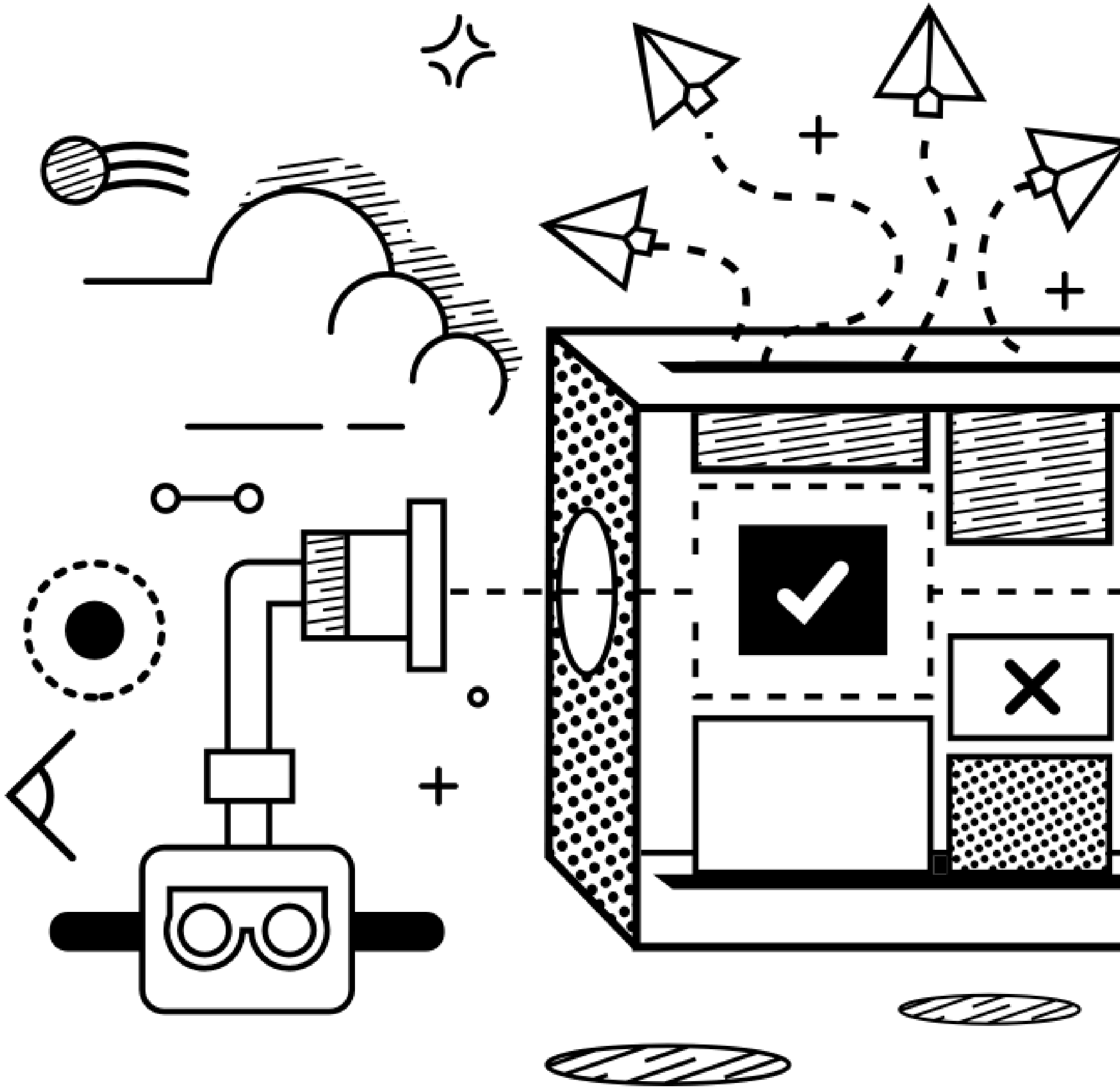
candour

Advanced robots.txt

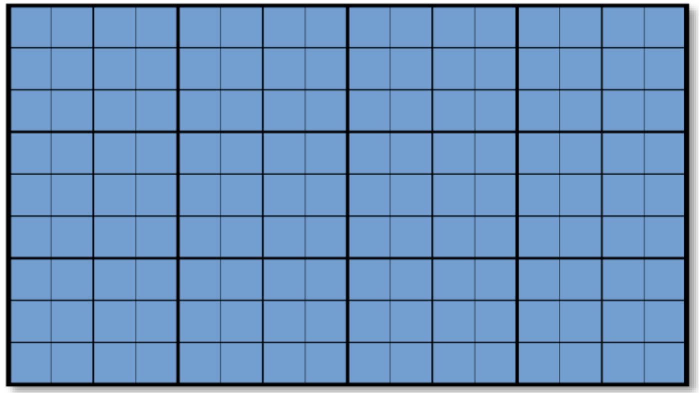
Trainer:

