Knowledge Discovery and Data Mining

Class 2 Data Collecting, Data Sampling and Missing Values

Xuan Song
songx@sustech.edu.cn

1.Data and Data Types

2.Data Quality

3.Data Preprocessing

4.Similarity and Dissimilarity

What is Data?

What is Data?

- Collection of **data objects** and their **attributes**

Attributes

Objects

id	height	weight	score	type
1	157	61	28.7	A
2	155	50	7.0	A
3	155	63	17.8	B
4	154	44	15.1	B
5	153	70	15.1	B
6	151	60	13.6	B
7	158	38	29.6	C
8	152	44	8.1	D
9	149	57	2.1	D
10	153	41	27.0	D

This is data.

Attributes

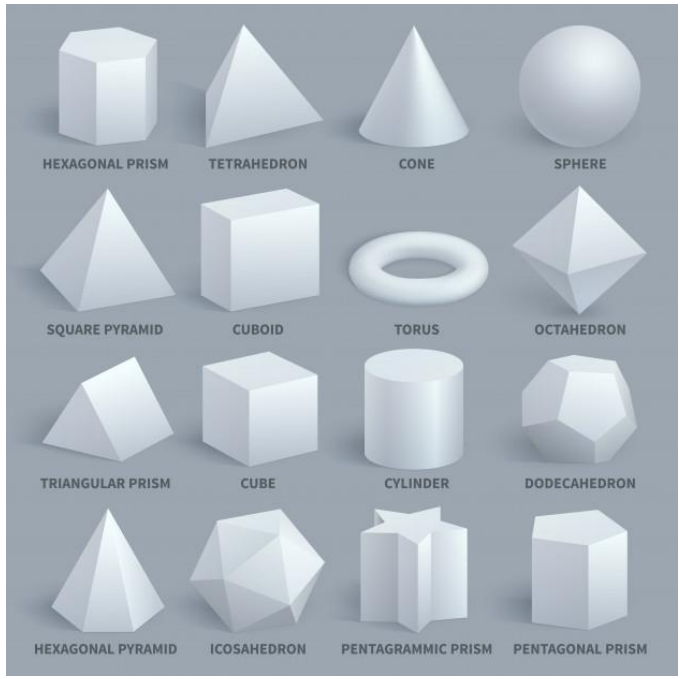
Objects

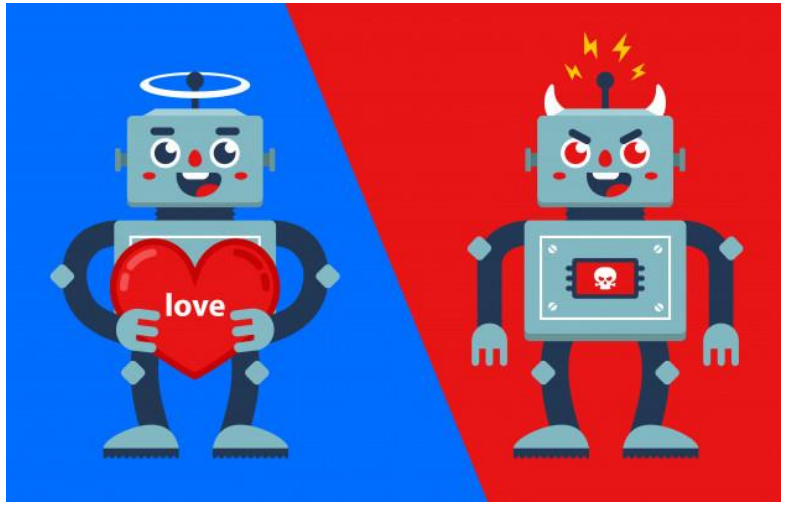
```
{ “name”:”Alice”,
  “age”:18,
  “favorite_color”:[“red”,”blue”,”green”]
}
{ “name”:”Bob”,
  “age”:25,
  “favorite_color”:[“red”,”black”],
  “city”:”California”
}
{ “name”:”Dogg”,
  “age”:30,
  “favorite_color”:[“pink”,”purple”,”yellow”],
  “membership”:null
}
```

This is also data.

Attributes

- An **attribute** is a property or characteistic of an object.





Attribute is also known as “variable”, “field”, “characteristic”, “dimension”, or “feature”

Objects

- A collection of attributes describe an **object**

Bird

"has beak"

"has wing"

"feather"

"has head"

"has leg"

Cow

"has ear"

"has snout"

"furry"

"has head"

"has leg"

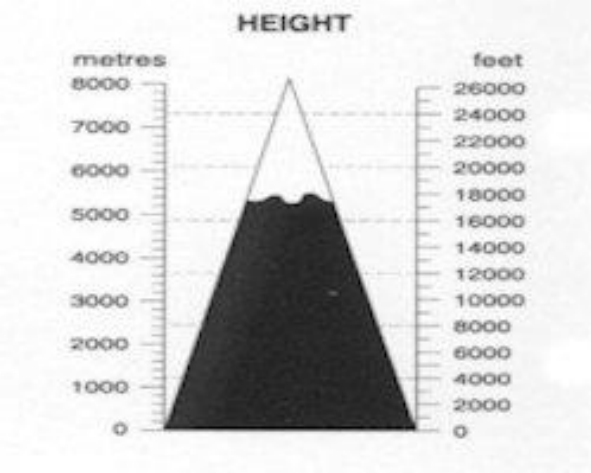
Attributes	Presence		Rating	
	walrus	polar bear	walrus	polar bear
Spot	no	no	less relevant	irrelevant
Blue	no	no	irrelevant	less relevant
Swim	yes	yes	highly relevant	relevant
Coastal	yes	yes	relevant	highly relevant

Object is also known as "record", "point", "case", "sample", "entity" or "instance".

1) Retrieved from https://www.ecse.rpi.edu/~cvrl/database/AttributeDataset.htm
2) Retrieved from http://seqamlab.com/2016/01/01/attribute-rating-for-classification-of-visual-objects/

Attribute Values

mountain_height = 8848 (meter)



mountain_height = 29032 (feet)

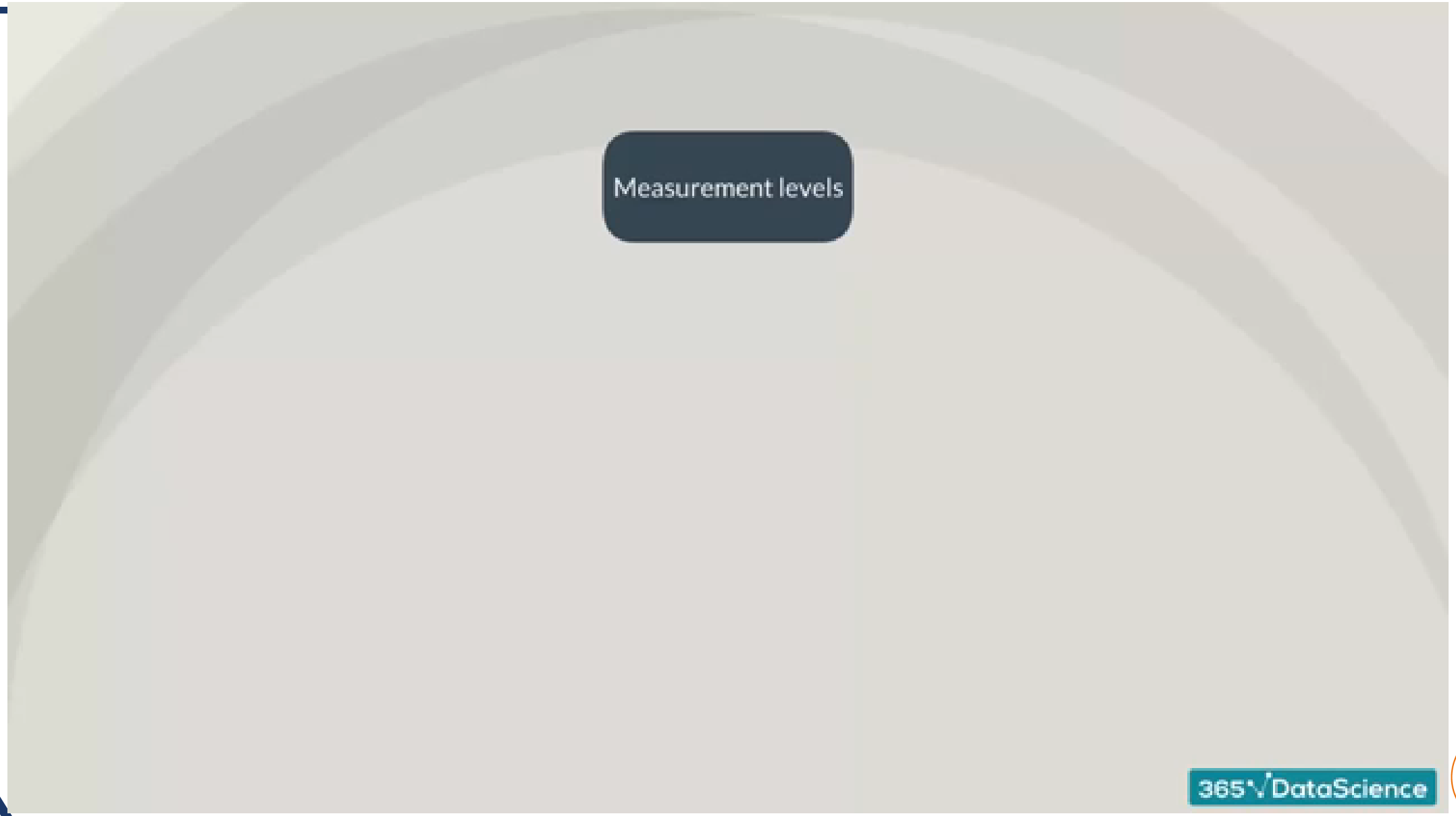
id is a Cardinal number value

id	18
name	Alice
age	18

age is a Ratio value

1) Retrieved from http://www.paulnoll.com/China/Metric/number-mountain-height.html

Types of Attributes

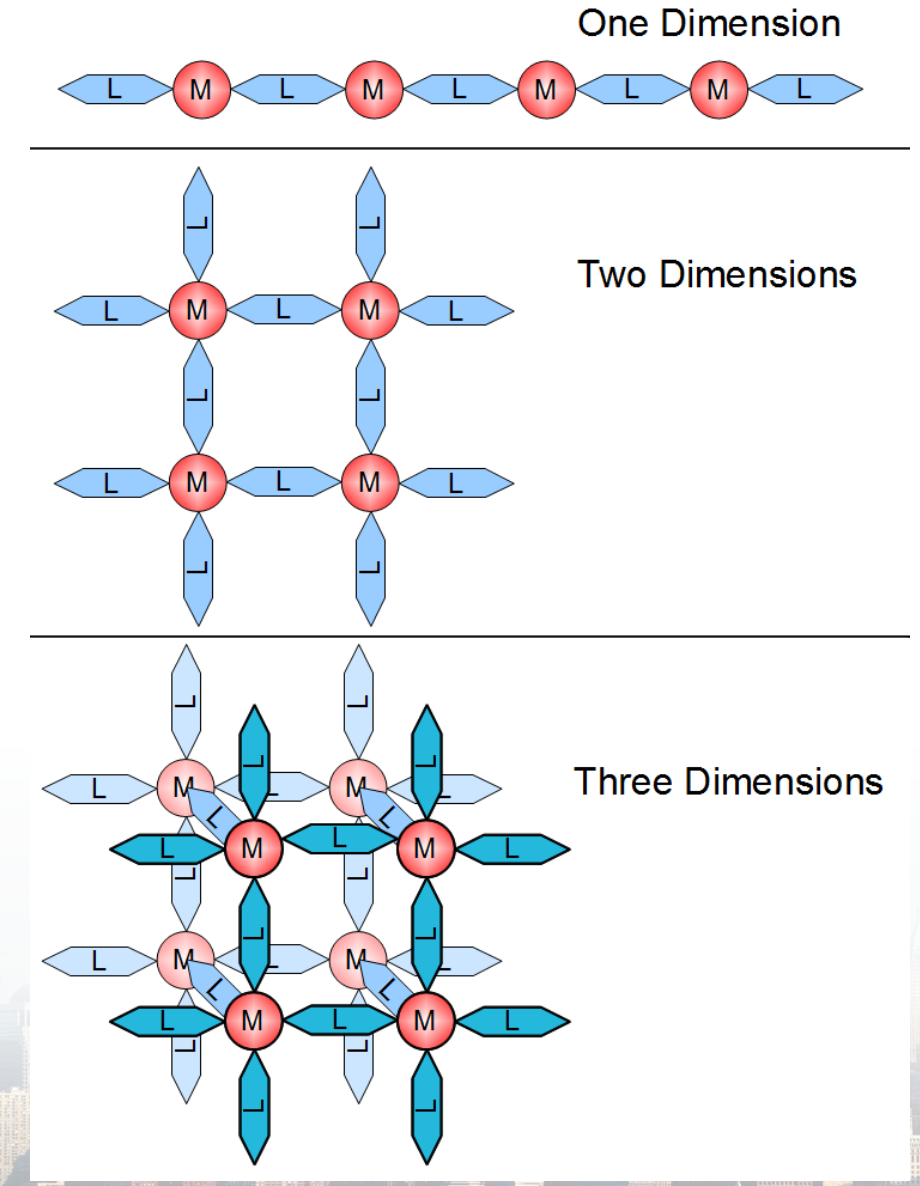


Properties of Attribute Values

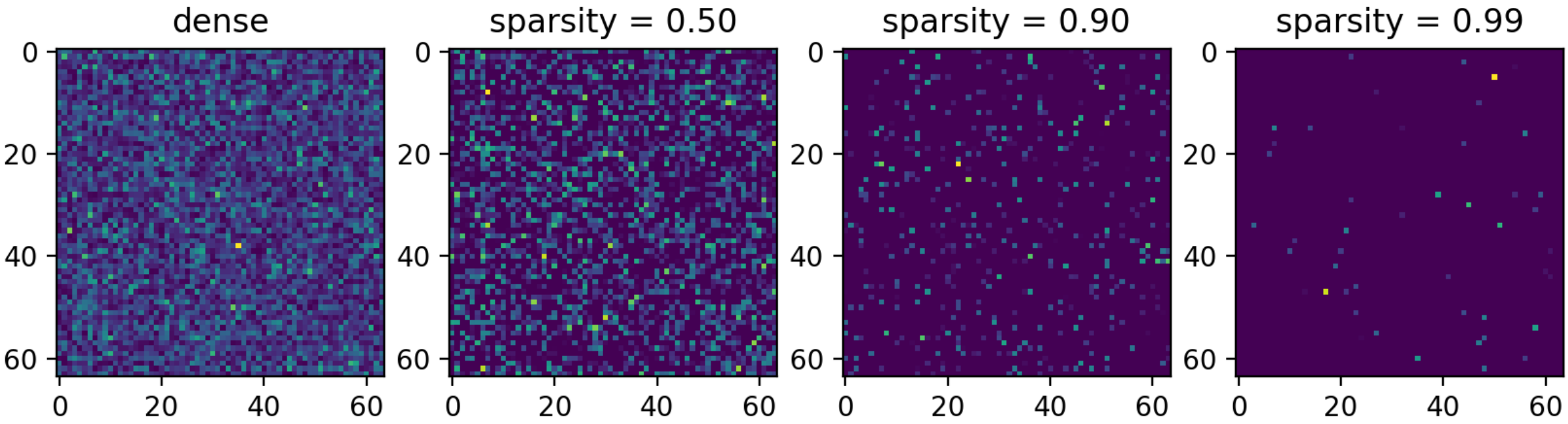
Rank	Incremental Progress	Measure Property	Mathematical Operators	Advanced Operations	Central Tendency	Type	Discreteness
1	Nominal（定类数据）	Classification, Membership	=, ≠	Grouping	Mode	Qualitative	Discrete
2	Ordinal（定序数据）	Comparison, Level	>, <	Sorting	Median	Qualitative	Discrete
3	Interval（定距数据）	Difference, Affinity	+, −	Yardstick	Mean, Deviation	Quantitative	Continuous
4	Ratio（定比数据）	Magnitude, Amount	*, /	Ratio	Geometric Mean, Coefficient of variation	Quantitative	Continuous

