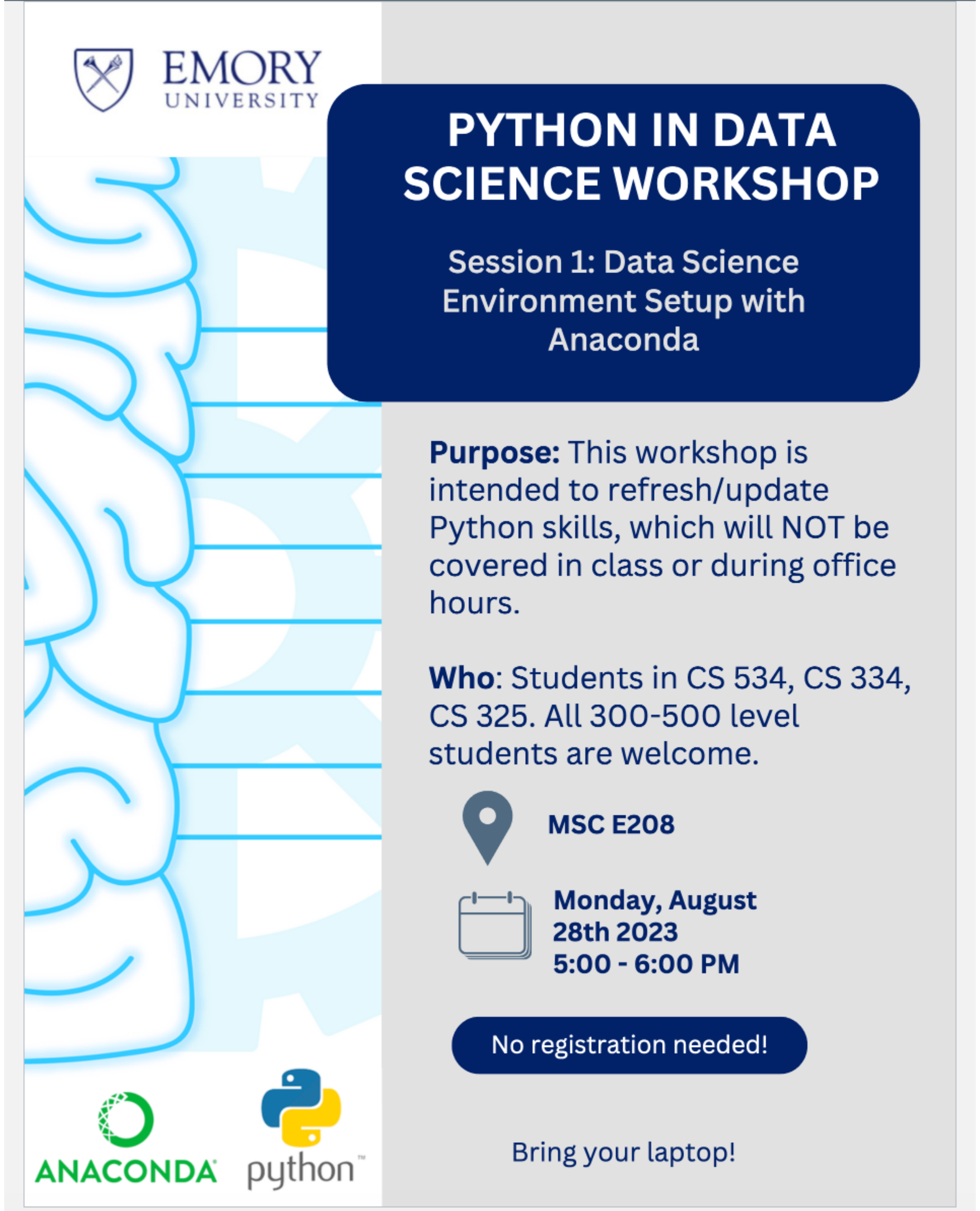
PYTHON WORKSHOP

- Session 1: environment setup, Monday, 8/28 5-6pm, E208
- Session 2: basic Python, Friday 9/1, 5-6:30pm
- (tentative) Session 3: Python ML workflow, Wednesday 9/20, 5:30-6:30pm
- (tentative) Session 4: beyond linear modeling, Monday 9/25, 5-6:30pm



COURSE OUTLINE

- Algorithms for supervised learning: nearest neighbors, decision trees, linear regression, logistic regression, neural networks, naïve bayes, ensembles, boosting, deep learning
- Algorithms for unsupervised learning: principal component analysis
- Model assessment and model selection
- New learning paradigms and emerging topics

K-NEAREST NEIGHBORS

CS 334: Machine Learning

Slides adapted from Intel ML Academy and Stanford CS231n

BREAKOUT ACTIVITY

NETFLIX PRIZE (2006-2009)

$1M prize for 10% improvement

NETFLIX DATASET

	Star Wars I: The Phantom Menace	Star Wars IV: A New Hope	Star Wars VII: The Force Awakens	Raiders of the Lost Arc	Casablanca	Singing in the Rain
Sam	3	4	3	4	1	2
Alice	4	5	5	4	2	1
Bob	1	2	3	2	5	3
Matt	2	3	3	1	4	4
Joyce	5	5	5	**?**	**?**	2

What are Joyce's missing ratings and why?

