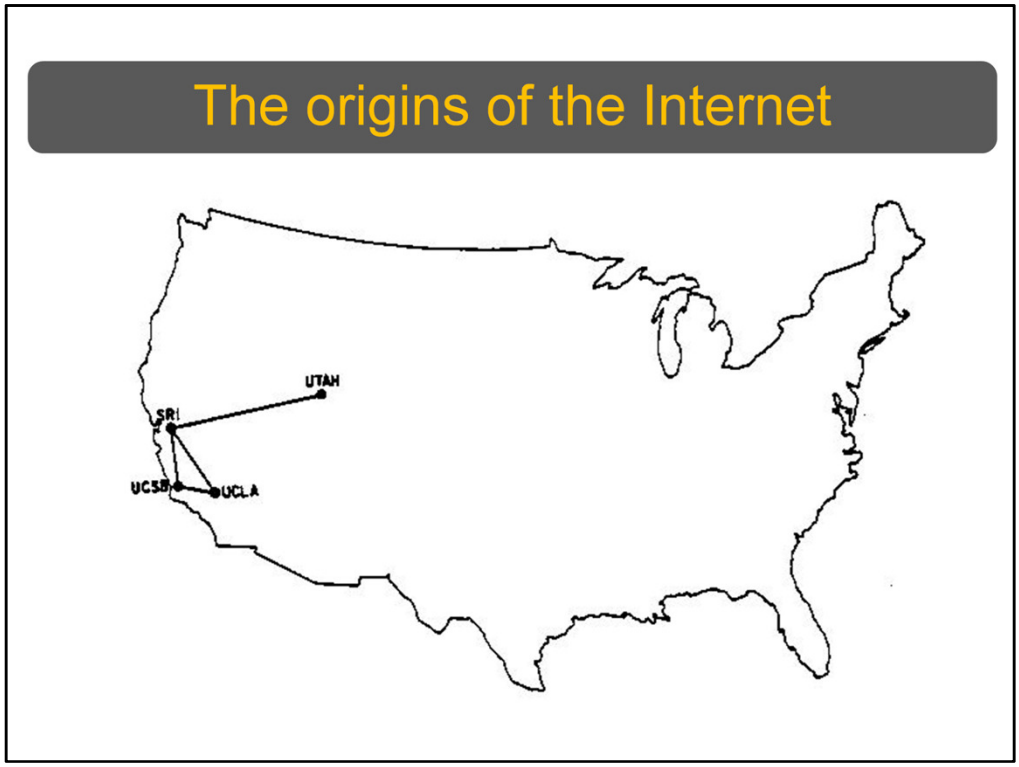
Seminario DeustoTech
18 Mayo 2011

Técnicas de minería de datos para la mejora de servicios en Internet

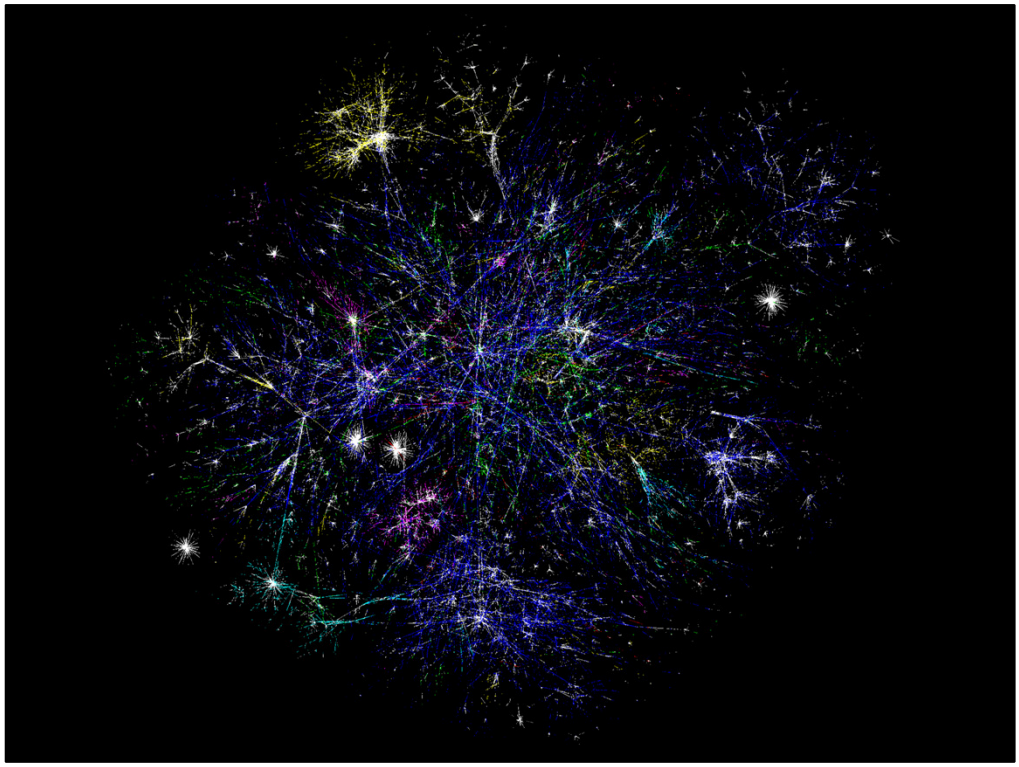
Igor Ruiz-Agundez

DeustoTech, Deusto Institute of Technology, University of Deusto



The origins of the Internet

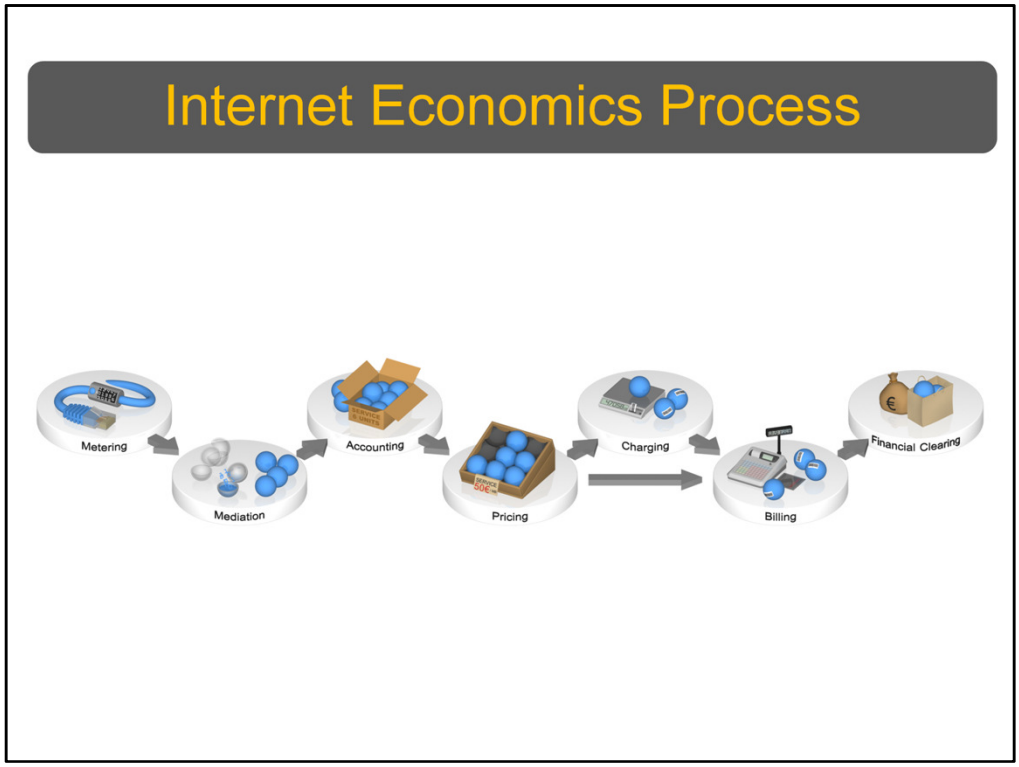
Grew into…



Breve historia de Internet desde el prisma de esta investigación
al principio militares: sus asuntos propios
después académicos: ciencia, conocimiento, experimentación
después grandes empresas: base de la globalización
después generalización en el primer mundo (primera explosión): el ordenador llego a nuestras casas
después modelo de consumo de las punto com: primer gran intento de monetizar la red
después generalización en dispositivos móviles (segunda explosión): Internet en todas partes
después modelo de consumo de los servicios y las app: segundo gran intento de monetizar la red

¿Hacia donde se dirige la economía en Internet?
Y sobre todo, ¿cómo podemos participar activamente en ella?

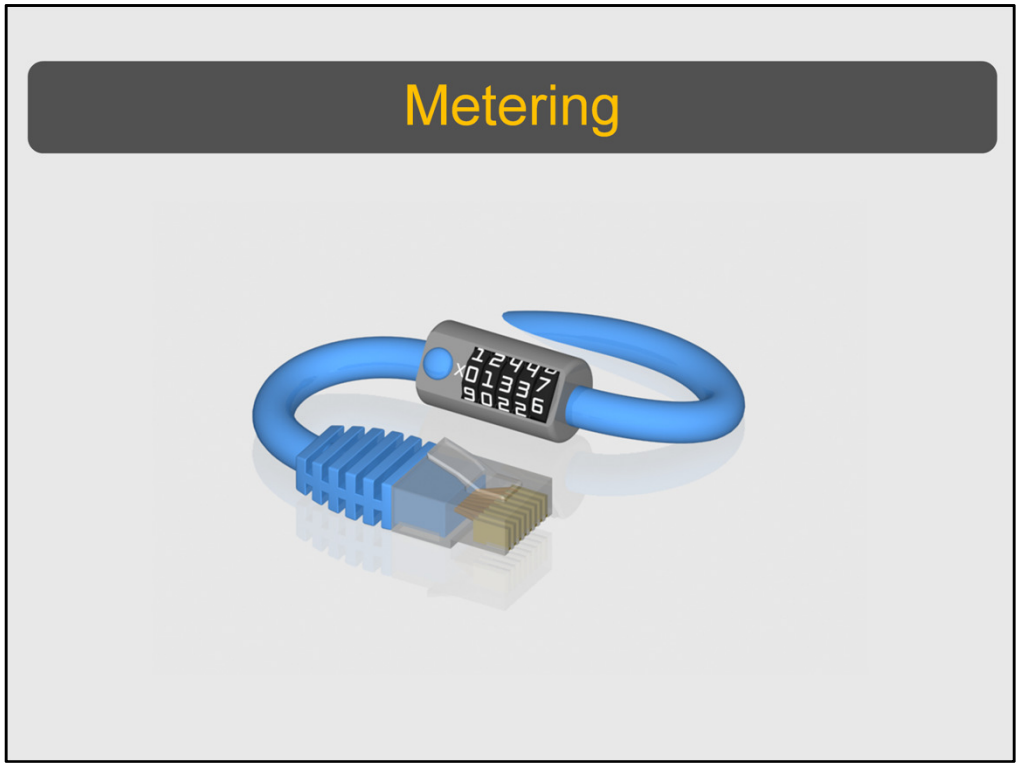
Primero necesitamos un marco de trabajo que defina el proceso de la economía en Internet desde un punto de vista tecnológico y de ingeniería.
Se hace necesario modelar el proceso económico en Internet.



