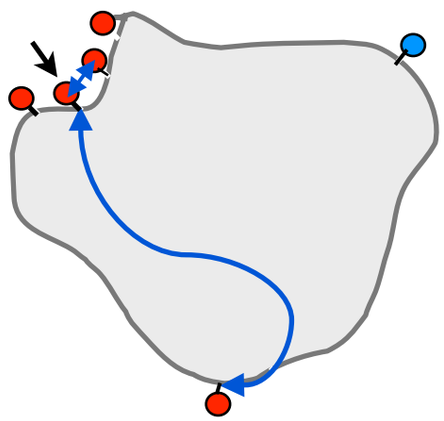
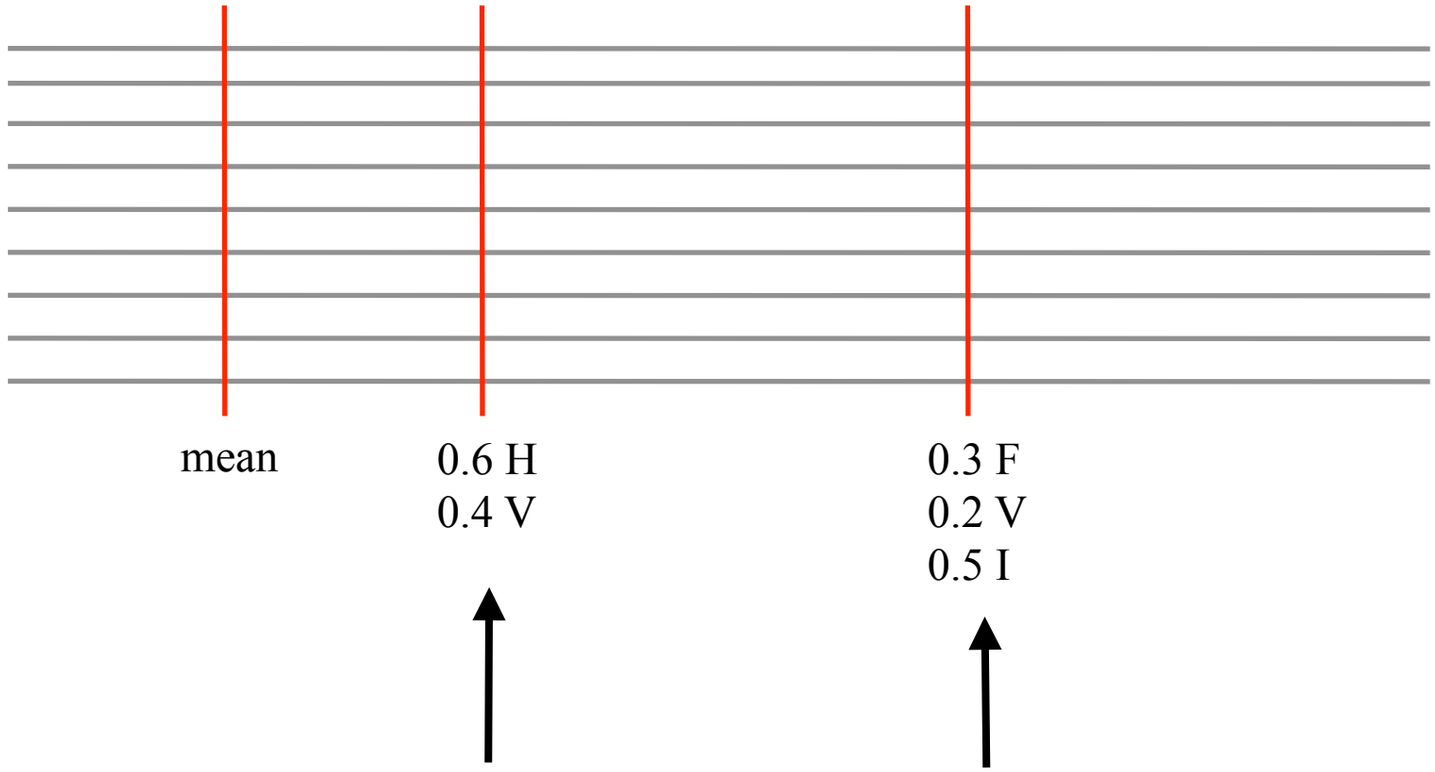
Considering the relationship between covariance, correlation, and mutual information.





j

Remember that our mathematical strategy is to examine interdependence between pairs of amino acid positions, with the goal of inferring shared functional constraints or interactions.

