Lecture 8: Non-Parametric Methods Part 2 (KNN and Decision Trees)

Feb 8, 2023

CIS 4190/5190

Spring 2023

Administrivia

- HW2 due tonight at 8 p.m.
- HW3 released tonight / tomorrow morning. (logistic regression, kNN, Decision trees)
    - PS: we will likely wrap up decision trees for first half of Monday

- Announcements on next quiz, and tomorrow's recitation tonight.

Optional Extra Readings: kNN and Decision Trees

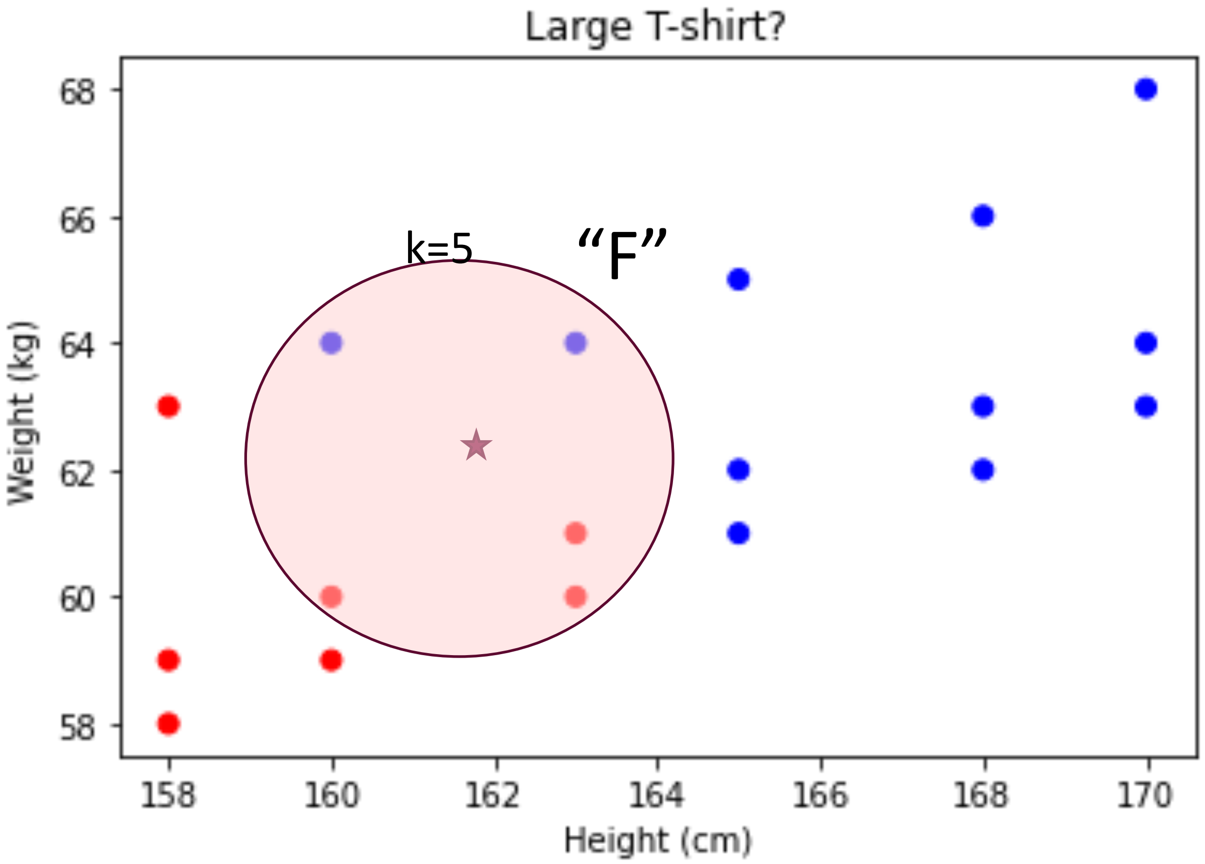
- Bishop, Pattern Recognition and Machine Learning, Ch 2.5:
  - https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf
- Tom Mitchell, Machine Learning Textbook, Ch 3: http://www.cs.cmu.edu/~tom/files/MachineLearningTomMitchell.pdf
- R2D3's visualizations:
  - Intro to decision trees: http://www.r2d3.us/visual-intro-to-machine-learning-part-1/
  - Bias and variance in the context of decision trees: http://www.r2d3.us/visual-intro-to-machine-learning-part-2/

Last Class: K-Nearest Neighbors

kNN Classification: To predict category label y of a new point $\boldsymbol{x}$:

Find k nearest neighbors

Assign the majority label



- Easy to implement
- Versatile in terms of modeling many functions
- Interpretable in terms of data

Based on data from https://www.listendata.com/2017/12/k-nearest-neighbor-step-by-step-tutorial.html

Scaling Issues with kNNs

- Irrelevant Features: Distances become unreliable.
- Too Many Features: “Curse of Dimensionality”
- Large datasets (high N or D): Computationally inefficient to make predictions!

Problem 1: Irrelevant Features

- Let's say we want to predict $y =$ t-shirt size for a person.
- What if my input features are:
  - $x_1 =$ height
  - $x_2 =$ weight
  - $x_3 =$ hair length
  - $x_4 =$ age
  - $x_5 =$ body temperature
  - $x_6 =$ what they ate for breakfast this morning

  …

Common distance functions implicitly value all input features equally.

As you add more irrelevant variables, distances get dominated by those irrelevant dimensions in $\boldsymbol{x}$.

i.e., your kNN model might make decisions more based on breakfast than on the height and weight!

Problem 2: "Curse of Dimensionality"

- Adding more dimensions makes lots of things weird and counterintuitive
  - For example, the percentage of the volume of a D-dimensional sphere with radius r, that lies beyond ℓ_2 distance $0.99r$ from the center is:
    - 3% at $D = 3$
    - 63% at $D = 100$
    - 99.99% at $D = 1000$

- Specifically for k-NN, the space is now so large that all points in any finite dataset are likely to be very far apart.
  - "Closest points" are almost as far away as the farthest away points. When "nearest neighbors" are far away, predictions are poor.

Problem 3: Computationally Expensive

- High N, D also makes it computationally expensive to compute neighbors.

- Naively, must compute N distances between D-dimensional data pairs to compute neighbors before classifying a single new point.

- O(ND) for each new sample

Scaling kNN to high D and N? An Overview

Beyond our scope, but a quick overview:

Indexing

- Use kd-trees and other multidimensional indices to capture the training data. Each lookup is O(log *n*) rather than O(n), but on disk

Parallelism (e.g., PANDA, LBL)

- Use multiple cores / processors, and either compare against in-memory data or kd trees

Approximation

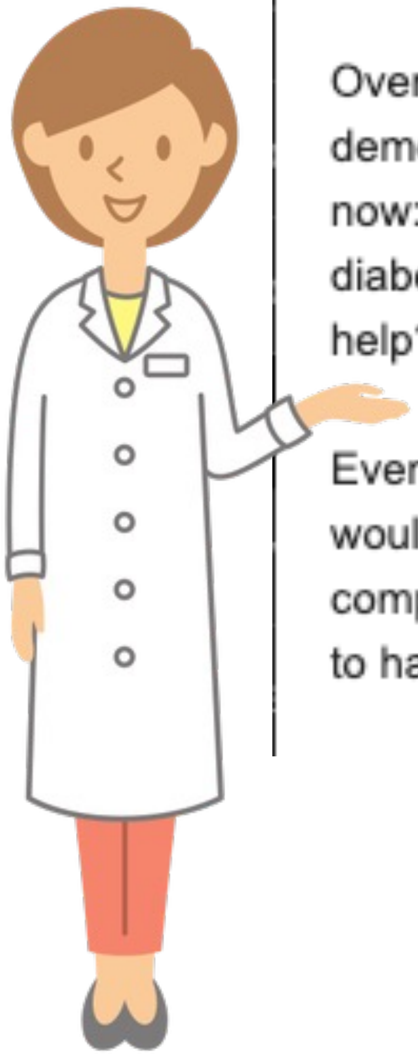
- https://scikit-learn.org/stable/modules/neighbors.html#nearest-neighbor-algorithms
- Libraries like FLANN: "Fast Library for Approximate Nearest Neighbors"
- For example, subsample the training dataset cleverly so that kNN mostly returns the same outputs
- See, e.g., https://www.kaggle.com/code/pawanbhandarkar/knn-vs-approximate-knn-what-s-the-difference/notebook

KNNs summary

- A simple and versatile ML approach, tied directly to the data.
- No training phase. Ready to make predictions the moment you have the dataset.
- “Non-parametric”. For KNNs, the data *are* the parameters.
- Scaling troubles, but still almost always worthwhile as your first algorithm for a new problem.

Decision Tree Models

(first, a new dataset from a physician friend)



Need help modeling diabetes risks!

I hope you are doing well in these weird times.

Over the years, I've collected data from lots of patients, recording their physical information, their demographic information, habits, and done their lab work to diagnose diabetes. I'm wondering now: from all this data, could I model the risk of other people with similar characteristics having diabetes given all this other information about them? And would your applied ML class be able to help? I've attached the data here for you to take a look.

Eventually, we'll want to explain our findings to patients, and point out any behavioral changes that would mitigate their risk for diabetes. Even if the risk factors we find are non-modifiable, insurance companies would be interested in understanding and estimating this risk. Either way, it'd be great to have something that we can understand and interpret well!

Diabetes Data

data matrix X

ID	AGE RIDAGEYR	WAIST	HEIGHT BMX	CHOLESTEROL	UPPER LEG LENGTH BMXLEG	WEIGHT	BMI BMXBMI	RACE	HIGH BP BPQ0	ALCOHOL USE	EDUCATION DMDEDUC2	GENDER	FAMILY INCOME RATIO INDFM	GLYCOHAEMOGLOBIN	TIC
73557	69.0	100.0	171.3	167.0	39.2	78.3	26.7	Non-Hispanic Black	yes	1.0	high school graduate / GED	male	0.84	13.9	yes
73558	54.0	107.6	176.8	170.0	40.0	89.5	28.6	Non-Hispanic White	yes	7.0	high school graduate / GED	male	1.78	9.1	yes
73559	72.0	109.2	175.3	126.0	40.0	88.9	28.9	Non-Hispanic White	yes	0.0	some college or AA degree	male	4.51	8.9	yes
73562	56.0	123.1	158.7	226.0	34.2	105.0	41.7	Mexican American	yes	5.0	some college or AA degree	male	4.79	5.5	no
73564	61.0	110.8	161.8	168.0	37.1	93.4	35.7	Non-Hispanic White	yes	2.0	college graduate or above	female	5.0	5.5	no
73566	56.0	85.5	152.8	278.0	32.4	61.8	26.5	Non-Hispanic White	no	1.0	high school graduate / GED	female	0.48	5.4	no
73567	65.0	93.7	172.4	173.0	40.0	65.3	22.0	Non-Hispanic White	no	4.0	9th-11th grade	male	1.2	5.2	no
73568	26.0	73.7	152.5	168.0	34.4	47.1	20.3	Non-Hispanic White	[illegible]	2.0	college graduate or above	female	5.0	5.2	no
73571	76.0	122.1	172.5	167.0	35.5	102.4	34.4	Non-Hispanic White	yes	2.0	college graduate or above	male	5.0	6.9	yes
73577	32.0	100.0	166.2	182.0	36.5	79.7	28.9	Mexican American	no	20.0	Less than 9th grade	male	0.29	5.3	no
73581	50.0	99.3	185.0	202.0	42.8	80.9	23.6	Other or Multi-Racial	no	0.0	college graduate or above	male	5.0	5.0	no
73585	28.0	90.3	175.1	198.0	40.5	92.2	30.1	Other or Multi-Racial	no	4.0	some college or AA degree	male	2.26	5.0	no
73589	35.0	94.6	172.9	192.0	39.1	78.3	26.2	Non-Hispanic White	no	2.0	high school graduate / GED	male	1.74	5.5	no
73595	58.0	114.8	175.3	165.0	40.1	96.0	31.2	Other Hispanic	no	1.0	some college or AA degree	male	3.09	7.7	no
73596	57.0	117.8	164.7	151.0	35.3	104.0	38.3	Other or Multi-Racial	yes	1.0	college graduate or above	female	5.0	5.9	no
73600	37.0	122.9	185.1	189.0	48.1	126.2	36.8	Non-Hispanic Black	yes	2.0	high school graduate / GED	male	0.63	6.2	yes
73604	69.0	96.6	156.9	203.0	37.0	59.5	24.2	Non-Hispanic White	no	1.0	some college or AA degree	female	2.44	5.4	no
73607	75.0	130.5	169.6	161.0	36.5	111.9	38.9	Non-Hispanic White	yes	0.0	high school graduate / GED	male	1.08	5.0	no
73610	43.0	102.6	176.8	200.0	38.8	90.2	28.9	Non-Hispanic White	no	5.0	college graduate or above	male	2.03	4.9	no
73613	60.0	113.6	163.8	203.0	41.6	104.9	39.1	Non-Hispanic Black	yes	2.0	9th-11th grade	female	5.0	6.1	no
73614	55.0	90.9	167.9	256.0	43.5	60.9	21.6	Non-Hispanic White	no	0.0	high school graduate / GED	female	1.29	5.0	no
73615	65.0	100.3	145.9	166.0	30.0	55.4	26.0	Other Hispanic	yes	1.0	Less than 9th grade	female	1.22	6.3	yes

label y_i

sample x_i

Data from NHANES 2013/14 survey

The Data

ID	AGE (RIDAGEYR)	WAIST	HEIGHT (BMX)	CHOLESTEROL	UPPER LEG LENGTH (MXLEG)	WEIGHT	BMI (BMXBMI)	RACE	HIGH BP (BPQ0)	ALCOHOL USE	EDUCATION (DMDEDUC2)	GENDER	FAMILY INCOME RATIO (INDFM)	GLYCOHAEMOGLOBIN	DIABETIC
73557	69.0	100.0	171.3	167.0	39.2	78.3	26.7	Non-Hispanic Black	yes	1.0	high school graduate / GED	male	0.84	13.9	yes
73558	54.0	107.6	176.8	170.0	40.0	89.5	28.6	Non-Hispanic White	yes	7.0	high school graduate / GED	male	1.78	9.1	yes
73559	72.0	109.2	175.3	126.0	40.0	88.9	28.9	Non-Hispanic White	yes	0.0	some college or AA degree	male	4.51	8.9	yes
73562	56.0	123.1	158.7	226.0	34.2	105.0	41.7	Mexican American	yes	5.0	some college or AA degree	male	4.79	5.5	no
73564	61.0								yes	2.0	college graduate or above	female	5.0	5.5	no
73566	56.0								no	1.0	high school graduate / GED	female	0.48	5.4	no
73567	65.0	93.7	172.4	173.0	40.0	65.3		White	no	4.0	9th-11th grade	male	1.2	5.2	no
73568	26.0	73.7	152.5	168.0	34.4	47.1	20.3	Non-Hispa	no	2.0	college graduate or above	female	5.0	5.2	no
73571	76.0	122.1	172.5	167.0	35.5	102.4	34.4	Non-Hispanic White	yes	2.0	college graduate or above	male	5.0	6.9	yes
73577		100.0	166.2	182.0	36.5	79.7	28.9	Mexican American	no	20.0	Less than 9th grade	male	0.29	5.3	no
73581								Multi-Racial	no	0.0	college graduate or above	male	5.0	5.0	no
73585								Multi-Racial	no	4.0	some college or AA degree	male	2.26	5.0	no
73589								anic White	no	2.0	high school graduate / GED	male	1.74	5.5	no
73595								spanic	no	1.0	some college or AA degree	male	3.09	7.7	no
73596	57.0	117.8	164.7	151.0	35.3	104.0	38.3	Other or Multi-Racial	yes	1.0	college graduate or above	female	5.0	5.9	no
73600	37.0	122.9	185.1	189.0	48.1	126.2	36.8	Non-Hispanic Black	yes	2.0	high school graduate / GED	male	0.63	6.2	yes
73604	69.0	96.6	156.9	203.0	37.0	59.5	24.2	Non-Hispanic White	no	1.0	some college or AA degree	female	2.44	5.4	no
73607	75.0	130.5	169.6	161.0	36.5	111.9	38.9	Non-Hispanic White	yes	0.0	high school graduate / GED	male	1.08	5.0	no
73610	43.0	102.6	176.8	200.0	38.8	90.2	28.9	Non-Hispanic White	no	5.0	college graduate or above	male	2.03	4.9	no
73613	60.0	113.6	163.8	203.0	41.6	104.9	39.1	Non-Hispanic Black	yes	2.0	9th-11th grade	female	5.0	6.1	no
73614	55.0	90.9	167.9	256.0	43.5	60.9	21.6	Non-Hispanic White	no	0.0	high school graduate / GED	female	1.29	5.0	no
73615	65.0	100.3	145.9	166.0	30.0	55.4	26.0	Other Hispanic	yes	1.0	Less than 9th grade	female	1.22	6.3	yes

Columns X_j denote features

Patient number: should this really be a feature?

