UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

UTPL
UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

Ciencias de Datos

Jose Aguilar

Febrero 2017

UNIVERSIDAD
DE LOS ANDES
MERIDA VENEZUELA

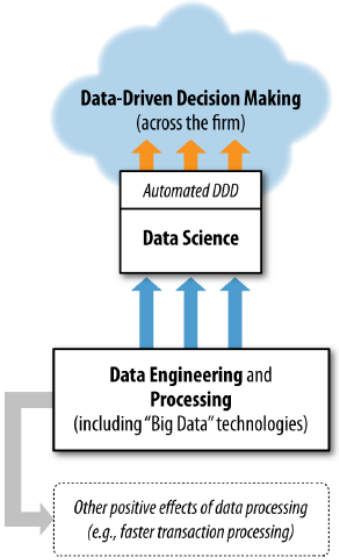
Contenido

- Ciencia de los Datos y operaciones ETL

- Tipos de tareas de Analítica de Datos.

La ciencia de los datos

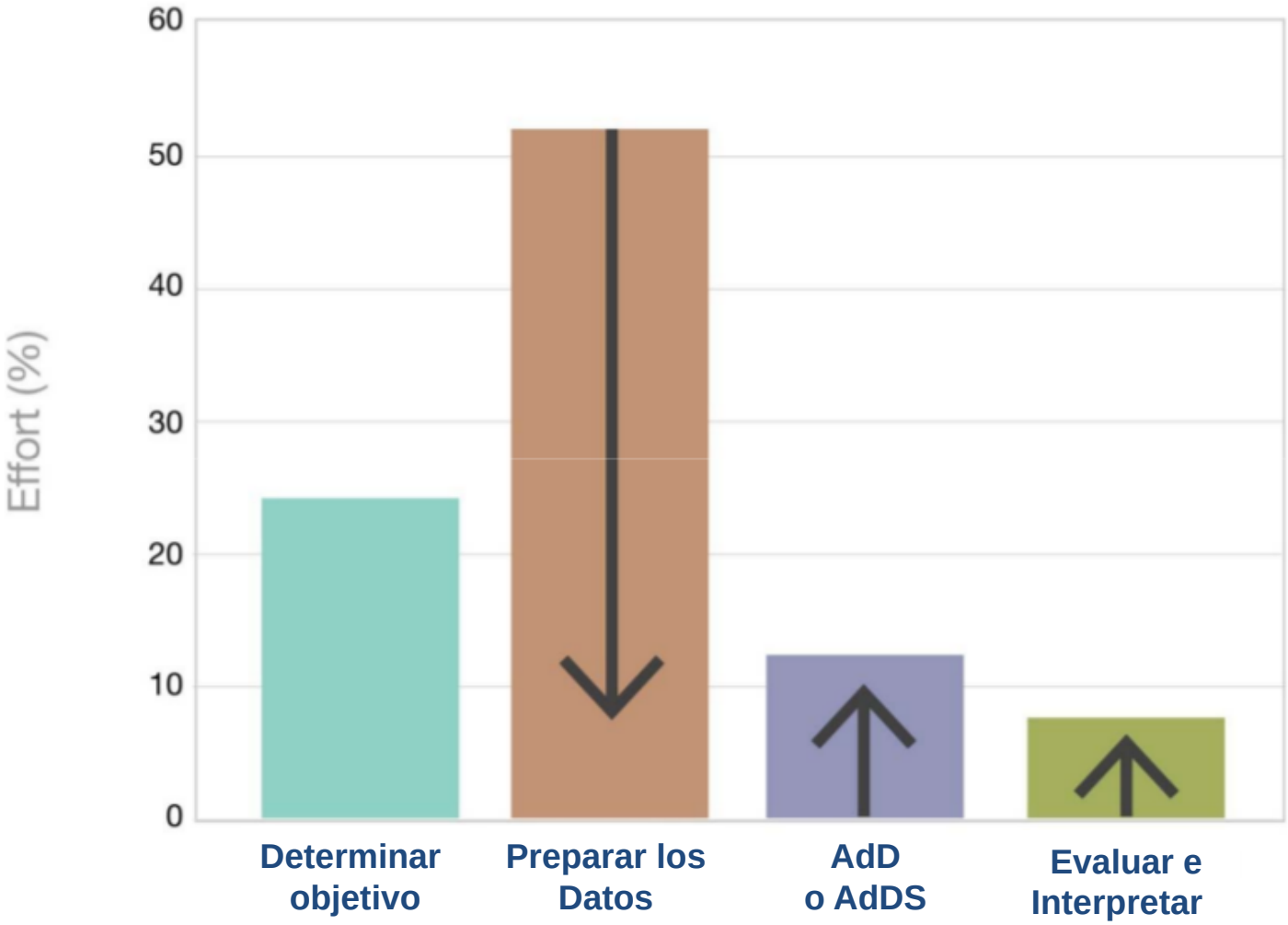
la ciencia de datos requiere de principios, procesos y técnicas para la comprensión de los fenómenos a través del análisis (automatizado) de los datos.



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.

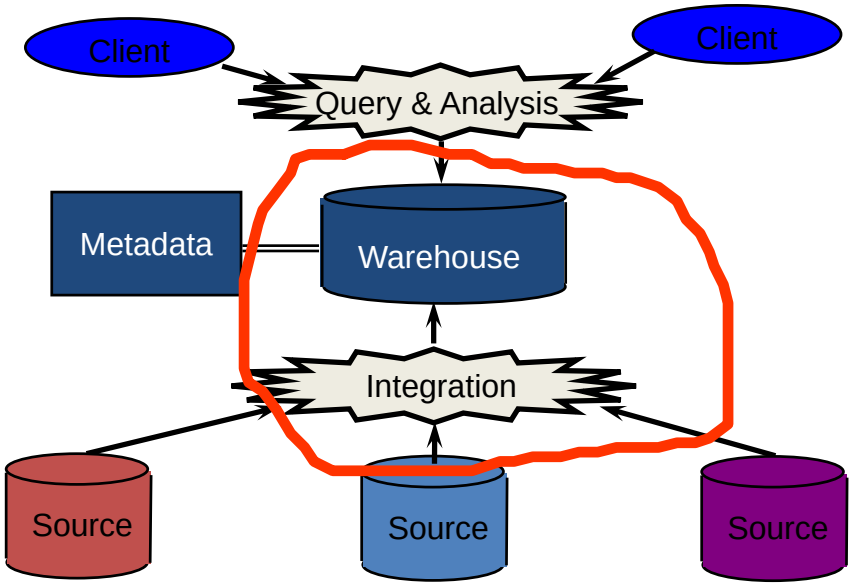
¿Que etapa lleva mas esfuerzo?



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

Integración

- Conocer y Selección de los datos
- Preparación de los datos
- Carga de datos



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.

Preparando los datos

Preparación de la entrada para una investigación de AdD 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**.

Preparación de los Datos

- **Recolección de datos**
  - Captura de la Información
- **Análisis**
  - Entender el contexto de la información
- **Tratamiento de los datos**
  - Hacer ciencia en los datos

http://www.youtube.com/watch?v=-xR5erOhkXo

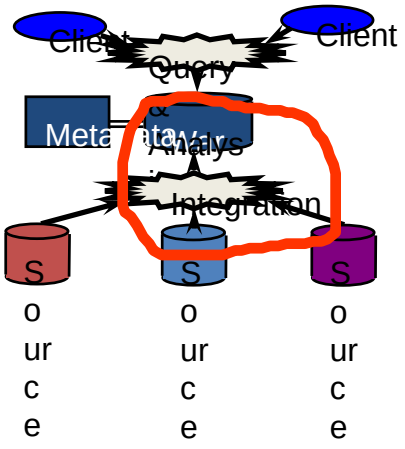
Proceso ETL

ETL (Extracción, Transformación y Carga)

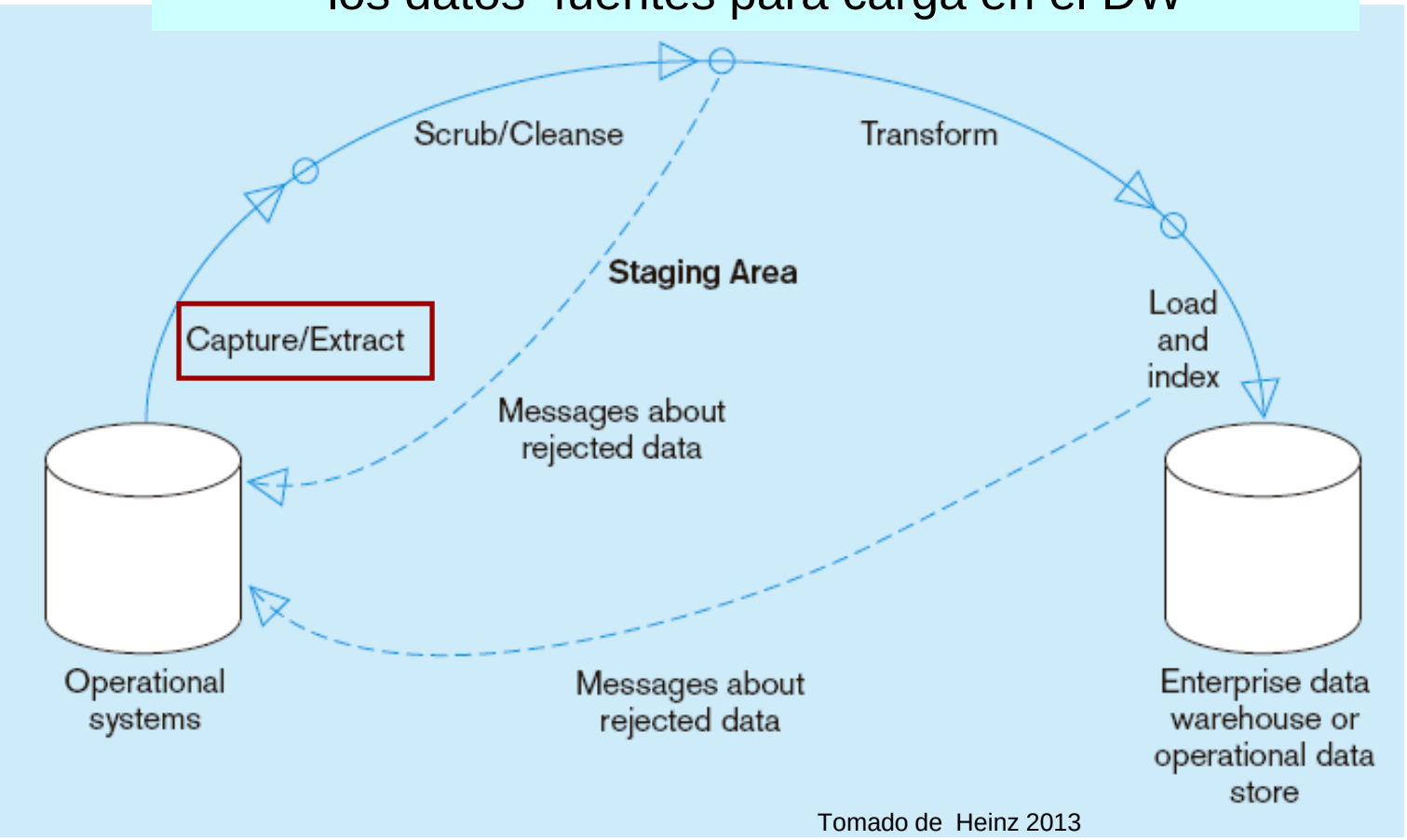
Extracción: Obtención de información de las distintas fuentes, tanto internas como externas.

Transformación: Filtrado, limpieza, depuración, homogeneización y agrupación de la información.

Carga: Organización y actualización de los datos y los metadatos en el DW.



