In [1]:
```
# imports
import numpy as np
import pandas as pd
```

In [2]:
```
# define the header
columns = [
    'age', 'workclass', 'fnlwgt', 'education', 'education-num',
    'marital-status', 'occupation', 'relationship', 'race', 'sex',
    'capital-gain', 'capital-loss', 'hpw', 'country', 'classification'
]
```

In [3]:
```
# for labels
def label_converter(x):
    return 0 if x == ' <=50K' else 1
```

In [4]:
```
# load the raw data
# define non numeric columns as categories for one hot encoding
df = pd.read_csv(
    './adult.data',
    header=None,
    names=columns,
    dtype={
        'workclass': 'category',
        'education': 'category',
        'marital-status': 'category',
        'occupation': 'category',
        'relationship': 'category',
        'race': 'category',
        'sex': 'category',
        'country': 'category',
    },
    converters={
        'classification': label_converter
    }
)
```

In [5]:
```
# Construct np array from the dataframe
# get dummies changes the 'category' into a one hot encoding
arr = np.column_stack((
    df['age'].as_matrix(),
    pd.get_dummies(df['workclass']).as_matrix(),
    df['fnlwgt'].as_matrix(),
    pd.get_dummies(df['education']).as_matrix(),
    df['education-num'].as_matrix(),
    pd.get_dummies(df['marital-status']).as_matrix(),
    pd.get_dummies(df['occupation']).as_matrix(),
    pd.get_dummies(df['relationship']).as_matrix(),
    pd.get_dummies(df['race']).as_matrix(),
    pd.get_dummies(df['sex']).as_matrix(),
    df['capital-gain'].as_matrix(),
    df['capital-loss'].as_matrix(),
    df['hpw'].as_matrix(),
    pd.get_dummies(df['hpw']).as_matrix(),
    pd.get_dummies(df['country']).as_matrix(),
    df['classification'].as_matrix()
))
```

In [6]:
```
# save in npy format for use by classifiers
np.save('adult.npy', arr)
```