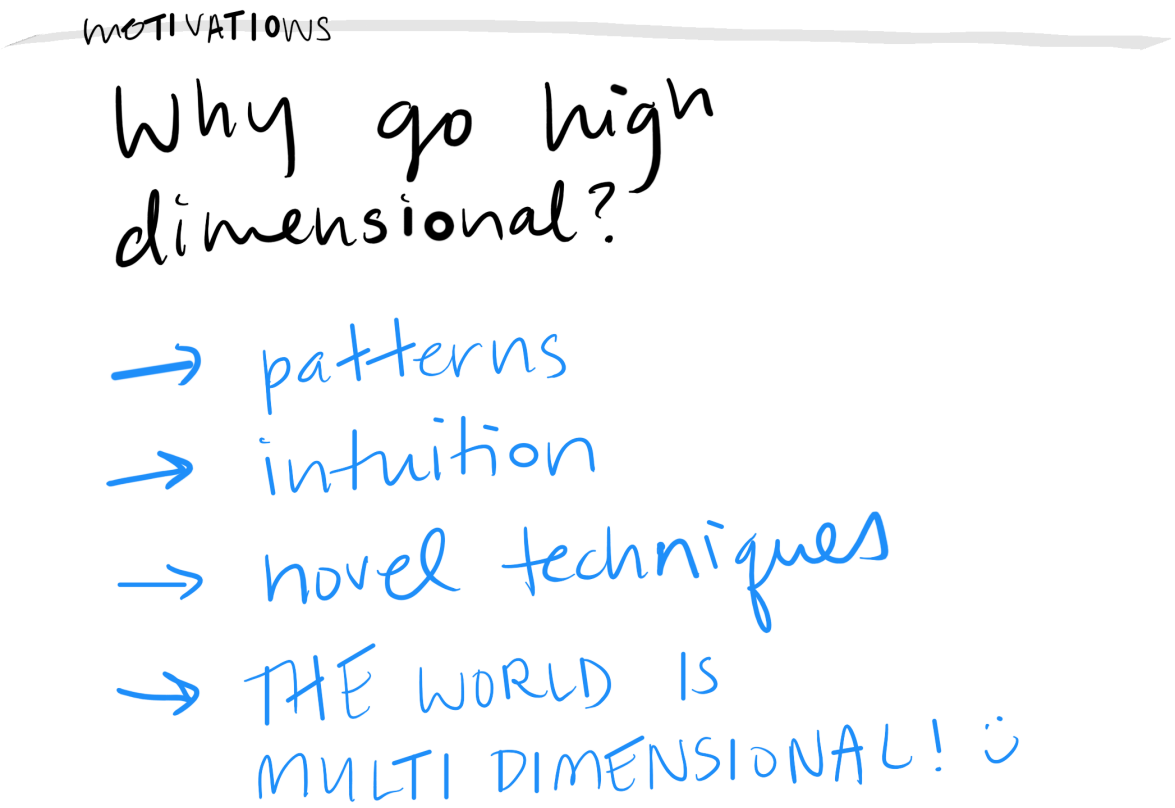
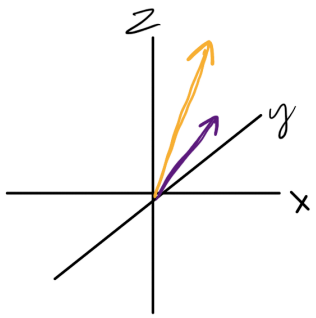
Correlation and Cosine Similarity: An Introduction to Thinking Geometrically



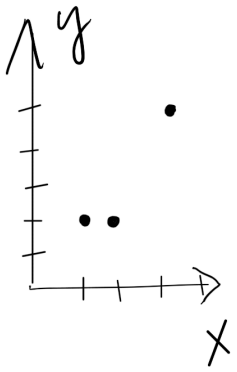
DATA + QUESTIONS

How similar or related are two sets of data?



