

# **INTRODUCTION TO AUTOMATA THEORY**

Reading: Chapter 1



# WHAT IS AUTOMATA THEORY?

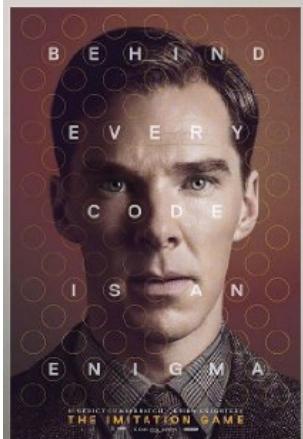
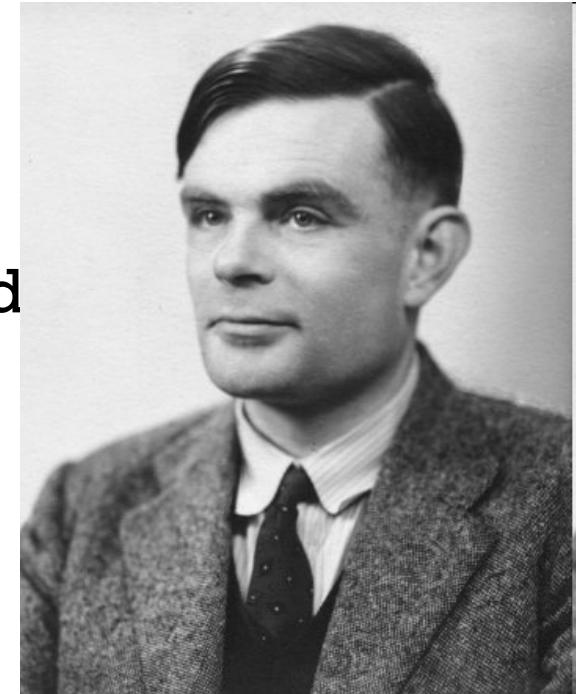
- *Study of abstract computing devices, or “machines”*
- **Automaton = an abstract computing device**
  - Note: A “device” need not even be a physical hardware!
- **A fundamental question in computer science:**
  - Find out what different models of machines can do and cannot do
  - The *theory of computation*
- **Computability vs. Complexity**



(A pioneer of automata theory)

# ALAN TURING (1912-1954)

- Father of Modern Computer Science
- English mathematician
- Studied abstract machines called **Turing machines** even before computers existed
- Heard of the Turing test?



# THEORY OF COMPUTATION: A HISTORICAL PERSPECTIVE

1930s	<ul style="list-style-type: none"><li>• Alan Turing studies <b>Turing machines</b></li><li>• <b>Decidability</b></li><li>• <b>Halting problem</b></li></ul>
1940-1950s	<ul style="list-style-type: none"><li>• “<b>Finite automata</b>” machines studied</li><li>• Noam Chomsky proposes the <b>“Chomsky Hierarchy”</b> for formal languages</li></ul>
1969	Cook introduces “intractable” problems or “ <b>NP-Hard</b> ” problems
1970-	Modern computer science: <b>compilers</b> , <b>computational &amp; complexity theory</b> evolve



# LANGUAGES & GRAMMARS

An **alphabet** is a set of symbols:

{0,1}

Or “**words**”

→ **Sentences** are strings of symbols:

0,1,00,01,10,1,...

A **language** is a set of sentences:

$L = \{000,0100,0010,\dots\}$

A **grammar** is a finite list of rules defining a language.

$S \rightarrow 0A$

$B \rightarrow 1B$

$A \rightarrow 1A$

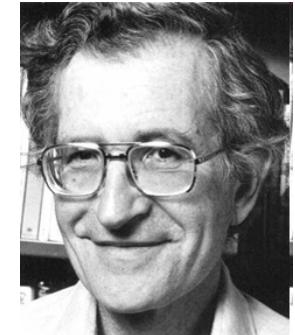
$B \rightarrow 0F$

$A \rightarrow 0B$

$F \rightarrow \epsilon$

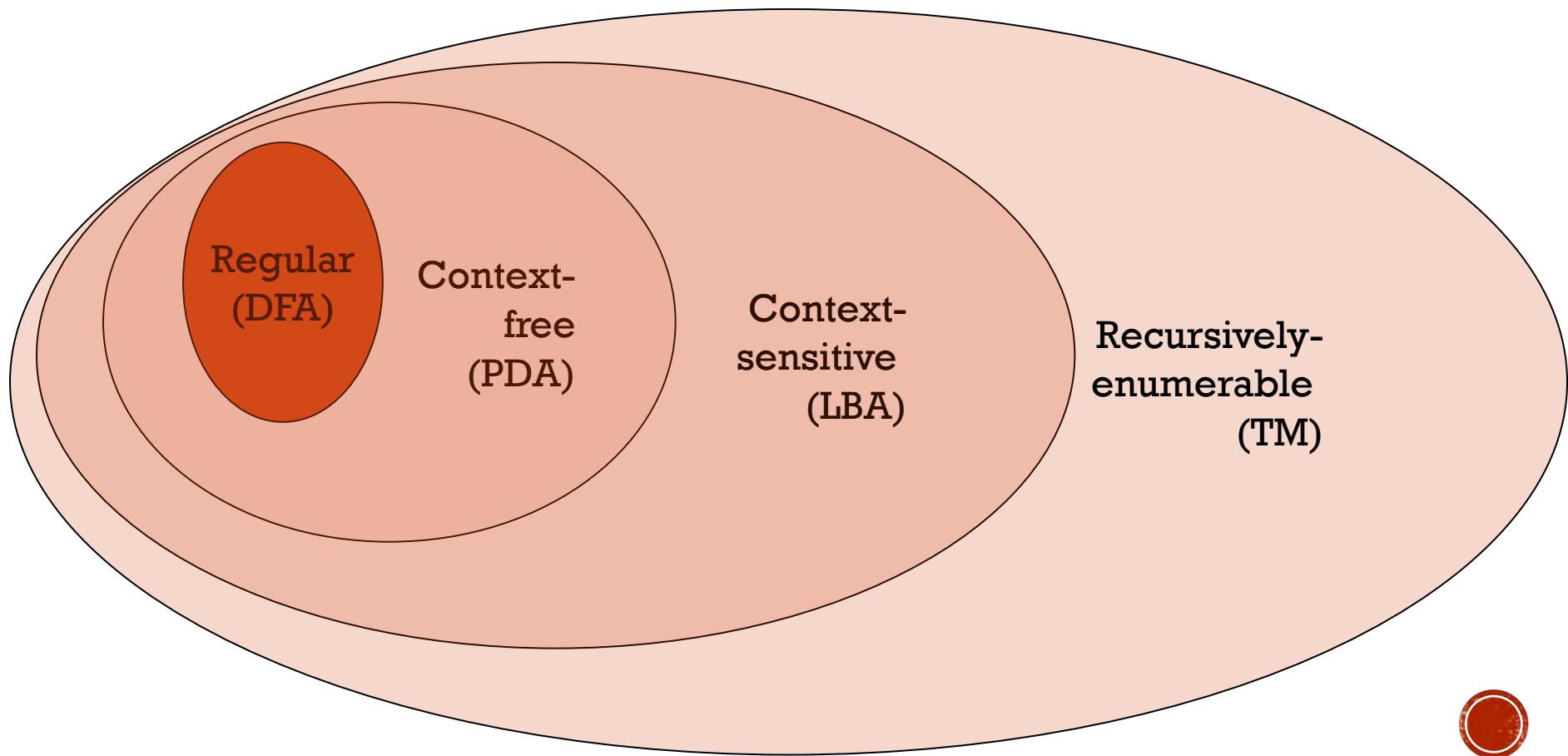
- **Languages:** “*A language is a collection of sentences of finite length all constructed from a finite alphabet of symbols*”
- **Grammars:** “*A grammar can be regarded as a device that enumerates the sentences of a language*” - nothing more, nothing less
- **N. Chomsky, Information and Control, Vol 2, 1959**



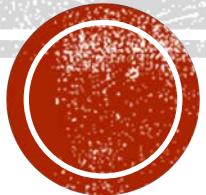


# THE CHOMSKY HIERARCHY

- A containment hierarchy of classes of formal languages



# **THE CENTRAL CONCEPTS OF AUTOMATA THEORY**



# ALPHABET

*An alphabet is a finite, non-empty set of symbols*

- We use the symbol  $\Sigma$  (sigma) to denote an alphabet
- Examples:
  - Binary:  $\Sigma = \{0, 1\}$
  - All lower case letters:  $\Sigma = \{a, b, c, \dots, z\}$
  - Alphanumeric:  $\Sigma = \{a-z, A-Z, 0-9\}$
  - DNA molecule letters:  $\Sigma = \{a, c, g, t\}$
  - ...



# STRINGS

*A string or word is a finite sequence of symbols chosen from  $\Sigma$*

- **Empty string is  $\varepsilon$  (or “epsilon”)**
- Length of a string  $w$ , denoted by “ $|w|$ ”, is equal to the *number of (non-  $\varepsilon$ ) characters in the string*
  - E.g.,  $x = 010100$   $|x| = 6$
  - $x = 01 \varepsilon 0 \varepsilon 1 \varepsilon 00 \varepsilon$   $|x| = ?$
- $xy$  = concatenation of two strings  $x$  and  $y$



# POWERS OF AN ALPHABET

Let  $\Sigma$  be an alphabet.

- $\Sigma^k$  = the set of all strings of length  $k$
- $\Sigma^* = \Sigma^0 \cup \Sigma^1 \cup \Sigma^2 \cup \dots$
- $\Sigma^+ = \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \dots$



# LANGUAGES

*L is said to be a language over alphabet  $\Sigma$ , only if  $L \subseteq \Sigma^*$*

→ this is because  $\Sigma^*$  is the set of all strings (of all possible length including 0) over the given alphabet  $\Sigma$

Examples:

1. Let L be the language of all strings consisting of n 0's followed by n 1's:

$$L = \{\epsilon, 01, 0011, 000111, \dots\}$$

2. Let L be the language of all strings of with equal number of 0's and 1's:

$$L = \{\epsilon, 01, 10, 0011, 1100, 0101, 1010, 1001, \dots\}$$

→ Canonical ordering of strings in the language

**Definition:**  $\emptyset$  denotes the Empty language

- Let  $L = \{\epsilon\}$ ; Is  $L = \emptyset$ ?

NO



# THE MEMBERSHIP PROBLEM

*Given a string  $w \in \Sigma^*$  and a language  $L$  over  $\Sigma$ , decide whether or not  $w \in L$ .*

Example:

Let  $w = 100011$

Q) Is  $w \in$  the language of strings with equal number of 0s and 1s?



# **FINITE AUTOMATA**

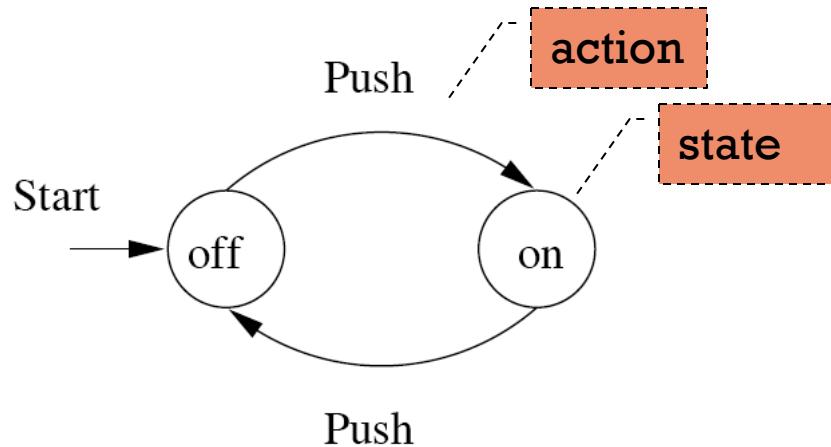
## **■ Some Applications**

- Software for designing and checking the behavior of digital circuits
- Lexical analyzer of a typical compiler
- Software for scanning large bodies of text (e.g., web pages) for pattern finding
- Software for verifying systems of all types that have a finite number of states (e.g., stock market transaction, communication/network protocol)

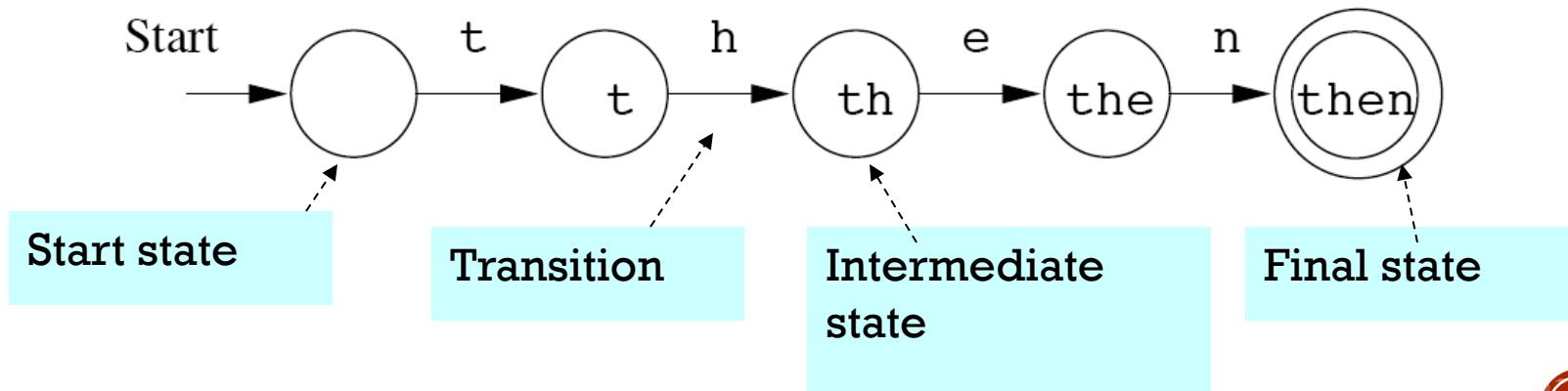


# FINITE AUTOMATA : EXAMPLES

- On/Off switch



- Modeling recognition of the word “then”



# STRUCTURAL EXPRESSIONS

- Grammars
- Regular expressions
  - E.g., unix style to capture city names such as “Palo Alto CA”:

▪  $[A-Z][a-z]^*([ ][A-Z][a-z]^*)^*[ ][A-Z][A-Z]$

Start with a letter

A string of other letters (possibly empty)

Should end w/ 2-letter state code

Other space delimited words  
(part of city name)



# FORMAL PROOFS



# DEDUCTIVE PROOFS

*From the given statement(s) to a conclusion statement (what we want to prove)*

- Logical progression by direct implications

Example for parsing a statement:

- “If  $y \geq 4$ , then  $2^y \geq y^2$ .”

*given*

*conclusion*

(there are other ways of writing this).



# EXAMPLE: DEDUCTIVE PROOF

Let Claim 1: If  $y \geq 4$ , then  $2^y \geq y^2$ .

Let  $x$  be any number which is obtained by adding the squares of 4 positive integers.

Claim 2:

Given  $x$  and assuming that **Claim 1** is true, prove that  $2^x \geq x^2$

- Proof:
  - 1) Given:  $x = a^2 + b^2 + c^2 + d^2$
  - 2) Given:  $a \geq 1, b \geq 1, c \geq 1, d \geq 1$
  - 3)  $\rightarrow a^2 \geq 1, b^2 \geq 1, c^2 \geq 1, d^2 \geq 1$  (by 2)
  - 4)  $\rightarrow x \geq 4$  (by 1 & 3)
  - 5)  $\rightarrow 2^x \geq x^2$  (by 4 and Claim 1)

“implies” or “follows”



# ON THEOREMS, LEMMAS AND COROLLARIES

We typically refer to:

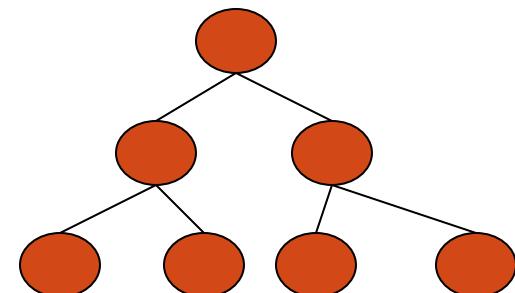
- A major result as a “**theorem**”
  - An intermediate result that we show to prove a larger result as a “**lemma**”
  - A result that follows from an already proven result as a “**corollary**”
- 

An example:

**Theorem:** *The height of an  $n$ -node binary tree is at least  $\text{floor}(\lg n)$*

**Lemma:** *Level  $i$  of a perfect binary tree has  $2^i$  nodes.*

**Corollary:** *A perfect binary tree of height  $h$  has  $2^{h+1}-1$  nodes.*



# QUANTIFIERS

*“For all” or “For every”*

- Universal proofs
- Notation=

$\forall$

*“There exists”*

- Used in existential proofs
- Notation=

$\exists$

Implication is denoted by =>

- E.g., “IF A THEN B” can also be written as “ $A \Rightarrow B$ ”



# PROVING TECHNIQUES

- **By contradiction**

- Start with the statement contradictory to the given statement
- E.g., To prove  $(A \Rightarrow B)$ , we start with:
  - $(A \text{ and } \sim B)$
  - ... and then show that could never happen

What if you want to prove that " $(A \text{ and } B \Rightarrow C \text{ or } D)$ "?

- **By induction**

- (3 steps) Basis, inductive hypothesis, inductive step

- **By contrapositive statement**

- If  $A$  then  $B$     $\equiv$     If  $\sim B$  then  $\sim A$



# PROVING TECHNIQUES...

- **By counter-example**
  - Show an example that disproves the claim
- **Note: There is no such thing called a “proof by example”!**
  - So when asked to prove a claim, an example that satisfied that claim is *not* a proof



# DIFFERENT WAYS OF SAYING THE SAME THING

- “If H then C”:
  - i. H implies C
  - ii.  $H \Rightarrow C$
  - iii. C if H
  - iv. H only if C
  - v. Whenever H holds, C follows



# **“IF-AND-ONLY-IF” STATEMENTS**

- “A if and only if B”      ( $A \iff B$ )
  - (*if part*) if B then A      ( $\leq$ )
  - (*only if part*) A only if B    ( $\geq$ )  
(same as “if A then B”)
- “If and only if” is abbreviated as “iff”
  - i.e., “A iff B”
- Example:
  - Theorem: *Let x be a real number. Then floor of x = ceiling of x if and only if x is an integer.*
- Proofs for iff have two parts
  - One for the “if part” & another for the “only if part”



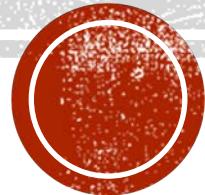
# SUMMARY

- Automata theory & a historical perspective
- Chomsky hierarchy
- Finite automata
- Alphabets, strings/words/sentences, languages
- Membership problem
- Proofs:
  - Deductive, induction, contrapositive, contradiction, counterexample
  - If and only if
- Read chapter 1 for more examples and exercises



# **FINITE AUTOMATA**

Reading: Chapter 2



# FINITE AUTOMATON (FA)

- Informally, a state diagram that comprehensively captures all possible states and transitions that a machine can take while responding to a stream or sequence of input symbols
  - Recognizer for “Regular Languages”
- 
- Deterministic Finite Automata (DFA)
    - The machine can exist in only one state at any given time
  - Non-deterministic Finite Automata (NFA)
    - The machine can exist in multiple states at the same time

# DETERMINISTIC FINITE AUTOMATA - DEFINITION

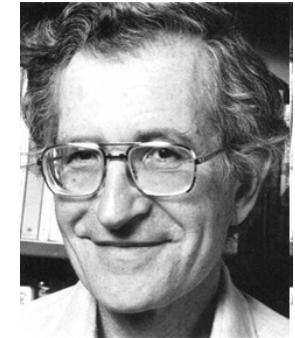
- A Deterministic Finite Automaton (DFA) consists of:
  - $Q \implies$  a finite set of states
  - $\Sigma \implies$  a finite set of input symbols (alphabet)
  - $q_0 \implies$  a start state
  - $F \implies$  set of accepting states
  - $\delta \implies$  a transition function, which is a mapping between  $Q \times \Sigma \implies Q$
- A DFA is defined by the 5-tuple:
  - $\{Q, \Sigma, q_0, F, \delta\}$

# WHAT DOES A DFA DO ON READING AN INPUT STRING?

- Input: a word  $w$  in  $\Sigma^*$
- Question: Is  $w$  acceptable by the DFA?
- Steps:
  - Start at the “start state”  $q_0$
  - For every input symbol in the sequence  $w$  do
    - Compute the next state from the current state, given the current input symbol in  $w$  and the transition function
  - If after all symbols in  $w$  are consumed, the current state is one of the accepting states ( $F$ ) then accept  $w$ ;
  - Otherwise, *reject*  $w$ .

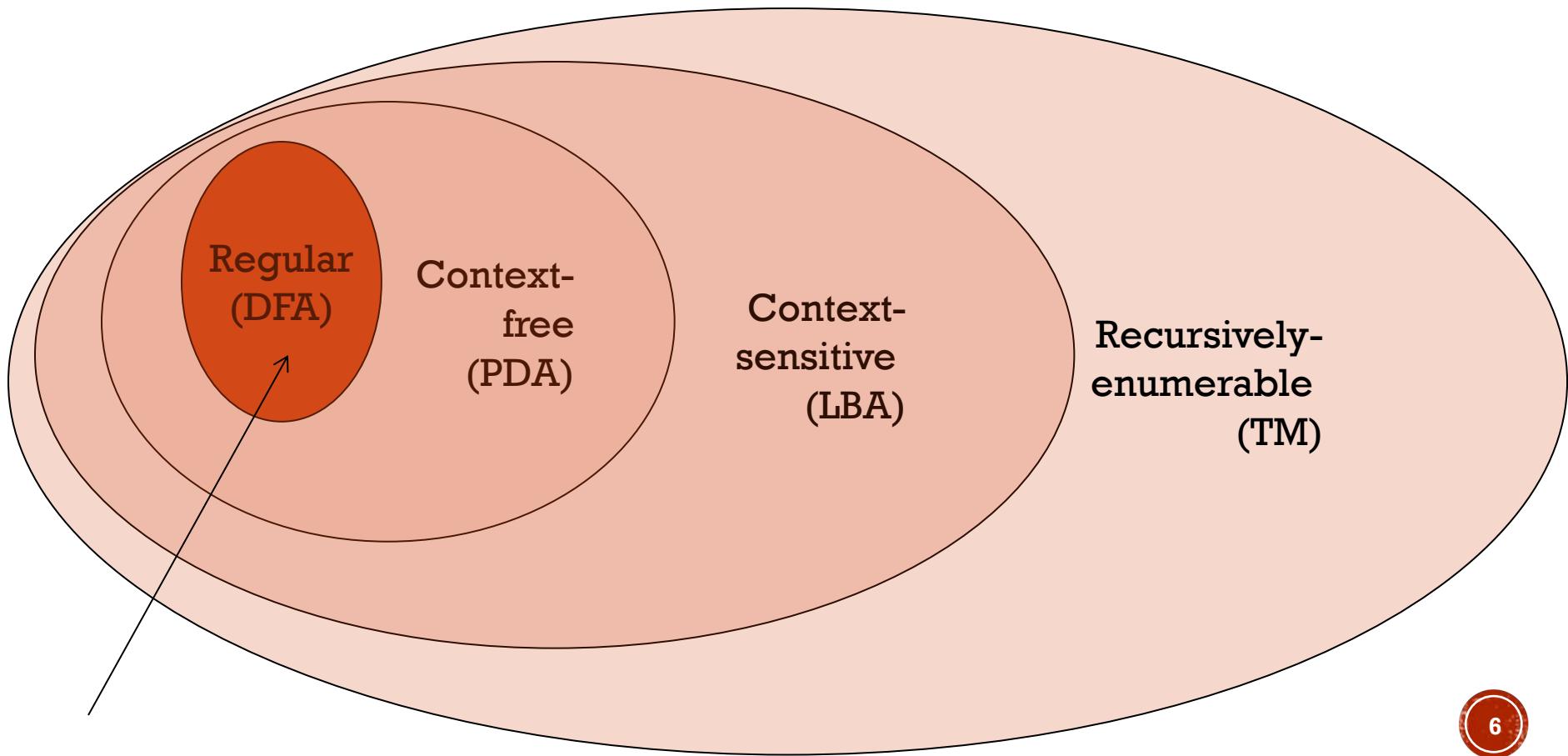
# REGULAR LANGUAGES

- Let  $L(A)$  be a language *recognized* by a DFA A.
  - Then  $L(A)$  is called a “*Regular Language*”.
- Locate regular languages in the Chomsky Hierarchy



# THE CHOMSKY HIERARCHY

- A containment hierarchy of classes of formal languages



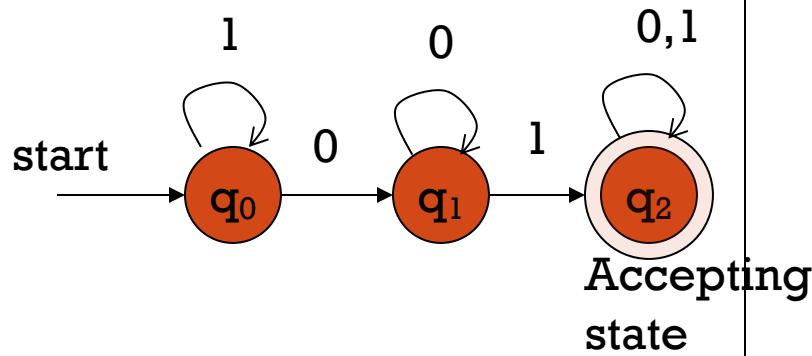
# EXAMPLE #1

- Build a DFA for the following language:
  - $L = \{w \mid w \text{ is a binary string that contains } 01 \text{ as a substring}\}$
- Steps for building a DFA to recognize L:
  - $\Sigma = \{0,1\}$
  - Decide on the states: Q
  - Designate start state and final state(s)
  - $\delta$ : Decide on the transitions:
- “Final” states == same as “accepting states”
- Other states == same as “non-accepting states”

Regular expression:  $(0+1)^*01(0+1)^*$

# DFA FOR STRINGS CONTAINING 01

- What makes this DFA deterministic?



- What if the language allows empty strings?

- $Q = \{q_0, q_1, q_2\}$

- $\Sigma = \{0, 1\}$

- start state =  $q_0$

- $F = \{q_2\}$

- Transition table

$\delta$	0	1
states	$q_0$	$q_1$
$q_0$	$q_1$	$q_0$
$q_1$	$q_1$	$q_2$
$*q_2$	$q_2$	$q_2$

# EXAMPLE #2

## Clamping Logic:

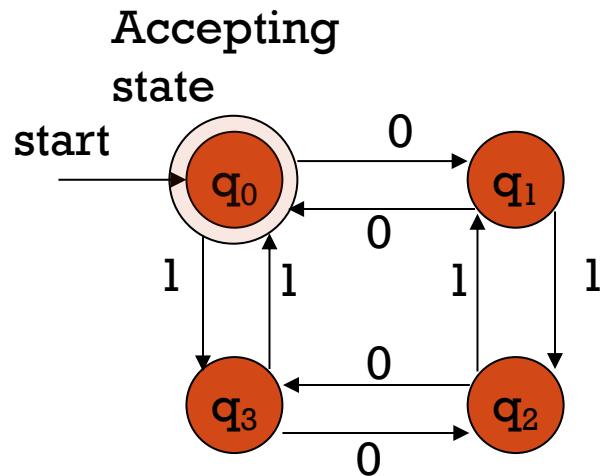
- A clamping circuit waits for a "1" input, and turns on forever. However, to avoid clamping on spurious noise, we'll design a DFA that waits for *two consecutive 1s* in a row before clamping on.
- Build a DFA for the following language:  
 $L = \{ w \mid w \text{ is a bit string which contains the substring } 11\}$

## ▪ State Design:

- $q_0$ : start state (initially off), also means the most recent input was not a 1
- $q_1$ : has never seen 11 but the most recent input was a 1
- $q_2$ : has seen 11 at least once

# EXAMPLE #3

- Build a DFA for the following language:  
 $L = \{ w \mid w \text{ is a binary string that has even number of 1s and even number of 0s}\}$



# EXTENSION OF TRANSITIONS ( $\Delta$ ) TO PATHS ( $\hat{\Delta}$ )

- $\hat{\delta}(q, w) = \text{destination state from state } q \text{ on input } \underline{\text{string}} \text{ } w$
- $\hat{\delta}(q, wa) = \hat{\delta}(\hat{\delta}(q, w), a)$
- Work out example #3 using the input sequence  $w=10010$ ,  $a=1$ :
  - $\hat{\delta}(q_0, wa) = ?$

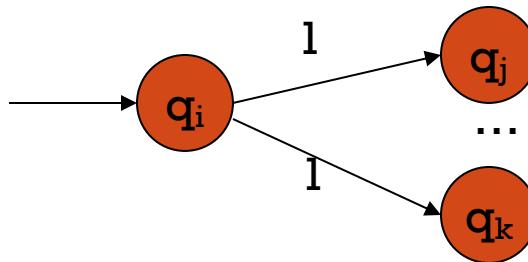
# LANGUAGE OF A DFA

A DFA  $A$  accepts string  $w$  if there is a path from  $q_0$  to an accepting (or final) state that is labeled by  $w$

- i.e.,  $L(A) = \{ w \mid \hat{\delta}(q_0, w) \in F \}$
- I.e.,  $L(A)$  = all strings that lead to an accepting state from  $q_0$

# NON-DETERMINISTIC FINITE AUTOMATA (NFA)

- A Non-deterministic Finite Automaton (NFA)
  - is of course “non-deterministic”
    - Implying that the machine can exist in more than one state at the same time
    - Transitions could be non-deterministic



- Each transition function therefore maps to a set of states

# NON-DETERMINISTIC FINITE AUTOMATA (NFA)

- A Non-deterministic Finite Automaton (NFA) consists of:
  - $Q \implies$  a finite set of states
  - $\Sigma \implies$  a finite set of input symbols (alphabet)
  - $q_0 \implies$  a start state
  - $F \implies$  set of accepting states
  - $\delta \implies$  a transition function, which is a mapping between  $Q \times \Sigma \implies$  subset of  $Q$
- An NFA is also defined by the 5-tuple:
  - $\{Q, \Sigma, q_0, F, \delta\}$

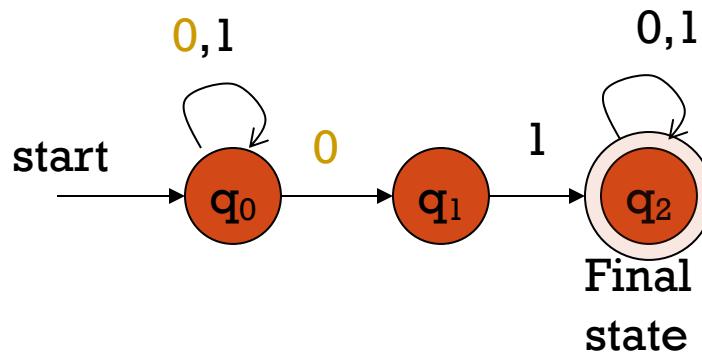
# HOW TO USE AN NFA?

- Input: a word  $w$  in  $\Sigma^*$
- Question: Is  $w$  acceptable by the NFA?
- Steps:
  - Start at the “start state”  $q_0$
  - For every input symbol in the sequence  $w$  do
    - Determine **all possible next states from all current states**, given the current input symbol in  $w$  and the transition function
  - If after all symbols in  $w$  are consumed and if at least **one** of the current states is a final state then *accept w*;
  - Otherwise, *reject w*.

Regular expression:  $(0+1)^*01(0+1)^*$

# NFA FOR STRINGS CONTAINING 01

Why is this non-deterministic?



What will happen if at state  $q_1$  an input of 0 is received?

- $Q = \{q_0, q_1, q_2\}$

- $\Sigma = \{0, 1\}$

- $\text{start state} = q_0$

- $F = \{q_2\}$

- Transition table

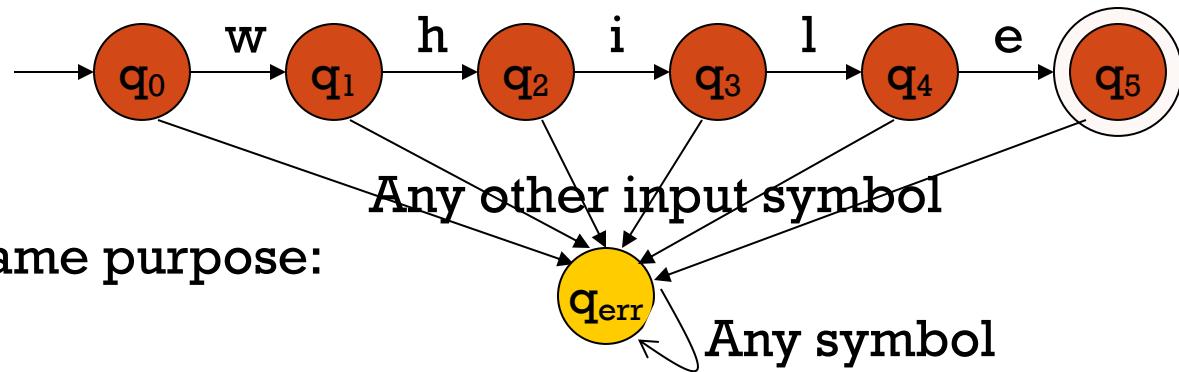
symbols

$\delta$	0	1
states		
$q_0$	$\{q_0, q_1\}$	$\{q_0\}$
$q_1$	$\Phi$	$\{q_2\}$
$*q_2$	$\{q_2\}$	$\{q_2\}$

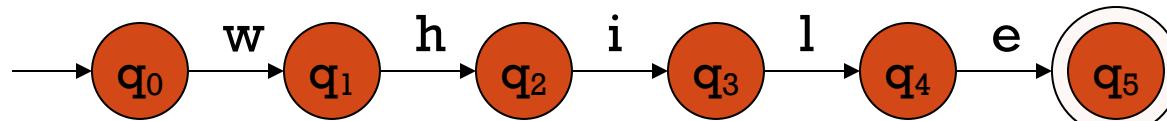
Note: Omitting to explicitly show error states is just a matter of design convenience (one that is generally followed for NFAs), and i.e., this feature should not be confused with the notion of non-determinism.

# WHAT IS AN ‘ERROR STATE’?

- A DFA for recognizing the key word “while”



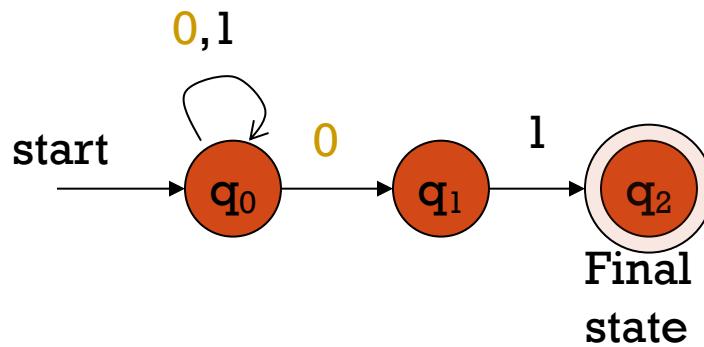
- An NFA for the same purpose:



*Transitions into a dead state are implicit*

# EXAMPLES

- Build an NFA for the following language:  
 $L = \{ w \mid w \text{ ends in } 01\}$



- Other examples
  - Keyword recognizer (e.g., if, then, else, while, for, include, etc.)
  - Strings where the first symbol is present somewhere later on at least once

# EXTENSION OF $\Delta$ TO NFA PATHS

- Basis:  $\hat{\delta}(q, \varepsilon) = \{q\}$
- Induction:
  - Let  $\hat{\delta}(q_0, w) = \{p_1, p_2, \dots, p_k\}$
  - $\delta(p_i, a) = S_i \quad \text{for } i=1, 2, \dots, k$
  - Then,  $\hat{\delta}(q_0, wa) = S_1 U S_2 U \dots U S_k$

# LANGUAGE OF AN NFA

- An NFA accepts  $w$  if *there exists at least one* path from the start state to an accepting (or final) state that is labeled by  $w$
- $L(N) = \{ w \mid \stackrel{\wedge}{\delta}(q_0, w) \cap F \neq \emptyset \}$

# ADVANTAGES & CAVEATS FOR NFA

- Great for modeling regular expressions
  - String processing - e.g., grep, lexical analyzer
- Could a non-deterministic state machine be implemented in practice?
  - Probabilistic models could be viewed as extensions of non-deterministic state machines  
(e.g., toss of a coin, a roll of dice)
    - They are not the same though
  - A parallel computer could exist in multiple “states” at the same time

# TECHNOLOGIES FOR NFAS

- Micron's Automata Processor (introduced in 2013)
- 2D array of MISD (multiple instruction single data) fabric w/ thousands to millions of processing elements.
- 1 input symbol = fed to all states (i.e., cores)
- Non-determinism using circuits
- <http://www.micronautomata.com/>



But, DFAs and NFAs are equivalent in their power to capture languages !!

# DIFFERENCES: DFA VS. NFA

- **DFA**

1. All transitions are deterministic
  - Each transition leads to exactly one state
2. For each state, transition on all possible symbols (alphabet) should be defined
3. Accepts input if the last state visited is in F
4. Sometimes harder to construct because of the number of states
5. Practical implementation is feasible

- **NFA**

1. Some transitions could be non-deterministic
  - A transition could lead to a subset of states
2. Not all symbol transitions need to be defined explicitly (if undefined will go to an error state – this is just a design convenience, not to be confused with “non-determinism”)
3. Accepts input if one of the last states is in F
4. Generally easier than a DFA to construct
5. Practical implementations limited but emerging (e.g., Micron automata processor)

# EQUIVALENCE OF DFA & NFA

- Theorem:
  - A language  $L$  is accepted by a DFA if and only if it is accepted by an NFA.

Should be  
true for  
any  $L$

## Proof:

1. If part:
  - Prove by showing every NFA can be converted to an equivalent DFA (in the next few slides...)
2. Only-if part is trivial:
  - Every DFA is a special case of an NFA where each state has exactly one transition for every input symbol. Therefore, if  $L$  is accepted by a DFA, it is accepted by a corresponding NFA.