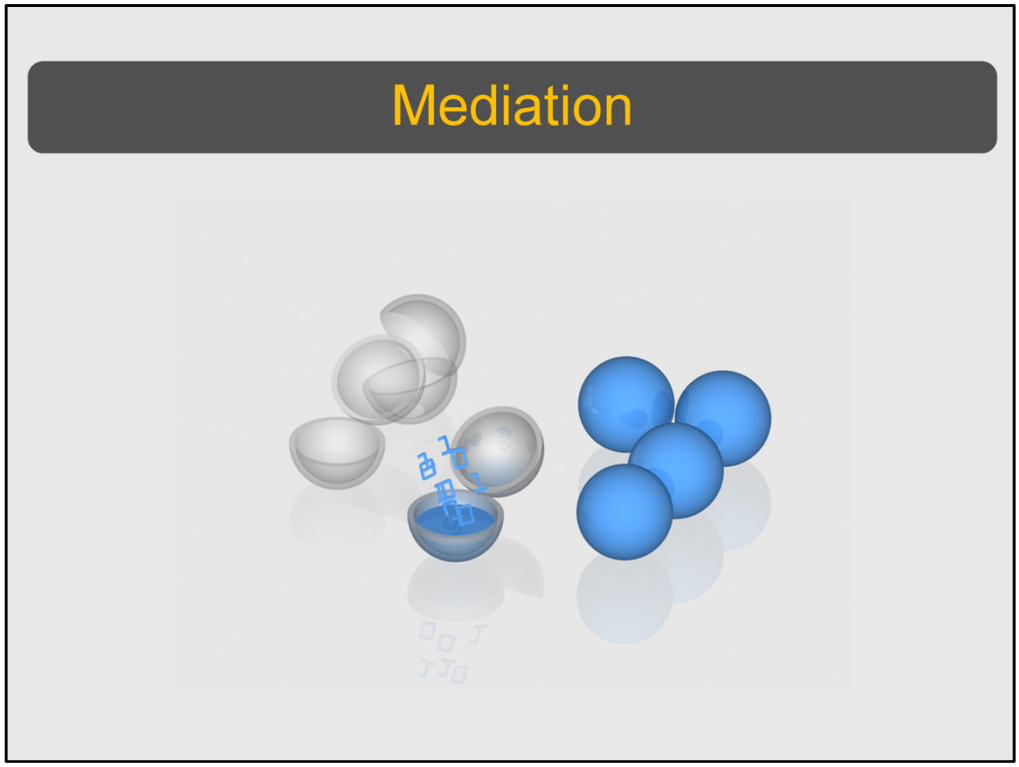
Internet economics is a research area which covers such diverse aspects as economics, engineering, and policy. It overlaps concepts and procedures of theses areas creating new semantics and ideas from them.

The related work in Internet economics is vast and covers many areas of knowledge. From economics to engineering, the different projects, publications and products that group in Internet economics have a big impact in the scientific community.

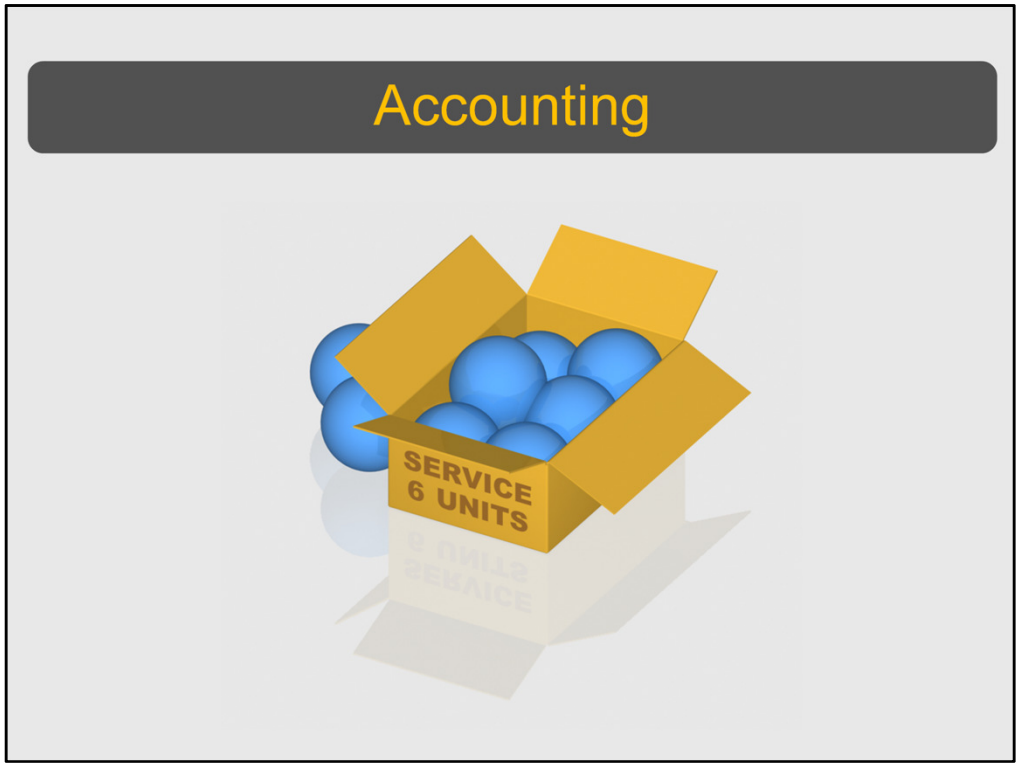
In order to integrate all the concepts related to Internet accounting that are present in the literature, we introduce a taxonomy representing the economics process. We detailed all the functions involved in it, looking at all the relationships between them. We believe that the presented taxonomy contributes to the learning, training and assessing in the area of accounting because it gives an integrated vision of the process. As it defines a common vocabulary, it is also useful for the definition of the economics requirements among different actors.



Metering is the function that collects the information flow regarding the resource usage of a certain service by a consumer and its usage. This measurement data is formed by service usage metrics provided by the monitoring function.



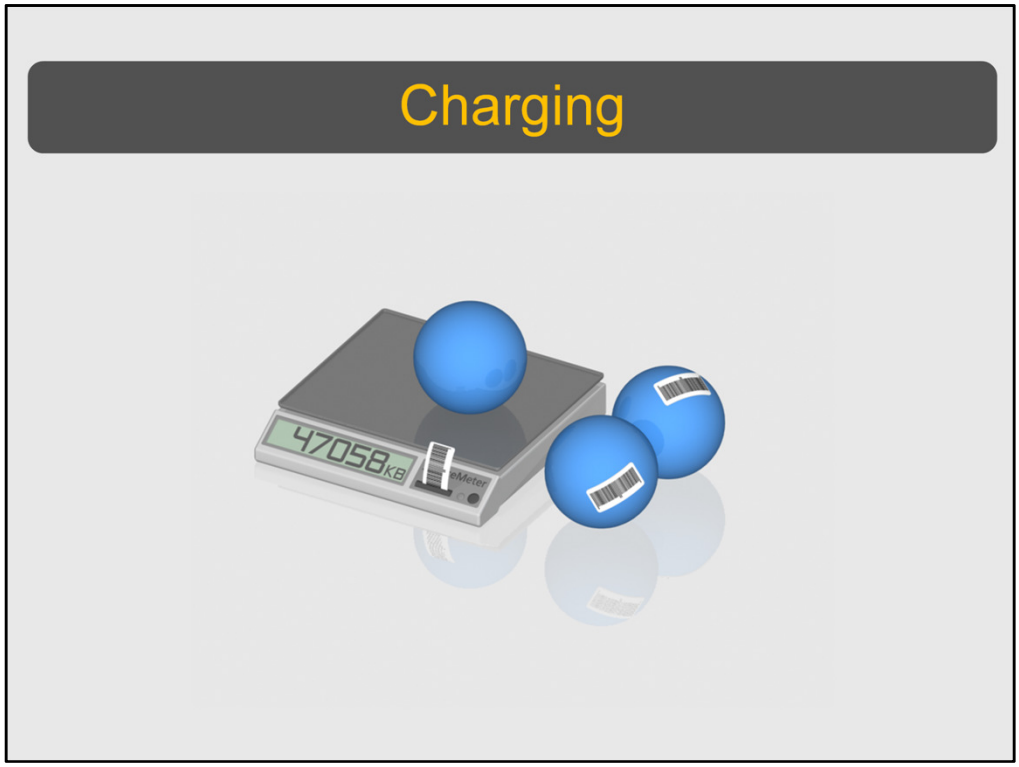
Mediation is intended to filter, collect, generate, aggregate, correlate, and reconcile raw technical data by transforming these metering records into a data format that can be used for storing and further processing. In this way, data processing is easier and the different functions of the economic process require less mash-ups and conversions, resulting in a better performance.



Accounting is the process of filtering, collecting and aggregating the information that reflects a resource usage by a certain consumer. This process will generate session records whose format will depend on the service infrastructure and the service provider. The session records represent the resource usage over a session.

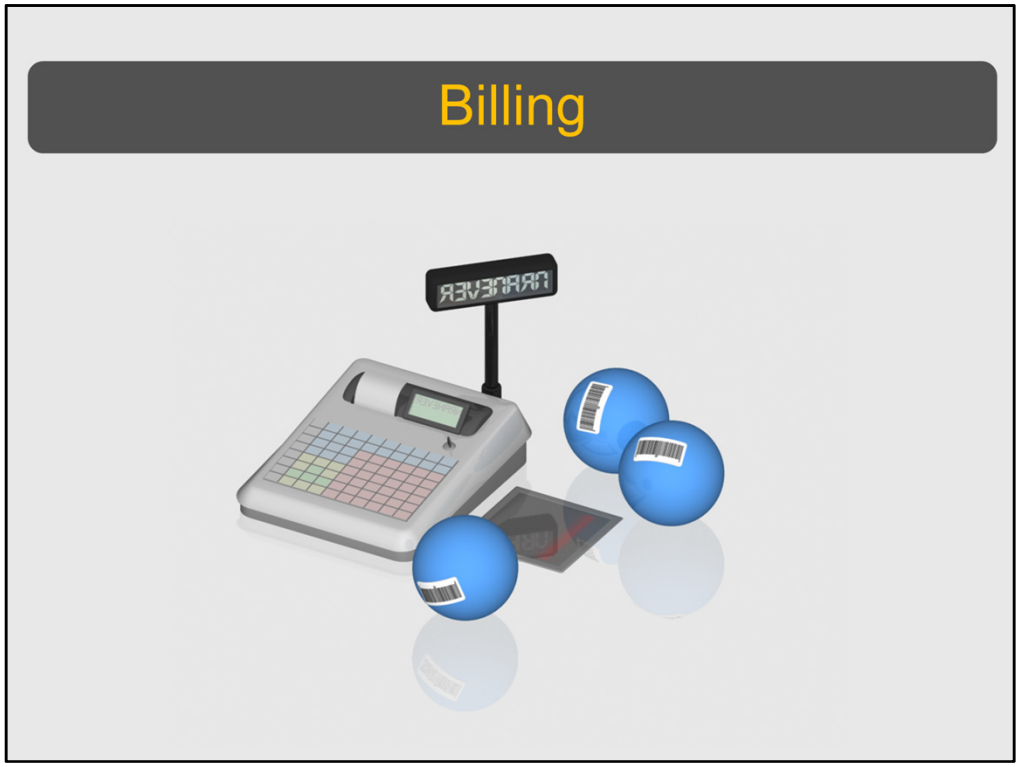
Accounting gateways creating the session records may do so by processing interim accounting events or accounting events from several devices serving the same user.

