



# **DATA MINING E DESCOBERTA DE CONHECIMENTO**

Hugo Peixoto

2019 – 2020 Universidade do Minho



# Data Mining





# DATA MINING vs KNOWLEDGE DISCOVERY

KDD refers to the overall process of discovering useful knowledge from data

Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process.



# DATA MINING vs KNOWLEDGE DISCOVERY

- Many Definitions

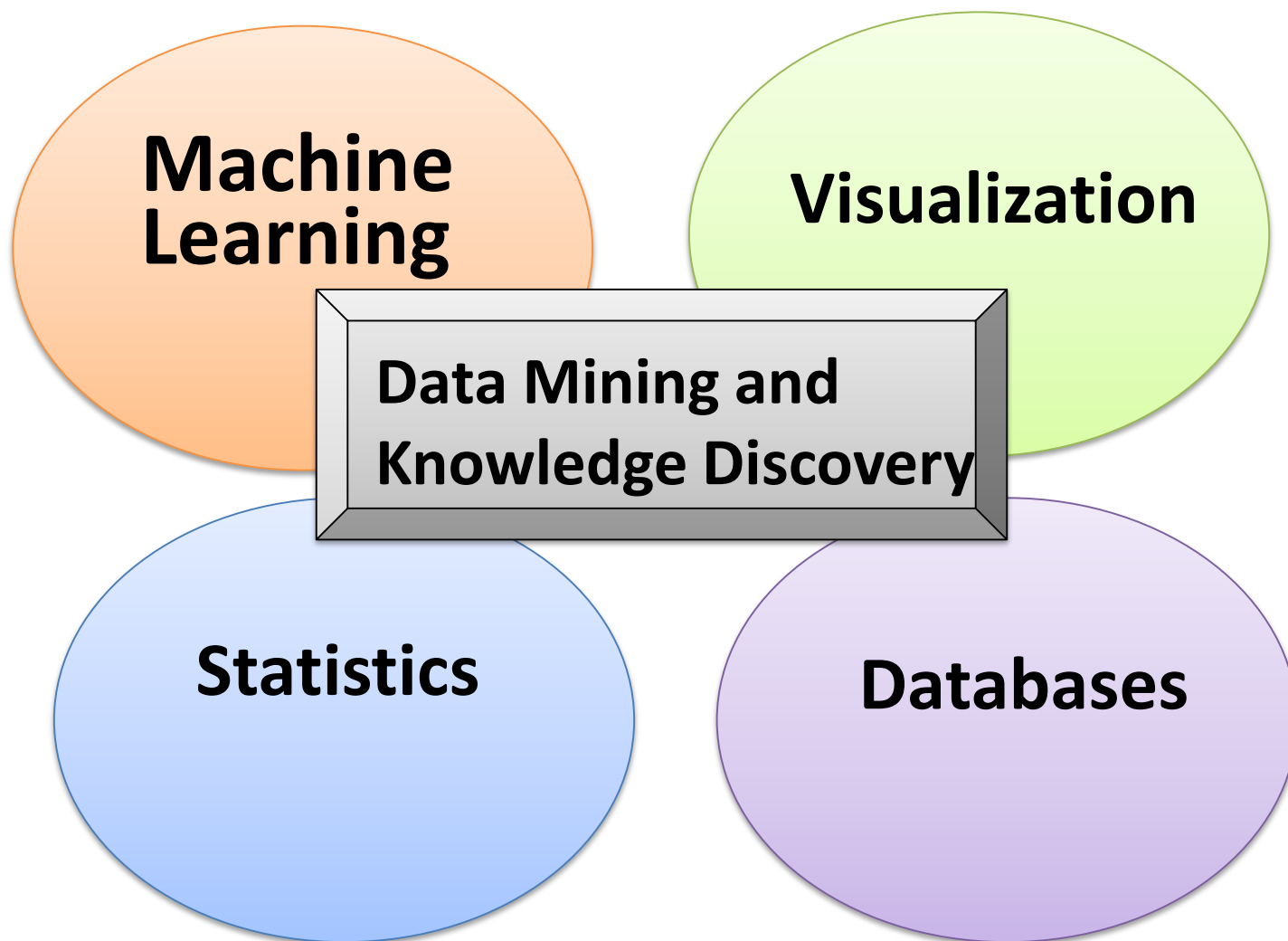
- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data;

from Advances in Knowledge Discovery and Data Mining, Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, (Chapter 1), AAAI/MIT Press 1996

- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns



# Related Fields





# Related Fields

- **Statistics:**
  - more theory-based
  - more focused on testing hypotheses
- **Machine Learning**
  - more heuristics than theory-based
  - focused on improving performance of a learning algorithms
- **Data Mining and Knowledge Discovery**
  - Data Mining one step in the Knowledge Discovery process (applying the Machine Learning algorithm)
  - Knowledge Discovery, the whole process including data cleaning, learning, and integration and visualization of results
- **Distinctions are fuzzy**



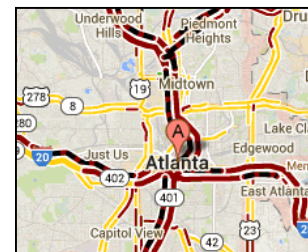
# DATA IS EVERYWHERE



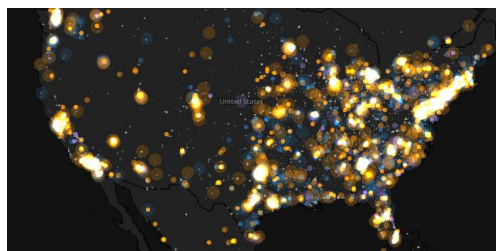
**Cyber Security**



**E-Commerce**



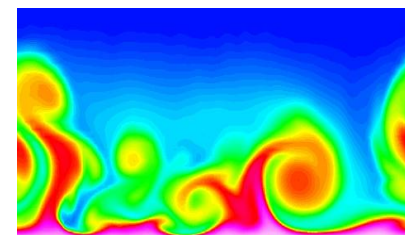
**Traffic Patterns**



**Social Networking: Twitter**



**Sensor Networks**



**Computational Simulations**



# DATA IS EVERYWHERE

There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies

## **New mantra**

Gather whatever data you can whenever and wherever possible.

## **Expectations**

Gathered data will have value either for the purpose collected or for a purpose not envisioned.





# WHERE TO FIND?



# COMERCIAL VIEWPOINT

Google



PayPal

AliExpress™

facebook

Microsoft

Apple

NETFLIX



# COMERCIAL VIEWPOINT

- Lots of data is being collected and warehoused
  - Web data
    - Google/Twitter has Peta Bytes of web data
    - Facebook has billions of active users
  - purchases at department/grocery stores, e-commerce
    - Amazon handles millions of visits/day
  - Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)



# COMERCIAL VIEWPOINT





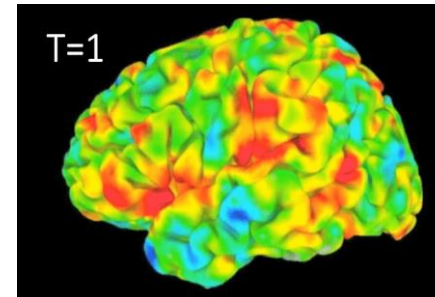
# COMERCIAL VIEWPOINT

- British political consulting firm which combined misappropriation of digital assets, data mining, data brokerage, and data analysis with strategic communication during the electoral processes (wikipedia).
- [2018] The **New York Times** and **The Observer** reported that the company had acquired and used personal data about **Facebook** users from an external researcher who had told Facebook he was collecting it for academic purposes

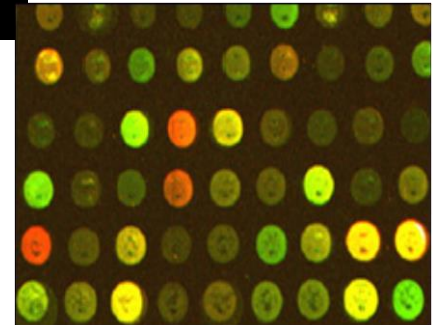


# SCIENTIFIC VIEWPOINT

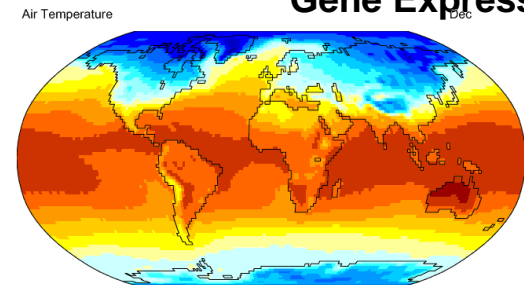
- Data collected and stored at enormous speeds
  - remote sensors on a satellite
    - NASA EOSDIS archives over petabytes of earth science data / year
  - telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - scientific simulations
    - terabytes of data generated in a few hours
- Data mining helps scientists
  - in automated analysis of massive datasets
  - In hypothesis formation



