# 2.2
# Regular Expressions

# Regular Expressions

A way to denote regular languages

- simple patterns to describe related strings
- useful in
  - text search (editors, Unix/grep, emacs)
  - compilers: lexical analysis
  - compact way to represent interesting/useful languages
  - dates back to 50's: Stephen Kleene
    who has a star names after him.

# Inductive Definition

A regular expression r over an alphabet $\Sigma$ is one of the following:

**Base cases:**

- $\emptyset$ denotes the language $\emptyset$
- $\epsilon$ denotes the language $\{\epsilon\}$.
- $a$ denote the language $\{a\}$.

**Inductive cases:** If $\mathrm{r}_1$ and $\mathrm{r}_2$ are regular expressions denoting languages $R_1$ and $R_2$ respectively then,

- $(\mathrm{r}_1 + \mathrm{r}_2)$ denotes the language $R_1 \cup R_2$
- $(\mathrm{r}_1 \bullet \mathrm{r}_2) = r_1 \bullet r_2 = (\mathrm{r}_1\mathrm{r}_2)$ denotes the language $R_1R_2$
- $(\mathrm{r}_1)^*$ denotes the language $R_1^*$

# Inductive Definition

A regular expression $\mathsf{r}$ over an alphabet $\Sigma$ is one of the following:

**Base cases:**

- $\emptyset$ denotes the language $\emptyset$
- $\epsilon$ denotes the language $\{\epsilon\}$.
- $a$ denote the language $\{a\}$.

**Inductive cases:** If $\mathsf{r}_1$ and $\mathsf{r}_2$ are regular expressions denoting languages $R_1$ and $R_2$ respectively then,

- $(\mathsf{r}_1 + \mathsf{r}_2)$ denotes the language $R_1 \cup R_2$
- $(\mathsf{r}_1 \bullet \mathsf{r}_2) = r_1 \bullet r_2 = (\mathsf{r}_1 \mathsf{r}_2)$ denotes the language $R_1 R_2$
- $(\mathsf{r}_1)^*$ denotes the language $R_1^*$

# Regular Languages vs Regular Expressions

| **Regular Languages** | **Regular Expressions** |
|---|---|
| $\emptyset$ regular | $\emptyset$ denotes $\emptyset$ |
| $\{\epsilon\}$ regular | $\epsilon$ denotes $\{\epsilon\}$ |
| $\{a\}$ regular for $a \in \Sigma$ | a denote $\{a\}$ |
| $R_1 \cup R_2$ regular if both are | $\mathrm{r}_1 + \mathrm{r}_2$ denotes $R_1 \cup R_2$ |
| $R_1 R_2$ regular if both are | $\mathrm{r}_1 \bullet \mathrm{r}_2$ denotes $R_1 R_2$ |
| $R^*$ is regular if $R$ is | $\mathrm{r}^*$ denote $R^*$ |

Regular expressions denote regular languages — they explicitly show the operations that were used to form the language

# Notation and Parenthesis

- For a regular expression r, $L(r)$ is the language denoted by r. Multiple regular expressions can denote the same language!
  **Example:** $(0+1)$ and $(1+0)$ denote same language $\{0,1\}$
- Two regular expressions $r_1$ and $r_2$ are equivalent if $L(r_1) = L(r_2)$.
- Omit parenthesis by adopting precedence order: $*$, concatenate, $+$.
  **Example:** $r^*s + t = ((r^*)s) + t$
- Omit parenthesis by associativity of each of these operations.
  **Example:** $rst = (rs)t = r(st)$, $r + s + t = r + (s + t) = (r + s) + t$.
- Superscript $+$. For convenience, define $r^+ = rr^*$. Hence if $L(r) = R$ then $L(r^+) = R^+$.
- Other notation: $r + s$, $r \cup s$, $r|s$ all denote union. $rs$ is sometimes written as $r \bullet s$.

# Notation and Parenthesis

- For a regular expression r, $\boldsymbol{L}(\mathrm{r})$ is the language denoted by r. Multiple regular expressions can denote the same language!
  **Example:** $(0+1)$ and $(1+0)$ denote same language $\{0,1\}$
- Two regular expressions $\mathrm{r}_1$ and $\mathrm{r}_2$ are equivalent if $\boldsymbol{L}(\mathrm{r}_1) = \boldsymbol{L}(\mathrm{r}_2)$.
- Omit parenthesis by adopting precedence order: $*$, concatenate, $+$.
  **Example:** $r^*s + t = ((r^*)s) + t$
- Omit parenthesis by associativity of each of these operations.
  **Example:** $rst = (rs)t = r(st)$, $r + s + t = r + (s + t) = (r + s) + t$.
- Superscript $+$. For convenience, define $\mathrm{r}^+ = \mathrm{rr}^*$. Hence if $\boldsymbol{L}(\mathrm{r}) = R$ then $\boldsymbol{L}(\mathrm{r}^+) = R^+$.
- Other notation: $r + s$, $r \cup s$, $r|s$ all denote union. $rs$ is sometimes written as $r \bullet s$.

