Аномалии в потоковых данных

алгоритмы под капотом

Обо мне

- Старший специалист по машинному обучению
- deep learning engineer
- NLP, CV, anomaly detection
- Open source contributor
- Амбассадор Яндекс Практикума

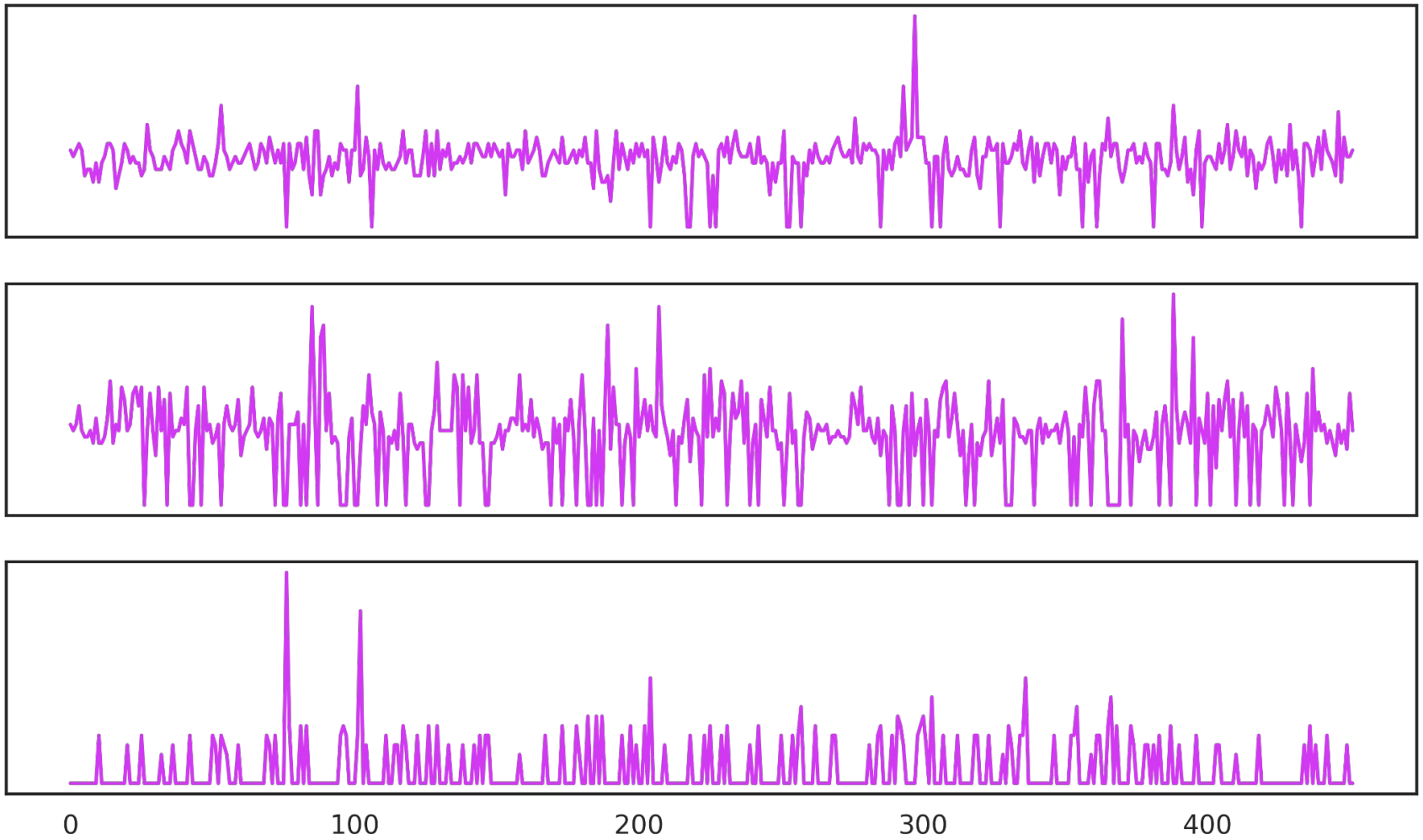
Аномалии


SCAMMER

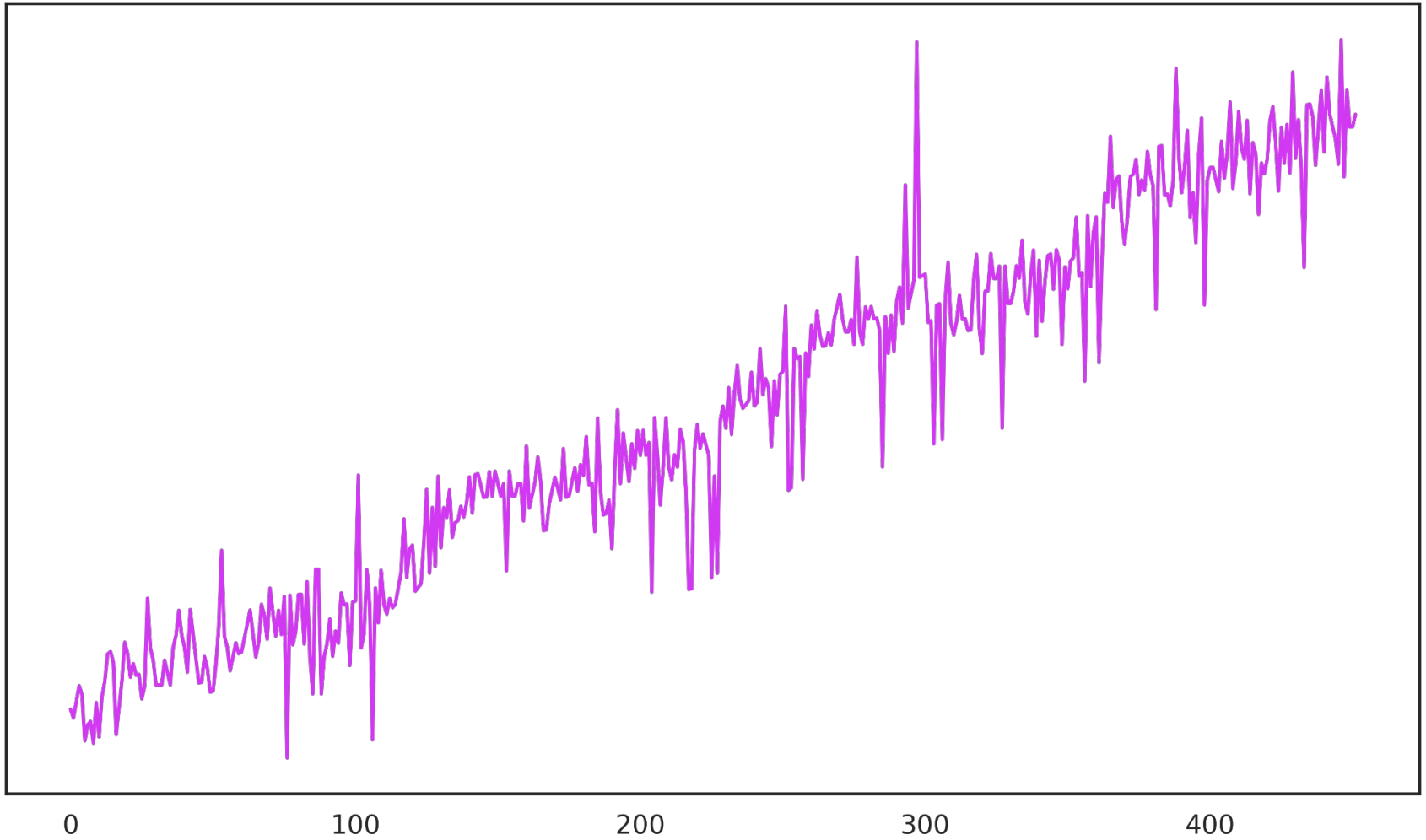
попугаи	**удавы**
3.248357	-2.874754
2.930868	-2.826776
3.323844	-3.340012
3.761515	-2.883873
2.882923	-2.853464
2.882932	-3.357176
3.789606	-2.067113
…	…

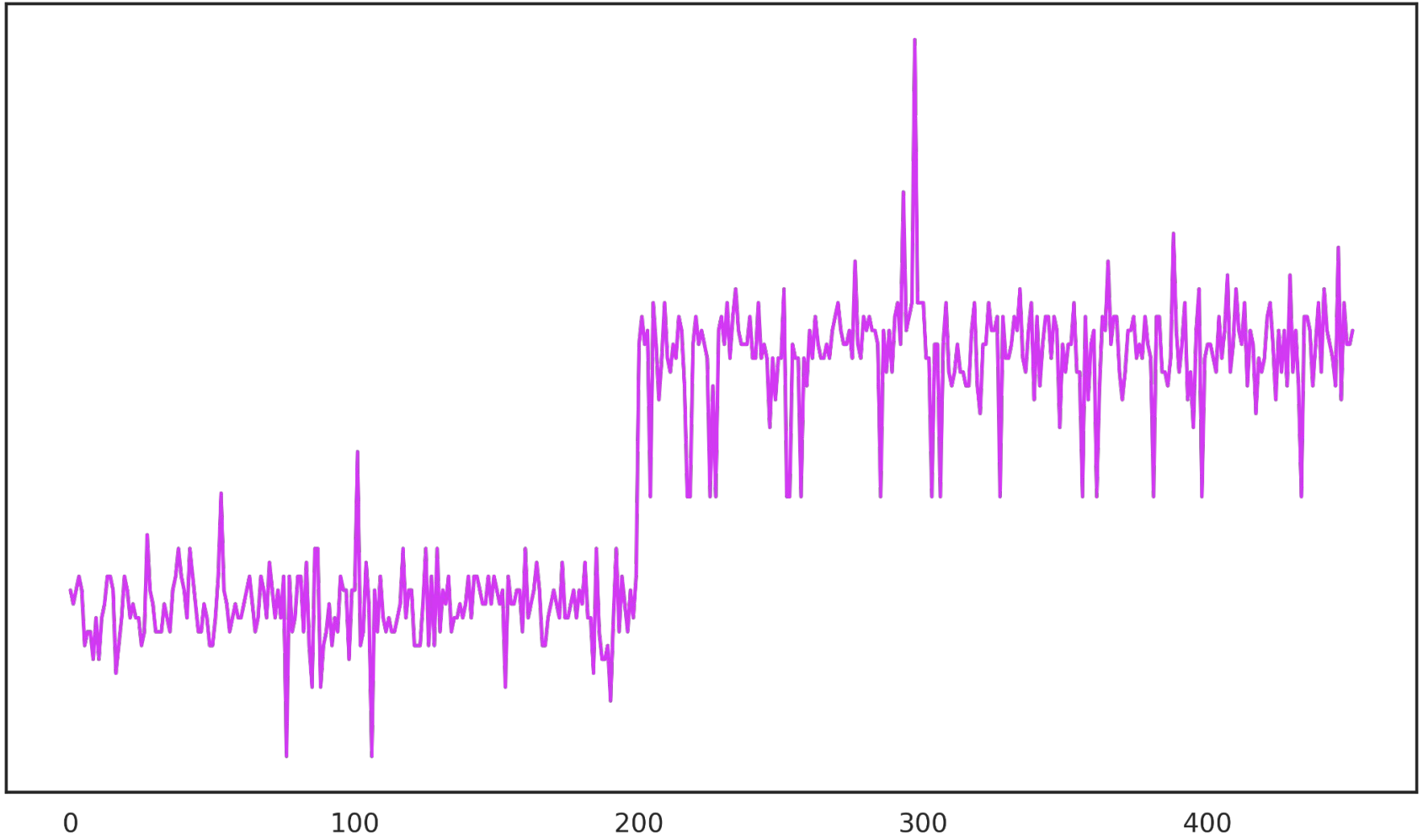
попугаи	**удавы**
3.248357	-2.874754
2.930868	-2.826776
3.323844	-3.340012
3.761515	-2.883873
2.882923	-2.853464
2.882932	-3.357176
3.789606	-2.067113
…	…

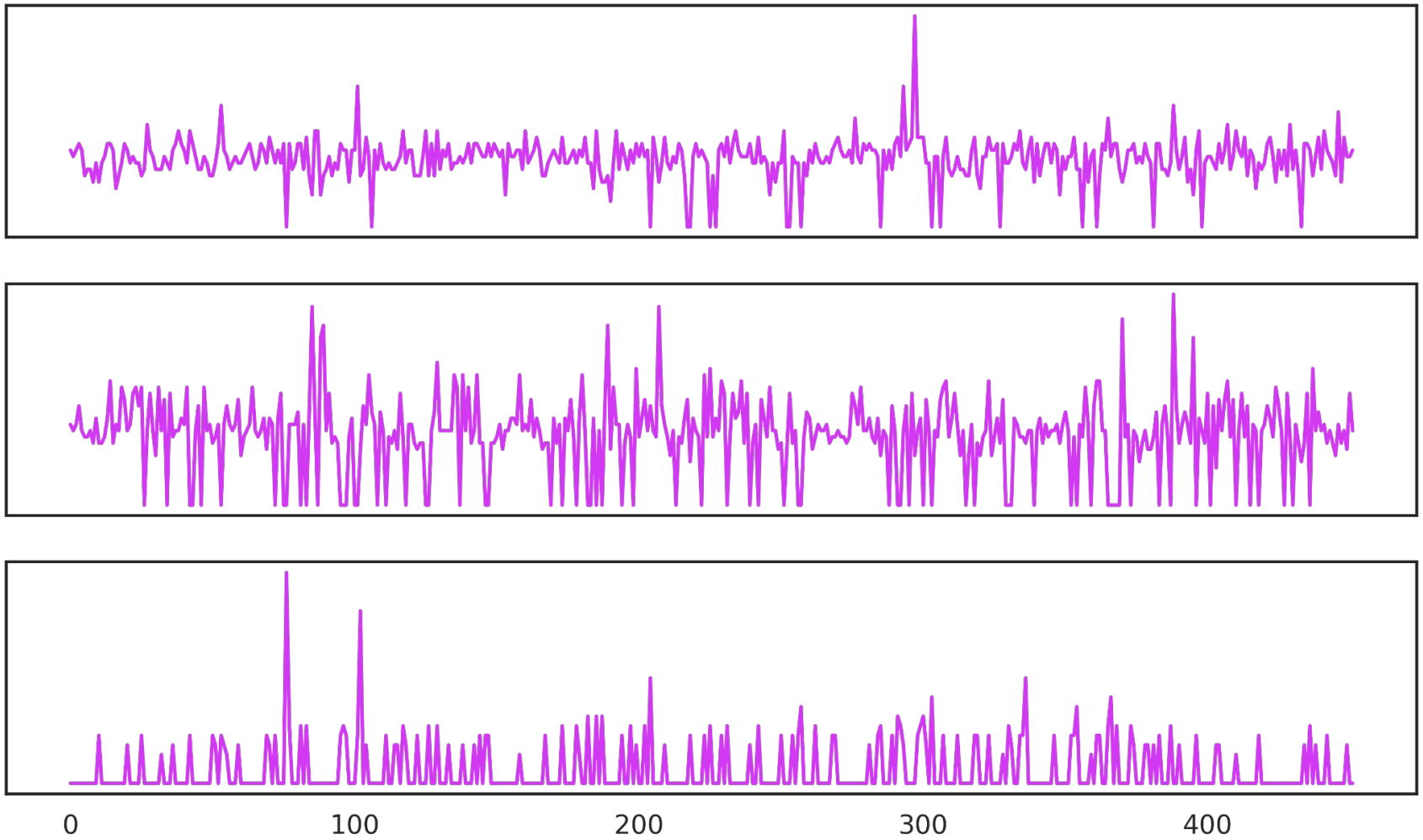
индекс аномальности
-0.119150
-0.131275
-0.108670
-0.065472
-0.128972
-0.120056
-0.012170
…

