

Pumping Lemma for Context-free Languages

Take an **infinite** context-free language



Generates an infinite number
of different strings

Example: $S \rightarrow ABE \mid bBd$

$A \rightarrow Aa \mid a$

$B \rightarrow bSD \mid cc$

$D \rightarrow Dd \mid d$

$E \rightarrow eE \mid e$

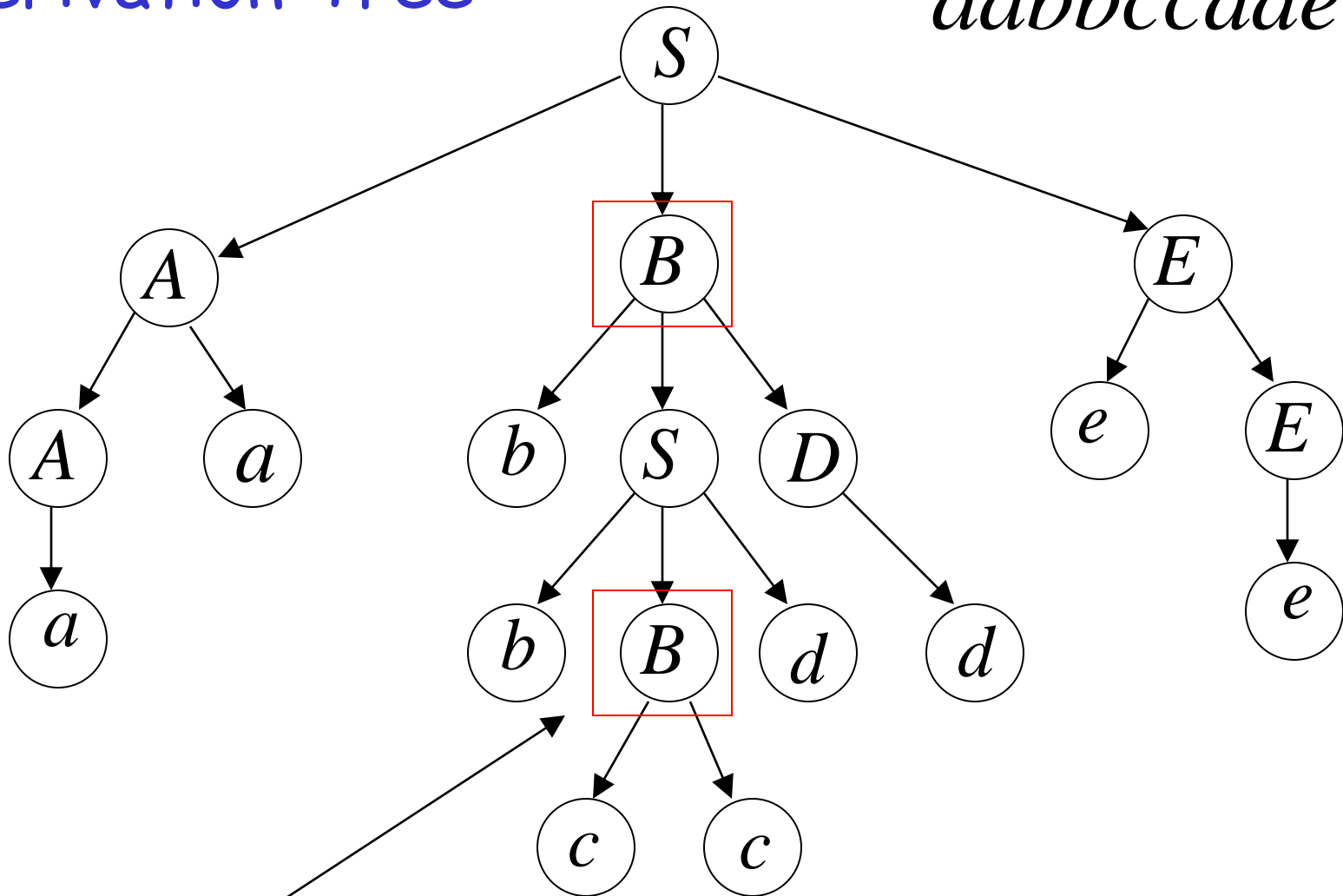
In a derivation of a "long" enough string, variables are repeated

A possible derivation:

$$\begin{aligned} S &\Rightarrow ABE \Rightarrow AaBE \Rightarrow aaBE \\ &\Rightarrow aabSDE \Rightarrow aabbBdDE \Rightarrow \\ &\Rightarrow aaabbccdDE \Rightarrow aabbccddeE \\ &\Rightarrow aabbccddeE \Rightarrow aabbccddeee \end{aligned}$$

Derivation Tree

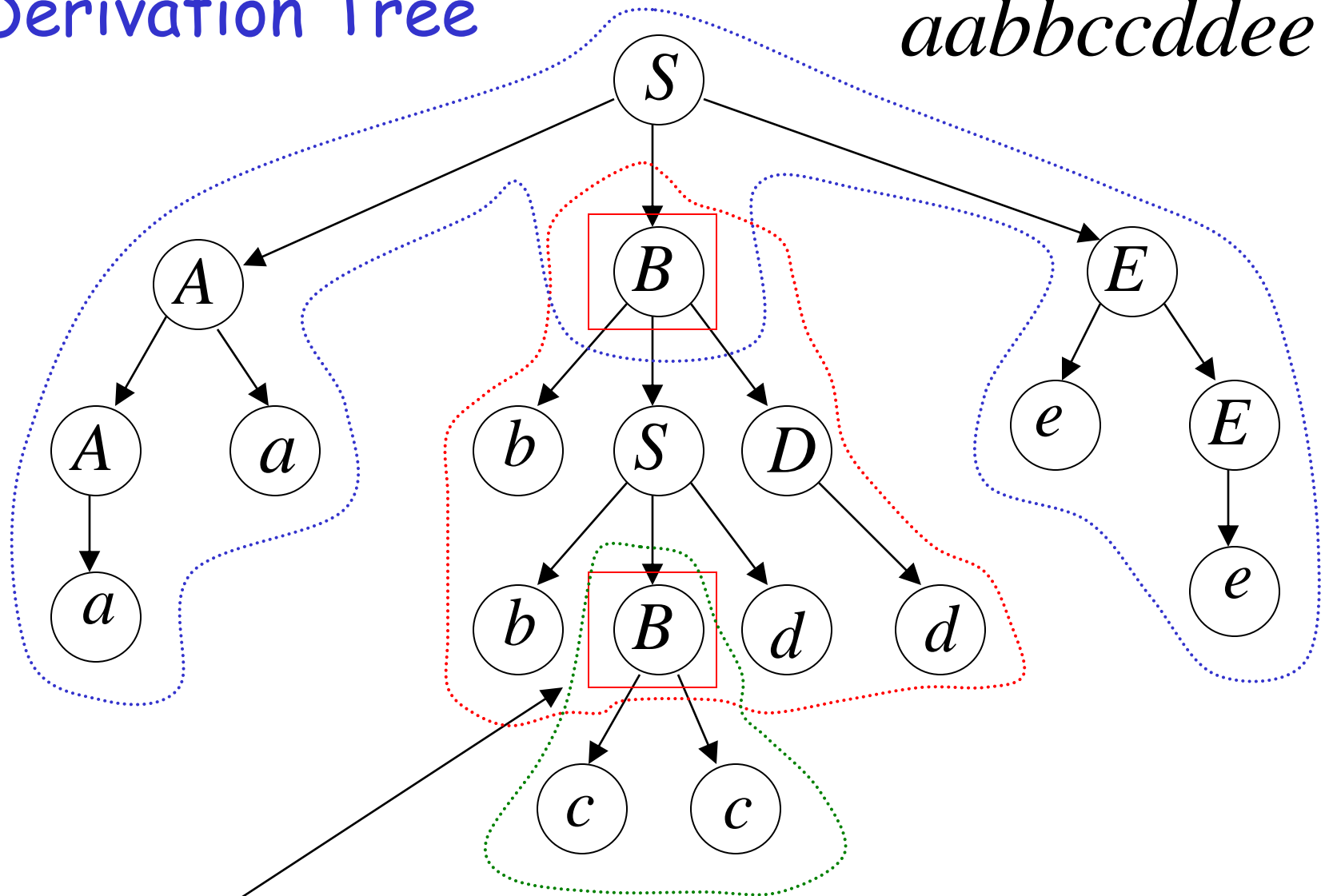
aabbccdde



Repeated
variable

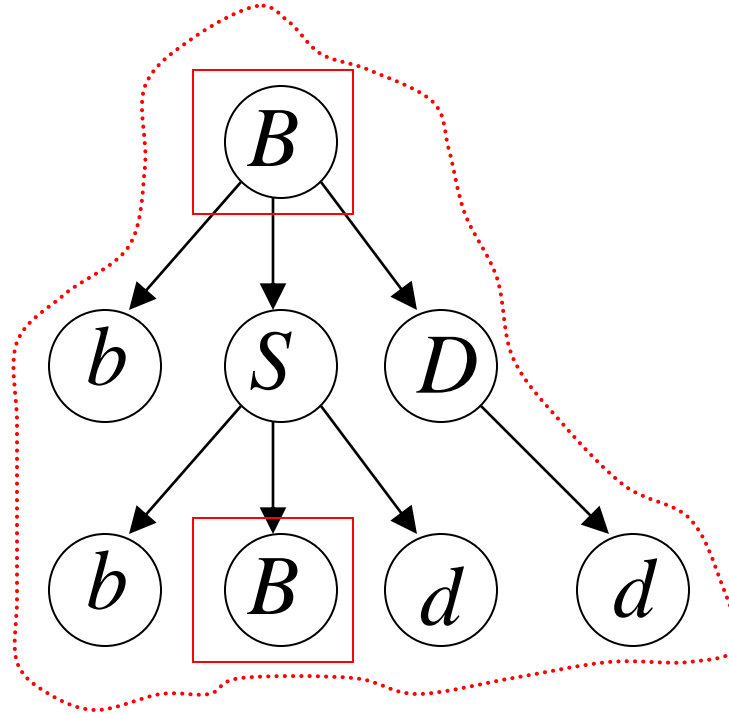
Derivation Tree

aabbccdde



Repeated
variable

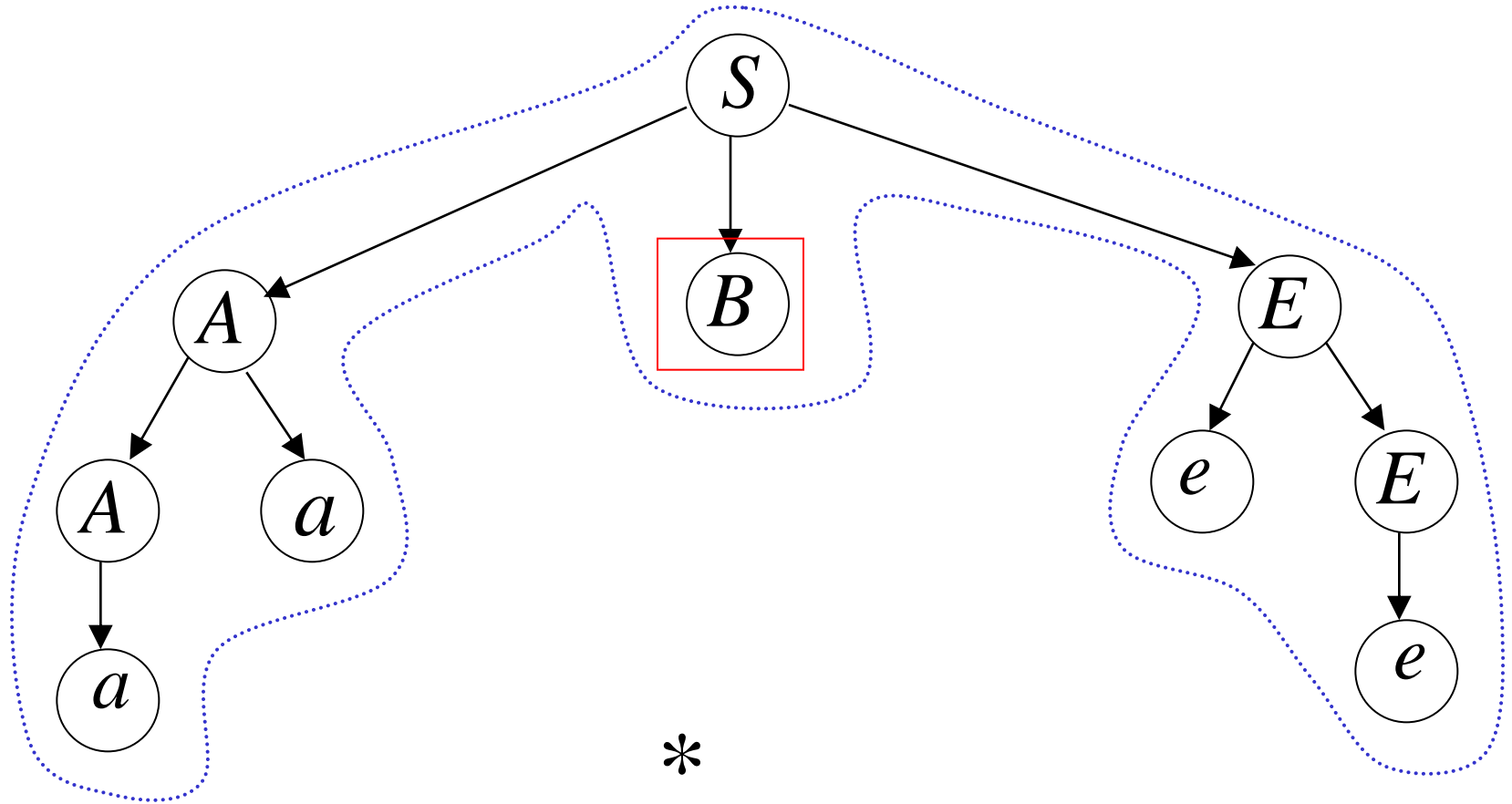
$$B \Rightarrow bSD \Rightarrow bbBdD \Rightarrow bbBdd$$



*

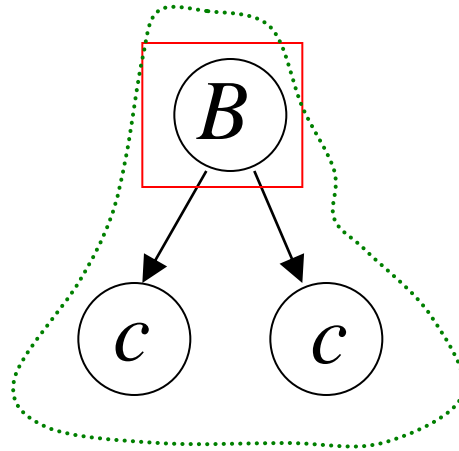
$$B \Rightarrow bbBdd$$

$S \Rightarrow ABE \Rightarrow AaBE \Rightarrow aaBE \Rightarrow aaBeE \Rightarrow aaBee$



