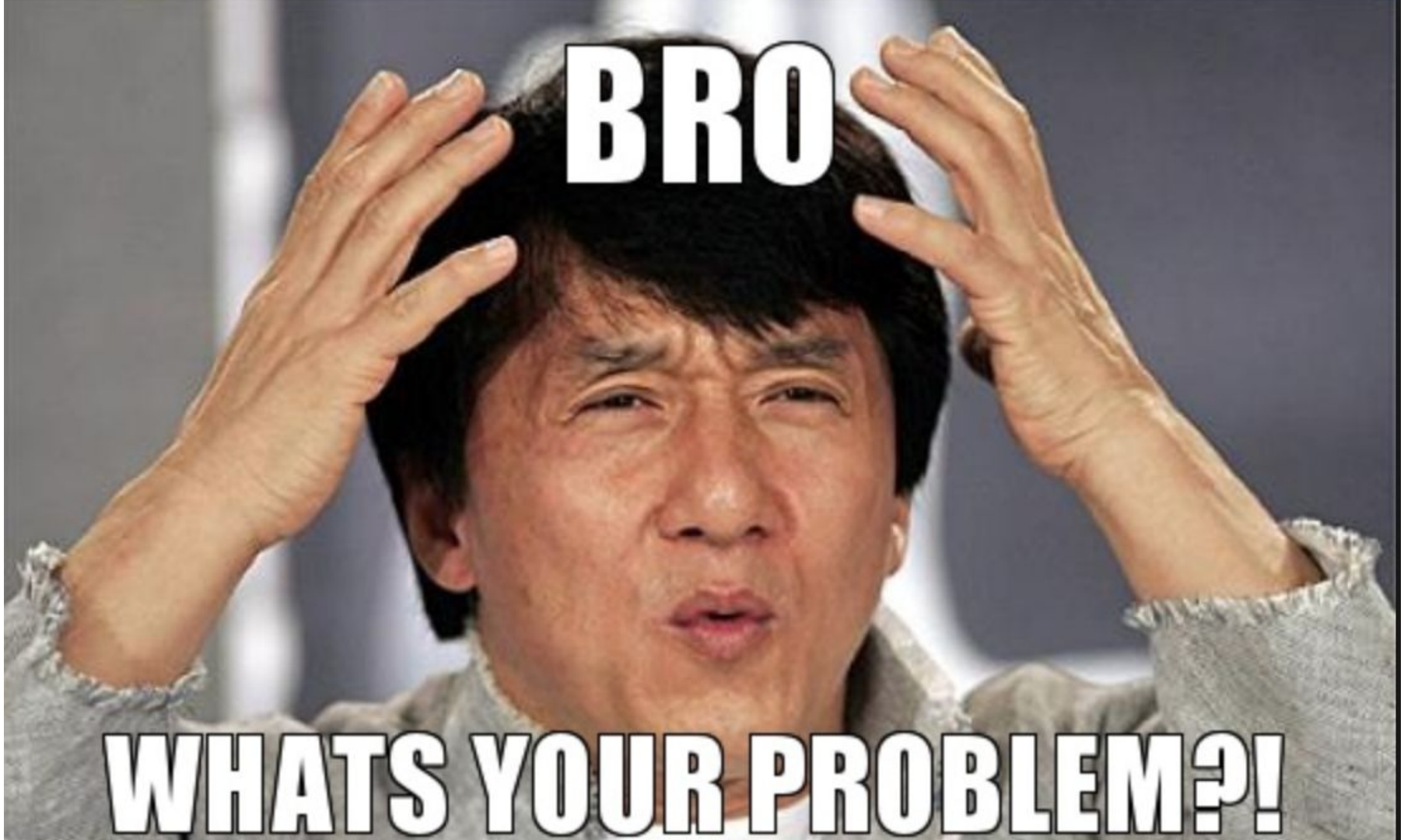
Similarity measurements


BRO
WHATS YOUR PROBLEM?!

- Is there any other day that had similar weather?
    - What was the sales rate on that day?
- What two days have the most similar weather?
- What is the most common weaher?

