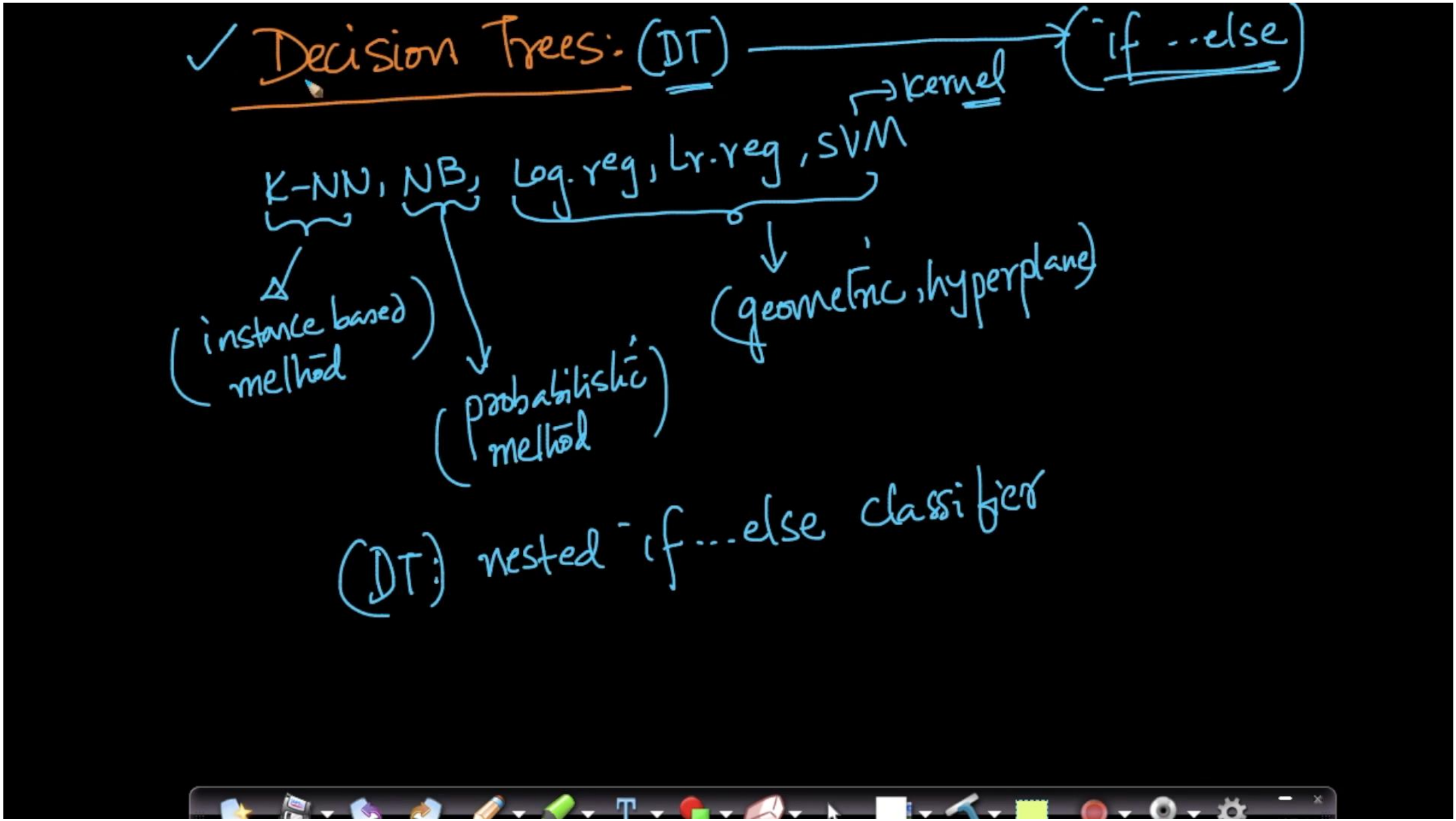

✓ Decision Trees: (DT) → (if ...else)
K-NN, NB, Log. reg, Lr. reg, SVM
kernel
(instance based method)
(probabilistic method)
(geometric, hyperplane)
(DT): nested if...else classifier

EDA : Iris dataset

↳ $y_i = \{1, 2, 3\}$

(SL, SW, PL, PW)

simple { if $PL < 5$ then $y_i = 1$ }

✓ nested if .. else conditions

↓

diagram (Tree)

$x_i = \langle SL, PL, SW, PL \rangle$

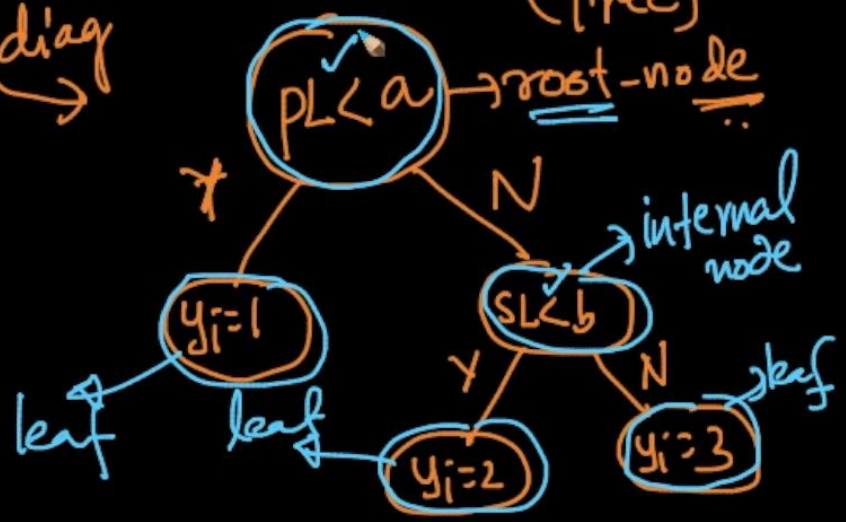
model [if $PL < a$ [$y_i = 1$]

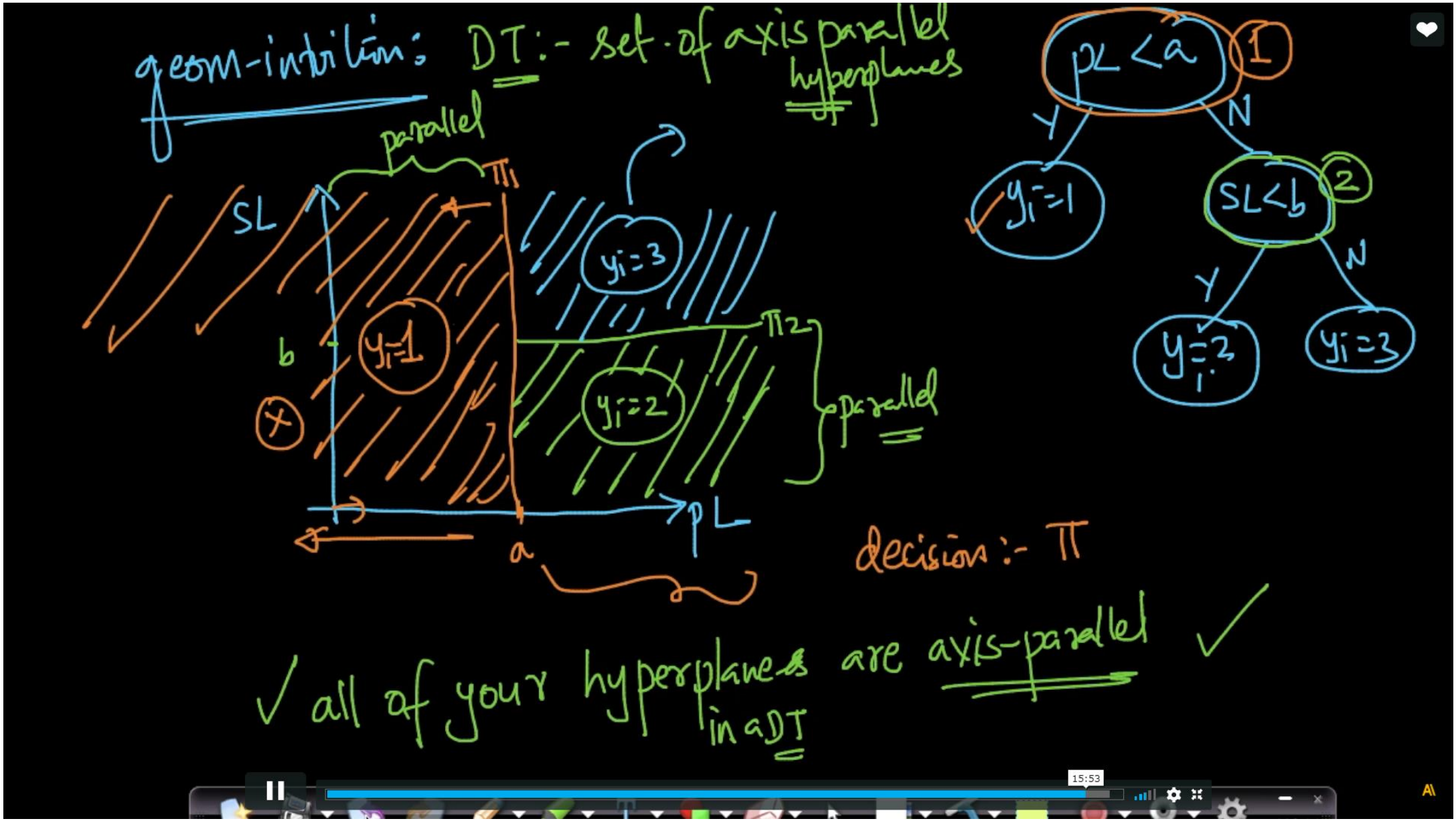
else : [if $SL < b$ class = 2

else class = 3]]

diag →

(Tree)




geom-intuition: DT:- set of axis parallel hyperplanes
parallel
π_1
SL
$y_i=3$
$y_i=1$
b
x
π_2
$y_i=2$
parallel
pL
a
decision:- π
pL<a
1
Y
N
$y_i=1$
SL<b
2
Y
N
$y_i=2$
$y_i=3$
✓ all of your hyperplanes are axis-parallel in a DT ✓
15:53

node — vertex

root-node

leaf-node

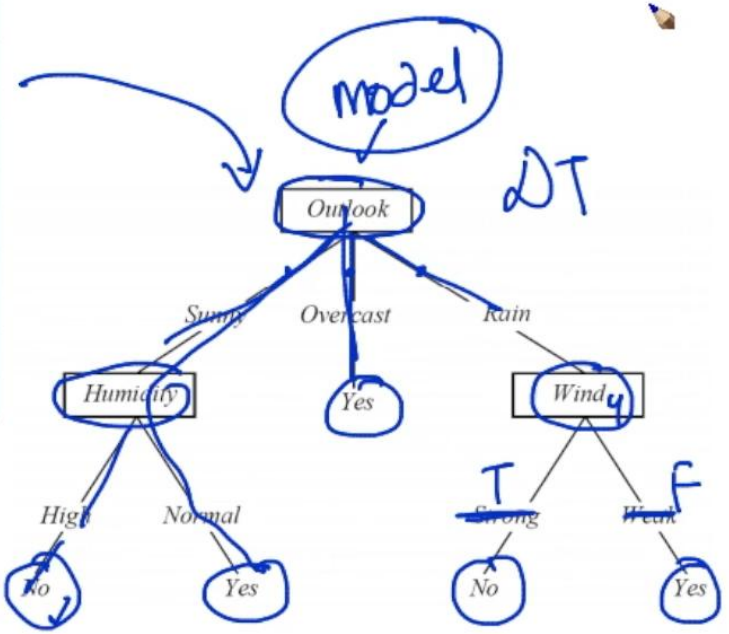
internal-node

non-leaf nodes :- decisions in a DT

Contents - Google Docs | 32-decision-trees.pdf | Chekuri Srikan...
homepage.cs.uri.edu/faculty/hamel/courses/2016/spring2016/csc581/lecture-notes/32-decision-trees.pdf

Play Tennis Example

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No



05:30
67 / 67
APPLIED COURSE

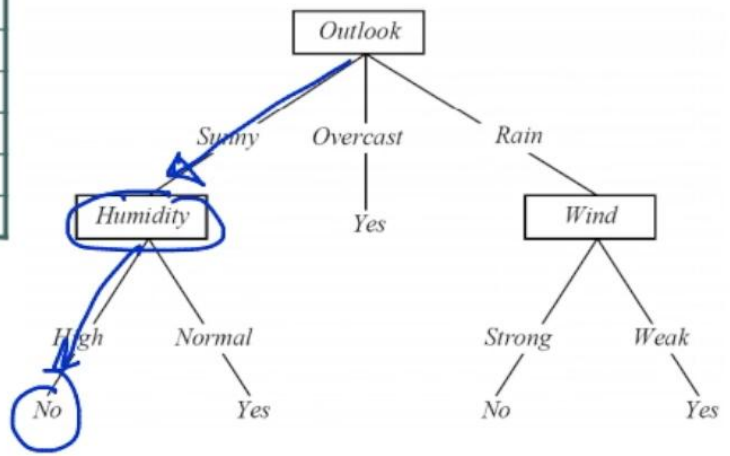
Contents - Google Docs | 32-decision-trees.pdf | Chekuri Srikan...
homepage.cs.uri.edu/faculty/hamel/courses/2016/spring2016/csc581/lecture-notes/32-decision-trees.pdf

Play Tennis Example

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

$x_4 = [\text{Sunny}, \text{Hot}, \text{High}, \text{True}]$

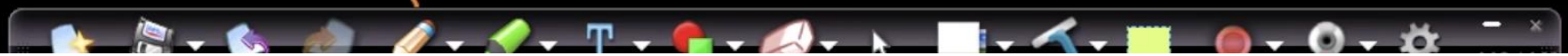
$y_4 = \text{No}$



✓ Building a DT: Entropy → information theory

→ Physics, info thy ...

✓ $D_{Train} \rightarrow DT$



Contents - Google Docs | 32-decision-trees.pdf | Chekuri Srikan...
homepage.cs.uri.edu/faculty/hamel/courses/2016/spring2016/csc581/lecture-notes/32-decision-trees.pdf

Play Tennis Example

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

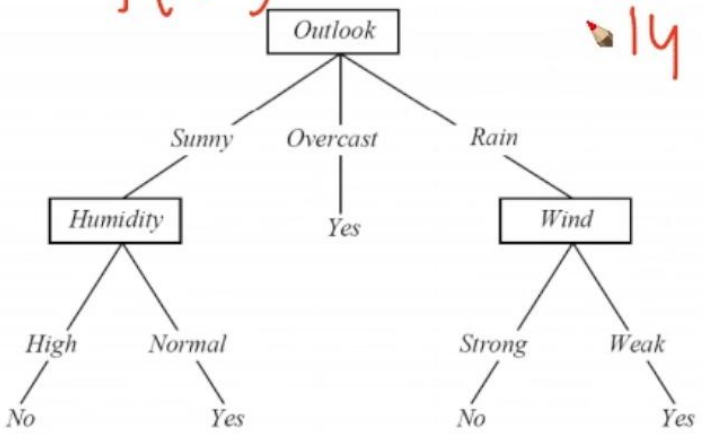
Y

Y : play Tennis

Y_+, Y_- $\{P(Y_+) = \frac{9}{14}\}$

$P(Y_-) = 1 - P(Y_+) = \frac{5}{14}$

DT



r.v $Y \longrightarrow y_1, y_2, y_3, \ldots y_k$

entropy

$$H(Y) = -\sum_{i=1}^{K} p(y_i) \log_b \left(p(y_i) \right)$$

$b = 2$ ✓

or

$b = e = 2.718$

$p(y_i) =$

$p(y_i) = P(Y = y_i)$

$\log_2 = \lg$

$\log_e = \ln$

$$H(Y) = -\sum_{i=1}^{K} p(y_i) \log_2(p(y_i))$$

$$H(Y) = -\frac{9}{14}\log\left(\frac{9}{14}\right) - \frac{5}{14}\log\left(\frac{5}{14}\right) = 0.94$$ ✓

$\frac{5}{14}$ → % age of -ve pts in D

$\frac{9}{14}$ → $P(Y+)$

$P(Y-)$

$\frac{\#\text{ +ve pts}}{\text{Total \# pts}}$ = % age of +ve pts in D

$$H(Y) = -\sum_{i=1}^{K} p(y_i) \log_2(p(y_i))$$

$$H(Y) = -\frac{9}{14}\log\left(\frac{9}{14}\right) - \frac{5}{14}\log\left(\frac{5}{14}\right) = 0.94$$ ✓

$\frac{5}{14}$ → % age of -ve pts in D

$\frac{9}{14}$ → $P(Y+)$

$P(Y-)$

$\frac{\#\text{ +ve pts}}{\text{Total \# pts}}$ = % age of +ve pts in D

Properties: $Y \rightarrow y_+, y_-$ (2 class, 2 category)

Case 1:

