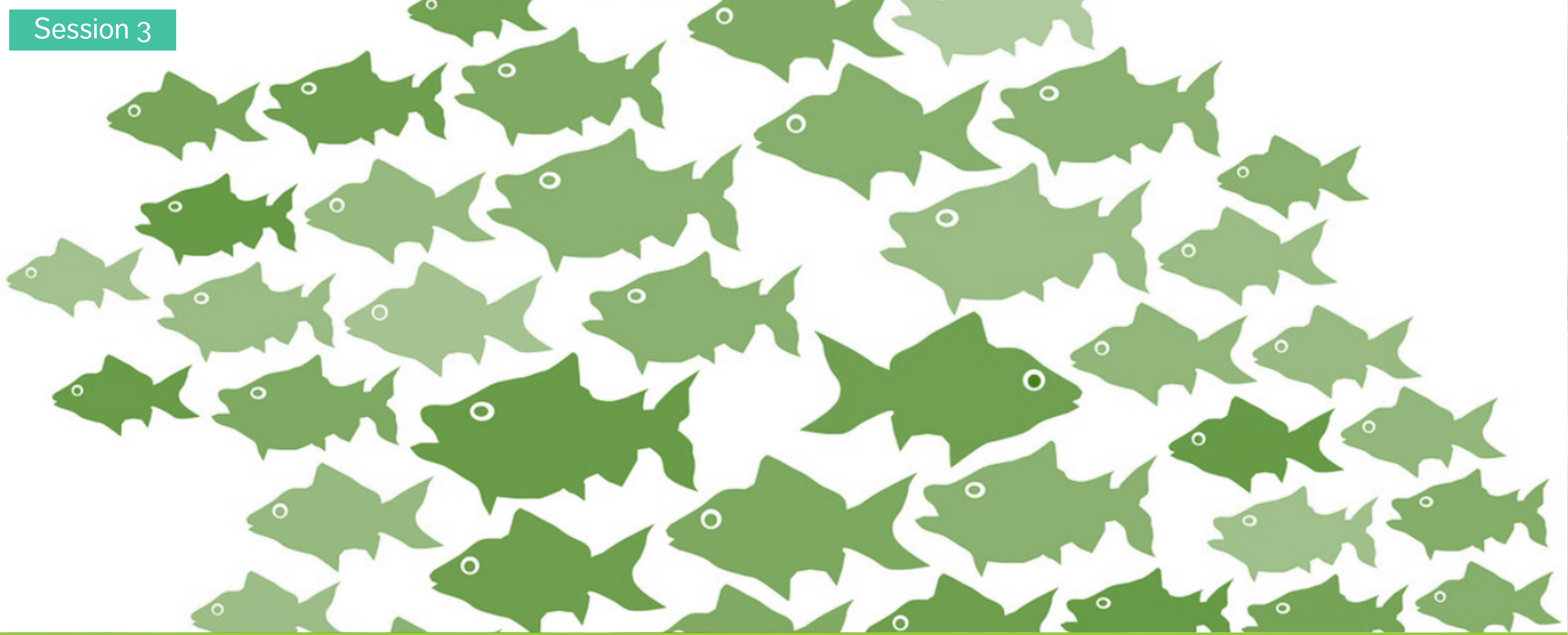


9. Anomalous Observations

Anomalous Observations

In practice, an **anomalous observation** may arise as

- a **“bad” object/measurement:** data artifacts, spelling mistakes, poorly imputed values, etc.
- a **misclassified observation:** according to the existing data patterns, the observation should have been labeled differently;
- an observation whose measurements are found in the **distribution tails** of a large enough number of features;
- an **unknown unknown**: a completely new type of observations whose existence was heretofore unsuspected.

Anomalous Observations

Observations could be anomalous in one context, but not in another:

- A 6-foot tall adult male is in the 86th percentile for Canadian males (tall, but not unusual)
- in Bolivia, the same man would be in the 99.9th percentile (very tall and unusual)

Anomaly detection points towards **interesting questions** for analysts and subject matter experts: in this case, why is there such a large discrepancy in the two populations?

Outliers

Outlying observations are data points which are **atypical** in comparison to

- the unit's remaining features (*within-unit*),
- the field measurements for other units (*between-units*)

Outliers are observations which are **dissimilar to other cases** or which **contradict known dependencies** or rules.

