Applications of Suffix Trees: Longest common substrings

Lecture 3.3

by Marina Barsky

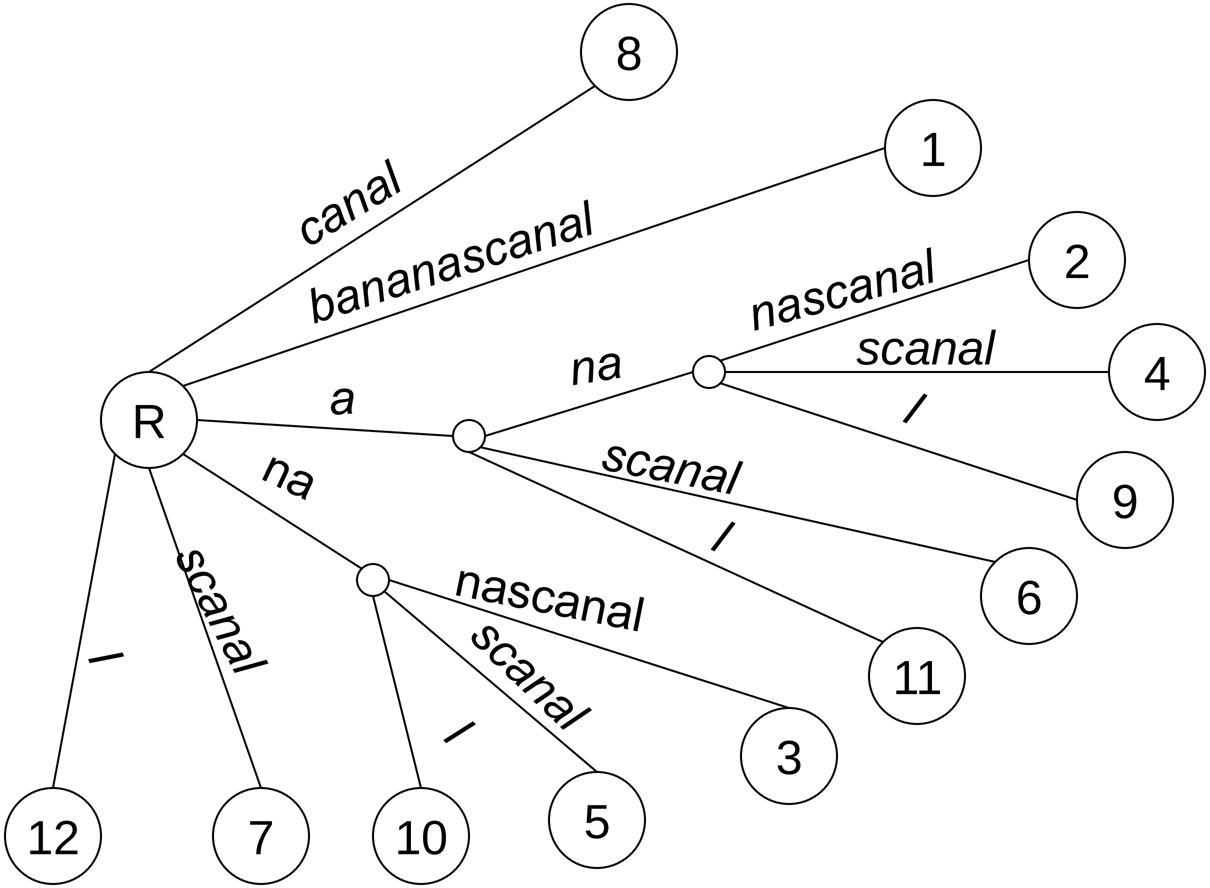
The longest common substring of several strings

- The problem: find the longest substring common to two given strings I and II.
  - For example, if I=*superiorcalifornialives* and II=*sealiver*, then the longest common substring of I and II is *alive*.
- 1970 – Knuth conjectured that the linear-time solution to the longest common substring problem would be impossible

The longest common substring for 2 strings in linear time

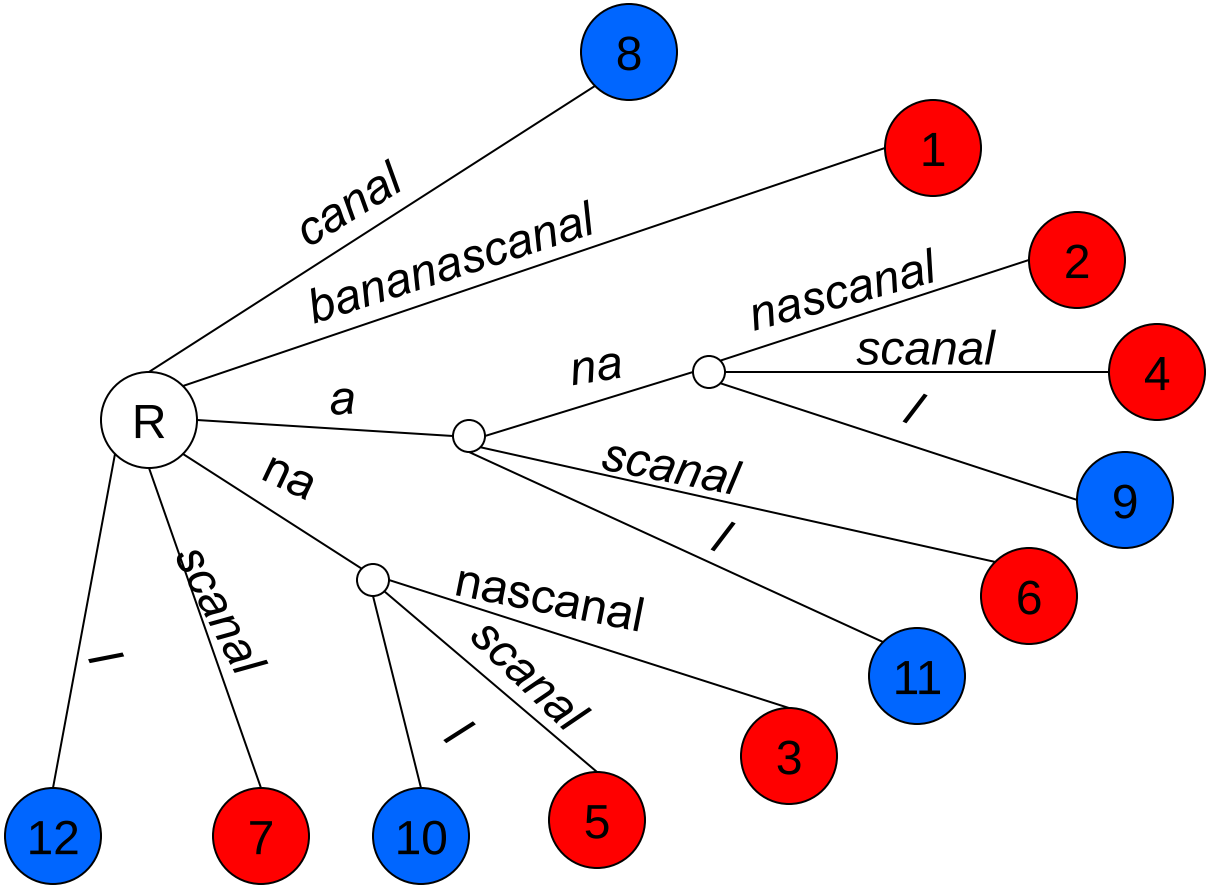
- Concatenate 2 strings and build the suffix tree for the concatenated string
- Label each leaf with the corresponding suffix start position, plus the ID of the string (I or II)
- Perform the depth-first traversal and mark each internal node by I, II or both, depending what suffixes are found in the subtree for this node
- Find the deepest internal node which is marked by both I and II