$D \rightarrow y_+ \rightarrow 99\%$

$D \rightarrow y_- \rightarrow 1\%$

$H(Y) = -0.99 \lg 0.99 - 0.01 \lg 0.01$

$= 0.0801$

Case 2:

$D \rightarrow y_+ \rightarrow 50\%$

$D \rightarrow y_- \rightarrow 50\%$

$H(Y) = -0.5 \lg 0.5 - 0.5 \lg 0.5$

$= 1$

Case 3:

$\times y_+ \rightarrow 0\%$

$\checkmark y_- \rightarrow 100\%$

$H(Y) = 0$

Properties: $Y \rightarrow y_+, y_-$ (2 class, 2 category)

Case 1:

$D \rightarrow y_+ \rightarrow 99\%$

$D \rightarrow y_- \rightarrow 1\%$

$H(Y) = -0.99 \lg 0.99 - 0.01 \lg 0.01$

$= 0.0801$

Case 2:

$D \rightarrow y_+ \rightarrow 50\%$

$D \rightarrow y_- \rightarrow 50\%$

$H(Y) = -0.5 \lg 0.5 - 0.5 \lg 0.5$

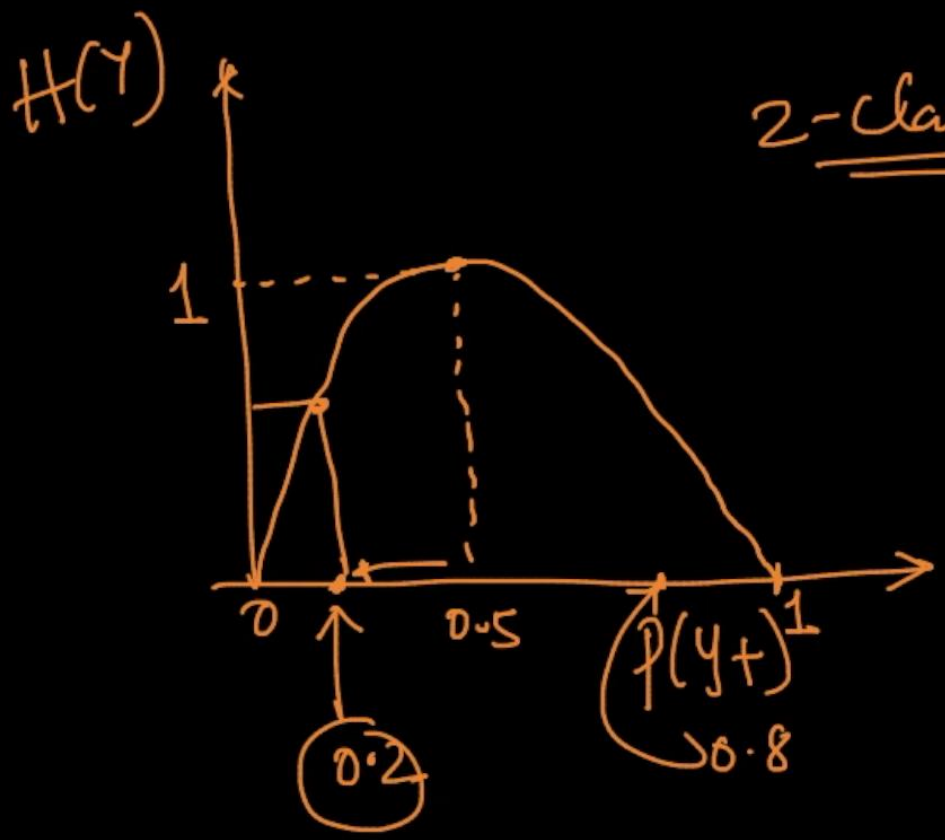
$= 1$

Case 3:

$\times\ y_+ \rightarrow 0\%$

$\checkmark\ y_- \rightarrow 100\%$

$H(Y) = 0$



2-class case

$$-0.2\log 0.2 - 0.8\log 0.8$$

$$= -0.8\log 0.8 - 0.2\log 0.2$$

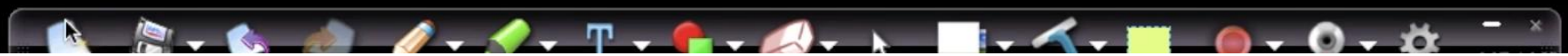
$$\left[p(y+) = 1 - p(y-) \right]$$

$$\begin{cases} 0.2 \leftarrow y+ \\ 0.8 \leftarrow y- \end{cases}$$

$Y \rightarrow y_1, y_2, \ldots y_k$

equi-probable $\rightarrow$ entropy is maximum

$y_1 \rightarrow$ most possible
$y_2, y_3 \ldots \rightarrow$ '0' $\rightarrow$ entropy is minimum



$Y \rightarrow \{y_1, y_2, \ldots y_k\}$

equi-probable $\rightarrow$ entropy is maximum ✓

histogram

PDF

1

v-peaked

$y_1 \rightarrow$ most possible

$y_2, y_3 \ldots \rightarrow$ '0'

$\rightarrow$ entropy is minimum

not peaked

discrete

y_1 y_2 y_3 y_4 y_{16} Y

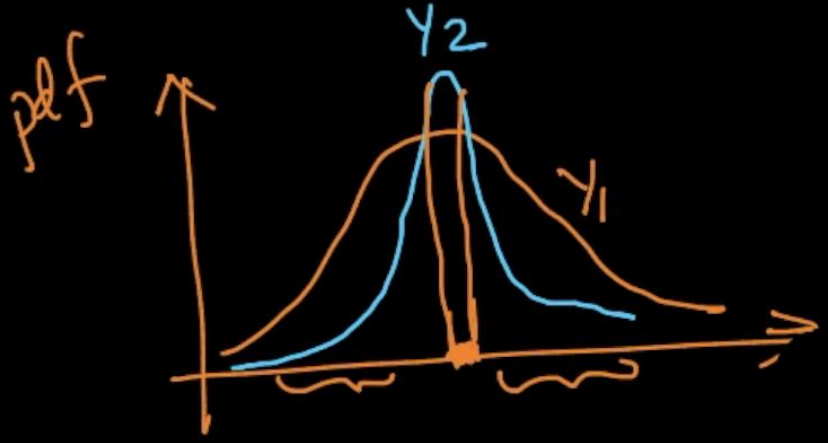
PDF

more peaked a dist is, less is its entropy

PDF

→ uniform dist

entropy is max



y_1 → less peaked

y_2 → very peaked

$H(y_2) < H(y_1)$

Contents - Google Docs | 32-decision-trees.pdf | Chekuri Srikan...
homepage.cs.uri.edu/faculty/hamel/courses/2016/spring2016/csc581/lecture-notes/32-decision-trees.pdf

Partitioning the Data Set

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Sunny	Mild	Normal	True	Yes

E = .97

Sunny

Outlook

Overcast

Outlook	Temperature	Humidity	Windy	PlayTennis
Overcast	Hot	High	False	Yes
Overcast	Cool	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes

E = 0

Rainy

Outlook	Temperature	Humidity	Windy	PlayTennis
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Rainy	Mild	Normal	False	Yes
Rainy	Mild	High	True	No

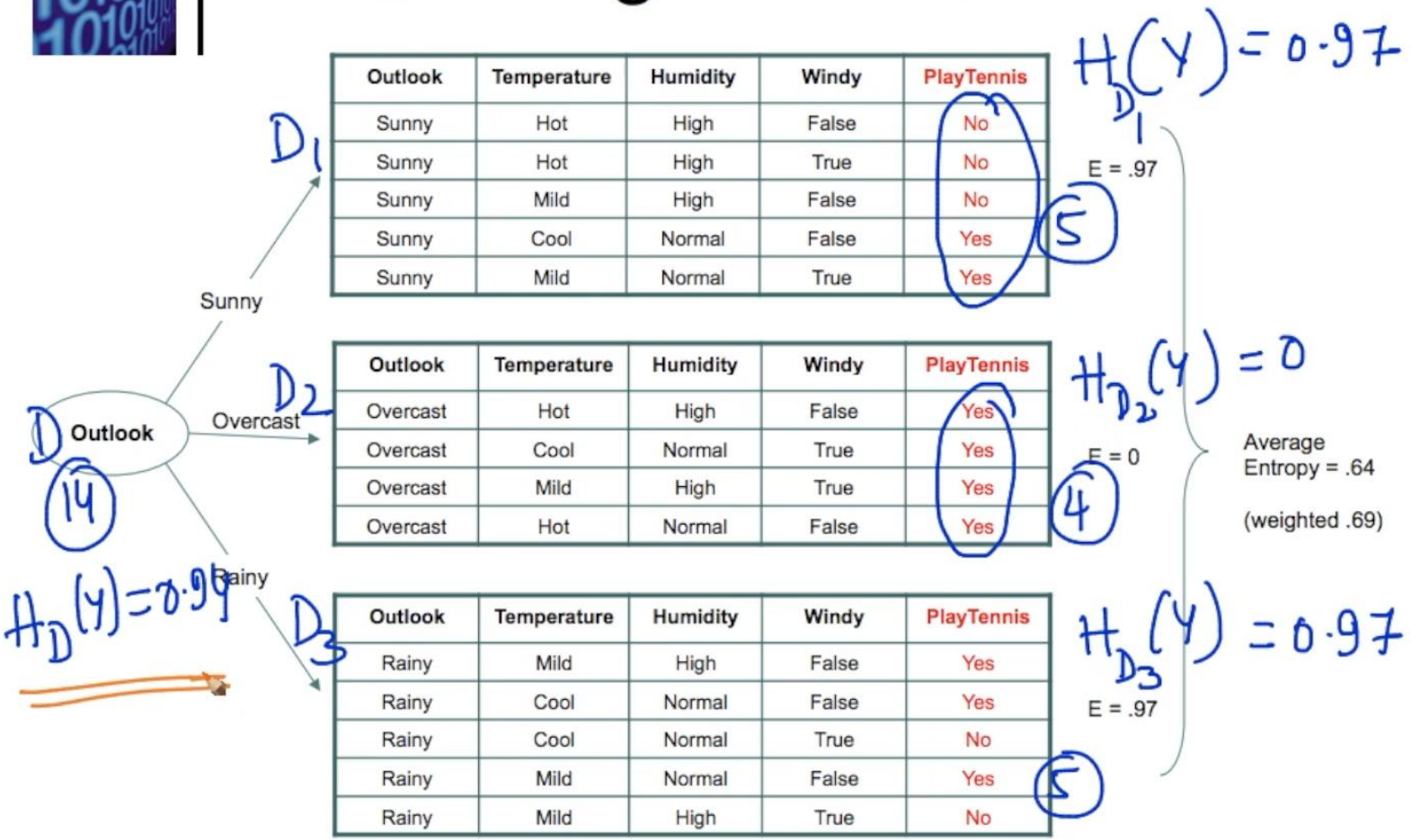
E = .97

Average Entropy = .64

(weighted .69)

Contents - Google Docs × 32-decision-trees.pdf ×
Chekuri Srikan...
homepage.cs.uri.edu/faculty/hamel/courses/2016/spring2016/csc581/lecture-notes/32-decision-trees.pdf

Partitioning the Data Set

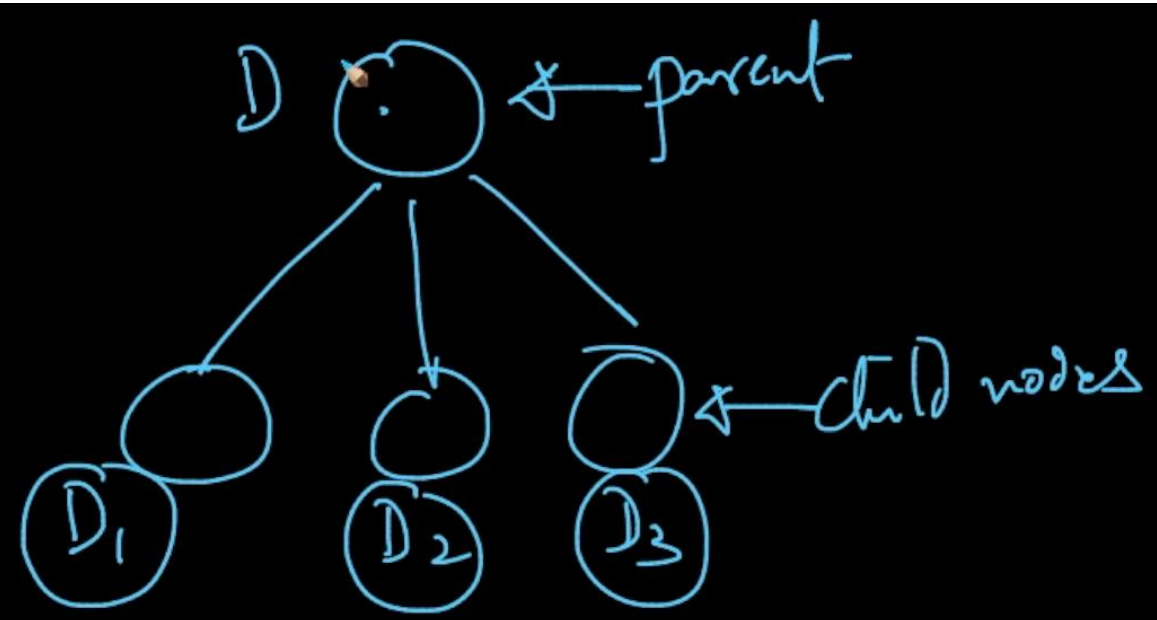


$IG(Y, \text{outlook}) = \left(\frac{5}{7} \times 0.97 \right) - 0.94 = IG$

Weighted entropy after D_1, D_2, D_3 $\leftarrow$ $\left(\frac{5}{14} \times 0.97\right) + \left(\frac{4}{14} \times 0\right) + \left(\frac{5}{14} * 0.97\right)$

$\frac{5}{7} \times 0.97$

$\frac{5}{14} \times 0.97 \times \frac{1}{7}$



$$-H_D(Y) + \left(\frac{|D_1|}{|D|} * H_{D_1}(Y) + \frac{|D_2|}{|D|} * H_{D_2}(Y) + \frac{|D_3|}{|D|} * H_{D_3}(Y) \right)$$

