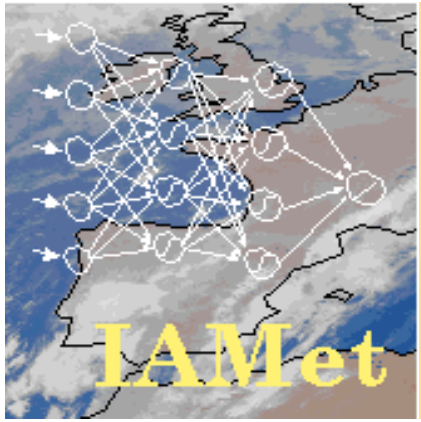


Data Mining.
Extracción de Conocimiento en Grandes Bases de Datos

http://etsiso2.macc.unican.es/~meteo

José M. Gutiérrez
Dpto. de Matemática Aplicada,
Universidad de Cantabria, Santander

http://personales.unican.es/gutierjm

UC UNIVERSIDAD DE CANTABRIA

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

El Mundo de la Información y sus Problemas.

- **Cada vez se genera más información** y se hace más fácil el acceso masivo a la misma (existen gran cantidad de bases de datos on-line)
  - ✓ Transacciones bancarias, Internet y la Web, observaciones científicas (biología, altas energías, etc.) "tranNASA's EOS (Earth Observation System)".
- La **tecnología es barata** y los **sistemas de gestión de bases de datos** son capaces de trabjar con cantidades masivas de datos (Terabytes).

Los datos contienen información útil **"CONOCIMIENTO"** !!!

- Necesitamos **extraer** información (**conocimiento**) de estos datos:
  - ✓**Rapidez y confiabilidad.**
  - ✓**Capacidad de modelización y escalabilidad.**
  - ✓**Explicación e Interpretación de los resultados (visualización, …).**

WalMart captura transacciones de 2900 tiendas en 6 países. Esta información e acumula en una base de datos masiva de 7.5 terabyte. WalMart permite que más de 3500 proveedores accedan a los datos relativos a sus productos para realizar distintos análisis. Así pueden identificar clientes, patrones de compras, etc. En 1995, WalMart computers procesó más de un millón de consultas complejas.

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Datos, Información y Conocimiento.

¿Qué diferencias hay entre información, datos y conocimiento?

✓ Informalmente se utilizan indistintamente, con pequeños matices.

✓ ***información*** y ***datos*** se pueden referir a cualquier cosa, aunque "Datos" suele referir a la "evidencia".

✓ ***Conocimiento*** es subjetivo:

depende de las intenciones (**objetivo del aprendizaje**).

debe ser ***inteligible*** para el que aprende o el que encarga el **aprendizaje** (usuario).

¿Qué es aprendizaje?

✓ (visión genérica, Mitchell 1997) es **mejorar el comportamiento** a partir de la experiencia. Aprendizaje = Inteligencia.

✓ (visión más estática) es la *identificación de* ***patrones*** , de ***regularidades***, existentes en la evidencia.

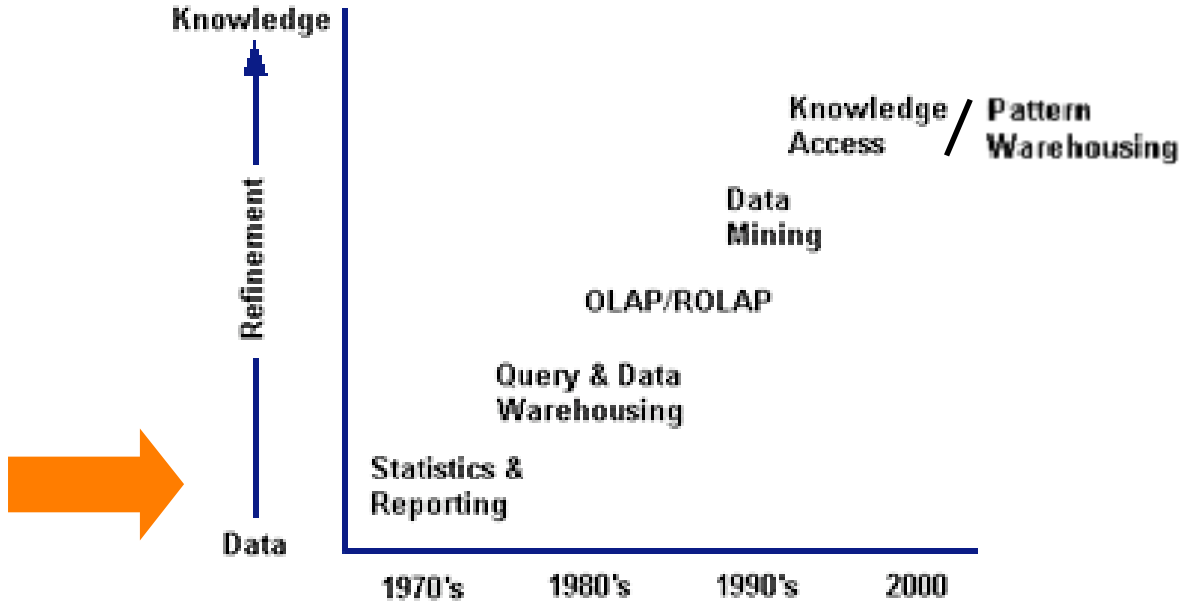
✓ (visión externa) es la ***predicción*** de observaciones futuras con ***plausibilidad***.

✓ (visión teórico- informacional, Solomonoff 1966) es
eliminación de redundancia = compresión de información .

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Acceso a los Datos. Evolución histórica.

La necesidad de almacenar información ha motivado históricamente el desarrollo de sistemas más eficientes, con mayor capacidad y más baratos de almacenamiento.



- ***Bases de datos relacionales.***
- ***DBMS*** (Data Base Management Systems) ***y repositorios de información:***
  - Bases de datos orientadas a objetos y objeto-relacionales.
  - Bases de datos espaciales (geográficas).
  - Bases de datos de texto y multimedia.
  - WWW.

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

OLAP (On-Line Analytical Processing)

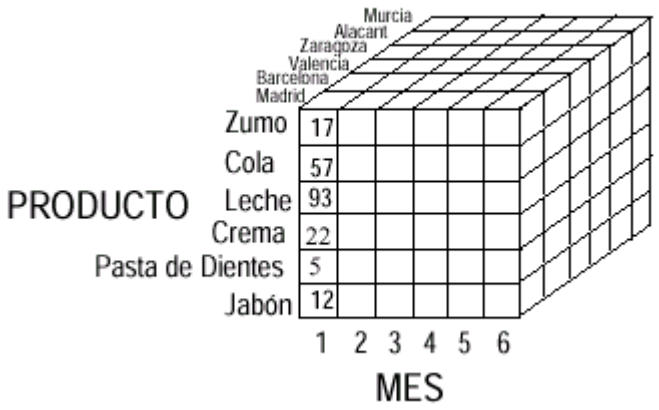
Sobre estas mismas bases de datos de trabajo ya se puede extraer conocimiento (visión tradicional).

- ✓**Se mantiene el trabajo transaccional** diario de los sistemas de información originales (conocido como **OLTP,** *On- Line Transactional Processing*).
- ✓Se hace **análisis de los datos en tiempo real** sobre la misma base de datos(conocido como **OLAP, *On- Line Analytical Processing***),

Según la organización de la información copiada se distingue:

- ✓**ROLAP** (Relational OLAP): el almacén de datos es relacional.
- ✓**MOLAP** (Multidim OLAP): el almacén de datos es una matriz multidimensional.

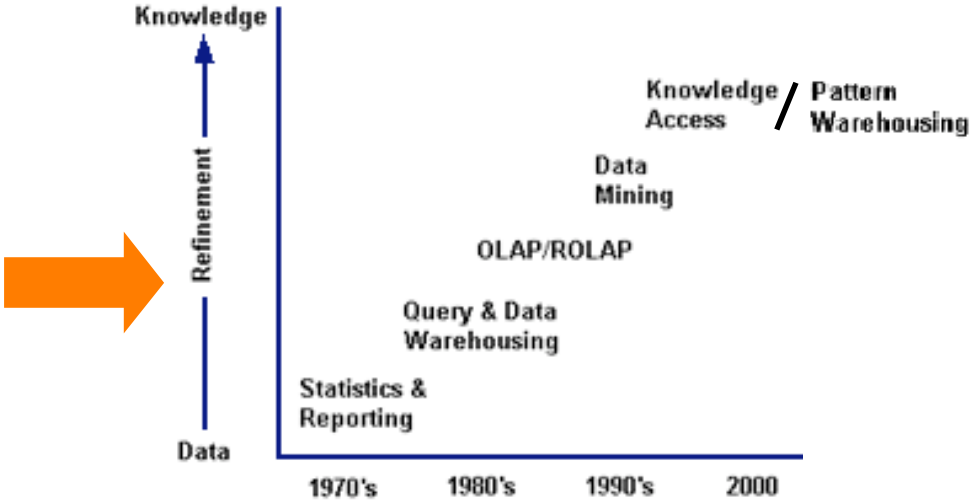
Cada atributo relevante se establece en una dimensión, que se puede agregar o desagregar.



Las dimensiones se agregan:

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Data Warehouses. Génesis.



PROBLEMAS:

✓ Disturba el trabajo transaccional diario de los sistemas de información originales (“ *killer queries* ”). Se debe hacer por la noche o en fines de semana.

✓ La base de datos está diseñada para el trabajo transaccional, no para el análisis de los datos. Generalmente no puede ser en tiempo real (era AP pero no OLAP).

Para poder operar eficientemente con esos datos y debido a que los costes de almacenamiento masivo y conectividad se han reducido drásticamente en los últimos años, parece razonable recoger (copiar) los datos en un sistema unificado.

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Data Warehouses

DATA-WAREHOUSES (Almacenes de Datos): Se separan de los datos a analizar con respecto a sus fuentes transaccionales (se copia/ almacena toda la información histórica).

Existe toda una tecnología creciente de cómo organizarlos y sobretodo de cómo tenerlos actualizados (cargas periódicas) respecto a los datos originales

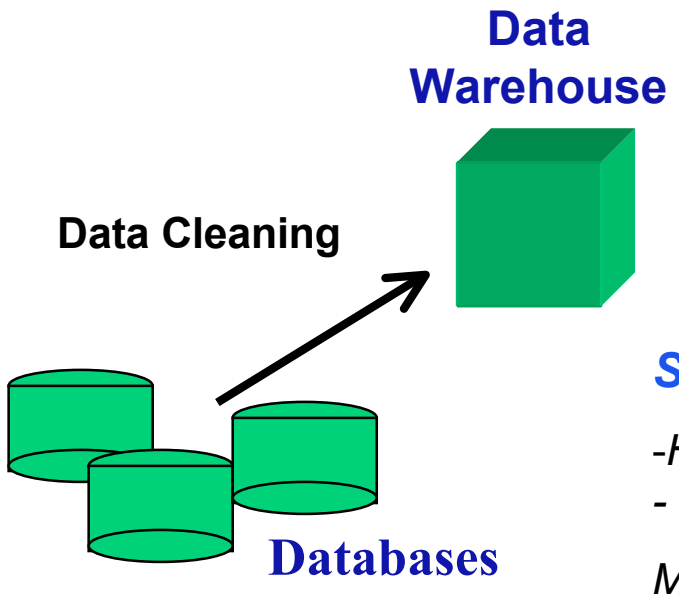
VENTAJAS:

- ✓Facilita el análisis de los datos en tiempo real (**OLAP**),
- ✓No disturba el **OLTP** de las bases de datos originales.

A partir de ahora diferenciaremos entre bases de datos para OLTP (tradicional) y almacenes de datos (KDD sobre data warehouses).

	OLTP	Data Warehouse
Purpose	Run day-to-day operations	Information retrieval and analysis
Structure	RDBMS	RDBMS
Data Model	Normalised	Multi-dimensional
Access	SQL	SQL plus data analysis extensions
Type of Data	Data that runs the business	Data that analyses the business
Condition of Data	Changing, incomplete	Historical, descriptive

Construcción de un Data Warehouse



Limpieza y criba selección de datos:

Se deben elmininar el mayor número posible de datos erróneos o ***inconsistentes*** *(limpieza) e* ***irrelevantes*** *(criba).*

Se aplican métodos estadísticos:

-Histogramas (detección de datos anómalos).
- Redefinición de atributos (agrupación o separación).

Muy relacionado con la disciplina de "Calidad de Datos".

Acciones ante datos anómalos (outliers):

- **Ignorar**: algunos algoritmos son robustos a datos anómalos.
- **Filtrar** (eliminar o reemplazar) la columna: solución extrema.
- **Discretizar**: transformar un valor continuo en uno discreto (p. ej. muy alto, alto, etc.) hace que los outliers caigan en 'muy alto' o 'muy bajo' sin mayores problemas.

Acciones ante datos faltantes (missing values):

- **Ignorar**: algunos algoritmos son robustos a datos faltantes.
- **Filtrar** (eliminar o reemplazar) la columna
- **Reemplazar** el valor: por medias. A veces se puede *predecir* a partir de otros datos, utilizando cualquier técnica de ML.

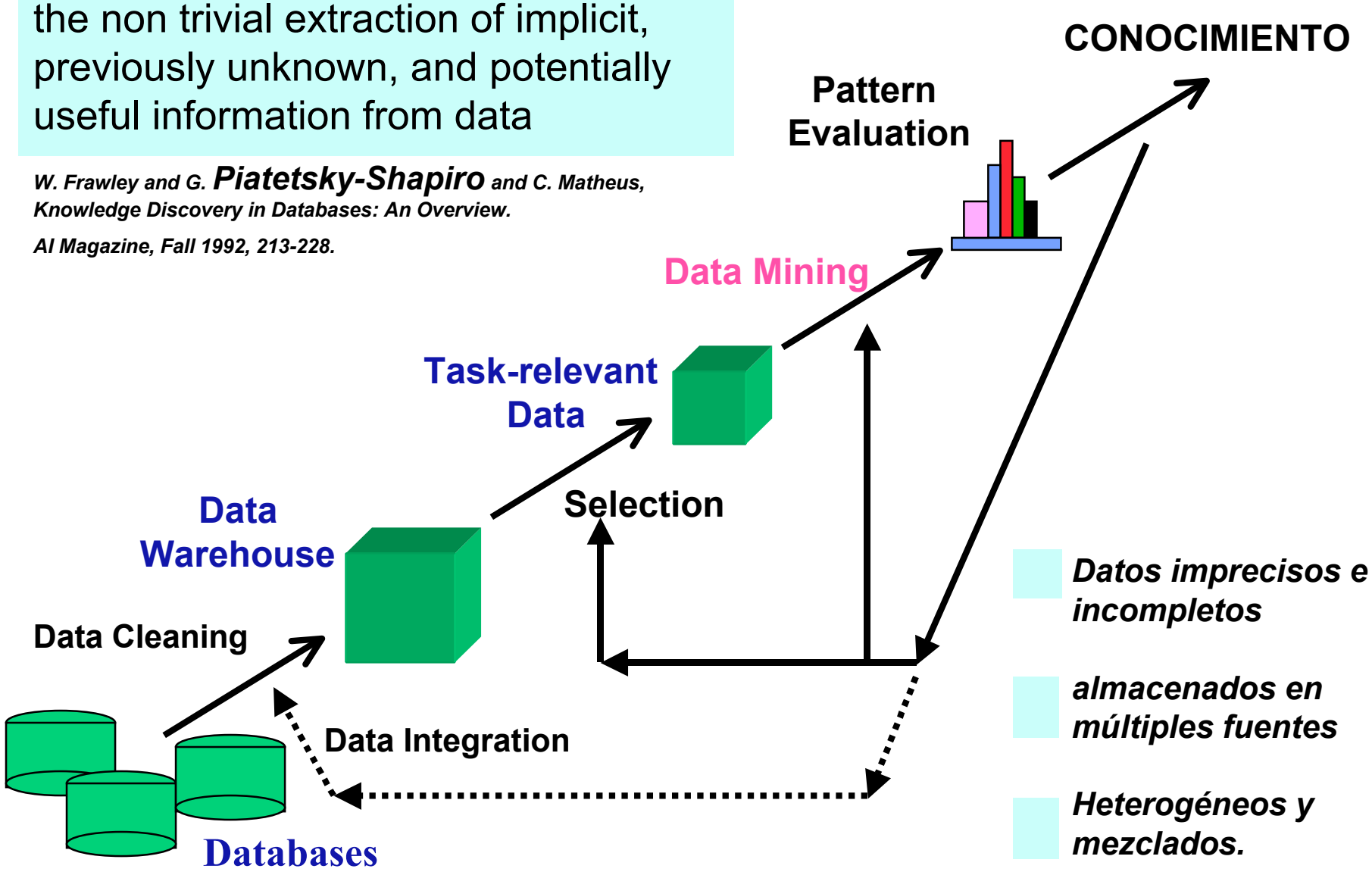
José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

¿Qué es Data Mining (minería de datos)?

the non trivial extraction of implicit, previously unknown, and potentially useful information from data

*W. Frawley and G. **Piatetsky-Shapiro** and C. Matheus, Knowledge Discovery in Databases: An Overview.*

AI Magazine, Fall 1992, 213-228.



Datos imprecisos e incompletos

almacenados en múltiples fuentes

Heterogéneos y mezclados.

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Diferencias entre DBMS y Data Mining

En los sistemas estándar de gestión de bases de datos las consultas se resuelven accediendo a distintos conjuntos de datos almacenados:

- ✓*Ventas del último mes de un producto.*
- ✓*Ventas agrupadas por la edad del comprador.*

Los sistemas de data mining infieren conocimiento de la base de datos en forma de estructuras y patrones. Este conocimiento supone un nuevo conjunto de información en base a la cual se responden las consultas:

- ✓por qué es tan rentable la división Iberoamericana de Telefónica?
- ✓¿qué clientes son potenciales compradores de un producto?
- ✓¿cuál será el beneficio de la compañía el mes próximo?

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

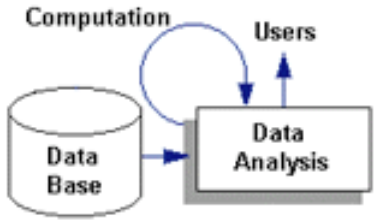
Acceso a Datos vs. Acceso a Conocir

Information DISCOVERY INC.
http://www.datamining.com

Paradigma de Acceso a Datos:

El usuario solicita datos y procesa los datos recibidos en busca de "conocmiento".

SQL + **algoritmos de data mining**.

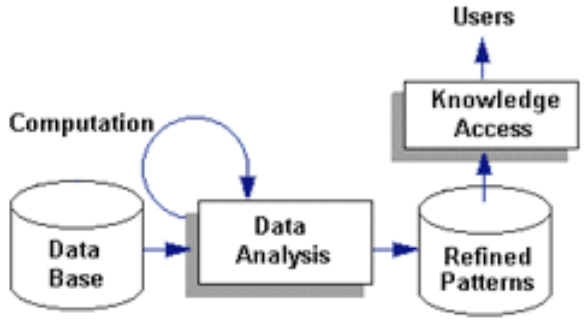


The Data Analysis Approach

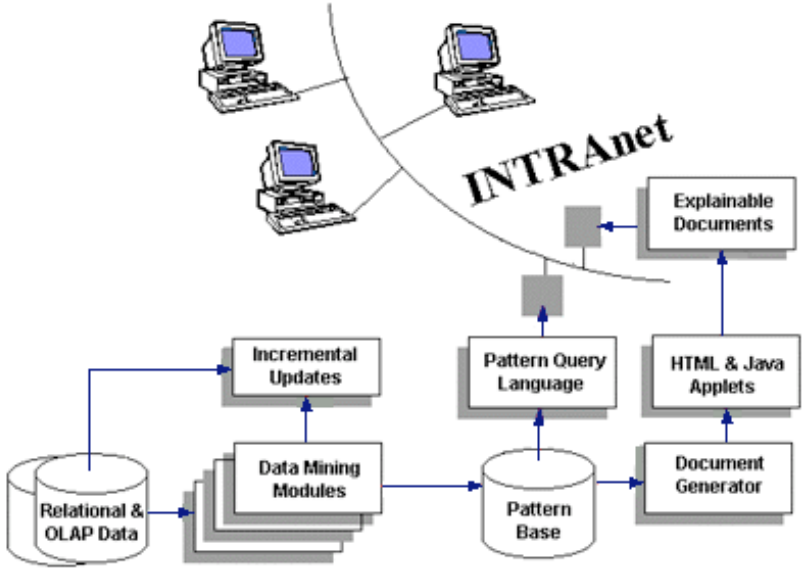
Paradigma de Acceso a Conocimiento:

El sistema genera automáticamente patrones de conocimiento refinados y el usuario accede directamente a los mismos.

PQL = Pattern Query Languaje



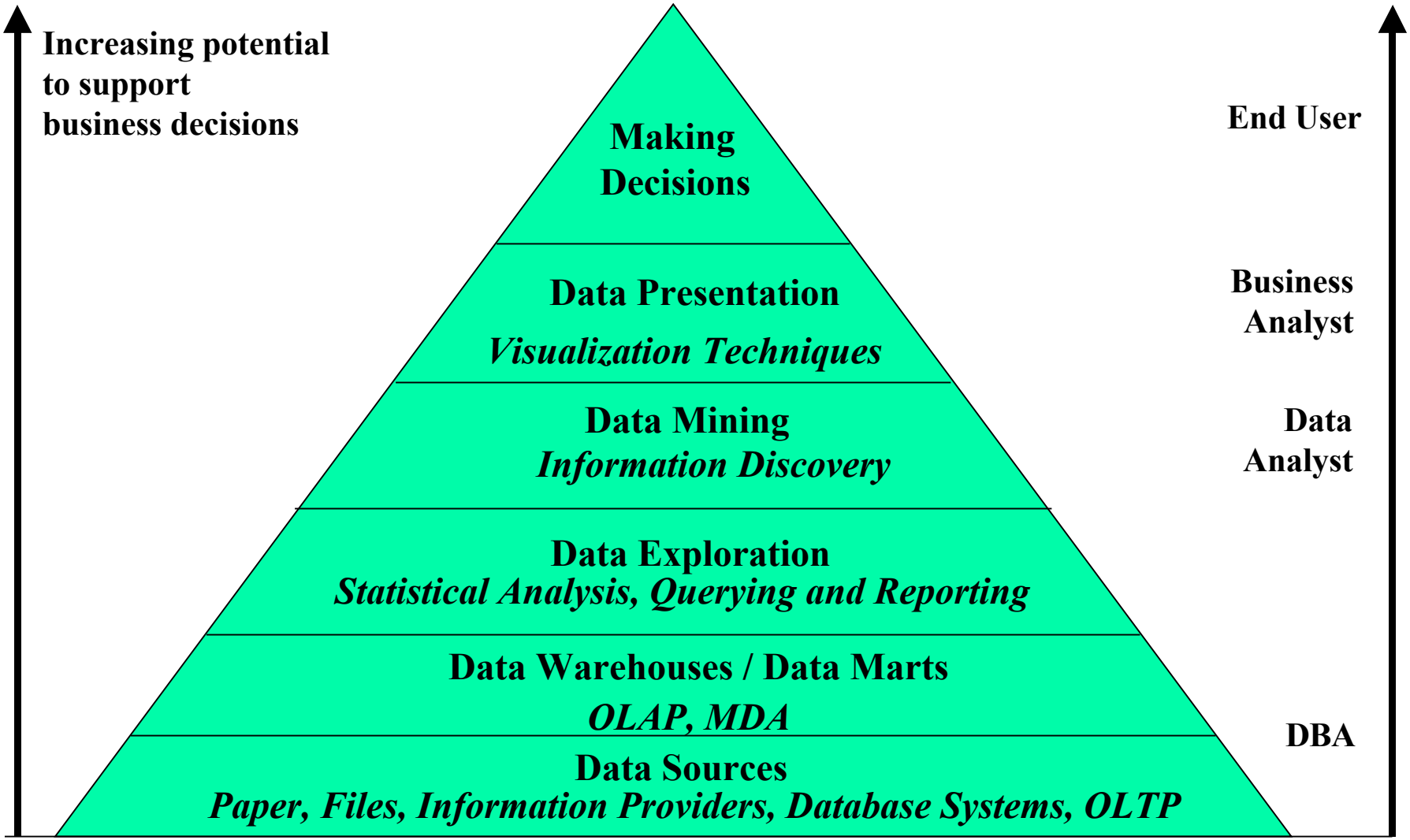
The Knowledge Access Approach



PQL was designed to access patterns just as **SQL was designed to access data**. **PQL resembles SQL**, works atop existing SQL engines.
Information Discovery uses a *Pattern Warehouse*^{TM} of refined information and PQL works on patterns just as SQL works on a datawarehouse. While SQL relies on the relational algebra, **PQL uses the "pattern algebra"**. PQL allows pattern-based queries just as SQL allows data-based queries. And, PQL uses SQL as part of its operation, i.e. **PQL queries are decomposed into a set of related SQL queries**, the Pattern Warehouse is accessed with these queries and the results are re-combined for display to the user. The user accesses these patterns using a web browser.