**fMRI Data from Brain**



**Gene Expression Data**



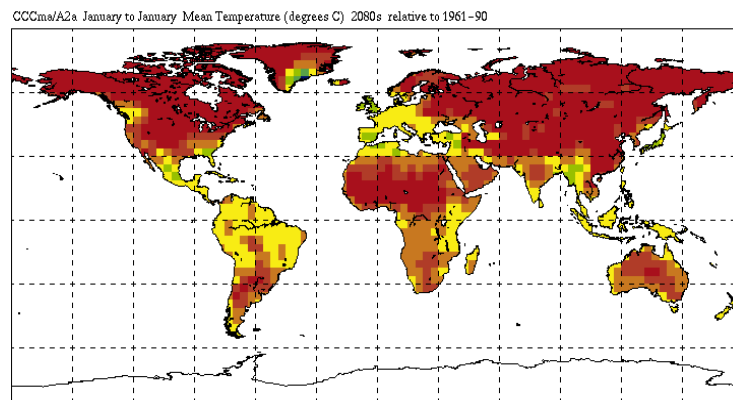
**Surface Temperature of Earth**



# OPPORTUNITIES



**Improving health care and reducing costs**



**Predicting the impact of climate change**



**Finding alternative/ green energy sources**



**Reducing hunger and poverty by increasing agriculture production**



# Application examples

- **Customer Relationship Management (CRM)**
  - Based on a data base with client information and behavior try to select other potential consumers of a product.
- **Profiling tax cheaters**
  - Based on the profile of the tax payer and some figures from the tax (electronic) form try to product tax cheating.





# Application examples

- **Health care**
  - Given the patient profile and the diagnoses try to predict the number of hospital days. Information is used in planning system.
- **Industry**
  - Job shop planning. Based on already accepted jobs, try to product the delivery time of a new offered job.



# Application examples

- **Data analysis and decision support**
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)
- **Other Applications**
  - Text mining (news group, email, documents) and Web mining
  - Medical data mining
  - Bioinformatics and bio-data analysis



# DATA MINING TECHNIQUES

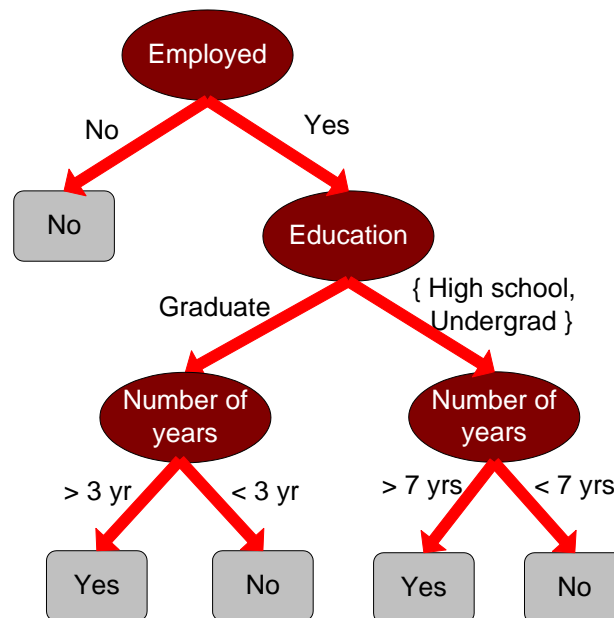
## Examples:

1. **Classification**
2. **Regression**
3. **Association Rules**
4. **Clustering**



# DATA MINING TECHNIQUES

**1. Classification.** Is a complex technique that forces to collect various attributes together into discernable categories, which you can then use to draw further conclusions, or serve some function.





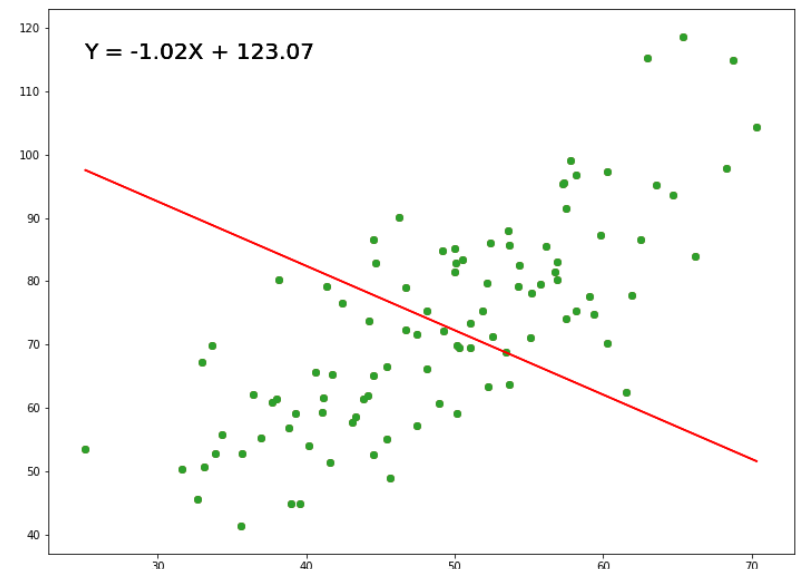
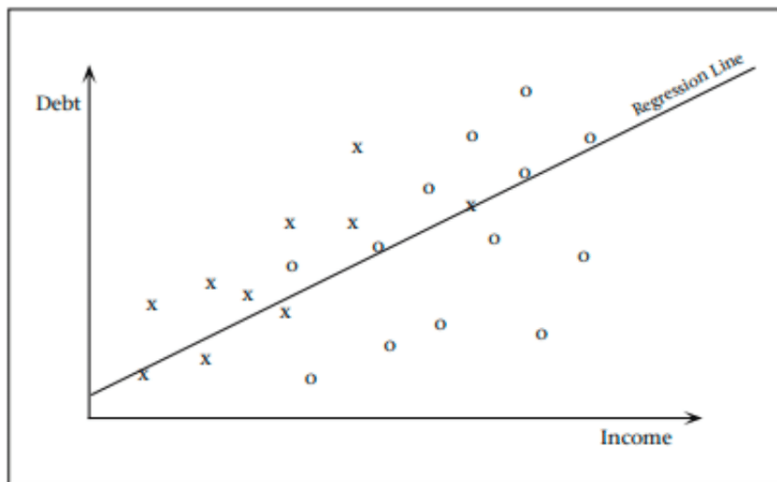
# CLASSIFICATION EXAMPLES

- Classifying credit card transactions as legitimate or fraudulent
- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data
- Categorizing news stories as finance, weather, entertainment, sports, etc
- Identifying intruders in the cyberspace
- Predicting tumor cells as benign or malignant
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil



# DATA MINING TECHNIQUES

**2. Regression.** Regression, used primarily as a form of planning and modeling, is used to identify the likelihood of a certain variable, given the presence of other variables.





# REGRESSION EXAMPLES

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Extensively studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.



# DATA MINING TECHNIQUES

**3. Association Rules.** Look for specific events or attributes that are highly correlated with another event or attribute.

Example: when your customers buy a specific item, they also often buy a second, related item. This is usually what's used to populate “people also bought” sections of online stores.







# ASSOCIATION EXAMPLES

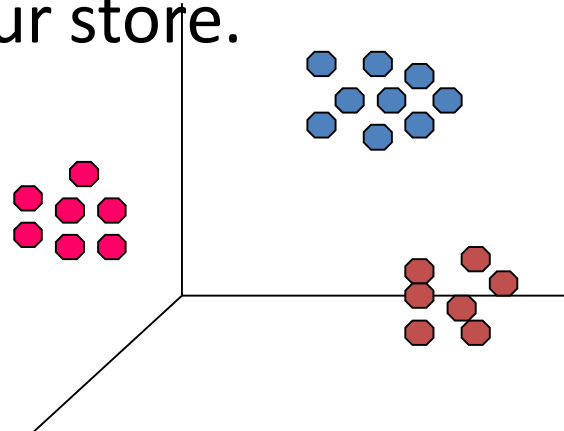
- Market-basket analysis
  - Rules are used for sales promotion, shelf management, and inventory management
- Telecommunication alarm diagnosis
  - Rules are used to find combination of alarms that occur together frequently in the same time period
- Medical Informatics
  - Rules are used to find combination of patient symptoms and test results associated with certain diseases



# DATA MINING TECHNIQUES

**4. Clustering.** Clustering is very similar to classification but involves grouping chunks of data together based on their similarities.

