# DATA 607
# **Meetup 3**

Andy Catlin

CUNY School of Professional Studies

CUNY

# tonight's class

- Georgia Galanopoulos's *Data Science in Context* presentation
- Regular Expressions
- Review of week 2 assignments

# Regular expressions

Some people, when confronted with a problem, think "I know, I'll use regular expressions." Now they have two problems.

Source: Jamie Zawinski

# Regular Expressions

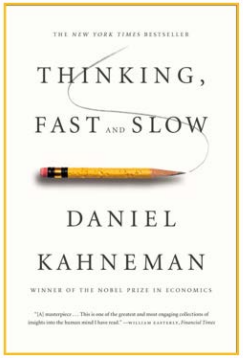

Title text: Wait, forgot to escape a space. Wheeeeee[taptaptap]eeeeee.

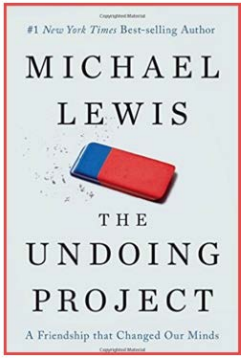
Source: https://xkcd.com/208/

# gaining proficiency with regex

- Understanding regex concepts
  - finding regex matches
  - Replacing regex matches
- Understanding R language-specific implementations of regex concepts
  - R supports two regex flavors:  ~POSIX and PERL (PCRE)
  - base R or stringr package regex
- Learning and applying recurring regex patterns

# availability heuristic




In Tversky & Kahneman's first examination of availability heuristics, subjects were asked, "If a random word is taken from an English text, is it more likely that the word starts with a K, or that K is the third letter?" They argue that English-speaking people would immediately think of many words that begin with the letter "K" (kangaroo, kitchen, kale), but that it would take a more concentrated effort to think of any words in which "K" is the third letter (acknowledge, ask). Results indicated that participants overestimated the number of words that began with the letter "K" and underestimated the number of words that had "K" as the third letter. Tversky and Kahneman concluded that people answer questions like these by comparing the availability of the two categories and assessing how easily they can recall these instances. In other words, it is easier to think of words that begin with "K", more than words with "K" as the third letter. Thus, people judge words beginning with a "K" to be a more common occurrence.

Source: https://en.wikipedia.org/wiki/Availability_heuristic

# Are fruit words more fun?

Some fruit names have **repeated pairs of letters**:

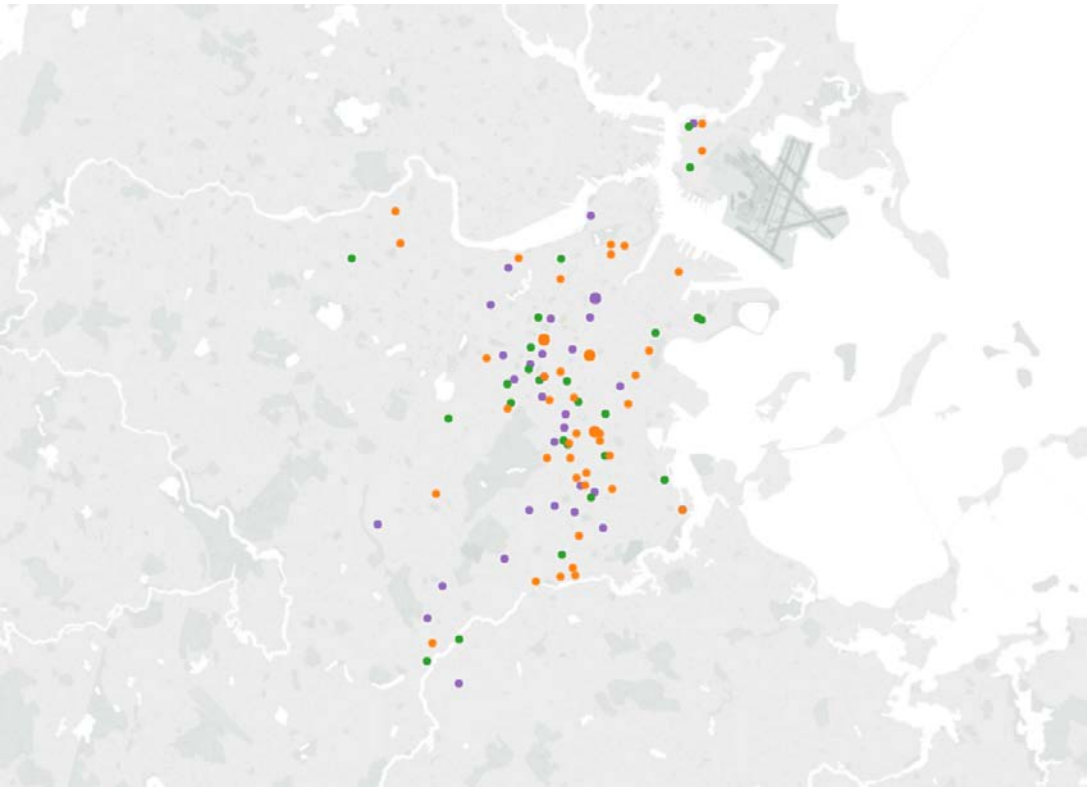
- coconut, papaya

Are fruit words more likely to have repeated pairs of letters than common words overall in English?