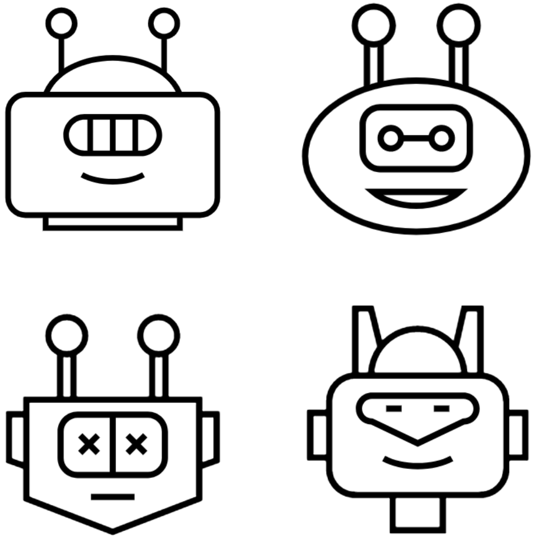
Mark Williams-Cook
Candour



PREMIER Google Partner

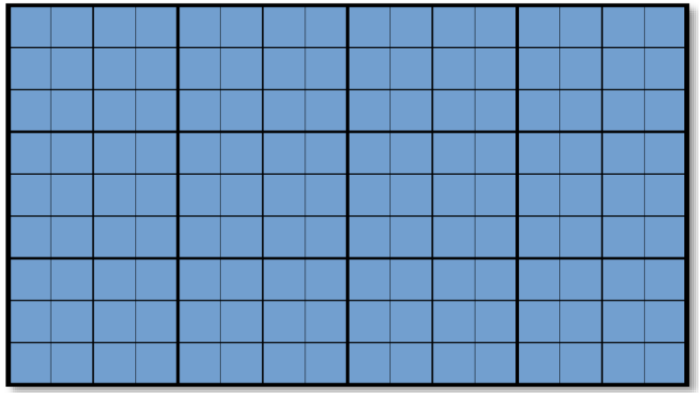
Filtered navigation

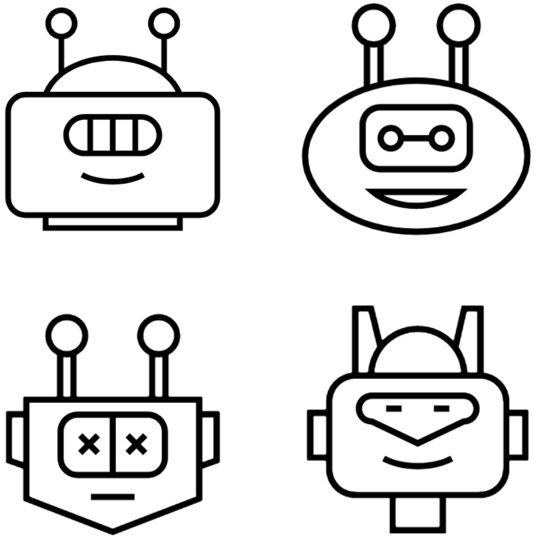




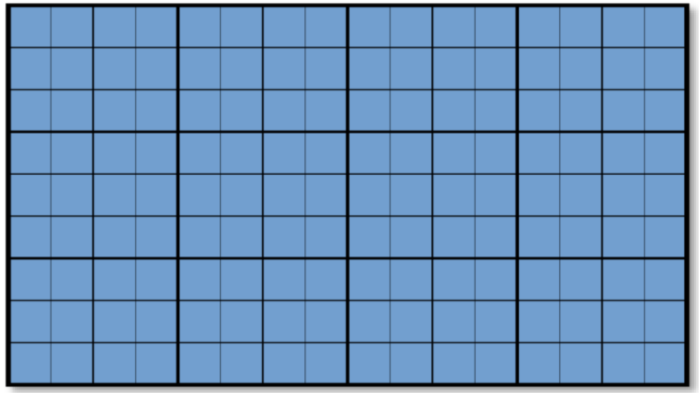
Filtered navigation