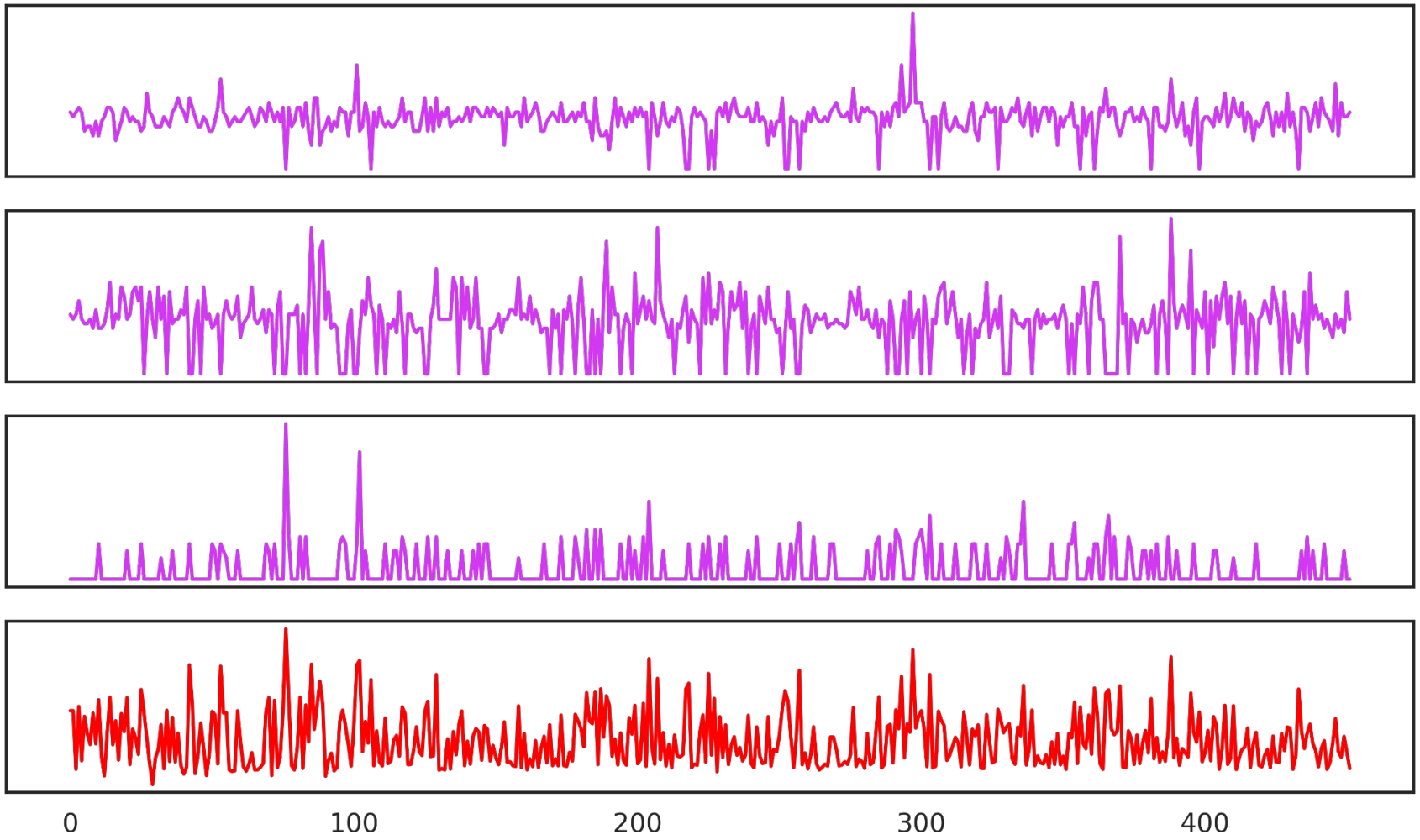

0
100
200
300
400

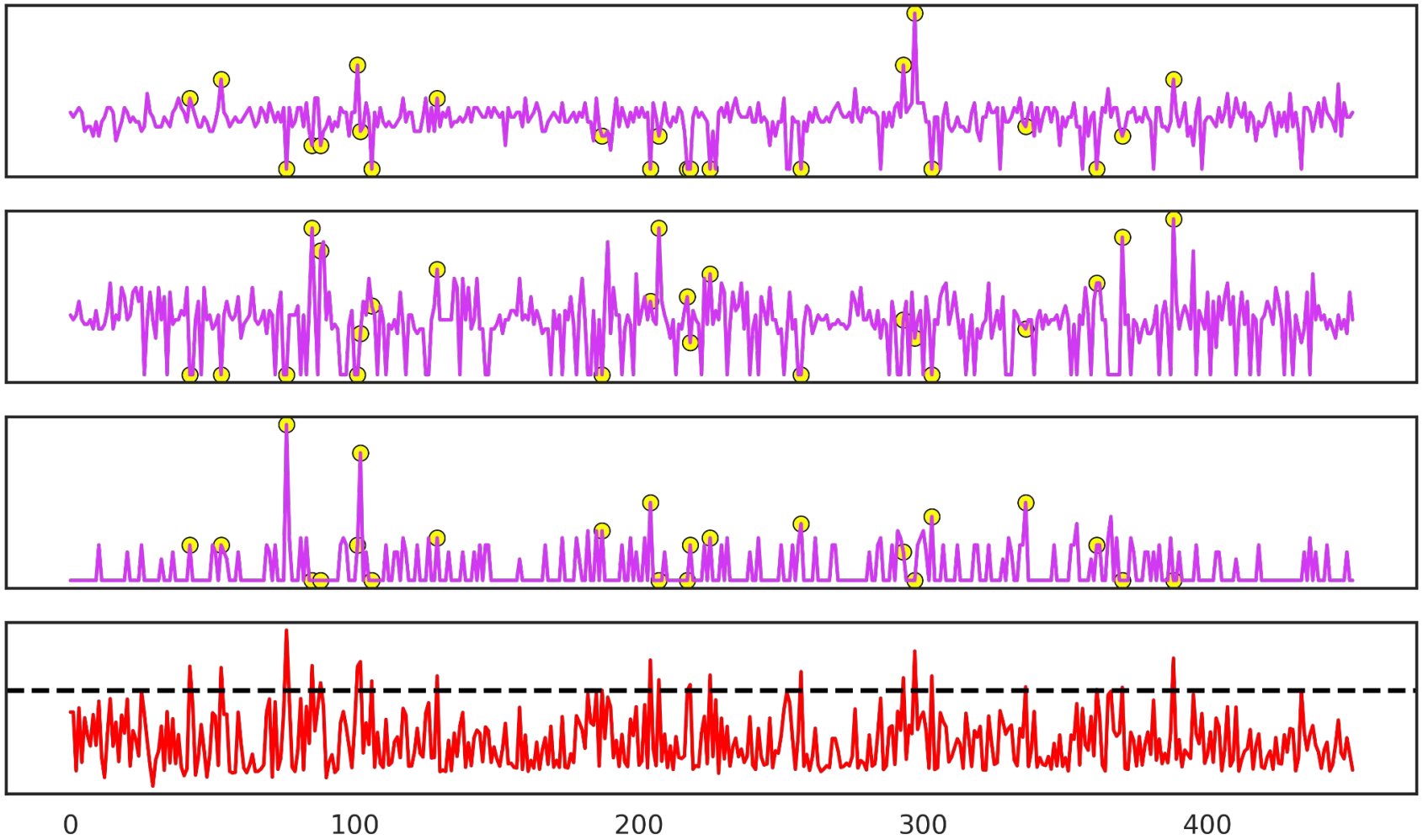
Threshold


0
100
200
300
400


0
100
200
300
400


0
100
200
300
400


0
100
200
300
400


0
100
200
300
400

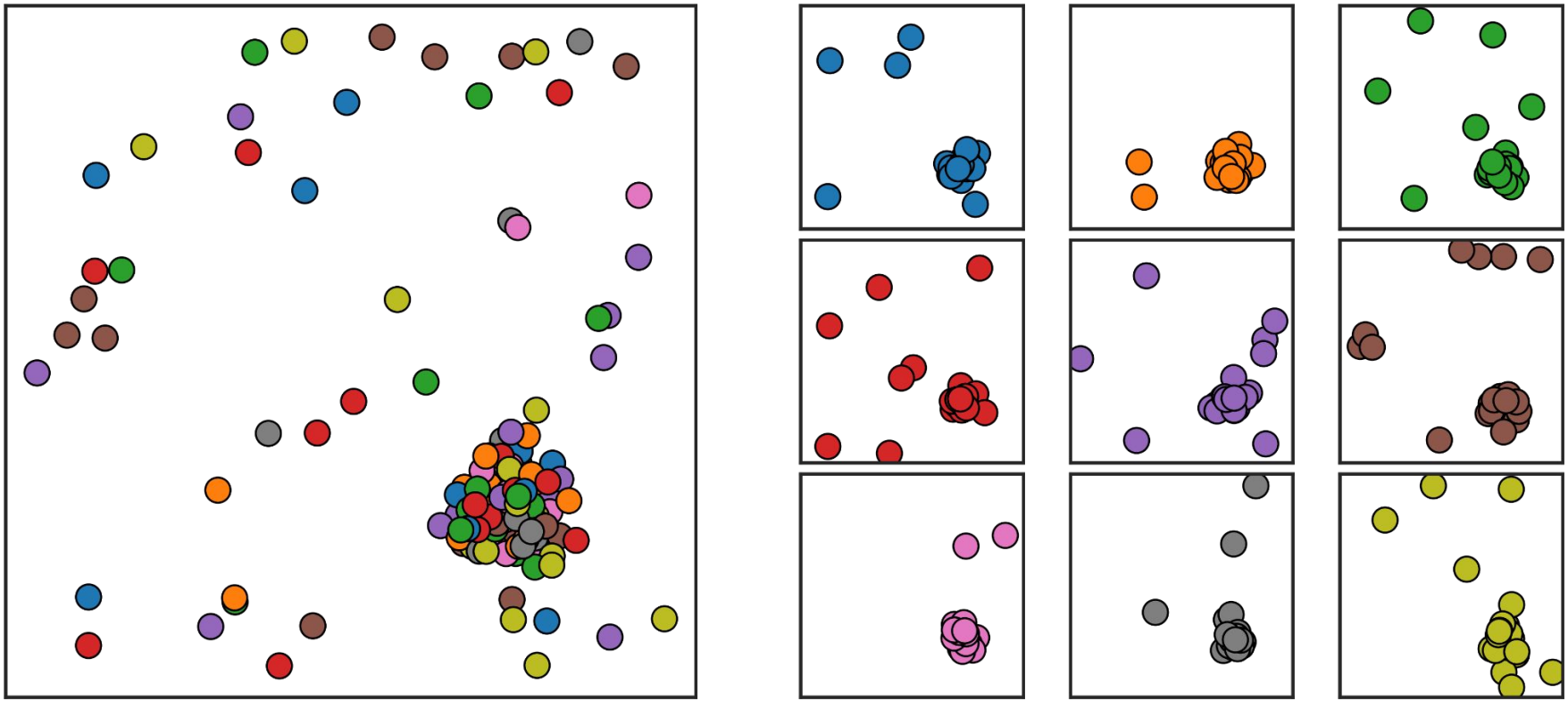
Потоковые данные

pip install pysad

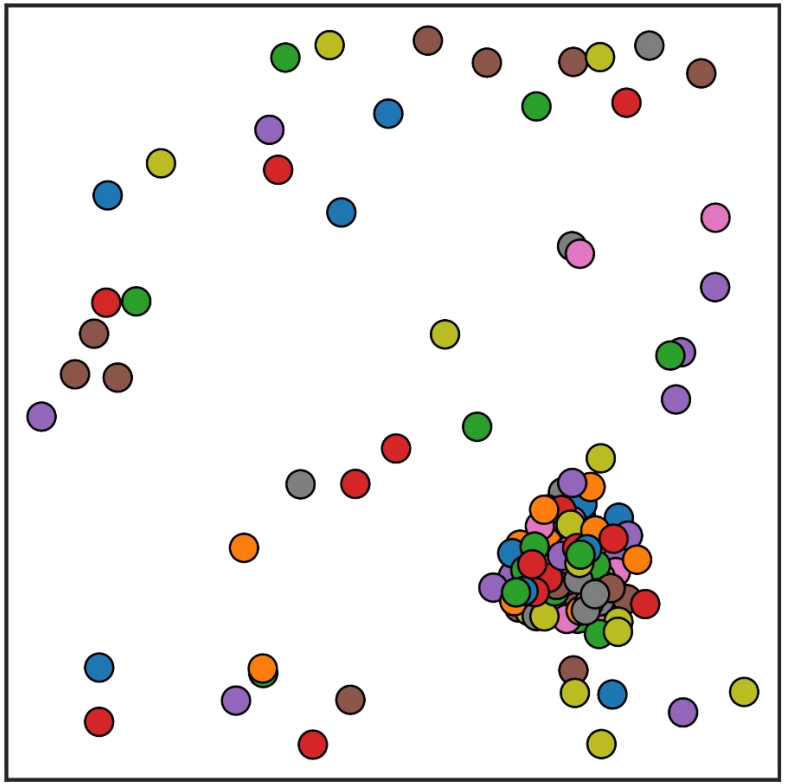
```
from pysad.models import IForestASD

model = IForestASD()

for x in streaming_data:
    anomaly_score = model.fit_score_partial(x)
```