# PROOF FOR THE IF-PART

- If-part: A language  $L$  is accepted by a DFA if it is accepted by an NFA
  - rephrasing...
  - Given any NFA  $N$ , we can construct a DFA  $D$  such that  $L(N)=L(D)$
- 
- How to convert an NFA into a DFA?
    - Observation: In an NFA, each transition maps to a *subset* of states
    - Idea: Represent:  
each “subset of NFA\_states” → a single “DFA\_state”

Subset construction

# NFA TO DFA BY SUBSET CONSTRUCTION

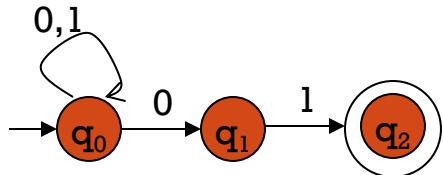
- Let  $N = \{Q_N, \Sigma, \delta_N, q_0, F_N\}$
- Goal: Build  $D = \{Q_D, \Sigma, \delta_D, \{q_0\}, F_D\}$  s.t.  
 $L(D) = L(N)$
- Construction:
  1.  $Q_D$  = all subsets of  $Q_N$  (i.e., power set)
  2.  $F_D$  = set of subsets  $S$  of  $Q_N$  s.t.  $S \cap F_N \neq \emptyset$
  3.  $\delta_D$ : for each subset  $S$  of  $Q_N$  and for each input symbol  $a$  in  $\Sigma$ :
    - $\delta_D(S, a) = \bigcup_{p \text{ in } S} \delta_N(p, a)$

Idea: To avoid enumerating all of power set, do “lazy creation of states”

# NFA TO DFA CONSTRUCTION: EXAMPLE

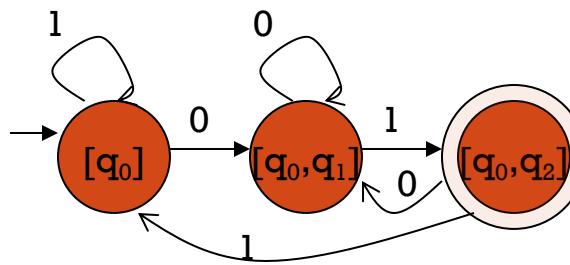
- $L = \{w \mid w \text{ ends in } 01\}$

NFA:



$\delta_N$	0	1
$q_0$	{ $q_0, q_1$ }	{ $q_0$ }
$q_1$	$\emptyset$	{ $q_2$ }
$*q_2$	$\emptyset$	$\emptyset$

DFA:



$\delta_D$	0	1
$\emptyset$	$\emptyset$	$\emptyset$
$[q_0]$	{ $q_0, q_1$ }	{ $q_0$ }
$[q_1]$	$\emptyset$	{ $q_2$ }
$*[q_2]$	$\emptyset$	$\emptyset$
$[q_0, q_1]$	{ $q_0, q_1$ }	{ $q_0, q_2$ }
$*[q_0, q_2]$	{ $q_0, q_1$ }	{ $q_0$ }
$*[q_1, q_2]$	$\emptyset$	{ $q_2$ }
$*[q_0, q_1, q_2]$	{ $q_0, q_1$ }	{ $q_0, q_2$ }

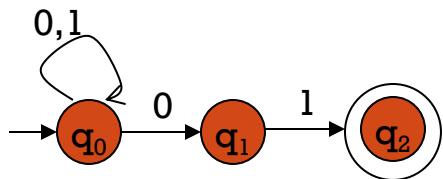
$\delta_D$	0	1
$[q_0]$	[ $q_0, q_1$ ]	[ $q_0$ ]
$[q_0, q_1]$	[ $q_0, q_1$ ]	[ $q_0, q_2$ ]
$*[q_0, q_2]$	[ $q_0, q_1$ ]	[ $q_0$ ]

0. Enumerate all possible subsets
1. Determine transitions
2. Retain only those states  
reachable from { $q_0$ }

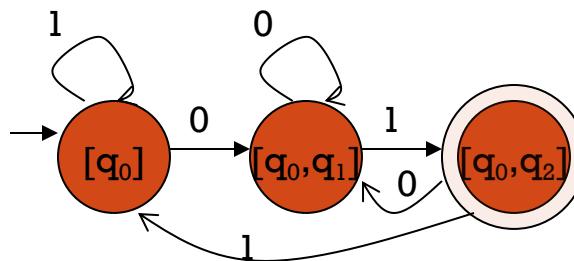
# NFA TO DFA: REPEATING THE EXAMPLE USING *LAZY CREATION*

- $L = \{w \mid w \text{ ends in } 01\}$

NFA:



DFA:



$\delta_N$	0	1
$q_0$	$\{q_0, q_1\}$	$\{q_0\}$
$q_1$	$\emptyset$	$\{q_2\}$
$*q_2$	$\emptyset$	$\emptyset$

$\delta_D$	0	1
$[q_0]$	$[q_0]$	$[q_0]$

Main Idea:

Introduce states as you go  
(on a need basis)

# CORRECTNESS OF SUBSET CONSTRUCTION

Theorem: If  $D$  is the DFA constructed from NFA  $N$  by subset construction, then  $L(D) = L(N)$

- Proof:

- Show that  $\widehat{\delta}_D(\{q_0\}, w) \equiv \widehat{\delta}_N(q_0, w)$ , for all  $w$
- Using induction on  $w$ 's length:
  - Let  $w = xa$
  - $\widehat{\delta}_D(\{q_0\}, xa) \equiv \widehat{\delta}_D(\widehat{\delta}_N(q_0, x), a) \equiv \widehat{\delta}_N(q_0, w)$

# A BAD CASE WHERE $\# \text{STATES}(\text{DFA}) >> \# \text{STATES}(\text{NFA})$

- $L = \{w \mid w \text{ is a binary string s.t., the } k^{\text{th}} \text{ symbol from its end is a } 1\}$
- NFA has  $k+1$  states
- But an equivalent DFA needs to have at least  $2^k$  states

## (Pigeon hole principle)

- $m$  holes and  $>m$  pigeons
  - $\Rightarrow$  at least one hole has to contain two or more pigeons

# APPLICATIONS

- **Text indexing**
  - inverted indexing
  - For each unique word in the database, store all locations that contain it using an NFA or a DFA
- **Find pattern P in text T**
  - Example: Google querying
- **Extensions of this idea:**
  - PATRICIA tree, suffix tree

# A FEW SUBTLE PROPERTIES OF DFAS AND NFAS

- The machine never really terminates.
  - It is always waiting for the next input symbol or making transitions.
- The machine decides when to consume the next symbol from the input and when to ignore it.
  - (but the machine can never skip a symbol)
- => A transition can happen even *without* really consuming an input symbol (think of consuming  $\epsilon$  as a free token) – if this happens, then it becomes an  $\epsilon$ -NFA (see next few slides).
- A single transition *cannot consume more than one (non- $\epsilon$ ) symbol.*

# FA WITH $\epsilon$ -TRANSITIONS

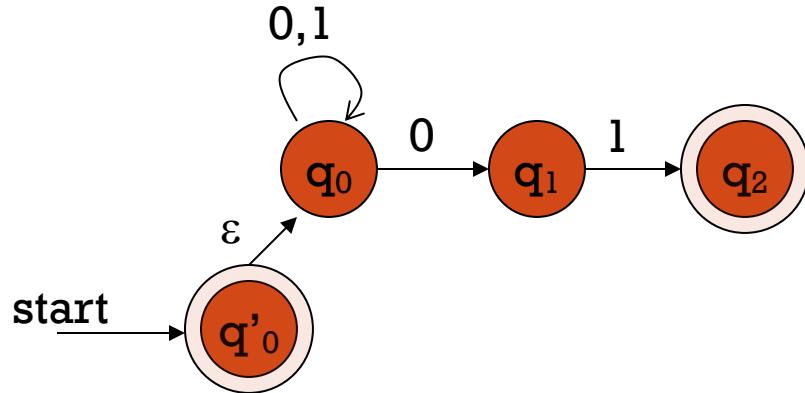
- We can allow explicit  $\epsilon$ -transitions in finite automata
  - i.e., a transition from one state to another state without consuming any additional input symbol
  - Explicit  $\epsilon$ -transitions between different states introduce non-determinism.
  - Makes it easier sometimes to construct NFAs

**Definition:  $\epsilon$  -NFAs are those NFAs with at least one explicit  $\epsilon$ -transition defined.**

- $\epsilon$  -NFAs have one more column in their transition table

# EXAMPLE OF AN $\varepsilon$ -NFA

$L = \{w \mid w \text{ is empty, or if non-empty will end in } 01\}$



$\delta_E$	0	1	$\varepsilon$	
$*q'_0$	$\emptyset$	$\emptyset$	$\{q'_0, q_0\}$	$ECLOSE(q'_0)$
$q_0$	$\{q_0, q_1\}$	$\{q_0\}$	$\{q_0\}$	$ECLOSE(q_0)$
$q_1$	$\emptyset$	$\{q_2\}$	$\{q_1\}$	$ECLOSE(q_1)$
$*q_2$	$\emptyset$	$\emptyset$	$\{q_2\}$	$ECLOSE(q_2)$

- $\varepsilon$ -closure of a state  $q$ ,  $ECLOSE(q)$ , is the set of all states (including itself) that can be reached from  $q$  by repeatedly making an arbitrary number of  $\varepsilon$ -transitions.

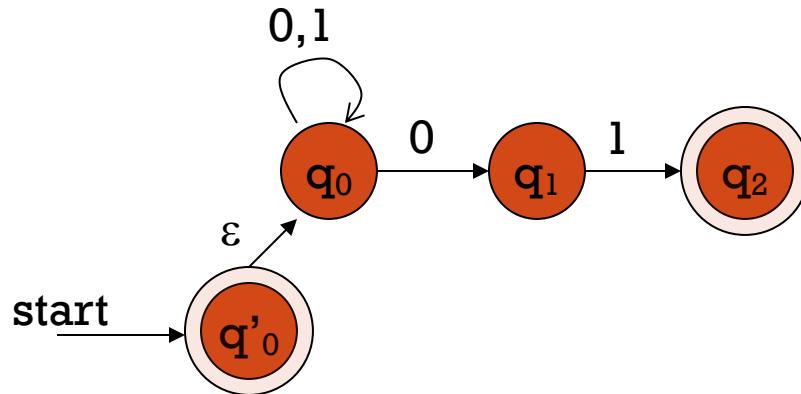
To simulate any transition:

Step 1) Go to all immediate destination states.

Step 2) From there go to all their  $\epsilon$ -closure states as well.

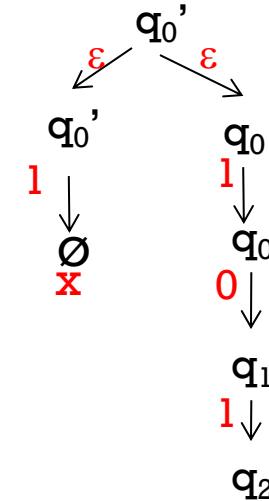
# EXAMPLE OF AN $\epsilon$ -NFA

$L = \{w \mid w \text{ is empty, or if non-empty will end in } 01\}$



Simulate for  $w=101$ :

$\delta_E$	0	1	$\epsilon$
$*q'_0$	$\emptyset$	$\emptyset$	$\{q'_0, q_0\}$
$q_0$	$\{q_0, q_1\}$	$\{q_0\}$	$\{q_0\}$
$q_1$	$\emptyset$	$\{q_2\}$	$\{q_1\}$
$*q_2$	$\emptyset$	$\emptyset$	$\{q_2\}$

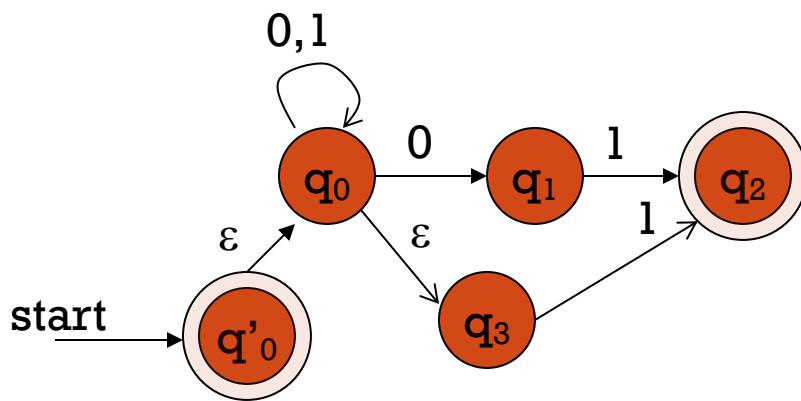


To simulate any transition:

Step 1) Go to all immediate destination states.

Step 2) From there go to all their  $\epsilon$ -closure states as well.

# EXAMPLE OF ANOTHER $\epsilon$ -NFA



Simulate for w=101:

?

$\delta_E$	0	1	$\epsilon$
$*q'_0$	$\emptyset$	$\emptyset$	$\{q'_0, q_0, q_3\}$
$q_0$	$\{q_0, q_1\}$	$\{q_0\}$	$\{q_0, q_3\}$
$q_1$	$\emptyset$	$\{q_2\}$	$\{q_1\}$
$*q_2$	$\emptyset$	$\emptyset$	$\{q_2\}$
$q_3$	$\emptyset$	$\{q_2\}$	$\{q_3\}$

# EQUIVALENCY OF DFA, NFA, $\epsilon$ -NFA

- Theorem: A language  $L$  is accepted by some  $\epsilon$ -NFA if and only if  $L$  is accepted by some DFA
- Implication:
  - $DFA \equiv NFA \equiv \epsilon\text{-NFA}$
  - (all accept Regular Languages)

# ELIMINATING $\varepsilon$ -TRANSITIONS

Let  $E = \{Q_E, \Sigma, \delta_E, q_0, F_E\}$  be an  $\varepsilon$ -NFA

Goal: To build DFA  $D = \{Q_D, \Sigma, \delta_D, \{q_D\}, F_D\}$  s.t.  $L(D) = L(E)$

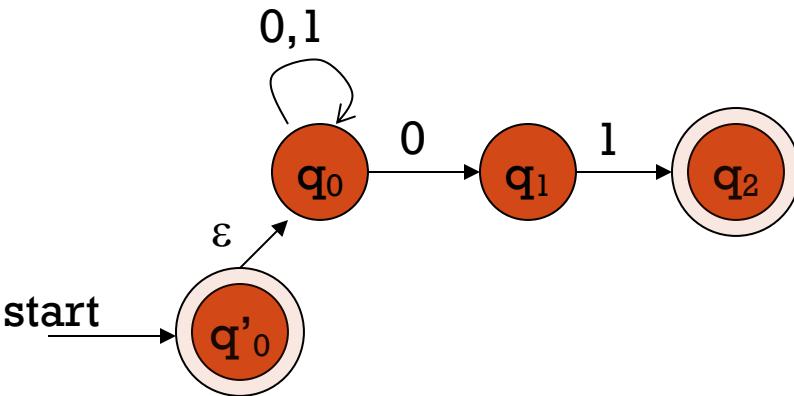
Construction:

1.  $Q_D$  = all reachable subsets of  $Q_E$  factoring in  $\varepsilon$ -closures
2.  $q_D = ECLOSE(q_0)$
3.  $F_D$  = subsets  $S$  in  $Q_D$  s.t.  $S \cap F_E \neq \emptyset$
4.  $\delta_D$ : for each subset  $S$  of  $Q_E$  and for each input symbol  $a \in \Sigma$ :
  - Let  $R = \bigcup_{p \text{ in } S} \delta_E(p, a)$  // go to destination states
  - $\delta_D(S, a) = \bigcup_{r \text{ in } R} ECLOSE(r)$  // from there, take a union of all their  $\varepsilon$ -closures

Reading: Section 2.5.5 in book

# EXAMPLE: $\epsilon$ -NFA $\rightarrow$ DFA

$L = \{w \mid w \text{ is empty, or if non-empty will end in } 01\}$

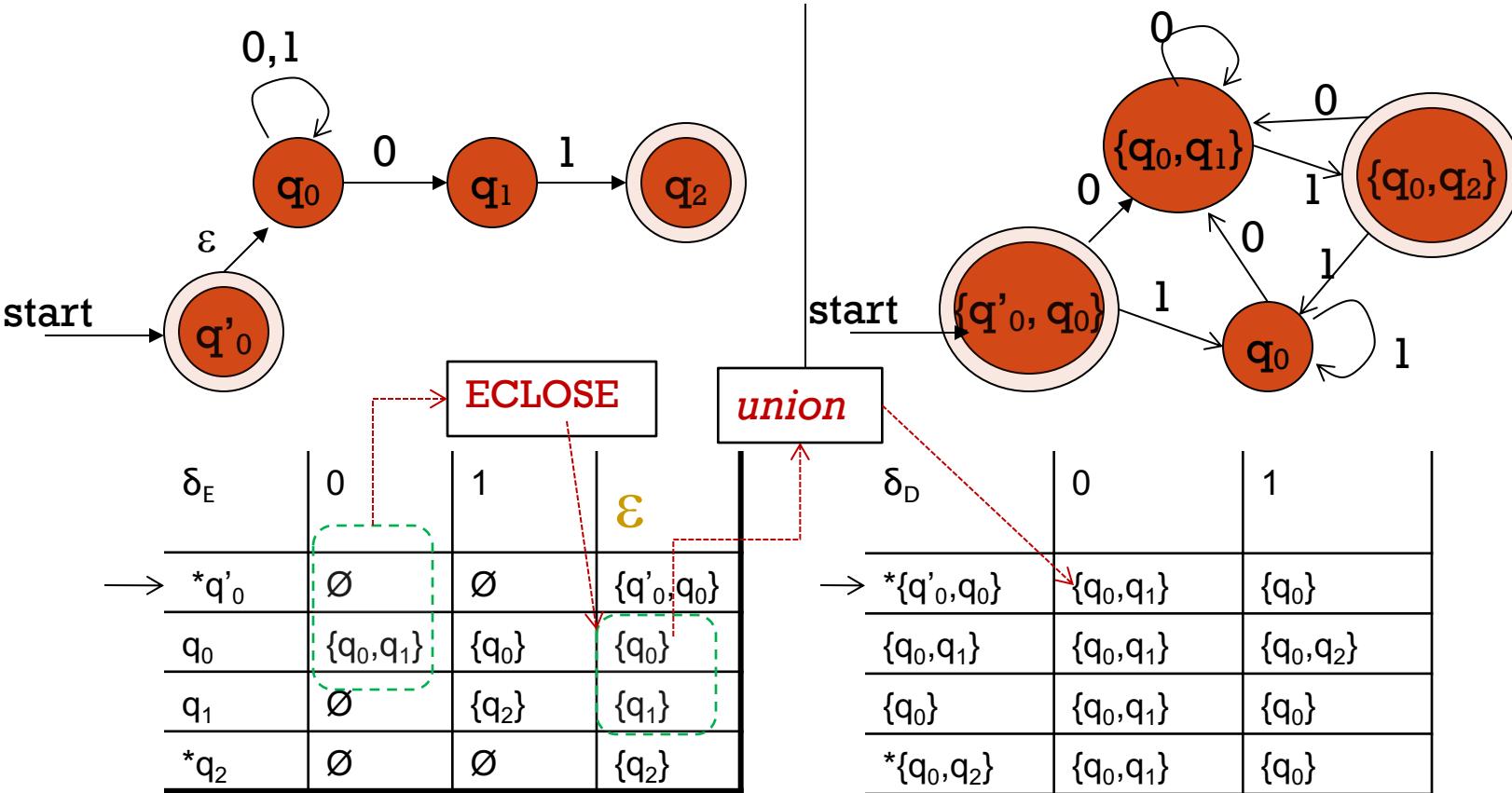


$\delta_E$	0	1	$\epsilon$
$\rightarrow *q'_0$	$\emptyset$	$\emptyset$	$\{q'_0, q_0\}$
$q_0$	$\{q_0, q_1\}$	$\{q_0\}$	$\{q_0\}$
$q_1$	$\emptyset$	$\{q_2\}$	$\{q_1\}$
$*q_2$	$\emptyset$	$\emptyset$	$\{q_2\}$

$\delta_D$	0	1
$\rightarrow *q'_0, q_0\}$		
...		

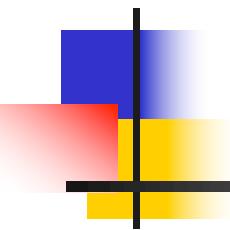
# EXAMPLE: $\varepsilon$ -NFA $\rightarrow$ DFA

$L = \{w \mid w \text{ is empty, or if non-empty will end in } 01\}$



# SUMMARY

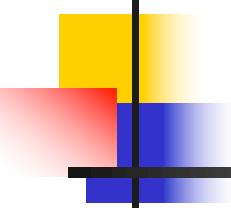
- DFA
  - Definition
  - Transition diagrams & tables
- Regular language
- NFA
  - Definition
  - Transition diagrams & tables
- DFA vs. NFA
- NFA to DFA conversion using subset construction
- Equivalency of DFA & NFA
- Removal of redundant states and including dead states
- $\epsilon$ -transitions in NFA
- Pigeon hole principles
- Text searching applications



# Regular Expressions

---

Reading: Chapter 3

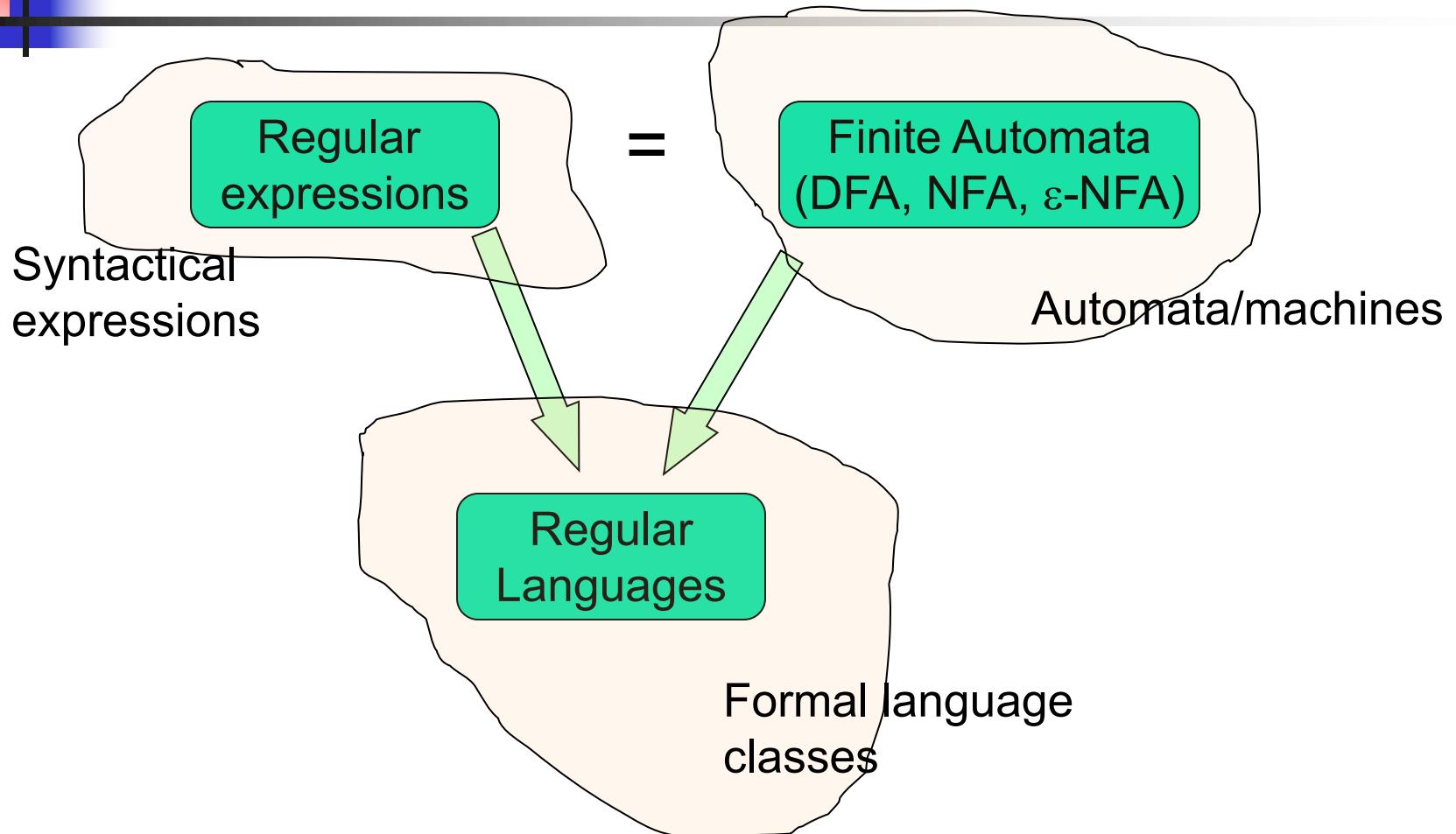


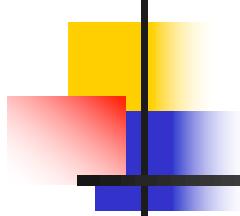
# Regular Expressions vs. Finite Automata

---

- Offers a declarative way to express the pattern of any string we want to accept
  - E.g.,  $01^* + 10^*$
- Automata => more machine-like
  - < input: string , output: [accept/reject] >
- Regular expressions => more program syntax-like
- Unix environments heavily use regular expressions
  - E.g., bash shell, grep, vi & other editors, sed
- Perl scripting – good for string processing
- Lexical analyzers such as Lex or Flex

# Regular Expressions





# Language Operators

---

- Union of two languages:
  - **L U M** = all strings that are either in L or M
  - Note: A union of two languages produces a third language
  
- Concatenation of two languages:
  - **L . M** = all strings that are of the form  $xy$   
s.t.,  $x \in L$  and  $y \in M$
  - The *dot* operator is usually omitted
    - i.e., **LM** is same as  $L.M$

“i” here refers to how many strings to concatenate from the parent language L to produce strings in the language  $L^i$

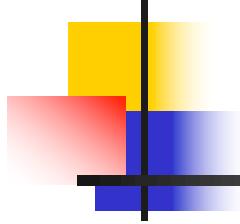
# Kleene Closure (the \* operator)

- Kleene Closure of a given language L:

- $L^0 = \{\epsilon\}$
- $L^1 = \{w \mid \text{for some } w \in L\}$
- $L^2 = \{w_1 w_2 \mid w_1 \in L, w_2 \in L \text{ (duplicates allowed)}\}$
- $L^i = \{w_1 w_2 \dots w_i \mid \text{all } w\text{'s chosen are } \in L \text{ (duplicates allowed)}\}$
- (Note: the choice of each  $w_i$  is independent)
- $L^* = \bigcup_{i \geq 0} L^i$  (arbitrary number of concatenations)

## Example:

- Let  $L = \{1, 00\}$ 
  - $L^0 = \{\epsilon\}$
  - $L^1 = \{1, 00\}$
  - $L^2 = \{11, 100, 001, 0000\}$
  - $L^3 = \{111, 1100, 1001, 10000, 000000, 00001, 00100, 0011\}$
  - $L^* = L^0 \bigcup L^1 \bigcup L^2 \bigcup \dots$

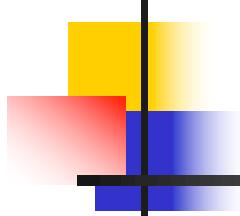


# Kleene Closure (special notes)

- $L^*$  is an infinite set iff  $|L| \geq 1$  and  $L \neq \{\epsilon\}$  **Why?**
- If  $L = \{\epsilon\}$ , then  $L^* = \{\epsilon\}$  **Why?**
- If  $L = \emptyset$ , then  $L^* = \{\epsilon\}$  **Why?**

$\Sigma^*$  denotes the set of all words over an alphabet  $\Sigma$

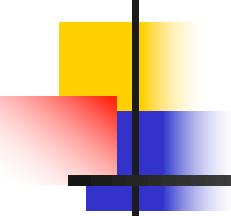
- Therefore, an abbreviated way of saying there is an arbitrary language  $L$  over an alphabet  $\Sigma$  is:
  - $L \subseteq \Sigma^*$



# Building Regular Expressions

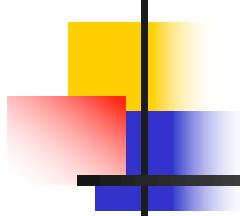
---

- Let  $E$  be a regular expression and the language represented by  $E$  is  $L(E)$
- Then:
  - $(E) = E$
  - $L(E + F) = L(E) \cup L(F)$
  - $L(E F) = L(E) L(F)$
  - $L(E^*) = (L(E))^*$



# Example: how to use these regular expression properties and language operators?

- $L = \{ w \mid w \text{ is a binary string which does not contain two consecutive 0s or two consecutive 1s anywhere}\}$ 
  - E.g.,  $w = 01010101$  is in  $L$ , while  $w = 10010$  is not in  $L$
- Goal: Build a regular expression for  $L$
- Four cases for  $w$ :
  - Case A:  $w$  starts with 0 and  $|w|$  is even
  - Case B:  $w$  starts with 1 and  $|w|$  is even
  - Case C:  $w$  starts with 0 and  $|w|$  is odd
  - Case D:  $w$  starts with 1 and  $|w|$  is odd
- Regular expression for the four cases:
  - Case A:  $(01)^*$
  - Case B:  $(10)^*$
  - Case C:  $0(10)^*$
  - Case D:  $1(01)^*$
- Since  $L$  is the union of all 4 cases:
  - Reg Exp for  $L = (01)^* + (10)^* + 0(10)^* + 1(01)^*$
- If we introduce  $\epsilon$  then the regular expression can be simplified to:
  - Reg Exp for  $L = (\epsilon + 1)(01)^*(\epsilon + 0)$



# Precedence of Operators

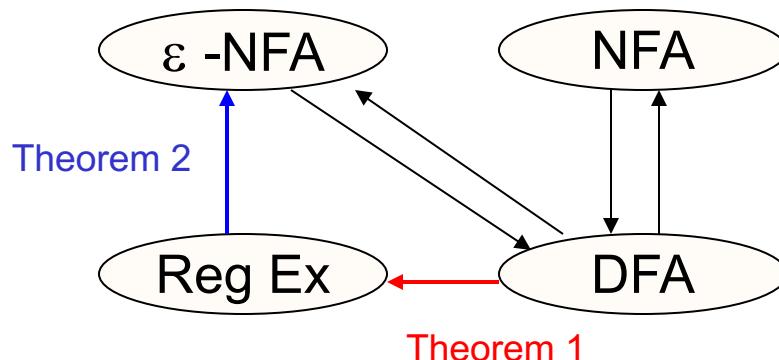
---

- Highest to lowest
  - \* operator (star)
  - . (concatenation)
  - + operator
- Example:
  - $01^* + 1 = (0 . ((1)^*)) + 1$

# Finite Automata (FA) & Regular Expressions (Reg Ex)

Proofs  
in the book

- To show that they are interchangeable, consider the following theorems:
  - Theorem 1:** For every DFA  $A$  there exists a regular expression  $R$  such that  $L(R)=L(A)$
  - Theorem 2:** For every regular expression  $R$  there exists an  $\varepsilon$ -NFA  $E$  such that  $L(E)=L(R)$



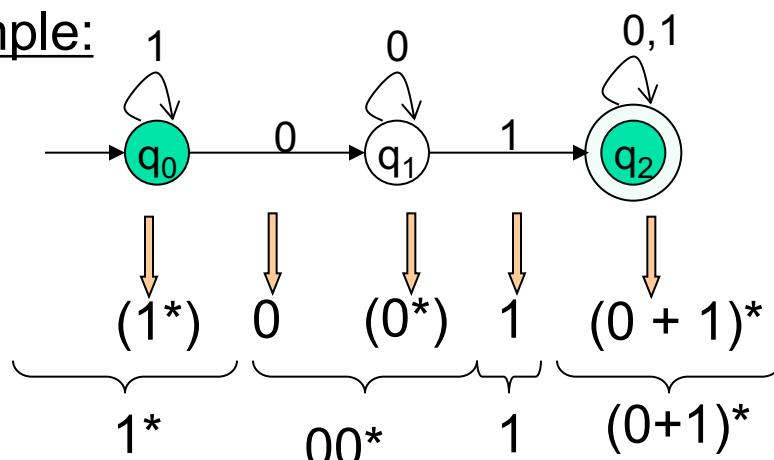
**Kleene Theorem**



# DFA to RE construction

Informally, trace all distinct paths (traversing cycles only once) from the start state to *each of the* final states and enumerate all the expressions along the way

Example:



$$1^*00^*1(0+1)^*$$

Q) What is the language?

Reg Ex

$\epsilon$ -NFA

Theorem 2

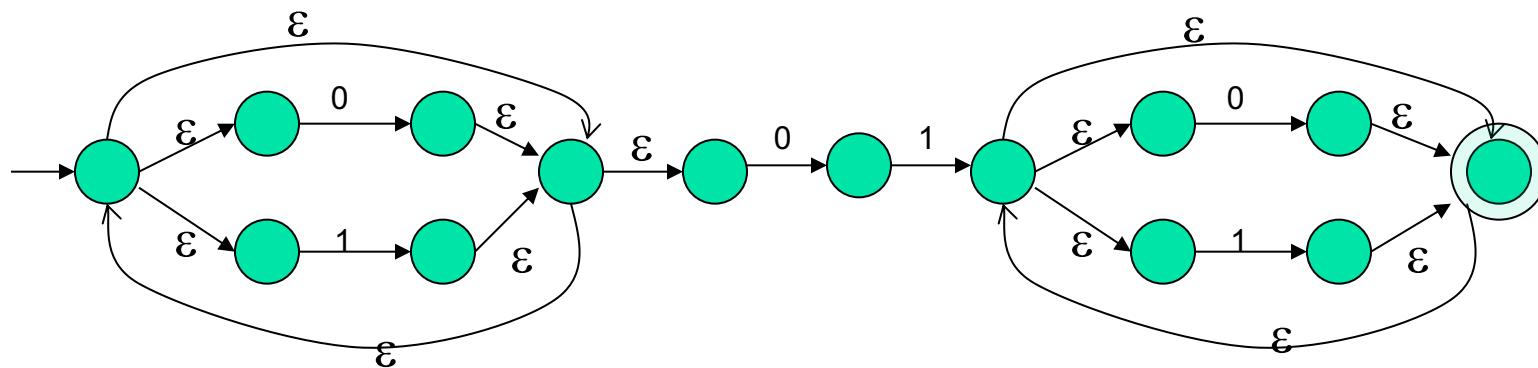
# RE to $\epsilon$ -NFA construction

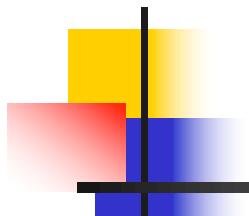
Example:  $(0+1)^*01(0+1)^*$

$(0+1)^*$

01

$(0+1)^*$

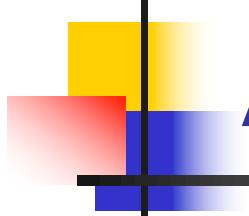




# Algebraic Laws of Regular Expressions

---

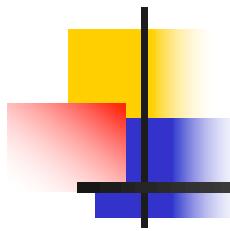
- Commutative:
  - $E+F = F+E$
- Associative:
  - $(E+F)+G = E+(F+G)$
  - $(EF)G = E(FG)$
- Identity:
  - $E+\Phi = E$
  - $\varepsilon E = E \varepsilon = E$
- Anniliator:
  - $\Phi E = E\Phi = \Phi$



# Algebraic Laws...

---

- Distributive:
  - $E(F+G) = EF + EG$
  - $(F+G)E = FE+GE$
- Idempotent:  $E + E = E$
- Involving Kleene closures:
  - $(E^*)^* = E^*$
  - $\Phi^* = \varepsilon$
  - $\varepsilon^* = \varepsilon$
  - $E^+ = EE^*$
  - $E? = \varepsilon + E$



# True or False?

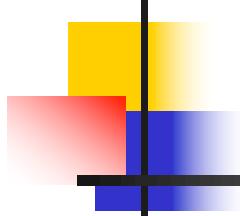
---

Let  $R$  and  $S$  be two regular expressions. Then:

1.  $((R^*)^*)^* = R^*$  ?

2.  $(R+S)^* = R^* + S^*$  ?

3.  $(RS + R)^* RS = (RR^*S)^*$  ?



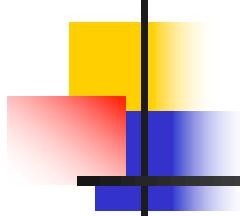
# Summary

---

- Regular expressions
- Equivalence to finite automata
- DFA to regular expression conversion
- Regular expression to  $\epsilon$ -NFA conversion
- Algebraic laws of regular expressions
- Unix regular expressions and Lexical Analyzer

# Properties of Regular Languages

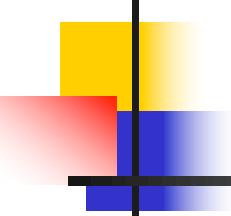
Reading: Chapter 4



# Topics

---

- 1) How to prove whether a given language is regular or not?
- 2) Closure properties of regular languages
- 3) Minimization of DFAs



# Some languages are *not* regular

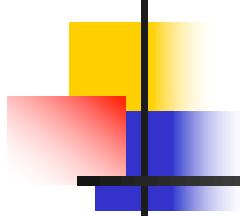
---

When is a language is regular?

if we are able to construct one of the  
following: DFA or NFA or  $\epsilon$ -NFA or regular  
expression

When is it not?

If we can show that no FA can be built for a  
language



# How to prove languages are *not* regular?

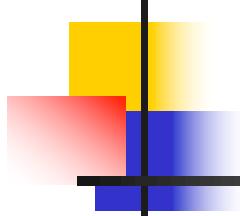
---

What if we cannot come up with any FA?

- A) Can it be language that is not regular?
- B) Or is it that we tried wrong approaches?

How do we *decisively* prove that a language is not regular?

*“The hardest thing of all is to find a black cat in a dark room,  
especially if there is no cat!”*      -Confucius

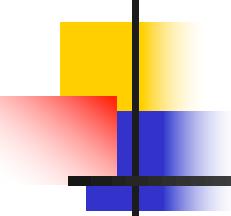


# Example of a non-regular language

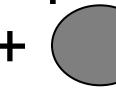
---

Let  $L = \{w \mid w \text{ is of the form } 0^n1^n, \text{ for all } n \geq 0\}$

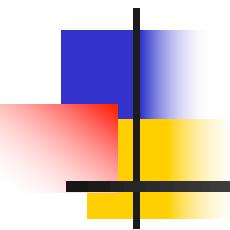
- Hypothesis:  $L$  is not regular
- Intuitive rationale: How do you keep track of a running count in an FA?
- A more formal rationale:
  - By contradiction, if  $L$  is regular then there should exist a DFA for  $L$ .
  - Let  $k$  = number of states in that DFA.
  - Consider the special word  $w = 0^k1^k \Rightarrow w \in L$
  - DFA is in some state  $p_i$ , after consuming the first  $i$  symbols in  $w$



# Rationale...

- Let  $\{p_0, p_1, \dots, p_k\}$  be the sequence of states that the DFA should have visited after consuming the first  $k$  symbols in  $w$  which is  $0^k$
- But there are only  $k$  states in the DFA!
- ==> at least one state should repeat somewhere along the path (by  +  Principle)
- ==> Let the repeating state be  $p_i = p_j$  for  $i < j$
- ==> We can fool the DFA by inputting  $0^{(k-(j-i))}1^k$  and still get it to accept (note:  $k-(j-i)$  is at most  $k-1$ ).
- ==> DFA accepts strings w/ unequal number of 0s and 1s, implying that the DFA is wrong!





# The Pumping Lemma for Regular Languages

## **What it is?**

The Pumping Lemma is a property of all regular languages.

## **How is it used?**

A technique that is used to show that a given language is not regular

# Pumping Lemma for Regular Languages

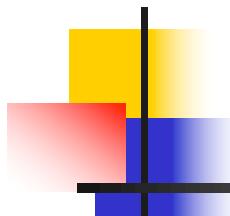
Let  $L$  be a regular language

Then there exists some constant  $N$  such that for every string  $w \in L$  s.t.  $|w| \geq N$ , there exists a way to break  $w$  into three parts,  $w = xyz$ , such that:

1.  $y \neq \varepsilon$
2.  $|xy| \leq N$
3. For all  $k \geq 0$ , all strings of the form  $xy^kz \in L$

This property should hold for all regular languages.

**Definition:**  $N$  is called the “Pumping Lemma Constant”

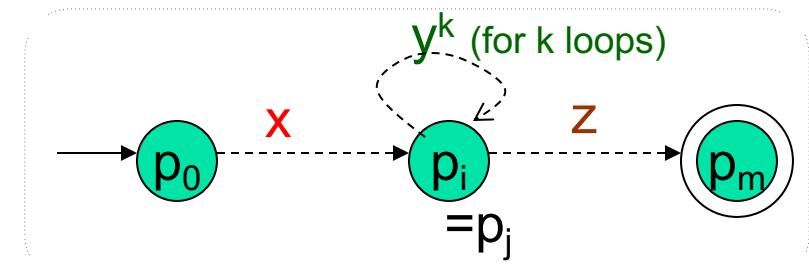


# Pumping Lemma: Proof

- $L$  is regular  $\Rightarrow$  it should have a DFA.
  - Set  $N$  := number of states in the DFA
- Any string  $w \in L$ , s.t.  $|w| \geq N$ , should have the form:  $w = a_1 a_2 \dots a_m$ , where  $m \geq N$
- Let the states traversed after reading the first  $N$  symbols be:  $\{p_0, p_1, \dots, p_N\}$ 
  - $\Rightarrow$  There are  $N+1$  p-states, while there are only  $N$  DFA states
  - $\Rightarrow$  at least one state has to repeat i.e.,  $p_i = p_j$  where  $0 \leq i < j \leq N$  (by PHP)

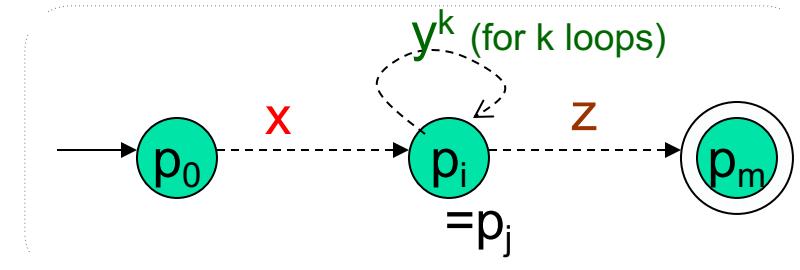
# Pumping Lemma: Proof...

- => We should be able to break  $w=xyz$  as follows:
  - $x=a_1a_2..a_i$ ;       $y=a_{i+1}a_{i+2}..a_j$ ;       $z=a_{j+1}a_{j+2}..a_m$
  - $x$ 's path will be  $p_0..p_i$
  - $y$ 's path will be  $p_i p_{i+1}..p_j$  (but  $p_i=p_j$  implying a loop)
  - $z$ 's path will be  $p_j p_{j+1}..p_m$
- Now consider another string  $w_k=xy^kz$ , where  $k \geq 0$
- Case  $k=0$ 
  - DFA will reach the accept state  $p_m$
- Case  $k>0$ 
  - DFA will loop for  $y^k$ , and finally reach the accept state  $p_m$  for  $z$
- In either case,  $w_k \in L$       This proves part (3) of the lemma

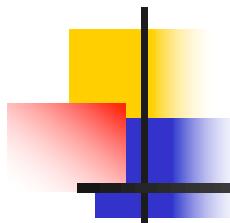


# Pumping Lemma: Proof...

- For part (1):
  - Since  $i < j$ ,  $y \neq \varepsilon$
- For part (2):
  - By PHP, the repetition of states has to occur within the first  $N$  symbols in  $w$
  - $\implies |xy| \leq N$



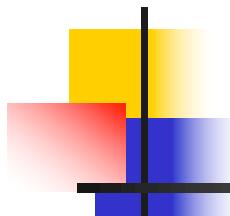
□



# The Purpose of the Pumping Lemma for RL

---

- To prove that some languages *cannot* be regular.

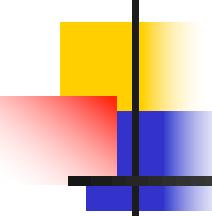


# How to use the pumping lemma?

---

Think of playing a 2 person game

- Role 1: **We** claim that the language cannot be regular
- Role 2: An **adversary** who claims the language is regular
- We show that the adversary's statement will lead to a contradiction that implies pumping lemma *cannot* hold for the language.
- We win!!



# How to use the pumping lemma? (The Steps)

1. (we) L is not regular.
2. (adv.) Claims that L is regular and gives you a value for N as its P/L constant
3. (we) Using N, choose a string  $w \in L$  s.t.,
  1.  $|w| \geq N$ ,
  2. Using w as the template, construct other words  $w_k$  of the form  $xy^kz$  and show that at least one such  $w_k \notin L$

=> this implies we have successfully broken the pumping lemma for the language, and hence that the adversary is wrong.

(Note: In this process, we may have to try many values of k, starting with k=0, and then 2, 3, .. so on, until  $w_k \notin L$ )

Note: We don't have any control over  $N$ , except that it is positive.

We also don't have any control over how to split  $w=xyz$ ,  
but  $xyz$  should respect the P/L conditions (1) and (2).

# Using the Pumping Lemma

- What WE do?
- What the Adversary does?
  1. Claims  $L$  is regular
  2. Provides  $N$
- 3. Using  $N$ , we construct our template string  $w$
- 4. Demonstrate to the adversary, either through pumping up or down on  $w$ , that some string  $w_k \notin L$   
(this should happen regardless of  $w=xyz$ )

Note: This  $N$  can be anything (need not necessarily be the #states in the DFA.  
It's the adversary's choice.)

## Example of using the Pumping Lemma to prove that a language is not regular

**Let  $L_{eq} = \{w \mid w \text{ is a binary string with equal number of 1s and 0s}\}$**

- Your Claim:  $L_{eq}$  is not regular
- Proof:
  - By contradiction, let  $L_{eq}$  be regular → adv.
  - P/L constant should exist → adv.
    - Let  $N$  = that P/L constant
  - Consider input  $w = 0^N 1^N$  → you  
*(your choice for the template string)*
  - By pumping lemma, we should be able to break → you  $w=xyz$ , such that:
    - 1)  $y \neq \epsilon$
    - 2)  $|xy| \leq N$
    - 3) For all  $k \geq 0$ , the string  $xy^k z$  is also in  $L$

Template string  $w = 0^N 1^N = 00 \dots \xleftarrow{N} 011 \dots \xrightarrow{N} 1$

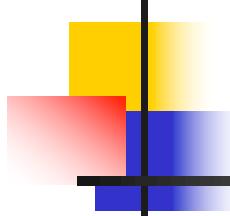
# Proof...

