# Edit Distance

張智星 (Roger Jang)

*jang@mirlab.org*

*http://mirlab.org/jang*

多媒體資訊檢索實驗室

台灣大學 資訊工程系

# Edit Distance

- Edit distance
  - The minimum total cost of the basic operations (delete with cost 1, insert with cost 1, substitute with cost 2) that are required to convert a string into another.
  - Used in “diff” command of unix.

Quiz!

# DP for Edit Distance: 3-step Formula

Three-step DP formula for computing $ed\left(\vec{A},\vec{B}\right)$

1. Optimum-value function

   $ed\left(\vec{p},\vec{q}\right)$ is the edit distance between string $\vec{p}$ and $\vec{q}$.

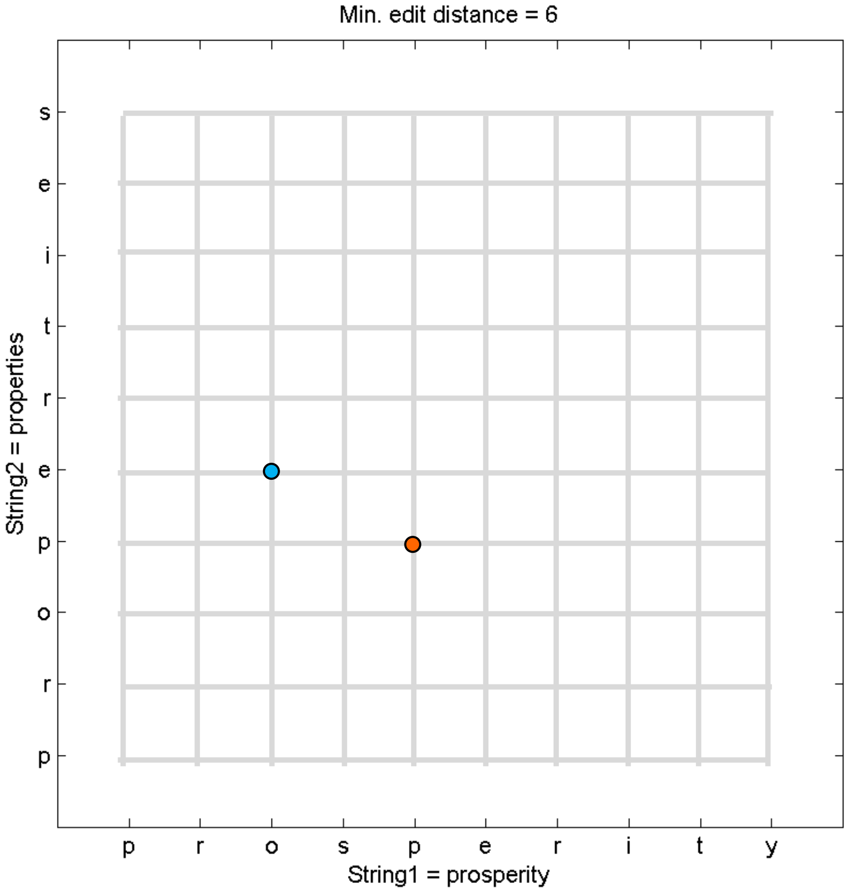
2. Recurrent formula

$$ed\left(\vec{a}x,\vec{b}y\right)=\begin{cases} ed\left(\vec{a},\vec{b}\right), \text{ if } x=y \\ \min\begin{cases} ed\left(\vec{a}x,\vec{b}\right)+1 \\ ed\left(\vec{a},\vec{b}y\right)+1, \text{ if } x\neq y \\ ed\left(\vec{a},\vec{b}\right)+2 \end{cases} \end{cases}$$

   Boundary condition: $ed\left(\vec{a},[]\right)=len\left(\vec{a}\right), ed\left([],\vec{b}\right)=len\left(\vec{b}\right)$

