METIS

Day 7: Trees and Forests

John Navarro
john.navarro@thisismetis.com
https://www.linkedin.com/in/johnnavarro/

Group Exercise

- **Situation:** A friend of yours who owns a popular restaurant tells you that she's interested in applying data science to help optimize her business. Specifically, she's interested in reducing waste and improving the flow through the restaurant, as well as any other ideas you may have.

- **Your task:** Come up with 2 or 3 recommendations as to how she might accomplish these goals using data science techniques. In your recommendations, be sure to think about what data is available, what features you might collect, and how you'd assess performance.

METIS

Regression Trees

The Data

The dataset is composed of 7 input measurement features (kilograms per meter cubed of concrete):

- Cement
- Slag
- Fly ash
- Water
- SP
- Coarse Aggr.
- Fine Aggr.

And 3 output measurements:

- SLUMP (cm)
- FLOW (cm)
- 28-day Compressive Strength (Mpa)

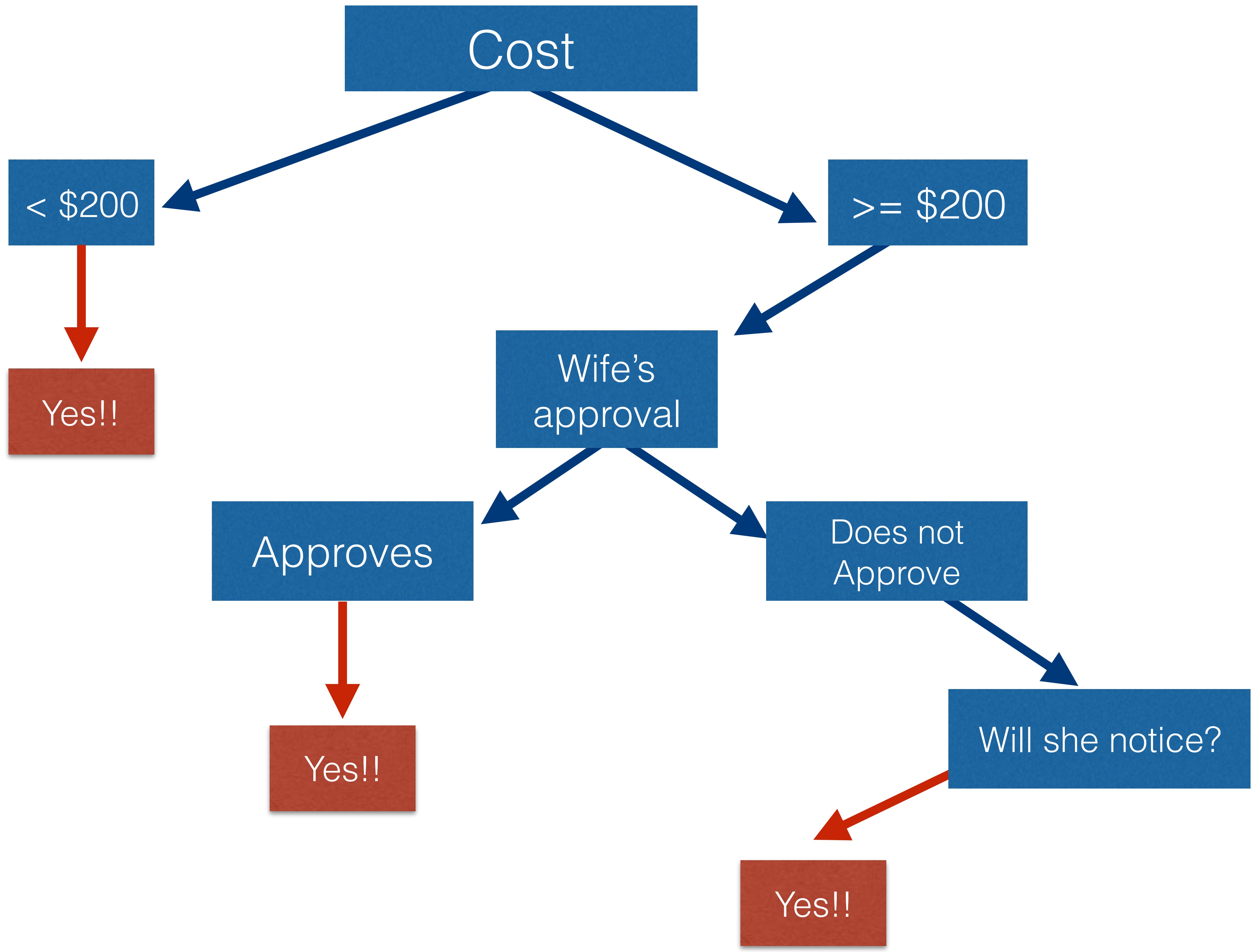
The imports

```
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import BaggingRegressor,
RandomForestRegressor
from sklearn.metrics import mean_squared_error,
accuracy_score
from sklearn.externals.six import StringIO
from IPython.display import Image
import pydotplus
from sklearn.tree import export_graphviz
```

METIS

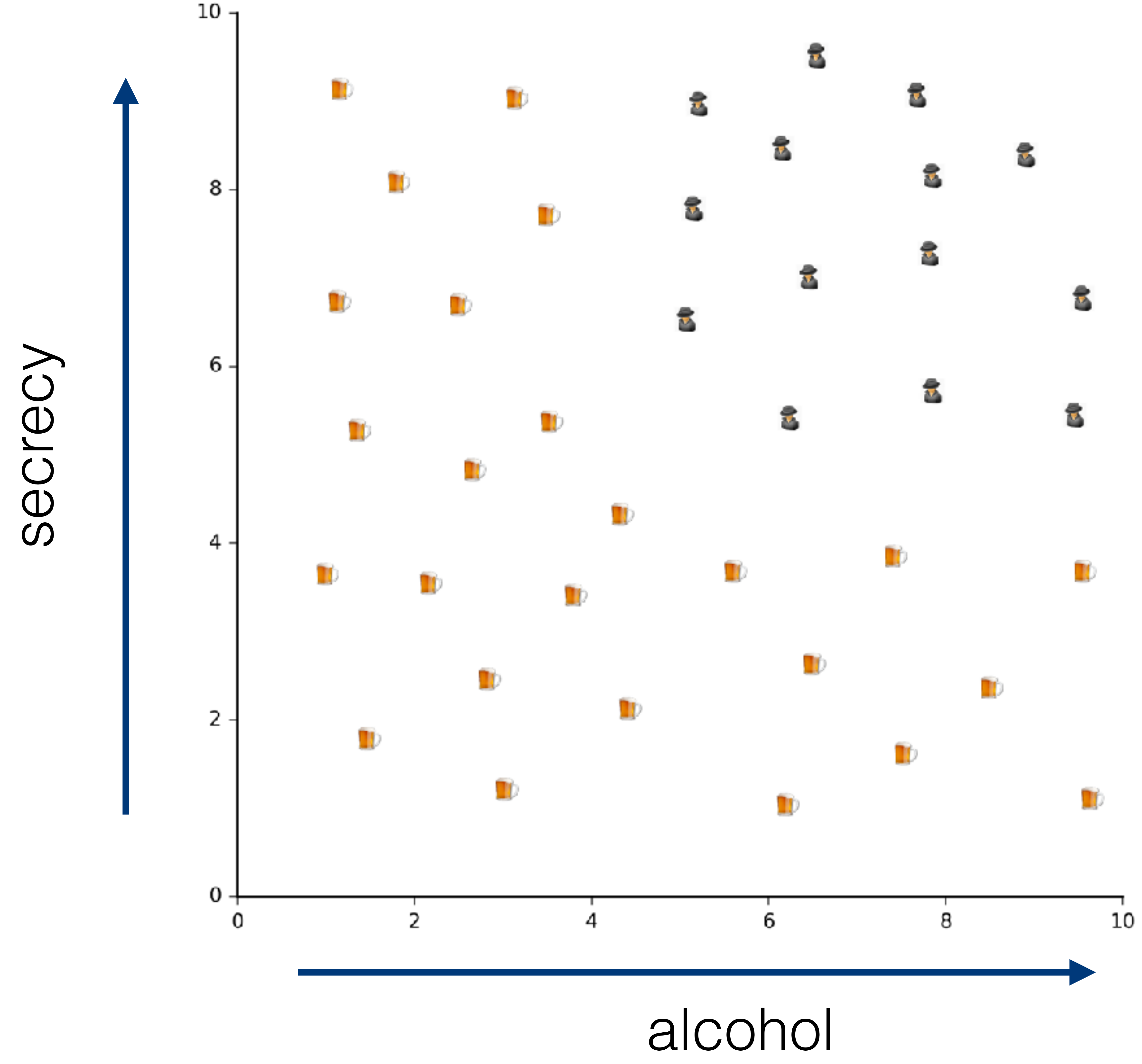
What is a Tree?

Should I buy a new tech gadget?