Example: *I=bananas II=canal*



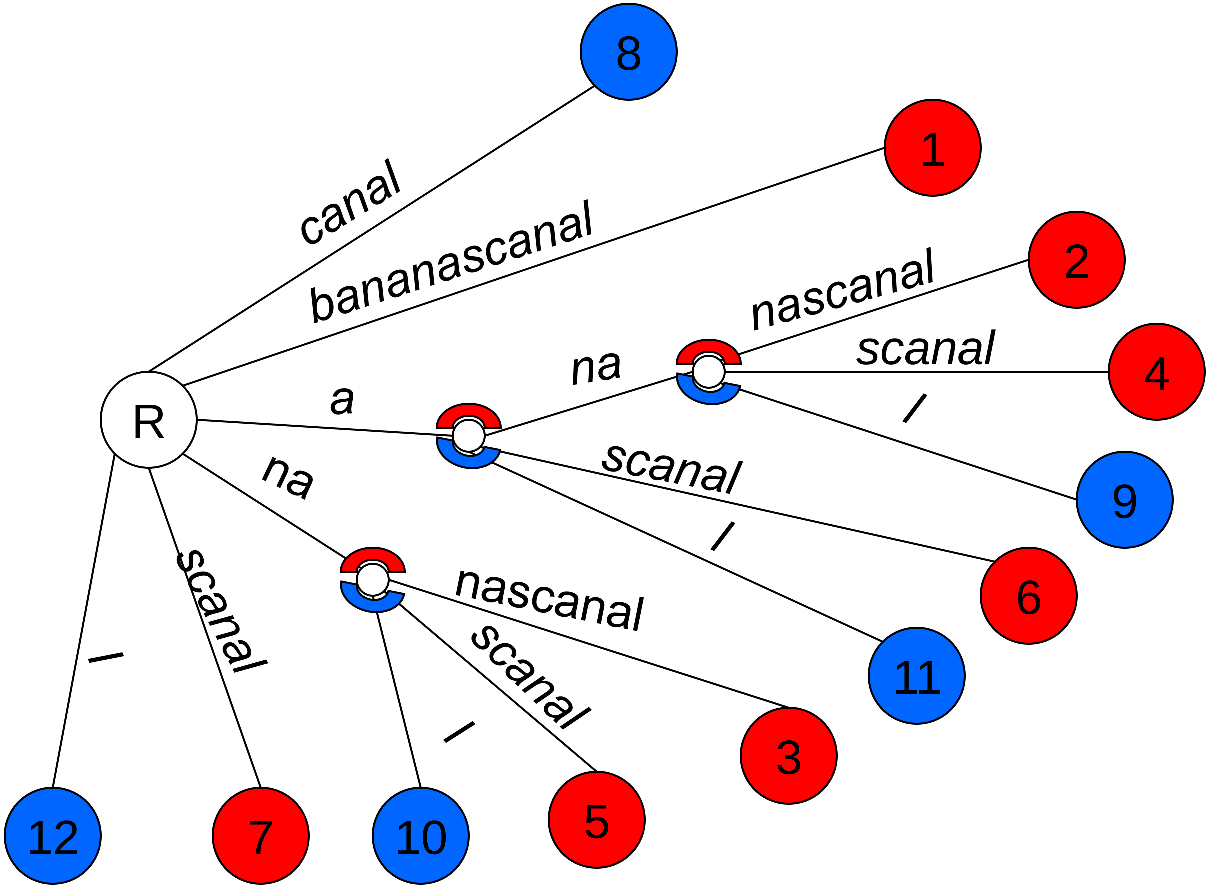
b	*a*	*n*	*a*	*n*	*a*	*s*	*c*	*a*	*n*	*a*	*l*
1	2	3	4	5	6	7	8	9	10	11	12

Example: *I=bananas II=canal*



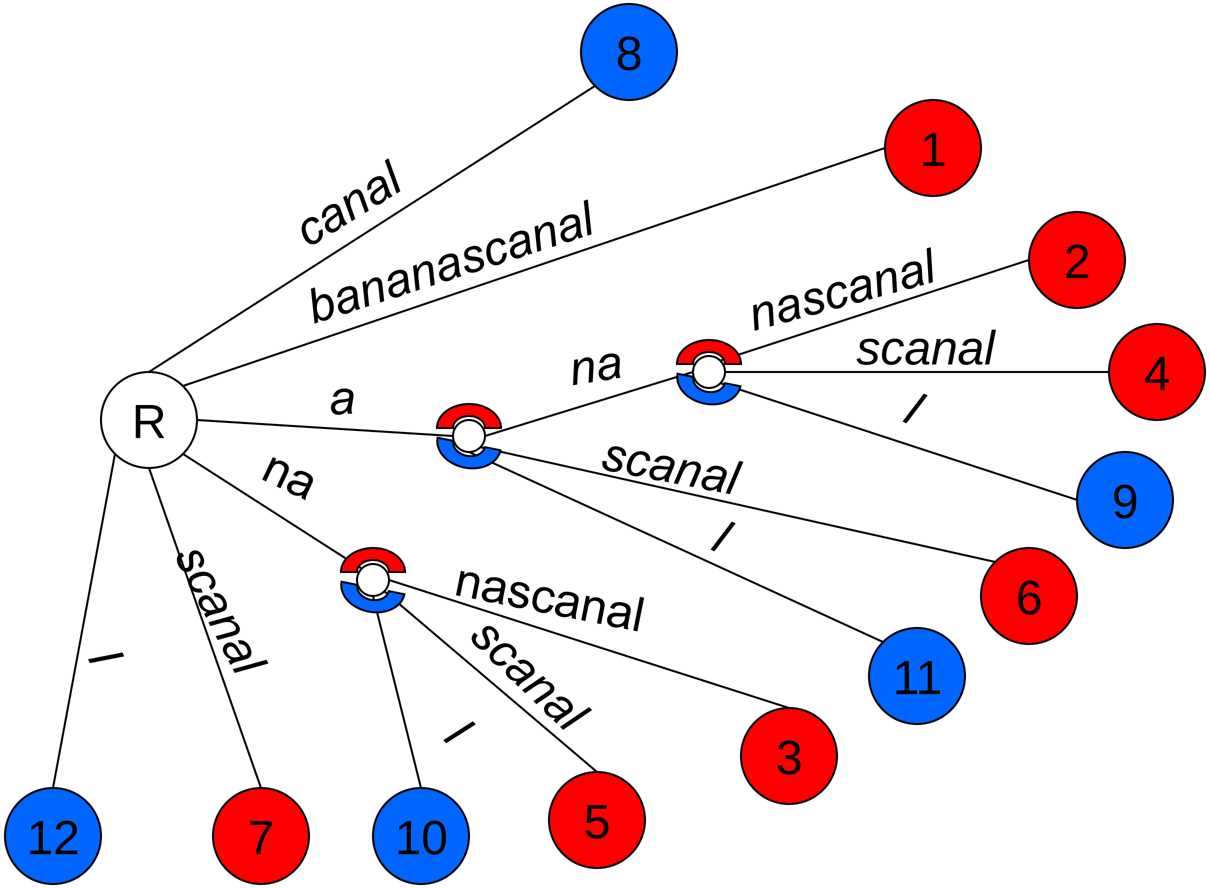
b	*a*	*n*	*a*	*n*	*a*	*s*	*c*	*a*	*n*	*a*	*l*
1	2	3	4	5	6	7	8	9	10	11	12