site.com/cat/womens-shoes

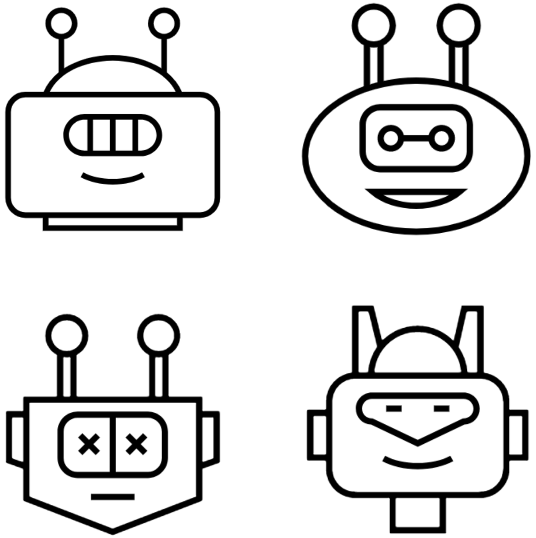




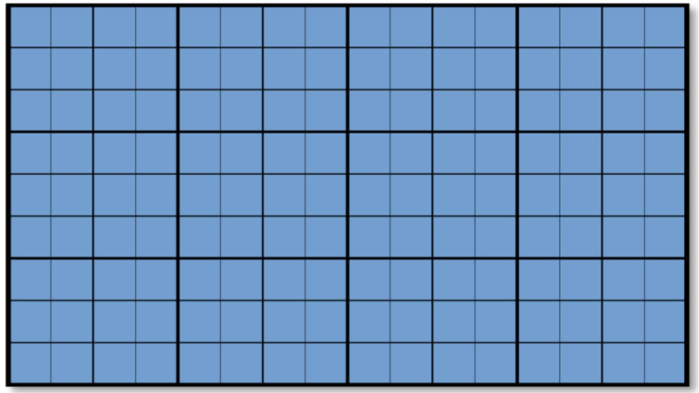
Filtered navigation



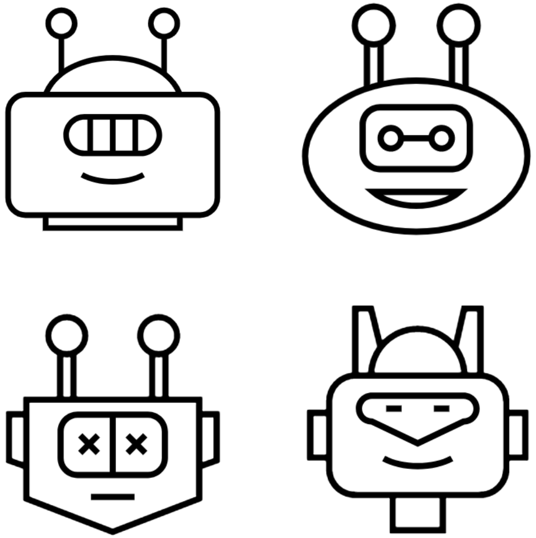
site.com/cat/womens-shoes?size=10



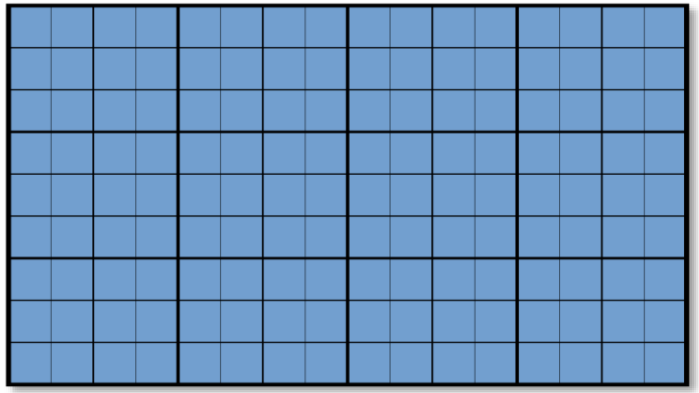
