Regular Expressions

Recap from Last Time

Regular Languages

- A language L is a ***regular language*** if there is a DFA D such that $\mathscr{L}(D) = L$.
- ***Theorem:*** The following are equivalent:
  - L is a regular language.
  - There is a DFA for L.
  - There is an NFA for L.

Language Concatenation

- If $w \in \Sigma^*$ and $x \in \Sigma^*$, then wx is the ***concatenation*** of w and x.
- If L_1 and L_2 are languages over Σ, the ***concatenation*** of L_1 and L_2 is the language L_1L_2 defined as

$$\boldsymbol{L_1L_2 = \{ wx \mid w \in L_1 \text{ and } x \in L_2 \}}$$

- Example: if $L_1 = \{$ `a`, `ba`, `bb` $\}$ and $L_2 = \{$ `aa`, `bb` $\}$, then

$$L_1L_2 = \{ \texttt{aaa}, \texttt{abb}, \texttt{baaa}, \texttt{babb}, \texttt{bbaa}, \texttt{bbbb} \}$$

New Stuff!

Lots and Lots of Concatenation

- Consider the language L = { `aa`, `b` }
- LL is the set of strings formed by concatenating pairs of strings in L.

{ `aaaa`, `aab`, `baa`, `bb` }

- LLL is the set of strings formed by concatenating triples of strings in L.

{ `aaaaaa`, `aaaab`, `aabaa`, `aabb`, `baaaa`, `baab`, `bbaa`, `bbb`}

- $LLLL$ is the set of strings formed by concatenating quadruples of strings in L.

{ `aaaaaaaa`, `aaaaaab`, `aaaabaa`, `aaaabb`, `aabaaaa`, `aabaab`, `aabbaa`, `aabbb`, `baaaaaa`, `baaaab`, `baabaa`, `baabb`, `bbaaaa`, `bbaab`, `bbbaa`, `bbbb`}

Language Exponentiation

- We can define what it means to "exponentiate" a language as follows:
- $L^0 = \{\varepsilon\}$
  - The set containing just the empty string.
  - Idea: Any string formed by concatenating zero strings together is the empty string.
- $L^{n+1} = LL^n$
  - Idea: Concatenating $(n+1)$ strings together works by concatenating n strings, then concatenating one more.
- ***Question:*** Why define $L^0 = \{\varepsilon\}$?

The Kleene Closure

- An important operation on languages is the ***Kleene Closure***, which is defined as

$$\boldsymbol{L^* = \{ w \in \Sigma^* \mid \exists n \in \mathbb{N}.\ w \in L^n \}}$$

- Mathematically:

$$\boldsymbol{w \in L^* \quad \textbf{iff} \quad \exists n \in \mathbb{N}.\ w \in L^n}$$

- Intuitively, all possible ways of concatenating zero or more strings in L together, possibly with repetition.

The Kleene Closure

If L = { **a**, **bb** }, then L^* = {

ε,

a, **bb**,

aa, **abb**, **bba**, **bbbb**,

aaa, **aabb**, **abba**, **abbbb**, **bbaa**, **bbabb**, **bbbba**, **bbbbbb**,

…

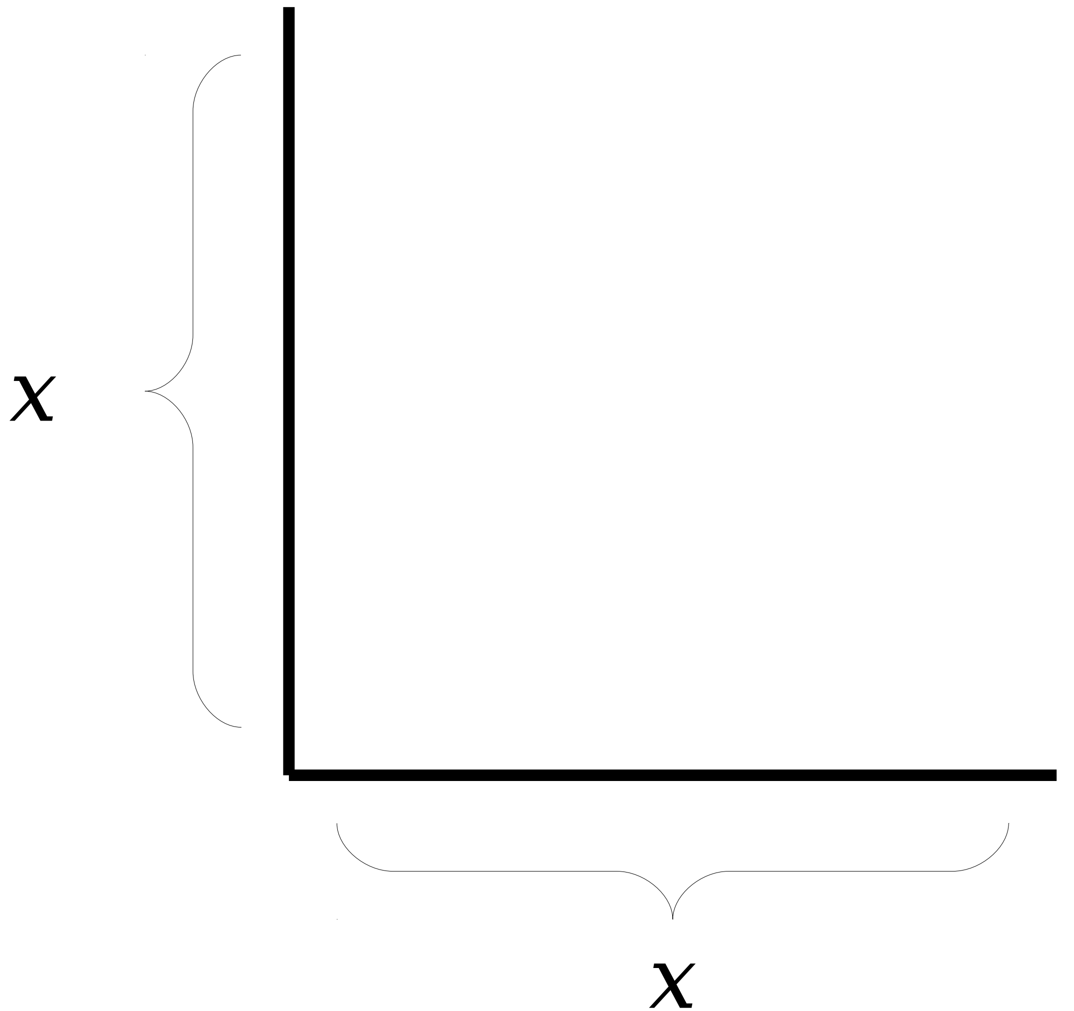
}

Think of L* as the set of strings you can make if you have a collection of stamps – one for each string in L – and you form every possible string that can be made from those stamps.