Example: Marking internal nodes



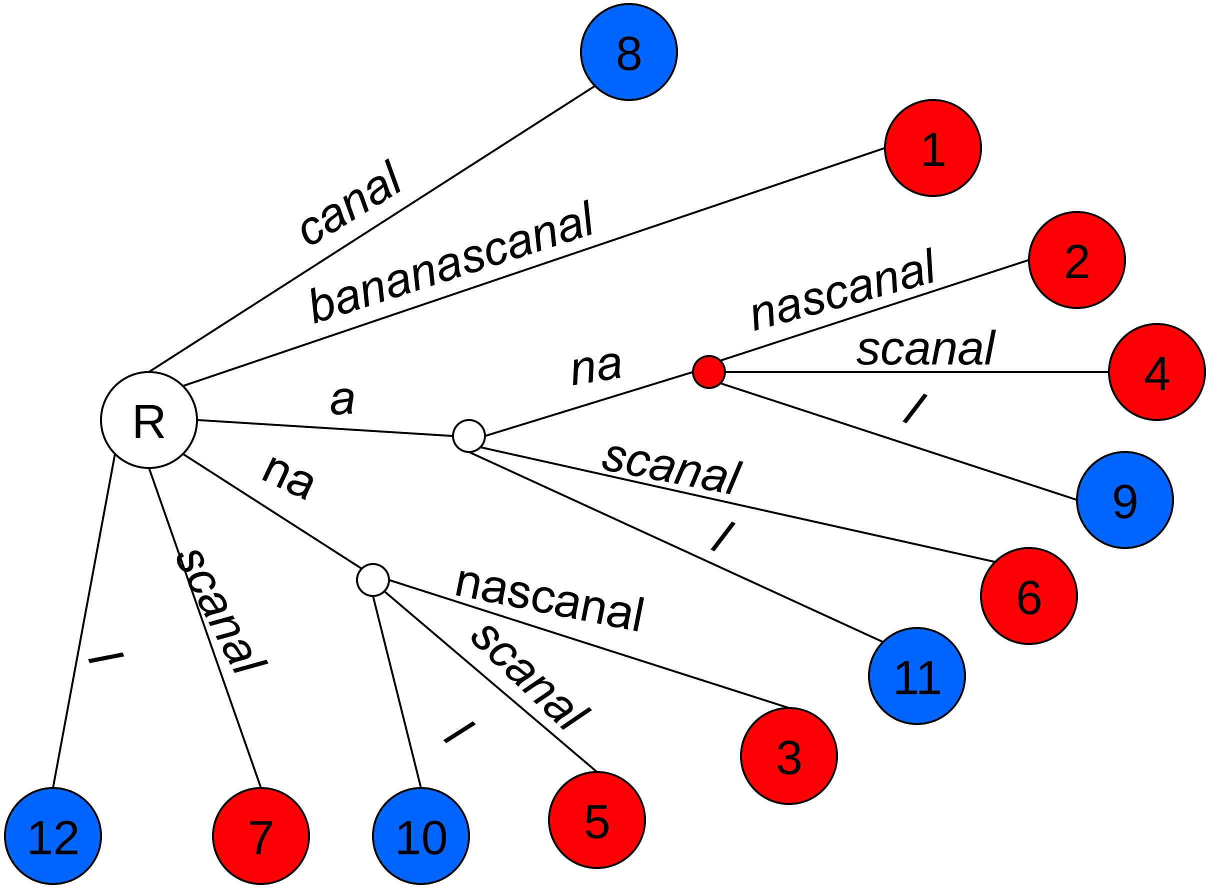
b	a	n	a	n	a	s	c	a	n	a	l
1	2	3	4	5	6	7	8	9	10	11	12

Example: What is the longest common substring ?