Data Mining and Business Intelligence

Jiawei Han
Intelligent Database System Research Lab
http://www.cs.sfu.ca/~han

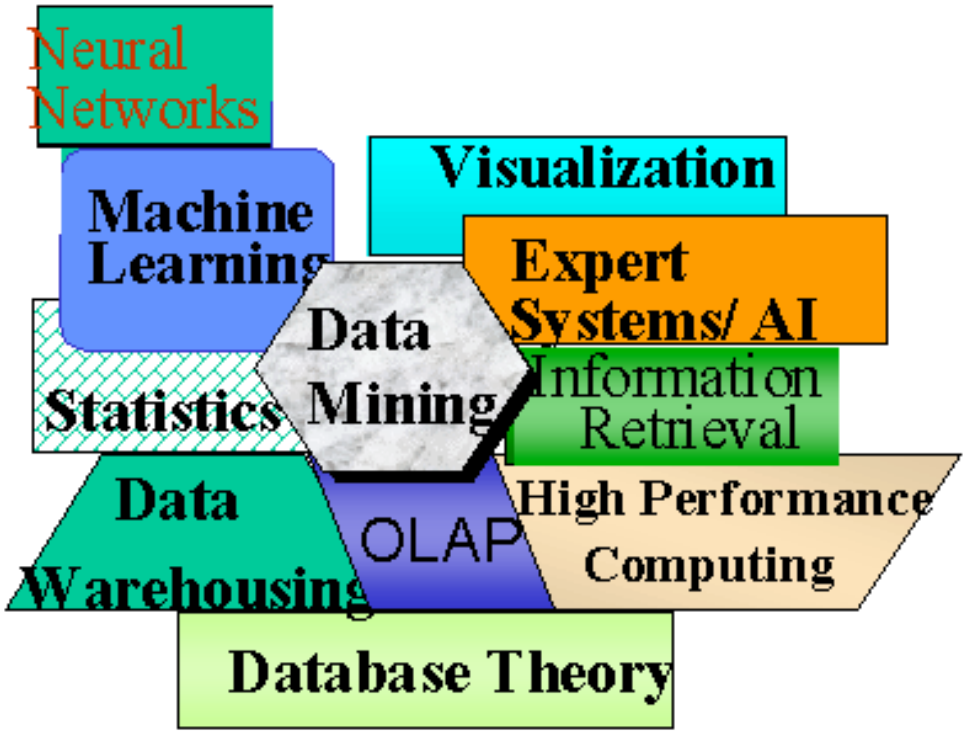


José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Multidisciplinar. Areas y Técnicas Involucradas

variety of techniques to identify nuggets of information or decision-making knowledge in bodies of data, and extracting these in such a way that they can be put to use in the areas such as decision support, prediction, forecasting and estimation. The data is often voluminous, but as it stands of low value as no direct use can be made of it; it is the hidden information in the data that is useful.

Areas Involucradas



- ✓ ***Componentes Principales:***
  compresión de la información.
- ✓ ***Componentes Independientes:***
  extracción de características.
- ✓ ***Modelado de Dependencias:***
  hallar asociaciones entre variables.
  redes Bayesianas
- ✓ ***Agrupación:***
  hallar grupos de elementos.
- ✓ ***Clasificación:***
  asignar elementos a clases.
- ✓ ***Predicción:***
  estimación de valores.
- ✓ ***Visualización:***
  representación gráfica.

Redes Neuronales

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Estadística y Ciencias de la Computación

- **Estadística**
  - *1970: EDA, estimación Bayesiana, modelos flexibles, EM, etc*
  - *Conciencia sobre el papel de la computación en el análisis de datos.*
- **Reconocimiento de Patrones e Inteligencia Artificial**
  - *Atención dirigiga a problemas de percepción (e.g., habla, visión)*
  - *1960: división en técnicas estadísticas y no estadísticas (gramáticas, etc.)*
  - *Convergencia de estadística aplicada e ingeniería (análisis imágenes, Geman)*
- **Aprendizaje Automático y Redes Neuronales**
  - *1980: fracaso de las téncias de aprendizaje no estadísticas*
  - *Aparición de modelos flexibles (árboles, redes)*
  - *Convergencia de estadística aplicada y aprendizaje*
    - *e.g., trabajos de Friedman, Spiegelhalter, Jordan, Hinton*

IA / Apredizaje Automático

Extracción automática de conocimiento

1989 KDD workshop

ACM SIGKDD 2000

Bases de Datos

Bases de datos masivas

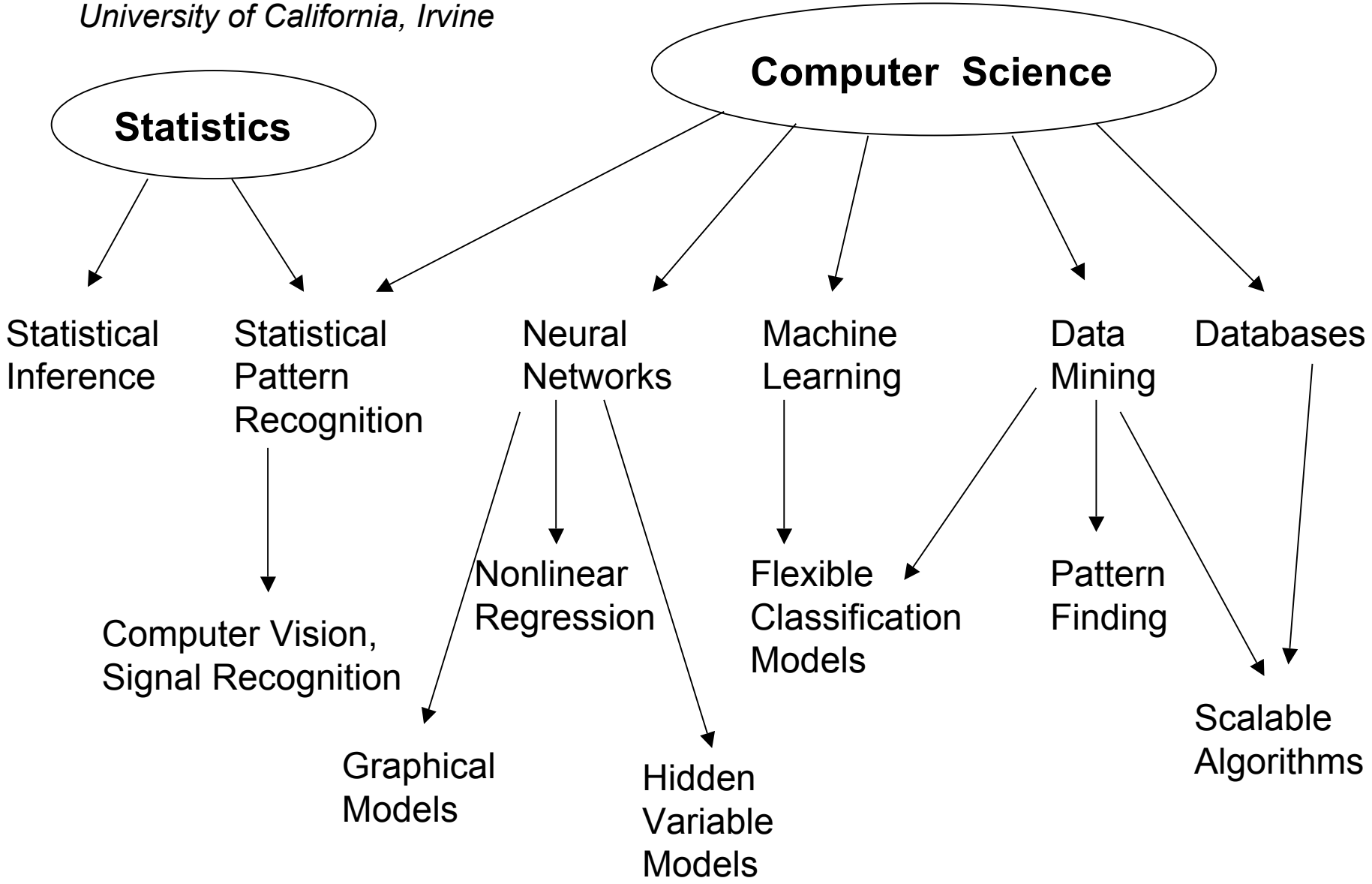
Reglas de asociación

Algoritmos escalables

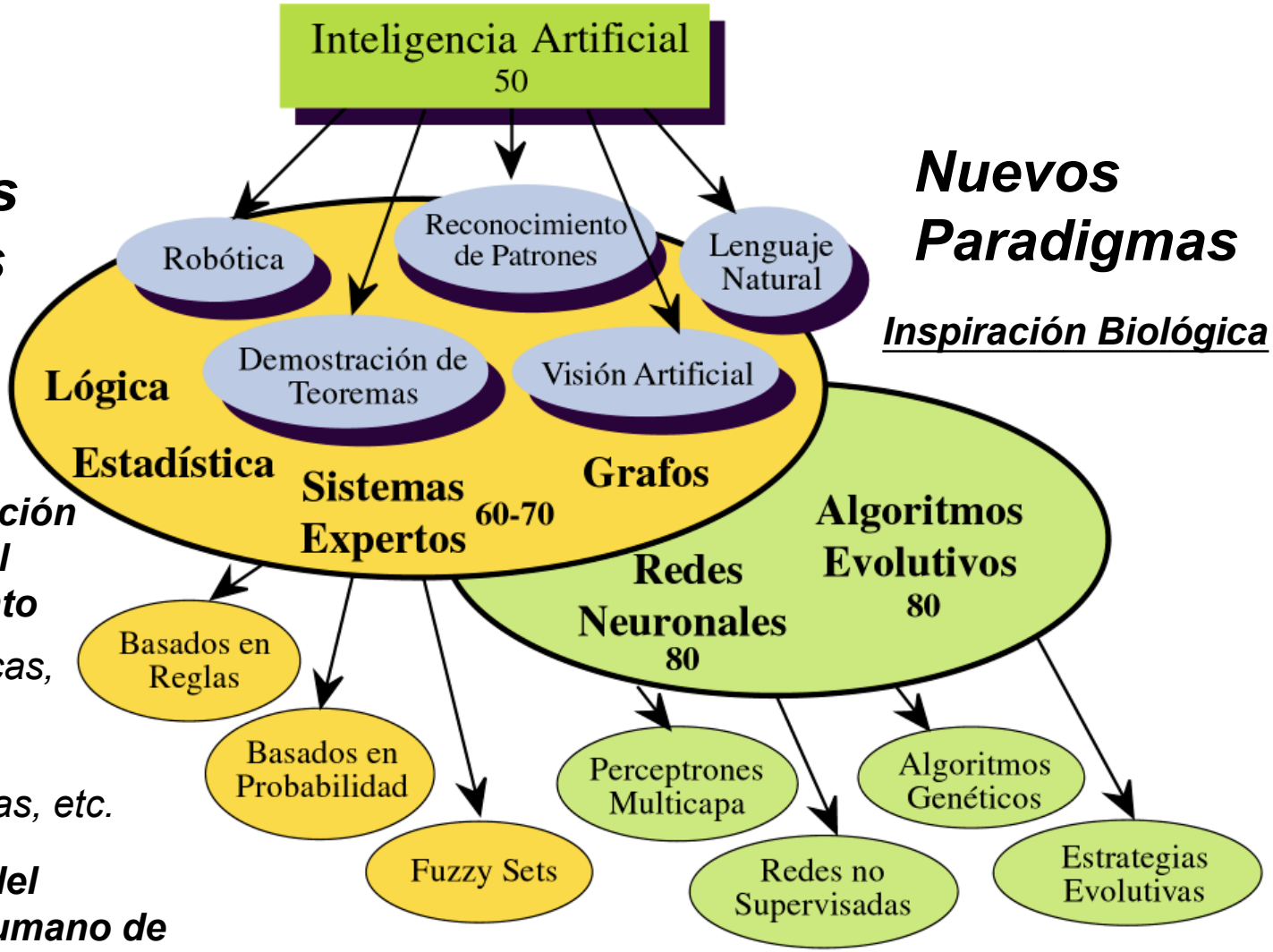
MINERIA DE DATOS

Focus Areas

Padhraic Smyth. Information and Computer Science
University of California, Irvine



Inteligencia Artificial



Representación explícita del conocimiento

sentencias lógicas,
reglas,
grafos,
redes semánticas, etc.

Imitación del proceso humano de razonamiento

Inferencia lógica,
búsqueda en grafos

Procesamiento en serie de la información

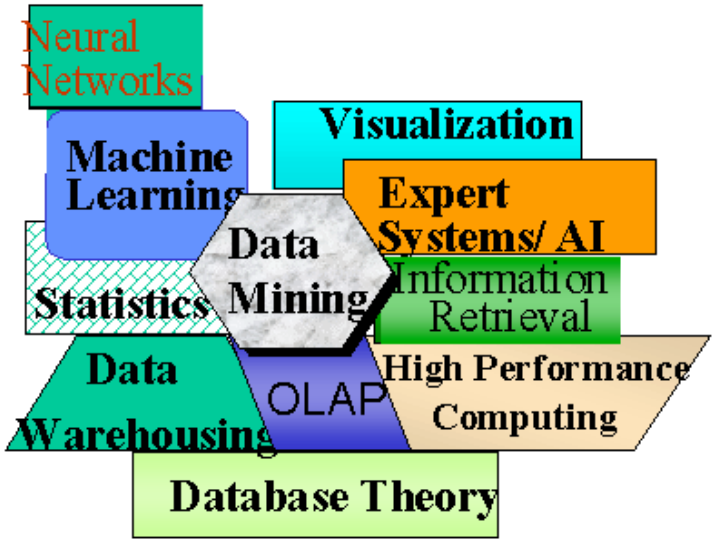
José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

Areas y Técnicas Involucradas

variety of techniques to identify *nuggets of information or decision-making knowledge in bodies of data, and* ***extracting*** *these in such a way that they can be put to use in the areas such as decision support, prediction, forecasting and estimation. The data is often voluminous, but as it stands of low value as no direct use can be made of it;* ***it is the hidden information in the data that is useful****.*

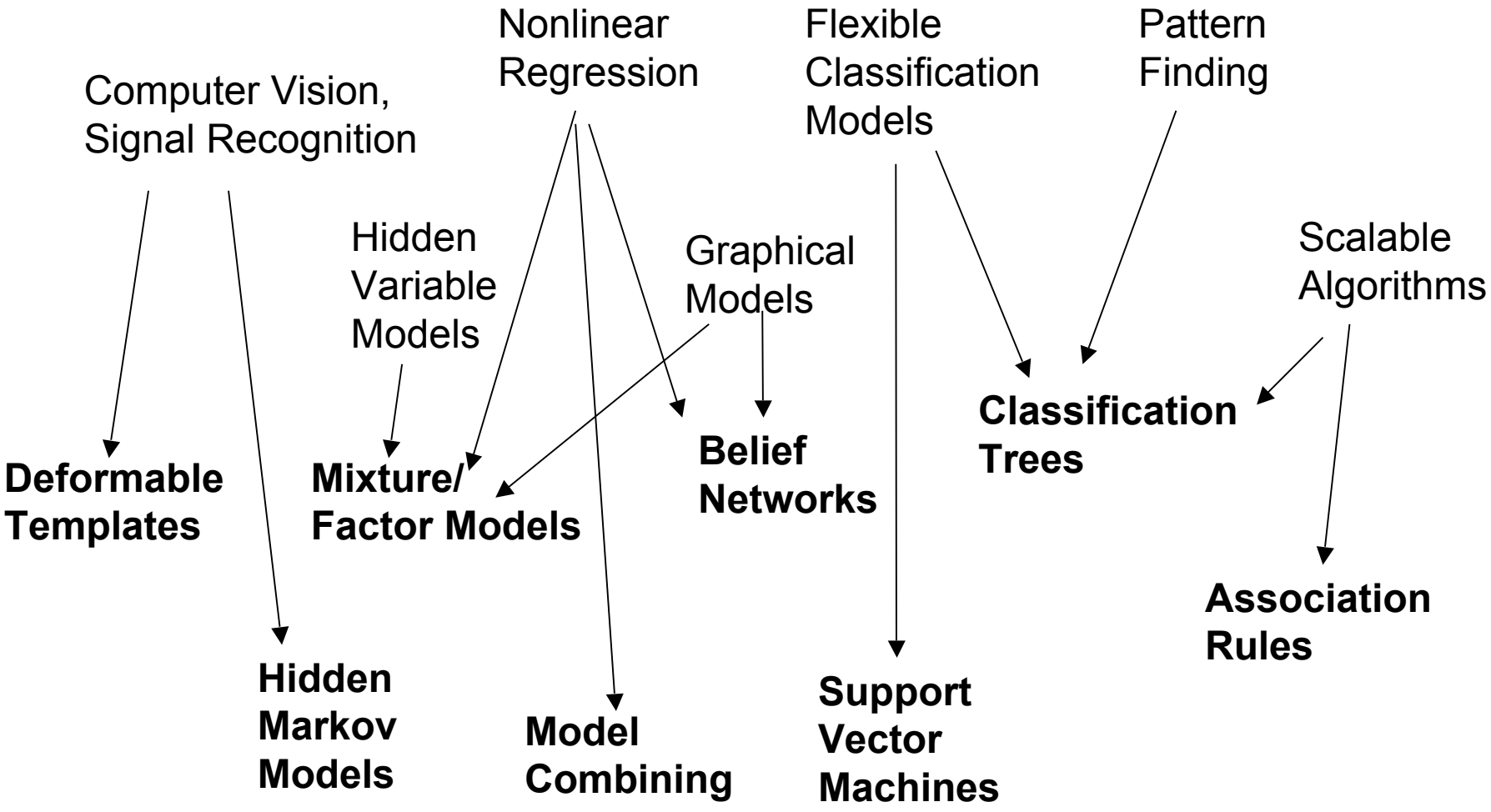
Técnicas Involucradas



- ✓ ***Modelado de Dependencias:***
  asociaciones entre variables.
  reglas y grafos *(redes Bayesianas).*
- ✓ ***Componentes Principales:***
  compresión *de la información.*
- ✓ ***Componentes Independientes:***
  extracción de ***características****.*
- ✓ ***Agrupación:***
  hallar grupos *de elementos.*
- ✓ ***Clasificación:***
  asignar elementos a clases.
- ✓ ***Predicción:***
  estimación *de valores.*
- ✓ ***Visualización:***
  representación gráfica.

Hot Topics (Statistics and Machine Learning)

Padhraic Smyth
Information and Computer Science
University of California, Irvine



Objetivos. Un Primer Ejemplo

Asociación:

Interesa obtener automáticamente reglas que relacionen unos atributos de la base de datos con otros, en base a alguna asociación:

Ejemplo - Base de datos de clientes de un banco.

Regla de Asociación:

if STATUS = married and INCOME > 10000 and HOUSE_OWNER = yes

then INVESTMENT_TYPE = good

Clasificación:

Un sistema de minería de datos aprende de los datos cómo particionar o calsificar los mismos en base a reglas de clasificación:

Ejemplo - Base de datos de clientes de un banco.

Pregunta - ***Un cliente que solicita un préstamo, es una buena inversión?***

Regla típica formulada:

if STATUS = married and INCOME > 10000 and HOUSE_OWNER = yes

then INVESTMENT_TYPE = good

José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

Aplicaciones de la Minería de Datos.

Gran cantidad de información (financiera, servicios, empresas, universidades, libros y hobbies), con complejas interrelaciones.
El 99% de la información no le interesa al 99% de la gente.

Ambiente dinámico

En Internet

- ✓***E-bussines.*** ***Perfiles de clientes, publicidad dirigida, fraude.***
- ✓***Buscadores "inteligentes".*** ***Generación de jerarquías, bases de conocimiento web.***
- ✓***Gestión del tráfico de la red.*** ***Control de eficiencia y errores.***

- ***Reglas de asociación:***
  El 60% de las personas que esquían viajan frecuentemente a Europa.
- ***Clasificación:***
  Personas menores de 40 años y salario superior a 2000$ compran on-line frecuentemente.
- ***Clustering:***
  Los usuarios A y B tienen gustos parecidos (acceden URLs similares).
- ***Detección de "outliers"***
  El usuario A navega en Internet más del doble del tiempo promedio.

La publicidad en Internet es uno de los tópicos más actuales de Data Mining.

Los data warehouse de las empresas contienen enormes cantidades de información sobre sus clientes y gestiones.

El Mundo de los Negocios

- ✓***Banca.*** ***Grupos de clientes, préstamos, oferta de productos.***
- ✓***Compañías de seguros.*** ***Detección de fraude, administración de recursos.***
- ✓***Marketing.*** ***Publicidad dirigida, estudios de competencia.***

José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

La cantidad de información generada en proyectos científicos ha sido enorme: Genoma Humano, datos geofísicos, altas energías, etc.

En Biología, Meteorología, etc.

- ✓ ***Bio-Informática.*** *Búsqueda de patrones en ADN, consultas inteligentes.*
- ✓ ***Meteorología.*** *Teleconexiones (asociaciones espaciales), predicción.*
- ✓ ***Física (altas energías).*** *Datos de colisiones de partículas (búsqueda de patrones).*

crossgrid

WP4: TESTBED

http://www.ifca.unican.es/crossgrid/

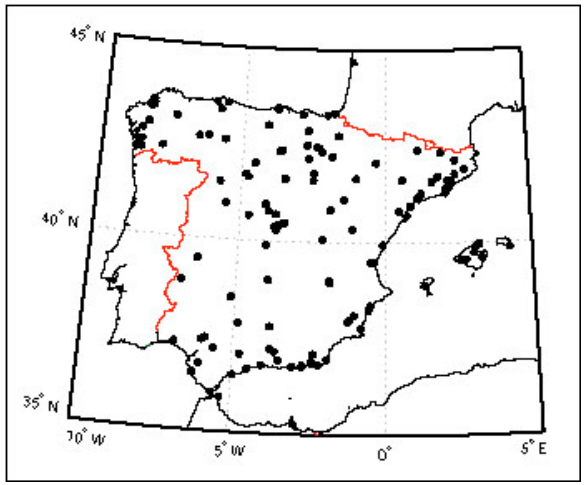
EJEMPLO !!!!!!!!!!!!!!!!!