Careful study is needed to determine whether outliers should be retained or removed from the dataset.

Detecting Anomalies

Outliers may be anomalous along any of the unit's variables, or in combination.

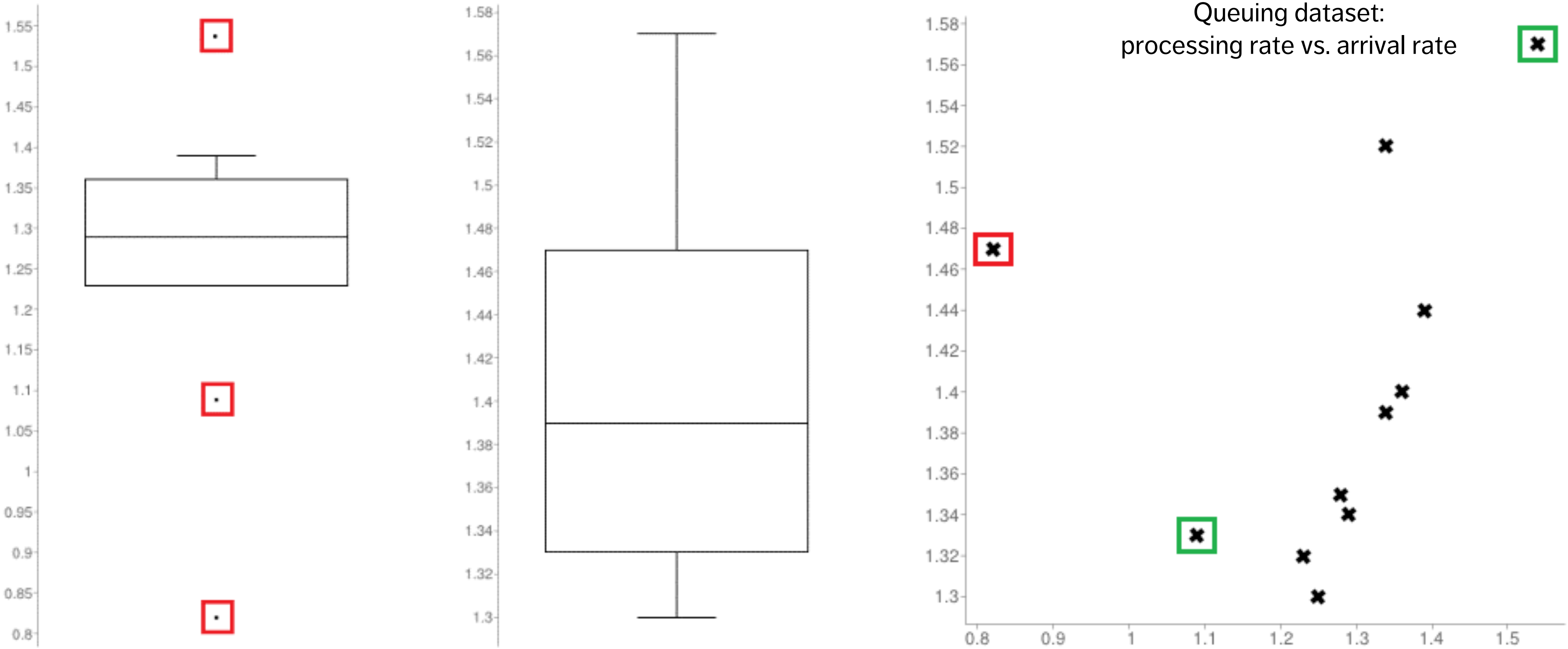
Anomalies are by definition **infrequent**, and typically shrouded in **uncertainty** due to small sample sizes.

Differentiating anomalies from noise or data entry errors is **hard**.

Boundaries between normal and deviating units may be **fuzzy**.

Anomalies associated with malicious activities are typically **disguised**.

Visual Outlier Detection



Detecting Anomalies

Numerous methods exist to identify anomalous observations; **none of them are foolproof** and judgement must be used.

Graphical methods are easy to implement and interpret:

- **Outlying Observations**

  box-plots, scatterplots, scatterplot matrices, 2D tour, Cooke's distance, normal qq plots
- **Influential Data**

  some level of analysis must be performed (leverage)

Careful: once anomalous observations have been removed from the dataset, previously “regular" units may become anomalous.