b	a	n	a	n	a	s	c	a	n	a	l
1	2	3	4	5	6	7	8	9	10	11	12

Example: LCS=*ana*

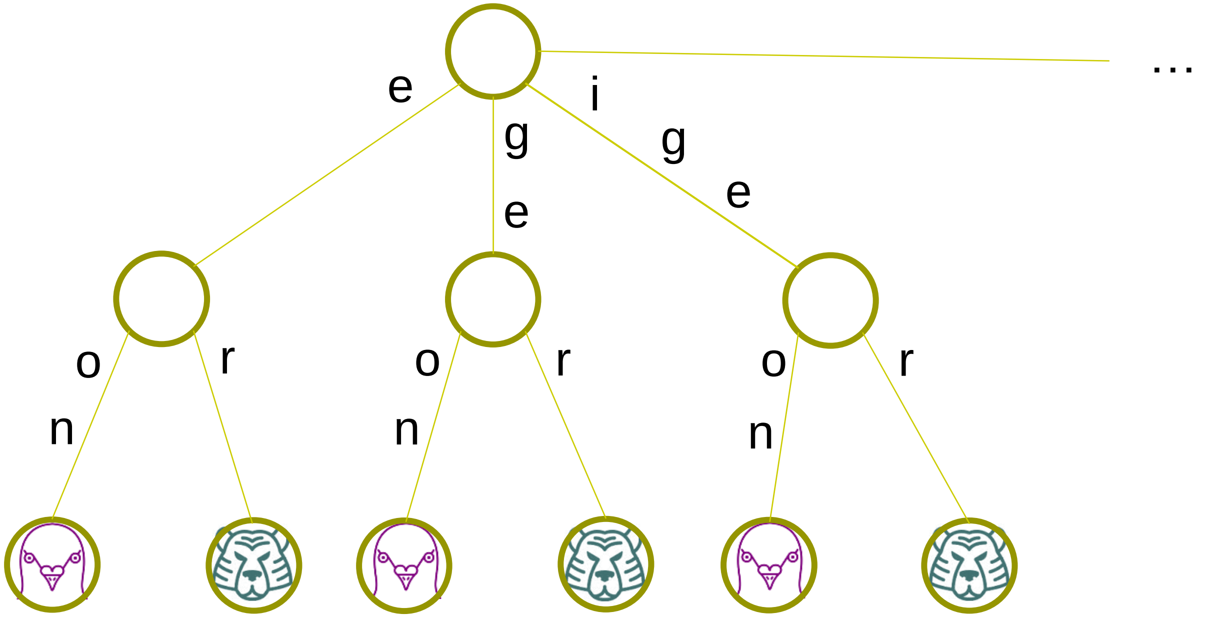


b	*a*	*n*	*a*	*n*	*a*	*s*	*c*	*a*	*n*	*a*	*l*
1	2	3	4	5	6	7	8	9	10	11	12

I

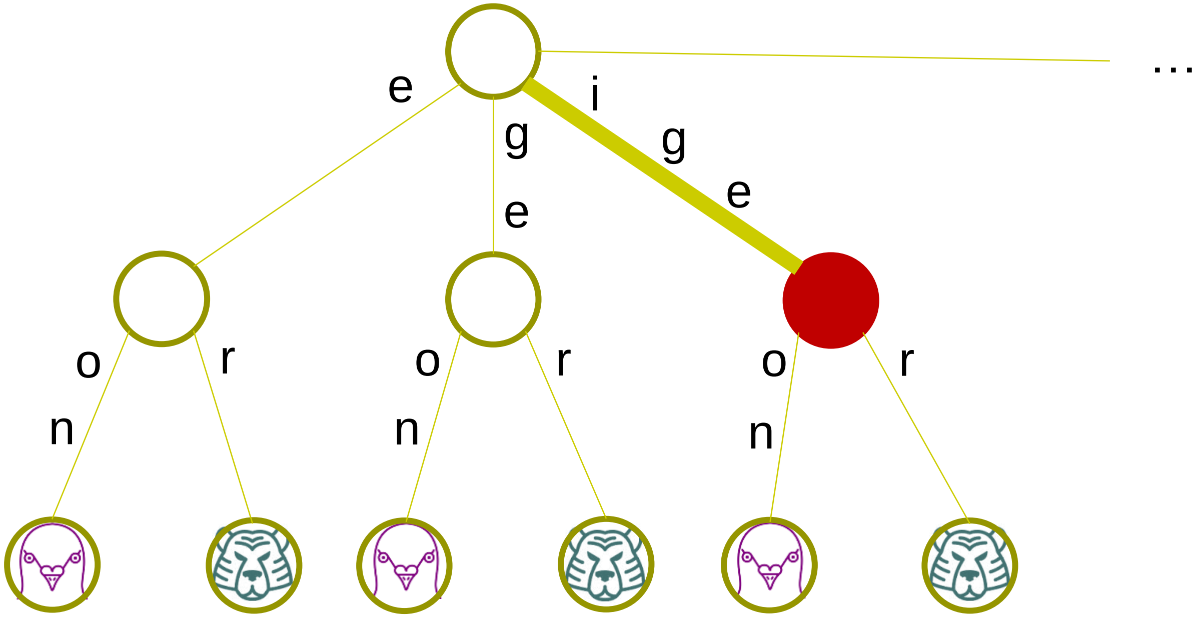
II

Longest common substrings: example



Query: what do *tiger* and *pigeon* have in common?

Longest common substrings: example



Query: what do *tiger* and *pigeon* have in common?

Common substrings for a set of DNA sequences

Insert suffixes of multiple strings into one tree

- Discover substrings common to viruses and humans
- Discover substrings unique to cancer

Used in the identification of the remains of US military personnel

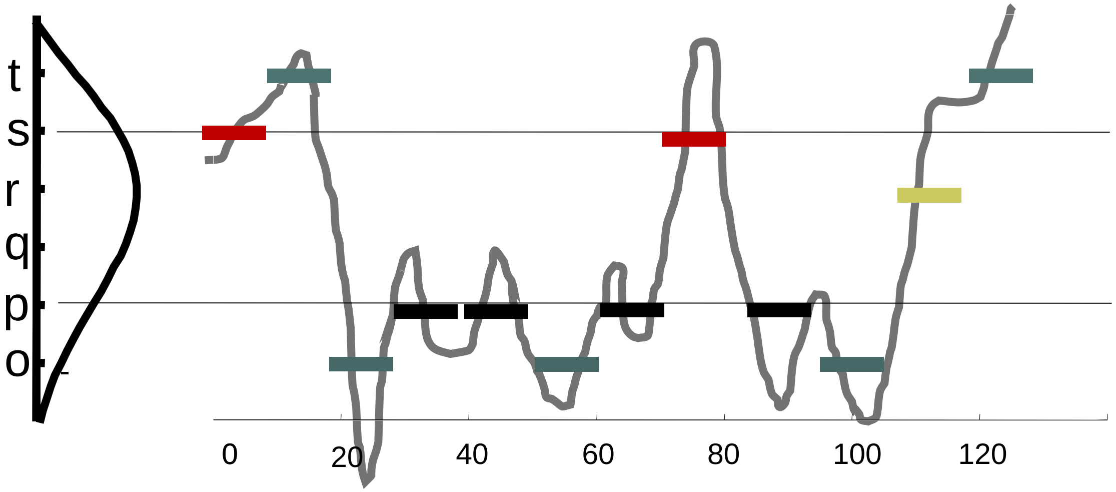
- Mitochondrial DNA from live person is collected, sequenced and the sequences are stored in the database (I)
- Later, the DNA is extracted from the remains (II), and the longest common substring of I and II helps to narrow down the search

Potential applications of suffix trees for other types of sequential data

Time series as strings

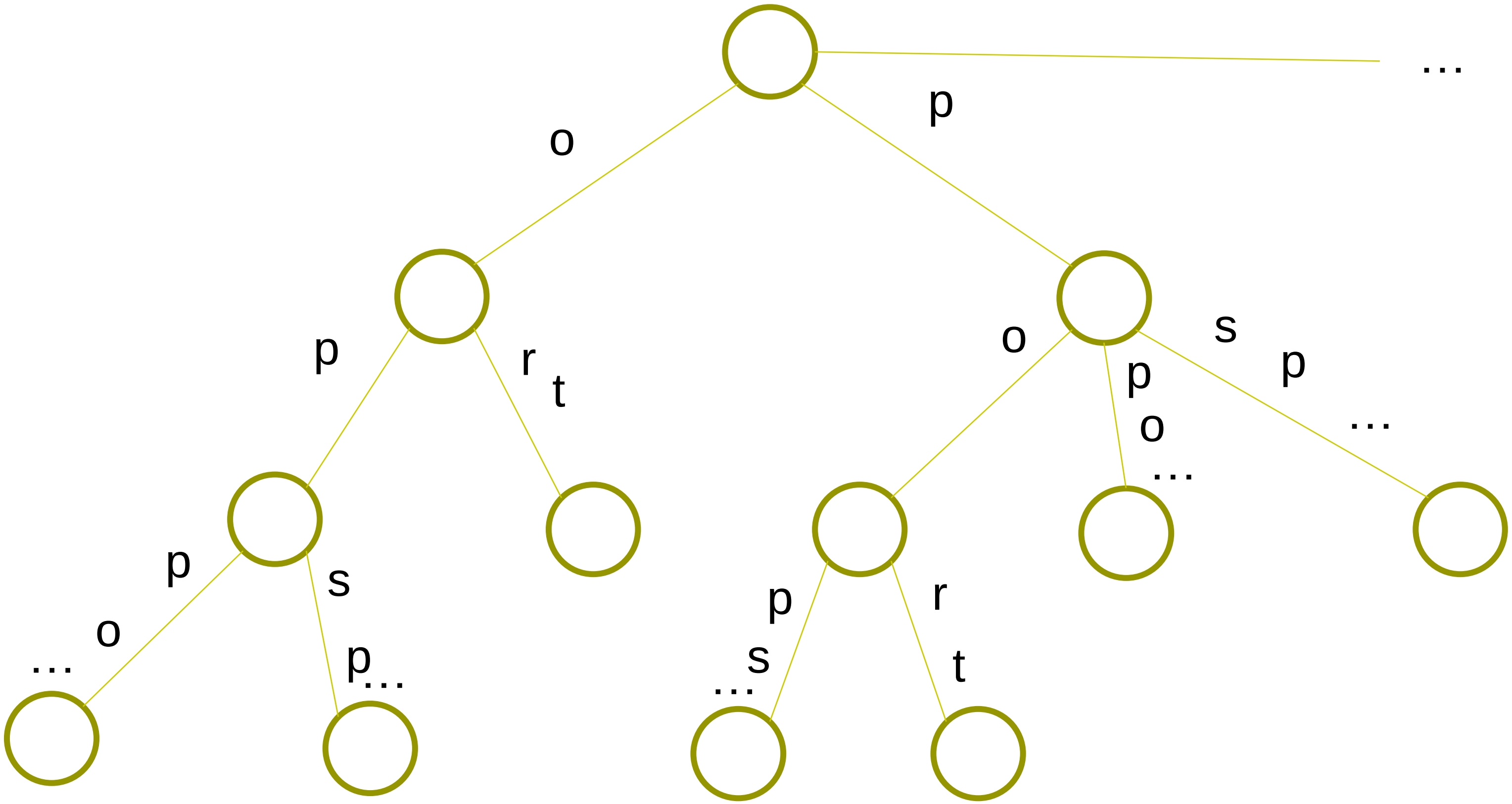
SAX - Symbolic Aggregate approximation

(by Eamon Keough, 2001)

