Power of Ensembles

Amit Kapoor

@amitkaps

Bargava Subramanian

@bargava

The Blind Men & the Elephant

“And so these men of Indostan
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong,
Though each was partly in the right,
And all were in the wrong.”

— *John Godfrey Saxe*

Model Abstraction

"All models are wrong, but some are useful"

— *George Box*

Building Many Models

"All models are wrong, ~~but~~ some are useful, **and their combination may be better**"

Ensembles

(noun) /änˈsämbəl/
a group of items viewed as a whole rather than individually.

Machine Learning Process

Frame: Problem definition

Acquire: Data ingestion

Refine: Data wrangling

Transform: Feature creation

Explore: Feature selection

Model: Model selection

Insight: Solution communication

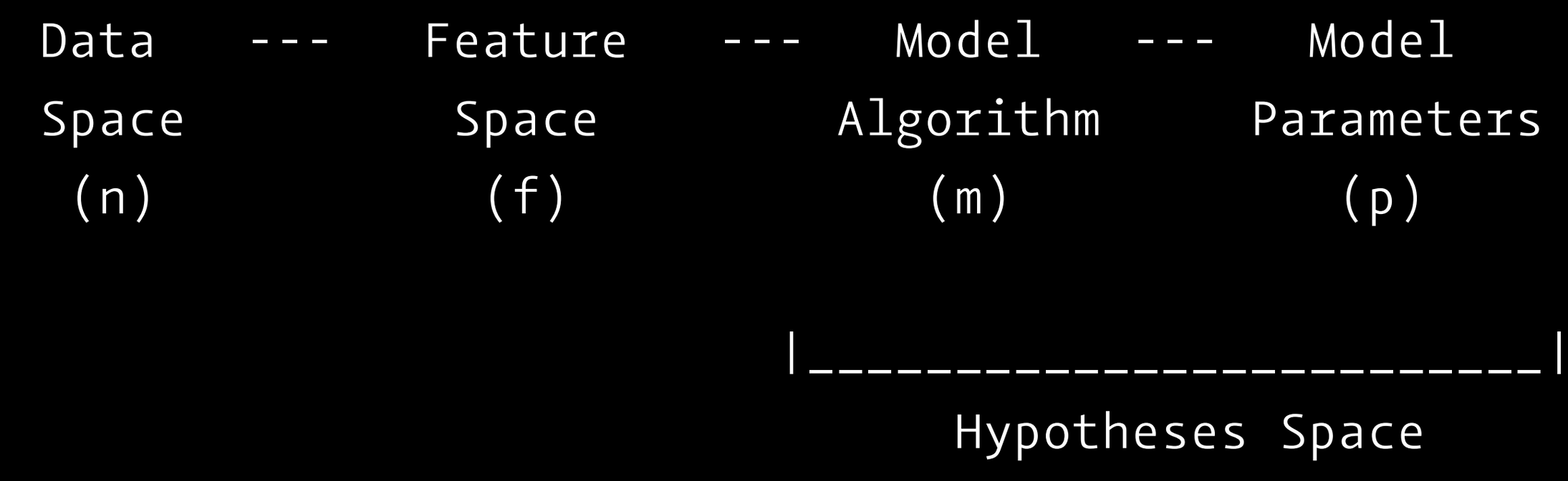
Machine Learning Model Pipeline

Data Space (n)	---	Feature Space (f)	---	Model Algorithm (m)	---	Model Parameters (p)

Machine Learning Challenge

"...the task of searching through a hypotheses space to find a suitable hypothesis that will make good predictions for a particular problem"

Hypotheses Space



Search Hypotheses Space

The key question is how do we efficiently search this *Hypotheses Space*?

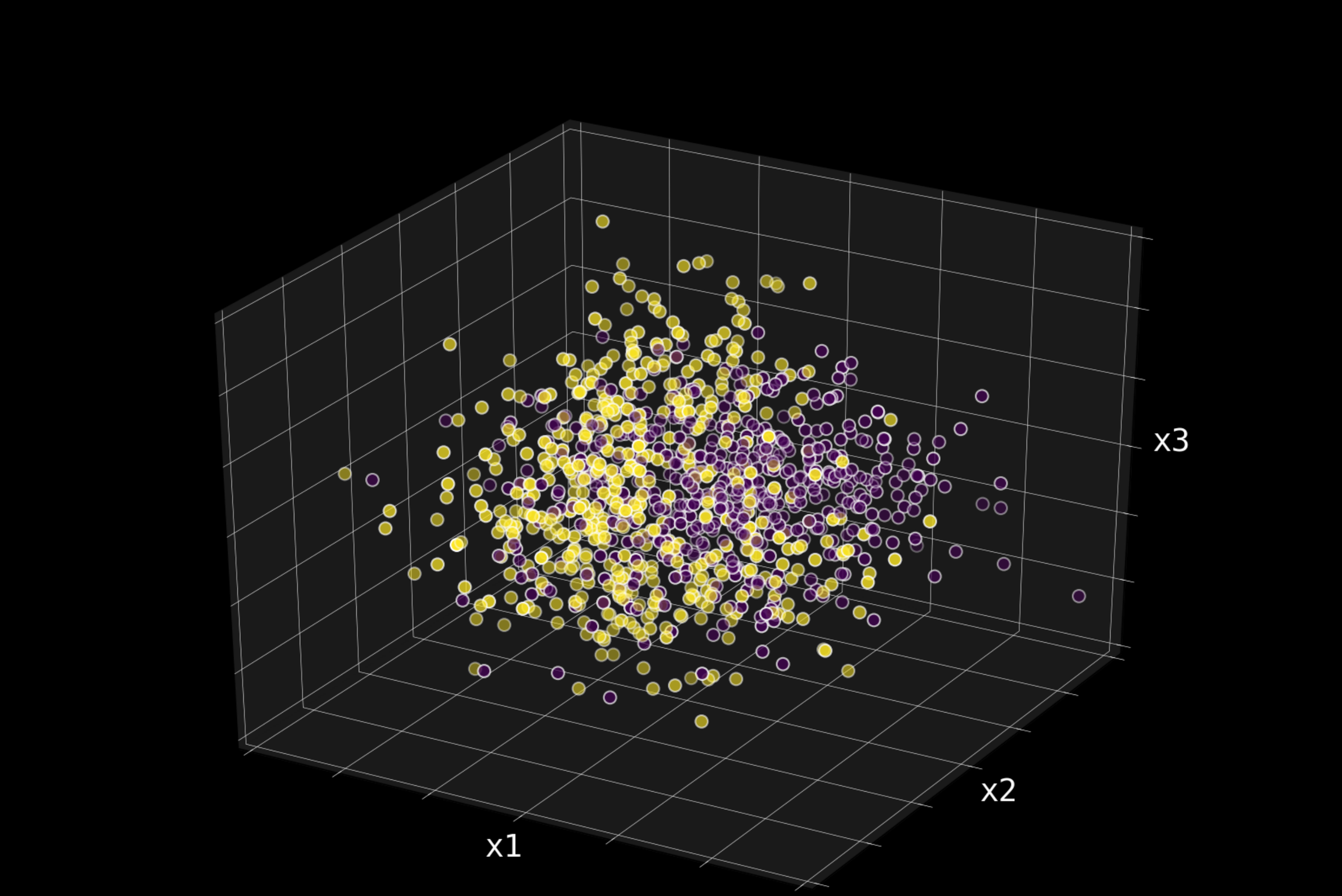
Illustrative Example

Binary Classification Problem

Classes: c = 2

Features: f = x1, x2, x3

Observations: n = 1,000

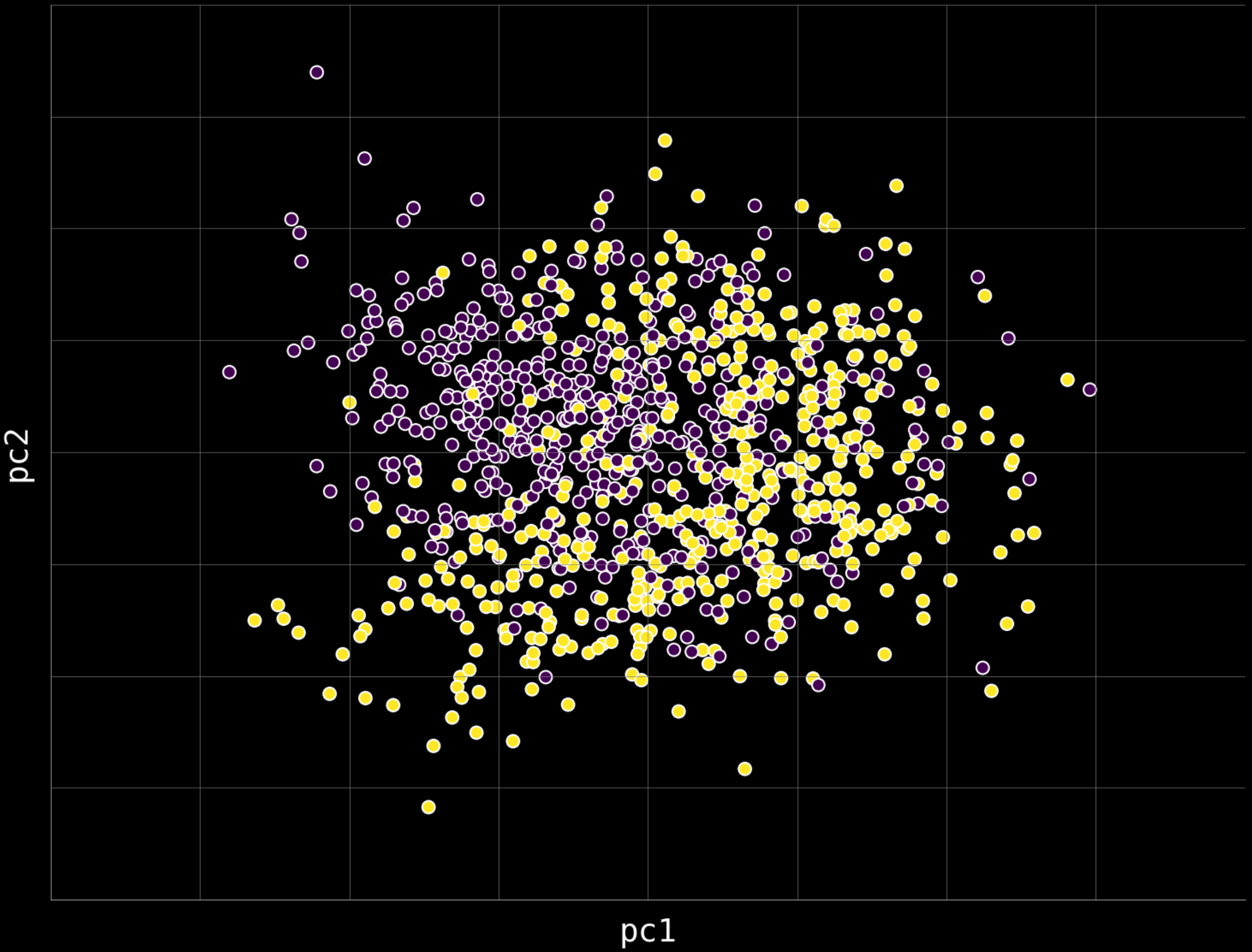

x3
x2
x1

Dimensionality Reduction

"Easier to visualise the data space"

