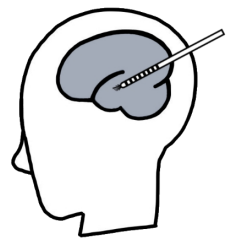
CCA from scratch

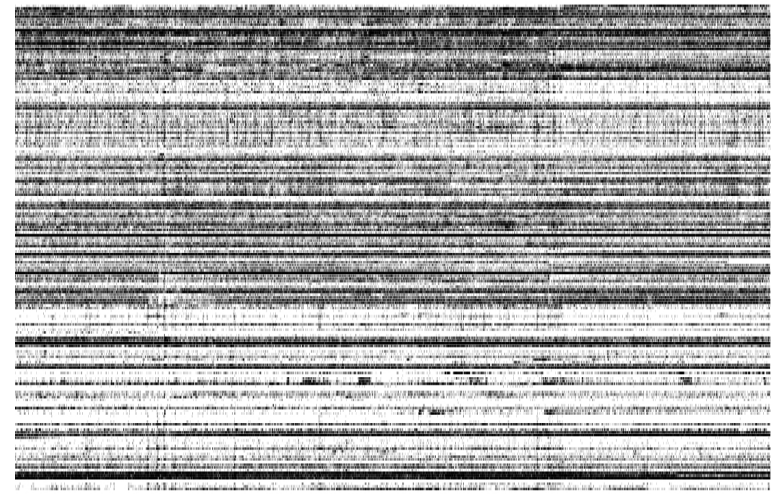
Alana Darcher

MLCoLearn Pitch

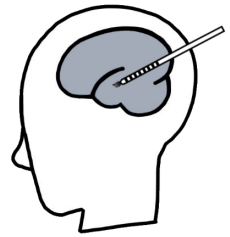
28 Feb 24

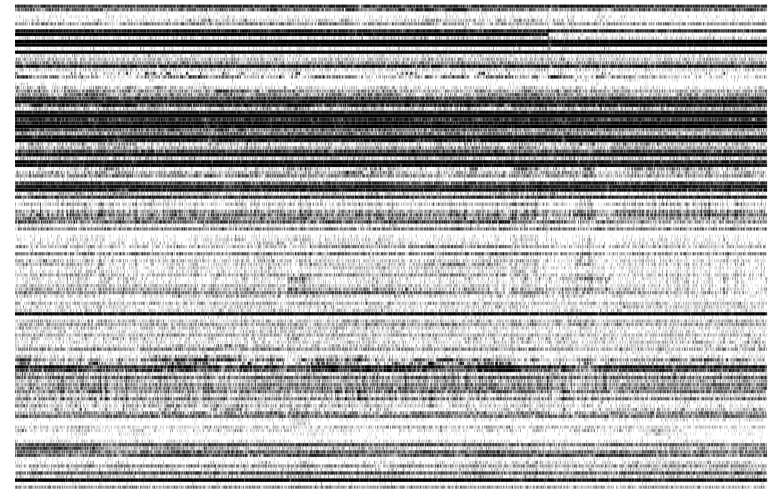
A





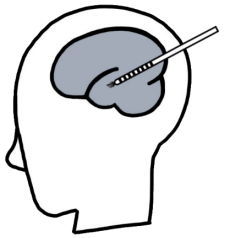
B

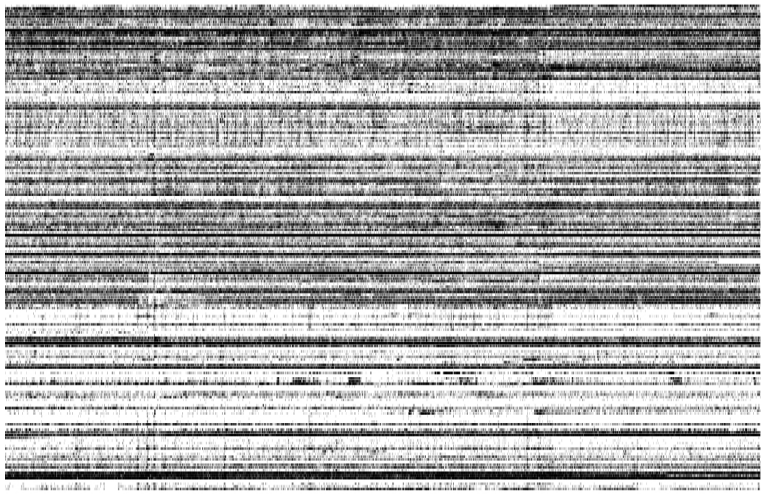




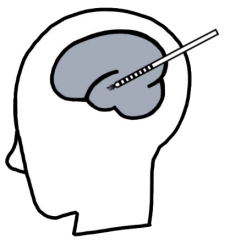
Can you tell that these two people watched the same movie, based on the neural activity?

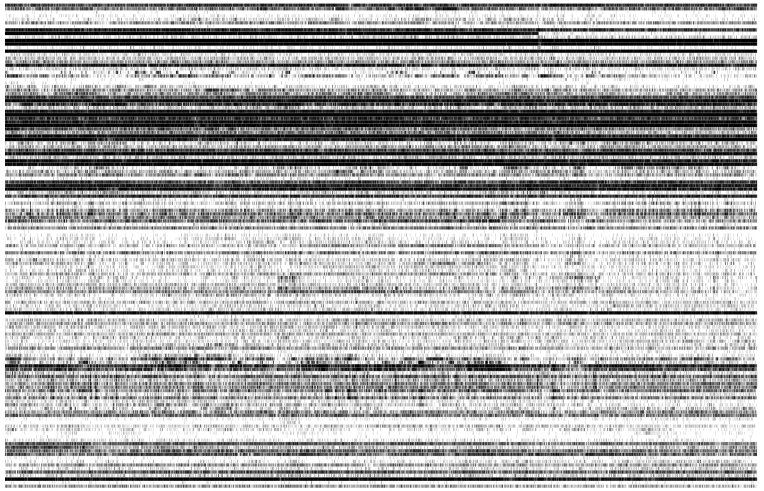
A





B





Canonical Correlation Analysis (CCA)