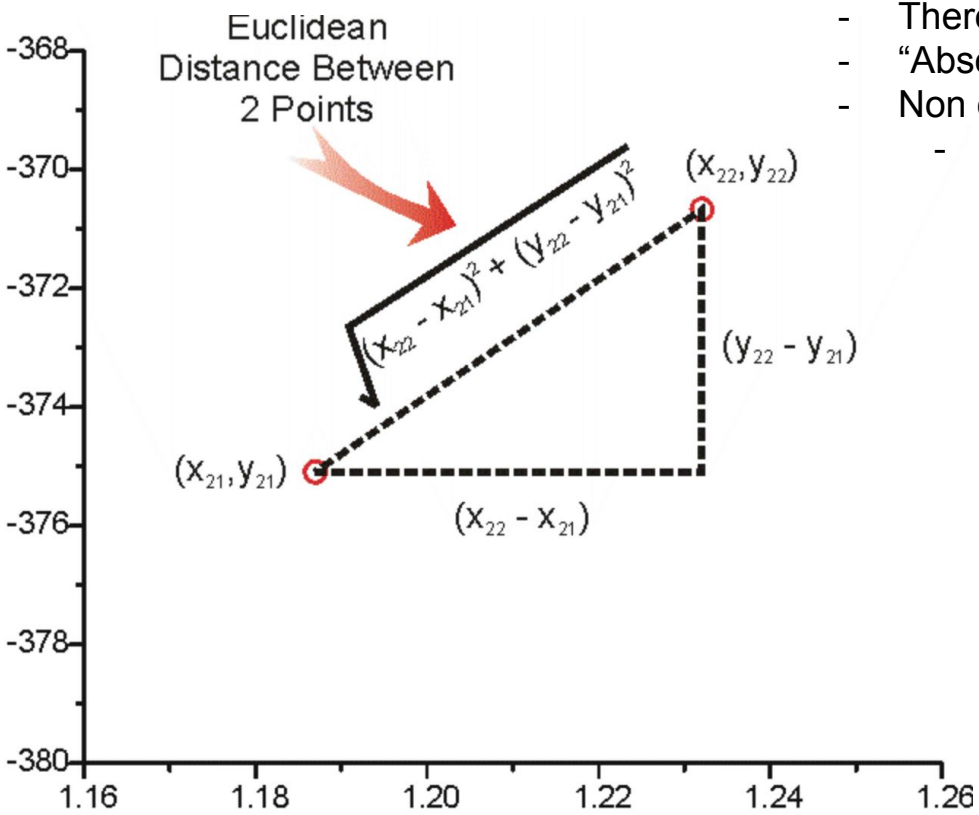
2009-11-22;37.0;57.0;0.19;0.0;0.0;5.82;70;17.0;70;23.04;Yes;No;Yes;Yes;No;No;No;No;No;No;No;No;No;Yes;No;No;No
2009-11-27;36.0;54.0;0.0;0.0;0.0;7.38;250;21.03;270;31.09;No;No;No;Yes;No;No;No;No;No;No;No;No;No;No;No;No;No
2009-12-01;30.0;55.9;0.0;0.0;0.0;2.01;240;8.05;230;12.08;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No
2009-12-08;33.1;46.0;0.45;0.0;0.0;6.49;80;16.11;80;21.92;Yes;No;Yes;Yes;No;No;No;Yes;No;No;No;No;No;Yes;No;No;No
2009-12-11;25.0;36.0;0.0;0.0;0.0;3.36;290;12.08;290;16.11;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No
2009-12-16;28.9;48.9;0.0;0.0;0.0;2.91;10;10.07;350;16.11;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No
2009-12-17;28.0;45.0;0.0;0.0;0.0;2.01;50;10.07;20;14.09;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No
2009-12-21;24.1;48.9;0.0;0.0;0.0;2.24;240;14.09;270;17.9;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No
2009-12-23;25.0;53.1;0.0;0.0;0.0;0.45;50;6.04;160;10.07;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No
2010-01-07;19.9;48.0;0.0;0.0;0.0;5.37;210;12.97;230;17.9;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No
2010-01-08;21.9;42.1;0.05;0.0;0.0;8.95;240;23.04;230;33.11;Yes;No;Yes;Yes;No;No;No;Yes;No;No;No;No;No;Yes;No;No;No

Euclidean distance

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

Figure 2: Euclidean distance [6]

