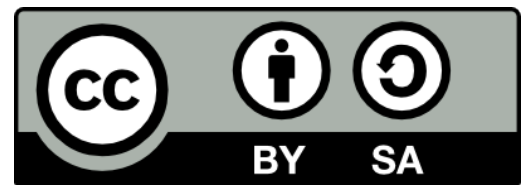
Big Data I:

Ingeniería de datos

Felipe Ortega
Dpto. de Estadística e Investigación Operativa
Universidad Rey Juan Carlos

March 24, 2015

Universidad
Rey Juan Carlos

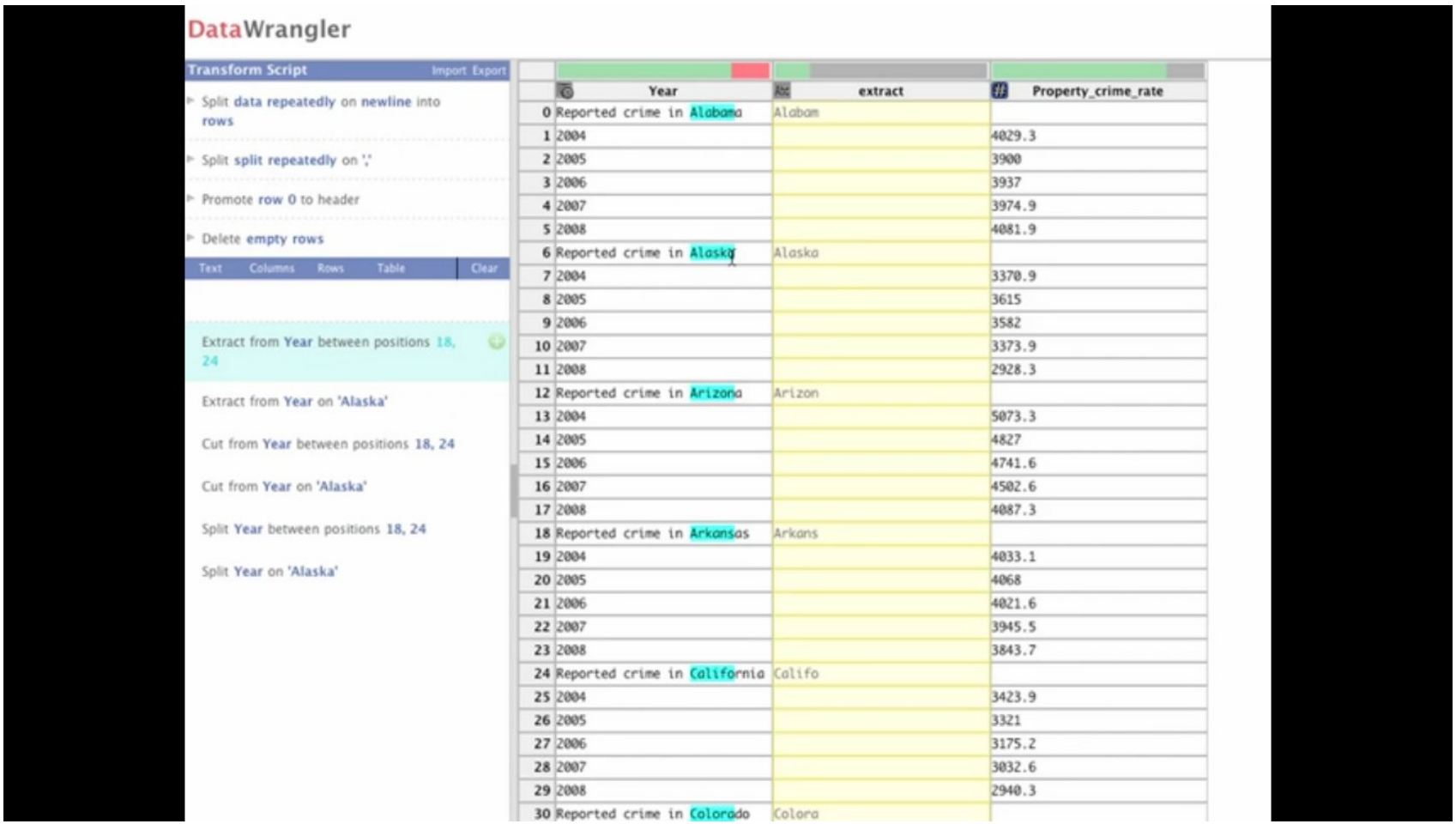


Preparación y transformación de datos

- Limpieza de datos.
- Datos no disponibles.
  - Gestión de valores vacíos.
  - Imputación de datos no disponibles.
- Transformación de datos.
  - *Data munging* o *data wrangling*.
  - Pasar los datos a otro formato o dejarlos preparados para luego poder analizarlos más fácilmente.

Data Wrangler

- Ejemplo de este tipo de herramientas (UW Interactive Data Lab).

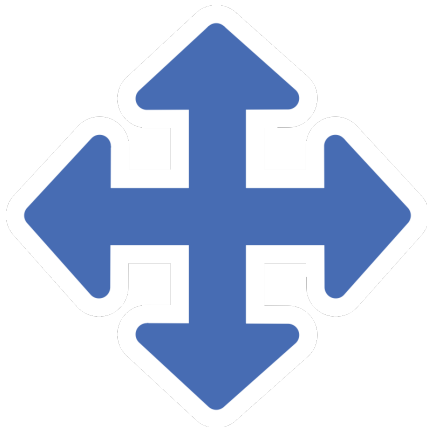


Preparación de datos

- En primer lugar, debemos comprobar que no existen valores extraños ni datos omitidos.
  - Utilizar técnicas básicas de resumen de datos.
  - Técnicas de visualización de datos omitidos.
- Después, tenemos dos opciones:
  - Descartar los casos que contengan variables con datos omitidos.
  - Imputar valores para los datos que faltan, utilizando técnicas avanzadas de imputación de múltiples valores.

Transformación de datos

- Otro paso crucial antes de comenzar nuestro análisis es comprobar la distribución de valores de los parámetros implicados.