NETFLIX DATASET

	Star Wars I: The Phantom Menace	Star Wars IV: A New Hope	Star Wars VII: The Force Awakens	Raiders of the Lost Arc	Casablanca	Singing in the Rain
Sam	3	4	3	4	1	2
Alice	4	5	5	4	2	1
Bob	1	2	3	2	5	3
Matt	2	3	3	1	4	4
Joyce	5	5	5	**?**	**?**	2

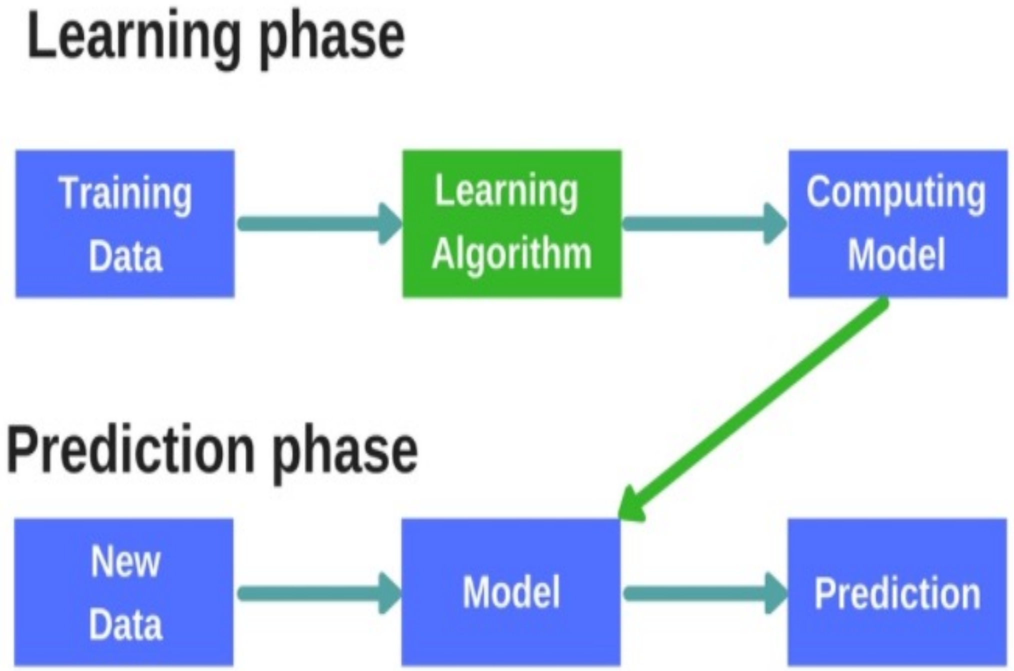
Most similar

What are Joyce's missing ratings and why?

EXAMPLE: IMAGE RECOGNITION

Distance Metric  $\to \mathbb{R}$

Nearest Neighbor (NN) Classifier



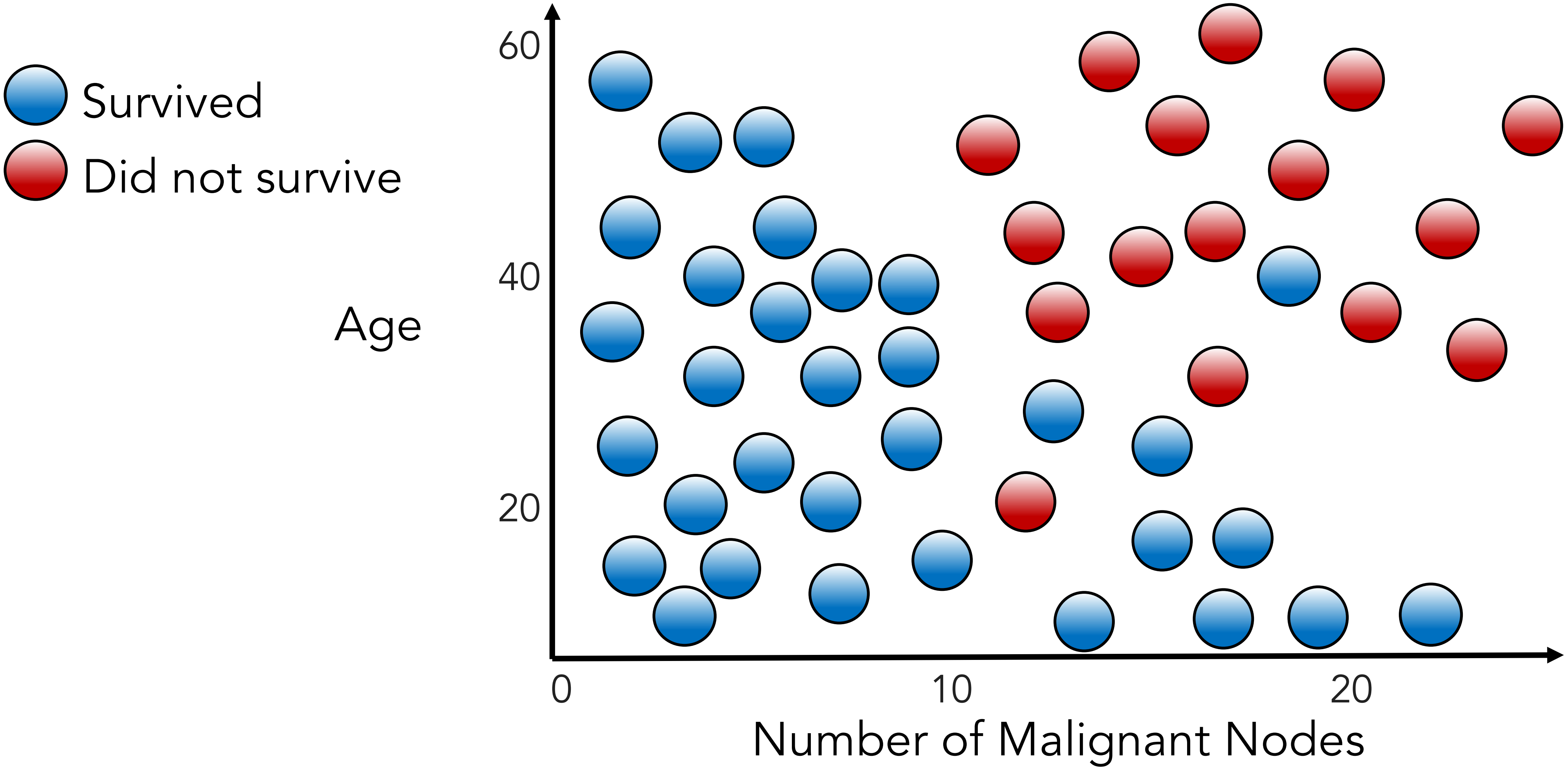

```python
def train(images, labels):
  # Machine learning!
  return model
```

