# Chapter DM:I (continued)

# On Data [Tan et al. 2005]

- An object $o \in O$ is described by a set of attributes.
  An object is also known as record, point, case, sample, entity, or instance.
- An attribute $A$ is a property of an object.
  An attribute is also known as variable, field, characteristic, or feature.
- A measurement scale is a system (often a convention) of assigning a numerical or symbolic value to an attribute of an object.

Attributes

Objects

| ID | Check | Status | Income | Risk |
|---|---|---|---|---|
| 1 | + | single | 125 000 | No |
| 2 | - | married | 100 000 | No |
| 3 | - | single | 70 000 | No |
| 4 | + | married | 120 000 | No |
| 5 | - | divorced | 95 000 | Yes |
| 6 | - | married | 60 000 | No |
| 7 | + | divorced | 220 000 | No |
| 8 | - | single | 85 000 | Yes |
| 9 | - | married | 75 000 | No |
| 10 | - | single | 90 000 | Yes |

# On Data [Tan et al. 2005]

- An object $o \in O$ is described by a set of attributes.
  An object is also known as record, point, case, sample, entity, or instance.
- An attribute $A$ is a property of an object.
  An attribute is also known as variable, field, characteristic, or feature.
- A measurement scale is a system (often a convention) of assigning a numerical or symbolic value to an attribute of an object.

Attributes

Objects

| ID | Check | Status | Income | Risk |
|---|---|---|---|---|
| 1 | + | single | 125 000 | No |
| 2 | - | married | 100 000 | No |
| 3 | - | single | 70 000 | No |
| 4 | + | married | 120 000 | No |
| 5 | - | divorced | 95 000 | Yes |
| 6 | - | married | 60 000 | No |
| 7 | + | divorced | 220 000 | No |
| 8 | - | single | 85 000 | Yes |
| 9 | - | married | 75 000 | No |
| 10 | - | single | 90 000 | Yes |

# On Data [Tan et al. 2005]

- An object $o \in O$ is described by a set of attributes.
  An object is also known as record, point, case, sample, entity, or instance.
- An attribute $A$ is a property of an object.
  An attribute is also known as variable, field, characteristic, or feature.
- A measurement scale is a system (often a convention) of assigning a numerical or symbolic value to an attribute of an object.

Attributes

Objects

| ID | Check | Status | Income | Risk |
|---|---|---|---|---|
| 1 | + | single | 125 000 | No |
| 2 | - | married | 100 000 | No |
| 3 | - | single | 70 000 | No |
| 4 | + | married | 120 000 | No |
| 5 | - | divorced | 95 000 | Yes |
| 6 | - | married | 60 000 | No |
| 7 | + | divorced | 220 000 | No |
| 8 | - | single | 85 000 | Yes |
| 9 | - | married | 75 000 | No |
| 10 | - | single | 90 000 | Yes |

# On Data [Tan et al. 2005]

- Attribute values may vary from one object to another or one time to another.
- The same attribute can be mapped to different attribute values.
  Example: height can be measured in feet or meters.
- Different attributes can be mapped to the same set of values.
  Example: attribute values for person ID and age are integers.

# On Data [Tan et al. 2005]

- Attribute values may vary from one object to another or one time to another.
- The same attribute can be mapped to different attribute values.
  Example: height can be measured in feet or meters.
- Different attributes can be mapped to the same set of values.
  Example: attribute values for person ID and age are integers.

The way an attribute is measured may not match the attribute's properties:

# On Data [Tan et al. 2005]

## Types of Attributes

| Type | | Comparison | Statistics | Examples |
|---|---|---|---|---|
| *categorical (qualitative)* | nominal | values are names, only information to distinguish objects<br>$=$ $\neq$ | mode, entropy, contingency, correlation, $\chi^2$ test | zip codes, employee IDs, eye color, gender: {male, female} |

# On Data [Tan et al. 2005]

## Types of Attributes

| Type | | Comparison | Statistics | Examples |
|---|---|---|---|---|
| *categorical (qualitative)* | nominal | values are names, only information to distinguish objects<br>$=$ $\neq$ | mode, entropy, contingency, correlation, $\chi^2$ test | zip codes, employee IDs, eye color, gender: {male, female} |
| | ordinal | enough information to order objects<br>$<$ $>$ $\leq$ $\geq$ | median, percentiles, rank correlation, run tests, sign tests | hardness of minerals, grades, street numbers, quality: {good, better, best} |

