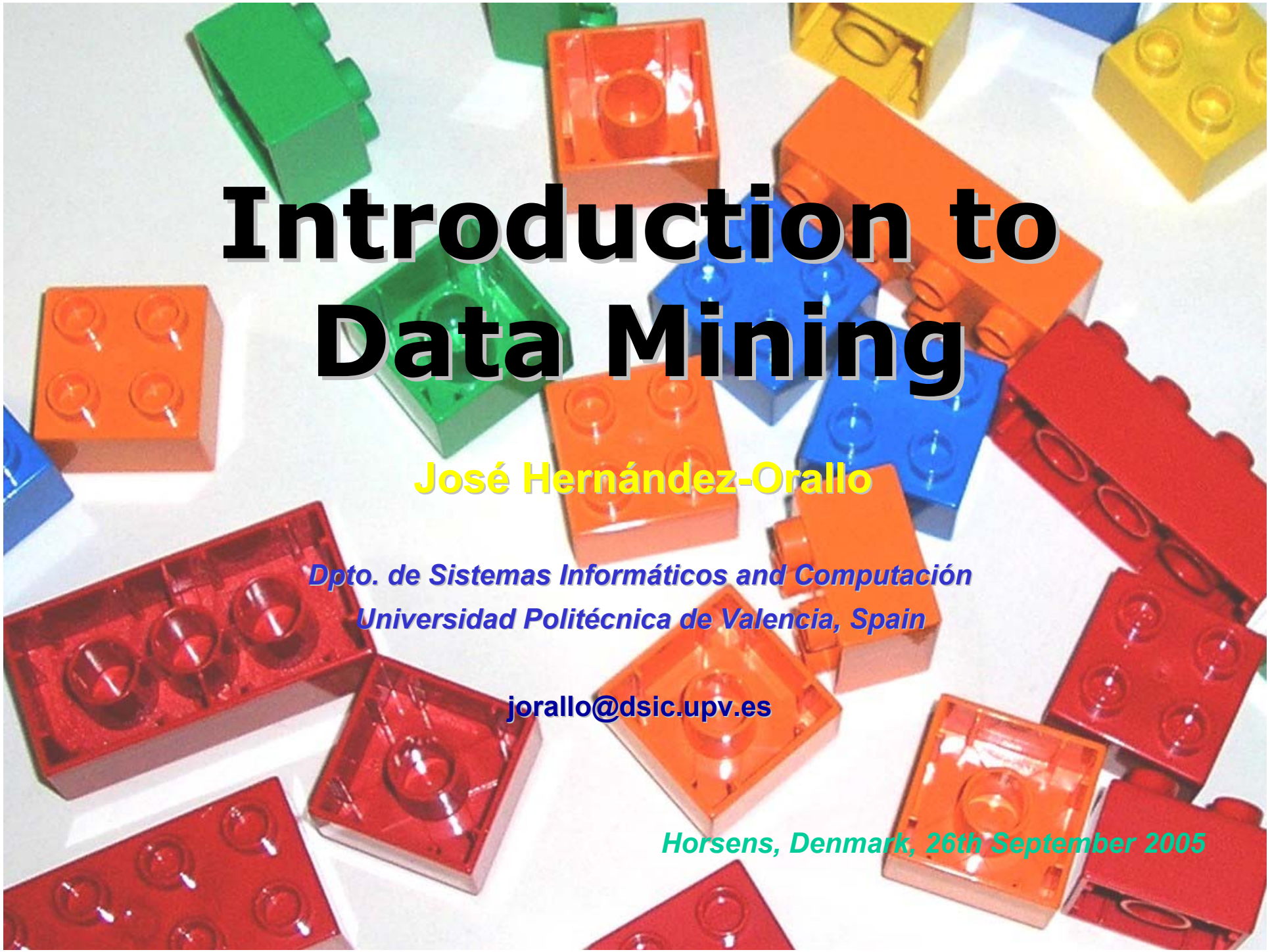


Introduction to Data Mining

José Hernández-Orallo

Dpto. de Sistemas Informáticos and Computación

Universidad Politécnica de Valencia, Spain

jorallo@dsic.upv.es

Horsens, Denmark, 26th September 2005

Outline

- Motivation. BI: Old Needs, New Tools.
- Some DM Examples.
- Data Mining: Definition and Applications
- The KDD Process
- Data Mining Techniques
- Development and Implementation

Relevance

- **The volume and variety** of the information which is stored in digital databases has grown spectacularly in the last decade.
- Most of this **information is historical**, i.e., represents transactions or situations which have happened.
- Apart from its function as "state of the organisation", and ultimately "memory of the organisation",

the historical information is also useful
to predict future information.

