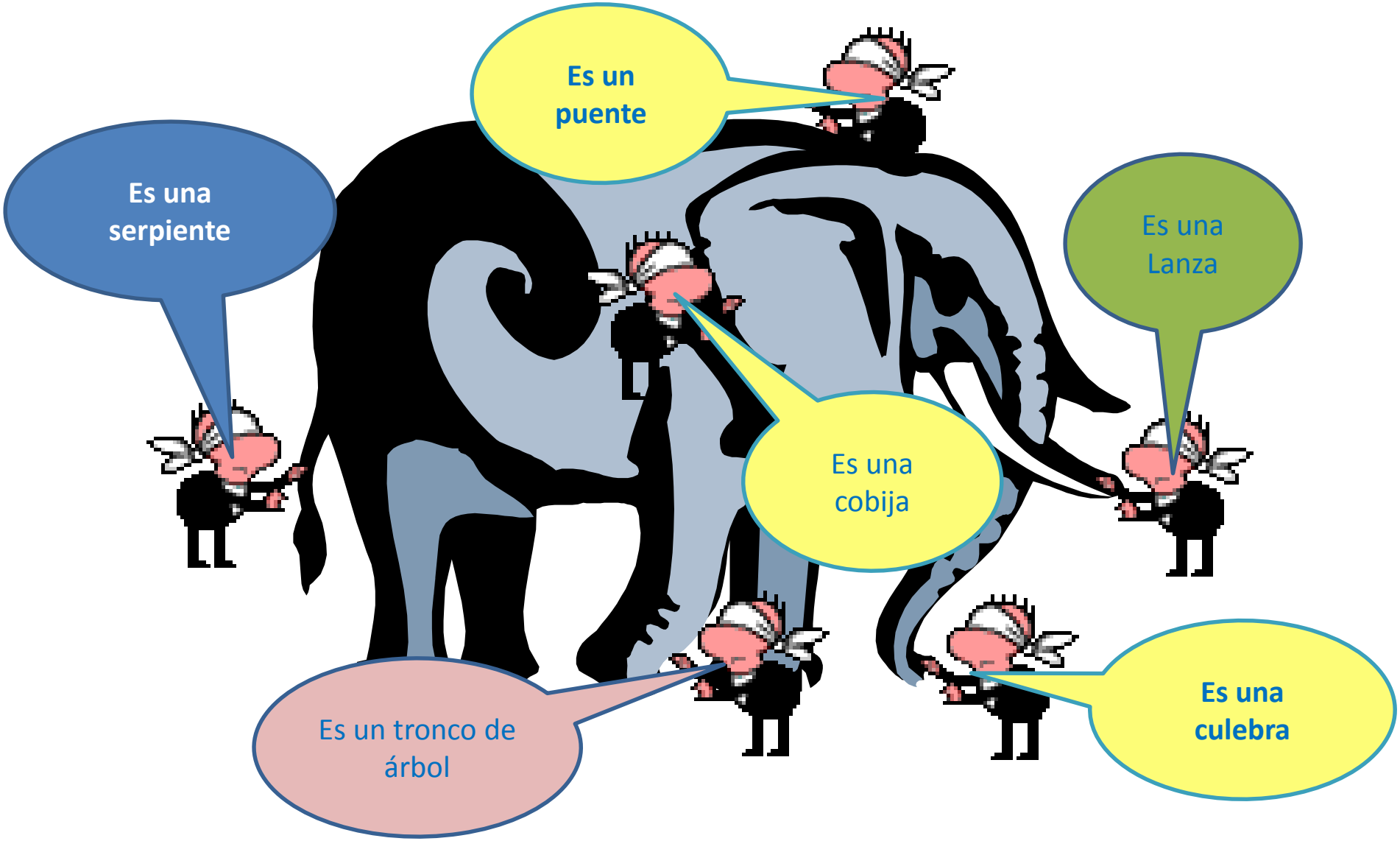
Ingeniería de Conocimiento: Analítica de Datos

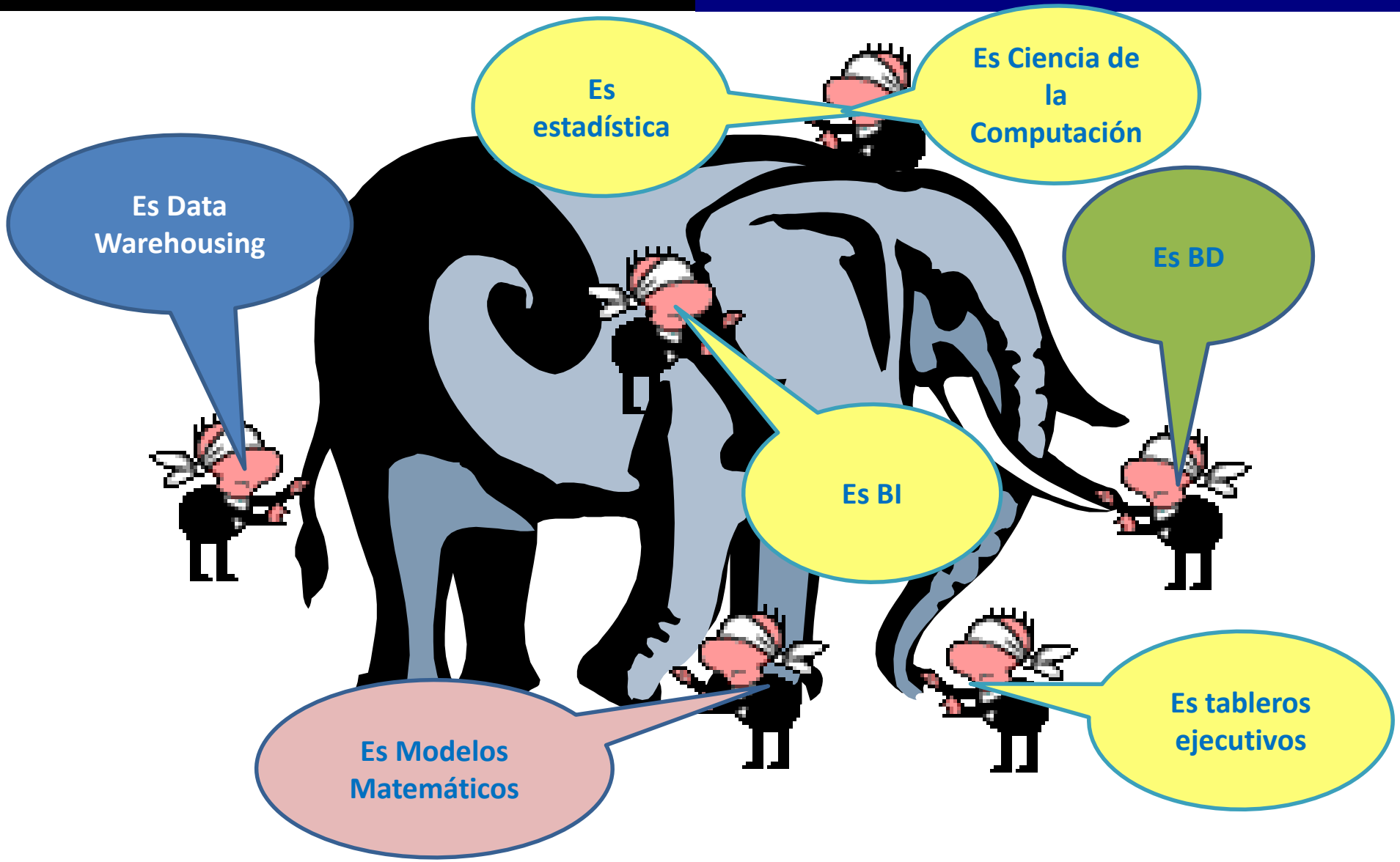
Jose Aguilar

Un ciego describiendo un Elefante



Un ciego describiendo un Elefante

AD



Analítica de Datos

- Con las grandes cantidades de datos disponibles actualmente, las organizaciones se centran en la explotación de los datos para obtener una ventaja competitiva.
- En el pasado, las oganizaciones podrían emplear equipos de estadísticos y analistas para explorar conjuntos de datos de forma manual, pero el volumen y la variedad de datos han superado con creces la capacidad de análisis manual.
- Las computadoras son más poderosas, el trabajo en red es omnipresente, y se han desarrollado algoritmos que pueden conectar conjuntos de datos para permitir análisis más amplios y profundos que antes era imposible.