Principal Component Analysis

Dimensions = 2 --> (pc1, pc2)


pc2
pc1

Approach to Build Models

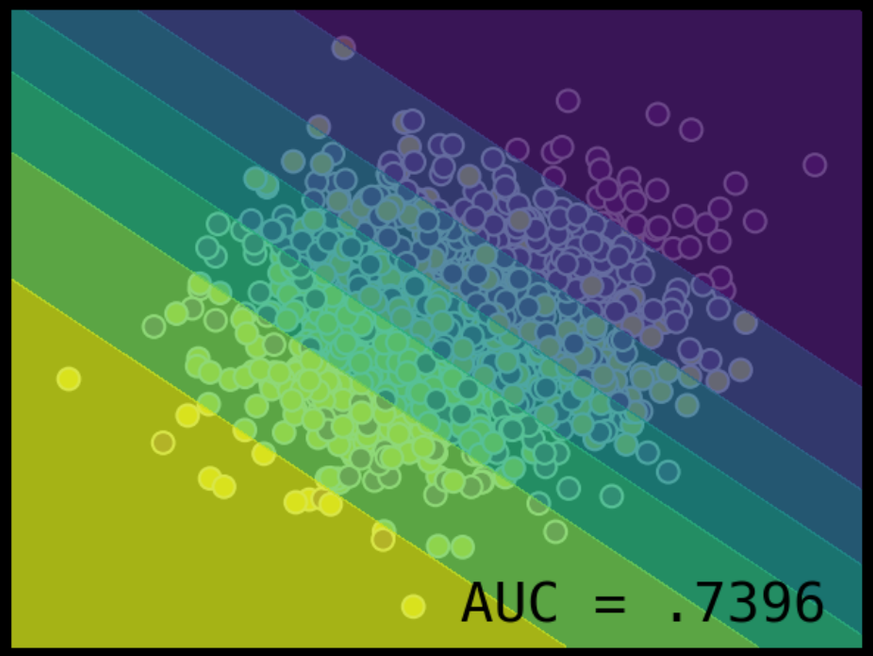
---

1. Train a **model**
2. Predict the **probabilities** for each class
3. Score the model on **AUC** via **5-Fold Cross-Validation**
4. Show the **Decision Space**

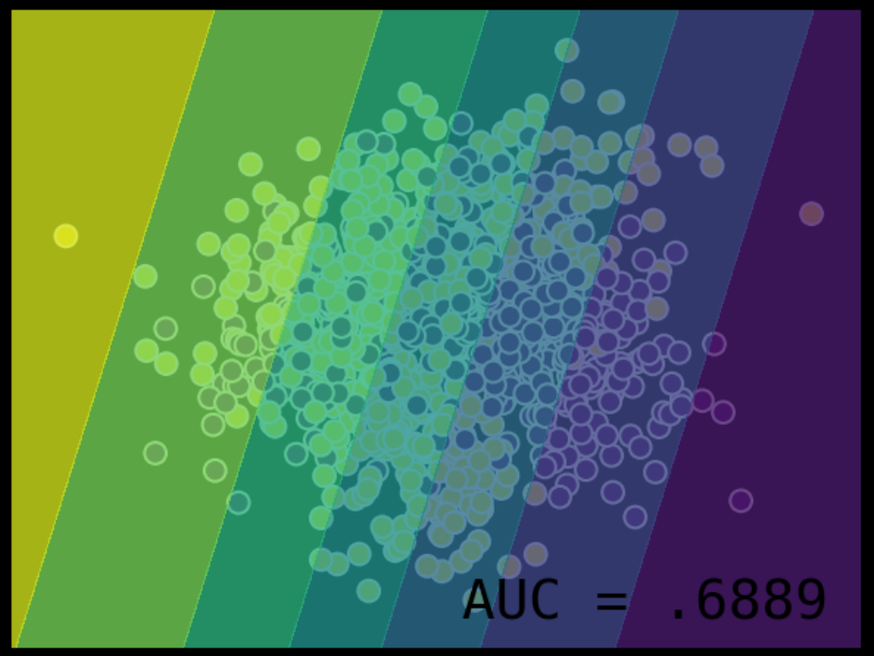
Simple Model with Different Features

Data Space (n)	Feature Space (f)	Model Algorithm (m)	Model Parameters (p)
100%	x1,x2	Logistic	C=1
100%	x2,x3	Logistic	C=1
100%	x1,x3	Logistic	C=1
100%	pc1,pc2	Logistic	C=1

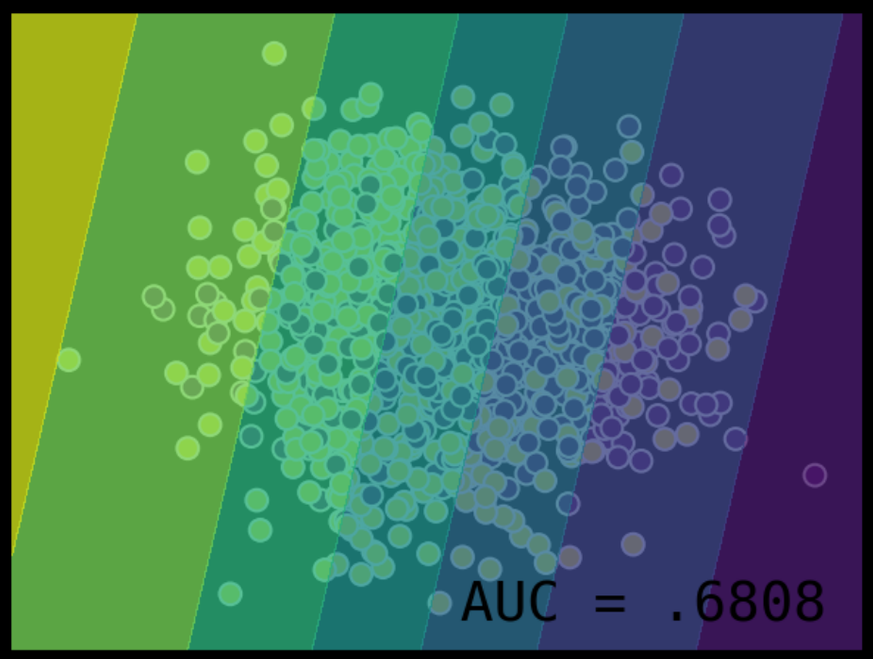
Logistic (x1,x2)



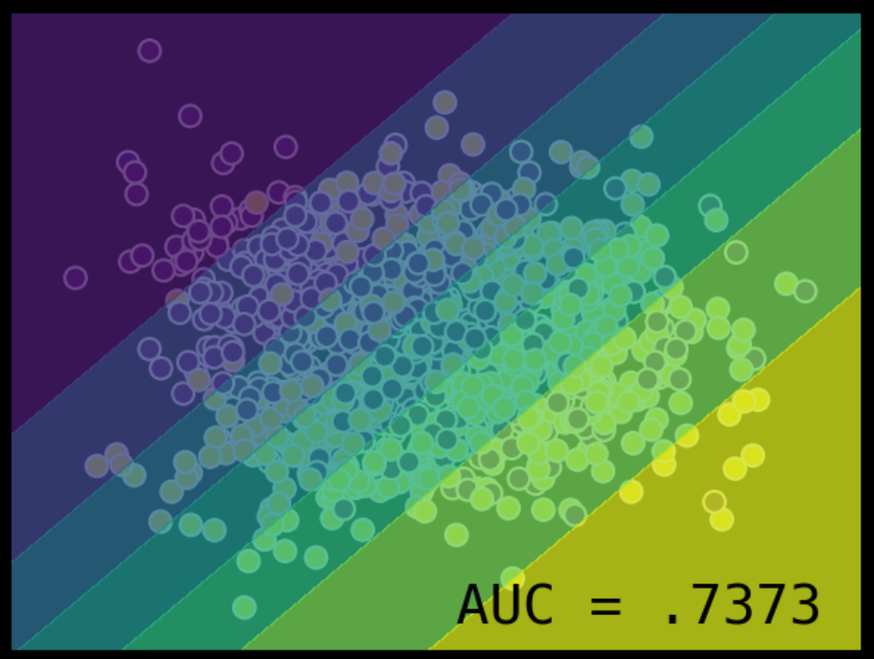
Logistic (x2,x3)



Logistic (x1,x3)



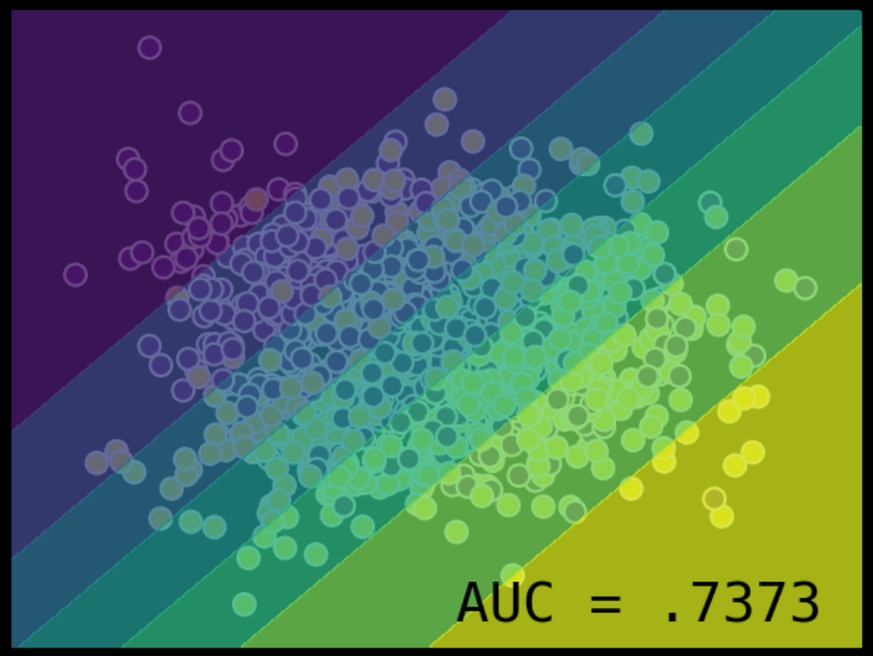
Logistic (pc1,pc2)



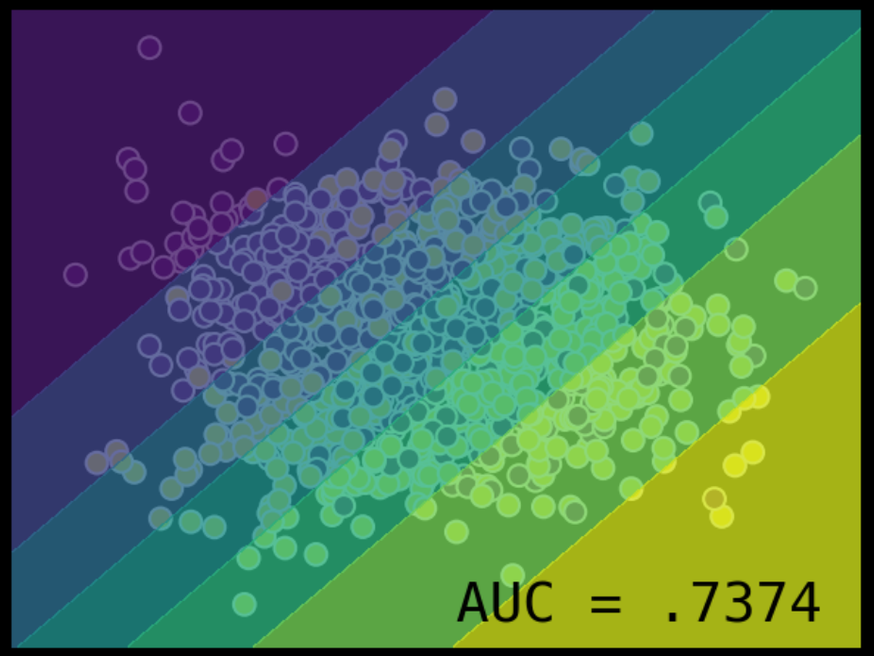
Tune the Model Parameters

Data Space (n)	---	Feature Space (f)	---	Model Algorithm (m)	---	Model Parameters (p)
100%		pc1,pc2		Logistic		C=1
100%		pc1,pc2		Logistic		C=1e-1
100%		pc1,pc2		Logistic		C=1e-3
100%		pc1,pc2		Logistic		C=1e-5

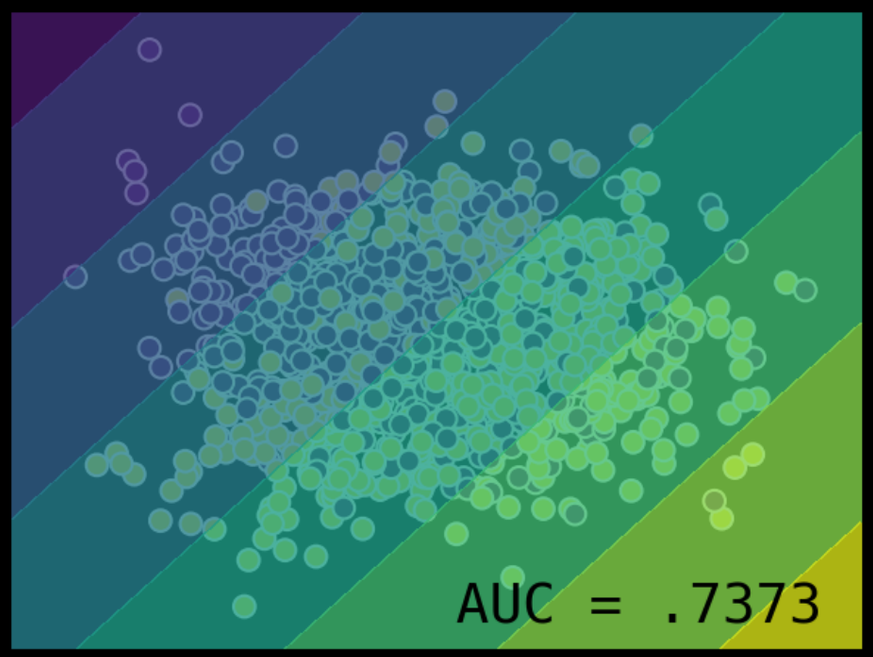
Logistic (C=1)



