Isolation Forest

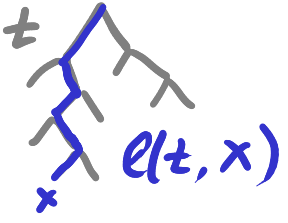
A. Grow the trees

n-dimensional data

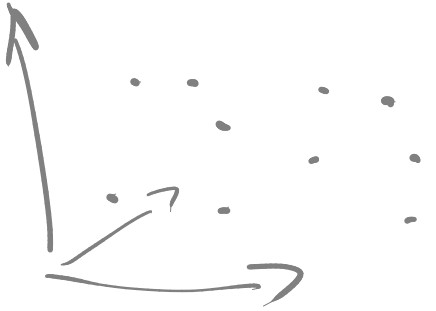
1. take some random feature
2. take some random value
3. split the data
4. repeat

B. Calculate the scores

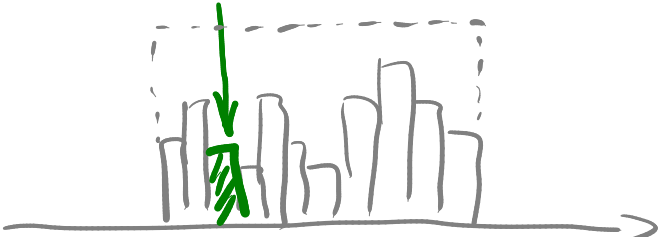
$$\bar{\ell}(x) = \frac{1}{N_t} \sum_t \ell(t, x),$$

