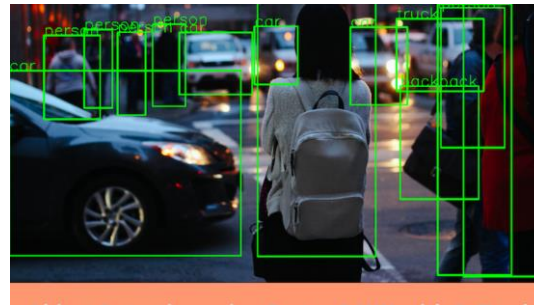
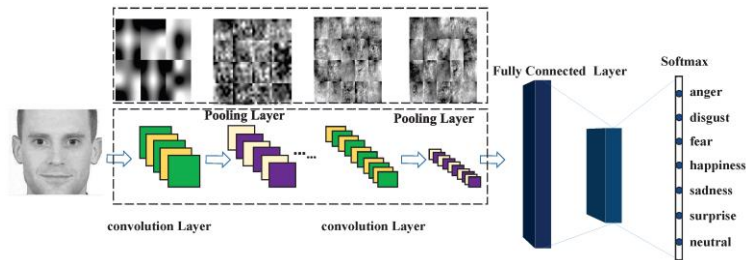


# Основы искусственного интеллекта



## Лекция 6

Ансамбли моделей на примере решающих деревьев

к.ф.-м.н., доцент кафедры ИСиЦТ  
Корнаева Е.П.

# Обучение с учителем: Решающие деревья

## Понятие смещения (bias) и разброса (variance)

**Смещение (bias)** измеряет ожидаемое отклонение от истинного значения функции или параметра.  
**Разброс (variance)** – мера отклонения от ожидаемого значения оценки в произвольной выборке данных.

$$E \left[ (y - \hat{f}(x))^2 \right] = \left( \text{Bias} [\hat{f}(x)] \right)^2 + \text{Var} [\hat{f}(x)] + \sigma^2$$

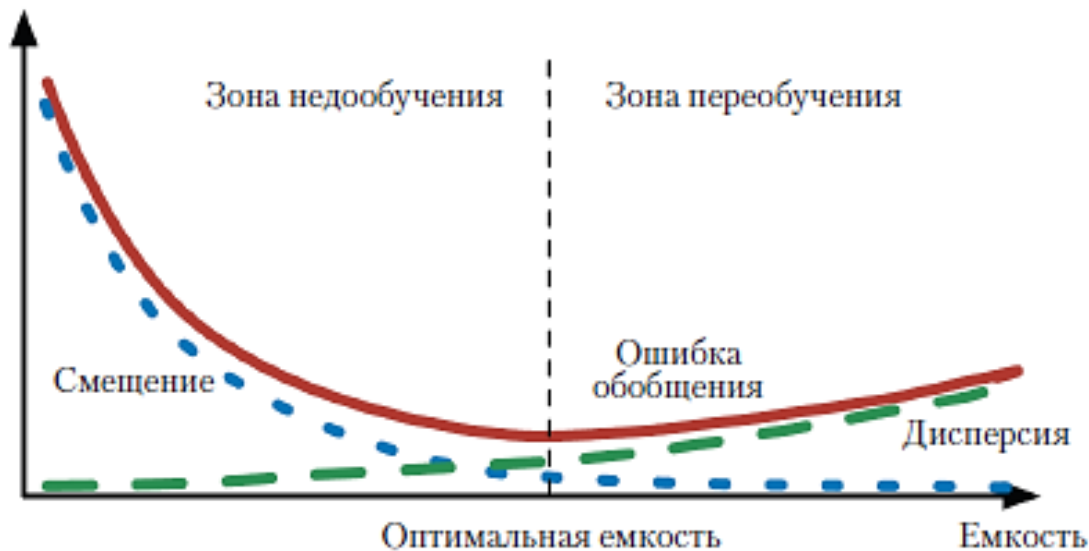


Рис.1. Ошибка модели [1]:  
красная – общая ошибка модели;  
синяя – смещение; зеленая – разброс.

