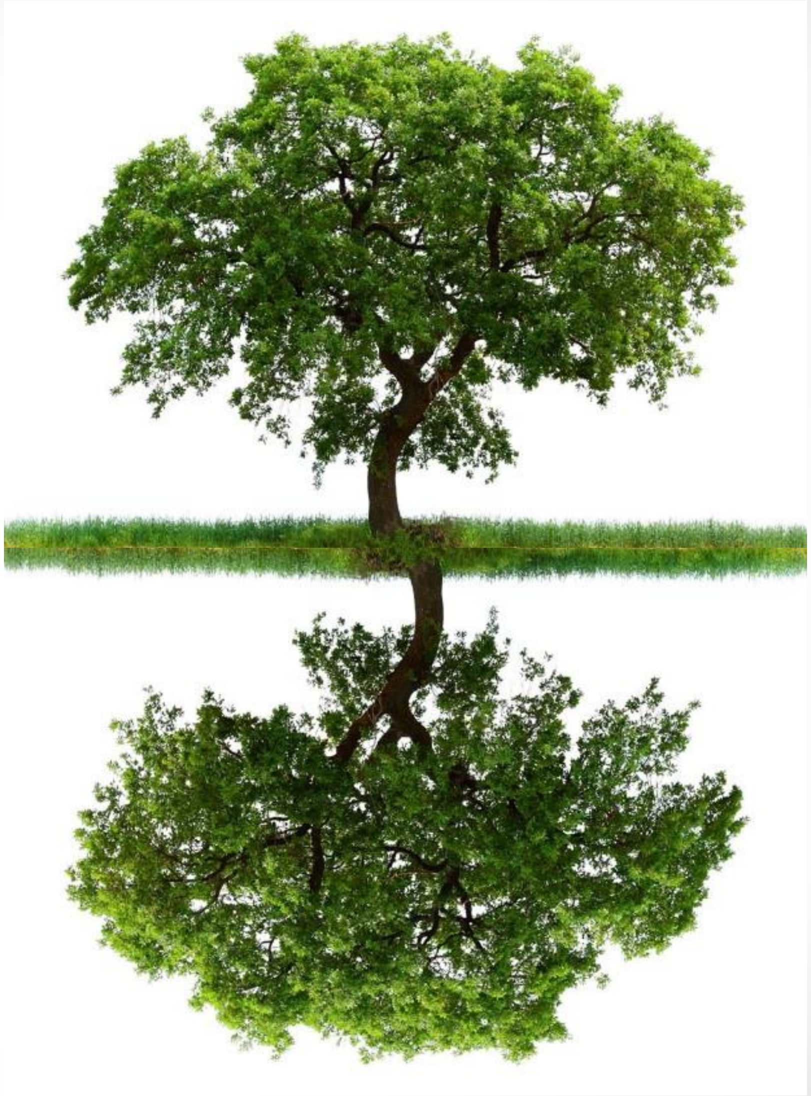
"0" Training data
"1" Validation data
"2" Testing data

What happens here?



Analytic Models

- **Decision trees**
- Regression trees
- Random forests
- Boosted trees
- Logistic regression
- Naïve Bayes
- K-Means Clustering



Decision Trees

Goal: Classify or predict an outcome based on a set of predictors

The output is a set of **rules** represented by tree diagrams

Key Ideas

Recursive partitioning: Repeatedly split the records into two subsets so as to achieve maximum homogeneity within the new subsets (or, equivalently, with the greatest dissimilarity between the subsets)

Pruning the tree: Simplify the tree by pruning peripheral branches to avoid overfitting

Recursive Partitioning Idea

- Pick one of the predictor variables, x_i
- Pick a value of x_i, say s_i, that divides the training data into two (not necessarily equal) portions
- Measure how dissimilar each of the resulting portions are
- Try different values of x_i, and s_i to maximize the dissimilarity in the initial split
- After the first split, repeat the process for a second split, and so on

The Riding Mowers

- **Goal**: Classify 24 households as owning or not owning riding mowers

- **Predictors**: Income, Lot Size

	Income	Lot_Size	Ownership
1	60	18.4	owner
2	85.5	16.8	owner
3	64.8	21.6	owner
4	61.5	20.8	owner
5	87	23.6	owner
6	110.1	19.2	owner
7	108	17.6	owner
8	82.8	22.4	owner
9	69	20	owner
10	93	20.8	owner
11	51	22	owner
12	81	20	owner
13	75	19.6	non-owner
14	52.8	20.8	non-owner
15	64.8	17.2	non-owner
16	43.2	20.4	non-owner
17	84	17.6	non-owner
18	49.2	17.6	non-owner
19	59.4	16	non-owner
20	66	18.4	non-owner
21	47.4	16.4	non-owner
22	33	18.8	non-owner
23	51	14	non-owner
24	63	14.8	non-owner

Splitting on Categorical Variables

- Examine all possible ways in which the categories can be split.
- E.g., nominal categories A, B, C can be split 3 ways

  {A} and {B, C}

  {B} and {A, C}

  {C} and {A, B}

- With many categories, # of potential splits becomes huge

Splitting on Categorical Variables

- For ordinal data (ordered categories) there is an option for the splits to respect ordering
- Example: An ordinal predictor takes on the values 1, 2, 3, or 4
- The data can be split 3 ways:

{1} and {2, 3, 4}

{1, 2} and {3, 4}

{1, 2, 3} and {4}

Splitting on Continuous Variables

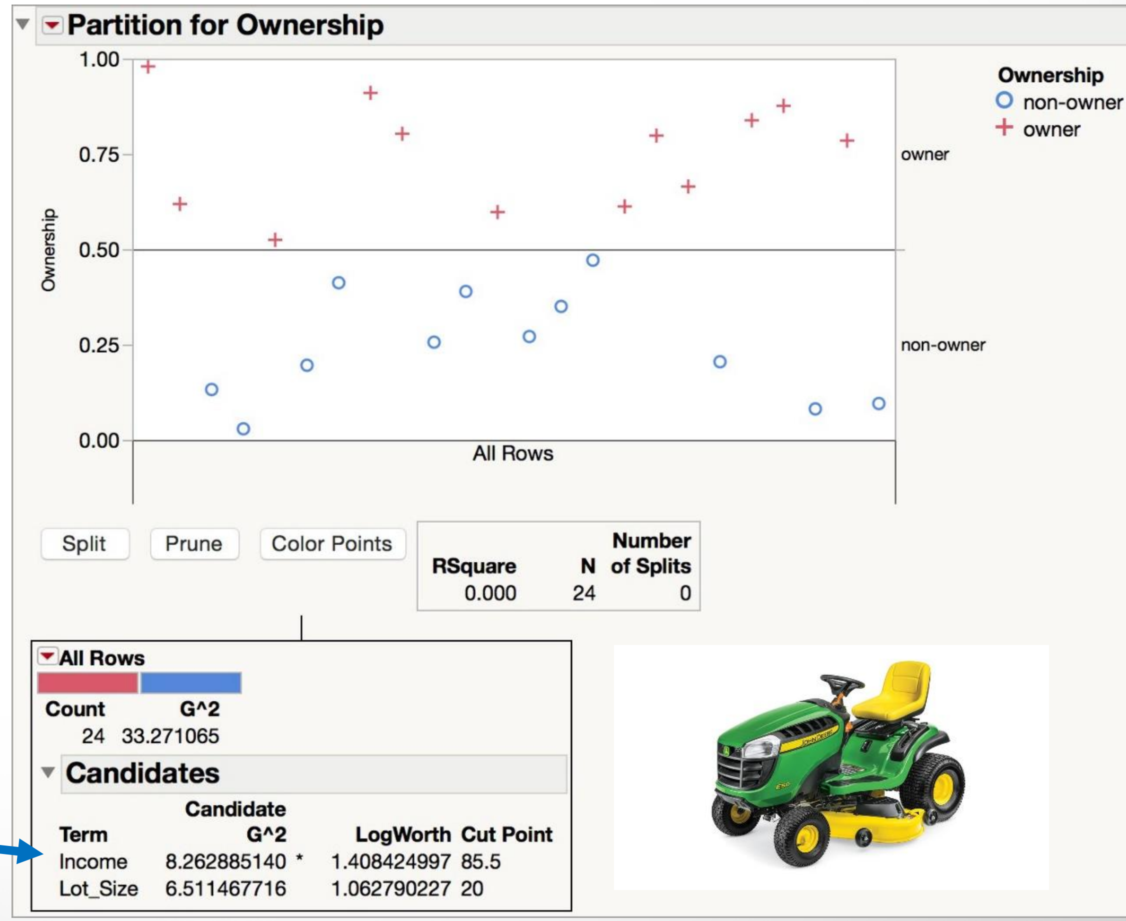
- Order records according to one variable, say lot size
- Split at the first value
- Measure the dissimilarity between the two subsets
- Split at the next value, and continue
- Repeat for the other variable(s)
- For all variables, the split value that drives the greatest dissimilarity in propensities (or probabilities) is selected as the split point

The Riding Mowers

Before splitting (50% are owners and 50% are non-owners)

All splits are considered (see Candidates)

The first split variable is Income, and the cut point is 85.5



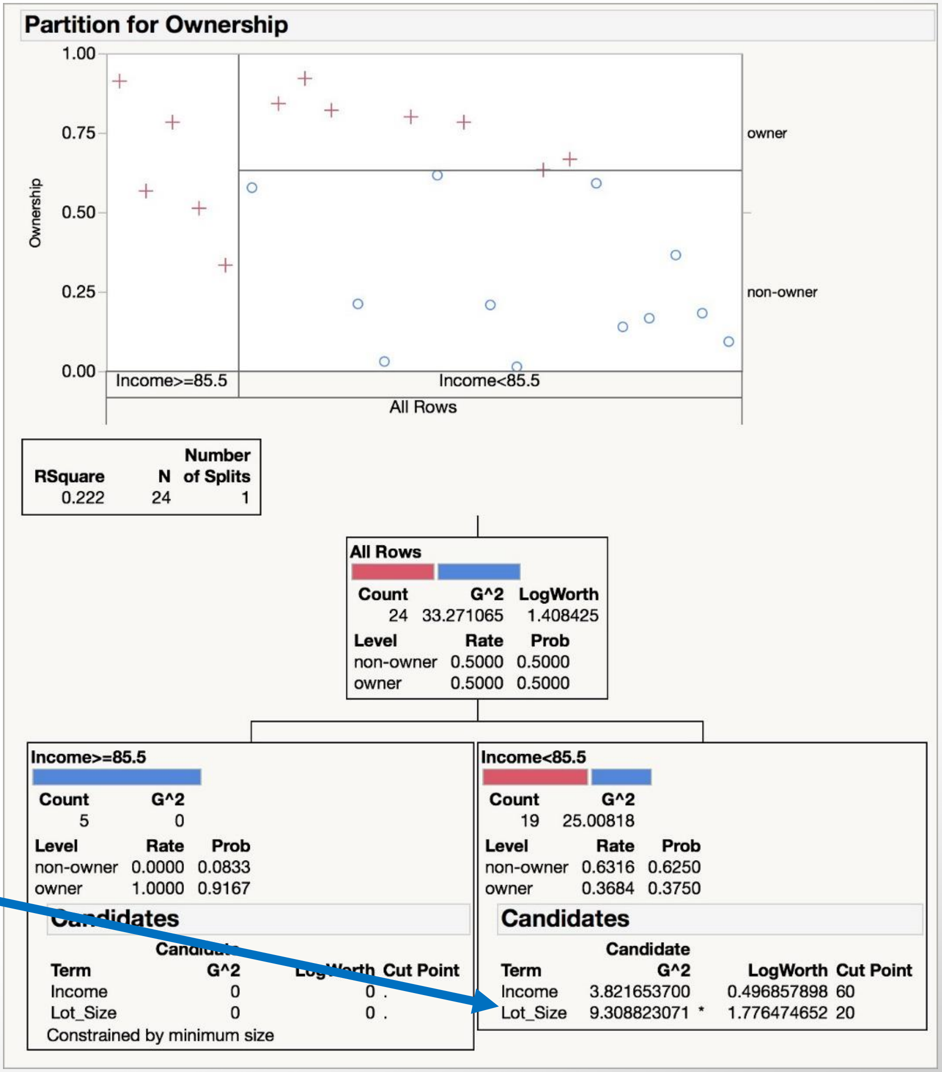
KPA

The Riding Mowers

When Income >= 85.5, all of the households were Owners (this “node” is “pure”).

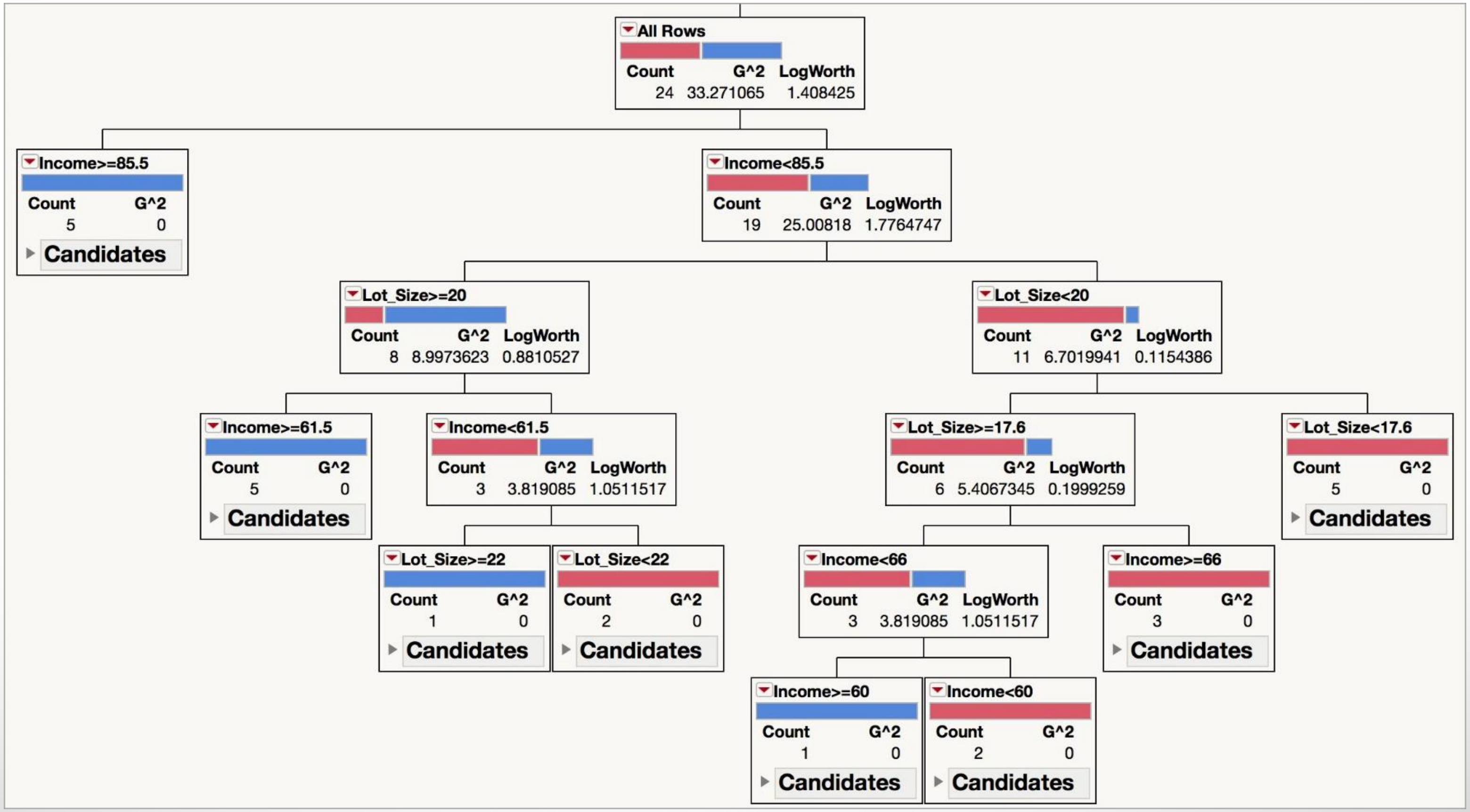
The next split is Lot Size when Income < 85.5.

The cut point is 20.



The Riding Mowers

The final tree after 7 splits (probabilities are hidden)



Tree Structure

- Split points become nodes on the tree
- Leaves are the terminal nodes (there are no further splits)
- Read down tree to derive the decision rule

  E.g., Income < 85.5, Lot Size is >= 20, and Income >=61.5 , the probability that a household is an owner is 0.9185.

- Records within each node are from the training data (validation data are not used in building the tree)
- Default cutoff = 0.5 is used for classification

  In the previous example, the record would be classified as an owner.

The Riding Mowers

The leaf report provides a summary the splits

It displays the rules for classifying outcomes

For example, If Income < 85.5, Lot Size is < 17.6, the probability that a household is an owner is 0.0752. This record will be classified as a non-owner.

Leaf Report

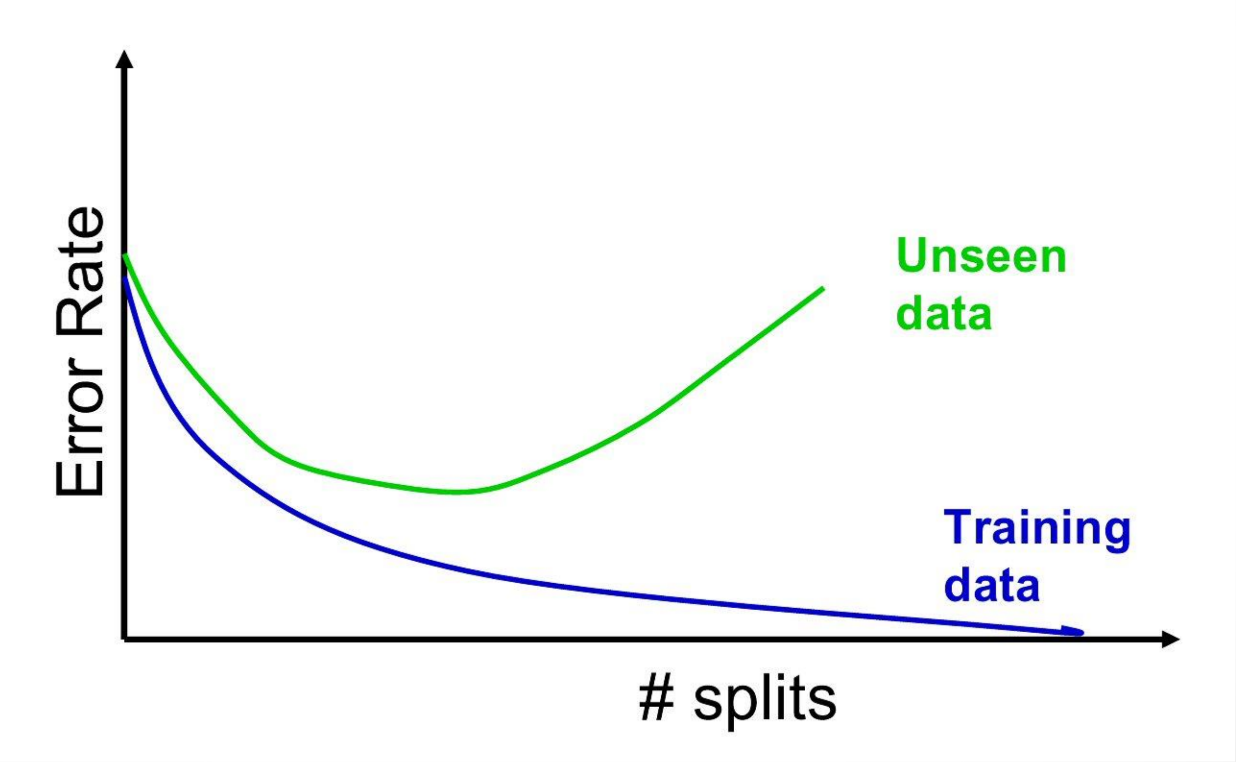
Response Prob

Leaf Label	non-owner	.2 .4 .6 .8	owner	
Income>=85.5	0.0833		0.9167	
Income<85.5&Lot_Size>=20&Income>=61.5	0.0815		0.9185	
Income<85.5&Lot_Size>=20&Income<61.5&Lot_Size>=22	0.2512		0.7488	
Income<85.5&Lot_Size>=20&Income<61.5&Lot_Size<22	0.8342		0.1658	
Income<85.5&Lot_Size<20&Lot_Size>=17.6&Income<66&Income>=60	0.2901		0.7099	
Income<85.5&Lot_Size<20&Lot_Size>=17.6&Income<66&Income<60	0.8601		0.1399	
Income<85.5&Lot_Size<20&Lot_Size>=17.6&Income>=66	0.8933		0.1067	
Income<85.5&Lot_Size<20&Lot_Size<17.6	0.9248		0.0752	

Stopping Tree Growth

- Natural end of process is 100% purity in each leaf
- This **overfits** the data, which end up fitting noise in the data
- Overfitting leads to low predictive accuracy of new data
- Past a certain point, the error rate for the validation data starts to increase

Full Tree Error Rate



CART - Classification and regression trees

- CART lets tree grow to full extent, then prunes it back
- Idea is to find that point at which the validation error begins to rise
- Generate successively smaller trees by pruning leaves
- At each pruning stage, multiple trees are possible
- Use *cost complexity* to choose the best tree at that stage

Cost Complexity

$$CC(T) = Err(T) + \alpha L(T)$$

CC(T) = cost complexity of a tree

Err(T) = proportion of misclassified records

L(T) – size of tree

α = penalty factor attached to tree size (set by user)

Among trees of given size, choose the one with lowest CC

Do this for each size of tree

CART - Classification and regression trees

- Nonparametric (no probabilistic assumptions)
- Automatically performs variable selection
- Uses any combination of continuous/discrete variables
  - Very nice feature: ability to automatically bin massively categorical variables into a few categories (zip code, business class, make/model…)
- Invariant to monotonic transformations of predictive variable
- Unlike regression, not sensitive to outliers in predictive variables

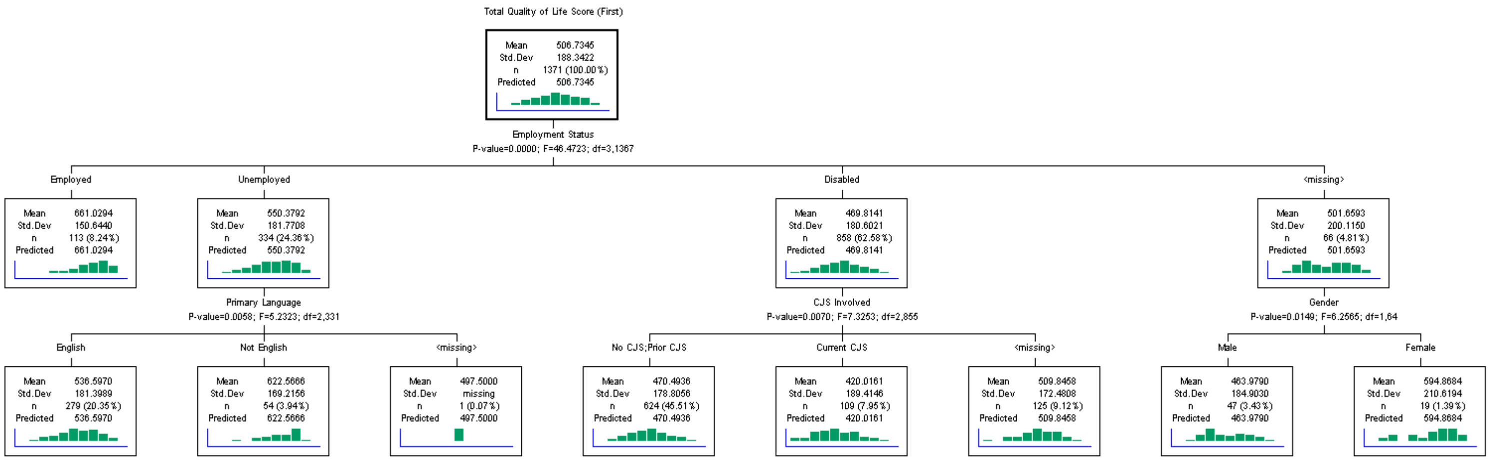
CART overview

- Classification and Regression Trees are an easily understandable and transparent method for predicting or classifying new records
- A tree is a graphical representation of a set of rules
- Trees must be pruned to avoid over-fitting of the training data
- As trees do not make any assumptions about the data structure, they usually require large samples

CHAID - Chi-squared automatic interaction detector

- CHAID, older than CART, uses chi-square statistical test to limit tree growth
- Splitting stops when purity improvement is not statistically significant

CHAID - Chi-squared automatic interaction detector



- CHAID is a non-binary decision tree.
- The decision or split made at each node is still based on a single variable, but can result in multiple branches.
- The split search algorithm is designed for categorical variables.

Classification Trees: CART versus CHAID

At each split, the CHAID algorithm looks for the predictor variable that if split, most "explains" the category response variable. In order to decide whether to create a particular split based on this variable, the CHAID algorithm tests a hypothesis regarding dependence between the split variable and the categorical response (using the chi-squared test for independence). Using a pre-specified significance level, if the test shows that the split variable and the response are independent, the algorithm stops the tree growth. Otherwise the split is created, and the next best split is searched. In contrast, the CART algorithm decides on a split based on the amount of homogeneity within class that is achieved by the split. The split is reconsidered based on considerations of over-fitting.

CHAID is most useful for **analysis**, whereas CART is more suitable for **prediction**. In other words, CHAID should be used when the goal is to describe or understand the relationship between a response variable and a set of explanatory variables, whereas CART is better suited for creating a model that has high prediction accuracy of new cases.

How JMP limits tree size

JMP uses a combination of limiting tree growth and pruning the tree after it has grown

- **Minimum Split Size**: Controls the minimum number of records in terminal nodes
- **Validation**: The tree is grown, and pruned back to maximize the RSquare on the validation data

When validation is used, the “Go” option automates tree growth and pruning

The tree with the maximum Validation Rsquare has 8 splits

The tree is grown to 18 splits, and is pruned back to 8 splits

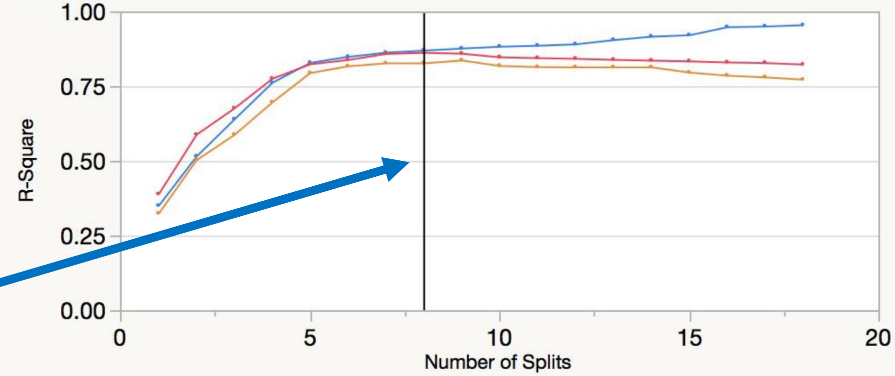
Validation error rate and confusion matrix for the final tree (cutoff for classification = 0.50)

Partition for Personal Loan

Split | Prune | Go | Color Points

	RSquare	N	Number of Splits
Training	0.872	2500	8
Validation	0.864	1500	
Test	0.829	1000	

Split History



Validation Data in Red
Test Data in Orange

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.8715	0.8642	0.8289	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.9041	0.8984	0.8705	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.0406	0.0429	0.0541	∑ -Log(ρ[j])/n
RMSE	0.1107	0.1095	0.1245	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.0252	0.0235	0.0263	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0172	0.0127	0.0190	∑ (ρ[j]≠ρMax)/n
N	2500	1500	1000	n

Confusion Matrix

Training

Actual Personal Loan	Predicted 0	Predicted 1
0	2247	13
1	30	210

Validation

Actual Personal Loan	Predicted 0	Predicted 1
0	1355	1
1	18	126

Test

Actual Personal Loan	Predicted 0	Predicted 1
0	902	2
1	17	79

Leaf Report

Leaf Report

Response Prob

Leaf Label	0	1
Income>=99&Education(2, 3)&Income>=118	0.0051	0.9949
Income>=99&Education(2, 3)&Income<118&CCAvg>=2.9	0.4605	0.5395
Income>=99&Education(2, 3)&Income<118&CCAvg<2.9	0.9048	0.0952
Income>=99&Education(1)&Family>=3&Income>=119	0.0226	0.9774
Income>=99&Education(1)&Family>=3&Income<119	0.6950	0.3050
Income>=99&Education(1)&Family<3	0.9966	0.0034
Income<99&CCAvg>=3&Income>=82	0.7205	0.2795
Income<99&CCAvg>=3&Income<82	0.9447	0.0553
Income<99&CCAvg<3	1.0000	0.0000

Response Counts

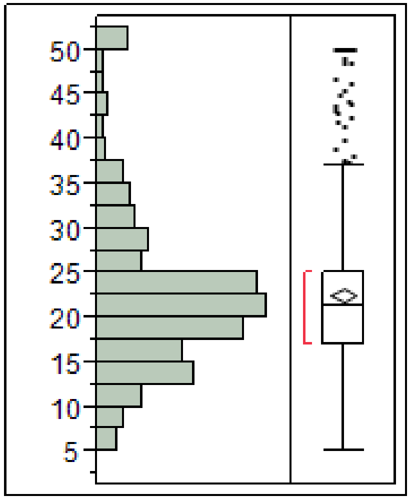
Leaf Label	0	1
Income>=99&Education(2, 3)&Income>=118	0	159
Income>=99&Education(2, 3)&Income<118&CCAvg>=2.9	13	16
Income>=99&Education(2, 3)&Income<118&CCAvg<2.9	58	6
Income>=99&Education(1)&Family>=3&Income>=119	0	35
Income>=99&Education(1)&Family>=3&Income<119	11	5
Income>=99&Education(1)&Family<3	331	1
Income<99&CCAvg>=3&Income>=82	38	15
Income<99&CCAvg>=3&Income<82	52	3
Income<99&CCAvg<3	1757	0

Regression Trees for Prediction

- Used with continuous outcome variable
- Procedure similar to classification tree
- Many splits attempted, choose the one that maximizes the difference between subgroup means
- Difference measured as the sum of squared deviations
- Prediction is the **average** of the numerical target variable (rather than a probability)

Boston Housing Data

mvalue



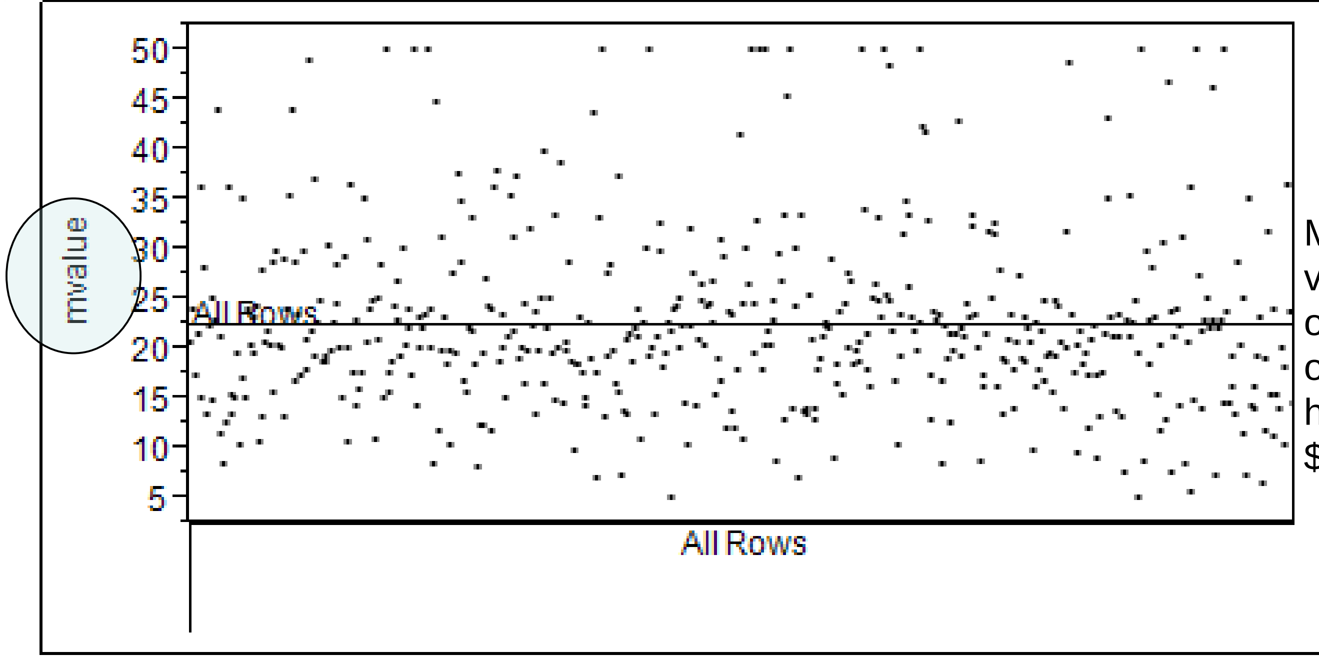
Quantiles

100.0%	maximum	50
99.5%		50
97.5%		50
90.0%		34.9
75.0%	quartile	25
50.0%	median	21.2
25.0%	quartile	16.95
10.0%		12.7
2.5%		8.235
0.5%		5.321
0.0%	minimum	5