- *Relations Between Two Sets of Variates,* Harold Hotelling, 1936
- multivariate statistical method used to analyze correlations between 2 datasets
- identifies the sources of common variation in two high-dimensional sets of variables, from two measurement domains
- makes no assumptions about directionality/flow of information

Variable sets could be:

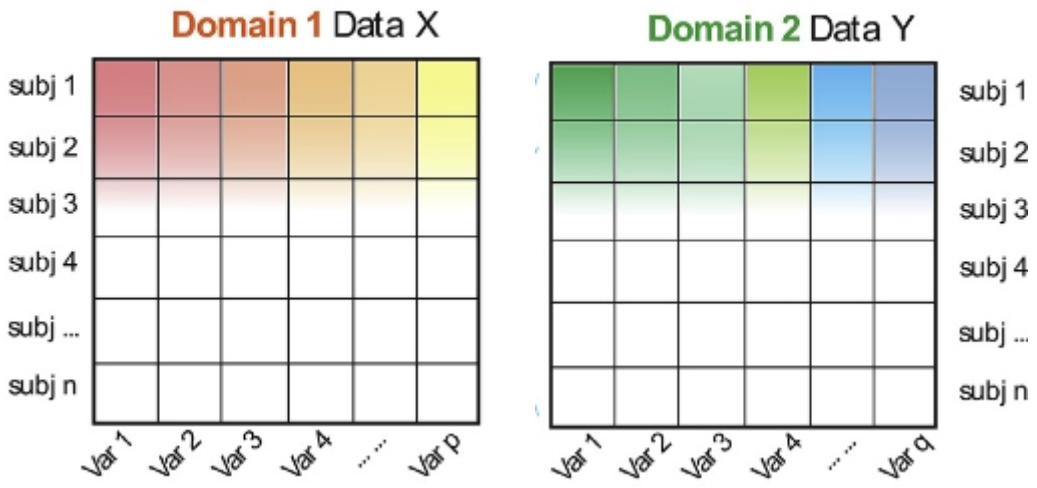
neural activity X behavioral measure

region A X region B

patient 1 X patient 2

spikes X calcium imaging

....