Filtered navigation



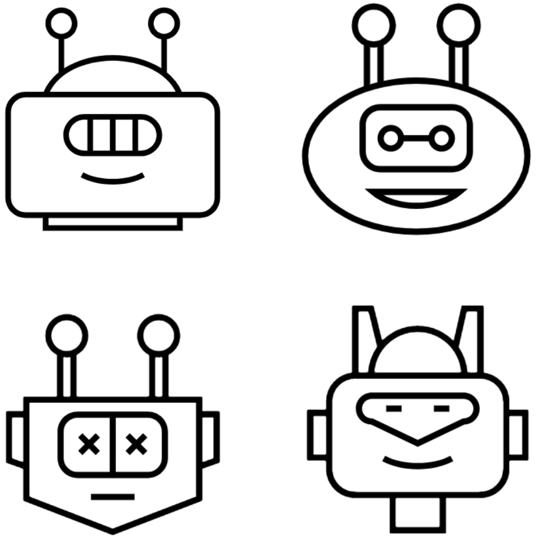
site.com/cat/womens-shoes?size=10&colour=red



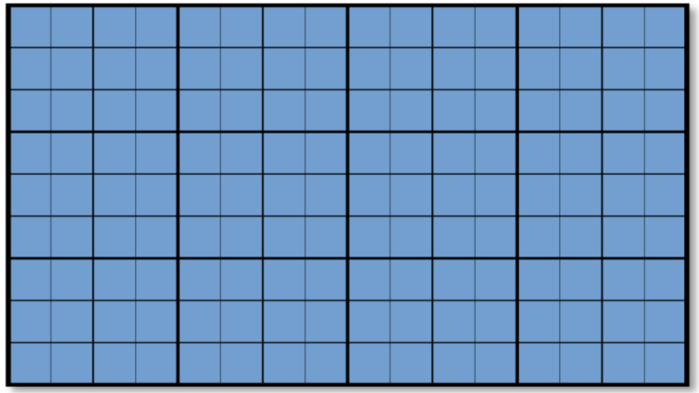
Filtered navigation



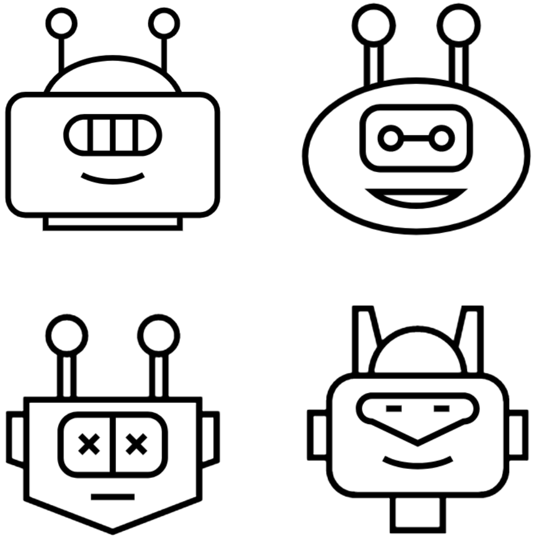
site.com/cat/womens-shoes?size=10&colour=red&price=30