La convergencia de estos fenómenos ha dado lugar a la aplicación generalizada cada vez mayor de los principios de la analítica de datos.

La Ubicuidad de los Datos

- La AD se utiliza en general para analizar comportamientos de clientes, de empleados, etc..
- La industria financiera utiliza la AD para la detección del fraude, administración de personal, para decidir créditos, etc.
- Walmart, Amazon usan la AD en sus negocios, marketing, en la gestión de la cadena de gestión.
- Muchas empresas se están moviendo a la ciencia de los datos, hasta el punto de evolucionar hacia empresas de AD.

Los objetivos principales de AD son ayudar a ver los problemas de negocio desde una perspectiva de los datos, y comprender los principios de extracción de conocimiento útil a partir de los datos.

Analítica de Datos

Análisis de los datos **es la ciencia** de la recogida, almacenamiento, extracción, limpieza, transformación, agregación y análisis de datos, **con el fin de descubrir información y el conocimiento.**

- Analítica utiliza modelos descriptivos, de identificación y predictivos, con el fin de **producir conocimiento a partir de datos**, que se utilizará para guiar la **toma de decisiones**.
- El **alto grado de datificación incrustado en la sociedad** exige nuevas herramientas y mecanismos para la manipulación y la representación de los datos que facilitan la **extracción de conocimiento significativo**.

Analítica de Datos

Los datos son el nuevo petróleo de la economía

Es la ciencia que examina datos en bruto con el propósito de buscar conocimiento, sacar conclusiones, generar información, entre otras cosas.

Es usado en muchos ámbitos:

- La industria para tomar mejores decisiones empresariales
- Las ciencias para verificar o reprobar modelos o teorías existentes.
- …

Análisis de los datos no es simplemente la minería de datos

La minería de datos navega a través de grandes conjuntos de datos utilizando un software sofisticado para identificar patrones.

El análisis de datos se centra en la inferencia para sacar una conclusión sobre la base de lo que se conoce, para descubrir relaciones ocultas y establecer nuevos

Análisis de los datos no es simplemente la minería de datos

Datos disponibles para el agricultor:

1. Los patrones climáticos históricos

2. Los datos de cultivo de plantas y la productividad de cada cepa

3. Las especificaciones de fertilizantes

4. Especificaciones de Plaguicidas

5. Los datos de productividad del suelo

6. datos del ciclo de plagas

7. coste, fiabilidad, costes de fallo y los datos de las máquinas

8. Los datos de conducción de aguas

9. Los datos de oferta y demanda históricos

10. Mercado de precios al contado y de futuros datos

Análisis de los datos no es simplemente la minería de datos

Analítica Datos es la aplicación de un procedimiento algorítmico para obtener conocimiento, desde un conjuntos de datos.

- En cuanto a los datos del tiempo y plagas vemos que existe una alta correlación de un cierto tipo de hongo cuando el nivel de humedad alcanza un cierto punto.
- Las futuras proyecciones meteorológicas para los próximos meses predicen un bajo nivel de humedad y, por tanto, el riesgo bajo de ese hongo.

Para el agricultor que esto podría significar ser capaz de plantar un determinado tipo de fresa, para mayor rendimiento y mayor precio de mercado, sin tener que comprar un determinado fungicida.

En MD los datos se reúnen para descubrir tendencias, anomalías, correlaciones como patrones, previamente desconocidos.

Todos los datos se usan para responder a:
¿Cuál es el patrón de siembra de las fresas para obtener el mejor precio?
¿cuáles son los tipos de fresa con mejor rendimiento?

Análisis de los datos no es simplemente la minería de datos

Análisis de datos: cualquier intento de dar sentido a los datos puede ser llamado como el análisis de datos. Actividad heurística, donde se realización una exploración a través de todos los datos, tal que el analistas ganan conocimiento sobre el negocio

Es la ciencia o el proceso de análisis de datos para sacar conclusiones acerca de lo que está pasando en el proceso, evento, etc. que los datos están describiendo.

Minería de datos: se refiere a la ciencia de la recopilación de todos los datos del pasado y luego la búsqueda de patrones en los datos. Una vez que se encuentren, se validan mediante la aplicación de los patrones detectados a nuevos subconjuntos de datos.

es el proceso de buscar a través de conjuntos de datos existentes, relaciones entre ellos.

Analítica de Datos

- Enfoques de des-síntesis de los datos e información para responder a preguntas

- Método de crear datos, cifras (conocimiento) para resolver problemas

- Proceso sistemático de utilizar los datos para hacer frente a cuestiones

- Desglose de los temas a través de la utilización de los datos y la información conocida

Analítica de Datos

Utilizando datos para **aumentar el valor del negocio**

Dato =

- Grande y pequeño
- Interno y externo
- Estructurado y no estructurado
- Tradicional y "Nuevo“
- "Libre" y comprado

Analítica de Datos

Los datos se pueden "hablar“

Un análisis contiene algunos aspectos del razonamiento científico:

* Define
* Interpreta
* Evalua
* Ilustra
* Discute
* Explica
* Clarifica
* Compara
* Contrasta

Objetivo de un análisis:

- Para explicar los fenómenos de causa y efecto
- Para relacionar la investigación con el mundo real
- Para predecir / pronosticar en el mundo real fenómenos
- Para encontrar respuestas a un problema particular
- Para concluir acerca de eventos del mundo real basado en el problema
- Para aprender de un problema

Guía básica para el análisis de datos:

- Analizar es “NO“ narrar
- Descomponen objetivos en preguntas de investigación
- Identificar los fenómenos que han de investigarse
- Visualizar las respuestas "esperadas“
- Validar las respuestas con los datos
- No diga algo que no es soportado por datos

- Al analizar:
  - Sea objetivo
  - Preciso
  - Cierto

- Separar los hechos y la opinión

- Evitar el "mal" razonamiento. Por ejemplo. errores en la interpretación.

