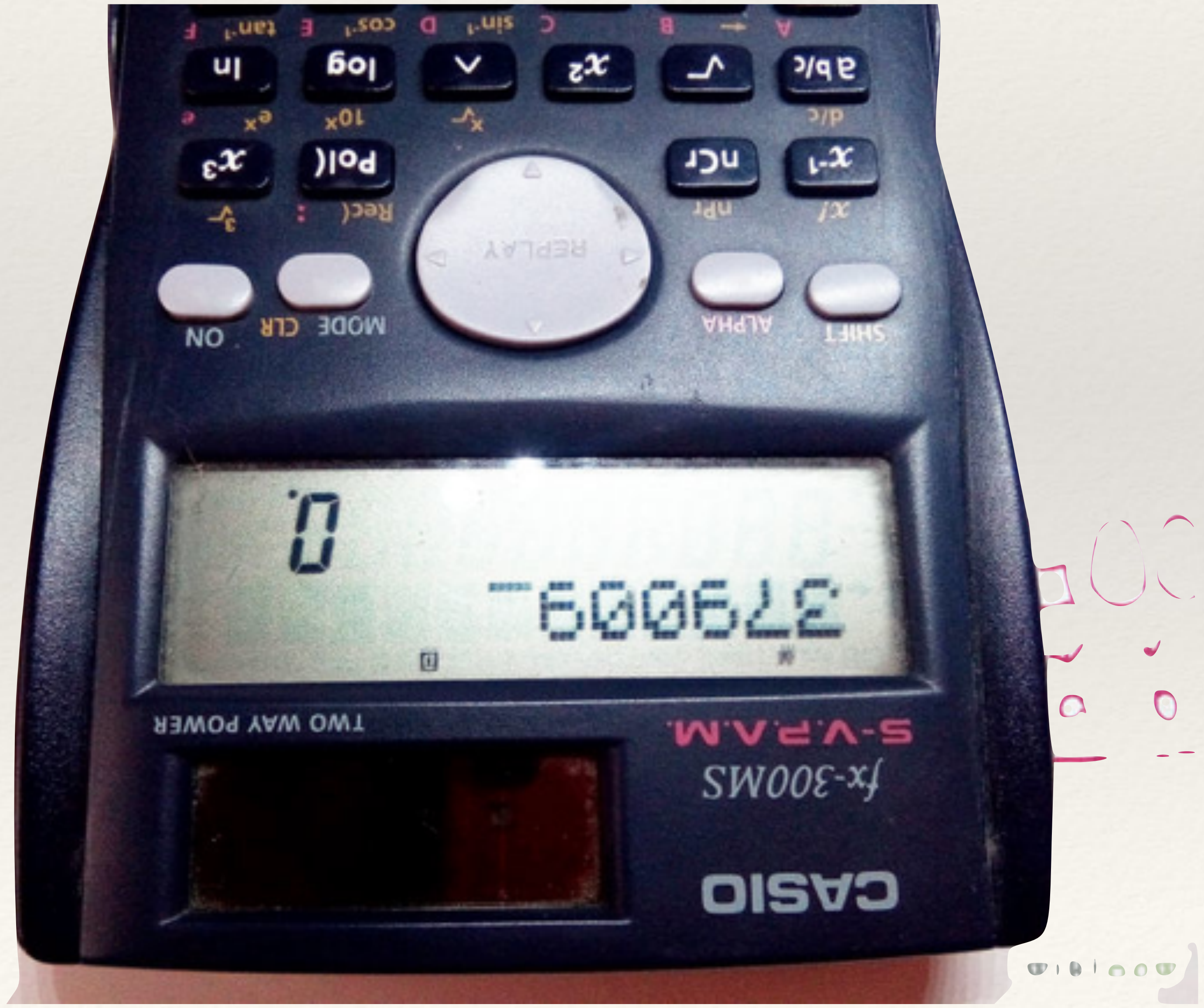
Advanced Programming Techniques

Regular Expressions

Christopher Moretti

bEghILOSZ

bEghILOSZ

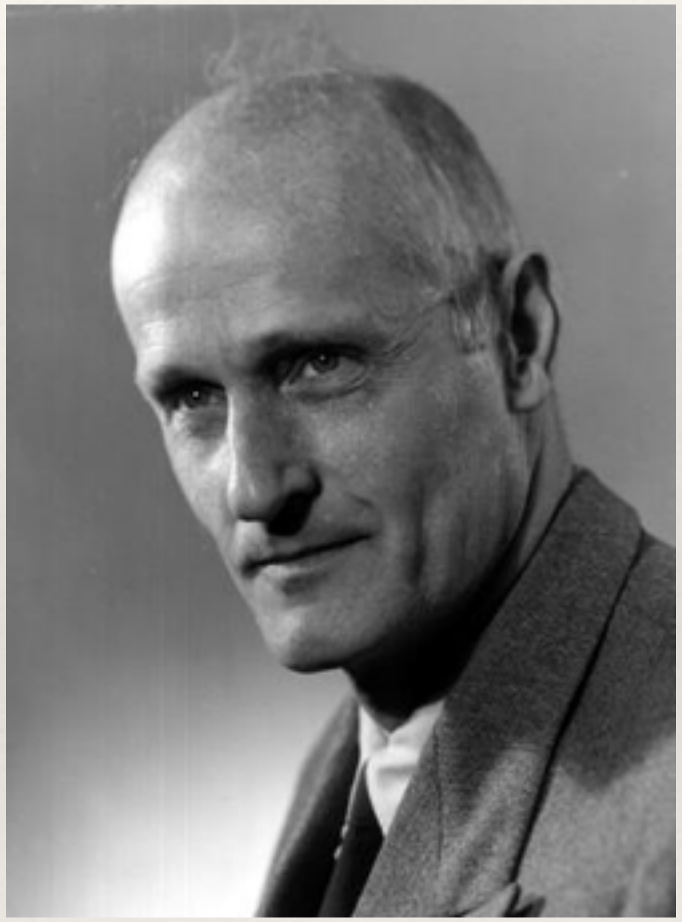


“It's not a silly question if you can't answer it.”