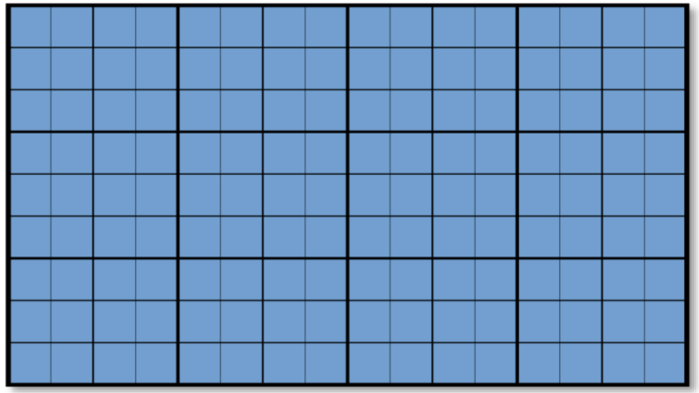
Filtered navigation



site.com/cat/womens-shoes?size=10&colour=red&price=30&order=desc

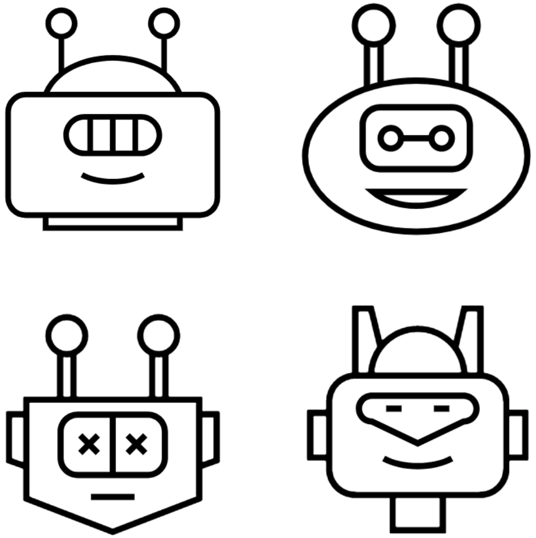


Filtered navigation

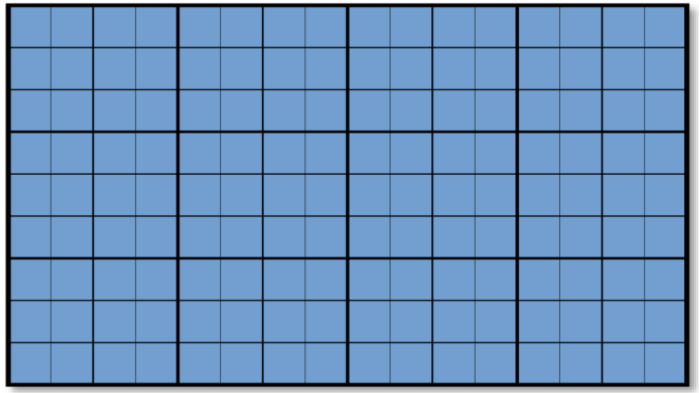


site.com/cat/womens-shoes?size=10
&colour=red&price=30&order=desc

12 sizes & 10 colours & 6 price brackets



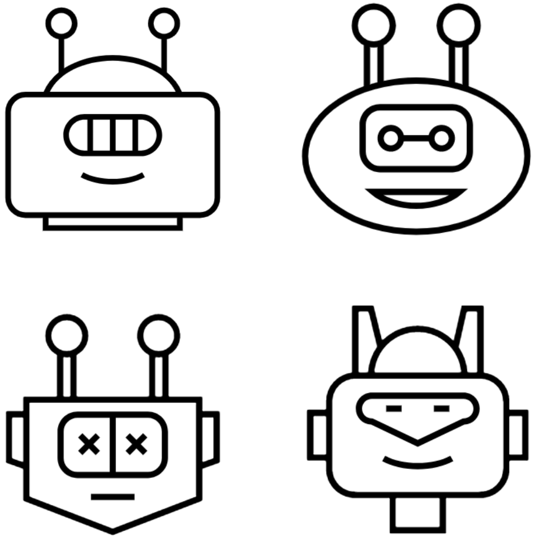
Filtered navigation



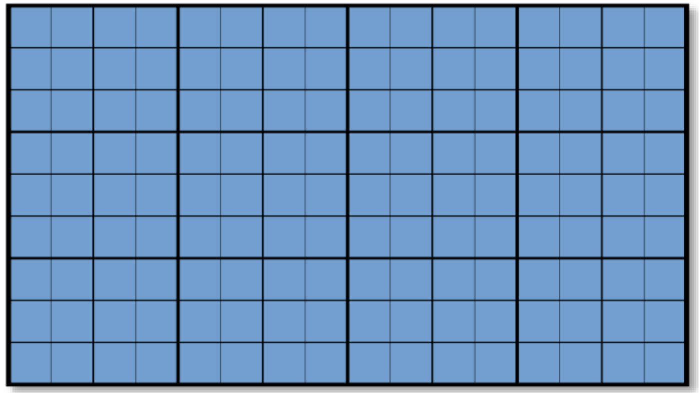
site.com/cat/womens-shoes?size=10&colour=red&price=30&order=desc

12 sizes & 10 colours & 6 price brackets

This is 10,000s of URL variations!



