Data Mining: Data

Lecture Notes for Chapter 2

Data Mining

https://ml-graph.github.io/winter-2025/

Yu Wang, Ph.D.
yuwang@uoregon.edu
Assistant Professor
Computer Science
University of Oregon

CS 453/553 – Winter 2025

Course Lecture is very heavily based on
“Introduction to Data Mining”
by Tan, Steinbach, Karpatne, Kumar

Outline

- Attributes and Objects
- Types of Data
- Data Quality
- Similarity and Distance
- Data Preprocessing

What is Data?

- Collection of *data objects* and their *attributes*
- An *attribute* is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describe an *object*
  - Object is also known as record, point, case, 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

Attribute Values

- ***Attribute values*** **are numbers or symbols assigned to an attribute for a particular object**

- **Distinction between attributes and attribute values**
  - **Same attribute -》 different attribute values**
    - **Example: height can be measured in feet or meters**

  - **Different attributes -》 same set of values**
    - **Example: Attribute values for ID and age are integers**

Types of Attributes

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio (Interval + 0 point)
    - Examples: length, counts, elapsed time (e.g., time to run a race)

Properties of Attribute Values

- **The type of an attribute depends on which of the following properties/operations it possesses:**
  - **Distinctness: $= \ \neq$**
  - **Order: $< \ >$**
  - **Differences are meaningful : $+ \ -$**
  - **Ratios are meaningful $* \ /$**

  - **Nominal attribute: distinctness**
  - **Ordinal attribute: distinctness & order**
  - **Interval attribute: distinctness, order & meaningful differences**
  - **Ratio attribute: all 4 properties/operations**

Difference Between Ratio and Interval

- **Is it physically meaningful to say that a temperature of 10 ° is twice that of 5° on**
  - **the Celsius scale (0°C: The freezing point of water at standard atmospheric pressure.)?**
  - **the Fahrenheit scale?**
  - **the Kelvin scale?**

- **Consider measuring the elapsed time for race**

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ2 test
	Ordinal	Ordinal attribute values also 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, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
	Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

This categorization of attributes is due to S. S. Stevens

	Attribute Type	Transformation	Comments
Categorical Qualitative	Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
	Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Numeric Quantitative	Interval	*new_value = a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
	Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

This categorization of attributes is due to S. S. Stevens

Any Question?

1. "Judge a man by his questions rather than by his answers."
   – Voltaire
2. "If I had an hour to solve a problem, I'd spend 55 minutes thinking about the problem and 5 minutes thinking about solutions."
   – Albert Einstein
3. "The art and science of asking questions is the source of all knowledge."
   – Thomas Berger
4. "Asking the right questions takes as much skill as giving the right answers."
   – Robert Half
5. "The wise man doesn't give the right answers, he poses the right questions."
   – Claude Lévi-Strauss
6. "Great questions make great companies."
   – Peter Drucker

Discrete and Continuous Attributes

- **Discrete Attribute**
  - Finite or countably infinite set of values
  - Counts, the set of words
  - Integer variables.
- **Continuous Attribute**
  - Real numbers
  - Examples: temperature, height, or weight.
  - Represented using a finite number of digits.
  - Floating-point variables.

Key Messages for Attribute Types

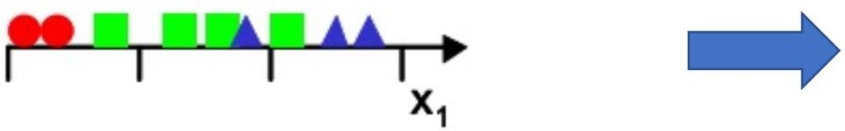
- **The types of operations <-> the type of data you have**
  - **Not only Distinctness, order, meaningful intervals, and meaningful ratios**
  - **Textual Strings – may not capture all the properties or may suggest properties that are not present**
  - **Statistical analyses depend only on the distribution**
  - **What is meaningful can be specific to problem**

Important Characteristics of Data

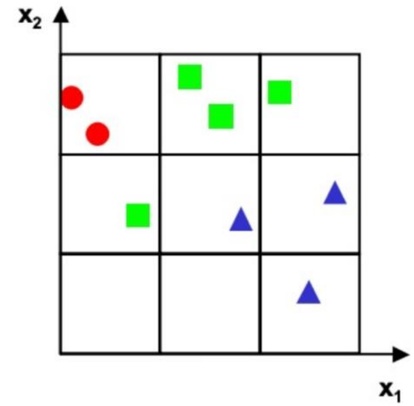
- Dimensionality (number of attributes)
  - High dimensional data brings a number of challenges – Curse of Dimensionality
- Sparsity – Recommender Systems
  - Only presence counts
- Resolution – Time-series Data
  - Patterns depend on the scale
- Size
  - Type of analysis may depend on size of data

Curse of Dimensionality

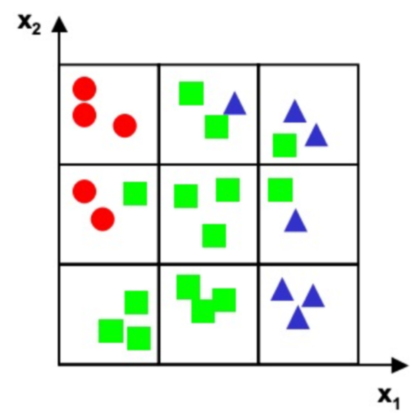
- We add a second feature.



Constant # examples



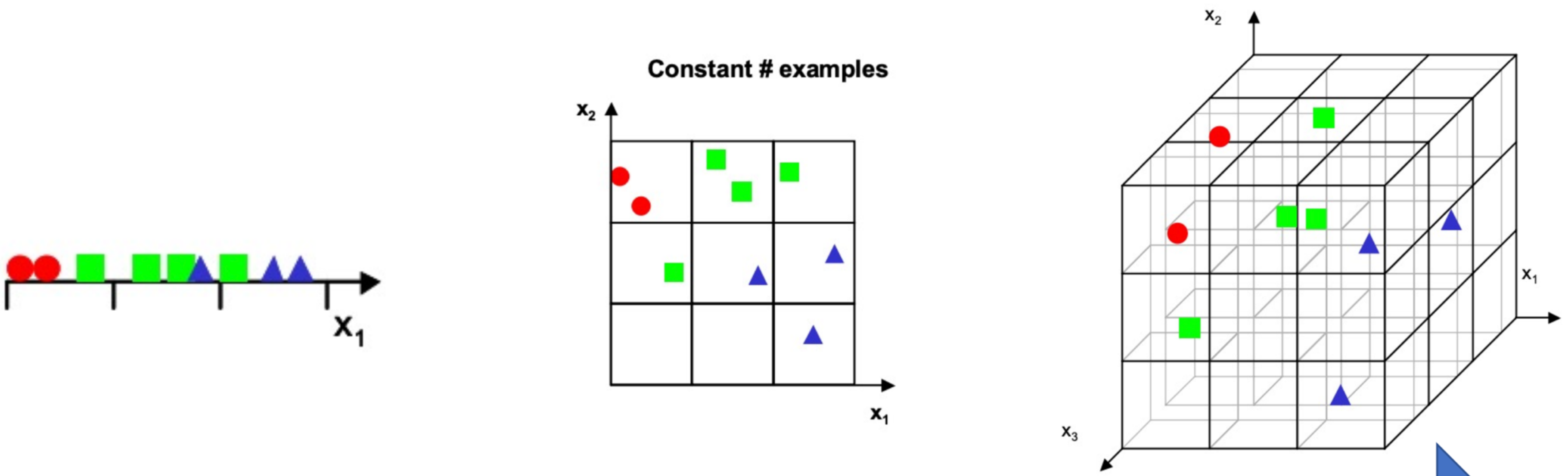
Constant density



- How many samples do we need if we wanted to keep the average density per segment constant?