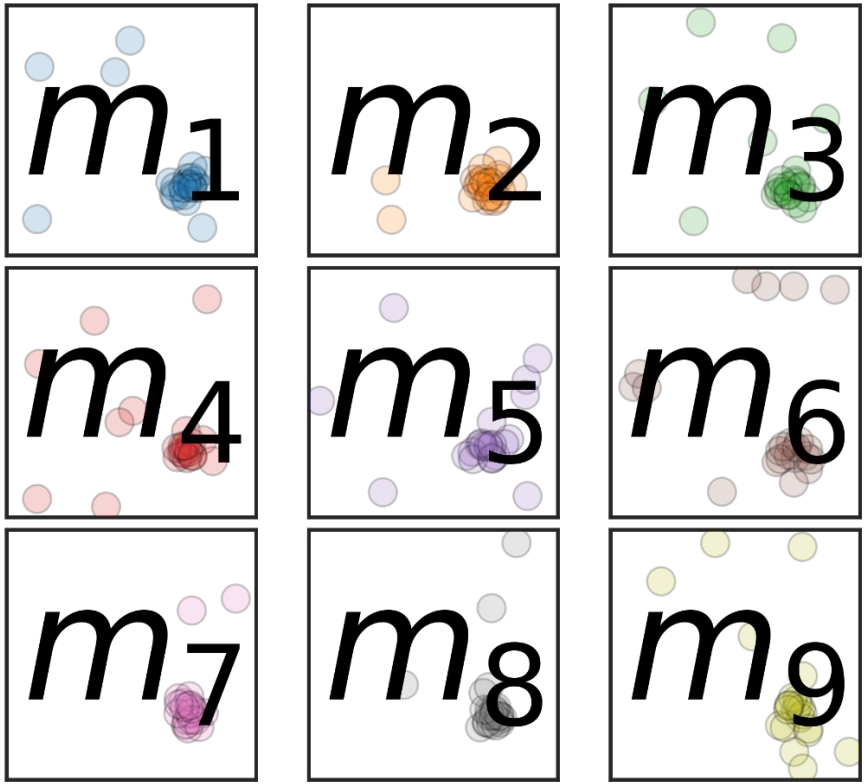
Ансамбли

Делим

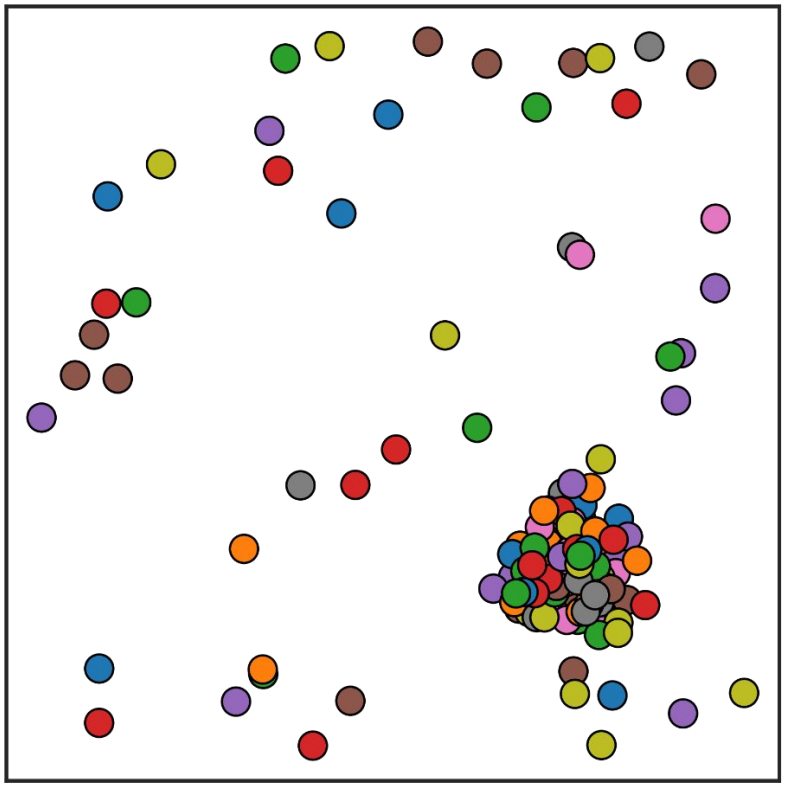


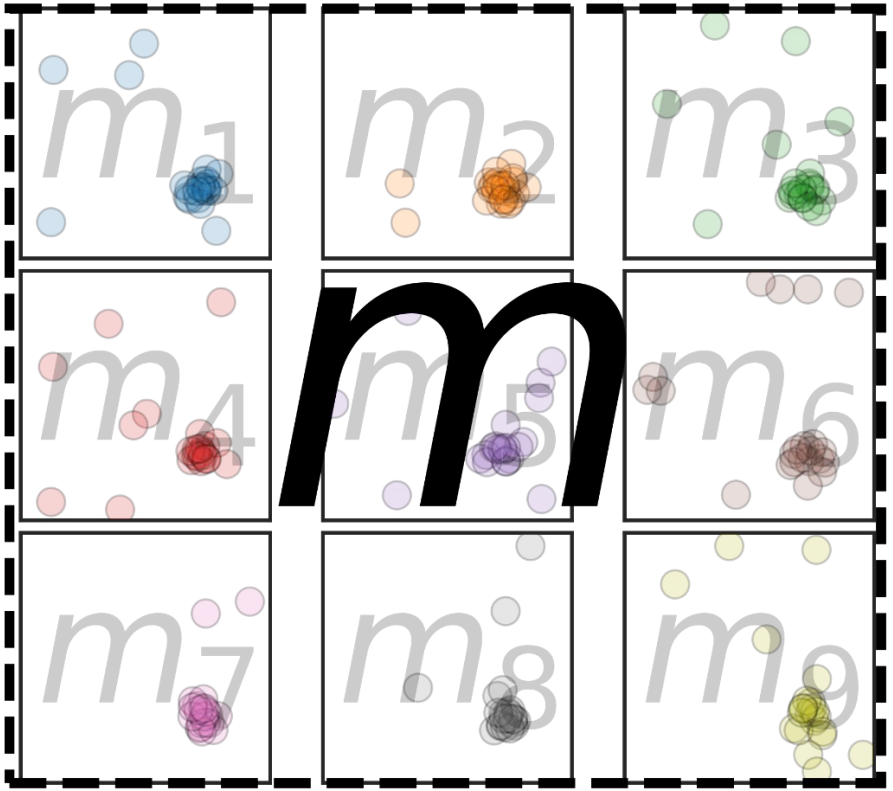
Обучаем

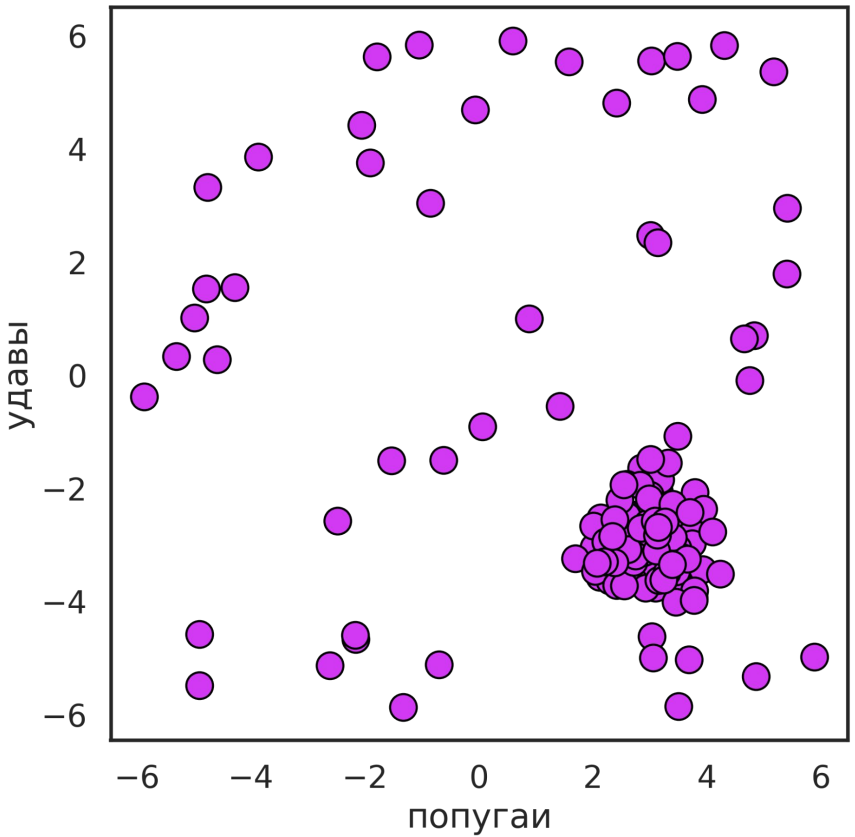


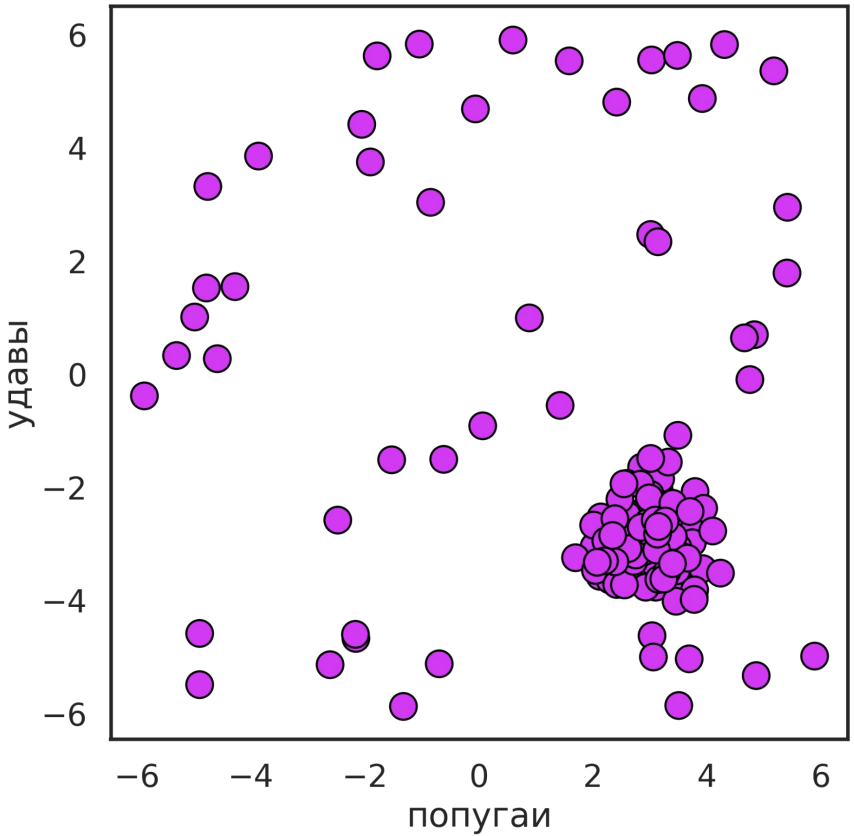


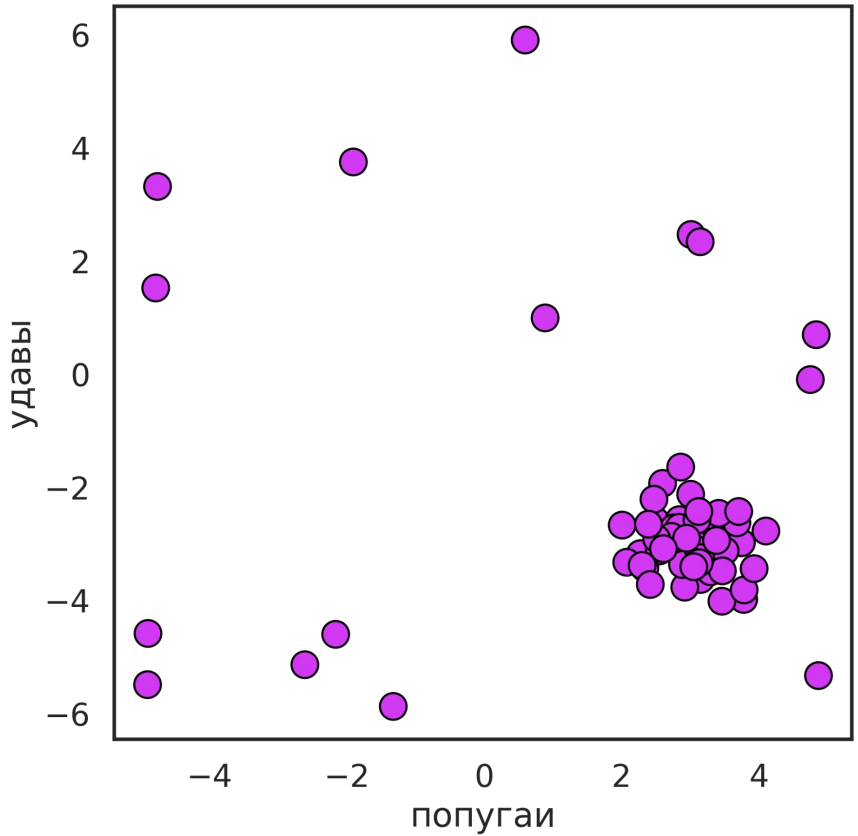
Усредняем



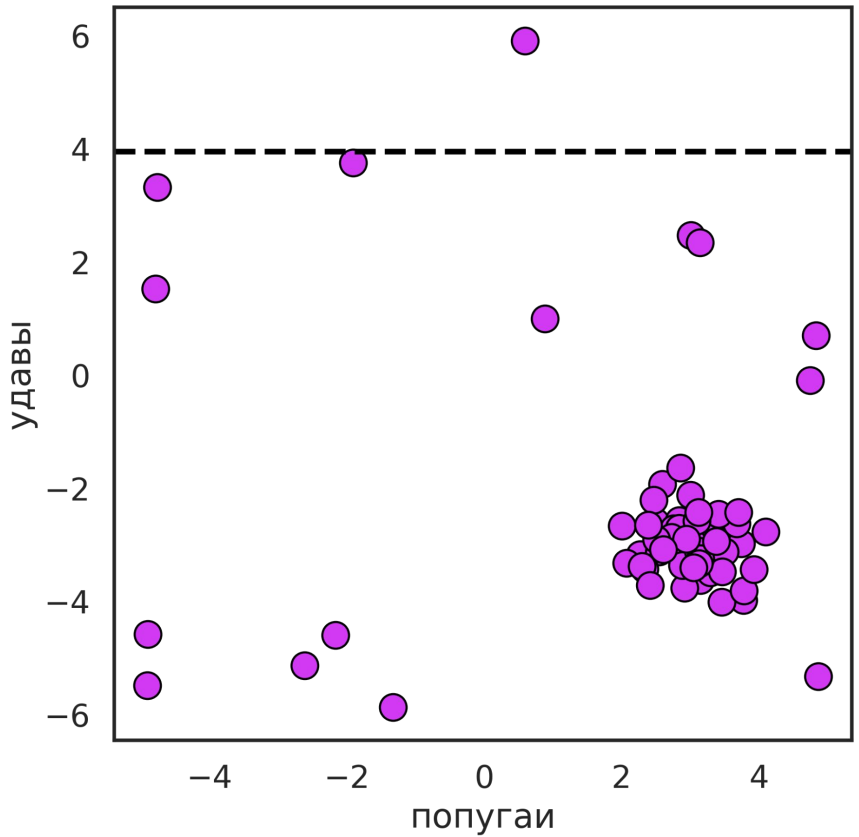



6
4
2
0
−2
−4
−6
удавы
−6
−4
−2
0
2
4
6
попугаи


6
4
2
0
−2
−4
−6
удавы
−6
−4
−2
0
2
4
6
попугаи


6
4
2
0
−2
−4
−6
удавы
−4
−2
0
2
4
попугаи

```
In [1]: import random

In [2]: axes = ['попугаи', 'удавы']

In [3]: random.choice(axes)

Out[3]: 'удавы'

In [4]: random.uniform(df['удавы'].min(),