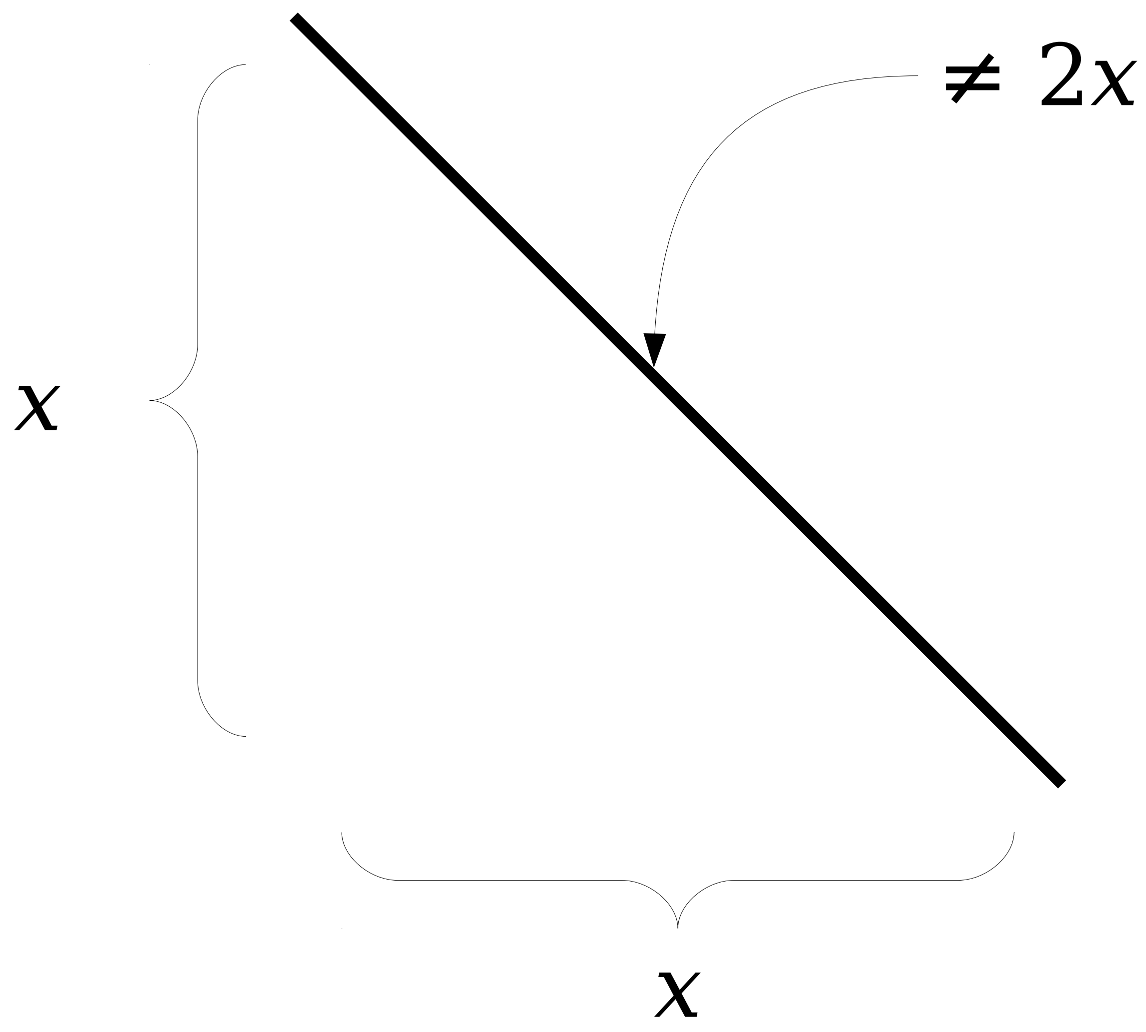
Reasoning about Infinity

- If L is regular, is L^* necessarily regular?
- **⚠ A Bad Line of Reasoning: ⚠**
  - $L^0 = \{ \varepsilon \}$ is regular.
  - $L^1 = L$ is regular.
  - $L^2 = LL$ is regular
  - $L^3 = L(LL)$ is regular
  - …
  - Regular languages are closed under union.
  - So the union of all these languages is regular.