Curse of Dimensionality

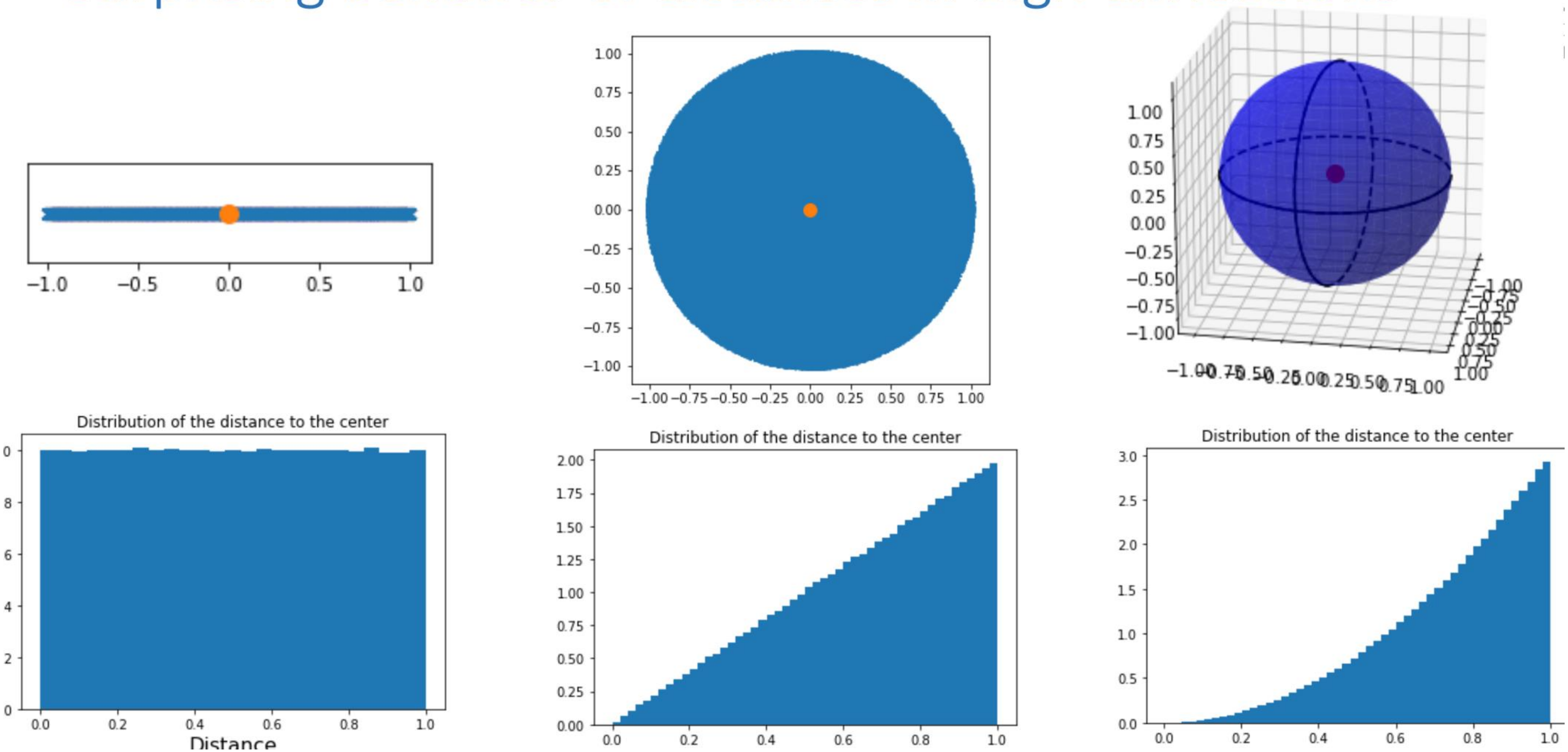
- Lets add a third feature:



