Global Model for Hierarchical Multi-Label Text Classification

MURAWAKI Yūgo

Kyoto University ⇨ Kyushu University

15 October 2013 IJCNLP 2013@Nagoya, Japan

Outline

- Task Settings
- Previous Work
- Proposed Method
- Results and Discussion
- Conclusion

Outline

- Task Settings
- Previous Work
- Proposed Method
- Results and Discussion
- Conclusion

Task: Multi-Label Text Classification

Input: Document	
	Journal of Wood Science
	Harvested wood products accounting in the post Kyoto commitment period
	An increase in the amount of harvested wood products (HWP) from sustainable forestry would help to reduce levels of atmospheric carbon. In the first commitment period of the Kyoto Protocol (2008-2012), this carbon stock effect of HWP is ignored, and forest harvesting is treated as an instantaneous emission of carbon dioxide. However, in the next commitment period of the United Nations Framework Convention on Climate Change from 2013, the carbon stock changes resulting from HWP will be taken into account in the national greenhouse gas inventories…

Task: Multi-Label Text Classification

Input:
Document

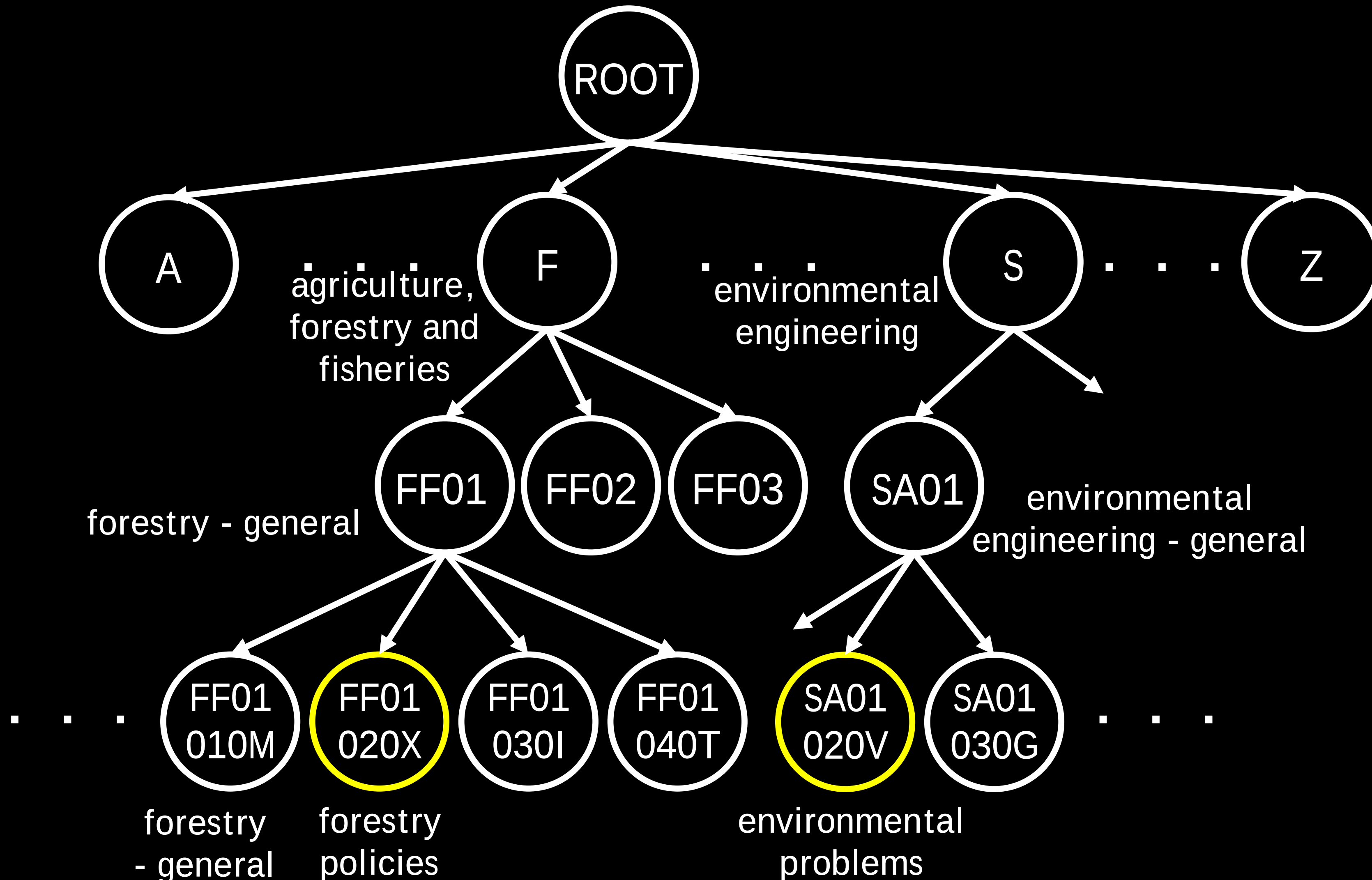
Journal of Wood Science
Harvested wood products accounting in the post Kyoto commitment period
An increase in the amount of harvested wood products (HWP) from sustainable forestry would help to reduce levels of atmospheric carbon. In the first commitment period of the Kyoto Protocol (2008-2012), this carbon stock effect of HWP is ignored, and forest harvesting is treated as an instantaneous emission of carbon dioxide. However, in the next commitment period of the United Nations Framework Convention on Climate Change from 2013, the carbon stock changes resulting from HWP will be taken into account in the national greenhouse gas inventories…

Output:
A set of labels

- FF01020X (forestry policies)
- SA01020V (environmental problems)

Task: Hierarchically Organized Labels

Task: Hierarchically Organized Labels



Outline

- Task Settings
- Previous Work
- Proposed Method
- Results and Discussion
- Conclusion

Question: What is the Hierarchy for Text Classification?

Question: What is the Hierarchy for Text Classification?

1. Flat: Nothing. Just ignore the hierarchy.

Question: What is the Hierarchy for Text Classification?

1. **Flat**: Nothing. Just ignore the hierarchy.
2. **Pruning**: Top-down pruning for speed, often in exchange for accuracy

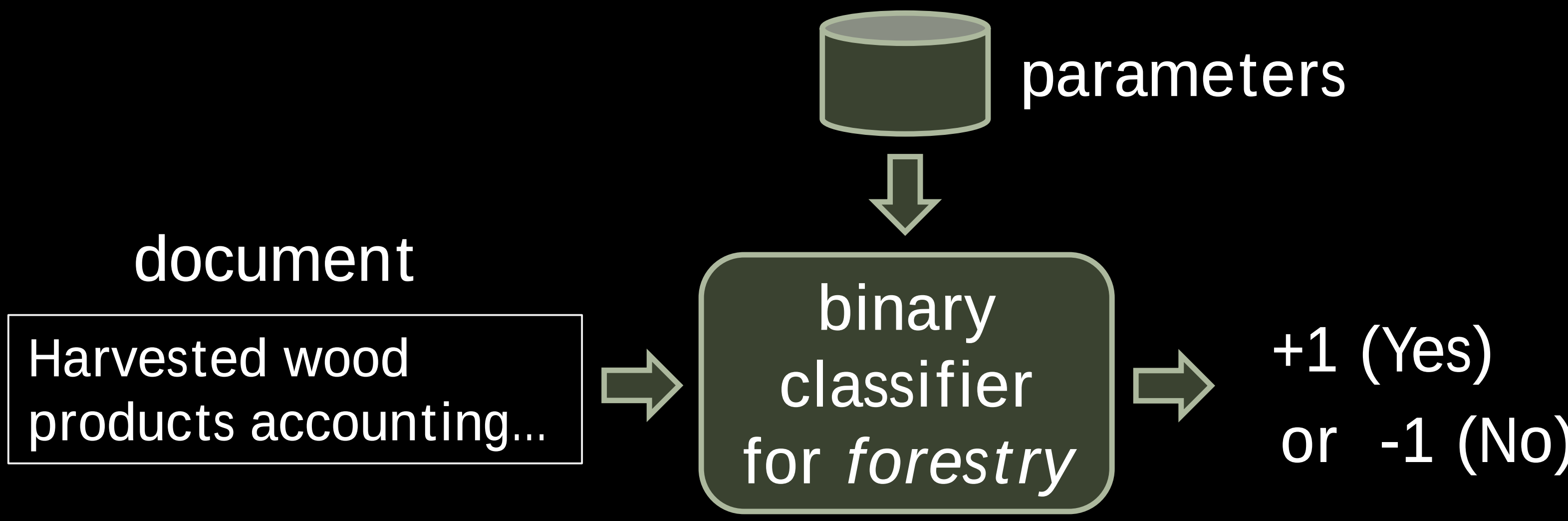
Question: What is the Hierarchy for Text Classification?

1. **Flat**: Nothing. Just ignore the hierarchy.
2. **Pruning**: Top-down pruning for speed, often in exchange for accuracy
3. **Parameter Sharing**: Share parameters according to the hierarchy to improve accuracy

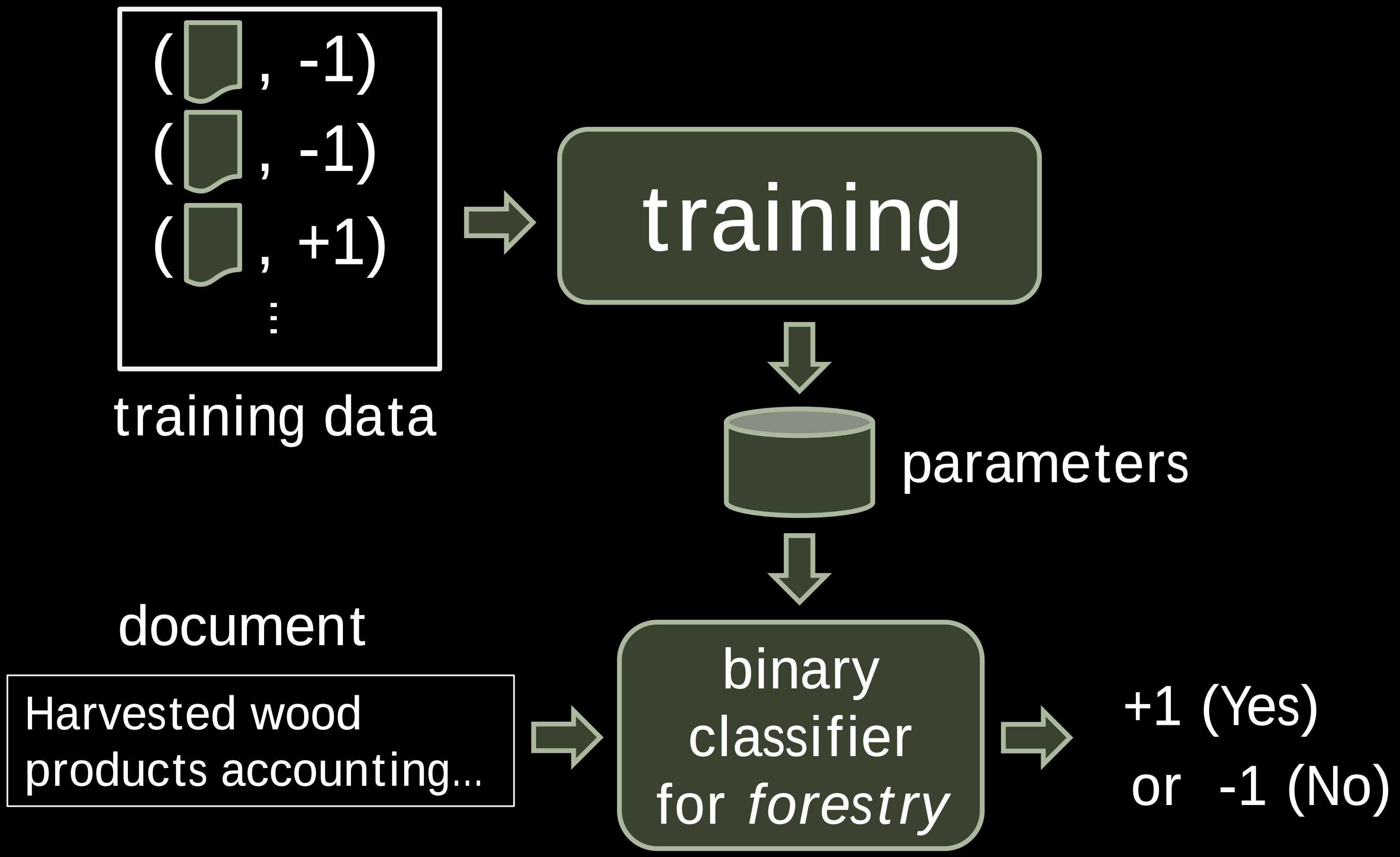
Question: What is the Hierarchy for Text Classification?

1. **Flat**: Nothing. Just ignore the hierarchy.
2. **Pruning**: Top-down pruning for speed, often in exchange for accuracy
3. **Parameter Sharing**: Share parameters according to the hierarchy to improve accuracy
4. **Proposed Method**: To capture dependencies among multiple labels to be output

Building Block: Binary Linear Classifier 1/2



Building Block: Binary Linear Classifier 1/2



Building Block: Binary Linear Classifier 1/2

document

Harvested wood products accounting...

Building Block: Binary Linear Classifier 1/2

feature vector

$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix}$$

document

Harvested wood products accounting...

Building Block: Binary Linear Classifier 1/2

document

Harvested wood
products accounting...

feature vector

$$\begin{array}{cl} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix} & \begin{array}{l} \text{algorithm} \\ \text{dolphin} \\ \text{wood} \\ \text{products} \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \end{array}$$

Building Block: Binary Linear Classifier 1/2

document

Harvested wood products accounting...

weight vector

(-0.5, -0.2, 0.8, ...)

algorighm dolphin wood

feature vector

$$
\begin{array}{cl}
\left[\begin{array}{c} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \end{array}\right] &
\begin{array}{l} \text{algorithm} \\ \text{dolphin} \\ \text{wood} \\ \text{products} \\ \\ \\ \\ \\ \\ \\ \\ \end{array}
\end{array}
$$

Building Block: Binary Linear Classifier 1/2

document

Harvested wood
products accounting...

weight vector

(-0.5, -0.2, 0.8, ...)

algorighm dolphin wood

feature vector

$$\begin{array}{cl} \left[\begin{array}{c} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \end{array}\right] & \begin{array}{l} \text{algorithm} \\ \text{dolphin} \\ \text{wood} \\ \text{products} \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \end{array}$$

>0 ? then +1
otherwise -1

Building Block: Binary Linear Classifier 1/2

This is the parameters to be learned

document

Harvested wood products accounting...

weight vector

(-0.5, -0.2, 0.8, ...)

algorighm dolphin wood

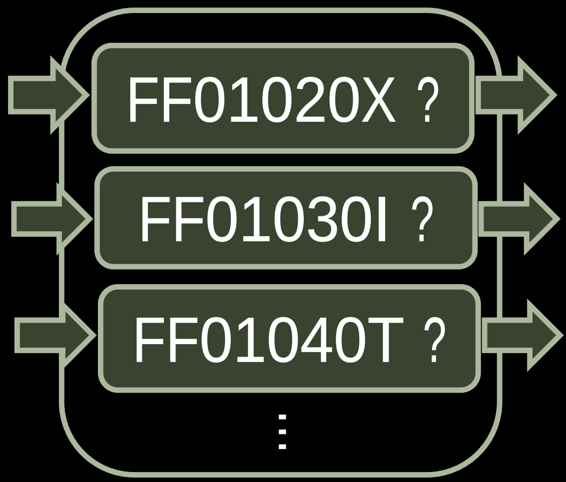
feature vector

$$\begin{array}{cl} \left[\begin{array}{c} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \end{array}\right] & \begin{array}{l} \text{algorithm} \\ \text{dolphin} \\ \text{wood} \\ \text{products} \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \end{array}$$

>0 ? then +1

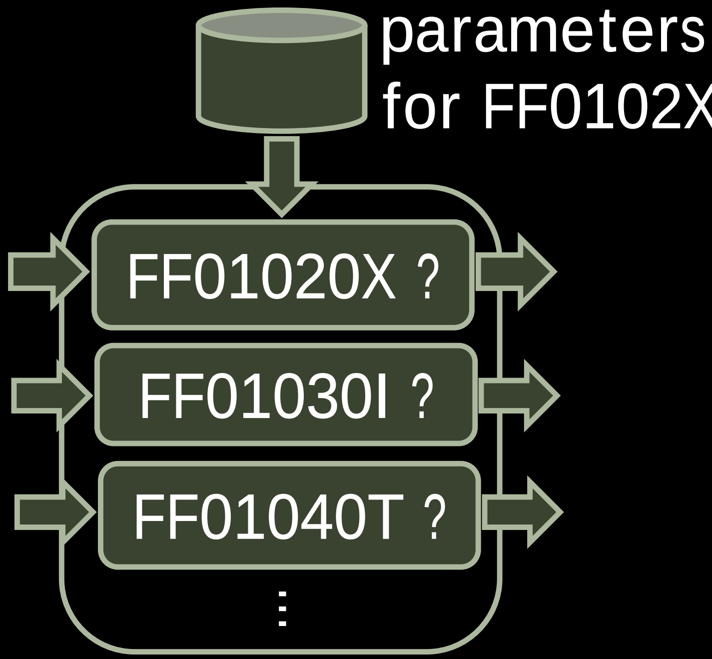
otherwise -1

Flat Model (ignores label hierarchy)



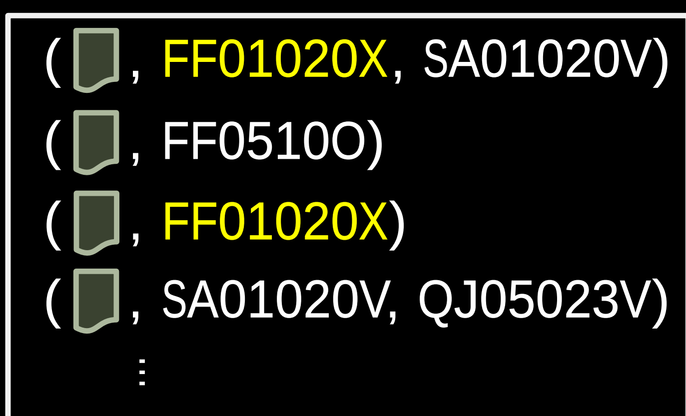
set of binary classifiers

Flat Model (ignores label hierarchy)

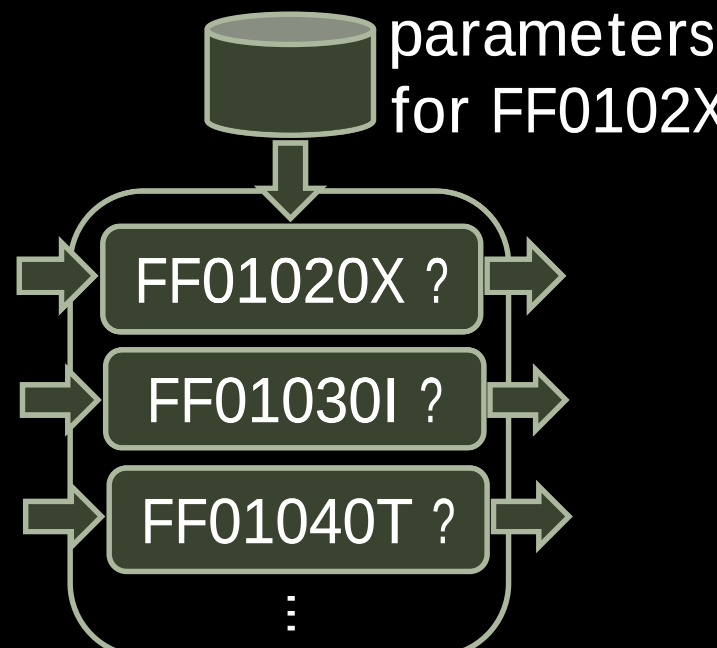


set of binary classifiers

Flat Model (ignores label hierarchy)

