ISEDC

Anomaly Detection and Outlier Analysis

P.Boily (uOttawa, IACS, DAL)
with Y.Cissokkho, S.Fadel, R.Millson, R.Pourhasan

Fall 2020

Outline

With the advent of automatic data collection, it is now possible to store and process large troves of data. There are technical issues associated to massive data sets, such as the speed and efficiency of analytical methods, but there are also problems related to the detection of **anomalous observations** and the **analysis of outliers**.

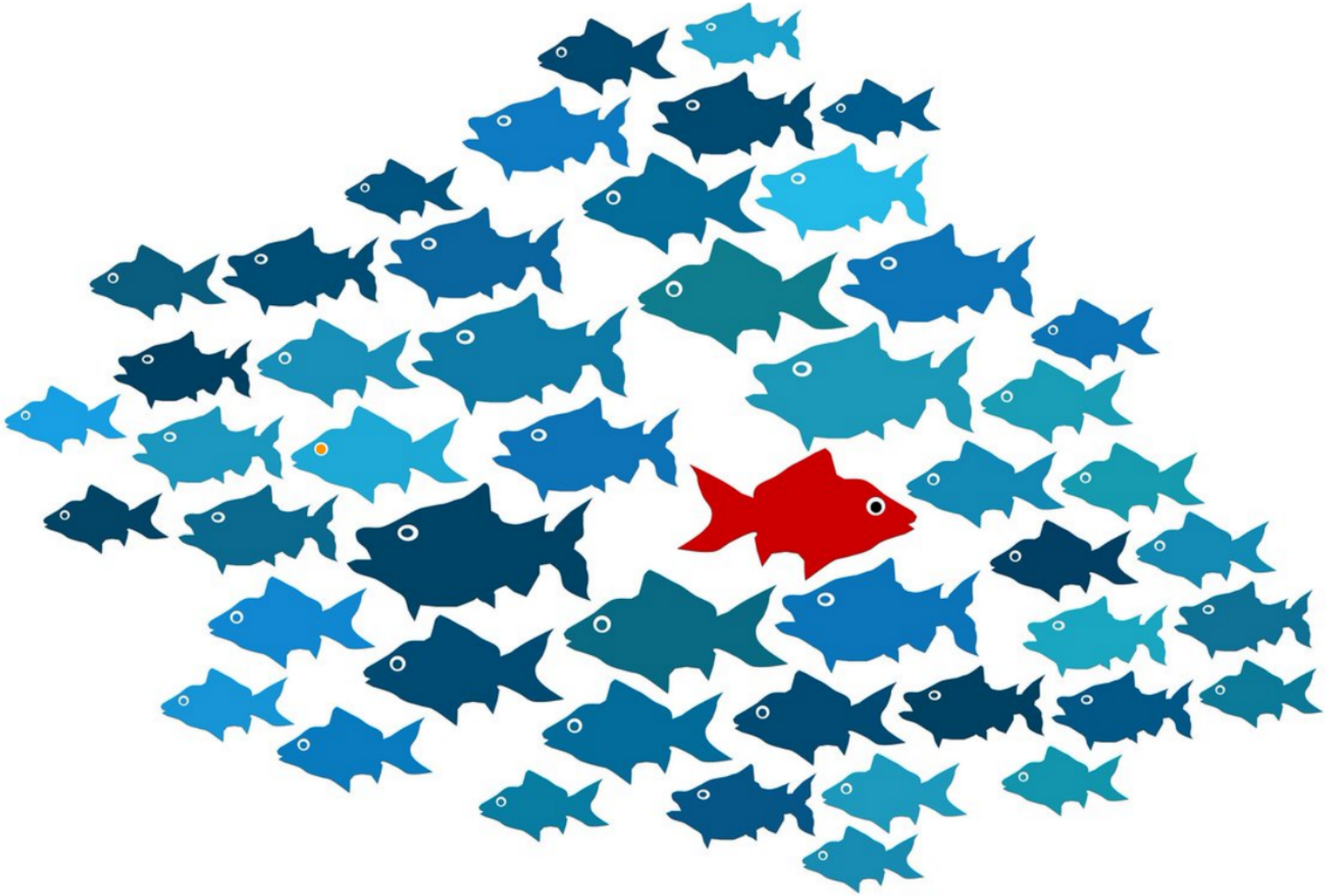
Unexpected observations can spoil analysises and/or be indicative of data collection and data processing issues.

Extreme and irregular values behave very differently from the majority of observations: they can represent criminal attacks, fraud attempts, targeted attacks, or data collection errors. As a result, anomaly detection and outlier analysis plays a crucial role in cyber-security, quality control, etc.

5.5 – Advanced Topics (p.212)

- Outlier Ensembles (p.213)
- Anomalies in Text Datasets (p.224)

References and other details can be found in Cissokho, Y., Fadel, S., Millson, R., Pourhasan, R., Boily, P. [2020], *Anomaly Detection and Outlier Analysis*, Data Science Report Series, Data Action Lab.



5.1 – Basic Notions and Overview

Isaac Asimov, the prolific American author, once wrote that

> The most exciting phrase to hear [...], the one that heralds the most discoveries, is not "Eureka!" but "That's funny...".

Important Goals: establish anomaly detection protocols and to identify strategies to deal with such observations.

Outlying observations: data points that are atypical within-unit or between-units, or as part of a collective subset of observations.

In other words, outliers are observations which are **dissimilar to other cases** or which contradict **known dependencies** or rules.

Outlying observations may be anomalous along any of the individual variables, or in combination.

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

- an adult male who is 6-foot tall falls in the 86th percentile among Canadian males $\Longrightarrow$ tall, but not unusually so;
- in Bolivia, the same man would land in the 99.9th percentile $\Longrightarrow$ extremely tall; a rarity.

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?

What's an **outlier/anomalous observation?** (reprise)

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

A common mistake that analysts make when dealing with outlying observations is to remove them from the dataset without carefully studying whether they are **influential data points**.

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

Points can be influential for one analytical methods, but not for another.

Remedial measures (data transformation strategies, etc.) may need to be applied to minimize any undue effect.

Outliers may be influential, and influential data points may be outliers, but the conditions are **neither necessary nor sufficient**.

Anomalies

Anomalies are **infrequent** and typically shrouded in **uncertainty** due to their relatively low numbers.

This makes it difficult to differentiate anomalies from banal **noise** or **data collection errors**.

The boundary between normal and deviant observations is usually **fuzzy**.

Example: before the advent of e-shops, a purchase which was recorded at 3AM (local time) would probably raise a red flag for a credit card company; but with online shops, that is not necessarily the case.

If anomalies are actually associated with **malicious activities**, they are often **disguised** to blend in with normal observations $\Longrightarrow$ this obviously complicates the detection process.

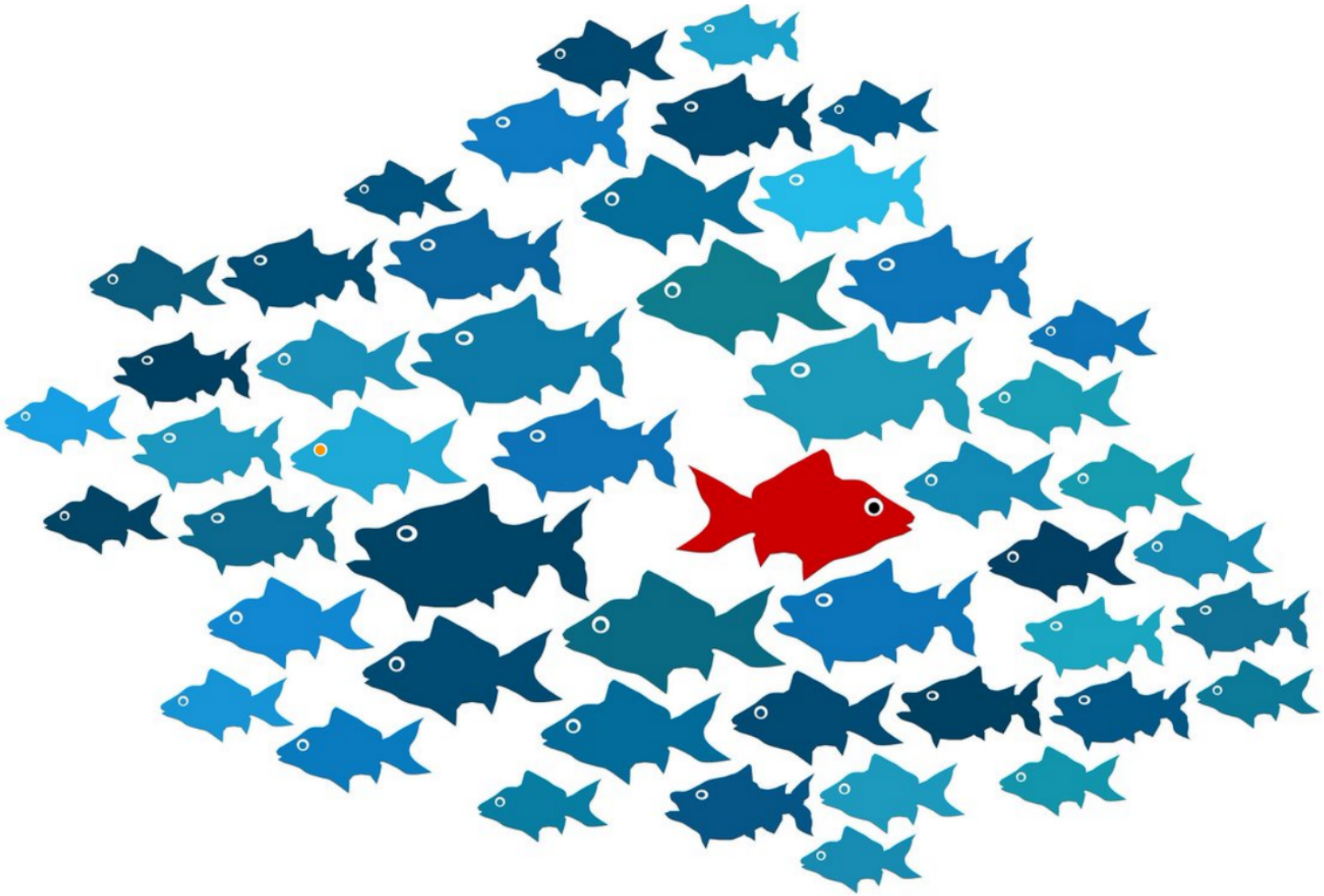
Numerous methods exist to identify anomalous observations; **none of them are foolproof** and judgement must be used.

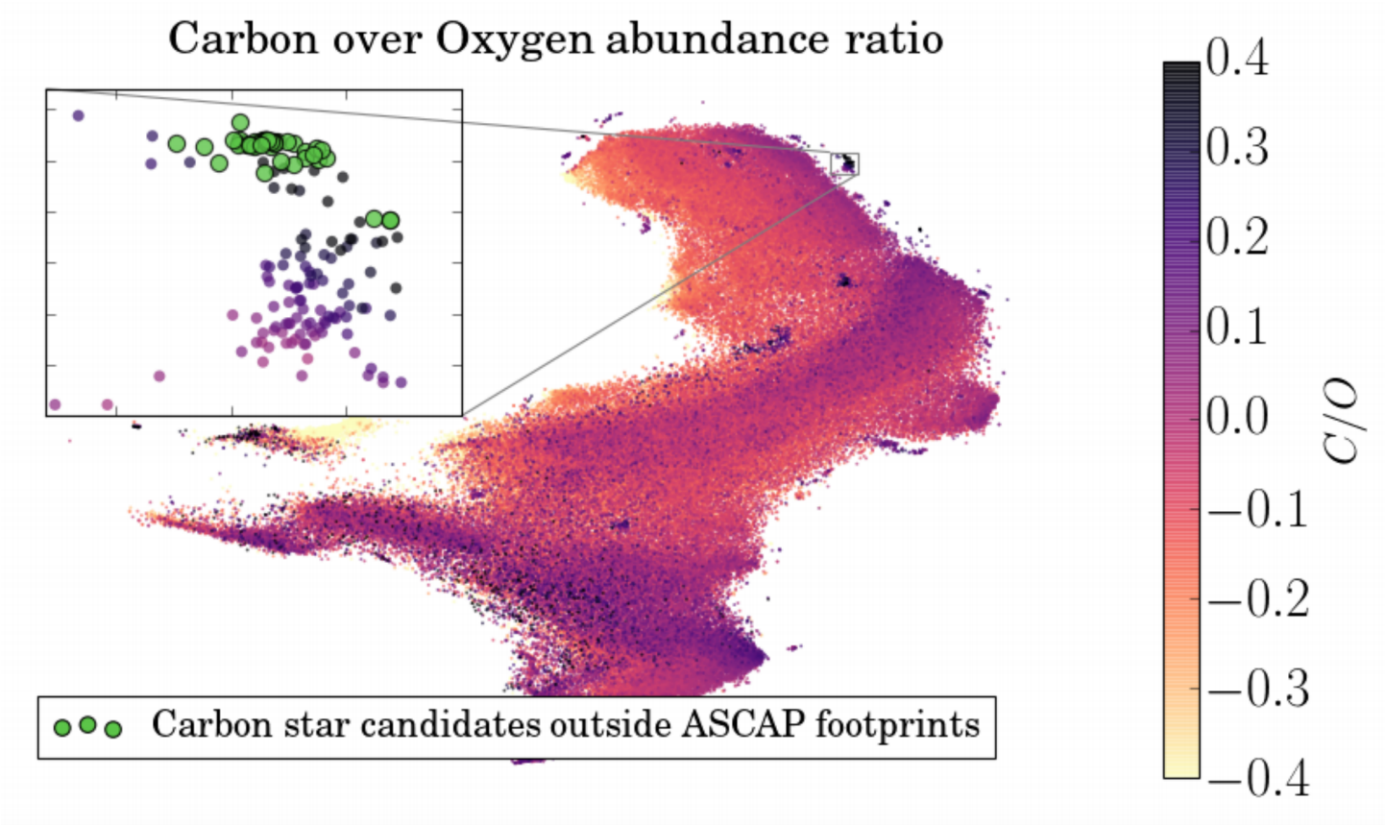
Graphical methods to identify outliers are particularly easy to implement:

- boxplots, scatterplots, scatterplot matrices, and 2D tours

usually require a low-dimensional setting for **interpretability**.

They also usually find those anomalies that "**shout the loudest**" [Baron].





Derived-score anomaly detection may help (... or it may not) [Baron]

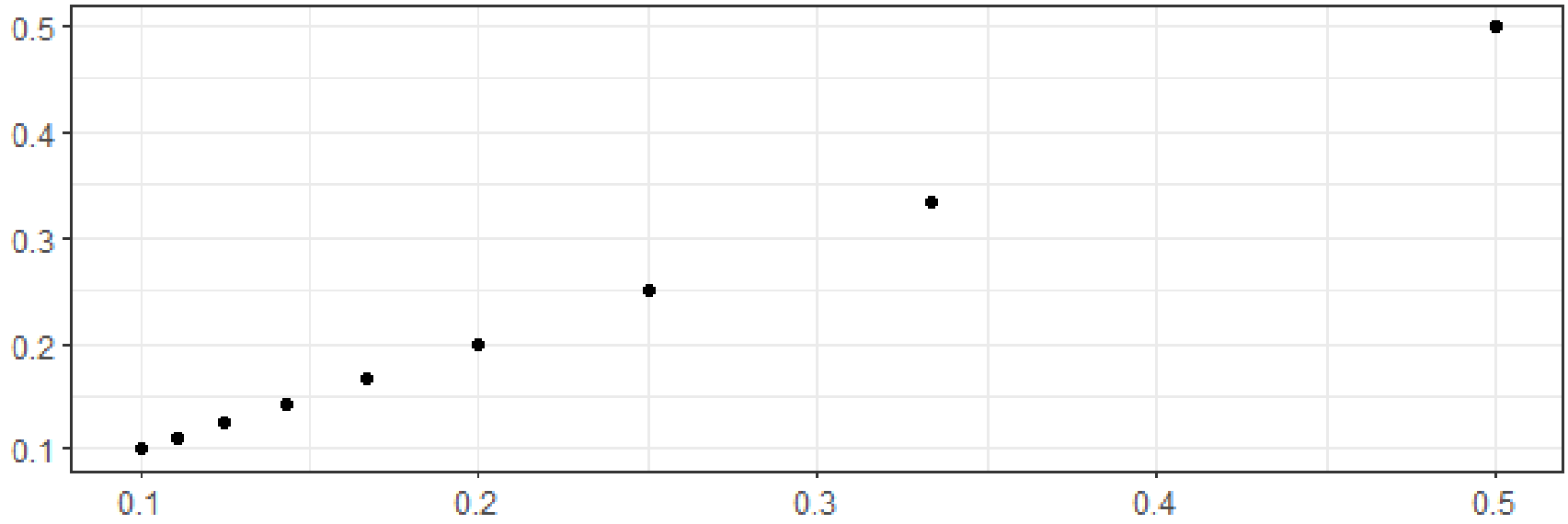
Simple analytical methods using Cooke's or Mahalanobis' distances are sometimes used, but more sophisticated analysis is usually required, especially when trying to identify influential points (*cf.* **leverage**).