The number of bins grows exponentially -> We need exponentially more samples

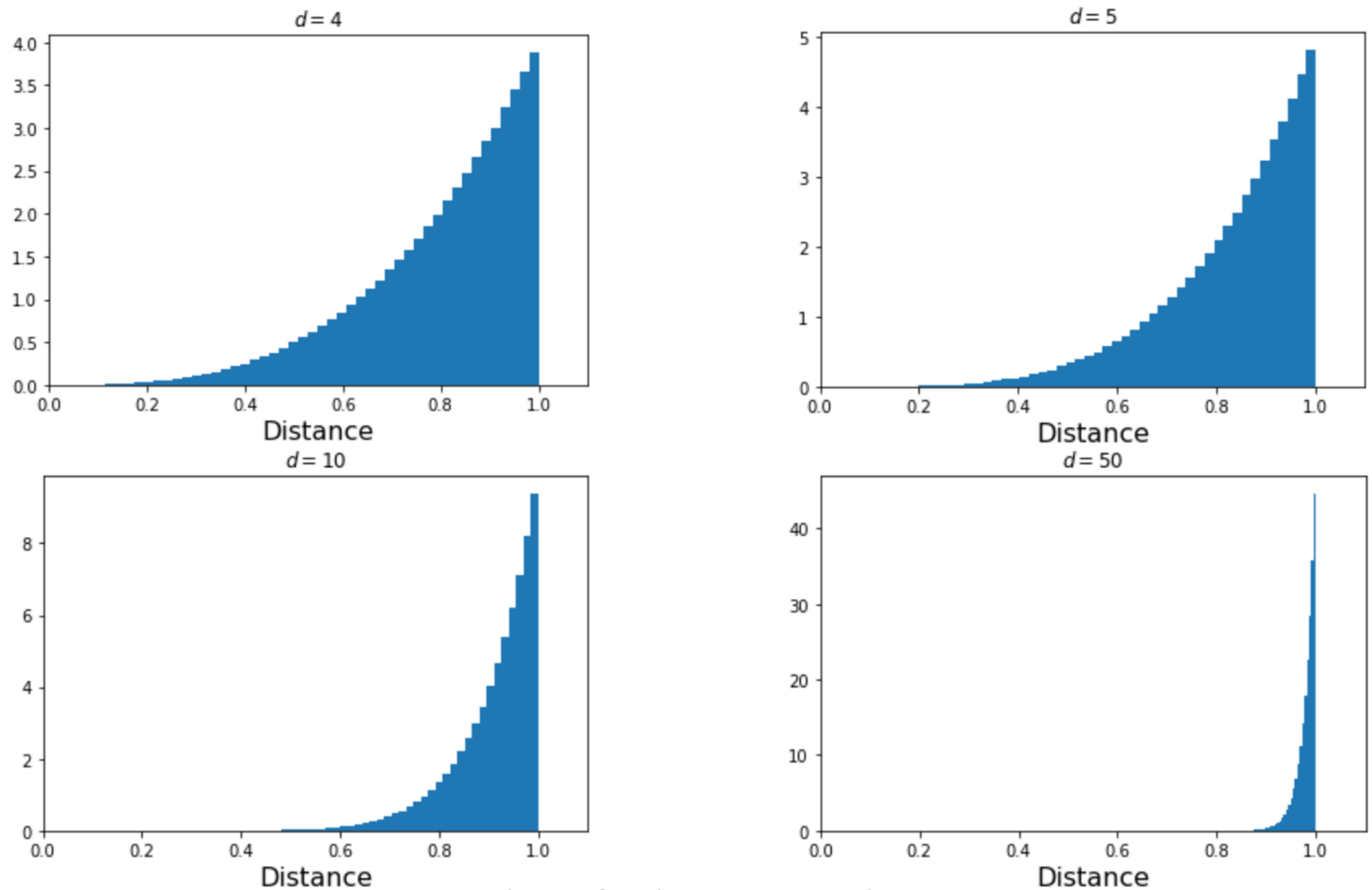
Curse of Dimensionality

Surprising behavior of distances in high dimensions



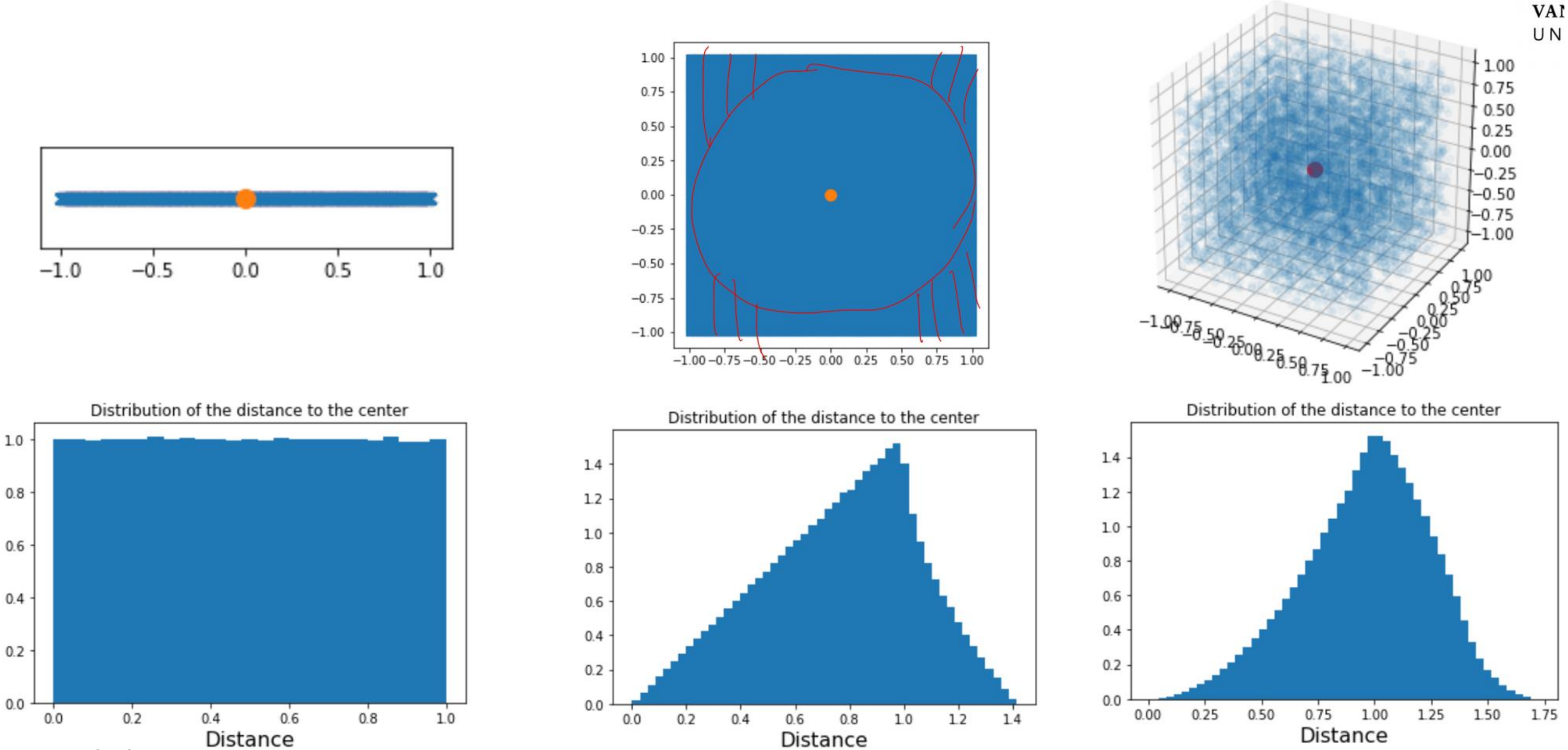
Curse of Dimensionality

Surprising behavior of distances in high dimensions



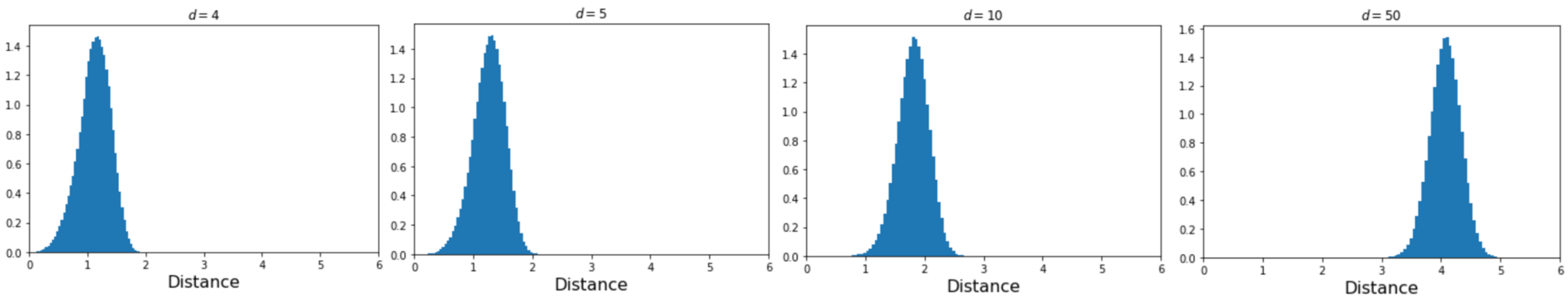
Curse of Dimensionality

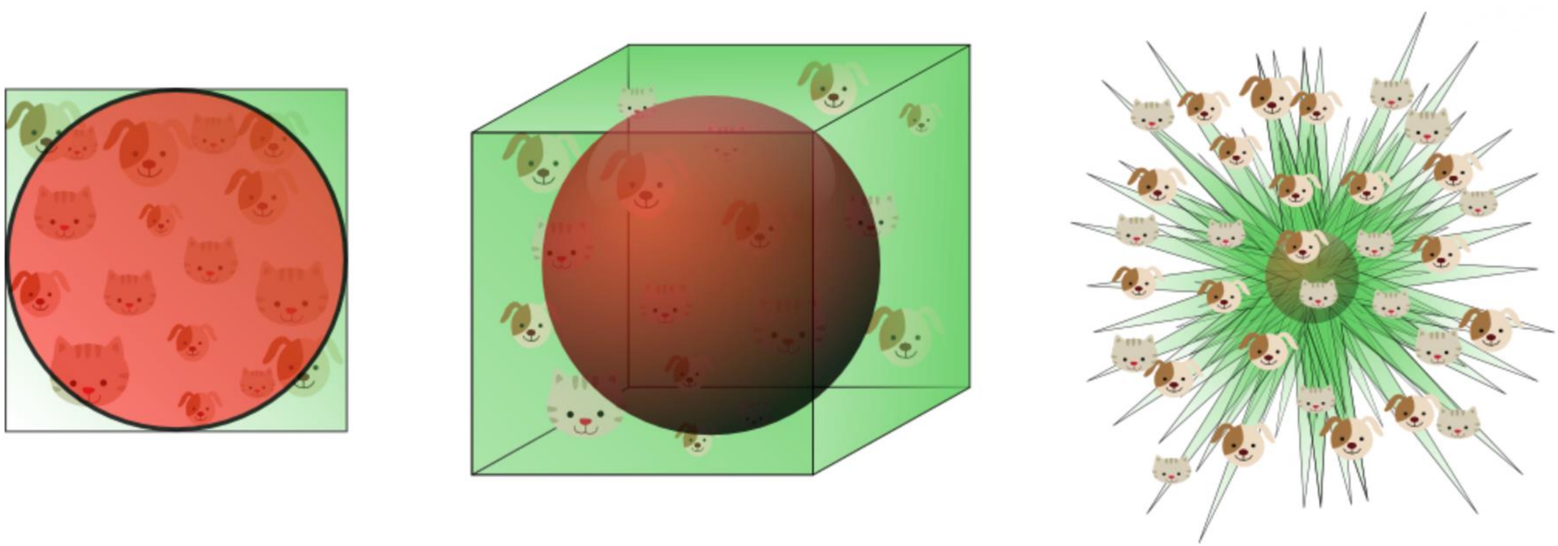
Surprising behavior of distances in high dimensions



