Chapter 8, Strings:

- General introduction; string sorting (last week Friday)
- Tries (Monday)
- Regular expression (**Today**)

Today:

- What regular expressions are
- How to use regular expressions practically
- Why regular expressions are important theoretically

End-of-semester important dates

- Mon, Dec 9: Last project assigned ✓
- Tues, Dec 10: Last "normal" running of project grading script ✓
- Wed, Dec 11: Test 3 & 4 Review sheet distributed, Test 4 practice problems made available. ✓
- Thurs, Dec 12: Review lab (pick practice problems for Test 4)
- Fri, Dec 13, AM: "Two-minute warning" running of project grading script (Canvas gradebook will not be updated—see project report in your turn-in file)
  Note that Fri, Dec 13 is the *Last Day of Classes*.
- Fri, Dec 13, midnight: Official project deadline
- Sat, Dec 14, when I wake up: Permissions to turn-in folders turned off
- Mon, Dec 16: Project grading script run for final/semester grades
- Wed, Dec 18, 10:30am-12:30pm: Tests 3 and 4 (in lab)
  - Test 3: On paper (like Test 1) covering BSTs (ch 5), DP (Ch 6), hashtables (Ch 7) and strings (ch 8).
  - Test 4: At a computer (like Test 2) covering DP (Ch 6), hashtables (Ch 7) and strings (ch 8).


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.
GERU
TAP
TAP

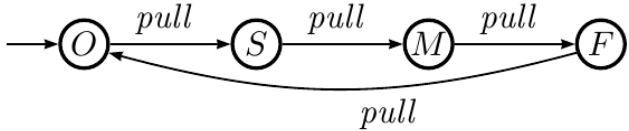
- An **alphabet** is a set of symbols, Σ.
- A **string** over an alphabet is a sequence of symbols from that alphabet. $\Sigma*$ is the set of all strings over alphabet Σ.
- A **language** over an alphabet is a set of strings, that is, a subset of $\Sigma*$.

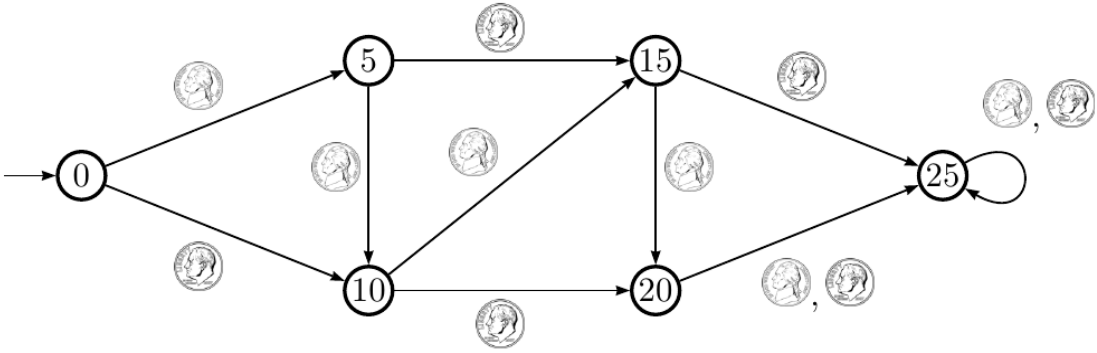
- **Regular expressions** constitute a system for specifying languages; a regular expression denotes a language.

base cases	$\emptyset$	the empty set of strings
	ε	the set containing the empty string, $\{""\}$
	a	the set containing only the string with only a, for some $a \in \Sigma$, $\{"a"\}$
recursive cases	rs	the set of strings made from concatening strings from r and s, $\{x + y \mid x \in r \wedge y \in s\}$, for some regular expressions r and s
	$r\|s$	the set of strings from r or s, $r \cup s$ for some regular expressions r and s
	$r*$	the set of strings made from concatenating 0 or more strings from r for some regular expression r

