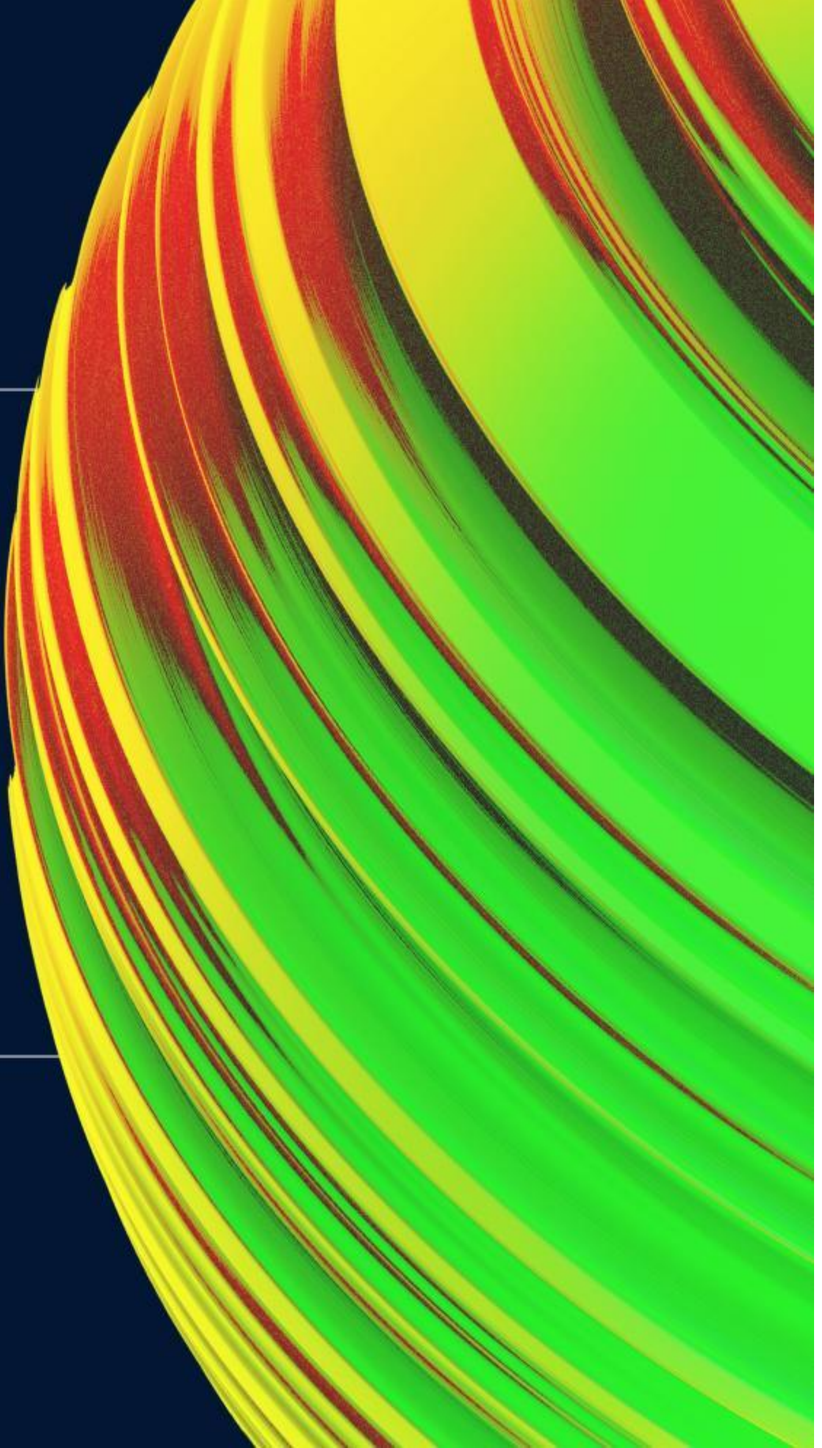
УНИВЕРСИТЕТ ИННОПОЛИС

Ансамбли моделей

Воробьёва Мария

- maria.vorobyova.ser@gmail.com
- @SparrowMaria



План лекции

INNOPOLIS UNIVERSITY

1) Что такое ансамбли моделей
2) Виды ансамблей
3) Случайный лес

Что вы уже знаете о машинном обучении

INNOPOLIS UNIVERSITY

1) k-ближайших соседей
2) линейная регрессия, логистическая регрессия
3) SVM
4) дерево решений
5) кластеризация

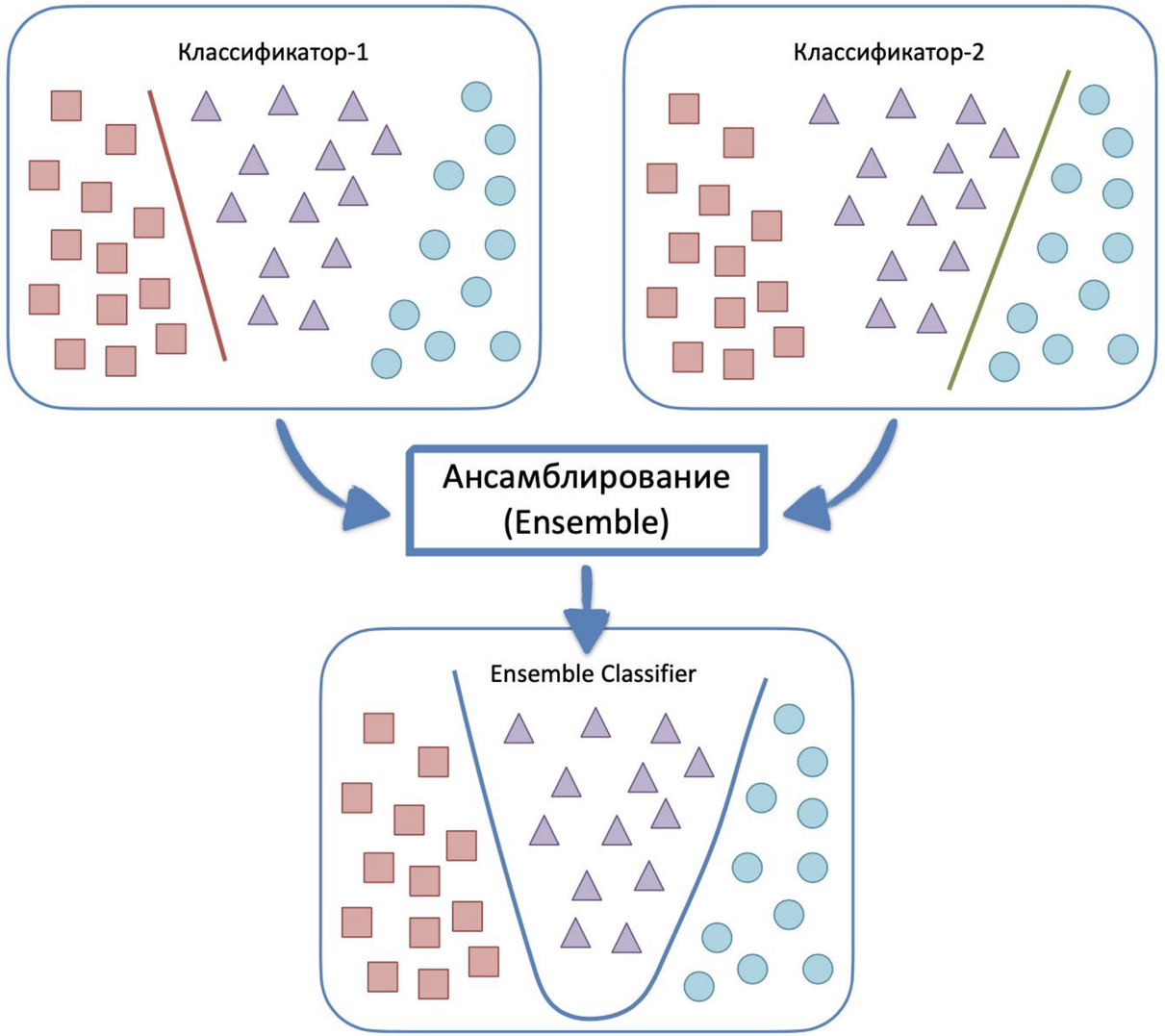
Идея ансамблей моделей

Что делать, если перепробовали все известные методы, а модель все равно с низкими метриками качества?

Есть ли шанс улучшить модель?