Relevance

- Most ***decisions*** in companies, organisations and institutions are based on information from past experience, which are extracted from very different sources.
- **Collective decisions** use to entail most critical consequences, especially economical and, recently, must be based on **data volumes which overflow human capacity**.

The area of the (semi-)automatic knowledge extraction from databases has recently acquired an unusual scientific and economical significance.

Relevance

- The final user is not an expert in data analysis tools (statistics, machine learning, …).
- The user cannot lose more time on analysing the data inefficiently:
  - industry: competitive advantages, more effective decisions.
  - science: data never analysed, data banks never related, etc.
  - personal: “information overload”...

The classical statistical packages are not easy to use and are not scalable to the size and type of data usual in databases.

Relation of DM and other disciplines

KDD arrives on the scene…

- *Knowledge Discovery from Databases*.

"*non-trivial process of identifying valid, novel, potentially useful and ultimately comprehensible patterns from data*".

(Fayyad et al. 1996)

- The discipline integrates techniques from many other different disciplines, without prejudices.

Relation of DM and other disciplines

KDD appears as an interface between and is fed from several disciplines:

- machine learning / AI.
- statistics.
- information systems / databases.
- data visualisation.
- parallel/distributed computation.
- natural language interfaces to databases.

Typical Application Areas

KDD for decision making (Dilly 96)

Retail/Marketing:	- Identify customer purchase patterns.
	- Find associations between customers and demographical features.
	- Predict the response to a *mailing campaign*.
	- Analyse shopping baskets.
Bank:	- Detect patterns of unlawful credit card use.
	- Identify loyal clients.
	- Predict customers with risk of churn.
	- Determine the credit card spending by several groups.
	- Find correlations between financial indicators.
	- Identify stock market rules from historical data.
Insurance / Private Health Care:	- Analyse medical procedures which are demanded together.
	- Predict which customers buy new insurance policies.
	- Identify behaviour patterns for customers with high risk.
	- Identify illegitimate behaviour.
Transportation:	- Determine the schedule for store delivering.
	- Analyse load patterns.

Typical Application Areas

KDD for decision making

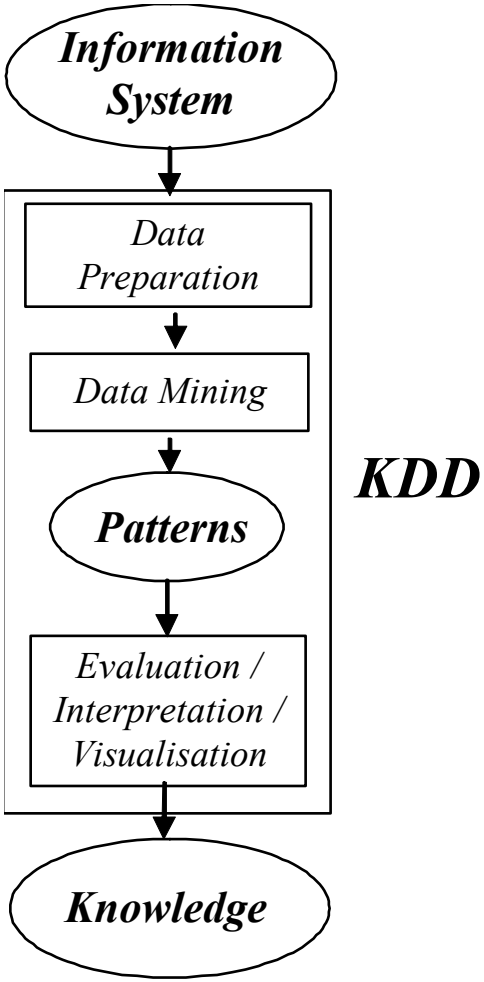
Medicine:

- Identify satisfactory medical therapies for several illnesses.
- Find symptom association and provide differential classifications for several pathologies.
- Study risk/health factors (genetic, precedents, habits, dietary, etc.) in different pathologies.
- Patient segmentation (clustering) for a more “intelligent” (specialised) attention according to each cluster/group.
- Temporal predictions in healthcare centres for a better use of resources, visits, wards and rooms.
- Epidemiological studies, throughput analysis of information campaigns, prevention, drug substitution, etc.

