Introduction to Formal Languages and Automata Theory

Dr. Mousumi Dutt

Lecture 2

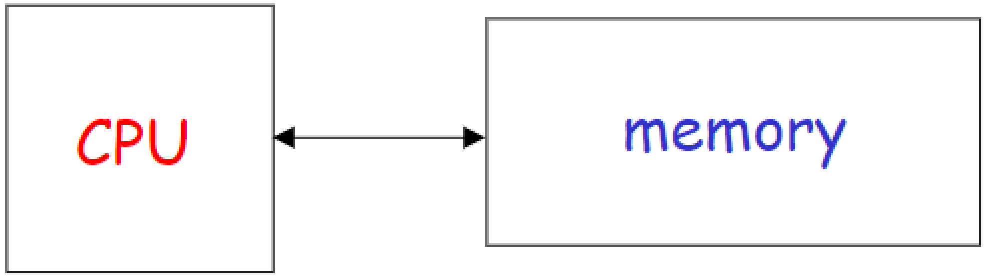
What is Automata

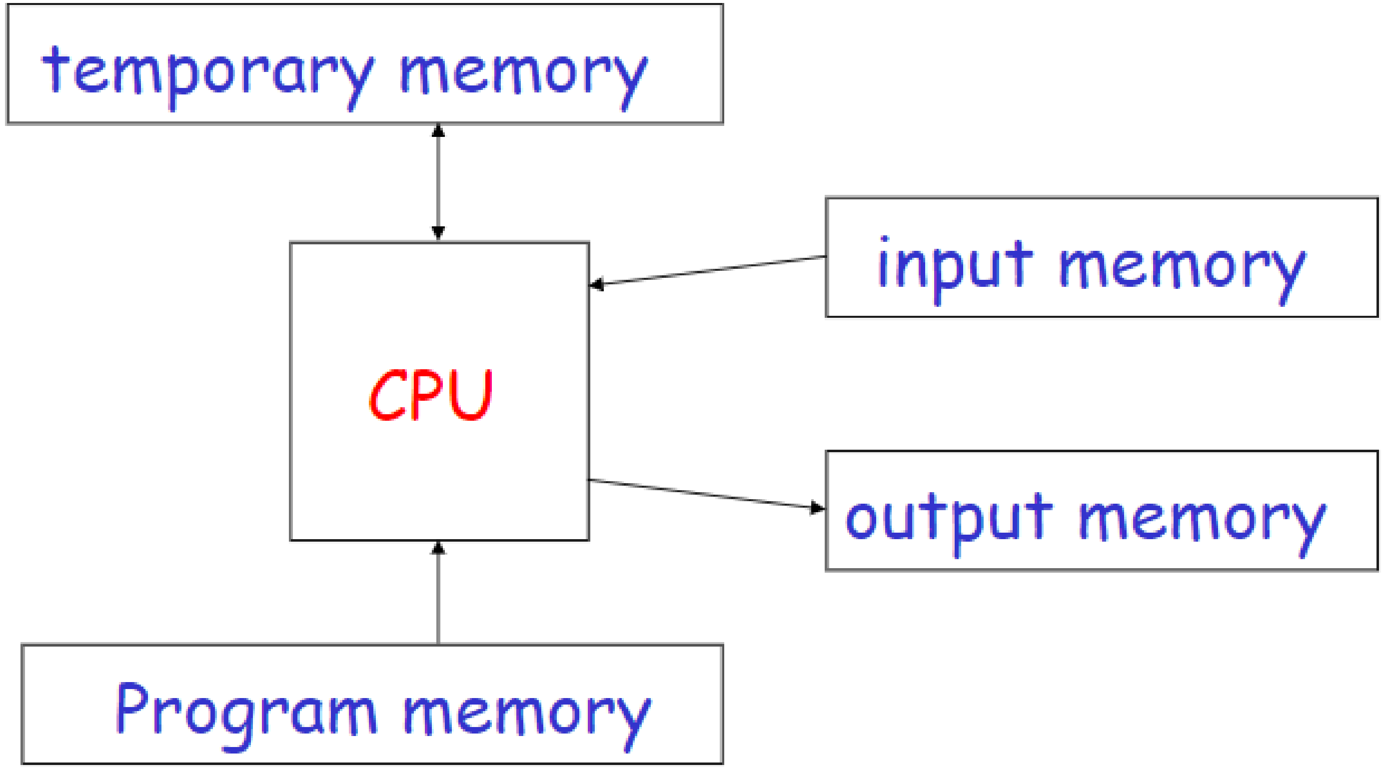
- Automaton = A self-operating machine or mechanism (Dictionary definition), plural is Automata.
- Automata = abstract computing devices
- Automata theory = the study of abstract machines (or more appropriately, abstract 'mathematical' machines or systems), and the computational problems that can be solved using these machines.
  - Mathematical models of computation
  - Finite automata
  - Push-down automata
  - Turing machines

