Data preprocessing

Violaine Antoine

ISIMA / LIMOS

January, 2020

Outline

Data types

Complex data

- image
- data stream : video, audio, text
- graph

Simple data

- qualitative data
  - ordinal (level of appreciation : good, fair, poor)
  - nominal or categorical (brand, colors)
- quantitative data
  - discrete (number population)
  - continuous (temperature)

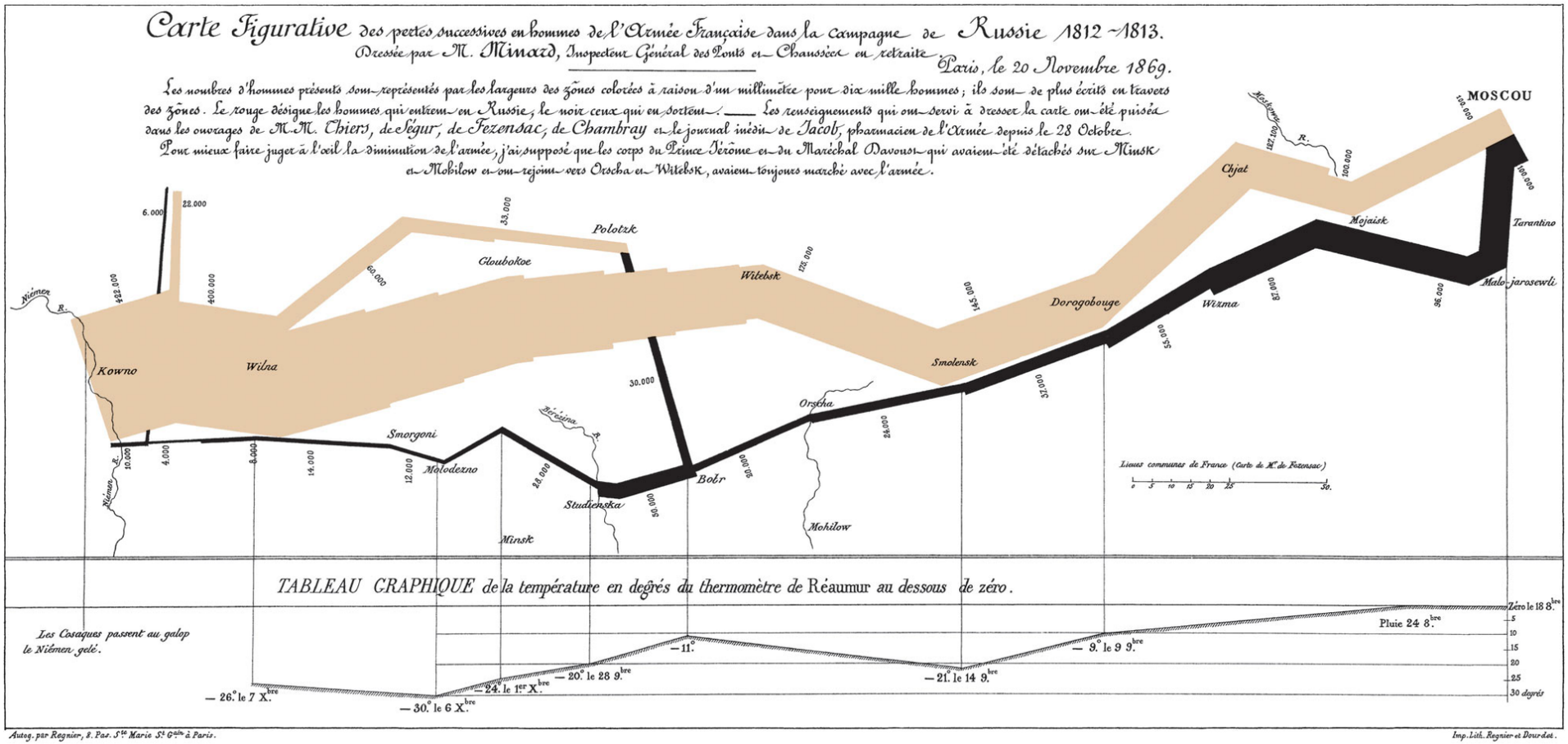
Outline

Data visualization

Data visualization enables to perceive information in various way to :

- make assumptions about the data
- understand large data
- detect of errors and outliers
- identify patterns
- improve decision-making