данные	оценка модели	смещение	разброс	шум
--------	---------------	----------	---------	-----

# Обучение с учителем: Решающие деревья

## Понятие смещения (bias) и разброса (variance)

$$\mathbb{E} \left[ (y - \hat{f}(x))^2 \right] = \left( \text{Bias} [\hat{f}(x)] \right)^2 + \text{Var} [\hat{f}(x)] + \sigma^2$$

**Смещение (bias)** оценивает насколько в среднем модель хорошо предсказывает целевую переменную

**Разброс (variance)** оценивает устойчивость модели к изменениям в обучающей выборке

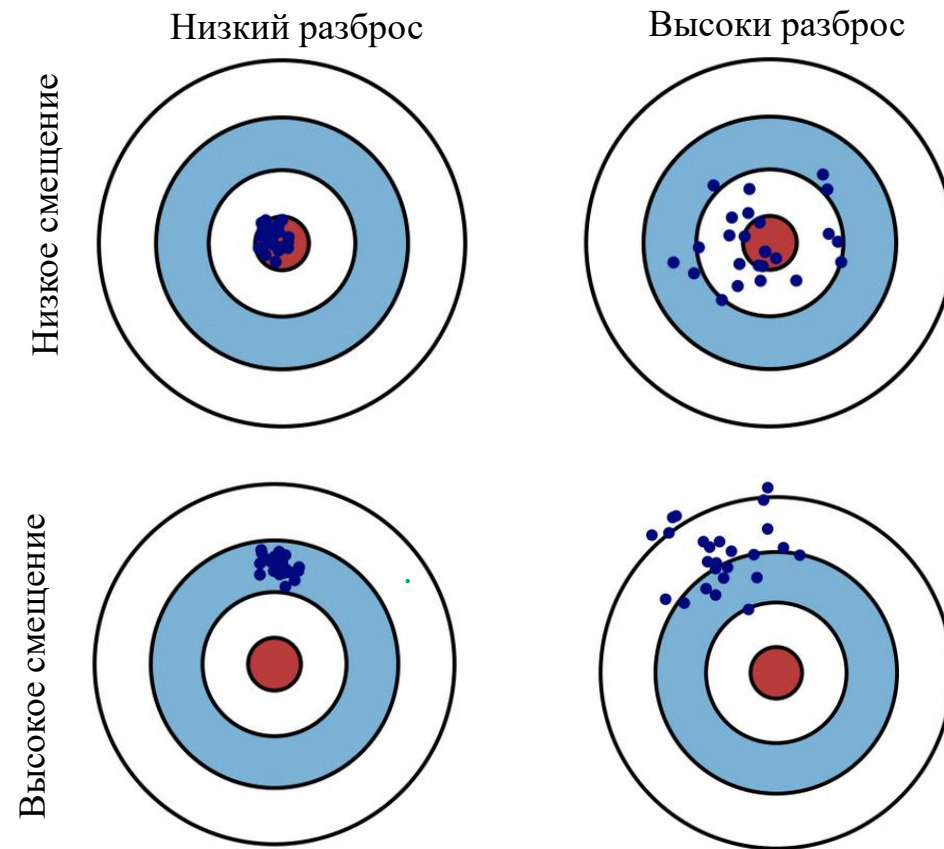


Рис.2. Смещение и разброс [2]