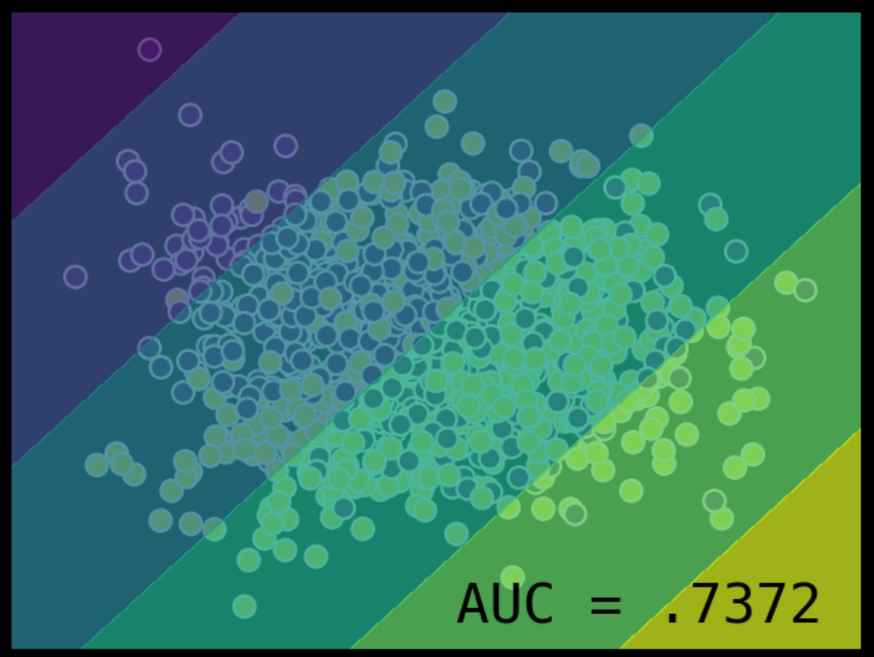
Logistic (C=1e-1)



Logistic (C=1e-3)



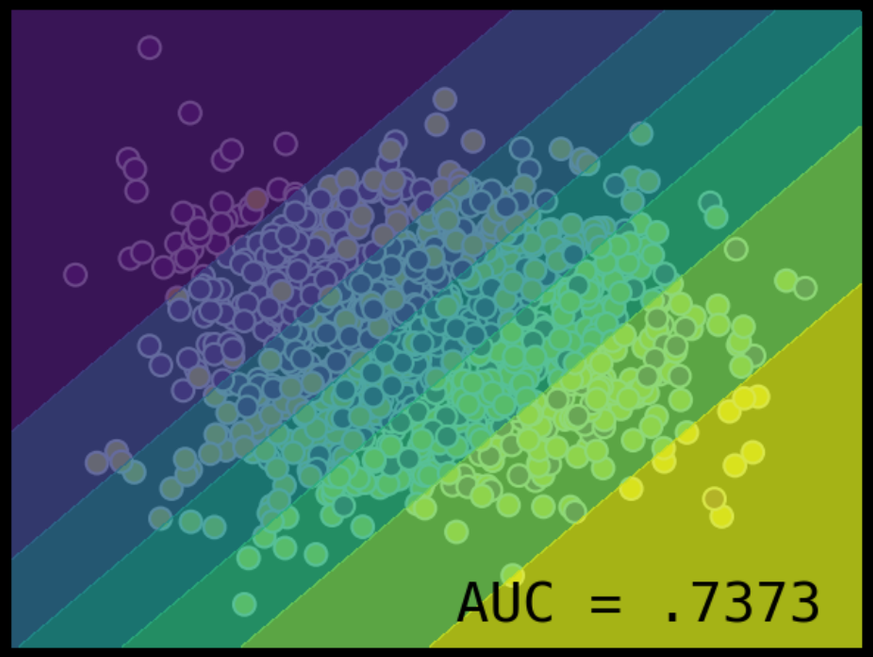
Logistic (C=1e-5)



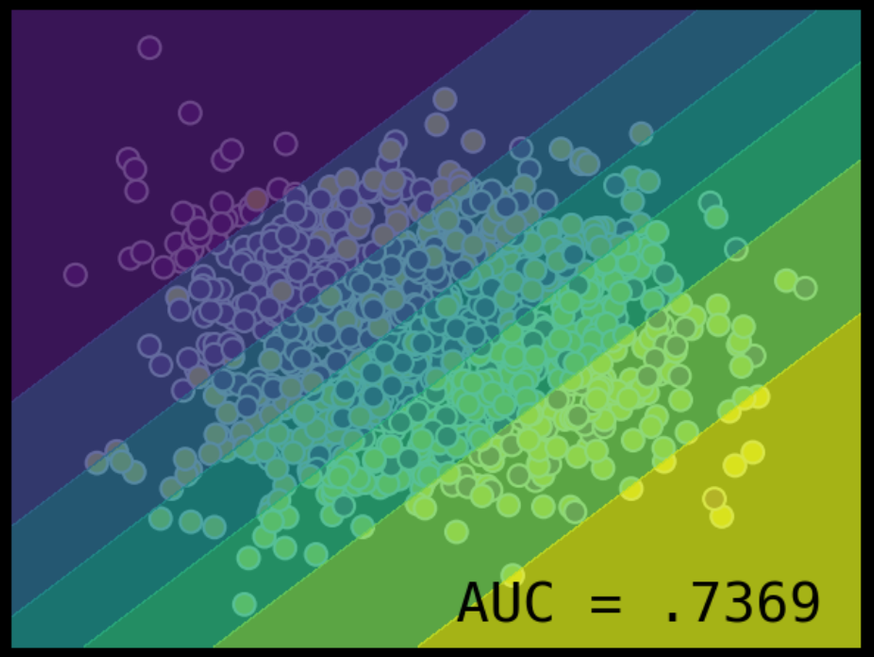
Make Many Simple Models

Data Space (n)	Feature Space (f)	Model Algorithm (m)	Model Parameters (p)
100%	pc1,pc2	Logistic	C=1
100%	pc1,pc2	SVM-Linear	prob=True
100%	pc1,pc2	Decision Tree	d=full
100%	pc1,pc2	KNN	nn=3

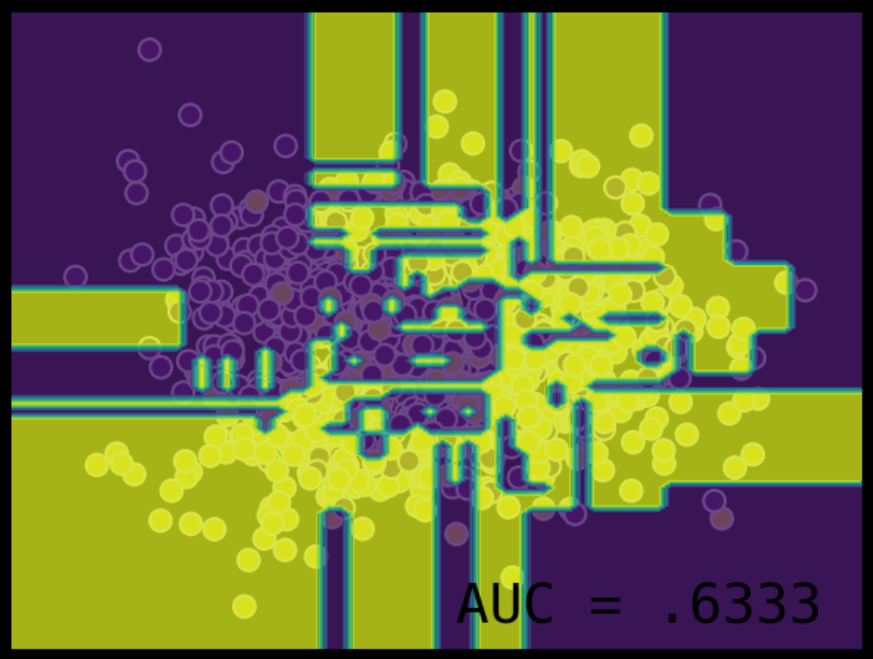
Logistic



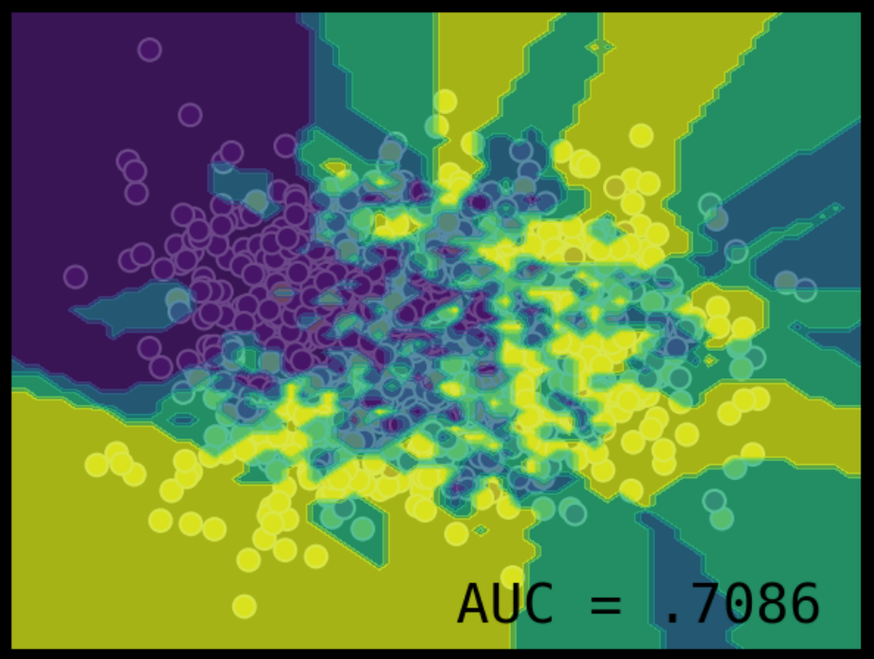
SVM-Linear



Decision Tree



K Nearest Neighbour



Hypotheses Search Approach

Exhaustive search is impossible

Compute as a proxy for human IQ

Clever algorithmic way to search the solution space

Ensemble Thought Process

"The goal of ensemble methods is to combine the predictions of several base estimators built with (one or multiple) model algorithms in order to improve generalisation and robustness over a single estimator."