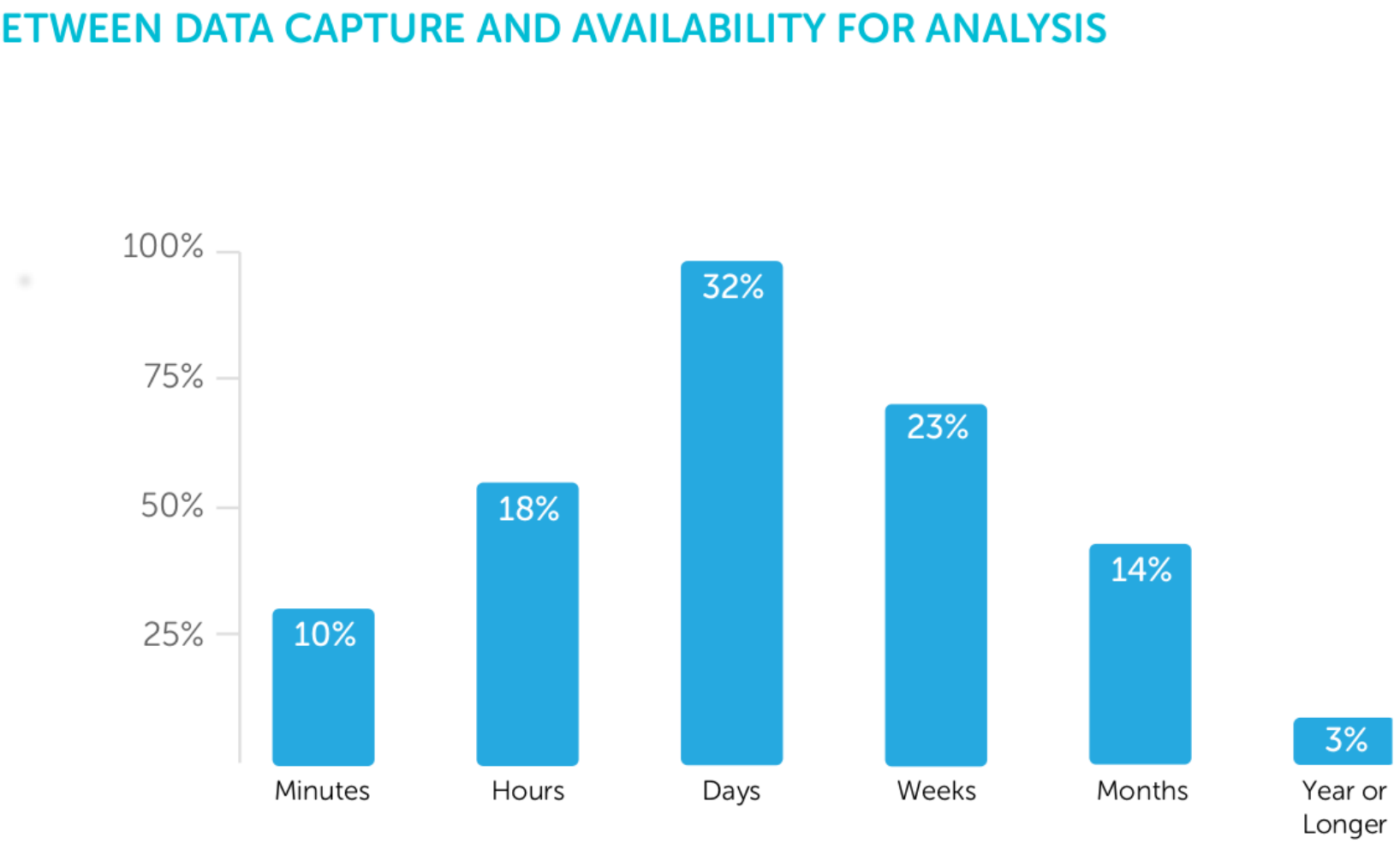
Data visualization



French engineer and civil servant Charles Joseph Minard in 1861

Simple data visualization : qualitative data

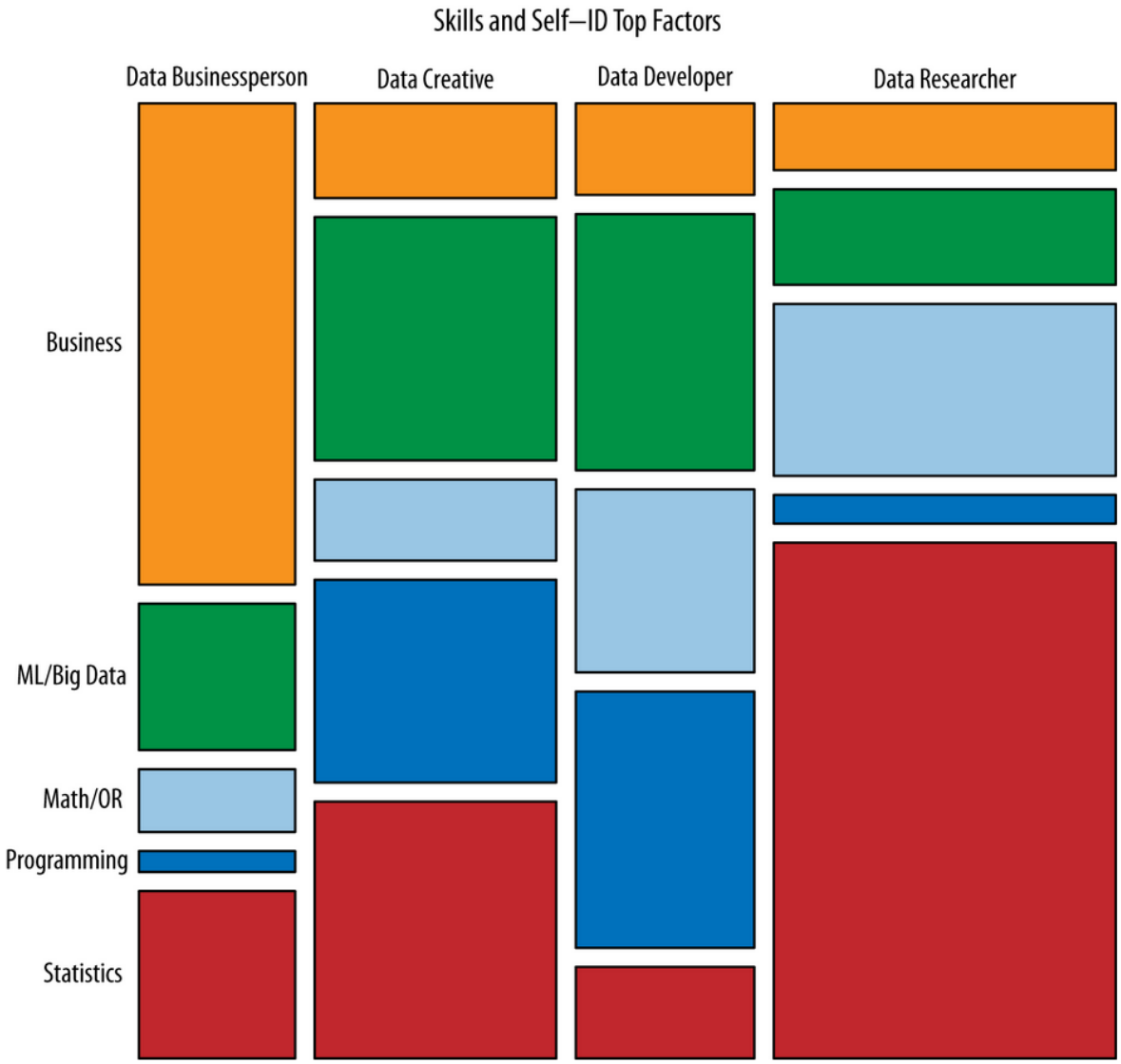
Histogram



Data Science Survey, 2015, Rexer Analytics

Simple data visualization : qualitative data

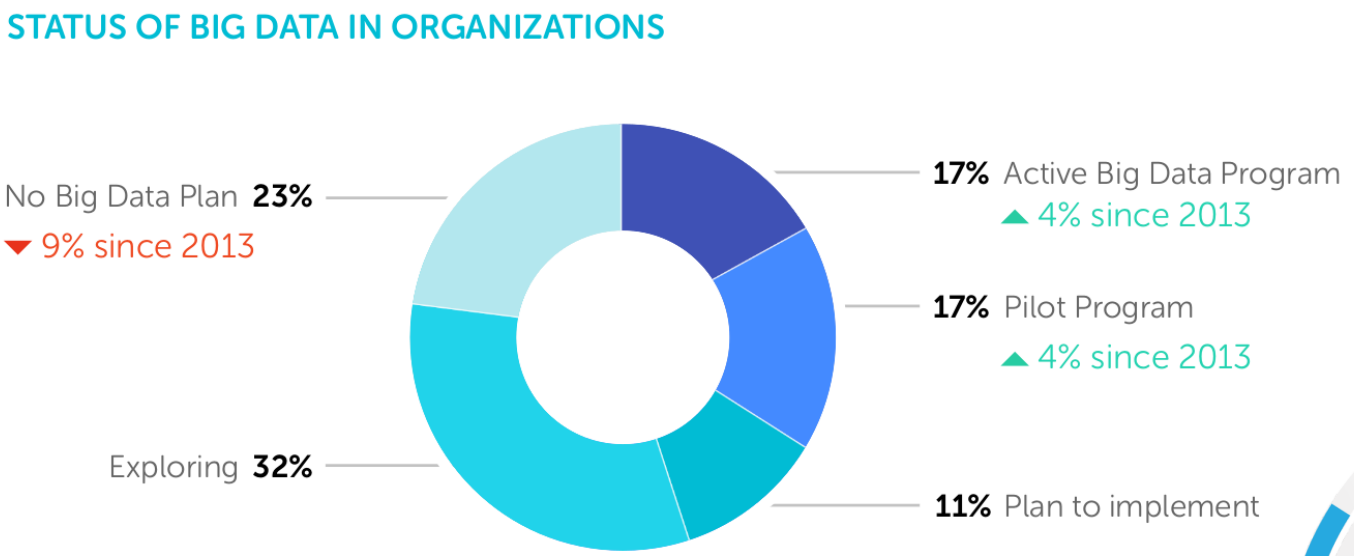
Mosaic plot

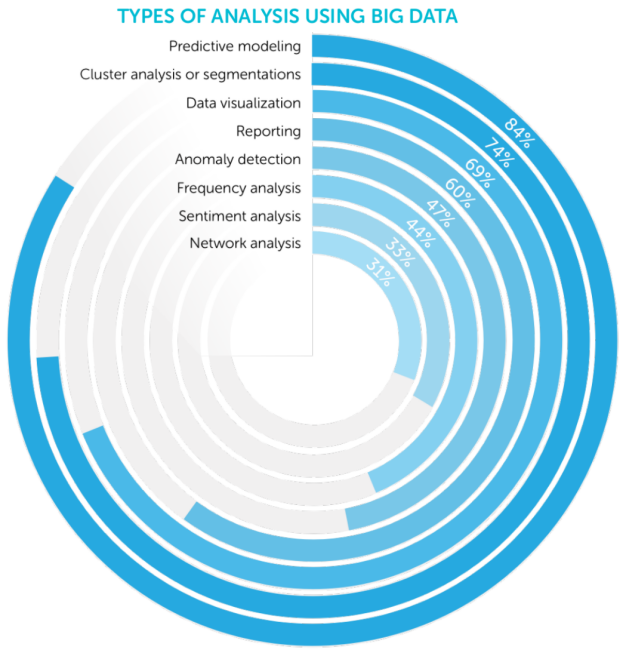


https ://jeremiahstanghini.com/2017/07/30/what-is-data-science/

Simple data visualization : qualitative data

Pie chart and multilevel pie chart

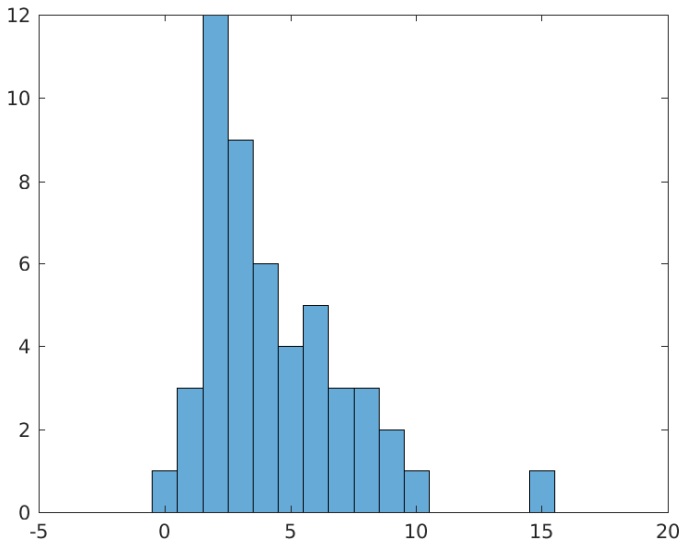




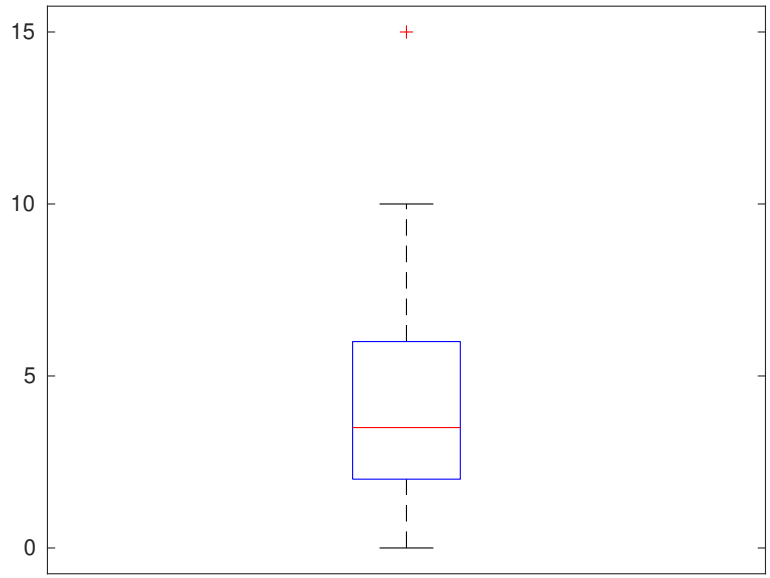
Data Science Survey, 2015, Rexer Analytics

Simple data visualization : quantitative data

Histogram and boxplot



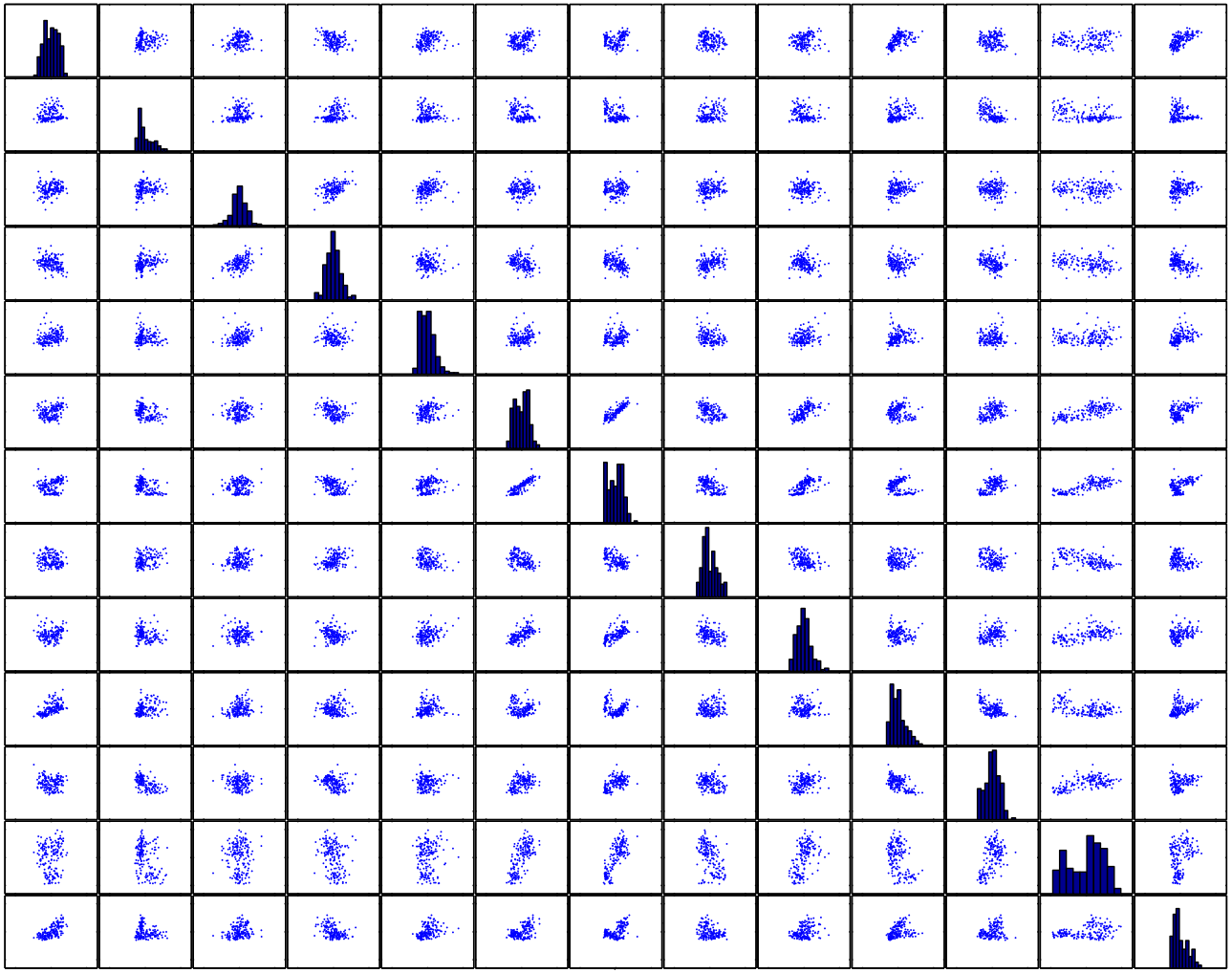
Distribution, range, max and min values



Median, quartiles, interquartile range (IQR), lowest and highest data still within 1.5*IQR, outliers

Simple data visualization : quantitative data

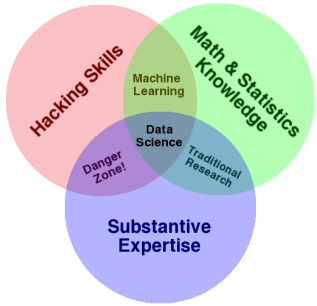
Scatterplot



Wine dataset

Simple data visualization

- bubble chart, ven diagram,
- table,
- line chart,
- time line...

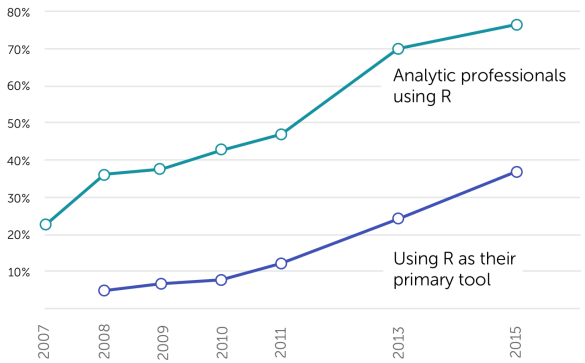


http ://drewconway.com

CONCURRENT TOOL USE

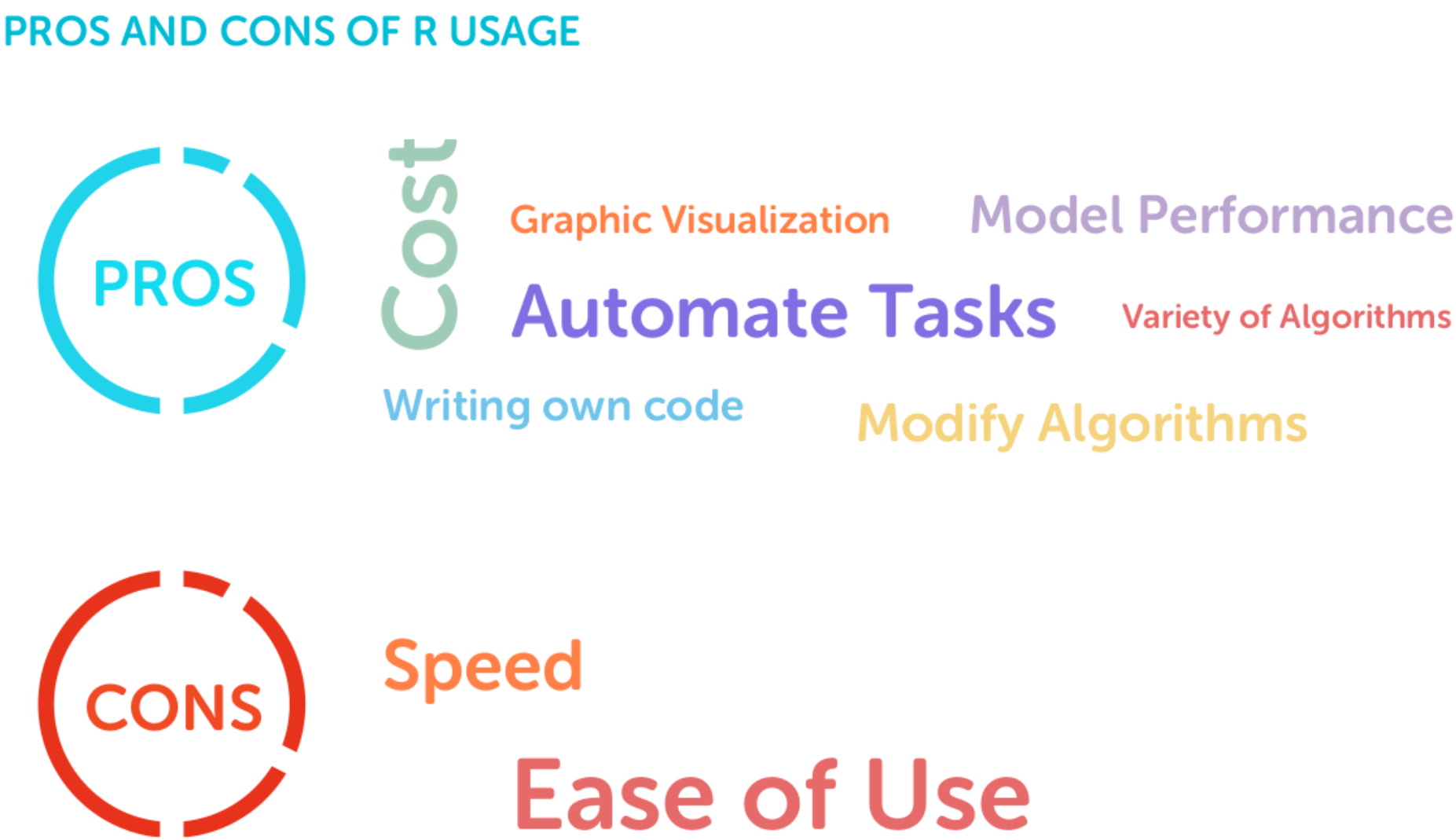
Antecedent \ Consequent	R	IBM SPSS Statistics	SAS	MS Excel Data Mining	Tableau	IBM SPSS Modeler	Weka	Matlab	KNIME (free)	Rapid Miner (free)	MS SQL Server Data Mining	SAS Enterprise Miner	STATISTICA
R		32%	31%	30%	28%	23%	22%	22%	22%	19%	20%	18%	11%
IBM SPSS Statistics	78%		35%	40%	28%	56%	20%	24%	19%	23%	25%	22%	15%
SAS	79%	37%		31%	35%	30%	15%	23%	18%	20%	24%	55%	14%
MS Excel Data Mining	80%	44%	32%		35%	29%	22%	33%	24%	24%	46%	21%	18%
Tableau	86%	36%	42%	41%		27%	23%	27%	22%	25%	29%	26%	11%
IBM SPSS Modeler	75%	75%	38%	36%	28%		18%	24%	24%	25%	27%	28%	14%
Weka	88%	33%	23%	33%	30%	22%		36%	41%	41%	21%	18%	17%
Matlab	85%	37%	34%	46%	33%	28%	34%		29%	29%	29%	19%	22%
KNIME (free)	88%	30%	27%	35%	28%	29%	41%	30%		40%	26%	19%	14%
Rapid Miner (free)	86%	42%	35%	41%	35%	34%	47%	35%	46%		33%	23%	19%
MS SQL Server Data Mining	84%	44%	39%	73%	40%	35%	23%	33%	28%	31%		28%	14%
SAS Enterprise Miner	79%	39%	92%	33%	36%	37%	20%	21%	20%	22%	28%		14%
STATISTICA	67%	37%	33%	41%	21%	25%	26%	34%	21%	25%	19%	19%	