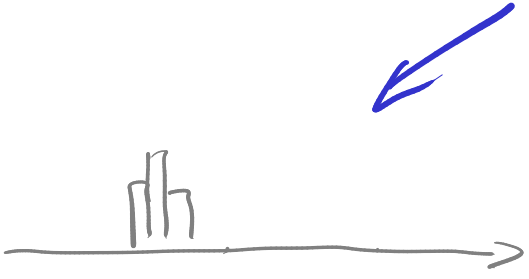


Isolation Forest



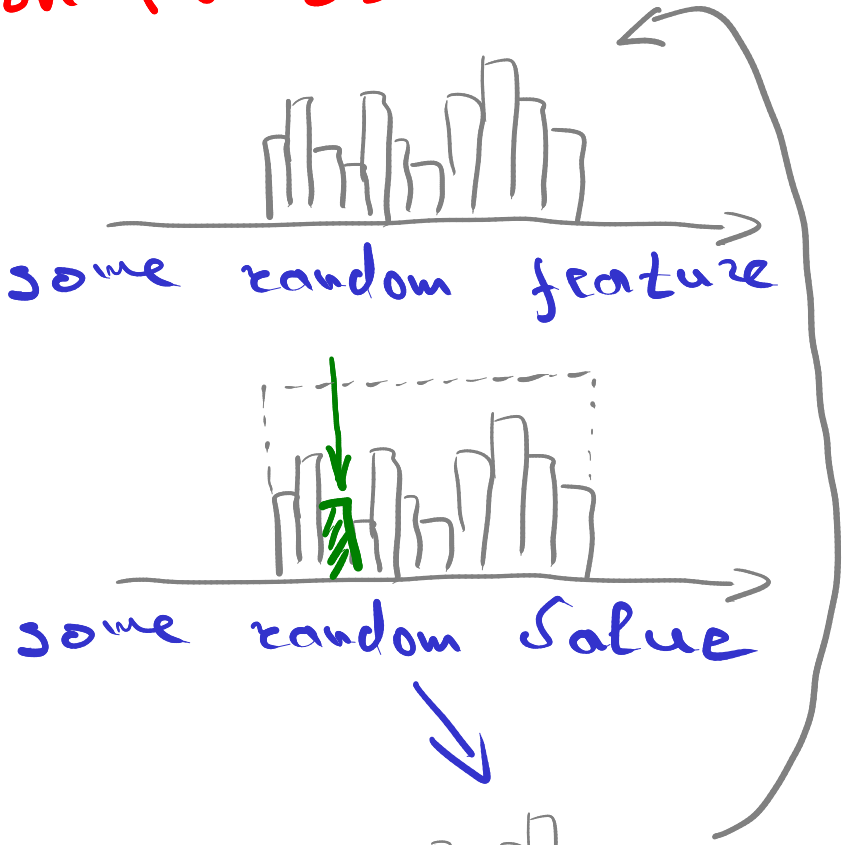
n-dimensional data



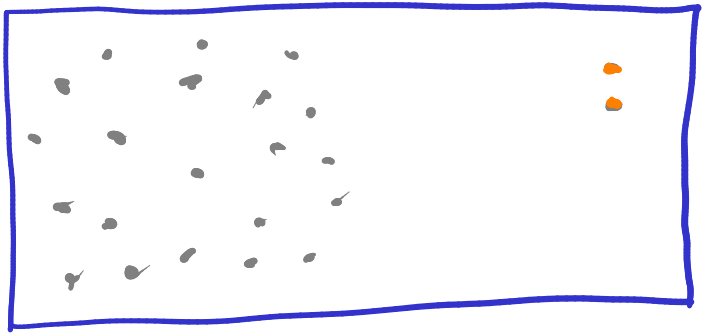


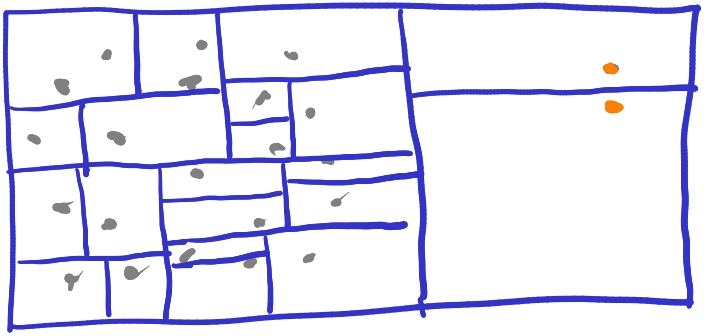
Isolation Forest

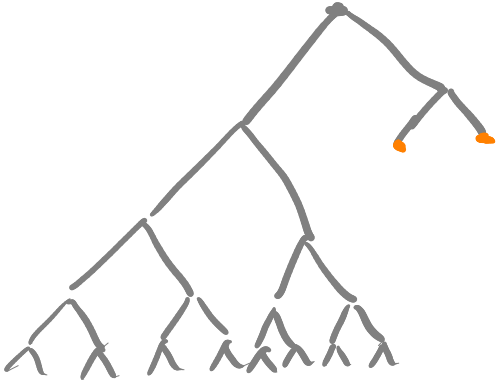
n-dimensional data



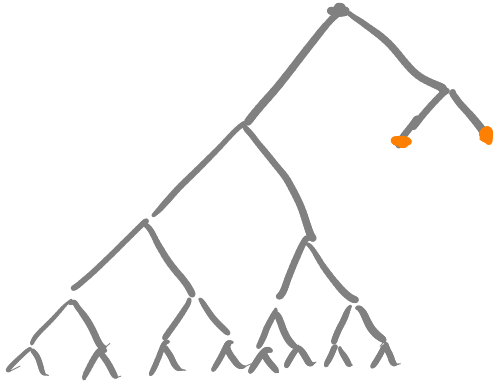
Isolation Forest

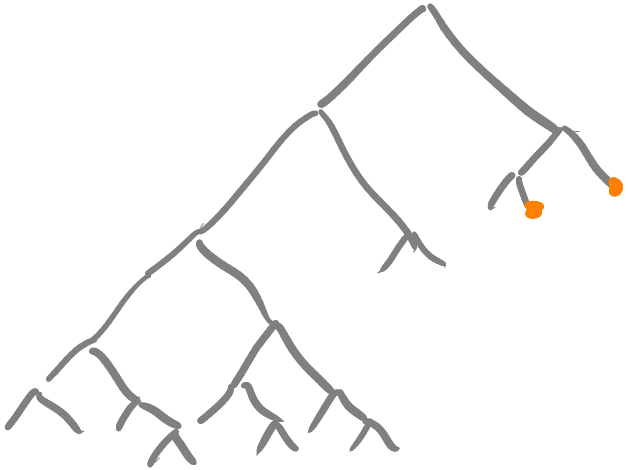


