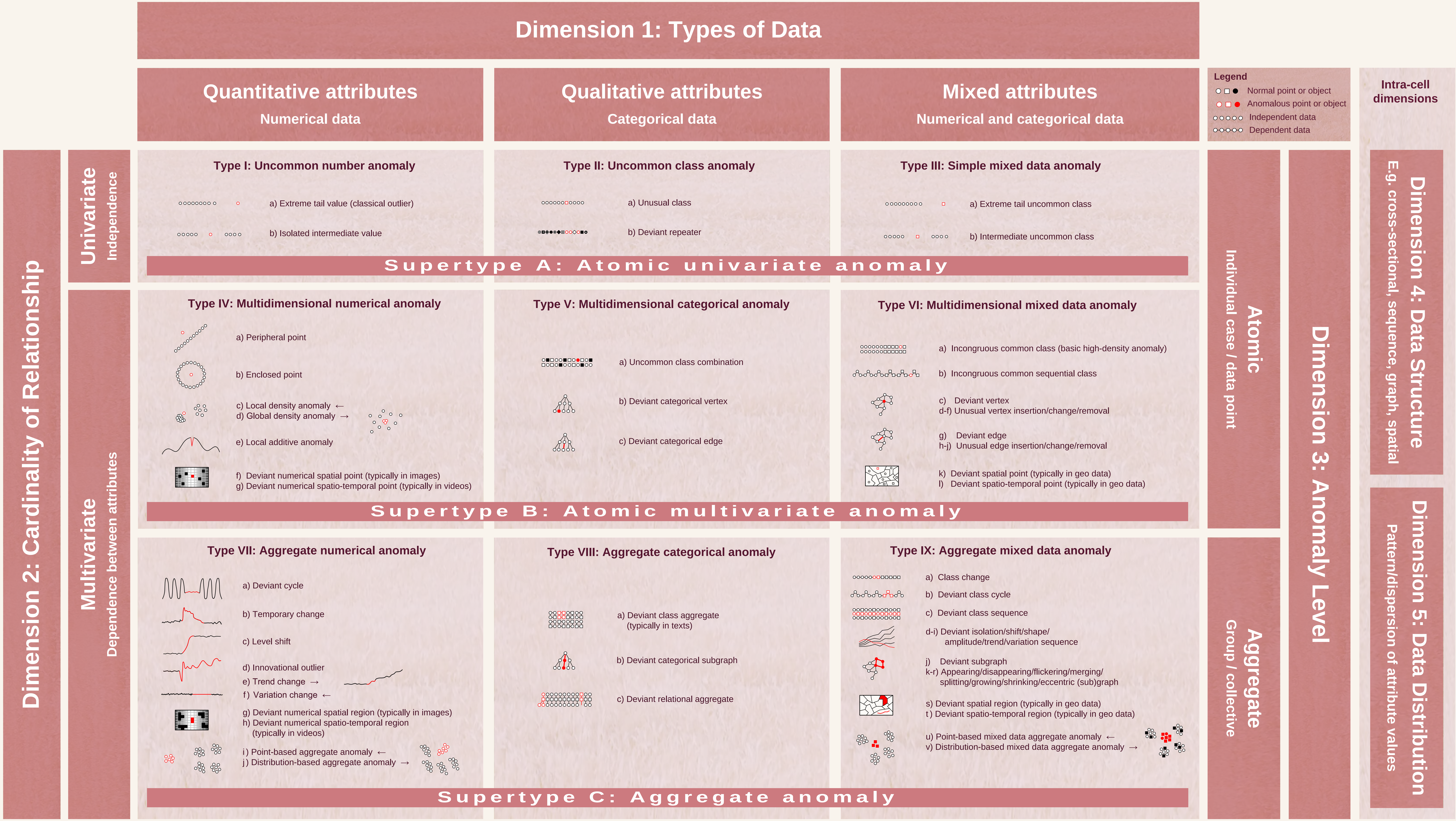
Typology of Outliers and Other Anomalies

A full overview of the kinds of anomalies that can be encountered in datasets. For a detailed explanation of all groups, types and subtypes, download the free open access publication: Foorthuis, R.M. (2021). *On the Nature and Types of Anomalies: A Review of Deviations in Data*. International Journal of Data Science and Analytics, Vol. 12, No. 4.



