DATA ANALYSIS

DIMENSION REDUCTION

DIMENSIONALITY OF DATA

In data analysis, the **dimension** of the data is the number of attributes that are collected in a dataset, represented by the **number of columns**.

We can think of the number of variables used to describe each object (row) as a vector describing that object: the dimension is simply the **size** of that vector.

(**Note:** "dimension" is used differently in business intelligence contexts)

HIGH DIMENSIONALITY AND BIG DATA

Datasets can be “big” in a variety of ways:

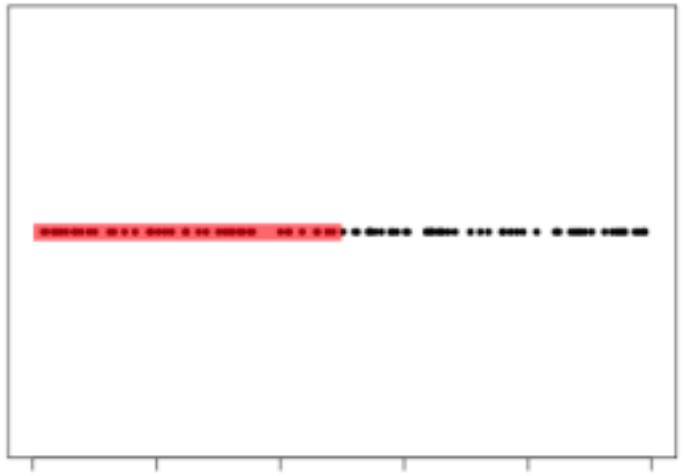
- too large for the **hardware** to handle (cannot be stored, accessed, manipulated properly due to # of observations, # of features, the overall size)
- dimensions can go against **modeling assumptions** (# of features ≫ # observations)

Examples:

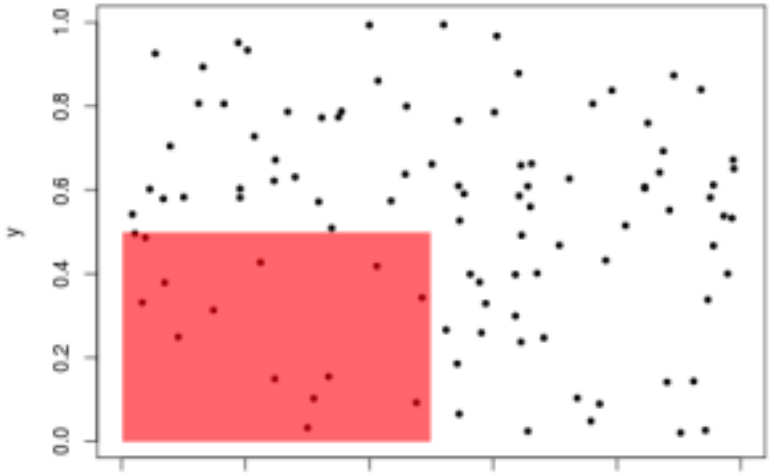
- multiple sensors recording 100+ observations per second in a large geographical area over a long time period = **very big dataset**
- in a corpus’ *Term Document Matrix* (cols = terms, rows = documents), the number of terms is usually substantially higher than the number of documents, leading to **sparse data**

CURSE OF DIMENSIONALITY

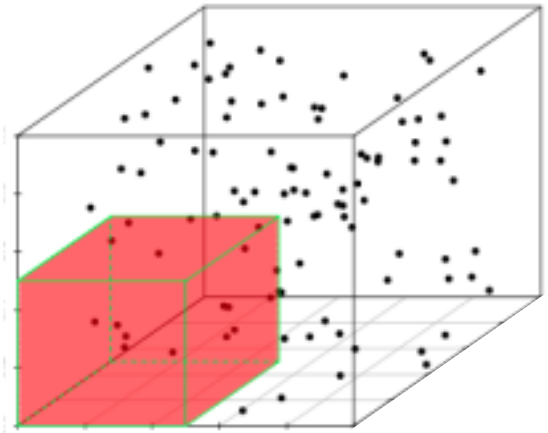
42% of data is captured



14% of data is captured



7% of data is captured



$N = 100$ observations, uniformly distributed on $[0,1]^d, d = 1,2,3.$
% of observations captured by $[0,1/2]^d, d = 1,2,3.$

SAMPLING OBSERVATIONS

Question: does every row of the dataset need to be used?

If rows are selected randomly (with or without replacement), the resulting sample might be **representative** of the entire dataset.

Drawbacks:

- if the signal of interest is rare, sampling might drown it altogether
- if aggregation is happening down the road, sampling will necessarily affect the numbers (passengers vs. flights)
- even simple operations on a large file (finding the # of lines, say) can be taxing on the memory – **prior information on the dataset structure can help**

FEATURE SELECTION

Removing **irrelevant/redundant** variables is a common data processing task.

Motivations:

- modeling tools do not handle these well (variance inflation due to multicolinearity, etc.)
- dimension reduction (# variables ≫ # observations)

Approaches:

- filter vs. wrapper
- unsupervised vs. supervised

DATA ANALYSIS

DIMENSION REDUCTION