Ensemble Approach

Different **training sets**

Different **feature sampling**

Different **algorithms**

Different **(hyper) parameters**

Clever **aggregation**

Ensemble Methods

---

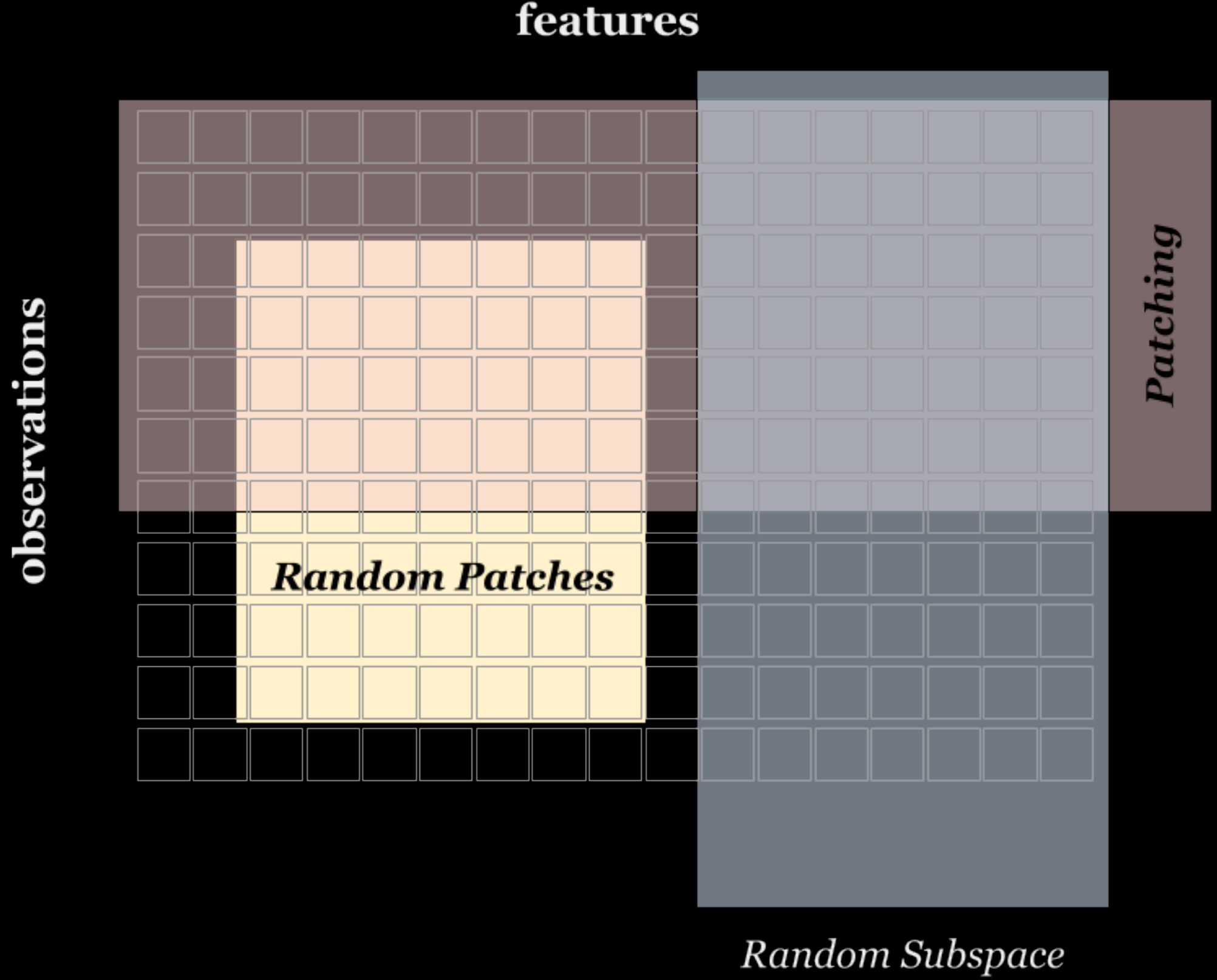
[1] Averaging
[2] Boosting
[3] Voting
[4] Stacking

[1] Averaging: Concept

"Build several estimators independently and then average their predictions to reduce model variance"

To ensemble several good models to produce a less variance model.

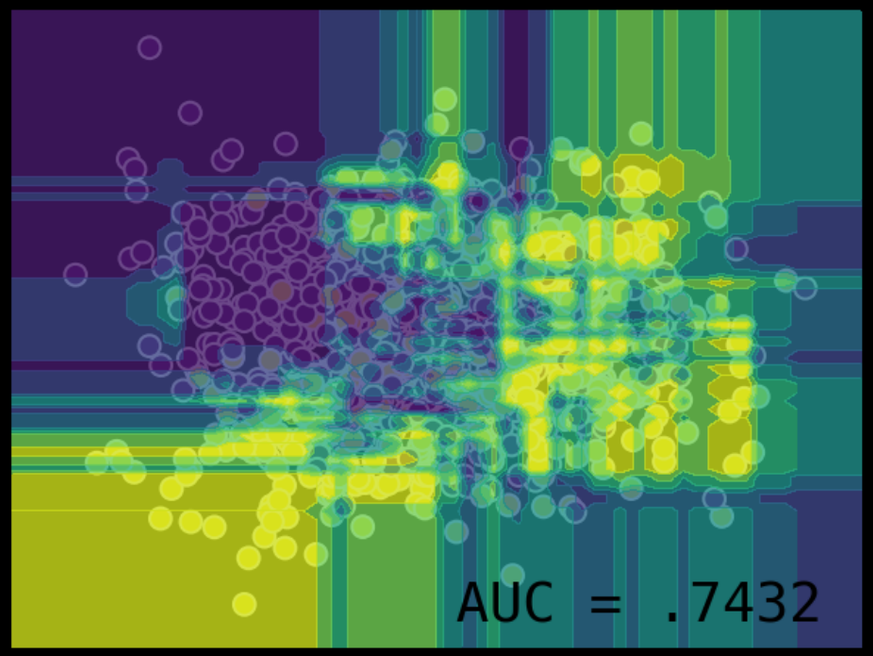
[1] Averaging: Approaches



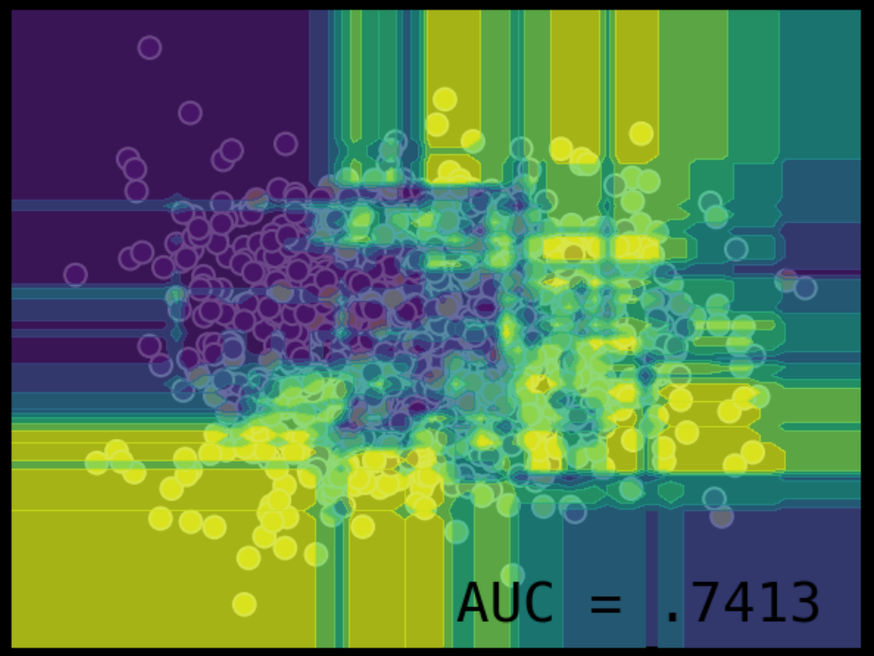
[1] Averaging: Simple Models

Data Space (n)	Feature Space (f)	Model Algorithm (m)	Model Parameters (p)
R[50%]	pc1,pc2	Decision Tree	n_est=10
R[50%,r]	pc1,pc2	Decision Tree	n_est=10
100%	R[50%]	Decision Tree	n_est=10
R[50%]	R[50%]	Decision Tree	n_est=10

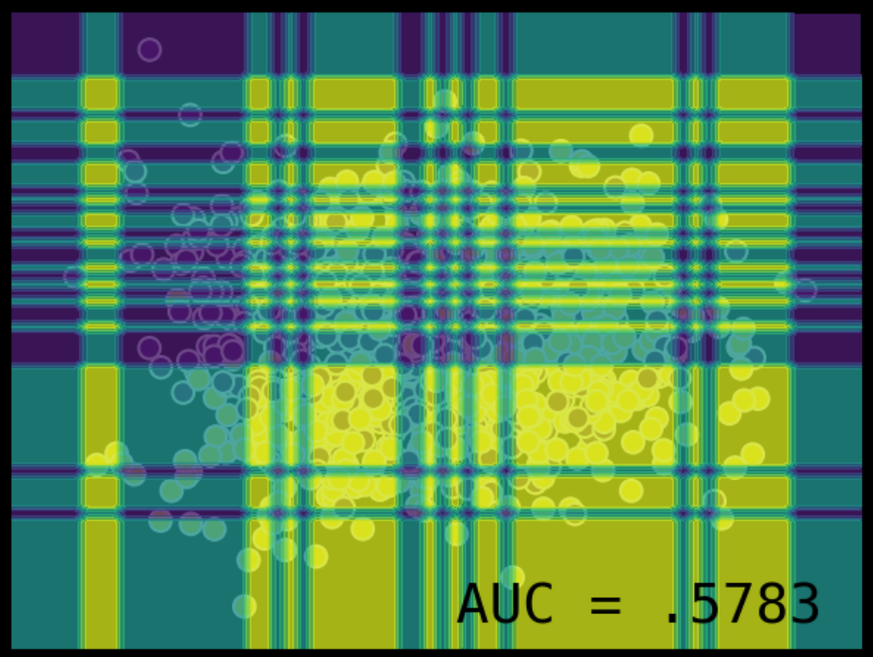
Patching



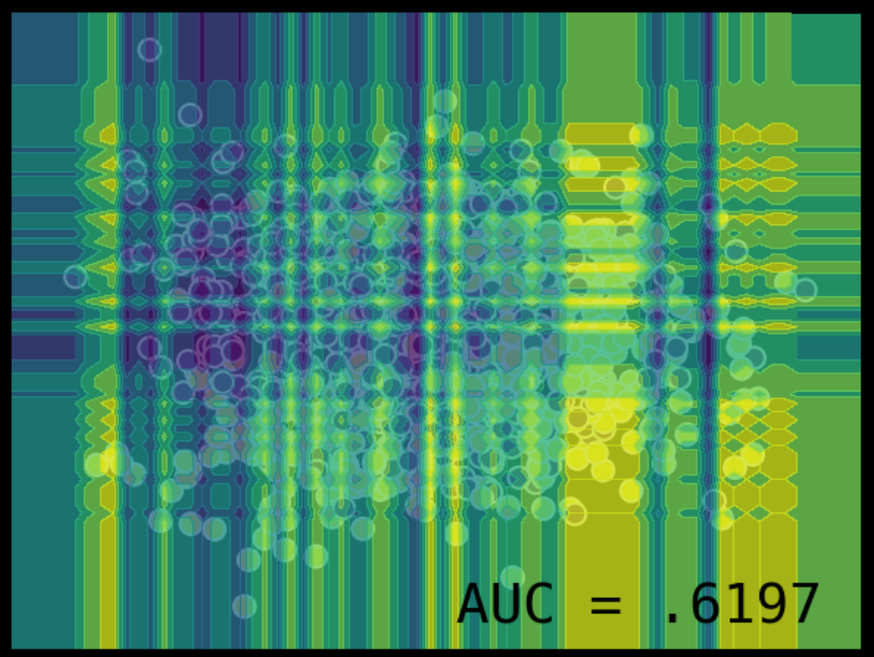
Bagging



Random Subspace



Random Patches



[1] Averaging: Extended Models

Use *perturb-and-combine techniques*

Bagging: Bootstrap Aggregation

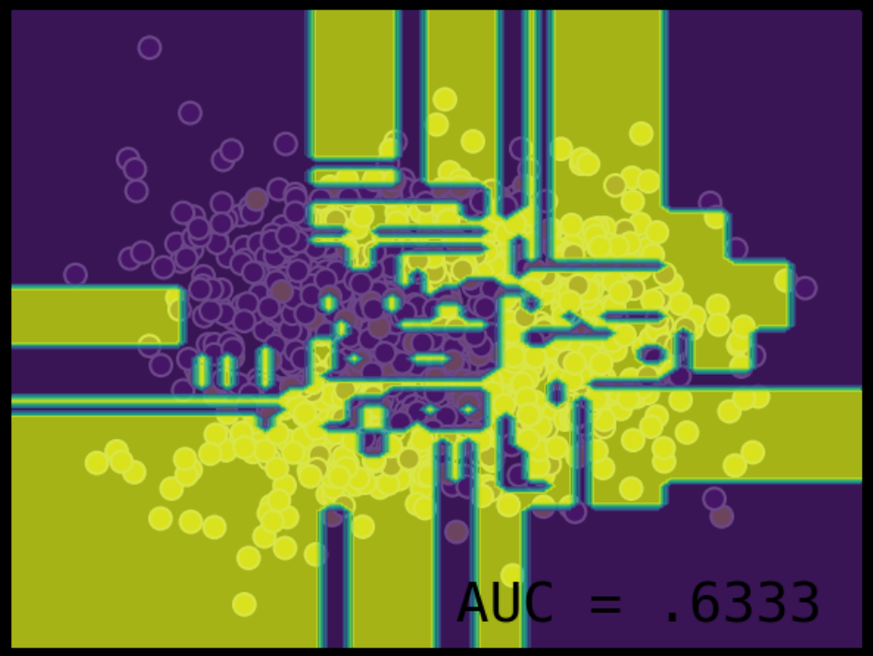
Random Forest: Best split amongst random features

