Event Coding with Machine Learning

Javier Osorio

School of Government and Public Policy
University of Arizona

Prepared for the
Winter Institute in Computational Social Science (WICSS)
University of Arizona

Virtual event, January 7th, 2021

Route

PLAN AHEAD

The Plan Ahead

The Plan for this session:

- Substantive tasks:
  - Terrorist attacks
  - Type of terrorist activity
- R package
- Cautionary note for NLP
- Machine Learning models:
  - Support Vector Machine
  - Boosting
  - Bagging
  - Random Forest
- Plot ML model performance
- Classify types of terrorism events
  - Only with SVM

Terrorist Attacks

South Asia Terrorism Portal (SATP)

- Short narratives of terrorist activity
- `https://satp.org/`

Date	Incidents
Dec - 1	On December 1, a Liberation Tribal Tiger (LTT) militant identified as Khuptinthang Khongsai was arrested from Pangei bazar in Imphal East District of Manipur, reports E Pao. One 9 mm pistol along with a magazine loaded with 7 live rounds was recovered the arrested militant. Elsewhere in the District, a gang member involved in extortion activities was arrested from Yaingangpokpi bazar along with a .32 pistol along with a magazine loaded with two live rounds. The extortionist was identified as Phungreiso Masangva.
Dec - 1	A massive search operation was launched in the Ajnala sector near the Kot Rajada border outpost (BoP) in Amritsar District of Punjab on December 1 after an unmanned aerial vehicle (drone) was reportedly sighted in the area, reports The Tribune. The alert troops of the 73 Battalion of the Border Security Force (BSF) patrolling near the International Border (IB) fired indiscriminately in the air after they heard the sound of movement of drones at 1 am. However, it managed to return taking the advantage of foggy conditions.
Dec - 1	A special National Investigation Agency (NIA) court of Mumbai (Maharashtra), conducting trial in the 2008 Malegaon blast case on December 1, directed all the seven accused, including BJP Member of Parliament (MP), Pragya Singh Thakur and Lieutenant Colonel Prasad Purohit, to appear before it on December 3, 2020, reports NDTV. The court's direction came on a plea of one of the victim's family, seeking day-to-day trial in the case, which is being probed by the NIA. Judge PR Sitre asked all the accused to remain present in the court on December 2. The court had framed terror charges against Purohit, Pragya Singh Thakur and five other accused in October 2018.

Terrorist Attacks

Two tasks:

- Identify relevant stories about terrorist attacks
- Identify the type of terrorist action

Data:

- Pre-annotated data using Tagtog
- Manually coded 2,251 stories
- Terrorist attacks:
  - Balanced data
  - File `"satp_terrorism.csv"`
- Type of terrorist action:
  - Focus on the action of the event (not source or target)
  - Actions: Kidnapping, Bombing, Armed Assault, and Other.
  - File `"satp_actions.csv"`

The RTextTools package

R package: `RTextTools`

- Machine Learning models for NLP
- Integrates several other packages into a single environment
- Nine algorithms:
  - Support Vector Machine, SVM (from `"e1071"`)
  - Boosting (from `"caTools"`)
  - Bagging (from `"ipred"`)
  - Random Forest, RF (from `"randomForest"`)
  - Neural Networks, NN (from `"nnet"`)
  - Generalized Linear Models (from `"glmnet"`)
  - Maximum Entropy (from `"maxent"`)
  - Scaled Linear Discriminant Analysis, LSDA (from `"ipred"`)
  - Classification or Regression trees (from `"tree"`)
- One-stop shop with few steps (default options)
- To tune specific parameters, use source packages

The RTextTools package

R package: `RTextTools`

- Machine Learning models for NLP
- Integrates several other packages into a single environment
- Nine algorithms:
  - Support Vector Machine, SVM (from `"e1071"`)
  - Boosting (from `"caTools"`)
  - Bagging (from `"ipred"`)
  - Random Forest, RF (from `"randomForest"`)
  - Neural Networks, NN (from `"nnet"`)
  - Generalized Linear Models (from `"glmnet"`)
  - Maximum Entropy (from `"maxent"`)
  - Scaled Linear Discriminant Analysis, LSDA (from `"ipred"`)
  - Classification or Regression trees (from `"tree"`)
- One-stop shop with few steps (default options)
- To tune specific parameters, use source packages

The RTextTools package

Machine Learning Workflow:

1. Split the data
   - Training data
   - Testing data

The RTextTools package

Machine Learning Workflow:

1. Split the data
   - Training data
   - Testing data
2. Create Document Term Matrix
   - Turn text as data

The RTextTools package

Machine Learning Workflow:

1. Split the data
   - Training data
   - Testing data
2. Create Document Term Matrix
   - Turn text as data
3. Configure containers
   - Training container
   - Testing container

The RTextTools package

Machine Learning Workflow:

1. Split the data
   - Training data
   - Testing data
2. Create Document Term Matrix
   - Turn text as data
3. Configure containers
   - Training container
   - Testing container
4. Train ML model

The RTextTools package

Machine Learning Workflow:

1. Split the data
   - Training data
   - Testing data
2. Create Document Term Matrix
   - Turn text as data
3. Configure containers
   - Training container
   - Testing container
4. Train ML model
5. Test ML model

The RTextTools package

Machine Learning Workflow:

1. Split the data
   - Training data
   - Testing data
2. Create Document Term Matrix
   - Turn text as data
3. Configure containers
   - Training container
   - Testing container
4. Train ML model
5. Test ML model
6. Assess model accuracy

Text as Data

Consider the following sentences:

```
"I like event coding"
"Event coding is fun!"
"I like text analysis"
"Text analysis is also fun"
"I like text, I like fun!"
```

The Document Term Matrix (DTM) approach:

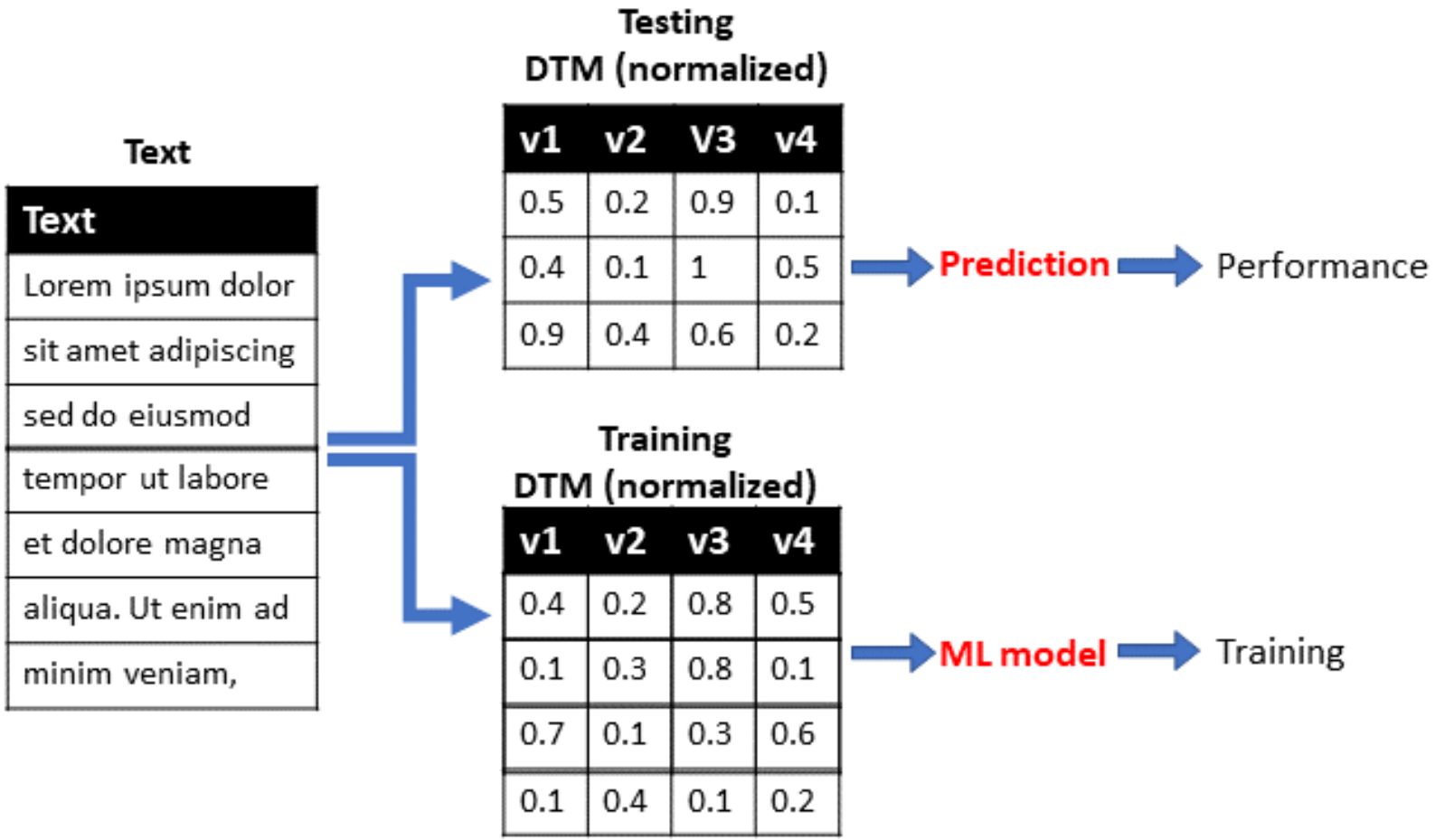
	I	like	event	coding	text	is	analysis	also	fun
I like event coding	1	1	1	1	0	0	0	0	0
Event coding is fun	0	0	1	1	0	1	0	0	1
I like text analysis	1	1	0	0	1	0	1	0	0
Text analysis is also fun	0	0	0	0	1	1	1	1	1
I like text I like fun	2	2	0	0	1	0	0	0	1

DTM also normalizes the data

OVERALL STRATEGY OF MACHINE LEARNING

Machine Learning with Data

Machine Learning with Text as Data



Let's go to R

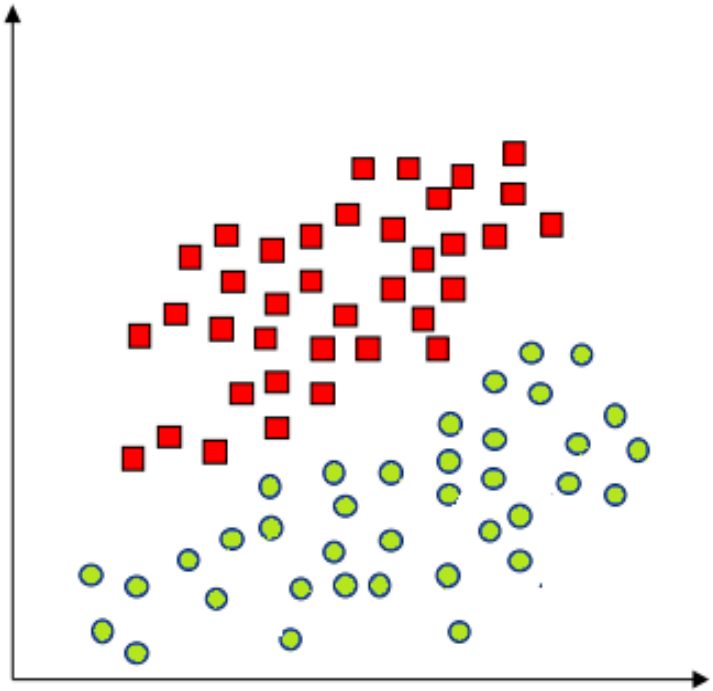
SUPPORT VECTOR MACHINE

Support Vector Machine

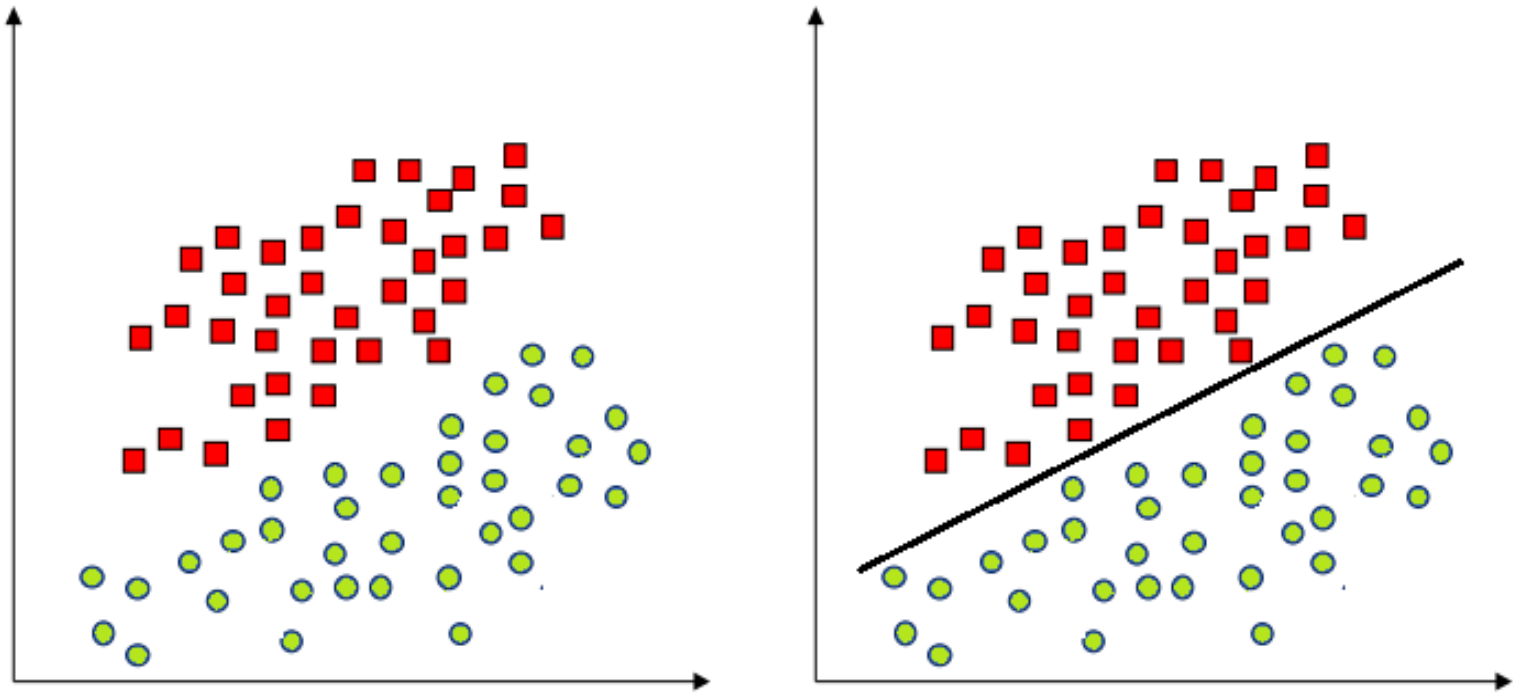
Support Vector Machine, SVM

- SVM finds a way to divide the data into two classes
- The power behind SVM rests on its capacity to create hyper-planes to divide the data
- It is capable of doing so in multiple dimensions