In today's project check-in you considered covariance between amino acid types over pairs of positions.

A reminder on the definition of covariance:

$$cov(x, y) = \sum_{i=1}^{N} p_i(x_i - E(x))(y_i - E(y))$$

$$cov(x, y) = \frac{\sum_{i=1}^{N}(x_i - E(x))(y_i - E(y))}{N}$$ (When probabilities are all uniform)

$$cov(x, y) = \frac{\sum_{i=1}^{N} xy - xE(y) - E(x)y + E(x)E(y)}{N}$$

$$cov(x, y) = E(x, y) - E(x)E(y)$$

In today's project check-in you considered covariance between amino acid types over pairs of positions.

A reminder on the definition of covariance:

$$cov(x,y) = \sum_{i=1}^{N} p_i(x_i - E(x))(y_i - E(y))$$

$$cov(x,y) = \frac{\sum_{i=1}^{N}(x_i - E(x))(y_i - E(y))}{N}$$ (When probabilities are all uniform)

$$cov(x,y) = \frac{\sum_{i=1}^{N} xy - xE(y) - E(x)y + E(x)E(y)}{N}$$

$$cov(x,y) = E(x,y) - E(x)E(y)$$

In the case of amino acid a at site i and amino acid b at site j:

$$C_{ij}^{(ab)} = \left[f_{ij}^{(ab)} - f_i^{(a)} f_j^{(b)} \right]$$

(From here you likely compressed the covariance between amino acids to a positional measure)

Now, the covariance tells you about the direction of a linear relationship between two variables. *But it does not tell you about the magnitude or strength of the relationship*.

For that you need the correlation coefficient:

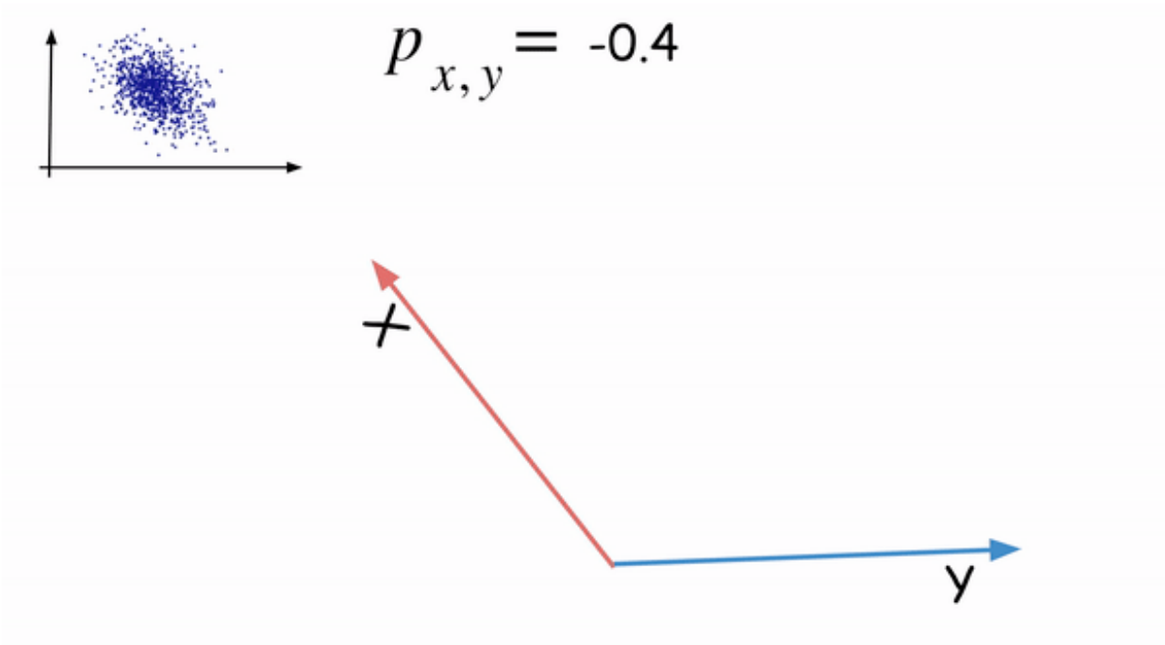
$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$ (Pearson's correlation coefficient).

