# Diagramas de Conceptos Formales

## Del álgebra a la minería de datos, pasando por la intuición visual

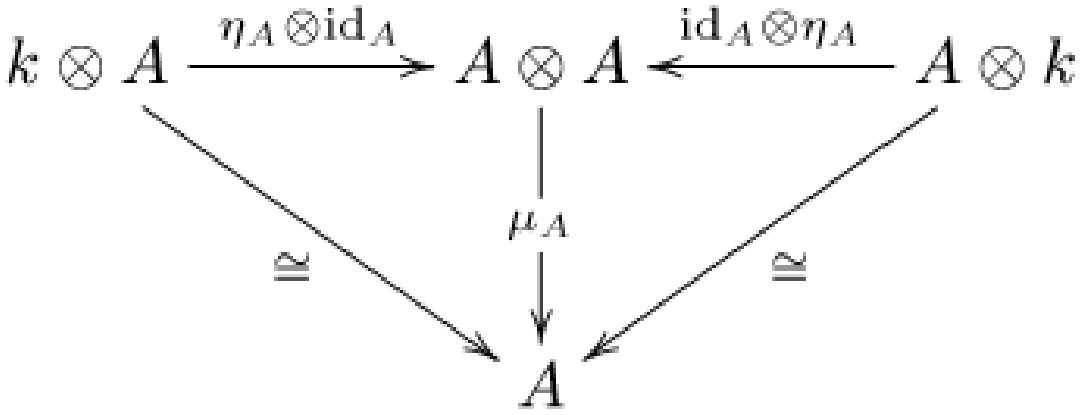
José L. Balcázar

Departamento de Matemáticas, Estadística y Computación
Universidad de Cantabria

Taller "Matemáticas en Acción", 18 de Febrero de 2009

# Diagramas de Conceptos Formales

## Nuestro paseo de hoy




Del álgebra...

...a la minería de datos...

...¿y la intuición visual?

# Álgebra

## ¡En su origen fue otra cosa!

"En esto fueron razonando los dos, hasta que llegaron a un pueblo donde fue ventura hallar un algebrista [...]"

Miguel de Cervantes Saavedra

El ingenioso hidalgo Don Quijote de la Mancha, II parte, cap. XV

# Álgebra

## ¡En su origen fue otra cosa!

"En esto fueron razonando los dos, hasta que llegaron a un pueblo donde fue ventura hallar un algebrista [...]"

Miguel de Cervantes Saavedra
El ingenioso hidalgo Don Quijote de la Mancha, II parte, cap. XV

La palabra "álgebra" [...] (al-dejaber), proviene por lo tanto del árabe y significa "reducción", operación de cirugía por la cual se reducen los huesos luxados o fraccionados [...]

Wikipedia

# Álgebra

## ¡En su origen fue otra cosa!

"En esto fueron razonando los dos, hasta que llegaron a un pueblo donde fue ventura hallar un algebrista [...]"

Miguel de Cervantes Saavedra
El ingenioso hidalgo Don Quijote de la Mancha, II parte, cap. XV

La palabra "álgebra" [...] (al-dejaber), proviene por lo tanto del árabe y significa "reducción", operación de cirugía por la cual se reducen los huesos luxados o fraccionados [...]