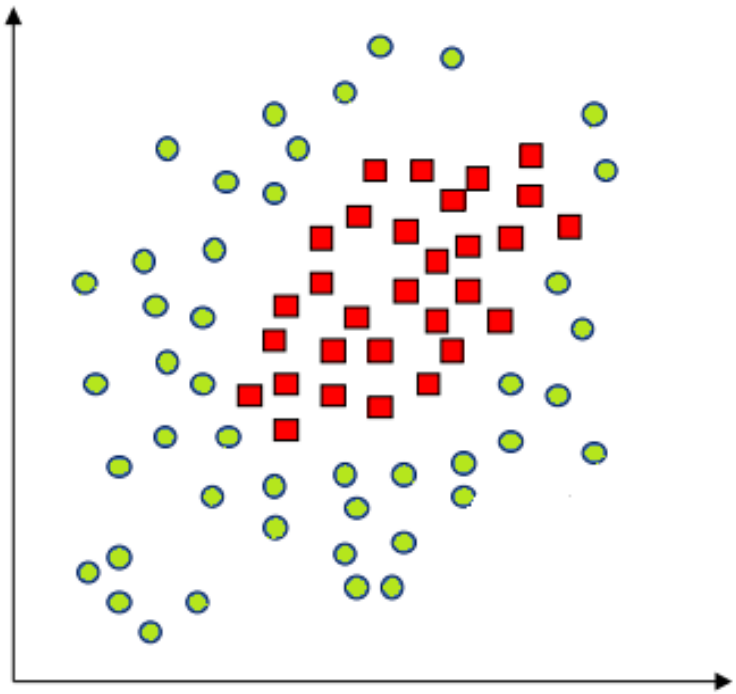
Support Vector Machine



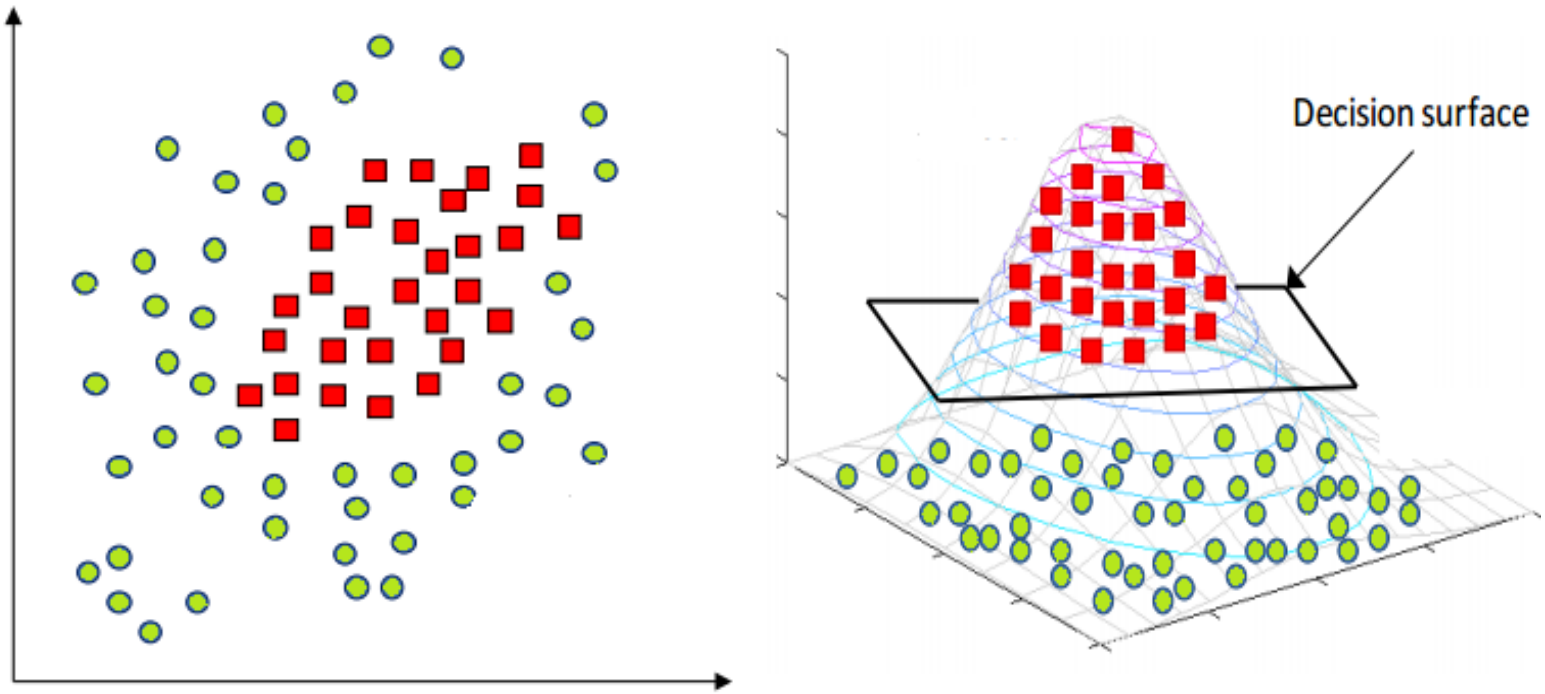
Support Vector Machine



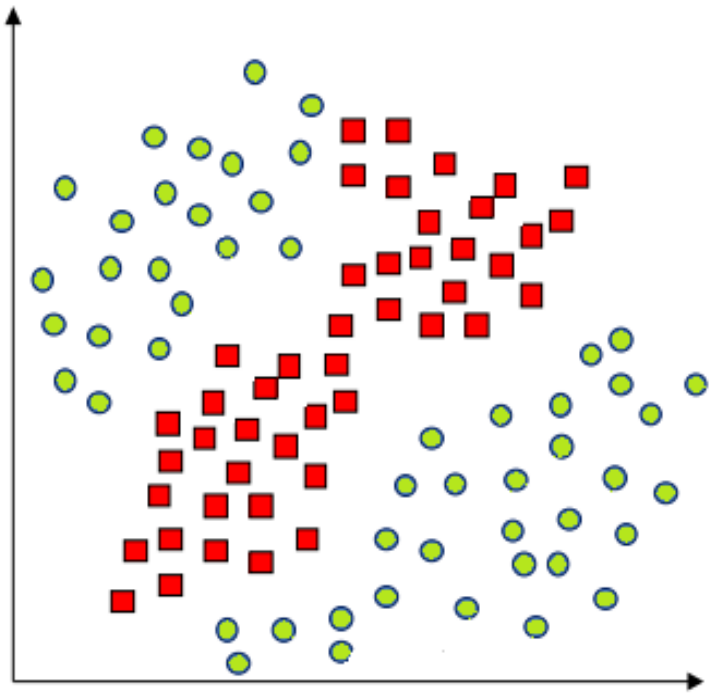
Support Vector Machine



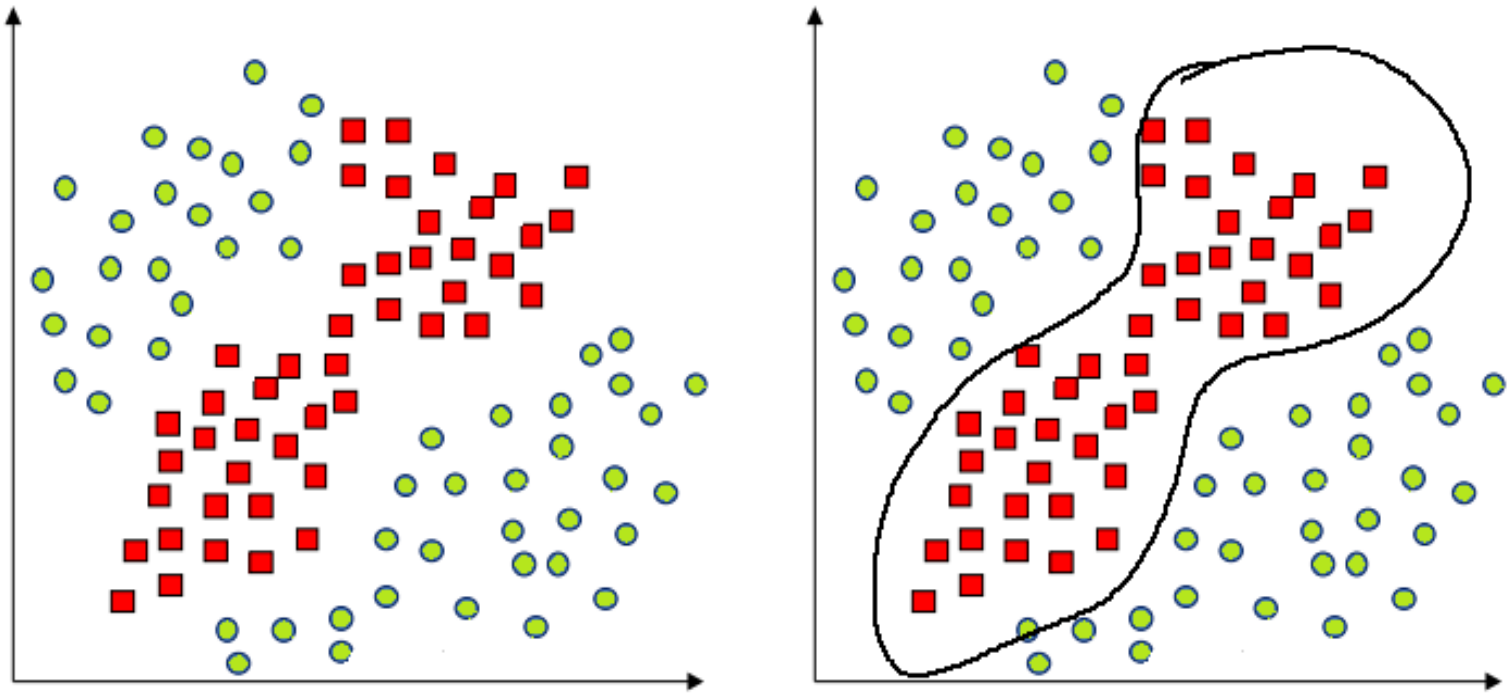
Support Vector Machine



Support Vector Machine



Support Vector Machine



Let's go to R

BOOSTING

Boosting

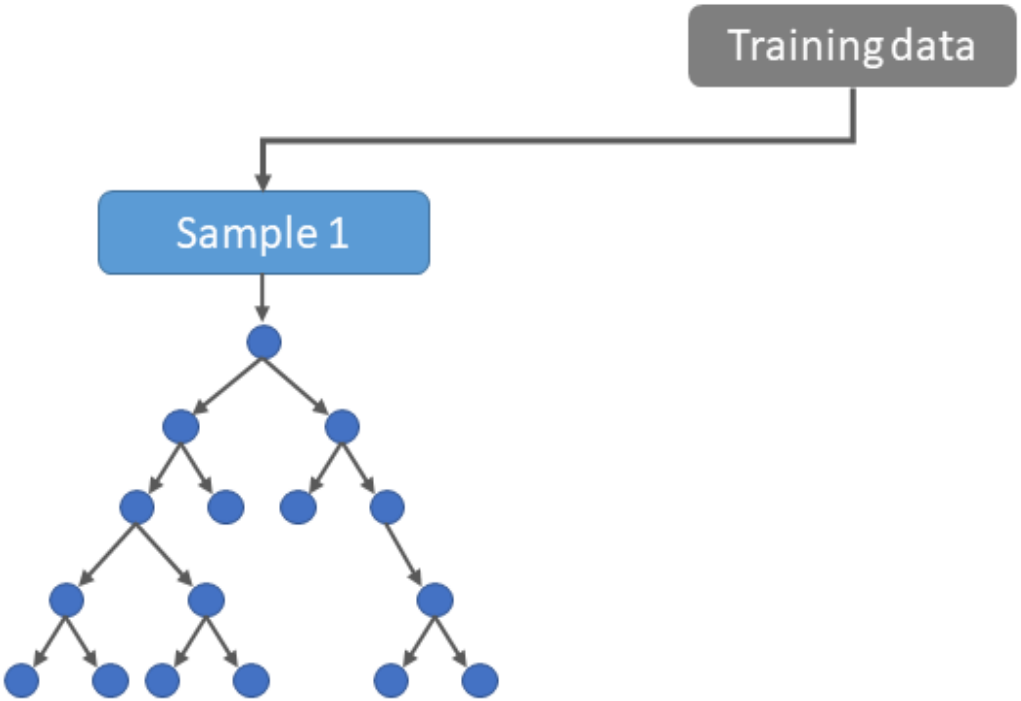
Boosting

- Learns from its own mistakes
  - Conducts an initial classification round
  - Oversamples from misclassified cases
  - Conducts another classification round
  - Oversamples from misclassified cases
  - Conducts another classification round and so on
- Ensemble technique
- Uses mediocre models to generate a better model