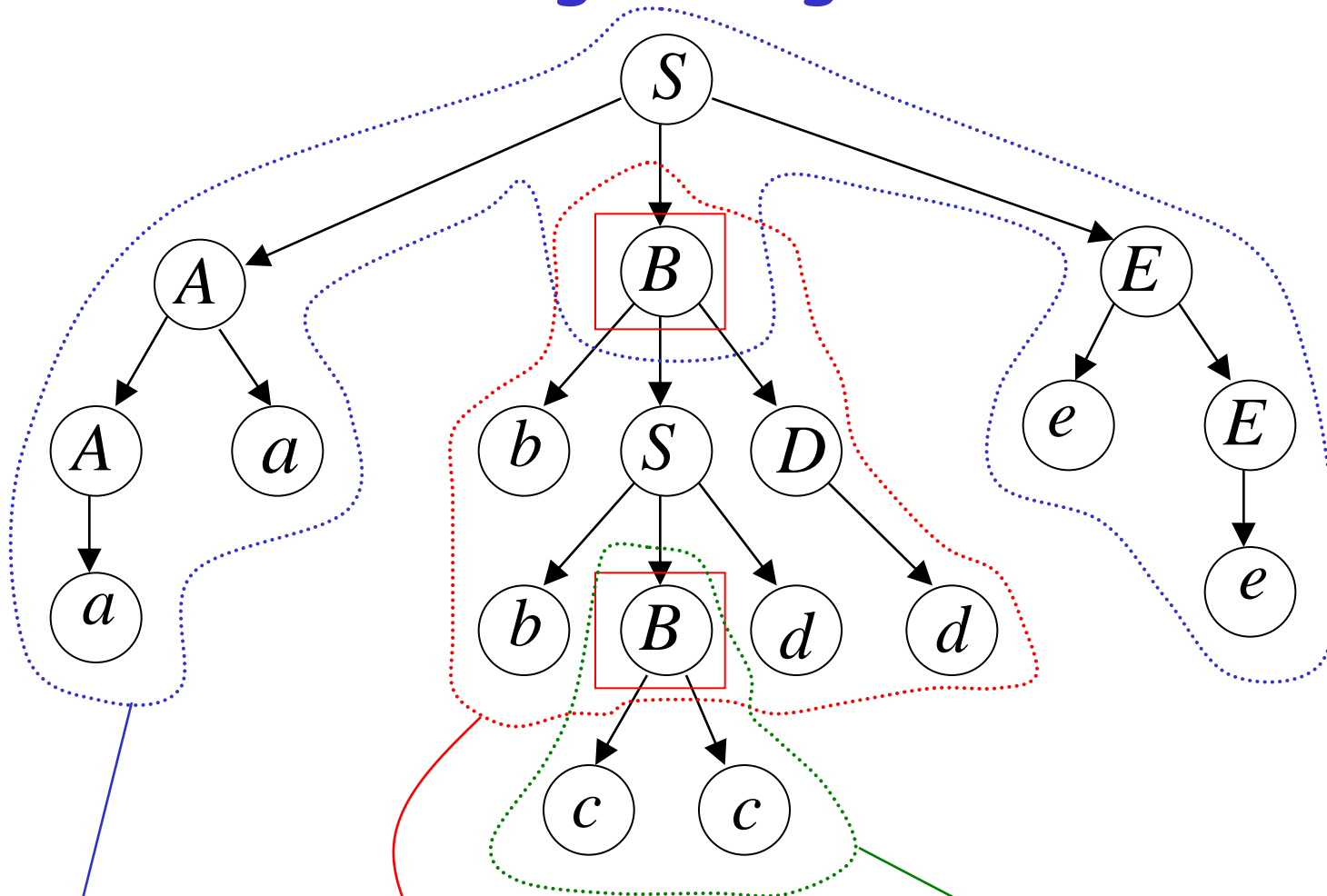
*

$S \Rightarrow aaBee$



$$B \Rightarrow cc$$

Putting all together

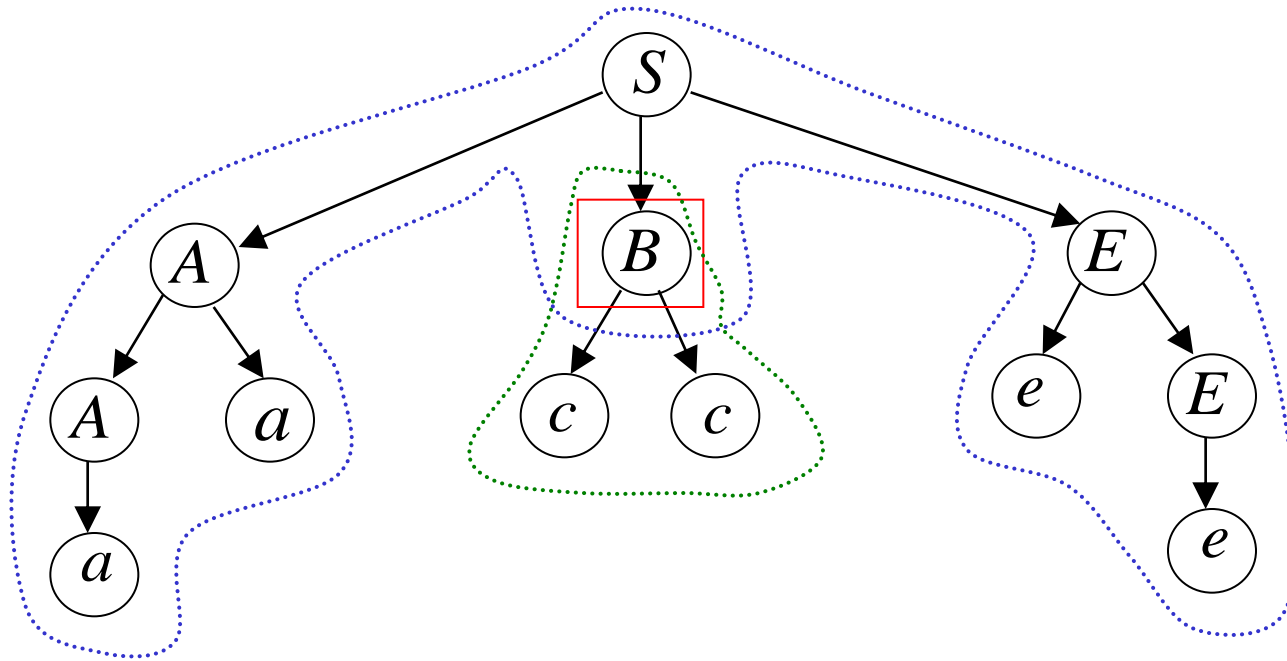


*
 $S \Rightarrow aaBee$

*
 $B \Rightarrow bbBdd$

$B \Rightarrow cc$

We can remove the middle part



$$* \\ S \Rightarrow aa(bb)^0 cc(dd)^0 ee$$

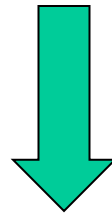
$$S \Rightarrow aaBee$$

$$B \Rightarrow bbBdd$$

$$B \Rightarrow cc$$

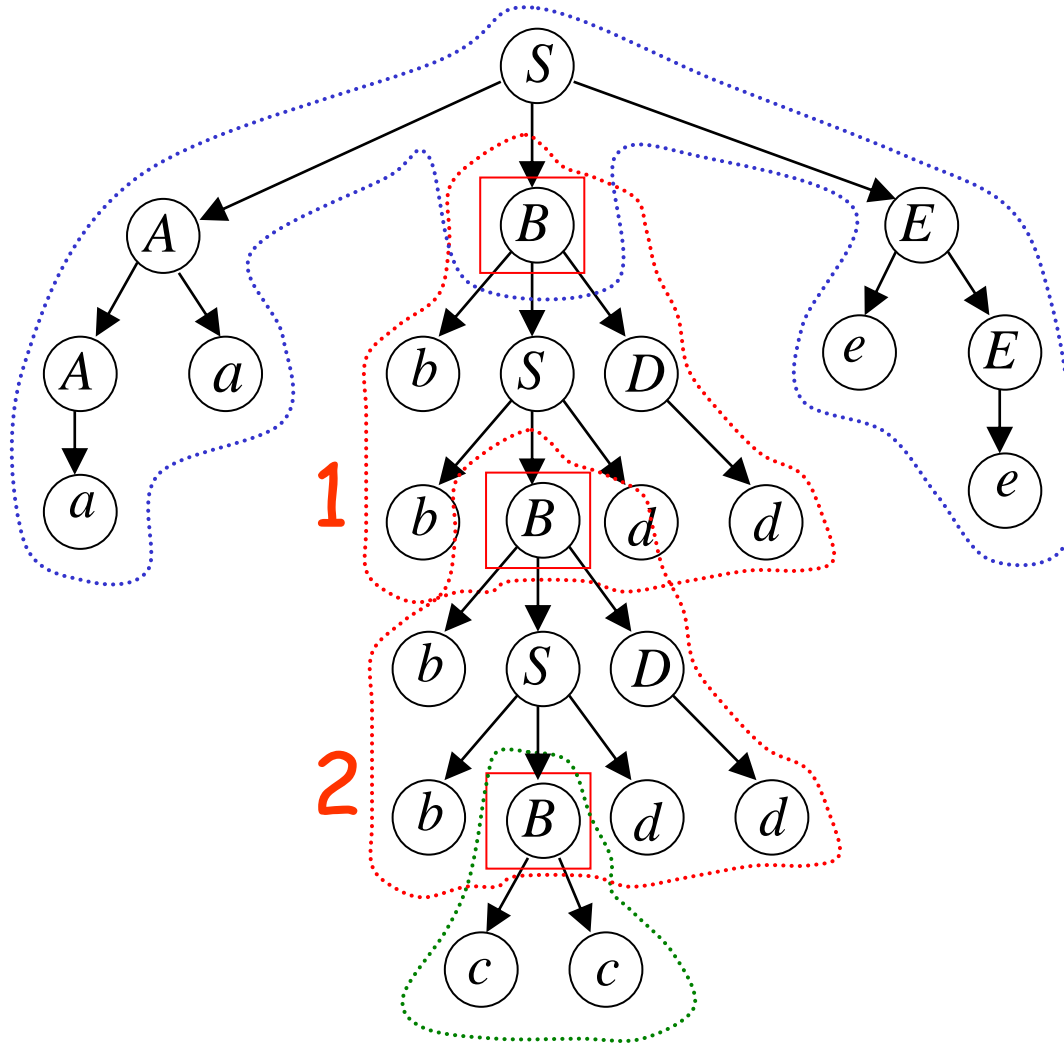


$$S \Rightarrow aaBee \Rightarrow aaccee = aa(bb)^0 cc(dd)^0 ee$$



$$aa(bb)^0 cc(dd)^0 ee \in L(G)$$

We can repeated middle part two times



*

$$S \Rightarrow aa(bb)^2 cc(dd)^2 ee$$

$$S \Rightarrow aaBee$$

$$B \Rightarrow bbBdd$$

$$B \Rightarrow cc$$



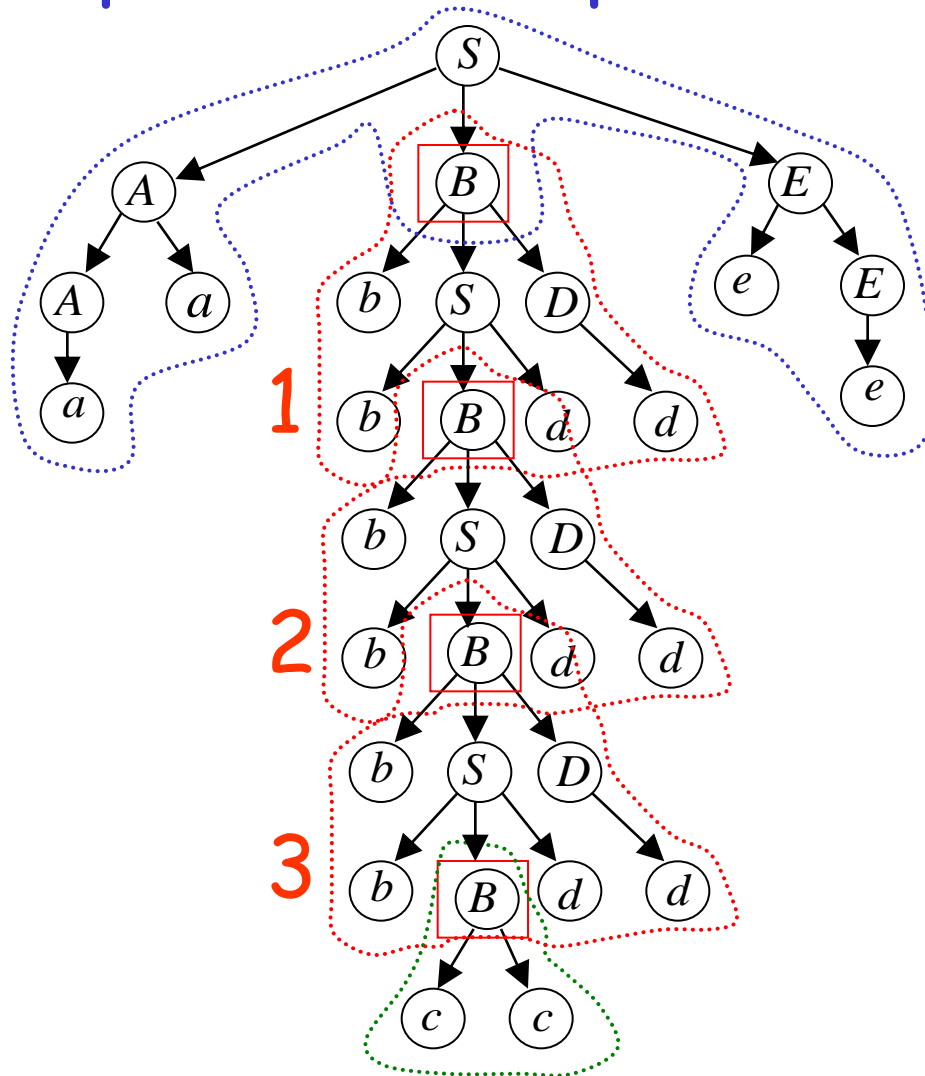
$$S \Rightarrow aaBee \Rightarrow aabbBddee$$

$$\Rightarrow aa(bb)^2 B(dd)^2 ee \Rightarrow aa(bb)^2 cc(dd)^2 ee$$



$$aa(bb)^2 cc(dd)^2 ee \in L(G)$$

We can repeat middle part three times



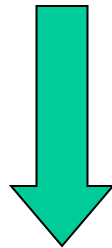
*

$$S \Rightarrow aa(bb)^3cc(dd)^3ee$$

$$* \\ S \Rightarrow aaBee$$

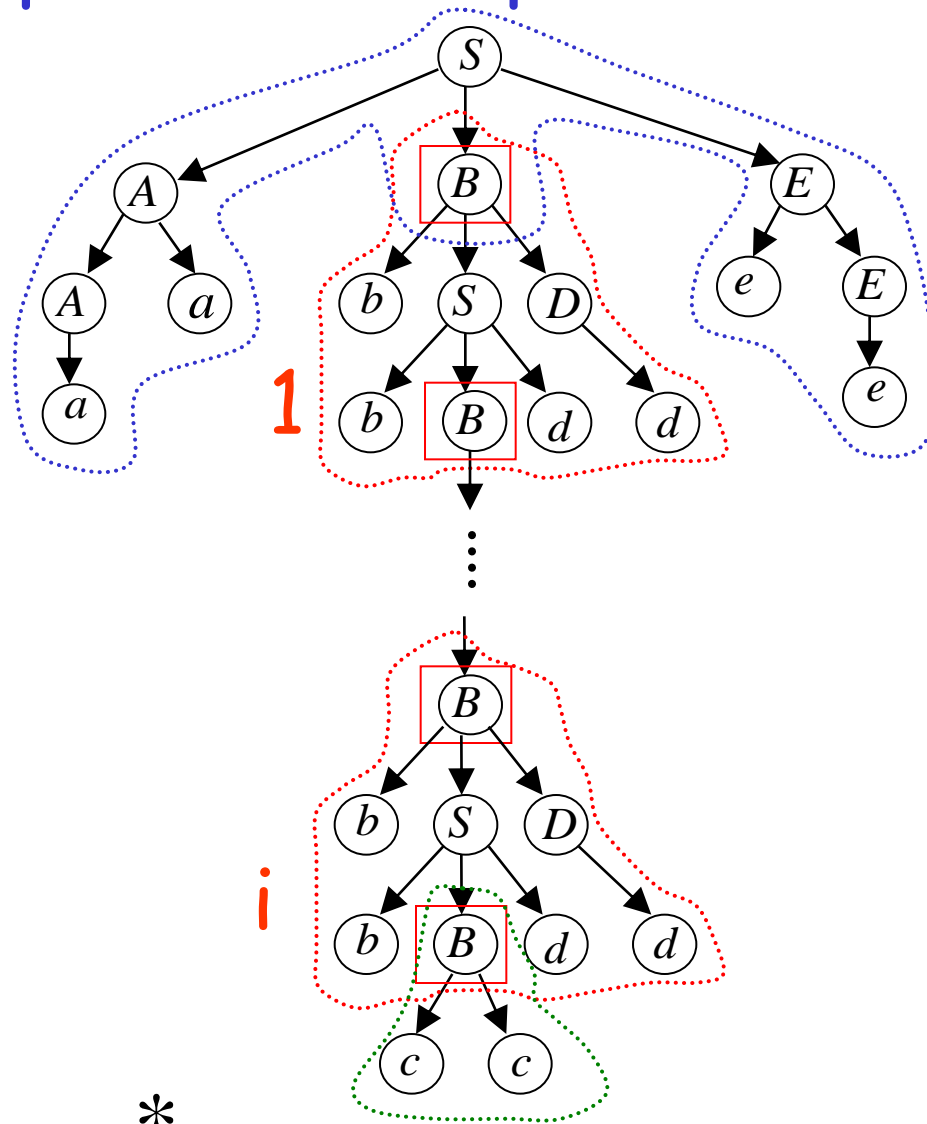
$$* \\ B \Rightarrow bbBdd$$

$$B \Rightarrow cc$$



$$* \\ S \Rightarrow aa(bb)^3 cc(dd)^3 ee \in L(G)$$

Repeat middle part i times



$$S \Rightarrow aa(bb)^i cc(dd)^i ee$$

$$* \\ S \Rightarrow aaBee$$

$$* \\ B \Rightarrow bbBdd$$

$$B \Rightarrow cc$$



$$* \\ S \Rightarrow aa(bb)^i cc(dd)^i ee \in L(G)$$

For any $i \geq 0$

From Grammar

$$S \rightarrow ABE \mid bBd$$

$$A \rightarrow Aa \mid a$$

$$B \rightarrow bSD \mid cc$$

$$D \rightarrow Dd \mid d$$

$$E \rightarrow eE \mid e$$

and given string

$$aabbccdde \in L(G)$$

We inferred that a family of strings is in $L(G)$

*

$$S \Rightarrow aa(bb)^i cc(dd)^i ee \in L(G) \text{ for any } i \geq 0$$

Arbitrary Grammars

Consider now an arbitrary **infinite context-free** language L

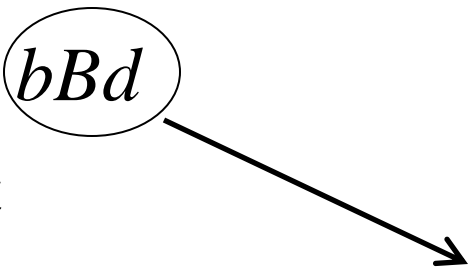
Let G be the grammar of $L - \{\lambda\}$

