BUS41204 Review Session 4

HW1, Classification and ROC

Jingyu He
jingyuhe@chicagobooth.edu

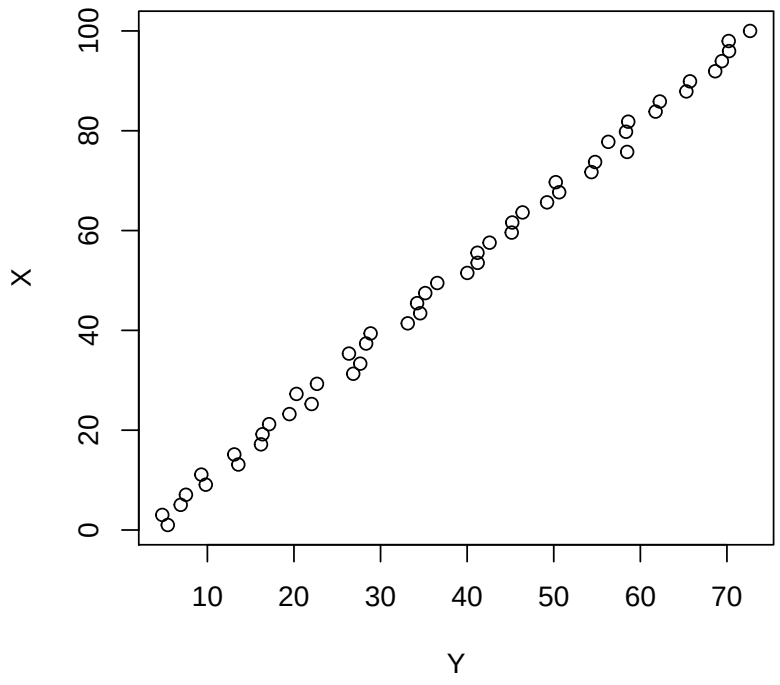
01/28/2017

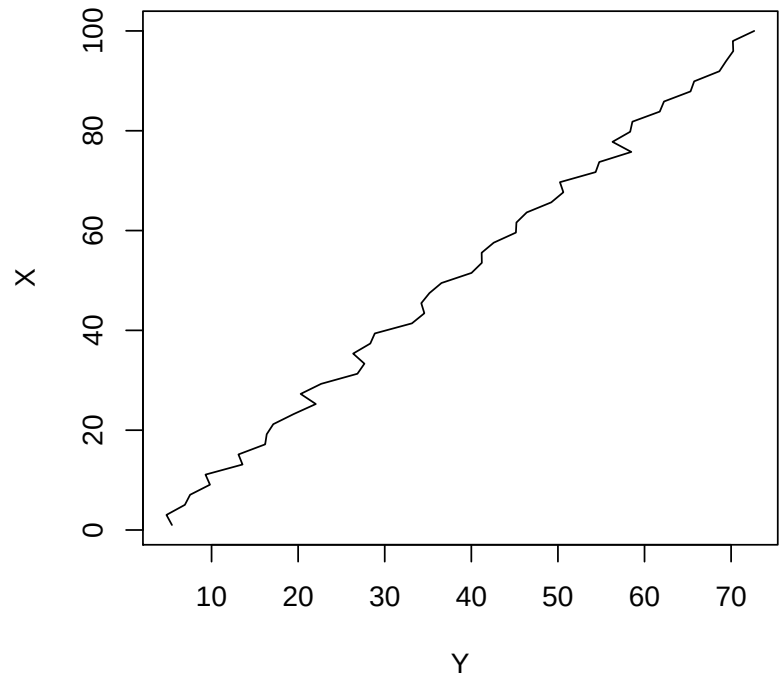
Plan

1. Show how to draw “blue dashed line”.
2. Go over homework 1
3. Show demos of classification by logistic regression, random forest and boosting, how to draw ROC curves to select models.