Types of Data

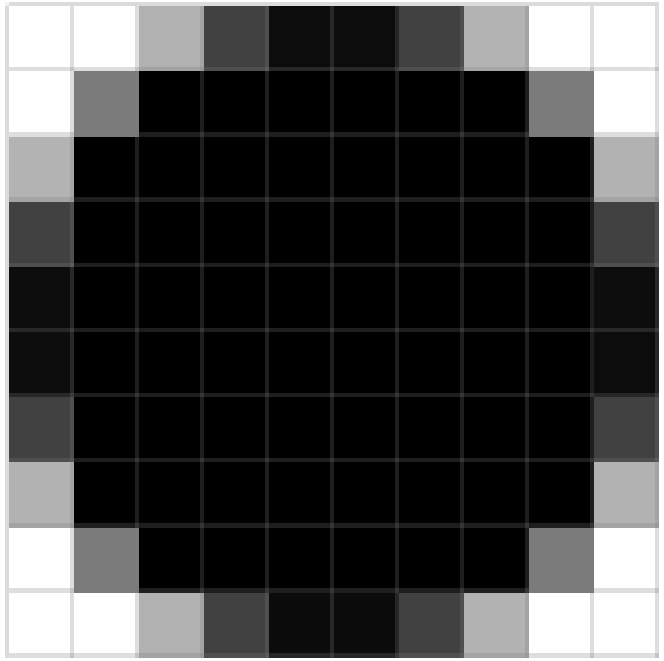
Important Characteristics of Data



Important Characteristics of Data

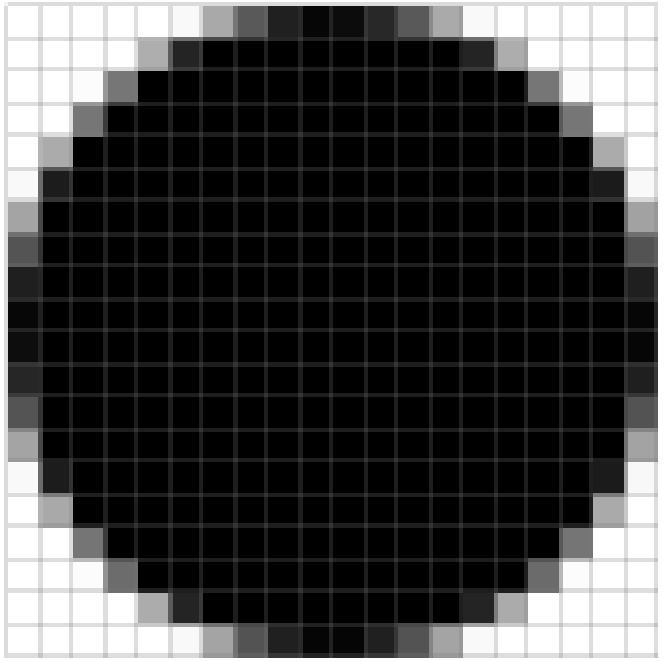


Important Characteristics of Data



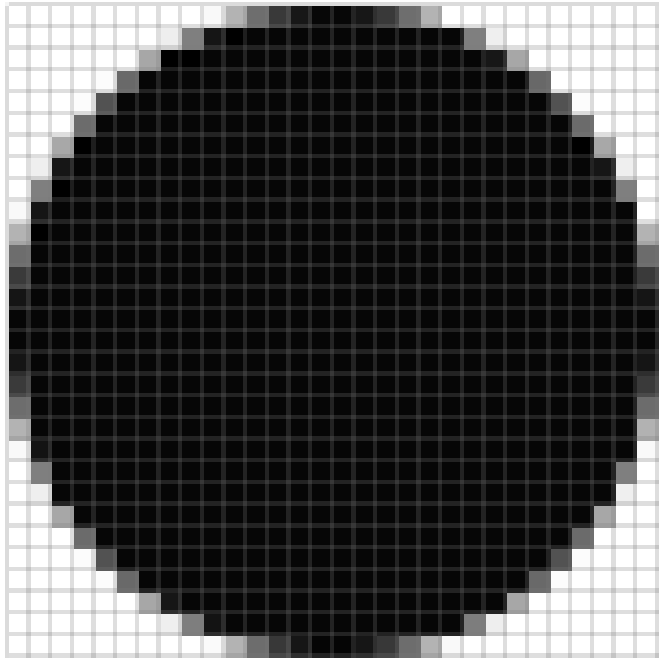
1x

(10 x 10 px)



2x

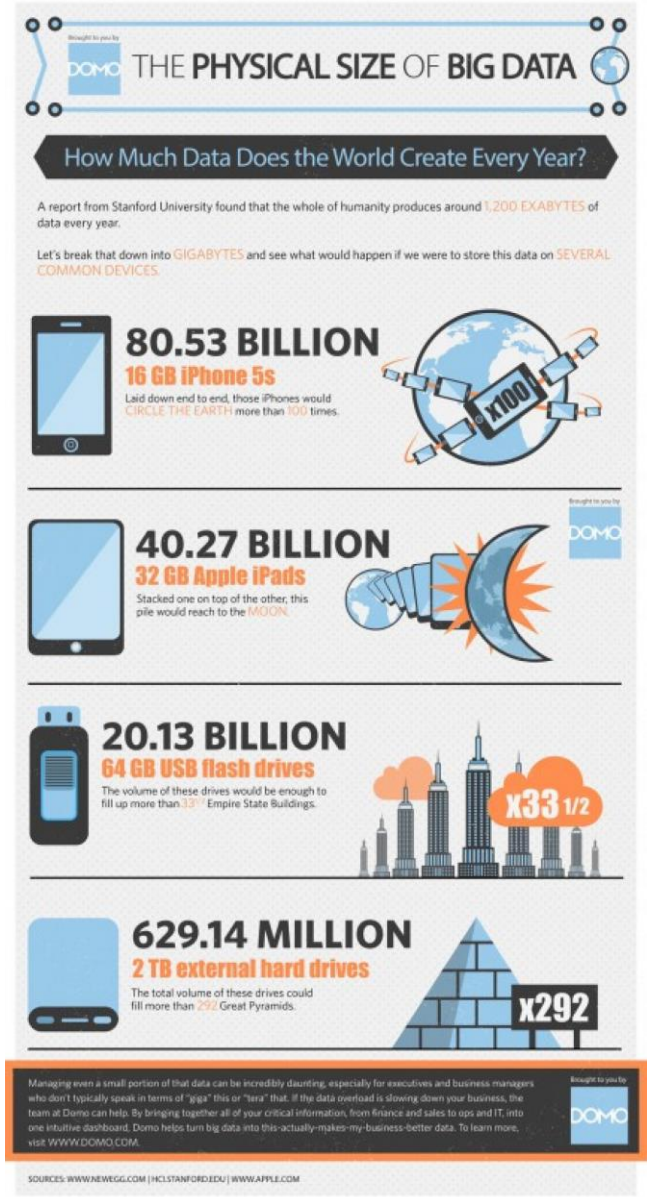
(20 x 20 px)



3x

(30 x 30 px)

Important Characteristics of Data

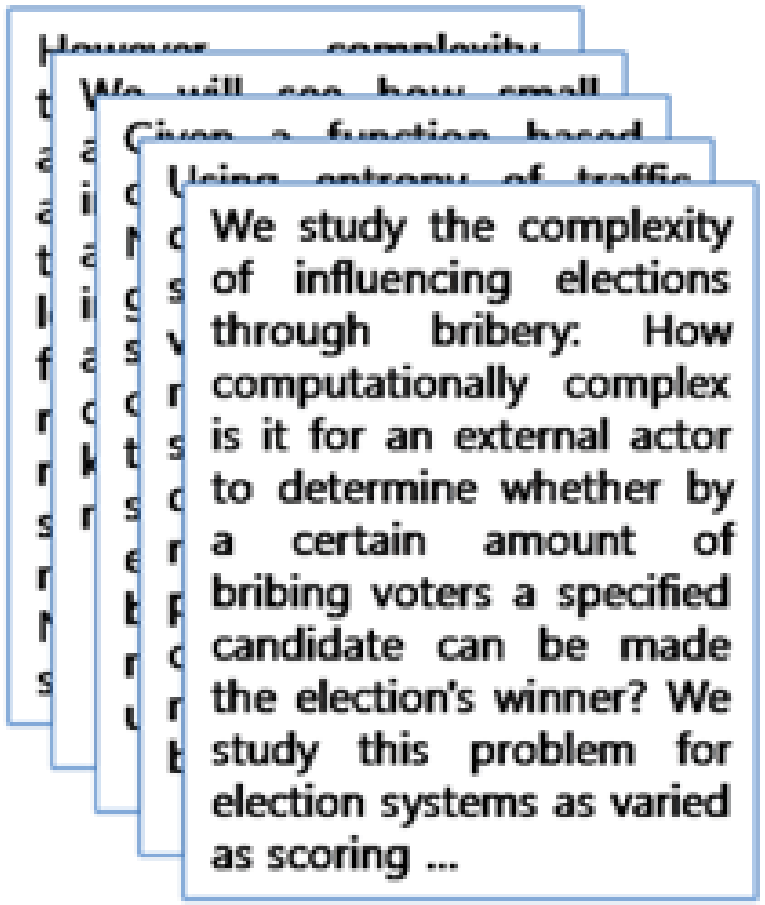


Basic Data Types

Types of Data Sets - Record

id	height	weight	score	type
1	157	61	28.7	A
2	155	50	7.0	A
3	155	63	17.8	B
4	154	44	15.1	B
5	153	70	15.1	B
6	151	60	13.6	B
7	158	38	29.6	C
8	152	44	8.1	D
9	149	57	2.1	D
10	153	41	27.0	D