Data from NHANES 2013/14 survey

Feature Types

numeric | nominal | ordinal | binary

ID	AGE RIDAGEYR	WAIST	HEIGHT	CHOLESTEROL	UPPER LEG LENGTH MXLEG	WEIGHT	BMI BMXBMI	RACE	HIGH BP BPQ0	ALCOHOL USE	EDUCATION DMDEDUC2	GENDER	FAMILY INCOME RATIO INDFM	GLYCOHAEMOGLOBIN	DIABETIC
73557	69.0	100.0	171.3	167.0	39.2	78.3	26.7	Non-Hispanic Black	yes	1.0	high school graduate / GED	male	0.84	13.9	yes
73558	54.0	107.6	176.8	170.0	40.0	89.5	28.6	Non-Hispanic White	yes	7.0	high school graduate / GED	male	1.78	9.1	yes
73559	72.0	109.2	175.3	126.0	40.0	88.9	28.9	Non-Hispanic White	yes	0.0	some college or AA degree	male	4.51	8.9	yes
73562	56.0	123.1	158.7	226.0	34.2	105.0	41.7	Mexican American	yes	5.0	some college or AA degree	male	4.79	5.5	no
73564	61.0	110.8	161.8	168.0	37.1	93.4	35.7	Non-Hispanic White	yes	2.0	college graduate or above	female	5.0	5.5	no
73566	56.0	85.5	152.8	278.0	32.4	61.8	26.5	Non-Hispanic White	no	1.0	high school graduate / GED	female	0.48	5.4	no
73567	65.0	93.7	172.4	173.0	40.0	65.3	22.0	Non-Hispanic White	no	4.0	9th-11th grade	male	1.2	5.2	no
73568	26.0	73.7	152.5	168.0	34.4	47.1	20.3	Non-Hispanic White	no	2.0	college graduate or above	female	5.0	5.2	no
73571	76.0	122.1	172.5	167.0	35.5	102.4	34.4	Non-Hispanic White	yes	2.0	college graduate or above	male	5.0	6.9	yes
73577	32.0	100.0	166.2	182.0	36.5	79.7	28.9	Mexican American	no	20.0	Less than 9th grade	male	0.29	5.3	no
73581	50.0	99.3	185.0	202.0	42.8	80.9	23.6	Other or Multi-Racial	no	0.0	college graduate or above	male	5.0	5.0	no
73585	28.0	90.3	175.1	198.0	40.5	92.2	30.1	Other or Multi-Racial	no	4.0	some college or AA degree	male	2.26	5.0	no
73589	35.0	94.6	[illegible]	192.0	39.1	[illegible]	26.2	Non-Hispanic White	no	2.0	high school graduate / GED	male	1.74	5.5	no
73595	58.0	114.8	175.3	165.0	40.1	96.0	31.2	Other Hispanic	no	1.0	some college or AA degree	male	3.09	7.7	no
73596	57.0	[illegible]	[illegible]	[illegible]	35.3	[illegible]	[illegible]	Other or Multi-Racial	yes	1.0	college graduate or above	female	5.0	5.9	no
73600	37.0	122.9	185.1	189.0	[illegible]	[illegible]	36.8	Non-Hispanic Black	yes	2.0	high school graduate / GED	male	0.63	6.2	yes
73604	69.0	96.6	156.9	203.0	37.0	59.5	24.2	Non-Hispanic White	no	1.0	some college or AA degree	female	2.44	5.4	no
73607	75.0	130.5	169.6	[illegible]	[illegible]	[illegible]	38.9	Non-Hispanic White	yes	0.0	high school graduate / GED	male	1.08	5.0	no
73610	43.0	102.6	176.8	200.0	38.8	90.2	28.9	Non-Hispanic White	no	5.0	college graduate or above	male	2.03	4.9	no
73613	60.0	113.6	163.8	203.0	41.6	104.9	39.1	Non-Hispanic Black	yes	2.0	9th-11th grade	female	5.0	6.1	no
73614	55.0	90.9	167.9	256.0	43.5	60.9	21.6	Non-Hispanic White	no	0.0	high school graduate / GED	female	1.29	5.0	no
73615	65.0	100.3	145.9	166.0	30.0	55.4	26.0	Other Hispanic	yes	1.0	Less than 9th grade	female	1.22	6.3	yes

This column seems binary, but also has "refused to answer" and "don't know" categories

Data from NHANES 2013/14 survey

Data Dictionary

- Data sets are often accompanied by a data dictionary that describes each feature
- It is critical to understand the data!
- The dictionary for our data: https://wwwn.cdc.gov/nchs/nhanes/Default.aspx

ID (SEQN)	AGE (RIDAGEYR)	WAIST_CIRCUM (BMXWAIST)	HEIGHT (BMXHT)	CHOLESTEROL (LBXTC)	UPPER_LEG_LEN (BMXLEG)	WEIGHT (BMXWT)	BMI (BMXBMI)	RACE (RIDRETH1)	HIGH_BP (BPQ020)	ALCOHOL_USE (ALQ120Q)	EDUCATION (DMDEDUC2)	GENDER (RIAGENDR)	FAMILY_INCOME_RATIO (INDFMPIR)	GLYCOHEMOGLOBIN (LBXGH)	DIABETIC
73557	69.0	100.0	171.3	167.0	39.2	78.3	26.7	Non-Hispanic Black	yes	1.0	high school graduate / GED	male	0.84	13.9	yes
73558	54.0	107.6	176.8	170.0	40.0	89.5	28.6	Non-Hispanic White	yes	7.0	high school graduate / GED	male	1.78	9.1	yes
73559	72.0	109.2	175.3	126.0	40.0	88.9	28.9	Non-Hispanic White	yes	0.0	some college or AA degree	male	4.51	8.9	yes
73562	56.0	123.1	158.7	226.0	34.2	105.0	41.7	Mexican American	yes	5.0	some college or AA degree	male	4.79	5.5	no
73564	61.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.0	college graduate or above	female	5.0	5.5	no
73566	56.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.0	high school graduate / GED	female	0.48	5.4	no
73567	65.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4.0	9th-11th grade	male	1.2	5.2	no
73568	26.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.0	college graduate or above	female	5.0	5.2	no
73571	76.0	122.1	172.5	167.0	35.5	102.4	34.4	[illegible]	[illegible]	2.0	college graduate or above	male	5.0	6.9	yes
73577	32.0	100.0	166.2	182.0	36.5	79.7	28.9	Mexican American	no	20.0	Less than 9th grade	male	0.29	5.3	no
73581	50.0	99.3	185.0	202.0	42.8	80.9	23.6	Other or Multi-Racial	no	0.0	college graduate or above	male	5.0	5.0	no

777 = refused; 999 = don't know

Data from NHANES 2013/14 survey

Decision Trees for People

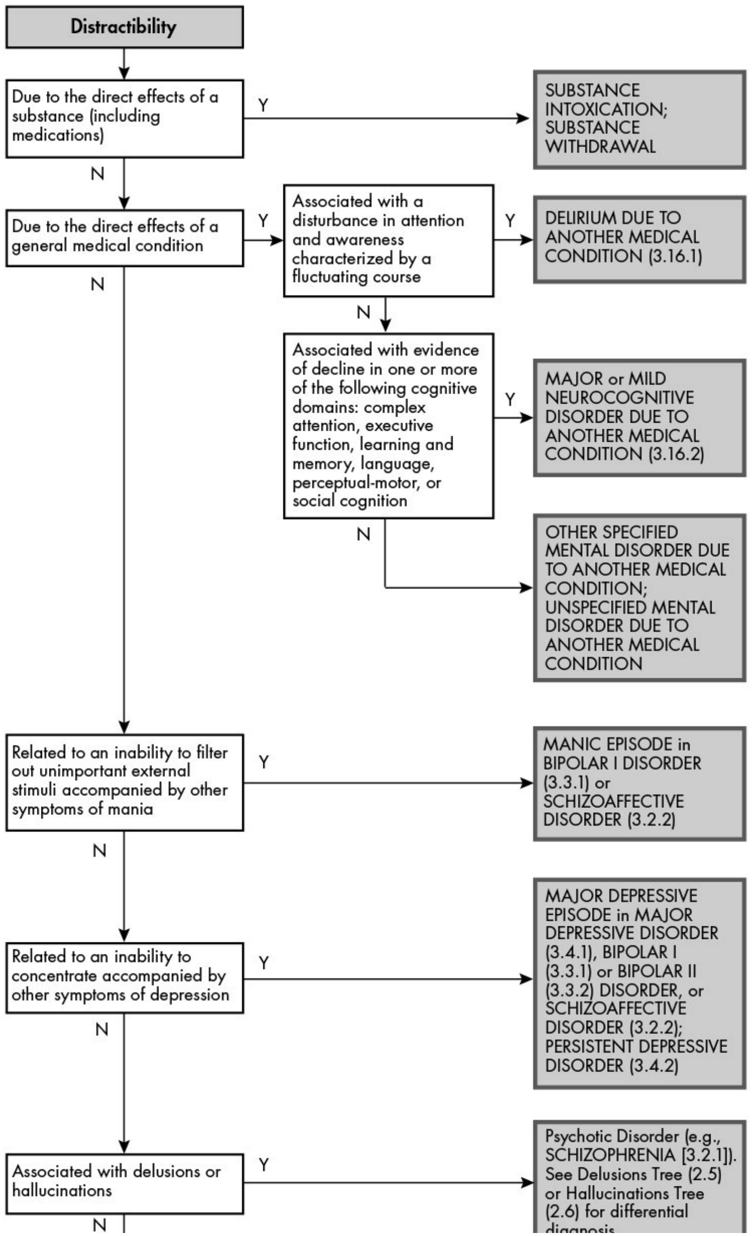
How do we train a human to make a diagnosis?

- Often, a kind of flowchart based on tests! "Decision Tree"
  - e.g., how we train psychiatrists to make diagnoses? →
- "Explainable" in a clear way, easy to evaluate

Idea: Let's create decision trees by looking at example input->output pairs i.e. learning!

First, let's formalize what we mean by a decision tree…

APA DSM Library



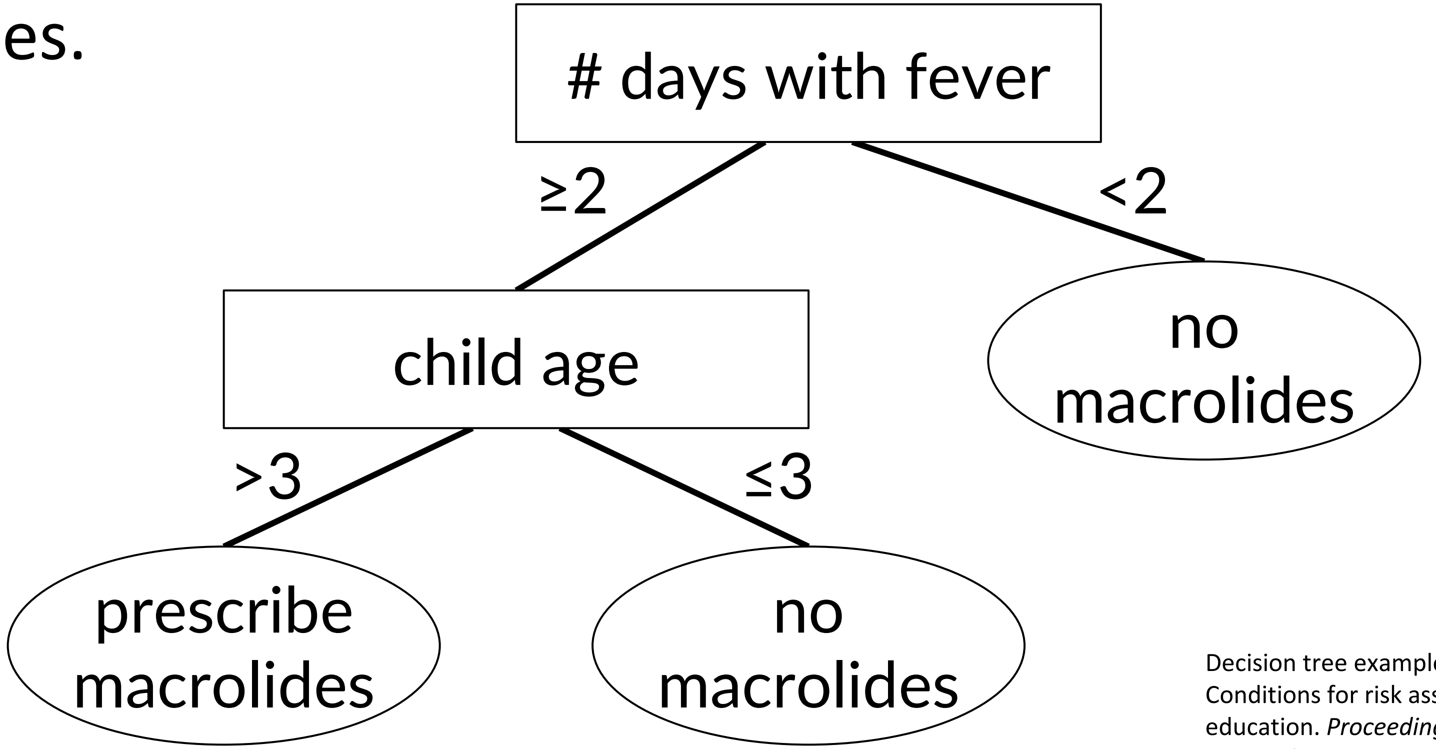
A Decision Tree Based on Boolean Tests

For continuous features, we'll restrict our study to internal nodes that make binary decisions* based on a single feature:

- e.g. is a real-valued feature above or below some threshold?

