Data and Data Exploration

Xiaochun MAI

Shenzhen University

Outline

1. Data Attribute Types
2. Types of Data Sets
3. Characteristics of Structured Data
4. Data Preprocessing

Data Attribute Types

- Collection of **data objects** and their **attributes**
- An **attribute** is a property/characteristic of an object
  - Examples: eye color of a person, temperature, heart beat, blood pressure, cholesterol etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an **object**
  - **Object** is also known as **record**, **observation**, **point**, **case**, **example**, **sample**, **entity**, or **instance**

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Data Attribute Types

- **Attribute values** are *numbers* (numerical) or *symbols/*strings (categorical) assigned to an attribute
- Distinction between **attributes** and **attribute values**
  - 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 attribute values
    - Example: Students' grades for different subjects are same. Attribute values for ID and age are integers
    - But properties of attribute values can be different
      - ID has no limit, but age has a maximum and minimum value

Types of Attributes

- There are 4 different types of attributes (more detailed)
  - **Nominal**: describe *qualitative* aspects of an object (can distinguish)
    - Barely enough to tell one object from another
    - Examples: ID numbers, eye color, zip codes
  - **Ordinal**
    - **"Order"** has meaning (can compare better/higher or worse/lower)
    - Examples: rankings (e.g., credit risk ratings {B-, B, B+, A, AA, …}, taste of potato chips on a scale from 1-10), grades {A,B,C,D}, height in {tall, medium, short}
  - **Interval**
    - **"Difference"** has meaning (can compare and conduct +, -)
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - **Ratio**
    - **"Ratio"** has meaning (can compare, conduct +, -, * /)
    - Examples: length, time, counts, temperature in Kelvin
    - Can't compare Aug 20 and 27 to get a ratio

https://www.youtube.com/watch?v=hZxnzfnt5v8

Properties of Attribute Values

- The *properties* of attribute values
  1. Distinctness: $= \ \neq$
  2. Order: $< \ >$
  3. Addition: $+ \ -$
  4. Multiplication: $* \ /$
- The type of an attribute has what properties?
  – Nominal attribute: distinctness
  – Ordinal attribute: distinctness, order
  – Interval attribute: distinctness, order, addition
  – Ratio attribute: **all 4 properties**

Let us take a look at more examples

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, $\neq$)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	**mode**, entropy, contingency correlation, χ^2 test
	Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, **street numbers**	**median**, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative	Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	**mean**, standard deviation, *t* and *F* tests, Pearson's correlation,
	Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in **Kelvin**, monetary quantities, counts, age, mass, length, electrical current	**geometric mean**, harmonic mean, percent variation

Mode of a set of data values is the value that appears most often

Median is the value separating the higher half from the lower half of a data sample

Mean is the central value of a discrete set of numbers

Geometric mean is defined as the *n*-th root of the product of *n* numbers

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite (or countably infinite) set of values
  - Examples: zip codes, or the set of words in a collection of documents
  - Often represented as *integer* variables (some algos can only take numbers).
  - Binary attributes are a special case of discrete attributes

- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Continuous attributes are typically represented as *floating-point* variables.

Outline

1. Data Attribute types
2. Types of Data Sets
3. Characteristics of Structured Data
4. Data Preprocessing

Types of Data Sets

- **Record**
  - Data Matrix
  - Document Data
  - Transaction Data
- **Graph**
  - World Wide Web
  - Molecular Structures
- **Ordered**
  - Spatial Data
  - Sequential Data
  - Sequence Data

Record Data

- Data that consists of a collection of records, each of which consists of a *fixed* set of attributes

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Record Data - Data Matrix

- If data **objects** have the same fixed set of **numeric attributes**, then the data objects can be thought of as **points** in a multi-dimensional space, where each dimension represents a distinct attribute.
- Such data set can be represented by an ***m*** by ***n*** matrix, where there are ***m*** rows (one for each object), and ***n*** columns (one for each attribute)

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

$$\begin{pmatrix} 10.23 & 5.27 & 15.22 & 27 & 1.2 \\ 12.65 & 6.25 & 16.22 & 22 & 1.1 \end{pmatrix}$$

Record Data - Document Data

- Also called text file
- Example

  "A text file is a kind of computer file that is structured as a sequence of lines of electronic text. A text file exists within a computer file system. The end of a text file is often denoted by placing one or more special characters, known as an end-of-file marker, after the last line in a text file"
- We typically want to *represent each document or text file into a feature vector or **"term"** vector*, in many applications such as document clustering, classification etc.

Record Data-Document Data & TF Representation

- Each document becomes a **"term"** vector,
  - Each term (mostly word, sometimes can be phrases in *n*-gram) is a component (attribute) of the vector,
  - **TF (Term Frequency based representation)**: The value of each component is the number of times the corresponding term (word) occurs in the document (Bag of words, ignoring the sequences), e.g.

document1 “analytics is what analytics is”
document2 “what is analytics”
document3 “analytics is a tool”

	"a"	"tool"	"is"	"analytics"	"what"
document1	0	0	2	2	1
document2	0	0	1	1	1
document3	1	1	1	1	0

TF representation

The number of all terms in data decides vector dimension, i.e. 5d. What if we are given huge corpus?