Gini Impurity ~ similar to Entropy

$$I_G(Y) = 1 - \sum_{i=1}^{k} (p(y_i))^2$$

$Y \rightarrow y+$, $y-$

$Y \rightarrow y_1, y_2, y_3, \ldots y_k$

Case 1: $p(y+) = 0.5$, $p(y-) = 0.5$

$I_G(Y) = 1 - (0.25 + 0.25) = 0.5$

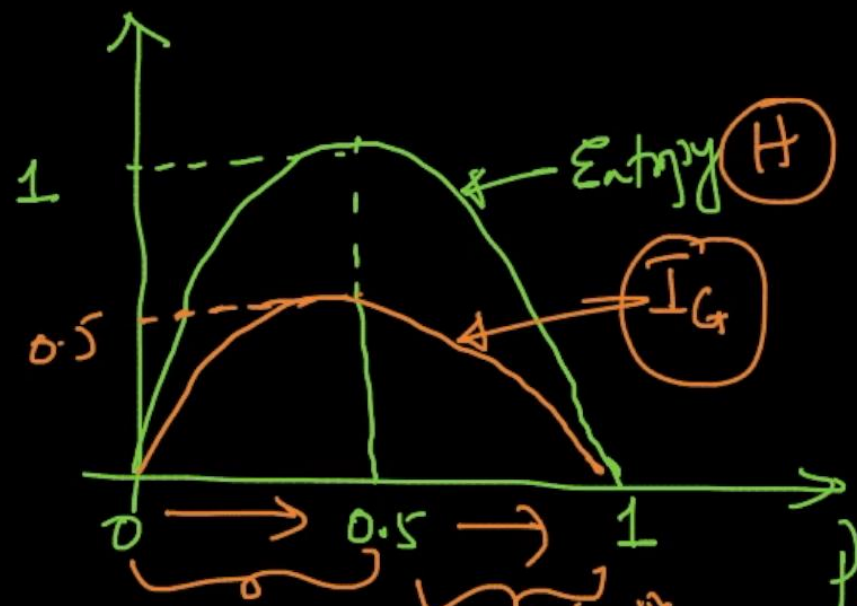
$H(Y) = 1$

Case 2: $p(y+) = 1$, $p(y0) = 0$

$I_G(Y) = 1 - (1 + 0) = 0$

$H(Y) \downarrow$

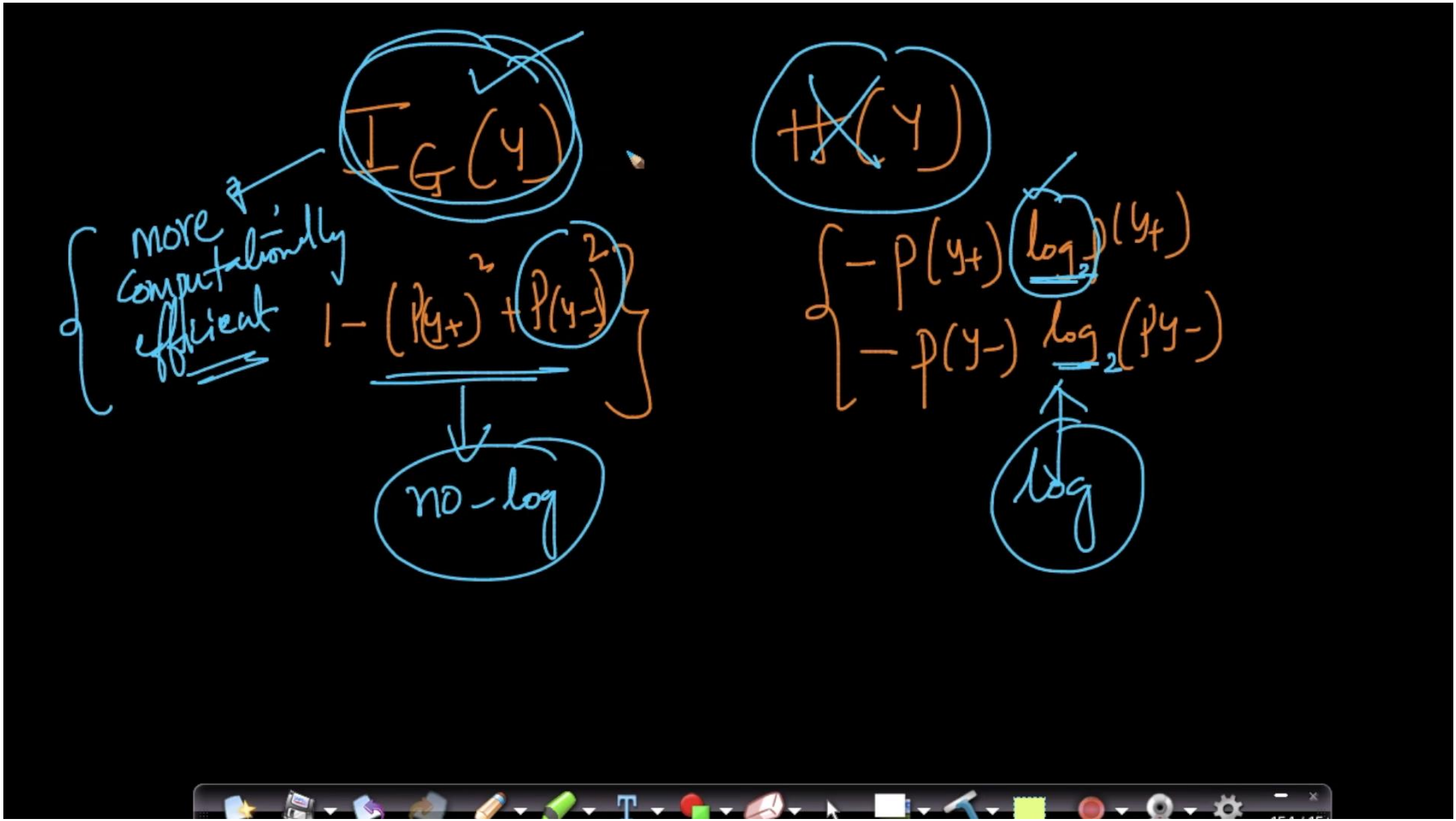
2-Category Case :- y_+ , $y-$ $\quad P(y_+) = 1 - P(y-)$

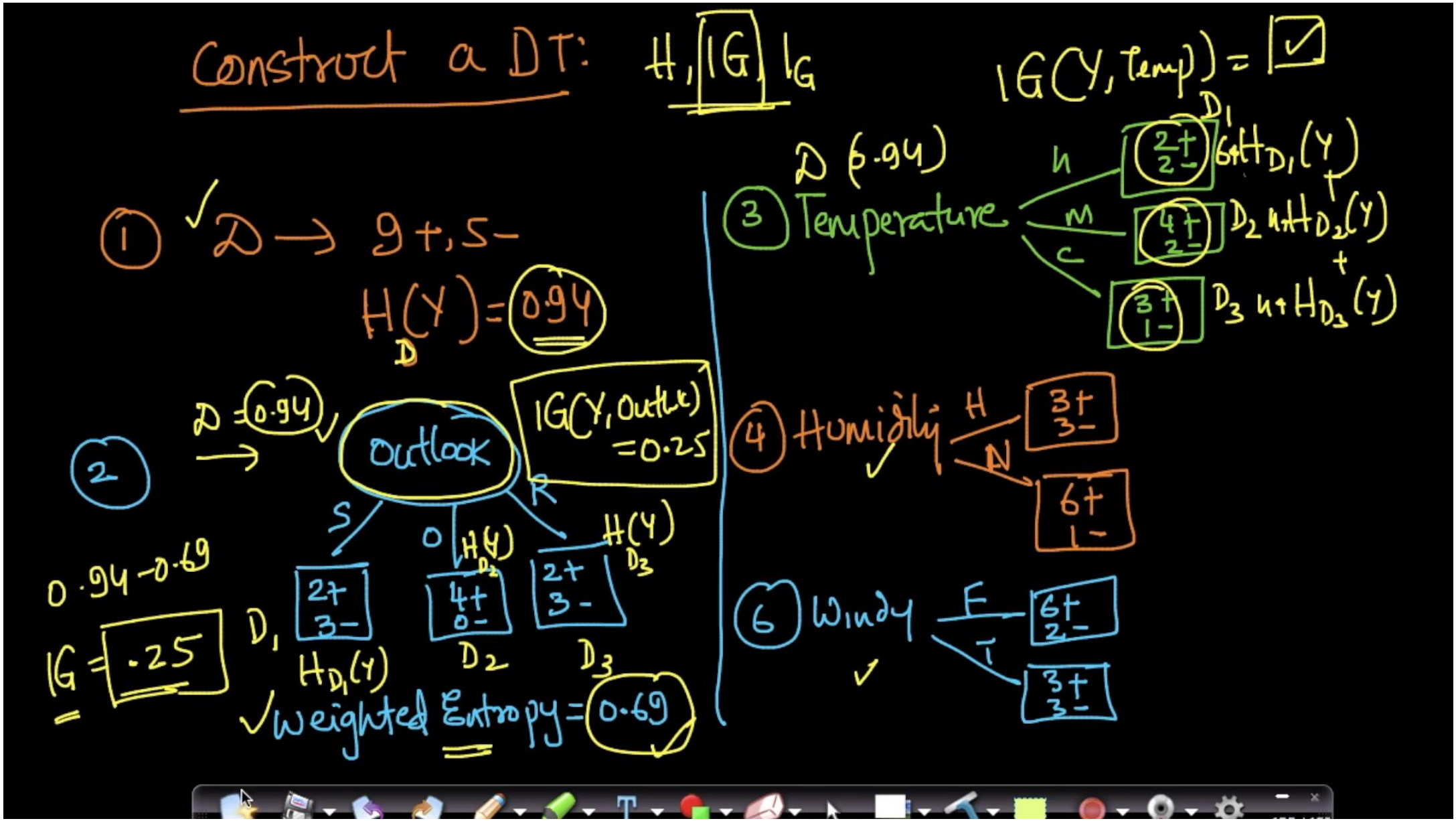


P(y+) ↑

P(y+) ↑
H ↑
I_G ↑

P(y+) ↑
H ↓
I_G ↓


$I_G(Y)$
more computationally efficient
$1 - (P(y+)^2 + P(y-)^2)$
no-log
$H(Y)$
$-P(y+)\log_2(y+)$
$-P(y-)\log_2(Py-)$
log


Construct a DT: H, IG, IG
IG(Y, Temp) = ☑
① ✓ D → 9+, 5−
H(Y) = 0.94
D
D (0.94)
③ Temperature
H
M
C
2+ 2− D1 6+ H_D1(Y)
4+ 2− D2 4+ H_D2(Y)
3+ 1− D3 4+ H_D3(Y)
+
+
② D = 0.94 ✓
Outlook
IG(Y, Outlook) = 0.25
S
O
R
H(Y) D2
H(Y) D3
0.94 − 0.69
IG = .25
D1 2+ 3− H_D1(Y)
D2 4+ 0− D2
2+ 3− D3
✓ Weighted Entropy = 0.69
④ Humidity ✓
H
N
3+ 3−
6+ 1−
⑥ Windy ✓
F
T
6+ 2−
3+ 3−

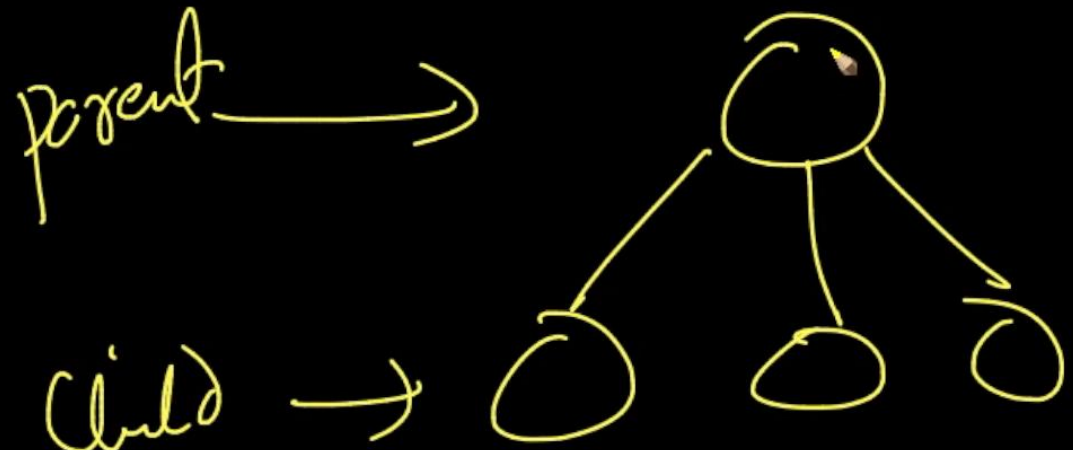
$$IG(Y,f) = H_D(Y) - \sum_{i=1}^{2} \frac{|D_i|}{|D|} * H_{D_i}(Y)$$

→ Choosing the root node

D →

- IG(Y, outlook) = 0.25
- IG(Y, Temp) = —
- IG(Y, Humidity) = —
- IG(Y, Windy) = —

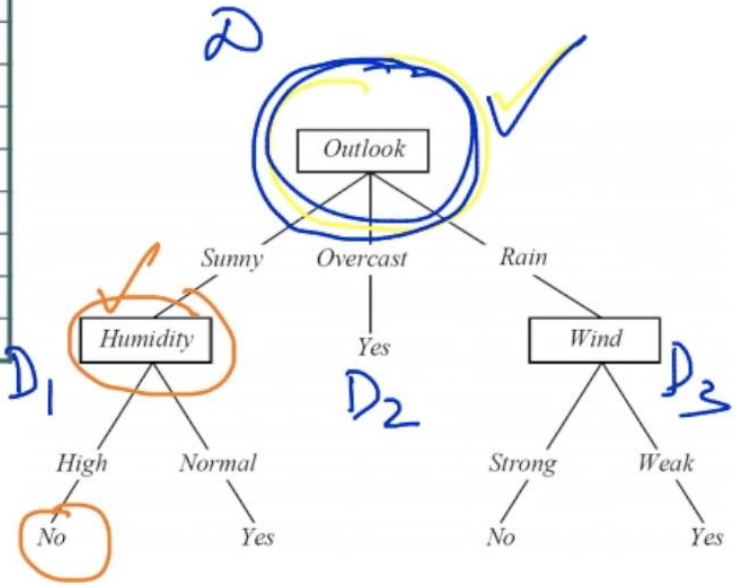
$IG(Y, f) =$ entropy @ parent level
− weighted entropy @ child level

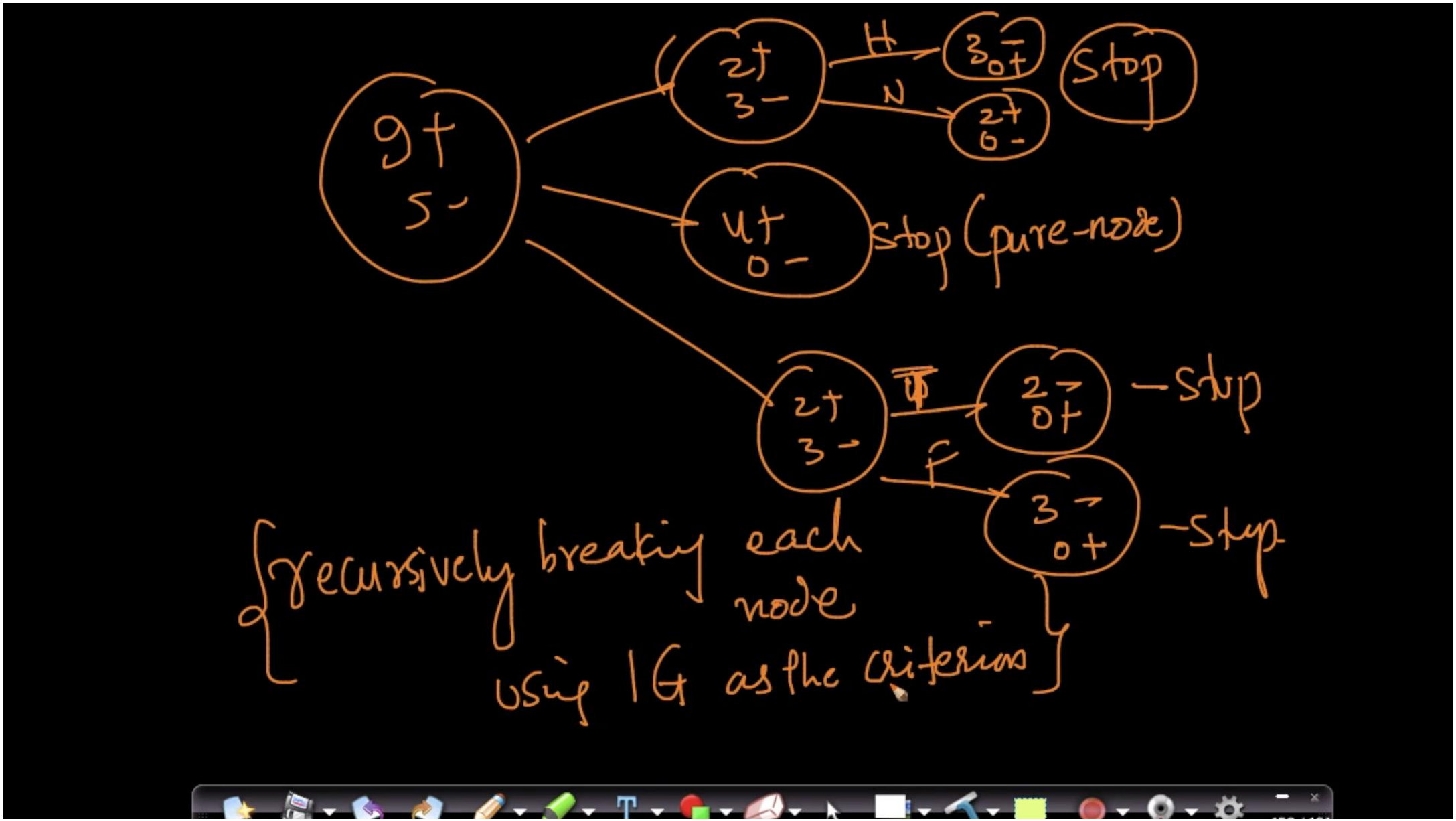


Contents - Google Docs | 32-decision-trees.pdf | play tennis decision trees | Decision Tree Learning | Chekuri Srikan...
homepage.cs.uri.edu/faculty/hamel/courses/2016/spring2016/csc581/lecture-notes/32-decision-trees.pdf

Play Tennis Example

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No



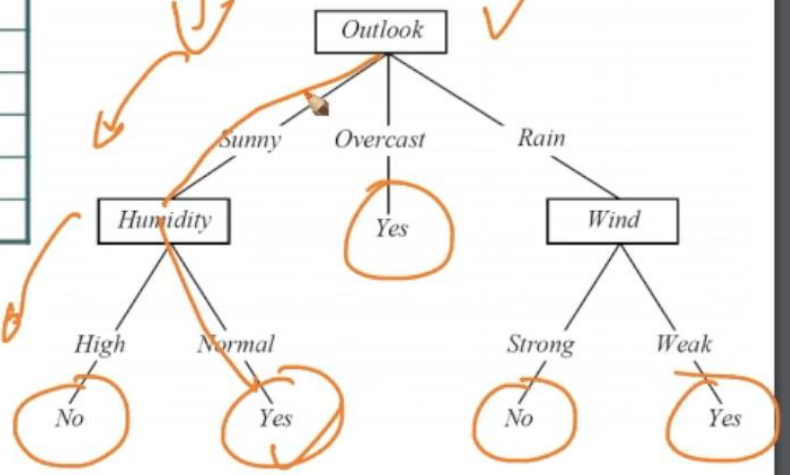

9+ 5-
2+ 3-
H
N
3- 0+
2+ 0-
Stop
4+ 0-
Stop (pure-node)
2+ 3-
T
F
2- 0+
-stop
3- 0+
-stop
{recursively breaking each node using IG as the criterion}