Captura / Extrae... obtiene un subconjunto de los datos fuentes para carga en el DW



Tomado de Heinz 2013

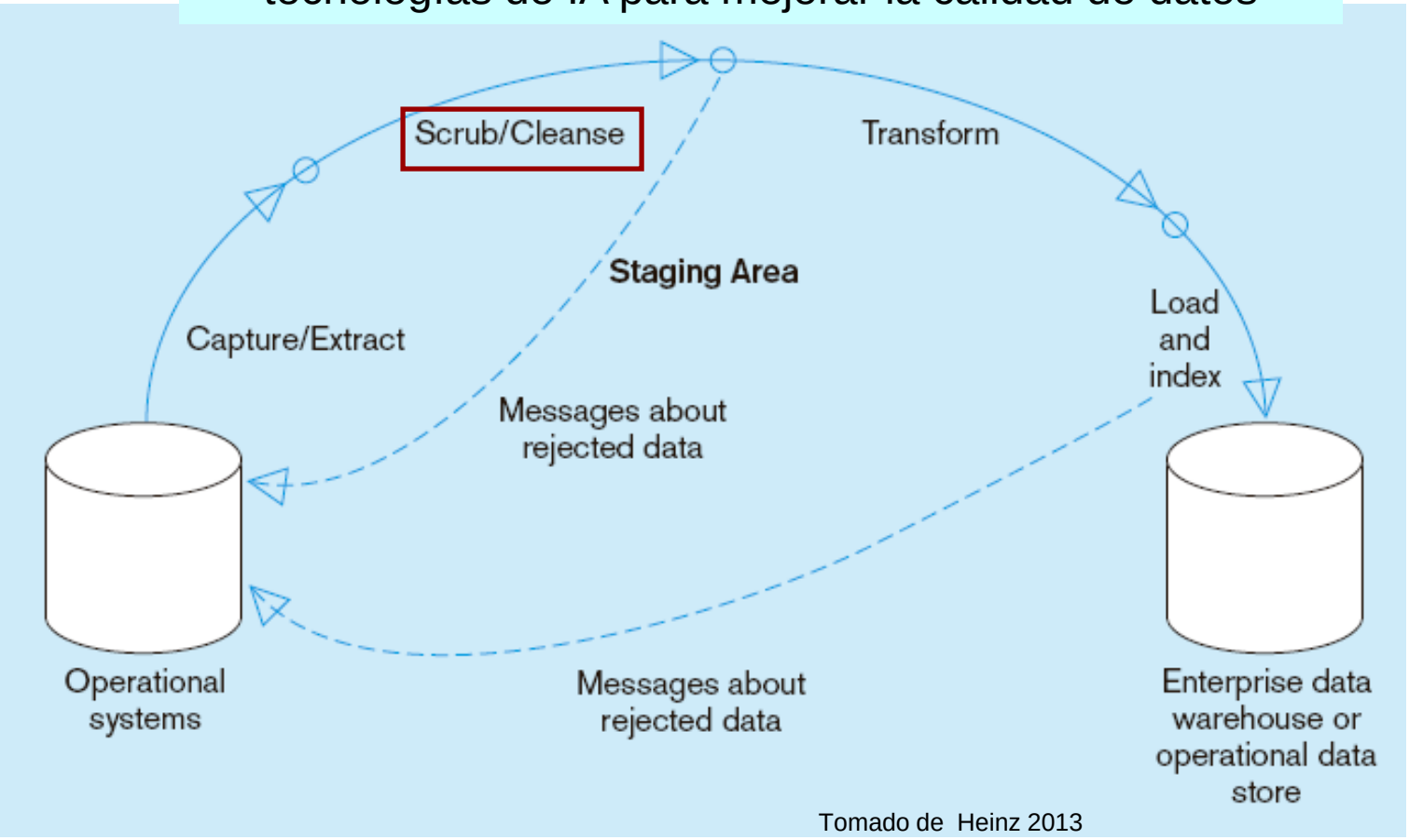
Extracción estática: captura los datos en un momento puntual

Extracción Incremental: captura cambios que se van produciendo

Extracción

- Obtener datos de múltiples fuentes externas , heterogéneas
- Periódica
- Claves:
  - Manipular los datos sin interrumpir ni paralizar los OLTP, ni tampoco el DW.
  - Facilitar la integración de las diversas fuentes, internas y externas.

Limpieza... utiliza reconocimiento de patrones y tecnologías de IA para mejorar la calidad de datos



Tomado de Heinz 2013

Solución de errores: faltas de ortografía, fechas erróneas, uso de campo incorrecto direcciones no coinciden, datos faltantes, datos duplicados, inconsistencias

También decodifica, reformatea, convierte, genera claves, fusiona, detecta errores registro, localiza datos faltantes

Limpieza

se refiere a una serie de procesos en los cuales la **calidad de los datos es mejorada**, enfrentando los problemas 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.

Pasos en la Limpieza de datos

1. Análisis (parsing)
2. Corrección
3. Estandarización
4. Mapeo (matching)
5. Consolidación

Análisis (Parsing)

Localiza e identifica **elementos individuales** en los archivos de origen y luego los aísla.

- **Ejemplos**
  - Análisis del primer nombre, segundo nombre y apellido;
  - Analizar número y nombre de la calle;
  - Analizar la ciudad y el estado.

Corregir

Corrige los componentes de **datos individuales** utilizando algoritmos sofisticados y fuentes de datos secundarias.

Acciones típicas con Datos Anómalos (Outliers):

- Ignorarlos.
- Eliminar la columna.
- Filtrar la columna.
- Filtrar la fila errónea, ya que a veces su origen, se debe a casos especiales.
- Reemplazar el valor.

Acciones contra Datos Faltantes (Missing Values):

- Ignorarlos.
- Eliminar la columna.
- Filtrar la columna.
- Filtrar la fila errónea, ya que a veces su origen, se debe a casos especiales.
- Reemplazar el valor.
- Esperar hasta que los datos

Estandarización

- Aplica **rutinas de conversión** para transformar los datos en formatos preferidos (y coherentes), utilizando reglas de estándarización.

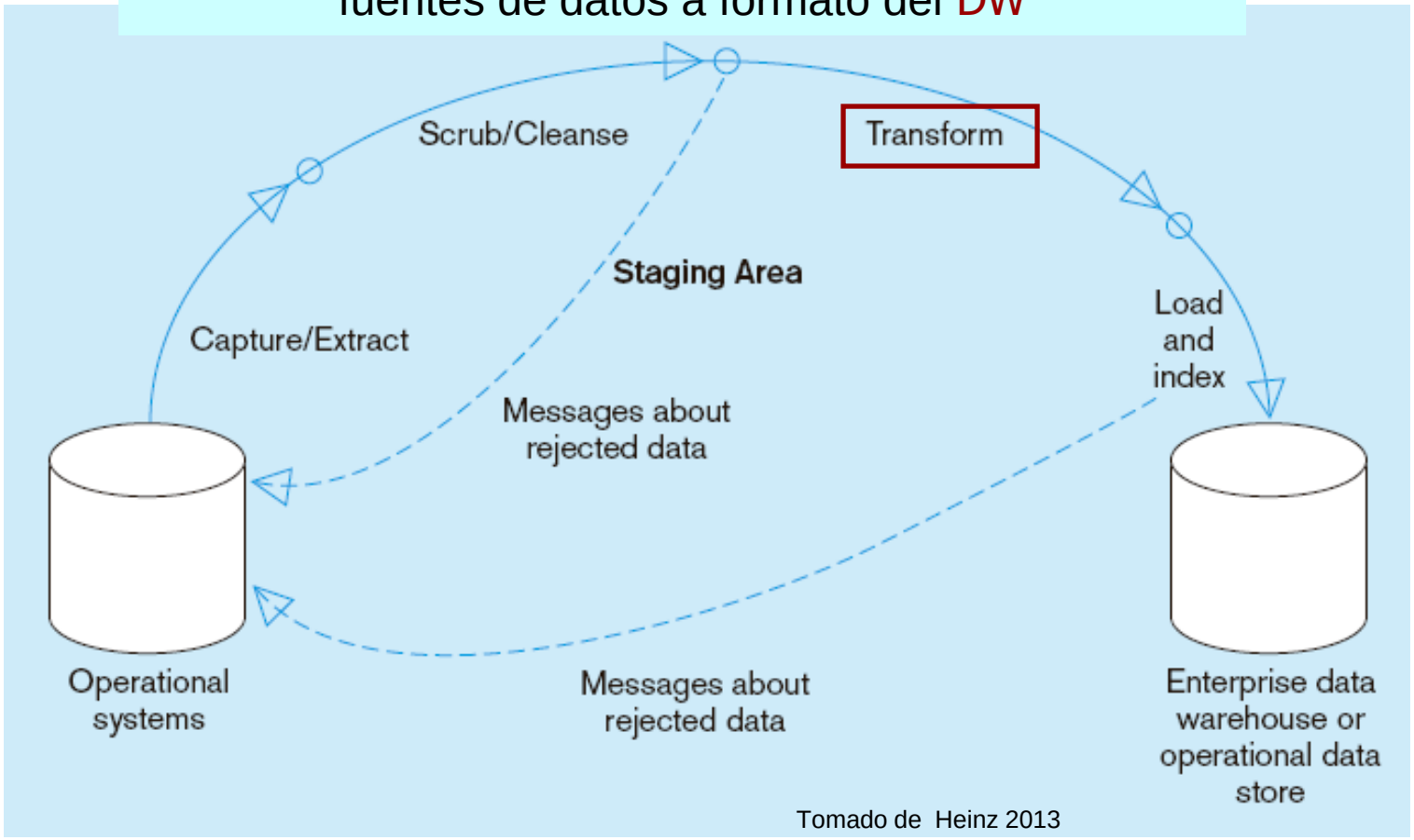
- **Ejemplos:** la sustitución de un apodo, uso de un nombre de calle preferido, etc.

Mapeo (matching)

- búsqueda de **coincidentes en registros** en los datos analizados, corregidos y estandarizados, basados en reglas predefinidas para eliminar la duplicación.

- **Ejemplos:** identificación de nombres y direcciones similares.

Transforma… convierten los datos desde las fuentes de datos a formato del DW



Tomado de Heinz 2013

A nivel de registro:

- Partición de Datos (selección)
- Juntar datos (combinación)
- Resumir datos (agregación)

A nivel de campo:

- de un solo campo: de un campo a un campo
- multi-campo: de muchos campos a uno, o uno a muchos campos

Transformación de datos

- Transforma los datos de acuerdo con reglas y normas establecidas

- Clásicos problemas:
  - Codificación.
  - Medida de atributos.
  - Convenciones de nombramiento.
  - Fuentes múltiples,

ordena **calcula**

resume **consolida**

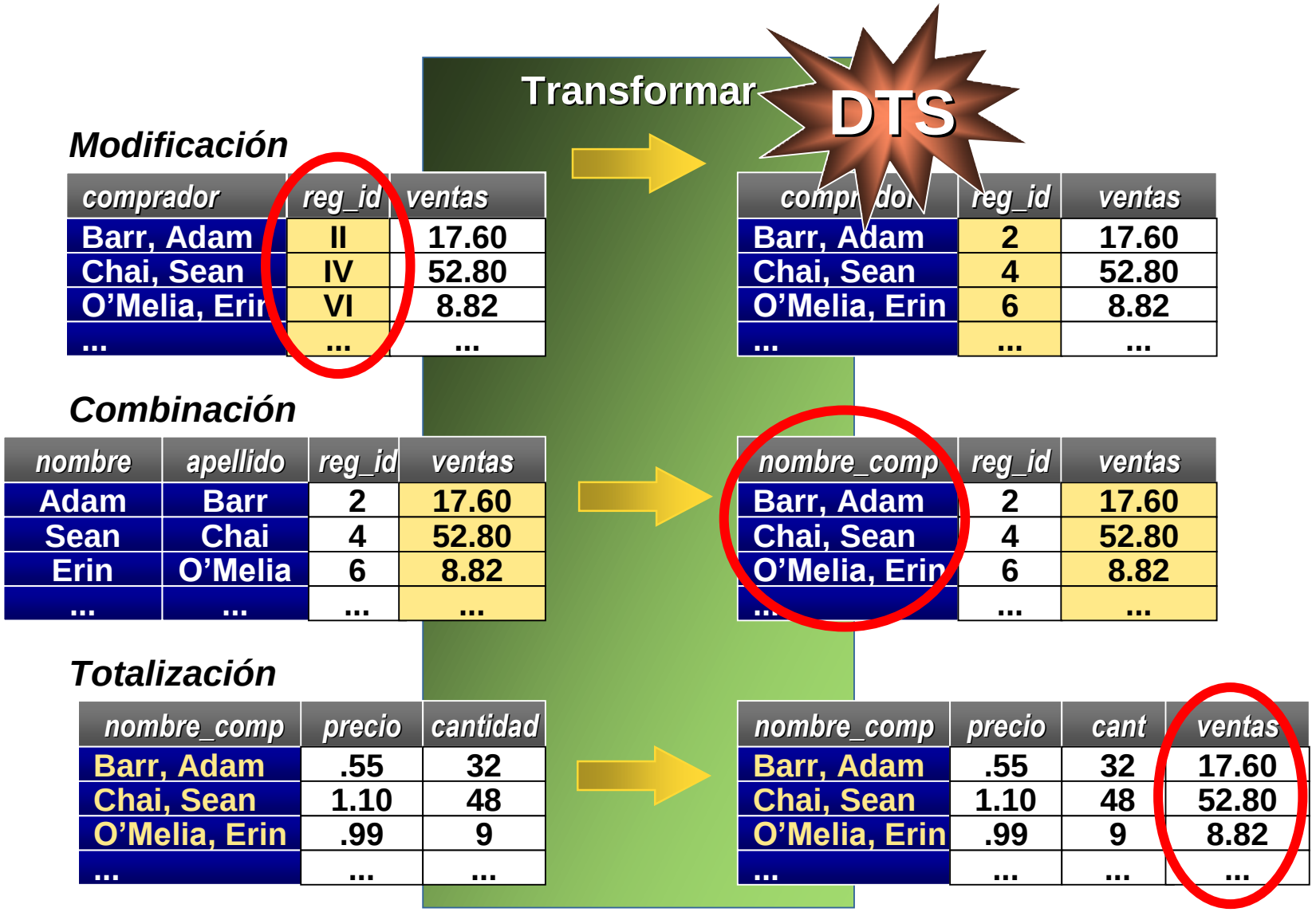
Transformación

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.