- e.g. is a binary-valued feature true or false?

* for discrete-valued features we will usually create as many splits as the number of values.



Decision tree example from: Martignon and Monti. (2010). Conditions for risk assessment as a topic for probabilistic education. *Proceedings of the Eighth International Conference on Teaching Statistics* (ICOTS8).

Each Internal Tree Node "Splits" Training Data

ColorOfCoat	TypeOfHorse
black	thoroughbred
bay	Arabian
black	thoroughbred
chestnut	quarter
black	Arabian

N=5; 3 classes

ColorOfCoat = 'black'

ColorOfCoat	TypeOfHorse
black	thoroughbred
black	thoroughbred
black	Arabian

N=3; 2 classes

ColorOfCoat	TypeOfHorse
bay	Arabian
chestnut	quarter

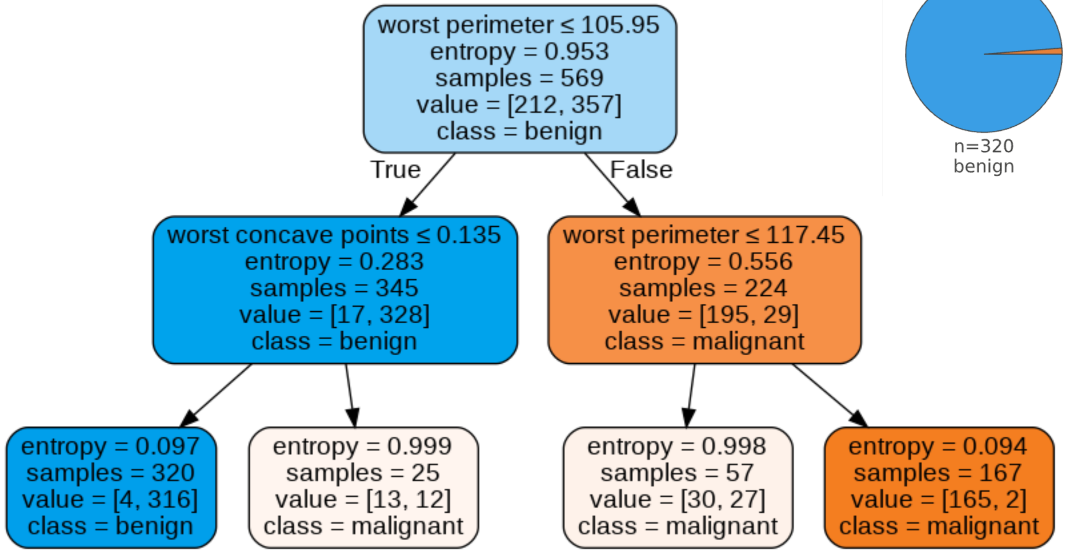
N=2; 2 classes

Representing Decision Trees

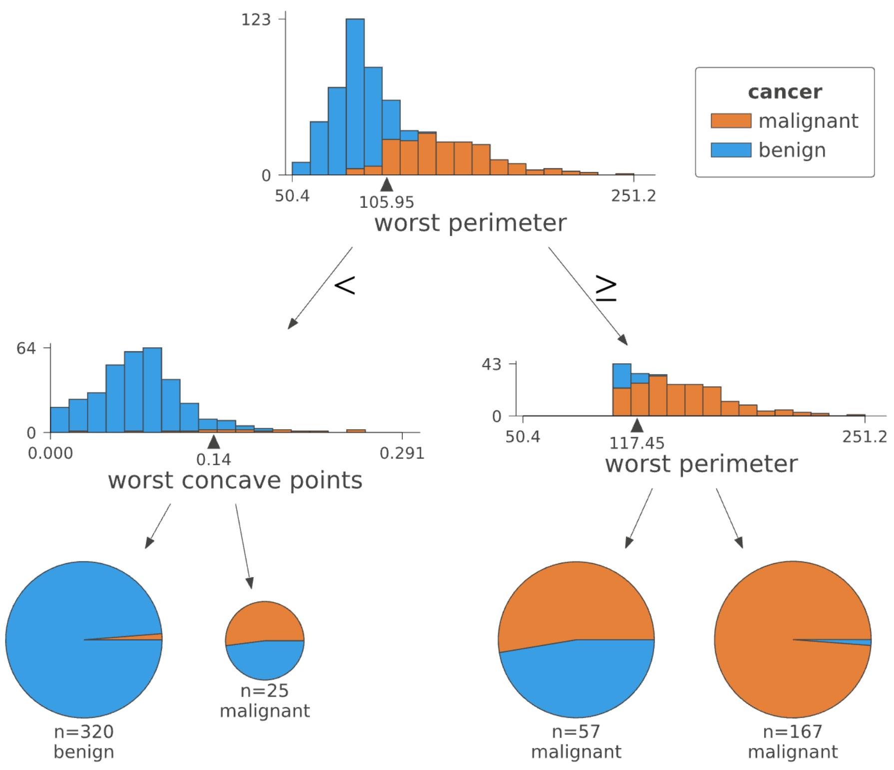
sklearn text

```
|--- worst perimeter <= 105.95
|   |--- worst concave points <= 0.135
|   |   |--- class: benign
|   |--- worst concave points > 0.135
|   |   |--- class: malignant
|--- worst perimeter > 105.95
|   |--- worst perimeter <= 117.45
|   |   |--- class: malignant
|   |--- worst perimeter > 117.45
|   |   |--- class: malignant
```

sklearn graphviz



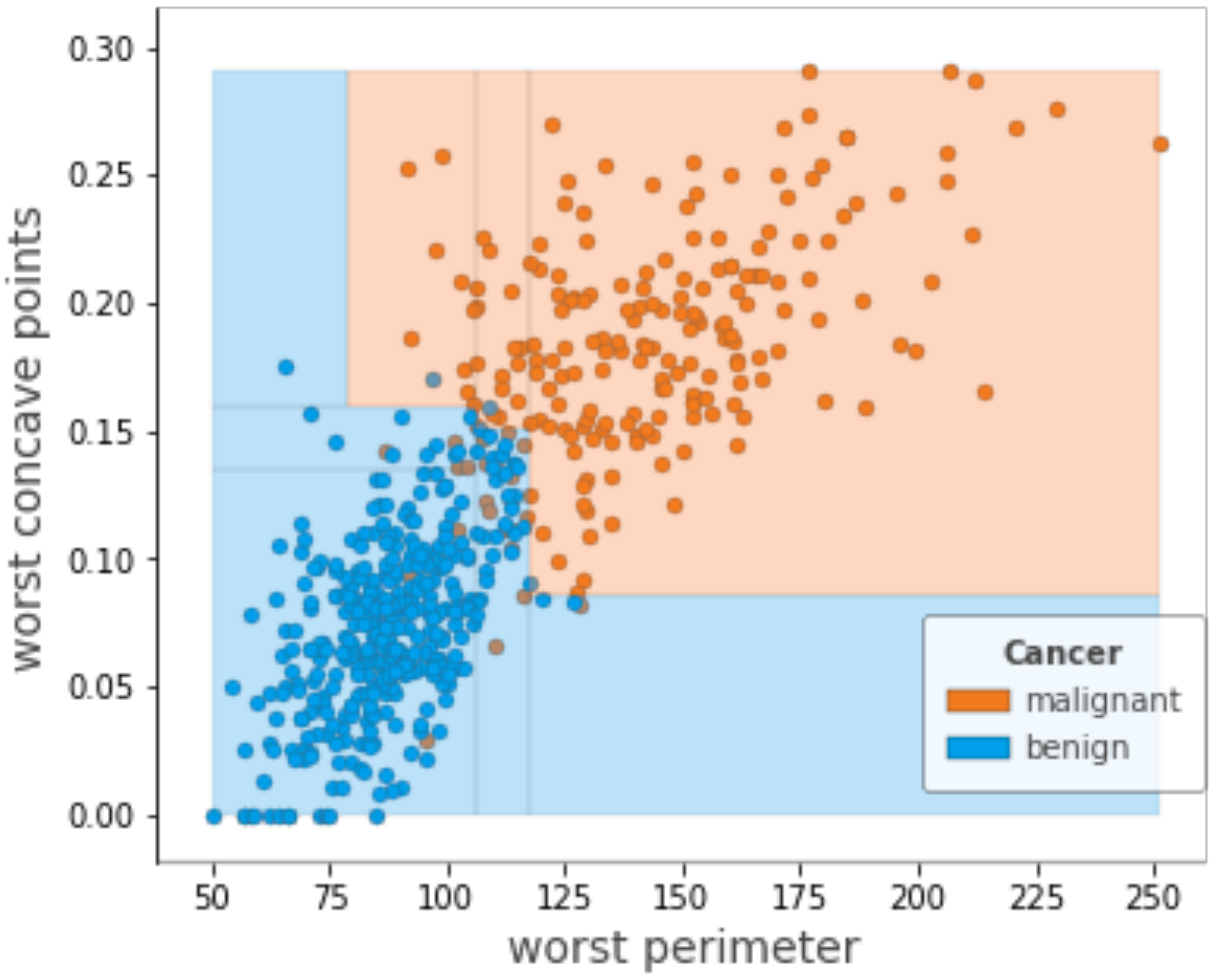
dtreeviz



Decisions trees generated on Wisconsin Breast Cancer dataset in sklearn

Decision Tree – Induced Partition

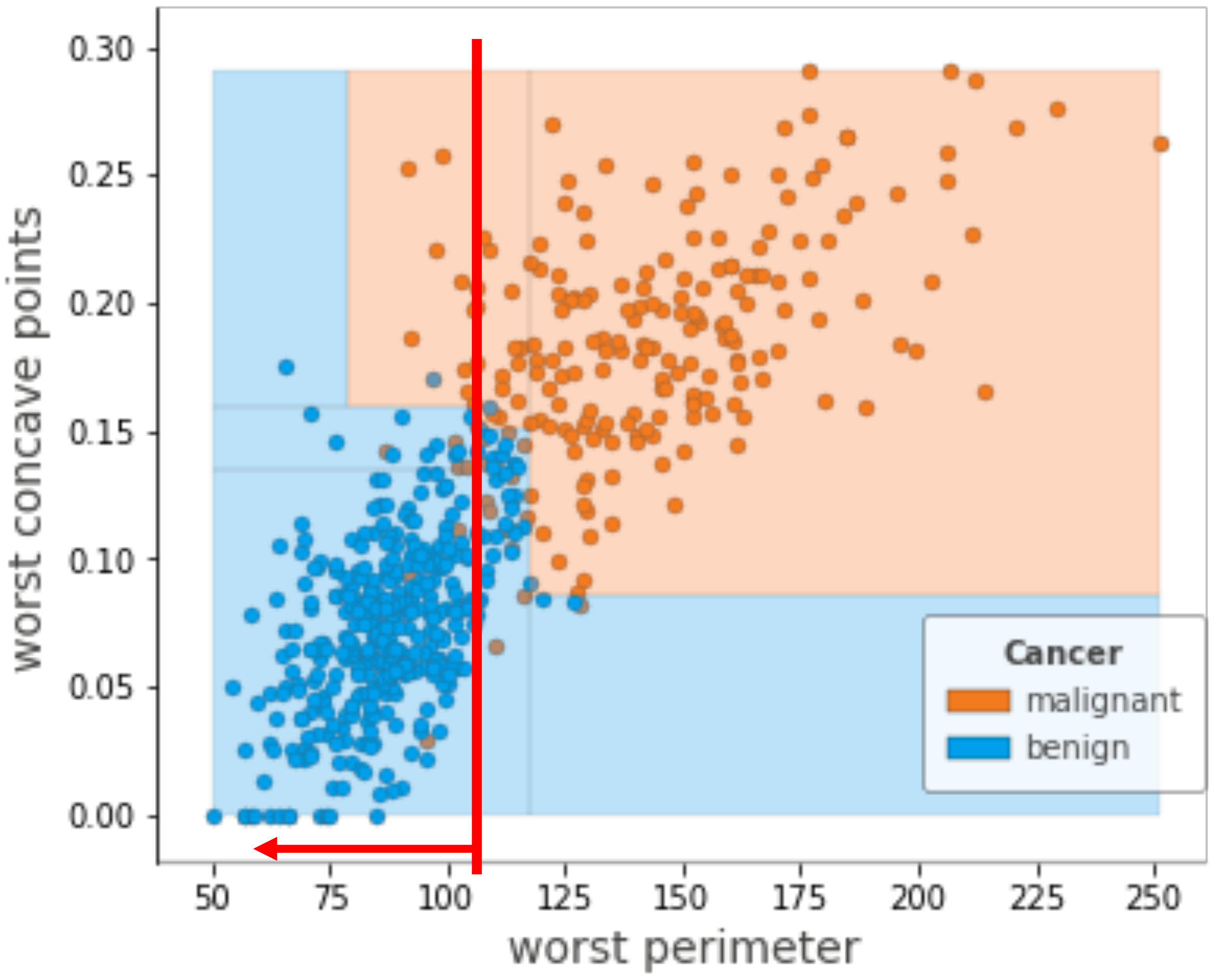
```
|--- worst perimeter <= 105.95
| |--- worst concave points <= 0.135
| | |--- class: benign
| |--- worst concave points > 0.135
| | |--- worst concave points < 0.16
| | | |--- class: benign
| | |--- worst concave points > 0.16
| | | | --- worst perimeter > 80
| | | | | --- class: malignant
| | | | --- worst perimeter < 80
| | | | | --- class: benign
                    …
                    …
```