La gestión usando AD

El éxito de la analítica sólo puede medirse en términos de lo bien que ayudan a lograr objetivos estratégicos

Por lo tanto, se debe:

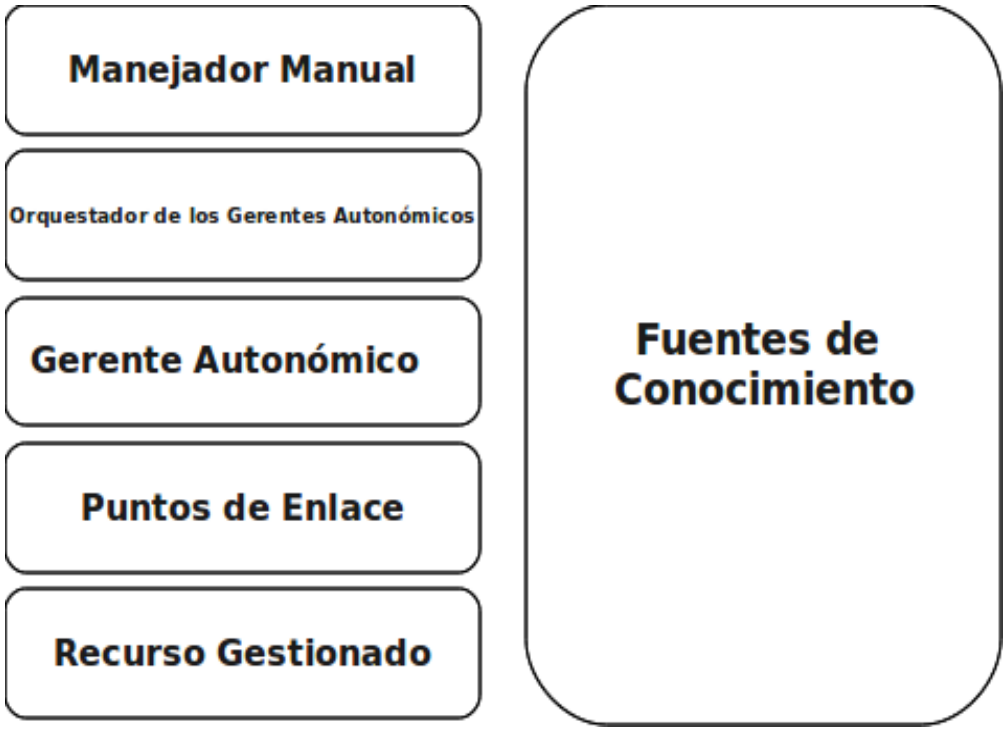
- Identificar los objetivos de negocio
- Recoger los datos necesarios para medir sus objetivos
- Analizar los datos
- Sacar conclusiones sobre la base de la información generada

Un análisis debe tener cuatro elementos:

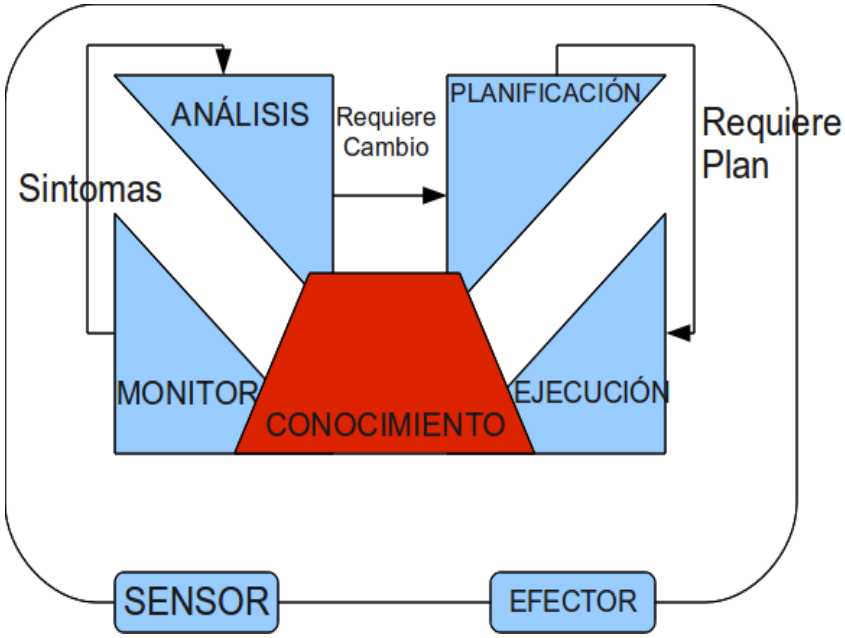
- Los datos/información
- Razonamiento Científico (¿qué? ¿quien? ¿dónde? ¿qué pasa?)
- Hallazgo (¿qué resultados?)
- Lección/conclusión (¿y qué? ¿cómo?, por lo tanto,…)

Ciclo Autonómico

Arquitectura Computación Autonómica

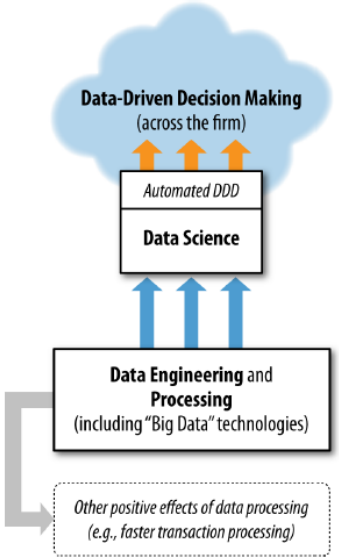


Computación Autonómica basada en arquitectura MAPE-K:

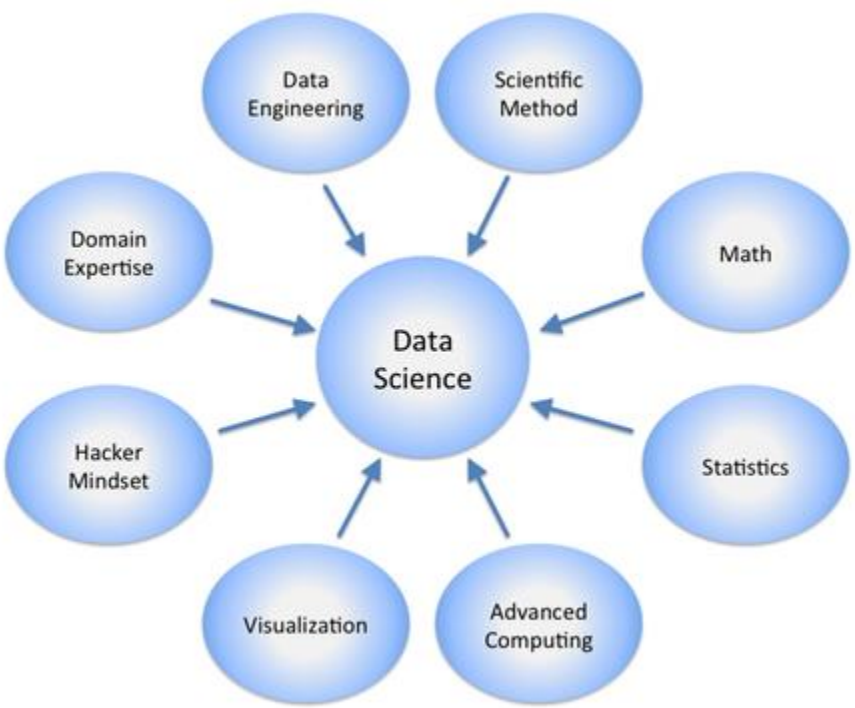


La ciencia de los datos

la ciencia de datos requiere de principios, procesos y técnicas para la comprensión de los fenómenos a través del análisis (automatizado) de los datos.

