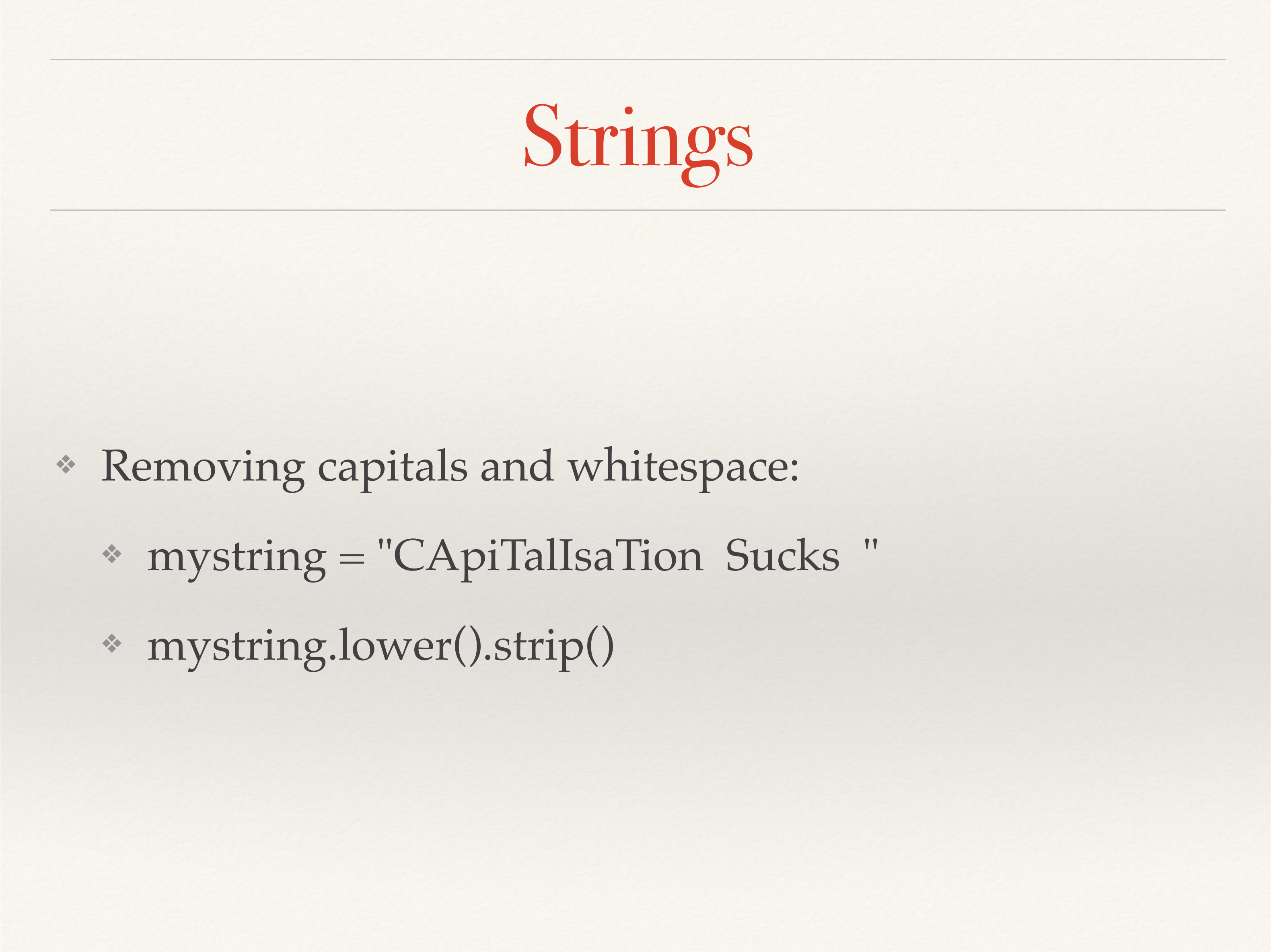
DataScience for Development and Social Change, 2015

Cleaning Data part ii

Getting data into usable formats

Data Cleaning

- The right format (e.g. CSV)
- The right shape (e.g. no ‘extra’ rows on top, one row per piece of data)
- Consistent labels
- Consistent “no data” values
- No junk symbols, whitespace etc.



Strings

- Removing capitals and whitespace:
  - mystring = "CApiTalIsaTion Sucks "
  - mystring.lower().strip()

Regular Expressions

- Really powerful library for cleaning up strings

```
import re

string1 = “This is a! sentence&&with junk!@“

cleanstring1 = re.sub(r'[^\w ]', '', s)

string2 = “comma , list ,  with , extra , spaces”

cleanstring2 = re.sub(r'\s+,\s+', ',', string2)
```

Date/Times

- European vs American? Name of month vs number? Python comes with a bunch of date reformatting libraries that can convert between these. For example:

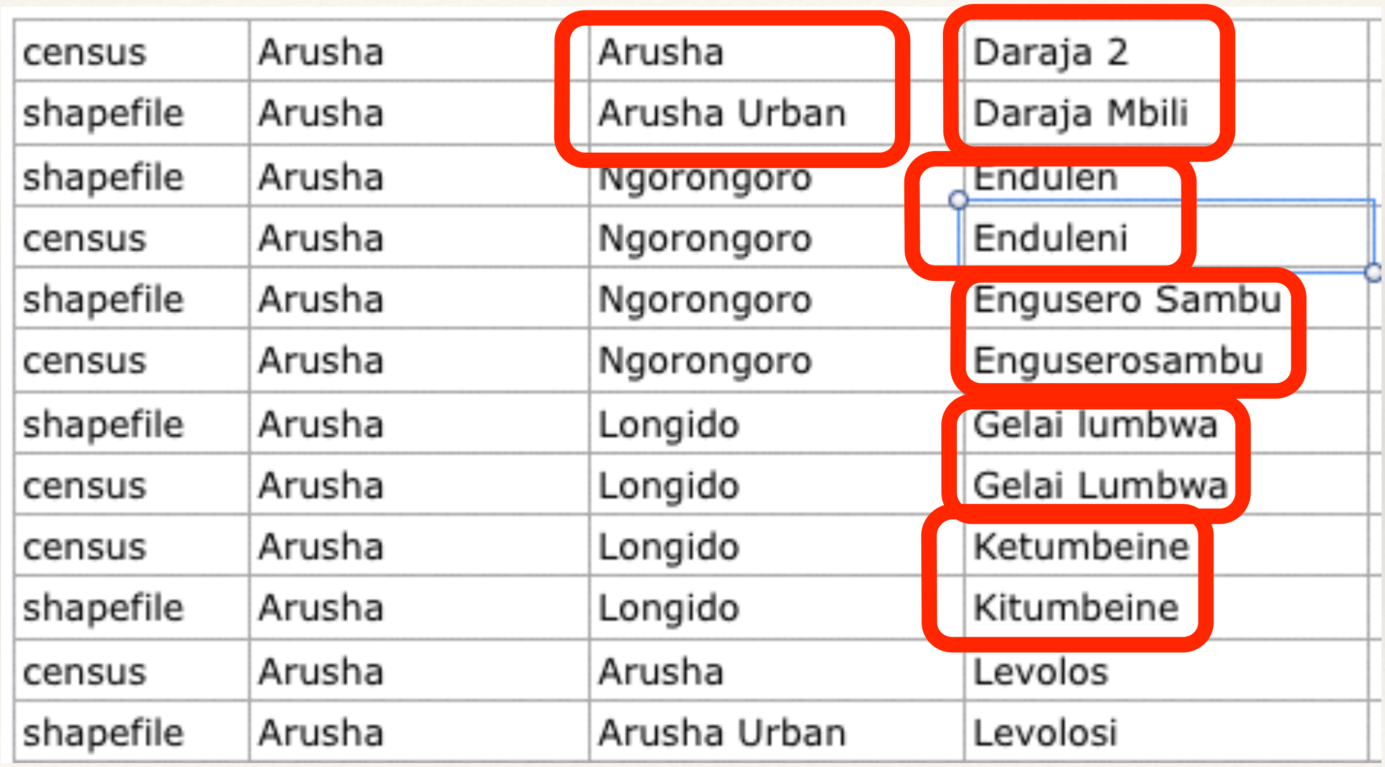
```
import datetime

date_string = “14/03/48"

datetime.datetime.strptime(date_string, ‘%m/%d/%y').strftime('%m/
%d/%Y')
```

- More about date formats at section 8.17 of https://docs.python.org/2/library/datetime.html

Merging Datasets



census	Arusha	Arusha	Daraja 2
shapefile	Arusha	Arusha Urban	Daraja Mbili
shapefile	Arusha	Ngorongoro	Endulen
census	Arusha	Ngorongoro	Enduleni
shapefile	Arusha	Ngorongoro	Engusero Sambu
census	Arusha	Ngorongoro	Enguserosambu
shapefile	Arusha	Longido	Gelai lumbwa
census	Arusha	Longido	Gelai Lumbwa
census	Arusha	Longido	Ketumbeine
shapefile	Arusha	Longido	Kitumbeine
census	Arusha	Arusha	Levolos
shapefile	Arusha	Arusha Urban	Levolosi

Use Standards

DR Congo in data.UN.org: “Congo, Democratic Republic of the”, “Congo Democratic”, “Democratic Republic of the Congo”, “Congo (Democratic Republic of the)”, “Congo, Dem. Rep.”, “Congo Dem. Rep.”, “Congo, Democratic Republic of”, “Dem. Rep. of Congo”, “Dem. Rep. of the Congo”

DR Congo in common standards: “Democratic Republic of the Congo” (UN Stats), “Congo, The Democratic Republic of the” (ISO3166), “Congo, Democratic Republic of the” (FIPS10, Stanag), “180” (UN Stats), “COD” (ISO3166, Stanag), “CG” (FIPS10)

Google Open Refine

- Data cleaning tool
- Watch the Explore Data video on http://openrefine.org!

More on Data Cleaning

- http://schoolofdata.org/handbook/courses/data-cleaning/