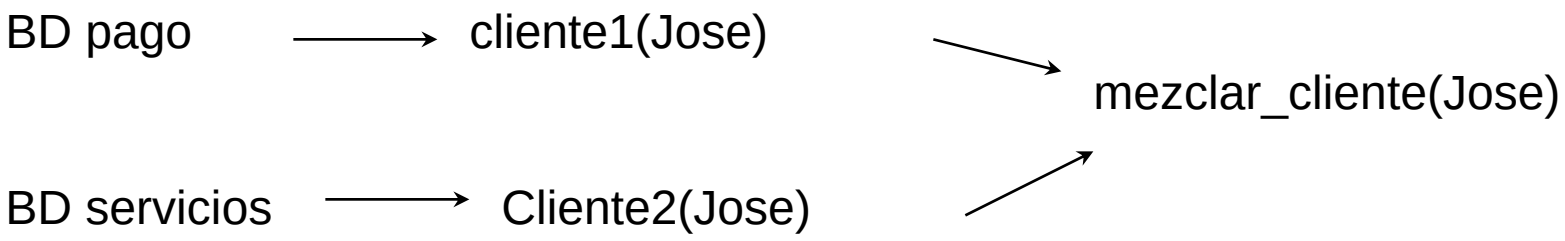
Proceso ETL; Consideraciones



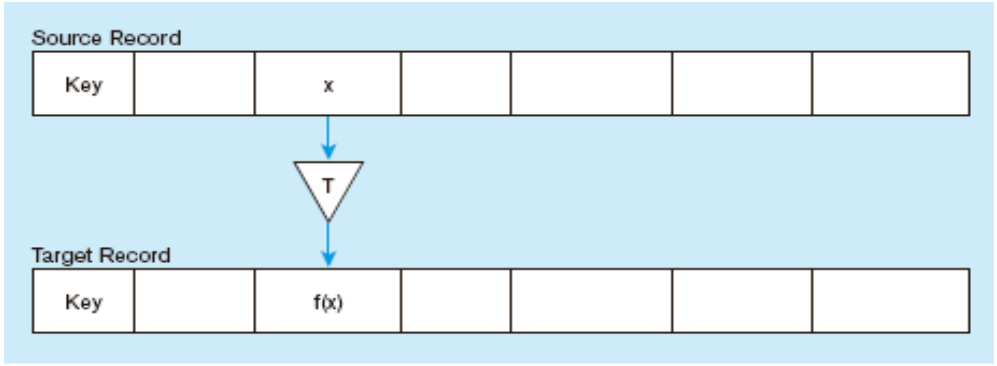
ETL: Transformación de datos

Posibles tareas

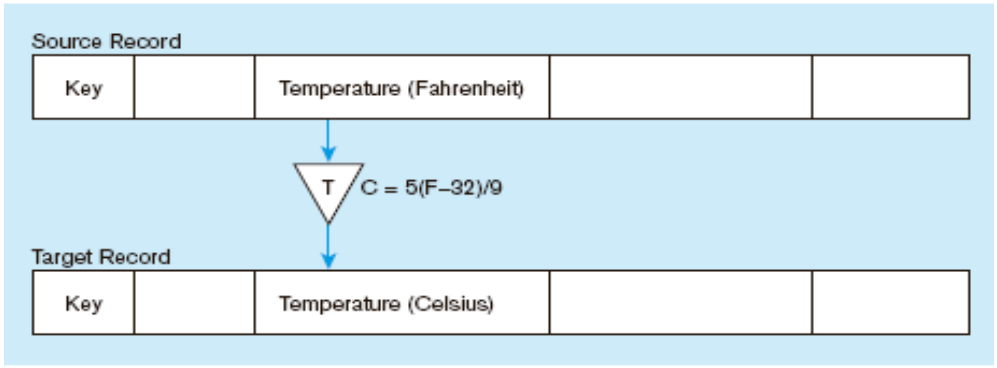
- **Migrar** (por ejemplo, yen a dólares)
- **Refinar**: utilizar el conocimiento específico de dominio (por ejemplo, números de seguro social)
- **Fusionar** (por ejemplo, lista de correo con la de clientes)



Transformación de un solo campo



Una función de transformación traduce los datos de una forma antigua a una nueva



Transformación algorítmica: utiliza una fórmula o expresión lógica

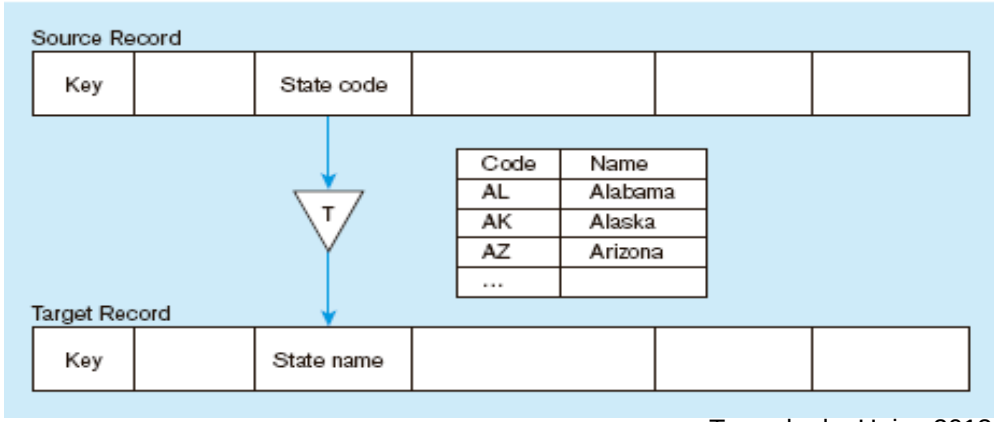
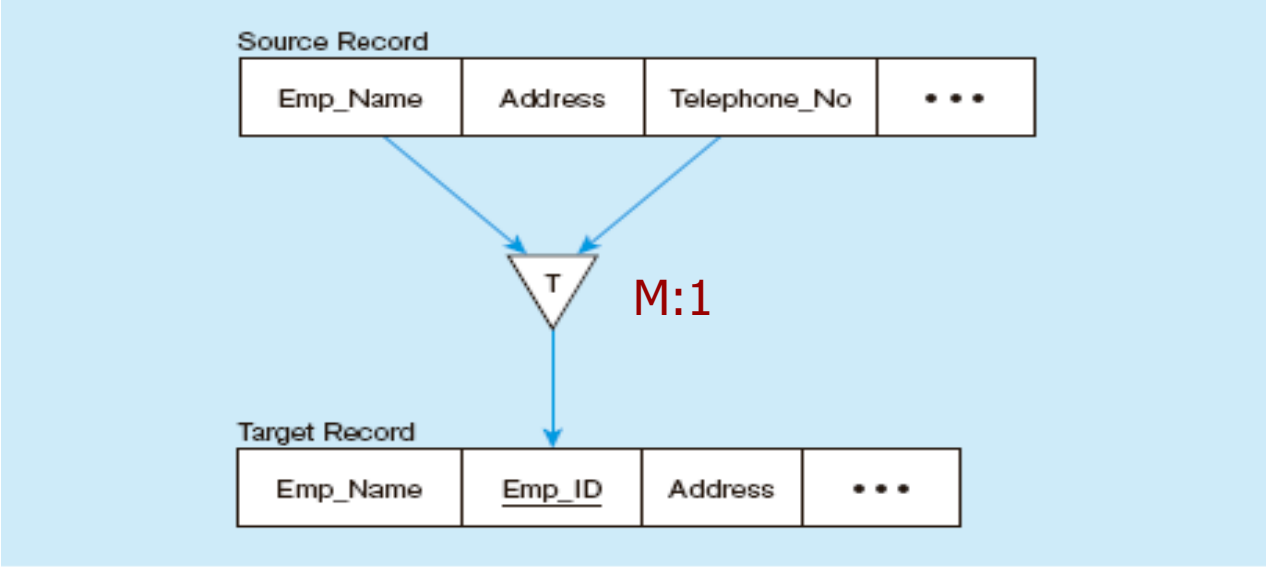
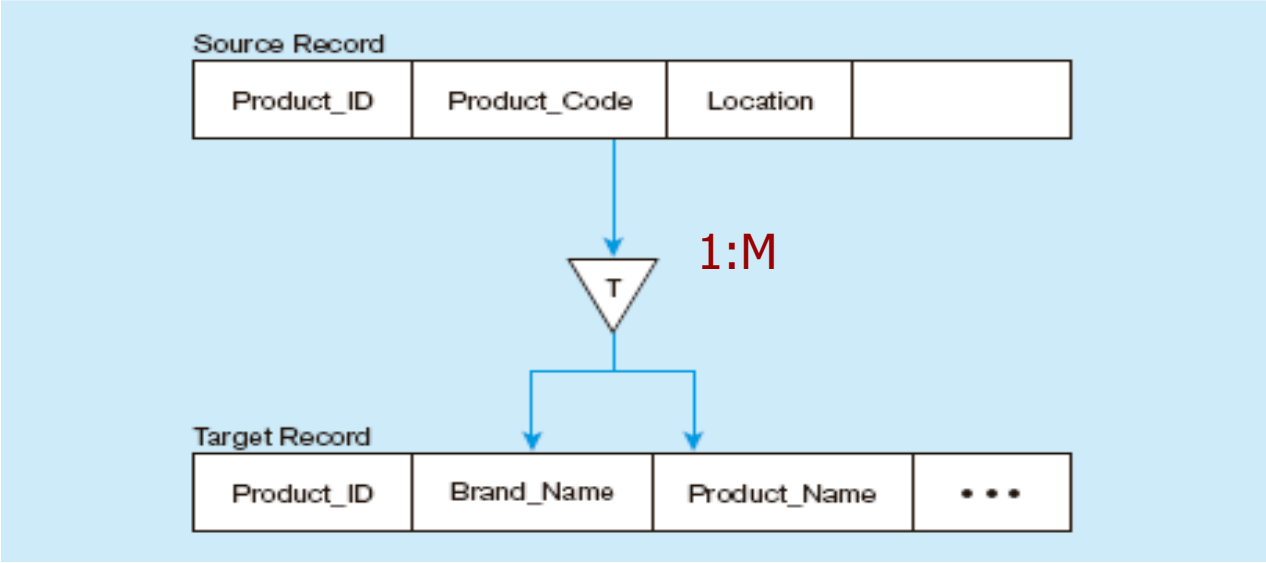


Table lookup: utiliza una tabla separada basada en códigos

Tomado de Heinz 2013

Transformación multicampos





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"}

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

Las fuentes 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 AdD, los errores y omisiones de inmediato comienzan a tomar significancia.

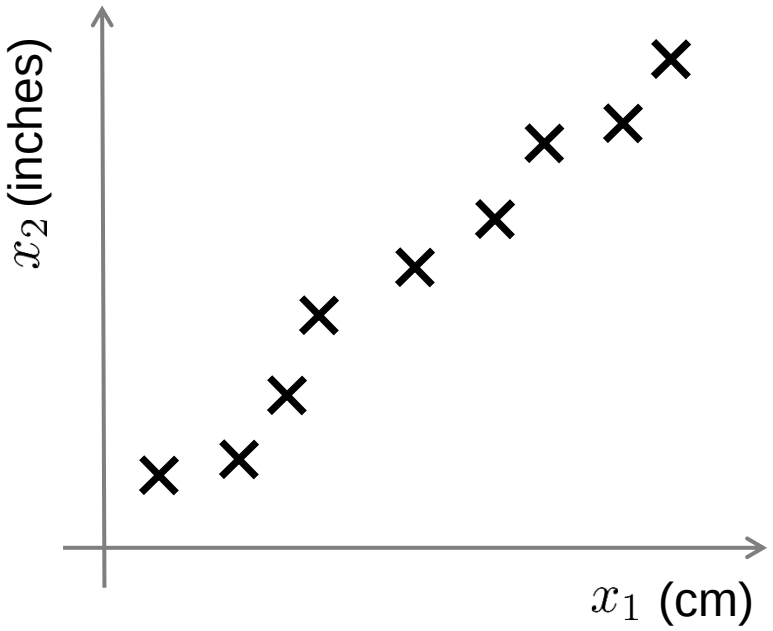
Por ejemplo:

- 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

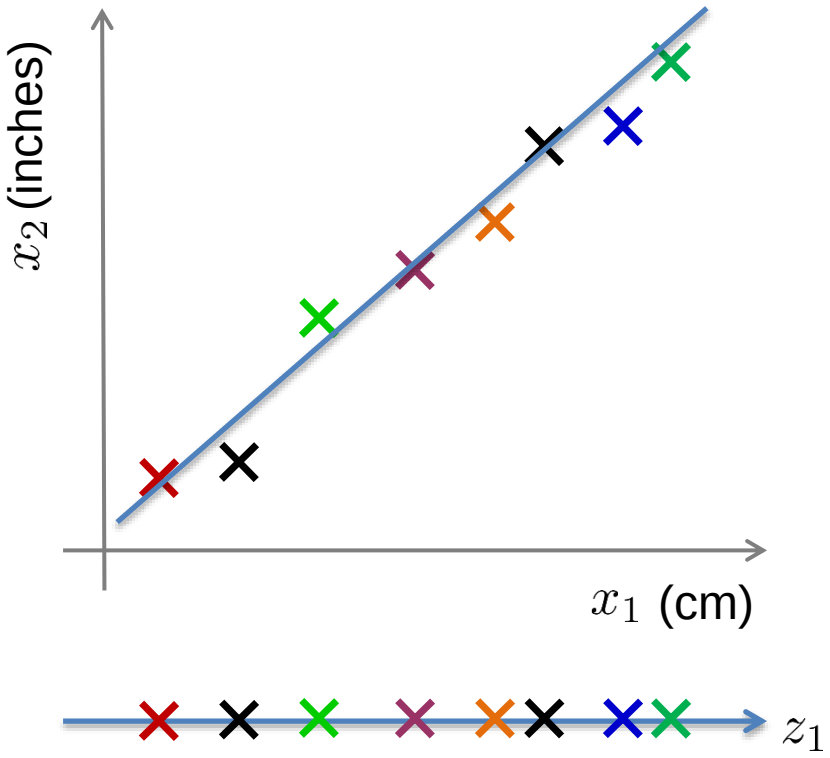
compresión de datos es la reducción del volumen de datos tratables para representar una determinada información empleando una menor cantidad de espacio.