Summary Statistics

Mean	22.532806
Std Dev	9.1971041
Std Err Mean	0.4088611
Upper 95% Mean	23.336085
Lower 95% Mean	21.729528
N	506

Partition for mvalue



Median value of owner-occupied homes in $1000

RSquare	RMSE	N	Number of Splits	AICc
0.000	.	506	0	0

All Rows

Count	506
Mean	22.532806
Std Dev	9.1971041

Boston Housing Data

All Rows

Count	506
Mean	22.532806
Std Dev	9.1971041

Candidates

Term	Candidate SS		LogWorth
crim	8266.17273		32.6638216
zn	6669.06251		24.9773486
indus	11083.22547		48.7519537
chas	1312.07927		4.1110954
nox	9536.22405		39.5670978
rooms	19339.55503	*	118.7473483
age	5573.64765		19.6751451
distance	4994.54054		17.1453361
radial	6708.64333		24.6205659
tax	8618.08428		34.5266980
pt	10438.69478		44.8775094
b	5259.31980		18.2910466
lstat	18896.19401		113.7427626

All Rows

		LogWorth	Difference
Count	506	118.74735	17.3044
Mean	22.532806		
Std Dev	9.1971041		

rooms<6.943

Count	430
Mean	19.933721
Std Dev	6.3534806

Candidates

Term	Candidate SS		LogWorth
crim	4300.967311		38.57528016
zn	1961.912781		13.93948488
indus	3552.756728		29.65539469
chas	533.165511		3.56806955
nox	4806.344267		45.22939006
rooms	2498.676569		18.68959899
age	3618.341104		30.39395326
distance	3526.248005		29.35482815
radial	2778.264622		21.29849865
tax	3487.174824		28.92548472
pt	3808.647013		32.66254455
b	2454.655577		18.26837433
lstat	7311.852356	*	88.35256425

rooms>=6.943

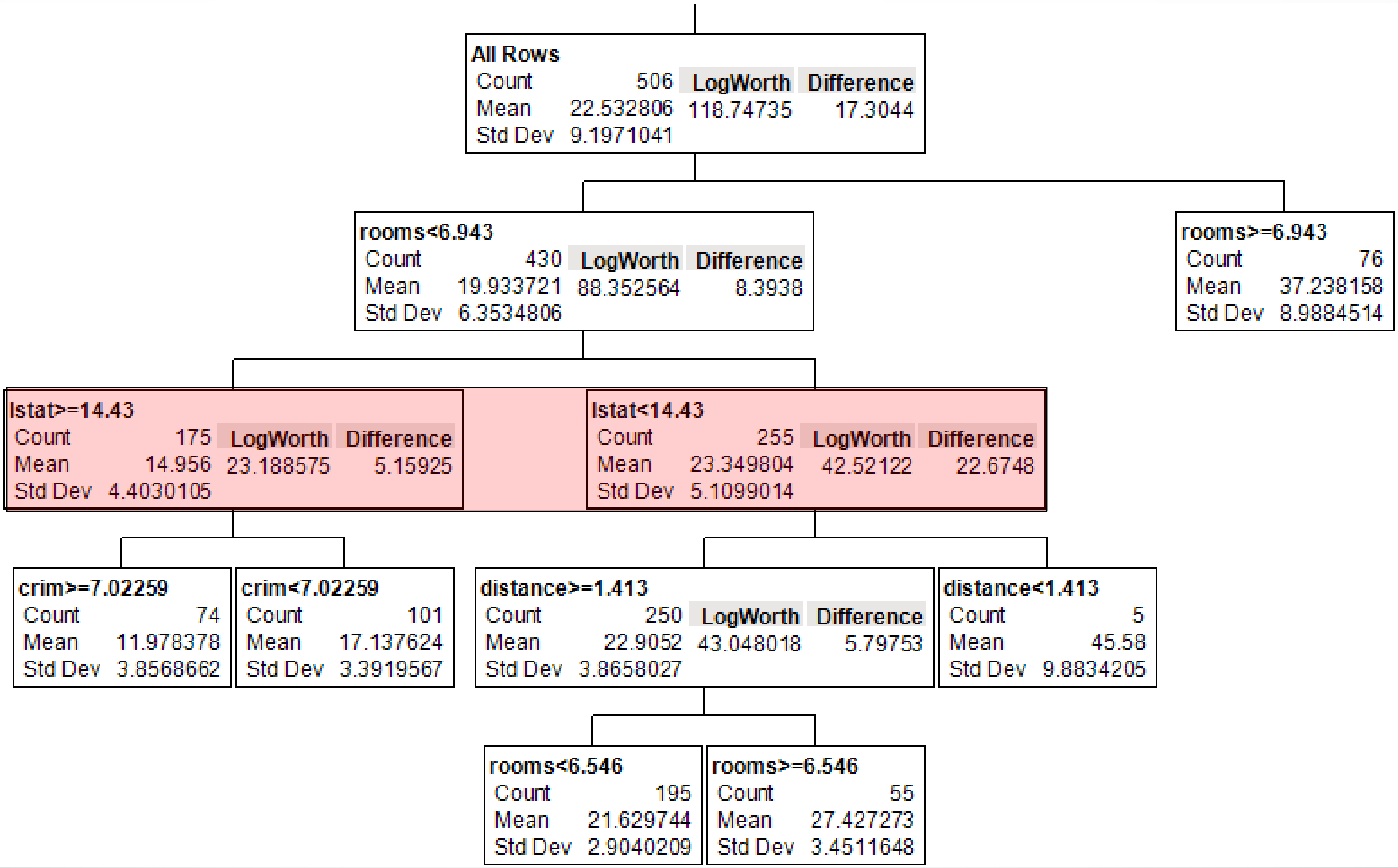
Count	76
Mean	37.238158
Std Dev	8.9884514

Candidates

Term	Candidate SS		LogWorth
crim	1296.353462		4.24150833
zn	154.894267		0.16015922
indus	650.180018		1.45829879
chas	97.802924		0.53155728
nox	510.976998		0.97911866
rooms	3060.957502	*	19.65116632
age	106.820174		0.05293436
distance	210.835800		0.20608146
radial	1296.353462		4.68218182
tax	1296.353462		4.30278667
pt	1514.119195		5.52903675
b	750.759998		1.79989185
lstat	2011.069265		8.73682304

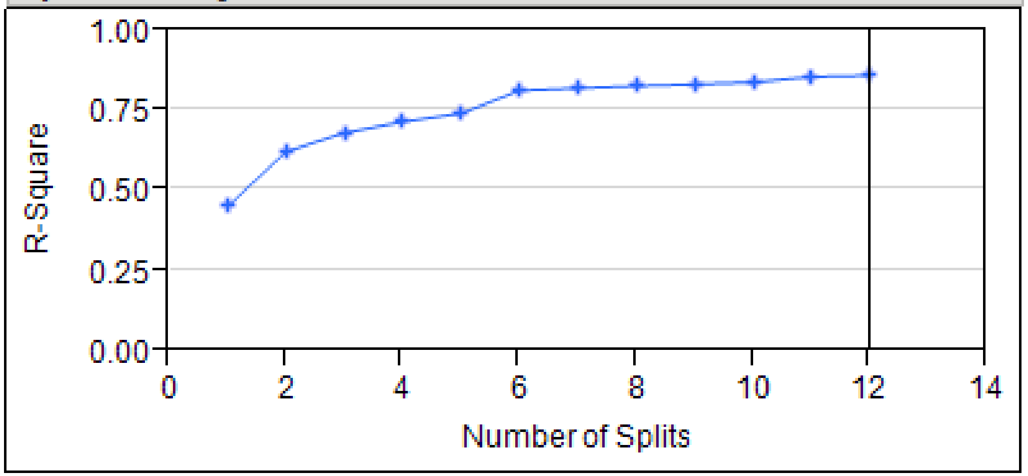
$-\log_{10}(p\text{-value})$

Boston Housing Data



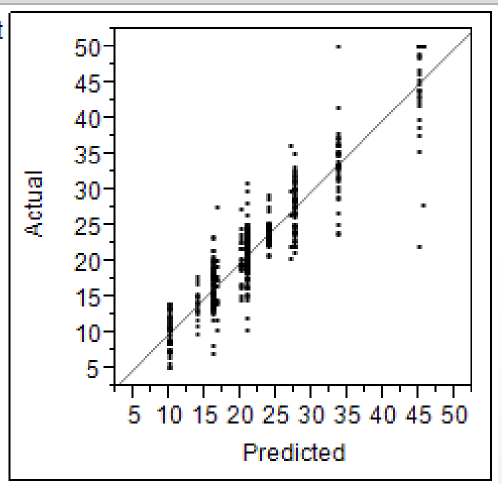
Boston Housing Data

Split History



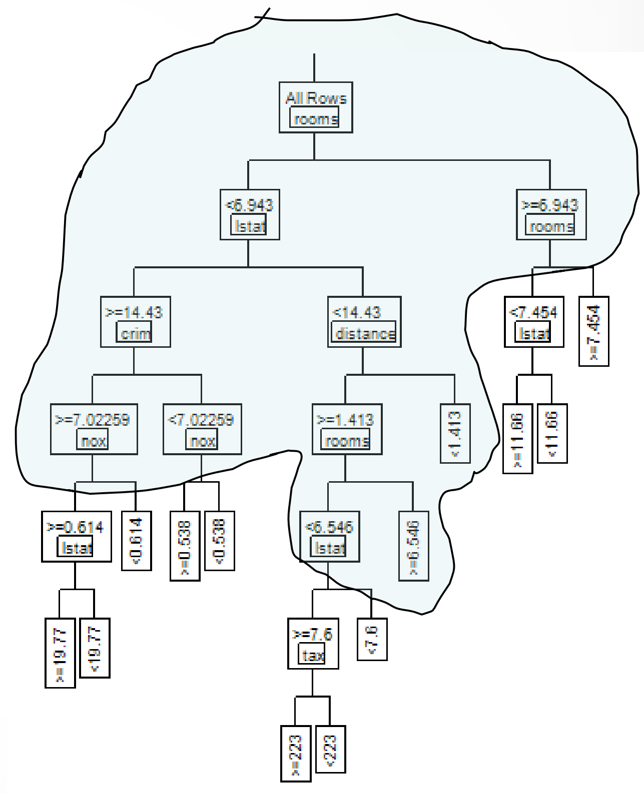
Actual by Predicted Plot

Training Set



Column Contributions

Term	Number of Splits	SS
crim	1	1136.809
zn	0	0.000
indus	0	0.000
chas	0	0.000
nox	2	572.045
rooms	3	23842.439
age	0	0.000
distance	1	2520.326
radial	0	0.000
tax	1	181.942
pt	0	0.000
b	0	0.000
lstat	4	8544.783

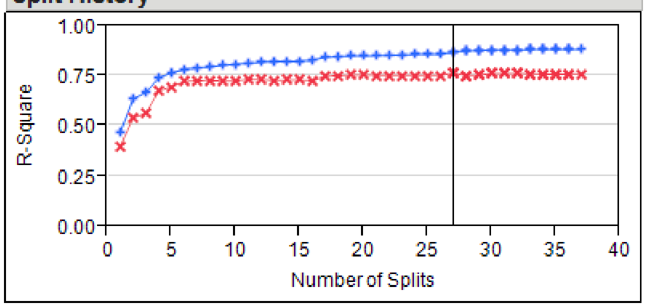


Boston Housing Data

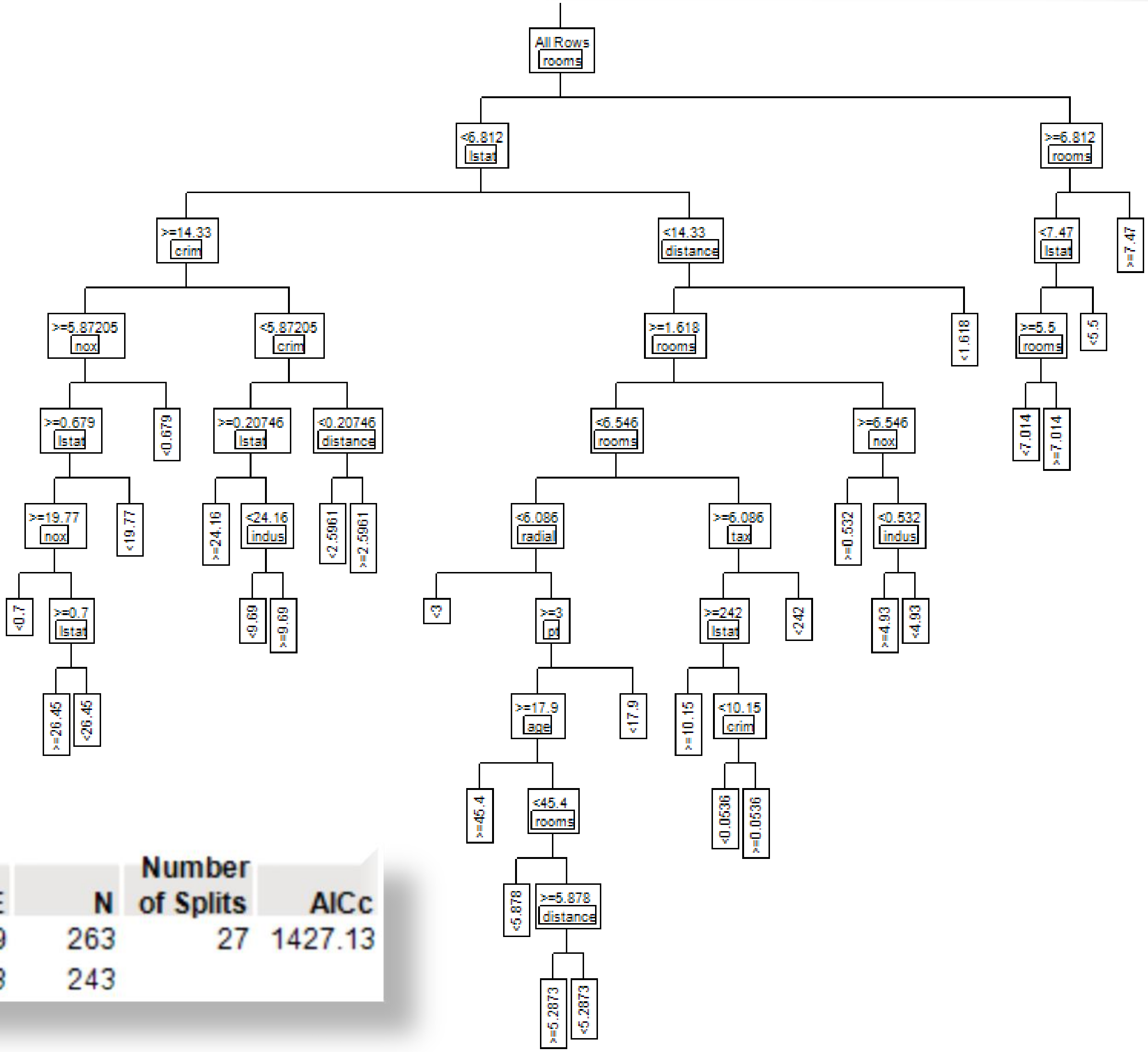
50% validation data with automatic splitting

Term	Number of Splits	SS
crim	3	924.046
zn	0	0.000
indus	2	86.329
chas	0	0.000
nox	3	265.824
rooms	6	12285.998
age	1	50.482
distance	3	642.560
radial	1	39.724
tax	1	68.077
pt	1	40.686
b	0	0.000
lstat	6	4131.903

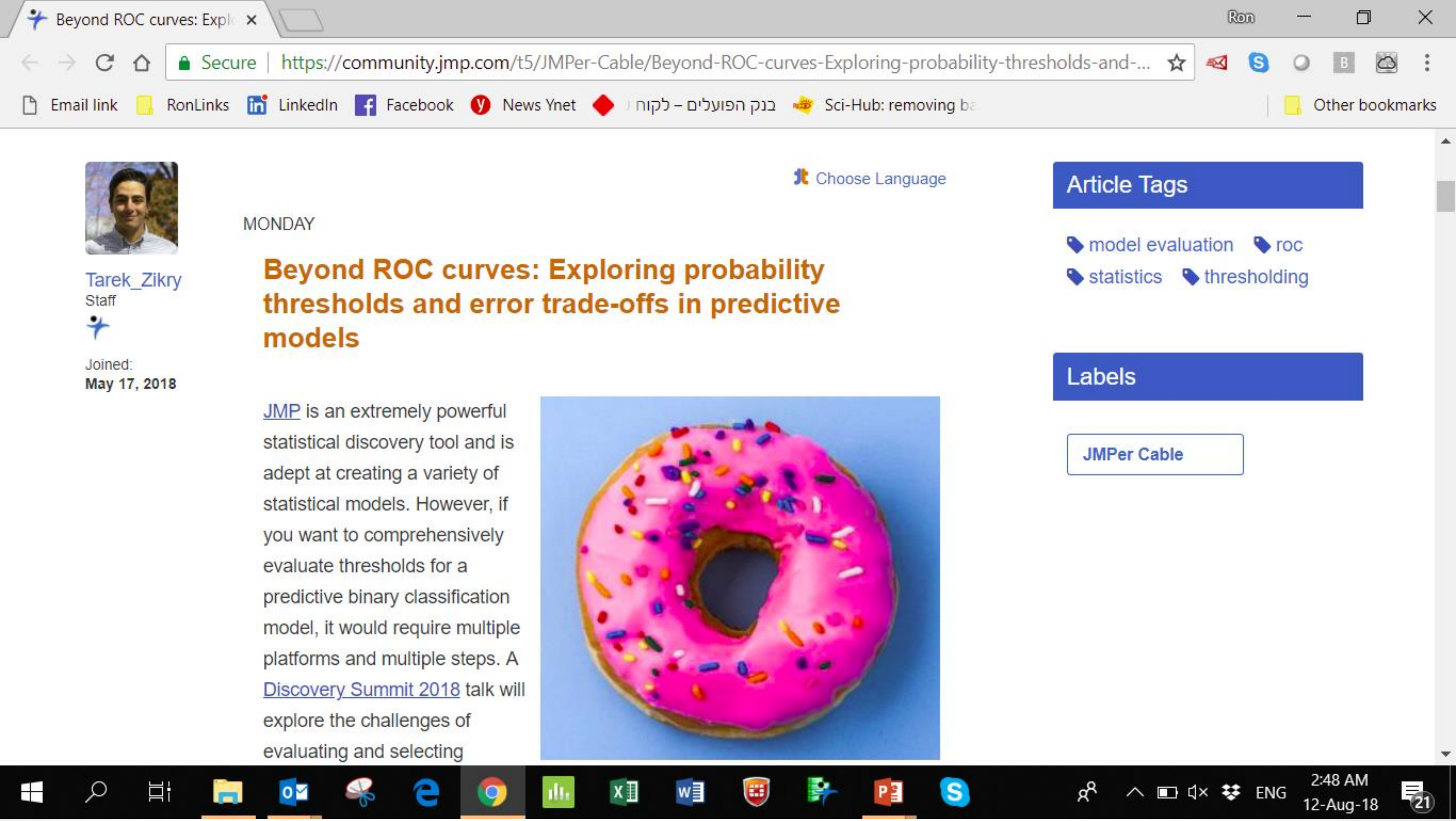
Split History



Validation Data in Red



	RSquare	RMSE	N	Number of Splits	AICc
Training	0.872	3.2212389	263	27	1427.13
Validation	0.769	4.5128963	243		



Donuts Blog Example.jmp

Model Classification Explorer.jmpaddin

Advantages of Trees

- Easy to use, understand
- Produce rules that are easy to interpret & implement
- Variable selection & reduction is automatic
- Do not require the assumptions of statistical models
- Can work without extensive handling of missing data (this is an option in the Partition dialog in JMP)

Disadvantages of Trees

- May not perform well where there is structure in the data that is not well captured by horizontal or vertical splits
- Since the process deals with one variable at a time, no way to capture interactions between variables

Improving Trees

- Single trees may not have good predictive ability.
- Results from multiple trees can be combined to improve performance
- The resulting model is an “ensemble” model
- Two multi-tree approaches in JMP Pro:
  - **Bootstrap Forests** (a variant of Random Forests)
  - **Boosted Trees**

Ensemble Tree Methods

- *Bootstrap aggregation*, or *bagging*, is a general-purpose procedure for reducing the variance of a statistical learning method; we introduce it here because it is particularly useful and frequently used in the context of decision trees.

- Recall that given a set of n independent observations $Z_1, \ldots, Z_n$, each with variance σ^2, the variance of the mean $\bar{Z}$ of the observations is given by σ^2/n.

- In other words, *averaging a set of observations reduces variance*. Of course, this is not practical because we generally do not have access to multiple training sets.

Ensemble Tree Methods

- Instead, we can bootstrap, by taking repeated samples from the (single) training data set.
- In this approach we generate B different bootstrapped training data sets. We then train our method on the bth bootstrapped training set in order to get $\hat{f}^{*b}(x)$, the prediction at a point x. We then average all the predictions to obtain

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x).$$

This is called *bagging*.

- *Random forests* provide an improvement over bagged trees by way of a small tweak that *decorrelates* the trees. This reduces the variance when we average the trees.
- As in bagging, we build a number of decision trees on bootstrapped training samples.
- But when building these decision trees, each time a split in a tree is considered, *a random selection of m predictors* is chosen as split candidates from the full set of p predictors. The split is allowed to use only one of those m predictors.
- A fresh selection of m predictors is taken at each split, and typically we choose $m \approx \sqrt{p}$ — that is, the number of predictors considered at each split is approximately equal to the square root of the total number of predictors

Ensemble Tree Methods

Bootstrap Forests

1. A random sample is drawn with replacement from the data set (bootstrapping)
2. Predictors are randomly drawn from the candidate list of predictors
3. A small tree is fit (a “weak learner”)
4. The process is repeated
5. The final model is the average of all of the trees, producing a “Bootstrap aggregated” (or “bagged) model

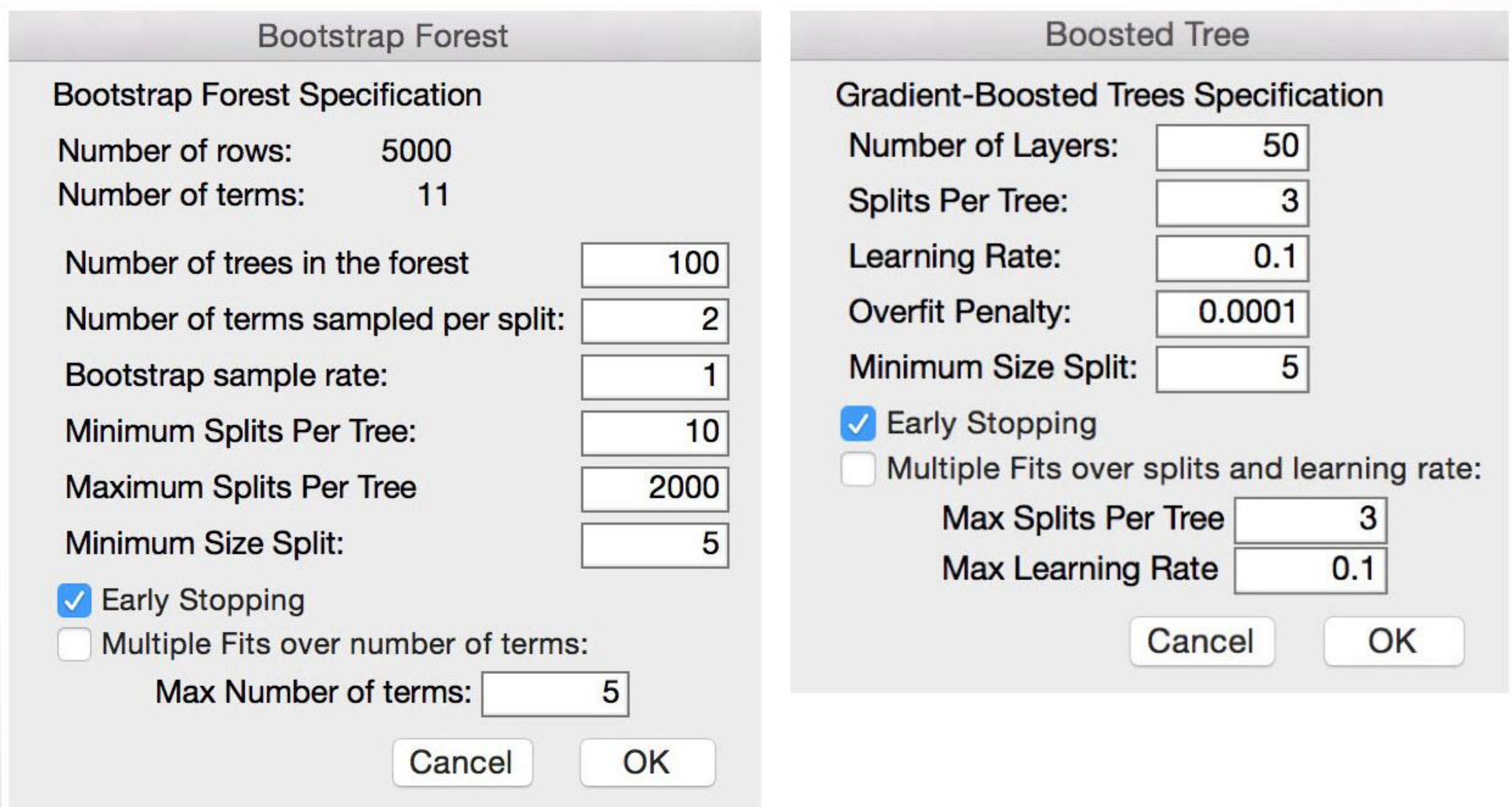
Ensemble Tree Methods

Boosted Trees

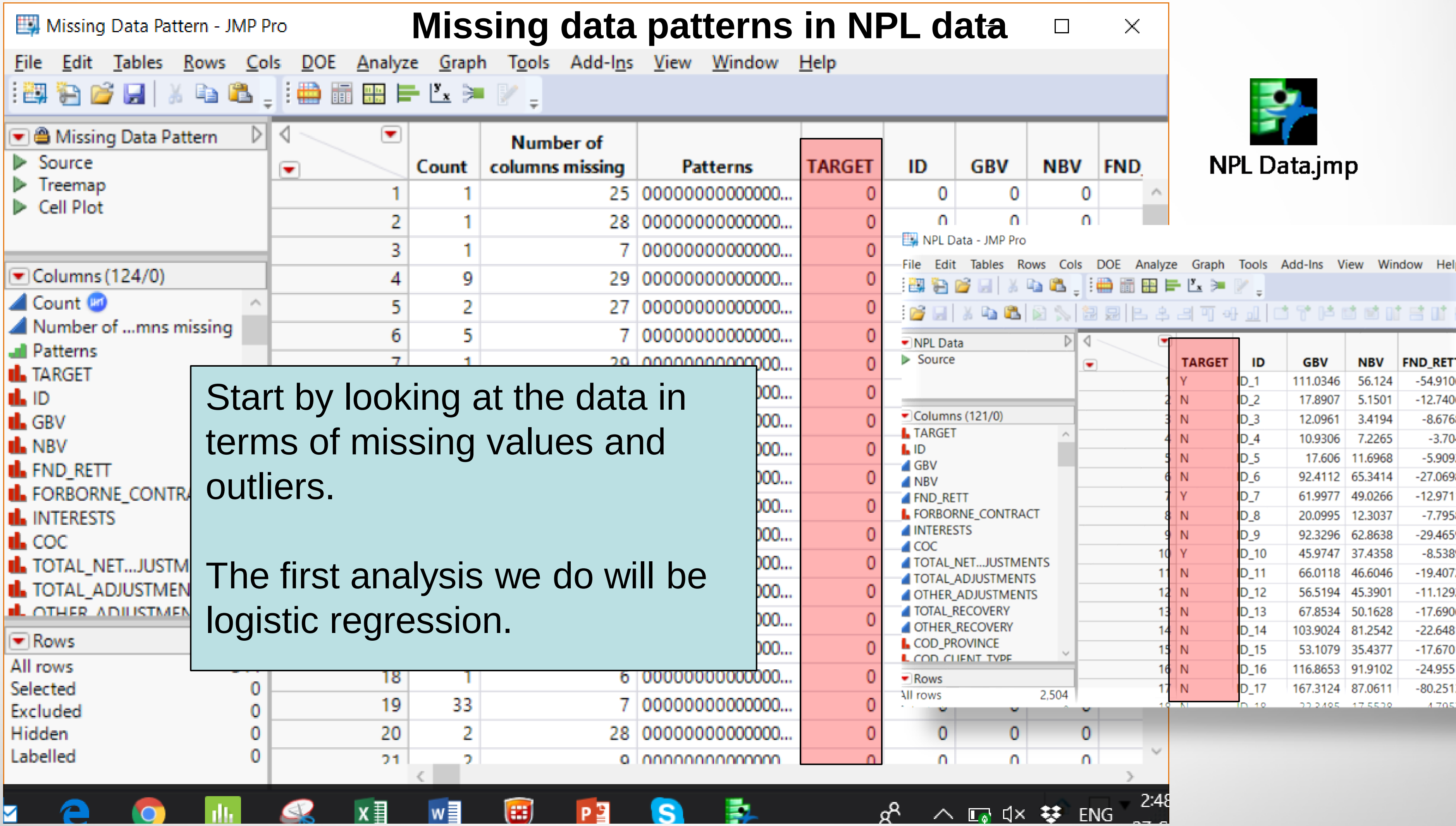
1. A simple (small) tree is fit to the data with a random sample of the predictors
2. The scaled residuals from this tree are calculated
3. A new simple tree is fit to these scaled residuals with another random sample of predictors
4. This process continues
5. The final boosted model is the sum of the models for the individual trees

1. The *number of trees* B. Unlike bagging and random forests, boosting can overfit if B is too large, although this overfitting tends to occur slowly if at all. We use cross-validation to select B.
2. The *shrinkage parameter* λ, a small positive number. This controls the rate at which boosting learns. Typical values are 0.01 or 0.001, and the right choice can depend on the problem. Very small λ can require using a very large value of B in order to achieve good performance.
3. The *number of splits* d in each tree, which controls the complexity of the boosted ensemble. Often $d = 1$ works well, in which case each tree is a *stump*, consisting of a single split and resulting in an additive model. More generally d is the *interaction depth*, and controls the interaction order of the boosted model, since d splits can involve at most d variables.

Ensemble Tree Methods



The non paying loan (NPL) case study



Start by looking at the data in terms of missing values and outliers.

The first analysis we do will be logistic regression.

Outliers

The bank can evaluate outlying cases and determine possible data entry errors or special circumstances. Here we used all data.

Explore Outliers

Quantile Range Outliers

Outliers are values Q times the interquantile range past the lower and upper quantiles.