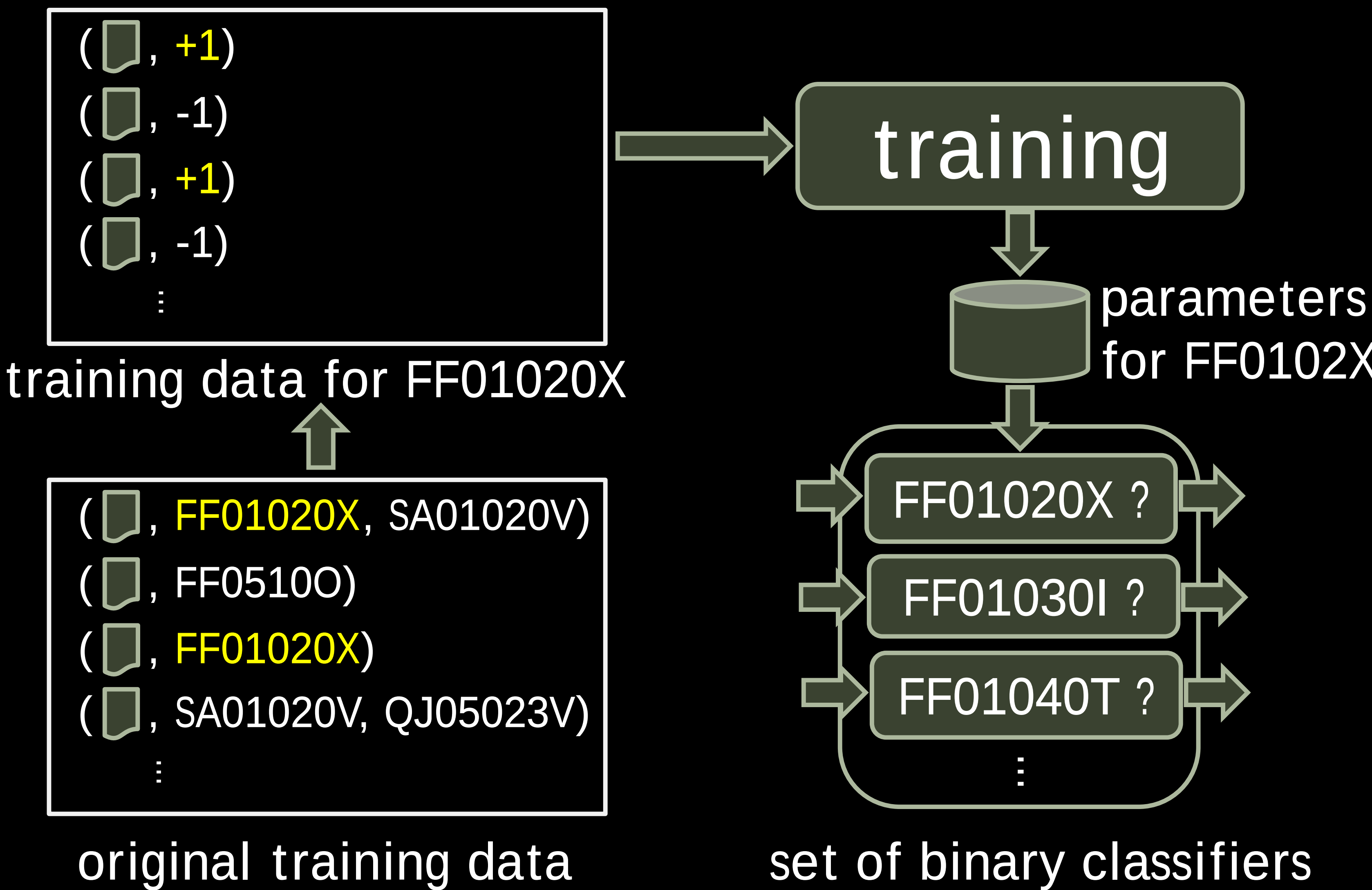


original training data

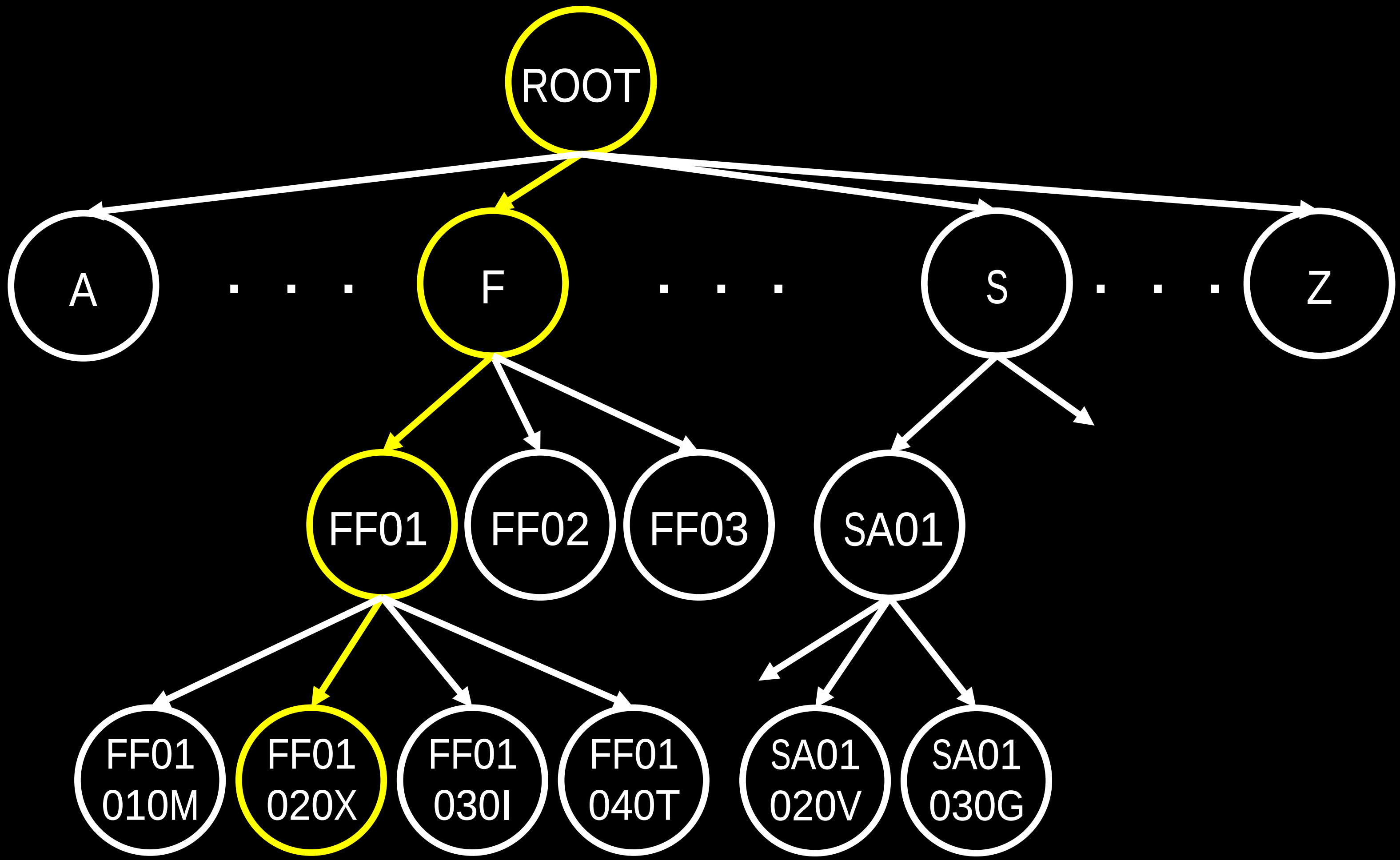


set of binary classifiers

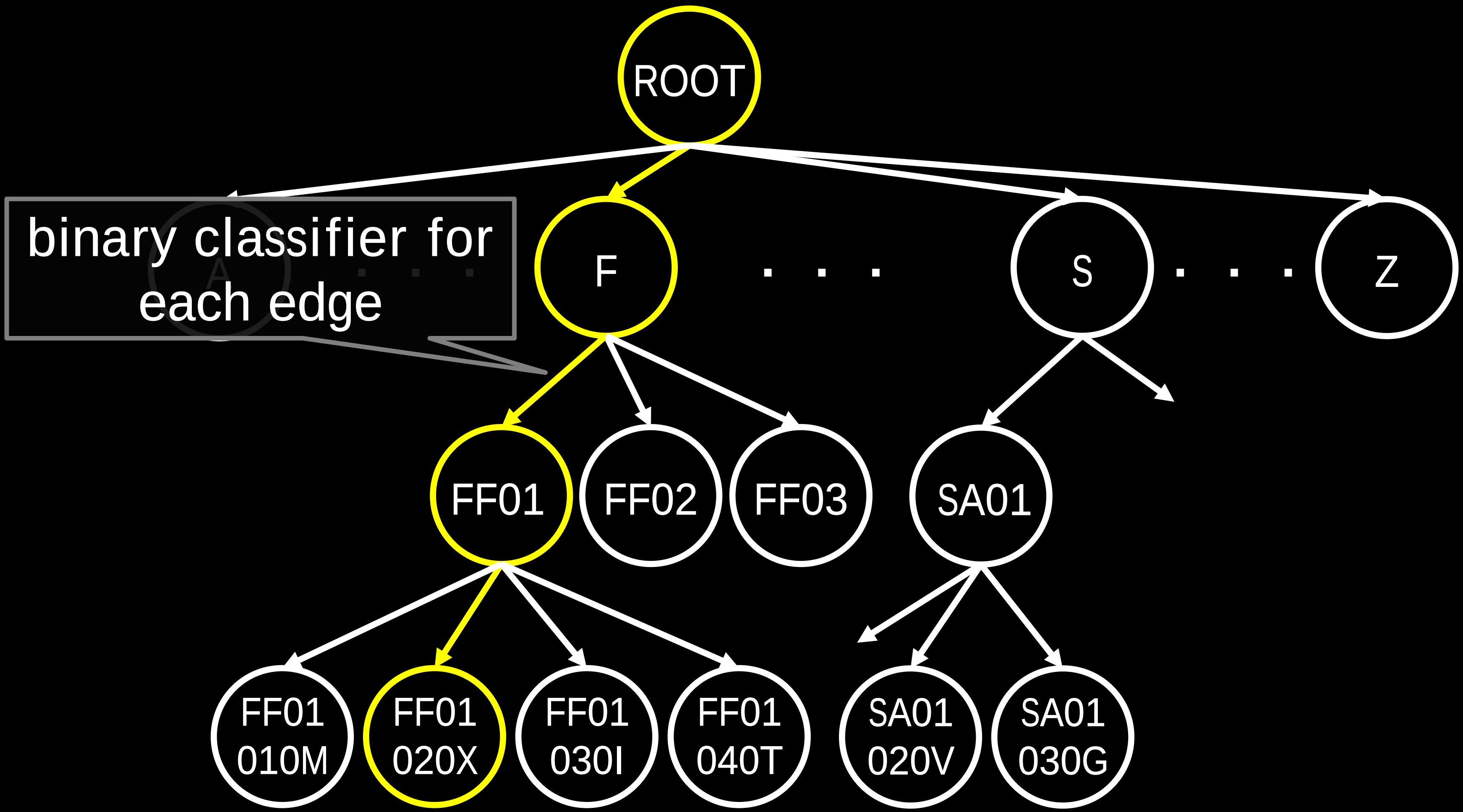
Flat Model (ignores label hierarchy)



