Universidad Técnica Federico Santa María
Departamento de Informática

Capítulo 1:
"Análisis Inteligente de Datos"

II Semestre 2010

Profesor: Héctor Allende (hallende@inf.utfsm.cl)
Página: *www.inf.utfsm.cl/~hallende*

Syllabus AIDA-2010

- **Análisis Inteligente de Datos**
- Profesor encargado: Héctor Allende
- Relatores : Invitados (C. Valle, R. Ñanculef)
- Horario Propuesto Viernes: 6-7 y 8

- **Evaluación:**
- 1 Prueba
- Promedio Tareas
- Proyecto Final

Syllabus AIDA-2010

- **Cálculo de la Nota final: promedio lineal de las calificaciones**

Temas de Proyecto Tentativos :

- **Clasificación automática**
- **Extracción de características**
- **Selección de características**
- **Algoritmos de Asociación**
- **Pronóstico**

Objetivos General

- Al aprobar la asignatura el alumno será capaz de conocer, comprender y aplicar los principios de análisis de datos utilizando técnicas avanzadas de inteligencia computacional, de manera tal que, el descubrimiento de patrones, sirva como sistema de apoyo a la toma de decisiones.

Objetivos: Syllabus AIDA-2010

- El alumno deberá ser capaz de:
- Comprender el papel que juegan los modelos de datos en la ingeniería (ICI)
- Seleccionar y aplicar métodos de aprendizaje estadístico en modelado de datos
- Conocer los principales métodos estadísticos de aprendizaje computacional en clasificación; regresión y pronóstico
- Comprender y dominar las técnicas de extracción de reglas de sistemas de inferencia (clásicos y Borrosos)
- Aplicar los métodos estadísticos computacionales a problemas de modelado que surgen en la ingeniería, Física Biología y Ciencias Sociales)

Contenidos

- Introducción a la Minería de Datos DM
- Conceptos estadísticos del proceso de aprendizaje:
- Técnicas de inteligencia computacional aplicados al análisis de datos: ANN; SVM CPA
- Lógica difusa: Conjuntos fuzzy y lógica fuzzy, Sistemas de Inferencia fuzzy
- Aplicaciones en Clustering ; Clasificación , Regresión y Pronóstico

Bibliografia

- Michael Berthold, David J. Hand, *Intelligent Data Analysis*, Spinger Verlag 2003
- D. Hand, H. Mannila, P. Smyth, *Principles of Data Mining,* The MIT Press 2001
- B.D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press 2000
- C. Bishop *Pattern and Classification*, John Wiley & Sons Inc. 2006

Profesor: H.Allende 7

¿ Qué es un dato?

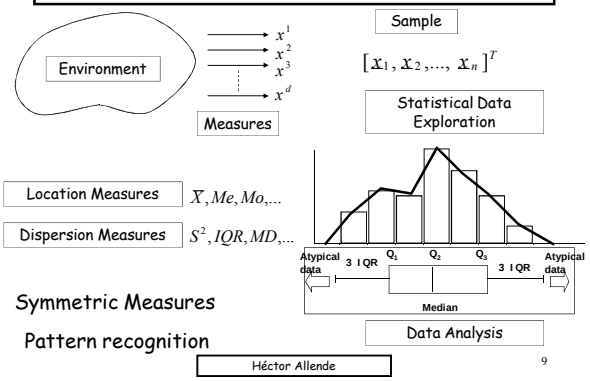
Dato es: Número; Vector ;nombre; cualidad, etc.

Pero también podría ser :

Imágen
Símbolo
Señal acústica
Electrocardiograma
Documento
Función
Matriz
Tensor,
etc.

Profesor: H.Allende

Extracción de Característica



Héctor Allende 9

Análisis exploratorio de datos (EDA)

- Es el proceso de explorar los datos sin tener ideas previas y claras respecto a lo que estamos buscando.
- Técnicas de EDA consisten en calcular una serie de indicadores que resuman y se deriven a partir de los datos. Pueden ser interactivos y visuales.
- Medias ; localización; dispersión ; simetría; forma, etc.