Pricing is the function of giving a price to a certain resource usage. It is a critical function for the full economics process because it defines the price that a basic quantity of the service will cost.



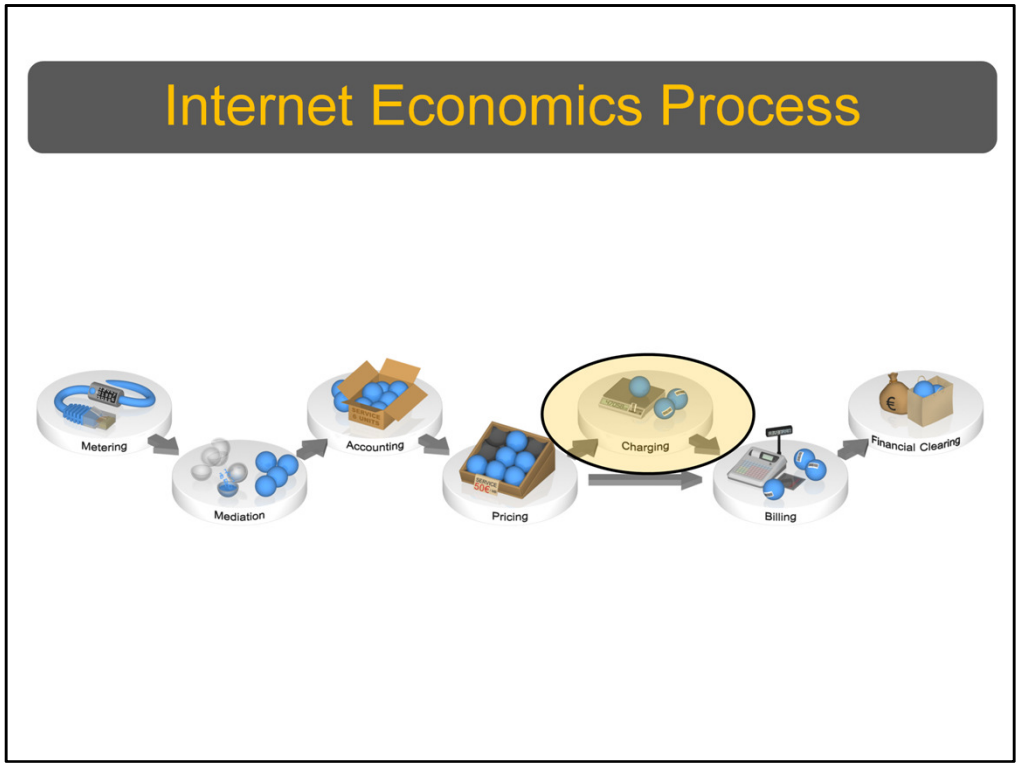
Charging is the process of calculating the cost of a resource usage, the function that translates technical values into monetary units by applying a pricing function to the session records. It correlates session records, from the accounting function, and resource usage unit price to generate charge records.

These charge records are formed by the technical quantities of a resource usage and their corresponding monetary units. The records can be used for multiple purposes of support systems: statistical analysis, data mining, auditing, revenue estimation, financial planning, structure dimensioning, or any other support system operation.



Billing, or invoicing, is the process of transforming charge records into the final bill, or invoice, summarizing the charge records of a certain time period (usually a month) and indicating the amount of monetary units to be paid by the customer.

The financial clearing function includes activities from a commitment for a transaction to its settlement. In the case of resource accounting, this function implies the payment of a bill. Payment is the function of transferring the money of the client to the service provider. The amount to transfer is defined by the bill.



De las posibles áreas de trabajo, decidimos centrarnos en el charging.

Esta función es la que más nos interesa ya que es donde tienen lugar la mayoría de las operaciones de soporte del negocio.

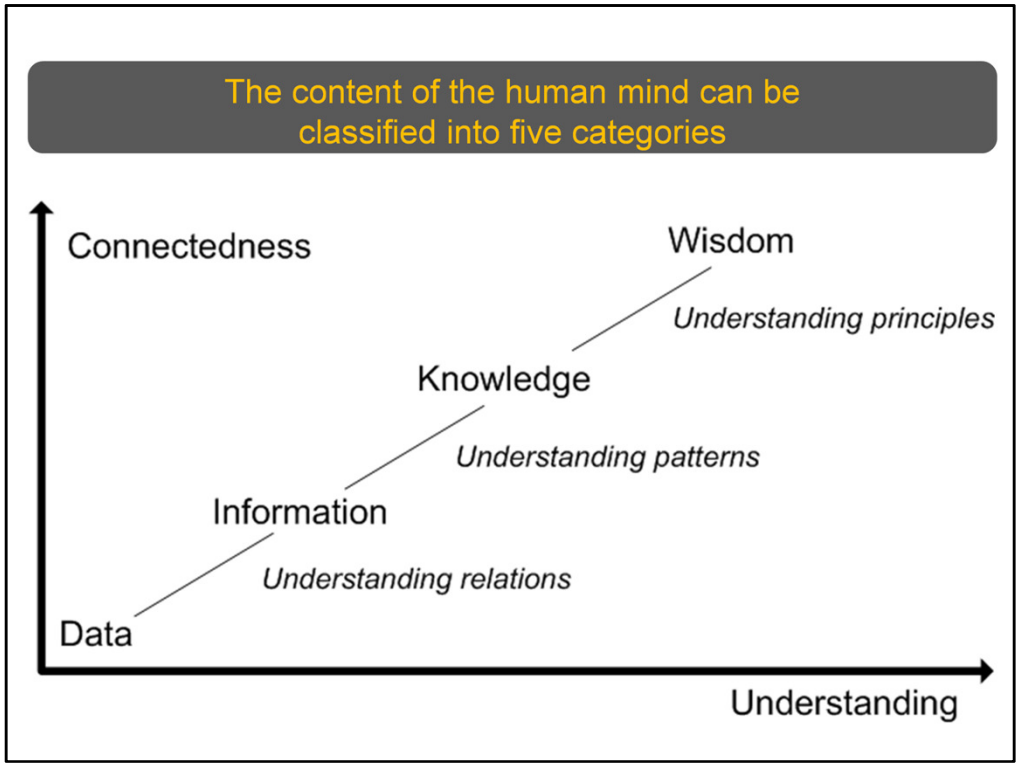
De ellas, nos centraremos en la minería de datos.

En la función de charging, se gestionan los aspectos principales de un servicio.

Y lo más importante para mejorar un servicio son los datos.

Podemos tener los mejores servicios del mundo y la mejor orquestación posible de ellos. Pero, si no analizamos los datos de su uso, estamos perdiendo una información muy valiosa.

Por ello, hemos de analizar tres cosas: los datos, los datos y los datos.
Los datos son la base del conocimiento, los pilares que nos van a ayudar a mejorar nuestros servicios.



Veamos en detalle la evolución de los datos.

Data: symbols

Information: data that are processed to be useful; provides answers to "who", "what", "where", and "when" questions

Knowledge: application of data and information; answers "how" questions

Understanding: appreciation of "why"

Wisdom: evaluated understanding.

La minería de datos (DM, Data Mining) se engloba dentro de las técnicas de Knowledge Discovery (Descubrimiento de Conocimiento).

Consiste en la extracción no trivial de información que reside de manera implícita en los datos.

Dicha información era previamente desconocida y podrá resultar útil para algún proceso.

En otras palabras, la minería de datos prepara, sondea y explora los datos para sacar la información oculta en ellos.

Bajo el nombre de minería de datos se engloba todo un conjunto de técnicas encaminadas a la extracción de conocimiento procesable, implícito en las bases de datos.

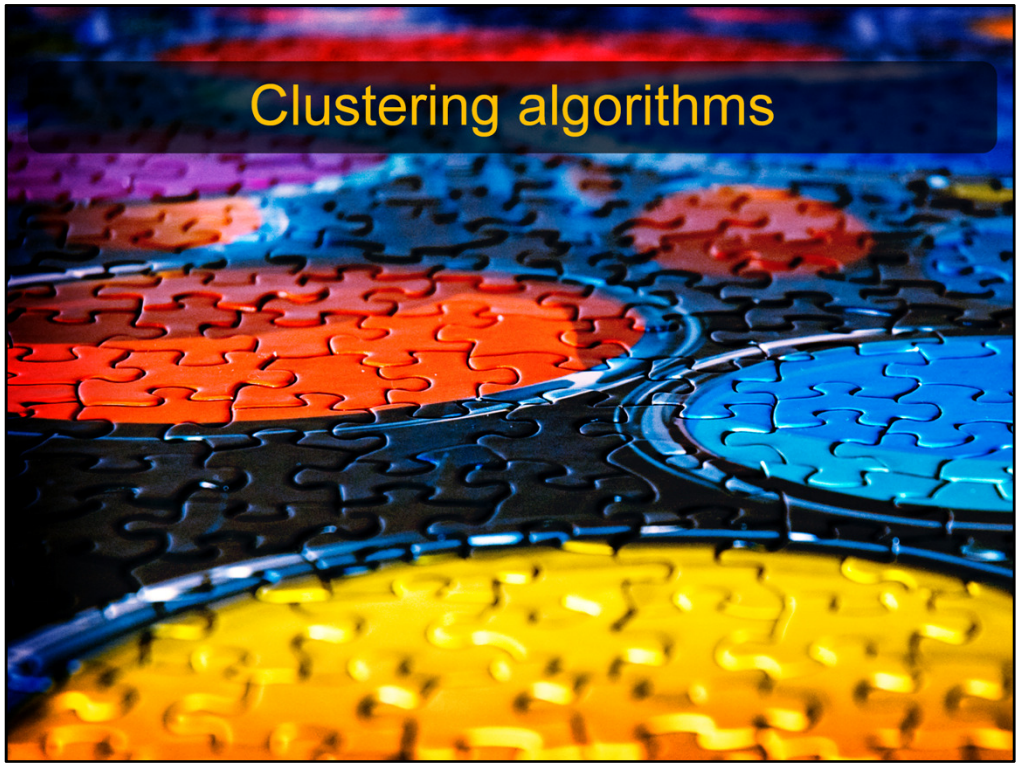
Está fuertemente ligado con la supervisión de procesos industriales ya que resulta muy útil para aprovechar los datos almacenados en las bases de datos.