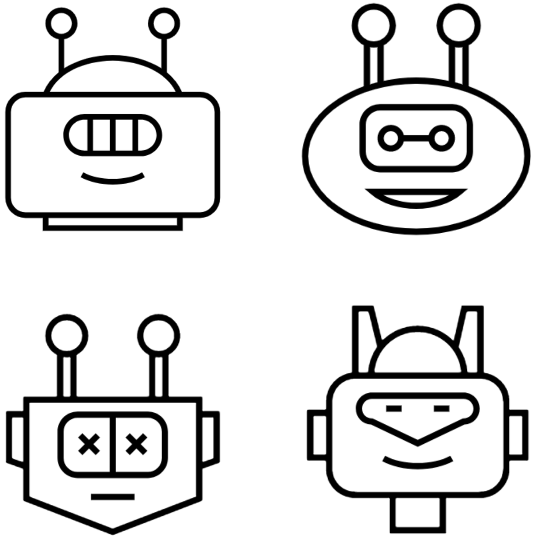
Filtered navigation



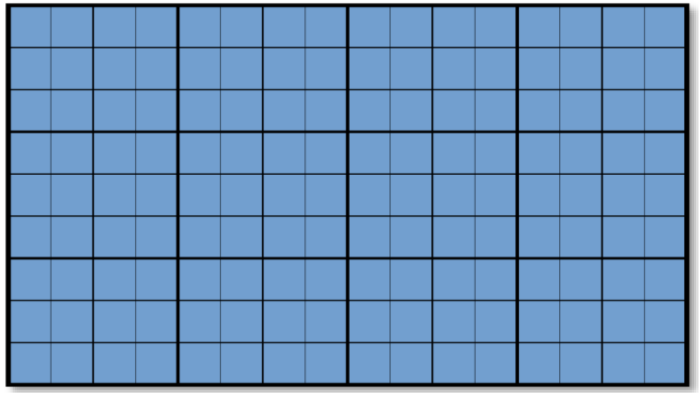
site.com/cat/womens-shoes?size=10

site.com/cat/womens-shoes?colour=red

site.com/cat/womens-shoes?price=30



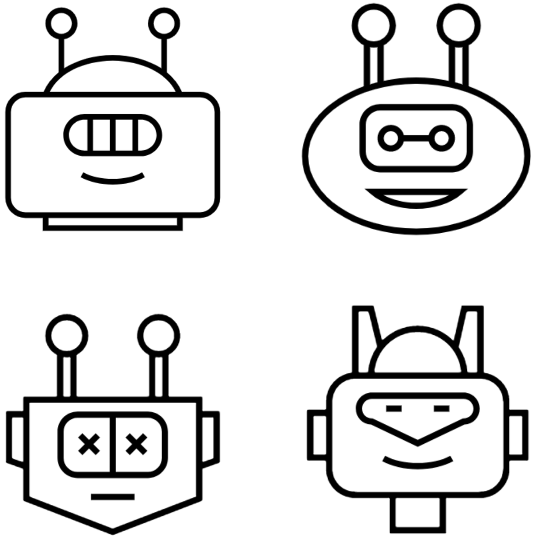
Filtered navigation



site.com/cat/womens-shoes?size=10

site.com/cat/womens-shoes?colour=red

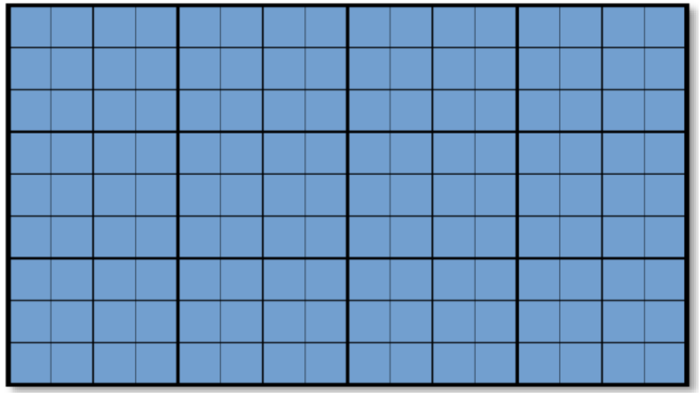
site.com/cat/womens-shoes?price=30



12 sizes & 10 colours & 6 price brackets

= 12 + 10 + 6 = 28 URLs

Filtered navigation

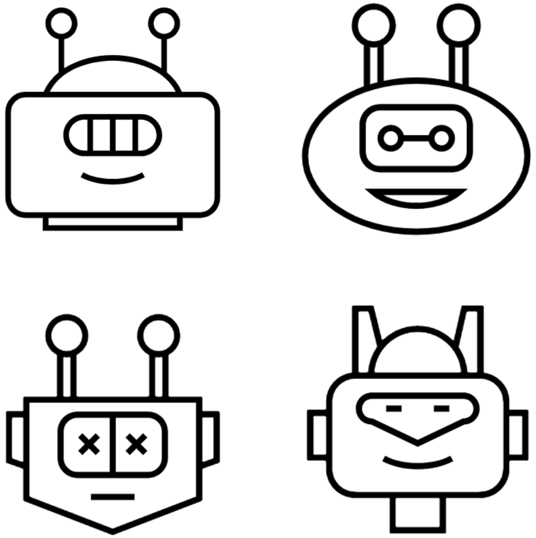


We want to allow: any single query string

site.com/cat/womens-shoes?size=10

site.com/cat/womens-shoes?colour=red