Types of Data Sets - Record



Vector-space representation

	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

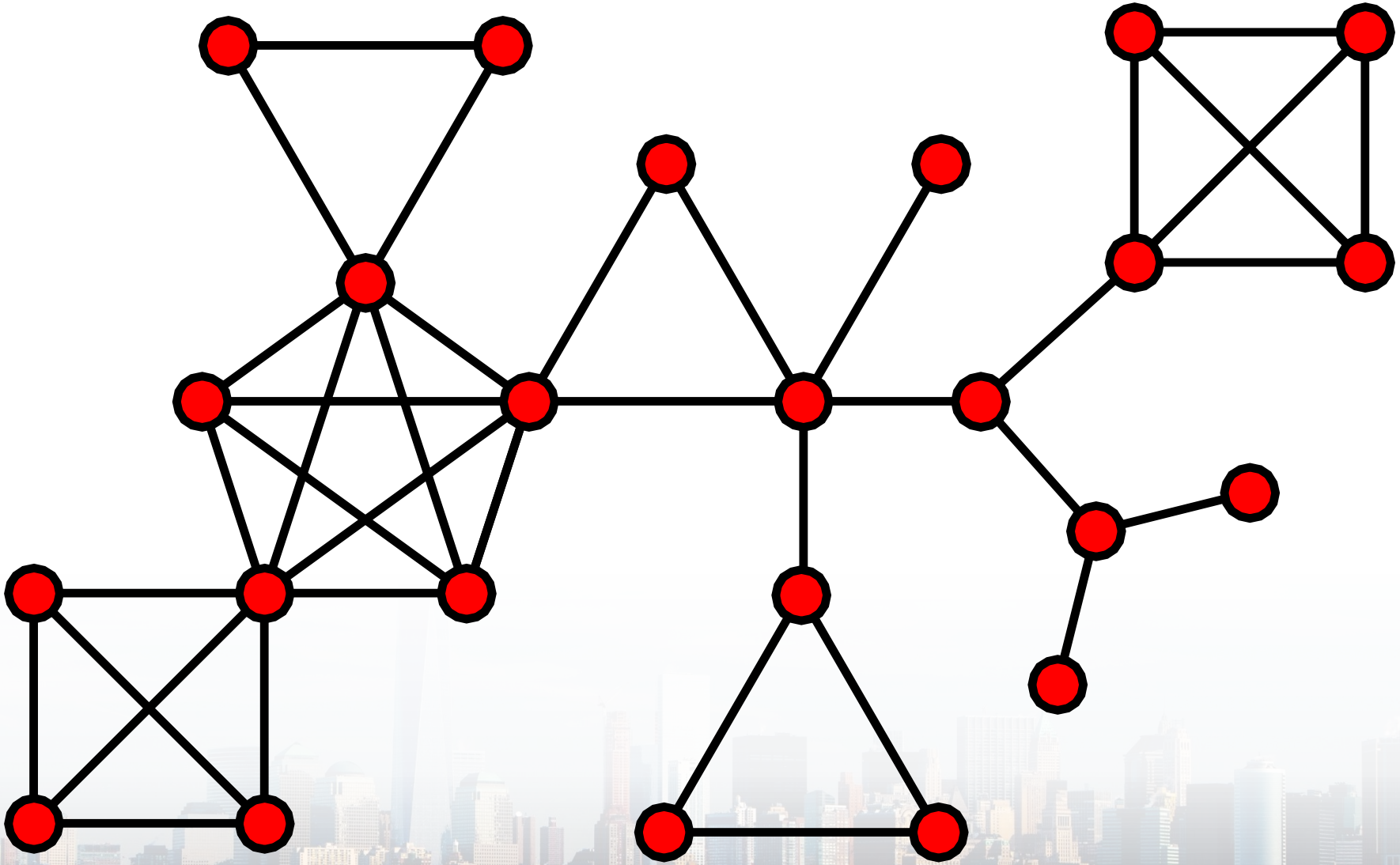
Term-document matrix

Types of Data Sets - Record

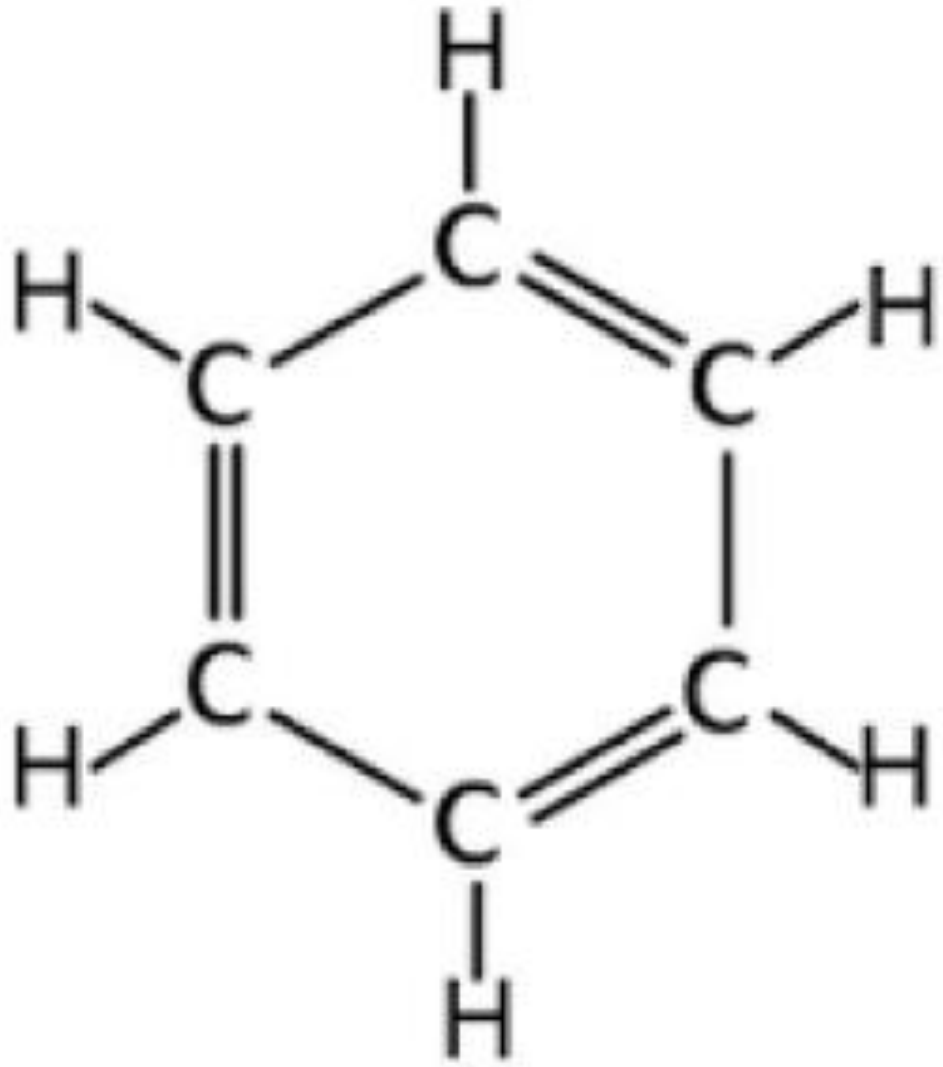
Account	Transaction Date	Transaction Type	Transaction Amount
12345	3/1/2017	Initial Deposit	$1,000
12345	3/3/2017	Payroll Deposit	$500
12345	3/3/2017	ATM Withdrawal	$100
12345	3/7/2017	Check	$75
12345	3/10/2017	Payroll Deposit	$500
12345	3/11/2017	ATM	$250
12345	3/17/2017	Payroll Deposit	$500
12345	3/20/2017	Check	$110
12345	3/22/2017	Web Payment	$135
12345	3/24/2017	Payroll Deposit	$500
12345	3/24/2017	Web Payment	$90
12345	3/24/2017	ATM Withdrawal	$125
12345	3/28/2017	Check	$50
12345	3/30/2017	ATM Withdrawal	$65
12345	3/31/2017	Payroll Deposit	$500

1) Retrieved from https://www.nuwavesolutions.com/snapshot-fact-tables/

Types of Data Sets - Graph

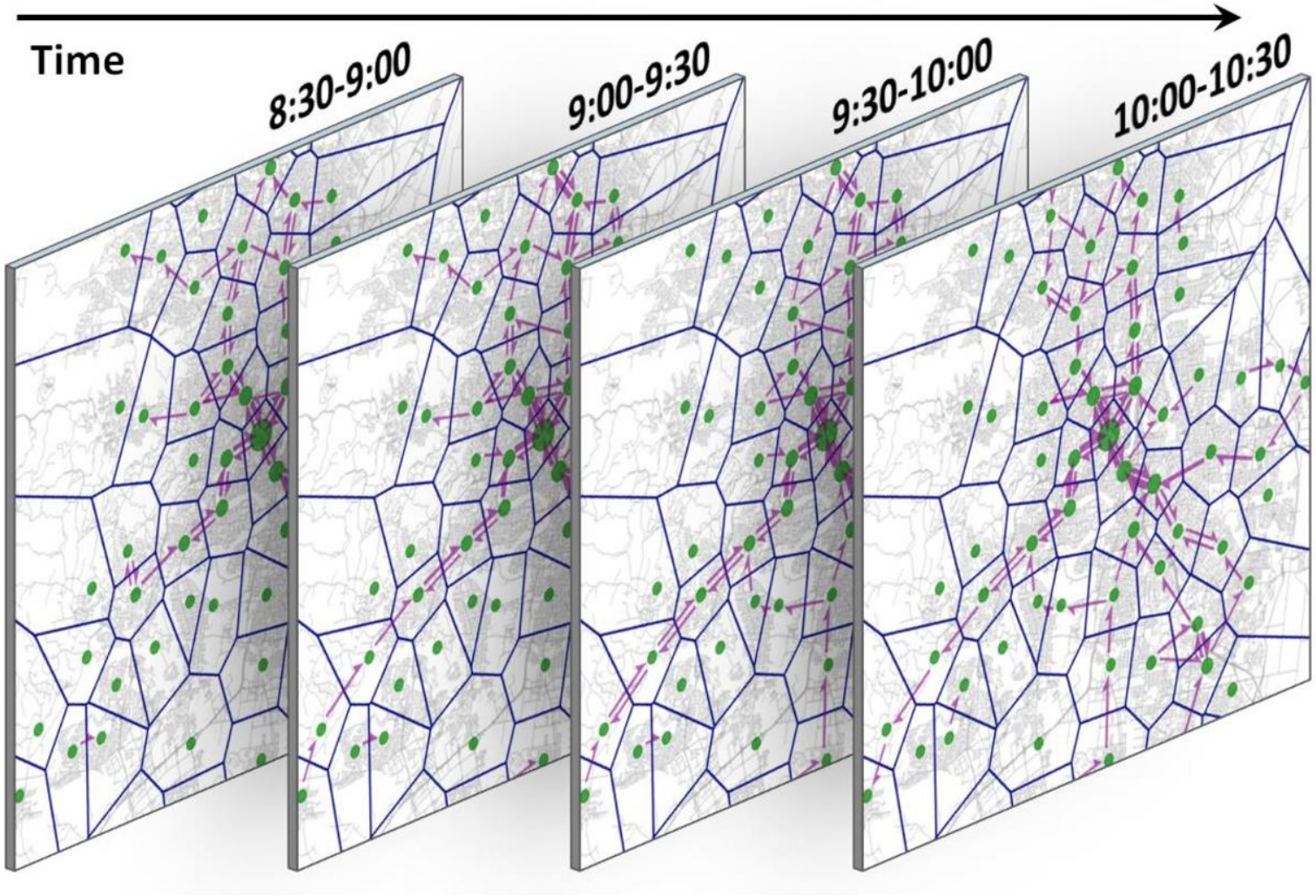


Types of Data Sets - Graph



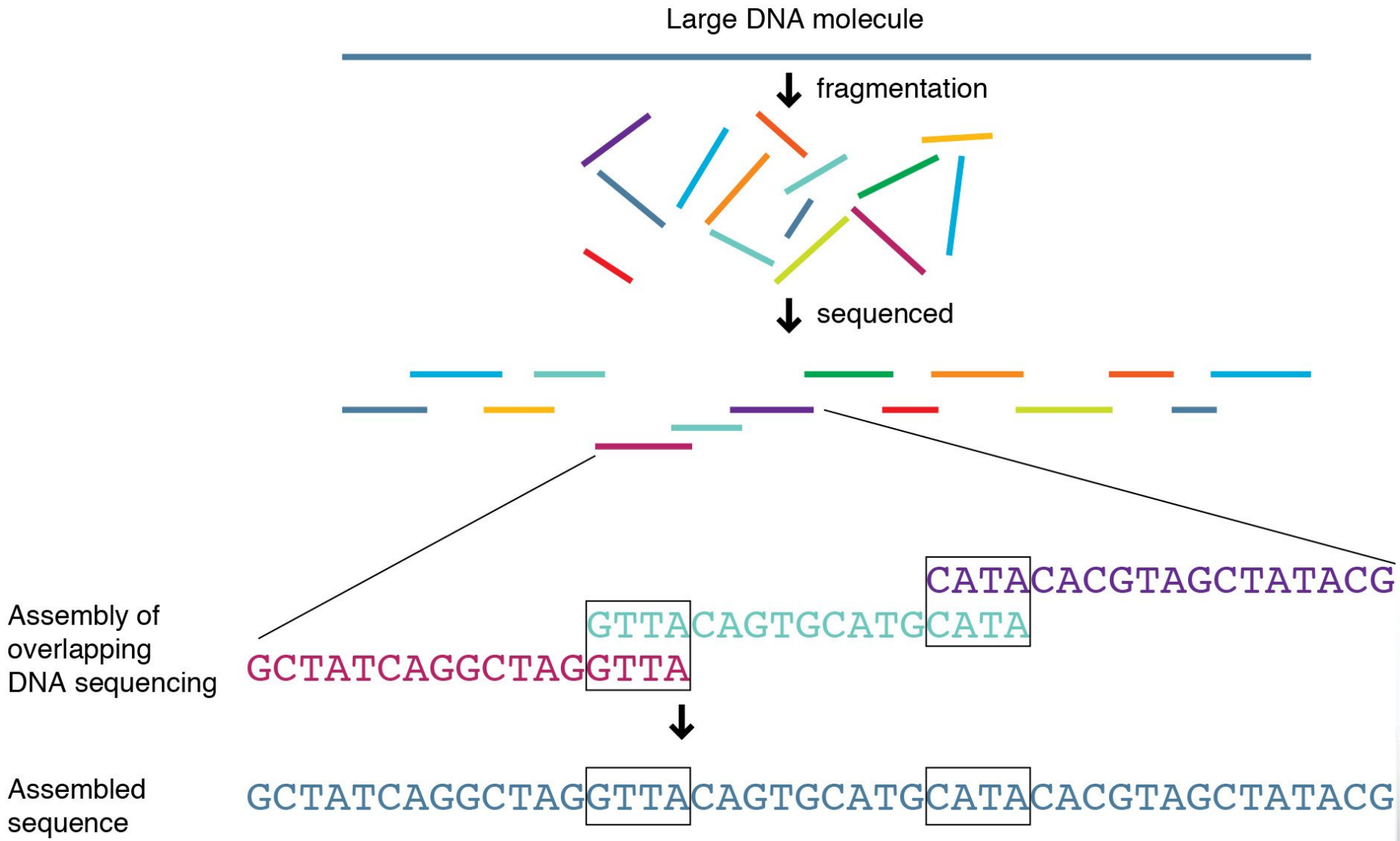
1) Retrieved from https://regenesis.com/en/glossary/benzene/