Take G so that it has no unit-productions
and no λ -productions
(remove them)

Let r be the number of variables

Let t be the maximum right-hand size
of any production

Example: $S \rightarrow ABE \mid bBd$ $r = 5$
 $A \rightarrow Aa \mid a$
 $B \rightarrow bSD \mid cc$ $t = 3$
 $D \rightarrow Dd \mid d$
 $E \rightarrow eE \mid e$



Claim:

Take string $w \in L(G)$ with $|w| > t^r$.
Then in the derivation tree of w
there is a path from the root to a leaf
where a variable of G is repeated

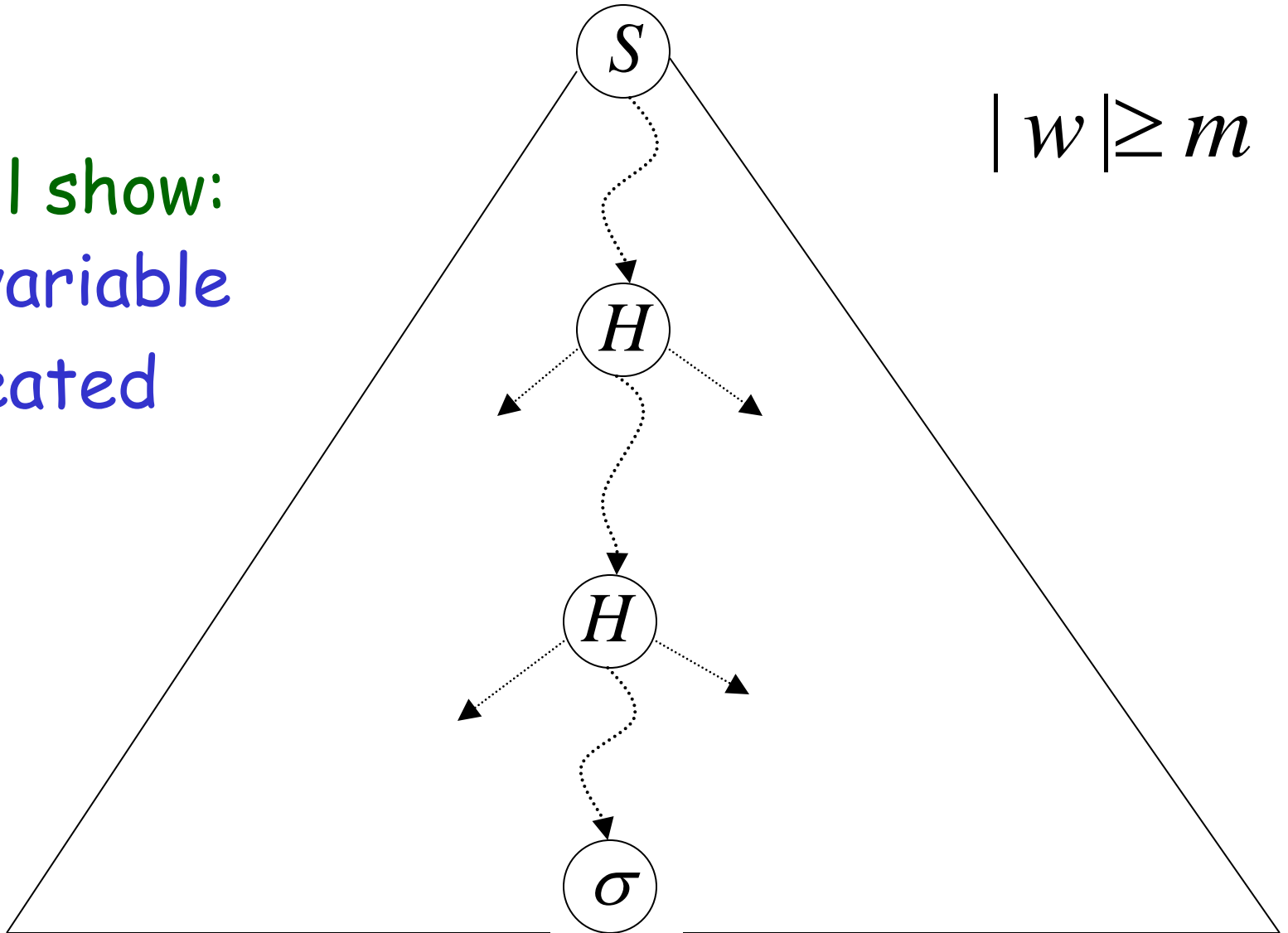
Proof:

Proof by contradiction

Derivation tree of w

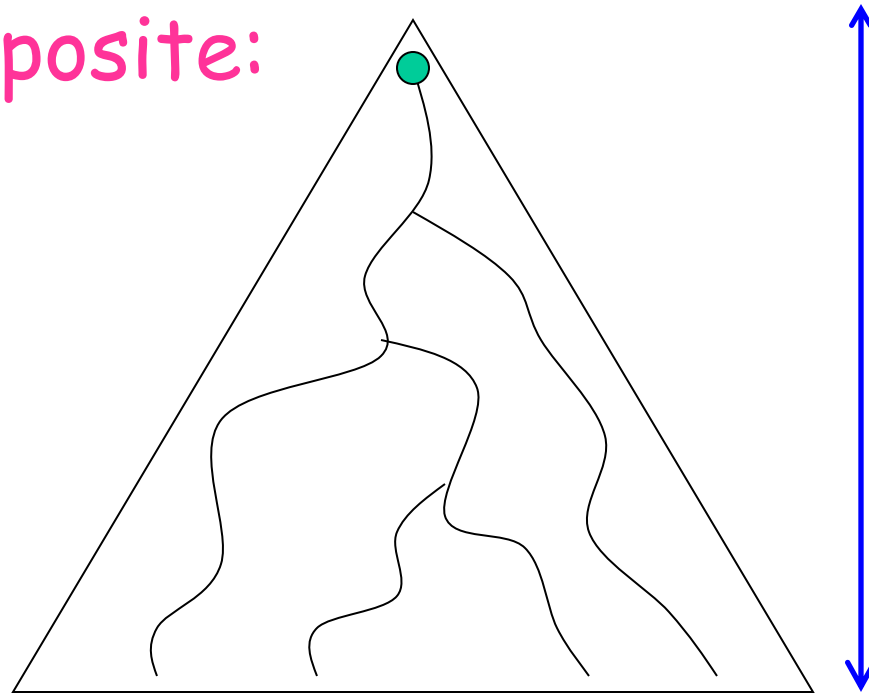
We will show:
some variable
is repeated

$$|w| \geq m$$



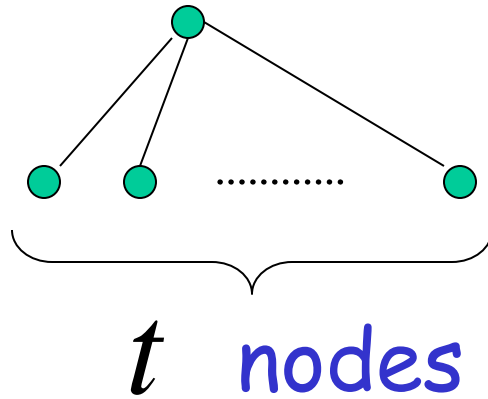
First we show that the tree of w
has at least $r + 2$ levels of nodes

Suppose the opposite:



At most
 $r + 1$
Levels

Maximum number of nodes per level

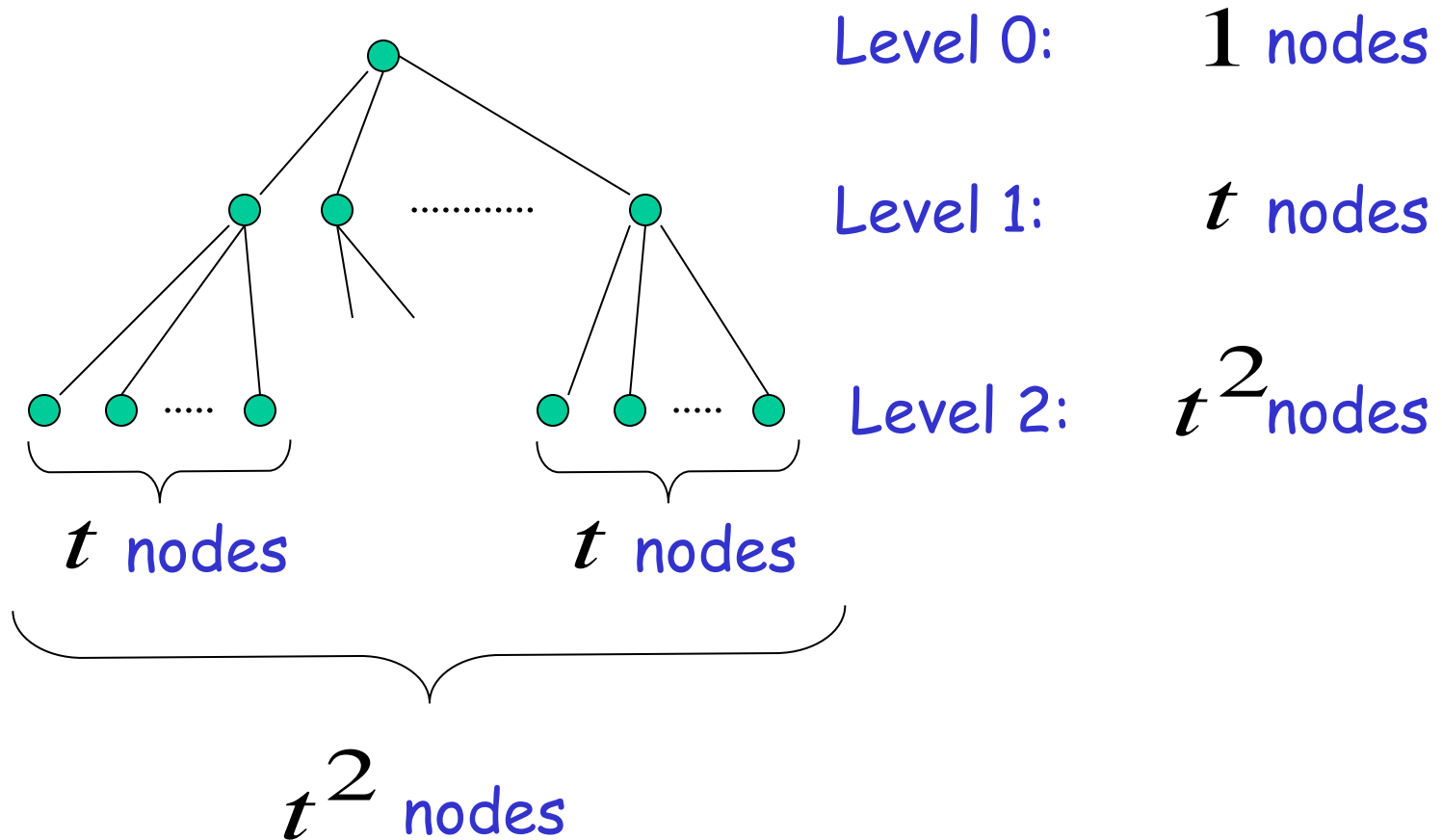


Level 0: **1** nodes

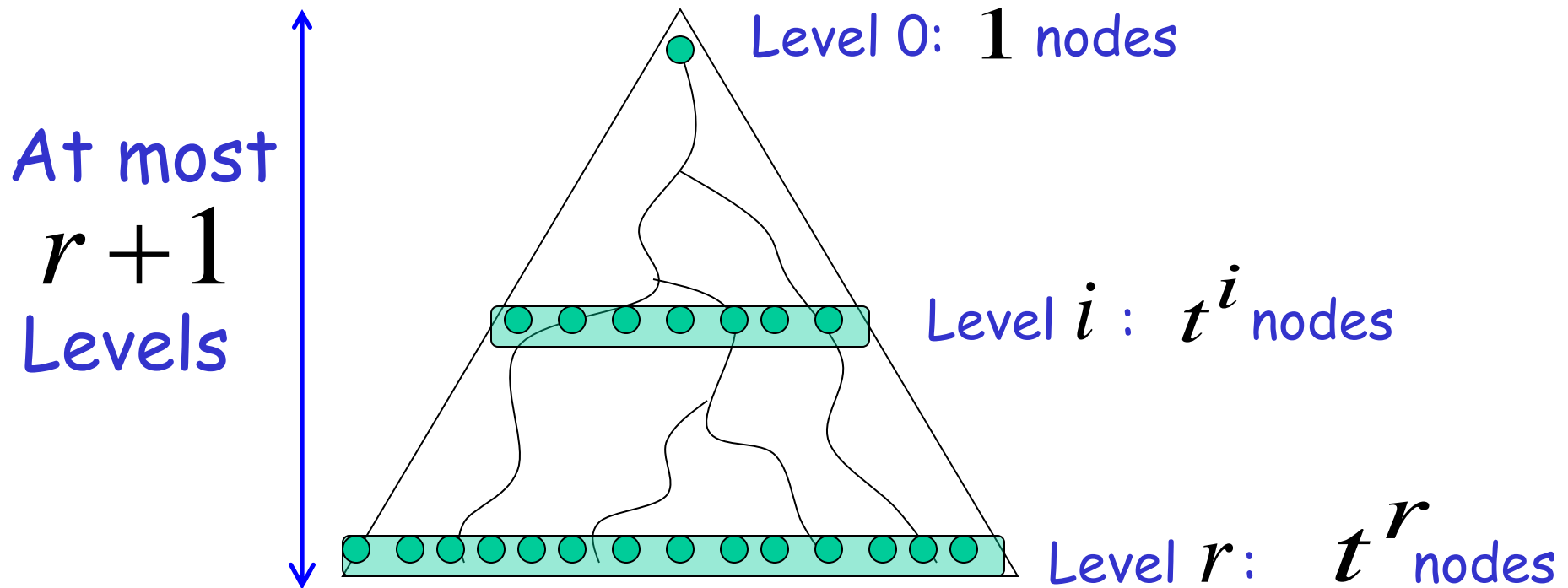
Level 1: t nodes

The maximum right-hand side of any production

Maximum number of nodes per level



Maximum number of nodes per level



Maximum possible string length

= max nodes at level $r = t^r$

Therefore,

maximum length of string $w : |w| \leq t^r$

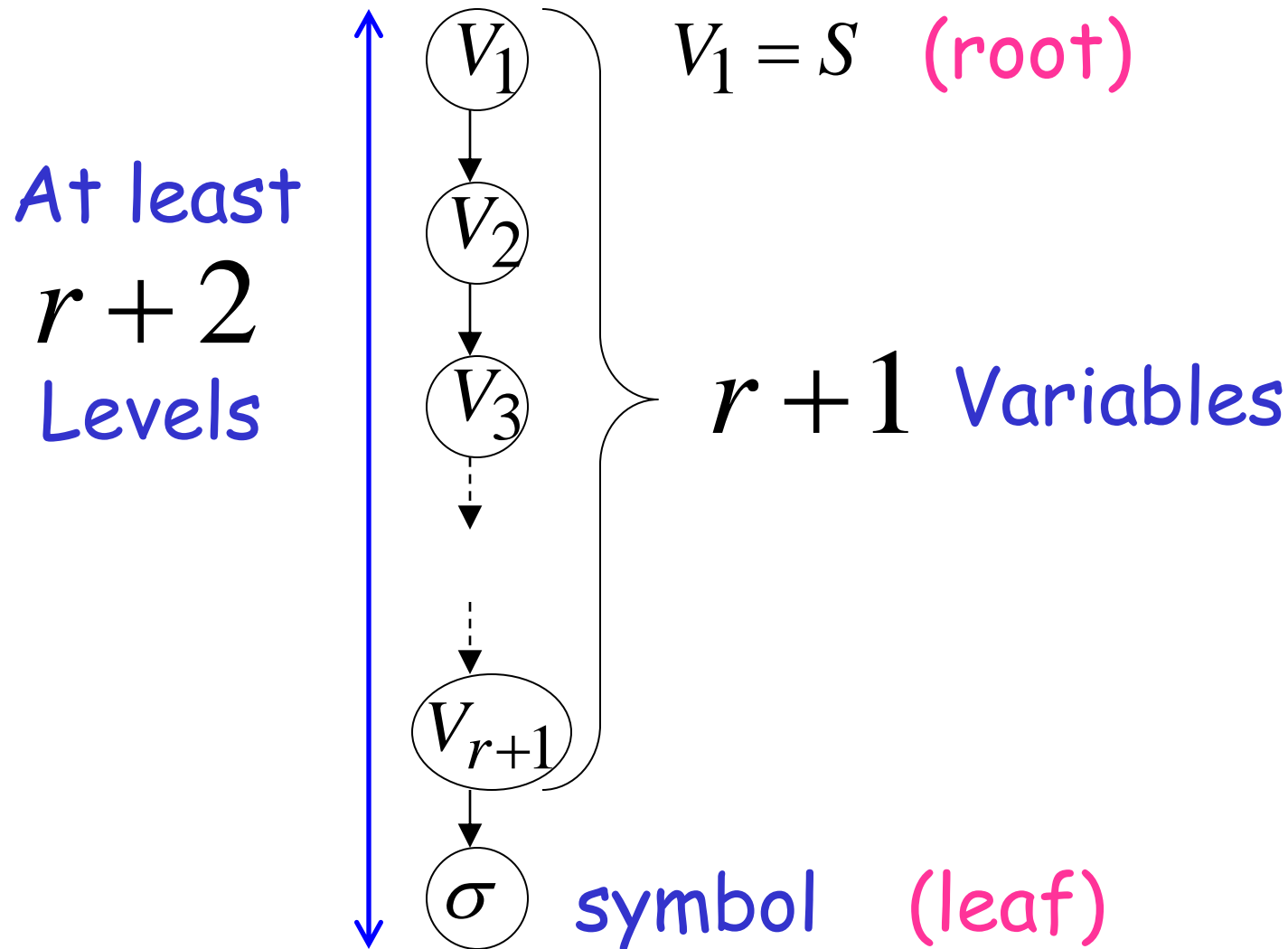
However we took, $|w| > t^r$

Contradiction!!!

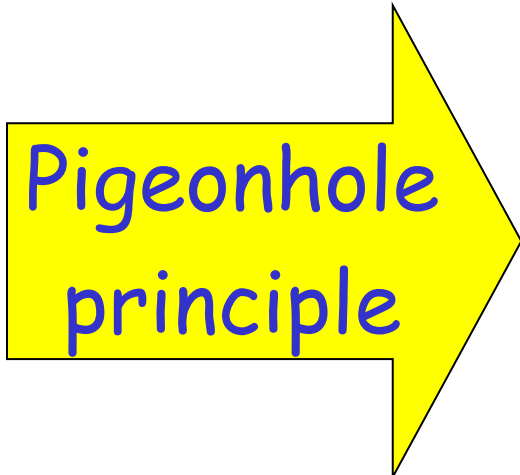
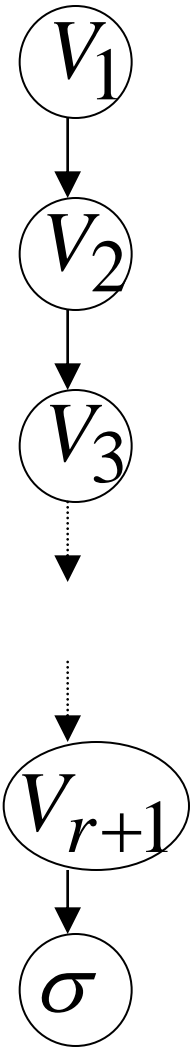
Therefore,

the tree must have at least $r + 2$ levels

Thus, there is a path from the root to a leaf with at least $r + 2$ nodes



Since there are at most r different variables, some variable is repeated

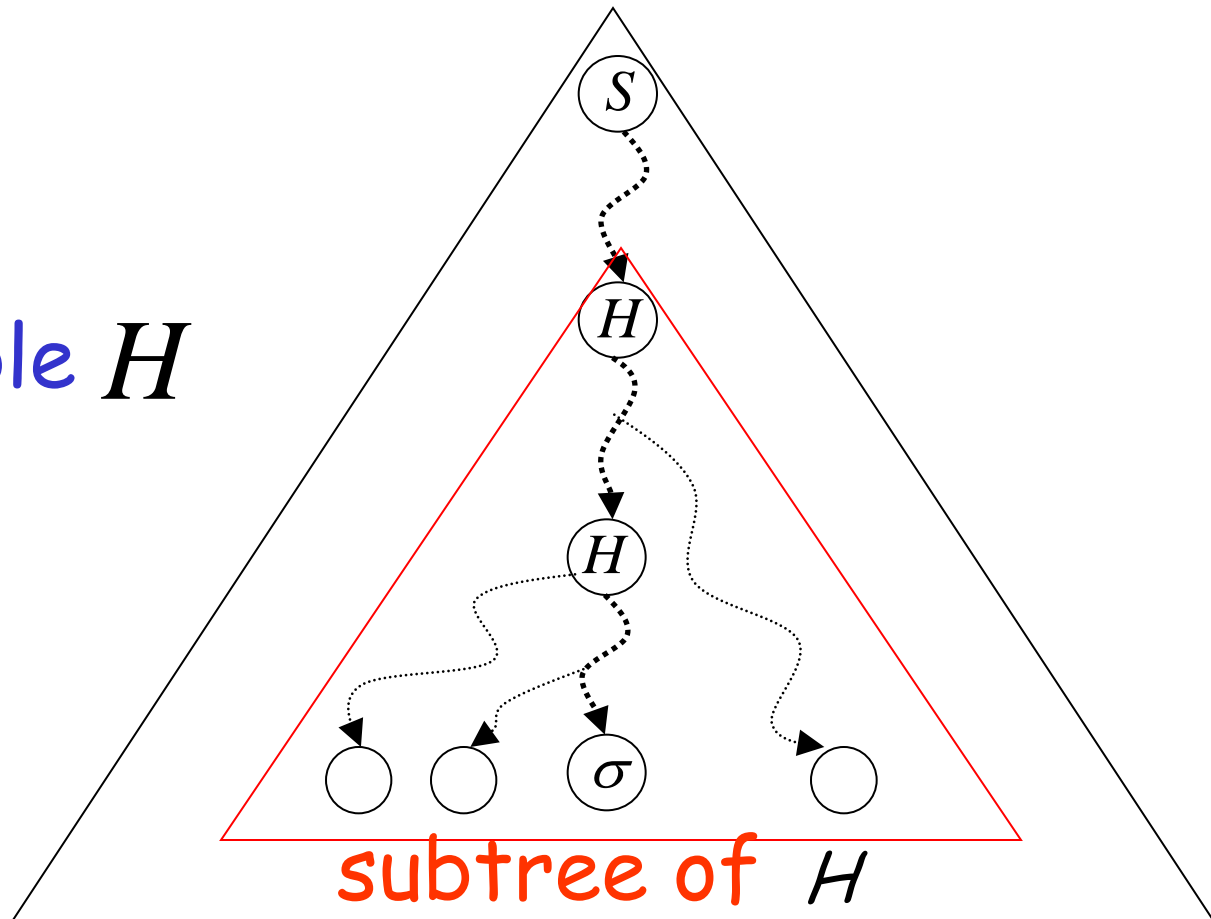


END OF CLAIM PROOF

Take now a string w with $|w| > t^r$

From claim:

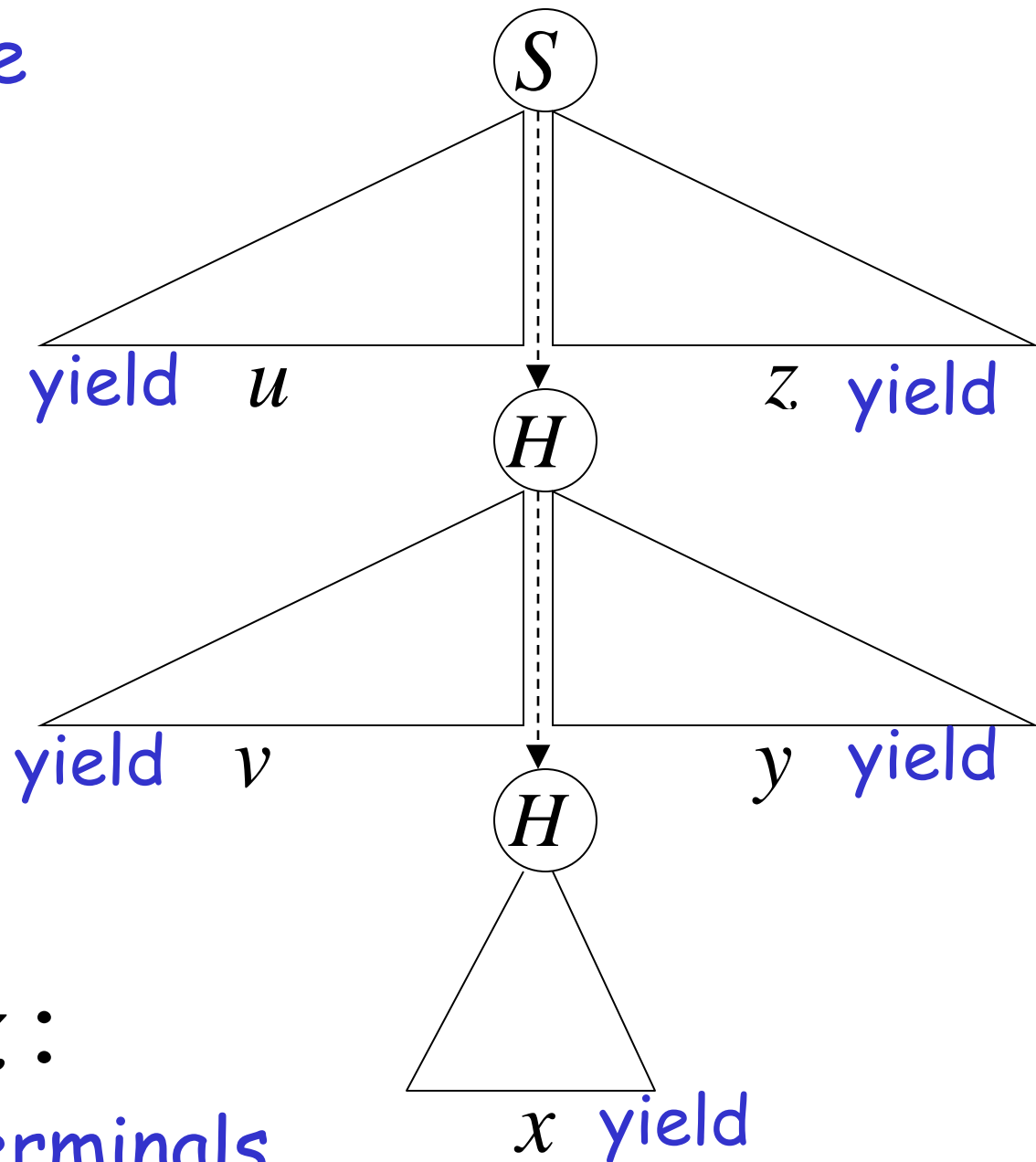
some variable H
is repeated



Take H to be the deepest, so that
only H is repeated in subtree

We can write

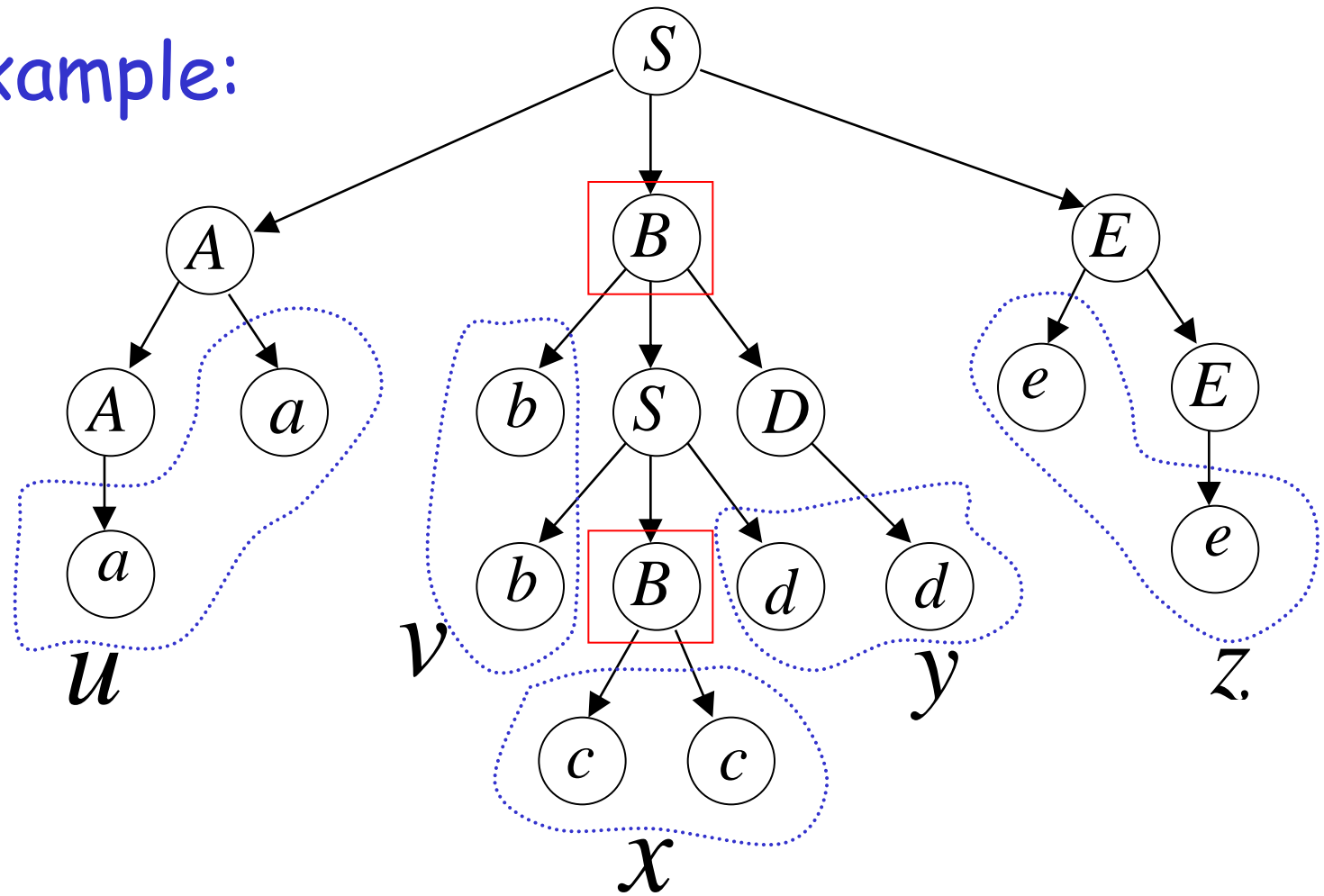
$$w = uvxyz$$



u, v, x, y, z :