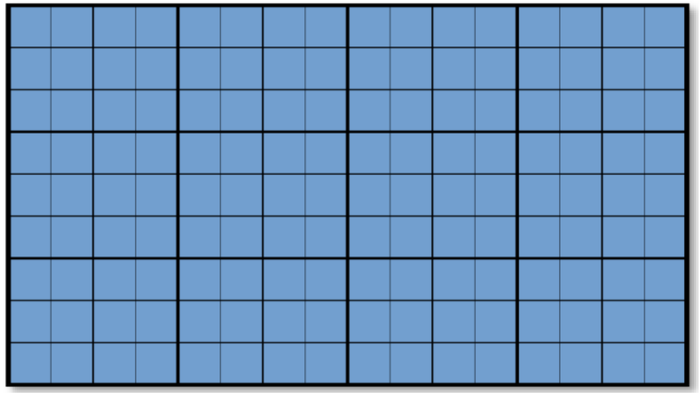
site.com/cat/womens-shoes?price=30



We want to disallow any multiple query string:

site.com/cat/womens-shoes?price=30&size=10

Filtered navigation



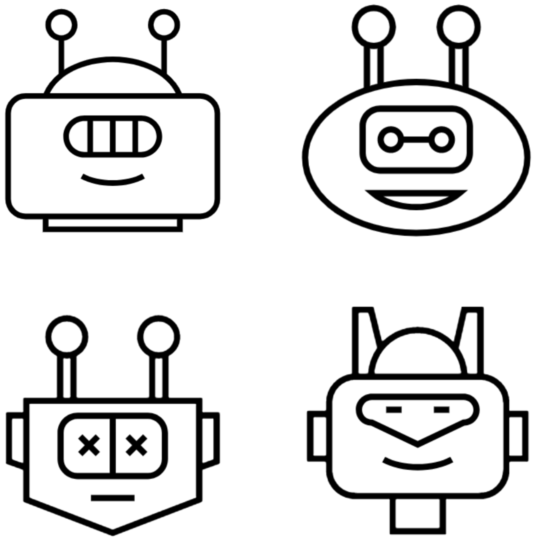
Hint:

Use wildcards (*) within your robots.txt rules

Wildcards are 0 or more of any valid character

E.g.

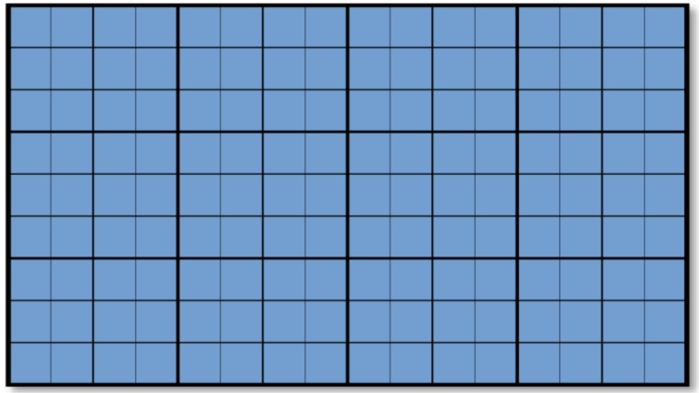
Disallow: /cat/*/



Challenge

Write out a robots.txt file that will allow any category URL with a single query string to be crawled but will block crawling of any multiple query string URLs.

Key concept



Robots.txt is one way to fix such issues – don’t rush off to implement anything yet!