–Jostein Gaarder, Sophie’s World

Sounds like 1134? Not in this font.

- Based on ideas from automata theory in regular languages pioneered by Stephen Kleene *34
- Practically applied in late-60's in various settings (compilers, editors) and started to enter into languages in the 70's (e.g., AWK)
- Mechanized text filtering and pattern match
  - Pervasive in *nix tools (e.g. sed, grep)
  - Key built-in for scripting syntax (pattern-action languages)
  - Available in just about any other language as a library



grep Regular Expressions

`c`	Any character matches itself, except for meta-characters: `. [ ] ^ $ * \`
`rs`	Matches r followed by s
`.`	Matches any single character
`[rst]`	Matches one of r, s, or t — range shorthands are allowed: [0-9a-z]
`[^rst]`	Matches one character other than r, s, or t
`^`	Matches start of the line when at start of pattern (not special otherwise)
`$`	Matches end of the line when at end of pattern (not special otherwise)
`x*`	Kleene closure: 0 or more repetitions of x
`\c`	Matches c unless c is ( ) or a digit — \ is the escape character
`\(x\)`	Matches the pattern x and saves the substring that matches (x′) as \#
`x{m,n}`	Matches x repeated between m and n times

Reasonable Toy Examples

xy	xy anywhere in string
^xy	xy at beginning of string
xy$	xy at end of string
^xy$	string that contains only xy
^	matches any string, even empty
^$	empty string
.	non-empty, i.e., at least 1 char
xy.$	xy plus any char at end of string
xy\.$	xy. at end of string
\\xy\\	\xy\ anywhere in string
[xX]y	xy or Xy anywhere in string
^xy[0-9][^0-9]$	xy followed by digit, then non-digit
xy1.*xy2	xy1 then any text then xy2
^xy1.*xy2$	xy1 at beginning and xy2 at end

RE-Building our Vocabulary

1. Contains only the letters beghilosz
2. Can use the letters in any order
3. Can use any letter multiple times
4. Assume no length restriction

```
^[beghilosz]*$
```

grep in Action

```
tars: wc -l /usr/share/dict/words
  235886 /usr/share/dict/words

tars: grep '^[beghilosz]*$' words | wc -l
     451


opus: wc -l /usr/share/dict/words
479828 /usr/share/dict/words

opus: grep '^[beghilosz]*$' words | wc -l
975
```