RISE OF R USAGE



Data Science Survey, 2015, Rexer Analytics

Big data visualization : words cloud



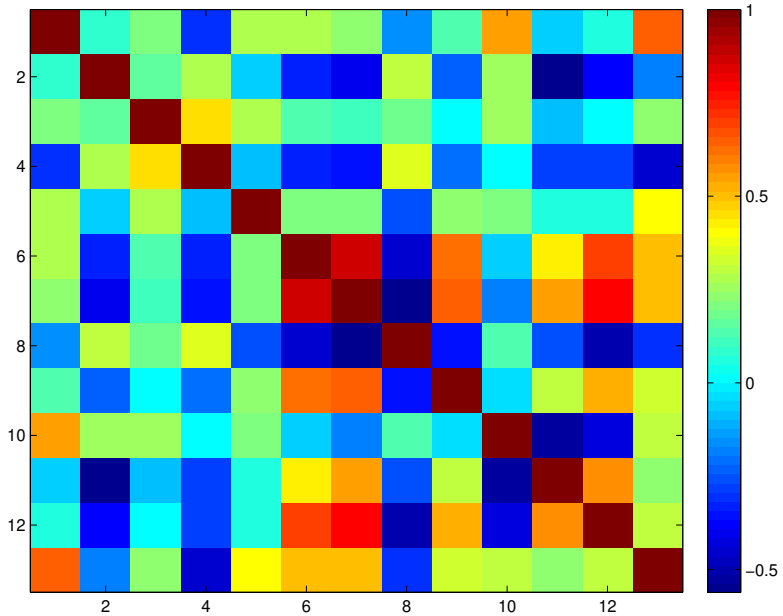
Data Science Survey, 2015, Rexer Analytics

Big data visualization : correlations

Let $\mathbf{X} = (x_1, \ldots x_n)$, $\mathbf{Y} = (y_1, \ldots, y_n)$ be 2 variables

- $\bar{x}$, $\bar{y}$ are the means of $\mathbf{X}$, $\mathbf{Y}$,
- σ_x, σ_y the standard deviation of $\mathbf{X}$, $\mathbf{Y}$.

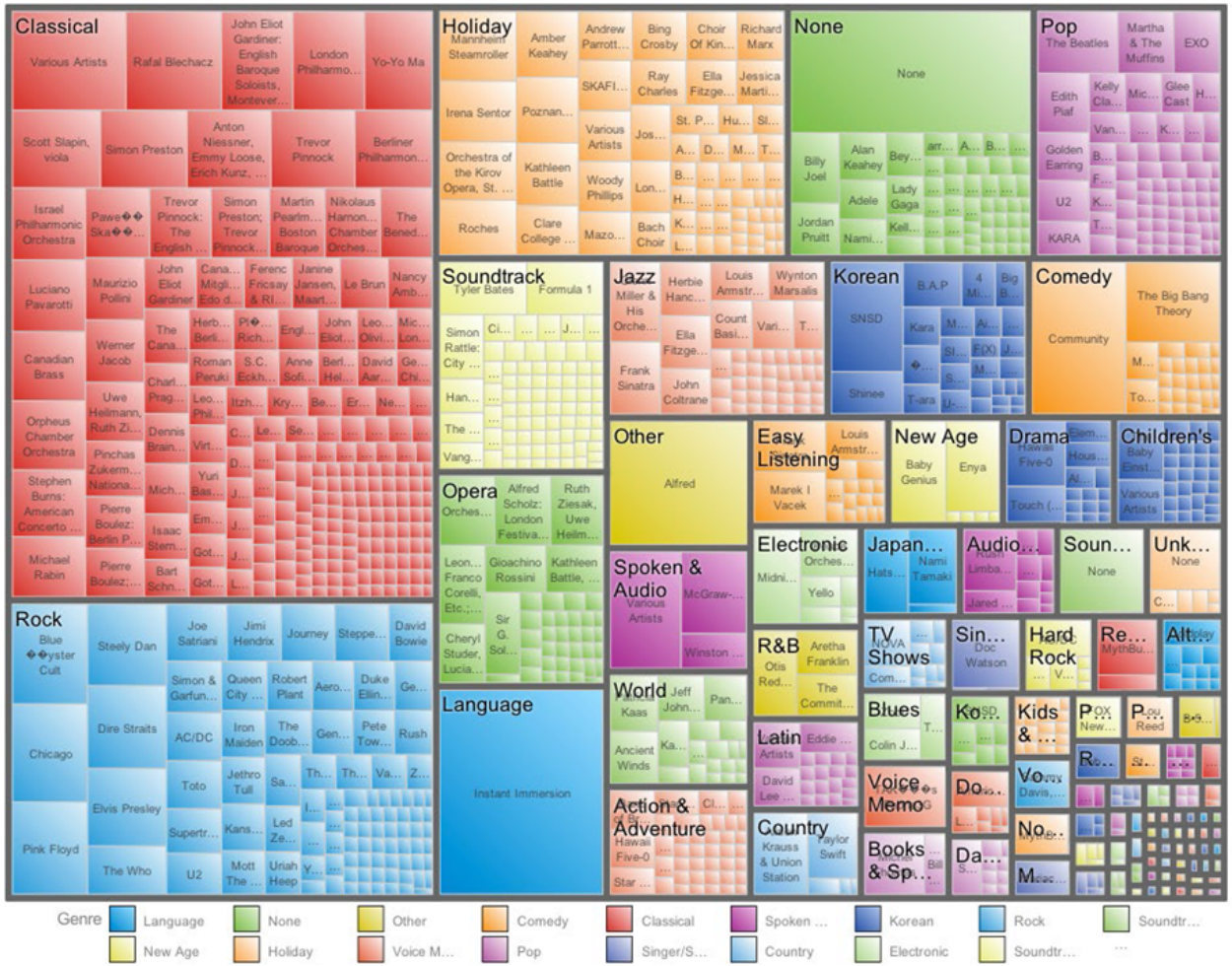
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$$



Wine dataset

Big data visualization : tree map

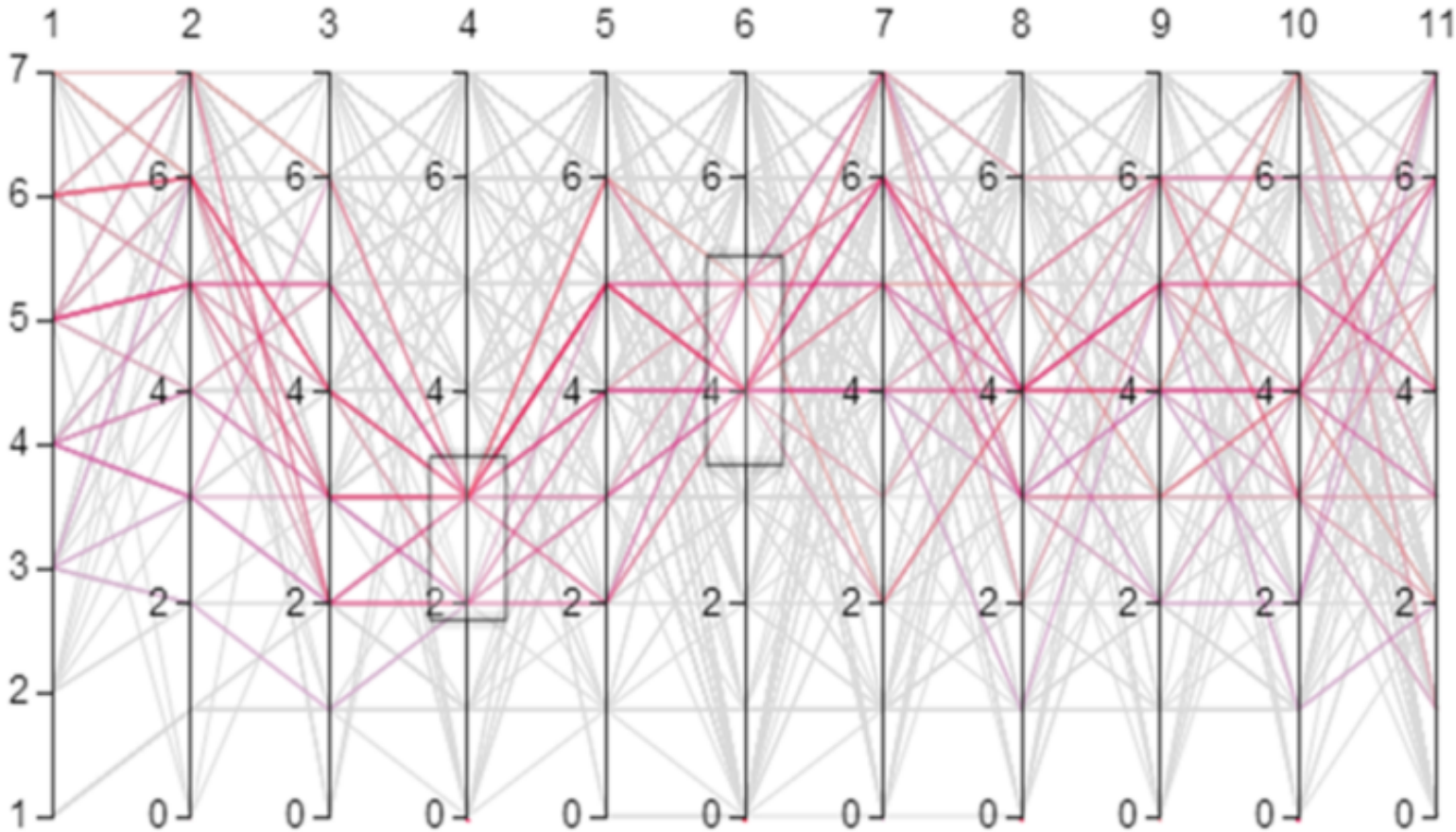
Social network's track selections from a streaming media service



T. Keahey, **Using visualization to understand big data**, Technical Report, IBM Corporation, 2013.

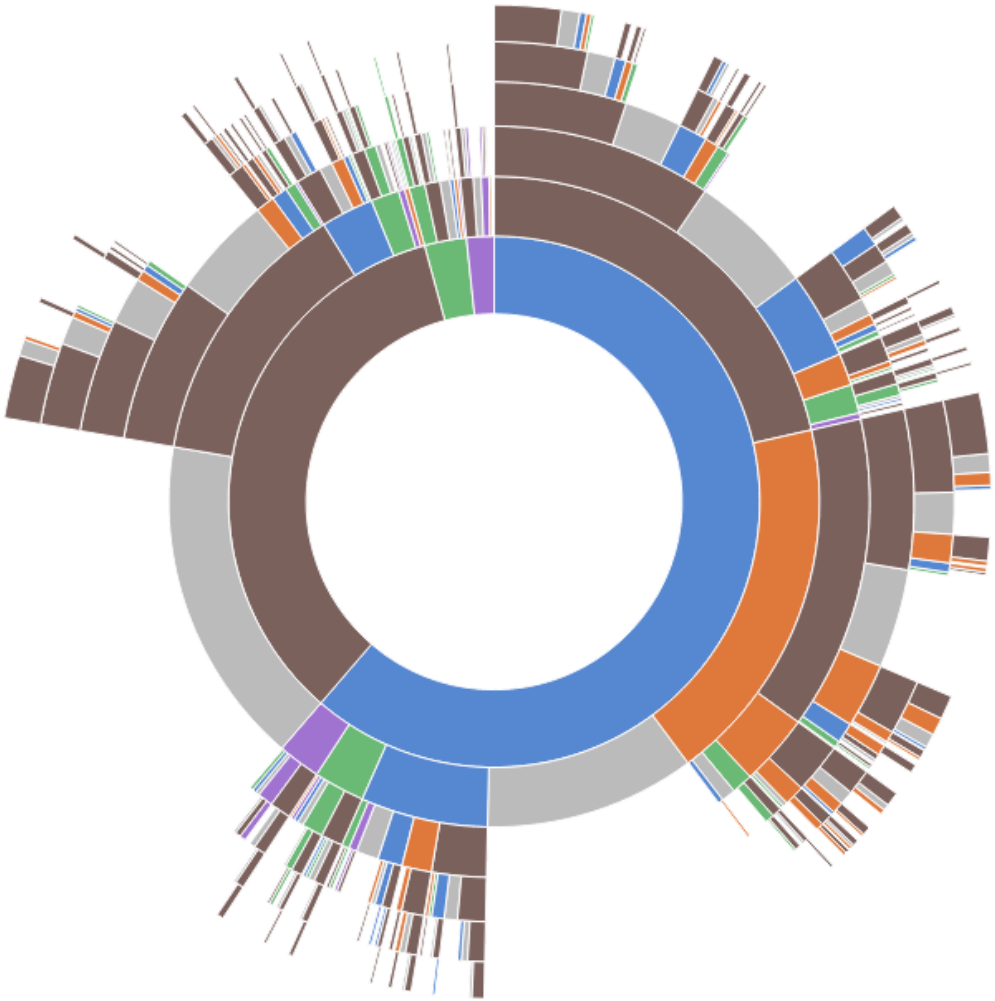
Big data visualization : Parallel coordinates

Plot a datum across many dimensions



C. Chen, C. Zhang, **Data-intensive applications, challenges, techniques and technologies : A survey on Big Data**, Information Sciences, 2014.