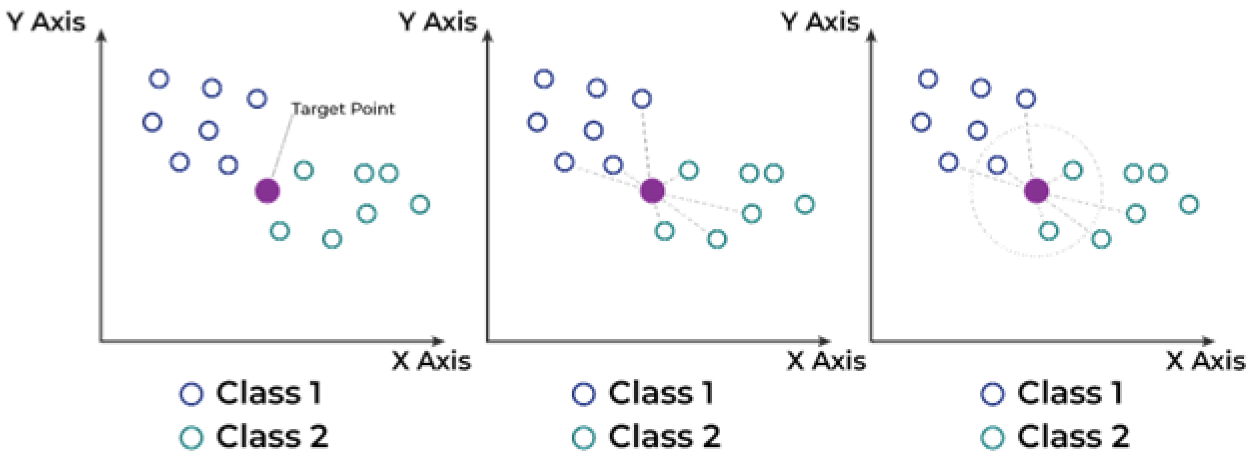
Curse of Dimensionality

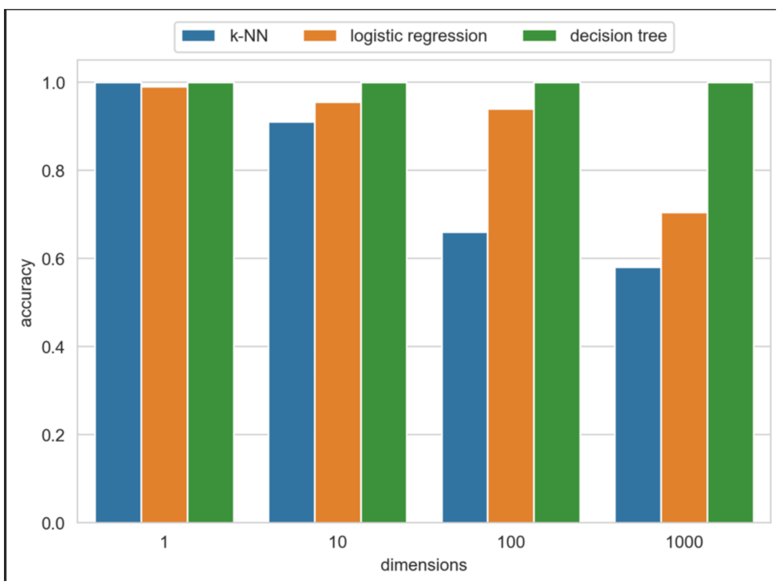
Distribution of distances of samples in a d-dimensional cube from the origin.