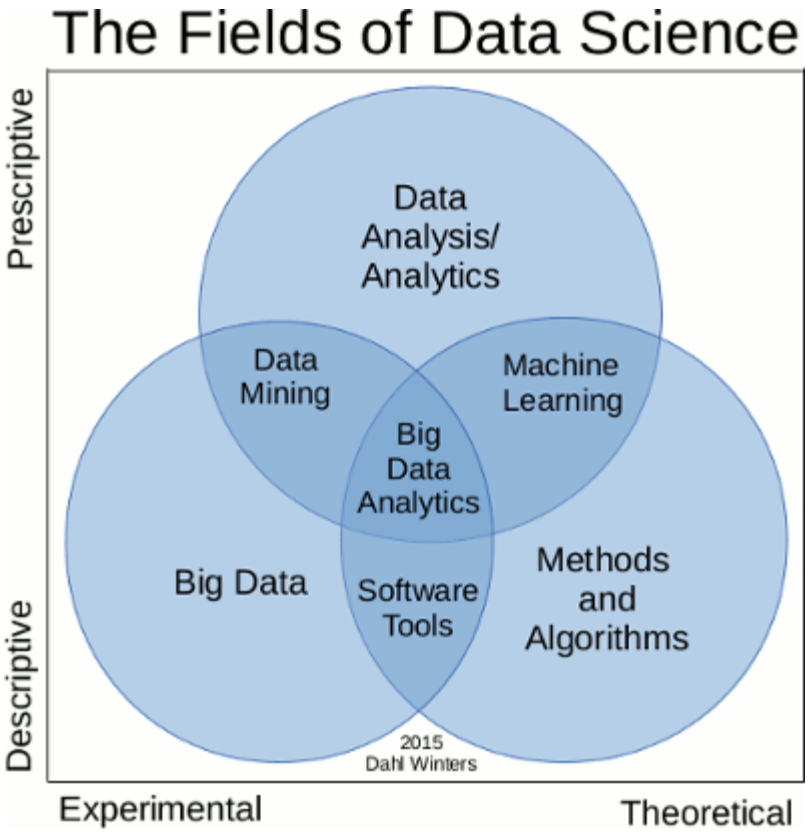


La ciencia de los datos



Combinación de las matemáticas, estadísticas, etc., para resolver el problema de captura de datos, además de la limpieza, la preparación y la alineación de los datos.

La ciencia de los datos



Preparando la Entrada

Preparación de entrada para una investigación de AD suele **consumir la mayor parte del esfuerzo invertido en el proceso**.

Los datos deben pasar por **procesos de ensamblaje, integración, limpieza, agregación y preparación general**.

Recopilación de los datos en conjunto

Conocer los datos

Expertos de dominio deben ser consultados para explicar las anomalías, los valores perdidos, el significado de los números enteros que representan categorías en lugar de cantidades numéricas, y así sucesivamente.

La ciencia de datos es un procedimiento que **consume tiempo y requieren mucho trabajo**, pero que es absolutamente necesario para la AD con éxito.

Limpieza

se refiere a una serie de procesos en los cuales la **calidad de los datos es mejorada**, enfrentando los problemas mencionados como datos mal capturados, anómalos y vacíos, ya sea por características obvias que el dato no cumple con ciertos parámetros del estándar, o porque el experto del proceso ya tiene identificado anomalías comunes en el almacenamiento de los datos.

- normalización de formatos,
- remoción de anomalías,
- corrección de errores
- eliminación de duplicados.

Limpieza

Ejemplos anomalías que presenta la base de datos:

- Unidades de las entradas
- Abreviaciones
- Convenciones de nombres
- Representaciones diferentes
- Variaciones de Ortografía
- Elementos repetidos
- Datos no guardados

Para buscar anomalías :

- Estudiar la representación de cada una de las variables.
- Buscar anomalías de representación.
- Definir alguna estrategia de limpieza para erradicar dichas
- Realizar las operaciones con un software para limpieza de datos.

Transformación

En esta etapa se **transforman las variables de entrada en nuevas variables de interés**, esto se realiza a través de diversos métodos, los cuales se deben escoger en caso de ser pertinente alguna transformación de alguna de las variables.

Una transformación de variables puede ser la combinación entre variables

- concatenación de cadenas,
- multiplicación entre variables,
- otras operaciones aritméticas, etc.

Proceso Transformación

- Estudiar las representaciones de cada una de las variables
- Identificar las representaciones que se puedan transformar en otra representación más conveniente o fácil de utilizar para AD.
- Ordenar dichas transformaciones que se desean aplicar en una tabla, para observar las equivalencias
- Aplicar la transformación con el software seleccionado
- Identificar las variables que potencialmente se pueden normalizar
- Definir la función(es) de normalización para cada una de las variables seleccionadas en el paso anterior y ordenarla en tablas.
- Aplicar la función(es) de normalización en las variables seleccionada
- Combinar variables por un método seleccionado tal como el PCA (del inglés *Principal Component Analysis*) que es considerado también un método para reducción de variables.
- Describir en tablas cada una de las transformaciones realizadas.

La ciencia de los datos

Reducción

Consiste en **decidir qué datos deben ser utilizados para el análisis**. El criterio que se sigue para realizar reducción de variables incluye la relevancia con respecto a los objetivos que se persiguen, y limitaciones técnicas tales como los volúmenes máximos de datos o bien tipos de datos concretos.

Así que en este paso se reduce la cantidad de variables a sólo las necesarias para modelar el proceso en estudio.

- Realizar análisis estadísticos para reducir variables que posean una alta relación lineal, como por ejemplo un análisis de correlación.
- Identificar las posibles variables que se pueden reducir.
- Justificar la reducción de las mismas
- Construir la nueva vista minable con las nuevas variables reducidas

Datos Dispersos

- La mayoría de los atributos tienen un valor de 0
  - Los registros de datos de la cesta de compras realizadas por los clientes de los supermercados

- Puede ser poco práctico representar cada elemento de una matriz dispersa de forma explícita:
  - 0, 26, 0, 0, 0, 0, 63, 0, 0, 0, "clase A"
  - 0, 0, 0, 42, 0, 0, 0, 0, 0, 0, "clase B"

En cambio, los atributos no nulos pueden ser explícitamente identificado por el número de atributo y su valor declarado:

{1 26, 6 63, 10 "clase A"}
{3 42, 10 "clase B"}

Los tipos de atributo

Existen dos tipos básicos de atributos nominales y numéricos.

Los atributos string y date son atributos nominales y numéricos, respectivamente.

Aunque previamente las cadenas pueden ser convertidas en la forma numérica como un vector de palabras

Valores Perdidos

Los Valores perdidos son frecuentemente indicados por fuera del rango de entradas, tal vez un número negativo (-1).

A veces, diferentes tipos de valores perdidos se distinguen (por ejemplo, valores desconocidos: sin grabar vs irrelevante) y que pueden estar representados por diferentes números enteros negativos (-1, -2)

Pueden ocurrir por varias razones:

- Equipos de medición funcionando de manera incorrecta
- Los cambios en el diseño experimental durante la recolección de datos
- Los entrevistados en una encuesta pueden negarse a responder a ciertas preguntas
- En un estudio arqueológico, un espécimen tal como un cráneo pueden ser dañados de forma que algunas variables no se pueden medir.

Valores inexactos

La base de datos debe ser revisada cuidadosamente.

Cuando es recolectada la información, es probable que no importe si hay campos en blanco o sin restricción en los archivos.

Si la misma base de datos es utilizada para AD, los errores y omisiones de inmediato comienzan a tomar significancia.