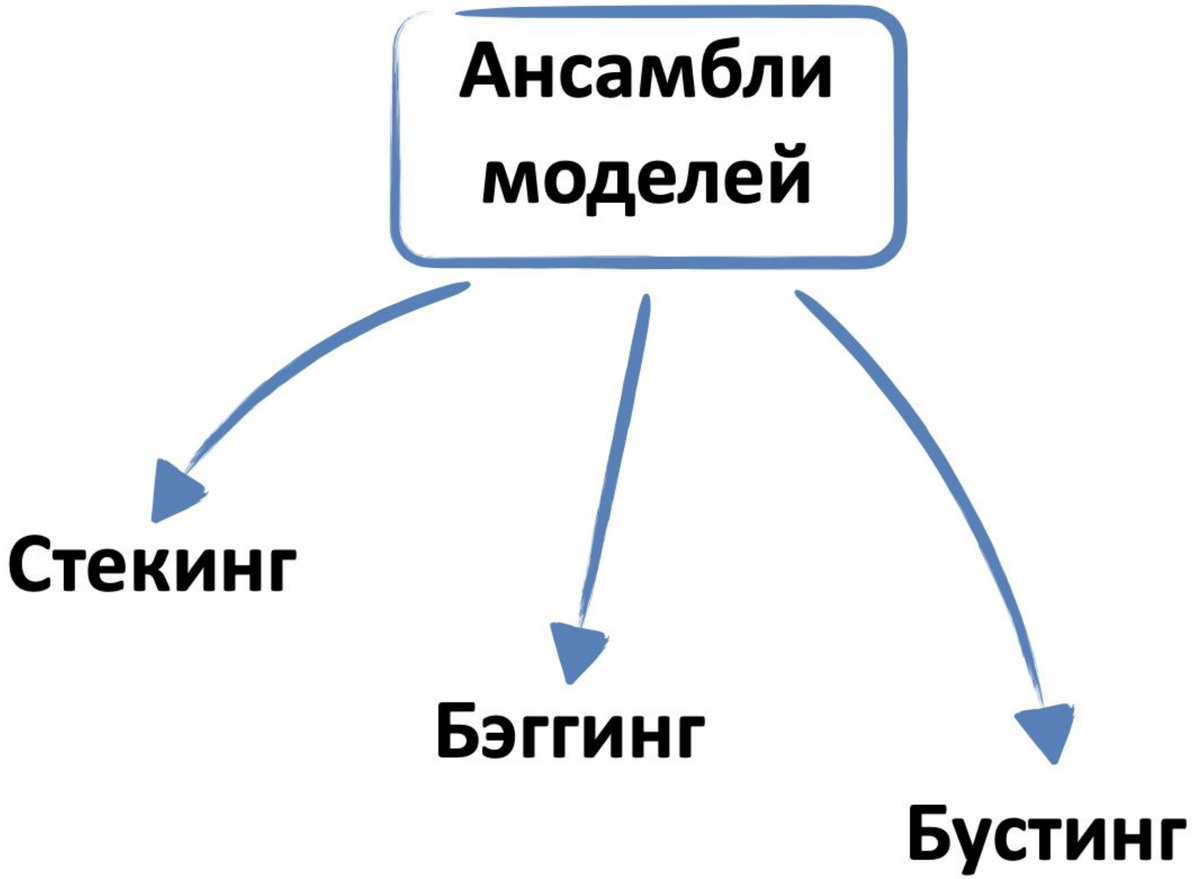
Идея ансамблей моделей

Основная идея заключается в том, что отдельно обучаются **несколько моделей**, а далее их предсказания ***комбинируются***



Виды ансамблей моделей

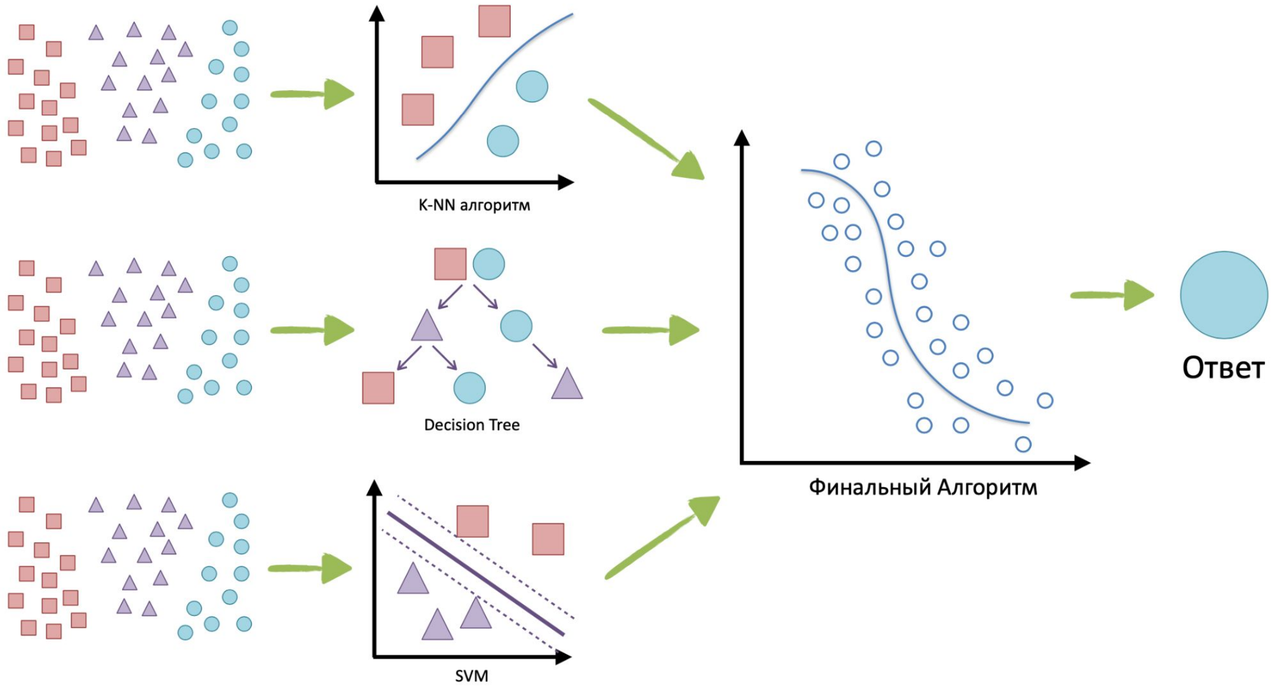
- Стекинг
- Бэггинг
- Бустинг



Стекинг (Stacking)

Идея - возьмем **предсказания базовых моделей** в роли **новых признаков**, а поверх их обучим какую-либо еще ML- модель.

Например:



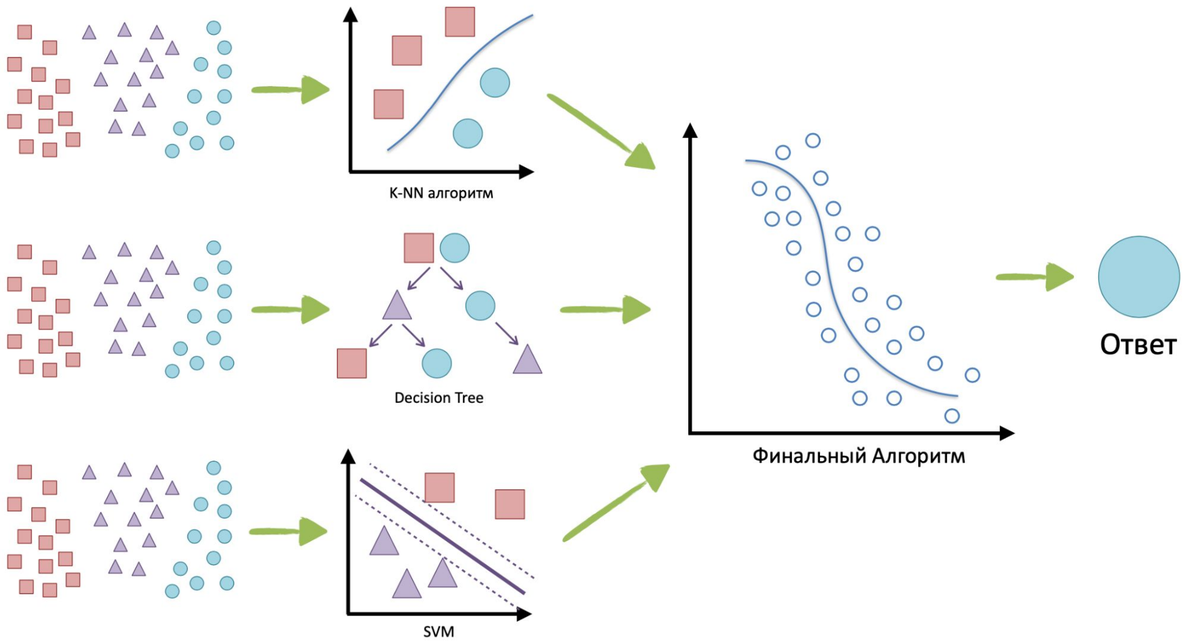
Стекинг (Stacking). Алгоритм

Разбиение данных: Исходные данные разбиваются на две или более части, например на обучающую, валидационную и тестовую

Обучение базовых моделей: На обучающей части данных обучаются несколько различных базовых моделей.