- Because  $|xy| \leq N$ ,  $xy$  should contain only 0s  $\rightarrow$  you
- (This and because  $y \neq \epsilon$ , implies  $y=0^+$ )
- Therefore  $x$  can contain *at most*  $N-1$  0s
- Also, all the  $N$  1s must be inside  $z$
- By (3), any string of the form  $xy^k z \in L_{eq}$  for all  $k \geq 0$
- Case  $k=0$ :  $xz$  has at most  $N-1$  0s but has  $N$  1s
- Therefore,  $xy^0 z \notin L_{eq}$
- This violates the P/L (a contradiction)  $\cancel{\rightarrow}$

Setting  $k=0$  is referred to as "pumping down"

Setting  $k>1$  is referred to as "pumping up"

Another way of proving this will be to show that if the #0s is arbitrarily pumped up (e.g.,  $k=2$ ), then the #0s will become exceed the #1s



## Exercise 2

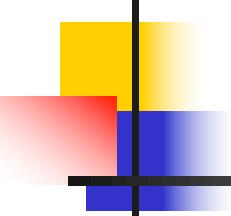
---

*Prove  $L = \{0^n 1 0^n \mid n \geq 1\}$  is not regular*

Note: This  $n$  is not to be confused with the pumping lemma constant  $N$ . That can be different.

In other words, the above question is same as proving:

- $L = \{0^m 1 0^m \mid m \geq 1\}$  is not regular



# Example 3: Pumping Lemma

**Claim:  $L = \{ 0^i \mid i \text{ is a perfect square} \}$  is not regular**

■ Proof:

- By contradiction, let  $L$  be regular.
- P/L should apply
- Let  $N = P/L$  constant
- Choose  $w=0^{N^2}$
- By pumping lemma,  $w=xyz$  satisfying all three rules
- By rules (1) & (2),  $y$  has between 1 and  $N$  0s
- By rule (3), any string of the form  $xy^kz$  is also in  $L$  for all  $k \geq 0$
- Case  $k=0$ :

- $\# \text{zeros}(xy^0z) = \# \text{zeros}(xyz) - \# \text{zeros}(y)$
- $N^2 - N \leq \# \text{zeros}(xy^0z) \leq N^2 - 1$
- $(N-1)^2 < N^2 - N \leq \# \text{zeros}(xy^0z) \leq N^2 - 1 < N^2$
- $xy^0z \notin L$
- But the above will complete the proof ONLY IF  $N > 1$ .
- ... (proof contd.. Next slide)

# Example 3: Pumping Lemma

- (proof contd...)
  - If the adversary pick  $N=1$ , then  $(N-1)^2 \leq N^2 - N$ , and therefore the #zeros( $xy^0z$ ) could end up being a perfect square!
  - This means that pumping down (i.e., setting  $k=0$ ) is not giving us the proof!
  - So lets try pumping up next...
- Case  $k=2$ :
  - $\#zeros(xy^2z) = \#zeros(xy^0z) + \#zeros(y)$
  - $N^2 + 1 \leq \#zeros(xy^2z) \leq N^2 + N$
  - $N^2 < N^2 + 1 \leq \#zeros(xy^2z) \leq N^2 + N < (N+1)^2$
  - $xy^2z \notin L$  
- (Notice that the above should hold for all possible  $N$  values of  $N>0$ . Therefore, this completes the proof.)

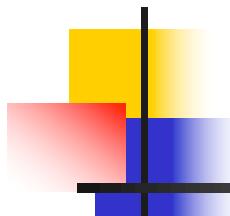
# Closure properties of Regular Languages

# Closure properties for Regular Languages (RL)

- Closure property:
  - If a set of regular languages are combined using an operator, then the resulting language is also regular
- Regular languages are closed under:
  - Union, intersection, complement, difference
  - Reversal
  - Kleene closure
  - Concatenation
  - Homomorphism
  - Inverse homomorphism

This is different from Kleene closure

Now, lets prove all of this!



# RLs are closed under union

---

- IF L and M are two RLs THEN:
  - they both have two corresponding regular expressions, R and S respectively
  - $(L \cup M)$  can be represented using the regular expression  $R+S$
  - Therefore,  $(L \cup M)$  is also regular

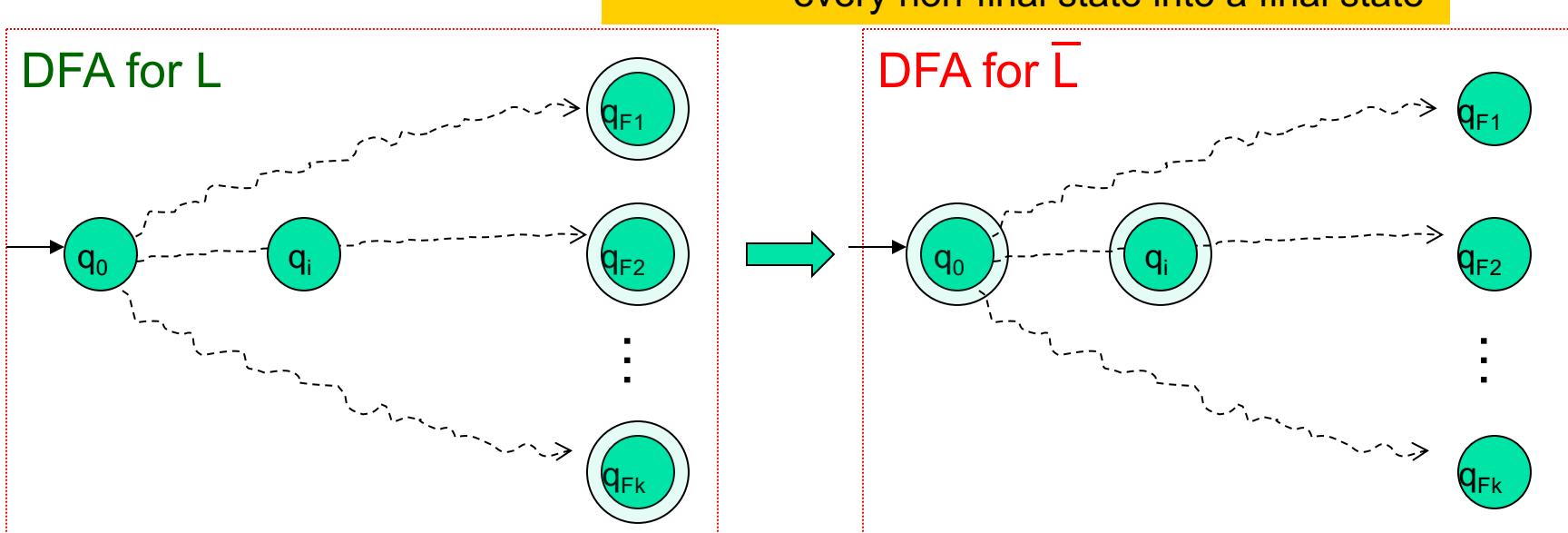
□

How can this be proved using FAs?

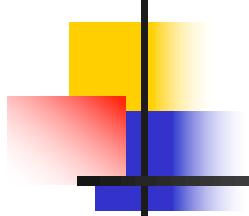
# RLs are closed under complementation

- If  $L$  is an RL over  $\Sigma$ , then  $\overline{L} = \Sigma^* - L$ 
  - To show  $\overline{L}$  is also regular, make the following construction

Convert every final state into non-final, and every non-final state into a final state



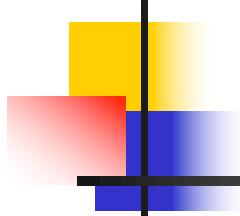
Assumes  $q_0$  is a non-final state. If not, do the opposite.



# RLs are closed under intersection

---

- A quick, indirect way to prove:
  - By DeMorgan's law:
$$L \cap M = \overline{\overline{L} \cup \overline{M}}$$
  - Since we know RLs are closed under union and complementation, they are also closed under intersection
- A more direct way would be construct a finite automaton for  $L \cap M$

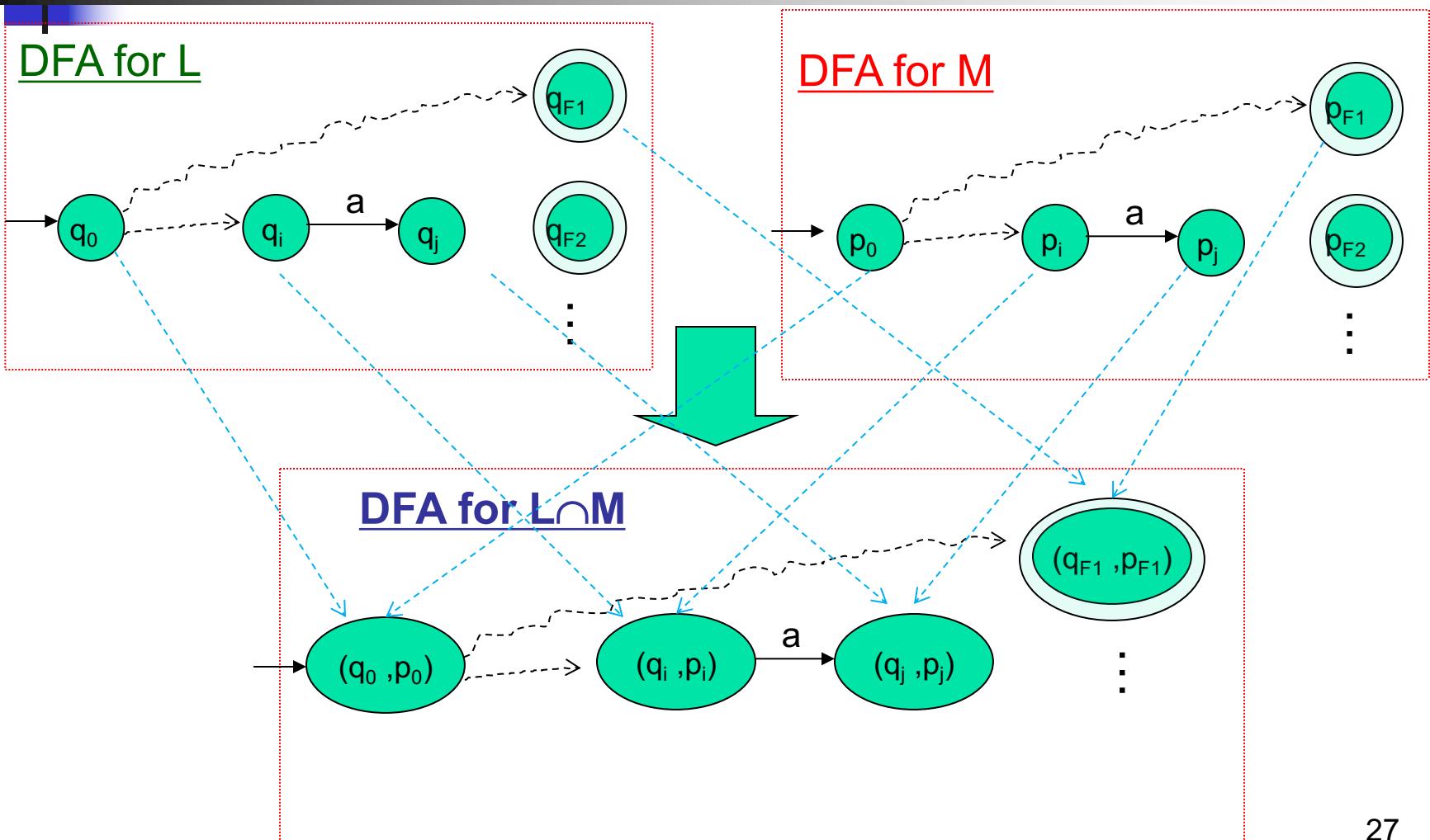


# DFA construction for $L \cap M$

---

- $A_L = \text{DFA for } L = \{Q_L, \Sigma, q_L, F_L, \delta_L\}$
- $A_M = \text{DFA for } M = \{Q_M, \Sigma, q_M, F_M, \delta_M\}$
- Build  $A_{L \cap M} = \{Q_L \times Q_M, \Sigma, (q_L, q_M), F_L \times F_M, \delta\}$   
such that:
  - $\delta((p, q), a) = (\delta_L(p, a), \delta_M(q, a))$ , where  $p$  in  $Q_L$ ,  
and  $q$  in  $Q_M$
- This construction ensures that a string  $w$  will  
be accepted if and only if  $w$  reaches an  
accepting state in both input DFAs.

# DFA construction for $L \cap M$

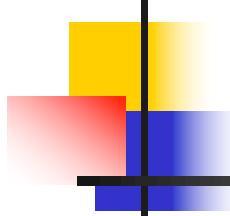


# RLs are closed under set difference

- We observe:
  - $L - M = L \cap \overline{M}$
- Therefore,  $L - M$  is also regular

Closed under intersection

Closed under  
complementation



# RLs are closed under reversal

---

Reversal of a string  $w$  is denoted by  $w^R$

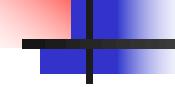
- E.g.,  $w=00111$ ,  $w^R=11100$

Reversal of a language:

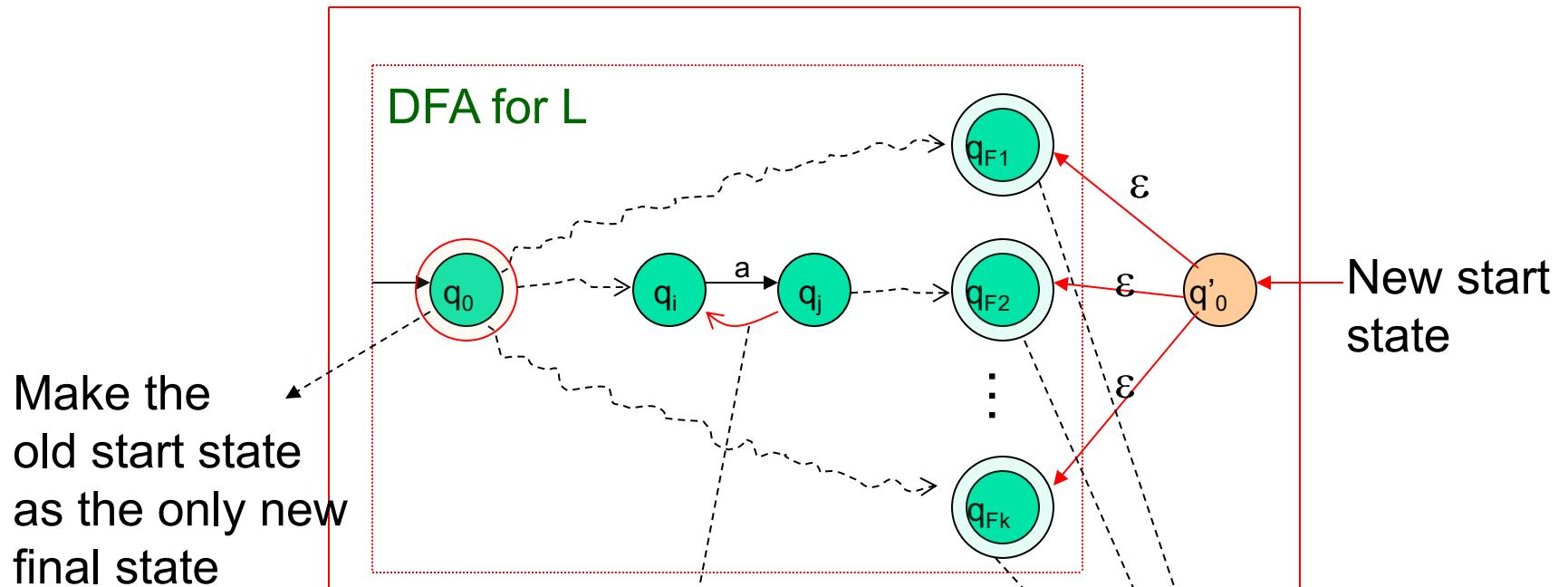
- $L^R =$  The language generated by reversing all strings in  $L$

Theorem: If  $L$  is regular then  $L^R$  is also regular

# $\epsilon$ -NFA Construction for $L^R$



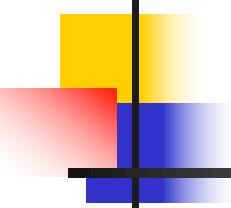
New  $\epsilon$ -NFA for  $L^R$



What to do if  $q_0$  was one of the final states in the input DFA?

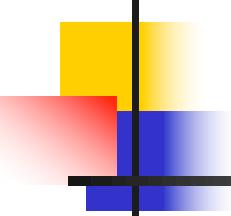
Reverse all transitions

Convert the old set of final states into non-final states



# If $L$ is regular, $L^R$ is regular (proof using regular expressions)

- Let  $E$  be a regular expression for  $L$
- Given  $E$ , how to build  $E^R$ ?
- Basis: If  $E = \varepsilon, \emptyset$ , or  $a$ , then  $E^R = E$
- Induction: Every part of  $E$  (refer to the part as “ $F$ ”) can be in only *one* of the three following forms:
  1.  $F = F_1 + F_2$ 
    - $F^R = F_1^R + F_2^R$
  2.  $F = F_1 F_2$ 
    - $F^R = F_2^R F_1^R$
  3.  $F = (F_1)^*$ 
    - $(F^R)^* = (F_1^R)^*$



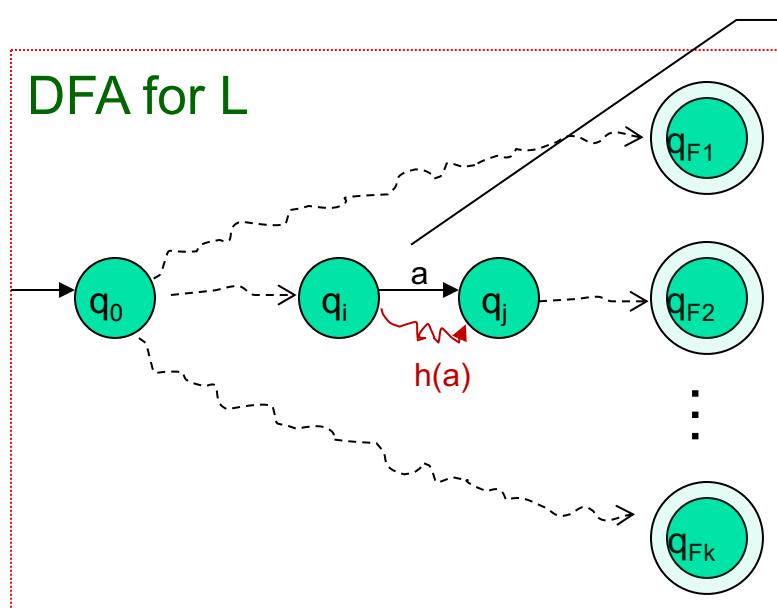
# Homomorphisms

---

- Substitute each symbol in  $\Sigma$  (main alphabet) by a corresponding string in  $T$  (another alphabet)
  - $h: \Sigma \rightarrow T^*$
- Example:
  - Let  $\Sigma = \{0, 1\}$  and  $T = \{a, b\}$
  - Let a homomorphic function  $h$  on  $\Sigma$  be:
    - $h(0)=ab, h(1)=\varepsilon$
    - If  $w=10110$ , then  $h(w) = \varepsilon ab\varepsilon\varepsilon ab = abab$
- In general,
  - $h(w) = h(a_1) h(a_2) \dots h(a_n)$

Given a DFA for  $L$ , how to convert it into an FA for  $h(L)$ ?

# FA Construction for $h(L)$



Replace every edge “ $a$ ” by a path labeled  $h(a)$  in the new DFA

- Build a new FA that simulates  $h(a)$  for every symbol  $a$  transition in the above DFA
- The resulting FA may or may not be a DFA, but will be a FA for  $h(L)$

Given a DFA for M, how to convert it into an FA for  $h^{-1}(M)$ ?

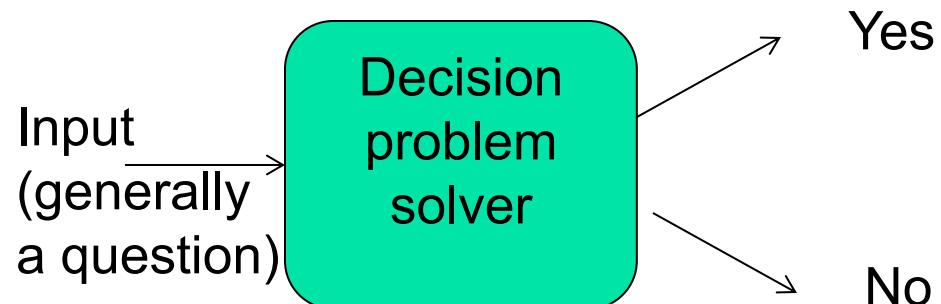
The set of strings in  $\Sigma^*$  whose homomorphic translation results in the strings of M

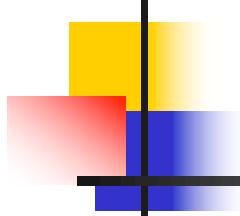
# Inverse homomorphism

- Let  $h: \Sigma \rightarrow T^*$
  - Let M be a language over alphabet T
  - $h^{-1}(M) = \{w \mid w \in \Sigma^* \text{ s.t., } h(w) \in M\}$
- Claim: If M is regular, then so is  $h^{-1}(M)$*
- Proof:
    - Let A be a DFA for M
    - Construct another DFA A' which encodes  $h^{-1}(M)$
    - A' is an exact replica of A, except that its transition functions are s.t. for any input symbol  $a$  in  $\Sigma$ , A' will simulate  $h(a)$  in A.
      - $\delta(p,a) = \hat{\delta}(p,h(a))$

# Decision properties of regular languages

Any “decision problem” looks like this:

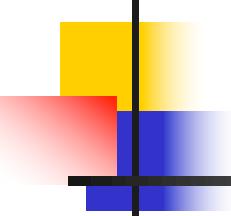




# Membership question

---

- Decision Problem: Given  $L$ , is  $w$  in  $L$ ?
- Possible answers: Yes or No
- Approach:
  1. Build a DFA for  $L$
  2. Input  $w$  to the DFA
  3. If the DFA ends in an accepting state, then yes; otherwise no.



# Emptiness test

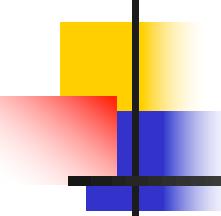
---

- Decision Problem: Is  $L = \emptyset$  ?
- Approach:

On a DFA for  $L$ :

1. From the start state, run a *reachability* test, which returns:
  1. success: if there is at least one final state that is reachable from the start state
  2. failure: otherwise
2.  $L = \emptyset$  if and only if the reachability test fails

How to implement the reachability test?



# Finiteness

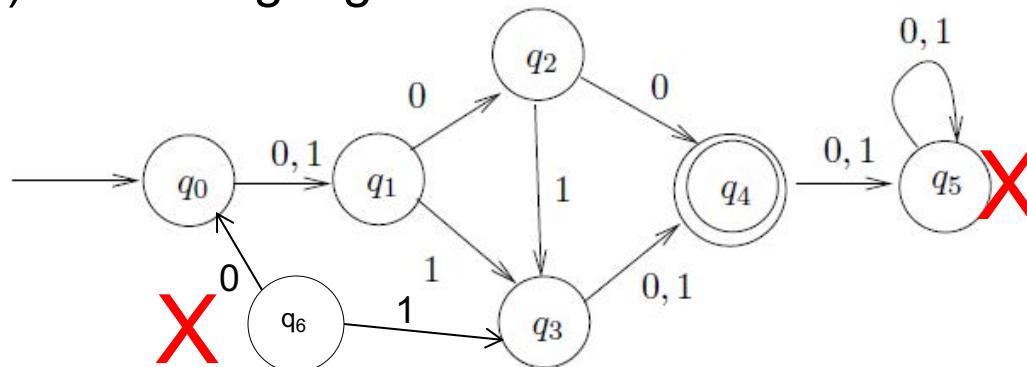
---

- Decision Problem: Is  $L$  finite or infinite?
- Approach:
  - On a DFA for  $L$ :
    1. Remove all states unreachable from the start state
    2. Remove all states that cannot lead to any accepting state.
    3. After removal, check for cycles in the resulting FA
    4.  $L$  is finite if there are no cycles; otherwise it is infinite
  - Another approach
    - Build a regular expression and look for Kleene closure

How to implement steps 2 and 3?

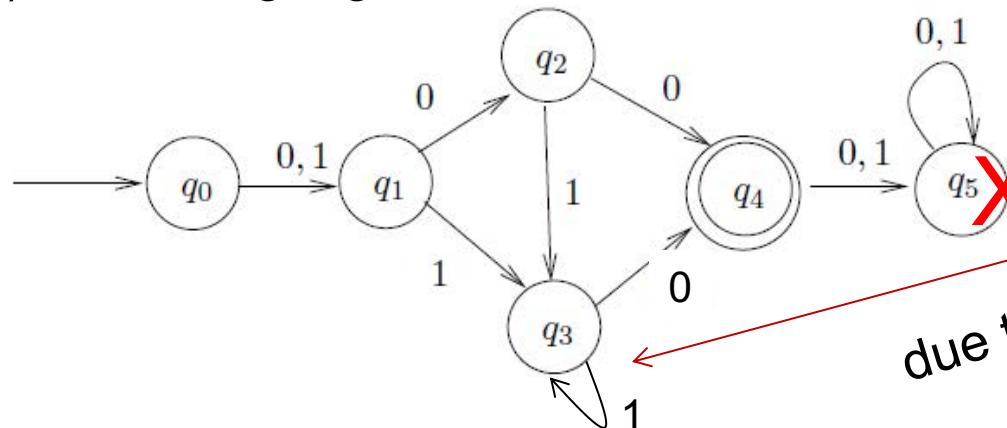
# Finiteness test - examples

Ex 1) Is the language of this DFA finite or infinite?



FINITE

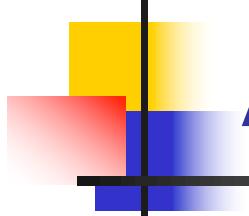
Ex 2) Is the language of this DFA finite or infinite?



INFINITE

due to this

# Equivalence & Minimization of DFAs



# Applications of interest

---

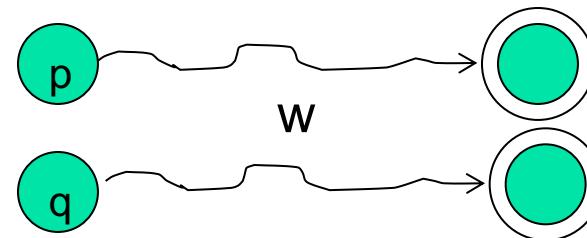
- Comparing two DFAs:
  - $L(\text{DFA}_1) == L(\text{DFA}_2)?$
- How to minimize a DFA?
  1. Remove unreachable states
  2. Identify & condense equivalent states into one

# When to call two states in a DFA “equivalent”?

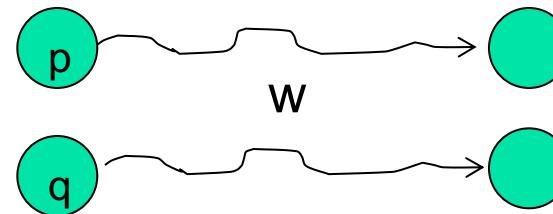
Two states  $p$  and  $q$  are said to be *equivalent* iff:

- i) Any string  $w$  accepted by starting at  $p$  is also accepted by starting at  $q$ ;

AND



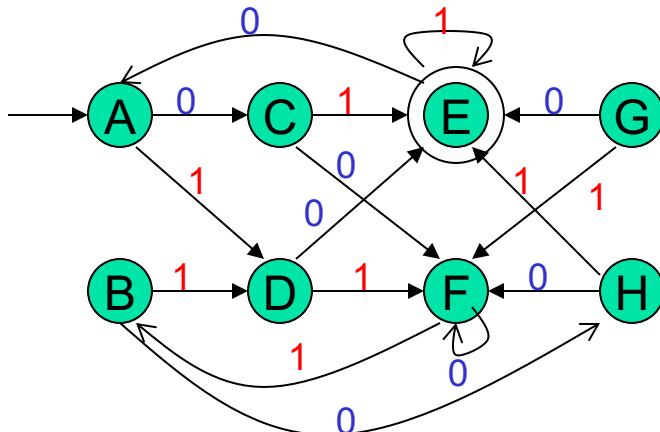
- i) Any string  $w$  rejected by starting at  $p$  is also rejected by starting at  $q$ .



→  $p \equiv q$

# Computing equivalent states in a DFA

## Table Filling Algorithm



### Pass #0

1. Mark accepting states ≠ non-accepting states

### Pass #1

1. Compare every pair of states
2. Distinguish by one symbol transition
3. Mark = or ≠ or blank(tbd)

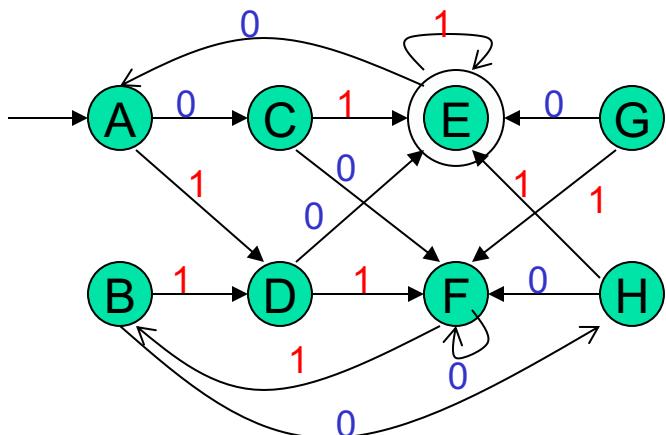
### Pass #2

1. Compare every pair of states
2. Distinguish by up to two symbol transitions (until different or same or tbd)

....  
(keep repeating until table complete)

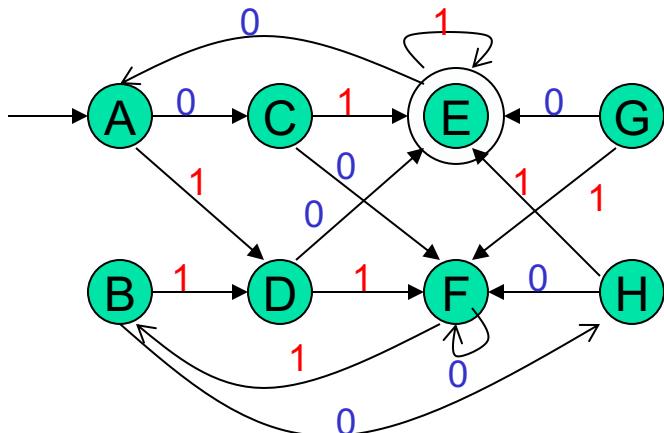
A	=							
B	=							
C	X	X	=					
D	X	X	X	=				
E	X	X	X	X	=			
F	X	X	X	X	X	=		
G	X	X	X	=	X	X	=	
H	X	X	=	X	X	X	X	=
	A	B	C	D	E	F	G	H

# Table Filling Algorithm - step by step



A	=							
B		=						
C			=					
D				=				
E					=			
F						=		
G							=	
H								=
	A	B	C	D	E	F	G	H

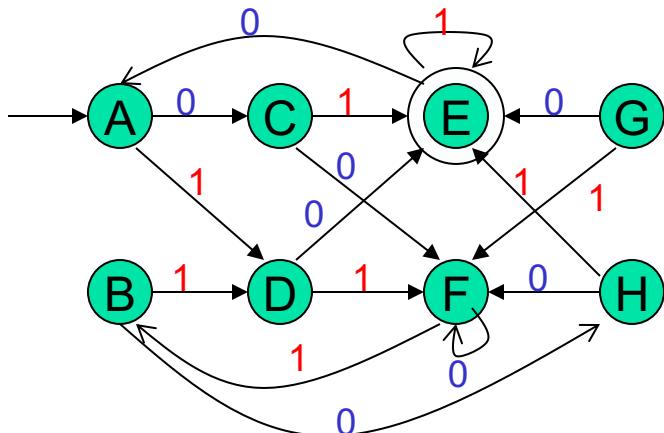
# Table Filling Algorithm - step by step



1. Mark X between accepting vs. non-accepting state

A	=						
B		=					
C			=				
D				=			
E	X	X	X	X	=		
F				X	=		
G				X		=	
H				X			=
	A	B	C	D	E	F	G
							H

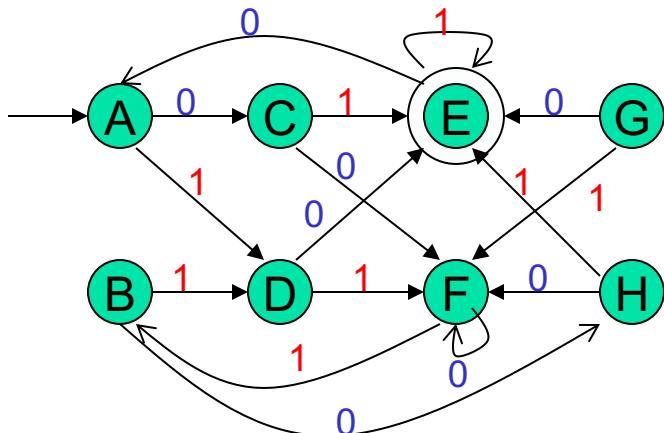
# Table Filling Algorithm - step by step



1. Mark **X** between accepting vs. non-accepting state
2. Look 1- hop away for distinguishing states or strings

A	=						
B		=					
C	X		=				
D	X			=			
E	X	X	X	X	=		
F				X	=		
G	X			X		=	
H	X			X			=
	A	B	C	D	E	F	G

# Table Filling Algorithm - step by step

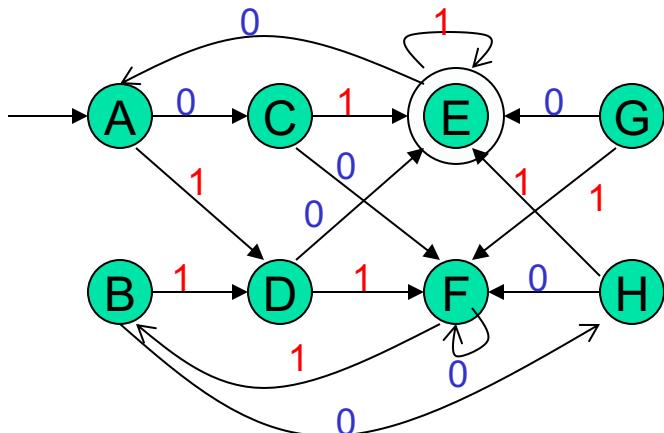


1. Mark X between accepting vs. non-accepting state
2. Look 1-hop away for distinguishing states or strings

A	=						
B		=					
C	X	X	=				
D	X	X		=			
E	X	X	X	X	=		
F				X	=		
G	X	X		X		=	
H	X	X		X			=
A	B	C	D	E	F	G	H

↑

# Table Filling Algorithm - step by step

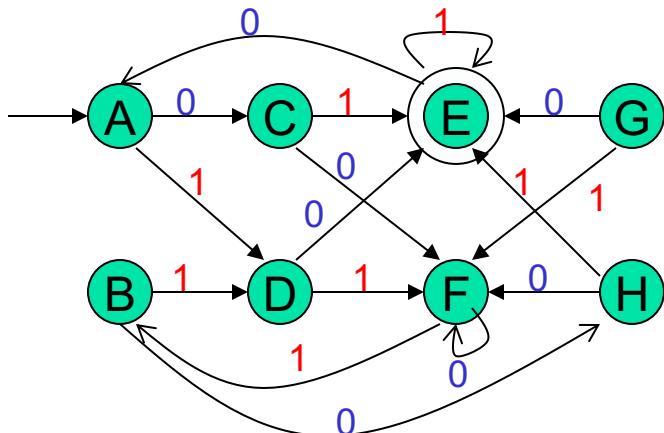


1. Mark X between accepting vs. non-accepting state
2. Look 1-hop away for distinguishing states or strings

A	=							
B		=						
C	X	X	=					
D	X	X	X	=				
E	X	X	X	X	=			
F			X		X	=		
G	X	X	X		X		=	
H	X	X	=		X			=
	A	B	C	D	E	F	G	H

A blue arrow points upwards from the letter C to the row labeled C in the table.

# Table Filling Algorithm - step by step

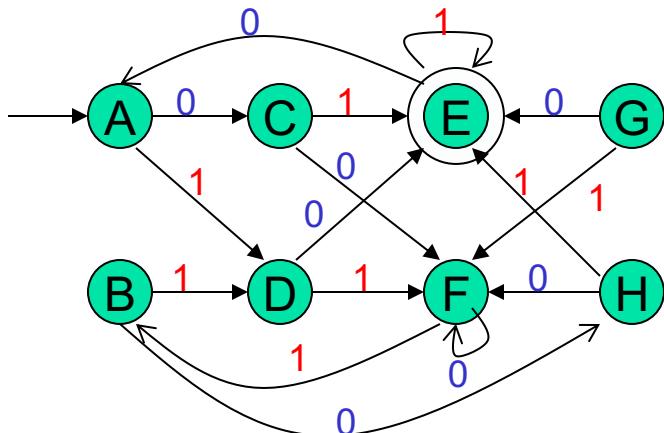


1. Mark X between accepting vs. non-accepting state
2. Look 1-hop away for distinguishing states or strings

A	=							
B		=						
C	X	X	=					
D	X	X	X	=				
E	X	X	X	X	=			
F			X	X	X	=		
G	X	X	X	=	X		=	
H	X	X	=	X	X		=	
	A	B	C	D	E	F	G	H

↑

# Table Filling Algorithm - step by step

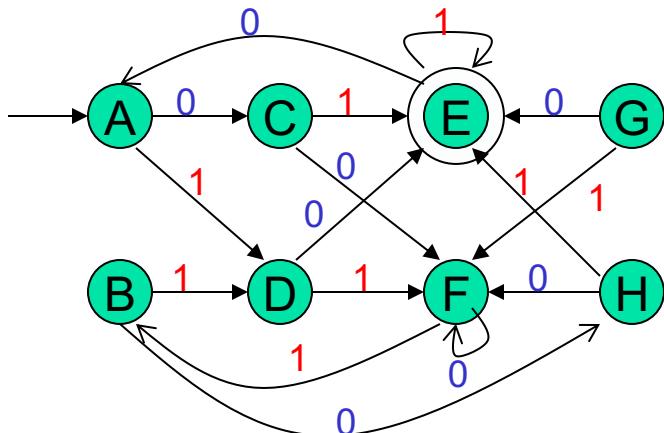


1. Mark X between accepting vs. non-accepting state
2. Look 1-hop away for distinguishing states or strings

A	=							
B		=						
C	X	X	=					
D	X	X	X	=				
E	X	X	X	X	=			
F			X	X	X	=		
G	X	X	X	=	X	X	=	
H	X	X	=	X	X	X		=
	A	B	C	D	E	F	G	H

A blue arrow points upwards from the letter F in the row header to the state F in the table body.

# Table Filling Algorithm - step by step

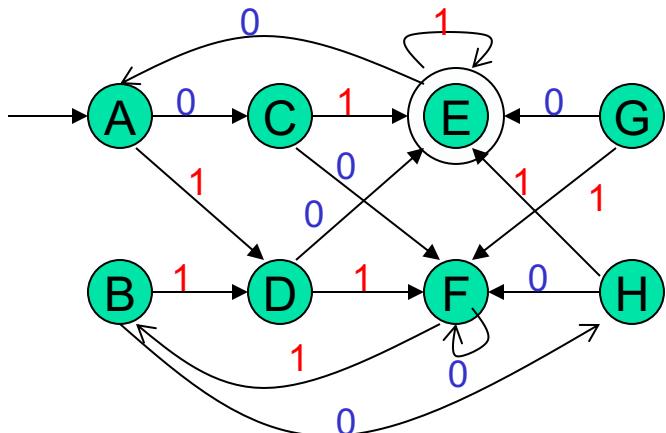


1. Mark X between accepting vs. non-accepting state
2. Look 1-hop away for distinguishing states or strings

A	=						
B		=					
C	X	X	=				
D	X	X	X	=			
E	X	X	X	X	=		
F			X	X	X	=	
G	X	X	X	=	X	X	=
H	X	X	=	X	X	X	=
	A	B	C	D	E	F	G
							H

A blue arrow points from the bottom row to the letter H.

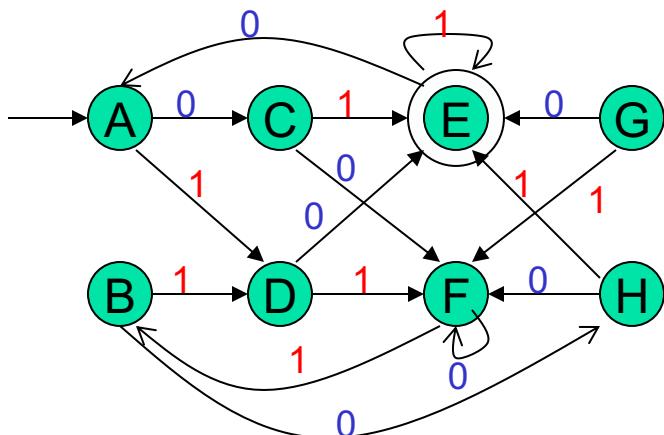
# Table Filling Algorithm - step by step



1. Mark X between accepting vs. non-accepting state
2. Pass 1:  
Look 1-hop away for distinguishing states or strings
3. Pass 2:  
Look 1-hop away again for distinguishing states or strings  
continue....

A	=							
B	=	=						
C	X	X	=					
D	X	X	X	=				
E	X	X	X	X	=			
F	X	X	X	X	X	=		
G	X	X	X	=	X	X	=	
H	X	X	=	X	X	X	X	=
	A	B	C	D	E	F	G	H

# Table Filling Algorithm - step by step



1. Mark **X** between accepting vs. non-accepting state
2. Pass 1:  
Look 1-hop away for distinguishing states or strings
3. Pass 2:  
Look 1-hop away again for distinguishing states or strings  
continue....

