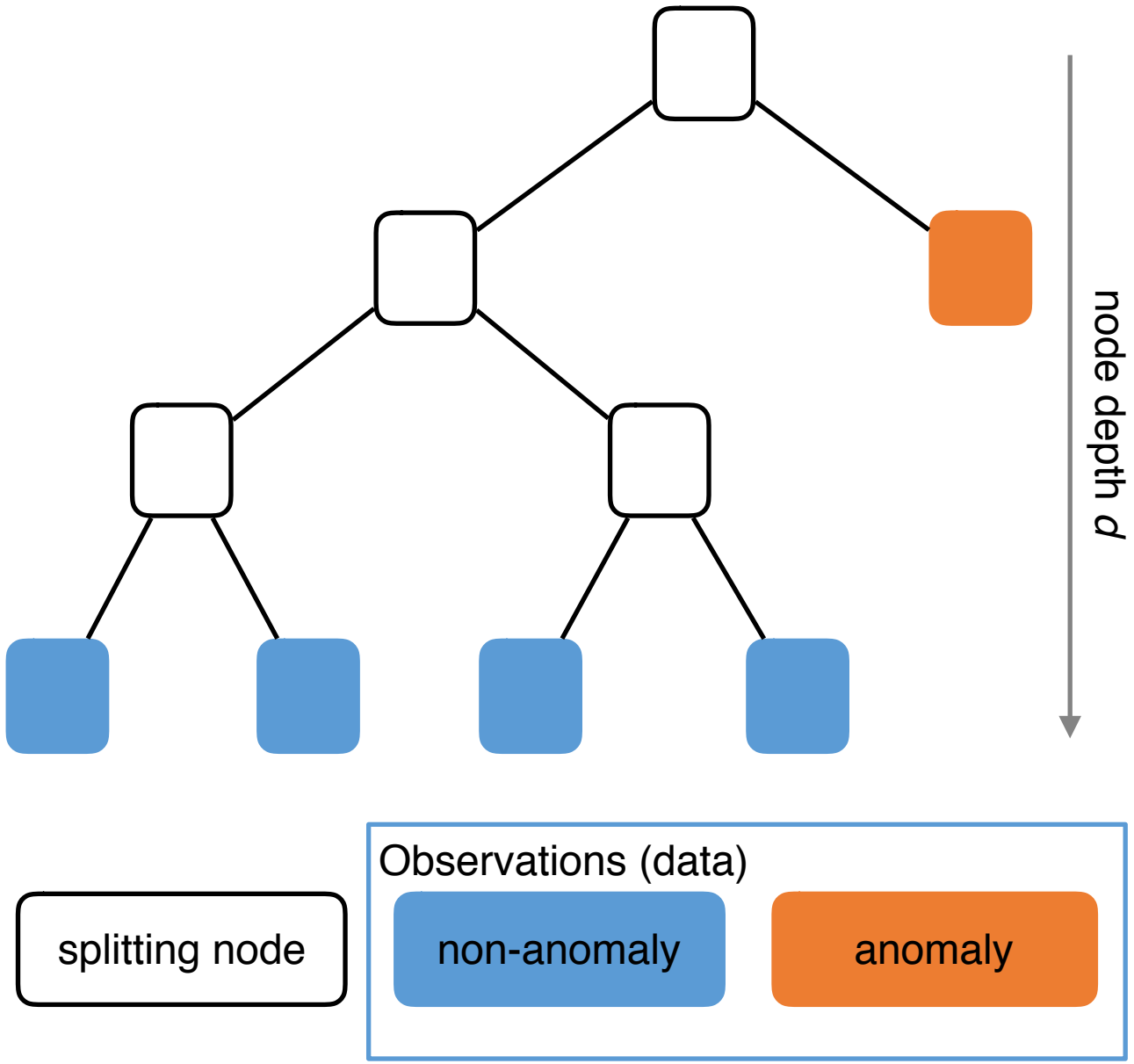
Isolation Forest

H2O Meetup@Amsterdam, Dec 5 2018

Václav Belák, Ph.D.
Data Scientist
vaclav@h2o.ai

H2O.ai

Building an Isolation Tree by Random Splitting

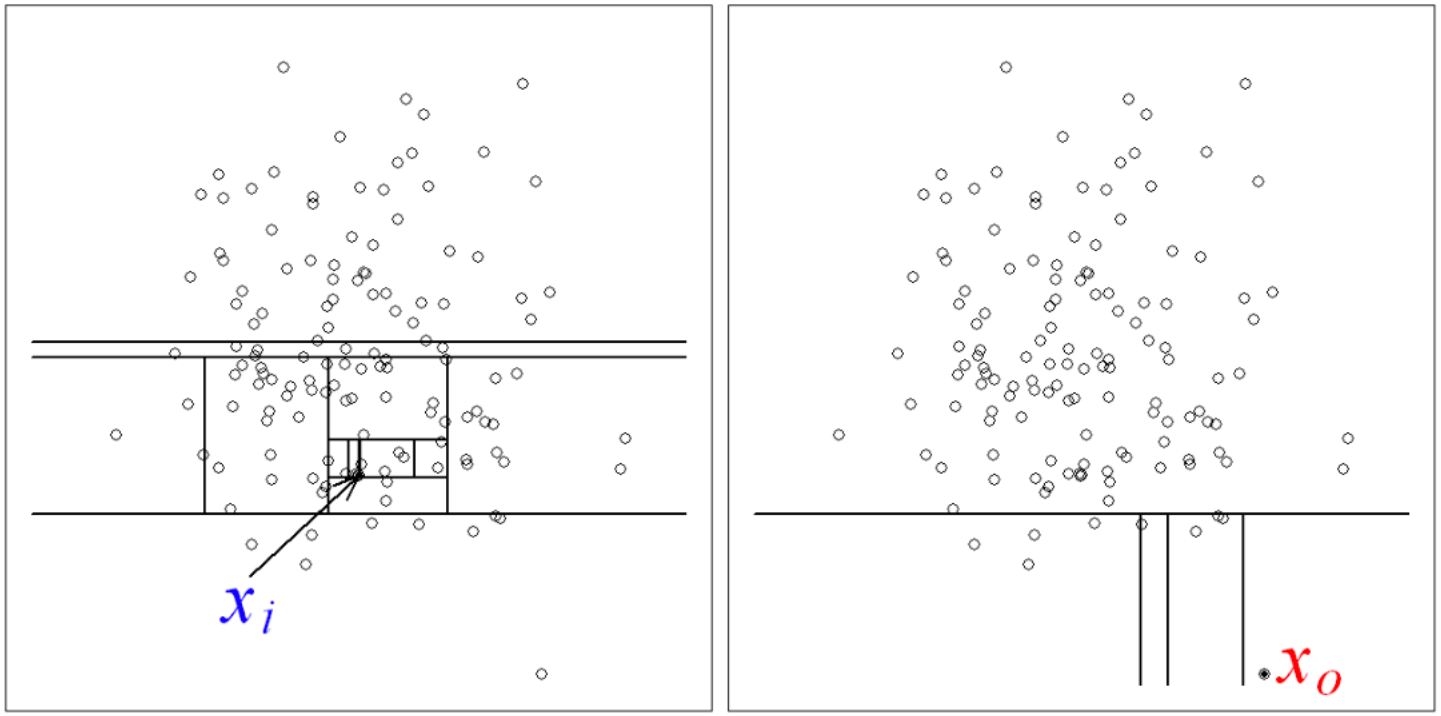


Recursive partitioning:

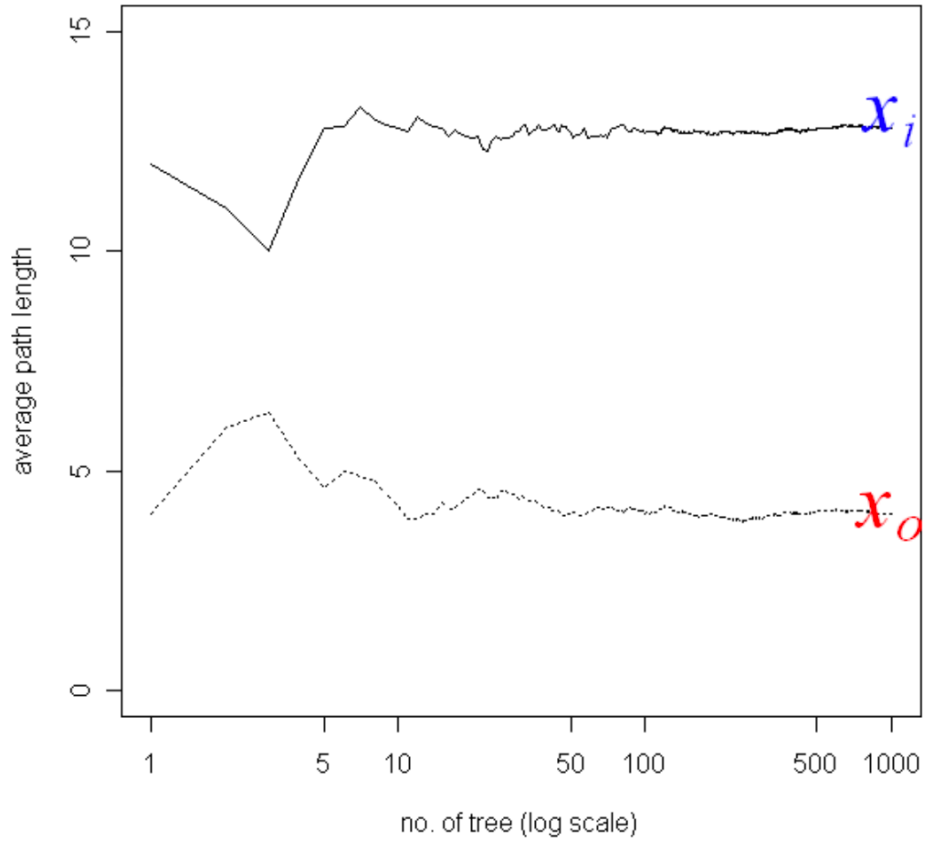
1. Choose randomly a feature *f*
2. Split the data by a random threshold *t* within *[min(f), max(f)]*
3. Stop when:
   - Tree depth limit reached
   - All samples have the same values
   - No data left to be partitioned

