**Youtube Video:** (https://youtu.be/rc3YDj5GiVM)

Image segmentation: input is pixel values and processed features. Target variable (y) is a class for every pixel.

Text processing: input is words. Output are the labels of words (e.g. person, organisation, etc.).

If you have features with a lot of redundancy by using naïve bayes you are ignoring the correlation. This can lead to a skewed distribution. However, if you start adding edges to the graphical model it will lead to a densely connected model.

CRFs: we don’t care about the distribution over x. Instead we model p(y|x) instead of p(y,x).

CRFs are similar to the logistic model. The logisitic model is a very simple CRF.

Because we are modelling as a conditional distribution it focuses on how the x’s come together to effect y. By focuses just on the target variable it means we can ignore correlations.

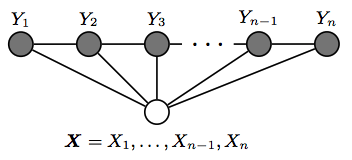
CRFs for language. We have correlated feautes, e.g. capitalised word will be correlated with the fact that word is in the Atlas, or follows ‘Mrs’. We don’t distribution over words in sentence but rather p(label | words).

CRF is similar to a Gibbs distribution but is normalised so that we have a conditional distribution.

**Blog Post:** (<http://www.davidsbatista.net/blog/2017/11/13/Conditional_Random_Fields/>)

MEMMs have a model to compute the probability of the next state, given the current state and the observation. CRFs compute all state transitions globally in a single model. The idea behind CRFs is to drop the local per state normalisation and replace it with a global per sequence normalisation.

A CRF can be seen as an undirected graphical model, or Markov Random Field, globally conditioned on X (the random variable representing the observation sequence)

****

Inference with a CRF is the same as computing the sequence that maximises the following equation:

The denominator, also called the partition function, is useful to compute a marginal probability. This is useful for measuring the model’s confidence in it’s predicted labelling over a segment of input.

Parameter estimation: we maximise the conditional likelihood of the training data:

The most important realisation about CRF features is that they can be arbitrarily complex. This does not affect computational performance, because at inference time, the x is always observed,

**A helpful example:** (<https://ermongroup.github.io/cs228-notes/representation/undirected/>)

**An Introduction to CRFs (Sutton and McCallum):**

CRFs combine the ability to compactly model multivariate data with the ability to leverage a large number of input features for prediction. The advantage to a conditional model is that dependencies that involve only variables in x play no role in the conditional model, so that the conditional model can be much less complicated than the joint model.

**Definitions:**

Hidden Markov Chains: a Markov process with unobservable states.

Markov Random Field: (also Markov network or undirected graphical model)

Information Entropy: is the average rate at which information is produced by a stochastic source. When the source produces a low-probability value, the event. Carries more information.