Now, the covariance tells you about the direction of a linear relationship between two variables. *But it does not tell you about the magnitude or strength of the relationship*.

For that you need the correlation coefficient:

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$ (Pearson's correlation coefficient).

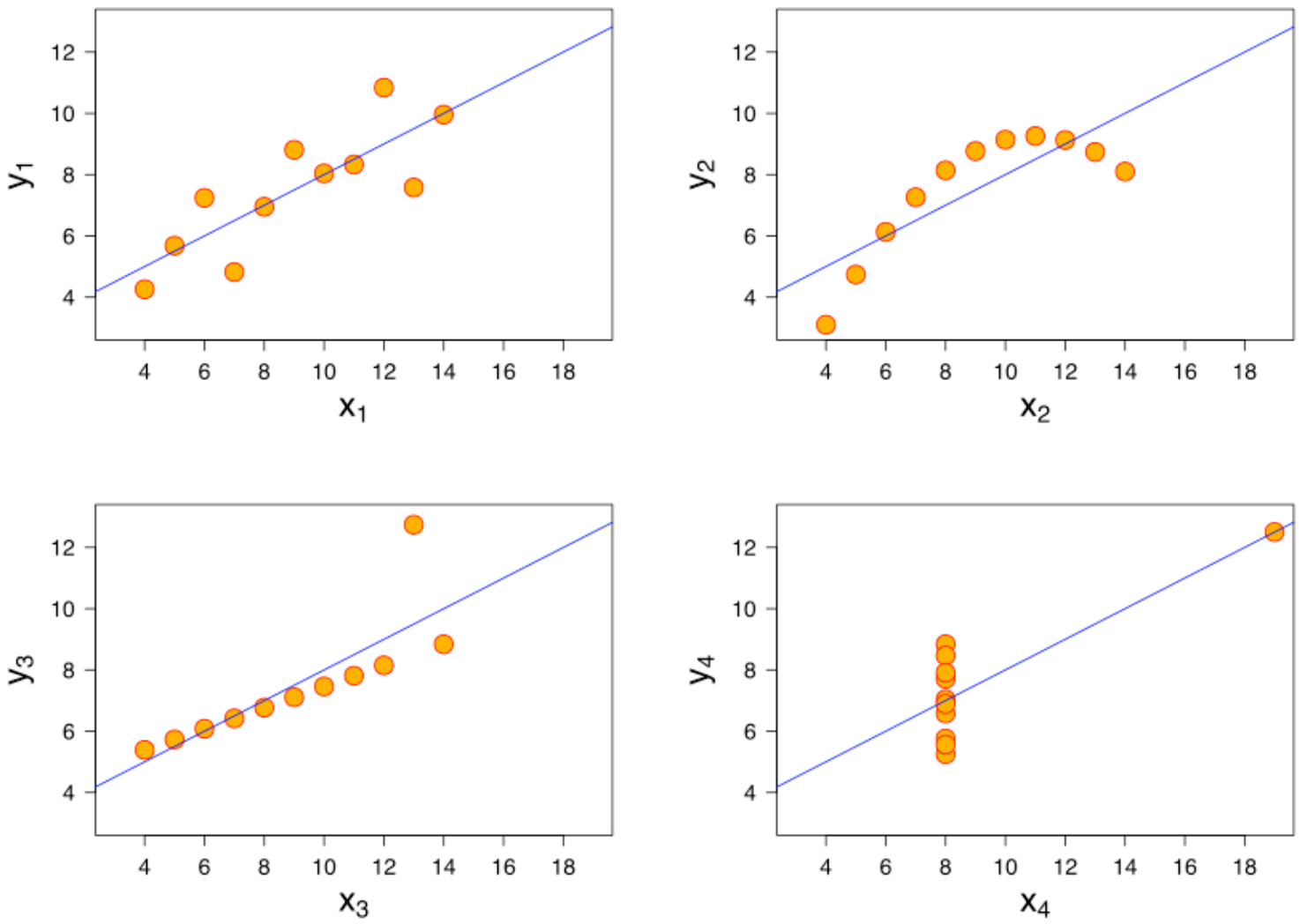
<u>One nice graphical interpretation</u>: it is the cosine of the angle between 2 vectors representing variation in each variable.