Tail Quantile 0.1

Q 3

☐ Restrict search to integers

☐ Show only columns with outliers

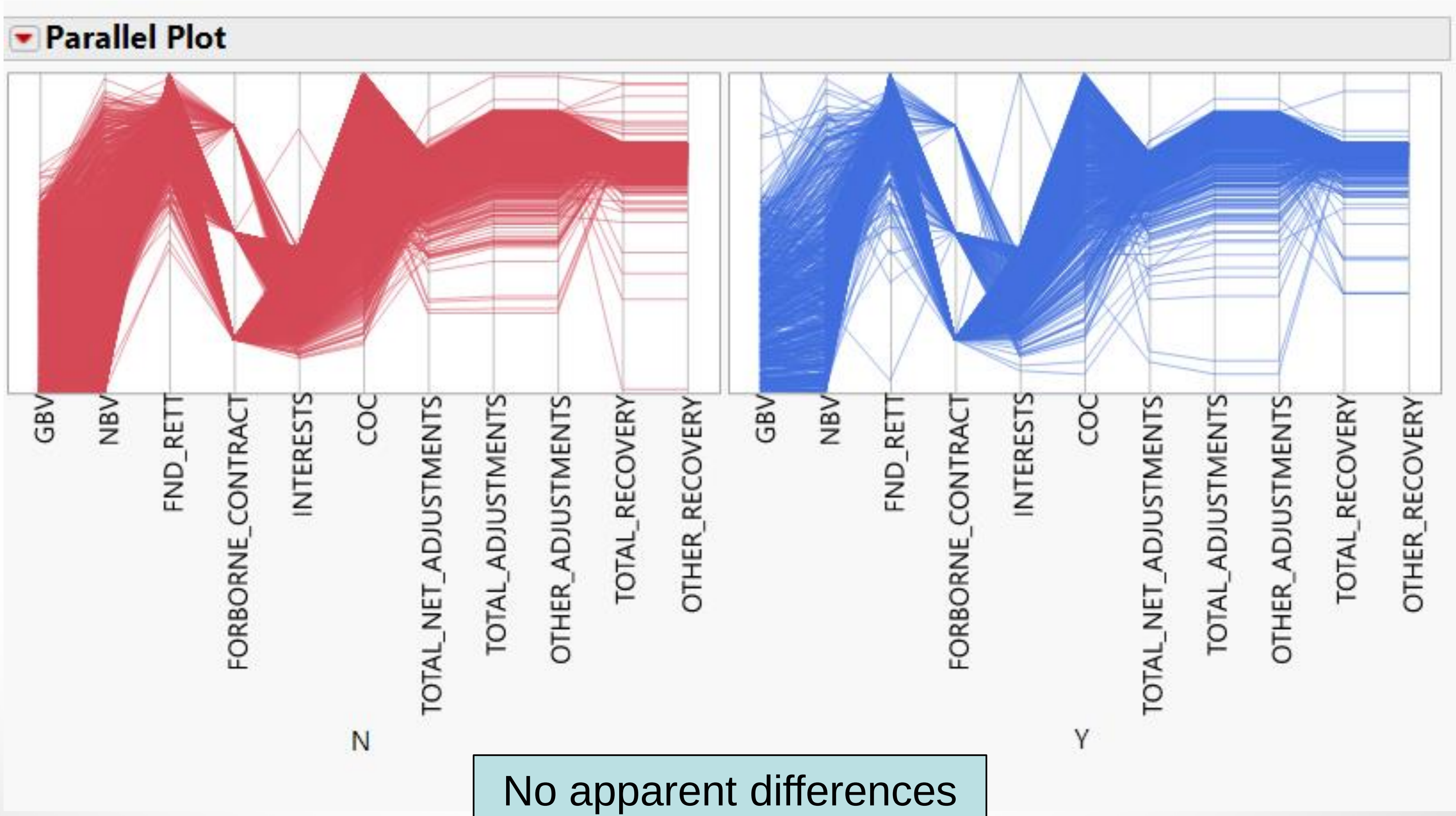
Select columns and choose an action.

Some quantiles were stretched to avoid a large group at the median.

Column	10% Quantile	90% Quantile	Low Threshold	High Threshold	Number of Outliers	Outliers (Count)
GBV	9.4955	122.132	-328.41	460.042	0	
NBV	3.91635	82.2023	-230.94	317.06	0	
FND_RETT	-53.665	-3.7921	-203.28	145.827	1	-216.35
INTERESTS	-0.1304	-0.0076	-0.499	0.361	0	
COC	-0.1304	-0.0077	-0.4987	0.3606	0	
TOTAL_NET_ADJUSTMENTS	-28.687	-1.2657	-110.95	80.9987	2	-130.5706 -124.0946
TOTAL_ADJUSTMENTS	-27.3	-0.8594	-106.62	78.4625	2	-130.5706 -124.0946
OTHER_ADJUSTMENTS	-27.3	-0.8594	-106.62	78.4625	2	-130.5706 -124.0946
TOTAL_RECOVERY	-2.8666	0.09217	-11.743	8.96834	10	-34.4724 -21.8526 -21.0934 -21.031 -20.9955 -18.2491 -16.3949 -16.1417 -15.9399 -15.3047
OTHER_RECOVERY	-2.8666	0.09217	-11.743	8.96834	10	-34.4724 -21.8526 -21.0934 -21.031 -20.9955 -18.2491 -16.3949 -16.1417 -15.9399 -15.3047
AGE	29.875	58.655	-56.465	144.995	0	

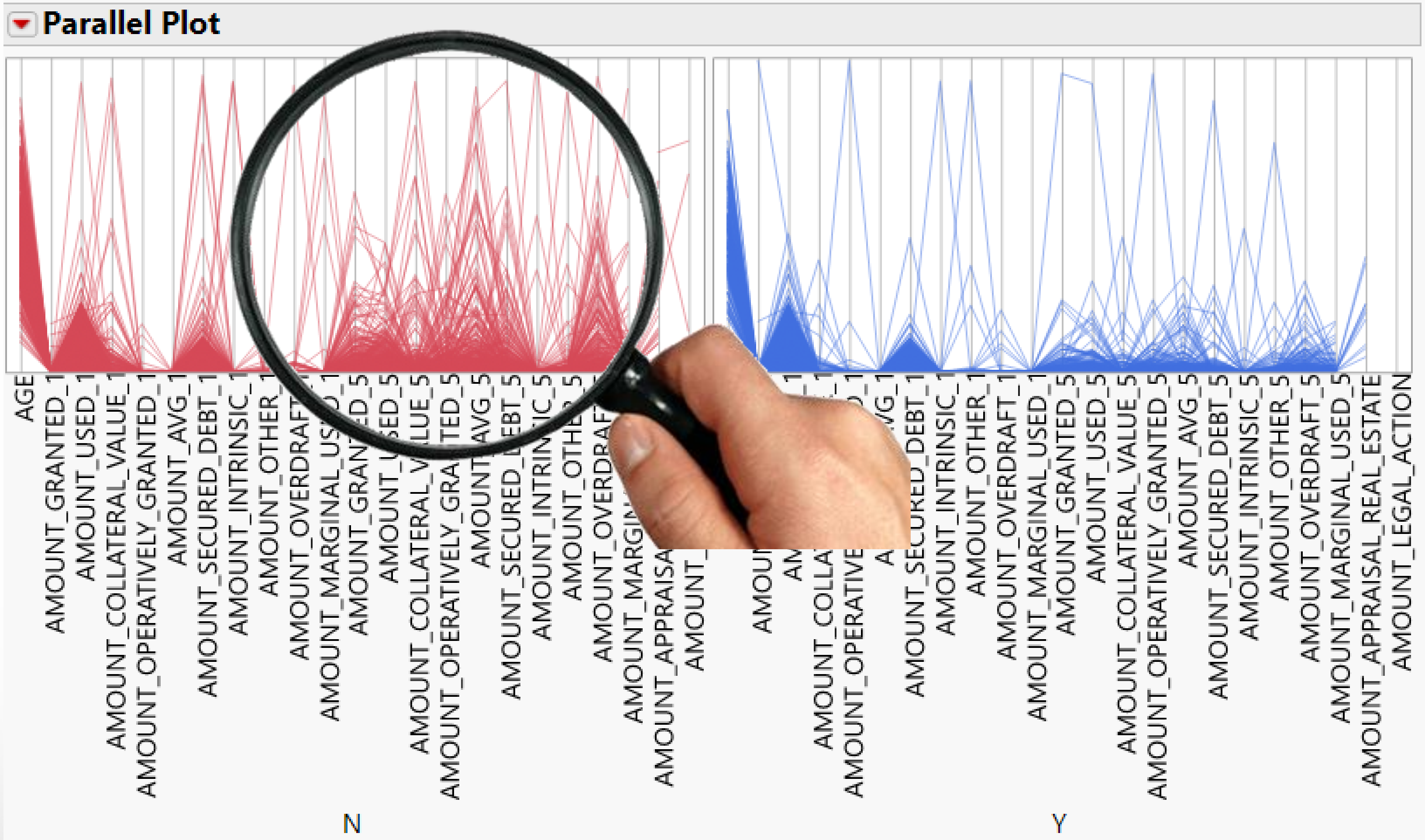
Total recovery with many outliers

Parallel plots



No apparent differences on first 11 variables

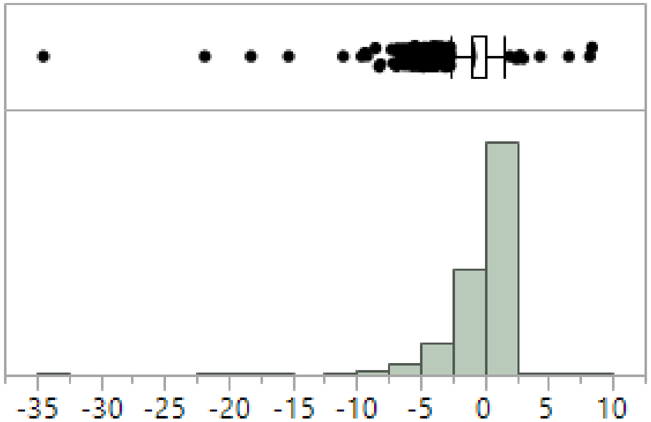
Apparent differences on amounts



A more in depth analysis could reveal some patterns and help redefine variables. We use all variables.

Distributions TARGET=N

TOTAL_RECOVERY



Quantiles

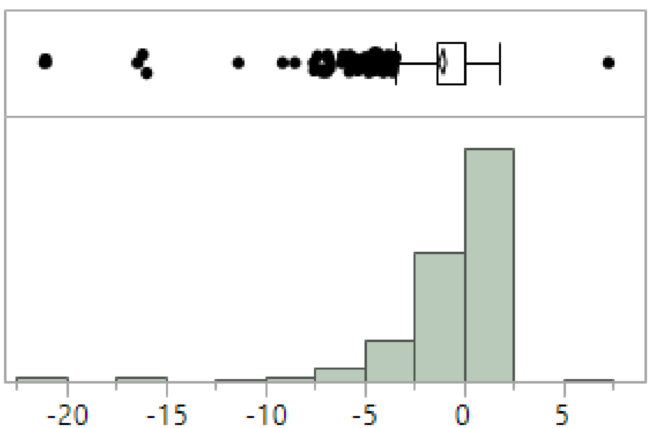
100.0%	maximum	8.4241
99.5%		1.396663
97.5%		0
90.0%		0
75.0%	quartile	0
50.0%	median	0
25.0%	quartile	-1.0715
10.0%		-2.7032
2.5%		-5.22362
0.5%		-8.482124
0.0%	minimum	-34.4724

Summary Statistics

Mean	-0.789656
Std Dev	1.8385048
Std Err Mean	0.0430834
Upper 95% Mean	-0.705158
Lower 95% Mean	-0.874154
N	1821

Distributions TARGET=Y

TOTAL_RECOVERY



Quantiles

100.0%	maximum	7.2869
99.5%		0.57659
97.5%		0
90.0%		0
75.0%	quartile	0
50.0%	median	0
25.0%	quartile	-1.3743
10.0%		-3.29788
2.5%		-6.75773
0.5%		-19.063248
0.0%	minimum	-21.0934

Summary Statistics

Mean	-1.030319
Std Dev	2.3390971
Std Err Mean	0.0895031
Upper 95% Mean	-0.854584
Lower 95% Mean	-1.206054
N	683

Logistic Regression

- Extends idea of linear regression to situation where outcome variable is categorical
- Widely used, particularly where a structured model is useful to explain (=*profiling*) or to predict
- We focus on binary classification

  i.e. Y=0 or Y=1

The Logit

Goal: Find a function of the predictor variables that relates them to a 0/1 outcome

- Instead of *Y* as outcome variable (like in linear regression), we use a function of Y called the ***logit***
- Logit can be modeled as a linear function of the predictors
- The logit can be mapped back to a probability, which, in turn, can be mapped to a class

Step 1: Logistic Response Function

p = probability of belonging to class 1

Need to relate *p* to predictors with a function that guarantees 0 <= *p* <=1

Standard linear function (as shown below) does not constrain the probability:

$$p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_q x_q$$

q = number of predictors

Step 1: Logistic Response Function

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_q x_q)}}$$

KPA

Step 2: Calculate the odds

$$Odds = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q}$$

$$Odds = \frac{p}{1-p}$$

***p* = probability of event**

Or, given the odds of an event, the probability of the event can be computed by:

$$p = \frac{Odds}{1+Odds}$$

Step 3: Take log on both sides

This gives us the logit:

$$\log(Odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q$$

$$\log(Odds) = logit$$

The logit is a linear function of predictors x_1, x_2, …
that takes values from -infinity to +infinity

Personal loan (Universal Bank)

Outcome variable: accept bank loan (no = 0/yes = 1)

Predictors: Demographic info, and info about the customer relationship with the bank

Data Preprocessing

- Partition 60% training, 40% validation
- The data set includes four 2-level categorical predictors that have been coded as 0/1 dummy variables– these variables have the Continuous modeling type

$$\text{Securities Account} = \begin{cases} 1 & \text{if customer has securities account in bank} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{CD Account} = \begin{cases} 1 & \text{if customer has CD account in bank} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Online} = \begin{cases} 1 & \text{if customer uses online banking} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{CreditCard} = \begin{cases} 1 & \text{if customer holds Universal Bank credit card} \\ 0 & \text{otherwise} \end{cases}$$

Single Predictor Model

Modeling loan acceptance on income (*x*)

$$\text{Prob}(Personal\ Loan = Yes \mid Income = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

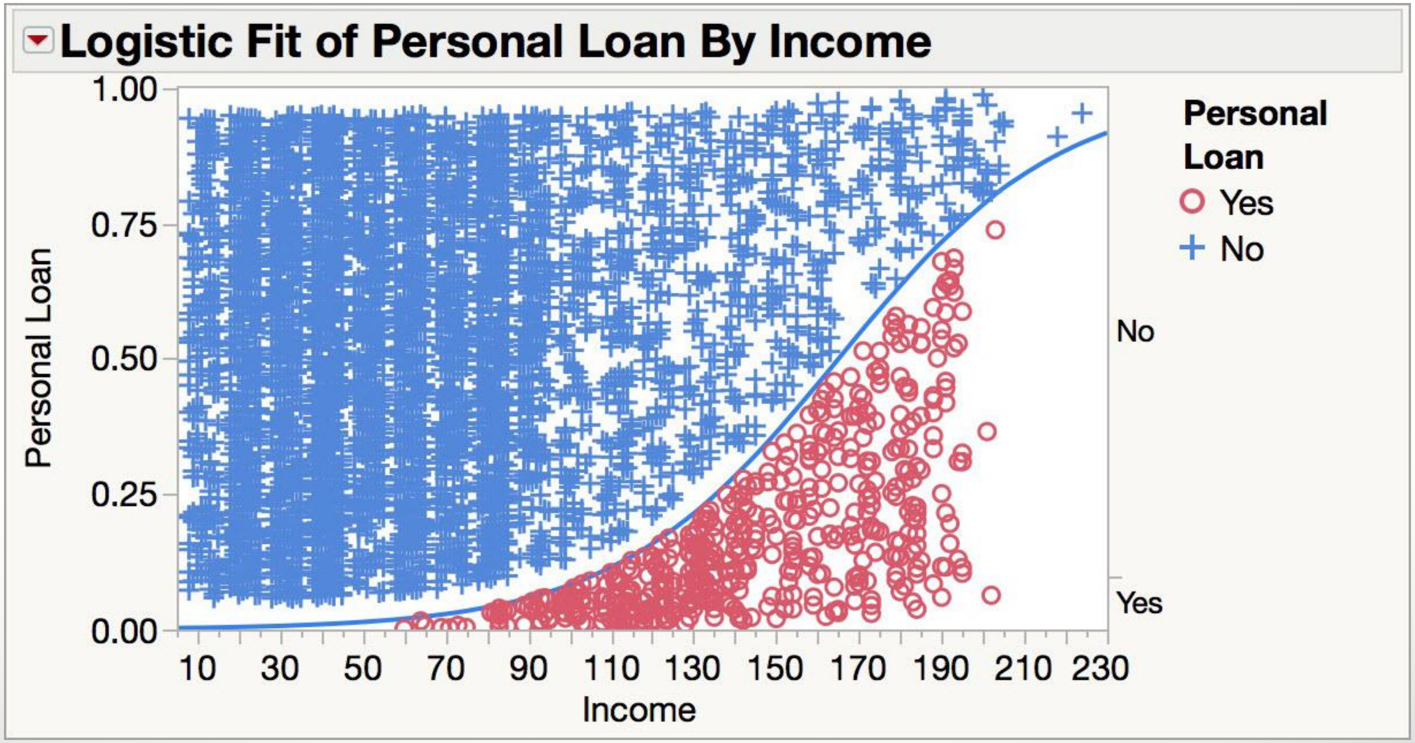
Fitted coefficients (more later): b_0 = -6.3525,

$$P(Personal\ Loan = Yes \mid Income = x) = \frac{1}{1 + e^{6.3525 - 0.0392x}}$$

KPA

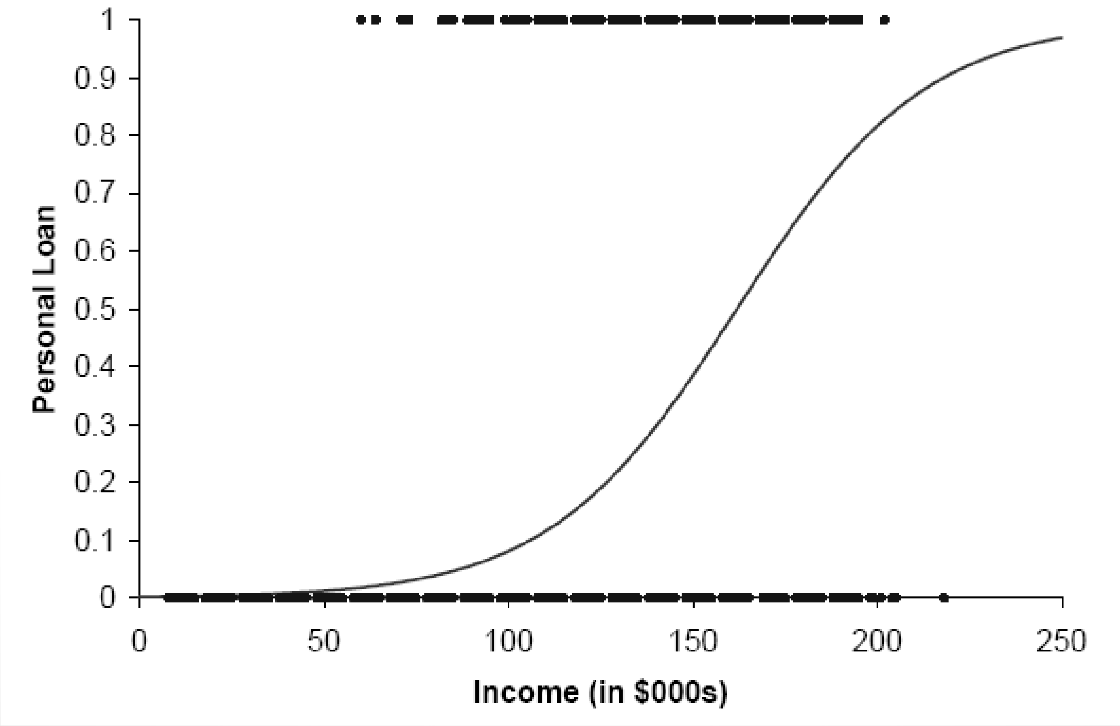
Seeing the Relationship (JMP)

$$P(Personal\ Loan = Yes \mid Income = x) = \frac{1}{1 + e^{6.3525 - 0.0392x}}$$



Seeing the Relationship

Note that the logistic curve is often represented like the one below (in other software packages)



Last step - classify

The logistic model produces an estimated probability of being a yes (or a 1)*.

- Convert to a classification by comparing the estimated probability to a cutoff value
- The default cutoff value is 0.50
- If the estimated probability > 0.50, classify as “yes”

*Note: By default JMP will model the probability of the first category (alphanumerically). To model the probability of 1 rather than the probability of 0, use the Value Ordering column property. In JMP 13 the target category can be specified in the platform.

Ways to determine cutoff

- A cutoff of 0.50 is the default
- Additional considerations
  - Maximize classification accuracy
  - Maximize sensitivity (subject to min. level of specificity)
  - Minimize false positives (subject to max. false negative rate)
  - Minimize expected cost of misclassification (need to specify costs)

Universal Bank example, continued

- Estimates of β's are derived through an iterative process called *maximum likelihood estimation*
- We now fit a full model including all predictors
- JMP reports coefficients for the logit in the Parameter Estimates Table
- Options like Odds Ratios are available under the red triangle

Universal Bank example, continued

- Estimated coefficients

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-10.164069	2.4497848	17.21	<.0001*
Age	-0.044547	0.090961	0.24	0.6243
Experience	0.05658147	0.0900536	0.39	0.5298
Income	0.06576067	0.0042213	242.68	<.0001*
Family	0.57155568	0.1011896	31.90	<.0001*
CCAvg	0.18723439	0.0615372	9.26	0.0023*
Education[Undergrad]	-3.0372506	0.2432931	155.85	<.0001*
Education[Graduate]	1.55179759	0.1752704	78.39	<.0001*
Mortgage	0.00175308	0.0008038	4.76	0.0292*
Securities Account	-0.8548708	0.4186376	4.17	0.0411*
CD Account	3.46902866	0.4489309	59.71	<.0001*
Online	-0.843563	0.2283237	13.65	0.0002*
CreditCard	-0.9640741	0.2825423	11.64	0.0006*

For log odds of Yes/No

Universal Bank example, continued

When the logit is saved to the data table, JMP calculates estimated probabilities, and uses a 0.50 cutoff to classify records (in the Most Likely column)

	Personal Loan	Validation	Lin[Yes]	Prob[Yes]	Prob[No]	Most Likely Personal Loan
1	No	Training	-9.30521373	0.0000909405	0.9999090595	No
2	No	Validation	-10.75437659	0.0000213513	0.9999786487	No
3	No	Validation	-12.6077736	3.345893e-6	0.9999966541	No
4	No	Training	-2.009027973	0.1182582966	0.8817417034	No
5	No	Training	-5.2501519	0.005219337	0.994780663	No
6	No	Training	-5.82861105	0.0029335297	0.9970664703	No
7	No	Validation	-4.130395058	0.015822161	0.984177839	No
8	No	Validation	-8.437624603	0.0002165171	0.9997834829	No
9	No	Training	-3.113222558	0.0425651205	0.9574348795	No
10	Yes	Training	4.3908814129	0.9877618247	0.0122381753	Yes
11	No	Validation	0.2729614358	0.5678197882	0.4321802118	Yes
12	No	Training	-5.772169835	0.0031033348	0.9968966652	No
13	No	Validation	-1.019050966	0.265212302	0.734787698	No
14	No	Validation	-4.888765476	0.0074744258	0.9925255742	No
15	No	Validation	-6.409780202	0.0016426833	0.9983573167	No

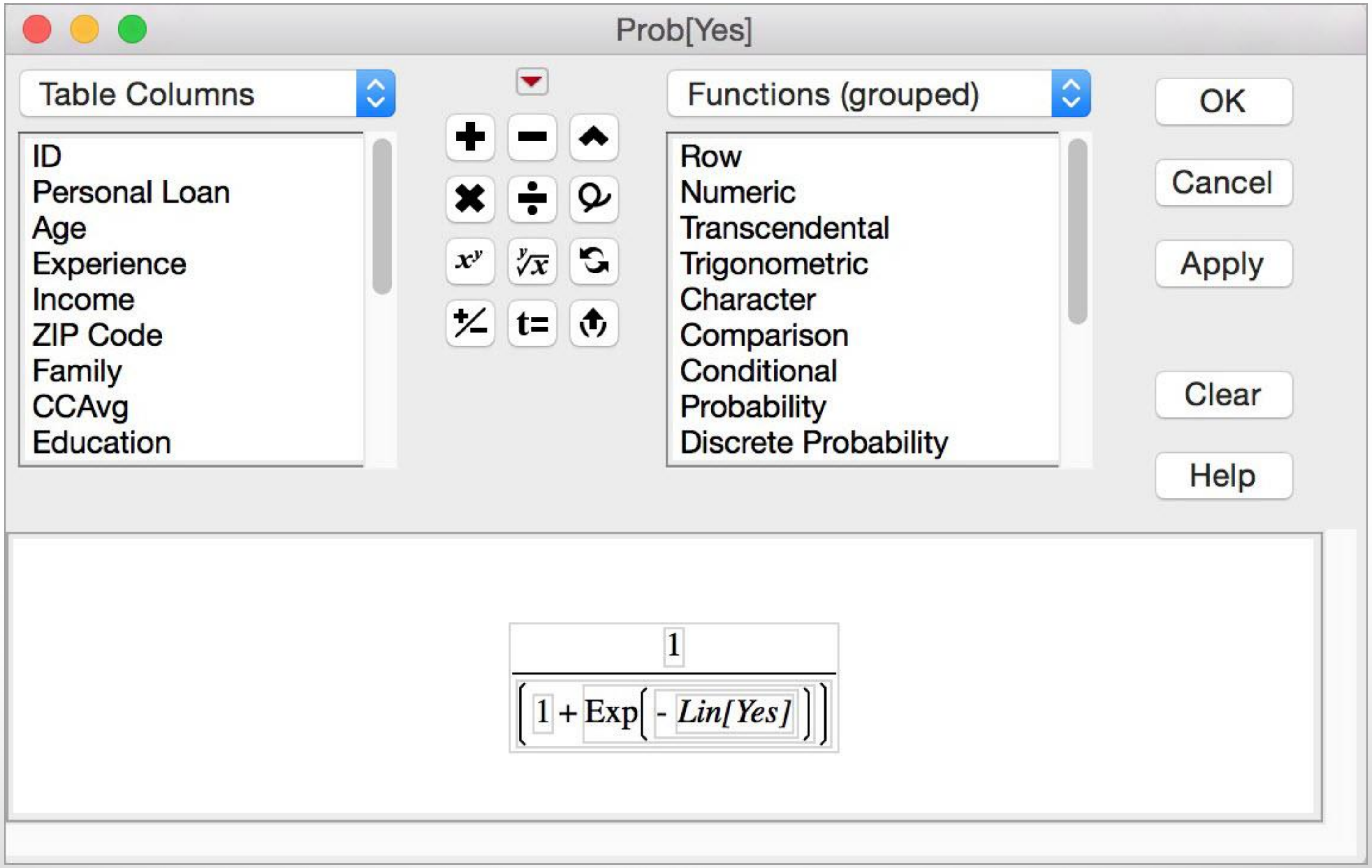
Universal Bank example, continued

- Estimated equation for the logit

$$
\begin{aligned}
&-10.164069254528 \\
&+ -0.0445470057798 * \mathit{Age} \\
&+ 0.05658146902618 * \mathit{Experience} \\
&+ 0.06576067129506 * \mathit{Income} \\
&+ 0.57155567973585 * \mathit{Family} \\
&+ 0.18723439396087 * \mathit{CCAvg} \\
&+ \mathrm{Match}\left(\mathit{Education}\right) \begin{cases} 1 & \Rightarrow -3.0372506132999 \\ 2 & \Rightarrow 1.55179758968979 \\ 3 & \Rightarrow 1.4854530236101 \\ \text{else} & \Rightarrow . \end{cases} \\
&+ 0.00175308271912 * \mathit{Mortgage} \\
&+ -0.8548708298344 * \mathit{Securities\ Account} \\
&+ 3.46902865946834 * \mathit{CD\ Account} \\
&+ -0.8435630344762 * \mathit{Online} \\
&+ -0.9640741062968 * \mathit{CreditCard}
\end{aligned}
$$

Universal Bank example, continued

The logistic response function is used to calculate the probabilities (propensities)



Evaluating classification performance

Performance measures: Confusion matrix and % of misclassifications for the validation set

Measure	Training	Validation	Definition
Entropy RSquare	0.6544	0.5810	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.7228	0.6566	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.1088	0.1334	∑ -Log(ρ[j])/n
RMSE	0.1717	0.1853	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.0607	0.0644	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0367	0.0470	∑ (ρ[j]≠ρMax)/n
N	3000	2000	n

Confusion Matrix

Training

Actual Personal Loan	Predicted Count Yes	Predicted Count No
Yes	201	85
No	25	2689

Validation

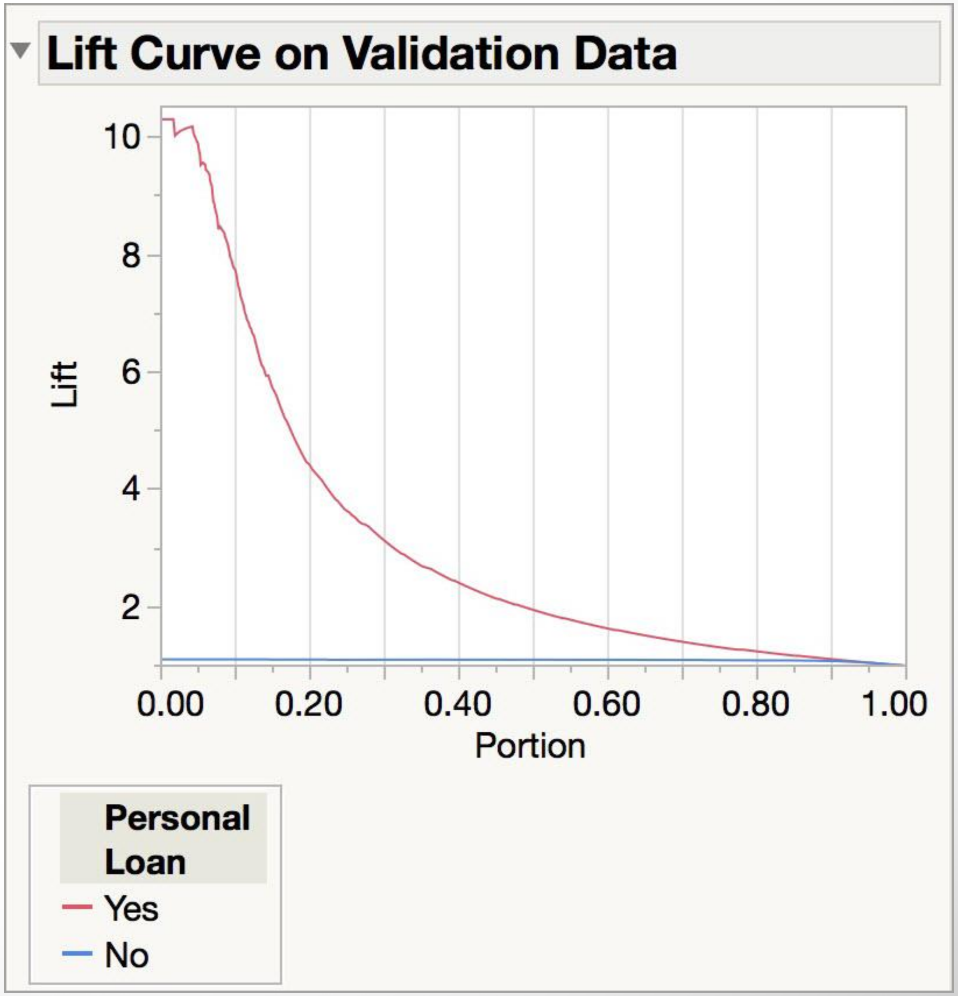
Actual Personal Loan	Predicted Count Yes	Predicted Count No
Yes	126	68
No	26	1780

Evaluating classification performance

The rate for the target category is low (<10%)

So, more useful in this example is: **lift**

The lift for the top 10% of the sorted probabilities (Yes) = 7.7



Multicollinearity

Problem: As in linear regression, if one predictor is a linear combination of other predictor(s), model estimation will fail

– Note that in such a case, we have at least one redundant predictor

Solution: Remove extreme redundancies (by dropping predictors via variable selection or by data reduction methods such as PCA)

Variable selection

This is the same issue as in linear regression:

- The number of correlated predictors can grow when we create derived variables such as **interaction terms** (e.g. *Income x Family)*, to capture more complex relationships
- **Problem**: Overly complex models have the danger of overfitting
- **Solution**: Reduce variables via automated selection of variable subsets (as with linear regression)
  - Data preparation strategies (e.g. grouping or binning) can also reduce the number parameters to be estimated

P-values for predictors

- Test null hypothesis that coefficient = 0
- Useful for review to determine whether to include variable in model
- Key in profiling tasks, but less important in predictive classification

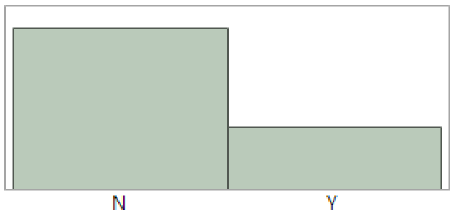
Logistic regression overview

- Logistic regression is similar to linear regression, except that it is used with a categorical response
- It can be used for explanatory tasks (=profiling) or predictive tasks (=classification)
- The predictors are related to the response Y via a nonlinear function called the *logit*
- As in linear regression, reducing predictors can be done via variable selection
- Logistic regression can be generalized to more than two classes (ordinal or multinomial)

NPL logistic regression

Distributions

TARGET

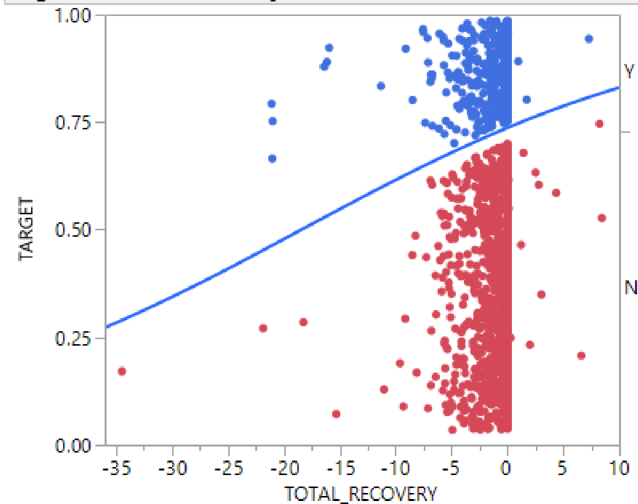


Frequencies

Level	Count	Prob
N	1821	0.72724
Y	683	0.27276
Total	2504	1.00000

N Missing 0
2 Levels

Logistic Fit of TARGET By TOTAL_RECOVERY

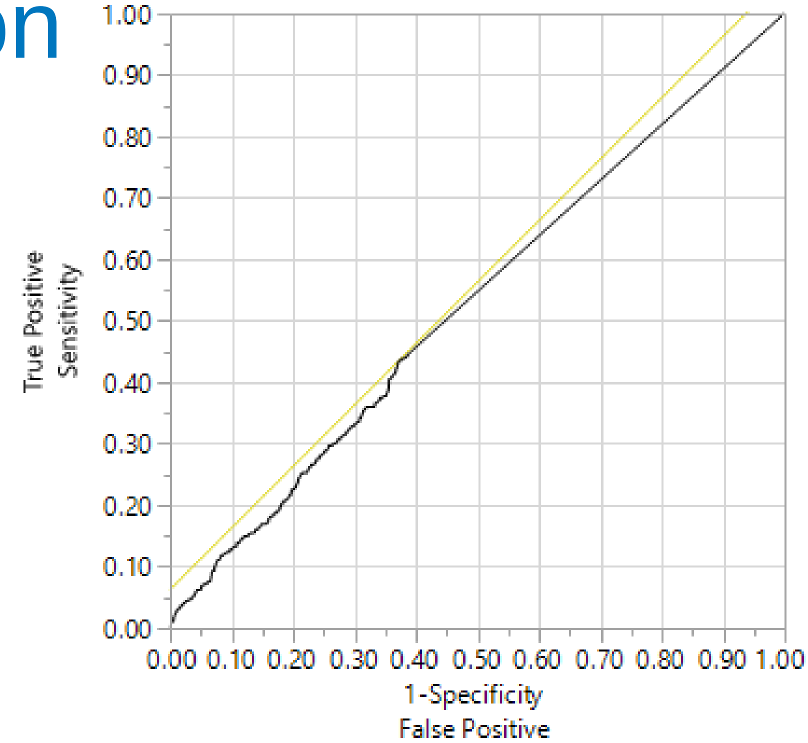


Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	1.03078919	0.0491292	440.21	<.0001*
TOTAL_RECOVERY	0.05586896	0.0213757	6.83	0.0090*

For log odds of N/Y

Receiver Operating Characteristic



Using TARGET='Y' to be the positive level

AUC
0.52903

Simple logistic regression on total recovery is not informative because of little spread. Transforming the data could prove more informative.