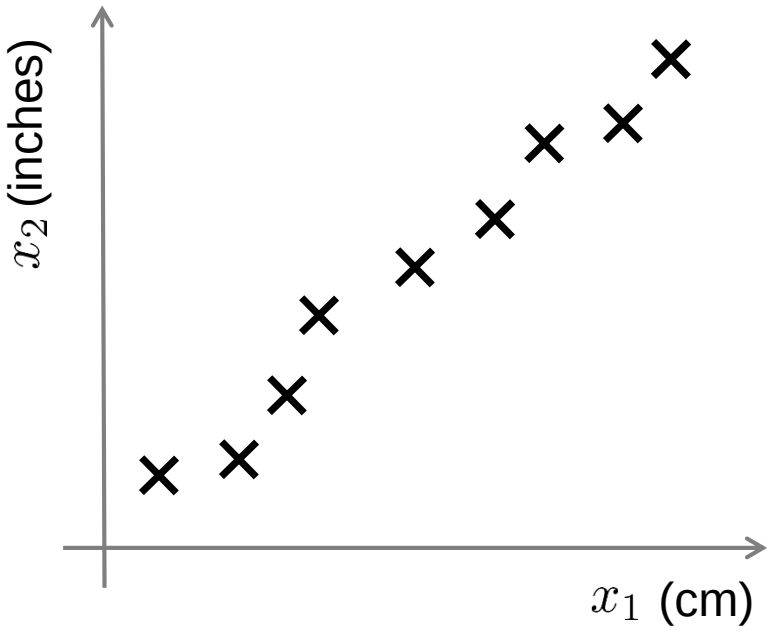
Por ejemplo:

- Datos duplicados son una fuente de error
- Datos que se convierten en obsoletos, hay que considerar si los datos a usar en minería son todavía validos o actuales

Compresión de los datos

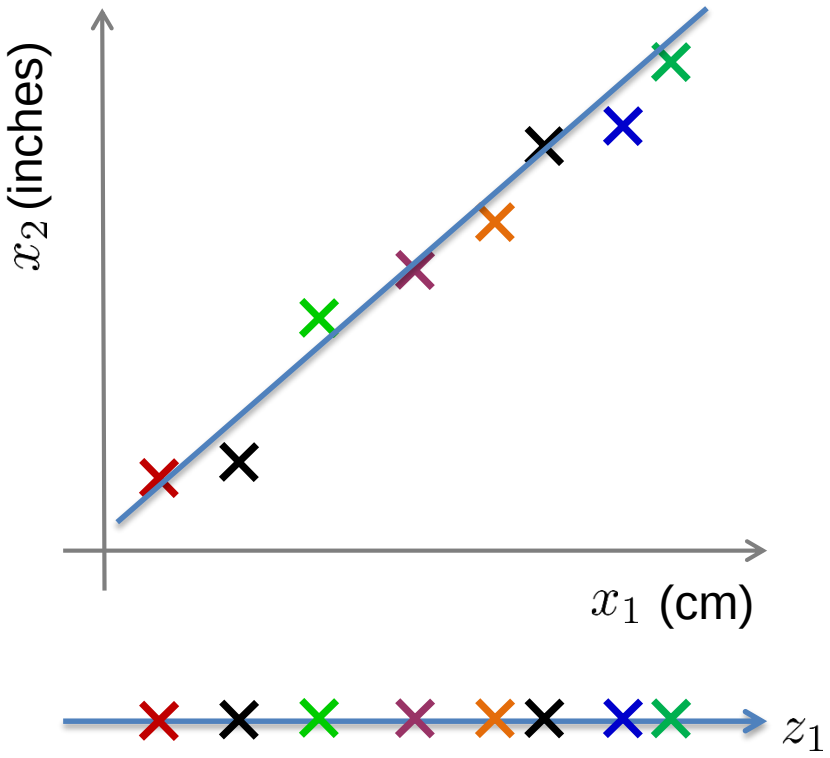
compresión de datos es la reducción del volumen de datos tratables para representar una determinada información empleando una menor cantidad de espacio.

Data Compression



Reduce data from 2D to 1D

Data Compression



Reduce data from 2D to 1D

$$x^{(1)} \in R^2 \rightarrow z^{(1)} \in R$$

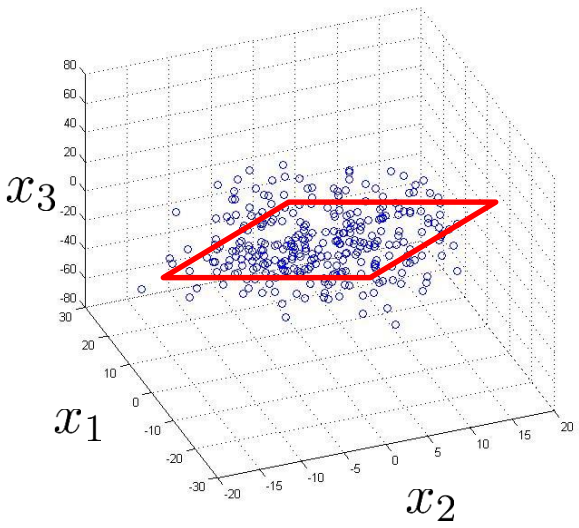
$$x^{(2)} \in R^2 \rightarrow z^{(2)}$$

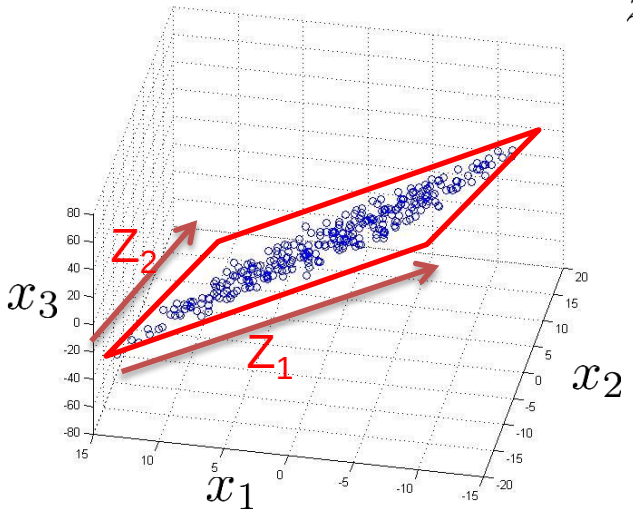
$$\vdots$$

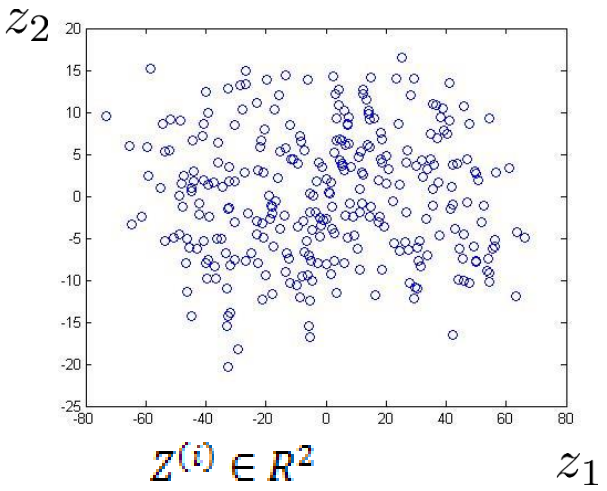
$$x^{(m)} \in R^2 \rightarrow z^{(m)} \in R$$