Scatter Plot and Line

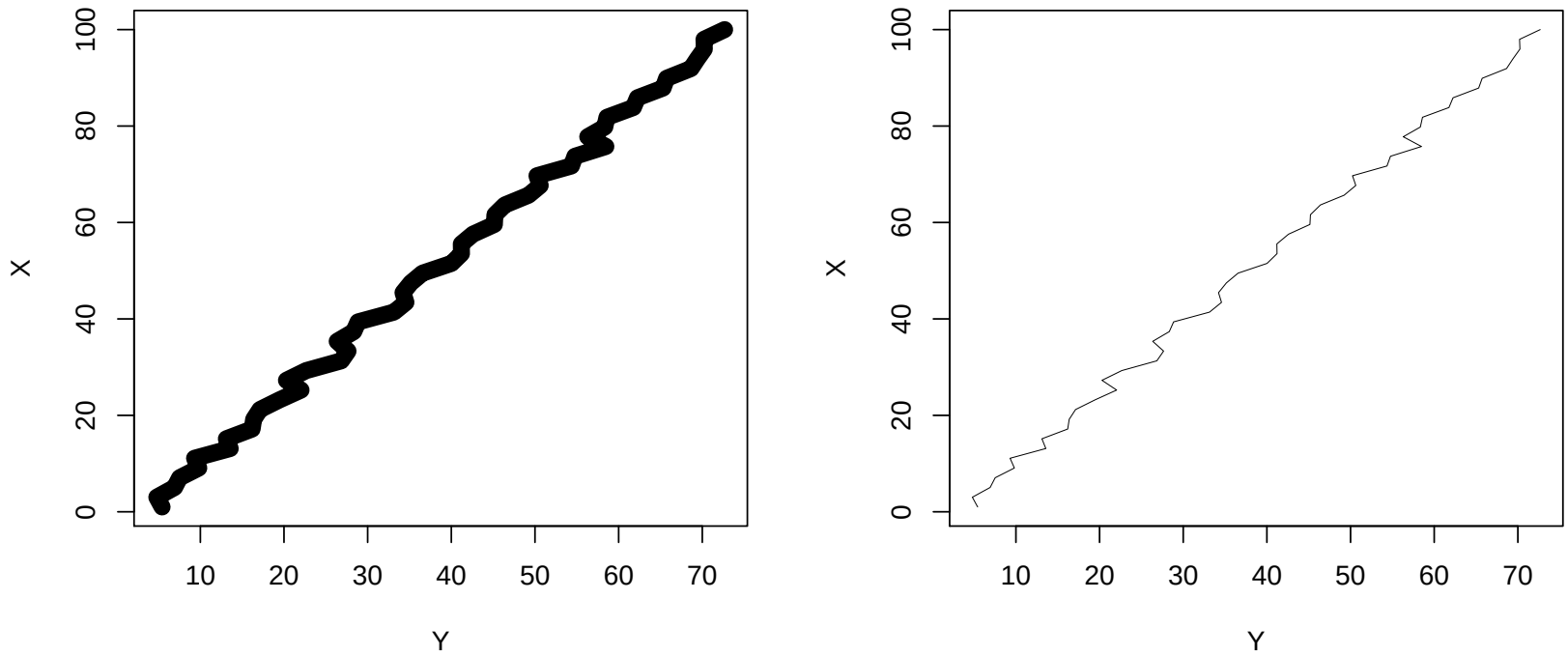
```
par(mfrow = c(1,2))
X = seq(1, 100, length.out =  50);
Y = 0.7 * X + 3 + rnorm(length(X))
plot(Y, X)
plot(Y, X, type = "l") # "l" means lines
```





