AIIPCC2019

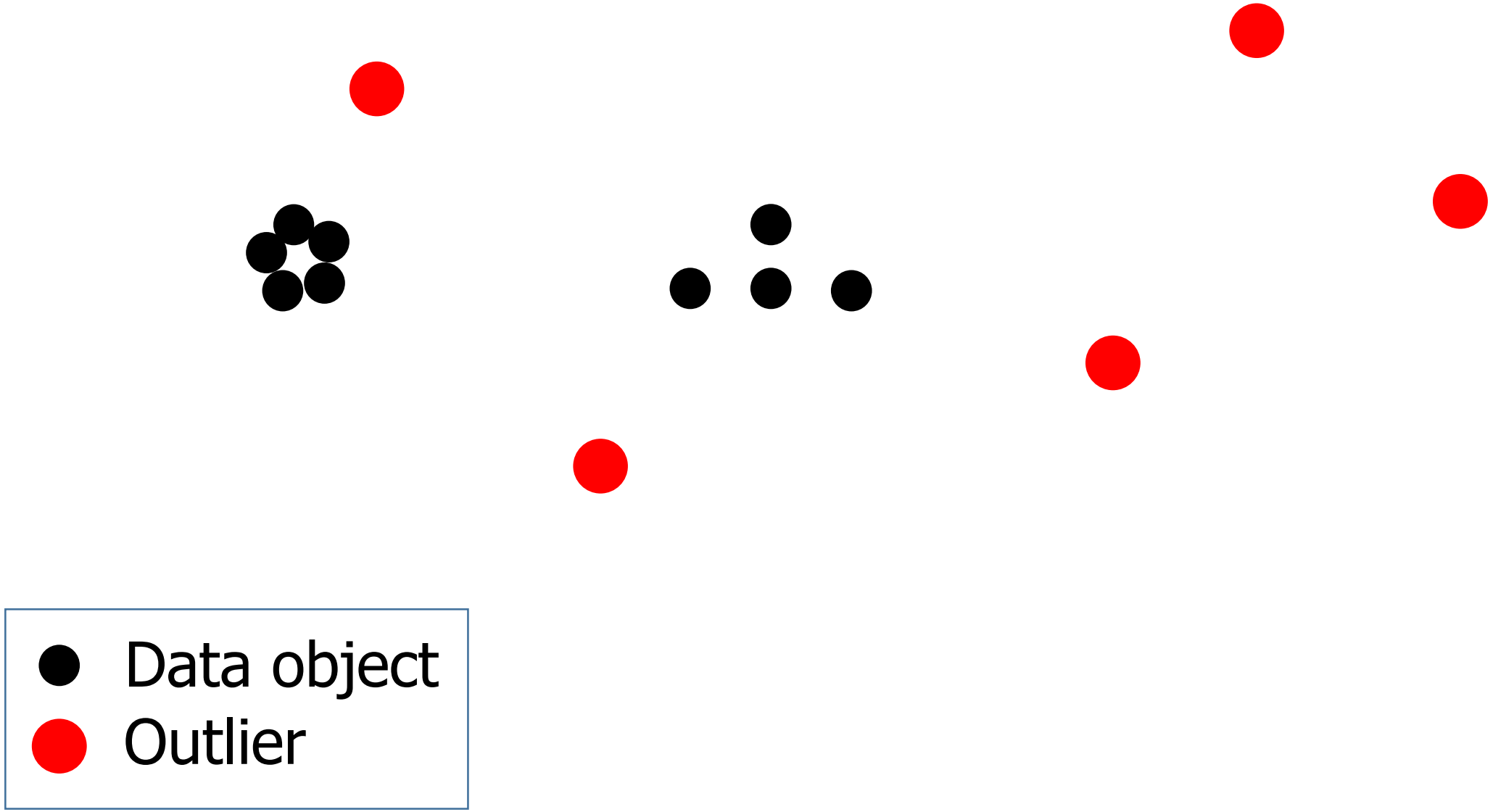
UNIVERSITY OF EASTERN FINLAND

Outlier Detection: How to Threshold Outlier Scores?

