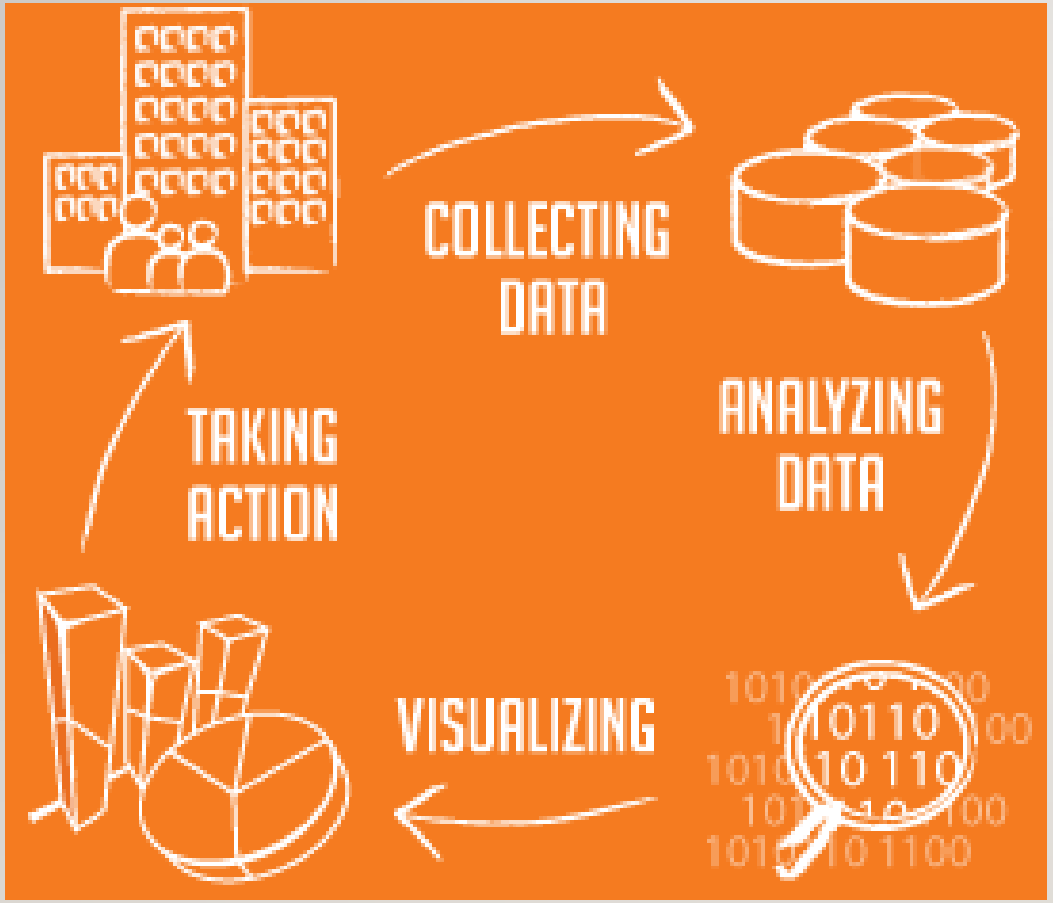
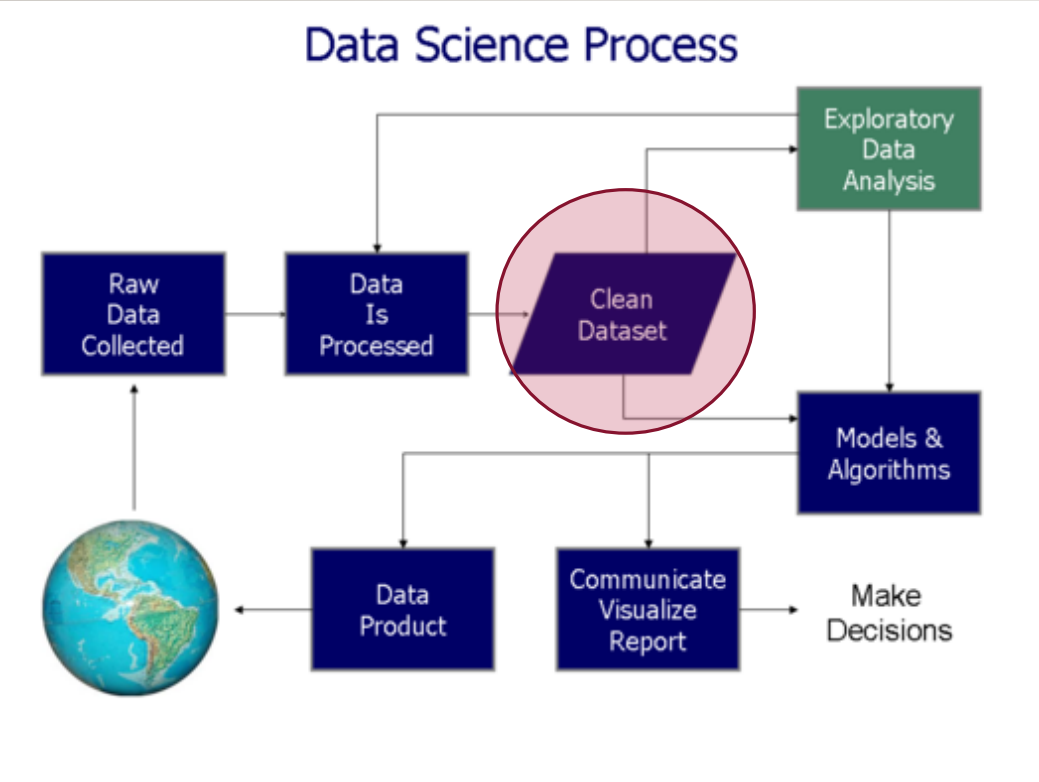
PREPROCESAMIENTO DE DATOS