Types of Data Sets - Graph



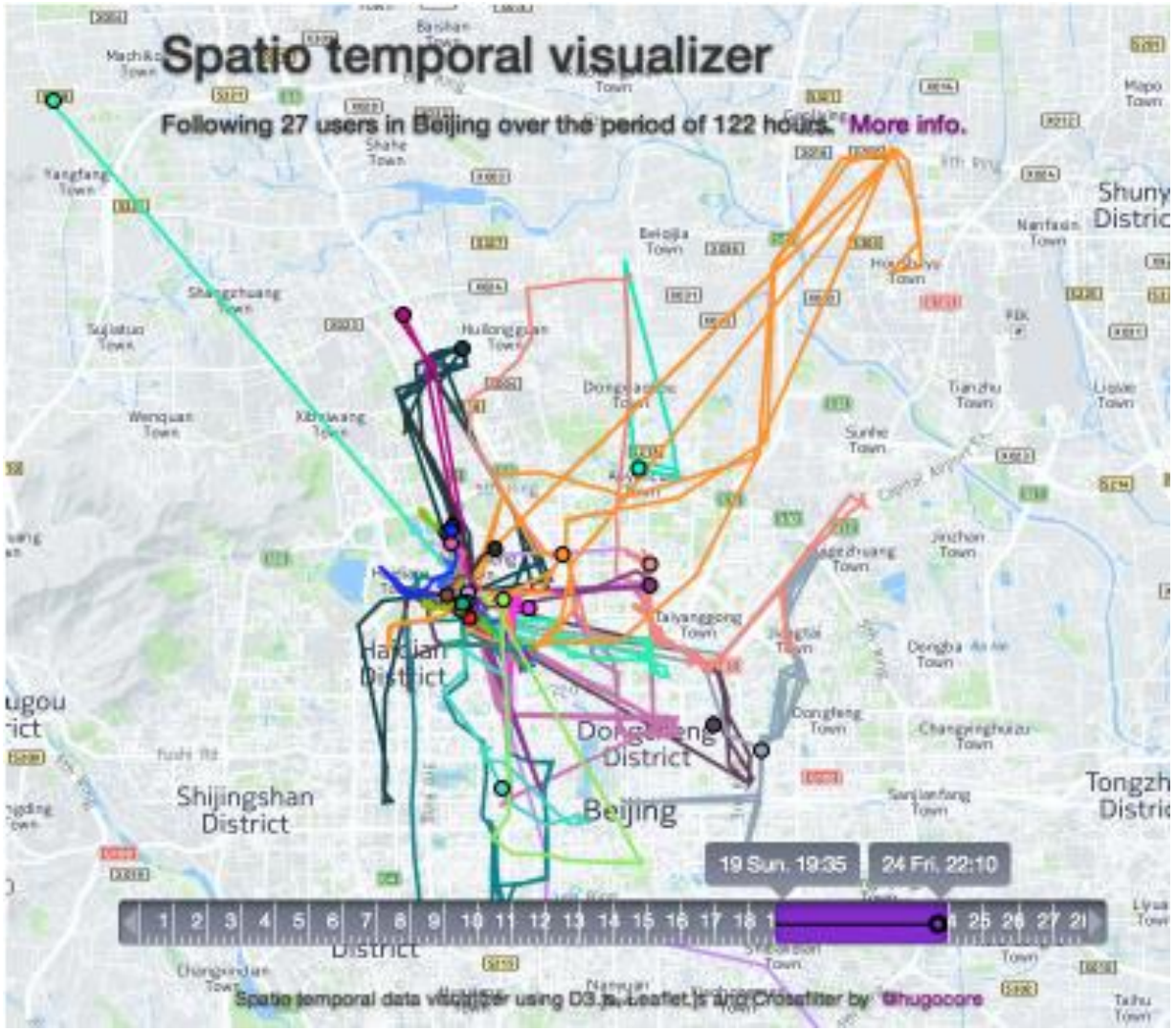
1) Retrieved from https://www.semanticscholar.org/paper/Graph-based-analysis-of-city-wide-traffic-dynamics-Kim-Zheng/1d6ad012d3129bf016e00f443e7bd0ce7a5463d5

Types of Data Sets - Ordered Data



1) Retrieved from https://knowgenetics.org/whole-genome-sequencing/

Types of Data Sets - Ordered Data



1) Retrieved from https://www.researchgate.net/figure/This-work-tool-to-visualize-spatio-temporal-data_fig1_266734204

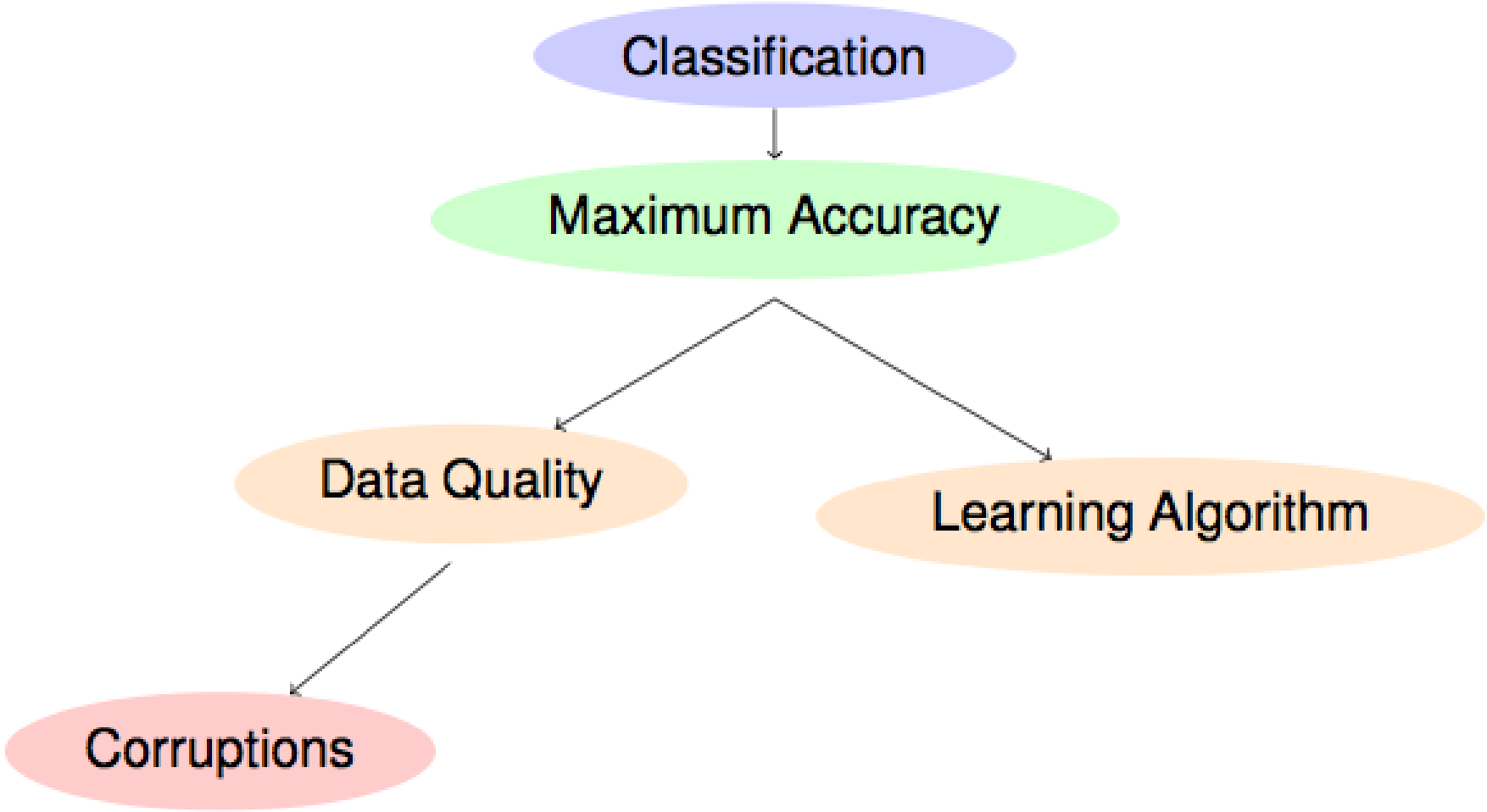
Quality of Data

Data Quality

Data Quality Common Problems

- What problems should we worry about?
- How can we detect problems with the data?
- What can we do about these problems?

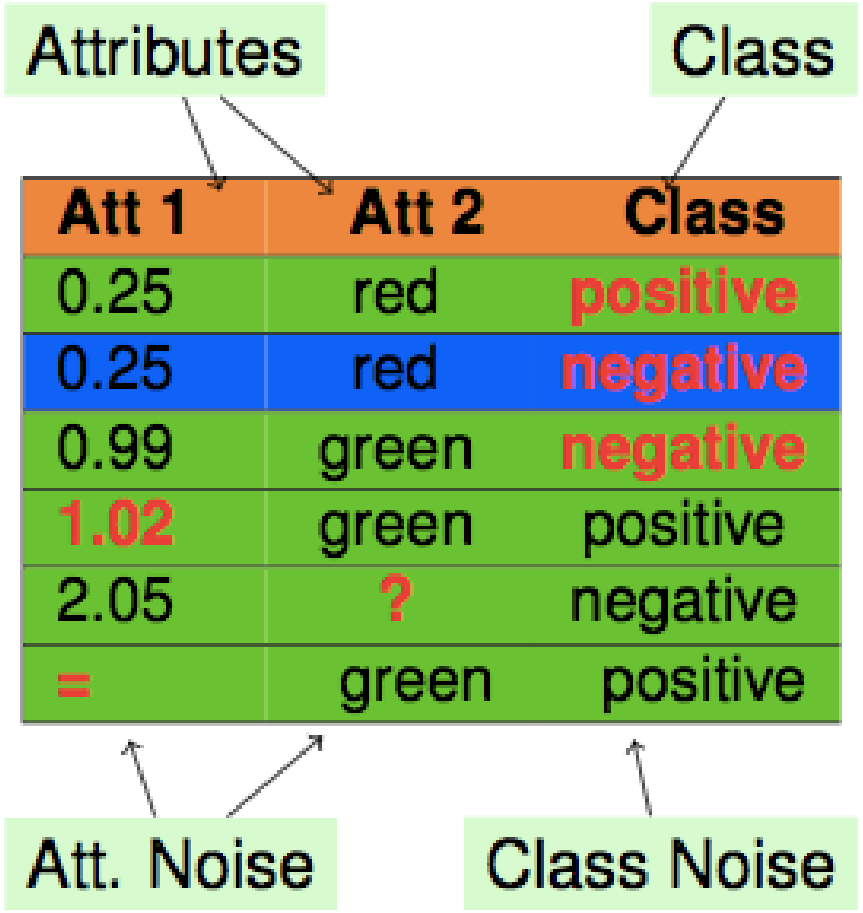
Noise



1) Retrieved from https://sci2s.ugr.es/noisydata

Noise

Information Sources



Kinds of Noise

Class Noise

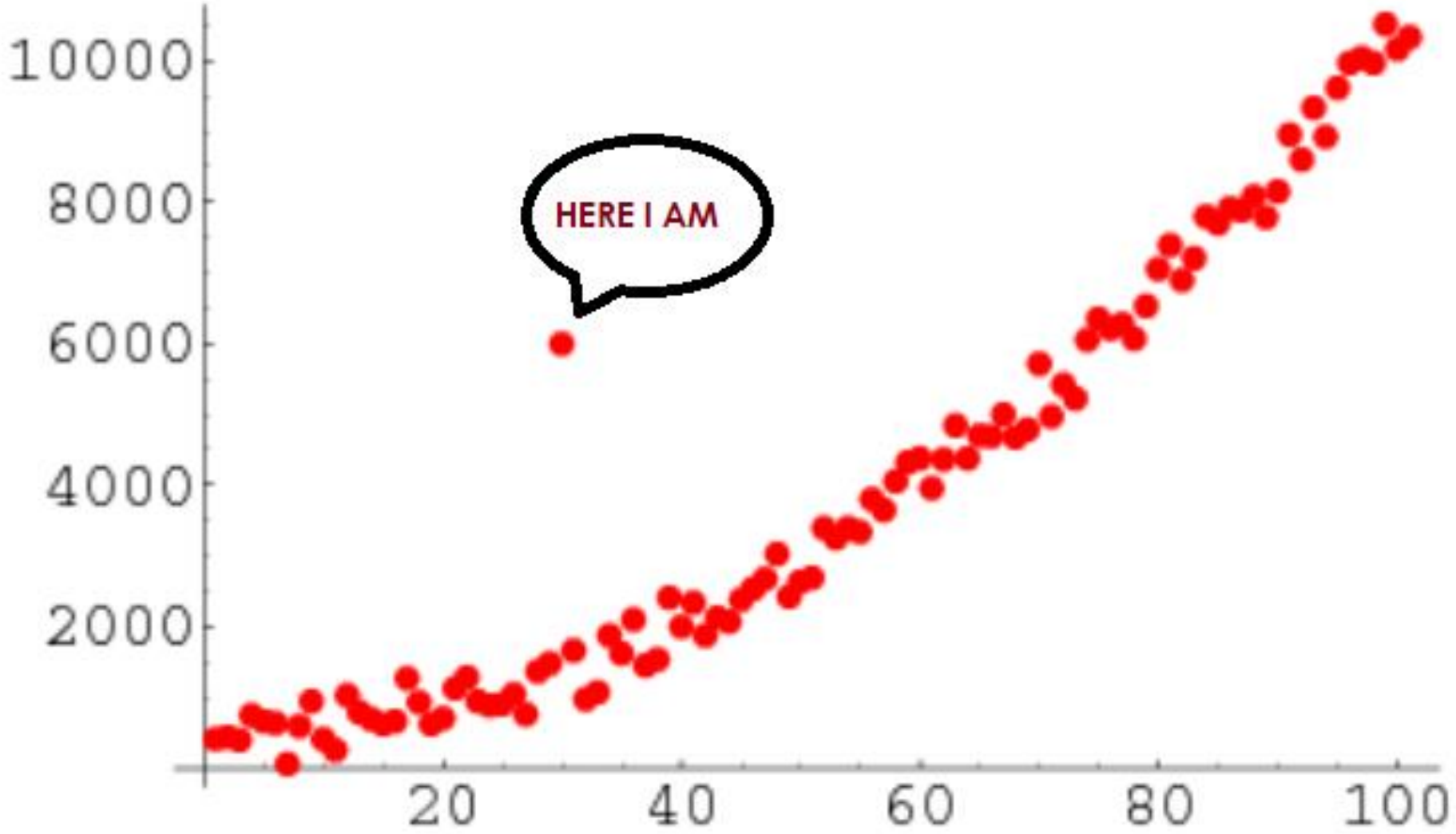
- Contradictory examples
- Mislabeled examples

Attribute Noise

- Erroneous values
- Missing values
- Don't care values

1) Retrieved from https://sci2s.ugr.es/noisydata

Outliers



1) Retrieved from https://medium.com/analytics-vidhya/its-all-about-outliers-cbe172aa1309