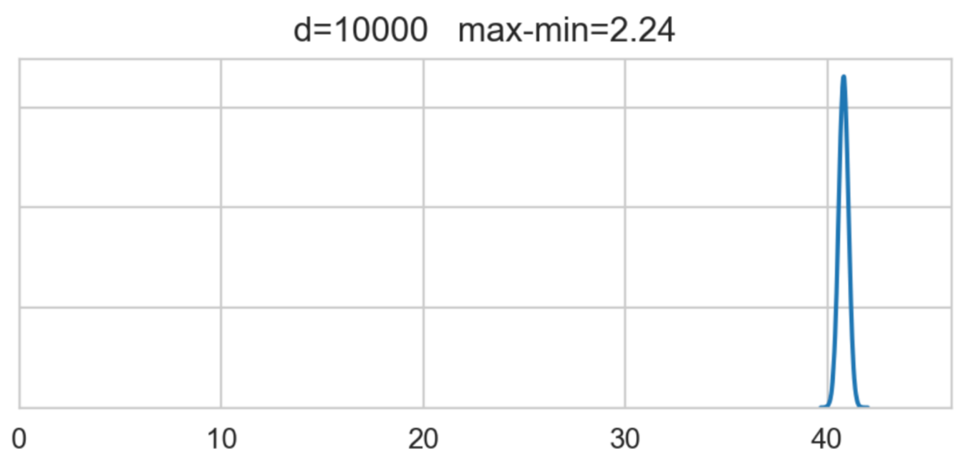




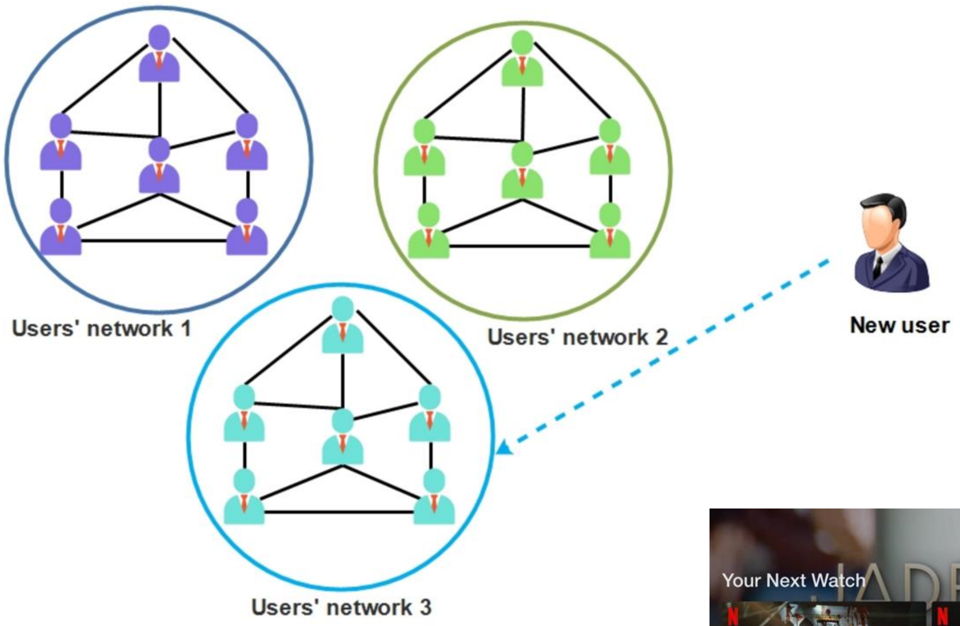
Curse of Dimensionality



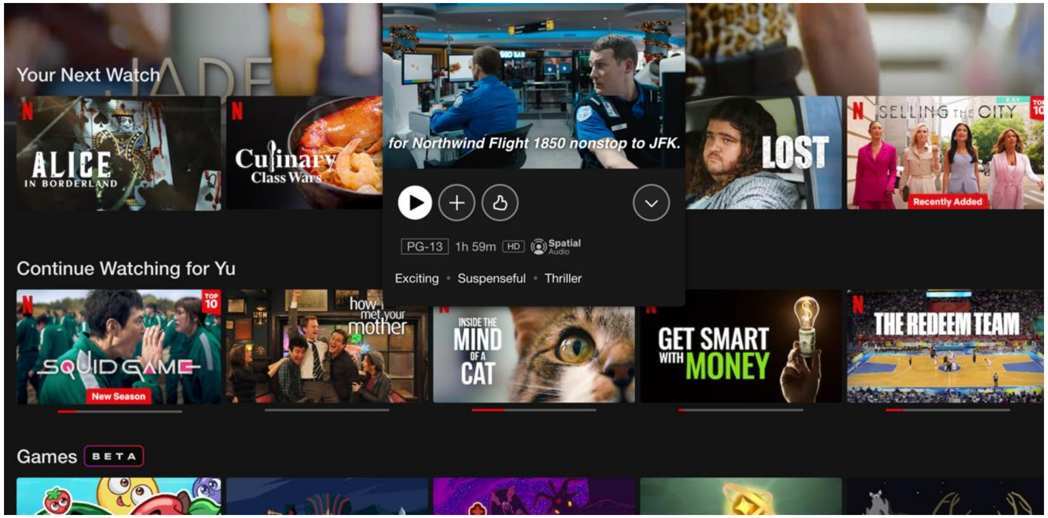




Sparsity

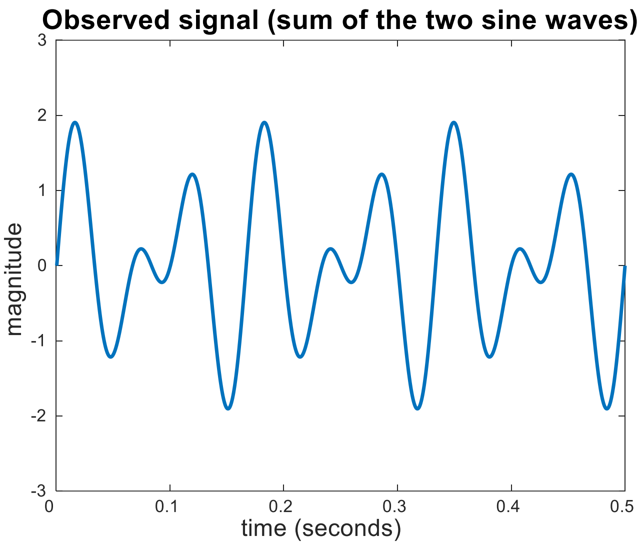


Users' network 1
Users' network 2
New user
Users' network 3

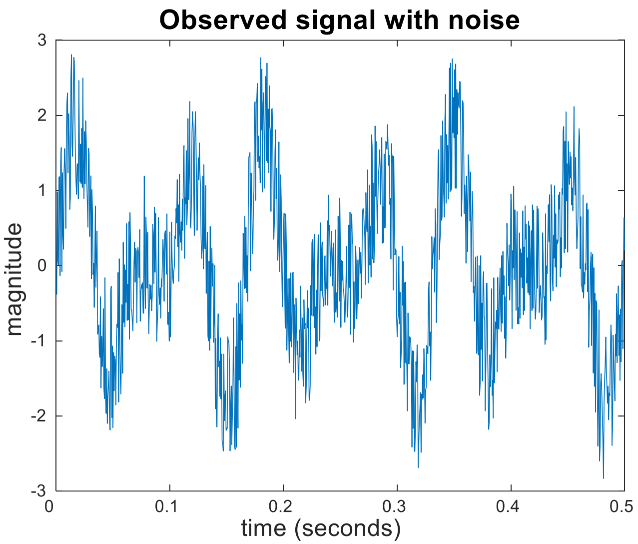


Your Next Watch
ALICE IN BORDERLAND
Culinary Class Wars
for Northwind Flight 1850 nonstop to JFK.
LOST
SELLING THE CITY
Recently Added
PG-13
1h 59m
HD
Spatial Audio
Exciting • Suspenseful • Thriller
Continue Watching for Yu
SQUID GAME
New Season
how I met your mother
INSIDE THE MIND OF A CAT
GET SMART WITH MONEY
THE REDEEM TEAM
Games
BETA

Resolution



Observed signal (sum of the two sine waves)
magnitude
time (seconds)



Observed signal with noise
magnitude
time (seconds)

Types of data sets

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

Data Matrix

- **Data with same fixed set of numeric attributes**

  Points in a multi-dimensional space

- **Such a 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

Document Data

- **Each document becomes a 'term' vector**
  - **Each term is a component (attribute) of the vector**
  - **The value of each component is the number of times the corresponding term occurs in the document.**

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

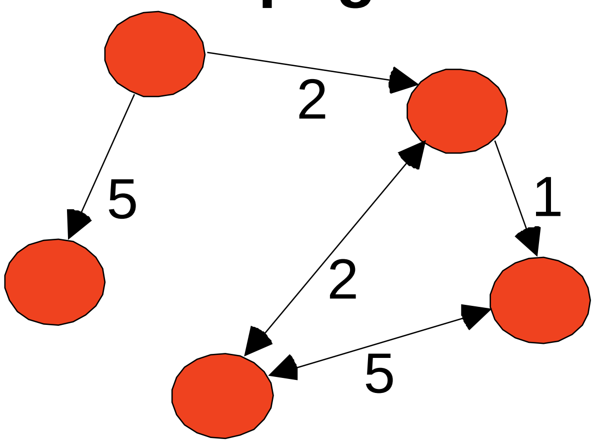
Transaction Data

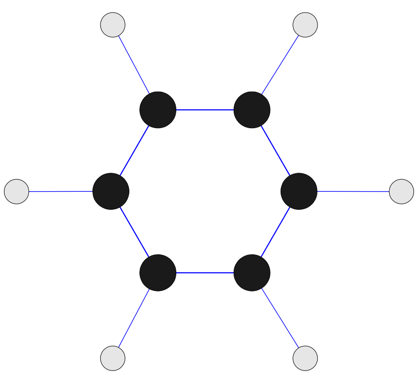
- A special type of data, where
  - Each transaction involves a set of items.
  - The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.
  - Can represent transaction data as record data