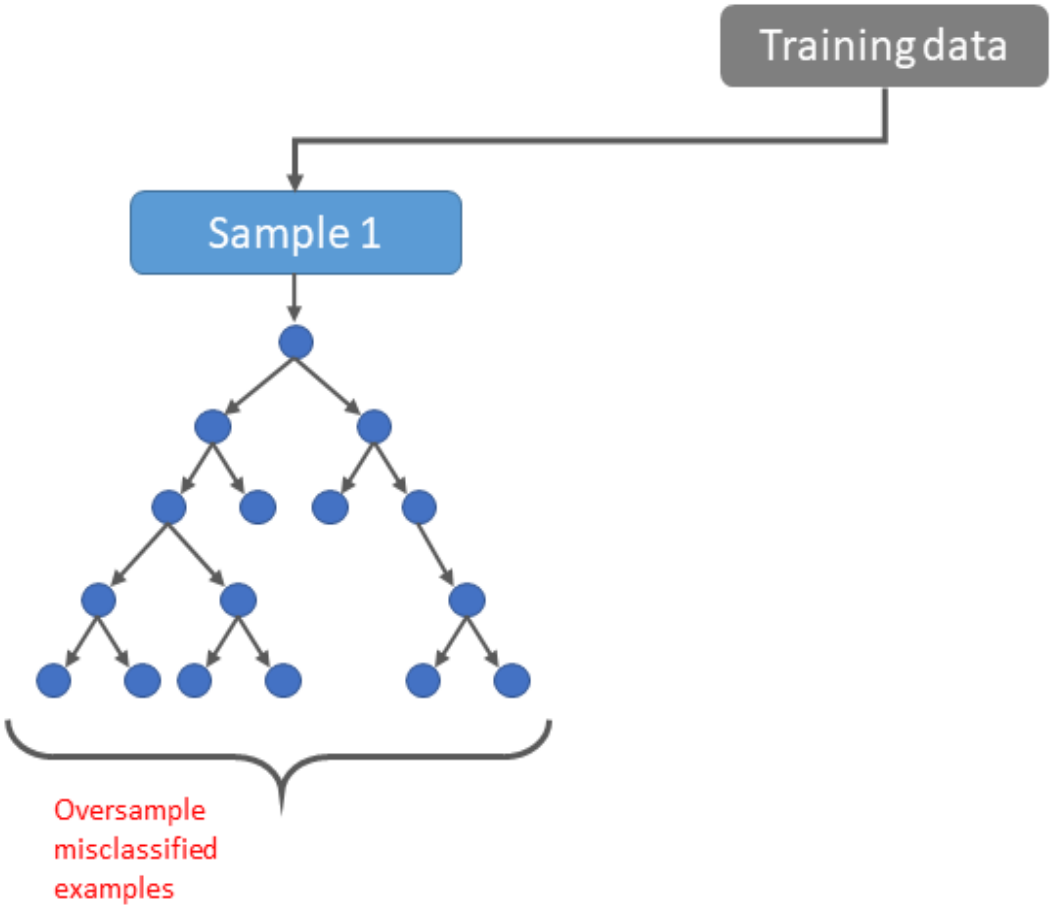
Boosting

Boosting

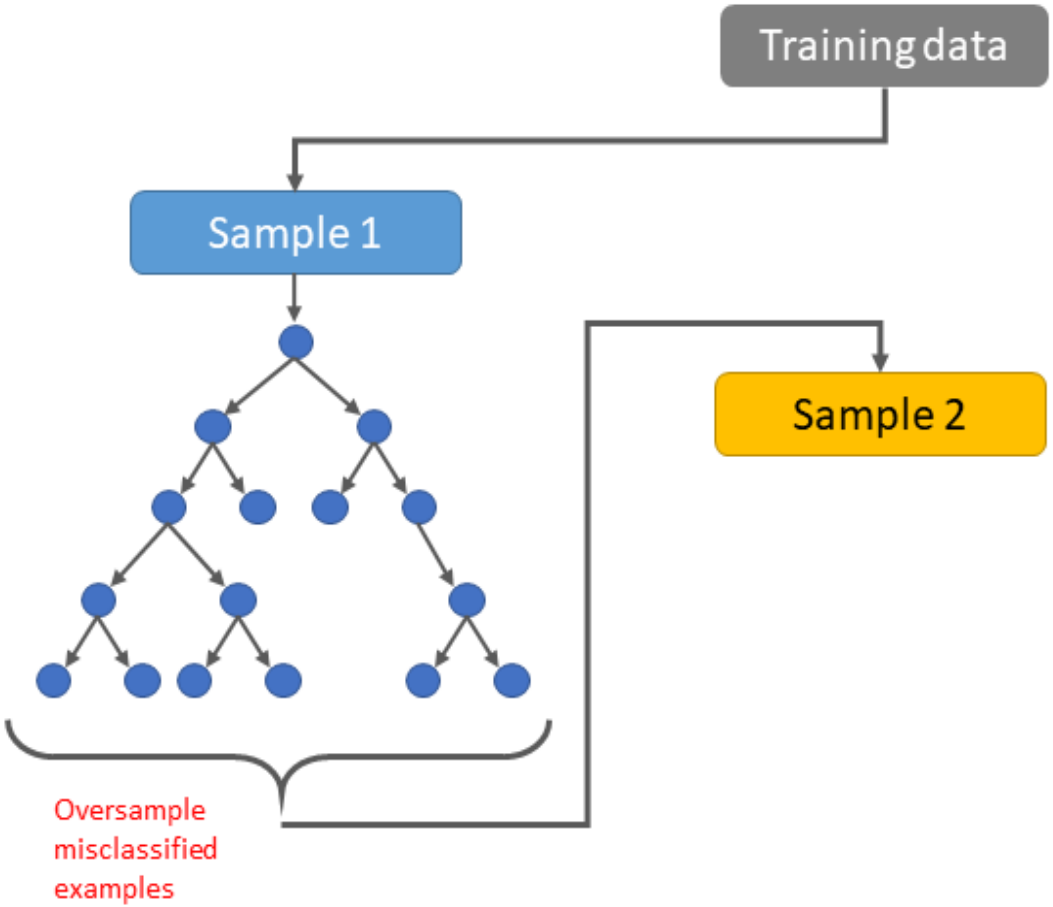
Boosting



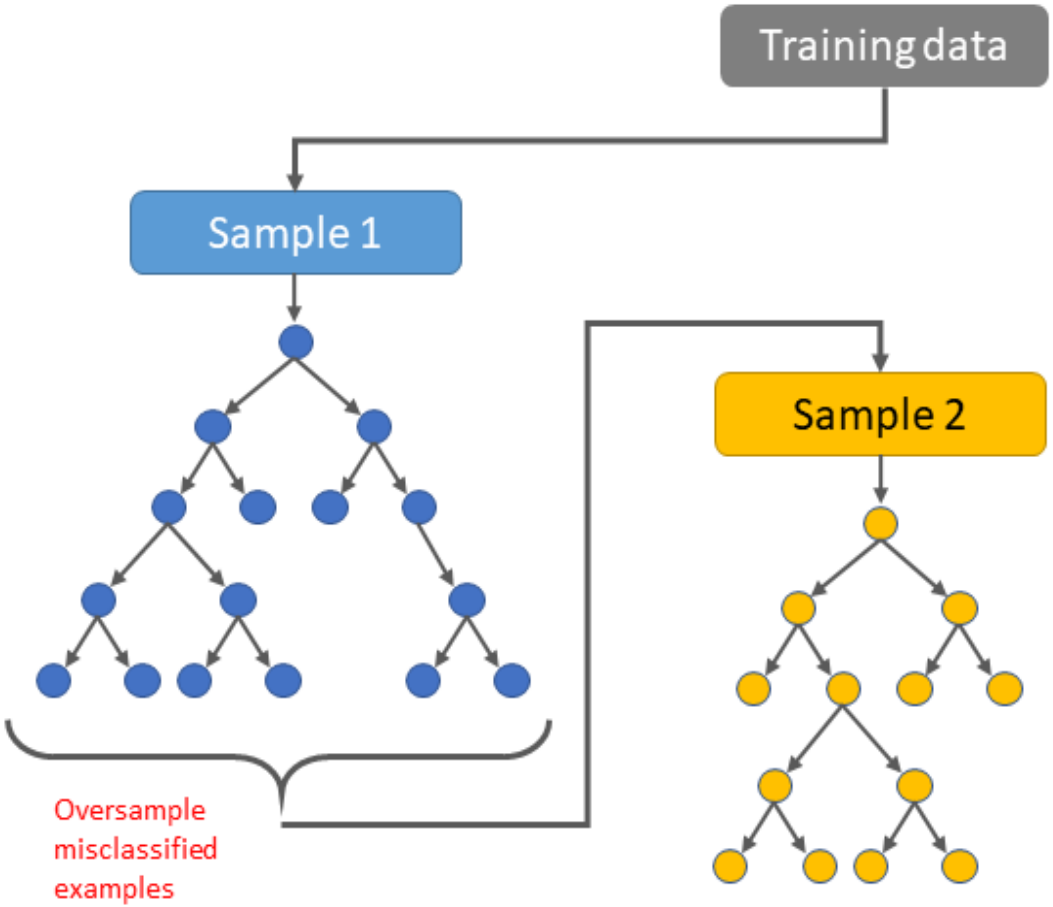
Boosting


Training data
Sample 1
Oversample
misclassified
examples

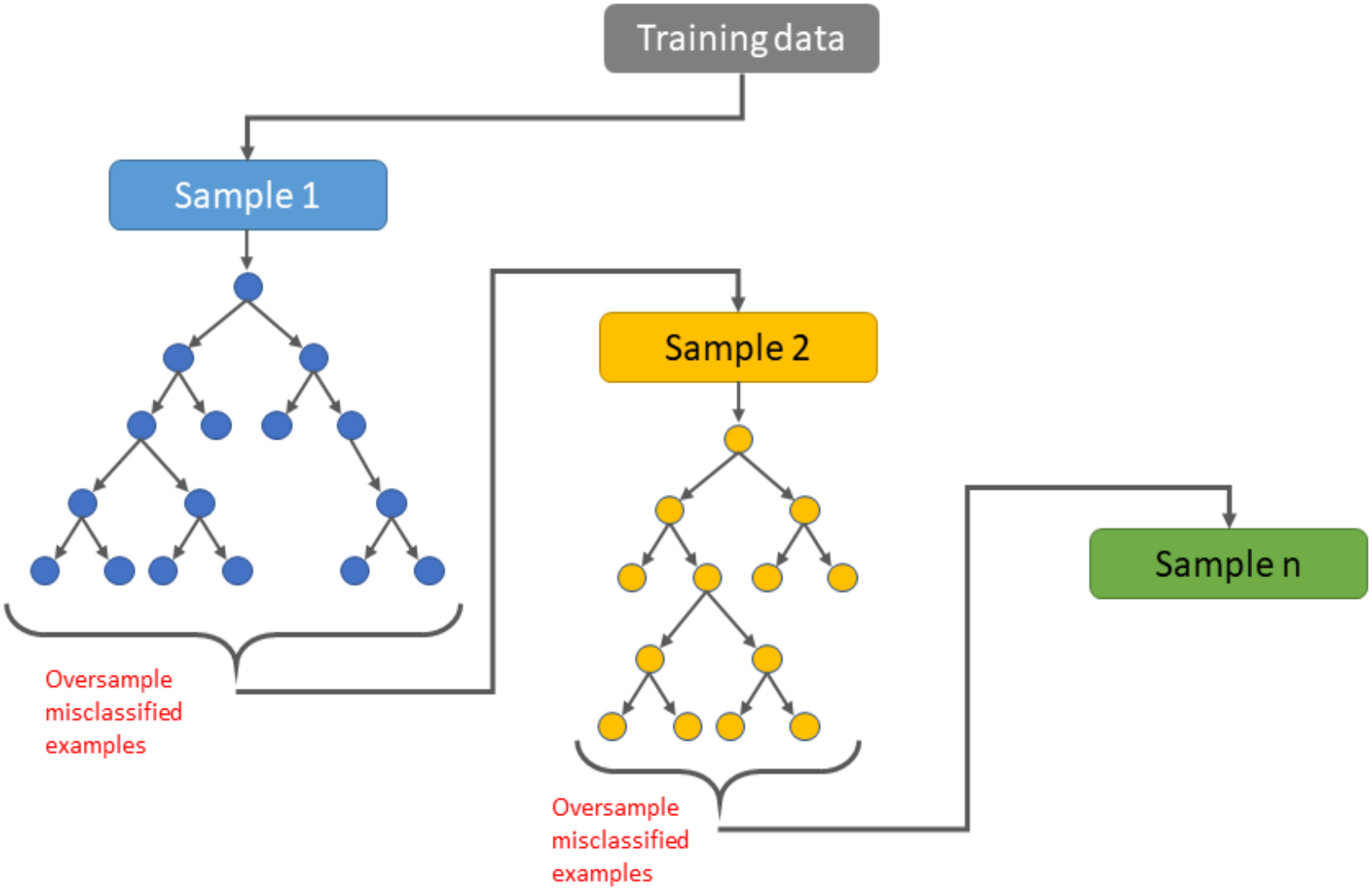
Boosting


Training data
Sample 1
Sample 2
Oversample
misclassified
examples

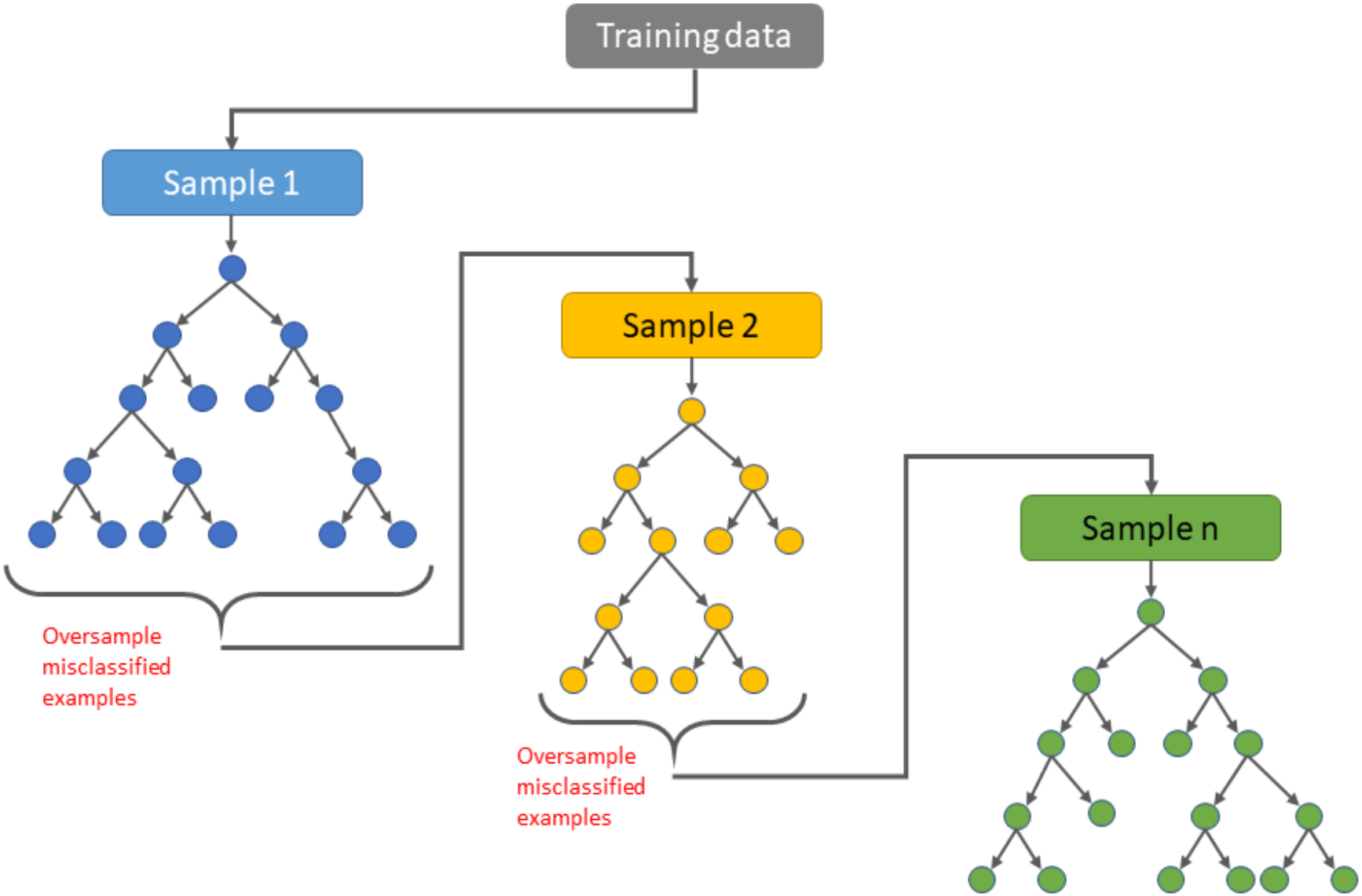
Boosting


Training data
Sample 1
Sample 2
Oversample
misclassified
examples

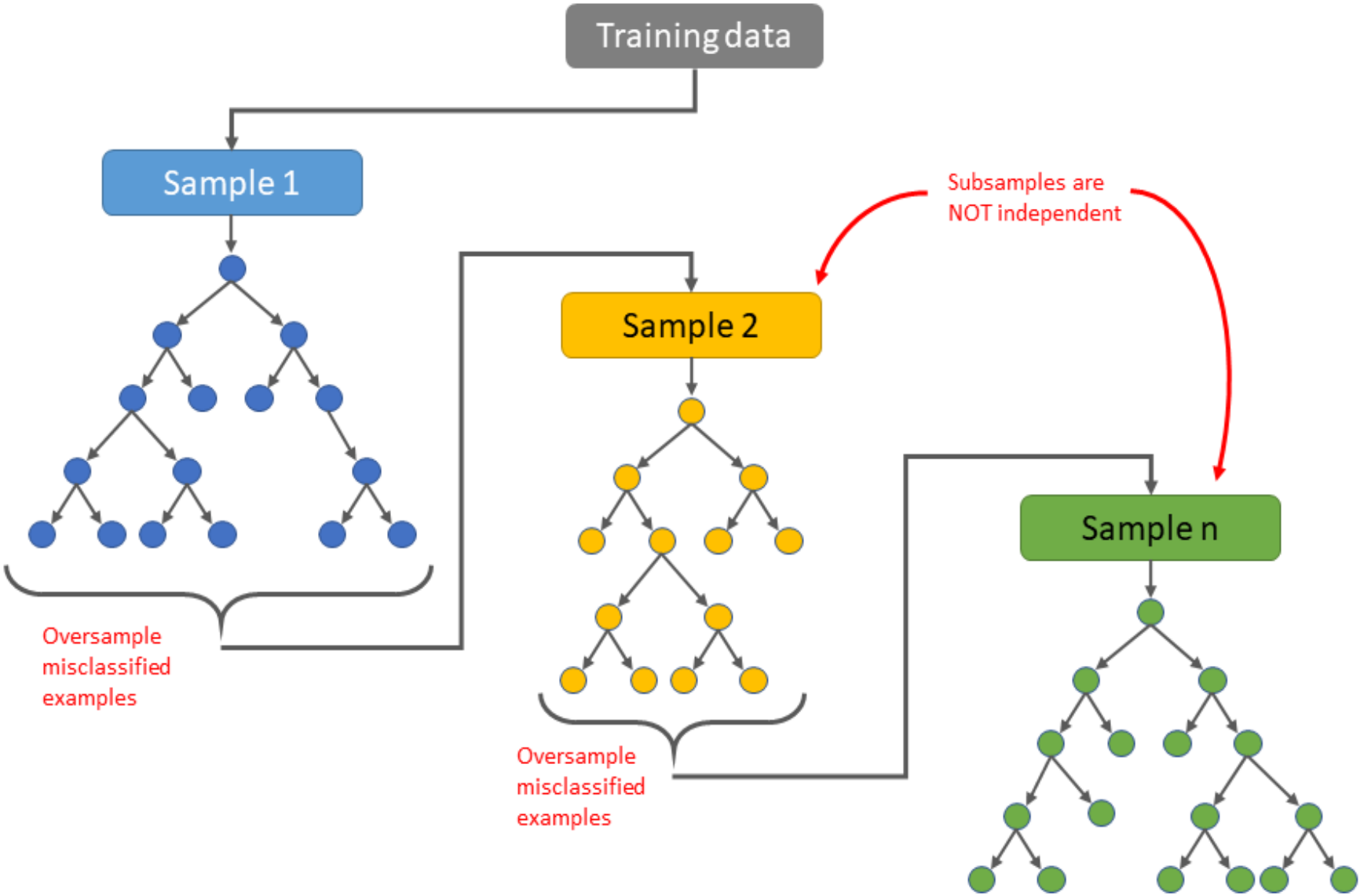
Boosting


Training data
Sample 1
Sample 2
Sample n
Oversample misclassified examples
Oversample misclassified examples

Boosting


Training data
Sample 1
Oversample misclassified examples
Sample 2
Oversample misclassified examples
Sample n

Boosting


Training data
Sample 1
Subsamples are
NOT independent
Sample 2
Sample n
Oversample
misclassified
examples
Oversample
misclassified
examples

Let's go to R

BAGGING

Bagging

This model will take time and might crash your computer.

So, let´s run it while we review the intuition behind this model.

