Correlation

- Measures the linear relationship between objects

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

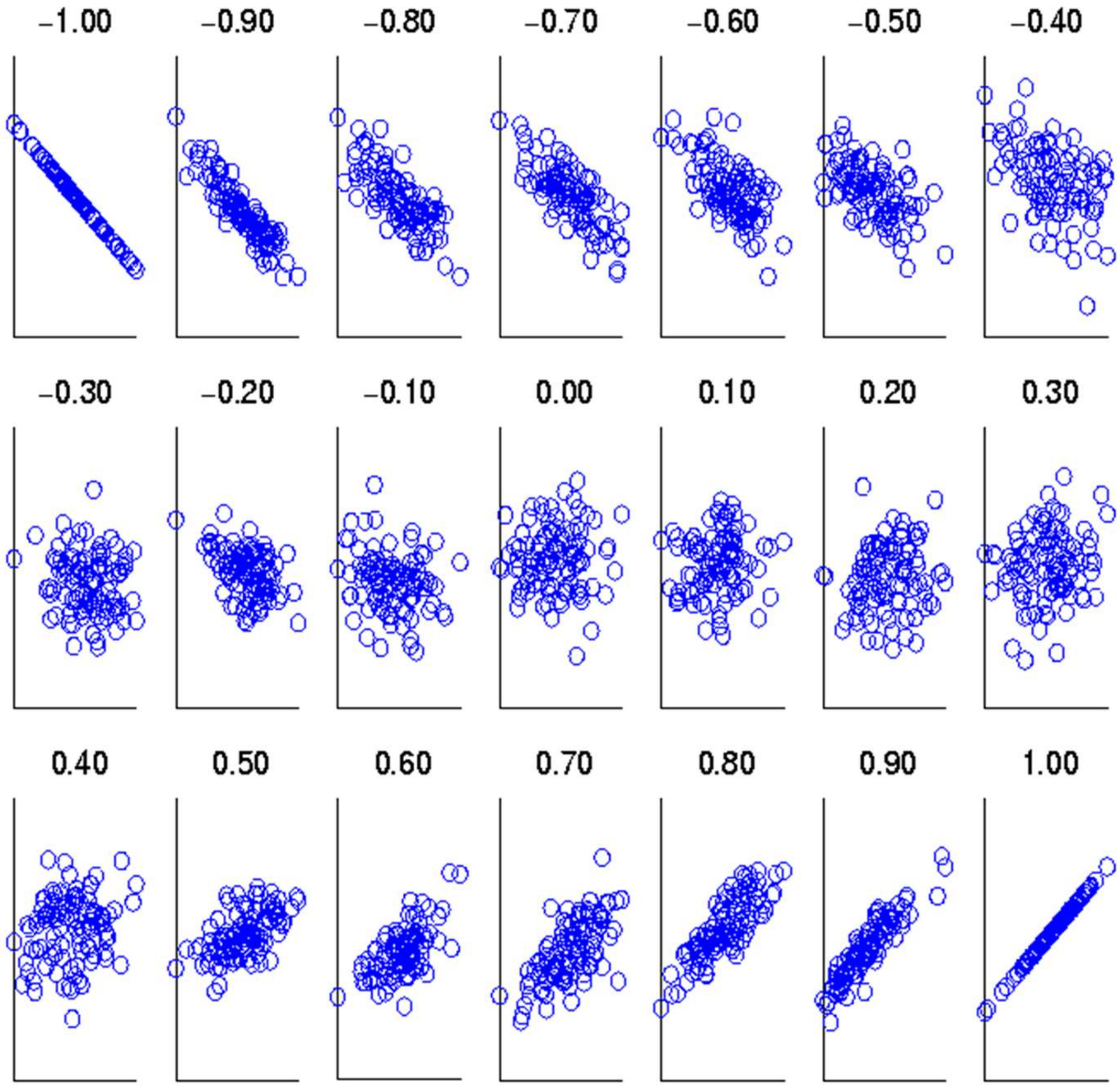
where we are using the following standard statistical notation and definitions

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

$$\begin{aligned}
\mathrm{standard_deviation}(\mathbf{x}) &= s_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\overline{x})^2} \\
\mathrm{standard_deviation}(\mathbf{y}) &= s_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k-\overline{y})^2}
\end{aligned}$$

$$\begin{aligned}
\overline{x} &= \frac{1}{n}\sum_{k=1}^{n} x_k \text{ is the mean of } \mathbf{x} \\
\overline{y} &= \frac{1}{n}\sum_{k=1}^{n} y_k \text{ is the mean of } \mathbf{y}
\end{aligned}$$

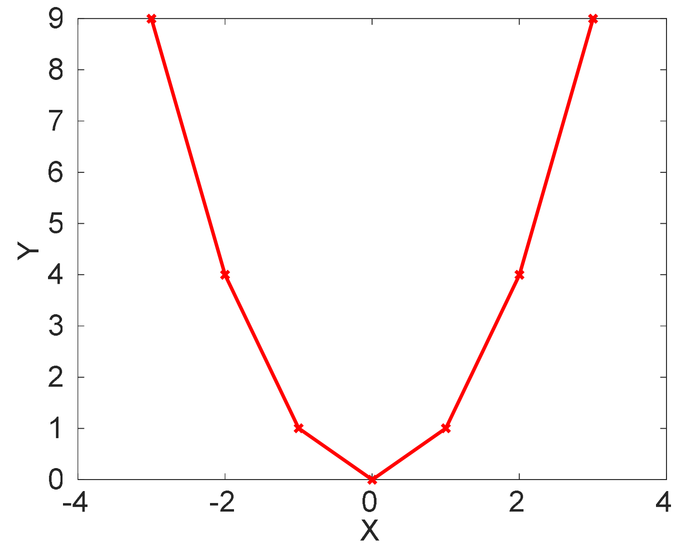
Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

Drawback of Correlation

- $\mathbf{x} = (-3, -2, -1, 0, 1, 2, 3)$
- $\mathbf{y} = (9, 4, 1, 0, 1, 4, 9)$



$y_i = x_i^2$

- $\text{mean}(\mathbf{x}) = 0$, $\text{mean}(\mathbf{y}) = 4$
- $\text{std}(\mathbf{x}) = 2.16$, $\text{std}(\mathbf{y}) = 3.74$

- 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)
  - $\mathbf{y_s} = \mathbf{y} * \mathbf{2}$ (scaled version of y), $\mathbf{y_t} = \mathbf{y} + \mathbf{5}$ (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