Classification with nearest neighbors

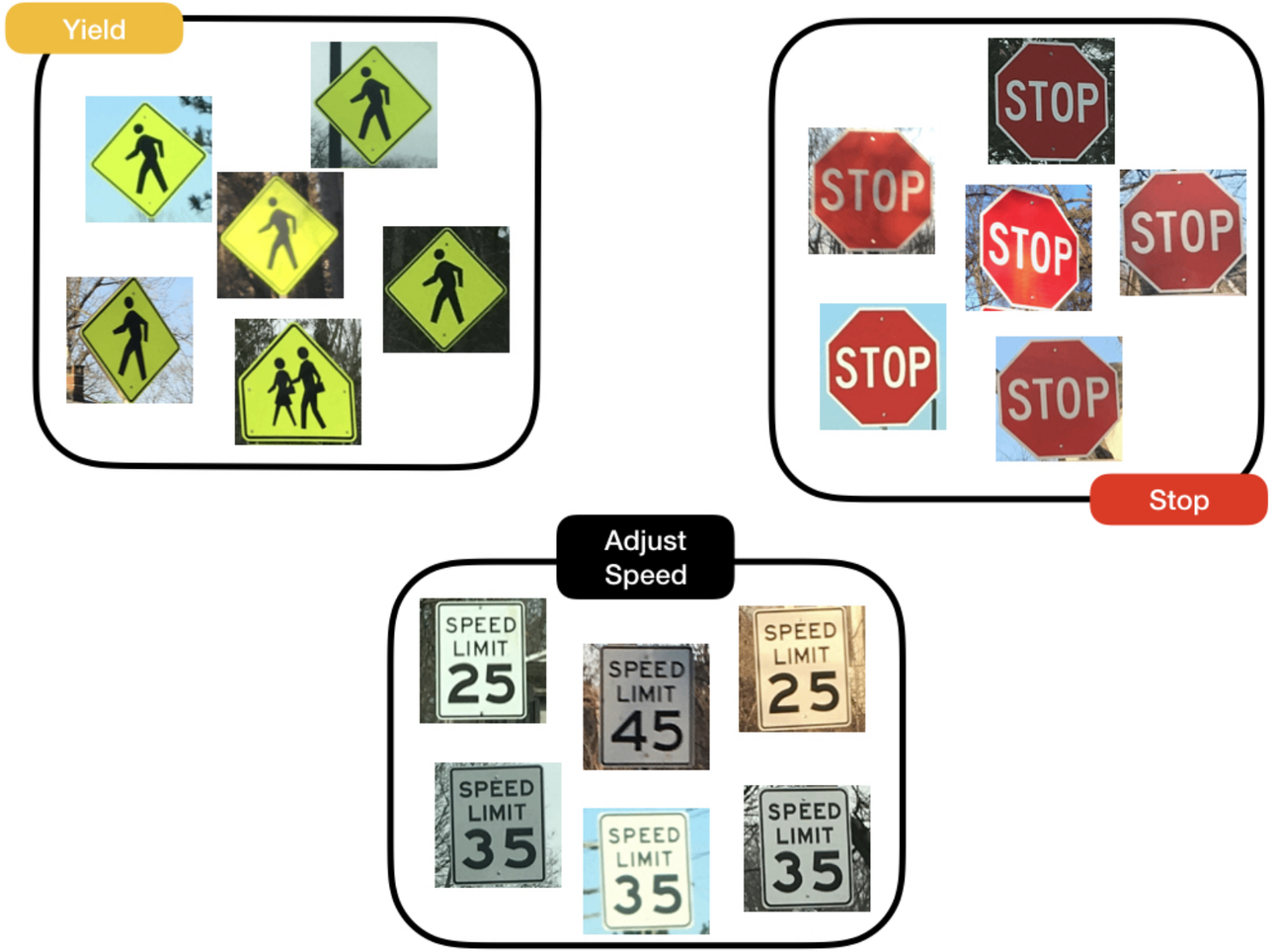
SUPERVISED LEARNING IN R: CLASSIFICATION