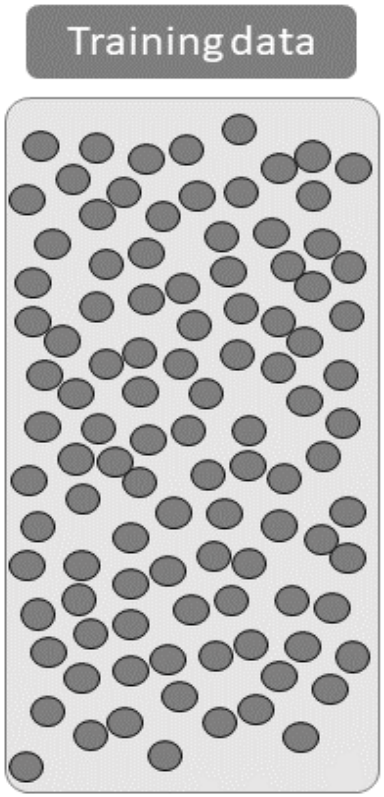
Let's go to R

Bagging

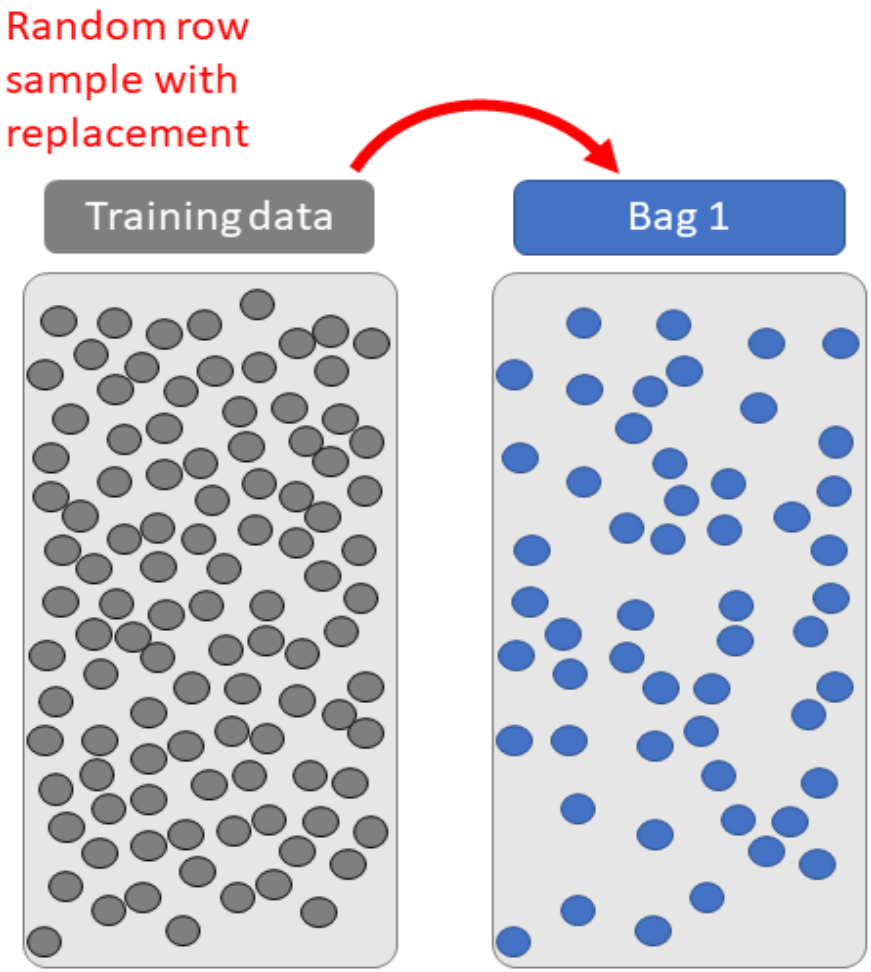
Bagging - Bootstrap Aggregation

- Bootstrapping
- Learning model
- Ensemble technique

Bagging



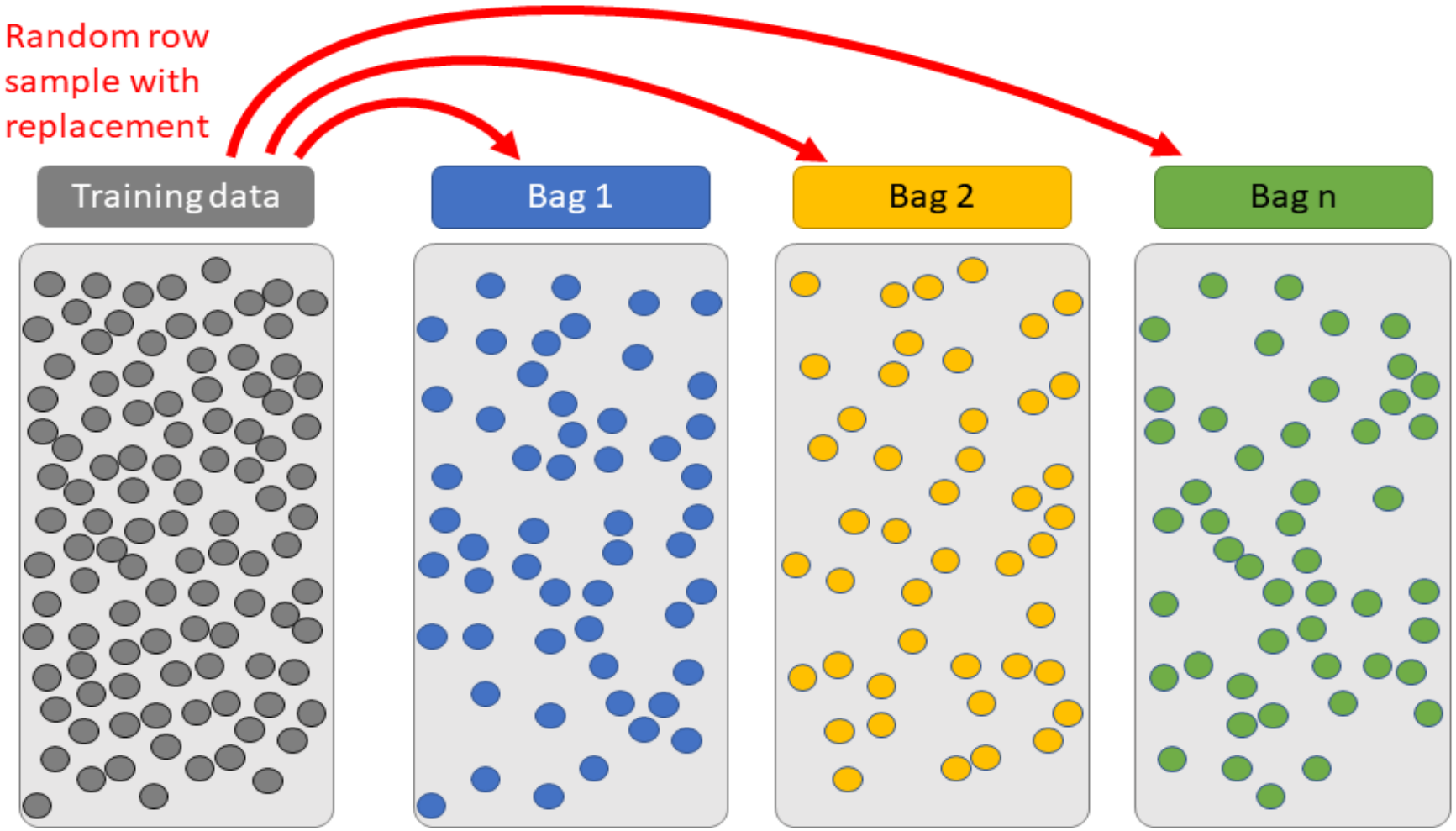
Bagging



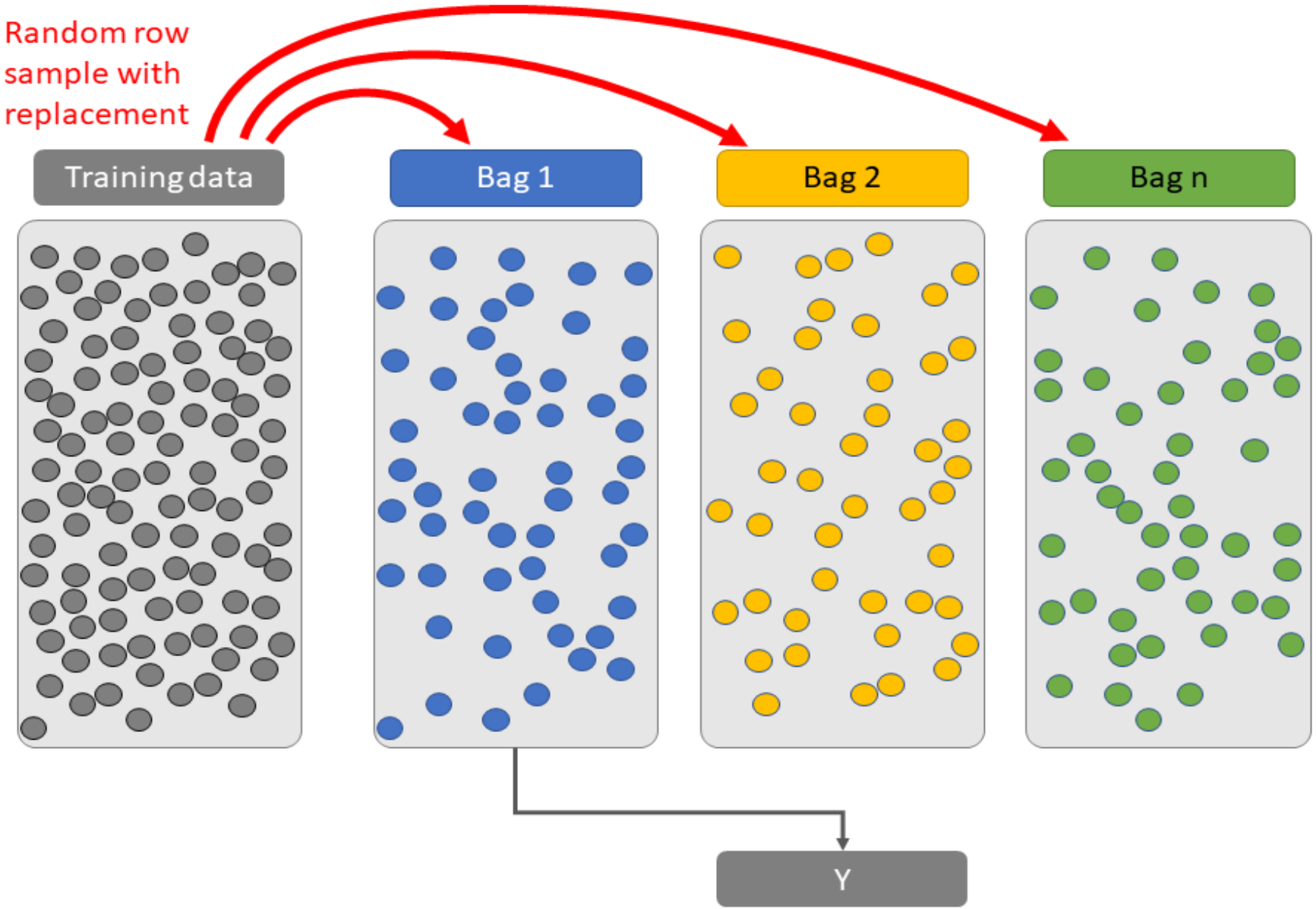
Bagging


Random row
sample with
replacement
Training data
Bag 1
Bag 2

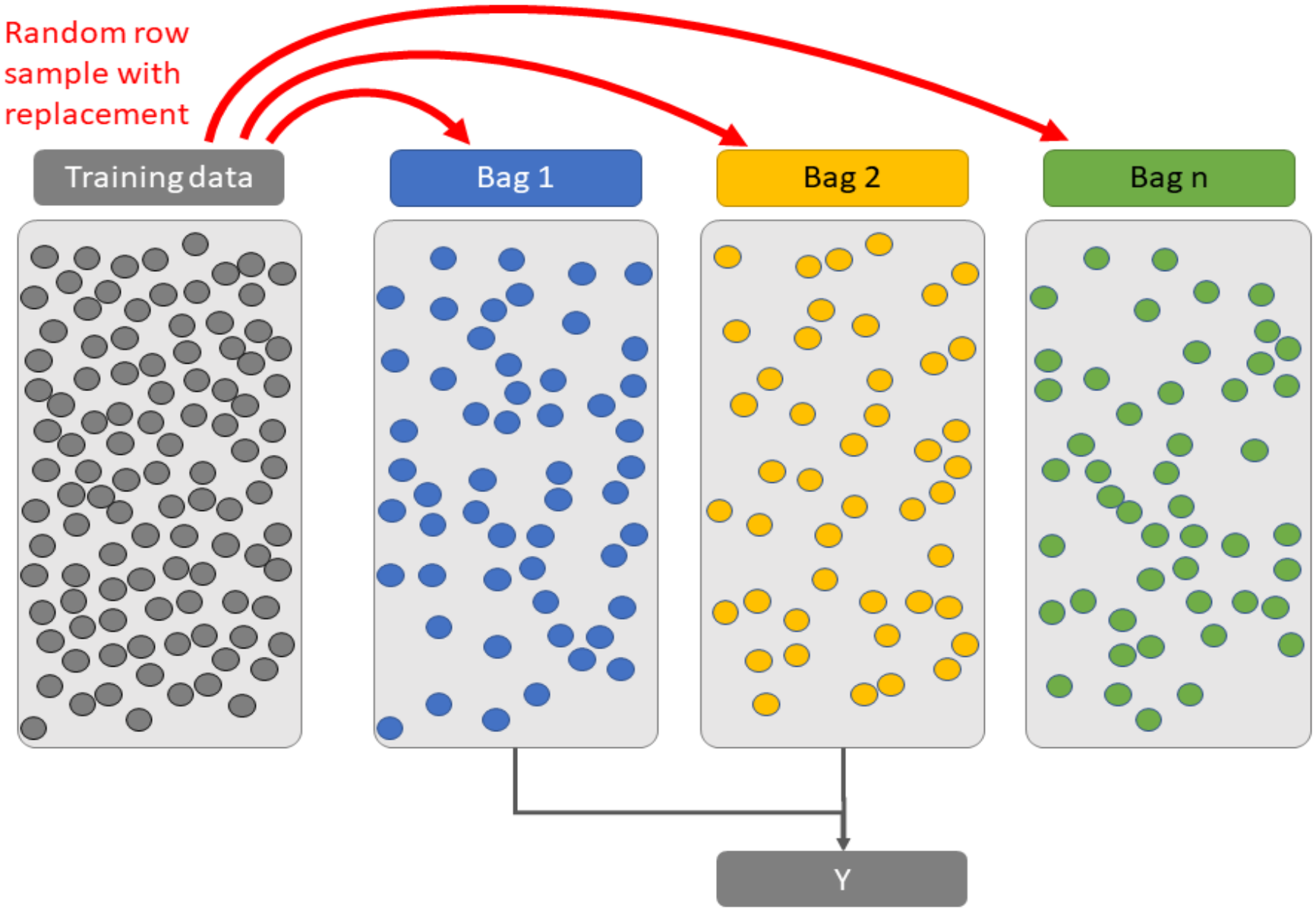
Bagging


Random row
sample with
replacement
Training data
Bag 1
Bag 2
Bag n

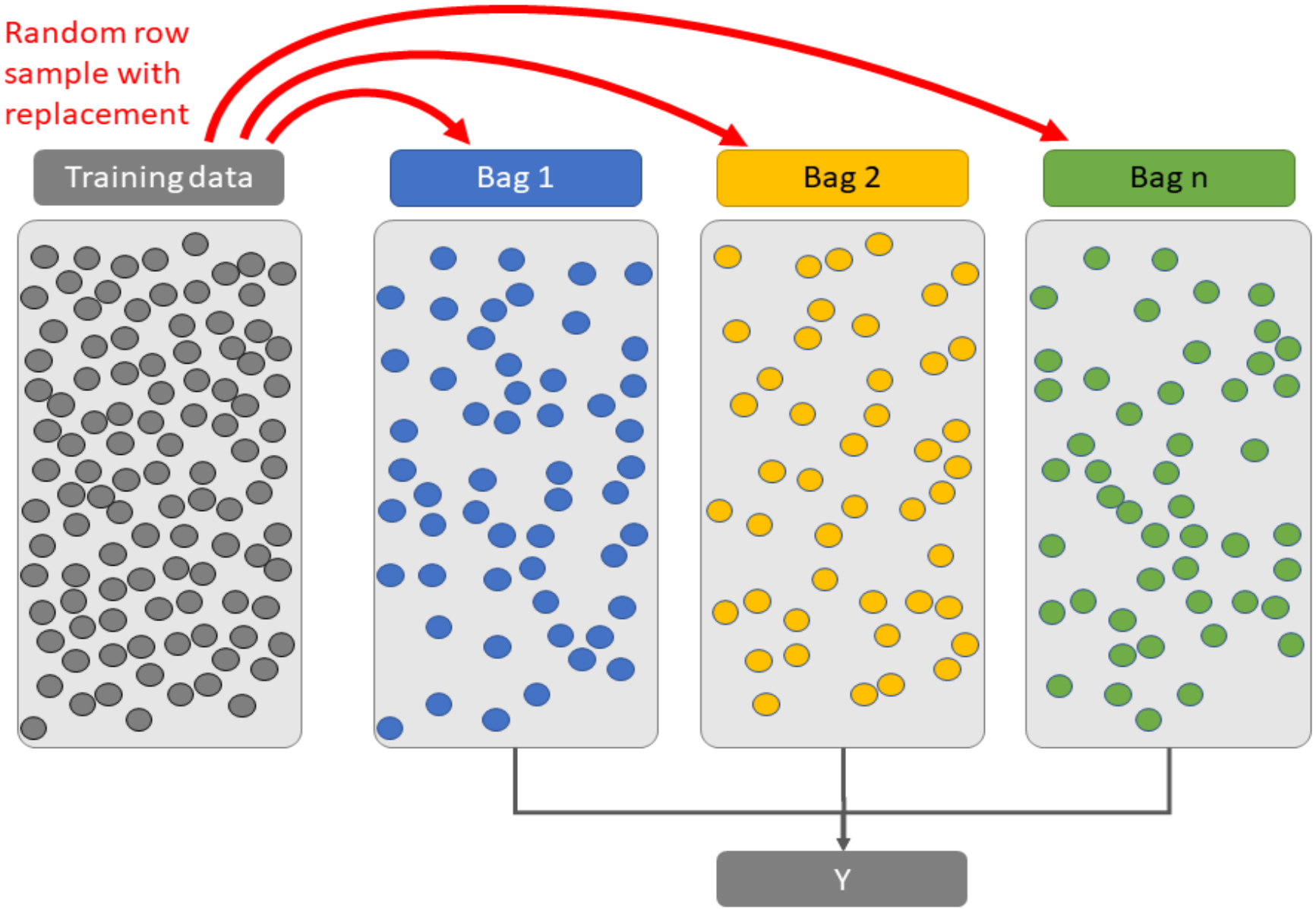
Bagging


Random row sample with replacement
Training data
Bag 1
Bag 2
Bag n
Y

Bagging


Random row sample with replacement
Training data
Bag 1
Bag 2
Bag n
Y

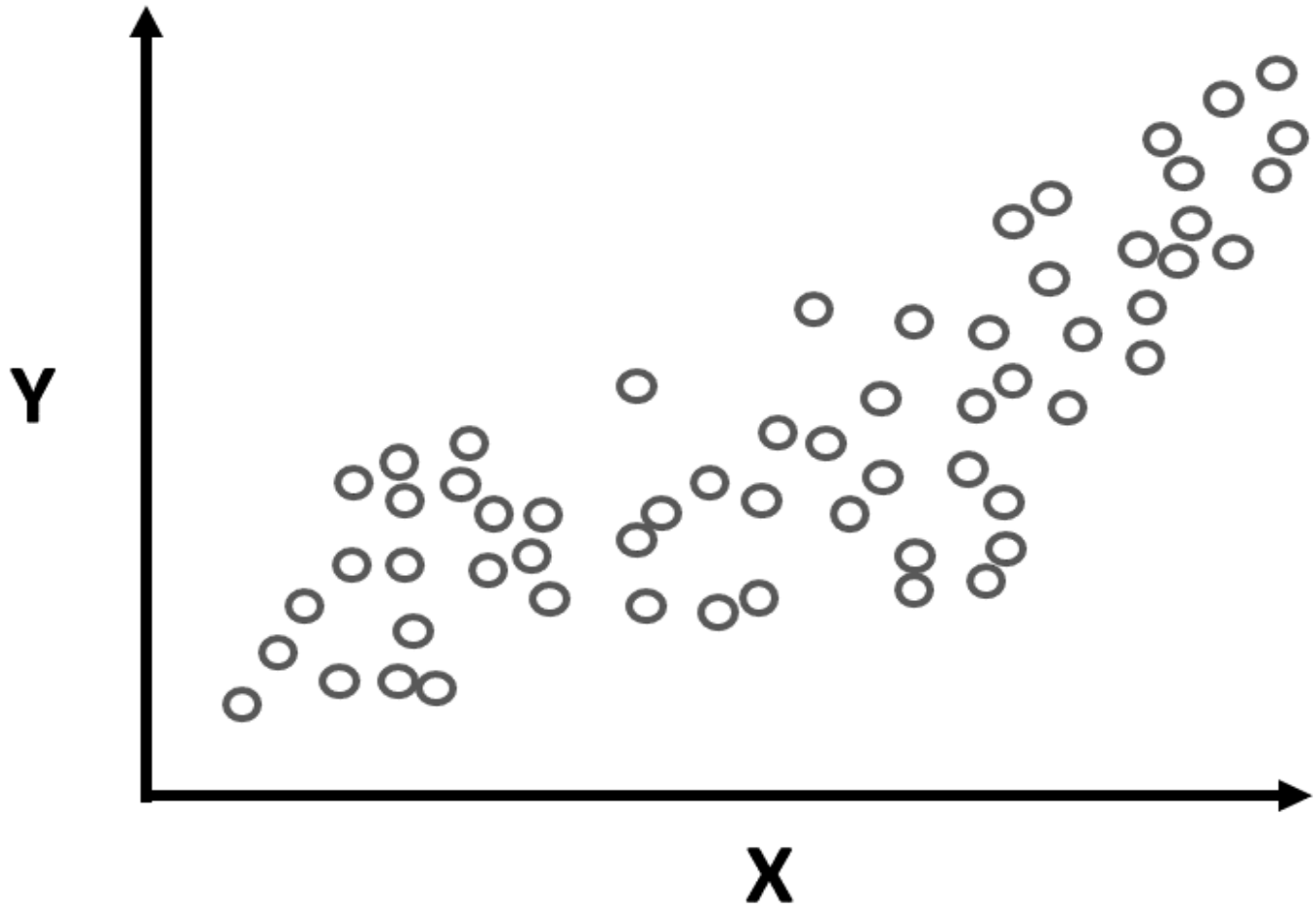
Bagging


Random row sample with replacement
Training data
Bag 1
Bag 2
Bag n
Y

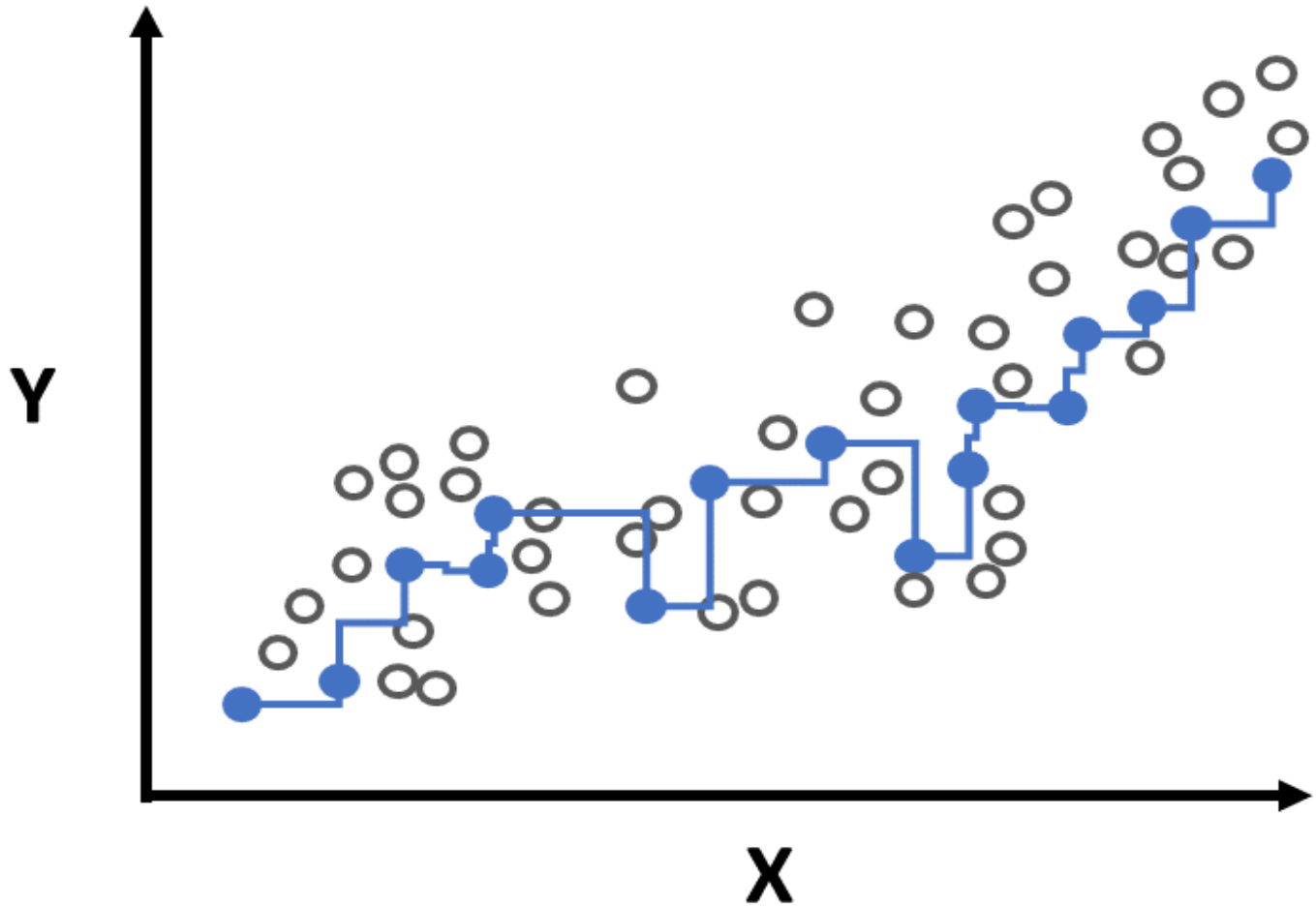
Bagging


Random row sample with replacement
Training data
Bag 1
Bag 2
Bag n
Aggregation