Reasoning about Infinity



Reasoning about Infinity



Reasoning about Infinity

$$0.9 < 1$$

Reasoning about Infinity

$$0.9999\overline{9} < 1$$

Reasoning about Infinity

$$0.9999\overline{9} \not< 1$$

Reasoning about Infinity

0 is finite

Reasoning about Infinity

∞ is finite

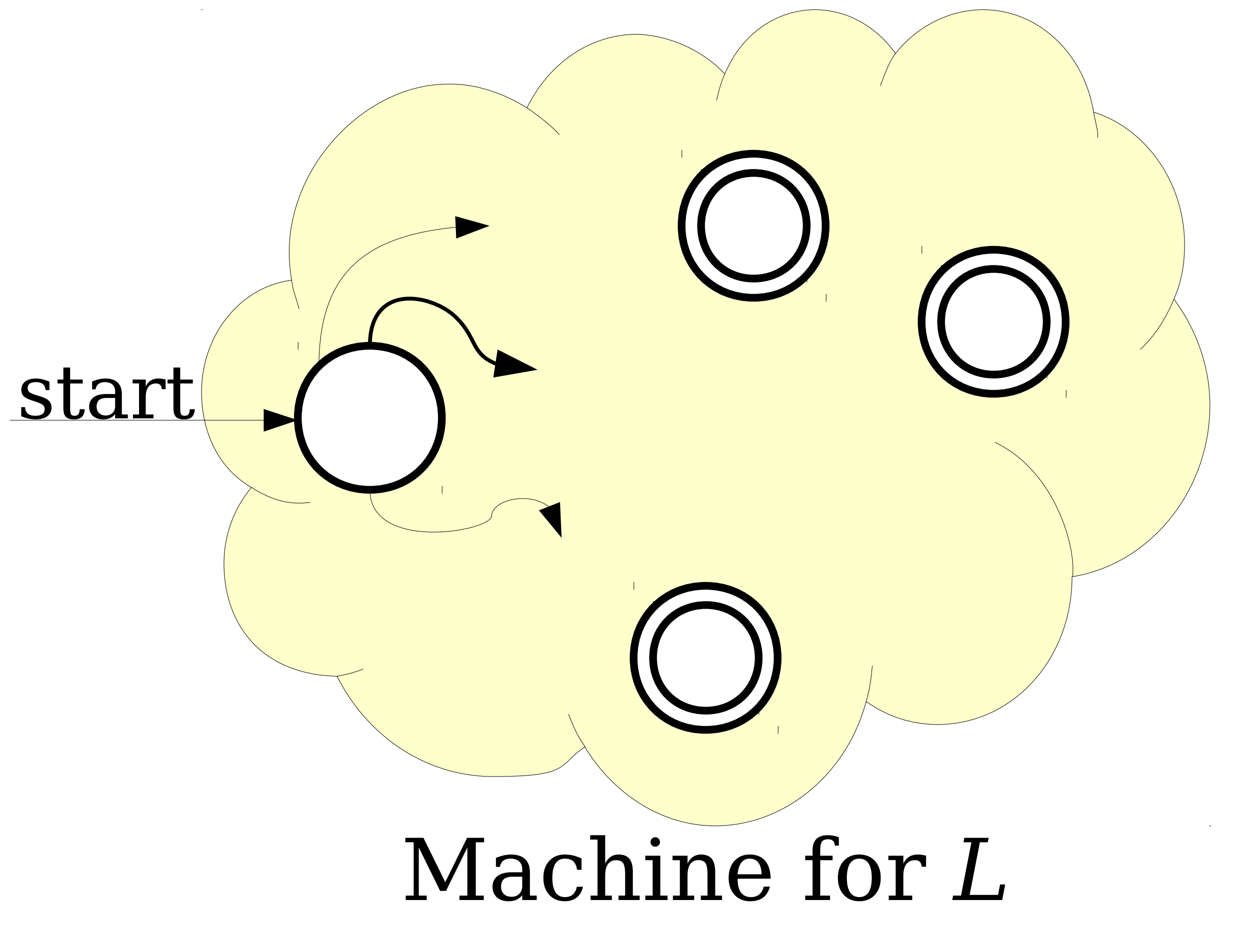
^ not

Reasoning About the Infinite

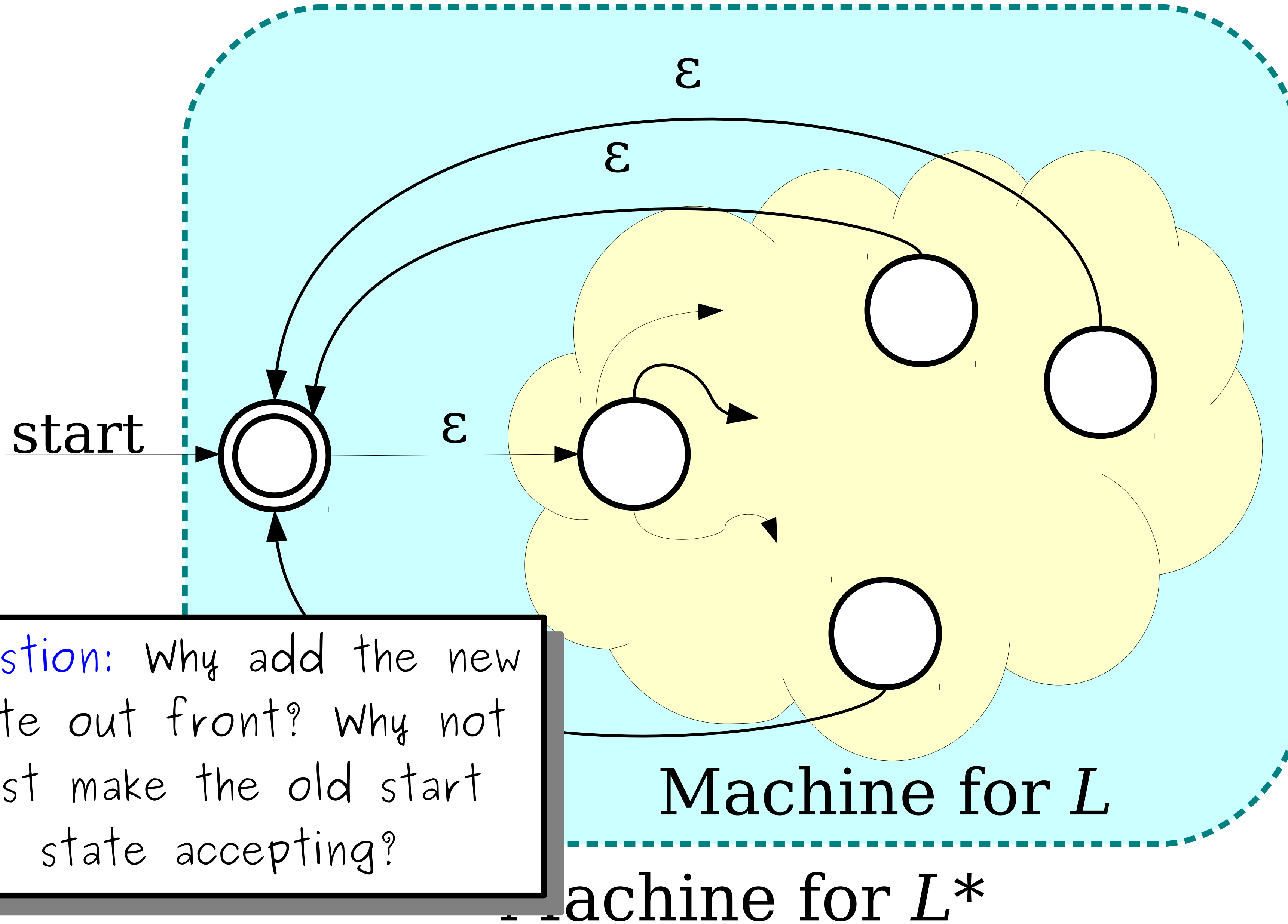
- If a series of finite objects all have some property, the "limit" of that process *does not* necessarily have that property.
- In general, it is not safe to conclude that some property that always holds in the finite case must hold in the infinite case.
  - (This is why calculus is interesting).

Idea: Can we directly convert an NFA for language L to an NFA for language L^*?

The Kleene Star



The Kleene Star



Question: Why add the new state out front? Why not just make the old start state accepting?

Closure Properties

- ***Theorem:*** If L_1 and L_2 are regular languages over an alphabet Σ, then so are the following languages:
  - $\overline{L_1}$
  - $L_1 \cup L_2$
  - $L_1 \cap L_2$
  - L_1L_2
  - L_1^*
- These properties are called ***closure properties of the regular languages***.

Another View of Regular Languages

Rethinking Regular Languages

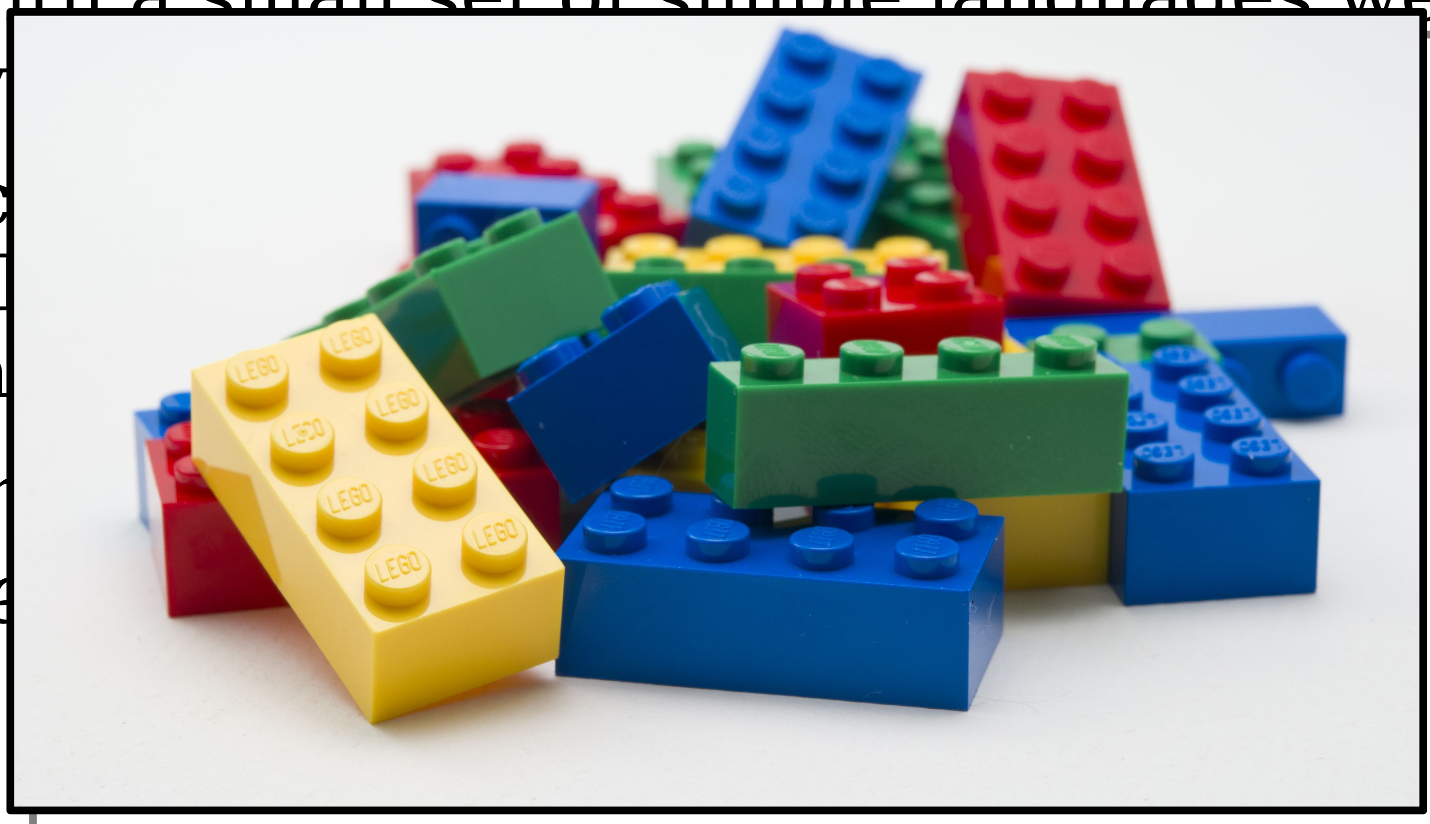
- We currently have several tools for showing a language is regular.
  - Construct a DFA for it.
  - Construct an NFA for it.
  - Apply closure properties to existing languages.
- We have not spoken much of this last idea.