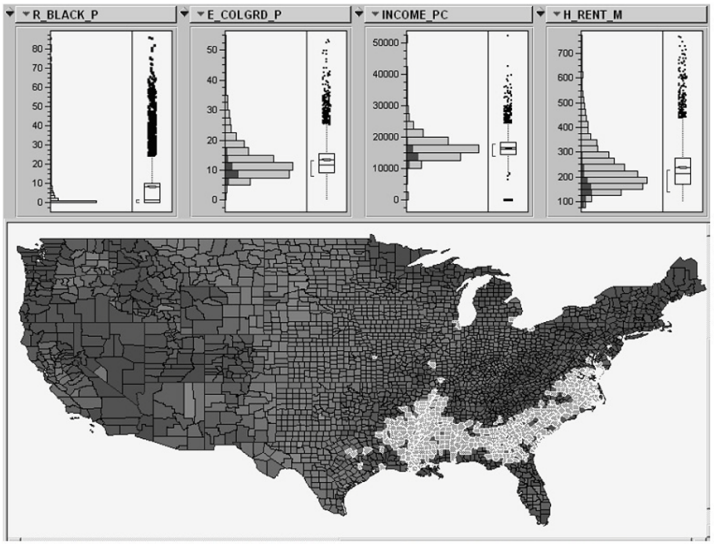
Big data visualization : Sequence sunburst



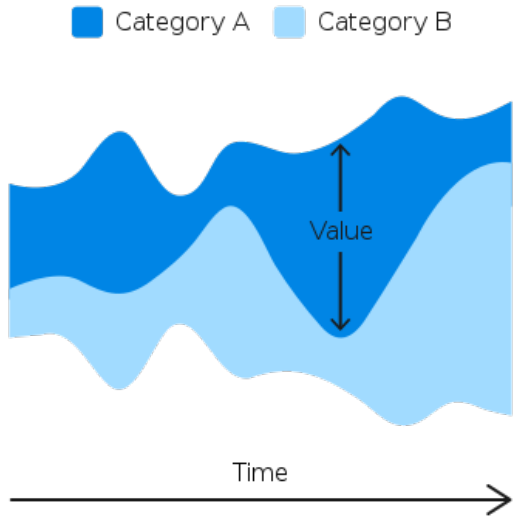
https ://bl.ocks.org/kerryrodden/7090426

Big data visualization

- all simple data visualization,
- Interactive visualization,
- time series : stream graph,...
- networks



M. Khan, S. Khan, **Data and Information Visualization Methods and Interactive Mechanisms : A Survey**, International Journal of Computer Applications, 2011.

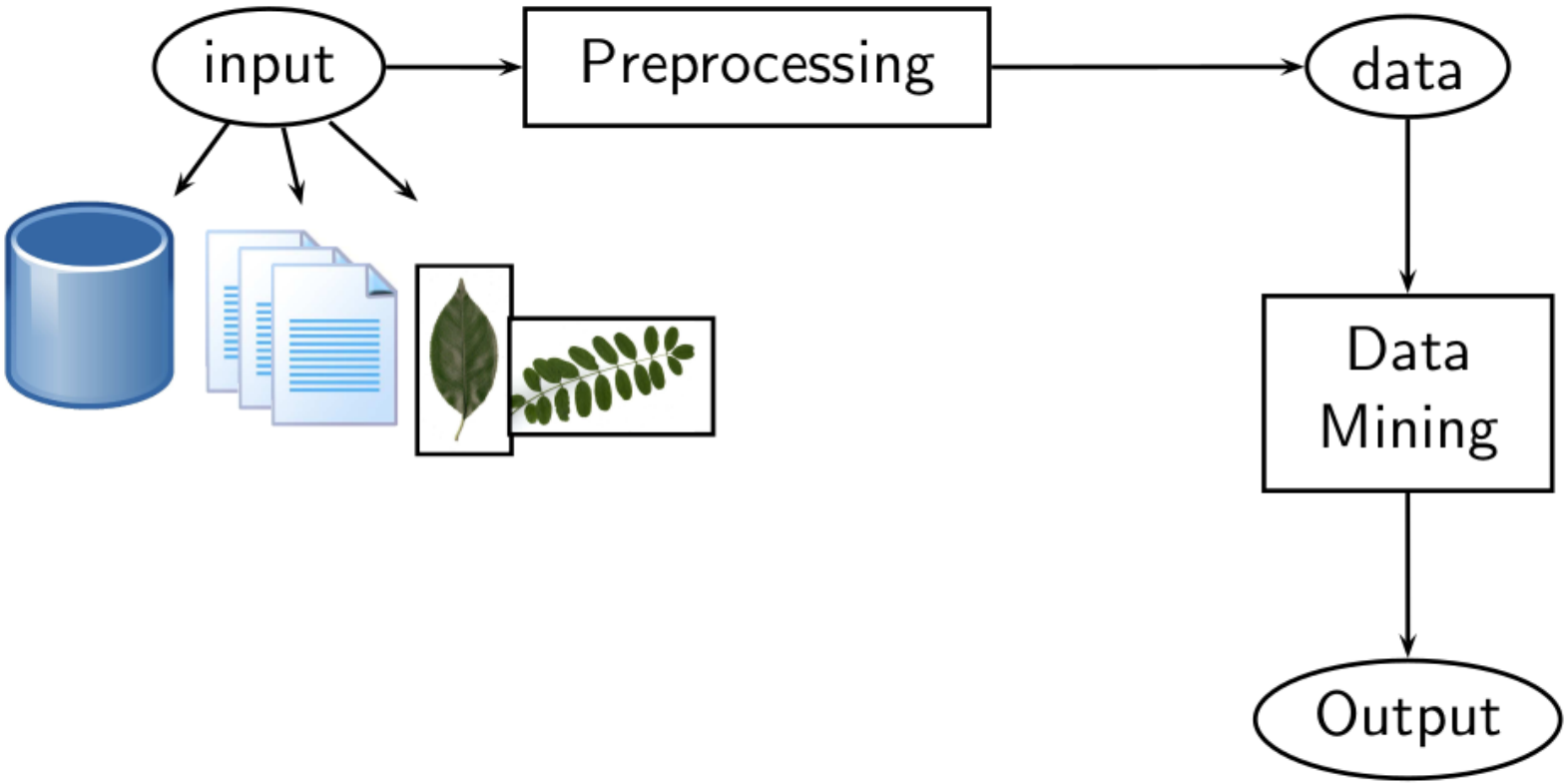


https ://datavizcatalogue.com/methods/stream_graph.html

Softwares for visualization

- Professional
  - Tableau
  - Qlikview
  - PowerBI
- Free and open source
  - Superset
  - Metabase

Data Preprocessing



Why preprocess data ?

Data project steps

D. Pyle , Data Preparation for Data Mining, Morgan Kaufmann Publishers Inc., 1999

- Explore problems : identify problems to solve

Why preprocess data ?

Data project steps

D. Pyle , Data Preparation for Data Mining, Morgan Kaufmann Publishers Inc., 1999

- Explore problems : identify problems to solve
- Explore solutions : identify known solutions, expected solutions

Why preprocess data ?

Data project steps

D. Pyle , Data Preparation for Data Mining, Morgan Kaufmann Publishers Inc., 1999

- Explore problems : identify problems to solve
- Explore solutions : identify known solutions, expected solutions
- Create added value on the discoveries

  know a priori how to get advantages of discories (sofware, decision making)

Why preprocess data ?

Data project steps

D. Pyle , Data Preparation for Data Mining, Morgan Kaufmann Publishers Inc., 1999

- Explore problems : identify problems to solve
- Explore solutions : identify known solutions, expected solutions
- Create added value on the discoveries

  know a priori how to get advantages of discories (sofware, decision making)
- Data preprocessing

Why preprocess data ?

Data project steps

D. Pyle , Data Preparation for Data Mining, Morgan Kaufmann Publishers Inc., 1999

- Explore problems : identify problems to solve
- Explore solutions : identify known solutions, expected solutions
- Create added value on the discoveries

  know a priori how to get advantages of discories (sofware, decision making)
- Data preprocessing
- Data surveying

  Check the good connection between the problems to solve and the dataset

Why preprocess data ?

Data project steps

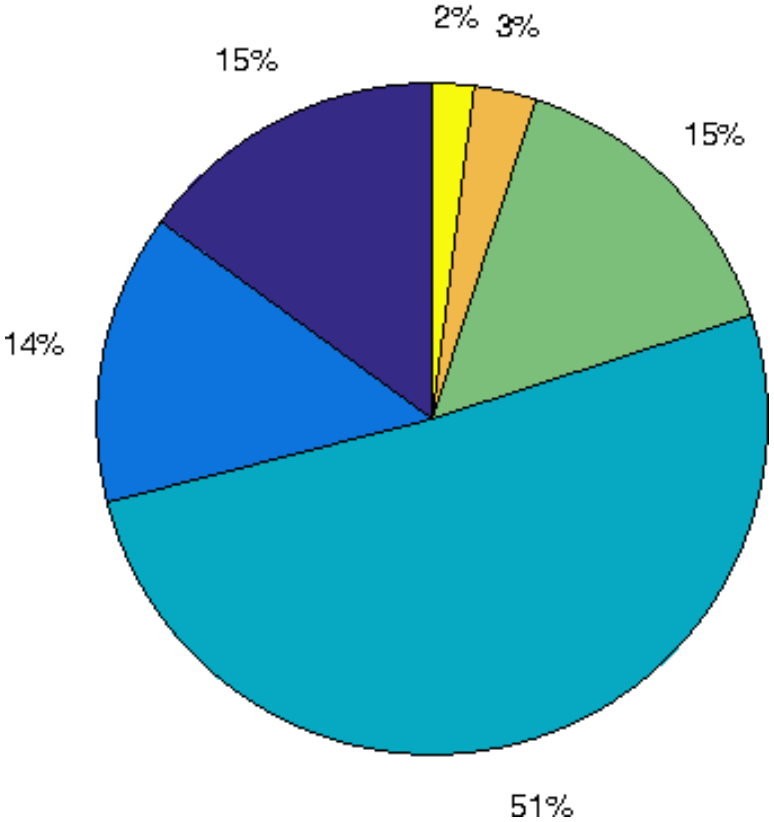
D. Pyle , Data Preparation for Data Mining, Morgan Kaufmann Publishers Inc., 1999

- Explore problems : identify problems to solve
- Explore solutions : identify known solutions, expected solutions
- Create added value on the discoveries

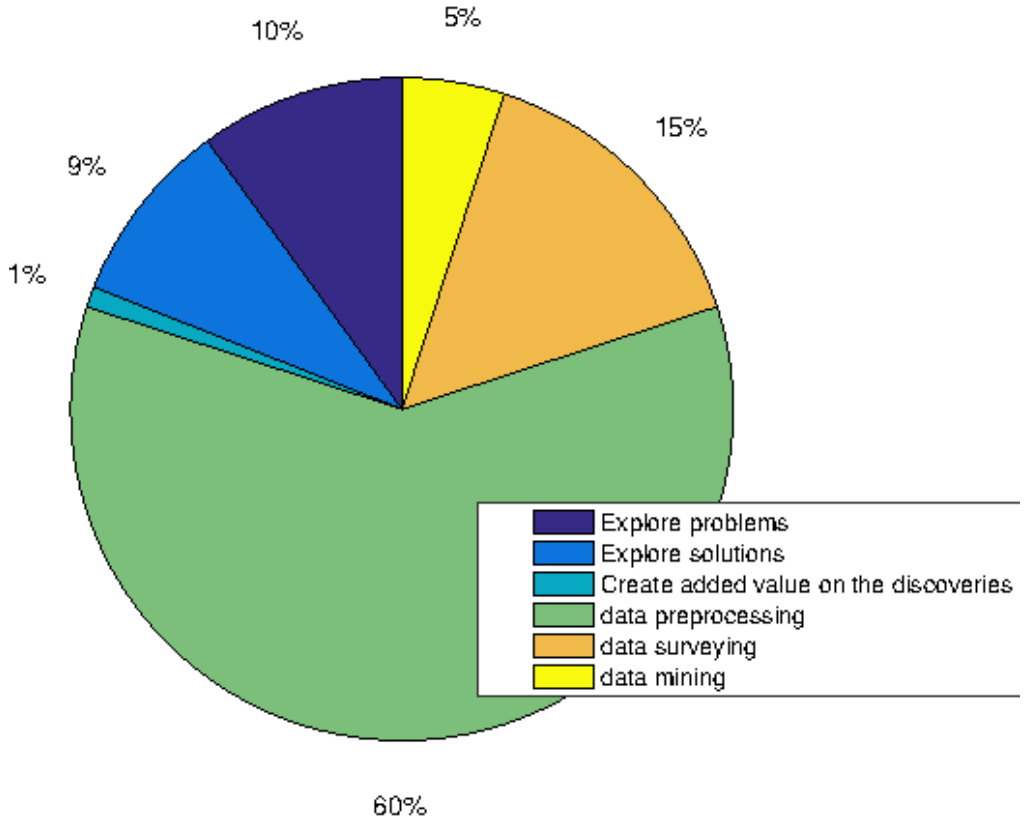
  know a priori how to get advantages of discories (sofware, decision making)
- Data preprocessing
- Data surveying

  Check the good connection between the problems to solve and the dataset
- Data mining : clustering, association rules,. . ..

Why preprocess data ?



importance of the project success



time to complete

Notations

Let us consider a dataset with quantitative variables such that $\mathbf{X} = (x_i^j)$

- $\mathbf{x}_i$ the object i,
- $\mathbf{x}^j$ the variable j,
- n the number of objects,
- p the number of attributes.

Outline

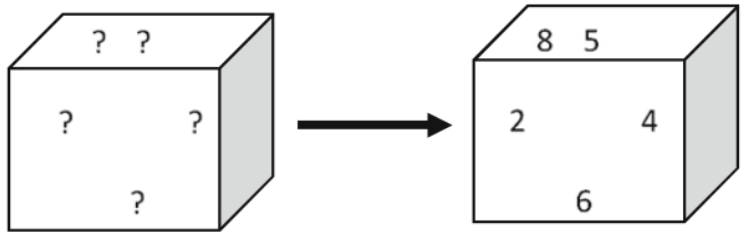
Cleaning methods

S. Garcia & al, Big data preprocessing : methods and prospects. Big Data Analytics, 2016.

Raw data have

- inconsistencies,
- missing values,
- noise,
- redundancies,. . .