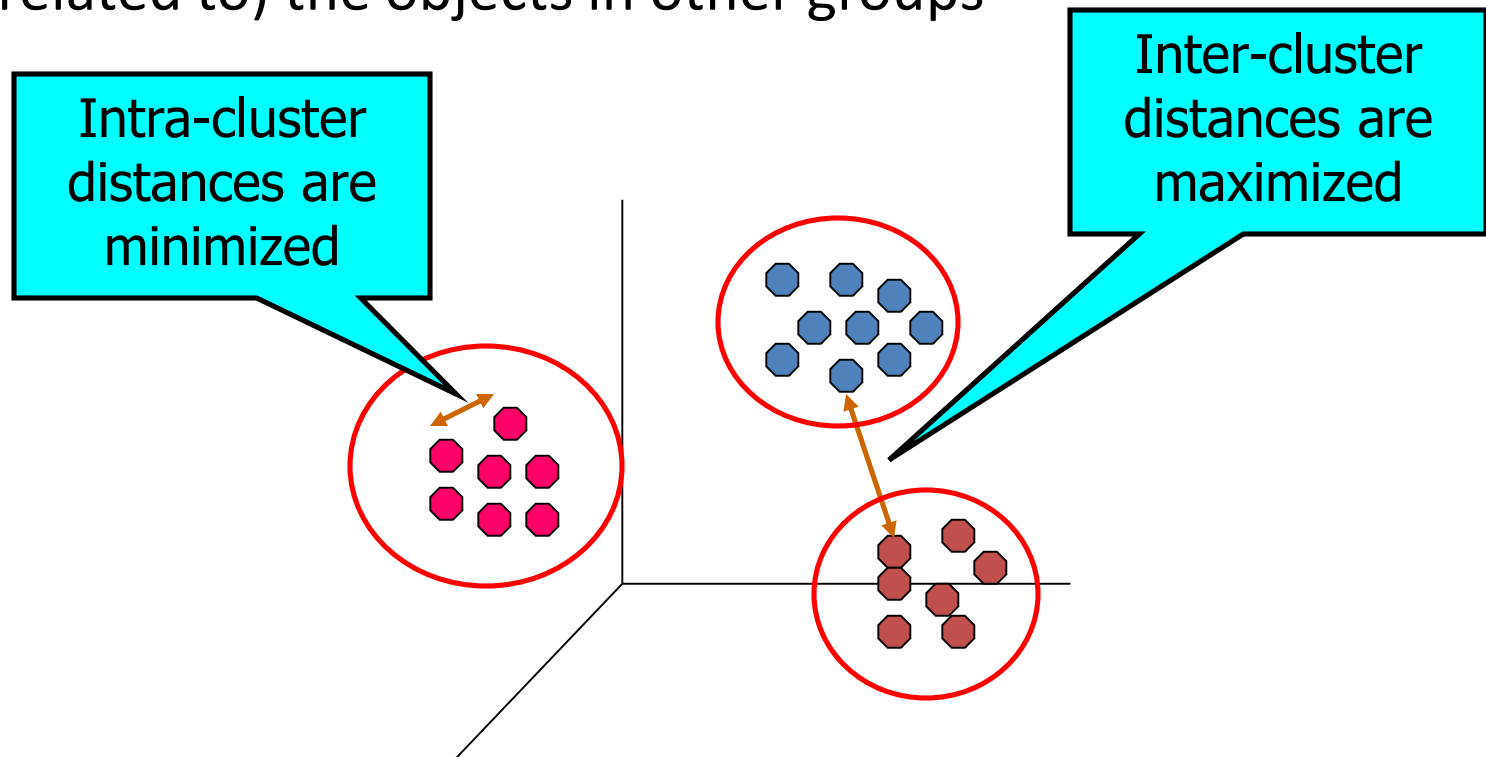
Example: cluster different demographics of your audience into different packets based on how much disposable income they have, or how often they tend to shop at your store.

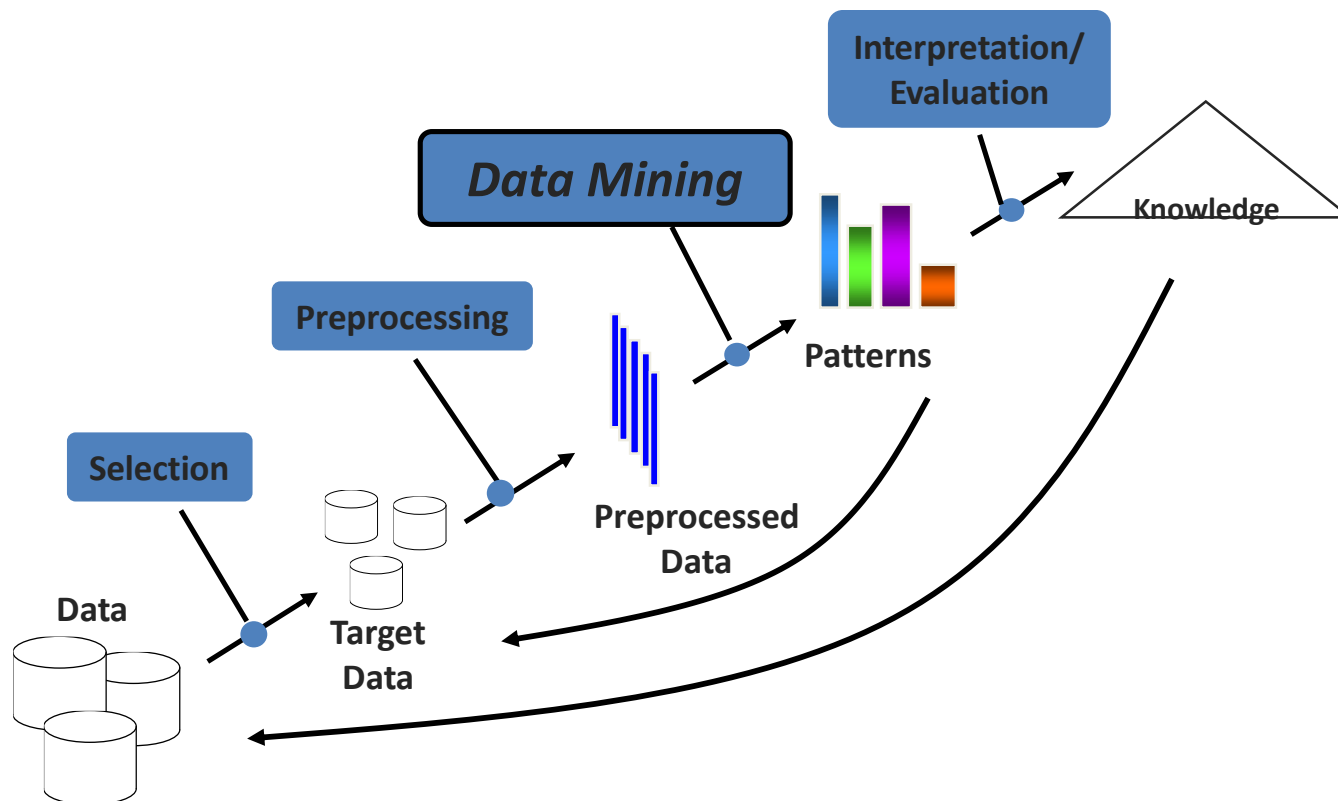




# CLUSTERING EXAMPLES

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups







# CRISP-DM

**C**ross Industry **S**tandard **P**rocess for **D**ata **M**ining



# CRISP-DM

European Community funded effort to develop framework for data mining tasks

## Goals:

- Encourage interoperable tools across entire data mining process

- Take the mystery/high-priced expertise out of simple data mining tasks



# CRISP-DM

The data mining process must be reliable and repeatable by people with little data mining background !!