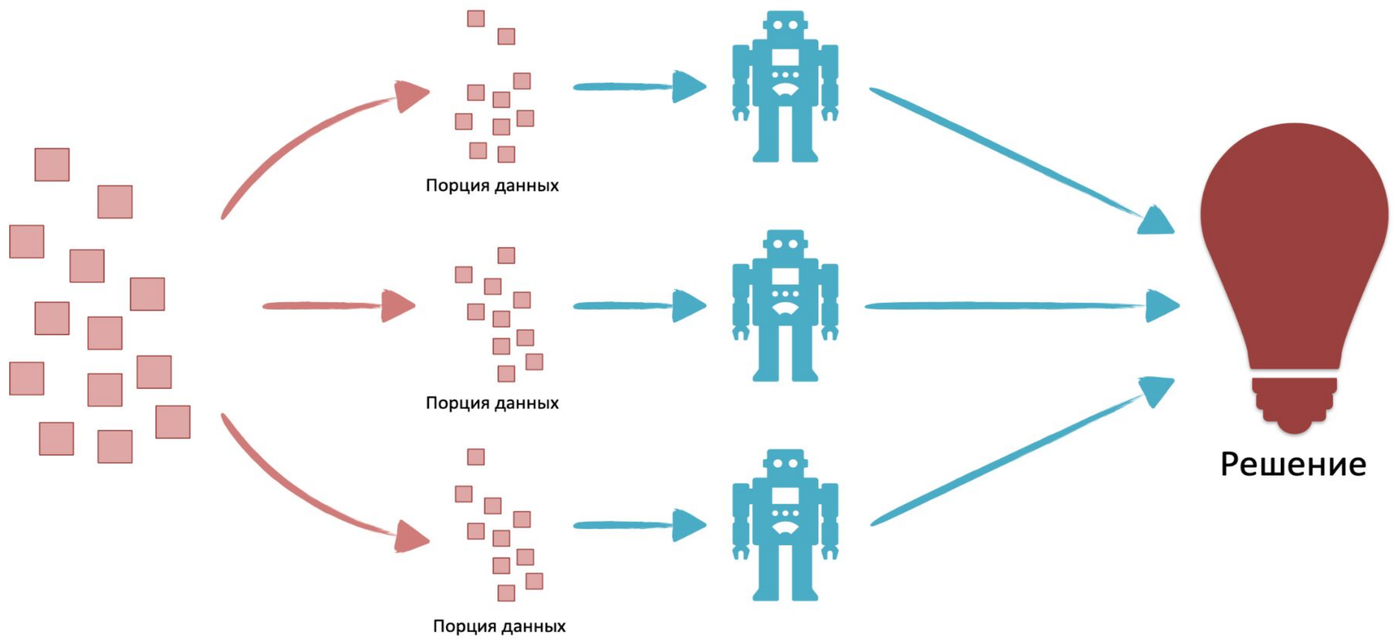
Создание прогнозов: Обученные базовые модели используются для генерации прогнозов на валидационной части данных.

Обучение метамодели: Создается метамодель, которая принимает прогнозы базовых моделей в качестве входных данных и обучается на фактических значениях целевой переменной (так называемый "истинный" результат).



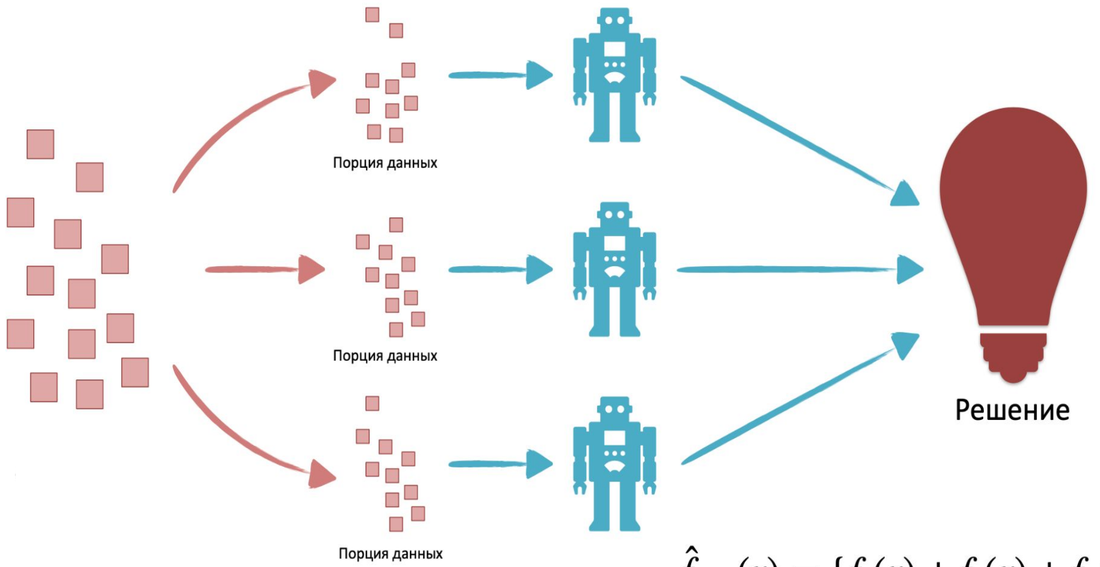
Bagging (bootstrap **agg**regat**ing)**

Идея - несколько моделей обучаются **независимо** на **различных подмножествах данных**, далее прогнозы объединяются



Бэггинг (Bagging). Алгоритм

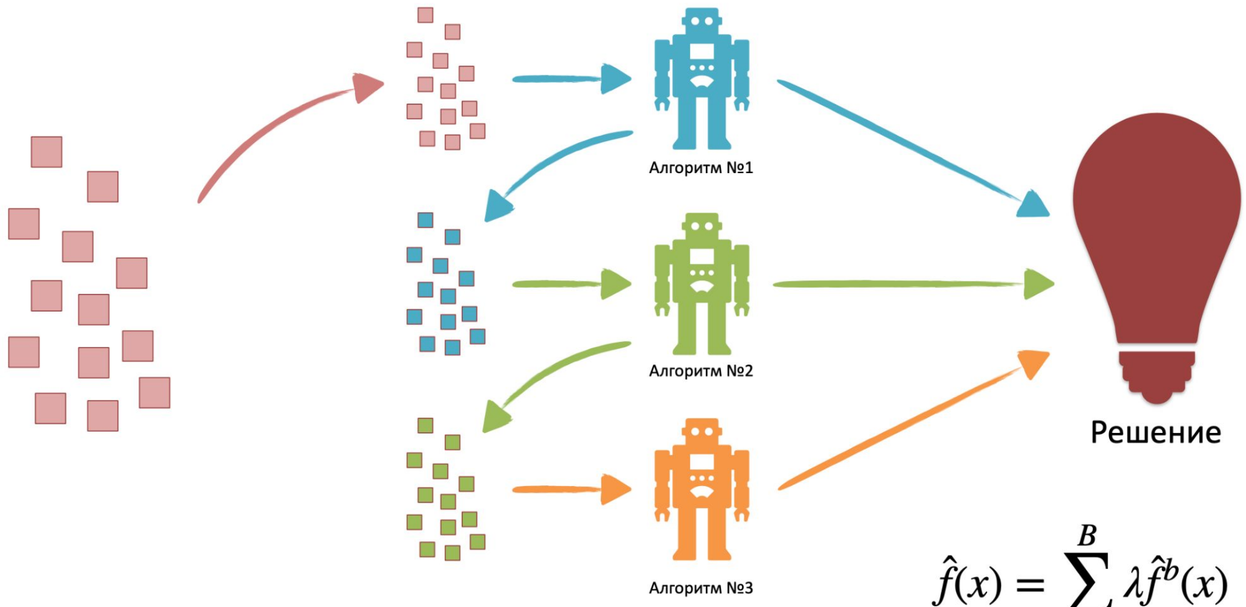
1) **Создание подвыборки:** Из обучающего набора данных случайным образом выбираются подвыборки (выборки с возвращением), которые могут пересекаться.
2) **Обучение базовых моделей:** На каждой из подвыборок данных обучается своя базовая модель (например, решающее дерево)
3) **Прогнозы**: Обученные модели используются для генерации прогнозов на новых данных
4) **Агрегация прогнозов:** Прогнозы, сгенерированные разными моделями, объединяются, например, путем голосования для классификации или усреднения для регрессии.