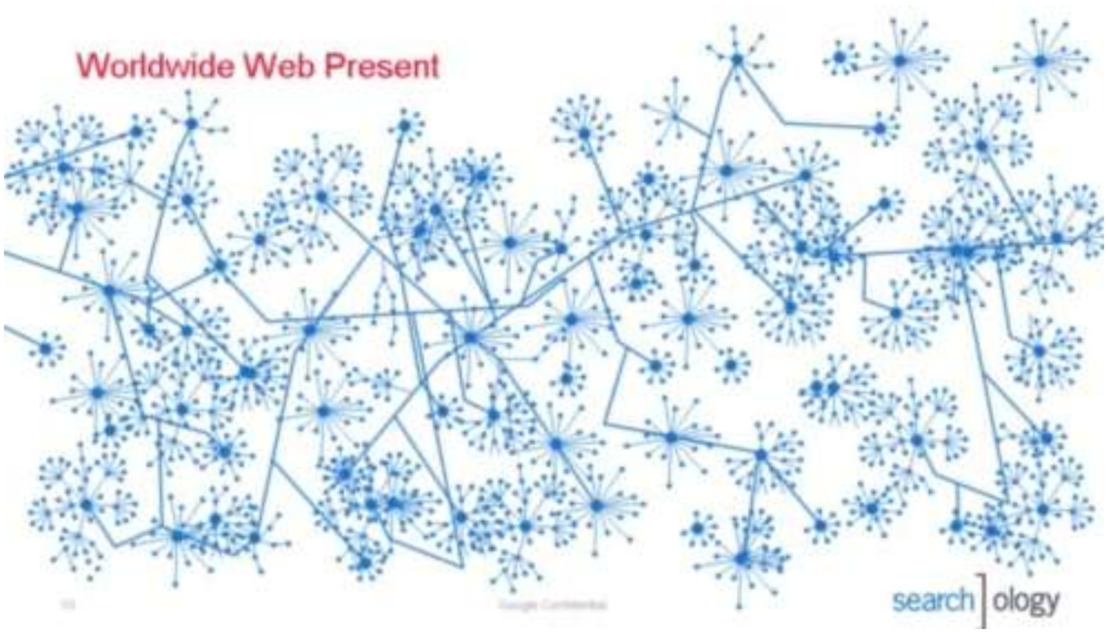
Record Data -Transaction Data

- A special type of record data, where
  - Each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

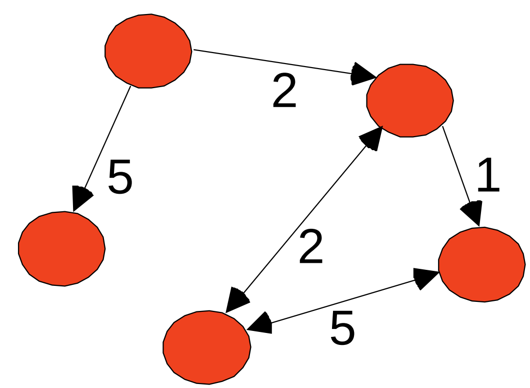
TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Graph Data – Web Link Data

- Example:
  - Generic WWW graph and HTML Links



The WWW Graph by Google's View



Ordered Data

- Sequences of transactions

Items/Events

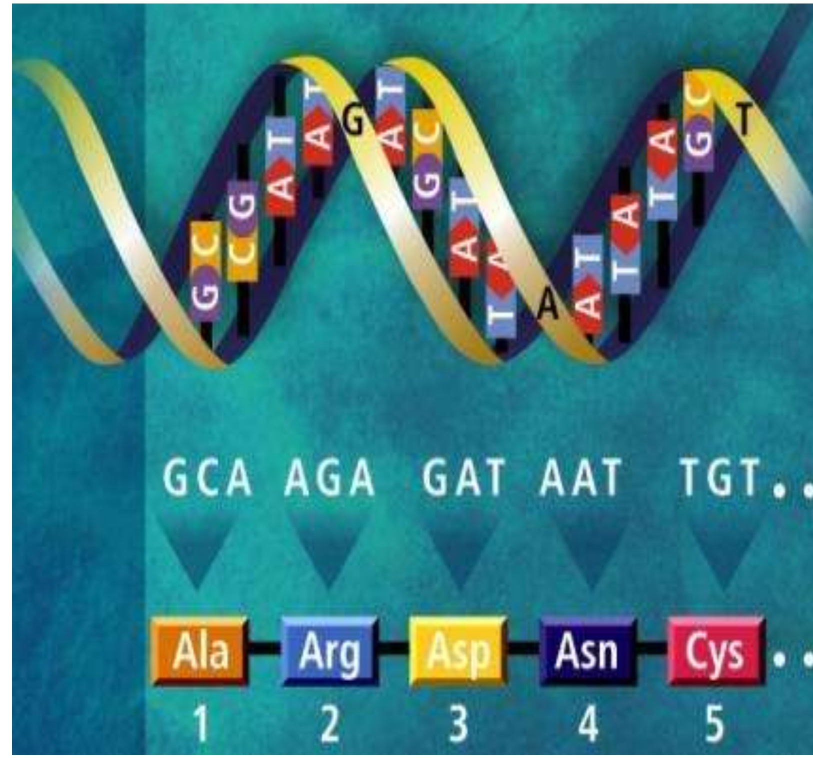
(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

An element of the sequence

Ordered Data

- Genomic sequence data

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

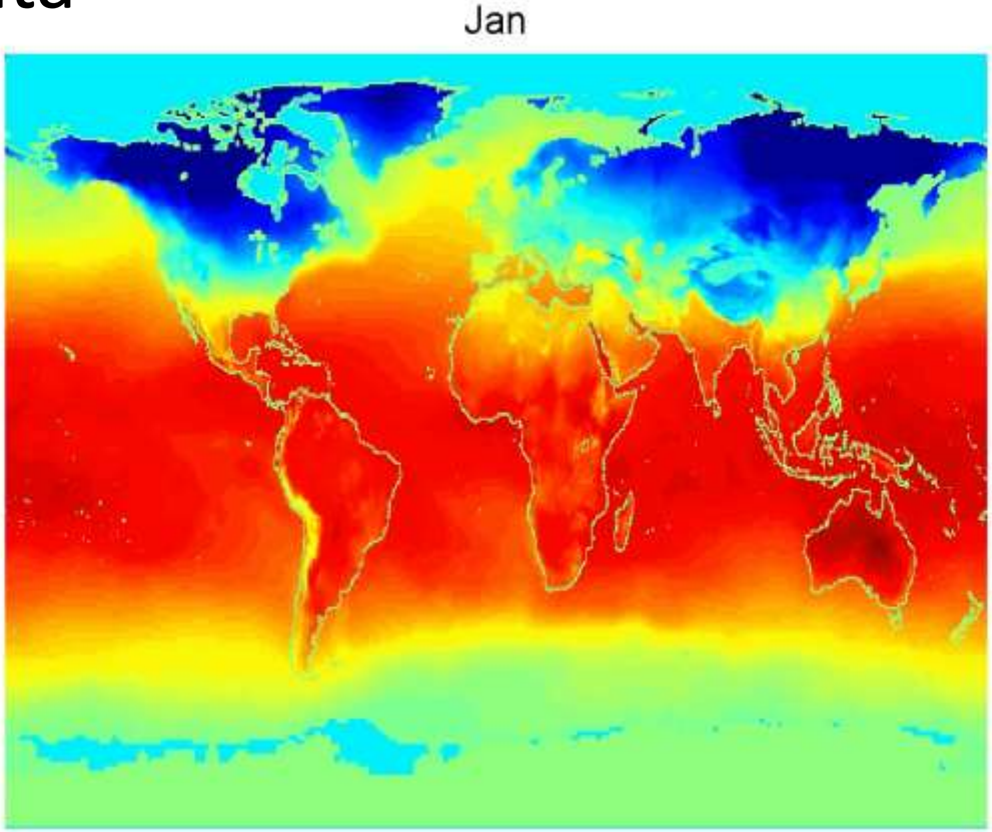


A,C,G,T stands for the four nucleic acids that make up DNA, a creature's genetic code. We can compare two person’s DNA to identify biological father/mother, identify criminals, discover disease genes, and compute the likelihood to get a specific disease