3. Answer: $ed\left(\vec{A},\vec{B}\right)$

# DP for Edit Distance: Table Filling (1/2)

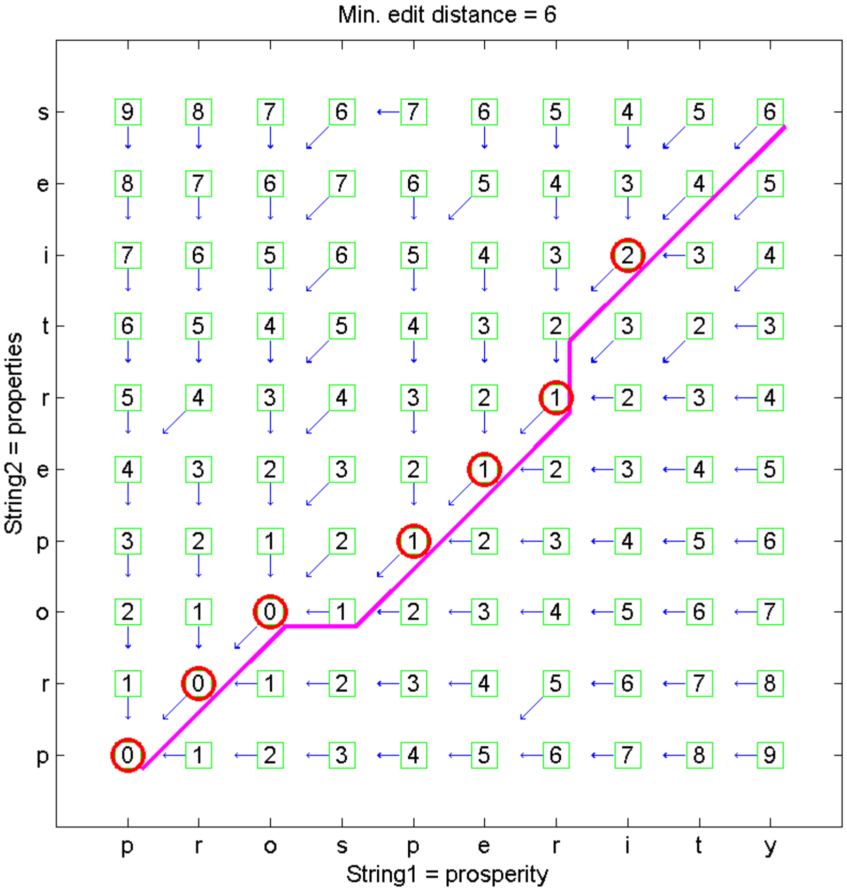


- All DP problems can be visualized as path finding…

● $ed(prosp, prop) = ed(pros, pro\ )$

● $ed(pro, prope) = \min \begin{cases} ed(pro, prop) + 1 \\ ed(pr, prope) + 1 \\ ed(pr, prop) + 2 \end{cases}$

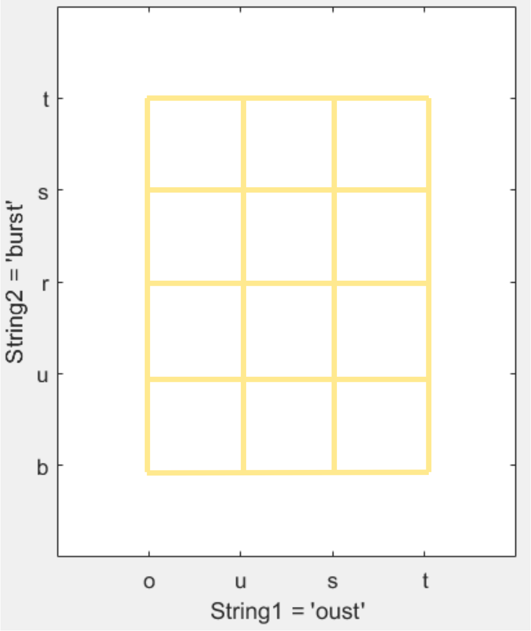
# DP for Edit Distance: Table Filling (2/2)