Jiawei Yang
Susanto Rahardja
Pasi Fränti

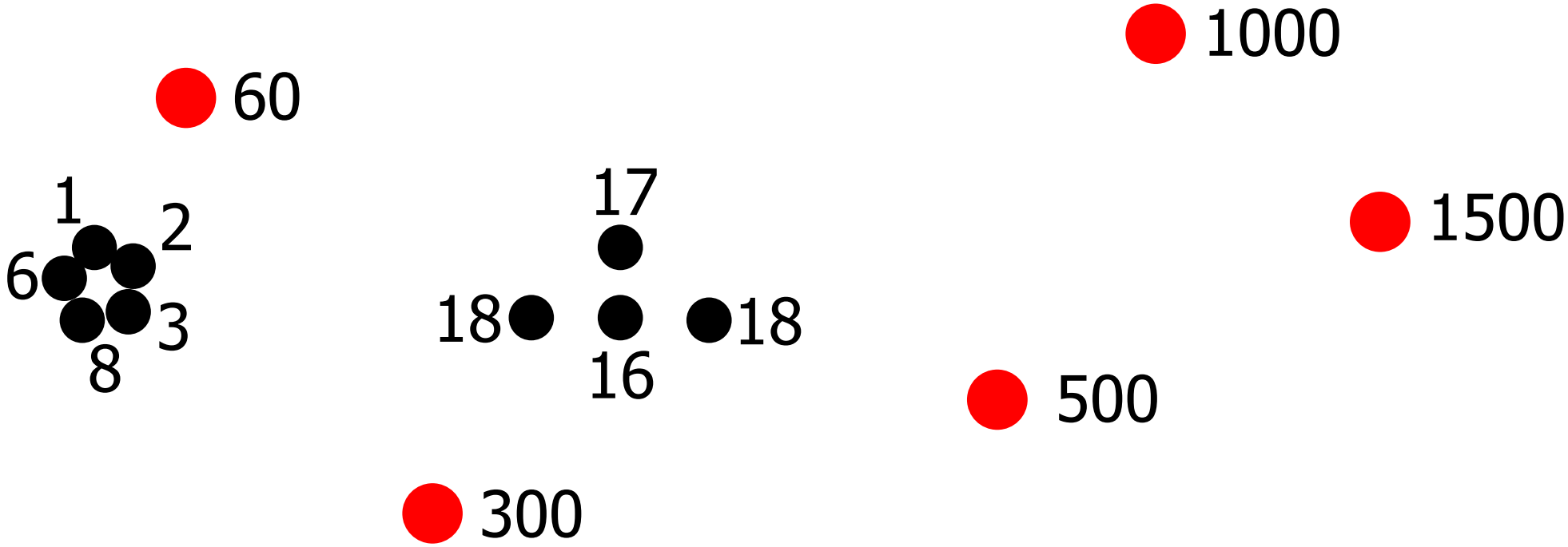
20.12.2019

Data with outliers



Outlier detection steps

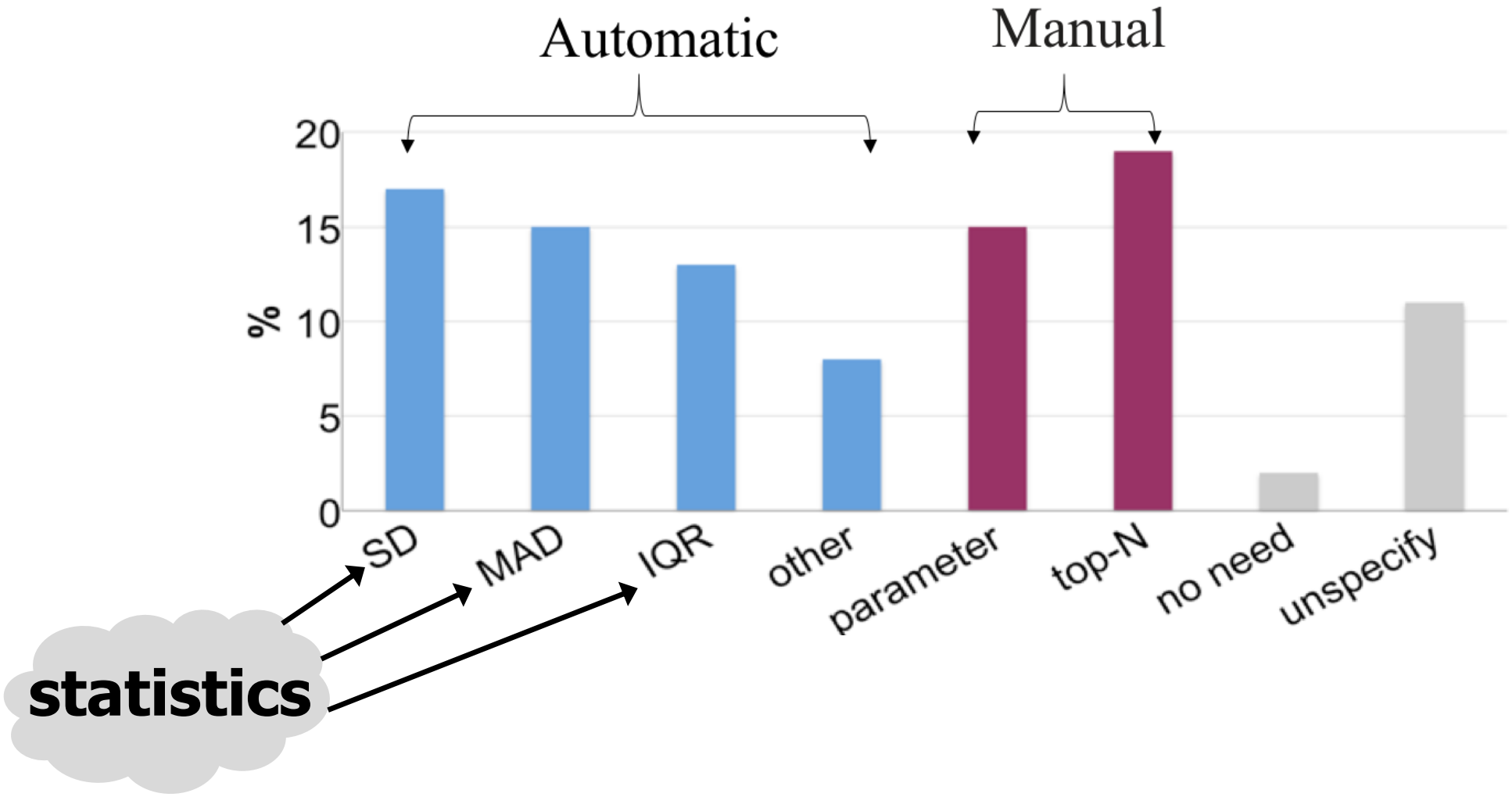
Scoring:



Thresholding: 1, 2, 3, 6, 8, 16, 17, 18, 18, 60, 300, 500, 1000, 1500

Thresholding techniques

Based on literature 6/2016 – 6/2018



Equations

SD:

$$\mathrm{T} = \mathrm{mean} + a * \mathrm{SD};$$

MAD:

$$\mathrm{T} = \mathrm{median}(\mathrm{X}) + a * \mathrm{MAD};$$

$$\mathrm{MAD} = \mathrm{b} * \mathrm{median}(|\mathrm{X} - \mathrm{median}(\mathrm{X})|);$$