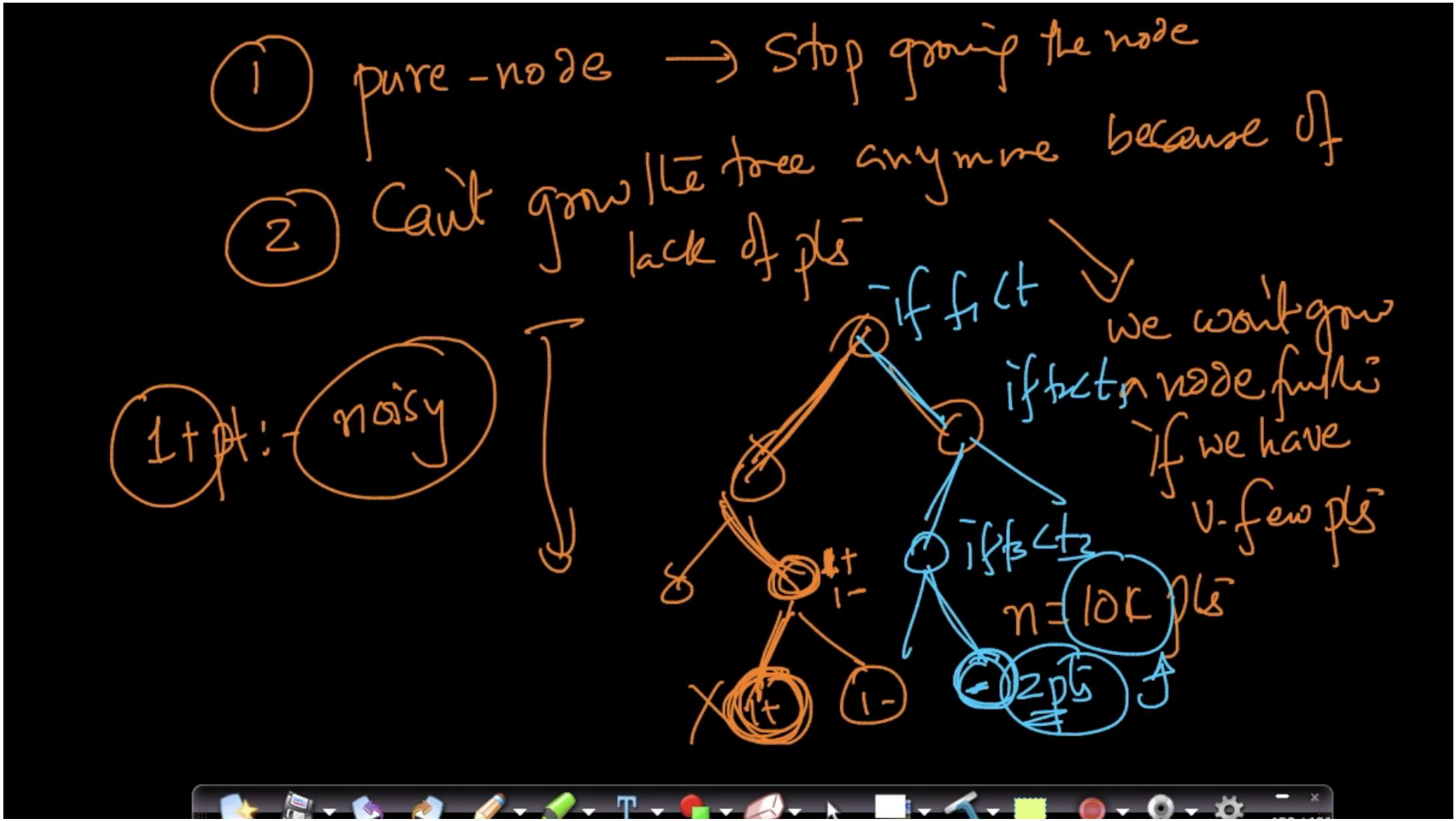
Contents - Google Docs | 32-decision-trees.pdf | play tennis decision trees | Decision Tree Learning | Chekuri Srikan...
homepage.cs.uri.edu/faculty/hamel/courses/2016/spring2016/csc581/lecture-notes/32-decision-trees.pdf

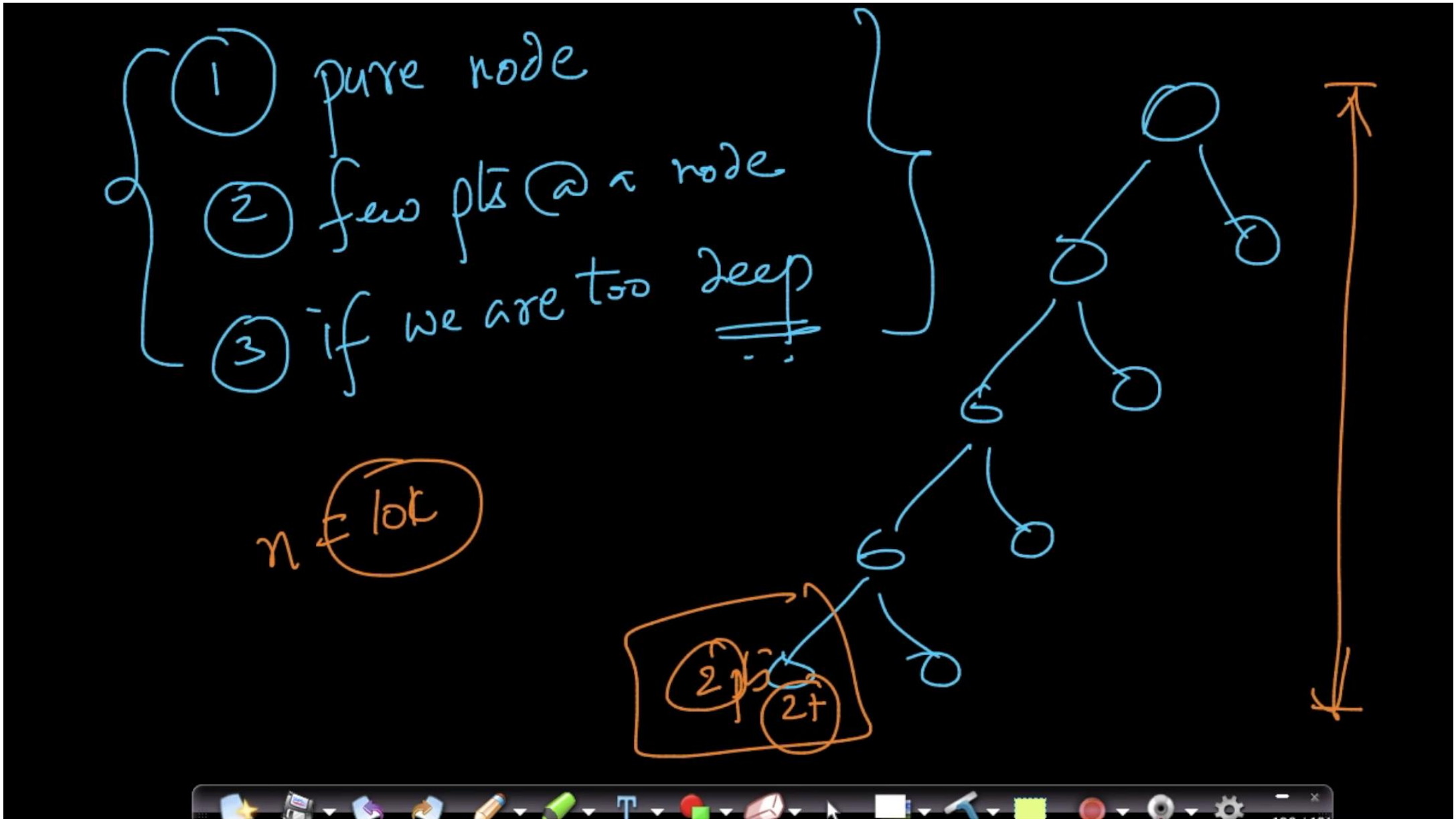
Play Tennis Example

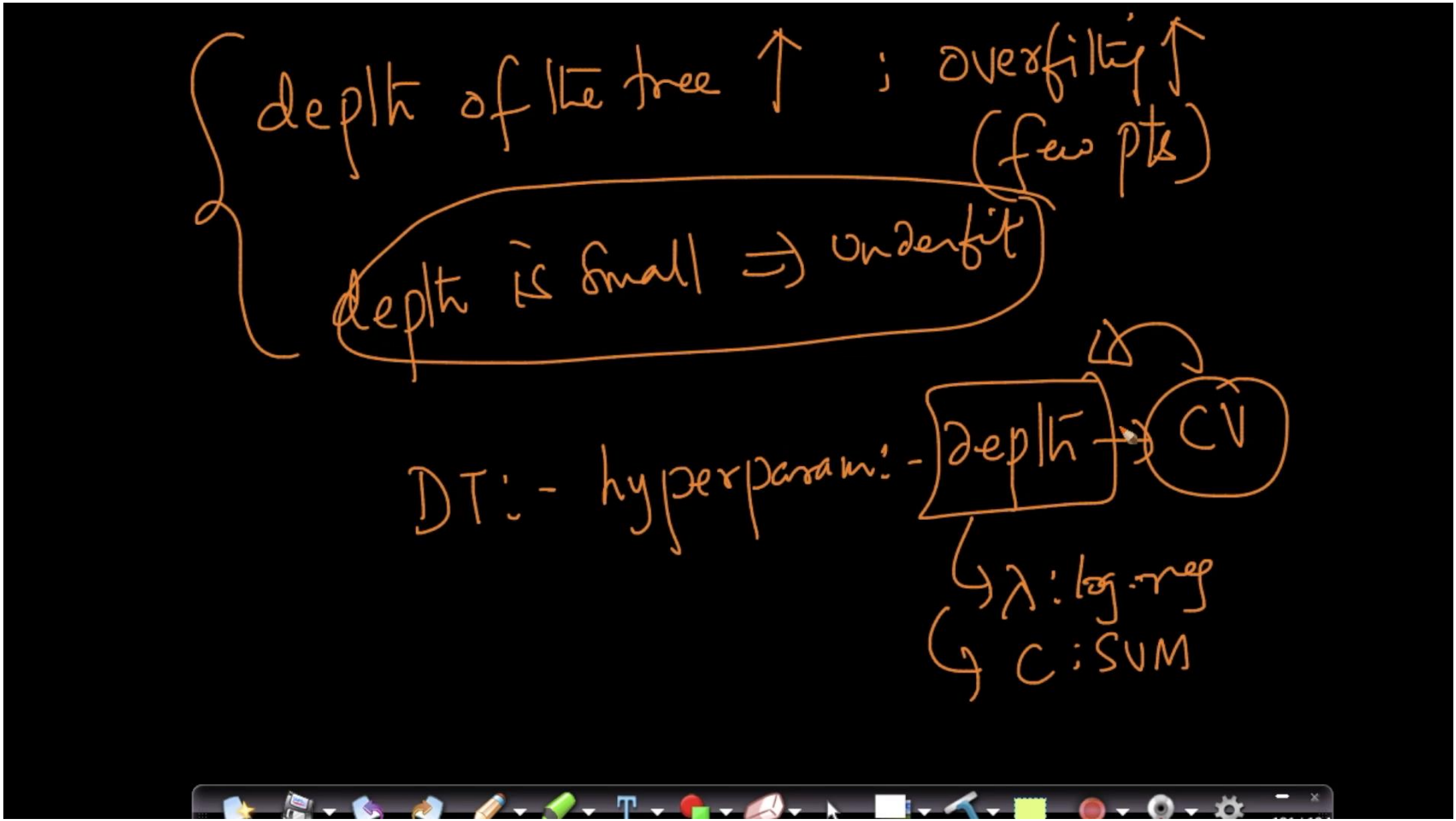
Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

IG as the criterion




1 pure-node → stop growing the node
2 Can't grow the tree anymore because of lack of pts
if fr < t
we won't grow if the current node fulfils
if we have v. few pts
if fr < t2
n = 10k pts
2 pts
1 pt :- noisy
1+
1-
1+
1-


1 pure node
2 few ptrs @ a node
3 if we are too deep
n ∈ lot
2 ptrs
2+


depth of the tree ↑ ; overfitting ↑
(few pts)
depth is small ⇒ underfit
DT :- hyperparam :- depth → CV
λ : log-reg
C : SVM

Splitting numerical features

Construct a DT :- Splitting a node → IG

IG : → entropy
→ Gini impurity → computationally efficient

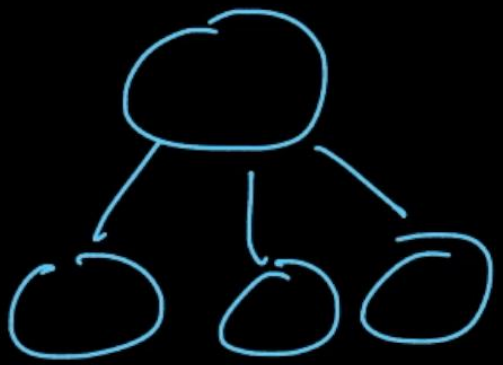
discrete r.v (or) categorical features

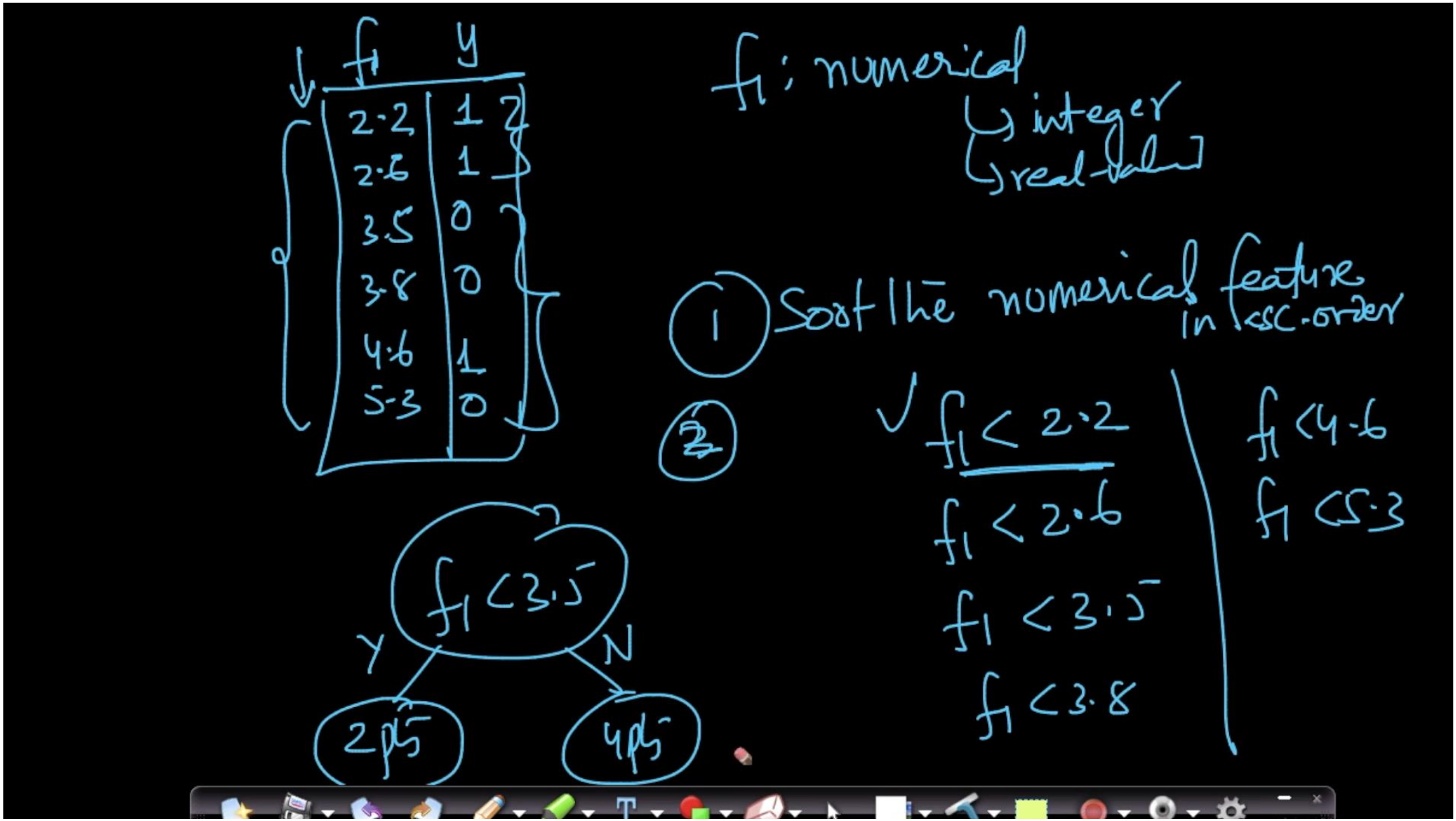
f_1	y
2.2	1
2.6	1
3.5	0
3.8	0
4.6	1
5.3	0