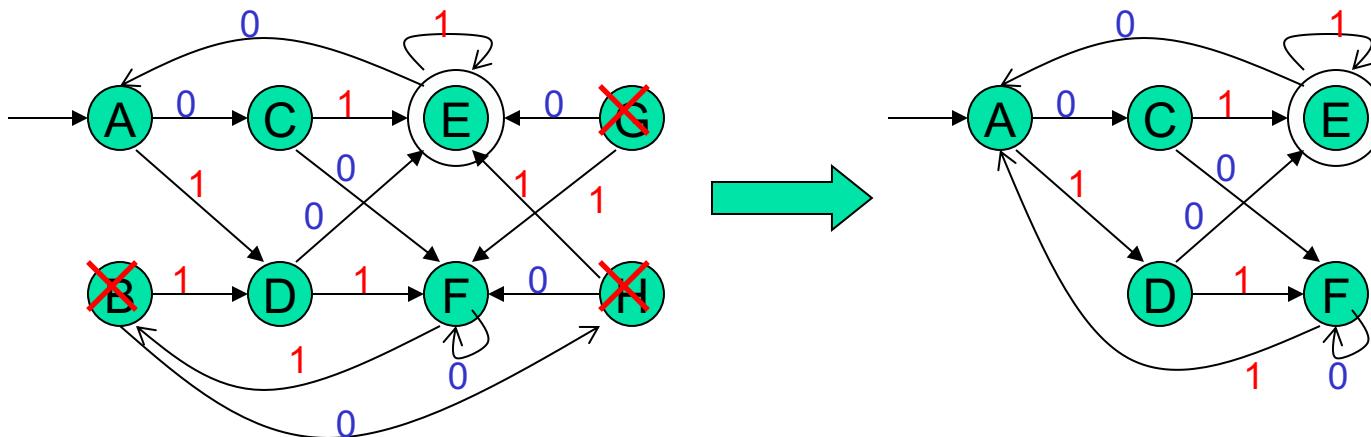
A	=						
B	=						
C	X	X		=			
D	X	X	X	=			
E	X	X	X	X	=		
F	X	X	X	X	X	=	
G	X	X	X	=	X	X	=
H	X	X	=	X	X	X	=

A B C D E F G H

**Equivalences:**

- A=B
- C=H
- D=G

# Table Filling Algorithm - step by step

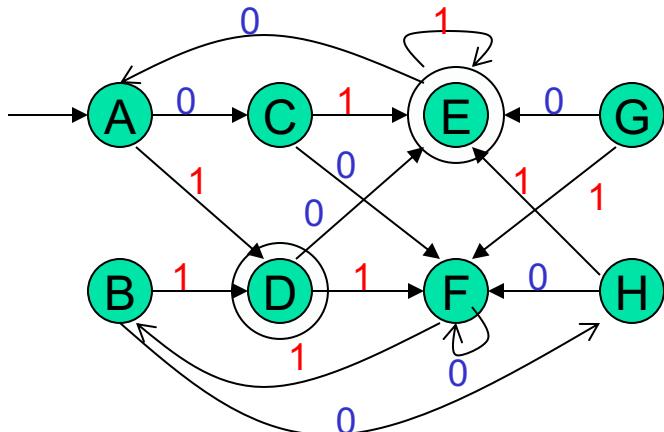


Retrain only one copy for  
each equivalence set of states

## Equivalences:

- A=B
- C=H
- D=G

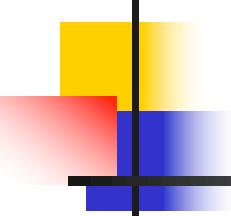
# Table Filling Algorithm – special case



Q) What happens if the input DFA has more than one final state?  
Can all final states initially be treated as equivalent to one another?

A	=							
B		=						
C			=					
D				=				
E				?	=			
F						=		
G							=	
H								=
	A	B	C	D	E	F	G	H

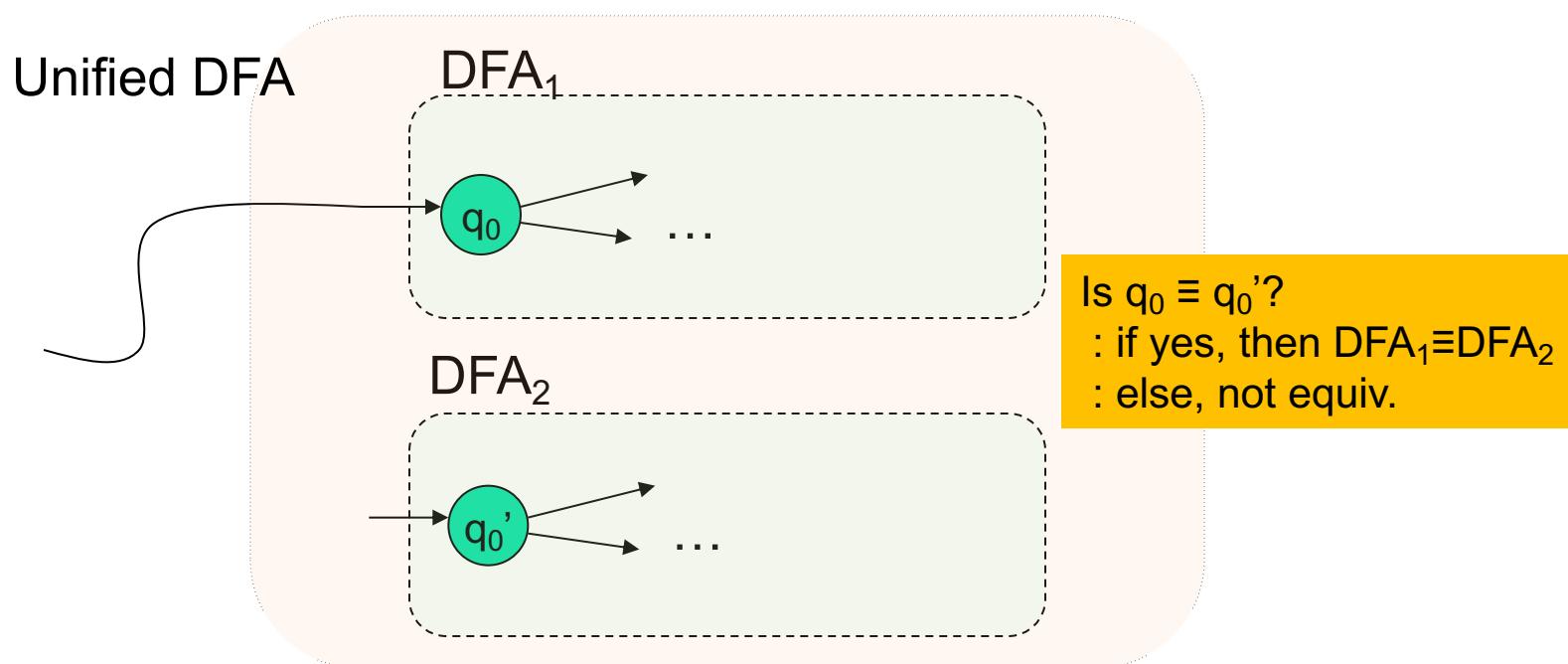
Putting it all together ...



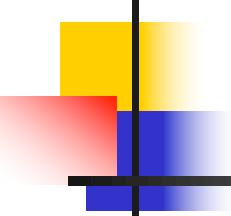
## How to minimize a DFA?

- Goal: Minimize the number of states in a DFA
  - Algorithm:
    - 1. Eliminate states unreachable from the start state
    - 2. Identify and remove equivalent states
    - 3. Output the resultant DFA
- Depth-first traversal from the start state
- Table filling algorithm

# Are Two DFAs Equivalent?



1. Make a new dummy DFA by just putting together both DFAs
2. Run table-filling algorithm on the unified DFA
3. *IF* the start states of both DFAs are found to be equivalent,  
    *THEN*:  $\text{DFA}_1 \equiv \text{DFA}_2$   
    *ELSE*: different



# Summary

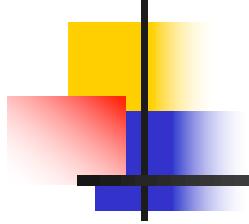
---

- How to prove languages are not regular?
  - Pumping lemma & its applications
- Closure properties of regular languages
- Simplification of DFAs
  - How to remove unreachable states?
  - How to identify and collapse equivalent states?
  - How to minimize a DFA?
  - How to tell whether two DFAs are equivalent?

# Context-Free Languages & Grammars (CFLs & CFGs)

---

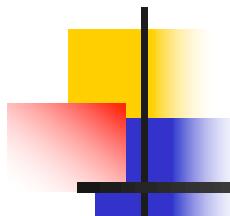
Reading: Chapter 5



# Not all languages are regular

---

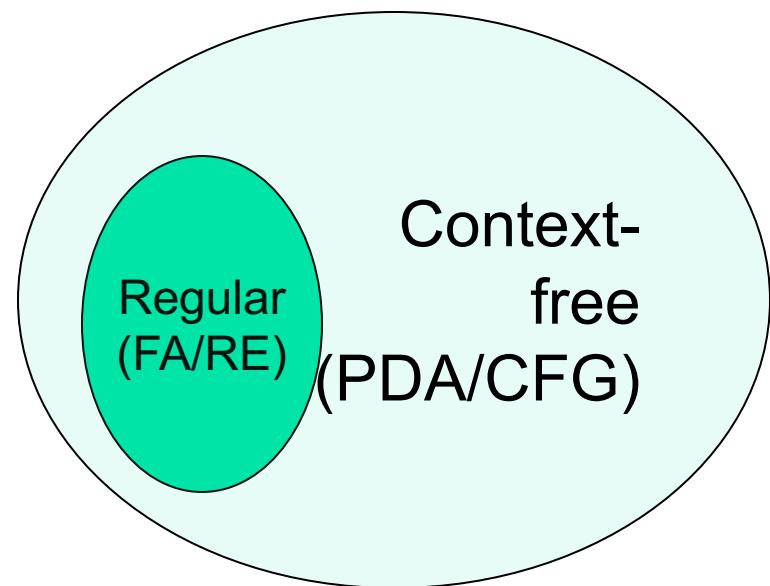
- So what happens to the languages which are not regular?
- Can we still come up with a language recognizer?
  - i.e., something that will accept (or reject) strings that belong (or do not belong) to the language?

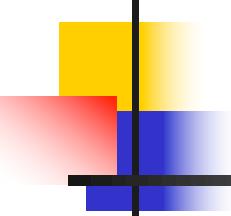


# Context-Free Languages

---

- A language class larger than the class of regular languages
- Supports natural, recursive notation called “context-free grammar”
- Applications:
  - Parse trees, compilers
  - XML





# An Example

- A palindrome is a word that reads identical from both ends
  - E.g., madam, redivider, malayalam, 010010010
- Let  $L = \{ w \mid w \text{ is a binary palindrome}\}$
- Is  $L$  regular?
  - No.
  - Proof:
    - Let  $w=0^N10^N$  (assuming  $N$  to be the p/l constant)
    - By Pumping lemma,  $w$  can be rewritten as  $xyz$ , such that  $xy^kz$  is also  $L$  (for any  $k \geq 0$ )
    - But  $|xy| \leq N$  and  $y \neq \epsilon$
    - $\Rightarrow y = 0^+$
    - $\Rightarrow xy^kz$  will NOT be in  $L$  for  $k=0$
    - $\Rightarrow$  Contradiction

# But the language of palindromes...

is a CFL, because it supports recursive substitution (in the form of a CFG)

- This is because we can construct a “grammar” like this:

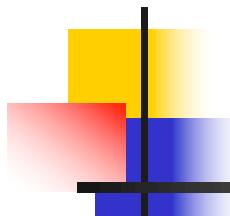
- Productions
1.  $A \implies \epsilon$
  2.  $A \implies 0$
  3.  $A \implies 1$
  4.  $A \implies 0A0$
  5.  $A \implies 1A1$

Terminal

Variable or non-terminal

Same as:  
 $A \Rightarrow 0A0 \mid 1A1 \mid 0 \mid 1 \mid \epsilon$

How does this grammar work?



# How does the CFG for palindromes work?

An input string belongs to the language (i.e., accepted) iff it can be generated by the CFG

- Example:  $w=01110$
- G can generate w as follows:

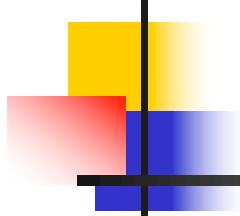
1.  $A \Rightarrow 0A0$
2.  $\Rightarrow 01A10$
3.  $\Rightarrow 01110$

G:

$A \Rightarrow 0A0 \mid 1A1 \mid 0 \mid 1 \mid \epsilon$

## Generating a string from a grammar:

1. Pick and choose a sequence of productions that would allow us to generate the string.
2. At every step, substitute one variable with one of its productions.



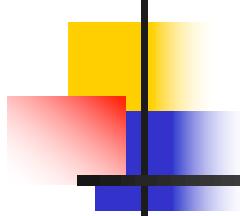
# Context-Free Grammar: Definition

- A context-free grammar  $G=(V,T,P,S)$ , where:
  - $V$ : set of variables or non-terminals
  - $T$ : set of terminals (= alphabet  $\cup \{\epsilon\}$ )
  - $P$ : set of *productions*, each of which is of the form
$$V ==> \alpha_1 \mid \alpha_2 \mid \dots$$
    - Where each  $\alpha_i$  is an arbitrary string of variables and terminals
  - $S ==>$  start variable

CFG for the language of binary palindromes:

$$G=(\{A\}, \{0,1\}, P, A)$$

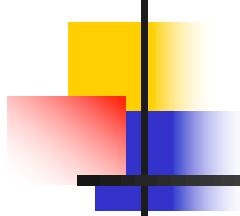
$$P: A ==> 0A0 \mid 1A1 \mid 0 \mid 1 \mid \epsilon$$



# More examples

---

- Parenthesis matching in code
- Syntax checking
- In scenarios where there is a general need for:
  - Matching a symbol with another symbol, or
  - Matching a count of one symbol with that of another symbol, or
  - Recursively substituting one symbol with a string of other symbols



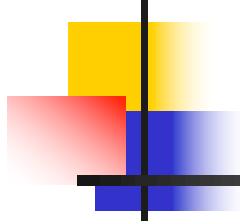
## Example #2

---

- Language of balanced parenthesis  
e.g.,  $()((((()))))(((())))\dots$
- CFG?

G:  
 $S \Rightarrow (S) \mid SS \mid \epsilon$

How would you “interpret” the string “ $((((())())())()$ ” using this grammar?  
 $S \rightarrow SS \rightarrow (S) \rightarrow (SS) \rightarrow ((S)SS) \rightarrow (((S))(S)(S)) \rightarrow (((((())())())())()$



## Example #3

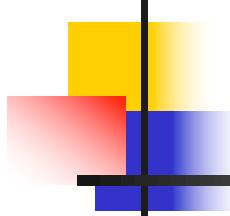
---

- A grammar for  $L = \{0^m 1^n \mid m \geq n\}$
- CFG?

G:  
 $S \Rightarrow 0S1 \mid A$   
 $A \Rightarrow 0A \mid \epsilon$

How would you interpret the string “00000111” using this grammar?

$S \rightarrow 0S1 \rightarrow 00S1 \rightarrow 000S1 \rightarrow 0000A1 \rightarrow 00000A1 \rightarrow 00000111$



## Example #4

---

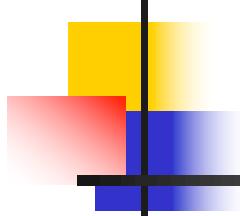
A program containing **if-then(-else)** statements

**if Condition then Statement else Statement**

(Or)

**if Condition then Statement**

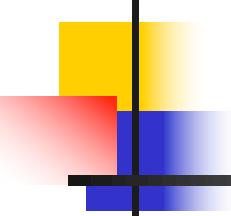
CFG?



## More examples

---

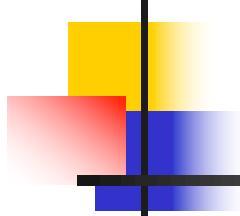
- $L_1 = \{0^n \mid n \geq 0\}$
- $L_2 = \{0^n \mid n \geq 1\}$
- $L_3 = \{0^i 1^j 2^k \mid i=j \text{ or } j=k, \text{ where } i, j, k \geq 0\}$
- $L_4 = \{0^i 1^j 2^k \mid i=j \text{ or } i=k, \text{ where } i, j, k \geq 1\}$



# Applications of CFLs & CFGs

---

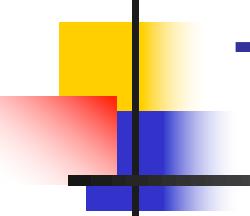
- Compilers use parsers for syntactic checking
- Parsers can be expressed as CFGs
  - 1. Balancing parenthesis:
    - $B \Rightarrow BB \mid (B) \mid Statement$
    - $Statement \Rightarrow \dots$
  - 2. If-then-else:
    - $S \Rightarrow SS \mid if\ Condition\ then\ Statement\ else\ Statement \mid if\ Condition\ then\ Statement \mid Statement$
    - $Condition \Rightarrow \dots$
    - $Statement \Rightarrow \dots$
  - 3. C parenthesis matching { ... }
  - 4. Pascal begin-end matching
  - 5. YACC (Yet Another Compiler-Compiler)



# More applications

---

- Markup languages
  - Nested Tag Matching
    - HTML
      - <html> ... <p> ... <a href=...> ... </a> </p> ... </html>
    - XML
      - <PC> ... <MODEL> ... </MODEL> .. <RAM> ... </RAM> ... </PC>



# Tag-Markup Languages

Roll ==> <ROLL> Class Students </ROLL>

Class ==> <CLASS> Text </CLASS>

Text ==> Char Text | Char

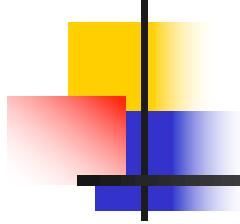
Char ==> a | b | ... | z | A | B | .. | Z

Students ==> Student Students |  $\epsilon$

Student ==> <STUD> Text </STUD>

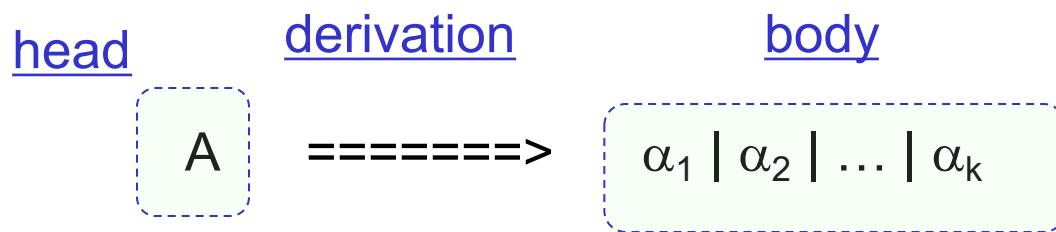
Here, the left hand side of each production denotes one non-terminals  
(e.g., “Roll”, “Class”, etc.)

Those symbols on the right hand side for which no productions (i.e.,  
substitutions) are defined are terminals (e.g., ‘a’, ‘b’, ‘|’, ‘<’, ‘>’, “ROLL”,  
etc.)



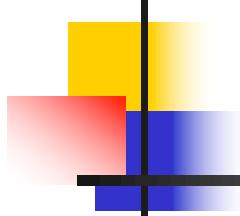
# Structure of a production

---



The above is same as:

1.  $A \Rightarrow \alpha_1$
2.  $A \Rightarrow \alpha_2$
3.  $A \Rightarrow \alpha_3$
- ...
- K.  $A \Rightarrow \alpha_k$



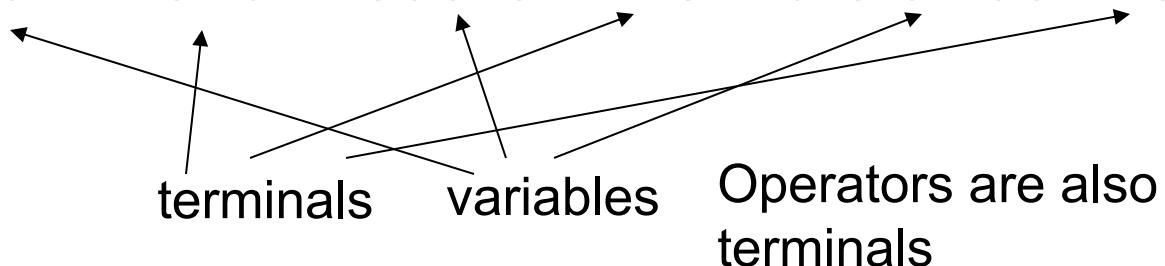
# CFG conventions

---

- Terminal symbols  $\leqslant a, b, c\dots$
- Non-terminal symbols  $\leqslant A, B, C, \dots$
- Terminal or non-terminal symbols  $\leqslant X, Y, Z$
- Terminal strings  $\leqslant w, x, y, z$
- Arbitrary strings of terminals and non-terminals  $\leqslant \alpha, \beta, \gamma, \dots$

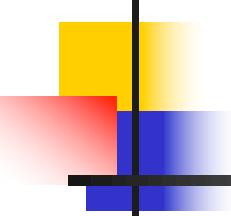
# Syntactic Expressions in Programming Languages

*result = a\*b + score + 10 \* distance + c*



Regular languages have only terminals

- Reg expression = [a-z][a-z0-1]\*
- If we allow only letters a & b, and 0 & 1 for constants (for simplification)
  - Regular expression = (a+b)(a+b+0+1)\*



# String membership

How to say if a string belong to the language defined by a CFG?

1. Derivation
  - Head to body
2. Recursive inference
  - Body to head

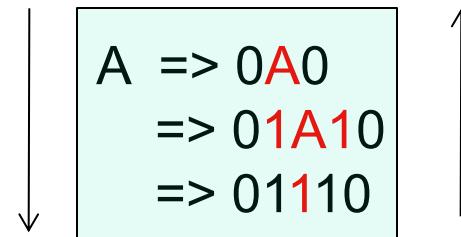
Example:

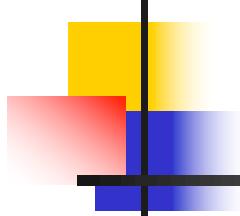
- $w = 01110$
- Is  $w$  a palindrome?



Both are equivalent forms

G:  
 $A \Rightarrow 0A0 \mid 1A1 \mid 0 \mid 1 \mid \epsilon$

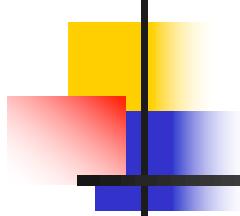




# Simple Expressions...

---

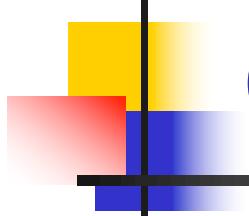
- We can write a CFG for accepting simple expressions
- $G = (V, T, P, S)$ 
  - $V = \{E, F\}$
  - $T = \{0, 1, a, b, +, *, (, )\}$
  - $S = \{E\}$
  - $P:$ 
    - $E \implies E+E \mid E^*E \mid (E) \mid F$
    - $F \implies aF \mid bF \mid 0F \mid 1F \mid a \mid b \mid 0 \mid 1$



# Generalization of derivation

---

- Derivation is *head ==> body*
  
- $A ==> X$                     (A derives X in a single step)
- $A ==>^*_G X$             (A derives X in a multiple steps)
  
- Transitivity:  
    IF  $A ==>^*_G B$ , and  $B ==>^*_G C$ , THEN  $A ==>^*_G C$



# Context-Free Language

---

- The language of a CFG,  $G=(V,T,P,S)$ , denoted by  $L(G)$ , is the set of terminal strings that have a derivation from the start variable  $S$ .
  - $L(G) = \{ w \text{ in } T^* \mid S \Rightarrow^* G w \}$

# Left-most & Right-most Derivation Styles

G:  
 $E \Rightarrow E+E \mid E^*E \mid (E) \mid F$   
 $F \Rightarrow aF \mid bF \mid 0F \mid 1F \mid \epsilon$

Derive the string  $a^*(ab+10)$  from G:

$$E \stackrel{*}{\Rightarrow}_G a^*(ab+10)$$

Left-most derivation:

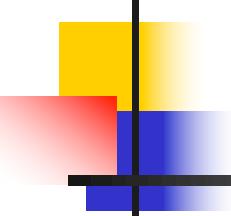
Always substitute leftmost variable

- E
- $\Rightarrow E^*E$
- $\Rightarrow F^*E$
- $\Rightarrow aF^*E$
- $\Rightarrow a^*E$
- $\Rightarrow a^*(E)$
- $\Rightarrow a^*(E+E)$
- $\Rightarrow a^*(F+E)$
- $\Rightarrow a^*(aF+E)$
- $\Rightarrow a^*(abF+E)$
- $\Rightarrow a^*(ab+E)$
- $\Rightarrow a^*(ab+F)$
- $\Rightarrow a^*(ab+1F)$
- $\Rightarrow a^*(ab+10F)$
- $\Rightarrow a^*(ab+10)$

- E
- $\Rightarrow E^*E$
- $\Rightarrow E^*(E)$
- $\Rightarrow E^*(E+E)$
- $\Rightarrow E^*(E+F)$
- $\Rightarrow E^*(E+1F)$
- $\Rightarrow E^*(E+10F)$
- $\Rightarrow E^*(E+10)$
- $\Rightarrow E^*(F+10)$
- $\Rightarrow E^*(aF+10)$
- $\Rightarrow E^*(abF+0)$
- $\Rightarrow E^*(ab+10)$
- $\Rightarrow F^*(ab+10)$
- $\Rightarrow aF^*(ab+10)$
- $\Rightarrow a^*(ab+10)$

Right-most derivation:

Always substitute rightmost variable



# Leftmost vs. Rightmost derivations

---

Q1) For every leftmost derivation, there is a rightmost derivation, and vice versa. True or False?

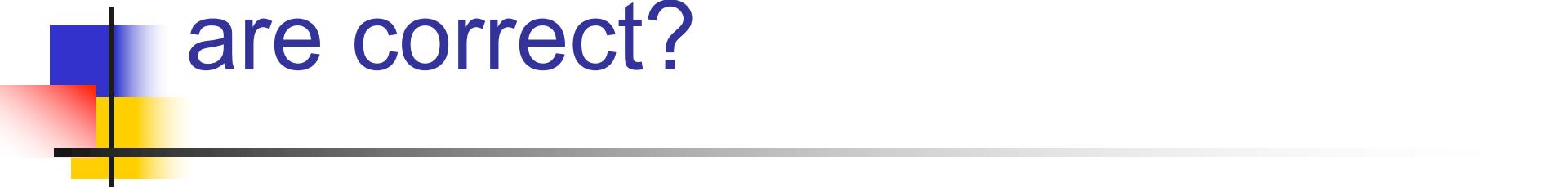
True - will use parse trees to prove this

Q2) Does every word generated by a CFG have a leftmost and a rightmost derivation?

Yes – easy to prove (reverse direction)

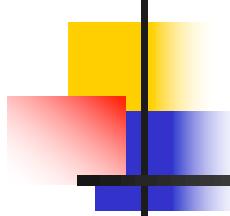
Q3) Could there be words which have more than one leftmost (or rightmost) derivation?

Yes – depending on the grammar



# How to prove that your CFGs are correct?

(using induction)



# CFG & CFL

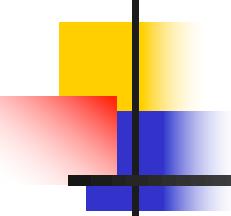
G<sub>pal</sub>:

A => 0A0 | 1A1 | 0 | 1 | ε

- Theorem: A string w in  $(0+1)^*$  is in  $L(G_{\text{pal}})$ , if and only if, w is a palindrome.
- Proof:
  - Use induction
    - on string length for the IF part
    - On length of derivation for the ONLY IF part



# Parse trees



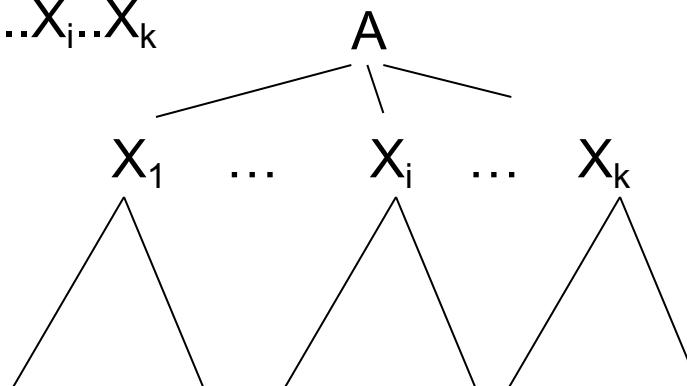
# Parse Trees

---

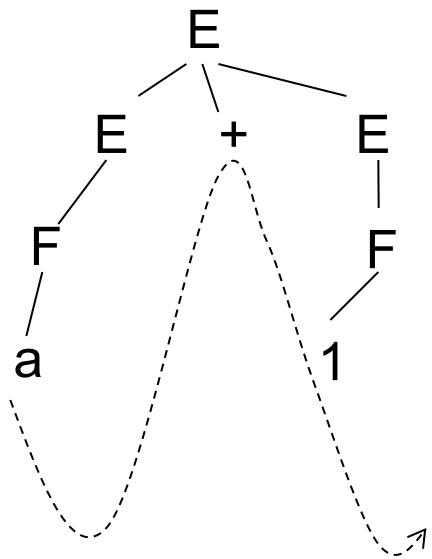
- Each CFG can be represented using a *parse tree*:
    - Each internal node is labeled by a variable in  $V$
    - Each leaf is terminal symbol
    - For a production,  $A \Rightarrow X_1 X_2 \dots X_k$ , then any internal node labeled  $A$  has  $k$  children which are labeled from  $X_1, X_2, \dots, X_k$  from left to right
- 

Parse tree for production and all other subsequent productions:

$$A \Rightarrow X_1 \dots X_i \dots X_k$$



# Examples



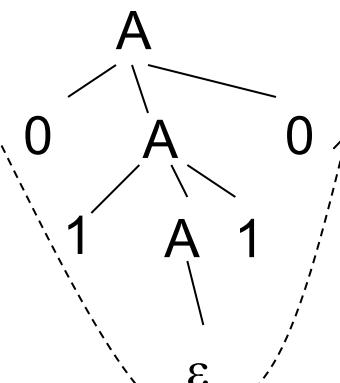
Parse tree for a + 1

[G:

$E \Rightarrow E+E \mid E^*E \mid (E) \mid F$

$F \Rightarrow aF \mid bF \mid 0F \mid 1F \mid 0 \mid 1 \mid a \mid b$

Recursive inference ↑



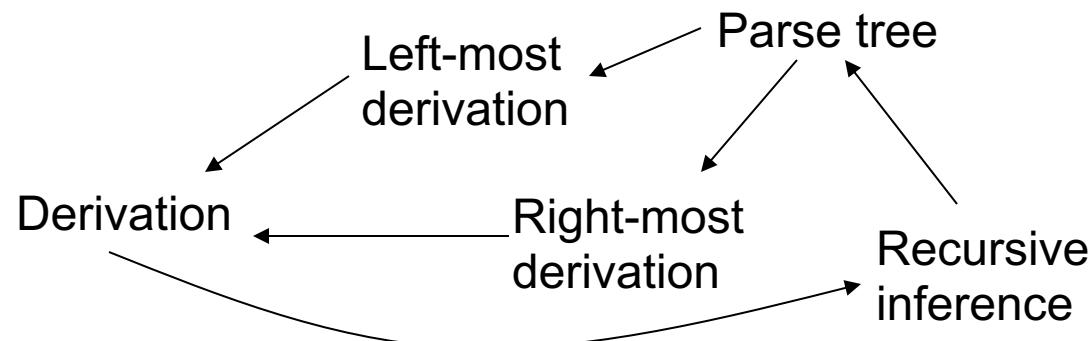
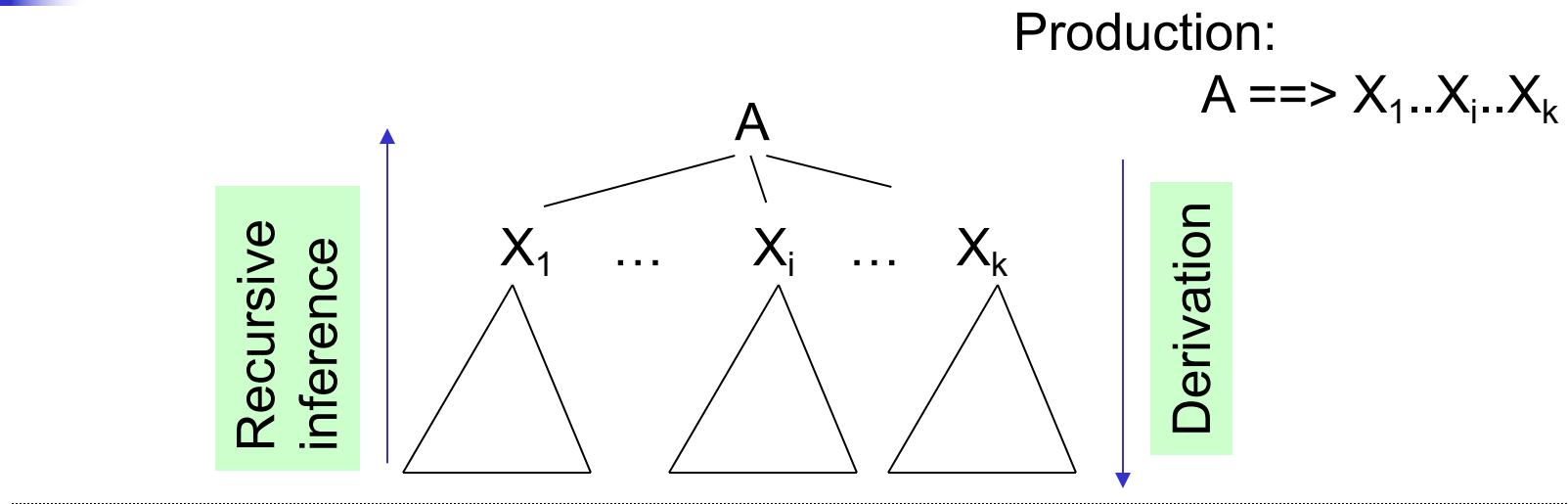
Parse tree for 0110

[G:

$A \Rightarrow 0A0 \mid 1A1 \mid 0 \mid 1 \mid \epsilon$

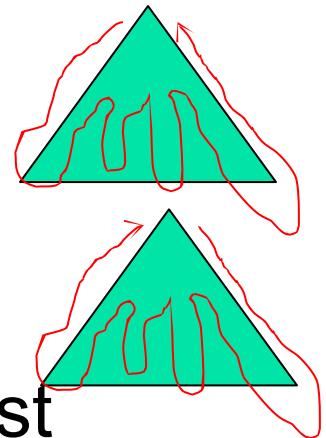
Derivation ↓

# Parse Trees, Derivations, and Recursive Inferences



# Interchangeability of different CFG representations

- Parse tree ==> left-most derivation
  - DFS left to right
- Parse tree ==> right-most derivation
  - DFS right to left
- ==> left-most derivation == right-most derivation
- Derivation ==> Recursive inference
  - Reverse the order of productions
- Recursive inference ==> Parse trees
  - bottom-up traversal of parse tree



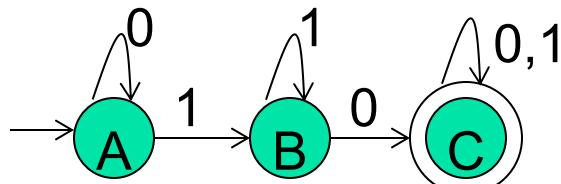
# Connection between CFLs and RLS

What kind of grammars result for regular languages?

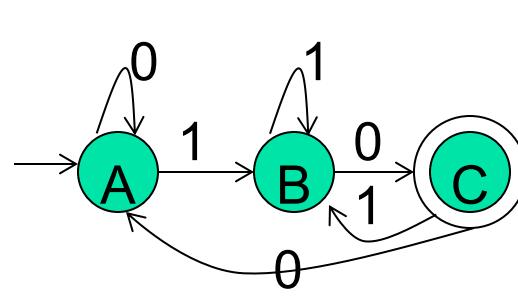
# CFLs & Regular Languages

- A CFG is said to be *right-linear* if all the productions are one of the following two forms:  $A \Rightarrow wB$  (or)  $A \Rightarrow w$ 
  - Where:
    - A & B are variables,
    - w is a string of terminals
- Theorem 1: Every right-linear CFG generates a regular language
- Theorem 2: Every regular language has a right-linear grammar
- Theorem 3: Left-linear CFGs also represent RLs

# Some Examples



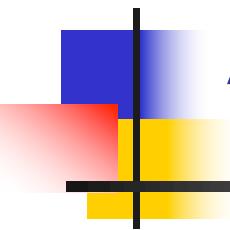
Right linear CFG?



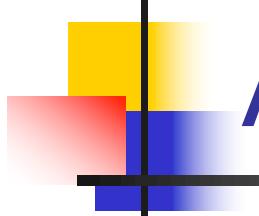
Right linear CFG?

➤  $A \Rightarrow 01B \mid C$   
 $B \Rightarrow 11B \mid 0C \mid 1A$   
 $C \Rightarrow 1A \mid 0 \mid 1$

Finite Automaton?



# Ambiguity in CFGs and CFLs



# Ambiguity in CFGs

- A CFG is said to be *ambiguous* if there exists a string which has more than one left-most derivation

Example:

$$\begin{aligned} S & \Rightarrow AS \mid \epsilon \\ A & \Rightarrow A1 \mid 0A1 \mid 01 \end{aligned}$$

Input string: 00111

Can be derived in two ways

LM derivation #1:

$$\begin{aligned} S & \Rightarrow AS \\ & \Rightarrow 0A1S \\ & \Rightarrow 0A11S \\ & \Rightarrow 00111S \\ & \Rightarrow 00111 \end{aligned}$$

LM derivation #2:

$$\begin{aligned} S & \Rightarrow AS \\ & \Rightarrow A1S \\ & \Rightarrow 0A11S \\ & \Rightarrow 00111S \\ & \Rightarrow 00111 \end{aligned}$$

# Why does ambiguity matter?

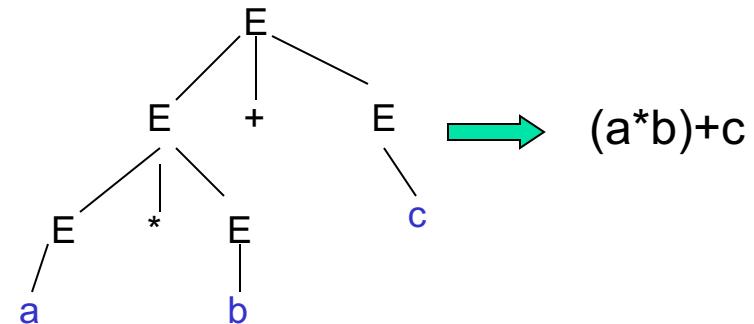
$E \Rightarrow E + E \mid E * E \mid (E) \mid a \mid b \mid c \mid 0 \mid 1$

Values are different !!!

*string = a \* b + c*

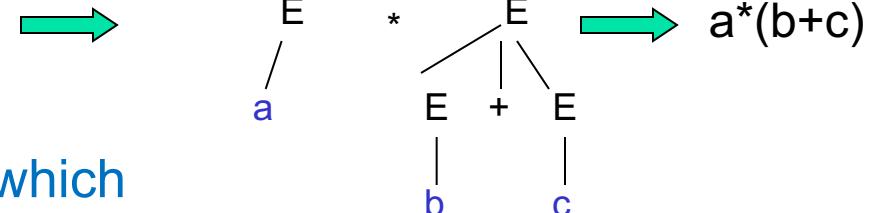
- LM derivation #1:

•  $E \Rightarrow E + E \Rightarrow E * E + E \Rightarrow^* a * b + c$

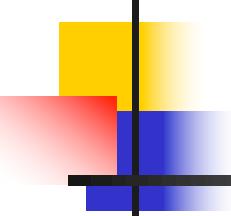


- LM derivation #2

•  $E \Rightarrow E * E \Rightarrow a * E \Rightarrow a * E + E \Rightarrow^* a * b + c$



The calculated value depends on which of the two parse trees is actually used.



# Removing Ambiguity in Expression Evaluations

- It MAY be possible to remove ambiguity for some CFLs
  - E.g., in a CFG for expression evaluation by imposing rules & restrictions such as precedence
  - This would imply rewrite of the grammar

Modified unambiguous version:

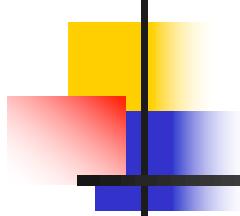
- Precedence: (), \*, +

$$\begin{aligned}E &\Rightarrow E + T \mid T \\T &\Rightarrow T * F \mid F \\F &\Rightarrow I \mid (E) \\I &\Rightarrow a \mid b \mid c \mid 0 \mid 1\end{aligned}$$

Ambiguous version:

$$E \Rightarrow E + E \mid E * E \mid (E) \mid a \mid b \mid c \mid 0 \mid 1$$

How will this avoid ambiguity?



# Inherently Ambiguous CFLs

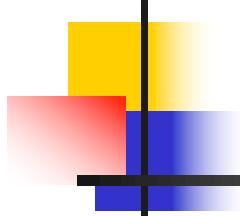
---

- However, for some languages, it may not be possible to remove ambiguity
- A CFL is said to be *inherently ambiguous* if every CFG that describes it is ambiguous

Example:

- $L = \{ a^n b^n c^m d^m \mid n, m \geq 1 \} \cup \{ a^n b^m c^m d^n \mid n, m \geq 1 \}$
- $L$  is inherently ambiguous
- Why?