Note the single-quotes around the REs —
we don’t want the shell to take matters into its own hands

```
opus: grep ‘^[beghilosz]*$’ words | column –c 220
b               biggies                 bogglebo        egises                  ghis            gobies                  hello           hoo             leo             logoes          oilish          shi             slobs
bb              biggish                 boggles                 ego             ghole           gobo            helloes                 hool            leos            logoi           oilless                 shiel           sloe
bbl             bilbi           bogglish        egoize                  gi              goboes                  hellos          hoolee          les             logos           oils            shiels          sloes
bbls            bilbie                  boghole                 egol            gib             gobos                   hells           hoolie          lese            logs            ol              shies           slog
bbs             bilbies                 bogie           egoless                 gibbles                 gobs            heloe           hoose           less            loli            ole             shih            slogs
be              bilbo           bogies                  egos            gibbol                  goebbels        helosis                 hoosh           lessee          loll            oleo            shill           sloo
bebless                 bilboes                 bogle           eh              gibbose                 goel            helzel          hose            lessees                 lollies                 oleos           shilloo                 sloosh
bebog           bilbos                  bogles                  eigh            gibe            goes            heo             hosel           lesses          lolls           oleose          shills          slosh
beboss                  bile            bogo            eisegeses       gibel           gog             hes             hoseless        lez             loo             oles            shish           sloshes
bee             biles           bogs            eisegesis       gibes           goggle                  hg              hosels          lezes           loob            olio            sho             so
beebee                  bilge           boh             eisell                  gibleh                  goggles                 hi              hoses           lezzie          loobies                 olios           shoe            sob
beebees                 bilges                  boho            el              gibli           gogo            hie             hoss            lezzies                 looie           ologies                 shoebill        sobole
beeish                  bilio           boil            elb             gibs            gogos           hies            hs              lg              looies          ooh             shoebills       soboles
beele           bill            boils                   elegies         gie             goi             higgle          i               lh              loos            oohs            shoeless        sobs
beelol                  billie                  boise           elegise         gies            gois                    higgles                 ib              lhb             loose           oollies                 shoes           soe
bees            billies                 bol             elegises        giesel                  gol             high            ibis            li              looses          oologies        shog            sog
beg             bills           bole            elegize                 gig             golee           highhole        ibisbill        lib             loosish                 oologize        shoggie                 soh
begiggle        bilo            boles           elegizes        gigge           goles           highish                 ibises          libel           lose            oooo            shoggle                 soho
begloze                 bilobe                  bolis           elhi            giggish                 goli            highs           ie              libelee                 losel           oos             shogi           soil
bego            bilos           boll            eli             giggle          goll                    hile            ieee            libelees        losels          oose            shogs           soilless
begob           bilsh           bollies                 eligible        giggles                 goloe                   hili            igloo           libellee        loses           ooze            shole           soils
begobs                  bio             bolls           eligibles       gigglish        golosh                  hill            igloos          libellees       losh            oozes           shoo            sol
begs            biog            bolo            ell             gighe           goloshe                 hillbillies     ihi             libels          loss            os              shoogle                 sole
beige           biol            bolos           elle            giglio                  goloshes        hillo           ihs             libs            losses          ose             shooi           solei
beigel                  biologese       bolshie                 ells            gigolo                  goo             hillos          ii              lie             lossless        oselle          shool           soleil
beiges                  biologies       bolshies        eloge           gigolos                 goog            hills           iii             liebig          ls              oses            shools          soleless
bel             biologize               boo     els             gigs            googlies        his             il              liege           o               osi             shoos           soles
belee           bios            boob            else            gile            googol                  hish            ile             liegeless       ob              oslo            si              solgel
belibel                 biose           boobie                  elses           giles           googols                 hisis           ill             lieges          obb             osse            sib             soli
belie           biosis                  boobies                 eo              gilgie                  gool            hiss            illegible       lies            obe             oz              sibb            solio
belies                  bis             boobish                 eole            gill            gools                   hissel          illess          liesh           obeish          ozs             sibbs           solo
bell            bise            booboisie       eos             gillie                  goos            hisses          illish          lig             obeli           s               sibs            solos
belle           bises           booboo                  es              gillies                 goose                   hizz            ills            lige            obelise                 sb              sie             sols
belles                  bish            booboos                 ese             gills           goosebill       hizzie          io              ligge           obelises        se              siege           soogee
belli           bisie           boobs           eses            gilo            goosegog        hl              ios             lile            obelize                 see             sieges          sool
bellies                 bisso           boogie                  ess             gils            gooses                  ho              is              liles           obelizes        seege           sig             sooloos
bello           biz             boogies                 esse            gilse           goosish                 hob             ise             lilies          obes            seel            sigh            soosoo
bells           bize            boohoo                  essee           gio             goozle                  hobbies                 ish             lill            obese           seels           sighless        sos
bels            bizel           boohoos                 esses           gis             gos             hobbil          isis            lis             obi             sees            sighs           sosh
bes             bizes           bool            g               gise            gosh            hobble                  isize           lish            obis            seesee          sigil           sosie
besee           bizz            boolies                 ge              gisel           goss            hobbles                 isl             lisle           obl             seg             sigill          soso
beseige                 bl              boos            geb             gish            gozell                  hoblob          isle            lisles          obli            seggio          sigils          sosoish
beshell                 bleb            boose           gebbie                  gizz            gozill                  hobo            isleless        liss            oblige          seghol          sigloi          soss
besiege                 blebs           boosies                 gee             gl              gs              hoboe           isles           lisses          obligee                 sego            siglos          sossiego
besieges        blee            booze           gees            glb             h               hoboes                  isls            lizzie          obligees        segol           sil             sossle
besigh                  bleeze          boozes                          geese           glebe           hb              hobos           iso             ll              obliges                 segos           sile            sozzle
besoil                  bleo            bos             geez            glebeless       he              hobs            isogloss        llb             oboe            segs            sill            ss
bess            bless           bose            gegg            glebes                  hebe            hoe             isoglosses      lo              oboes           sei             sillibib        z
bessel                  blesse                  bosh            geggee                  glee            hebes           hoes            isohel          lob             obol            seige           sillibibs       zee
besses                  blesses                 boshes                  gel             glees           hee             hog             isohels                 lobbies                 obole           seis            sillies                 zees
bezel           blibe           boss            gelee           gleg            heel            hogg            isolog          lobbish                 oboles          seise           sills           zeiss
bezels                  bliss           bosses                  gelees                  glib            heelless        hoggee          isologs                 lobe            oboli           seises          silo            zel
bezil           blisses                 bossies                 gell            gliosis                 heels           hoggie          issei           lobeless        obolos          seize           silos           zho
bezils          blissless               bozo            gelose                  gliss           heeze                   hoggish                 isseis          lobes           obols           seizes          siol            zig
bezzi           blizz           bozos           gelosie                 glob            heezes                  hoggs           isz             lobi            obs             sel             sis             zigs
bezzle                  blo             bozze           gels            globe           heezie                  hogo            izle            loblollies      obsess          sele            sise            zill
bezzo           blob            bs              geo             globes                  heh             hogs            l               lobo            obsesses        sell            sisel           zills
bg              blobs           bsh             geobios                 globose                 hehs            hoho            lb              lobolo          oe              selle           sises           zizel
bi              blooie                  bz              geog            globs           hei             hoi             lbs             lobolos                 oes             selles          sish            zizz
bib             bls             e               geol            glogg           heigh           hoise                   le              lobos           oesogi          sellie          sisi            zizzle
bibb            bo              ebb             geologies       gloggs                  heii            hoises          lebes           lobose          og              sells           siss            zizzles
bibble                  bob             ebbs            geologise       glos            heil            hol             lee             lobs            ogee            sels            sissies                 zo
bibbs           bobbie                  eblis           geologize       gloss           heils           hole            lees            loe             ogees           sess            sissoo          zobo
bibi            bobbies                 eboe            gess            glosses                 heishi                  holeless        leese           loeil           ogle            sessile                 size            zogo
bibl            bobbish                 ee              gesso           glossies        heize           holes           leg             loess           ogles           sg              sizes           zoll
bible           bobble                  eel             gessoes         glossless               hel             holi            lege            loesses                 oh              sh              sizz            zolle
bibles                  bobbles                 eelbob                  gez             glossologies            helbeh          holies          leges           log             ohelo           she             sizzle          zoo
bibless                 bobo            eels            ghee            gloze           hele            holl            legge           loge            ohio            shee            sizzles         zoogeog
bibliog                 bobol           eg              ghees           glozes                  helio                   hollies                 legible                 loges           oho             sheel           sl              zool
bibliologies            bobs            egg             ghess           go              helios                  hollo           legis           loggie          ohs             shell           slee            zoologies
biblos                  bobsleigh       eggless                 ghi             gob             heliosis        holloes                 legless                 loggish                 oie             shells          sleigh          zoologize
bibs            boe             eggs            ghibli                  gobbe           hell            holloo          legs            logie           oii             sheol           sleighs                 zoos
big             bog             eggshell        ghiblis                 gobble                  hellhole        holloos                 lei             logis           oil             sheols          slish           zoosis
bigg            boggish                 eggshells       ghillie                 gobbles                 hellholes       hollos          leis            loglog          oilhole                 shes            slob            zoozoo
biggie                  boggle                  egis            ghillies        gobi            hellish                 hols            leiss           logo            oilholes        shh             slobbish        zs
```

Back-references

`\(rst\)` Matches rst and saves that matching string as \1 to be used later in pattern