Naïve Bayes: The basic idea

For a given new record to be classified:

- Find other records like it (i.e., same values for the predictors)
- Identify the prevalent class among those records
- Assign that class to your new record

Usage

- Requires categorical variables
- Numerical variable must be binned and converted to categorical
- Can be used with very large data sets
- Example: Spell check programs assign your misspelled word to an established "class" (i.e., correctly spelled word)

Exact Bayes classifier

- Relies on finding other records that share <u>same predictor values</u> as record-to-be-classified.
- Want to find "probability of belonging to class *C*, given specified values of predictors."
- Even with large data sets, may be hard to find other records that **exactly match** your record, in terms of predictor values.

Solution – Naïve Bayes

- Assume independence of predictor variables (within each class)
- Use multiplication rule
- Find same probability that record belongs to class C, given predictor values, <u>without</u> limiting calculation to records that share all those same values

Naïve Bayes procedure

Take a record, and note its predictor values:

1. Find the probabilities those predictor values occur across all records in C1
2. Multiply them together, then by the proportion of records belonging to C1
3. Repeat steps 1 and 2 for each class
4. The probability of belonging to C1 is value from step (3) divide by sum of all such values C1 … Cn
5. Establish and adjust a “cutoff” prob. for class of interest

Example: financial fraud

Target variable:

- Audit finds fraud, no fraud

Predictors:

- Prior pending legal charges (yes/no)
- Size of firm (small/large)

	Prior Legal Trouble	Company Size	Status
1	Yes	Small	Truthful
2	No	Small	Truthful
3	No	Large	Truthful
4	No	Large	Truthful
5	No	Small	Truthful
6	No	Small	Truthful
7	Yes	Small	Fraudulent
8	Yes	Large	Fraudulent
9	No	Large	Fraudulent
10	Yes	Large	Fraudulent

Exact Bayes calculations

Goal: classify (as "fraudulent" or as "truthful") a small firm with charges filed

- There are 2 firms like that, one fraudulent and the other truthful
- P(fraud | charges=y, size=small) = ½ = 0.50

Note: calculation is limited to the two firms matching those characteristics

KPA

Naïve Bayes calculations

Same goal as before

Compute 2 quantities:

– Proportion of “charges = y” among frauds, times proportion of “small” among frauds, times proportion frauds = 3/4 * 1/4 * 4/10 = 0.075

– Prop “charges = y” among frauds, times prop. “small” among truthfuls, times proportion truthfuls = 1/6 * 4/6 * 6/10 = 0.067

P(fraud | charges, small) = 0.075/(0.075+0.067)

= 0.53

Naïve Bayes, continued

- Note that probability **estimate** does not differ greatly from **exact**
- All records are used in calculations, not just those matching predictor values
- This makes calculations practical in most circumstances
- Relies on assumption of independence between predictor variables within each class

Independence assumption

- Not strictly justified (variables often correlated with one another)
- Often "good enough"

Naïve Bayes advantages

- Handles purely categorical data well
- Works well with very large data sets
- Simple and computationally efficient

Naïve Bayes shortcomings

- Requires large number of records
- Problematic when a predictor category is not present in training data
  - Assigns 0 probability of response, ignoring information in other variables

On the other hand…

- Probability <u>rankings</u> are more accurate than the actual probability estimates
  - Good for applications using lift (e.g. response to mailing), less so for applications requiring probabilities (e.g. credit scoring)

Naïve Bayes overview

$$P(y \mid x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots x_n \mid y)}{P(x_1, \dots, x_n)}$$

$$P(y \mid x_1, \dots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \dots, x_n)}$$

- No statistical models involved
- Naïve Bayes (like KNN) pays attention to complex interactions and local structure
- Computational challenges remain

NPL Naïve Bayes

Assumes independence of predictors

Naive Bayes

TARGET

Training Set

Count	Misclassification Rate	Misclassifications
1748	0.66648	1165

Validation Set

Count	Misclassification Rate	Misclassifications
756	0.71561	541

Confusion Matrix

Training Set

Actual TARGET	Predicted Count N	Predicted Count Y
N	118	1152
Y	13	465

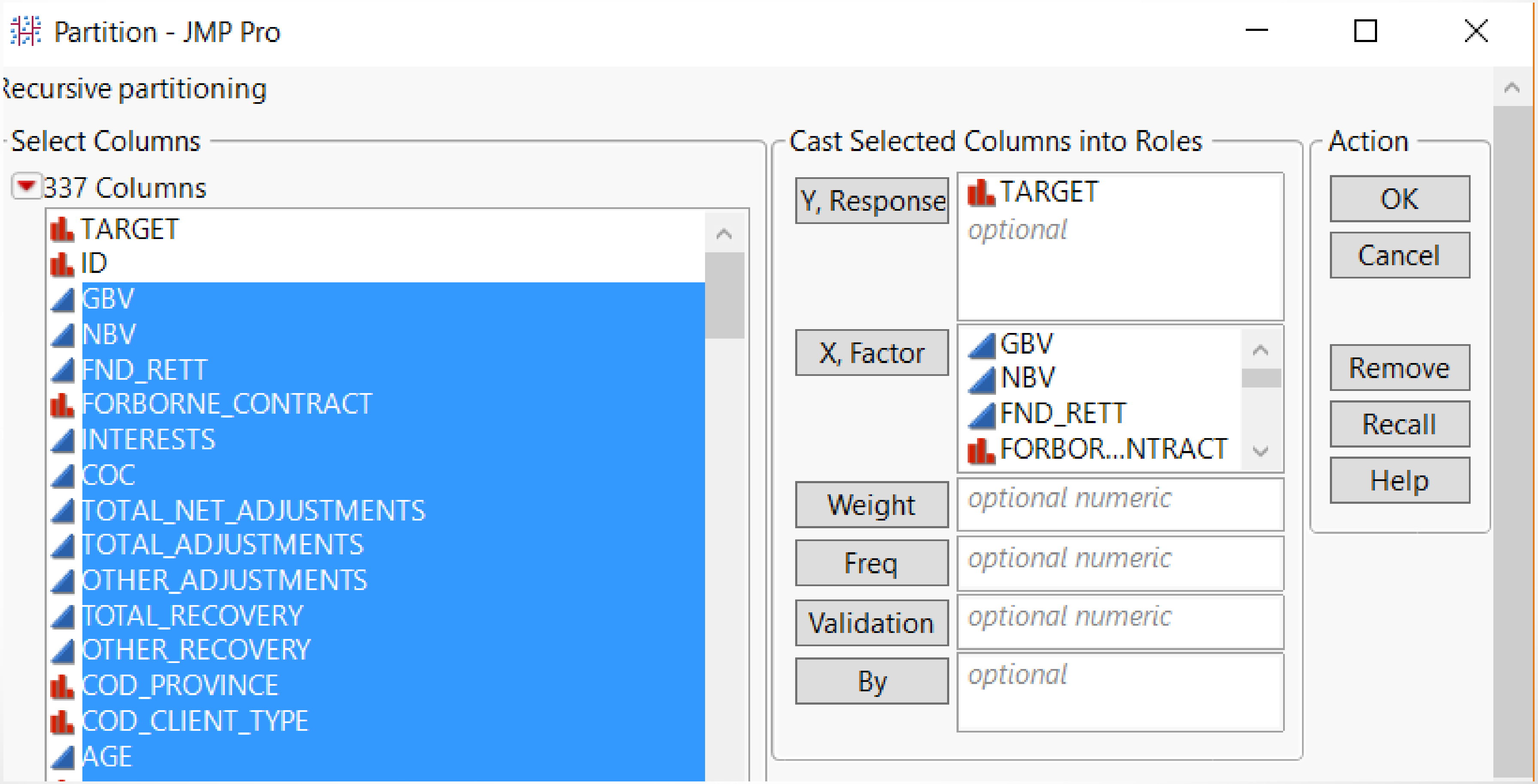
Validation Set

Actual TARGET	Predicted Count N	Predicted Count Y
N	36	515
Y	26	179

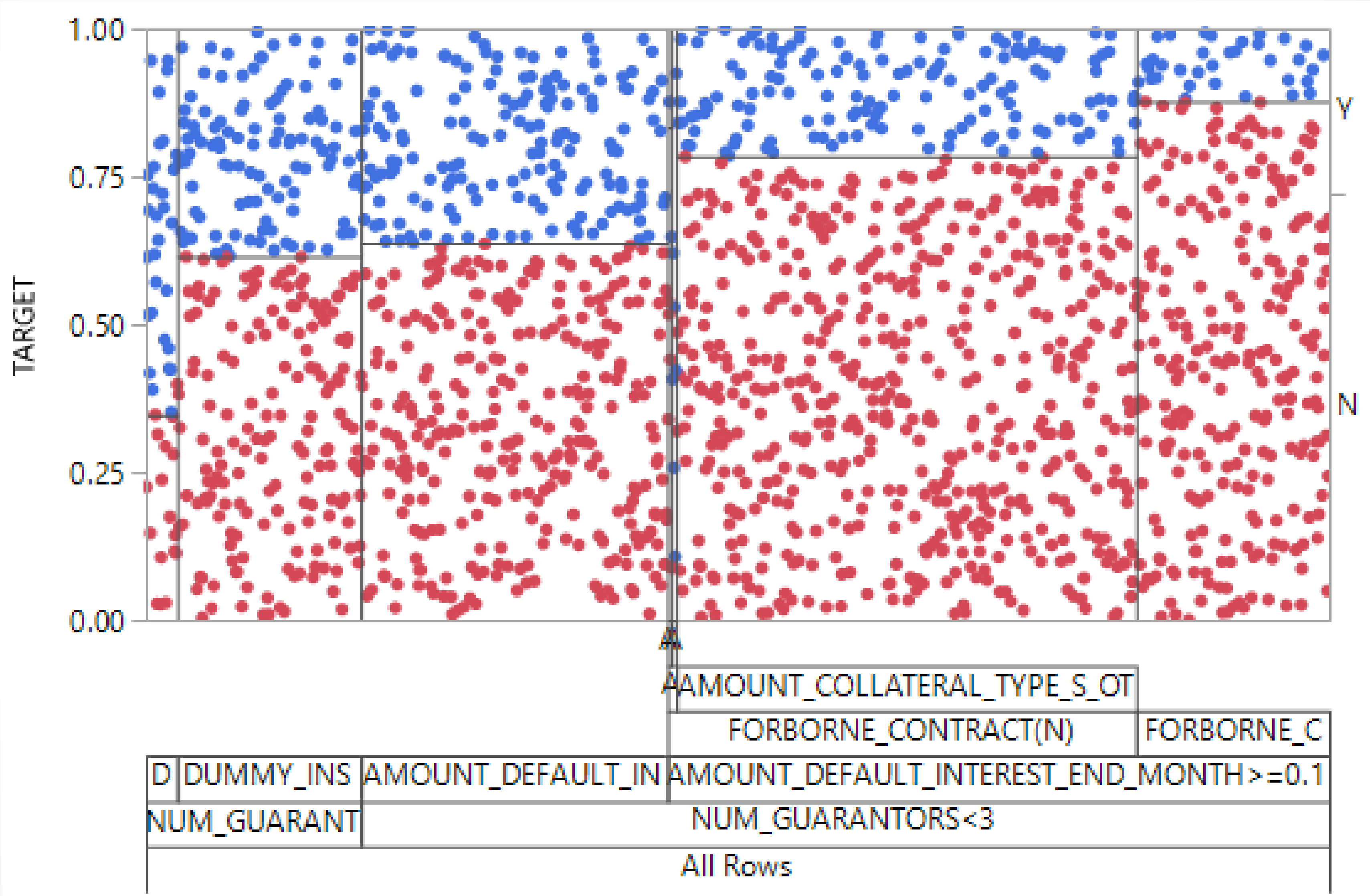
Sensitivity (N classified as N) = 9.3%

Specificity (Y classified as Y) = 97.3%

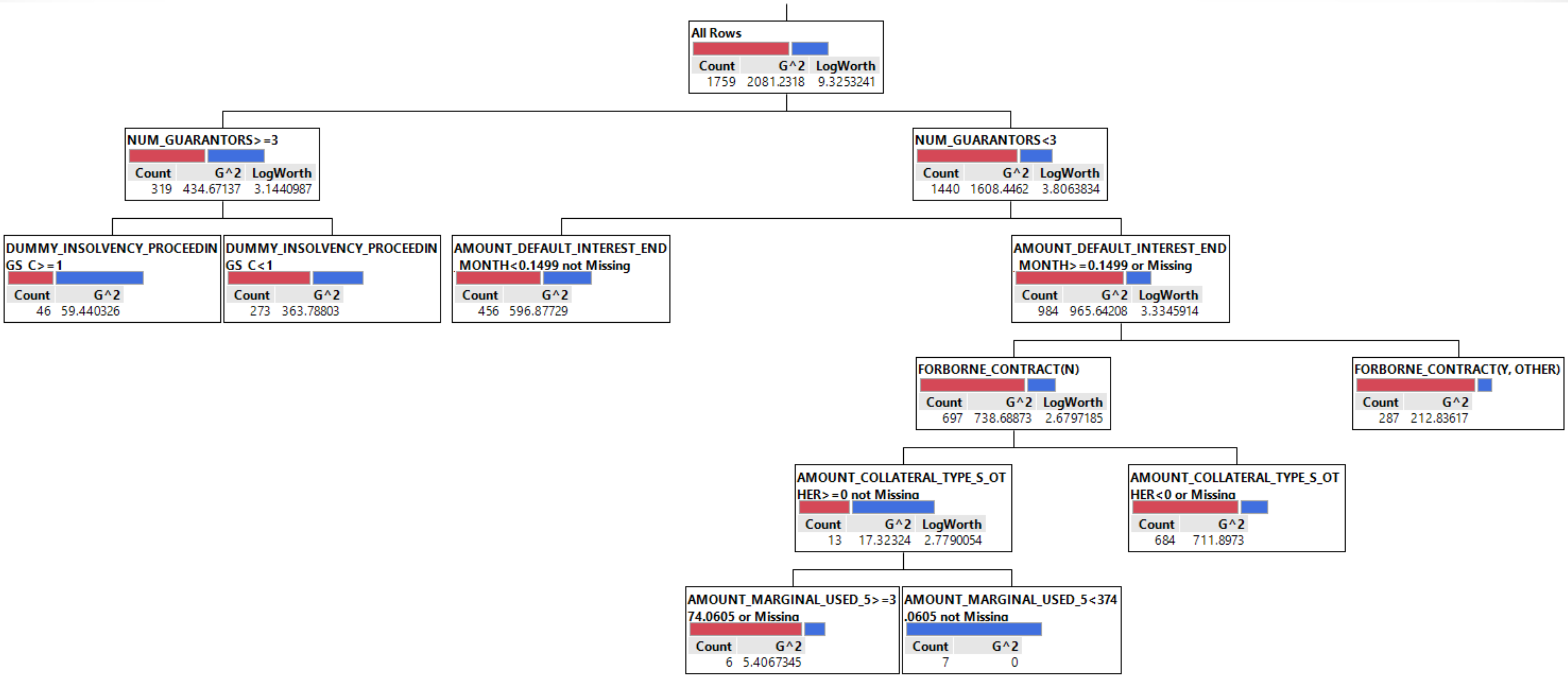
NPL decision trees



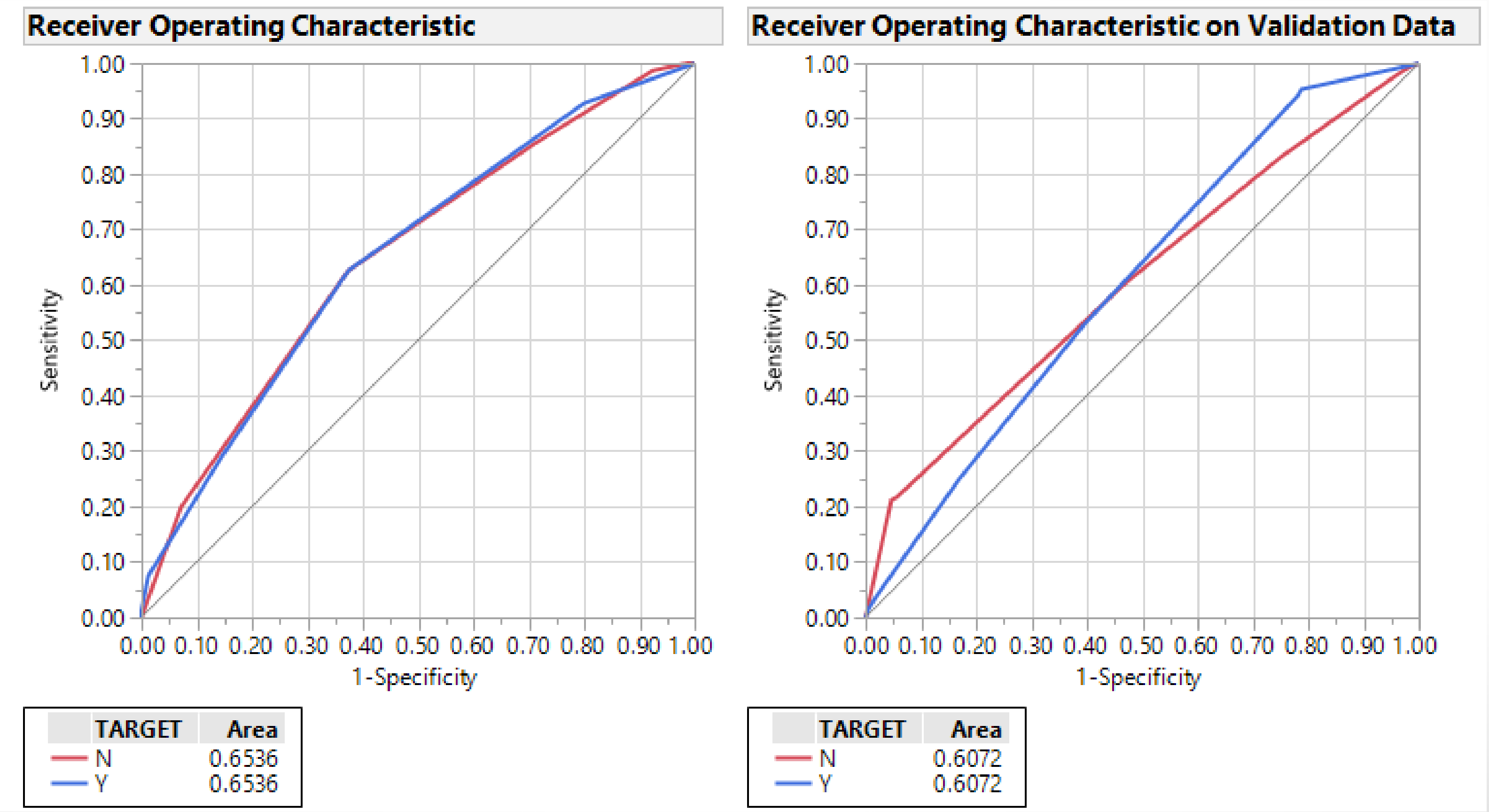
KPA


1.00
0.75
0.50
0.25
0.00
TARGET
Y
N
AMOUNT_COLLATERAL_TYPE_S_OT
FORBORNE_CONTRACT(N)
FORBORNE_C
D
DUMMY_INS
AMOUNT_DEFAULT_IN
AMOUNT_DEFAULT_INTEREST_END_MONTH>=0.1
NUM_GUARANT
NUM_GUARANTORS<3
All Rows

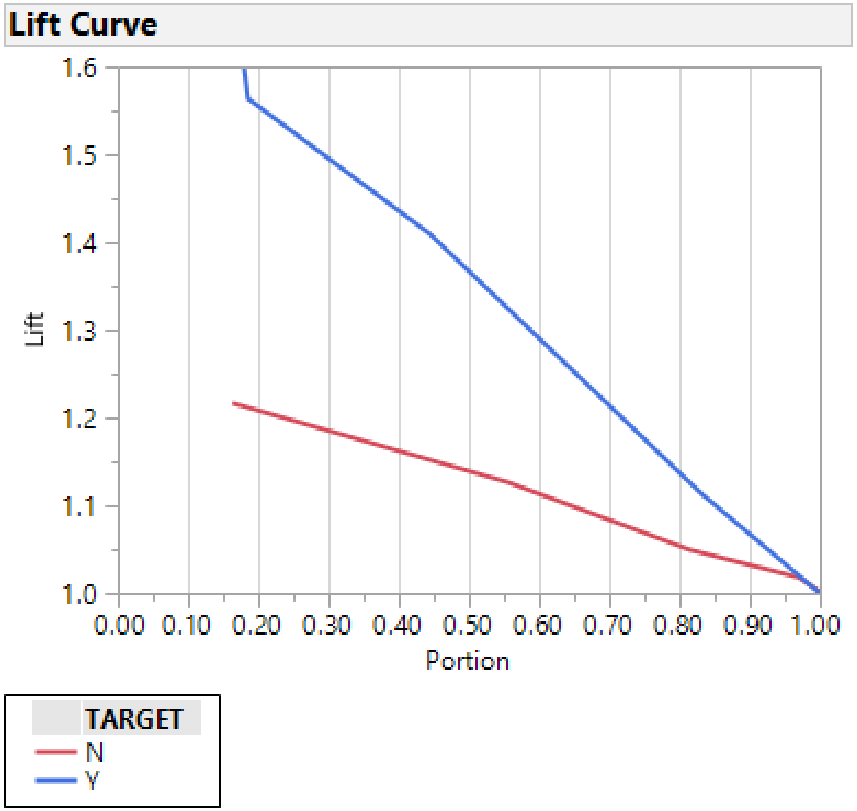
	RSquare	N	Number of Splits
Training	0.062	1759	6
Validation	0.013	745	

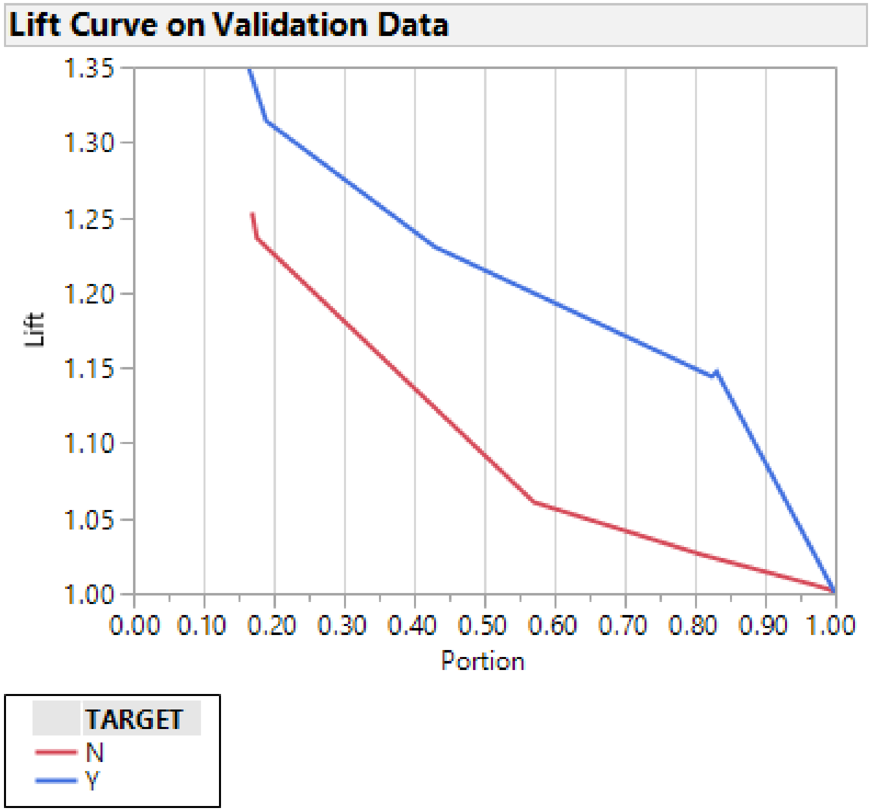


ROC of decision tree



Lift of decision tree





Variable contributions to decision tree

Column Contributions

Term	Number of Splits	G^2	Portion
AMOUNT_DEFAULT_INTEREST_END_MONTH	1	45.926837	0.3506
NUM_GUARANTORS	1	38.1141701	0.2910
FORBORNE_CONTRACT	1	14.1171833	0.1078
AMOUNT_MARGINAL_USED_5	1	11.916505	0.0910
DUMMY_INSOLVENCY_PROCEEDINGS_C	1	11.4430186	0.0874
AMOUNT_COLLATERAL_TYPE_S_OTHER	1	9.468194	0.0723
GBV	0	0	0.0000
NBV	0	0	0.0000
FND_RETT	0	0	0.0000
INTERESTS	0	0	0.0000
COC	0	0	0.0000
TOTAL_NET_ADJUSTMENTS	0	0	0.0000
TOTAL_ADJUSTMENTS	0	0	0.0000
OTHER_ADJUSTMENTS	0	0	0.0000
TOTAL_RECOVERY	0	0	0.0000
OTHER_RECOVERY	0	0	0.0000
COD_PROVINCE	0	0	0.0000
COD_CLIENT_TYPE	0	0	0.0000
AGE	0	0	0.0000
COD_ATECO_100VAL	0	0	0.0000

Performance of decision tree

Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.0623	0.0132	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.1024	0.0219	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5547	0.5645	∑ -Log(ρ[j])/n
RMSE	0.4314	0.4361	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.3727	0.3748	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2666	0.2644	∑ (ρ[j]≠ρMax)/n
N	1759	745	n

Confusion Matrix

Training

Actual	Predicted Count	
TARGET	N	Y
N	1253	16
Y	453	37

Validation

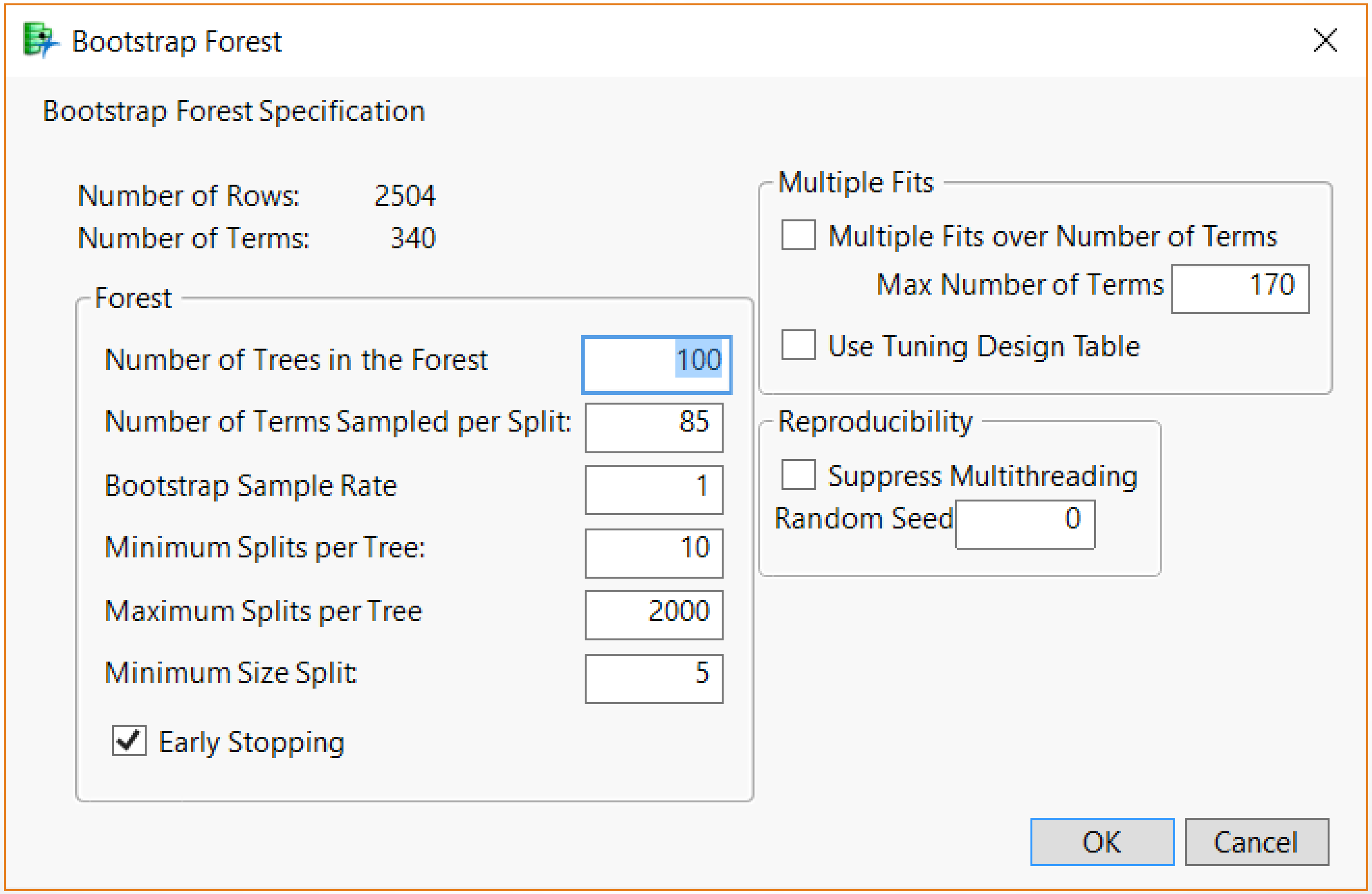
Actual	Predicted Count	
TARGET	N	Y
N	540	12
Y	185	8