Intuition: Anomalies should be well separated from the regular observations which corresponds to a small depth

Mean Depth of Nodes Converges in Isolation Forest



(a) Isolating x_i (b) Isolating x_o

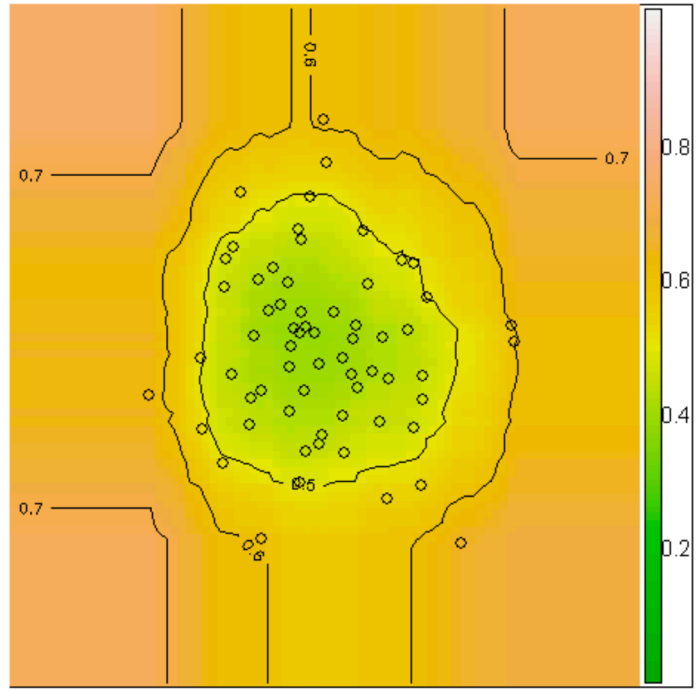


Source: Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation forest." *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008.

H2O.ai

How do I decide if the observation *x* is an anomaly?

anomaly score	Is anomaly?
close to 1	highly likely
much smaller than 0.5	highly unlikely
0.5 for all observations	data does not exhibit distinct anomalies



Source: Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation forest." *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008.

Why?

- Isolation tree is structurally the same as binary search tree (BST)
- Average depth of observation $E(d(x))$ is therefore the same as average path of unsuccessful search in BST
- This can be approximated as $c(n) = 2H(n-1) - (2(n-1)/n)$, where $H(i)$ is the harmonic number and n the number of observations
- This can be used then to normalise the average depth of an observation in the Isolation Forest
- The **anomaly score** can be then defined as

$$s(x, n) = 2^{-\frac{E(d(x))}{c(n)}}$$

Isolation Forest in H2O

Load Data

```
import h2o
h2o.init()

df = h2o.import_file("creditcard.csv")
```

Training Isolation Forest

```
seed = 12345
ntrees = 100
isoforest = h2o.estimators.H2OIsolationForestEstimator(
    ntrees=ntrees, seed=seed)
isoforest.train(x=df.col_names[0:31], training_frame=df)
predictions = isoforest.predict(df)

predictions
```

predict	mean_length
0.0284238	6.82

H2O.ai