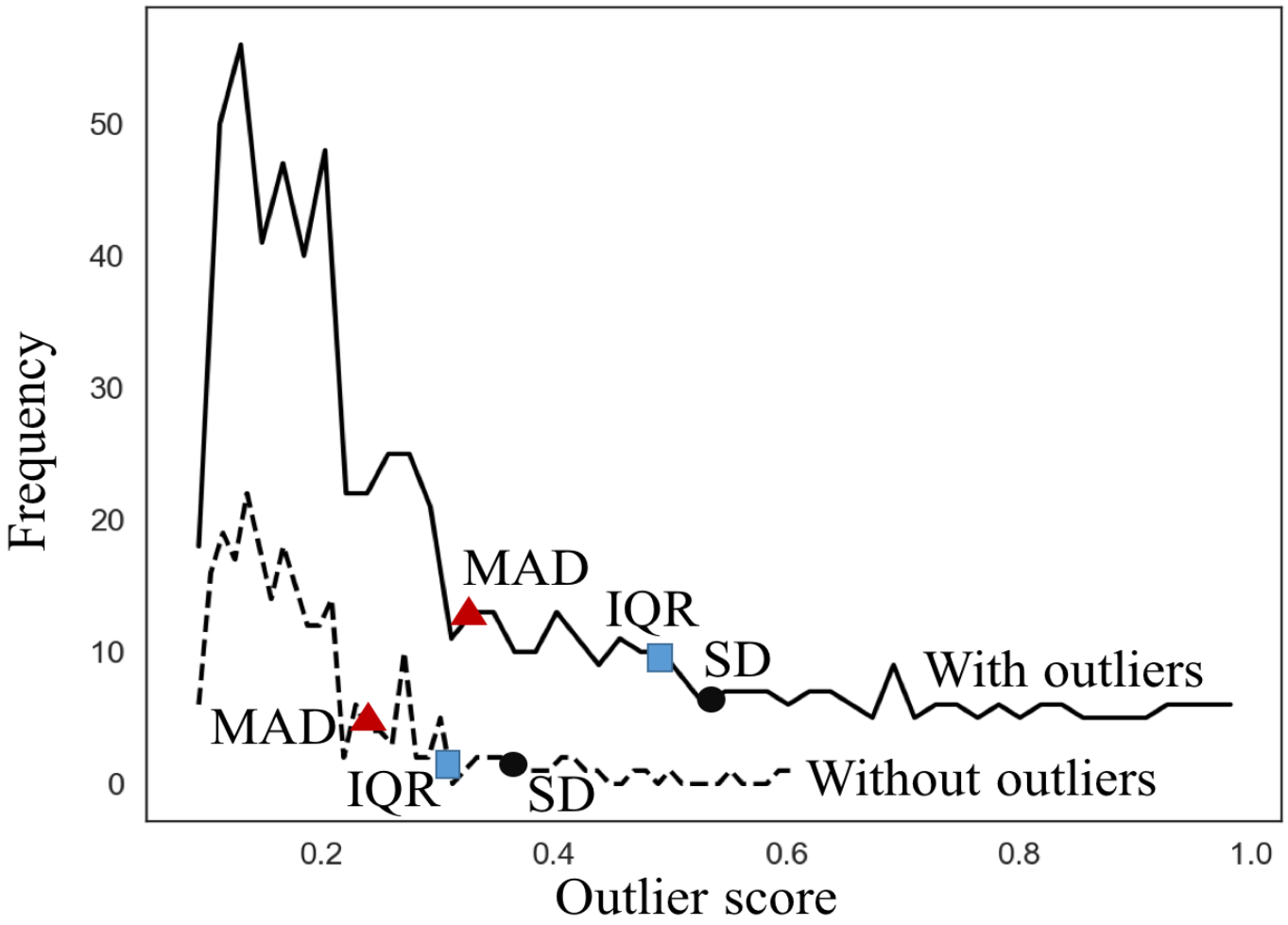
IQR:

$$\mathrm{T} = \mathrm{Q3} + \mathrm{c} * \mathrm{IQR};$$

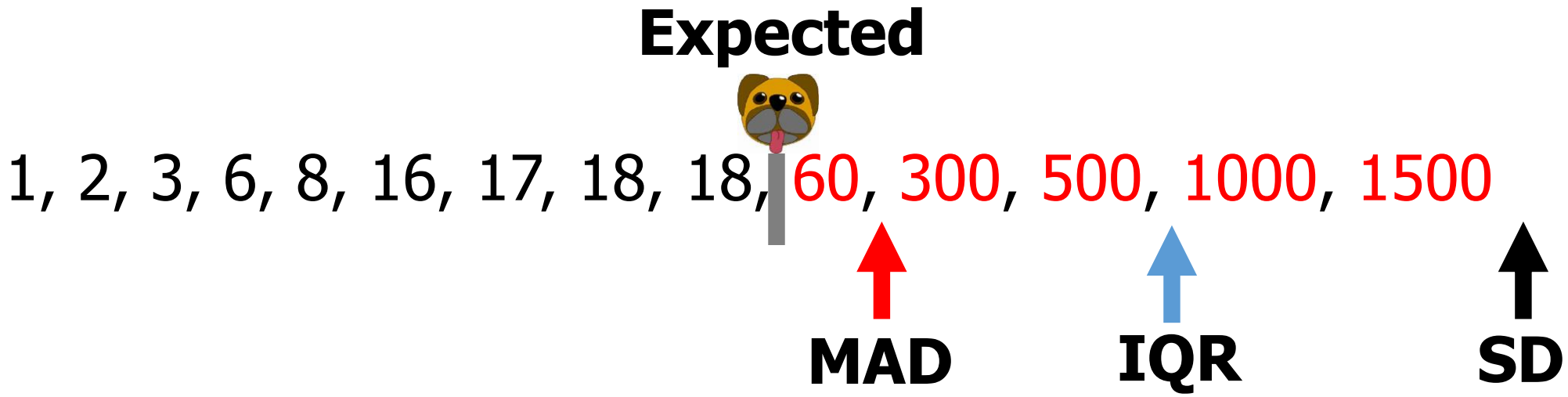
$$\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1};$$