Sensitivity (N classified as N) = 98.7%

Specificity (Y classified as Y) = 7.5%

KPA

Random forests



Random forests

Overall Statistics

Measure	Training	Validation	Definition
Entropy RSquare	0.5497	0.0564	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6870	0.0935	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.2617	0.5633	∑ -Log(ρ[j])/n
RMSE	0.2575	0.4352	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.2155	0.3682	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0403	0.2749	∑ (ρ[j]≠ρMax)/n
N	1762	742	n

Confusion Matrix

Training

Actual	Predicted Count	
TARGET	N	Y
N	1290	0
Y	71	401

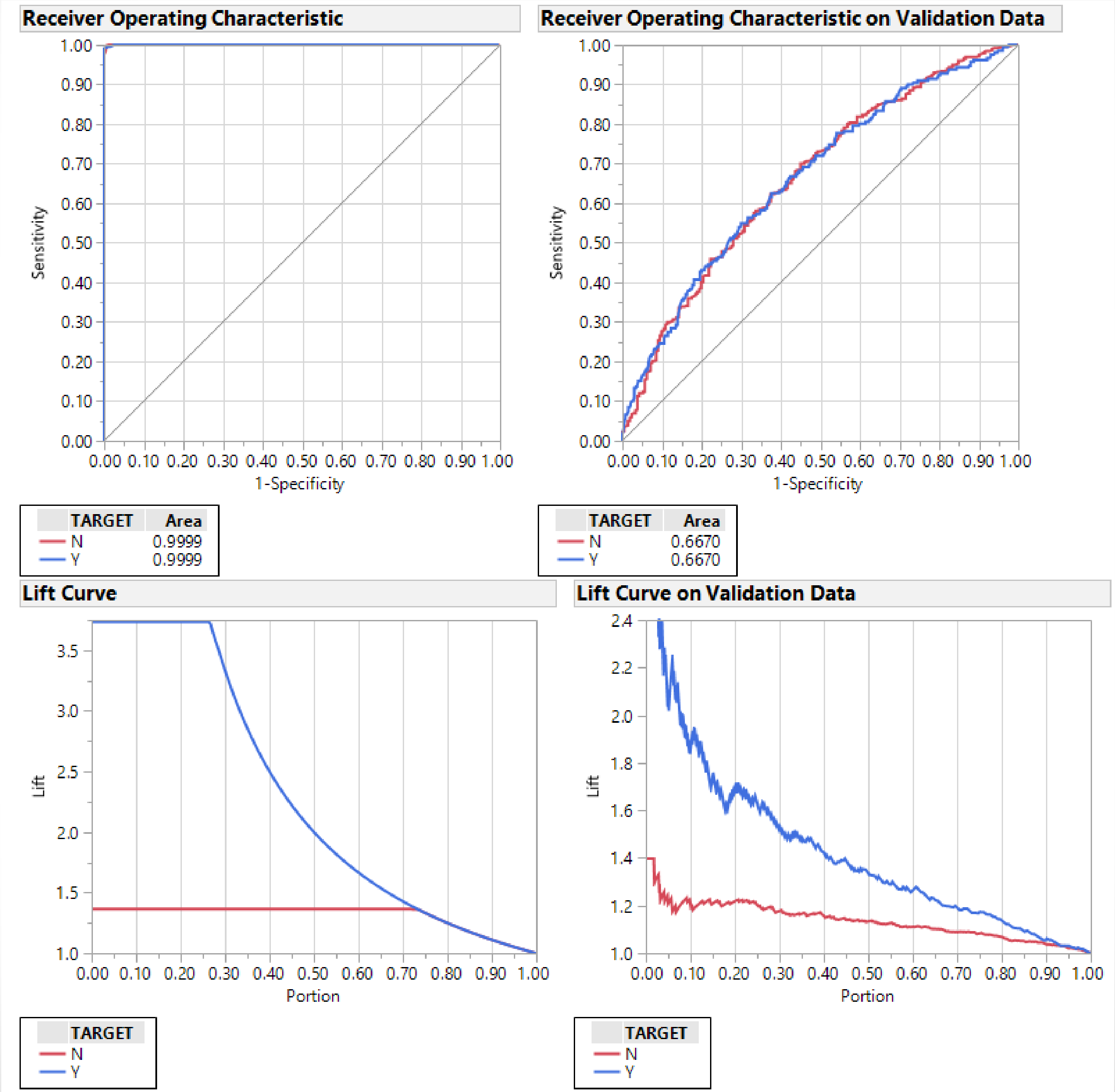
Validation

Actual	Predicted Count	
TARGET	N	Y
N	527	4
Y	200	11

Sensitivity (N classified as N) = 100%

Specificity (Y classified as Y) = 84.9%

KPA


Receiver Operating Characteristic
Sensitivity
1-Specificity
0.00 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00
TARGET Area
N 0.9999
Y 0.9999
Receiver Operating Characteristic on Validation Data
Sensitivity
1-Specificity
0.00 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00
TARGET Area
N 0.6670
Y 0.6670
Lift Curve
Lift
1.0 1.5 2.0 2.5 3.0 3.5
Portion
0.00 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00
TARGET
N
Y
Lift Curve on Validation Data
Lift
1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4
Portion
0.00 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00
TARGET
N
Y

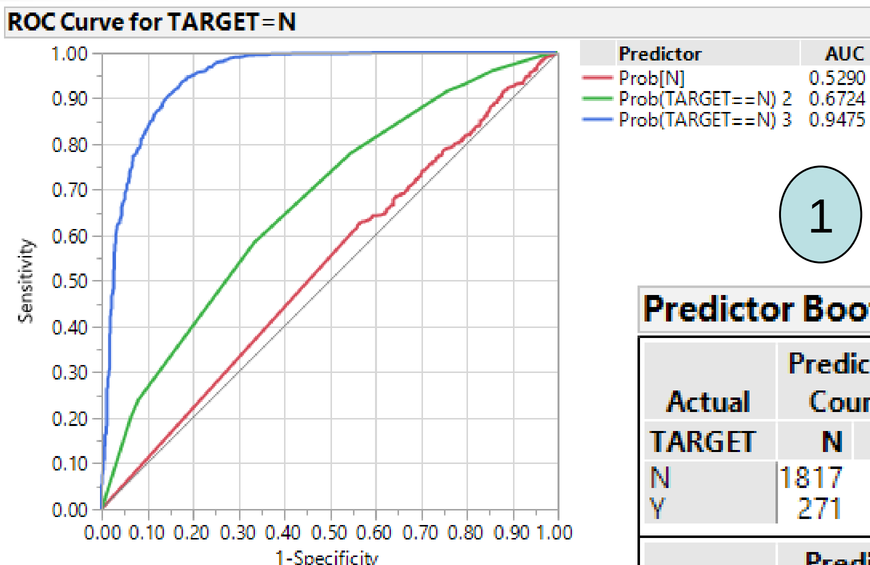
Variable contributions to forest

Column Contributions

Term	Number of Splits	G^2	Portion
COD_PROVINCE	543	104.771204	0.1081
COD_ATECO_100VAL	376	47.9837687	0.0495
COD_ISTAT_ATECO_07	297	43.2221623	0.0446
AMOUNT_DEFAULT_INTEREST_END_MONTH	381	36.4398855	0.0376
NUM_MONTHS_FROM_FIRST_CONTRACT	356	24.8622738	0.0256
COD_RAE	180	21.5436371	0.0222
VINTAGE_LEGAL_PROCEDURE	343	21.0771415	0.0217
VINTAGE_STRANDING	310	19.1737245	0.0198
GEOGRAPHICAL_POSITION	320	17.5453309	0.0181
AMOUNT_BANK_ACCOUNT_GUARANTORS	155	14.4818336	0.0149
TOTAL_NET_ADJUSTMENTS	229	14.1764942	0.0146
TOTAL_ADJUSTMENTS	223	14.1484413	0.0146
GBV	219	13.3053047	0.0137
AMOUNT_COLLATERAL_TYPE_X_REAL_ACTUALIZED	212	13.2631144	0.0137
OTHER_ADJUSTMENTS	222	13.1912139	0.0136
COD_SAE	190	13.0545203	0.0135
NBV	199	12.9392247	0.0133
AMOUNT_COLLATERAL_TYPE_X_REAL	204	12.7548631	0.0132
AMOUNT_COLLATERAL	201	12.6033107	0.0130
FND_RETT	209	12.5426841	0.0129
INTERESTS	218	12.510933	0.0129
AMOUNT_USED_1	189	12.1616324	0.0125
AMOUNT_SECURED_DEBT_1	191	12.0697238	0.0124
MEDIUM_LONG_TERM_LOANS_RECORDED_ARREARS	177	11.6562735	0.0120
AMOUNT_USED_5	191	11.5505344	0.0119
AMOUNT_SECURED_DEBT_5	192	11.4543736	0.0118
FORBORNE_CONTRACT	216	11.4490915	0.0118
COC	190	11.3884388	0.0117
AMOUNT_COLLATERAL_FAIR_VALUE	200	11.2502439	0.0116
MEDIUM_LONG_TERM_LOANS_OVERDUE_DEBT	174	11.0867439	0.0114

Measures of Fit for TARGET

Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N	AUC
Logistic		0.0023	0.0039	0.5846	0.4447	0.3956	0.2724	2504	0.5290
Partition		0.0691	0.1128	0.5455	0.4270	0.3652	0.2668	2504	0.6724
Bootstrap Forest		0.4009	0.5432	0.3511	0.3206	0.2607	0.1098	2504	0.9475



Random Forests provide the best performance

1

Predictor Bootstrap Forest

Actual TARGET	Predicted Count N	Predicted Count Y
N	1817	4
Y	271	412

Actual TARGET	Predicted Rate N	Predicted Rate Y
N	0.998	0.002
Y	0.397	0.603

2

Predictor Partition

Actual TARGET	Predicted Count N	Predicted Count Y
N	1695	126
Y	542	141

Actual TARGET	Predicted Rate N	Predicted Rate Y
N	0.931	0.069
Y	0.794	0.206

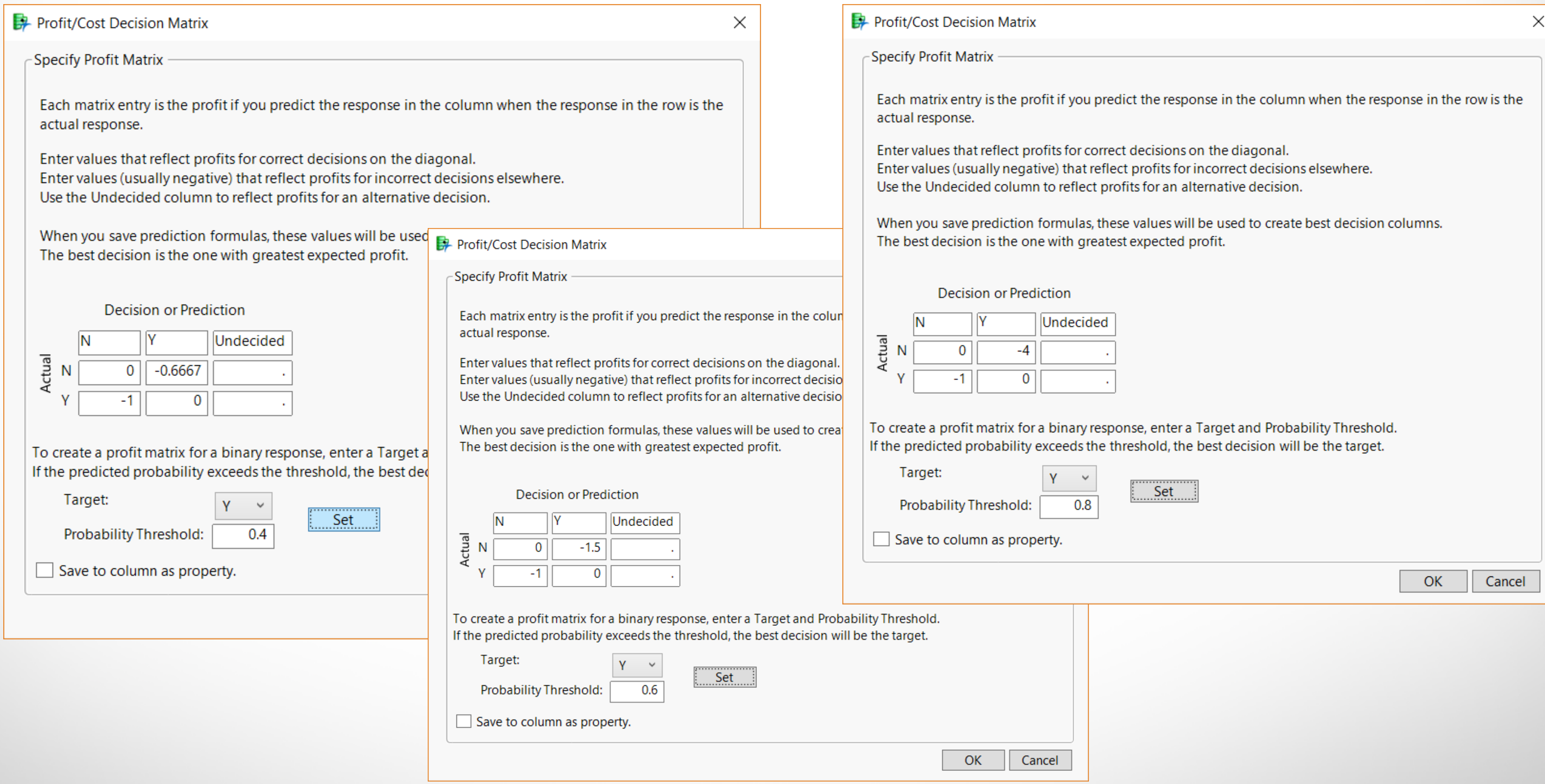
3

Predictor Logistic

Actual TARGET	Predicted Count N	Predicted Count Y
N	1819	2
Y	680	3

Actual TARGET	Predicted Rate N	Predicted Rate Y
N	0.999	0.001
Y	0.996	0.004

Performance of random forest with cutoff on Y = 0.4, 0.6, 0.8



Performance of random forest with cutoff on Y = 0.4

Confusion Matrix

Training

Actual	Predicted Count	
TARGET	N	Y
N	1277	0
Y	100	363

Validation

Actual	Predicted Count	
TARGET	N	Y
N	540	4
Y	211	9

Decision Matrix

Training

Actual	Decision Count	
TARGET	N	Y
N	1275	2
Y	20	443

Actual	Decision Rate	
TARGET	N	Y
N	0.998	0.002
Y	0.043	0.957

Misclassification Rate
0.0126

Validation

Actual	Decision Count	
TARGET	N	Y
N	500	44
Y	177	43

Actual	Decision Rate	
TARGET	N	Y
N	0.919	0.081
Y	0.805	0.195

Misclassification Rate
0.2893

Specified Profit Matrix

	Decision	
Actual	N	Y
N	0	-0.667
Y	-1	0

Sensitivity (N classified as N) = 99.9%

Specificity (Y classified as Y) = 95.7%

Performance of random forest with cutoff on Y = 0.6

Confusion Matrix

Training

Actual	Predicted Count	
TARGET	N	Y
N	1277	0
Y	100	363

Validation

Actual	Predicted Count	
TARGET	N	Y
N	540	4
Y	211	9

Decision Matrix

Training

Actual	Decision Count	
TARGET	N	Y
N	1277	0
Y	278	185

Actual	Decision Rate	
TARGET	N	Y
N	1.000	0.000
Y	0.600	0.400

Misclassification Rate
0.1598

Validation

Actual	Decision Count	
TARGET	N	Y
N	544	0
Y	218	2

Actual	Decision Rate	
TARGET	N	Y
N	1.000	0.000
Y	0.991	0.009

Misclassification Rate
0.2853

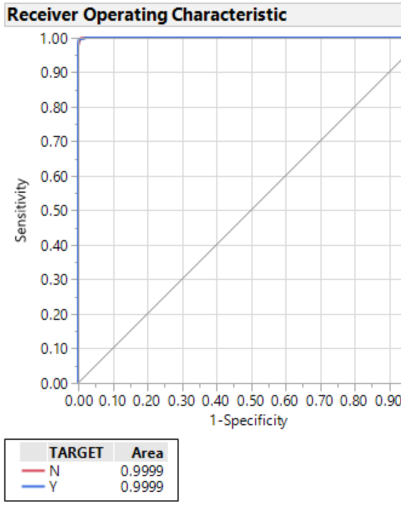
Specified Profit Matrix

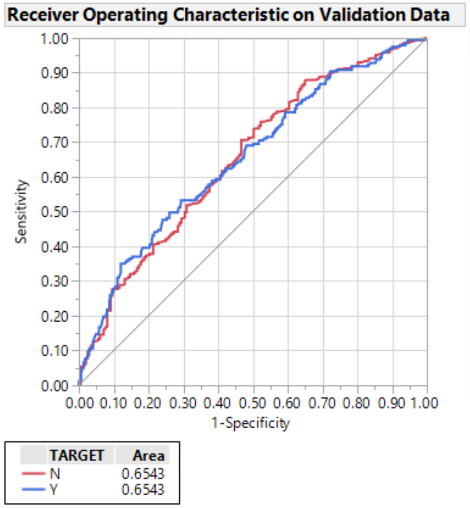
	Decision	
Actual	N	Y
N	0	-1.5
Y	-1	0

Sensitivity (N classified as N) = 100%

Specificity (Y classified as Y) = 40%

Performance of random forest with cutoff on Y = 0.8





Confusion Matrix

Training

Actual	Predicted Count	
TARGET	N	Y
N	1306	0
Y	74	412

Validation

Actual	Predicted Count	
TARGET	N	Y
N	510	5
Y	187	10

Decision Matrix

Training

Actual	Decision Count	
TARGET	N	Y
N	1306	0
Y	486	0

Actual	Decision Rate	
TARGET	N	Y
N	1.000	0.000
Y	1.000	0.000

Misclassification Rate
0.2712

Validation

Actual	Decision Count	
TARGET	N	Y
N	515	0
Y	197	0

Actual	Decision Rate	
TARGET	N	Y
N	1.000	0.000
Y	1.000	0.000

Misclassification Rate
0.2767

Specified Profit Matrix

Actual	Decision	
	N	Y
N	0	-4
Y	-1	0

Sensitivity (N classified as N) = 100%

Specificity (Y classified as Y) = 0%

KPA

The NPL case study

- Random Forest with informative missing data imputation
- Number of trees in forest =100 with 10-2000 splits and no multithreading
- Validation set consisting of 30% randomly selected cased
- With Cutoff=0.5 one gets Sensitivity=100% and Specificity=85%
- Sensitivity of cutoff needs to be evaluated with economic parameters

The NPL case study

- We have not considered an option of “undecided”
- We have focused on individual classifications not ranking of cases for prioritizing action items
- Sensitivity and specificity are performance measures from the bank’s perspective (not misclassification rates)
- Missing data and outliers should be investigated
- Clustering and event driven predictive analytics for risk mitigation can be considered
- Costs in profit matrix need to be justified by bank
- The economic impact of the model needs to be evaluated
- Customers could be segmented with different models being applied to different segments

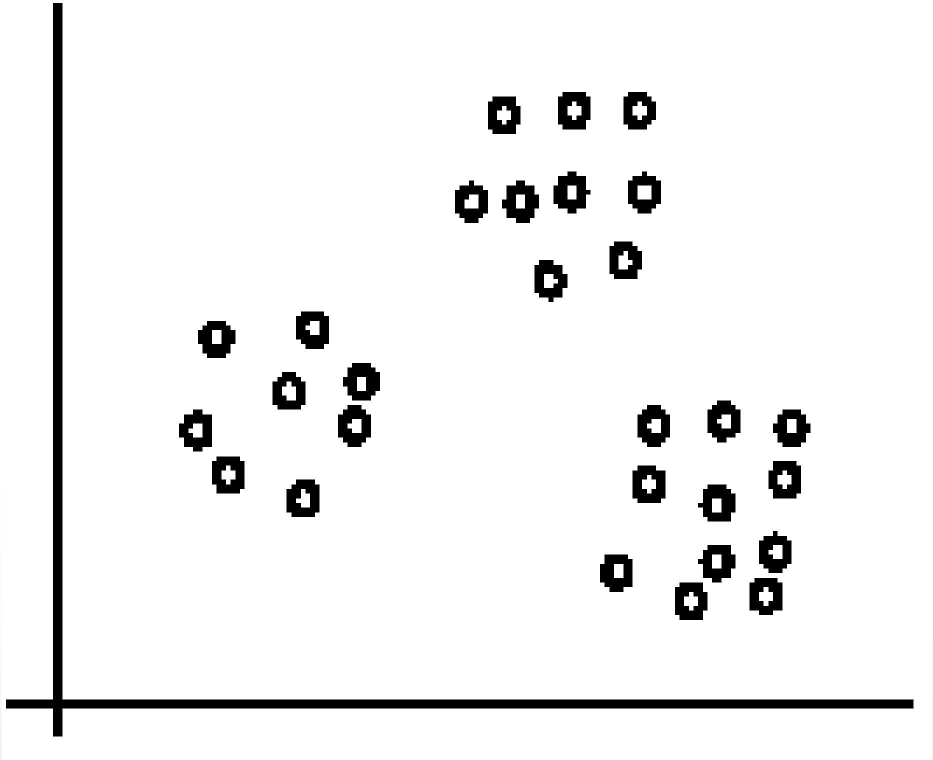
Unsupervised Learning

Clustering

- Clustering is a technique for finding similarity groups in data, called **clusters**. I.e.,
  - it groups data instances that are similar to (near) each other in one cluster and data instances that are very different (far away) from each other into different clusters.
- Clustering is often called an **unsupervised learning** task as no class values denoting an *a priori* grouping of the data instances are given, which is the case in supervised learning.

An illustration

- The data set has three natural groups of data points, i.e., 3 natural clusters.



Aspects of clustering

- A clustering algorithm
  - Partitional clustering
  - Hierarchical clustering
  - …
- A distance (similarity, or dissimilarity) function
- Clustering quality
  - Inter-clusters distance $\Rightarrow$ maximized
  - Intra-clusters distance $\Rightarrow$ minimized
- The quality of a clustering result depends on the algorithm, the distance function, and the application.

K-means clustering

- K-means is a partitional clustering algorithm
- Let the set of data points (or instances) *D* be

  $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$,

  where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$ is a vector in a real-valued space $X \subseteq R^r$, and r is the number of attributes (dimensions) in the data.
- The *k*-means algorithm partitions the given data into *k* clusters.
  - Each cluster has a cluster **center**, called **centroid**.
  - *k* is specified by the user

K-means algorithm

- Given *k*, the *k-means* algorithm works as follows:
  1) Randomly choose *k* data points (seeds) to be the initial centroids, cluster centers
  2) Assign each data point to the closest centroid
  3) Re-compute the centroids using the current cluster memberships.
  4) If a convergence criterion is not met, go to 2).

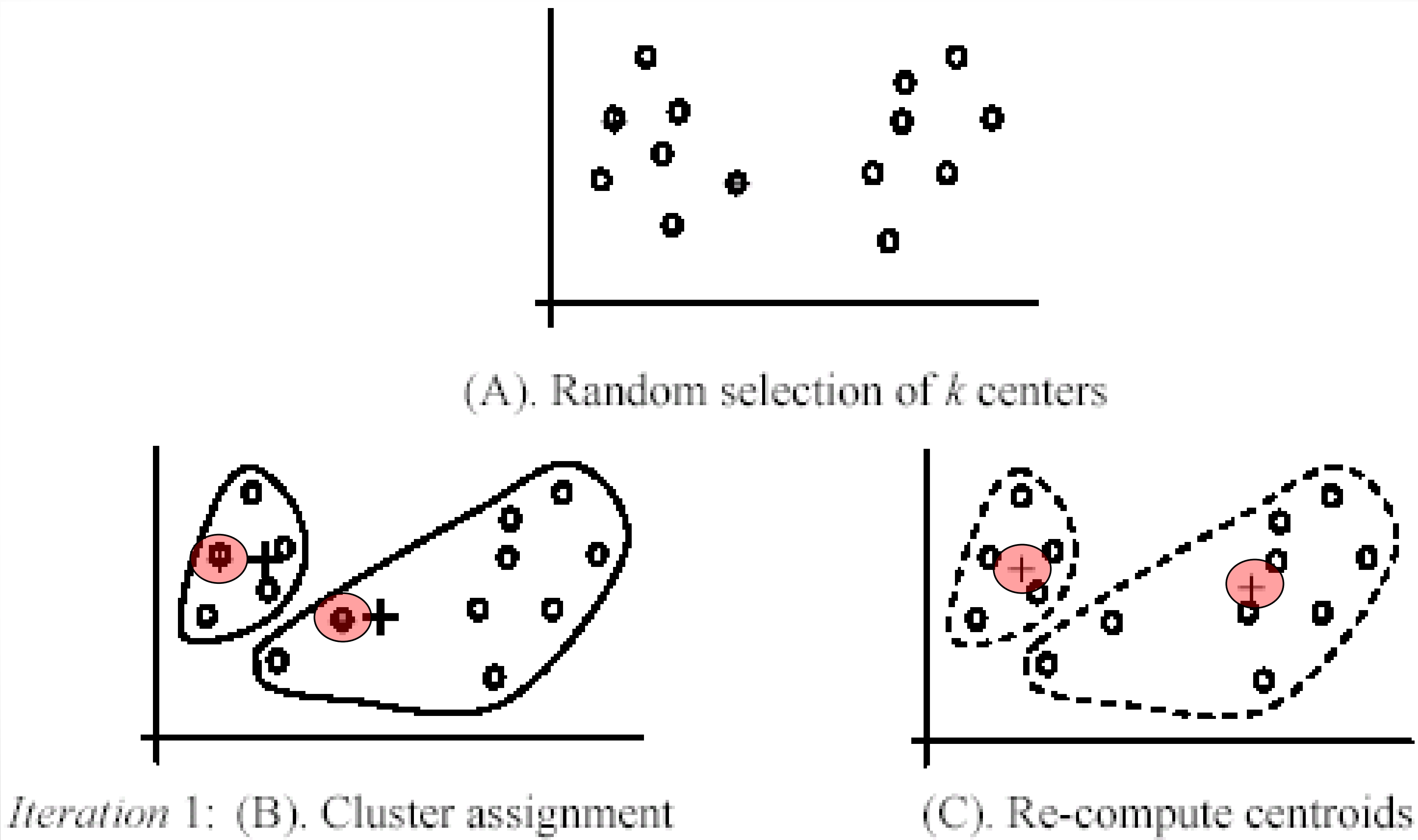
Stopping/convergence criterion

1. no (or minimum) re-assignments of data points to different clusters,
2. no (or minimum) change of centroids, or
3. minimum decrease in the **sum of squared error** (SSE),

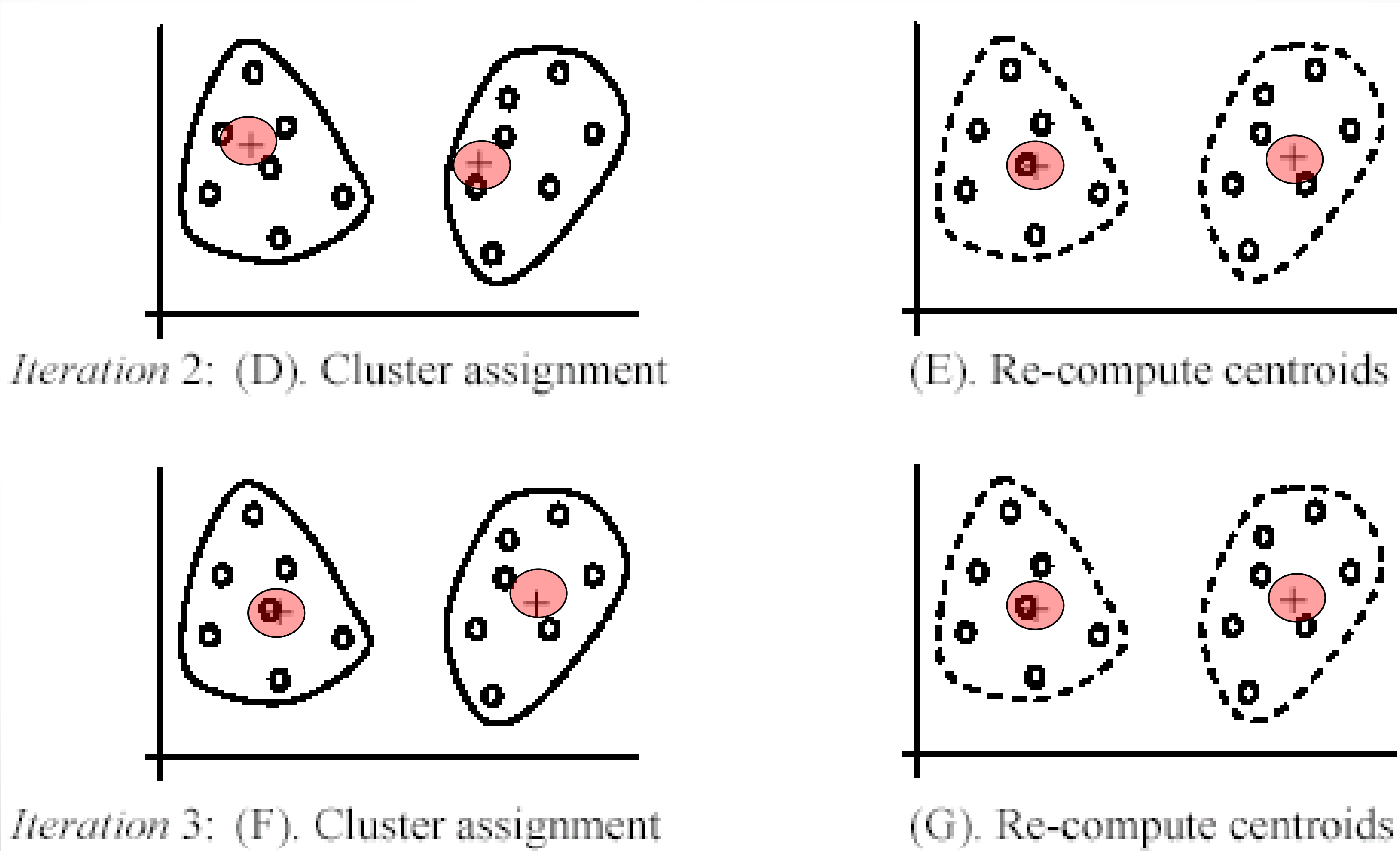
$$SSE = \sum_{j=1}^{k} \sum_{\mathbf{x} \in C_j} dist(\mathbf{x}, \mathbf{m}_j)^2 \quad (1)$$

– C_i is the jth cluster, $\mathbf{m}_j$ is the centroid of cluster C_j (the mean vector of all the data points in C_j), and $dist(\mathbf{x}, \mathbf{m}_j)$ is the distance between data point $\mathbf{x}$ and centroid $\mathbf{m}_j$.

An example


(A). Random selection of k centers
Iteration 1: (B). Cluster assignment
(C). Re-compute centroids

An example (cont …)



Strengths of K-means

- Strengths:
  - Simple: easy to understand and to implement
  - Efficient: Time complexity: $O(tkn)$,

    where n is the number of data points,

    k is the number of clusters, and

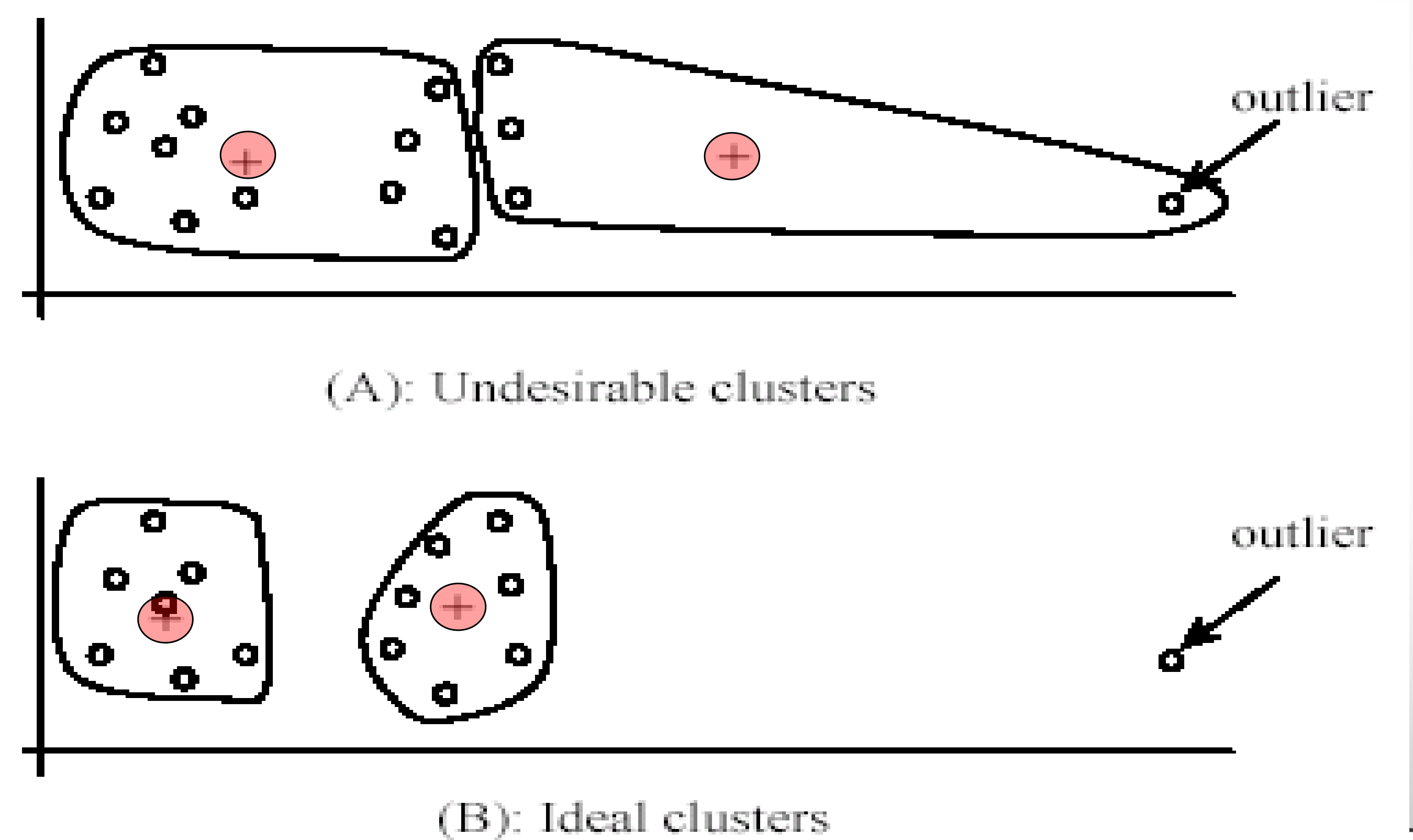
    t is the number of iterations.
  - Since both k and t are small. k-means is considered a linear algorithm.
- K-means is the most popular clustering algorithm.
- Note that: it terminates at a local optimum if SSE is used. The global optimum is hard to find due to complexity.