# Обучение с учителем: Решающие деревья

## Понятие смещения (bias) и разброса (variance)

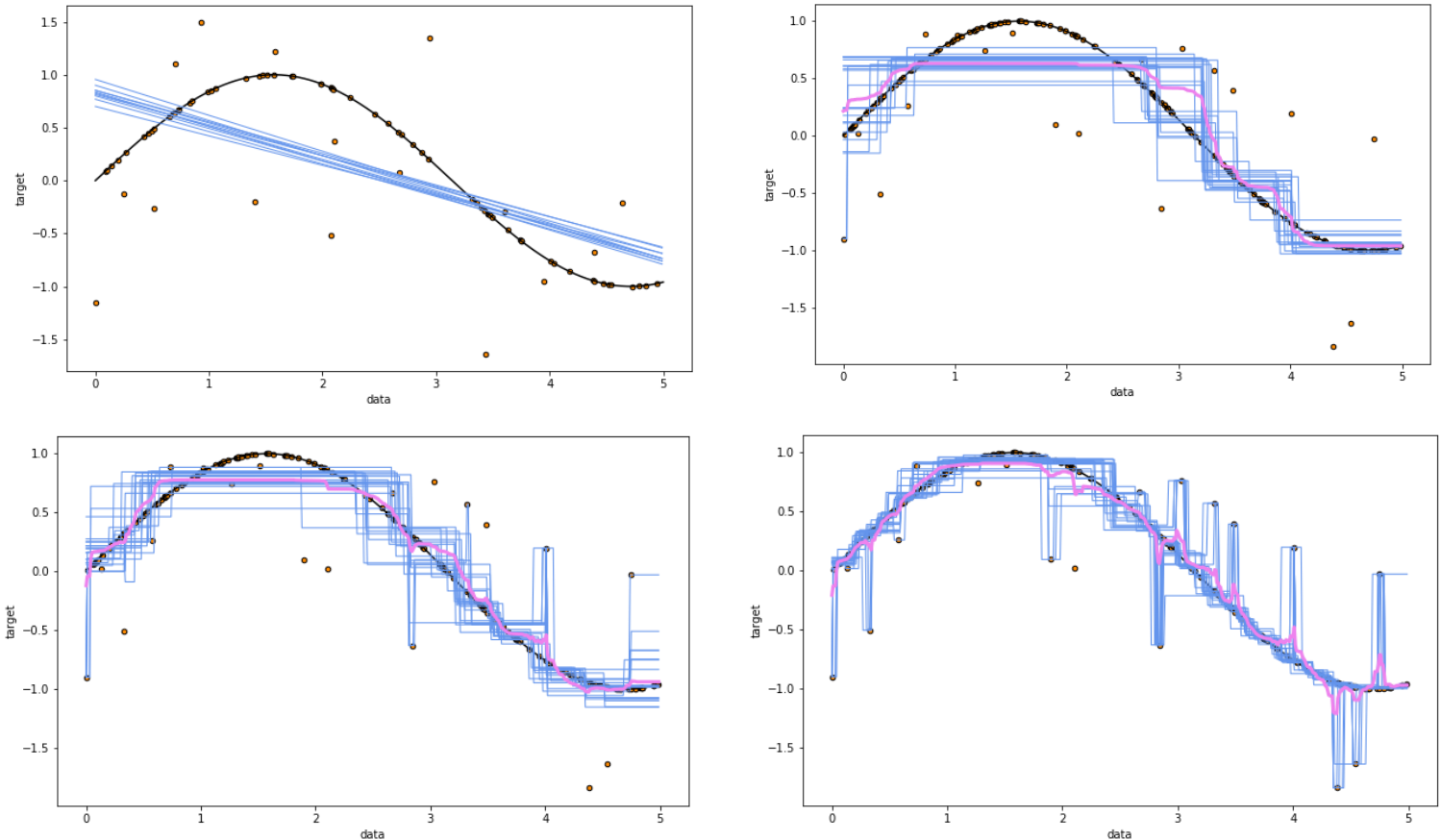
$$\mathbb{E} \left[ (y - \hat{f}(x))^2 \right] = \left( \text{Bias} [\hat{f}(x)] \right)^2 + \text{Var} [\hat{f}(x)] + \sigma^2$$

Например, для регрессии:

$$y(x) = f(x) + \varepsilon$$

$$\text{Bias} [\hat{f}(x)] = \mathbb{E} [\hat{f}(x) - f(x)]$$

$$\text{Var} [\hat{f}(x)] = \mathbb{E}[\hat{f}(x)^2] - \left( \mathbb{E}[\hat{f}(x)] \right)^2$$