Héctor Allende 10

Análisis exploratorio de datos

- El progreso logrado por la tecnología de hardware permite hoy en día a los sistemas computacionales de almacenamiento que pueden guardar una gran cantidad de datos de alta dimensionalidad.
- Los datos son recolectados porque creemos que son una fuente de información de gran valor, lo que puede proporcionar una ventaja competitiva para cualquier Organización

Héctor Allende 11

Análisis exploratorio de datos

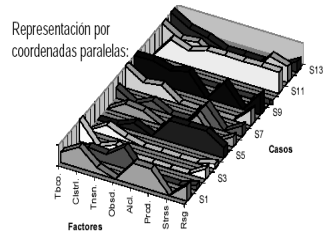
- Encontrar información valiosa escondida en los datos, no es una tarea facil.
- La visualización de la información y análisis de datos visuales, puede ayudar a mejorar la cantidad y calidad de la información.
- Hay un gran numero de técnicas para visualización de la información que han sido desarrolladas en las ultimas décadas para soportar el análisis exploratorio de grandes conjuntos de datos.

Héctor Allende 12

Visualización de datos.

- Ejemplo 1. Dados ciertos atributos de pacientes (tabaquismo, colesterol, tensión, obesidad, alcoholismo, precedentes, estrés) y su riesgo (muy bajo, bajo, medio, alto, muy alto) de enfermedades coronarias:

Tbco.	Clstrl	Tmsn	Obsd	Alcl	Prod	Strs	Rsg
Med.	Alto	8	No	Si	Si	No	Alto
Bajo	Med	9	Si	No	No	No	Bajo
Alto	Bajo	8,5	No	No	No	No	Med
Bajo	Med	7	No	No	No	No	Bajo
Bajo	Bajo	8,5	No	Si	Si	Si	Med
Bajo	Med	9	No	No	Si	No	Med
Med.	Bajo	9	No	No	Si	No	Med
Alto	Med	11	No	No	No	No	Alto
Alto	Alto	13	Si	No	Si	No	M.A.
Bajo	Bajo	7	No	No	No	No	M.B.
Bajo	Alto	12	Si	Si	Si	Si	M.A.
Alto	Med	11	No	No	No	Si	Alto
Alto	Med	8	No	No	No	No	Med



Héctor Allende

13

Densidad de píxeles.

- Ejemplo2: La técnica de los segmentos circulares: La idea central de esta técnica es desplegarlas dimensiones de los datos como segmentos de un circulo.

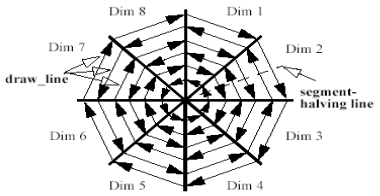


Figure 1: 'Circle Segments' Technique for 8-dimensional Data

14

Preguntas Relevantes del Análisis de Datos

- ¿ Porqué análisis Inteligente de Datos?
- ¿ Existe alguna estructura en los datos?
- ¿Existen datos anómalos (Outliers)?
- ¿Se pueden fusionar (sintetizar) los datos de otra manera más conveniente?
- ¿Se pueden Desagregar los datos de otra manera más conveniente
- ¿Es éste grupo diferente de este otro?
- ¿Este atributo es dinámico (cambia en el tiempo) ?
- ¿Se puede predecir el valor del atributo basado en las mediciones de otros valores ?

Profesor: H.Allende

15

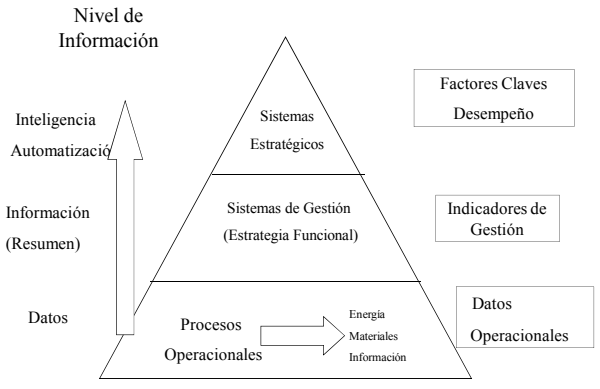
¿ Qué es el análisis de Datos ?