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

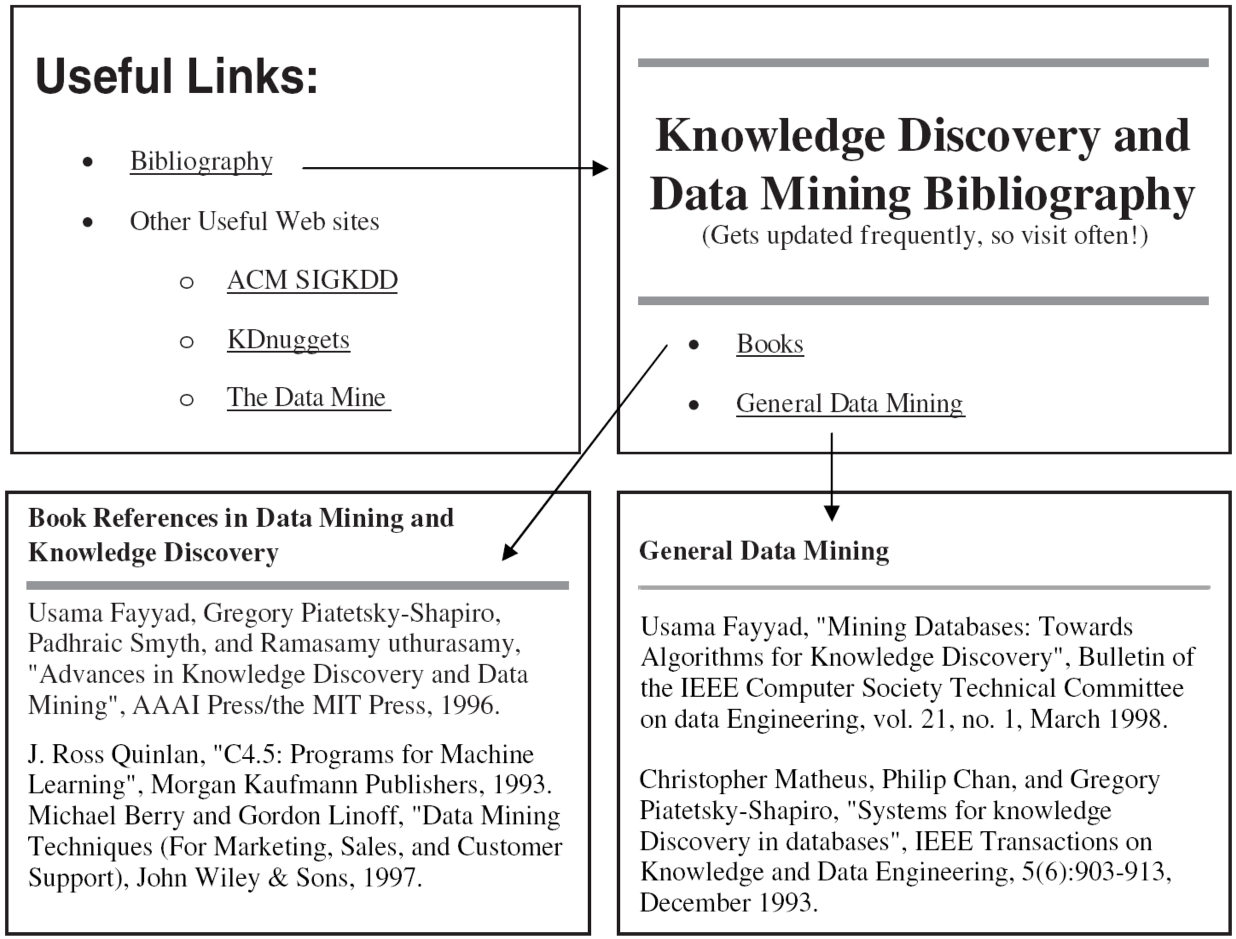
Graph Data

- Examples: Generic graph, a molecule, and webpages





Benzene Molecule: C_6H_6



Ordered Data

- **Sequences of transactions**

Items/Events

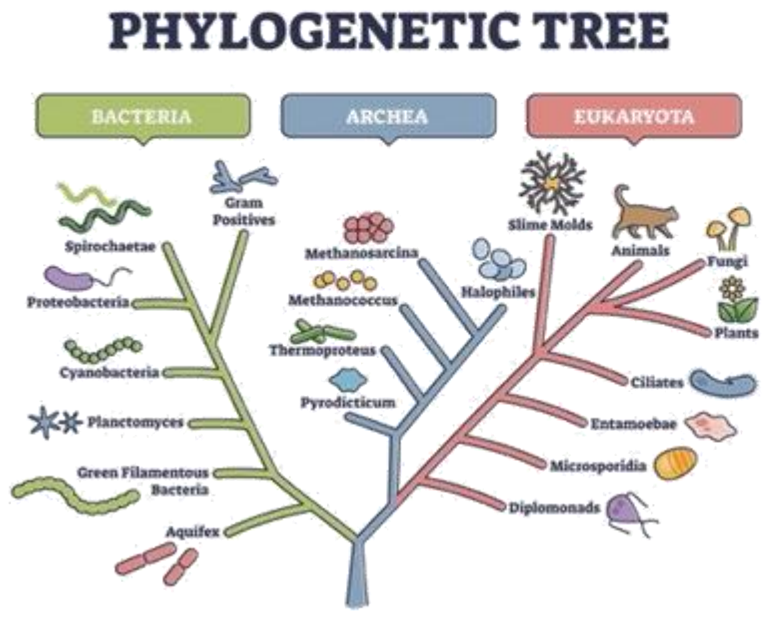
(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

An element of the sequence

Ordered Data

- Genomic sequence data

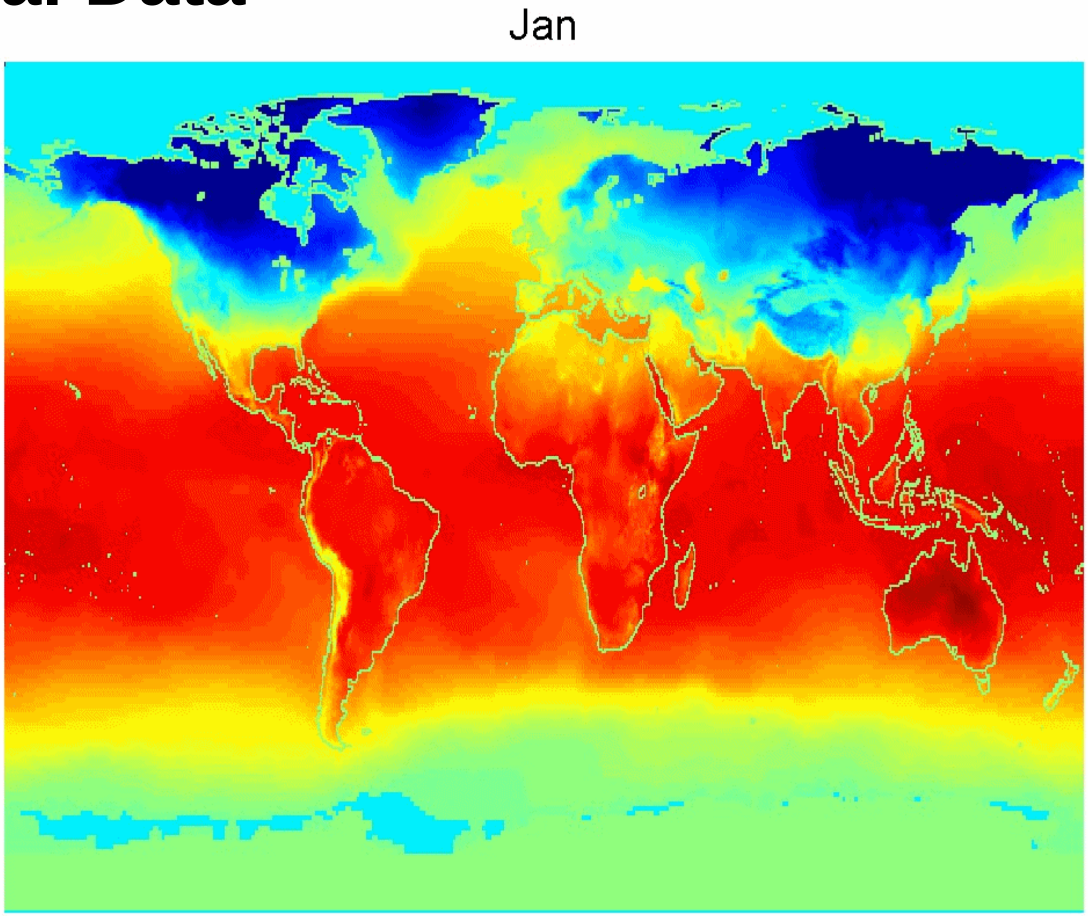
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG



Ordered Data

- **Spatio-Temporal Data**

Average Monthly Temperature of land and ocean



Data Quality

- Poor data quality negatively affects many data processing efforts

- Data mining example: a classification model for detecting people who are loan risks is built using poor data
  - Some credit-worthy candidates are denied loans
  - More loans are given to individuals by default

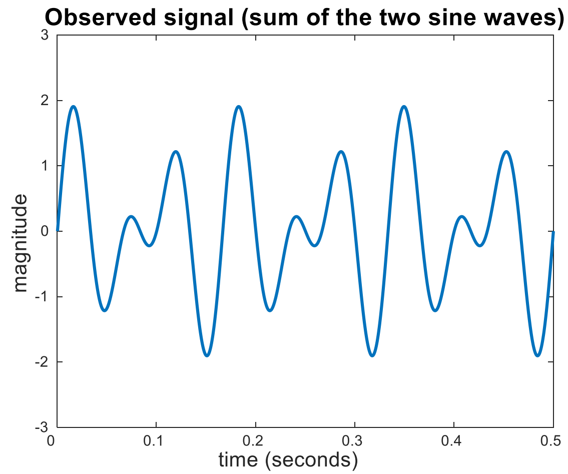
Data Quality …

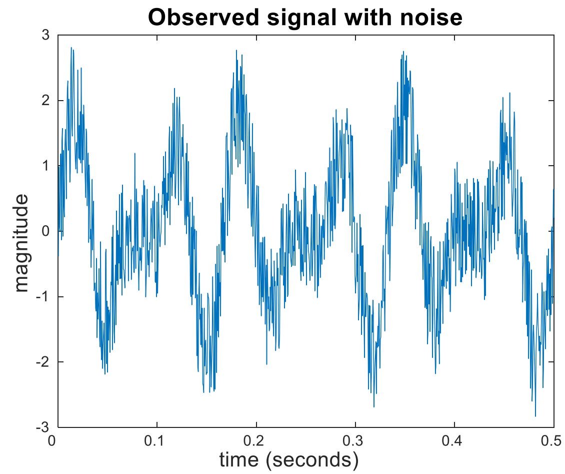
- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

- Examples of data quality problems:
  - Noise and outliers
  - Fake data
  - Missing values
  - Duplicate data