In small datasets, detection can be conducted on a case-by-case basis.

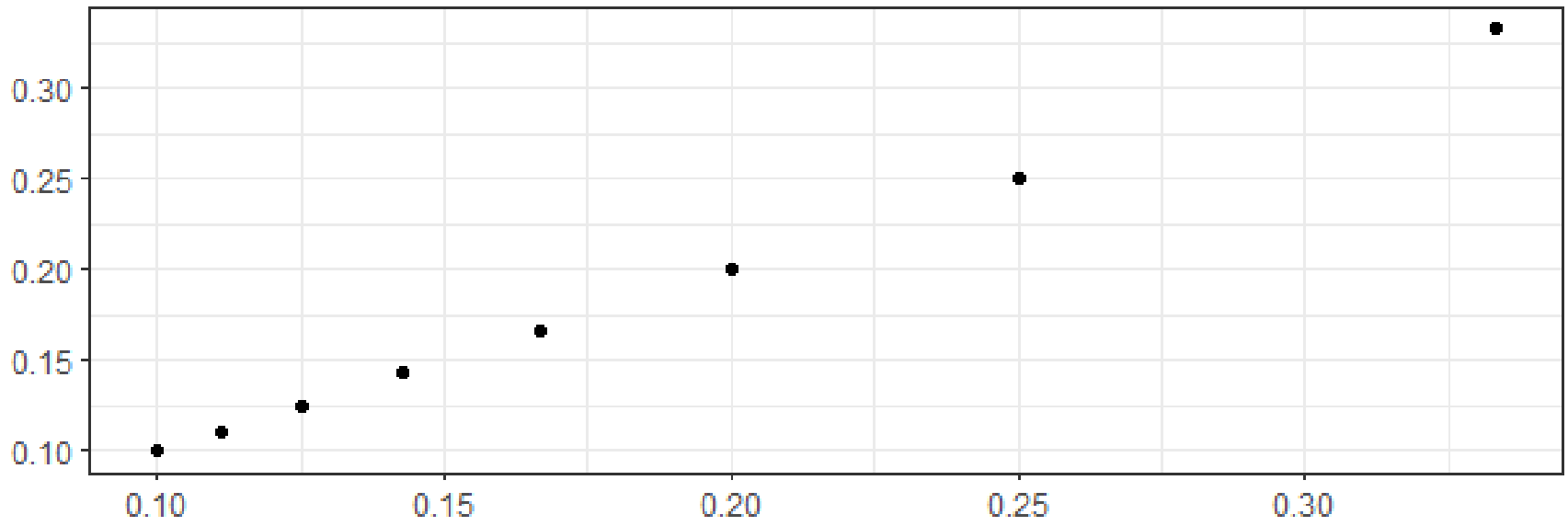
Questions: how many anomalies are too many to find? How many cases are you willing to inspect manually?

It is tempting to use **automated detection/removal** with large datasets, but doing so may be catastrophic from a data analysis perspective!

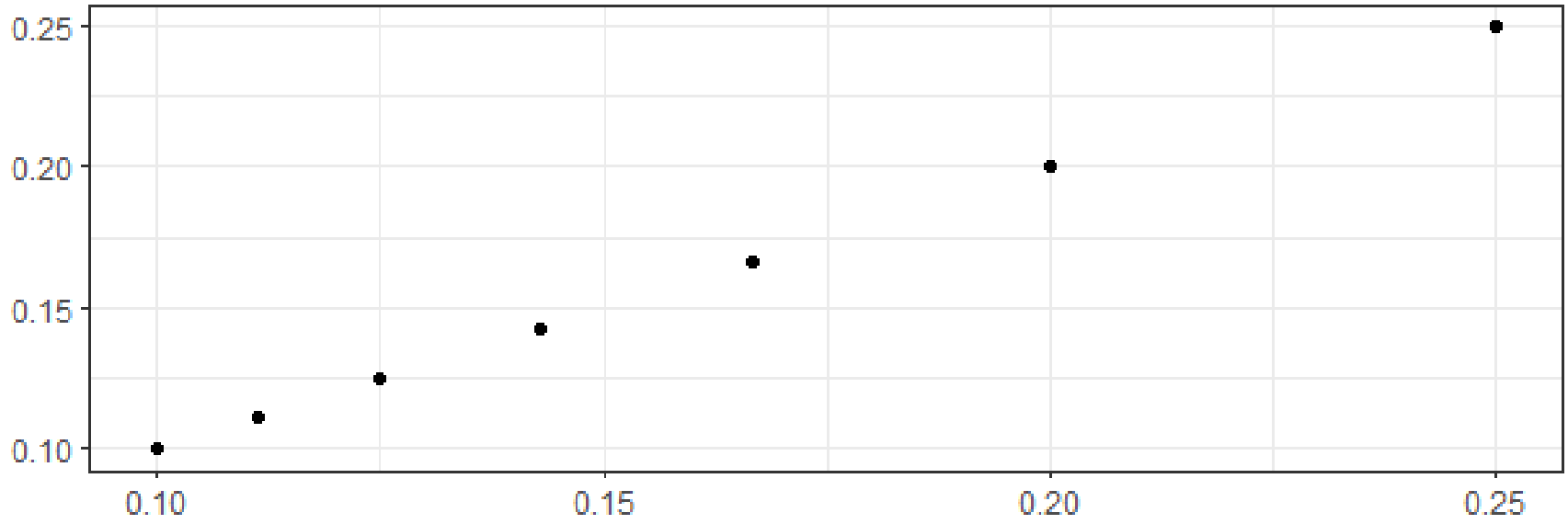
If once "anomalous" observations have been removed from the dataset, previously "regular" observations can become anomalous in turn in the smaller dataset – when does the runaway train?



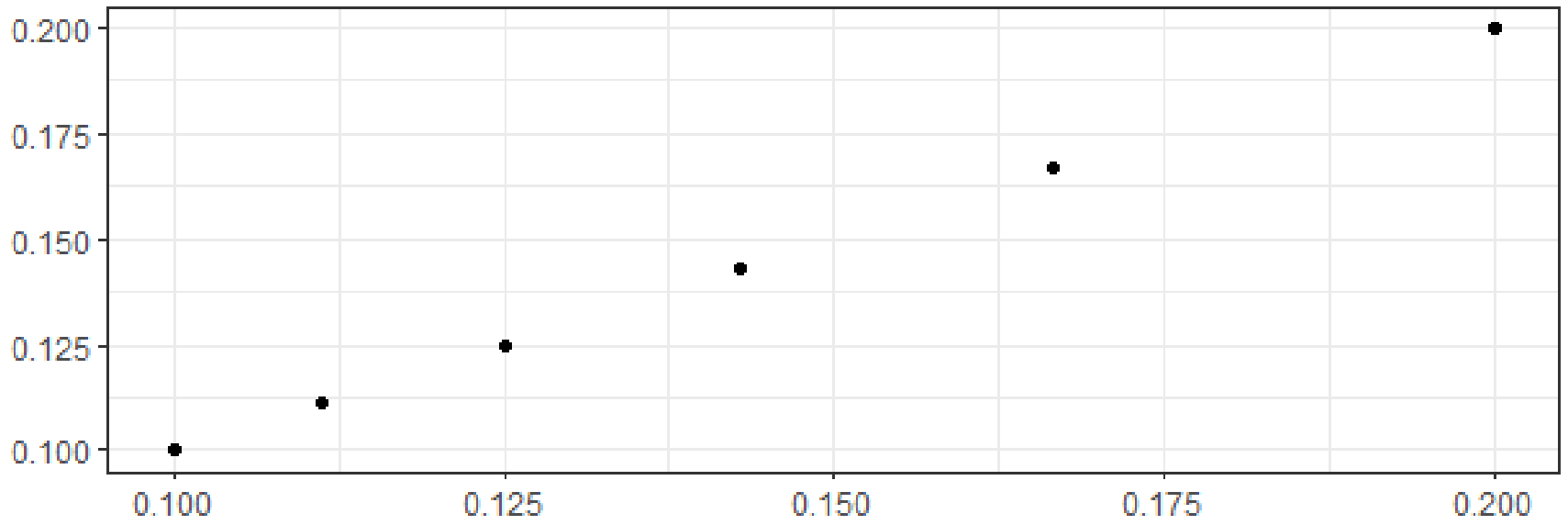
Automatic renewal of anomaly (step 1)



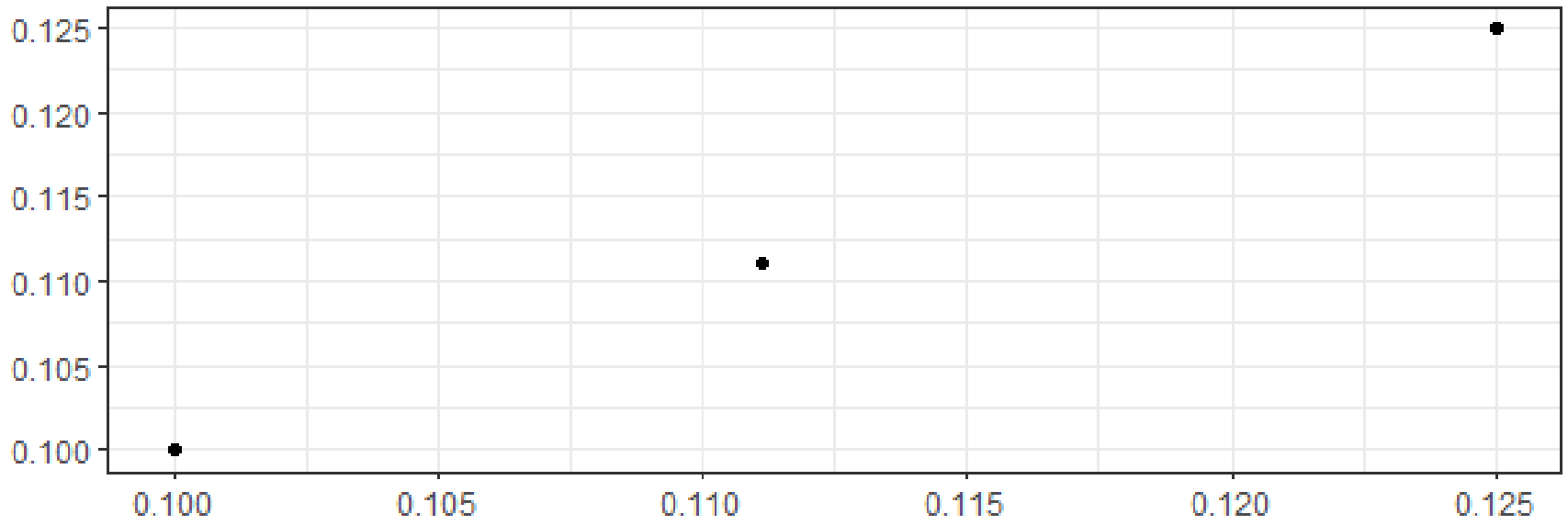
Automatic renewal of anomaly (step 2)



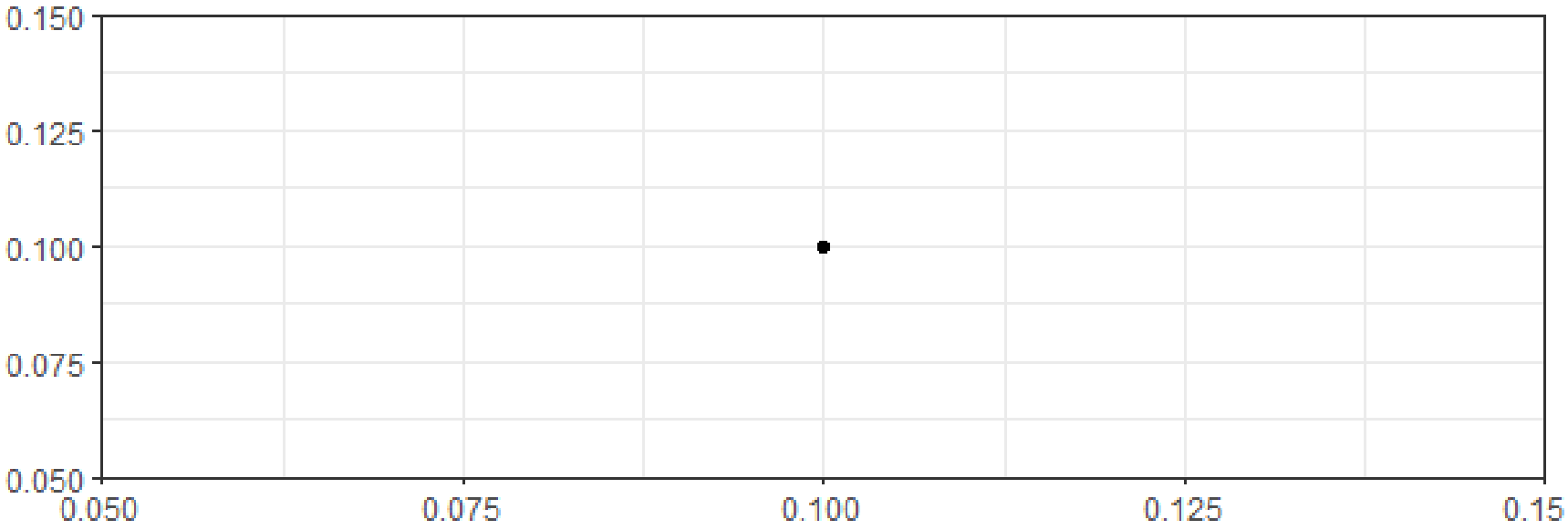
Automatic renewal of anomaly (step 3)



Automatic renewal of anomaly (step 4)



Automatic renewal of anomaly (step n)



Automatic renewal of anomaly (step 9)

In the early stages of anomaly detection, we use **simple data analyses:**

- descriptive statistics,
- $1-$ and $2-$way tables, and
- traditional visualizations.

The goal is to **help identify anomalous observations** and to **obtain insights about the data**.

This leads to more sophisticated anomaly detection methods and could also could eventually lead to modifications of the analysis plan.

THIS IS NEVER AN UNWELCOME DEVELOPMENT!

