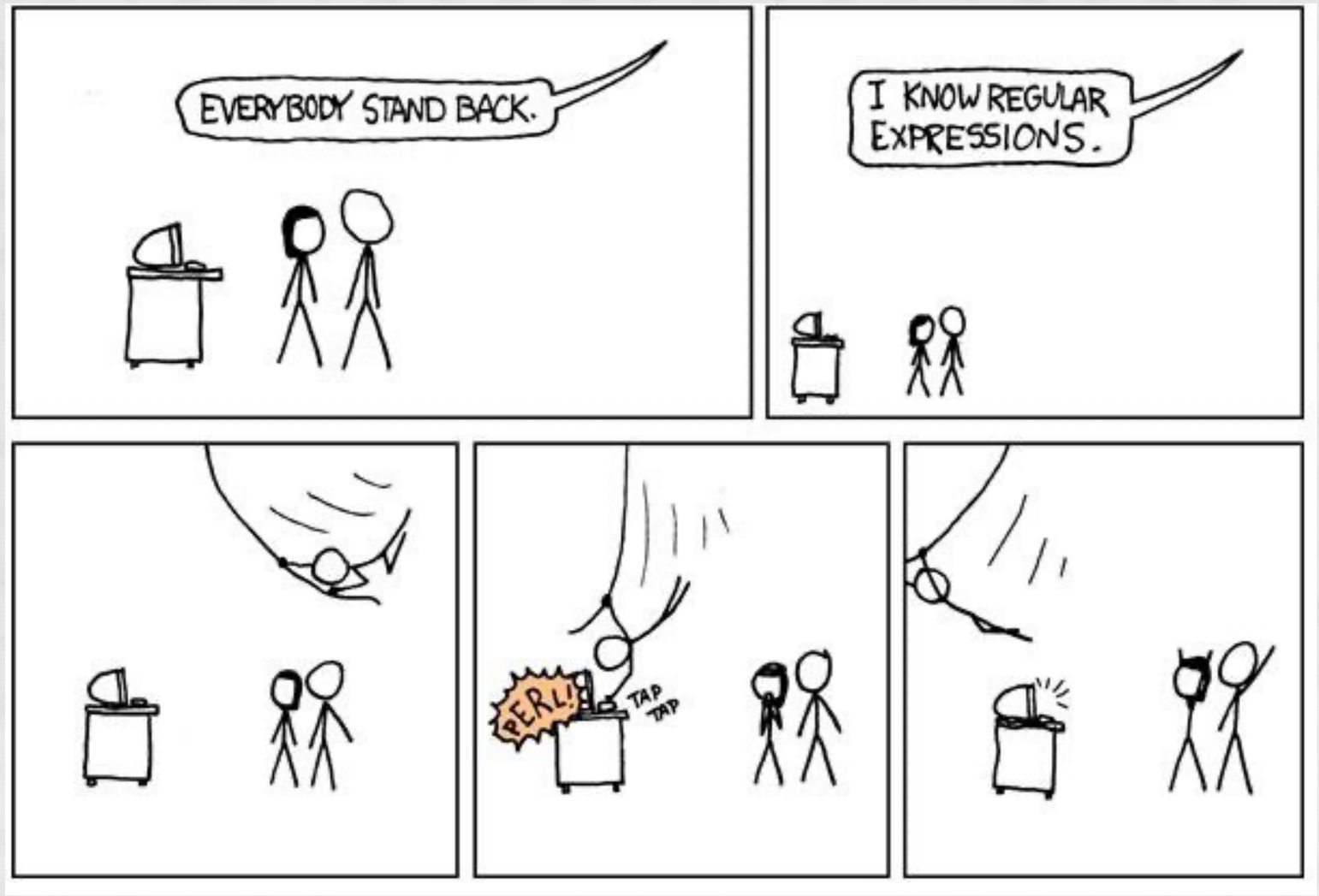


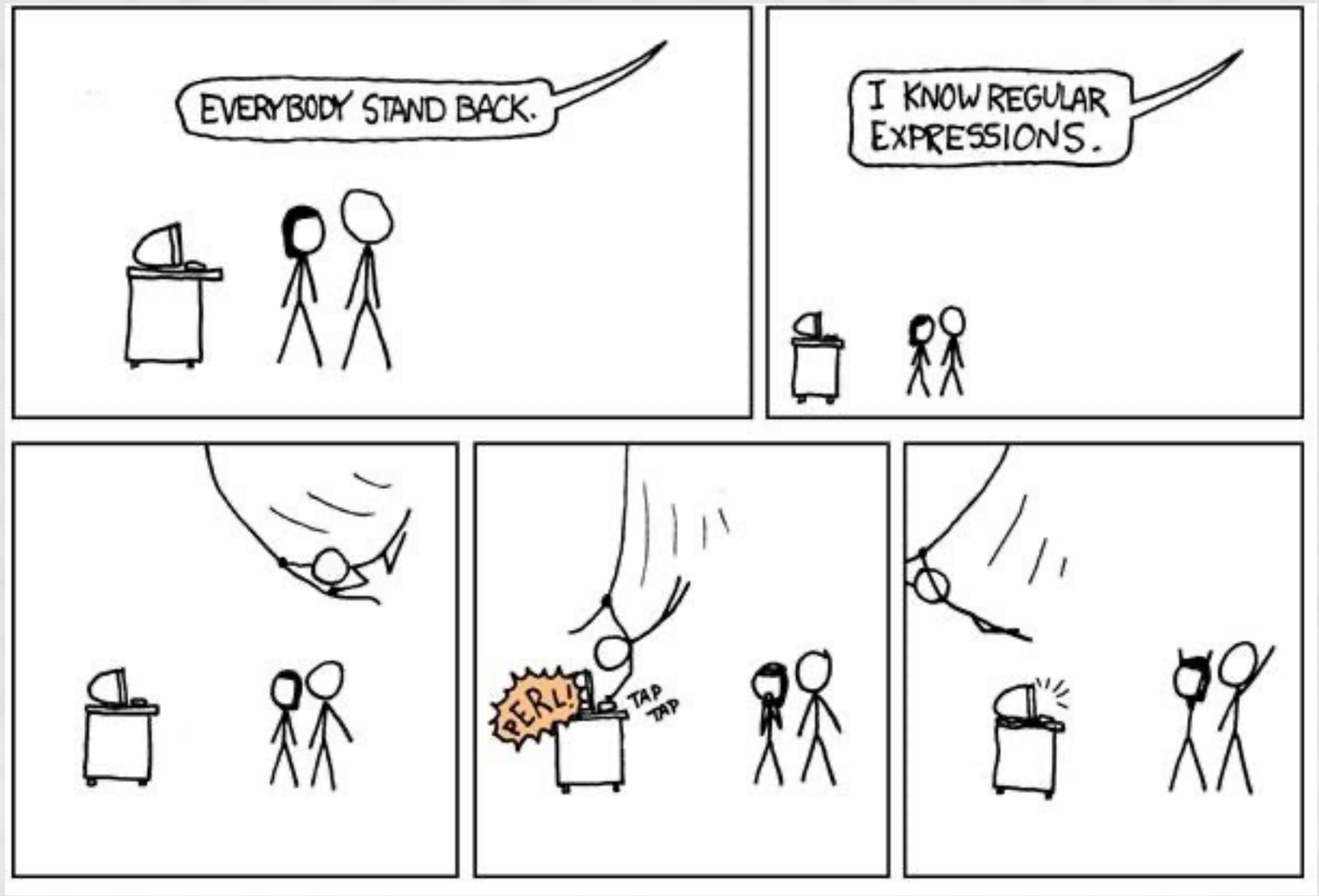
Regular Expressions - Regex

DSA8022 Frontiers in Analytics

Dr Gary McKeown

School of Psychology, David Keir Building

0G3.536



Regular Expressions - Regex

DSA8022 Frontiers in Analytics

Dr Gary McKeown

School of Psychology, David Keir Building

0G3.536

TEXT PROCESSING

^[Reg]ular[Ex]pressions$

REGULAR EXPRESSIONS

- Regular expressions are a bit of dark art
- Allow you to define search patterns for text
- Very powerful, can be difficult to interpret
- Very flexible and also precise, but comes with health warnings.
- Many formulations
  - PERL - very powerful, difficult to interpret
  - UNIX/POSIX
    - grep - **g**lobally search for a **r**egular **e**xpression and **p**rint matching lines
  - In R stringr from Tidyverse, in Python re.py
  - All are variations on a theme

ESCAPE CHARACTERS

- Computer languages reserve certain characters for their compilers to understand.
- Therefore we cannot use them normally
- For example, " and ' are reserved to let a computer know we are about to start a string of letters or numbers
  - "This is a string"
- To use them inside a strong we have to use the escape letter \
  - "He said \”this is a string\" to the class”
- We can also do the same for the \ command as it too is a reserved character.
  - “If you want to put the backslash in a string \\ do this”