df['удавы'].max())

Out[4]: 3.954197818641566
```



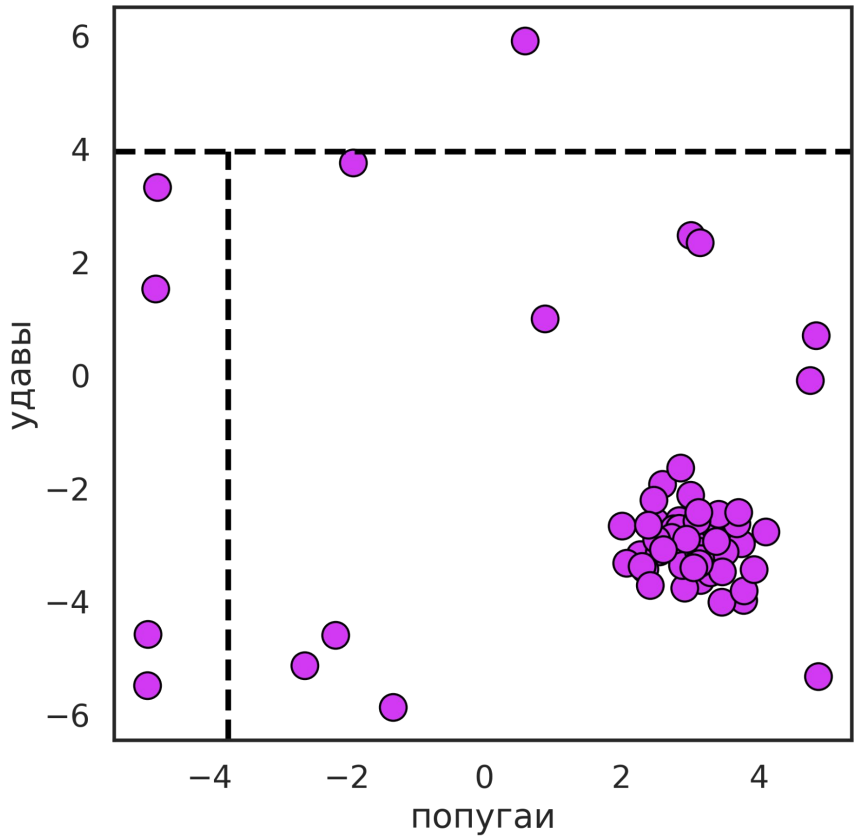
…

In [5]: random.choice(axes)

Out[5]: 'попугаи'

In [6]: random.uniform(df_b['попугаи'].min(), df_b['попугаи'].max())

Out[6]: -3.7345546743319455



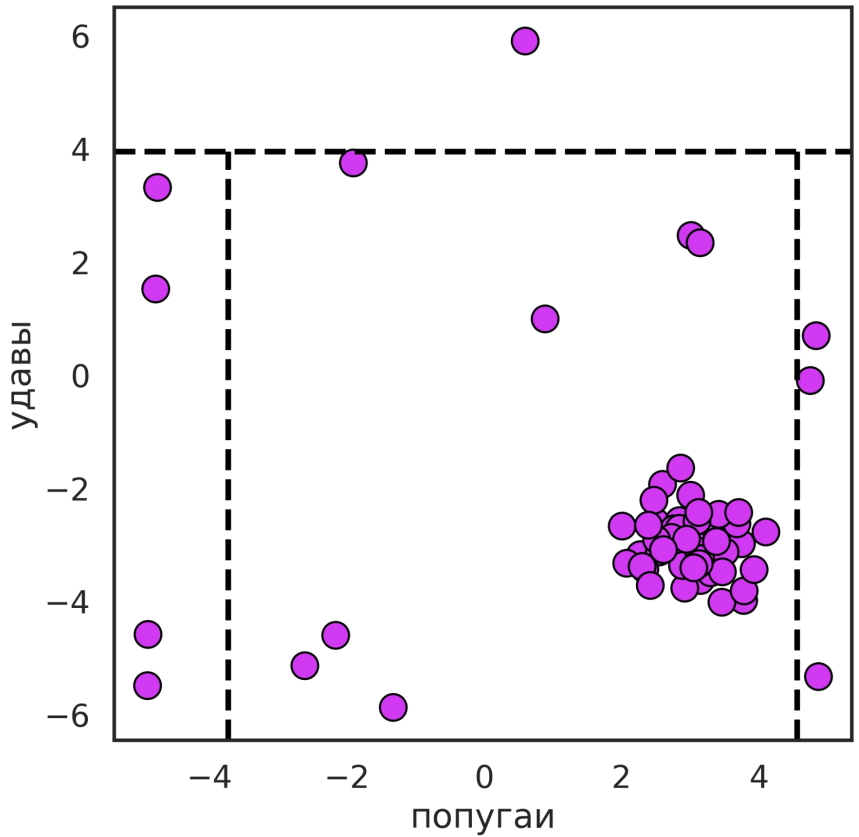
…

In [7]: random.choice(axes)

Out[7]: 'попугаи'

In [8]: random.uniform(df_bl['попугаи'].min(), df_bl['попугаи'].max())

Out[8]: 4.55352143693694



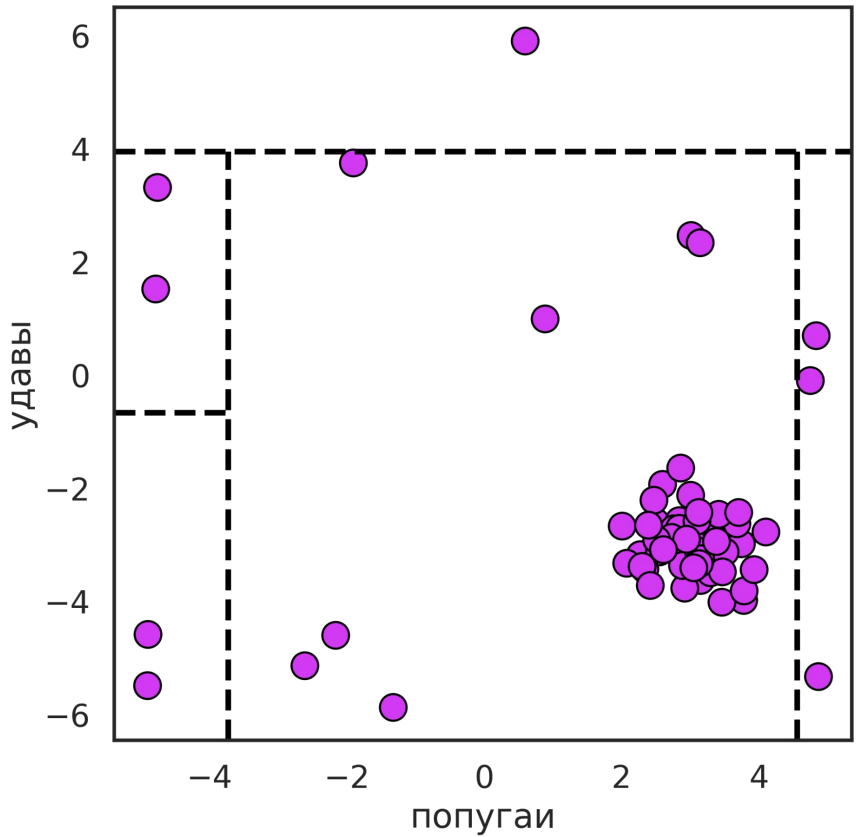
…

In [9]: random.choice(axes)

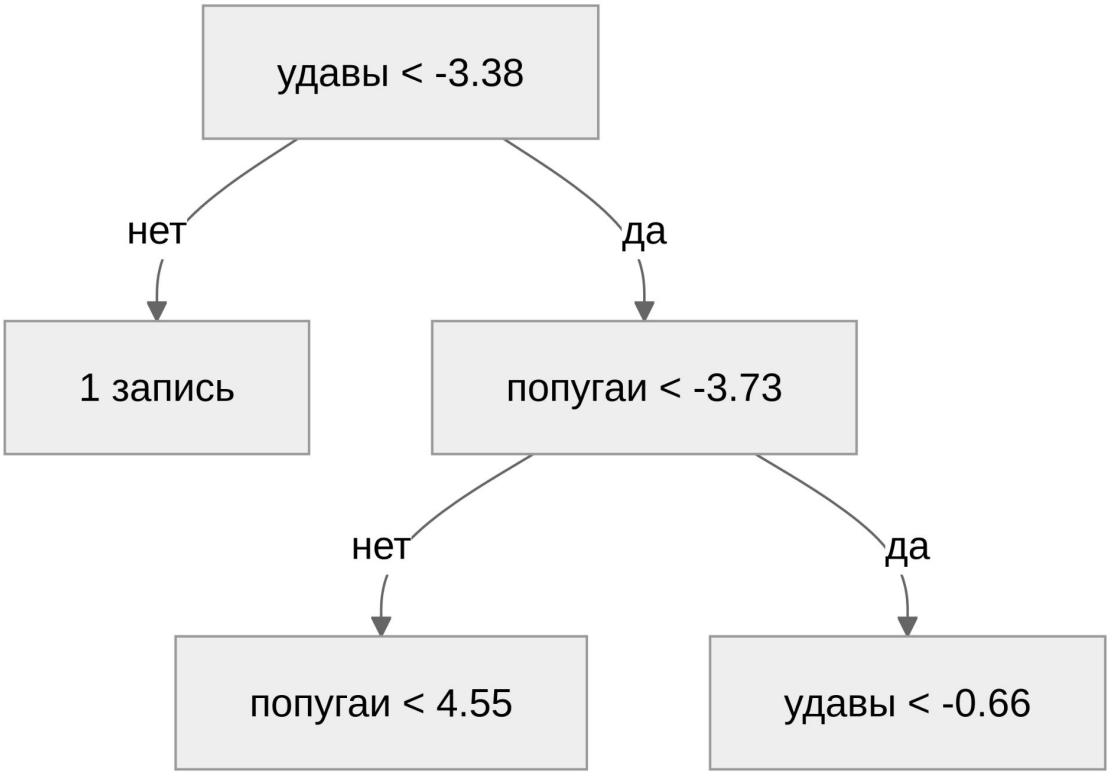
Out[9]: 'удавы'

In [10]: random.uniform(df_br['удавы'].min(), df_br['удавы'].max())