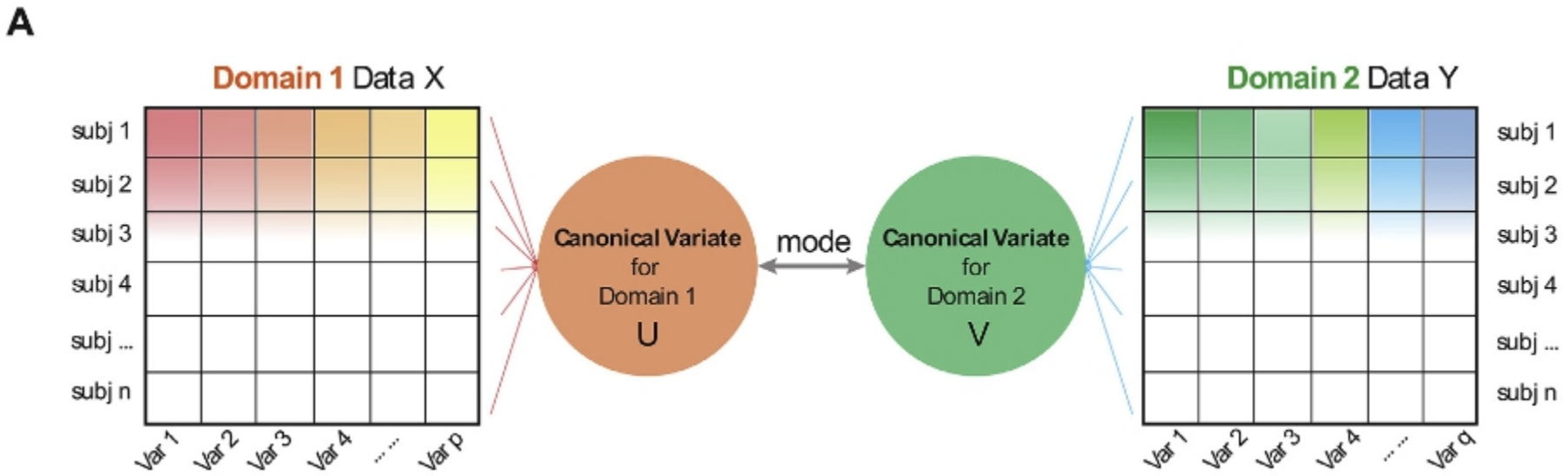
2 datasets, different measurement domains.

Goal:
re-express the datasets as multiple pairs
of highly-correlated latent embeddings.