Las bases de la minería de datos se encuentran en la inteligencia artificial y en el análisis estadístico.

Mediante los modelos extraídos utilizando técnicas de minería de datos se aborda la solución a problemas de predicción, clasificación y segmentación.

El conocimiento extraído de los datos nos ayudará a mejorar los servicios.

En ciertas ocasiones, esta mejora puede ser más valiosa que el oro.



De las múltiples técnicas de minería de datos, nos centramos en los algoritmos de clustering.

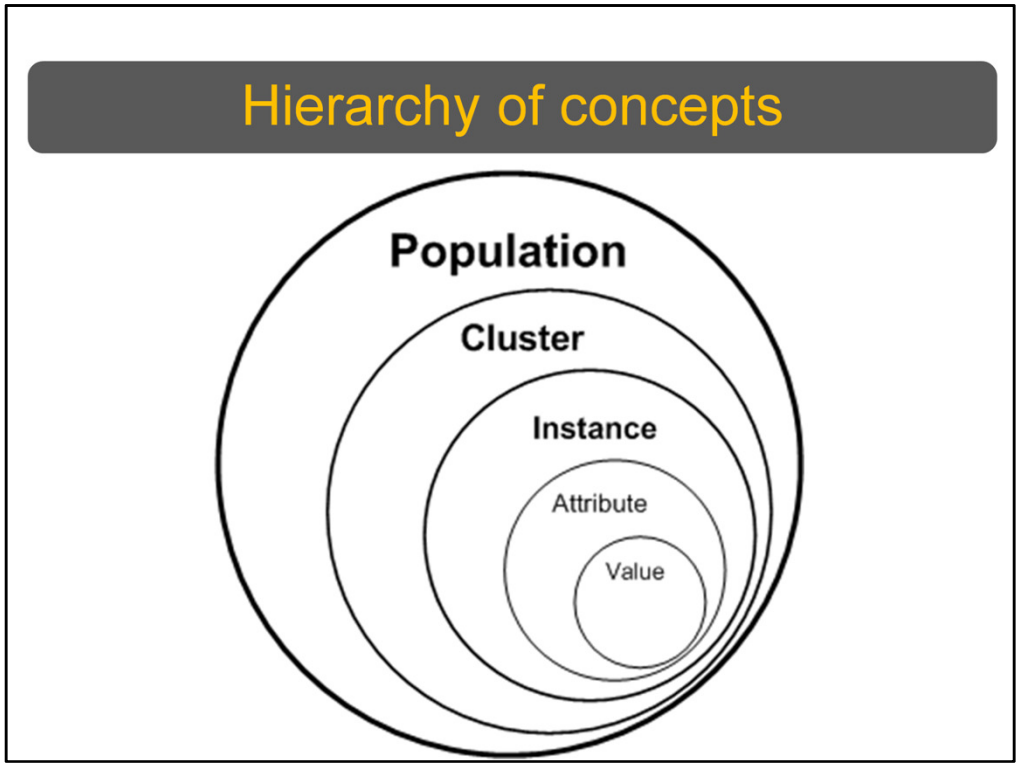
Los algoritmos de clustering tienen como objetivo agrupar la información.

Agrupan datos de observaciones, vectores o cualquier otro elemento, buscando su estructura natural.

Definiciones más formales presentan los algoritmos de clustering como una técnica de organización estadística.
Realizan una comparación cuantitativa de los elementos de una población, agrupándolos en grupos.

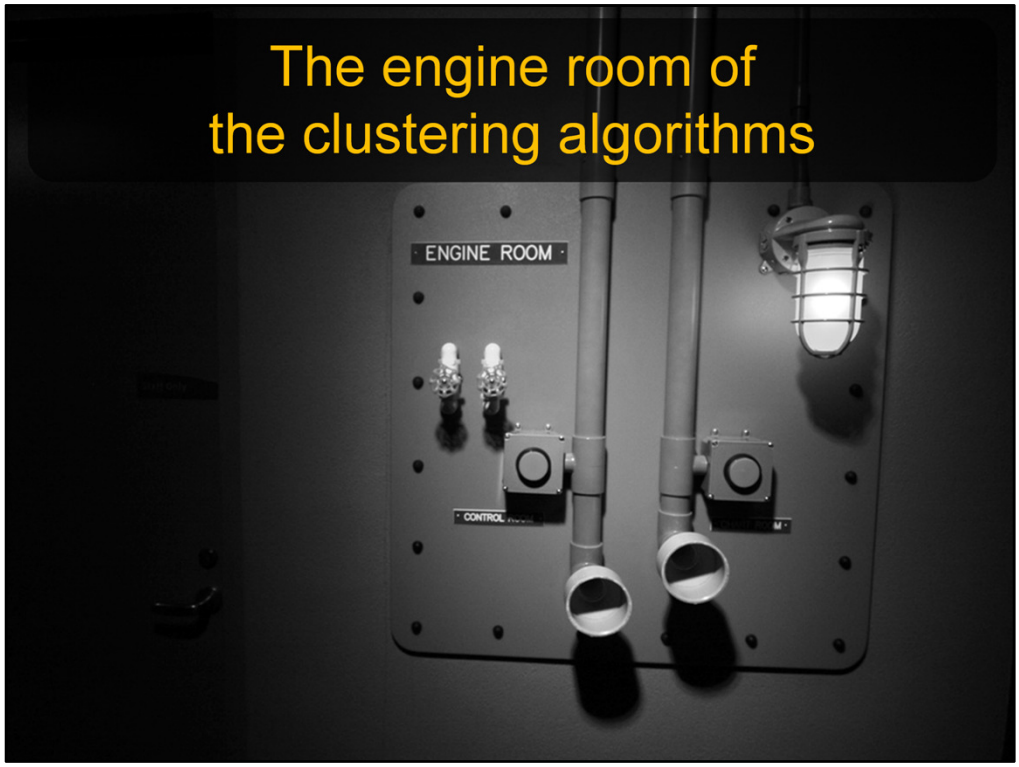
Los grupos (clusters) generados son representativos y útiles ya que representan la estructura natural de los datos.

Gracias a esta estructura, obtenemos conocimiento en el dominio de los datos de origen.
Los algoritmos de clustering se utilizan en áreas tan diversas como la psicología, biología, estadística, aprendizaje automático, etc.



Para entender el funcionamiento de los algoritmos de clustering hemos de definir primero la jerarquía de concentos subyacente.

Población
Cluster
Instancia
Atributo
Valor



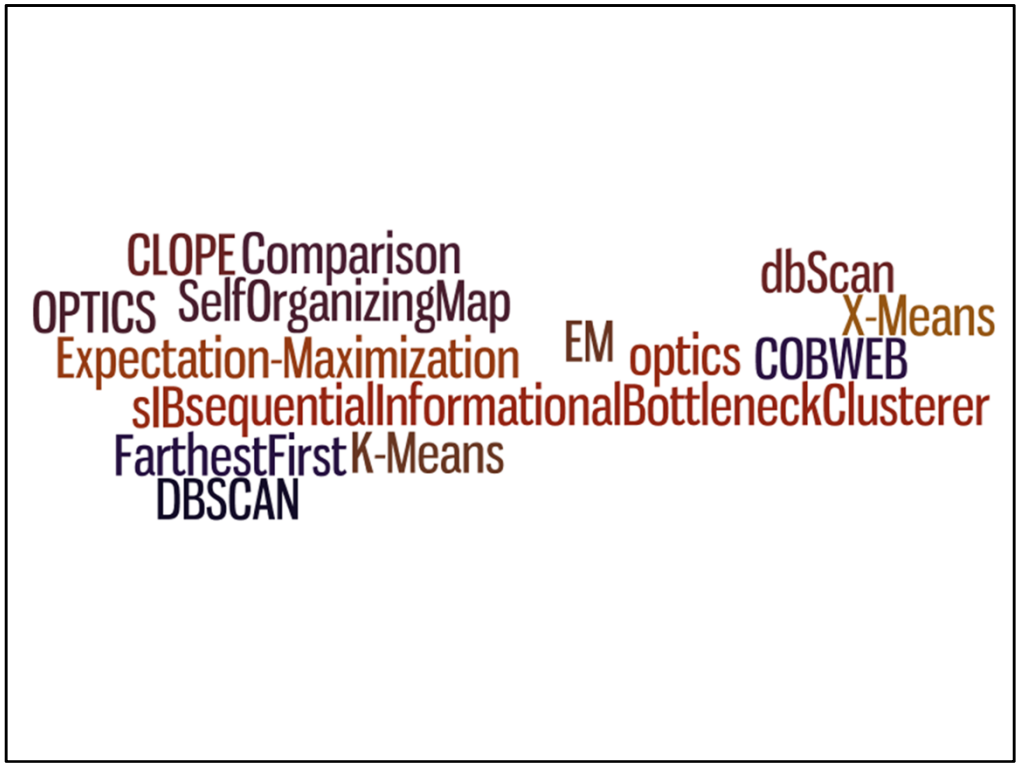
Para buscar los clusters, se utilizará los elementos internos de nuestra población.

Se utilizarán los valores que describen los atributos de las instancias en busca de relaciones entre ellas.

El objetivo es asignar cada instancia a al menos un cluster.
Además, los elementos dentro de un cluster serán parecidos entre si y distintos respecto al resto de los clusters.
De esta forma, cada cluster tendrá su propia idiosincrasia, sus propias características.