Decisions trees generated on Wisconsin Breast Cancer dataset in sklearn

Decision Tree – Induced Partition

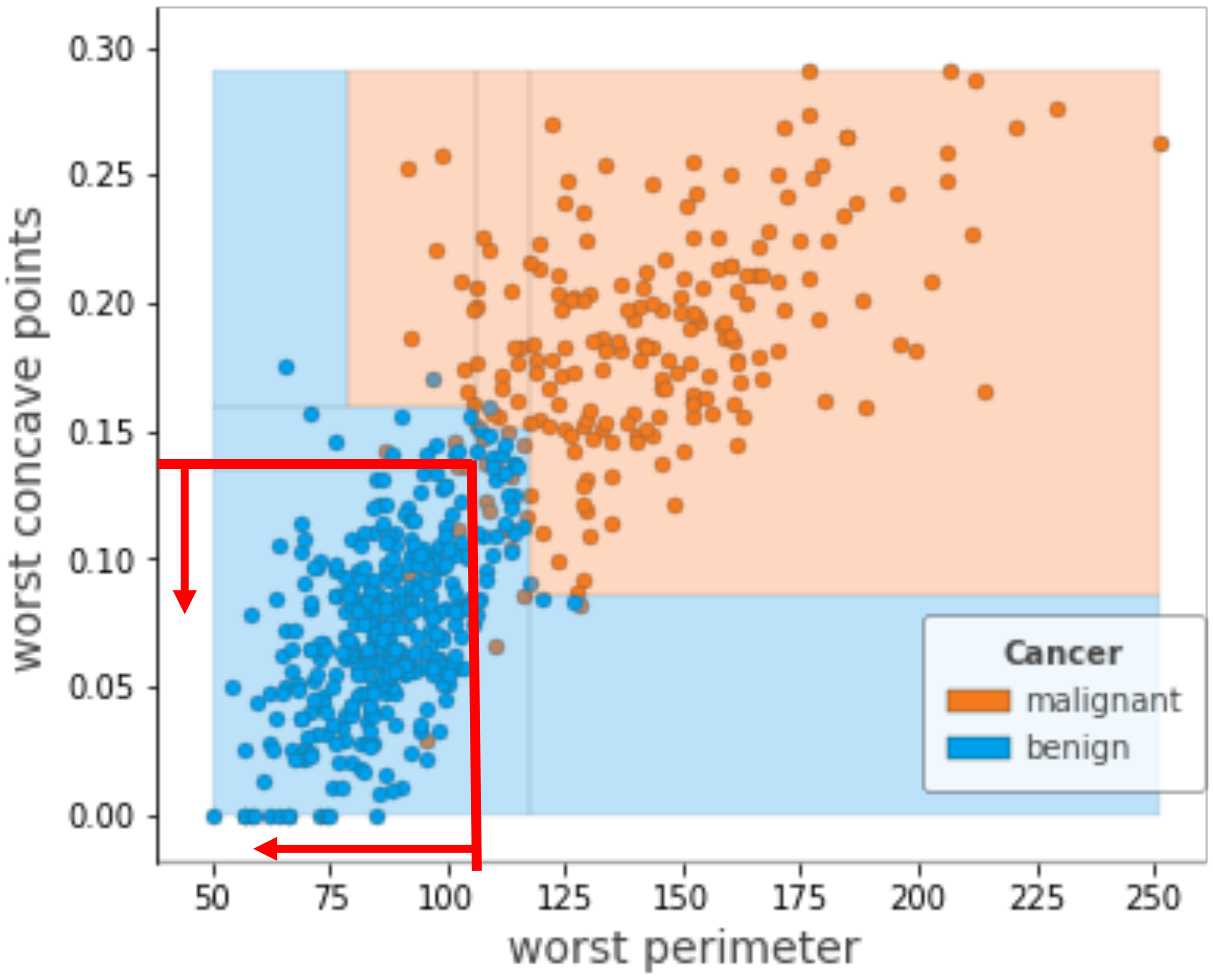
```
|--- worst perimeter <= 105.95
| |--- worst concave points <= 0.135
| | |--- class: benign
| |--- worst concave points > 0.135
| | |--- worst concave points < 0.16
| | | |--- class: benign
| | |--- worst concave points > 0.16
| | | | --- worst perimeter > 80
| | | | | --- class: malignant
| | | | --- worst perimeter < 80
| | | | | --- class: benign
                    …
                    …
```



Decisions trees generated on Wisconsin Breast Cancer dataset in sklearn

Decision Tree – Induced Partition

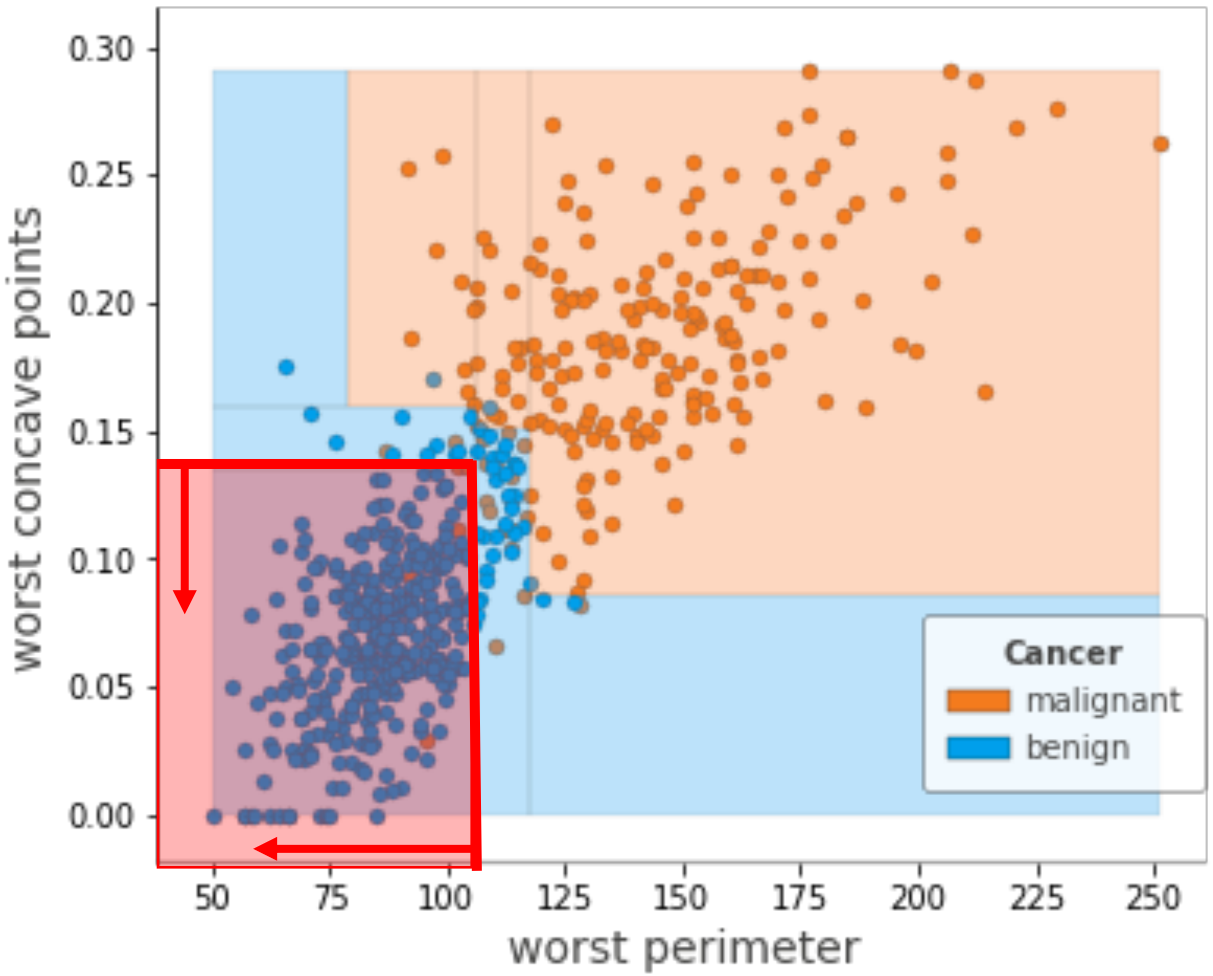
```
|--- worst perimeter <= 105.95
|  |--- worst concave points <= 0.135
|  |  |--- class: benign
|  |--- worst concave points > 0.135
|  |  |--- worst concave points < 0.16
|  |  |  |--- class: benign
|  |  |--- worst concave points > 0.16
|  |  |  |  --- worst perimeter > 80
|  |  |  |  |  --- class: malignant
|  |  |  |  --- worst perimeter < 80
|  |  |  |  |  --- class: benign
                         …
                         …
```



Decisions trees generated on Wisconsin Breast Cancer dataset in sklearn

Decision Tree – Induced Partition

```
|--- worst perimeter <= 105.95
|  |--- worst concave points <= 0.135
|  |  |--- class: benign
|  |--- worst concave points > 0.135
|  |  |--- worst concave points < 0.16
|  |  |  |--- class: benign
|  |  |--- worst concave points > 0.16
|  |  |  | --- worst perimeter > 80
|  |  |  |  | --- class: malignant
|  |  |  | --- worst perimeter < 80
|  |  |  |  | --- class: benign
                    …
                    …
```



So what is the hypothesis class expressed by a DT?

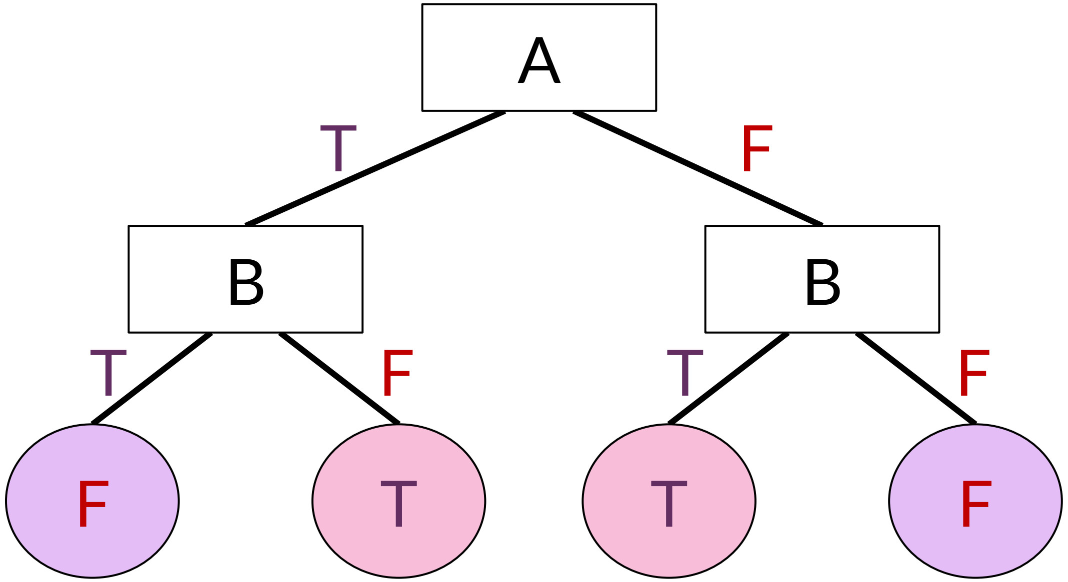
Decision trees divide the feature space into axis-aligned “hyperrectangles”