Tree Models 1/2



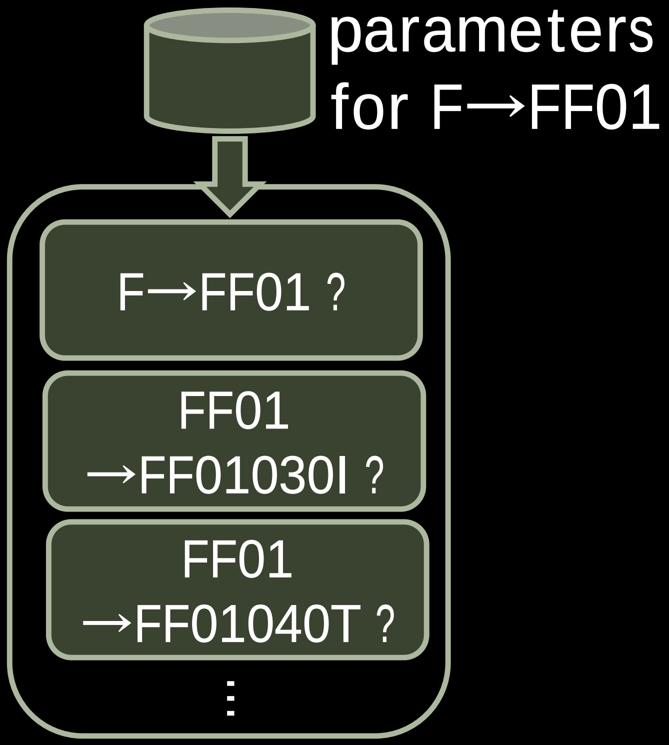
Tree Models 1/2



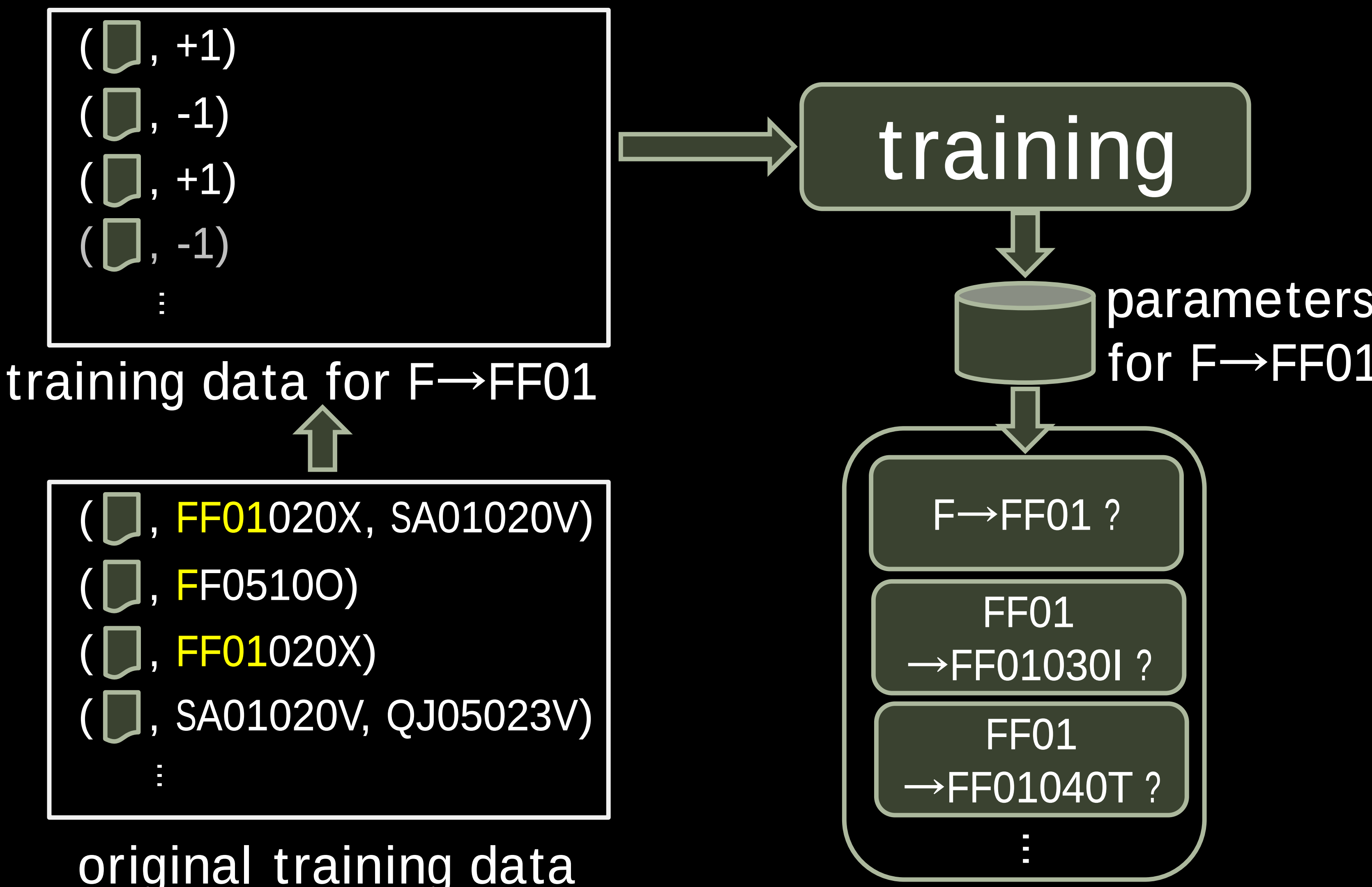
Tree Models 2/2

(, FF01020X, SA01020V)
(, FF0510O)
(, FF01020X)
(, SA01020V, QJ05023V)
⋮

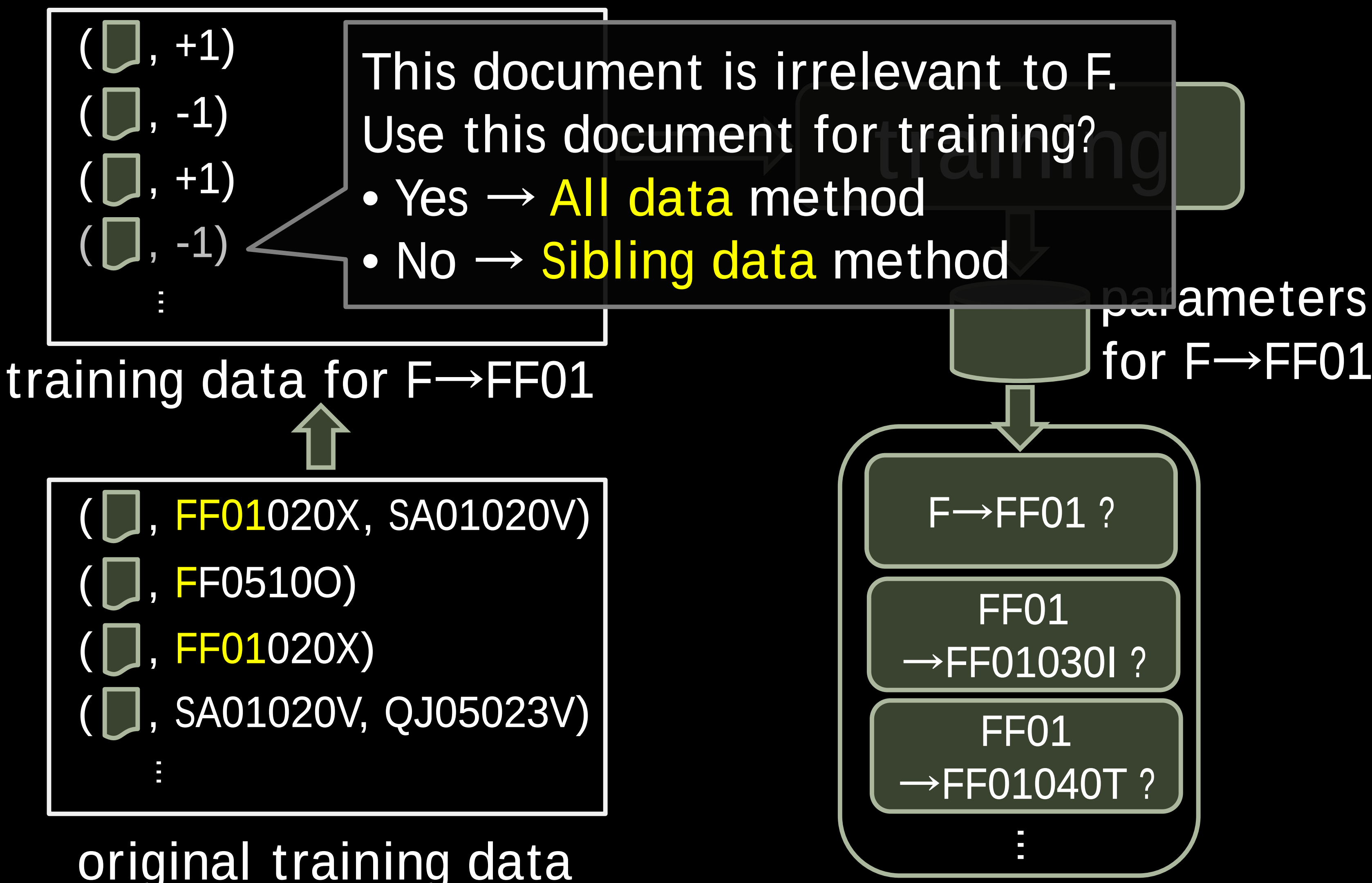
original training data



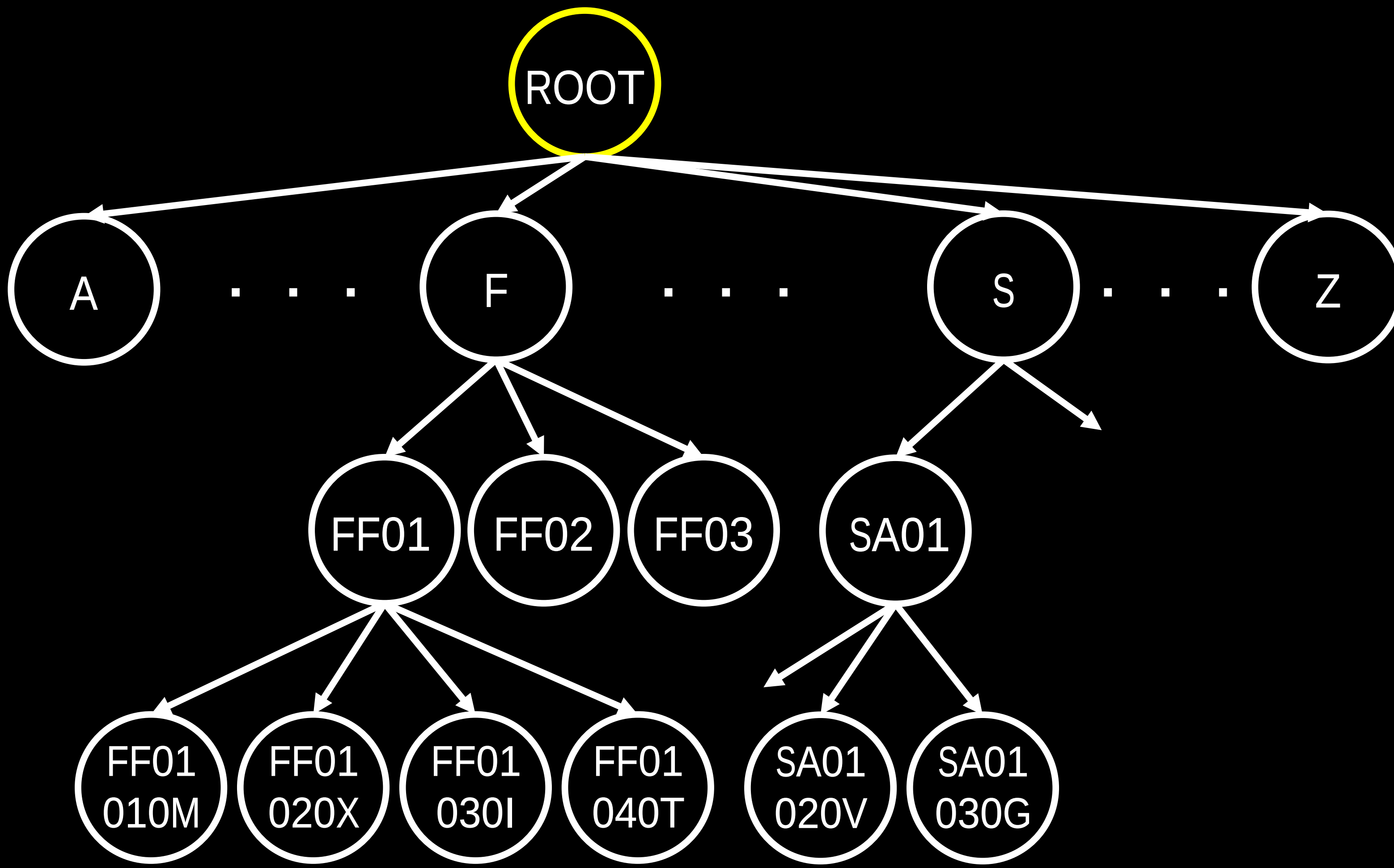
Tree Models 2/2



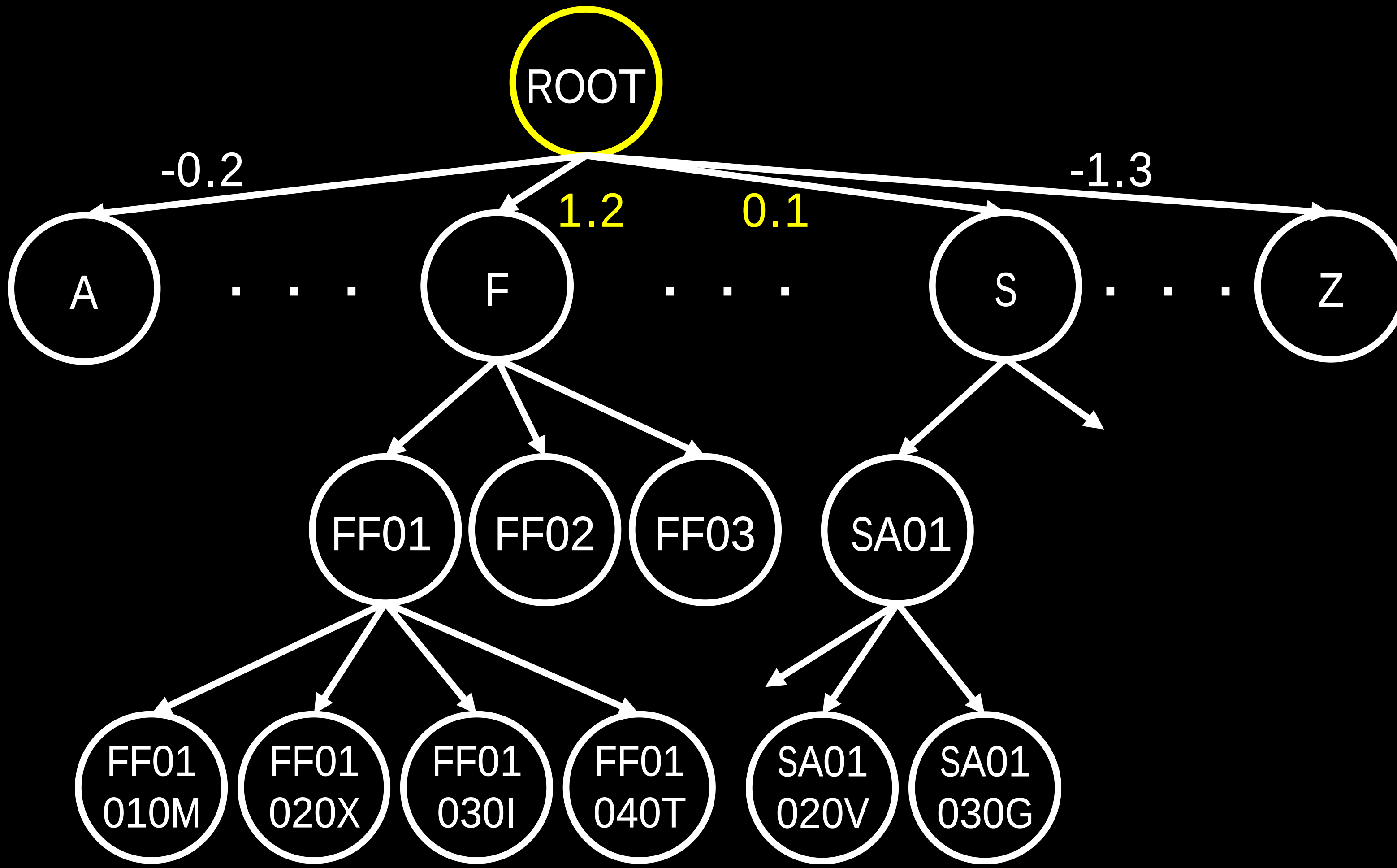
Tree Models 2/2



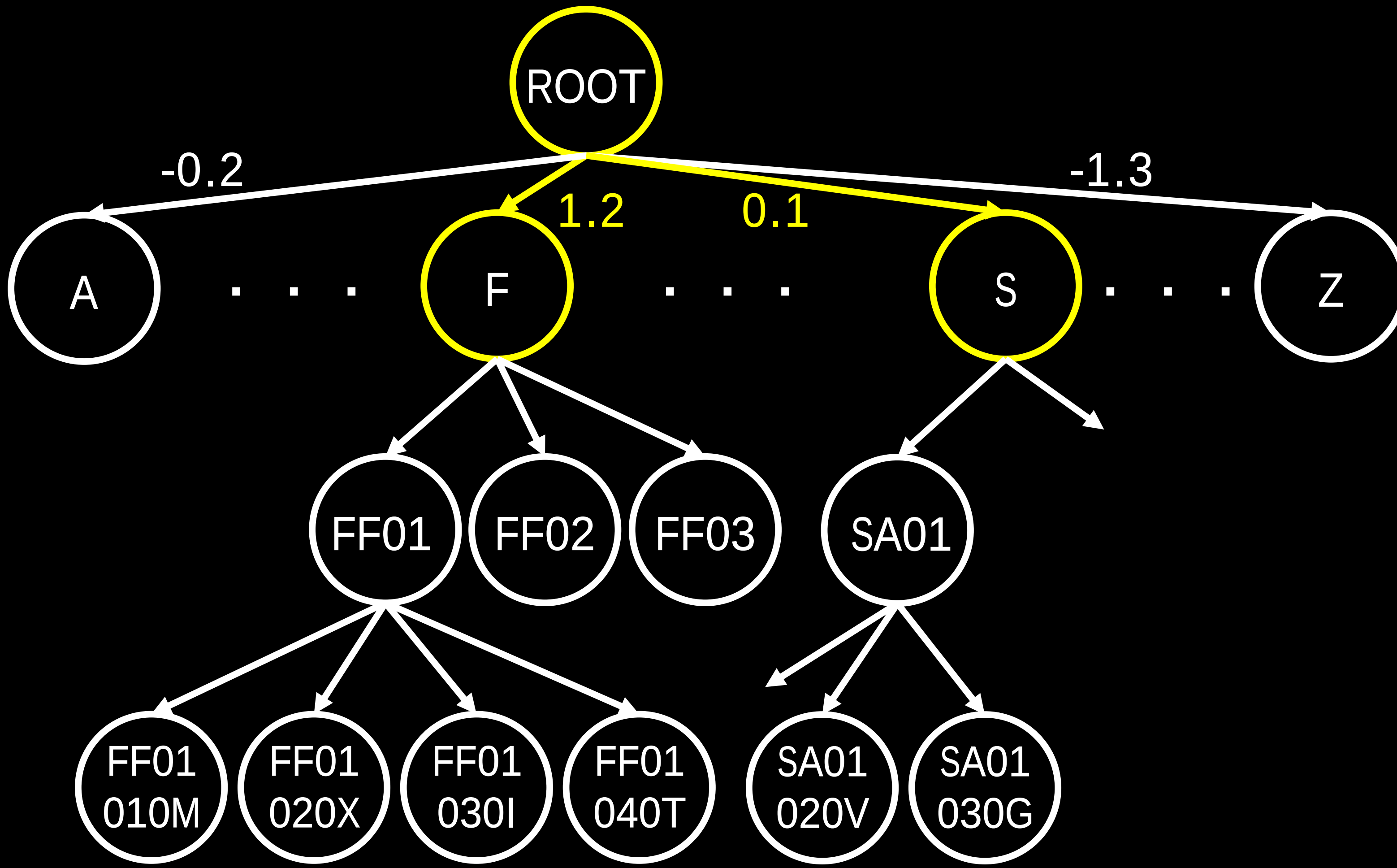
Top-down Pruning



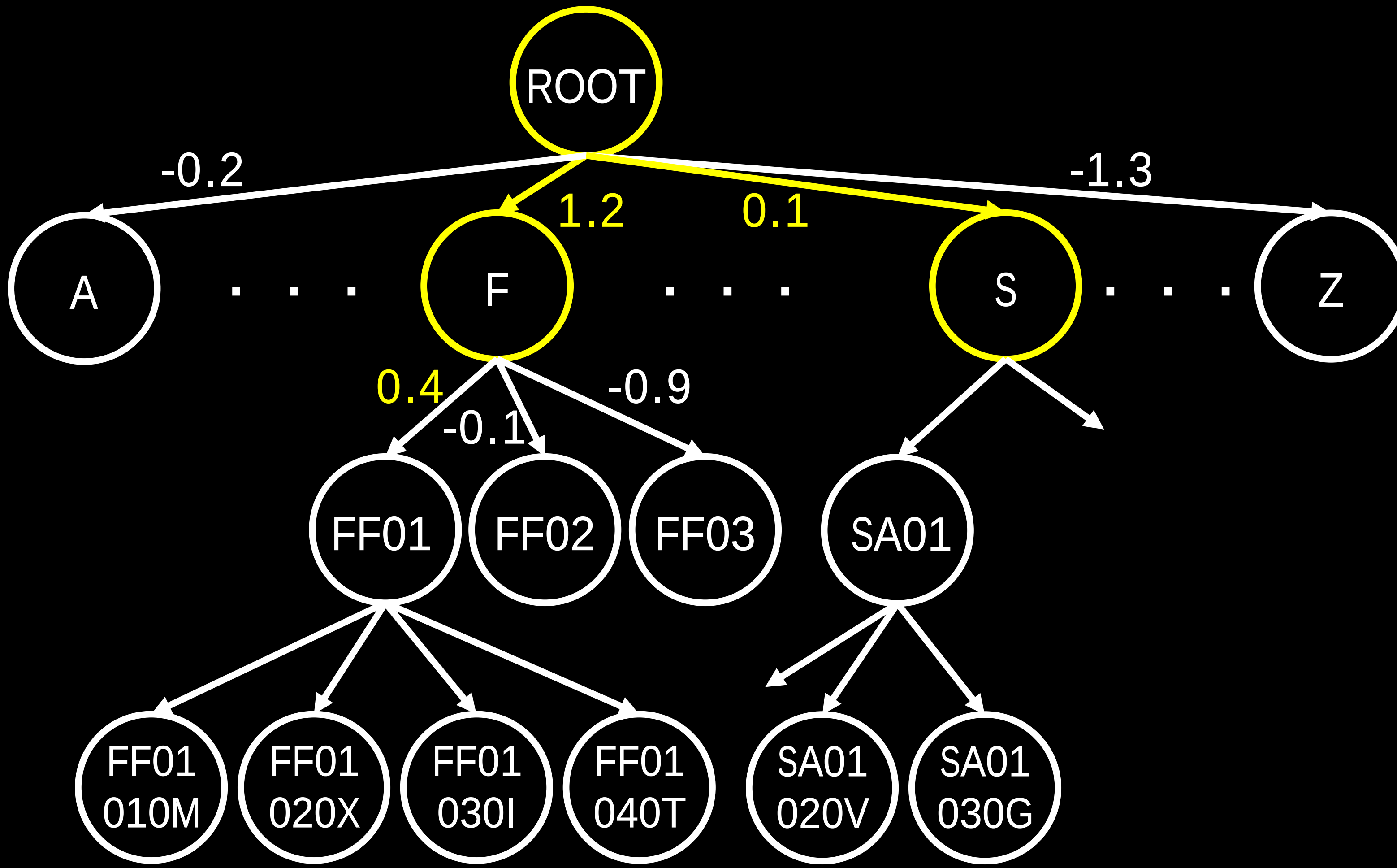
Top-down Pruning



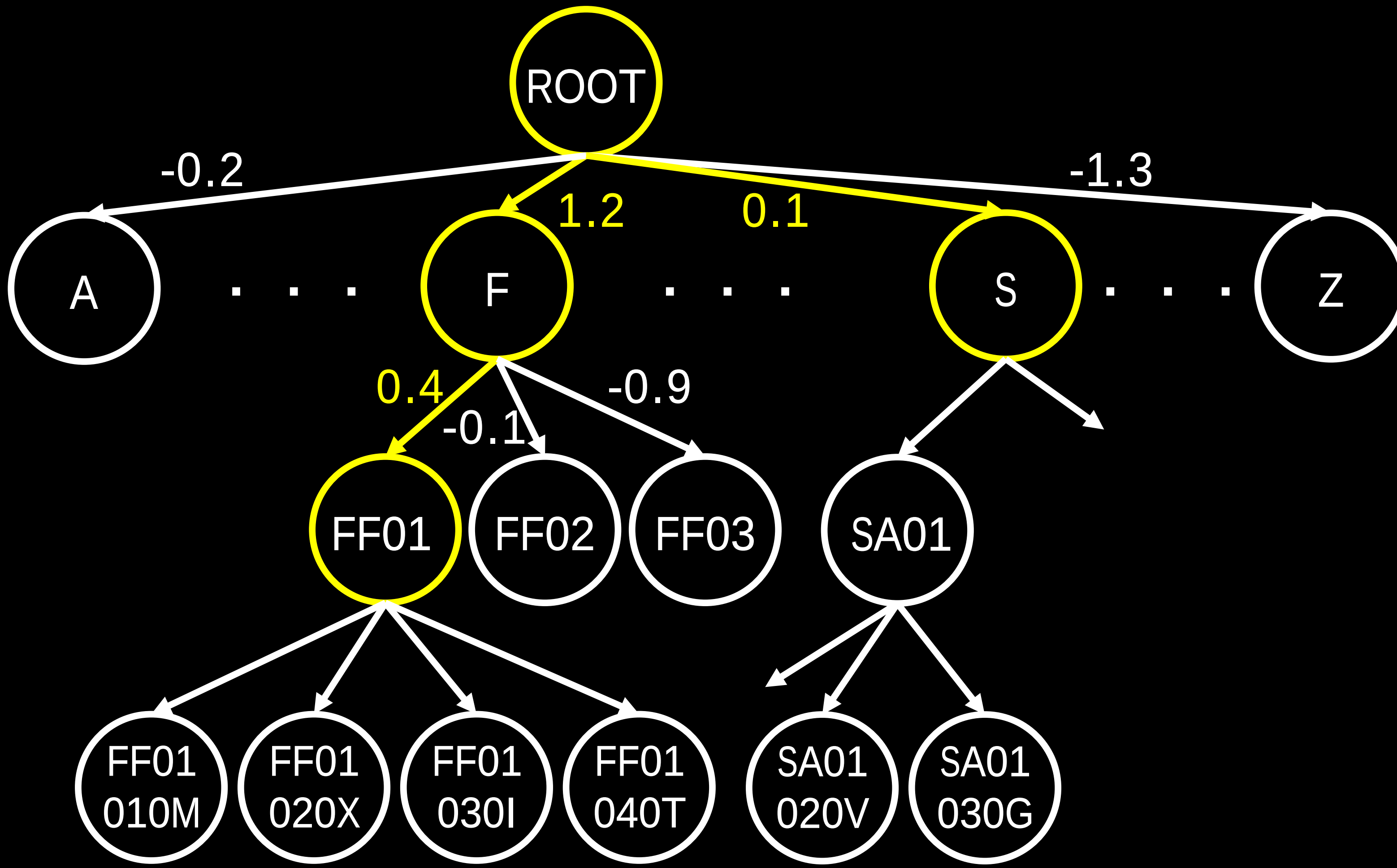
Top-down Pruning



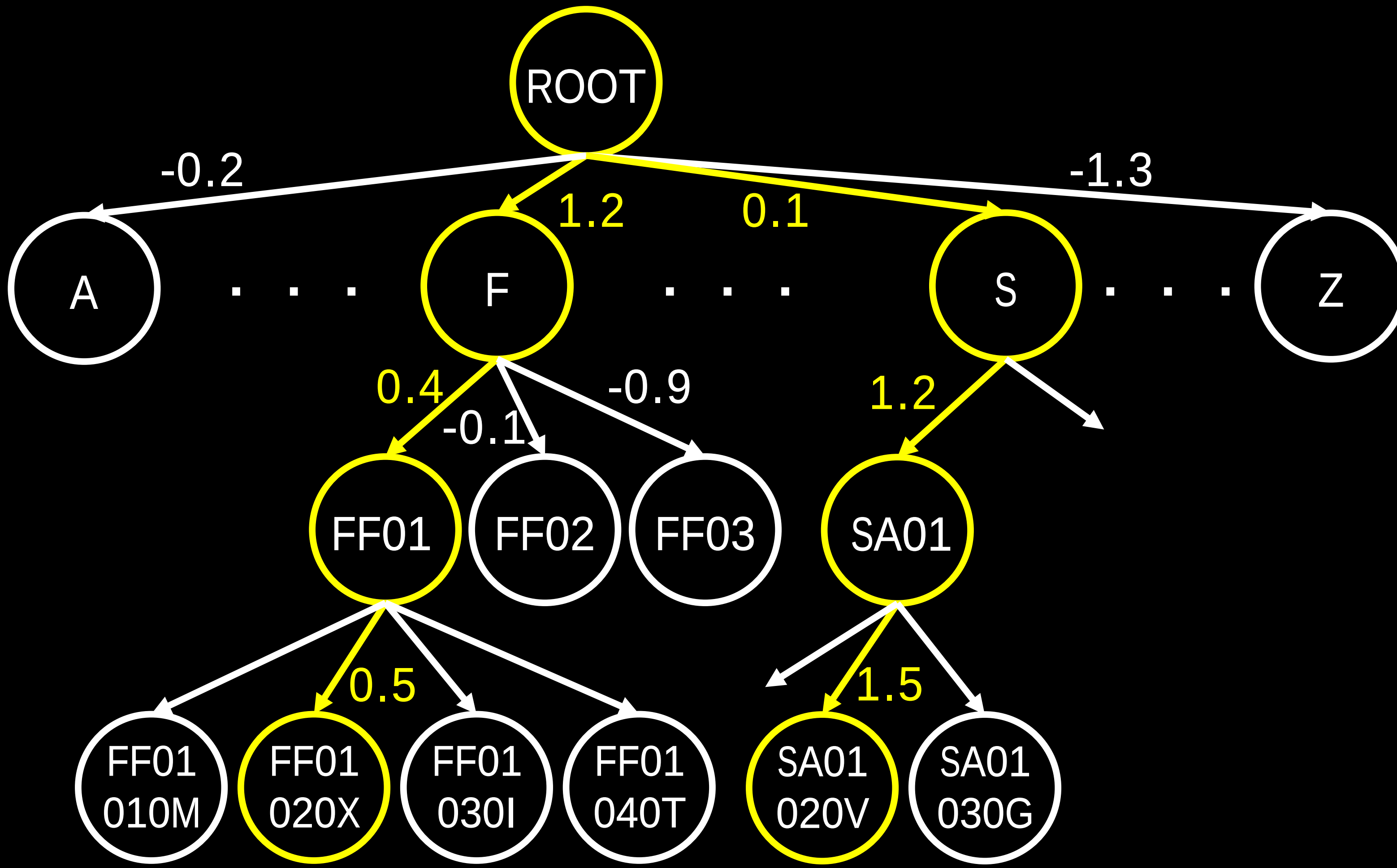
Top-down Pruning



Top-down Pruning



Top-down Pruning