Constructing Regular Languages

- ***Idea:*** Build up all regular languages as follows:
  - Start with a small set of simple languages we already know to be regular.
  - Using closure properties, combine these simple languages together to form more elaborate languages.
- *A bottom-up approach to the regular languages.*

Constructing Regular Languages

- ***Idea:*** Build up all regular languages as follows:
  - Start with a small set of simple languages we already
  - Using c simple l elabora
- *A bottom language*



Regular Expressions

- ***Regular expressions*** are a way of describing a language via a string representation.
- Used extensively in software systems for string processing and as the basis for tools like `grep` and `flex`.
- Conceptually, regular languages are strings describing how to assemble a larger language out of smaller pieces.

Atomic Regular Expressions

- The regular expressions begin with three simple building blocks.
- The symbol **Ø** is a regular expression that represents the empty language Ø.
- For any **a** ∈ Σ, the symbol **a** is a regular expression for the language {**a**}.
- The symbol **ε** is a regular expression that represents the language {ε}.
  - ***Remember: {ε} ≠ Ø!***
  - ***Remember: {ε} ≠ ε!***

Compound Regular Expressions

- If R_1 and R_2 are regular expressions, $\boldsymbol{R_1R_2}$ is a regular expression for the *concatenation* of the languages of R_1 and R_2.
- If R_1 and R_2 are regular expressions, $\boldsymbol{R_1 \cup R_2}$ is a regular expression for the *union* of the languages of R_1 and R_2.
- If R is a regular expression, $\boldsymbol{R^*}$ is a regular expression for the *Kleene closure* of the language of R.
- If R is a regular expression, $\boldsymbol{(R)}$ is a regular expression with the same meaning as R.

Operator Precedence

- Regular expression operator precedence:

$$(R)$$

$$R^*$$

$$R_1R_2$$

$$R_1 \cup R_2$$

- So `ab*c∪d` is parsed as `((a(b*))c)∪d`

Regular Expression Examples

- The regular expression **trick∪treat** represents the regular language { **trick**, **treat** }.
- The regular expression **booo*** represents the regular language { **boo**, **booo**, **boooo**, … }.
- The regular expression **candy!(candy!)*** represents the regular language { **candy!**, **candy!candy!**, **candy!candy!candy!**, … }.

Regular Expressions, Formally

- The ***language of a regular expression*** is the language described by that regular expression.
- Formally:
  - $\mathscr{L}(\varepsilon) = \{\varepsilon\}$
  - $\mathscr{L}(\emptyset) = \emptyset$
  - $\mathscr{L}(\mathbf{a}) = \{\mathbf{a}\}$
  - $\mathscr{L}(R_1R_2) = \mathscr{L}(R_1)\ \mathscr{L}(R_2)$
  - $\mathscr{L}(R_1 \cup R_2) = \mathscr{L}(R_1) \cup \mathscr{L}(R_2)$
  - $\mathscr{L}(R^*) = \mathscr{L}(R)^*$
  - $\mathscr{L}((R)) = \mathscr{L}(R)$

Worthwhile activity: Apply this recursive definition to

a(b∪c)((d))

and see what you get.

Designing Regular Expressions

- Let $\Sigma = \{$**`0`**, **`1`**$\}$
- Let $L = \{\ w \in \Sigma^* \mid w$ contains **`00`** as a substring $\}$

$$\mathbf{(0 \cup 1)^*00(0 \cup 1)^*}$$

11011100101
0000
1111011110011111

Designing Regular Expressions

- Let $\Sigma = \{$**`0`**, **`1`**$\}$
- Let $L = \{ w \in \Sigma^* \mid w$ contains **`00`** as a substring $\}$

$$\Sigma^*00\Sigma^*$$

11011100101
0000
11111011110011111

Designing Regular Expressions

Let $\Sigma = \{\mathbf{0}, \mathbf{1}\}$

Let $L = \{ w \in \Sigma^* \mid |w| = 4 \}$

The length of a string w is denoted |w|

Designing Regular Expressions

- Let $\Sigma = \{$**`0`**, **`1`**$\}$
- Let $L = \{ w \in \Sigma^* \mid |w| = 4 \}$

$$\Sigma\Sigma\Sigma\Sigma$$

0000
1010
1111
1000

Designing Regular Expressions

- Let $\Sigma = \{\mathtt{0}, \mathtt{1}\}$
- Let $L = \{ w \in \Sigma^* \mid |w| = 4 \}$

$$\mathbf{\Sigma^4}$$

0000
1010
1111
1000

Designing Regular Expressions

- Let Σ = {**0**, **1**}
- Let L = { $w \in \Sigma^*$ | w contains at most one **0** }

$$\mathbf{1^*(0 \cup \varepsilon)1^*}$$

11110111
111111
0111
0

Designing Regular Expressions

- Let $\Sigma = \{$**`0`**, **`1`**$\}$
- Let $L = \{ w \in \Sigma^* \mid w$ contains at most one **`0`** $\}$

1*0?1*

11110111
111111
0111
0

A More Elaborate Design

- Let Σ = { `a`, `.`, `@` }, where `a` represents "some letter."
- Let's make a regex for email addresses.

aa*(.aa*)*@aa*.aa*(.aa*)*

cs103@cs.stanford.edu
first.middle.last@mail.site.org
dot.at@dot.com

A More Elaborate Design

- Let Σ = { **a**, **.**, **@** }, where **a** represents "some letter."
- Let's make a regex for email addresses.

$$\mathbf{a^+ \; (.a^+)^* \; @ \; a^+.a^+ \; (.a^+)^*}$$

cs103@cs.stanford.edu
first.middle.last@mail.site.org
dot.at@dot.com

A More Elaborate Design

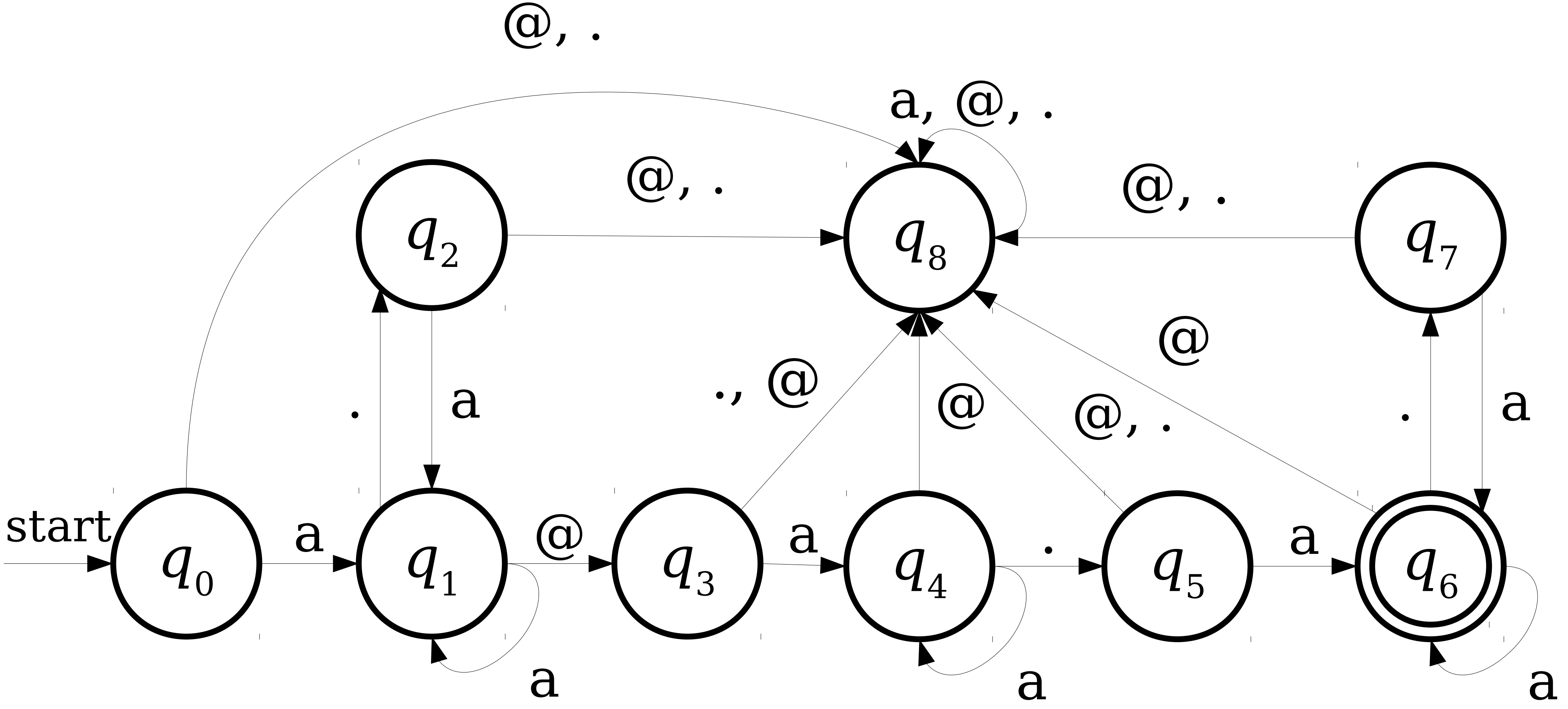
- Let Σ = { **a**, **.**, **@** }, where **a** represents "some letter."
- Let's make a regex for email addresses.