|   | p | r | o | s | p | e | r | i | t | y |
|---|---|---|---|---|---|---|---|---|---|---|
| s | 9 | 8 | 7 | 6 | 7 | 6 | 5 | 4 | 5 | 6 |
| e | 8 | 7 | 6 | 7 | 6 | 5 | 4 | 3 | 4 | 5 |
| i | 7 | 6 | 5 | 6 | 5 | 4 | 3 | 2 | 3 | 4 |
| t | 6 | 5 | 4 | 5 | 4 | 3 | 2 | 3 | 2 | 3 |
| r | 5 | 4 | 3 | 4 | 3 | 2 | 1 | 2 | 3 | 4 |
| e | 4 | 3 | 2 | 3 | 2 | 1 | 2 | 3 | 4 | 5 |
| p | 3 | 2 | 1 | 2 | 1 | 2 | 3 | 4 | 5 | 6 |
| o | 2 | 1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| r | 1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| p | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |

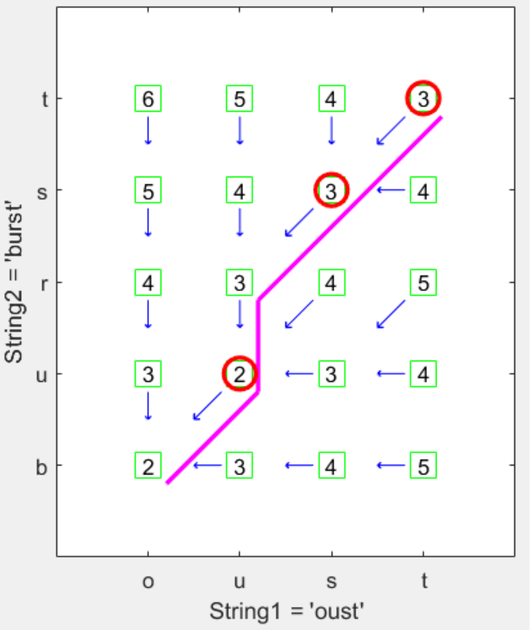
- To create this plot
  - Download Machine Learning Toolbox
  - Run editDistance('prosperity', 'properties', [2 1 1], 1) under MATLAB
  - A 45-degree back-tracking path indicates a “substitute” or “equal”.

# DP for Edit Distance: Table Filling



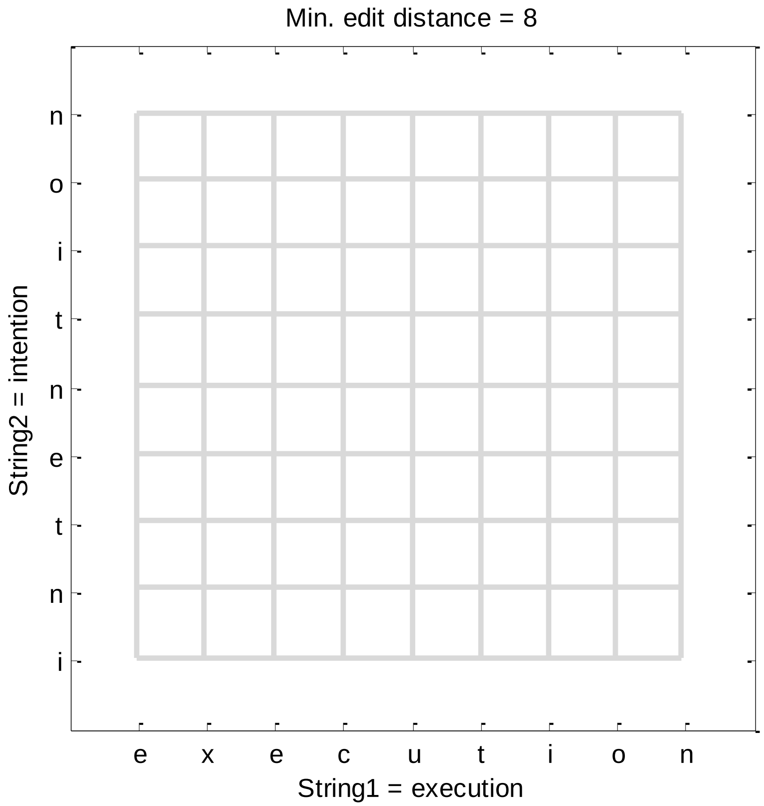
- Hints
  - Create a (m+1)*(n+1) matrix for table filling
  - Fill row 0 and column 0 first to establish the base cases of boundary conditions
  - Fill all the other elements in a layer-by-layer manner.