Wikipedia

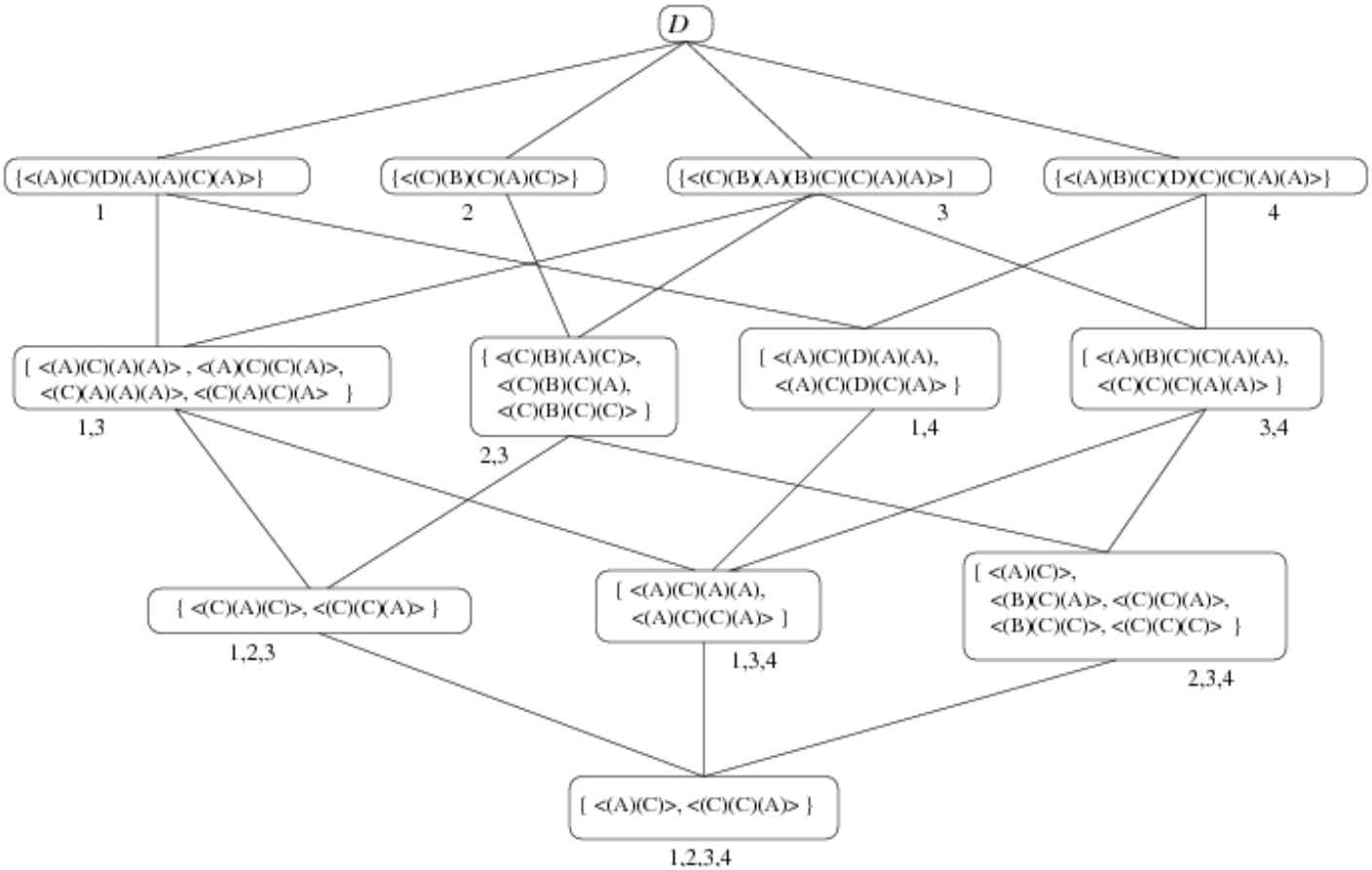
Actualmente, el álgebra es la rama de las matemáticas que estudia las estructuras, relaciones y cantidades [...]

Wikipedia

# Álgebra en Minería de Datos

## Retículos, esencialmente

En nuestro caso, las estructuras a que nos referiremos se conocen como retículos.



# Minería de Datos

## Data Mining, Knowledge Discovery

> Bajo el nombre de minería de datos se engloba todo un conjunto de técnicas encaminadas a la extracción de conocimiento procesable, implícito en las bases de datos. [...] Las bases de la minería de datos se encuentran en la inteligencia artificial y en el análisis estadístico.
>
> Wikipedia

# Minería de Datos

## Data Mining, Knowledge Discovery

Bajo el nombre de minería de datos se engloba todo un conjunto de técnicas encaminadas a la extracción de conocimiento procesable, implícito en las bases de datos. [...] Las bases de la minería de datos se encuentran en la inteligencia artificial y en el análisis estadístico.