$$\mathbf{a^+(.a^+)^*@a^+(.a^+)^+}$$

cs103@cs.stanford.edu
first.middle.last@mail.site.org
dot.at@dot.com

Regular Expressions are Awesome

$$\mathbf{a^+(.a^+)^*@a^+(.a^+)^+}$$



Shorthand Summary

- R^n is shorthand for $RR \ldots R$ (n times).
  - Edge case: define $R^0 = \varepsilon$.
- Σ is shorthand for "any character in Σ."
- $R?$ is shorthand for $(R \cup \varepsilon)$, meaning "zero or one copies of R."
- R^+ is shorthand for RR^*, meaning "one or more copies of R."

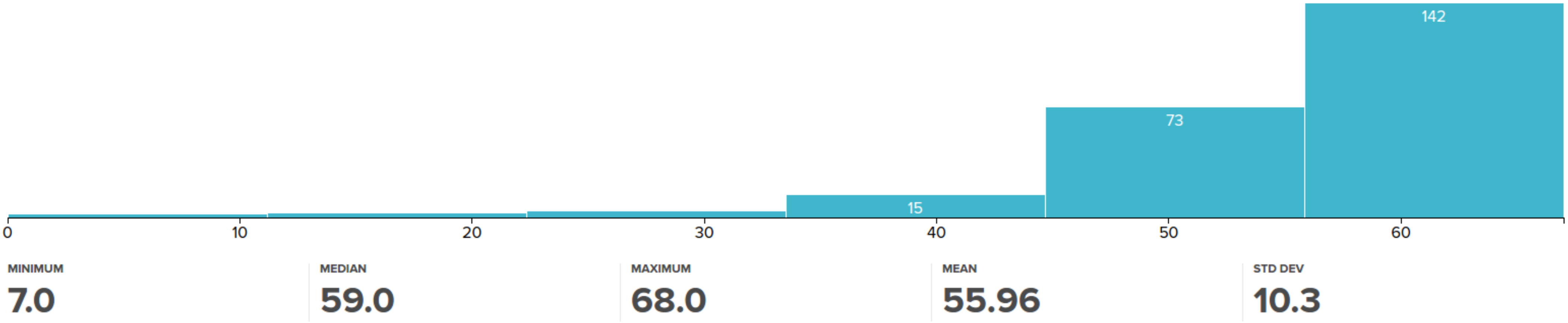
Time-Out for Announcements!

Problem Sets

- Problem Set Five was due at the start of today’s lecture.
  - Need more time? Use your 24-hour late days!
- Problem Set Six goes out today. It’s due next Friday at the start of class.
  - Design DFAs and NFAs for various languages!
  - Explore properties of formal language theory!
  - See applications of the concepts!

Problem Set Four

- Problem Set Four is now graded. Here's the score distribution:



- As always, feel free to stop by office hours to discuss your problem sets or the feedback. We're happy to help out!

Your Questions

“what do the numberings of the handouts mean”

They're numbered sequentially. Anything ending in an S is a solutions set, anything in C is a checkpoint solutions set, and anything in an R is a regrade form.

"Why did you delete a top-ranking question about handout numbers?"

Oops. I didn't mean to do that. Sorry!

“What is Stanford CS doing increase faculty diversity? (in terms of gender, race, class, etc.)”

We've significantly stepped up efforts to address this recently. I'll take this one in class.

Back to CS103!

The Power of Regular Expressions

Theorem: If R is a regular expression, then $\mathscr{L}(R)$ is regular.

Proof idea: Use induction!

- The atomic regular expressions all represent regular languages.
- The combination steps represent closure properties.
- So anything you can make from them must be regular!

Thompson’s Algorithm

- In practice, many regex matchers use an algorithm called ***Thompson's algorithm*** to convert regular expressions into NFAs (and, from there, to DFAs).
  - Read Sipser if you’re curious!
- ***Fun fact:*** the “Thompson” here is Ken Thompson, one of the co-inventors of Unix!

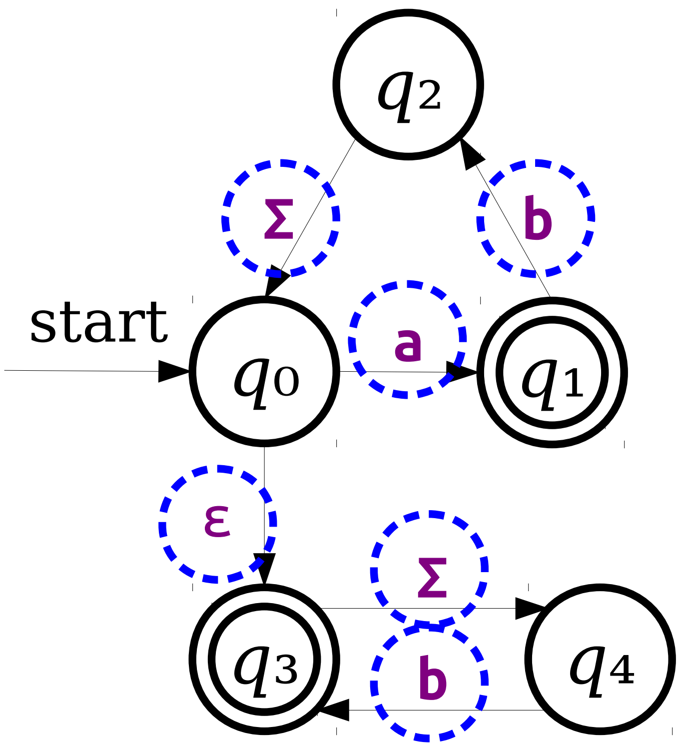
The Power of Regular Expressions

Theorem: If L is a regular language, then there is a regular expression for L.

This is not obvious!

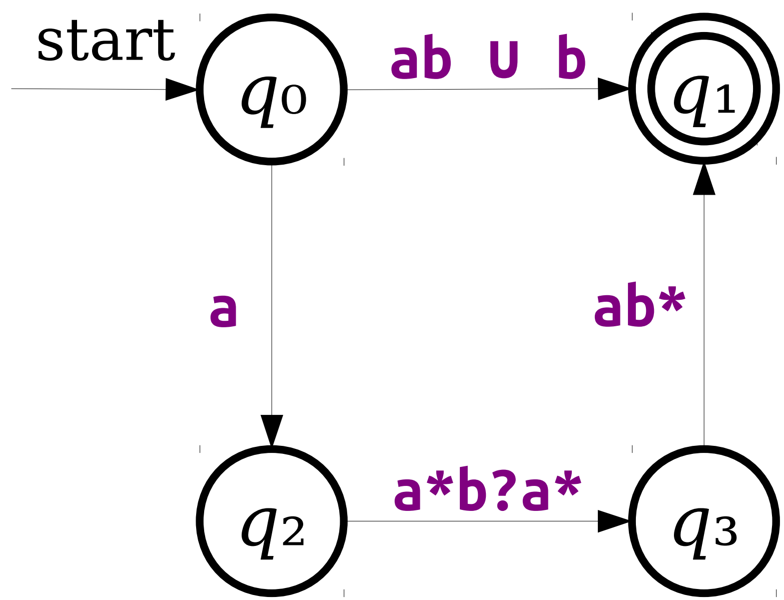
Proof idea: Show how to convert an arbitrary NFA into a regular expression.