Memorize all data and labels

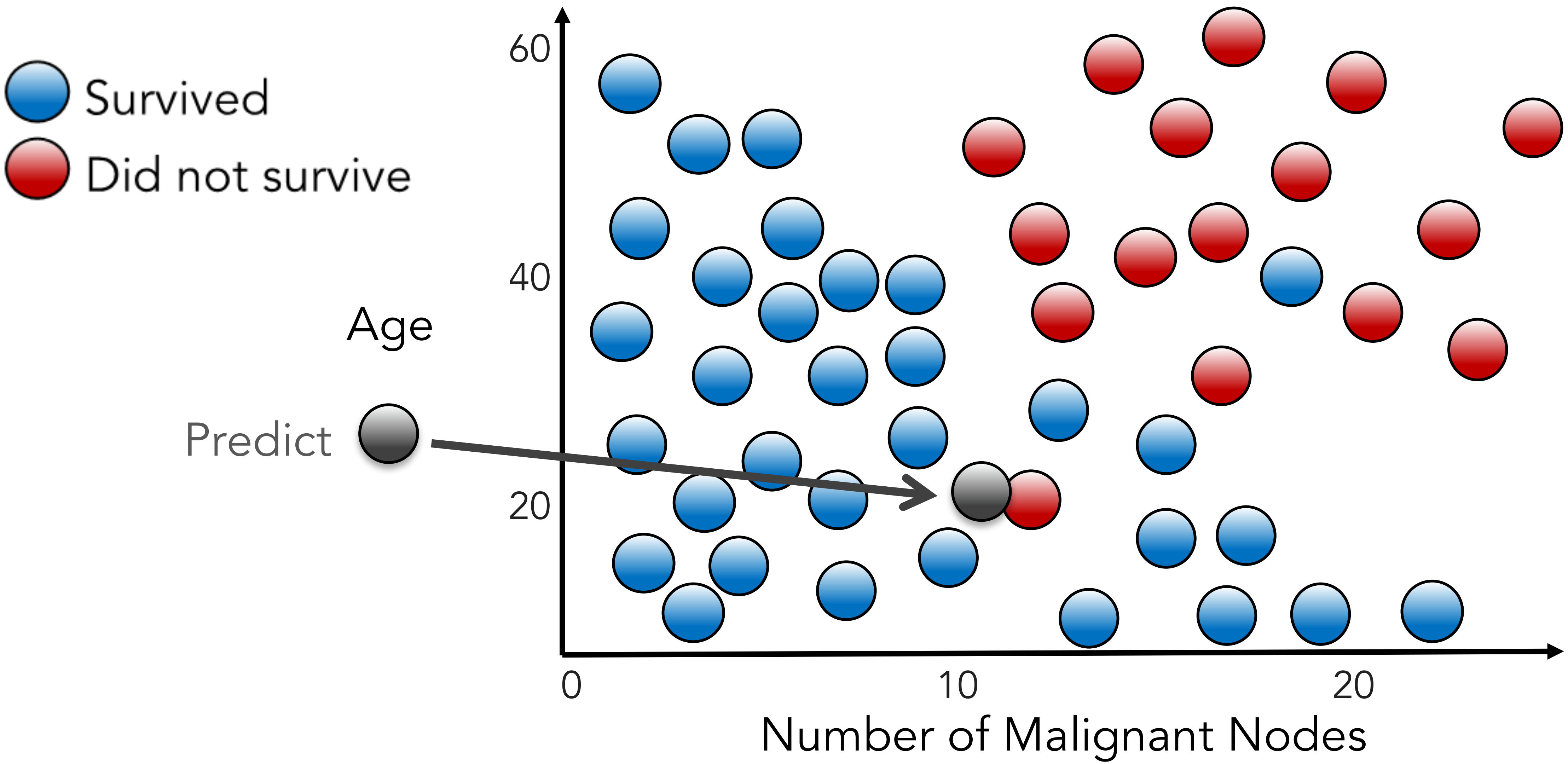
```python
def predict(model, test_images):
  # Use model to predict labels
  return test_labels
```