Wikipedia

Algunas de las definiciones existentes especifican tres propiedades deseables sobre la información obtenida:

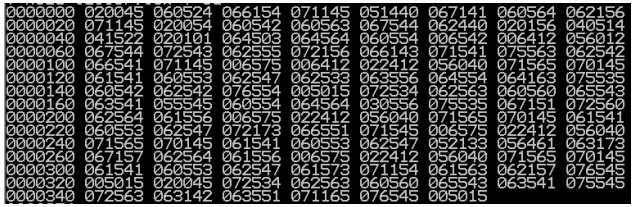
### Deseamos identificar hechos:

- correctos,
- novedosos o previamente desapercibidos,
- "accionables", que permiten generar nuevas actuaciones.

# Sofisticación conceptual

## ¿Qué busca un proyecto de Minería de Datos?

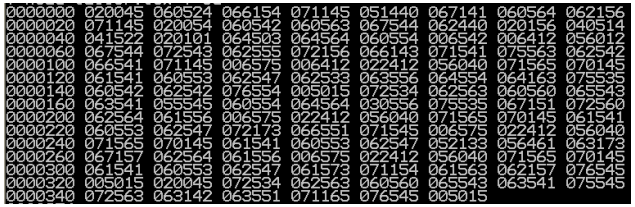
Datos?...

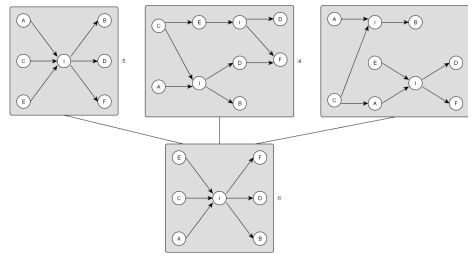


# Sofisticación conceptual

## ¿Qué busca un proyecto de Minería de Datos?

Datos?...



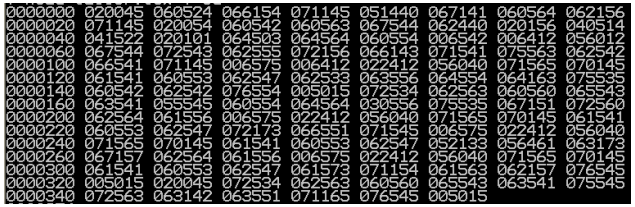



...Información?...

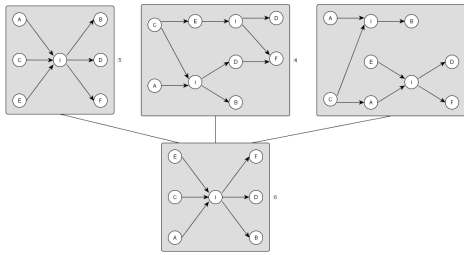
# Sofisticación conceptual

## ¿Qué busca un proyecto de Minería de Datos?

Datos?...




...Información?...

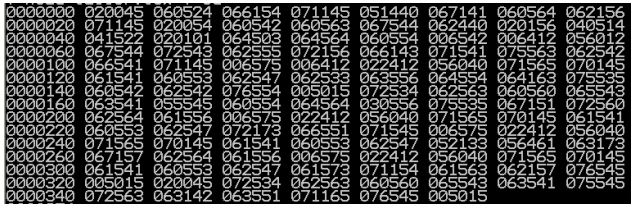


...Conocimiento?...

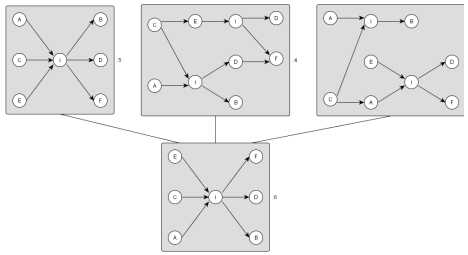
# Sofisticación conceptual

## ¿Qué busca un proyecto de Minería de Datos?

Datos?...