# Ансамбли моделей на примере решающих деревьев

## Деревья классификации и регрессии (Classification and regression trees (CART))

**Недостаток решающих деревьев:** чувствительность к шумам в данных

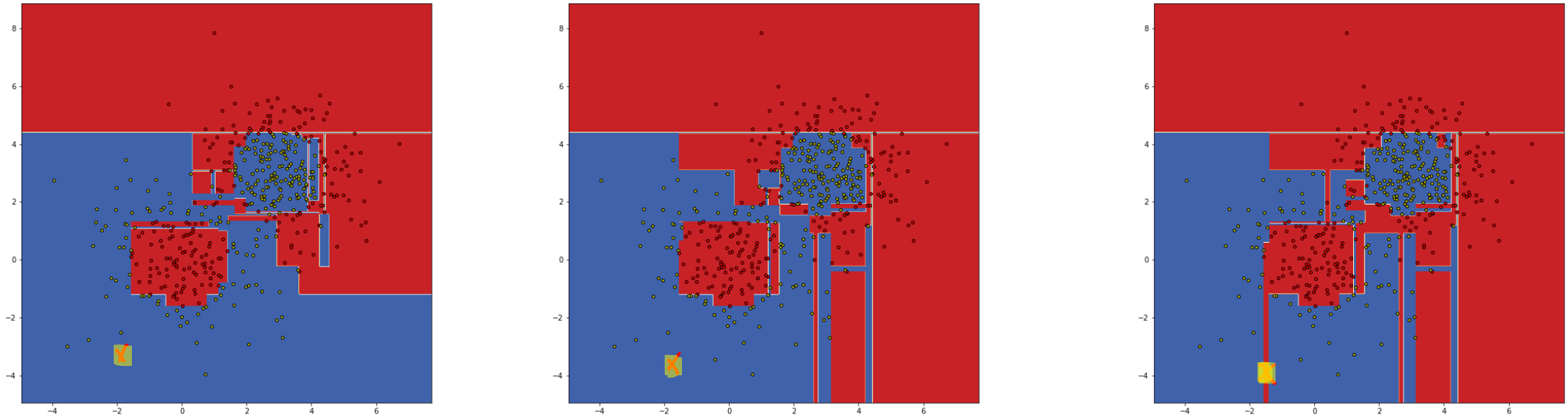


Рис.4. Чувствительность решающих деревьев к изменениям в исходных данных [3]

# Ансамбли моделей на примере решающих деревьев

## Деревья классификации и регрессии (Classification and regression trees (CART))

**Недостаток решающих деревьев:** чувствительность к шумам в данных

Композиция моделей:

$$\tilde{y}(x) = \arg \max_{y \in \mathbb{Y}} \sum_{k=1}^N [\tilde{y}_k(x) = y]$$

$\tilde{y}_k(x)$  - решающие деревья, построенные на подвыборках

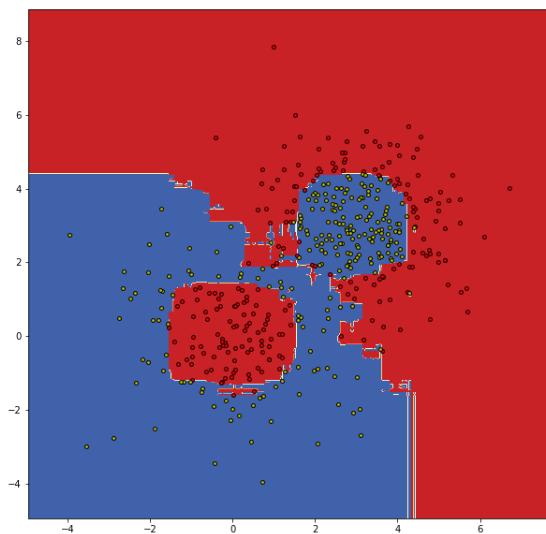


Рис.5. Композиция моделей [3]



