Data Visualization

Part 2

Jonathan Arriaga, PhD

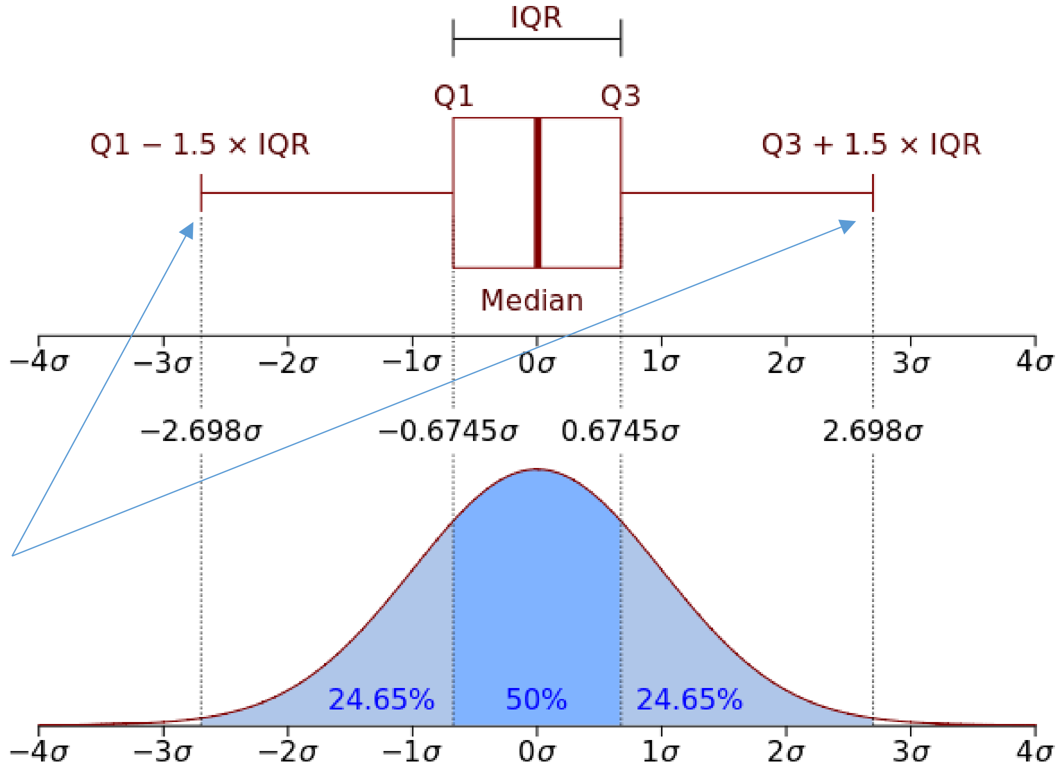
Index

- Box Plots
- Histograms
- Density Plots
- Principal Components Analysis to Visualize High-Dimensional Data

Box Plots

- Non-parametric: no assumptions are made about the distribution of data.
- Depict numerical data using its quartiles.