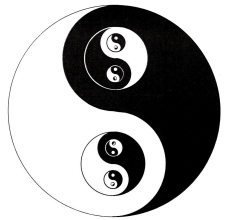



...Información?...



...Conocimiento?...

...Sabiduría!?



# No siempre es posible

## ¿Procesamos datos o torturamos datos?

La validez estadística de la “información” obtenida no siempre se analiza con el cuidado suficiente.

# No siempre es posible

## ¿Procesamos datos o torturamos datos?

La validez estadística de la “información” obtenida no siempre se analiza con el cuidado suficiente.

Cada posible hipótesis en juego requiere disponer de más y más datos para validar el proceso.
Sin embargo, en la práctica hay que conformarse con los que se tengan.

# No siempre es posible

## ¿Procesamos datos o torturamos datos?

La validez estadística de la "información" obtenida no siempre se analiza con el cuidado suficiente.

Cada posible hipótesis en juego requiere disponer de más y más datos para validar el proceso.
Sin embargo, en la práctica hay que conformarse con los que se tengan.

### Un malentendido habitual:

Si tenemos pocos datos, deduciremos de ellos menos información.

# No siempre es posible

## ¿Procesamos datos o torturamos datos?

La validez estadística de la "información" obtenida no siempre se analiza con el cuidado suficiente.

Cada posible hipótesis en juego requiere disponer de más y más datos para validar el proceso.
Sin embargo, en la práctica hay que conformarse con los que se tengan.

### Un malentendido habitual:

Si tenemos pocos datos, deduciremos de ellos menos información.

¡No es así!

# No siempre es posible

## ¿Procesamos datos o torturamos datos?

La validez estadística de la "información" obtenida no siempre se analiza con el cuidado suficiente.

Cada posible hipótesis en juego requiere disponer de más y más datos para validar el proceso.
Sin embargo, en la práctica hay que conformarse con los que se tengan.

### Un malentendido habitual:

Si tenemos pocos datos, deduciremos de ellos menos información.

### ¡No es así!

Si tenemos pocos datos, deduciremos de ellos más información.

# No siempre es posible

## ¿Procesamos datos o torturamos datos?

La validez estadística de la "información" obtenida no siempre se analiza con el cuidado suficiente.

Cada posible hipótesis en juego requiere disponer de más y más datos para validar el proceso.
Sin embargo, en la práctica hay que conformarse con los que se tengan.

### Un malentendido habitual:

Si tenemos pocos datos, deduciremos de ellos menos información.

### ¡No es así!

Si tenemos pocos datos, deduciremos de ellos más información.
Pero será más falsa.

# Pero tenemos algoritmos muy astutos

Theorem: There is no free lunch

## Otro malentendido frecuente:

Lo importante es inventar el algoritmo adecuado para extraer la información.

# Pero tenemos algoritmos muy astutos

## Theorem: There is no free lunch

### Otro malentendido frecuente:

Lo importante es inventar el algoritmo adecuado para extraer la información.

¡No es así!

# Pero tenemos algoritmos muy astutos

## Theorem: There is no free lunch

### Otro malentendido frecuente:

Lo importante es inventar el algoritmo adecuado para extraer la información.

### ¡No es así!

"There are circumstances in which the outputs of all procedures solving a particular type of problem are statistically identical". ("No free lunch theorem", Wolpert (1996), en: The Lack of A Priori Distinctions between Learning Algorithms.)

# Pero tenemos algoritmos muy astutos

## Theorem: There is no free lunch

### Otro malentendido frecuente:

Lo importante es inventar el algoritmo adecuado para extraer la información.

### ¡No es así!

"There are circumstances in which the outputs of all procedures solving a particular type of problem are statistically identical". ("No free lunch theorem", Wolpert (1996), en: The Lack of A Priori Distinctions between Learning Algorithms.)