- Área dedicada al estudio sistemático de los datos
  - Transforma datos en información
  - Contribuye al descubrimiento de nuevo conocimiento (KDD), Base de conocimiento
  - Ayuda al Reconocimiento de Patron (PR)
- Papel de las Máquinas de Aprendizaje:
  - Métodos para aprender de los datos
  - Desarrollo de métodos de aprendizaje automáticos y/o semiautomáticos

Profesor: H.Allende

16

Tiempo de respuesta una variable critica



Tipos de datos

- ❖**Estructurados**
  - **•Cuantitativos**
  - **•Cualitativos**
  - **•Simbólicos**
  - **•Ordenados jerárquicamente**
- ❖**Bloques de datos binarios**
  - **Imágenes**
  - **Sonido**
- ❖**No Estructurados**
  - **•Textos**

Profesor: H.Allende

18

Aplicaciones del Análisis de Datos

Problemas que están relacionados con AIDA :

•Identificar un rostro en una imagen

•Convertir un texto hablado en uno escrito

•Establecer un diagnóstico médico a partir de un ECG

-En cada uno de estos problemas tienen propósitos específicos.
-Estos propósitos determinan la forma en que los datos deben ser procesados.

-Esto implica que todo proceso de datos está precedido por un proceso de modelado del problema que necesitamos resolver.

Profesor: H.Allende

Aplicaciones del Análisis de Datos

- Pronóstico de magnitudes máxima de terremotos
- Pronóstico de perspectiva de yacimientos minerales
- Pronóstico de tormentas ionosféricas
- Regionalización sísmica
- Diagnóstico diferencial de enfermedades
- Evaluación de pacientes
- Lectura diagnóstica de señales (EEG, ECG,IC, etc.)
- Clasificación automática de hongos (Bio-lixiviación)
- Clasificación automática de Clientes

Profesor: H.Allende 20

Aplicaciones del Análisis de Datos

- Identificación de huellas digitales
- Identificación de caligrafías
- Identificación de rostros (estáticos, en movimiento, enmascarados, etc.)
- Identificación de interlocutores
- Identificación de objetos mediante sonidos (aviones, vehículos)
- Identificación de objetos mediante rastros (balística)
- Dispositivos de acceso por identificación iriológica
- Reconocimiento de placas de vehículos

Profesor: H.Allende 21

Aplicaciones del Análisis de Datos

- Caracterización socio política de colectivos sociales
- Pronóstico de surgimiento de fenómenos sociales
- Caracterización del modus operandi de un terrorista
- Caracterización del modus operandi de un delincuente
- Análisis de las causas de la delincuencia juvenil (u otro fenómeno social)
- Clasificación jerárquica de delitos
- Evaluación de la gravedad delictiva

Profesor: H.Allende 22

Aplicaciones del Análisis de Datos

En la práctica, casi siempre tenemos que vernos con datos difusos con el propósito de extraer de ellos información que nos sea útil.

En el caso particular del AIDA, aunque no haya una división exacta en el procesamiento desde los datos difusos por un lado hasta las conclusiones por el otro, un modelo útil de AIDA ser dividido en cuatro etapas.

Profesor: H.Allende

Etapas del Análisis Inteligente de Datos

4 Etapas del procesamiento de datos :

1. ***Adquisición***
2. ***Preprocesamiento***
3. ***Representación-descripción de objetos***
4. ***Análisis de datos***

Profesor: H.Allende

Etapas AIDA : *Adquisición*

1. - Adquisición

Este proceso está caracterizado por el hecho que la entrada esta constituida por los datos originales tomados de las fuentes originales y la salida son los datos difusos de ellos podemos extraer información (reglas) que nos puede ser útil.

Profesor: H.Allende

Etapas AIDA : *Adquisición*

La adquisición de datos puede ser tan simple como tomar los datos sin ruidos, limpios, listos para ser procesados.

Observe que en la entrada de esta etapa del proceso tenemos una fuente, por ejemplo un electrocardiógrafo, a partir del cual tomamos una señal, el ECG del paciente, un acelerograma de un sismo, etc.

Esa señal ECG casi siempre tiene ruidos, no está lo suficientemente limpia por lo que no es siempre posible la lectura de lo queremos extraer sin errores.

Profesor: H.Allende

Etapas AIDA : *Preprocesamiento*

2.- Preprocesamiento: La etapa del *Preprocesamiento* está caracterizada por el hecho que ambas, la entrada y la salida son datos de la misma naturaleza, es decir, significa casi la misma cosa.

Por ejemplo, ambas son señales, imágenes, jeroglíficos, matrices, n-tuplas de valores de un cierto rasgo o caracteristica etc.