Strings of terminals

Example:



$u = aa$

$v = bb$

$x = cc$

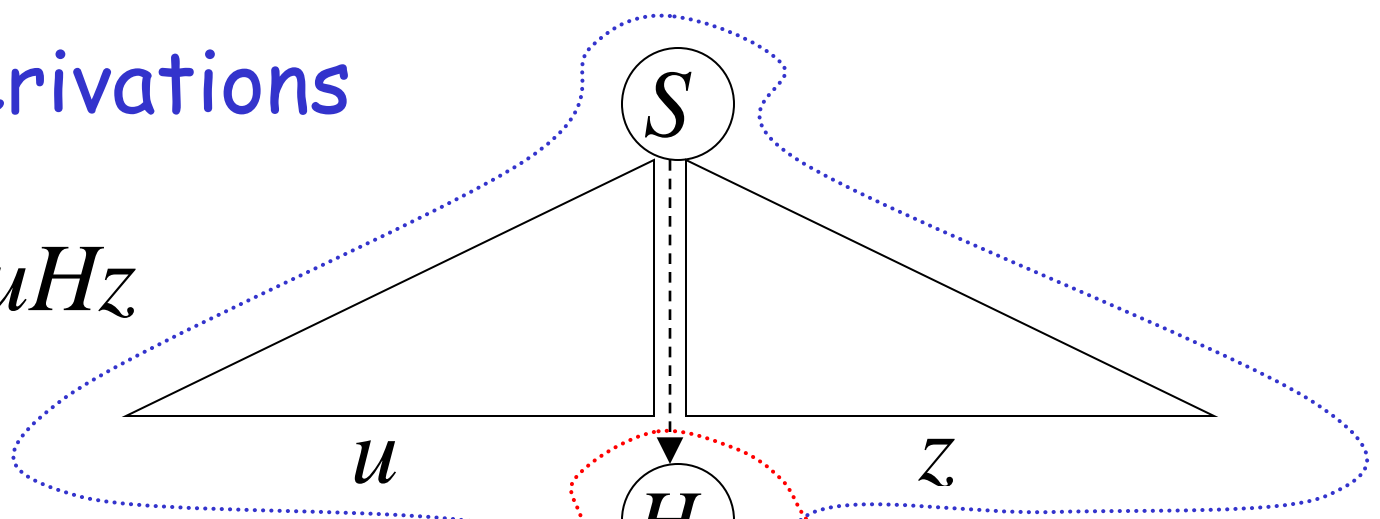
$y = dd$

$z = ee$

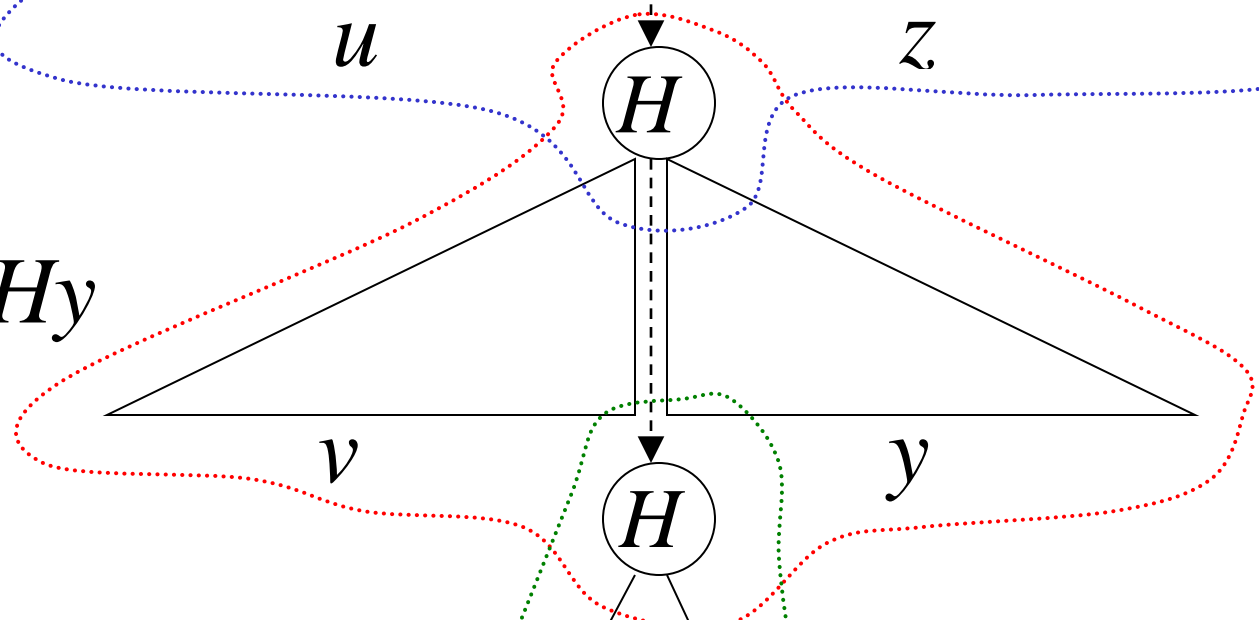
B corresponds to H

Possible derivations

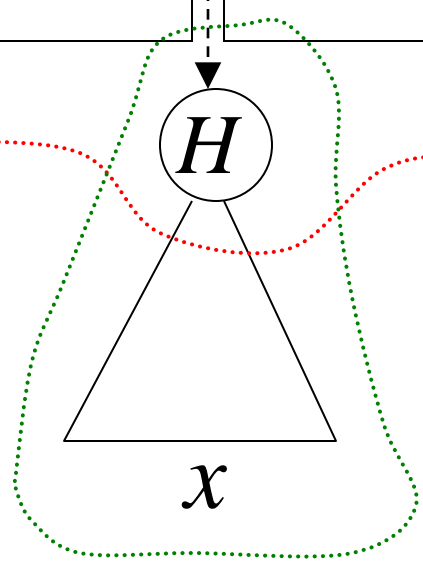
$$* \\ S \Rightarrow uHz$$



$$* \\ H \Rightarrow vHy$$



$$* \\ H \Rightarrow x$$



Example:

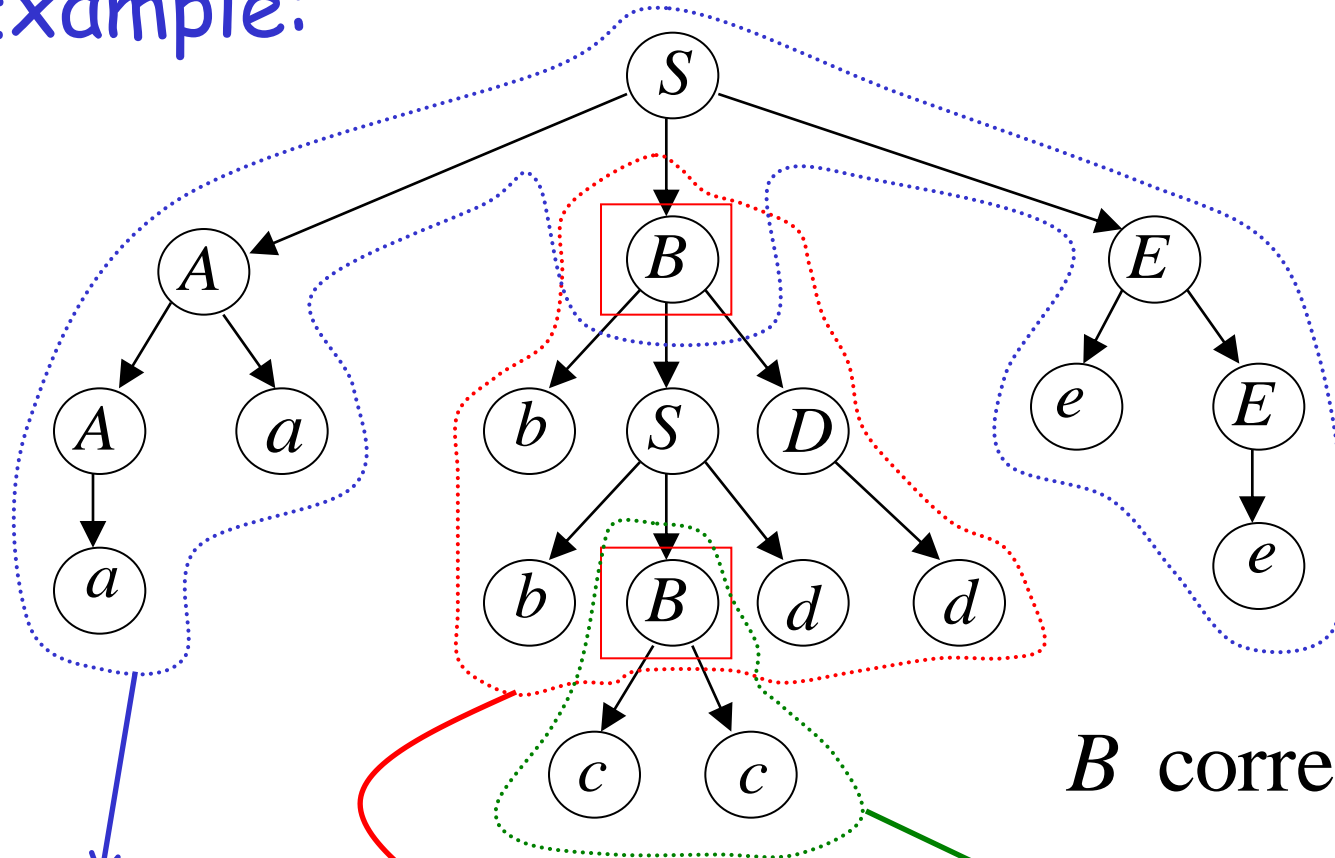
$u = aa$

$v = bb$

$x = cc$

$y = dd$

$z = ee$



B corresponds to H

$S \Rightarrow^* uHz$

$H \Rightarrow^* vHy$

$H \Rightarrow^* x$

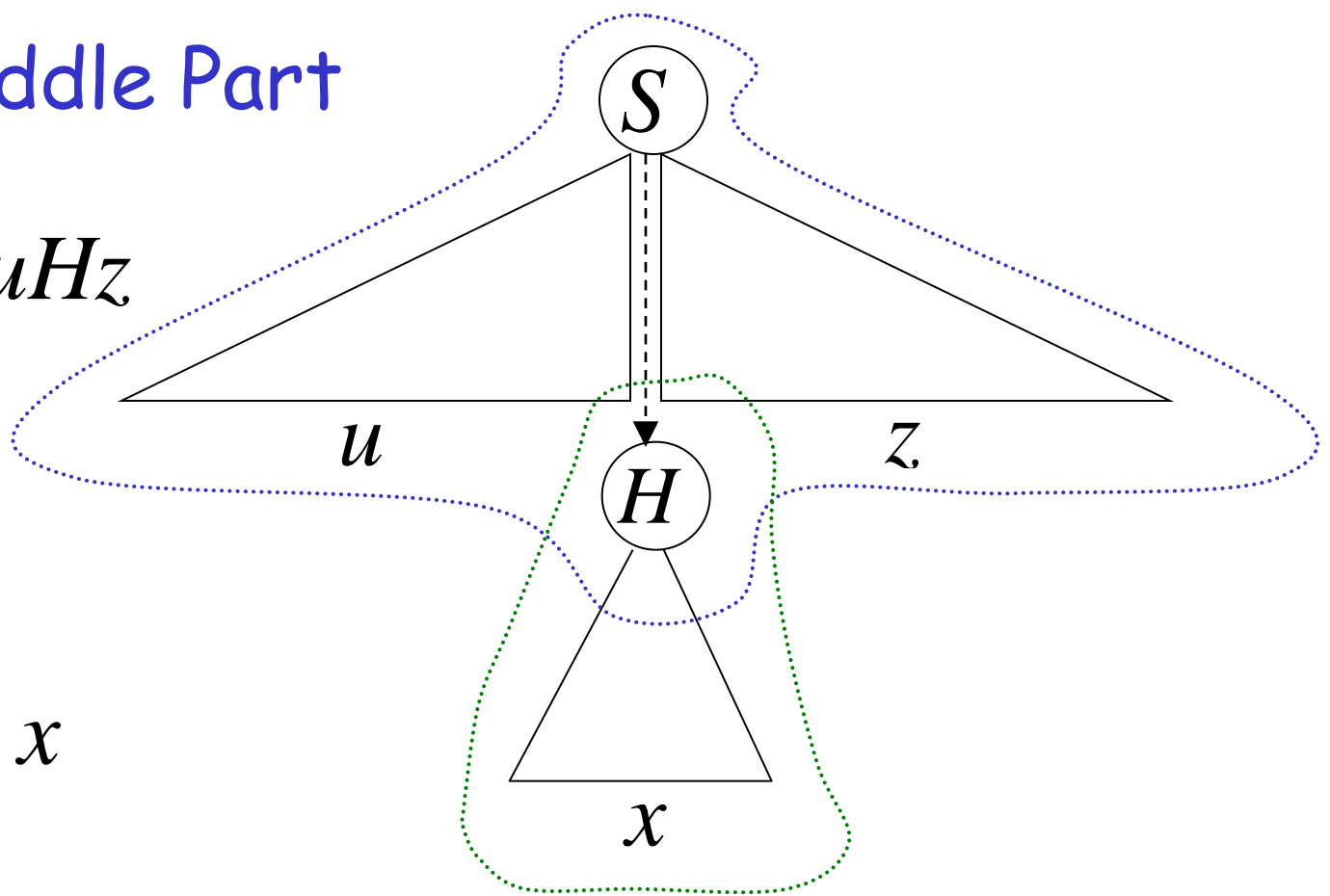
$S \Rightarrow^* aaBee$

$B \Rightarrow^* bbBdd$

$B \Rightarrow^* cc$

Remove Middle Part

$$\overset{*}{S} \Rightarrow uHz$$



$$\overset{*}{H} \Rightarrow x$$

Yield: $UXZ = UV^0XY^0Z$

$$\overset{*}{S} \Rightarrow \overset{*}{uHz} \Rightarrow \overset{*}{UXZ} = UV^0XY^0Z \in L(G)$$