Euclidean distance



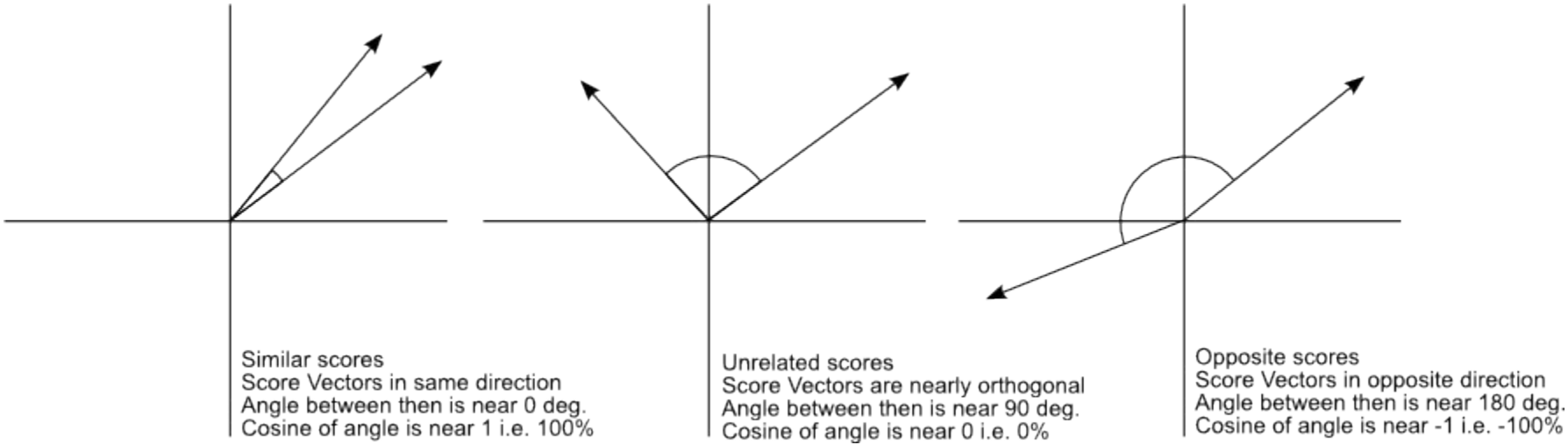
- There is a notion of average
- “Absolute” location in space
- Non euclidean distance is based on properties of points
  - But no concept of location in space

Cosine similarity

$$cos(x,y) = \frac{(x \bullet y)}{||x|| ||y||}$$

Figure 4: Cosine similarity [6]

Cosine similarity



Similar scores
Score Vectors in same direction
Angle between then is near 0 deg.
Cosine of angle is near 1 i.e. 100%

Unrelated scores
Score Vectors are nearly orthogonal
Angle between then is near 90 deg.
Cosine of angle is near 0 i.e. 0%

Opposite scores
Score Vectors in opposite direction
Angle between then is near 180 deg.
Cosine of angle is near -1 i.e. -100%

Inner product (dot product)

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^{\mathrm{T}} \mathbf{b}$$

$$= \begin{pmatrix} a_1 & a_2 & \cdots & a_n \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