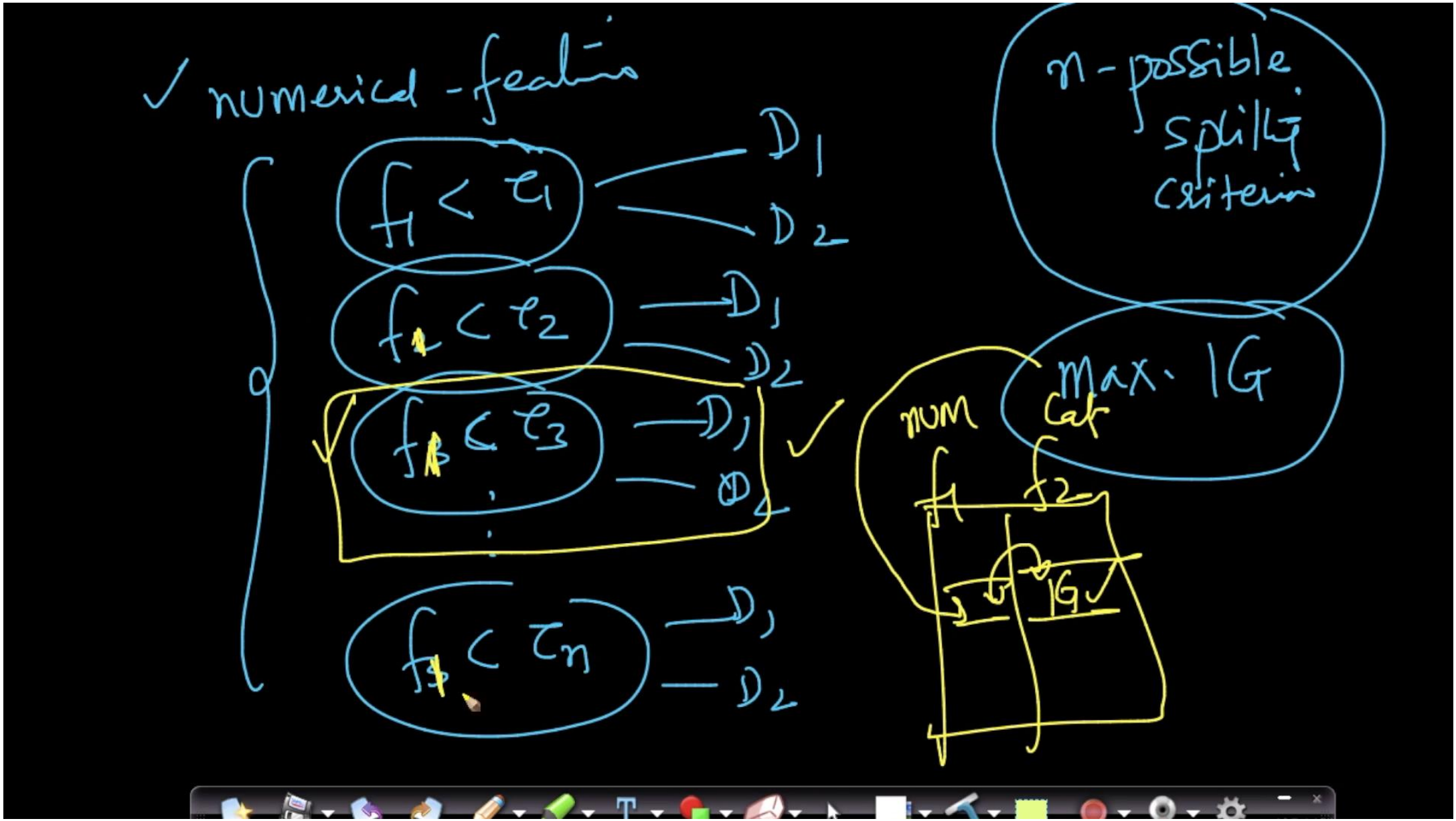
f_1: numerical
- integer
- real-valued

→ Split based on Cat. Var

f_2: 3 Categories



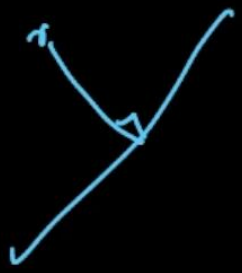

f1 y
2·2 1
2·6 1
3·5 0
3·8 0
4·6 1
5·3 0
f1 ; numerical
integer
real-valued
1 Sort the numerical feature in asc. order
2
f1 < 2·2
f1 < 2·6
f1 < 3·5
f1 < 3·8
f1 < 4·6
f1 < 5·3
f1 < 3·5
Y
N
2 pts
4 pts

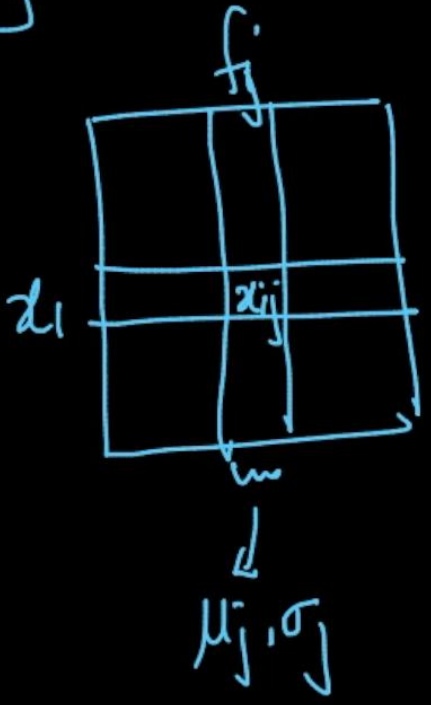

numerical - feature
$f_1 < v_1$
$f_1 < v_2$
$f_1 < v_3$
$f_1 < v_n$
D_1
D_2
n - possible splitting criterion
max. IG
num
cat
IG

Feature Standardization:

dist

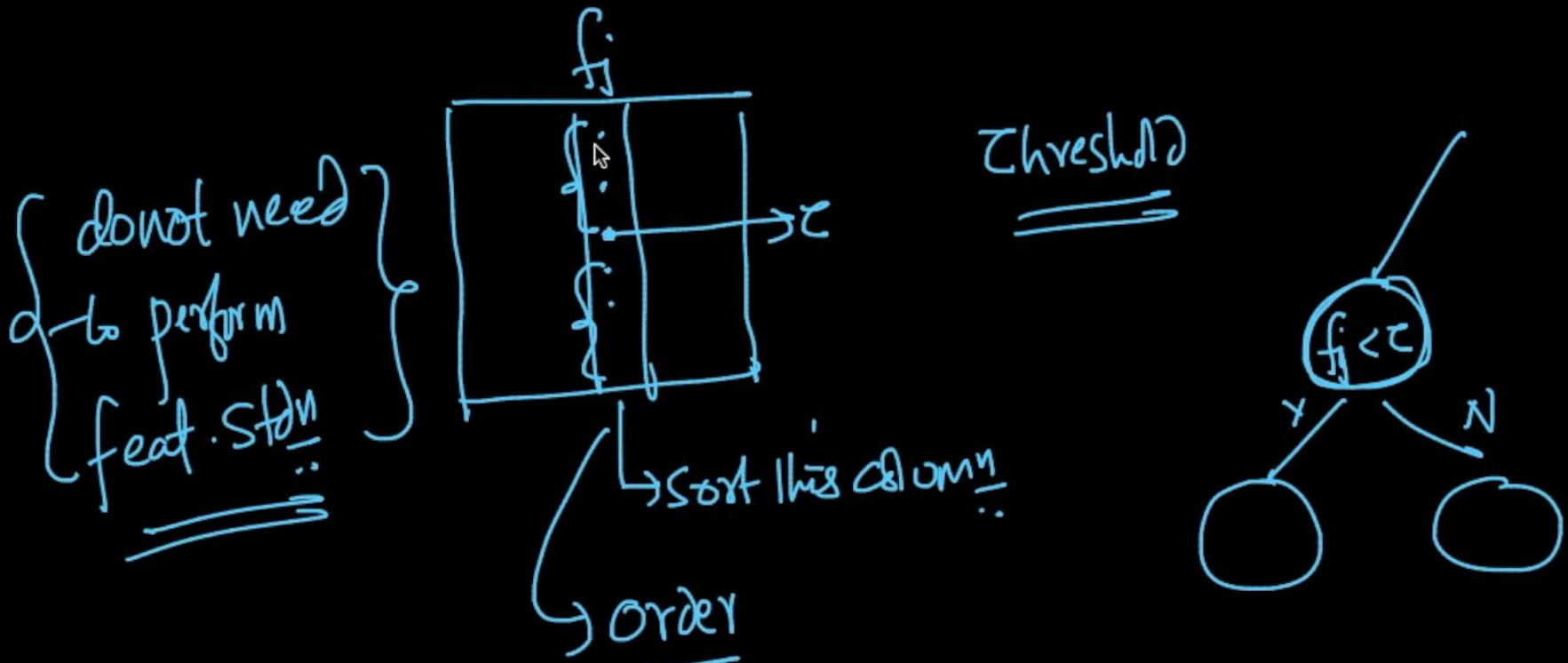
{Logistic regsn
SVM
KNN} → feature stdn





$$x'_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

Decision Trees:- not a distance based method



Categorical features with many categories

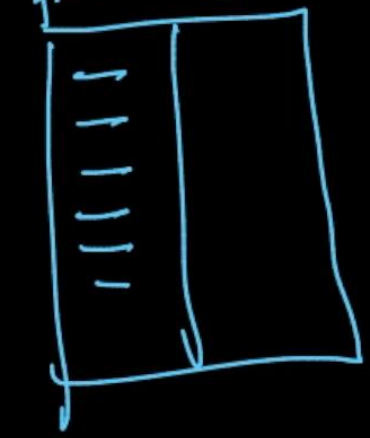
levels

Pincode / Zipcode

1000's

Zipcode

Pincode y



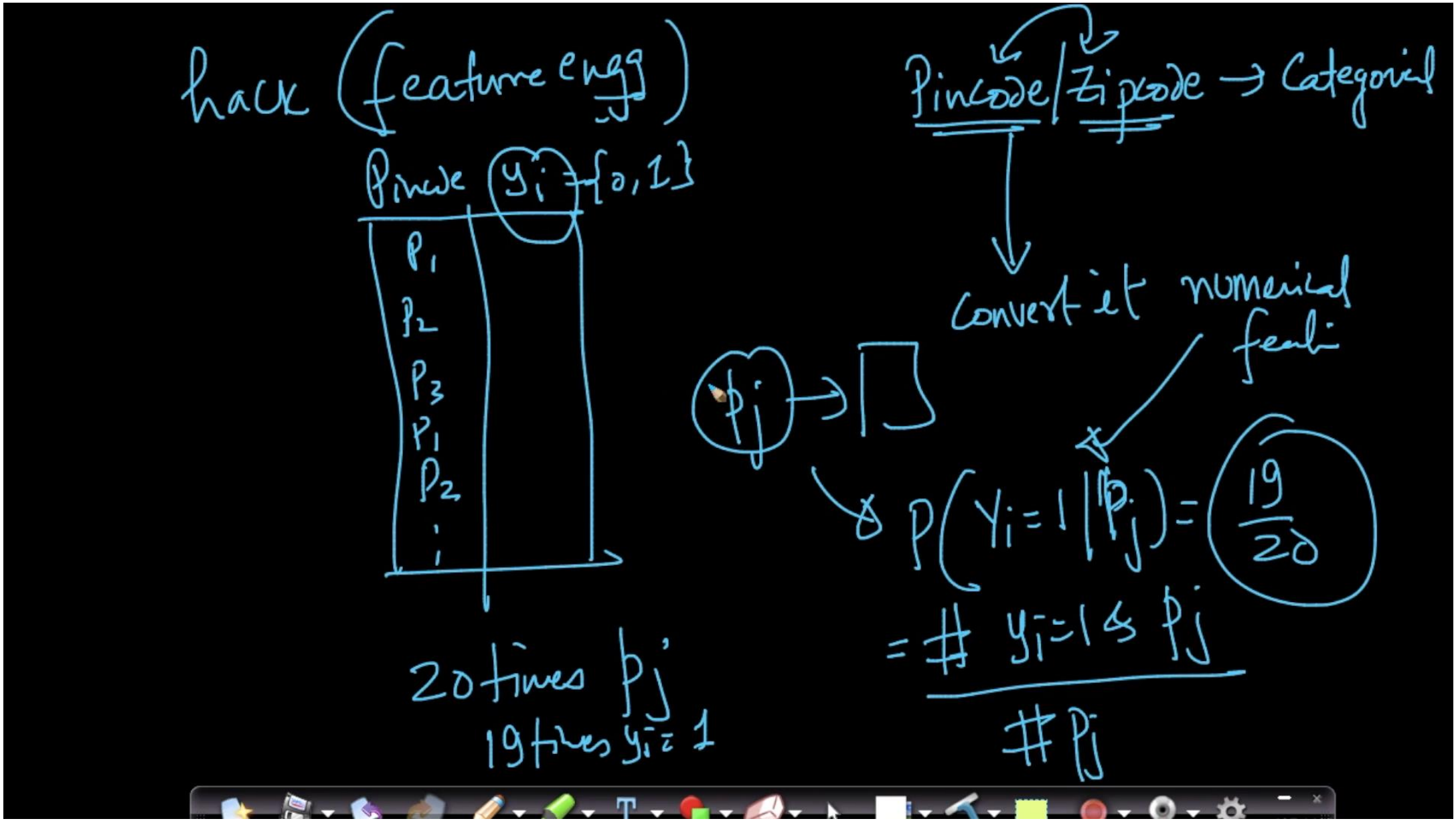
District

Zipcode

D_1 D_2 D_3 ... D_{1000}

Small

ENG 00:29 20-09-2019

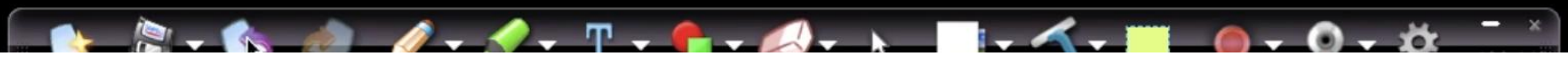

hack (feature engg)
Pincode y_i = {0, 1}
P_1
P_2
P_3
P_1
P_2
20 times P_j
19 times $y_i = 1$
Pincode/Zipcode → Categorical
Convert it numerical feat
$P(Y_i = 1 | P_j) = \frac{19}{20}$
$= \frac{\#\ y_i = 1 \text{ \& } P_j}{\#\ P_j}$

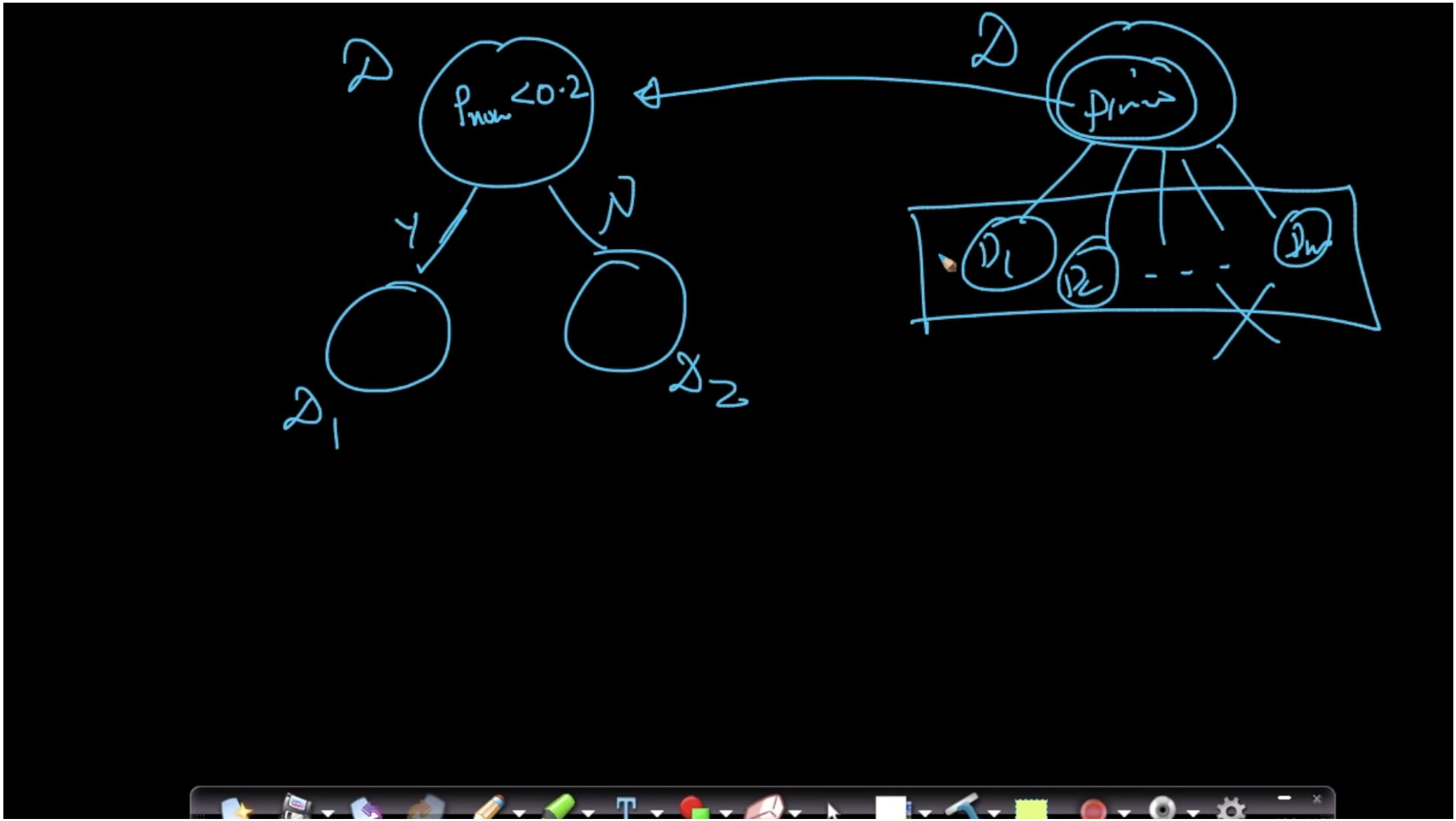
$p_j \rightarrow$ numerical feature $= P(y_i = 1 \mid p_j)$