Outline

- Motivation. BI: Old Needs, New Tools.
- Some DM Examples.
- Data Mining: Definition and Applications
- The KDD Process
- Data Mining Techniques
- Development and Implementation

KDD Process: Stages.



1. Determine the sources of information which can be useful and where to find them.
2. Design a common data repository (data warehouse) which can unify in an operative way all the gathered information.
3. Implement the data warehouse which allows for data "navigation" and prior visualisation, to determine which issues deserve analysis.
4. Data selection, cleansing and transformation of the data which will be analysed. Construction of the *minable views*.
5. Choose and apply the most appropriate data mining method(s).
6. Interpretation, transformation and representation of the extracted patterns.
7. Spread and use (deployment) of the new knowledge.
8. Monitoring and revision (in case).

KDD stages: Data Integration

- The first KDD stages determine whether the subsequent stages are able to extract valid and useful knowledge from the original information.
- Generally, the information which requires analysis in the organisation may be found:
  - in databases and other **highly diverse sources**,
  - both internal and **external**.
  - many of these sources are those used for the **transactional work**.

The subsequent analysis will be much simpler if the source is **unified**, **accessible** (internal) and disconnected from the **transactional work**.

KDD stages: Data Integration

- Consequently, the subsequent data mining process **depends highly on the source**:
  - Database or plain files.
  - OLAP or OLTP.
  - Datawarehouse or just a copy of the original transactional database schema.
  - ROLAP or MOLAP.
  - The design of the “datamarts”: normalised, star, snowflake, …

KDD stages: Data selection, cleansing and transformation

Data cleansing and selection:

- *Cleansing*: we must eliminate as many erroneous or inconsistent data as we can.
- *Selection*: we must eliminate as many irrelevant data as we can.

  - Many **statistical methods**.
    - histograms (anomalous data detection).
    - data selection:
      - sampling (eliminating rows).
      - feature selection (eliminating columns).
    - attribute redefinition (grouping or split).

KDD stages: Data Selection, Cleansing and Transformation

Field Transformation:

- Numerisation
  - Advantages:
    - Many data mining methods (especially from statistics) only deal with numerical data. Gender (M/F) → 0/1.
    - In some cases the original nominal attributes represent an order (*shared flat, flat, apartment, semi-detached house, house, villa, castle*).
  - Disadvantages:
    - we need knowledge to know which data is initially non-numerical (the quantity or the order is irrelevant)
    - in some cases, the model is biased.

KDD stages: Data Selection, Cleansing and Transformation

Field Transformation:

- Discretisation (*binning*):
  - Advantages:
    - Space is usually reduced. Example: 0..10 ⇒ (small, medium, large).
    - (in some cases). More comprehensible rules (age > 18.53) ⇒ (age = ‘adult’).
  - Disadvantages:
    - In some cases, the model is biased.

KDD stages: Data Selection, Cleansing and Transformation

Transformation: Pick & Mix:

- Some data suggest the construction of new fields (columns) through *pick & mix*.
- Examples:
  - height^2/weight (obesity index)
  - debt/earnings
  - passengers * miles
  - credit limit - balance
  - population / area
  - minutes of use / number of telephone calls
  - activation_date - application_date
  - number of web pages visited / total amount purchased

KDD stages: data mining

Patterns to discover:

- Once the interesting data are collected, an "explorer" can decide which kinds of patterns s/he likes to discover.
- Depending on this search for knowledge (and for specific data mining problems), we can distinguish between:
  - *Directed data mining*: we know in advance which analysis problem we want to solve. This even allows automatic data mining to be included in software applications.
  - *Undirected data mining*: we want to extract patterns, but there is no specific task at hand. We work with the data (*until they confess!*). In this case, the use of OLAP tools is more necessary.
- In either case, the kind of knowledge which is needed determines the data mining *technique (algorithm)* to use.