Generalizing NFAs



These are all regular expressions!

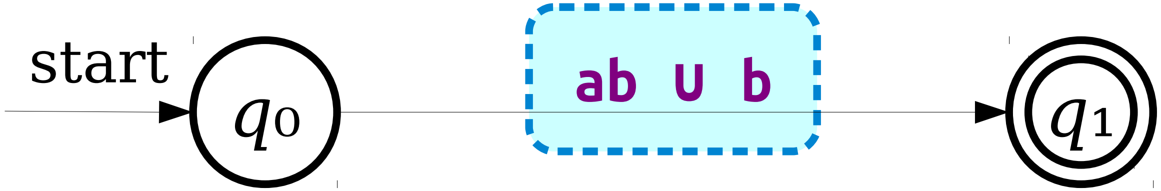
Generalizing NFAs



Note: Actual NFAs aren't allowed to have transitions like these. This is just a thought experiment.

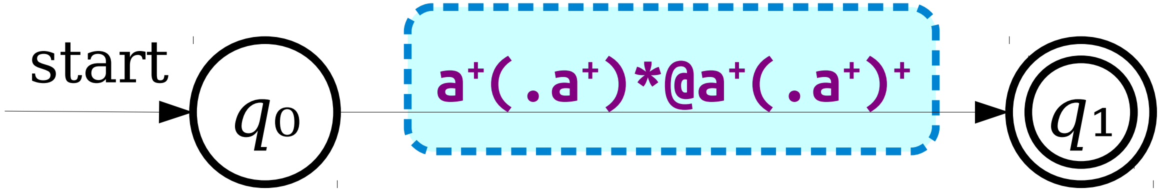
Key Idea 1: Imagine that we can label transitions in an NFA with arbitrary regular expressions.

Generalizing NFAs



Is there a simple regular expression for the language of this generalized NFA?

Generalizing NFAs

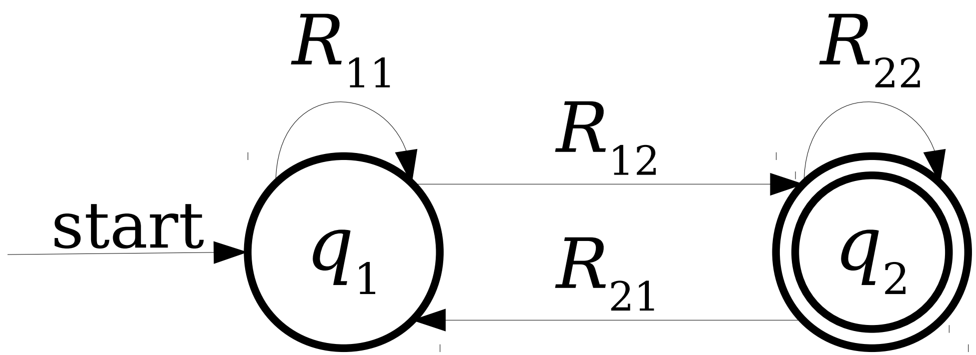


Is there a simple regular expression for the language of this generalized NFA?

Key Idea 2: If we can convert an NFA into a generalized NFA that looks like this...

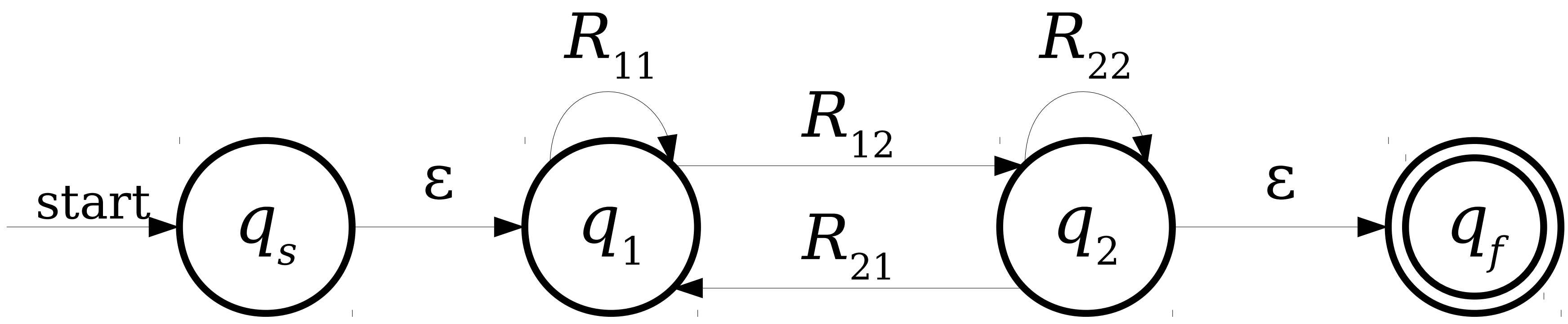
...then we can easily read off a regular expression for that NFA.

From NFAs to Regular Expressions

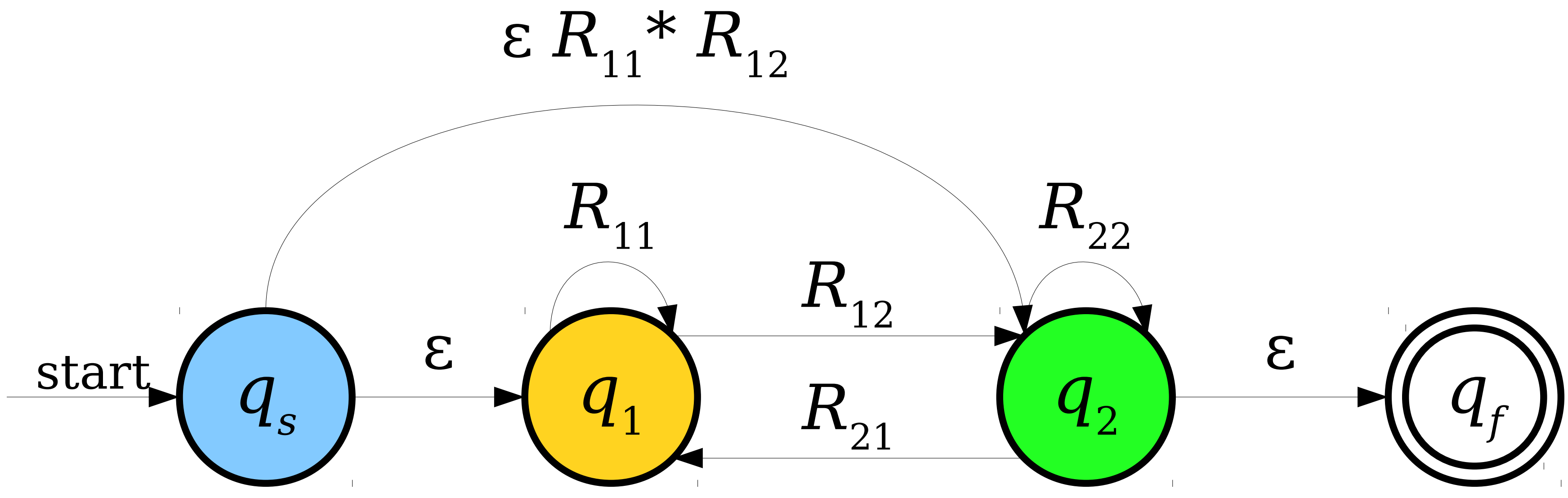


Key Idea 3: Somehow transform this NFA so that it looks like this:

From NFAs to Regular Expressions

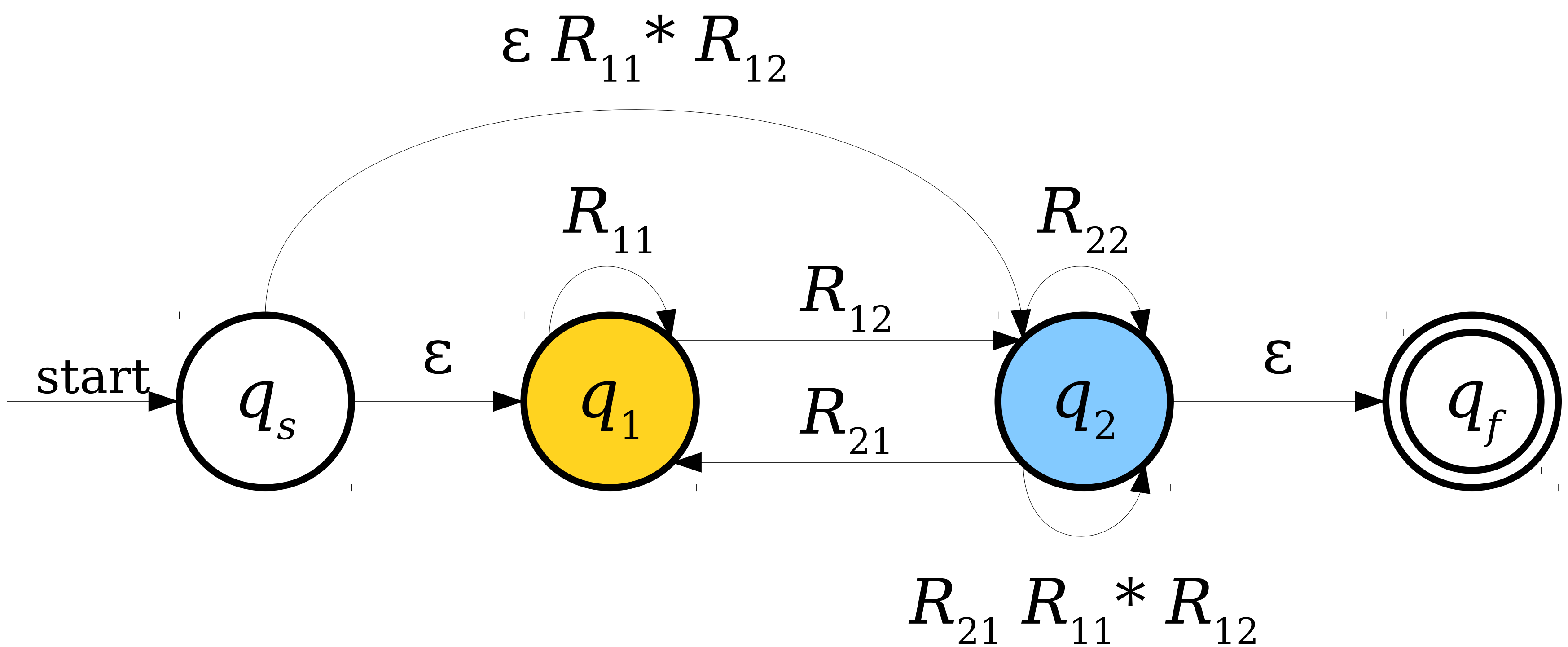


From NFAs to Regular Expressions

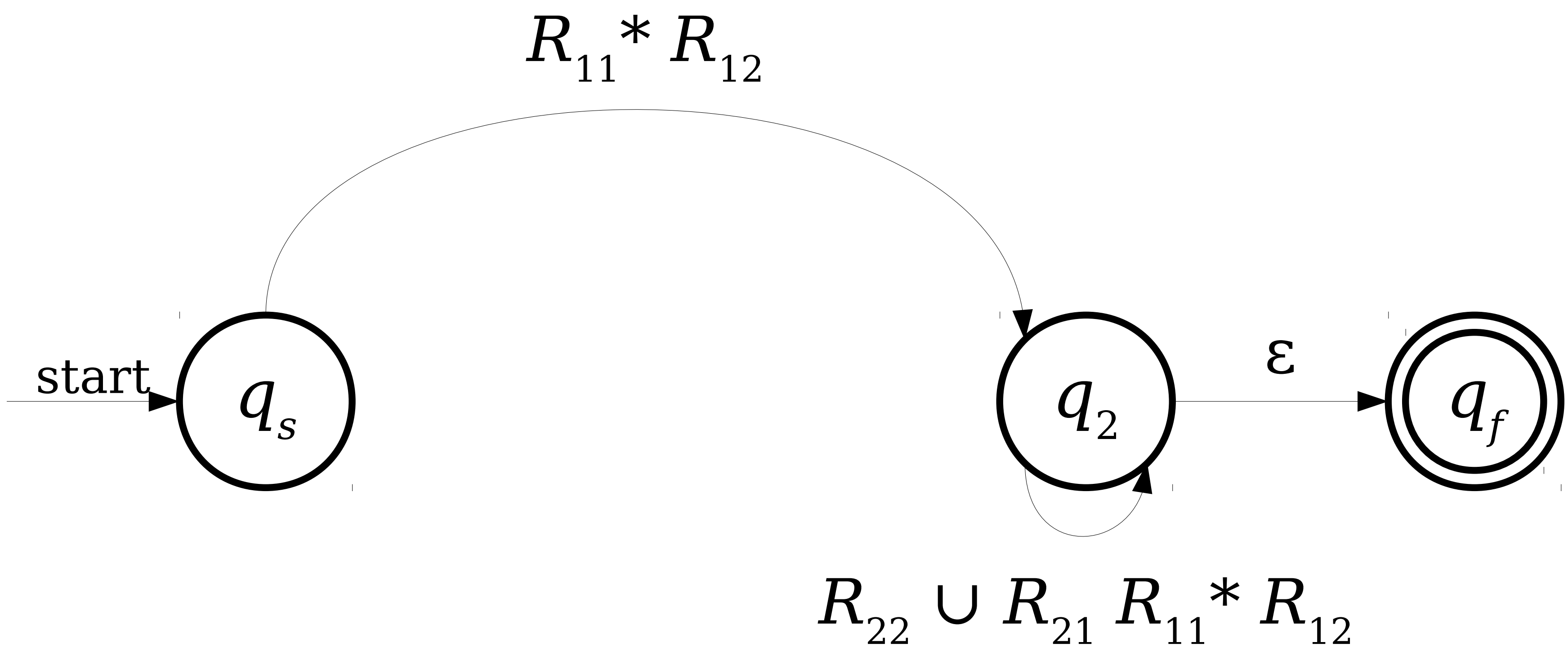


Note: We're using **concatenation** and **Kleene closure** in order to skip this state.