D → remove pincodes feature

numerical feature

p_i is p_j

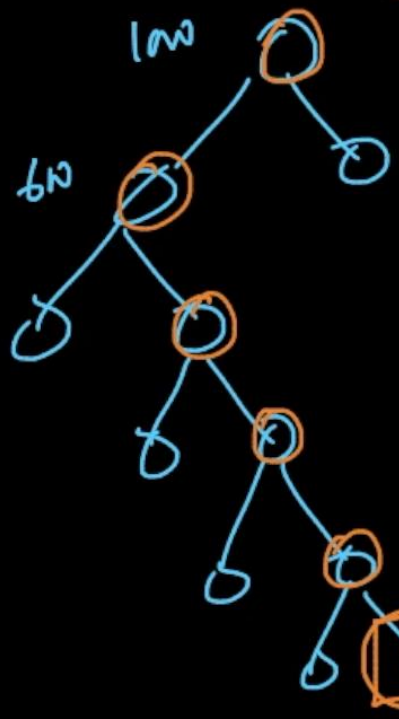



D
$p_{max} < 0.2$
Y
N
D_1
D_2
D
p_{max}
D_1
D_2
D_w

Overfitting & Underfitting:

depth ↑ :- possibility of having very few pts @ a leaf node ↑

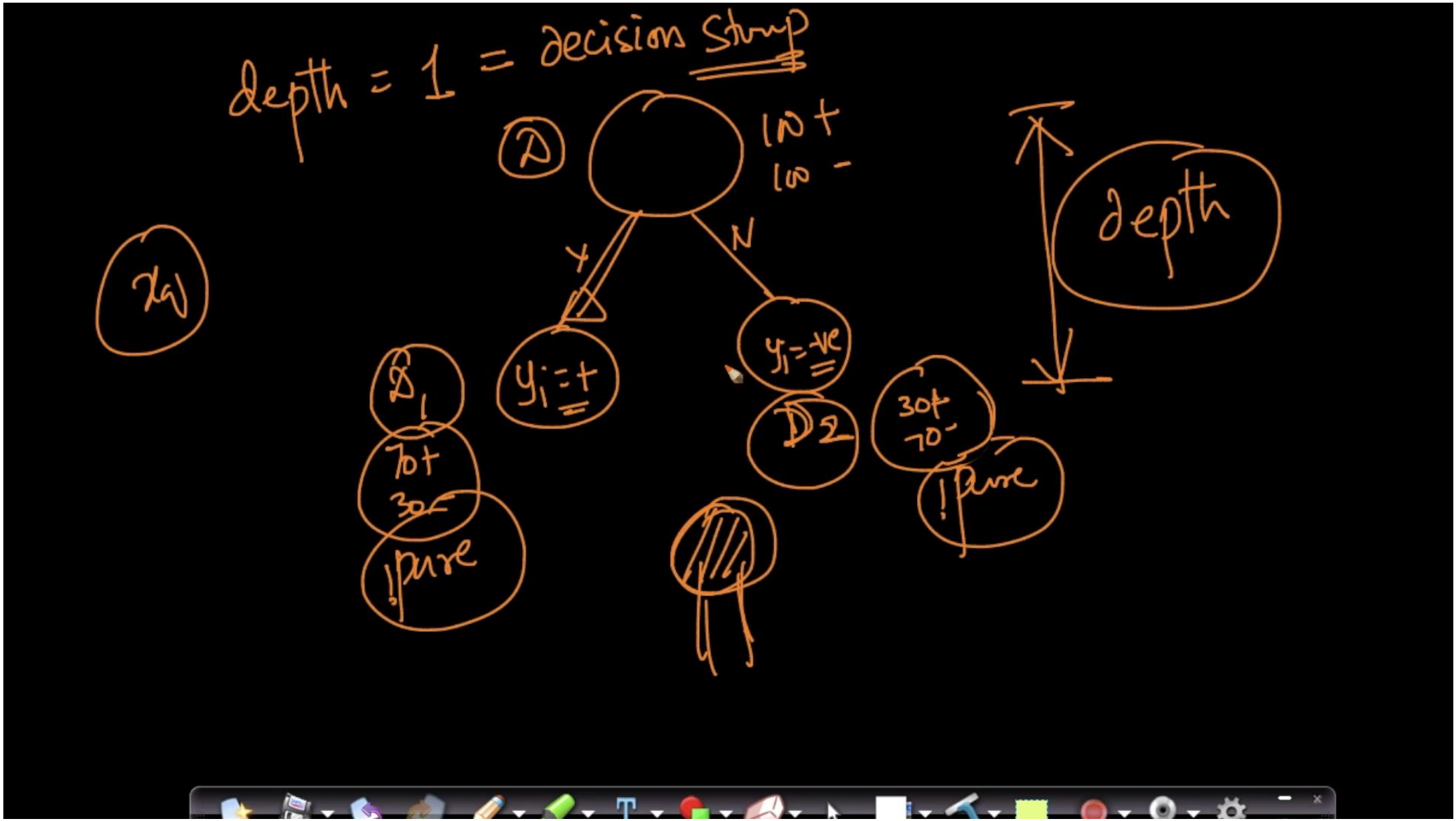
→ interpretability of the model ↓



if ( ) & ( ) & ( ) & ( )

$x_q \rightarrow y_q$

2 pts → noisy → overfitting my model to noise

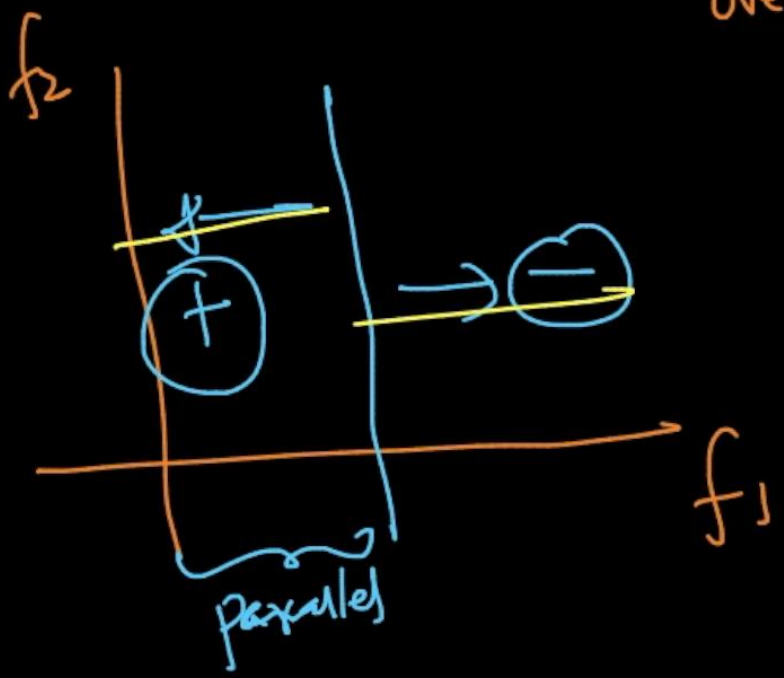

depth = 1 = decision Stump
100 +
100 -
Y
N
depth
y_i = +
y_i = -ve
70+
30-
!pure
30+
70-
!pure

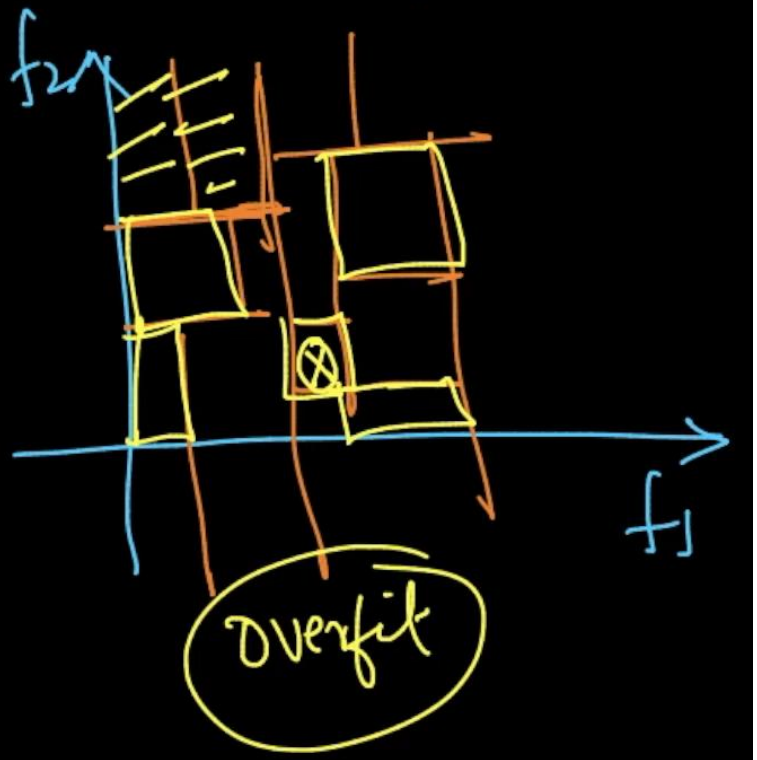
depth :- CV

too deep
↓
overfit

too shallow
↓
underfit

geom:





Train & Run time complx:

Train: $\sim O(n(\lg n)d)$ (Time)

sorting, evaluate IG

$n = \#\text{pts } D_{\text{Train}}$

$d = \dim$.

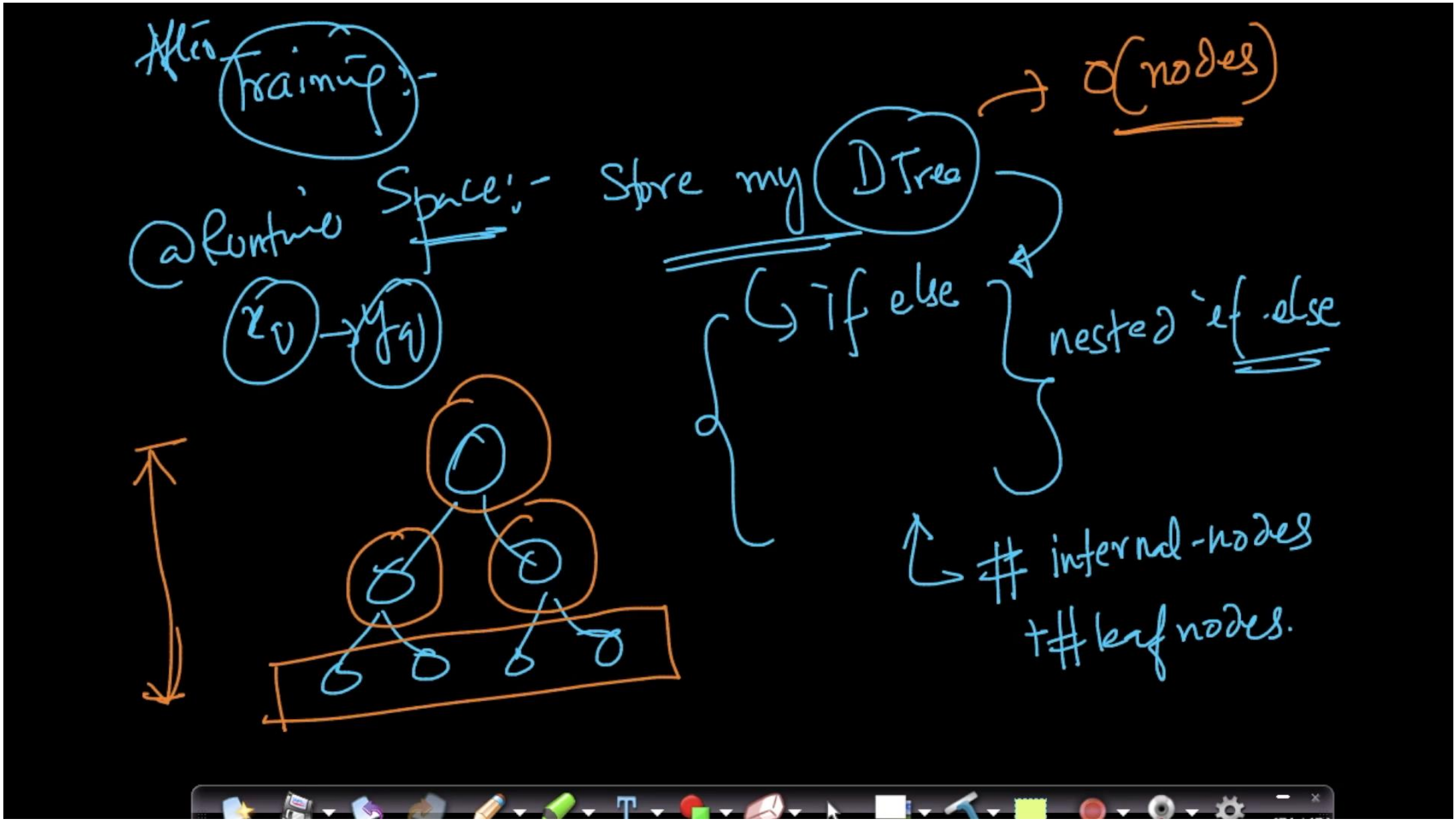
numerical features: - (threshold)

↳ algorithmic methods

$n \lg n$ → sorting

$O(n \lg n d)$

large d


After training :-
@Runtime Space :- Store my DTree
O(nodes)
if else
nested if else
internal-nodes
+# leaf nodes.

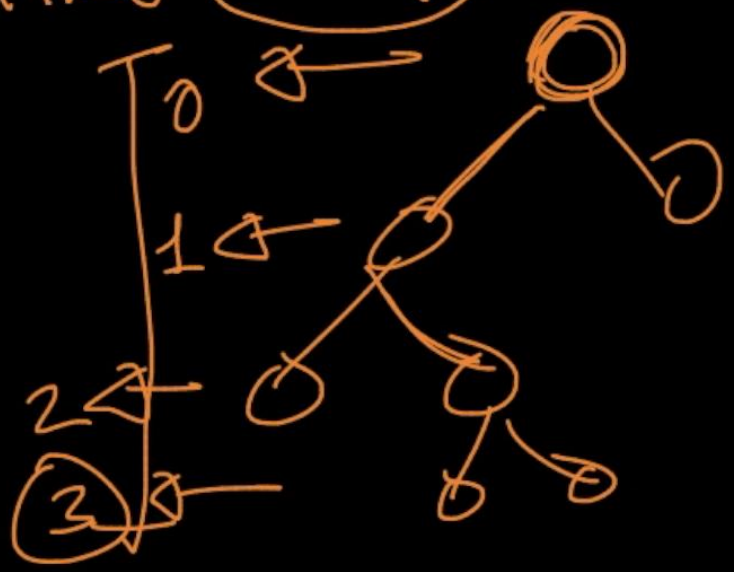
depth = 300

depth ↑ ⇒ interpret ↓

runtime Space :- reasonable.