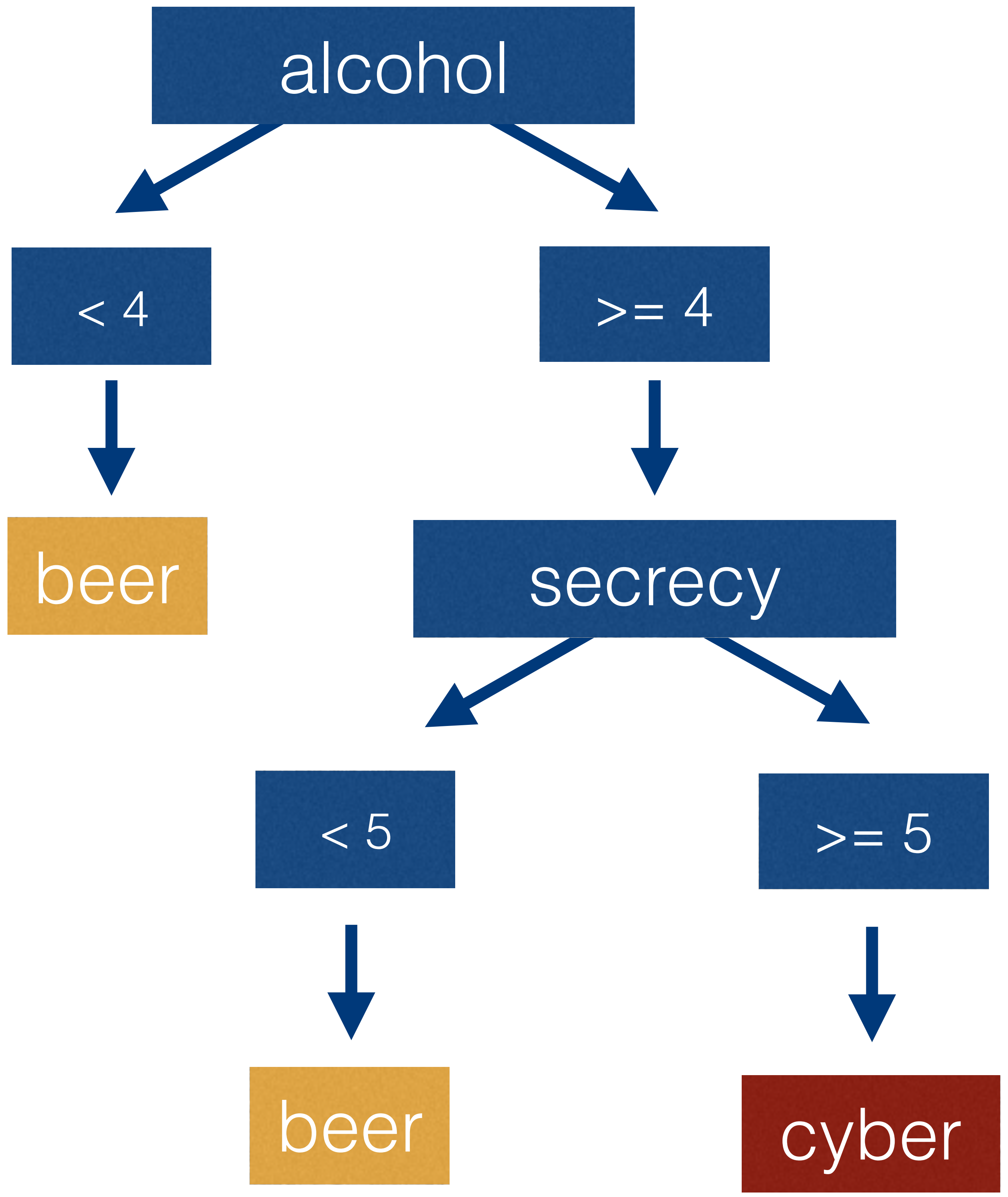
METIS

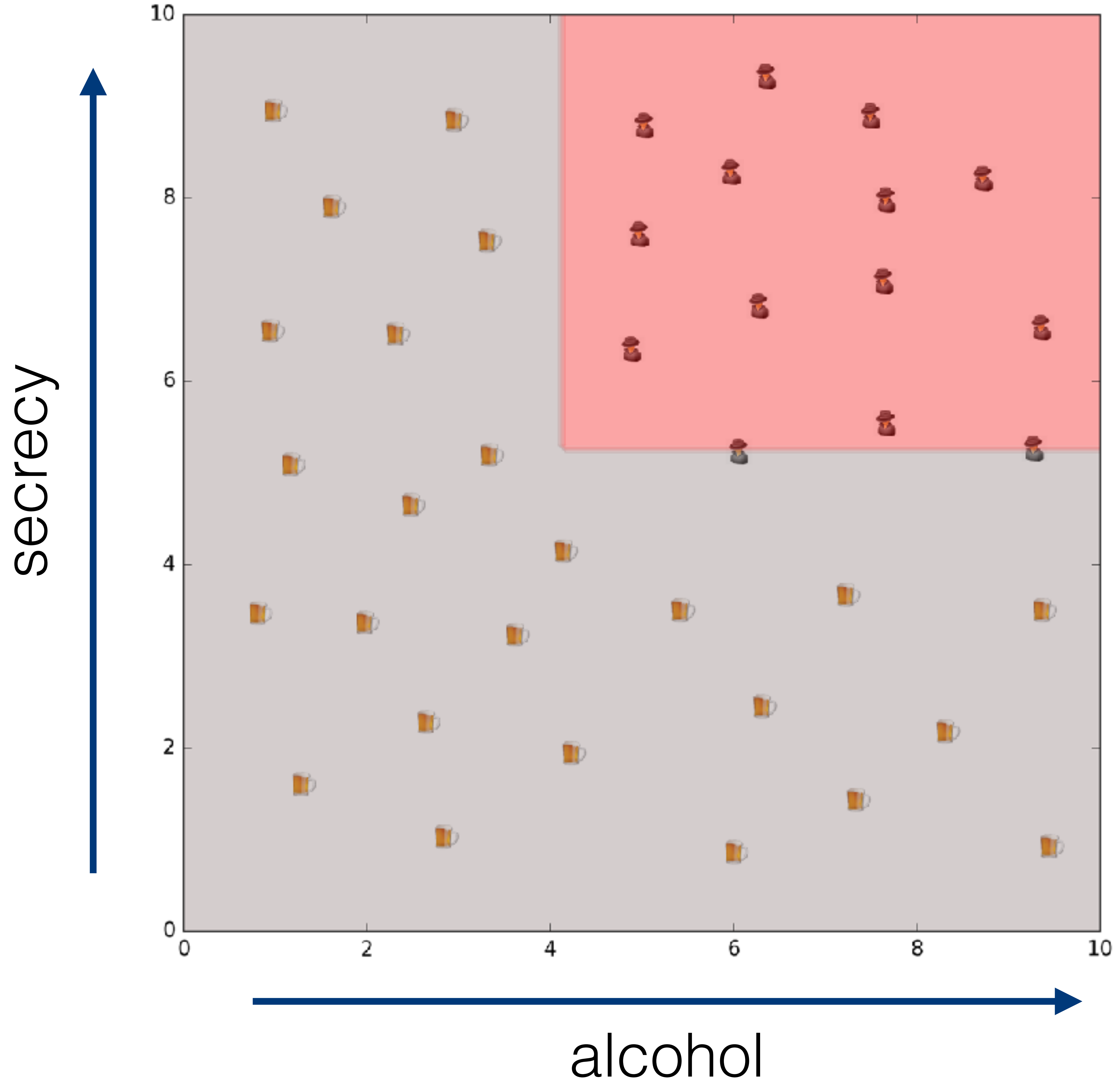
How can we separate beer and cyber?



METIS

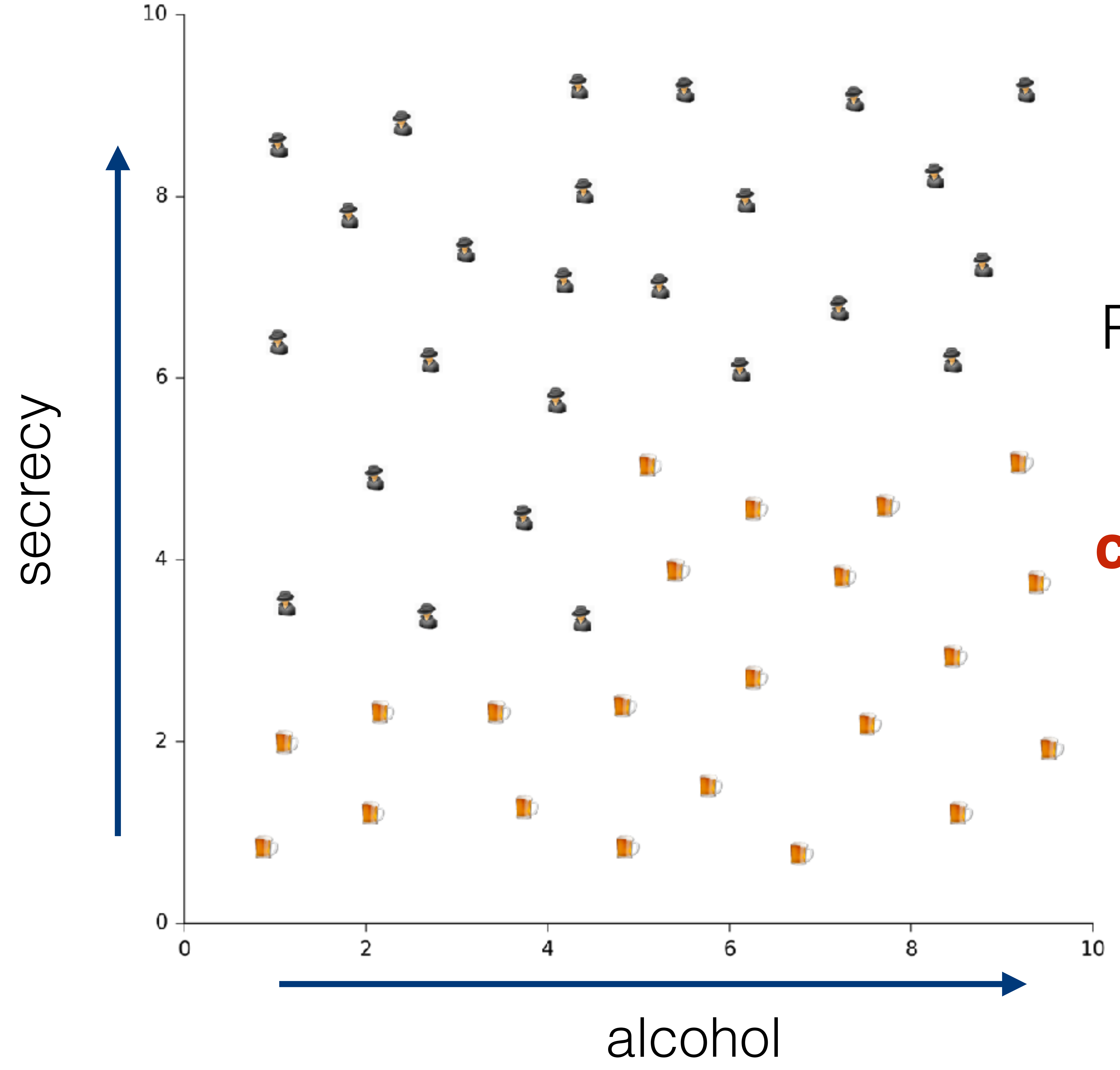
How can we separate beer and cyber?





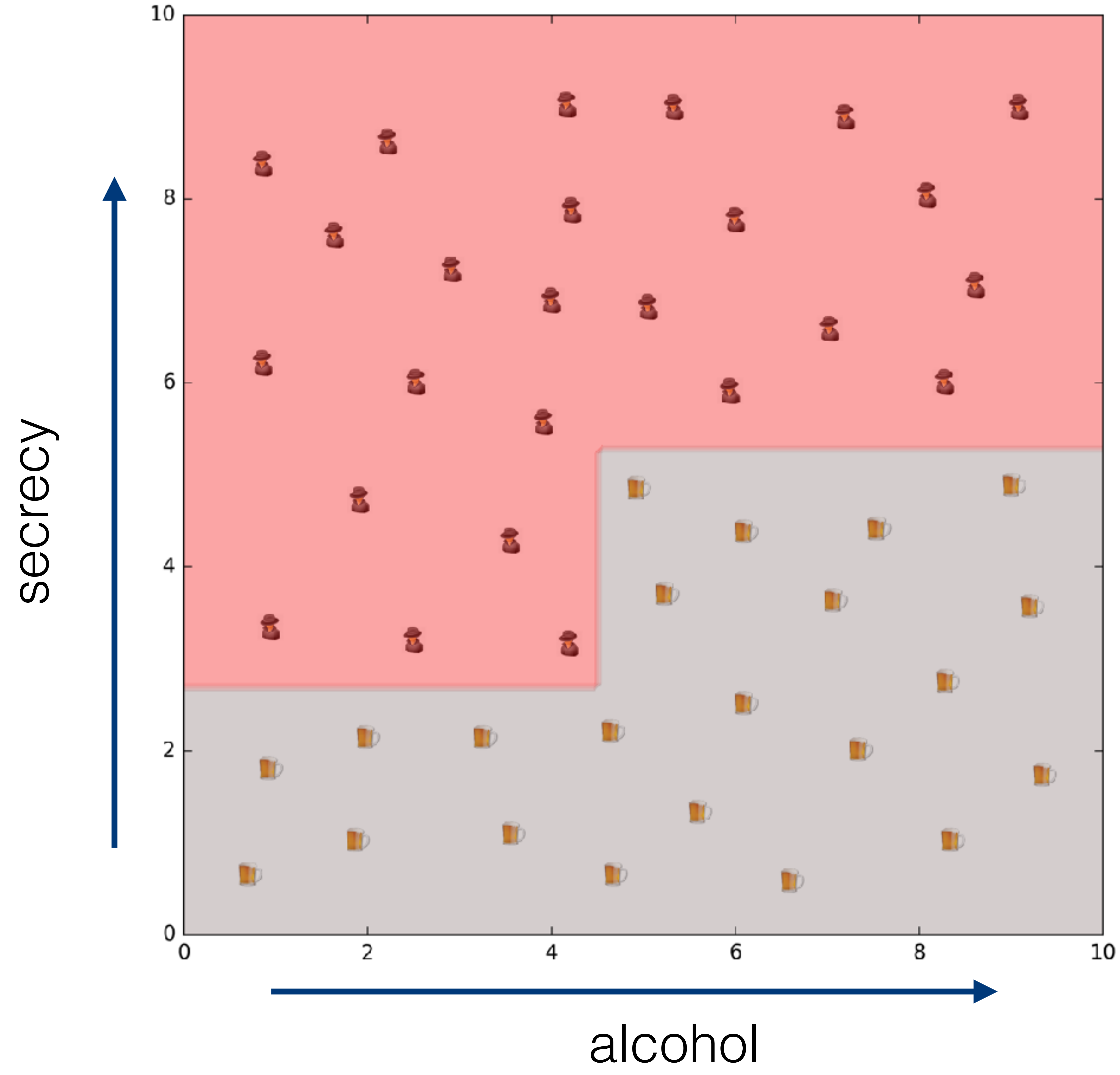
METIS

Pen and paper worksheet: solve!