# CRISP-DM

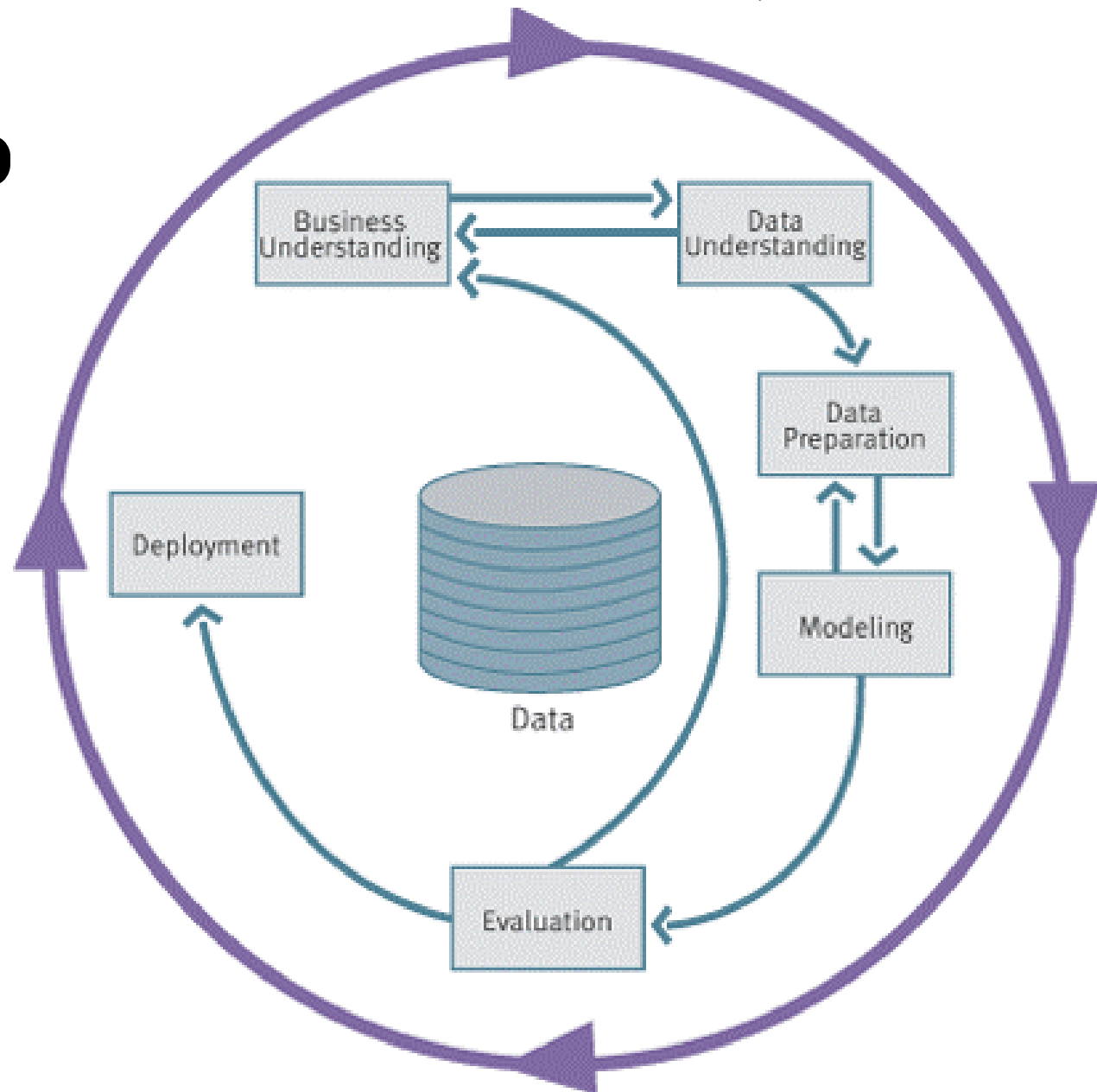
## FEATURES:

- Framework for recording experience
- Allows projects to be replicated
- Aid to project planning and management
- “Comfort factor” for new adopters
- Demonstrates maturity of Data Mining
- Reduces dependency on “stars”





# CRISP-DM





# CRISP-DM

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
<b>Determine Business Objectives</b> <i>Background Business Objectives Business Success Criteria</i>	<b>Collect Initial Data</b> <i>Initial Data Collection Report</i>	<i>Data Set Data Set Description</i>	<b>Select Modeling Technique</b> <i>Modeling Technique Modeling Assumptions</i>	<b>Evaluate Results</b> <i>Assessment of Data Mining Results w.r.t. Business Success Criteria Approved Models</i>	<b>Plan Deployment</b> <i>Deployment Plan</i>
<b>Situation Assessment</b> <i>Inventory of Resources Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits</i>	<b>Describe Data</b> <i>Data Description Report</i>	<b>Select Data</b> <i>Rationale for Inclusion/ Exclusion</i>	<b>Generate Test Design</b> <i>Test Design</i>	<b>Review Process</b> <i>Review of Process</i>	<b>Plan Monitoring and Maintenance</b> <i>Monitoring and Maintenance Plan</i>
<b>Determine Data Mining Goal</b> <i>Data Mining Success Criteria</i>	<b>Explore Data</b> <i>Data Exploration Report</i>	<b>Clean Data</b> <i>Data Cleaning Report</i>	<b>Build Model</b> <i>Parameter Settings Models Model Description</i>	<b>Determine Next Steps</b> <i>List of Possible Actions Decision</i>	<b>Produce Final Report</b> <i>Final Report Final Presentation</i>
<b>Produce Project Plan</b> <i>Project Plan Initial Assessment of Tools and Techniques</i>	<b>Verify Data Quality</b> <i>Data Quality Report</i>	<b>Construct Data</b> <i>Derived Attributes Generated Records</i>	<b>Assess Model</b> <i>Model Assessment Revised Parameter Settings</i>	<b>Review Project</b> <i>Experience Documentation</i>	
		<b>Integrate Data</b> <i>Merged Data</i>			
		<b>Format Data</b> <i>Reformatted Data</i>			



# CRISP-DM

- **Business Understanding**
  - Understanding project objectives and requirements
  - Data mining problem definition
- **Data Understanding**
  - Initial data collection and familiarization
  - Identify data quality issues
  - Initial, obvious results
- **Data Preparation**
  - Record and attribute selection
  - Data cleansing
- **Modeling**
  - Run the data mining tools
- **Evaluation**
  - Determine if results meet business objectives
  - Identify business issues that should have been addressed earlier
- **Deployment**
  - Put the resulting models into practice
  - Set up for repeated/continuous mining of the data



## CRISP-DM PHASE 1 - BUSINESS UNDERSTANDING

- Statement of Business Objective  
States goal in business terminology
- Statement of Data Mining objective  
States objectives in technical terms
- Statement of Success Criteria

Focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives

What the client really wants to accomplish?

Uncover important factors (constraints, competing objectives)



# CRISP-DM PHASE 1 - BUSINESS UNDERSTANDING

## Determine business objectives

- Key persons and their roles? Is there a steering committee. Internal sponsor (financial, domain expert).
- Business units impacted by the project (sales, finance,...)? Business success criteria and who assesses it?
- Users' needs and expectations.
- Describe problem in general terms. Business questions, Expected benefits.

## Assess situation

- Are they already using data mining.
- Identify hardware and software available. Identify data sources and their types (online, experts, written documentation).
- Identify knowledge sources and types (online, experts, written documentation)
- Describe the relevant background.



# CRISP-DM PHASE 1 - BUSINESS UNDERSTANDING

## Determine data mining goals

- Translate the business questions to data mining goals  
*(e.g., a marketing campaign requires segmentation of customers in order to decide whom to approach in this campaign; the level/size of the segments should be specified).*
- Specify data mining problem type  
(e.g., classification, description, prediction and clustering).
- Specify criteria for model assessment.

## Produce project plan

- Define initial process plan; discuss its feasibility with involved personnel.
- Put identified goals and selected techniques into a coherent procedure.
- Estimate effort and resources needed; Identify critical steps.



## CRISP-DM PHASE 2 – DATA UNDERSTANDING

- Acquire the data
- Explore the data (query & visualization)
- Verify the quality

Starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.



# CRISP-DM PHASE 2 – DATA UNDERSTANDING

## Collect data

- List the datasets acquired (locations, methods used to acquire, problems encountered and solutions achieved).

## Describe data

- Check data volume and examine its gross properties.
- Accessibility and availability of attributes. Attribute types, range, correlations, the identities.
- Understand the meaning of each attribute and attribute value in business terms.
- For each attribute, compute basic statistics (e.g., distribution, average, max, min, standard deviation, variance, mode, skewness).