`\(abc\)\(rst\)\1\(xyz\)\3\2\1`

```
echo abcrstabcxyzxyzrstabc | grep '\(abc\)\(rst\)\1\(xyz\)\3\2\1'
abcrstabcxyzxyzrstabc
```

`\(ious\)\1`

```
grep '\(ious\)\1' words
homoiousious
```

`\(ious\).*\1`

```
grep '\(ious\).*\1' words
homoiousious
```

`\(.\).*\1.*\1.*\1.*\1.*\1.*\1.*\1`

```
grep '\(.\).*\1.*\1.*\1.*\1.*\1.*\1.*\1' words
bras-dessus-bras-dessous
humuhumunukunukuapuaa
pneumonoultramicroscopicsilicovolcanoconiosis
possessionlessness
```

Nested Back-references

```
\(r\(.\)\2\).*\1
```

So we have an r, then the same character twice,
then any string,
then the r and **same** same character twice

```
grep '\(r\(.\)\2\).*\1' words
broomroot
free-reed
greegree
greegrees
Greentree
groo-groo
proof-proof
proofroom
reel-to-reel
six-three-three
three-reel
tree-creeper
```

POSIX Extended Standard

`(x)`	Grouping
`x+`	Matches x repeated 1 or more times
`x?`	Matches x 0 or 1 times
`a|b`	Matches a OR b
`[:alnum:]`	Alphanumeric characters: [A-Za-z0-9]
`[:alpha:]`	Alphabetic characters: [A-Za-z]
`[:blank:]`	Space and Tab

`[:digit:] [:lower:] [:upper:] [:xdigit:]`

egrep

egrep (equivalent to ‘grep -E’) is “extended” grep

Some changes/non-standards/caveats:

```
echo foo | egrep '[:digit:]'
grep: character class syntax is [[:space:]], not [:space:]
echo foo | egrep '[[:digit:]]'
echo f00 | egrep '[[:digit:]]'
f00

echo tartar | grep ‘(a..a)*'
echo 't(arta)r' | grep '(a..a)*'
t(arta)r
echo tartar | egrep '(a..a)*'
tartar

echo tartar | grep '\(t..\)\1'
tartar
echo tartar | egrep '\(t..\)\1'
grep: Invalid back reference
echo tartar | egrep '(t..)\1'
tartar
```

Party Tricks

Words with all the vowels in order?

```
grep 'a.*e.*i.*o.*u.*y' words
abstemiously
adventitiously
anticensoriously
anticeremoniously
antireligiously
auteciously
autoeciously
facetiously
half-ingeniously
half-rebelliously
half-seriously
nonabstemiously
nonadventitiously
nonfacetiously
nonsacrilegiously
overabstemiously
pancreaticoduodenostomy
pareciously
paroeciously
pseudosacrilegiously
quasi-conscientiously
quasi-enviously
quasi-mysteriously
quasi-rebelliously
quasi-religiously
quasi-seriously
sacrilegiously
unabstemiously
unfacetiously
unsacrilegiously
```

No, what I really meant was …

```
grep 'a[^aeiouy]*e[^aeiouy]*i[^aeiouy]*o[^aeiouy]*u[^aeiouy]*y' words
     abstemiously
      facetiously
   half-seriously
  nonabstemiously
   nonfacetiously
 overabstemiously
      pareciously
   unabstemiously
    unfacetiously
```

No, what I really, *really* meant was …

```
grep '^[^aeiouy]*a[^aeiouy]*e[^aeiouy]*i[^aeiouy]*o[^aeiouy]*u[^aeiouy]*y[^aeiouy]*$' words
  abstemiously
   facetiously
half-seriously
   pareciously
```

Words with all their letters in order?

```
grep '^a*b*c*d*e*f*g*h*i*j*k*l*m*n*o*p*q*r*s*t*u*v*w*x*y*z*$' words | wc -l
1370
grep $pattern_above words | awk '{print length, $0;}' | sort -n | tail | column
6 knoppy  7 alloquy 7 begorry 7 billowy   7 egilops
6 knotty  7 beefily 7 belloot 7 deglory   8 aegilops
```

I before E? Except after C? And in words like "neighbor" and "weigh"?

```
grep ’ei' words | wc -l
5677
grep 'ie' words | wc -l
16690
```

2.94:1

```
grep 'cei' words | wc -l
300
grep 'cie' words | wc -l
863
```

2.87:1 … that difference begets a named exception?

```
grep 'eigh' words | wc -l
429
grep 'iegh' words | wc -l
3
```