Please take 10 minutes and complete **printed Worksheet - create a Decision Tree by hand!**

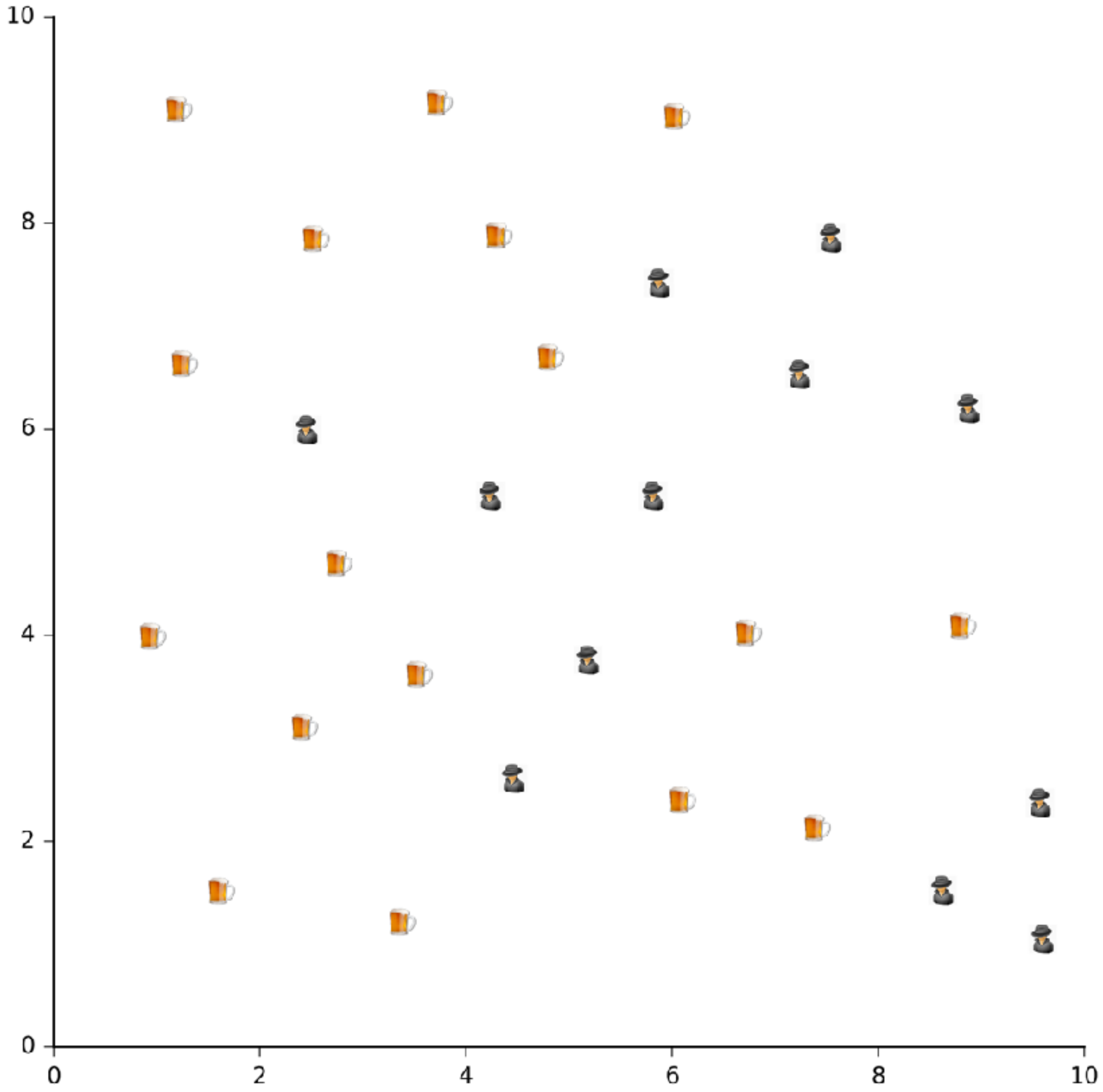
Pen and paper worksheet: solve!



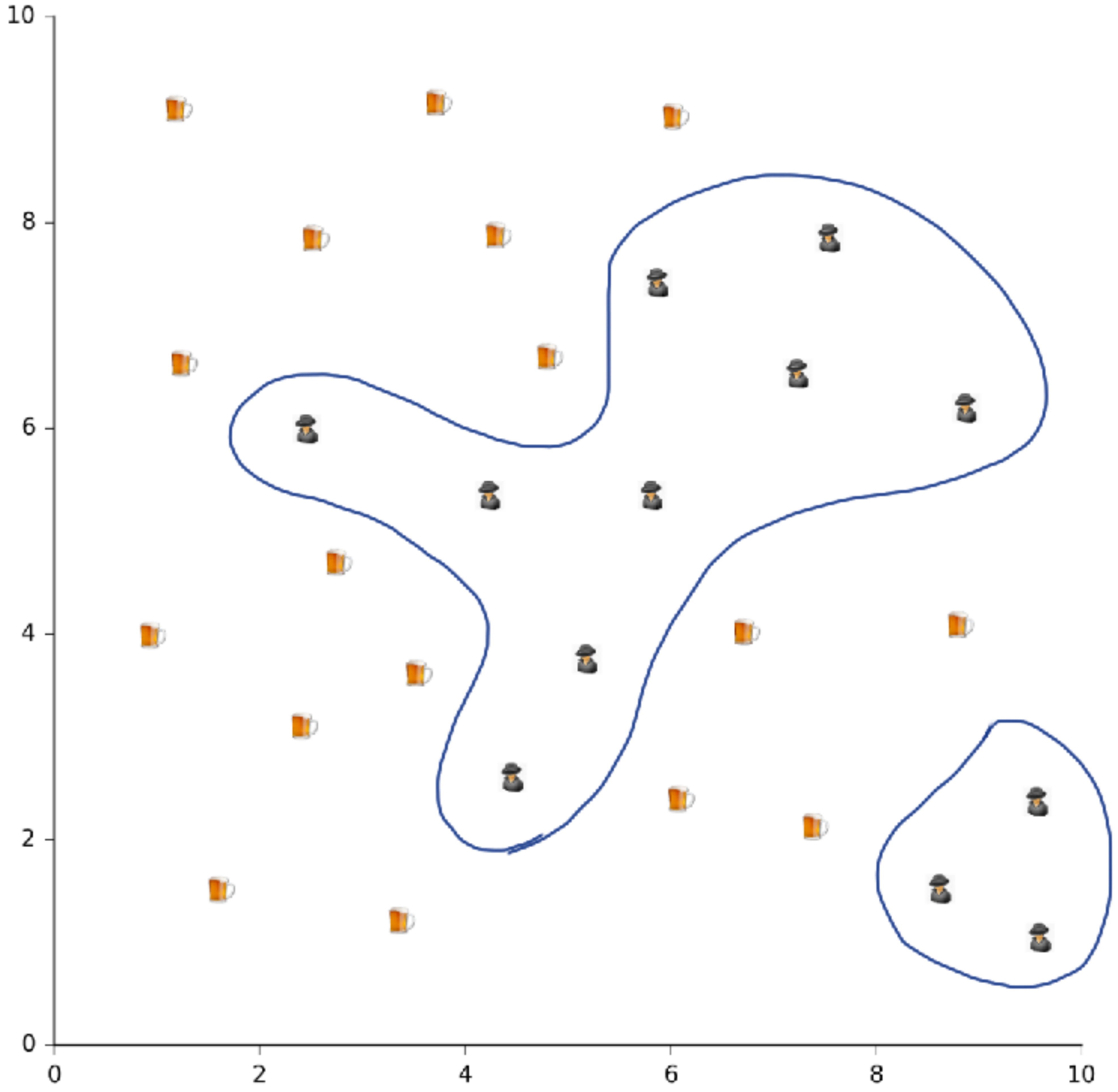
Answer:

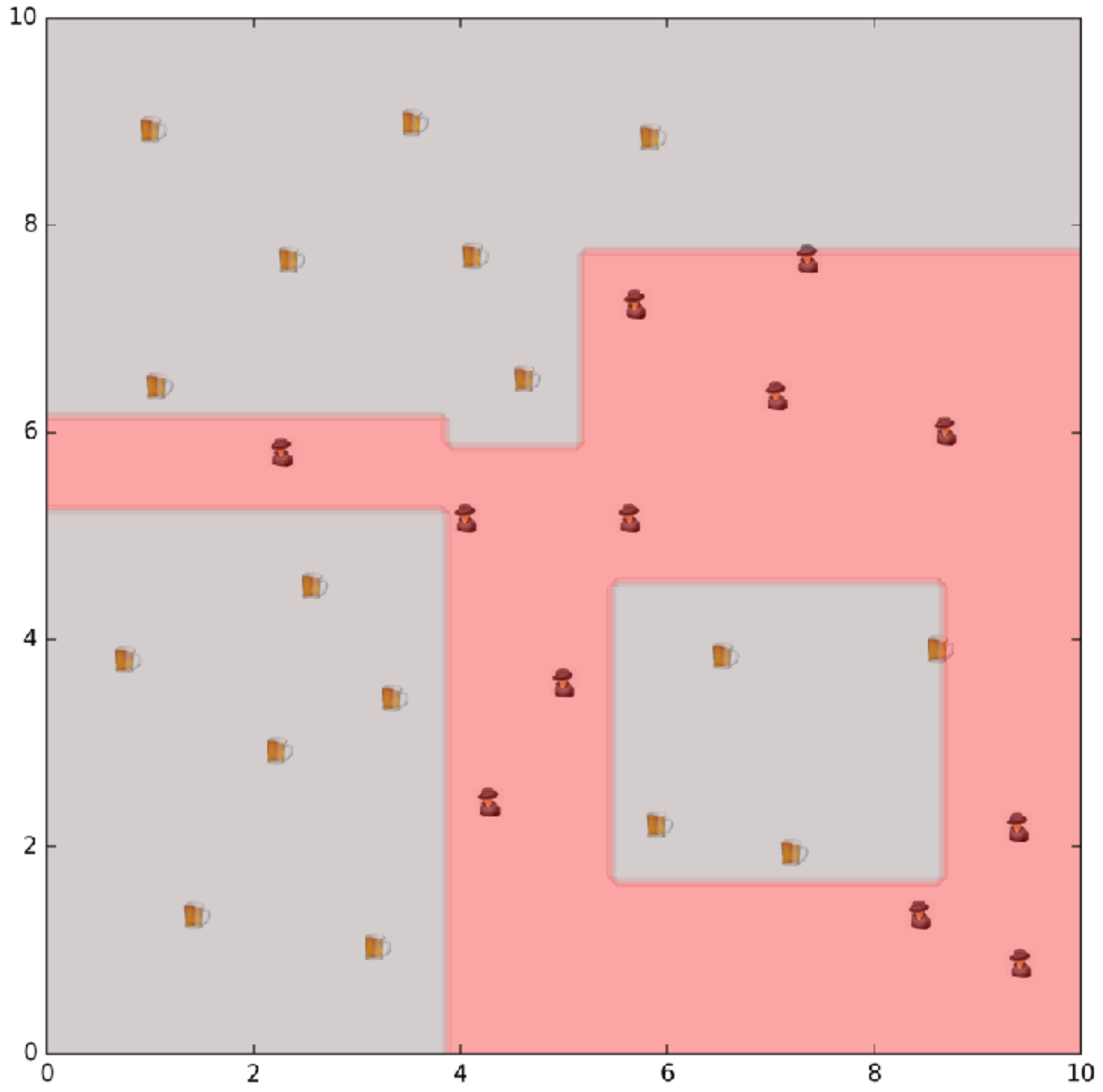
1. secrecy threshold 5.27
2. alcohol threshold 4.47
3. secrecy threshold 2.70

METIS Ultimate last challenge! Who can solve it?