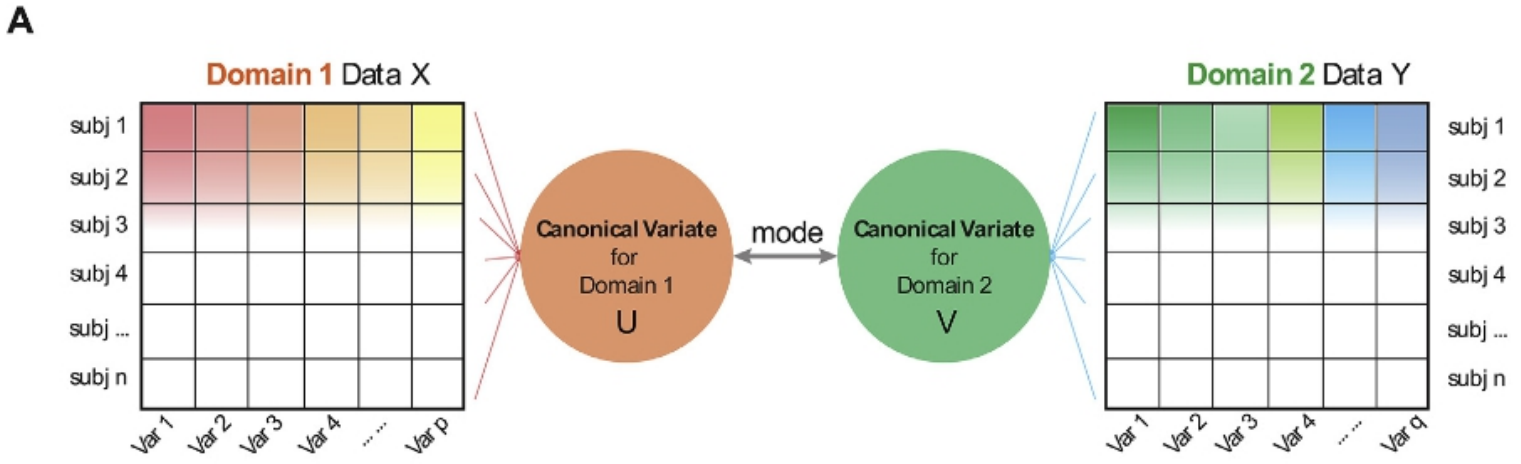
Wang et al. 2020

2 datasets, different measurement domains.

Goal:
re-express the datasets as multiple pairs
of highly-correlated latent embeddings.
(i.e., common decomposition of 2 matrices)



Wang et al. 2020



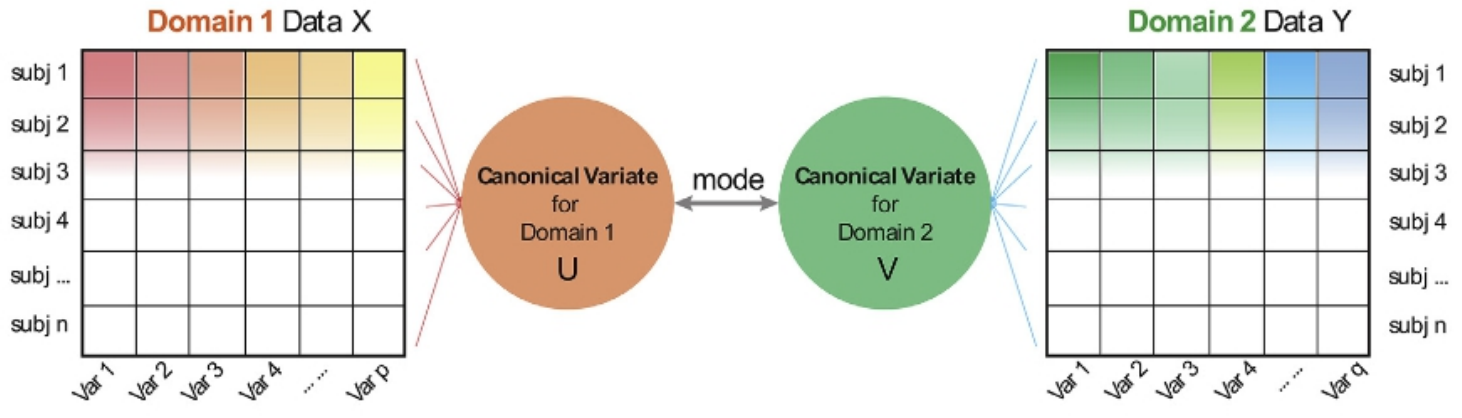
Canonical Variate (U,V):
a linear combination of a set of variables from 1 of the domains.