# Boston crime

Boston, MA has made available a list of crime incident reports (45 MB, 268K):
https://data.cityofboston.gov/Public-Safety/Crime-Incident-Reports-July-2012-August-2015-Sourc/7cdf-6fgx

- How would you approach the problem of populating an R data frame that contains the information needed to map the crimes for a range of dates?



| FROMDATE | Location |
|---|---|
| 7/8/2012 6:00 | (42.34638135, -71.10379454) |
| 7/8/2012 6:03 | (42.31684135, -71.07458456) |
| 7/8/2012 6:26 | (42.34284135, -71.09698955) |
| 7/8/2012 6:56 | (42.3164411, -71.06582908) |
| 7/8/2012 7:15 | (42.27051636, -71.11989955) |
| 7/8/2012 7:32 | (42.31328183, -71.0530059) |
| 7/8/2012 7:50 | (42.32425136, -71.08620956) |
| 7/8/2012 7:50 | (42.34924634, -71.06378456) |
| 7/8/2012 7:53 | (42.35174635, -71.16590953) |
| 7/8/2012 8:05 | (42.25938275, -71.11729354) |
| 7/8/2012 8:10 | (42.34180635, -71.09707955) |
| 7/8/2012 8:15 | (42.3092417, -71.05033304) |
| 7/8/2012 9:00 | (42.34924634, -71.06378456) |
| 7/8/2012 9:30 | (42.33490135, -71.07516956) |
| 7/8/2012 9:30 | (42.35973634, -71.06796956) |
| 7/8/2012 9:45 | (42.34456135, -71.13768454) |
| 7/8/2012 9:47 | (42.36179134, -71.05277456) |

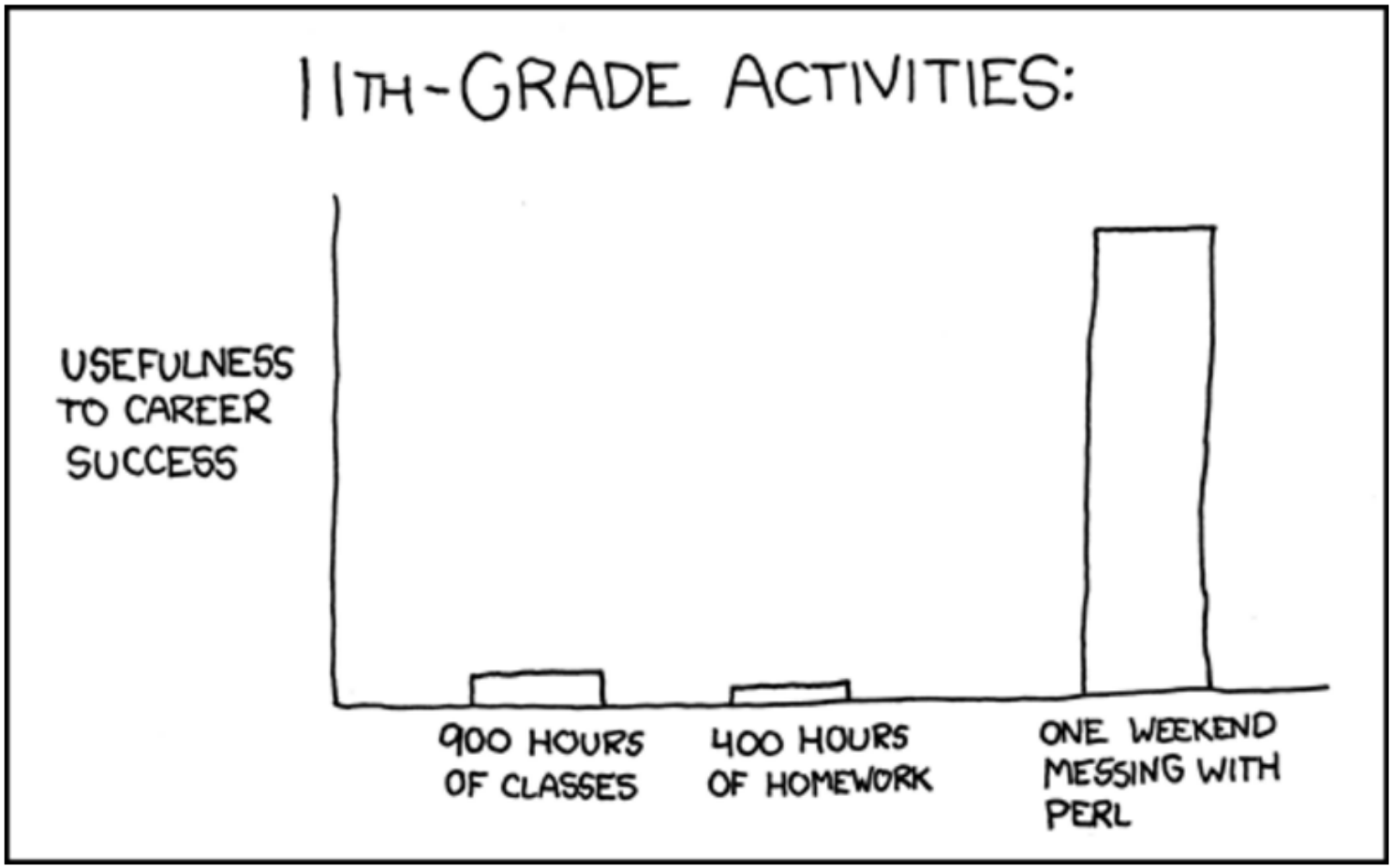
- What if you wanted to extend this analysis of Boston crime data to also include the type of crime committed?