# Ансамбли моделей на примере решающих деревьев

## Композиция моделей на примере деревьев решений

**Бэггинг (Bagging от Bootstrap aggregation)** - композиция моделей, обученных независимо на случайных подмножествах объектов

Пусть  $\tilde{y}_k(x)$  -  $k$ -ая модель в композиции  $\tilde{y}(x)$  из  $N$  моделей

Смещение  $\tilde{y}(x)$  такое же, как у  $\tilde{y}_k(x)$

$$\text{разброс}(\tilde{y}(x)) = \frac{1}{N} (\text{разброс}(\tilde{y}_k(x))) + \text{ковариация}(\tilde{y}_k(x), \tilde{y}_q(x))$$

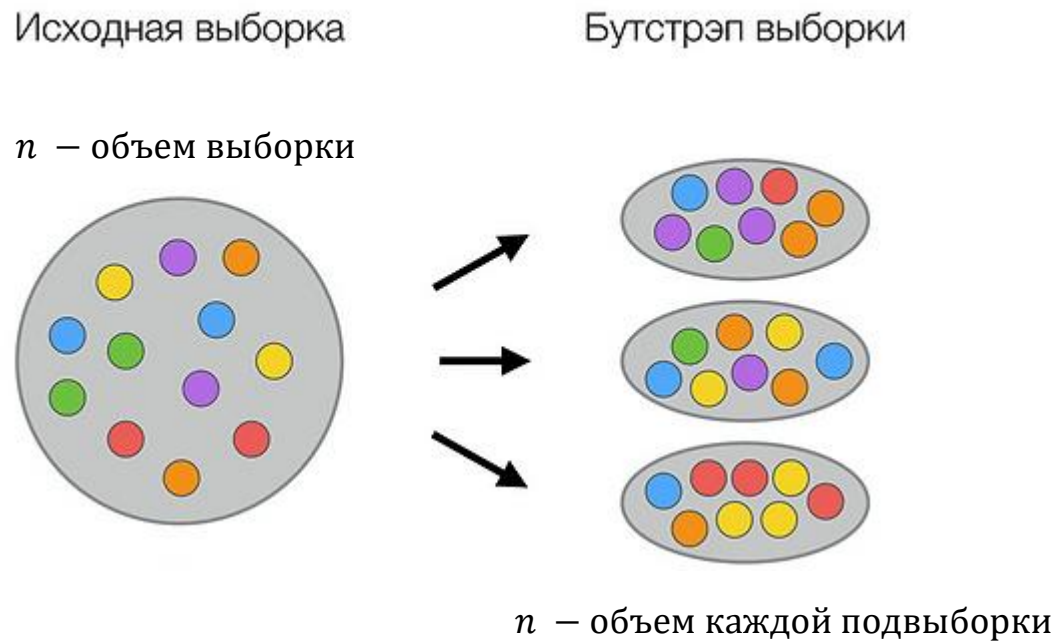
Если базовые модели  $\tilde{y}_k(x)$  независимы, то разброс уменьшается в  $N$  раз!

# Ансамбли моделей на примере решающих деревьев

## Композиция моделей на примере деревьев решений

**Бэггинг (Bagging от Bootstrap aggregation)** - композиция моделей, обученных независимо на случайных подмножествах объектов

Процедура бутстрэпа на рис. 6.



Примерно 37% примеров остаются вне выборки бутстрэпа и не используются при построении  $k$ -го дерева