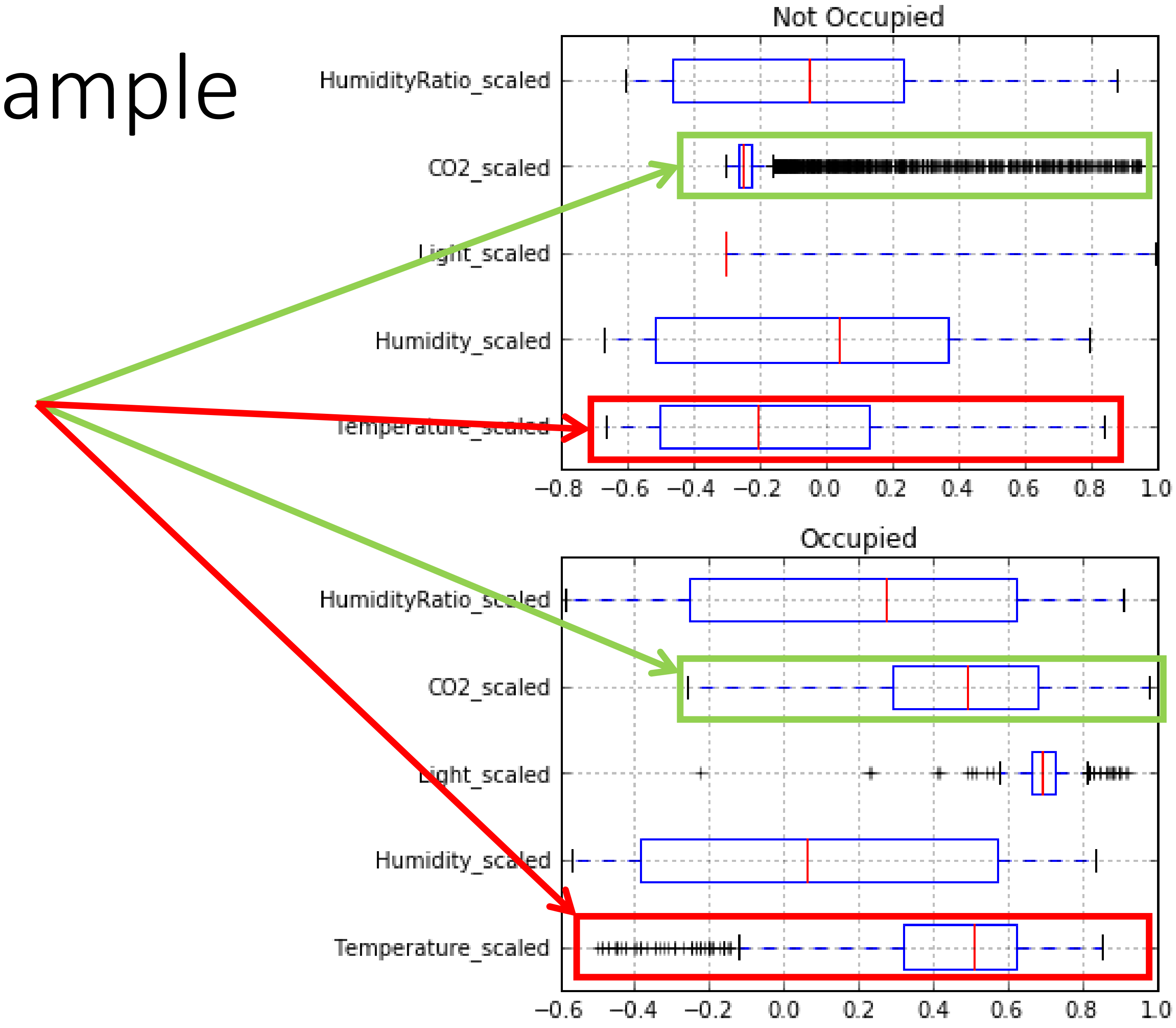
The end of the whiskers represent the minimum and maximum data points within 1.5 IQR from the lower quartile and the upper quartile respectively



https://en.wikipedia.org/wiki/Box_plot#/media/File:Boxplot_vs_PDF.svg

Box Plots Example

Easily spot variations in the sampled values of different groups.



Dataset: Occupancy Detection

Features used:

- HumidityRatio
- CO2
- Light
- Humidity
- Temperature
- Occupancy

Source:

http://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+

Histograms

- Very useful to visualize the distribution of data.
- The entire range of values is divided in a series of intervals, or “bins”;

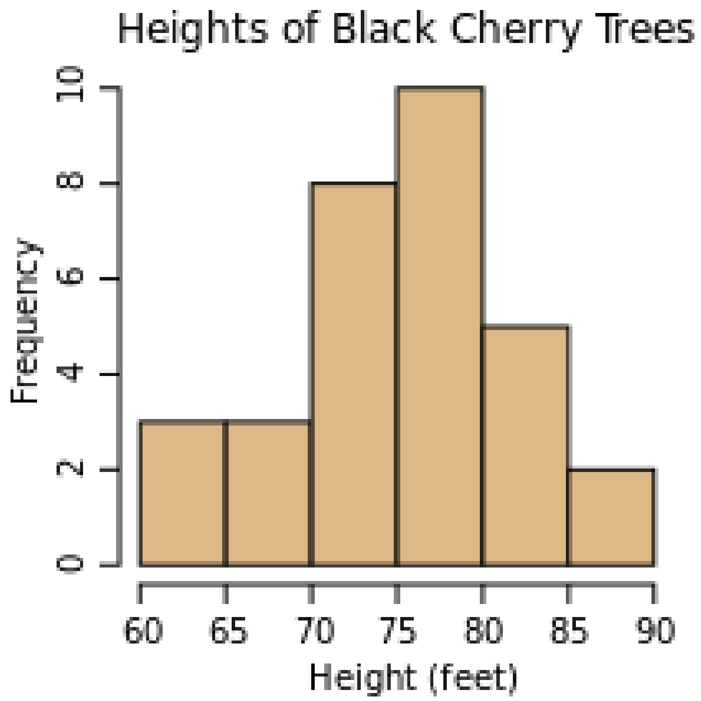
then count how many values fall within each bin.