Noise

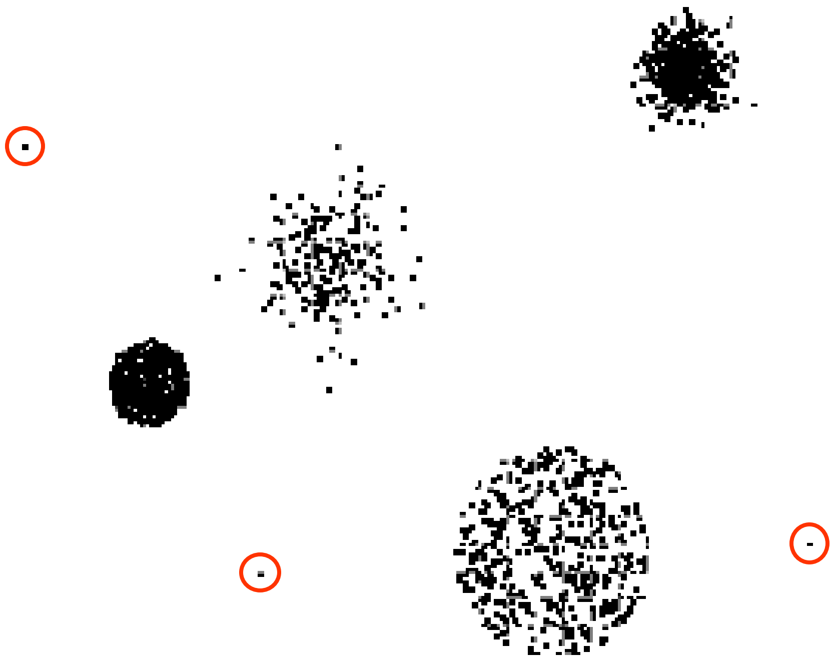
- **Extraneous Objects**
- **Attributes, noise refers to modification of original values**
  - **Examples: distortion of a person's voice on poor phone**
  - **Two sine waves of the same magnitude and different frequencies**





Outliers

- *Outliers* are data objects with characteristics that are considerably different than most of the other data objects in the data set
  - Case 1: Outliers are noise that interferes with data analysis
  - Case 2: Outliers are the goal
    - Credit card fraud
    - Intrusion detection



- Causes?

Missing Values

- **Reasons for missing values**
  - **Information is 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)**

- **Handling missing values**
  - **Eliminate data objects or variables**
  - **Estimate missing values**
    - **Example: time series of temperature**
    - **Example: node attribute**
  - **Ignore the missing value during analysis**

Duplicate Data

- **Data set may include data objects that are duplicates, or almost duplicates of one another**
  - **Major issue when merging data from heterogeneous sources**

- **Examples:**
  - **Same person with multiple email addresses**

- **Data cleaning**
  - **Process of dealing with duplicate data issues**

- **When should duplicate data not be removed?**

Similarity and Dissimilarity Measures

- **Similarity measure**
  - **Numerical measure of how alike two data objects are.**
  - **Is higher when objects are more alike.**

- **Dissimilarity measure**
  - **Numerical measure of how different two data objects are**
  - **Lower when objects are more alike**

Similarity/Dissimilarity for Simple Attributes

The following table shows the similarity and dissimilarity between two objects, *x* and *y*, with respect to a single, simple attribute.

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

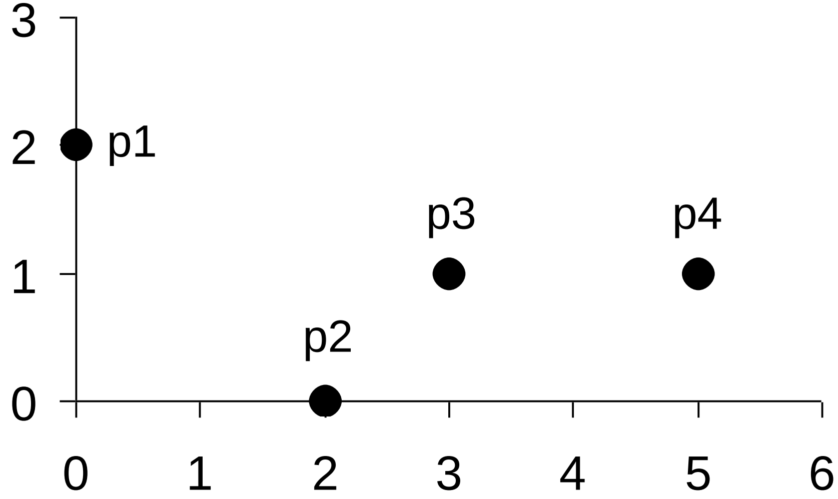
- **Euclidean Distance**

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

where *n* is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects $\mathbf{x}$ and $\mathbf{y}$.

- Standardization is necessary, if scales differ.

Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

Where r is a parameter, n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects x and y.

Minkowski Distance: Examples

- ***r* = 1. City block (Manhattan, taxicab, L_1 norm) distance.**

- ***r* = 2. Euclidean distance**

- **Do not confuse *r* with *n*, i.e., all these distances are defined for all numbers of dimensions.**

Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.

    1. $d(x, y) \geq 0$ for all *x* and *y* and $d(x, y) = 0$ if and only if $x = y$.
    2. $d(x, y) = d(y, x)$ for all x and y. (Symmetry)
    3. $d(x, z) \leq d(x, y) + d(y, z)$ for all points x, y, and z. (Triangle Inequality)

  where $d(x, y)$ is the distance (dissimilarity) between points (data objects), x and y.

- A distance that satisfies these properties is a metric

Common Properties of a Similarity

- Similarities, also have some well known properties.

1. $s(\mathrm{x}, \mathrm{y}) = 1$ (or maximum similarity) only if $\mathrm{x} = \mathrm{y}$.
   (does not always hold, e.g., cosine)
2. $s(\mathrm{x}, \mathrm{y}) = s(\mathrm{y}, \mathrm{x})$ for all x and y. (Symmetry)

where $s(\mathrm{x}, \mathrm{y})$ is the similarity between points (data objects), x and y.

Similarity Between Binary Vectors

- Common situation is that objects, $\mathbf{x}$ and $\mathbf{y}$, have only binary attributes

- Compute similarities using the following quantities

  f_{01} = the number of attributes where $\mathbf{x}$ was 0 and $\mathbf{y}$ was 1
  f_{10} = the number of attributes where $\mathbf{x}$ was 1 and $\mathbf{y}$ was 0
  f_{00} = the number of attributes where $\mathbf{x}$ was 0 and $\mathbf{y}$ was 0
  f_{11} = the number of attributes where $\mathbf{x}$ was 1 and $\mathbf{y}$ was 1

- Simple Matching and Jaccard Coefficients

  SMC = number of matches / number of attributes
  $= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$

SMC versus Jaccard: Example

x = 1 0 0 0 0 0 0 0 0 0

y = 0 0 0 0 0 0 1 0 0 1

$f_{01} = 2$ (the number of attributes where x was 0 and y was 1)

$f_{10} = 1$ (the number of attributes where x was 1 and y was 0)

$f_{00} = 7$ (the number of attributes where x was 0 and y was 0)

$f_{11} = 0$ (the number of attributes where x was 1 and y was 1)

$$\text{SMC} = (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$$
$$= (0+7) / (2+1+0+7) = 0.7$$

Cosine Similarity

- If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = \langle d_1,d_2\rangle / ||d_1|| \, ||d_2|| ,$$

where $\langle d_1,d_2\rangle$ indicates inner product or vector dot product of vectors, d_1 and d_2, and $|| d ||$ is the length of vector d.