Predict the label of the most similar training image

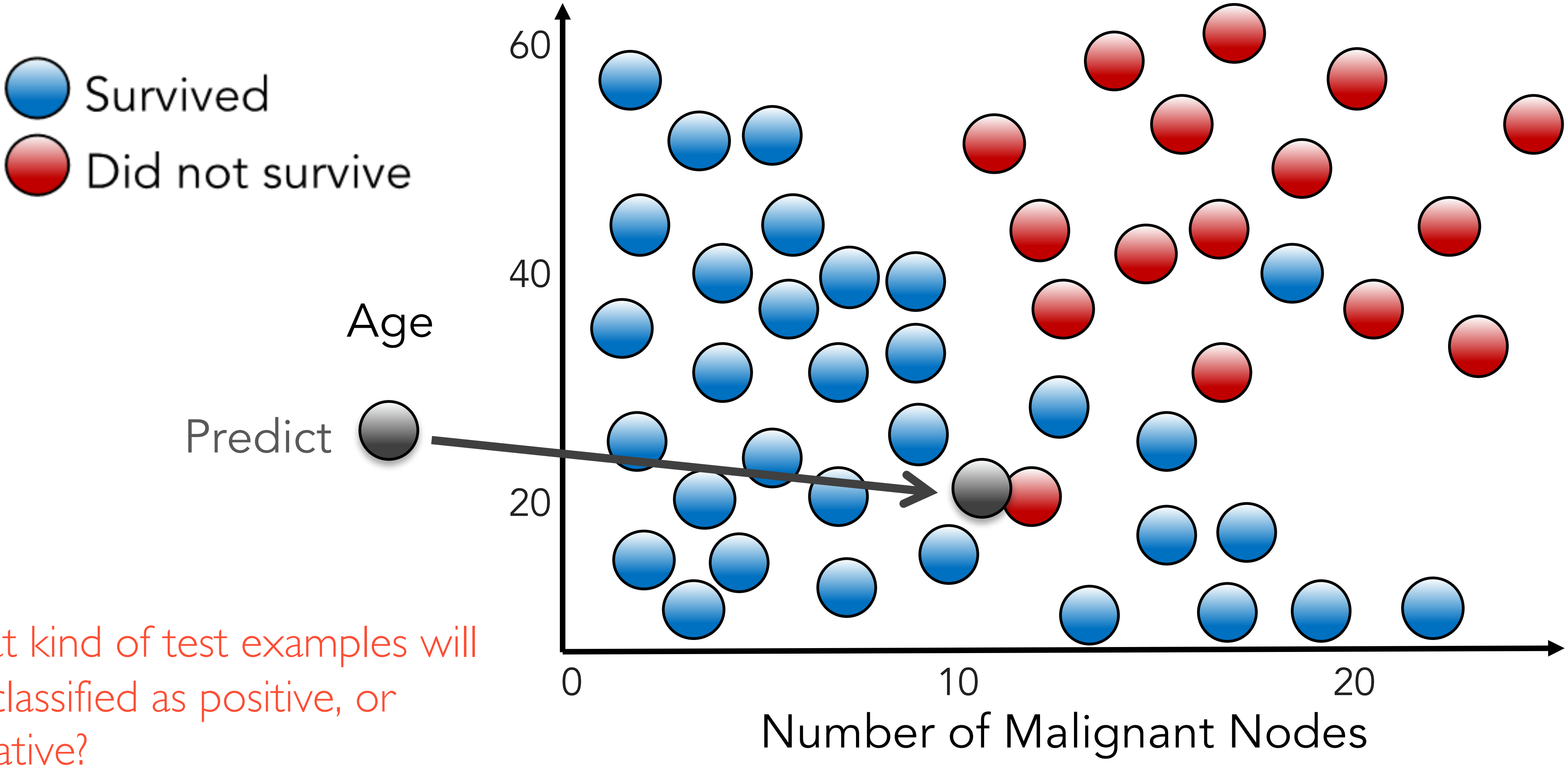
EXAMPLE: CANCER SURVIVAL



EXAMPLE: CANCER SURVIVAL

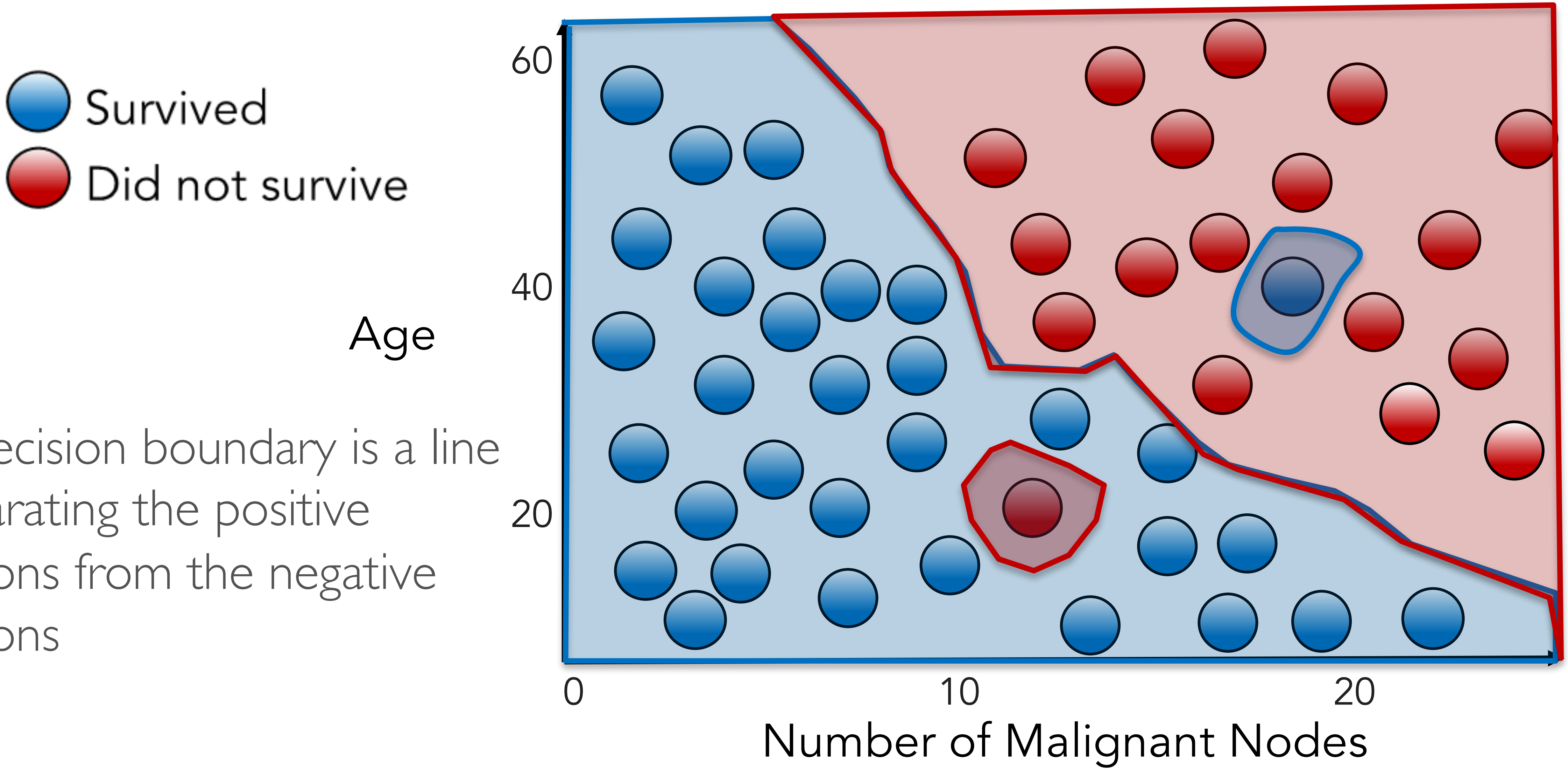


EXAMPLE: CANCER SURVIVAL



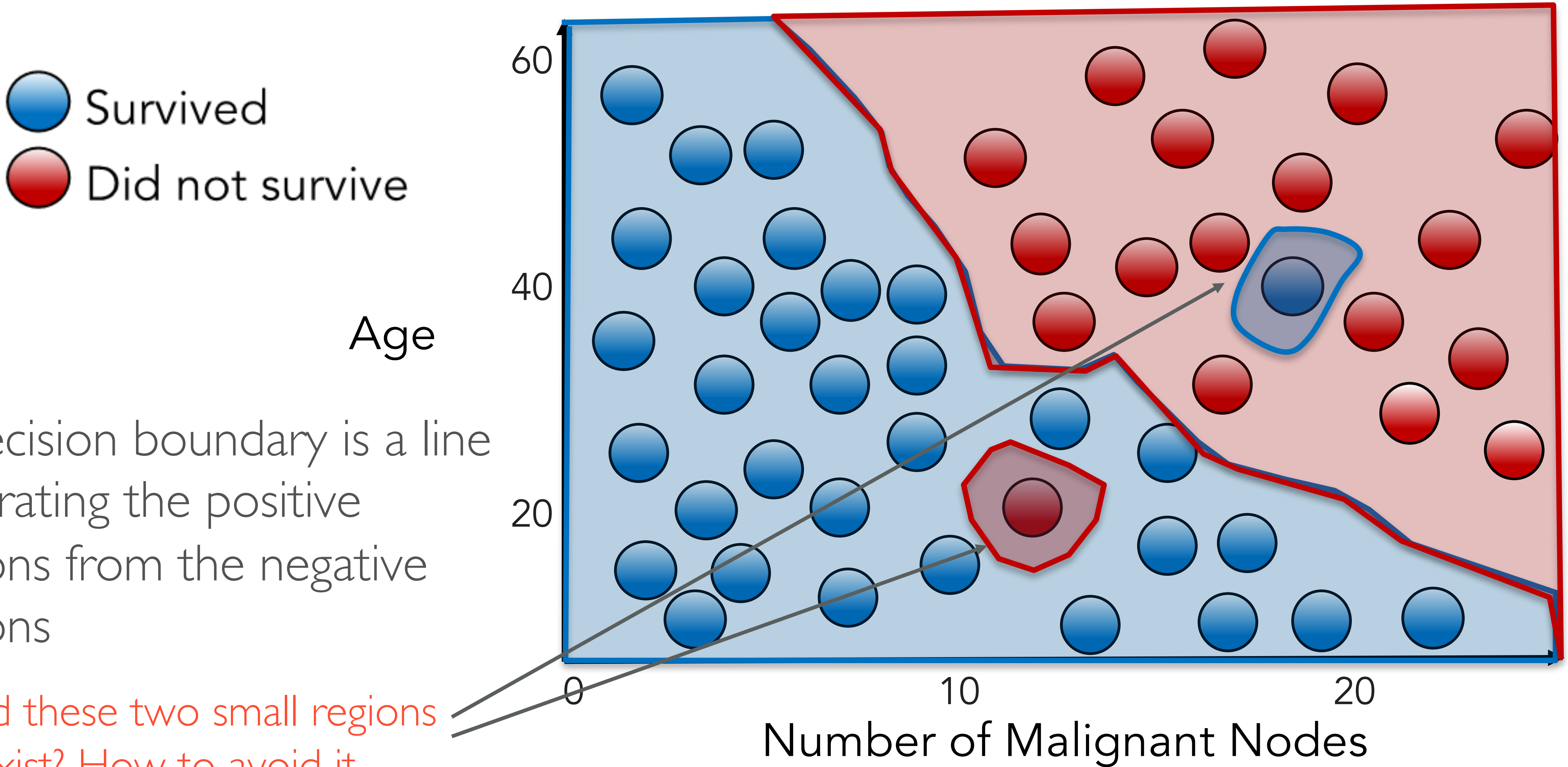
what kind of test examples will be classified as positive, or negative?