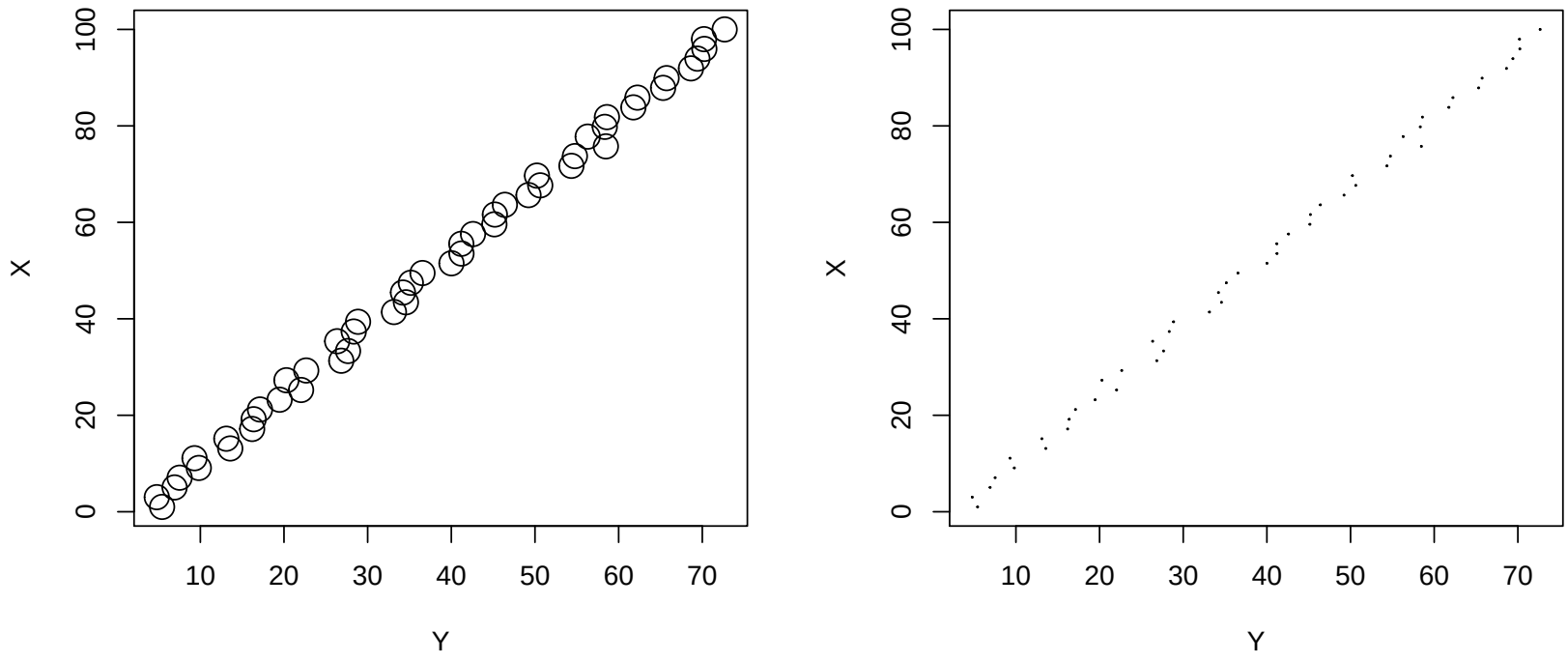
Change Line Width

```
par(mfrow = c(1,2))
# lwd controls width of line
plot(Y, X, type = "l", lwd = 10) # bold line
plot(Y, X, type = "l", lwd = 0.4)
```