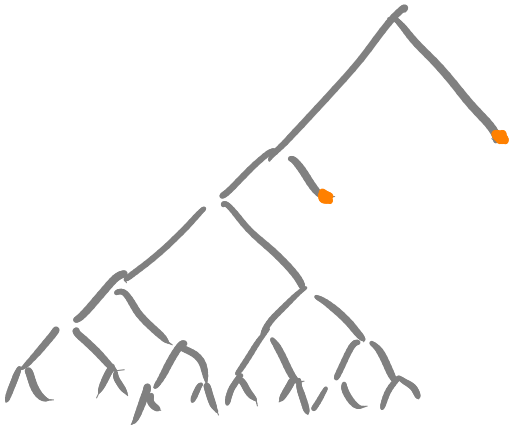




Isolation Forest



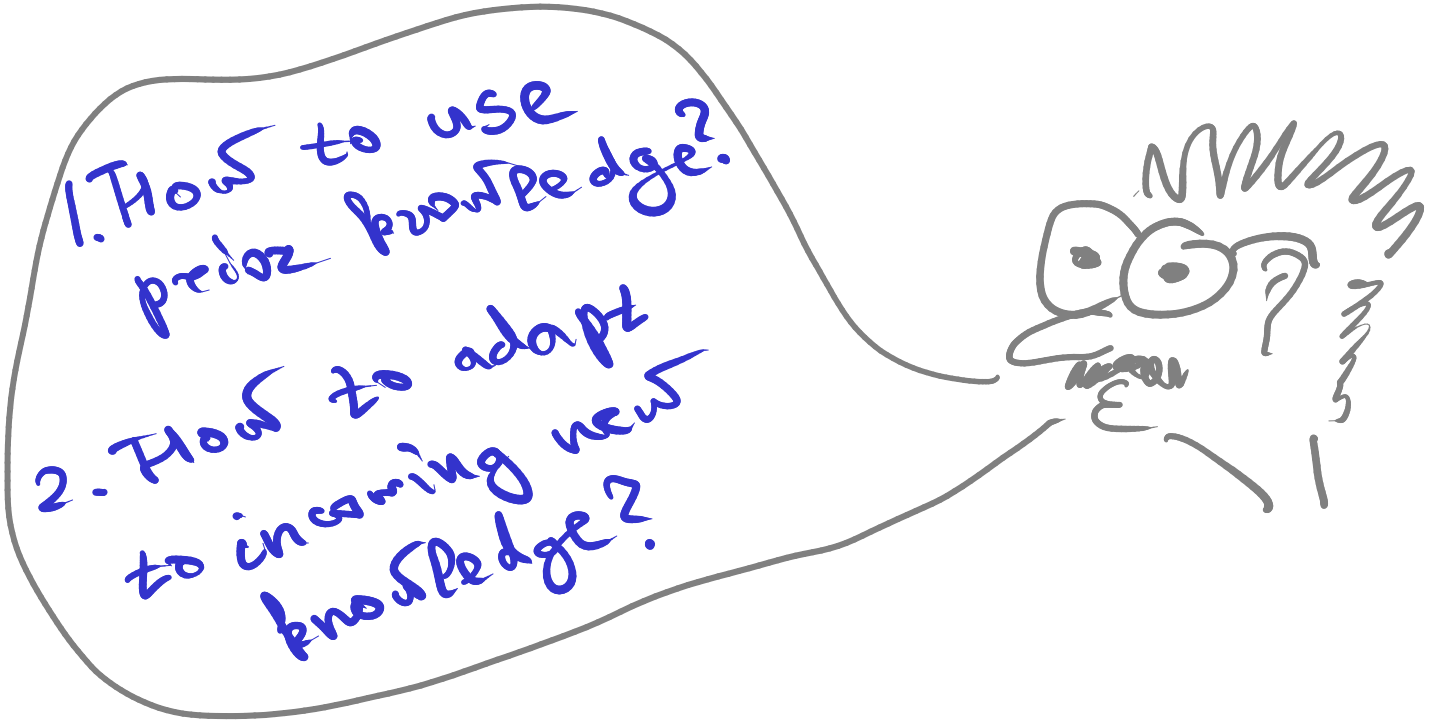




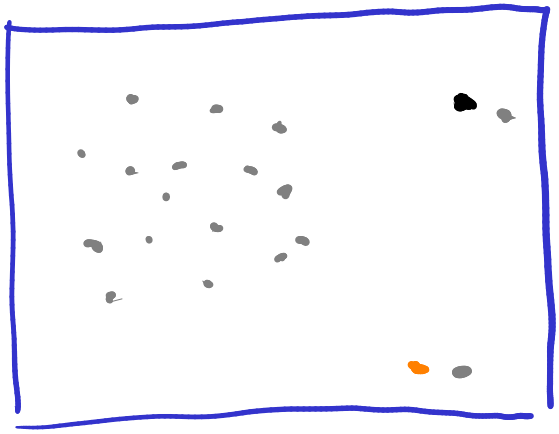
(. . .)

Isolation Forest

Any problems?



Pineforest



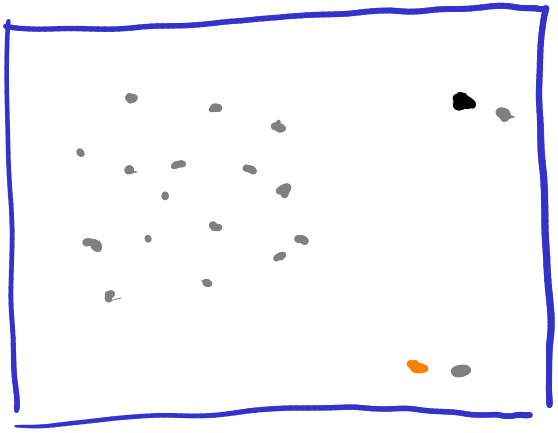
A. Let's score the trees

$$\lambda(t) = \sum_i y_i \, e(t, x_i)$$

y_i - label, $y_i = \begin{cases} -1, \text{false anomaly} \\ 0, \text{unknown} \\ +1, \text{true anomaly} \end{cases}$