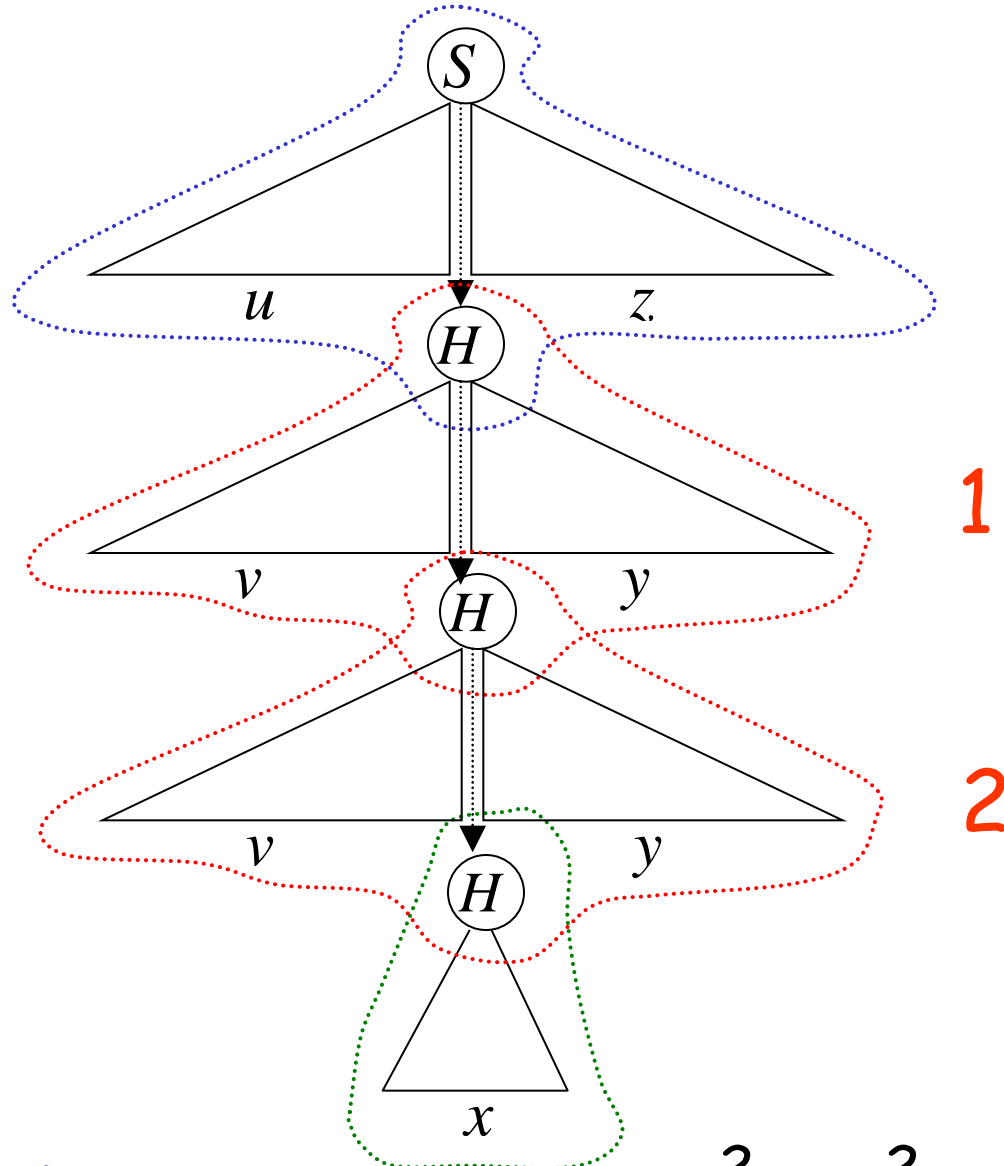
Repeat Middle part two times

$$* \\ S \Rightarrow uHz$$

$$* \\ H \Rightarrow vHy$$

$$* \\ H \Rightarrow vHy$$

$$* \\ H \Rightarrow x$$



Yield: $UVVXYYZ = UV^2XY^2Z$

$$S \stackrel{*}{\Rightarrow} uHz$$

$$H \stackrel{*}{\Rightarrow} vHy$$

$$H \stackrel{*}{\Rightarrow} x$$



$$S \stackrel{*}{\Rightarrow} uHz \stackrel{*}{\Rightarrow} uvHyz \stackrel{*}{\Rightarrow} uvvHyyz$$

$$\stackrel{*}{\Rightarrow} uvvxyyz = uv^2xy^2z \in L(G)$$

$$S \overset{*}{\Rightarrow} uHz$$

$$H \overset{*}{\Rightarrow} vHy$$

$$H \overset{*}{\Rightarrow} x$$



$$S \overset{*}{\Rightarrow} uHz \overset{*}{\Rightarrow} uvHyZ \overset{*}{\Rightarrow} uvvHyyz \overset{*}{\Rightarrow}$$

$$\overset{*}{\Rightarrow} \dots$$

$$\overset{*}{\Rightarrow} uv^i Hy^i z \overset{*}{\Rightarrow} uv^i xy^i z \in L(G)$$

Therefore,

$$|w| \geq t^r$$

If we know that: $w = uvxyz \in L(G)$

then we also know: $uv^i xy^i z \in L(G)$

For all $i \geq 0$

since

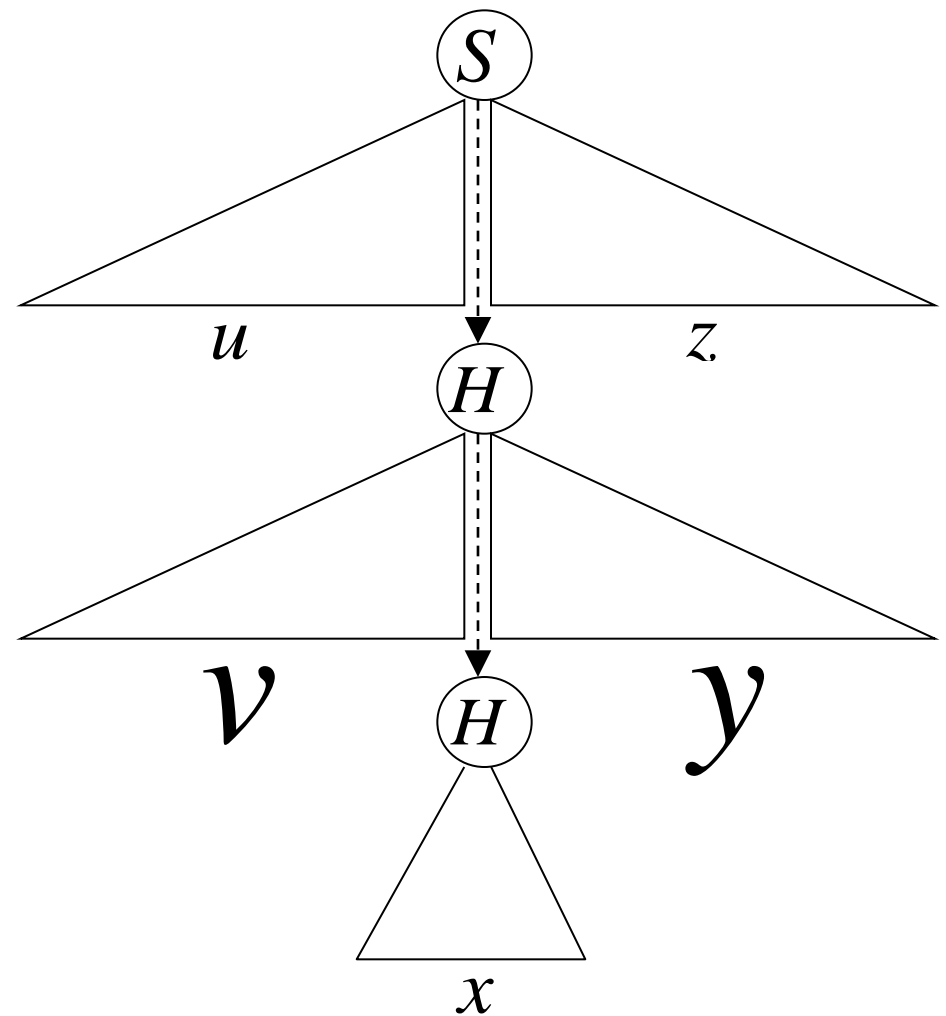
$$L(G) = L - \{\lambda\}$$

$$uv^i xy^i z \in L$$

Observation 1:

$$|vy| \geq 1$$

Since G has no
unit and
 λ -productions

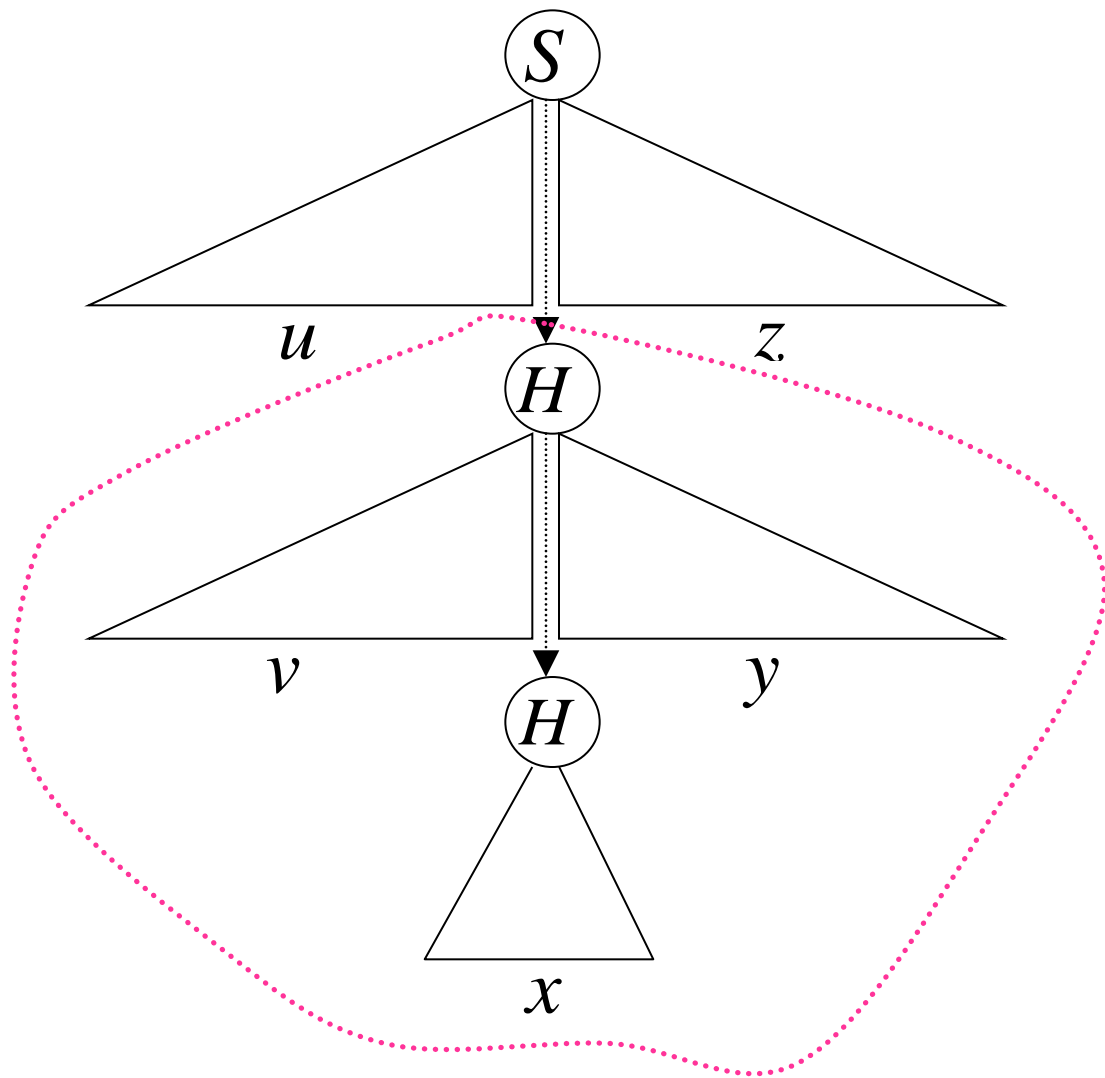


At least one of v or y is not λ

Observation 2:

$$|vxy| \leq t^{r+1}$$

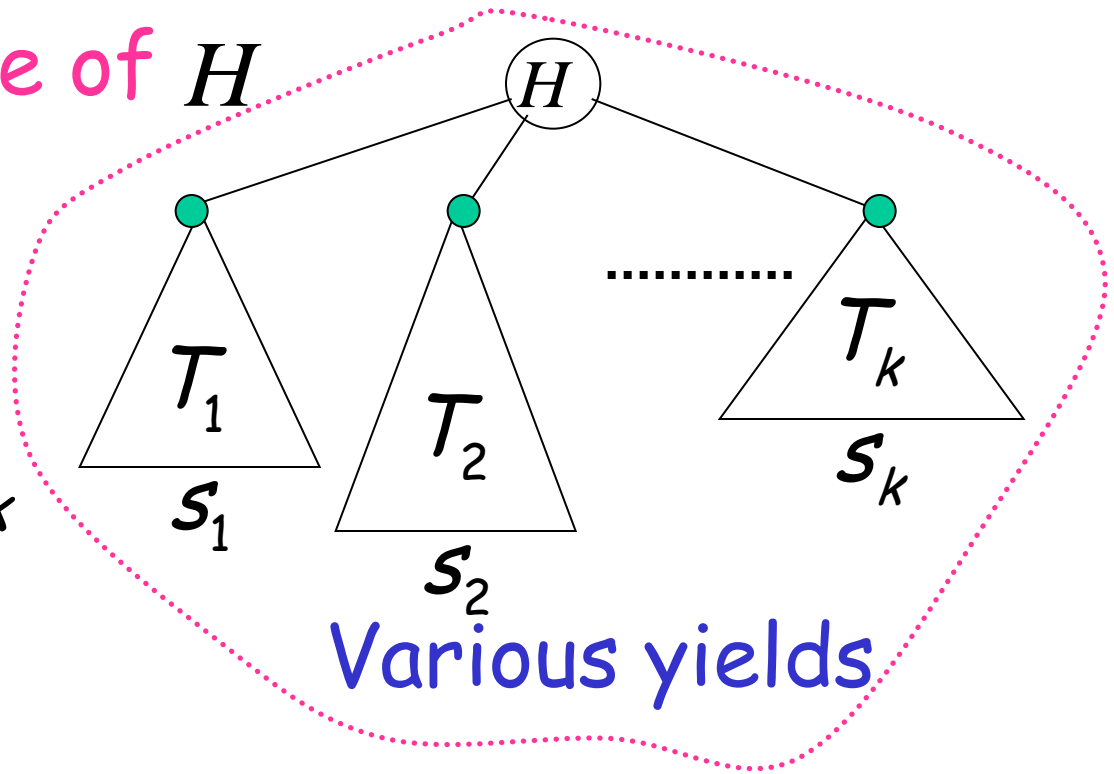
since in subtree
only variable H
is repeated



subtree of H

Explanation follows....

subtree of H



$$vxy = s_1 s_2 \cdots s_k$$

Various yields

$|s_j| \leq t^r$ since no variable is repeated in T_j

$$|vxy| = \sum_{j=1}^k |s_j| \leq k \cdot t^r \leq \underbrace{t}_{\substack{\text{Maximum right-hand} \\ \text{side of any production}}} \cdot t^r = t^{r+1}$$

Maximum right-hand side of any production

Thus, if we choose **critical length**

$$m = t^{r+1} > t^r$$

then, we obtain the pumping lemma for context-free languages

The Pumping Lemma:

For any infinite context-free language L

there exists an integer m such that

for any string $w \in L, |w| \geq m$

we can write $w = uvxyz$

with lengths $|vxy| \leq m$ and $|vy| \geq 1$

and it must be that:

$$uv^i xy^i z \in L, \quad \text{for all } i \geq 0$$

Applications of The Pumping Lemma

Non-context free languages

$$\{a^n b^n c^n : n \geq 0\}$$

Context-free languages

$$\{a^n b^n : n \geq 0\}$$

Theorem: The language

$$L = \{a^n b^n c^n : n \geq 0\}$$

is **not** context free

Proof: Use the Pumping Lemma
for context-free languages

$$L = \{a^n b^n c^n : n \geq 0\}$$

Assume for contradiction that L
is context-free

Since L is context-free and infinite
we can apply the pumping lemma

$$L = \{a^n b^n c^n : n \geq 0\}$$

Let m be the critical length
of the pumping lemma

Pick any string $w \in L$ with length $|w| \geq m$

We pick: $w = a^m b^m c^m$

$$L = \{a^n b^n c^n : n \geq 0\}$$

$$w = a^m b^m c^m$$

From pumping lemma:

we can write: $w = uvxyz$

with lengths $|vxy| \leq m$ and $|vy| \geq 1$

$$L = \{a^n b^n c^n : n \geq 0\}$$

$$w = a^m b^m c^m$$

$$w = uvxyz \quad |vxy| \leq m \quad |vy| \geq 1$$

Pumping Lemma says:

$$uv^i xy^i z \in L \quad \text{for all } i \geq 0$$

$$L = \{a^n b^n c^n : n \geq 0\}$$

$$w = a^m b^m c^m$$

$$w = uvxyz \quad |vxy| \leq m \quad |vy| \geq 1$$

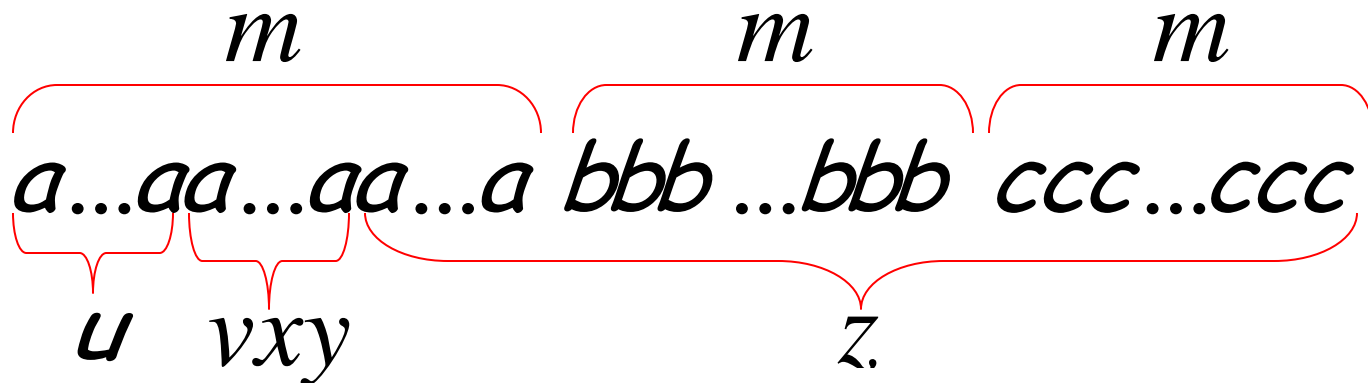
We examine all the possible locations
of string vxy in w