Statistics are biased

Reason: presence of the outliers



Performance of biases statistics



Method	Threshold	Detected Outliers
SD	1574.81	{}
MAD	84.22	{300, 500, 1000, 1500}
IQR	590.25	{1000, 1500}

Two-stage Thresholding (2T)

2T Algorithm

Select initial threshold **T**=(SD, MAD or IQR);

REPEAT

1. Remove biggest outlier scores
2. Re-calculate **T**=(SD, MAD, IQR)

UNTIL Stop condition

RETURN **T**

2T demonstration

1. Initial threshold MAD=84.22
2. Remove scores
3. Re-calculate threshold MAD=39.13

Expected

1, 2, 3, 6, 8, 16, 17, 18, 18, 60, 300, 500, 1000, 1500

MAD

Performance of 2T

Expected

1, 2, 3, 6, 8, 16, 17, 18, 18, 60, 300, 500, 1000, 1500

MAD IQR

SD

Method	Threshold	Detected Outliers
SD	1574.81	{}
MAD	39.13	{60, 300, 500, 1000, 1500}
IQR	38.00	{60, 300, 500, 1000, 1500}

If first stage fails, 2T will fail!

Experiments

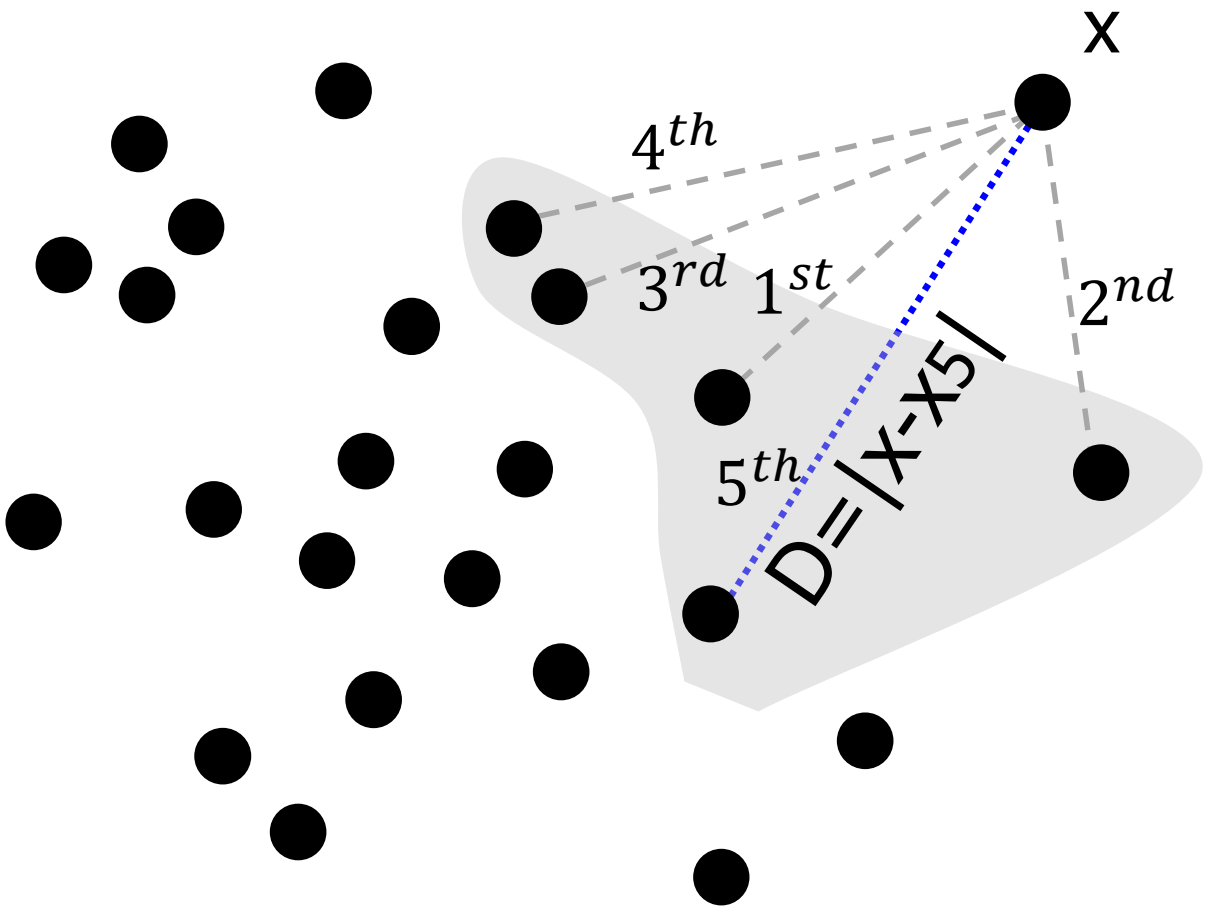
Datasets

Dataset	Size	Outliers	Dim	Outlier Object
KDDCup99	60632	246	38	Network attack
Wilt	4839	261	5	Diseased trees
Stamps	340	31	9	Forged stamps
PageBlocks	5473	560	10	Pictures or graphics
Cardiotocography	2126	471	21	Patients
Pima	768	268	8	Patients
SpamBase	4601	1,813	57	Spam email
HeartDisease	270	120	13	Patients
Arrhythmia	450	206	259	Affected patients
Parkinson	195	147	22	Patients

Campos et al: "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study",
Data Mining and Knowledge Discovery, 2016.