| NatureCode | INCIDENT_TYPE_DESCRIPTION | MAIN_CRIMECODE | FROMDATE | Location |
|---|---|---|---|---|
| BERPTA | RESIDENTIAL BURGLARY | 05RB | 7/8/2012 6:00 | (42.34638135, -71.10379454) |
| PSHOT | AGGRAVATED ASSAULT | 04xx | 7/8/2012 6:03 | (42.31684135, -71.07458456) |
| ARMROB | ROBBERY | 03xx | 7/8/2012 6:26 | (42.34284135, -71.09698955) |
| ALARMC | COMMERCIAL BURGLARY | 05CB | 7/8/2012 6:56 | (42.3164411, -71.06582908) |
| ARMROB | ROBBERY | 03xx | 7/8/2012 7:15 | (42.27051636, -71.11989955) |
| SHOT | ROBBERY | 03xx | 7/8/2012 7:32 | (42.31328183, -71.0530059) |
| ARMROB | ROBBERY | 03xx | 7/8/2012 7:50 | (42.32425136, -71.08620956) |
| THREAT | SIMPLE ASSAULT | 08xx | 7/8/2012 7:50 | (42.34924634, -71.06378456) |
| REQP | MedAssist | MedAssist | 7/8/2012 7:53 | (42.35174635, -71.16590953) |
| ALARMI | MedAssist | MedAssist | 7/8/2012 8:05 | (42.25938275, -71.11729354) |
| BEIP | BENoProp | BENoProp | 7/8/2012 8:10 | (42.34180635, -71.09707955) |
| IVMV | VAL | VAL | 7/8/2012 8:15 | (42.3092417, -71.05033304) |
| LARCRT | FRAUD | 11xx | 7/8/2012 9:00 | (42.34924634, -71.06378456) |
| FDPROP | PropFound | PropFound | 7/8/2012 9:30 | (42.33490135, -71.07516956) |
| FDPROP | InvPer | InvPer | 7/8/2012 9:30 | (42.35973634, -71.06796956) |
| IVMV | TOWED | TOWED | 7/8/2012 9:45 | (42.34456135, -71.13768454) |
| ILLPRK | VAL | VAL | 7/8/2012 9:47 | (42.36179134, -71.05277456) |

- What would be the advantage of categorizing crimes with Uniform Crime Reporting (UCR) stats? Could we transform the provided data into UCR codes, and if so, how might we go about this?
- See also: https://fivethirtyeight.com/features/which-cities-share-the-most-crime-data/

# regex resources

- stringr
  - R for Data Science, Chapter 14, http://r4ds.had.co.nz/strings.html. See also http://stringr.tidyverse.org/ and Hadley Wickham, "stringr: modern, consistent string processing," *The R Journal*, Dec 2010. https://journal.r-project.org/archive/2010-2/RJournal_2010-2_Wickham.pdf and "Regular Expressions in R," http://stat545.com/block022_regular-expression.html
- Base R
  - "Regular Expressions with The R Language," http://www.regular-expressions.info/rlanguage.html base R methods
- General regex
  - Dan Bickford , "5 Tools You Didn't Know That Use Regular Expressions," https://www.codeschool.com/blog/2015/07/30/5-tools-you-didnt-know-that-use-regular-expressions/. Jul 30, 2015.
  - Regex for twitter emoticon tokens, http://sentiment.christopherpotts.net/tokenizing.html#emoticons
  - Google training material on Regular Expressions in Python, https://developers.google.com/edu/python/regular-expressions




Source: https://xkcd.com/519/

# review of week 2 assignments