### Pero no seamos tan pesimistas.

Conservemos un poco de optimismo, trabajemos duro y veamos a dónde somos capaces de llegar.

# Causalidad

## La relación causa-efecto como proceso cognitivo

La generalidad de los humanos nos encontramos casi permanentemente “detectando” relaciones de causalidad.

# Causalidad

## La relación causa-efecto como proceso cognitivo

La generalidad de los humanos nos encontramos casi permanentemente "detectando" relaciones de causalidad.

- Nos apoyamos en ellas para estructurar nuestras percepciones.
- Nos apoyamos en ellas para generar nuestras expectativas.
- Un motor para la Ciencia: no basta con predecir, hay que construir teorías que expliquen relaciones de causa y efecto entre los hechos.

  > "5.1361 Der Glaube an den Kausalnexus ist der Aberglaube"
  > (LW, T L-Ph)

- La piedra angular de la Lógica, y el gran avance de Aristóteles sobre Platón: 2350 años de estudio.

# Un estudio computacional

## Recursos de cómputo

Allá donde exista un proceso de información, podemos ver un cálculo o una computación.

Allá donde haya una computación, podemos medir los recursos computacionales que consume:

- número de operaciones elementales (tiempo de cálculo),
- cantidad de memoria requerida.

Problemas indecidibles: nuestra noción de cálculo no permite resolverlos.

Problemas no factibles: existen cálculos que los resuelven, pero requieren recursos excesivos.

# Intuiciones de amplia aplicación

## Ejemplo: la habitación china

Encerrado en tu habitación, haces creer que sabes chino:

Te van pasando por una rendija hojas con preguntas en chino.

Tú respondes con la respuesta correcta.

# Intuiciones de amplia aplicación

## Ejemplo: la habitación china

Encerrado en tu habitación, haces creer que sabes chino:

Te van pasando por una rendija hojas con preguntas en chino.

Tú respondes con la respuesta correcta.

(Sí, es el Turing Test.)

# Intuiciones de amplia aplicación

## Ejemplo: la habitación china

Encerrado en tu habitación, haces creer que sabes chino:

Te van pasando por una rendija hojas con preguntas en chino.

Tú respondes con la respuesta correcta.

(Sí, es el Turing Test.)

¡Pero, en realidad, lo que tienes es una pila de diccionarios en los que buscas las respuestas!

# Intuiciones de amplia aplicación

## Ejemplo: la habitación china

Encerrado en tu habitación, haces creer que sabes chino:

Te van pasando por una rendija hojas con preguntas en chino.

Tú respondes con la respuesta correcta.

(Sí, es el Turing Test.)

¡Pero, en realidad, lo que tienes es una pila de diccionarios en los que buscas las respuestas!

### La Teoría de la Complejidad plantea:

Sí, pero, ¿cuánto te ocupan los diccionarios?

# Intuiciones de amplia aplicación

## Ejemplo: la habitación china

Encerrado en tu habitación, haces creer que sabes chino:

Te van pasando por una rendija hojas con preguntas en chino.

Tú respondes con la respuesta correcta.

(Sí, es el Turing Test.)

¡Pero, en realidad, lo que tienes es una pila de diccionarios en los que buscas las respuestas!

### La Teoría de la Complejidad plantea:

Sí, pero, ¿cuánto te ocupan los diccionarios?

Quédate con un ejemplo: ¿Qué animal es el más grande de entre diez, tomados de una lista de cien?

# Aplicación a la causalidad

## La lógica y la computación están muy próximas

Comprobar si una afirmación es consecuencia de otra,

- en cualquier lógica "razonable" (como primer orden), es desafortunadamente indecidible;
- incluso en lógicas muy "limitadas" (como la proposicional) es desafortunadamente intratable.

### ¿Qué podemos hacer?

Si nos limitamos un poco más aún, encontramos una lógica que es, poco más o menos, "lo más expresiva posible" bajo la condición de tratabilidad de la noción de consecuencia lógica.