Anomaly Detection Algorithms

Supervised methods use a historical record of labeled anomalous observations:

- domain expertise is required to tag the data
- classification or regression task
- rare occurrence problem

		Predicted Class	
		Normal	Anomaly
Actual Class	Normal	*TN*	*FP*
	Anomaly	*FN*	*TP*

Unsupervised methods don’t use external information:

- traditional methods and tests
- can also be seen as a clustering or association rules problem

Anomaly Detection Algorithms

The mis-classification cost is often assumed to be symmetrical, which can lead to **technically correct but useless** outputs.

For instance, most (99.999+%) air passengers do not bring weapons with them on flights; a model that predicts that no passenger is smuggling a weapon would be 99.999+% accurate, but it would miss the point completely.

For the **security agency**, the cost of wrongly thinking that a passenger is:

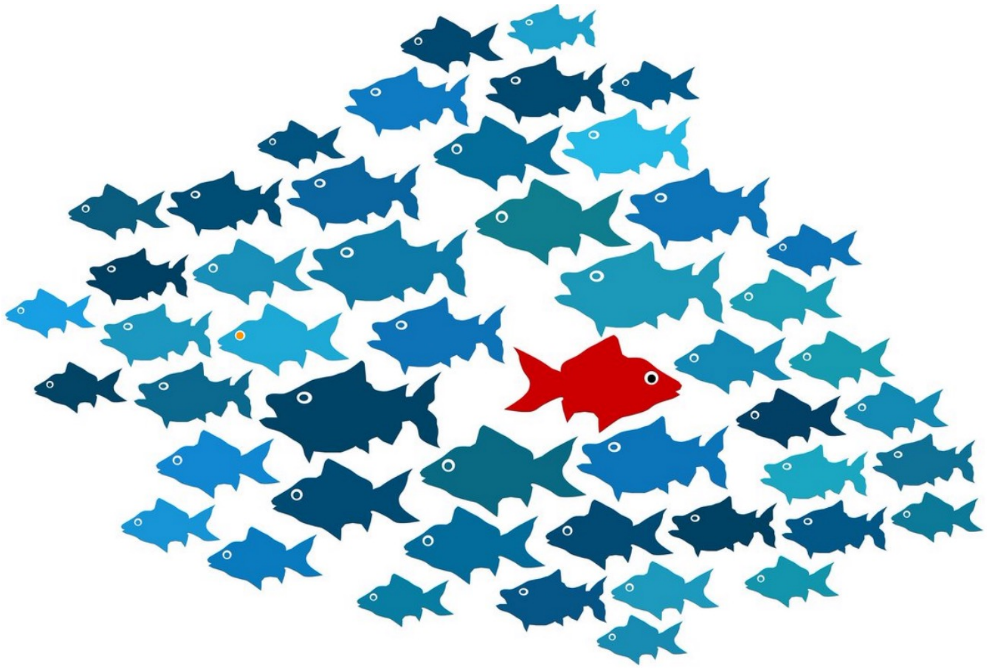
- smuggling a weapon $\Longrightarrow$ cost of a single search
- NOT smuggling a weapon $\Longrightarrow$ catastrophe (potentially)

The wrongly targeted individuals may have a different take on this!

Anomaly Detection Algorithms

If all participants in a workshop except for one can view the video conference lectures, then the one individual/internet connection/computer is **anomalous** – it behaves in a manner which is different from the others.

But this **DOES NOT MEAN** that the different behaviour is necessarily the one we are interested in…