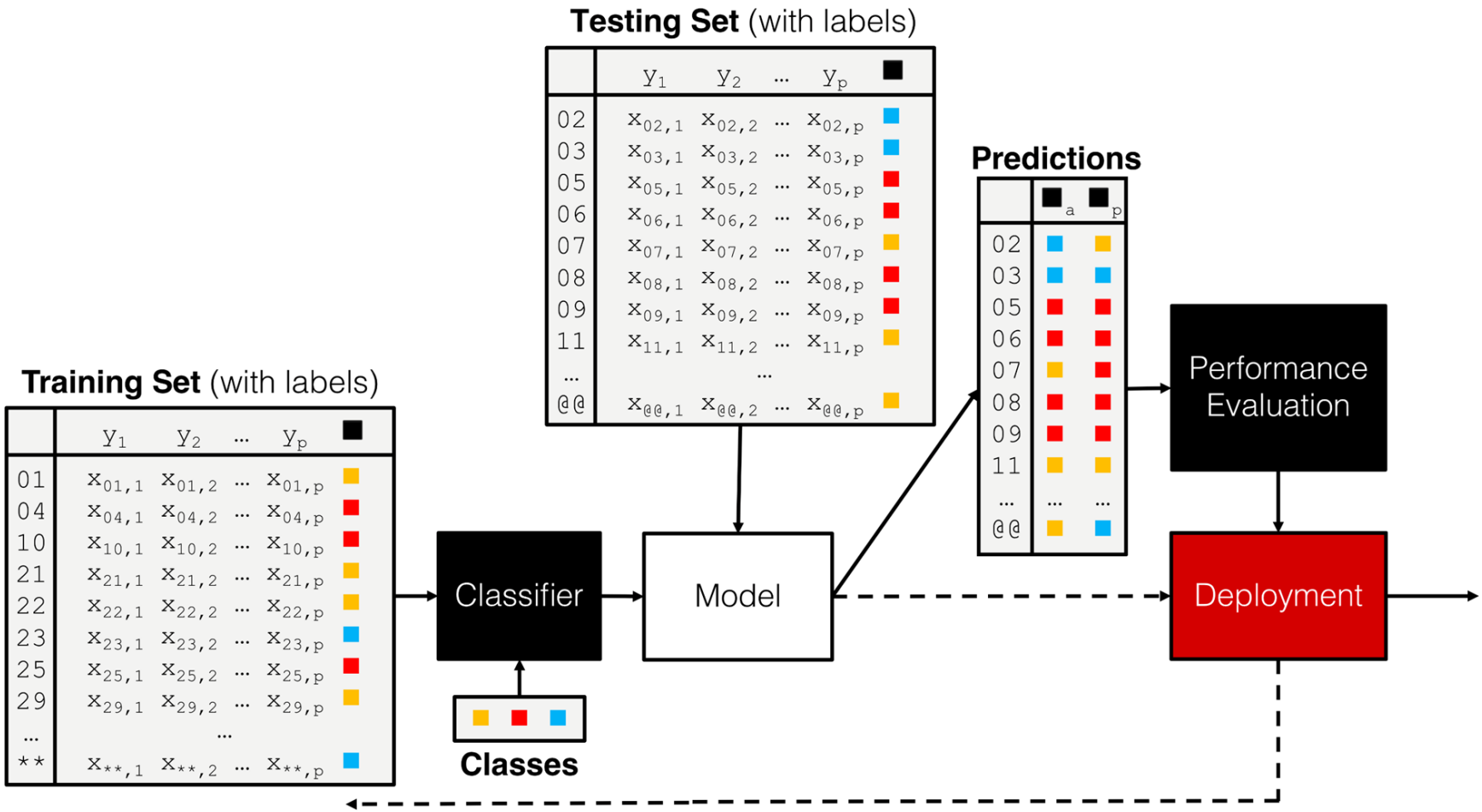
Learning Framework

How are outliers detected, in practice?

Methods come in two flavours:

- **supervised**, and
- **unsupervised**.

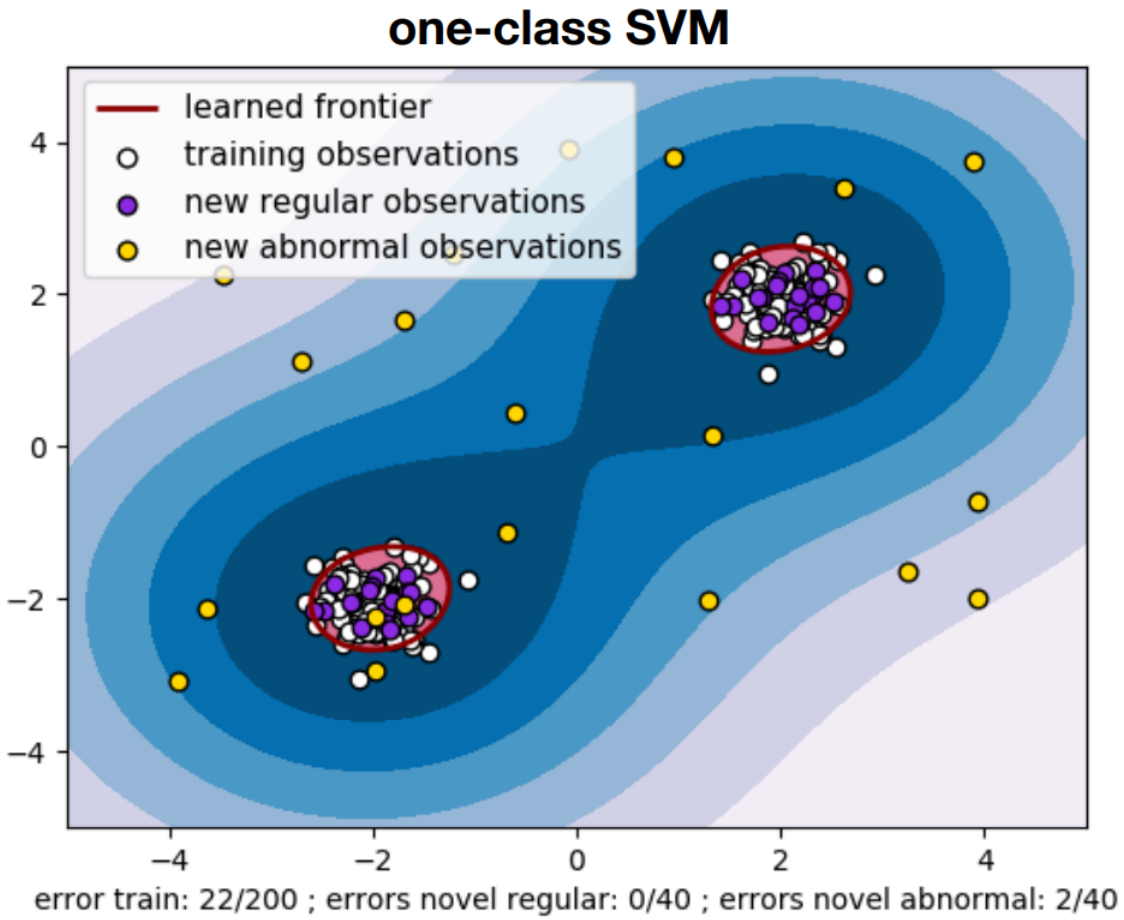
Supervised methods (SL) use a historical record of **previously identified anomalous observations** to build a **predictive classification or regression model** which estimates the probability that a unit is anomalous.


Testing Set (with labels)
Predictions
Training Set (with labels)
Performance Evaluation
Classifier
Model
Deployment
Classes

SL Challenges:

- domain expertise and resources are required to tag the data;
- since anomalies are typically **infrequent**, these models often also have to accommodate the **rare occurrence** (or class imbalance) problem, and
- SL methods need to minimize a **loss function** (cost of making a mistake) which is usually symmetrical (in the anomaly detection context, this is not usually a valid assumption).

Even more than in traditional analysis settings, anomaly detection can lead to **technically correct but ultimately useless** (non-actionable) **results**.



Learning an **anomaly frontier** [Baron].

Example: The vast majority ($99.999+\%$) of air passengers **do not** bring weapons with them on flights.

A model that predicts that no passenger is ever attempting to smuggle a weapon on board a flight 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 ... somewhat different take on this, from a societal and personal perspective.

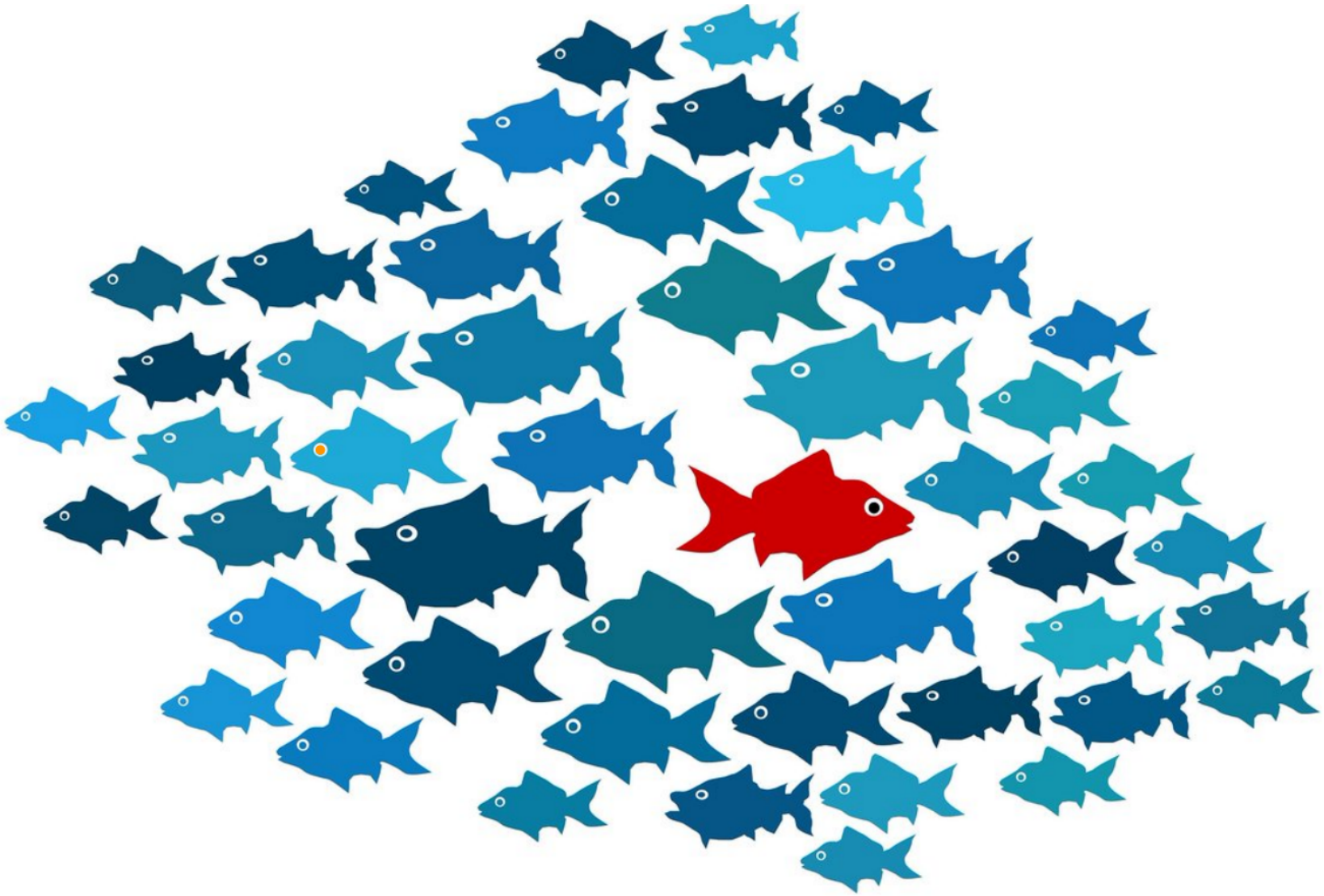
Unsupervised methods (UL)

- use no previously labeled (anomalous/non-anomalous) data, and
- try to determine if an observation anomalous solely by comparing its behaviour to that of the other observations.

Example: if all workshop participants 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.

VERY IMPORTANT NOTE: this **DOES NOT** mean that the different behaviour is the one we are actually interested in/searching for!

Be weary: this is true of anomaly detection in data and in real-life.



Traditional Outlier Detection Tests

The most commonly-used test is **Tukey's** (univariate) **boxplot test**. Let Q_1 and Q_3 represent an observed feature's 1^{st} and 3^{rd} quartile, respectively.

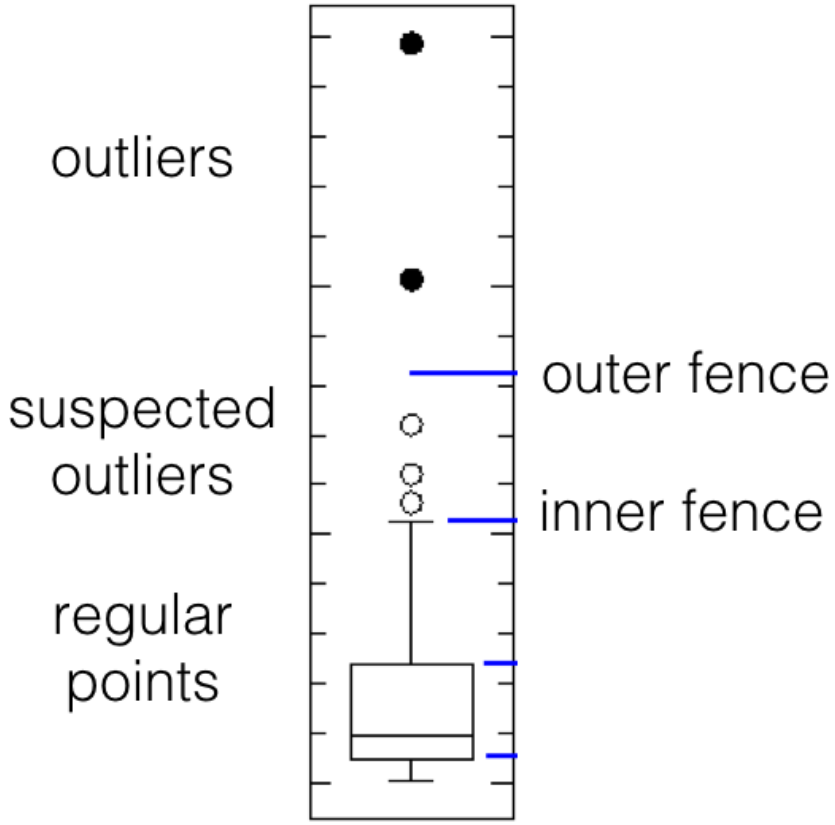
For **normally distributed** measurements, regular observations typically lie between the **inner fences**

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

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

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

Points beyond the outer fences are identified as **outliers**.



Suspected outliers are marked by white disks, outliers by black disks.
Use `boxplot()` and `boxplot.stats()` in `R` to plot and identify outliers in normally distributed data.

The **Grubbs test** is another univariate test:

H_0 : no outlier in the data vs. H_1 : **exactly one** outlier in the data.

- let x_i be the value of feature X for the i^{th} unit, $1 \leq i \leq N$,
- let $(\overline{x}, s_x)$ be the mean and standard deviation of feature X,
- let α be the desired significance level, and
- let $T(\alpha/2N; N)$ be the critical value of the Student t-distribution.

The test statistic is

$$G = \frac{\max_i\{|x_i - \overline{x}|\}}{s_x} = \frac{|x_{i^*} - \overline{x}|}{s_x}.$$

Under H_0, G follows a special distribution with critical value

$$\ell(\alpha; N) = \frac{N-1}{\sqrt{N}} \sqrt{\frac{T^2(\alpha/2N, N)}{N - 2 + T^2(\alpha/2N, N)}}.$$

At significance level α, we reject the null hypothesis in favour of the alternative (i.e. x_{i^*} is the outlier) if $G \geq \ell(\alpha; N)$.