$$\hat{f}_{bag}(x) = \{f_1(x) + f_2(x) + f_B(x)\}/B$$

Бустинг (Boosting)

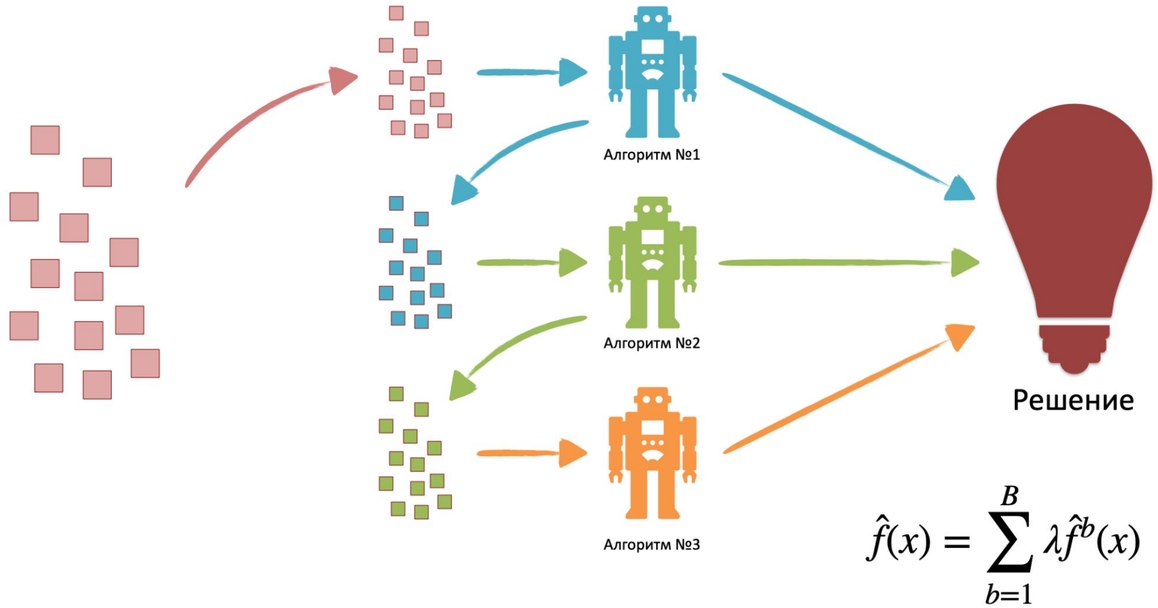
Идея: базовые модели строятся **последовательно**, причем каждая следующая базовая модель **уточняет** предсказание предыдущих (при этом всех).



$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$$

Бустинг (Boosting). Алгоритм

- **Обучение базовой модели:** обучается базовая модель на исходных данных
- **Вычисление ошибок:** Для каждого образца в обучающем наборе рассчитывается ошибка базовой модели, сравнивая её предсказание с фактическим значением.
- **Создание новой модели:** Создаётся новая модель, которая пытается скорректировать ошибки предыдущей модели. Она уделяет больше внимания тем образцам, на которых предыдущая модель ошиблась.
- **Взвешенное голосование:** Прогнозы всех созданных моделей комбинируются с разными весами. Обычно модели, справившиеся с ошибками лучше, получают больший вес.



Bias-variance. Ошибка модели

Model error = Bias*Bias + Variance + Irreducible error

Ошибку модели можно разложить на три составляющие:

- смещение (bias)
- разброс (variance)
- неконтролируемая ошибка

https://academy.yandex.ru/handbook/ml/article/bias-variance-decomposition

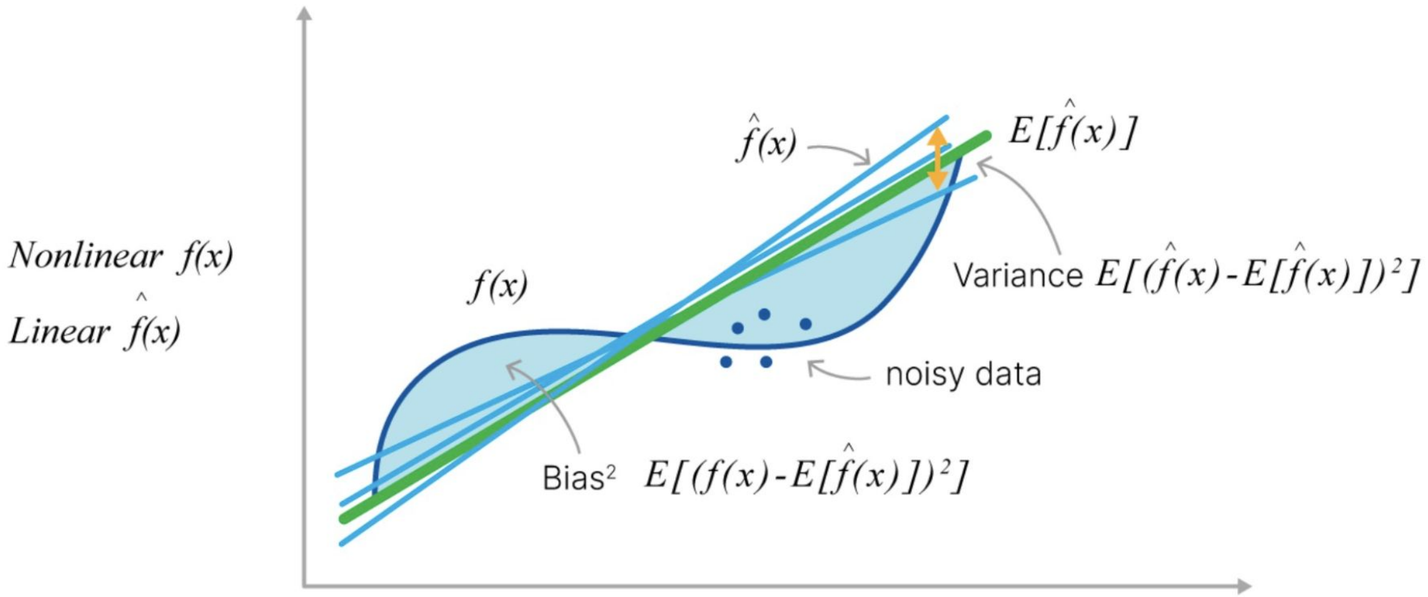
Bias. Смещение

INNOPOLIS UNIVERSITY

Model error = Bias*Bias + Variance + Irreducible error

Смещение (Bias): Смещение отражает разницу между прогнозом модели и фактическим "истинным" значением в данных.

Модель с **высоким смещением** недостаточно сложна, то есть **недообучена**. Высокое смещение может привести к систематическим ошибкам, которые повторяются во всех предсказаниях



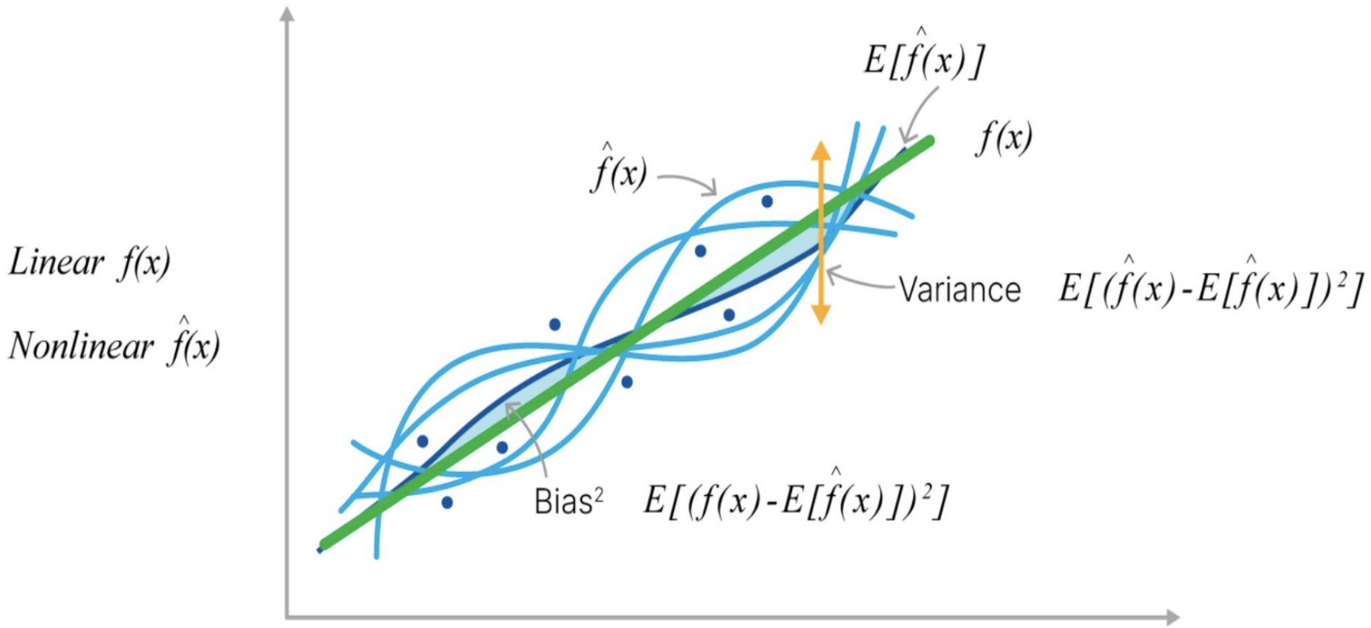
Variance. Разброс

INNOPOLIS UNIVERSITY

Model error = Bias*Bias + Variance + Irreducible error

Дисперсия (Variance): Дисперсия описывает, насколько разнообразны прогнозы модели при различных обучающих наборах данных.