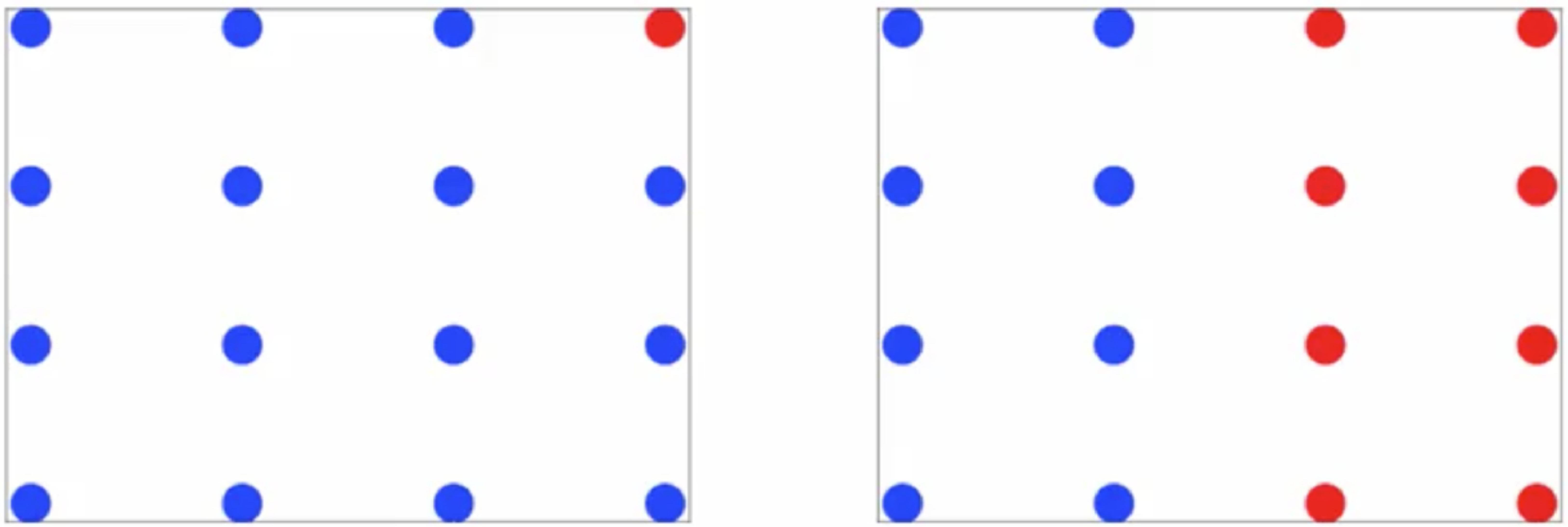
Such boundaries can be a sign of over-fitting!





Lastly: How does a Decision Tree exactly decide to split?

A **B**

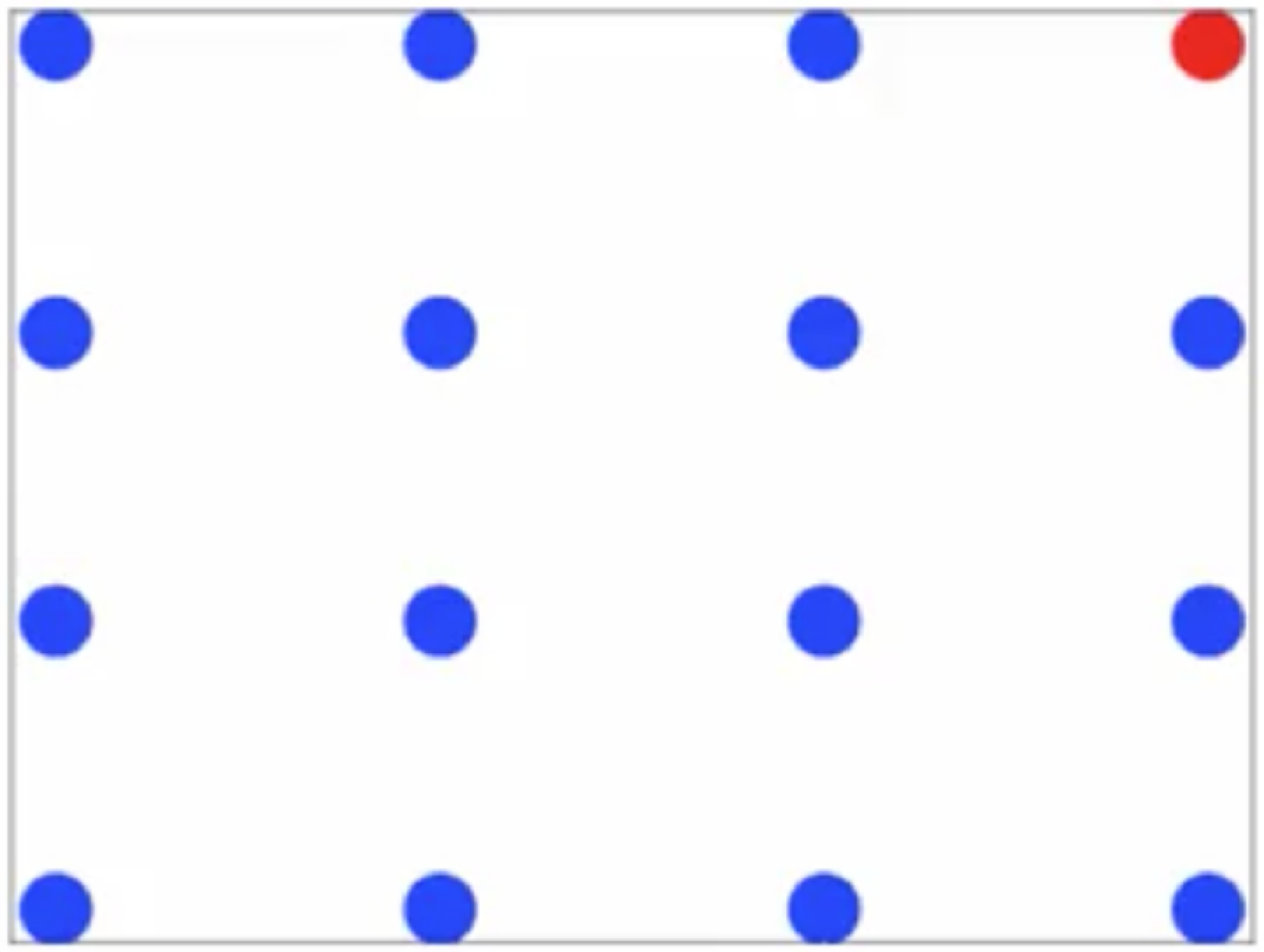


Quiz: By intuition which example appears to be more pure, that is, separates the two classes better?

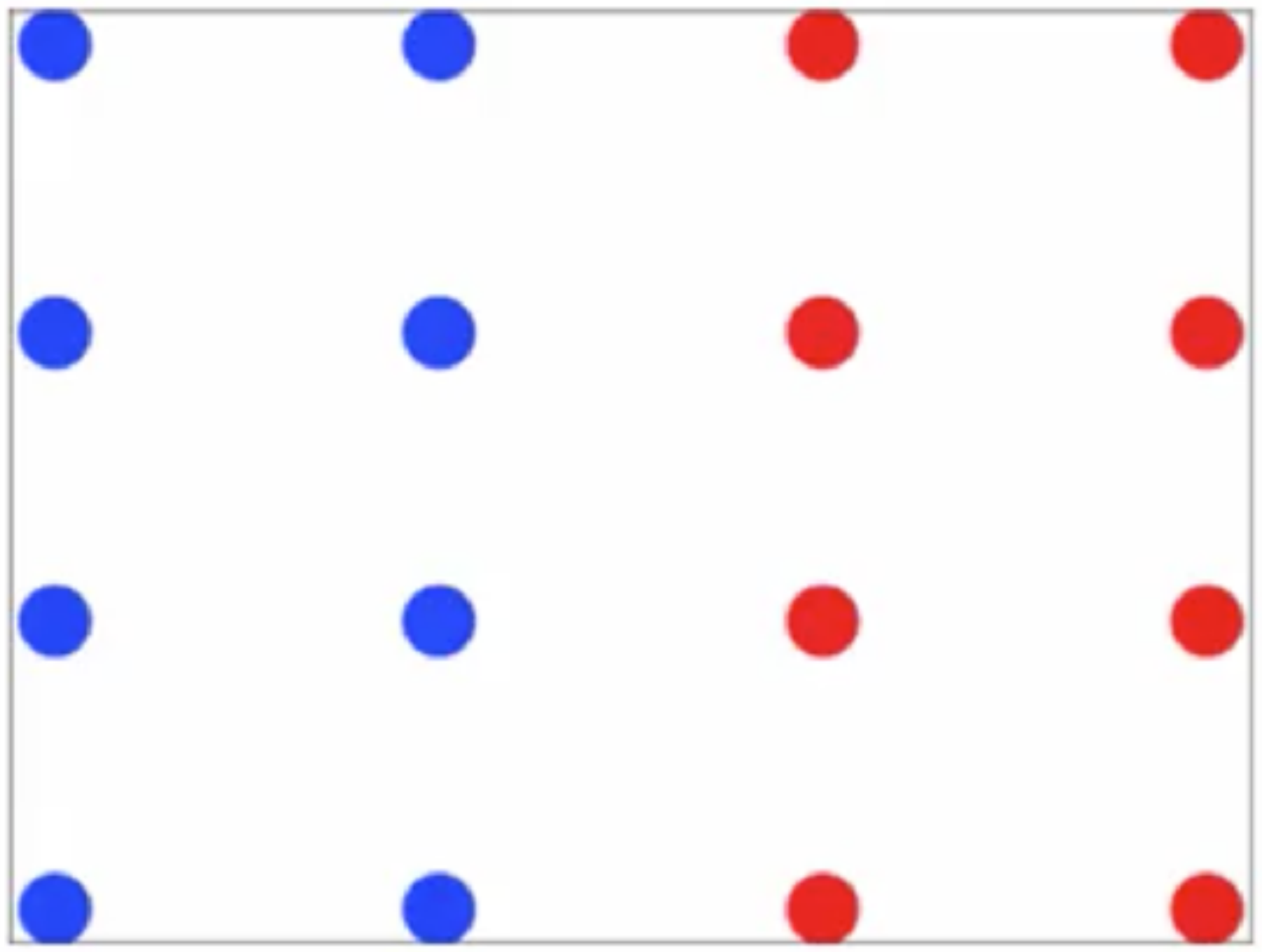
Lastly: How does a Decision Tree exactly decide to split?