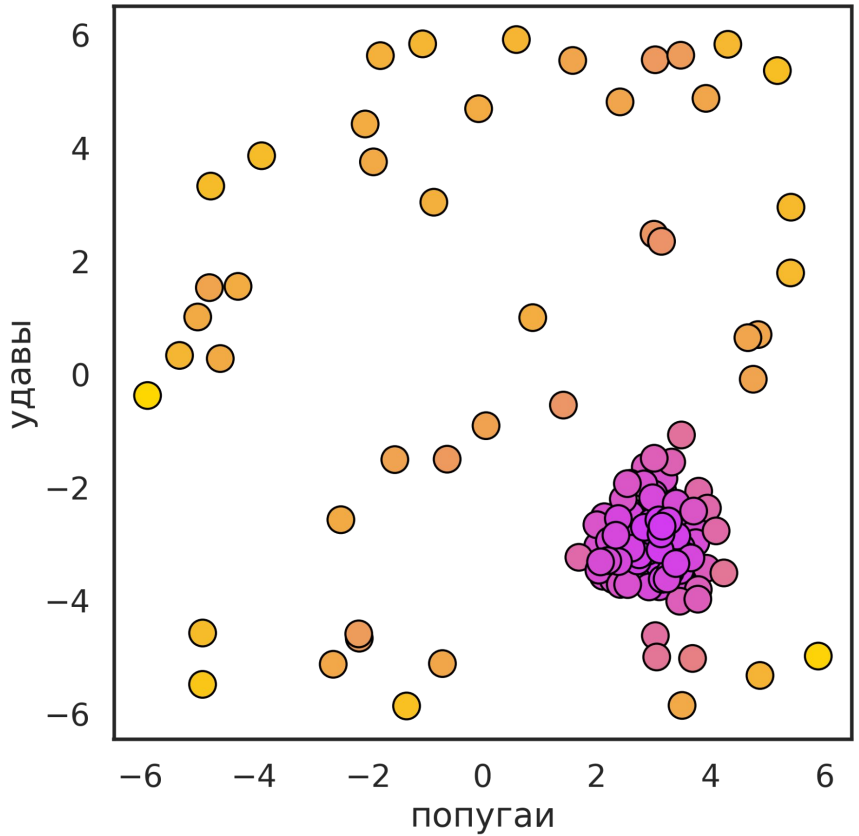
Out[10]: -0.6572094533790986



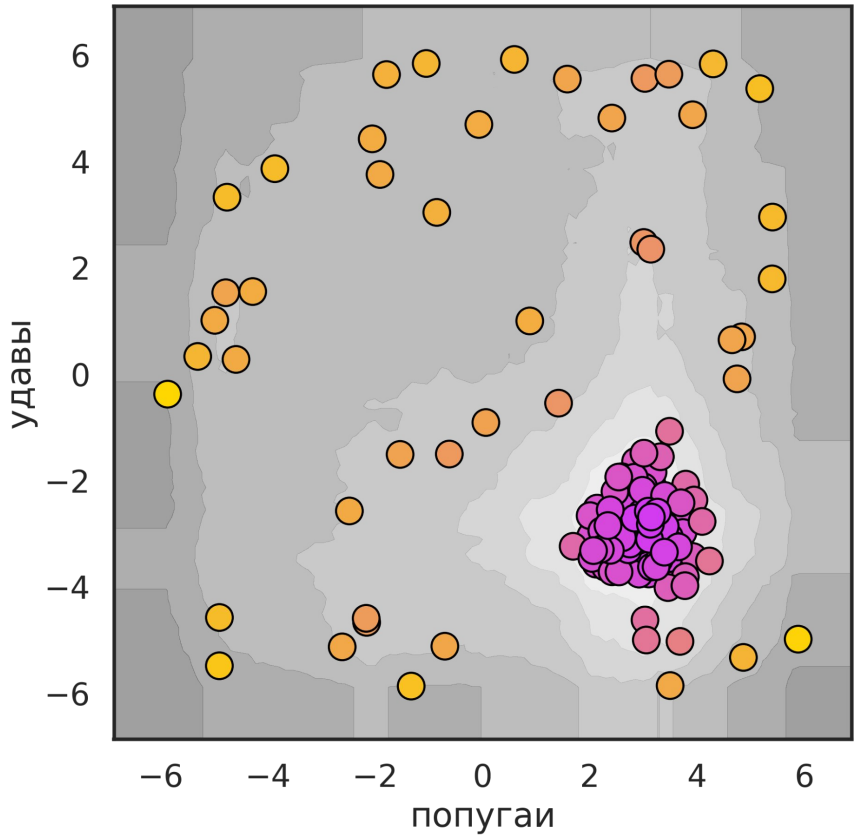
iTree


удавы < -3.38
нет
да
1 запись
попугаи < -3.73
нет
да
попугаи < 4.55
удавы < -0.66