Brett Lantz
Instructor

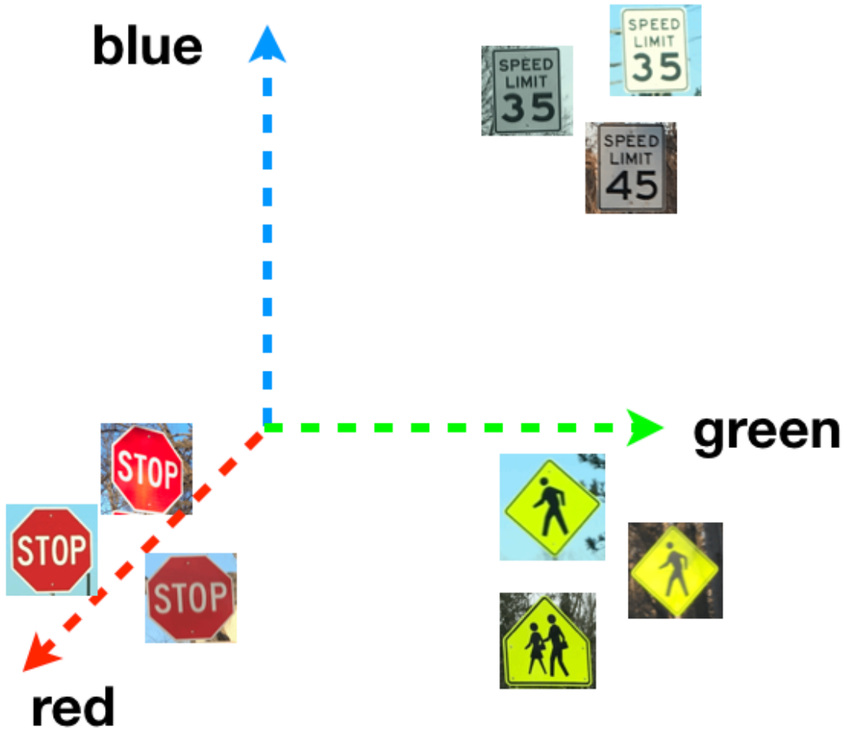
datacamp

Classification tasks for driverless cars

Understanding Nearest Neighbors



Measuring similarity with distance



$$\text{dist}(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2}$$

Applying nearest neighbors in R

```
library(class)
pred <- knn(training_data, testing_data, training_labels)
```

Let's practice!

SUPERVISED LEARNING IN R: CLASSIFICATION

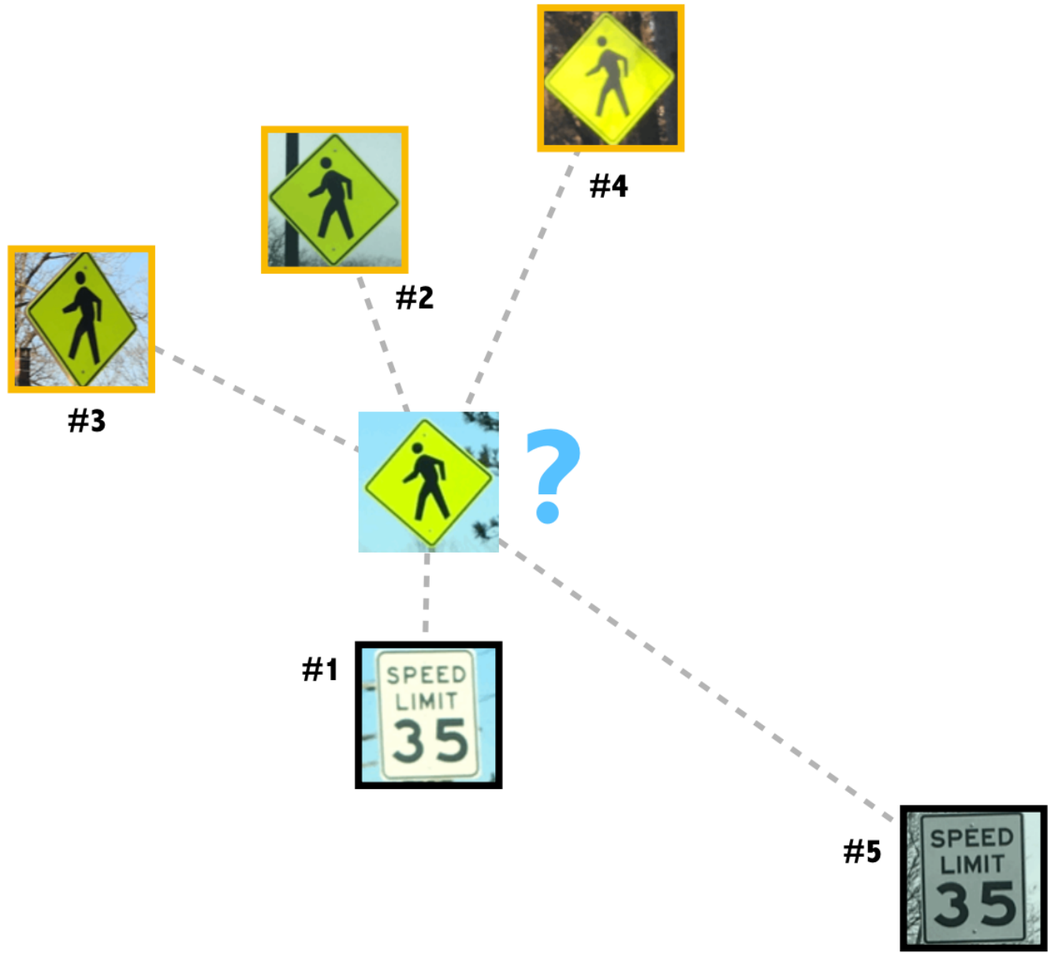
What about the 'k' in kNN?