# On Data [Tan et al. 2005]

## Types of Attributes

| Type | | Comparison | Statistics | Examples |
|---|---|---|---|---|
| *categorical (qualitative)* | nominal | values are names, only information to distinguish objects<br>$=$ $\neq$ | mode, entropy, contingency, correlation, $\chi^2$ test | zip codes, employee IDs, eye color, gender: {male, female} |
| | ordinal | enough information to order objects<br>$<$ $>$ $\leq$ $\geq$ | median, percentiles, rank correlation, run tests, sign tests | hardness of minerals, grades, street numbers, quality: {good, better, best} |
| *numeric (quantitative)* | interval | differences are meaningful, a unit of measurement exists<br>$+$ $-$ | mean, standard deviation, Pearson's correlation, $t$-test, $F$-test | calendar dates, temperature in Celsius, temperature in Fahrenheit |

# On Data [Tan et al. 2005]

## Types of Attributes

| Type | | Comparison | Statistics | Examples |
|---|---|---|---|---|
| *categorical (qualitative)* | nominal | values are names, only information to distinguish objects<br>$=$ $\neq$ | mode, entropy, contingency, correlation, $\chi^2$ test | zip codes, employee IDs, eye color, gender: {male, female} |
| | ordinal | enough information to order objects<br>$<$ $>$ $\leq$ $\geq$ | median, percentiles, rank correlation, run tests, sign tests | hardness of minerals, grades, street numbers, quality: {good, better, best} |
| *numeric (quantitative)* | interval | differences are meaningful, a unit of measurement exists<br>$+$ $-$ | mean, standard deviation, Pearson's correlation, $t$-test, $F$-test | calendar dates, temperature in Celsius, temperature in Fahrenheit |
| | ratio | differences and ratios are meaningful<br>$*$ $/$ | geometric mean, harmonic mean, percent variation | temperature in Kelvin, monetary quantities, counts, age, length, electrical current |

# On Data [Tan et al. 2005]

## Types of Attributes

| Type | | Permissible transformation | Comment |
|---|---|---|---|
| *categorical (qualitative)* | nominal | any one-to-one mapping, permutation of values | A reassignment of employee ID numbers will not make any difference. |

# On Data [Tan et al. 2005]

## Types of Attributes

| Type | | Permissible transformation | Comment |
|---|---|---|---|
| *categorical (qualitative)* | nominal | any one-to-one mapping, permutation of values | A reassignment of employee ID numbers will not make any difference. |
| | ordinal | any order-preserving change of values: $x \mapsto f(x)$, where $f$ is a monotonic | An attribute encompassing the notion of {good, better, best} can be represented equally well by the values $\{1, 2, 3\}$. |

# On Data [Tan et al. 2005]

## Types of Attributes

| Type | | Permissible transformation | Comment |
|---|---|---|---|
| *categorical (qualitative)* | nominal | any one-to-one mapping, permutation of values | A reassignment of employee ID numbers will not make any difference. |
| | ordinal | any order-preserving change of values: $x \mapsto f(x)$, where $f$ is a monotonic | An attribute encompassing the notion of {good, better, best} can be represented equally well by the values $\{1, 2, 3\}$. |
| *numeric (quantitative)* | interval | $x \mapsto a \cdot x + b$, where $a$ and $b$ are constants | The Fahrenheit and Celsius temperature scales differ in terms of where their zero value is, as well as the size of a unit (degree). |

# On Data [Tan et al. 2005]

## Types of Attributes

| Type | | Permissible transformation | Comment |
|---|---|---|---|
| *categorical (qualitative)* | nominal | any one-to-one mapping, permutation of values | A reassignment of employee ID numbers will not make any difference. |
| | ordinal | any order-preserving change of values: $x \mapsto f(x)$, where $f$ is a monotonic | An attribute encompassing the notion of {good, better, best} can be represented equally well by the values $\{1, 2, 3\}$. |
| *numeric (quantitative)* | interval | $x \mapsto a \cdot x + b$, where $a$ and $b$ are constants | The Fahrenheit and Celsius temperature scales differ in terms of where their zero value is, as well as the size of a unit (degree). |
| | ratio | $x \mapsto a \cdot x$, where $a$ is a constant | Length can be measured in meters or feet. |