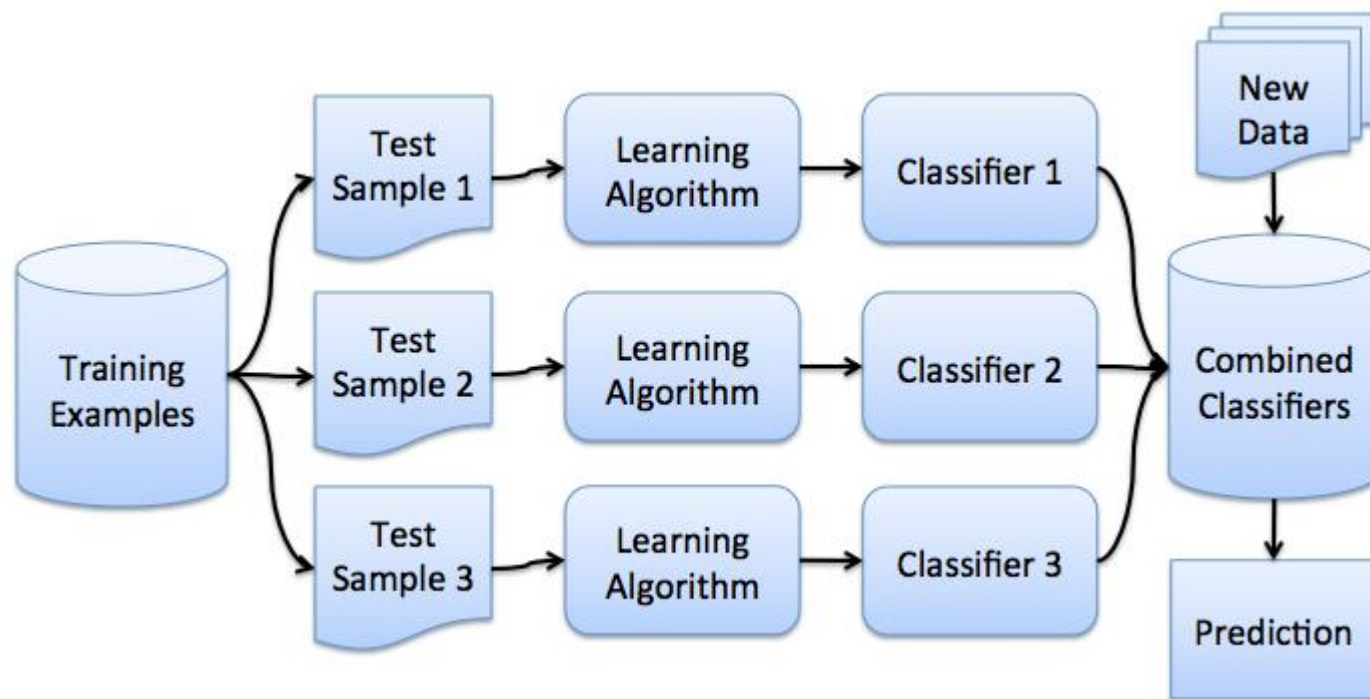
Рис.6. Бутстрэп [4]



# Ансамбли моделей на примере решающих деревьев

## Композиция моделей на примере деревьев решений

**Бэггинг (Bagging от Bootstrap aggregation)** - композиция моделей, обученных независимо на случайных подмножествах объектов



Для классификации:

$$\tilde{y}(x) = \arg \max_{y \in \mathbb{Y}} \sum_{k=1}^N [\tilde{y}_k(x) = y]$$

Для регрессии:

$$\tilde{y}(x) = \frac{1}{N} \sum_{k=1}^N \tilde{y}_k(x)$$

# Ансамбли моделей на примере решающих деревьев

## Композиция моделей на примере деревьев решений

**Бэггинг (Bagging от Bootstrap aggregation)** - композиция моделей, обученных независимо на случайных подмножествах объектов

- Бэггинг не меняет смещение;
- Бэггинг понижает разброс;
- Для эффективного понижения разброса надо строить независимые базовые модели композиции.