Extremely Randomised: Random threshold for split

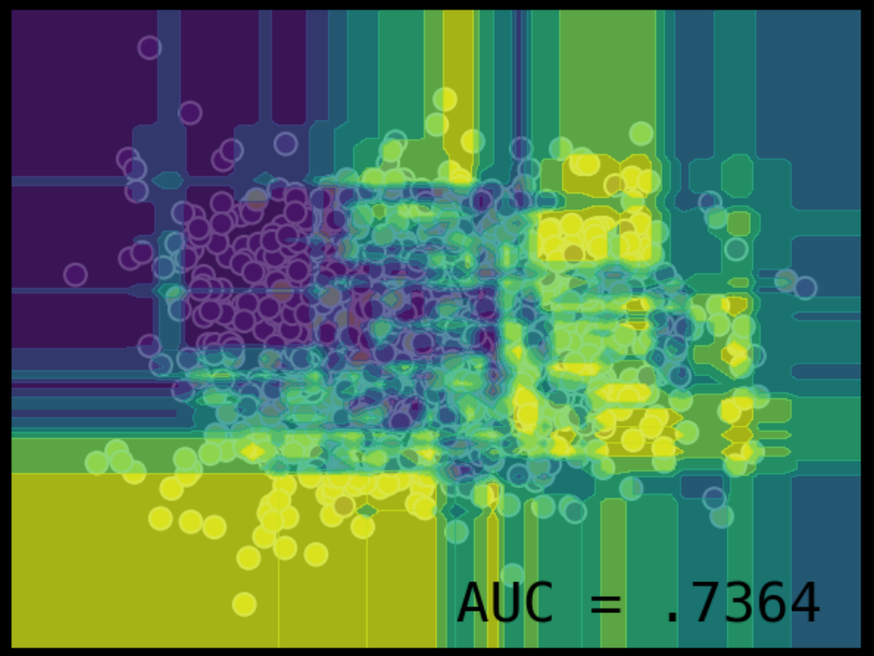
[1] Averaging: Extended Models

Data Space (n)	---	Feature Space (f)	---	Model Algorithm (m)	---	Model Parameters (p)
100%		pc1,pc2		Decision Tree		d=full
R[50%,r]		pc1,pc2		Decision Tree		n_est=10
R[50%,r]		R[Split]		Decision Tree		n_est=10
R[50%,r]		R[Split,thresh]		Decision Tree		n_est=10

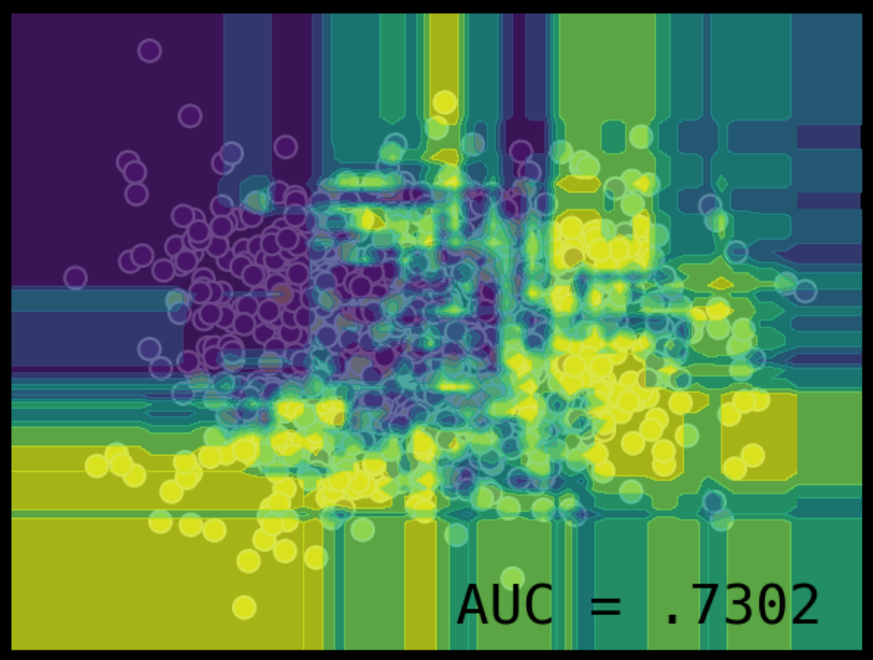
Decision Tree



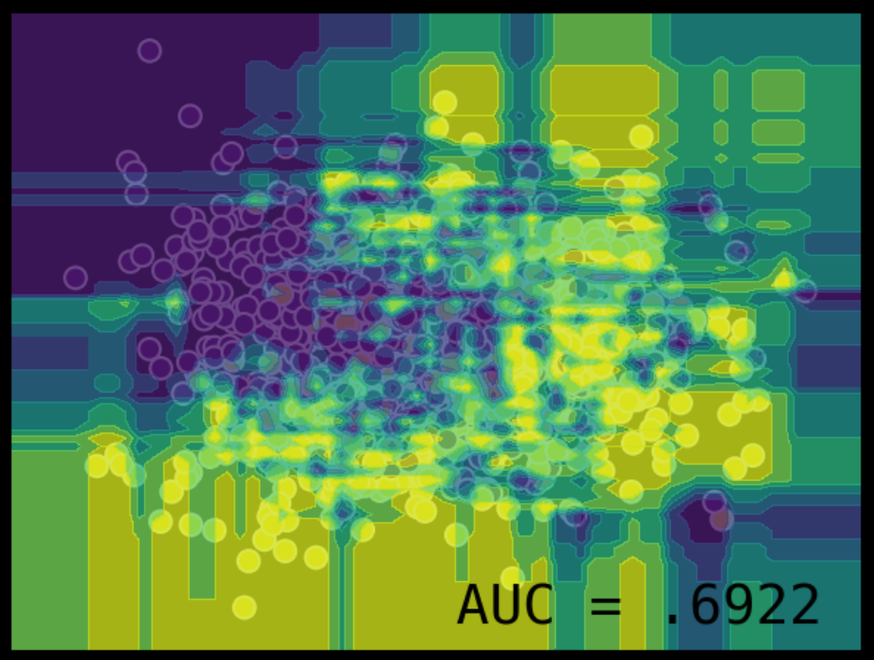
Bagging



Random Forest



Extremely Random

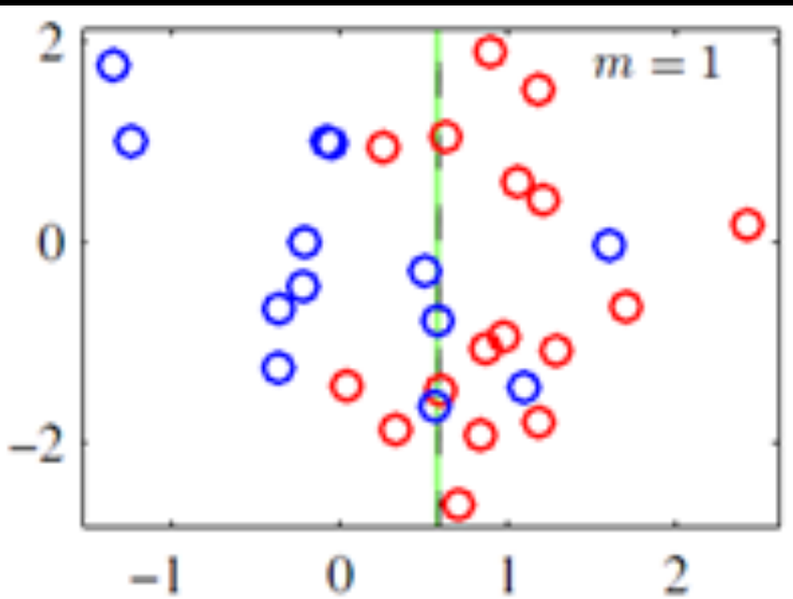


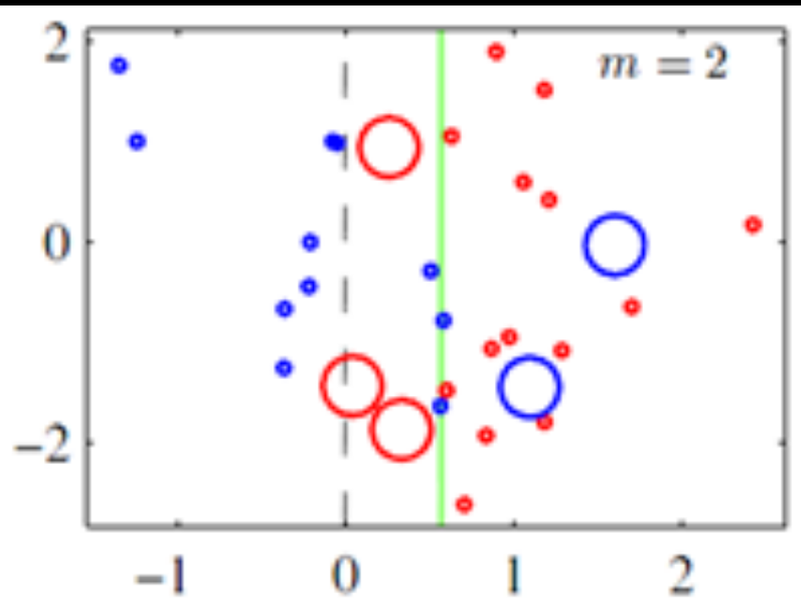
[2] Boosting: Concept

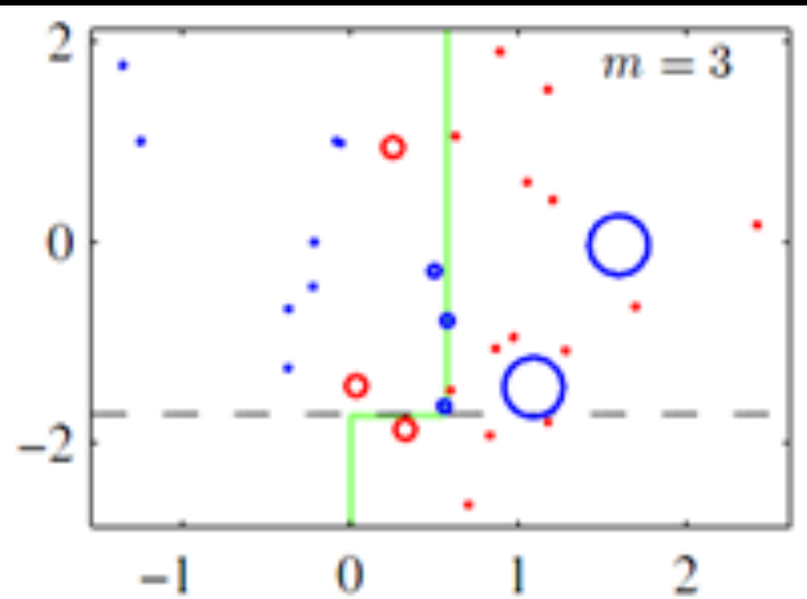
"Build base estimators sequentially and then try to reduce the bias of the combined estimator."

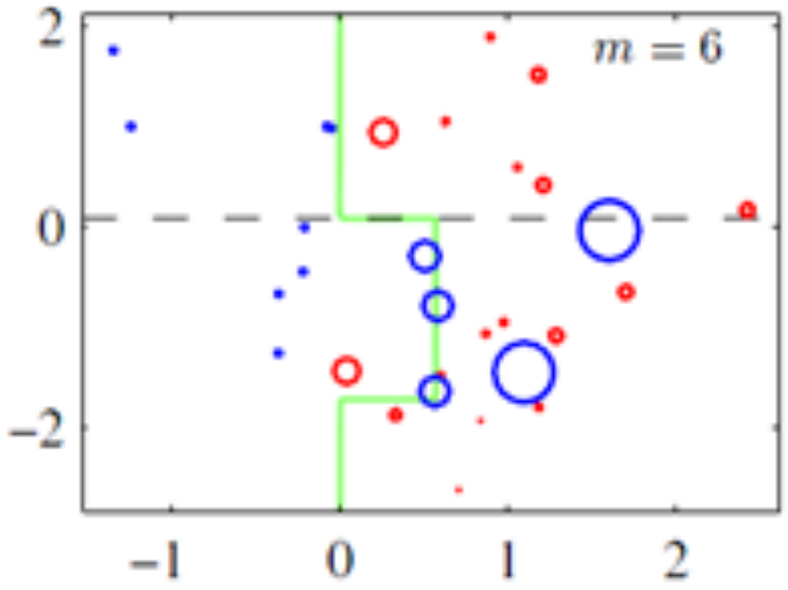
Combine several weak models to produce a powerful ensemble.