Canonical Vector (a,b):
coefficients of the above linear combination

mode:
a pair of latent embeddings between each canonical variate

Wang et al. 2020

A



$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

X, Y chosen such that

$$p \leq q$$

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_q \end{pmatrix}$$

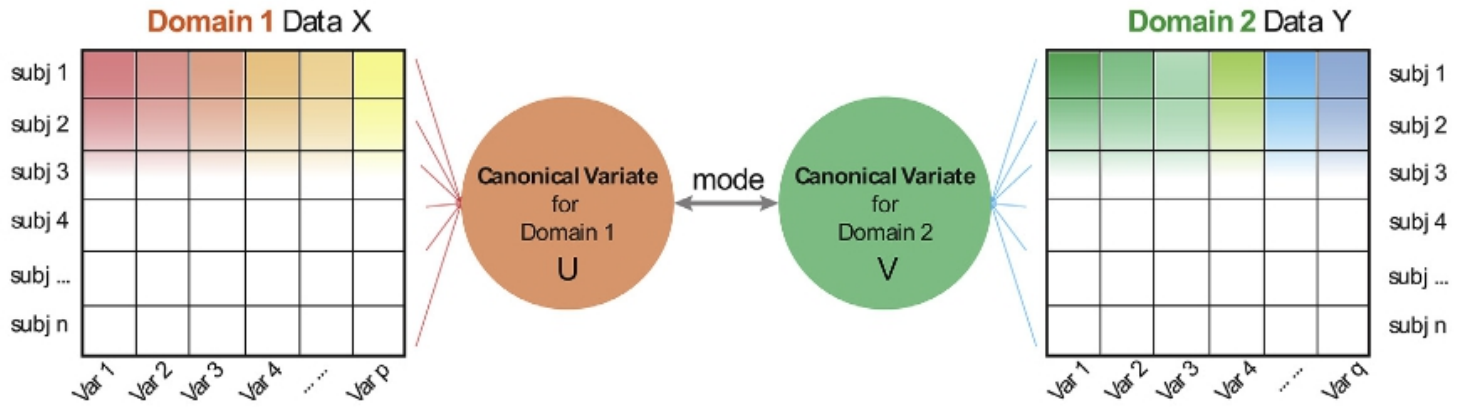
Canonical Variate (U,V):
a linear combination of a set of variables from 1 of the domains.

Canonical Vector (a,b):
coefficients of the above linear combination

mode:
a pair of latent embeddings between each canonical variate

Wang et al. 2020

A



Canonical Variate (U,V):
a linear combination of a set of variables from 1 of the domains.

Canonical Vector (a,b):
coefficients of the above linear combination

mode:
a pair of latent embeddings between each canonical variate

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_q \end{pmatrix}$$

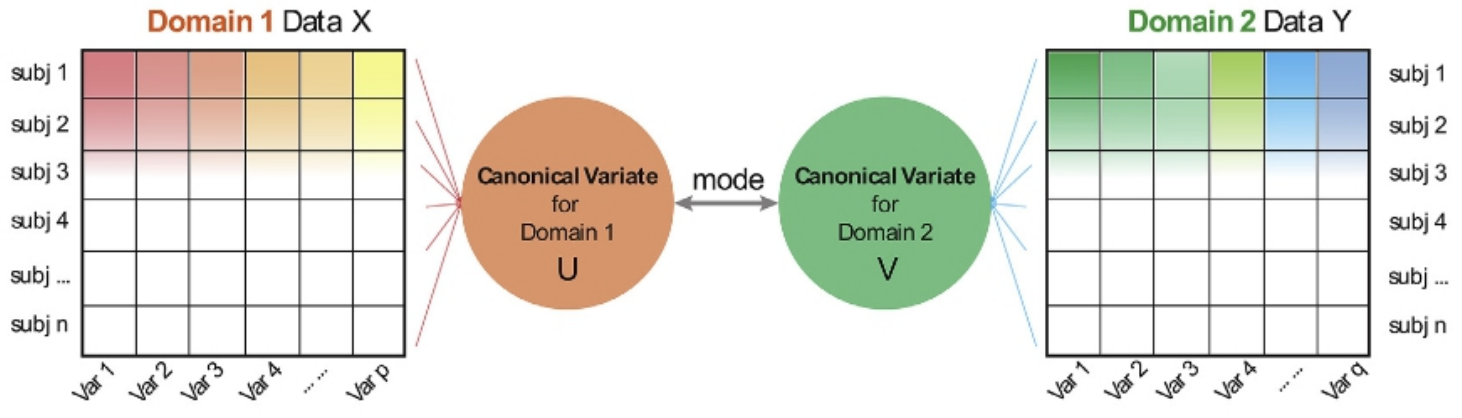
$$\begin{aligned} U_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ U_2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\vdots \\ U_p &= a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \end{aligned}$$