Decision Trees with Boolean Variables

Decision Trees and Boolean Functions

- Decision trees can represent any Boolean function of the features

A	B	A xor B
T	T	F
T	F	T
F	T	T
F	F	F

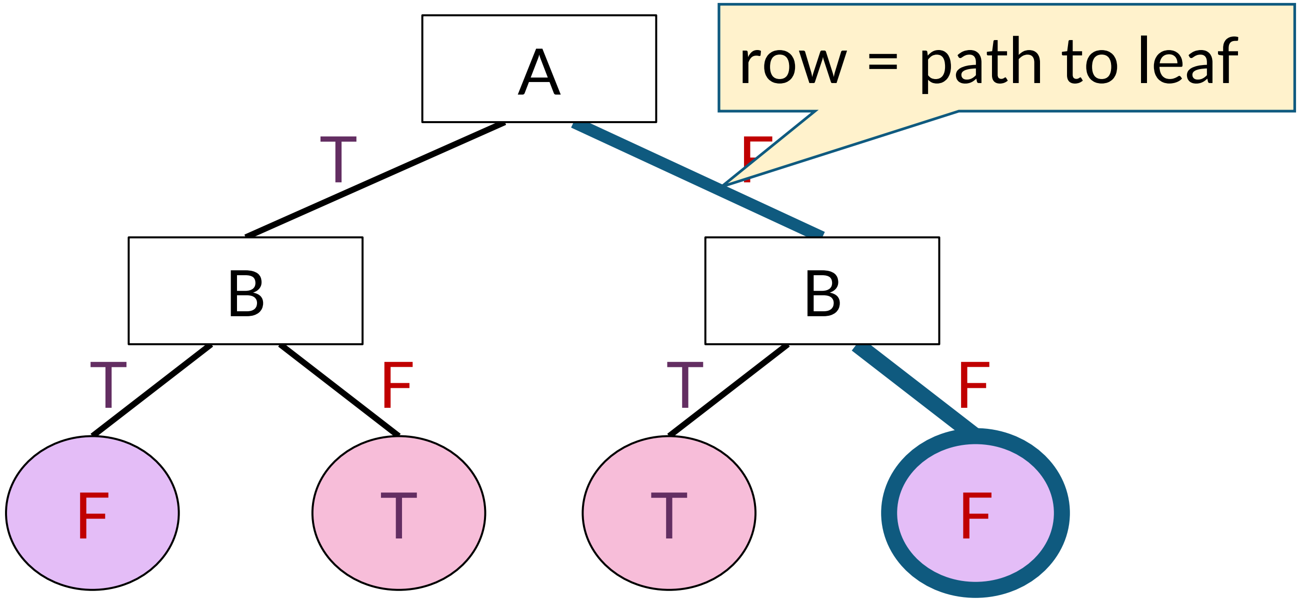


- In the worst case, the tree will require exponentially many nodes

Decision Trees and Boolean Functions

- Decision trees can represent any boolean function of the features

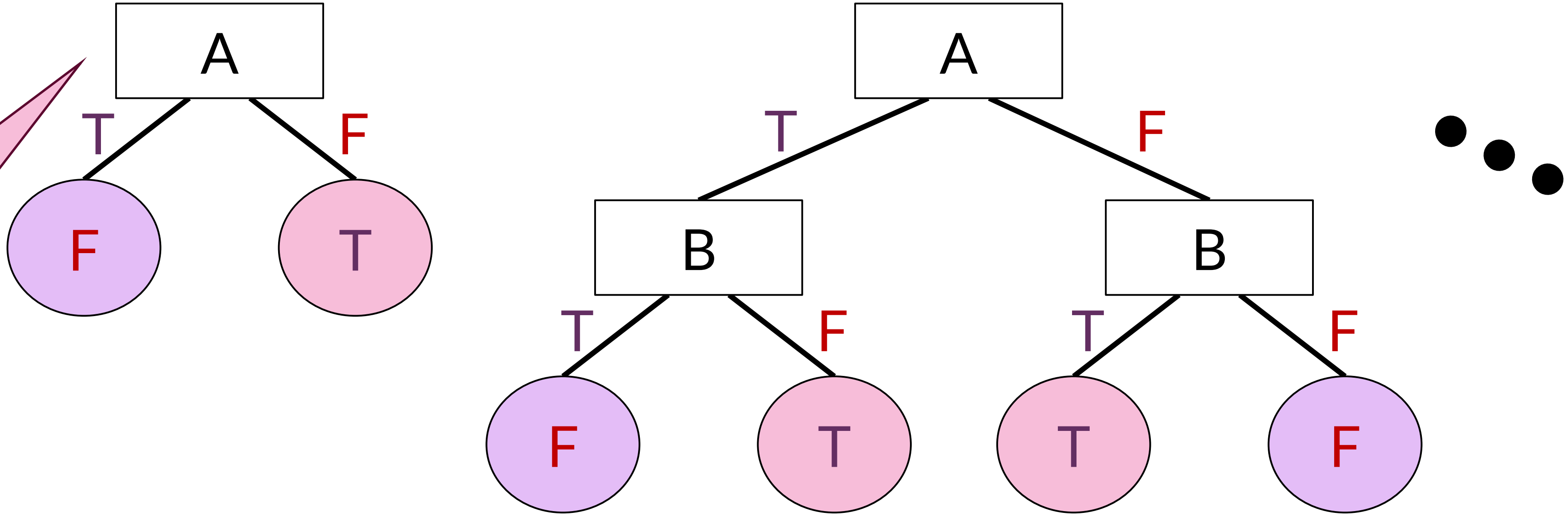
A	B	A xor B
T	T	F
T	F	T
F	T	T
F	F	F



Decision Trees and Boolean Functions

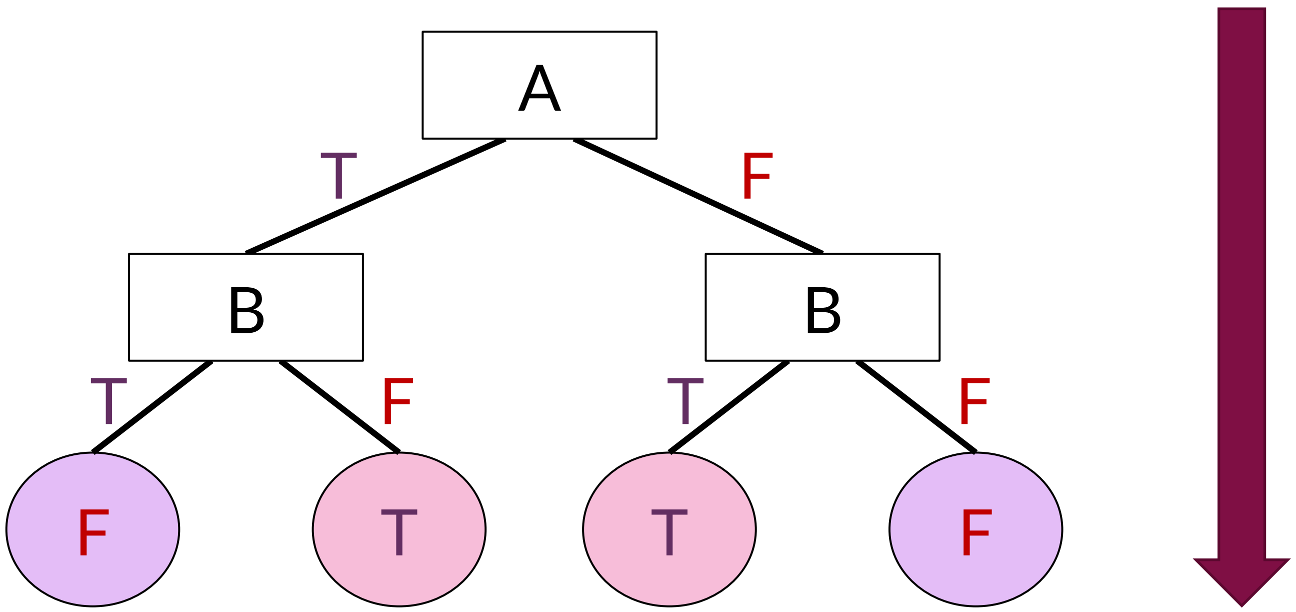
- DTs have a variable-sized hypothesis space based on their depth
  - Depth 1: any boolean function based on one feature
  - Depth 2: any boolean function based on two features
  - . . .

DTs of depth 1 are also called decision stumps



Training Decision Trees

Top-Down Decision Tree Training – Grow top down



Top-Down Decision Tree Induction

[ID3 (1986), C4.5(1993) by Quinlan]

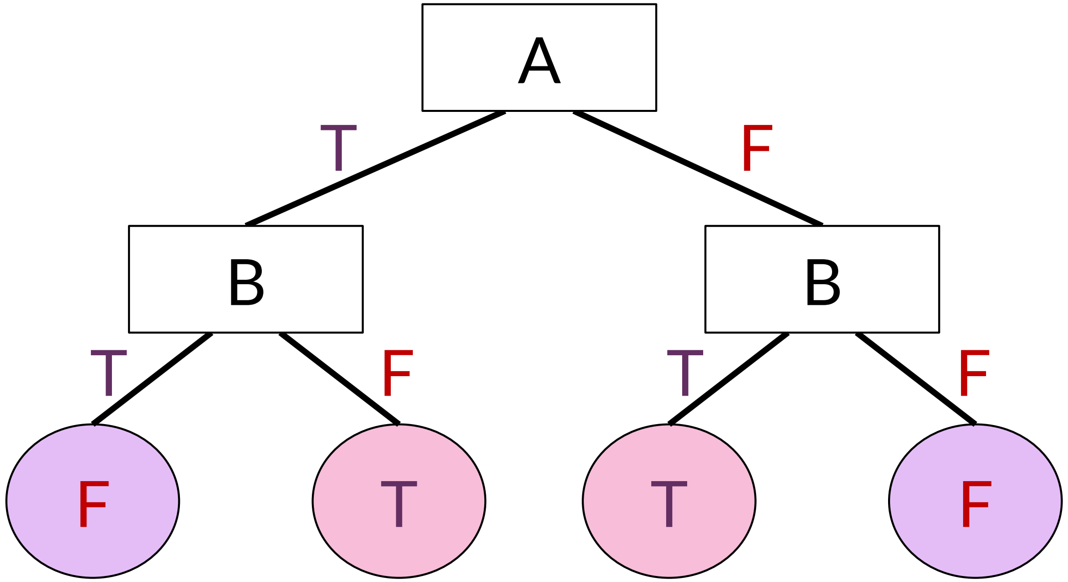
Let $\mathcal{D}$ be a set of labeled instances; $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N} = [X_{N\times D}, \boldsymbol{y}_{N\times 1}]$

Let $\mathcal{D}[X_j = v]$ be the subset of $\mathcal{D}$ where feature X_j has value v

function `train_tree(`$\mathcal{D}$`)`

1. If data $\mathcal{D}$ all have the same label y, return new `leaf_node(`y`)`
2. Pick the "best" feature X_j to partition $\mathcal{D}$
3. Set `node = new decision_node(`X_j`)`
4. For each value v that X_j can take
   - Recursively create a new child `train_tree(`$\mathcal{D}[X_j = v]$`)` of `node`
5. Return `node`

Top-Down Decision Tree Training



Top-Down Decision Tree Induction

[ID3, C4.5 by Quinlan]

Let $\mathcal{D}$ be a set of labeled instances; initially $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N} = [X_{N\times D}, \boldsymbol{y}_{N\times 1}]$

Let $\mathcal{D}[X_j = v]$ be the subset of $\mathcal{D}$ where feature X_j has value v

How do we choose which feature is best?

function `train_tree(`$\mathcal{D}$`)`

1. If data $\mathcal{D}$ all have the same label y, return new `leaf_node(`y`)`
2. Pick the "best" feature X_j to partition $\mathcal{D}$
3. Set `node = new decision_node(`X_j`)`
4. For each value v that X_j can take
   - Recursively create a new child `train_tree(`$\mathcal{D}[X_j = v]$`)` of `node`
5. Return `node`

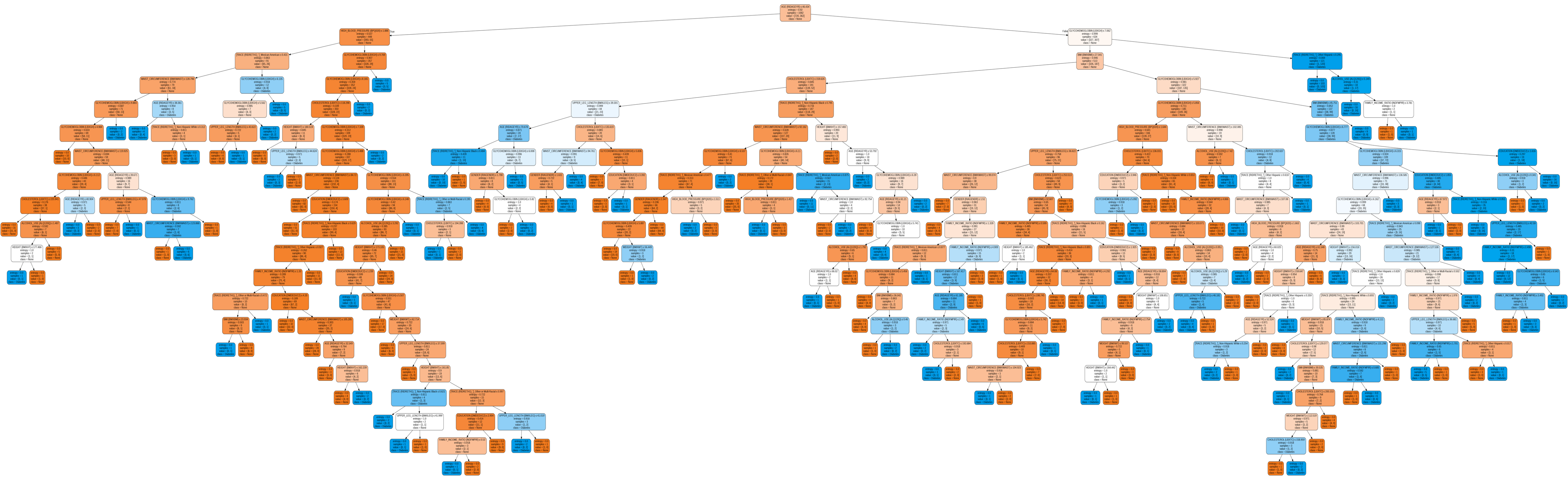
Choosing the "Best Feature"

Key problem: how should we choose which feature to split the data?

Possibilities:

Random
Choose any feature at random?

Diabetes DT – Random Features



Is this really the best way to choose decision nodes?

Choosing the Best Feature

Key problem: how should we choose which feature to split the data?

Possibilities:

Random	?
Choose any feature at random	

Choosing the Best Feature

Key problem: how should we choose which feature to split the data?

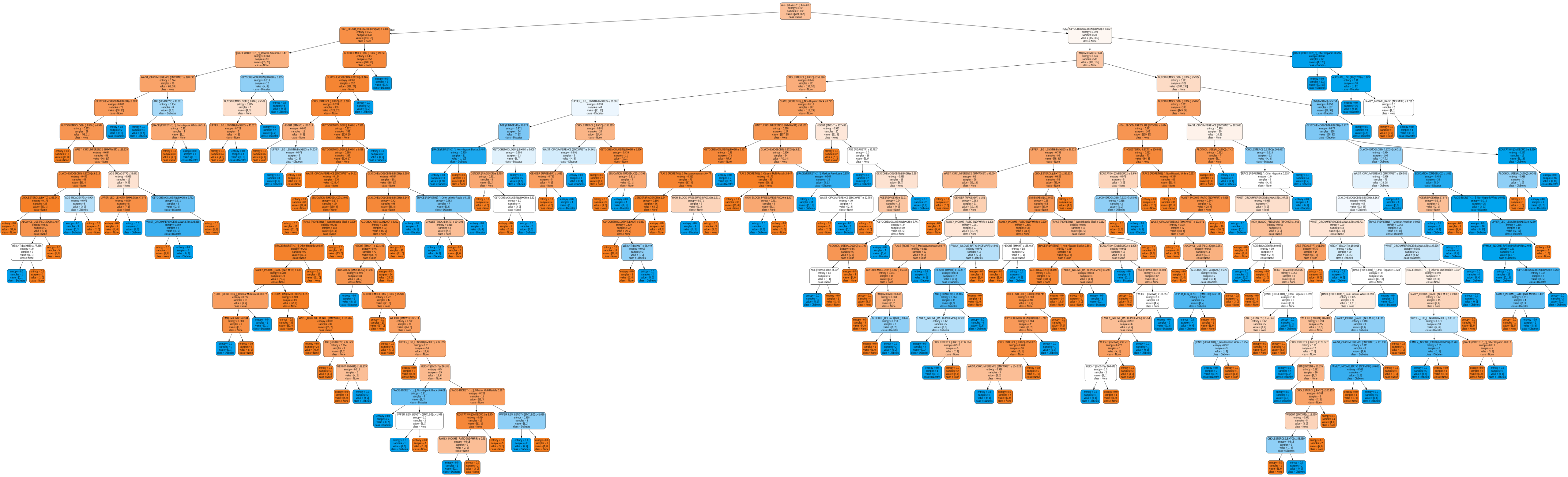
Possibilities:

Random	Max-Gain
Choose any feature at random	Choose the feature with the largest expected *information gain*

i.e., the feature that is expected to result in the shortest subtree

Learning Smaller Models

Recap: DT with random features



Recall: We like Simple Models!

This is why we studied Bias-Variance Tradeoffs, Regularization, Feature Selection etc.

Learning bias: Occam’s Razor

Principle stated by William of Ockham (1285-1347)

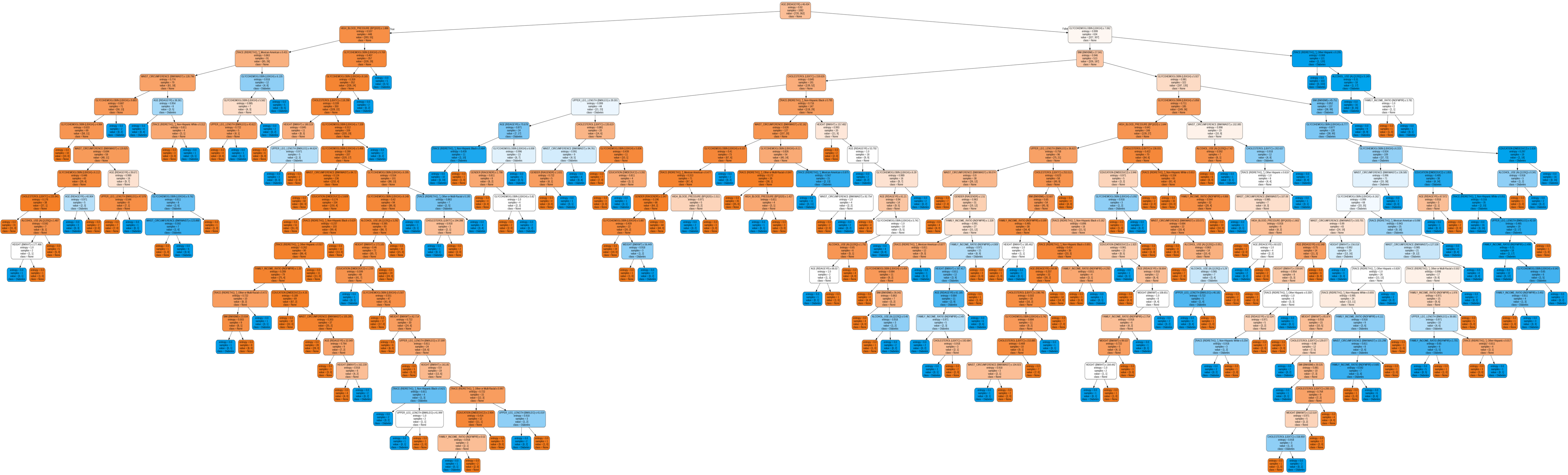
- “non sunt multiplicanda entia praeter necessitatem”
- entities are not to be multiplied beyond necessity
- also called Ockham’s Razor, Law of Economy, or Law of Parsimony

Key Idea: The simplest consistent explanation is the best

(Recall: this is also why we used “regularization” in linear and logistic regression.)

https://en.wikipedia.org/wiki/William_of_Ockham

DT with random features



How could we make smaller trees (and keep Occam happy)?