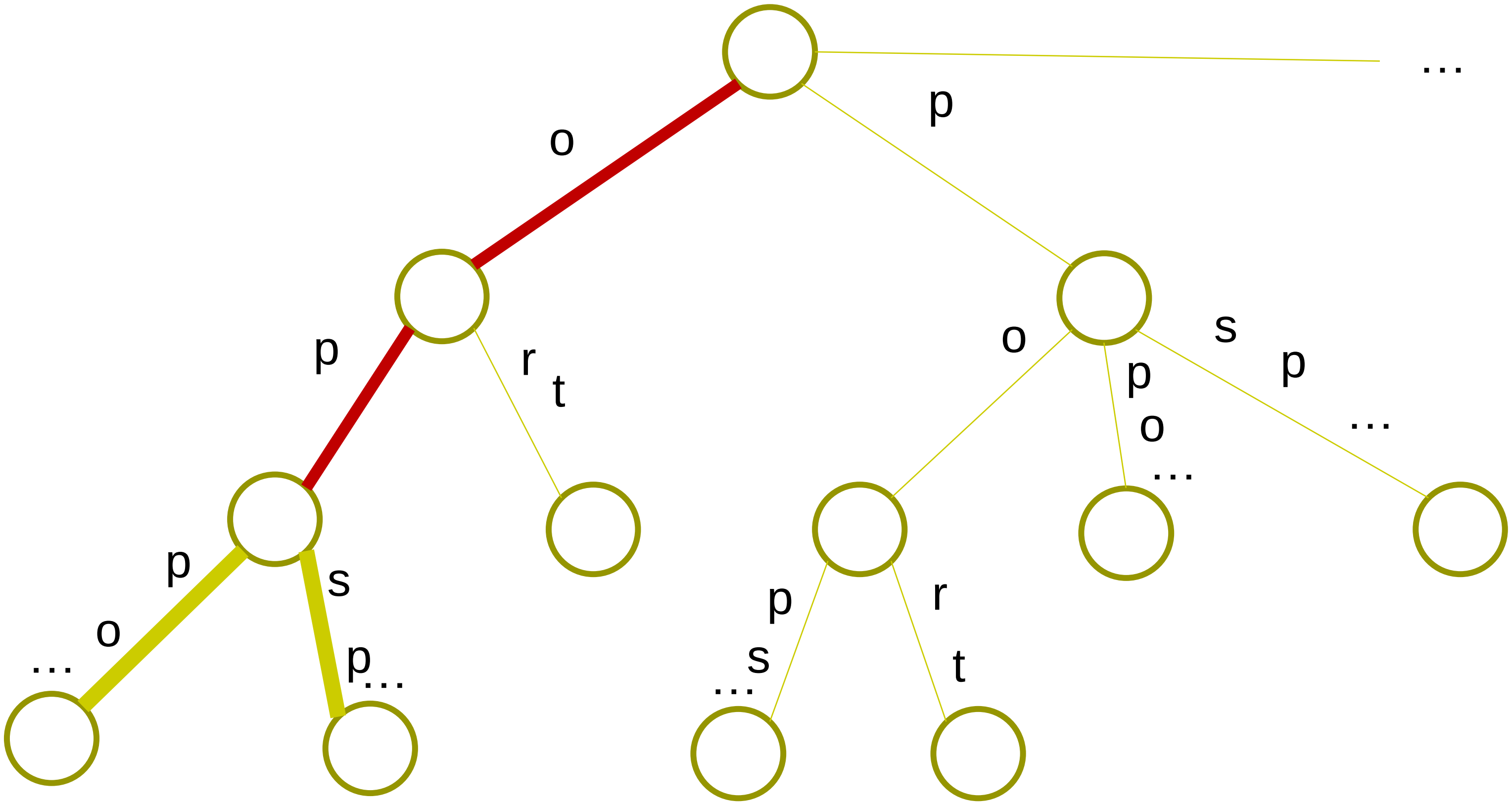


stoppopsport

Suffix trees for time series



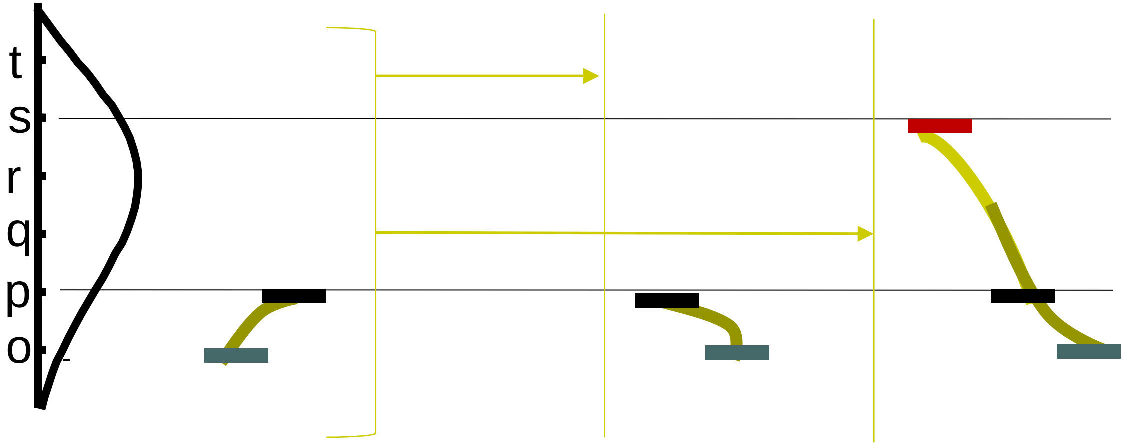
Suffix trees for time series



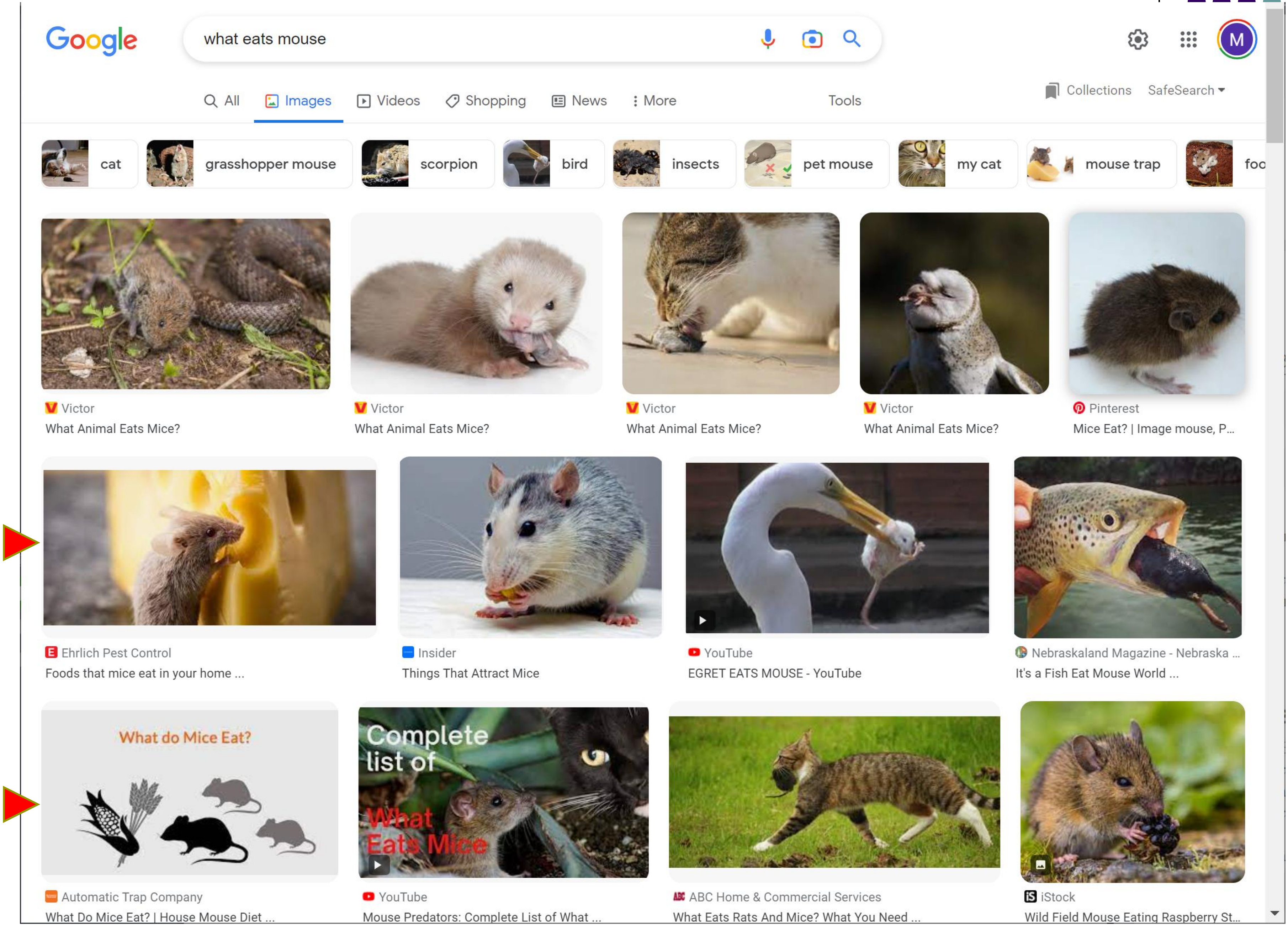
Query: what happened after *op*?

Suffix trees for time series: rise and fall of stocks

50% *po*, 50% *spo*



Query: what happened after *op*?



what eats mouse

Victor
What Animal Eats Mice?

Victor
What Animal Eats Mice?

Victor
What Animal Eats Mice?

Victor
What Animal Eats Mice?

Pinterest
Mice Eat? | Image mouse, P...

Ehrlich Pest Control
Foods that mice eat in your home ...