Profesor: H.Allende

Etapas AIDA : *Preprocesamiento*

Filtrado de señales o imágenes, incrementar la resolución o el contraste de una imagen, restaurar una imagen, eliminarle el ruido, ajustar los datos de una variable, validar los datos, escalarlos, transformarlos etc. Son ejemplos de procesamiento de datos.

Observe que en la entrada de esta estapa tenemos por ejemplo una señal, el ECG de un paciente, y en la salida casi que el mismo ECG sólo que quizás, con menos ruidos, más limpio, más claro, en el cual es más simple leer la información relevante que estamos buscando.

Profesor: H.Allende

Etapas AIDA: *Representación - descripción*

3. Representación-descripción: En esta etapa los datos originales previamente pre-procesados son transformados en una nueva forma que es la adecuada para el procesamiento posterior.

Esta etapa está caracterizada por el hecho que las entrada y salida son diferentes, al menos en su significado. Este es un proceso en el que los objetos involucrados en los datos originales son descritos.

Profesor: H.Allende

Etapas AIDA: *Representación - descripción*

- Segmentación (particionar la imagen o la señal en regiones similares disjuntas)
- Selección de Características
- Representación de una imagen mediante wavelets
- Representación de una imagen mediante una matriz digital
- Representación de la voz mediante una señal de audio

Son ejemplos de representación-descripción de datos.

Profesor: H.Allende

30

Etapas AIDA: *Representación - descripción*

Por ejemplo, una Imagen la podemos describir en términos complejos como una señal bidimensional, en diferentes formatos : Amplitud, Intensidad , Compleja, Polarimétrica etc, y podemos decir que el Imagen de intensidad es normal, pero que la Imágen de amplitud, no lo es etc.

En este caso, la salida es una secuencia de atributos de la mencionada Imagen.

Profesor: H.Allende

PA Etapas AIDA: Análisis

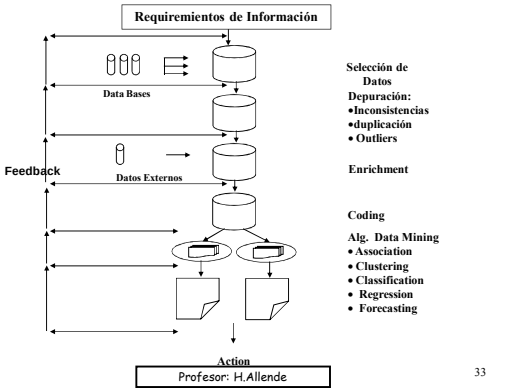
4. Análisis

Finalmente, la etapa del *análisis* es un proceso en el cual encontramos el significado de los datos originales o al menos a una parte de ellos.

Podemos reconocer la ocurrencia de cierta asociación, correlación causalidad a partir de las representaciones previamente almacenada y podemos tomar una decisión, extarer una regla sacar una conclusión.

Profesor: H.Allende

Etapas del Análisis Inteligente de Datos



Profesor: H.Allende

33

▪Problemas : Análisis Inteligente de Datos

Problemas Tipos son :
de Asociación y/o correlación
de causalidad
de Interpretación,
de Caracterización,
de Clasificación,
de Clusterización
de Reconocimiento
de Pronóstico

Estos son ejemplos de análisis de datos.

Profesor: H.Allende

Problemas : Análisis Inteligente de Datos

En el caso de la señal ECG podemos determinar la normalidad del paciente desde el punto de vista del estado de su sistema cardiovascular, si nosotros tenemos el suficiente conocimiento de Cardiología (base de conocimiento.

En el caso de una imagen, podemos reconocer , personas examinando sus rostros, incluso si tuviésemos suficiente conocimiento previo, pudiéramos autentificar a cada una de esas personas.