For more see: <u>https://medium.com/swlh/a-deep-conceptual-guide-to-mutual-information-a5021031fad0</u>

Now the correlation coefficient works well to describe the linear relationship between two variables, *but it is less appropriate to describe non-linear relationships*.

Anscombe's quartet:

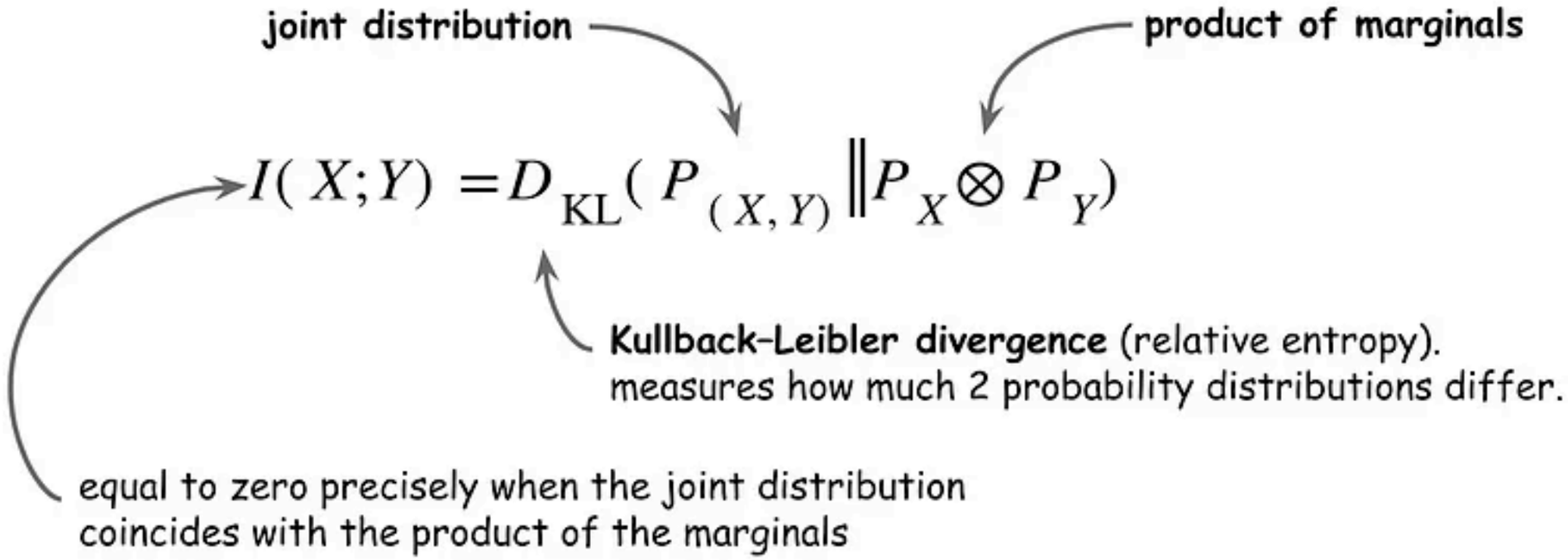


In each case the correlation between x and y is 0.816.
The linear regression line is y = 3 + 0.5x.

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Mutual information between two variables does not presume a linear relationship.

