UNIVERSITY OF GLOUCESTERSHIRE

REGULAR EXPRESSIONS (REGEX)

- In a typical data analytics project, data usually come under different forms and formats.
  - Combination of numbers and characters (e.g., 2016-09-21, APPL; 16.52; 35.66; UP)
- The question becomes: how can we extract the data that we want in a quick and efficient way?
- We can use regular expressions to achieve this.
- Python provides this functionality through the ***re*** module.

UNIVERSITY OF GLOUCESTERSHIRE

REGULAR EXPRESSIONS (REGEX)

```
import re

strTarget = "There are 2 guitarists and 1 bassist in" +
 " a 4 member rock band."

iBandCount = re.findall(r'\d+', strTarget)

print ("Total number of band members: ", iBandCount[2])

print ("Number of guitarists: ", iBandCount[0])

print ("Number of bassists: ", iBandCount[1])
```

UNIVERSITY OF GLOUCESTERSHIRE

REGULAR EXPRESSIONS (REGEX)

Table: Some regex metacharacters and what they mean

Regex	What it means
*	looks for zero or more occurrences of a set of characters.
+	looks for one or more occurrences of a set of characters.
\d	Matches any number
\s	Matches any whitespace character
\w	Matches any alphanumeric character

NB: the capital letters of the regex (e.g., \S) represents the complementary of their small letter counterparts

UNIVERSITY OF GLOUCESTERSHIRE

REGULAR EXPRESSIONS (REGEX)

Given a string "aabaaa456baaa666abab71:22bccaaa123"

Table: Regex inputs and their outputs

Regex	What we will get
\d	'4', '5', '6', '6', '6', ...
\d+	'456', '666','71','22', '123'
\D+	'aabaaa', 'baaa', 'abab', ':', 'bccaaa'
[\d] + $	'123'
[a|b]{3} [\d]+	'aaa456', 'aaa666', 'bab71', 'aaa123'

UNIVERSITY OF GLOUCESTERSHIRE

REGULAR EXPRESSIONS (REGEX)

```
import re

strTargetString = "aabaaa456baaa666abab71:22bccaaa123"

print "Target string", strTargetString

strRegEx = input('Enter regular expression: ')

print(re.findall(strRegEx, strTargetString))
```