Weaknesses of K-means

- The algorithm is only applicable if the mean is defined.
  – For categorical data, *k*-mode - the centroid is represented by most frequent values.
- The user needs to specify *k*.
- The algorithm is sensitive to **outliers**
  – Outliers are data points that are very far away from other data points.
  – Outliers could be errors in the data recording or some special data points with very different values.

KPA

Weaknesses of K-means: Outliers

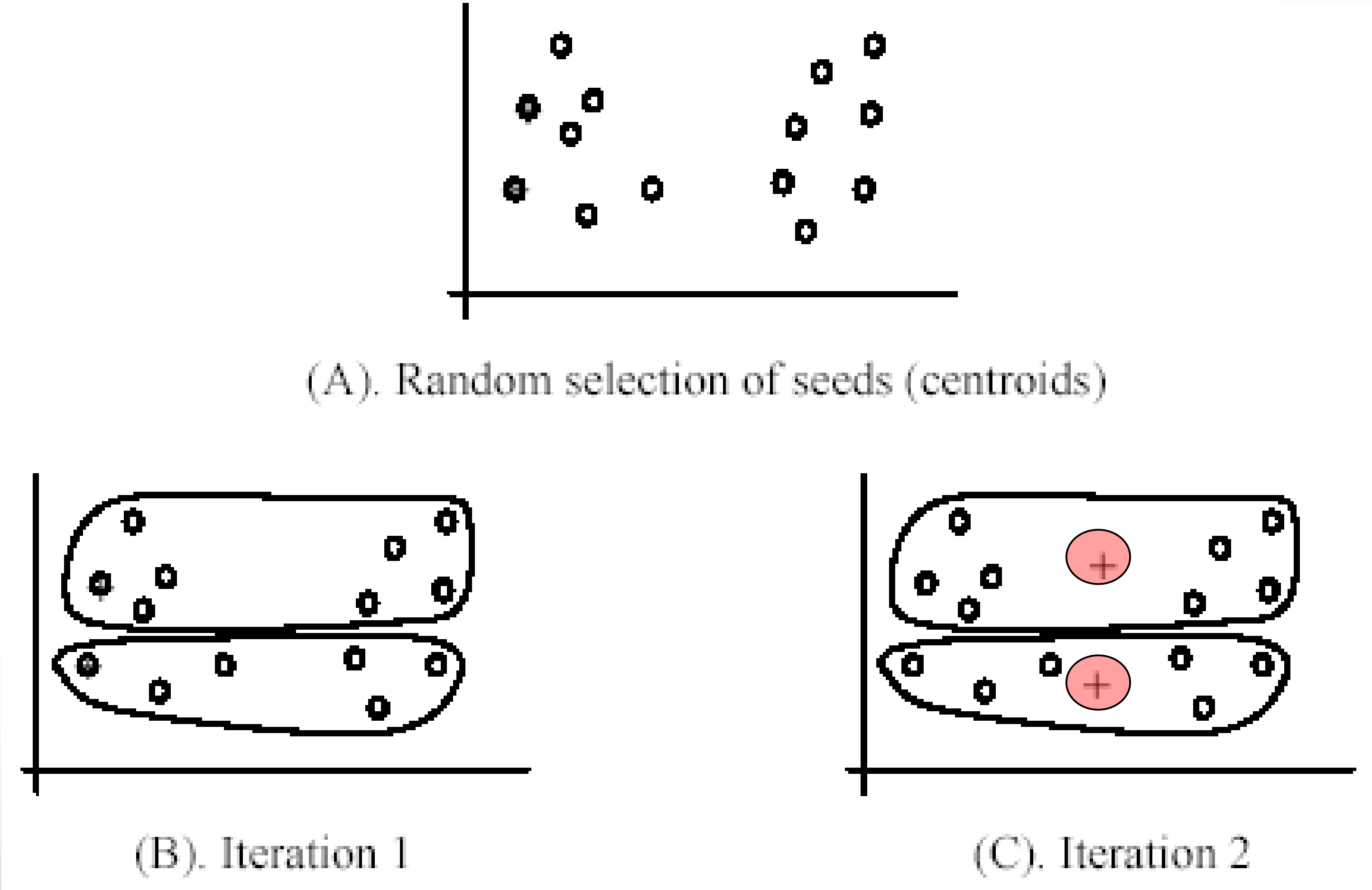


Weaknesses of K-means: Outliers

- One method is to remove some data points in the clustering process that are much further away from the centroids than other data points.
  - Monitor possible outliers over a few iterations and then decide to remove them.
- Another method is to perform random sampling. Since in sampling we only choose a small subset of the data points, the chance of selecting an outlier is very small.
  - Assign the rest of the data points to the clusters by distance or similarity comparison, or classification

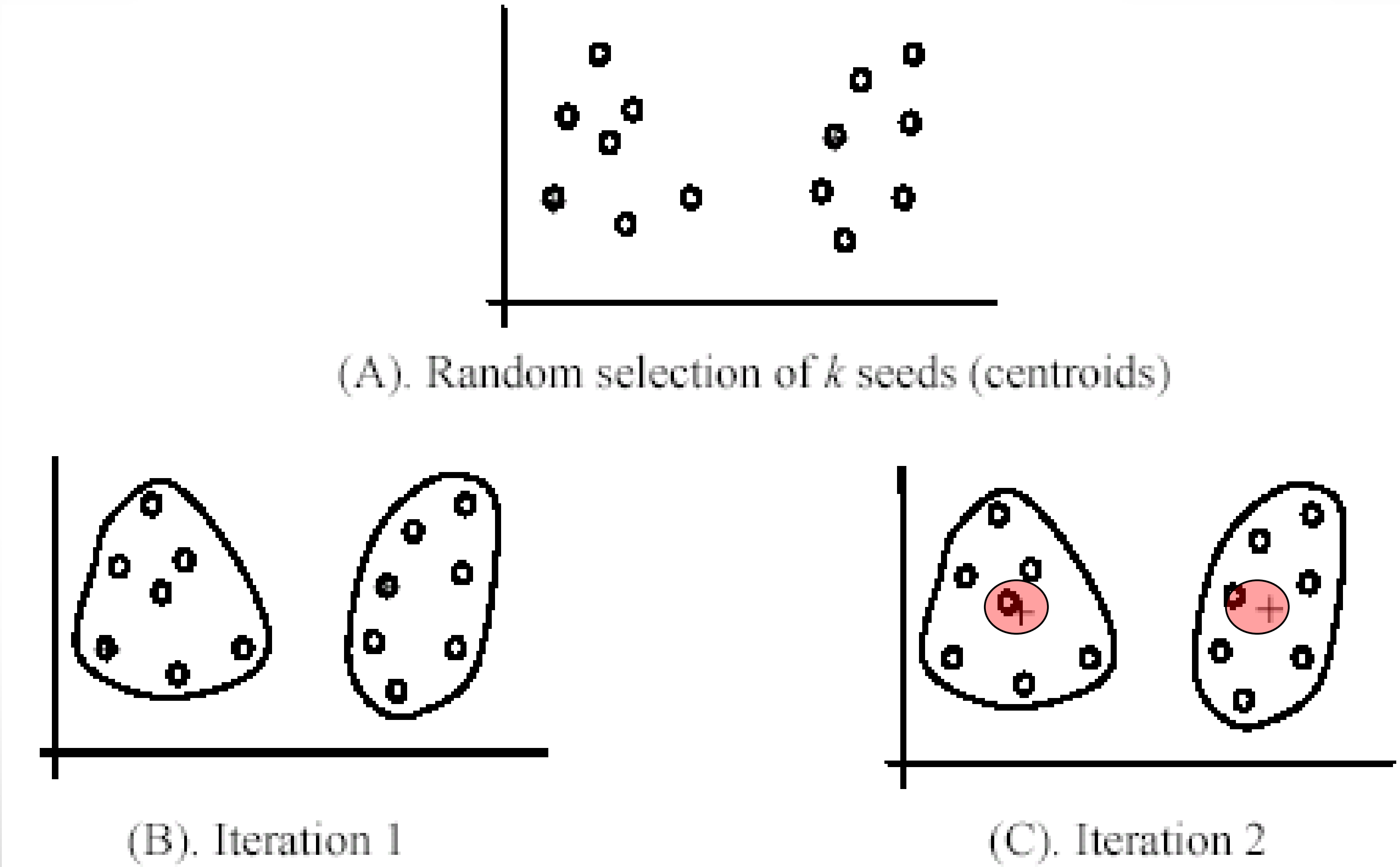
Weaknesses of K-means (cont …)

- The algorithm is sensitive to initial seeds.



Weaknesses of K-means (cont …)

- If we use different seeds: good results



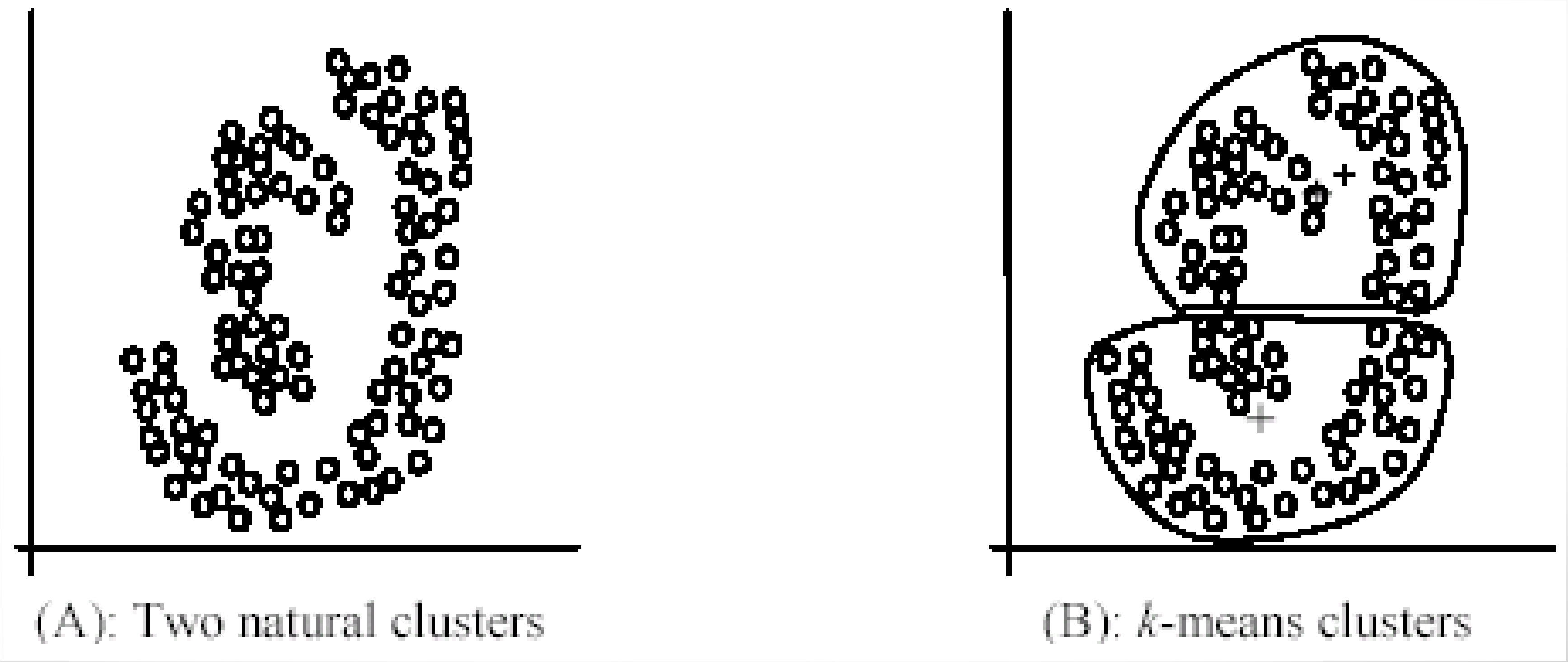
(A). Random selection of *k* seeds (centroids)

(B). Iteration 1

(C). Iteration 2

Weaknesses of K-means (cont …)

- The *k*-means algorithm is not suitable for discovering clusters that are not hyper-ellipsoids (or hyper-spheres).



K-means summary

- Despite weaknesses, *k*-means is still the most popular algorithm due to its simplicity, efficiency and
  - other clustering algorithms have their own lists of weaknesses.
- No clear evidence that any other clustering algorithm performs better in general
  - although they may be more suitable for some specific types of data or applications.
- Comparing different clustering algorithms is a difficult task. No one knows the correct clusters!

Common ways to represent clusters

- Use the centroid of each cluster to represent the cluster.
  - compute the radius and
  - standard deviation of the cluster to determine its spread in each dimension

  - The centroid representation alone works well if the clusters are of the hyper-spherical shape.
  - If clusters are elongated or are of other shapes, centroids are not sufficient

Hierarchical methods

Agglomerative Methods

– Begin with n-clusters (each record its own cluster)

– Keep joining records into clusters until one cluster is left (the entire data set)

– Most popular

Divisive Methods

– Start with one all-inclusive cluster

– Repeatedly divide into smaller clusters

Distance between two records

Euclidean Distance is most popular:

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

Normalizing

Problem: Raw distance measures are highly influenced by scale of measurements

Solution: normalize (standardize) the data first

- Subtract mean, divide by std. deviation
- Also called **z-scores**

Other distance measures

- Correlation-based similarity
- Statistical distance (Mahalanobis)
- Manhattan distance (absolute differences)
- Maximum coordinate distance
- Gower's similarity (for mixed variable types: continuous & categorical)

Minimum distance (Cluster A to Cluster B)

- Also called **single linkage**

- Distance between two clusters is the distance between the pair of records A_i and B_j that are closest

Maximum distance (Cluster A to Cluster B)

- Also called **complete linkage**
- Distance between two clusters is the distance between the pair of records A_i and B_j that are farthest from each other

Average distance

- Also called **average linkage**

- Distance between two clusters is the average of all possible pair-wise distances

Centroid distance

- Distance between two clusters is the distance between the two cluster centroids
- Centroid is the vector of variable averages for all records in a cluster

Ward’s method

- Considers loss of information when observations are clustered together
- Uses error sum of squares (ESS) to measure the difference between observations and the centroid
- The *Fast Ward* method in JMP is more efficient, and is used automatically for large data sets

The Hierarchical Clustering (using agglomerative method)

Steps:

1. Start with *n* clusters (each record is its own cluster)
2. Merge two closest records into one cluster
3. At each successive step, the two clusters closest to each other are merged

Dendrogram, from left to right, illustrates the process

Interpreting clusters

Goal: obtain meaningful and useful clusters

Caveats:

(1) Random chance can often produce apparent clusters

(2) Different cluster methods produce different results

Solutions:

- Obtain summary statistics
- Also review clusters in terms of variables **not** used in clustering
- Label the cluster (e.g. clustering of financial firms in 2008 might yield label like “midsize, sub-prime loser”)

Desirable cluster features

Stability

- Are clusters and cluster assignments sensitive to slight changes in inputs?
- Are cluster assignments in partition B similar to partition A?

Separation

- check ratio of between-cluster variation to within-cluster variation (higher is better)

K-Means clustering algorithm

1. Choose # of clusters desired, *k*
2. Start with a partition into k clusters

   Often based on random selection of k centroids
3. At each step, move each record to cluster with closest centroid
4. Recompute centroids, repeat step 3
5. Stop when moving records increases within-cluster dispersion

K-means algorithm: choosing k and initial partitioning

Choose *k* based on the how results will be used

- e.g., “How many market segments do we want?”

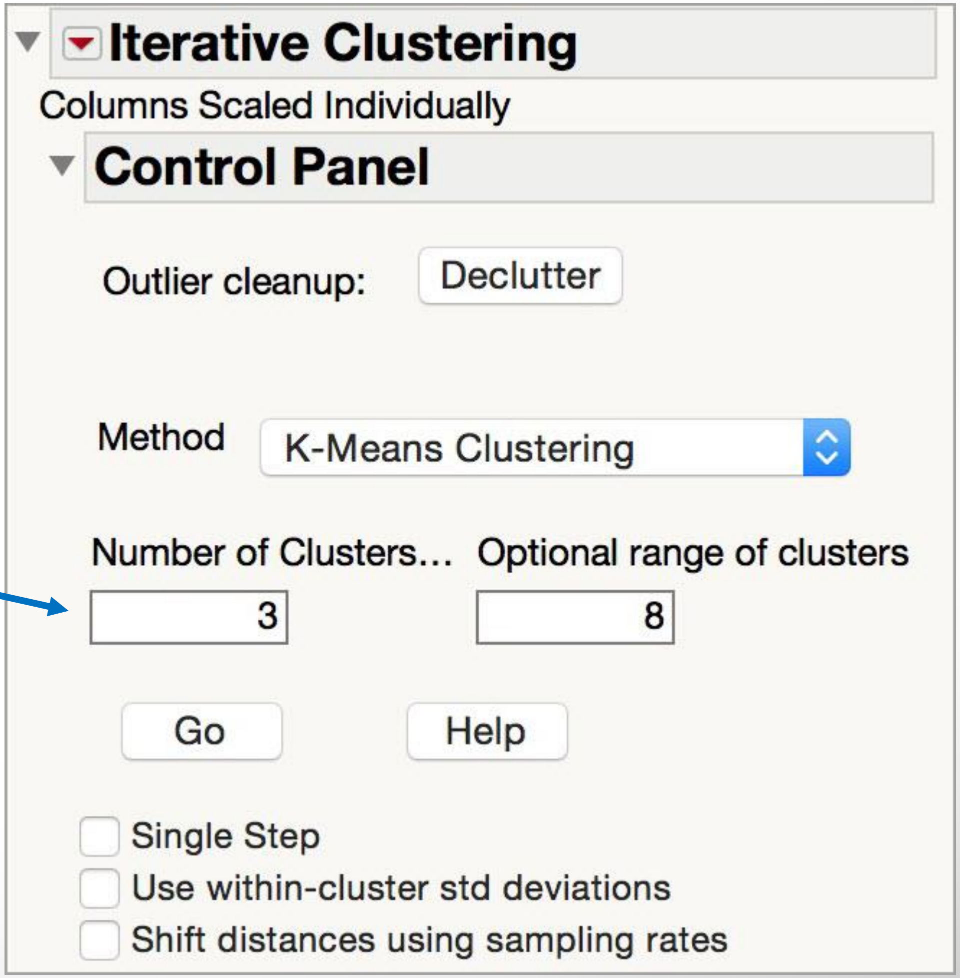
Also experiment with slightly different *k*’s

Initial partition into clusters can be random, or based on domain knowledge

- If random partition, repeat the process with different random partitions

K-means dialog in JMP

Try a range of values for k



K-means output (k = 6)

K Means NCluster=6

Columns Scaled Individually

Cluster Summary

Cluster	Count
1	1
2	6
3	6
4	5
5	1
6	3

Step	Criterion
2	0

Cluster Means

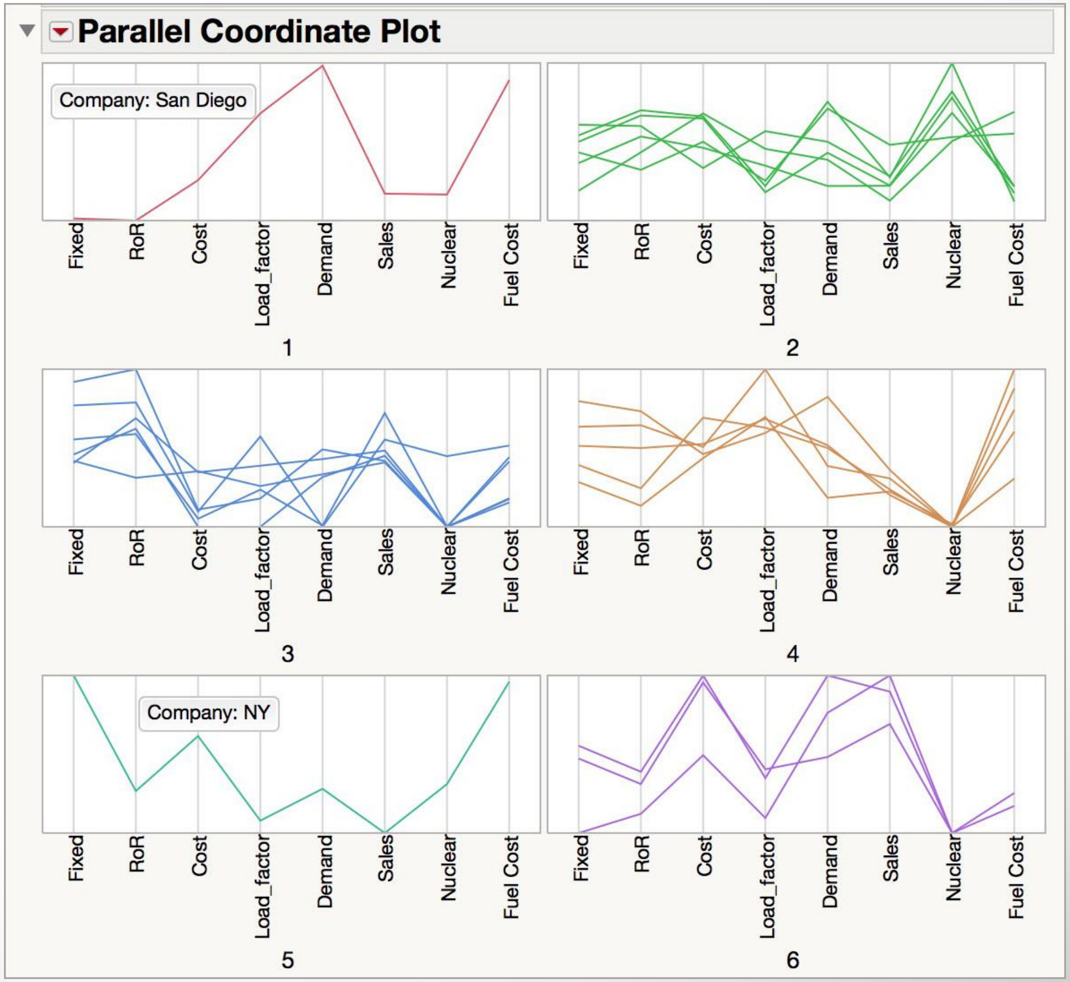
Cluster	Fixed	RoR	Cost	Load_factor	Demand	Sales	Nuclear	Fuel Cost
1	0.76	6.4	136	61.9	9	5714	8.3	1.92
2	1.075	11.2833333	181.333333	55.8	3.5	7087.5	36.1166667	0.90216667
3	1.185	12.4	120.833333	54.65	0.8	10456	3.75	0.8765
4	1.138	10.46	177.8	62.64	3.3	7064	0.18	1.5854
5	1.49	8.8	192	51.2	1	3300	15.6	2.044
6	1.00333333	8.86666667	223.333333	54.8333333	6.33333333	15504.6667	0	0.56566667

Cluster Standard Deviations

Cluster	Fixed	RoR	Cost	Load_factor	Demand	Sales	Nuclear	Fuel Cost
1	0	0	0	0	0	0	0	0
2	0.10045729	1.18520978	19.694895	2.44131112	2.11108187	1523.19946	8.59794872	0.38490147
3	0.14244882	1.87705443	21.6749984	3.15052906	2.19012937	1523.20736	8.38525492	0.26527329
4	0.13332667	2.06455806	14.0057131	2.56561883	2.37402612	835.435216	0.36	0.43376196
5	0	0	0	0	0	0	0	0
6	0.18080069	1.00774776	35.7055862	2.39211668	2.4115463	1812.47719	0	0.19128397

Visualizing clusters

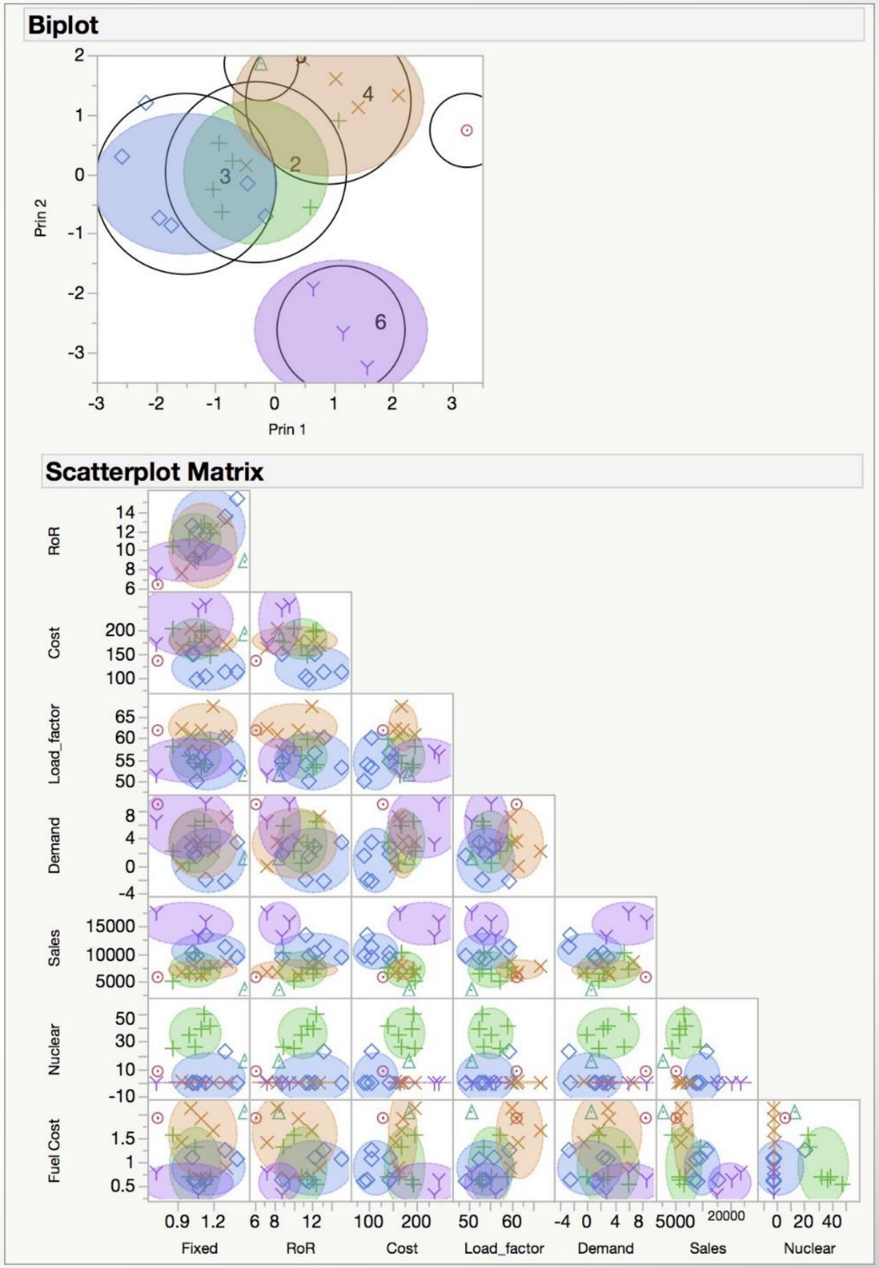
- Parallel Plot shows the number of records per cluster, and the profile of the clusters across the variables



KPA

Visualizing clusters

- Biplot shows separation and overlap of the clusters
- Scatterplot matrix shows separation of the clusters across the variables



Clustering overview

- Cluster analysis is an exploratory tool
- It is useful only when it produces **meaningful** clusters
- **Hierarchical** clustering gives visual representation of different levels of clustering
- **Non-hierarchical** clustering is computationally cheap and more stable (good with larger data sets); requires user to set *k*
- Can use both methods
- Be wary of chance results; data may not have definitive “real” clusters

An Introduction to Text Mining with examples from an Annual Report

KPA

What is Text Mining?

- Text mining: semi-automated process of detecting patterns (useful information and knowledge) from large amounts of *unstructured* data sources
- Text analytics: methods used for intelligent analyses of textual data; a larger set of activities around inference steps of discovering information, grouping documents, summarizing information, etc.
- In order to analyze text in a systematic and structured way, we first need to develop a numerical representation of the text.
- Obviously, there is not a unique solution to this problem. The appropriate mapping of text->numbers depends on the goal of the study.

Text Mining Flow

Define Problem Statement

- Determine clear study objectives and end-state
- Identify relevant data sources to answer research questions

Collect and Extract Data

- Organize files, scrape internet with web crawling and social media tools
- Extract text from disparate file types (pptx, doc, txt, pdf, html)
- Strip off code, figures, extraneous characters

Process and Filter Text

- Clean manually with character functions, queries, filters, R&R
- Remove punctuation, numbers, stop words
- Stem and tokenize text, change to lowercase, identify multiwords.

Transform Text

- Create document term matrix (DTM)
- Weight matrix based on analysis objectives
- Use Singular Value Decomposition to get structured data

Text Mining Exploration

- Discover topics and common themes
- Group like documents and words
- Subset documents and link concepts

Predictive Analytics

- Combine with structured data
- Visualize exploitable patterns
- Understand sentiments and trends

A Simple Example

Car Accidents
Slid on ice into a curb.
Driving too fast in a dust storm, hit the curb.
Low-budget tires failed after bumping curb.

Bag of Words Approach

- Using a "bag of words" approach, we disregard the ordering of the words in each document as well as their grammatical properties.
- While this may seem simplistic, it has been shown to give excellent results in many applications.

Vocabulary

- Document: a string of words.
- Corpus: a collection of documents.
- In the text mining literature, “words,” “terms,” and “tokens” all describe roughly the same idea. There are some subtleties to their use: we will use them interchangeably to mean words that have been extracted from a document and processed.

Processing Text

- Within each document, we will first
  - Isolate individual words
  - Remove punctuation
  - Normalize case (convert all characters to lowercase)
  - Remove numbers
- Later, we will discuss further processing of the words.

Natural Language Processing

- After extracting the tokens from a document, it is typically useful to
  - Remove stopwords (most frequent words).
  - Stem the text.
  - Remove words with character length below a minimum or above a maximum.
  - Remove words that appear in only a few documents (most infrequent words).

Isolate Words

Document 1	Document 2	Document 3
Slid	Driving	Low-budget
on	too	tire
ice	fast	failed
into	in	after
a	a	bumping
curb.	dust	curb.
	storm,	
	hit	
	the	
	curb.	

Notice that punctuation is concatenated to adjacent terms.