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

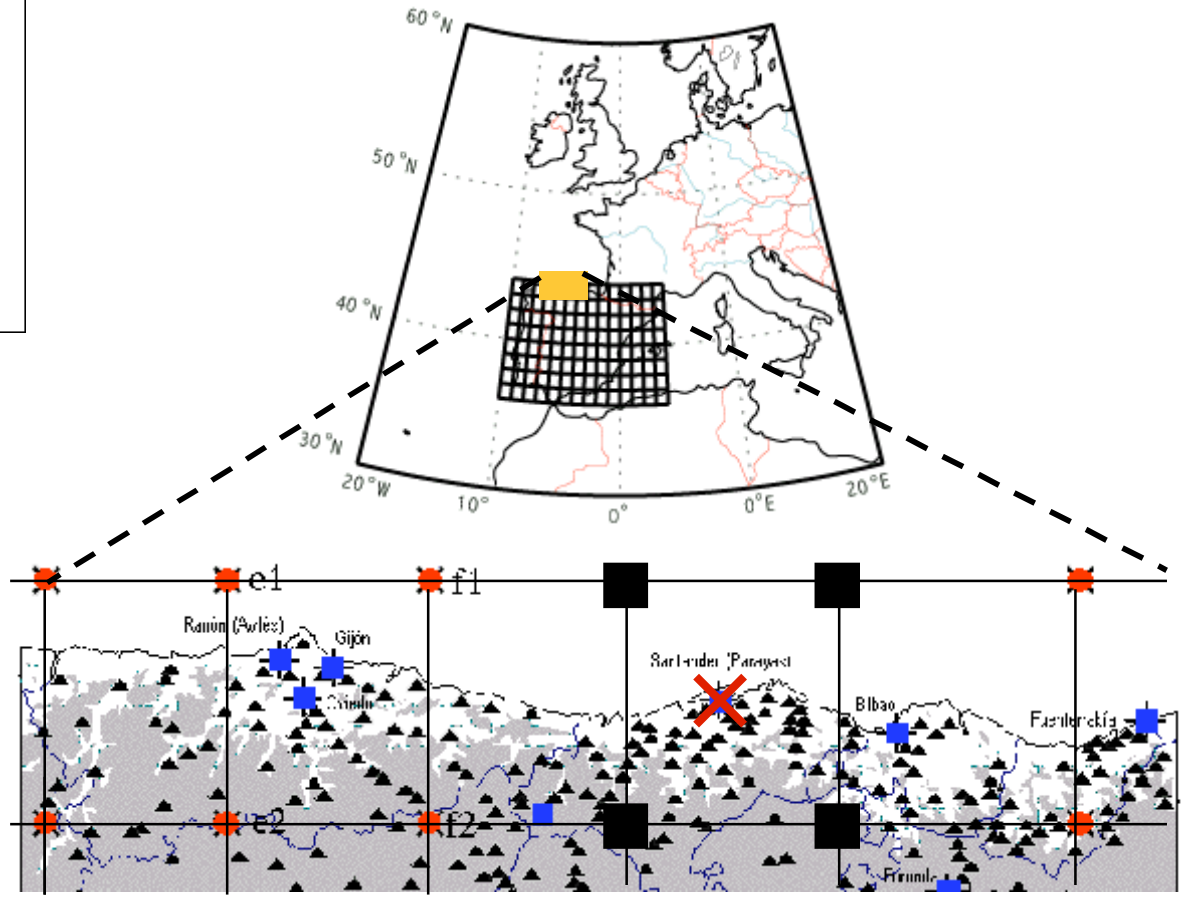
Ejemplo. Meteorología.

✓***Meteorología.*** ***Teleconexiones (asociaciones espaciales), predicción.***

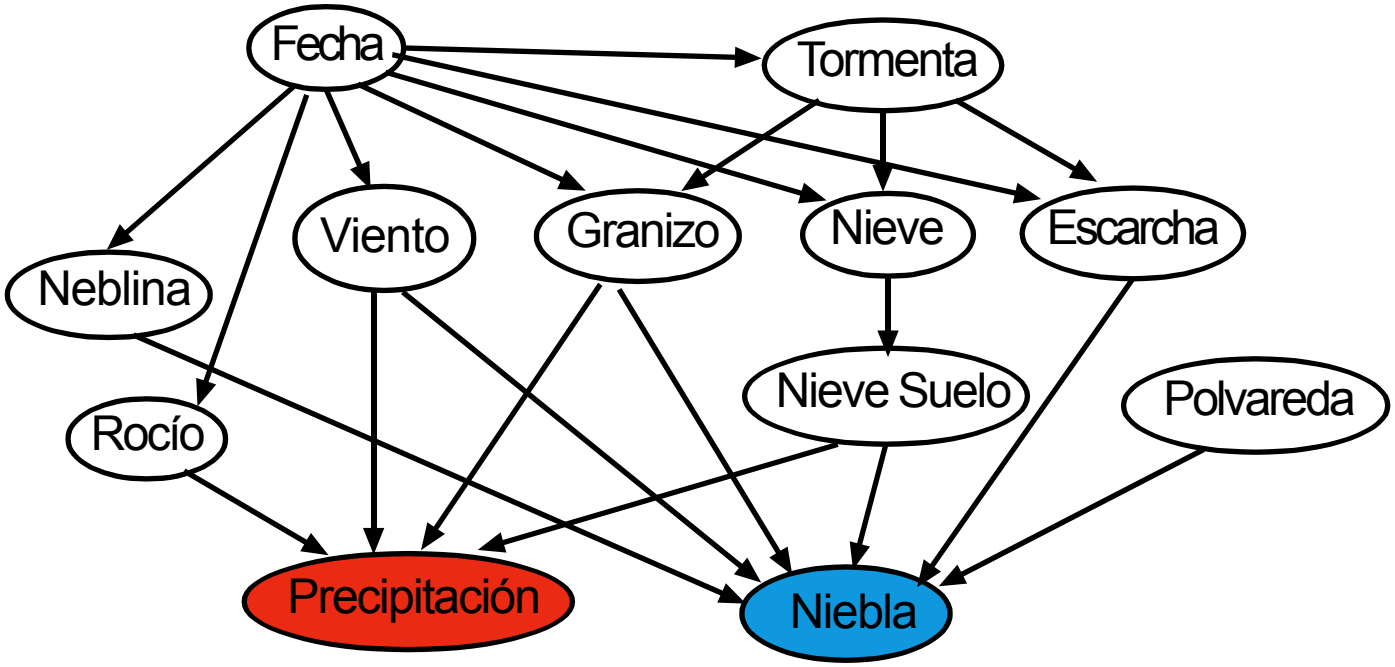
Existen bases de datos con simulaciones de los campos atmosféricos en rejillas dadas.



Se dispone de gran cantidad de información en observatorios locales: Precipitación, temperatura, Viento, etc.



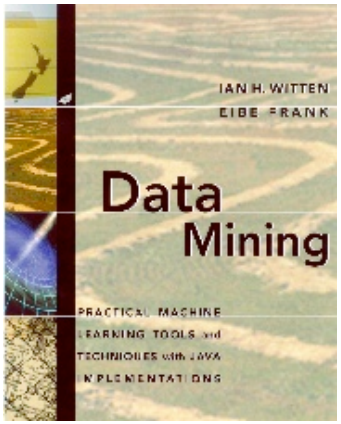
José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm



8601015000000000010
8601021000000000010
8601035001000000010
8601045000000000010
8601051011000000010
8601061011000000010
8601073001000000010
8601085000000000010
8601095000000001000
8601100000001001100
8601110010000000000

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Libros y Material de Consulta

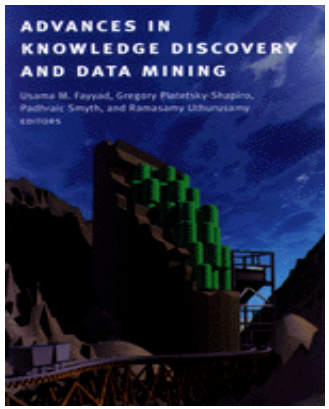


Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations

Ian H. Witten, Eibe Frank

Machine Learning and Data Mining
Open Soure Tools in Java

`http://www.cs.waikato.ac.nz/~ml/weka/`



Advances in Knowledge Discovery and Data Mining

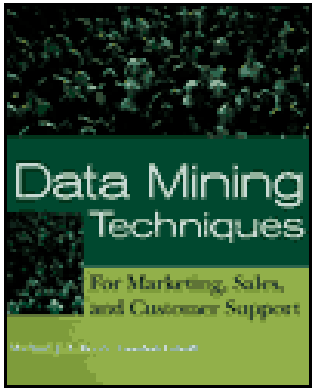
Edited by U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy

The AAAI Press

Data Mining Techniques: For Marketing, Sales, and Customer Support

By Michael J. Berry,Gordon Linoff

Wiley, John & Sons,

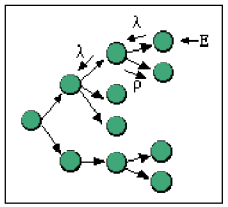


`http://www1.fatbrain.com/FindItNow/Services/home.cl?from=cbs169&store=1`

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Libros disponibles en Internet

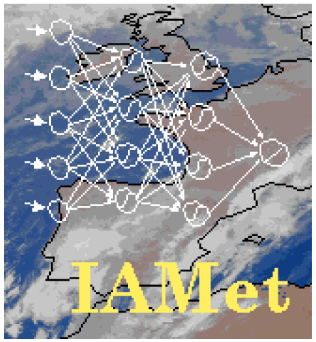
Artificial Intelligence Research Group



Bayesian and Neural-Based Networks

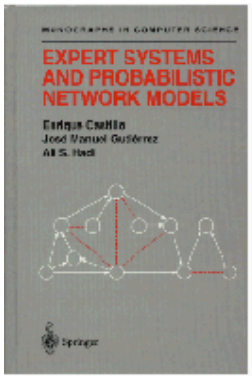
CORNELL UNIVERSITY

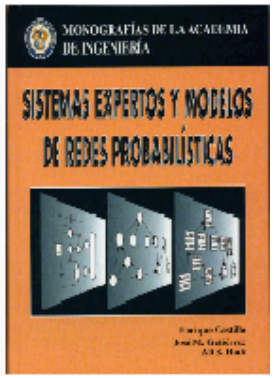
UC Cantabria University



CSIC

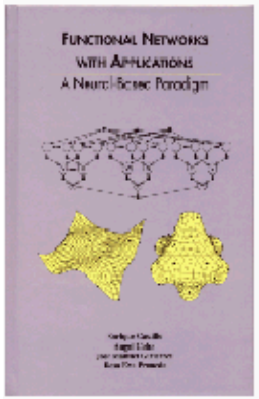
UC





Expert Systems and Probabilistic Network Models.
E. Castillo, J.M. Gutiérrez, y A.S. Hadi
Springer-Verlag, New York.

Monografías de la Academia Española de Ingeniería



An Introduction to Functional Networks
E. Castillo, A. Cobo, J.M. Gutiérrez and E. Pruneda
Kluwer Academic Publishers (1999).

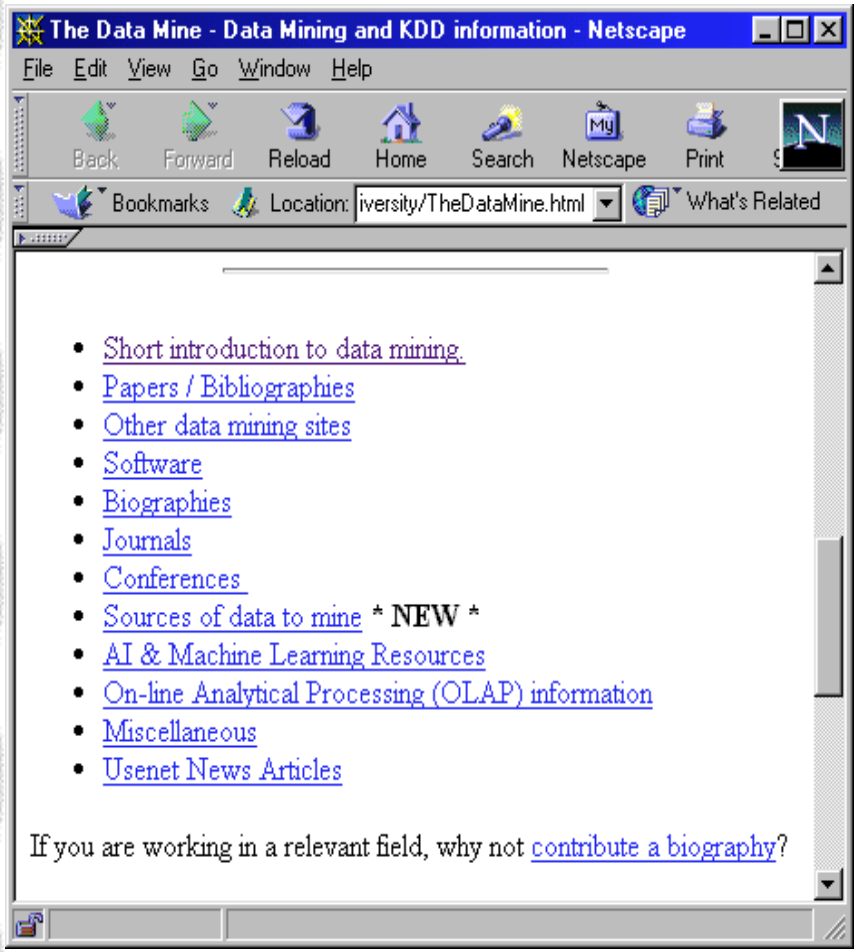
Paraninfo/International Thomson Publishing

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Enlaces Interesantes y Revistas

The Data Mine provides information about Data Mining and Knowledge Discovery in Databases (KDD).

`http://www.cs.bham.ac.uk/~anp/TheDataMine.html`



`http://www.data-miners.com/`

`http://www.kdcentral.com/Software/Data_Mining/`

`http://www.andypryke.com/university/software.html`

`http://www.galaxy.gmu.edu/stats/syllabi/DMLIST.html`

Journals

Data Mining and Knowledge Discovery.
`http://www.wkap.nl/journalhome.htm/1384-5810`

Intelligent Data Analysis
`http://www.iospress.nl/site/html/1088467x.html`

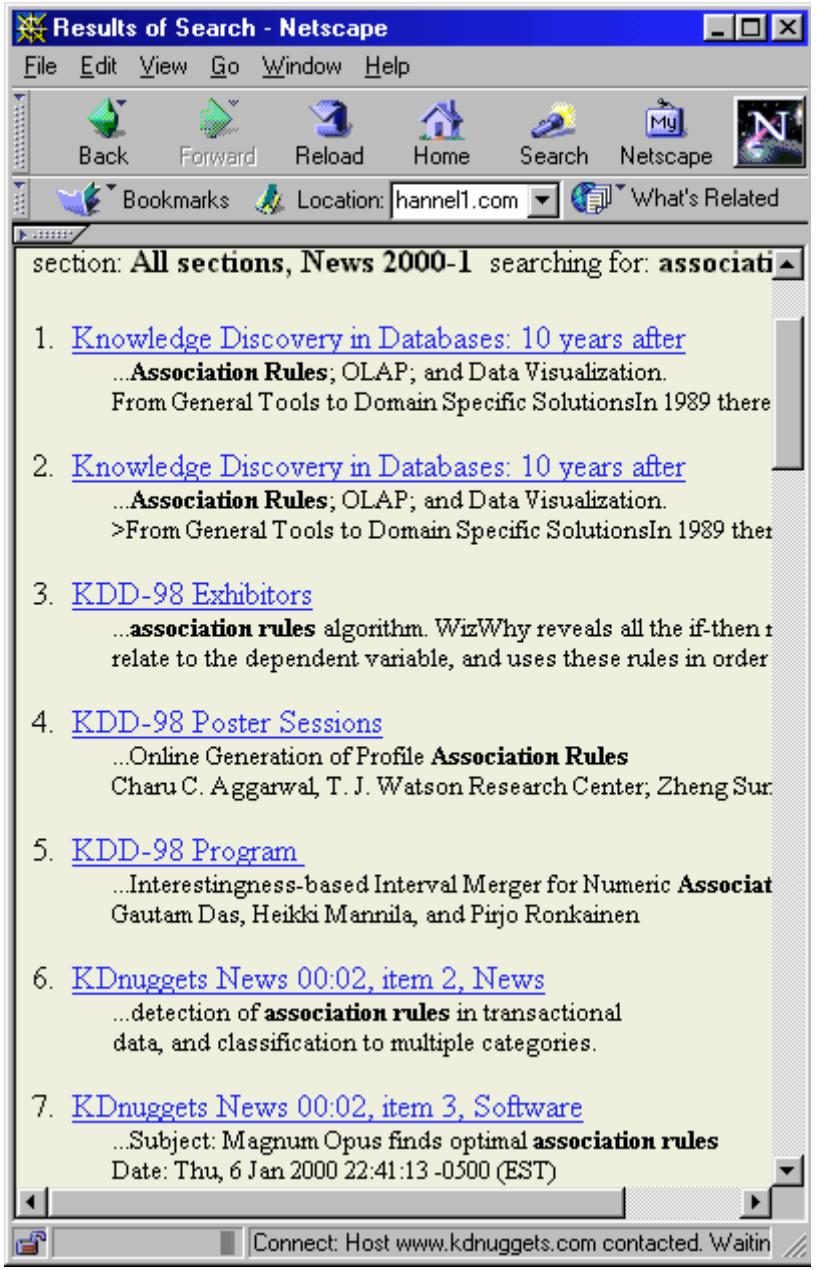
IEEE Trans. on Knowledge and Data Engineering
`http://www.iospress.nl/site/html/1088467x.html`

Related Journals (from IDA)
`http://www.ida-society.org/journals.html`

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

El Portal KDnuggets: http://www.kdnuggets.com/

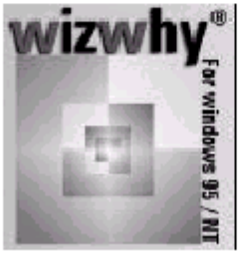
Portal dedicado a Data Mining, Web Mining y Búsqueda de Conocimiento.



José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

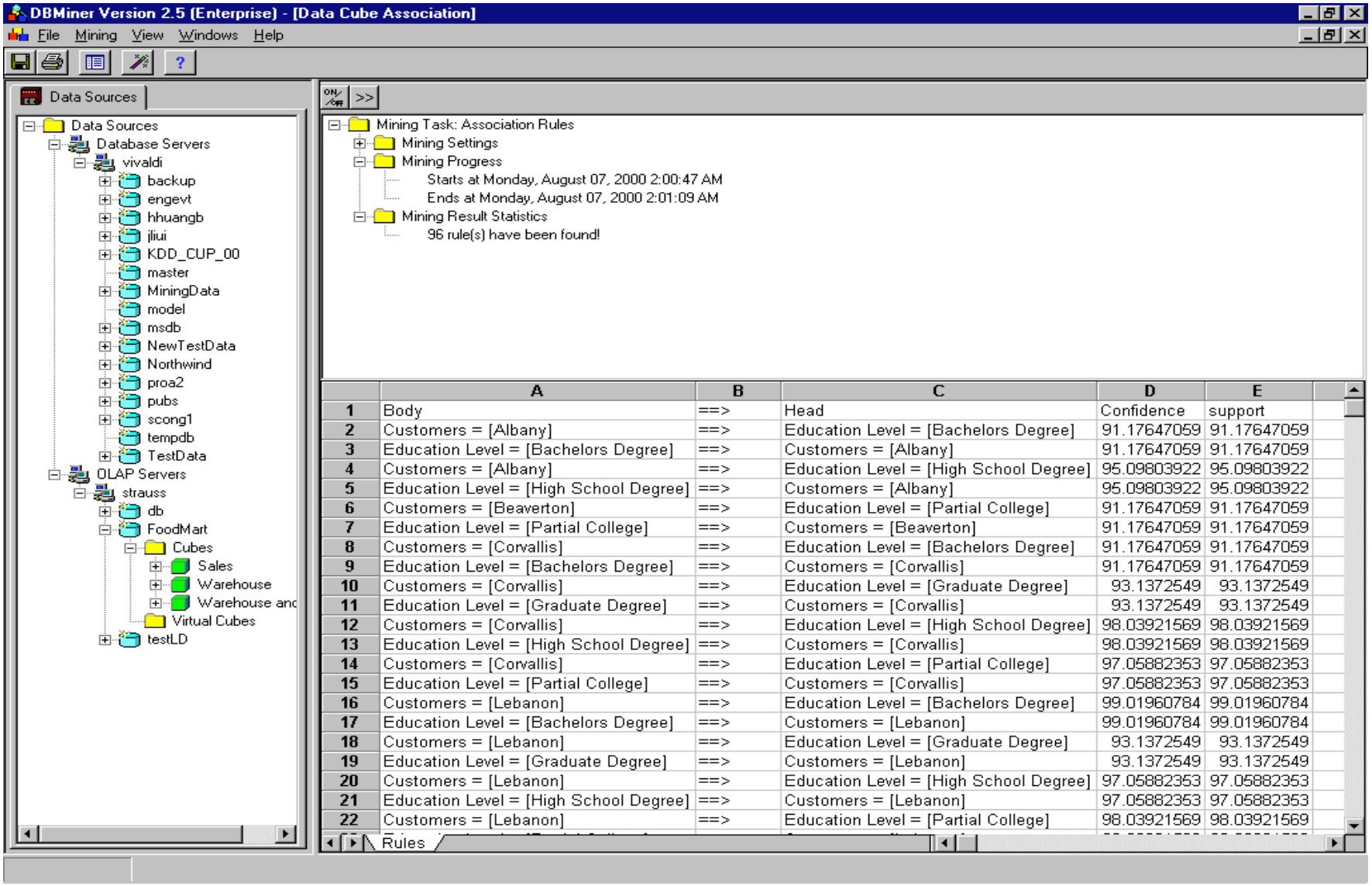
Productos Comerciales

Model 1



José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

Un Ejemplo: DBMiner. http://www.dbminer.com



	A	B	C	D	E
1	Body	==>	Head	Confidence	support
2	Customers = [Albany]	==>	Education Level = [Bachelors Degree]	91.17647059	91.17647059
3	Education Level = [Bachelors Degree]	==>	Customers = [Albany]	91.17647059	91.17647059
4	Customers = [Albany]	==>	Education Level = [High School Degree]	95.09803922	95.09803922
5	Education Level = [High School Degree]	==>	Customers = [Albany]	95.09803922	95.09803922
6	Customers = [Beaverton]	==>	Education Level = [Partial College]	91.17647059	91.17647059
7	Education Level = [Partial College]	==>	Customers = [Beaverton]	91.17647059	91.17647059
8	Customers = [Corvallis]	==>	Education Level = [Bachelors Degree]	91.17647059	91.17647059
9	Education Level = [Bachelors Degree]	==>	Customers = [Corvallis]	91.17647059	91.17647059
10	Customers = [Corvallis]	==>	Education Level = [Graduate Degree]	93.1372549	93.1372549
11	Education Level = [Graduate Degree]	==>	Customers = [Corvallis]	93.1372549	93.1372549
12	Customers = [Corvallis]	==>	Education Level = [High School Degree]	98.03921569	98.03921569
13	Education Level = [High School Degree]	==>	Customers = [Corvallis]	98.03921569	98.03921569
14	Customers = [Corvallis]	==>	Education Level = [Partial College]	97.05882353	97.05882353
15	Education Level = [Partial College]	==>	Customers = [Corvallis]	97.05882353	97.05882353
16	Customers = [Lebanon]	==>	Education Level = [Bachelors Degree]	99.01960784	99.01960784
17	Education Level = [Bachelors Degree]	==>	Customers = [Lebanon]	99.01960784	99.01960784
18	Customers = [Lebanon]	==>	Education Level = [Graduate Degree]	93.1372549	93.1372549
19	Education Level = [Graduate Degree]	==>	Customers = [Lebanon]	93.1372549	93.1372549
20	Customers = [Lebanon]	==>	Education Level = [High School Degree]	97.05882353	97.05882353
21	Education Level = [High School Degree]	==>	Customers = [Lebanon]	97.05882353	97.05882353
22	Customers = [Lebanon]	==>	Education Level = [Partial College]	98.03921569	98.03921569

IBM DB2 Intelligent Miner

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

IBM Advanced Scout. http://www.research.ibm.com/scout

Using data mining software called Advanced Scout to prepare for a game, a coach can quickly review countless stats: shots attempted, shots blocked, assists made, personal fouls. But Advanced Scout can also detect patterns in these statistics that a coach may not have known about. So during a game, a coach can know exactly which plays are most effective with which players and under what circumstances.