143:1 … okay, this one's pretty legit

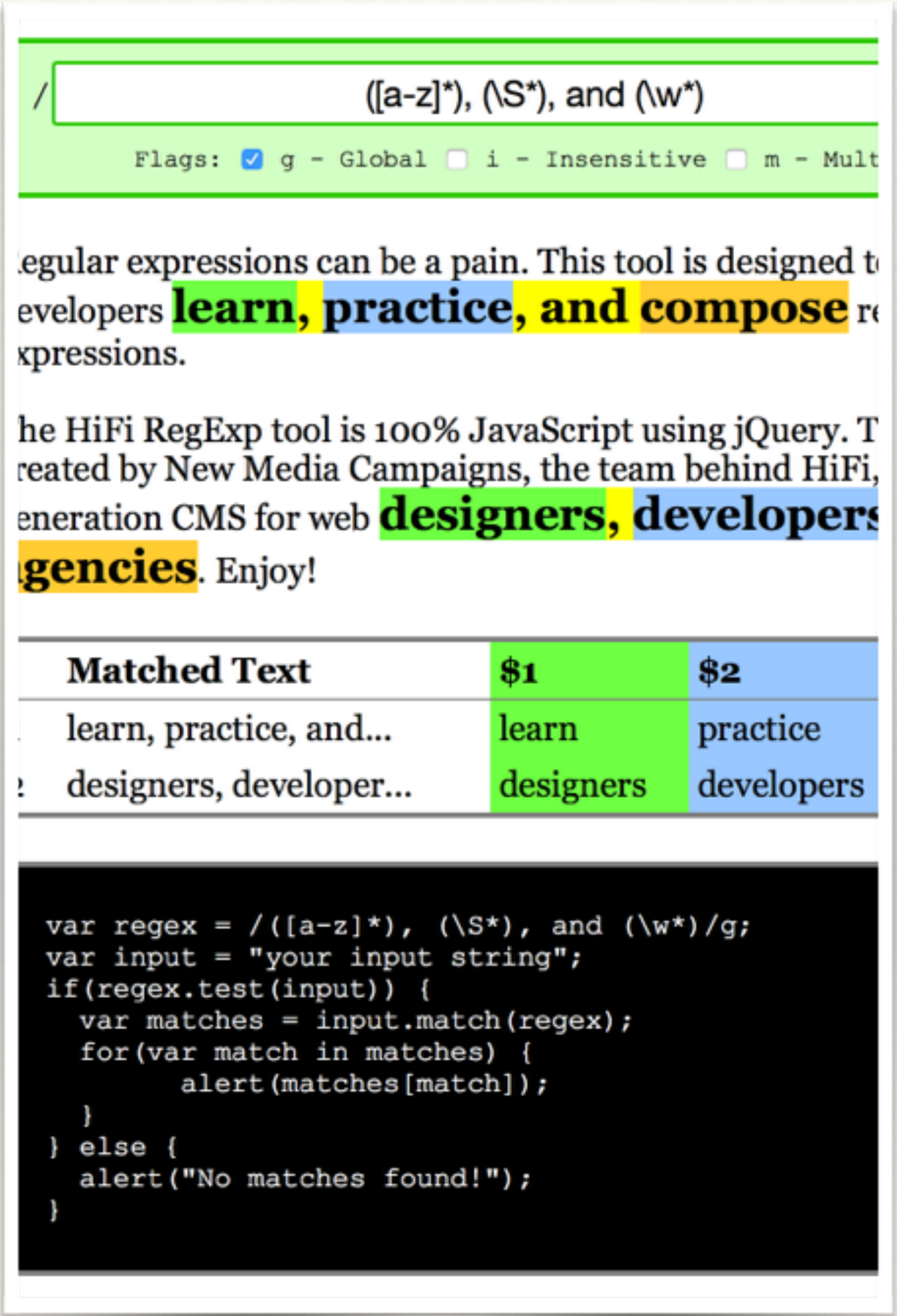
```
grep 'iegh' words
abiegh
driegh
skiegh
```

Floating point number?

```
[-+]?([0-9]+\.?[0-9]*|\.[0-9]+)([Ee][-+]?[0-9]+)?
```

http://www.gethifi.com/tools/regex

Need Practice?



/ ([a-z]*), (\S*), and (\w*)

Flags: ☑ g - Global ☐ i - Insensitive ☐ m - Mult

egular expressions can be a pain. This tool is designed t evelopers **learn, practice, and compose** re xpressions.

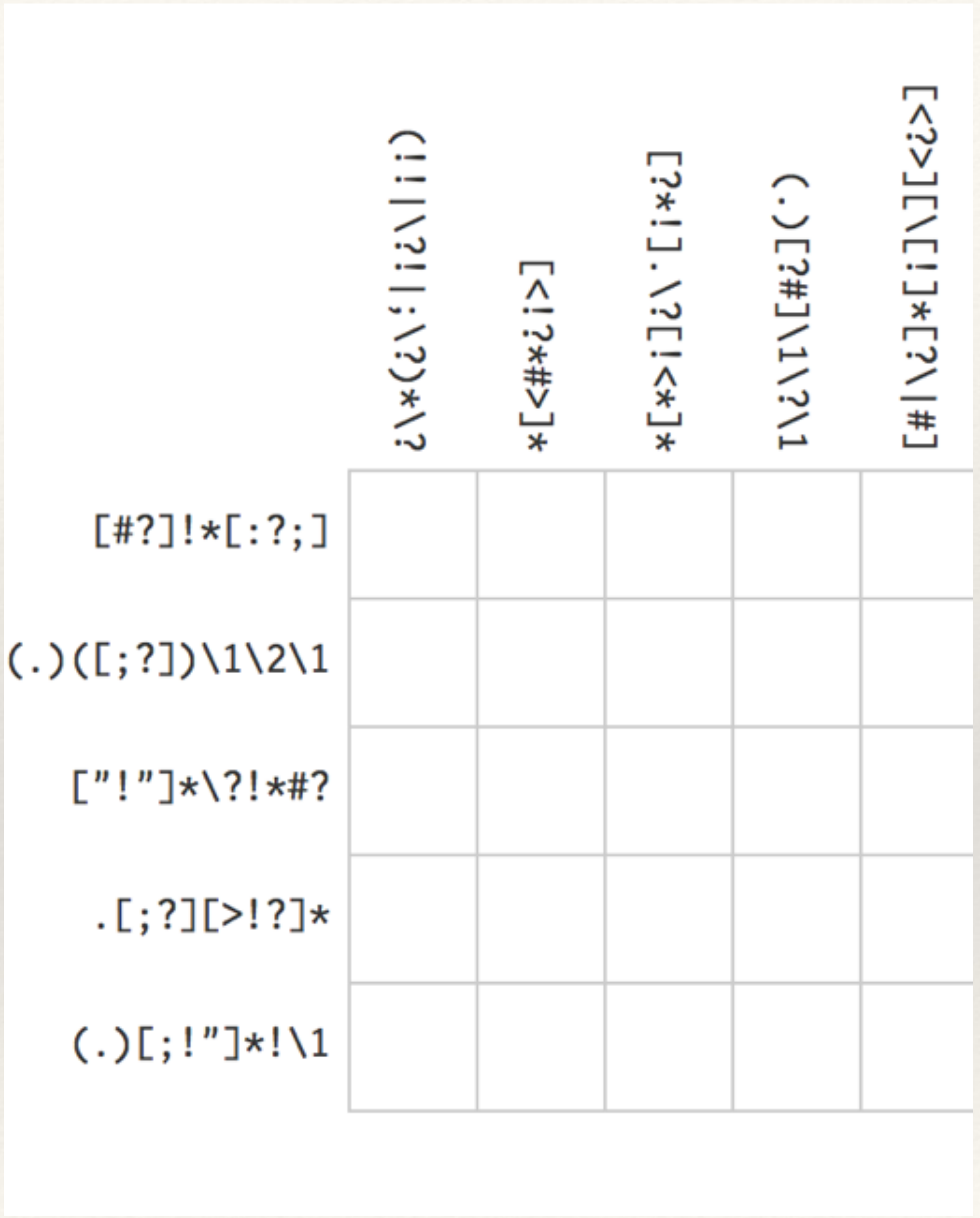
he HiFi RegExp tool is 100% JavaScript using jQuery. T reated by New Media Campaigns, the team behind HiFi, eneration CMS for web **designers, developers gencies**. Enjoy!