If looking for more than one outlier, it can be tempting to classify every observation i for which

$$\frac{|x_i - \overline{x}|}{s_x} \geq \ell(\alpha; N)$$

as an outlier, but this is **NOT RECOMMENDED**.

Other generalizations are also problematic (cf. outlier sequence).

Other common tests include:

- the **Mahalanobis distance**, which is linked to the leverage of an observation (a measure of influence), can also be used to find multi-dimensional outliers, when all relationships are linear (or nearly linear);
- the **Tietjen-Moore** test, which is used to find a specific number of outliers (this is similar to Grubbs' test, replacing H_1 by H_k);
- the **generalized extreme studentized deviate** test, the preferred extension to Grubbs' test if the number of outliers is unknown;
- the **chi-square** test, when outliers affect the goodness-of-fit;
- DBSCAN and other clustering-based outlier detection methods.

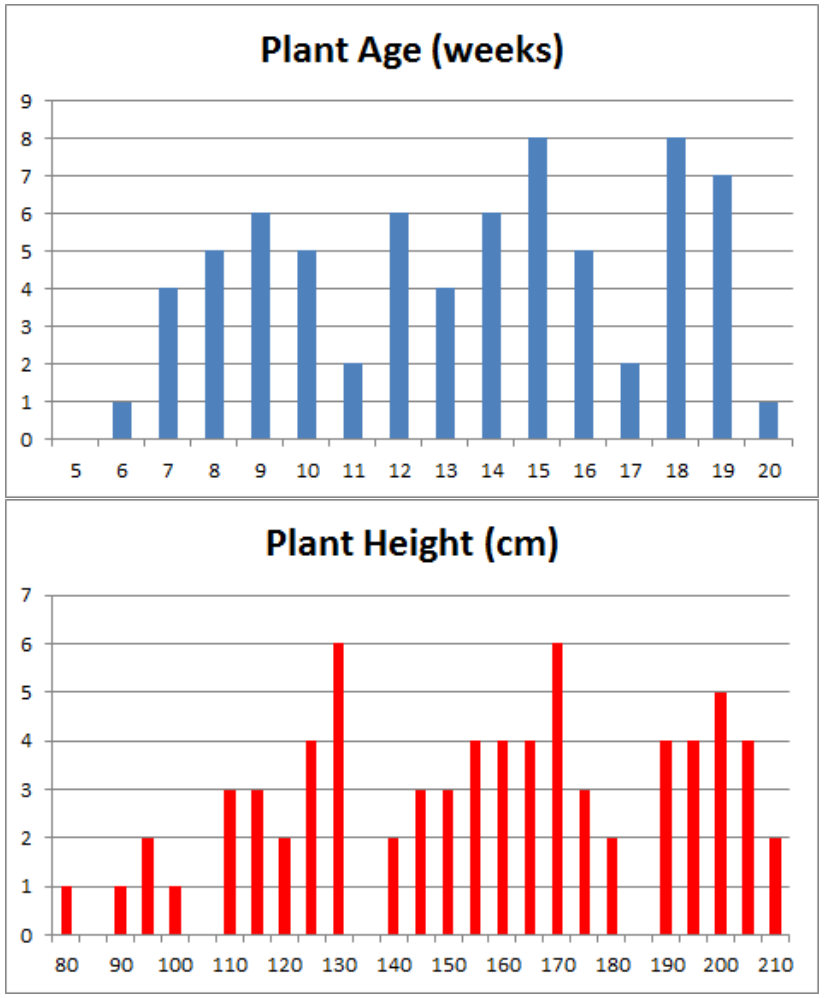
Visual Outlier Detection

The following simple examples illustrate the principles underlying **visual outlier and anomaly detection**.

Example 1: on a specific day, the **height** of several plants in a nursery are measured. The records also show each plant's **age** (the number of weeks since the seed has been planted).

Very little can be said about the data at that stage:

- the age of the plants (controlled by the nursery staff) seems to be somewhat haphazard,
- as does the response variable (height).


Plant Age (weeks)
9
8
7
6
5
4
3
2
1
0
5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Plant Height (cm)
7
6
5
4
3
2
1
0
80 90 100 110 120 130 140 150 160 170 180 190 200 210

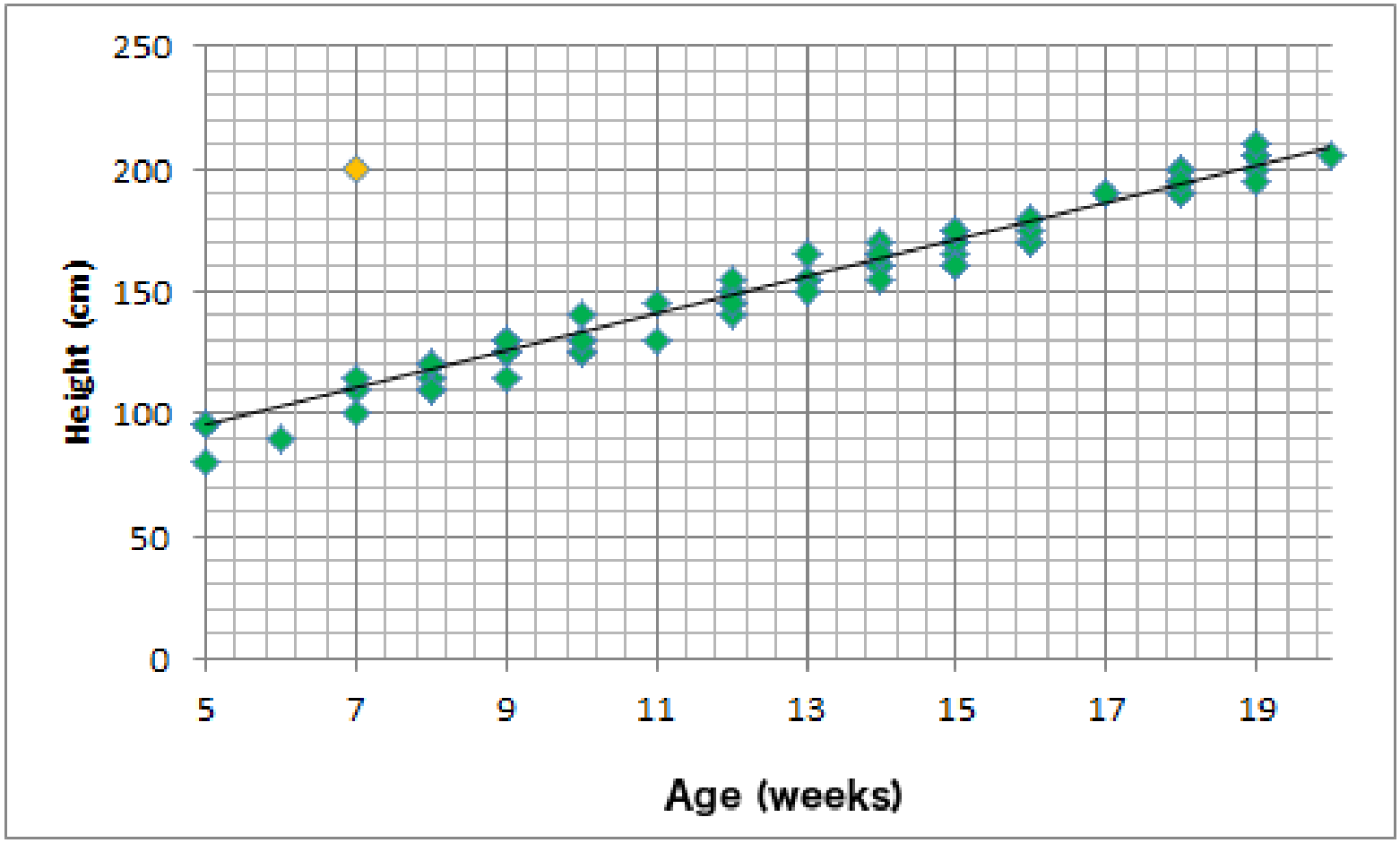
A scatter plot of the data reveals that **growth is strongly correlated with age** for the observations in the dataset; points clutter around a linear trend.

One point (in yellow) is easily identified as an **outlier**.

There are (at least) two possibilities:

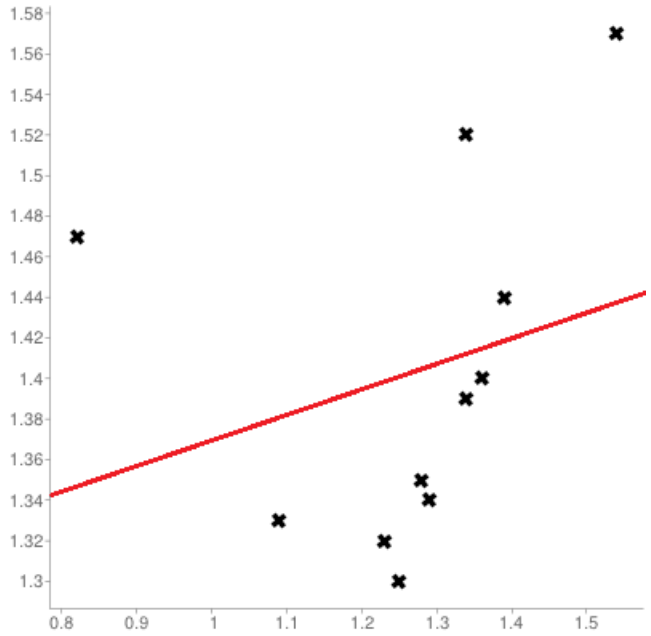
- either that measurement was botched or mis-entered in the database (representing an invalid entry), or
- that one specimen has experienced unusual growth (outlier).

Either way, the analyst has to investigate further.

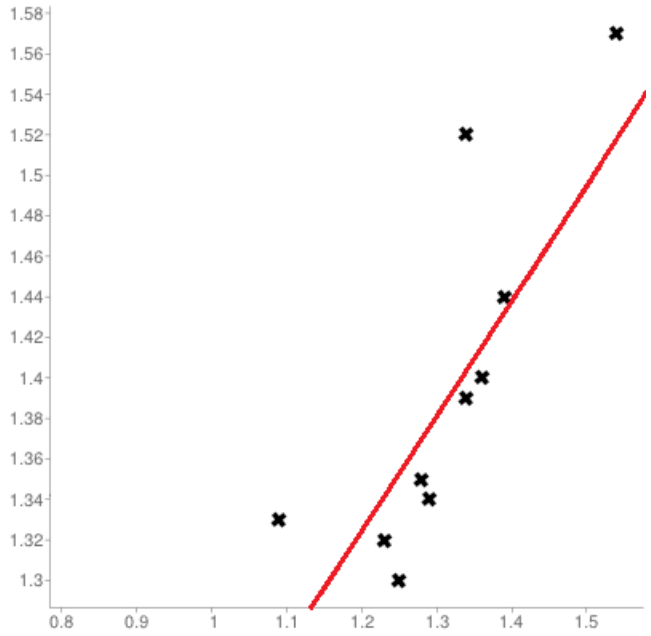

250
200
150
100
50
0
Height (cm)
5
7
9
11
13
15
17
19
Age (weeks)

Example 2: a government department has 11 service points. The monthly average arrival and service rates per teller for each service point are available.

The scatter plot of the service rate per teller (y axis) against the arrival rate per teller (x axis), with linear regression trend, is shown below.



A similar chart, but with the left-most point removed from consideration, is shown below.



The trend still slopes upward, but the fit is significantly improved.

This suggests that the removed observation is unduly **influential** (or anomalous) – a better understanding of the relationship between arrivals and services is afforded if it is set aside.

Any attempt to fit that data point into the model must take this information into consideration.

The status of an influential observations **depends on the analysis that is ultimately conducted** – a point may be influential for one analysis, but not for another.

Note that setting aside an influential observation does not mean that the observation is removed from the dataset – only that it will not be used in a specific analysis.

Example 3: Measurements of the length of the appendage of a certain species of insect have been made on 71 individuals. Descriptive statistics have been computed; the results are shown below.

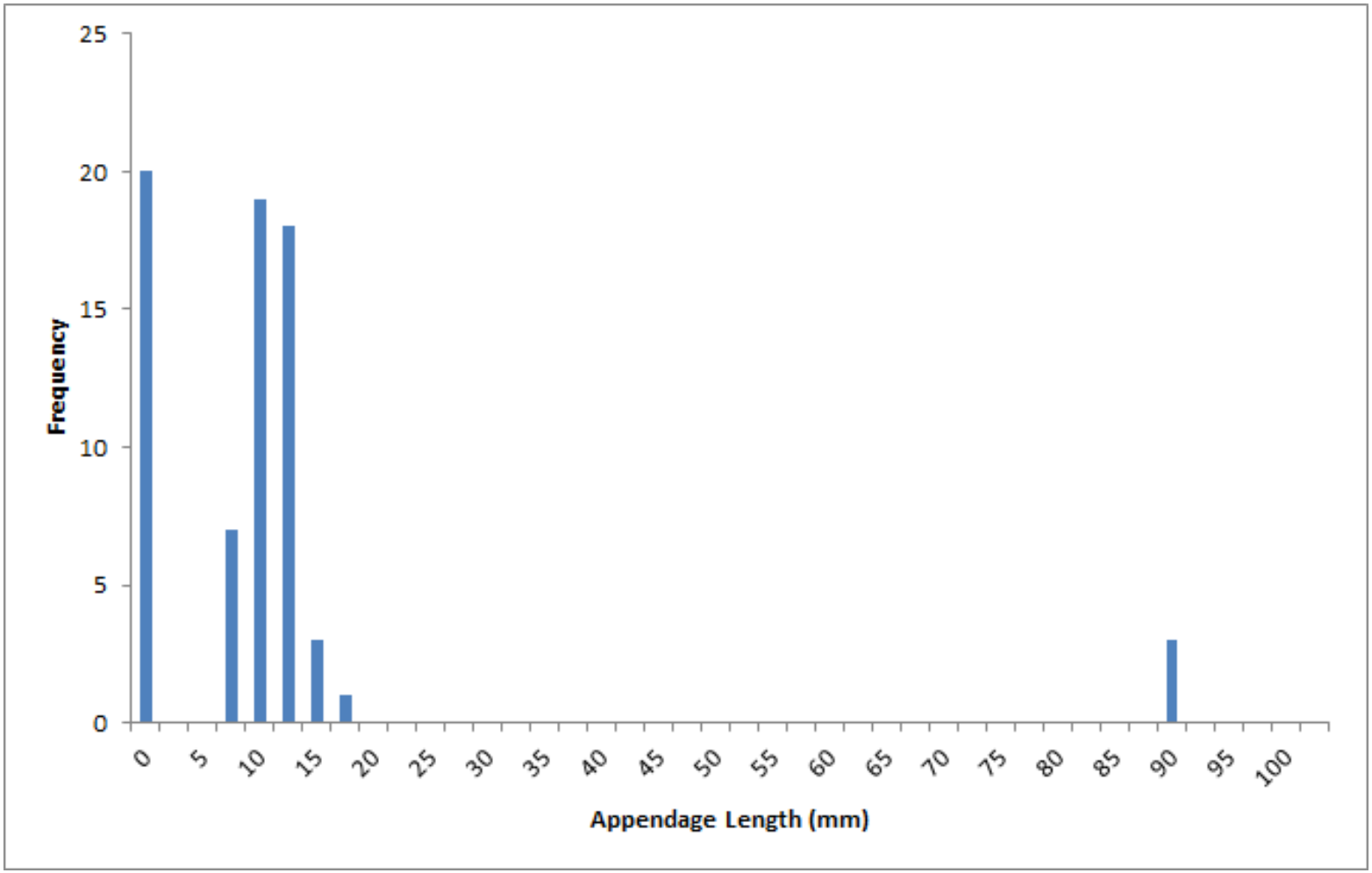
Appendage length (mm)	
Mean	10.35
Standard Deviation	16.98
Kurtosis	16.78
Skewness	4.07
Minimum	0
First Quartile	0
Median	8.77
Third Quartile	10.58
Maximum	88
Range	88
Interquartile Range	10.58
Mode	0
Count	71

The descriptive statistics might help the analyst recognize the tell-tale signs that the distribution of appendage lengths is likely to:

- be **asymmetrical** (since skewness is non-negligible), and
- have a **"fat" tail** (since large kurtosis, range $\gg$ interquartile range, and max $\gg Q_3$)

The mode, min, and Q_1 belong to individuals without appendages $\Longrightarrow$ at least two sub-groups in the population (perhaps split along the lines of juveniles/adults, or males/females).