Missing values

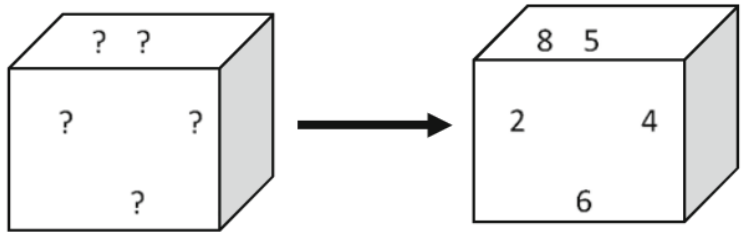


Missing value examples

- optional fields (in a questionaire)
- person refuse to answer
- data processing errors set to MV
  - abnormal values
  - data flow interrupted

			How many hours per day do you…			
	genre	age	...sleep ?	...watch TV ?	play sports ?	sport practiced
1	M	210	20	2	2	undefined
2	F	22				
3	F		8	1	0	undefined
4	M	20	7	3	1	baseball

Missing values



Missing value examples

- optional fields (in a questionaire)
- person refuse to answer
- data processing errors set to MV
  - abnormal values
  - data flow interrupted

			How many hours per day do you…			
	genre	age	...sleep ?	...watch TV ?	play sports ?	sport practiced
1	M	**MV**	**MV**	2	2	**MV**
2	F	22	**MV**	**MV**	**MV**	**MV**
3	F	**MV**	8	1	0	**MV**
4	M	20	7	3	1	baseball

Types of Missing Values

- Missing completely at random (MCAR)
  - missing value x neither depends on y or z
- Missing at random (MAR)
  - missing value x depends on y but not z
- Not missing at random (NMAR)
  - Missingness depends on unobserved predictors
  - Missingness depends on the missing value itself

Types of Missing Values

- Missing completely at random (MCAR)
  - missing value x neither depends on y or z
- Missing at random (MAR)
  - missing value x depends on y but not z
- Not missing at random (NMAR)
  - Missingness depends on unobserved predictors
  - Missingness depends on the missing value itself

Test the type of missing value :

- t-tests for MCAR but not totally accurate
- General impossibility to prove that data are MAR or NMAR !

Types of Missing Values

- MCAR is very unlikely to occur
- In empirical studies data are often NMAR but in an afterthought

$\Rightarrow$ Data are usually treated as MAR

Missing values methods

- Discard instances/variables
  - Listwise deletion
  - Pairwise deletion
- Imput a value
  - Single imputation methods
    - the mean / mode [of a group]
    - hot deck, KNN
    - the regression
  - Model-based method
    - multiple imputation method
    - maximum likelihood method

Missing values methods

Not handling MV may lead to

- biased estimates,
- incorrect standard errors,
- ...

Goal of MV methods are

- NOT correct prediction of MV
- but obtain accurate parameters estimates.

Discard values : listwise deletion

Remove lines or columns containing MV :

+ simple
- reduces the dataset
- create bias if MV are not MCAR

			How many hours per day do you…			
	gender	age	...sleep ?	...watch TV ?	play sports ?	sport practiced
1	M	**MV**	7	2	2	hockey
2	F	22	7	**MV**	**MV**	**MV**
3	F	23	8	1	0,5	dance
4	M	20	7	3	1	baseball
5	M	25	8	3	0,5	treck
6	M	30	8	3	1	baseball
7	F	**MV**	7	2	1	football
8	F	22	5	1	2	basketball
9	F	24	9	0	5	basketball
10	M	20	7	4	0	hockey

Discard values : pairwise deletion

Keep MV for the analysis :

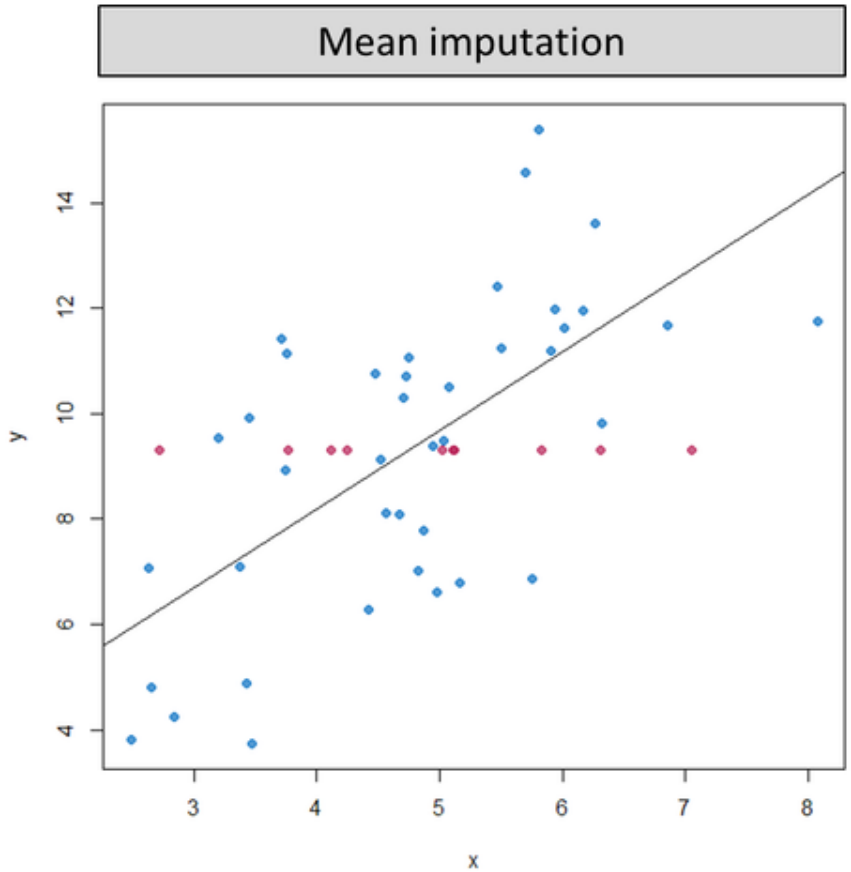
+ keeps all information
- analysis methods should handle the MV

			How many hours per day do you…			
	gender	age	...sleep ?	...watch TV ?	play sports ?	sport practiced
1	M	**MV**	7	2	2	hockey
2	F	22	7	**MV**	**MV**	**MV**
3	F	23	8	1	0,5	dance
4	M	20	7	3	1	baseball
5	M	25	8	3	0,5	treck
6	M	30	8	3	1	baseball
7	F	**MV**	7	2	1	football
8	F	22	5	1	2	basketball
9	F	24	9	0	5	basketball
10	M	20	7	4	0	hockey

Single imputation values : mean/mode [of a group]

			How many hours per day do you…			
	gender	age	...sleep ?	...watch TV ?	play sports ?	sport practiced
1	M	**MV**	7	2	2	hockey
2	F	22	7	**MV**	**MV**	**MV**
3	F	23	8	1	0,5	dance
4	M	20	7	3	1	baseball
5	M	25	8	3	0,5	treck
6	M	30	8	3	1	baseball
7	F	**MV**	7	2	1	football
8	F	22	5	1	2	basketball
9	F	24	9	0	5	basketball
10	M	20	7	4	0	hockey
mean		23,25	7,30	2,11	1,44	
median		22,5	7	2	1	
mode		20	7	3	1	

Single imputation values : mean/mode [of a group]



www.iriseekhout.com/missing-data/missing-data-methods/imputation-methods

\+ simple

\- underestimates the standard deviation

\- ignore possible relationship with other variables

Single imputation values : hot deck and KNN

KNN

- Pick a value from the observation the closest to the object with MV

Hot deck

- Select at random a value in a set closed observation

			How many hours per day do you…			
	gender	age	...sleep ?	...watch TV ?	play sports ?	sport practiced
1	M	**MV**	7	2	2	hockey
2	F	22	7	**MV**	**MV**	**MV**
3	F	23	8	1	0,5	dance
4	M	20	7	3	1	baseball
5	M	25	8	3	0,5	treck
6	M	30	8	3	1	baseball
7	F	25	7	2	1	football
8	F	22	5	1	2	basketball
9	F	24	9	0	5	basketball
10	M	20	7	4	0	hockey

- suffers from little variation in the dataset

Single imputation values : hot deck and KNN

KNN

- Pick a value from the observation the closest to the object with MV

Hot deck

- Select at random a value in a set closed observation

			How many hours per day do you…			
	gender	age	...sleep ?	...watch TV ?	play sports ?	sport practiced
1	M	**30**	7	2	2	hockey
2	F	22	7	**1**	**2**	**basketball**
3	F	23	8	1	0,5	dance
4	M	20	7	3	1	baseball
5	M	25	8	3	0,5	treck
6	M	30	8	3	1	baseball
7	F	25	7	2	1	football
8	F	22	5	1	2	basketball
9	F	24	9	0	5	basketball
10	M	20	7	4	0	hockey

- suffers from little variation in the dataset

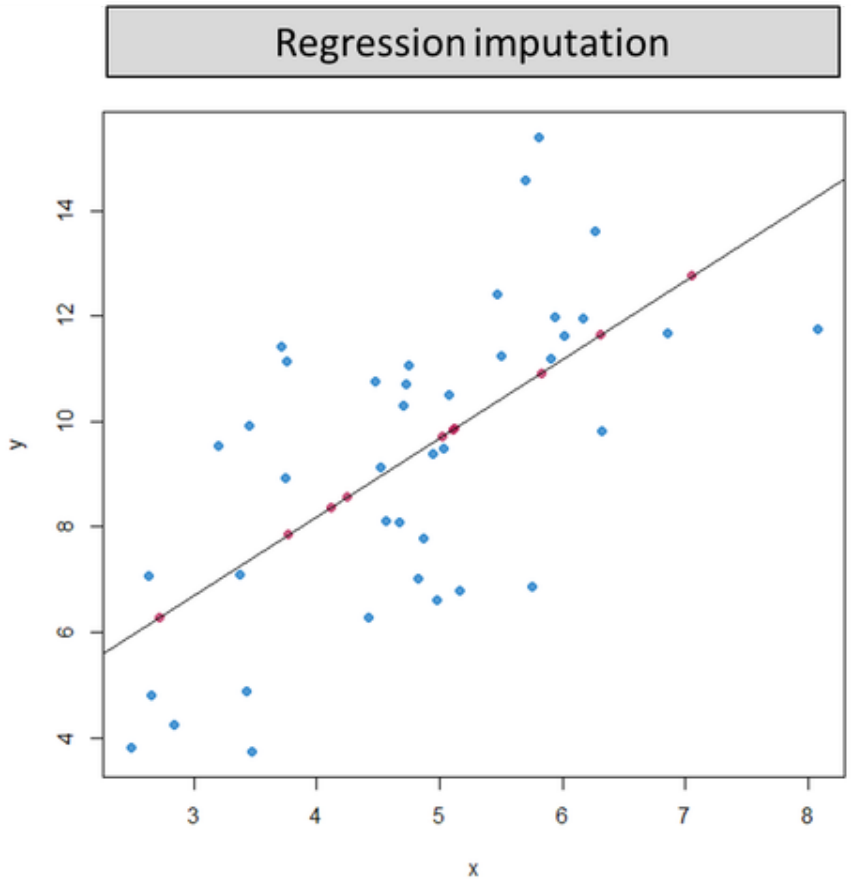
Single imputation values : regression

$$z = a_0 + \sum_{j=1}^{J} a_j x^j,$$

s.t.

- z the variable with missing values,
- $x^j \in [1 \dots J]$ the set of variables fully observed,
- $a_j \in [0 \dots J]$ the coefficients for linear regression.

Single imputation values : regression



www.iriseekhout.com/missing-data/missing-data-methods/imputation-methods

\+ preserve correlation with other variables

\- variability of missing values is underestimated

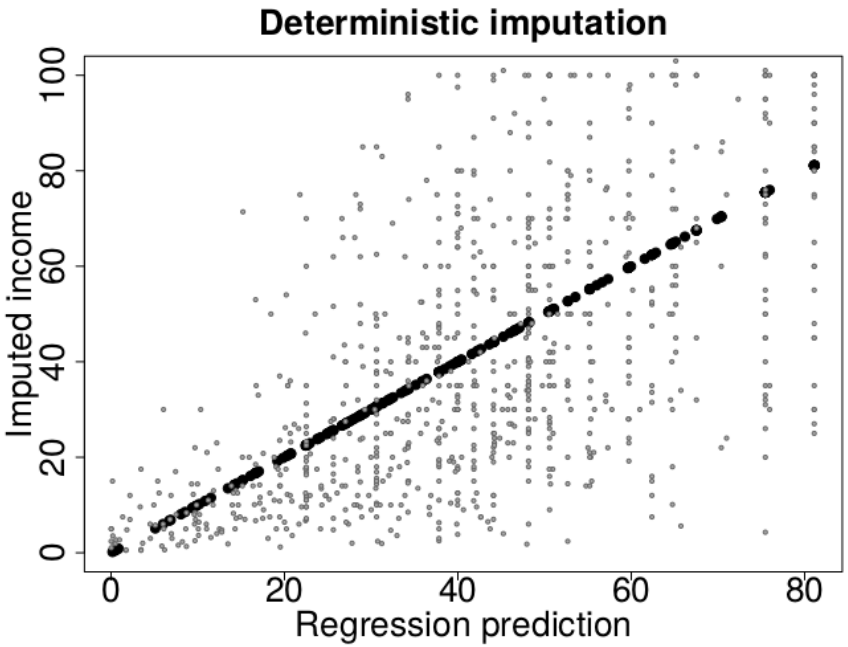
Random regression imputation

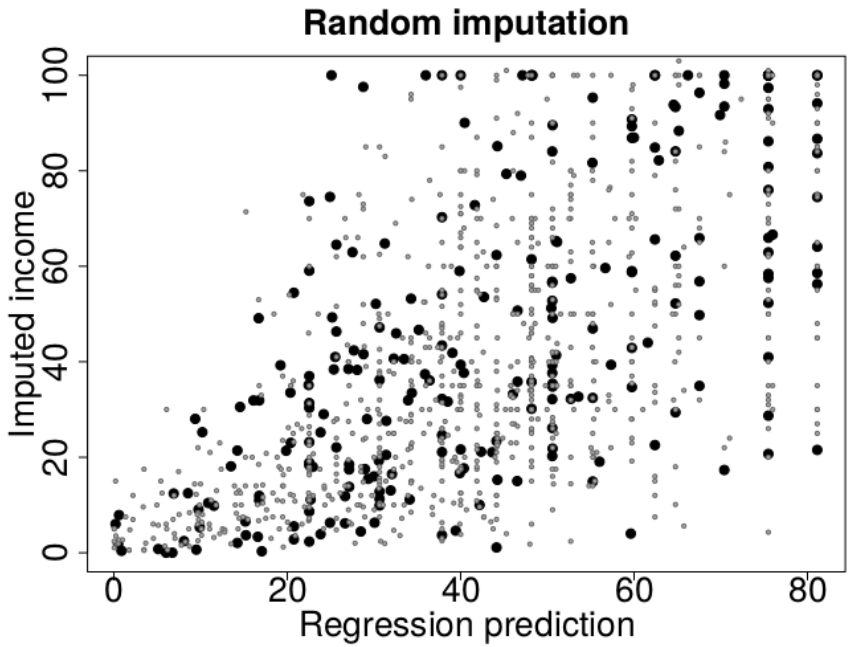
$$z = a_0 + \left(\sum_{j=1}^{J} a_j x^j \right) + sE,$$

s.t.