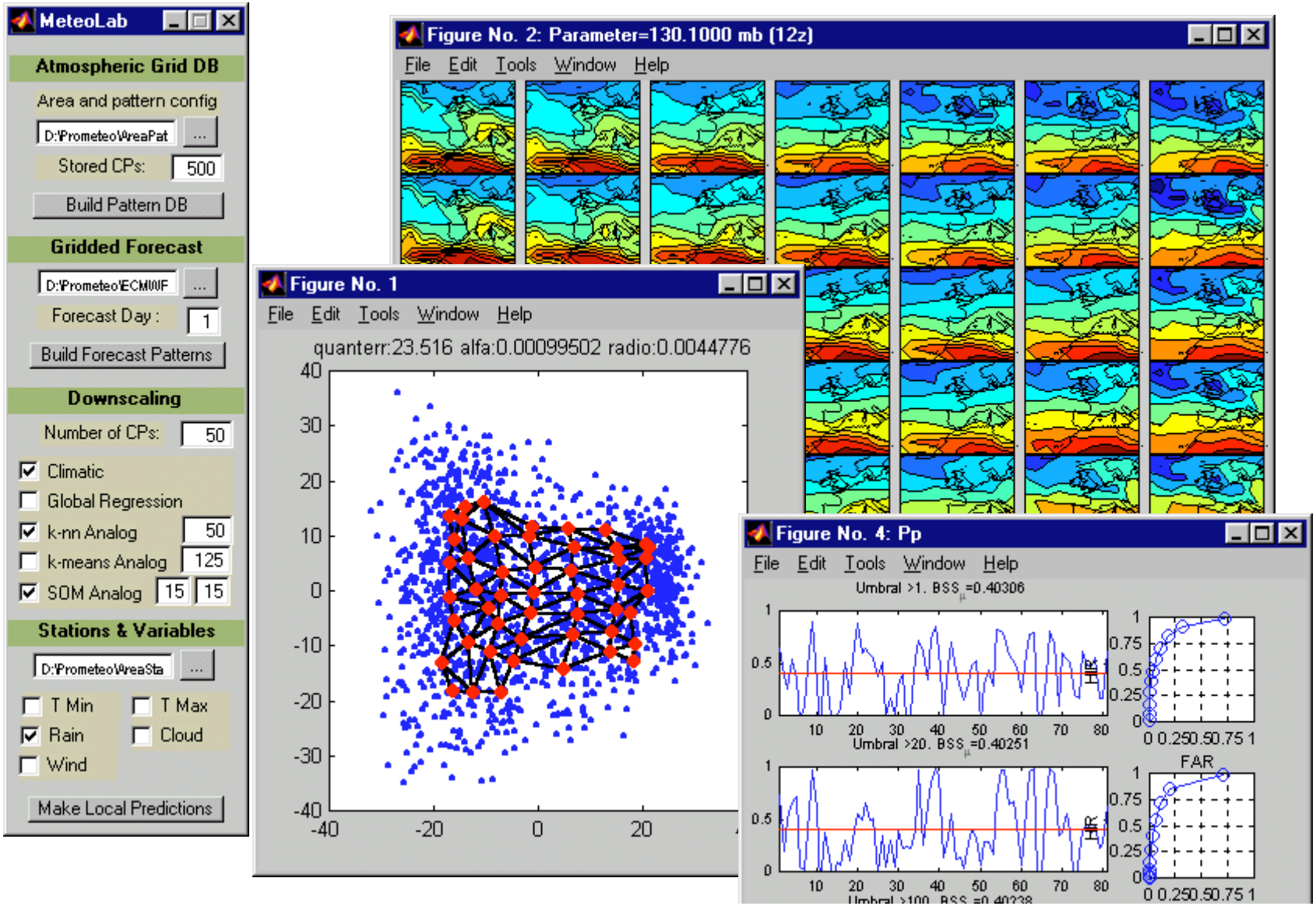
"attribute focusing" finds conditional ranges on attributes where the distributions differ from the norm.

An analysis of the data from a game played between the New York Knicks and the Charlotte Hornets revealed that when "Glenn Rice played the shooting guard position, he shot 5/6 (83%) on jump shots."

Through data mining, Advanced Scout identified a certain player (Rice), playing a certain position (shooting guard), shooting at a certain rate (83%), on a certain type of shot (jump shots). Advanced Scout not only finds this pattern, but points out that it is interesting because it differs considerably from the average shooting percentage of 54% for the Charlotte Hornets during that game.

The Toolbox "MeteoLab" Data Mining in Meteorology

http://etsiso2.macc.unican.es/~meteo



José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

Modelado de Dep. (reglas de asociación)

✓***Modelado de Dependencias:***
asociaciones entre variables.
reglas y grafos.

✓***Componentes Principales:***
compresión *de la información.*

✓***Componentes Independientes:***
extracción de ***características***.

✓***Agrupación:***
hallar grupos *de elementos.*

✓***Clasificación:***
asignar elementos a clases.

✓***Predicción:***
estimación *de valores.*

✓***Visualización:***
representación gráfica.

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Relaciones entre atributos. Fórmulas y Reglas.

*Una de las técnicas más habituales en data mining consiste en **extraer las relaciones relevantes** que existan entre conjuntos de variables **(itemsets)** de la base de datos.*

*De esta forma se pueden detectar **errores**, **fraudes**, e **inconsistencias** fácilmente.*

En el caso de **bases de datos relacionales** trabajaríamos con ***conjuntos*** formados por **pares (atributo # valor)** utilizando los registros de la base de datos.

{Cliente = Pepe, Precio > 10$}

{Producto = Café}

Estas relaciones de asociación se pueden establecer en distintas formas:

✓**Reglas if-then "reglas de asociación"**

*Son implicaciones de la forma **X=>Y***
if (X1= a, X3= c, X5= d) then (X4= b, X2= a)

*La **fiabilidad** [confidence] es la proporción de Aquellos registros con X que también contienen también a Y.*

*La **relevancia** [support] es la proporción de registros que contienen tanto X como Y.*

If **Cliente** is **Pepe**
and **Precio** is lower than **10$**
Then
Producto = **Café**

confidence: **0.98**
The rule exists in **102** *records*
Significance level: error prob < 0.001

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

✓Asociaciones

Se buscan asociaciones de la forma:
(X1= a) <=> (X4= b)
*De los **n** registros de la tabla, las dos igualdades*
Son verdaderas o falsas simultáneamente
*en **rc** casos:*
fiabilidad de la asociación** = **rc /n

*The value **Pepe** in the **Cliente** field is associated with the value **Café** in the **Producto** field*

*Rule's **fiab**:* **0.8**

Ejemplo:

DNI	Renta Familiar	Ciudad	Profesion	Edad	Hijos	Obeso	Casado
11251545	5.000.000	Barcelona	Ejecutivo	45	3	S	S
30512526	1.000.000	Melilla	Abogado	25	0	S	N
22451616	3.000.000	Len	Ejecutivo	35	2	S	S
25152516	2.000.000	Valencia	Camarero	30	0	S	S
23525251	1.500.000	Benidorm	Animador Parque Temático	30	0	N	N

Reglas de Asociación:

(Hijos > 0) => Casado (100%, 2 casos).
Casado => Obeso (100%, 3 casos).

Asociaciones:

Casado e (Hijos > 0) están asociados (80%, 4 casos).
Obeso y casado están asociados (80%, 4 casos)

José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

✓Fórmulas

Relaciones matemáticas ***X=f(Y,Z)=Y*Z***

La ***fiabilidad*** *denota el cociente entre el número de casos en que se cumple la fórmula (suponiendo un cierto error de redondeo) y el número total de casos.*

A = B * C

Where: **A = Total**

B = Cantidad

C = Precio

Rule's Accuracy Level: **0.99**

The rule exists in **1890** *records*

✓Reglas de hortografía.

Estas reglas permiten detectar errores de ortografía. Un nombre es similar a otro pero la frecuencia en que aparecen ambos es muy diferente.

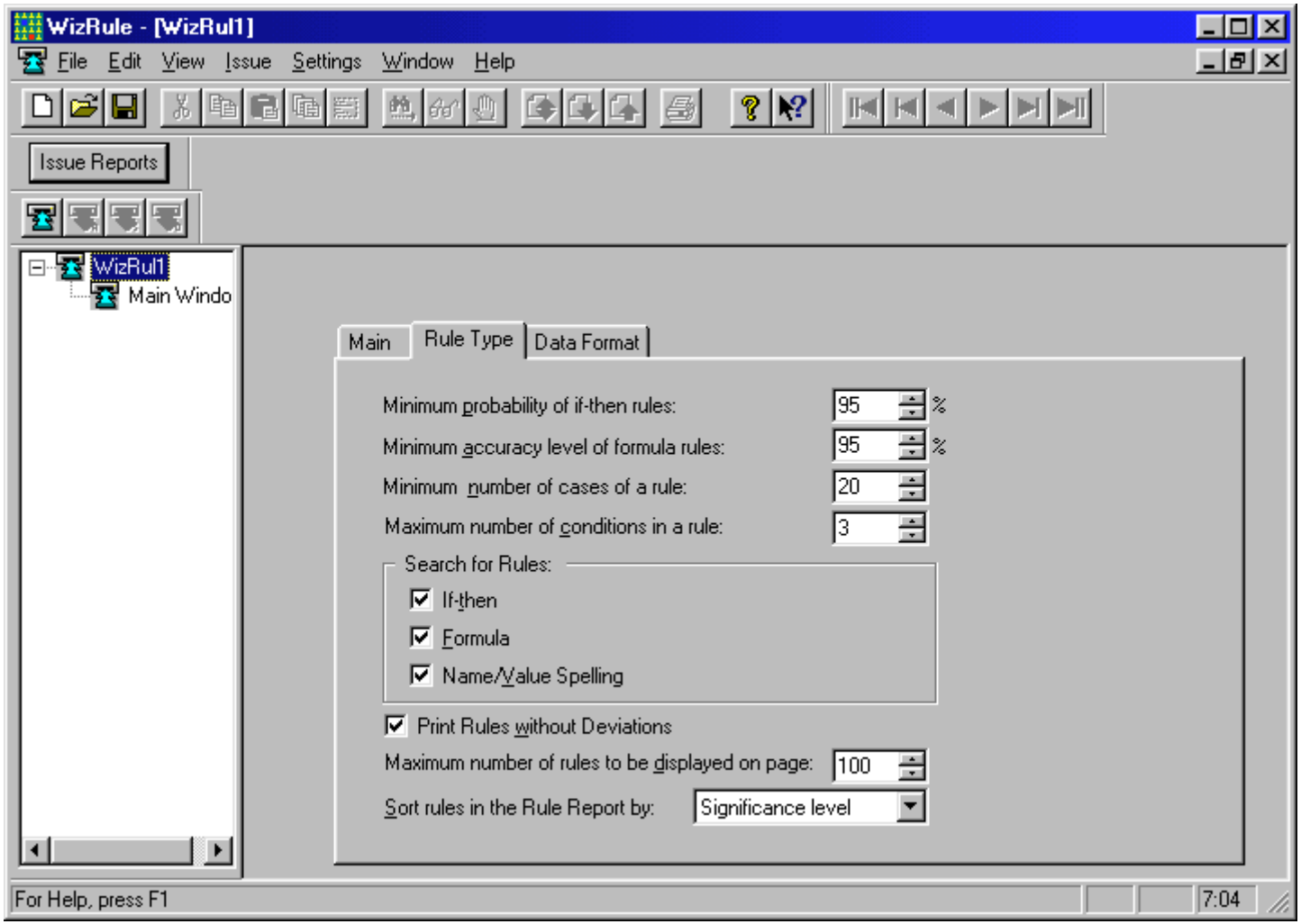
(Text Mining)

The value **Pepe** *appears* **52** *times in the* ***Cliente*** *field.*

There are **2** *case(s) containing similar value(s)* ***{Repr, Repe}***

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Ejemplo



José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Búsqueda de Reglas de Asociación

La mayoría se basa en descomponer el problema en dos fases:

- **FASE A: BÚSQUEDA DE GRANDES CONJUNTOS DE ATRIBUTOS**.

*Se buscan conjuntos de atributos con **relevancia >= umbral**. De momento no se busca separarlos en parte izquierda y parte derecha.*

- **FASE B: ESCLARECIMIENTO DE DEPENDENCIAS** (REGLAS).

*Se hacen particiones binarias y disjuntas de los conjuntos hallados y se calcula la confianza de cada uno. Se retienen aquellas reglas que tienen **confianza >= umbral***

Propiedad: cualquier subconjunto de un conjunto grande es también grande.

***AIS** es el primer algoritmo que se desarrolló para obtener reglas de asociación.*

***X=>Y [s,c]** donde*

- ✓ ***Y** es un único atributo,*
- ✓ ***s** es la relevancia y*
- ✓ ***c** su fiabilidad.*

AIS [Agrawal, Imielinski & Swami]
R. Agrawal, T. Imielinsky & A. Swami
IBM Almaden Research Center, 1993

José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

Fase A: Selección Grandes de Atributos

Dada una relevancia mínima ***Rmin***:

1. **i** = 1 (tamaño de los conjuntos)
2. Generar un conjunto unitario en **S1** para cada atributo.
3. Comprobar la relevancia de todos los conjuntos en **Si**.
   Eliminar aquellos cuya relevancia < ***Rmin***.
4. Combinar los conjuntos en **Si**
   creando conjuntos de tamaño i+1 en **Si+1**.

*Este paso se lleva a cabo secuencialmente, recorriendo los registros de la base de datos siguiendo el contador **i**. Tras leer un registro de la base de datos, se hallan los conjuntos relevantes **Si** contenidos en el mismo. **Si+1** se genera extendiendo los conjuntos hallados con otros atributos del registro.*

5. **Si** Si no es vacío **entonces** **i:= i+ 1**. Ir a 3.
6. **Si no** , retornar ***S2*** ∪ ***S3*** ∪ ... ∪ ***Si***

Dados n registros y m atributos $o(m\,2^{m-1})$ reglas posibles.

Complejidad computacional $o(n\,m\,2^{m})$

Ejemplo

Fila	1	2	3	4	5
1	x		x	x	
2		x	x		x
3	x	x	x		x
4		x			x

relevancia = 2

confianza = 0.75

FASE A:

S1 = {{1}, {2}, {3}, {4}, {5}}

S2 = {{1,2},{1,3},{1,5},{2, 3},{2, 5},{3, 5}}

S3 = {{1,2, 3}, {1,2, 5}, {1,3, 5}, {2,3, 5}}

S'1:rel = {{1}:2, {2}:3, {3}:3, {5}:3}

S'2:rel = {{1,3}:2, {2,3}:2, {2,5}:3, {3,5}:2}

S'3:rel = {{2,3,5}:2}

Sfinal = *S'2* □ *S'3* = **{{1, 3}, {2, 3}, {2, 5}, {3, 5}, {2,3,5}}**

FASE B:

{1} □ {3} : 1
{2} □ {3} : 0.67
{2} □ {5} : 1
{3} □ {5} : 0.67
{2,3} □ {5} : 1

{3} □ {1} : 0.67
{3} □ {2} : 0.67
{5} □ {2} : 1
{5} □ {3} : 0.67
{2,5} □ {3} : 0.67

{3,5} □ {2} : 1

El Algoritmo APRIORI

- **Fk** : Set of frequent itemsets of size k
- **Ck** : Set of candidate itemsets of size k

F1 = {single attribute sets} with minimum support

for (k=2; Fk != 0; k++) do {

  Ck+1 = *New candidates generated from Fk*

  foreach entry t in the database do

    Increment the count of all candidates in Ck+1 contained in t

    Fk+1 = Candidates in Ck+1 with minimum support

  }

Answer = **Uk Fk**

- Every subset of a frequent itemset is also frequent

  => a candidate itemset in Ck+1 can be pruned if even one of its subsets is not contained in Fk

José Manuel Gutiérrez, *Universidad de Cantabria, (2001)*
http://personales.unican.es/gutierjm

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Fase de Combinación

Este algoritmo realizan múltiples pasadas sobre la base de datos para obtener los conjuntos de atributos relevantes.

En la primera pasada, se obtienen los items individuales cuya relevancia alcanza el umbral mínimo preestablecido: L[1] de conjuntos relevante.
En las siguientes iteraciones, se utiliza el último conjunto L[k] obtenido para generar un conjuntos de (k+1) atributos potencialmente relevantes (el conjunto de candidatos C[k+1]) y se obtiene la relevancia de estos candidatos para quedarnos sólo con aquéllos que son relevantes, que incluimos en el conjunto L[k+1]. Este proceso se repite hasta que no se encuentran más itemsets relevantes.

En el algoritmo AIS, los candidatos se generaban sobre la marcha, conforme se iban leyendo registros de la base de datos. Se generan innecesariamente conjuntos candidatos que de por sí nunca pueden llegar a ser relevantes.
Por su parte, en Apriori los candidatos se generan a partir de los conjuntos relevantes encontrados en la iteración anterior, única y exclusivamente. La idea subyacente es que, dado un itemset relevante, cualquier subconjunto suyo también es relevante.
Por lo tanto, los conjuntos de k atributos candidatos del conjunto C[k] pueden generarse a partir del conjunto L[k-1].

Ejemplo

Database D

TID	Items
1	{1, 3, 4}
2	{2, 3, 5}
3	{1, 2, 3, 5}
4	{2, 5}

Scan D →

C_1

Itemset	Sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

F_1

Itemset	Sup.
{2}	3
{3}	3
{5}	3

C_2

Itemset
{2, 3}
{2, 5}
{3, 5}

Scan D →

C_2

{2, 3}	2
{2, 5}	3
{3, 5}	2

F_2

Itemset	Sup.
{2, 5}	3

José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

Lógica

La lógica proporciona un entorno para representar conocimiento en el que es fácil razonar.

eg1. John is a human
every human are mortals
therefore
John is mortal.

In logic:
human(John)
□h(human(h) □ mortal(h))
therefore: human(John) □ mortal(John) □ **elim. rule**
therefore: mortal(John) □ **elim. rule**

Las expresiones lógicas se construyen en base a un conjunto reducido de símbolos y cuantificadores.

- *Símbolos lógicos*
  ~ NOT □ AND OR □ IMPLIES
- *Cuantificadores*
  □ FOR ALL □ THERE EXISTS

José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

Lógica. Representación de Conocimiento con L_{PC}

A language of PC, call it L_{PC} is defined by the following rules:

1. Variables *p, q, r,...* are in L_{PC}. We call the above variables: ***undeterminate statements***.

2. If a statement *A* is in L_{PC} and a statement *B* is in L_{PC}, then the statement ***(A&B)*** is in L_{PC} .Similarly for the symbols: , □ .

3. If a statement *A* is in L_{PC}, then the statement **~A** is in L_{PC} .

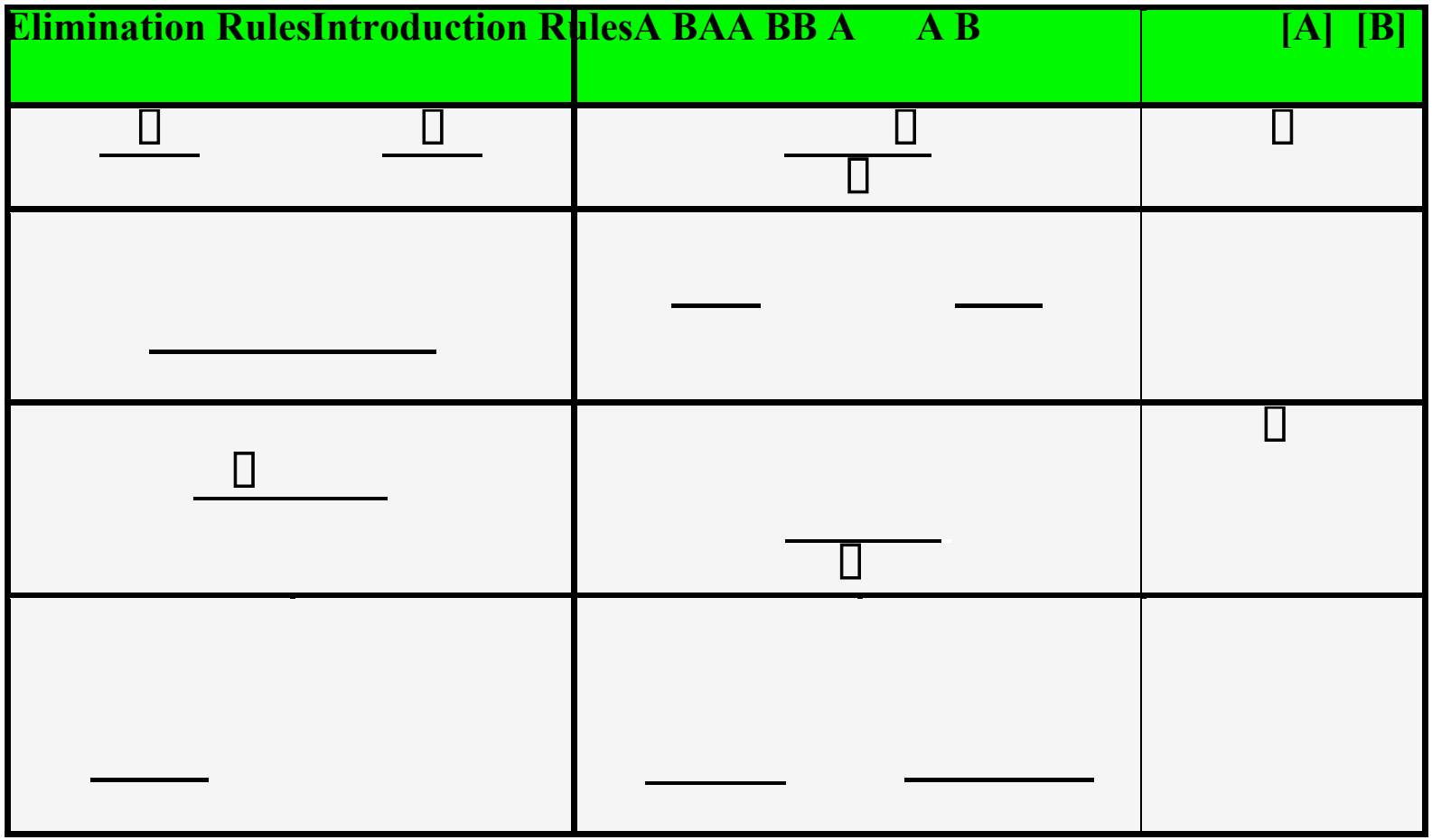
L_{PC} is a set of statements which represent useful logical expressions for a given problem

- (~A□ B)
- (((A□ B)&(A B)□ B)

Using the above rules and some other logical inference techniqes it is easy to reason on a given problem.

Inferencia Lógica. Deducción natural.

Natural deduction uses the definition of logical symbols for eliminating, or introducing, knowledge on a given expression.



José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

Tablas de Verdad y Leyes Lógicas

- ~(~P) = P
- (P Q) = (~P Q) [or (~ P Q) = (P Q)]

P	Q	~P	~P ∨ Q	P → Q	(~P ∨ Q)= (P → Q)
T	T	T	T	T	T
T	F	F	T	F	F
F	T	F	T	T	F
F	F	F	F	T	T

- De Morgan's laws:
  ~(P Q) = (~P ~Q)
  ~(P Q) = (~P ~Q)
- Distributive laws:
  P (Q R) = (P Q) (P R)
  P (Q R) = (P Q) (P R)

José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

Reglas de Inferencia Lógica.

- Modus ponens

  If P is true and P → Q is true

  then Q is true

- Modus tolens

  if P → Q is true and Q is false or ~Q is true

  then ~P is true

  e. g., sick(student) → not_ attend_ lecture(student)

  ~not_ attend_ lecture(student)

  produces: ~sick(student)

- Elimination

  if P → Q is true

  then P is true and Q is true

José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

Modelado de Dep. (redes Bayesianas)

- ✓ ***Componentes Principales:***
  compresión de la información.
- ✓ ***Componentes Independientes:***
  extracción de características.
- ✓ ***Modelado de Dependencias:***
  hallar asociaciones entre variables
  redes Bayesianas
- ✓ ***Agrupamiento:***
  hallar grupos de elementos
- ✓ ***Clasificación:***
  asignar elementos a clases
- ✓ ***Predicción:***
  estimación de valores
- ✓ ***Visualización:***
  representación gráfica.

Redes Neuronales

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Redes Probabilísticas. Redes Bayesianas

Algunos problemas involucran gran número de variables y se conocen ciertas relaciones de independencia entre ellas.

Obtener un modelo probabilístico

Cto. de relaciones	Cto. de variables aleatorias
$I(X,Y\|Z)_M$	$\{X_1, X_2, ..., X_n\}$

→ Función de probabilidad conjunta $P(X_1, X_2, ..., X_n)$

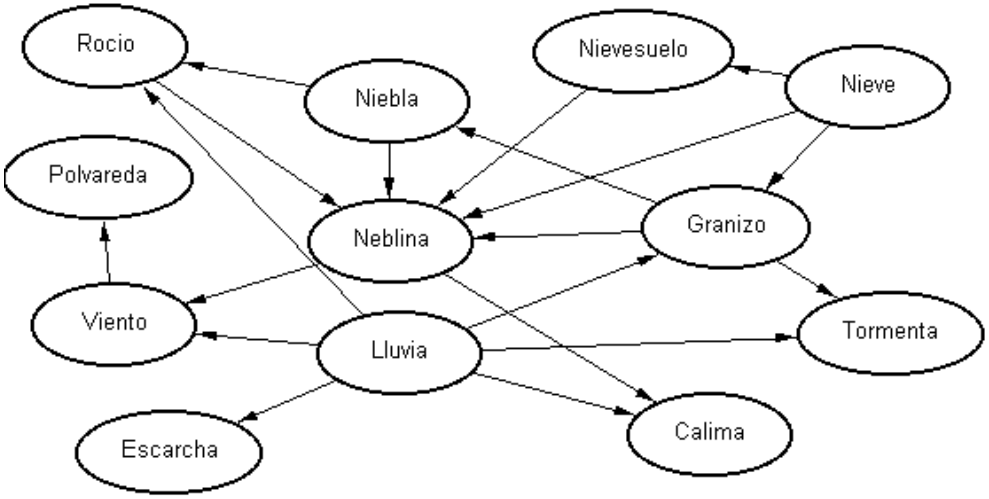
Factorización de la probabilidad !!