Ordered Data

- Spatio-Temporal Data

Average Monthly Temperature of land and ocean



Earth science data sets that record the temperature or pressure measured at points (grid cells) on latitude-longitude spherical grids of various resolutions (also at different time period)

Outline

1. Data Attribute Types
2. Types of Data Sets
3. Characteristics of Structured Data
4. Data Preprocessing

Key Characteristics of Structured Data or Tables

- **Dimensionality**
  - The number of dimensions/attributes: low or high
  - Curse of Dimensionality [from *wiki*]: when the *dimensionality increases*, the volume of the space increases so fast that the *available data become sparse*.
  - In order to obtain a statistically sound and reliable result, the amount of data needed to support the result often grows exponentially with the dimensionality. In high dimensional data, all objects appear to be sparse and dissimilar in many ways.

Key Characteristics of Structured Data or Tables

- **Sparsity**
  - Loosely distributed in the space
  - The number of non-zero values: sparse or dense
  - In many ML algorithms, only non-zero values need to be stored and manipulated
- **Resolution**
  - Data can be collected at different levels of resolution (sensor data collected in different sampling rates)
  - Properties of data differ at different resolutions
  - Patterns depend on the levels of resolution, e.g., surface of earth seems very uneven at a resolution of a few meters, but is relatively smooth at a resolution of tens of kilometers.

Outline

1. Data Attribute types
2. Types of Data Sets
3. Characteristics of Structured Data
4. Data Preprocessing

Data Preprocessing

1) Data Quality Issues
2) Data Preprocessing to address quality issues
    1) Data Cleaning
    2) Aggregation
    3) Sampling
    4) Dimensionality Reduction
    5) Feature Subset Selection
    6) Feature Generation/Creation
    7) Discretization and Binarization
    8) Attribute Transformation

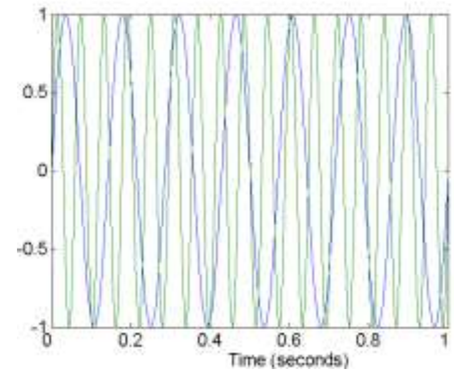
Data Quality Issues

- What **kinds** of data quality problems do we have?
  - Examples of data quality problems:
    - Noise
    - Outliers
    - Missing values
    - Duplicate data
- How can we **detect problems** with the data?
- What can we **do** about these problems?

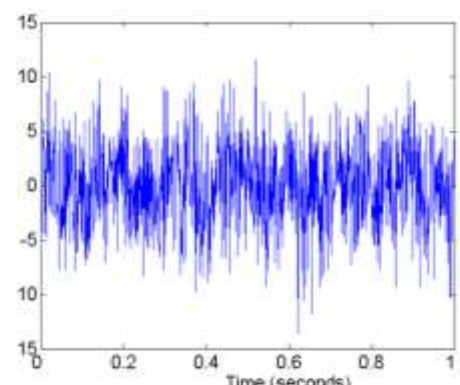
Data Quality Issues - Noise

- Noise refers to **modification** of original values
  - Examples: distortion of a person’s voice when talking on a poor phone

- What causes noise?
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention



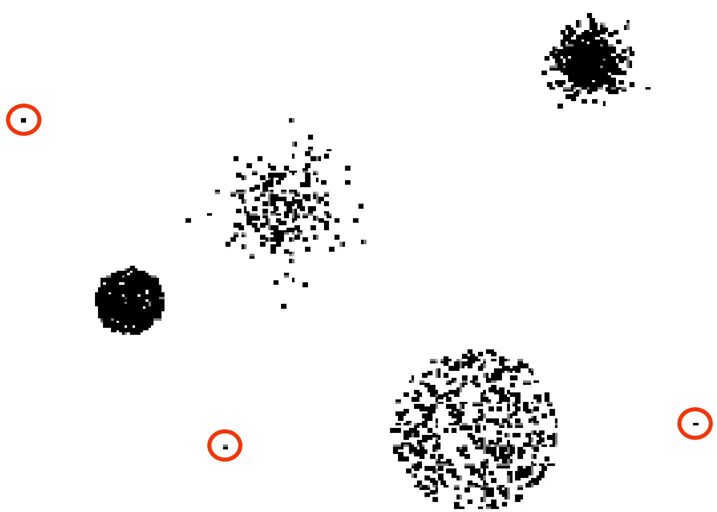
Two Sine Waves



Two Sine Waves + Noise

Data Quality Issues - Outliers

- Outliers are data objects with characteristics that are **considerably different** than **most** of the other data objects in the data set



Should we remove the outliers?

Data Quality Issues - Missing Values

- **Data values are not always available**
  - E.g., many tuples have no recorded value for **several attributes**, such as customer income in sales data
- **Missing data may be due to**
  - Equipment malfunction (faulty sensors)
  - Information was not collected (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
  - Data not entered due to misunderstanding
  - Certain data may not be considered important at the time of entry

Data Quality Issues - Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogonous sources
- Examples (email addresses are not unique IDs)
  - Same person with multiple email addresses
- Data de-duplication
  - Process of dealing with duplicate data issues
  - 1) Even if you know that there are two objects that actually represent a single object, values of the corresponding attributes may be different (some may be obsolete); these inconsistent values must be resolved.
  - 2) Avoid accidentally combining data objects that are similar, but not duplicates, e.g., two distinct people with identical names.

Data Preprocessing
1) Data Cleaning

- Importance
  - Data cleaning (cleansing) is the number1 problem in data preprocessing
- Data cleaning tasks (could be time-consuming)
  - Handle the missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

How to Handle Missing Data?