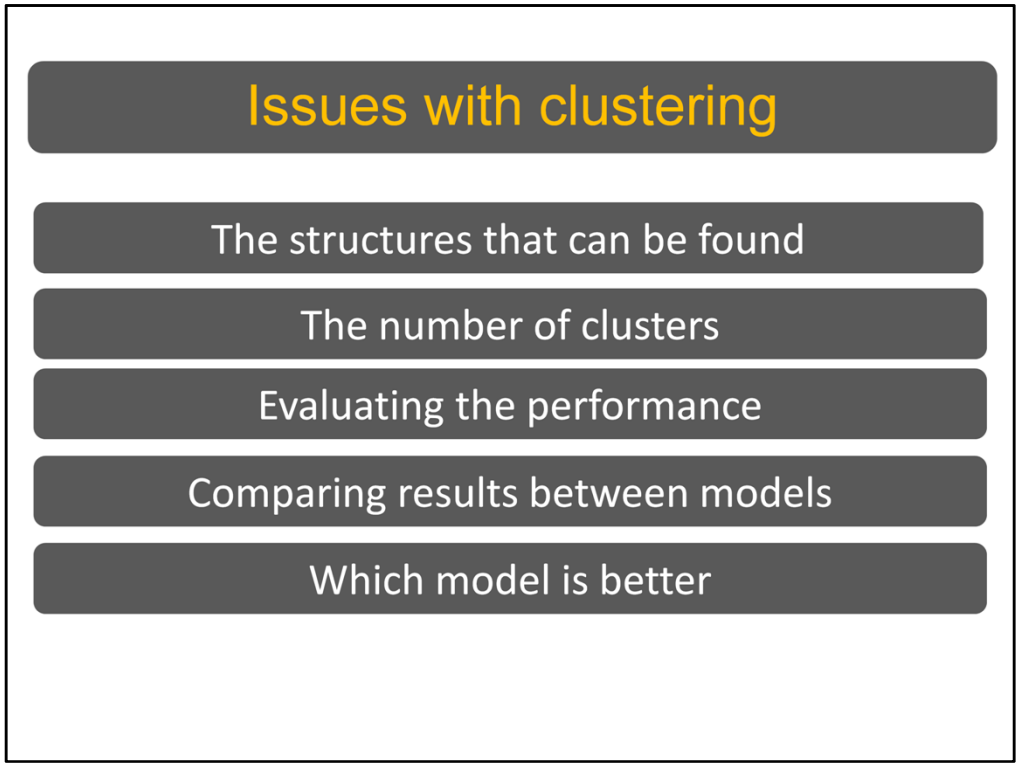
La mayoría de los algoritmos de clustering asignan a cada instancia a un único cluster, de forma que no hay solapamiento entre grupos.
De esta forma, cuanto más se parecen las instancias de un cluster más homogéneo y significativo será.
Cabe mencionar que existen algoritmos de clustering difuso que pueden asignar las instancias a más de un cluster.

El clusterizado asigna a cada instancia una etiqueta que identifica al cluster al que pertenece.
Por ello, algunos autores consideran que la clusterización es un tipo de clasificación no supervisada.
La clusterización no utiliza etiquetas de antemano, salvo para realizar posibles validaciones.



Existen muchos algoritmos de clustering, entre los que cabe mencionar:

CLOPE
COBWEB
DBSCAN
Expectation-Maximization (EM)
FarthestFirst
OPTICS
sIB
K-Means
X-Means
Comparison
SelfOrganizingMap
sequentialInformationalBottleneckClusterer
optics_dbScan



Different surveys in clustering evaluation identify concrete questions in cluster evaluation:

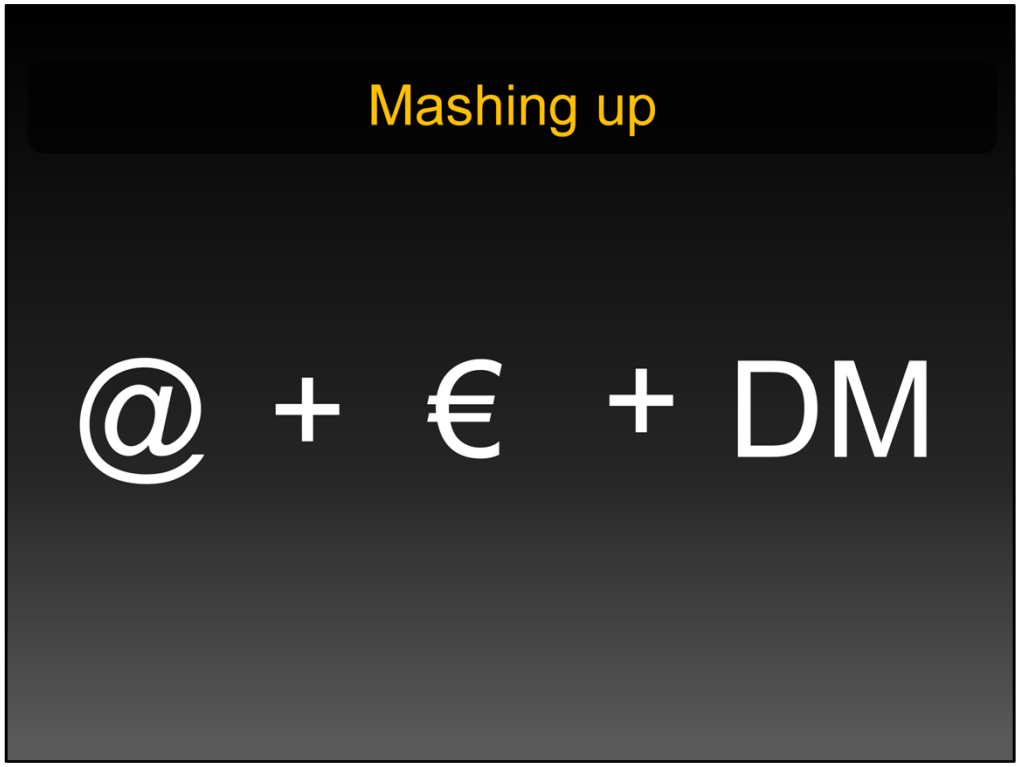
Determining the structures that can be found in the dataset (which a domain expert may know) from random structures.

Finding out the number of clusters in which the dataset is divided, especially when there is not a domain expert. In this cases we use unsupervised clustering algorithms and we have no indication of the number of clusters present on a dataset.

Evaluating the performance model with new records or entries before having them.

Comparing the resulting model results with an external test-set.

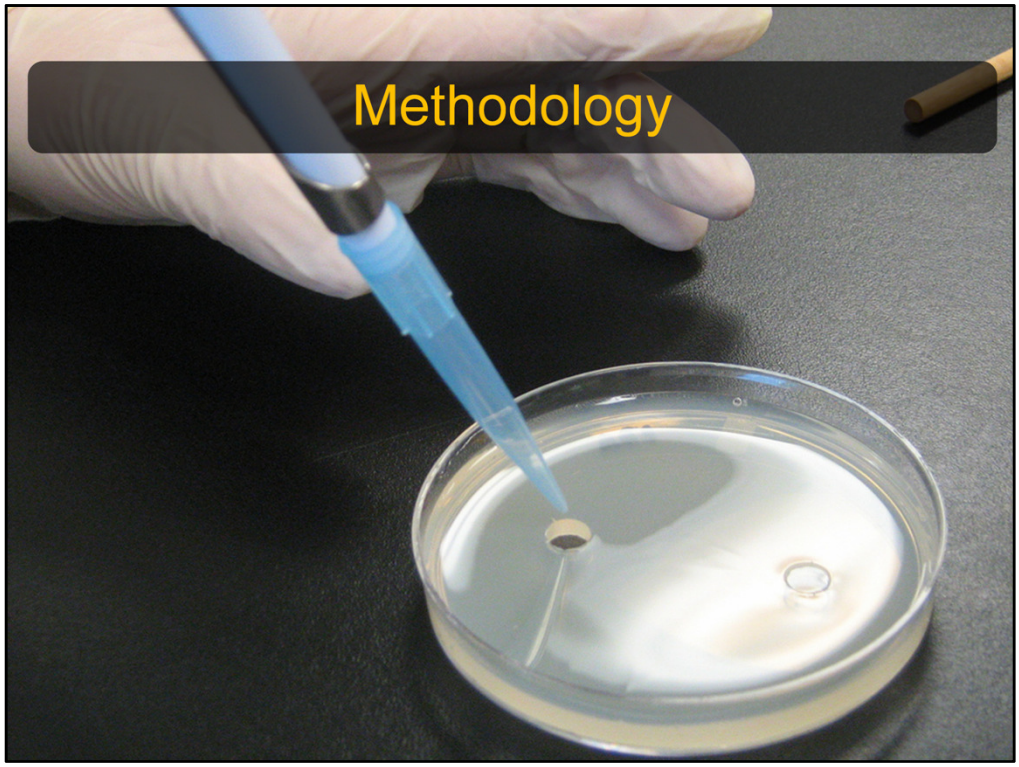
Having two cluster models, determining which model is better than the other. We define better as being more appropriate to the requirements of the problem.



Mashing up

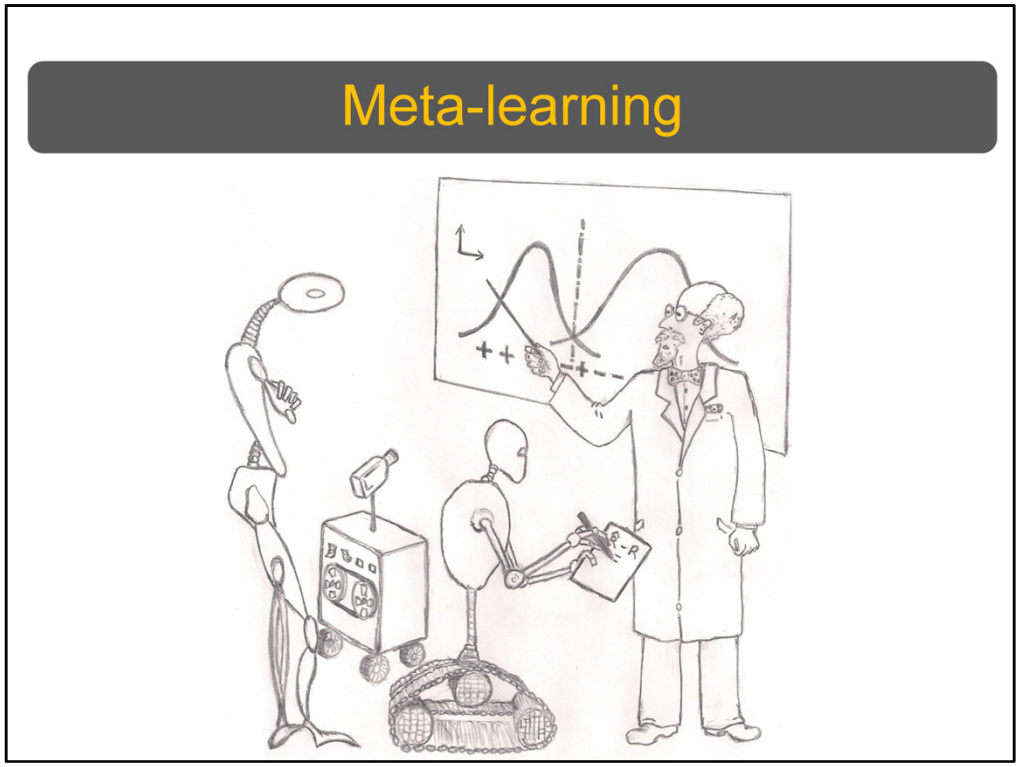
Internet services + Economics + Data mining

We designed a methodology



The first attend to describe the steps that form a knowledge discovery process come out with an interactive and iterative process involving steps that included data selection, preprocessing, integration and transformation (we define all of them as data acquisition), data mining algorithm selection and result interpretation. This process is repeated changing the data and the algorithm until the results are acceptable. Other proposals define more detailed methods, including the identification of the required resources and knowledge post-processing steps. In general, most of the authors define the required steps as extracting knowledge by choosing the features we want to study, selecting the proper clustering algorithm, validating the results and interpreting them.