Criterion for best splits: **Information Gain**

A



B



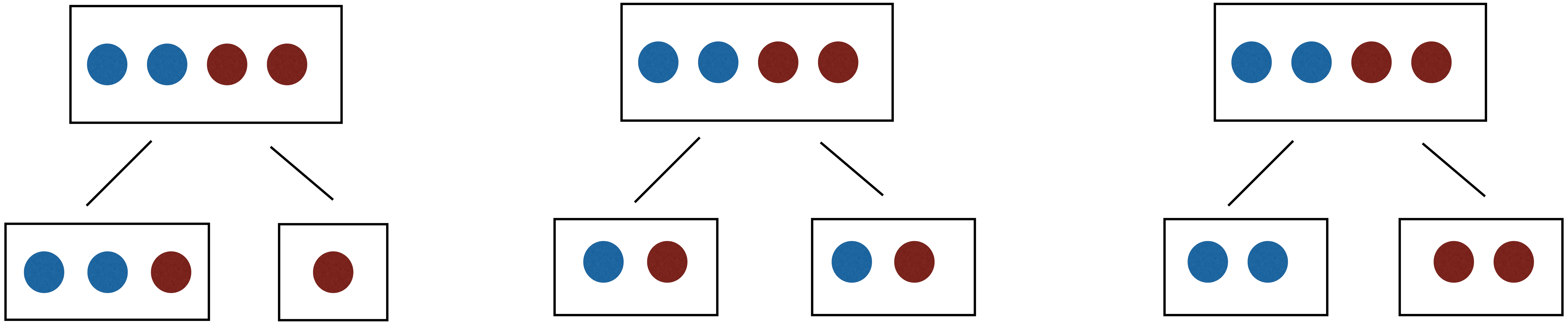
Answer: A

Which feature below gives the best split?

Information Gain = entropy(parent) - [weighted average]entropy(children)

entropy(parent)=1.0
most impure binary system

choosing next best feature to split next



feature 1 feature 2 feature 3

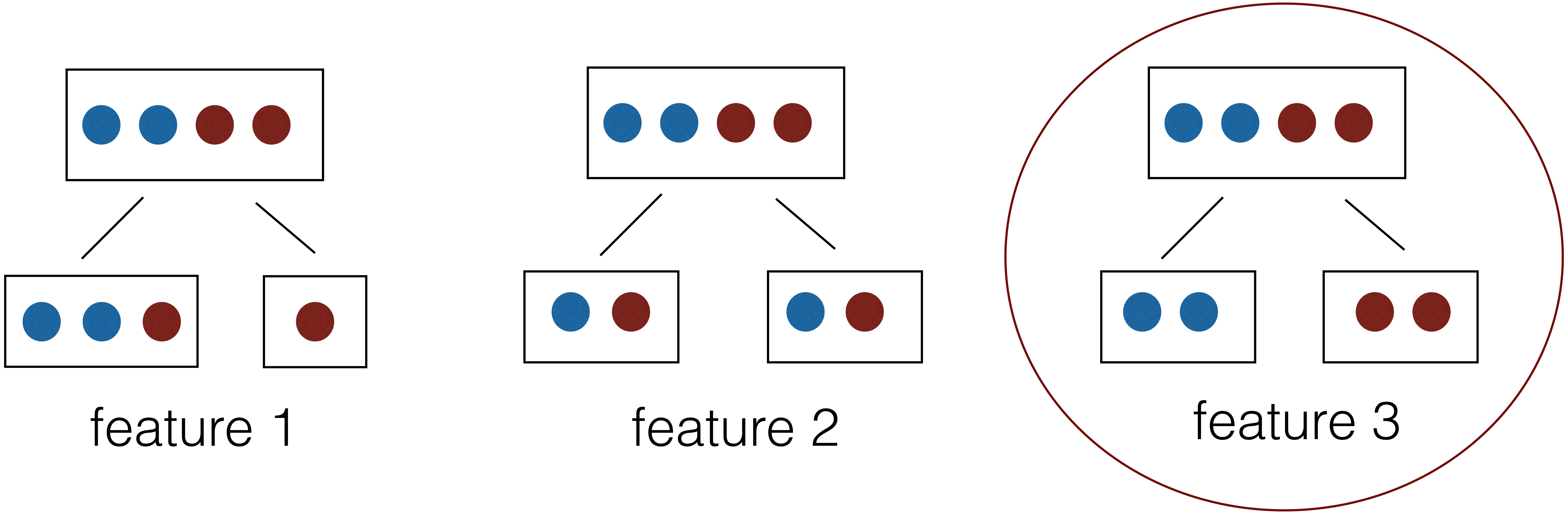
?

Which feature below gives the best split?

Information Gain = entropy(parent) - [weighted average]entropy(children)

entropy(parent)=1.0
most impure binary system

choosing next best feature to split next



Information Gain	**0.3112**	**0**	**1**

METIS

Regression Trees

Lets build a model using just these two features that can give reasonable predictions on `compressive_strength`. We will build it as follows:

- Segment the whole space of `cement`/`fly_ash` possibilities into distinct regions
- Use the mean compressive_strength in each region as the predicted `compressive_strength` for that combination of `cement` to `fly_ash` for future concrete samples.
- Intuitively, we want to **maximize** the similarity (or "homogeneity") of `compressive_strength` within a given region, and **minimize** the similarity of `compressive_strength` between regions. So, more similar colors within a region, distinct colors across regions.

Regression Trees

We will follow some strict rules for segmenting the whole space:

- You can only use **straight lines**
- Your lines must either be **vertical or horizontal**.
- Every line **stops** when it hits an existing line.

Regression Trees

```
decision_tree = DecisionTreeRegressor(max_depth=2)
decision_tree.fit(X_train,y_train)

Decision Tree RMSE: 5.32229449641
```

How?

Recursive binary splitting:

Begin at the top of the tree.

- For **every feature**, examine **every possible cutpoint**, and choose the feature and cutpoint such that the resulting tree has the lowest possible mean squared error (MSE). Make that split.

- Examine the two resulting regions, and again make a **single split** (in one of the regions) to minimize the MSE.

Keep repeating the previous step until a **stopping criterion** is met:

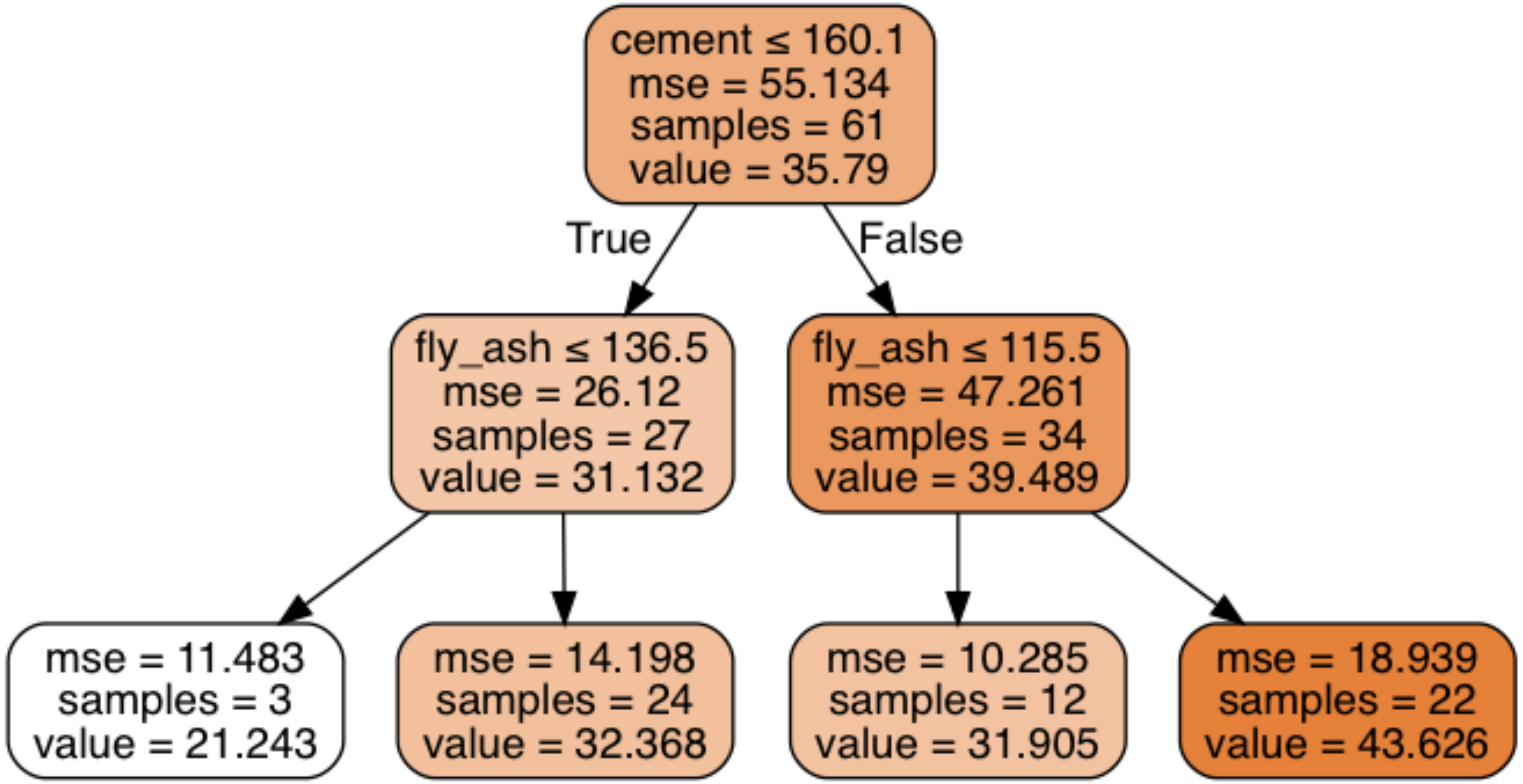
- maximum tree depth is reached (maximum number of splits required to arrive at a leaf, in our case it was 2)

- minimum number of observations in a leaf (default is 2)

Regression Trees

```
decision_tree = DecisionTreeRegressor(max_depth=2)
decision_tree.fit(X_train,y_train)

Decision Tree RMSE: 5.32229449641
```