Data Compression

Reduce data from 3D to 2D







$X^{(i)} \in R^3$

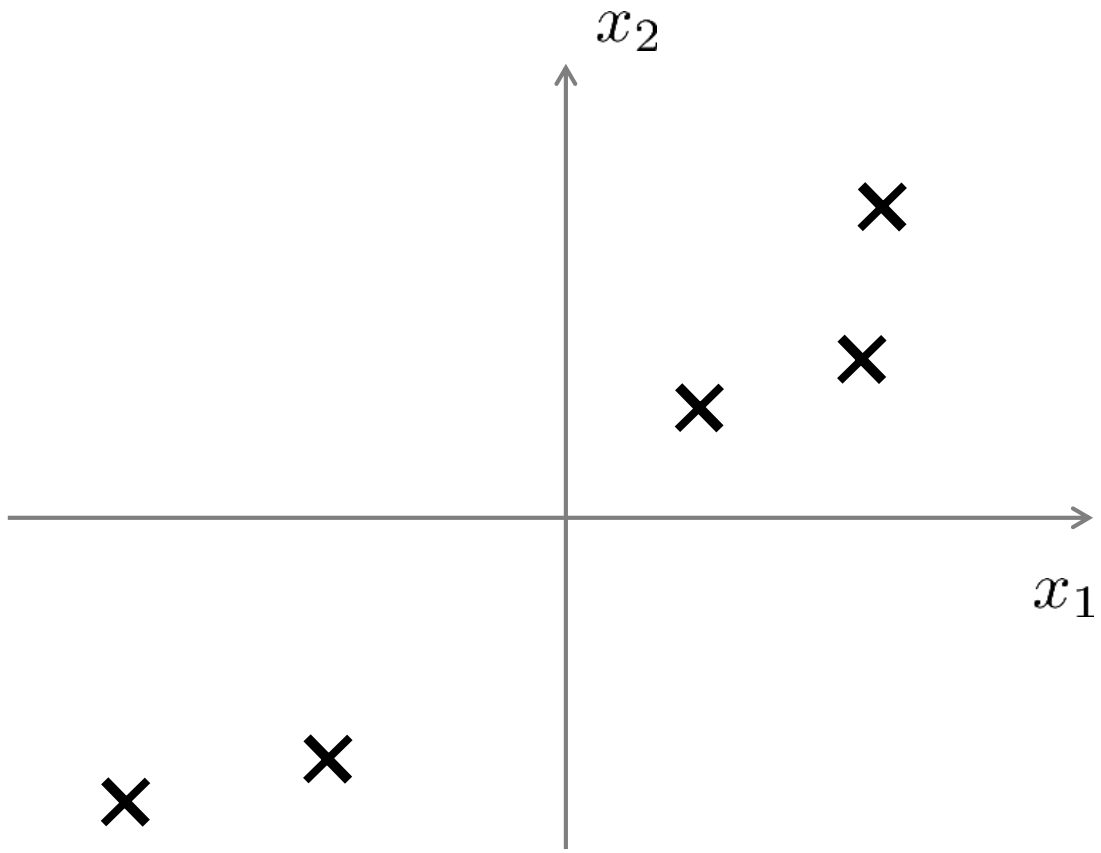
$Z^{(i)} \in R^2$

Formulación del Problema PCA (Principal Component Analysis)

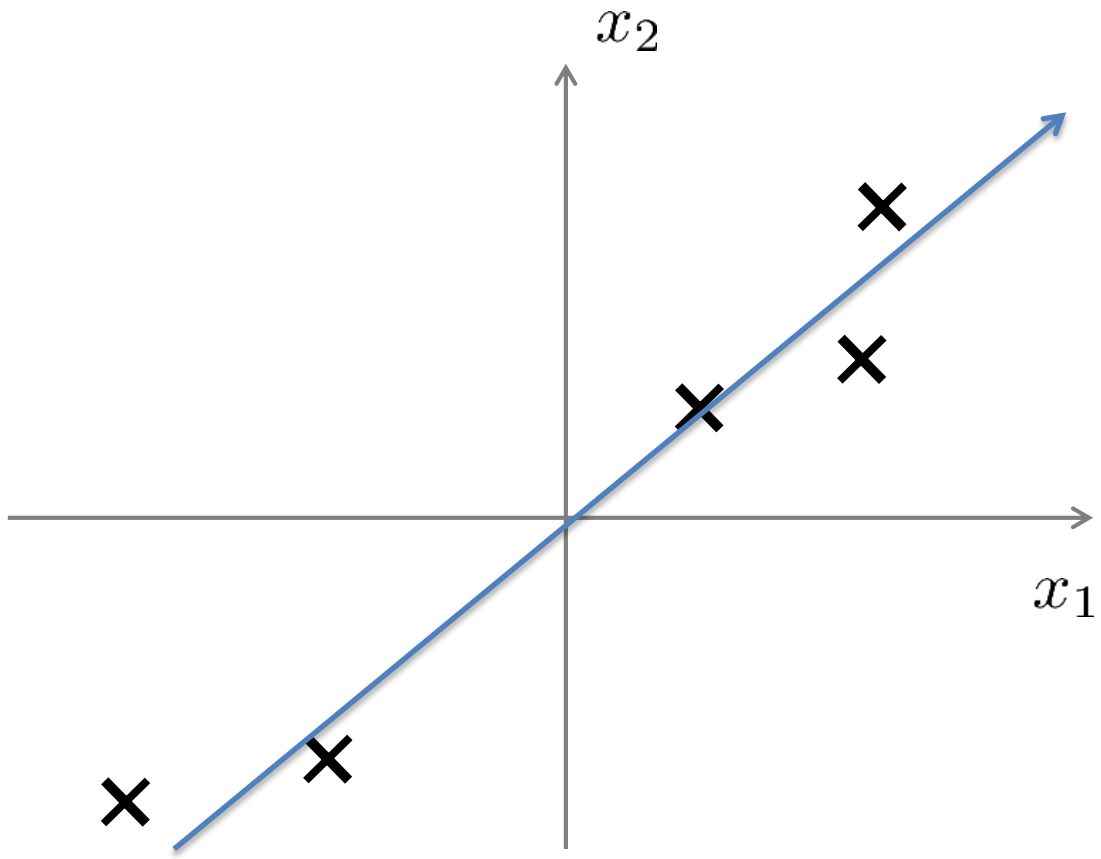
El **análisis de componentes principales** (en español **ACP**) es una técnica utilizada para reducir la dimensionalidad de un conjunto de datos.

Intuitivamente la técnica sirve para hallar las causas de la variabilidad de un conjunto de datos y ordenarlas por importancia.

