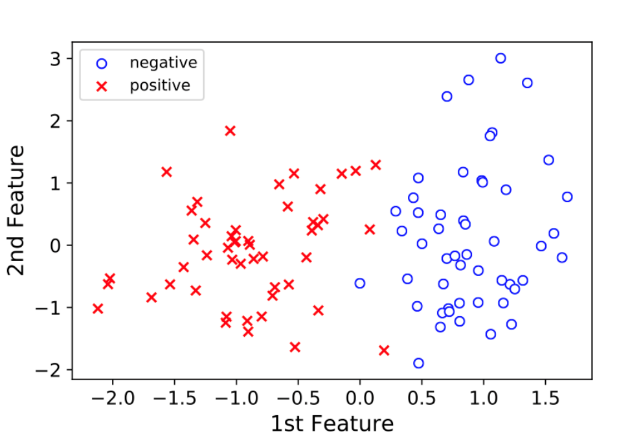
## Week 1

1. What is necessary for supervised machine learning?
   1. A model
   2. Labelled training data
   3. Learning from data
2. What decision boundary can logistic regression provide?
   1. Linear
3. What is the primary advantage of using multiple filters?
   1. This allows the model to look for subtypes of the classification.
4. What best describes transfer learning in the context of document analysis?
   1. Parameters at the bottom of the model are transferable across all people and documents, while the parameters at the top are different between individuals.
5. Given the following image of data classifications, what model would you choose?



* 1. Logistic regression

1. What new feature did neural networks acquire in 2010?
   1. A new name: Deep Learning
2. What is convolved with layer 2 features, or sub-motifs?
   1. Layer 1 feature map
3. What gives the best conceptual meaning of convolution?
   1. Shifting a filter to every location in an image.
4. What does transfer learning mean in the context of medical imaging?
   1. Weights of convolutional layers learned from ImageNet transfer to medical images, so we only need learn new parameters at the top of the network.
5. What is the primary advantage of having a deep architecture?
   1. The model shares knowledge between motifs through their shared substructures.

## Week 2

1. What does the equation for the loss function do conceptually?
2. Penalise overconfidence
3. What is overfitting?
4. Model complexity fits too well to training data and will not generalize in the real-world.
5. Why should the test set only be used once?
   1. More than one use can lead to bias.
6. Which two of the following describe the purpose of a validation set?
   1. To estimate the performance of a model.
   2. To pick the best performing model.
7. How do we learn our network?
8. Gradient descent
9. What technique is used to minimize loss for a large data set?
10. Stochastic gradient descent
11. What are benefits of stochastic gradient descent?
12. Stochastic gradient descent gets near the solution quickly.
13. Stochastic gradient descent can update many more times than gradient descent.
14. With stochastic gradient descent, the update time does not scale with data size.
15. Why is gradient descent computationally expensive for large data sets?
16. Calculating the gradient requires looking at every single data point.
17. What are the two main benefits of early stopping?
18. It helps save computation cost.
19. It performs better in the real world.
20. Why are optimisation and validation at odds?
21. Optimization seeks to do as well as possible on a training set, while validation seeks to generalize to the real world.

## Week 3

1. What indicates whether a doctor or machine is doing well at finding positive examples in a data set?
   1. Sensitivity
2. What is used to distinguish the false positive rate from the false negative rate?
   1. Specificity
3. What is the best conceptual definition of one-dimensional convolution?
   1. “Sliding” of two signals, where a matched feature gives a high value of convolution.
4. What can a user choose when designing a convolutional layer?
   1. Filter stride
   2. Filter number
   3. Filter size
5. What is a fully connected readout?
   1. A layer with a single neuron for each output class.
6. Why are nonlinear activation functions preferable?
   1. Nonlinear activation functions increase the functional capacity of the neural network by allowing the representation of nonlinear relationships between features in input.
7. What are benefits of pooling?
   1. Reduces computational complexity.
   2. Combats overfitting.
   3. Encourages translational invariance.
8. How are parameters that minimize the loss function found in practice?
   1. Stochastic gradient descent
9. What is an advantage of hierarchical representation of image features?
   1. Better leveraging all training data.
10. Why does transfer learning work?
    1. Top-level features are specialized for a particular task, while low-level features are universal to all images.

## Week 4

1. What is meant by “word vector”?
   1. A vector of numbers associated with a word.
2. What word is a synonym for “word vector”?
   1. Embedding
3. What is the term for a set of vectors, with one vector for each word in the vocabulary?
   1. Codebook
4. What is natural language processing?
   1. Taking natural text and making inferences and predictions.
5. What is the goal of learning word vectors?
   1. Given a word, predict which words are in its vicinity.
6. What function is the generalization of the logistic function to multiple dimensions?
   1. Softmax function
7. What is the continuous bag of words (CBOW) approach?
   1. Vectors for the neighbourhood of words are averaged and used to predict word n.
8. What is the Skip-Gram approach?
   1. Word n is used to predict the words in the neighbourhood of word n.
9. What is the goal of the recurrent neural network?
   1. Synthesise a sequence of words.
10. Which model is the state-of-the-art for text synthesis?
    1. Long short-term memory