Since max $\gg$ other observations, might it belong to an **outlier**? The histogram of the measurements shows 3 individuals with long appendages.


25
20
15
10
5
0
Frequency
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
Appendage Length (mm)

It is plausible that these individuals belong to another species who were **erroneously added** to the dataset.

On its own, the chart does not constitute a proof of such an error, but it **raises the possibility of an error**, which is often the best that an analyst can do in the absence of subject matter expertise.

This traditional approach to anomaly detection is difficult to apply to high-dimensional datasets because it is nearly impossible to visualize them directly $\Longrightarrow$ fundamentally different approaches are needed.

Dimension reduction methods can be used to provide a **low-dimensional representation** of the data on which to apply visual detection (see autoencoder example), but some information always gets lost in the process.

5.1.1 – Anomaly Detection as Statistical Learning

Fraudulent behaviour is not always easily identifiable, even after the fact.

Example: credit card fraudsters try to disguise their transactions as regular and banal, and try to avoid outlandish behaviour.

Their goal: fool human observers into confusing **plausible** (or possible) with **probable** (or at least, **not improbable**).

It is plausible that a generic 40-something father of 3 might purchase a new TV; is it probable that THIS particular father of 3 would do so?

But it's unlikely that a generic father of 3 who resides in North America would purchase a round of drinks at a dance club in Kiev.

Anomaly detection is really a problem in **applied probability**. Let I be what is known about the dataset/situation:

- behaviour of individual observations,
- behaviour of observations as a whole,
- anomalous/normal verdict for a number of similar observations, etc.

Main Question: is $P(\text{obs. is anomalous} \mid I) > P(\text{obs. is normal} \mid I)$?

Anomaly detection models assume **stationarity of regular observations**: that the underlying mechanism that generates regular data does not change much over time.

For time series data, this means that it may be necessary to first perform **trend and seasonality extraction**.

Example: supply chains play a crucial role in the transportation of goods from one part of the world to another – as the saying goes, "a given chain is only as strong as its weakest link."

Say that marine cargo departing Shanghai in Feb'13 took two more days, on average, to arrive in Vancouver than those departing in Jul'17.

- Has the shipping process improved in the intervening years?
- Do departures in Feb usually take longer to reach Vancouver?
- Are either the Feb'13 or the Jul'17 performance anomalous?

Seasonal variability is relevant to supply chain monitoring: quantifying and accounting for impact severity is of great interest.

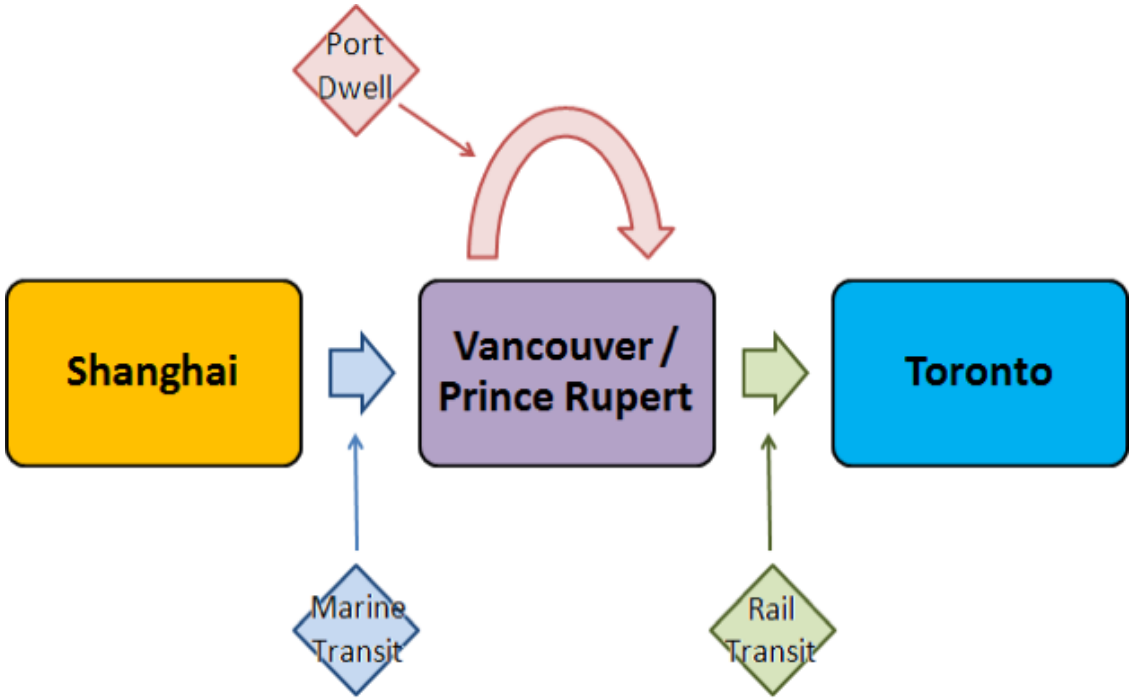
Potential Solution: create an **index** to track container transit times.

This index should depict

- **reliability** and
- **variability** of transit times,
- and allow for performance comparison between differing time periods.

Consider the scenario where we want to compare the monthly performance, irrespective of the transit season, of the corridor

Shanghai → Port Metro Vancouver/Prince Rupert → Toronto.



For each of the three segments (Marine Transit, Port Dwell, Rail Transit), the data consists of:

- monthly empirical distribution of transit/dwell times
- built from sub-samples (assumed to be randomly selected and fully representative) of all containers entering the appropriate segment.

Specific containers are not followed from Shanghai to Toronto: no covariance information about the various transit/dwell times is available.

Each segment's performance is measured using **fluidity indicators**, which are computed using various statistics of the transit/dwell time distributions for each of the supply chain segments.

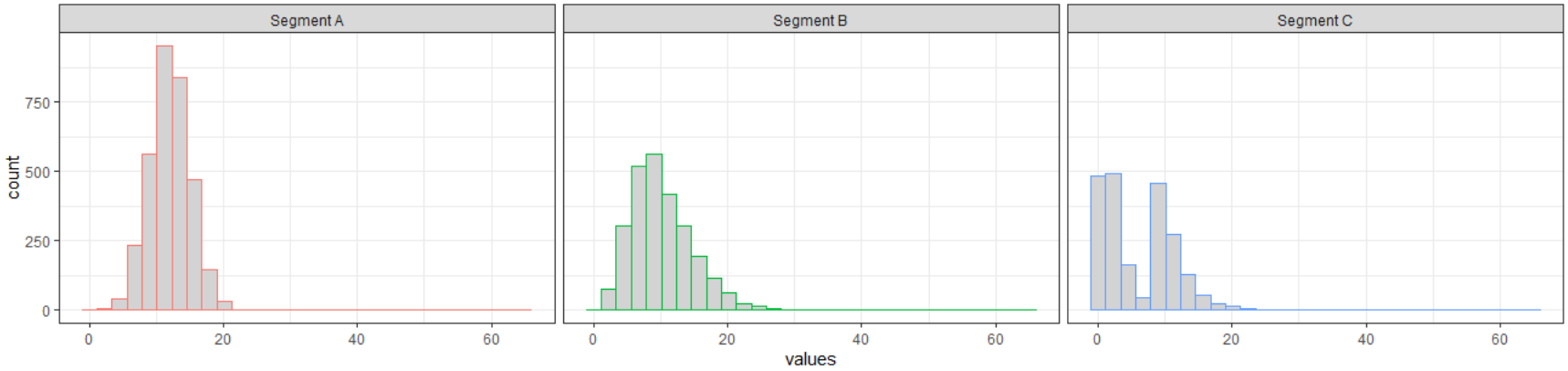
Reliability Indicator (RI) – the ratio of the 95^{th} percentile to the 5^{th} percentile of transit/dwell times.

A high RI indicates high volatility, whereas a low RI (≈ 1) indicates a reliable corridor.

Buffer Index (BI) – the ratio of the positive difference between the 95^{th} percentile and the mean, to the mean.

A small BI (≈ 0) indicates only slight variability in the upper (longer) transit/dwell times; a large BI indicates that the variability of the longer transit/dwell times is high, and that outliers might be found there;

Coefficient of Variation (CV) – the ratio of the standard deviation of transit/dwell times to the mean transit/dwell time.

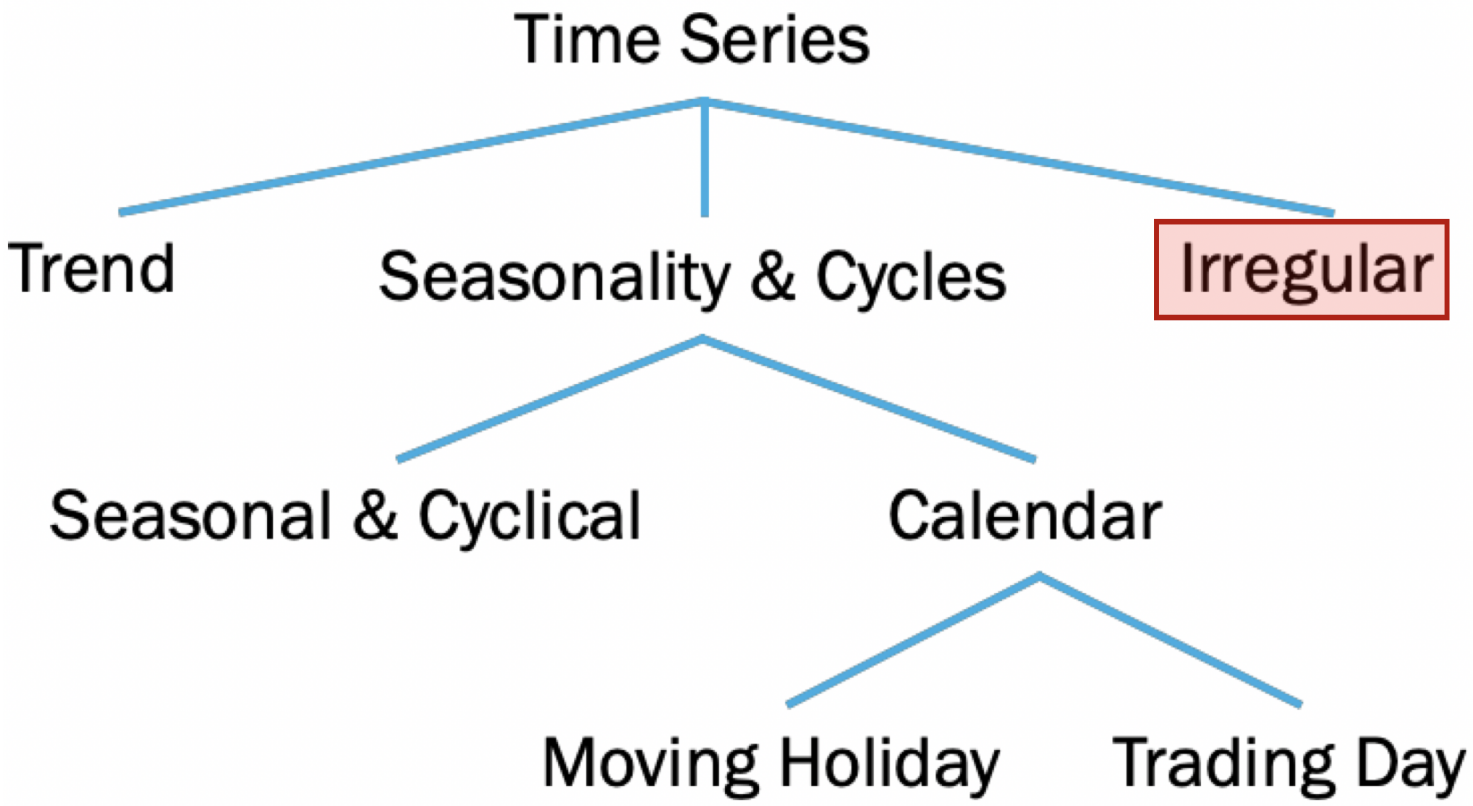


Segmt	Freq	Mean	SD	C05	C95	RI	BI	CV
A	3286	12.10	3.33	7.06	17.00	2.41	0.41	0.27
B	2594	10.09	4.43	3.88	18.20	4.69	0.80	0.44
C	2142	5.96	5.08	0.19	14.40	77.12	1.41	0.85

The time series of monthly fluidity indicators are then **decomposed** into:

- trend $\Longrightarrow$ **expected behaviour**
- seasonal component (seasonality, trading-day, moving-holiday) $\Longrightarrow$ **expected behaviour**
- structural breaks, $\Longrightarrow$ **explained unexpected behaviour** and
- irregular component $\Longrightarrow$ chain **volatility**

A high irregular component at a given time indicates a poor performance against expectations $\Longrightarrow$ an **anomalous observation**.



Conceptual time series decomposition (after structural breaks are removed); potential anomalous behaviour should be searched for in the **irregular component**.

In general, the decomposition follows a model which is

- multiplicative;
- additive, or
- pseudo-additive.

The choice of a model is driven by data behaviour and choice of assumptions; the X12 model automates some of the aspects of the decomposition, but manual intervention and diagnostics are still required.

IMPORTANT NOTE: anomaly detection often requires modeling choices/ assumptions.

The **additive model**, for instance, assumes that:

1. the seasonal component S_t and the irregular component I_t are independent of the trend T_t;

2. the seasonal component S_t remains stable from year to year; and

3. there is no seasonal fluctuation: $\sum_{j=1}^{12} S_{t+j} = 0$.

Mathematically, the model is expressed as:

$$O_t = T_t + S_t + I_t$$

All components share the same dimensions and units.