Duplicate Data

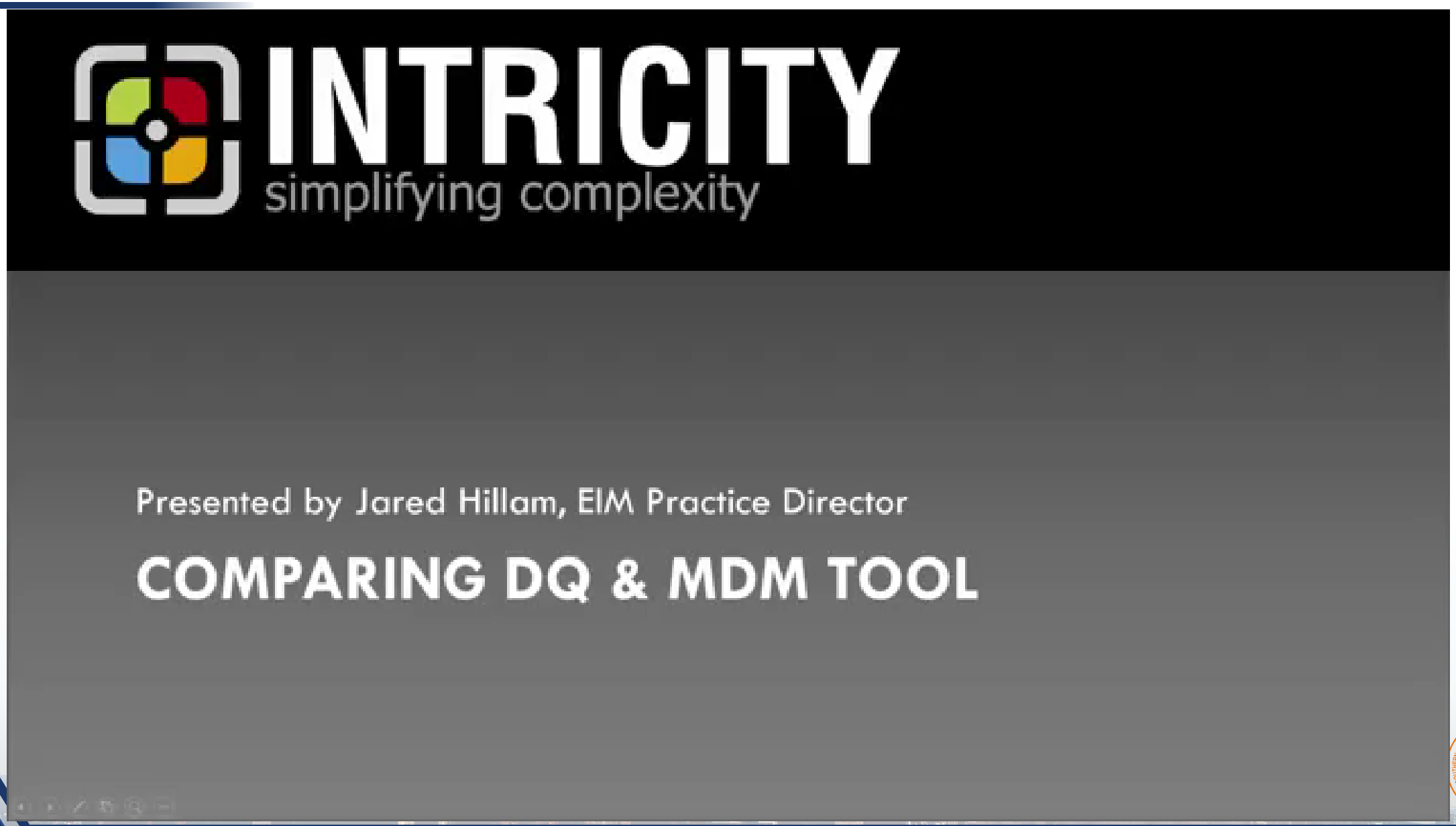
1) Retrieved from http://www.datamartist.com/duplicate-data-and-removing-duplicate-records

Missing Values

1) Retrieved from https://towardsdatascience.com/practical-strategies-to-handle-missing-values-626f9c43870b

Data Quality

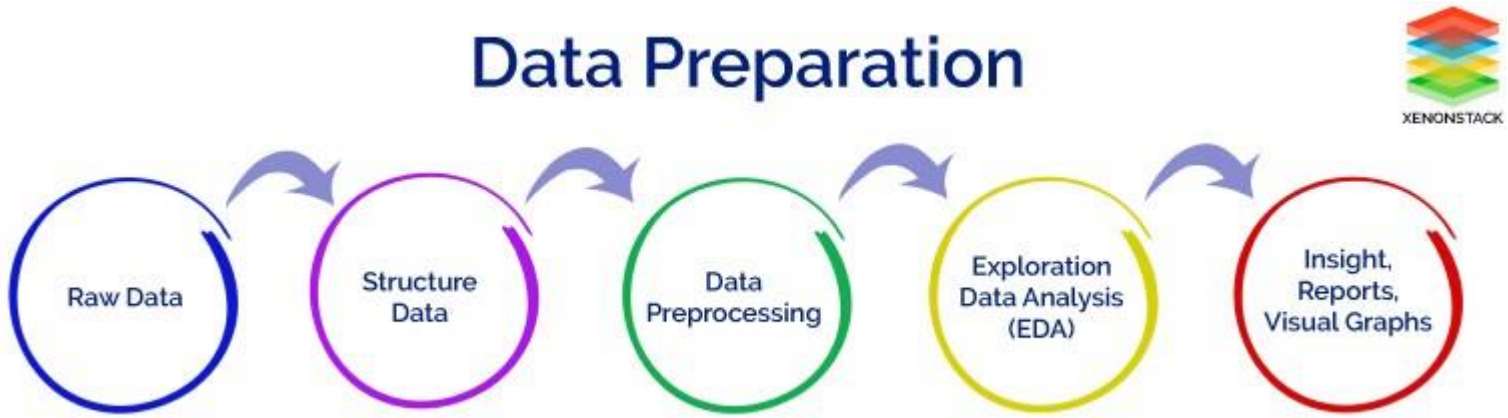
Data Quality



Preprocessing

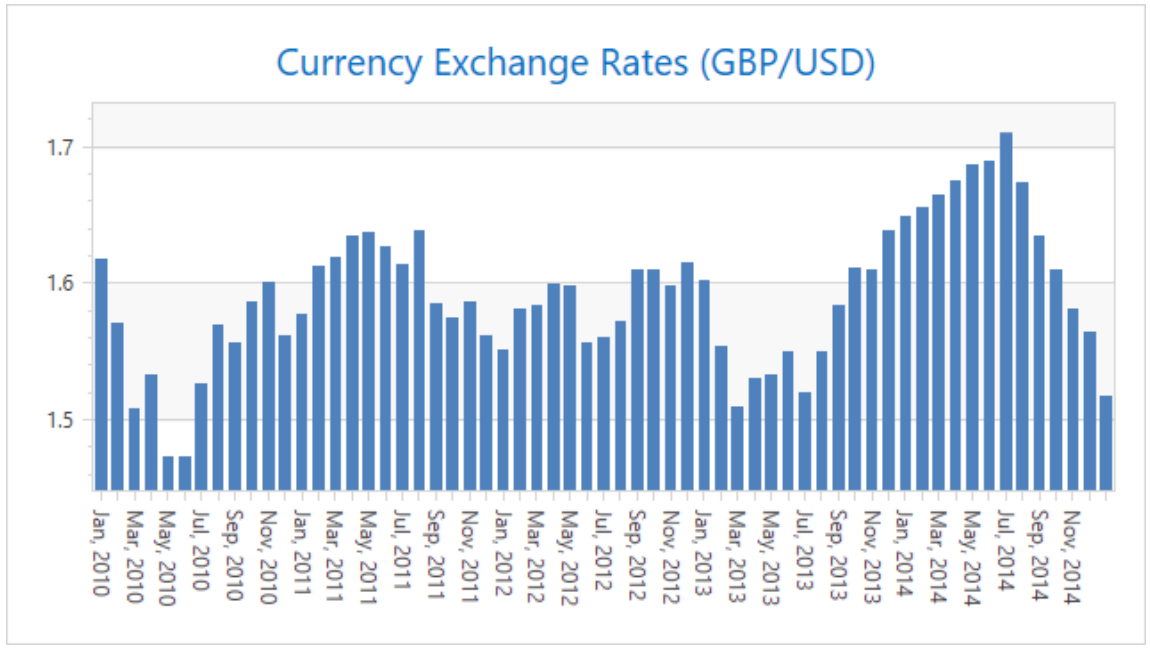
Data Preprocessing

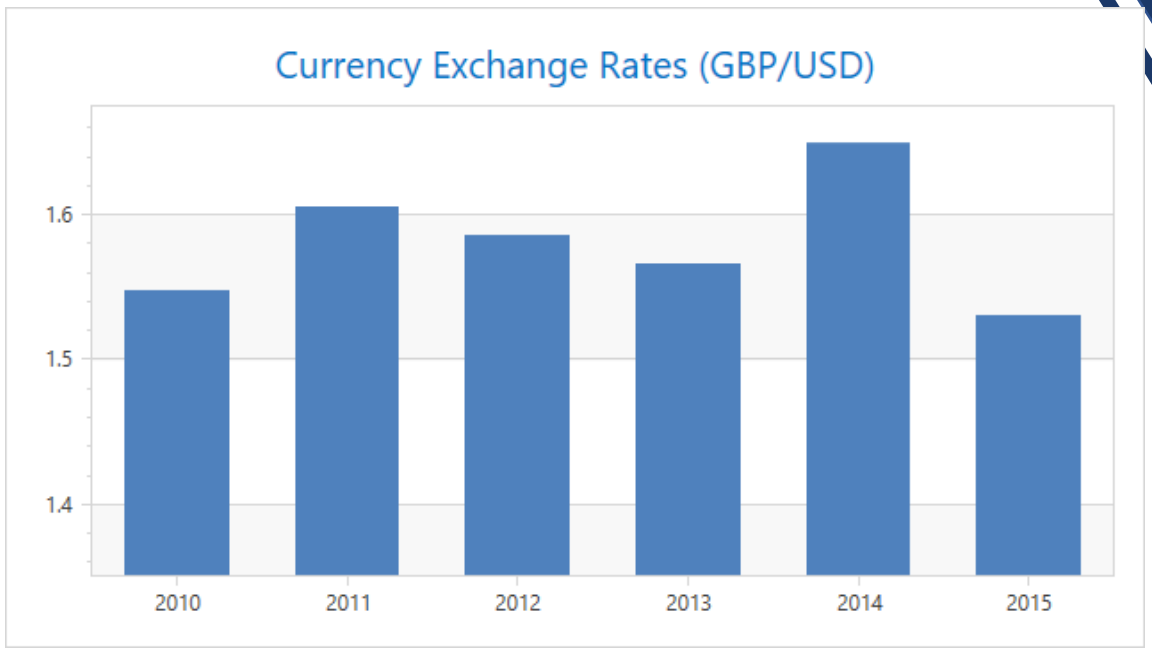
- Aggregation
- Sampling
- Dimensionality Reduction
- Feature Subset Selection
- Feature Creation
- Discretization and Binarization
- Attribute Transformation



1) Retrieved from https://medium.com/@mallrishabh52/data-preprocessing-and-workflow-of-a-machine-learning-project-bc6ca9e6f3ad

Data Aggregation



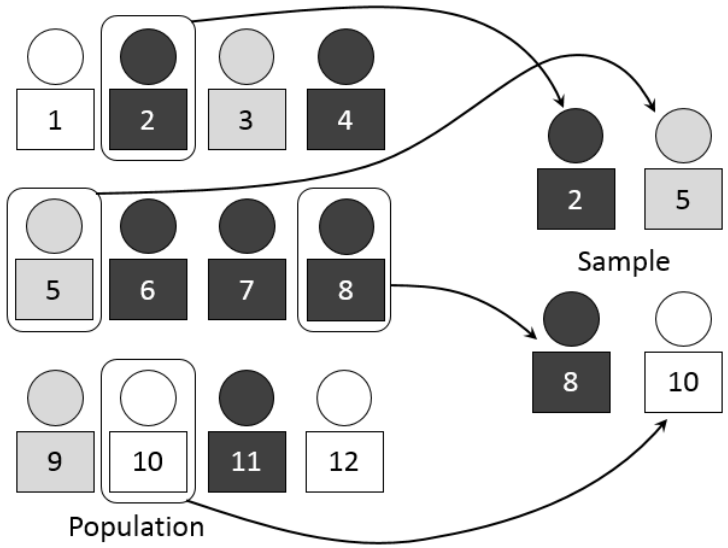


Aggregate by Year

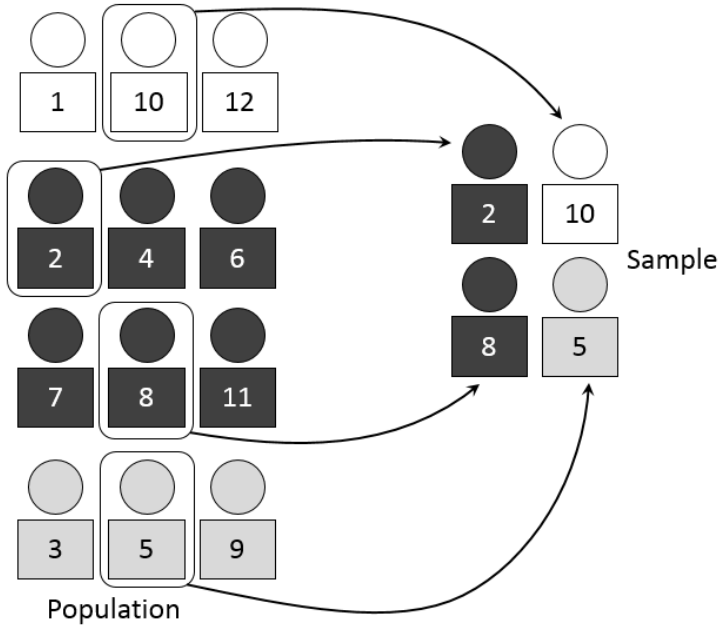
1) Retrieved from https://docs.devexpress.com/WPF/16846/controls-and-libraries/charts-suite/chart-control/providing-data/data-aggregation

