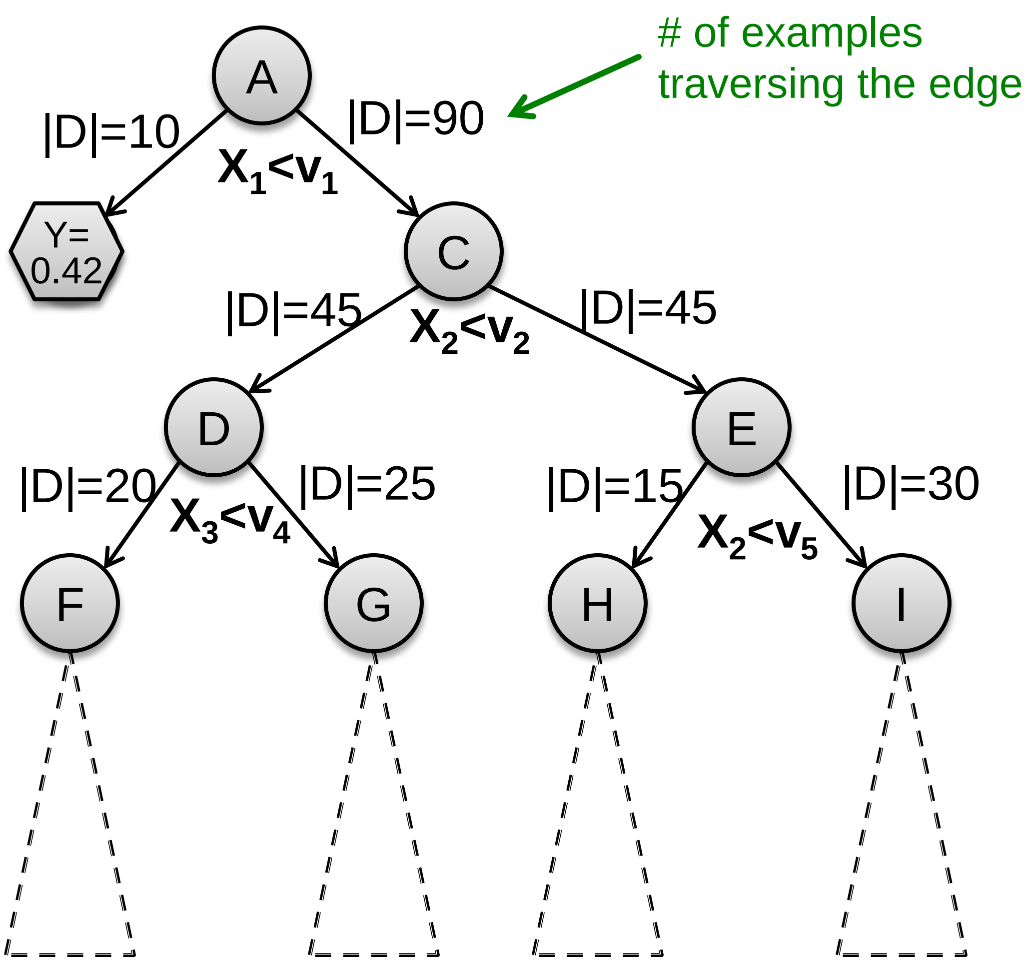
How to construct a tree?

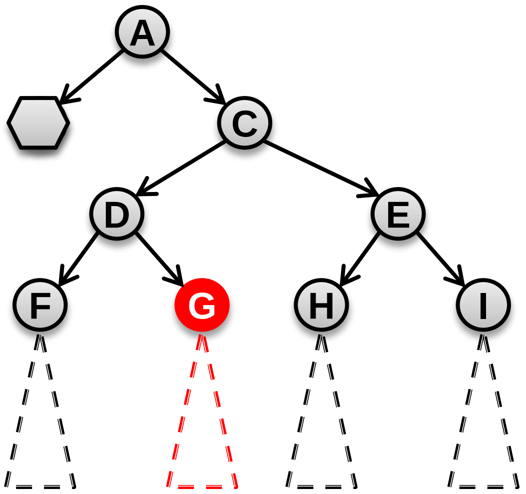
Mining of Massive Datasets
Leskovec, Rajaraman, and Ullman
Stanford University

How to construct a tree?