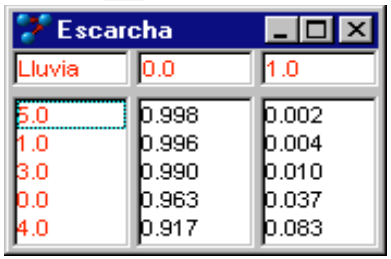
$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P_i(x_i \mid \Pi_i)$$

Lluvia	*Nieve*	*Granizo*	*Tormenta*	*Niebla*	...
5	0	0	0	0	...
1	0	0	0	0	...
5	0	0	1	0	...

Relaciones de dependencia

Mediante un grafo dirigido donde cada variable tiene sus antecedentes.





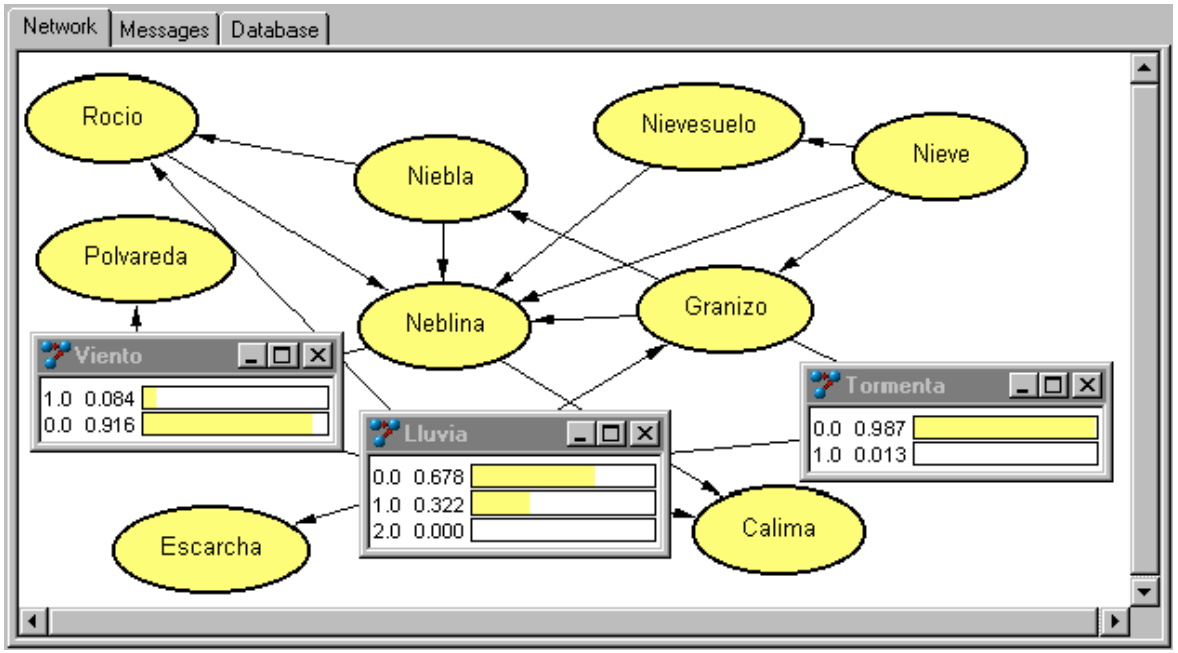
Cuantificación

Funciones de prob. condicionada.

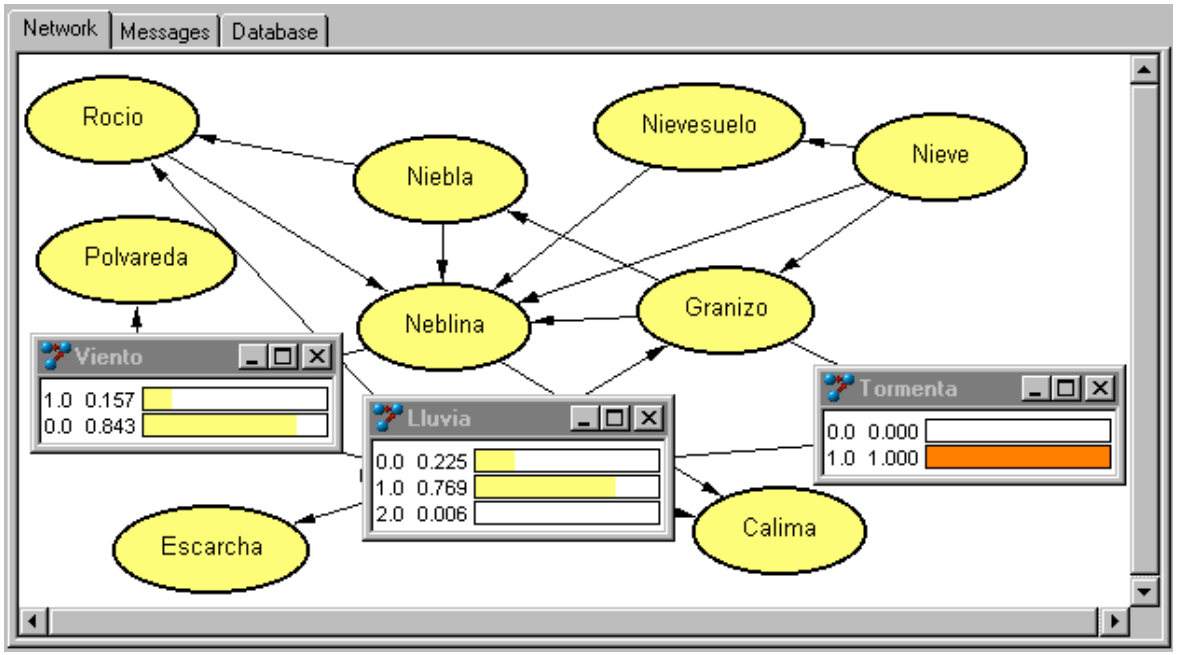
José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Cálculo de probabilidades

*Inicialmente los distintos estados de las variables de la red tienen probabilidades que corresponden al estado de conocimiento inicial (**sin evidencia**).*



Cuando se tiene alguna evidencia, las nuevas probabilidades condicinadas dan la influencia de esta información en el resto de variables ***Tormenta = 1***



José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

Componentes Principales e Indepenedientes

✓***Modelado de Dependencias:***

asociaciones entre variables.
reglas y grafos.

✓***Componentes Principales:***

compresión de la información.

✓***Componentes Independientes:***

extracción de características.

✓***Agrupación:***

hallar grupos de elementos.

✓***Clasificación:***

asignar elementos a clases.

✓***Predicción:***

estimación de valores.

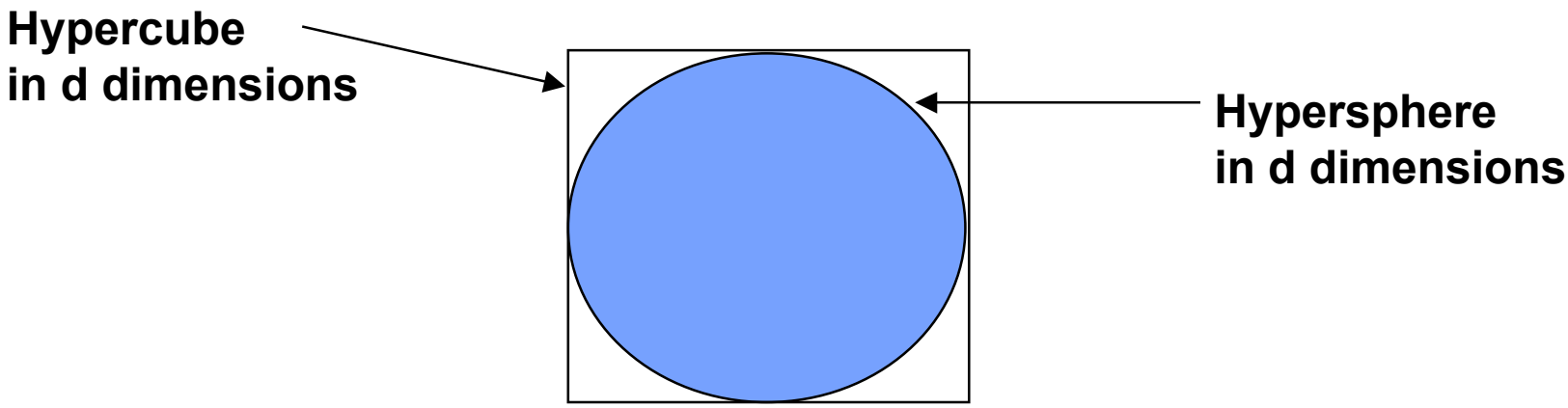
✓***Visualización:***

representación gráfica.

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Problemas con datos de alta dimensionalidad

(David Scott, *Multivariate Density Estimation,* Wiley, 1992)



Volume of sphere relative to cube in d dimensions?

Dimension	2	3	4	5	6	7
Rel. Volume	0.79	0.53	0.31	0.16	0.08	0.04

- *high-d, uniform => most data points will be "out" at the corners*
- ***high-d space is sparse: and non-intuitive***

José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

Ejemplos y casos a estudiar

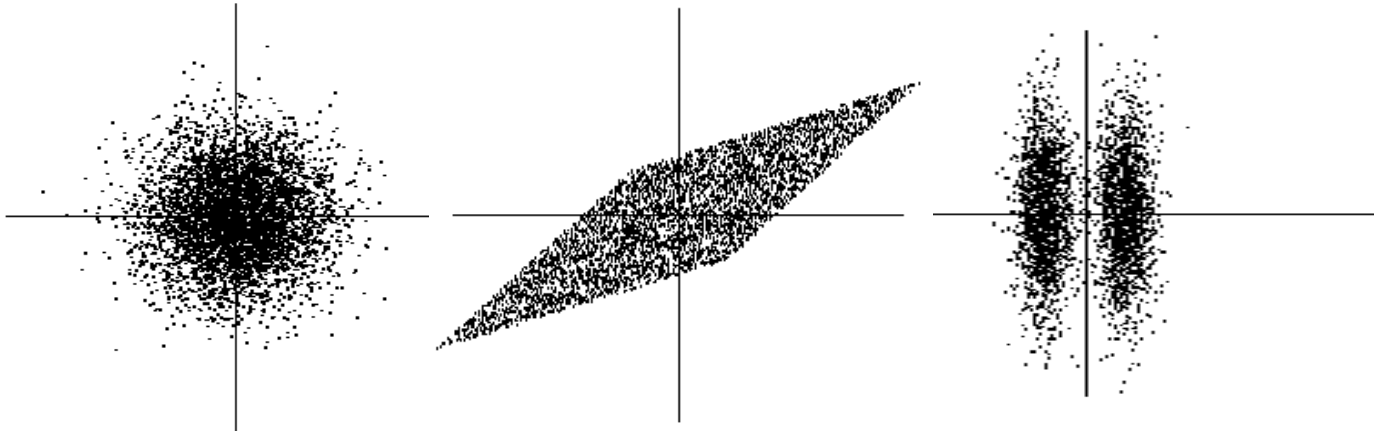
$$Y^k = M\, X^k$$

Datos Aleatorios Gaussianos
Análisis de Componentes Principales

Maximizar varianza

$$X^k = \sum_{i=1}^{d} x_i^k\, E_i$$

$$\sum_{i=1}^{r} c_i^k\, V_i,$$



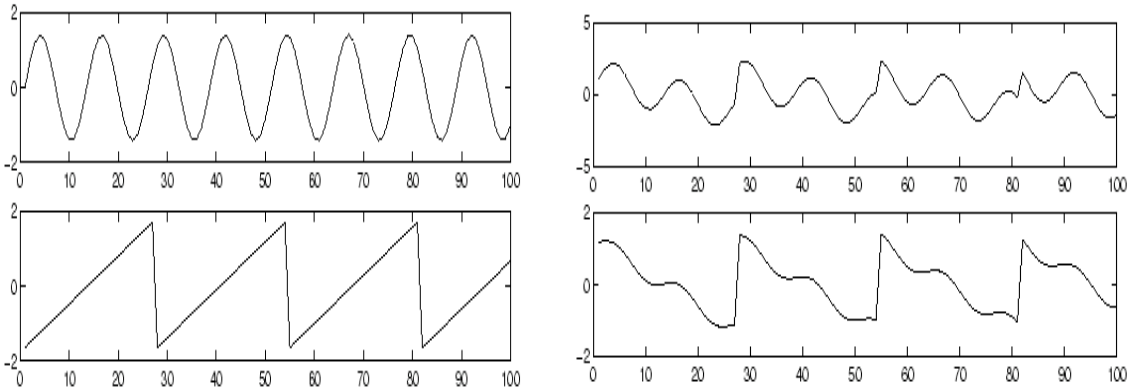
Datos Aleatorios NO-Gaussianos
Análisis de Componentes Independientes

Maximizar independencia

$$X^k = M\, S^k$$

X es la mezcla de m señales S Independientes. Dada X:

$$y_i^k = w_i^T\, X^k$$ Indep.



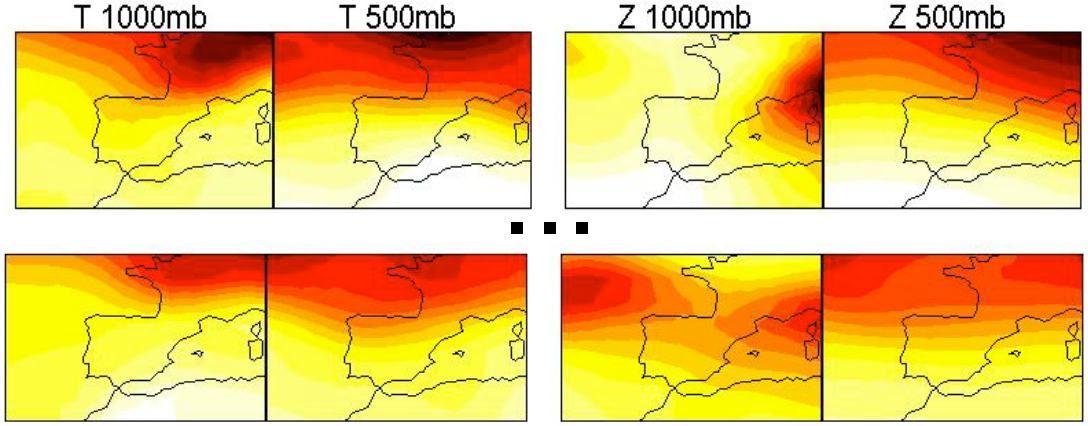
José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Ejemplos y casos a estudiar

$$Y^k = M\,X^k$$

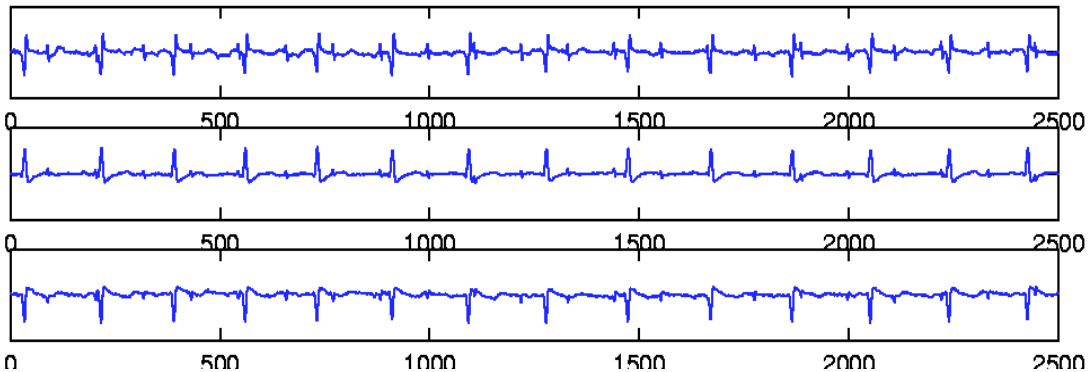
Datos Aleatorios Gaussianos
Análisis de Componentes Principales

Maximizar varianza



Datos Aleatorios NO-Gaussianos
Análisis de Componentes Independientes

Maximizar independencia



Sistemas Deterministas (Caos determinista)
Estimación No-Paramétrica

$$Y^k = F(X^k)$$

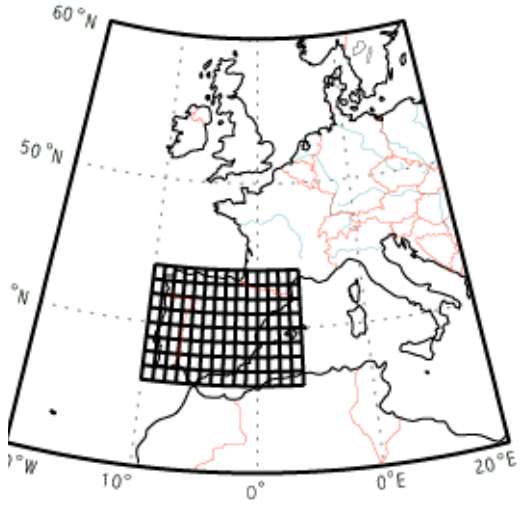
Estimar **F**

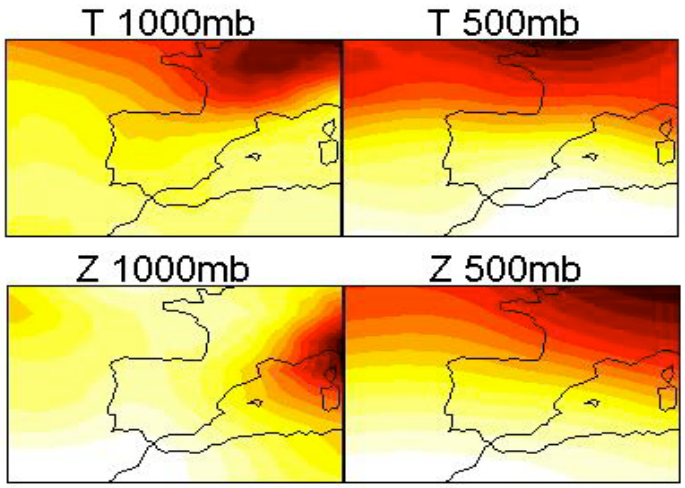
Base de Datos de Re-Análisis del Centro Europeo

*El **Reanálisis del ECMWF** proporciona una base de datos de salidas del modelo numérico.*

- Serie diaria **1979-1993** a las **0**, **6**, **12** y **18** horas.
- En cada uno de los nodos
  - **5 variables** Z, T, U , V y H
  - **7 niveles** de presión

$$\begin{cases} \dfrac{\partial u}{\partial t} = -u\dfrac{\partial u}{\partial x} - v\dfrac{\partial u}{\partial y} - \omega\dfrac{\partial u}{\partial p} - \dfrac{\partial \Phi}{\partial x} + fv + F_x \\ \dfrac{\partial v}{\partial t} = -u\dfrac{\partial v}{\partial x} - v\dfrac{\partial v}{\partial y} - \omega\dfrac{\partial v}{\partial p} - \dfrac{\partial \Phi}{\partial y} - fu + F_y \\ \dfrac{\partial \Phi}{\partial p} = -\dfrac{RT}{p} \\ \dfrac{\partial T}{\partial t} = -u\dfrac{\partial T}{\partial x} - v\dfrac{\partial T}{\partial y} + \omega\left(\dfrac{\kappa T}{p} - \dfrac{\partial T}{\partial p}\right) + \dfrac{\dot{H}}{c_p} \\ \dfrac{\partial \omega}{\partial p} = -\left(\dfrac{\partial u}{\partial x} + \dfrac{\partial v}{\partial y}\right) \end{cases}$$





José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

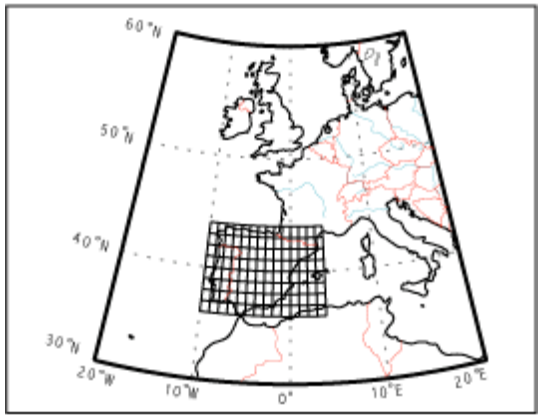
José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Componentes Principales. Primera Opción

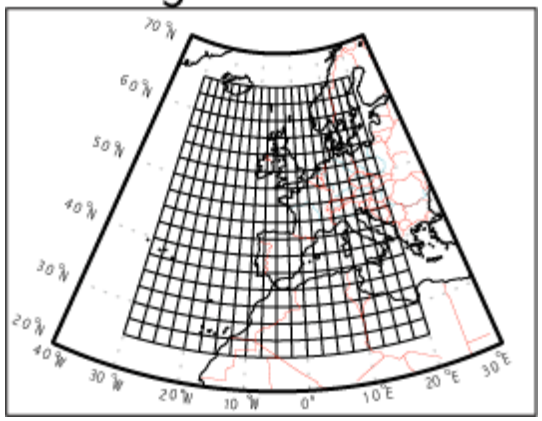
We used atmospheric circulation patterns at 1200 UTC of ERA-15 (1979-1993) reanalysis data

P=(**T**(1ooo mb),..., **T**(500 mb); **Z**(1ooo mb),..., **Z**(500 mb); .....; **H**(1ooo mb),..., **H**(500 mb))

Limited Area Grid

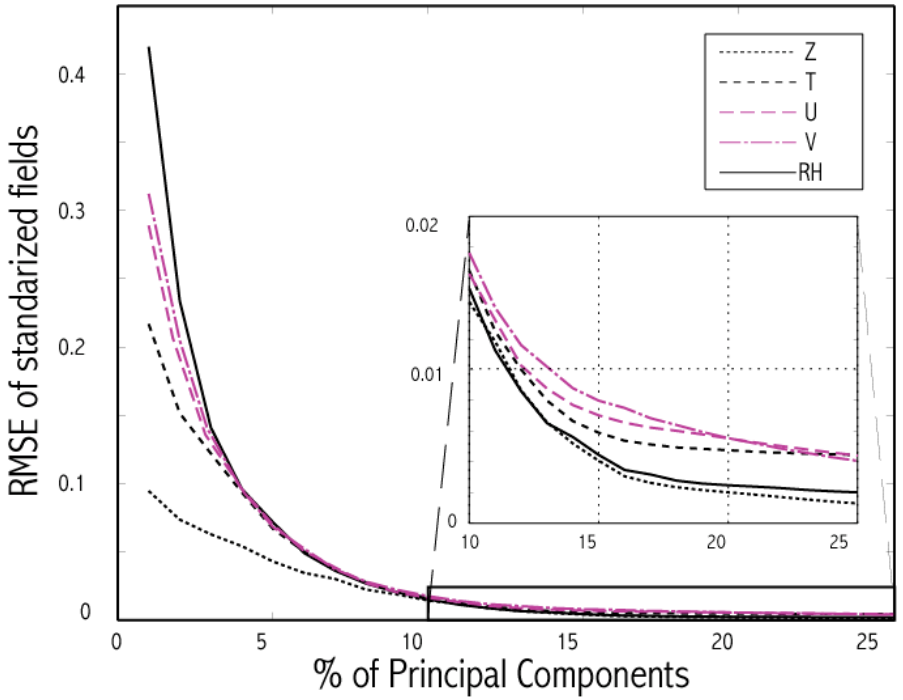


Large Scale Grid



P is 6000 dimensional !!!!