Isolation Forest



Isolation Forest

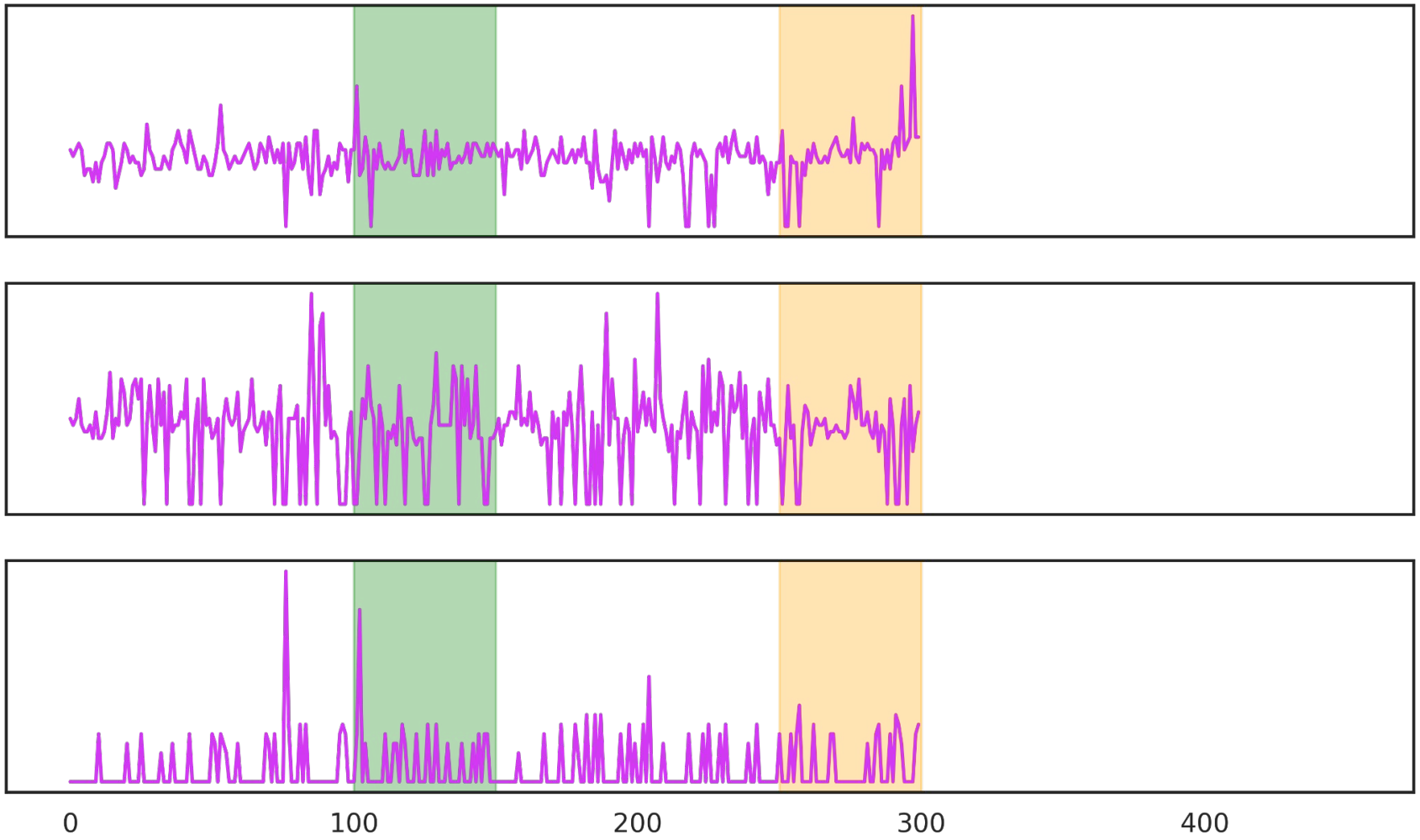


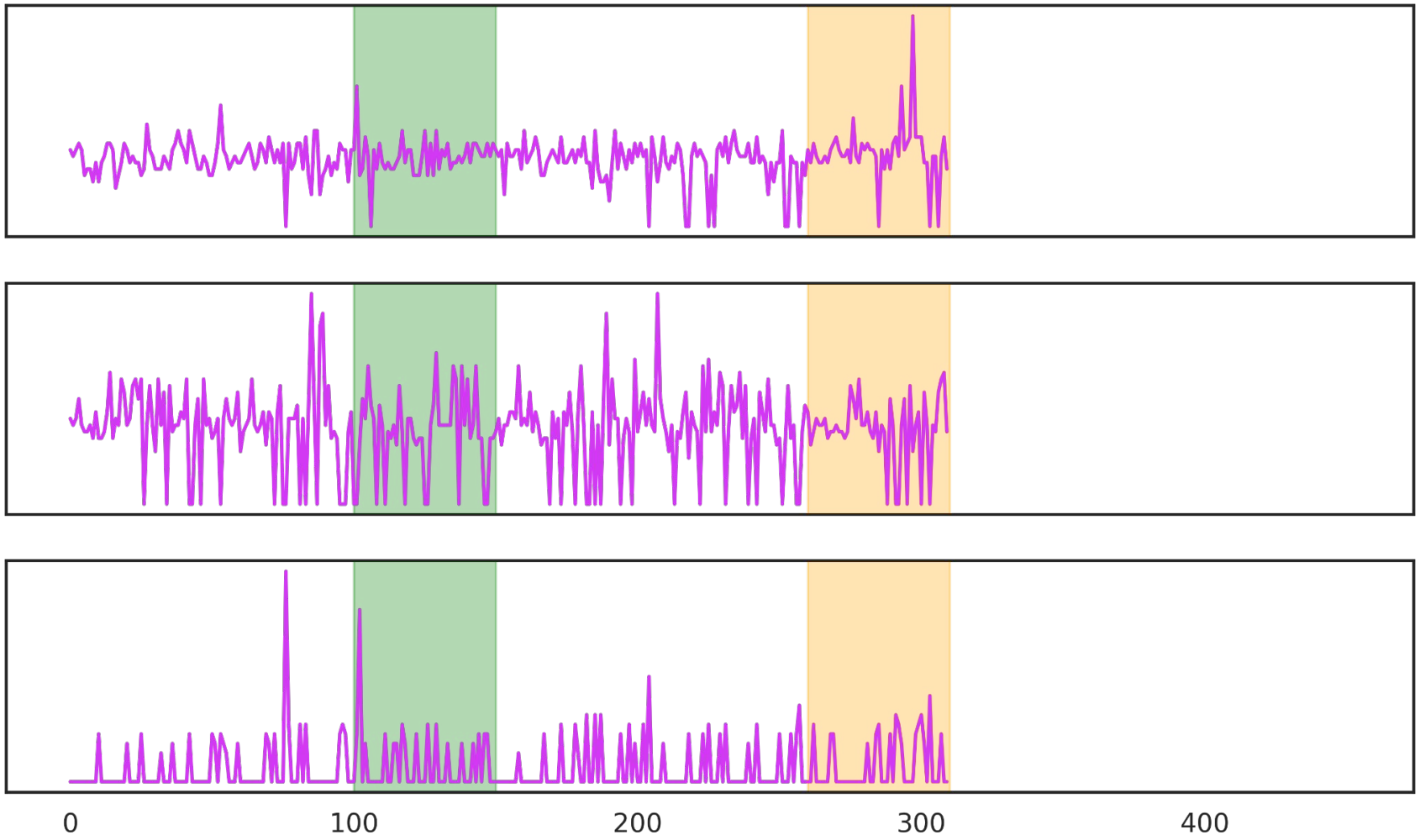
Алгоритм

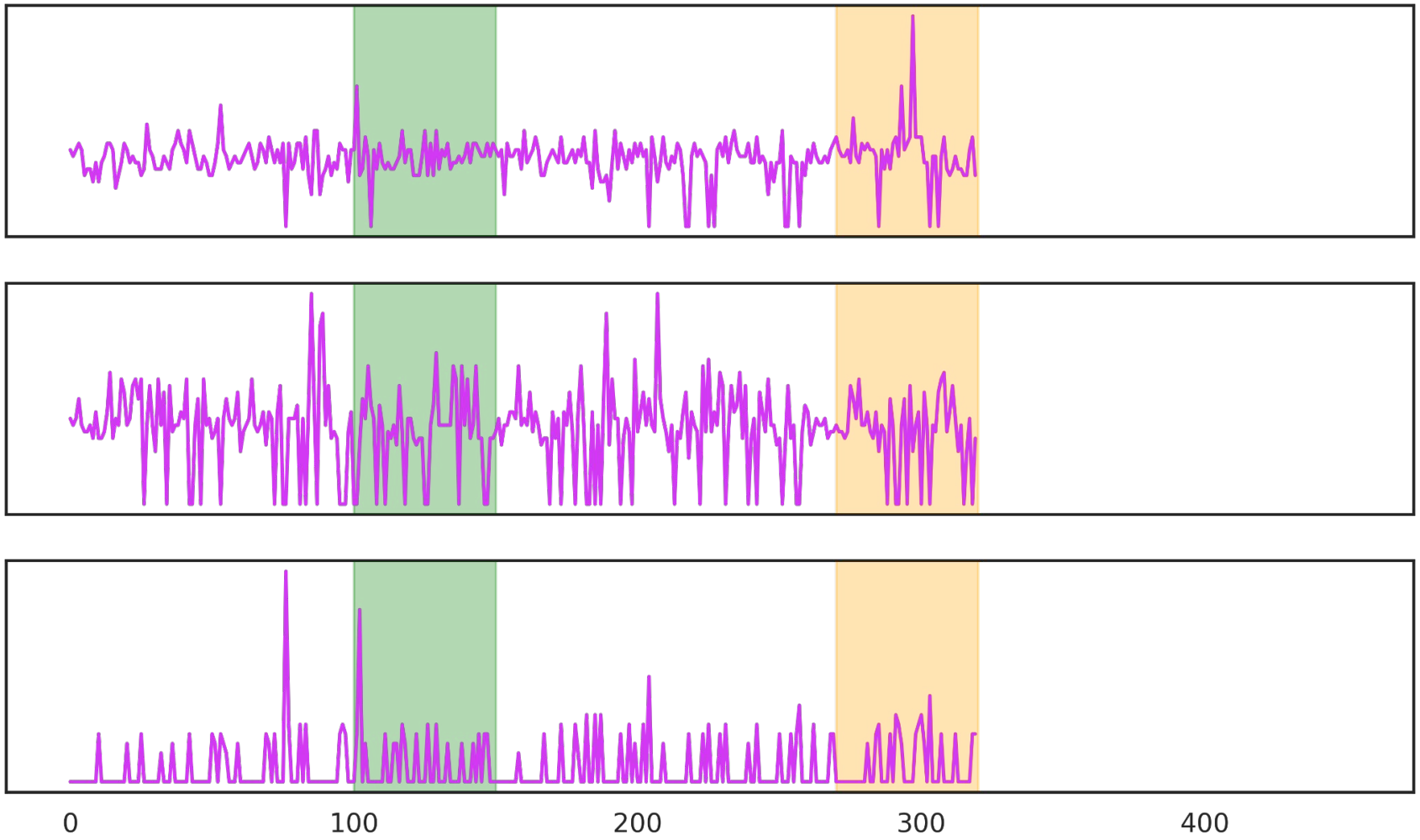
Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, 413–422. IEEE, 2008.

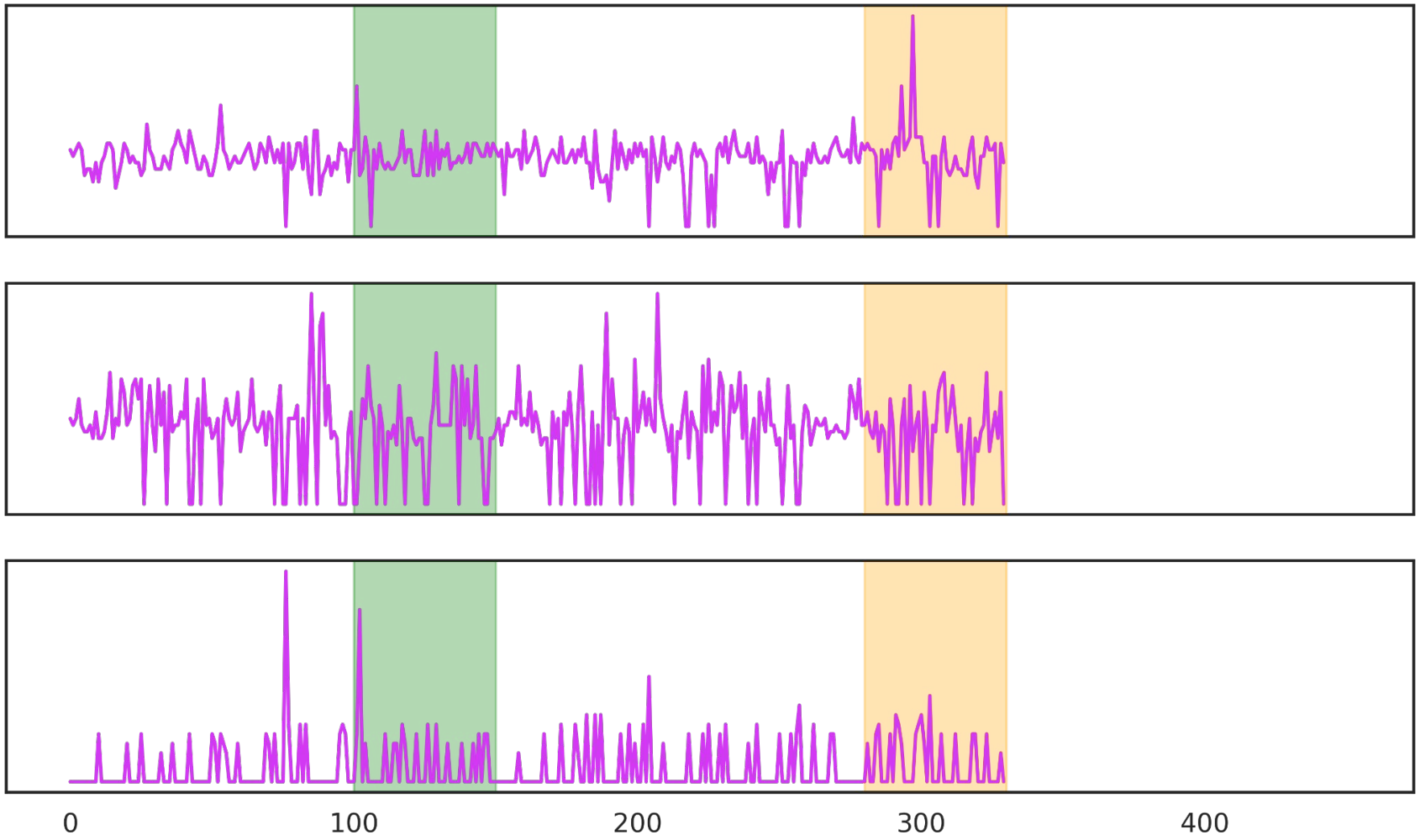
Потоковые данные

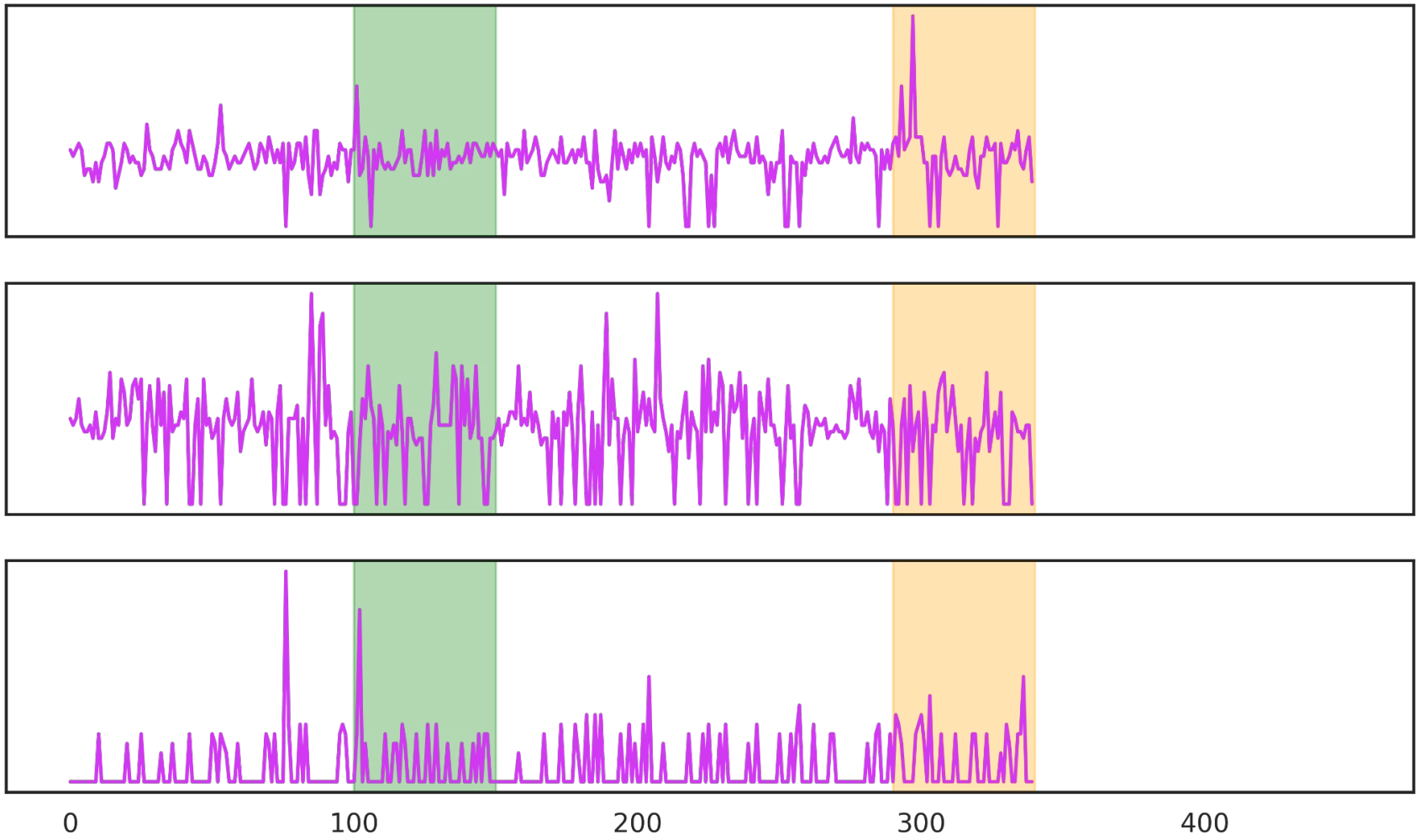
Zhiguo Ding and Minrui Fei. An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. *IFAC Proceedings Volumes*, 46(20):12–17, 2013.