- **Training dataset D*, |D*|=100 examples**



How to construct a tree?



- Imagine we are currently at some node ***G***
  - Let D_G be the data that reaches ***G***
- **There is a decision we have to make: Do we continue building the tree?**
  - **If yes**, which variable and which value do we use for a **split**?
    - Continue building the tree recursively
  - **If not**, how do we make a prediction?
    - We need to build a **"predictor node"**

How to construct a tree?

Algorithm 1 BuildSubtree

Require: Node n, Data $D \subset D^*$

$(n \rightarrow \text{split}, D_L, D_R) = \boxed{\text{FindBestSplit}(D)}$ (1)

if $\boxed{\text{StoppingCriteria}(D_L)}$ **then** (2)

  $n \rightarrow \text{left\_prediction} = \boxed{\text{FindPrediction}(D_L)}$ (3)

else

  BuildSubtree $(n \rightarrow \text{left}, D_L)$

if $\text{StoppingCriteria}(D_R)$ **then**

  $n \rightarrow \text{right\_prediction} = \text{FindPrediction}(D_R)$

else

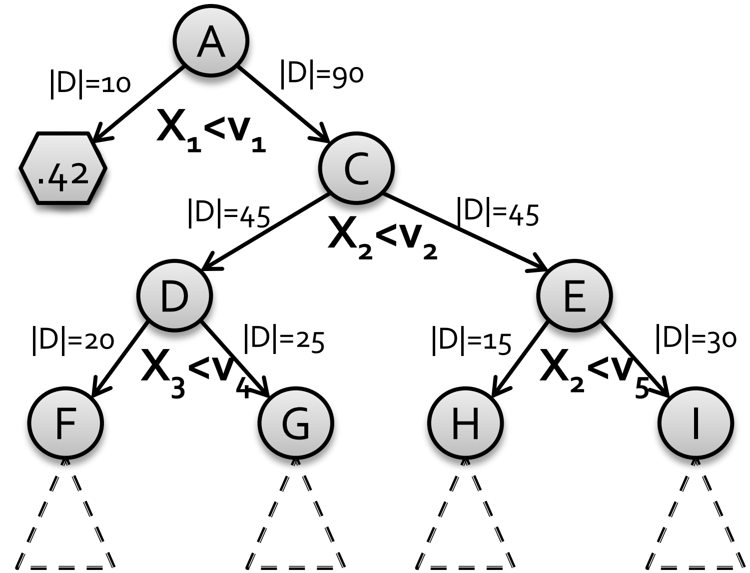
  BuildSubtree $(n \rightarrow \text{right}, D_R)$

- **Requires at least a single pass over the data!**

How to construct a tree?

(1) How to split? Pick attribute & value that optimizes some criterion

- **Classification: Information Gain**
  - **Measures how much a given attribute X tells us about the class Y**
  - **IG(Y | X)** : We must transmit **Y** over a binary link. How many bits on average would it save us if both ends of the line knew **X**?



Back to: How to construct a tree?

Algorithm 1 BuildSubtree

Require: Node n, Data $D \subset D^*$

$(n \rightarrow \text{split}, D_L, D_R) = \boxed{\text{FindBestSplit}(D)}$ **(1)**

if $\boxed{\text{StoppingCriteria}(D_L)}$ then **(2)**

  $n \rightarrow \text{left\_prediction} = \boxed{\text{FindPrediction}(D_L)}$ **(3)**