Data Sampling

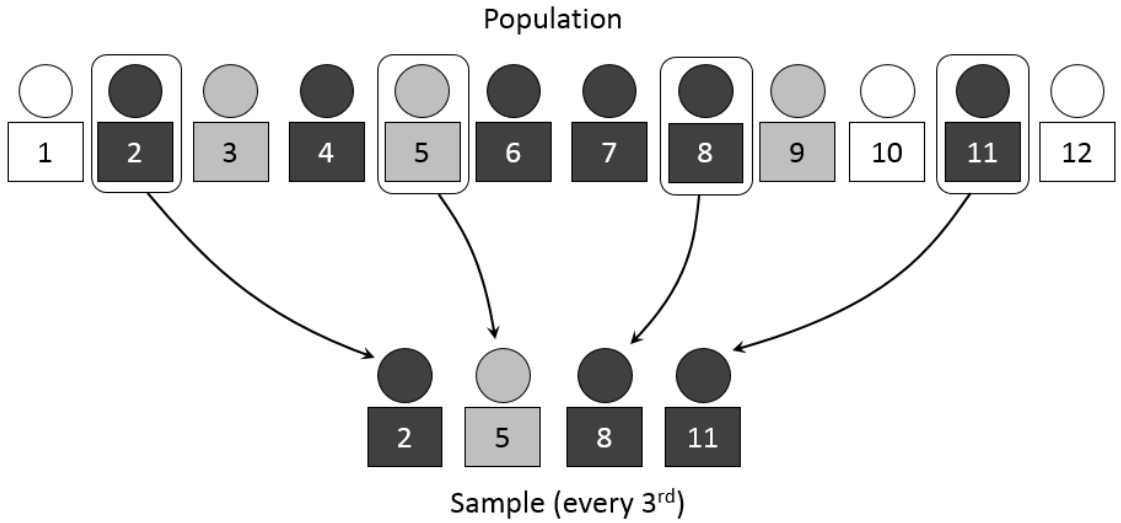
Types of Data Sampling



Simple Random Sampling

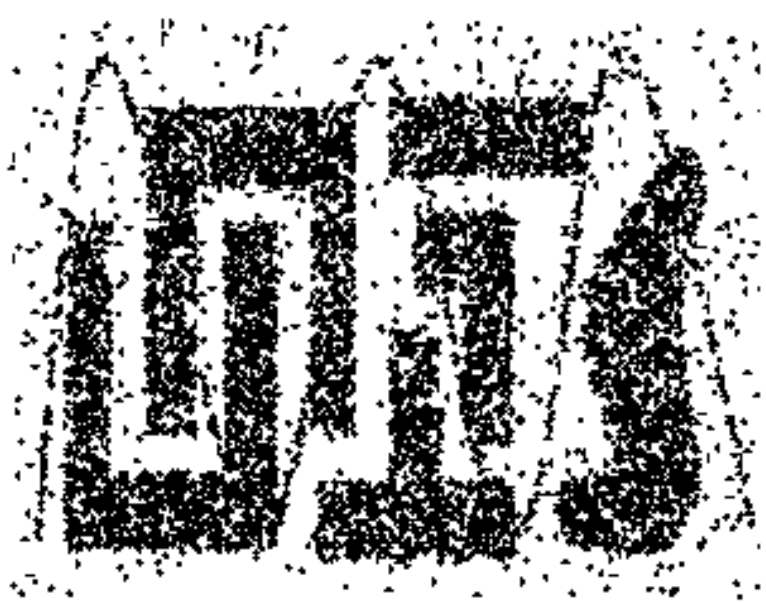


Stratified Sampling

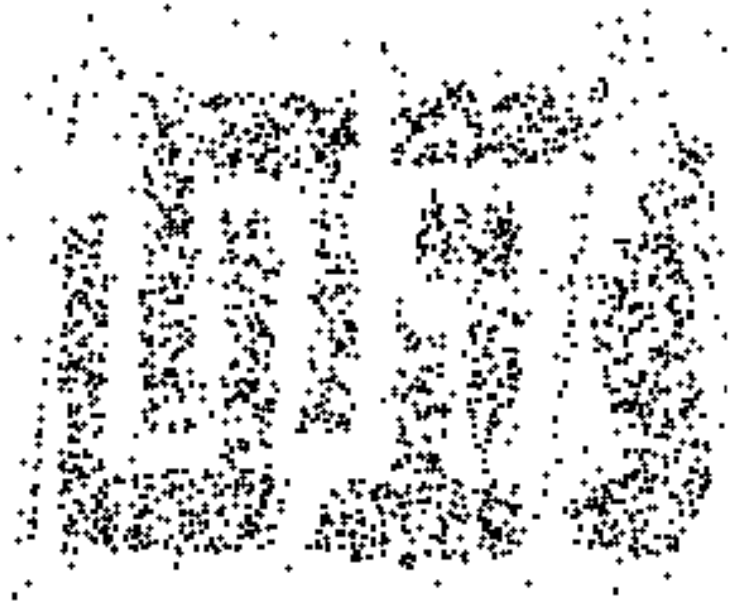


Systematic Sampling

Sampling Size



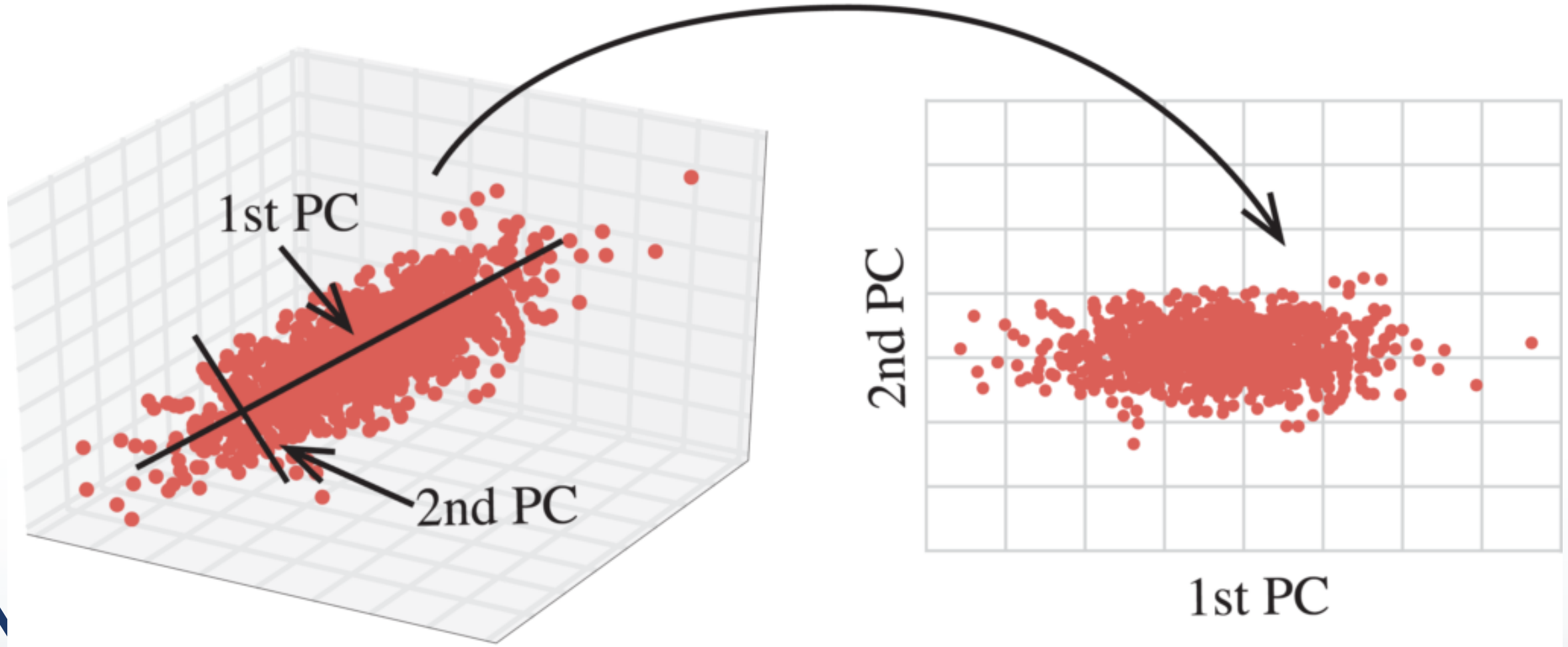
8000 points



2000 Points

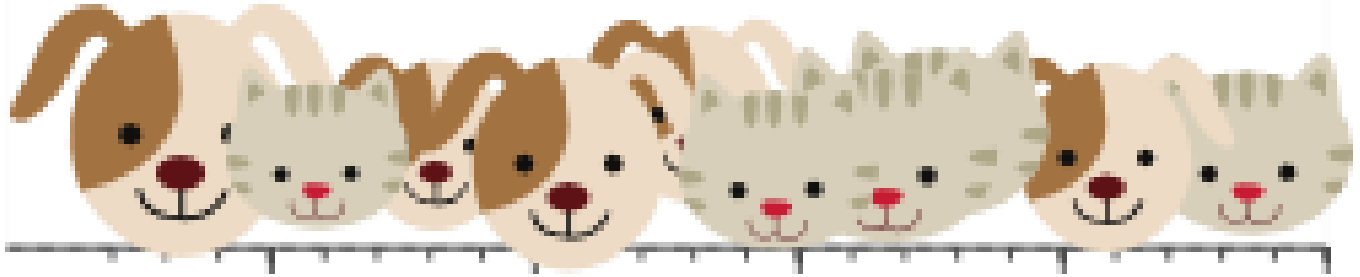
500 Points

Dimension Reduction



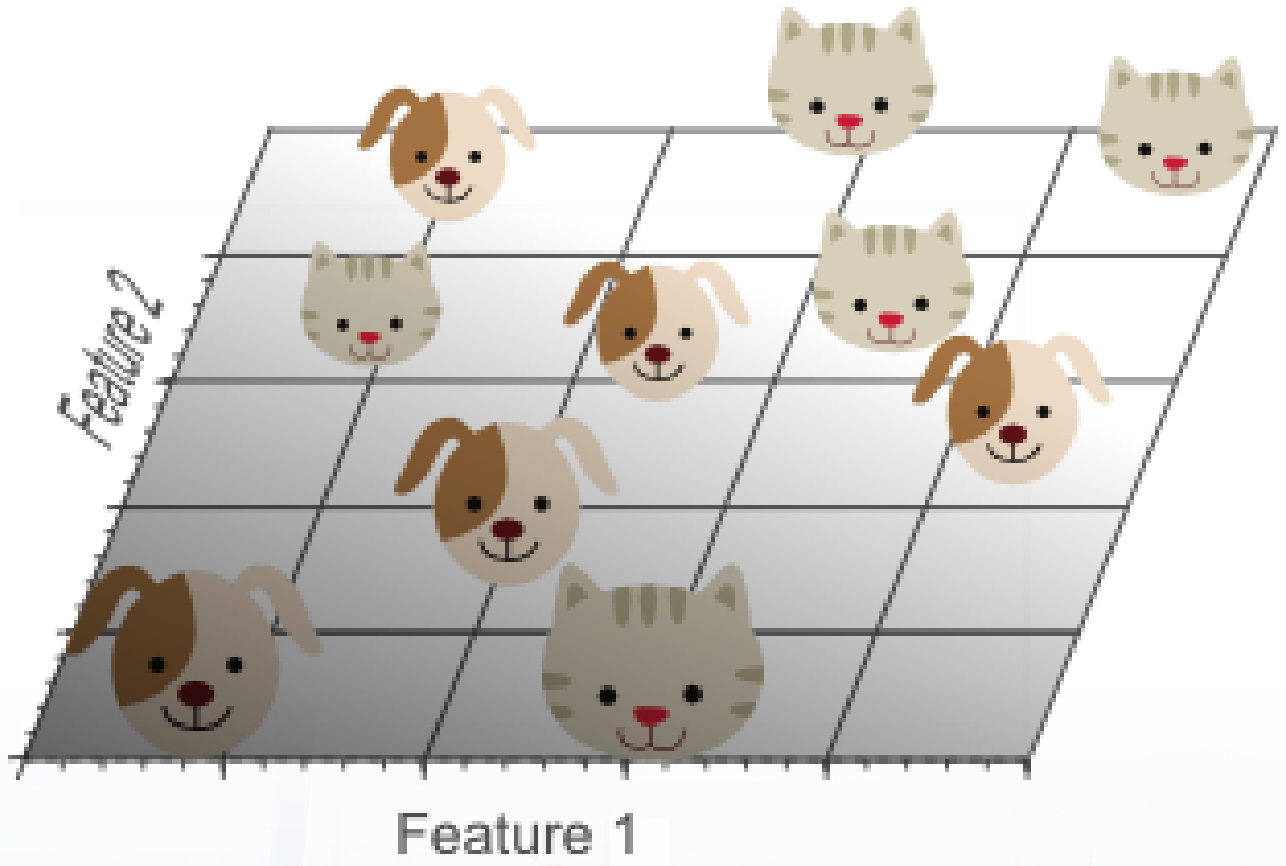
1) Retrieved from https://medium.com/@TheDataGyan/dimensionality-reduction-with-pca-and-t-sne-in-r-2715683819

Dimension Reduction Example



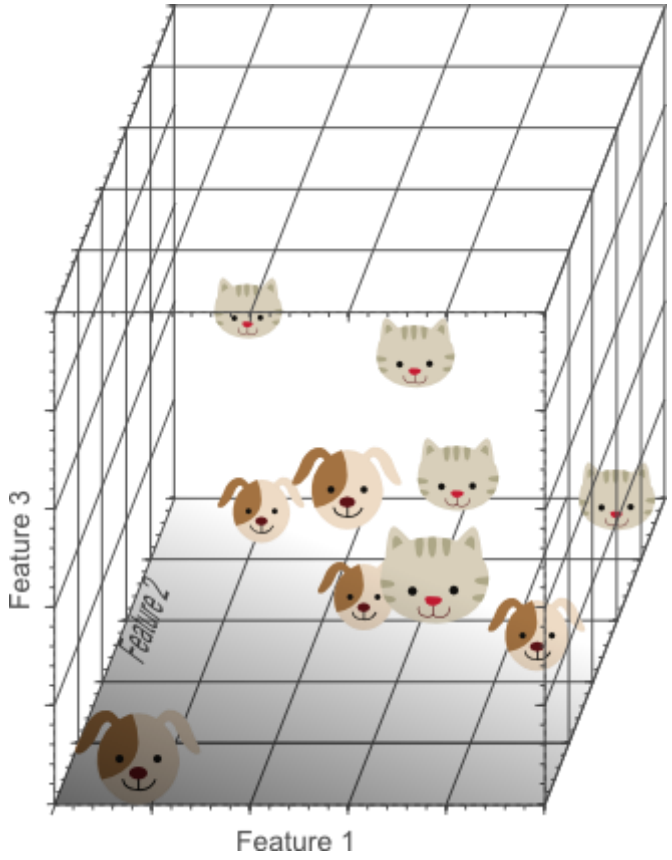
1) Retrieved from https://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/