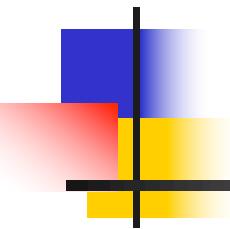
Input string:  $a^n b^n c^n d^n$



# Summary

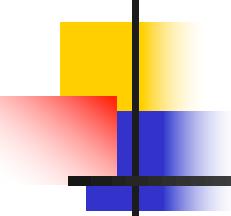
---

- Context-free grammars
- Context-free languages
- Productions, derivations, recursive inference, parse trees
- Left-most & right-most derivations
- Ambiguous grammars
- Removing ambiguity
- CFL/CFG applications
  - parsers, markup languages



# Pushdown Automata (PDA)

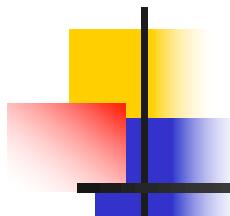
Reading: Chapter 6



# Introduction

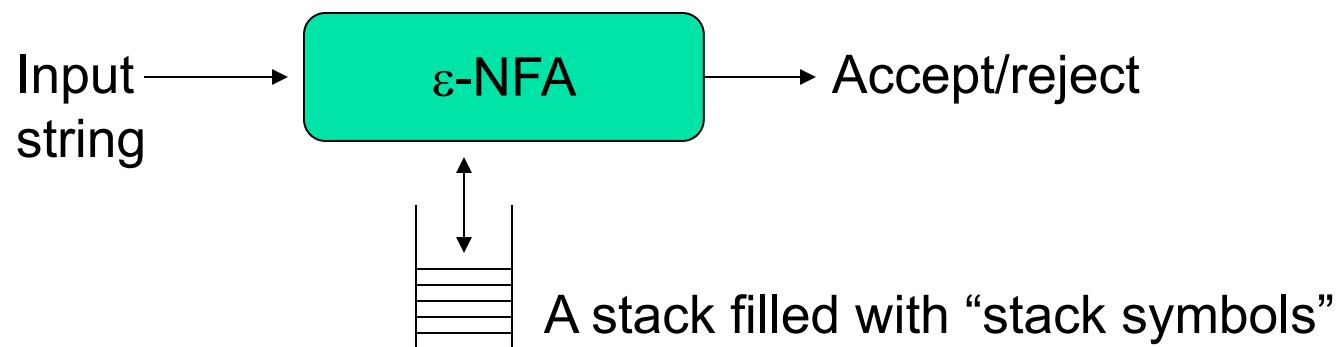
---

- Pushdown automata are used to determine what can be computed by machines.
- More capable than finite-state machines but less capable than Turing machines.
- A type of automaton that uses a stack.
- A pushdown automaton (PDA) differs from a finite state machine in two ways:
  - It can use the top of the stack to decide which transition to take.
  - It can manipulate the stack as part of performing a transition.



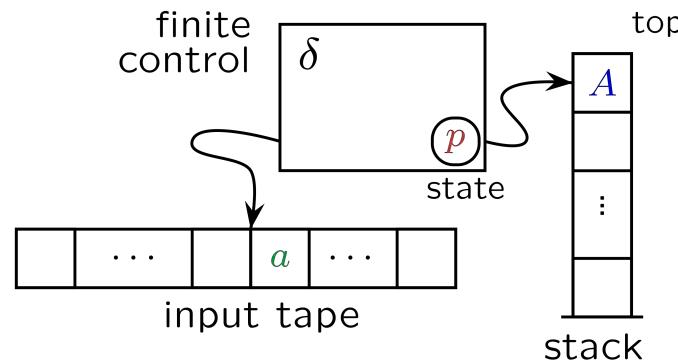
# PDA - the automata for CFLs

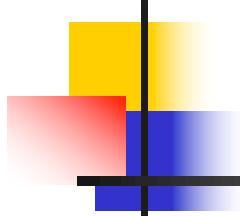
- What is?
  - FA to Reg Lang, PDA is to CFL
- PDA == [  $\epsilon$ -NFA + “a stack” ]
- Why a stack?



# PDA

- PDA reads a given input string from left to right.
- In each step, it chooses a transition by indexing a table by input symbol, current state, and the symbol at the top of the stack.
- A PDA can also manipulate the stack, as part of performing a transition.
  - The manipulation can be to push a particular symbol to the top of the stack
  - Pop off the top of the stack.
  - The automaton can alternatively ignore the stack, and leave it as it is.





# DPDA & NPDA

---

- Given an input symbol, current state, and stack symbol, the automaton can follow a transition to another state, and optionally manipulate (push or pop) the stack.
  - If at most one such transition action is possible, then the automaton is called a deterministic pushdown automaton (DPDA).
  - If several actions are possible, then the automaton is called nondeterministic, (NPDA).
- In case of NPDA, if one of the action leads to an accepting state after reading the complete input string, then language is accepted by the automaton.

# Pushdown Automata -

## Definition

- A PDA  $P := ( Q, \Sigma, \Gamma, \delta, q_0, Z_0, F )$ :
  - $Q$ : states of the  $\epsilon$ -NFA
  - $\Sigma$ : input alphabet
  - $\Gamma$ : stack symbols
  - $\delta$ : transition function
  - $q_0$ : start state
  - $Z_0$ : Initial stack top symbol
  - $F$ : Final/accepting states

$$\delta : Q \times \Sigma \times \Gamma \Rightarrow Q \times \Gamma$$

*old state*    *input symb.*    *Stack top*    *new state(s)* *new Stack top(s)*

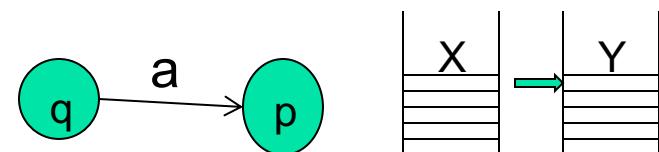
# $\delta$ : The Transition Function

$$\delta(q, a, X) = \{(p, Y), \dots\}$$

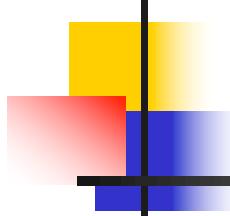
- 1. state transition from  $q$  to  $p$
- 2.  $a$  is the next input symbol
- 3.  $X$  is the current stack *top* symbol
- 4.  $Y$  is the replacement for  $X$ ; it is in  $\Gamma^*$  (a string of stack symbols)

- i. Set  $Y = \epsilon$  for: Pop( $X$ )
- ii. If  $Y=X$ : stack top is unchanged
- iii. If  $Y=Z_1Z_2\dots Z_k$ :  $X$  is popped and is replaced by  $Y$  in reverse order (i.e.,  $Z_1$  will be the new stack top)

**Non-determinism**



$Y = ?$	Action
$Y=\epsilon$	Pop( $X$ )
$Y=X$	Pop( $X$ ) Push( $X$ )
$Y=Z_1Z_2\dots Z_k$	Pop( $X$ ) Push( $Z_k$ ) Push( $Z_{k-1}$ ) ... Push( $Z_2$ ) Push( $Z_1$ )



# Example

---

Let  $L_{wwr} = \{ww^R \mid w \text{ is in } (0+1)^*\}$

- CFG for  $L_{wwr}$  :  $S \Rightarrow 0S0 \mid 1S1 \mid \epsilon$

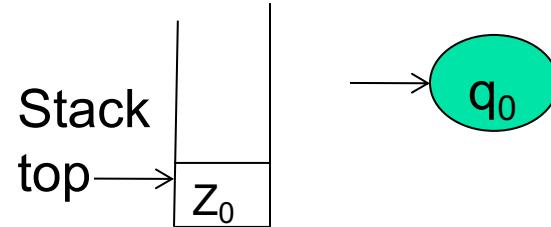
- PDA for  $L_{wwr}$  :

- $P := (Q, \Sigma, \Gamma, \delta, q_0, Z_0, F)$

$$= ( \{q_0, q_1, q_2\}, \{0, 1\}, \{0, 1, Z_0\}, \delta, q_0, Z_0, \{q_2\} )$$

# PDA for $L_{wwr}$

Initial state of the PDA:



$$\begin{array}{ll} 1. & \delta(q_0, 0, Z_0) = \{(q_0, 0Z_0)\} \\ 2. & \delta(q_0, 1, Z_0) = \{(q_0, 1Z_0)\} \end{array} \quad \} \quad \text{First symbol push on stack}$$

$$\begin{array}{ll} 3. & \delta(q_0, 0, 0) = \{(q_0, 00)\} \\ 4. & \delta(q_0, 0, 1) = \{(q_0, 01)\} \\ 5. & \delta(q_0, 1, 0) = \{(q_0, 10)\} \\ 6. & \delta(q_0, 1, 1) = \{(q_0, 11)\} \end{array} \quad } \quad \text{Grow the stack by pushing new symbols on top of old (w-part)}$$

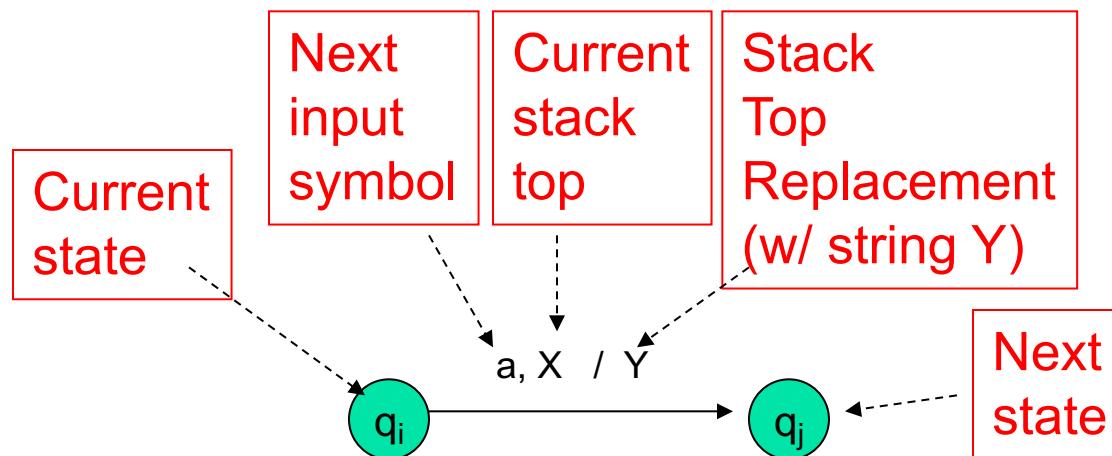
$$\begin{array}{ll} 7. & \delta(q_0, \varepsilon, 0) = \{(q_1, 0)\} \\ 8. & \delta(q_0, \varepsilon, 1) = \{(q_1, 1)\} \\ 9. & \delta(q_0, \varepsilon, Z_0) = \{(q_1, Z_0)\} \end{array} \quad } \quad \text{Switch to popping mode, nondeterministically (boundary between w and w<sup>R</sup>)}$$

$$\begin{array}{ll} 10. & \delta(q_1, 0, 0) = \{(q_1, \varepsilon)\} \\ 11. & \delta(q_1, 1, 1) = \{(q_1, \varepsilon)\} \end{array} \quad } \quad \text{Shrink the stack by popping matching symbols (w<sup>R</sup>-part)}$$

$$12. \quad \delta(q_1, \varepsilon, Z_0) = \{(q_2, Z_0)\} \quad } \quad \text{Enter acceptance state}$$

# PDA as a state diagram

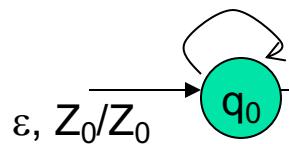
$$\delta(q_i, a, X) = \{(q_j, Y)\}$$



# PDA for $L_{wwr}$ : Transition Diagram

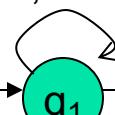
Grow stack

0,  $Z_0/Z_0$   
1,  $Z_0/Z_0$   
0, 0/0  
0, 1/01  
1, 0/10  
1, 1/11

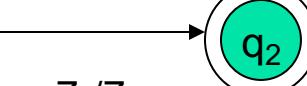


Pop stack for matching symbols

0, 0/  $\epsilon$   
1, 1/  $\epsilon$



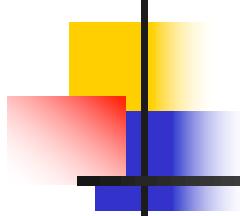
$\Sigma = \{0, 1\}$   
 $\Gamma = \{Z_0, 0, 1\}$   
 $Q = \{q_0, q_1, q_2\}$



Go to acceptance

Switch to popping mode

This would be a non-deterministic PDA



# PDA's Instantaneous Description (ID)

A PDA has a configuration at any given instance:  
 $(q, w, y)$

- $q$  - current state
- $w$  - remainder of the input (i.e., unconsumed part)
- $y$  - current stack contents as a string from top to bottom of stack

---

If  $\delta(q, a, X) = \{(p, A)\}$  is a transition, then the following are also true:

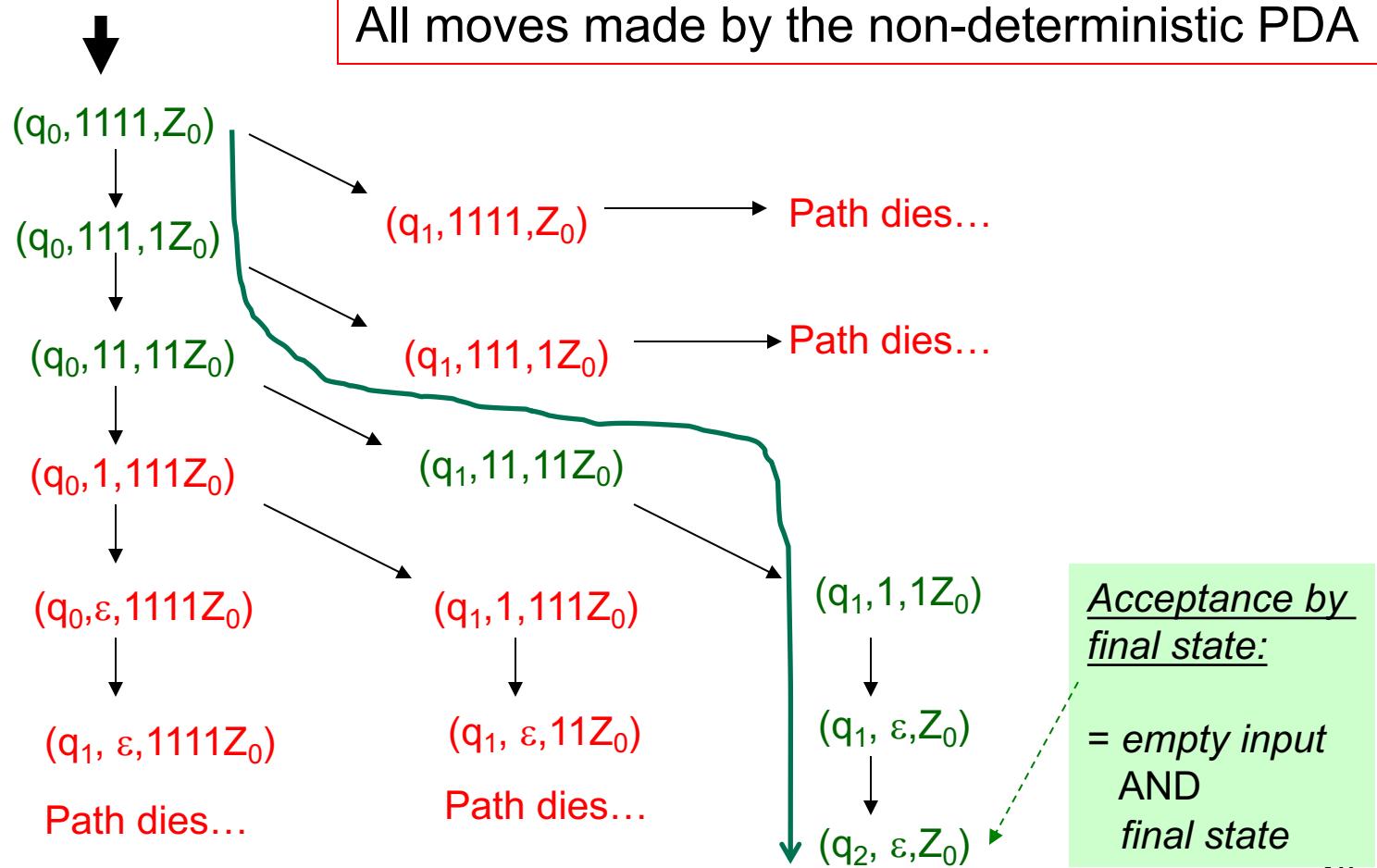
- $(q, a, X) \dashv (p, \epsilon, A)$
- $(q, aw, XB) \dashv (p, w, AB)$

---

$\dashv$  sign is called a “turnstile notation” and represents one move

$\dashv^*$  sign represents a sequence of moves

# How does the PDA for $L_{wwr}$ work on input “1111”?



There are two types of PDAs that one can design:  
those that accept by final state or by empty stack

# Acceptance by...

## ■ PDAs that accept by **final state**:

- For a PDA  $P$ , the language accepted by  $P$ , denoted by  $L(P)$  by *final state*, is:
  - $\{w \mid (q_0, w, Z_0) \xrightarrow{\cdot\cdot\cdot}^* (q, \varepsilon, A)\}$ , s.t.,  $q \in F$

Checklist:

- input exhausted?
- in a final state?

## ■ PDAs that accept by **empty stack**:

- For a PDA  $P$ , the language accepted by  $P$ , denoted by  $N(P)$  by *empty stack*, is:
  - $\{w \mid (q_0, w, Z_0) \xrightarrow{\cdot\cdot\cdot}^* (q, \varepsilon, \varepsilon)\}$ , for any  $q \in Q$ .

Q) Does a PDA that accepts by empty stack  
need any final state specified in the design?

Checklist:

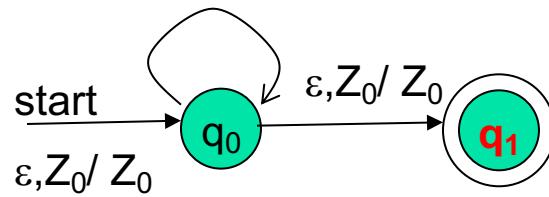
- input exhausted?
- is the stack empty?

# Example: L of balanced parenthesis

PDA that accepts by final state

$$\mathbf{P}_F:$$

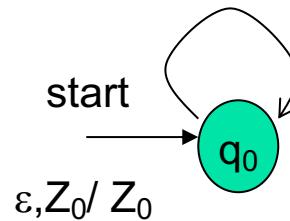
$(, Z_0 / ( Z_0$   
 $(, ( / ( ($   
 $), ( / \varepsilon$



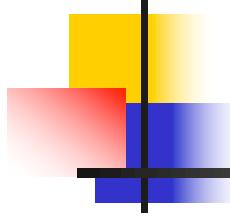
An equivalent PDA that accepts by empty stack

$$\mathbf{P}_N:$$

$(, Z_0 / ( Z_0$   
 $(, ( / ( ($   
 $), ( / \varepsilon$   
 $\varepsilon, Z_0 / \varepsilon$



*How will these two PDAs work on the input: ((( ))( )) ()*



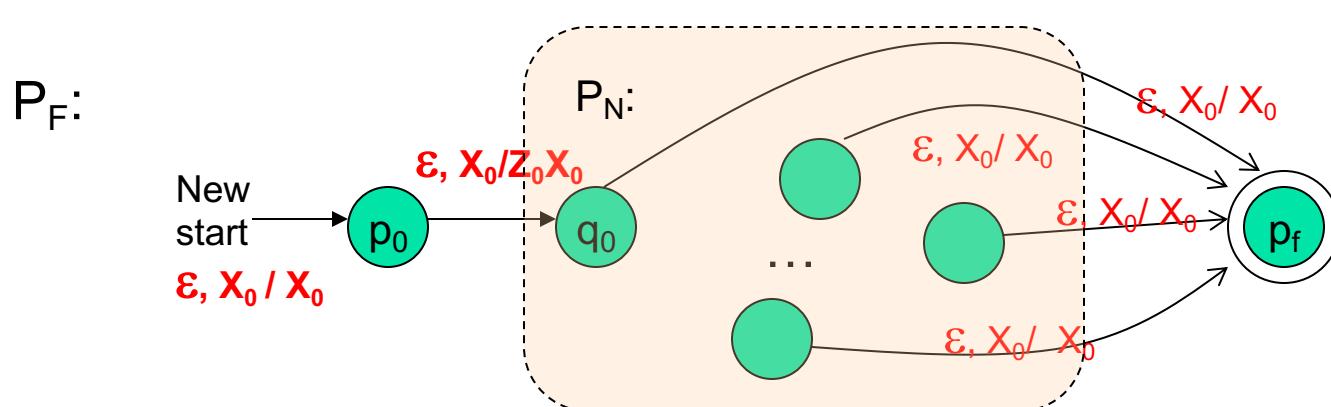
# PDAs accepting by final state and empty stack are equivalent

---

- $P_F \leq PDA$  accepting by final state
  - $P_F = (Q_F, \Sigma, \Gamma, \delta_F, q_0, Z_0, F)$
- $P_N \leq PDA$  accepting by empty stack
  - $P_N = (Q_N, \Sigma, \Gamma, \delta_N, q_0, Z_0)$
- Theorem:
  - $(P_N \Rightarrow P_F)$  For every  $P_N$ , there exists a  $P_F$  s.t.  $L(P_F) = L(P_N)$
  - $(P_F \Rightarrow P_N)$  For every  $P_F$ , there exists a  $P_N$  s.t.  $L(P_F) = L(P_N)$

## $P_N \Rightarrow P_F$ construction

- Whenever  $P_N$ 's stack becomes empty, make  $P_F$  go to a final state without consuming any addition symbol
- To detect empty stack in  $P_N$ :  $P_F$  pushes a new stack symbol  $X_0$  (not in  $\Gamma$  of  $P_N$ ) initially before simulating  $P_N$



$$P_F = (Q_N \cup \{p_0, p_f\}, \Sigma, \Gamma \cup \{X_0\}, \delta_F, p_0, X_0, \{p_f\})$$

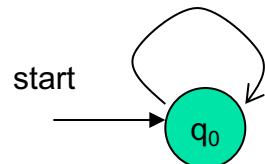
# Example: Matching parenthesis “(” “)”

$P_N:$   $(\{q_0\}, \{(,\)\}, \{Z_0, Z_1\}, \delta_N, q_0, Z_0)$

$\delta_N:$

$$\begin{aligned}\delta_N(q_0, (, Z_0) &= \{ (q_0, Z_1 Z_0) \} \\ \delta_N(q_0, (, Z_1) &= \{ (q_0, Z_1 Z_1) \} \\ \delta_N(q_0, ), Z_1 &= \{ (q_0, \varepsilon) \} \\ \delta_N(q_0, \varepsilon, Z_0) &= \{ (q_0, \varepsilon) \}\end{aligned}$$

$(, Z_0 / Z_1 Z_0$   
 $(, Z_1 / Z_1 Z_1$   
 $), Z_1 / \varepsilon$   
 $\varepsilon, Z_0 / \varepsilon$



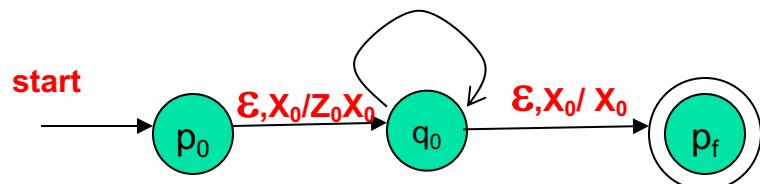
Accept by empty stack

$P_f:$   $(\{p_0, q_0, p_f\}, \{(,\)\}, \{X_0, Z_0, Z_1\}, \delta_f, p_0, X_0, p_f)$

$\delta_f:$

$$\begin{aligned}\delta_f(p_0, \varepsilon, X_0) &= \{ (q_0, Z_0) \} \\ \delta_f(q_0, (, Z_0) &= \{ (q_0, Z_1 Z_0) \} \\ \delta_f(q_0, (, Z_1) &= \{ (q_0, Z_1 Z_1) \} \\ \delta_f(q_0, ), Z_1 &= \{ (q_0, \varepsilon) \} \\ \delta_f(q_0, \varepsilon, Z_0) &= \{ (q_0, \varepsilon) \} \\ \delta_f(p_0, \varepsilon, X_0) &= \{ (p_f, X_0) \}\end{aligned}$$

$(, Z_0 / Z_1 Z_0$   
 $(, Z_1 / Z_1 Z_1$   
 $), Z_1 / \varepsilon$   
 $\varepsilon, Z_0 / \varepsilon$



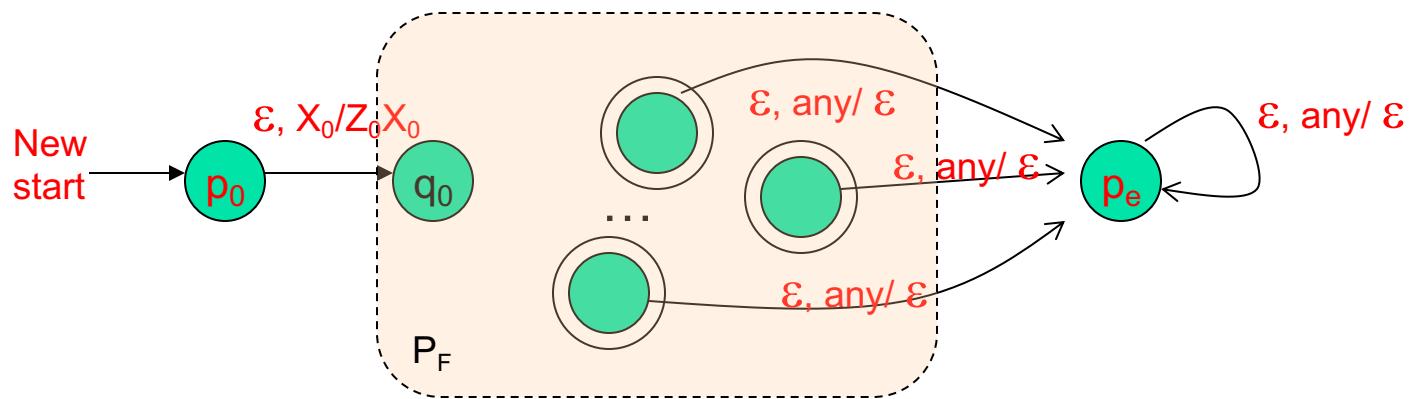
Accept by final state

## $P_F \Rightarrow P_N$ construction

- Main idea:
  - Whenever  $P_F$  reaches a final state, just make an  $\epsilon$ -transition into a new end state, clear out the stack and accept
  - Danger: What if  $P_F$  design is such that it clears the stack midway *without* entering a final state?
    - ➔ to address this, add a new start symbol  $X_0$  (not in  $\Gamma$  of  $P_F$ )

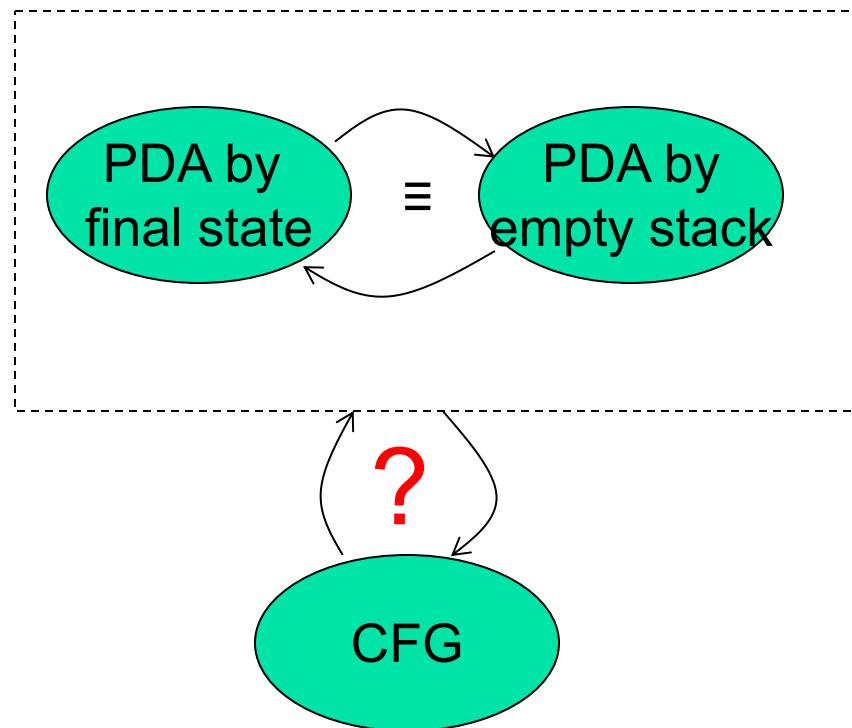
$$P_N = (Q \cup \{p_0, p_e\}, \Sigma, \Gamma \cup \{X_0\}, \delta_N, p_0, X_0)$$

$P_N$ :



# Equivalence of PDAs and CFGs

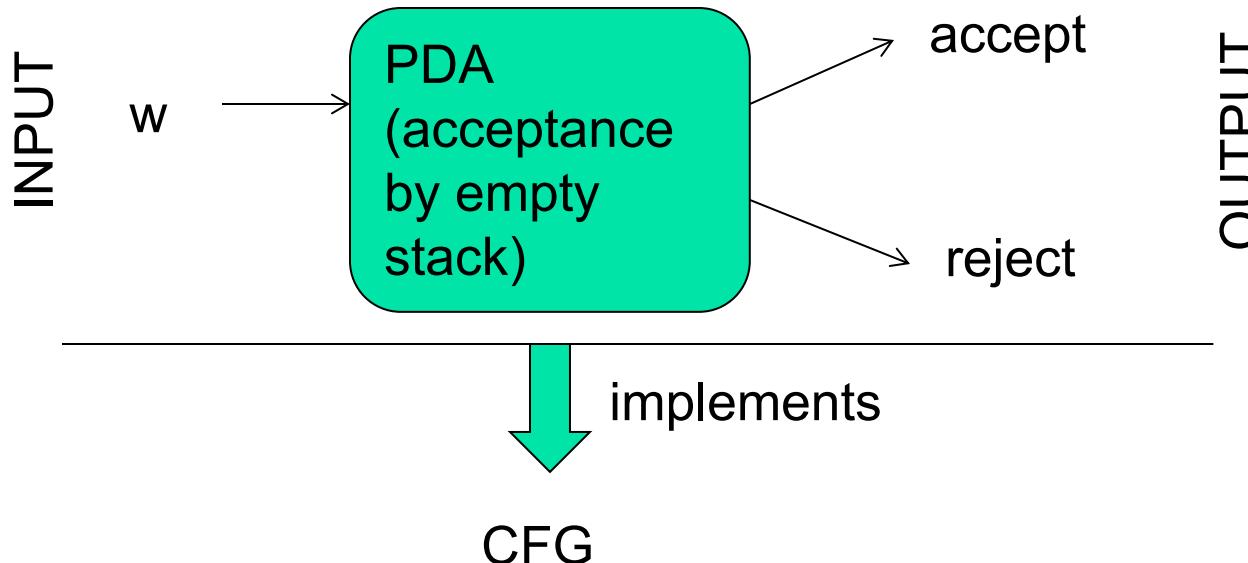
# CFGs == PDAs ==> CFLs



This is same as: “implementing a CFG using a PDA”

# Converting CFG to PDA

Main idea: The PDA simulates the leftmost derivation on a given  $w$ , and upon consuming it fully it either arrives at acceptance (by empty stack) or non-acceptance.



This is same as: “implementing a CFG using a PDA”

# Converting a CFG into a PDA

Main idea: The PDA simulates the leftmost derivation on a given  $w$ , and upon consuming it fully it either arrives at acceptance (by empty stack) or non-acceptance.

## Steps:

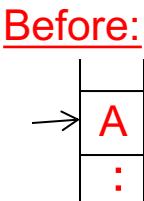
1. Push the right hand side of the production onto the stack, with leftmost symbol at the stack top
2. If stack top is the leftmost variable, then replace it by all its productions (each possible substitution will represent a distinct path taken by the non-deterministic PDA)
3. If stack top has a terminal symbol, and if it matches with the next symbol in the input string, then pop it

State is inconsequential (only one state is needed)

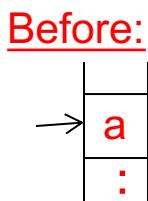
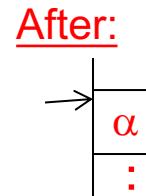
# Formal construction of PDA from CFG

Note: Initial stack symbol (S) same as the start variable in the grammar

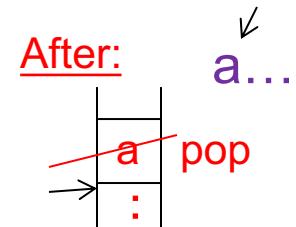
- Given:  $G = (V, T, P, S)$
- Output:  $P_N = (\{q\}, T, V \cup T, \delta, q, S)$
- $\delta$ :



- For all  $A \in V$ , add the following transition(s) in the PDA:
  - $\delta(q, \varepsilon, A) = \{ (q, \alpha) \mid "A \Rightarrow \alpha" \in P \}$

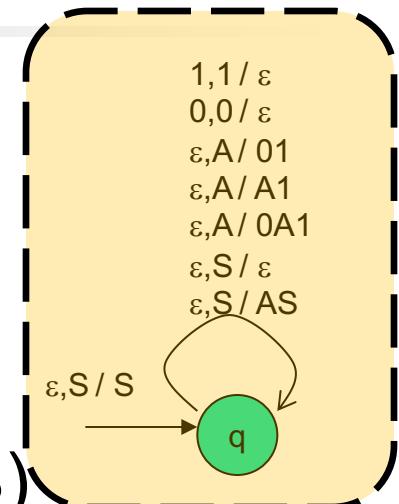


- For all  $a \in T$ , add the following transition(s) in the PDA:
  - $\delta(q, a, a) = \{ (q, \varepsilon) \}$



# Example: CFG to PDA

- $G = (\{S, A\}, \{0, 1\}, P, S)$
- $P:$ 
  - $S \implies AS \mid \epsilon$
  - $A \implies 0A1 \mid A1 \mid 01$
- $PDA = (\{q\}, \{0, 1\}, \{0, 1, A, S\}, \delta, q, S)$
- $\delta:$ 
  - $\delta(q, \epsilon, S) = \{(q, AS), (q, \epsilon)\}$
  - $\delta(q, \epsilon, A) = \{(q, 0A1), (q, A1), (q, 01)\}$
  - $\delta(q, 0, 0) = \{(q, \epsilon)\}$
  - $\delta(q, 1, 1) = \{(q, \epsilon)\}$



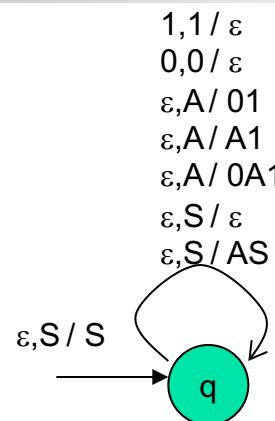
How will this new PDA work?

Lets simulate string 0011

# Simulating string 0011 on the new PDA ...

PDA ( $\delta$ ):

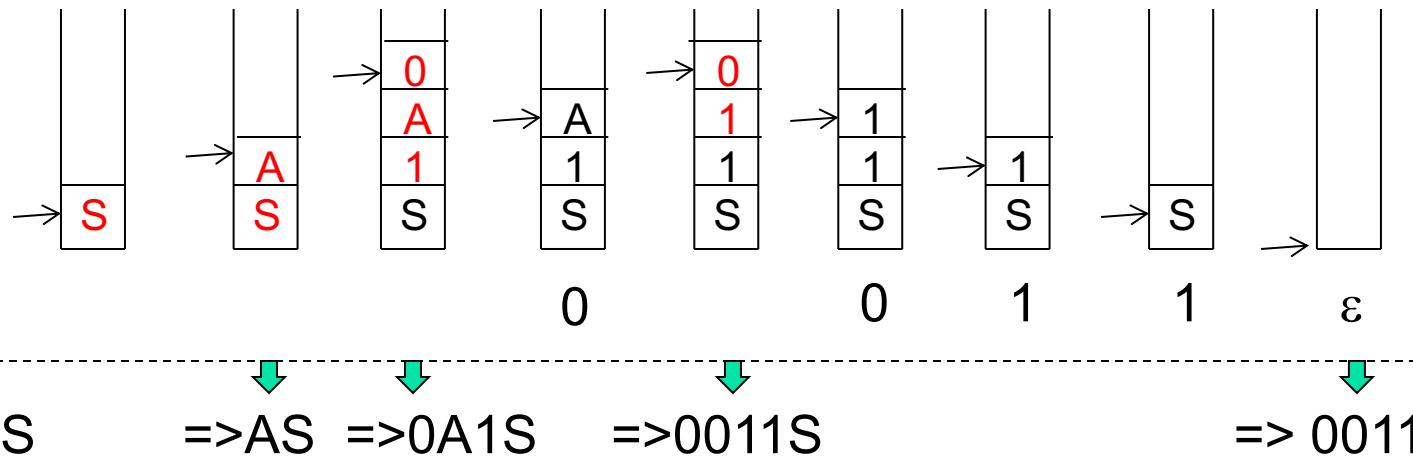
$$\begin{aligned}\delta(q, \varepsilon, S) &= \{ (q, AS), (q, \varepsilon) \} \\ \delta(q, \varepsilon, A) &= \{ (q, 0A1), (q, A1), (q, 01) \} \\ \delta(q, 0, 0) &= \{ (q, \varepsilon) \} \\ \delta(q, 1, 1) &= \{ (q, \varepsilon) \}\end{aligned}$$



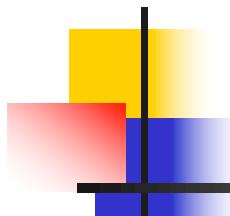
Leftmost deriv.:

$$\begin{aligned}S &\Rightarrow AS \\ &\Rightarrow 0A1S \\ &\Rightarrow 0011S \\ &\Rightarrow 0011\end{aligned}$$

Stack moves (shows only the successful path):



Accept by  
empty stack



# Converting a PDA into a CFG

- Main idea: Reverse engineer the productions from transitions

If  $\delta(q, a, Z) \Rightarrow (p, Y_1 Y_2 Y_3 \dots Y_k)$ :

1. State is changed from q to p;
2. Terminal a is consumed;
3. Stack top symbol Z is popped and replaced with a sequence of k variables.

- Action: Create a grammar variable called “[qZp]” which includes the following production:

- $[qZp] \Rightarrow a[pY_1q_1] [q_1Y_2q_2] [q_2Y_3q_3] \dots [q_{k-1}Y_kq_k]$

- Proof discussion (in the book)

# Example: Bracket matching

- To avoid confusion, we will use  $b=“(“$  and  $e=“)”$

$P_N: (\{q_0\}, \{b,e\}, \{Z_0, Z_1\}, \delta, q_0, Z_0)$

1.  $\delta(q_0, b, Z_0) = \{ (q_0, Z_1 Z_0) \}$
2.  $\delta(q_0, b, Z_1) = \{ (q_0, Z_1 Z_1) \}$
3.  $\delta(q_0, e, Z_1) = \{ (q_0, \epsilon) \}$
4.  $\delta(q_0, \epsilon, Z_0) = \{ (q_0, \epsilon) \}$

0.  $S \Rightarrow [q_0 Z_0 q_0]$
1.  $[q_0 Z_0 q_0] \Rightarrow b [q_0 Z_1 q_0] [q_0 Z_0 q_0]$
2.  $[q_0 Z_1 q_0] \Rightarrow b [q_0 Z_1 q_0] [q_0 Z_1 q_0]$
3.  $[q_0 Z_1 q_0] \Rightarrow e$
4.  $[q_0 Z_0 q_0] \Rightarrow \epsilon$

If you were to directly write a CFG:

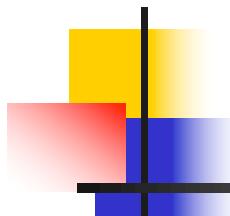
$S \Rightarrow b S e S | \epsilon$

Let  $A=[q_0 Z_0 q_0]$   
Let  $B=[q_0 Z_1 q_0]$

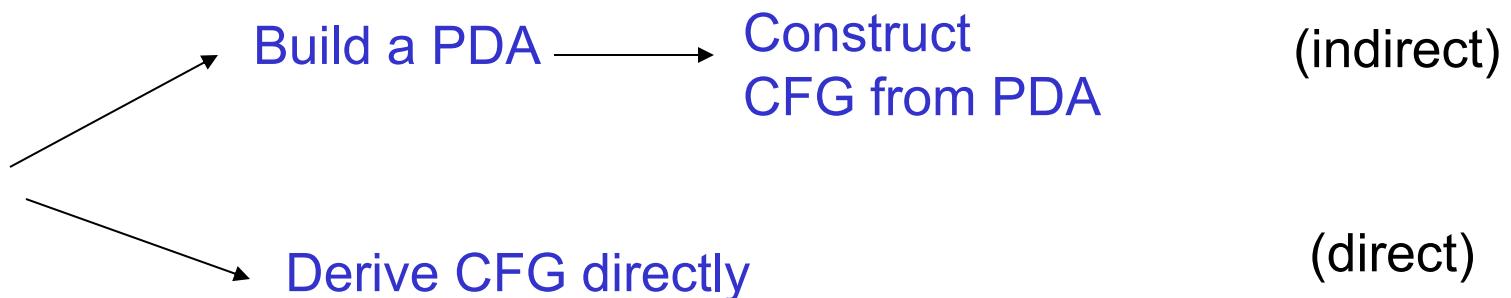
0.  $S \Rightarrow A$
1.  $A \Rightarrow b B A$
2.  $B \Rightarrow b B B$
3.  $B \Rightarrow e$
4.  $A \Rightarrow \epsilon$

Simplifying,

0.  $S \Rightarrow b B S | \epsilon$
1.  $B \Rightarrow b B B | e$

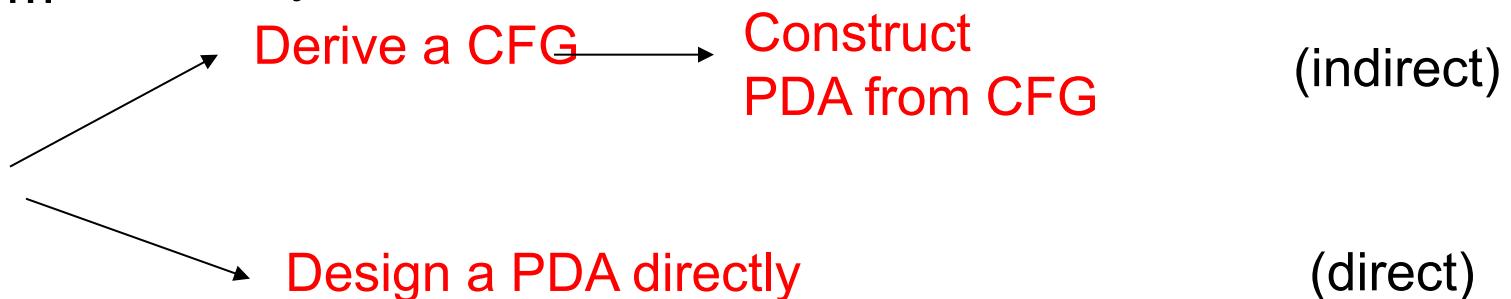


# Two ways to build a CFG



---

Similarly... Two ways to build a PDA

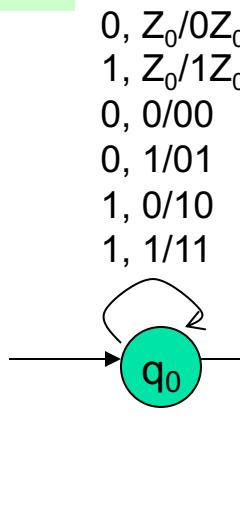




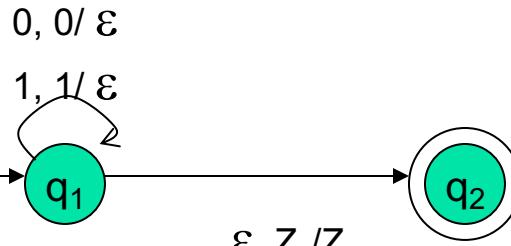
# Deterministic PDAs

# This PDA for $L_{wwr}$ is non-deterministic

Grow stack



Pop stack for matching symbols



Why does it have to be non-deterministic?