Remarks:

- Identifying, considering, and measuring an attribute $A$ of an object $o \in O$ is the heart of model formation and always goes along with a sort of abstraction. Formally, this abstraction is operationalized by a model formation function $\alpha : O \to \mathbf{X}$, where $\mathbf{X}$ is called the feature space. [ML Introduction]
- The terms “attribute” and “feature” can be used synonymously. However, a slight distinction is the following: attributes are often associated with objects, $O$, while features usually designate the dimensions of the feature space $\mathbf{X}$.
- The type of an attribute is also referred to as the type of a *measurement scale* or *level of measurement*.
- We call a transformation of an attribute *permissible* if its meaning is unchanged after the transformation.
- Distinguish between *discrete* attributes and *continuous* attributes. The former can only take a finite or countably infinite set of values, the latter can be measured in infinitely small units. Be careful when deriving from this distinction an attribute’s type.
- We will encode attributes of interval type or ratio type by real numbers. Note that attributes of nominal type and ordinal type can also be encoded by real numbers.
- Particular learning methods require particular attribute types.

# On Data [Tan et al. 2005]

## Types of Data Sets

Data sets may not be a homogeneous collection of objects but come along with differently intricate characteristics:

1. Inhomogeneity of *attributes*:

2. Inhomogeneity of *objects*:

3. Inhomogeneity of *distributions*:

4. Resolution:

5. Curse of dimensionality:

# On Data [Tan et al. 2005]

## Types of Data Sets

Data sets may not be a homogeneous collection of objects but come along with differently intricate characteristics:

1. Inhomogeneity of *attributes*:
   Consider the combination of different attribute types within a single object.

2. Inhomogeneity of *objects*:
   Consider the combination of different objects in a single data set.

3. Inhomogeneity of *distributions*:
   The correlation between attributes varies in the sample space.

4. Resolution:
   The attributes may be given at different resolutions.

5. Curse of dimensionality:
   Attribute number and object density stand in exponential relation.

# On Data [Tan et al. 2005]

## Types of Data Sets: Record Data

Collection of records, each of which consists of a fixed set of attributes:

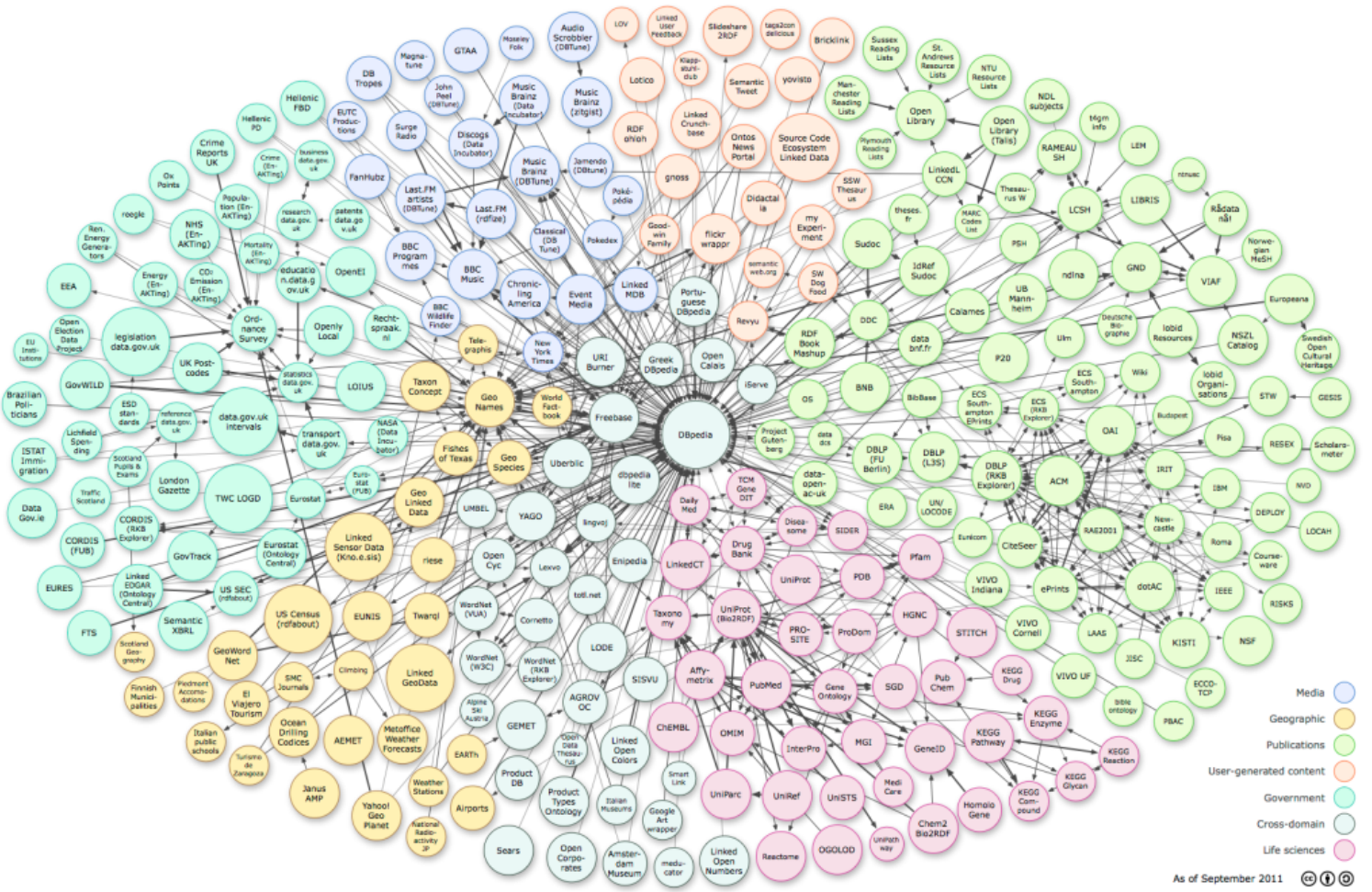
| ID | Check | Status | Income | Risk |
|---|---|---|---|---|
| 1 | + | single | 125 000 | No |
| 2 | - | married | 100 000 | No |
| 3 | - | single | 70 000 | No |
| 4 | + | married | 120 000 | No |
| 5 | - | divorced | 95 000 | Yes |
| 6 | - | married | 60 000 | No |
| 7 | + | divorced | 220 000 | No |
| 8 | - | single | 85 000 | Yes |
| 9 | - | married | 75 000 | No |
| 10 | - | single | 90 000 | Yes |

- If all elements in a data set have the same fixed set of numeric attributes, they can be thought of as points in a multi-dimensional space.
- Such data can be represented by a matrix, where each row stores an object and each column stores an attribute.
  Example: term-document matrices in information retrieval.