SUPERVISED LEARNING IN R: CLASSIFICATION

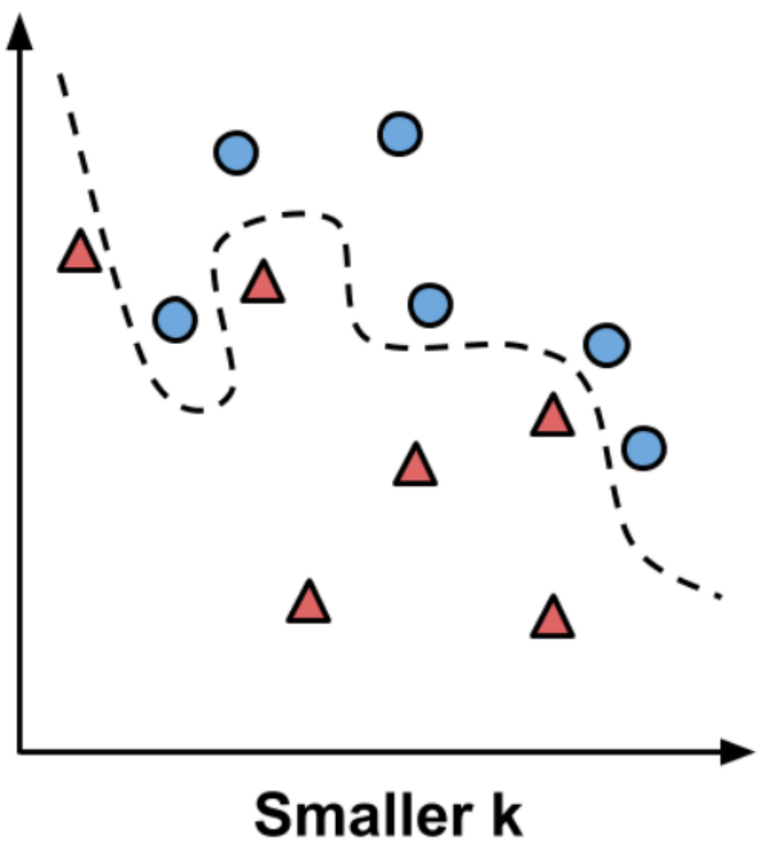
Brett Lantz
Instructor

datacamp

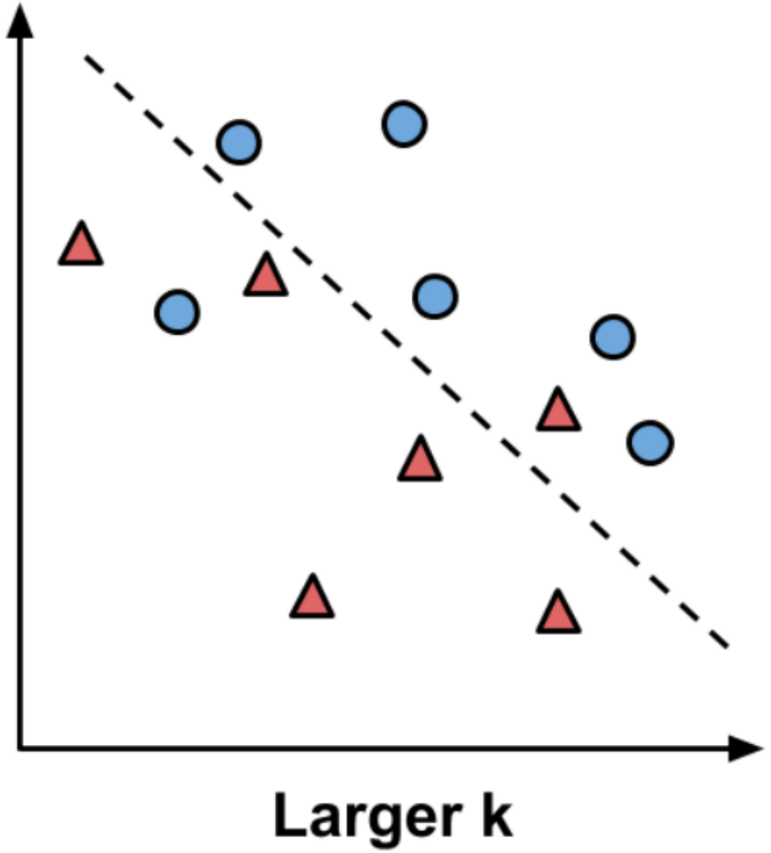
Choosing 'k' neighbors



Bigger 'k' is not always better



Smaller k



Larger k

Choosing 'k'

Let's practice!