# Aplicación a la causalidad

## La lógica y la computación están muy próximas

Comprobar si una afirmación es consecuencia de otra,

- en cualquier lógica "razonable" (como primer orden), es desafortunadamente indecidible;
- incluso en lógicas muy "limitadas" (como la proposicional) es desafortunadamente intratable.

### ¿Qué podemos hacer?

Si nos limitamos un poco más aún, encontramos una lógica que es, poco más o menos, "lo más expresiva posible" bajo la condición de tratabilidad de la noción de consecuencia lógica.

Se trata de la Lógica de Horn, y sus variantes han sido redescubiertas muchas veces de manera independiente.

# Algunas formulaciones

## Aparentemente dispares

En todos estos temas aparece, en mayor o menor centralidad, un concepto esencialmente equivalente al de la lógica de Horn:

i. Semiretículos y retículos.
ii. Conexiones de Galois.
iii. Análisis de Conceptos Formales.
iv. Geometrías convexas.
v. Antimatroides.
vi. Operadores de clausura (monotonía, extensividad e idempotencia).
vii. Teoría de la normalización en Bases de Datos.
viii. "Knowledge compilation".

# Modus Operandi

## Lógica de Horn y variantes en Minería de Datos

Propiedades: **a**, **b**, **c**, **d**

Observaciones: $m_1$ , $m_2$ , $m_3$ ,

| **Id** | **a** | **b** | **c** | **d** |
|---|---|---|---|---|
| $m_1$ | 1 | 1 | 0 | 1 |
| $m_2$ | 0 | 1 | 1 | 1 |
| $m_3$ | 0 | 1 | 0 | 1 |

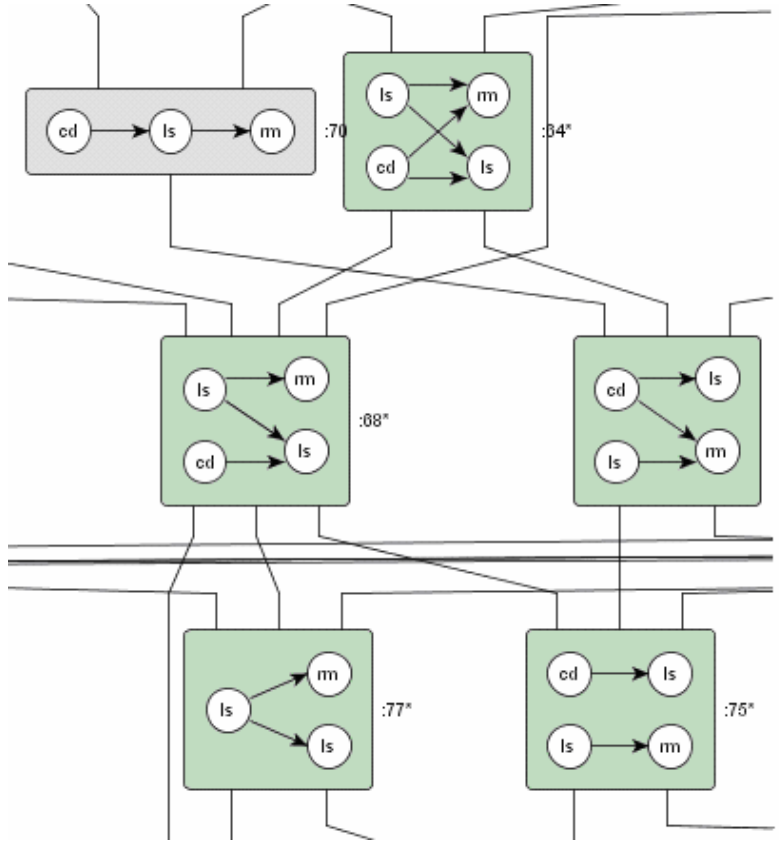
$$a \Rightarrow b, d$$
$$d \Rightarrow b$$
$$a, b \Rightarrow d$$
$$\ldots$$

Admite interesantes extensiones.