Модель с **высокой дисперсией** чувствительна к малым изменениям в обучающих данных и может давать сильно отличающиеся прогнозы на разных наборах данных. Это может произойти, когда используется **слишком сложная модель**, которая "запоминает" обучающие данные и не обобщает свои знания на новые данные, то есть модель



Bias-Variance

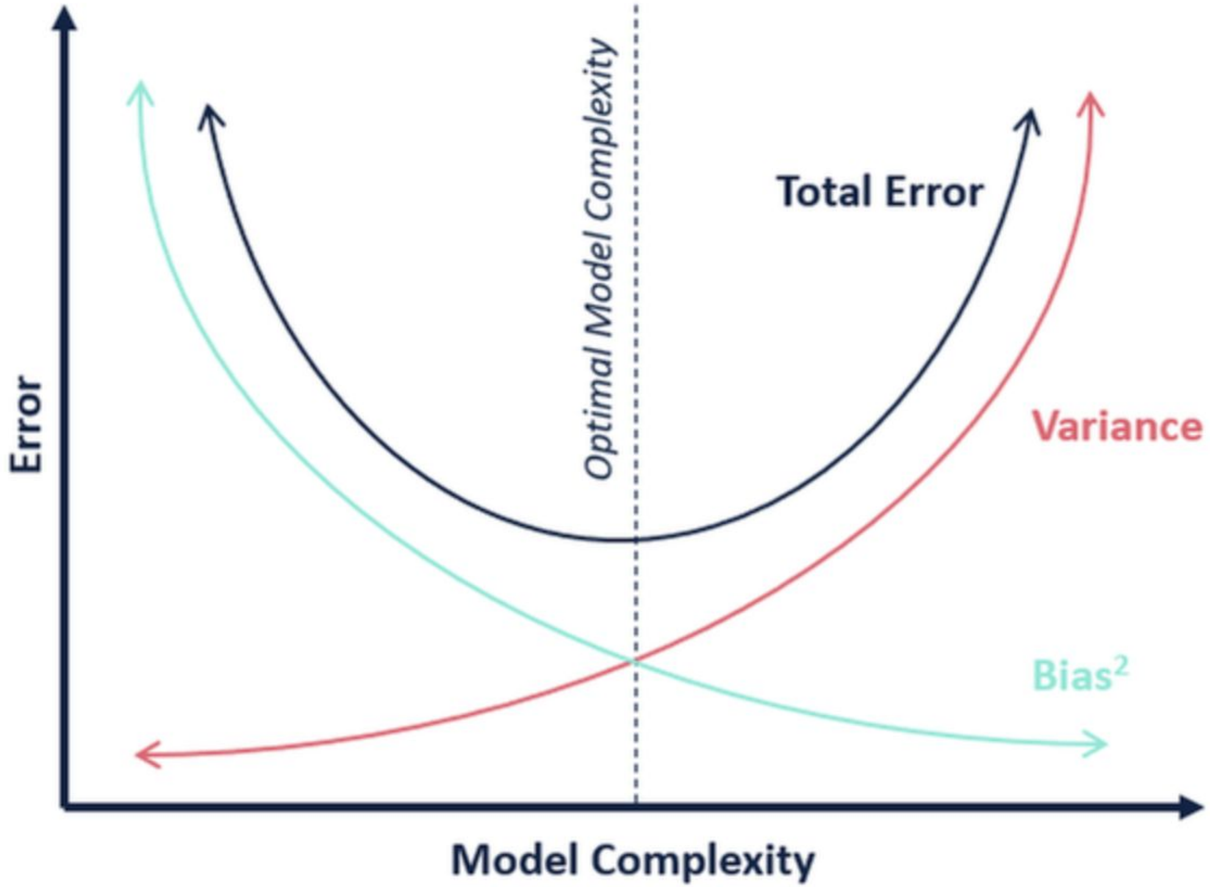
Model error = Bias*Bias + Variance + Irreducible error

Проблема bias-variance

Компромисс между смещением и дисперсией: Цель состоит в том, чтобы достичь баланса между смещением и дисперсией.

Слишком простая модель может иметь низкую дисперсию, но высокое смещение.

Слишком сложная модель может иметь низкое смещение, но высокую дисперсию.



Идеально, мы хотели бы иметь модель, которая способна "захватывать" сложные зависимости в данных, но при этом не переобучается и обобщает на новые данные.

Проблема bias-variance

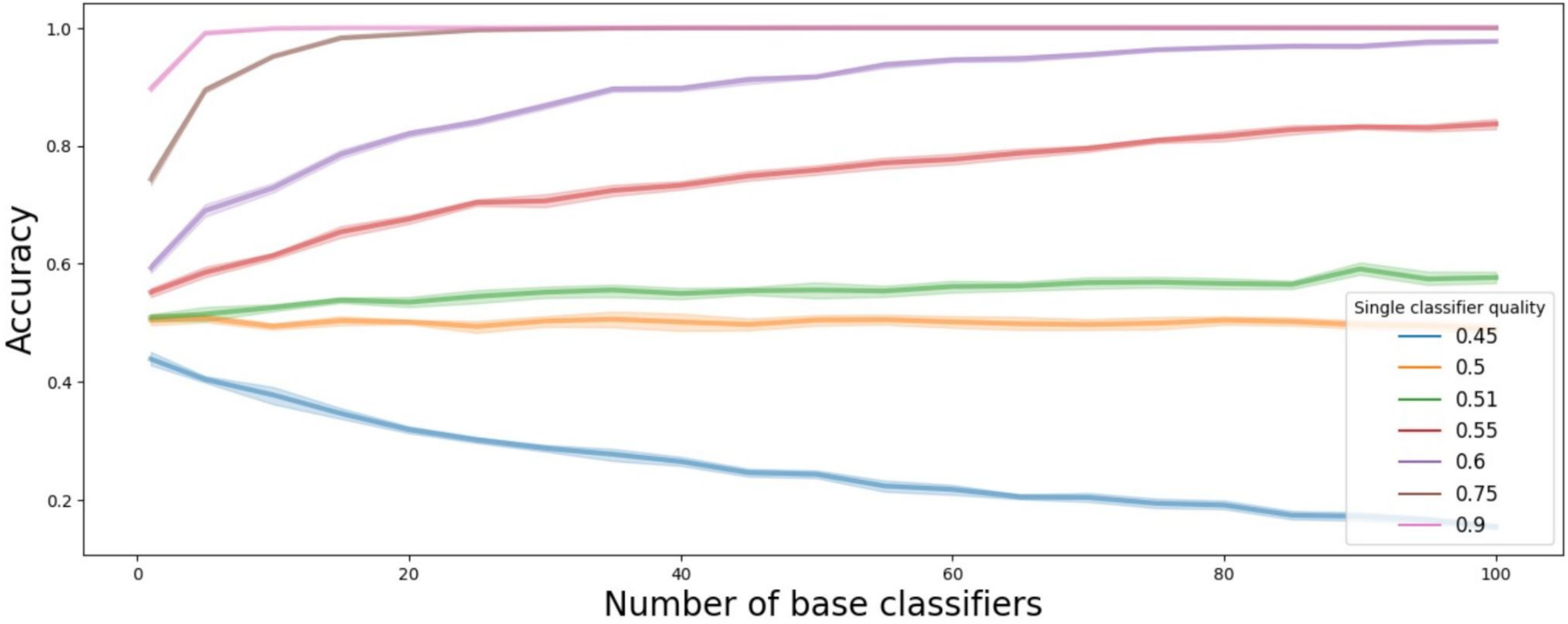
- В **бэггинге** рассматриваются базовые модели с **низким смещением, но высоким разбросом** (например, переобученные деревья).

  Усреднение в бэггинге будет уменьшать разброс.