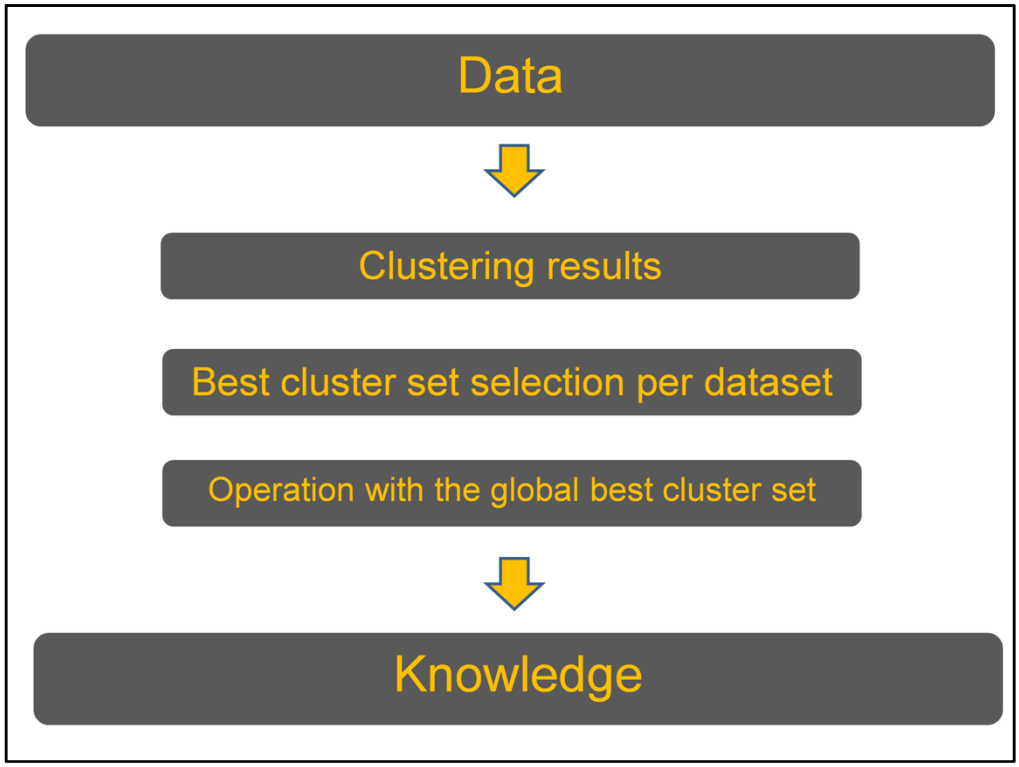
One of the hardest steps on clustering is the validation step. Normally, it is performed by a domain expert that studies the algorithms parameters and their results and selects the best model based on her knowledge. She will analyse the attributes, the ranges, the distances between clusters, among many other aspects. Nevertheless, and due to the specifications of most of the problems, in which we do not have a domain expert, we need to find an non-supervised way to validate the results of clustering algorithms.



Meta learning is a subfield of Machine learning where automatic learning algorithms are applied on meta-data about machine learning experiments. Although different researchers hold different views as to what the term exactly means, the main goal is to use such meta-data to understand how automatic learning can become flexible in solving different kinds of learning problems, hence to improve the performance of existing learning algorithms.

Flexibility is very important because each learning algorithm is based on a set of assumptions about the data, its inductive bias. This means that it will only learn well if the bias matches the data in the learning problem. A learning algorithm may perform very well on one learning problem, but very badly on the next. From a non-expert point of view, this poses strong restrictions on the use of machine learning or data mining techniques, since the relationship between the learning problem (often some kind of database) and the effectiveness of different learning algorithms is not yet understood.

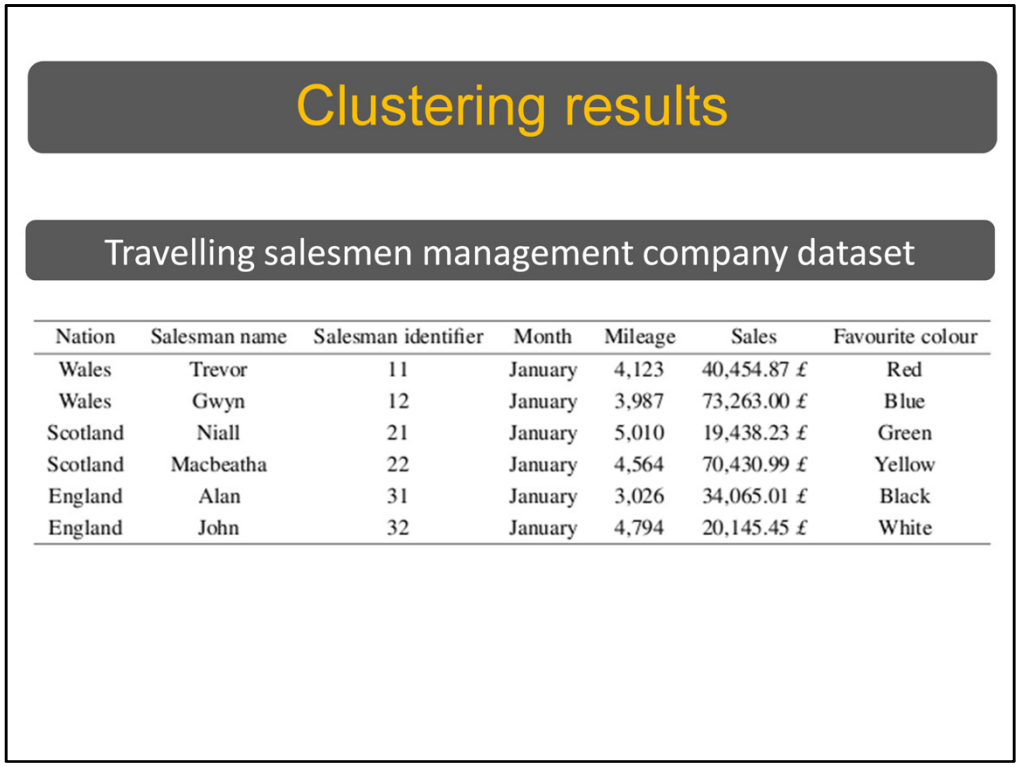
By using different kinds of meta-data, like properties of the learning problem, algorithm properties (like performance measures), or patterns previously derived from the data, it is possible to select, alter or combine different learning algorithms to effectively solve a given learning problem. Critiques of meta learning approaches bear a strong resemblance to the critique of metaheuristic, which can be said to be a related problem.



From data to Knowledge

Clustering results
Best cluster set selection per dataset
Operation with the global best cluster set

Knowledge



Nation	Salesman name	Salesman identifier	Month	Mileage	Sales	Favourite colour
Wales	Trevor	11	January	4,123	40,454.87 £	Red
Wales	Gwyn	12	January	3,987	73,263.00 £	Blue
Scotland	Niall	21	January	5,010	19,438.23 £	Green
Scotland	Macbeatha	22	January	4,564	70,430.99 £	Yellow
England	Alan	31	January	3,026	34,065.01 £	Black
England	John	32	January	4,794	20,145.45 £	White

Clustering results is the phase devoted to the execution of the clustering algorithms.

They are fed with the available datasets to be data-mined, the action plan to get ready the dataset, and the parameters that each algorithm requires (we assume all datasets share the same structure).

The resulting cluster sets are gathered in a database that will be used in the next step.

Please note that this phase will generate a different distribution set for each algorithm and parameter configuration.

Let us enrich the explanation of the methodology with an example: think of a travelling salesmen management company operating in Britain. In that case, we would have 3 separate datasets (England, Scotland, and Wales), recording all the journeys of their respective salesmen within a certain period of time.

Best cluster set selection per dataset

Number of instances per cluster metric

Model identifier \ Cluster	Cluster0	Cluster1	Cluster2	Cluster3
1	3	3		
2	2	3	1	
3	2	1	1	2

Bigger difference of number of instances

Model identifier \ Algorithm	SimpleKMeans
1	0
2	2
3	1

The best cluster set selection module picks out a base attribute in these cluster sets that allows it to compare them.

It must be present in all the algorithms and may be a meta-data of the algorithms (e.g. number of instances) or an attribute of the datasets (as in our example, where it would be the mileage).

The metric will use it as the fitness function to measure the cluster sets performance (e.g. number of instances per cluster or mileage per month). This metric may include some constrains to be applied (e.g. monthly mileage must be under 5,000 miles). Finally, a criterion sets the ranking rules (e.g. lowest mileage per month).

The ultimate goal here is to select for each dataset the best cluster set according to the criterion and also to determine which one of them is best suited to represent the overall behaviour. Please note that with “cluster set” we mean algorithm model, result information, cluster assignments to the instances, and other output information. In the example, after applying the lowest mileage per month criterion to journeys above 5,000 miles, we would obtain several distribution sets for England, Scotland, and Wales ranked from the lowest (and also best) to the biggest miles.

Now we are already able to perform a global validation of the cluster set of each dataset by comparing their metric results. This task only shows whether the selected best model is correct; it does not provide any further “useful” information.

Moreover, this information will depend on the initial datasets. Thus, it could differ from one to another as this global validation does not include a deep attribute analysis.

Operation with the global best cluster set

Selected dataset best cluster set analysed by nation attribute, in percentage

Nation Attribute \ Cluster label	Cluster0	Cluster1	Cluster2
Wales	0	16,66	16,66
Scotland	33,33	0	0
England	0	33,33	0
Total	33,33	50	16,66

The point here is modelling the general representation for all the datasets and use it to extract data-independent knowledge (i.e. applicable to the overall problem).

For this purpose, in the operation stage we will use the best cluster set obtained in each dataset focusing on a certain operation type (e.g. commercial efficiency of the travelling salesmen).

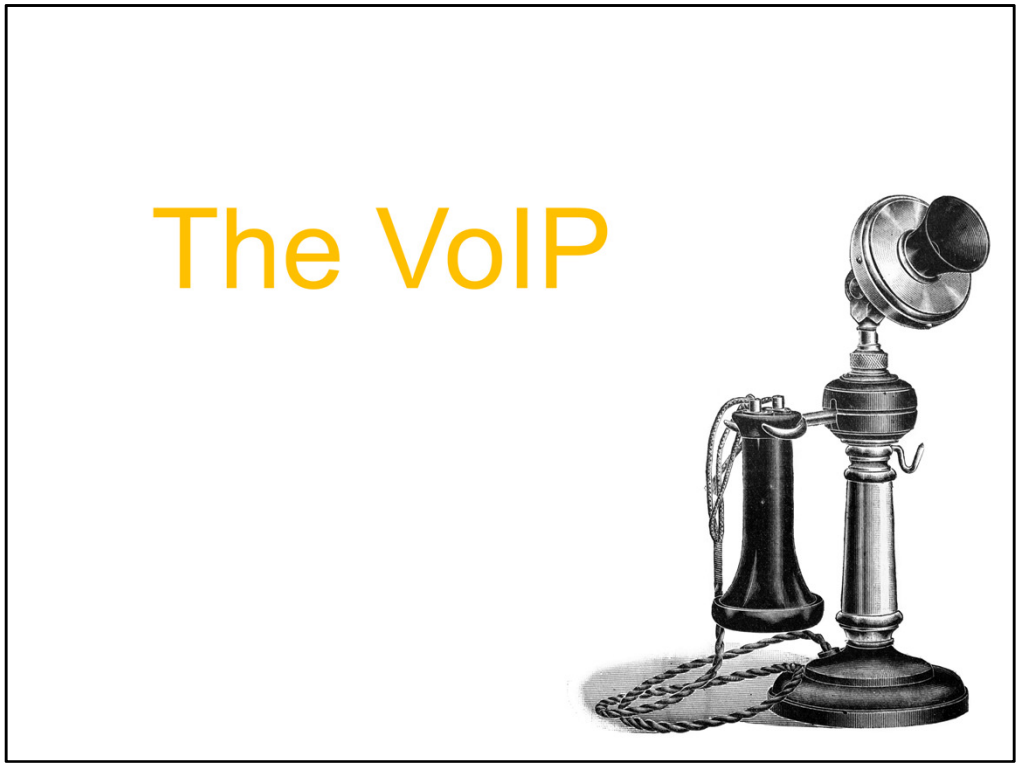
This operation type will be specified by the operation attribute that will be the measure unit (e.g. sales).