# Notation and Parenthesis

- For a regular expression r, $\boldsymbol{L}(\mathrm{r})$ is the language denoted by r. Multiple regular expressions can denote the same language!
  **Example:** $(0+1)$ and $(1+0)$ denote same language $\{0,1\}$
- Two regular expressions $\mathrm{r}_1$ and $\mathrm{r}_2$ are equivalent if $\boldsymbol{L}(\mathrm{r}_1) = \boldsymbol{L}(\mathrm{r}_2)$.
- Omit parenthesis by adopting precedence order: $*$, concatenate, $+$.
  **Example:** $r^*s + t = ((r^*)s) + t$
- Omit parenthesis by associativity of each of these operations.
  **Example:** $rst = (rs)t = r(st)$, $r + s + t = r + (s + t) = (r + s) + t$.
- Superscript $+$. For convenience, define $\mathrm{r}^+ = \mathrm{r}\mathrm{r}^*$. Hence if $\boldsymbol{L}(\mathrm{r}) = R$ then $\boldsymbol{L}(\mathrm{r}^+) = R^+$.
- Other notation: $r + s$, $r \cup s$, $r|s$ all denote union. $rs$ is sometimes written as $r \bullet s$.

# Notation and Parenthesis

- For a regular expression $\mathrm{r}$, $L(\mathrm{r})$ is the language denoted by $\mathrm{r}$. Multiple regular expressions can denote the same language!
  **Example:** $(0+1)$ and $(1+0)$ denote same language $\{0,1\}$
- Two regular expressions $\mathrm{r}_1$ and $\mathrm{r}_2$ are equivalent if $L(\mathrm{r}_1) = L(\mathrm{r}_2)$.
- Omit parenthesis by adopting precedence order: $*$, concatenate, $+$.
  **Example:** $r^*s + t = ((r^*)s) + t$
- Omit parenthesis by associativity of each of these operations.
  **Example:** $rst = (rs)t = r(st)$, $r+s+t = r+(s+t) = (r+s)+t$.
- Superscript $+$. For convenience, define $\mathrm{r}^+ = \mathrm{r}\mathrm{r}^*$. Hence if $L(\mathrm{r}) = R$ then $L(\mathrm{r}^+) = R^+$.
- Other notation: $r+s$, $r \cup s$, $r|s$ all denote union. $rs$ is sometimes written as $r \bullet s$.

# Notation and Parenthesis

- For a regular expression $\mathrm{r}$, $\boldsymbol{L}(\mathrm{r})$ is the language denoted by $\mathrm{r}$. Multiple regular expressions can denote the same language!
  **Example:** $(0+1)$ and $(1+0)$ denote same language $\{0,1\}$
- Two regular expressions $\mathrm{r}_1$ and $\mathrm{r}_2$ are equivalent if $\boldsymbol{L}(\mathrm{r}_1) = \boldsymbol{L}(\mathrm{r}_2)$.
- Omit parenthesis by adopting precedence order: $*$, concatenate, $+$.
  **Example:** $r^*s + t = ((r^*)s) + t$
- Omit parenthesis by associativity of each of these operations.
  **Example:** $rst = (rs)t = r(st)$, $r + s + t = r + (s + t) = (r + s) + t$.
- Superscript $+$. For convenience, define $\mathrm{r}^+ = \mathrm{r}\mathrm{r}^*$. Hence if $\boldsymbol{L}(\mathrm{r}) = \boldsymbol{R}$ then $\boldsymbol{L}(\mathrm{r}^+) = \boldsymbol{R}^+$.
- Other notation: $r + s$, $r \cup s$, $r|s$ all denote union. $rs$ is sometimes written as $r \bullet s$.

# Notation and Parenthesis

- For a regular expression $\mathrm{r}$, $\boldsymbol{L}(\mathrm{r})$ is the language denoted by $\mathrm{r}$. Multiple regular expressions can denote the same language!
  **Example:** $(0+1)$ and $(1+0)$ denote same language $\{0,1\}$
- Two regular expressions $\mathrm{r}_1$ and $\mathrm{r}_2$ are equivalent if $\boldsymbol{L}(\mathrm{r}_1) = \boldsymbol{L}(\mathrm{r}_2)$.
- Omit parenthesis by adopting precedence order: $*$, concatenate, $+$.
  **Example:** $r^*s + t = ((r^*)s) + t$
- Omit parenthesis by associativity of each of these operations.
  **Example:** $rst = (rs)t = r(st)$, $r + s + t = r + (s + t) = (r + s) + t$.
- Superscript $+$. For convenience, define $\mathrm{r}^+ = \mathrm{r}\mathrm{r}^*$. Hence if $\boldsymbol{L}(\mathrm{r}) = \boldsymbol{R}$ then $\boldsymbol{L}(\mathrm{r}^+) = \boldsymbol{R}^+$.
- Other notation: $r + s$, $r \cup s$, $r|s$ all denote union. $rs$ is sometimes written as $r \bullet s$.

# Skills

- Given a language $L$ "in mind" (say an English description) we would like to write a regular expression for $L$ (if possible)
- Given a regular expression r we would like to "understand" $L(r)$ (say by giving an English description)

# Skills

- Given a language $L$ "in mind" (say an English description) we would like to write a regular expression for $L$ (if possible)
- Given a regular expression $r$ we would like to "understand" $L(r)$ (say by giving an English description)

# THE END

...

# (for now)