NN DECISION BOUNDARIES



A decision boundary is a line separating the positive regions from the negative regions

NN DECISION BOUNDARIES

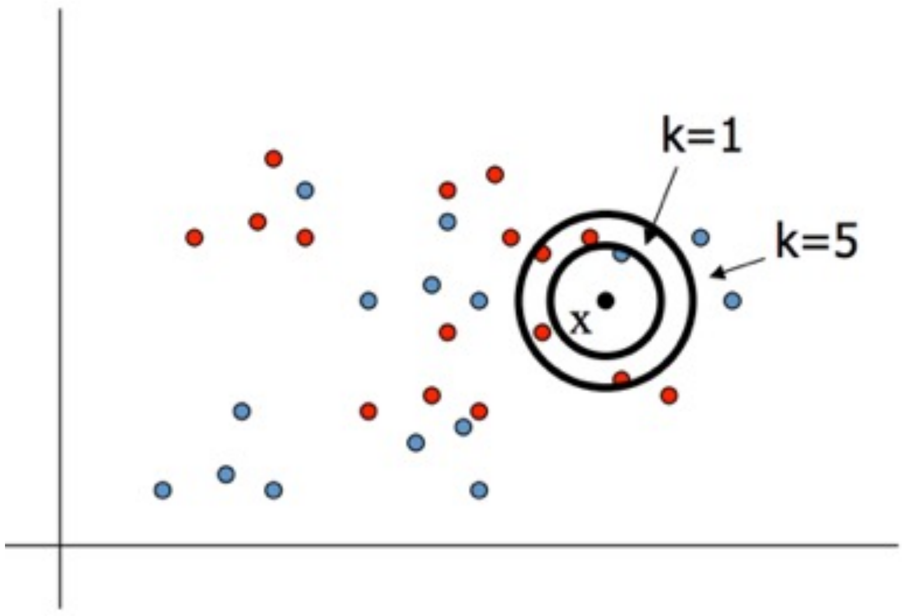


A decision boundary is a line separating the positive regions from the negative regions