Insider
Things That Attract Mice

YouTube
EGRET EATS MOUSE - YouTube

Nebraskaland Magazine - Nebraska ...
It's a Fish Eat Mouse World ...

Automatic Trap Company
What Do Mice Eat? | House Mouse Diet ...

YouTube
Mouse Predators: Complete List of What ...

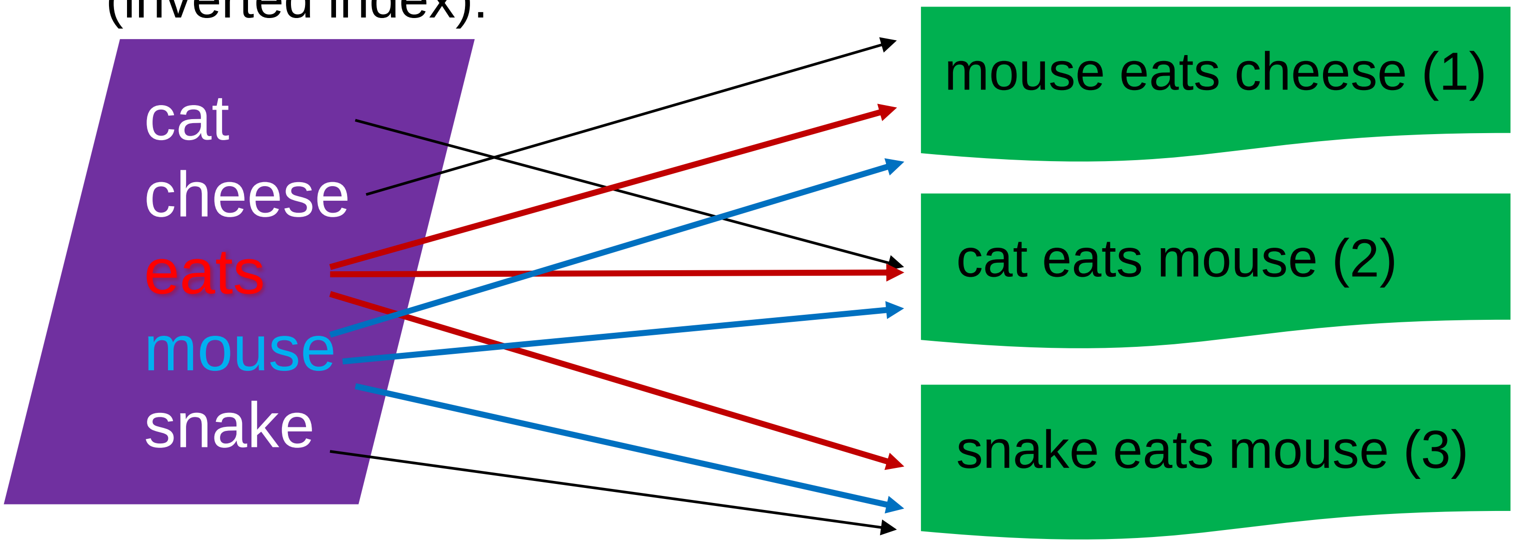
ABC Home & Commercial Services
What Eats Rats And Mice? What You Need ...

iStock
Wild Field Mouse Eating Raspberry St...

Inverted index

Query: What animal “**eats mouse**”

WORD-BASED INDEX
(inverted index):



cat
cheese
eats
mouse
snake

mouse eats cheese (1)

cat eats mouse (2)