Data Compression



Reduce data from 2D to 1D

Data Compression



Reduce data from 2D to 1D

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

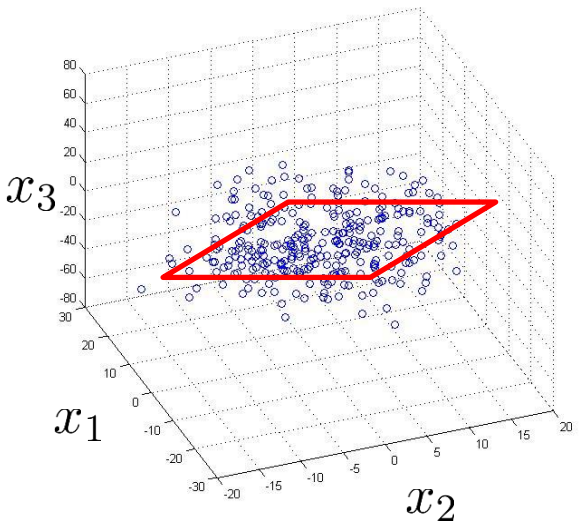
$$x^{(2)} \in R^2 \rightarrow z^{(2)}$$

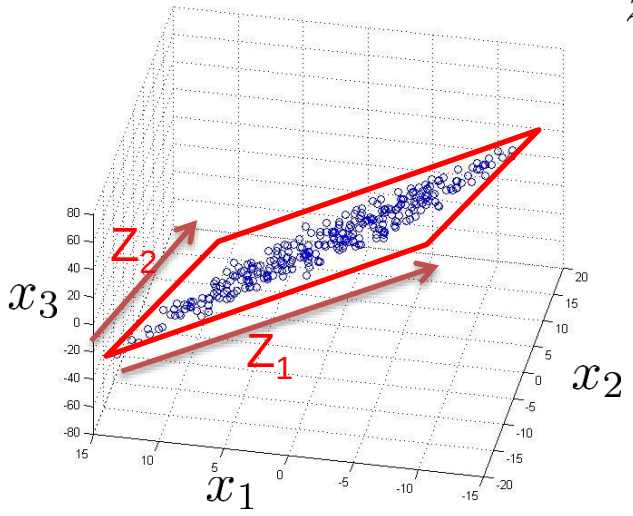
$$\vdots$$

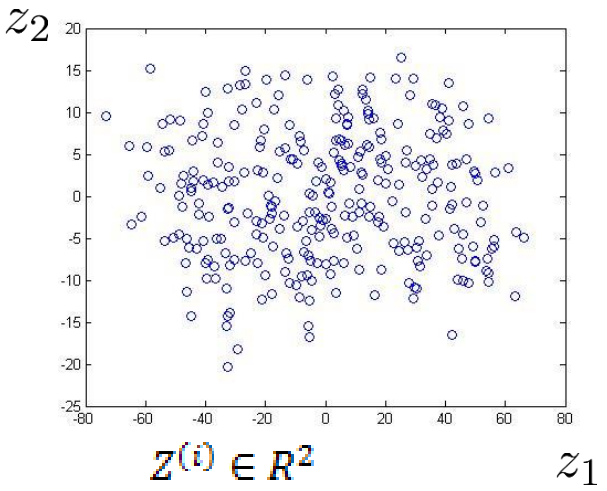
$$x^{(m)} \in R^2 \rightarrow z^{(m)} \in R$$

Data Compression

Reduce data from 3D to 2D







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

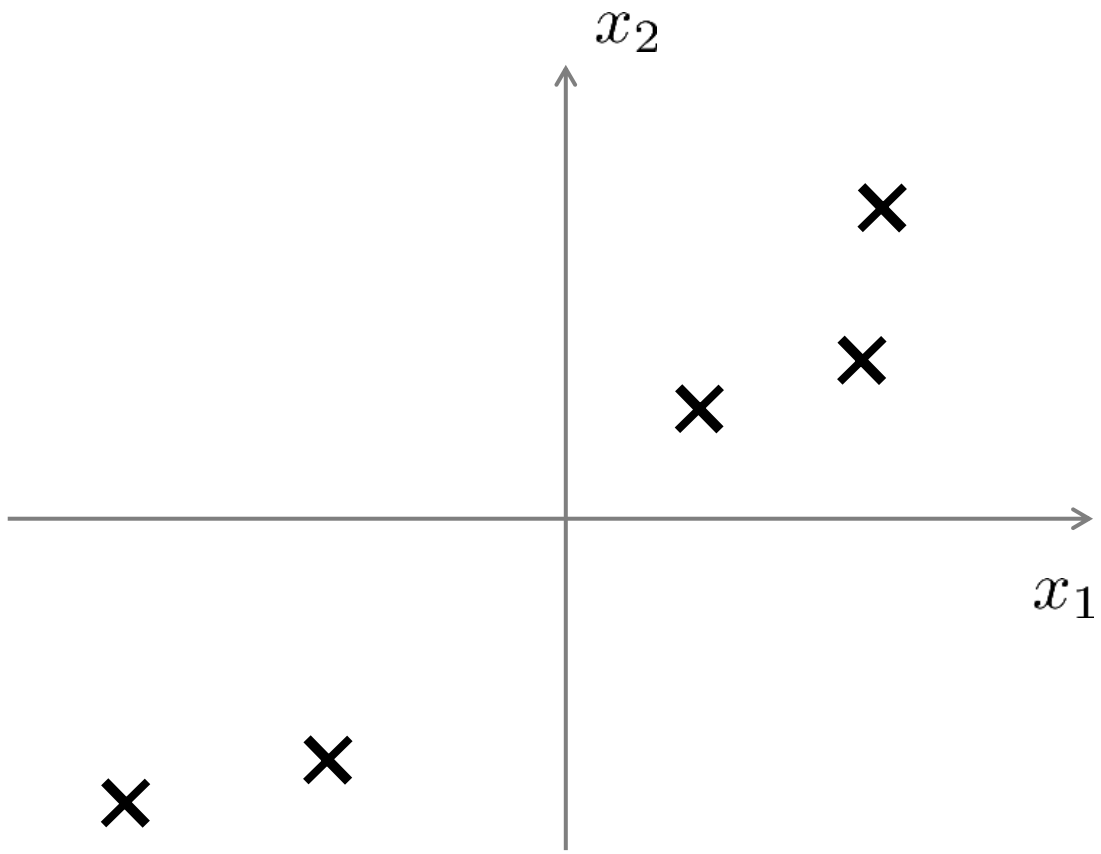
$Z^{(i)} \in R^2$

Formulación del Problema PCA (Principal Component Analysis)

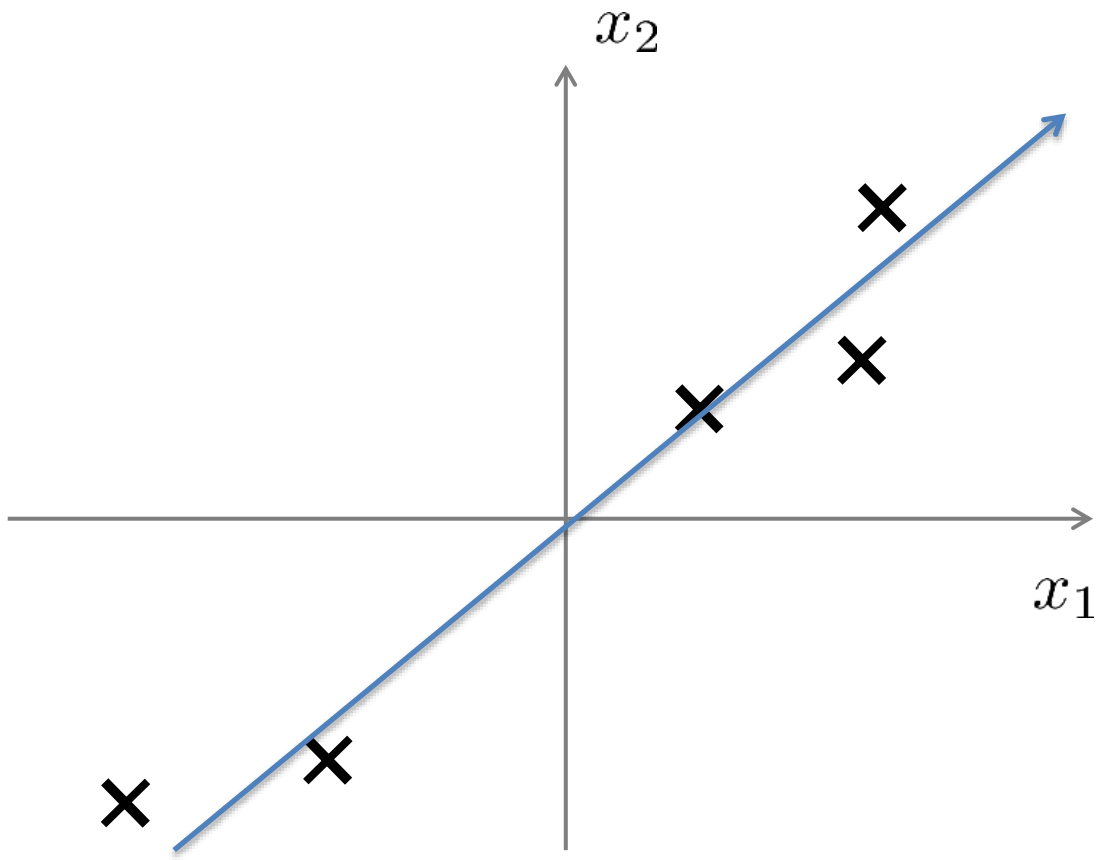
El **análisis de componentes principales** (en español **ACP**) es una técnica utilizada para reducir la dimensionalidad de un conjunto de datos.

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

Formulación del Problema PCA

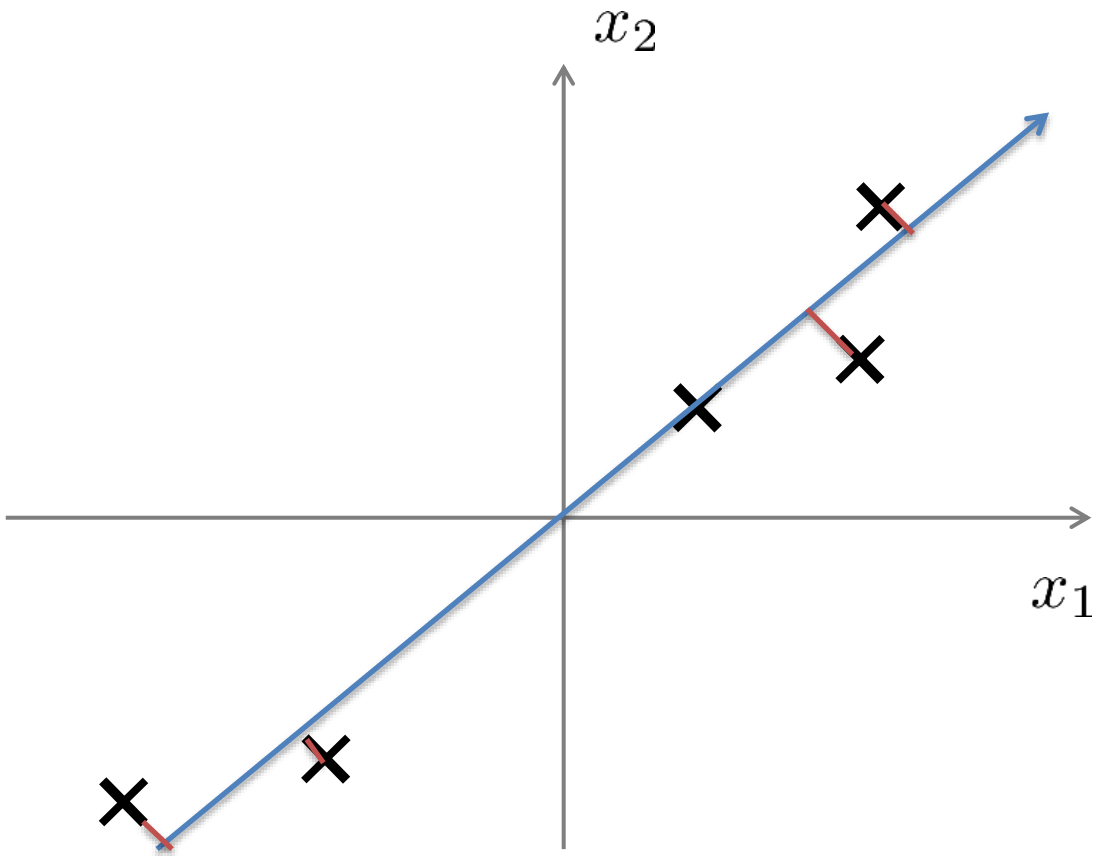


Formulación del Problema PCA



Formulación del Problema PCA

Error de proyección



Formulación del Problema PCA

Error de proyección

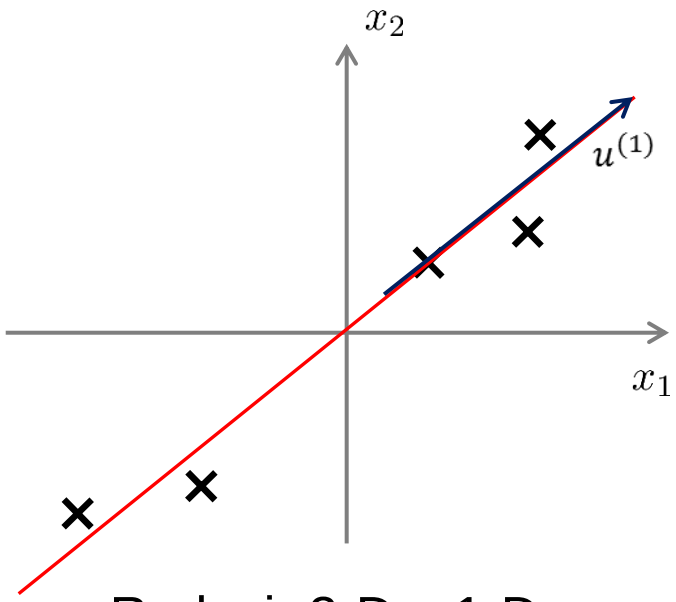
Para reducir de 2- dimensiones a 1-dimensión: Buscar un vector
Para proyectar la data, el cual minimice el error de proyección.