$$L = \{a^n b^n c^n : n \geq 0\}$$

$$w = a^m b^m c^m$$

$$w = uvxyz \quad |vxy| \leq m \quad |vy| \geq 1$$

Case 1: vxy is in a^m

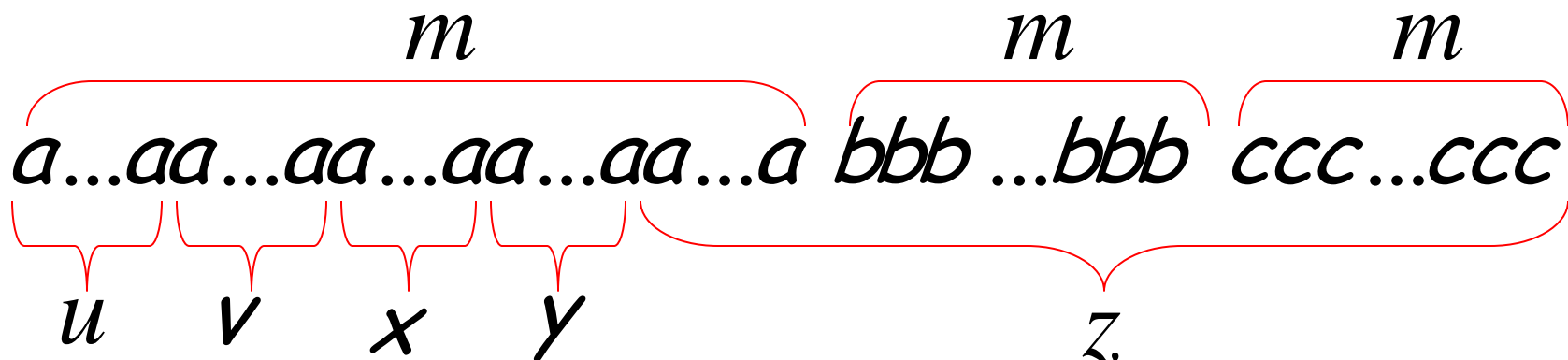


$$L = \{a^n b^n c^n : n \geq 0\}$$

$$w = a^m b^m c^m$$

$$w = uvxyz \quad |vxy| \leq m \quad |vy| \geq 1$$

$$v = a^{k_1} \quad y = a^{k_2} \quad k_1 + k_2 \geq 1$$

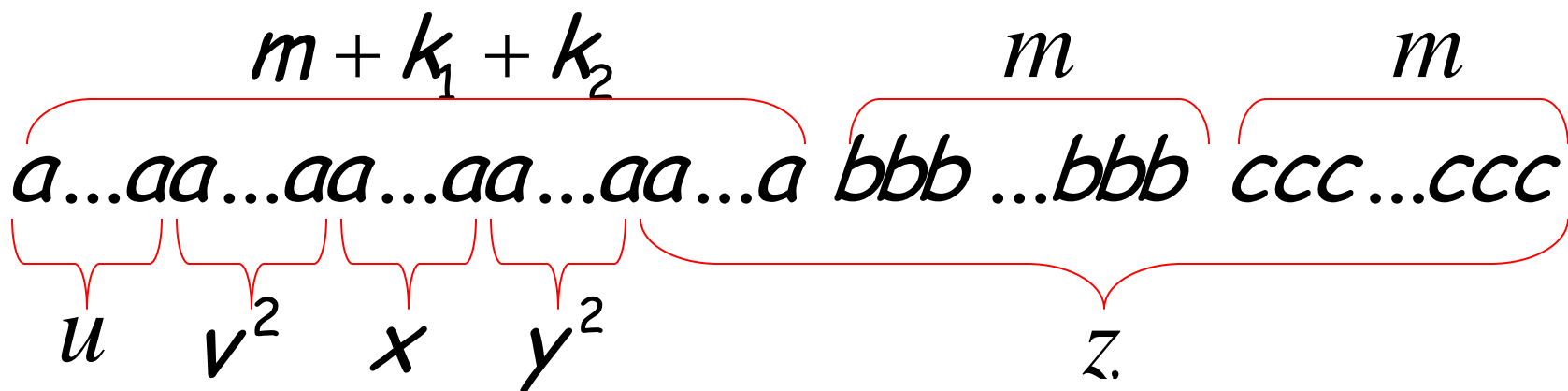


$$L = \{a^n b^n c^n : n \geq 0\}$$

$$w = a^m b^m c^m$$

$$w = uvxyz \quad |vxy| \leq m \quad |vy| \geq 1$$

$$v = a^{k_1} \quad y = a^{k_2} \quad k_1 + k_2 \geq 1$$



$$L = \{a^n b^n c^n : n \geq 0\}$$

$$w = a^m b^m c^m$$

$$w = uvxyz \quad |vxy| \leq m \quad |vy| \geq 1$$

From Pumping Lemma: $uv^2xy^2z \in L$

$$k_1 + k_2 \geq 1$$

However: $uv^2xy^2z = a^{m+k_1+k_2} b^m c^m \notin L$

Contradiction!!!