The lower $\lambda(t)$ — the better.

Pineforest



A. Let's score the trees

$$\lambda(t) = \sum_i y_i \, e(t, x_i)$$

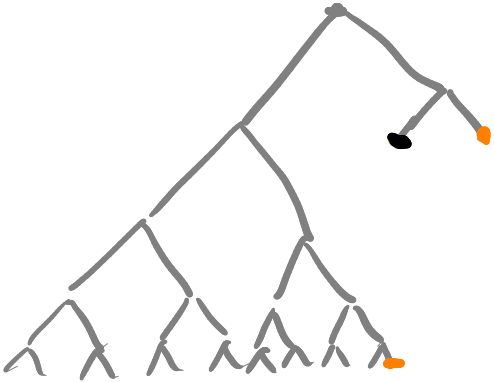
y_i - label, $y_i = \begin{cases} -1, \text{false anomaly} \\ 0, \text{unknown} \\ +1, \text{true anomaly} \end{cases}$

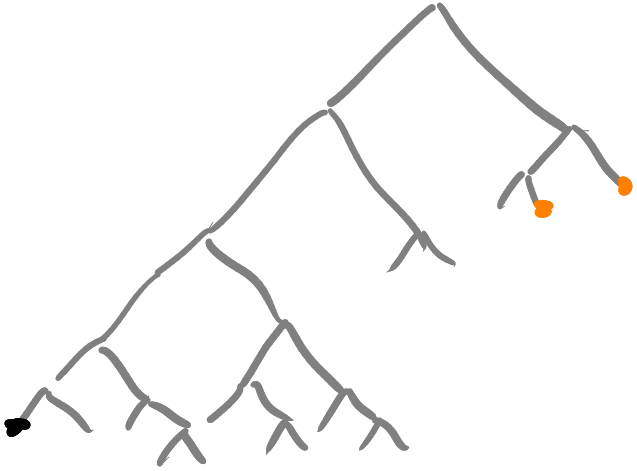
The lower $\lambda(t)$ — the better.