Dimension 1: Types of Data

Quantitative attributes
Numerical data

Qualitative attributes
Categorical data

Mixed attributes
Numerical and categorical data

Legend

- ○ □ ● Normal point or object
- ○ □ ● Anomalous point or object
- ○○○○○ Independent data
- ○-○-○-○-○ Dependent data

Intra-cell dimensions

Dimension 2: Cardinality of Relationship

Univariate
Independence

Multivariate
Dependence between attributes

Type I: Uncommon number anomaly
a) Extreme tail value (classical outlier)
b) Isolated intermediate value

Type II: Uncommon class anomaly
a) Unusual class
b) Deviant repeater

Type III: Simple mixed data anomaly
a) Extreme tail uncommon class
b) Intermediate uncommon class

Supertype A: Atomic univariate anomaly

Type IV: Multidimensional numerical anomaly
a) Peripheral point
b) Enclosed point
c) Local density anomaly ←
d) Global density anomaly →
e) Local additive anomaly
f) Deviant numerical spatial point (typically in images)
g) Deviant numerical spatio-temporal point (typically in videos)

Type V: Multidimensional categorical anomaly
a) Uncommon class combination
b) Deviant categorical vertex
c) Deviant categorical edge

Type VI: Multidimensional mixed data anomaly
a) Incongruous common class (basic high-density anomaly)
b) Incongruous common sequential class
c) Deviant vertex
d-f) Unusual vertex insertion/change/removal
g) Deviant edge
h-j) Unusual edge insertion/change/removal
k) Deviant spatial point (typically in geo data)
l) Deviant spatio-temporal point (typically in geo data)

Supertype B: Atomic multivariate anomaly

Type VII: Aggregate numerical anomaly
a) Deviant cycle
b) Temporary change
c) Level shift
d) Innovational outlier
e) Trend change →
f) Variation change ←
g) Deviant numerical spatial region (typically in images)
h) Deviant numerical spatio-temporal region (typically in videos)
i) Point-based aggregate anomaly ←
j) Distribution-based aggregate anomaly →

Type VIII: Aggregate categorical anomaly
a) Deviant class aggregate (typically in texts)
b) Deviant categorical subgraph
c) Deviant relational aggregate

Type IX: Aggregate mixed data anomaly
a) Class change
b) Deviant class cycle
c) Deviant class sequence
d-i) Deviant isolation/shift/shape/amplitude/trend/variation sequence
j) Deviant subgraph
k-r) Appearing/disappearing/flickering/merging/splitting/growing/shrinking/eccentric (sub)graph
s) Deviant spatial region (typically in geo data)
t) Deviant spatio-temporal region (typically in geo data)
u) Point-based mixed data aggregate anomaly ←
v) Distribution-based mixed data aggregate anomaly →

Supertype C: Aggregate anomaly

Atomic
Individual case / data point

Aggregate
Group / collective

Dimension 3: Anomaly Level

Dimension 4: Data Structure
E.g. cross-sectional, sequence, graph, spatial

Dimension 5: Data Distribution
Pattern/dispersion of attribute values

The typology of anomalies presents a full overview of the kinds of anomalies that can be encountered in datasets. Detecting anomalies can help with increasing the quality of your data, or with identifying risks, errors, opportunities and other interesting phenomena. The typology describes 3 broad groups, 9 basic types, and 63 subtypes of anomalies.

The typology's rows represent **3 supertypes** (broad groups) of anomalies

Each supertype features three basic types of anomalies, resulting in a total of **9 basic types**. Within these types the **63 concrete subtypes** specify the deviation that is explainable in terms of **5 fundamental data dimensions** (types of data, cardinality of relationship, anomaly level, data structure and data distribution)

Atomic univariate anomalies are single cases with a deviant value for an individual attribute (possibly multiple attributes, but each individual value is deviant in its own right). They are relatively easy to describe and detect because the individual values are unusual, and relationships between attributes or cases are not relevant.

I. Uncommon number anomaly: This is a case with an extremely high, low or otherwise unusual value for a single quantitative attribute (or multiple, but these are then anomalies in their own right). These deviant numbers often manifest themselves as an *extreme tail value* (ST-Ia), i.e. an unusually high or low number. An example is an unrealistically high numerical value, such as a person reported to be 269 cm high. However, it can also be located in the middle of the value range, thus an *isolated intermediate value* (ST-Ib) in between the majority of the data.

II. Uncommon class anomaly: This can manifest itself as an *unusual class* (ST-IIa), i.e. a case with a unique or rare categorical value for a single qualitative variable (or for multiple, but these are then anomalies in their own right). For example, the label 'monkey' will be unusual if all other labels represent insects and birds. Alternatively, it can be a *deviant repeater* (ST-IIb), i.e. a frequently occuring value in a set in which most values are unique, such in a collection of identifying codes.

III. Simple mixed data anomaly: This is a case that is both a Type I and a Type II anomaly, i.e. with at least one isolated numerical value and one uncommon class. An example is a data point with a unique class label that lies at an isolated intermediate location in the numerical space, a so-called *intermediate uncommon class* (ST-IIIb). However, like Type I and II anomalies, analyzing the attributes jointly is not necessary because the case in question is not multivariately anomalous. This type involves a set of individually deviant attribute values.

Atomic multivariate anomalies are single cases whose deviant nature lies in their relations, with the individual values not being anomalous. In independent data this will manifest itself in the unusual combination of a case's own attribute values, such as a 10 year old person with a body length of 180 cm. However, the multivariate nature also allows defining and detecting deviations in dependent data, i.e. in the relation with the other cases to which the given case is linked. An example is a time series temperature measurement that is unusually high for winter, but that would have been normal in summer.