If the distance between the results of the dataset best cluster sets and the optimal global one shows a similarity above a pre-defined threshold, we can conclude that the selected optimal global cluster set is representative of the whole problem data.

In the example, we would calculate the commercial efficiency (i.e. sales/miles) of the best cluster set in England, Wales, and Scotland.

Assuming for instance the best cluster set of Wales as the optimal one, and that it shows a similarity with England and Scotland above the given threshold; we will conclude that the obtained sales/miles information for each of the groups of the cluster set can be exported to the whole Britain.

Llevamos esta metodología a un caso de uso real



The general banner of VoIP services groups a number of technologies aiming at enabling voice communication over an IP network. Services range from a simple call between two users to more complex practices including multiple user teleconference, call transferring, call-center functionalities, and so on.

The data about user identifier, call duration, receivers, call time and type, channels, etc. is recorded and can be processed by data mining techniques to obtain valuable knowledge such as call sinks (i.e. very frequently- called users), call niche (i.e. very frequently calling users), diverse calling trends (e.g. call type concentration or peak hours), and so on.

Many applications may take advantage from this information. For example, congestion prediction tries to alleviate situations in which the resources are saturated. Depending on the user types and their behaviour, companies may address different marketing strategies. In the same way, network planning intends to design and organise the infrastructures in an optimal way to adequately respond to the system's requirements. We focus here on this latter. Specifically, we have access to the Call Detail Records (CDR) and our objective in this work is to classify the different types of users with respect to their calling behaviour. We presume this information (e.g. peak hours or regular calling trends) may help us optimise the trunk-lines design infrastructure.

In this use case, we apply the methodology introduced to a specific VoIP service. The knowledge that we may obtain from the VoIP datasets are relevant to different applications.

Nevertheless, among these applications, this use case focused on the peculiarities of the network planning.

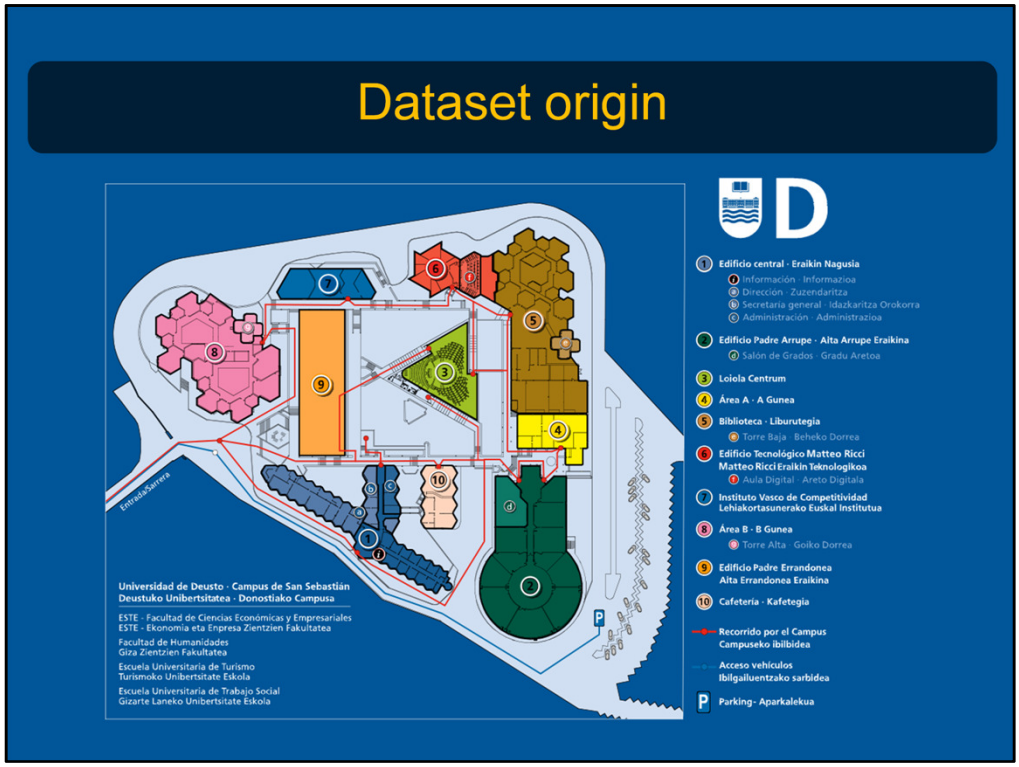
Networking planning may be accomplished a priori, before infrastructures deployment, or on the fly, in order to improve them. To this end, we may focus on the channels (i.e. used physical trunk-line), call length, call source and destination, context (e.g. land-line, mobile, international) or call time (i.e. time of the call).

Among the attributes of our dataset, we focus on the call behaviour over the day. From the point of view of the service client (say the corporation that contracts a VoIP service), this information is crucial to devise the optimal infrastructure in operation time (i.e. on the fly).

The diverse clustered groups arising from this analysis will show unique trends regarding their activity period, so the network administrator will be able to accordingly improve the performance for instance by implementing traffic balancing techniques or optimising the use of the existing trunk-lines or selecting the most suitable pricing scheme with the provider.

From the point of view of the service provider, this knowledge may also be useful to

improve the revenue management, prevent fraud by the analysis of anomalous deviations in the client behaviour, planning infrastructure re-dimensioning to answer the needs of client's quality of service or better aim marketing strategies (e.g. customer-specific campaigns to increase client loyalty) .



Our experiment studies the behaviour of VoIP users in a certain corporation. More specifically, we use the accounting records generated by a VoIP Private Branch Exchange (PBX) with more than 1,300 users as our experimental dataset. A PBX is a telephone exchange that serves a particular corporation; all CDRs belong to the users of this corporation. PBXs are used to connect the internal telephones of a corporation and to connect to the public switched telephone network (PSTN) by means of trunk lines (i.e. bunch of lines shared by many clients). They can include modems, fax machines, and other telephone devices.

Among the possibilities that the VoIP PBXs offers, we focus on the Call Detail Records logs. These records are produced when two or more participants communicate over a partial or complete Internet-based voice connection.

The call is normally initiated by one of the participants (i.e., the call initiator) and is received by one or more participants (i.e., call recipients). The call can be IP to IP, PSTN to IP, IP to PSTN, mobile to IP or any other possible combination. In each case, the resulting CDR reflects the nature of the call and provides all its details. The data correspond to all of the CDRs recorded in 2008, which number more than 700,000 entries.

We obtained data from the “Asterisk” PBX database records of a corporation. Asterisk is a VoIP PBX based on free software. It allows linked telephones to make regular calls and to connect to other telephone services, including the PSTN and other VoIP services.

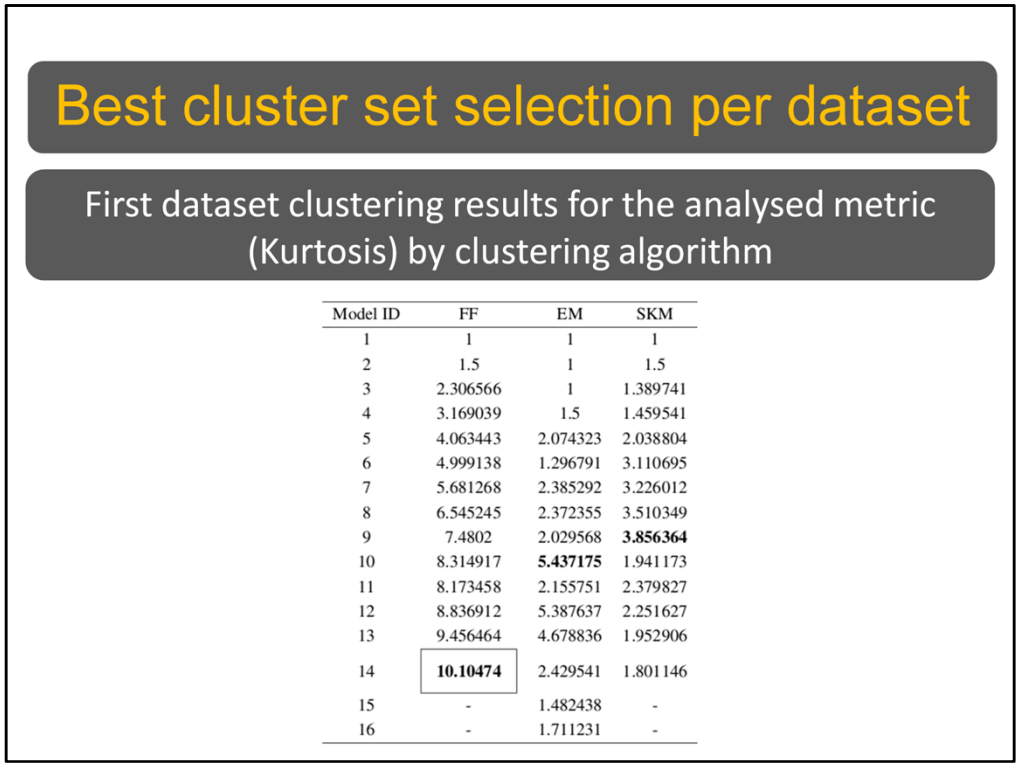
We decided to use this data source because it represents a medium-sized corporation with numerous types of users and terminals. To address data privacy concerns, all confidential information was made anonymous, guaranteeing the rights of users under compliance with the existing laws (under the Spanish jurisdiction) on data protection.