Simple Outlier Tests

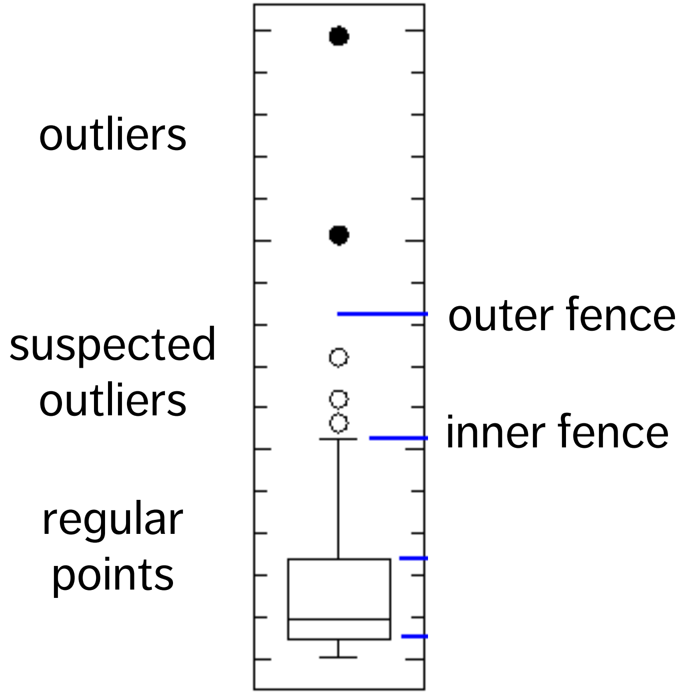
Tukey's Boxplot test: for normally distributed data, regular observations typically lie between the **inner fences**

$$Q_1 - 1.5\times(Q_3 - Q_1) \text{ and } Q_3 + 1.5\times(Q_3 - Q_1).$$

Suspected outliers lie between the **inner fences** and the **outer fences**

$$Q_1 - 3\times(Q_3 - Q_1) \text{ and } Q_3 + 3\times(Q_3 - Q_1).$$

Outliers lie beyond the **outer fences**.



Simple Outlier Tests

The **Dixon Q Test** is used in experimental sciences to find outliers in (extremely) small datasets (dubious validity).

The **Mahalanobis Distance** (linked to the leverage) can be used to find multi-dimensional outliers (when relationships are linear).

Other simple tests:

- **Grubbs** (univariate)
- **Tietjen-Moore** (for a specific # of outliers)
- **generalized extreme studentized deviate** (for unknown # of outliers)
- **chi-square** (outliers affecting goodness-of-fit)

Sophisticated Anomaly Detection

- **DBSCAN**, **OR$_h$**, and **LOF** (unsupervised outlier detection)
- **rank-power** method (supervised outlier detection)
- **distance** or **density-based** methods (with exotic distance measures)
- **autoencoders and reconstruction error** (deep learning method)
- **rare-occurrence** methods (oversampling, undersampling, CREDOS, PN, SHRINK, SMOTE, DRAMOTE, SMOTEBoost, RareBoost, MetaCost, AdaCost, CSB, SSTBoost, etc.)
- **AVF**, **Greedy** algorithms (categorical data)
- **PCA**, **DOBIN**, and other **projection** methods (for high-dimensional data)
- **subspace** methods and **ensemble** methods

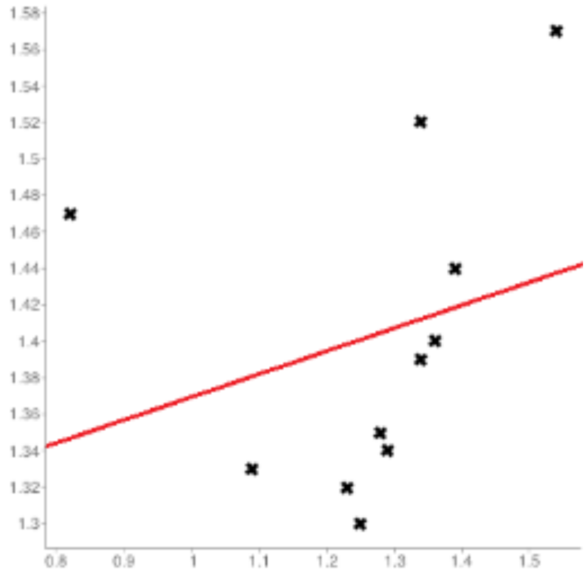
Influential Observations

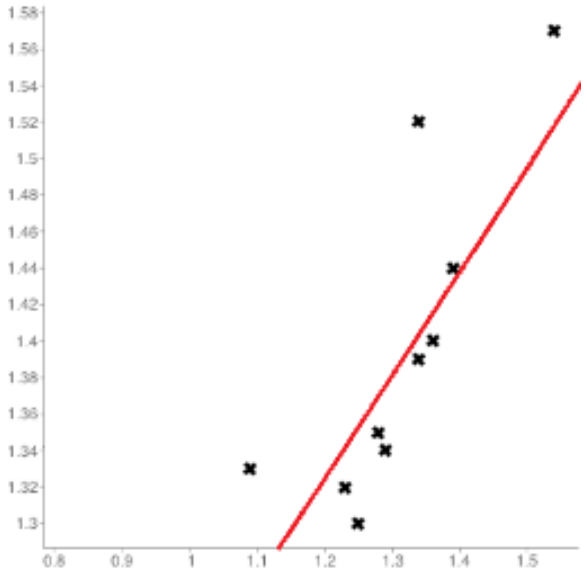
Influential data points are observations whose absence leads to **markedly different** analysis results.