Matched Text	$1	$2
learn, practice, and...	learn	practice
designers, developer...	designers	developers

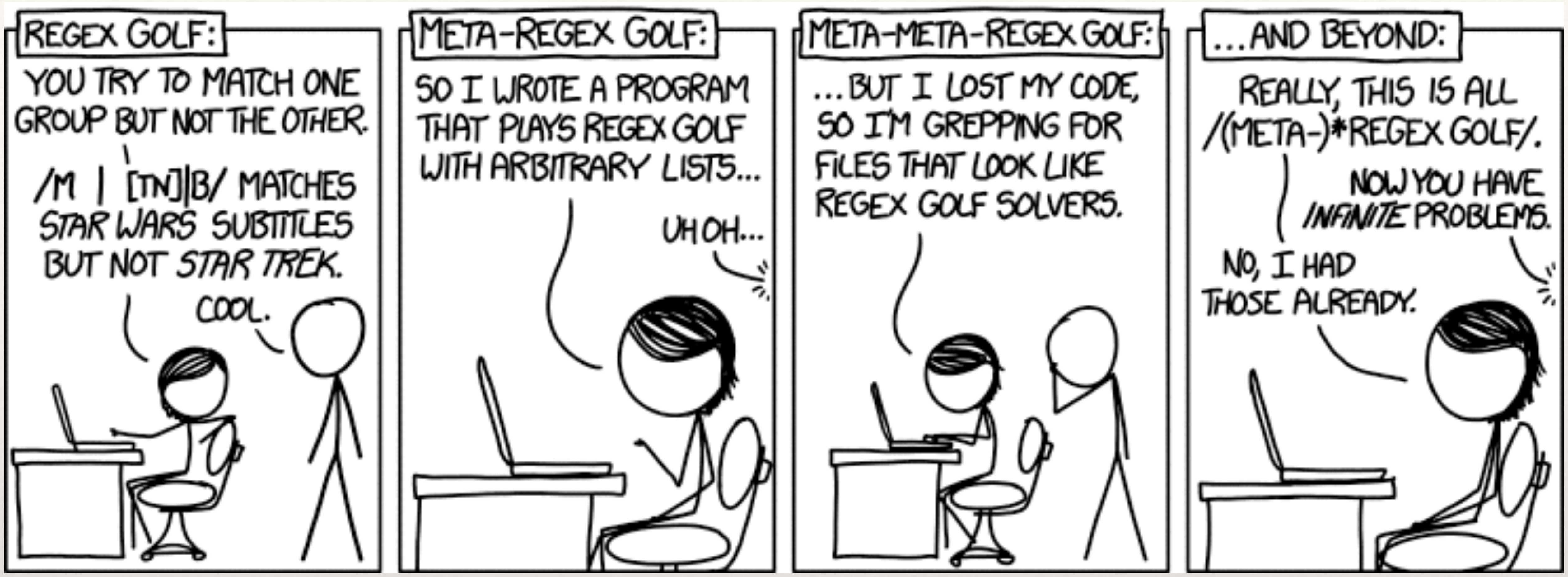
```
var regex = /([a-z]*), (\S*), and (\w*)/g;
var input = "your input string";
if(regex.test(input)) {
  var matches = input.match(regex);
  for(var match in matches) {
        alert(matches[match]);
  }
} else {
  alert("No matches found!");
}
```

https://regexcrossword.com/challenges/

Just a Masochist?



	(!!|\?!|;\?)*\?	[<!?*#>]*	[?*!].\?[!<*]*	(.)[?#]\1\?\1	[<?>][\[!]*[?\\|#]
[#?]!*[:?;]					
(.)([;?])\1\2\1					
["!"]*\?!*#?					
.[;?][>!?]*					
(.)[;!"]*!\1					


REGEX GOLF:
YOU TRY TO MATCH ONE GROUP BUT NOT THE OTHER.
/M | [TN]|B/ MATCHES STAR WARS SUBTITLES BUT NOT STAR TREK.
COOL.
META-REGEX GOLF:
SO I WROTE A PROGRAM THAT PLAYS REGEX GOLF WITH ARBITRARY LISTS...
UH OH...
META-META-REGEX GOLF:
...BUT I LOST MY CODE, SO I'M GREPPING FOR FILES THAT LOOK LIKE REGEX GOLF SOLVERS.
...AND BEYOND:
REALLY, THIS IS ALL /(META-)*REGEX GOLF/.
NOW YOU HAVE INFINITE PROBLEMS.
NO, I HAD THOSE ALREADY.