# CRISP-DM PHASE 2 – DATA UNDERSTANDING

## Explore data

Analyze properties of interesting attributes in detail

Distribution, relations between pairs or small numbers of attributes, properties of significant sub-populations, simple statistical analyses

## Verify data quality

Identify special values and catalogue their meaning.

Does it cover all the cases required? Does it contain errors and how common are they?

Identify missing attributes and blank fields. Meaning of missing data.

Do the meanings of attributes and contained values fit together?

Check spelling of values (e.g., same value but sometime beginning with a lower case letter, sometimes with an upper case letter).

Check for plausibility of values, e.g. all fields have the same or nearly the same values.



# CRISP-DM PHASE 3 – DATA PREPARATION

## Construct data

Derived attributes.

Background knowledge .

How can missing attributes be constructed or imputed?

## Integrate data

Integrate sources and store result (new tables and records).

## Format Data

**Rearranging attributes** (Some tools have requirements on the order of the attributes, e.g. first field being a unique identifier for each record or last field being the outcome field the model is to predict).

**Reordering records** (Perhaps the modeling tool requires that the records be sorted according to the value of the outcome attribute).

**Reformatted within-value** (These are purely syntactic changes made to satisfy the requirements of the specific modeling tool, remove illegal characters, uppercase lowercase).



## CRISP-DM PHASE 4 – MODELING

- Select the modeling technique  
Based upon the data mining objective
- Generate test design  
Procedure to test model quality and validity
- Build model  
Parameter settings
- Assess model (rank the models)

Various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often necessary



# CRISP-DM PHASE 4 – MODELING

## Select modeling technique

- Select technique
- Identify any built-in assumptions made by the technique about the data (e.g. quality, format, distribution).
- Compare these assumptions with those in the Data Description Report and make sure that these assumptions hold.
- Preparation Phase if necessary.

## Generate test design

- Describe the intended plan for train, test and evaluate the models.
- How to divide the dataset into training, test and validation sets.
- Decide on necessary steps (number of iterations, number of folds etc.).
- Prepare data required for test



# CRISP-DM PHASE 4 – MODELING

## Build model

- Set initial parameters and document reasons for choosing those values.
- Run the selected technique on the input dataset. Post-process data mining results (eg. editing rules, display trees).
- Record parameter settings used to produce the model.
- Describe the model, its special features, behavior and interpretation.

## Assess model

- Evaluate result with respect to evaluation criteria. Rank results with respect to success and evaluation criteria and select best models.
- Interpret results in business terms. Get comments by domain experts.
- Check plausibility of model.
- Check model against given knowledge base (discovered info. novel and useful?)
- Check result reliability. Analyze potentials for deployment of each result.



## CRISP-DM PHASE 5 – EVALUATION

- More thoroughly evaluate model
- Decide how to use results
- Methods and criteria depend on model type:  
e.g., coincidence matrix with classification models, mean error rate with regression models

Interpretation of model: important or not, easy or hard depends on algorithm

Determine if there is some important business issue that has not been sufficiently considered.

A decision on the use of the data mining results should be reached



# CRISP-DM PHASE 5 – EVALUATION

## Evaluate results

- Understand data mining result. Check impact for data mining goal.
- Check result against knowledge base to see if it is novel and useful.
- Evaluate and assess result with respect to business success criteria
- Rank results according to business success criteria. Check result impact on initial application goal.
- Are there new business objectives? (address later in project or new project?)
- State conclusions for future data mining projects.

## Review of process

- Summarize the process review (activities that missed or should be repeated).
- Overview data mining process. Is there any overlooked factor or task?
- (did we correctly build the model? Did we only use attributes that we are allowed to use and that are available for future analyses?)
- Identify failures, misleading steps, possible alternative actions, unexpected paths
- Review data mining results with respect to business success



# CRISP-DM PHASE 5 – EVALUATION

## Determine next steps

- Analyze potential for deployment of each result. Estimate potential for improvement of current process.
- Check remaining resources to determine if they allow additional process iterations (or whether additional resources can be made available).
- Recommend alternative continuations. Refine process plan.

## Decision

- According to the results and process review, it is decided how to proceed to the next stage (remaining resources and budget)
- Rank the possible actions. Select one of the possible actions.
- Document reasons for the choice.





# CRISP-DM

## WHY?

The data mining process must be reliable and repeatable by people with little data mining skills

CRISP-DM provides a uniform framework for

- guidelines
- experience documentation

CRISP-DM is flexible to account for differences

- Different business/agency problems
- Different data



# EXERCÍCIO GRUPO

## DETERMINAR NECESSIDADE DE COMPONENTES SANGUÍNEOS

No bloco operatório de uma unidade de saúde é de vital importância determinar com a devida antecedência a potencial necessidade de um paciente, que vai ser intervencionado, vir a receber uma ou mais transfusões de componentes sanguíneos (sangue, plasma, etc).

Esta necessidade advém de diversos fatores, tais como:

- Custo unitário de cada fornecimento de componentes sanguíneos;
- Escassez de oferta de componentes sanguíneos;
- Correto tratamento do paciente em caso de necessidade de transfusão;
- Diminuir os desperdícios com o deteriorar de componentes sanguíneos afetos a cirurgias que não são usados;
- entre outros...



# EXERCÍCIO GRUPO

DETERMINAR NECESSIDADE DE COMPONENTES SANGUÍNEOS

## Business Understanding

*Business Objectives*

*Requirements, Assumptions, and Constraints*

*Risks and Contingencies*

*Costs and Benefits*

Data Mining Goals

Data Mining Success Criteria

## Data Understanding

*Collect Initial Data*

*Describe Data*

*Explore Data*

*Verify Data Quality*



# DATA MINING TOOLS





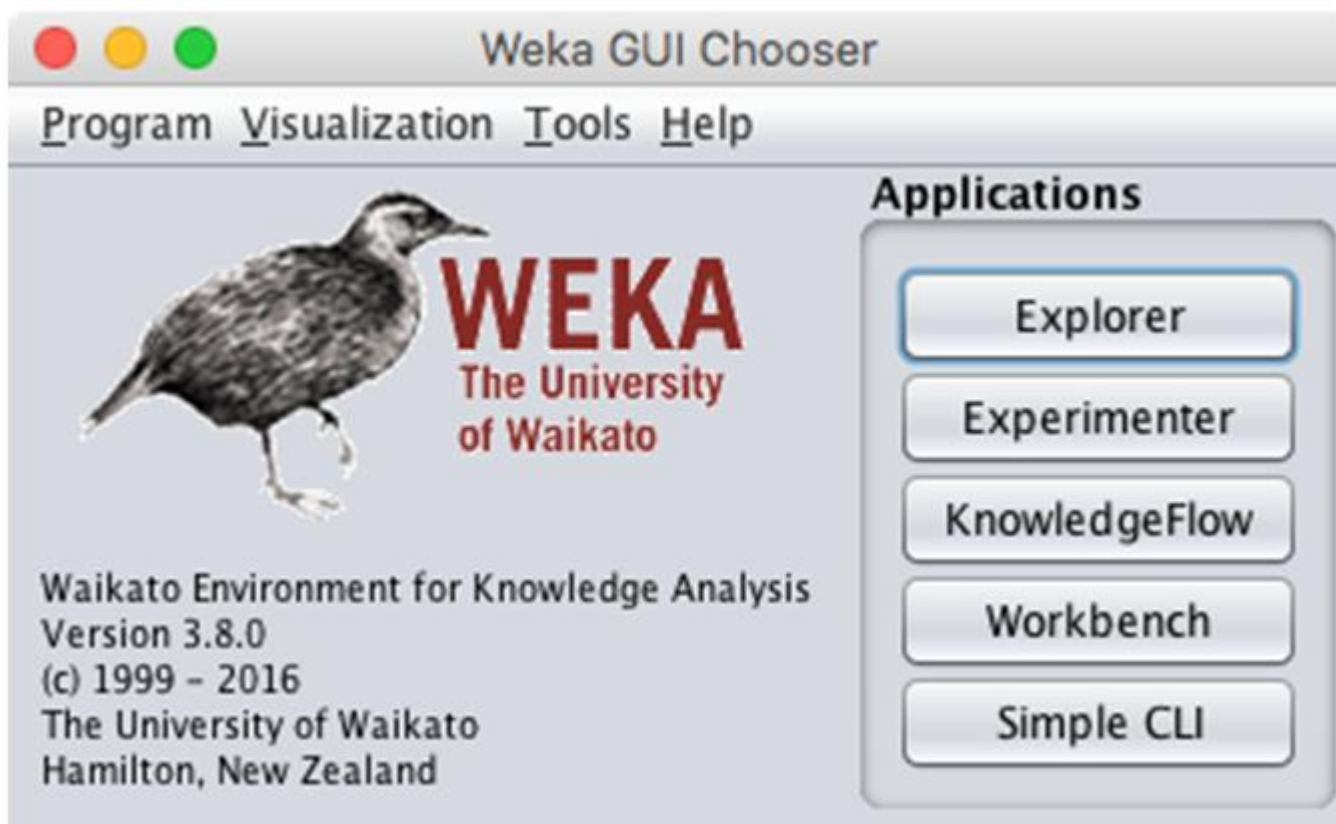
# WEKA



Weka is a collection of machine learning algorithms for data mining tasks.