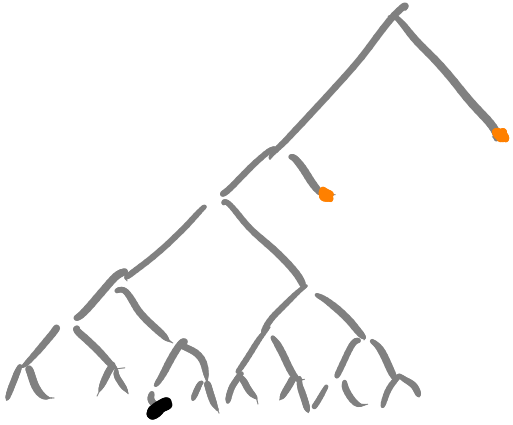
B. And filter bad trees out

$\lambda(t_1) = 3$, $\lambda(t_2) = 10$, $\lambda(t_3) = 80$

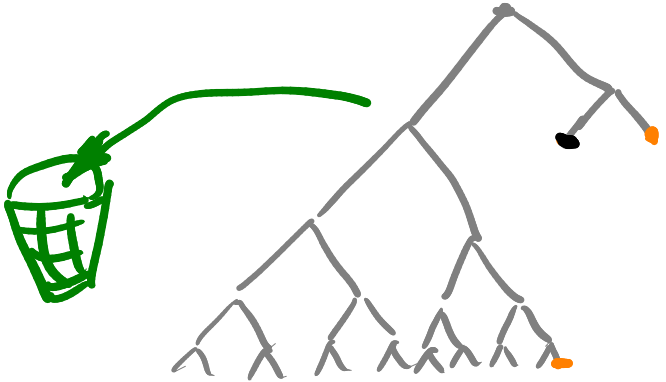
Pineforest

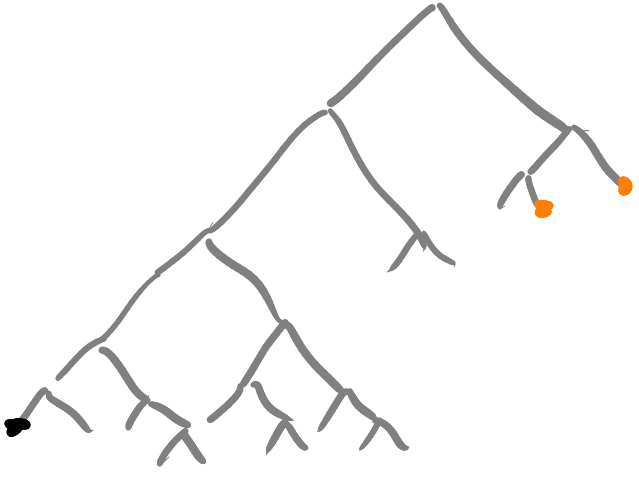


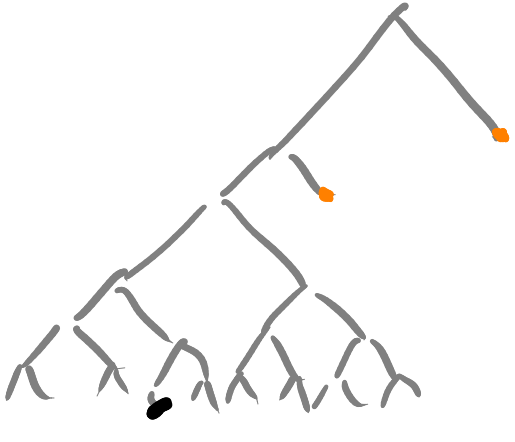