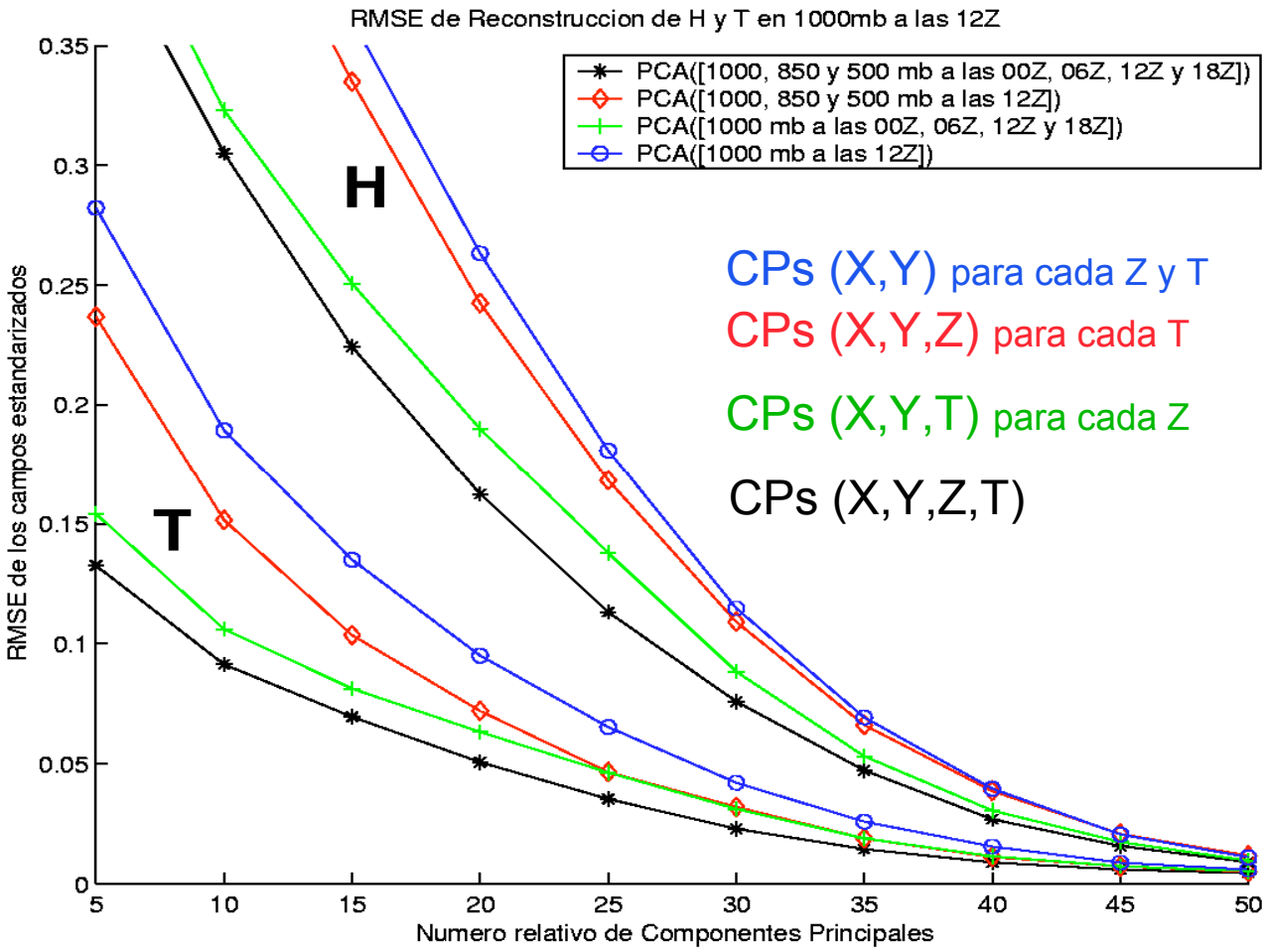
Using Principal Components the dimension can be reduced to 500 – 600.



José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Componentes Principales. Alternativas

La configuración atmosférica de un día concreto viene dada por un campo **(X,Y,Z)** para cada **T**=0, 6, 12 y 18 horas



CPs (X,Y) para cada Z y T

CPs (X,Y,Z) para cada T

CPs (X,Y,T) para cada Z

CPs (X,Y,Z,T)

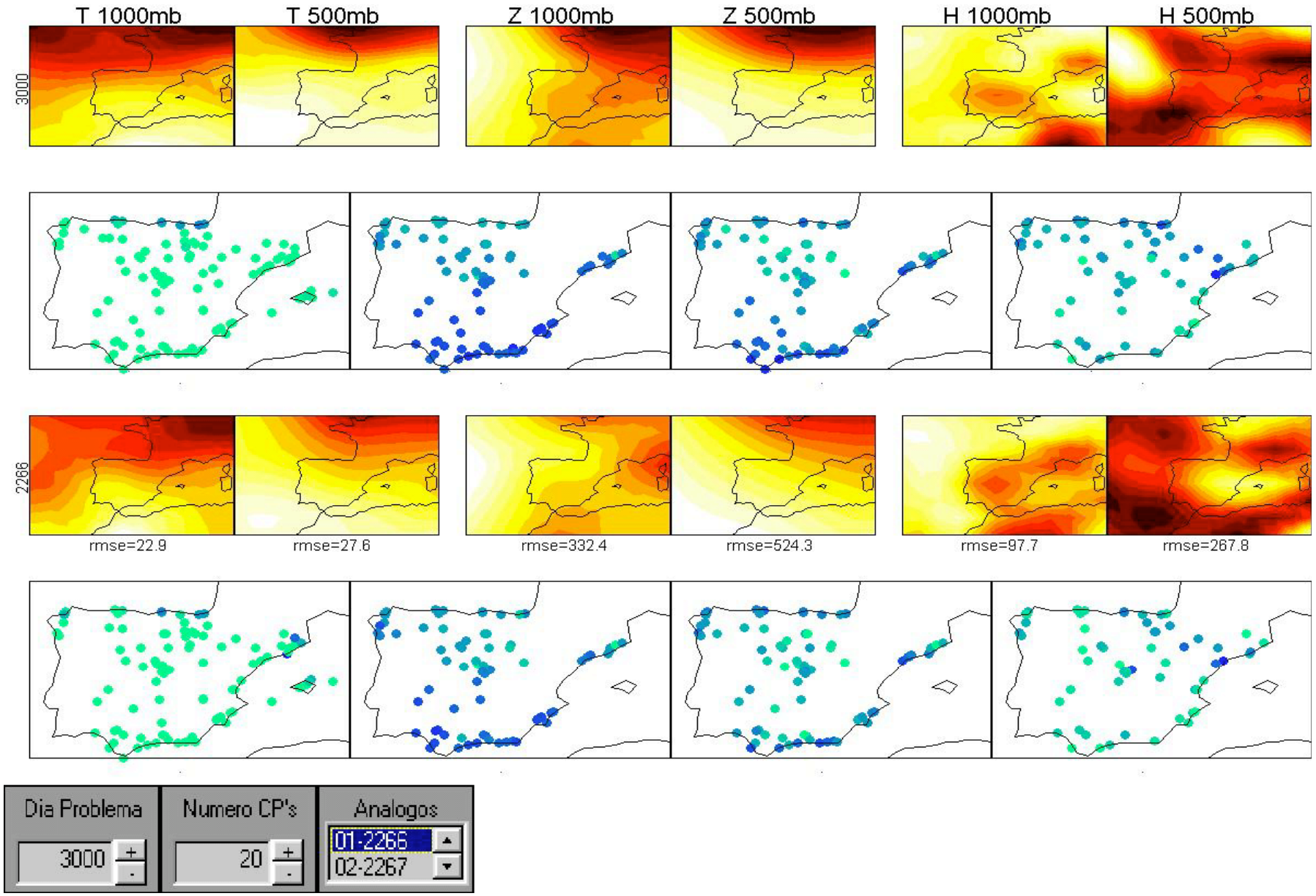
$$X^k = \sum_{i=1}^{d} v_i^k \, E_i$$

$$\sum_{i=1}^{r} c_i^k \, V_i,$$

$$k = 1, 2, ..., n$$

Si los vectores X^k *son realizaciones de una variable Gaussiana, los* V_i *óptimos son los autovectores de la matrix de covarianza.*

Componentes Principales con MeteoLab



José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

Agrupación y Clasificación

- ✓ ***Componentes Principales:***
  compresión de la información.
- ✓ ***Componentes Independientes:***
  extracción de características.
- ✓ ***Modelado de Dependencias:***
  hallar asociaciones entre variables
  redes Bayesianas
- ✓ ***Agrupación:***
  hallar grupos de elementos
- ✓ ***Clasificación:***
  asignar elementos a clases
- ✓ ***Predicción:***
  estimación de valores
- ✓ ***Visualización:***
  representación gráfica.

Redes Neuronales

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Técnicas de Aprendizaje Automático

Predictivas

- ***Interpolación***: una función continua sobre varias dimensiones
- ***Predicción secuencial***: las observaciones están ordenadas secuencialmente. Se predice el siguiente valor de la secuencia. Caso particular de interpol. con 2 dim., una discreta y regular.
- ***Aprendizaje supervisado***: cada observación incluye un valor de la clase a la que corresponde. Se aprende un clasificador. Caso particular de interpolación: la clase (imag. función) es discreta.

Deductivas

- ***Aprendizaje no supervisado***: el conjunto de observaciones no tienen clases asociadas. El objetivo es detectar regularidades en los datos de cualquier tipo: agrupaciones, contornos, asociaciones, valores anómalos.

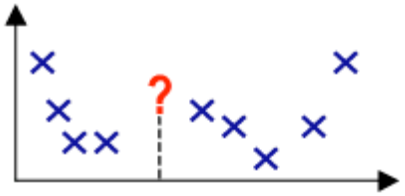
José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

Aprendizaje Automático. Ejemplos

Ejemplos:

Predictivos

- *Interpolacion*:



f(2.2)=?

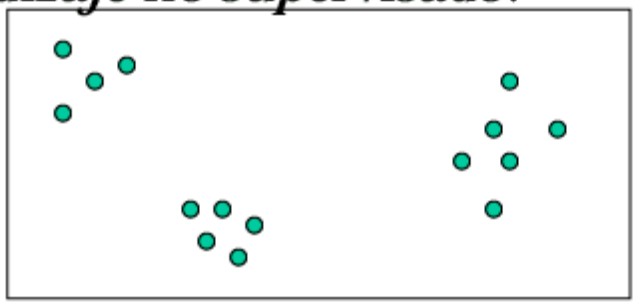
- *Prediccion secuencial*: 1, 2, 3, 5, 7, 11, 13, 17, 19, ... ?
- *Aprendizaje supervisado*:

1 3 -> 4.
3 5 -> 8.
7 2 -> 9.

4 2 -> ?

Descriptivos

- *Aprendizaje no supervisado*:



¿Cuántos grupos hay?

Clasificación. Planteamiento del Problema

- Classification is an important problem in artificial intelligence
- We have a special discrete-valued variable called the Class, C
  - C takes values in $\{c_1, c_2, \ldots\ldots c_m\}$
  - for now assume m=2, i.e., 2 classes: $c_1 = 1$, $c_2 = 2$
- Problem is to decide what class an object is
  - i.e., what value the class variable C is for a given object
  - given measurements on the object, e.g., X, Y, ….
  - These measurements are called “features”
  - **we wish to learn a mapping from Features -> Class**
- Notation:
  - C is the class
  - X, Y, etc (the measurements) are called the “features” (sometimes also called “attributes”)

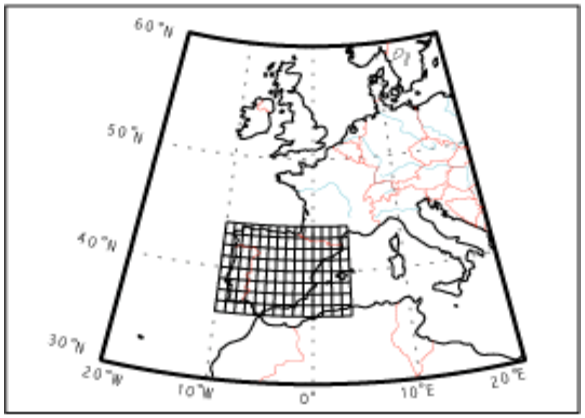
José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Clustering ERA-15 Data for **Selecting Analogues Ensembles**

We used atmospheric circulation patterns at 1200 UTC of ERA-15 (1979-1993) reanalysis data

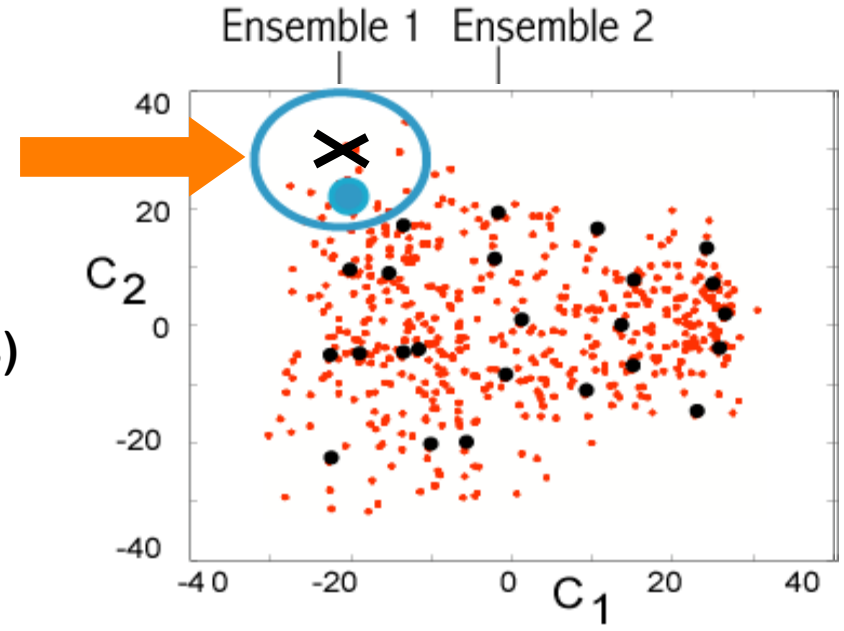
P=(**T**(1ooo mb),..., **T**(500 mb); **Z**(1ooo mb),..., **Z**(500 mb); .....; **H**(1ooo mb),..., **H**(500 mb))

Limited Area Grid



P is 6000 dimensional !!!!

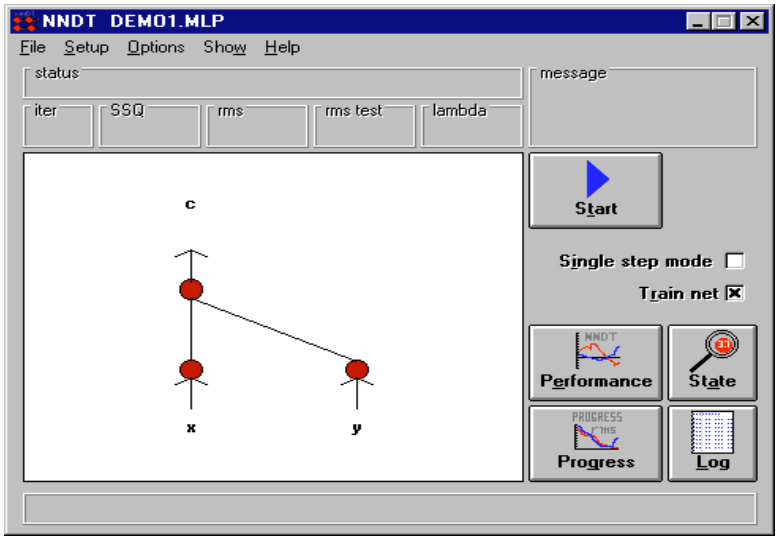
Using Principal Components the dimension can be reduced to 500 – 600.



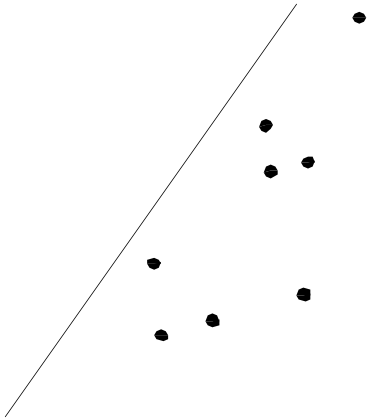
- **Projection of ERA-15 using the first 2 PCs**
- **Cluster centers (prototypes)**
- **Predicted pattern** to be downscaled

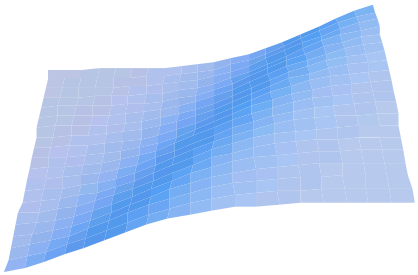
Clasificación con Redes Neuronales

Dada una nube de puntos en el plano correspondientes a dos clases distintas, se quiere obtener un criterio de clasificación automático, que extrapole la información de estos puntos.

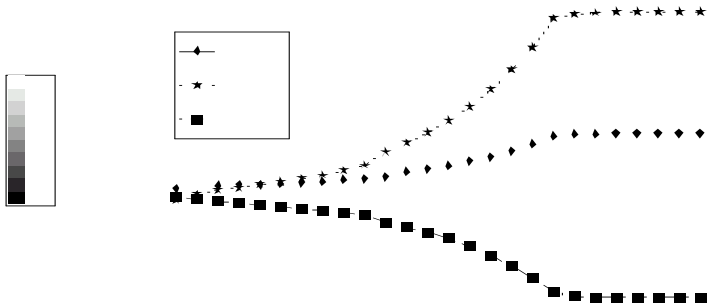


ERROR





PESOS

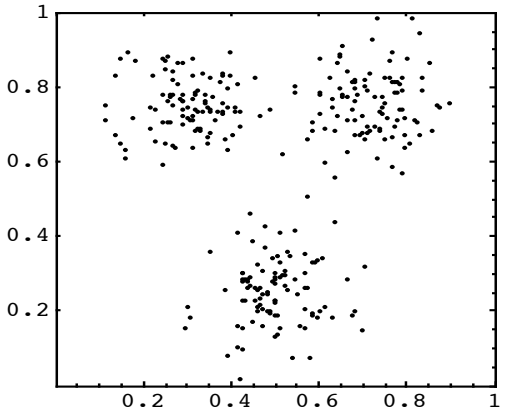


José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

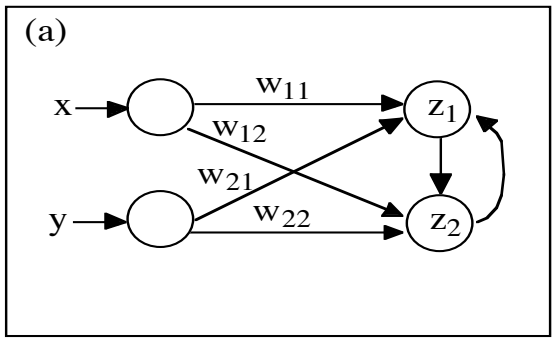
José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

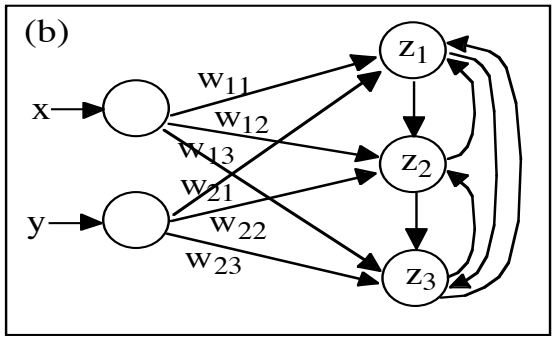
Clasificación no supervisada

Las redes competitivas permiten obtener criterios de clasificación no supervisados (el usuario no especifica el número de clases).

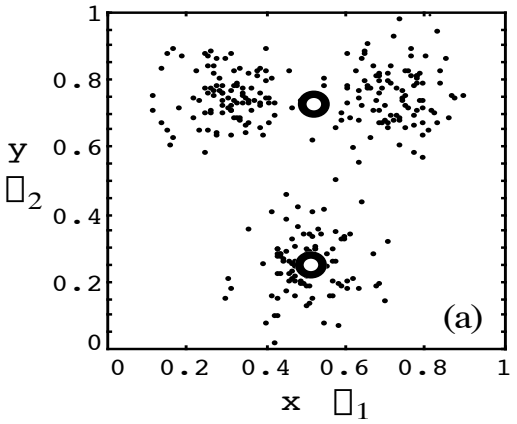


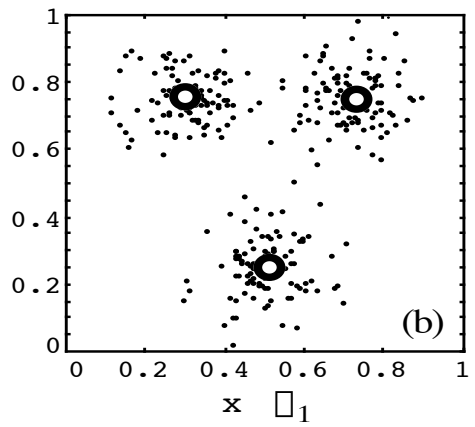
*Red con **dos** procesadores de salida (dos clases)*





*Red con **tres** procesadores d salida (tres clases)*





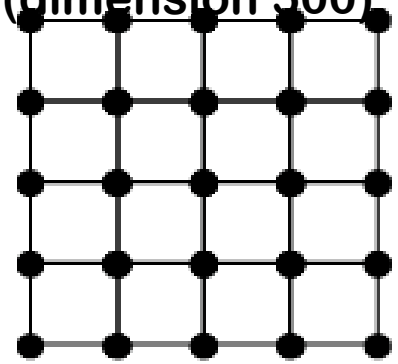
José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

Self-Organizing Maps. Preserving Data Topology

For medium and seasonal weather forecast, we need a measure of the dispersion of the forecast ensemble members.

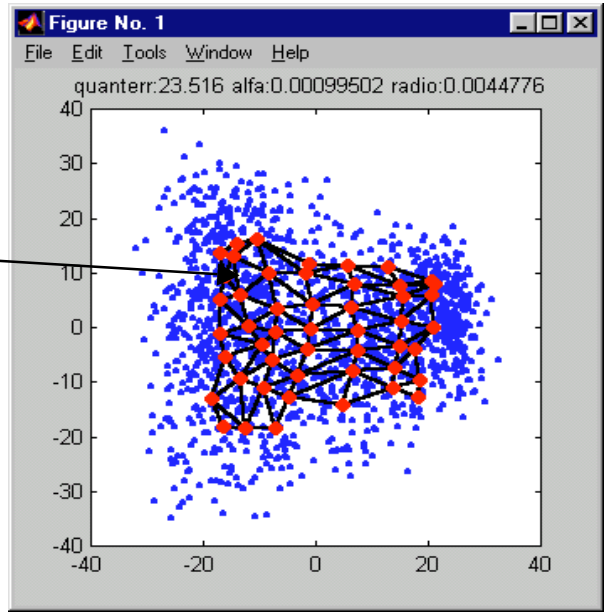
Therefore, we also need a measure of Cluster Dispersion.