IV. Multidimensional numerical anomaly: This is a case that does not fit the general patterns when the relationship between multiple quantitative attributes is taken into account, without showing unusual values for any of the individual attributes that partake in this relationship. The so-called *peripheral point* (ST-IVa) is an isolated datapoint that lies outside the dense multivariate cluster. An example is a person who is 182 cm tall and weighs 53 kilos, i.e. an unusual combination of normal individual values. A Type IV anomaly can also reside in dependent data such as time series. For example, the *local additive anomaly* (ST-IVe) is a short-lived spike that deviates from the local temporal neighborhood – e.g. the current season or trend – without exhibiting globally extreme values. Other Type IV anomalies can lie in spatial data.

V. Multidimensional categorical anomaly: This is a case that does not fit the general patterns when the relationship between multiple qualitative attributes is taken into account, without showing unusual values for any of the individual attributes that partake in this relationship. In short, a case with a rare or unique combination of class values. A subtype in independent data is the *uncommon class combination* (ST-Va), such as this curious combination of values from three attributes used to describe dogs: 'MALE', 'PUPPY' and 'PREGNANT'. A subtype in dependent data is the *deviant categorical vertex* (ST-Vb), which is an uncommon node in a tree or other form of graph data structure. An example is a node with a label that is different from the labels of the nodes it is connected to.

VI. Multidimensional mixed data anomaly: This is a case that does not fit the general patterns when the relationship between multiple quantitative and qualitative attributes is taken into account, without being an atomic univariate anomaly with regard to any of the individual attributes that partake in this relation. It concerns a case with an unusual combination of qualitative and quantitative attributes. In independent data this is typically the *incongruous common class* (ST-VIa), which has a class value (or combination thereof) that in itself is not rare in the overall dataset, but is only uncommon in its own neighborhood. Such cases thus seem to be mislabeled or misplaced. This anomaly is also the most basic form of a *high-density anomaly*. In dependent data a Type VI anomaly can manifest itself in many other ways. An *incongruous common sequential class* (ST-VIb), for example, is an individual deviant in a sequence of class values that exhibit a sequential pattern in which the anomaly does not fit. Other subtypes may lie in complex graph data structures or spatial datasets.

Aggregate anomalies are groups of cases that deviate as a collective, of which the constituent cases usually are not individually anomalous. Relationships between attributes and between cases not only position an occurrence in the set with dependent data, but also form a pre-defined or derived group. Owing to their complex and intricate nature, these anomalies are generally the most difficult to describe and detect. A deviant subsequence is one manifestation of an aggregate anomaly, such as a whole winter with many unusually high temperatures compared to other winters.

VII. Aggregate numerical anomaly: This is a group of related cases that deviate as a collective with regard to their quantitative attributes. Such anomaly subtypes are often found in time series data, in which they constitute a subsequence of the entire sequence. For example, the *deviant cycle* (ST-VIIa) occurs when the respective cycle (e.g. a season) follows a different pattern than the other cycles. The *temporary change* (ST-VIIb) is a rise or fall of the substantive value that requires a certain period to get back to the regular level. The *level shift* (ST-VIIc) is an abrupt structural change to a higher or lower value level. Type VII anomalies can also be found in spatial data. Independent datasets can host them as well, such as the *distribution-based aggregate anomaly* (ST-VIIj), which is a whole cluster with unusual characteristics at the collective level, such as a group of data points with a different orientation, (co)variance, mean or frequency.

VIII. Aggregate categorical anomaly: This is a group of related cases that deviate as a collective with regard to their qualitative attributes. A *deviant class aggregate* (ST-VIIIa) often manifests itself as a deviant text paragraph, section or document. The aggregate is for example a bag-of-words or a collection of sentences that stands out in terms of topic, style or tone. A *deviant relational aggregate* (ST-VIIIc) is a collective anomaly as a result of its relationships. Such complex aggregates are usually comprised of several domain-specific entities that are inter-related by one-to-many relationships, and are typically stored in different tables in a relational database.

IX. Aggregate mixed data anomaly: This is a group of related cases that deviate as a collective with regard to both numerical and categorical variables. Owing to its mixed data and potentially intricate relationships this type allows for a wide variety of complex anomaly subtypes. A *deviant class cycle* (ST-IXb) is an entire categorical subsequence that is anomalous, because it does not adhere to the overall cycle pattern. The *deviant shape sequence* (ST-IXf) is time series in a collection of multiple time series, with the anomaly located at the same position in the time space, but featuring an unusual form. There are many other Type IX anomalies, such as the *deviant subgraph* (ST-IXj), e.g. a group of linked vertices with significantly different substantive values than those observed in other parts of the graph. The *deviant spatio-temporal region* (ST-IXt) is a polygon or other aggregated object with substantive values that deviate when both its temporal pattern and spatial area are taken into account. An example is a region where the risk of suffering from a global disease outbreak is increasing significantly faster than elsewhere.

See the referenced publication for more details and examples, as well as for notes on how to use the typology for explainable data science and algorithm & software testing. See the Taming the Anomaly video for animated visual explanations.