- z the variable with missing values,
- $x^j \in [1 \dots J]$ the set of variables fully observed,
- $a_j \in [0 \dots J]$ the coefficients for linear regression,
- s estimated standard deviation
- E the random draw from a standard normal distribution $\mathcal{N}(0, 1)$

Random regression imputation





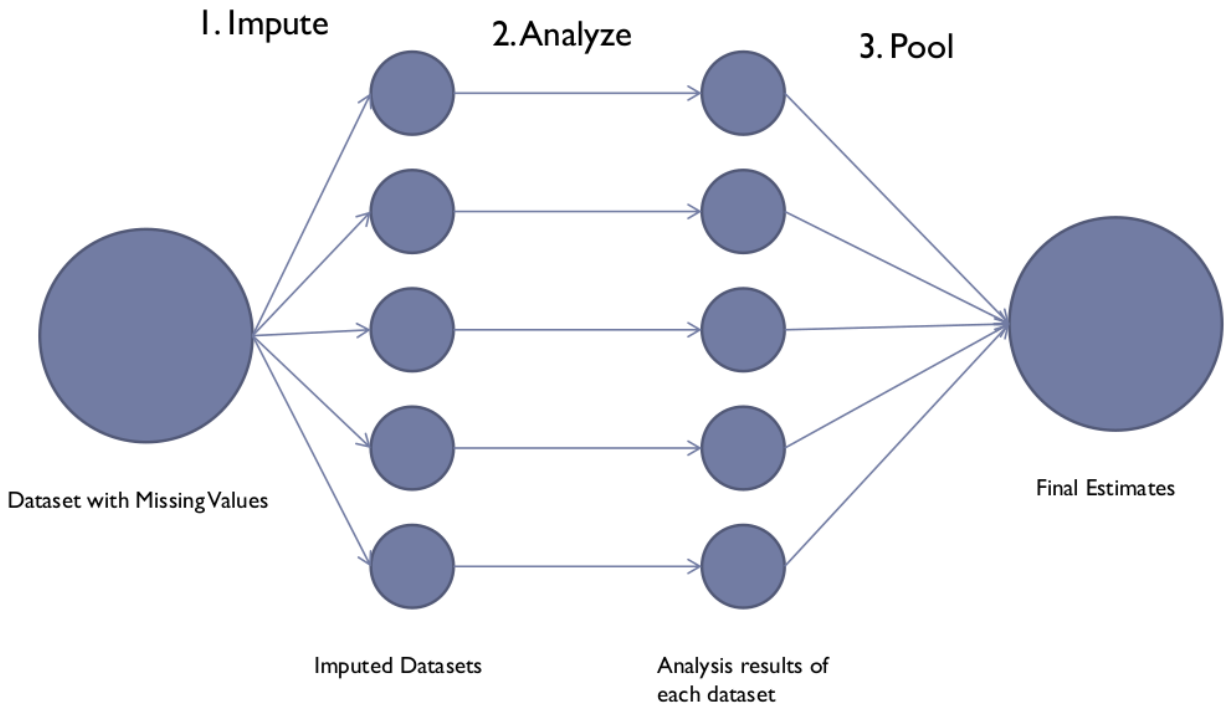
A. Gelman, J. Hill. Data Analysis Using Regression and Multilevel/Hierarchical Models, chapter 25

- increases noise particularly if the number of MV is large

Multiple imputation method

Method :

1. create m datasets with random regression imputation,
   usually $m \in [2 \dots 10]$
2. analyse the m datasets,
3. combine results
   e.g. estimate regression coefficients



M. Humphries, Missing Data & How to Deal : An overview of missing data

Maximum likelihood method

Identifies the set of parameter values θ that produces the maximal log-likelihood.

$$\mathcal{L}(\theta; \mathbf{x}_1, \ldots, \mathbf{x}_m) = \prod_{i=1}^{n} f(\mathbf{x}_i|\theta) \prod_{i=n+1}^{m} f^*(\mathbf{x}_i|\theta)$$

- $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are observed data
- $\mathbf{x}_{n+1}, \ldots, \mathbf{x}_m$ are objects with MV

\+ unbiased estimates and standard errors under MAR or MCAR

\+ deterministic compar to Multiple Imputation

Datasets with noise

Two types of noise :

- attribute noise
  - → errors giving outliers or inconsistant data
  - → missing values
  - → incomplete attributes
- class noise
  - → subjectivity during the labeling process
  - → inadequacy of the information used to label
  - → data entry errors

Two types of methods to handle noise :

- filter noise methods (i.e. eliminate noise)
- data polishing methods (i.e. correct noise)

Eliminate or correct noise enables more accurate analysis

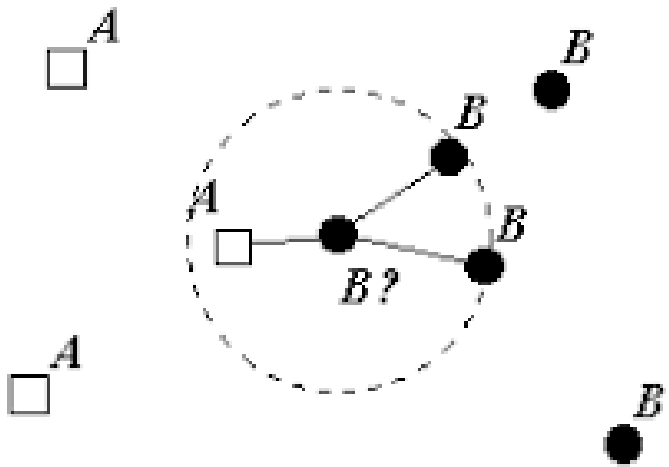
Attribute noise : detection methods

- outliers detection methods
- boxplots, histograms on attributes

Class noise : Edited Nearest Neighbor (ENN) method

For each instance $\mathbf{x}_i$ in the training set $\mathcal{L}$

1. Remove $\mathbf{x}_i$ from $\mathcal{L}$
2. Apply k-NN and new label of $\mathbf{x}_i$
3. Marked $\mathbf{x}_i$ for deletion/polishing if predicted class $\neq$ true class



Edited Nearest Neighbor (ENN) method

Many variants :

- RENN
  - repeats ENN until a stable set $\mathcal{L}$ is obtained

Edited Nearest Neighbor (ENN) method

Many variants :

- RENN
  - repeats ENN until a stable set $\mathcal{L}$ is obtained
- all k-NN
  - applies all ENN with i-NN rule s.t. $i \in [1 \dots k]$

Edited Nearest Neighbor (ENN) method

Many variants :

- RENN
  - repeats ENN until a stable set $\mathcal{L}$ is obtained
- all k-NN
  - applies all ENN with i-NN rule s.t. $i \in [1 \dots k]$
- Modified Edited Nearest Neighbor
  - rule to detect noisy objects also takes in account tying instances

Edited Nearest Neighbor (ENN) method

Many variants :

- RENN
  - repeats ENN until a stable set $\mathcal{L}$ is obtained
- all k-NN
  - applies all ENN with i-NN rule s.t. $i \in [1 \dots k]$
- Modified Edited Nearest Neighbor
  - rule to detect noisy objects also takes in account tying instances
- Multiedit
  - randomly breaks the initial training set $\mathcal{L}$ into different subsets

Edited Nearest Neighbor (ENN) method

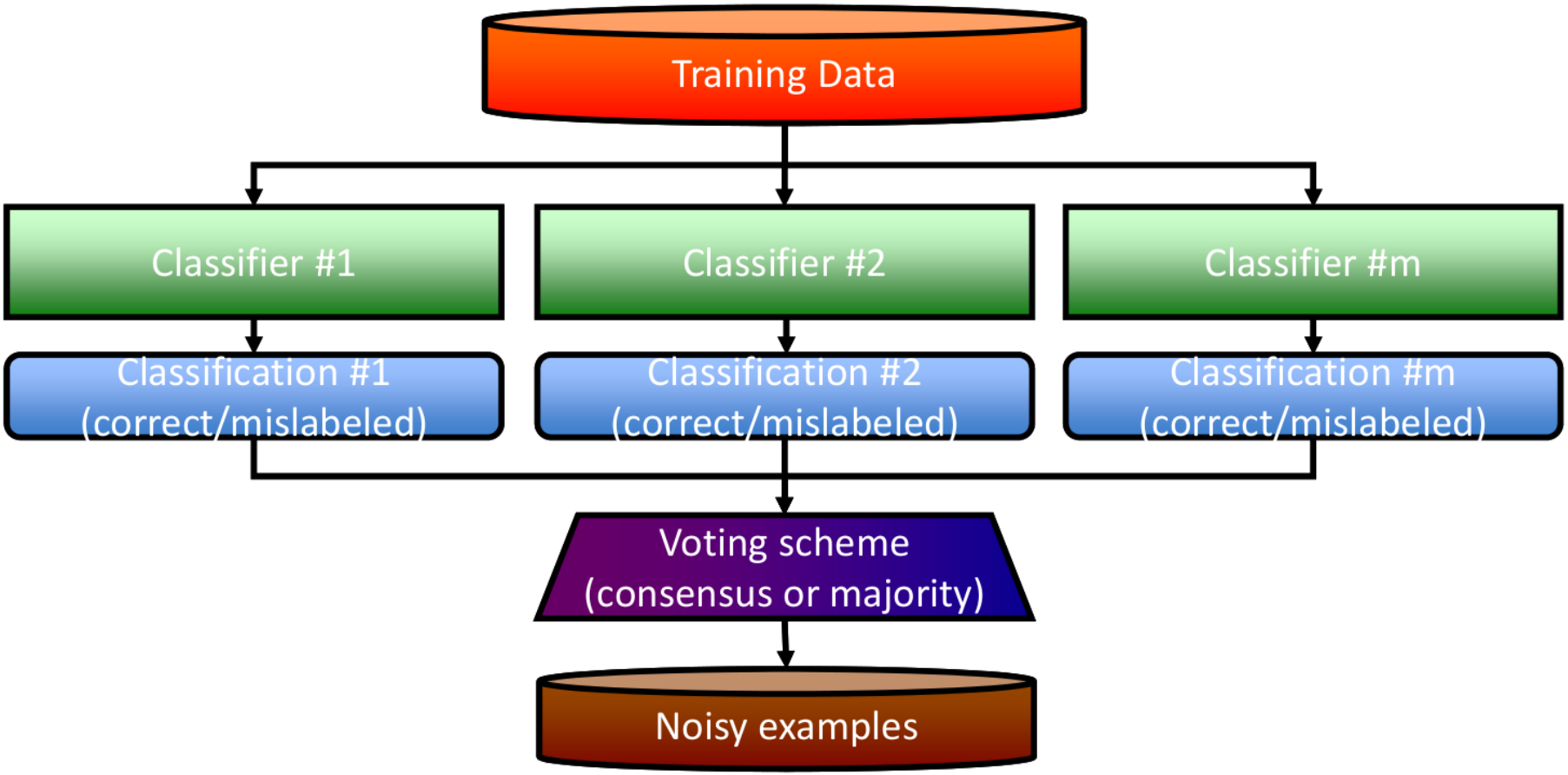
Many variants :

- RENN
  - repeats ENN until a stable set $\mathcal{L}$ is obtained
- all k-NN
  - applies all ENN with i-NN rule s.t. $i \in [1 \dots k]$
- Modified Edited Nearest Neighbor
  - rule to detect noisy objects also takes in account tying instances
- Multiedit
  - randomly breaks the initial training set $\mathcal{L}$ into different subsets
- Nearest Neighbor Editing Aided by Unlabelled Data
  - labels are first predicted for unlabelled instances

References and more details in http ://sci2s.ugr.es/noisydata

Ensemble Filter (EF) method

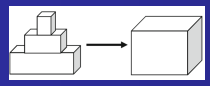
ENN variants have been generalized for any learning algorithm



F. Herrera, Data Mining Methods for Big Data Preprocessing, Summer school on ML, 2015

Outline

Normalization



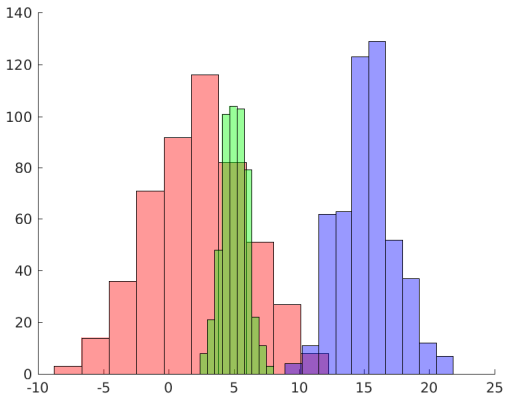
Let $\mathbf{x}$ be the j^{th} variable :

- Feature scaling : $z_i \in [0 \dots 1]$

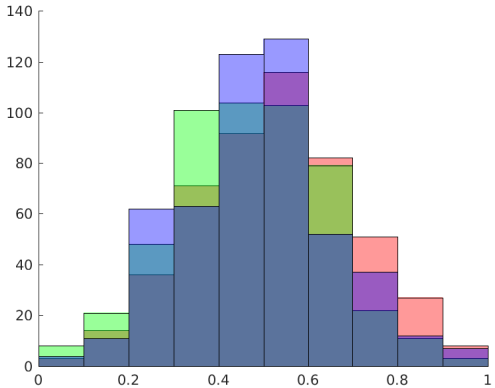
$$z_i = \frac{x_i - \min(\mathbf{x})}{\max(\mathbf{x}) - \min(\mathbf{x})}$$

- Student's t-statistic : $\bar{\mathbf{z}} = 0$, $\sigma_{\mathbf{z}} = 1$

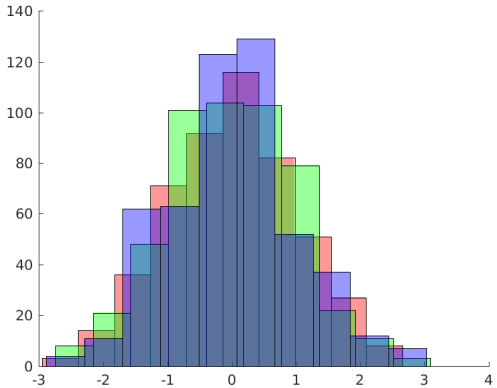
$$z_i = \frac{x_i - \bar{\mathbf{x}}}{\sigma_{\mathbf{x}}}$$



x



feature scaling



student's t-statistic

Generalization

Modify data to higher degrees of data by using concept hierarchies

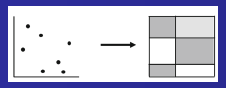
Examples :

- streets $\rightarrow$ cities
- animals $\rightarrow$ species

\+ data simplification and reduction

\+ data readability

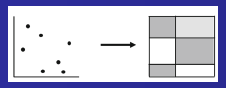
\- generate a loss of information

Discretization 

Type of generalization : quantitative data $\rightarrow$ qualitative data

			How many hours per day do you…			
	gender	age	...sleep ?	...watch TV ?	play sports ?	sport practiced
1	M	21	6	2	2	dance
2	F	12	5	0	0,2	swimming
3	F	14	8	1	0,5	baseball
4	M	20	7	3	1	baseball
1	M	21	6	2	2	hockey
2	F	25	7	0	3	basketball
3	F	10	8	1	0,1	softball
4	M	60	7	3	1	baseball

+ needed for algorithms which only accept discrete data, i.e association rules or DT.