Should these two small regions exist? How to avoid it

K-NEAREST NEIGHBOR (K-NN) CLASSIFIER

- Examine the k-“closest” training data points to new point x
- Assign the object the most frequently occurring class (majority vote) or the average value (regression)
- Can have weighted majority or weighted average



http://cs.nyu.edu/~dsontag/courses/ml13/slides/lecture11.pdf

K-Nearest Neighbor (kNN) Classifier

- Training:
  memorize/store the entire training set including features and labels

```python
def train(images, labels):
  # Machine learning!
  return model
```

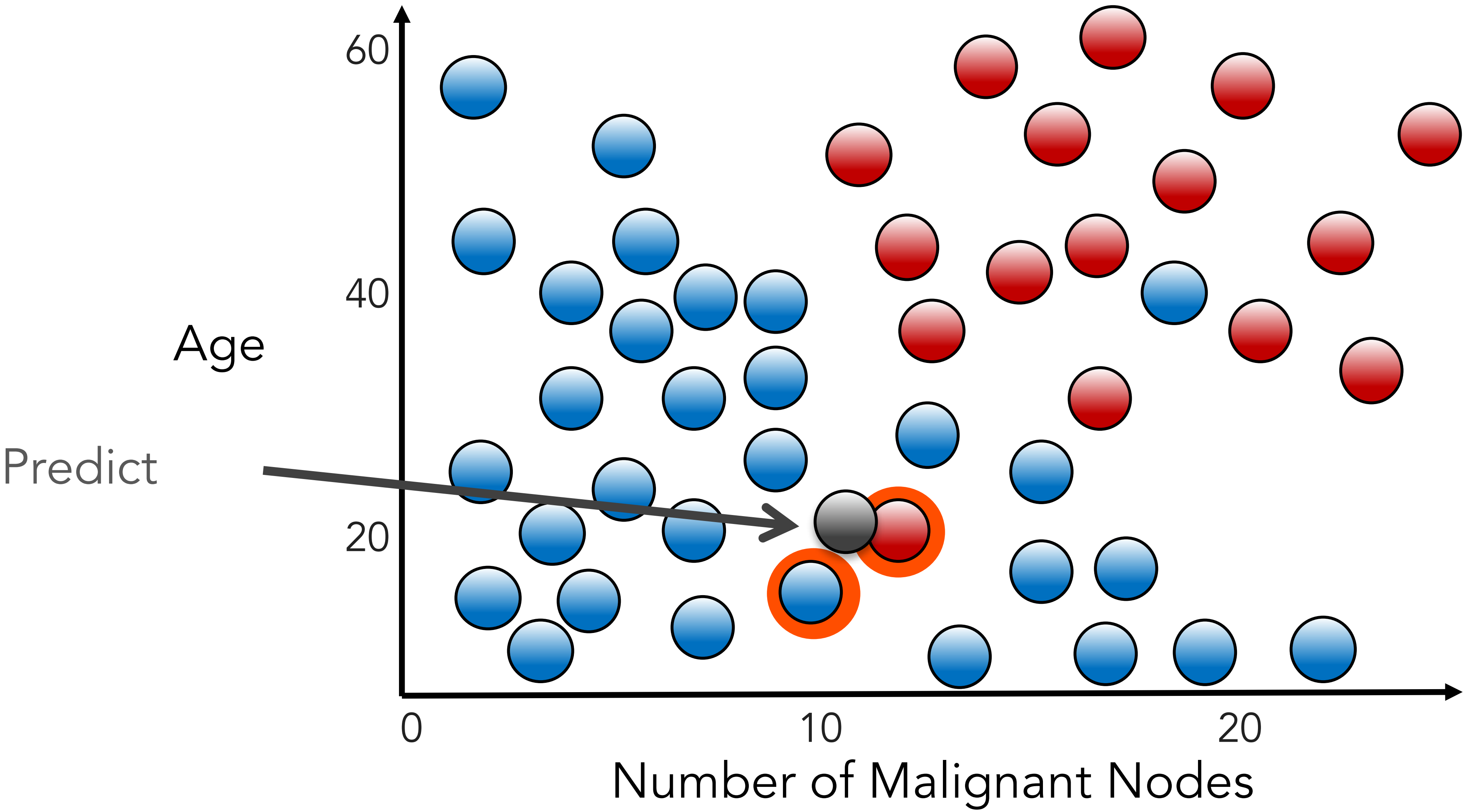
→ Memorize all data and labels

- Prediction:
  find the k-“closest” training data point to the test point
  return the majority class