- Example:

$$d_1 = 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0$$

$$d_2 = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2$$

$\langle d_1, d2\rangle = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$| d_1 || = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$|| d_2 || = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.449$

$\cos(d_1, d_2) = 0.3150$

Correlation measures the linear relationship between objects

$$\mathrm{corr}(\mathbf{x},\mathbf{y}) = \frac{\mathrm{covariance}(\mathbf{x},\mathbf{y})}{\mathrm{standard\_deviation}(\mathbf{x}) * \mathrm{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x\ s_y}, \quad (2.11)$$

where we are using the following standard statistical notation and definitions

$$\mathrm{covariance}(\mathbf{x},\mathbf{y}) = s_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y}) \quad (2.12)$$

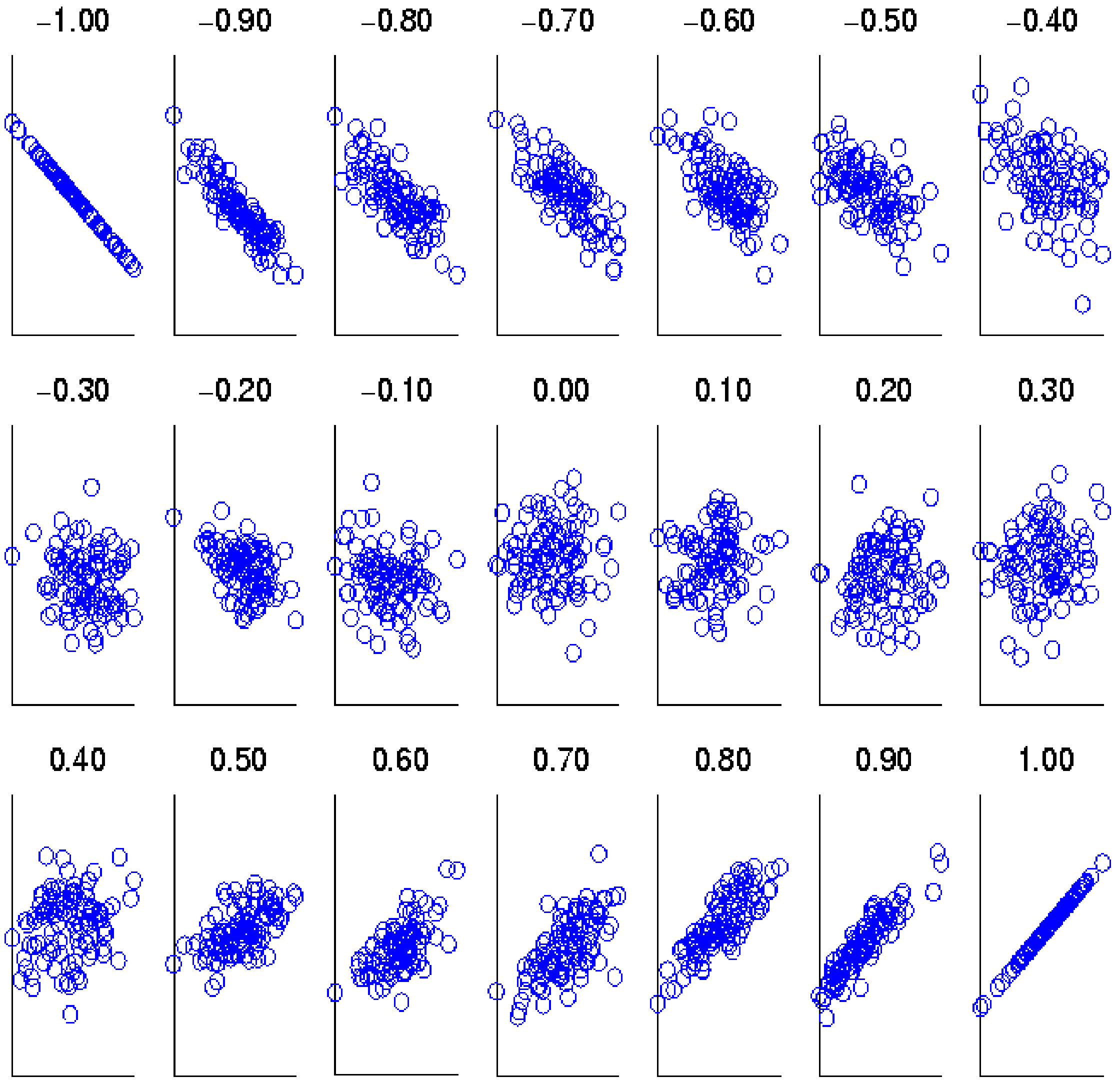
$$\mathrm{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^2}$$

$$\mathrm{standard\_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \overline{y})^2}$$

$$\overline{x} = \frac{1}{n}\sum_{k=1}^{n} x_k \text{ is the mean of } \mathbf{x}$$

$$\overline{y} = \frac{1}{n}\sum_{k=1}^{n} y_k \text{ is the mean of } \mathbf{y}$$

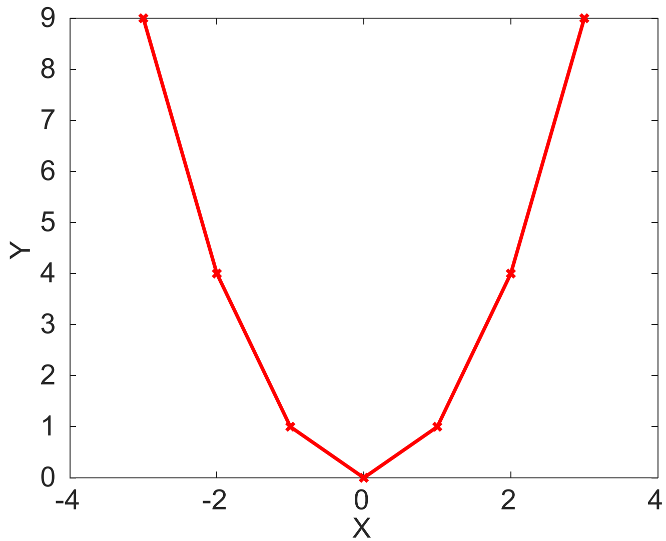
Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

Drawback of Correlation

- x = (-3, -2, -1, 0, 1, 2, 3)
- y = (9, 4, 1, 0, 1, 4, 9)



$y_i = x_i^2$

- mean(x) = 0, mean(y) = 4
- std(x) = 2.16, std(y) = 3.74

- corr = (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5) / (6 * 2.16 * 3.74)

  = 0

Correlation vs Cosine vs Euclidean Distance

- Compare the three proximity measures according to their behavior under variable transformation
  - scaling: multiplication by a value
  - translation: adding a constant

Property	Cosine	Correlation	Euclidean Distance
Invariant to scaling (multiplication)	Yes	Yes	No
Invariant to translation (addition)	No	Yes	No

- Consider the example
  - $x = (1, 2, 4, 3, 0, 0, 0)$, $y = (1, 2, 3, 4, 0, 0, 0)$
  - $y_s = y * 2$ (scaled version of y), $y_t = y + 5$ (translated version)

Measure	(x, y)	(x, y_s)	(x, y_t)
Cosine	0.9667	0.9667	0.7940
Correlation	0.9429	0.9429	0.9429
Euclidean Distance	1.4142	5.8310	14.2127

Correlation vs cosine vs Euclidean distance

- Choice of the right proximity measure depends on the domain
- What is the correct choice of proximity measure for the following situations?
  - Comparing documents using the frequencies of words
    - Documents are considered similar if the word frequencies are similar
  - Comparing the temperature in Celsius of two locations
    - Two locations are considered similar if the temperatures are similar in magnitude
  - Comparing two time series of temperature measured in Celsius
    - Two time series are considered similar if their “shape” is similar, i.e., they vary in the same way over time, achieving minimums and maximums at similar times, etc.

General Approach for Combining Similarities

- **Sometimes attributes are of many different types, but an overall similarity is needed.**

1: For the k^{th} attribute, compute a similarity, $s_k(\mathbf{x}, \mathbf{y})$, in the range [0, 1].

2: Define an indicator variable, δ_k, for the k^{th} attribute as follows:

3. Compute

$$\text{similarity}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^{n} \delta_k}$$

Using Weights to Combine Similarities

- May not want to treat all attributes the same.
  - Use non-negative weights ω_k
  - $similarity(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} \omega_k \delta_k s_k(\mathbf{x},\mathbf{y})}{\sum_{k=1}^{n} \omega_k \delta_k}$
- Can also define a weighted form of distance

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} w_k |x_k - y_k|^r \right)^{1/r}$$