- **Eliminate data objects or attributes**
  - Usually done when **class label** (what you want to predict) is missing
  - Effective when a data set has only **a few** objects with missing values
  - When there are many objects with missing values, a reliable analysis can be difficult or impossible.
- **Fill in the missing value manually**
  - Tedious and infeasible for large data with many missing values

Name	Age	Sex	Education	Major	Income	# Years experience	Loan
James	34	M	Bachelor	CS	5k	?	1
Alex	36	M	Bachelor	CS	?	7	1
Bruce	40	M	PhD	EEE	12k	20	0
Mary	35	F	Msc	Fintech	20k	18	1
Judice	22	F	Msc	Analytics	4k	1	?

How to Handle Missing Data?

- **Fill in it automatically** with
  - A global constant
    - e.g., “unknown” or “NA” --- this may confuse the machine learning algorithm to think that these tuples are common or similar
  - Attribute mean (e.g. mean income of all objects)
  - Attribute mean for all samples belonging to the *same class* (e.g. certain diagnostic value for class disease vs class normal, mean of objects with loan=1)
  - Most probable value
    - Mostly commonly occurring attribute values
    - Inference-based such as Bayesian formula, decision tree or other classification/regression models (build a model to predict the missing values), e.g. predict missing income based on education, age, experience etc.

How to Handle Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Binning Methods for Data Smoothing

- ***Sort*** data for price (in dollars):

  4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* ***Partition*** into equal-frequency (equi-depth) bins:

- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

* Smoothing by bin means:

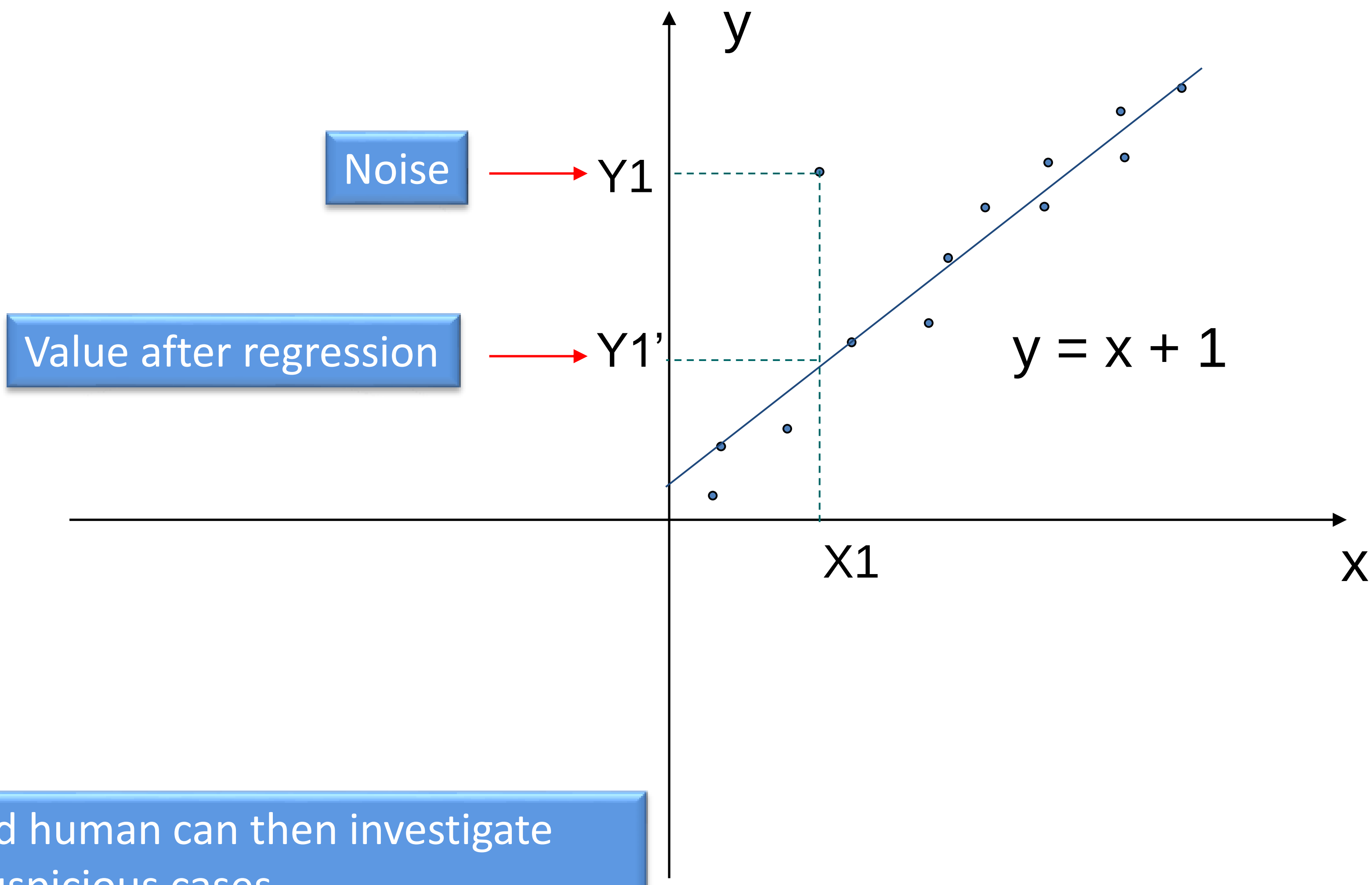
- Bin 1: 9, 9, 9, 9 (4+8+9+15)/4=9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

* Smoothing by bin boundaries (each value needs to check which boundary it nears to; or close boundary value):

- Bin 1: 4, **4, 4**, 15 [8 is near to 4, not 15; 9 is also same]
- Bin 2: 21, 21, **25**, 25
- Bin 3: 26, **26, 26**, 34