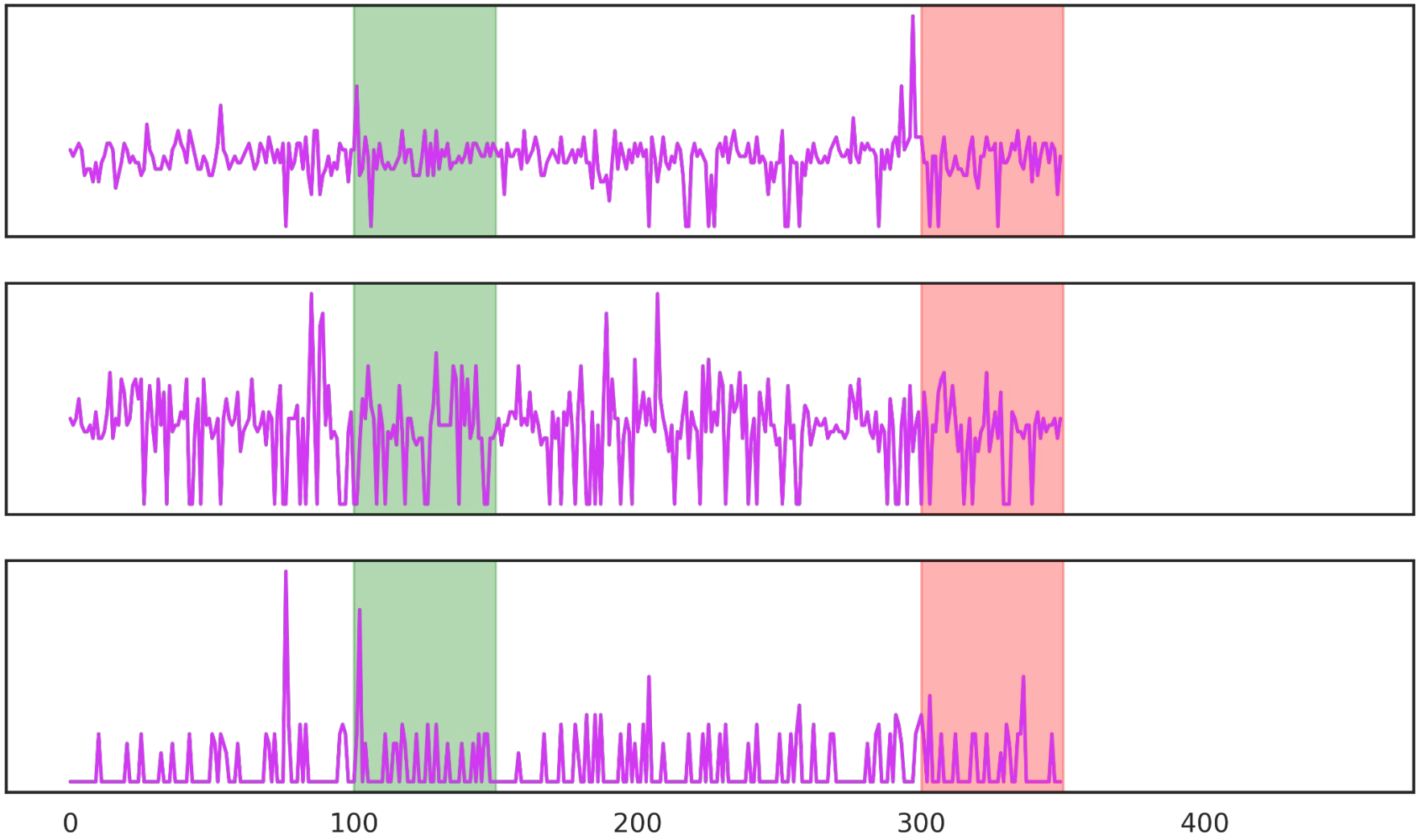

0
100
200
300
400

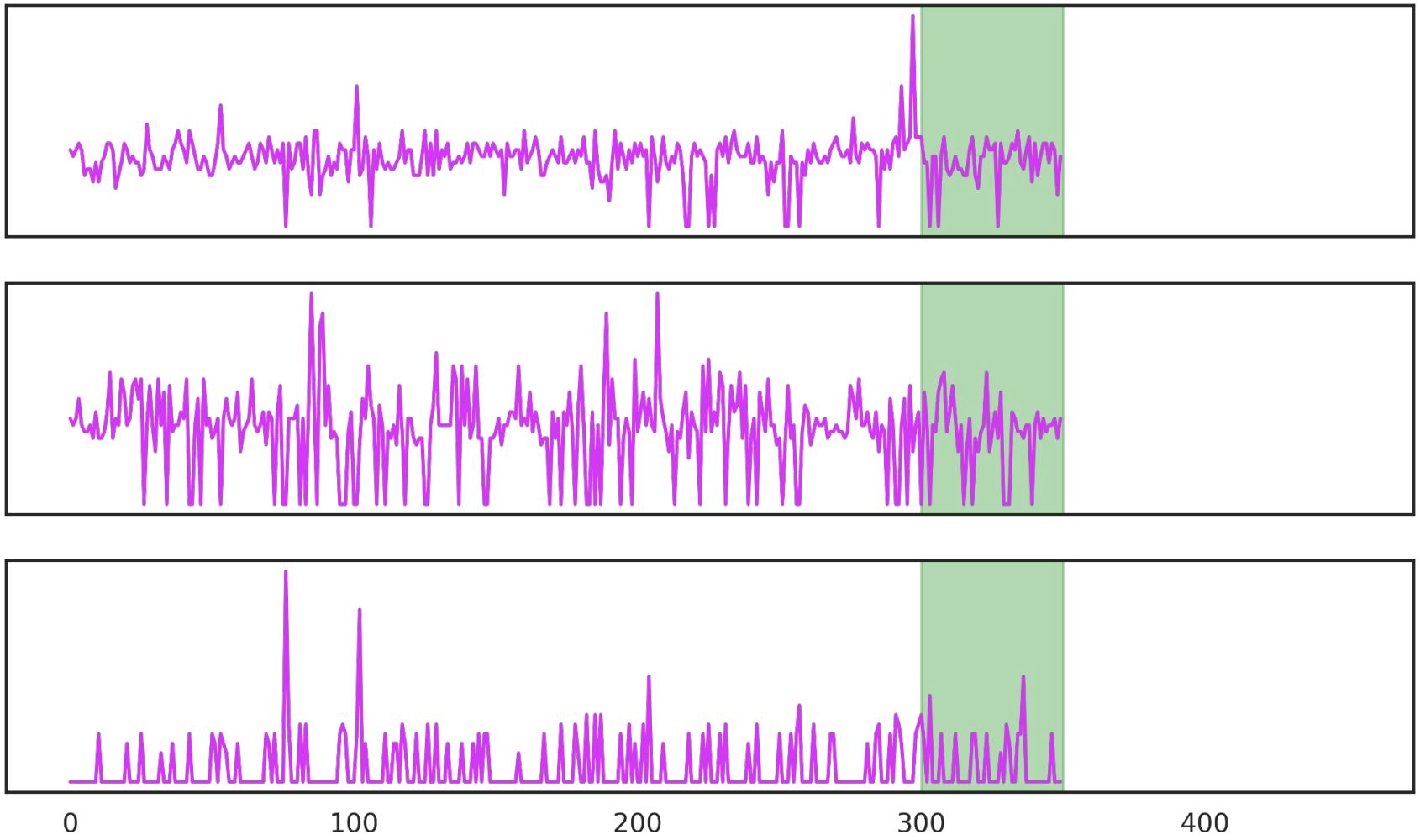

0
100
200
300
400


0
100
200
300
400

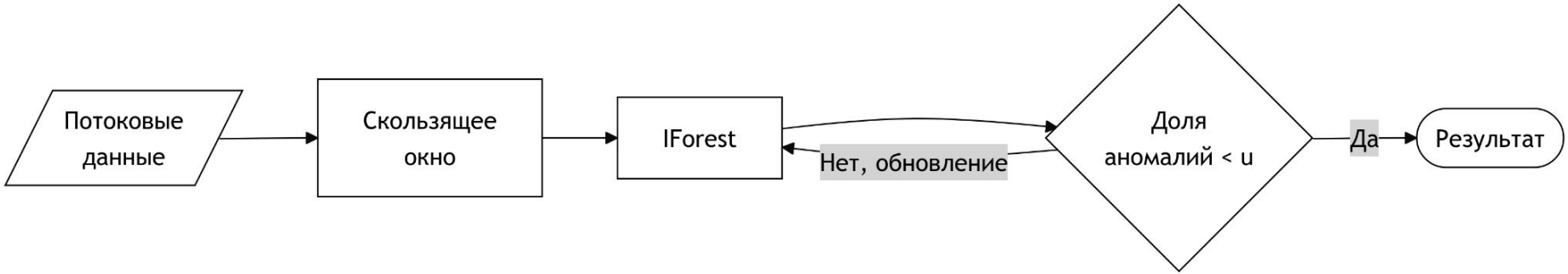

0
100
200
300
400

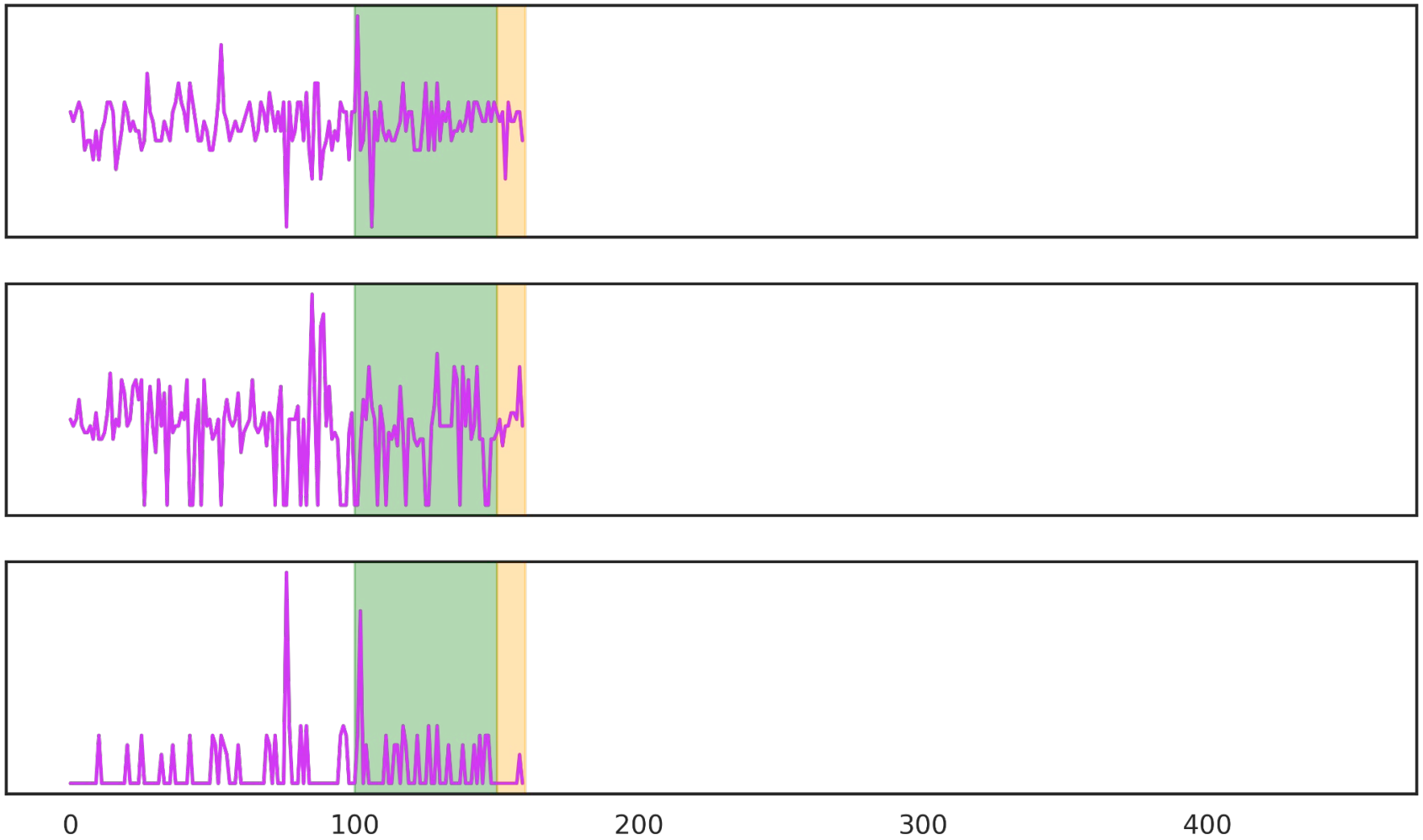

0
100
200
300
400

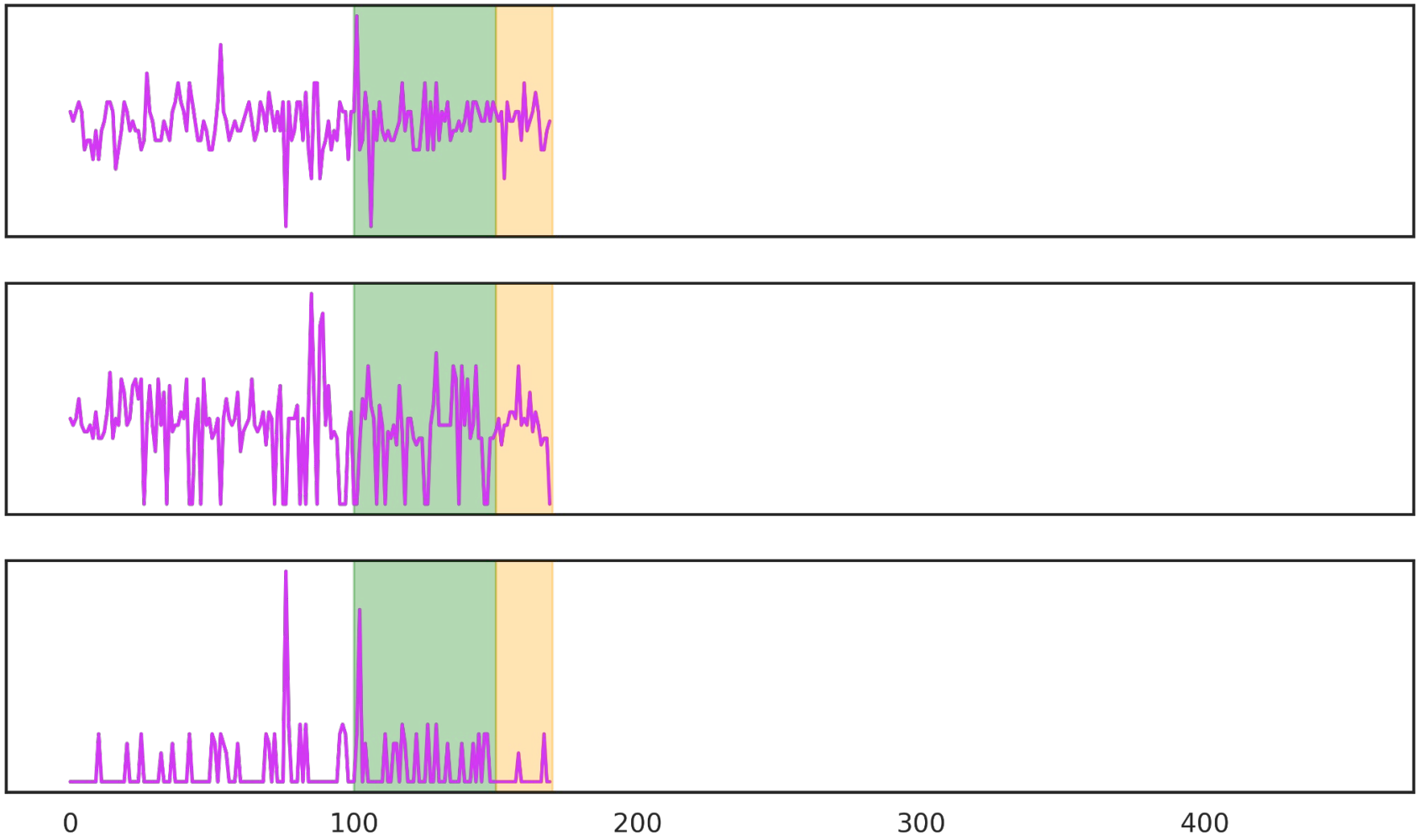

0
100
200
300
400

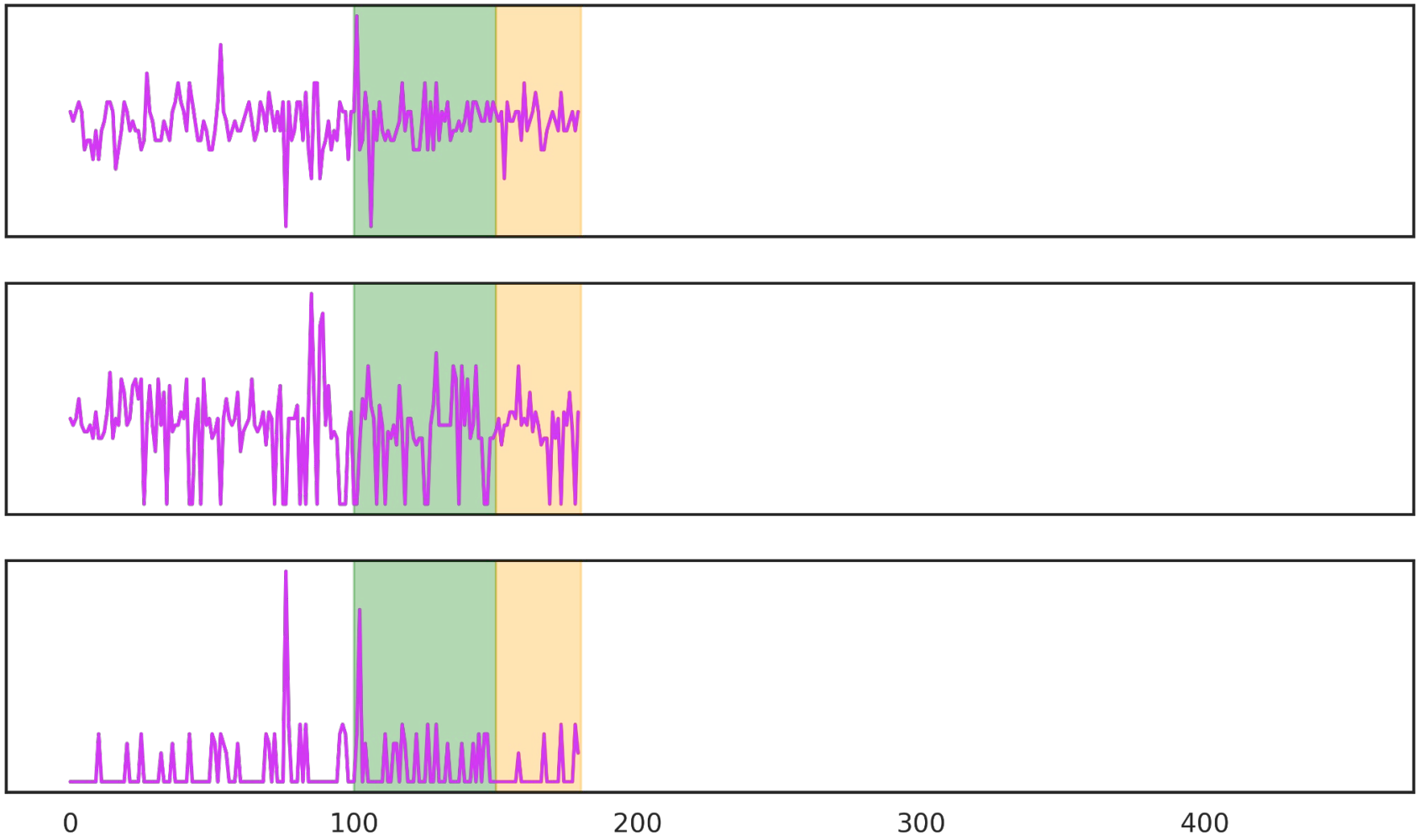

0
100
200
300
400

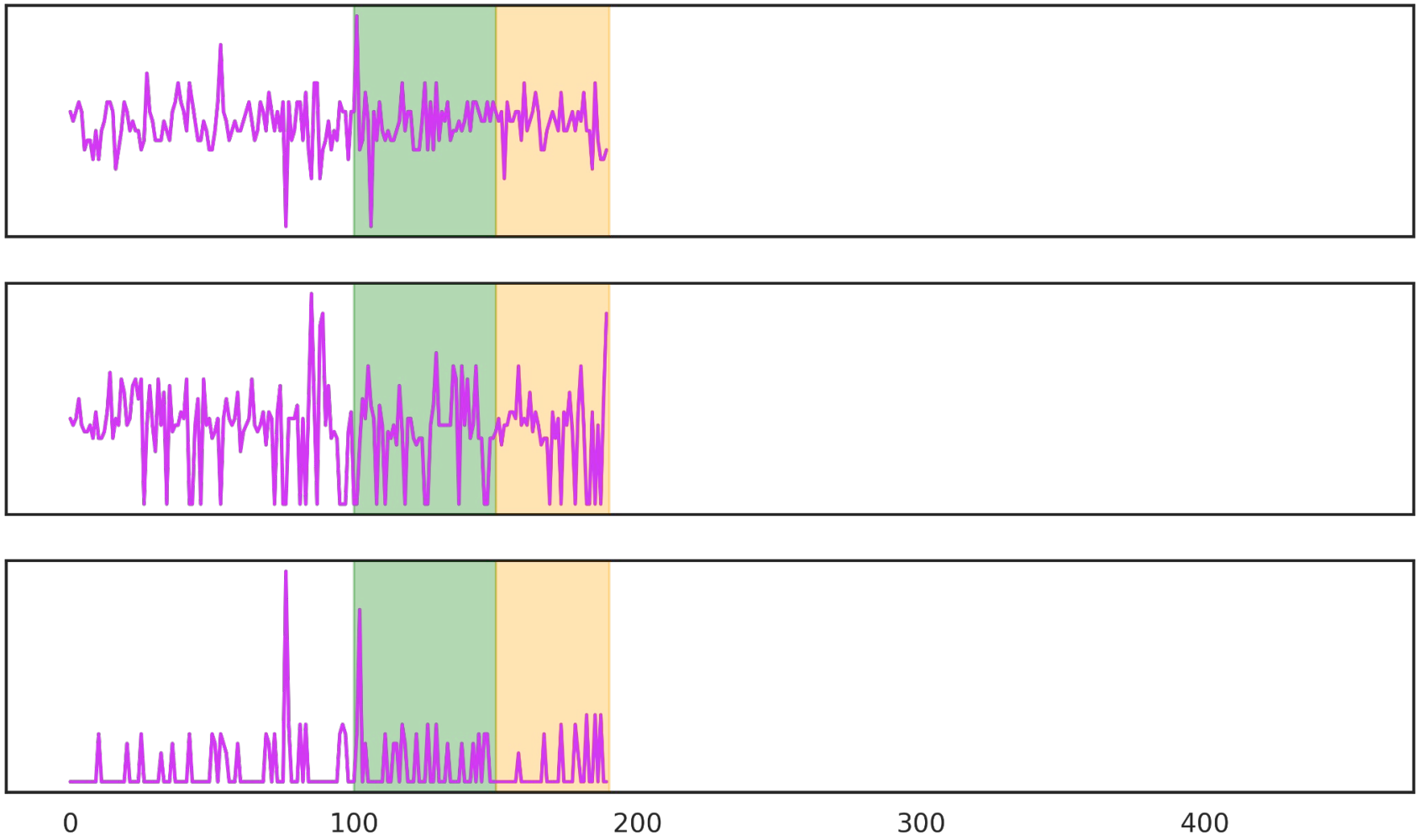
Алгоритм

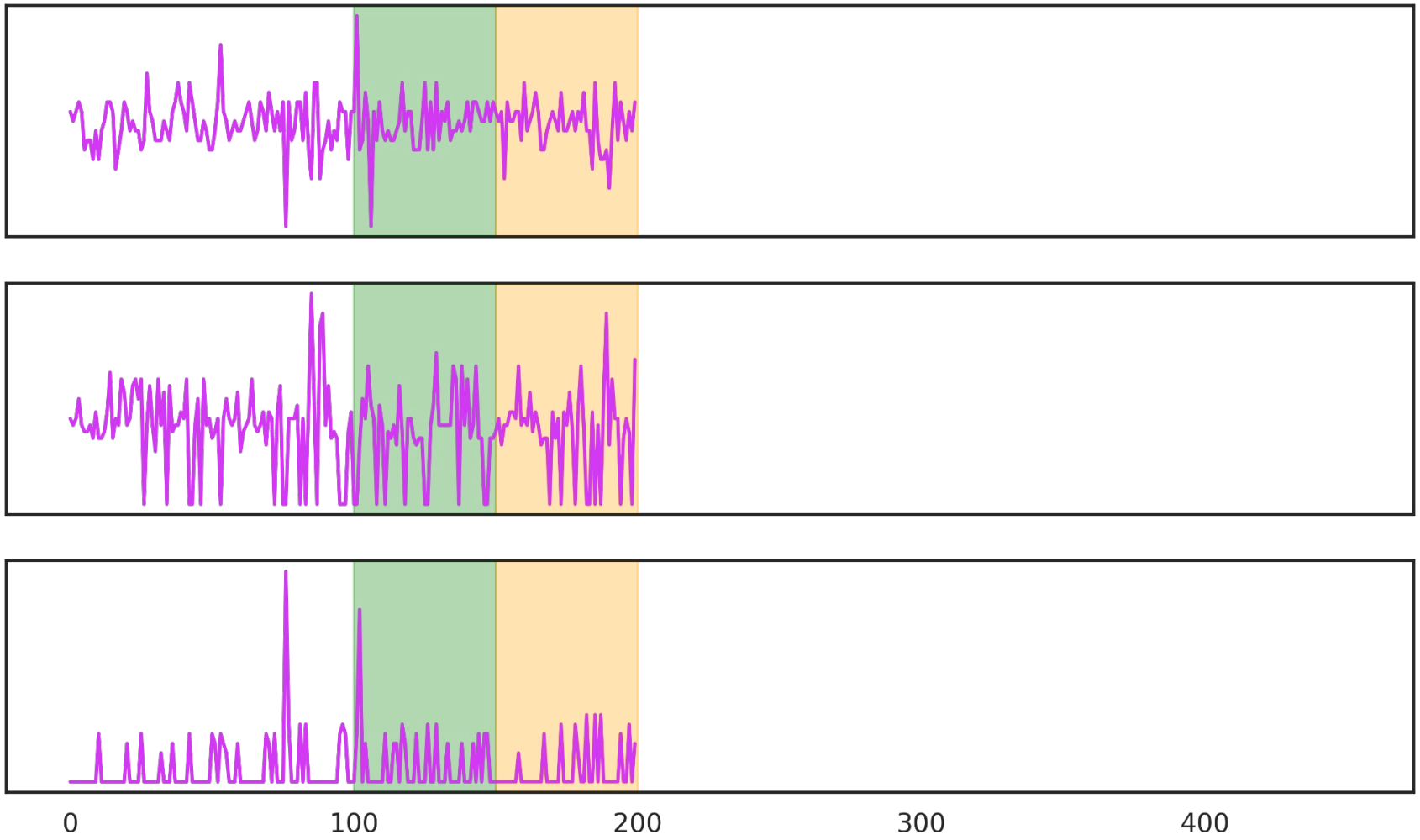