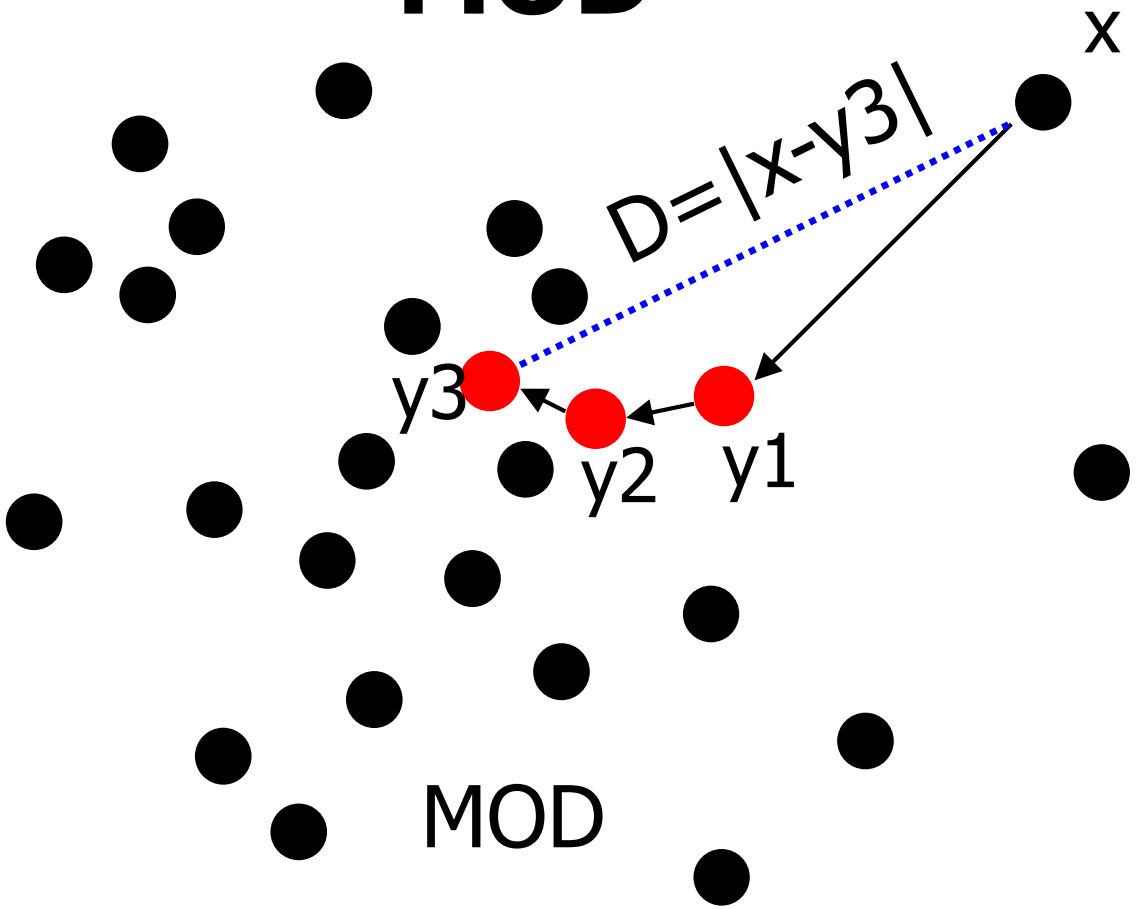
Outlier detectors

KNN



Ramaswamy et al. ”Efficient algorithms for mining outliers from large data sets”. *ACM SIGMOD Record*, 2000.

MOD



Yang et al. ”Mean-shift outlier detection”, FSDM 2018.

Results with KNN detector

F1-score

Dataset	SD			MAD		IQR	
	Original	2T	Clever	Original	2T	Original	2T
KDDCup.	0.54	0.46	0.00	0.43	0.37	0.48	0.45
Wilt	0.61	0.59	0.05	0.57	0.50	0.61	0.60
Stamps	0.58	0.72	0.08	0.75	0.61	0.59	0.69
PageB.	0.69	0.74	0.09	0.66	0.55	0.75	0.73
Card.	0.61	0.56	0.61	0.57	0.53	0.61	0.61
Pima	0.55	0.63	0.49	0.58	0.65	0.49	0.51
Spam.	0.42	0.48	0.42	0.47	0.51	0.41	0.43
HeartD.	0.52	0.59	0.38	0.53	0.59	0.42	0.46
Arrhy.	0.55	0.65	0.31	0.65	0.67	0.54	0.59
Parki.	0.31	0.42	0.30	0.39	0.46	0.31	0.34
AVG	0.53	0.58	0.33	0.57	0.54	0.51	0.54

Results with MOD detector

F1-score

Dataset	SD			MAD		IQR	
	Original	2T	Clever	Original	2T	Original	2T
KDDCup.	0.54	0.47	0.00	0.43	0.37	0.48	0.45
Wilt	0.61	0.53	0.05	0.52	0.47	0.59	0.56
Stamps	0.60	0.72	0.60	0.73	0.65	0.60	0.66
PageB.	0.68	0.73	0.09	0.66	0.55	0.75	0.72