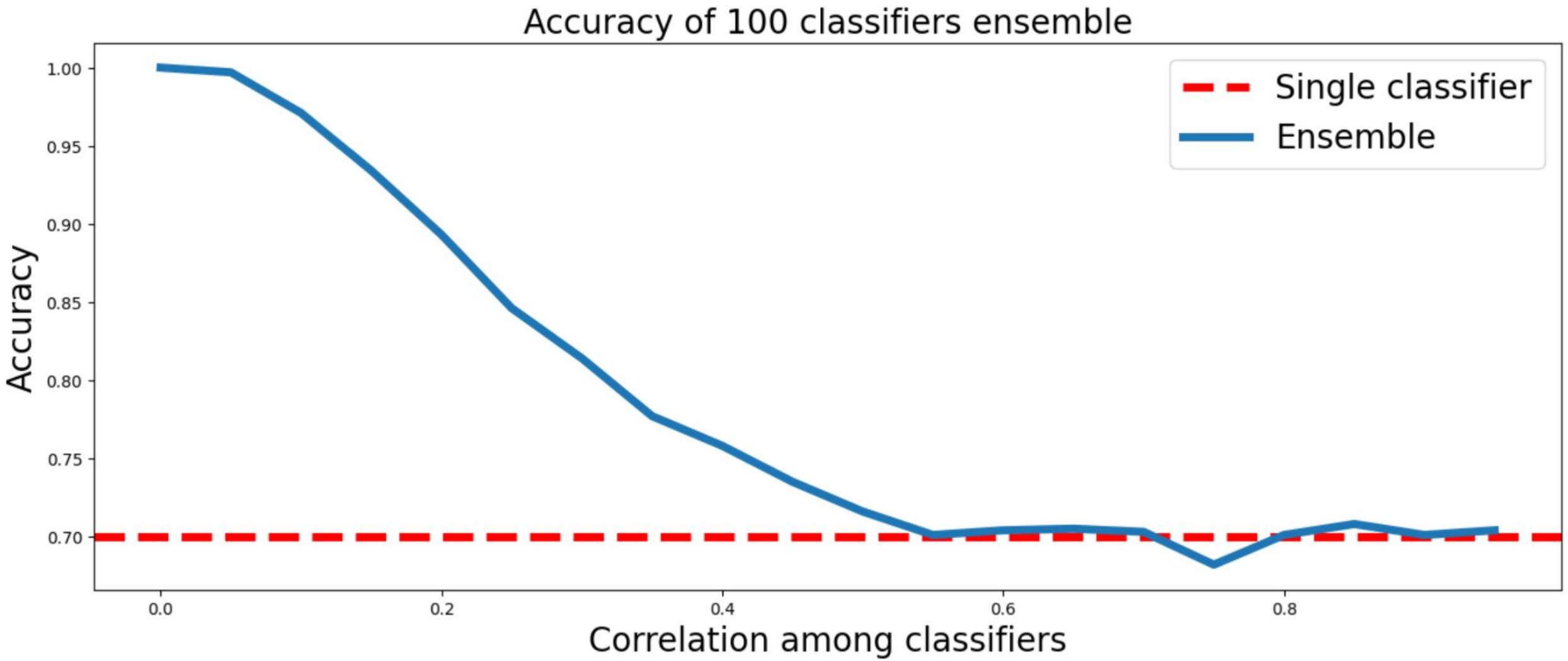
- В бустинге и стекинге рассматриваются **базовые модели с низким разбросом** (т. е. простые, "глупые"), но **высоким смещением** (например, бустинг часто начинают с неглубоких деревьев).

  Последовательное улучшение будет давать модели со меньшим смещением, чем исходная.

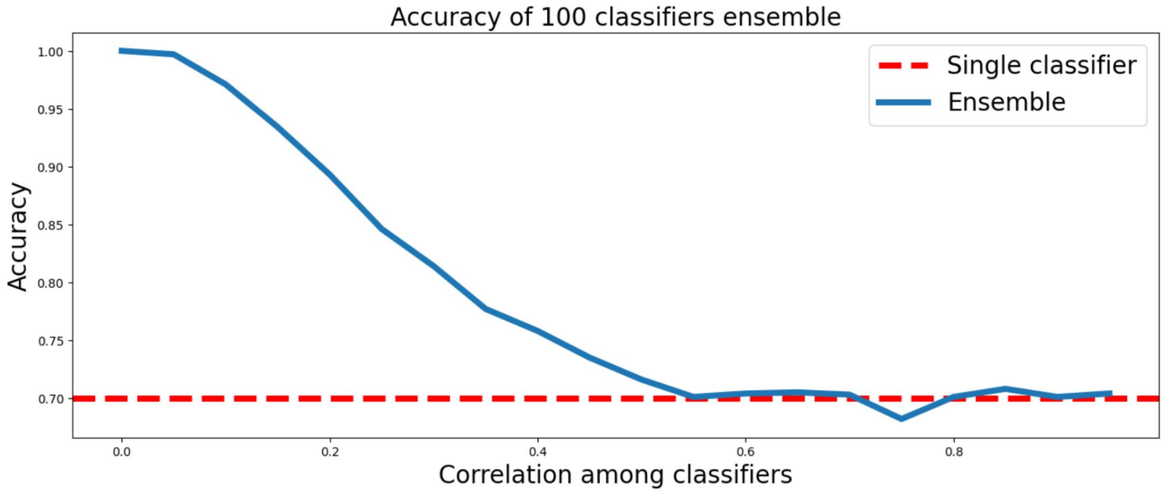
Идея ансамблей моделей

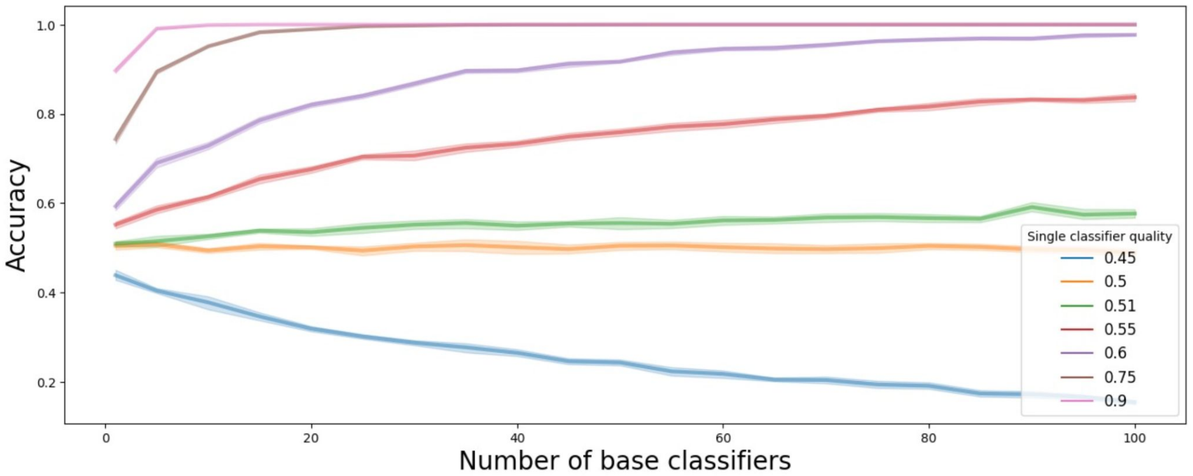


Идея ансамблей моделей. Зависимость алгоритмов



Идея ансамблей моделей. Зависимость алгоритмов





Вывод:

- базовые алгоритмы в ансамбле должны быть чуть лучше случайного
- базовые алгоритмы должны быть независимыми