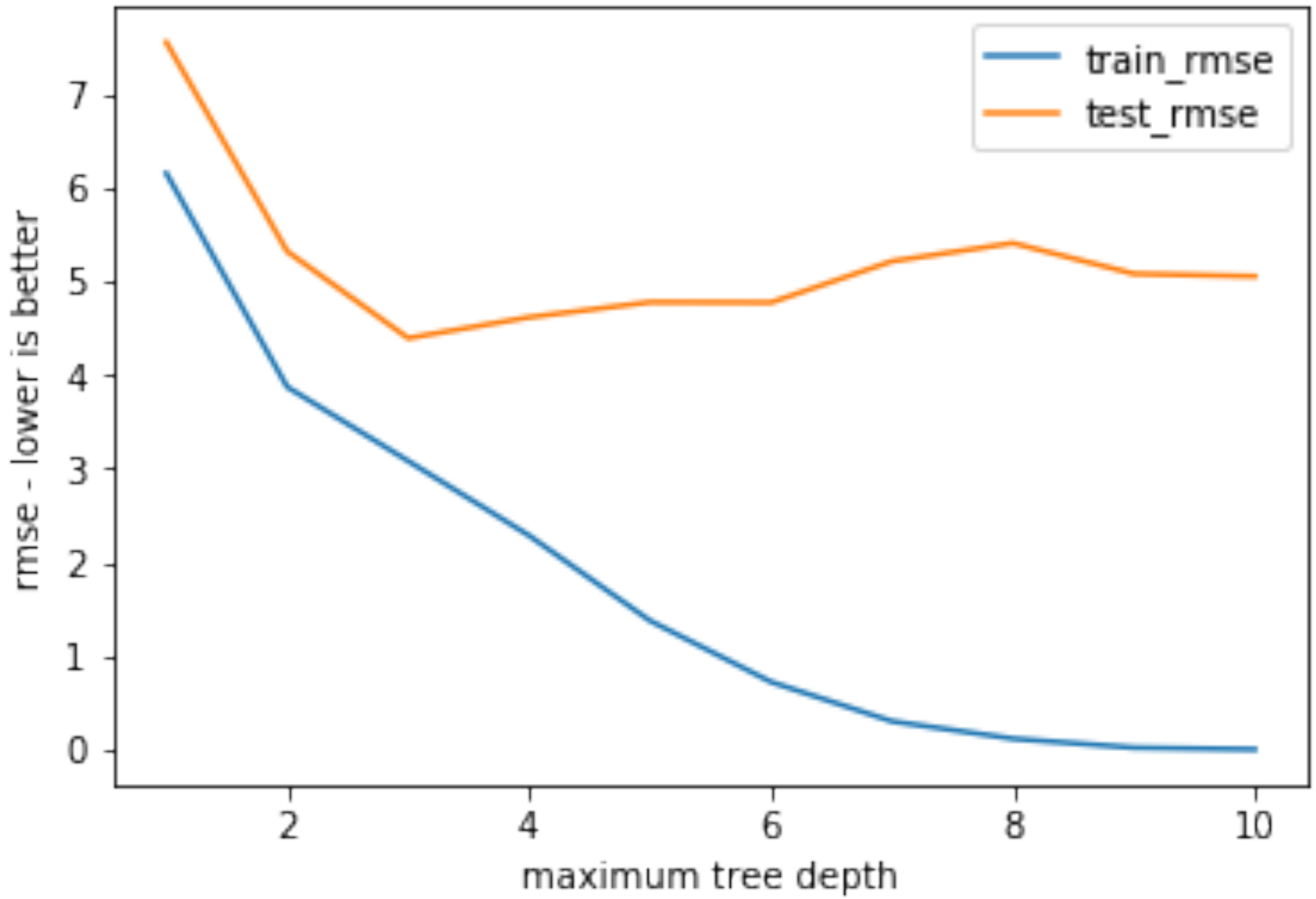
Exercise

- Build a decision tree model to predict slump given the input features.

- What is the test set RMSE?

METIS

Too Deep?



Feature Importance

```
pd.DataFrame({'feature':feature_names_cem,
'importance':best_single_tree.feature_importances_})
```

	feature	importance
0	cement	0.377599
1	slag	0.018253
2	fly_ash	0.501815
3	water	0.090038
4	sp	0.012295
5	coarse_aggr	0.000000
6	fine_aggr	0.000000

METIS

Exercise

Examine the feature importances of your slump model.

Are they the same as those from the last slide? Is the order of the features in terms of their importances the same?

METIS

Making Predictions

```
y_pred = best_single_tree.predict(X_test)
np.sqrt(mean_squared_error(y_test, y_pred))

>> 4.395463024956621
```

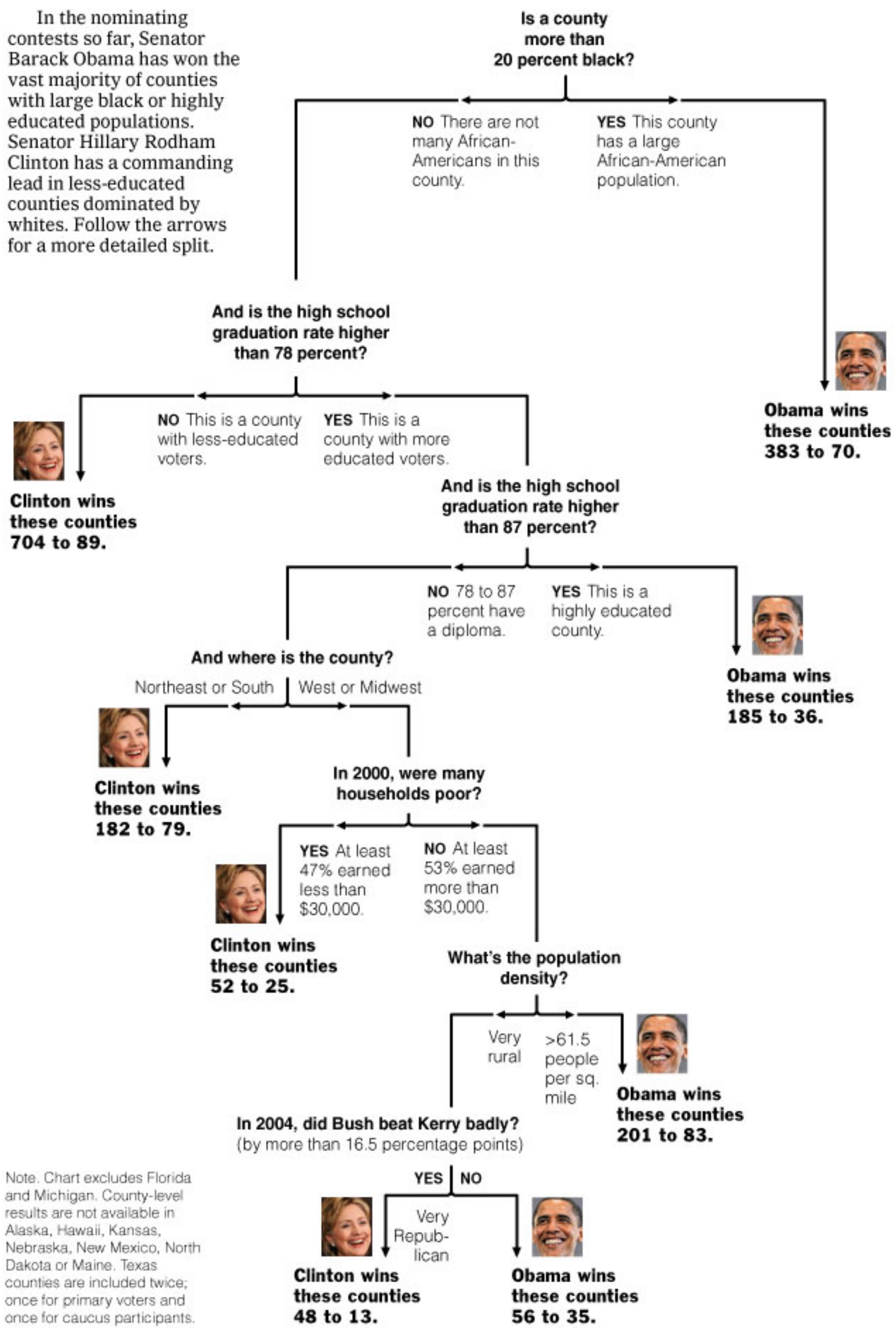
METIS

Classification Trees

METIS

Decision Tree: The Obama-Clinton Divide

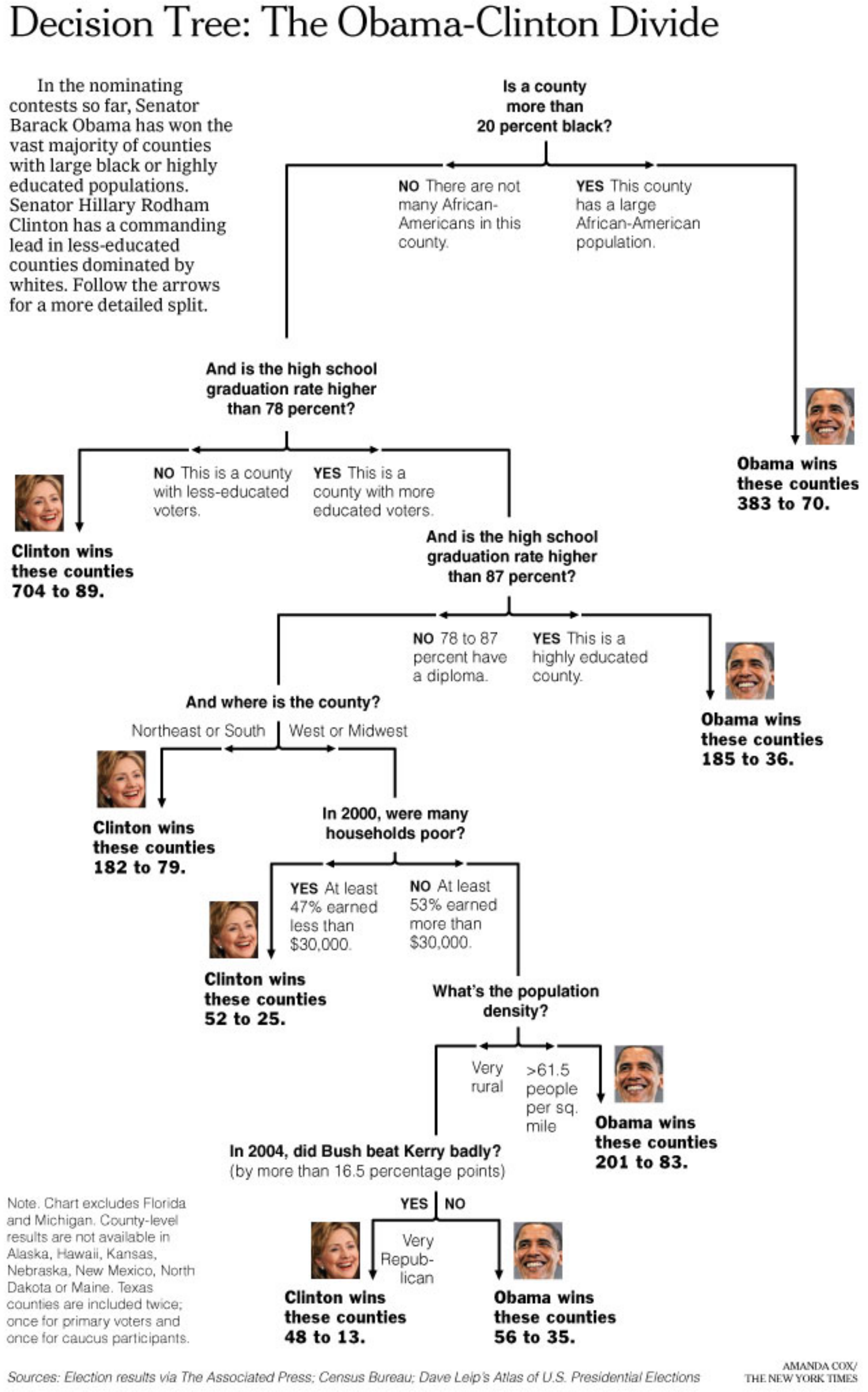
In the nominating contests so far, Senator Barack Obama has won the vast majority of counties with large black or highly educated populations. Senator Hillary Rodham Clinton has a commanding lead in less-educated counties dominated by whites. Follow the arrows for a more detailed split.