History

- 1930s : Alan Turing defined machines more powerful than any in existence, or even any that we could imagine – Goal was to establish the boundary between what was and was not computable.
- 1940s/1950s : In an attempt to model "Brain function" researchers defined finite state machines.
- Late 1950s : Linguist Noam Chomsky began the study of Formal Grammars.
- 1960s : A convergence of all this into a formal theory of computer science, with very deep philosophical implications as well as practical applications (compilers, web searching, hardware, A.I., algorithm design, software engineering,...)

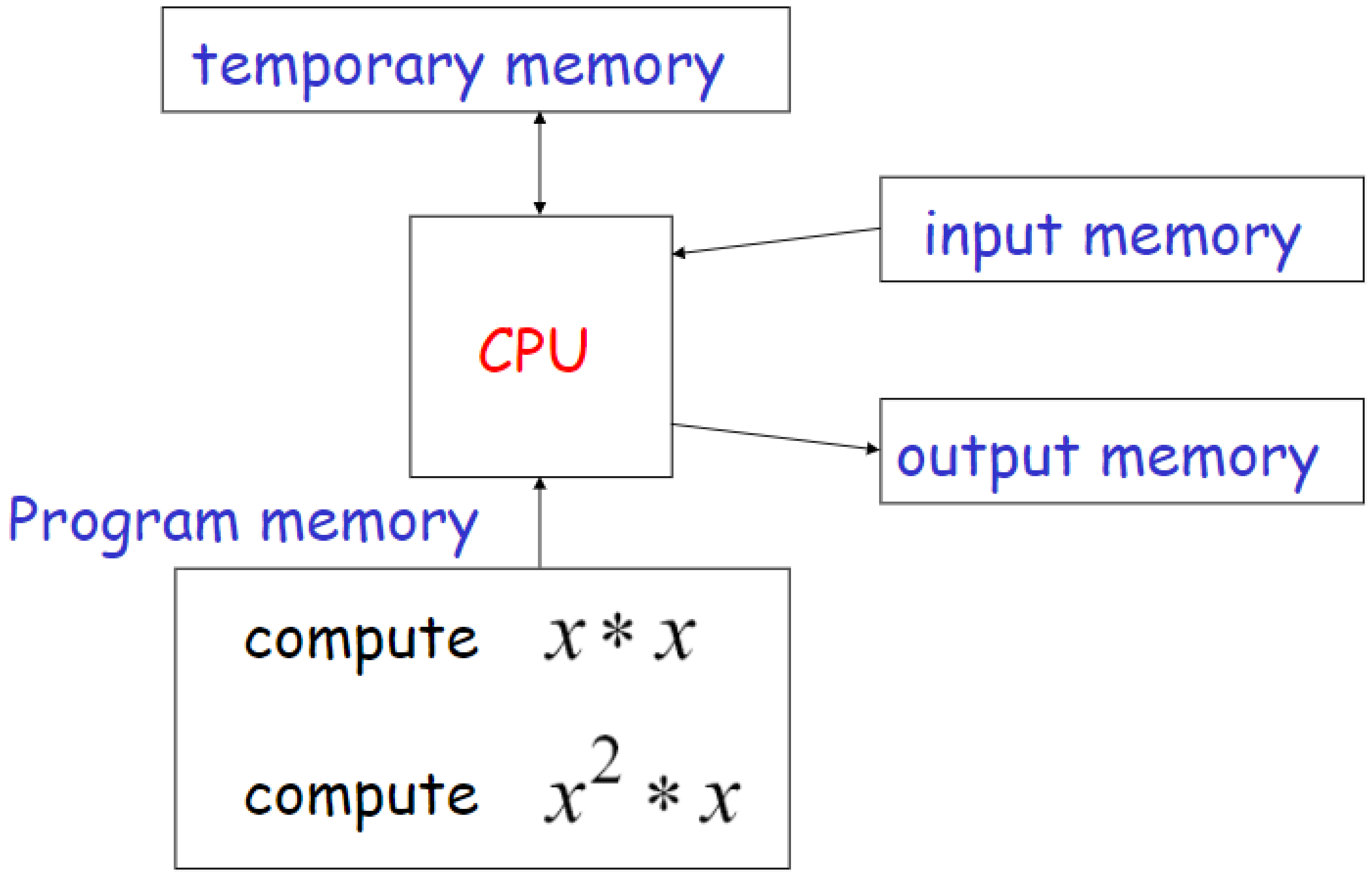
Computation





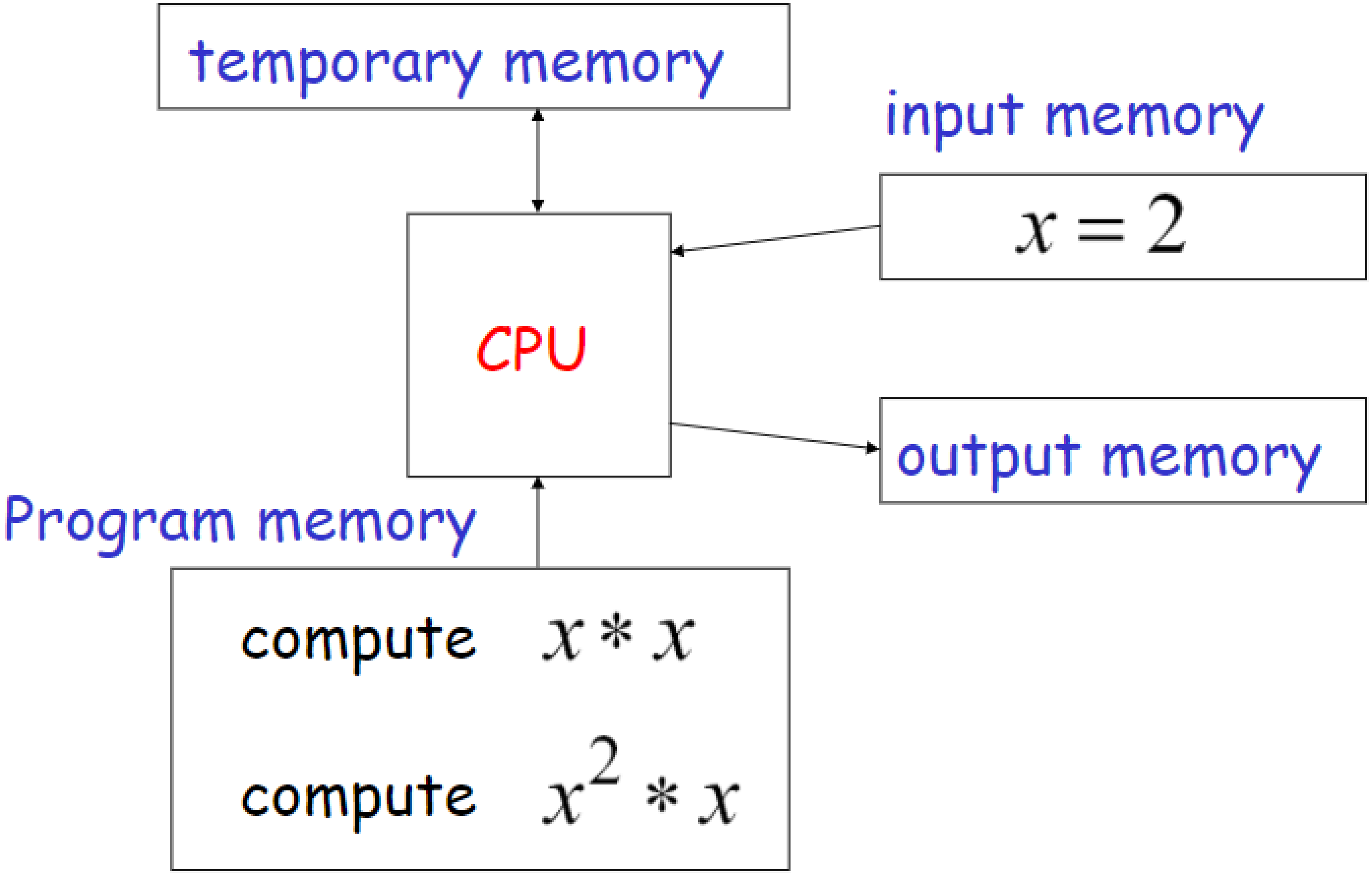
Computation

Example: $f(x) = x^3$

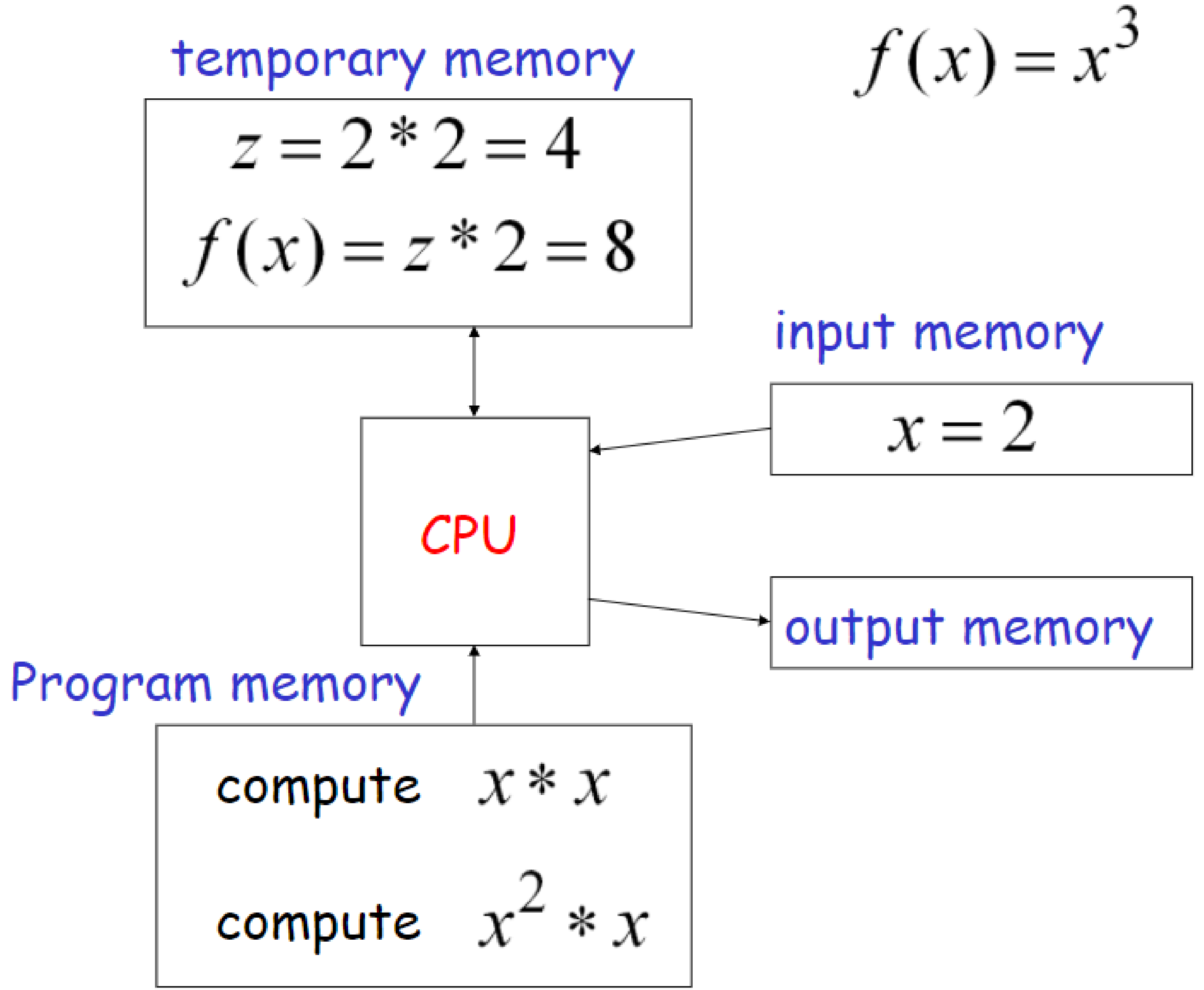


Computation

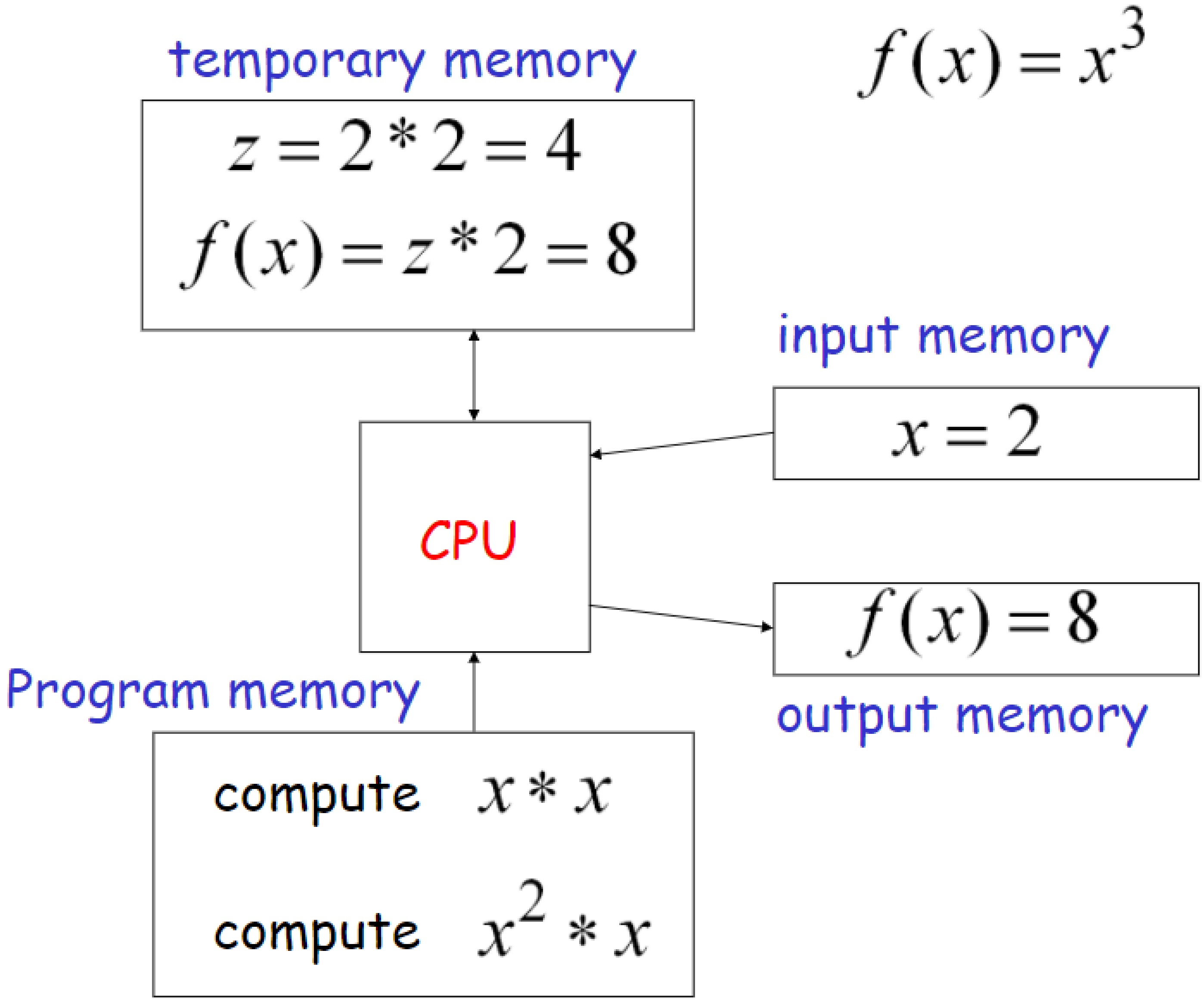
$$f(x) = x^3$$



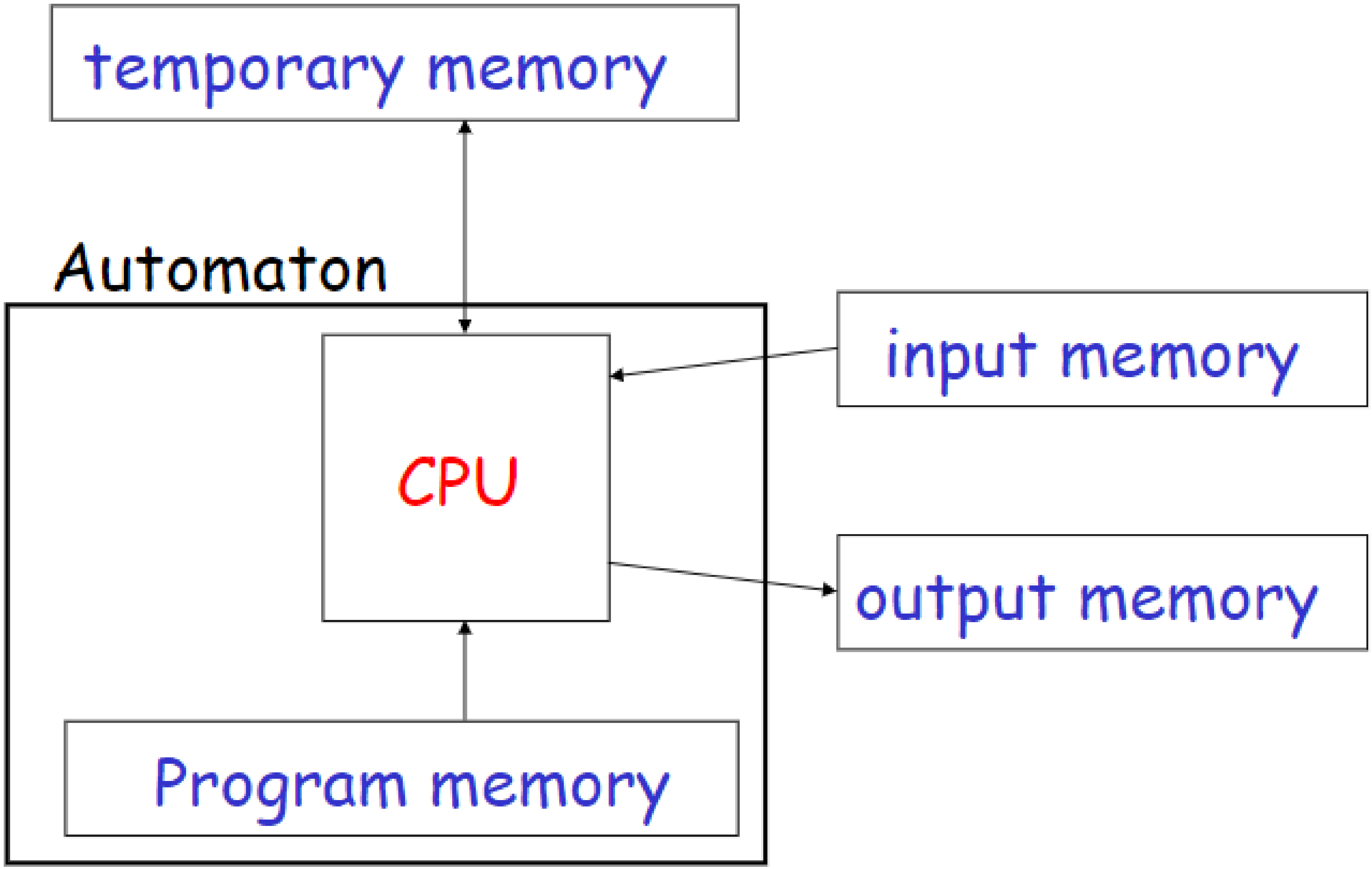
Computation



Computation



Automaton


temporary memory
Automaton
CPU
input memory
output memory
Program memory

Languages and Grammar

An alphabet is a set of symbols:

Or "**words**"

{0,1}

Sentences are strings of symbols:

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

A language is a set of sentences:

L = {000,0100,0010,..}

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

$S \longrightarrow 0A$	$B \longrightarrow 1B$
$A \longrightarrow 1A$	$B \longrightarrow 0F$
$A \longrightarrow 0B$	$F \longrightarrow \varepsilon$

Image source: Nowak et al. Nature, vol 417, 2002

- <u>Languages</u>: *"A language is a collection of sentences of finite length all constructed from a finite alphabet of symbols"*
- <u>Grammars</u>: *"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*

Alphabet

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

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

Strings

A string or word is a finite sequence of symbols chosen from ∑

- ***Empty string is ε (or "epsilon")***

- Length of a string *w*, denoted by "|*w*|", is equal to the *number of (non- ε) characters in the string*
  - *E.g., x = 010100* *|x| = 6*
  - *x = 01 ε 0 ε 1 ε 00 ε* *|x| = ?*

- *xy* = concatentation of two strings *x* and *y*

Empty Strings

The empty string is the string with no occurrences of symbols.