After seasonality adjustment,the seasonality adjusted series is:

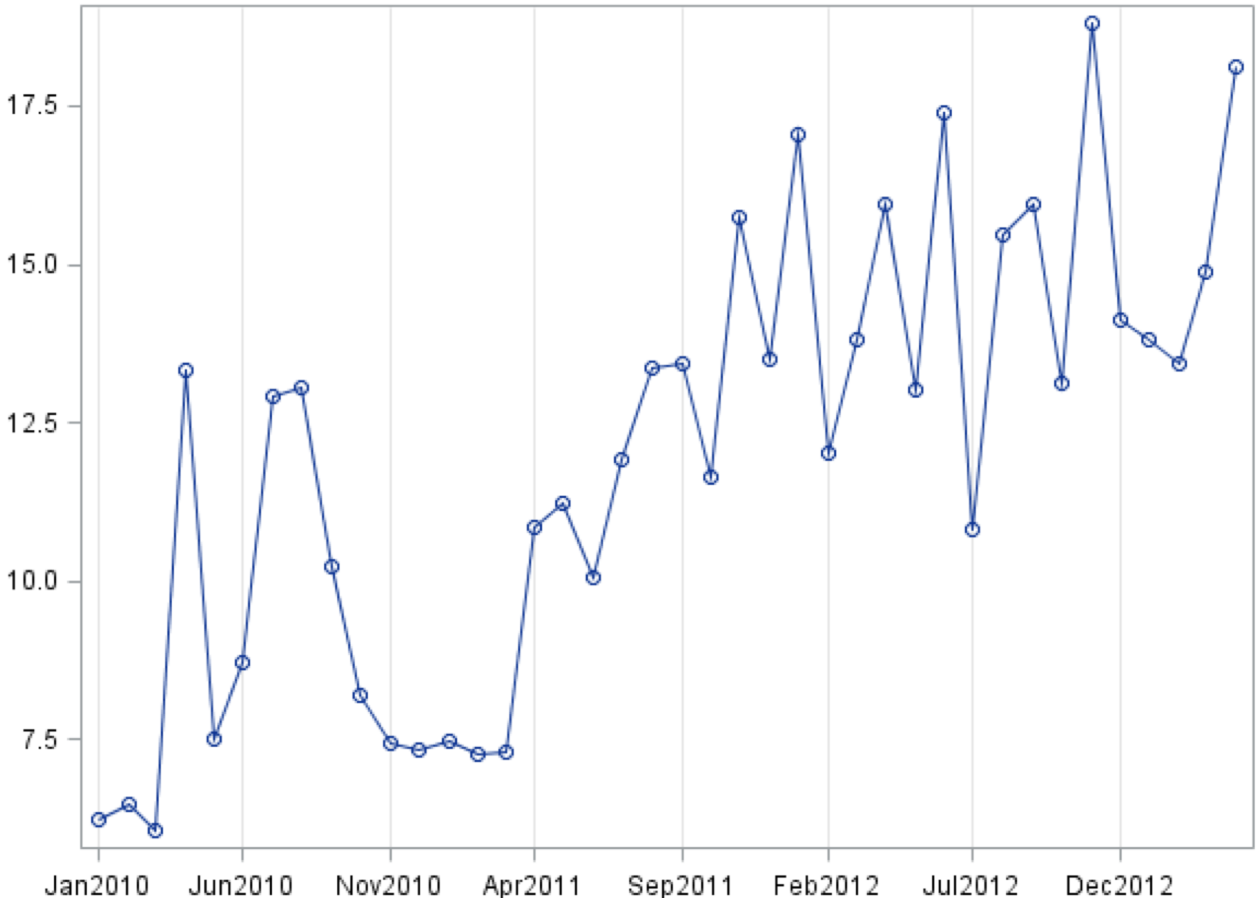
$$SA_t = O_t - S_t = T_t + I_t$$

The multiplicative and pseudo-additive models are defined in similar ways:

- if the size of S_t increases/decreases over time, use s multiplicative model;
- otherwise, use an additive model.

The data decomposition/preparation process is illustrated with the 40-month time series of marine transit CVs from 2010-2013.

The size of the peaks and troughs seems fairly constant with respect to the changing trend $\Longrightarrow$ use the additive model.

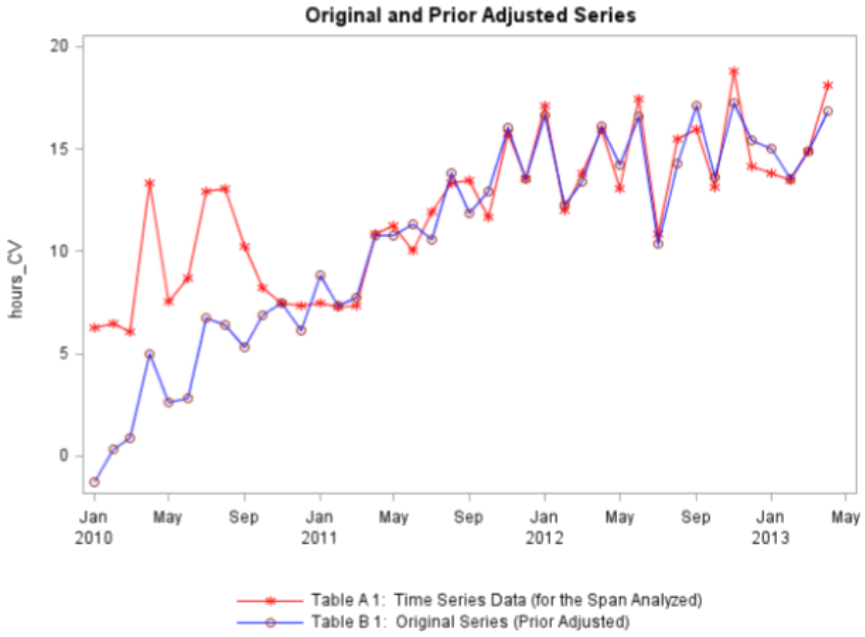


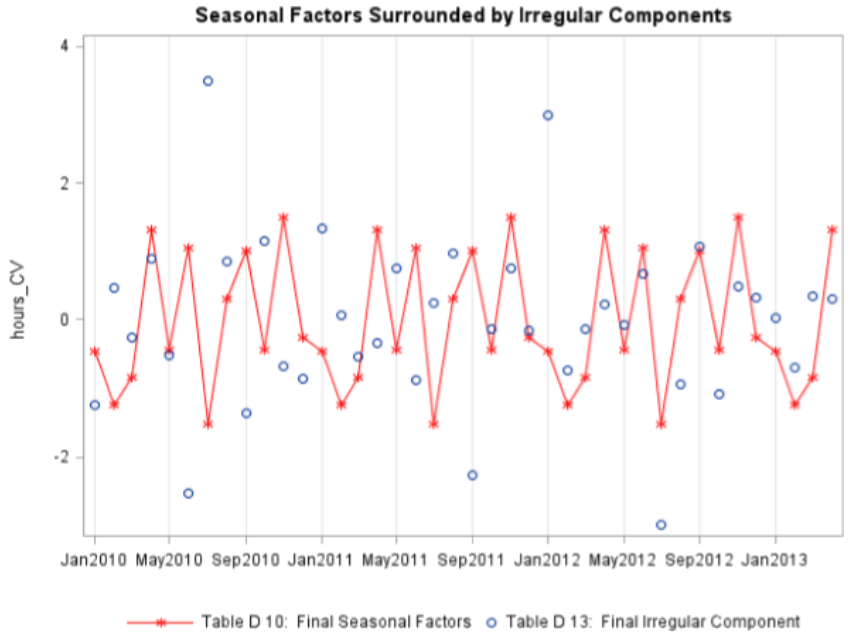
Estimation Summary	
For Variable hours_CV	
Number of Observations	40
Number of Residuals	27
Number of Parameters Estimated	3
Variance Estimate	5.6E-02
Standard Error Estimate	2.4E-01
Standard Error of Variance	1.5E-02
Log likelihood	0.4658
Transformation Adjustment	-69.5685
Adjusted Log likelihood	-69.1027
AIC	144.2053
AICC (F-corrected-AIC)	145.2488
Hannan Quinn	145.3613
BIC	148.0928

Results of Automatic Transformation Selection	
For Variable hours_CV	
AICC (with aicdiff=-2.00) prefers	No transformation
Adjustment will be	Additive

A structural break (trend level shift) is identified at OCT2010.

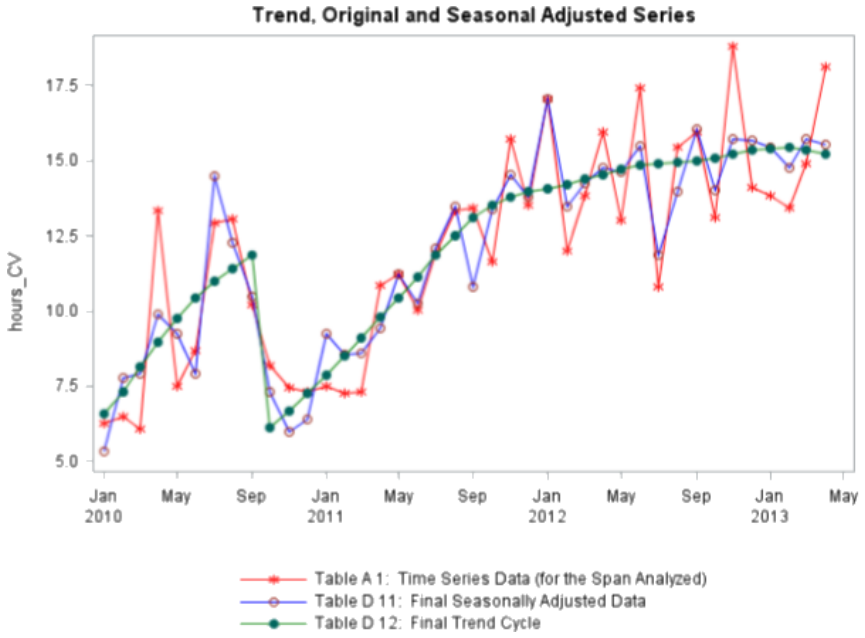
The SI (Seasonal Irregular) chart shows that there are more than one irregular component which exhibits volatility.

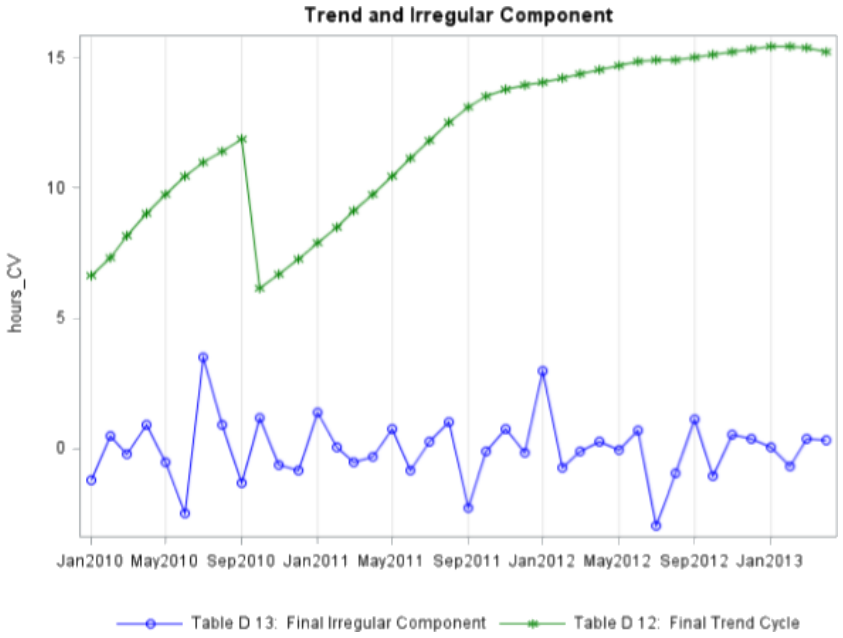




The adjusted series is shown below; the trend and irregular components are also shown separately for readability.

It is on the irregular component that detection anomaly would be conducted.





Given that the vast majority of observations in a general problem are typically "normal", another conceptually important approach is to view anomaly detection as a:

- **rare occurrence learning** classification problem, or
- **novelty detection** data stream problem.

While there a number of strategies that use regular classification/clustering algorithms for anomaly detection, they are rarely successful unless they are **adapted** or **modified for the anomaly detection context**.

MORAL OF THE STORY: anomaly detection is a difficult problem.

Basic Concepts

Generic systems (think of the monthly transit/dwell times from supply chains) may be realized in

- **normal** states, or
- **abnormal** states.

Normality is not confined to finding the most likely state – infrequently occurring states could still be normal or plausible under some interpretation of the system.

A system's states are the results of processes or behaviours that follow certain **natural rules** and **broad principles**; the observations are a manifestation of these states.

Data allows for inferences to be made about the underlying processes, which can be tested or invalidated by the collection of additional data.

When the inputs are perturbed, the corresponding outputs are likely to be perturbed as well.

If anomalies arise from perturbed processes, the **anomaly detection problem** could be helped along by being able to identify when the underlying process is abnormal.

Supervised anomaly detection algorithms require a

1. **training set of historical labeled data** on which to build the prediction model (usually costly to obtain), and

2. testing set on which to evaluate the model's performance in terms of

   - **True Positives** (TP) – detected anomalies that actually arise from process abnormalities;
   - **True Negatives** (TN) – predicted normal observations that indeed arise from normal processes;
   - **False Positives** (FP) – detected anomalies corresponding to regular processes, and
   - **False Negatives** (FN) – predicted normal observations that are in fact the product of an abnormal process.

This is often summarized in a **confusion matrix**:

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

Naïvely, one might look for an algorithm which maximizes the **accuracy**

$$a = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TN} + \mathrm{TP} + \mathrm{FN} + \mathrm{FP}}.$$

For rare occurrences, this is a losing strategy (see weapon smuggling ex.).

Better approach: try to minimize the FP rate and the FN rate under the assumption that the **cost of making a false negative error could be substantially higher than the cost of making a false positive error**.

For a testing set with $d = \mathsf{FN} + \mathsf{TP}$ **true outliers**, assume that an anomaly detection algorithm identifies $m = \mathsf{FP} + \mathsf{TP}$ **suspicious observations**, of which $n = \mathsf{TP}$ are **known** to be true outliers.

How well did the algorithm **perform**?

Precision: proportion of true outliers among suspicious observations

$$p = \frac{n}{m} = \frac{\mathsf{TP}}{\mathsf{FP} + \mathsf{TP}};$$

if most of the suspicious points are true outliers, $p \approx 1$;

Recall: proportion of true outliers detected by the algorithm

$$r = \frac{n}{d} = \frac{\mathsf{TP}}{\mathsf{FN} + \mathsf{TP}};$$

if most of the true outliers are identified by the algorithm, $r \approx 1$;

F_1-**Score:** harmonic mean of the algorithm's precision and its recall on the testing set

$$F_1 = \frac{2pr}{p + r} = \frac{2\mathsf{TP}}{2\mathsf{TP} + \mathsf{FP} + \mathsf{FN}}.$$

Question: precision, recall, and F_1-score do not incorporate TN in the evaluation process. Is this likely to be a problem?

Example: consider a test dataset with 5000 observations, 100 of which are anomalous.