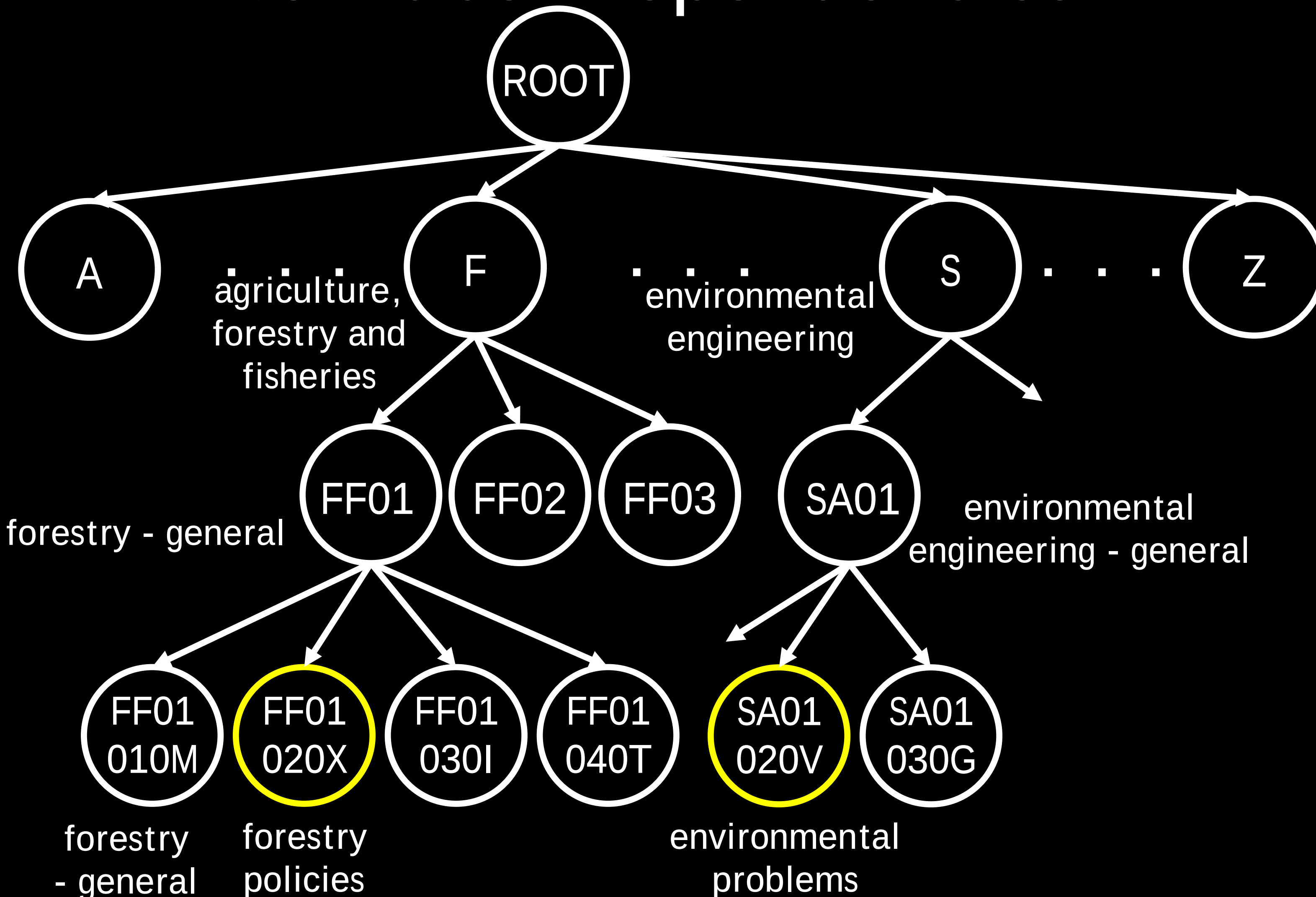
Outline

- Task Settings
- Previous Work
- Proposed Method
- Results and Discussion
- Conclusion

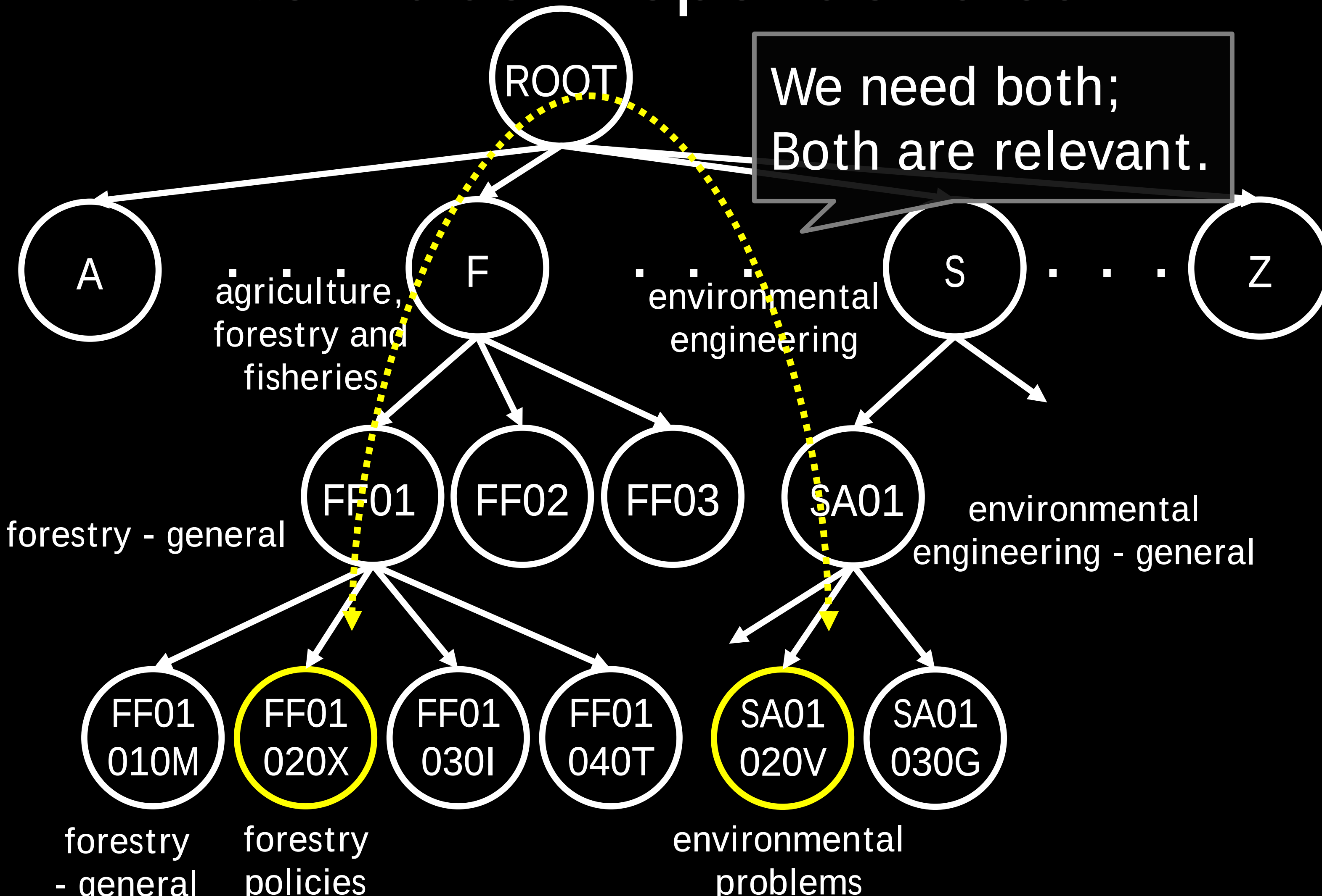
Proposed Method: Hierarchy for Multiple Labels

- Need to choose not just one best
- How many labels should we choose?
  A difficult question even for humans
- Our assumption: Human annotators do not select each label independently but consider the relative importance among competing labels by consulting the hierarchy

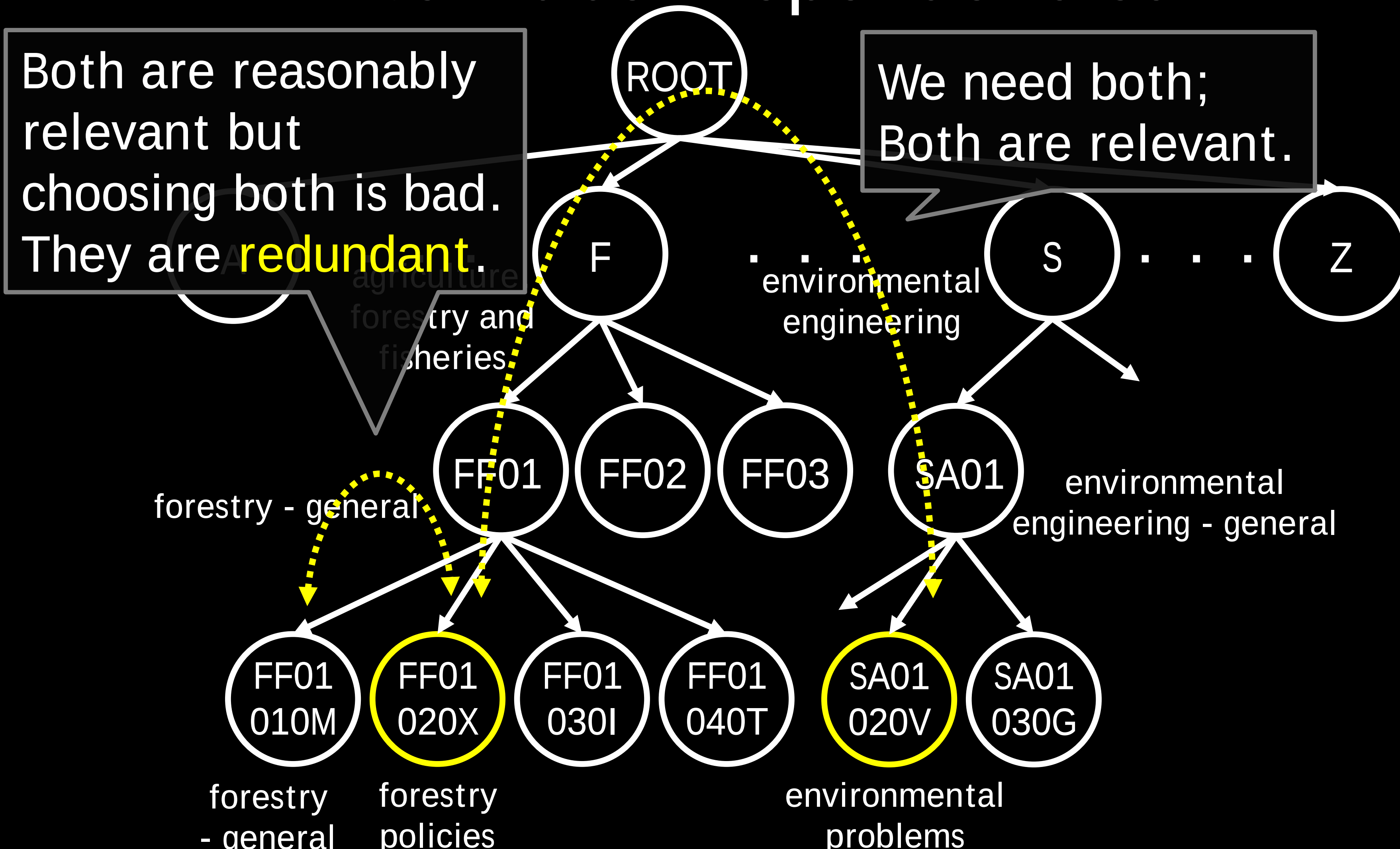
Intuition: Hierarchy Encodes Inter-label Dependencies



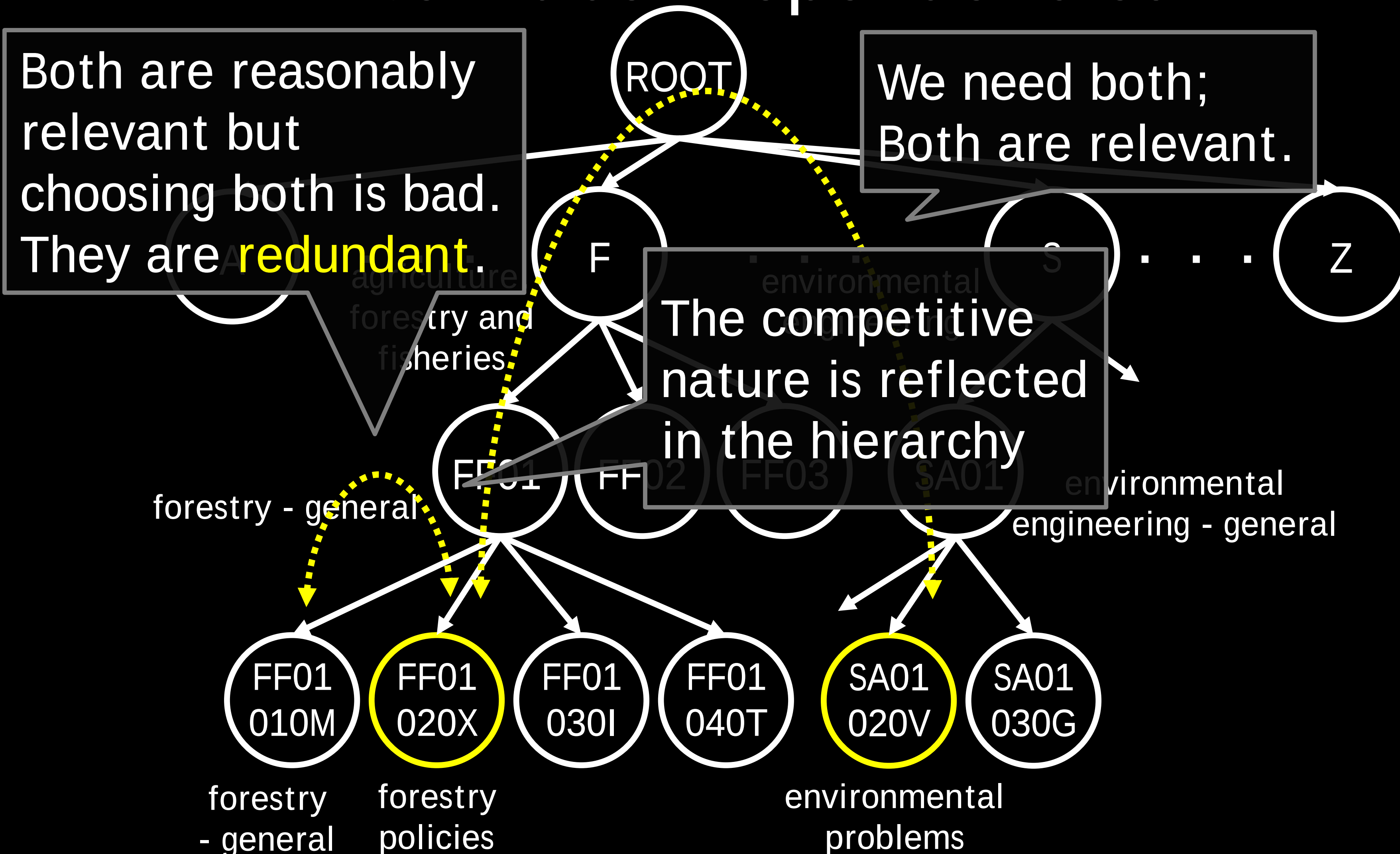
Intuition: Hierarchy Encodes Inter-label Dependencies