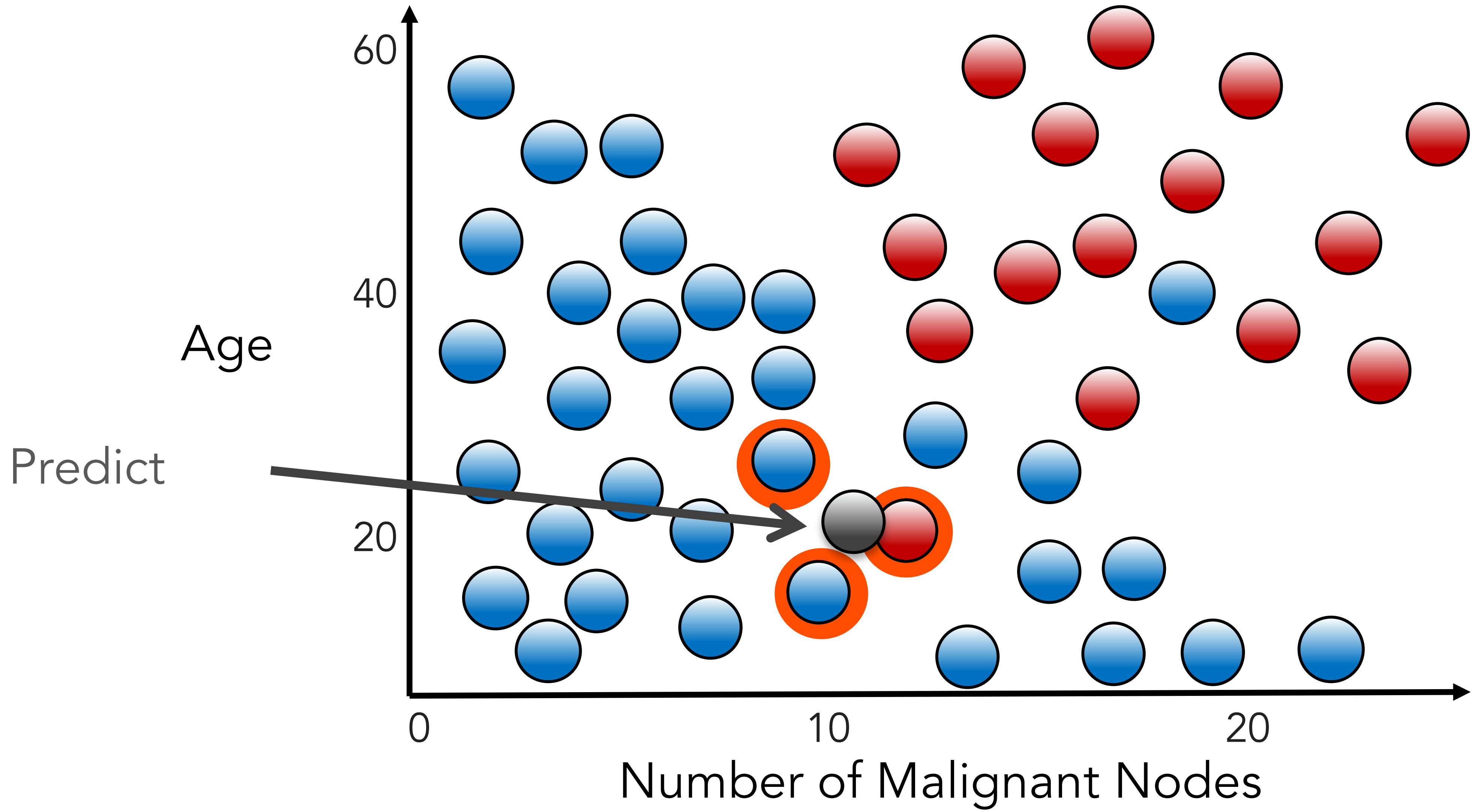
```python
def predict(model, test_images):
  # Use model to predict labels
  return test_labels
```