Switch to popping mode

Accepts by final state

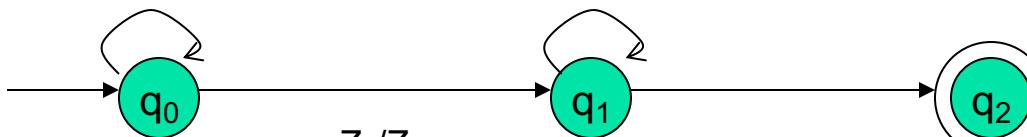
To remove guessing, impose the user to insert c in the middle

## Example shows that: Nondeterministic PDAs $\neq$ D-PDAs

D-PDA for  $L_{wcwr} = \{wcw^R \mid c \text{ is some special symbol not in } w\}$

Grow stack

0,  $Z_0/Z_0$   
1,  $Z_0/Z_0$   
0, 0/0  
0, 1/01  
1, 0/10  
1, 1/11



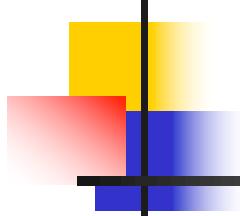
Pop stack for matching symbols

Note:

- all transitions have become deterministic

Switch to popping mode

Accepts by final state

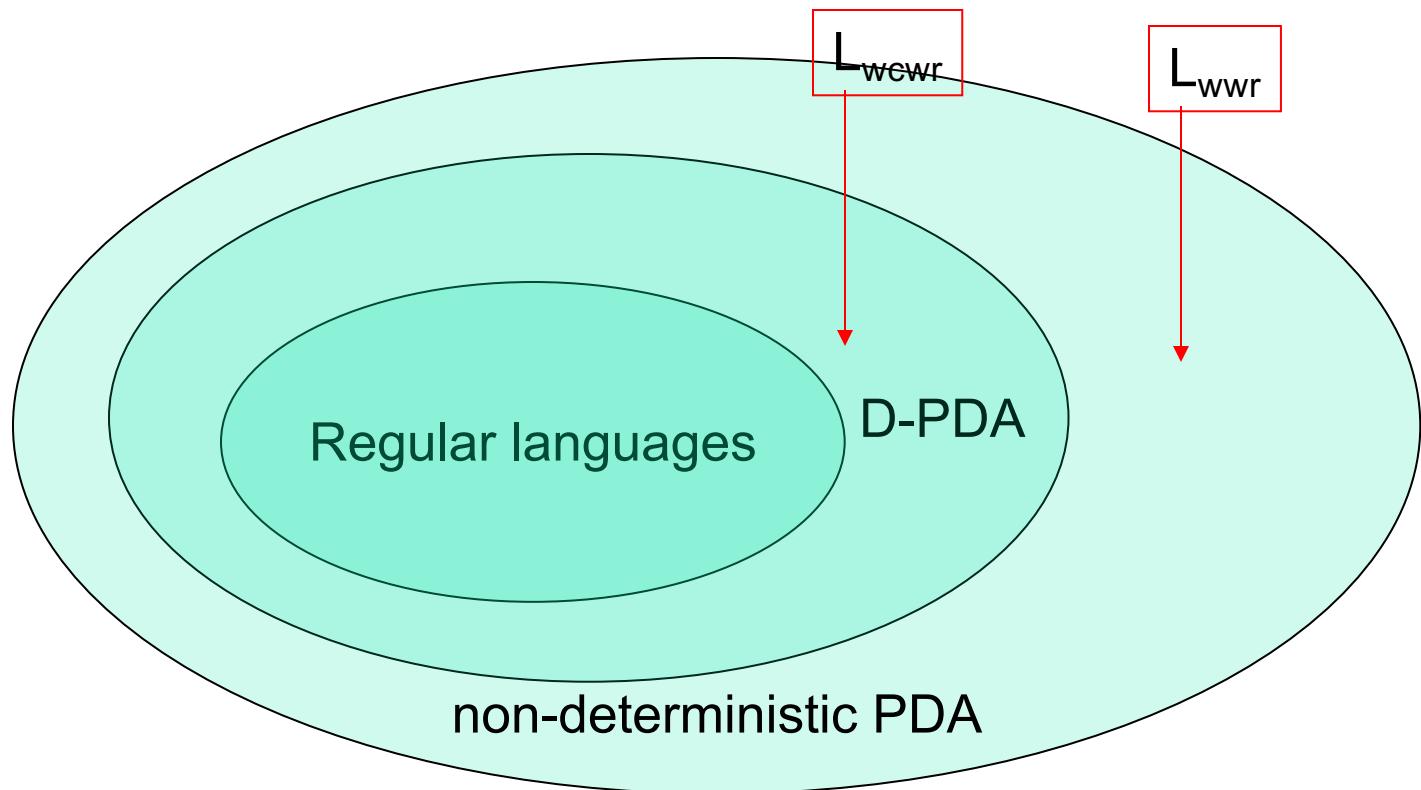


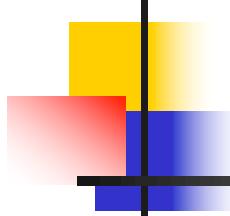
# Deterministic PDA: Definition

---

- A PDA is *deterministic* if and only if:
    1.  $\delta(q,a,X)$  has *at most one* member for any  $a \in \Sigma \cup \{\varepsilon\}$
- If  $\delta(q,a,X)$  is non-empty for some  $a \in \Sigma$ , then  $\delta(q, \varepsilon, X)$  must be empty.

# PDA vs DPDA vs Regular languages





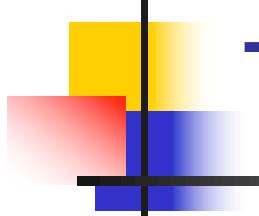
# Summary

---

- PDAs for CFLs and CFGs
  - Non-deterministic
  - Deterministic
- PDA acceptance types
  1. By final state
  2. By empty stack
- PDA
  - IDs, Transition diagram
- Equivalence of CFG and PDA
  - CFG  $\Rightarrow$  PDA construction
  - PDA  $\Rightarrow$  CFG construction

# Properties of Context-free Languages

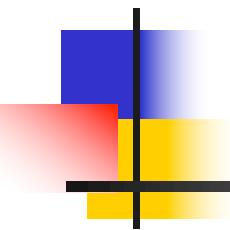
Reading: Chapter 7



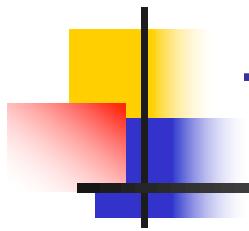
# Topics

---

- 1) Simplifying CFGs, Normal forms
- 2) Pumping lemma for CFLs
- 3) Closure and decision properties of CFLs



# How to “simplify” CFGs?



# Three ways to simplify/clean a CFG

---

*(clean)*

1. Eliminate *useless symbols*

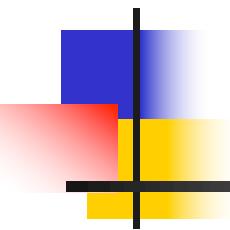
*(simplify)*

2. Eliminate  $\epsilon$ -productions

$A \not\Rightarrow \epsilon$

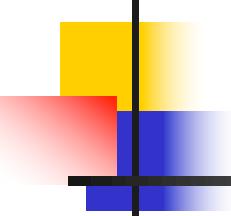
3. Eliminate *unit productions*

$A \not\Rightarrow B$



# Eliminating useless symbols

Grammar cleanup



# Eliminating *useless symbols*

---

A symbol X is reachable if there exists:

- $S \xrightarrow{*} \alpha X \beta$

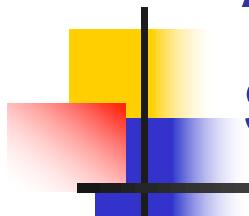
A symbol X is generating if there exists:

- $X \xrightarrow{*} w,$ 
  - for some  $w \in T^*$

For a symbol X to be “useful”, it has to be both  
reachable *and* generating

- $S \xrightarrow{*} \alpha X \beta \xrightarrow{*} w',$  for some  $w' \in T^*$

reachable      generating

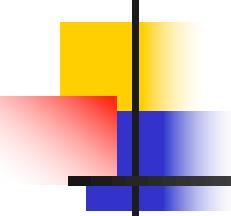


# Algorithm to detect useless symbols

---

1. First, eliminate all symbols that are *not* generating
2. Next, eliminate all symbols that are *not* reachable

Is the order of these steps important,  
or can we switch?



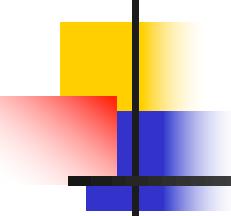
# Example: Useless symbols

- $S \rightarrow AB \mid a$
  - $A \rightarrow b$
1.  $A, S$  are generating
  2.  $B$  is *not generating* (and therefore  $B$  is useless)
  3. ==> Eliminating  $B$ ... (i.e., remove all productions that involve  $B$ )
    1.  $S \rightarrow a$
    2.  $A \rightarrow b$
  4. Now,  $A$  is *not reachable* and therefore is useless

5. Simplified G:
  1.  $S \rightarrow a$

What would happen if you reverse the order:  
i.e., test reachability before generating?

Will fail to remove:  
 $A \rightarrow b$

 $X \rightarrow^* w$ 

## Algorithm to find all generating symbols

---

- Given:  $G=(V,T,P,S)$
- Basis:
  - Every symbol in  $T$  is obviously generating.
- Induction:
  - Suppose for a production  $A \rightarrow \alpha$ , where  $\alpha$  is generating
  - Then,  $A$  is also generating

$$S \xrightarrow{*} \alpha X \beta$$

## Algorithm to find all reachable symbols

- Given:  $G = (V, T, P, S)$
- Basis:
  - $S$  is obviously reachable (from itself)
- Induction:
  - Suppose for a production  $A \xrightarrow{*} \alpha_1 \alpha_2 \dots \alpha_k$ , where  $A$  is reachable
    - Then, all symbols on the right hand side,  $\{\alpha_1, \alpha_2, \dots, \alpha_k\}$  are also reachable.

# Eliminating $\varepsilon$ -productions

$A \Rightarrow \varepsilon$

X

What's the point of removing  $\varepsilon$ -productions?

$$A \xrightarrow{\quad} \varepsilon$$

# Eliminating $\varepsilon$ -productions

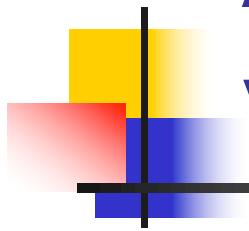
Caveat: It is *not* possible to eliminate  $\varepsilon$ -productions for languages which include  $\varepsilon$  in their word set

So we will target the grammar for the *rest of the language*

Theorem: If  $G=(V,T,P,S)$  is a CFG for a language  $L$ , then  $L \setminus \{\varepsilon\}$  has a CFG without  $\varepsilon$ -productions

Definition: A is “*nullable*” if  $A \xrightarrow{*} \varepsilon$

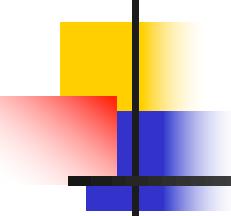
- If  $A$  is nullable, then any production of the form “ $B \xrightarrow{\quad} CAD$ ” can be simulated by:
  - $B \xrightarrow{\quad} CD \mid CAD$ 
    - This can allow us to remove  $\varepsilon$  transitions for  $A$



# Algorithm to detect all nullable variables

---

- Basis:
  - If  $A \rightarrow \varepsilon$  is a production in G, then A is nullable  
(note: A can still have other productions)
- Induction:
  - If there is a production  $B \rightarrow C_1C_2\dots C_k$ , where every  $C_i$  is nullable, then B is also nullable



# Eliminating $\epsilon$ -productions

Given:  $G=(V,T,P,S)$

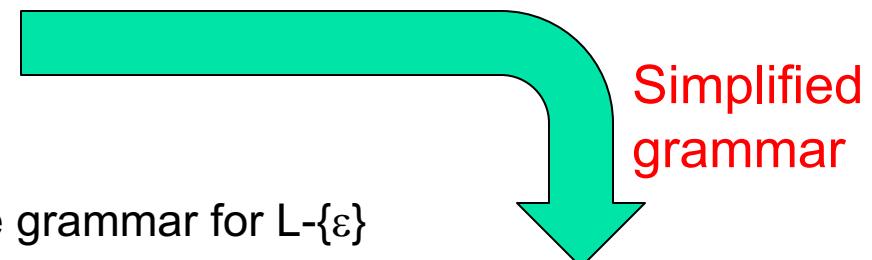
Algorithm:

1. Detect all nullable variables in  $G$
2. Then construct  $G_1=(V,T,P_1,S)$  as follows:
  - i. For each production of the form:  $A \rightarrow X_1X_2\dots X_k$ , where  $k \geq 1$ , suppose  $m$  out of the  $k$   $X_i$ 's are nullable symbols
  - ii. Then  $G_1$  will have  $2^m$  versions for this production
    - iii. i.e, all combinations where each  $X_i$  is either present or absent
  - iii. Alternatively, if a production is of the form:  $A \rightarrow \epsilon$ , then remove it

# Example: Eliminating $\epsilon$ -productions

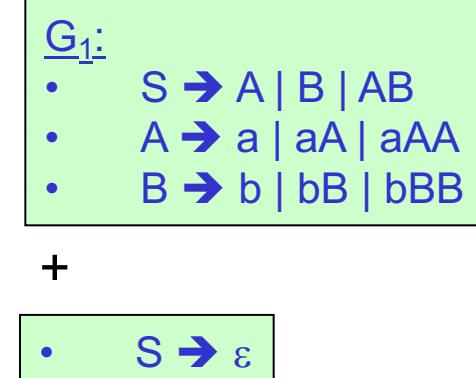
- Let  $L$  be the language represented by the following CFG  $G$ :

- i.  $S \rightarrow AB$
- ii.  $A \rightarrow aAA \mid \epsilon$
- iii.  $B \rightarrow bBB \mid \epsilon$



Goal: To construct  $G_1$ , which is the grammar for  $L - \{\epsilon\}$

- Nullable symbols:  $\{A, B\}$
- $G_1$  can be constructed from  $G$  as follows:
  - $B \rightarrow b \mid bB \mid bB \mid bBB$ 
    - $\Rightarrow B \rightarrow b \mid bB \mid bBB$
  - Similarly,  $A \rightarrow a \mid aA \mid aAA$
  - Similarly,  $S \rightarrow A \mid B \mid AB$
- Note:  $L(G) = L(G_1) \cup \{\epsilon\}$



# Eliminating unit productions

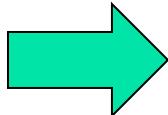
$A \Rightarrow B$  ← B has to be a variable  
X

What's the point of removing unit transitions ?

Will save #substitutions

E.g.,

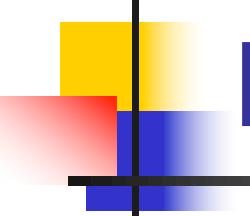
$A \Rightarrow B | \dots$   
 $B \Rightarrow C | \dots$   
 $C \Rightarrow D | \dots$   
 $D \Rightarrow xxx | yyy | zzz$



$A \Rightarrow xxx | yyy | zzz | \dots$   
 $B \Rightarrow xxx | yyy | zzz | \dots$   
 $C \Rightarrow xxx | yyy | zzz | \dots$   
 $D \Rightarrow xxx | yyy | zzz$

*before*

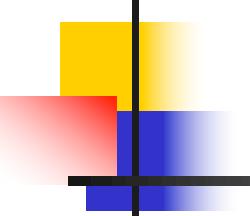
*after*



$A \rightarrow B$

# Eliminating unit productions

- Unit production is one which is of the form  $A \rightarrow B$ , where both A & B are variables
- E.g.,
  - 1.  $E \rightarrow T \mid E+T$
  - 2.  $T \rightarrow F \mid T^*F$
  - 3.  $F \rightarrow I \mid (E)$
  - 4.  $I \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
- How to eliminate unit productions?
  - Replace  $E \rightarrow T$  with  $E \rightarrow F \mid T^*F$
  - Then, upon recursive application wherever there is a unit production:
    - $E \rightarrow F \mid T^*F \mid E+T$  (substituting for T)
    - $E \rightarrow I \mid (E) \mid T^*F \mid E+T$  (substituting for F)
    - $E \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1 \mid (E) \mid T^*F \mid E+T$  (substituting for I)
    - Now, E has no unit productions
  - Similarly, eliminate for the remainder of the unit productions

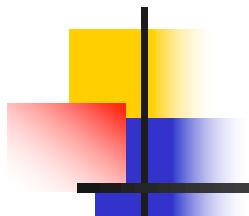


# The Unit Pair Algorithm: to remove unit productions

- Suppose  $A \rightarrow B_1 \rightarrow B_2 \rightarrow \dots \rightarrow B_n \rightarrow \alpha$
- Action: Replace all intermediate productions to produce  $\alpha$  directly
  - i.e.,  $A \rightarrow \alpha; B_1 \rightarrow \alpha; \dots; B_n \rightarrow \alpha;$

Definition:  $(A, B)$  to be a “**unit pair**” if  $A \rightarrow^* B$

- We can find all unit pairs inductively:
  - Basis: Every pair  $(A, A)$  is a unit pair (by definition). Similarly, if  $A \rightarrow B$  is a production, then  $(A, B)$  is a unit pair.
  - Induction: If  $(A, B)$  and  $(B, C)$  are unit pairs, and  $A \rightarrow C$  is also a unit pair.



# The Unit Pair Algorithm: to remove unit productions

---

Input:  $G = (V, T, P, S)$

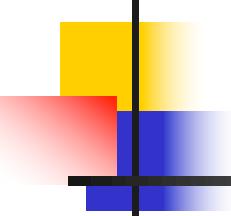
Goal: to build  $G_1 = (V, T, P_1, S)$  devoid of unit productions

Algorithm:

1. Find all unit pairs in  $G$
2. For each unit pair  $(A, B)$  in  $G$ :
  1. Add to  $P_1$  a new production  $A \rightarrow \alpha$ , for every  $B \rightarrow \alpha$  which is a *non-unit* production
  2. If a resulting production is already there in  $P$ , then there is no need to add it.

# Example: eliminating unit productions

Unit pairs		Only non-unit productions to be added to $P_1$
G:	1. $E \rightarrow T   E+T$ 2. $T \rightarrow F   T^*F$ 3. $F \rightarrow I   (E)$ 4. $I \rightarrow a   b   la   lb   l0   l1$	$(E,E)$ $(E,T)$ $(E,F)$ $(E,I)$ $(T,T)$ $(T,F)$ $(T,I)$ $(F,F)$ $(F,I)$ $(I,I)$
$G_1$ :	1. $E \rightarrow E+T   T^*F   (E)   a  b   la   lb   l0   l1$ 2. $T \rightarrow T^*F   (E)   a  b   la   lb   l0   l1$ 3. $F \rightarrow (E)   a  b   la   lb   l0   l1$ 4. $I \rightarrow a   b   la   lb   l0   l1$	$E \rightarrow E+T$ $E \rightarrow T^*F$ $E \rightarrow (E)$ $E \rightarrow a b la lb l0 l1$ $T \rightarrow T^*F$ $T \rightarrow (E)$ $T \rightarrow a b la lb l0 l1$ $F \rightarrow (E)$ $F \rightarrow a b la lb l0 l1$ $I \rightarrow a b la lb l0 l1$



# Putting all this together...

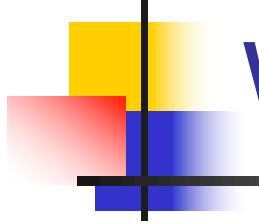
- Theorem: If  $G$  is a CFG for a language that contains at least one string other than  $\varepsilon$ , then there is another CFG  $G_1$ , such that  $L(G_1) = L(G) - \varepsilon$ , and  $G_1$  has:
  - no  $\varepsilon$ -productions
  - no unit productions
  - no useless symbols
- Algorithm:
  - Step 1) eliminate  $\varepsilon$ -productions
  - Step 2) eliminate unit productions
  - Step 3) eliminate useless symbols

Again,  
the order is  
important!

Why?



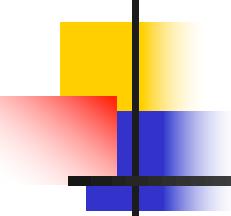
# Normal Forms



# Why normal forms?

---

- If all productions of the grammar could be expressed in the same form(s), then:
  - a. It becomes easy to design algorithms that use the grammar
  - b. It becomes easy to show proofs and properties



# Chomsky Normal Form (CNF)

Let G be a CFG for some L- $\{\varepsilon\}$

## Definition:

G is said to be in **Chomsky Normal Form** if all its productions are in one of the following two forms:

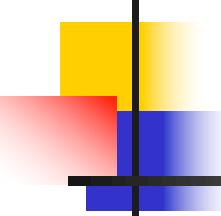
i.  $A \rightarrow BC$

where A, B, C are variables, or

ii.  $A \rightarrow a$

where a is a terminal

- G has no useless symbols
- G has no unit productions
- G has no  $\varepsilon$ -productions



# CNF checklist

Is this grammar in CNF?

G<sub>1</sub>:

1.  $E \rightarrow E+T \mid T^*F \mid (E) \mid Ia \mid Ib \mid I0 \mid I1$
2.  $T \rightarrow T^*F \mid (E) \mid Ia \mid Ib \mid I0 \mid I1$
3.  $F \rightarrow (E) \mid Ia \mid Ib \mid I0 \mid I1$
4.  $I \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$

Checklist:

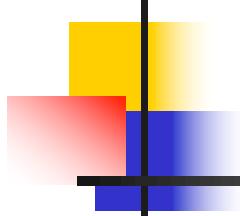
- G has no  $\epsilon$ -productions ✓
- G has no unit productions ✓
- G has no useless symbols ✓
- But...
  - the normal form for productions is violated



So, the grammar is not in CNF

# How to convert a G into CNF?

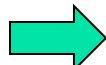
- Assumption: G has no  $\epsilon$ -productions, unit productions or useless symbols
  - 1) For every terminal  $a$  that appears in the body of a production:
    - i. create a unique variable, say  $X_a$ , with a production  $X_a \rightarrow a$ , and
    - ii. replace all other instances of  $a$  in G by  $X_a$
  - 2) Now, all productions will be in one of the following two forms:
    - $A \rightarrow B_1B_2\dots B_k \ (k \geq 3)$       or       $A \rightarrow a$
  - 3) Replace each production of the form  $A \rightarrow B_1B_2B_3\dots B_k$  by:  
$$\begin{array}{c} \xleftarrow{\hspace{1cm}} \\ B_2 \\ \xleftarrow{\hspace{1cm}} \end{array} \quad \begin{array}{c} \xleftarrow{\hspace{1cm}} \\ C_2 \\ \xleftarrow{\hspace{1cm}} \end{array} \quad \text{and so on...}$$
  
$$\begin{array}{c} \xleftarrow{\hspace{1cm}} \\ B_1 \\ \xleftarrow{\hspace{1cm}} \end{array} \quad \begin{array}{c} \xleftarrow{\hspace{1cm}} \\ C_1 \\ \xleftarrow{\hspace{1cm}} \end{array}$$
    - $A \rightarrow B_1C_1 \quad C_1 \rightarrow B_2C_2 \dots \quad C_{k-3} \rightarrow B_{k-2}C_{k-2} \quad C_{k-2} \rightarrow B_{k-1}B_k$



# Example

G:

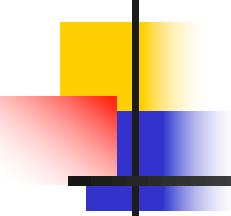
$S \Rightarrow AS \mid BABC$   
 $A \Rightarrow A1 \mid 0A1 \mid 01$   
 $B \Rightarrow 0B \mid 0$   
 $C \Rightarrow 1C \mid 1$



G in CNF:

$X_0 \Rightarrow 0$   
 $X_1 \Rightarrow 1$   
 $S \Rightarrow AS \mid BY_1$   
 $Y_1 \Rightarrow AY_2$   
 $Y_2 \Rightarrow BC$   
 $A \Rightarrow AX_1 \mid X_0Y_3 \mid X_0X_1$   
 $Y_3 \Rightarrow AX_1$   
 $B \Rightarrow X_0B \mid 0$   
 $C \Rightarrow X_1C \mid 1$

All productions are of the form:  $A \Rightarrow BC$  or  $A \Rightarrow a$

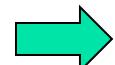


# Languages with $\epsilon$

- For languages that include  $\epsilon$ ,
  - Write down the rest of grammar in CNF
  - Then add production “ $S \Rightarrow \epsilon$ ” at the end

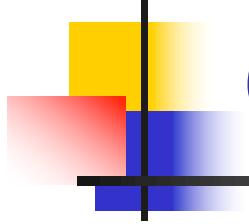
E.g., consider:

G:

$$\begin{aligned} S &\Rightarrow AS \mid BABC \\ A &\Rightarrow A1 \mid 0A1 \mid 01 \mid \epsilon \\ B &\Rightarrow 0B \mid 0 \mid \epsilon \\ C &\Rightarrow 1C \mid 1 \mid \epsilon \end{aligned}$$


G in CNF:

$$\begin{aligned} X_0 &\Rightarrow 0 \\ X_1 &\Rightarrow 1 \\ S &\Rightarrow AS \mid BY_1 \mid \epsilon \\ Y_1 &\Rightarrow AY_2 \\ Y_2 &\Rightarrow BC \\ A &\Rightarrow AX_1 \mid X_0Y_3 \mid X_0X_1 \\ Y_3 &\Rightarrow AX_1 \\ B &\Rightarrow X_0B \mid 0 \\ C &\Rightarrow X_1C \mid 1 \end{aligned}$$

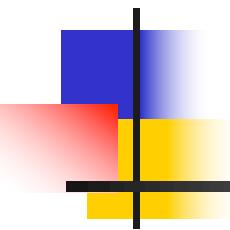


# Other Normal Forms

---

- Griebach Normal Form (GNF)
  - All productions of the form

$A \Rightarrow^* a \alpha$

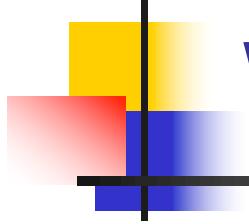


# Return of the Pumping Lemma !!

Think of languages that cannot be CFL

== think of languages for which a stack will not be enough

e.g., the language of strings of the form  $ww$



# Why pumping lemma?

---

- A result that will be useful in proving languages that *are not* CFLs
  - (just like we did for regular languages)
- But before we prove the pumping lemma for CFLs ....
  - Let us first prove an important property about parse trees

Observe that any parse tree generated by a CNF will be a binary tree, where all internal nodes have exactly two children (except those nodes connected to the leaves).

# The “parse tree theorem”

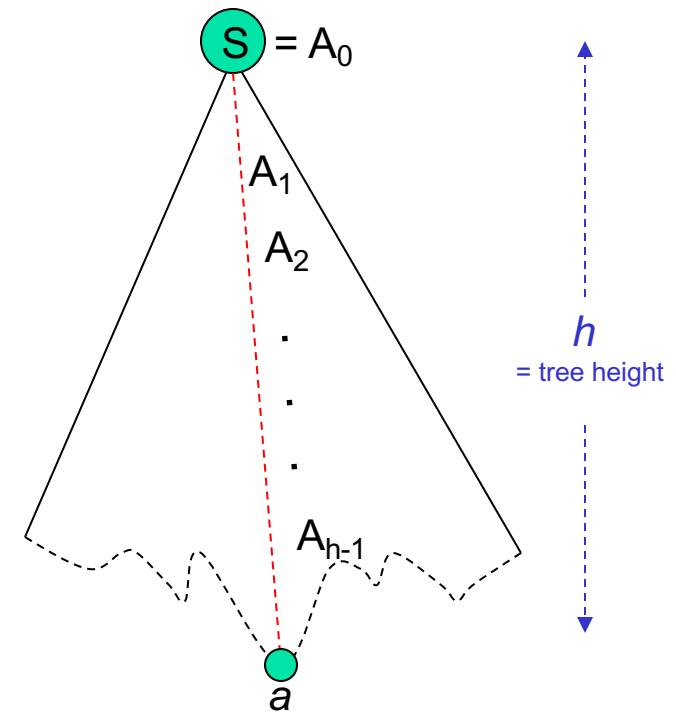
Given:

- Suppose we have a parse tree for a string  $w$ , according to a CNF grammar,  $G=(V,T,P,S)$
- Let  $h$  be the height of the parse tree

Implies:

- $|w| \leq 2^{h-1}$

Parse tree for  $w$



In other words, a CNF parse tree's string yield ( $w$ ) can no longer be  $2^{h-1}$

**To show:**  $|w| \leq 2^{h-1}$

# Proof... The size of parse trees

Proof: (using induction on h)

Basis:  $h = 1$

- Derivation will have to be " $S \rightarrow a$ "
- $|w| = 1 = 2^{1-1}$ .

Ind. Hyp:  $h = k-1$

- $|w| \leq 2^{k-2}$

Ind. Step:  $h = k$

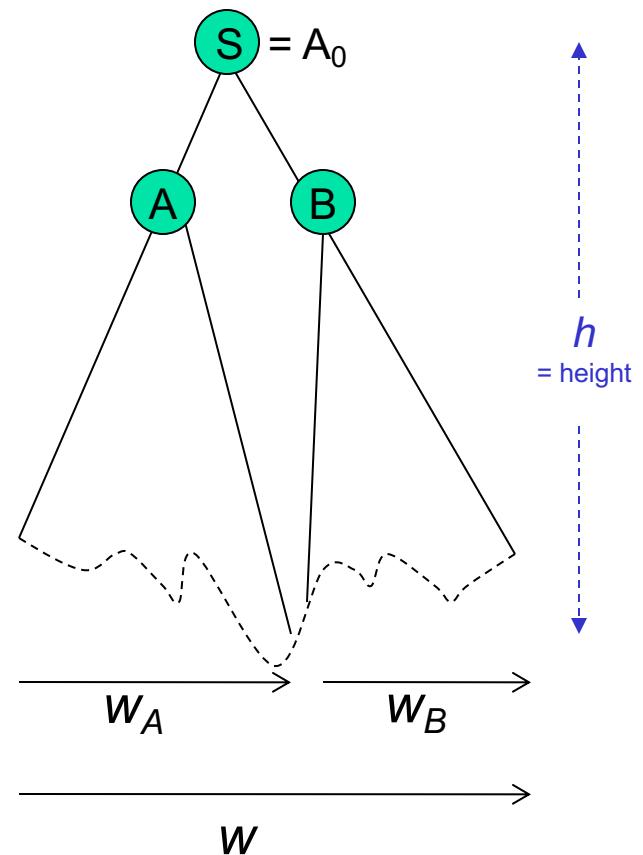
$S$  will have exactly two children:  
 $S \rightarrow AB$

- Heights of  $A$  &  $B$  subtrees are at most  $h-1$

- $w = w_A w_B$ , where  $|w_A| \leq 2^{k-2}$  and  $|w_B| \leq 2^{k-2}$

- $|w| \leq 2^{k-1}$

Parse tree for  $w$



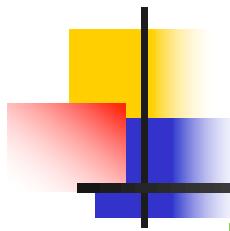
# Implication of the Parse Tree Theorem (assuming CNF)

## Fact:

- If the height of a parse tree is  $h$ , then
  - $\Rightarrow |w| \leq 2^{h-1}$

## Implication:

- If  $|w| \geq 2^m$ , then
  - Its parse tree's height is *at least*  $m+1$



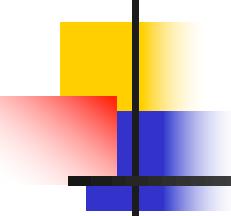
# The Pumping Lemma for CFLs

Let  $L$  be a CFL.

Then there exists a constant  $N$ , s.t.,

- if  $z \in L$  s.t.  $|z| \geq N$ , then we can write  $z = uvwx y$ , such that:
  1.  $|vwx| \leq N$
  2.  $vx \neq \epsilon$
  3. For all  $k \geq 0$ :  $uv^k w x^k y \in L$

Note: we are pumping in two places ( $v$  &  $x$ )



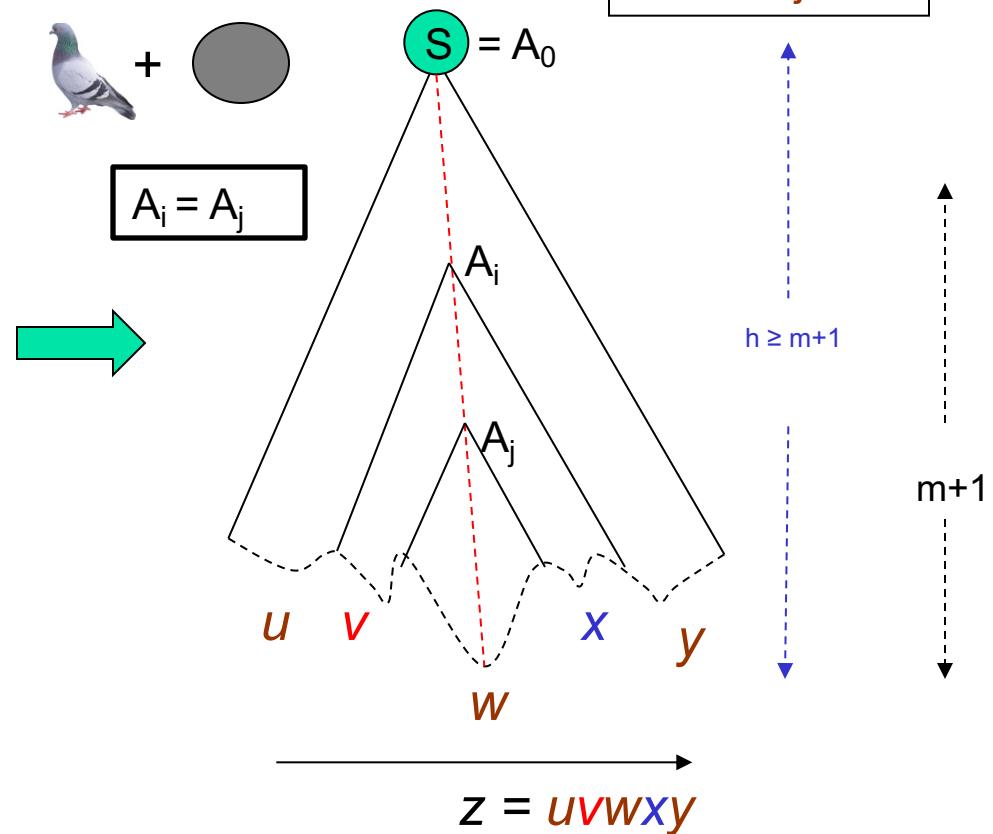
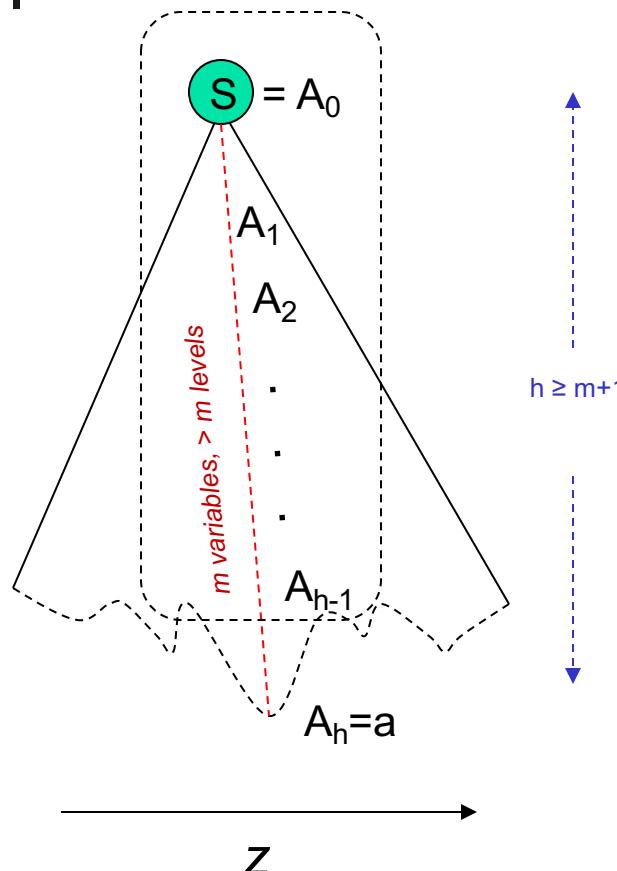
# Proof: Pumping Lemma for CFL

---

- If  $L = \emptyset$  or contains only  $\epsilon$ , then the lemma is trivially satisfied (as it cannot be violated)
- For any other  $L$  which is a CFL:
  - Let  $G$  be a CNF grammar for  $L$
  - Let  $m = \text{number of variables in } G$
  - Choose  $N = 2^m$ .
  - Pick any  $z \in L$  s.t.  $|z| \geq N$ 
    - ➔ the parse tree for  $z$  should have a height  $\geq m+1$   
(by the parse tree theorem)

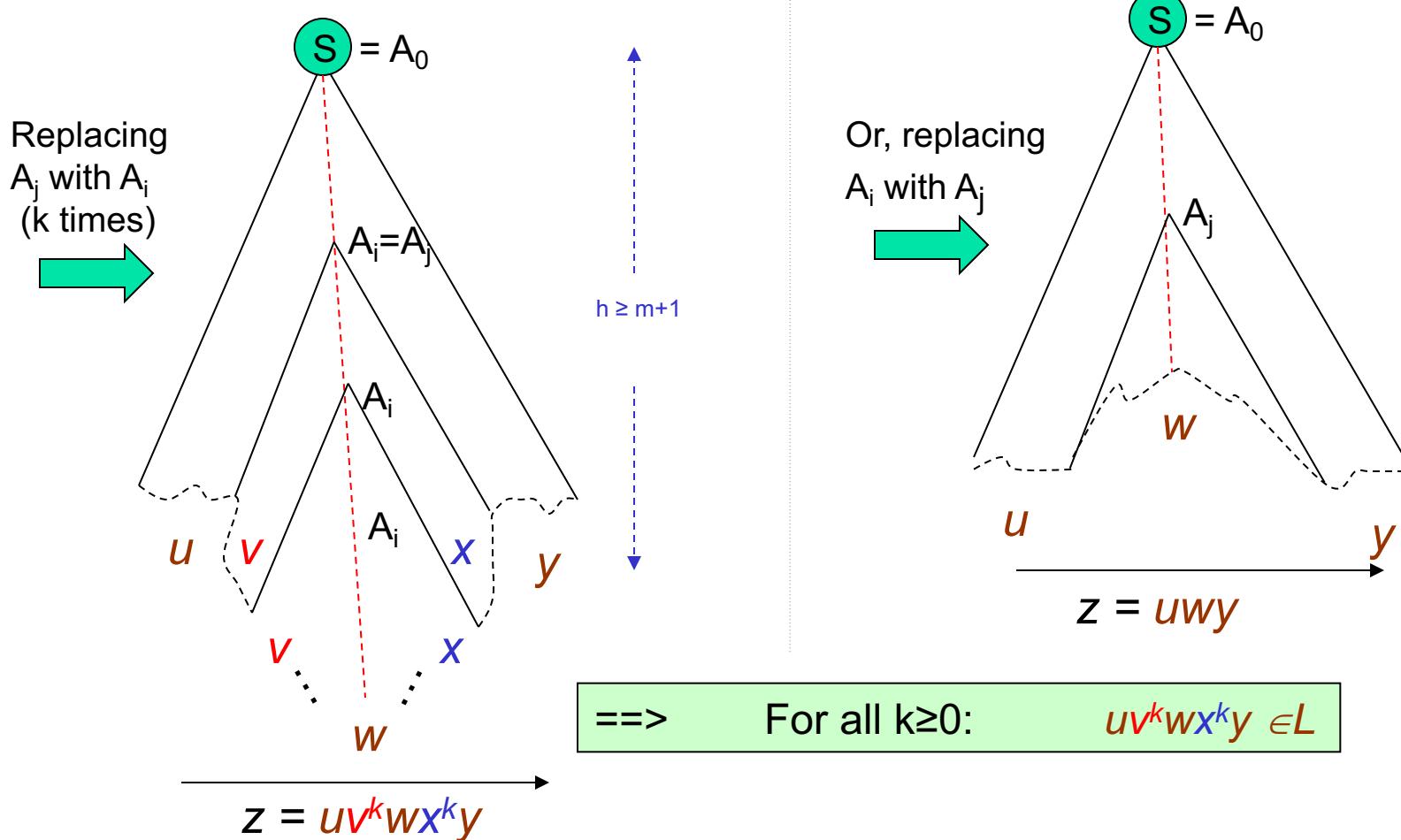
Meaning:  
Repetition in the last  $m+1$  variables

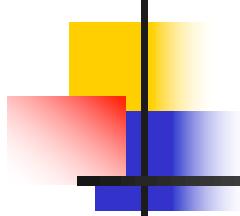
# Parse tree for $z$



- Therefore,  $vx \neq \epsilon$

# Extending the parse tree...





# Proof contd..

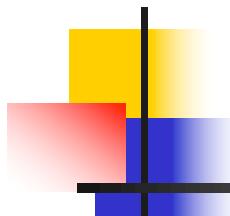
---

- Also, since  $A_i$ 's subtree no taller than  $m+1$   
==> the string generated under  $A_i$ 's subtree, which is  $\text{vwx}$ , cannot be longer than  $2^m (=N)$

But,  $2^m = N$

==>  $|\text{vwx}| \leq N$

This completes the proof for the pumping lemma.