$$\begin{aligned} V_1 &= b_{11}Y_1 + b_{12}Y_2 + \cdots + b_{1q}Y_q \\ V_2 &= b_{21}Y_1 + b_{22}Y_2 + \cdots + b_{2q}Y_q \\ &\vdots \\ V_p &= b_{p1}Y_1 + b_{p2}Y_2 + \cdots + b_{pq}Y_q \end{aligned}$$

$$(U_i, V_i)$$

Wang et al. 2020
https://online.stat.psu.edu/stat505/book/export/html/682

A



Canonical Variate (U,V):
a linear combination of a set of variables from 1 of the domains.

Canonical Vector (a,b):
coefficients of the above linear combination

mode:
a pair of latent embeddings between each canonical variate

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_q \end{pmatrix}$$

$$\begin{aligned} U_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ U_2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\vdots \\ U_p &= a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \end{aligned}$$

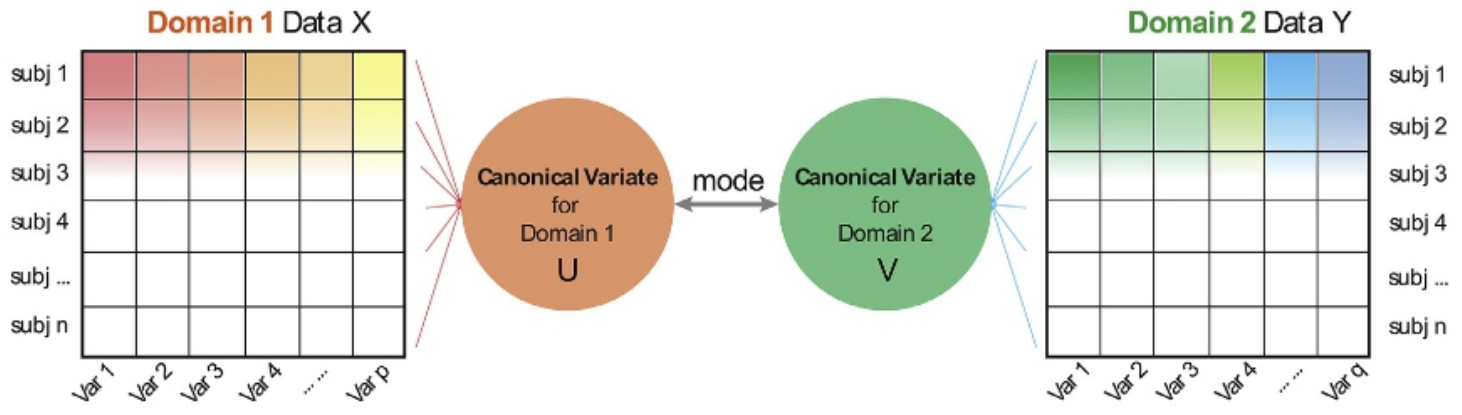
$$\begin{aligned} V_1 &= b_{11}Y_1 + b_{12}Y_2 + \cdots + b_{1q}Y_q \\ V_2 &= b_{21}Y_1 + b_{22}Y_2 + \cdots + b_{2q}Y_q \\ &\vdots \\ V_p &= b_{p1}Y_1 + b_{p2}Y_2 + \cdots + b_{pq}Y_q \end{aligned}$$