- One way (of many) to estimate the appropriate number

of bins k is

$$k = \sqrt{n}$$

where n is the number of observations.



CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=3483039

Histogram Example

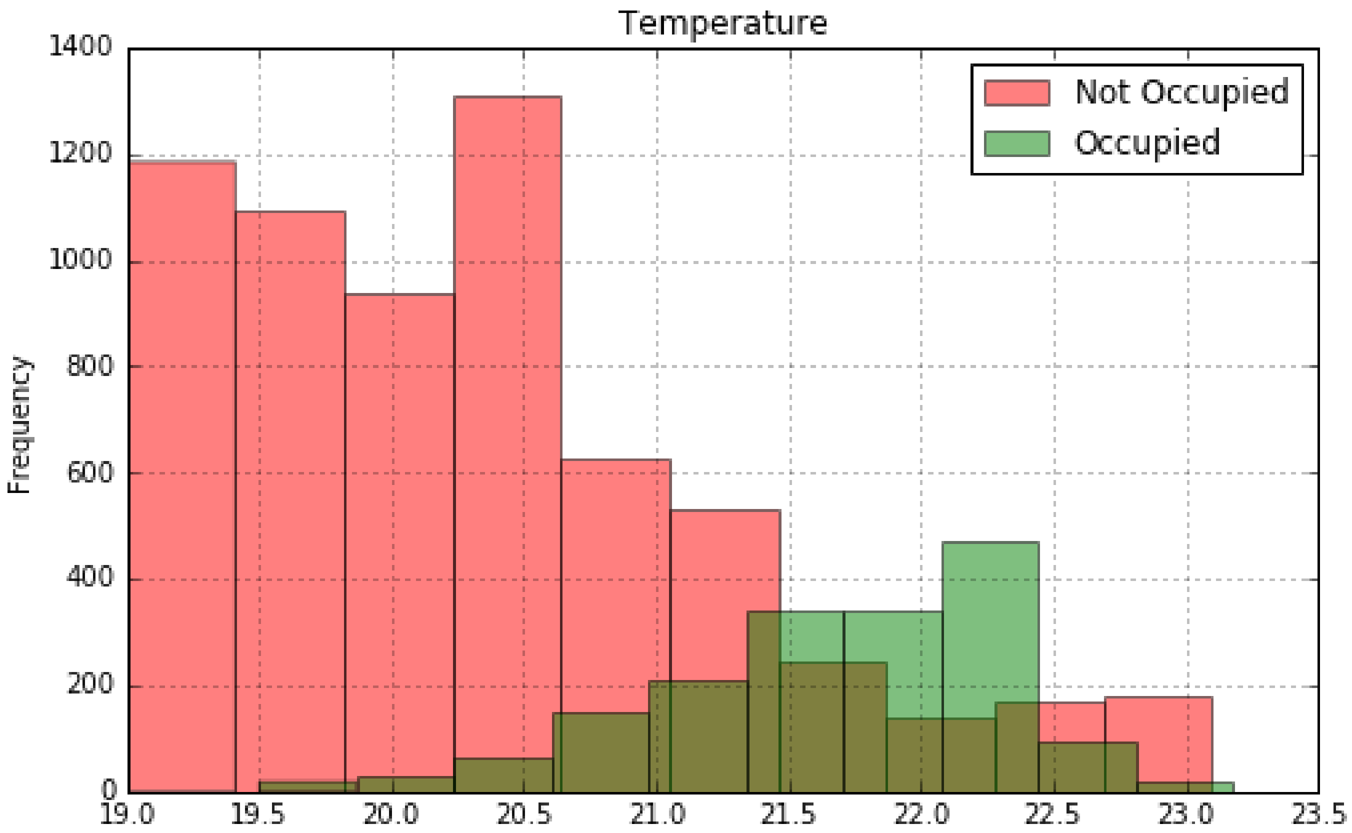
Dataset: Occupancy Detection

Features used:

- Temperature
- Occupancy

Source:

http://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+



Density Plots

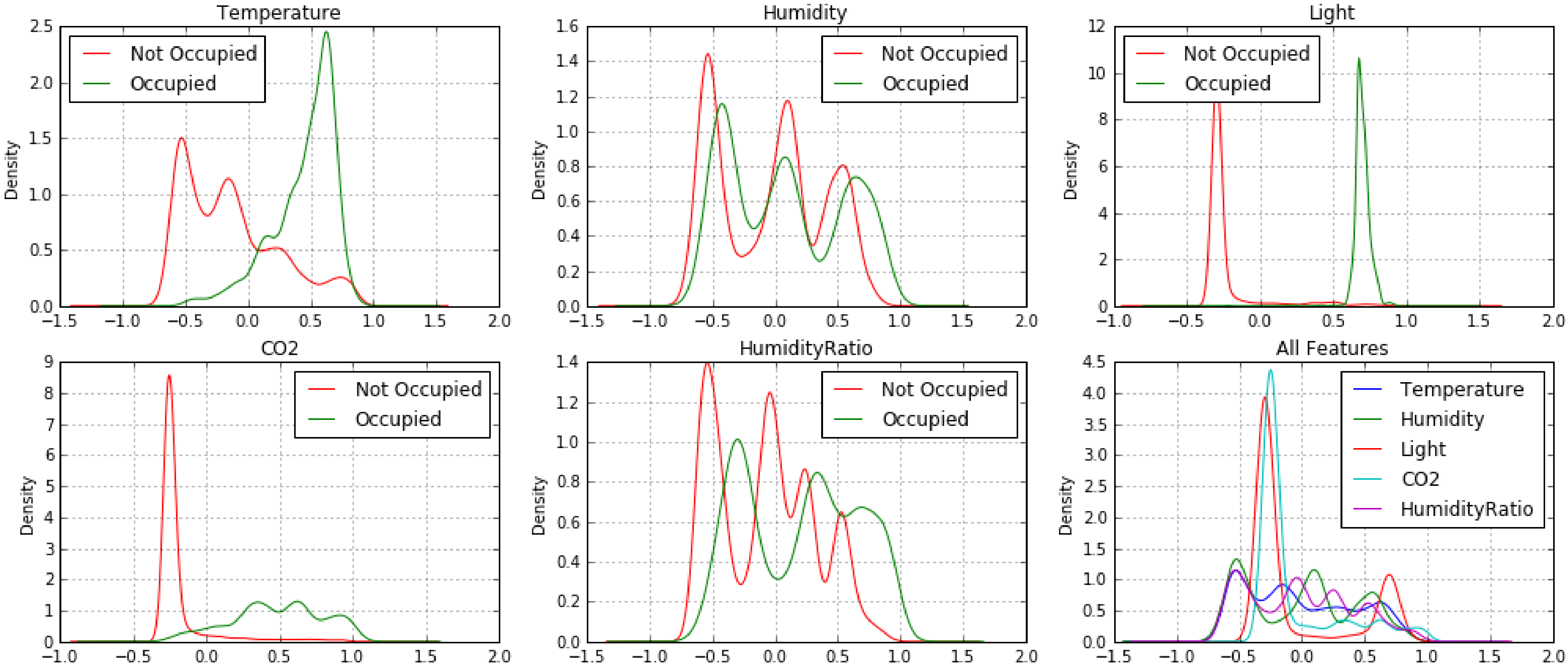
- The estimate of the underlying probability density function is constructed from observed data.

Dataset: Occupancy Detection

Features used:

- HumidityRatio
- CO2
- Light
- Humidity
- Temperature
- Occupancy

Source: http://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+



Principal Components Analysis (PCA) to Visualize High-Dimensional Data

- Statistical procedure to convert a set of observations of variables into a set of values of linearly uncorrelated variables called principal components.
- The first principal component accounts for as much variability as possible, succeeding components capture the highest possible variance under the constraint that they are orthogonal to preceding components.
- It is **not** a feature selection method.
- Should be used only to improve training time and ease of visualization.
- Quickly determine if the data can be segregated using the selected features.