else

  BuildSubtree $(n \rightarrow \text{left}, D_L)$

if StoppingCriteria(D_R) then

  $n \rightarrow \text{right\_prediction} = \text{FindPrediction}(D_R)$

else

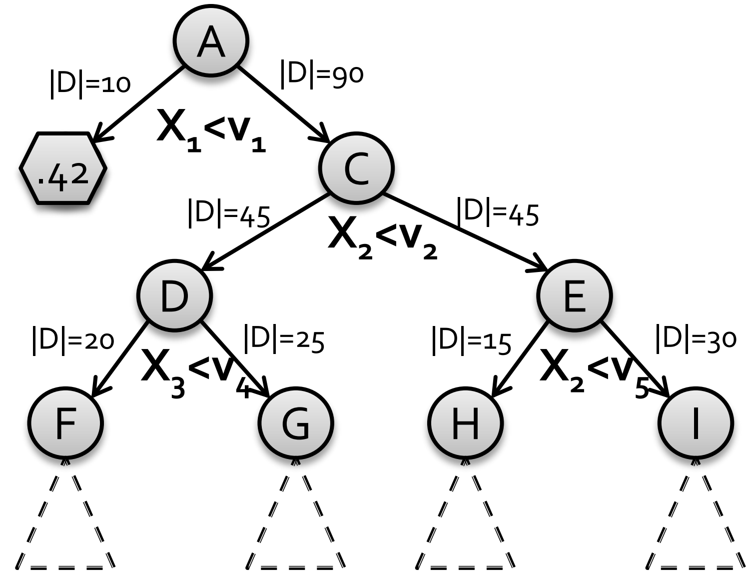
  BuildSubtree $(n \rightarrow \text{right}, D_R)$

- **Requires at least a single pass over the data!**

How to construct a tree?

(1): How to split?

- **Regression:**
  - Find split ***(X_i, v)*** that creates **D, D_L, D_R**: parent, left, right child datasets and **maximizes**:



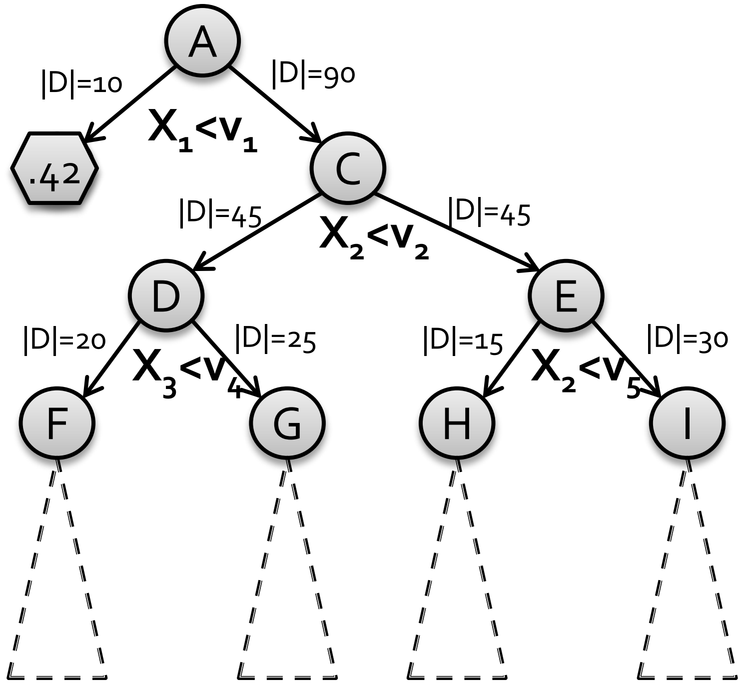
$$|D| \cdot Var(D) - \left(|D_L| \cdot Var(D_L) + |D_R| \cdot Var(D_R)\right)$$

  - $Var(D) = \frac{1}{n}\sum_{i=1}^{|D|}(y_i - \bar{y})^2$ … variance of y_i in D
  - For ordered domains sort **X_i** and consider a split between each pair of adjacent values
  - For categorical ***X_i*** find best split based on subsets

How to construct a tree?

(2) When to stop?

- Many different heuristic options
- **Two ideas:**
  - **(1) When the leaf is "pure"**
    - The target variable does not vary too much: $Var(y_i) < \varepsilon$
  - **(2) When # of examples in the leaf is too small**
    - For example, $|D| \leq 10$



How to construct a tree?

(3) How to predict?

- **Many options**
  - **Regression:**
    - Predict average y_i of the examples in the leaf
    - Build a linear regression model on the examples in the leaf
  - **Classification:**
    - Predict most common y_i of the examples in the leaf