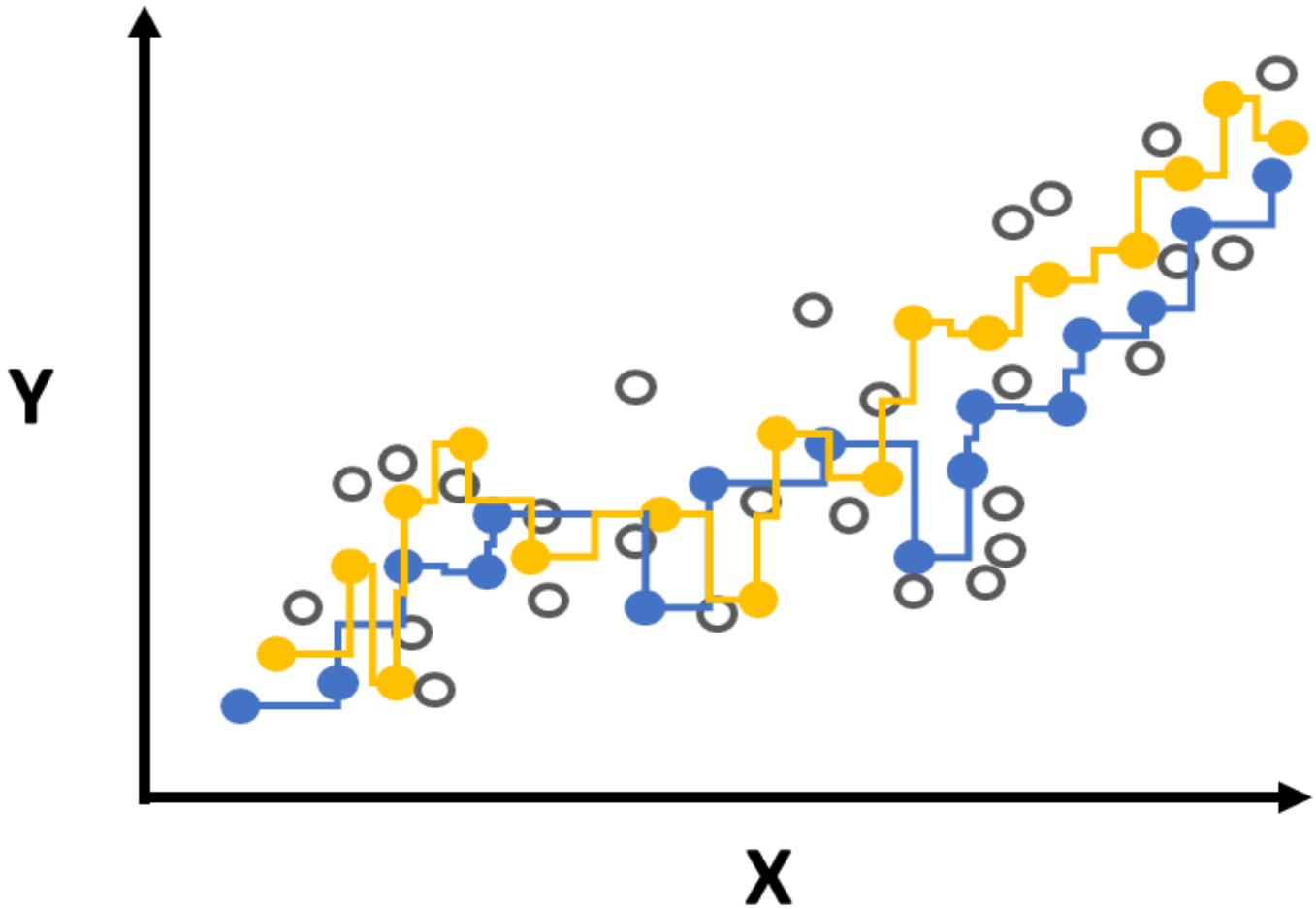
Bagging



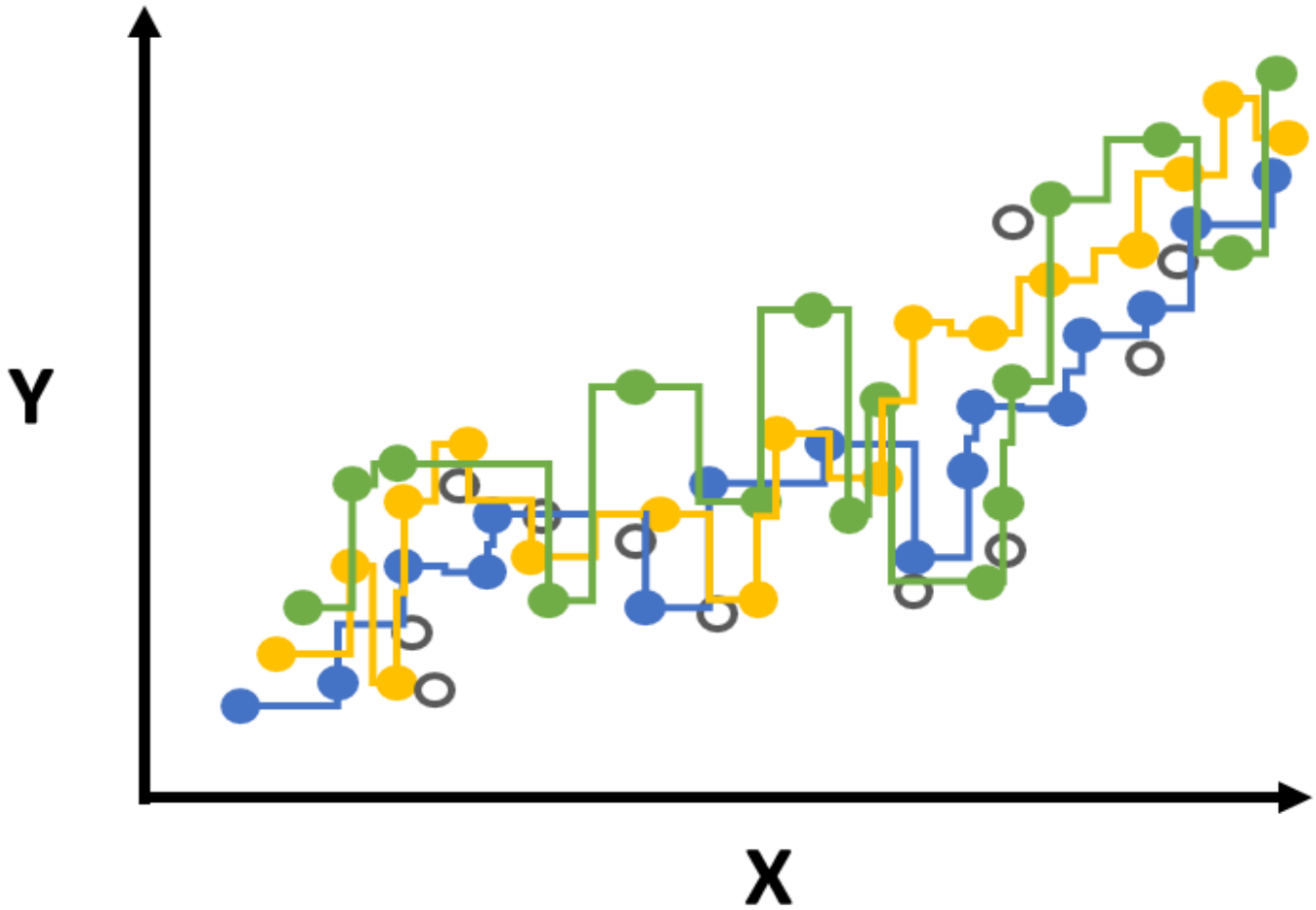
Bagging



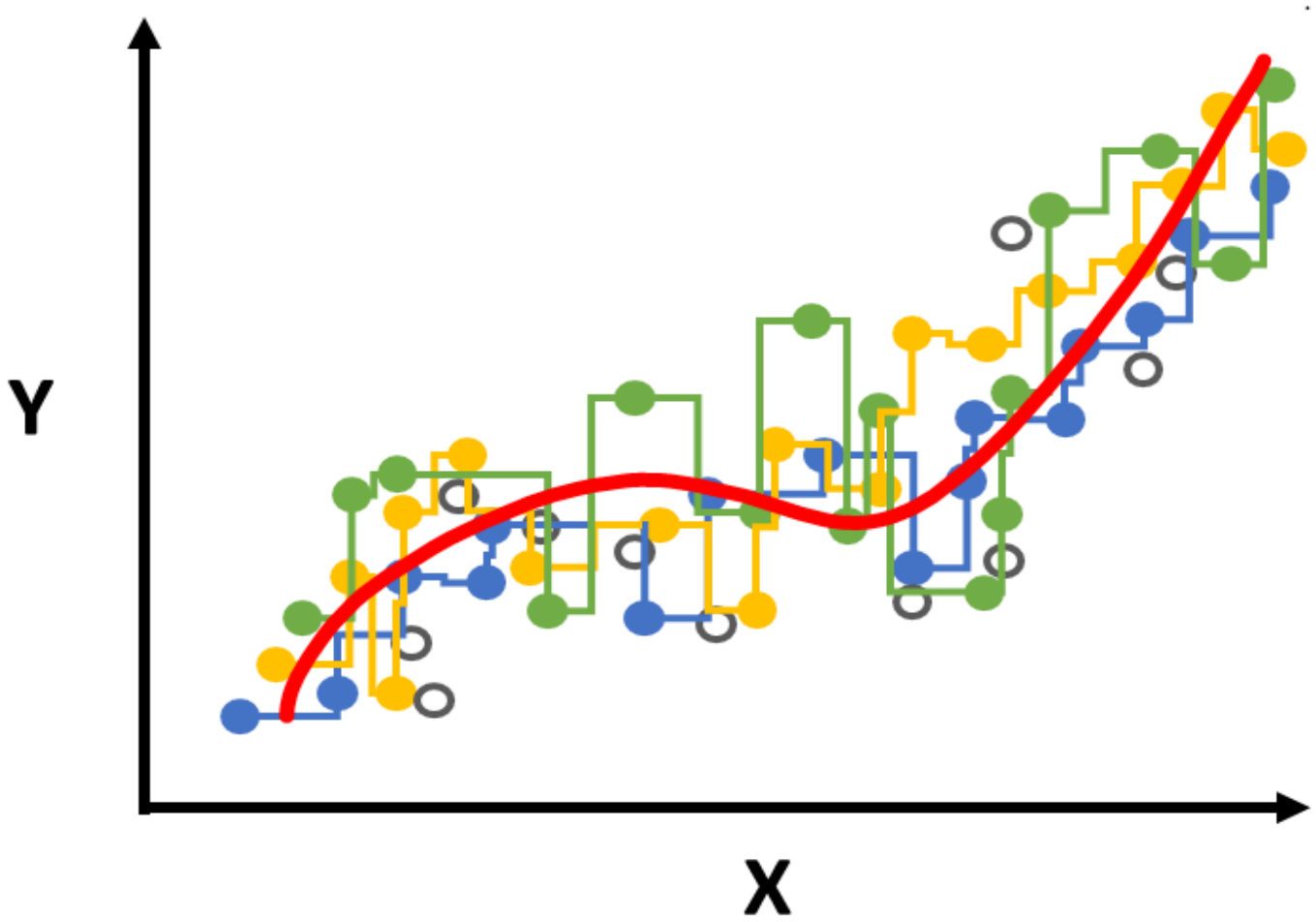
Bagging



Bagging



Bagging



Let's check the result in R

RANDOM FOREST

Bagging

This model will take time.

So, let´s run it while we review the intuition behind this model.

Let's go to R

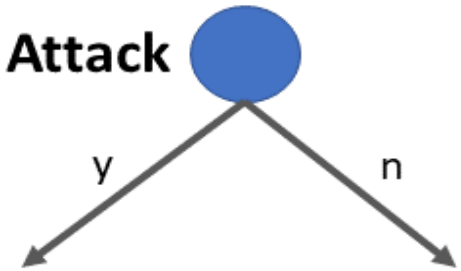
Random Forest

Random Forest

- The "forest" is an ensemble of decision trees
- that use the data features
- to predict the outcome

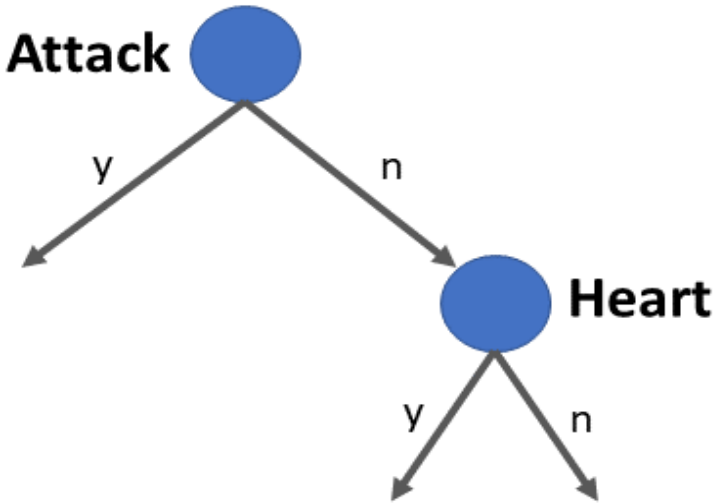
Decision Tree

Outcome: Terrorist attack?



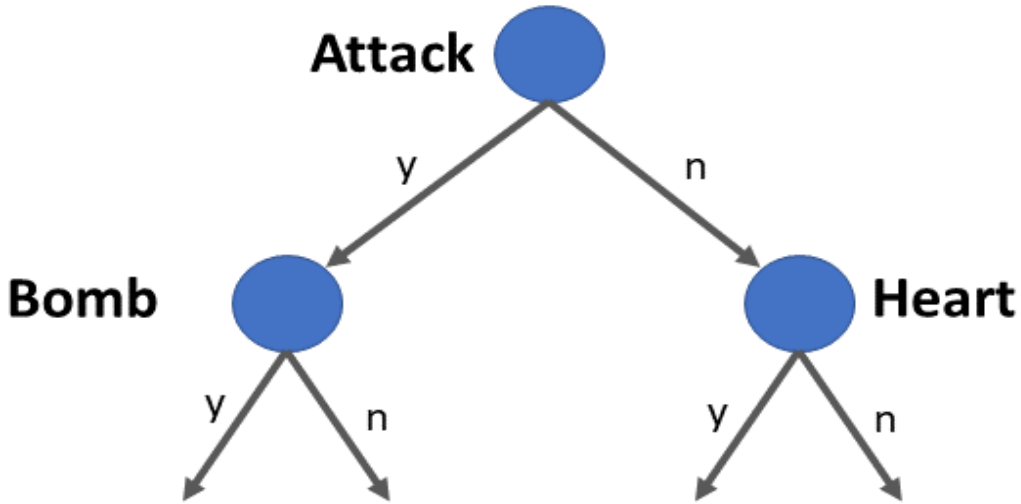
Decision Tree

Outcome: Terrorist attack?



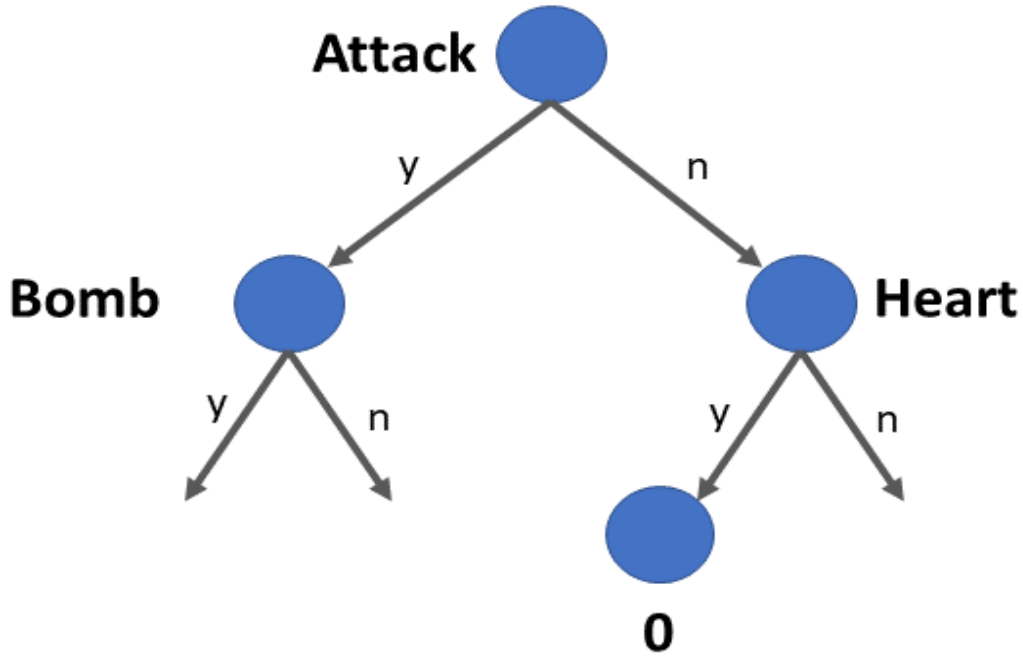
Decision Tree

Outcome: Terrorist attack?



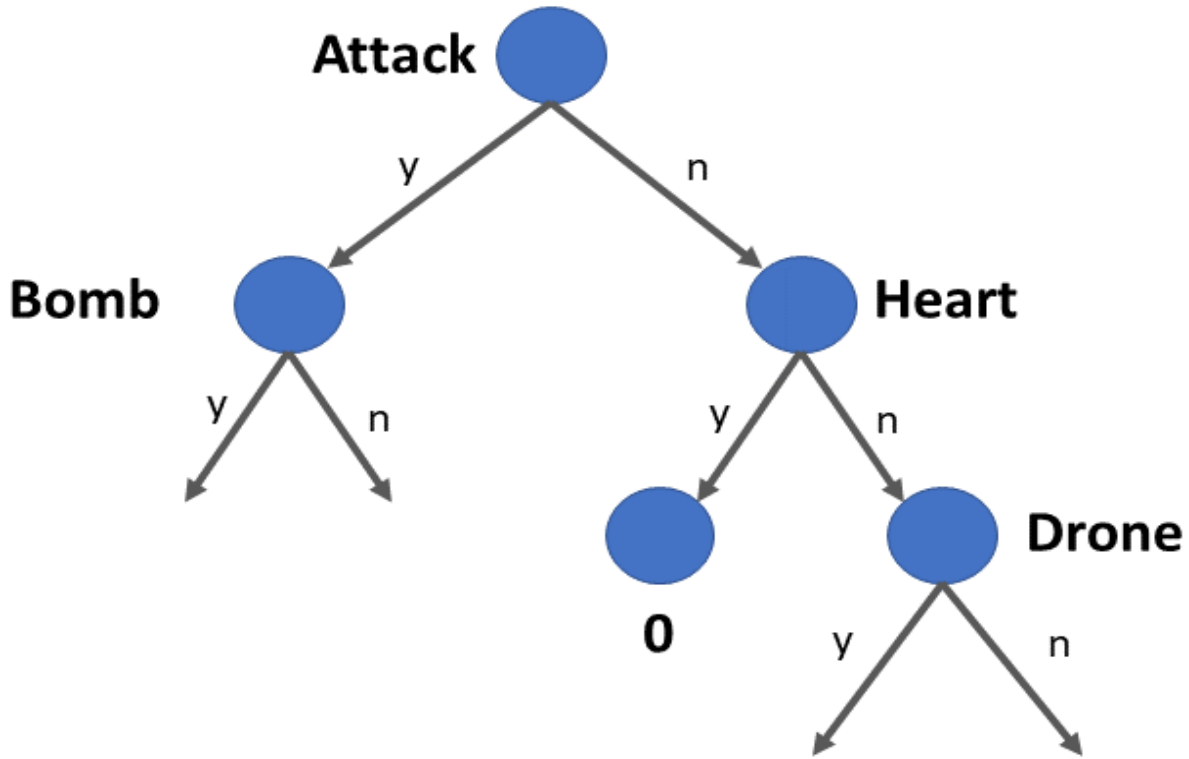
Decision Tree

Outcome: Terrorist attack?



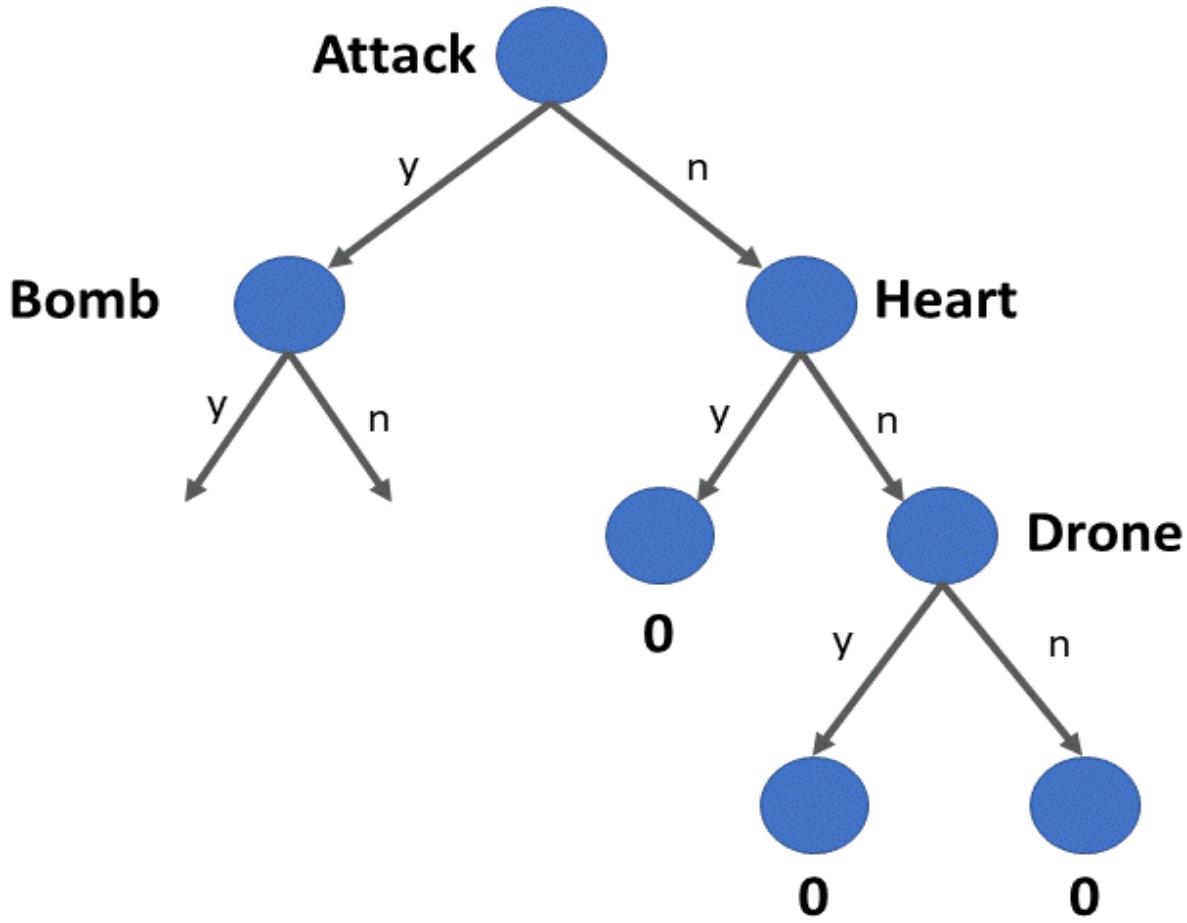
Decision Tree

Outcome: Terrorist attack?