# DP for Edit Distance: Solution

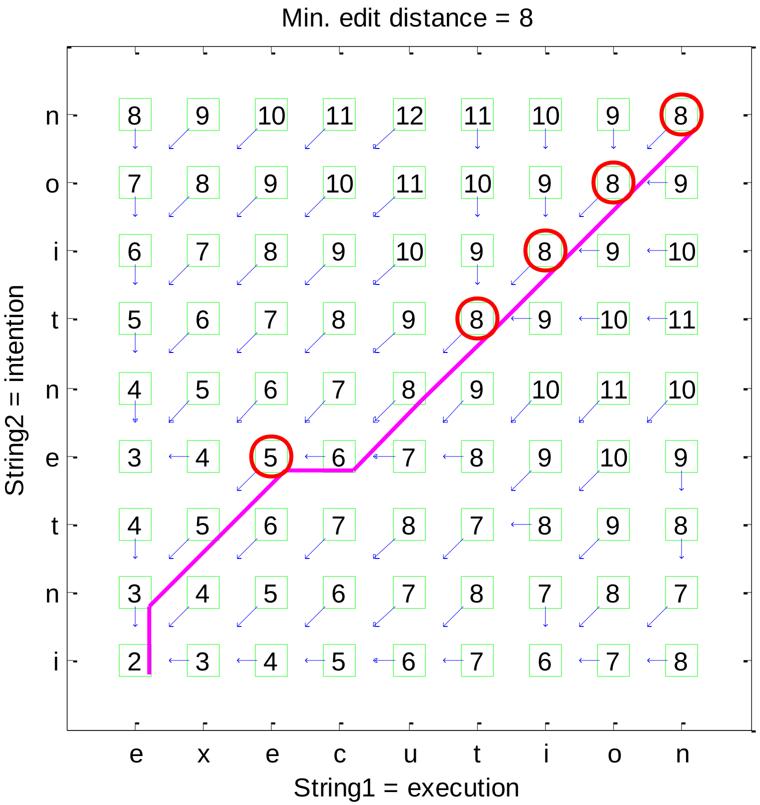


|  | o | u | s | t |
|---|---|---|---|---|
| t | 6 | 5 | 4 | 3 |
| s | 5 | 4 | 3 | 4 |
| r | 4 | 3 | 4 | 5 |
| u | 3 | 2 | 3 | 4 |
| b | 2 | 3 | 4 | 5 |

# Exercise


Min. edit distance = 8
n
o
i
t
n
e
t
n
i
String2 = intention
e
x
e
c
u
t
i
o
n
String1 = execution

# Solution



Min. edit distance = 8

| String2 = intention \ String1 = execution | e | x | e | c | u | t | i | o | n |
|---|---|---|---|---|---|---|---|---|---|
| n | 8 | 9 | 10 | 11 | 12 | 11 | 10 | 9 | 8 |
| o | 7 | 8 | 9 | 10 | 11 | 10 | 9 | 8 | 9 |
| i | 6 | 7 | 8 | 9 | 10 | 9 | 8 | 9 | 10 |
| t | 5 | 6 | 7 | 8 | 9 | 8 | 9 | 10 | 11 |
| n | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 10 |
| e | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 9 |
| t | 4 | 5 | 6 | 7 | 8 | 7 | 8 | 9 | 8 |
| n | 3 | 4 | 5 | 6 | 7 | 8 | 7 | 8 | 7 |
| i | 2 | 3 | 4 | 5 | 6 | 7 | 6 | 7 | 8 |