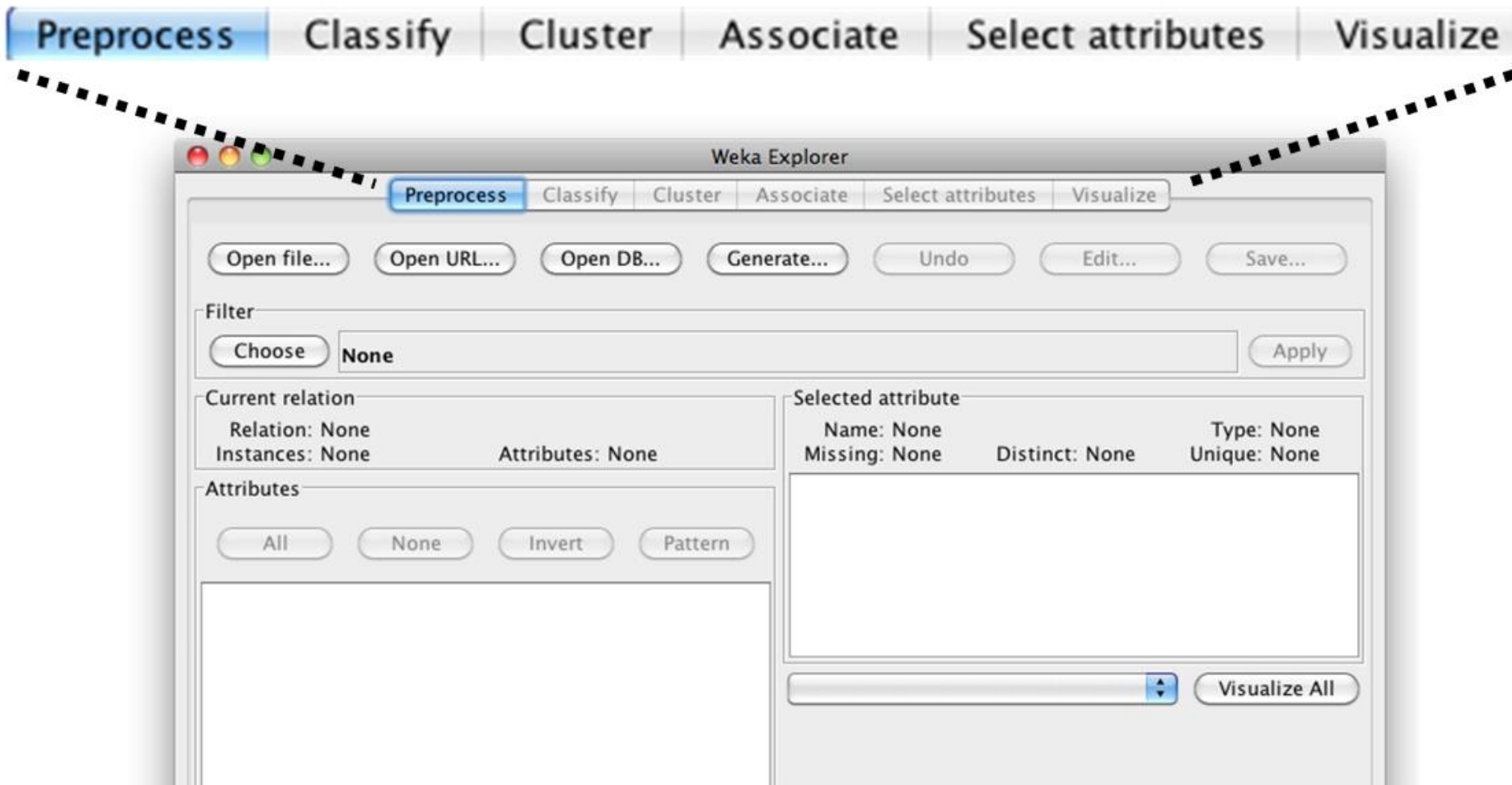


# WEKA



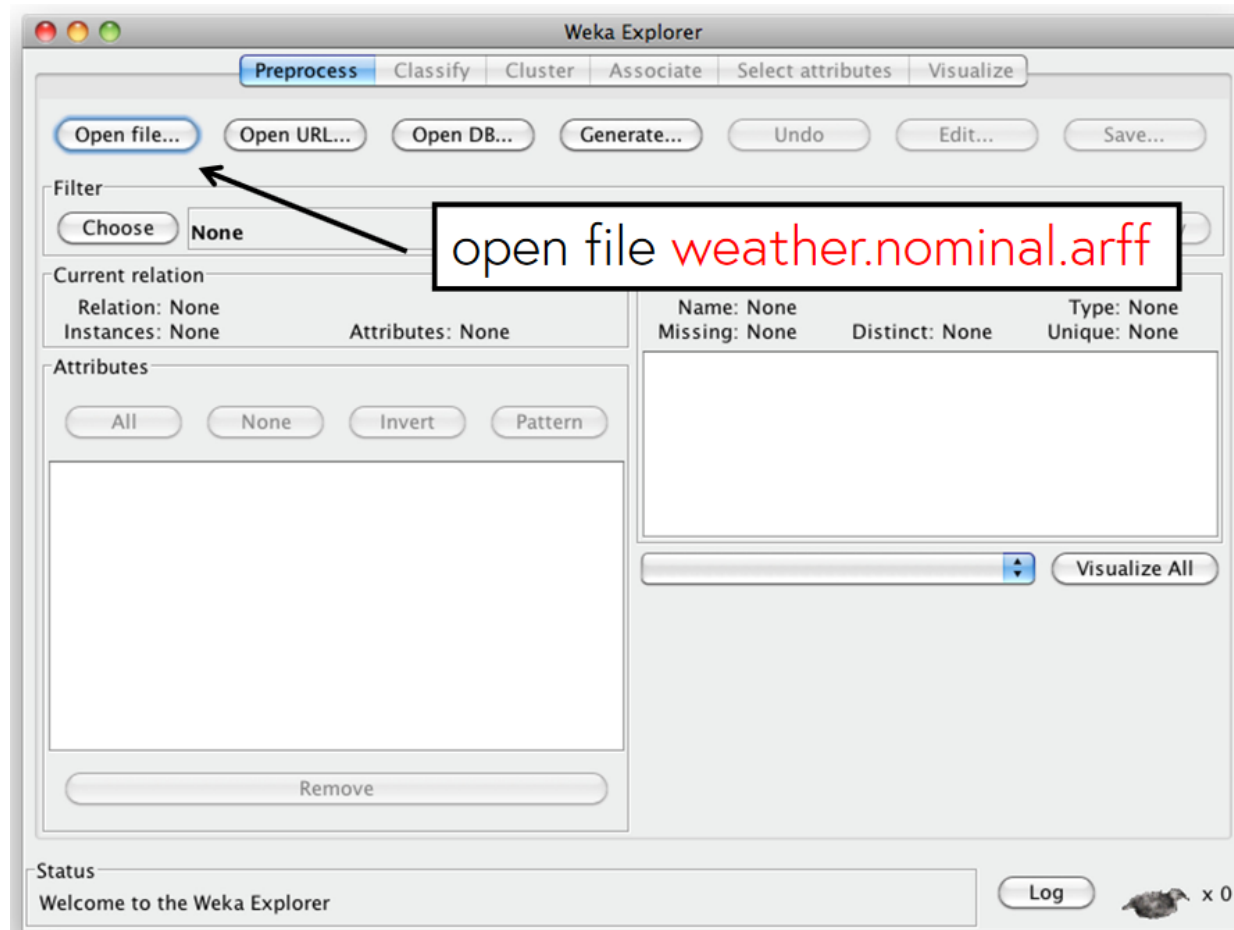


# WEKA





# WEKA







# WEKA

The screenshot shows the Weka Explorer application window. The 'Preprocess' tab is selected. The 'Attributes' list on the left contains: outlook, temperature, humidity, windy, and play. The 'Selected attribute' table on the right shows the distribution of the 'outlook' attribute. The 'Class' is set to 'play (Nom)'. The status bar at the bottom shows 'OK' and a 'Log' button.