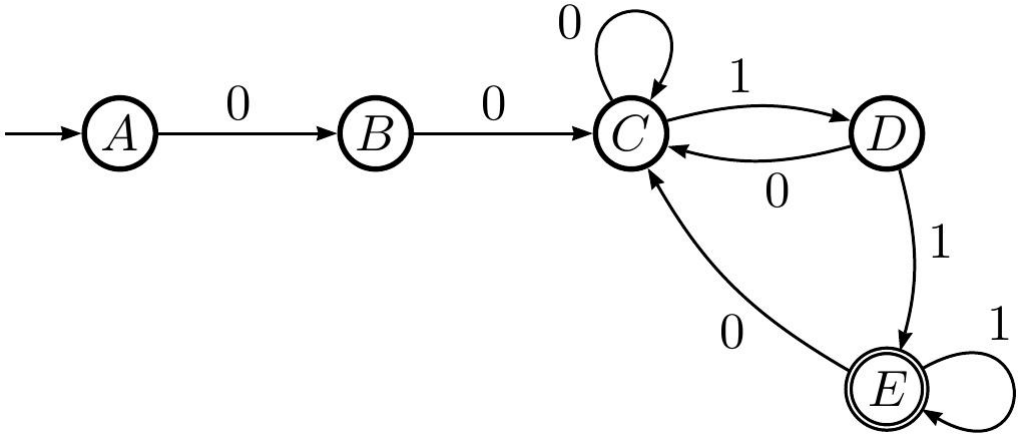
Abbreviation	**Meaning**	**Equivalence**
`[abc]`	One occurrence of any of these symbols	`(a|b|c)`
`[a-c]`	One occurrence of any symbol in this range	`(a|b|c)`
$r?$	Optionally an occurrence of a string defined by r	$(r\|\varepsilon)$
r^5	5 occurrences of a string defined by r	$rrrrr$
$r^{3,5}$	Between 3 and 5 occurrences of a string defined by r	$(rrr\|rrrr\|rrrrr)$
$r+$	One or more occurrences of a string defined by r	$rr*$

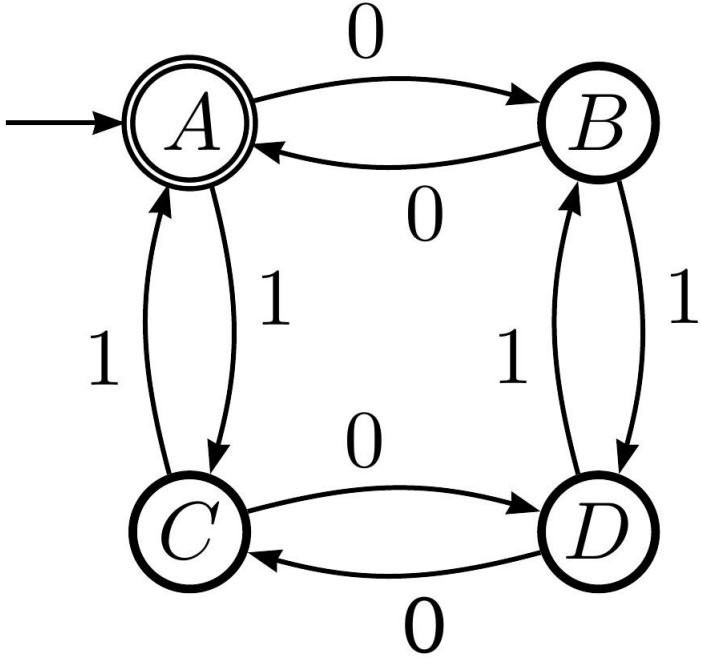
- *DNA sequences:* `(A|C|G|T)*`.
- *Identifiers:* `('?[A-Za-z][A-Za-z0-9_])|_`.
- *Phone numbers:* `[2-9][0-9]`^{2} – `[2-9][0-9]`^{2} – `[0-9]`^{4}.
- *Dates:* `((1[0-2])|[1-9])/(30|31|([12][0-9])|[1-9])/[1-9][0-9]`^{0,3}.
- *US Postal Addresses:* `[0-9]+ [NSEW]`^{0,2} `[A-Z][a-z]* (St|Ave|Rd|Ln|Dr|Blvd),([A-Z][a-z]*)*, [A-Z]`^{2}`[0-9]`^{5}.





Diagrams generated by VAUCANSON-G


A
B
C
D
E
0
0
0
1
0
1
0
1


A
B
C
D
0
0
1
1
1
1
0
0