SUPERVISED LEARNING IN R: CLASSIFICATION

Data preparation for kNN

SUPERVISED LEARNING IN R: CLASSIFICATION

Brett Lantz
Instructor

kNN assumes numeric data

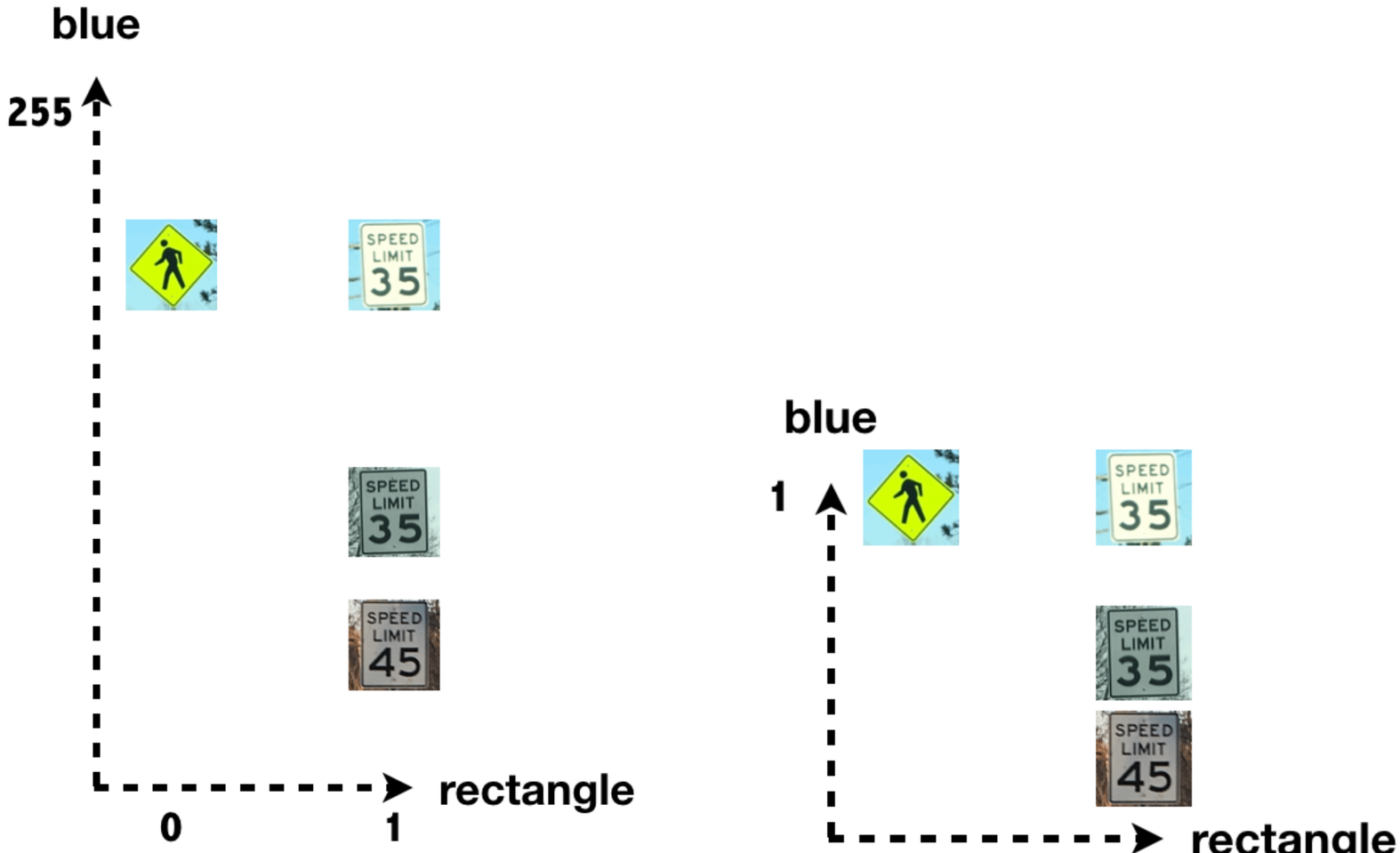
rectangle = 1

diamond = 0

rectangle = 0

diamond = 1

kNN benefits from normalized data



Normalizing data in R

```
# define a min-max normalize() function
normalize <- function(x) {
  return((x - min(x)) / (max(x) - min(x)))
}
```

```
# normalized version of r1
summary(normalize(signs$r1))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.1935  0.3528  0.4046  0.6129  1.0000
```

```
# un-normalized version of r1
summary(signs$r1)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    3.0    51.0    90.5   103.3   155.0   251.0
```

Let's practice!

SUPERVISED LEARNING IN R: CLASSIFICATION