# On Data [Tan et al. 2005]

## Types of Data Sets: Graph Data

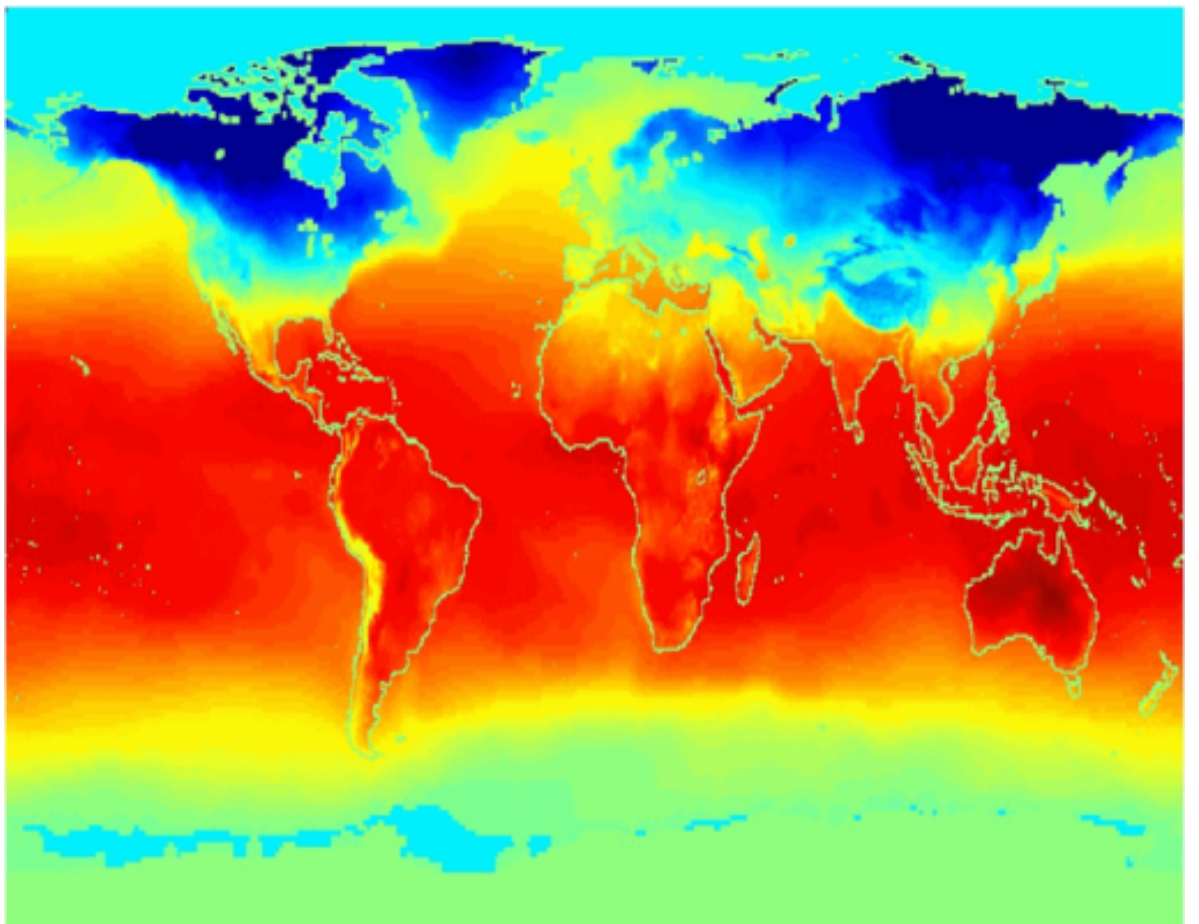
Graph of the Linked Open Data cloud [lod-cloud.net] :



# On Data [Tan et al. 2005]

## Types of Data Sets: Ordered Data

Average monthly temperature of land and ocean (= spatio-temporal data) :



# On Data [Tan et al. 2005]

## Data Quality

When repeating measurements of a quantity, measurement errors and data collection errors may occur during the measurement process. Questions:

1. What kinds of data quality problems exist?
2. How to detect data quality problems?
3. How to address data quality problems?

# On Data [Tan et al. 2005]

## Data Quality

When repeating measurements of a quantity, measurement errors and data collection errors may occur during the measurement process. Questions:

1. What kinds of data quality problems exist?
2. How to detect data quality problems?
3. How to address data quality problems?

**Definition 1 (Precision, Bias, Accuracy)**

Given a set of repeated measurements of the same quantity. Then, the closeness of the measurements to one another is called *precision*, a possible systematic variation is called *bias*, and the closeness to the true value is called *accuracy*.

# On Data [Tan et al. 2005]

## Data Quality

When repeating measurements of a quantity, measurement errors and data collection errors may occur during the measurement process. Questions:

1. What kinds of data quality problems exist?
2. How to detect data quality problems?
3. How to address data quality problems?

**Definition 1 (Precision, Bias, Accuracy)**

Given a set of repeated measurements of the same quantity. Then, the closeness of the measurements to one another is called *precision*, a possible systematic variation is called *bias*, and the closeness to the true value is called *accuracy*.