Card.	0.55	0.56	0.55	0.56	0.54	0.55	0.55
Pima	0.54	0.62	0.42	0.60	0.63	0.49	0.52
Spam.	0.55	0.49	0.38	0.56	0.52	0.42	0.43
HeartD.	0.52	0.54	0.38	0.51	0.54	0.40	0.42
Arrhy.	0.56	0.65	0.55	0.61	0.67	0.53	0.57
Parki.	0.35	0.49	0.34	0.42	0.48	0.34	0.36
AVG	0.55	0.58	0.34	0.56	0.54	0.51	0.52

The amount of detected outliers

MOD detector

Dataset	Outlier	SD	2T	Clever
KDDCup.	246	3509	9383	48105
Wilt	261	330	1073	4806
Stamps	31	35	79	334
PageB.	560	229	834	5378
Card.	471	183	522	2103
Pima	268	108	228	734
Spam.	1813	693	1047	4186
HeartD.	120	44	85	263
Arrhy.	206	63	136	428
Parki.	147	22	52	174

Time (s)

MOD detector

Dataset (Size)	SD	2T	Clever
KDDCup. (60632)	<0.01	0.12	811.40
Wilt (4839)	<0.01	0.01	8.39
Stamps (340)	<0.01	<0.01	0.06
PageB. (5473)	<0.01	0.01	10.54
Card. (2126)	<0.01	<0.01	1.69
Pima (768)	<0.01	<0.01	0.26
Spam. (4601)	<0.01	0.01	6.29
HeartD. (270)	<0.01	<0.01	0.05
Arrhy. (450)	<0.01	<0.01	0.10
Parki. (195)	<0.01	<0.01	0.02

Conclusions

Why to use:

- Simple but effective!

How it performs:

- Improve existing thresholding!

Usefulness:

- Almost no extra coding needed!

Thank you!