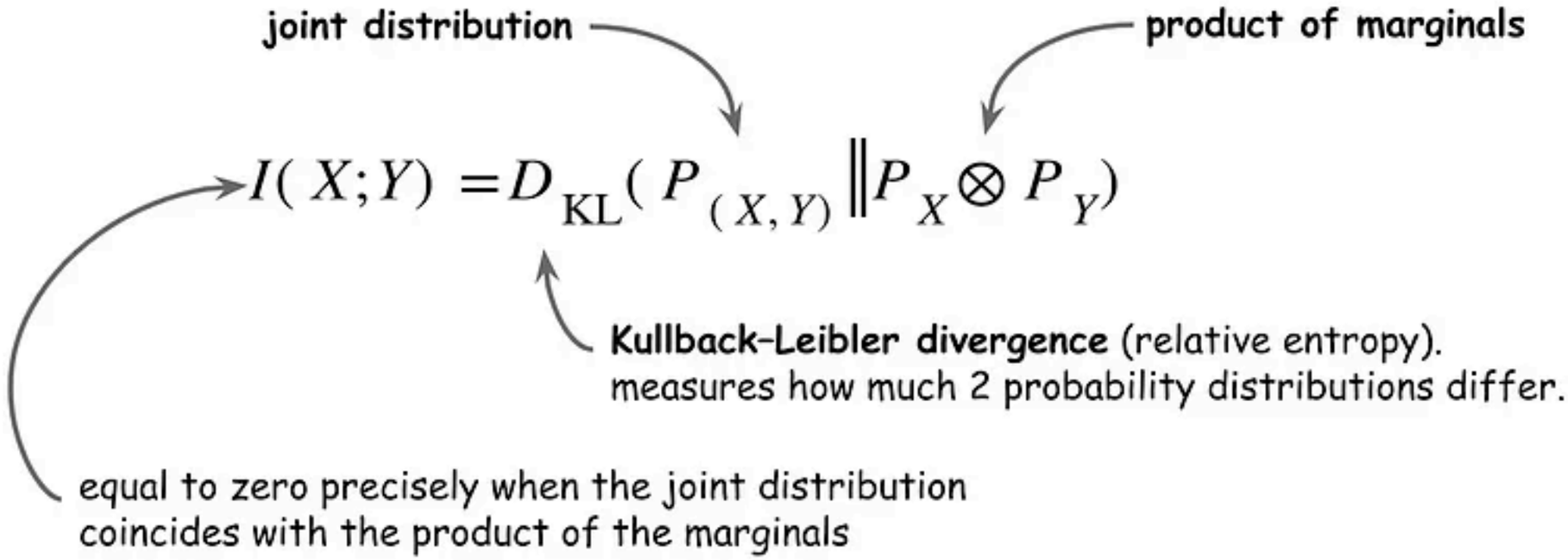
Instead you are computing the K-L divergence between two probability distributions:



For more see: https://medium.com/swlh/a-deep-conceptual-guide-to-mutual-information-a5021031fad0

Mutual information between two variables does not presume a linear relationship.

Instead you are computing the K-L divergence between two probability distributions:



In the case of n amino acids at site X and m amino acids at site Y:

$$\sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk} log \frac{p_{jk}}{p_j q_k}$$

(Note that this naturally gives a position-level measure of interdependence)

Fodor and Aldrich, (2004) Proteins 56:211

At the next project check-in (Feb 21) you will present:

Feb 21 – third project check-in. Analysis of mutual information between amino acid positions.

- Decide how you will compute mutual information (MI). Between individual amino acid types, or classes? How will you compress covariance between amino acid types to obtain a positional measure?
- Compute MI between all position pairs within each protein.
- Using the concatenated alignments, compute MI between all protein pairs.
- Plot the resulting MI matrix as a heatmap.
- Plot histograms of MI within and between individual domains. Again, visually inspect and describe these distributions.

Additionally: Compare your results with MI to covariance.

- Is the pattern of interaction more sparse for one measure than another?
- Does it show a different distribution throughout the sequence for one measure relative to another?