# Chapter 4 - Ex3: Combining Data

## Câu 1: Illinois Male Baby Names

### Cho dữ liệu 201*-baby-names-illinois.csv. Bộ dữ liệu này thống kê tần suất các tên gọi được đặt.

- Đọc dữ liệu
- Xem xét vấn đề về dữ liệu cần khắc phục
- Chuẩn lại dữ liệu để khắc phục vấn dề trên

In [1]:

```
import matplotlib.pyplot as plt
import pandas as pd
import glob
```

## Vấn đề cần khắc phục

- Dữ liệu được lưu trữ trên nhiều table/file.
- Biến “Year” được để trong tên tập tin

In [2]:

```
allFiles = glob.glob("201*-baby-names-illinois.csv")
frame = pd.DataFrame()
df_list= []
for file_ in allFiles:
    df = pd.read_csv(file_,index_col=None, header=0)
    print(file_, df.shape)
    df.columns = map(str.lower, df.columns)
    df["year"] = file_[0:4]
    df_list.append(df)
```

```
2014-baby-names-illinois.csv (101, 4)
2015-baby-names-illinois.csv (100, 4)
```

In [3]:

```
df4 = pd.concat(df_list)
df4.head(5)
```

Out[3]:

| | rank | name | frequency | sex | year |
|---|---|---|---|---|---|
| **0** | 1 | Noah | 837 | Male | 2014 |
| **1** | 2 | Alexander | 747 | Male | 2014 |
| **2** | 3 | William | 687 | Male | 2014 |
| **3** | 4 | Michael | 680 | Male | 2014 |
| **4** | 5 | Liam | 670 | Male | 2014 |

In [4]:

```
df4.shape
```

Out[4]:

```
(201, 5)
```