Categorization Methods: “Automatic”

- Use **human experts** to
  - Annotate data sets with **category labels** ➔ Training data
  - Provide a set of **features** to represent each text object that can potentially provide a “clue” about the category
- Use **machine learning** to learn “soft rules” for categorization from the training data
  - Figure out **which features are most useful** for separating different categories
  - **Optimally combine the features** to **minimize the errors** of categorization on the training data
  - The trained classifier can then be applied to a new text object to predict the most likely category (that a human expert would assign to it)