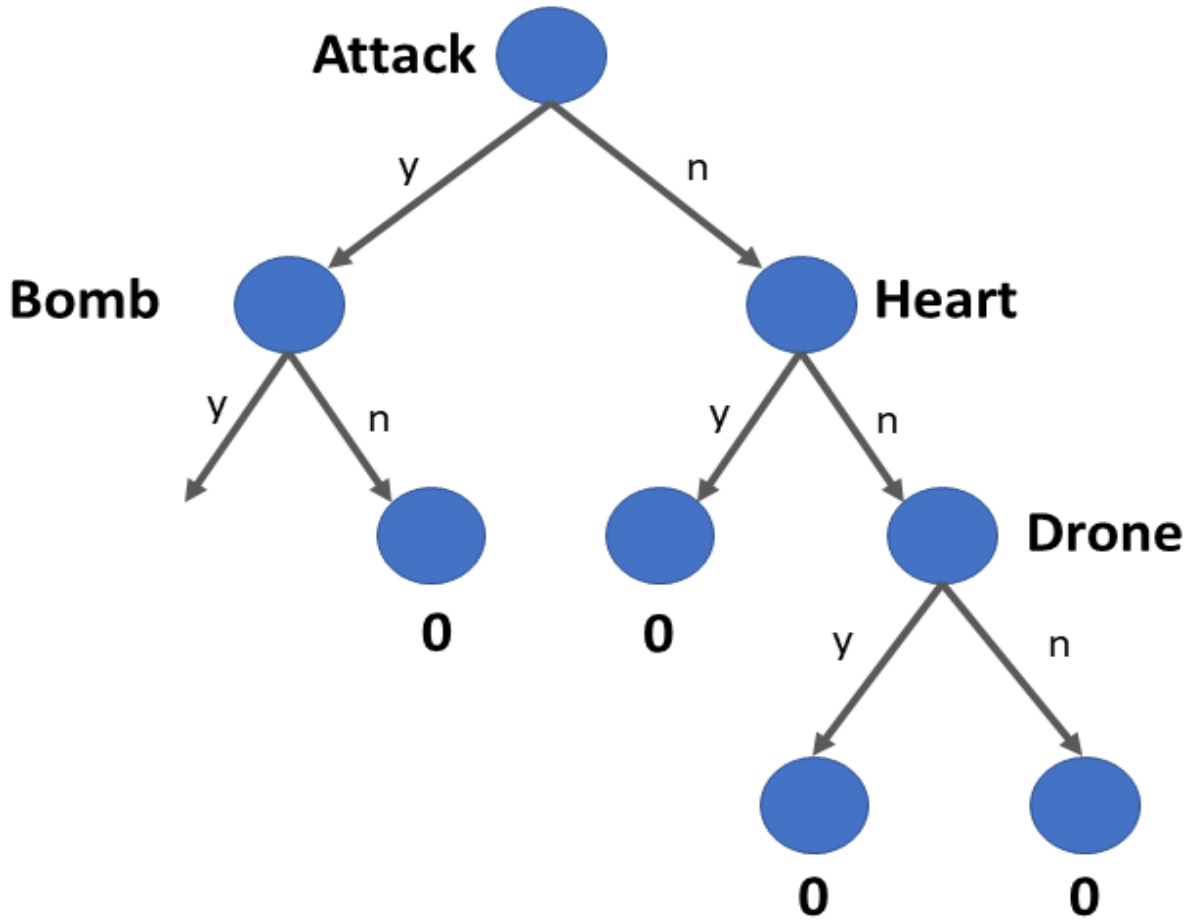
Decision Tree

Outcome: Terrorist attack?



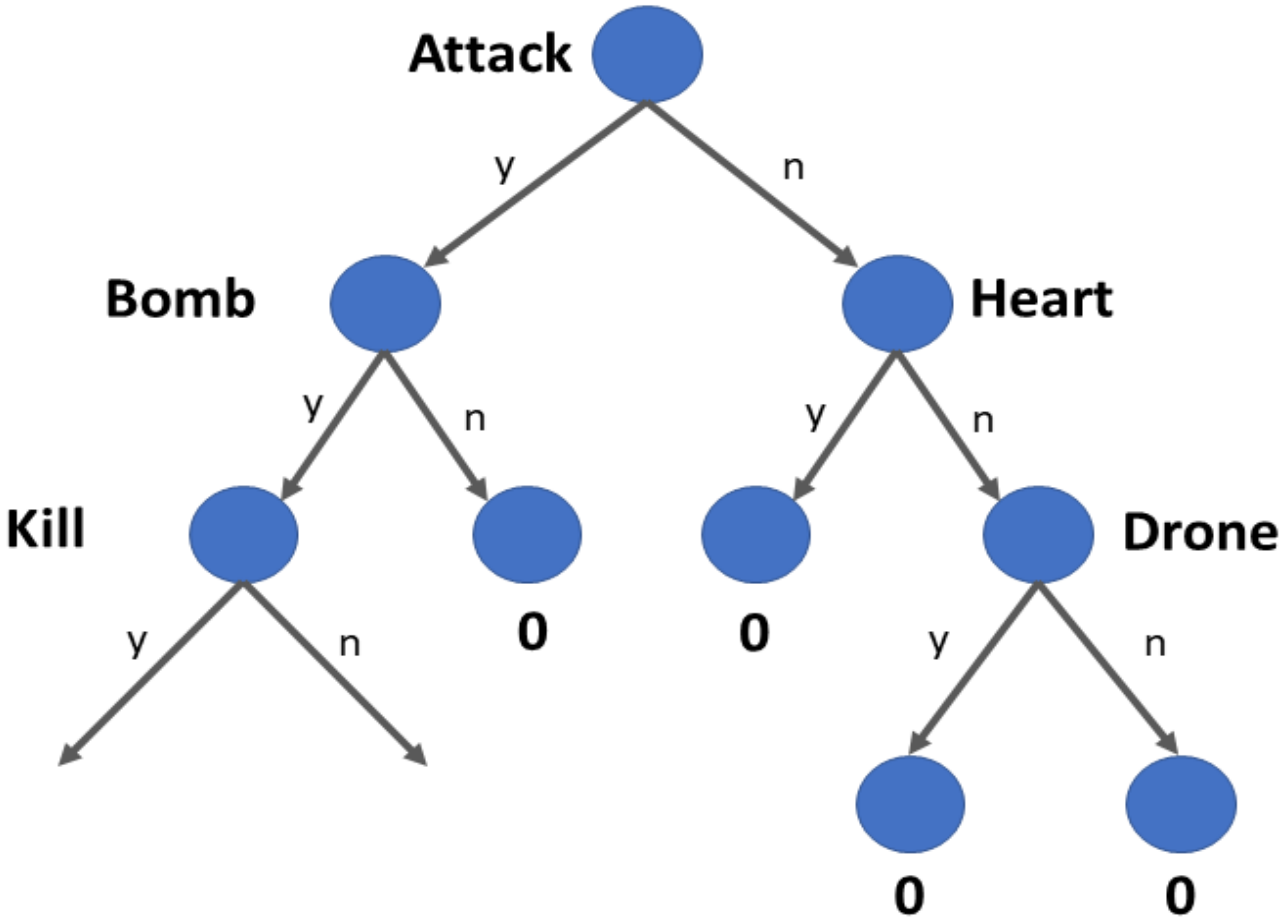
Decision Tree

Outcome: Terrorist attack?



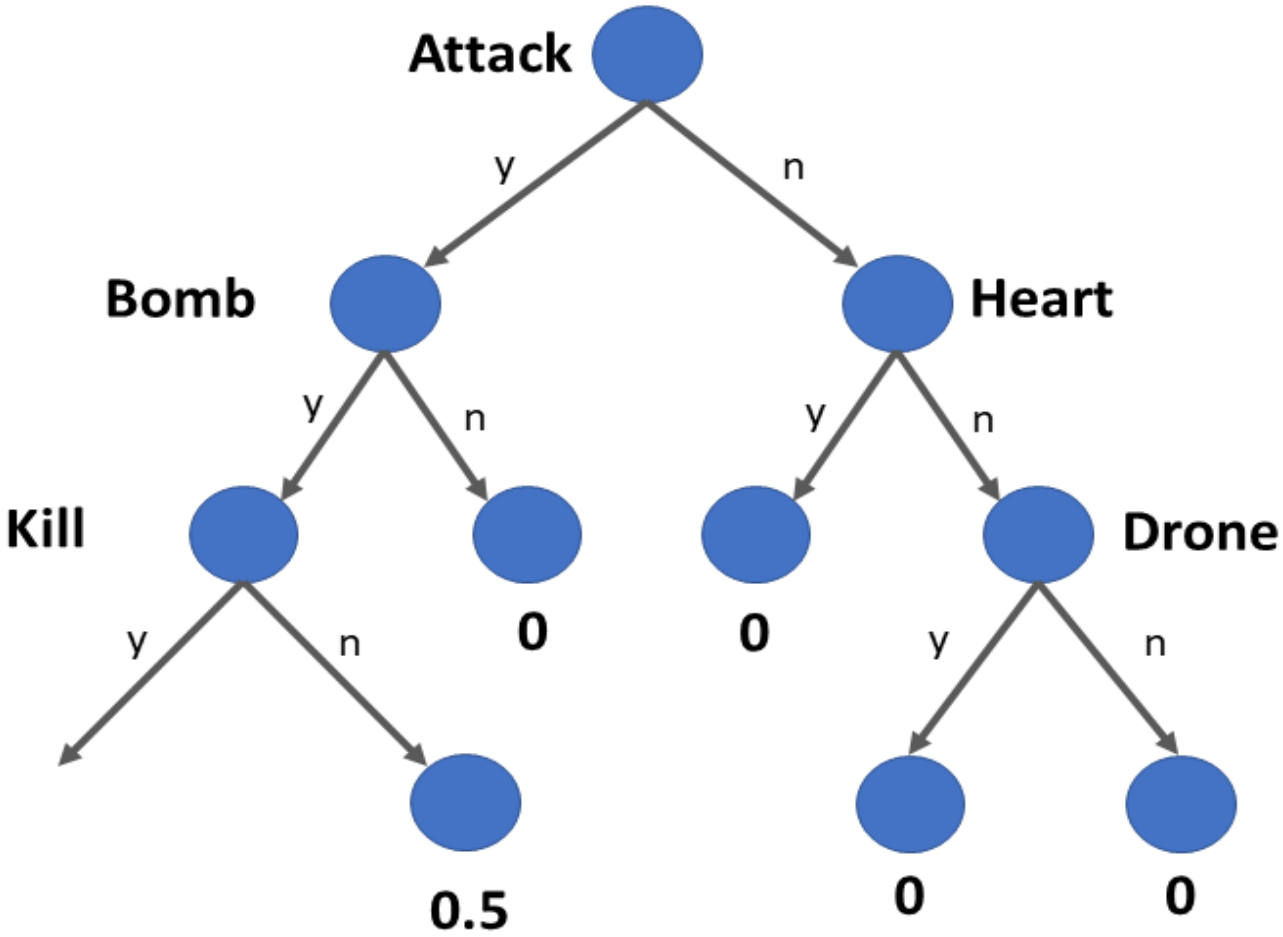
Decision Tree

Outcome: Terrorist attack?

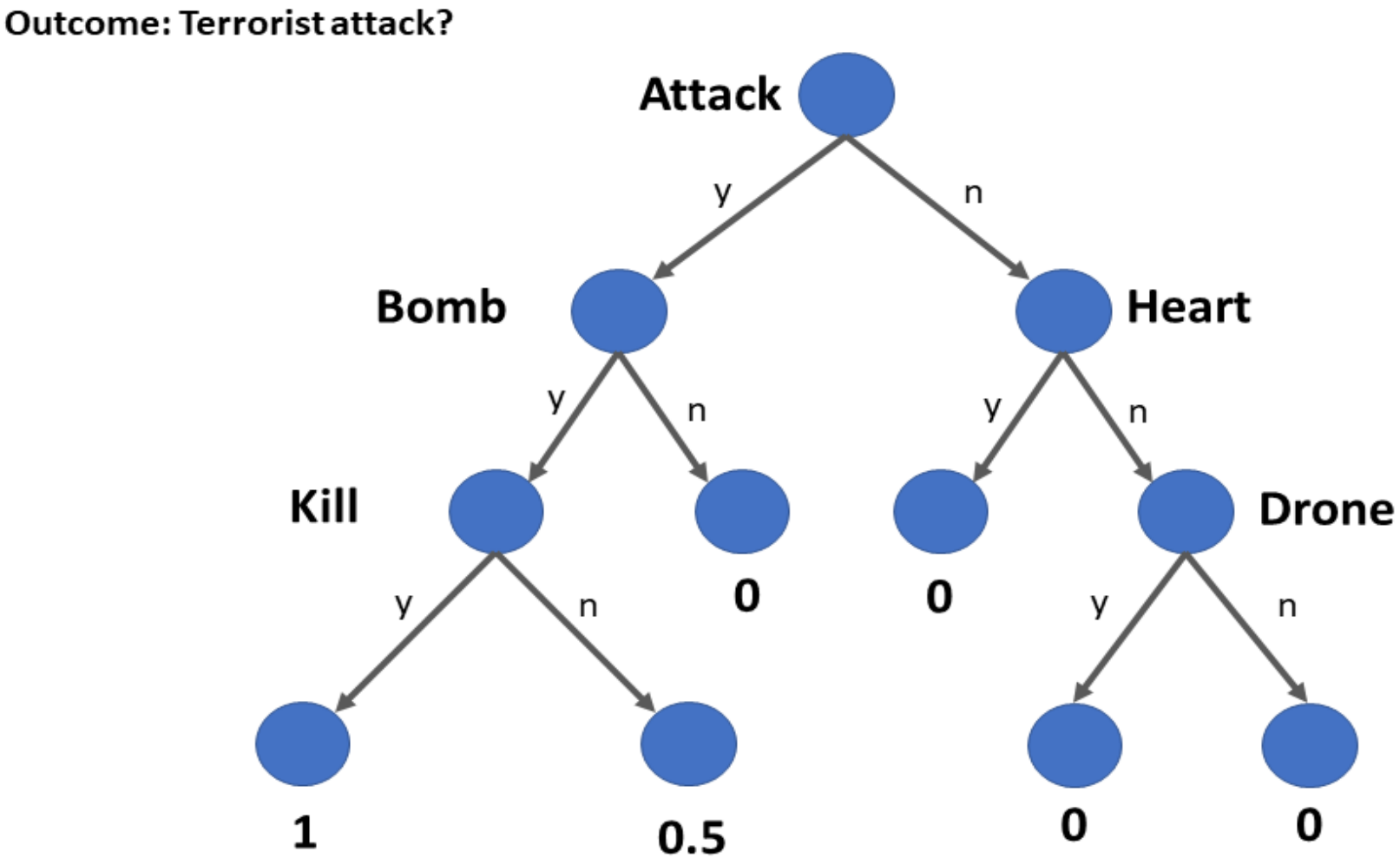


Decision Tree

Outcome: Terrorist attack?



Decision Tree



Random Forest

Training data

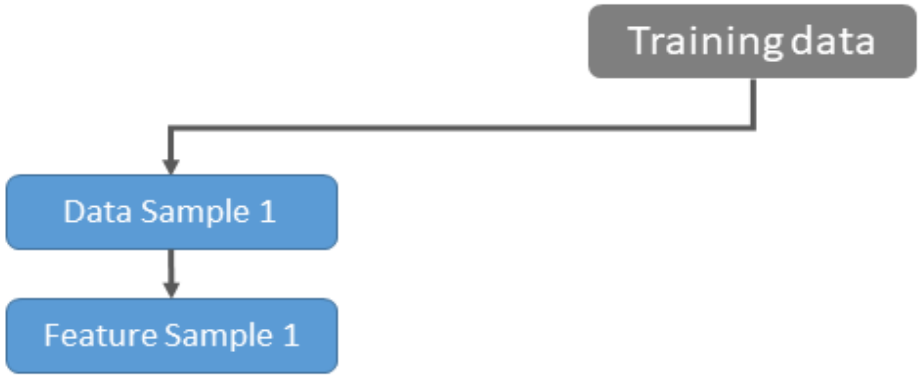
r	v1	v2	v3	v4	v5	v6	v7	v8
1	2	4	11	5	2	3	2	7
2	11	2	5	4	8	5	7	4
3	5	9	12	7	2	2	9	3
4	7	7	0	5	16	8	8	1
5	2	12	7	1	3	6	10	5
6	7	0	3	2	8	12	3	3
7	13	3	5	7	8	9	2	5
8	2	2	9	12	7	2	9	12
9	16	8	7	0	5	16	2	9
10	3	6	12	7	1	3	8	8

Random Forest

r	v1	v2	v3	v4	v5	v6	v7	v8
1	2	4	11	5	2	3	2	7
2	11	2	5	4	8	5	7	4
3	5	9	12	7	2	2	9	3
4	7	7	0	5	16	8	8	1
5	2	12	7	1	3	6	10	5
6	7	0	3	2	8	12	3	3
7	13	3	5	7	8	9	2	5
8	2	2	9	12	7	2	9	12
9	16	8	7	0	5	16	2	9
10	3	6	12	7	1	3	8	8

r	v1	v2	v3	v4	v5	v6	v7	v8
2	11	2	5	4	8	5	7	4
3	5	9	12	7	2	2	9	3
5	2	12	7	1	3	6	10	5
8	2	2	9	12	7	2	9	12
10	3	6	12	7	1	3	8	8

Random Forest

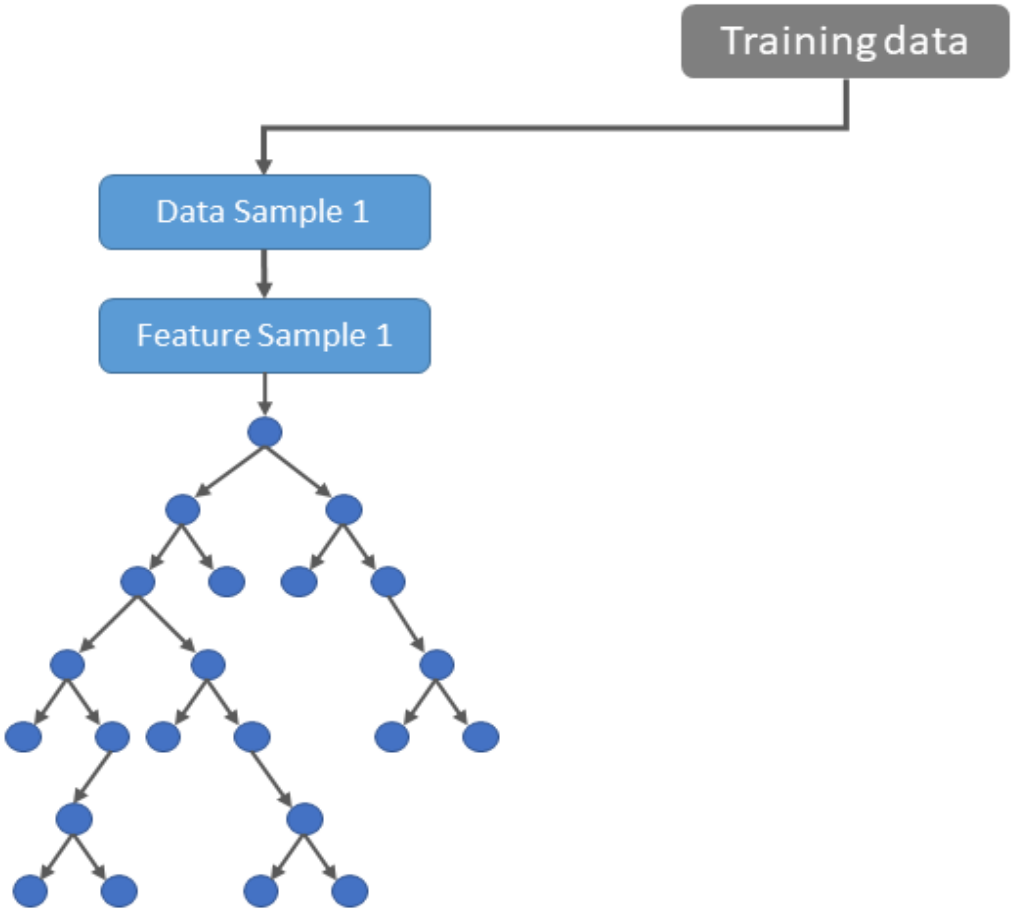


r	v1	v2	v3	v4	v5	v6	v7	v8
1	2	4	11	5	2	3	2	7
2	11	2	5	4	8	5	7	4
3	5	9	12	7	2	2	9	3
4	7	7	0	5	16	8	8	1
5	2	12	7	1	3	6	10	5
6	7	0	3	2	8	12	3	3
7	13	3	5	7	8	9	2	5
8	2	2	9	12	7	2	9	12
9	16	8	7	0	5	16	2	9
10	3	6	12	7	1	3	8	8