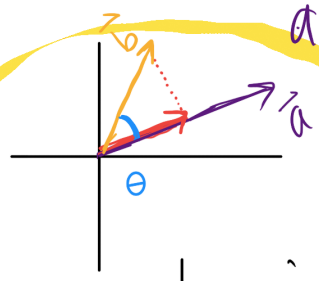
$$\vec{a} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

$$\vec{b} = \begin{bmatrix} 2 \\ 2 \\ 5 \end{bmatrix}$$



x	y
1	2
2	2
3	5

DOT PRODUCT REVIEW



$$\vec{a} \cdot \vec{b} = |\text{proj}_{\vec{a}} \vec{b}| \cdot |\vec{a}|$$

$$= |\vec{b}| \cos\theta \cdot |\vec{a}|$$

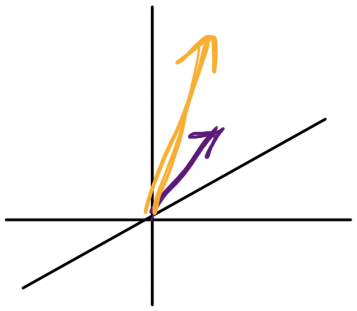
$$\cos\theta = \frac{|\text{proj}_{\vec{a}} \vec{b}|}{|\vec{b}|}$$

Force and Displacement

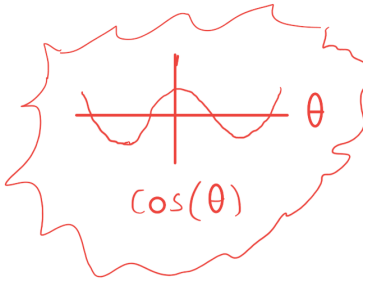
Work = $\vec{F} \cdot \Delta\vec{X}$

In other words, proportion of $\vec{b}$ in direction of $\vec{a}$.

DATA! (HOW SIMILAR ARE MY VECTORS?)



Similarity Metric: cosine θ



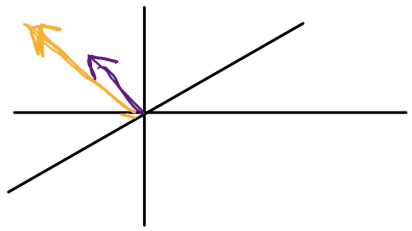
$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}|\,|\vec{b}|}$$

LET'S CALCULATE!

Let's ① mean-center the vectors and ② take cosine similarity

① $\vec{a} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ $\vec{a}_c = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$

$\vec{b} = \begin{bmatrix} 2 \\ 2 \\ 5 \end{bmatrix}$ $\vec{b}_c = \begin{bmatrix} -1 \\ -1 \\ 2 \end{bmatrix}$



②

$$\cos\theta = \frac{\vec{a}_c \cdot \vec{b}_c}{|\vec{a}|\,|\vec{b}|} = \frac{\begin{bmatrix}-1\\0\\1\end{bmatrix} \cdot \begin{bmatrix}-1\\-1\\2\end{bmatrix}}{\sqrt{(-1)^2+(1)^2}\sqrt{(-1)^2+(-1)^2+(2)^2}}$$

$$= \frac{\overbrace{(-1)(-1)}^{1} + \overbrace{0}^{0} + \overbrace{(1)(2)}^{2}}{\sqrt{2}\sqrt{6}}$$

$$= \sqrt{3}/2 \approx .86$$

DATA! (HOW SIMILAR ARE THESE DATASETS?)

x	y
1	2
2	5
3	5

Similarity Measure: Pearson R

$$r = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2}\sqrt{\Sigma(y_i-\bar{y})^2}}$$

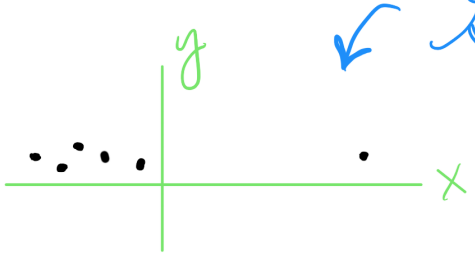
CORRELATION REVIEW!

$$r = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})^2}\ \sqrt{\Sigma (y_i - \bar{y})^2}}$$

covariance

s_x

s_y



Guesses About...

s_x? s_y? covariance?

LET'S CALCULATE!

x	1	2	3
y	2	2	5

$\bar{x} = 2$

$\bar{y} = 3$

$$r = \frac{\overbrace{(1-2)(2-3)}^{1} + \overbrace{(2-2)(2-3)}^{0} + \overbrace{(3-2)(5-3)}^{2}}{\sqrt{(1-2)^2+(2-2)^2+(3-2)^2}\sqrt{(2-3)^2+(2-3)^2+(5-3)^2}}$$

$$r = 3/\sqrt{12} = \sqrt{3}/2 = \boxed{.86}$$

CONCLUSION

Pearson's r $\equiv$ Cosine Similarity of mean-centered vectors

One example of the many ways we can think about our data geometrically ☺