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

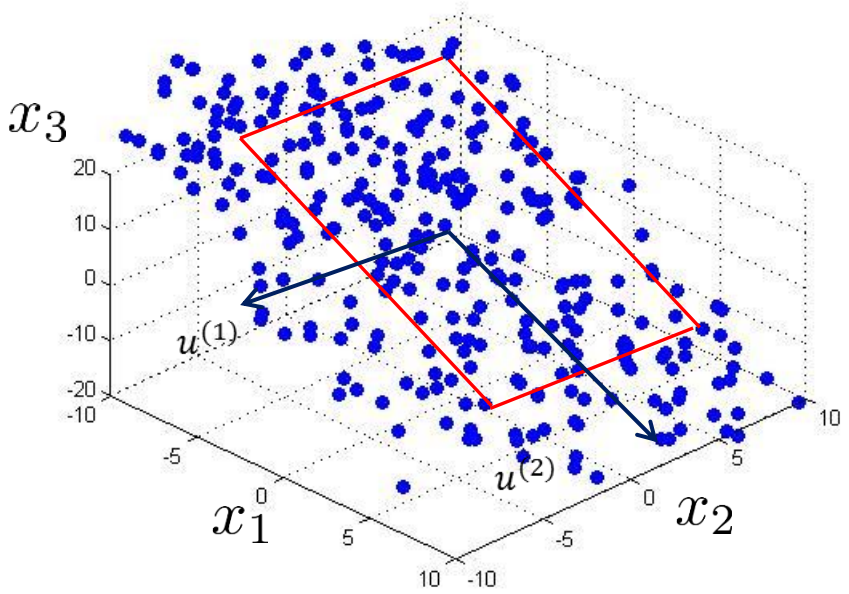
Para reducir de n-dimensiones a K-dimensiones: Buscar K vectores
Para proyectar la data, los cuales minimicen el error de proyección.

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

Formulación del Problema PCA

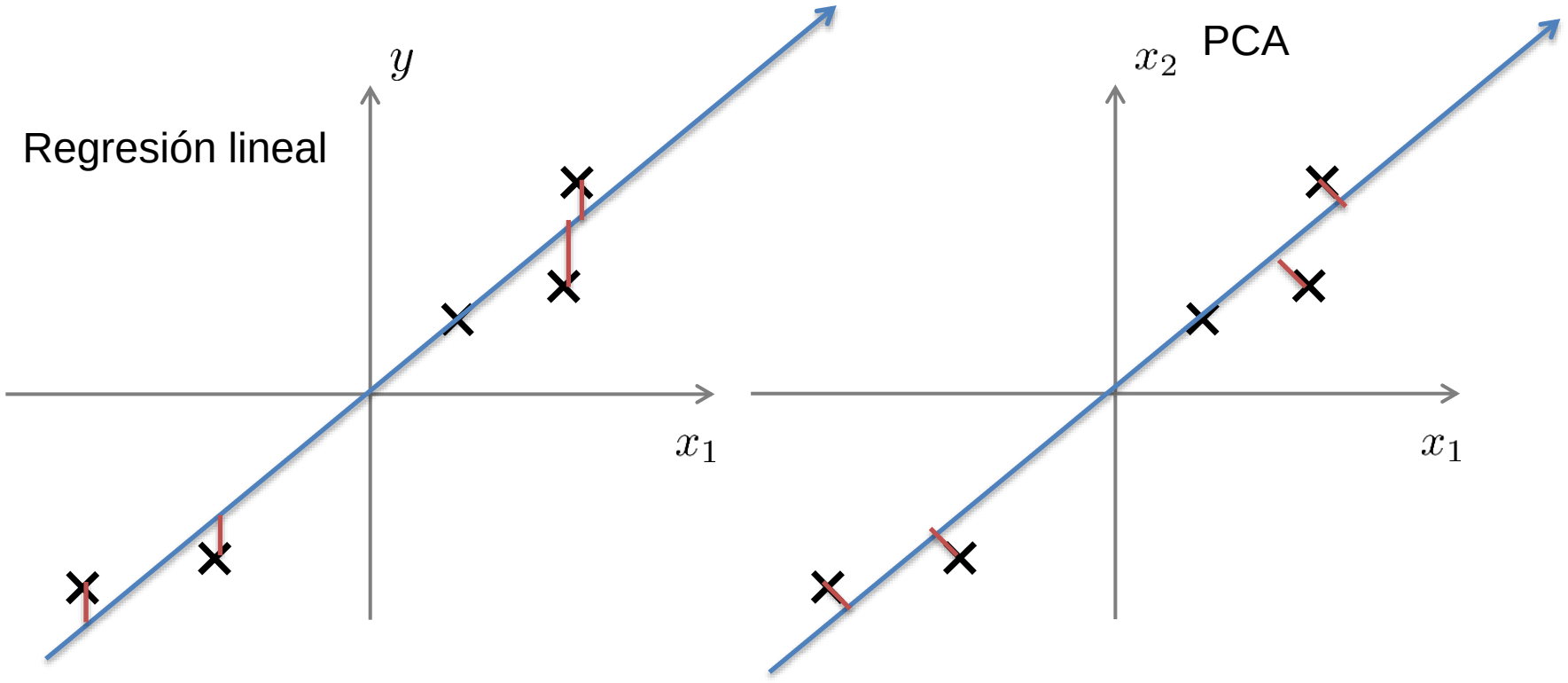


Reducir 2-D a 1-D



Reducir 3D a 2D

PCA no es regresión lineal



Preprocesamiento

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

Preprocesamiento (escalado de características/normalización):

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

Reemplazar cada $x_j^{(i)}$ con $x_j - \mu_j$

Si diferentes características en diferentes escalas (p.e., $x_1 =$ tamaño casa, $x_2 =$ número cuartos), escalar para tener rango de valores comparable

(PCA)

Reducir datos de n-dimensiones a k-dimensiones

Calcular “matriz covarianza”:

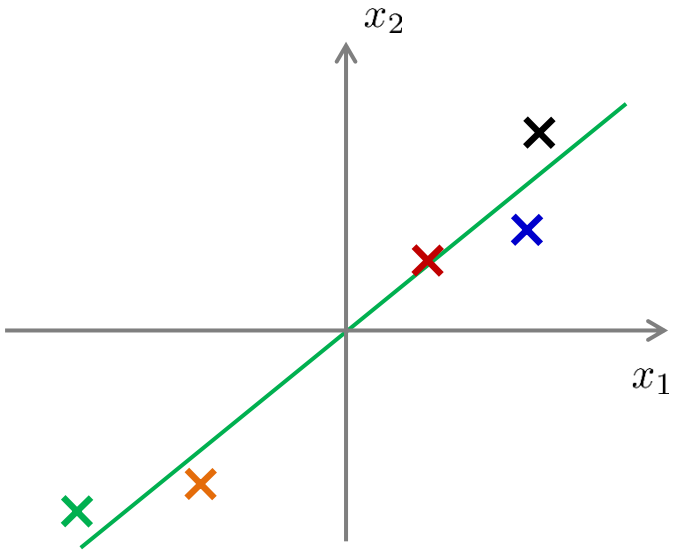
$$\Sigma = \frac{1}{m}\sum_{i=1}^{n}(x^{(i)})(x^{(i)})^T \longrightarrow \text{Sigma}$$

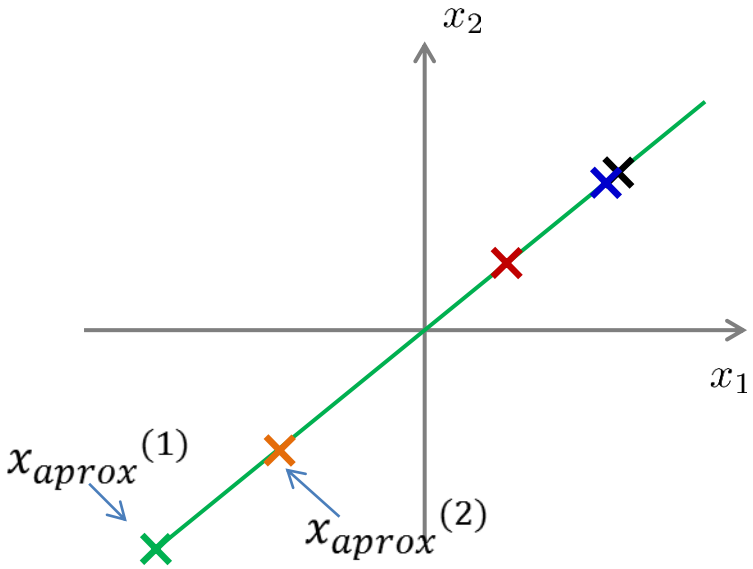
Calcular “Vectores propios” de la matriz Σ:

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

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

Reconstrucción

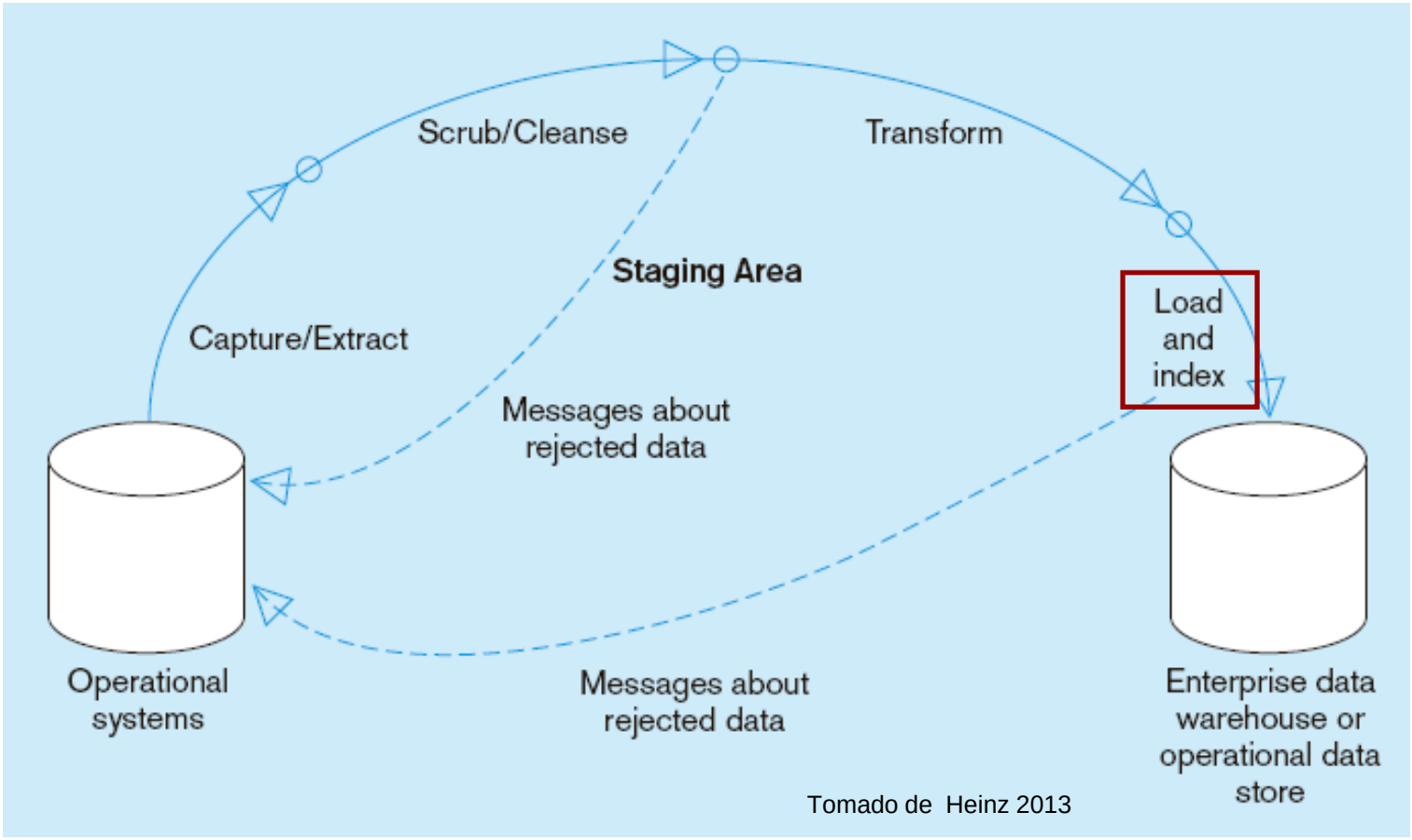




$$z = U_{reduce}^{T} x$$

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

Cargar/Indizar… Transforma datos y crea índices



Tomado de Heinz 2013

Modo de Actualización 1: reescritura masiva de datos en destino a intervalos periódicos

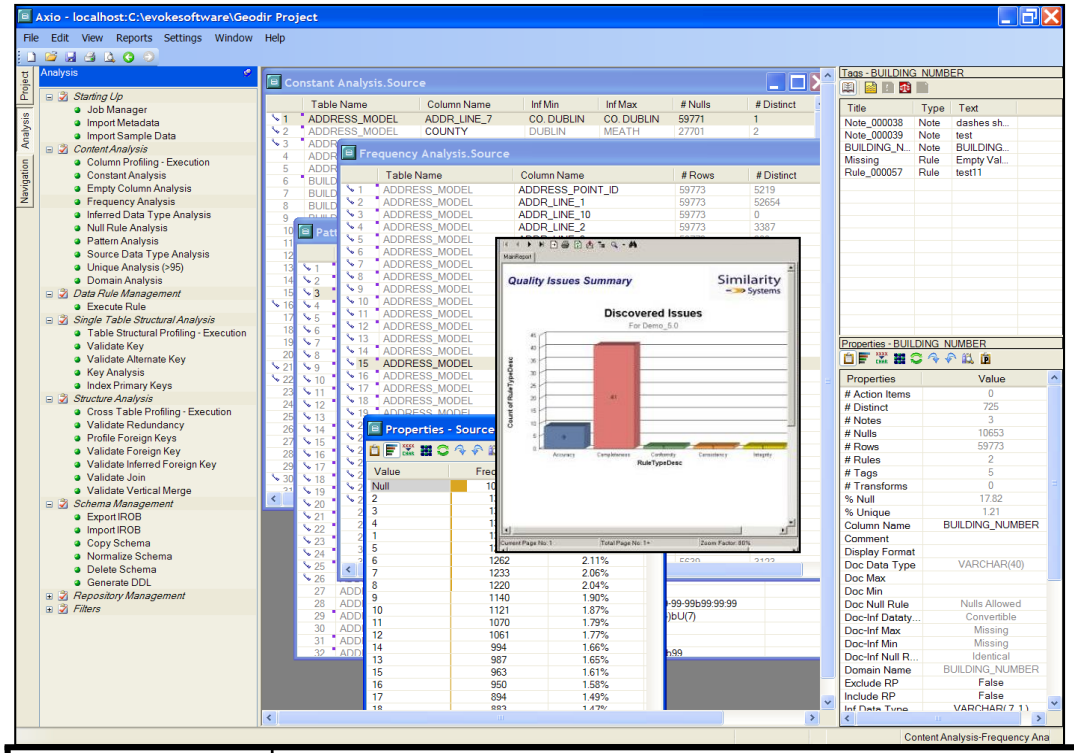
Modo de actualización 2: solamente cambios en los datos de origen se escriben en el warehouse

Carga de Datos

- Los datos físicamente se almacena en el almacén de datos
- La carga ocurre en una "ventana de carga“
- La tendencia cada vez mayor es actualizaciones en tiempo real

Preparación de los Datos

Perfil de los datos



Experto CD 	Perfila y etiqueta anomalías
Ùsuario 	Revisa anomalías

Catálogo de Problemas resueltos

- Completos
- Conformes
- Consistentes
- Precisos
- Duplicados
- Dependencias resueltas
- Correctas especificaciones y transformación
- **Íntegros**

¿Qué es lo Nuevo con AdDS?