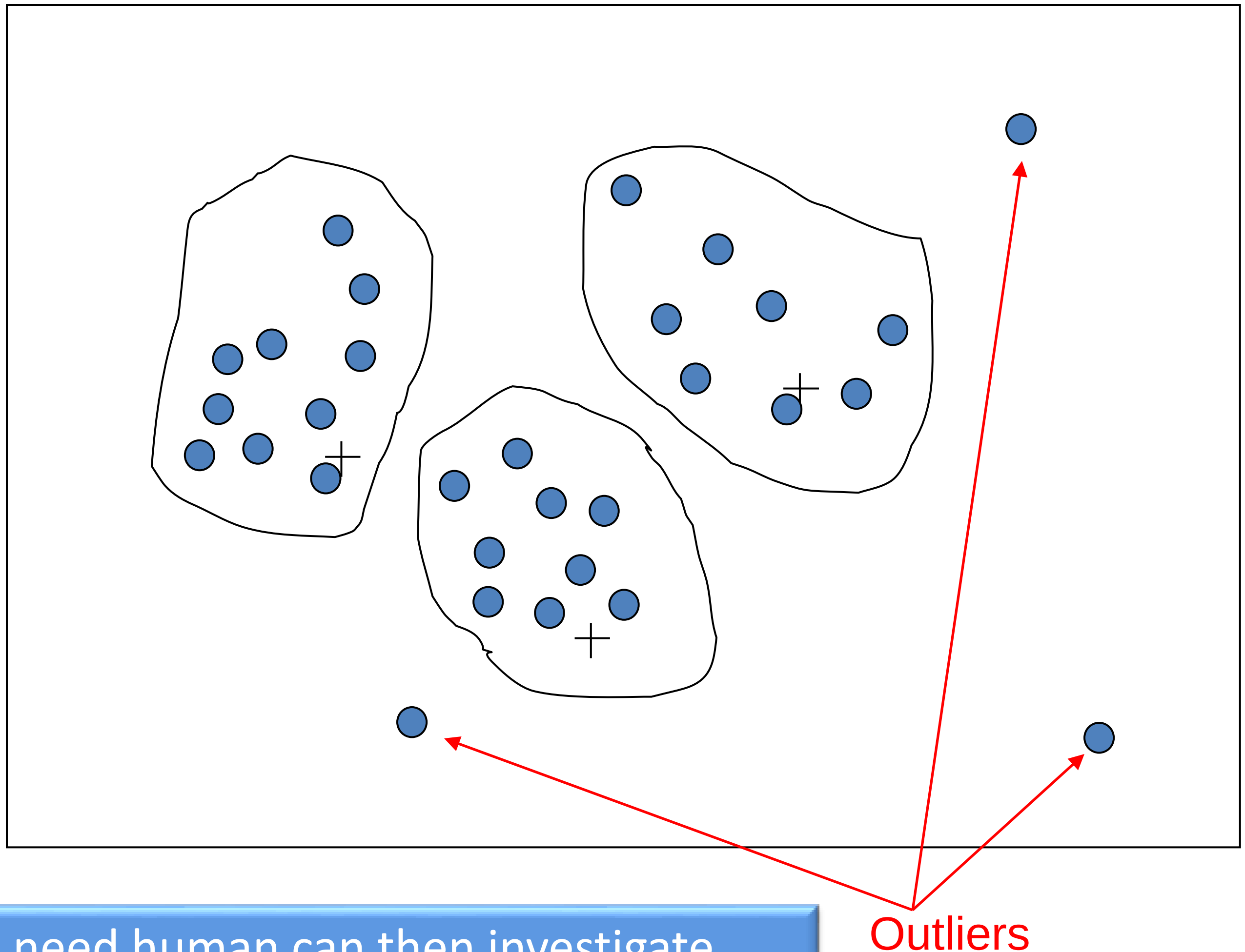
Bin boundaries preserve more information than the bin means

Regression



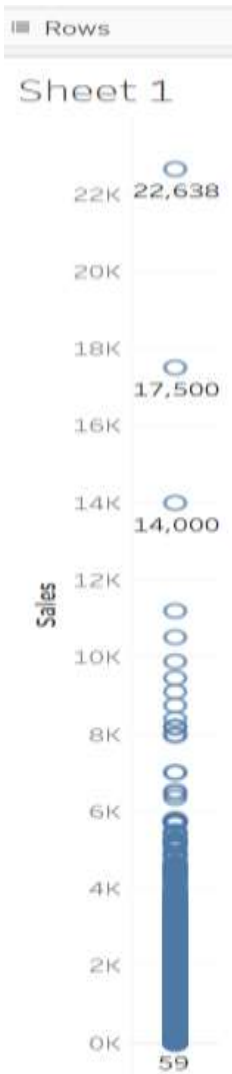
We need human can then investigate these suspicious cases.

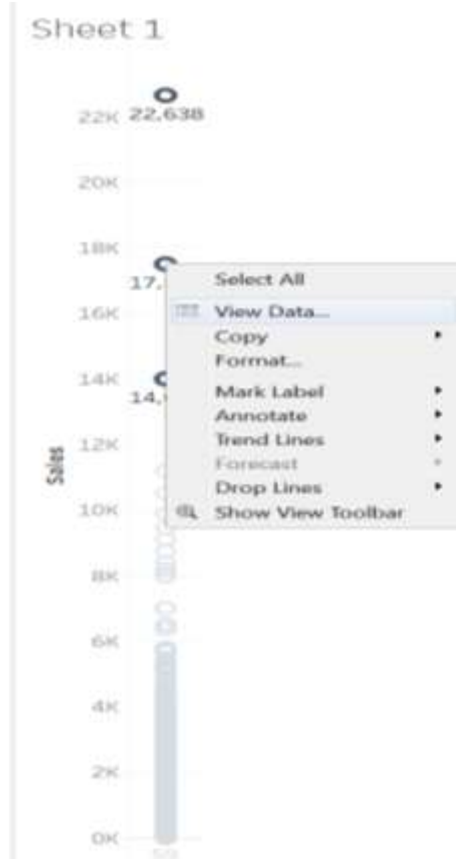
Cluster Analysis



We need human can then investigate these suspicious cases

Detect suspicious values and check by human (e.g., deal with possible outliers)