PCA Example

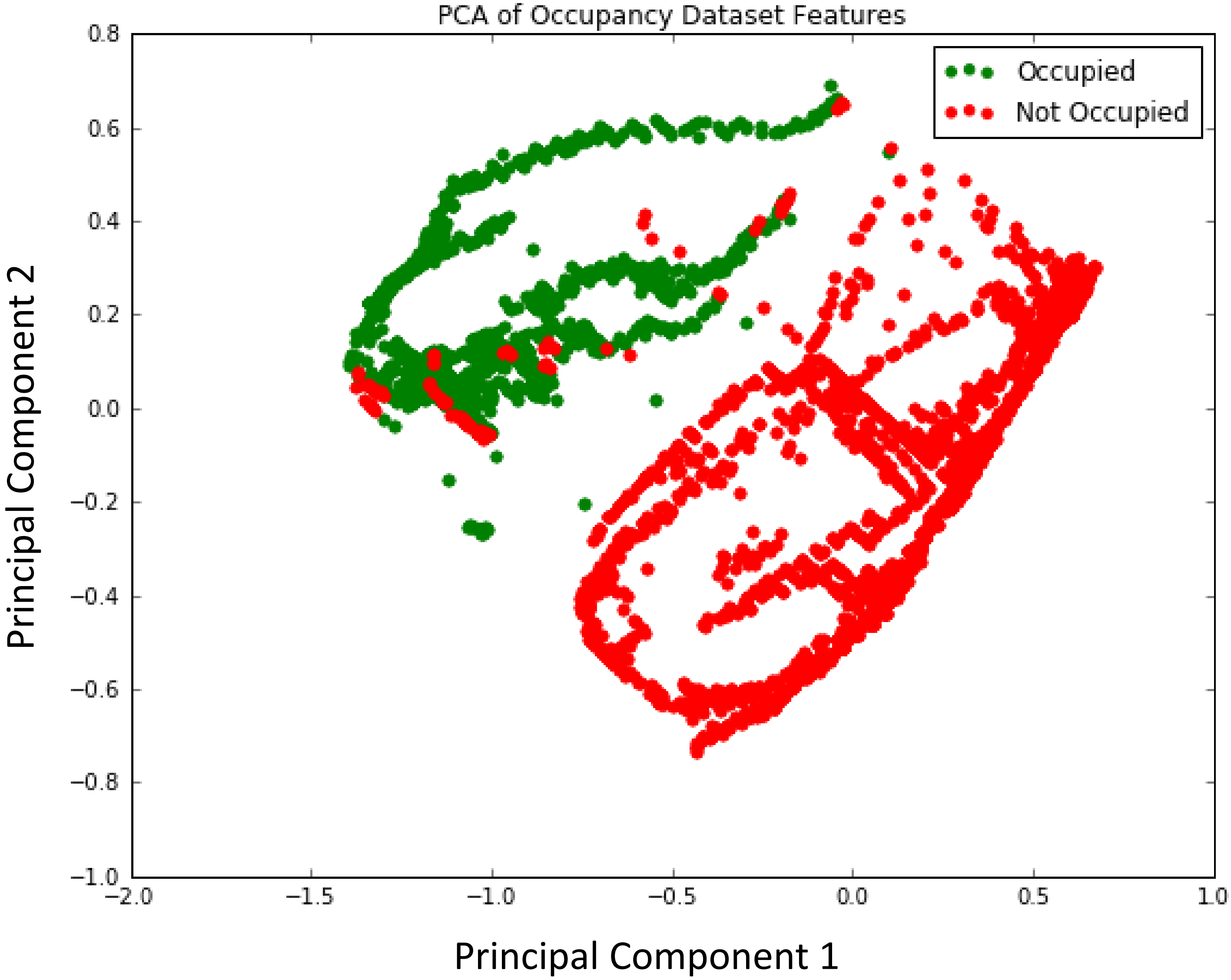
Dataset: Occupancy Detection

Features used:

- Temperature
- Light
- CO2
- Occupancy

Source:

http://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+



Datasets used for the examples in this lecture

- **Occupancy Detection Dataset**

http://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+