Note. Chart excludes Florida and Michigan. County-level results are not available in Alaska, Hawaii, Kansas, Nebraska, New Mexico, North Dakota or Maine. Texas counties are included twice; once for primary voters and once for caucus participants.

Sources: Election results via The Associated Press; Census Bureau; Dave Leip's Atlas of U.S. Presidential Elections

AMANDA COX/ THE NEW YORK TIMES



Exercise

Please answer the following questions about the Obama diagram:

- What are the observations? How many observations are there?
- What is the response variable?
- What are the features?
- What is the most predictive feature?
- Why does the tree split on high school graduation rate twice?
- What is the class prediction for the following counties:
  - 10% African-American, 50% high school graduation rate, located in the South, high poverty, high population density?
  - 18% African-American, 95% high school graduation rate, located in the South, high poverty, high population density?
  - What are the predicted probabilities for both of those counties?

METIS

regression trees	classification trees
predict a continuous response	predict a categorical response
predict using mean response of each leaf	predict using most commonly occuring class of each leaf
splits are chosen to minimize MSE	splits are chosen to minimize Gini index (discussed below)

METIS

Classification Trees

Common options for the splitting criteria when generating classification trees:

- **Classification error rate:** fraction of training observations in a region that don't belong to the most common class

- **Gini impurity:** measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset

METIS Classification Error

Pretend we are predicting whether someone buys an free-standing house or a condo:

- At a particular node, there are 30 observations (home buyers), of whom 10 bought free-standing homes and 20 bought condos.

- Since the majority class is **condos**, that's our prediction for all 25 observations, and thus the classification error rate is **10/30 = 33%**.

Our goal in making splits is to reduce the classification error rate.

METIS Classification Error

Let's try splitting on income:

- **Greater than 100k/year:** 8 free-standing and 3 condos, thus the predicted class is free-standing
- **Less than 100k/year:** 2 free-standing and 17 condos, thus the predicted class is condo
- Classification error rate after this split would be **5/30 = ~17%**

Compare that with a split on purchaser-type:

- **Married:** 4 free-standing and 6 condos, thus the predicted class is condo
- **Unmarried:** 6 free-standing and 14 condos, thus the predicted class is condo
- Classification error rate after this split would be **10/30 = ~33%** (it didnt change!)

The decision tree algorithm will try every possible split across all features, and choose the split that reduces the error rate the most.

METIS

Gini Impurity

$$1 - \left(\frac{freestanding}{Total}\right)^2 - \left(\frac{condo}{Total}\right)^2 = 1 - \left(\frac{10}{30}\right)^2 - \left(\frac{20}{30}\right)^2 = 0.44$$

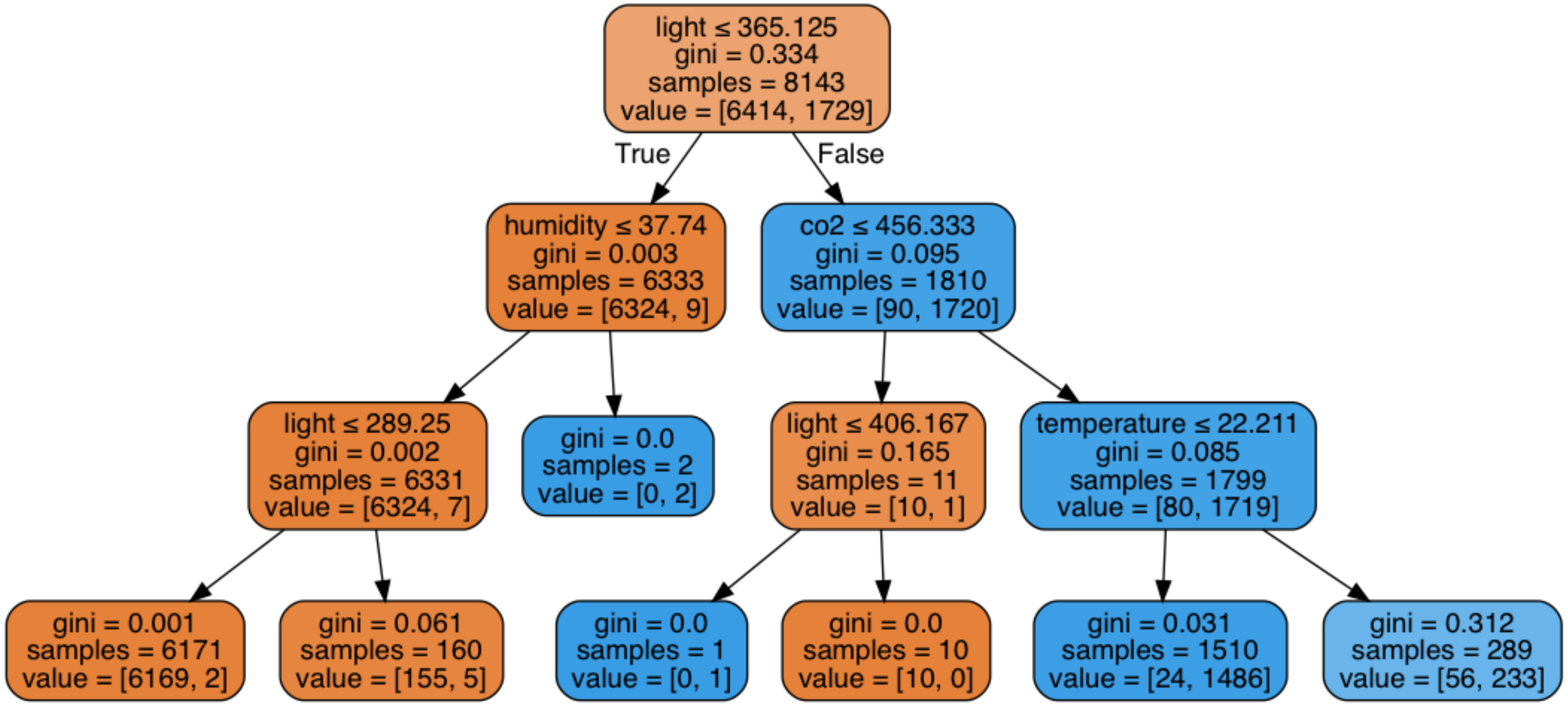
METIS

Decision Trees

```
from sklearn.tree import DecisionTreeClassifier
occupancy_tree = DecisionTreeClassifier(max_depth=3)
occupancy_tree.fit(X_occ_train, y_occ_train)
```

Decision Trees

```
from sklearn.tree import DecisionTreeClassifier
occupancy_tree = DecisionTreeClassifier(max_depth=3)
occupancy_tree.fit(X_occ_train, y_occ_train)
```



Exercise

- Evaluate the model using `accuracy_score` on the testing data.
- Is the accuracy score above chance? What is chance accuracy here?

Advantages of Decision Trees

- Can be used for regression or classification
- Can be displayed graphically
- Highly interpretable
- Can be specified as a series of rules, and more closely approximate human decision-making than other models
- Prediction is fast
- Features don't need scaling
- Automatically learns feature interactions (they are non-linear models)
- Tend to ignore irrelevant features (especially when there are lots of features)
- Because decision trees are non-linear models they will outperform linear models if the relationship between features and response is highly non-linear

Disadvantages of Decision Trees

- Performance is (generally) not competitive with the best supervised learning methods
- Can easily overfit the training data (tuning is required)
- Small variations in the data can result in a completely different tree (they are high variance models)
- Recursive binary splitting makes "locally optimal" decisions that may not result in a globally optimal tree
- Don't tend to work well if the classes are highly unbalanced
- Don't tend to work well with very small datasets

METIS

Ensembles

Ensemble learning is the process of combining several predictive models in order to produce a combined model that is more accurate than any individual model.

- **Regression:** take the average of the predictions
- **Classification:** take a vote and use the most common prediction, or take the average of the predicted probabilities

METIS

Ensembles

For ensembling to work well, the models must have the following characteristics:

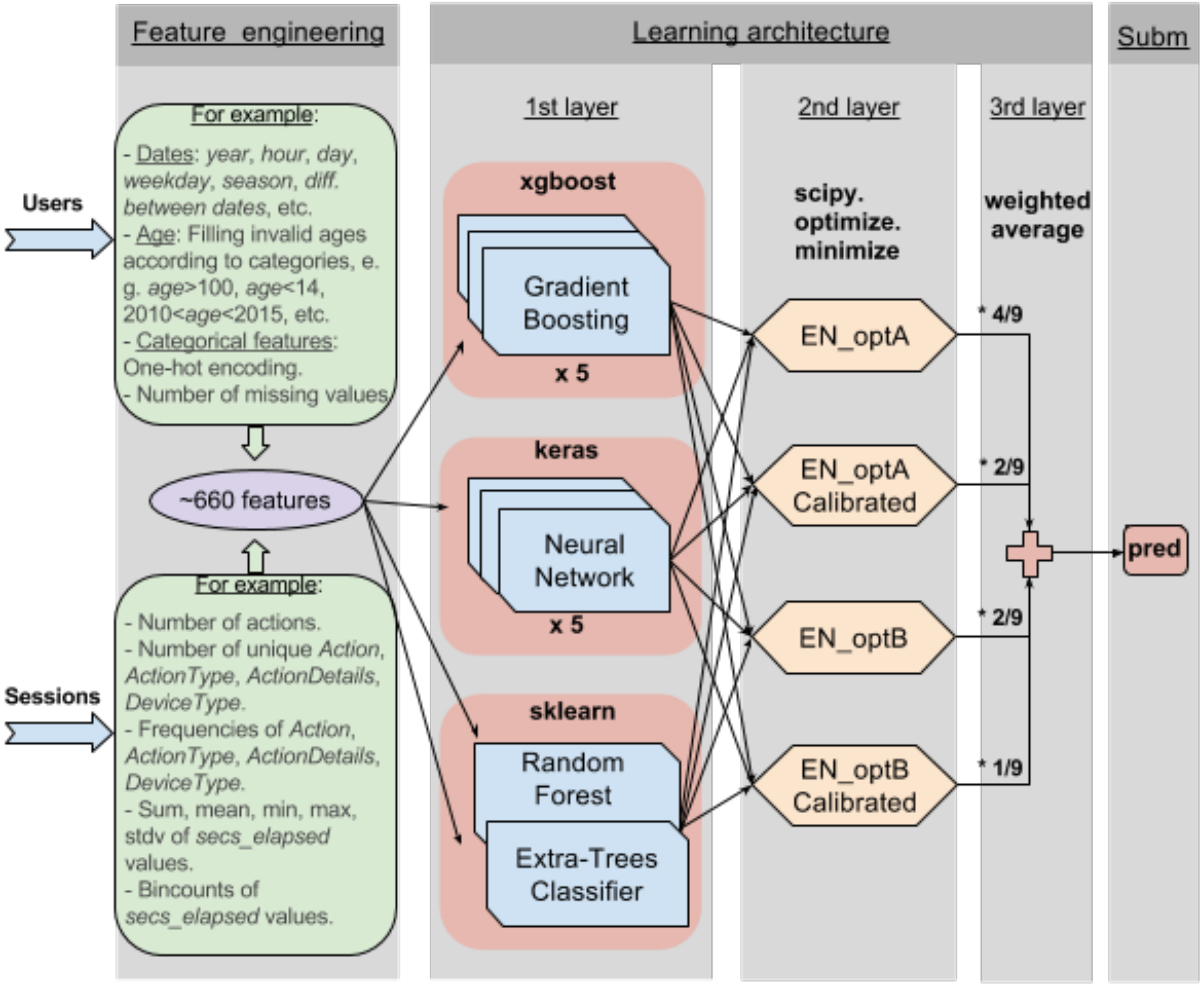
- **Accurate:** they outperform random guessing
- **Independent:** their predictions are generated using different processes

The big idea: If you have a collection of individually imperfect (and independent) models, the "one-off" mistakes made by each model are probably not going to be made by the rest of the models, and thus the mistakes will be discarded when averaging the models.