Typology of Data Mining Patterns

Kinds of knowledge:

- **Associations**: An association between two attributes happens when the frequency of two specific values to happen together is relatively high.
  - Example: in a supermarket we analyse whether nappies and baby jars are bought together.
- **Dependencies**: A functional dependency (approximate or absolute) is a pattern in which it is established that one or more attributes determine the value of the other. But, alert! There are many void (non-interesting) dependencies (inverse causalities).
  - Example: if a patient has been allocated to the maternity ward implies that their gender is female.

Typology of Data Mining Patterns

Kinds of knowledge (contd.):

- **Classification**: A classification can be seen as the clarification of a dependency, in which each dependent attribute can take a value from several classes, known in advance (supervised learning).
  - Example: we know (perhaps from a dependency study or factor analysis) that the attributes *age*, *myopic degree* and *astigmatism level* determine which patient may take a refractory surgical operation satisfactorily.
    - We can try to determine the exact rules which classify a case as positive or negative from these attributes.

Typology of Data Mining Patterns

Kinds of knowledge (contd.):

- **Clustering / Segmentation**: clustering is the detection of groups of individuals. It's different from classification in that we do not know in advance the classes (or even its number) (unsupervised learning). The goal is to determine groups or clusters which are different from the other.
  - Example: find types (clusters) of telephone calls or credit card purchases.
    - These groups are useful to design different policies for each group or to analyse one group in more detail.
    - Also useful to summarise the data.

Typology of Data Mining Patterns

Kinds of knowledge (cont.):

- **Trends / regression**: the goal is to predict the values of a continuous variable from the evolution of another continuous variable (generally time) or from other continuous or nominal variables.
  - Example: we need to know the number of future customers or patients, the incomes, calls, earnings, costs, etc. from previous results (day, weeks, months or years before).

- **Other types of knowledge:**
  - **Schema information:** (to discover alternative primary keys, integrity constraints, ...).
  - **General rules**: patterns which cannot be classified into the previous kinds. Other more complex models/patterns (dynamic, …).

Example

We have the following table with employee data:

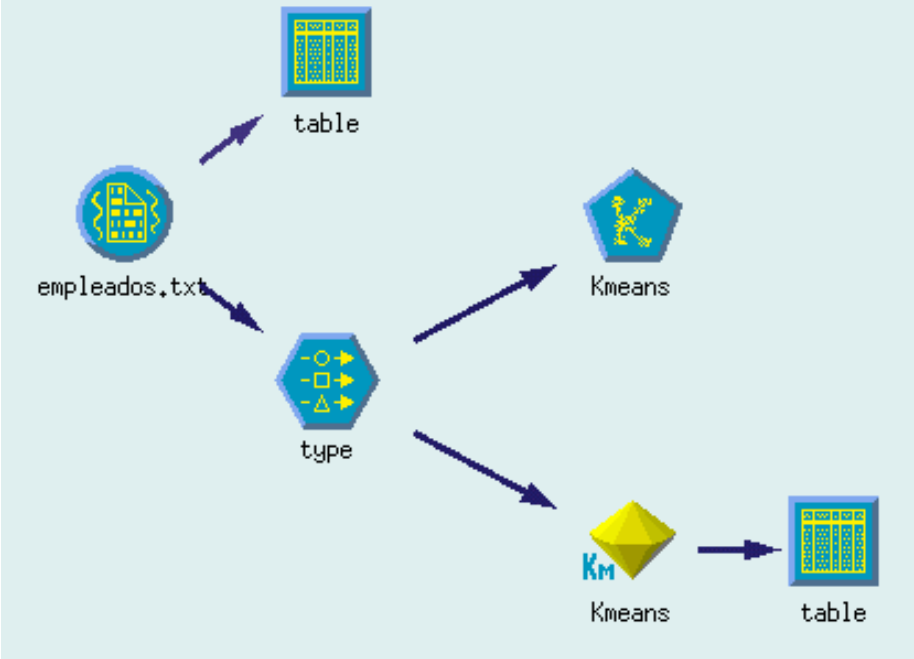
#	Salary	Marrd	Car	Chldrn	House	Union	offsick/Year	WorkYears	Gender
1	10000	Yes	No	0	Rent	No	7	15	M
2	20000	No	Yes	1	Rent	Yes	3	3	F
3	15000	Yes	Yes	2	Owner	Yes	5	10	M
4	30000	Yes	Yes	1	Rent	No	15	7	F
5	10000	Yes	Yes	0	Owner	Yes	1	6	M
6	40000	No	Yes	0	Rent	Yes	3	16	F
7	25000	No	No	0	Rent	Yes	0	8	M
8	20000	No	Yes	0	Owner	Yes	2	6	F
9	20000	Yes	Yes	3	Owner	No	7	5	M
10	30000	Yes	Yes	2	Owner	No	1	20	M
11	50000	No	No	0	Rent	No	2	12	F
12	8000	Yes	Yes	2	Owner	No	3	1	M
13	20000	No	No	0	Rent	No	27	5	F
14	10000	No	Yes	0	Rent	Yes	0	7	M
15	8000	No	Yes	0	Rent	No	3	2	M