/bu|[rn]t|[coy]e|[mtg]a|j|iso|n[hl]|[ae]d|lev|sh|[lnd]i|[po]o|ls/
matches the last names of elected US presidents but not their opponents.

http://xkcd.com/1313/

Day-to-day grep Workflows

- It is undeniably fun to find the perfect regexp.
  - It is undeniably more common to pipe them together until you exhaustively get exactly what you want
  - "Uh, name starts with J, upperclassman, BSE …"

```
. . . | grep J     | egrep '16|7' | egrep '(COS)|(ORF)|(ELE)|(CBE)|(MAE)|(CEE)'
      Janet 17 COS  Janet 17 COS  Janet 17 COS
      James 16 ORF  James 16 ORF  James 16 ORF
      Jenna 17 WWS  Jenna 17 WWS
      Julia 18 COS  Joaoa 16 HIS
      Jesse 19 BSE  …
      Joaoa 16 HIS
      …
```

Day-to-day grep "Workflows"

```
grep $p p.sh
^C (after 30 seconds of hanging)
```

```
grep '$p' p.sh
$python -c '#do nothing'
```

```
grep university cslist.txt  #immediately returns no matches
grep -i university cslist.txt
…
Johnson, Elizabeth Chief Reader   XAVIER UNIVERSITY
Leyzberg, DanielReader PRINCETON UNIVERSITY
Liu, David  Reader Indiana University Purdue University Fort Wayne
…

grep -i hu cslist.txt
Baker, Jeffrey   Reader Huntsville High School
Cunningham, Susie  Reader Indiana Academy of Science Math & Humanities
Hu, Chenglie  Reader CARROLL UNIVERSITY
Huggins, James   Question Leader  KETTERING UNIVERSITY
Hughes, Mary AnnReader Albert Gallatin Area School District
Wang, Huanjing   Reader WESTERN KENTUCKY UNIVERSITY

grep -w Hu cslist.txt
Hu, Chenglie  Reader CARROLL UNIVERSITY
```

http://public.wsu.edu/~brians/errors/backslash.html

It’s 2016, why do people still say this?

- Do you cringe every time you hear someone read a URL?
  - H-T-T-P … Colon …
  - Backslash Backslash
  - W-W-W
- No doubt these people are prone to type it that way, too. And you are going to end up responsible for cleaning up their mess

WHENEVER I LEARN A NEW SKILL I CONCOCT ELABORATE FANTASY SCENARIOS WHERE IT LETS ME SAVE THE DAY.
OH NO! THE KILLER MUST HAVE FOLLOWED HER ON VACATION!
BUT TO FIND THEM WE'D HAVE TO SEARCH THROUGH 200 MB OF EMAILS LOOKING FOR SOMETHING FORMATTED LIKE AN ADDRESS!
IT'S HOPELESS!
EVERYBODY STAND BACK.
I KNOW REGULAR EXPRESSIONS.
PERL!
TAP TAP

http://xkcd.com/208/

Everybody Stand Back. Uh, …

```
http://good.url
this is a tent! /\
\/ for \/endetta
http:\\bad.url
```

```
grep \ slashandburn
^C
```

```
grep '\' slashandburn
grep: trailing backslash (\)
```

```
grep '\\' slashandburn
this is a tent! /\
\/ for \/endetta
http:\\bad.url
```

```
grep '\\\' slashandburn
grep: trailing backslash (\)
```

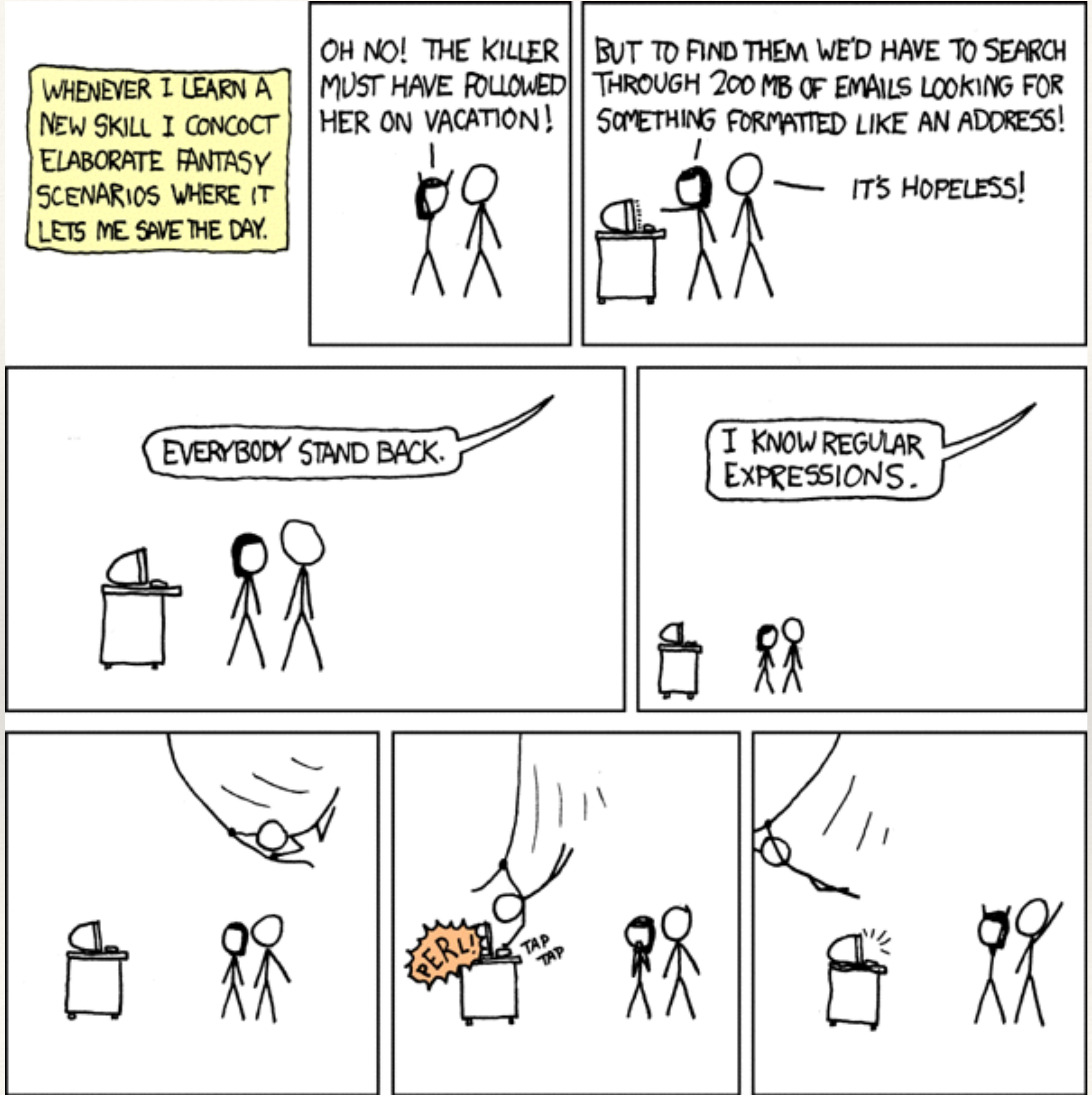
NB: you can use other delimiters to avoid having to escape every slash in your RE and replacement:

```
  sed s#'\\\\'#'//'#g
```

```
grep '\\\\' slashandburn
http:\\bad.url
```

```
sed s/'\\\\'/'//'/g slashandburn
sed: 1: "s/\\\\////g": bad flag in substitute command: '/'
```