View Data: Sheet 1

3 rows | Show aliases | Show all fields | Copy | Export All

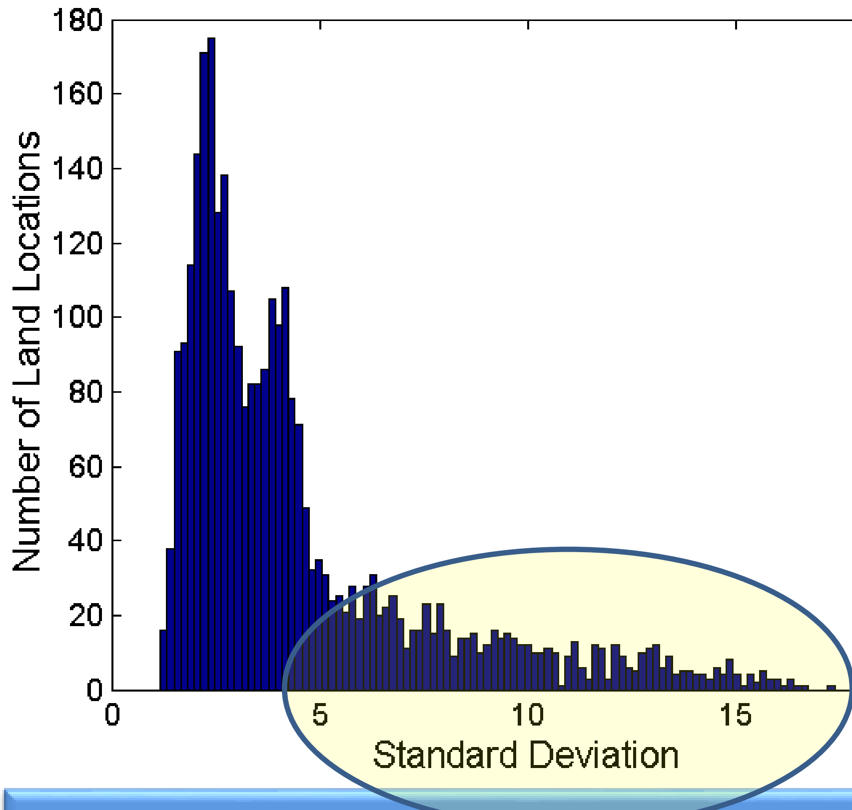
Customer Name	Market	Order Date	Order ID	Order Priority	Postal Code	Product ID	Product Name	Region	Row ID	Segment	Ship Date	Ship Mode	State	Sub-Category	Discount	Number of Records	Profit	Quantity	Sales	Shipping Cost
Tamara Chand	US	03-Oct-13	CA-2013-118689	Medium	47905	TEC-CO-10004722	Canon imageCLASS 2200 Advanced Copier	Central	38123	Corporate	10-Oct-13	Standard Class	Indiana	Copiers	0.000000	1	8,399.98	5	17,499.95	349.070
Sean Miller	US	18-Mar-11	CA-2011-145317	Medium	32216	TEC-MA-10002412	Cisco TelePresence System EX90 Videoconferencing Unit	South	33994	Home Office	23-Mar-11	Standard Class	Florida	Machines	0.500000	1	-1,811.08	6	22,638.48	24.287
Raymond Buch	US	24-Mar-14	CA-2014-140151	Medium	98115	TEC-CO-10004722	Canon imageCLASS 2200 Advanced Copier	West	39450	Consumer	26-Mar-14	First Class	Washington	Copiers	0.000000	1	6,719.98	4	13,999.96	20.001

2) Aggregation

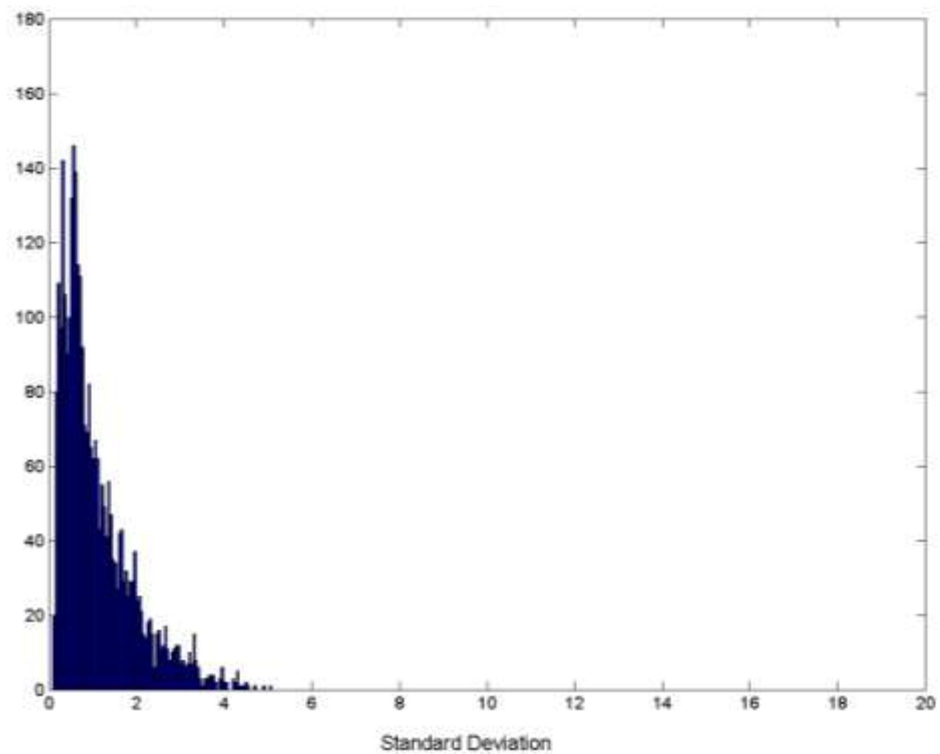
- Combining two or more attributes (or objects) into a single attribute (or object)
- Purposes
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc
  - More “stable” data
    - Aggregated data tends to have less variability

Aggregation

Example: Variation of Precipitation in Australia



Standard Deviation of Average **Monthly** Precipitation



Standard Deviation of Average **Yearly** Precipitation

If we have large SD, then the data is not stable.

3) Sampling

- Sampling is the main technique employed for data selection.
  - often used for both the preliminary investigation of the data and the final data analysis.
- Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.
- Sampling is used in machine learning because processing the entire set of data of interest is too expensive or time consuming. When you are writing a program to perform preprocessing and model building, it is good to perform sampling to get a subset a data.

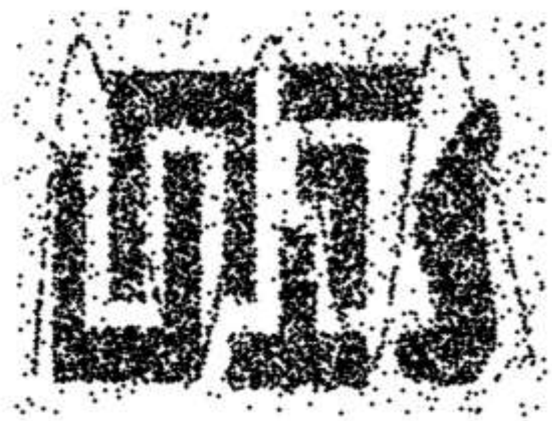
Sampling …

- The key principle for effective sampling:
  - Using a sample will work almost as well as using the entire data sets, if the sample is representative
  - A sample is representative if it has approximately the same property (of interest) as the original set of data.
  - Sample size vs representative