Dimension Reduction Example



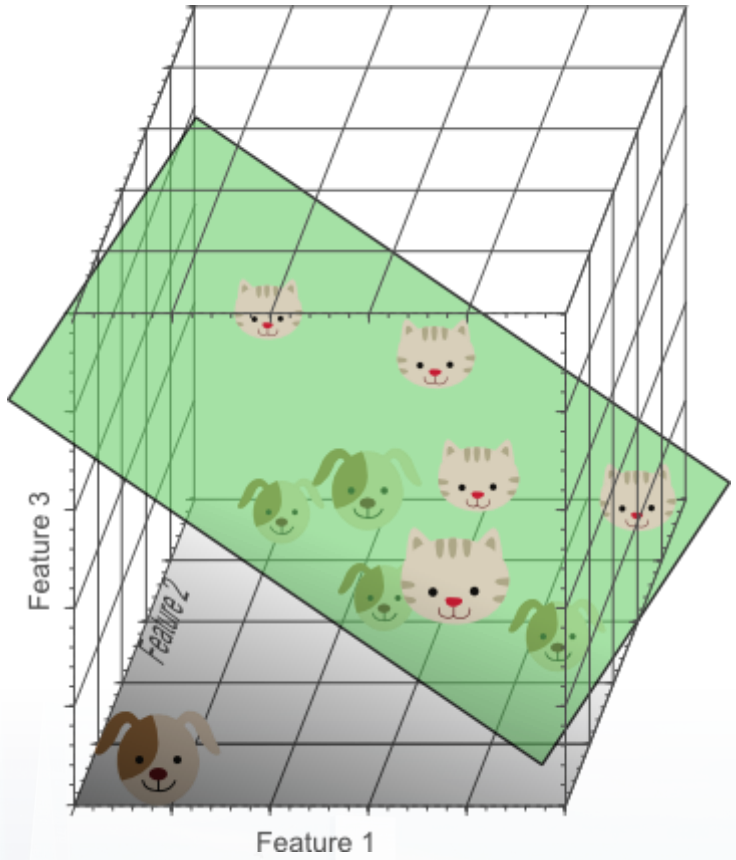
1) Retrieved from https://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/

Dimension Reduction Example



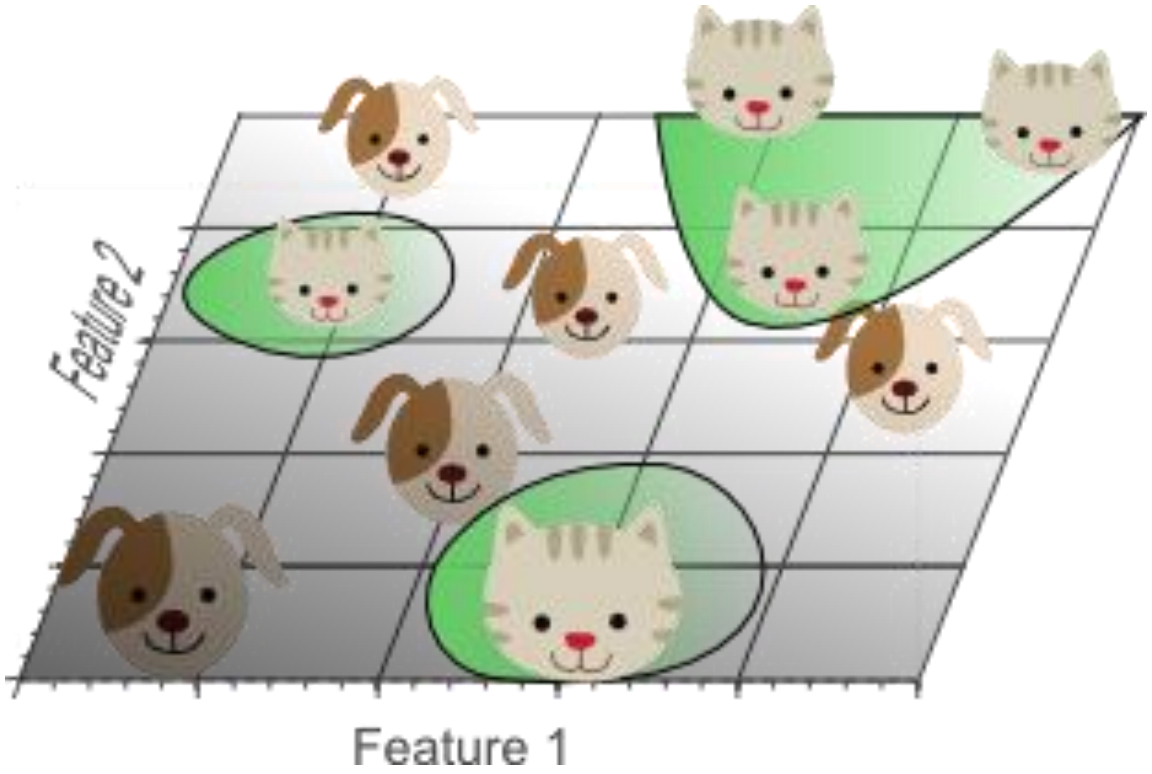
1) Retrieved from https://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/

Dimension Reduction Example



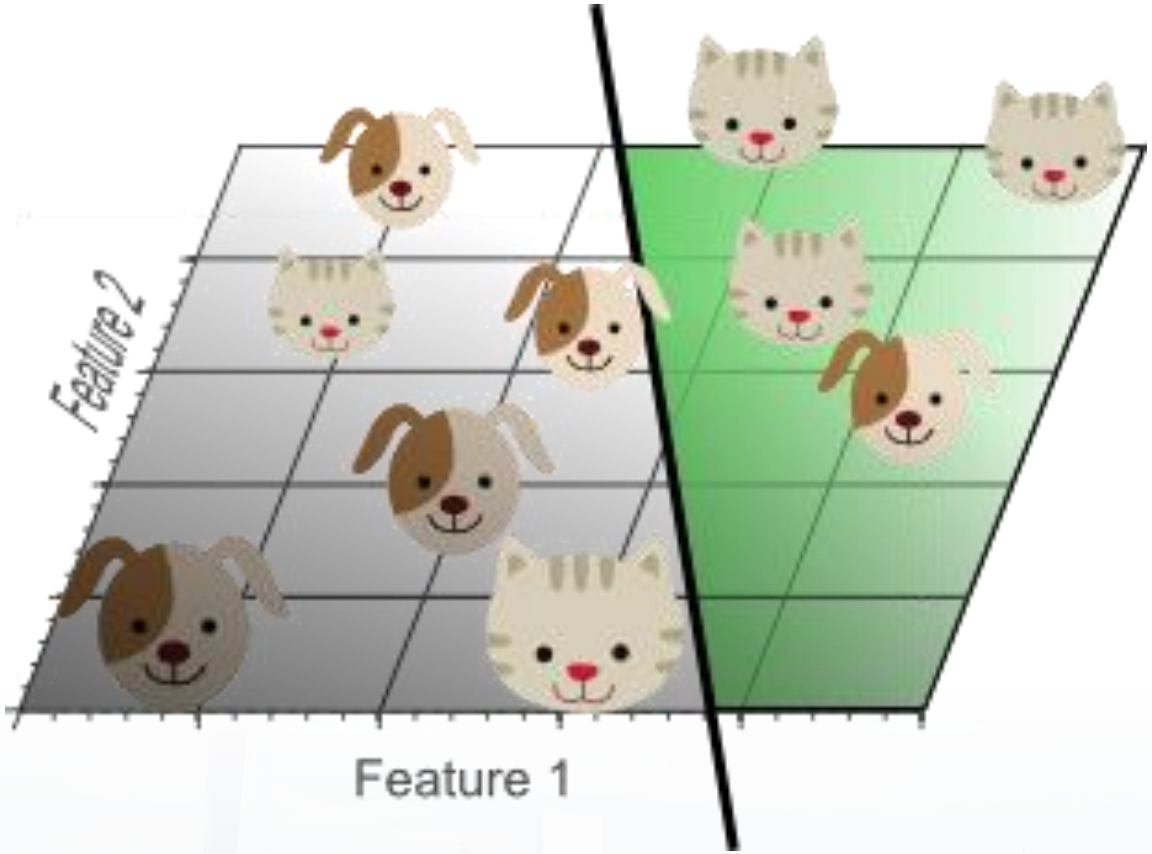
1) Retrieved from https://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/

Dimension Reduction Example



1) Retrieved from https://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/

Dimension Reduction Example

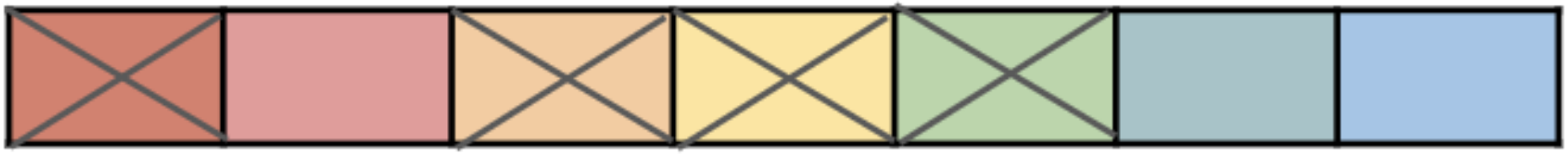


1) Retrieved from https://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/

Feature Subset Selection

All Features

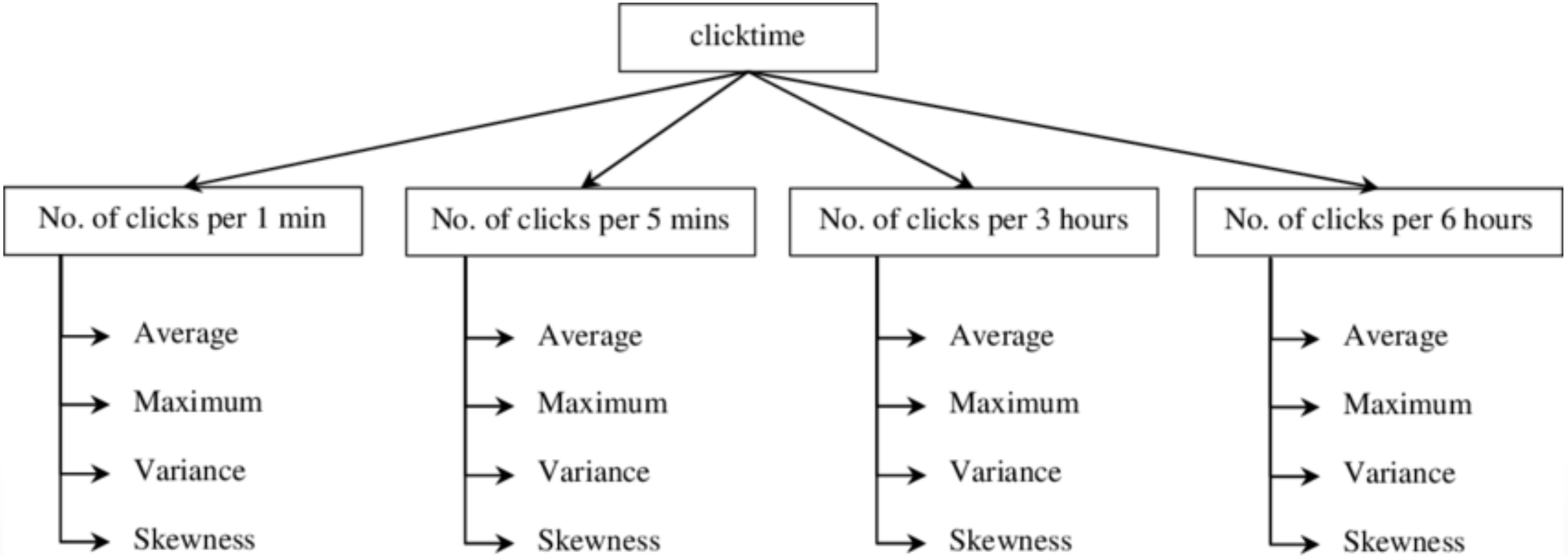
Feature Selection



Final Features

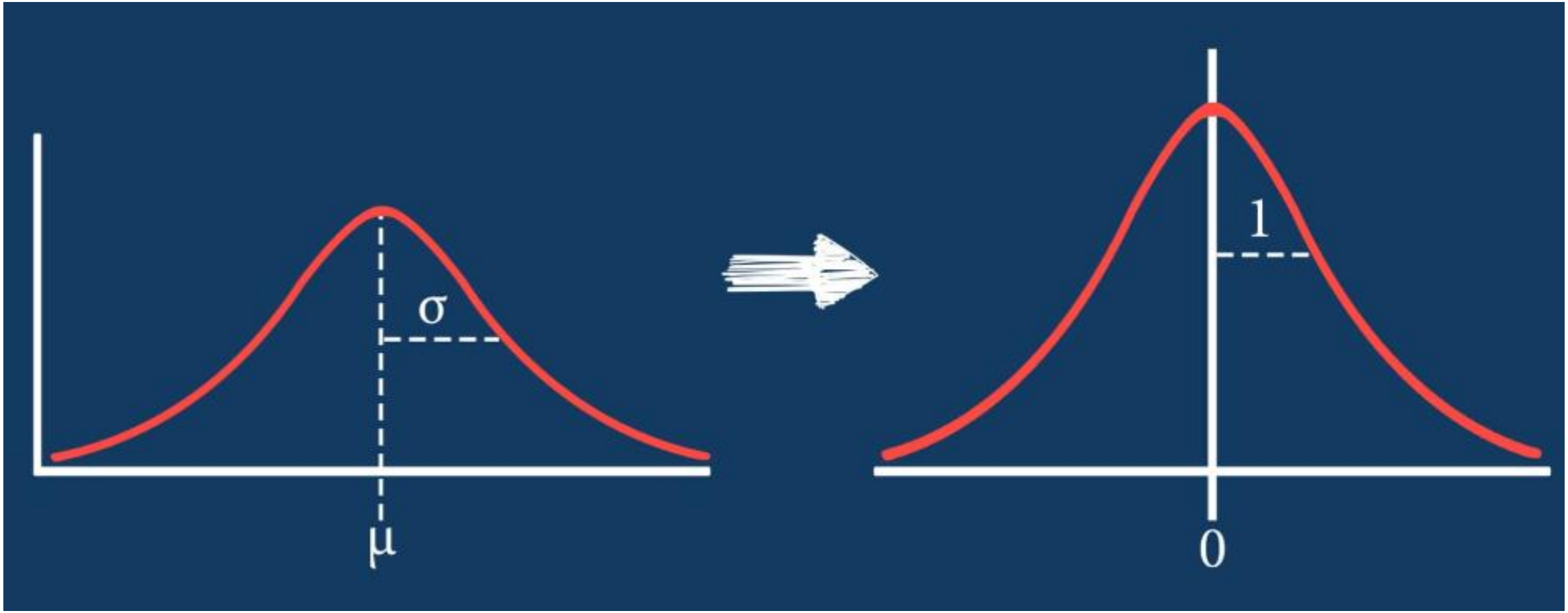
1) Retrieved from https://www.kdnuggets.com/2019/10/feature-selection-beyond-feature-importance.html