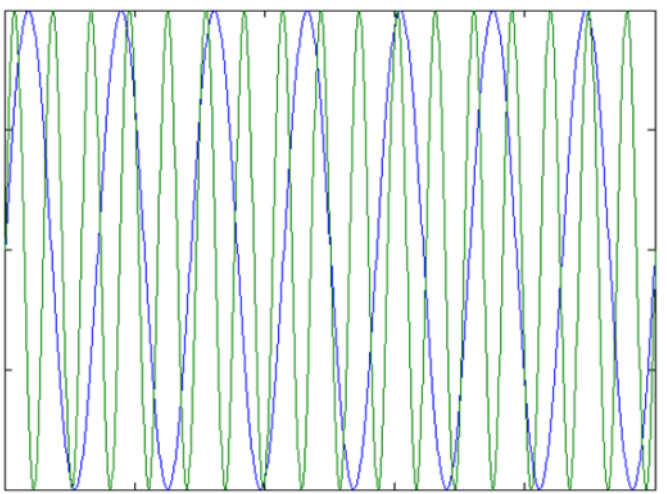
Examples for data quality problems:

- noise, artifacts, outliers
- missing values
- duplicate data

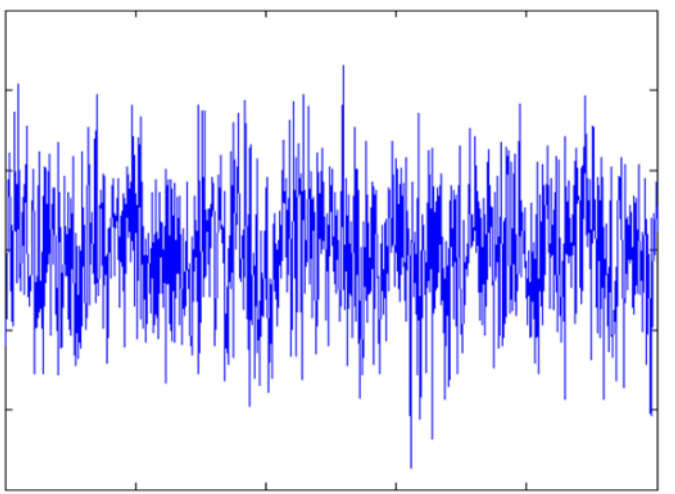
# On Data [Tan et al. 2005]

## Data Quality: Noise

Noise refers to random modifications of attributes that often have a spatial or temporal characteristics:



sine waves



sine waves with noise

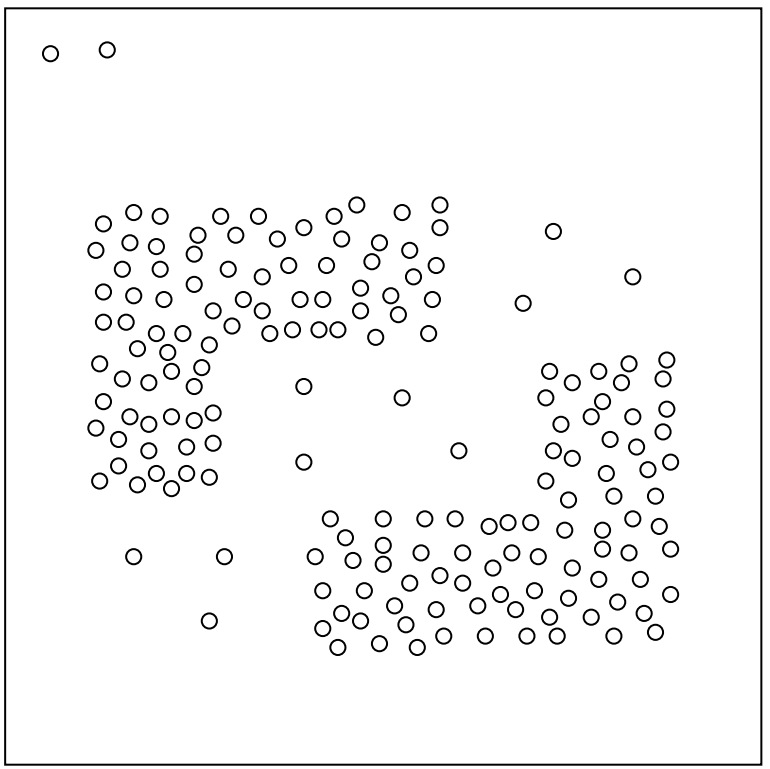
Noise represents the intrinsic variability of data. [Bishop 2006, p.47]

Artifacts refer to deterministic distortions of a measurement process.