Intuition: Hierarchy Encodes Inter-label Dependencies



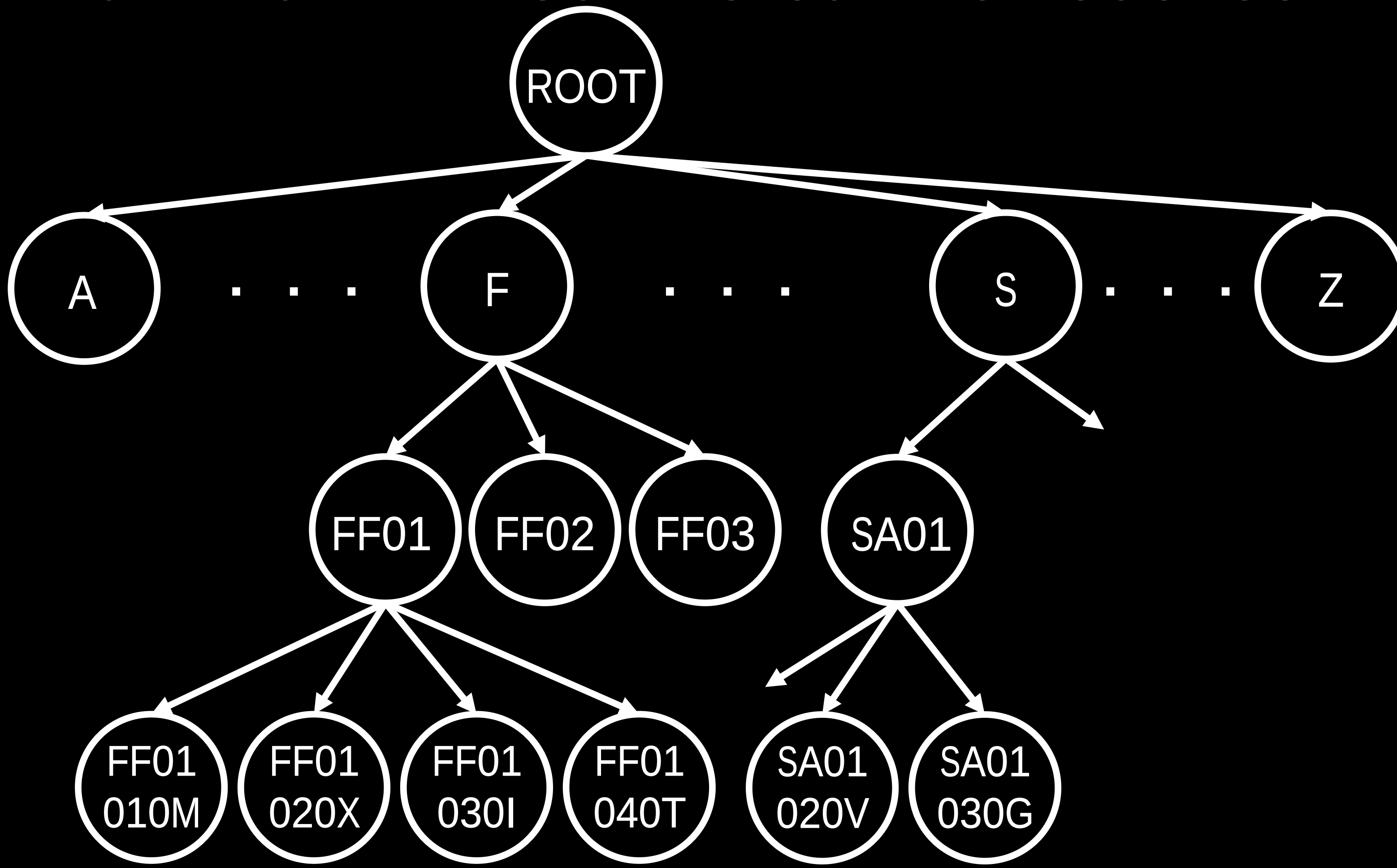
Intuition: Hierarchy Encodes Inter-label Dependencies



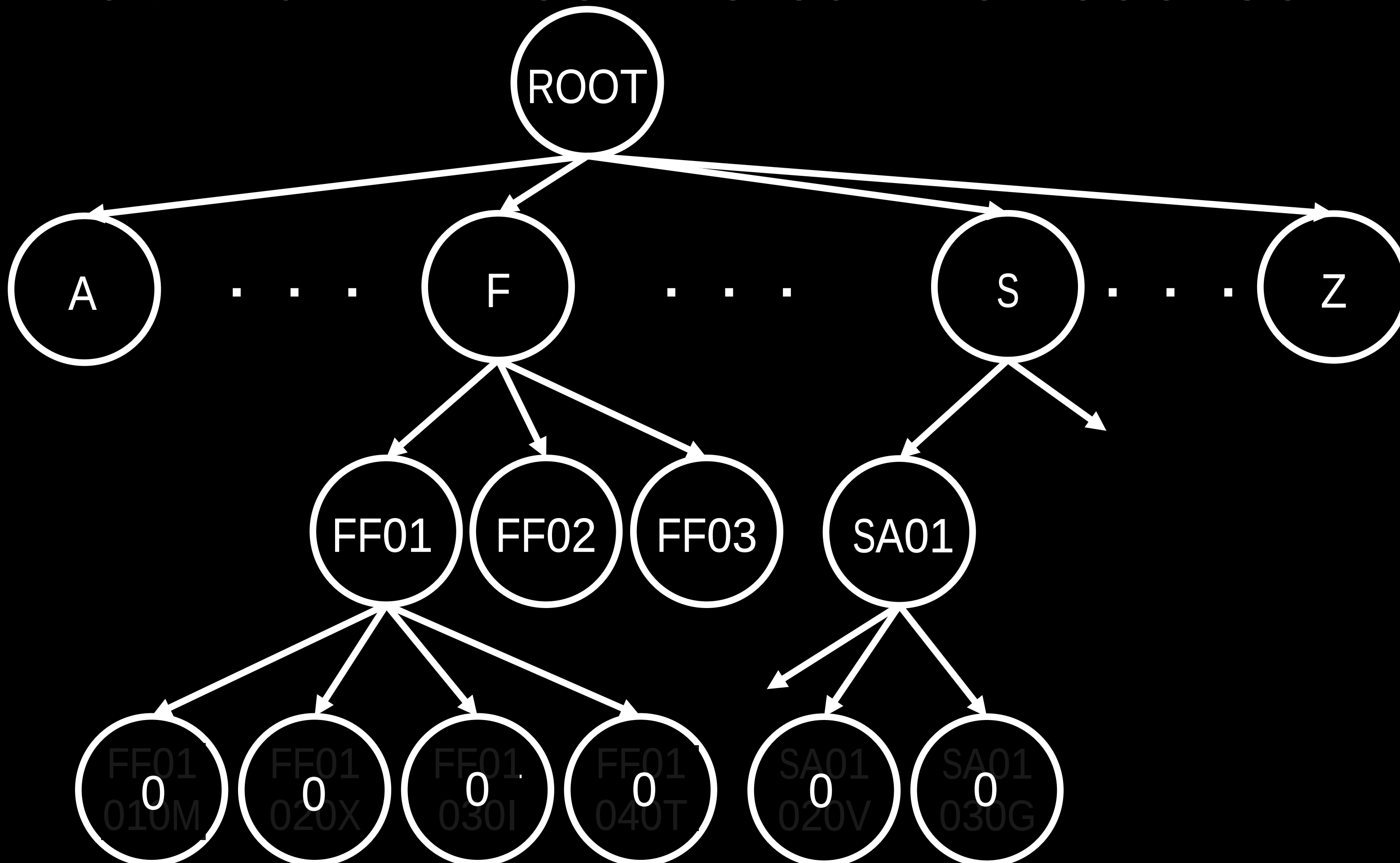
Preparation: Global Model

- Want to capture dependencies among multiple labels to be output
- First need to jointly predicts multiple labels
- Then capture inter-label dependencies by adding features to the global model

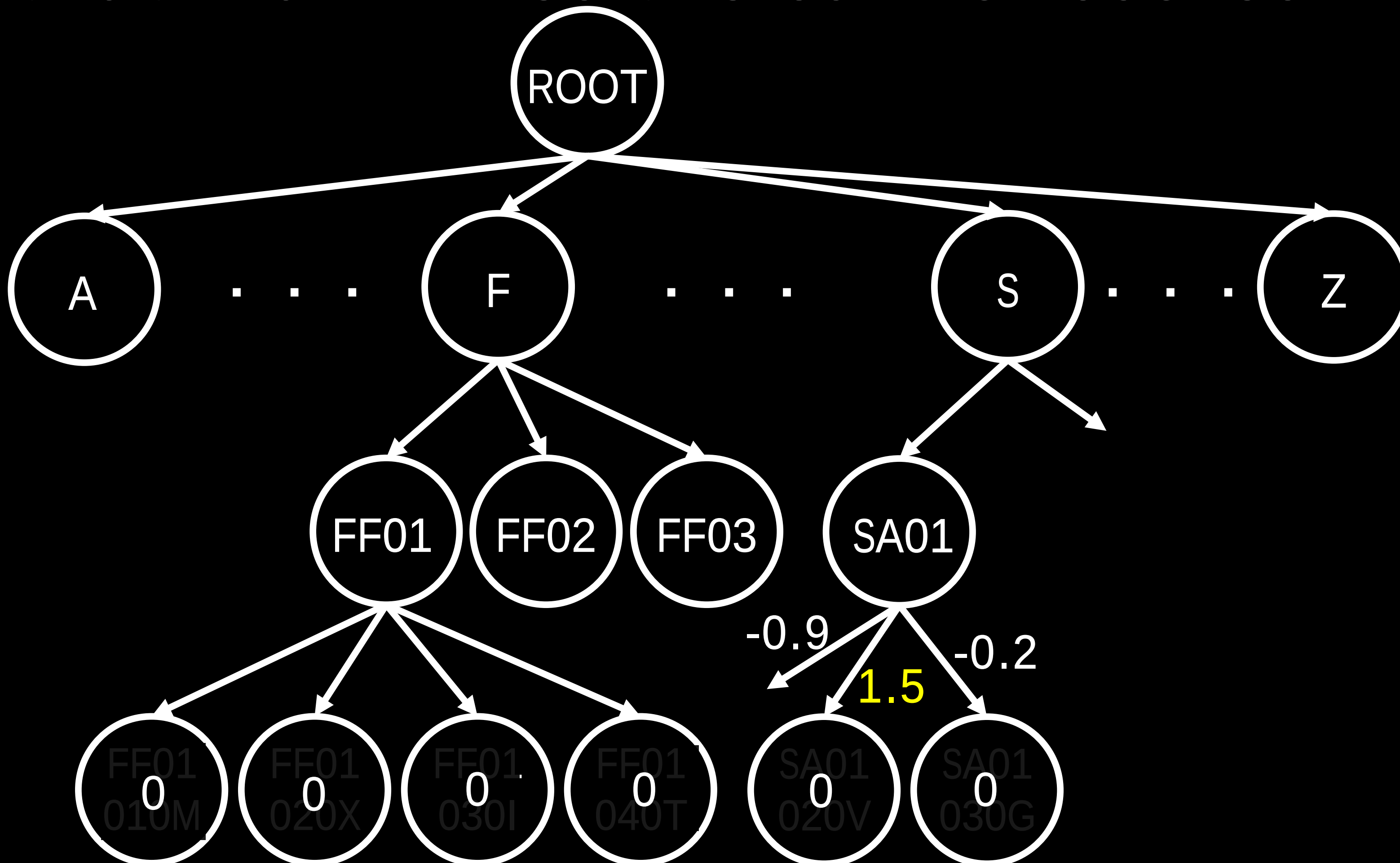
Global Model: Find a Subtree that maximizes the sum of scores



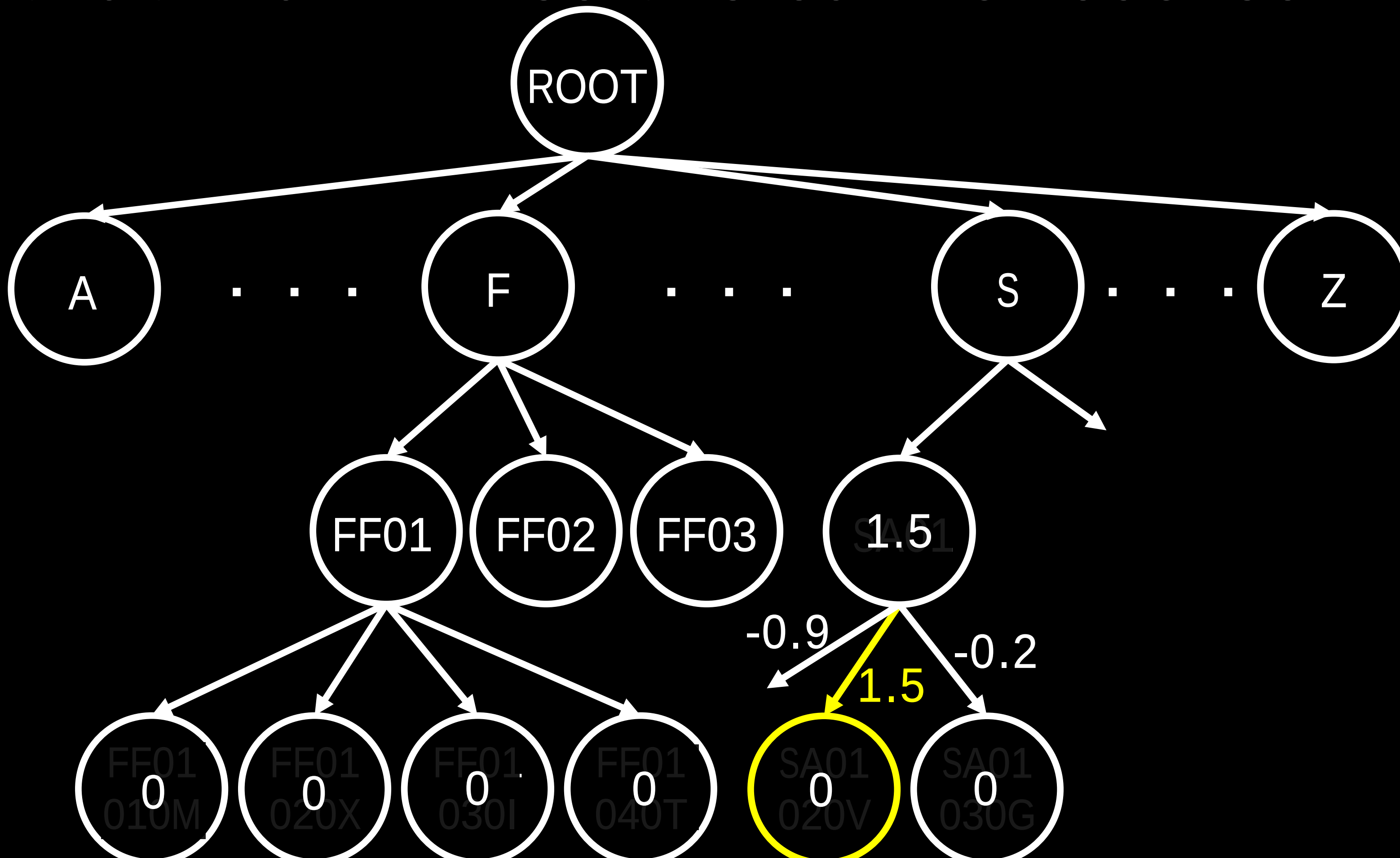
Global Model: Find a Subtree that maximizes the sum of scores



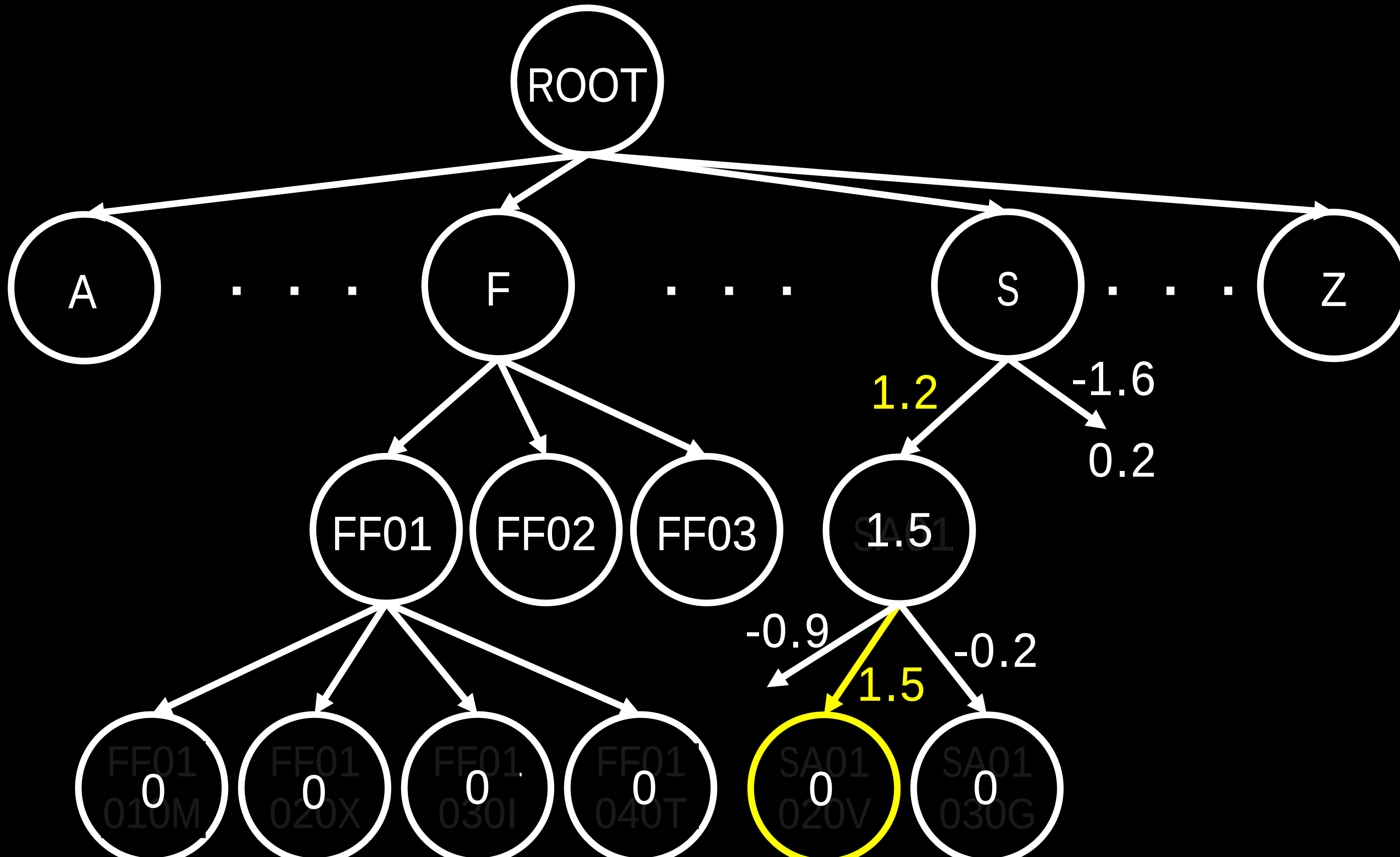
Global Model: Find a Subtree that maximizes the sum of scores



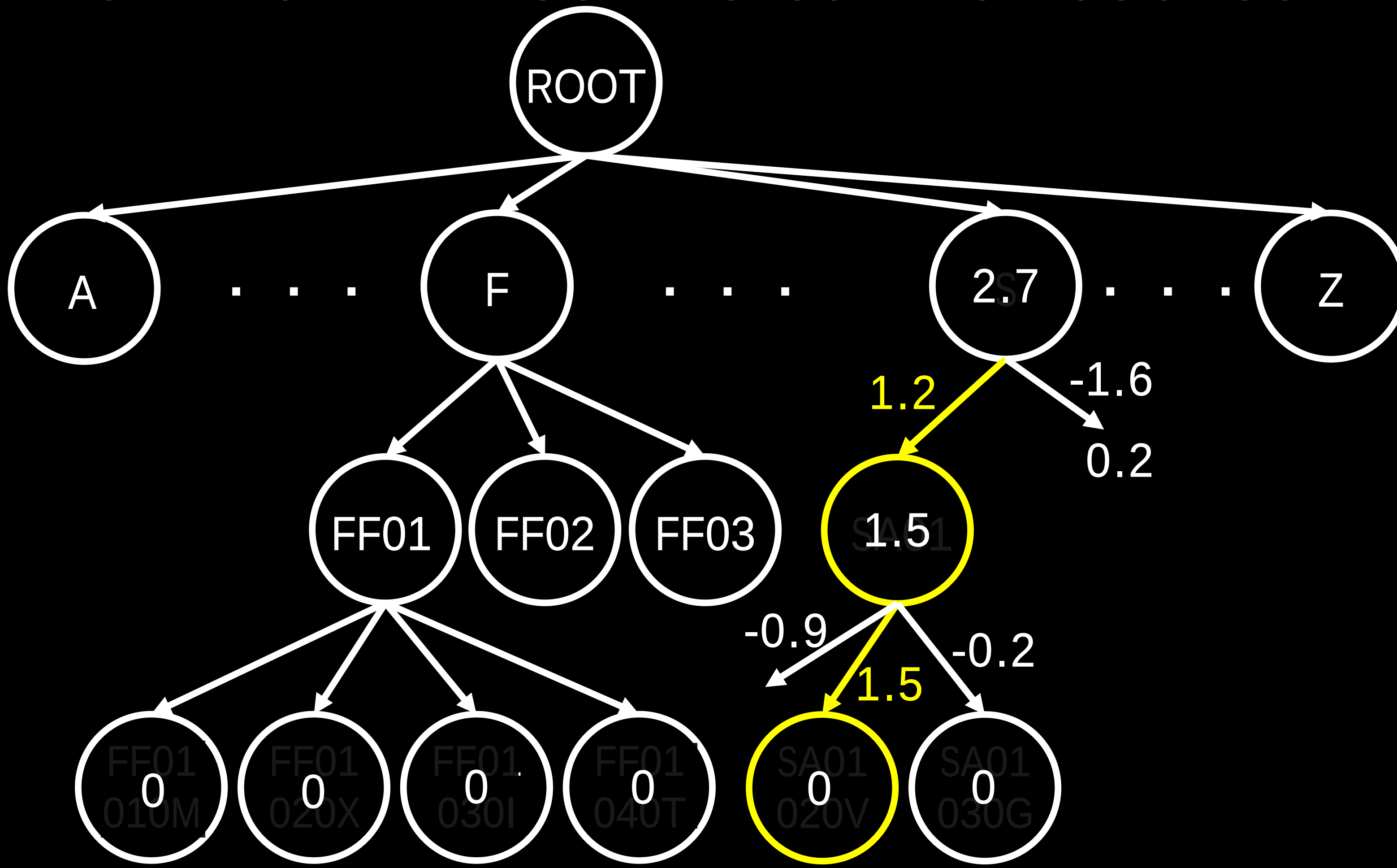
Global Model: Find a Subtree that maximizes the sum of scores