Remove Punctuation

Document 1	Document 2	Document 3
Slid	Driving	Lowbudget
on	too	tire
ice	fast	failed
into	in	after
a	a	bumping
curb	dust	curb
	storm	
	hit	
	the	
	curb	

KPA

Normalize Case

Document 1	Document 2	Document 3
slid	driving	lowbudget
on	too	tire
ice	fast	failed
into	in	after
a	a	bumping
curb	dust	curb
	storm	
	hit	
	the	
	curb	

Remove Stopwords

Document 1	Document 2	Document 3
slid	driving	lowbudget
ice	fast	tire
curb	dust	failed
	storm	bumping
	hit	curb
	curb	

KPA

Stem Text

Document 1	Document 2	Document 3
slid	drive	lowbudget
ice	fast	tire
curb	dust	fail
	storm	bump
	hit	curb
	curb	

KPA

Representing Text with Numbers

- To find clusters of documents or to use the information present in the documents in a predictive model, we need a numerical representation of the text.
- Using the bag of words approach, we create a document term matrix (DTM). Each document is represented by a row, and each token is represented by a column. The components of the matrix represent how many times each token appears in each document.

Document Term Matrix

Doc	bump	curb	drive	dust	fail	fast	hit	ice	lowbudget	slid	storm	tire
1	0	1	0	0	0	0	0	1	0	1	0	0
2	0	1	1	1	0	1	1	0	0	0	1	0
3	1	1	0	0	1	0	0	0	1	0	0	1

Properties of the DTM

- The DTM will typically be very sparse (most entries are 0).
- Even for modestly sized applications, the full DTM will be too large to hold in memory.
- Since most entries are 0, multiplying the matrix results in several multiplications by 0, which could be omitted.
- Special software and algorithms are available for storing and manipulating sparse matrices.

Transformations of the DTM

- Various transformations of the term-frequency counts in the DTM have been found to be useful.

Transformations of the DTM

- Frequency (local) weights
  - Binary: Useful if there is a lot of variance in the lengths of the documents in the corpus.
  - Ternary/Frequency: Some researchers have found that distinguishing between terms that appear only once in a document vs. those that appear multiple time can improve results.
  - Log: Dampens the presence of high counts in longer documents without sacrificing as much information as the binary weighting scheme.

Transformations of the DTM

- Term (global) weights
  - Term Frequency - Inverse Document Frequency (tf-idf)
    - Shrinks the weight of terms that appear in many documents while also inflating the weight of terms that appear in only a few documents
    - Sometimes makes interpretation of results more difficult, but can give better predictive performance. In practice, it is best to try different weighting schemes: there is no need to pick only one!

Inverse Document Frequency

- idf down-weights terms that appear in many documents. The idf for term *t* is

$$idf_t = \log_2\left(\frac{D}{df_t}\right)$$

- *D* is the number of documents in the corpus.
- df_t is the number of documents containing term *t.*
- If a term appears in every document, its idf is 0.

tf-idf

Doc	bump	curb	drive	dust	fail	fast	hit	ice	lowbudget	slid	storm	tire
1	0	0	0	0	0	0	0	1.585	0	1.585	0	0
2	0	0	1.585	1.585	0	1.585	1.585	0	0	0	1.585	0
3	1.585	0	0	0	1.585	0	0	0	1.585	0	0	1.585

Transformations of the DTM

- Normalizing each document
  - The term frequency weights in each document may be normalized so that the sum of each document vector is 1. This is done by dividing the term counts in each document (each row of the DTM) by the total number of words in each document (the row sums of the DTM).
  - This can be useful when the documents are of different lengths. An illustration of how this can help: if a document D’ is created by pasting two copies of a document D together, D and D’ will be identical after normalization.

Normalized Term-Frequency Document Term Matrix

Doc	bump	curb	drive	dust	fail	fast	hit	ice	low bud get	slid	storm	tire
1	0	0.333	0	0	0	0	0	0.333	0	0.333	0	0
2	0	0.167	0.167	0.167	0	0.167	0.167	0	0	0	0.167	0
3	0.2	0.2	0	0	0.2	0	0	0	0.2	0	0	0.2

Frequency Weighting Summary

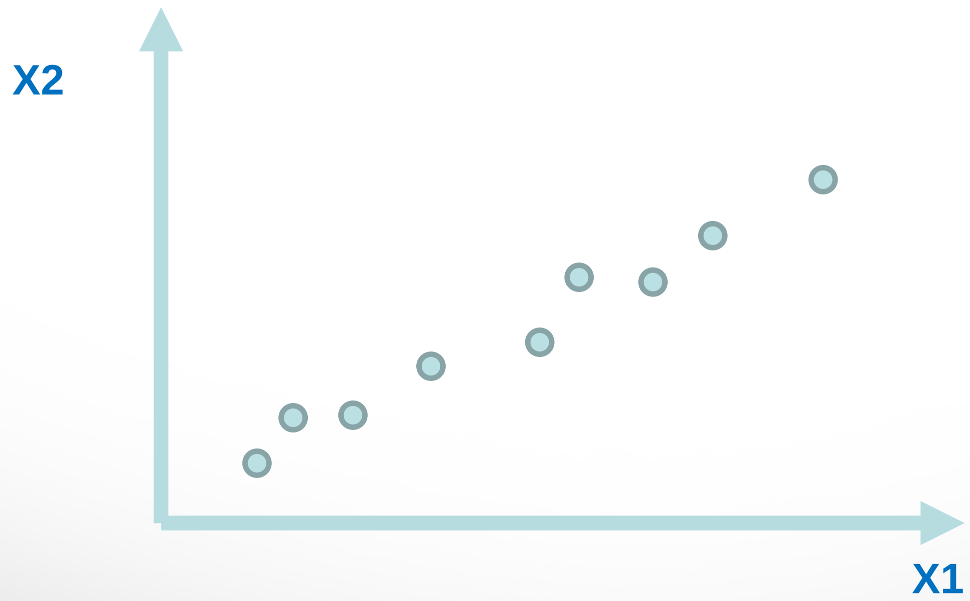
- There is no universally best weighting: take time to try different options.

Singular Value Decomposition

- The reduced-rank singular value decomposition (SVD) provides us with a dimensionality reduction technique.
- The SVD reduces the DTM to a (dense) matrix with fewer columns. The new (orthogonal) columns are linear combinations of the rows in the original DTM, selected to preserve as much of the structure of the original DTM as possible.

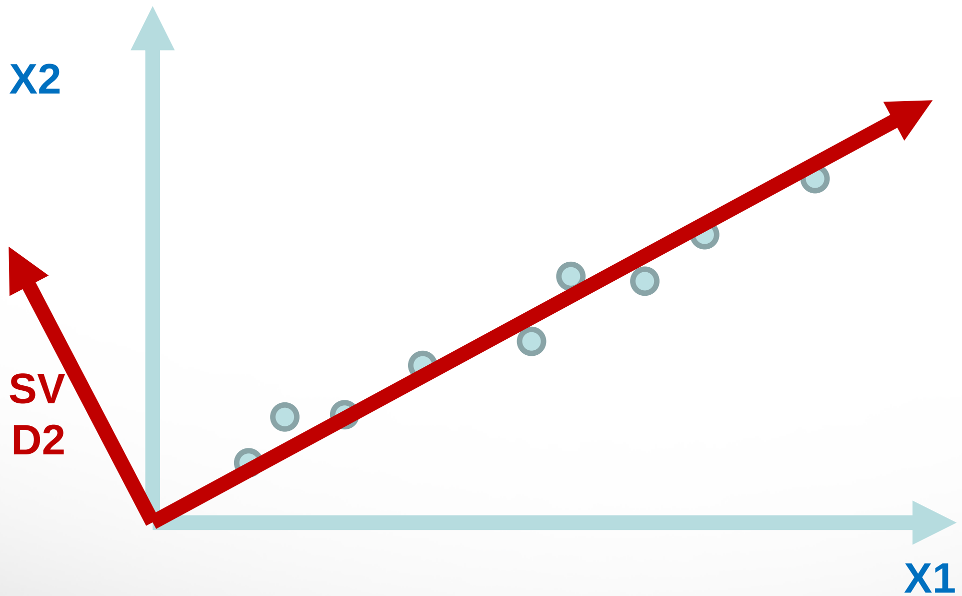
SVD Example

X1 and X2 describe the location of these points.
However, they appear to fall mostly along a line.



SVD Example

Roughly, the SVD finds a new set of orthogonal basis vectors such that each additional dimension accounts for as much of the variation of the data as possible.



Singular Value Decomposition

- For a DTM *X*, the SVD factorization is

$$X \approx UDV^t,$$

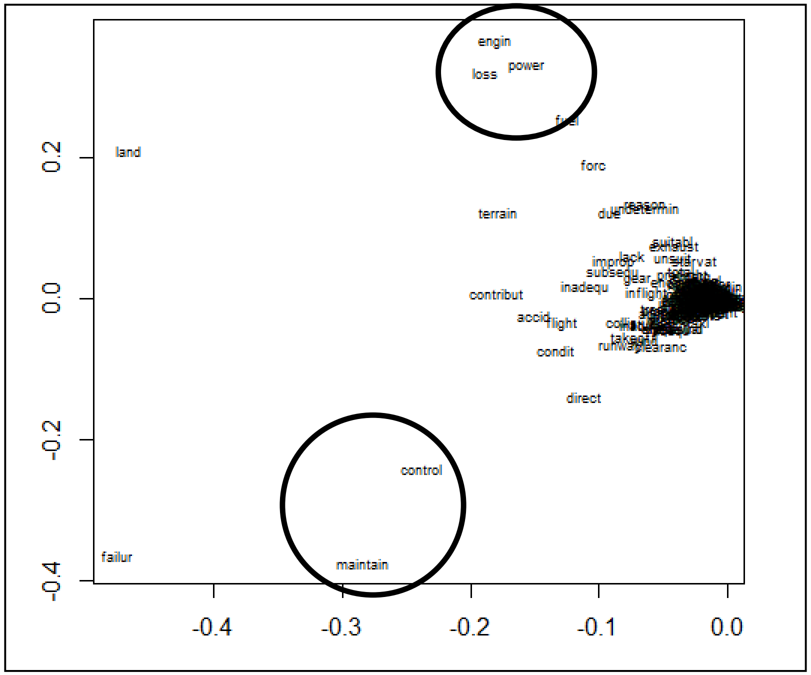
where

- *U* is a dense *d* by *s* orthogonal matrix ***U* gives us a new rank-reduced description of *documents***
- *D* is a diagonal matrix with nonnegative entries (the singular values).
- V^t is a dense *s* by *w* orthogonal matrix, where *s* is the rank of the SVD factorization (*s=1,…,min(d,w)*), and the superscript *t* indicates "transpose." ***V* gives us a new rank-reduced description of *terms*.**
- *d* is the number of documents
- *w* is the number of words
- *s* is the rank of the SVD factorization (*s=1,…,min(d,w)*).

Latent Semantic Analysis

- In natural language processing, the use of a rank-reduced SVD is referred to as latent semantic analysis (LSA).
- A popular LSA technique is to plot the corpus dictionary using the first two vectors resulting from the SVD.
- Similar words (words that either appear frequently in the same documents, or appear frequently with common sets of words throughout the corpus) are plotted together, and a rough interpretation can often be assigned to dimensions appearing in the plot.

SVD1 vs. SVD2



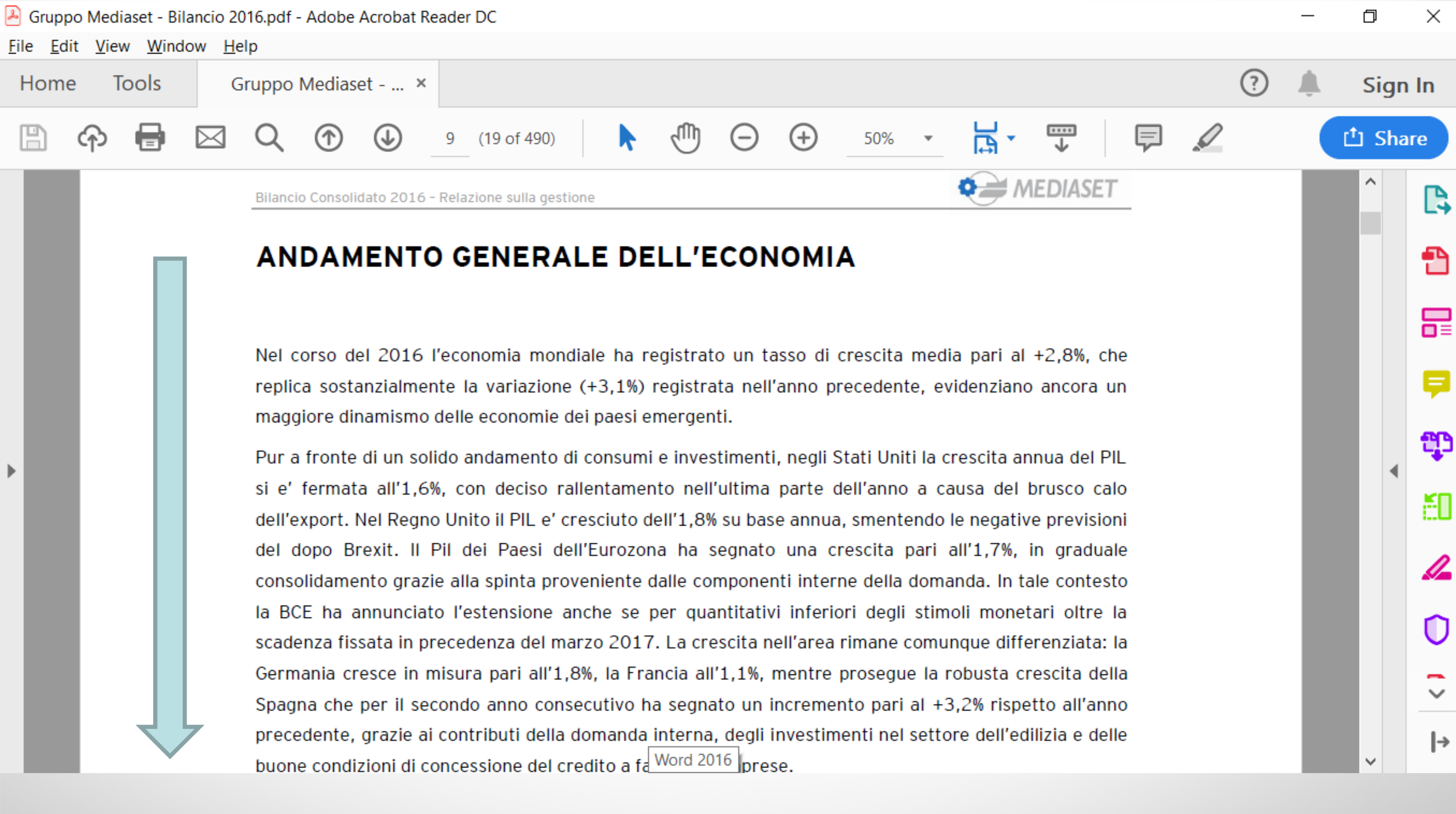
The words appearing close to each other appear together frequently (or appear independently with a common set of words) in documents in the corpus. We also look for themes describing the spread of terms in this plot (latent semantic analysis).

Clustering

- Once we have produced either a DTM or an SVD of a DTM, we may use the resulting numeric columns with clustering algorithms to answer questions such as
  - Which groups of documents are most similar?
  - Which documents are most similar to a particular document?
  - Which groups of terms tend to appear either together in the same documents or together with the same words?
  - Which terms are most similar to a particular term?
  - Are certain clusters of documents more strongly related to other variables (e.g. income, cost, fraudulent activity) than other clusters?

An example


Gruppo Mediaset
Bilancio Consolidato 2016
Relazione degli Amministratori sulla Gestione
MEDIASET



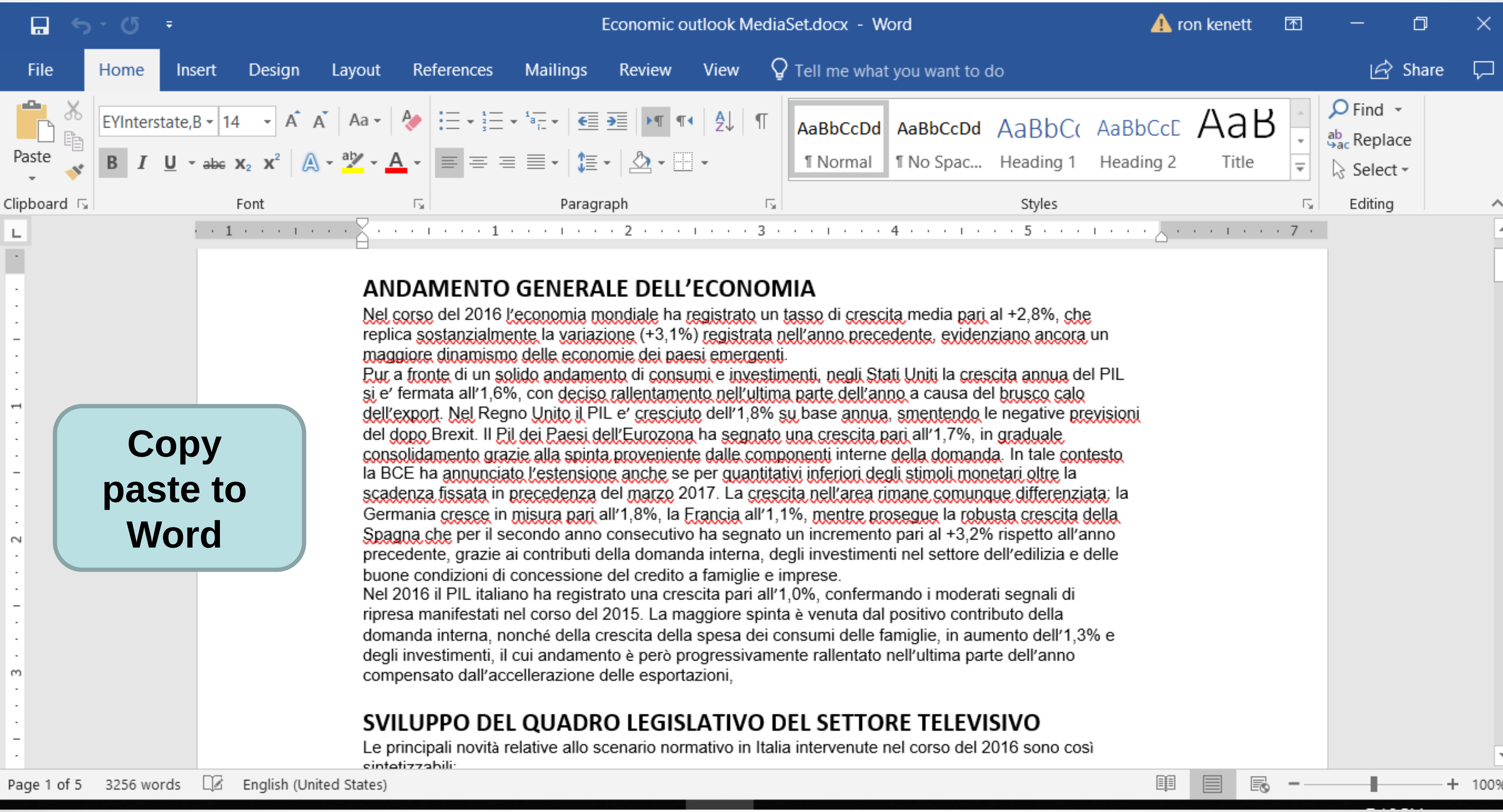
Bilancio Consolidato 2016 – Relazione sulla gestione

MEDIASET

ANDAMENTO GENERALE DELL'ECONOMIA

Nel corso del 2016 l'economia mondiale ha registrato un tasso di crescita media pari al +2,8%, che replica sostanzialmente la variazione (+3,1%) registrata nell'anno precedente, evidenziano ancora un maggiore dinamismo delle economie dei paesi emergenti.

Pur a fronte di un solido andamento di consumi e investimenti, negli Stati Uniti la crescita annua del PIL si e' fermata all'1,6%, con deciso rallentamento nell'ultima parte dell'anno a causa del brusco calo dell'export. Nel Regno Unito il PIL e' cresciuto dell'1,8% su base annua, smentendo le negative previsioni del dopo Brexit. Il Pil dei Paesi dell'Eurozona ha segnato una crescita pari all'1,7%, in graduale consolidamento grazie alla spinta proveniente dalle componenti interne della domanda. In tale contesto la BCE ha annunciato l'estensione anche se per quantitativi inferiori degli stimoli monetari oltre la scadenza fissata in precedenza del marzo 2017. La crescita nell'area rimane comunque differenziata: la Germania cresce in misura pari all'1,8%, la Francia all'1,1%, mentre prosegue la robusta crescita della Spagna che per il secondo anno consecutivo ha segnato un incremento pari al +3,2% rispetto all'anno precedente, grazie ai contributi della domanda interna, degli investimenti nel settore dell'edilizia e delle buone condizioni di concessione del credito a fa[...]prese.


Economic outlook MediaSet.docx - Word

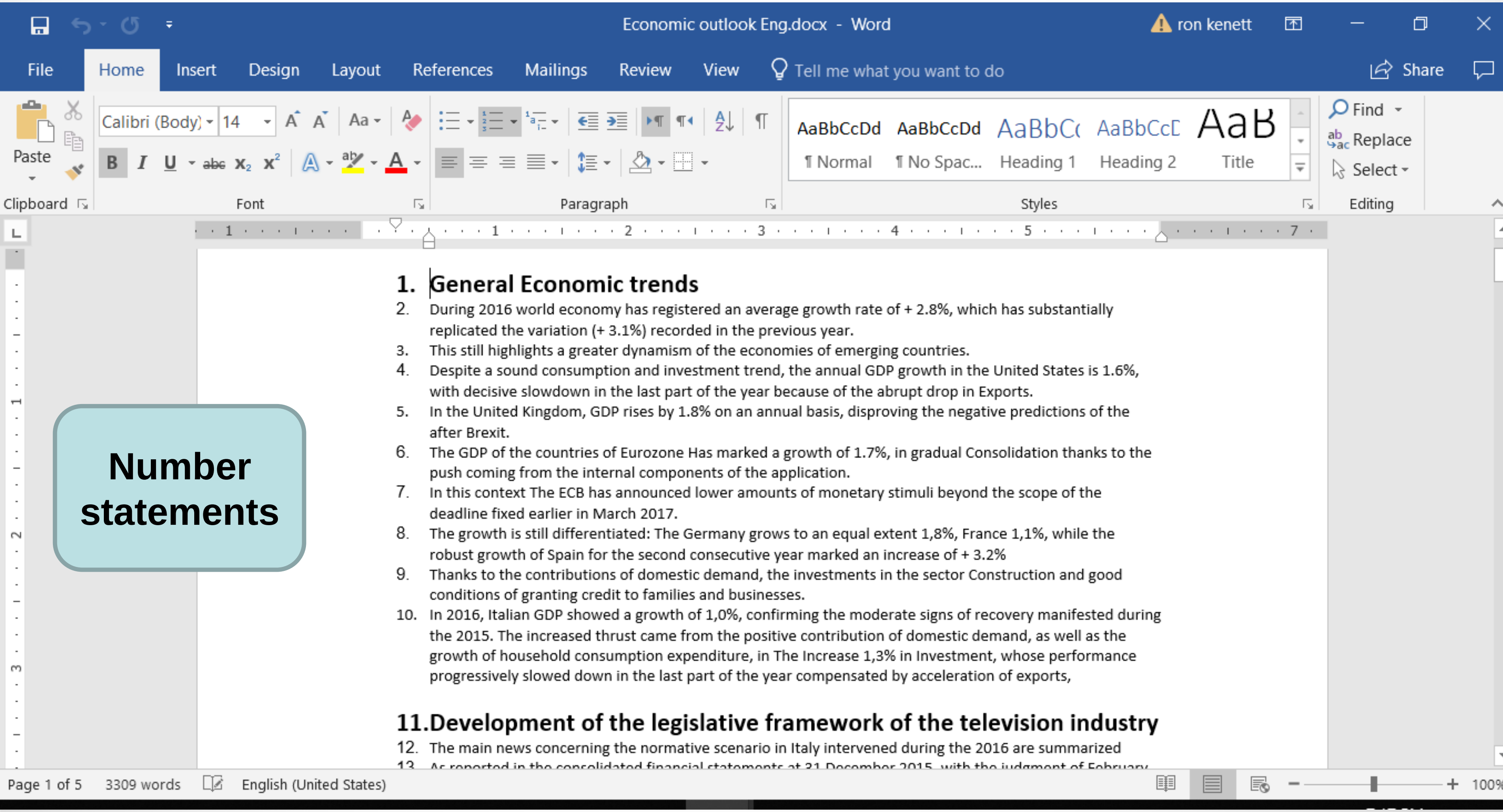
ANDAMENTO GENERALE DELL'ECONOMIA

Nel corso del 2016 l'economia mondiale ha registrato un tasso di crescita media pari al +2,8%, che replica sostanzialmente la variazione (+3,1%) registrata nell'anno precedente, evidenziano ancora un maggiore dinamismo delle economie dei paesi emergenti.
Pur a fronte di un solido andamento di consumi e investimenti, negli Stati Uniti la crescita annua del PIL si e' fermata all'1,6%, con deciso rallentamento nell'ultima parte dell'anno a causa del brusco calo dell'export. Nel Regno Unito il PIL e' cresciuto dell'1,8% su base annua, smentendo le negative previsioni del dopo Brexit. Il Pil dei Paesi dell'Eurozona ha segnato una crescita pari all'1,7%, in graduale consolidamento grazie alla spinta proveniente dalle componenti interne della domanda. In tale contesto la BCE ha annunciato l'estensione anche se per quantitativi inferiori degli stimoli monetari oltre la scadenza fissata in precedenza del marzo 2017. La crescita nell'area rimane comunque differenziata: la Germania cresce in misura pari all'1,8%, la Francia all'1,1%, mentre prosegue la robusta crescita della Spagna che per il secondo anno consecutivo ha segnato un incremento pari al +3,2% rispetto all'anno precedente, grazie ai contributi della domanda interna, degli investimenti nel settore dell'edilizia e delle buone condizioni di concessione del credito a famiglie e imprese.
Nel 2016 il PIL italiano ha registrato una crescita pari all'1,0%, confermando i moderati segnali di ripresa manifestati nel corso del 2015. La maggiore spinta è venuta dal positivo contributo della domanda interna, nonché della crescita della spesa dei consumi delle famiglie, in aumento dell'1,3% e degli investimenti, il cui andamento è però progressivamente rallentato nell'ultima parte dell'anno compensato dall'accellerazione delle esportazioni,

SVILUPPO DEL QUADRO LEGISLATIVO DEL SETTORE TELEVISIVO

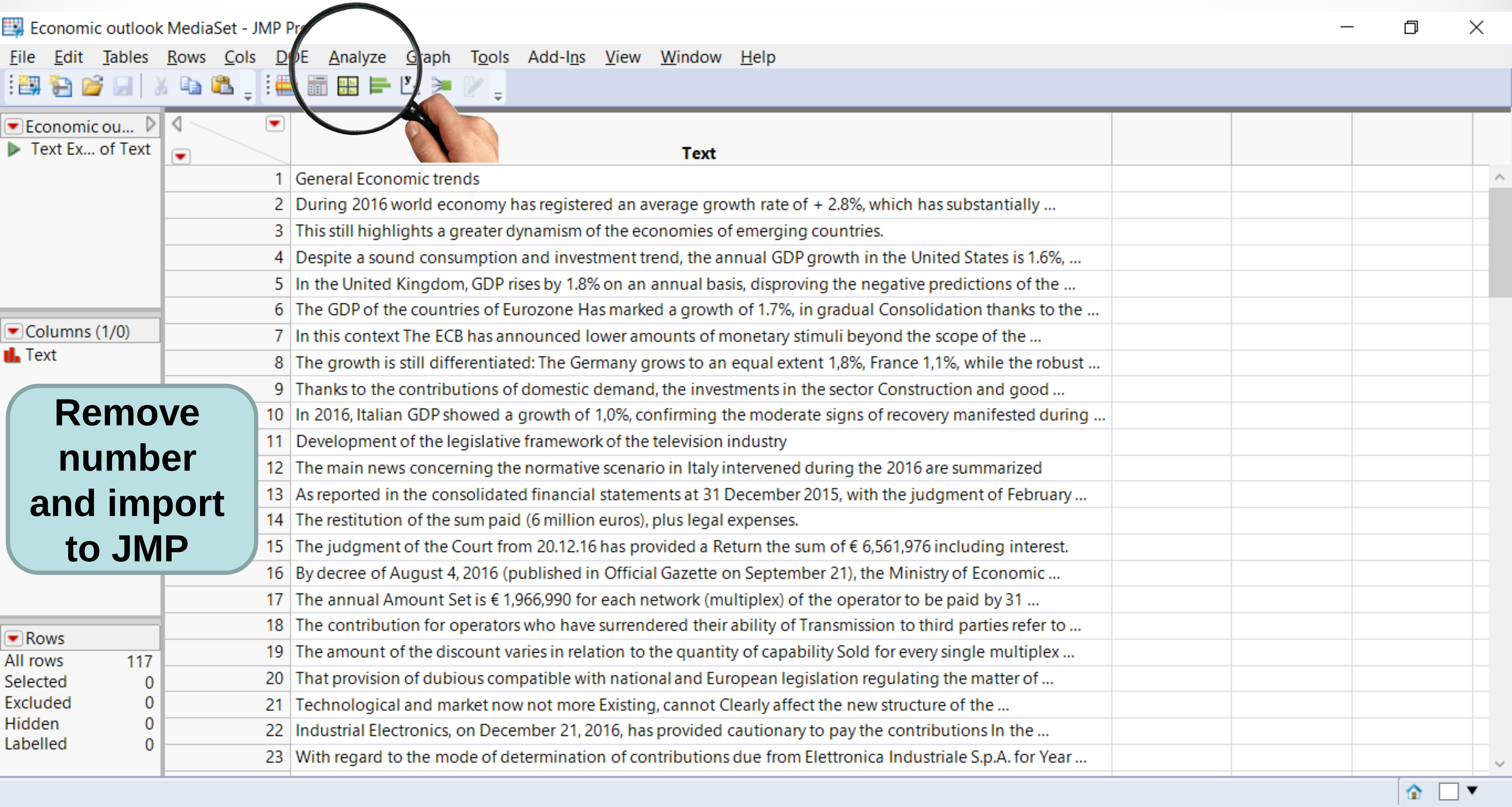
Le principali novità relative allo scenario normativo in Italia intervenute nel corso del 2016 sono così sintetizzabili:

Copy paste to Word



1. General Economic trends
2. During 2016 world economy has registered an average growth rate of + 2.8%, which has substantially replicated the variation (+ 3.1%) recorded in the previous year.
3. This still highlights a greater dynamism of the economies of emerging countries.
4. Despite a sound consumption and investment trend, the annual GDP growth in the United States is 1.6%, with decisive slowdown in the last part of the year because of the abrupt drop in Exports.
5. In the United Kingdom, GDP rises by 1.8% on an annual basis, disproving the negative predictions of the after Brexit.
6. The GDP of the countries of Eurozone Has marked a growth of 1.7%, in gradual Consolidation thanks to the push coming from the internal components of the application.
7. In this context The ECB has announced lower amounts of monetary stimuli beyond the scope of the deadline fixed earlier in March 2017.
8. The growth is still differentiated: The Germany grows to an equal extent 1,8%, France 1,1%, while the robust growth of Spain for the second consecutive year marked an increase of + 3.2%
9. Thanks to the contributions of domestic demand, the investments in the sector Construction and good conditions of granting credit to families and businesses.
10. In 2016, Italian GDP showed a growth of 1,0%, confirming the moderate signs of recovery manifested during the 2015. The increased thrust came from the positive contribution of domestic demand, as well as the growth of household consumption expenditure, in The Increase 1,3% in Investment, whose performance progressively slowed down in the last part of the year compensated by acceleration of exports,
11. Development of the legislative framework of the television industry
12. The main news concerning the normative scenario in Italy intervened during the 2016 are summarized
13. As reported in the consolidated financial statements at 31 December 2015, with the judgment of February

Number statements



Remove number and import to JMP

KPA

Economic outlook MediaSet - Text Explorer of Text - JMP Pro

Text Explorer for Text

Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-empty Cases	Portion Non-empty per Case
928	117	3315	28.3333	117	1.0000

Term and Phrase Lists

Term	Count
mediaset	41
date	29
capital	24
radio	21
shares	18
vivendi	18
million	17
share	16
equal	15
december	13
rti	13
euro	12
rights	11
july	10
year	10
group	9
growth	9
market	9
amount	8
april	8

Phrase	Count	N
share capital	11	2
million euros	6	2
date april	5	2
mediaset premium	5	2
board of directors	4	3
shareholders meeting	4	2
voting rights	4	2
last part of the year	3	5
part of the year	3	4
disposal of capacity	3	3
mediaset espa groupña	3	3
espa groupña	3	2
last part	3	2
mediaset espa	3	2
mediaset españa	3	2
million euro	3	2
ordinary shares	3	2
shares equal	3	2
rti s p a and advertisement 4 adventures slu mediaset ...	2	12
rti s p a and advertisement 4 adventures slu mediaset ...	2	11

