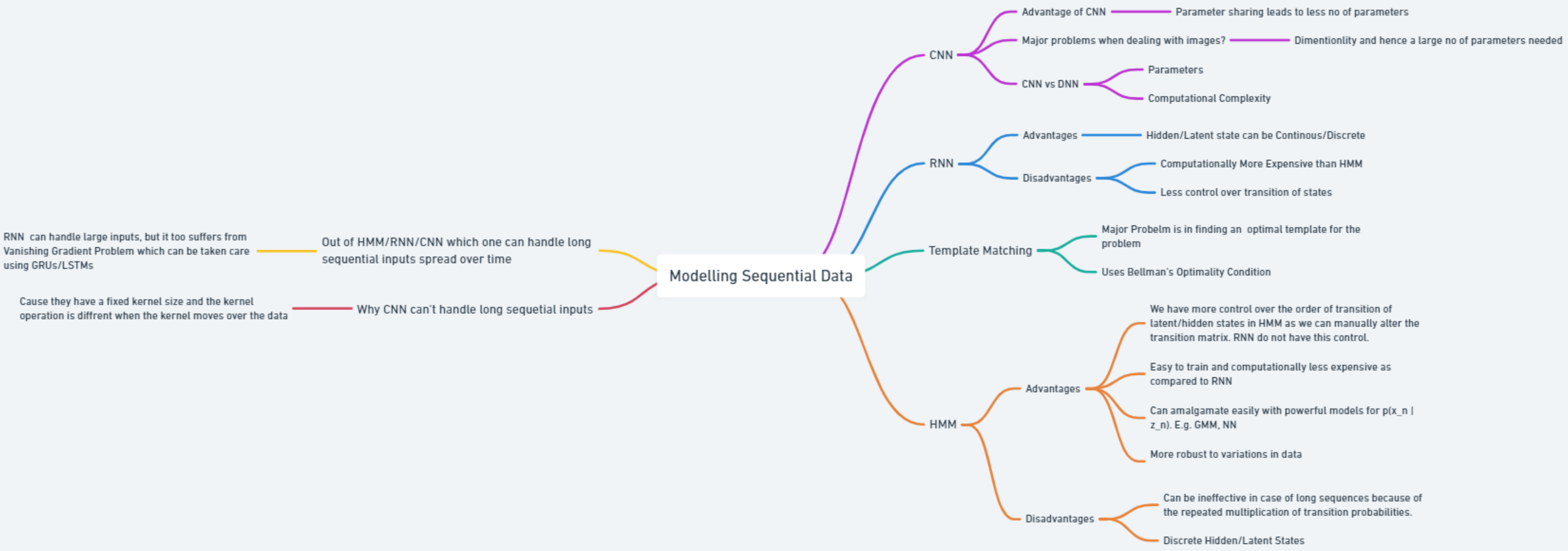

HMM
Sequence Prediction(Decoding)
Viterbi Algo
Baum-Welch Algo
Why its popularity decresing in comparison to NN based methods
GMM vs HMM
Both are proabilistic models
GMM cannot handle sequential inputs while HMM can handle sequential inputs
Both trained using EM algorithm, but the E and M steps are different for them
Training of these is done
Initial State Probability
Emission Probabilities
Transition Probabilities
What do we get after training?
Values of Initial State Probs, Trans Prob, Emission Probs
What do we train actually while training a model?
We optimize the value of Initial State Probs, Trans Prob, Emission Probs, to find our training data as good as possible.
We use Log Likelihood as a metric to judge the extent to which the HMM fits the training data


Modelling Sequential Data
CNN
Advantage of CNN
Parameter sharing leads to less no of parameters
Major problems when dealing with images?
Dimentionlity and hence a large no of parameters needed
CNN vs DNN
Parameters
Computational Complexity
RNN
Advantages
Hidden/Latent state can be Continous/Discrete
Disadvantages
Computationally More Expensive than HMM
Less control over transition of states
Template Matching
Major Probelm is in finding an optimal template for the problem
Uses Bellman's Optimality Condition
HMM
Advantages
We have more control over the order of transition of latent/hidden states in HMM as we can manually alter the transition matrix. RNN do not have this control.
Easy to train and computationally less expensive as compared to RNN
Can amalgamate easily with powerful models for p(x_n | z_n). E.g. GMM, NN
More robust to variations in data
Disadvantages
Can be ineffective in case of long sequences because of the repeated multiplication of transition probabilities.
Discrete Hidden/Latent States
Out of HMM/RNN/CNN which one can handle long sequential inputs spread over time
RNN can handle large inputs, but it too suffers from Vanishing Gradient Problem which can be taken care using GRUs/LSTMs
Why CNN can't handle long sequetial inputs
Cause they have a fixed kernel size and the kernel operation is diffrent when the kernel moves over the data