Cluster units are located on a 2D lattice, each one associate with a pattern prototype (dimension 500)



Adaptive Competitive Learning

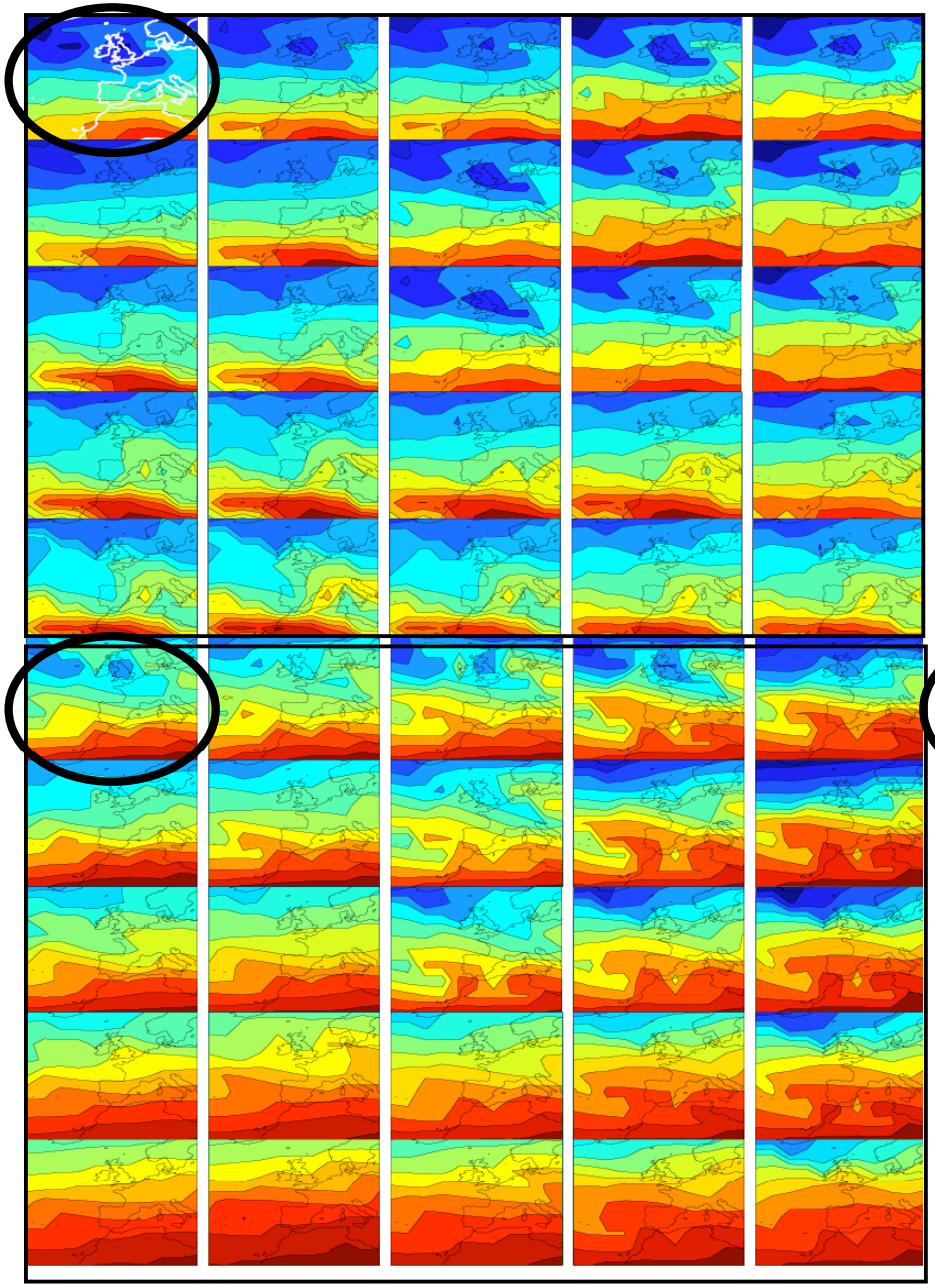
Dimension 500



Topology preserving transformation.

Close clusters in the lattice correspond to close prototypes in the high dimensional data space.

T 1000mb, T 500 mb Z,U,V 500mb patterns for a 5X5 SOM



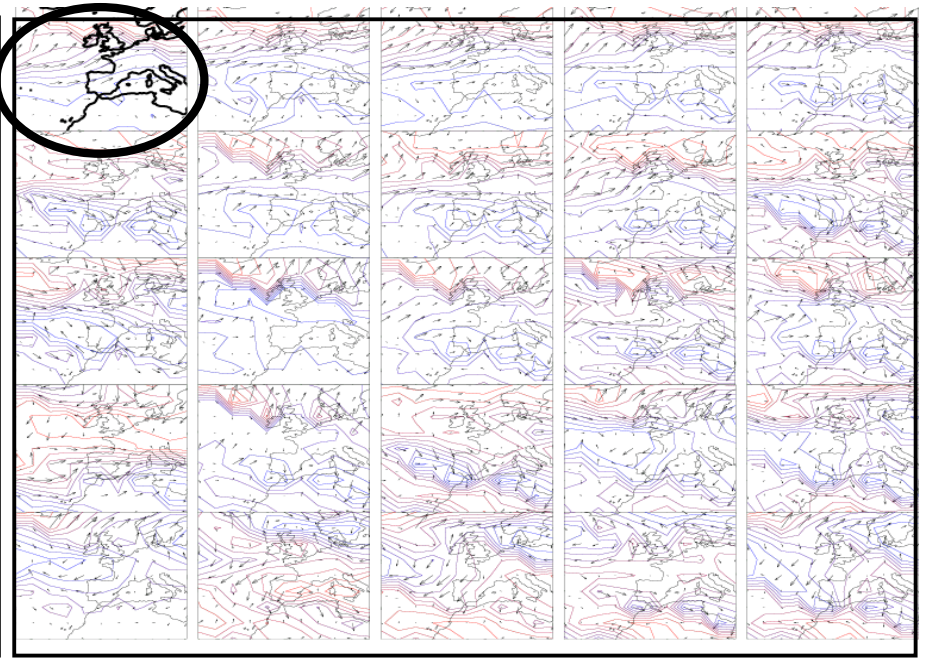
Prototypes for a trained SOM.

Close units in the lattice are associated with similar atmospheric patterns.

T 1000mb

T 500mb

Z, U, V 500mb



José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

Predicción. Regresión y Redes Neuronales

- ✓ ***Componentes Principales:***
  compresión de la información.
- ✓ ***Componentes Independientes:***
  extracción de características.
- ✓ ***Modelado de Dependencias:***
  hallar asociaciones entre variables
  redes Bayesianas
- ✓ ***Agrupación:***
  hallar grupos de elementos
- ✓ ***Clasificación:***
  asignar elementos a clases
- ✓ ***Predicción:***
  estimación de valores
- ✓ ***Visualización:***
  representación gráfica.

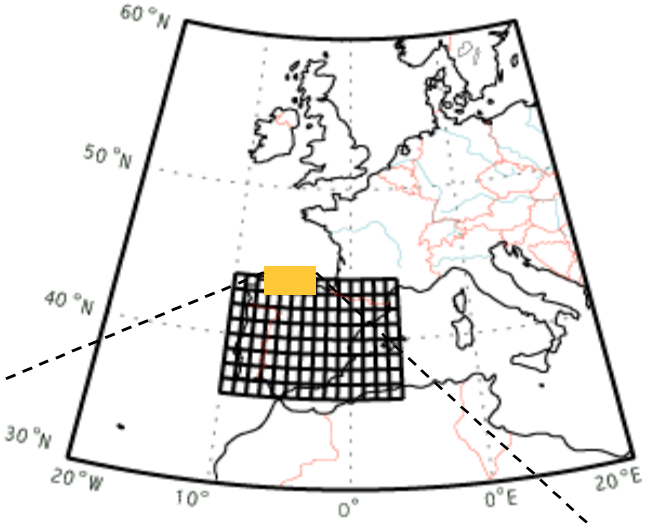
Redes Neuronales

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Weather Forecasting (Downscaling)

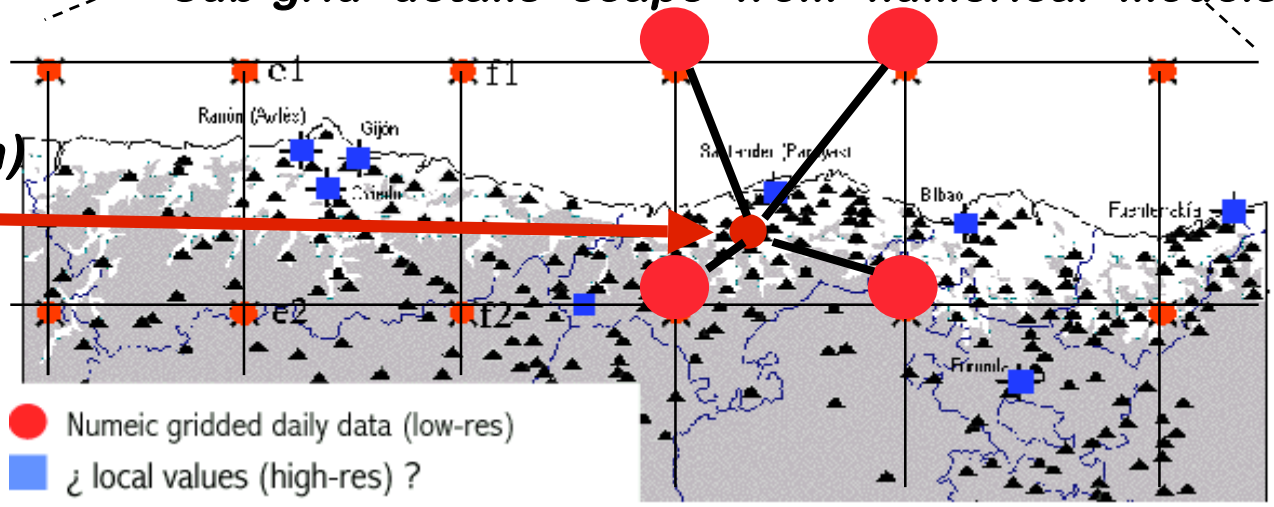
Numeric atmospheric models are the basic tools for operative forecasting.

$$
\begin{cases}
\dfrac{\partial u}{\partial t} = -u\dfrac{\partial u}{\partial x} - v\dfrac{\partial u}{\partial y} - \omega\dfrac{\partial u}{\partial p} - \dfrac{\partial \Phi}{\partial x} + fv + F_x \\
\dfrac{\partial v}{\partial t} = -u\dfrac{\partial v}{\partial x} - v\dfrac{\partial v}{\partial y} - \omega\dfrac{\partial v}{\partial p} - \dfrac{\partial \Phi}{\partial y} - fu + F_y \\
\dfrac{\partial \Phi}{\partial p} = -\dfrac{RT}{p} \\
\dfrac{\partial T}{\partial t} = -u\dfrac{\partial T}{\partial x} - v\dfrac{\partial T}{\partial y} + \omega\left(\dfrac{\kappa T}{p} - \dfrac{\partial T}{\partial p}\right) + \dfrac{\dot{H}}{c_p} \\
\dfrac{\partial \omega}{\partial p} = -\left(\dfrac{\partial u}{\partial x} + \dfrac{\partial v}{\partial y}\right)
\end{cases}
$$



Sub-grid details scape from numerical models

How can we get a prediction for a local point (station) of interest?



- ***Interpolating the gridded predictions.***
- ***Applying statistics to station historical records:***

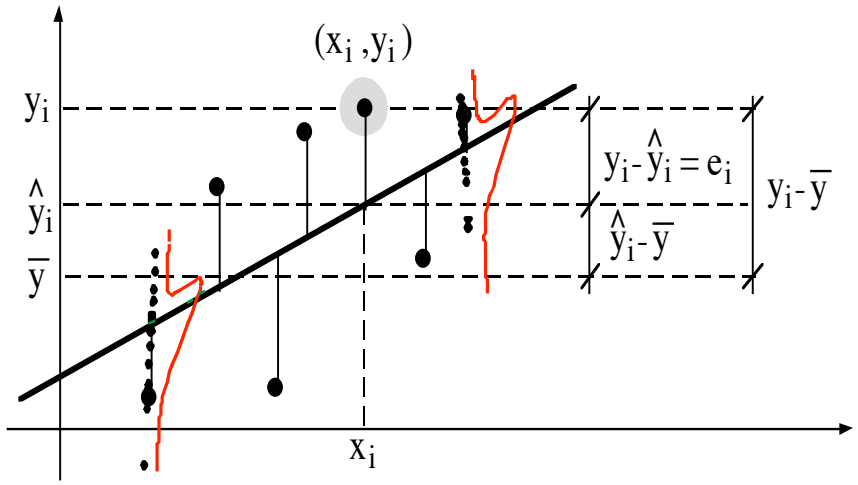
x_1, x_2, x_3, ...

STATISTICAL DOWNSCALING

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Técnicas Clásicas de Predicción

- **Regresión Lineal:** **Y = a + b X**
  - Los parámetros **a** y **b** definen la línea recta que más se ajusta al conjunto de datos disponible; estos parámetros se estiman **directamente** de los datos utilizando el criterio de mínimos cuadrados con los valores conocidos de Y1, Y2, …, X1, X2, ….



- **Regresión Múltiple:** una extensión de la regresión lineal para tratar vectores:

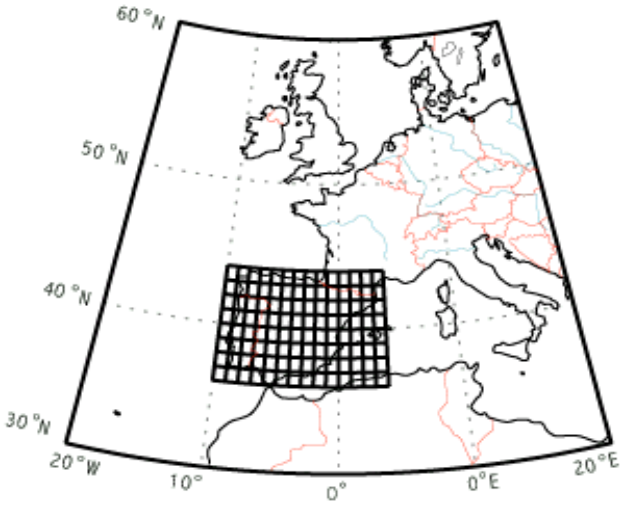
  Y = b0 + b1 X1 + b2 X2.

  Numerosas funciones no lineales puden transformarse en este tipo de ecuación.

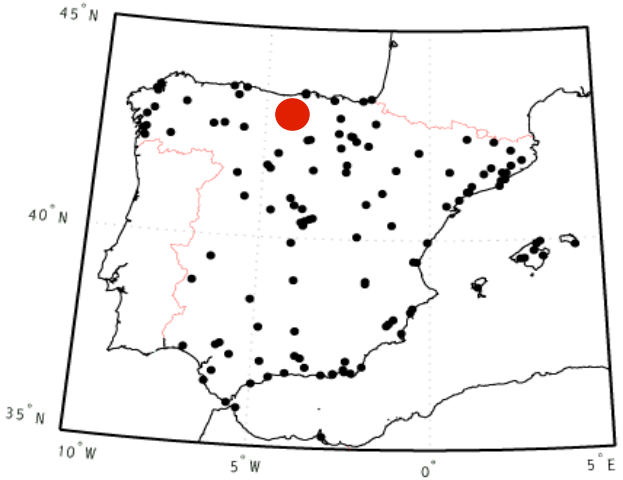
José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Regresión Lineal

Gridded atmospheric circulation patterns for day n



Observations at 122 stations for day n



$(\mathbf{T}(1000\ mb),\ldots,\mathbf{T}(500\ mb);$
$\mathbf{Z}(1000\ mb),\ldots,\mathbf{Z}(500\ mb);$
$\ldots\ldots;$
$\mathbf{H}(1000\ mb),\ldots,\mathbf{H}(500\ mb)) = \mathbf{X}_n$

$Y_n = F(x_n)$

Precipitation
Maximum wind speed
Sun light , ...
Y_n

A linear model $Y_n = W^T X_n$ ***can be obtained by estimating the coefficients*** $W = (w_1,\ldots, w_k)$ *from historical records from a period* $i=1,\ldots,N$ *where both* X_i *and* Y_i *are available.*

Given a gridded forecast X_{n+1} ***an estimation is obtained as:***

$$\hat{Y}_{n+1} = W^T X_{n+1}$$

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Nonlinearities in the relationship $Y_n = f(X_n)$

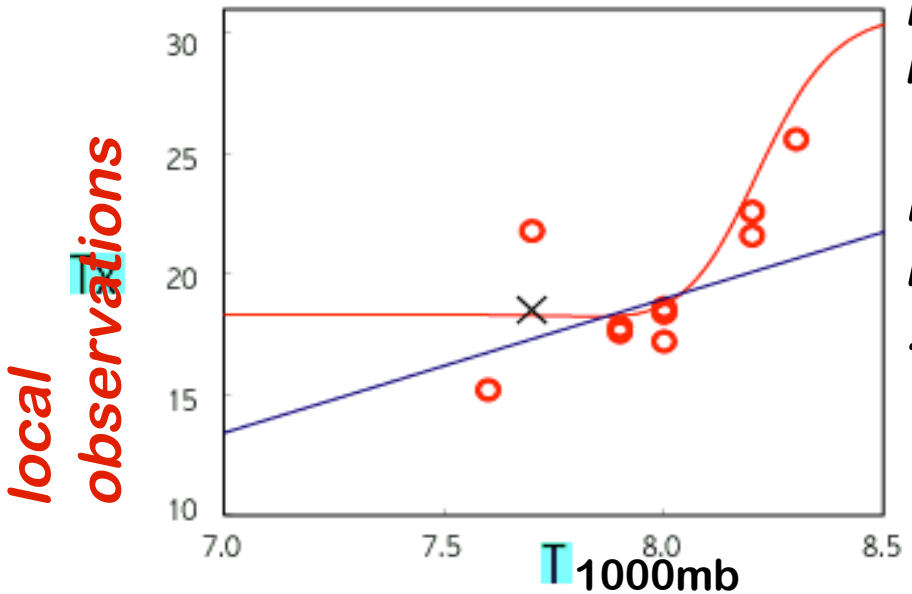
Let us consider a simple case:

$(T(1000\ mb),\ldots, T(500\ mb); Z(1000\ mb),\ldots, Z(500\ mb); \ldots\ldots; H(1000\ mb),\ldots, H(500\ mb)) = X_n$

$TMax_n = w\ T1000_n$ is the simplest case for the relationship

$$TMax_n = f(T1000mb_n),$$

if any, between local observations and gridded predictions.



Nonlinear neural network.

Linear regression .

This example shows how a simple linear model is not appropriate in some cases:

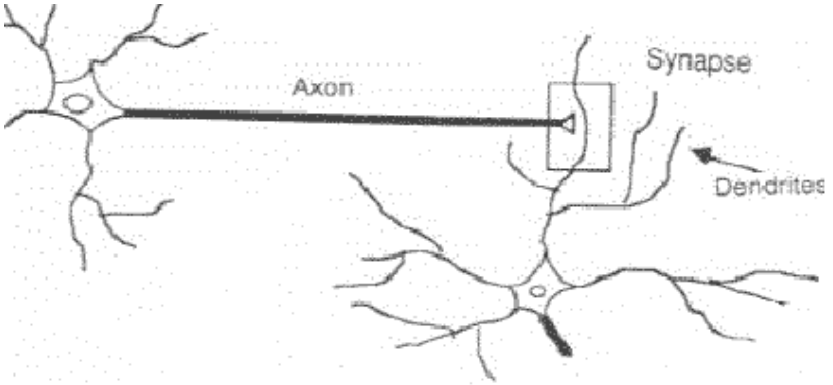
~~$TMax_n = a\ T_n$~~

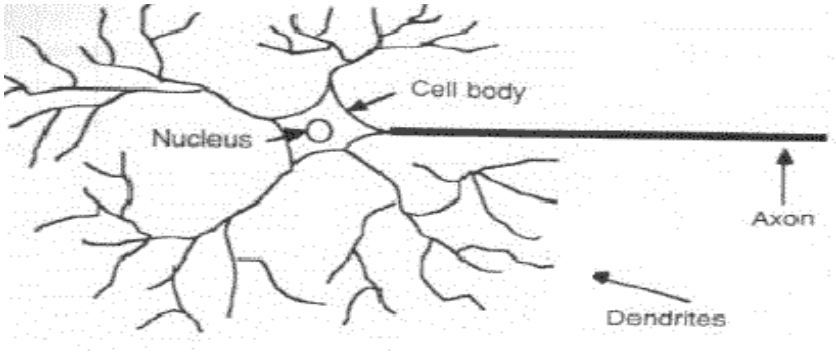
$TMax_n = f(T_n)$

Redes Neuronales. Inspiración en la Neurofisiología

El cerebro humano está formado por un gran número de neuronas (más de 100000 millones) conectadas entre sí de forma masivamente paralela

La actividad de cada neurona se basa en descargas electroquímicas, a partir de los estímulos recibidos por neuronas vecinas a las que esté conectada.





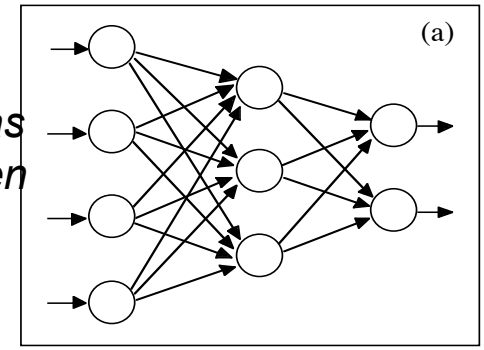
Neural Network Study (1988, AFCEA International Press, p. 60):
... a neural network is a system composed of many simple processing elements operating in parallel whose function is determined by network structure, connection strengths, and the processing performed at computing elements or nodes.

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Redes Neuronales Artificiales

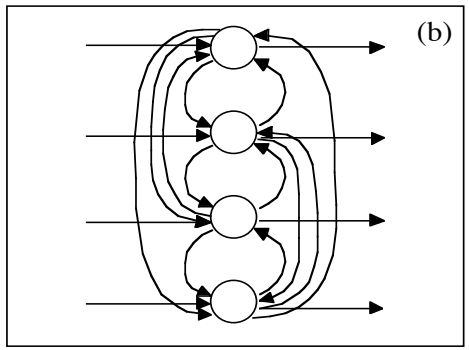
Siguiendo esta metodología se desarrollaron diversas topologías de red, para organizar la conexión de losprocesadores.

Redes Multicapa.
Con varias capas conectadas en serie con procesadores en paralelo.



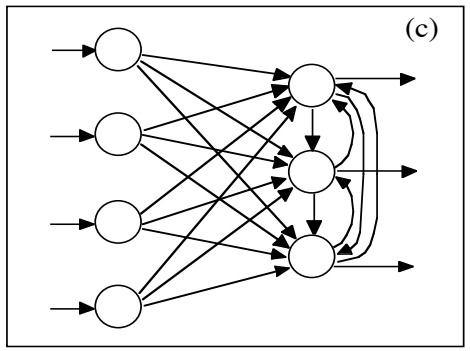
Reconocimiento de patrones
OCR
Proc. lenguaje natural
Interpolación y extrapolación
Ajuste de funciones

Redes de Hopfield.
Una sóla capa totalmente interconectada y paralela.



Reconocimiento de patrones
OCR
Almacenamiento con recuperación robusta al ruido

Redes competitivas.
Redes multicapa con conexiones laterales en la última capa.