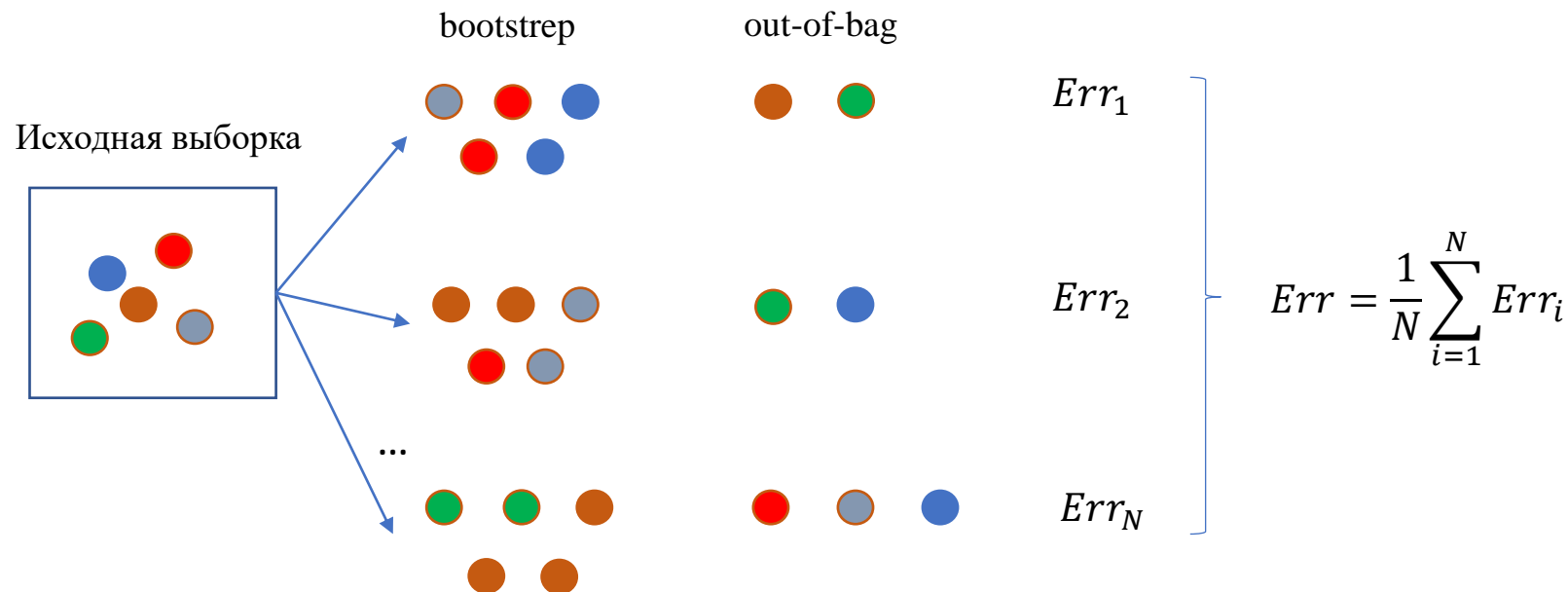
# Ансамбли моделей на примере решающих деревьев

## Композиция моделей на примере деревьев решений

**Бэггинг (Bagging от Bootstrap aggregation)** - композиция моделей, обученных независимо на случайных подмножествах объектов

**Out-of-bag** оценка - это усредненная оценка базовых алгоритмов на тех ~37% данных, на которых они не обучались

т.к. каждое дерево обучается примерно на 63% данных, то остальные объекты можно рассматривать как тестовую выборку для каждого дерева.



# Ансамбли моделей на примере решающих деревьев

## Композиция моделей на примере деревьев решений

**Случайный лес (random forest)**- ансамбль моделей, использующих метод случайного подпространства

Подпространство признаков  $\hat{X} \subseteq X$

	$X_1$	$X_2$	$X_3$	...	$X_l$	$X_m$	$Y$
1	$x_{11}$	$x_{12}$	$x_{13}$		$x_{1l}$	$x_{1m}$	$y_1$
2	$x_{21}$	$x_{22}$	$x_{23}$		$x_{2l}$	$x_{2m}$	$y_2$
...							...
n	$x_{n1}$	$x_{n2}$	$x_{n3}$		$x_{nl}$	$x_{nm}$	$y_n$

Выбор  $m'$  :

- для классификации  $m' = \sqrt{m}$
- для регрессии  $m' = \frac{m}{3}$

$m$  — мерность признакового пространства;

$m'$  — мерность признакового подпространства для разбиения в узлах каждого дерева.

