There are two stages of Mask RCNN. First, it generates proposals about the regions where there might be an object based on the input image. Second, it predicts the class of the object, refines the bounding box and generates a mask in pixel level of the object based on the first stage proposal. Both stages are connected to the backbone structure.

* The region-based Convolutional Neural Network family of models for object detection and the most recent variation called Mask R-CN
* **R-CNN**: Bounding boxes are proposed by the “*selective search*” algorithm, each of which is stretched and features are extracted via a deep convolutional neural network
* **Fast R-CNN**: Simplified design with a single model, bounding boxes are still specified as input, but a region-of-interest pooling layer is used after the deep CNN to consolidate regions and the model predicts both class labels and regions of interest directly.
* Matterport Mask R-CNN Project

What is backbone? Backbone is a [FPN](https://arxiv.org/abs/1612.03144) (Feature Pyramid Network, shows Using FPN in a basic Faster R-CNN system, our method achieves state-of-the-art single-model results on the COCO detection benchmark without bells and whistles, surpassing all existing single-model entries including those from the COCO 2016 challenge winners.) style deep neural network. It consists of a bottom-up pathway, a top-bottom pathway and lateral connections. Bottom-up pathway can be any ConvNet, usually ResNet or VGG, which extracts features from raw images

<https://machinelearningmastery.com/how-to-perform-object-detection-in-photographs-with-mask-r-cnn-in-keras/>

<https://medium.com/@alittlepain833/simple-understanding-of-mask-rcnn-134b5b330e95>