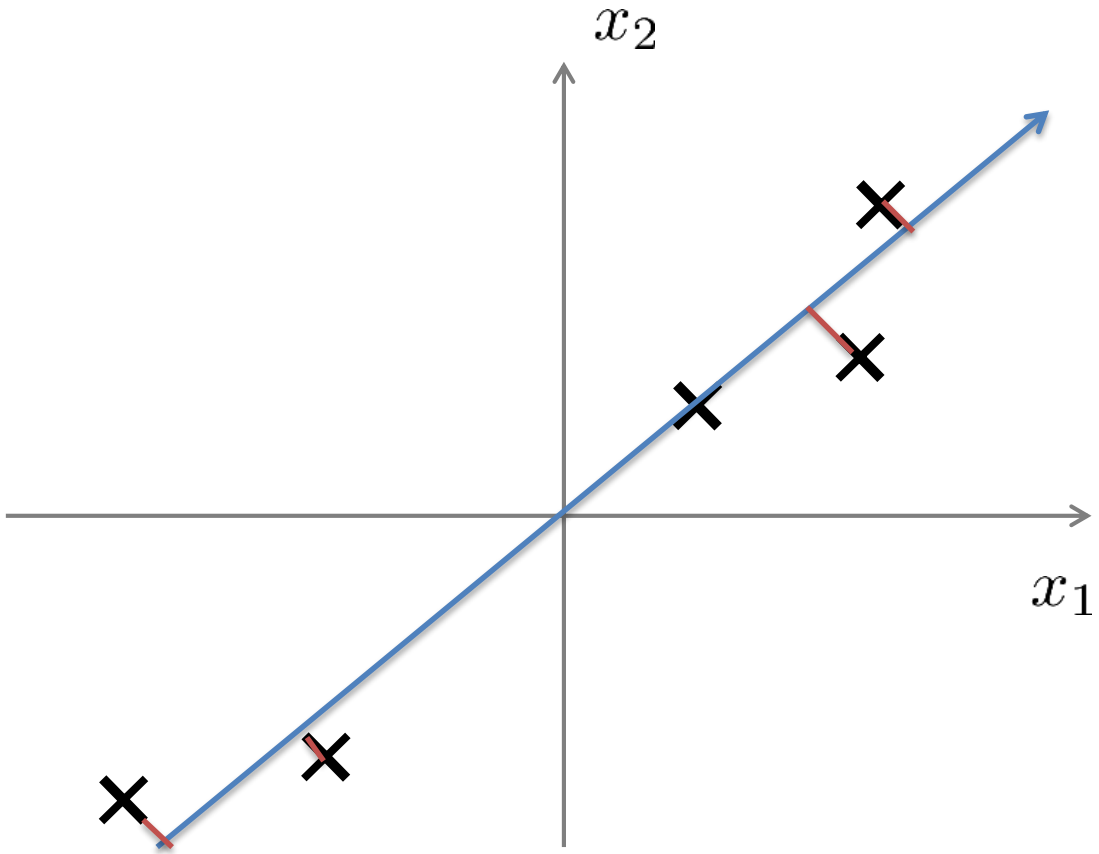
Formulación del Problema PCA (Principal Component Analysis)



Formulación del Problema PCA (Principal Component Analysis)



PCA (Principal Component Analysis) Error de proyección



PCA (Principal Component Analysis) Error de proyección

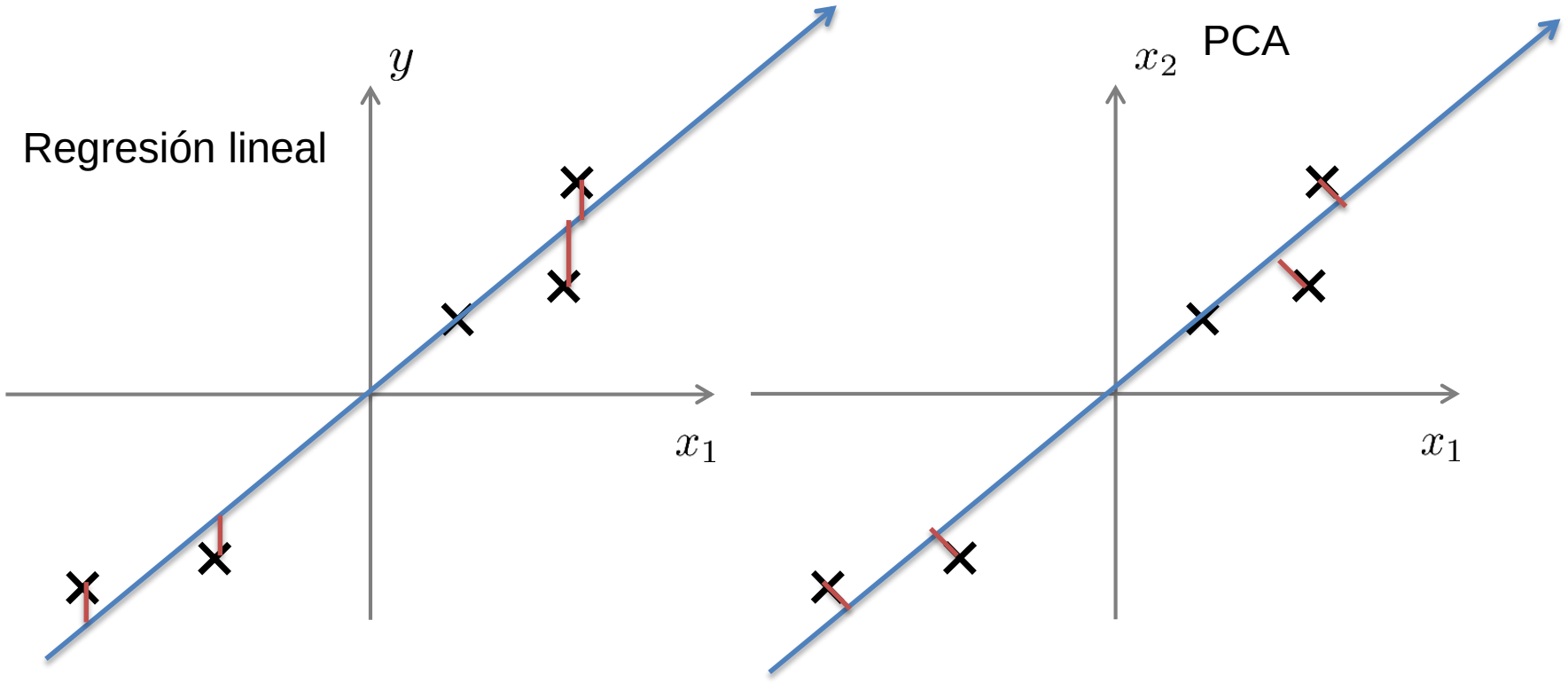
$$u^{(1)} \in \mathbb{R}^n$$

Para reducir de 2- dimensiones a 1-dimensión: Buscar un vector
Para proyectar la data, el cual minimice el error de proyección.

$$u^{(1)}, u^{(2)}, \ldots, u^{(k)}$$

Para reducir de n-dimensiones a K-dimensiones: Buscar K vectores
Para proyectar la data, los cuales minimicen el error de proyección.

PCA no es regresión lineal



Data preprocessing

Training set: $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$

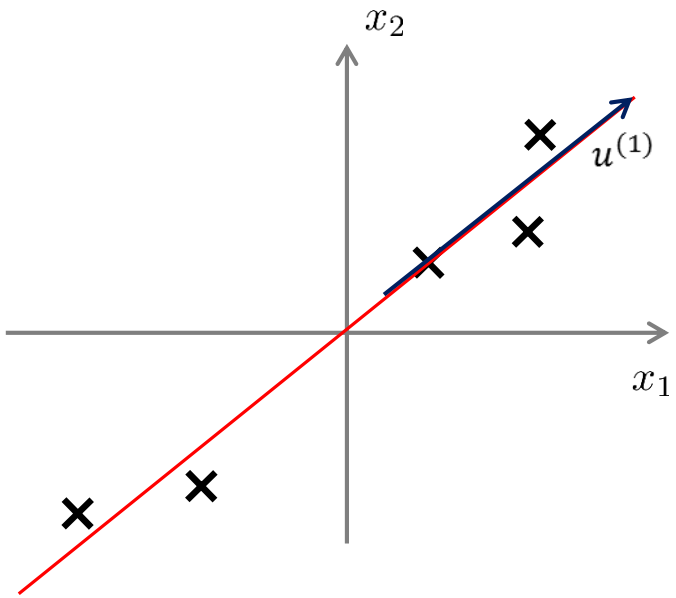
Preprocessing (feature scaling/mean normalization):