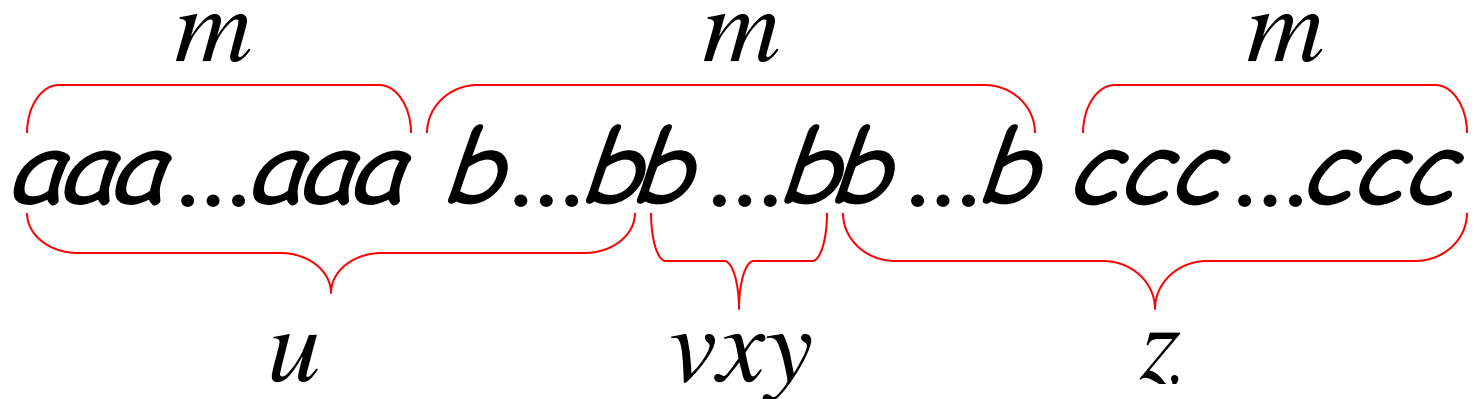
$$L = \{a^n b^n c^n : n \geq 0\}$$

$$w = a^m b^m c^m$$

$$w = uvxyz \quad |vxy| \leq m \quad |vy| \geq 1$$

Case 2: vxy is in b^m

Similar to case 1



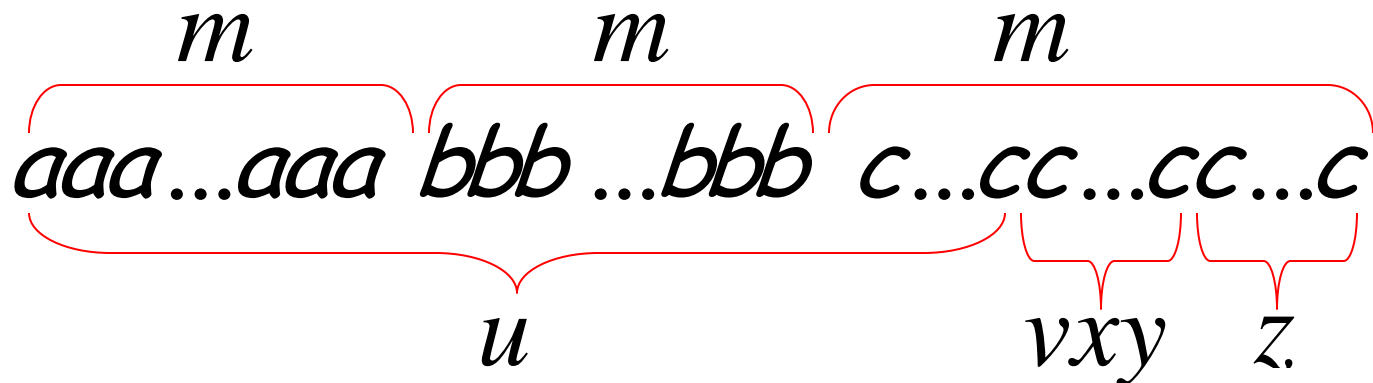
$$L = \{a^n b^n c^n : n \geq 0\}$$

$$w = a^m b^m c^m$$

$$w = uvxyz \quad |vxy| \leq m \quad |vy| \geq 1$$

Case 3: vxy is in c^m

Similar to case 1

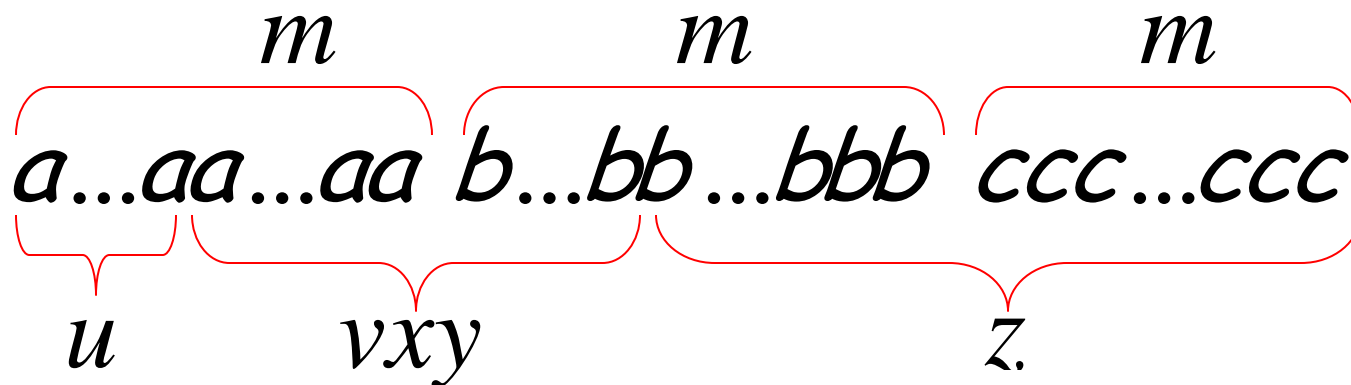


$$L = \{a^n b^n c^n : n \geq 0\}$$

$$w = a^m b^m c^m$$

$$w = uvxyz \quad |vxy| \leq m \quad |vy| \geq 1$$

Case 4: vxy overlaps a^m and b^m

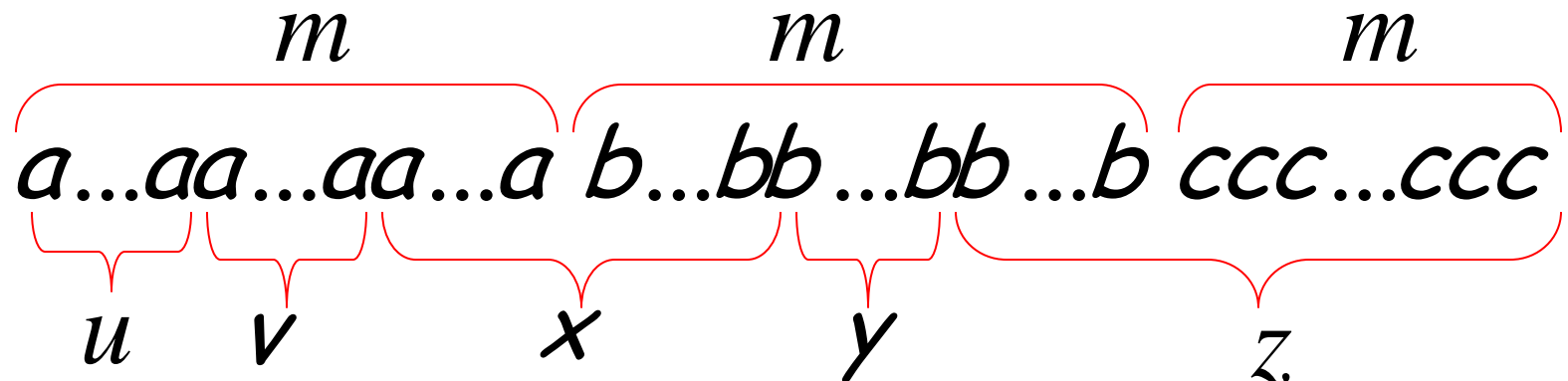


$$L = \{a^n b^n c^n : n \geq 0\}$$

$$w = a^m b^m c^m$$

$$w = uvxyz \quad |vxy| \leq m \quad |vy| \geq 1$$

Sub-case 1: v contains only a
 y contains only b

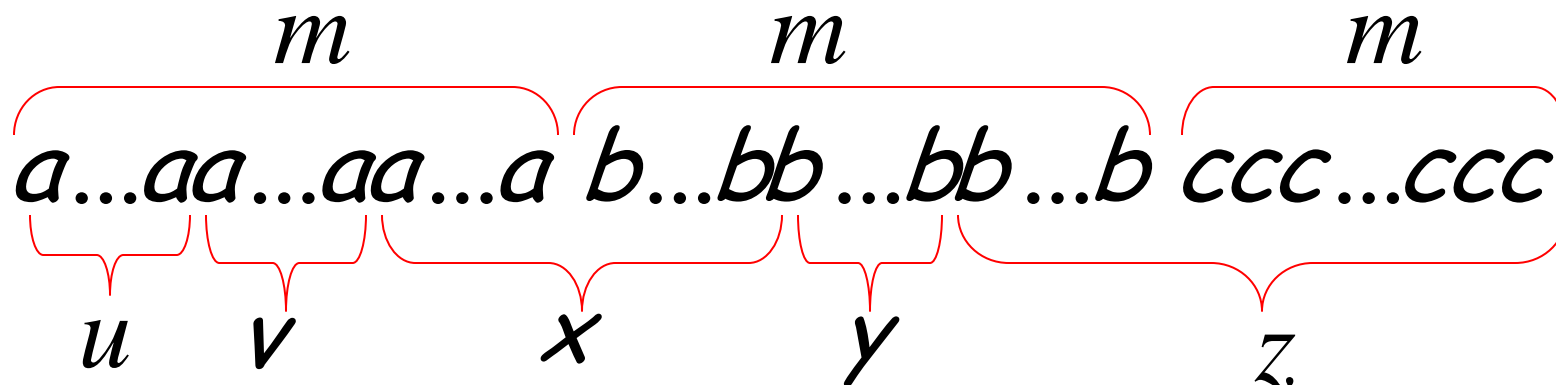


$$L = \{a^n b^n c^n : n \geq 0\}$$

$$w = a^m b^m c^m$$

$$w = uvxyz \quad |vxy| \leq m \quad |vy| \geq 1$$

$$v = a^{k_1} \quad y = a^{k_2} \quad k_1 + k_2 \geq 1$$

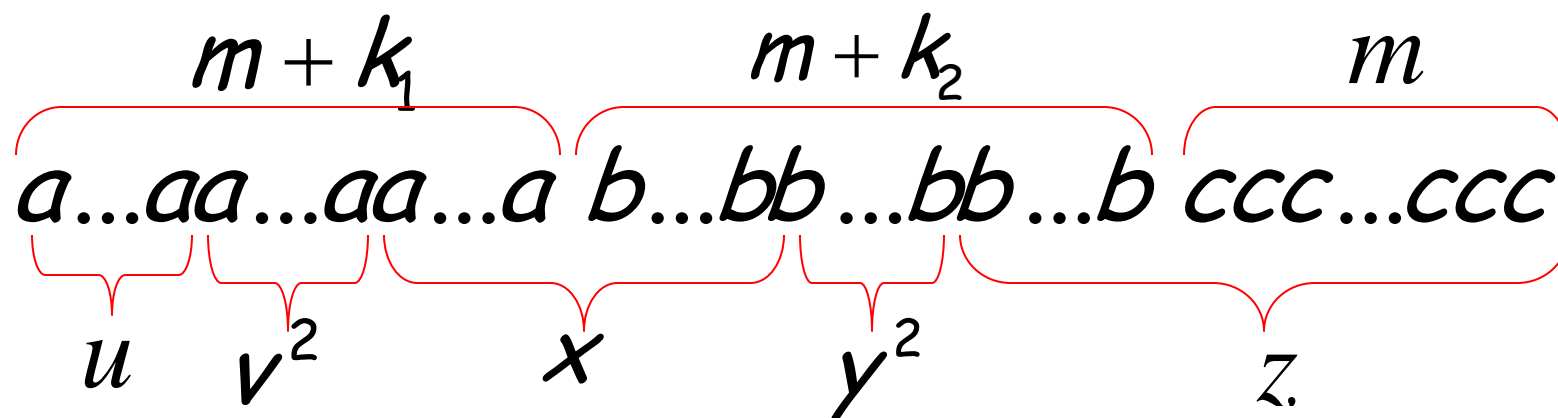


$$L = \{a^n b^n c^n : n \geq 0\}$$

$$w = a^m b^m c^m$$

$$w = uvxyz \quad |vxy| \leq m \quad |vy| \geq 1$$

$$v = a^{k_1} \quad y = a^{k_2} \quad k_1 + k_2 \geq 1$$



$$L = \{a^n b^n c^n : n \geq 0\}$$

$$w = a^m b^m c^m$$

$$w = uvxyz \quad |vxy| \leq m \quad |vy| \geq 1$$

From Pumping Lemma: $uv^2xy^2z \in L$

$$k_1 + k_2 \geq 1$$

However: $uv^2xy^2z = a^{m+k_1} b^{m+k_2} c^m \notin L$

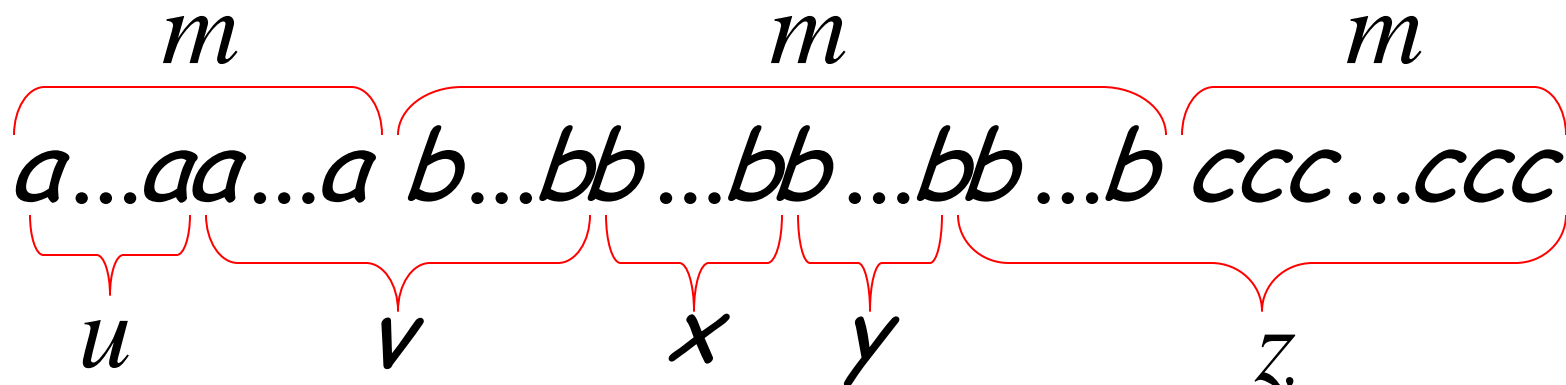
Contradiction!!!

$$L = \{a^n b^n c^n : n \geq 0\}$$

$$w = a^m b^m c^m$$

$$w = uvxyz \quad |vxy| \leq m \quad |vy| \geq 1$$

Sub-case 2: v contains a and b
 y contains only b



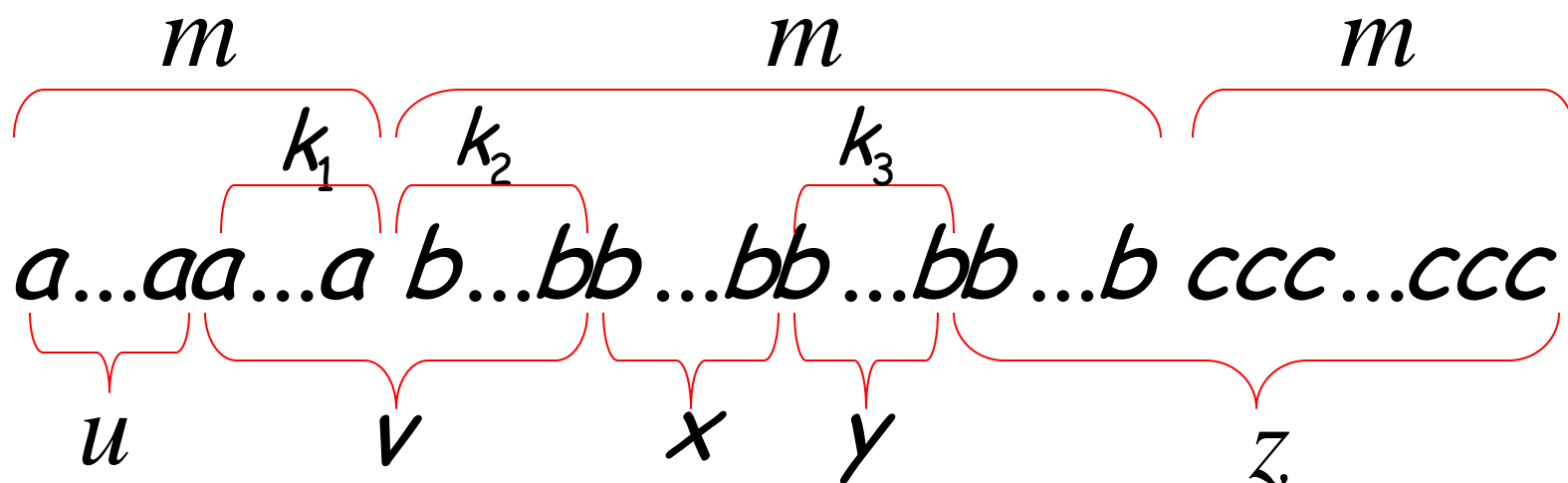
$$L = \{a^n b^n c^n : n \geq 0\}$$

$$w = a^m b^m c^m$$

$$w = uvxyz \quad |vxy| \leq m \quad |vy| \geq 1$$

By assumption

$$v = a^{k_1} b^{k_2} \quad y = a^{k_3} \quad k_1, k_2 \geq 1$$

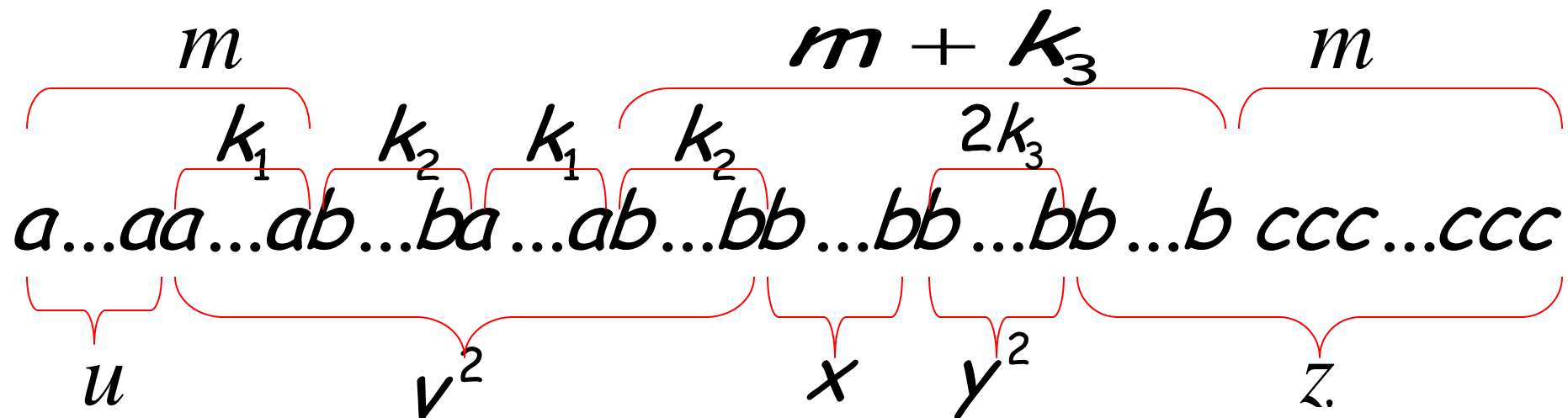


$$L = \{a^n b^n c^n : n \geq 0\}$$

$$w = a^m b^m c^m$$

$$w = uvxyz \quad |vxy| \leq m \quad |vy| \geq 1$$

$$v = a^{k_1} b^{k_2} \quad y = a^{k_3} \quad k_1, k_2 \geq 1$$



$$L = \{a^n b^n c^n : n \geq 0\}$$

$$w = a^m b^m c^m$$

$$w = uvxyz \quad |vxy| \leq m \quad |vy| \geq 1$$

From Pumping Lemma: $uv^2xy^2z \in L$

$$k_1, k_2 \geq 1$$

However: $uv^2xy^2z = a^m b^{k_2} a^{k_1} b^{m+k_3} c^m \notin L$

Contradiction!!!

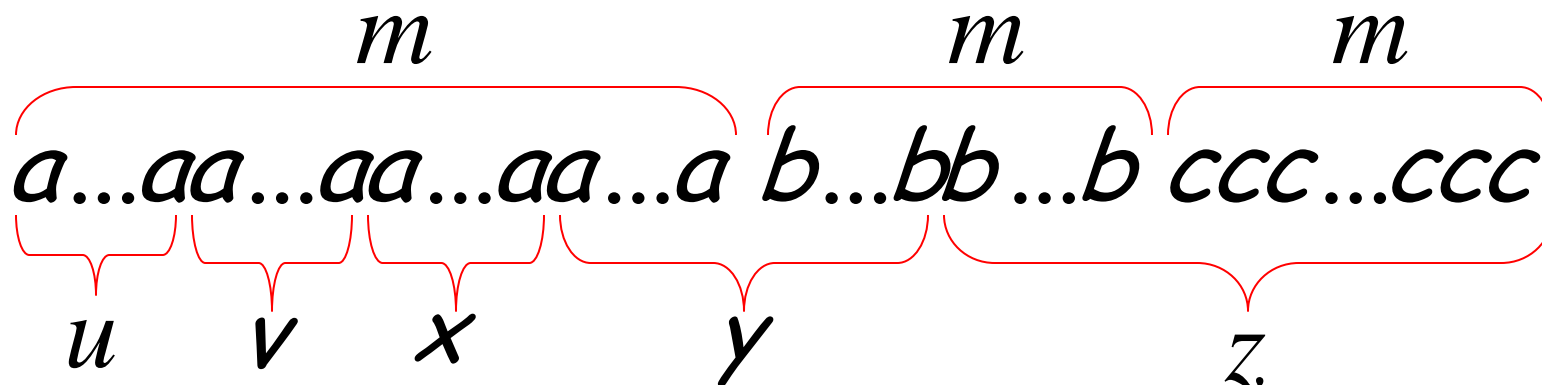
$$L = \{a^n b^n c^n : n \geq 0\}$$

$$w = a^m b^m c^m$$

$$w = uvxyz \quad |vxy| \leq m \quad |vy| \geq 1$$

Sub-case 3: v contains only a
 y contains a and b

Similar to sub-case 2



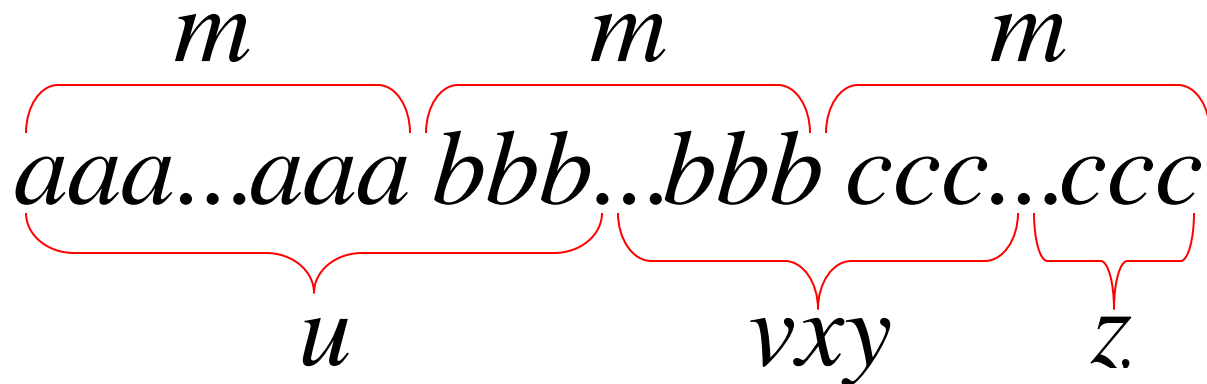
$$L = \{a^n b^n c^n : n \geq 0\}$$

$$w = a^m b^m c^m$$

$$w = uvxyz \quad |vxy| \leq m \quad |vy| \geq 1$$

Case 5: vxy overlaps b^m and c^m

Similar to case 4



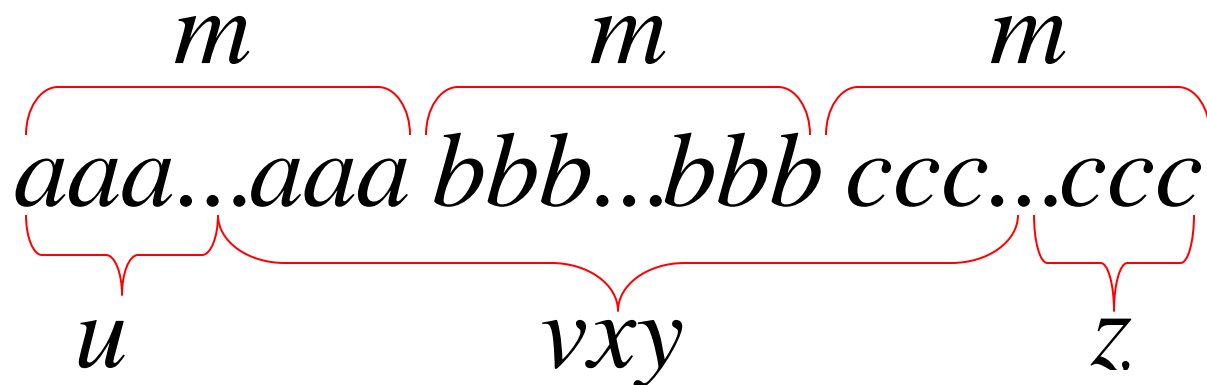
$$L = \{a^n b^n c^n : n \geq 0\}$$

$$w = a^m b^m c^m$$

$$w = uvxyz \quad |vxy| \leq m \quad |vy| \geq 1$$

Case 6: vxy overlaps a^m , b^m and c^m

Impossible!



In all cases we obtained a **contradiction**

Therefore: the original assumption that

$$L = \{a^n b^n c^n : n \geq 0\}$$

is context-free must be wrong

Conclusion: L is not context-free