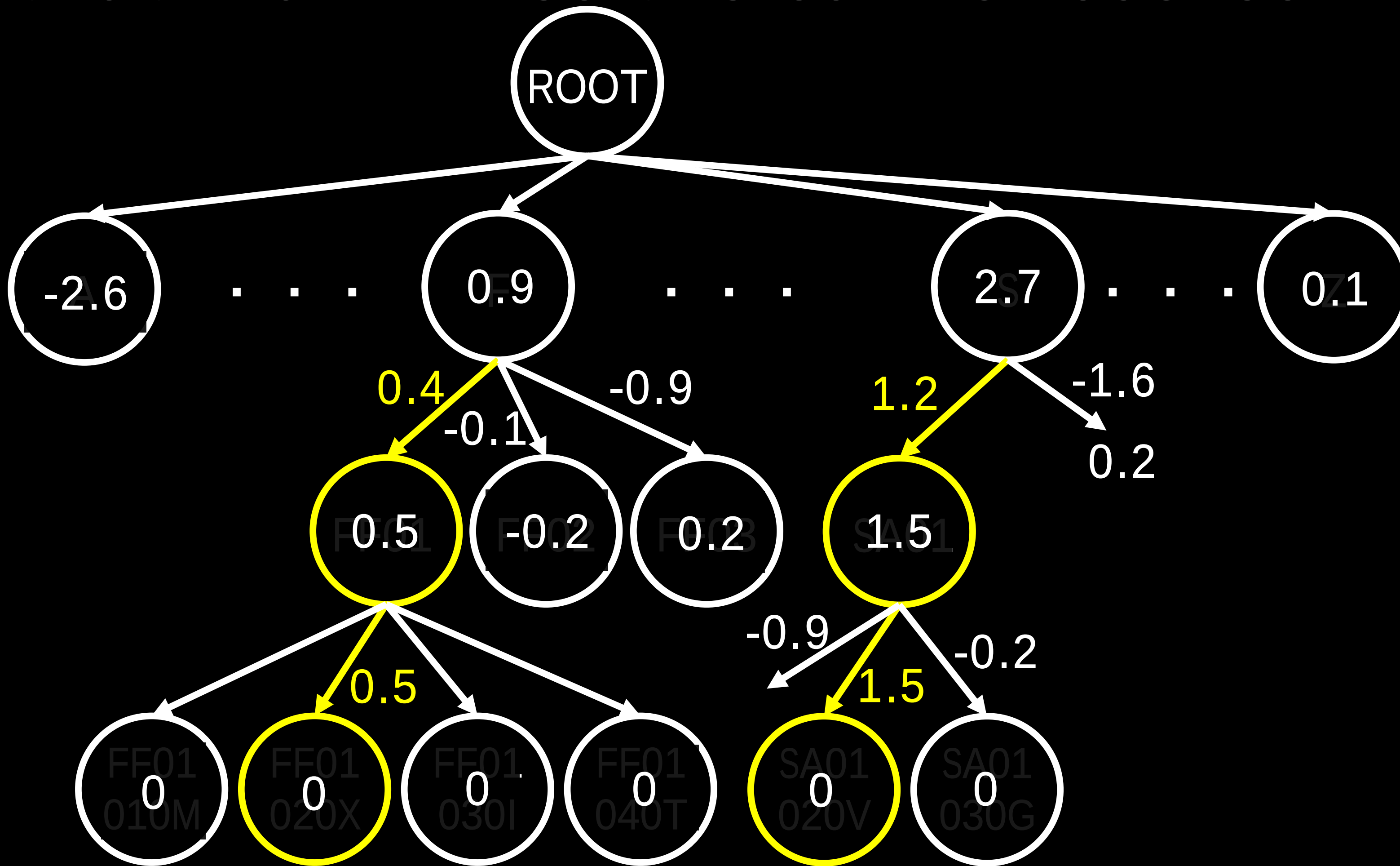
Global Model: Find a Subtree that maximizes the sum of scores



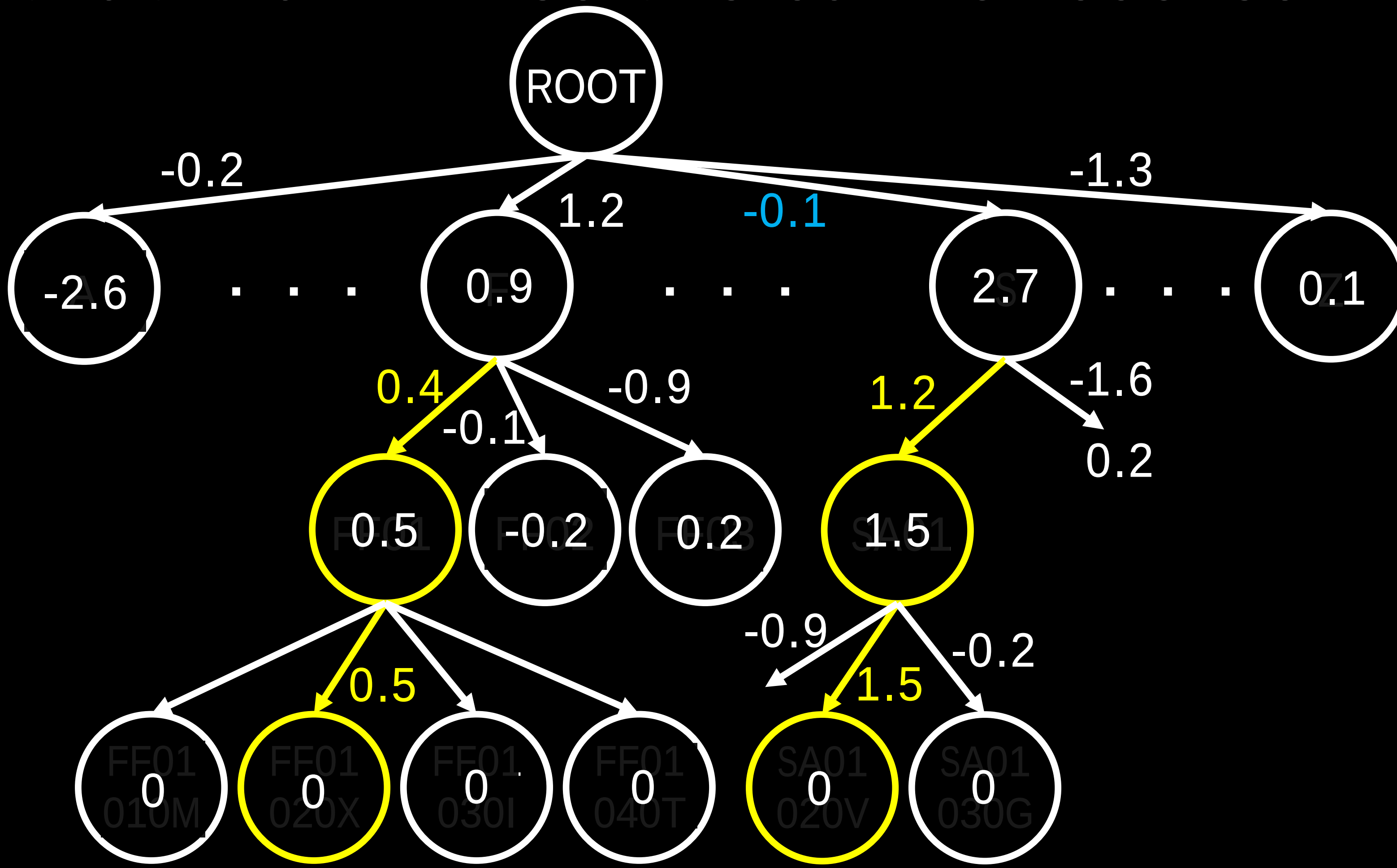
Global Model: Find a Subtree that maximizes the sum of scores



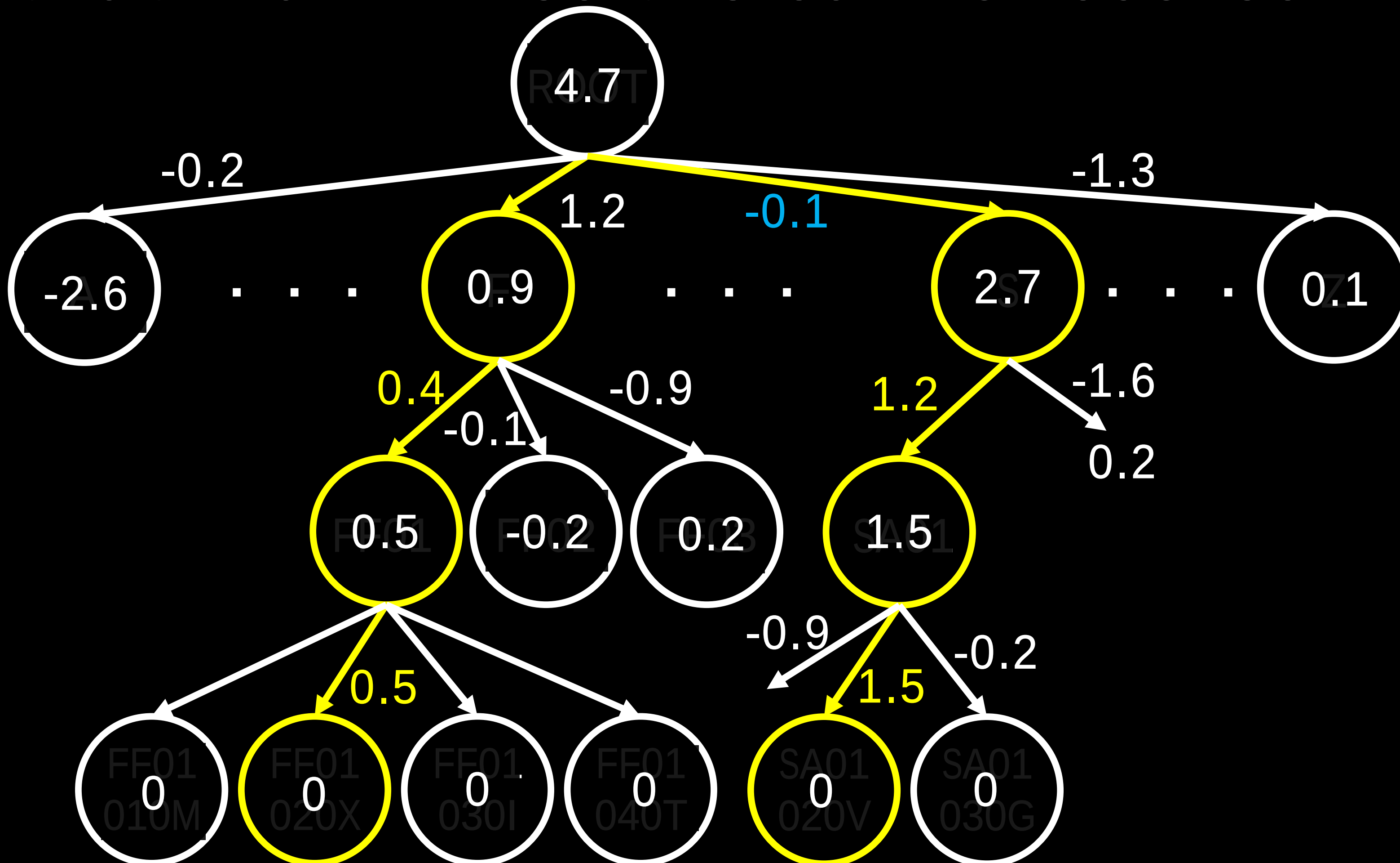
Global Model: Find a Subtree that maximizes the sum of scores



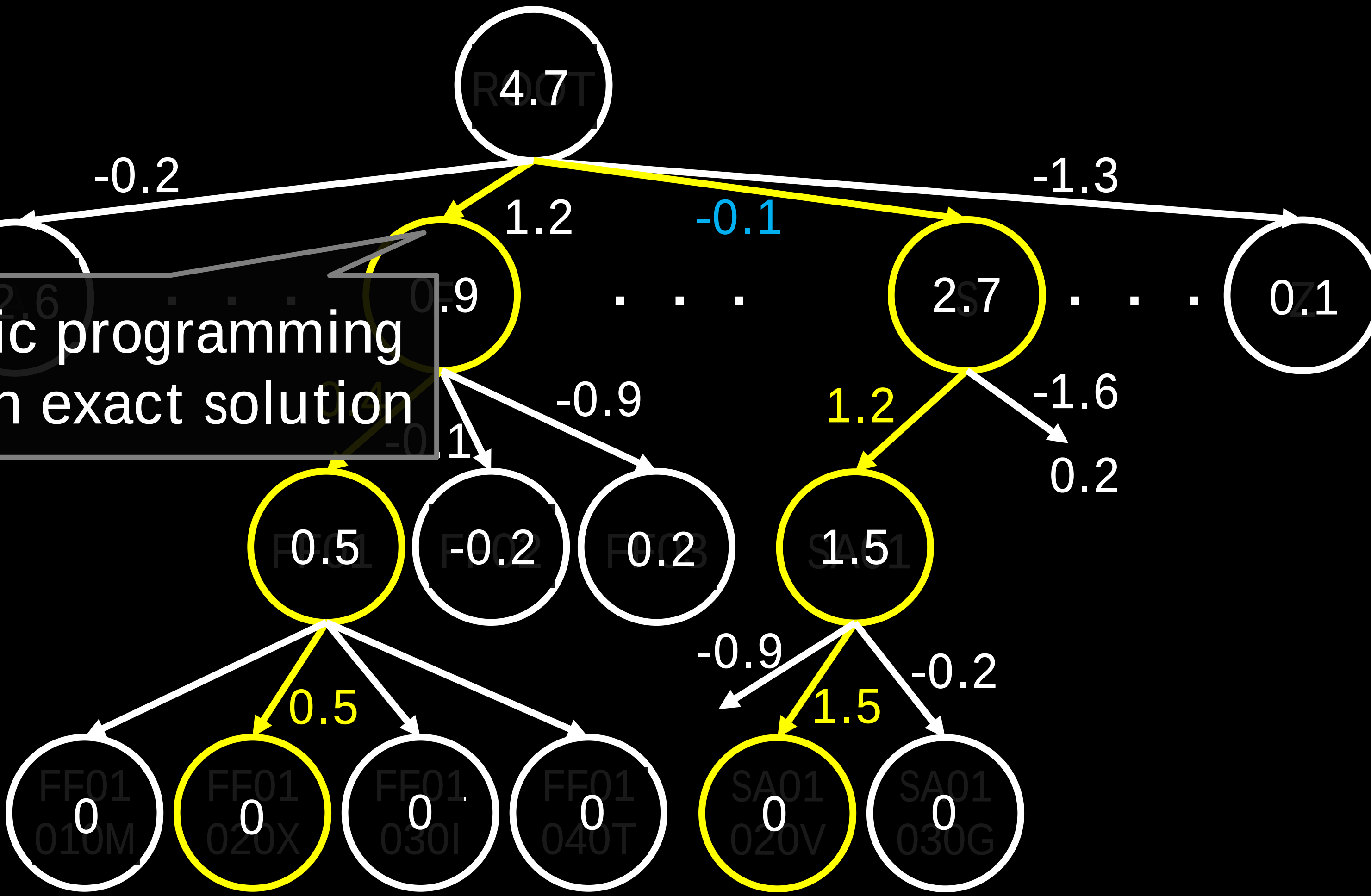
Global Model: Find a Subtree that maximizes the sum of scores