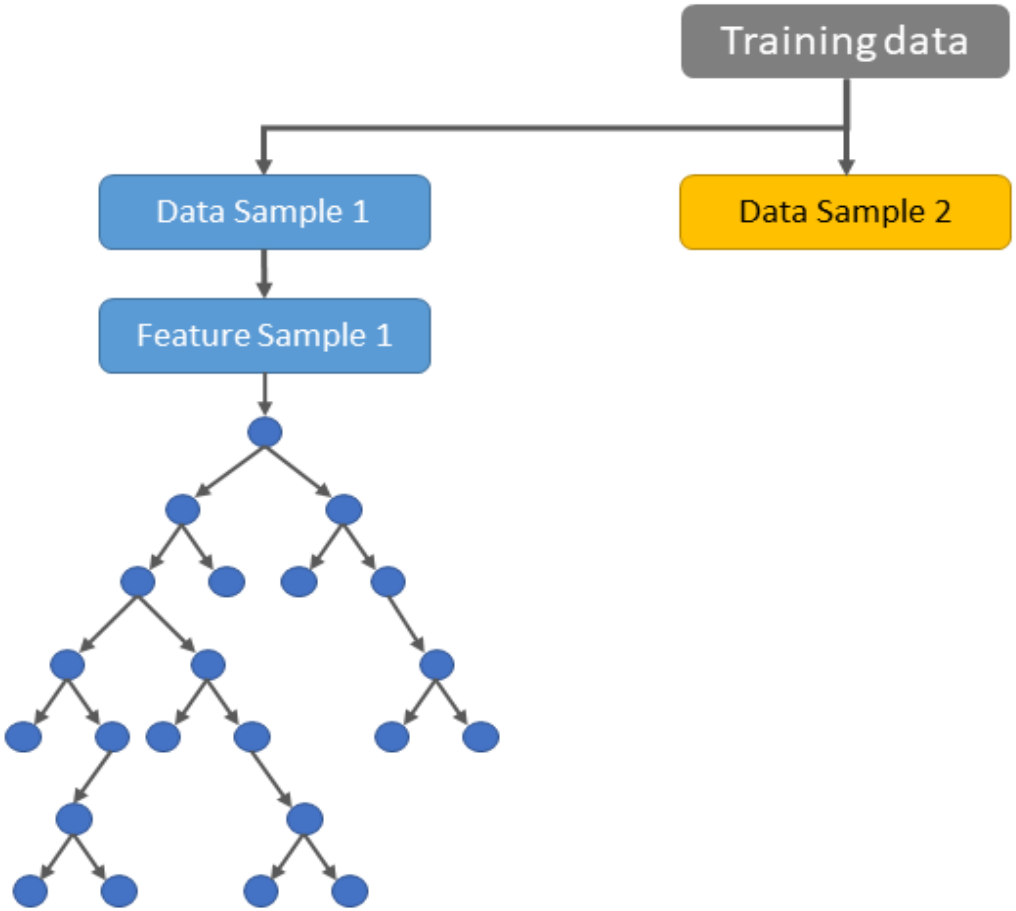
r	v1	v2	v3	v4	v5	v6	v7	v8
2	11	2	5	4	8	5	7	4
3	5	9	12	7	2	2	9	3
5	2	12	7	1	3	6	10	5
8	2	2	9	12	7	2	9	12
10	3	6	12	7	1	3	8	8