Want to find linear combinations that maximize the correlations between a given U_i and V_i.

$$(U_i, V_i)$$

Wang et al. 2020
https://online.stat.psu.edu/stat505/book/export/html/682

A



Canonical Variate (U,V):
a linear combination of a set of variables from 1 of the domains.

Canonical Vector (a,b):
coefficients of the above linear combination

mode:
a pair of latent embeddings between each canonical variate

Specific Goal: to find linear combinations that maximize the correlations between a given U_i and V_i.

$$(U_i, V_i) \qquad p \leq q$$

For a given *i*, compute the variance of U_i , V_i:

$$\mathrm{var}(U_i) = \sum_{k=1}^{p} \sum_{l=1}^{p} a_{ik} a_{il} cov(X_k, X_l)$$

For a given *i, j*, compute the covariance of U_i , V_j:

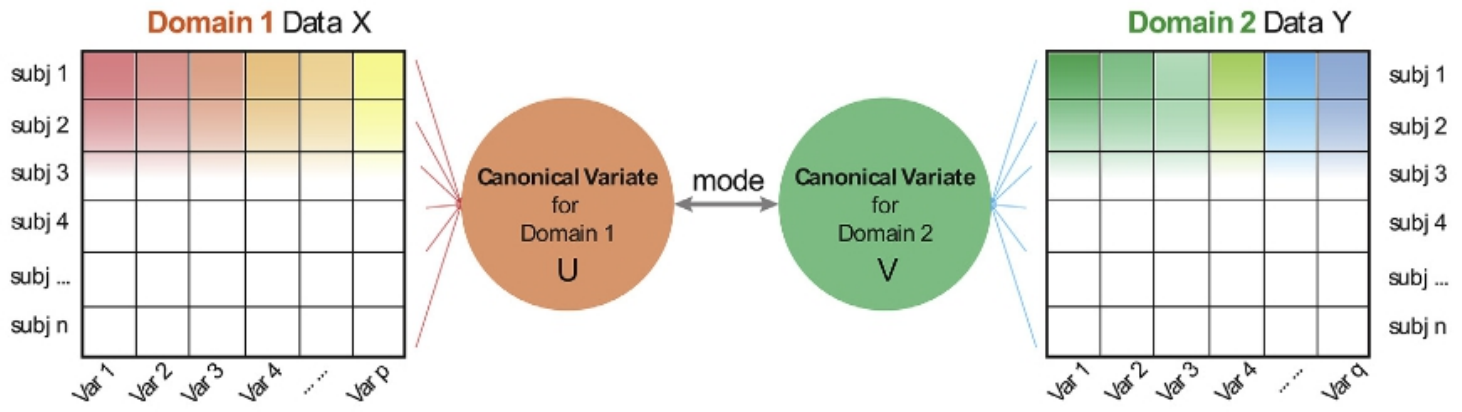
$$\mathrm{cov}(U_i, V_j) = \sum_{k=1}^{p} \sum_{l=1}^{q} a_{ik} b_{jl} \mathrm{cov}(X_k, Y_l)$$

Maximize the correlation between pairs of canonical variates:

$$\rho_i^* = \frac{\mathrm{cov}(U_i, V_i)}{\sqrt{\mathrm{var}(U_i)\mathrm{var}(V_i)}}$$

Wang et al. 2020
https://online.stat.psu.edu/stat505/book/export/html/682

A



Canonical Variate (U,V):
a linear combination of a set of variables from 1 of the domains.