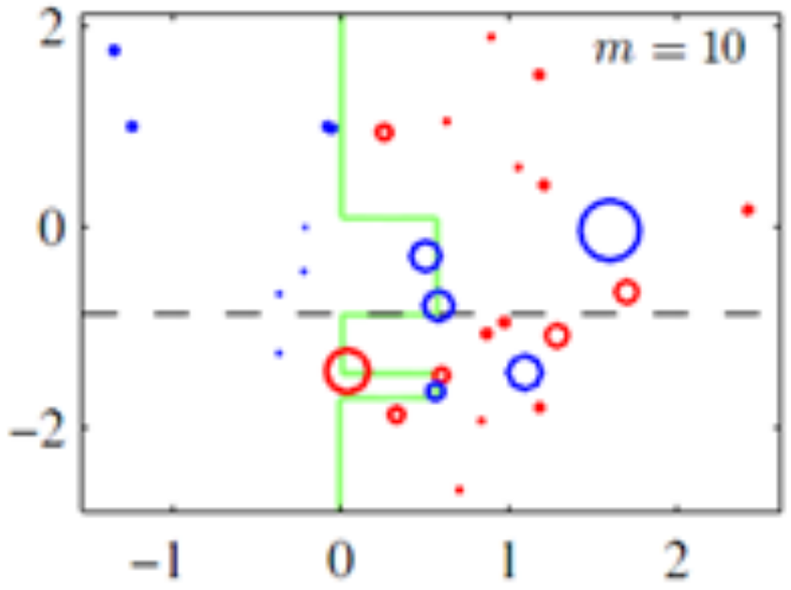
[2] Boosting: Concept

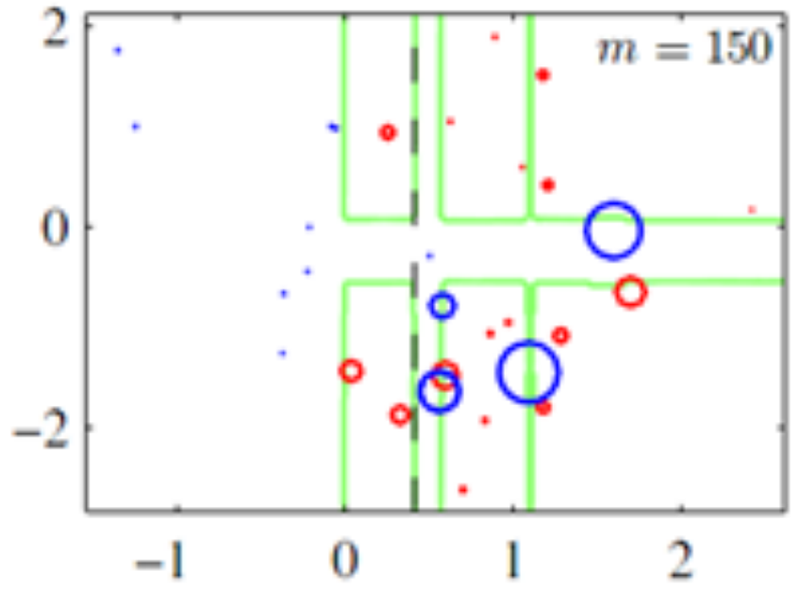










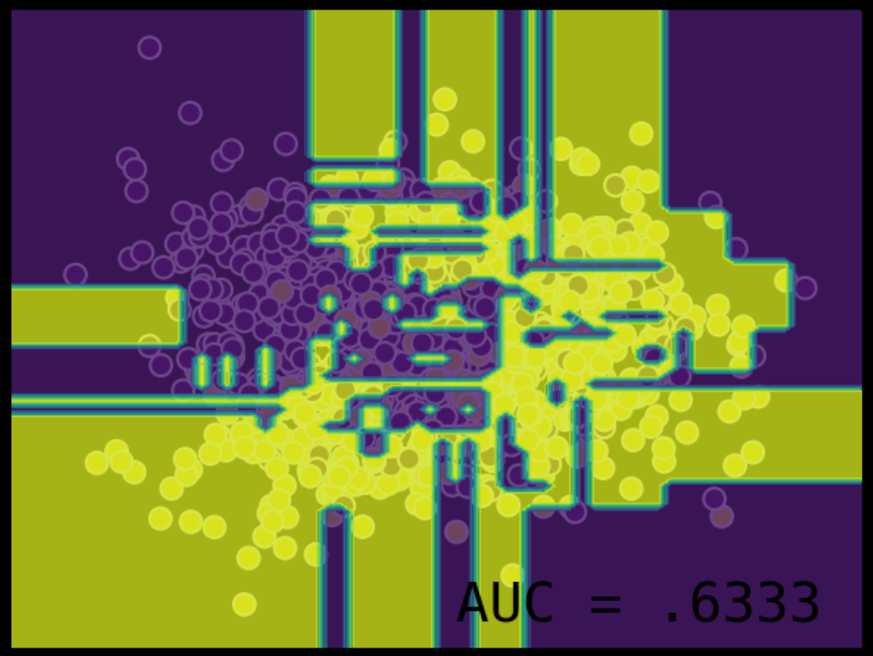


Source: Wang Tan

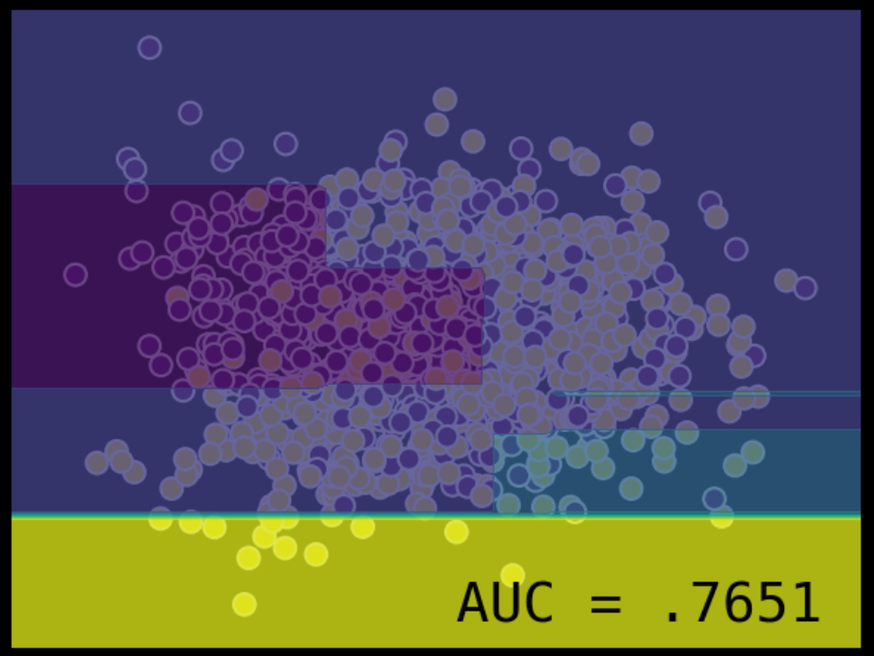
[2] Boosting: Models

Data Space (n)	Feature Space (f)	Model Algorithm (m)	Model Parameters (p)
100%	pc1,pc2	Decision Tree	d=full
100%	pc1,pc2	DTree Boost	n_est=10
100%	pc1,pc2	DTree Boost	n_est=10, d=1
100%	pc1,pc2	DTree Boost	n_est=10, d=2

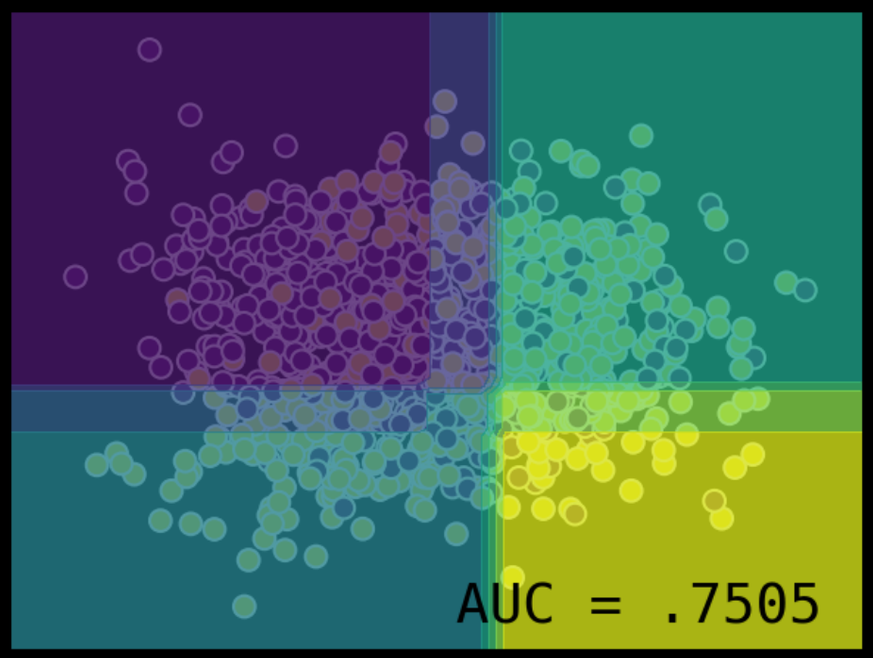
Decision Tree



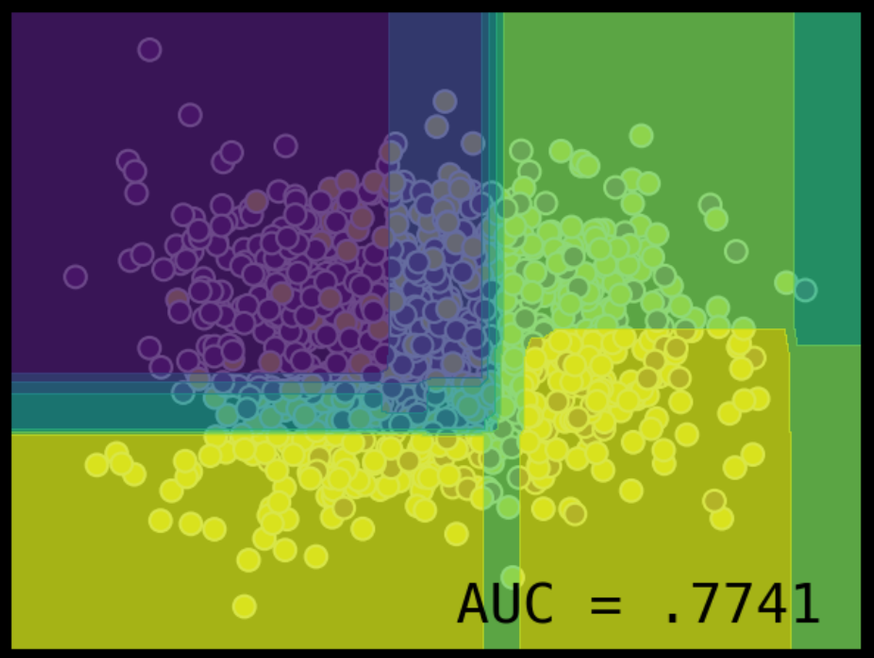
Ada Boost



Gradient Boosting (d=1)



Gradient Boosting (d=2)