r	v1	v3	v4	v7
2	11	5	4	7
3	5	12	7	9
5	2	7	1	10
8	2	9	12	9
10	3	12	7	8

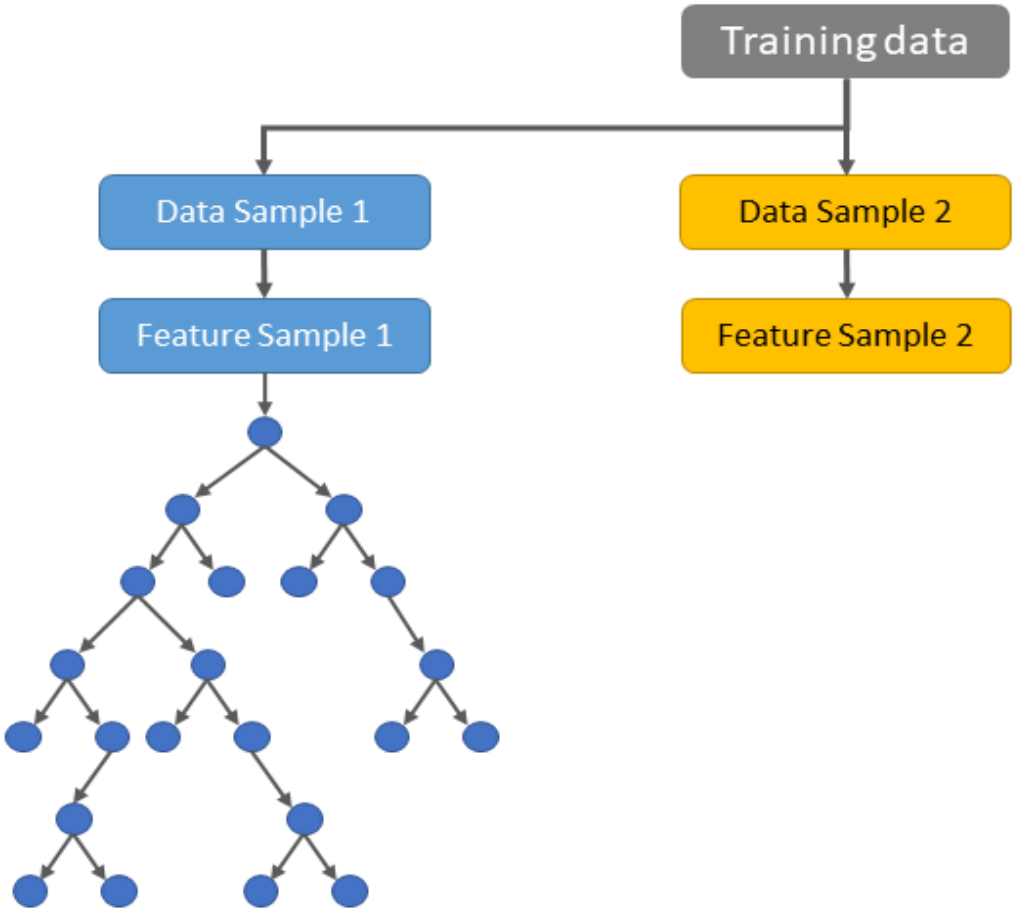
Random Forest



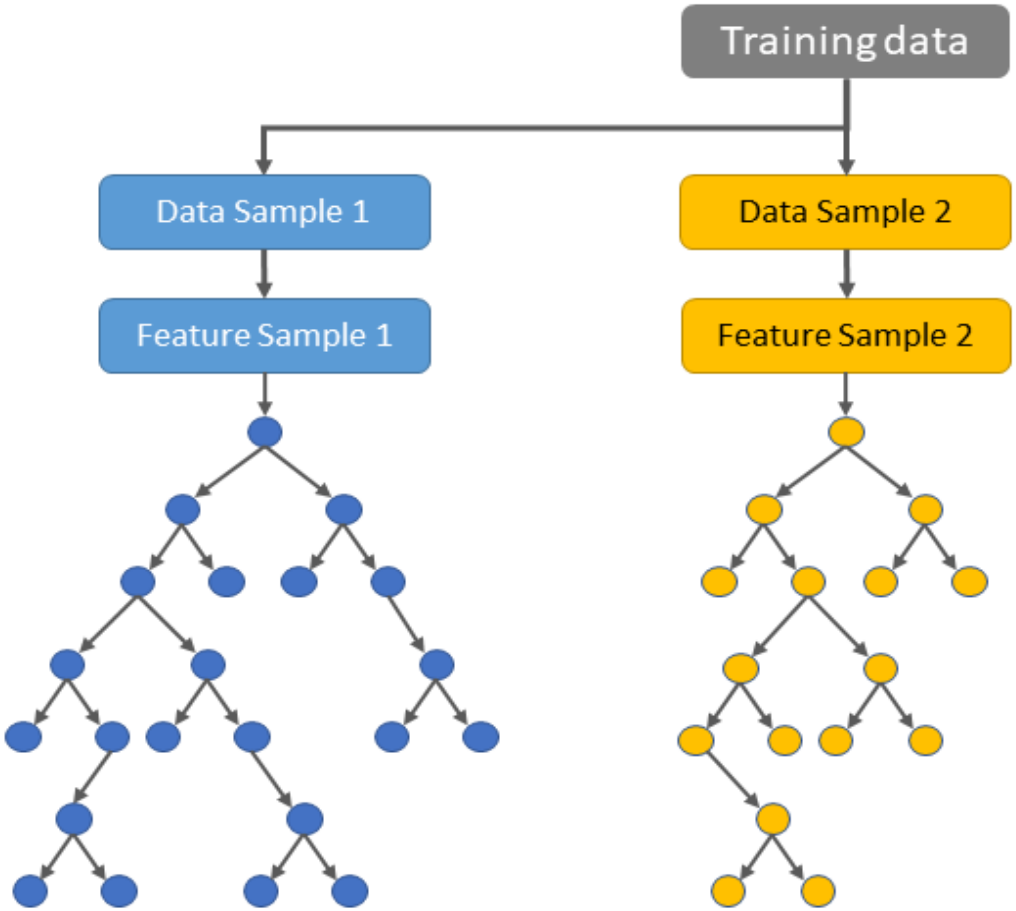
Random Forest



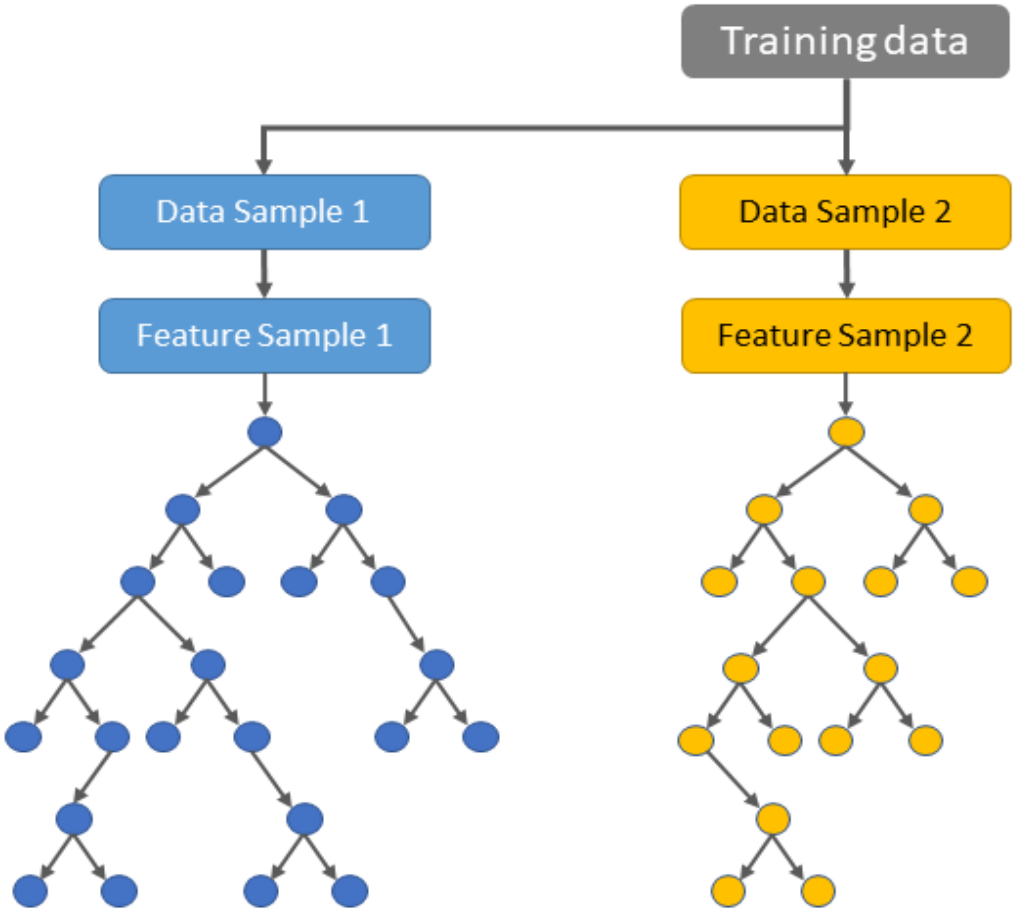
Random Forest



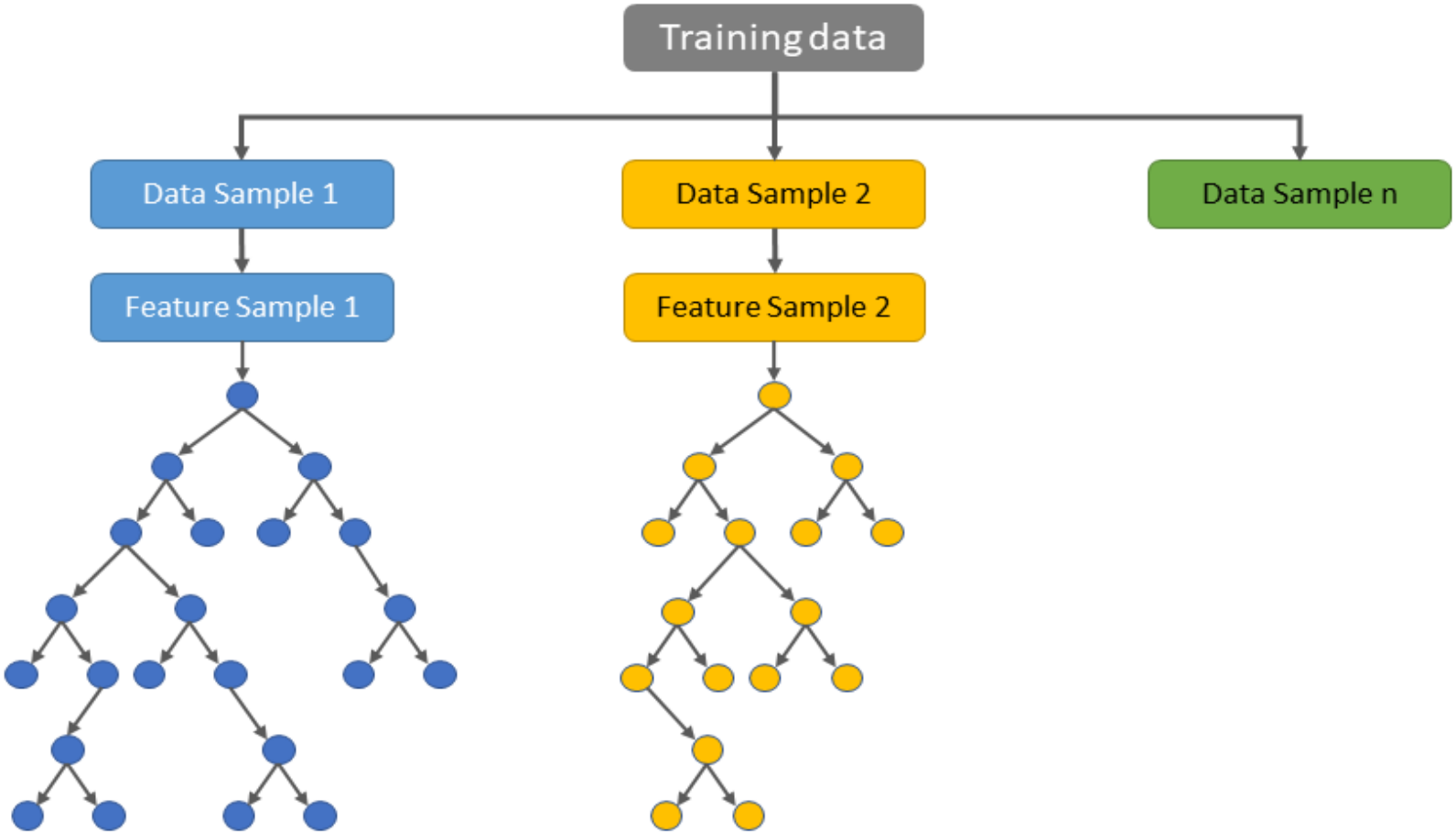
Random Forest



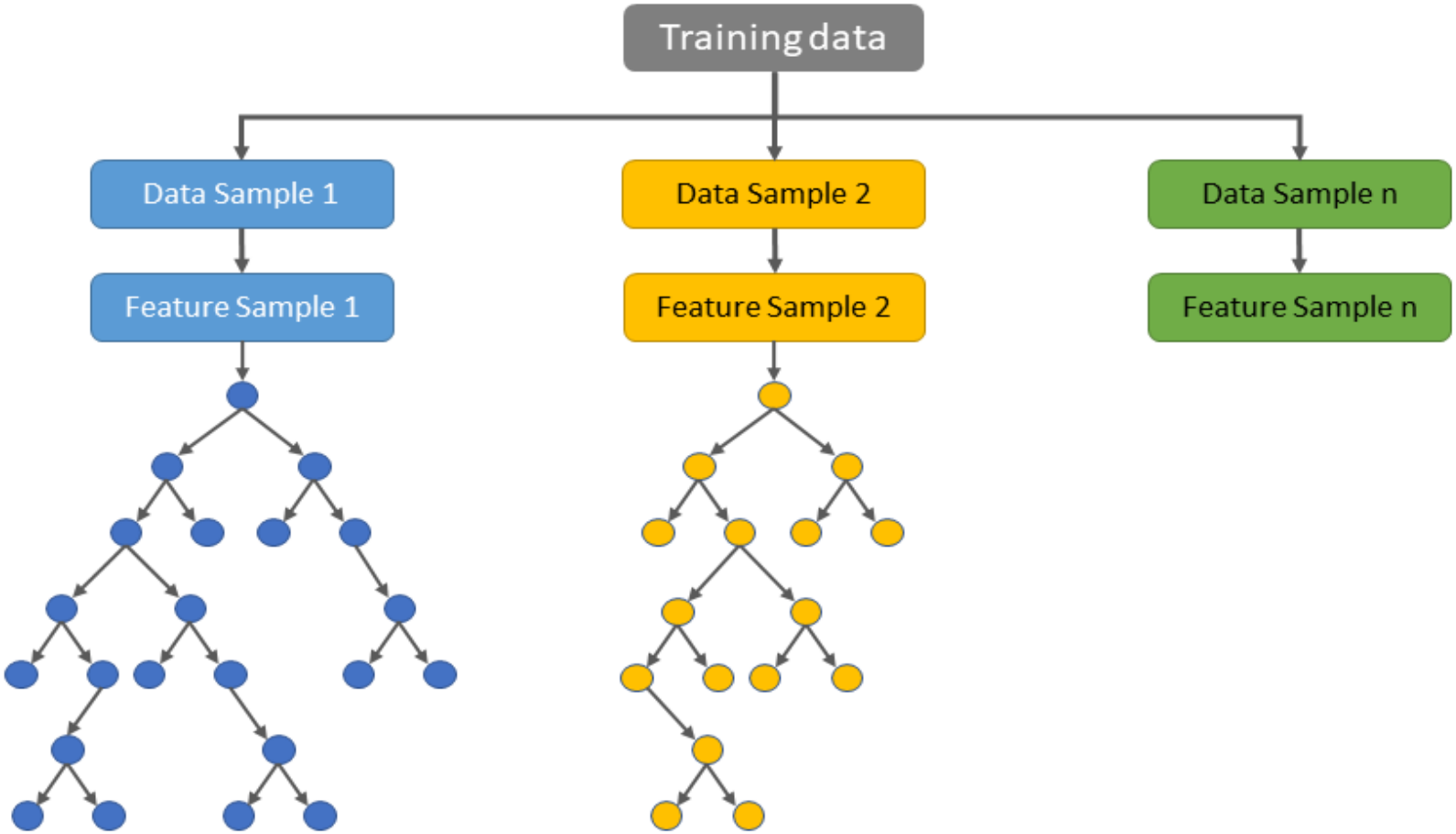
Random Forest



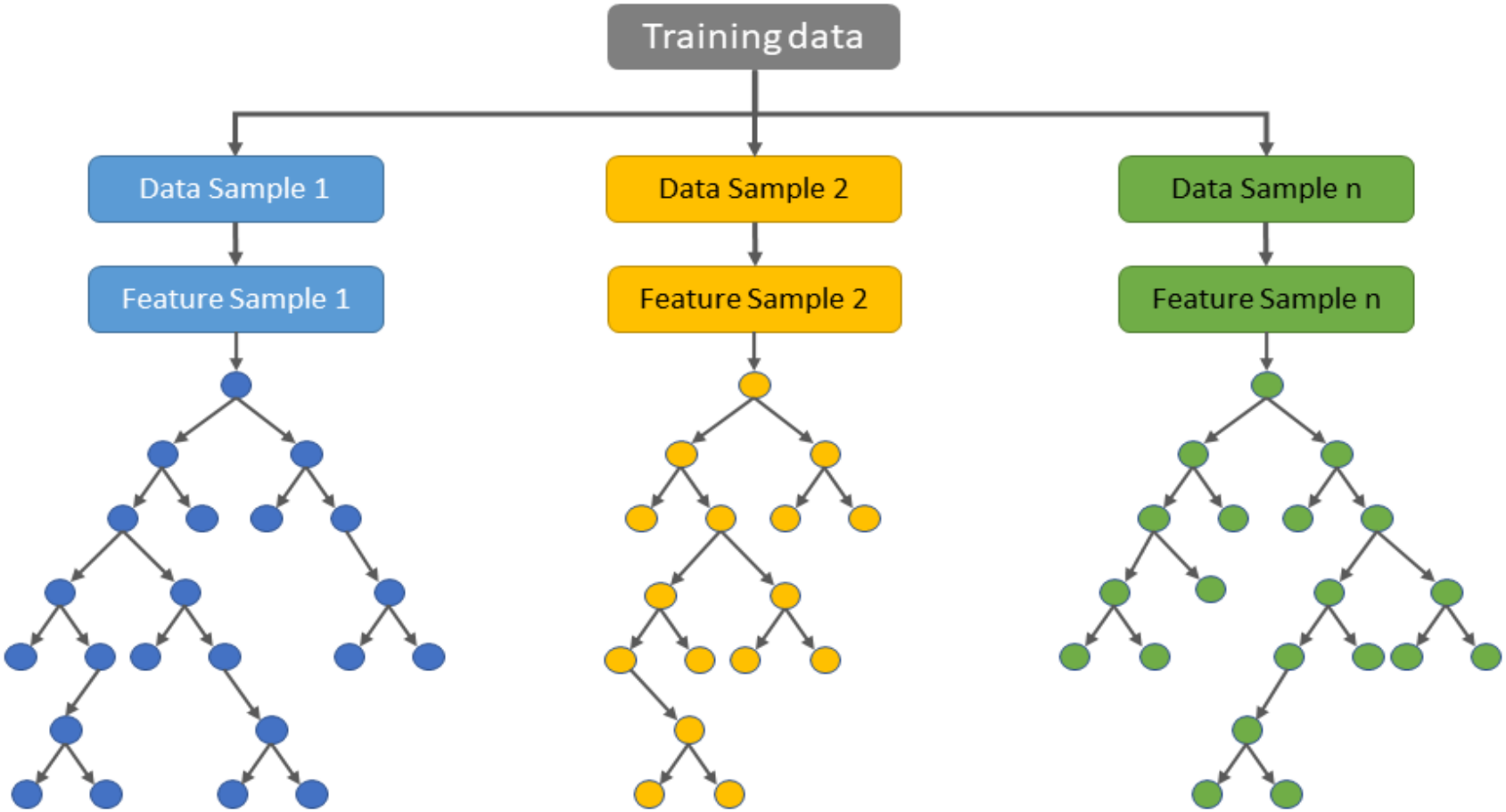
Random Forest



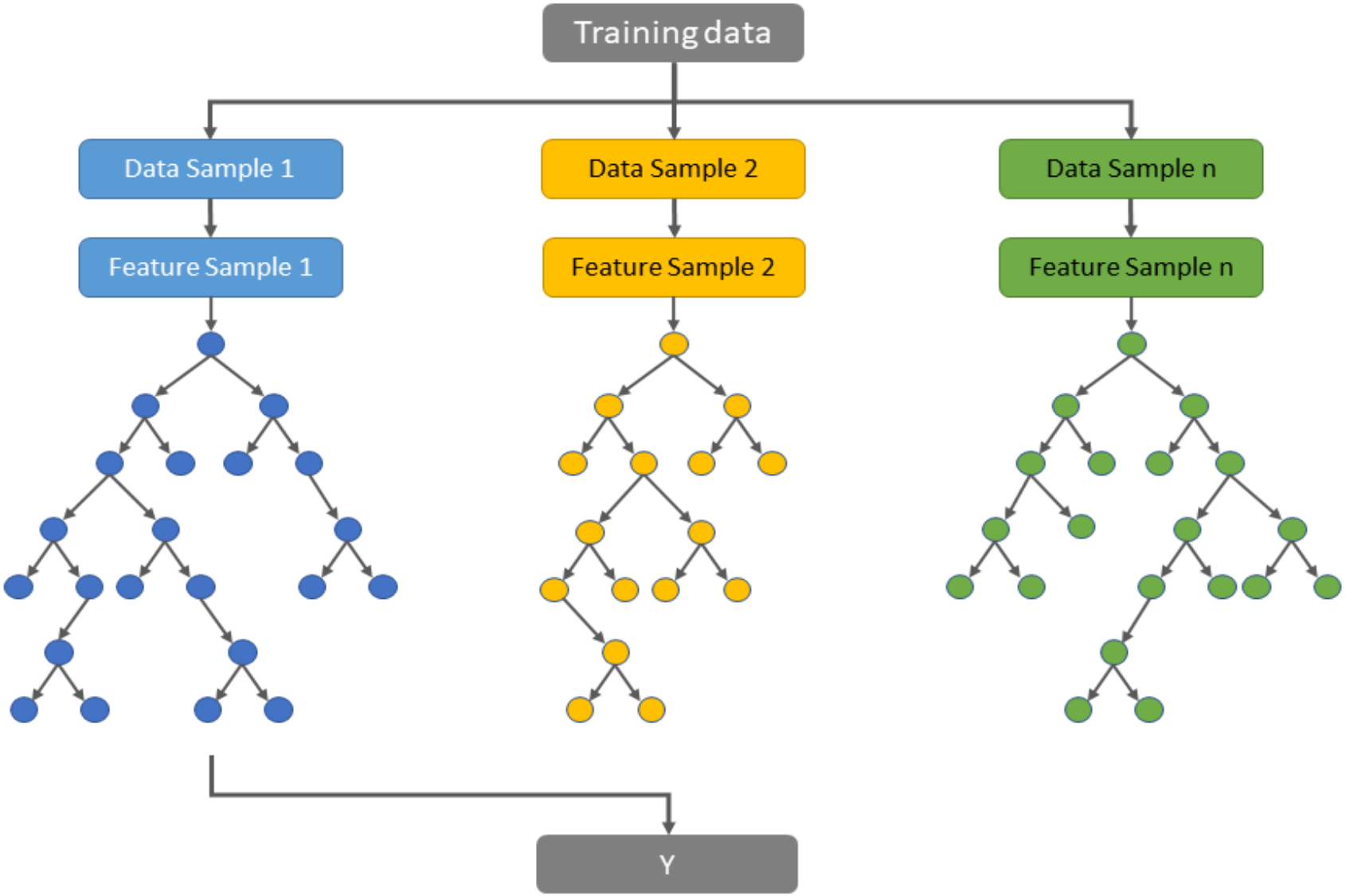
Random Forest



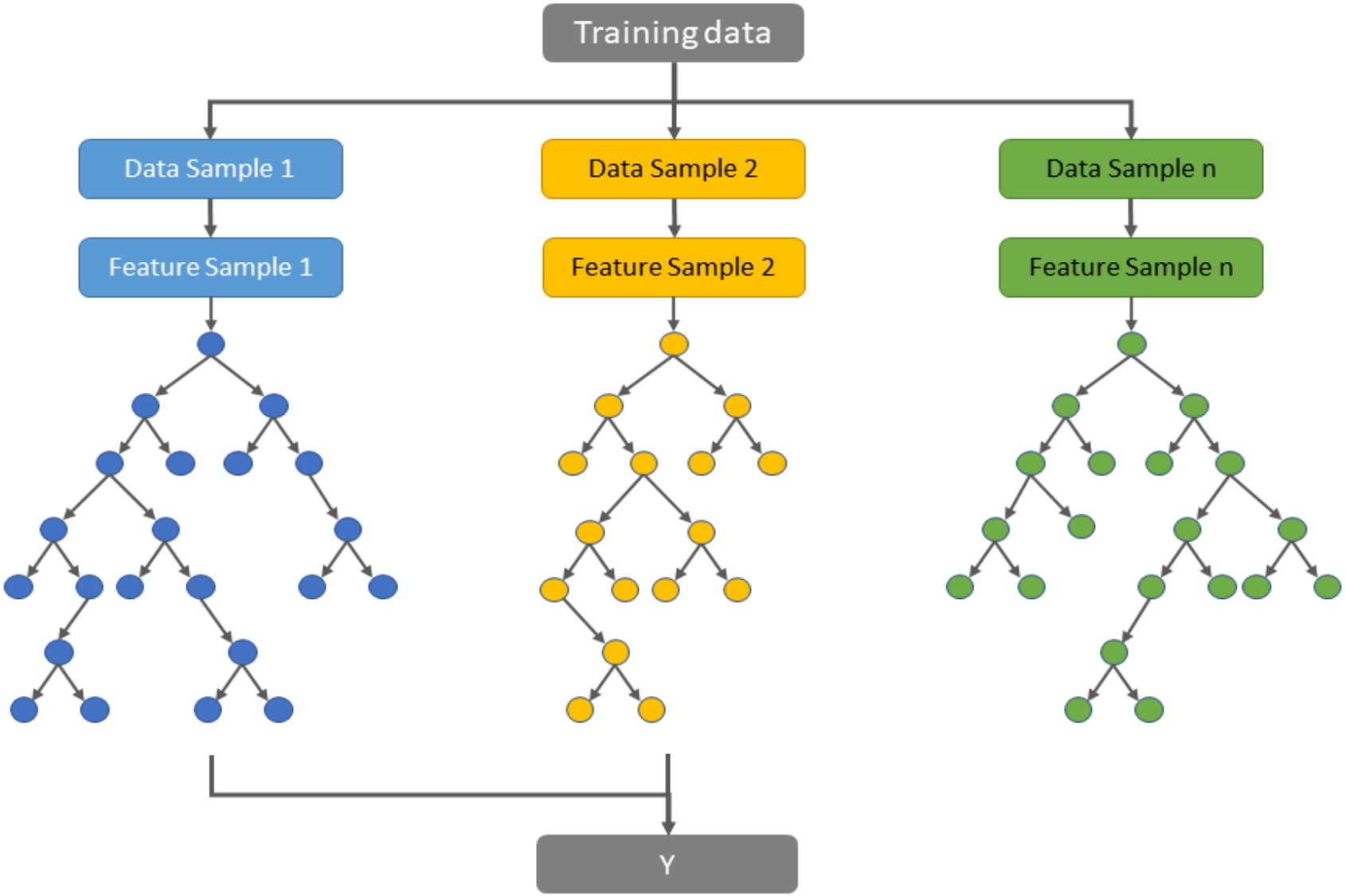
Random Forest



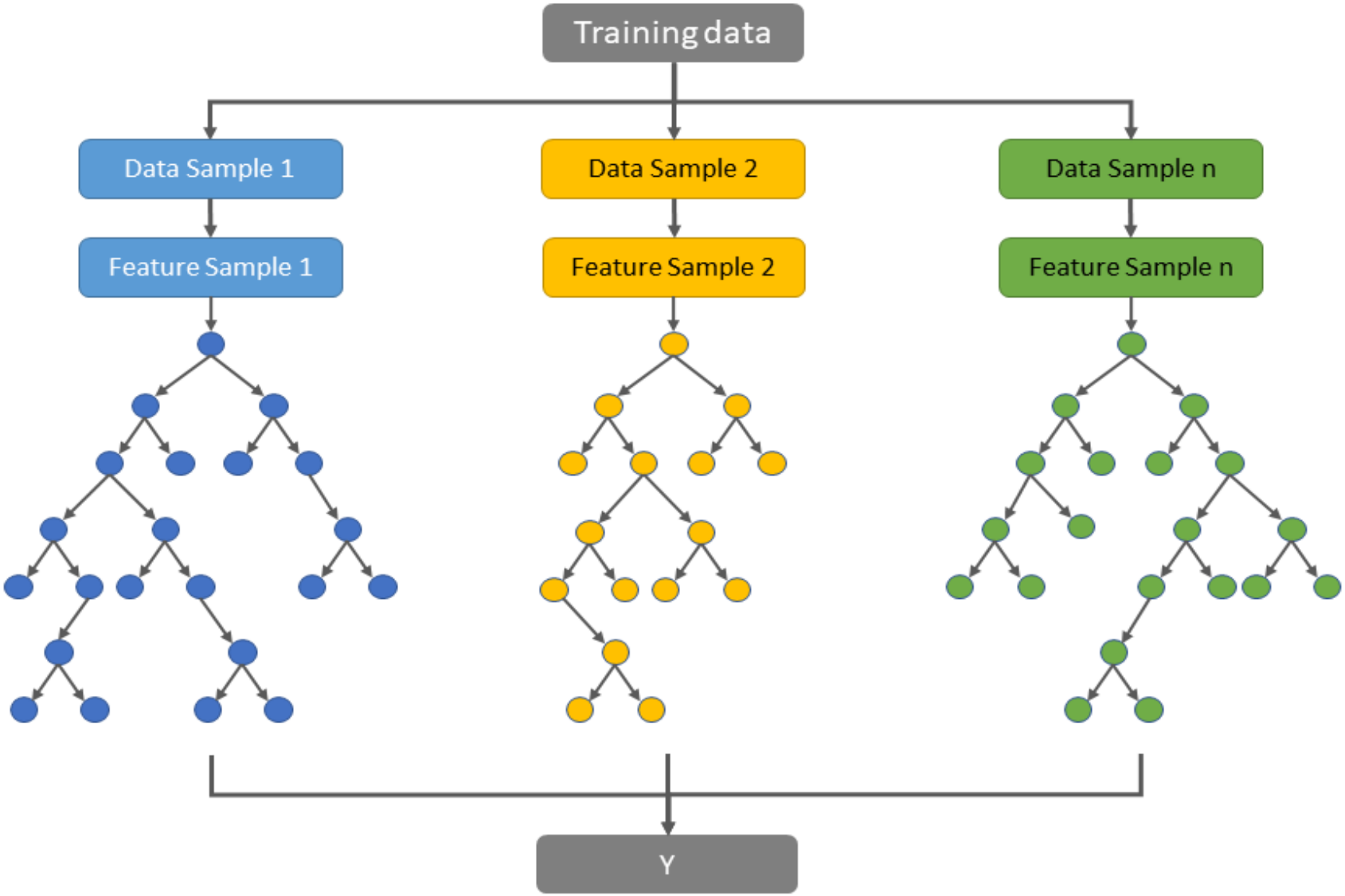
Random Forest



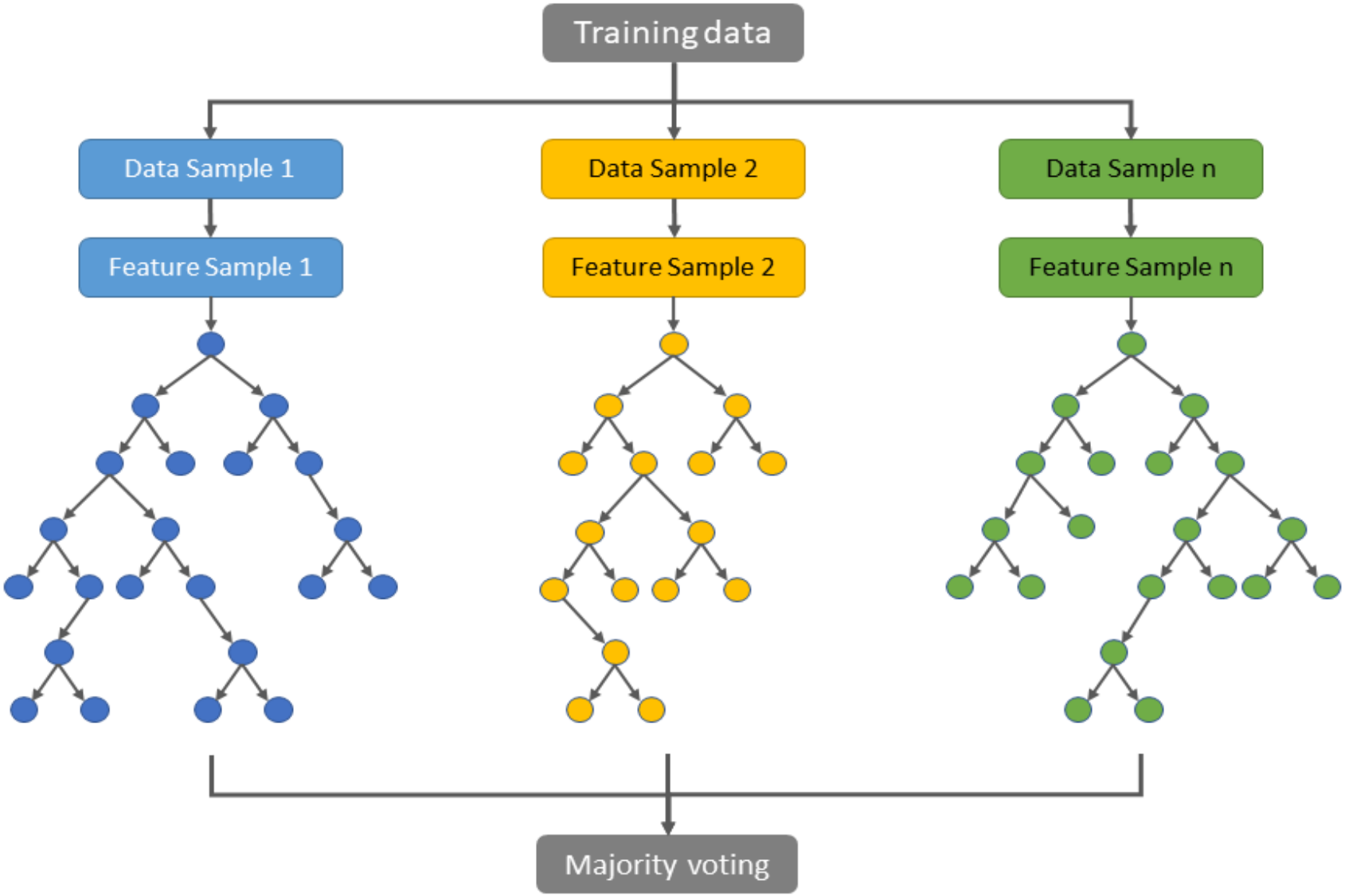
Random Forest



Random Forest



Random Forest



PLOT PERFORMANCE

Plot ML performance

Let's go to R

REPEAT FOR ACTIONS

Type of Terrorist Action

Let's go to R

Thanks :)

Javier Osorio

School of Government and Public Policy

University of Arizona