Comparing Two PDFs
Entropy
Cross Entropy
KL Divergence
Mutual Information
Binary Cross Entropy
BInary Entropy vs Cross Entropy
Binary Cross Entropy used with Multi-Label Classification
Cross Entropy used with Multi-Class Classification

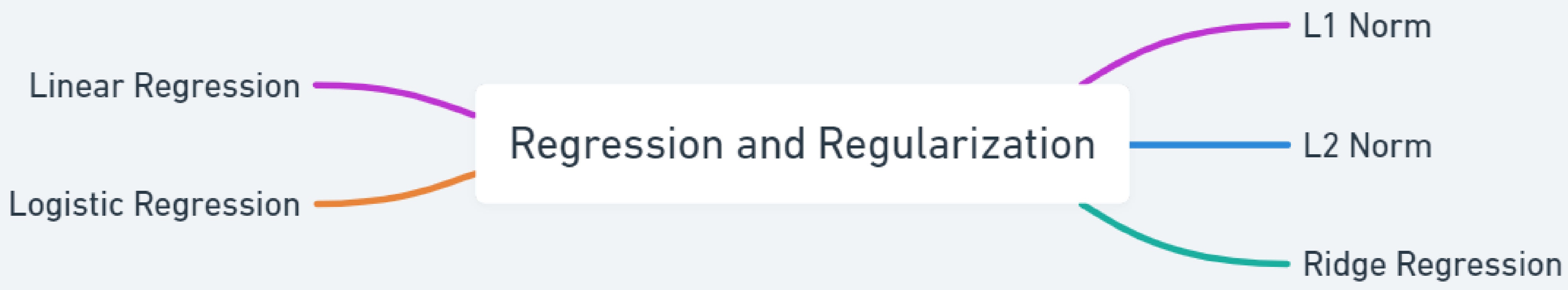
Estimation

- MLE
  - $$\mu, \sigma = \underset{\mu,\sigma}{\operatorname{argmax}} \ln p(t_1, \ldots, t_N | \mu, \sigma)|_{t_1=s_1,\ldots,t_N=s_N}$$
- MAP
  - $$\mu, \sigma = \underset{\mu,\sigma}{\operatorname{argmax}} \ln p(\mu, \sigma | t_1, \ldots, t_N)|_{t_1=s_1,\ldots,t_N=s_N}$$
- Bayersian Inference
  - Consider a probablistic model A which represent data with parameters mu and sigma which need to be estimated
    - Instead of estimating the exact value of these parameters mu and sigma using MAP/MLE , we instead model these parameters again as distributions, whose parameters we need to estimate
      - Find Likelihood of model(a)parameters