**attributes**

**attribute values**

No.	Label	Count
1	sunny	5
2	overcast	4
3	rainy	5

Class: play (Nom)

Status: OK

Log x 0



# WEKA

Weather.arff

		attributes			
		Outlook	Temp	Humidity	Windy
instances	1	Sunny	Hot	High	False
	2	Sunny	Hot	High	True
	3	Overcast	Hot	High	False
	4	Rainy	Hot	High	False
	5	Rainy	Mild	Normal	False
	6	Rainy	Mild	Normal	True
	7	Rainy	Cool	Normal	True
	8	Rainy	Mild	Normal	False
	9	Sunny	Cool	Normal	False
	10	Rainy	Mild	Normal	False
	11	Sunny	Mild	Normal	True
	12	Overcast	Mild	High	True
	13	Overcast	Hot	Normal	False
	14	Rainy	Mild	High	True

Classification problem:  
predict the “class” value



# WEKA

Weather.arff

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose None

Current relation: Relation: weather.symbolic Instances: 14 Attributes: 5

Attributes: All None Invert Pattern

No.	Name
1	outlook
2	temperature
3	humidity
4	windy
5	play

Remove

Name: outlook Missing: 0 (0%) Distinct: 3 Type: Nominal Unique: 0 (0%)

No.	Label	Count
1	sunny	5
2	overcast	4
3	rainy	5

Class: play (Nom) Visualize All

Status: OK Log x 0

attributes

class

attribute values

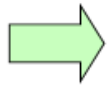


# WEKA

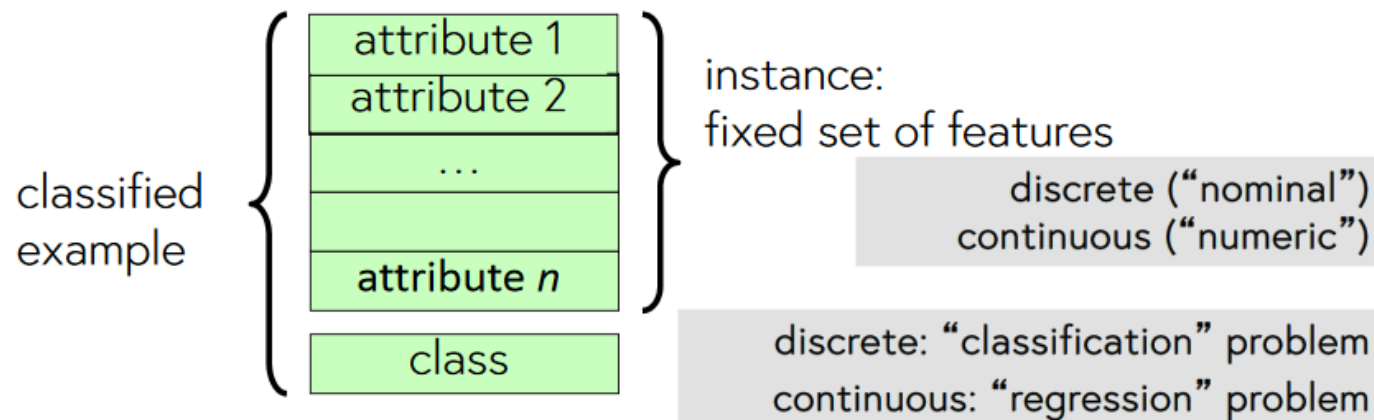
## Classification

sometimes called “supervised learning”

Dataset: classified examples



“Model” that classifies new examples





# WEKA

The screenshot shows the Weka Explorer window with the 'Preprocess' tab selected. The 'Open file...' button is highlighted with a callout box containing the text 'open file weather.numeric.arff'. The 'Attributes' list on the left is annotated with a box labeled 'attributes' pointing to the list and a box labeled 'class' pointing to the 'play' attribute. The 'Current relation' section shows 'Relation: weather.symbolic' and 'Instances: 14'. The 'Attributes' section shows a list of attributes: outlook, temperature, humidity, windy, and play. The 'Class' dropdown is set to 'play (Nom)'. The 'Visualize All' button is visible. The status bar at the bottom shows 'Status OK' and a 'Log' button.

open file weather.numeric.arff

attributes

class

attribute values

No.	Label	Count
1	sunny	5
2	overcast	4
3	rainy	5

Class: play (Nom)

Visualize All

Status OK

Log



# BUILD CLASSIFIER

## USE J48 TO ANALYZE THE GLASS DATASET

- Open file `glass.arff`
- Check the available classifiers
- Choose the J48 decision tree learner (trees>J48)
- Run it
- Examine the output
- Look at the correctly classified instances
  - ... and the confusion matrix



# BUILD CLASSIFIER

## INVESTIGATE J48

- Open the configuration panel
- Check the More information
- Examine the options
- Use an unpruned tree
- Look at leaf sizes
- Set minNumObj to 15 to avoid small leaves
- Visualize tree using right-click menu



# BUILD CLASSIFIER

## Pruning (decision trees)

is a technique in machine learning that reduces the size of **decision trees** by removing sections of the **tree** that provide little power to classify instances. **Pruning** reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting.







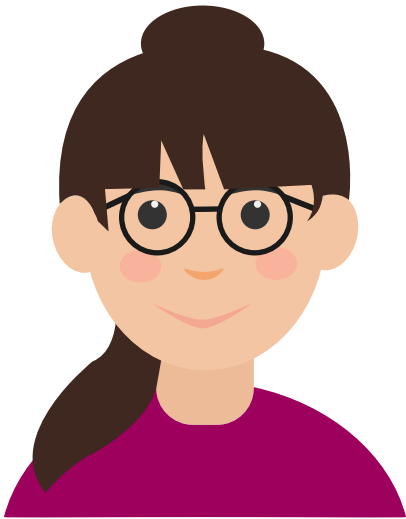
# CONTEXT AND PERSPECTIVE

Sarah is a regional sales manager for a nationwide supplier of fossil fuels for home heating.

Recent volatility in market prices for heating oil specifically, coupled with wide variability in the size of each order for home heating oil, has Sarah **concerned**.

Types of behaviors and other factors that may influence the demand for heating oil in the domestic market.

What factors are related to heating oil usage, and how might she use a knowledge of such factors to better manage her inventory, and anticipate demand.





# BUSINESS UNDERSTANDING

Sarah's goal is to **better understand how her company can succeed in the home heating oil market.**

She recognizes that there are many factors that influence heating oil consumption, and believes that by investigating the **relationship between a number of those factors**, she will be able to better monitor and respond to heating oil demand. She has selected correlation as a way to model the relationship between the factors she wishes to investigate.

**Correlation** is a statistical measure of how strong the relationships are between attributes in a data set.



# DATA UNDERSTANDING

**Insulation:** This is a density rating, ranging from one to ten, indicating the thickness of each home's insulation. A home with a density rating of one is poorly insulated, while a home with a density of ten has excellent insulation.

**Temperature:** This is the average outdoor ambient temperature at each home for the most recent year, measure in degree Fahrenheit.

**Heating\_Oil:** This is the total number of units of heating oil purchased by the owner of each home in the most recent year.

**Num\_Occupants:** This is the total number of occupants living in each home.

**Avg\_Age:** This is the average age of those occupants.

**Home\_Size:** This is a rating, on a scale of one to eight, of the home's overall size. The higher the number, the larger the home.

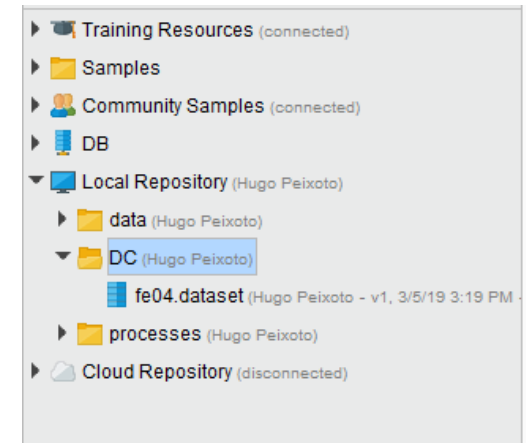


# DATA PREPARATION

Download csv: <http://hpeixoto.github.io/dc/pl05/pl05.dataset.csv>

Import csv to rapidminer repository.