Discretization 

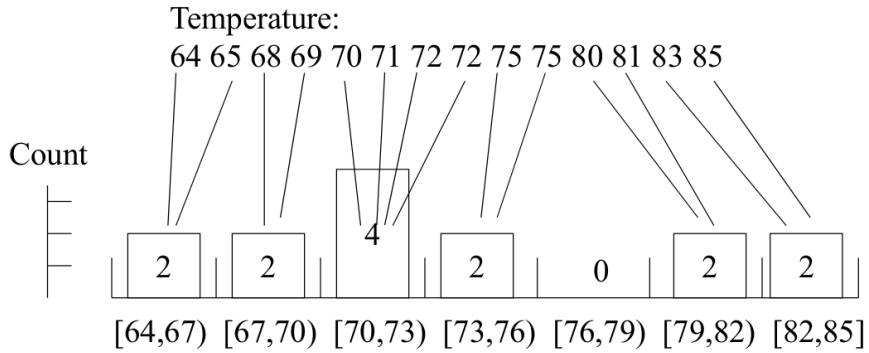
Type of generalization : quantitative data $\rightarrow$ qualitative data

			How many hours per day do you…			
	gender	age	...sleep ?	...watch TV ?	play sports ?	sport practiced
1	M	normal	6	2	2	dance
2	F	young	5	0	0,2	swimming
3	F	young	8	1	0,5	baseball
4	M	normal	7	3	1	baseball
1	M	normal	6	2	2	hockey
2	F	normal	7	0	3	basketball
3	F	young	8	1	0,1	softball
4	M	senior	7	3	1	baseball

+ needed for algorithms which only accept discrete data, i.e association rules or DT.

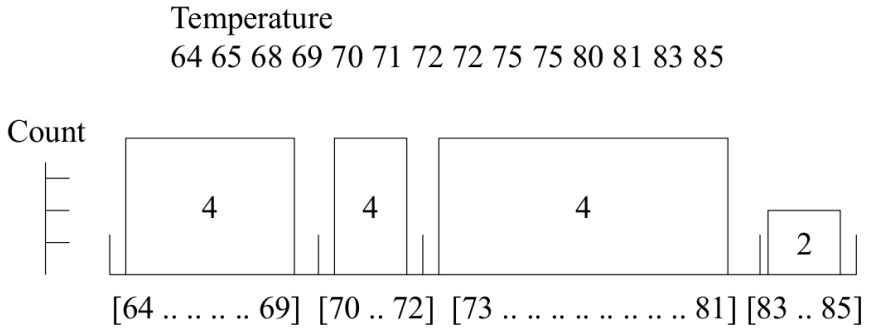
Discretization methods : unsupervised algorithms

- equal width (for normal or uniform distribution)



Equal width

- equal frequency



Equal frequency (height) = 4, except for the last box

F. Herrera, Data Mining Methods for Big Data Preprocessing, Summer school on ML, 2015

- clustering

Discretization methods : supervised algorithms

- Chi-squared
  - statistical test the relationship between a feature and classes
- Entropy
  - class information entropy
- DT
- Wrapper based (See feature selection)
- Evolutionary based

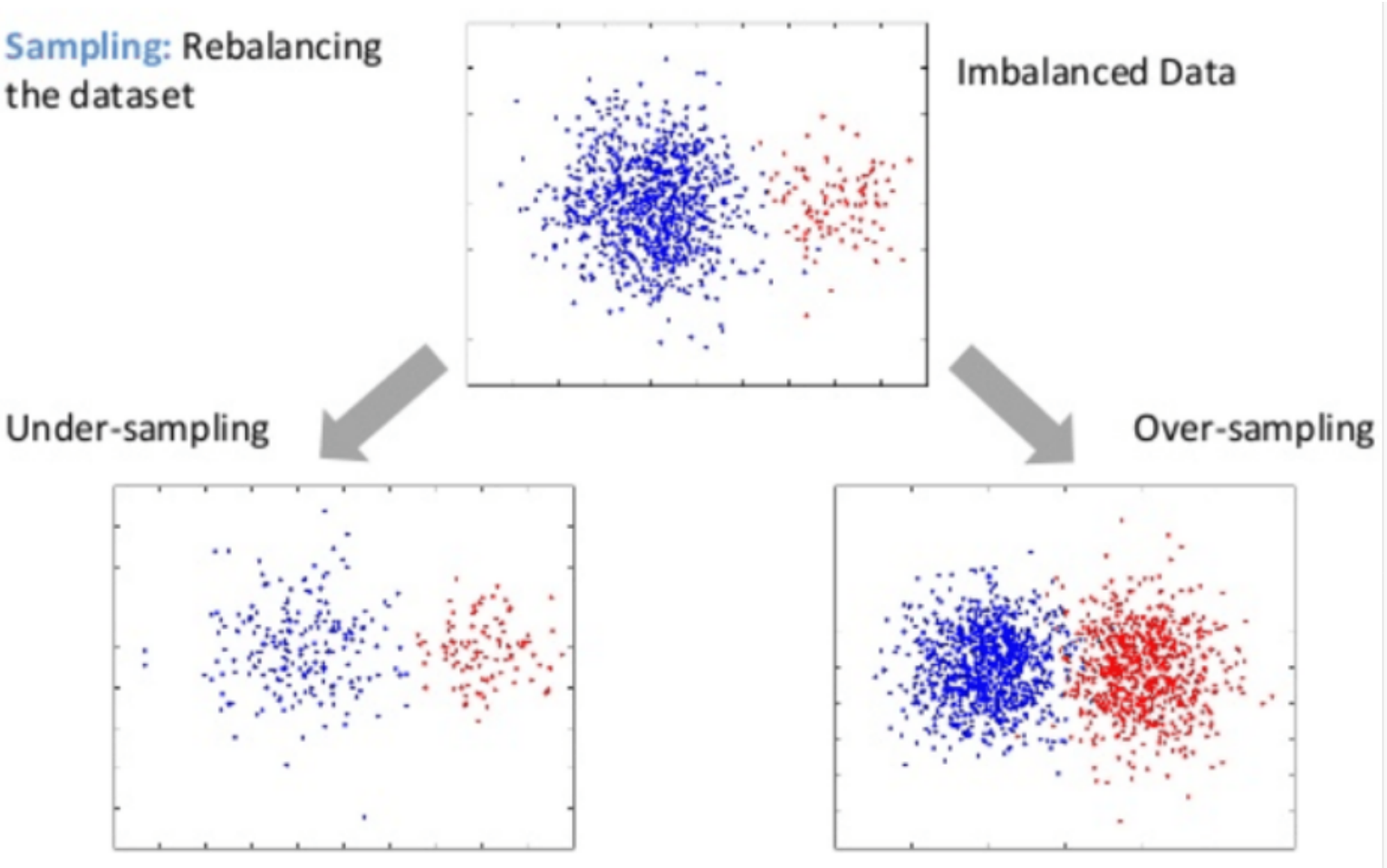
Instance generation / Instance selection

Goal : fill region with no samples, remove noisy data

- Instance generation (or prototype generation)
  - generates original data with new artificial data
- Instance selection
  - generate minimum data subset without losing performance
- Weighting instances

Undersampling / Oversampling

Goal : obtain balanced classes



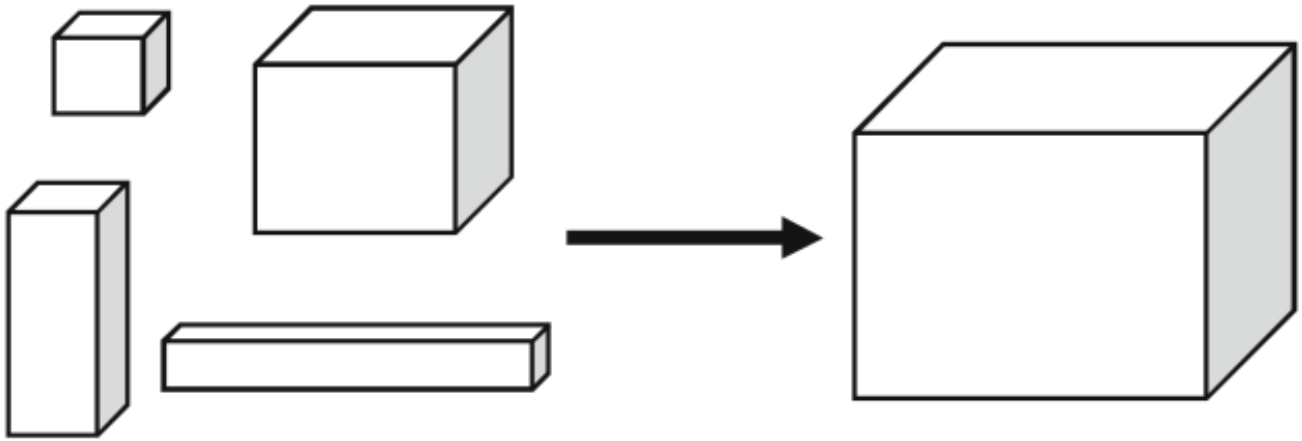
www.srutisj.in/blog/research/statisticalmodeling/balancing-techniques-for-unbalanced-datasets-in-python-r/

Undersampling / Oversampling methods

- Most popular oversampling algorithm : SMOTE
  1. Select a point $\mathbf{x}_i$ in the minority class,
  2. Select $\mathbf{x}_1, \ldots \mathbf{x}_k$ its KNN,
  3. Pick up a neighbor $\mathbf{x}_j$ at random,
  4. Create new point $\mathbf{y} = \mathbf{x}_i + \alpha(\mathbf{x}_i - \mathbf{x}_j)$, s.t. $\alpha \in [0, 1]$.
- Most popular undersampling : random undersampling

Integration

Combine data from different sources



Problems

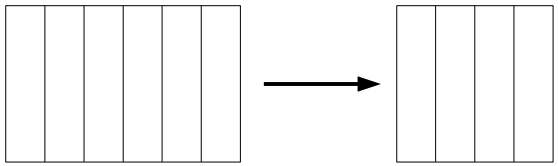
- identity of qualitative data (ex : basket, basketball, Basket...)
- data conflict
- data redundancy

Outline

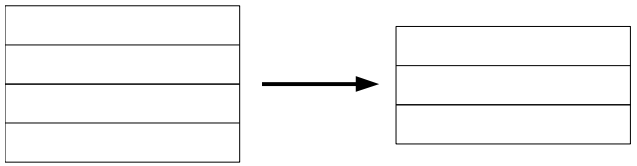
Dimensionality reduction

- Selection methods

attributes selection

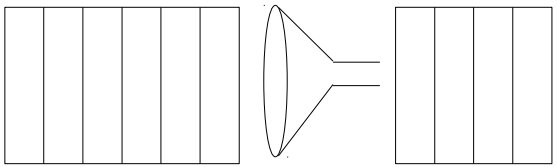


instance reduction



- extraction methods

attributes extraction



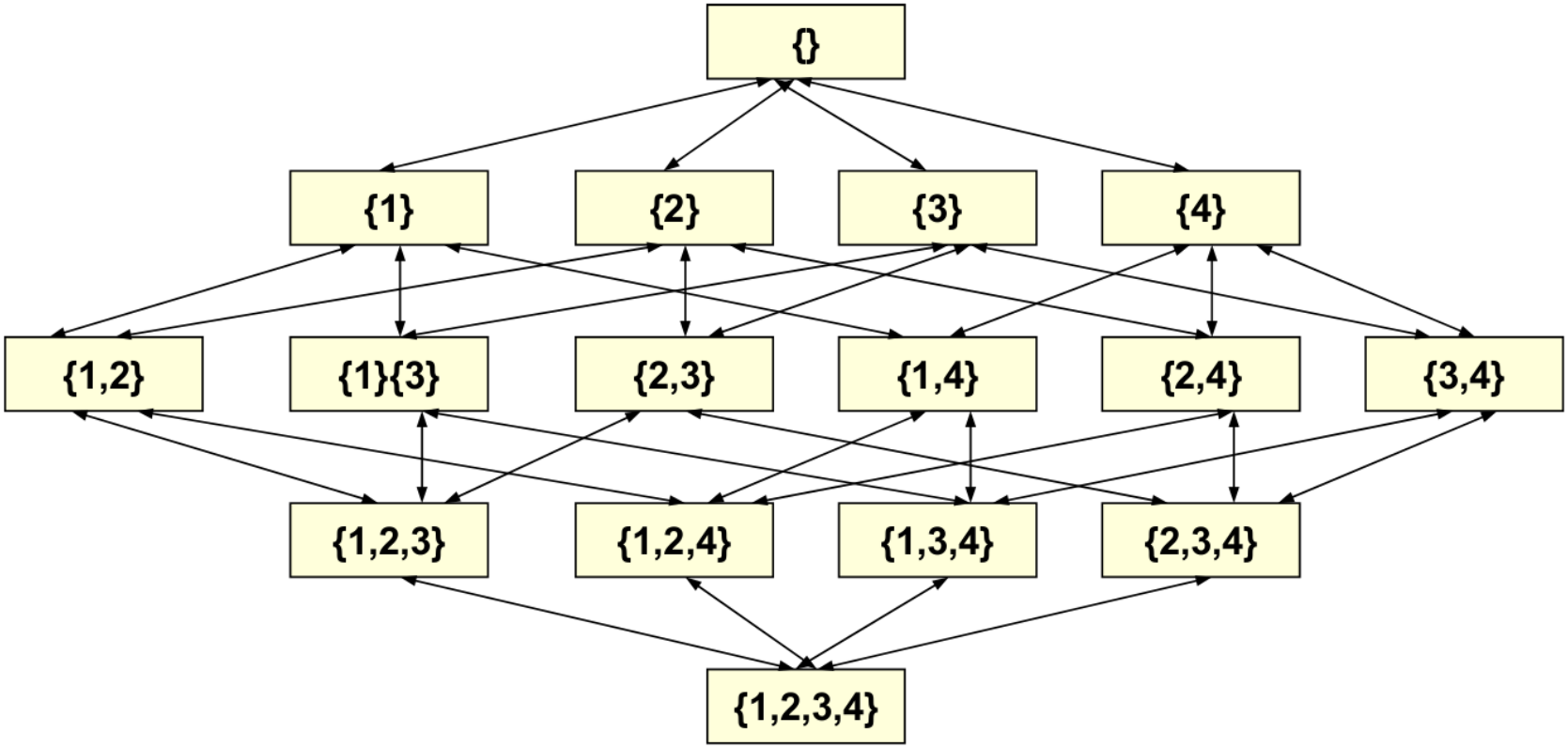
Attributes selection

Goal : rank/select a subset features that provides the best analysis.