TRATAMIENTO DE DATOS MASIVOS – RAFAEL CABALLERO



DÓNDE ESTÁ WALLY?

http://blog.cobia.net/cobiacomm/wp-content/uploads/2014/03/bigdata-action.png



AQUÍ, CASI, CASI

www.quora.com/What-is-the-workflow-or-process-of-a-data-scientist-What-tools-do-they-use

Combinación

¿Duplicados?

¿Similares?

Missing Data

¿Datos raros? (outliers)

Estandarización

Variables categóricas

Escalado

Proporción

Nuevas dimensiones

COMBINACIÓN, DUPLICADOS

- Lagos de datos → datos de muchas fuentes
- Identificarlos, buscar campos comunes → ¿clave primaria?
- Campos que faltan → dejar el hueco (missing data) o suprimir en todos
  - **No "inventar" valores.**
- Duplicados → eliminarlos
- Parecidos pero no iguales, criterios de unificación
  - Monedas, fechas: comprobar formatos y buscar uno único
  - Textos ¿similitud textual?

MISSING DATA

En algunas columnas puede faltar ocasionalmente un valor

- Si falta prácticamente siempre y no es muy importante → eliminar la columna
- Algunas herramientas permiten trabajar con datos dispersos → dejar el hueco
- En caso contrario, 2 posibilidades:
  - Eliminar las filas incompletas
  - Completar el dato ¿media de valores? ¿utilizar inferencias estadísticas o machine/learning?

Nunca poner valores "mágicos"!!!

DATOS RAROS

Datos erróneos o muy fuera de rango pueden estropear el futuro análisis

- Para detectarlos:
  - Intentar detectar la distribución de la columna ¿es realmente raro?
  - Mostrar los datos de esa columna (histogramas, boxplots)
- ¿Se puede comprobar que son erróneos o muy inusuales?
- Eliminarlos… a no ser que justo esos datos sean nuestro objetivo
  - Si es el caso → Idea: cambiar los datos normales por 0 y los raros por 1 (variable dummy)

ESTANDARIZACIÓN

Importante para muchos algoritmos que utilizan la noción de distancia.

- Distintas forma de hacerlo
- Comprobar en cada caso cuál es la más adecuada
- Se pueden requerir varias, una por método

$$x' = \frac{x - \text{average}(x)}{\max(x) - \min(x)}$$

$$x' = \frac{x - \bar{x}}{\sigma}$$

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

TIPOS DE DATOS

Datos numéricos → No hay problema

Ojo con los datos aparentemente numéricos (códigos, escalas de Likert)

Datos categóricos → dan lugar a n-1 variables dummy, con n el número de categorías

Ejemplo:

Codificar la columna "tipo de abono transporte" que puede tomar los valores "infantil", "estudiante", "adulto", "tercera edad"

OTRAS TRANSFORMACIONES

En el caso de sensores es muy normal encontrar series temporales →

El mismo valor se repite en periodos de tiempo (posiblemente constantes)

Requiere preprocesados específicos → diferencias de valores

SI QUEREIS IR REALMENTE EN SERIO

- Se debería conprobar la calidad de los datos, y medirla según se modifican (se quitan missing, se estandariza)
- Se deben guardar los datos originales y los scripts que realizan cada paso, de forma que se pueda reconstruir el proceso de transformación → workflow

ENLACES Y BIBLIOGRAFÍA DE INTERÉS

- Sobre calidad de datos: THE SIX PRIMARY DIMENSIONS FOR DATA QUALITY ASSESSMENT
- Sobre escalado, un buen blog
  - http://www.synergicpartners.com/precauciones-a-la-hora-de-normalizar-datos-en-data-science/
- Sobre variables dummy y cómo usarlas en regresión lineal:
  - https://www.youtube.com/watch?v=fTfMdCQJz4s
- Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data
  Publisher: SAGE Publications, Inc; 1 edition (January 10, 2012)
  Autor: Jason W. Osborne