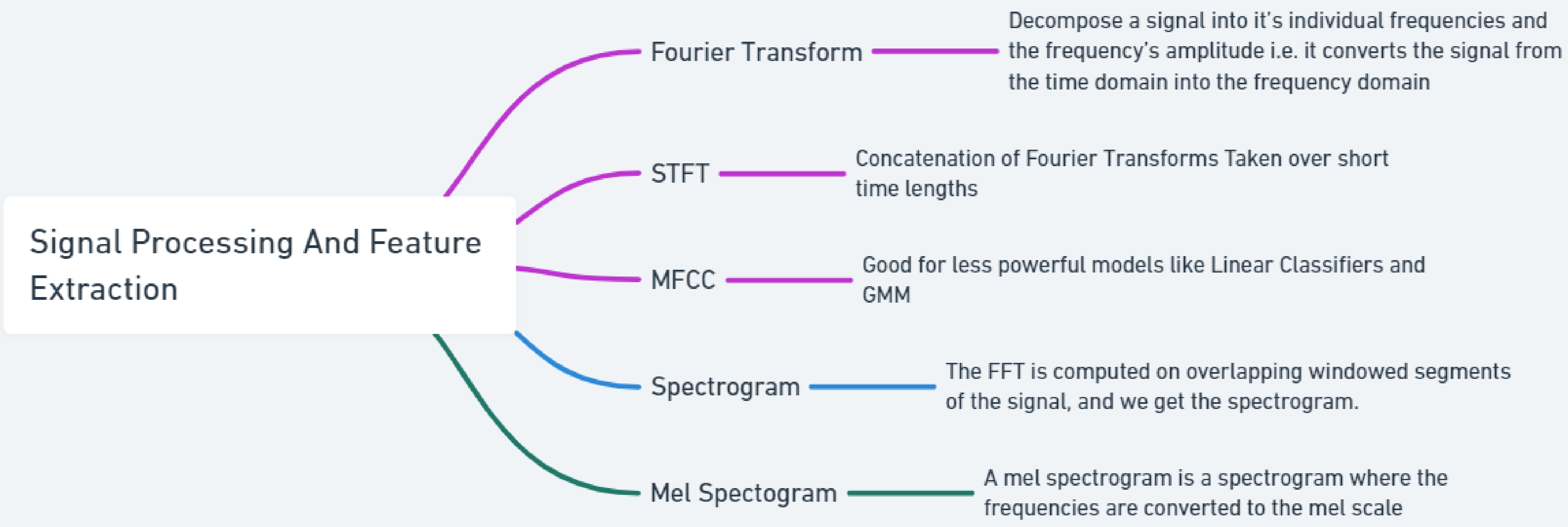

Methods for Clustered Data
Hard Clustering
K Means
Hard Clustering
Step 1: Choose the number of clusters k
Step 2: Select k random points from the data as centroids
Step 3: Assign all the points to the closest cluster centroid
Step 4: Recompute the centroids of newly formed clusters
Step 5: Repeat steps 3 and 4
Converge
Soft Clustering
GMM
Soft Clustering
Train using EM algorithm
Model Parameters
Trainable Parameters
Mean Vector
Covariance Matrix
Probabilities for a point to belong to cluster k
Hyperparameters
Number of clusters(k)
Find Trainable Parameters by MLE