# Application of Pumping Lemma for CFLs

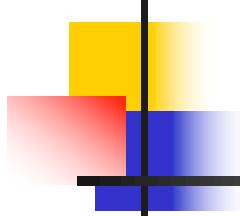
---

Example 1:  $L = \{a^m b^m c^m \mid m > 0\}$

Claim:  $L$  is not a CFL

Proof:

- Let  $N \leq P/L$  constant
- Pick  $z = a^N b^N c^N$
- Apply pumping lemma to  $z$  and show that there exists at least one other string constructed from  $z$  (obtained by pumping up or down) that is  $\notin L$



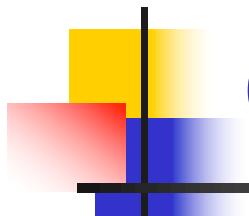
# Proof contd...

---

- $z = uvwxy$
- As  $z = a^N b^N c^N$  and  $|vwx| \leq N$  and  $vx \neq \epsilon$ 
  - $\Rightarrow v, x$  cannot contain all three symbols (a,b,c)
  - $\Rightarrow$  we can pump up or pump down to build another string which is  $\notin L$



# CFL Closure Properties



# Closure Property Results

- CFLs are closed under:
  - Union
  - Concatenation
  - Kleene closure operator
  - Substitution
  - Homomorphism, inverse homomorphism
  - reversal
- CFLs are *not* closed under:
  - Intersection
  - Difference
  - Complementation

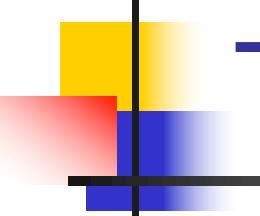
Note: Reg languages  
are closed  
under  
these  
operators

# Strategy for Closure Property Proofs

- First prove “closure under **substitution**”
- Using the above result, prove other closure properties
- CFLs are closed under:

- Union ←
- Concatenation ←
- Kleene closure operator ←
- Substitution →
- Homomorphism, inverse homomorphism ←
- Reversal

Prove  
this first



Note:  $s(L)$  can use  
a different alphabet

# The *Substitution* operation

For each  $a \in \Sigma$ , then let  $s(a)$  be a language

If  $w=a_1a_2\dots a_n \in L$ , then:

- $s(w) = \{ x_1x_2 \dots \} \in s(L)$ , s.t.,  $x_i \in s(a_i)$

## Example:

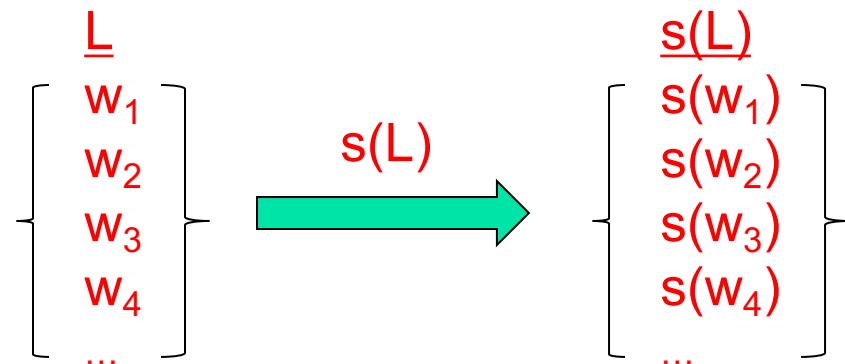
- Let  $\Sigma=\{0,1\}$
- Let:  $s(0) = \{a^n b^n \mid n \geq 1\}$ ,  $s(1) = \{aa, bb\}$
- If  $w=01$ ,  $s(w)=s(0).s(1)$ 
  - E.g.,  $s(w)$  contains  $a^1 b^1 aa$ ,  $a^1 b^1 bb$ ,  
 $a^2 b^2 aa$ ,  $a^2 b^2 bb$ ,  
... and so on.

# CFLs are closed under Substitution

IF  $L$  is a CFL and a substitution defined on  $L$ ,  $s(L)$ , is s.t.,  $s(a)$  is a CFL for every symbol  $a$ , THEN:

- $s(L)$  is also a CFL

What is  $s(L)$ ?



Note: each  $s(w)$  is itself a set of strings

# Substitution of a CFL: example

- Let  $L$  = language of binary palindromes s.t., substitutions for 0 and 1 are defined as follows:
  - $s(0) = \{a^n b^n \mid n \geq 1\}$ ,  $s(1) = \{xx, yy\}$
- Prove that  $s(L)$  is also a CFL.

CFG for  $L$ :

$S \Rightarrow 0S0 \mid 1S1 \mid \epsilon$

CFG for  $s(0)$ :

$S_0 \Rightarrow aS_0b \mid ab$

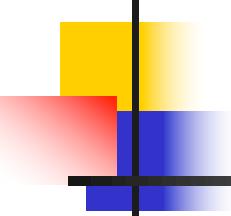
CFG for  $s(1)$ :

$S_1 \Rightarrow xx \mid yy$



**Therefore, CFG for  $s(L)$ :**

$S \Rightarrow S_0SS_0 \mid S_1S S_1 \mid \epsilon$   
 $S_0 \Rightarrow aS_0b \mid ab$   
 $S_1 \Rightarrow xx \mid yy$



# CFLs are closed under *union*

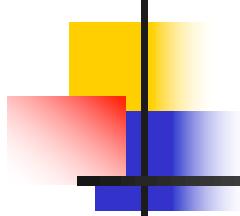
---

Let  $L_1$  and  $L_2$  be CFLs

To show:  $L_2 \cup L_2$  is also a CFL

Let us show by using the result of Substitution

- Make a new language:
  - $L_{\text{new}} = \{a,b\}$  s.t.,  $s(a) = L_1$  and  $s(b) = L_2$   
 $\implies s(L_{\text{new}}) == \text{same as} == L_1 \cup L_2$
- A more direct, alternative proof
  - Let  $S_1$  and  $S_2$  be the starting variables of the grammars for  $L_1$  and  $L_2$ 
    - Then,  $S_{\text{new}} \Rightarrow S_1 | S_2$



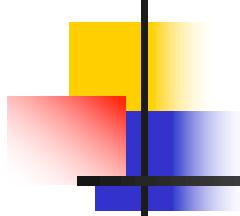
# CFLs are closed under *concatenation*

---

- Let  $L_1$  and  $L_2$  be CFLs

Let us show by using the result of Substitution

- Make  $L_{\text{new}} = \{ab\}$  s.t.,  
 $s(a) = L_1$  and  $s(b) = L_2$   
 $\implies L_1 L_2 = s(L_{\text{new}})$



# CFLs are closed under *Kleene Closure*

---

- Let  $L$  be a CFL
  
- Let  $L_{\text{new}} = \{a\}^*$  and  $s(a) = L_1$ 
  - Then,  $L^* = s(L_{\text{new}})$

# CFLs are closed under *Reversal*

- Let  $L$  be a CFL, with grammar  $G=(V,T,P,S)$
- For  $L^R$ , construct  $G^R=(V,T,P^R,S)$  s.t.,
  - If  $A \Rightarrow \alpha$  is in  $P$ , then:
    - $A \Rightarrow \alpha^R$  is in  $P^R$
    - (that is, reverse every production)

# CFLs are *not* closed under Intersection

- Existential proof:
  - $L_1 = \{0^n 1^n 2^i \mid n \geq 1, i \geq 1\}$
  - $L_2 = \{0^i 1^n 2^n \mid n \geq 1, i \geq 1\}$
- Both  $L_1$  and  $L_2$  are CFLs
- But  $L_1 \cap L_2$  *cannot* be a CFL
- We have an example, where intersection is not closed.
- Therefore, CFLs are not closed under intersection

# CFLs are not closed under complementation

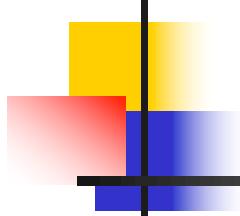
- Follows from the fact that CFLs are not closed under intersection
- $L_1 \cap L_2 = \overline{\overline{L_1} \cup \overline{L_2}}$

Logic: if CFLs were to be closed under complementation

- ➔ the whole right hand side becomes a CFL (because CFL is closed for union)
- ➔ the left hand side (intersection) is also a CFL
- ➔ but we just showed CFLs are NOT closed under intersection!
- ➔ CFLs cannot be closed under complementation.

# CFLs are not closed under difference

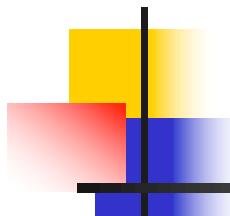
- Follows from the fact that CFLs are not closed under complementation
- Because, if CFLs are closed under difference, then:
  - $\overline{L} = \Sigma^* - L$
  - So  $\overline{L}$  has to be a CFL too
  - Contradiction



# Decision Properties

---

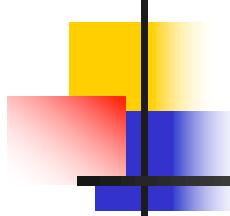
- Emptiness test
  - Generating test
  - Reachability test
- Membership test
  - PDA acceptance



# “Undecidable” problems for CFL

---

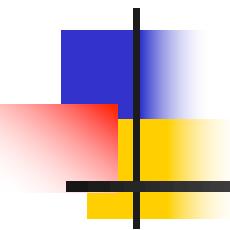
- Is a given CFG  $G$  ambiguous?
- Is a given CFL inherently ambiguous?
- Is the intersection of two CFLs empty?
- Are two CFLs the same?
- Is a given  $L(G)$  equal to  $\Sigma^*$ ?



# Summary

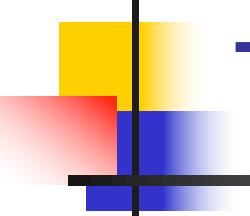
---

- Normal Forms
  - Chomsky Normal Form
  - Griebach Normal Form
  - Useful in proving P/L
- Pumping Lemma for CFLs
  - Main difference:  $z=uv^iwx^iy$
- Closure properties
  - Closed under: union, concatenation, reversal, Kleen closure, homomorphism, substitution
  - Not closed under: intersection, complementation, difference



# Turing Machines

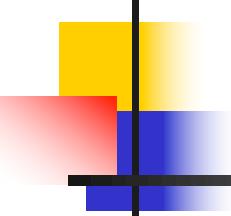
Reading: Chapter 8



# Turing Machines

---

- Invented by Alan Turing in 1936
- Like a CPU to control the data manipulation.
- Mathematical model of computation
  - Abstract Machine which manipulates the symbols on a strip of tape according to rules.
- Infinite memory tape
  - Divided into discrete cells
  - Head reads symbols on a tape
    - As per the instruction writes a symbol
    - decides to move left or right
    - Either moves to next symbol or halts



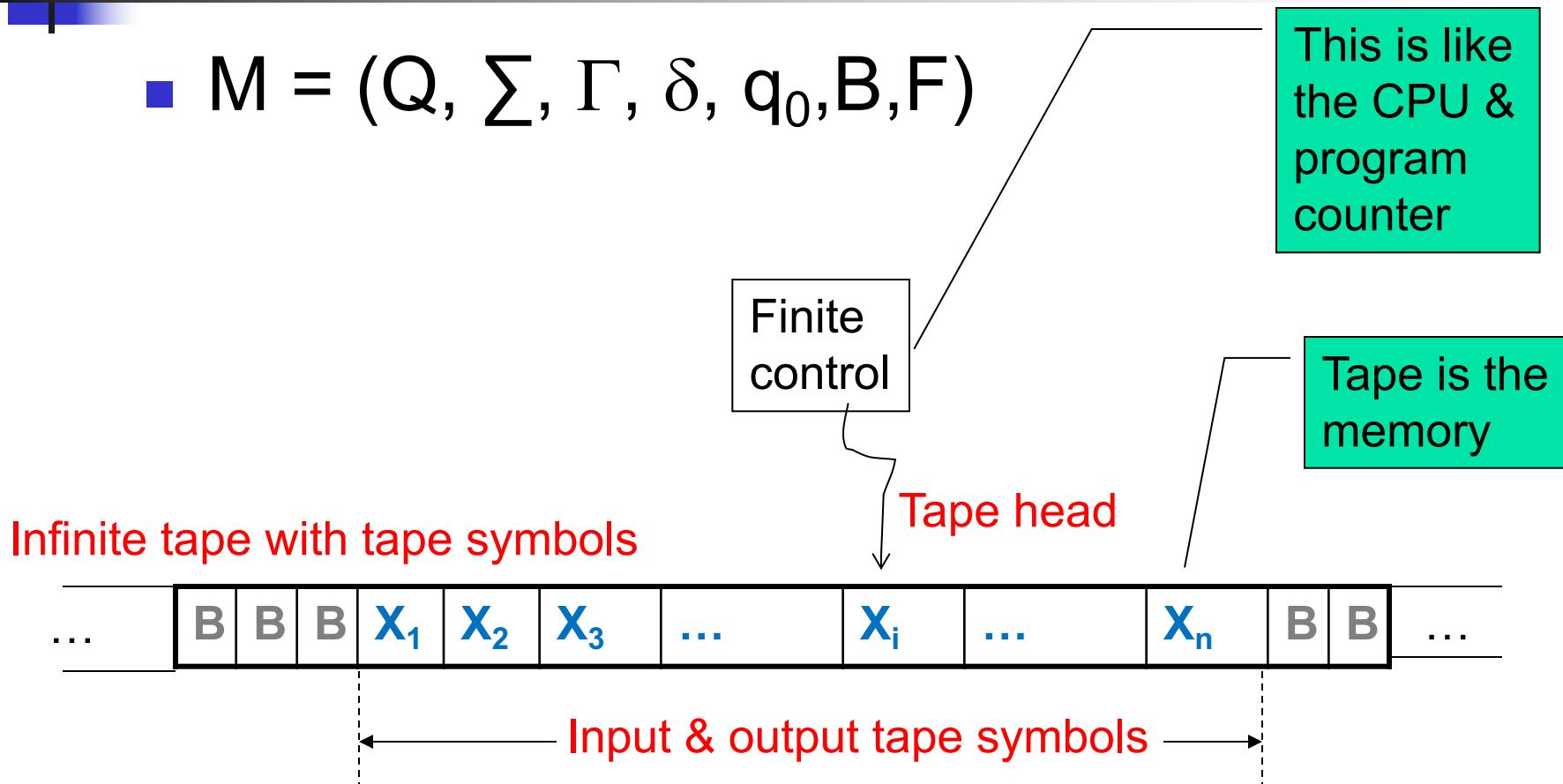
# Turing Machines are...

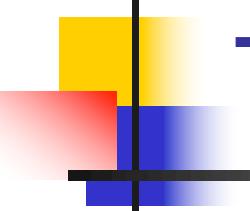
- Very powerful (abstract) machines that could simulate any modern day computer (although very, very slowly!)
- Why design such a machine?
  - If a problem cannot be “solved” even using a TM, then it implies that the problem is ***undecidable***
- Computability vs. Decidability

For every input,  
answer YES or NO

# A Turing Machine (TM)

- $M = (Q, \Sigma, \Gamma, \delta, q_0, B, F)$





You can also use:

→ for R

← for L

# Transition function

- One move (denoted by |---) in a TM does the following:

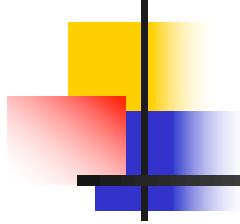
- $\delta(q, X) = (p, Y, D)$

- q is the current state
  - X is the current tape symbol pointed by tape head
  - State changes from q to p
  - After the move:

- X is replaced with symbol Y
    - If D="L", the tape head moves "left" by one position.

Alternatively, if D="R" the tape head moves "right" by one position.

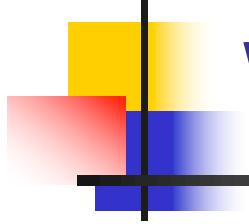




# ID of a TM

---

- Instantaneous Description or ID :
  - $X_1 X_2 \dots X_{i-1} q X_i X_{i+1} \dots X_n$   
means:
    - $q$  is the current state
    - Tape head is pointing to  $X_i$
    - $X_1 X_2 \dots X_{i-1} X_i X_{i+1} \dots X_n$  are the current tape symbols
- $\delta(q, X_i) = (p, Y, R)$  is same as:  
 $X_1 \dots X_{i-1} q X_i \dots X_n \xrightarrow{\text{---}} X_1 \dots X_{i-1} Y p X_{i+1} \dots X_n$
- $\delta(q, X_i) = (p, Y, L)$  is same as:  
 $X_1 \dots X_{i-1} q X_i \dots X_n \xrightarrow{\text{---}} X_1 \dots p X_{i-1} Y X_{i+1} \dots X_n$



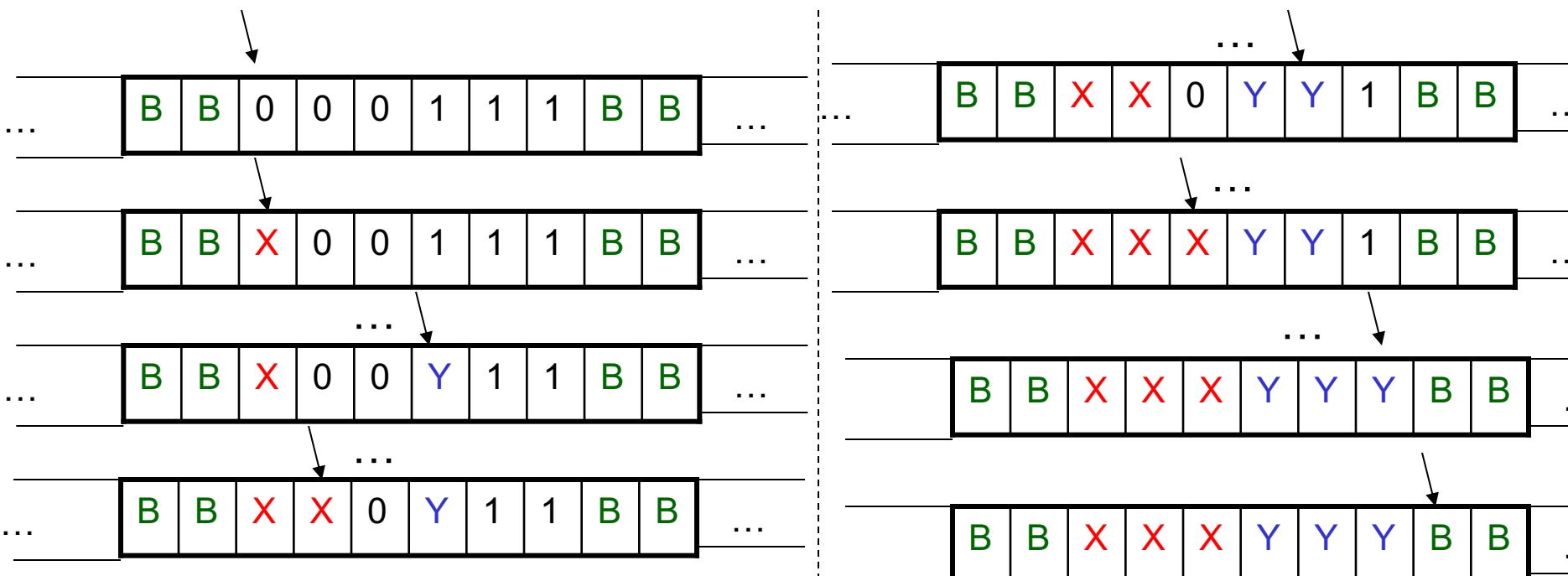
# Way to check for Membership

---

- Is a string  $w$  accepted by a TM?
  
- Initial condition:
  - The (whole) input string  $w$  is present in TM,  
preceded and followed by infinite blank symbols
- Final acceptance:
  - Accept  $w$  if TM enters final state and halts
  - If TM halts and not final state, then reject

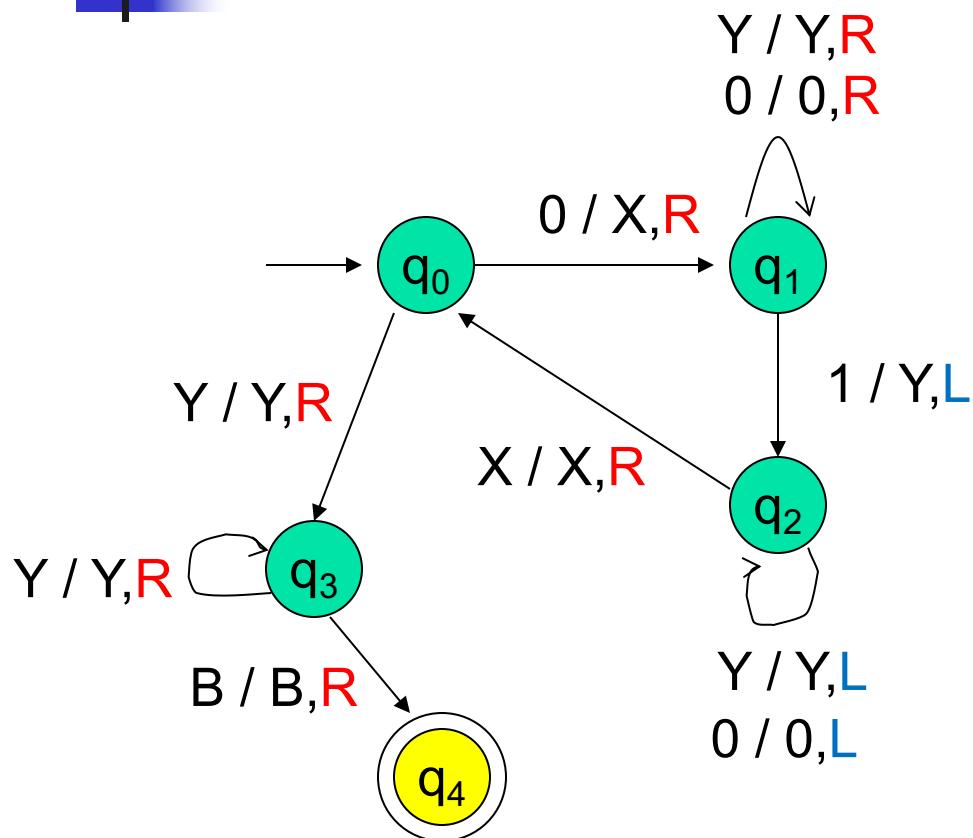
# Example: $L = \{0^n 1^n \mid n \geq 1\}$

- Strategy:  $w = 000111$

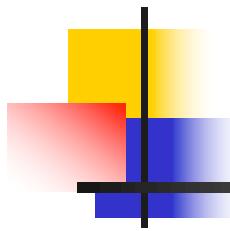


Accept

# TM for $\{0^n 1^n \mid n \geq 1\}$



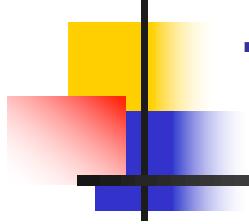
1. Mark next unread 0 with X and move right
2. Move to the right all the way to the first unread 1, and mark it with Y
3. Move back (to the left) all the way to the last marked X, and then move one position to the right
4. If the next position is 0, then goto step 1.  
Else move all the way to the right to ensure there are no excess 1s. If not move right to the next blank symbol and stop & accept.



# TM for $\{0^n 1^n \mid n \geq 1\}$

	Next Tape Symbol				
Curr. State	0	1	X	Y	B
→ $q_0$	$(q_1, X, R)$	-	-	$(q_3, Y, R)$	-
$q_1$	$(q_1, 0, R)$	$(q_2, Y, L)$	-	$(q_1, Y, R)$	-
$q_2$	$(q_2, 0, L)$	-	$(q_0, X, R)$	$(q_2, Y, L)$	-
$q_3$	-	-	-	$(q_3, Y, R)$	$(q_4, B, R)$
* $q_4$	-	--	-	-	-

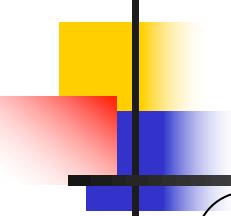
Table representation of the state diagram



# TMs for calculations

---

- TMs can also be used for calculating values
  - Like arithmetic computations
  - Eg., addition, subtraction, multiplication, etc.



## Example 2: monus subtraction

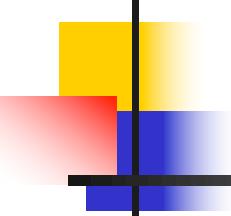
$$“m -\! n” = \max\{m-n, 0\}$$

$0^m 1 0^n \rightarrow \dots B 0^{m-n} B.. \text{ (if } m > n\text{)}$

$\dots BB\dots B.. \text{ (otherwise)}$

1. For every 0 on the left (mark X), mark off a 0 on the right (mark Y)
2. Repeat process, until one of the following happens:
  1. // No more 0s remaining on the left of 1  
Answer is 0, so flip all excess 0s on the right of 1 to Bs (and the 1 itself) and halt
  2. //No more 0s remaining on the right of 1  
Answer is  $m-n$ , so simply halt after making 1 to B

Give state diagram



# Example 3: Multiplication

- $0^m 1 0^n 1$  (input),  $0^{mn} 1$  (output)
- Pseudocode:
  1. Move tape head back & forth such that for every 0 seen in  $0^m$ , write n 0s to the right of the last delimiting 1
  2. Once written, that zero is changed to B to get marked as finished
  3. After completing on all m 0s, make the remaining n 0s and 1s also as Bs

Give state diagram

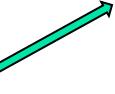
# Calculations vs. Languages

A “calculation” is one that takes an input and outputs a value (or values)



The “language” for a certain calculation is the set of strings of the form “<input, output>”, where the output corresponds to a valid calculated value for the input

A “language” is a set of strings that meet certain criteria



E.g., The language  $L_{add}$  for the addition operation

“<0#0,0>”

“<0#1,1>”

...

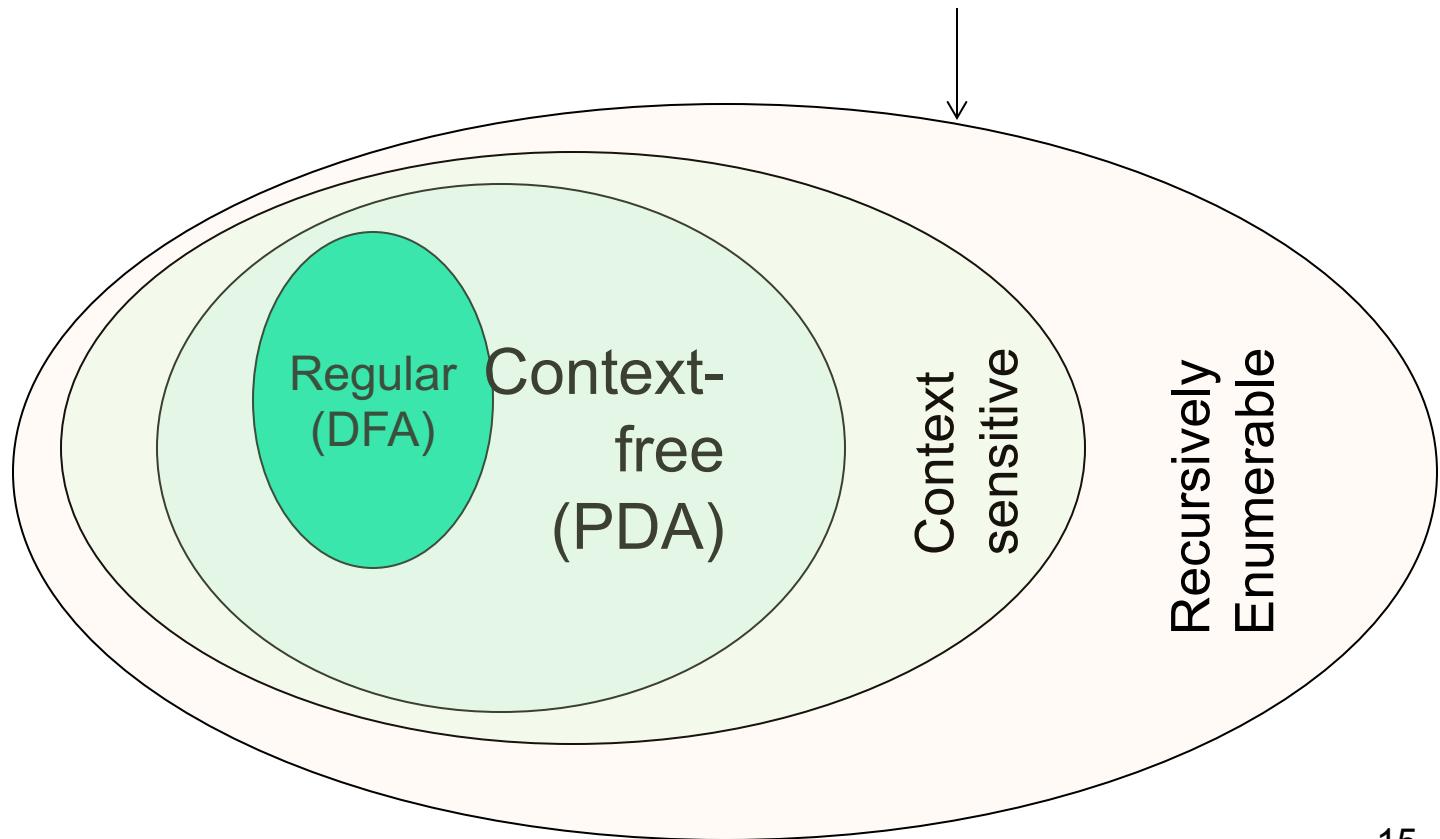
“<2#4,6>”

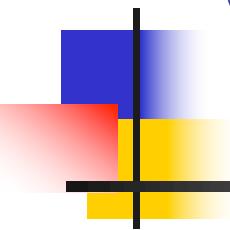
...

Membership question == verifying a solution  
e.g., is “<15#12,27>” a member of  $L_{add}$  ?

# Language of the Turing Machines

- *Recursive Enumerable (RE) language*



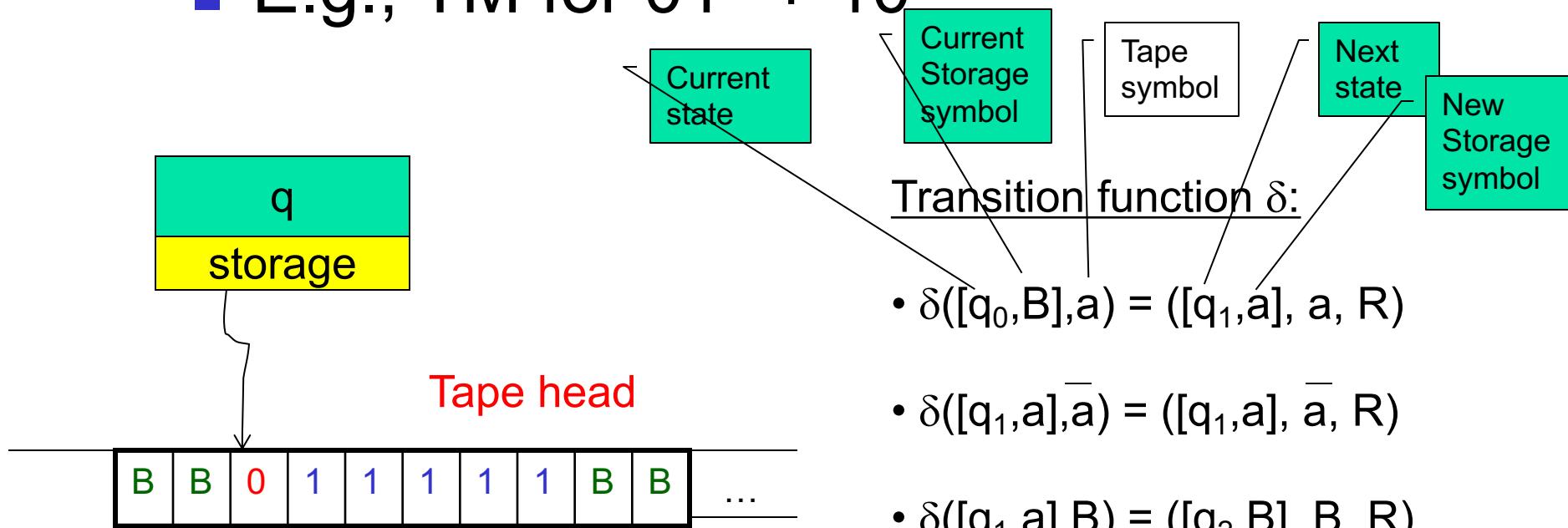


# Variations of Turing Machines

Generic description  
Will work for both  $a=0$  and  $a=1$

# TMs with storage

- E.g., TM for  $01^* + 10^*$



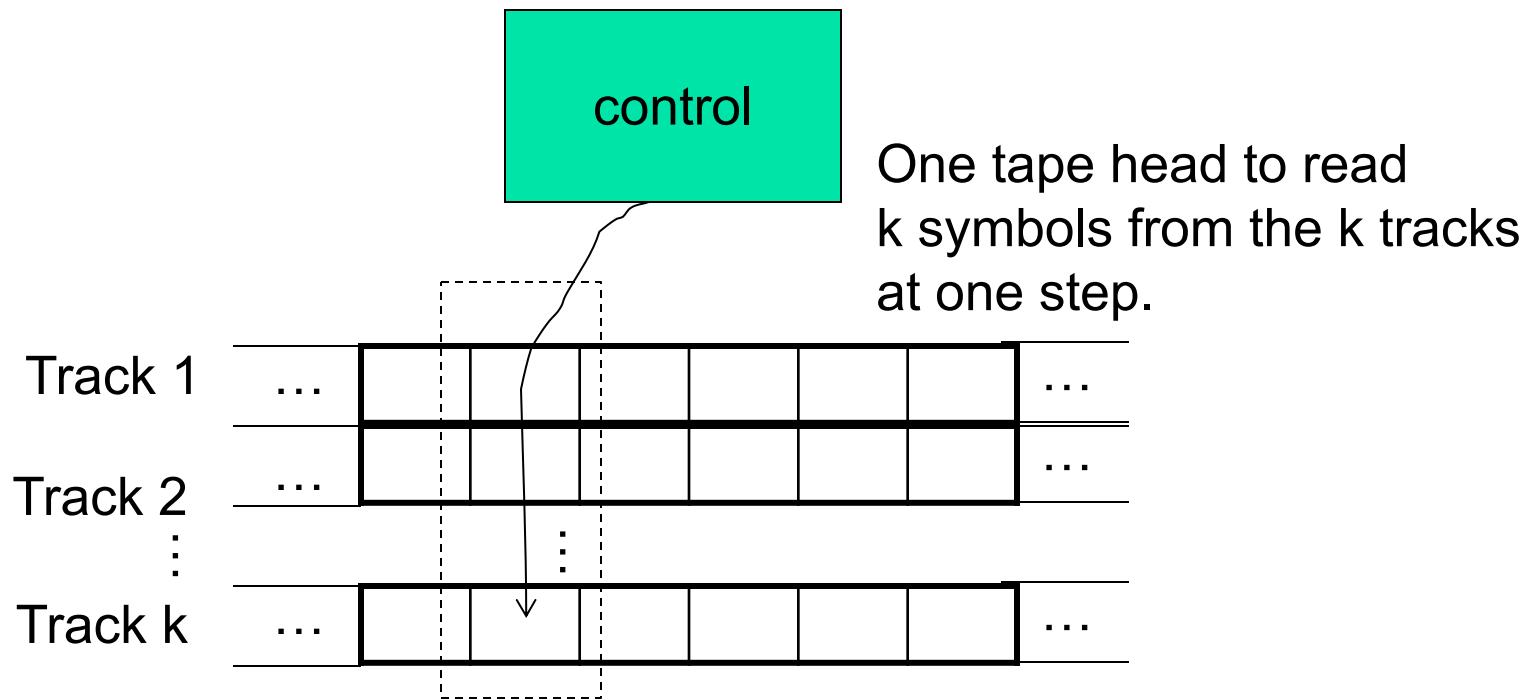
$[q, a]$ : where  $q$  is current state,  
 $a$  is the symbol in storage

Are the standard TMs  
equivalent to TMs with storage?