From NFAs to Regular Expressions

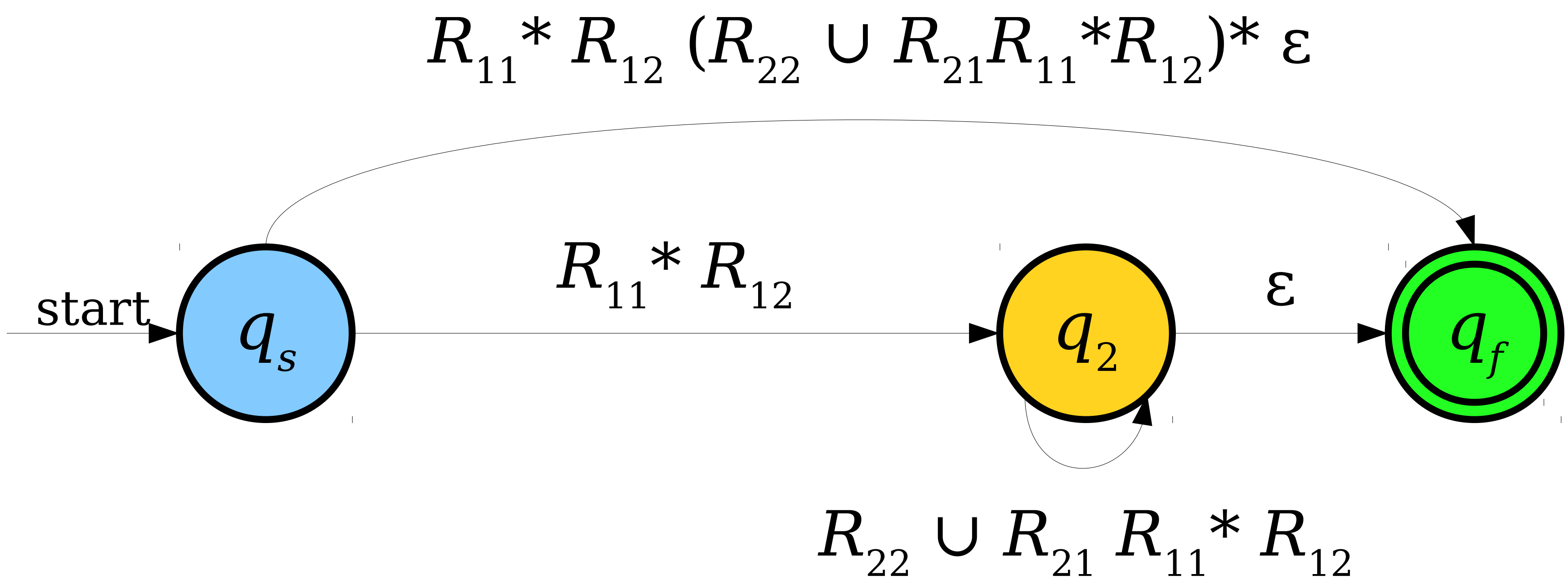


From NFAs to Regular Expressions

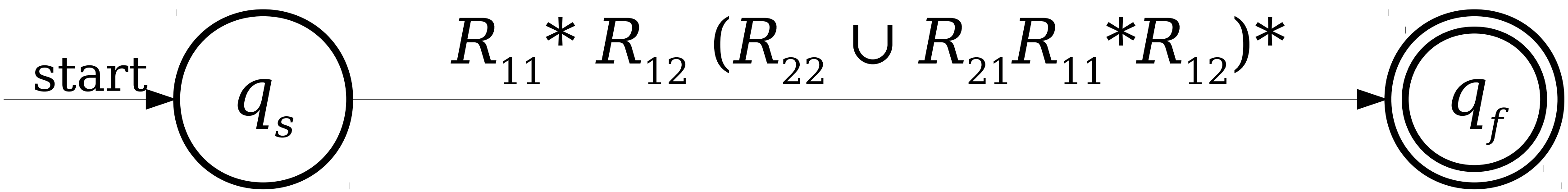


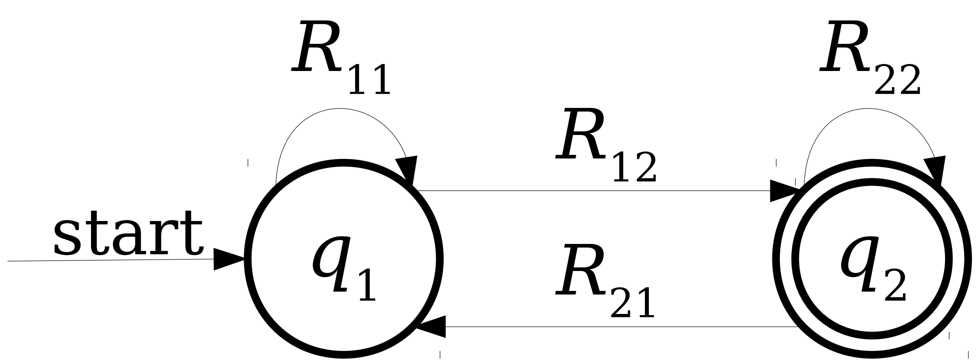
Note: We're using union to combine these transitions together.

From NFAs to Regular Expressions



From NFAs to Regular Expressions





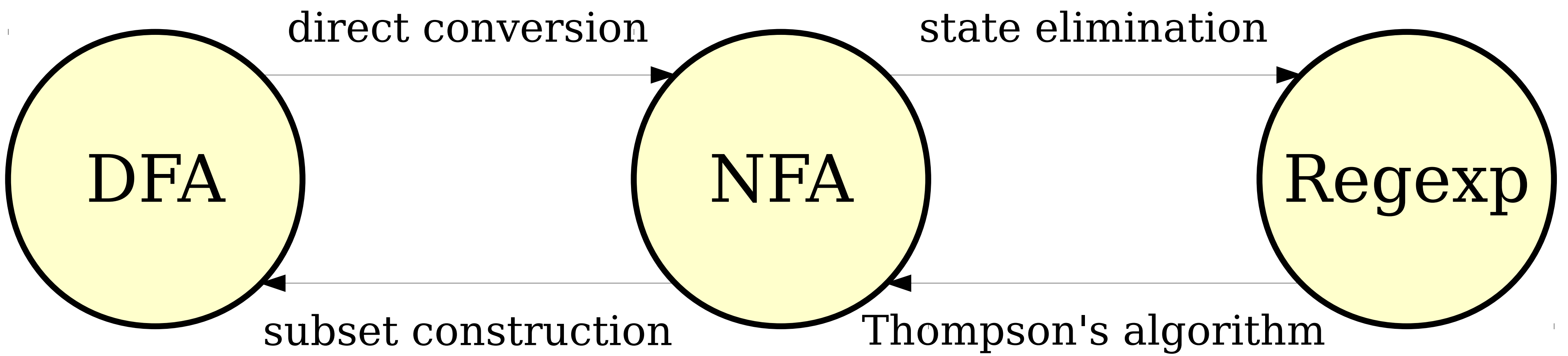
The Construction at a Glance

- Start with an NFA N for the language L.
- Add a new start state q_{s} and accept state q_{f} to the NFA.
  - Add an ε-transition from q_{s} to the old start state of N.
  - Add ε-transitions from each accepting state of N to q_{f}, then mark them as not accepting.
- Repeatedly remove states other than q_{s} and q_{f} from the NFA by "shortcutting" them until only two states remain: q_{s} and q_{f}.
- The transition from q_{s} to q_{f} is then a regular expression for the NFA.

Eliminating a State

- To eliminate a state q from the automaton, do the following for each pair of states q_0 and q_1, where there's a transition from q_0 into q and a transition from q into q_1:
  - Let R_{in} be the regex on the transition from q_0 to q.
  - Let R_{out} be the regex on the transition from q to q_1.
  - If there is a regular expression R_{stay} on a transition from q to itself, add a new transition from q_0 to q_1 labeled $(R_{in}(R_{stay})^*R_{out})$.
  - If there isn't, add a new transition from q_0 to q_1 labeled $(R_{in}R_{out})$
- If a pair of states has multiple transitions between them labeled R_1, R_2, …, R_k, replace them with a single transition labeled $R_1 \cup R_2 \cup \ldots \cup R_k$.

Our Transformations



Theorem: The following are all equivalent:

- L is a regular language.
- There is a DFA D such that $\mathscr{L}(D) = L$.
- There is an NFA N such that $\mathscr{L}(N) = L$.
- There is a regular expression R such that $\mathscr{L}(R) = L$.

Why This Matters

- The equivalence of regular expressions and finite automata has practical relevance.
  - Tools like `grep` and `flex` that use regular expressions capture all the power available via DFAs and NFAs.
- This also is hugely theoretically significant: the regular languages can be assembled “from scratch” using a small number of operations!

Next Time

- ***Applications of Regular Languages***
  - Answering “so what?”
- ***Intuiting Regular Languages***
  - What makes a language regular?
- ***The Myhill-Nerode Theorem***
  - The limits of regular languages.