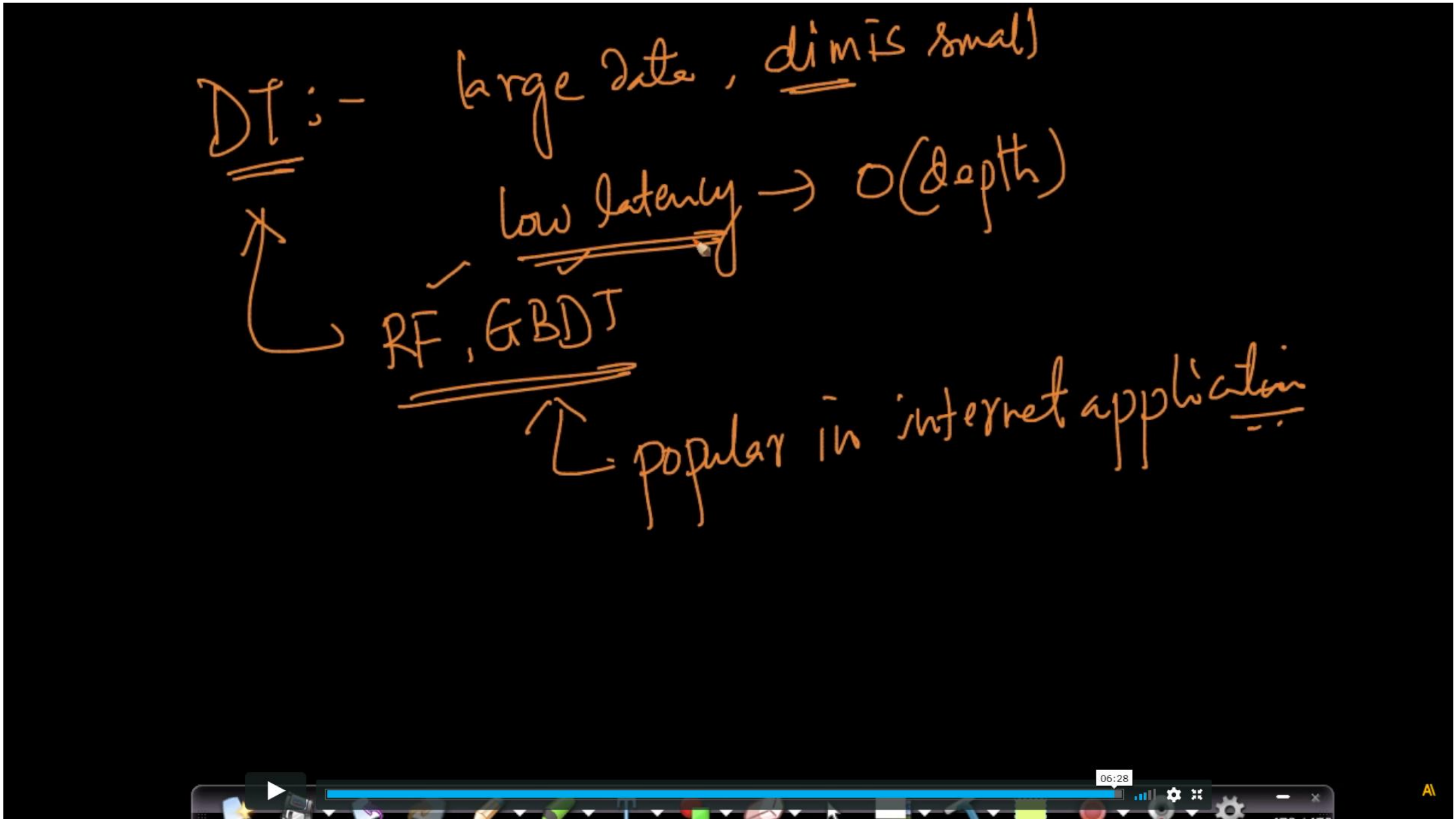
runtime Time Complexity :- $z_q \rightarrow y_{qv}$

O(depth)



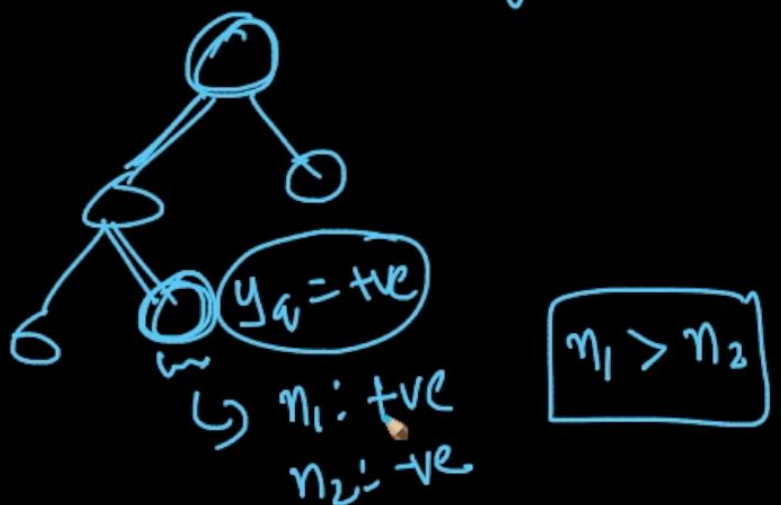
k = max-depth of any leaf-node

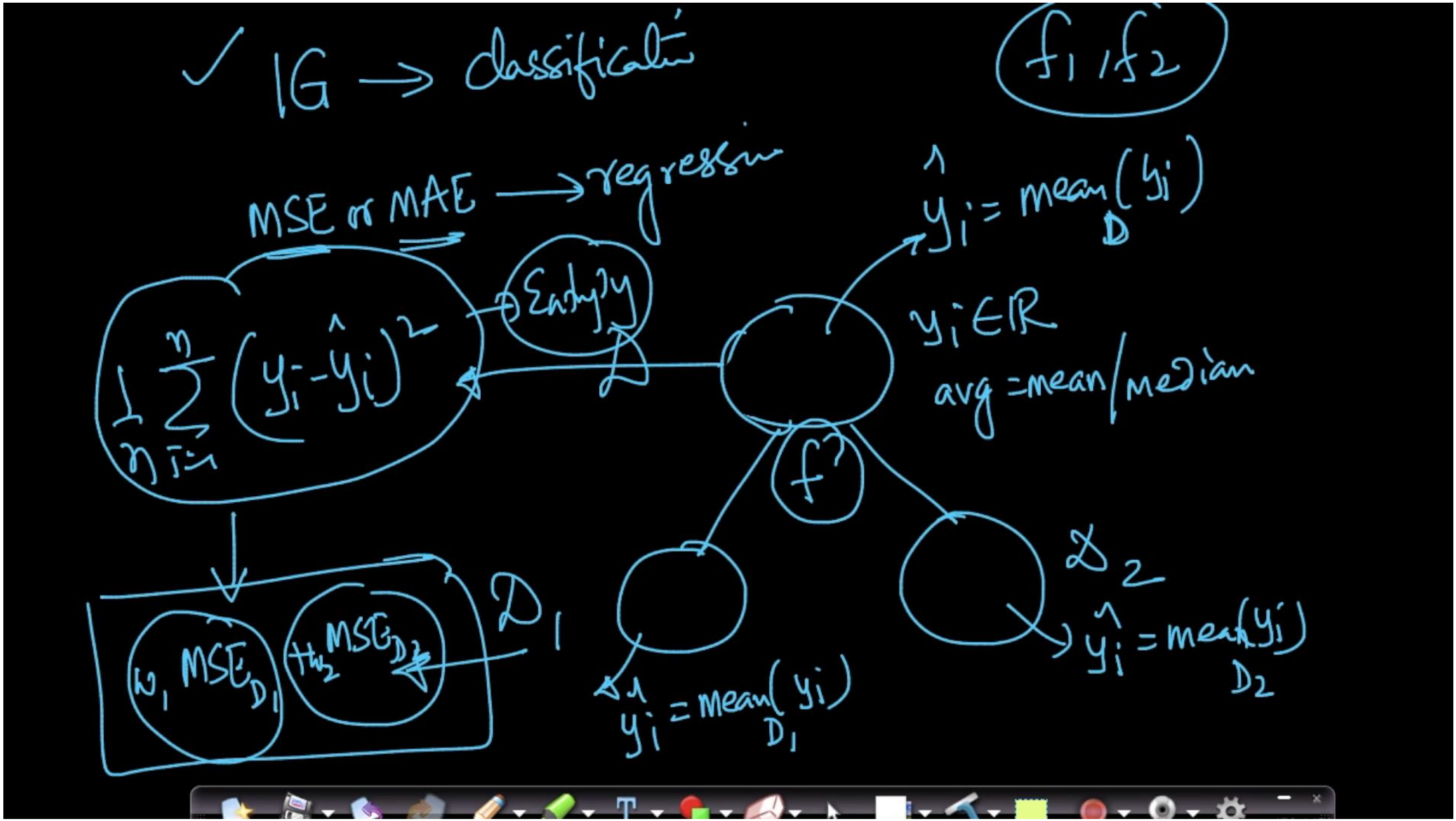
atmost 3 Comparisons


DT :- large data, dim is small
low latency → O(depth)
RF, GBDT
popular in internet application
06:28

Regression using Decision Trees:

DT → Classificati




IG → classification
f1 / f2
MSE or MAE → regression
Entropy
$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
$\hat{y}_i = \text{mean}_{D}(y_i)$
$y_i \in R$
avg = mean / median
f?
D_1
w_1 MSE$_{D_1}$
w_2 MSE$_{D_2}$
$\hat{y}_i = \text{mean}_{D_1}(y_i)$
D_2
$\hat{y}_i = \text{mean}_{D_2}(y_i)$

Classfn: fi :- reducing entropy :- "0"

regrsn: fi :- reduce MSE → "0"

$$MSE_J(y_i) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

MAE = MAD → "0"

$$\text{Median}(|y_i - \hat{y}_i|)$$

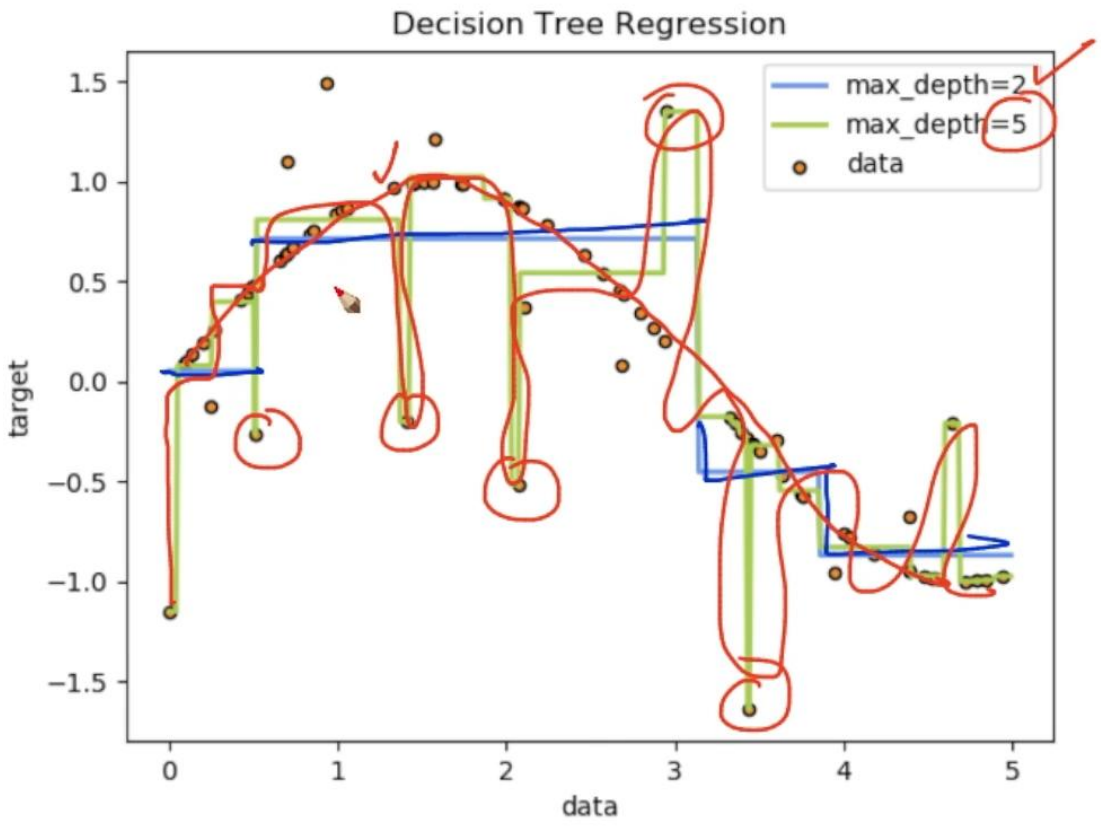
Contents - Google Dr | 32-decision-trees.pd | play tennis decision t | Decision Tree Learnir | 1.10. Decision Trees | sklearn.tree.Decision | Chekuri Srikan...

scikit-learn.org/stable/modules/tree.html

Classfn/Regn.

↓ depth = Underfit

depth↑ = overfit



Some advantages of decision trees are:

Cases: (DT)