Types of Sampling

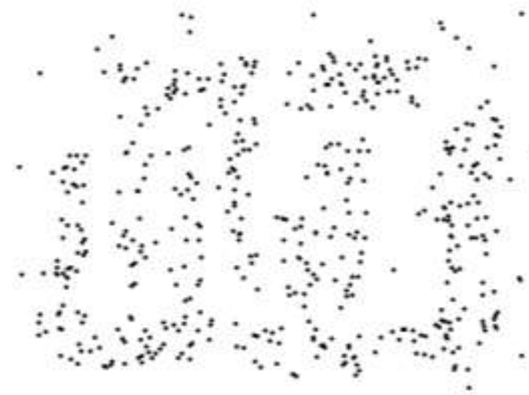
- **Simple Random Sampling**
  - There is an equal probability of selecting any particular item
- **Sampling without replacement**
  - Once an item is selected, it is removed from the population
- **Sampling with replacement**
  - Objects are not removed from the population when they are selected for the sample
  - The same object can be picked up more than once
- **Stratified sampling**
  - Split the data into several partitions; then draw random samples from each partition

Sample Size



8000 points

2000 Points

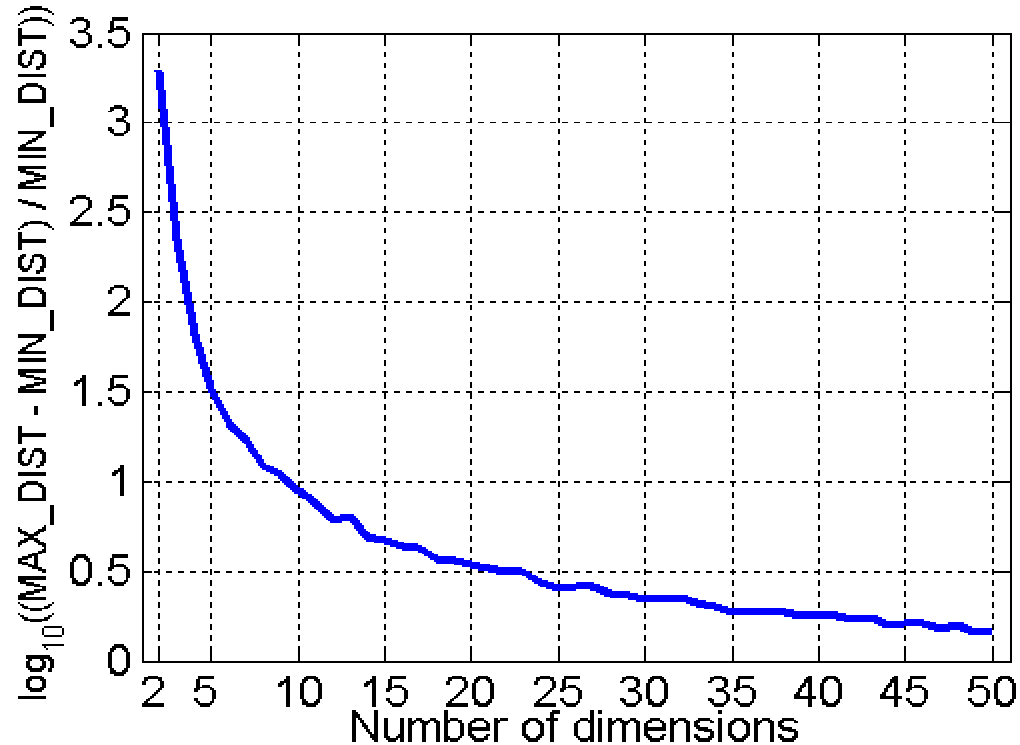


500 Points

Trade off “effectiveness” and “efficiency”

4) Curse of Dimensionality

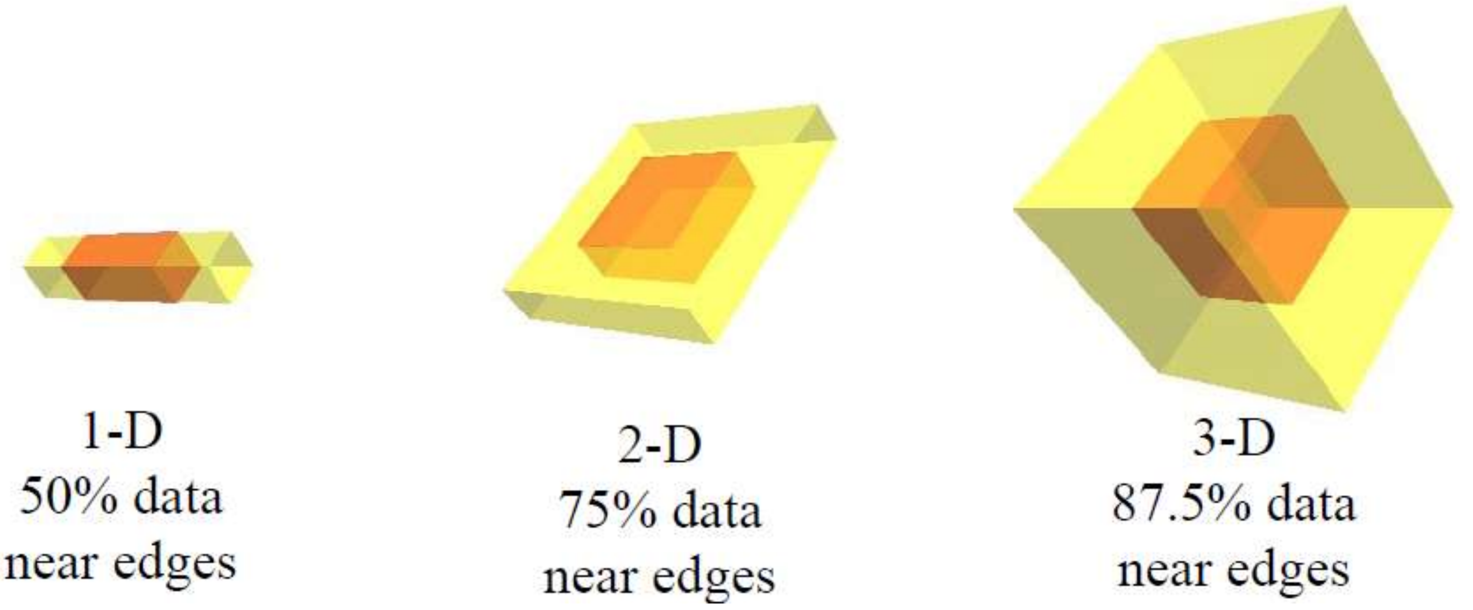
- When dimensionality increases, data becomes increasingly sparse in the space that it occupies.
- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful.