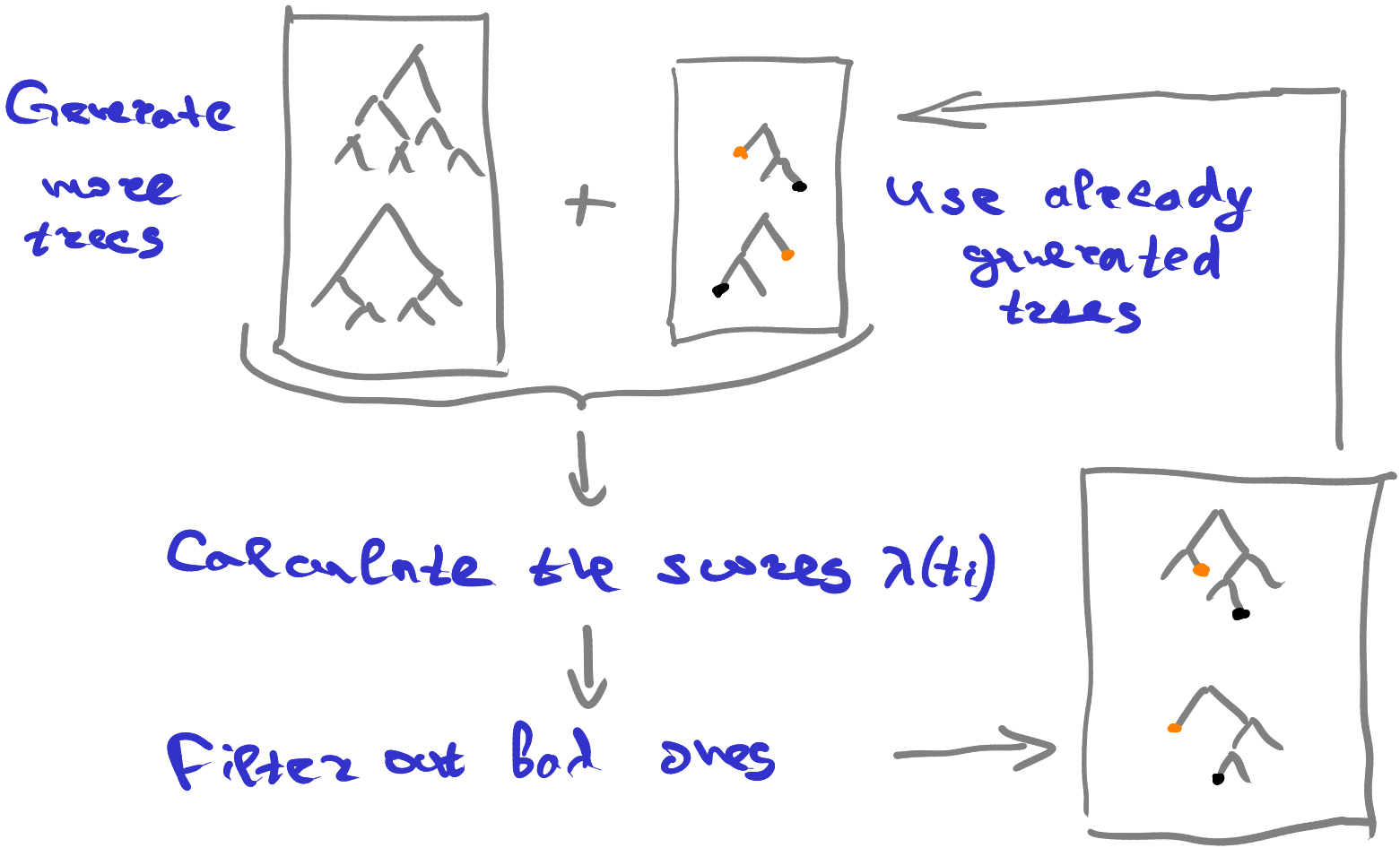
Pineforest







Pineforest


Generate more trees
+
Use already generated trees
Calculate the scores λ(ti)
Filter out bad ones

Pineforest

- [x] Can we use priors?
- [x] Can we adopt to new data?