$$\mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)}$$

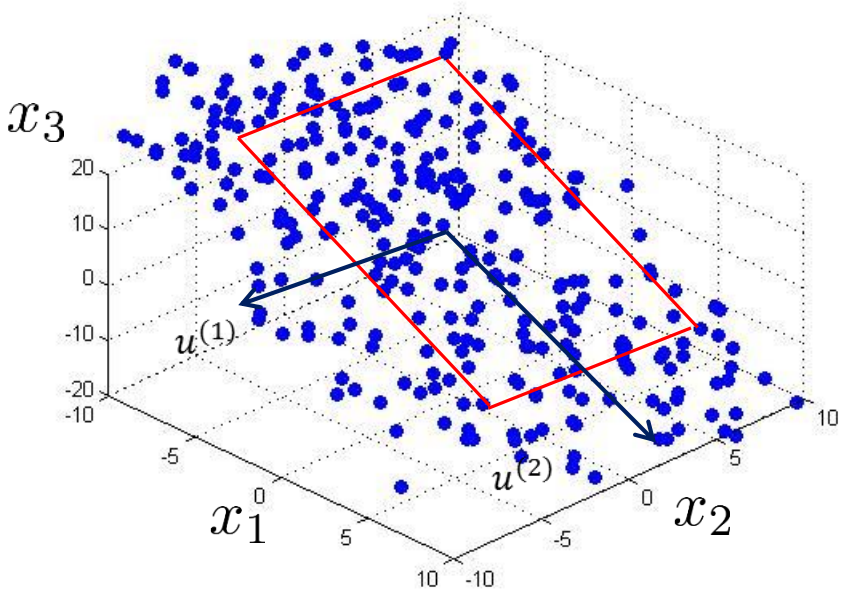
Replace each $x_j^{(i)}$ with $x_j - \mu_j$

If different features on different scales (e.g., $x_1 =$ size of house, $x_2 =$ number of bedrooms), scale features to have comparable range of values.

Principal Component Analysis (PCA)



Reduce data from 2D to 1D



Reduce data from 3D to 2D

Principal Component Analysis (PCA)

Reduce data from n-dimensions to k-dimensions

Compute "covariance matrix":

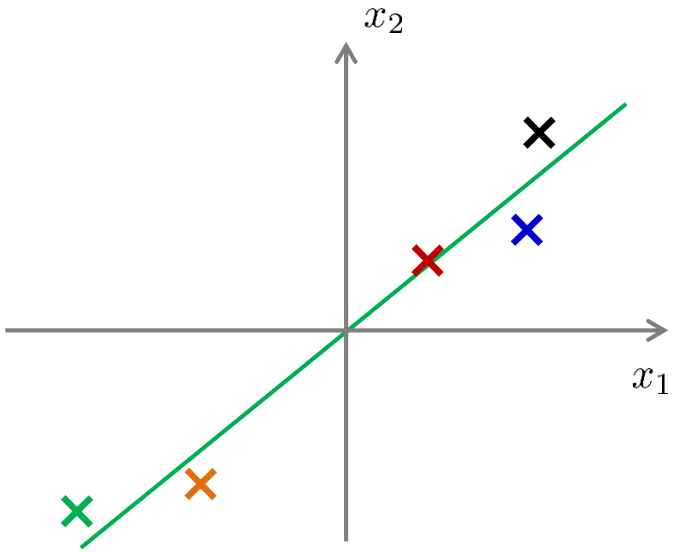
$$\Sigma = \frac{1}{m}\sum_{i=1}^{n}(x^{(i)})(x^{(i)})^T \longrightarrow \text{Sigma}$$

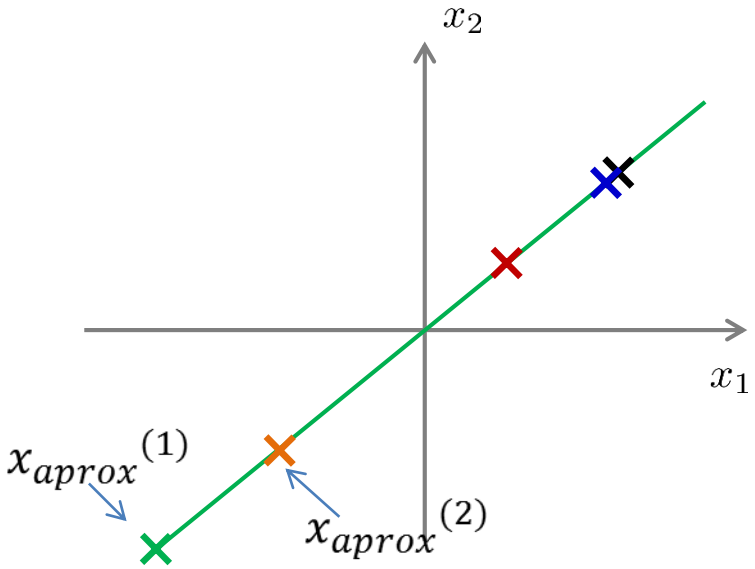
Compute "eigenvectors" of matrix Σ:

```
[U,S,V] = svd(Sigma);
```

El ACP comporta el cálculo de la descomposición en autovalores de la matriz de covarianza,

Reconstrucción





$$z = U_{reduce}^{T} x$$

$$X_{aprox} = U_{reduce} . Z$$

Formato ARFF

El formato ARRF, acrónimo de *Attribute-Relation File Format*. Este formato está compuesto por una estructura claramente diferenciada en tres partes:

1. **Cabecera.** Se define el nombre de la relación. Su formato es el siguiente:

    @relation <nombre-de-la-relación>

Donde <nombre-de-la-relación> es de tipo String*. Si dicho nombre contiene algún espacio será necesario expresarlo entrecomillado.

Formato ARFF

2. **Declaraciones de atributos.** En esta sección se declaran los atributos que compondrán nuestro archivo junto a su tipo. La sintaxis es la siguiente:

@attribute <nombre-del-atributo> <tipo>

3. **Sección de datos.** Declaramos los datos que componen la relación separando entre comas los atributos y con saltos de línea las relaciones.

@data

4,3.2

Formato ARFF

```
% ARFF file for the weather data with some numeric features
%
@relation weather

@attribute outlook { sunny, overcast, rainy }
@attribute temperature numeric
@attribute humidity numeric
@attribute windy { true, false }
@attribute play? { yes, no }

@data
%
% 14 instances
%
sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 86, false, yes
rainy, 70, 96, false, yes
rainy, 68, 80, false, yes
rainy, 65, 70, true, no
overcast, 64, 65, true, yes
sunny, 72, 95, false, no
sunny, 69, 70, false, yes
rainy, 75, 80, false, yes
sunny, 75, 70, true, yes
overcast, 72, 90, true, yes
overcast, 81, 75, false, yes
rainy, 71, 91, true, no
```

Peligros en AD

- Privacidad
- Seguridad
- Decisiones basado en datos incompletos
- Decisiones con datos inexactos
- Usando sólo los datos que apoyan nuestras decisiones
- Llegar a la conclusión errónea de los datos: por ejemplo, los precios de las acciones