  - Muchas técnicas y modelos asumen que los datos siguen una cierta distribución (e.g. Normal), pero puede no ser cierto.
  - De hecho, en la práctica nos encontramos muchas veces con distribuciones sesgadas (*skewed distributions*) o con diferentes apuntamientos (*kurtosis*).

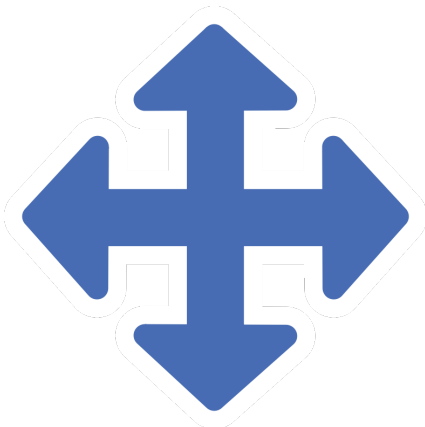


- Posibles objetivos:
  - Reducir la asimetría de la distribución de valores.
  - Transformar una o varias variables de forma que se parezcan más a una distribución Normal (univariante o multivariante).

Aspectos adicionales

- Tranformación entre diferentes formatos de datos
  - *Wide format* vx. *long format*.

```
# Wide format
subject sex control cond1 cond2
      1   M     7.9  12.3  10.7
      2   F     6.3  10.6  11.1
      3   F     9.5  13.1  13.8
      4   M    11.5  13.4  12.9
```



Aspectos adicionales

```
# Long format
subject sex condition measurement
       1   M   control         7.9
       1   M     cond1        12.3
       1   M     cond2        10.7
       2   F   control         6.3
       2   F     cond1        10.6
       2   F     cond2        11.1
       3   F   control         9.5
       3   F     cond1        13.1
       3   F     cond2        13.8
       4   M   control        11.5
       4   M     cond1        13.4
       4   M     cond2        12.9
```