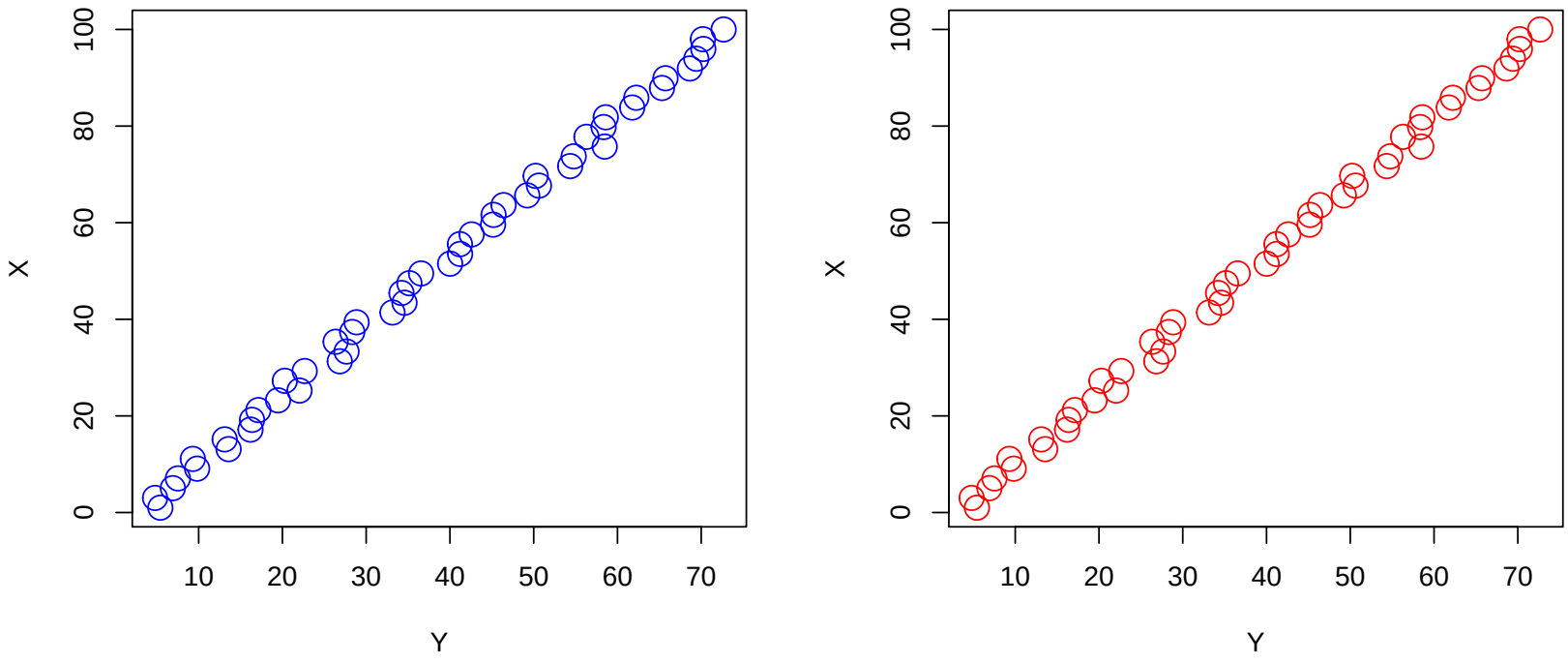
Change Point Size

```
par(mfrow = c(1,2))
# cex controls size of points
plot(Y, X, cex = 2)
plot(Y, X, cex = 0.1)
```