This string, denoted by ε, is a string that may be chosen from any alphabet whatsoever

Powers of an Alphabet

If ∑ is an alphabet, we can express the set of all strings of a certain length from that alphabet by using an exponentiation

Let ∑ be an alphabet.

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

Powers of an Alphabet

- $\sum^{0} = \{\varepsilon\}$

- If $\sum=\{0,1\}$, then
- $\sum^{1} = \{0,1\}$
- $\sum^{2} = \{00,11,01,10\}$

- $\{0,1\}^*= \{\varepsilon, 0, 1, 00, 01, 10, 11, 000, \ldots\}$

- $\sum^* = \sum^+ \cup \{\varepsilon\}$

Concatenation of Strings

Let x and y be strings

xy -> concatenation of string

If x is the string composed of i symbols $x=a_1a_2\ldots a_i$ and y is the string composed of j symbols $y=b_1b_2\ldots b_j$, then xy is the string of length i+j: $xy=a_1a_2\ldots a_ib_1b_2\ldots b_j$.

Substring of a String

A string v is a substring of a string w if and only if there are strings x & y such that, w=xvy

x & y could be ε

x=w & v=y= ε

ε is the substring of every string

w=xv for some x, then v is a suffix of w

w=vy for some y, then v is a prefix of w

More on String

For each string w and each natural number i, the string w^i is defined as

$w^0 = \varepsilon$, the empty string

$w^{i+1} = w^i \circ w$ for each $i >= 0$

Thus, $w^1 = w$ & $do^2 = dodo$

Reversal of a String

The reversal of a string w, denoted by w^R, is the string "spelled backward"

A formal definition can be given by induction on the length of a string

1) If w is a string of length 0, then $w^R = w = \varepsilon$
2) If w is a string of length n+1>0, then w=ua for $a \in \Sigma$, and $w^R = au^R$

Reversal of a String

Basis Steps:

$|x|=0$. Then $x=\varepsilon$, and $(wx)^R=(w\,\varepsilon)^R=w^R=\varepsilon\, w^R=\varepsilon^R\, w^R=x^R\, w^R$

Induction Hypothesis:

If $|x|<=n$, then $(wx)^R=x^R\, w^R$

Induction Step:

Let $|x|=n+1$

Then $x=ua$ for some $u \in \sum^*$ and $a \in \sum$ such that $|u|=n$

Reversal of a String

$(wx)^R = (w(ua))^R$ since $x=ua$

$=((wu)a)^R$ since concatenation is associative

$=a(wu)^R$ by the definition of reversal of $(wu)a$

$=au^R w^R$ by the induction hypothesis

$=(ua)^R w^R$ by the definition of reversal of ua

$=x^R w^R$ since $x=ua$

Palindrome of a String

A palindrome is a string which is the same whether written forward or backward

Eg., Malayalam

A palindrome of even length can be obtained by concatenation of a string and its reverse

Levi's Theorem

Let v, w, x, and y $\in \sum^*$ and vw=xy. Then:

1) There exist a unique string z in $\sum^*$ such that v=xz and y=zw if $|v|>|x|$

2) v=x, y=w, i.e., z= ε if $|v|=|x|$

3) There exists a unique string z in $\sum^*$ such that x=vz and w=zy if $|v|<|x|$

Languages

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

➔ this is because ∑* is the set of all strings (of all possible length including 0) over the given alphabet ∑

<u>Examples:</u>

1. Let L be *the* language of <u>all strings consisting of *n* 0's followed by *n* 1's</u>:

   L = {ε, 01, 0011, 000111,…}

2. Let L be *the* language of <u>all strings of with equal number of 0's and 1's</u>:

   L = {ε, 01, 10, 0011, 1100, 0101, 1010, 1001,…}

   Canonical ordering of strings in the language

Languages

A language over $\sum$ need not include strings with all the symbols of $\sum$, so once we have established that L is a language over $\sum$, we also know it is a language over any alphabet that is a superset of $\sum$

- A string in a language L will be called a sentence of L
- $\sum^*$ is a language for any alphabet $\sum$
- Ø, the empty language, is a language over any alphabet
- $\{\varepsilon\}$, the language consisting of only empty string, is also a language over any alphabet

Let $L = \{\varepsilon\}$; Is $L=\emptyset$?
No

More on next class…