Global Model: Find a Subtree that maximizes the sum of scores

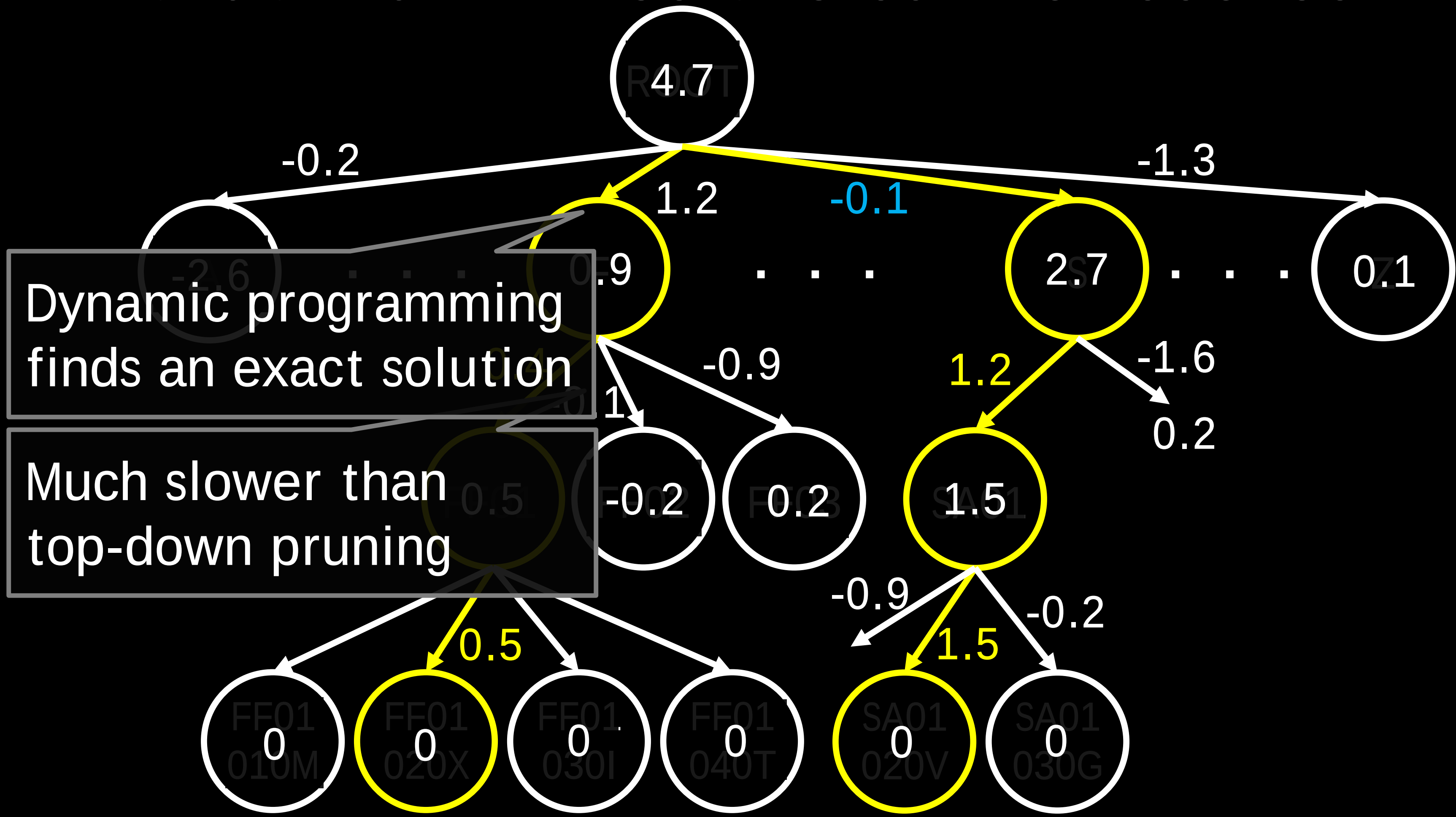


Global Model: Find a Subtree that maximizes the sum of scores

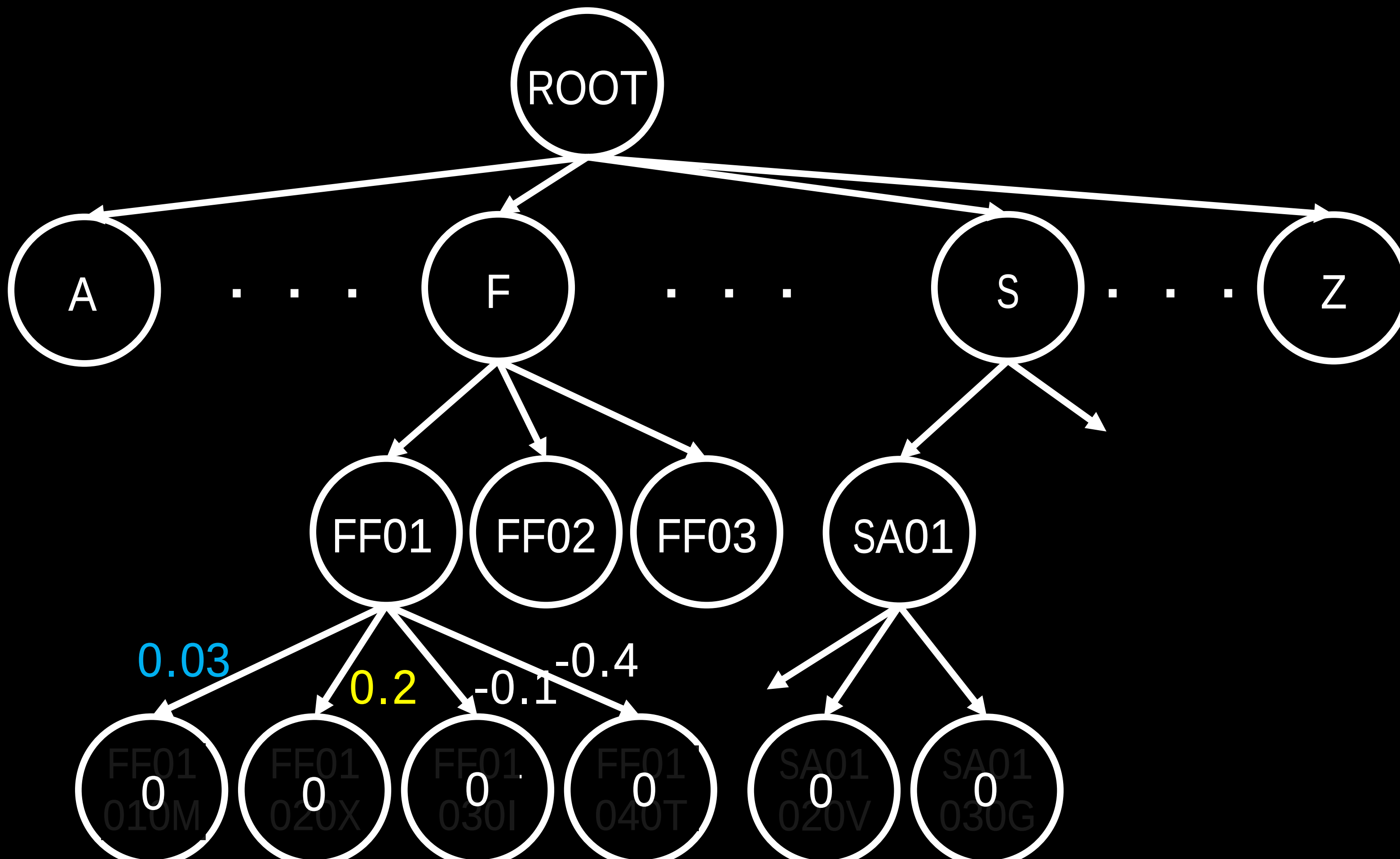


Dynamic programming finds an exact solution

Global Model: Find a Subtree that maximizes the sum of scores



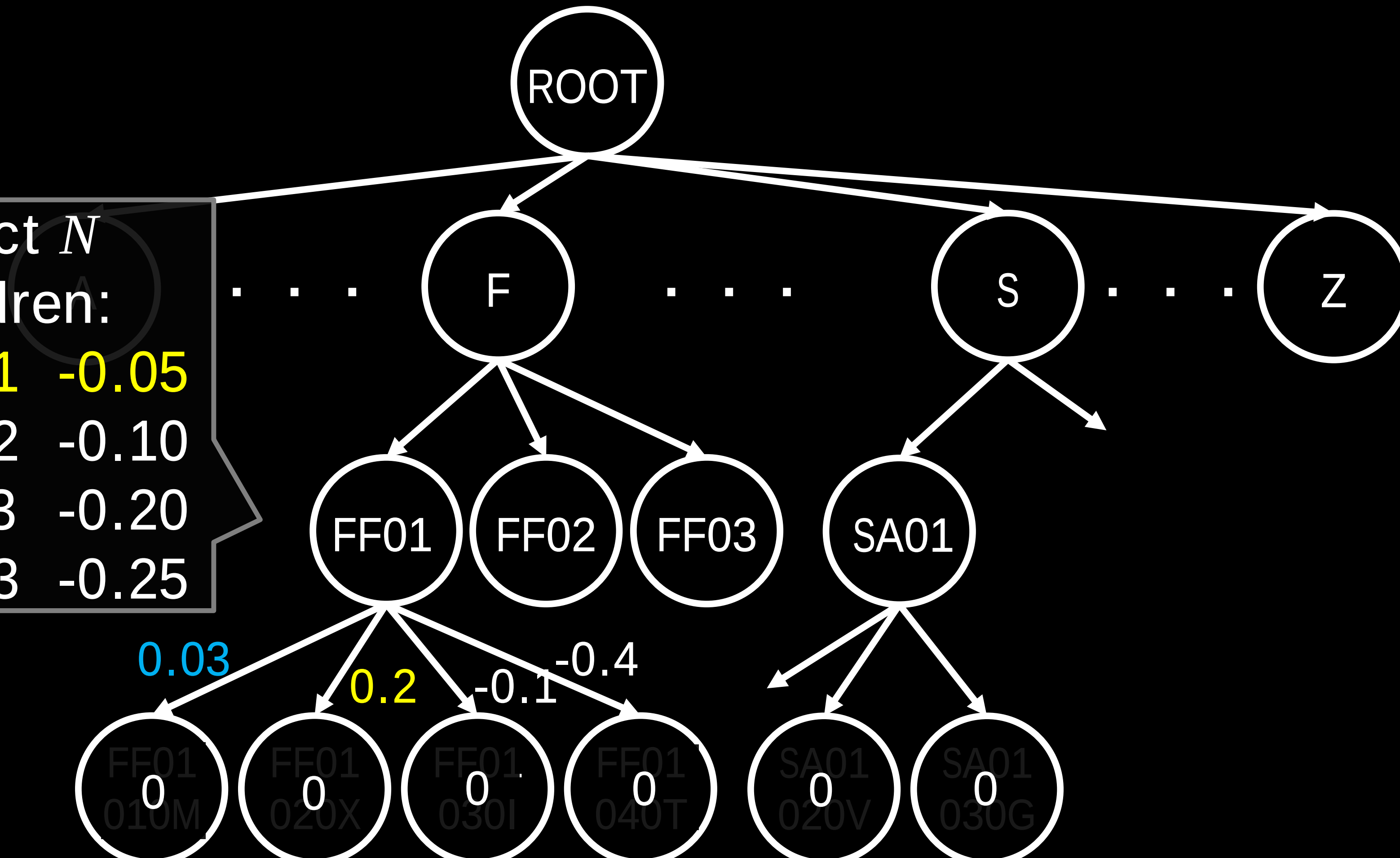
Additonal Scores by Branching Features



Additonal Scores by Branching Features

Select N children:

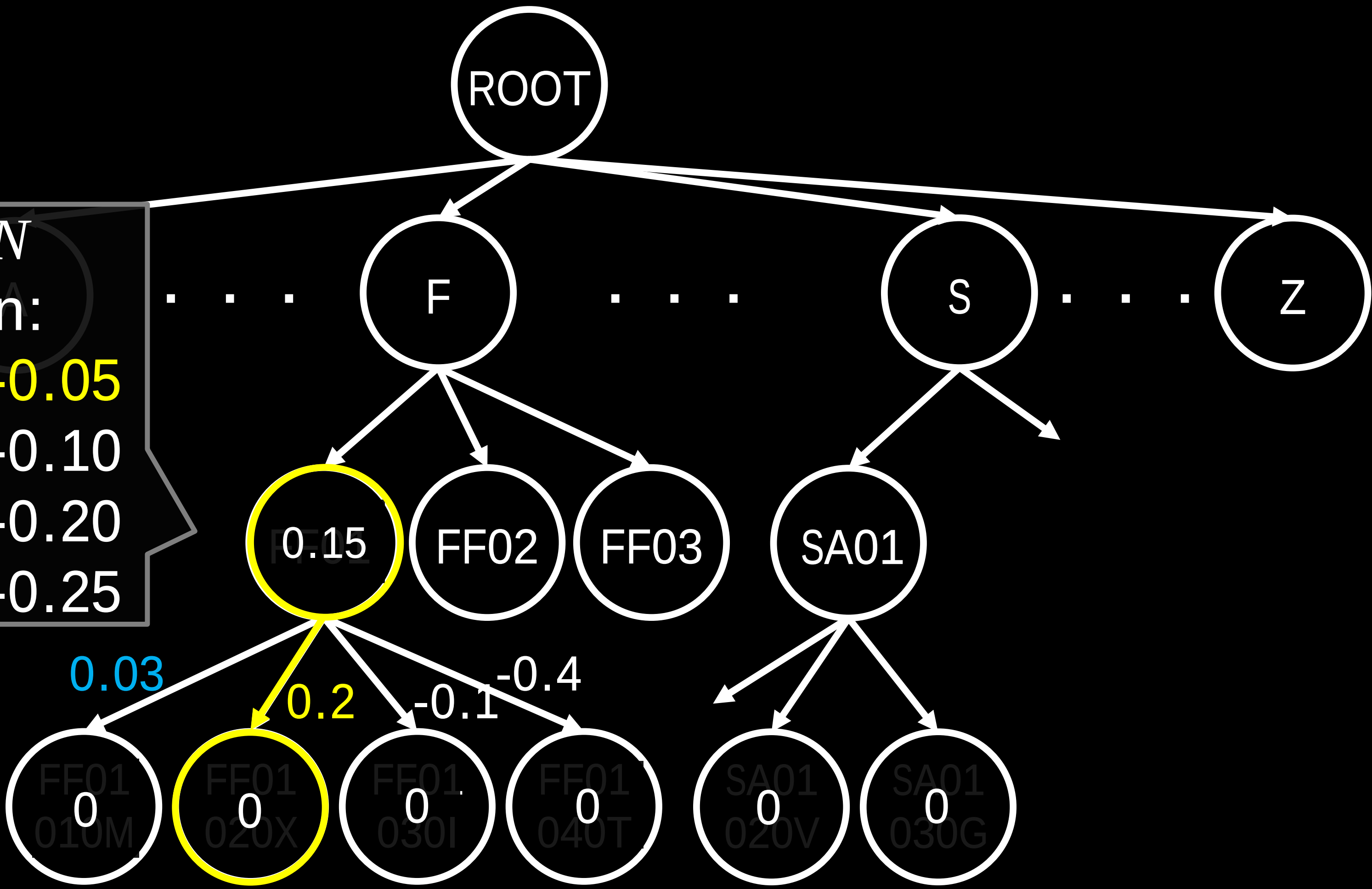
- N=1 -0.05
- N=2 -0.10
- N=3 -0.20
- N>3 -0.25



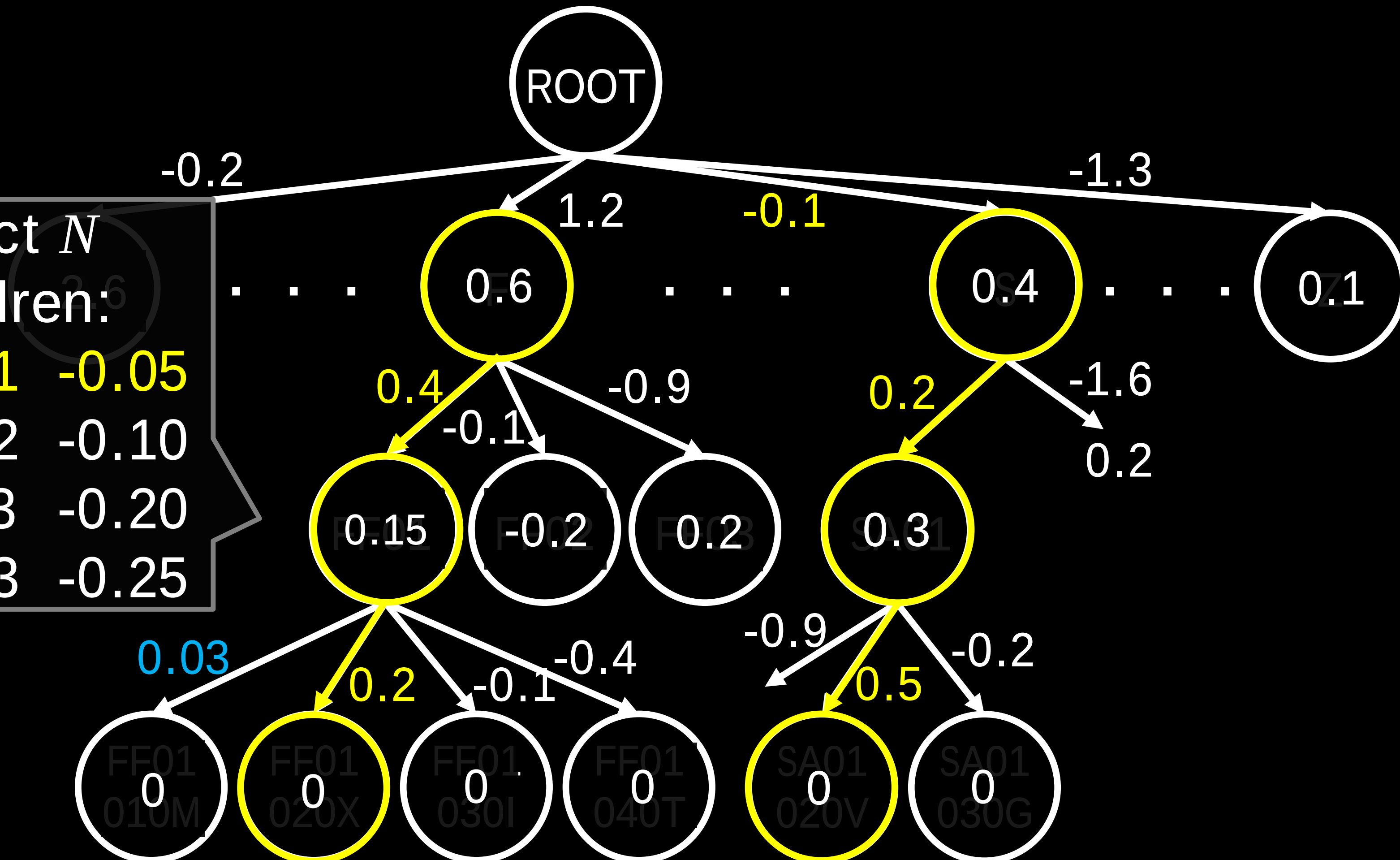
Additonal Scores by Branching Features

Select N children:

- N=1 -0.05
- N=2 -0.10
- N=3 -0.20
- N>3 -0.25



Additonal Scores by Branching Features



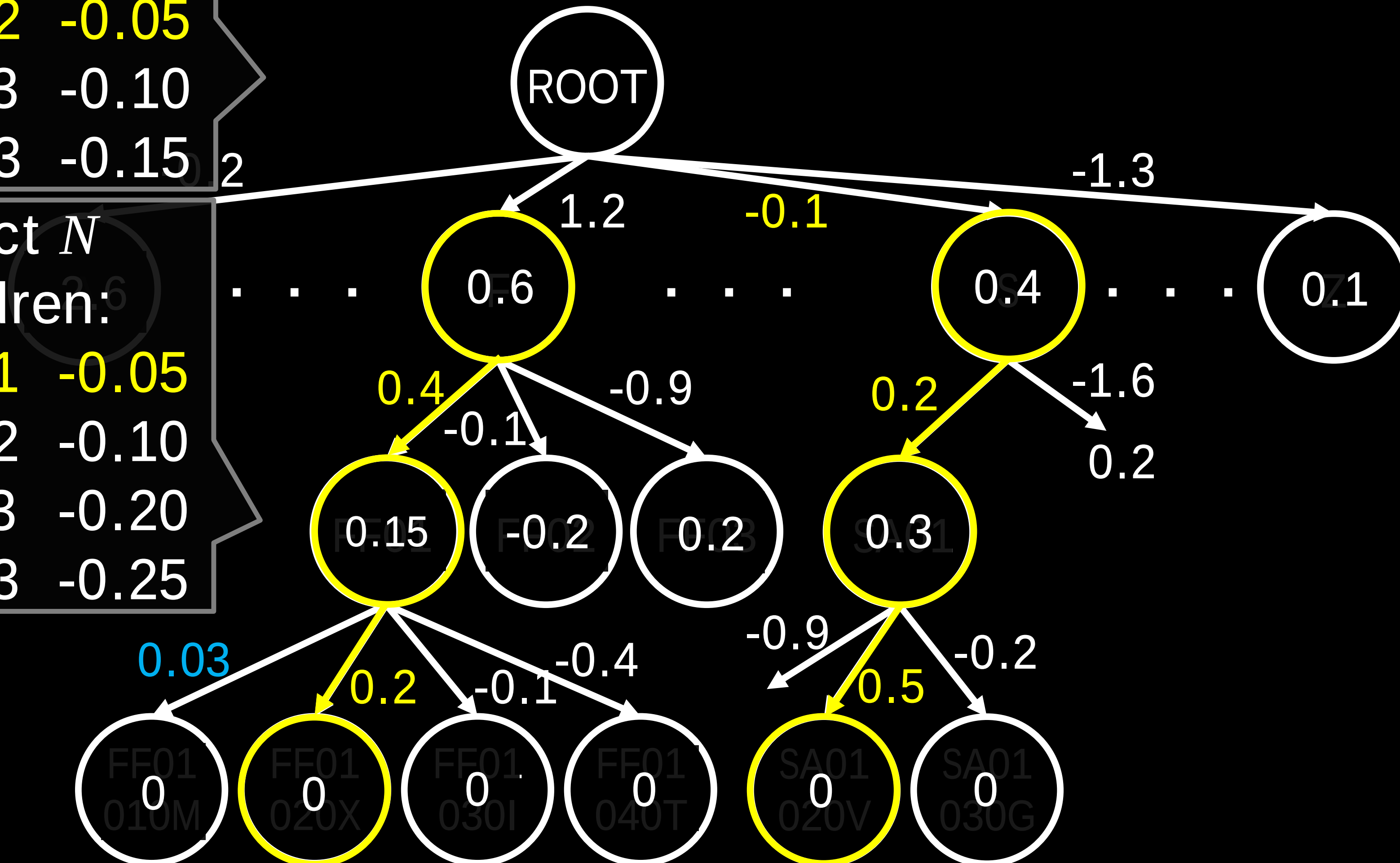
Select N children:

- N=1 -0.05
- N=2 -0.10
- N=3 -0.20
- N>3 -0.25

Additional Scores by Branching Features

Select N children:
- N=1 -0.02
- N=2 -0.05
- N=3 -0.10
- N>3 -0.15

Select N children:
- N=1 -0.05
- N=2 -0.10
- N=3 -0.20
- N>3 -0.25

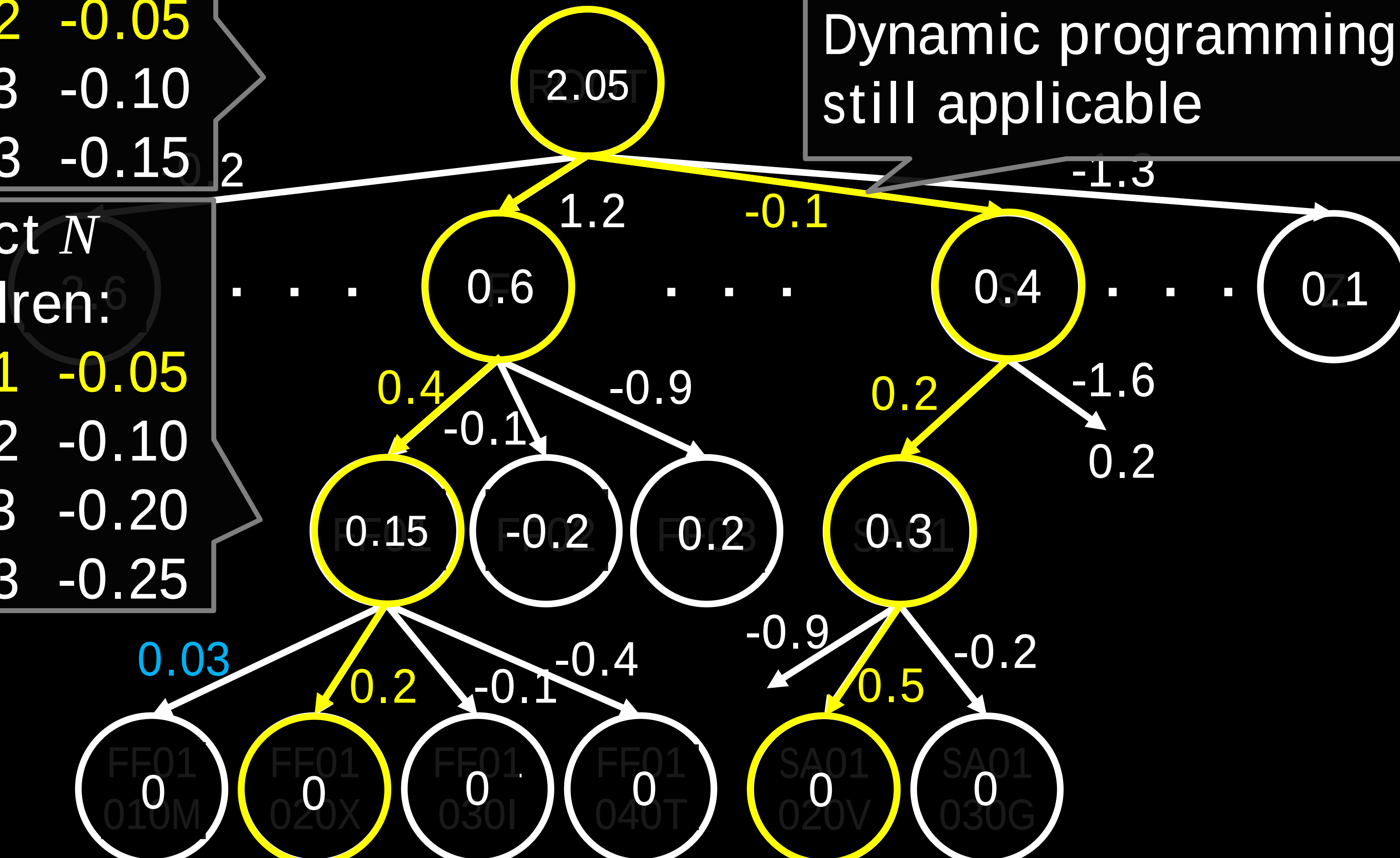


Additional Scores by Branching Features

Select N children:
- N=1 -0.02
- N=2 -0.05
- N=3 -0.10
- N>3 -0.15

Select N children:
- N=1 -0.05
- N=2 -0.10
- N=3 -0.20
- N>3 -0.25

Dynamic programming still applicable



Training

- Global optimization
  - Instead of training edge classifiers separately, we directly optimize the global model
- Parallelization by iterative parameter mixing (McDonald+, 2010)
  - Global optimization is too slow
  - Parallelize training by splitting training data into small “shards” and mix the models per iteration

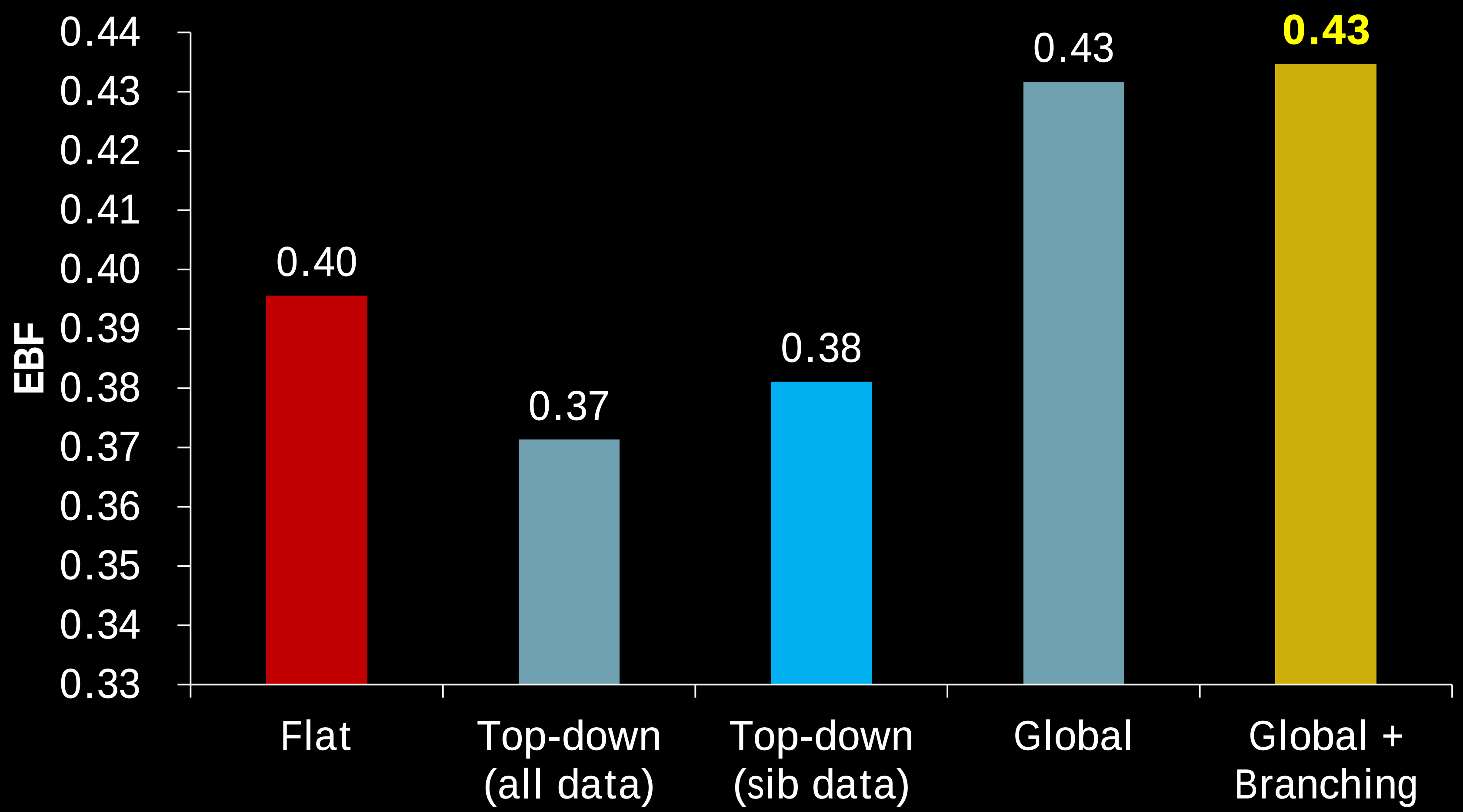
Outline

- Task Settings
- Previous Work
- Proposed Method
- Results and Discussion
- Conclusion

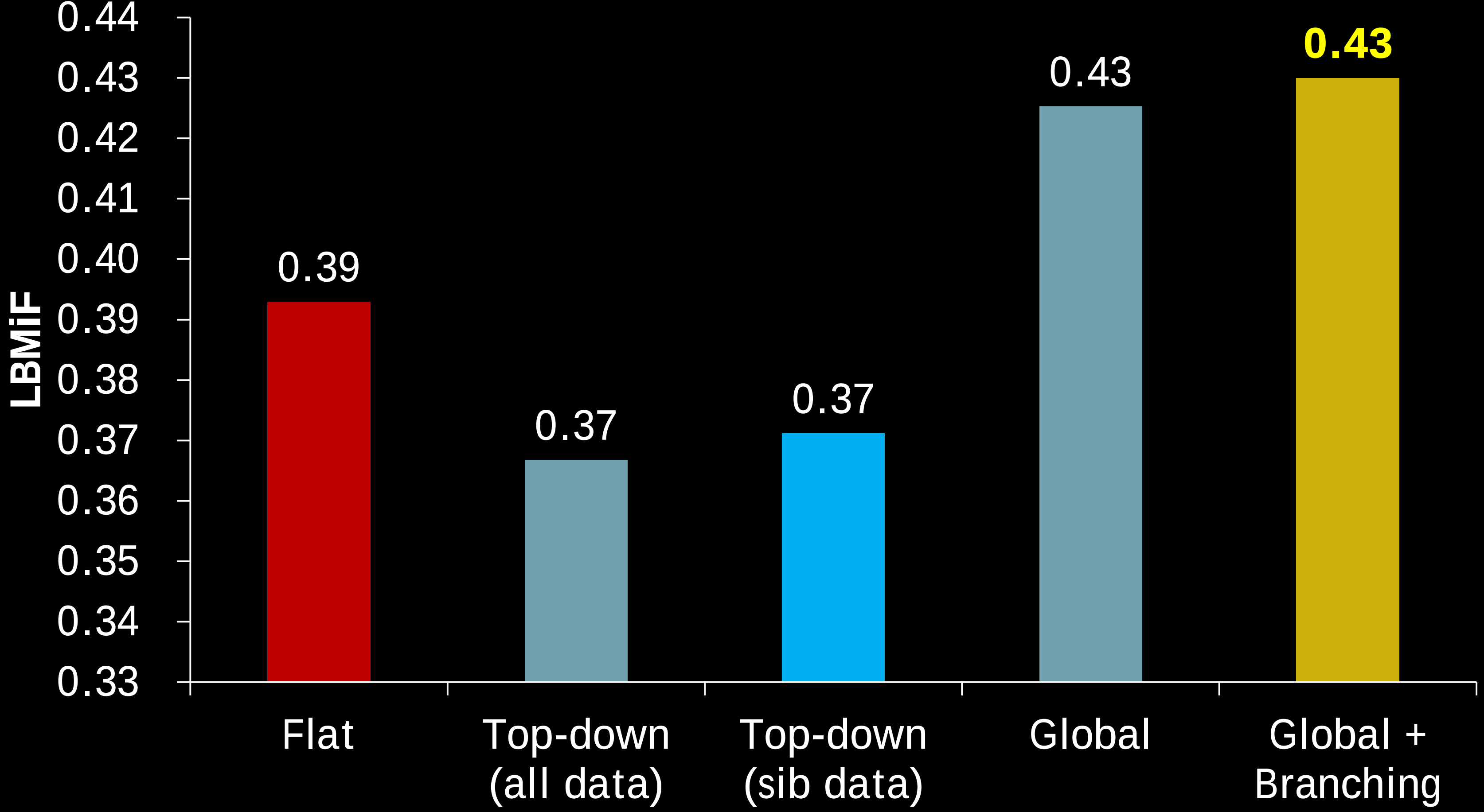
Experimental Settings

- Portion of JSTPlus (Japanese academic bibliographic database)
  – Training: 409,892 documents
  – Evaluation: 45,419 documents
- Features: content words & journal name
- Training:
  – Online passive-aggressive algorithm
  – 10 iterations

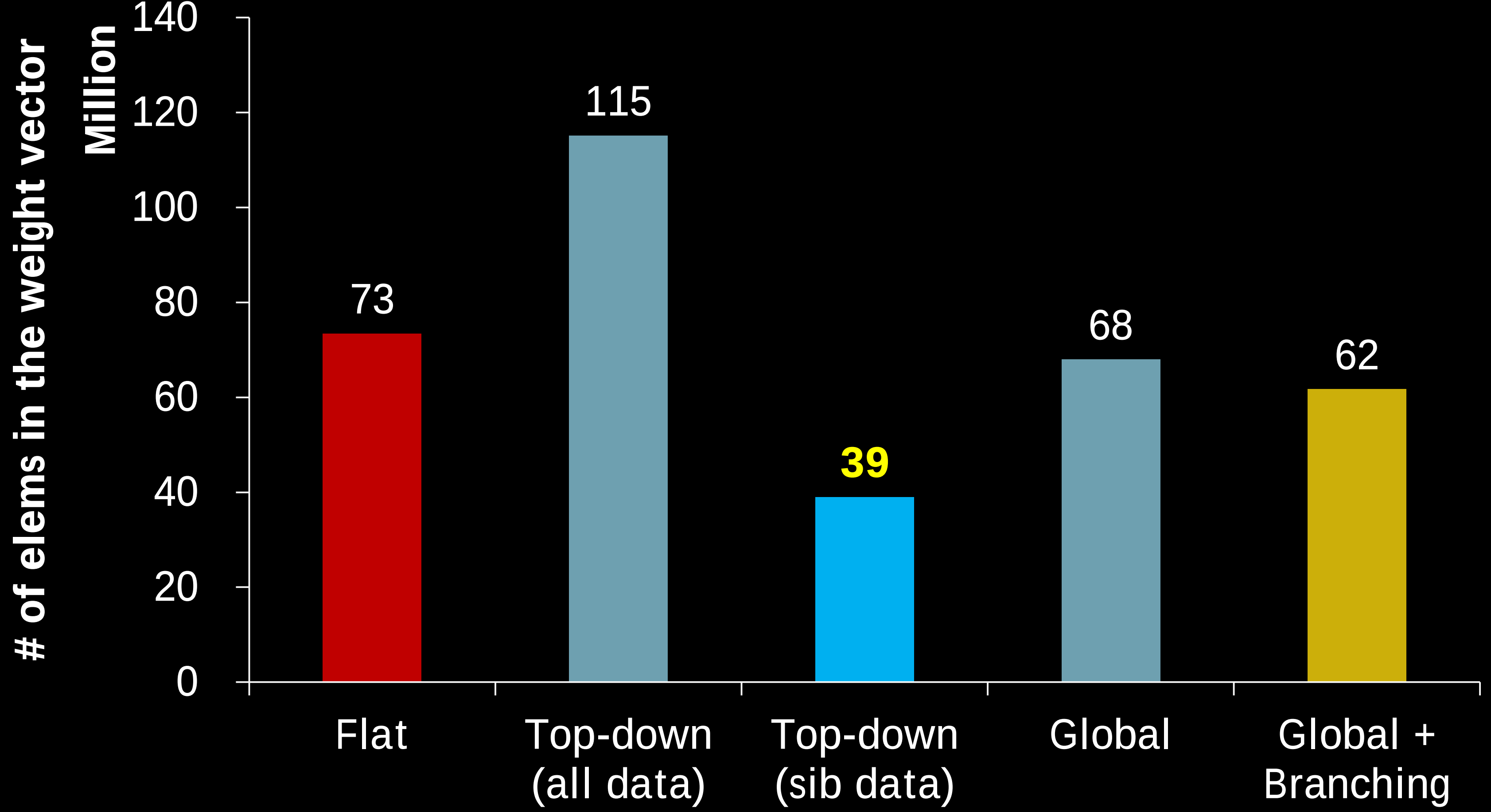
Example-based F Measure



Label-based Micro-average F Measure

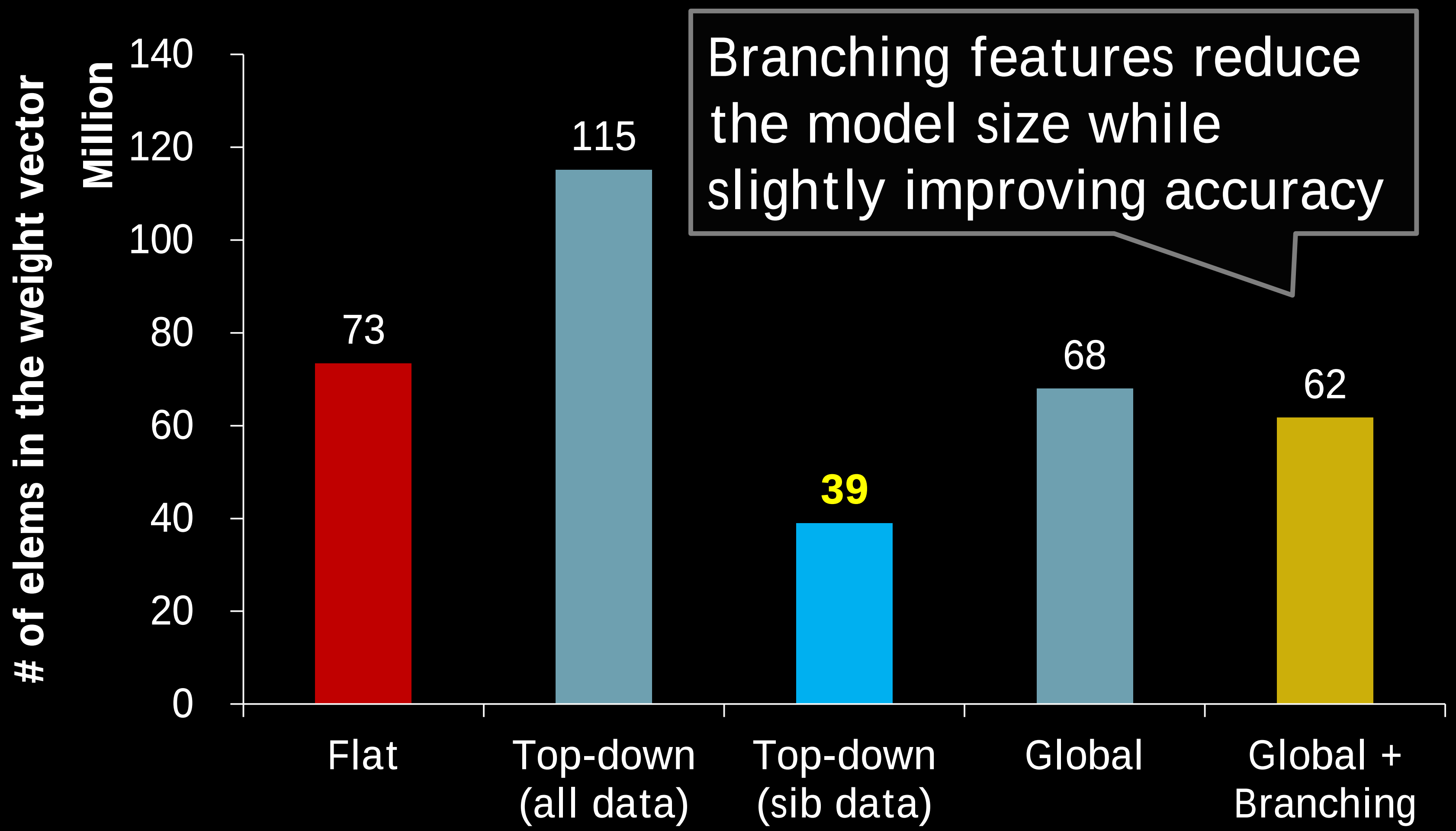


Model Size



# of elems in the weight vector
Million
140
120
100
80
60
40
20
0
73
115
39
68
62
Flat
Top-down (all data)
Top-down (sib data)
Global
Global + Branching

Model Size



# of elems in the weight vector
Million
140
120
100
80
60
40
20
0
73
115
39
68
62
Flat
Top-down (all data)
Top-down (sib data)
Global
Global + Branching
Branching features reduce the model size while slightly improving accuracy

Conclusions & Future Work

- Exploit the label hierarchy to capture inter-label dependencies
- Branching features reduce model size while slightly improving accuracy
- Future work
  - Extension to directed acyclic graphs
  - Improving scalability