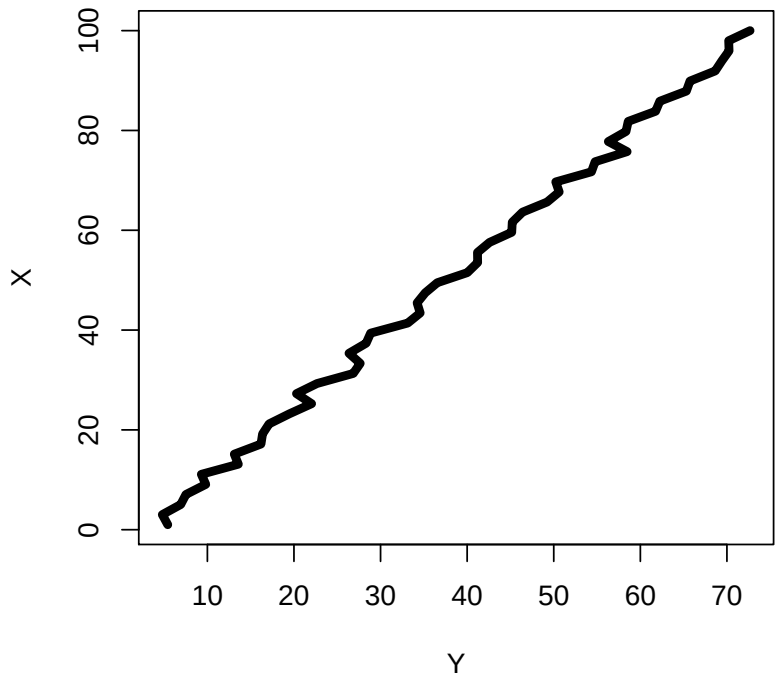
Change Color

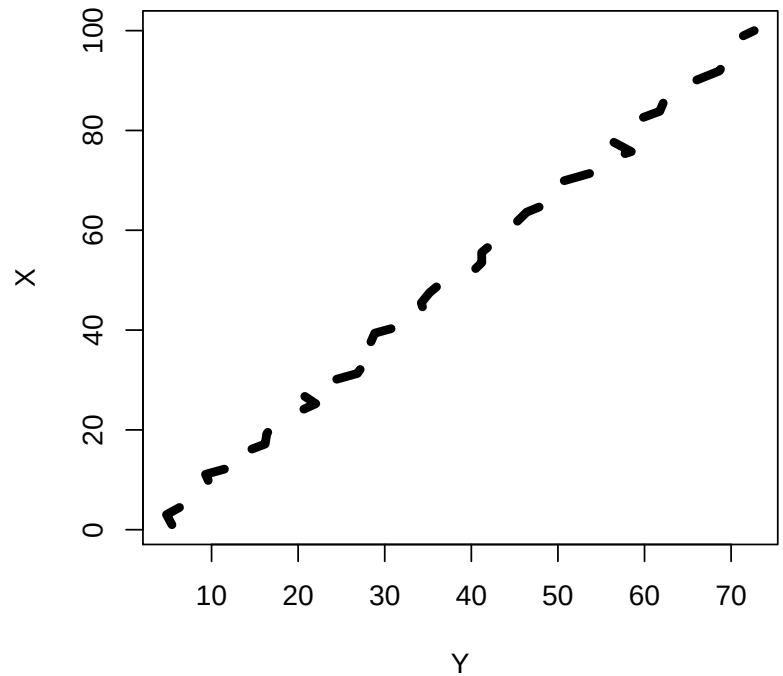
```
par(mfrow = c(1,2))
# col controls color
plot(Y, X, cex = 2, col = "blue")
plot(Y, X, cex = 2, col = "red")
```