# Asoma el álgebra

## Las implicaciones definen un retículo

Aplicaciones: datos médicos, censos, compras en grandes cadenas de supermercados, "clicks" en visitas a páginas web,




uso del "shell" en Unix...

# Reglas de asociación

## "Cláusulas de Horn" que permiten "excepciones"

Las implicaciones exigen absoluta ausencia de "contraejemplos". Frecuentemente esta condición es demasiado estricta.

### Por ejemplo:

En el conjunto de datos de censos estadounidenses, hay una tupla que incluye los dos items: "Husband" y "Female".

# Reglas de asociación

## "Cláusulas de Horn" que permiten "excepciones"

Las implicaciones exigen absoluta ausencia de "contraejemplos". Frecuentemente esta condición es demasiado estricta.

### Por ejemplo:

En el conjunto de datos de censos estadounidenses, hay una tupla que incluye los dos items: "Husband" y "Female".

Pero no es obvio cómo relajar la implicación.

# Reglas de asociación

## "Cláusulas de Horn" que permiten "excepciones"

Las implicaciones exigen absoluta ausencia de "contraejemplos". Frecuentemente esta condición es demasiado estricta.

### Por ejemplo:

En el conjunto de datos de censos estadounidenses, hay una tupla que incluye los dos items: "Husband" y "Female".

Pero no es obvio cómo relajar la implicación.

### Un problema tratado recientemente:

¿Cómo programar la exploración de todas las posibilidades para extraer, a partir de un conjunto grande de datos, las "asociaciones" (implicaciones relajadas) que aparecen entre ellos?

# Reglas de asociación

## "Cláusulas de Horn" que permiten "excepciones"

Las implicaciones exigen absoluta ausencia de "contraejemplos". Frecuentemente esta condición es demasiado estricta.

### Por ejemplo:

En el conjunto de datos de censos estadounidenses, hay una tupla que incluye los dos items: "Husband" y "Female".

Pero no es obvio cómo relajar la implicación.

### Un problema tratado recientemente:

¿Cómo programar la exploración de todas las posibilidades para extraer, a partir de un conjunto grande de datos, las "asociaciones" (implicaciones relajadas) que aparecen entre ellos?

### Un problema aún en estudio:

¿Cómo programar la exploración de todas las posibilidades para identificar, entre todas las "asociaciones", las que realmente deseamos?

# Métricas alternativas

## Para seleccionar reglas de asociación

### Por ejemplo:

En censos estadounidenses, encontramos que en más de 2/3 de los casos:

- $\Rightarrow$ United-States, White
- Husband $\Rightarrow$ Male, Married-civ-spouse
- Married-civ-spouse $\Rightarrow$ Husband, Male
- Not-in-family $\Rightarrow <= 50K$
- Black $\Rightarrow <= 50K$, United-States
- Adm-clerical, Private $\Rightarrow <= 50K$
- Self-emp-not-inc $\Rightarrow$ Male
- $<= 50K$, Sales $\Rightarrow$ Private
- hours-per-week:50 $\Rightarrow$ Male
- Female, Some-college $\Rightarrow <= 50K$
- Divorced $\Rightarrow <= 50K$

# En resumen

## Intuiciones a conservar

A medio camino entre el álgebra de retículos y la minería de reglas de asociación, existe un tema de trabajo matemático fascinante, con una teoría profunda e inmediatas aplicaciones prácticas.

- La lógica de Horn proposicional admite múltiples disfraces.
- Ofrece buenos algoritmos para ciertos problemas.
- Para otros... nos defendemos.
- En la práctica hay que admitir excepciones.
- A partir de ese punto las cosas se nos complican.
- Los avances recientes permiten extraer reglas bastante legibles, pero aún es muy poco lo que podemos garantizar matemáticamente que cumplen esas reglas en función de los datos originales.
- ¡Estoy convencido de que en pocos años veremos avances sorprendentes!