Nuevas aplicaciones de AdD

p.ej.,
propagación del virus Ebola

Detección de brotes
Dos semanas por delante

Nuevos modelos para estimar propagación del virus

“cuáles ciudades están en mayor riesgo"

El modelo e basa en
Varias fuentes de datos,
Tipos y análisis.

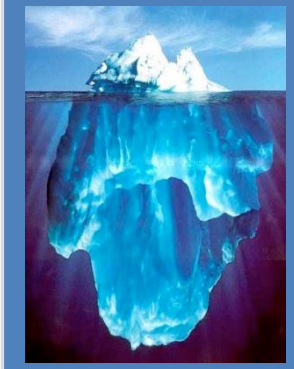
¿Qué es lo Nuevo con AdDS?

Captura, Curación y Agregación

- Ciencia de datos debe trabajar con:
  - Datos incompletos
  - Los datos suelen estar desordenados
  - Analizar los datos para ver qué información obtiene
  - Administrar Grandes conjuntos de datos

¿Qué es lo Nuevo con AdDS?

Todo está pasando en línea



Cada uno:
Hace clic
Ve anuncio
Factura un evento
Navega...
Solicita servidor
Realiza Transacción
Mensaje de error de red
...

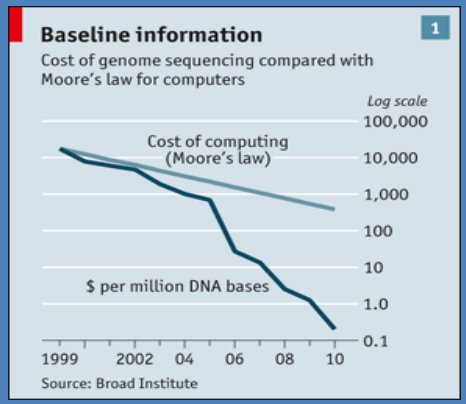
Generado por el usuario (Web y móvil)

...
..

Las fuentes

IoT

Computación Científica



Tipo de Datos

- Datos Relacionales (Tablas / Transacción / Datos Legados)
- Datos de texto (Web)
- Datos Semi-estructurados (XML)
- Grafos: Red social, Web semántica (RDF),
- Stream de datos

Tipo de Datos

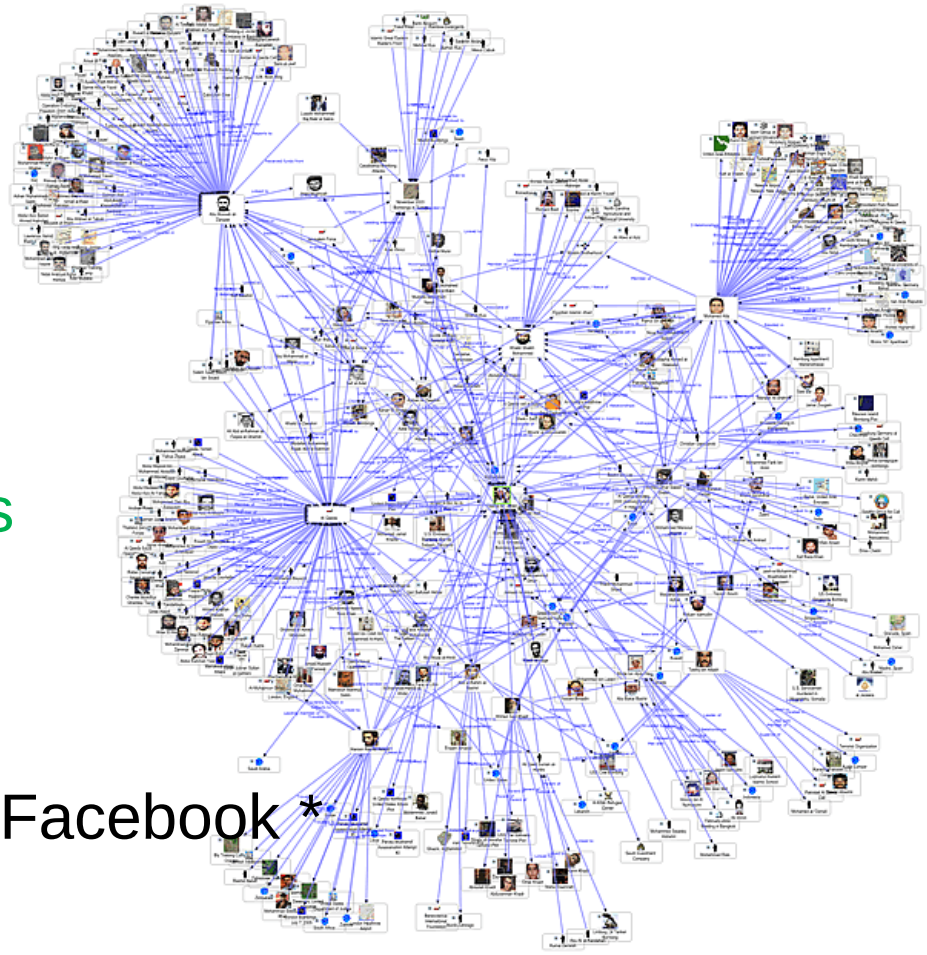
- Comercio electrónico
- Compras en tiendas de departamento / supermercado
- Transacciones bancarias / de tarjeta de crédito
- Red social
- Fotos, documentos,

Grafos:

Muchos datos interesantes
Tiene grafo:

- Redes sociales
- Redes de comunicación
- Red de computadoras
- Redes de carreteras
- Citas
- Colaboraciones / Relaciones

...



Algunos grafos pueden ser bastante grande (por ejemplo, Facebook * Gráfico de usuario)

Tipo de Datos

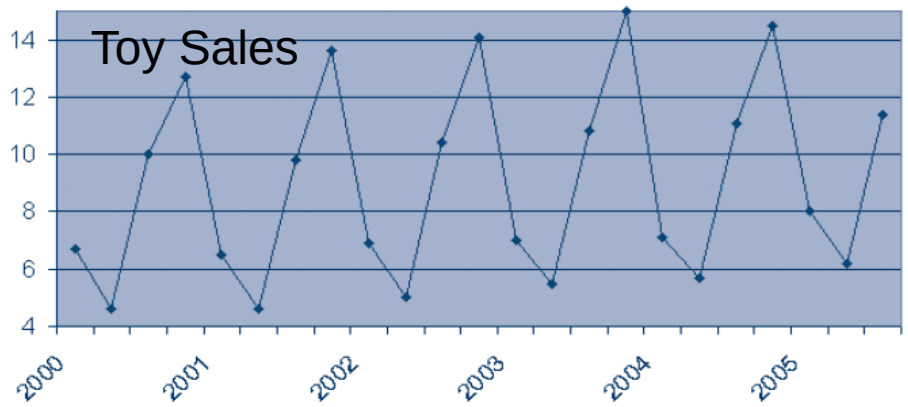
- **Red Social: Información de origen humano**
  - Redes sociales, Blogs, Documentos Personales, Imágenes, Vídeos, Búsquedas por Internet, Datos Móviles, Mapas generados por el usuario, E-mail
    Sistemas empresariales tradicionales: datos mediados por procesos

- **Agencias públicas (incluyendo registros médicos),**
  - producidas por negocios (transacciones comerciales, registros bancarios / de acciones, comercio electrónico, tarjetas de crédito)

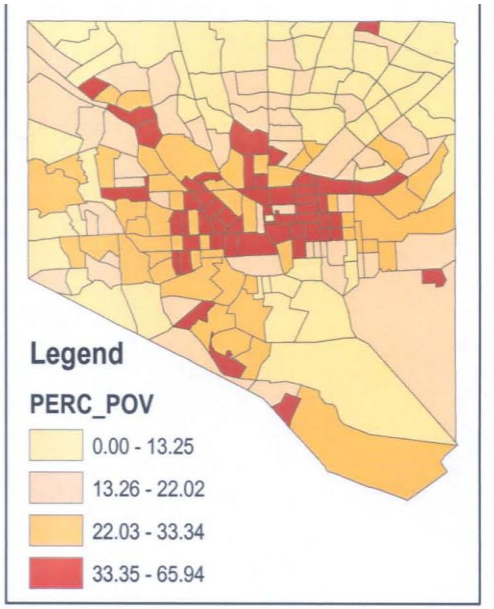
- **Internet: generado por la máquina**
  - Sensores fijos: domótica, sensor de tiempo / contaminación, tráfico, científico, seguridad / vigilancia
  - Sensores móviles: teléfono móvil, automóviles, imágenes de satélite
  - Sistemas informáticos: registros, registros web

Tipo de Datos

- **Series temporales:**
  - Modelar el precio del gas mediante una función de los precios recientes, la demanda, la geopolítica ...
  - Tendencias estacionales
- **SIG (sistemas de información geográfica)**
  - Longitud / latitud en la base de datos
  - Objetos: límites de ciudad/estado, ubicaciones de ríos, carreteras
  - Encontrar regiones con un exceso de cafeterías



from: Basic Statistics for Business and Economics, Lind et al (2009), Ch 16.



credit: Frank Curriero

Retos AdDS

- **Extraer, transformar, cargar (ETL)**
  - Necesitamos extraer datos de la (s) fuente (es)
  - Necesitamos cargar datos en el DW
  - Necesitamos transformarlos
  - Fuentes: archivo, base de datos, registro de eventos, sitio web,!

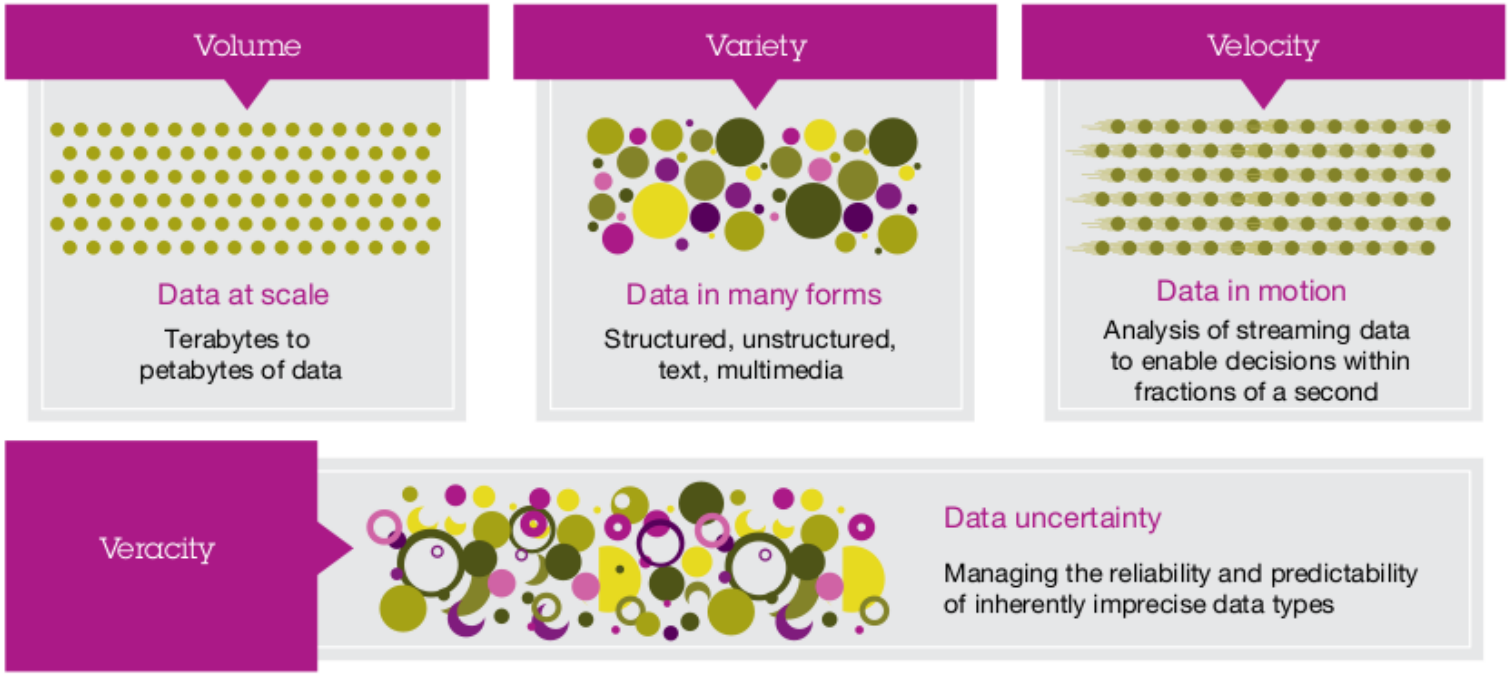
Retos AdDS

- **La construcción de un nuevo proceso de preparación de datos se realiza en muchas fases**
  - Caracterización de datos
  - Limpieza de datos
  - Integración de datos

- **Debemos mover los datos de manera eficiente en espacio y tiempo**
  - Transferencia de datos
  - Serialización de datos y deserialización (para archivos o red)

Retos AdDS

tracción de la percepción de un inmenso volumen, variedad y velocidad de los datos, en contexto, más allá de lo que era posible anteriormente.