An algorithm that predicts all observations to be anomalous yields

		Predicted Class: Normal	Predicted Class: Anomaly	Total
Actual Class	Normal	0	4900	4900
	Anomaly	0	100	100
	Total	0	5000	**5000**

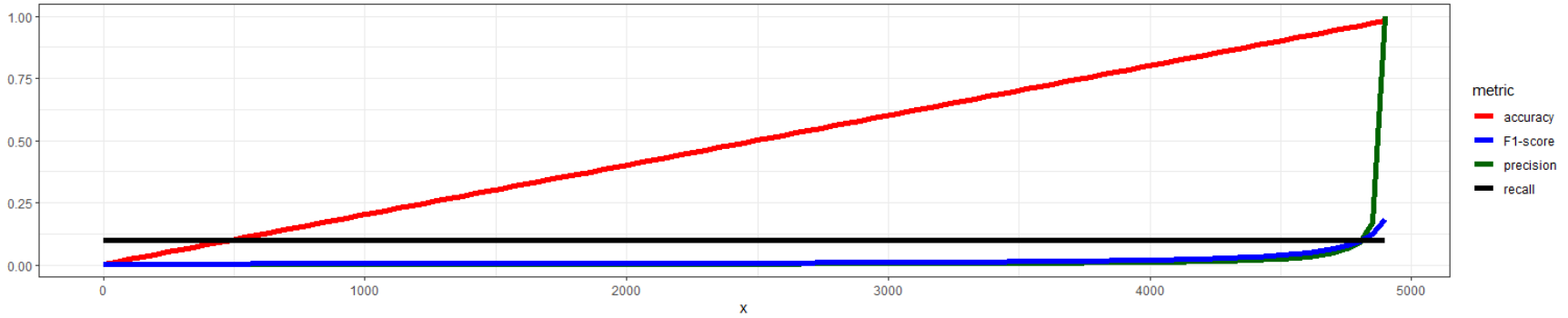
Accuracy	0.02
Precision	0.02
Recall	1.00
F1-Score	0.04

An algorithm that only detects 10 of the true outliers and x of the normal observations yields

		Predicted Class: Normal	Predicted Class: Anomaly	Total
Actual Class	Normal	x	$4900 - x$	4900
	Anomaly	90	10	100
	Total	$90 + x$	$4910 - x$	**5000**

$$0 \leq x \leq 4900$$

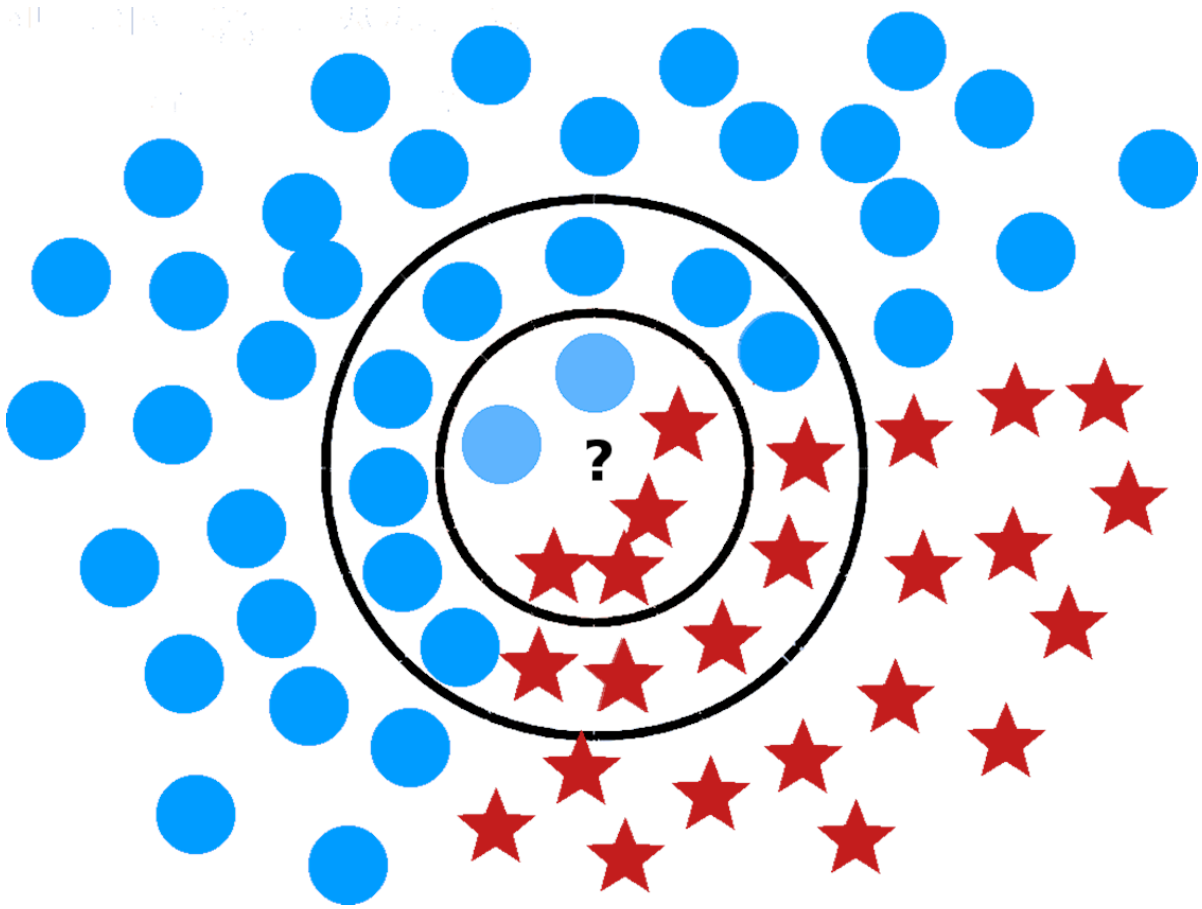
Accuracy	$(x + 10)/5000$
Precision	$10/(4910 - x)$
Recall	$1/10$
F1-Score	$20/(5010 - x)$



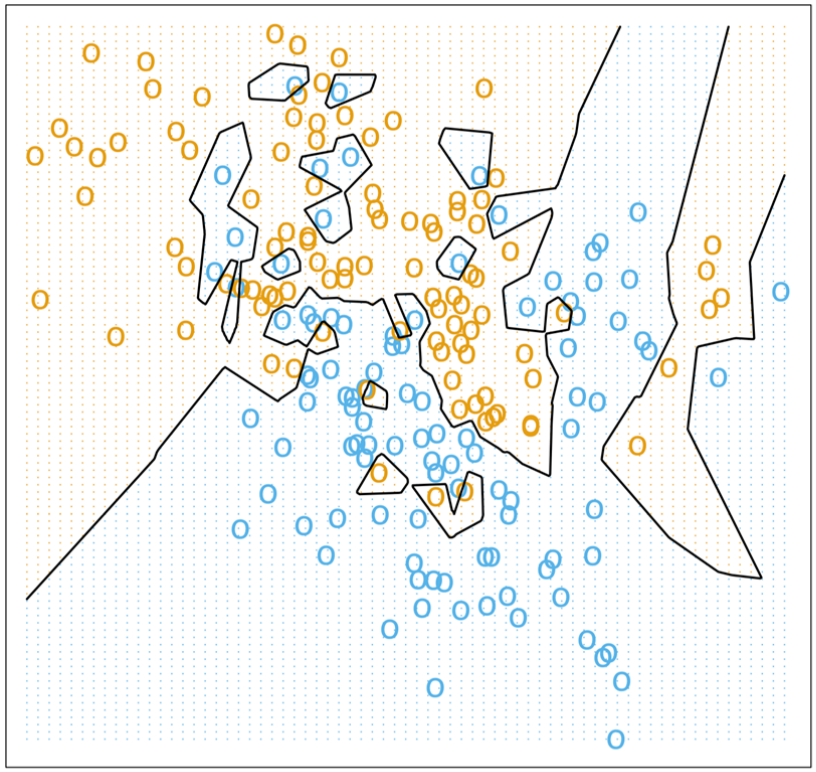
SL algorithms include:

- logistic regression
- naïve or optimal Bayes classifiers
- support vector machines
- neural networks (deep learning)
- decision trees, etc.

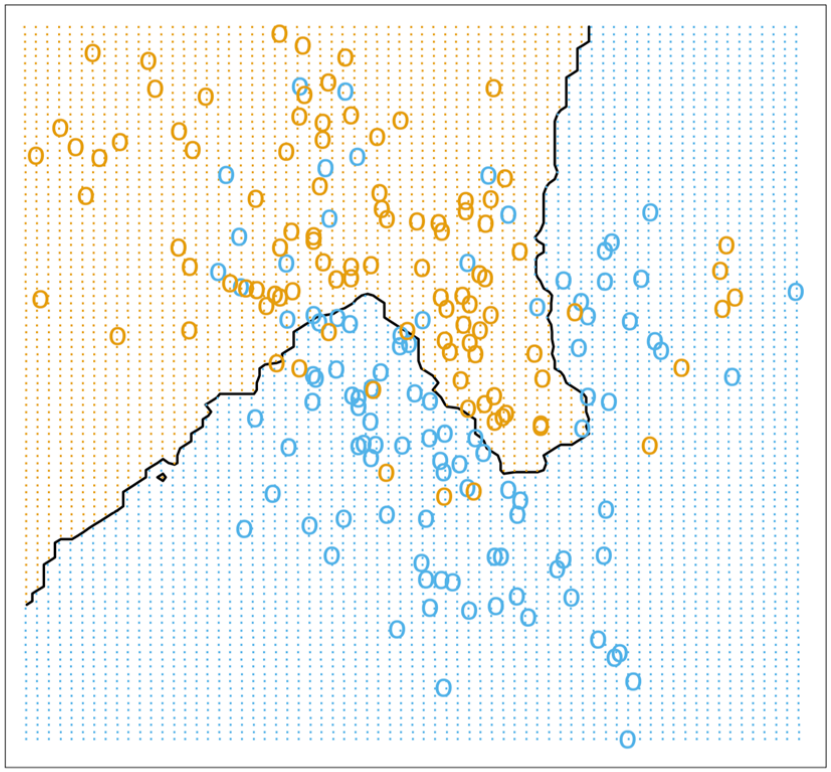
Such algorithms incorporate (and help us learn) historical patterns and rules.



k nearest neighbours

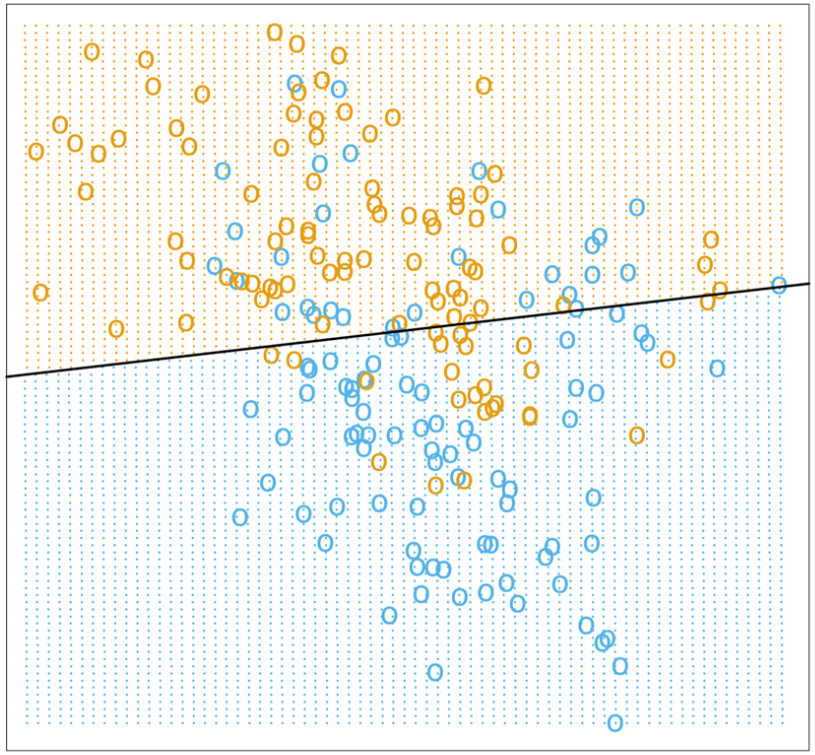


1NN Classifier

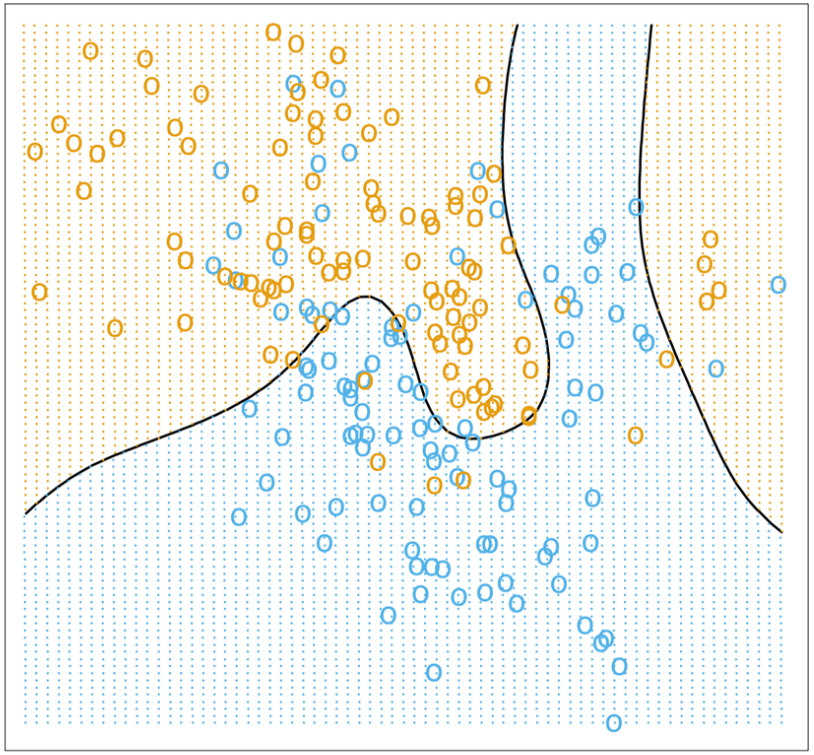


15NN Classifier

[Tibshirani, Hastie, Friedman]

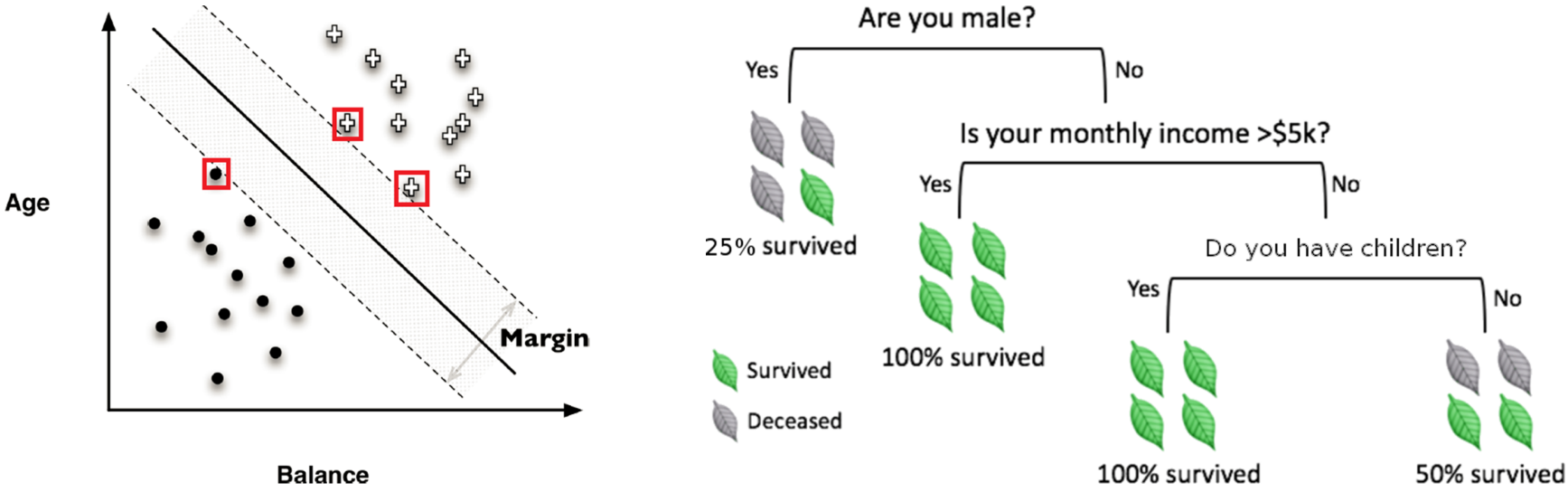


Linear Regression Classifier



Optimal Bayes Classifier

[Tibshirani, Hastie, Friedman]



Support vector machines (left), decision tree (right)
[Foster and Provost, Ng and Soo]

Another SL approach: estimate the **relative abnormality** of observations.

Estimating the probability that an observation $\mathbf{x}_1$ is anomalous $\Longrightarrow$ difficult;
Determining it is more likely to be anomalous than another $\mathbf{x}_2 \Longrightarrow$ easier.
(We write $\mathbf{x}_1 \succeq \mathbf{x}_2$.)

Let $k_i \in \{1, \ldots, m\}$ be the rank of the i^{th} **true outlier**, $i \in \{1, \ldots, n\}$, in the sorted list of suspicious observations

$$\mathbf{x}_1 \succeq \mathbf{x}_{k_1} \succeq \cdots \succeq \mathbf{x}_{k_i} \succeq \cdots \mathbf{x}_{k_n} \succeq \mathbf{x}_m, \quad n \leq m;$$

the **rank power** of the algorithm is

$$\mathsf{RP} = \frac{n(n+1)}{2\sum_{i=1}^{n} k_i}.$$

When the n actual anomalies are ranked in (or near) the top n suspicious observations, we have

$$\sum_{i=1}^{n} k_i \approx \sum_{i=1}^{n} i = \frac{n(n+1)}{2} \Longrightarrow \text{RP} \approx 1.$$

As with most performance evaluation metrics, a single raw number is meaningless – it needs to be compared to other algorithms.