Change Line Types

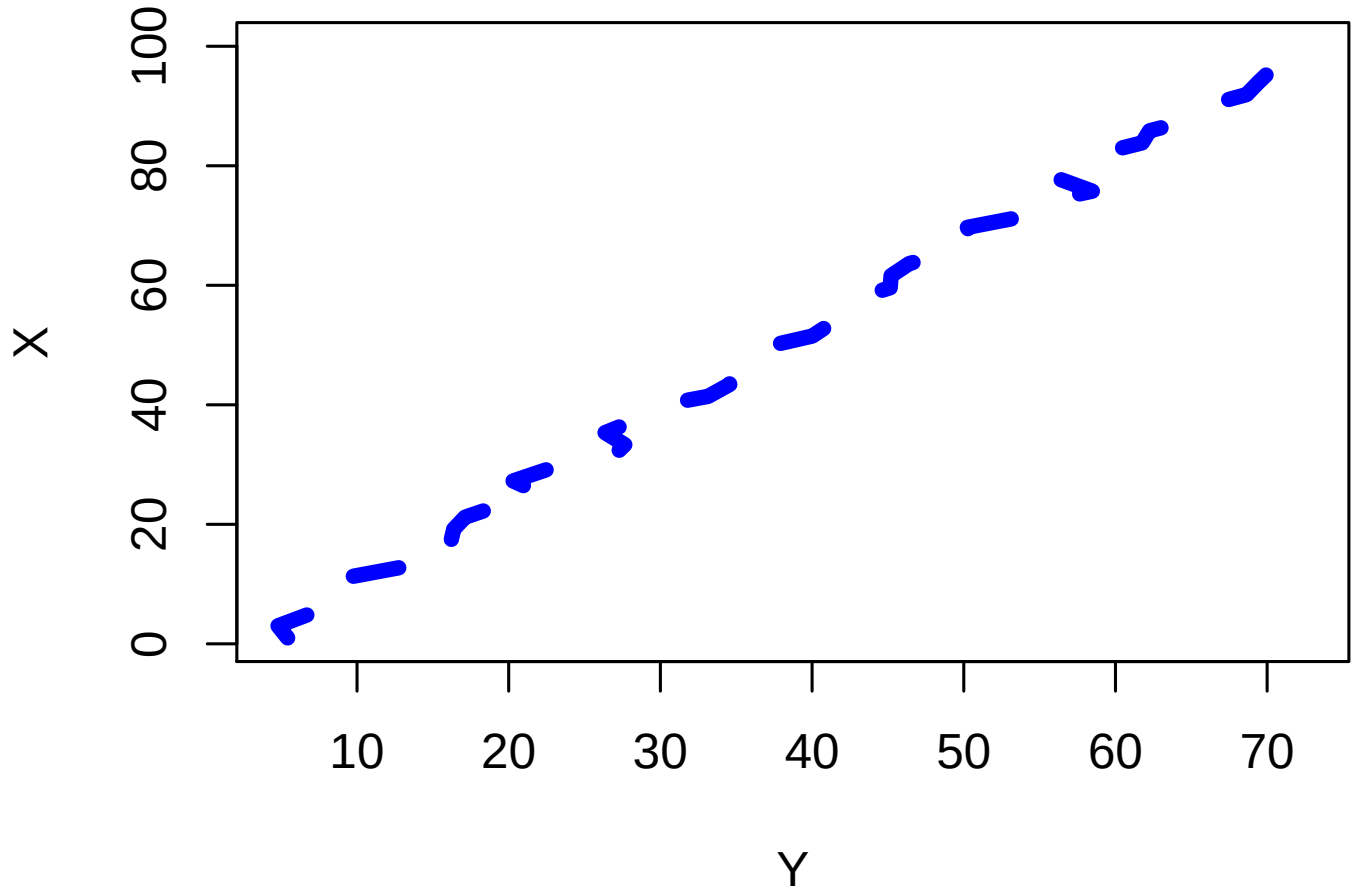
```
par(mfrow = c(1,2))
# lty = 1, solid line, lty = 2, dashed line
plot(Y, X, type = "l", lty = 1, lwd = 5)
plot(Y, X, type = "l", lty = 2, lwd = 5)
```





Blue Dashed Line !

```
plot(Y, X, type = "l", col = "blue", lty = 2, lwd = 5)
```



How to choose K

Sometimes it takes a long time to loop over all possible K (like Q2 in HW1). Here is the strategy:

1. Run the code using a K-vector with large range and big skips, like c(200, 400, 600, 800, 1000). Let's say you find K = 800 has smallest RMSE.
2. Then use finer grid for K around 800, such as c(700, 750, 800, 850, 900). So you can get optimal K with better precision.
3. However, because randomness, it's not necessary to try c(799, 800, 801, 802).