Recap: ID3 learning approach

Top-Down Decision Tree Induction

[ID3 (1986), C4.5(1993) by Quinlan]

Let $\mathcal{D}$ be a set of labeled instances; $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N} = [X_{N\times D}, \boldsymbol{y}_{N\times 1}]$

Let $\mathcal{D}[X_j = v]$ be the subset of $\mathcal{D}$ where feature X_j has value v

function `train_tree(`$\mathcal{D}$`)`

1. If data $\mathcal{D}$ all have the same label y, return new `leaf_node(`y`)`
2. Pick the "best" feature X_j to partition $\mathcal{D}$
3. Set `node = new decision_node(`X_j`)`
4. For each value v that X_j can take
   - Recursively create a new child `train_tree(`$\mathcal{D}[X_j = v]$`)` of `node`
5. Return `node`

The only way to stop growing a tree larger is to get to homogenous decision nodes where all samples have the same label

Decision Tree Classifier = “20-Questions”

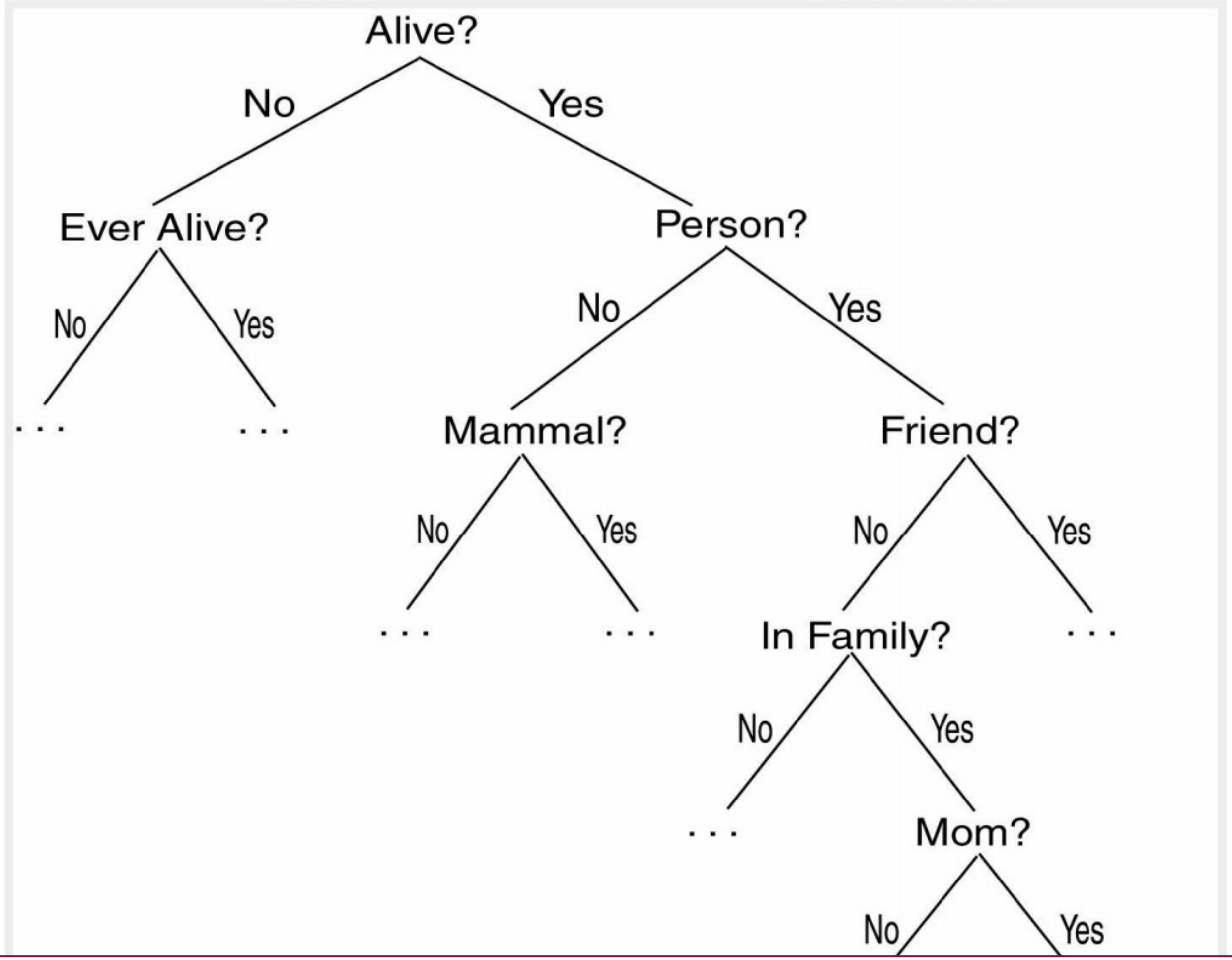
Alice has an object / person in mind

Bob can ask her up to 20 yes/no questions, must guess as quickly as possible

Questions ≈ Decision Tree nodes

Number of questions ≈ depth of tree

Identity ≈ Category Label



Intuitively, must ask questions such that we expect the answers to:

- “rule out as many category options as possible”
- “reveal as much information about the label as possible”

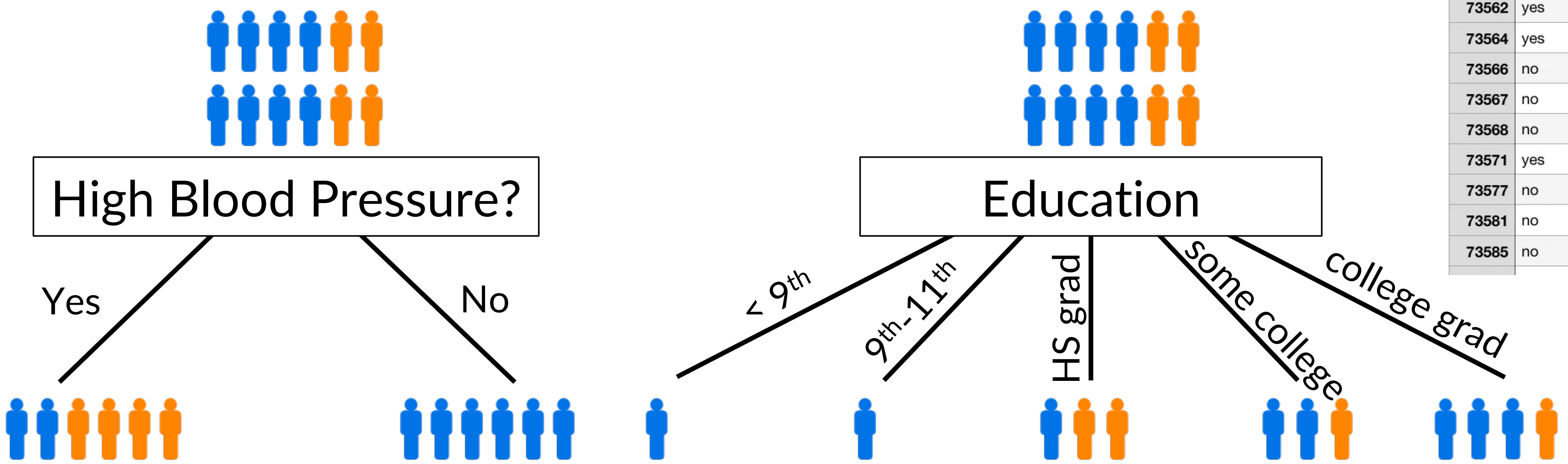
A Measure of Impurity

Choosing Features for Short Decision Trees

Key Idea: good features ideally partition the data into subsets that are either "all positive" (blue) or "all negative" (orange)

Subset of Data

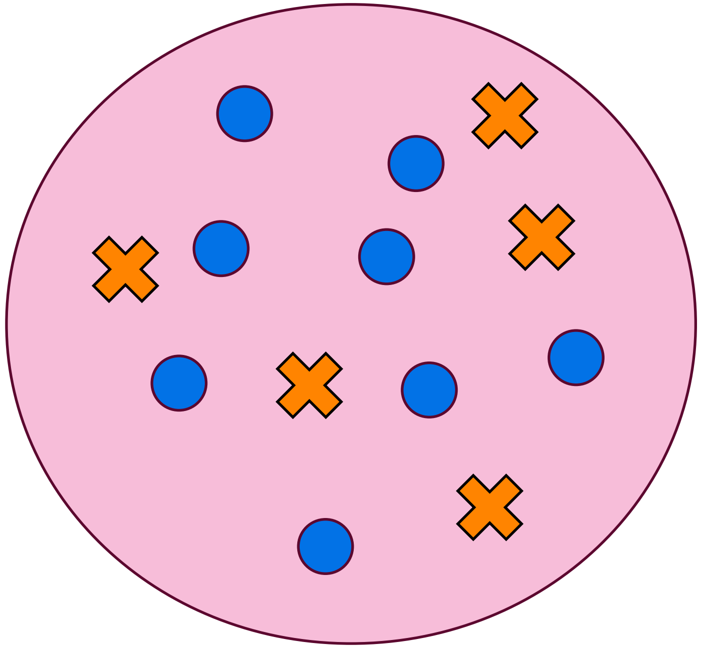
ID (SEQN)	HIGH_BP (BPQ020)	EDUCATION (DMDEDUC2)	DIABETIC
73557	yes	high school graduate / GED	yes
73558	yes	high school graduate / GED	yes
73559	yes	some college or AA degree	yes
73562	yes	some college or AA degree	no
73564	yes	college graduate or above	no
73566	no	high school graduate / GED	no
73567	no	9th-11th grade	no
73568	no	college graduate or above	no
73571	yes	college graduate or above	yes
73577	no	Less than 9th grade	no
73581	no	college graduate or above	no
73585	no	some college or AA degree	no

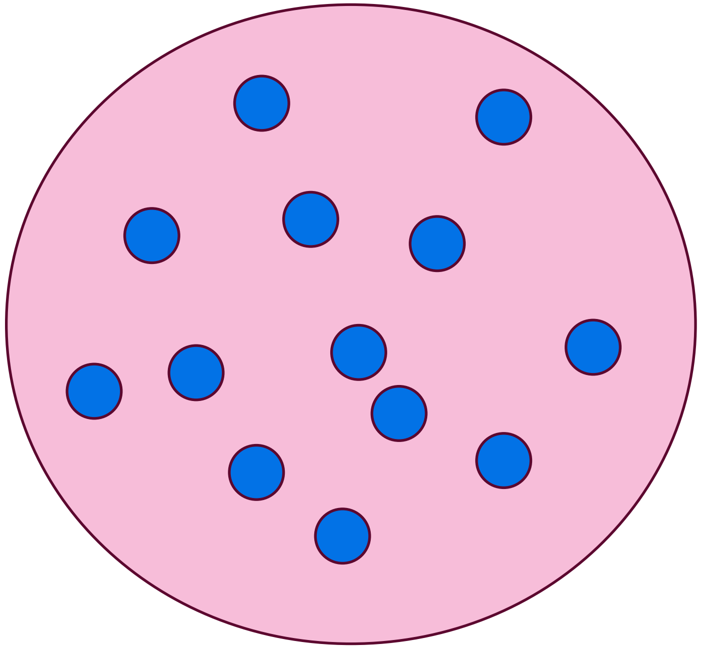


Which split is more informative?

Impurity

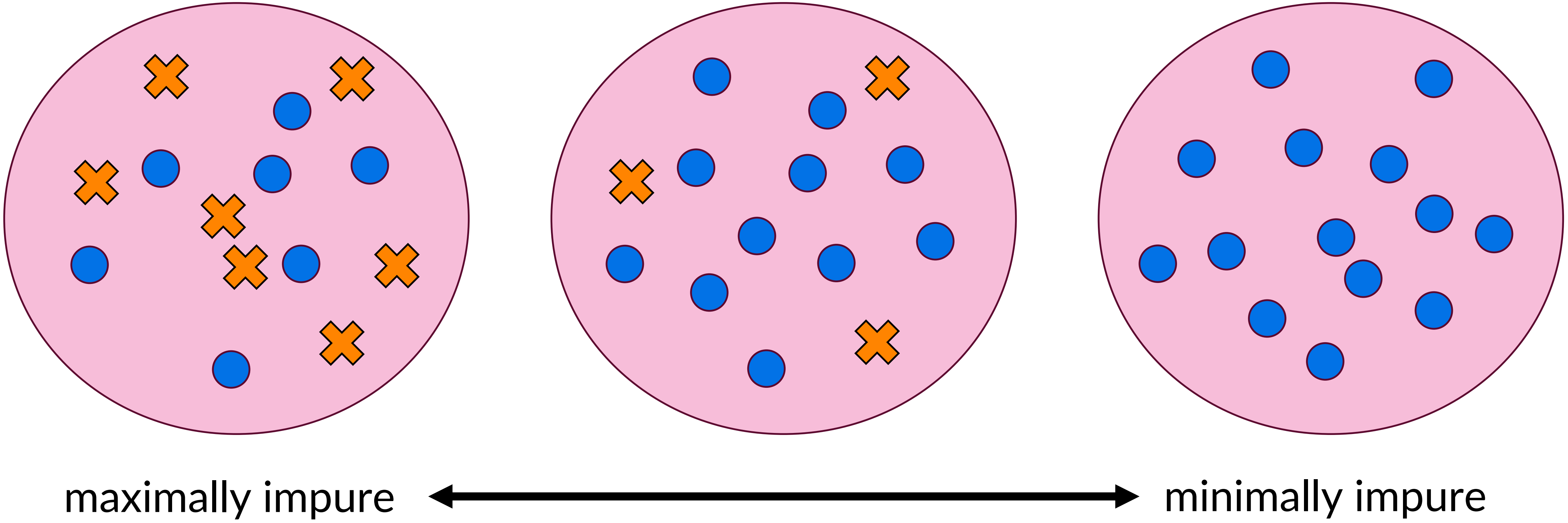
- Measures the level of impurity in a group of samples





Impurity

- Measures the level of impurity in a group of samples



Note: All x's is also "pure"

Could we come up with an "impurity function" of a set of samples?