We want to obtain representative subgroups.

Example

- We import the data into a data mining package, we give types (nominal or numerical) to the date, we check for anomalous data, etc.
- We apply the *k-means* algorithm to find clusters. We indicate the algorithm to find the three most significant groups.

The data mining process is as follows:



Example

After executing the algorithm we get a model, which shows three groups

cluster 1	cluster 2	cluster 3
5 examples	4 examples	6 examples
Salary : 226000 Married : No -> 0.8 Yes -> 0.2 Car : No -> 0.8 Yes -> 0.2 Children : 0 House : Rent -> 1.0 Union : No -> 0.8 Yes -> 0.2 Offsick/Year : 8 WorkYear : 8 Gender : M -> 0.6	Salary : 225000 Married : No -> 1.0 Car : Yes -> 1.0 Children : 0 House : Rent -> 0.75 Owner -> 0.25 Union : Yes -> 1.0 Offsick/Year : 2 WorkYear : 8 Gender : M -> 0.25 F -> 0.75	Salary : 188333 Married : Yes -> 1.0 Car : Yes -> 1.0 Children : 2 House : Rent -> 0.17 Owner -> 0.83 Union : No -> 0.67 Yes -> 0.33 Offsick/Year : 5 WorkYear : 8 Gender : M -> 0.83 F -> 0.17

How do we interpret these results?

Example

We can also see which employee has been assigned to each cluster:

table (15 records)

Table Generate

#Ej	Sueldo	Casado	Coche	Hijos	Alq/Prop	Sindic.	Bajas/Año	Antigüedad	Sexo	$KM-Kmeans	$KMD-Kmeans
1	100000	Sí	No	0	Alquiler	No	7	15	H	cluster-1	1.045
2	200000	No	Sí	1	Alquiler	Sí	3	3	M	cluster-2	0.512
3	150000	Sí	Sí	2	Prop	Sí	5	10	H	cluster-3	0.728
4	300000	Sí	Sí	1	Alquiler	No	15	7	M	cluster-3	1.324
5	100000	Sí	Sí	0	Prop	Sí	1	6	H	cluster-3	0.944
6	400000	No	Sí	0	Alquiler	Sí	3	16	M	cluster-2	0.696
7	250000	No	No	0	Alquiler	Sí	0	8	H	cluster-1	0.983
8	200000	No	Sí	0	Prop	Sí	2	6	M	cluster-2	0.804
9	200000	Sí	Sí	3	Prop	No	7	5	H	cluster-3	0.63
10	300000	Sí	Sí	2	Prop	No	1	20	H	cluster-3	0.814
11	500000	No	No	0	Alquiler	No	2	12	M	cluster-1	0.994
12	80000	Sí	Sí	2	Prop	No	3	1	H	cluster-3	0.629
13	200000	No	No	0	Alquiler	No	27	5	M	cluster-1	1.011
14	100000	No	Sí	0	Alquiler	Sí	0	7	H	cluster-2	0.854
15	80000	No	Sí	0	Alquiler	No	3	2	H	cluster-1	1.07

And assign new employees to the discovered clusters.