# On Data [Tan et al. 2005]

## Data Quality: Outliers

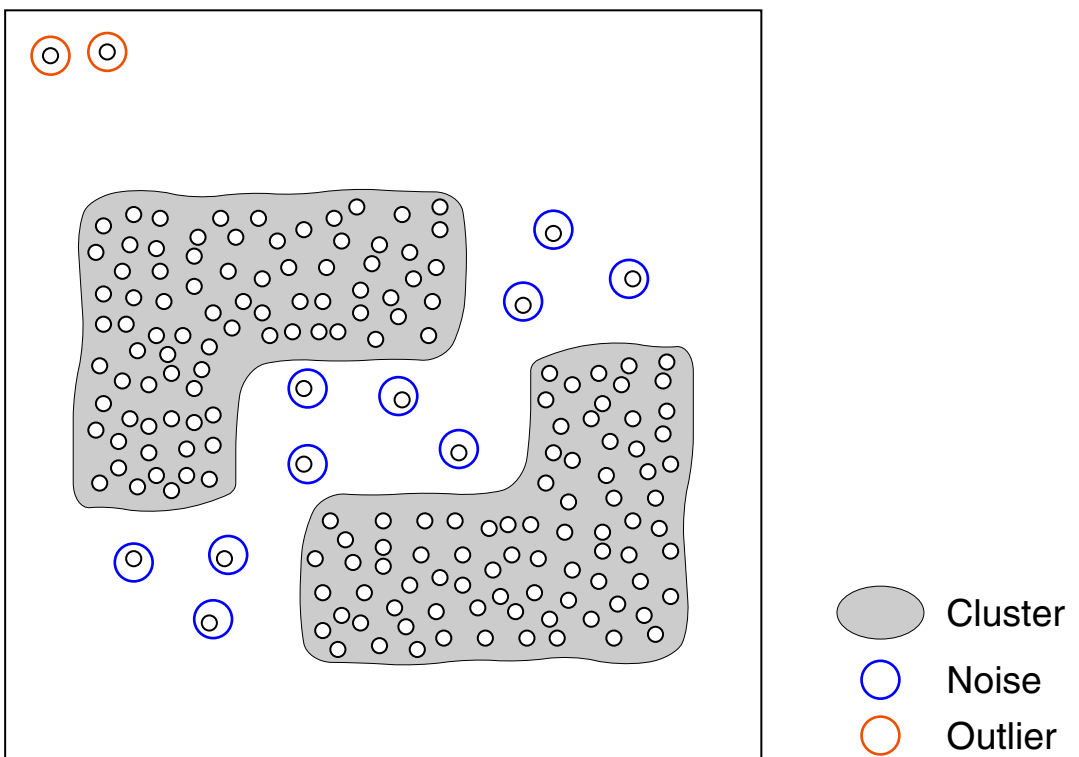
Outliers are members in the data set with characteristics that are considerably different than most of the other elements:



# On Data [Tan et al. 2005]

## Data Quality: Outliers

Outliers are members in the data set with characteristics that are considerably different than most of the other elements:



# On Data [Tan et al. 2005]

## Data Quality: Missing Values

Main reasons for missing values:

1. Information is not collected.
   Example: people decline to give their age or weight.
2. Attributes may not be applicable to all elements in $O$.
   Example: annual income is not applicable to children.
3. Information is not trustworthy.
   Example: profile data on Facebook is intentionally misleading.

Strategies for handling missing values:

- eliminate members of the data
- estimate missing values
- ignore the missing value during analysis
- replace with all possible values weighted by their probabilities

# On Data [Tan et al. 2005]

## Data Preprocessing

- sampling of object set $O$
- modeling of objects, $\alpha : O \to \mathbf{X}$
- sampling of the feature space $\mathbf{X}$ [Evaluating Effectiveness]
- selection of attributes (features) [attributes versus features]
- transformation of attributes (features)
- discretization and binarization of attributes (features)
- dimensionality reduction of the feature space $\mathbf{X}$