A Candidate For An "Impurity Function": Entropy

Shannon

- Let Y be any discrete random variable that can take on n values
- The entropy of Y is given by

$$H(Y) = -\sum_{i=1}^{n} P(Y=i) \log_2 P(Y=i)$$

Strictly, the entropy $H(Y)$ maps from a probability distribution (over the class label random variable Y) to an impurity score

$\Updownarrow$

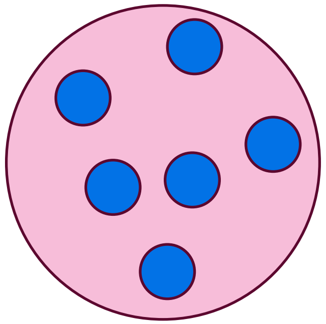
We'll denote $H(\mathcal{D})$ to map from a data subset $\mathcal{D}$ to the impurity score, by setting probability distribution $\approx$ distribution of labels Y in $\mathcal{D}$

Entropy of Binary Classes

Entropy $H(\mathcal{D}) = -\sum_c P(Y=c)\log_2 P(Y=c)$,
where different $c's$ correspond to different class labels

Min Impurity

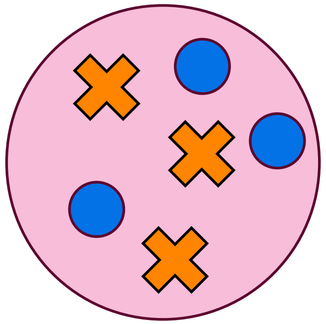
All instances in same class



$H(\mathcal{D}) = -1\log 1$
$= 0$

Max Impurity

Instances split evenly among classes



$H(\mathcal{D}) = -0.5\log 0.5 - 0.5\log 0.5$
$= 1$

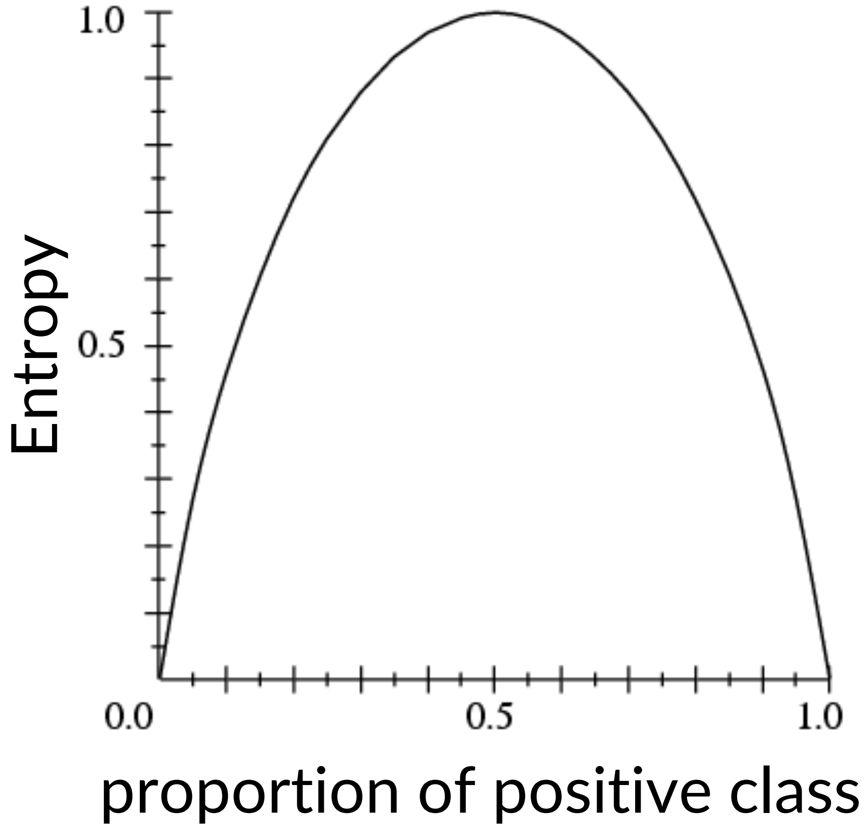
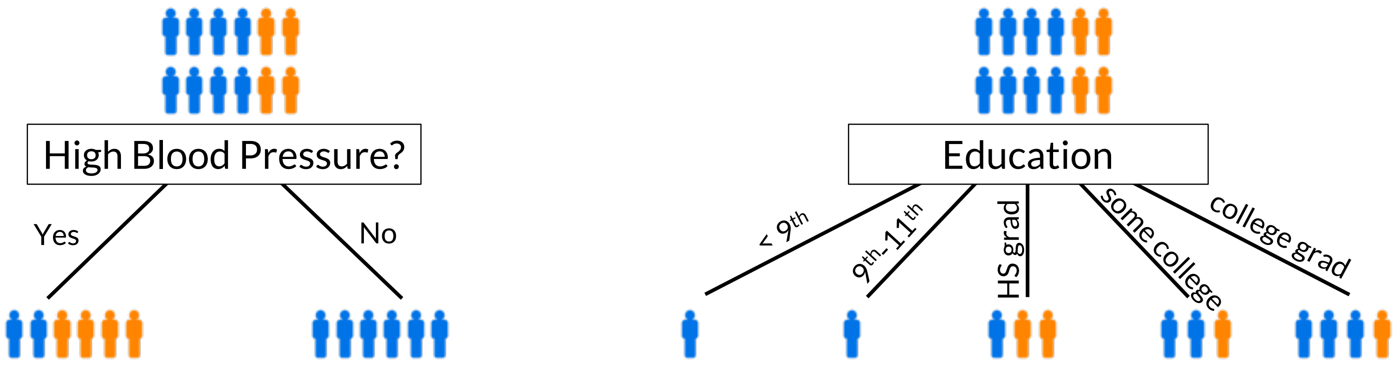


Image: Tom Mitchell

Choosing Features for Short Decision Trees



Recall: Ask questions such that the answers will reduce impurity in child nodes

When considering splitting on attribute / feature X_j,

- Need to estimate the "**expected** drop in impurity" after "getting the answer"/partitioning the data
- "Information Gain" based on our entropy function:

$$\text{IG}(\mathcal{D}, X_j) = H(\mathcal{D}) - \sum_v H(\mathcal{D}[X_j = v]) P(X_j = v)$$

Information Gain

Entropy $H(\mathcal{D}) = -\sum_c P(Y = c) \log_2 P(Y = c)$,
where different $c's$ correspond to different class labels

$$\text{IG}(\mathcal{D}, X_j) = H(\mathcal{D}) - \sum_v H(\mathcal{D}[X_j = v]) P(X_j = v)$$

- The second term is sometimes called the "conditional entropy":

$$H(\mathcal{D}|X_j) = \sum_v H(\mathcal{D}[X_j = v]) P(X_j = v)$$

$E[?]$

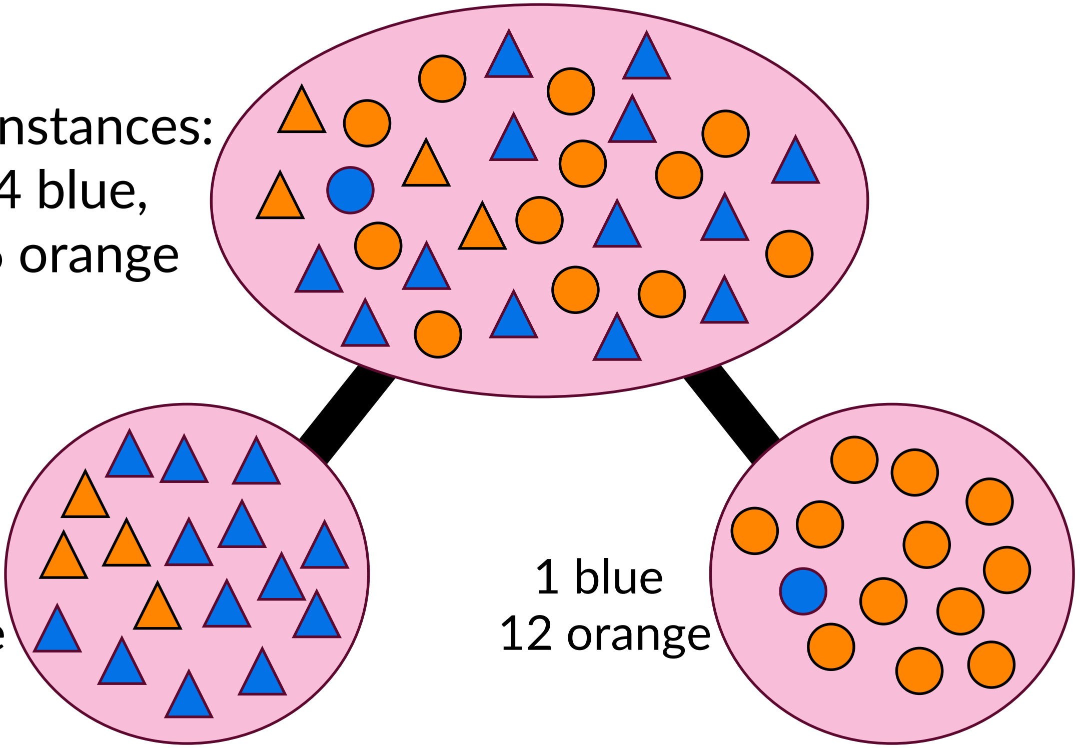
- The information gain may then also be written as:

$$IG(\mathcal{D}, X_j) = H(\mathcal{D}) - H(\mathcal{D}|X_j)$$

Example IG Calculation

$$\text{IG}(\mathcal{D}, X_j) = H(\mathcal{D}) - \sum_v H(\mathcal{D}[X_j = v])P(X_j = v)$$

30 instances:
14 blue,
16 orange



13 blue
4 orange

1 blue
12 orange

H(parent) =

$$-\left(\frac{14}{30}\log_2\frac{14}{30}\right) - \left(\frac{16}{30}\log_2\frac{16}{30}\right)$$
$$= 0.996$$

weighted_mean(H(children)) =

$$\frac{17}{30}\cdot 0.787 + \frac{13}{30}\cdot 0.391$$
$$= 0.615$$

H(child)) =

$$-\left(\frac{13}{17}\log_2\frac{13}{17}\right) - \left(\frac{4}{17}\log_2\frac{4}{17}\right)$$
$$= 0.787$$

H(child)) =

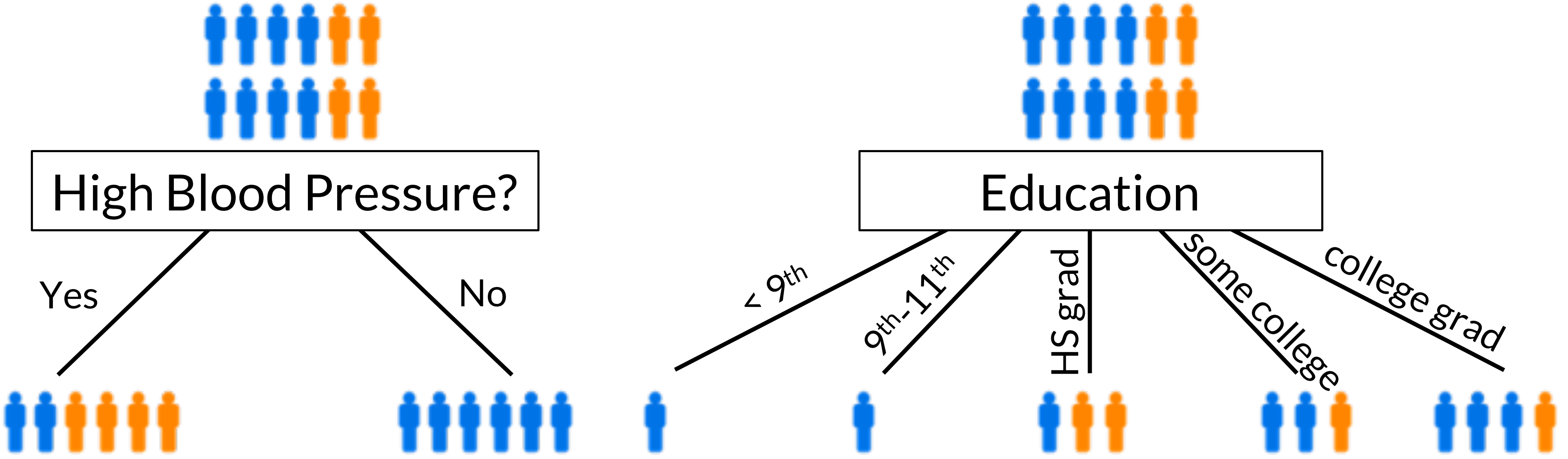
$$-\left(\frac{1}{13}\log_2\frac{1}{13}\right) - \left(\frac{12}{13}\log_2\frac{12}{13}\right)$$
$$= 0.391$$

IG = 0.996 – 0.615 = 0.381

Revisiting Our Diabetes Example

ID (SEQN)	HIGH_BP (BPQ020)	EDUCATION (DMDEDUC2)	DIABETIC
73557	yes	high school graduate / GED	yes
73558	yes	high school graduate / GED	yes
73559	yes	some college or AA degree	yes
73562	yes	some college or AA degree	no
73564	yes	college graduate or above	no
73566	no	high school graduate / GED	no
73567	no	9th-11th grade	no
73568	no	college graduate or above	no
73571	yes	college graduate or above	yes
73577	no	Less than 9th grade	no
73581	no	college graduate or above	no
73585	no	some college or AA degree	no

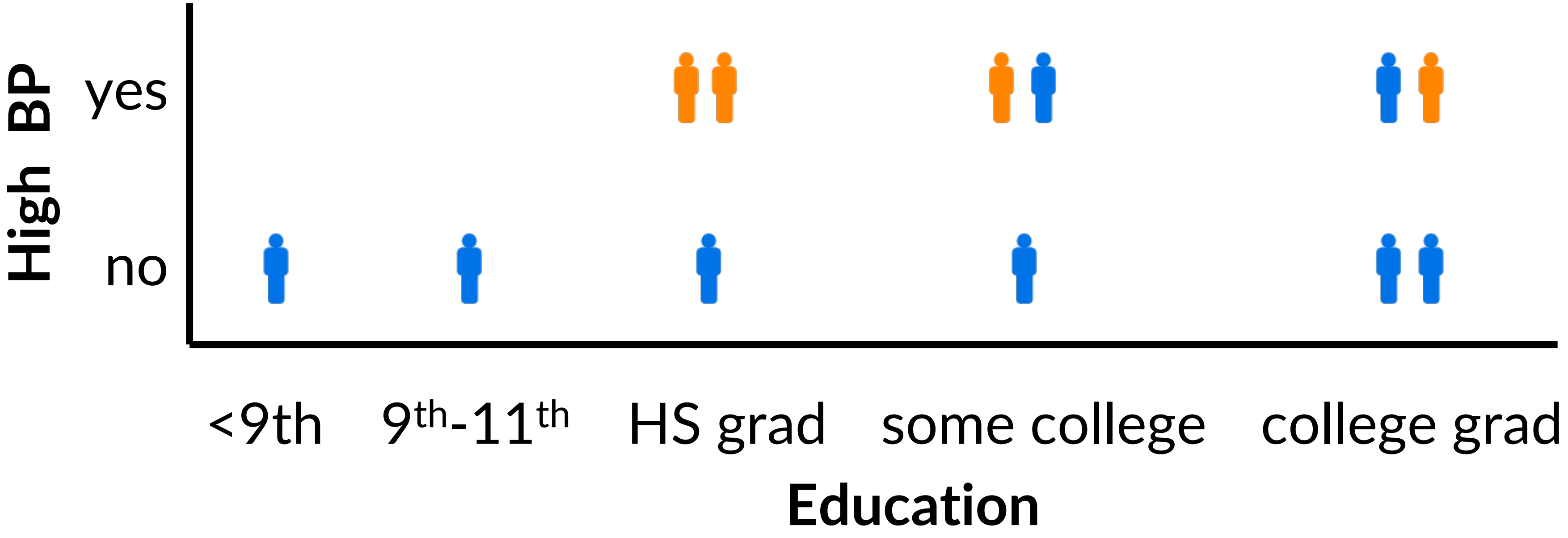
Which split is more informative?