Yes

# **Multi-track** Turing Machines

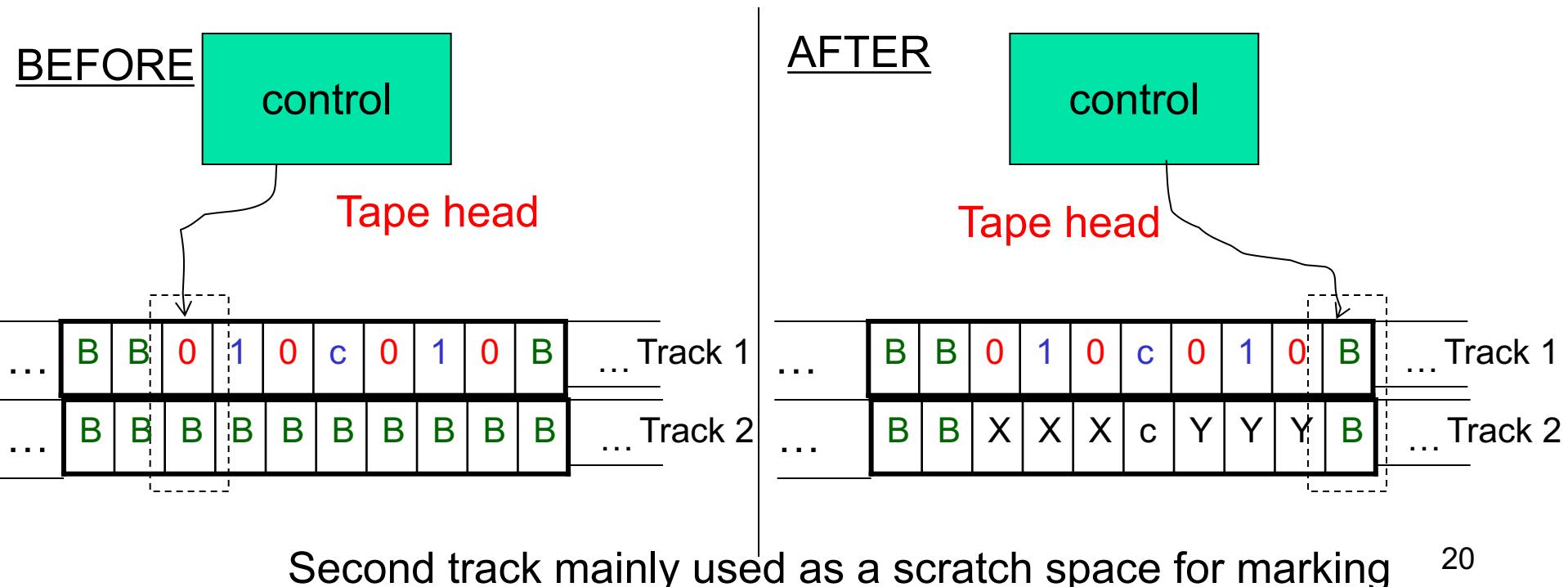
- TM with multiple tracks,  
but just one unified tape head



# Multi-Track TMs

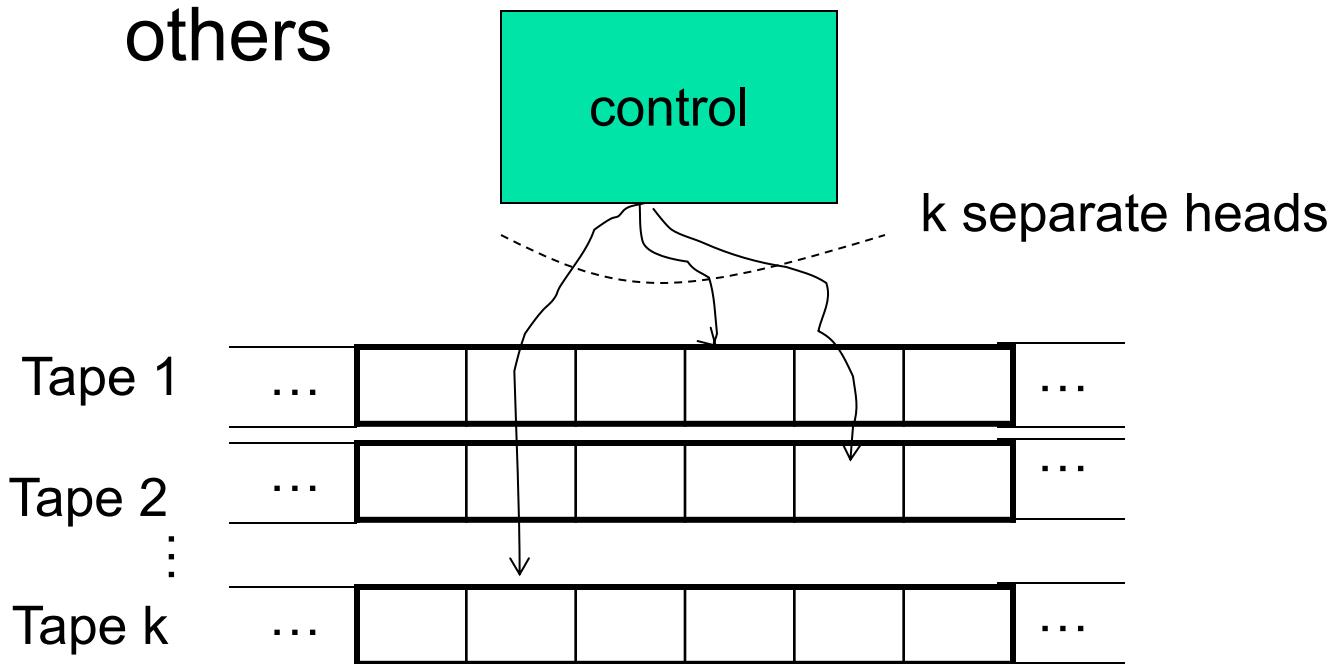
- TM with multiple “tracks” but just one head

E.g., TM for  $\{wcw \mid w \in \{0,1\}^*\}$   
but w/o modifying original input string



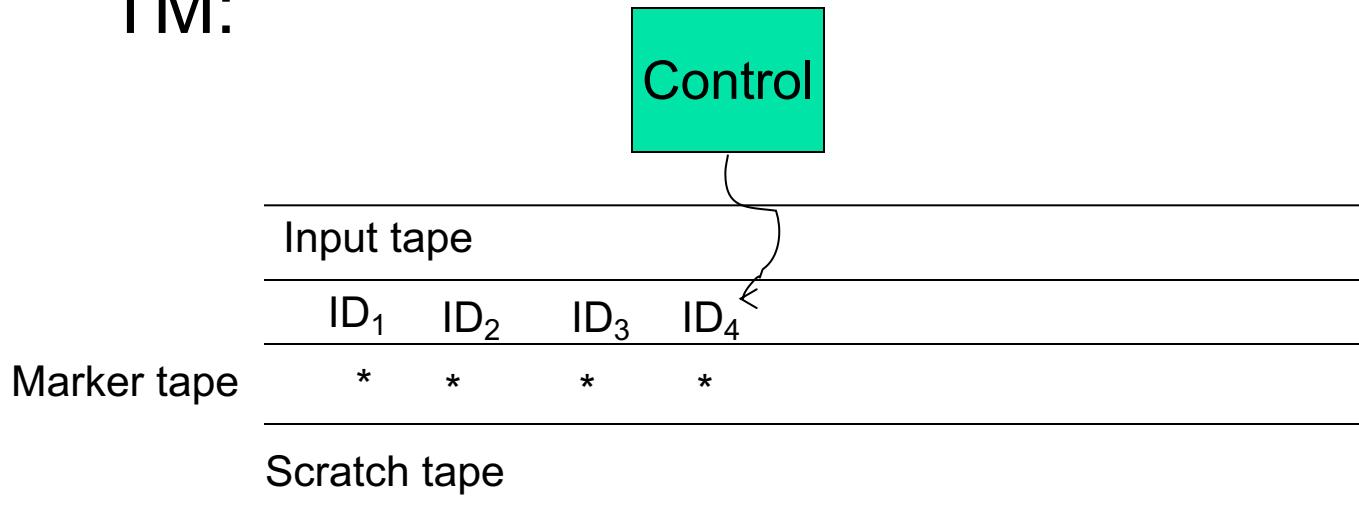
# **Multi-tape Turing Machines**

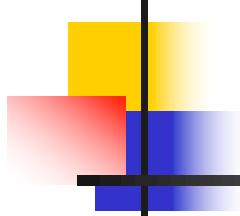
- TM with multiple tapes, *each tape with a separate head*
  - Each head can move independently of the others



# Non-deterministic TMs

- A TM can have non-deterministic moves:
  - $\delta(q,X) = \{ (q_1,Y_1,D_1), (q_2,Y_2,D_2), \dots \}$
- Simulation using a multitape deterministic TM:

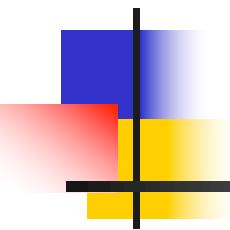




# Summary

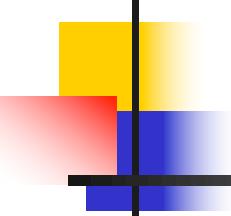
---

- TMs == Recursively Enumerable languages
- TMs can be used as both:
  - Language recognizers
  - Calculators/computers
- ***Basic TM is equivalent to all the below:***
  1. *TM + storage*
  2. *Multi-track TM*
  3. *Multi-tape TM*
  4. *Non-deterministic TM*
- TMs are like universal computing machines with unbounded storage



# Undecidability

Reading: Chapter 8 & 9

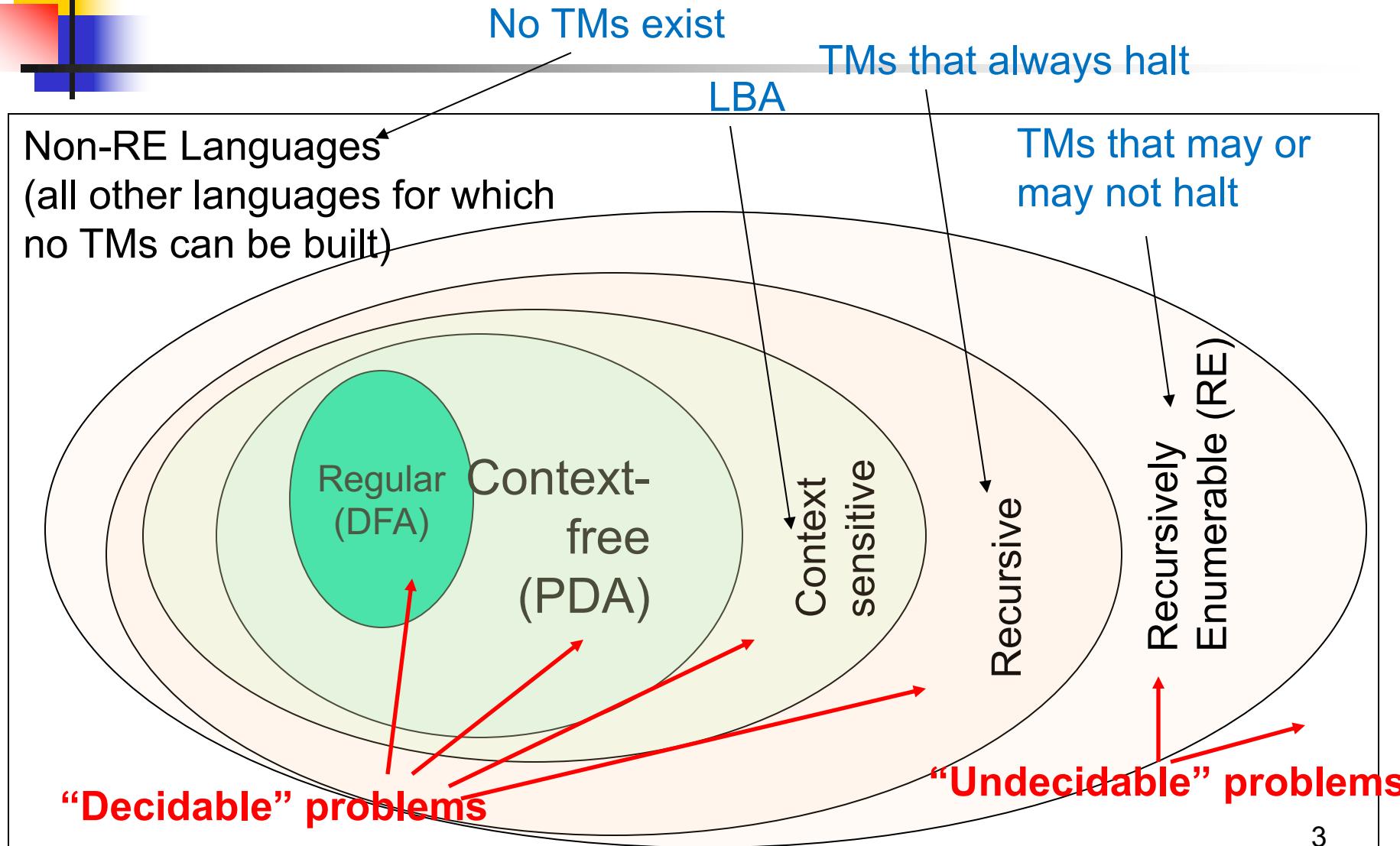


# Decidability vs. Undecidability

---

- There are two types of TMs (based on halting):  
*(Recursive)*  
TMs that **always halt**, no matter accepting or non-accepting ≡ **DECIDABLE PROBLEMS**
- *(Recursively enumerable)*  
TMs that **are guaranteed to halt only on acceptance**. If non-accepting, it may or may not halt (i.e., could loop forever).
- **Undecidability:**
  - Undecidable problems are those that are **not** recursive

# Recursive, RE, Undecidable languages



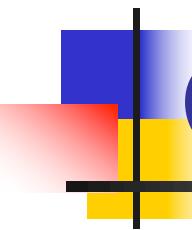
# Recursive Languages & Recursively Enumerable (RE) languages

- Any TM for a Recursive language is going to look like this:



- Any TM for a Recursively Enumerable (RE) language is going to look like this:



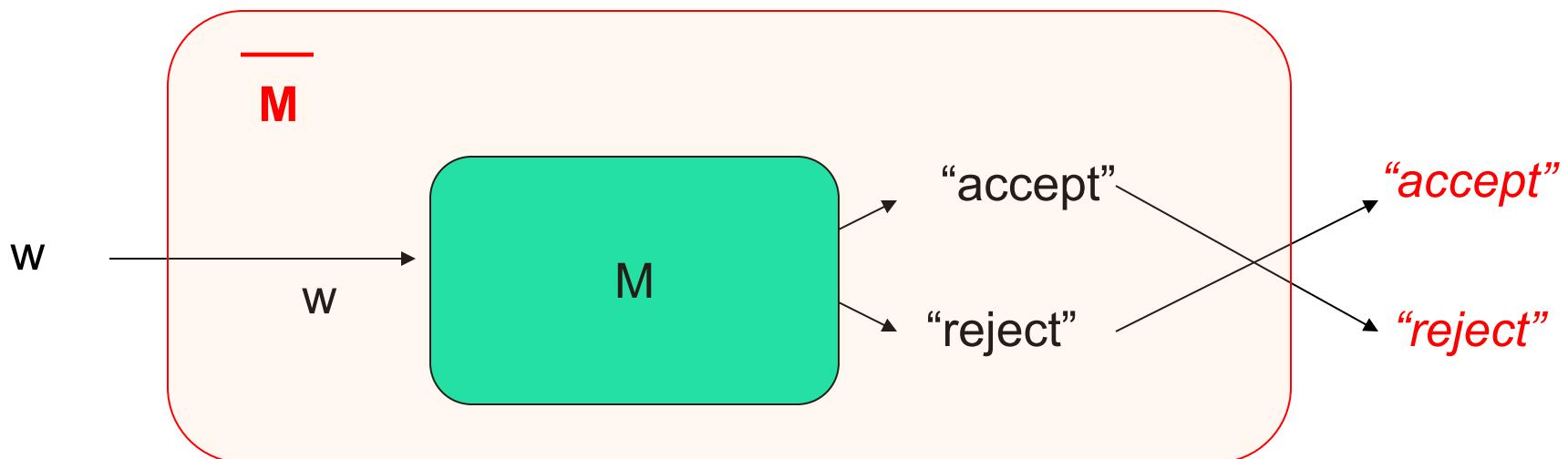


## Closure Properties of:

- the Recursive language class, and
- the Recursively Enumerable language class

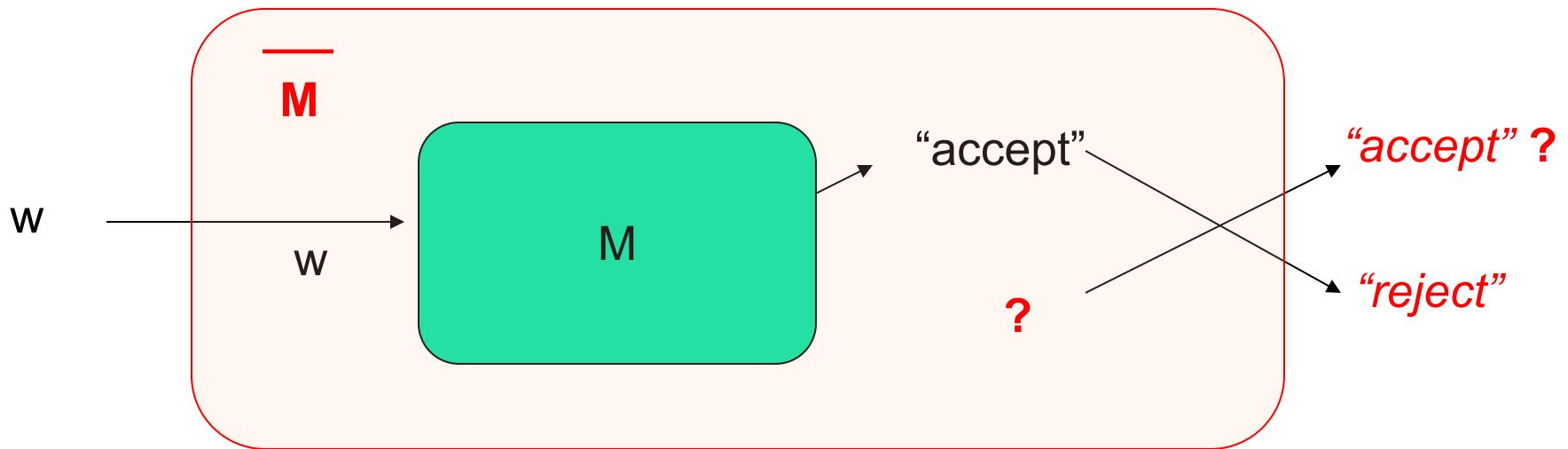
# Recursive Languages are closed under complementation

- If  $L$  is Recursive,  $\overline{L}$  is also Recursive



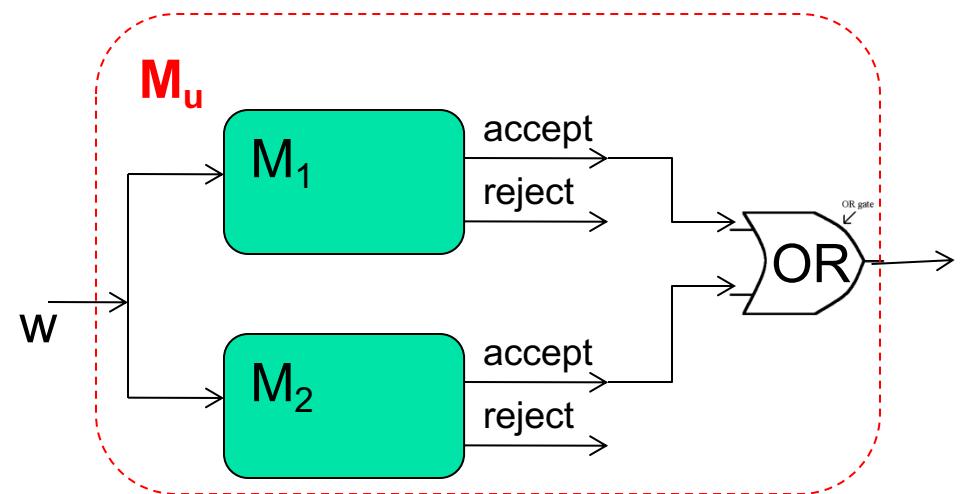
# Are Recursively Enumerable Languages closed under complementation? (NO)

- If  $L$  is RE,  $\overline{L}$  need not be RE



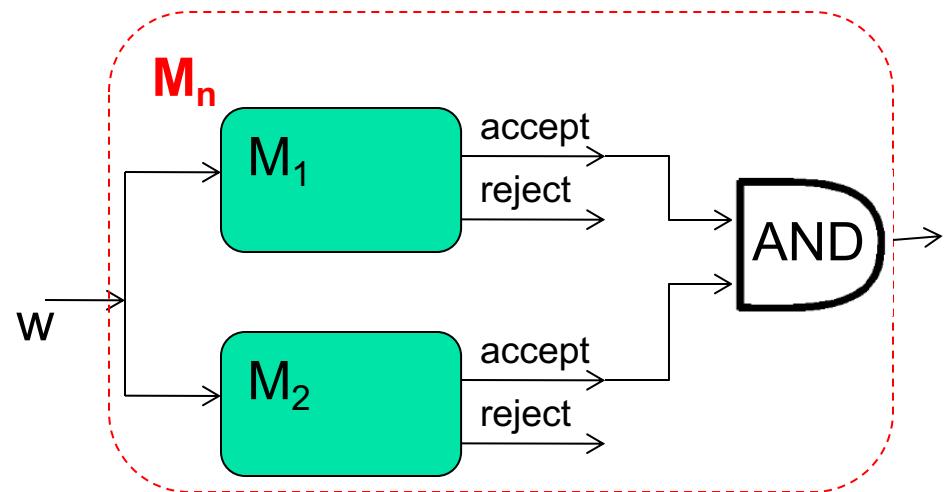
# Recursive Langs are closed under Union

- Let  $M_u = \text{TM for } L_1 \cup L_2$
- $M_u$  construction:
  - Make 2-tapes and copy input  $w$  on both tapes
  - Simulate  $M_1$  on tape 1
  - Simulate  $M_2$  on tape 2
  - If either  $M_1$  or  $M_2$  accepts, then  $M_u$  accepts
  - Otherwise,  $M_u$  rejects.



# Recursive Langs are closed under Intersection

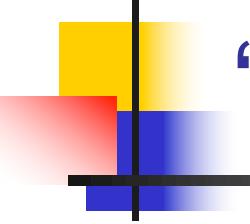
- Let  $M_n = \text{TM for } L_1 \cap L_2$
- $M_n$  construction:
  1. Make 2-tapes and copy input  $w$  on both tapes
  2. Simulate  $M_1$  on tape 1
  3. Simulate  $M_2$  on tape 2
  4. If  $M_1$  AND  $M_2$  accepts, then  $M_n$  accepts
  5. Otherwise,  $M_n$  rejects.



# Other Closure Property Results

---

- Recursive languages are also closed under:
  - Concatenation
  - Kleene closure (star operator)
  - Homomorphism, and inverse homomorphism
- RE languages are closed under:
  - Union, intersection, concatenation, Kleene closure
  
- RE languages are *not* closed under:
  - complementation



# “Languages” vs. “Problems”

A “language” is a set of strings

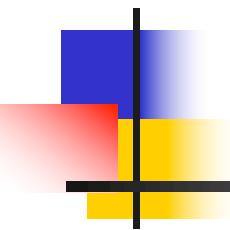
Any “problem” can be expressed as a set of all strings that are of the form:

- “<input, output>”

e.g., Problem  $(a+b)$   $\equiv$  Language of strings of the form { “ $a\#b$ ,  $a+b$  ” }

$\implies$  Every problem also corresponds to a language!!

Think of the language for a “problem”  $\equiv$  a *verifier* for the problem

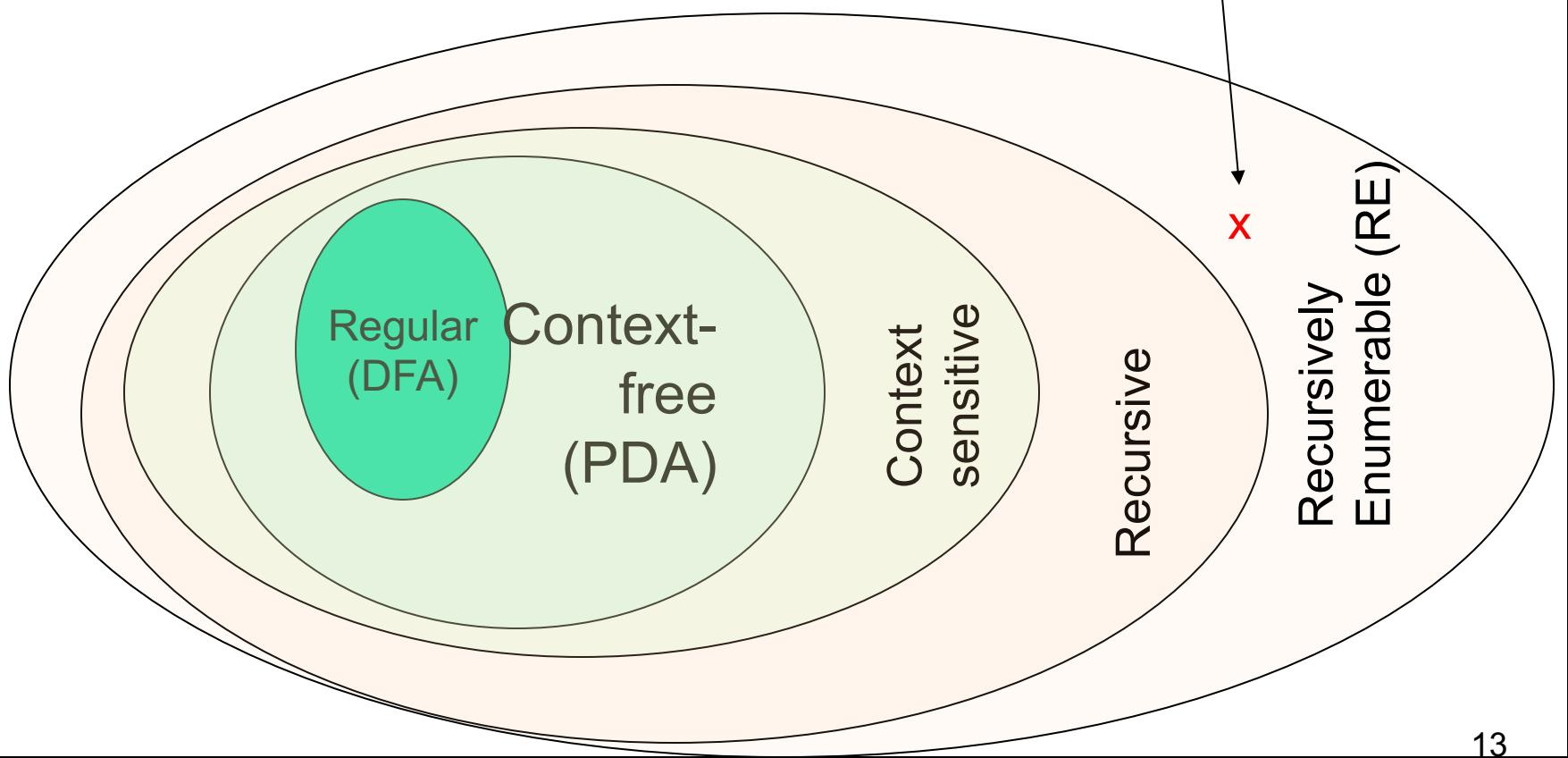


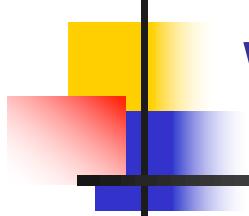
# *The Halting Problem*

An example of a recursive enumerable problem that is also undecidable

## The Halting Problem

Non-RE Languages



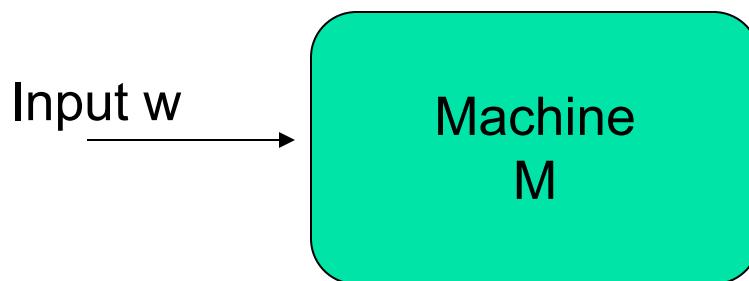


# What is the Halting Problem?

---

Definition of the “halting problem”:

- Does a given Turing Machine  $M$  halt on a given input  $w$ ?



A Turing Machine simulator

# The Universal Turing Machine

- A universal Turing machine (UTM) is a Turing machine that simulates an arbitrary Turing machine on arbitrary input.
- Given: TM  $M$  & its input  $w$
- Aim: Build another TM called “ $H$ ”, that will output:
  - “accept” if  $M$  accepts  $w$ , and
  - “reject” otherwise
- An algorithm for  $H$ :
  - Simulate  $M$  on  $w$

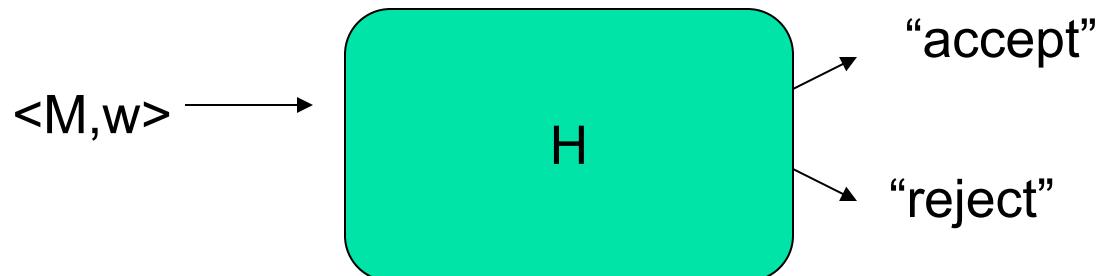
Implies:  $H$  is in RE

$$\text{■ } H(\langle M, w \rangle) = \begin{cases} \text{accept,} & \text{if } M \text{ accepts } w \\ \text{reject,} & \text{if } M \text{ does not accept } w \end{cases}$$

Question: If  $M$  does *not* halt on  $w$ , what will happen to  $H$ ?

# A Claim

- Claim: No H that is always guaranteed to halt, can exist!
- Proof: (Alan Turing, 1936)
  - By contradiction, let us assume H exists

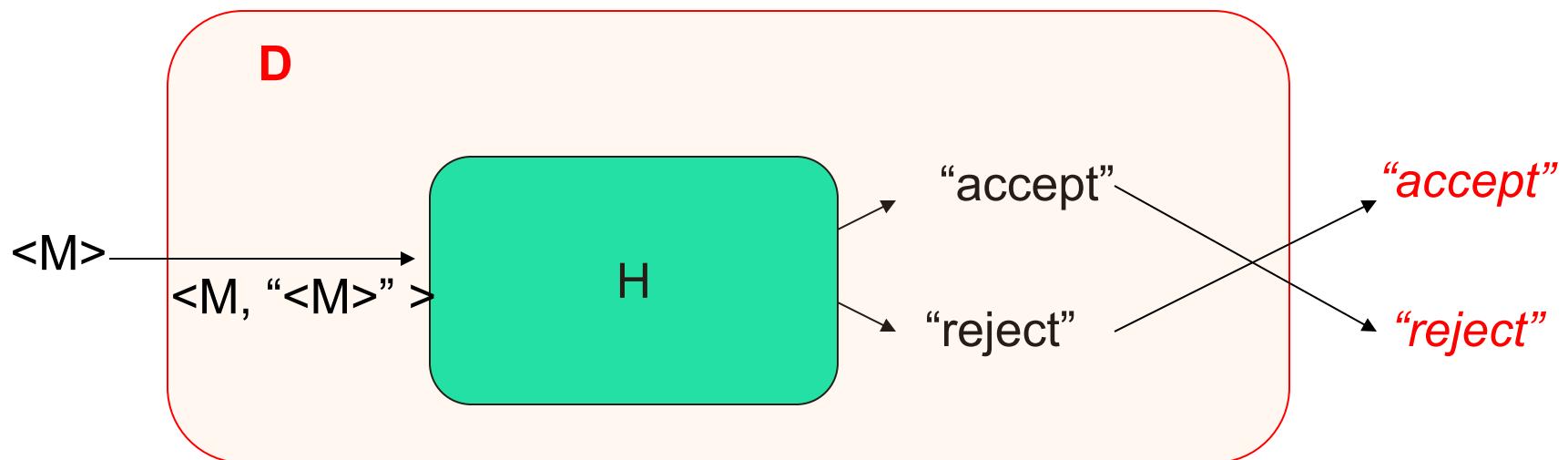


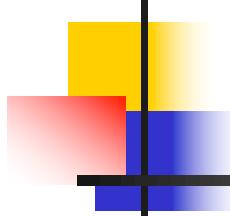
Therefore, if H exists  $\rightarrow$  D also should exist.

But can such a D exist? (if not, then H also cannot exist)

## HP Proof (step 1)

- Let us construct a new TM **D** using H as a subroutine:
  - On input  $\langle M \rangle$ :
    1. Run H on input  $\langle M, \langle M \rangle \rangle$ ; // (i.e., run M on M itself)
    2. Output the *opposite* of what H outputs;





# HP Proof (step 2)

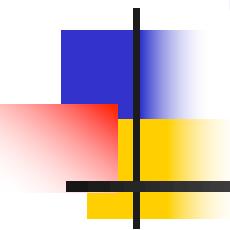
- The notion of inputting “ $\langle M \rangle$ ” to  $M$  itself
  - A program can be input to itself (e.g., a compiler is a program that takes any program as input)

$$D(\langle M \rangle) = \begin{cases} \text{accept,} & \text{if } M \text{ does not accept } \langle M \rangle \\ \text{reject,} & \text{if } M \text{ accepts } \langle M \rangle \end{cases}$$

Now, what happens if  $D$  is input to itself?

$$D(\langle D \rangle) = \begin{cases} \text{accept,} & \text{if } D \text{ does not accept } \langle D \rangle \\ \text{reject,} & \text{if } D \text{ accepts } \langle D \rangle \end{cases}$$

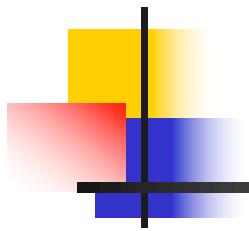
**A contradiction!!!** ==> Neither  $D$  nor  $H$  can exist.



# The Diagonalization Language

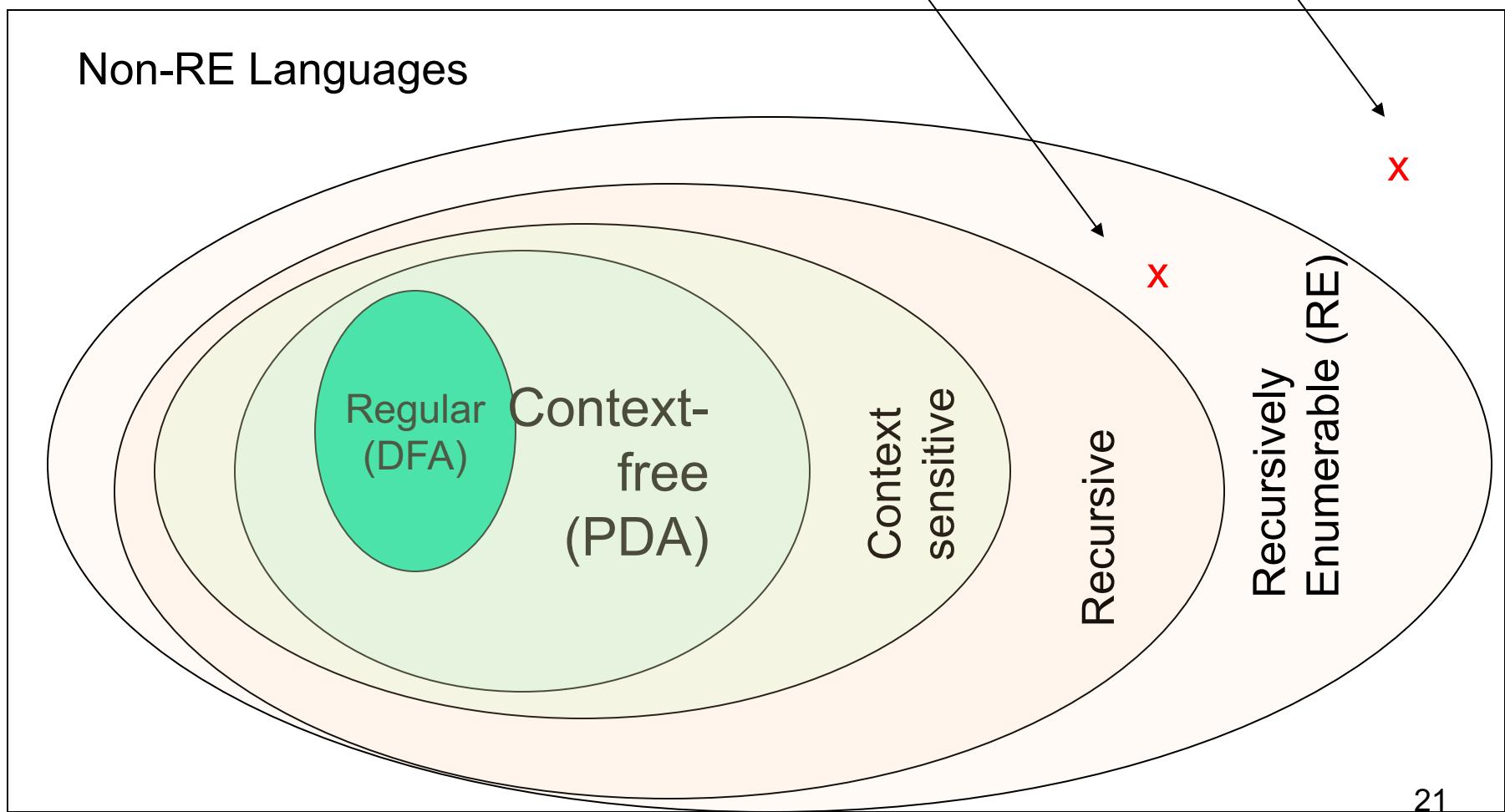
**Example of a language that is  
not recursive enumerable**

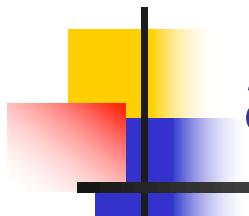
**(i.e, no TMs exist)**



## The Diagonalization language

### The Halting Problem

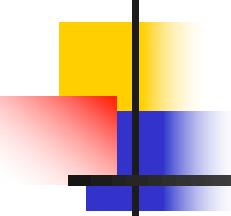




# A Language about TMs & acceptance

---

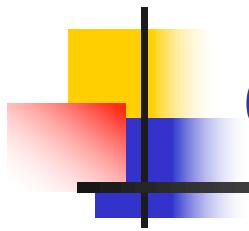
- Let  $L$  be the language of all strings  $\langle M, w \rangle$  s.t.:
  1.  $M$  is a TM (coded in binary) with input alphabet also binary
  2.  $w$  is a binary string
  3.  $M$  accepts input  $w$ .



# Enumerating all binary strings

---

- Let  $w$  be a binary string
- Then  $1w \equiv i$ , where  $i$  is some integer
  - E.g., If  $w=\varepsilon$ , then  $i=1$ ;
  - If  $w=0$ , then  $i=2$ ;
  - If  $w=1$ , then  $i=3$ ; so on...
- If  $1w \equiv i$ , then call  $w$  as the  $i^{\text{th}}$  word or  $i^{\text{th}}$  binary string, denoted by  $w_i$ .
- ==> A canonical ordering of all binary strings:
  - $\{\varepsilon, 0, 1, 00, 01, 10, 11, 000, 100, 101, 110, \dots\}$
  - $\{w_1, w_2, w_3, w_4, \dots, w_i, \dots\}$



# Any TM M can also be binary-coded

---

- $M = \{ Q, \{0,1\}, \Gamma, \delta, q_0, B, F \}$
- Map all states, tape symbols and transitions to integers ( $\Rightarrow$ binary strings)
- $\delta(q_i, X_j) = (q_k, X_l, D_m)$  will be represented as:
  - $\Rightarrow 0^i 1 0^j 1 0^k 1 0^l 1 0^m$
- Result: Each TM can be written down as a long binary string
- $\Rightarrow$  Canonical ordering of TMs:
  - $\{M_1, M_2, M_3, M_4, \dots, M_i, \dots\}$

# The Diagonalization Language

- $L_d = \{ w_i \mid w_i \notin L(M_i) \}$ 
  - The language of all strings whose corresponding machine does *not* accept itself (i.e., its own code)

		(input word w)						
		j	1	2	3	4	...	
		i	1	0	1	0	1	...
		2	1	1	0	0	...	
		3	0	1	0	1	...	
		4	1	0	0	1	...	
		⋮	⋮	⋮	⋮	⋮	⋮	

diagonal

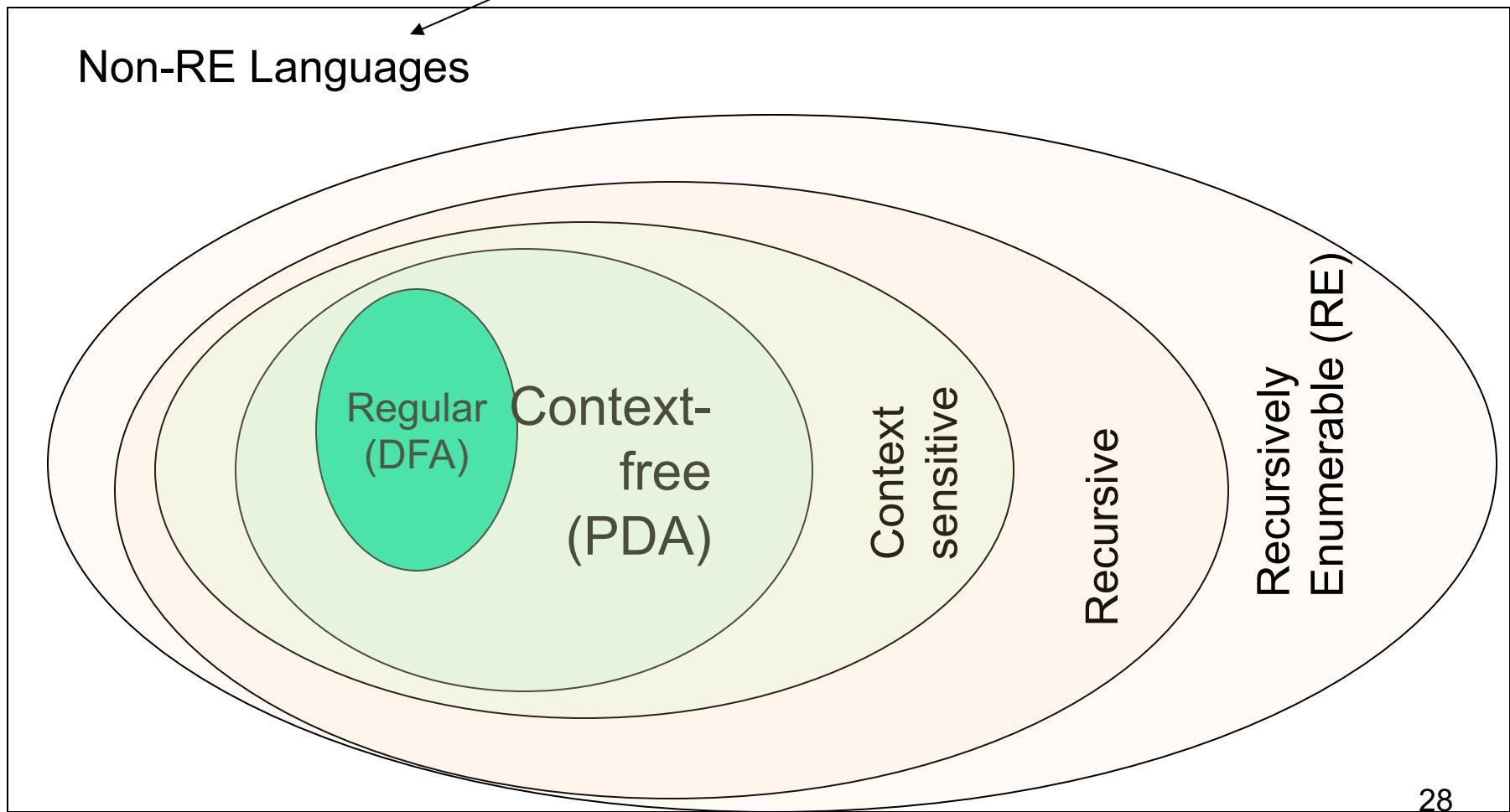
- Table:  $T[i,j] = 1$ , if  $M_i$  accepts  $w_j$   
= 0, otherwise.
- Make a new language called  
 $L_d = \{w_i \mid T[i,i] = 0\}$

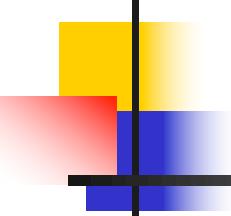
# Why should there be languages that do not have TMs?

We thought TMs can solve  
everything!!

# Non-RE languages

How come there are languages here?  
(e.g., diagonalization language)



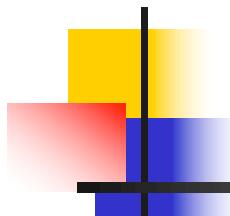


# One Explanation

---

***There are more languages than TMs***

- By pigeon hole principle:
- ==> some languages cannot have TMs
- But how do we show this?
- Need a way to “*count & compare*” two infinite sets (languages and TMs)



# How to count elements in a set?

---

Let A be a set:

- If A is finite ==> counting is trivial
- If A is infinite ==> how do we count?
- And, how do we compare two infinite sets by their size?

# Cantor's definition of set “size” for infinite sets (1873 A.D.)

Let  $N = \{1, 2, 3, \dots\}$  (all natural numbers)

Let  $E = \{2, 4, 6, \dots\}$  (all even numbers)

Q) Which is bigger?

- A) Both sets are of the same size
  - “**Countably infinite**”
  - Proof: Show by one-to-one, onto set correspondence from  
 $N \Rightarrow E$

i.e, for every element in  $N$ ,  
there is a unique element in  $E$ ,  
and vice versa.

$n$	$f(n)$
1	2
2	4
3	6
.	.
.	.

Really, really big sets!  
(even bigger than countably infinite sets)

# Uncountable sets

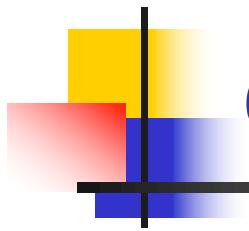
## Example:

- Let  $\mathbb{R}$  be the set of all real numbers
- Claim:  $\mathbb{R}$  is uncountable

$n$	$f(n)$
1	3 . <u>1</u> 4 1 5 9 ...
2	5 . 5 <u>5</u> 5 5 5 ...
3	0 . 1 2 <u>3</u> 4 5 ...
4	0 . 5 1 4 <u>3</u> 0 ...
.	.
.	.

Build  $x$  s.t.  $x$  cannot possibly occur in the table

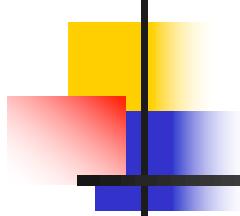
E.g.  $x = 0 . 2 6 4 4 \dots$



# Therefore, some languages cannot have TMs...

---

- The set of all TMs is countably infinite
- The set of all Languages is uncountable
- ==> There should be some languages without TMs ( by PHP)



# Summary

---

- Problems vs. languages
- Decidability
  - Recursive
- Undecidability
  - Recursively Enumerable
  - Not RE
  - Examples of languages
- The diagonalization technique
- Reducability