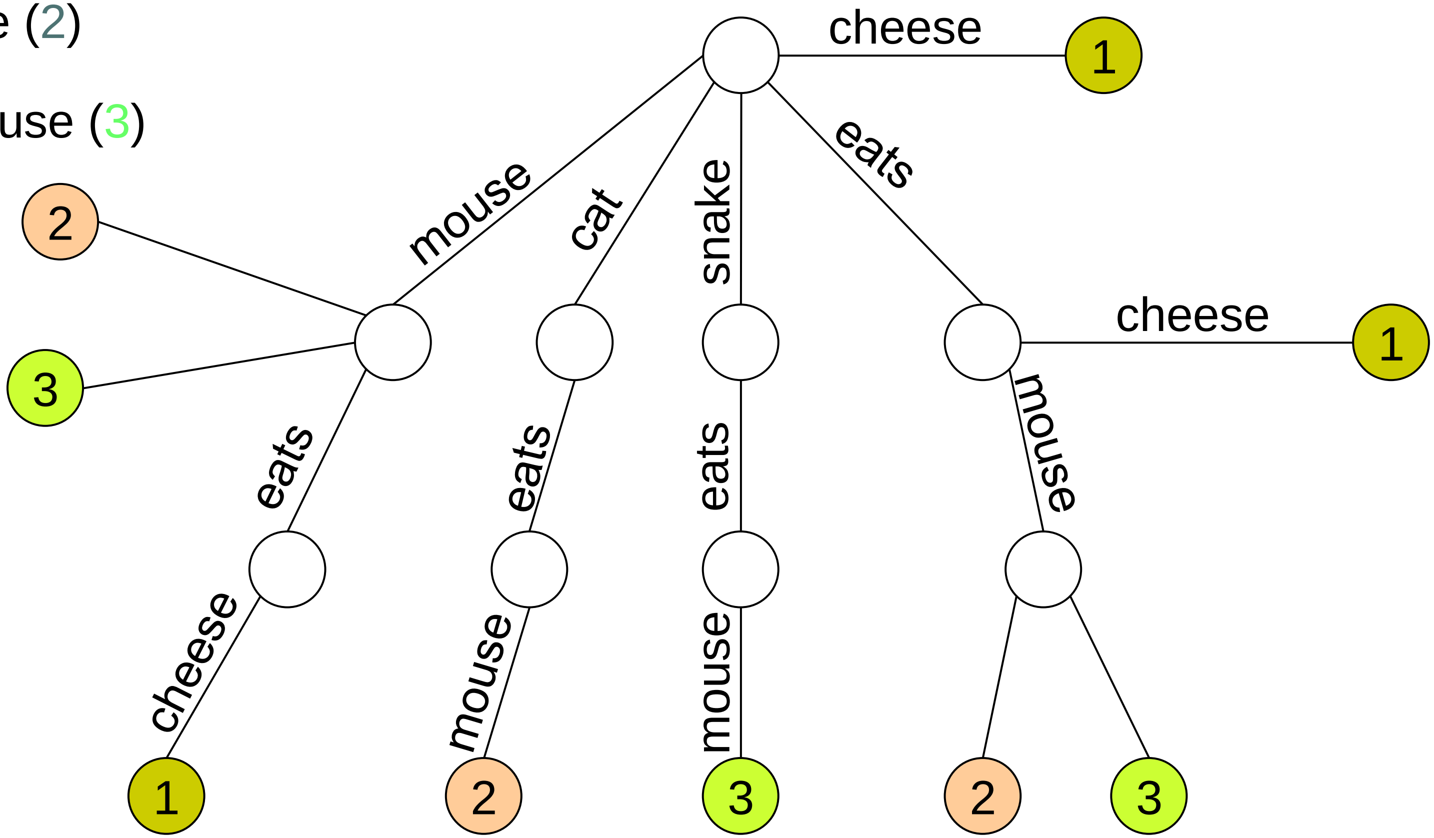
snake eats mouse (3)

Answer is in documents 1,2,3

Meaningful search: example

Collection of 1-sentence documents

- mouse eats cheese (1)
- cat eats mouse (2)
- snake eats mouse (3)

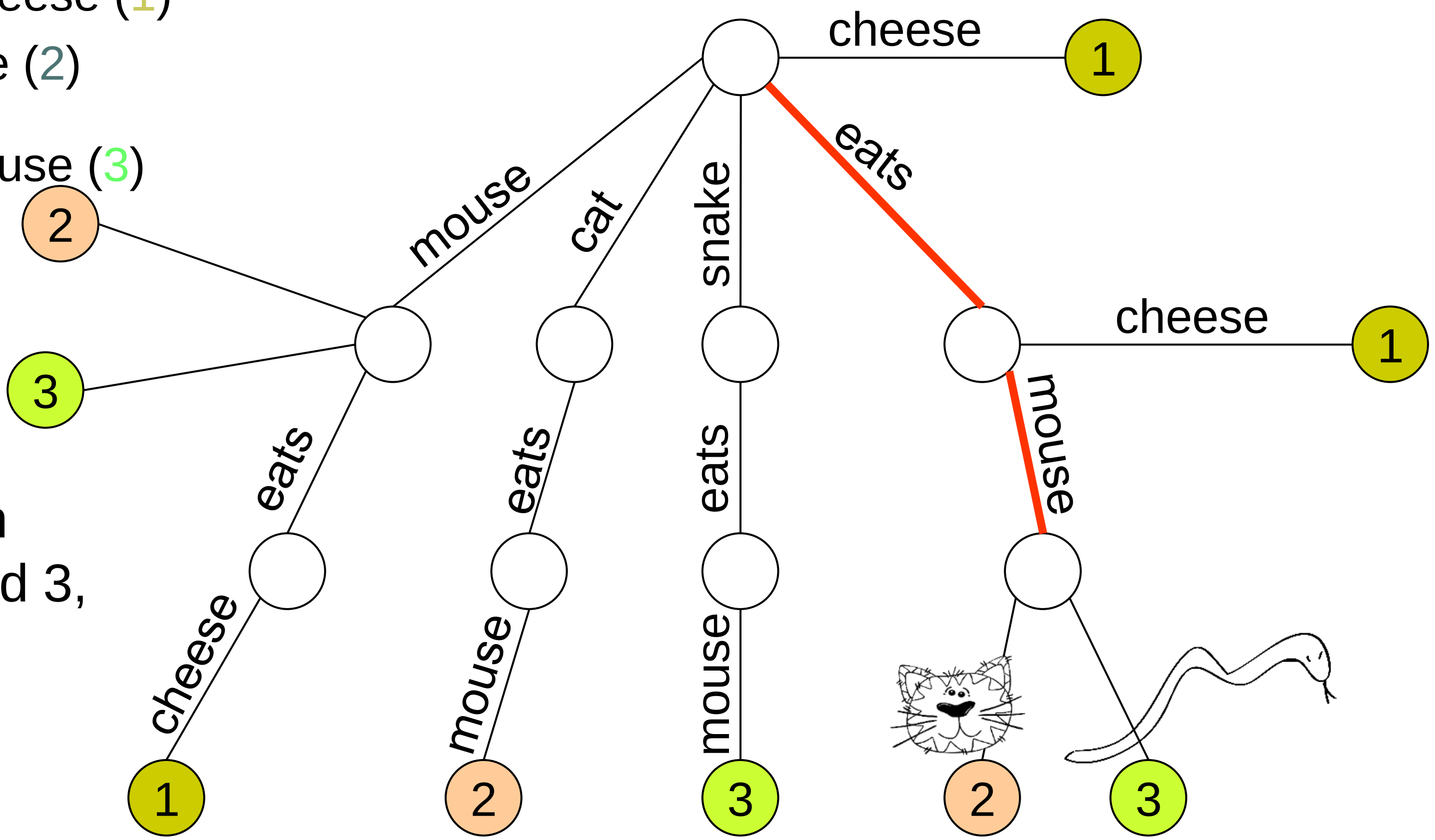


Meaningful search: example

Query: What animal "**eats mouse**"

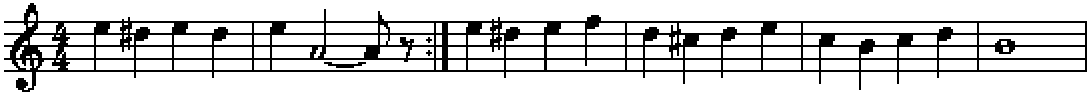
Collection of 1-sentence documents

- mouse eats cheese (1)
- cat eats mouse (2)
- snake eats mouse (3)

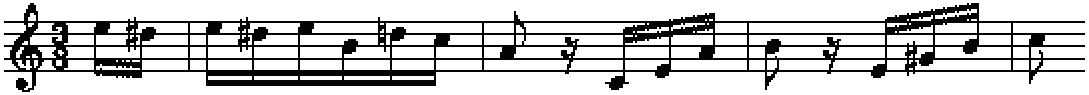


The answer is in documents 2 and 3, but not in 1

Suffix tree for melodies …



Saint-Saëns, Camille (1835-1921), Carnaval des Animaux, Orch. & 2 Pfts., Aquarium

