# CSCI568

## Discussion 7: Intro to Similarity, Dissimilarity

# Hello. I am a computer.

And I have no idea what love, happiness or similarity mean.

# Defining Similarity (to a computer)

Similarity between two objects is a numerical measure of the degree to which the two objects are alike.

# Dis/Similarity Values

Usually, use ranges `[-1, 1]` or `[0, 1]`.

(But not everyone does, so you may need to transform the similarity score.)

DM 66, 67

# Dis/similarity Between Two Attributes

| Type | Dissimilarity | Similarity |
|---|---|---|
| Nominal | | |
| Ordinal | | |
| Interval/Ratio | | |

# Dissimilarity of Single Attributes

- nominal: it is or it isn't
- ordinal
  - $d = |x - y| / (n-1)$
  - $s = 1 - d$
- continuous:
  - $d = |x - y|$
  - $s = 1/1+d$ (more, DM69)

# Proximity Calculation Issues

- attributes w/ different scales
  - (eg, age vs. income)
- heterogeneous attributes
  - (eg, nominal and interval attributes)
- attributes w/ different importance

# Euclidean Distance

Simple! Linear distance between two points.

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

$x_k$ and $y_k$ are values of $k^{th}$ attribute of objects x and y

DM 69 - 71

# Measuring Proximity of Data Objects

- Euclidean / Minkowski distance
- Simple Matching Coefficient (SMC)
- Jaccard / Tanimoto
- Cosine Similarity
- Pearson Correlation Coefficient
- Bregman Divergence

# Example: Movie Recommendations