Feature Creation



1) Retrieved from https://www.researchgate.net/figure/Feature-creation-from-the-clicktime-attribute_fig5_262337335

Attribute Transformation



1) Retrieved from https://towardsdatascience.com/normalization-vs-standardization-quantitative-analysis-a91e8a79cebf

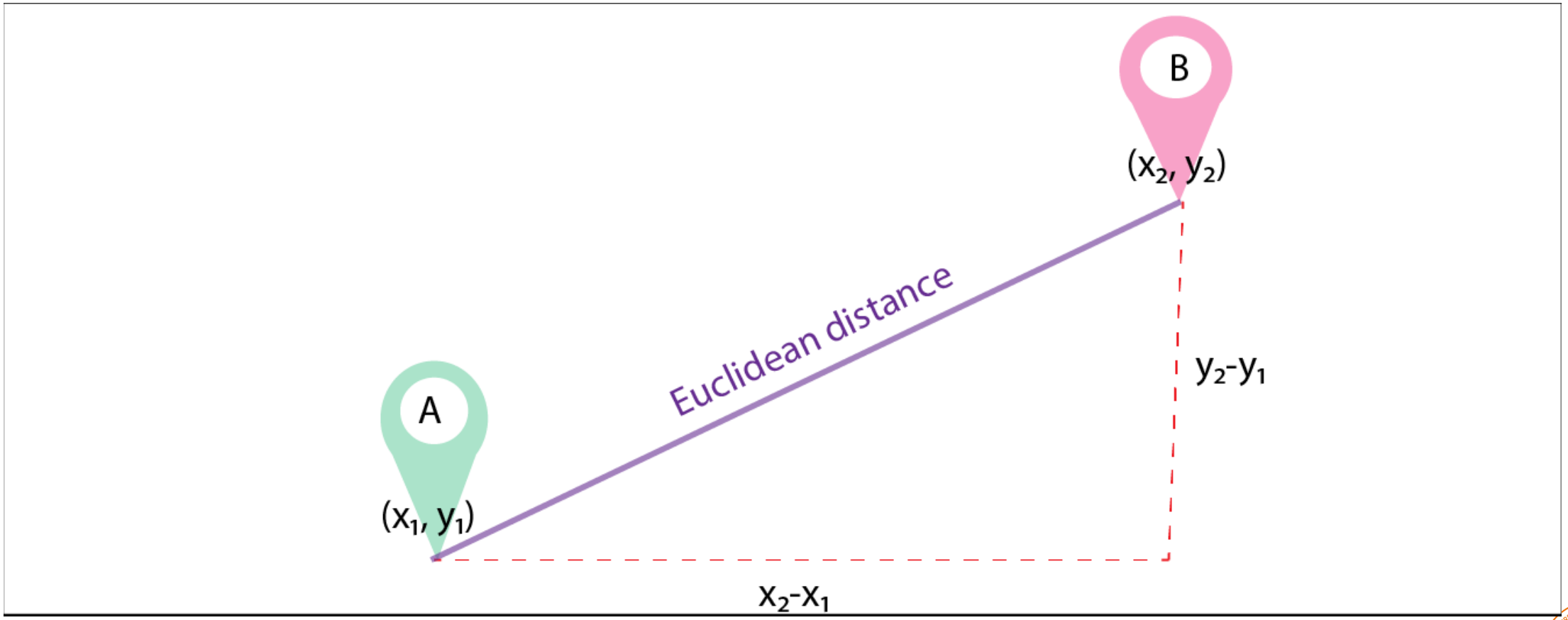
Calculating Similarity

Similarity/Dissimilarity

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x - y \rvert / (n-1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Euclidean Distance

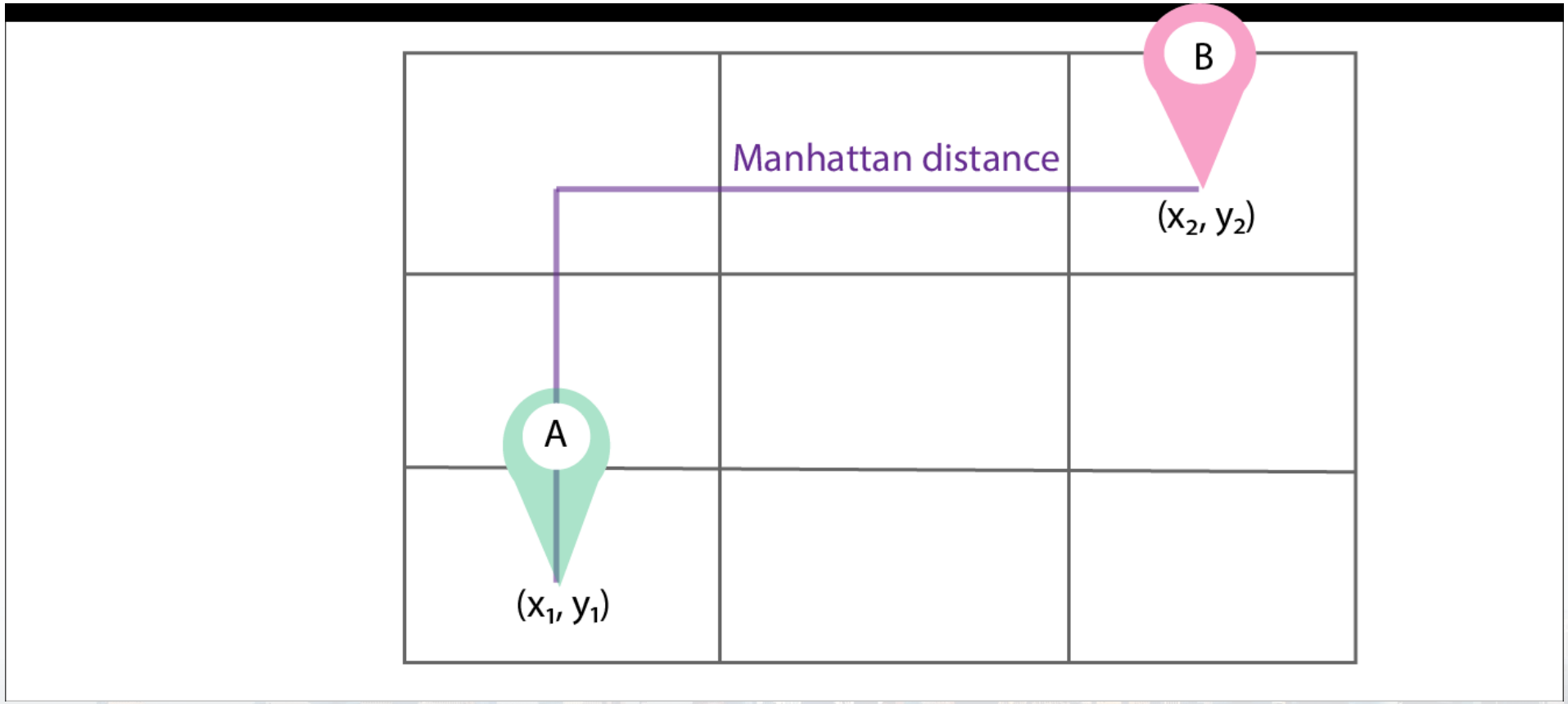
- Euclidean Distance represents the shortest distance between two points.



1) Retrieved from https://medium.com/analytics-vidhya/role-of-distance-metrics-in-machine-learning-e43391a6bf2e

Manhattan Distance

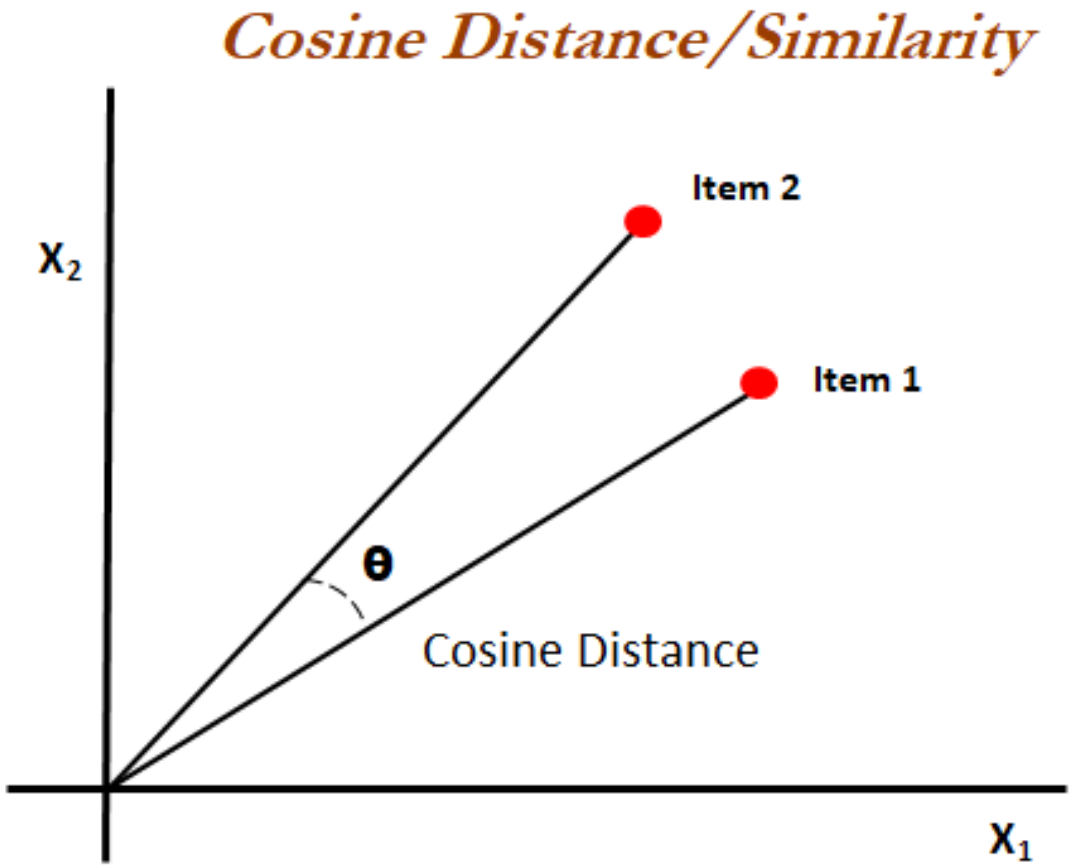
- Manhattan Distance is the sum of absolute differences between points across all the dimensions.



1) Retrieved from https://medium.com/analytics-vidhya/role-of-distance-metrics-in-machine-learning-e43391a6bf2e

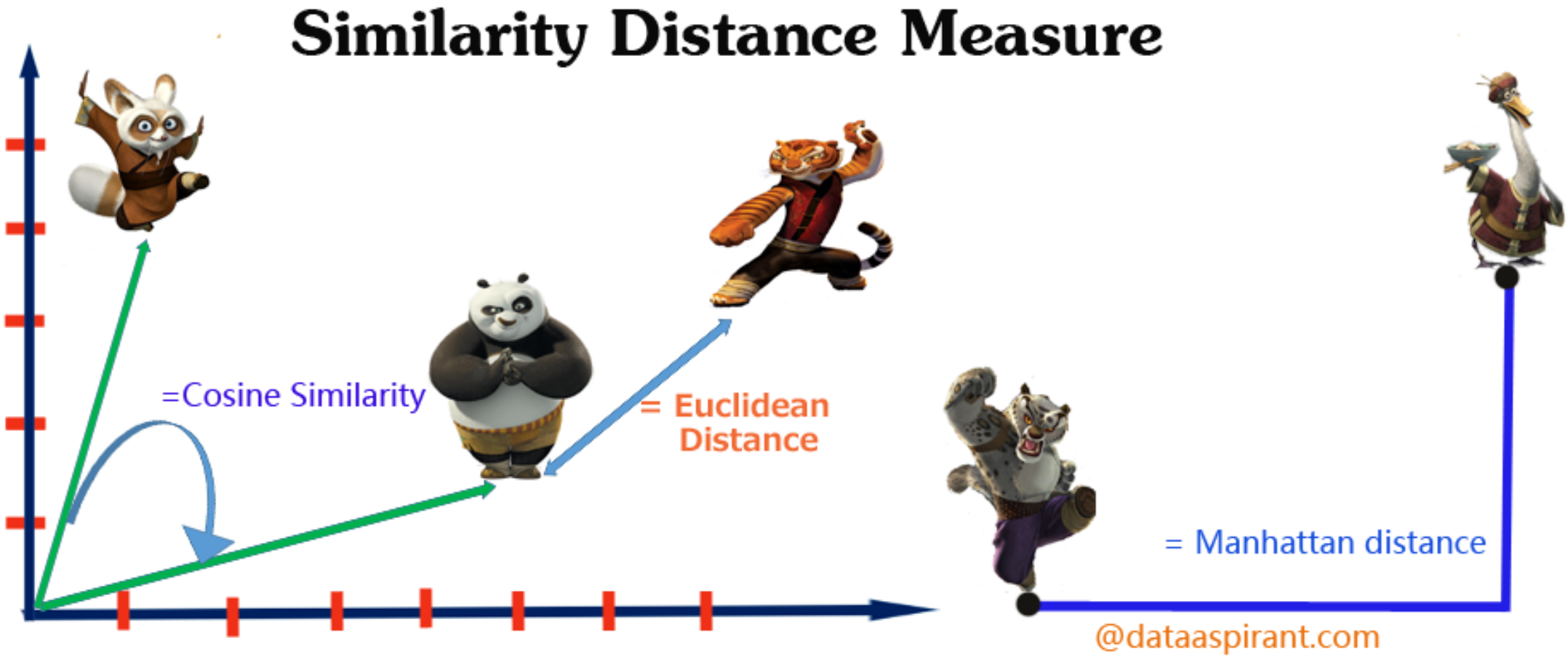
Cosine Similarity

- Consider similarity based on the angle between the two points



Statistics for Machine Learning by Pratap Dangeti

Summary of Distance Calculation



1) Retrieved from https://dataaspirant.com/five-most-popular-similarity-measures-implementation-in-python/

End of Class 2