✓ imbalanced data:- balance it → upsampling
↳ Recessut

90% +ve
10% -ve ↳ impacts Entropy calc / MSE

large 'd': → @ each node, Split

↓

each features IG

↓

Time complx to train DT increases

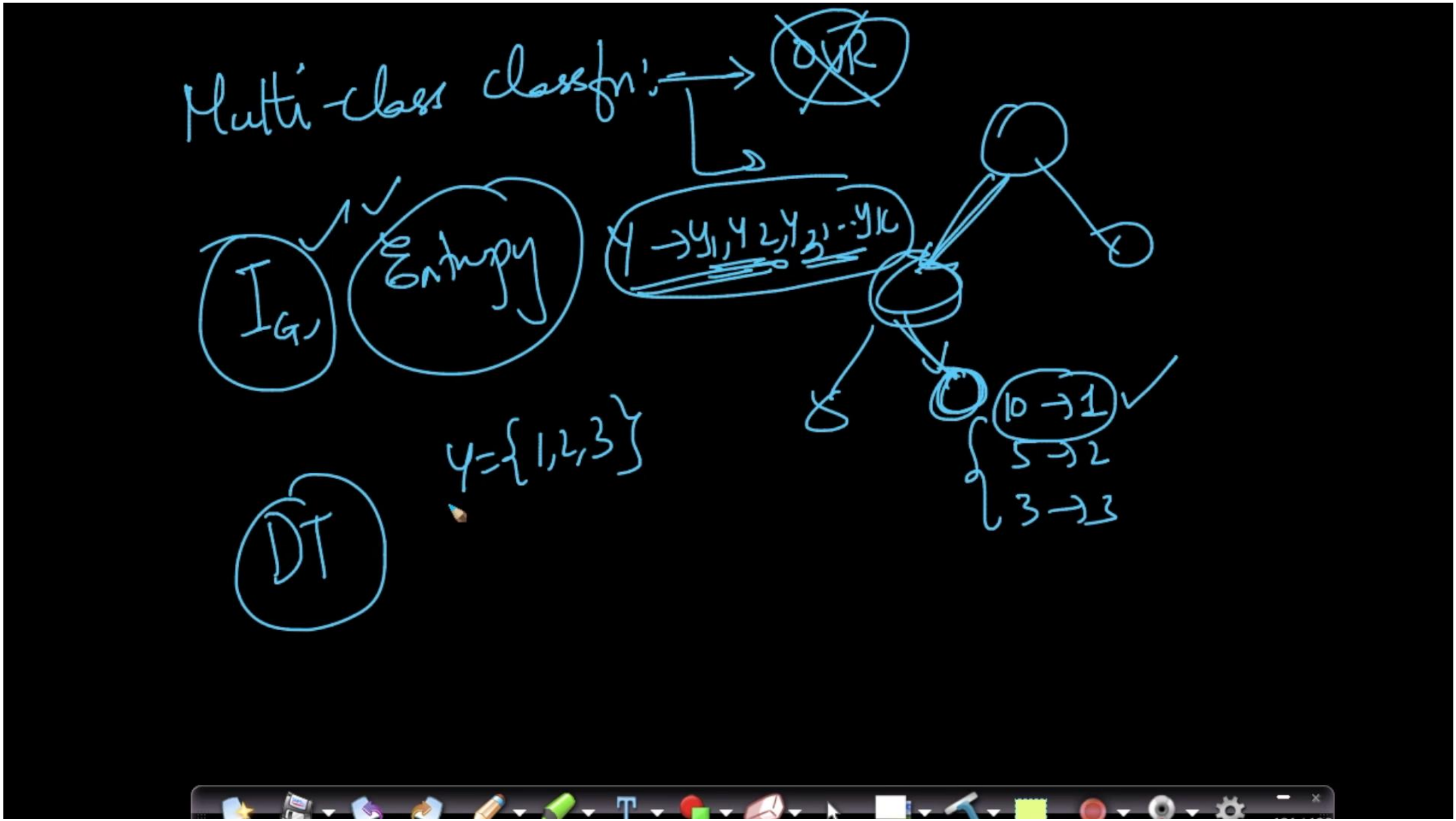
Cat: feat → avoid one-hot encoding

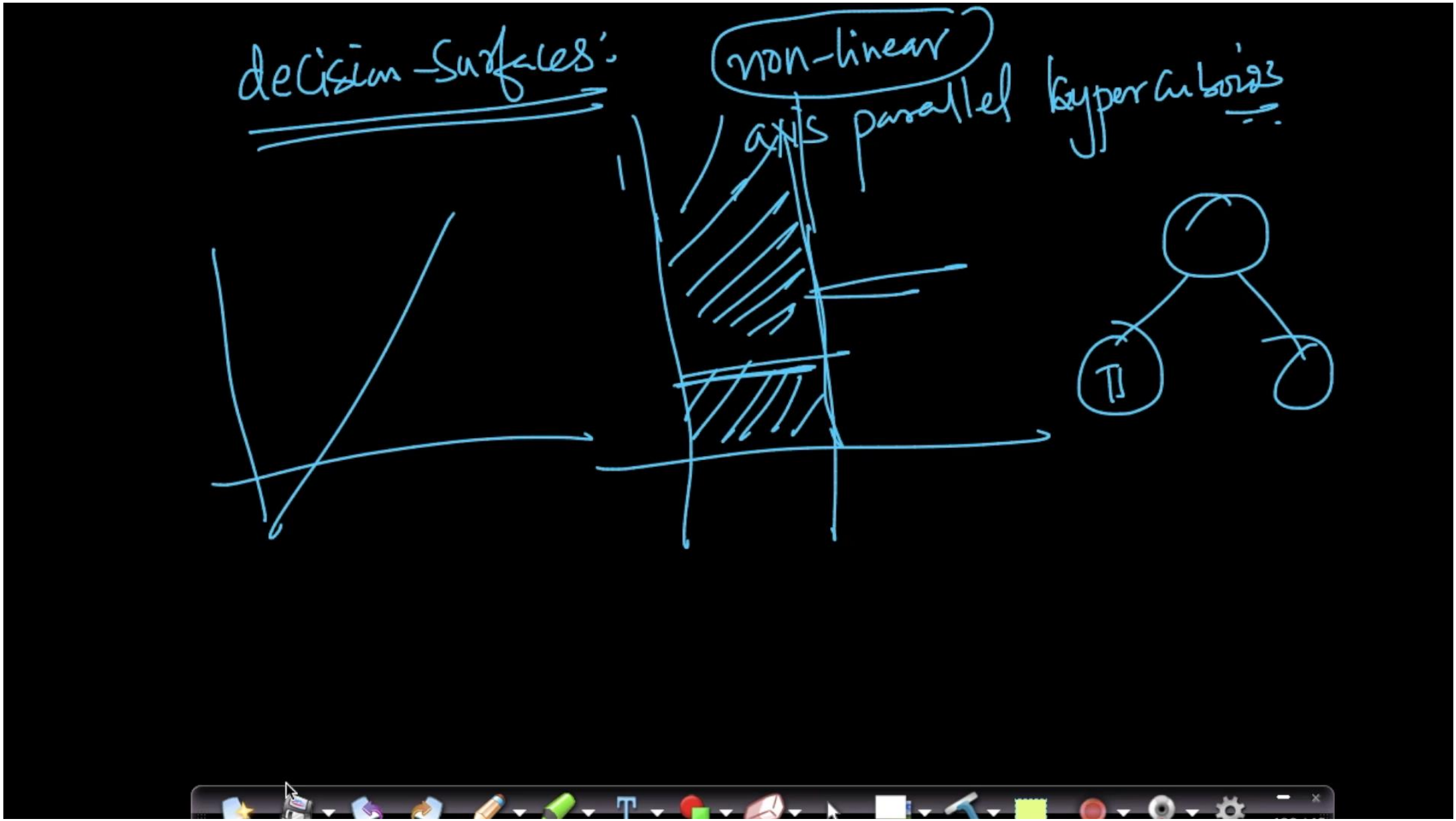
(f) 1 → 10 features → dim ↑

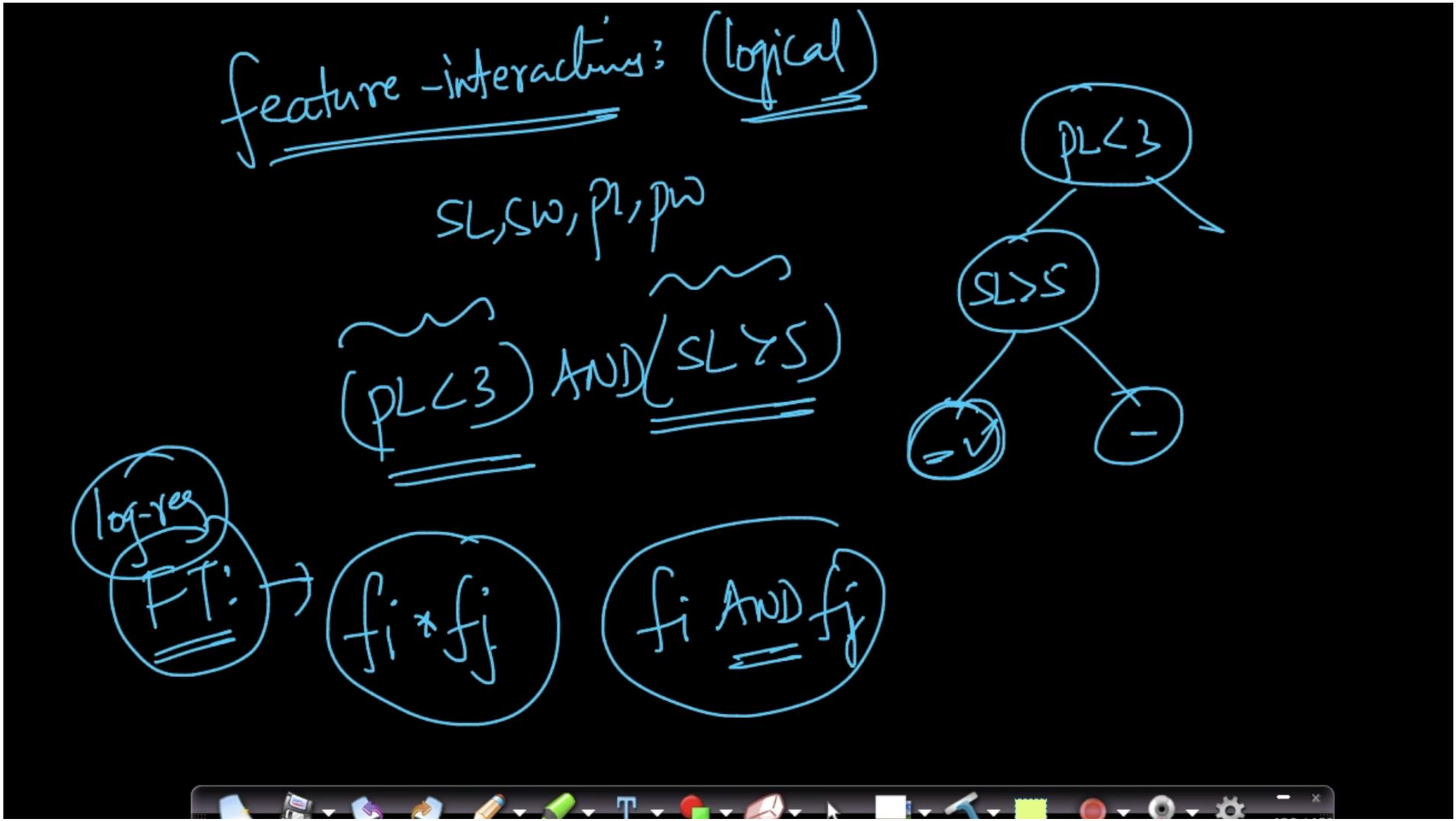
✓ Categorical feat → numerical feat
lots of levels

✓ $P(y_i=1 \mid f=c_1)$ ✓

Similarity Matrix :- DT need the features explicitly


Multi-class classfn:
OVR
Entropy
I_G
$Y \rightarrow y_1, y_2, y_3 \ldots y_k$
$Y = \{1,2,3\}$
DT
10 → 1
5 → 2
3 → 3

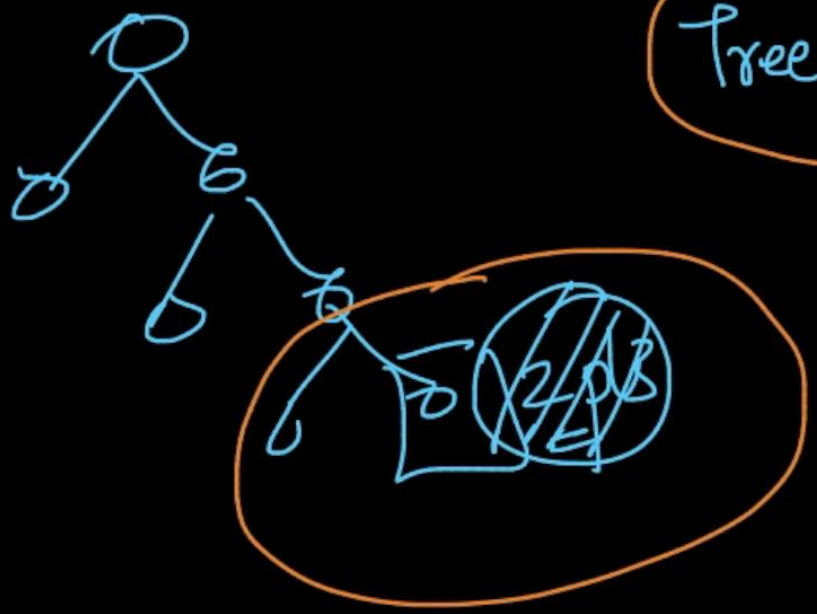

decision-surfaces:
non-linear
axis parallel hyper cuboids
T1


feature-interactions: (logical)
SL, SW, PL, PW
(PL<3) AND (SL>5)
PL<3
SL>5
-
-
log-reg
FT:
fi * fj
fi AND fj

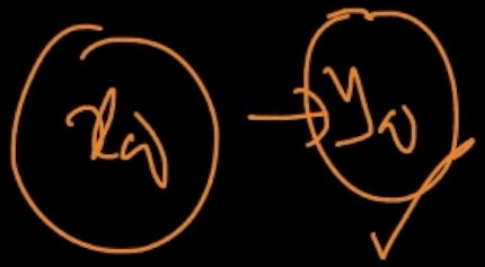
Outliers: depth ↑ :- outlier will impact

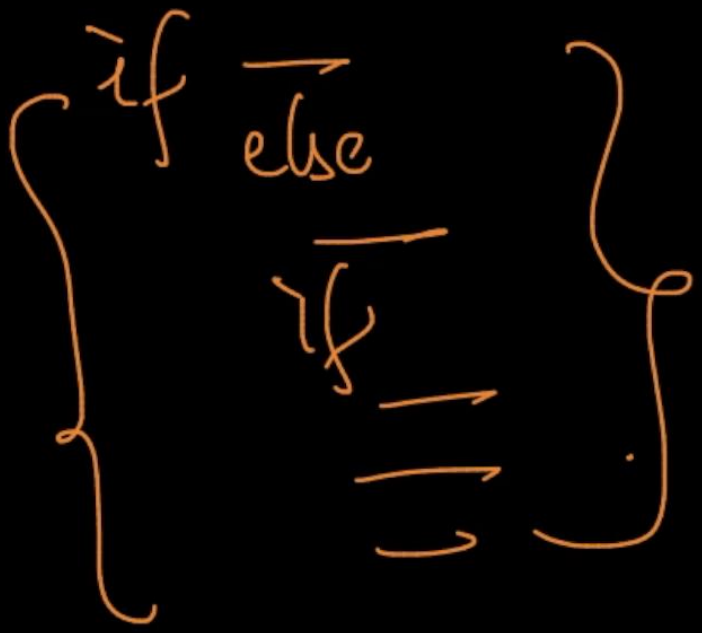
↓

Tree Unstable



Interpretability:-

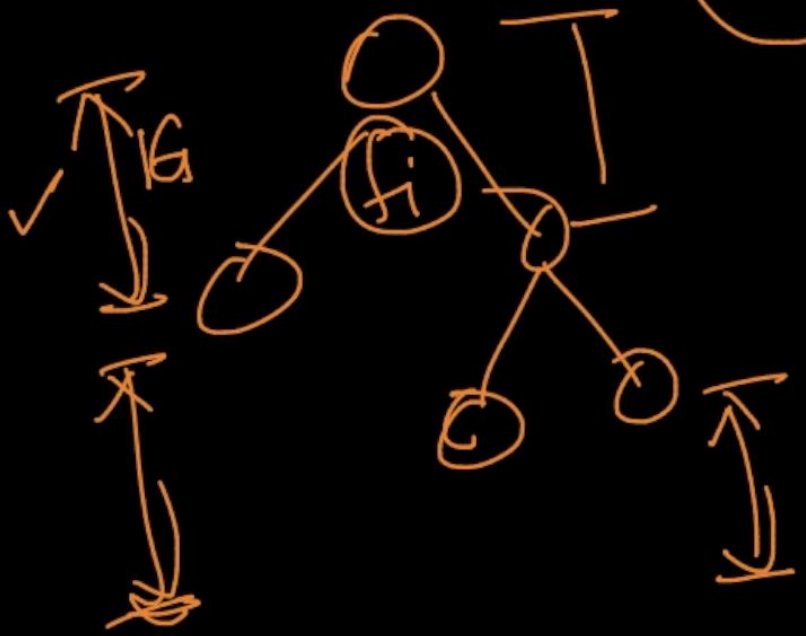




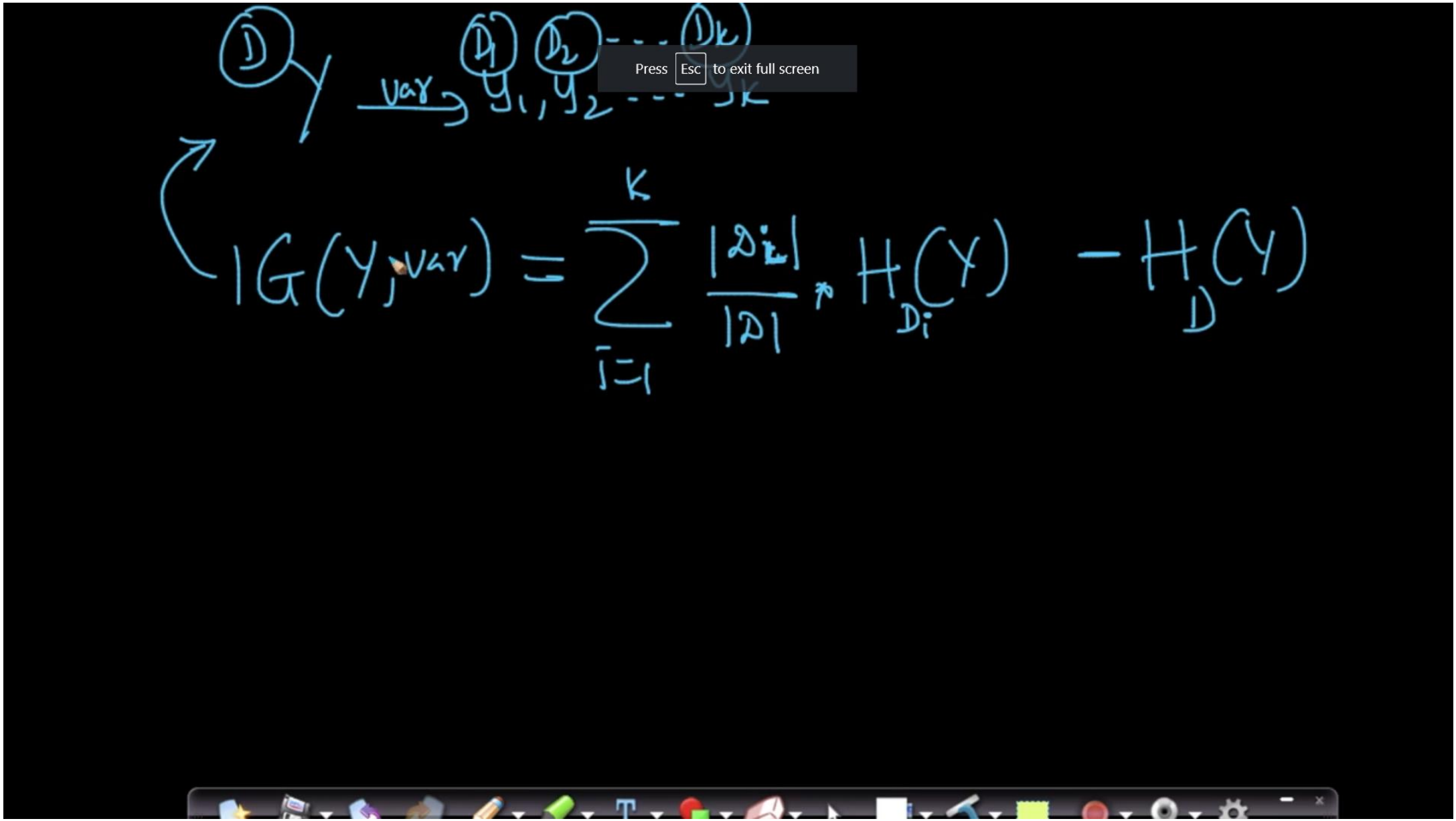
Feature Importance:

fi → reduction in H or I_G

{ Sum up (normalized) reductions in H due to fi



Sklearn → fi → F.I



Press Esc to exit full screen