- \+ reduce the complexity of the problem
- \+ remove useless attributes that may bias results
- \+ keep accurate and understandable attributes

Attributes selection

Exhaustive search corresponds to $2^p - 1$ subsets to evaluate !



F. Herrera, Data Mining Methods for Big Data Preprocessing, Summer school on ML, 2015

$\Rightarrow$ Heuristics have been developed in $O(p^2)$

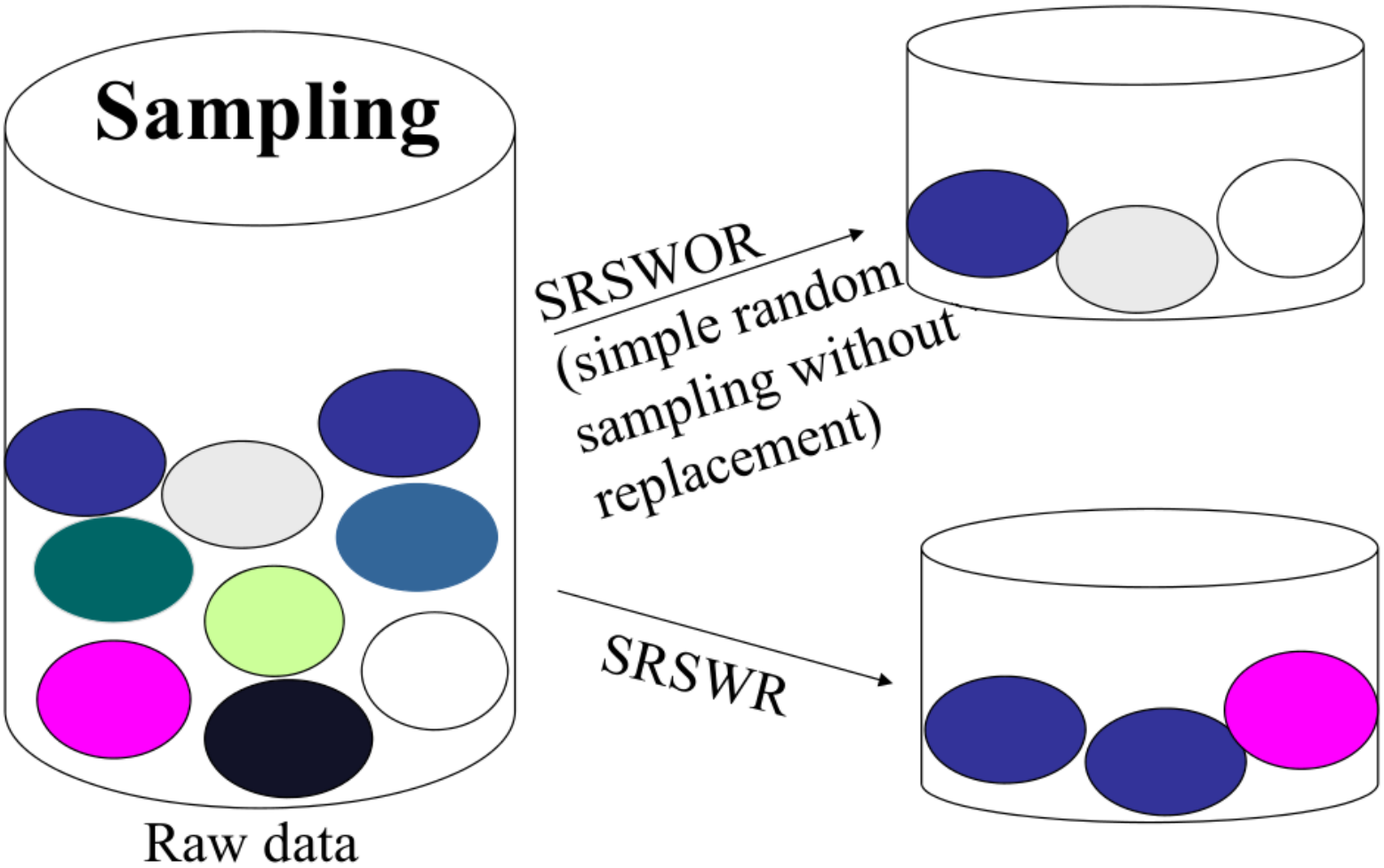
Instance selection

Select a representative subset of the data (random, filter, wrapper...)

Types of random sampling :

- Sampling without replacement
  - The object selected is removed from the initial dataset.
- Sampling with replacement
  - The object selected is not removed from the initial dataset.
- Stratified sampling
  - Partition the initial dataset to select objects

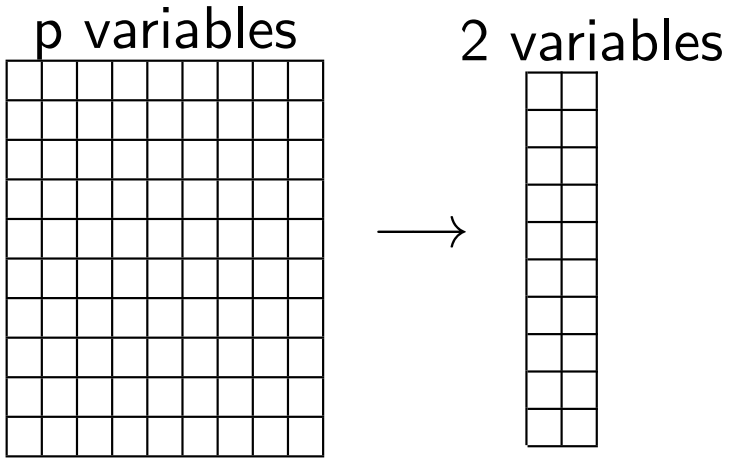
Instance selection



F. Herrera, Data Mining Methods for Big Data Preprocessing, Summer school on ML, 2015

Attributes extraction

Goal : concat variables while removing redondancy and keeping maximum of information.

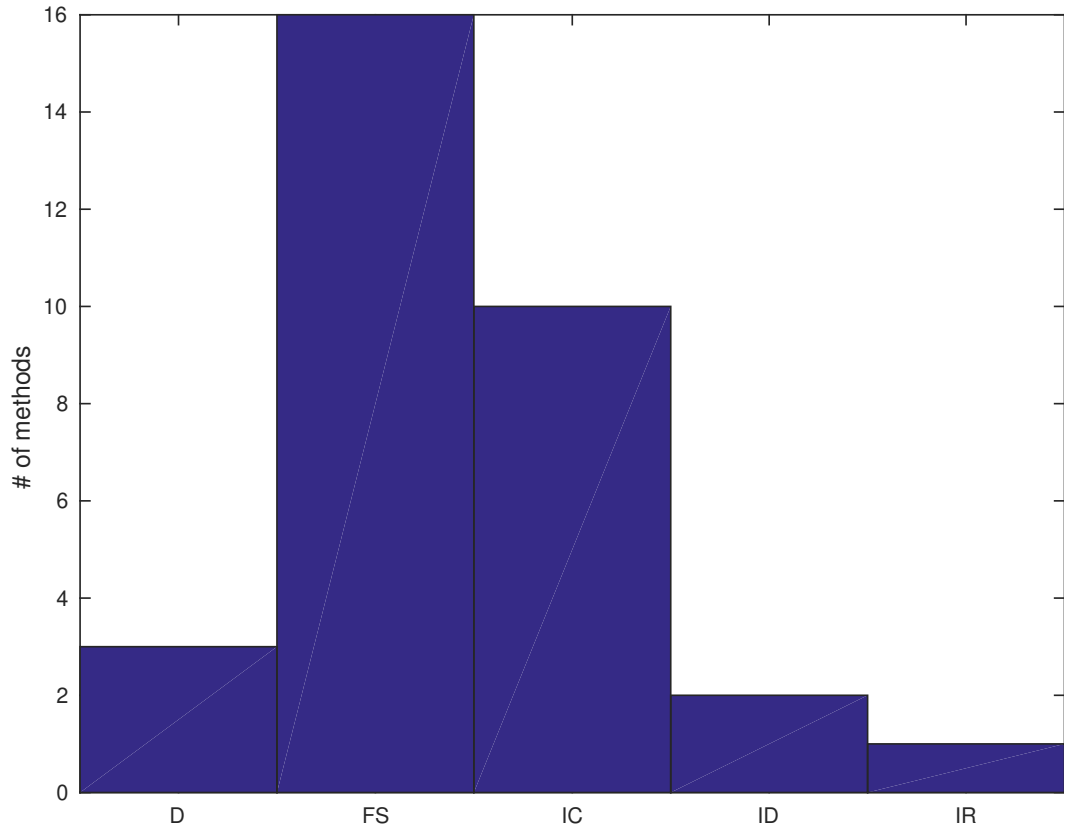


- linear methods
  - Principal Component analysis (PCA)
  - Multidimensional scaling (MDS)
- non linear methods
  - ISO Map
  - LLE
  - manyfold learning : Sammon mapping, SOM, autoencoders,...

Outline

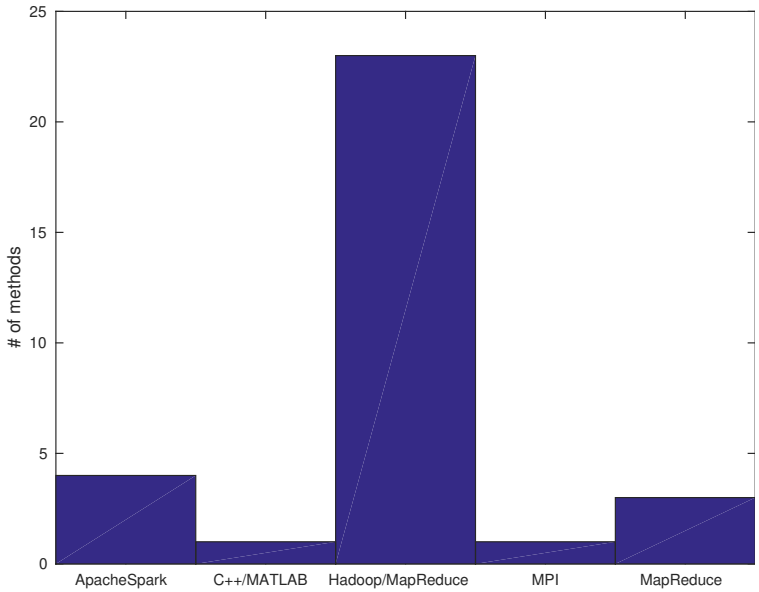
Big data and preprocessing

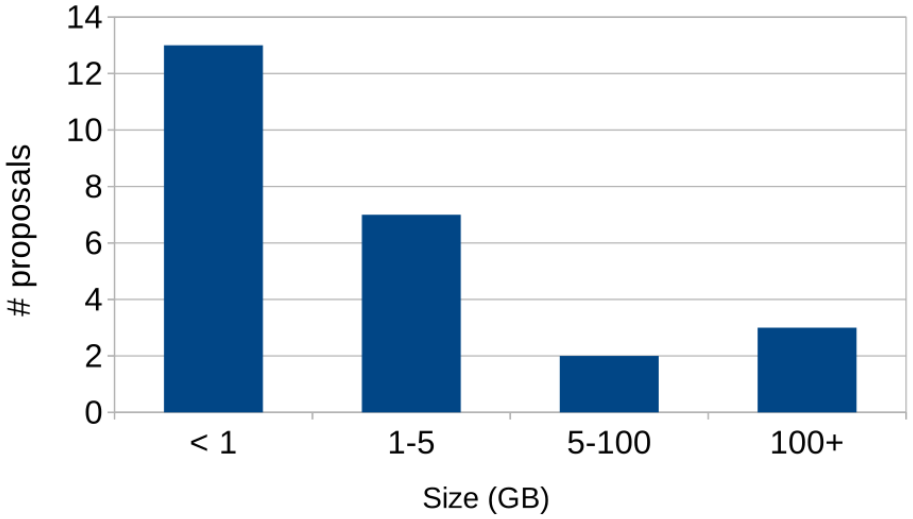
S. García & al, Big data preprocessing : methods and prospects, 2016



FS	Feature Selection	ID	Incomplete data
IR	Instance Reduction	D	discretization
IC	Imbalanced class		

Frameworks and data size





S.

García & al, Big data preprocessing : methods and prospects, 2016