Now we can solve it computationally via information gain

Information Gain For Diabetes Example

ID (SEQN)	HIGH_BP (BPQ020)	EDUCATION (DMDEDUC2)	DIABETIC
73557	yes	high school graduate / GED	yes
73558	yes	high school graduate / GED	yes
73559	yes	some college or AA degree	yes
73562	yes	some college or AA degree	no
73564	yes	college graduate or above	no
73566	no	high school graduate / GED	no
73567	no	9th-11th grade	no
73568	no	college graduate or above	no
73571	yes	college graduate or above	yes
73577	no	Less than 9th grade	no
73581	no	college graduate or above	no
73585	no	some college or AA degree	no



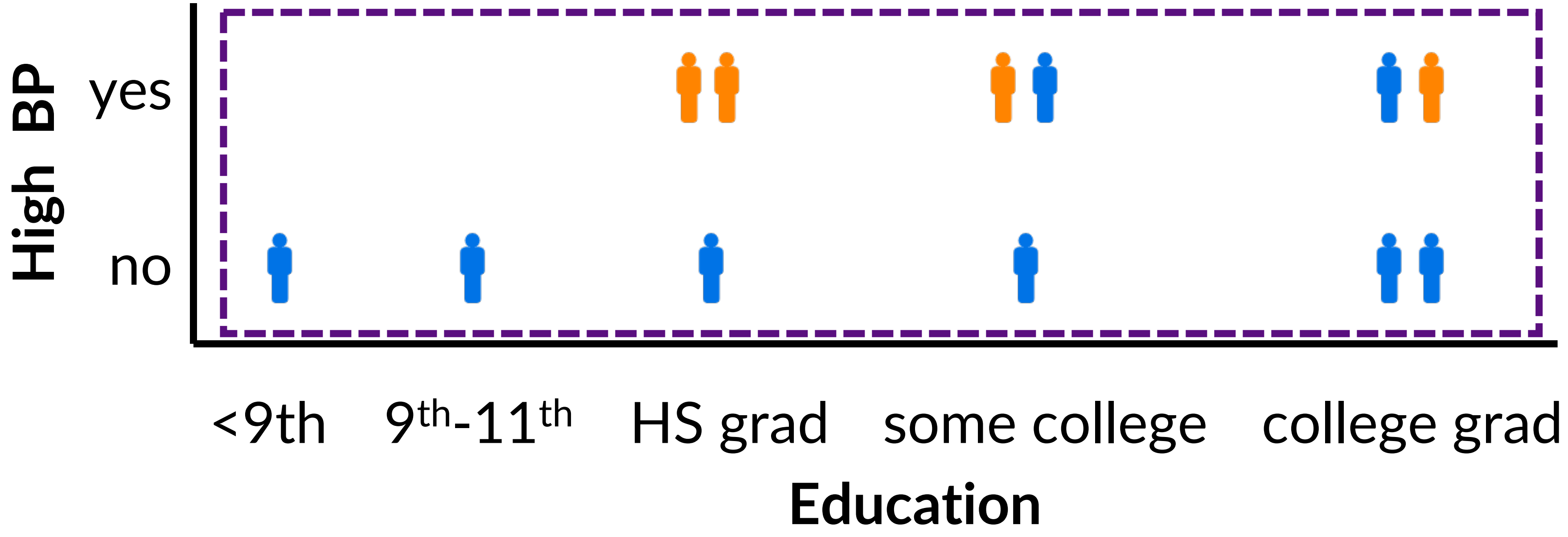
Need to compute:

$$IG(\mathcal{D}, High\ BP) = H(\mathcal{D}) - H(\mathcal{D} \mid High\ BP)$$

$$IG(\mathcal{D}, Education) = H(\mathcal{D}) - H(\mathcal{D} \mid Education)$$

Information Gain For Diabetes Example

ID (SEQN)	HIGH_BP (BPQ020)	EDUCATION (DMDEDUC2)	DIABETIC
73557	yes	high school graduate / GED	yes
73558	yes	high school graduate / GED	yes
73559	yes	some college or AA degree	yes
73562	yes	some college or AA degree	no
73564	yes	college graduate or above	no
73566	no	high school graduate / GED	no
73567	no	9th-11th grade	no
73568	no	college graduate or above	no
73571	yes	college graduate or above	yes
73577	no	Less than 9th grade	no
73581	no	college graduate or above	no
73585	no	some college or AA degree	no



Need to compute:

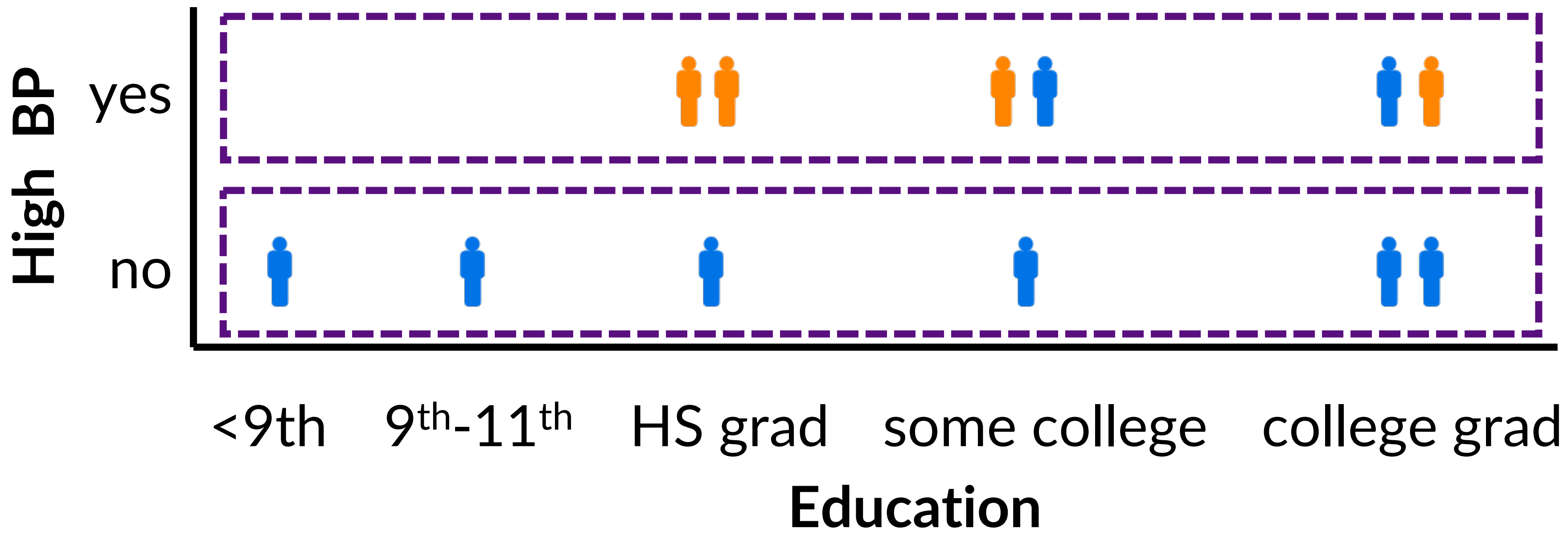
$$IG(\mathcal{D}, High\ BP) = H(\mathcal{D}) - H(\mathcal{D} \mid High\ BP)$$

$$IG(\mathcal{D}, Education) = H(\mathcal{D}) - H(\mathcal{D} \mid Education)$$

$$H(\mathcal{D}) = -4/12 \lg 4/12 - 8/12 \lg 8/12 = 0.918$$

Information Gain For Diabetes Example

ID (SEQN)	HIGH_BP (BPQ020)	EDUCATION (DMDEDUC2)	DIABETIC
73557	yes	high school graduate / GED	yes
73558	yes	high school graduate / GED	yes
73559	yes	some college or AA degree	yes
73562	yes	some college or AA degree	no
73564	yes	college graduate or above	no
73566	no	high school graduate / GED	no
73567	no	9th-11th grade	no
73568	no	college graduate or above	no
73571	yes	college graduate or above	yes
73577	no	Less than 9th grade	no
73581	no	college graduate or above	no
73585	no	some college or AA degree	no



Need to compute:

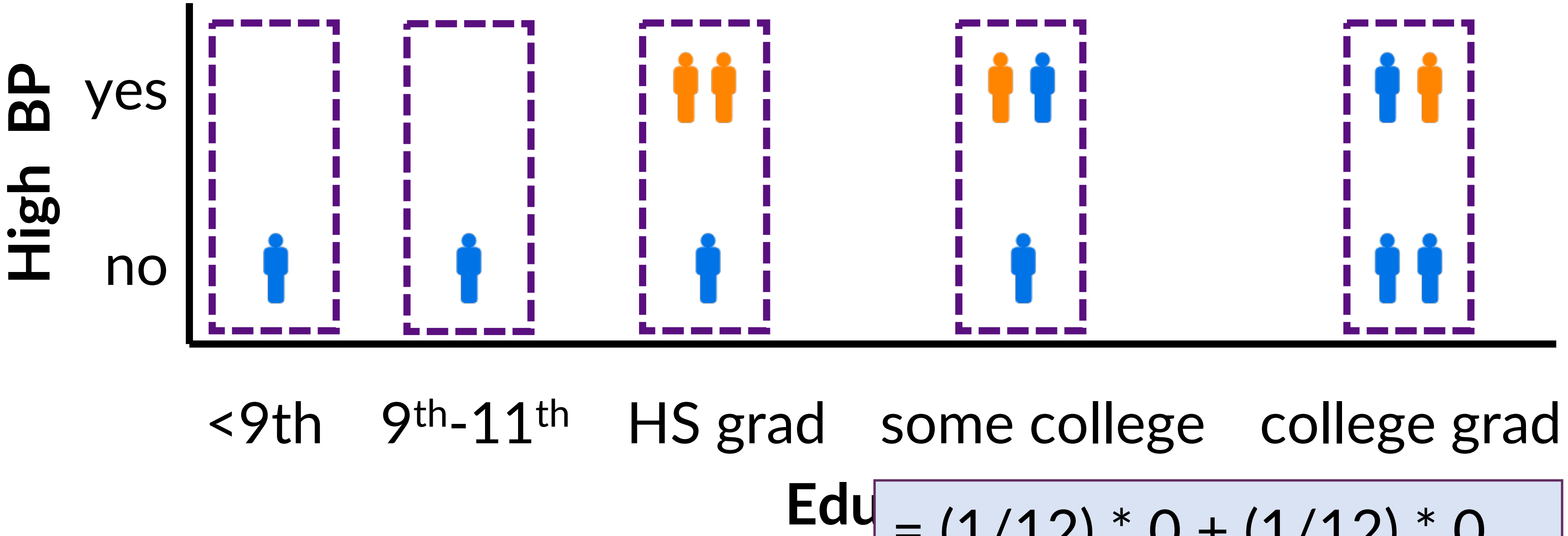
$$IG(\mathcal{D}, High\ BP) = H(\mathcal{D}) - H(\mathcal{D}\,|High\ BP)$$

$$IG(\mathcal{D}, Education) = H(\mathcal{D}) - H(\mathcal{D}|\,Education)$$

= (6/12) * (-2/6 lg 2/6
- 4/6 lg 4/6)
+ (6/12) * (0)
= 0.459

Information Gain For Diabetes Example

ID (SEQN)	HIGH_BP (BPQ020)	EDUCATION (DMDEDUC2)	DIABETIC
73557	yes	high school graduate / GED	yes
73558	yes	high school graduate / GED	yes
73559	yes	some college or AA degree	yes
73562	yes	some college or AA degree	no
73564	yes	college graduate or above	no
73566	no	high school graduate / GED	no
73567	no	9th-11th grade	no
73568	no	college graduate or above	no
73571	yes	college graduate or above	yes
73577	no	Less than 9th grade	no
73581	no	college graduate or above	no
73585	no	some college or AA degree	no



Need to compute:

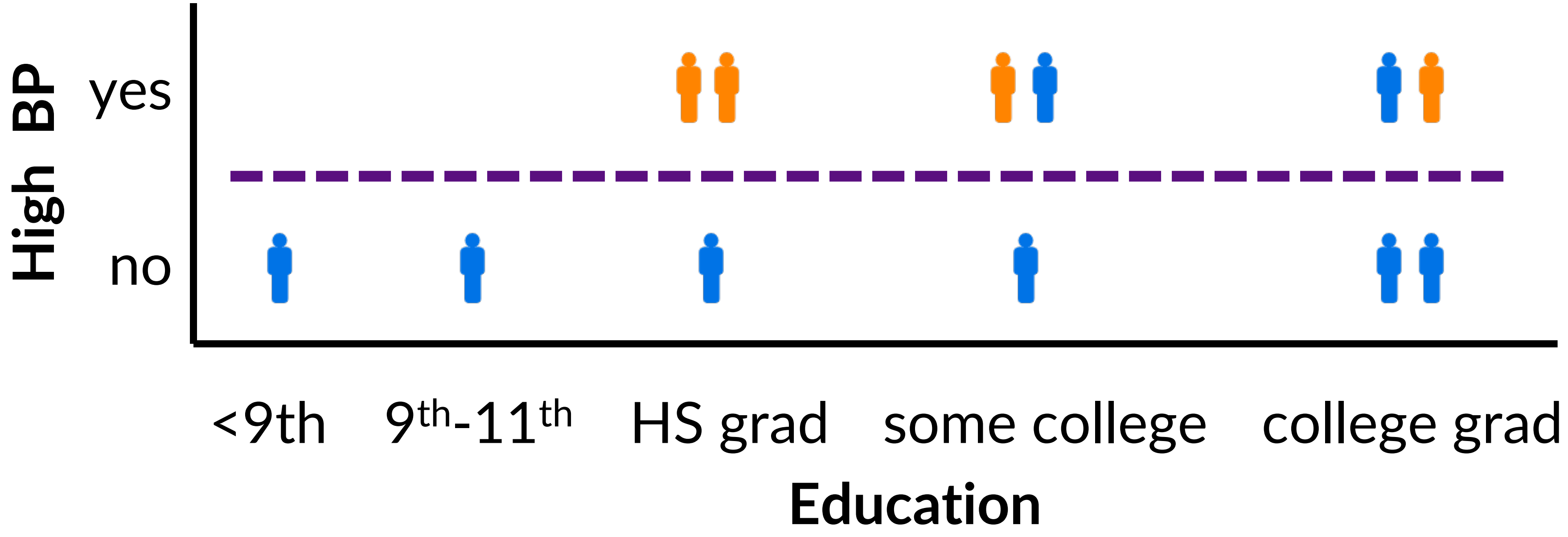
$$IG(\mathcal{D}, High\ BP) = H(\mathcal{D}) - H(\mathcal{D} \mid High\ BP)$$

$$IG(\mathcal{D}, Education) = H(\mathcal{D}) - H(\mathcal{D} \mid Education)$$

$$= (1/12) * 0 + (1/12) * 0$$
$$+ (3/12) * (-1/3 \lg 1/3 - 2/3 \lg 2/3)$$
$$+ (3/12) * (-2/3 \lg 2/3 - 1/3 \lg 1/3)$$
$$+ (4/12) * (-3/4 \lg 3/4 - 1/4 \lg 1/4)$$
$$= 0.730$$

Information Gain For Diabetes Example

ID (SEQN)	HIGH_BP (BPQ020)	EDUCATION (DMDEDUC2)	DIABETIC
73557	yes	high school graduate / GED	yes
73558	yes	high school graduate / GED	yes
73559	yes	some college or AA degree	yes
73562	yes	some college or AA degree	no
73564	yes	college graduate or above	no
73566	no	high school graduate / GED	no
73567	no	9th-11th grade	no
73568	no	college graduate or above	no
73571	yes	college graduate or above	yes
73577	no	Less than 9th grade	no
73581	no	college graduate or above	no
73585	no	some college or AA degree	no



Need to compute:

$$IG(\mathcal{D}, High\ BP) = H(\mathcal{D}) - H(\mathcal{D}\,|High\ BP) = 0.918 - 0.459 = 0.459$$

0.459 ★

$$IG(\mathcal{D}, Education) = H(\mathcal{D}) - H(\mathcal{D}|\ Education) = 0.918 - 0.730 = 0.188$$