$$= a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

$$= \sum_{i=1}^{n} a_i b_i,$$

L1-Norm

- Manhattan norm
- Taxicab norm

$$\|\boldsymbol{x}\|_1 := \sum_{i=1}^{n} |x_i| \,.$$

Manhattan distance

P-Norm

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

Pearson correlation

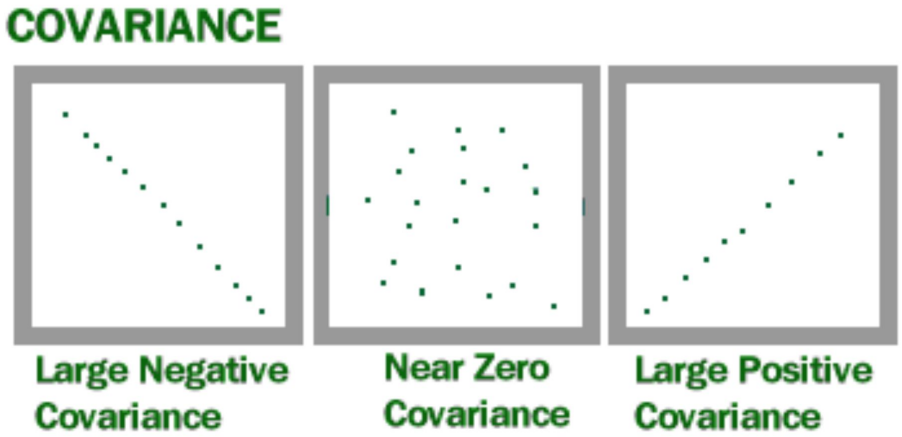
- Linear relationship between x and y

$$Pearson(x,y) = \frac{\Sigma(x,y)}{\sigma_x \times \sigma_y}$$

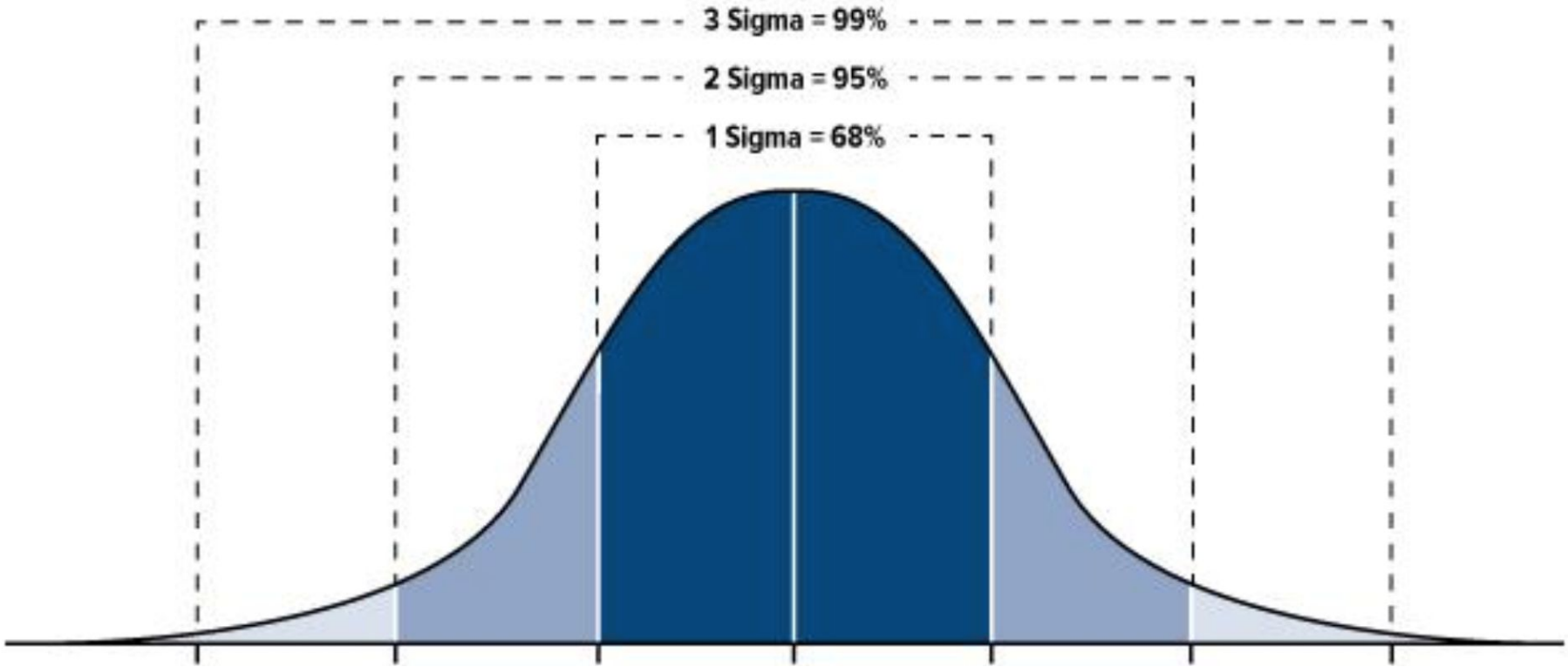
Figure 5: Pearson correlation [6]

Covariance

$$\text{Covariance} = \frac{\sum (x_i - x_{avg})(y_i - y_{avg})}{n-1}$$



Standard deviation



Standard deviation

$$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} |A_i - \mu|^2}$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} A_i.$$