Потоковые данные
Скользящее окно
IForest
Нет, обновление
Доля аномалий < u
Да
Результат

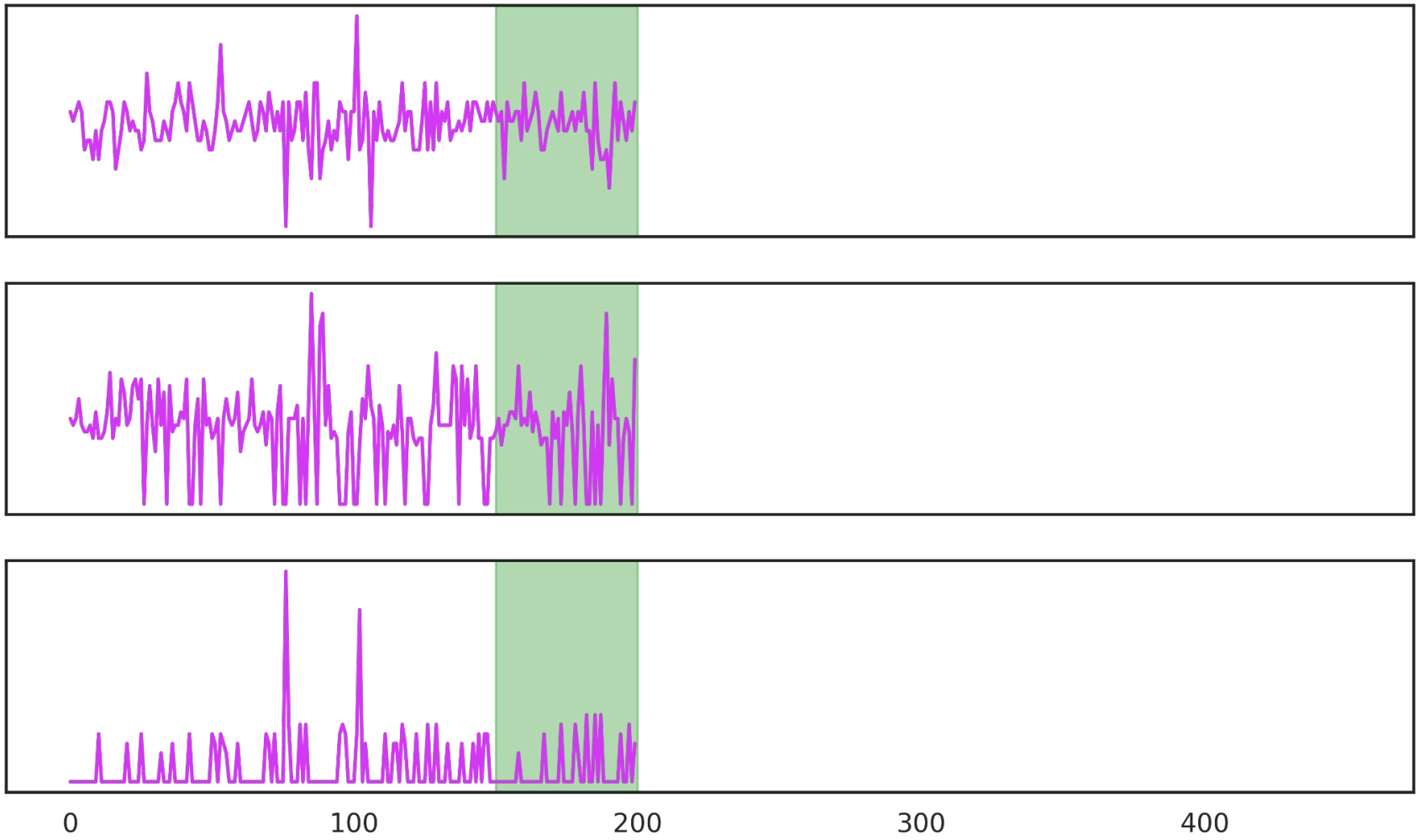

0
100
200
300
400


0
100
200
300
400


0
100
200
300
400


0
100
200
300
400


0
100
200
300
400


0
100
200
300
400

Алгоритм в pysad

Emaad Manzoor, Hemank Lamba, and Leman Akoglu. Xstream: outlier detection in feature-evolving data streams. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,* 1963–1972. 2018.

Резюме

Вопросы?