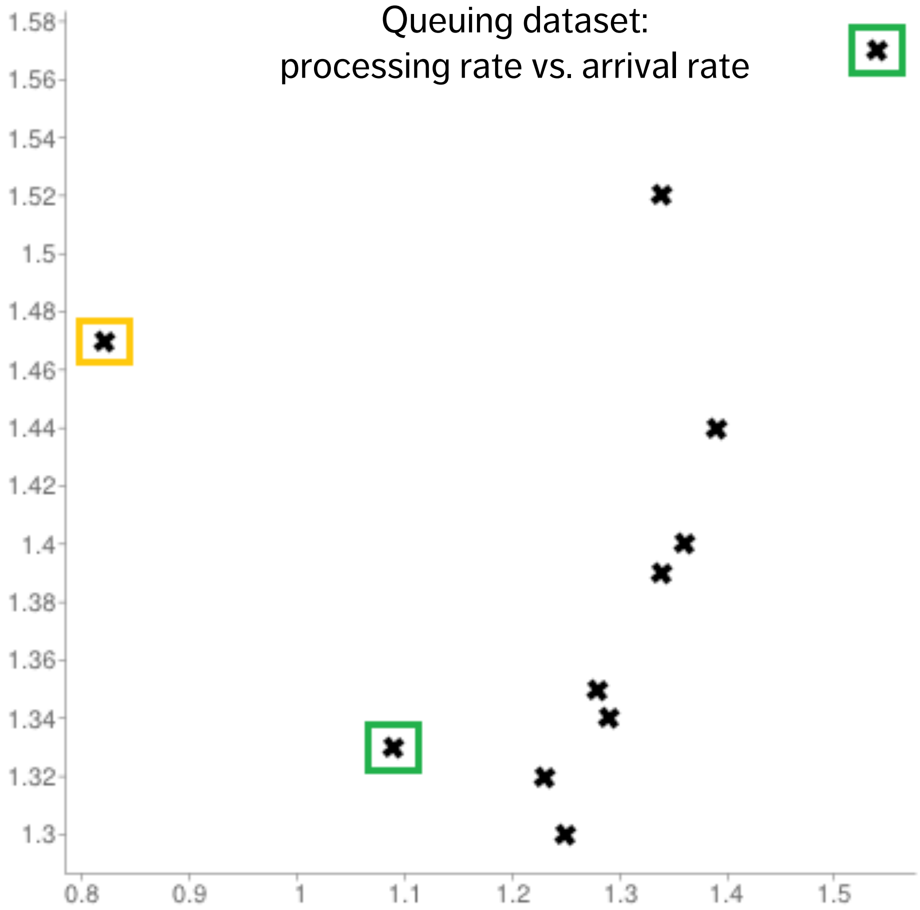
When influential observations are identified, **remedial measures** (such as data transformations) may be required to minimize their undue effects.

Outliers may be influential data points; influential data points need not be outliers (and *vice-versa*).

Influential Observations







Anomaly Detection Remarks

Identifying influential points is an **iterative process** as the various analyses have to be run numerous times.

Fully automated identification and removal of anomalous observations is **NOT recommended**.

Use data transformations if the data is **NOT normally distributed**.

Whether an observation is an outlier or not depends on **various factors**; what observations end up being influential data points depends on the **specific analysis to be performed**.

Suggested Reading

Anomalous Observations

Data Understanding, Data Analysis, Data Science
Data Preparation

Anomalous Observations

- Anomaly Detection
- Outlier Tests
- Visual Outlier Detection

***Anomaly Detection and Outlier Analysis** (advanced)

Exercises

Anomalous Observations

1. Recreate the anomaly detection process used in Example: Algae Bloom.
2. Find anomalous observations in the cities.txt and grades datasets (if applicable).
3. Find anomalous observations in a dataset of your choice.