Curse of Dimensionality

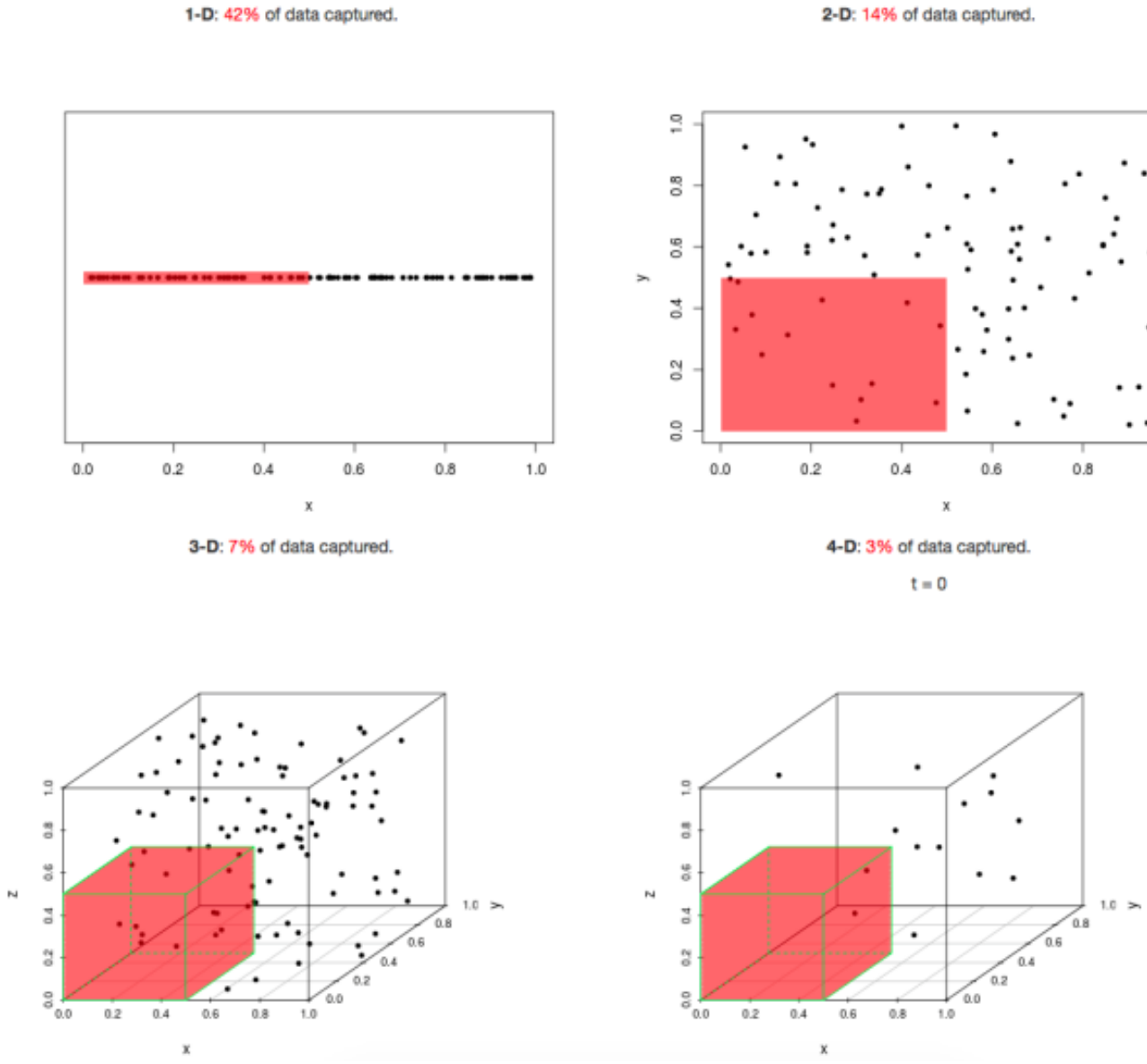


Figure 1:

Curse of Dimentionality

The volume ratio is

$$\left(\frac{0.5^d}{1^d}\right) \to 0 \text{ as } d \to \infty.$$

Data is sparse in high dimension space, aka, data points are far away (large distance) from each other.

Distance is a key concept in kNN. With rising dimensions, it's harder to find a "near neighbour".