Retos AdDS

¿Cómo se manejarán los datos?
¿Cómo se compartirán los datos?

Algunas reflexiones sobre "datos como un servicio“

- Establecimiento de normas y directrices. (Por ejemplo, arquitecturas abiertas)
- Creación de intercambios de datos específicos para la industria. (Por ejemplo, intercambios de datos sanitarios, intercambios de datos medioambientales, etc.)
- Creación de cruces de datos. (Por ejemplo, interactúan sin problemas datos ambientales con datos sanitarios)

Pasos de preparación

"acondicionamiento de datos“:

- La obtención de datos en un estado donde sea utilizable.
- Estamos viendo más datos en formatos que son más fáciles de consumir: servicios web, microformatos y otras nuevas tecnologías proporcionan datos en formatos que consumen directamente la máquina.
- Muchas fuentes de "datos salvajes" son extremadamente confusas.
- El acondicionamiento de datos puede implicar la limpieza de HTML desordenado, procesamiento de lenguaje natural para analizar texto.
- Es probable que se trate de una serie de fuentes de datos, todas en diferentes formas.

Pasos de preparación

Calidad de sus datos.

- Los datos suelen faltar o son incongruentes.
- Si faltan datos, simplemente ignorar lo faltante?

  Eso no siempre es posible.

- Si los datos son incongruentes,

  ¿usted decide que algo está mal?

- Si el problema involucra el lenguaje humano, la comprensión de los datos agrega otra dimensión al problema.

  La desambiguación nunca es una tarea fácil,

Capturar los datos

- Registrar datos Generados por diversas fuente.
- Mucho de esos datos no son interesantes
- Deben poder ser filtrados y comprimidos
- Datos recogidos espacial y temporalmente
- Reducción Inteligente de datos crudos
- Poder minimizar al humano la carga

Extracción y Curación

- Frecuentemente, la información recogida no esta en el formato listo para análisis.
- Expresa eso en un estructurado formar adecuada para el análisis.
- Debe ser correcta y completa
- Extracción es altamente compleja y dependiente
  p.ej., imagen Resonancia magnética diferente foto vigilancia foto)
  - Fidelidad
  - Ubicuidad
  - movimiento en el espacio

Integración, Agregación, Representación

- Dado heterogeneidad de los datos, no es suficiente grabarlos en un repositorio.
- Se requieren adecuados Metadatos,
- Para poder analizar esos Datos es importante Localizar, Identificar, comprender los datos.

Completamente Automatizado

- Diferentes estructuras de datos y semántica

Nadie sabe cómo curar datos

- Las fuentes de datos están fuera de control.
  - Se están extendiendo tanques de almacenamiento masivo de datos.
  - Te dicen muy poco, y a veces nada, sobre cómo utilizar los datos que se almacenan en ellos.
  - Lo que la gente realmente necesita son los datos y el conocimiento detrás de lo que los datos significan.

- El gran reto
  - Las instituciones tendrán que averiguar cómo recolectar datos a gran escala.
  - Saber 'cuándo' un trozo de datos cambia de significado es tan importante como saber 'qué' es ese elemento de datos.

Curar los datos

- Al igual que la mayoría de las cosas, 'limpiar' un conjunto de datos es tanto una habilidad como escribir una novela, tocar en una sinfonía
  - casi la mitad del tiempo se gasta en la limpieza y la preparación de los datos.
  - ¡La otra mitad se dedica a escribir los resultados!
  - "Hacer" el análisis real por lo general lleva sólo unos segundos.
- Reto- Si usted puede limpiar y preparar datos rápidamente, usted tendrá un inmenso éxito dentro.

Agregar y Desagregar

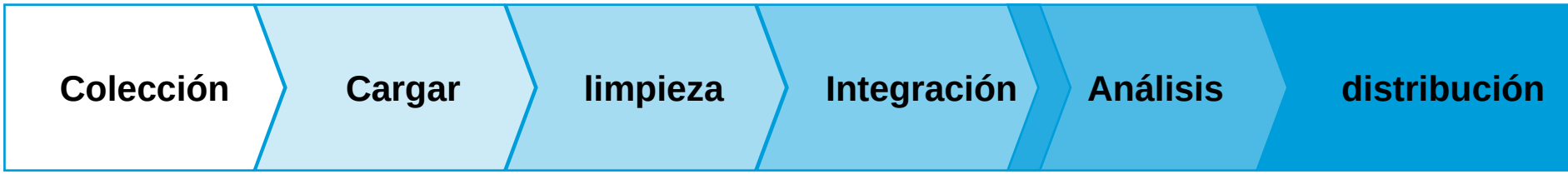
- Agregamos datos brutos para que surjan patrones.

- Desagregamos los datos para desenmascarar cosas, información.

  estamos tratando de encontrar patrones.

- Gran reto- la mayoría de las veces, hacer ambas cosas.

Cadena de Valor



Colección - Datos estructurados, no estructurados y semi-estructurados de múltiples fuentes

Cargar grandes cantidades de datos en un único almacén de datos

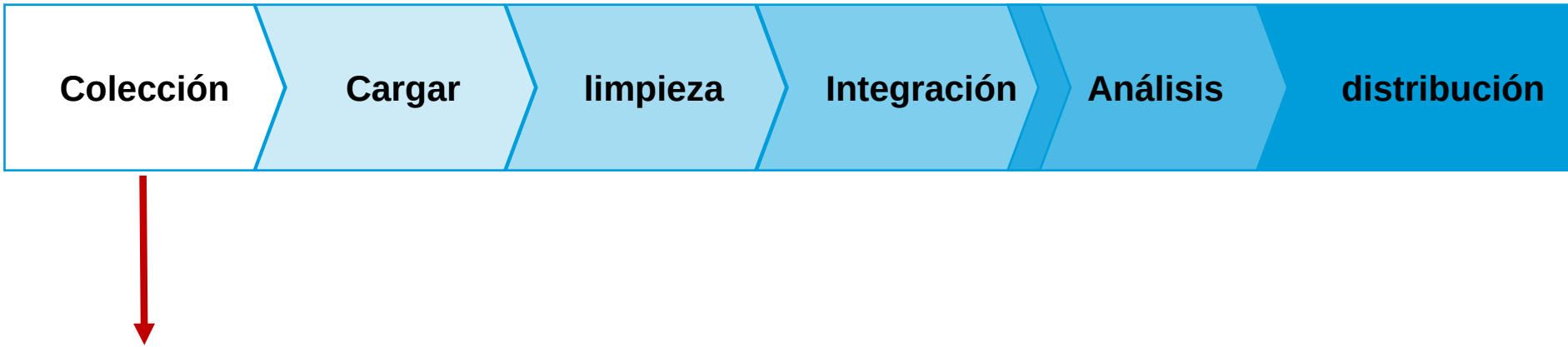
Limpieza- comprensión del formato y contenido; Limpieza y formateo

Integración - vinculación, extracción de entidades, resolución de entidades, indexación y fusión de datos

Análisis - Inteligencia, estadística, análisis predictivo y textual, aprendizaje automático

Entrega: consultas, visualización, entrega en tiempo real a la gerencia

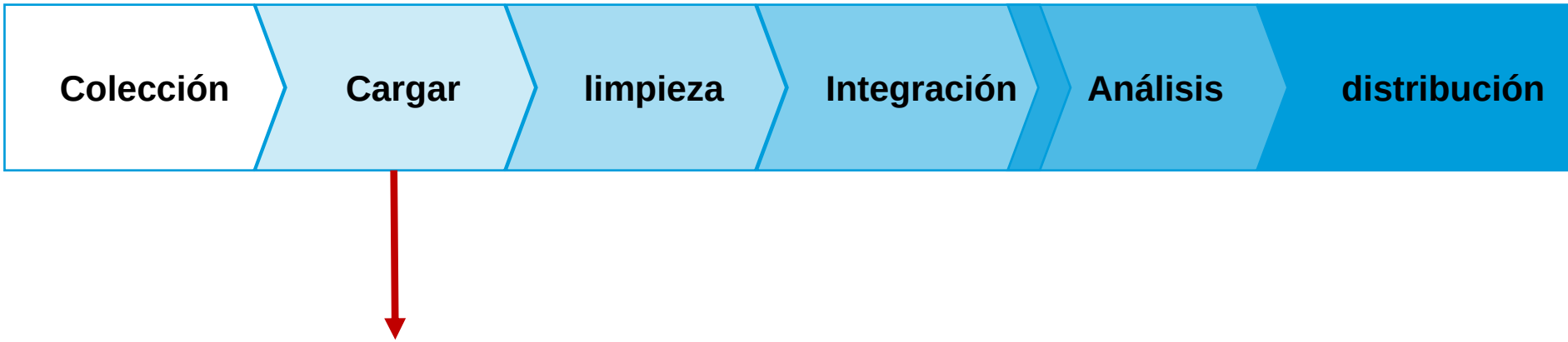
Cadena de Valor



Transacciones
Redes sociales
Flujos de datos

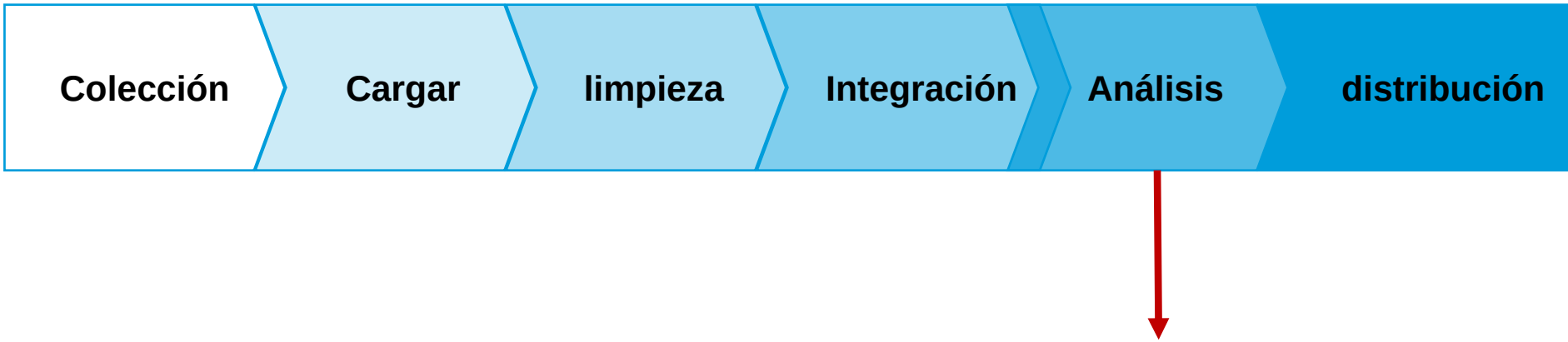
- Ambiental
- Industrial
- GPS
- Imagen / Video
- Datos de red
- Registros del sistema
- Datos financieros

Cadena de Valor



Seguridad
Calidad de los datos

Cadena de Valor



CA de tareas de AdD

Integridad de los datos

- valores faltantes
  - Cómo interpretar ¿no disponible? 0? Usar el medio
- Valores duplicados
  - Incluyendo cosas parciales (Jon Smith = John Smith?)
- inconsecuencia:
  - Varias direcciones por persona
- Datos desactualizados
- Uso inconsistente:
  - ¿Significa "destino" vuelo llegada?
  - Salarios que son negativos, o en los trillones

Interoperabilidad

- ¿Cómo pueden los datos ser comparados o combinados con otros?
- ¿Qué pasa si los datos se utilizan de manera diferente?
- Pensar en los registros médicos o de seguro
- Traducción / mapeo de términos
- Normas
- Unidades como ft / s, o galones, etc.
- Identificadores como SSN, UIN, ISBN

Curar los datos

- Rellenar los datos faltantes (valores de imputación)
  - Detección y eliminación de valores atípicos
- Suavizado
  - Eliminando el ruido promediando valores juntos
- Filtrado, muestreo
  - Manteniendo sólo valores representativos seleccionados
- Extracción de características
  - p.ej. En una base de datos de fotos, que la gente está usando gafas? Que tienen más de una persona? Que están al aire libre?

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

¿Qué genera la AdD?

- **Métodos Descriptivos**

Encontrar patrones interpretable que describen los datos.

- **Métodos de Predicción**

Utilizar algunas variables para predecir los valores desconocidos o futuros de otras variables.

MODELOS!!!