Profesor: H.Allende

Problemas: Análisis Inteligente de Datos

Basado en el análisis de inteligente datos podemos reconocer *patrones* de una cierta base de datos en particular con propósitos diversos :

- Clasificar,
- Caracterizar,
- Diagnósticar,
- Pronósticar,
- Descubrir la génesis de un fenómeno

Profesor: H.Allende

- AIDA es una disciplina con un marcado carácter aplicado e interdisciplinario, que tiene que ver con la Ingeniería, la Estadística y la Ciencia Computación para el procesamiento de datos acerca:

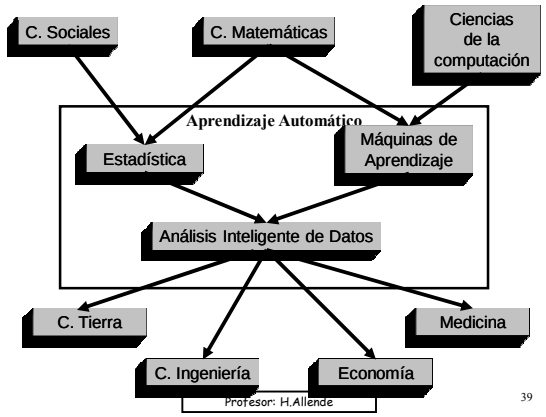
  Objetos fisicos (fotos, escrituras, jeroglíficos, símbolos, señales etc.) y/o

  Objetos abstractos (vectores de un cierto producto Cartesiano de conjuntos de ciertos tipos: duros, difusos, ruidosos)

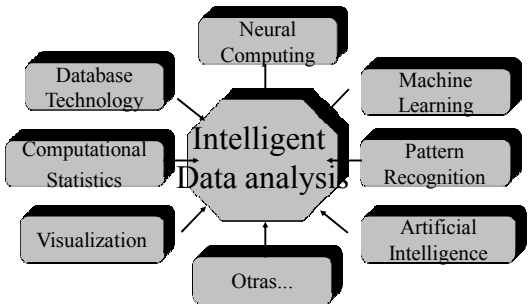
  Propósito mediante algoritmos obtener la información relevante y no evidente *que nos permita establecer propiedades (reglas) de ciertos subconjuntos objetos.*

AREAS : Análisis Inteligente de Datos

- Estadística
- *Ciencias de la Computación*
- Procesamiento de Señales
- Visión por Computacional
- Morfología Matemática
- Reconocimiento de patrones
- Máquinas de Aprendizaje Computacional
  -Redes Neuronales artificiales
  -Máquinas de soporte vectorial
  - Arboles de clasificación
- Entre otras.



Convergencia de Múltiples Disciplinas



Inteligencia Computacional

- Máquinas de Aprendizaje:
  - Capacidad del computador para aprender de la experiencia, es decir, modificar su procesamiento en base a la nueva información adquirida. (Oxford English Dictionary).
  - Proceso que causa que el sistema mejore con la experiencia (Mitchell 1997).
  - Uso de algoritmos computacionales para aprender de los datos. (Hutchinson 1995).
  - Programa de computación que puede aprender de la experiencia respecto a algún tipo de tarea y medida de desempeño (Mitchell 1997).

Aprendizaje

- Consiste en inducir funciones generales a partir de un conjunto específico de formas denominado patrones de entrenamiento

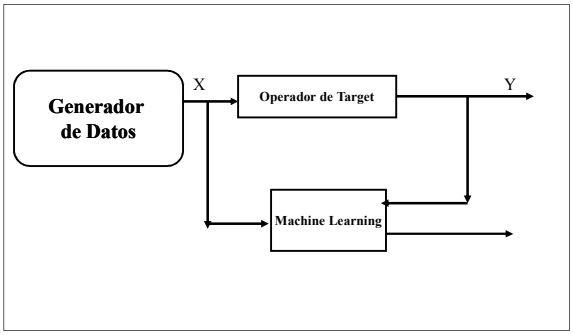
- Tipos de Aprendizaje:
  - Aprendizaje Supervisado
  - Aprendizaje Reforzado
  - Aprendizaje No-supervisado
  - Aprendizaje Semi-supervisado

Data – Text- Music – Opinion: Mining

- "Etapa de reconocimiento de patrones, a través de algoritmos automáticos o semiautomáticos de grandes bases de datos con el objeto de apoyar a la toma de decisiones dentro de una organización".
- Bigus(1996) : DM es el descubrimiento eficiente de información valiosa (nuevos hechos y relaciones) no evidente desde una gran base de datos.