Check results tab and inspect metadata view of the imported csv.

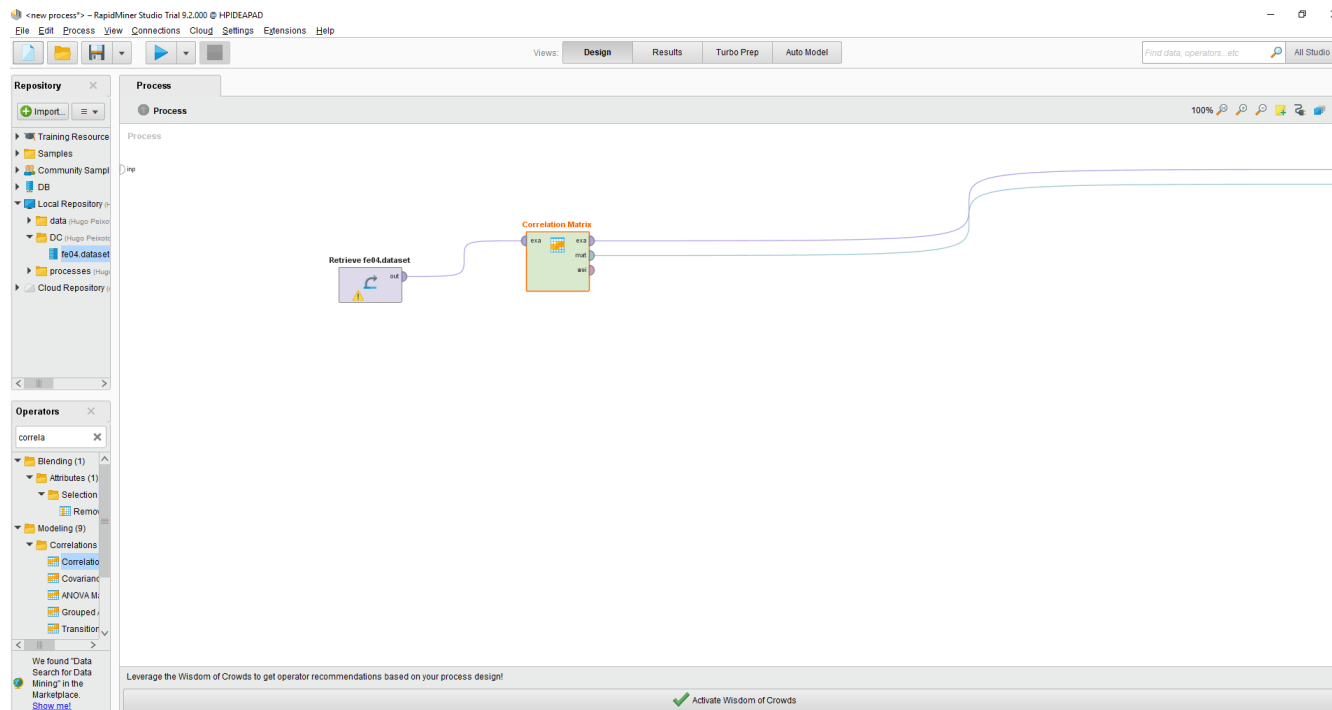




# MODELING

On the Operators tab in the lower left hand corner, use the search box and begin typing in the word *correlation*.

The tool we are looking for is called *Correlation Matrix*. Drag and drop and click *Run*.





# MODELING

## Correlation Matrix

Attribut...	Insulation	Temper...	Heating...	Num_O...	Avg_Age	Home_...
Insulation	1	-0.794	0.736	-0.013	0.643	0.201
Tempera...	-0.794	1	-0.774	0.013	-0.673	-0.214
Heating_...	0.736	-0.774	1	-0.042	0.848	0.381
Num_Oc...	-0.013	0.013	-0.042	1	-0.048	-0.023
Avg_Age	0.643	-0.673	0.848	-0.048	1	0.307
Home_S...	0.201	-0.214	0.381	-0.023	0.307	1



# EVALUATION

All correlation coefficients between 0 and 1 represent **positive correlations**, while all coefficients between 0 and -1 are **negative correlations**.

Keep in mind the direction of movement between the two attributes

consider the relationship between the **Heating\_Oil** consumption attribute, and the **Insulation** rating level attribute.

The coefficient there, as seen in our matrix, is 0.736. This is a positive number, and therefore, a positive correlation.





But what does that mean? **Correlations** that are **positive** mean that as one attribute's value rises, the other attribute's value also rises. But, a **positive correlation** also means that as one attribute's value falls, the other's also falls.







# EVALUATION

## Positive correlations

				
Heating Oil use rises	Insulation rating also rises		Heating Oil use falls	Insulation rating also falls

Whenever both attribute values move in the same direction, the correlation is positive.

## Negative correlations

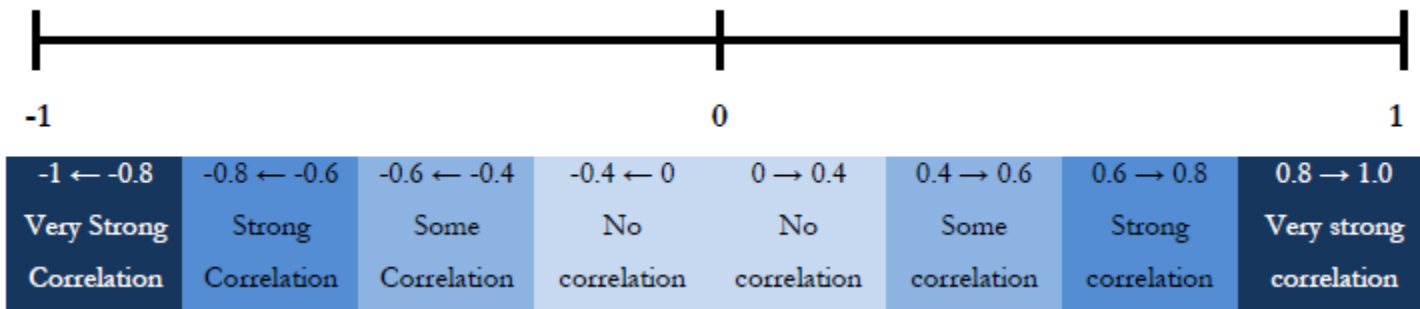
				
Temperature rises	Insulation rating falls		Temperature falls	Insulation rating rises

Whenever attribute values move in opposite directions, the correlation is negative.



# EVALUATION

## Correlations strengths





# DEPLOYMENT

The concept of deployment in data mining means doing something with what you've learned from your model; taking some action based upon what your model tells you:

We learned through our investigation, that the two most strongly correlated attributes in our data set are **Heating\_Oil** and **Avg\_Age**, with a coefficient of **0.848**

Thus, we know that in this data set, as the average age of the occupants in a home increases, so too does the heating oil usage in that home.



# DRAWBACKS

Consider the correlation coefficient between *Avg\_Age* and *Temperature*: -0.673 (strong negative correlation)

As the age of a home's residents increases, the average temperature outside decreases; and as the temperature rises, the age of the folks inside goes down.

Could the average age of a home's occupants have any effect on that home's average yearly outdoor temperature? **Certainly not.** If it did, we could control the temperature by simply moving people of different ages in and out of homes. This of course is silly.

While statistically, there is a correlation between these two attributes in our data set, there is no logical reason that movement in one *causes* movement in the other. The relationship is probably coincidental, but if not, there must be some other explanation that our model cannot offer.

Such limitations must be recognized and accepted in all data mining deployment decisions.



# DRAWBACKS

Another false interpretation about correlations is that the coefficients are percentages, as if to say that a correlation coefficient of 0.776 between two attributes is an indication that there is 77.6% shared variability between those two attributes. **This is not correct.**

While the coefficients do tell a story about the shared variability between attributes, the underlying mathematical formula used to calculate correlation coefficients solely measures strength, as indicated by proximity to 1 or -1, of the interaction between attributes. **No percentage is calculated or intended.**



# Biography

[Data Mining: Concepts and Techniques](#). Jiawei Han and Micheline Kamber. Morgan Kaufmann Publishers

[Data Mining Practical Machine Learning Tools and Techniques](#). Ian H. Witten and Eibe Frank. Morgan Kaufmann Publishers



# **DATA MINING E DESCOBERTA DE CONHECIMENTO**

Hugo Peixoto

2019 – 2020 Universidade do Minho