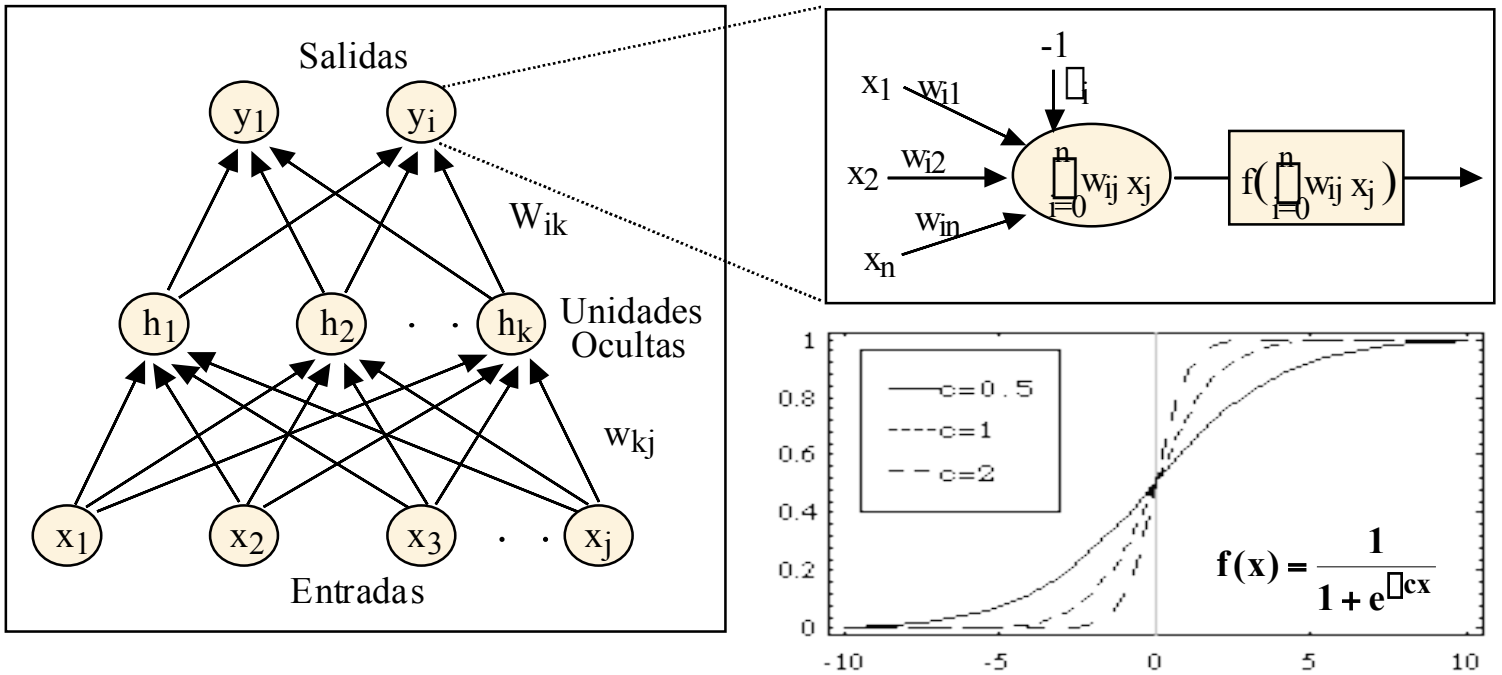
Clasificación autoorganizada
Reconstrucción topológica
Extracción de variables características

José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Redes Multicapa (Perceptrones)

*Las **redes neuronales** permiten obtener una aproximación funcional de un modelo dado en base a un conjunto de datos y a operadores sigmoidales.*



*Cada procesador realiza una actividad muy simple: **combinación lineal** de las actividades recibidas por la neurona **activación * peso**.*

*A continuación, se calcula su actividad aplicando una **función de activación** no lineal. La salida final es un **ajuste no lineal** de las entradas.*

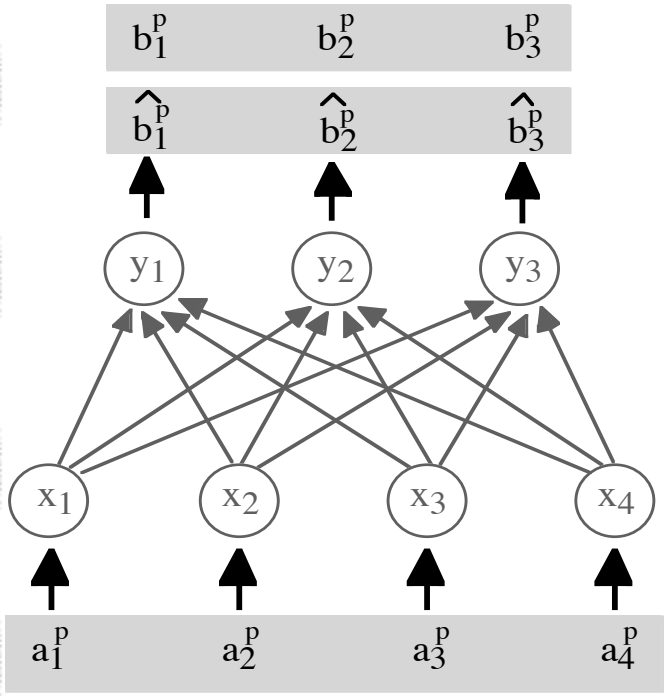
Estimar de los datos

$$y_i = f\left(\sum_k W_{ik}\, f\left(\sum_j w_{kj} x_j\right)\right)$$

José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

El Aprendizaje

¿Cómo se pueden hallar los pesos para que una red obtenga las salidas correctas a partir de las entradas de un conjunto de entradas-salidas dado?



Inicialmente se eligen valores aleatorios para los pesos.

Aprendizaje Hebbiano: Se modican los pesos acorde a la correlación entre las unidades.

Se eligen los patrones (a^p, b^p) de uno en uno y se modifican los pesos w_{ij} de los procesadores con salidas incorrectas:

$$\Delta w_{ij} = \eta (b_i^p - \hat{b}_i^p) a_j^p$$

Descenso de gradiente: Se modican los pesos acorde la dirección del gradiente de una función de error.

$$E(w_{ij}) = \sum_{i,p} (b_i^p - \hat{b}_i^p)^2$$

$$\Delta w_{ij} = -\eta \frac{\partial E(w_{ij})}{\partial w_{ij}} = \eta \sum_p (b_i^p - \hat{b}_i^p) a_j^p$$

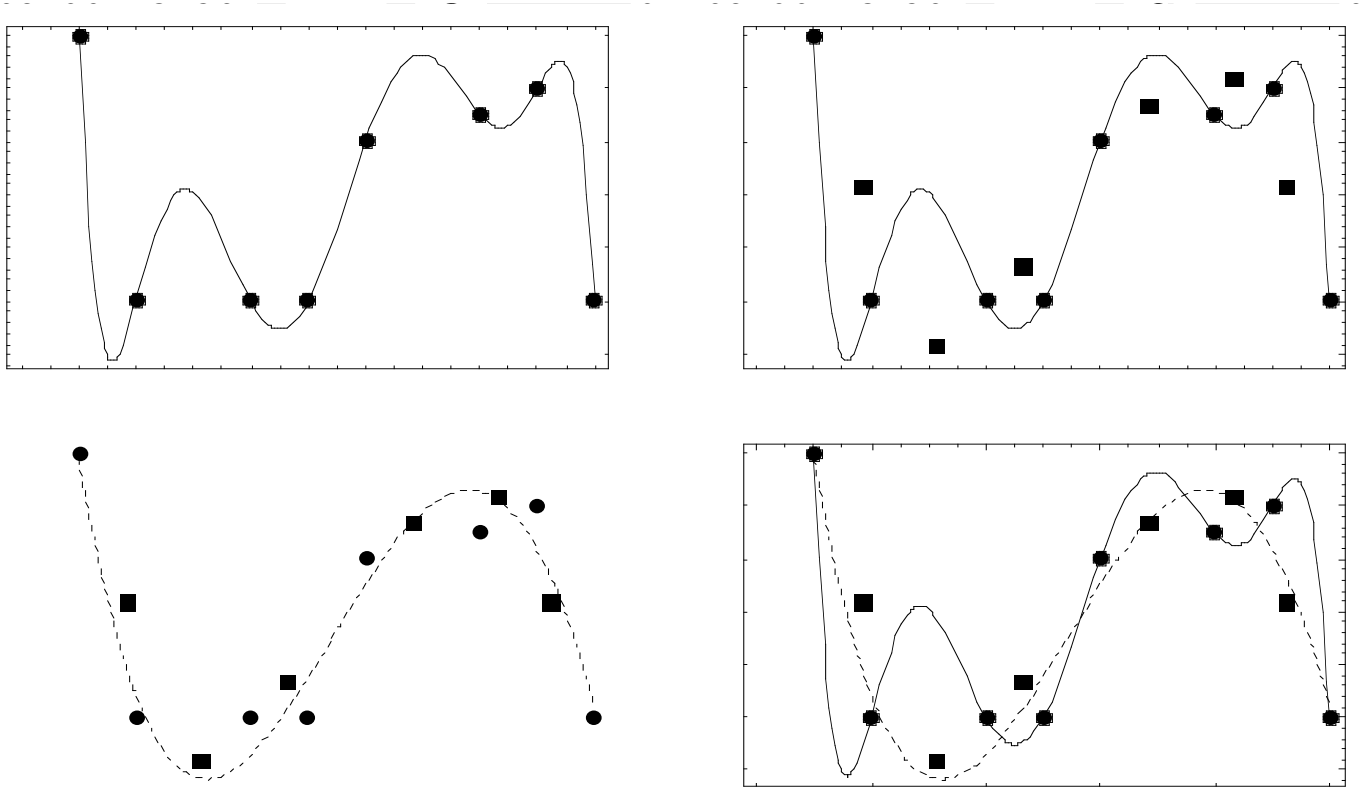
Regularización

$$E(w_{ij}) = \sum_{i,p} (b_i^p - \hat{b}_i^p)^2 + \lambda \sum_{i,j} w_{ij}^2$$

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Sobreajuste (balance incorrecto de parámetros y datos)

*En estadística es bien conocido que cuando se utiliza un modelo con **muchos parámetros** para ajustar un conjunto de datos procedente de proceso con **pocos grados de libertad**, el modelo obtenido puede no descubrir las tendencias reales del proceso original, aunque pueda presentar un error pequeño.*

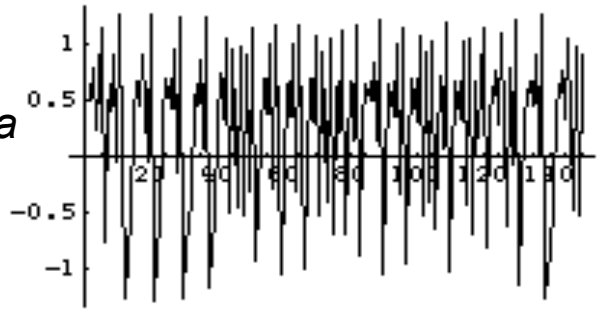


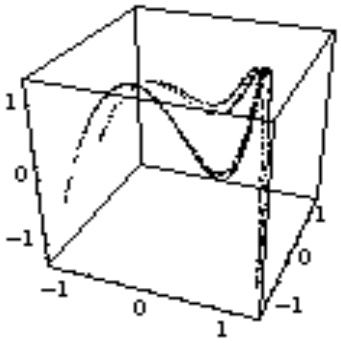
José Manuel Gutiérrez, Universidad de Cantabria. (2001)
http://personales.unican.es/gutierjm

Ejemplo: Predicción No Lineal

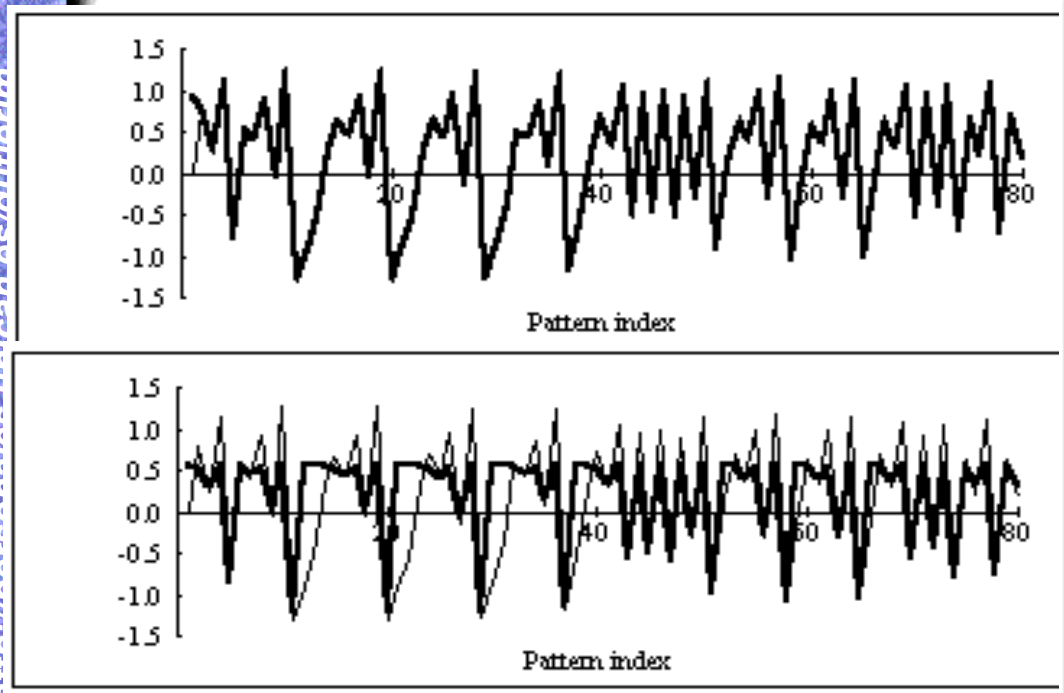
En ocasiones es necesario tratar con series temporales caóticas, que tienen apariencia estocástica y difícilmente precicible.

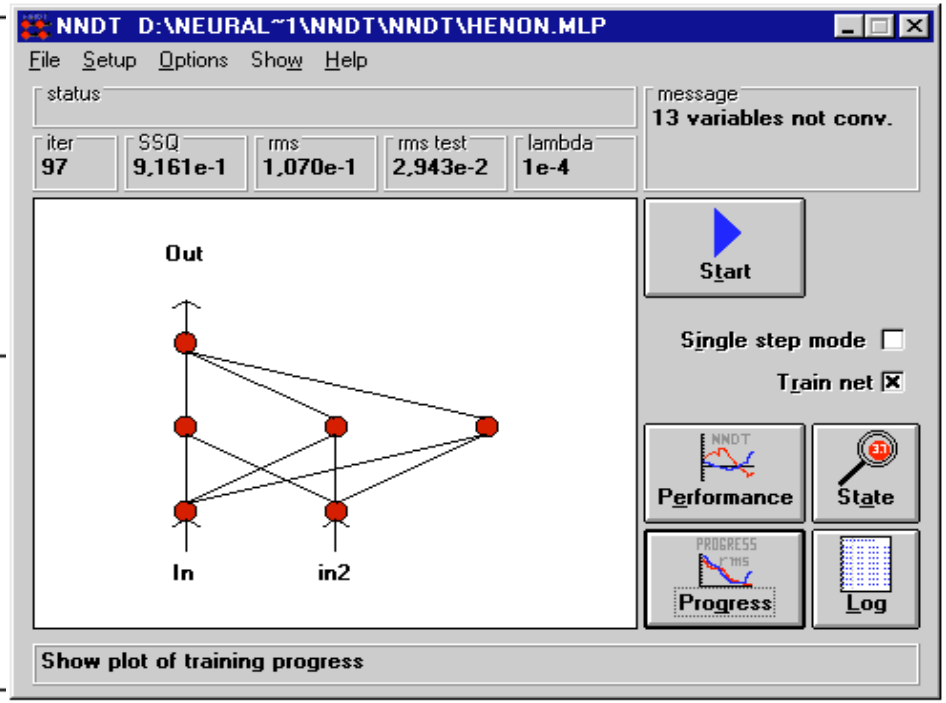
Un ejemplo es la serie de Henon:





$$x_{n+1} = 1 - 1.4\, x_n^2 + 0.3\, x_{n-1}$$

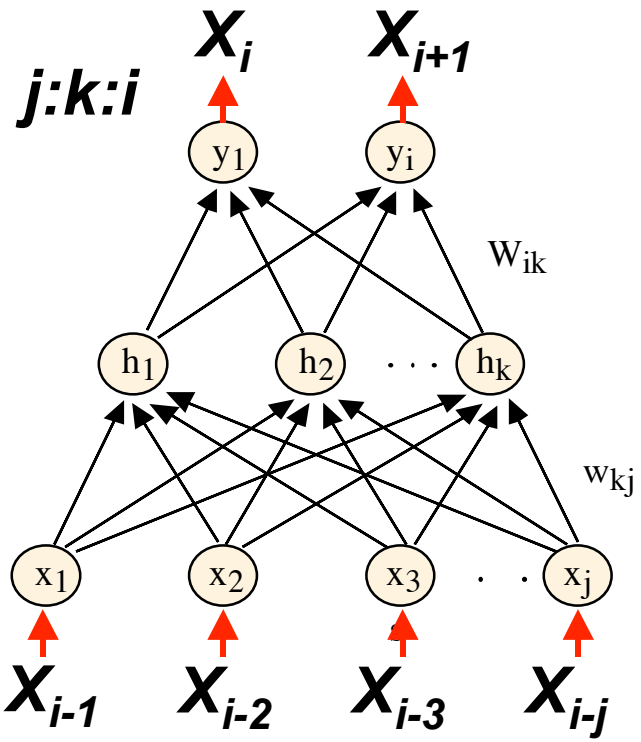


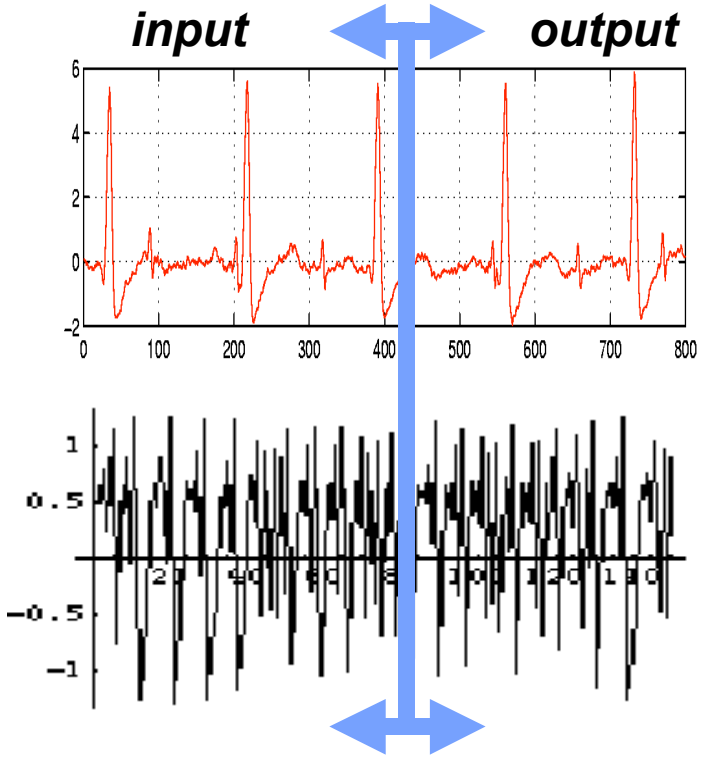


José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Modelización de Series Temporales

*Las **redes neuronales** permiten obtener una aproximación funcional de un modelo dado en base a un conjunto de datos y a operadores sigmoidales.*





Cada procesador realiza una actividad muy simple: combinación lineal de las actividades recibidas por la neurona.

A continuación, se calcula su actividad aplicando una función de activación al valor obtenido (simula el potencial de membrana de una neurona).

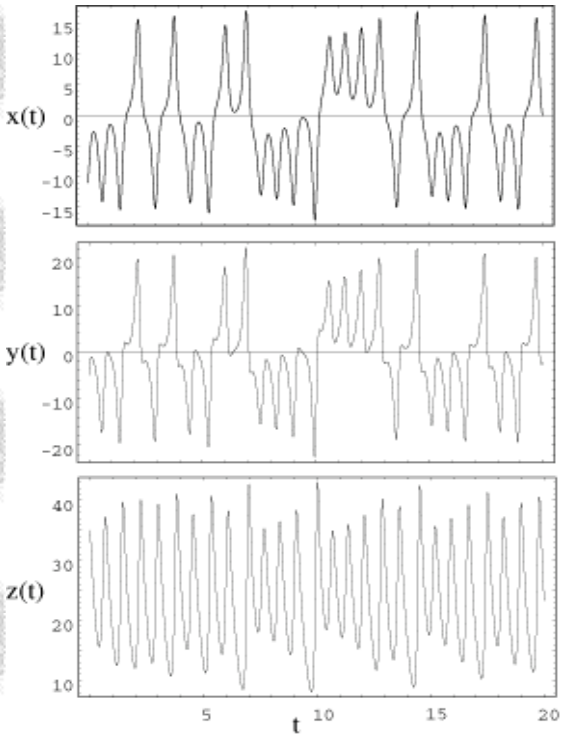
Finalmente, dados los valores de entrada, se obtienen las salidas de la red:

$$y_i = f\left(\sum_k W_{ik}\, f\left(\sum_j w_{kj} x_j\right)\right)$$

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

Modelos de Lorenz Aproximados

Dada una serie temporal de 2000 puntos (t=20), se prueban distintos modelos de red para comprobar el poder modelizador.



Tres variables ***(x,y,z)***

(x_n, y_n, z_n)

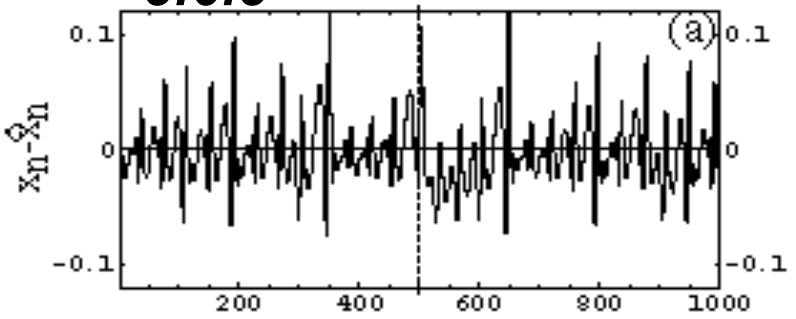
$\downarrow$

$(x_{n+1}, y_{n+1}, z_{n+1})$

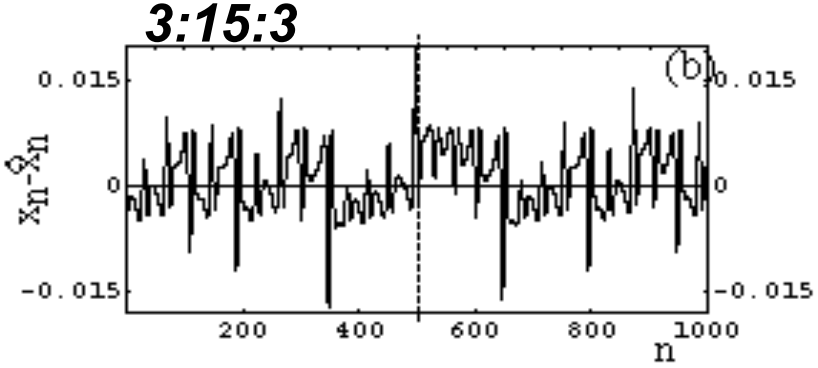
Sistema contínuo ***Red neuronal***

3:k:3

3:6:3



3:15:3



$$\hat{x}_{n+1} = -3768.18 - \frac{0.34}{1+e^{9.31+0.53\,x_n-0.68\,y_n-0.21\,z_n}} + \frac{0.92}{1+e^{7.64-0.121\,x_n-0.149\,y_n-0.13\,z_n}} - \frac{2.75}{1+e^{6.19+0.15\,x_n+0.0451\,y_n-0.09\,z_n}} - \frac{2.04}{1+e^{1.13+0.06\,x_n+0.0119\,y_n-0.06\,z_n}} + \frac{7164.31}{1+e^{-0.12+0.00021\,x_n-0.0002\,y_n+0.000021\,z_n}} - \frac{63.52}{1+e^{-0.24+0.08\,x_n-0.016\,y_n+0.0049\,z_n}},$$

José Manuel Gutiérrez, Universidad de Cantabria, (2001)
http://personales.unican.es/gutierjm

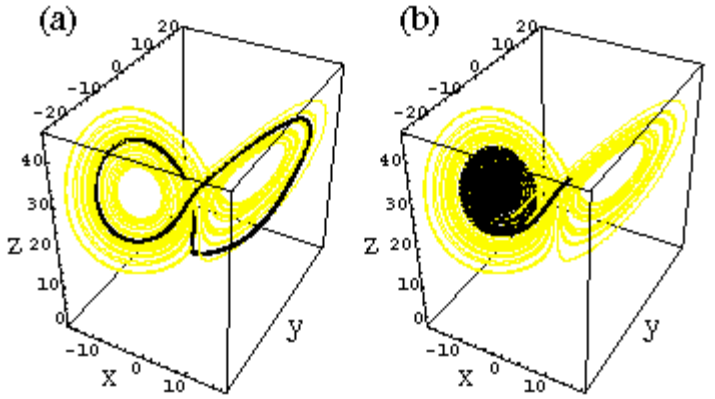
Comportamiento Dinámico

*Un **modelo simple** no captura de forma completa la estructura funcional y, por tanto, no reproduce la dinámica del sistema.*

*Un **modelo muy complicado** se sobreajusta al modelo y no reproduce la dinámica del sistema.*

*Sólo un **modelo intermedio**, con el número apropiado de parámetros puede ajustar apropiadamente la estructura funcional del sistema y, por tanto su dinámica.*

3:6:3



3:15:3