Modelos de Analítica

	Descriptivo	Predictivo	Prescriptivo
Preguntas	**Qué paso?** **Qué está pasando?** **Cuál es el problema?** **Qué acciones son necesarias?**	**Por qué esta pasando?** **Qué se producira?** **Por qué se producira?**	**Qué debería hacerse?** **Por qué debería hacerse?** **Qué pasa si se intenta eso?**
Habilitadores	• Reportes • Dashboards • Data Warehousing • Alertas	• Data Mining • Text Mining • Web/Media Mining • Forecasting	• Optimización • Simulación • Modelos de Decisión
Resultados	Bien definidos los problemas y oportunidades	Proyección de los futuros estados y condiciones	Mejores posibles decisiones y transacciones

Modelos de Analítica

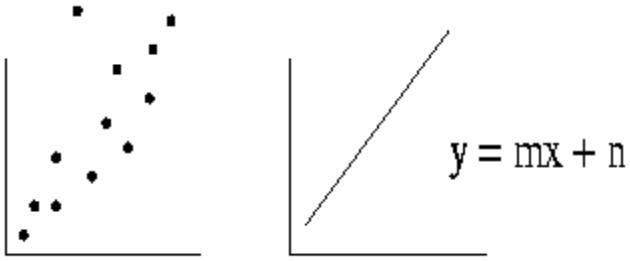
	Optimización	Identificación	Diagnóstico
Preguntas	**Qué puedo mejorar? Cómo mejorarlo?**	**Cómo es el modelo? Qué caracteriza a esos modelos?**	**Por qué sucede? Cuáles son las causas?**
Habilitadores	• Reportes • Modelos de mejora • Simulación	• Simulación • Formulas matemáticas	• Optimización • Simulación • Modelos de Decisión
Resultados	Mejores en la organización	Caracterización	Mejores posibles decisiones y transacciones

Definiciones iniciales

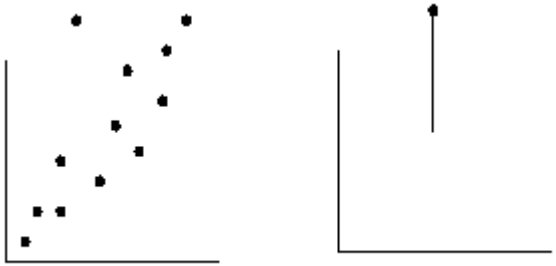
Conocimiento: Modelo vs. Patrón

Hand, Mannila y Smyth

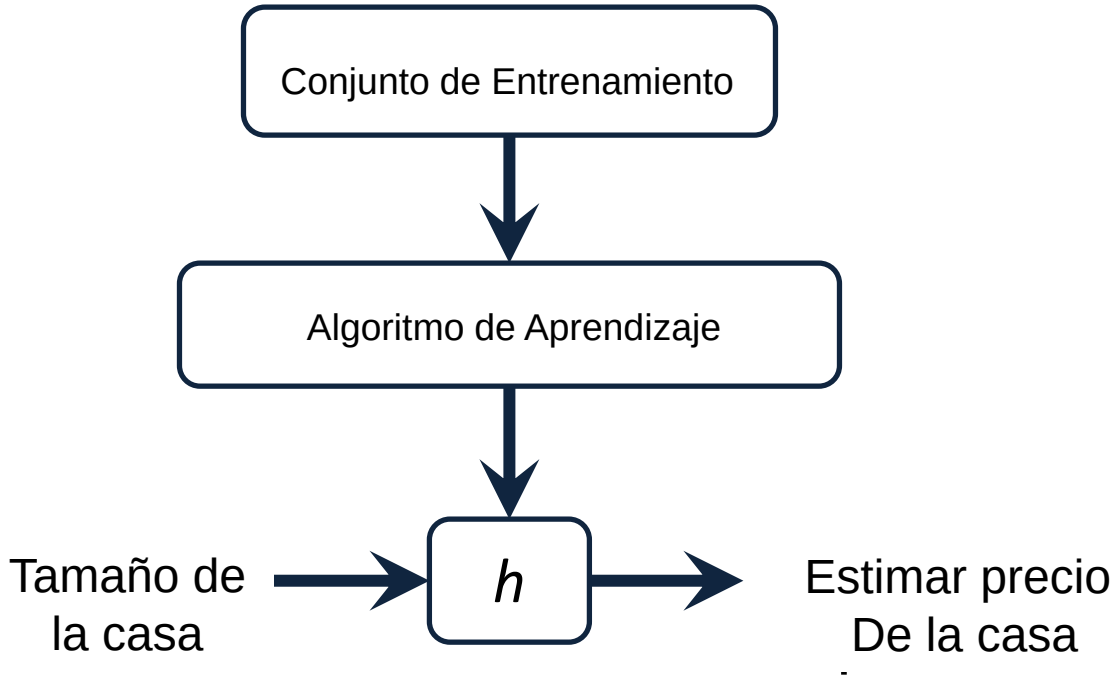
Modelo: Habla de todo el conjunto de datos



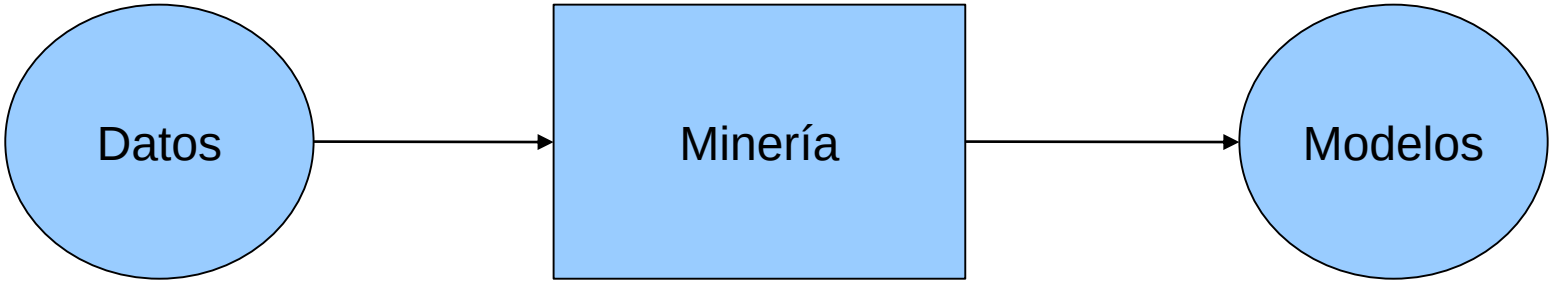
Patrón: Habla de una región particular de datos.



Construcción de modelos



Una visión simplificada de la minería de datos

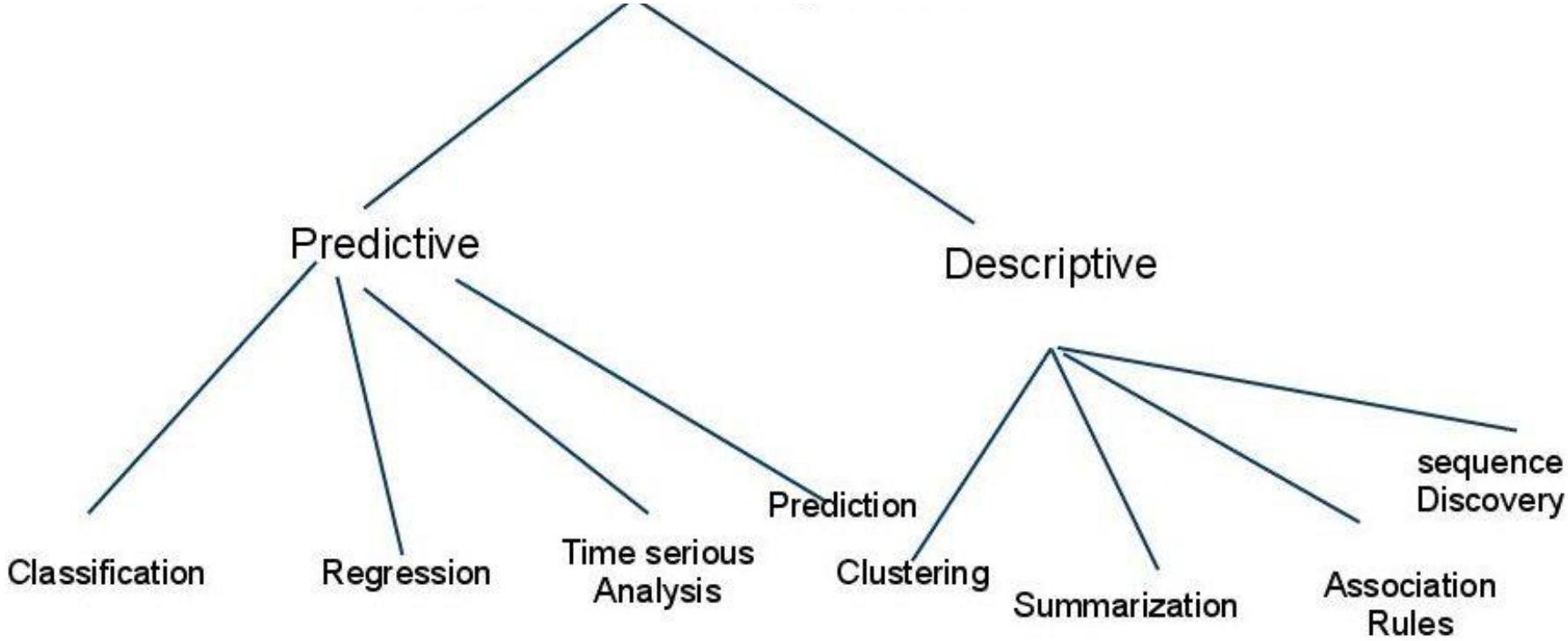


- Los “modelos” son el producto de la minería de...
- ...y dan soporte a las estrategias de decisión que se tomen

Tipos de Tarea de Analítica de Datos

Jose Aguilar

Las estrategias analíticas básicas:



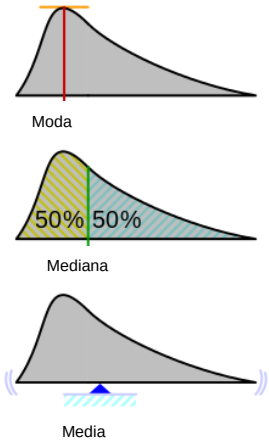
Herramientas de AD

- La minería de datos
- El análisis estadístico
- El análisis predictivo
- La Correlación
- La Regresión
- Pronosticar
- Modelado de procesos
- Optimización
- Simulación

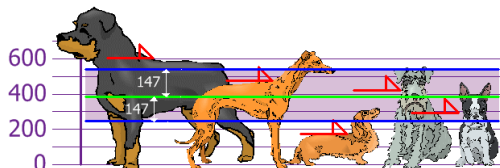
Dos categorías principales:
* Estadísticas descriptivas
* Estadística inferencial

Las estadísticas descriptivas básicas

- Usar **medidas de resumen** para describir la tendencia central de una distribución (media, moda, mediana)
- Utilizar la **dispersión o variabilidad** (desviación estándar, varianza, y el rango) para saber cómo se extienden los datos alrededor de la media.



- Frecuencias (contar)
- Porcentaje
- Media (suma de todos los valores ÷ no. de valores)
- Moda (valor más frecuente)
- Mediana (valor medio o posición central)
- Rango (intervalo entre el valor máximo y minimo)
- Desviación estándar (variación esperada con respecto a la media)
- Varianza (la esperanza del cuadrado de la desviación)
- Rankeo (clasificar, ordenar)



Compradores	Número
Hombre	
Viejo	6
Joven	4
Mujer	
Vieja	10
Joven	15

- **Más compradores femeninos que compradores masculinos**
- **Más jóvenes compradores femeninos que los compradores varones jóvenes**
- **Compradores masculinos jóvenes no están interesados en comprar en el centro comercial**

Las estrategias analíticas básicas:

Describiendo

Factorización

Agrupación

Comparando

Clasificación

encontrar puntos comunes

encontrar covarianza

Descartar alternativas

Las estrategias analíticas básicas:

- Clasificación
- Pronóstico (Predicción)
- Asociación (reglas)
- Agrupación o segmentación (Clustering)

Tipos de aplicaciones de la AD

- Clasificación [**predictivo** y descriptivo]
- Clustering [**descriptivo**]
- Descubrimiento de Regla Asociación [**descriptivo**]
  - Análisis de dependencia de datos
  - correlación y causalidad
- Descubrimiento Patrones Secuenciales [**descriptivo**]
  - Análisis de series de tiempo, asociaciones secuenciales
- Regresión [**predictivo**]
- Detección de Tendencia y Desviaciones [**predictivo**]
- Filtros Colaborativos [**predictivo** y descriptivo]
- Resumir [**descriptivo**]
- Descripción de Conceptos [**descriptivo**]
  - Descripción de características
  - Descripción de su identidad discriminante

Clasificación

Examinar las características de un nuevo objeto y **asignarle** una clase o categoría de acuerdo a un conjunto de tales objetos previamente clasificados.

- Ejemplos:
  - Clasificar los estudiantes en categorías según sus rendimiento: bajo, medio y alto
  - Detectar los estados operacionales de un sistema: con falla, seguro, inactivo.

Pronóstico

Predecir un valor futuro con base a valores pasados

Prognosis **Predicción**

- Ejemplos:
  - Predecir cuánto efectivo requerirá un cajero automático en un fin de semana

Asociación

Determinar cosas u objetos que van juntos

- Ejemplo:
  - Determinar que productos se adquieren conjuntamente en un supermercado

Agrupación o segmentación

Dividir una población en un número de grupos más homogéneos

- No depende de clases pre-definidas a diferencia de la clasificación

- Ejemplo:
  - Dividir la base de clientes de acuerdo con los hábitos de consumo
  - Establecer los grupos de estudiante según sus estilos de aprendizaje