Jaccard coefficient

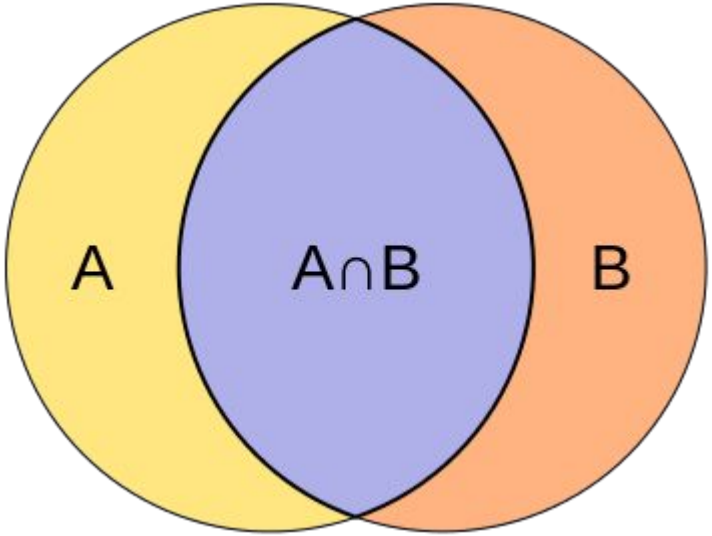
$$JC = \frac{M11}{M01 + M10 + M11}$$

M11 = Nr of items in both sets
M10 = Nr of items in set A
M01 = Nr of items in set B

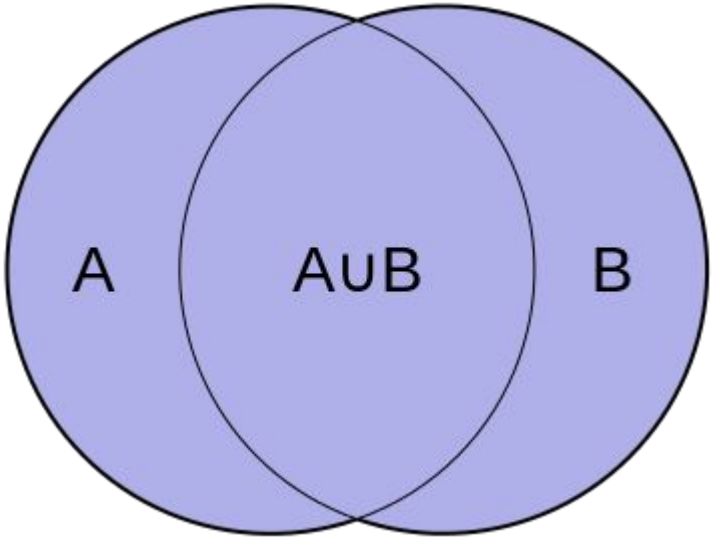
2009-11-22;37.0;57.0;0.19;0.0;0.0;5.82;70;17.0;70;23.04;Yes;No;Yes;Yes;No;No;No;No;No;No;No;No;No;Yes;No;No;No
2009-11-27;36.0;54.0;0.0;0.0;0.0;7.38;250;21.03;270;31.09;No;No;No;Yes;No;No;No;No;No;No;No;No;No;No;No;No;No
2009-12-01;30.0;55.9;0.0;0.0;0.0;2.01;240;8.05;230;12.08;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No
2009-12-08;33.1;46.0;0.45;0.0;0.0;6.49;80;16.11;80;21.92;Yes;No;Yes;Yes;No;No;No;Yes;No;No;No;No;No;Yes;No;No;No
2009-12-11;25.0;36.0;0.0;0.0;0.0;3.36;290;12.08;290;16.11;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No
2009-12-16;28.9;48.9;0.0;0.0;0.0;2.91;10;10.07;350;16.11;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No
2009-12-17;28.0;45.0;0.0;0.0;0.0;2.01;50;10.07;20;14.09;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No
2009-12-21;24.1;48.9;0.0;0.0;0.0;2.24;240;14.09;270;17.9;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No
2009-12-23;25.0;53.1;0.0;0.0;0.0;0.45;50;6.04;160;10.07;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No
2010-01-07;19.9;48.0;0.0;0.0;0.0;5.37;210;12.97;230;17.9;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No;No
2010-01-08;21.9;42.1;0.05;0.0;0.0;8.95;240;23.04;230;33.11;Yes;No;Yes;Yes;No;No;No;Yes;No;No;No;No;No;Yes;No;No;No

Jaccard coefficient

Intersection



Union



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Minkowski distance

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

Figure 3: Minkowski distance [6]

Minkowski distance

- r = 1 this is exactly the same as the euclidean distance.
- r = 2 we are measuring the manhattan distance
- $r \rightarrow \infty$ this corresponds to measuring the maximum difference between any dimension of x and y.