Profesor: H.Allende 43

Machine Learning



Profesor: H.Allende 44

Estadística v/s Máquinas de Aprendizaje

- Estadística moderna → Modelo
- Máquinas de Aprendizaje → Algoritmos
- Modelo:
  - Estructura propuesta, o una aproximación a una estructura de la cual se obtuvieron los datos.
  - Los Modelos pueden ser Empíricos o Mecanicistas.
    - Modelos Empíricos: Busca modelar las relaciones sin basarlas en alguna teoría subyacente.
    - Modelos Mecanicistas: Se construyen en base a algún mecanismo supuesto del proceso de generación de los datos.
  - Los Modelos pueden ser de Explicativos o de Pronóstico

Profesor: H.Allende 45

Modelos v/s Patrones

- Modelo:
  - Consiste en una estructura en gran escala que resume las relaciones sobre muchos casos.
- Patrón:
  - Consiste en una estructura local satisfecha por algunos pocos casos o en una pequeña región del espacio de los datos.

Profesor: H.Allende 46

Análisis de Datos

- Es el proceso de calcular varios resúmenes y valores derivados a partir de una colección de datos.
- La falsedad de la "Receta de Cocina" (Cookbook)
  - Las herramientas del análisis de datos poseen relaciones complejas.
  - Rara vez una pregunta de investigación es estipulada de manera precisa de manera tal que una aplicación simple y única de algún método será suficiente.
- El análisis de Datos es un proceso Iterativo.
  - Los datos se analizan utilizando herramientas analíticas, y modificando los procesos transformando o particionando y repitiendo estos procesos.

Profesor: H.Allende 47

¿Por qué Inteligente?

- Puesto que para poder extraer la estructura que subyace en los datos, entender lo que está sucediendo, aplicar en forma reiterada diversos métodos, refinar las preguntas que el investigador trata de responder requiere de mucho cuidado y sobre todo ***INTELIGENCIA***.
- El análisis inteligente de datos no es un método poco sistemático de aplicación de las herramientas análisis de datos y de reconocimiento de formas (Data Mining), no es un paseo aleatorio a través del espacio de las técnicas analíticas, sino que ***"Un proceso cuidadosamente planeado para decidir lo que será más útil y revelador."***

Profesor: H.Allende 48

Herramientas : Análisis de Datos

- Durante el transcurso del curso se analizarán una serie de técnicas modernas:
  - Modelos Bayesianos
  - Métodos de Kernel y Máquinas de vector soporte
  - Series Temporales, Cadenas de Markov ocultas
  - Reglas de Inducción
  - Redes Neuronales Artificiales
  - Lógica Difusa
  - Métodos de Búsqueda Estocástica
  - Visualización

Profesor: H.Allende 49

INTRODUCCIÓN

- Hace 30 años George BOX
  Creador del famoso modelo de pronóstico de Series de Tiempo (BOX and JENKINS) proclamó:

"Todos los modelos son erroneos, pero algunos son útiles". En mi opinión, tenía algo de razón.

- ¿Qué elección que tenemos en la era de los Petabytes?

- CORRELACION v/s CAUSALIDAD

Profesor: H.Allende 50

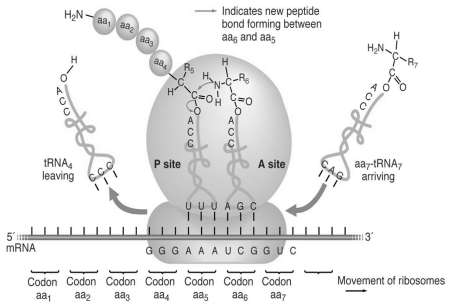
INTRODUCCIÓN

Fisicalismo: Sólo las teorías y modelos desde la cosmológica cuantitativa, parecen estar disponibles para explicarlos fenómenos del mundo que nos rodea (Fisicalismo nueva religión científica F. VARELA)

Biologísmo: Parece ser que la biología va en la misma dirección. Nosotros estudiamos en la escuela secundaria los genes recesivos y dominantes, genes direccionados por el paradigma mendeliano, el que debió ajustarse a los mecanismos de relojería de las Leyes fisicas de Newton.

Profesor: H.Allende 51

Traducción



Profesor: H.Allende 52

Los científicos están entrenados para reconoce que la correlación (estadística) no es casualidad, que no nos podemos conducir por un simple gráfico de correlación entre los patrones X e Y (éstos podrían, incluso, llegar a ser coincidentes), nuestra tarea es descubrir el mecanismo que subyace y que conecta ambos patrones.

A menudo, nosotros tenemos un modelo e, incluso, nosotros podemos conectar bases de datos con cierto grado de confianza, pero el mundo no es lineal, como muestra la racionalidad.

Profesor: H.Allende 53