Note: As you add more models to the voting process, the probability of error decreases, which is known as Condorcet's Jury Theorem.

METIS

Ensembles



Bagging

1. Grow *n* trees using *n* bootstrap samples from the training data.

2. Train each tree on its bootstrap sample and make predictions.

3. Combine the predictions:

    - Average the predictions for **regression trees**

    - Take a majority vote for **classification trees**

Bagging

```
bagreg = BaggingRegressor(DecisionTreeRegressor(),
n_estimators=500, bootstrap=True, oob_score=True)
bagreg.fit(X_train, y_train)
y_pred_bag = bagreg.predict(X_test)

>>> Bagged RMSE with 500 trees: 3.78785968654
```

Bagging

Here's how to calculate **"out-of-bag error":**

1. For every observation in the training data, predict its response value using **only** the trees in which that observation was out-of-bag. Average those predictions (for regression) or take a majority vote (for classification).

2. Compare all predictions to the actual response values in order to compute the out-of-bag error.

When *b* is sufficiently large, the **out-of-bag error** is an accurate estimate of **out-of-sample error**.

```
bagreg.oob_score_
```

METIS

Random Forest

Random Forests are a **slight variation of bagged trees** that have even better performance:

- Just like bagging, we create an ensemble of decision trees using bootstrapped samples of the training set.
- However, when building each tree, each time a split is considered, a **random sample of *m* features** is chosen as split candidates from the **full set of *p* features**. The split is only allowed to use **one of those *m* features**.
    - A new random sample of features is chosen for **every single tree at every single split**.
    - For **classification**, *m* is typically chosen to be the square root of *p* (the total number of features).
    - For **regression**, *m* is typically chosen to be somewhere between *p*/3 and *p*.

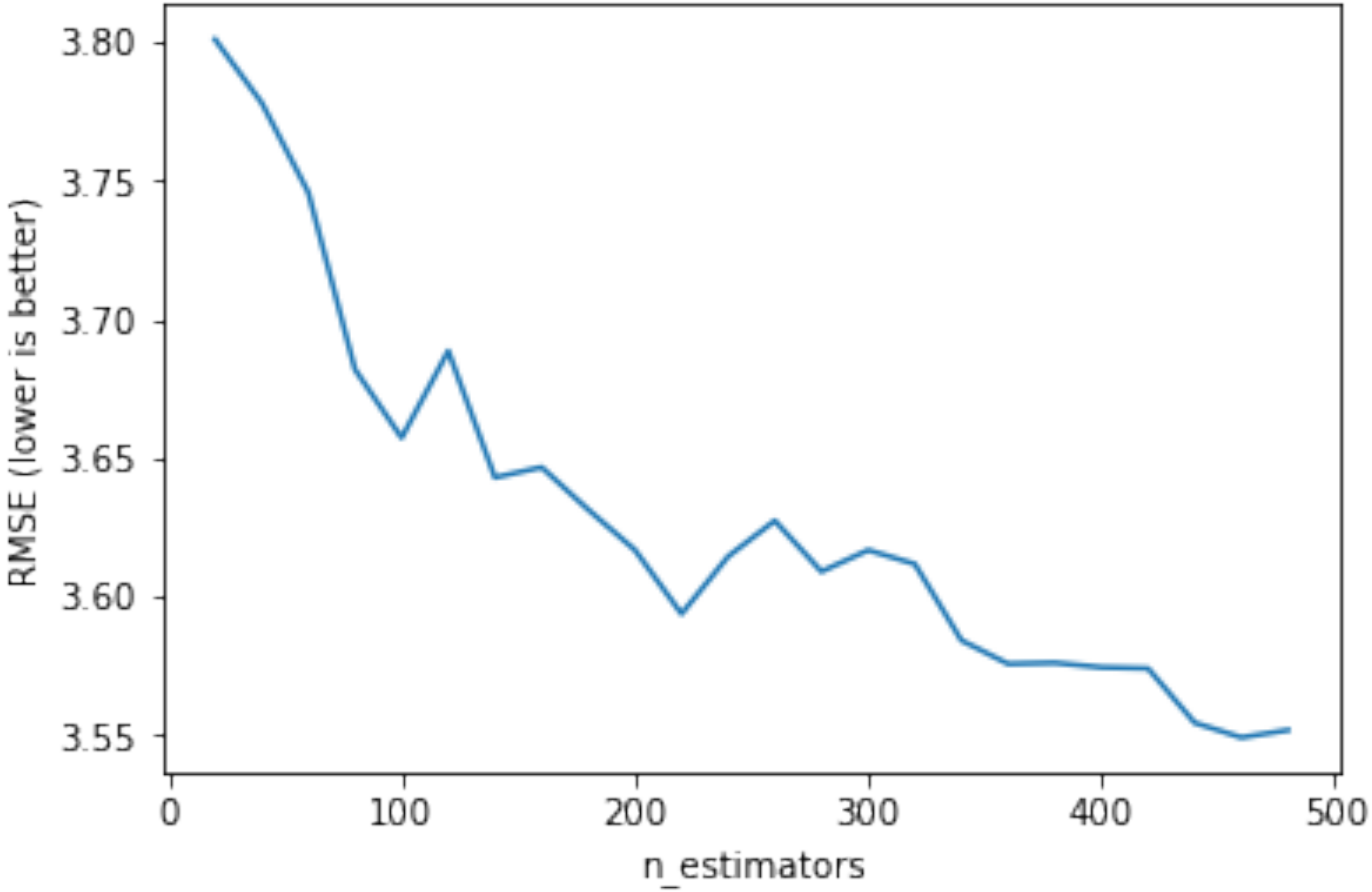
METIS

Tuning a Random Forest

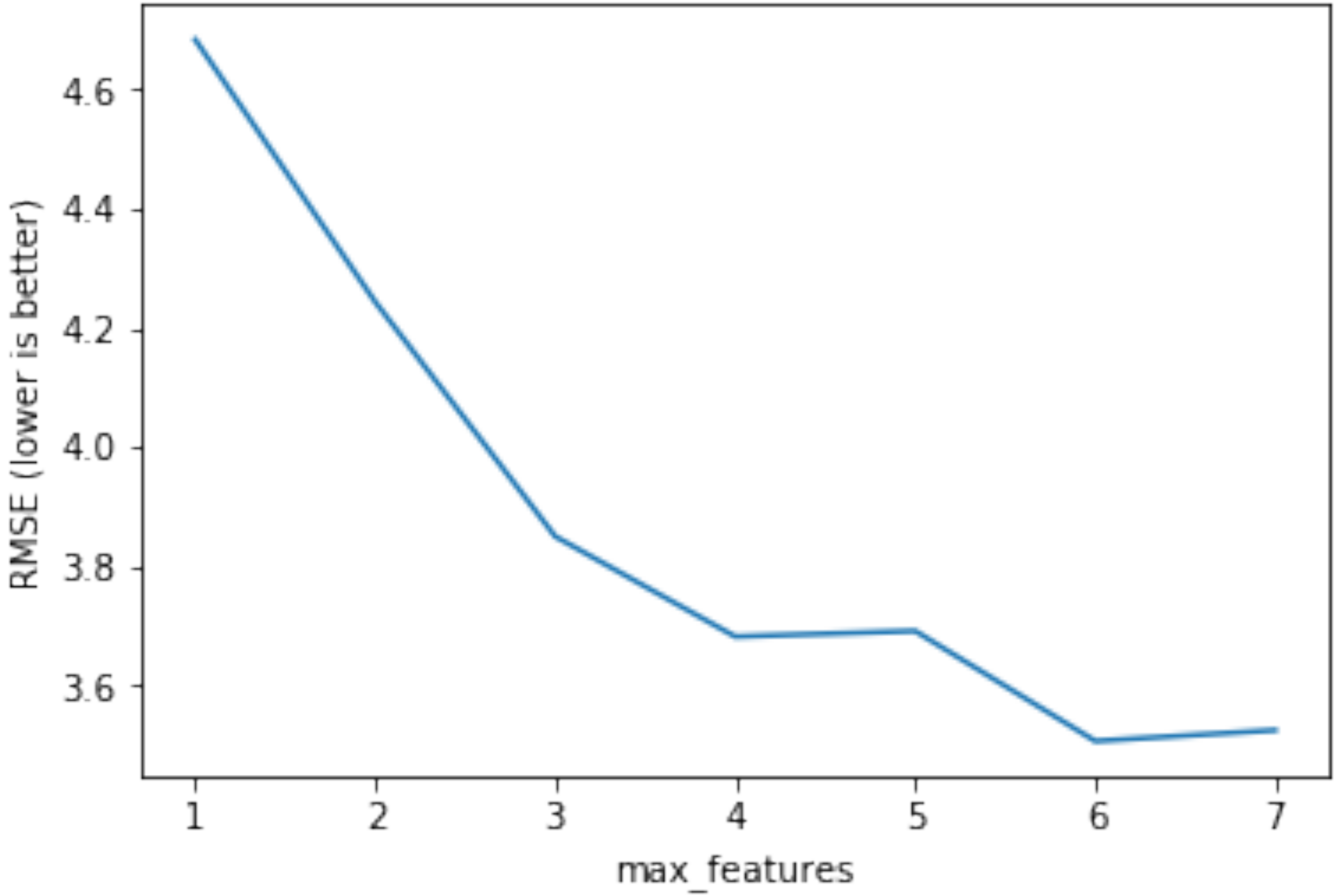
2 important parameters that should be tuned when creating a random forest model are:

- The number of trees to grow (called **n_estimators** in scikit-learn)
- The number of features that should be considered at each split (called **max_features** in scikit-learn)

Tuning a Random Forest



Tuning a Random Forest



Exercise

Build a Random Forest Regression Model to predict Flow using the concrete slump test dataset.

- What is the test-set RMSE?
- What are the feature importances? Are they in a different ranked order than those for the compressive_strength model?

Build a Random Forest Classification Model on the occupancy data, using the `datatraining.txt` data

- What is the test-set Accuracy when testing on `datatest.txt`? When testing on `datatest2.txt`?
- What are the feature importances?

METIS

Final Thoughts

Advantages of Random Forests:

- Performance is competitive with the best supervised learning methods
- Provides a more reliable estimate of feature importance
- Allows you to estimate out-of-sample error without using train/test split or cross-validation

Disadvantages of Random Forests:

- Less interpretable
- Slower to train
- Slower to predict

METIS

Final Thoughts

Advantages of Ensemble Learning:

- Increases predictive accuracy
- Easy to get started (especially with Random Forests)

Disadvantages of Ensemble Learning:

- Decreases interpretability
- Takes longer to train/predict
- More complex to automate and maintain
- Sometimes marginal gains in accuracy may not be worth the added complexity

Group Exercise

- **Situation:** You are a part of an Air Force team researching how to improve survivability of combat aircraft. At your disposal are mappings of damage on several thousand aircraft which have returned from sorties over enemy territory.

- **Your task:** The Air Force is seeking to optimize the amount of armor on the aircraft. Armor is heavy, and therefore, too much armor reduces the usefulness of the aircraft. The Air Force is seeking to determine the ideal location and amount of armor to place on the aircraft to maximize crew survivability and payload.