Canonical Vector (a,b):
coefficients of the above linear combination

mode:
a pair of latent embeddings between each canonical variate

Specific Goal: to find linear combinations that maximize the correlations between a given U_i and V_i.

$$(U_i, V_i) \qquad p \leq q$$

For a given *i*, compute the variance of U_i , V_i:

$$\mathrm{var}(U_i) = \sum_{k=1}^{p} \sum_{l=1}^{p} a_{ik} a_{il} cov(X_k, X_l)$$

For a given *i, j*, compute the covariance of U_i , V_j:

$$\mathrm{cov}(U_i, V_j) = \sum_{k=1}^{p} \sum_{l=1}^{q} a_{ik} b_{jl} \mathrm{cov}(X_k, Y_l)$$

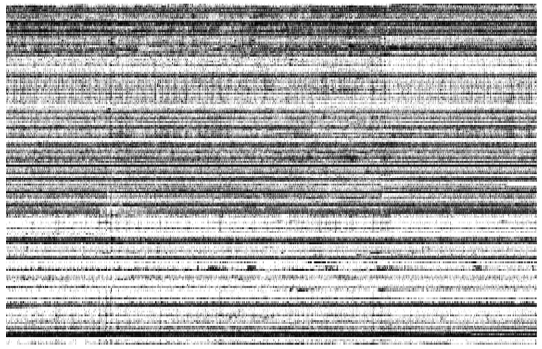
Maximize the correlation between pairs of canonical variates:

$$\rho_i^* = \frac{\mathrm{cov}(U_i, V_i)}{\sqrt{\mathrm{var}(U_i)\mathrm{var}(V_i)}}$$

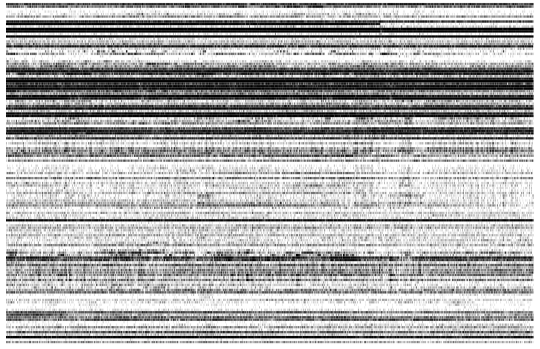
Wang et al. 2020
https://online.stat.psu.edu/stat505/book/export/html/682

Goal today:

- Implement CCA from scratch in Python.
- Test out the implementation with simulated data to see how reliably modes are recovered.
- Apply to human single-neurons recorded during a naturalistic stimulus.



```
[[0. 1. 7. ... 0. 0. 0.]
 [0. 1. 1. ... 0. 1. 1.]
 [1. 0. 6. ... 0. 0. 0.]
 ...
 [0. 1. 3. ... 1. 0. 0.]
 [0. 1. 6. ... 0. 0. 0.]
 [0. 0. 3. ... 1. 0. 0.]]
```



```
[[ 2.  1.  0. ...  0.  1.  1.]
 [10. 11.  3. ...  2.  8.  4.]
 [ 1.  1.  3. ...  3.  2.  1.]
 ...
 [ 0.  0.  0. ...  0.  0.  0.]
 [ 0.  1.  0. ...  2.  2.  1.]
 [ 1.  0.  1. ...  5.  3.  4.]]
```

Binned spiking data:
bin_size = 1 s
p x n = 30 x 4726
q x n = 109 x 4726

Resources:

- sklearn Implementations:
  - https://scikit-learn.org/stable/modules/generated/sklearn.cross_decomposition.CCA.html
  - https://scikit-learn.org/stable/modules/cross_decomposition.html#cross-decomposition

- background:
  - https://online.stat.psu.edu/stat505/book/export/html/682
  - https://www.sciencedirect.com/science/article/pii/S105381192030232 9

- solution:
  - https://gregorygundersen.com/blog/2018/07/17/cca/