Economic outlook MediaSet - Text Explorer of Text - JMP Pro

Text Explorer for Text

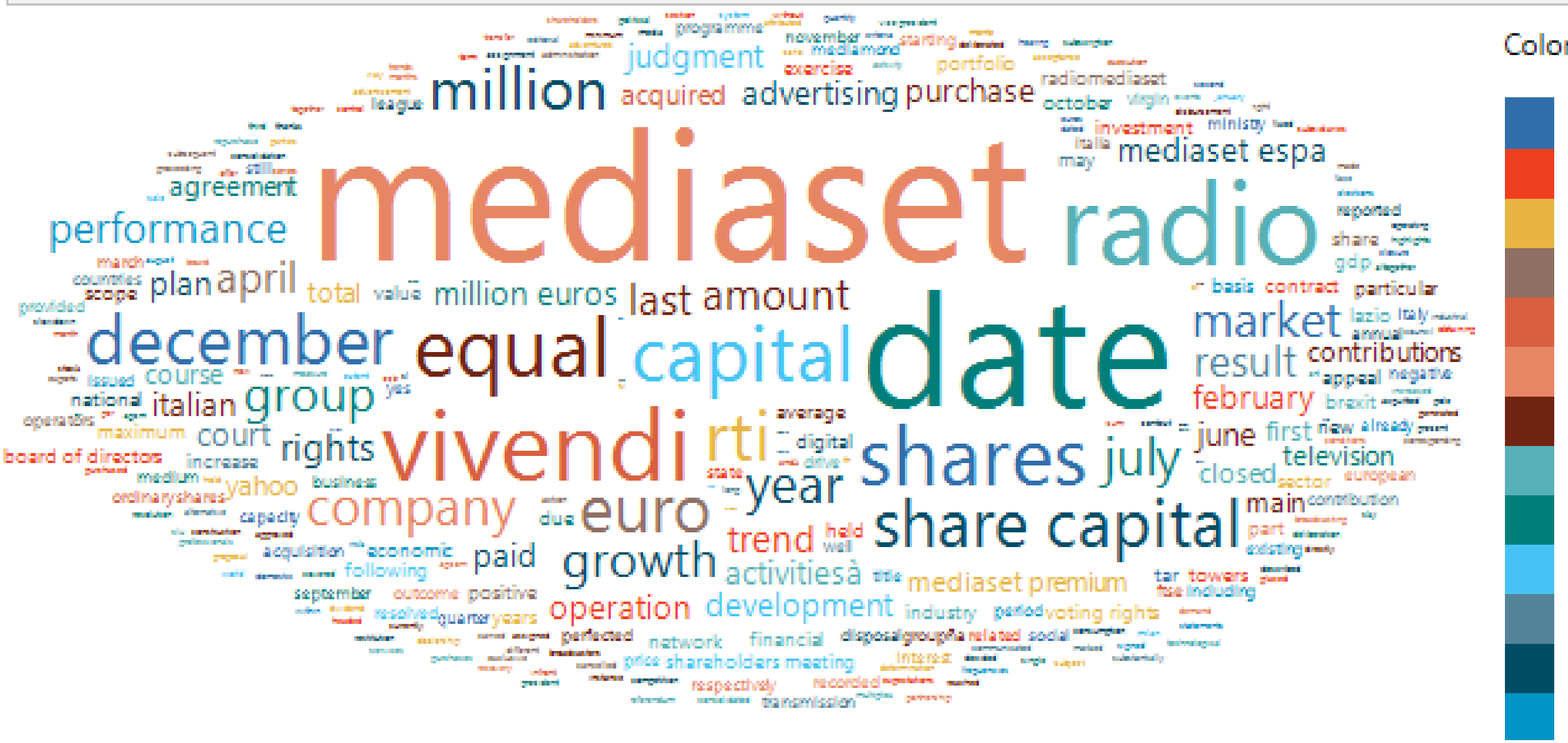
Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-empty Cases	Portion Non-empty per Case
925	117	3315	28.3333	117	1.0000

Term and Phrase Lists

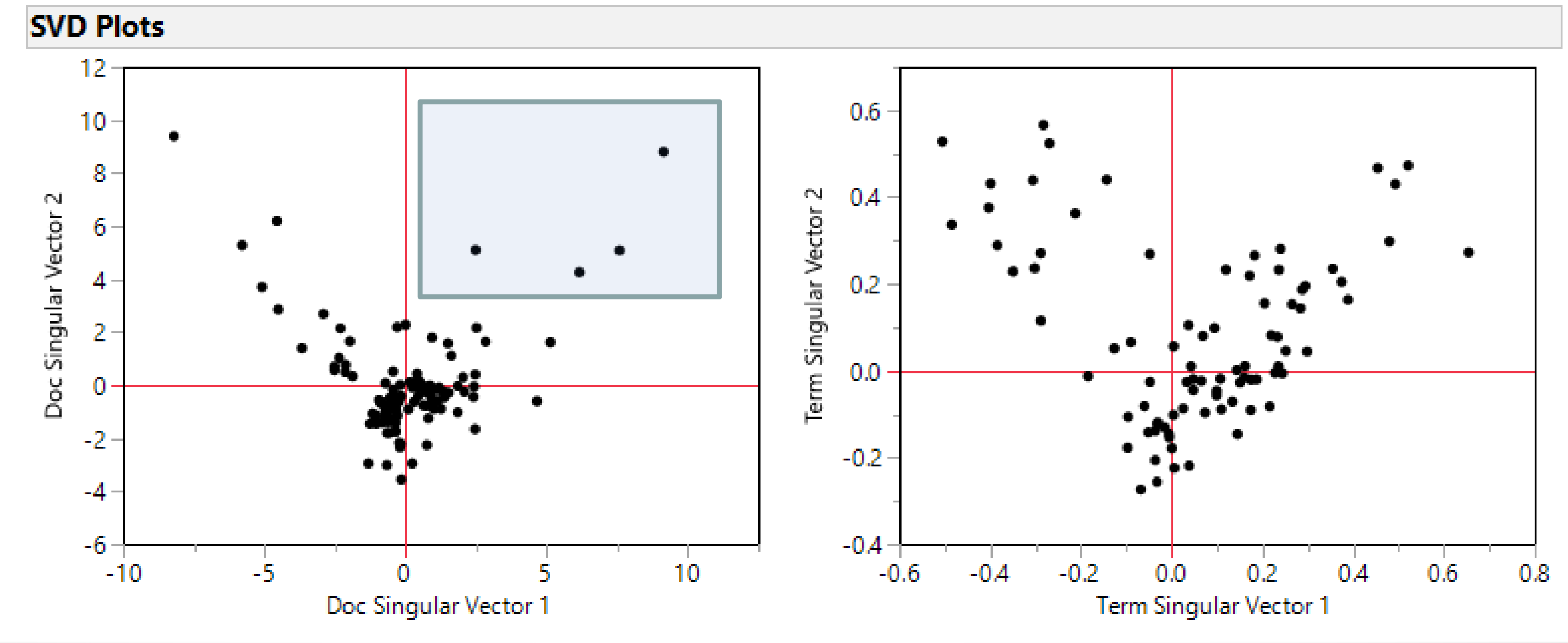
Term	Count
mediaset	33
date	29
radio	21
vivendi	18
equal	15
shares	15
capital	14
december	13
rti	13
euro	12
share capital	12
million	11
company	10
july	10
year	10
group	9
growth	9
market	9
amount	8
april	8

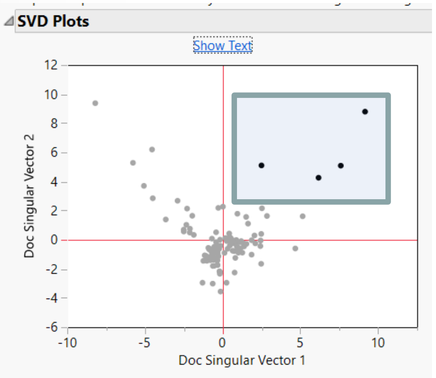
Phrase	Count	N
share capital	12	2
mediaset espa	6	2
million euros	6	2
date april	5	2
mediaset premium	5	2
board of directors	4	3
shareholders meeting	4	2
voting rights	4	2
last part of the year	3	5
part of the year	3	4
disposal of capacity	3	3
mediaset espa groupña	3	3
espa groupña	3	2
last part	3	2
million euro	3	2
ordinary shares	3	2
shares equal	3	2
rti s p a and advertisement 4 adventures slu mediaset ...	2	12
rti s p a and advertisement 4 adventures slu mediaset ...	2	11
rti s p a and advertisement 4 adventures slu mediaset	2	10

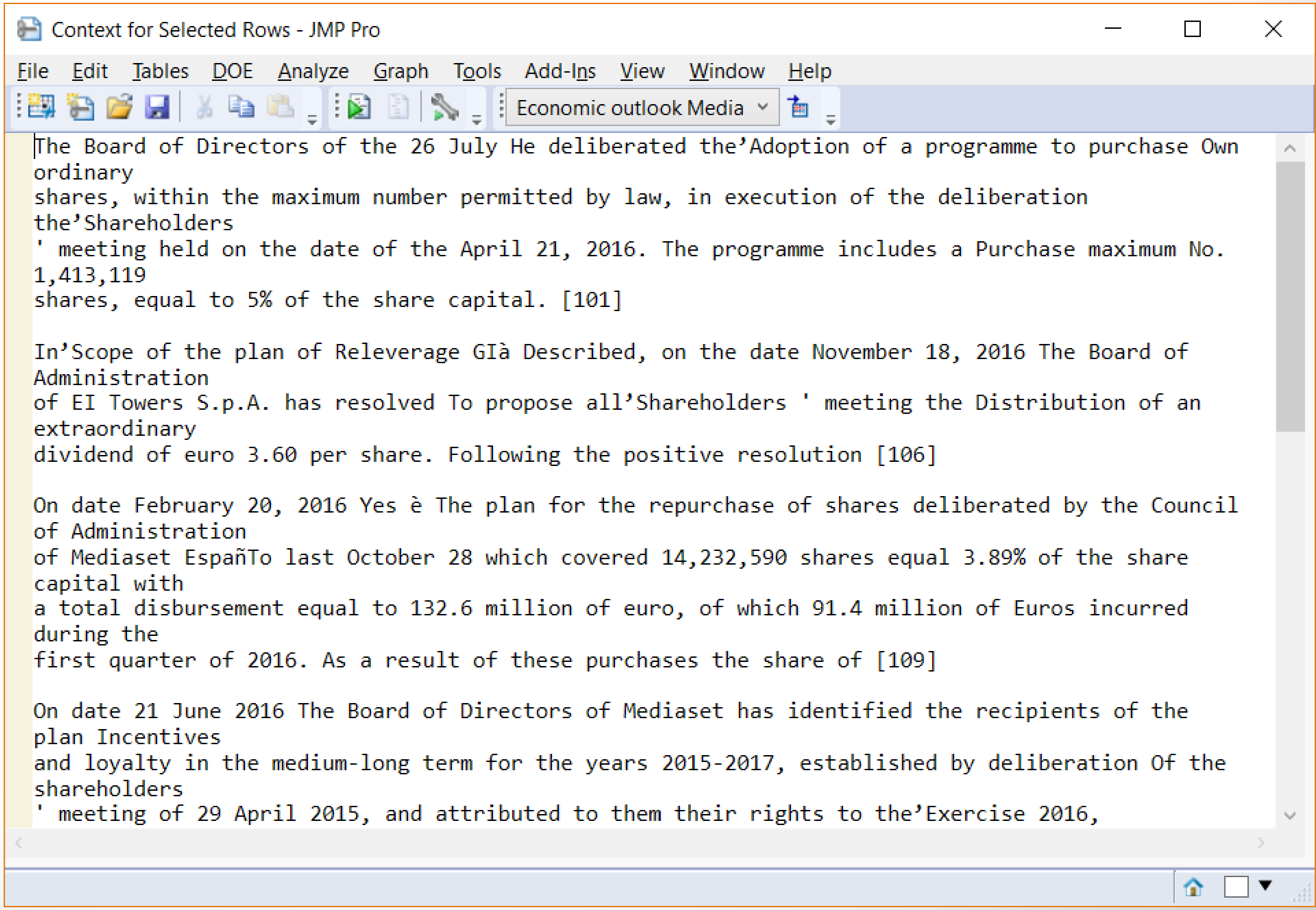
Word Cloud



Color
mediaset
radio
date
vivendi
december
equal
capital
shares
share capital
million
performance
april
company
euro
growth
rti
year
july
market
result
contributions
february
june
first
television
closed
main
group
italian
rights
court
trend
operation
development
activitiesà
mediaset premium
paid
million euros
last amount
judgment
acquired
advertising
purchase
october
mediaset espa
agreement
plan
total
value
course
national
yahoo
increase
board of directors
ordinary shares
capacity
acquisition
economic
following
september
outcome
positive
resolved
quarter
years
perfected
network
financial
shareholders meeting
industry
period
voting rights
disposal
related
social
respectively
recorded
transmission
investment
ministry
virgin
radiomediaset
portfolio
programme
november
mediamond
exercise
league
basis
contract
particular
lazio
italy
annual
appeal
negative
brexit
new
already
european
sector
contribution
part
existing
tax
towers
including
average
digital
state
held
well
due
business
medium
maximum
scope
countries
march
provided
issued
operators
share
gdp
reported
italia
may

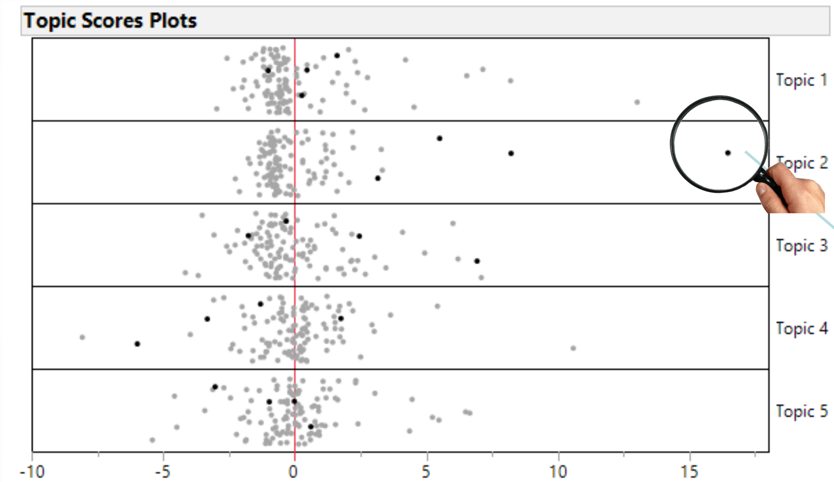

SVD Plots
Doc Singular Vector 2
Doc Singular Vector 1
12
10
8
6
4
2
0
-2
-4
-6
-10
-5
0
5
10
Term Singular Vector 2
Term Singular Vector 1
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8

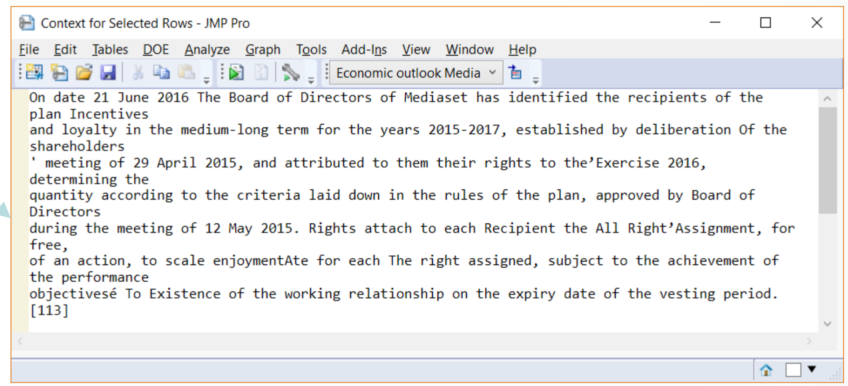

SVD Plots
Show Text
Doc Singular Vector 2
Doc Singular Vector 1
12
10
8
6
4
2
0
-2
-4
-6
-10
-5
0
5
10


Context for Selected Rows - JMP Pro
File Edit Tables DOE Analyze Graph Tools Add-Ins View Window Help
Economic outlook Media
The Board of Directors of the 26 July He deliberated the'Adoption of a programme to purchase Own ordinary
shares, within the maximum number permitted by law, in execution of the deliberation the'Shareholders
' meeting held on the date of the April 21, 2016. The programme includes a Purchase maximum No. 1,413,119
shares, equal to 5% of the share capital. [101]
In'Scope of the plan of Releverage GIà Described, on the date November 18, 2016 The Board of Administration
of EI Towers S.p.A. has resolved To propose all'Shareholders ' meeting the Distribution of an extraordinary
dividend of euro 3.60 per share. Following the positive resolution [106]
On date February 20, 2016 Yes è The plan for the repurchase of shares deliberated by the Council of Administration
of Mediaset EspañTo last October 28 which covered 14,232,590 shares equal 3.89% of the share capital with
a total disbursement equal to 132.6 million of euro, of which 91.4 million of Euros incurred during the
first quarter of 2016. As a result of these purchases the share of [109]
On date 21 June 2016 The Board of Directors of Mediaset has identified the recipients of the plan Incentives
and loyalty in the medium-long term for the years 2015-2017, established by deliberation Of the shareholders
' meeting of 29 April 2015, and attributed to them their rights to the'Exercise 2016,

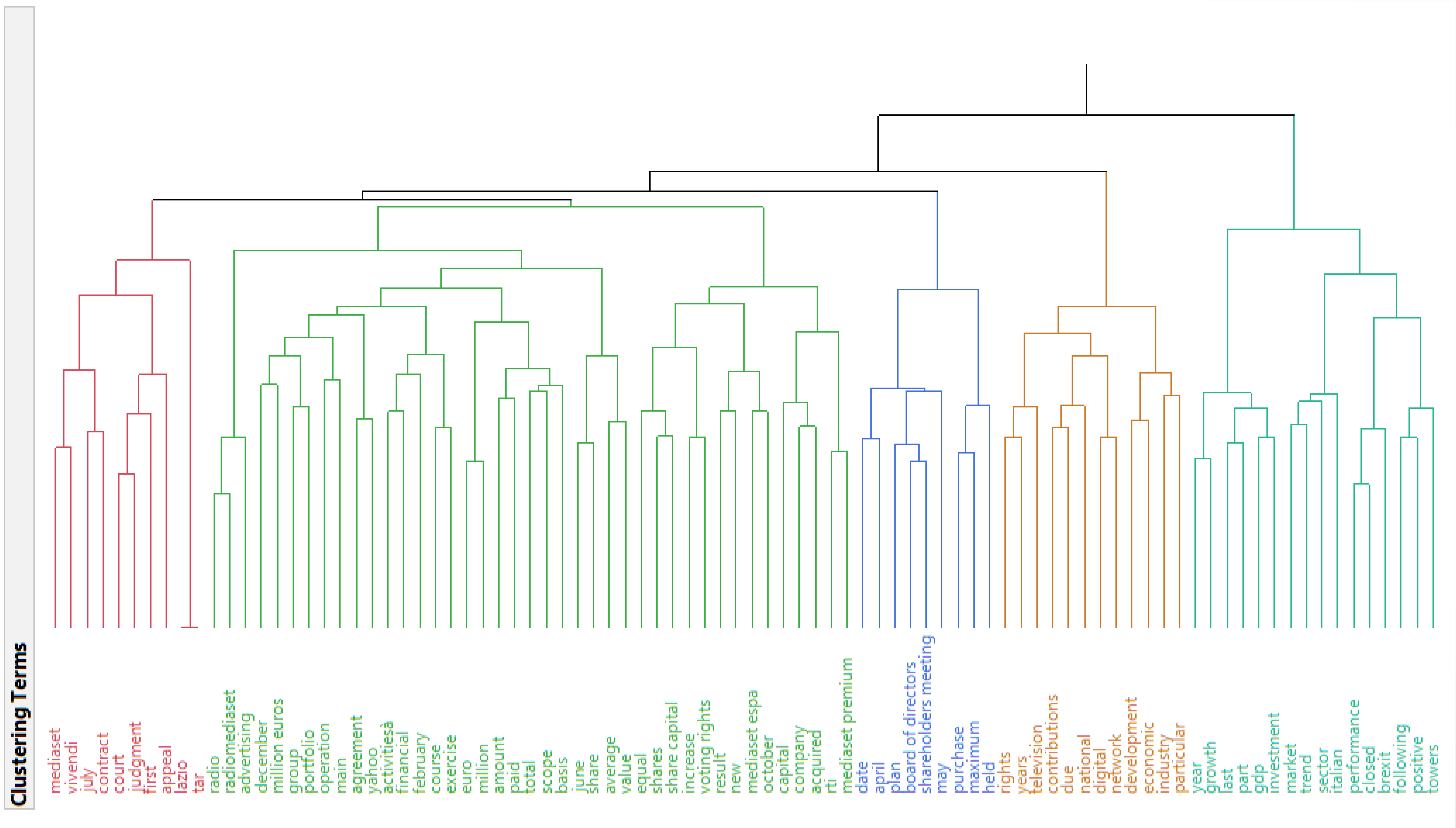
Top Loadings by Topic

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
Term	Loading	Term	Loading	Term	Loading	Term	Loading	Term	Loading
growth	0.76226	board of directors	0.79553	share capital	0.5893	digital	0.5248	judgment	0.5074
part	0.65251	shareholders meeting	0.75188	shares	0.5856	network	0.4772	lazio	0.5046
year	0.61201	plan	0.69908	rti	0.4906	first	-0.4756	tar	0.5046
performance	0.57413	date	0.60199	capital	0.4755	television	0.4634	july	0.4275
gdp	0.57119	rights	0.58254	mediaset espa	0.4621	equal	-0.4359	court	0.4201
positive	0.55028	april	0.57188	october	0.4376	years	0.4279	radio	-0.4143
investment	0.53719	years	0.52488	acquired	0.4083	july	-0.4134	group	-0.4007
last	0.50660	exercise	0.46498	voting rights	0.3639	court	-0.3928	activitiesà	-0.3649
increase	0.46461	may	0.45551	result	0.3551	contract	-0.3923	radiomediaset	-0.3598
trend	0.44960	june	0.31596	company	0.3508	vivendi	-0.3601	contributions	0.3495
italian	0.40199	maximum	0.30784	equal	0.3506	contributions	0.3541	due	0.3461
closed	0.35532			increase	0.3217	mediaset	-0.3524	advertising	-0.3333
brexit	0.34598			brexit	-0.2911	development	0.3460	contract	0.2990
						due	0.3386		
						national	0.3258		
						economic	0.3210		





On date 21 June 2016 The Board of Directors of Mediaset has identified the recipients of the plan Incentives
and loyalty in the medium-long term for the years 2015-2017, established by deliberation Of the shareholders
' meeting of 29 April 2015, and attributed to them their rights to the’Exercise 2016, determining the
quantity according to the criteria laid down in the rules of the plan, approved by Board of Directors
during the meeting of 12 May 2015. Rights attach to each Recipient the All Right’Assignment, for free,
of an action, to scale enjoymentAte for each The right assigned, subject to the achievement of the performance
objectivesé To Existence of the working relationship on the expiry date of the vesting period. [113]


Clustering Terms
mediaset
vivendi
july
contract
court
judgment
first
appeal
lazio
tar
radio
radiomediaset
advertising
december
million euros
group
portfolio
operation
main
agreement
yahoo
activitiesà
financial
february
course
exercise
euro
million
amount
paid
total
scope
basis
june
share
average
value
equal
shares
share capital
increase
voting rights
result
new
mediaset espa
october
capital
company
acquired
rti
mediaset premium
date
april
plan
board of directors
shareholders meeting
may
purchase
maximum
held
rights
years
television
contributions
due
national
digital
network
development
economic
industry
particular
year
growth
last
part
gdp
investment
market
trend
sector
italian
performance
closed
brexit
following
positive
towers

Text analytics task

- Again, in your task1 group
- Choose a text
- Use JMP text explorer to analyze it
- Prepare a brief report

Task 2 to get a pass/fail grade

	text
1	GREAT NEWS! #MAGA #KAG https://t.co/GXDE2IlGGu
2	THANK YOU! #MAGA #KAG https://t.co/igO1r1cTHS
3	https://t.co/BKo27n6tmz
4	“Congressman Van Drew (D-NJ) SLAMS Democrats for ‘fracturing the Nation’ with Impeachment prob...
5	Just finished a very good & cordial meeting at the White House with Jay Powell of the Federal R...
6	Just finished a very good & cordial meeting at the White House with Jay Powell of the Federal R...
7	....that I testify about the phony Impeachment Witch Hunt. She also said I could do it in writing. Even t...
8	Our Crazy, Do Nothing (where’s USMCA, infrastructure, lower drug pricing & much more?) Spea...
9	Never has the Republican Party been so united as it is now. 95% A.R. This is a great fraud being playe...
10	https://t.co/Mqj5tXaDAz
11	“All they do is bring up witnesses who didn’t witness anything.” @KatrinaPierson @SteveHiltonx Not...
12	“The Impeachment started before he even became President.” @greggutfeld @FoxNews
13	https://t.co/1Rg66Tn4uP
14	https://t.co/D66PEkuX6d
15	Where is the Fake Whistleblower?
16	https://t.co/ru2n7i2gzu
17	Republicans & others must remember, the Ukrainian President and Foreign Minister both said th...
18	The Crazed, Do Nothing Democrats are turning Impeachment into a routine partisan weapon. That is ...
19	Tell Jennifer Williams, whoever that is, to read BOTH transcripts of the presidential calls, & see th...
20	https://t.co/I3lO117SVh
21	Paul Krugman of @nytimes has been wrong about me from the very beginning. Anyone who has foll...
22	Schiff is a Corrupt Politician! https://t.co/DDBqlfIFLV
23	.@SteveScalise blew the nasty & obnoxious Chris Wallace (will never be his father, Mike!) away o...
24	.@SteveScalese blew the nasty & obnoxious Chris Wallace (will never be his father, Mike!) away ...
25	Thanks Eric! https://t.co/6Ai7bqto3P

DataTrumpTweets.docx

Argomento 1

Termine	Caricamento in corso
edwards	0,74424
louisiana	0,72874
bel	0,72672
eddierispone	0,63926
insurance	0,52384
governor	0,51783
runoff	0,47008
saturday	0,45531
taxes	0,41461
john	0,39798
amendment	0,39558
vote	0,38273
republican	0,37721
nd	0,37287

Argomento 2

Termine	Caricamento in corso
border	0,63691
southern	0,49385
drugs	0,44683
wall	0,41790
security	0,41658
trafficking	0,34897
loopholes	0,34027
human	0,33188
immigration	0,31876
laws	0,31558
fix	0,30936
democrats	0,29876
mexico	0,28726
crime	0,28340

Argomento 3

Termine	Caricamento in corso
schiff	0,56572
adam	0,48988
ukrainian	0,41966
whistleblower	0,39766
transcript	0,39038
call	0,39031
fraudulently	0,37387
ukraine	0,36185
read	0,35290
shifty	0,35149
phone call	0,33359
conversation	0,32117
perfect	0,27216
reading	0,26482

Argomento 4

Termine	Caricamento in corso
tariffs	0,58368
china	0,55925
dollars	0,53933
goods	0,52066
products	0,50491
billion	0,50052
25	0,47360
product	0,36854
buy	0,33346
farmers	0,32091
agricultural	0,30934
remaining	0,29829
deal	0,28796

Argomento 5

Termine	Caricamento in corso
hillary	0,58160
crooked	0,53172
clinton	0,43139
lover	0,41061
dnc	0,40199
fbi	0,39744
mueller	0,38504
mccabe	0,37384
lisa	0,37358
collusion	0,35268
page	0,34078
comey	0,32179
deleted	0,31924
steele	0,30419

Argomento 6

Termine	Caricamento in corso
reserve	0,59949
rates	0,59932
inflation	0,54620
fed	0,49679
interest	0,49279
federal	0,47777
tightening	0,47135
quantitative	0,44393
powell	0,41615
jay	0,39674
dollar	0,33229
raised	0,32911
low	0,32521

Argomento 7

Termine	Caricamento in corso
fake	0,58219
news	0,55081
media	0,50002
corrupt	0,28261
post	0,26236
story	0,25991
cnn	0,25308
sources	0,25029
times	0,23751
lamestream	0,21789
reporting	0,21251
totally	0,21223
even	0,21188

Argomento 8

Termine	Caricamento in corso
vets	0,61021
endorsement	0,57610
amendment	0,46005
military	0,44581
loves	0,44549
complete	0,41221
bishop	0,35811
mattbevin	0,35210
strong	0,35181
kentucky	0,33150
dan	0,32711
total	0,32581
vote	0,31553
crime	0,31467
second	0,30954

Argomento 9

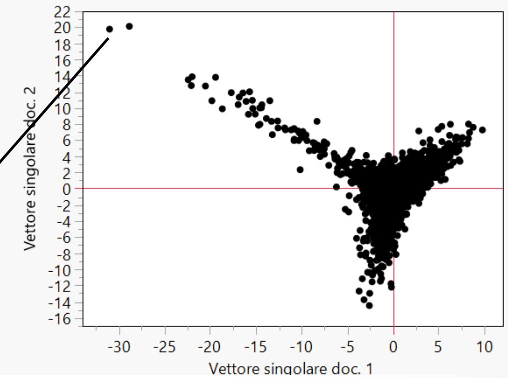
Termine	Caricamento in corso
kim	0,85027
jong	0,83155
un	0,83155
north korea	0,71666
chairman	0,46030
summit	0,41484
meeting	0,33783
vietnam	0,32744
nuclear	0,31247
forward	0,28862

Argomento 10

Termine	Caricamento in corso
turkey	0,57618
kurds	0,55861
wars	0,51127
isis	0,50448
endless	0,47681
syria	0,43427
fighters	0,39697
caliphate	0,34032
captured	0,33291
secured	0,32840
soldiers	0,32651

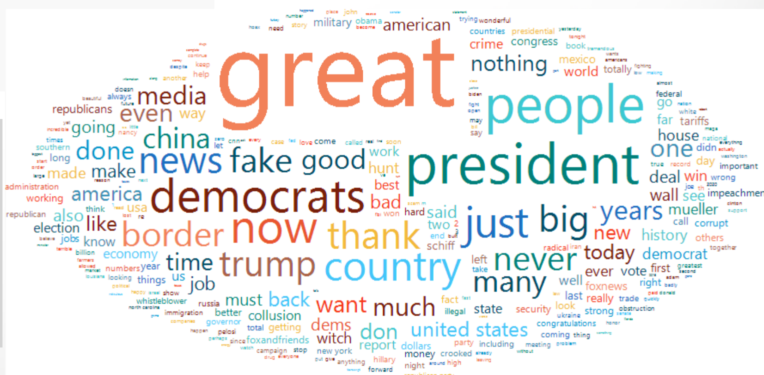
Diagrammi SVD

Mostra testo



Louisiana, get out and Vote Early for @EddieRispone as your next Governor. Lower Taxes and car insurance.
Will protect your 2nd Amendment. John Bel Edwards is always fighting our MAGA Agenda. Wants to raise your
taxes and car insurance to the sky. Vote for Republican Eddie R! [236]

Big Rally in Louisiana on Friday night. Must force a runoff with a Liberal Democrat Governor, John Bel Edwards, who has let your Taxes and Car Insurance get too high, and will never protect your 2nd Amendment.
Vote for one of our two great Republicans on Saturday, force a runoff! [719]



Task 3 to get a pass/fail grade

1. Develop a model to predict NPL
2. Explain what you did
3. Explain what you learned

Case: German Credit

The German Credit data set (available at ftp.ics.uci.edu/pub/machine-learning-databases/statlog/) contains observations on 30 variables for 1000 past applicants for credit. Each applicant was rated as "good credit" (700 cases) or "bad credit" (300 cases).

New applicants for credit can also be evaluated on these 30 "predictor" variables. We want to develop a credit scoring rule that can be used to determine if a new applicant is a good credit risk or a bad credit risk, based on values for one or more of the predictor variables. All the variables are explained in Table 1.1. (Note: The original data set had a number of categorical variables, some of which have been transformed into a series of binary variables so that they can be appropriately handled by XLMiner. Several ordered categorical variables have been left as is; to be treated by XLMiner as numerical. The data has been organized in the spreadsheet German CreditI.xls)

Var. #	Variable Name	Description	Variable Type	Code Description
1.	OBS#	Observation No.	Categorical	Sequence Number in data set
2.	CHK_ACCT	Checking account status	Categorical	0 : < 0 DM 1: 0 <= ...< 200 DM 2 : => 200 DM 3: no checking account
3.	DURATION	Duration of credit in months	Numerical	
4.	HISTORY	Credit history	Categorical	0: no credits taken

GermanCredit Data.jmp

KPA

Task 1: Information quality assessment of one case study

Task 2: Trump tweets text analysis

Task 3: German credit data analysis

ron@kpa-group.com

1/3/2020

Thank you for your attention