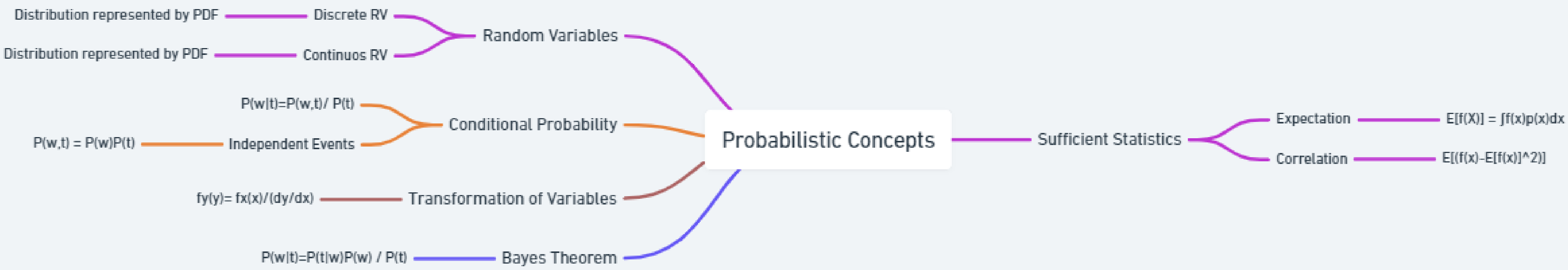
Latent Variables

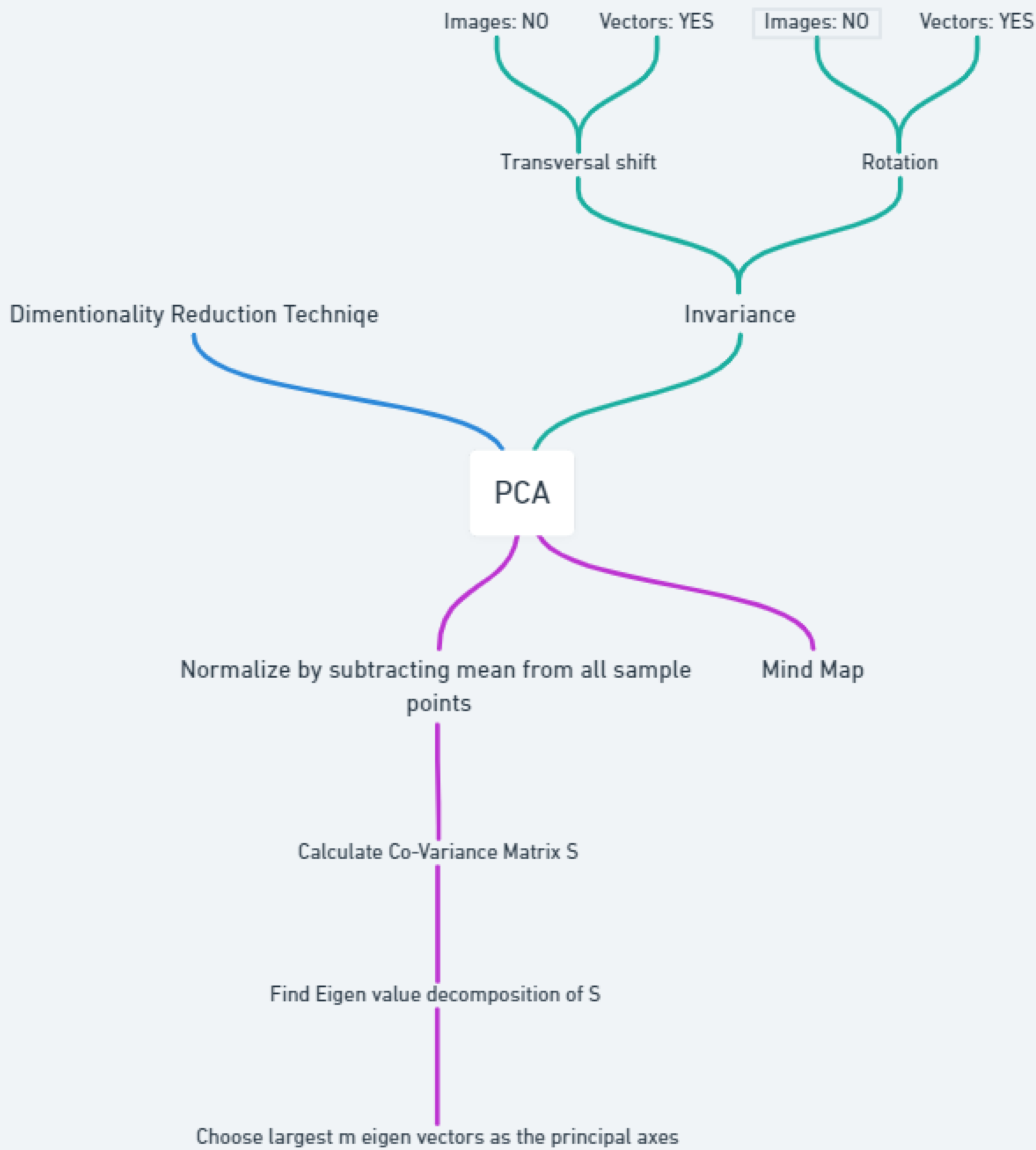
- Not directly observed but are rather inferred through a mathematical model from other variables that are observed
- Latent variables help complicated distributions to be formed from simpler components
- Mathematical models that aim to explain observed variables in terms of latent variables are called latent variable models.
  - GMM
  - HMM
- Physical Manifestation of Latent Variables
  - They represent the quantities to be estimated in HMMs when we insert the data


Regression and Regularization
Linear Regression
Logistic Regression
L1 Norm
L2 Norm
Ridge Regression


Non Linear Models
Multiclass classification
Out of multiple possible classes, only one class is true
Sigmoid Non-Linearity is applied on the last layer
Loss Function: Categorical Cross Entropy
Multilabel classification
Multiple classes can be present simultaneuosly
Softmax Non-Linearity is applied on the last layer
Loss Function: Binary Cross Entropy


Signal Processing And Feature Extraction
Fourier Transform
Decompose a signal into it's individual frequencies and the frequency's amplitude i.e. it converts the signal from the time domain into the frequency domain
STFT
Concatenation of Fourier Transforms Taken over short time lengths
MFCC
Good for less powerful models like Linear Classifiers and GMM
Spectrogram
The FFT is computed on overlapping windowed segments of the signal, and we get the spectrogram.
Mel Spectogram
A mel spectrogram is a spectrogram where the frequencies are converted to the mel scale


Probabilistic Concepts
Random Variables
Discrete RV
Distribution represented by PDF
Continuos RV
Distribution represented by PDF
Conditional Probability
P(w|t)=P(w,t)/ P(t)
Independent Events
P(w,t) = P(w)P(t)
Transformation of Variables
fy(y)= fx(x)/(dy/dx)
Bayes Theorem
P(w|t)=P(t|w)P(w) / P(t)
Sufficient Statistics
Expectation
E[f(X)] = ∫f(x)p(x)dx
Correlation
E[(f(x)-E[f(x)]^2)]


Images: NO
Vectors: YES
Images: NO
Vectors: YES
Transversal shift
Rotation
Dimentionality Reduction Techniqe
Invariance
PCA
Normalize by subtracting mean from all sample points
Mind Map
Calculate Co-Variance Matrix S
Find Eigen value decomposition of S
Choose largest m eigen vectors as the principal axes