- Randomly generate 500 points
- Compute difference of *max* and *min* distance (normalized by the *min* distance), between any pair of points

Curse of Dimensionality

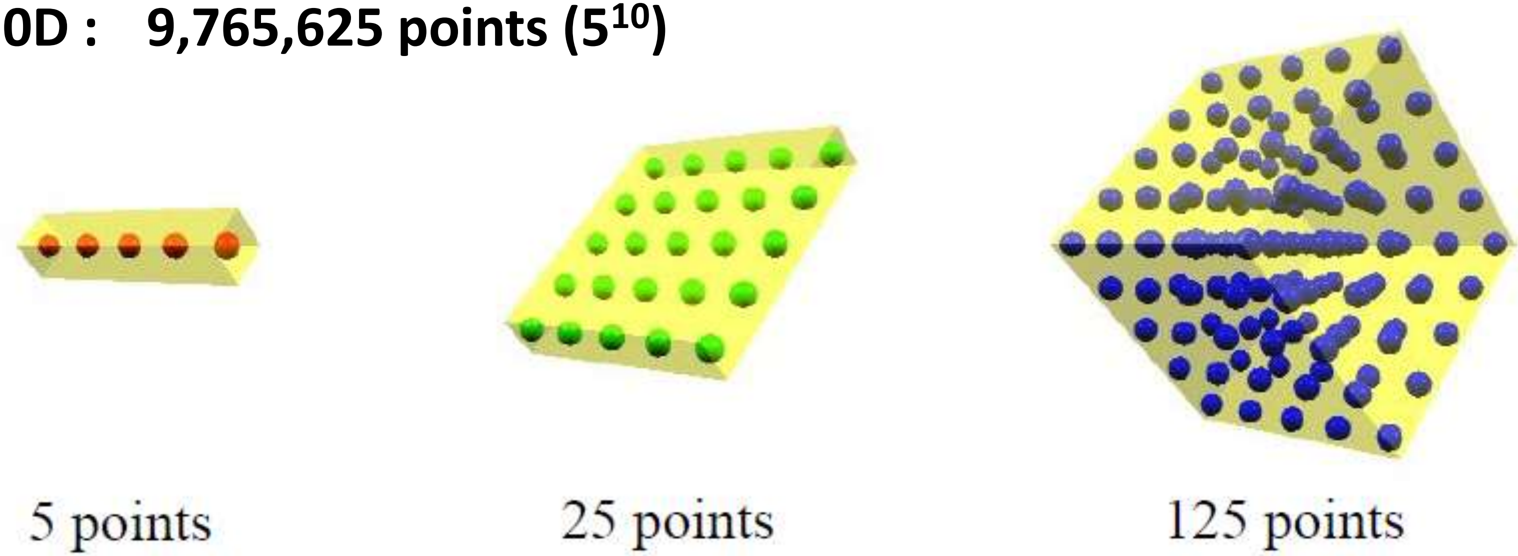
- Definitions of density and distance between points become less meaningful.
- In very high-Dimension, almost every point lies at the edge of the space, far away from the center.



Assume data points occupy *half* of each dimension (yellow)

Curse of Dimensionality

- One challenge of mining high-dimensional data is **insufficient data samples**
- **Suppose 5 samples/objects is considered enough in 1-D**
  - **1D : 5 points (5^1)**
  - **2D : 25 points (5^2)**
  - **3D : 125 points (5^3)**
  - **10D : 9,765,625 points (5^{10})**



4) Dimensionality Reduction Simply our data

- **Purposes:**
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by machine learning algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise
- **Techniques**
  - Principle Component Analysis (PCA).
  - Singular Value Decomposition (SVD)

Philosophy of PCA

- When a data set has too many variables that are **correlated**, how do you want to handle it?
- If we directly construct a model using all the correlated variables, then we could get low prediction results.
- We need to think some strategic method to **find few important uncorrelated variables** (in form of components) from a large set of original correlated variables available in a data set.
- PCA helps to overcome such challenges, which was by Pearson (1901) and Hotelling (1933).

https://www.youtube.com/watch?v=KO8VhTPmZPw

Philosophy of PCA

- Key idea is to extract low dimensional set of features from a high dimensional data set (>=3) with a motivation ***to capture as much information as possible***.
- Assume that a data set has dimension 1000 (*n*) × 100 (*p*), where *n* represents the number of observations/objects and *p* represents number of variables.
- One straightforward way to analyse the *correlation between variables* is to construct *scatter plots* for each pair of variable. Unfortunately, we will have *p*(*p*-1)/2 (499,500) variable pairs. It will be tedious job to perform exploratory analysis in such manner.

https://www.youtube.com/watch?v=HMOI_lkzW08

Philosophy of PCA

- Principal Component Analysis (PCA): Find a (linear) projection that

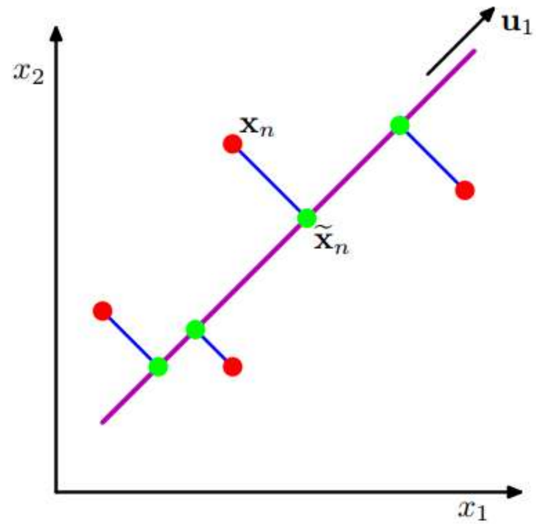
➢ Minimize reconstruction error (Pearson, 1901)

➢ Maximize the variance (signal) of the projected data

(Hotelling, 1933)

➢ Maximize the mutual information between original and projected data (Linsker, 1988)

Philosophy of PCA



From PRML (Bishop, 2006)

- Two-dimensional data $\boldsymbol{x} = [x_1, x_2]^\top$ projected onto a one-dimensional linear manifold (affine subspace) with direction $\boldsymbol{u}_1$.
- Red: Original data, Green: Projected data