Идея ансамблей моделей. Зависимость алгоритмов

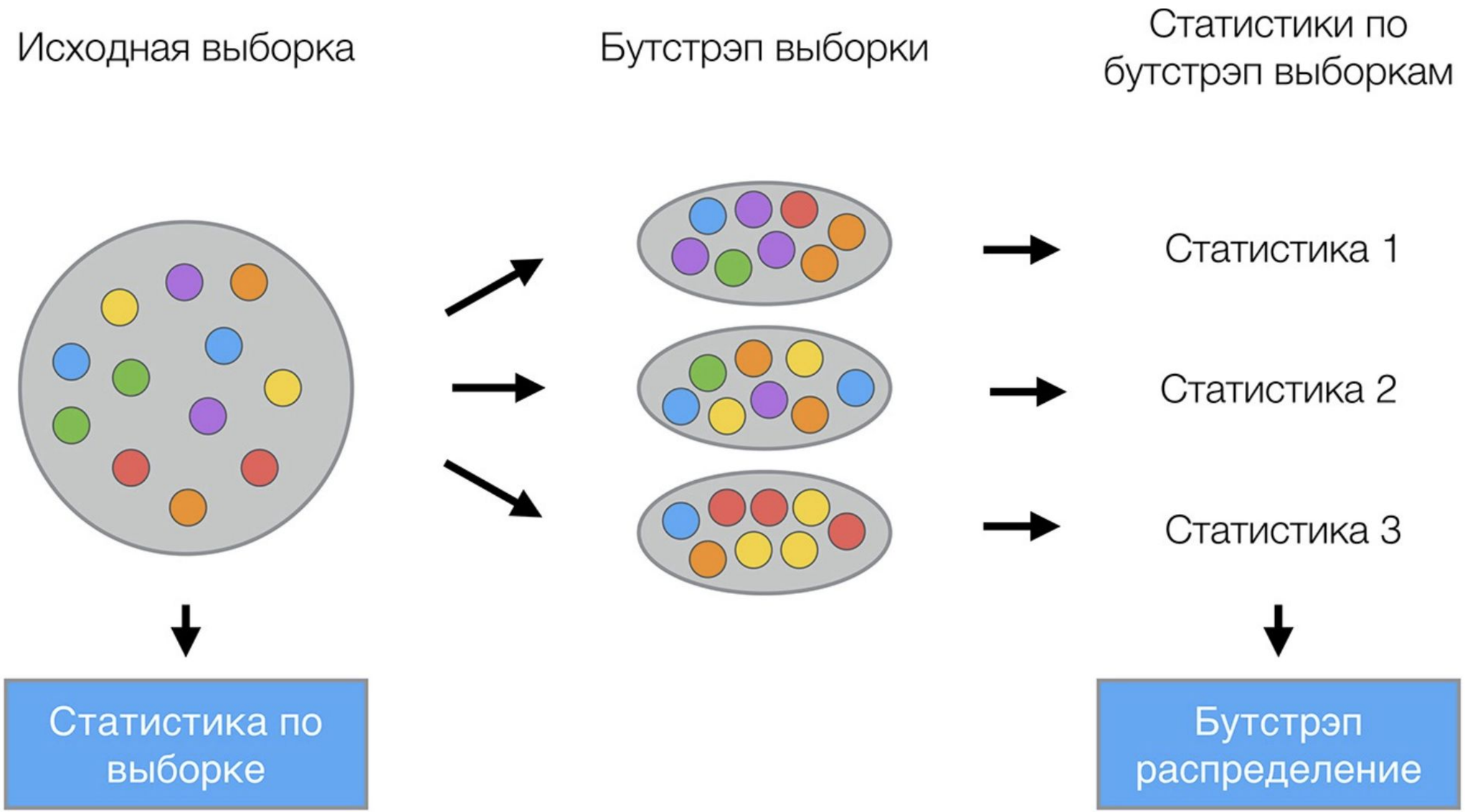
Вывод:

- базовые алгоритмы в ансамбле должны быть чуть лучше случайного
- базовые алгоритмы должны быть независимыми

Вопрос:

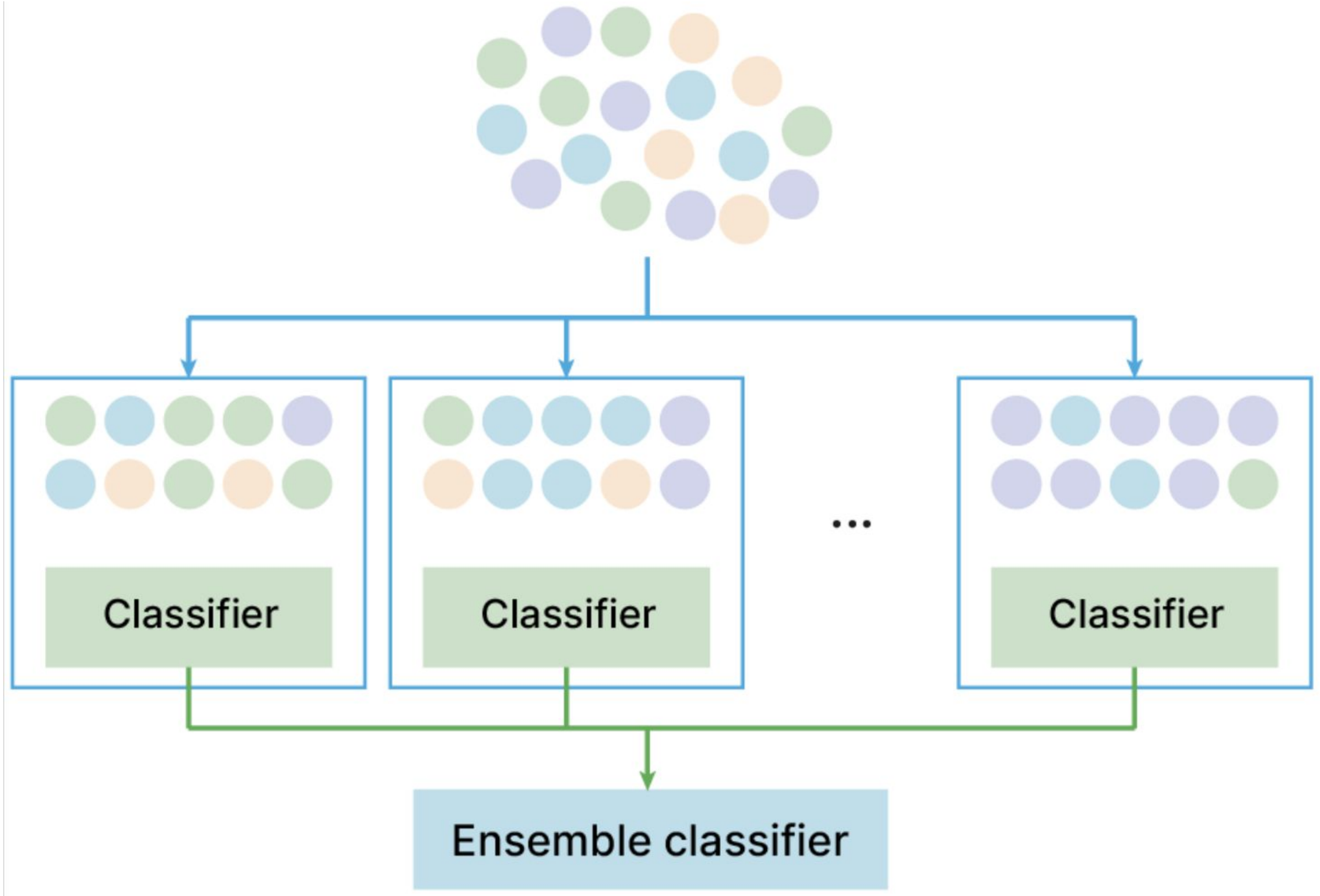
Как на одной выборке можно построить независимые алгоритмы?

Bootstrap


Исходная выборка
Бутстрэп выборки
Статистики по бутстрэп выборкам
Статистика 1
Статистика 2
Статистика 3
Статистика по выборке
Бутстрэп распределение

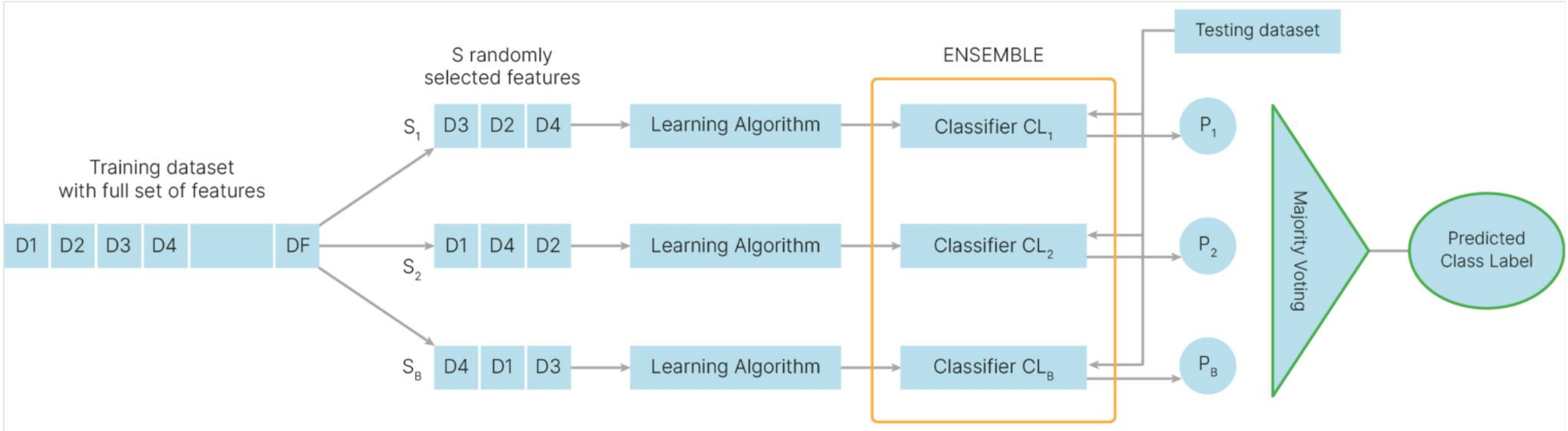
Bagging

Независимость за счет случайных подвыборок



Метод случайных подпространств

Независимость за счет случайных подпространств признаков



Случайный лес

Лео Брейман

RF (random forest) объединяет в себе 2 идеи:

1. бэггинг
2. случайные подпространства признаков

Используется:

1. для классификации
2. для регрессии

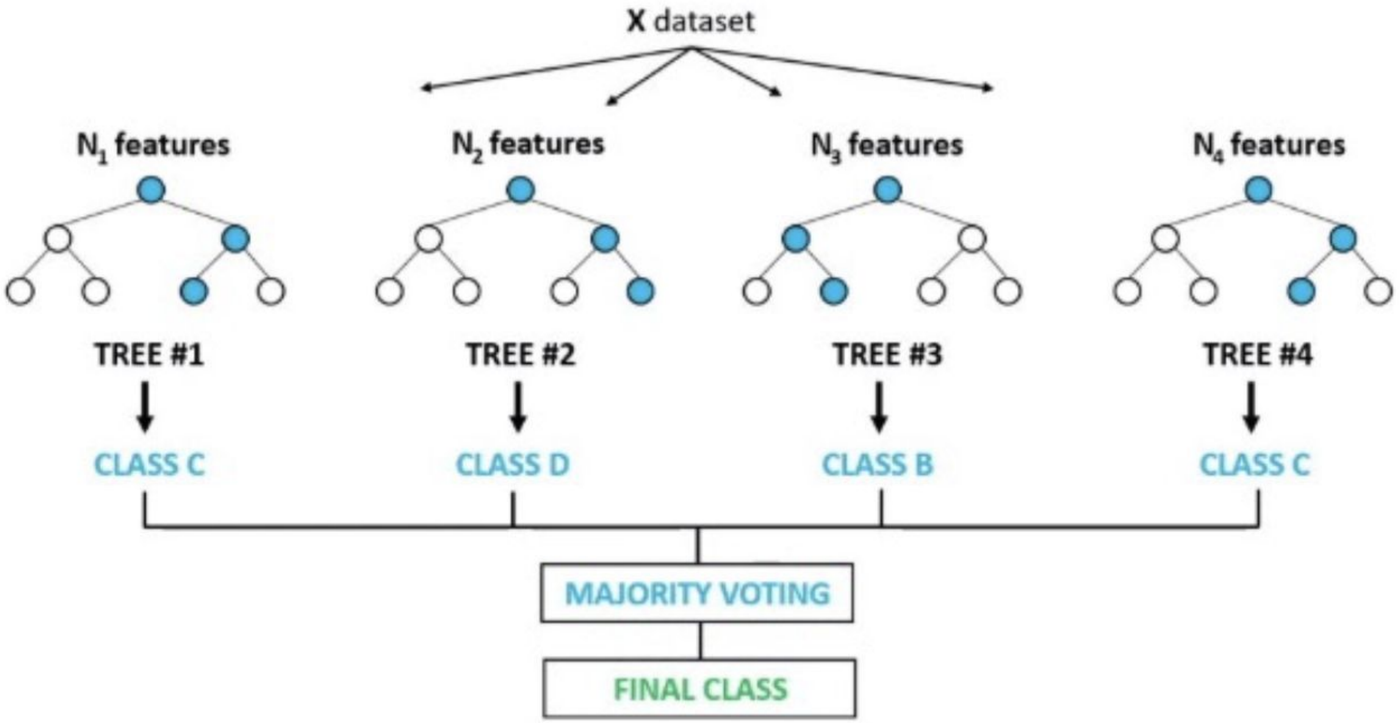
Случайный лес

RF (random forest) = Bagging + RSM

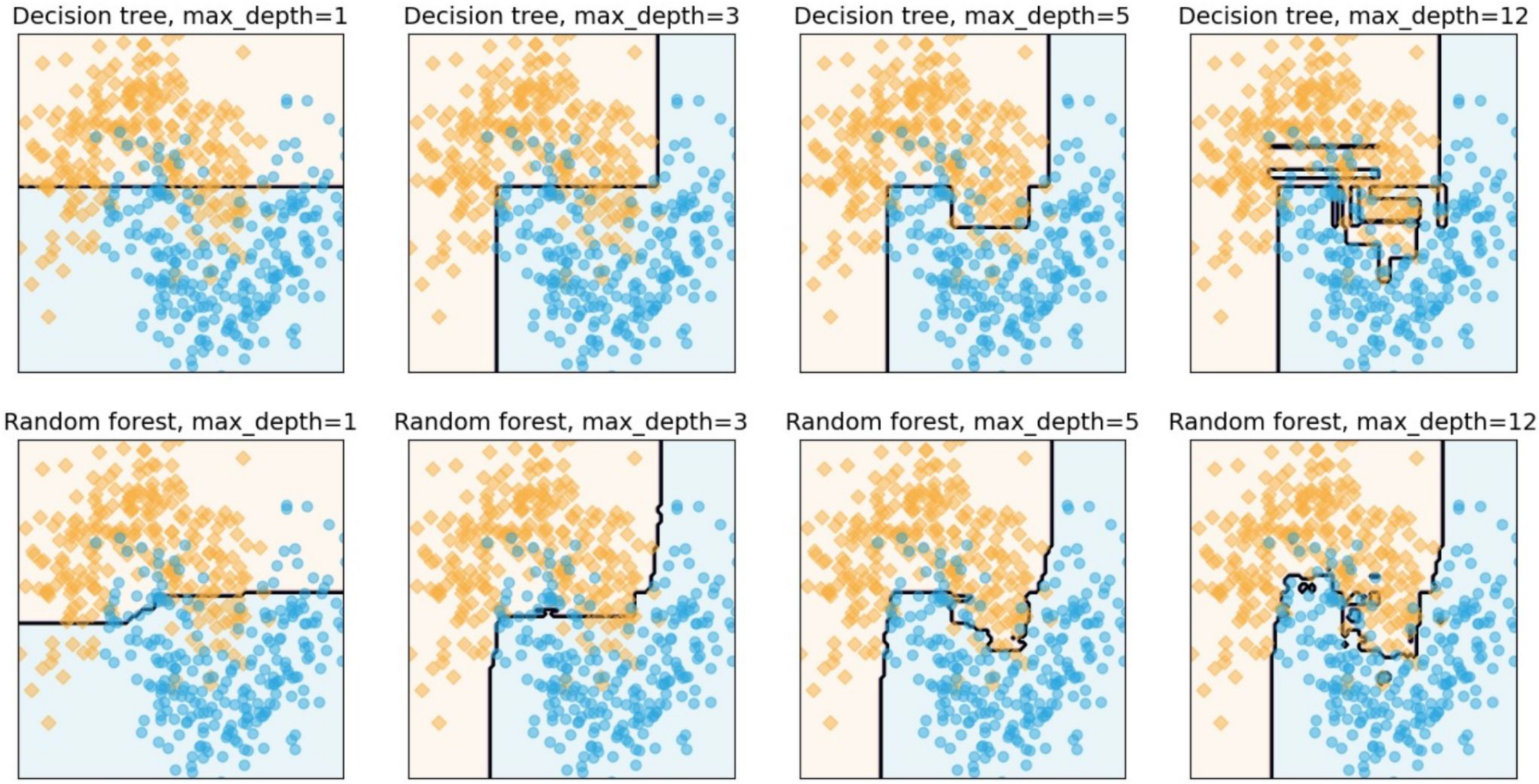
RF (random forest) — это множество **решающих деревьев**.

В задаче **регрессии** их ответы усредняются,

в задаче **классификации** принимается решение голосованием по большинству.



Random forest



Случайный лес

- Выбирается **подвыборка** обучающей выборки размера sample size (м.б. с возвращением) – по ней строится дерево. (для каждого дерева — своя **подвыборка**)
- Для построения каждого разделения в дереве просматриваем **max_features** случайных признаков (для каждого нового разделения — свои случайные признаки)
- Выбираем наилучшие признак и разделение по нему (по заранее заданному критерию). Дерево строится, как правило, до исчерпания выборки (пока в листьях не останутся представители только одного класса) - то есть **глубокие переобученные деревья.**

УНИВЕРСИТЕТ
ИННОПОЛИС

ВОПРОСЫ И ОТВЕТЫ