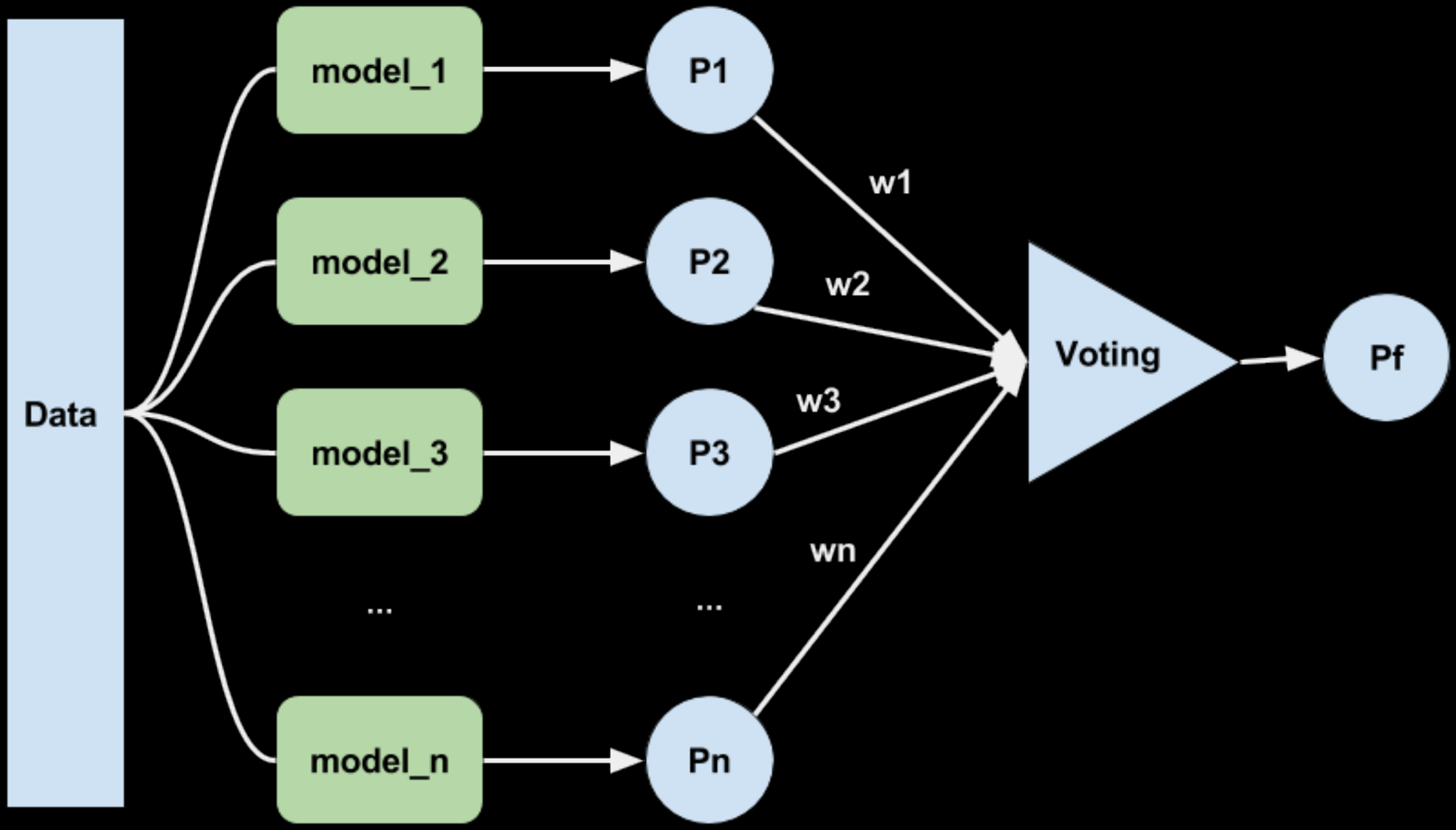


[3] Voting: Concept

"To use conceptually different model algorithms and use majority vote or soft vote (average probabilities) to combine."

To ensemble a set of equally well performing models in order to balance out their individual weaknesses.

[3] Voting: Approach



[3] Voting: Approach

Hard Voting: *the majority (mode) of the class labels predicted*

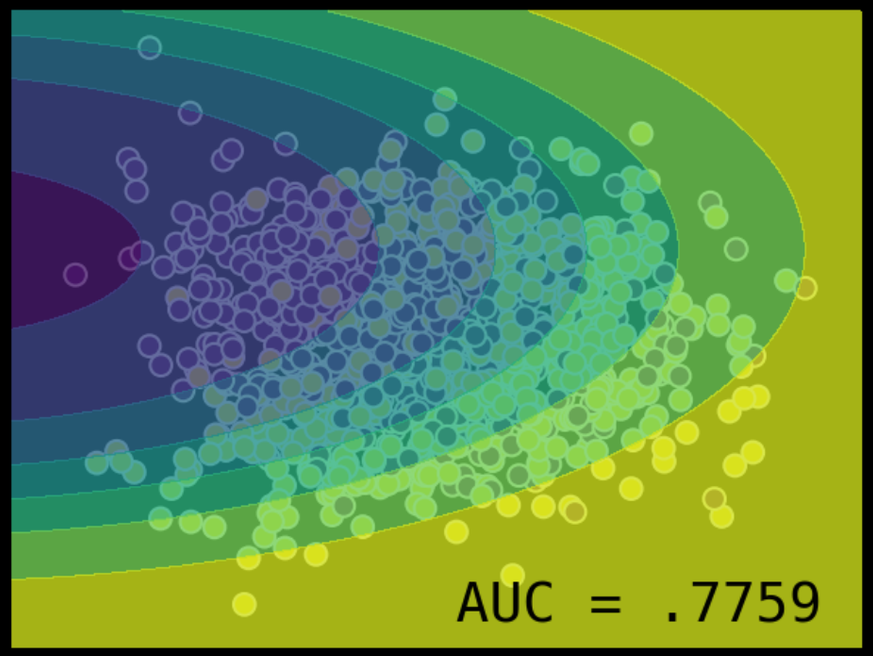
Soft Voting: *the argmax of the sum of predicted probabilities*

Weighted Voting: *the argmax of the weighted sum of predicted probabilities**

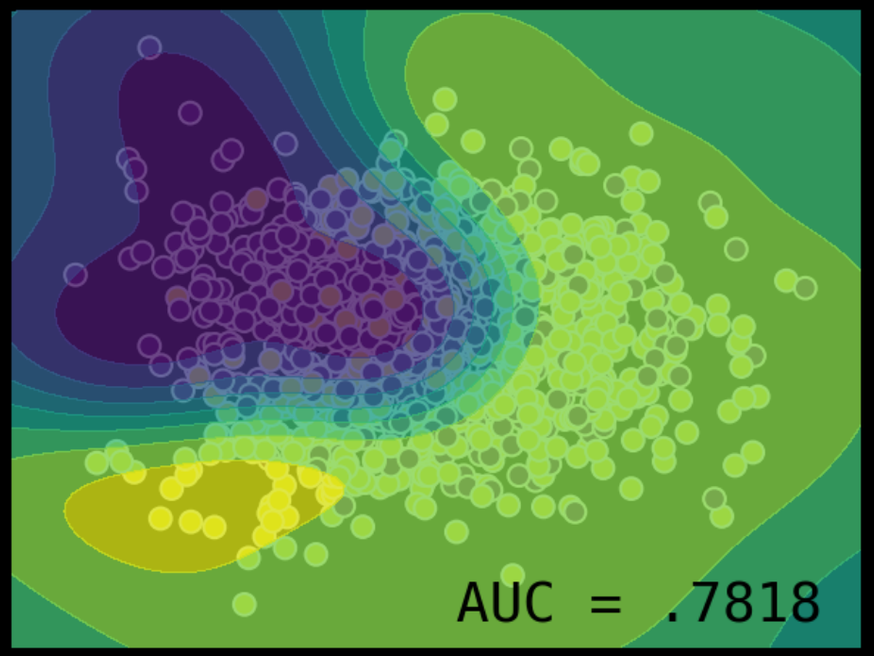
[3] Voting: Models

Data Space (n)	Feature Space (f)	Model Algorithm (m)	Model Parameters (p)
100%	pc1,pc2	Gaussian	-
100%	pc1,pc2	SVM - RBF	gamma=0.5, C=1
100%	pc1,pc2	Gradient Boost	n_est=10, d=2
-----	-------	-------------	-------------
100%	pc1,pc2	Voting	Soft