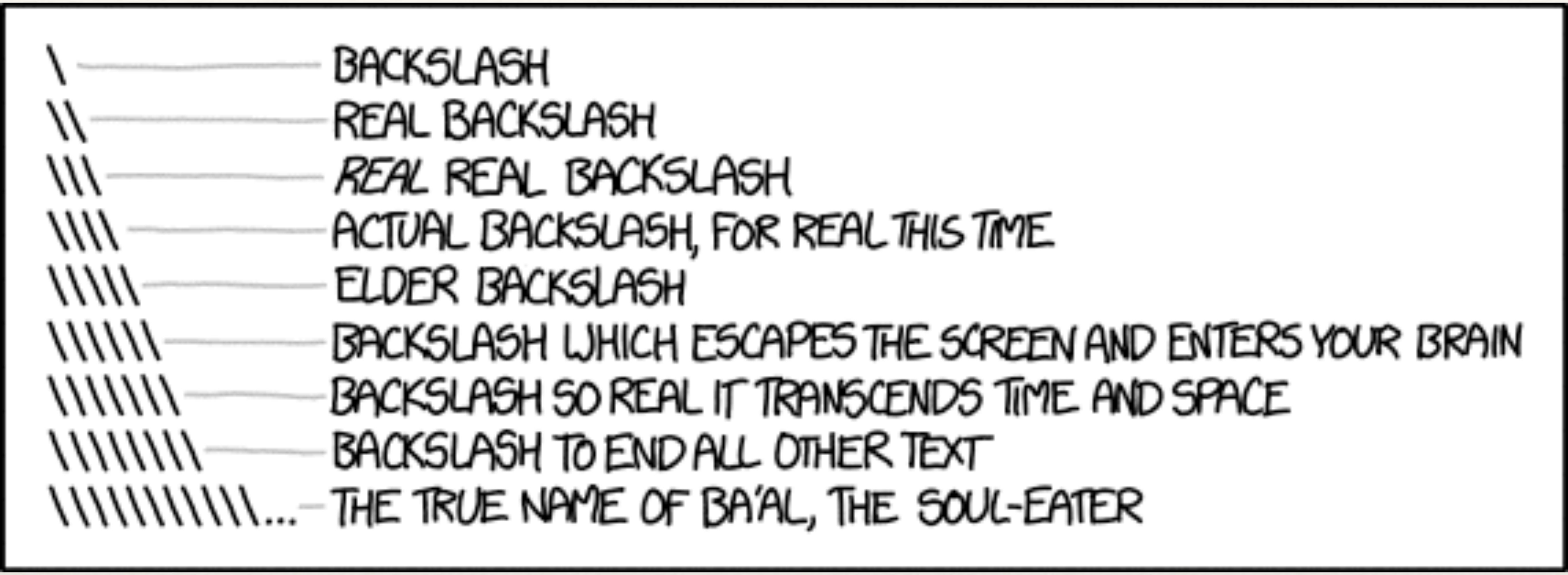
```
sed s/‘\\\\'/'\/\/'/g slashandburn
http://good.url
this is a tent! /\
\/ for \/endetta
http://bad.url
```


WHENEVER I LEARN A NEW SKILL I CONCOCT ELABORATE FANTASY SCENARIOS WHERE IT LETS ME SAVE THE DAY.
OH NO! THE KILLER MUST HAVE FOLLOWED HER ON VACATION!
BUT TO FIND THEM WE'D HAVE TO SEARCH THROUGH 200 MB OF EMAILS LOOKING FOR SOMETHING FORMATTED LIKE AN ADDRESS!
IT'S HOPELESS!
EVERYBODY STAND BACK.
I KNOW REGULAR EXPRESSIONS.
PERL!
TAP TAP

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

http://xkcd.com/208/

There's No Escaping Escaping


\ — BACKSLASH
\\ — REAL BACKSLASH
\\\ — REAL REAL BACKSLASH
\\\\ — ACTUAL BACKSLASH, FOR REAL THIS TIME
\\\\\ — ELDER BACKSLASH
\\\\\\ — BACKSLASH WHICH ESCAPES THE SCREEN AND ENTERS YOUR BRAIN
\\\\\\\ — BACKSLASH SO REAL IT TRANSCENDS TIME AND SPACE
\\\\\\\\ — BACKSLASH TO END ALL OTHER TEXT
\\\\\\\\\\\... — THE TRUE NAME OF BA'AL, THE SOUL-EATER

I searched my .bash_history for the line with the highest ratio of special characters to regular alphanumeric characters, and the winner was:

```
cat out.txt | grep -o "\\\[[(].*\\\[\])][^)\]]*$"
```

... I have no memory of this and no idea what I was trying to do, but I sure hope it worked.

http://xkcd.com/1638/

The grep Family

- fgrep
  - parallel search, but not actually full regexp (just fixed strings)
- agrep
  - "approximate" grep: search with errors permitted
- relatives that use similar regular expressions

ed	original Unix editor
sed	stream editor
vi, emacs, ...	editors
lex	lexical analyzer generator

- simpler variants
  - filename "wild cards" in Unix and other shells
  - "LIKE" operator in SQL, Visual Basic, etc.