Other SL performance evaluation metrics include:

- **AUC** – the probability of ranking a randomly chosen anomaly higher than a randomly chosen normal observation (higher is better);
- **probabilistic AUC** – a calibrated version of AUC.

The **rare occurrence** problem can be tackled by using:

- a **manipulated training set** (oversampling, undersampling, generating artificial instances);
- **specific SL AD algorithms** (CREDOS, PN, SHRINK);
- **boosting algorithms** (SMOTEBoost, RareBoost);
- **cost-sensitive classifiers** (MetaCost, AdaCost, CSB, SSTBoost),
- etc.

See A.Lazarević et al. [2004], *Data Mining for Analysis of Rare Events: A Case Study in Security, Financial and Medical Applications* for more information.

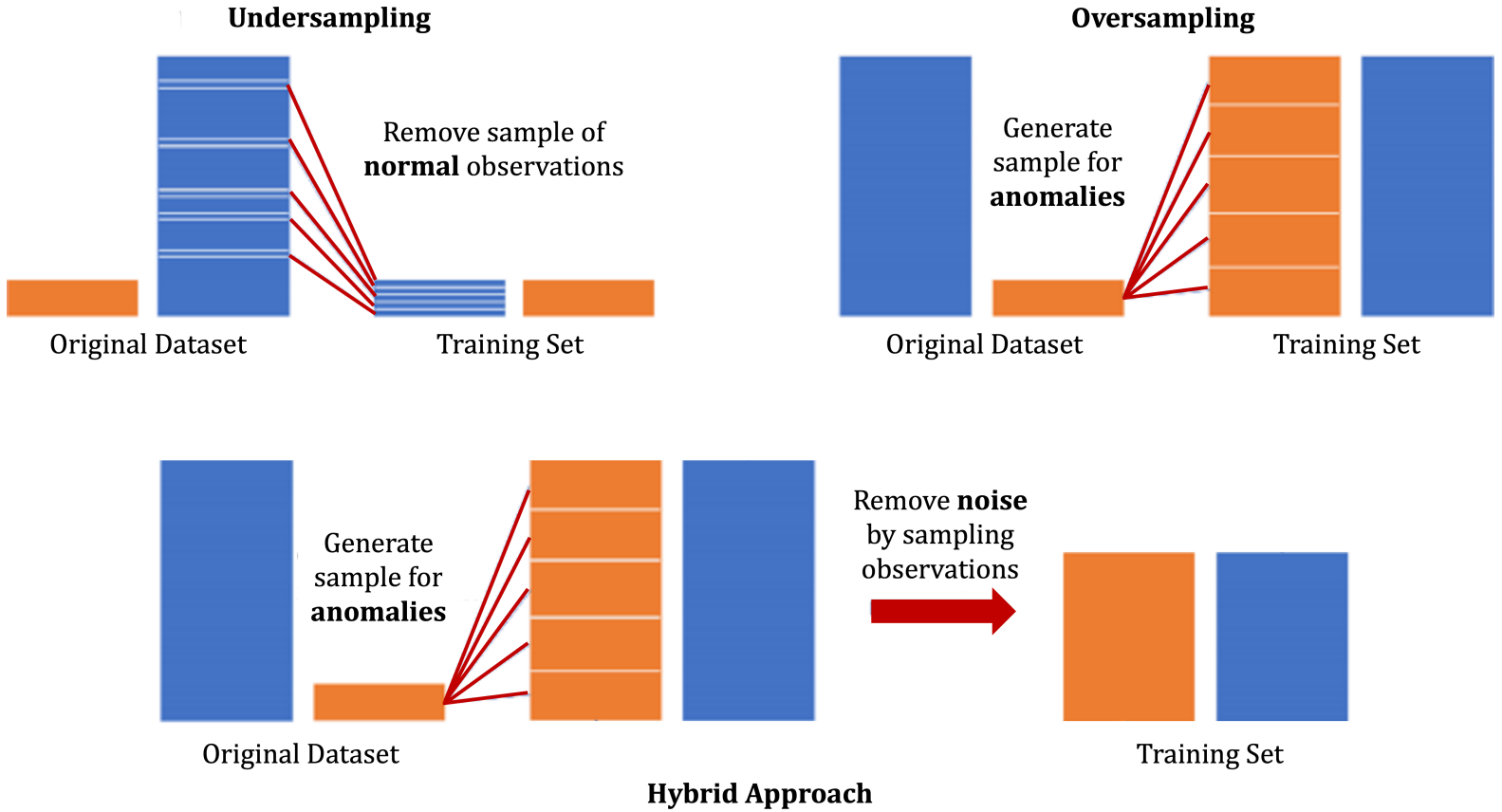
The rare (anomalous) class can be **oversampled** by duplicating the rare events until the data set is **balanced** (roughly the same number of anomalies and normal observations).

This does not increase the overall level of information, but it will increase the mis-classification cost.

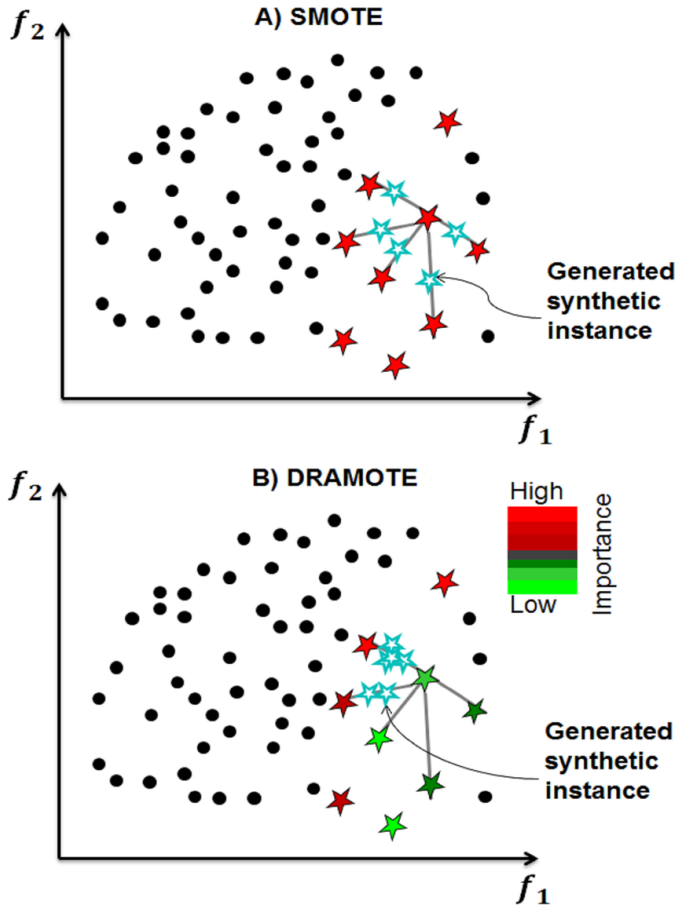
They majority class (normal observations) can also be **undersampled** by randomly removing:

- "near miss" observations or
- or observations far from anomalous observations.

Some loss of information has to be expected (and overly general rules).



[Le et al., A Hybrid Approach Using Oversampling Technique and Cost-Sensitive Learning for Bankruptcy Prediction]



Generating artificial cases with SMOTE and DRAMOTE [Soufan, et al.]

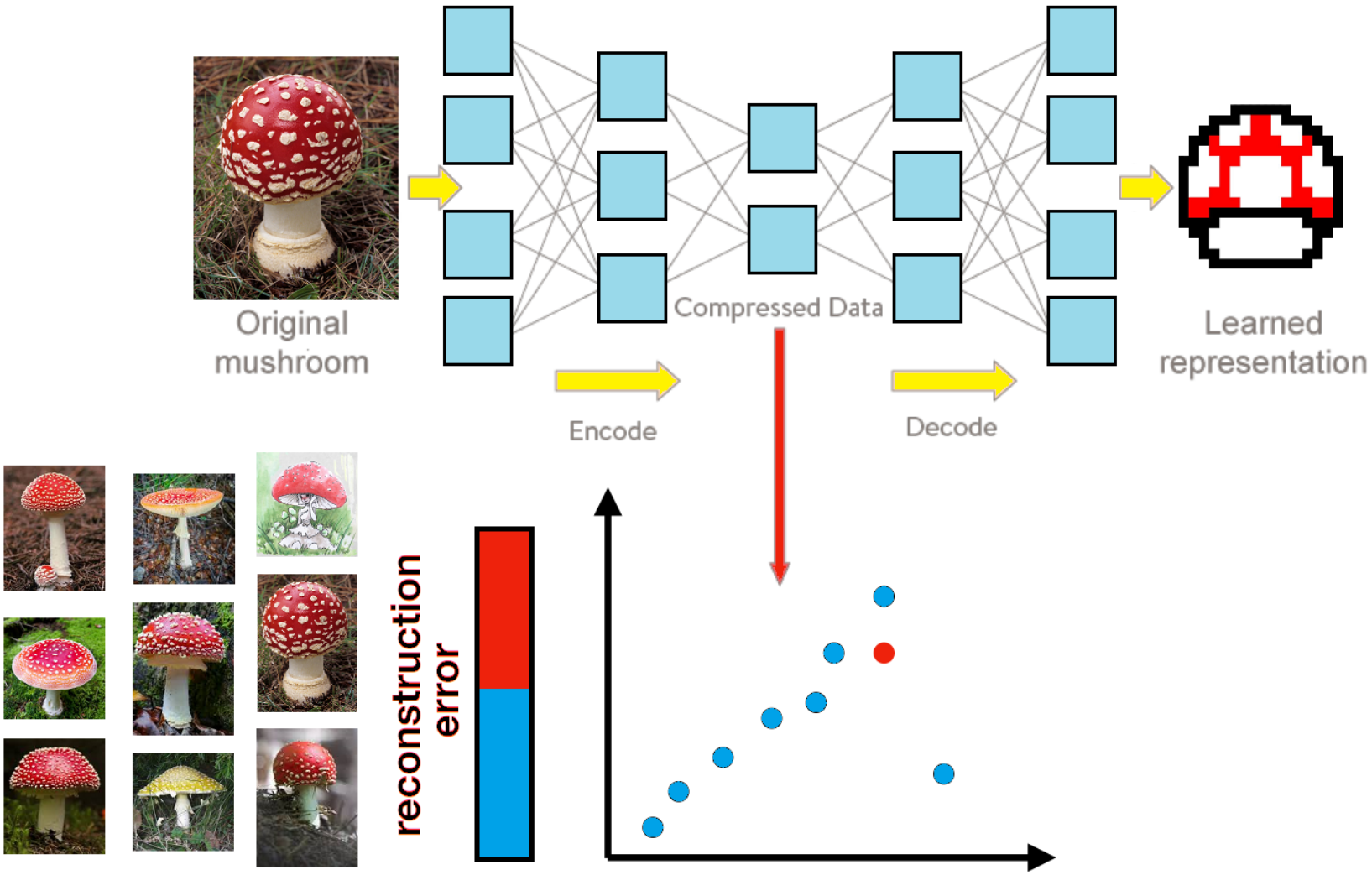
Autoencoders learn a compressed representation of the data (dimension reduction).

The **reconstruction error** measures (in a sense) how much information is lost in the compression.

Anomaly detection algorithms can be applied to the compressed data:

- look for anomalous patters, and/or
- anomalous reconstruction errors.

[Next slide adapted from Baron].

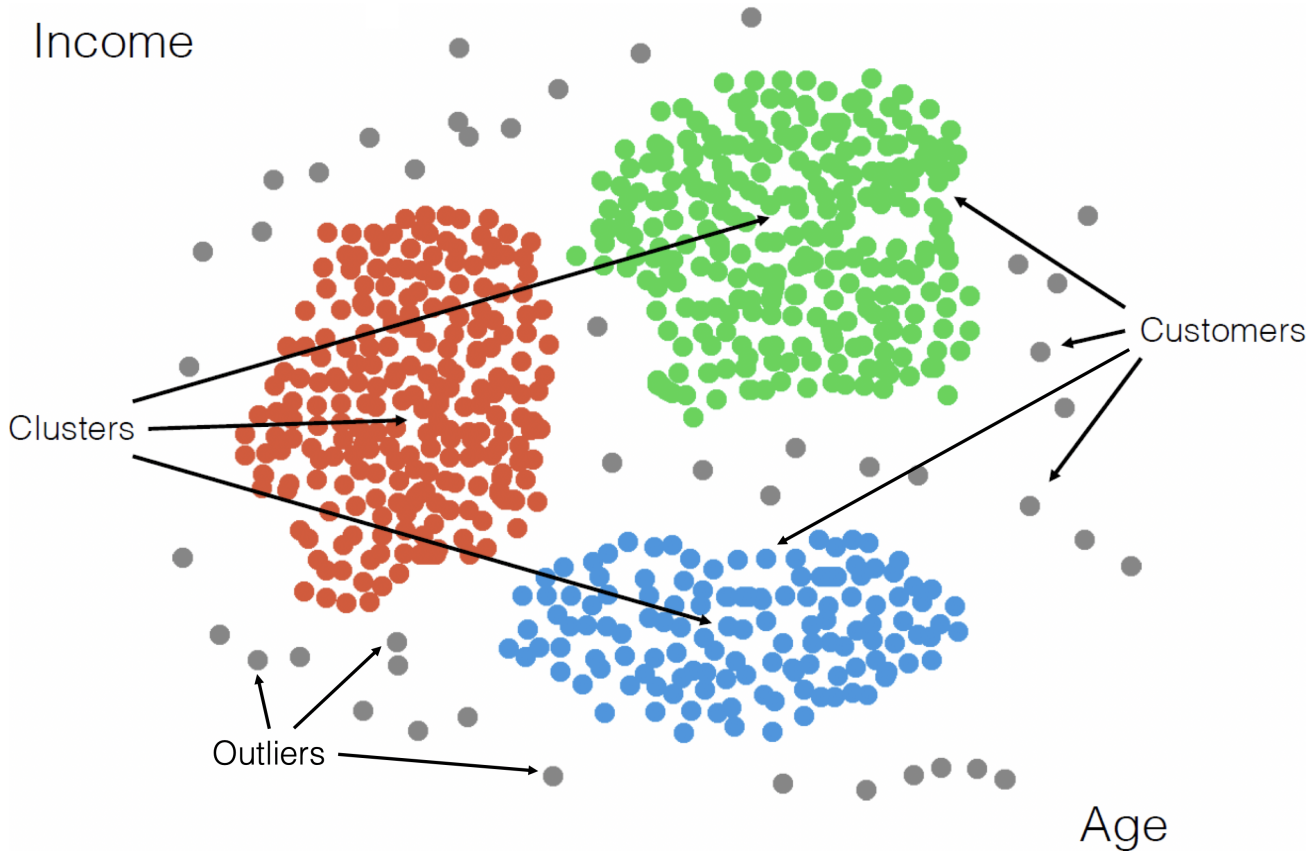

Original mushroom
Compressed Data
Learned representation
Encode
Decode
reconstruction error

On the **unsupervised** front, anomalous/normal labels are not used:

- anomalies are those observations that are dissimilar to other observations;
- **clusters** are groupings of similar observations, so
- observations without a natural cluster fit are potential anomalies.

Challenges:

- most clustering algorithms do not recognize potential outliers (DBSCAN and variants are exceptions), and
- finding an appropriate measure of similarity/dissimilarity of observations is difficult (different measures often lead to different cluster assignments).



Clusters of regular customers (red, green, blue) and potential anomalies/outliers (grey) in an artificial dataset.