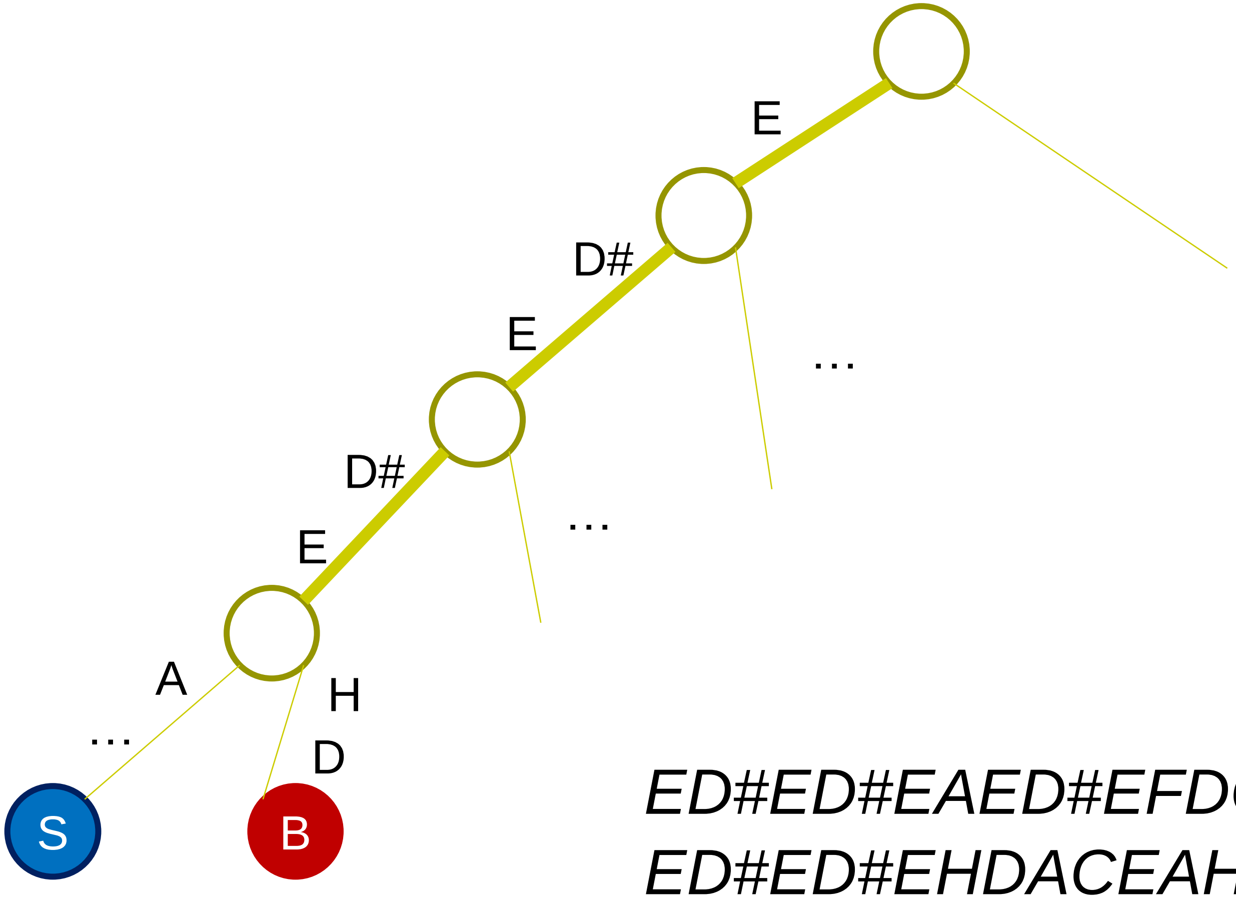


Beethoven, Ludwig Van (1770-1827), Für Elise, Pft.

ED#ED#EAED#EFDC#DECHCDH (**S-S**

ED#ED#EHDACEAHEG#C (**B**)

Suffix tree for melodies and plagiarism detection



ED#ED#EAED#EFDC#DECHCDH (**S-S**

ED#ED#EHDACEAHEG#C (**B**)

Indexing melodies…

Song 1

F [BALLADE]
[Zu Strassburg steht ein hohes Haus]
REG[Deutschland / Frankreich, Lothringen]
MEL[-5_ 1_.23_4_ 2_.31_
  -5_ 1_.23_4_ 2_231_
  3_ 5_5_5_66 5__2_
  2_ 5_4_3_2_ 1_-6_-5_
  -5_ 1_2_3_4_ 2__1_ //] >>
FCT[Ballade, Braut - Werbung, Erpressung]

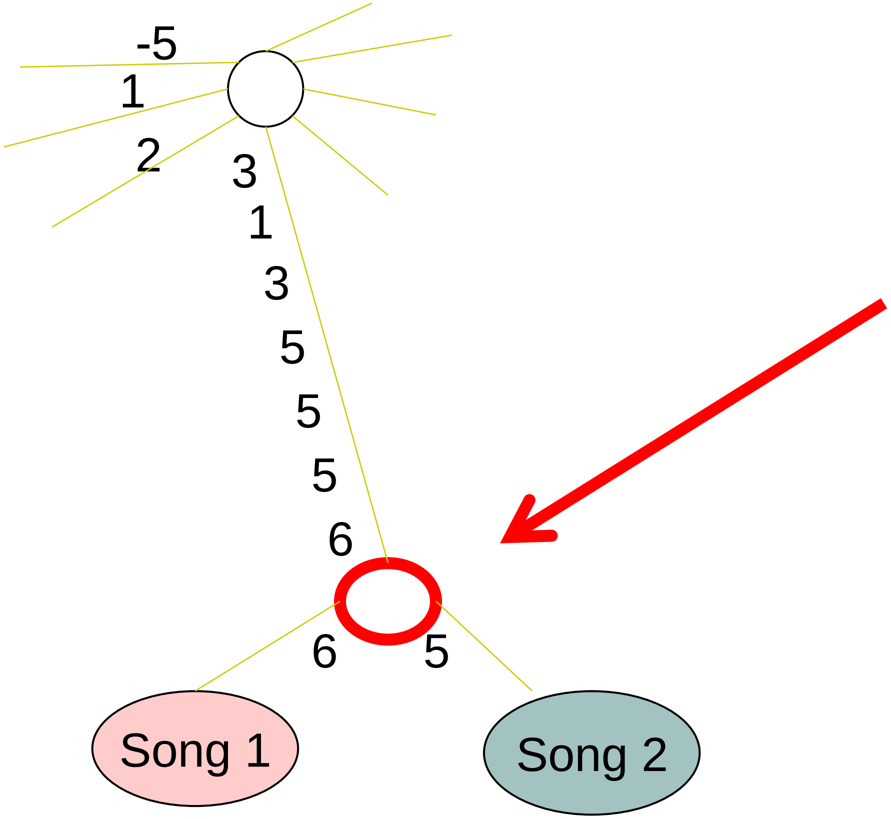
Song 2

F[KRIEGS]
[In Boehmen liegt ein staedtchen]
REG[Deutschland, Hessen, Marburg]
MEL[-5_ -5_.33_3_ 3__1_
  3_ 5_.55_6_ 5__0_
  5_ 7_.67_6_ 6__5_
  4_ 3_5_2_5_ 1__0_ //] >>
FCT[Staende -, Soldaten -, Kriegs – Lied]

2 folk songs from the Essen Associative Code (EsAC) database
http://www.esac-data.org/data/

…and plagiarism detection

Generalized suffix tree for two songs



The longest common substring

Set your imagination free☺