RESERVED CHARACTERS

- \! - !
- \? - ?
- \\ - \
- \(- (
- \) -)
- \{ - {
- \} - }
- \| - |
- \^ - ^
- Within stringr in R you need to add an extra escape character \
- See the cheatsheet: https://github.com/rstudio/cheatsheets/blob/master/strings.pdf

SPECIAL CHARACTERS

- \n - newline
- \t - tab
- \s - any whitespace
- \S - any non-whitespace
- \d - any digit
- \D - any non-digit
- \w - any word or character string
- \W - any non-word or non-character string

REGEX SPECIAL CHARACTERS

- . - the dot, matches any character except the newline
- ^ - ties the pattern match to the start of a line (if it appears at the start)
- $ - ties the pattern match to the end of a line
- [] - defines a set of characters to match
- [^abc] - caret inside square brackets means except
- | - pipe, or matches either or characters on either side
- **Repetition Qualifiers**
- * - asterisk, placed after a pattern it matches 0 or more repetitions
- + - plus, placed after a pattern it matches 1 or more repetitions
- ? - will match 0 or 1 of the preceding patterns

RESERVED CHARACTERS

- Lots of online tutorials
- https://regexone.com
  - Capturing
    - We can use parentheses to retain an element of a pattern
- https://www.rexegg.com/regex-quickstart.html
- https://regex101.com

DIFFERENT FLAVOURS

- POSIX - UNIX
- PCRE2 (PHP)
- PCRE (PHP)
- PERL
- ECMAScript (Javascript)
- Python
- GoLang
- Java 8