Clustering results

Algorithms execution parameters values

FF		EM			SKM	
Model ID	numClusters	ID	maxIterations	numClusters	ID	numClusters
1	2	1	100	-1	1	2
2	3	2	1000	-1	2	3
3	4	3	100	2	3	4
4	5	4	100	3	4	5
5	6	5	100	4	5	6
6	7	6	100	5	6	7
7	8	7	100	6	7	8
8	9	8	100	7	8	9
9	10	9	100	8	9	10
10	11	10	100	9	10	11
11	12	11	100	10	11	12
12	13	12	100	11	12	13
13	14	13	100	12	13	14
14	15	14	100	13	14	15
		15	100	14		
		16	100	15		
Common execution parameters by algorithm						
seed =	1	minStdDev =	1.0E-6		distance Func. =	Euclidean Dist.
		seed =	100		max Iterations =	500
					seed =	10

Algorithms execution parameters values by clustering algorithm



Model ID	FF	EM	SKM
1	1	1	1
2	1.5	1	1.5
3	2.306566	1	1.389741
4	3.169039	1.5	1.459541
5	4.063443	2.074323	2.038804
6	4.999138	1.296791	3.110695
7	5.681268	2.385292	3.226012
8	6.545245	2.372355	3.510349
9	7.4802	2.029568	**3.856364**
10	8.314917	**5.437175**	1.941173
11	8.173458	2.155751	2.379827
12	8.836912	5.387637	2.251627
13	9.456464	4.678836	1.952906
14	**10.10474**	2.429541	1.801146
15	-	1.482438	-
16	-	1.711231	-

First dataset clustering results for the analysed metric (Kurtosis) by clustering algorithm

Best cluster set selection per dataset

Validation of the results comparing the instances distribution per cluster for each dataset, in percentages

Cluster	Percentage of instances in dataset 1	Percentage of instances in dataset 2	Percentage of instances in dataset 3	Deviation
0	35,3190444021	30,608854556	33,0956923386	4,7101898461
1	2,0102765666	0,0498380264	2,2291394718	2,1793014454
2	0,2436078528	1,603123183	4,0365498544	3,7929420016
3	4,0226002579	5,6483096603	3,7553971282	1,892912532
4	5,5292841205	2,3340809037	2,028316096	3,5009680245
5	4,964277672	0,9303098264	5,6531780299	4,7228682035
6	3,3183893222	11,3298446715	5,07079024	8,0114553493
7	10,0861839546	9,0622144696	10,2018274927	1,1396130231
8	5,1628487789	3,9704294377	3,8457676474	1,3170811315
9	4,1413334971	1,7027992358	4,9603373833	3,2575381475
10	1,8219410837	6,8527286319	4,6992669947	5,0307875482
11	11,9060779136	5,0502533433	4,8197610202	7,0863168934
12	4,8046019366	10,8646897583	11,9891555377	7,1845536011
13	3,1505250875	4,3940526622	2,1387689527	2,2552837095
14	3,5190075539	5,5984716339	1,4760518124	4,1224198214
AVEDEV	4,974840836	4,7179998339	4,7048900492	0,2699507868
MEDIAN	4,1413334971	5,0502533433	4,6992669947	0,9089198462
STDEV	8,4654051627	7,470583131	7,8607872657	0,9948220317

Validation of the results comparing the instances distribution per cluster for each dataset, in percentages

Operation with the global best cluster set

Deviation between the instance distribution percentage

(A)/(C)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0	0	0	0	0	0	0	0	0	0	0	0,01	0	0	0
1	0,07	0	0	0	0,04	0	0	0	0	0	0	0,08	0,01	0,02	0
2	0,03	0	0	0	0,1	0	0	0	0	0	0	0,28	0	0,03	0
3	0,25	0	0	0	0,2	0	0	0	0	0	0	0,06	0,04	0,01	0
4	0,37	0	0	0	0,04	0	0	0	0	0	0	0	0	0,02	0
5	0,05	0	0	0	0,04	0	0	0	0	0	0,02	0	0	0	0
6	0,1	0	0	0,01	0	0	0	0	0	0,01	0	0,06	0	0,02	0
7	0,07	0	0	0,03	0,11	0,01	0	0,03	0,01	0,05	0,07	0,08	0,01	0,03	0,01
8	0,15	0,07	0,12	0,16	0,26	0,12	0,09	0,07	0,06	0,09	0,22	0,22	0,19	0,14	0,18
9	0,69	0,25	0,28	0,5	0,52	0,38	1,06	0,16	0,45	0,34	0,48	0,85	1,02	0,19	0,5
10	1,27	0,28	0,66	0,39	0,41	0,51	0,95	0,27	0,3	0,21	0,56	1,04	1,08	0,08	0,66
11	0,27	0,28	0,28	0,49	0,36	0,62	1,09	0,12	0,27	0,21	0,64	1,12	0,79	0,18	0,86
12	1,16	0,35	0,86	0,2	0,39	0,48	1,21	0,38	0,35	0,44	0,48	1,2	1,19	0,31	0,54
13	0,96	0,13	0,31	0,11	0,27	0,44	0,81	0,14	0,2	0,3	0,49	0,72	0,55	0,17	0,45
14	0,23	0,07	0,16	0,18	0,26	0,32	0,31	0,2	0,26	0,32	0,23	0,22	0,4	0,12	0,01
15	0,71	0,13	0,27	0,17	0,32	0,45	0,62	0,16	0,08	0,25	0,49	0,51	0,72	0,36	0,4
16	1,14	0,17	0,25	0,23	0,1	0,39	0,61	0,27	0,19	0,32	0,63	0,82	0,62	0,26	0,36
17	0,76	0,2	0,26	0,12	0,16	0,4	0,71	0,22	0,27	0,31	0,49	0,7	0,56	0,44	0,31
18	0,25	0,13	0,15	0,1	0,02	0,36	0,37	0,17	0,15	0,16	0,28	0,34	0,13	0,2	0,2
19	0,19	0,04	0,08	0	0,06	0,27	0,12	0,13	0,11	0,2	0,12	0,1	0,18	0,04	0,02
20	0,18	0,2	0,06	0,09	0,06	0,09	0,06	0,06	0,09	0,04	0,08	0,06	0,01	0,05	0,13
21	0,19	0,1	0,02	0,03	0,06	0,08	0,06	0,03	0,04	0,09	0,06	0,05	0,1	0,04	0,1
22	0,2	0,23	0	0,01	0,01	0,17	0,05	0,08	0,2	0,04	0,03	0	0,13	0,02	0,09
23	0,01	0,01	0	0	0	0,01	0	0,02	0	0	0	0	0	0	0
MAX	**1,27**	0,35	0,86	0,5	0,52	0,62	1,21	0,38	0,45	0,44	0,64	1,2	1,19	0,44	0,86

Deviation between the instance distribution percentage among the clusters analysed by the attribute (A) hour, representing the relative probability of a hour attribute value to belong to a certain cluster (C), in percentages

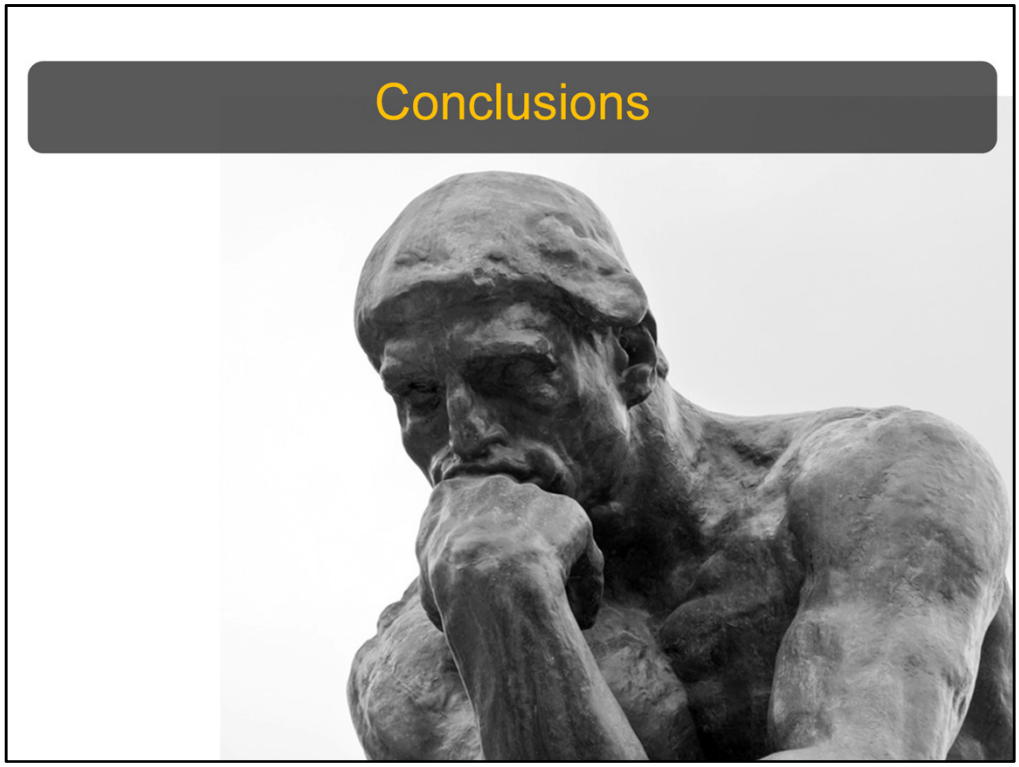
Operation with the global best cluster set

Summary of the calling hour attribute and the clusters which better represented each value

Cluster	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Hour		7	21	12	0		3	2	8	23	6	5	1	4	14
				13	16			9	15			22	20	10	
				17	18			11							
				19											

This is our finally extracted knowledge.

Summary of the calling hour attribute and the clusters which better represented each value



Hemos visto:

Un nuevo modelo de entender los aspectos económicos de Internet desde un punto de vista tecnológico.

Nos hemos centrado en el charging por ser donde se sitúan las operaciones de soporte del negocio.

Hemos visto la relevancia de los datos para poder mejorar servicios.

Hemos visto que aplicando técnicas de minería de datos en general y de clustering en particular, podemos extraer conocimiento para mejorar estos mismos servicios.

Hemos presentado una metodología que busca el modelo que, en la medida de lo posible, mejor represente los datos de un problema dado.

Hemos introducido un caso de uso real, el de la VoIP, para esta metodología.

Lo que para muchos es una leyenda urbana:

A famous story about association rule mining is the "beer and diaper" story.
A purported survey of behavior of supermarket shoppers discovered that customers (presumably young men) who buy diapers tend also to buy beer. This anecdote became popular as an example of how unexpected association rules might be found from everyday data. There are varying opinions as to how much of the story is true.

In 1992, Thomas Blischok, manager of a retail consulting group at Teradata, and his staff prepared an analysis of 1.2 million market baskets from about 25 Osco Drug stores. Database queries were developed to identify affinities. The analysis "did discover that between 5:00 and 7:00 p.m. that consumers bought beer and diapers". Osco managers did NOT exploit the beer and diapers relationship by moving the products closer together on the shelves.

Intellectual Property Rights

Seminario DeustoTech
18 Mayo 2011

Técnicas de minería de datos para la mejora de servicios en Internet

Igor Ruiz-Agundez

DeustoTech, Deusto Institute of Technology, University of Deusto