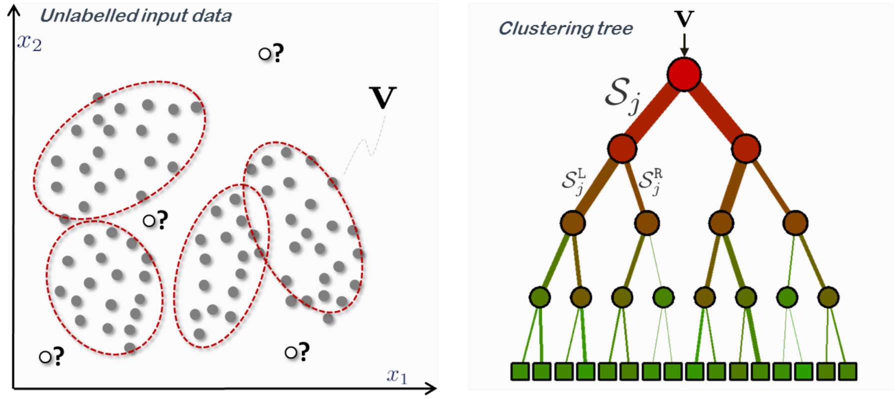

Fig. 5.1: **Input data and density forest training. (a)** Unlabelled data points using for training a density forest are shown as dark circles. White circles indicate previously unseen test data. **(b)** Density forests are ensembles of clustering trees.

*Given a set of unlabelled observations we wish to estimate the probability density function from which such data has been generated.*

Each input data point $\mathbf{v}$ is represented as usual as a multi-dimensional feature response vector $\mathbf{v} = (x_1, \cdots, x_d) \in \mathbb{R}^d$. The desired output is the entire probability density function $p(\mathbf{v}) \geq 0 \;\; s.t. \int p(\mathbf{v}) d\mathbf{v} = 1$, for any generic input $\mathbf{v}$. An explanatory illustration is shown in fig. 5.1a. Unlabelled training data points are denoted with dark circles, while white circles indicate previously unseen test data.

**What are density forests?** A density forest is a collection of randomly trained clustering trees (fig. 5.1b). The tree leaves contain simple prediction models such as Gaussians. So, loosely speaking a density forest can be thought of as a generalization of Gaussian mixture models (GMM) with two differences: (i) multiple hard clustered data partitions are created, one by each tree. This is in contrast to the single "soft"