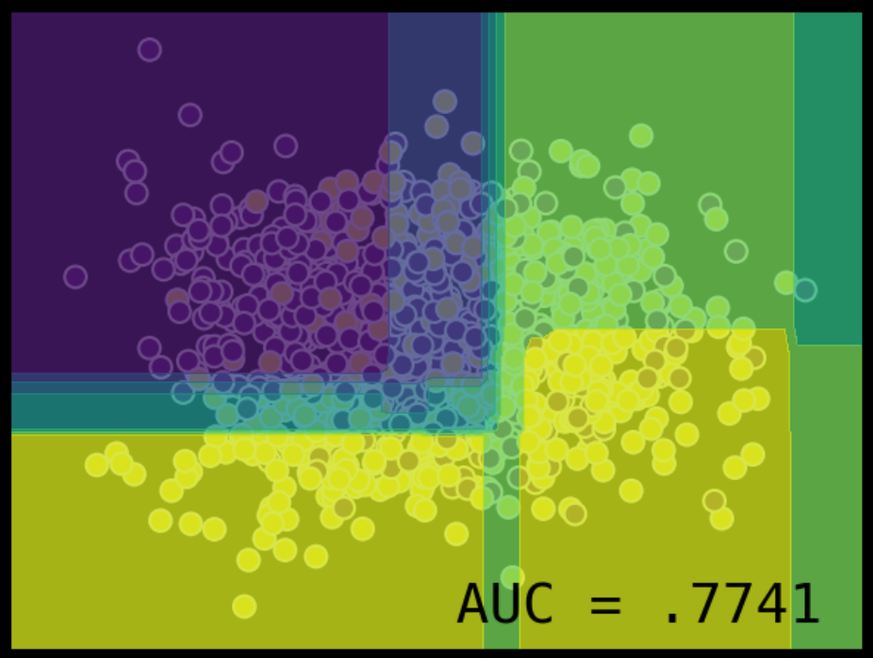
Gaussian



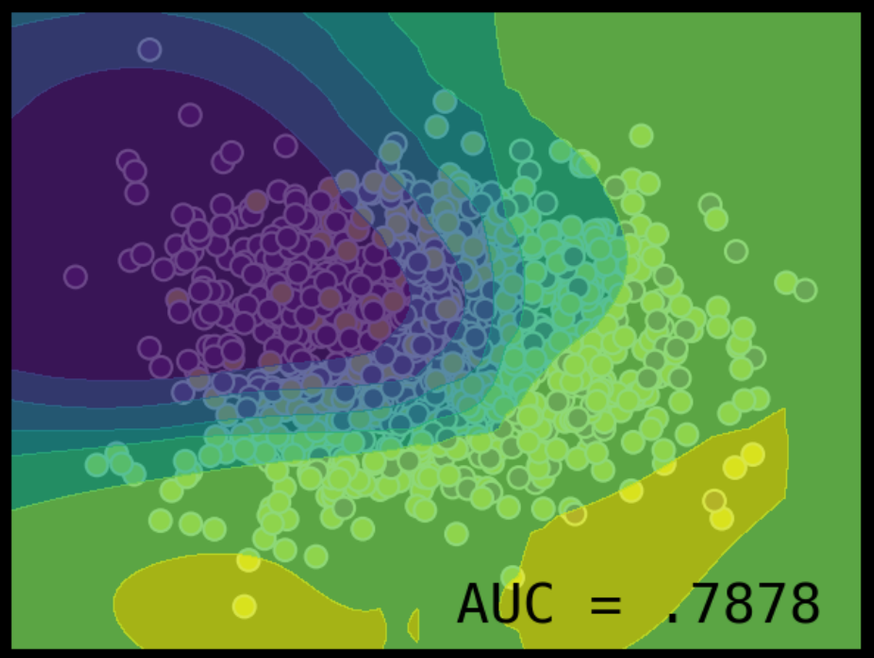
SVM - RBF



Gradient Boosting



Voting

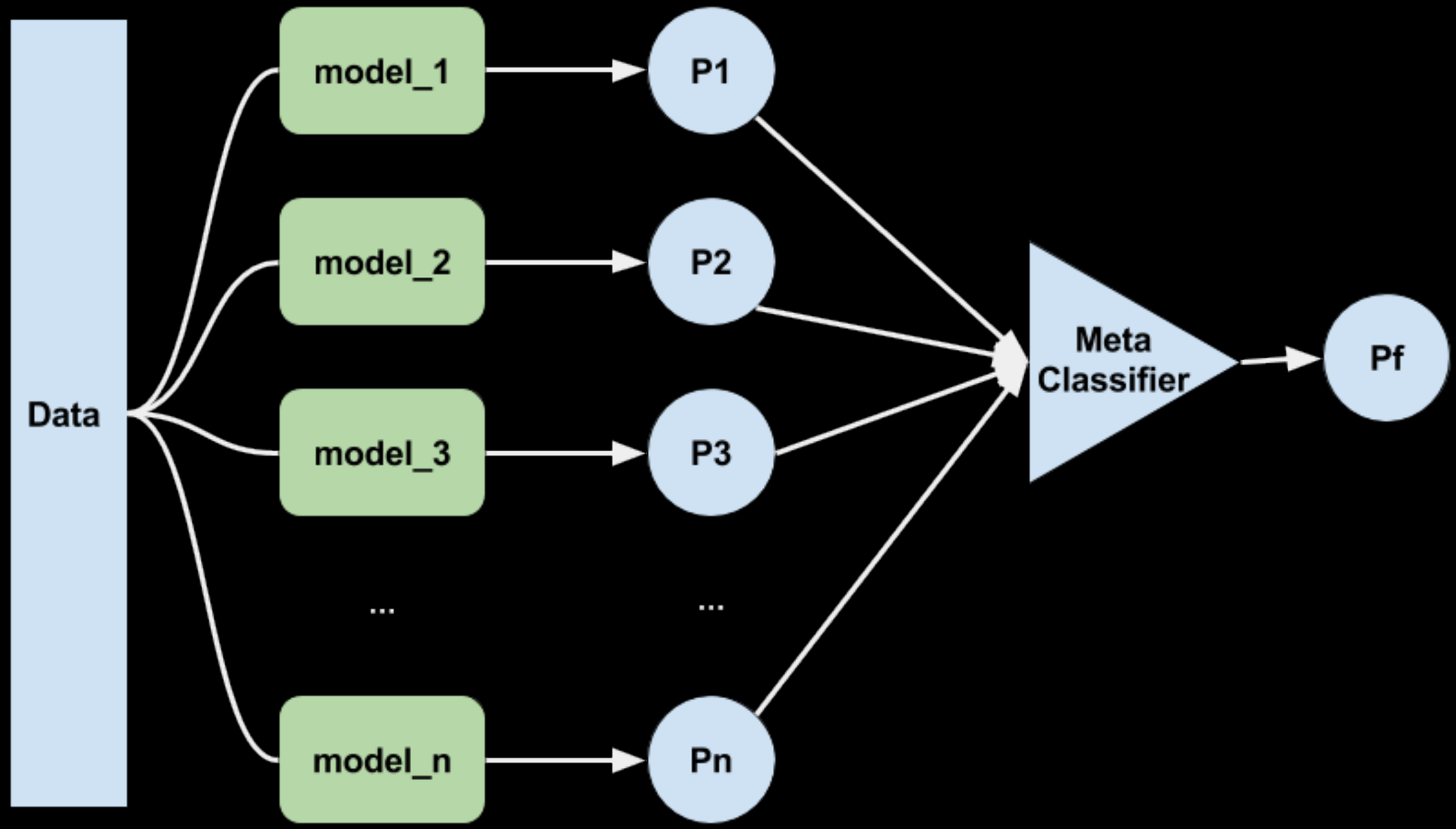


[4] Stacking

"To use output of different model algorithms as feature inputs for a meta classifier."

To ensemble a set of well performing models in order to uncover higher order interaction effects

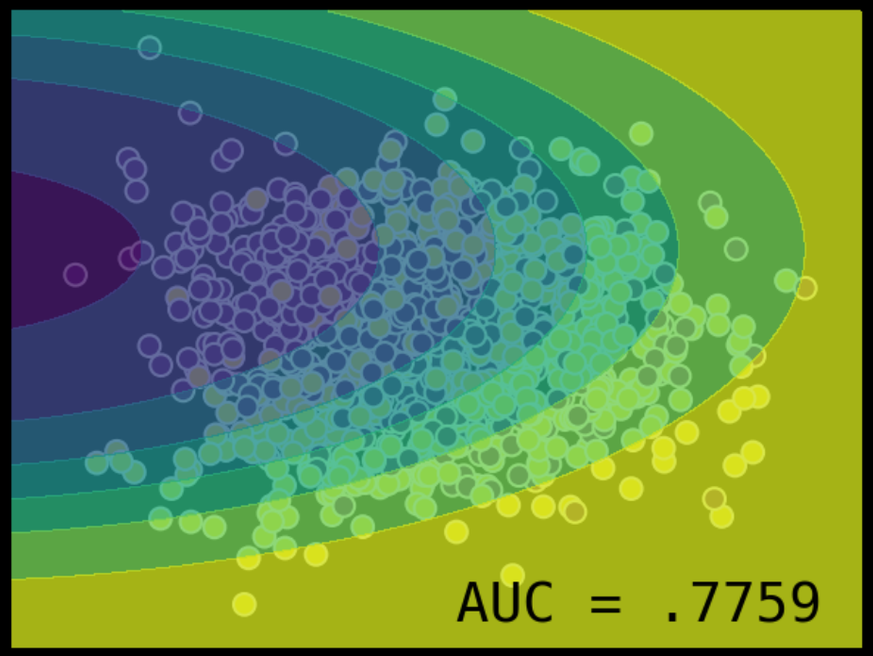
[4] Stacking: Approach



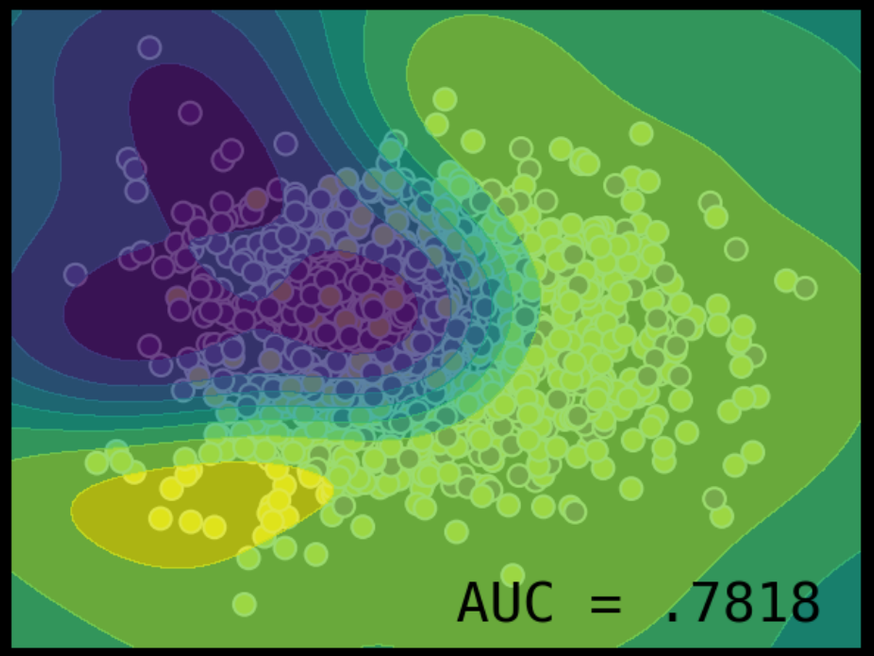
[4] Stacking: Models

Data Space (n)	Feature Space (f)	Model Algorithm (m)	Model Parameters (p)
100%	pc1,pc2	Gaussian	-
100%	pc1,pc2	SVM - RBF	gamma=0.5, C=1
100%	pc1,pc2	Gradient Boost	n_est=10, d=2
-----	-------	-------------	-------------
100%	pc1,pc2	Stacking	Logistic()

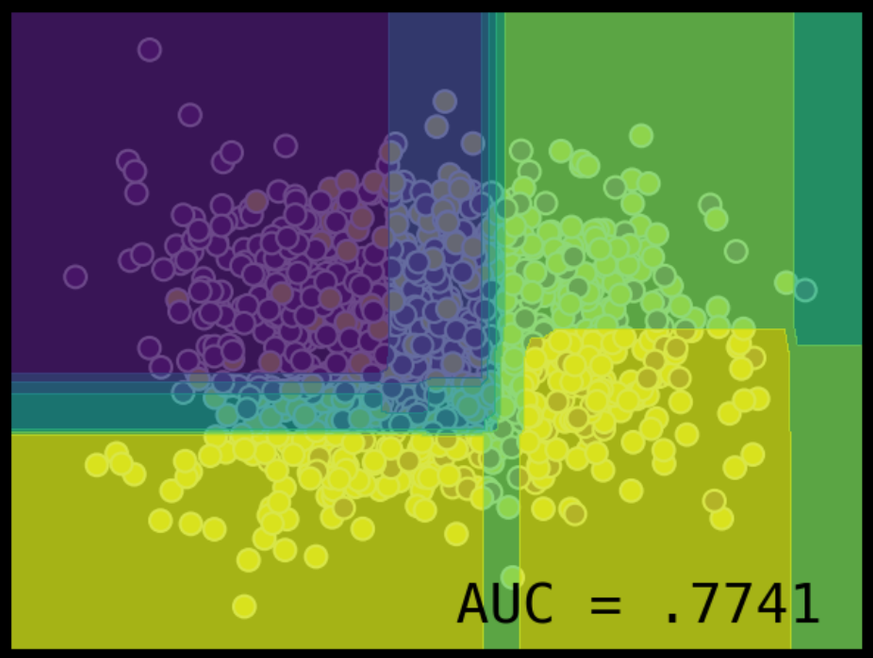
Gaussian



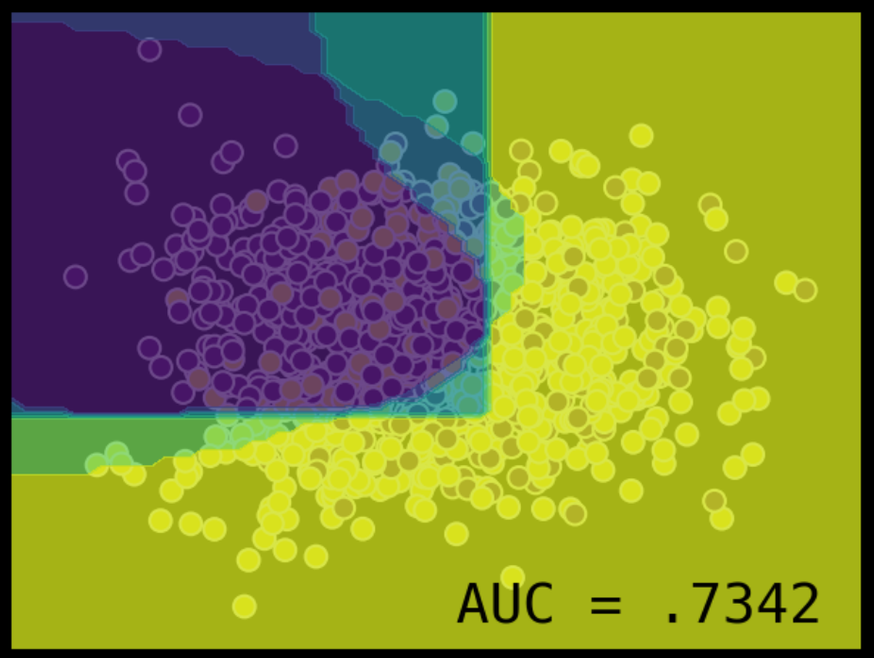
SVM - RBF



Gradient Boosting



Stacking



Ensemble Advantages

Improved **accuracy**

Robustness and better **generalisation**

Parallelization

Ensemble Disadvantages

Model human **readability** isn't great

Time/effort trade-off to improve accuracy may not make sense

Advanced Ensemble Approach

- *Grid Search* for voting weights
- *Bayesian Optimization* for weights & hyper-parameters
- *Feature Engineering* e.g. tree or cluster embedding

Ensembles

"It is the harmony of the diverse parts, their symmetry, their happy balance; in a word it is all that introduces order, all that gives unity, that permits us to see clearly and to comprehend at once both the ensemble and the details."

- **Henri Poincare**

Code and Slides

All the code and slides are available at https://github.com/amitkaps/ensemble

Contact

Amit Kapoor
@amitkaps
amitkaps.com

Bargava Subramanian
@bargava