→ Predict the label of the most similar training image

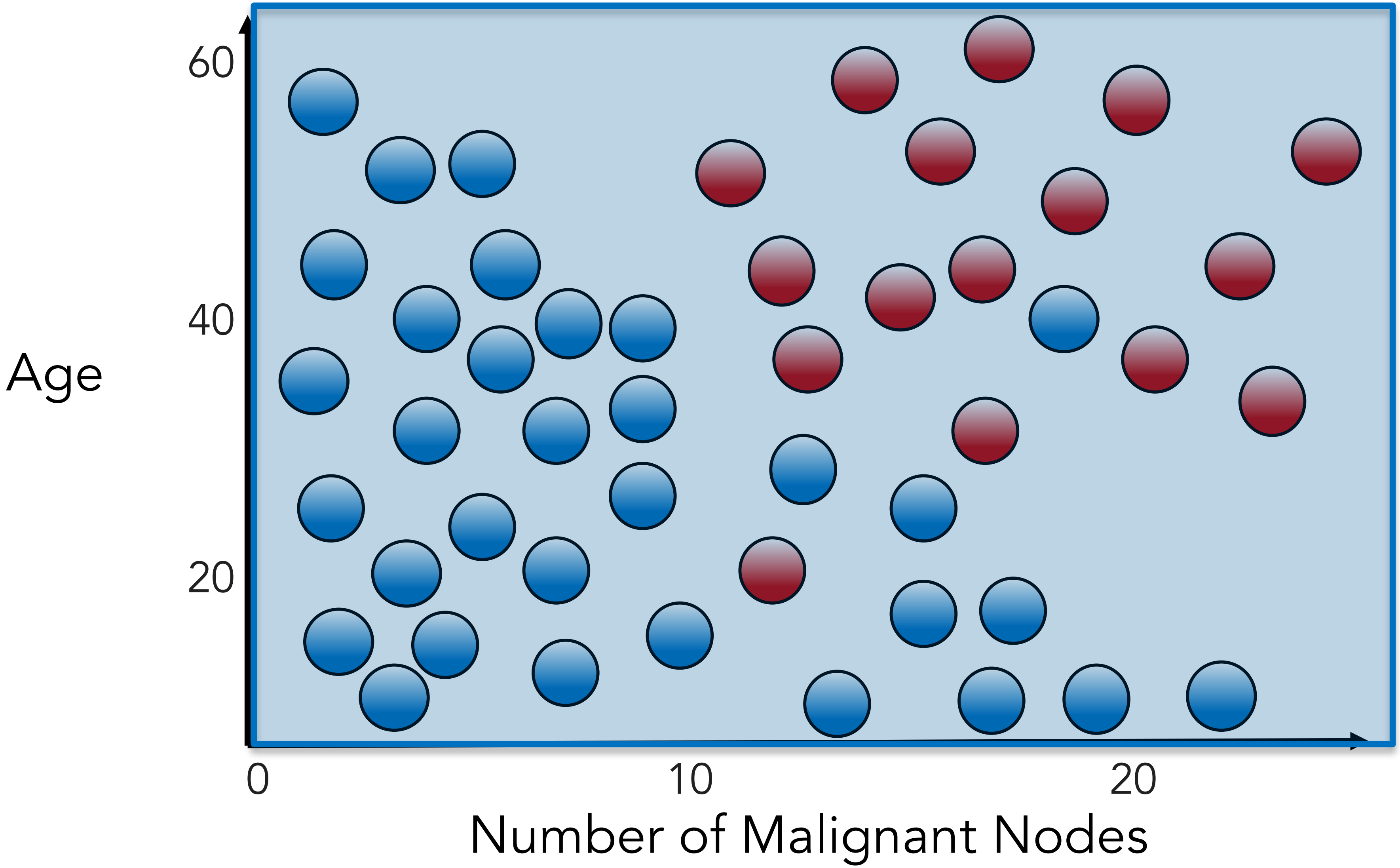
EXAMPLE: K=2



EXAMPLE: K=3

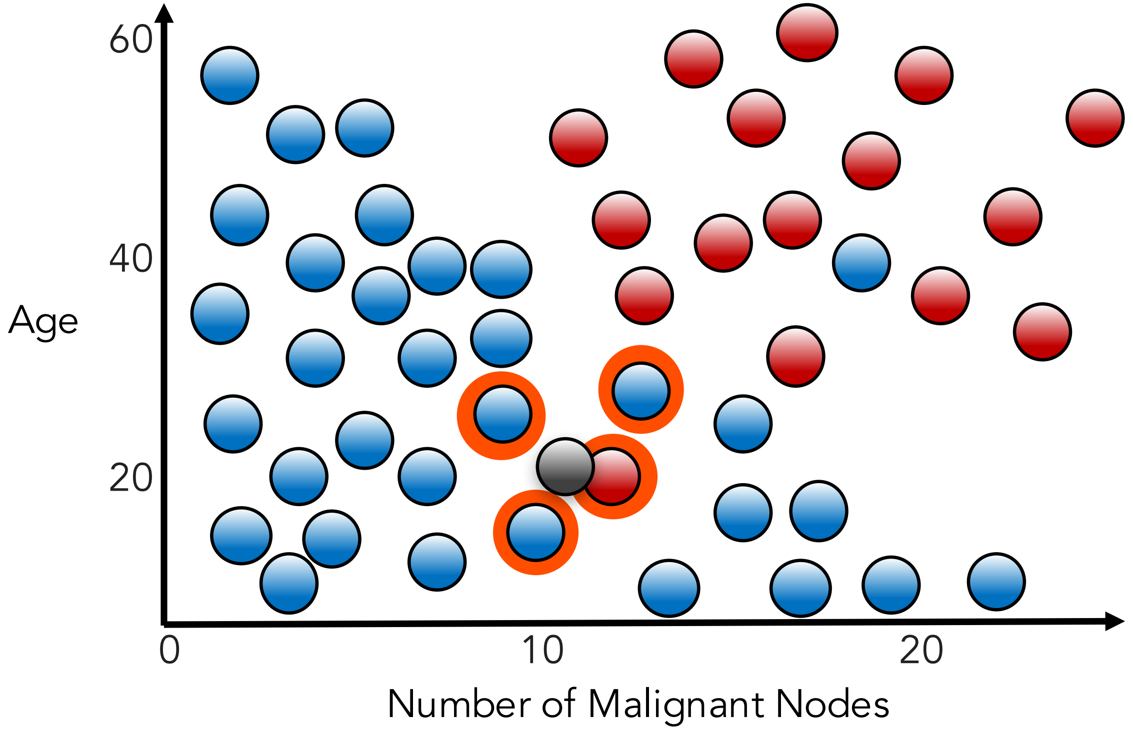


EXAMPLE: K=N

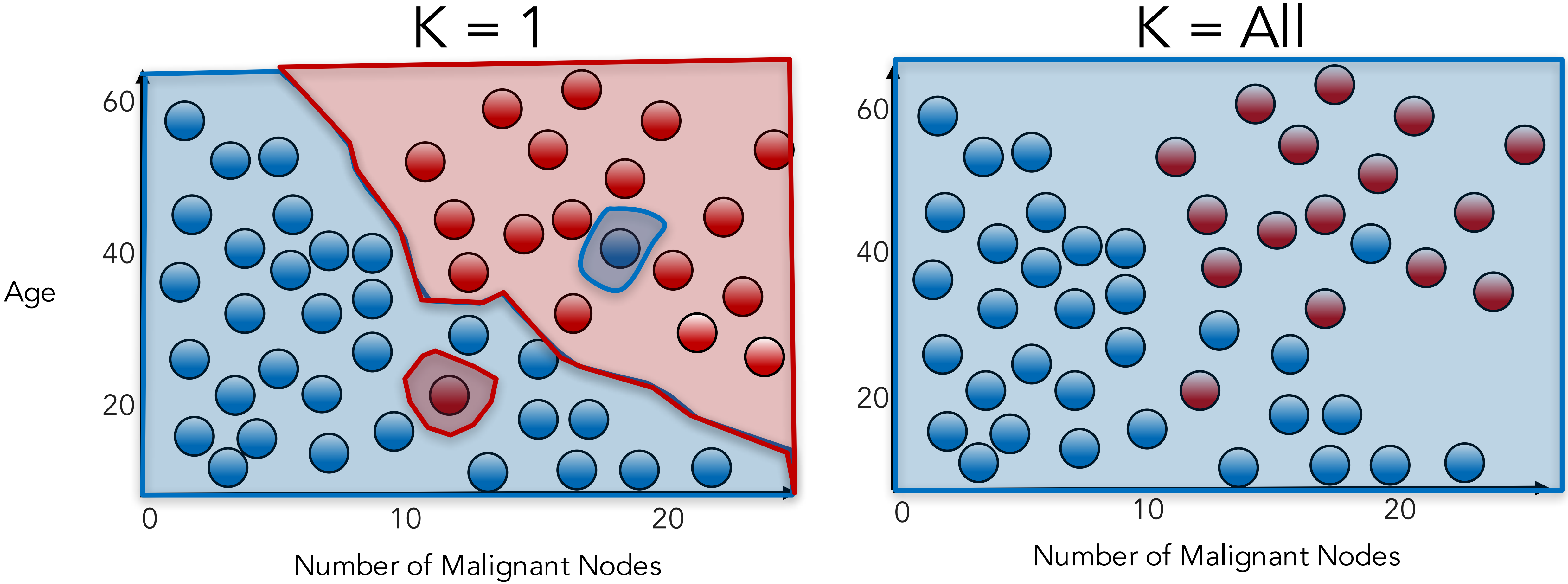


K-NN: PRACTICAL CHALLENGES

- How to pick k?
- What is the right measure of closeness/distance?

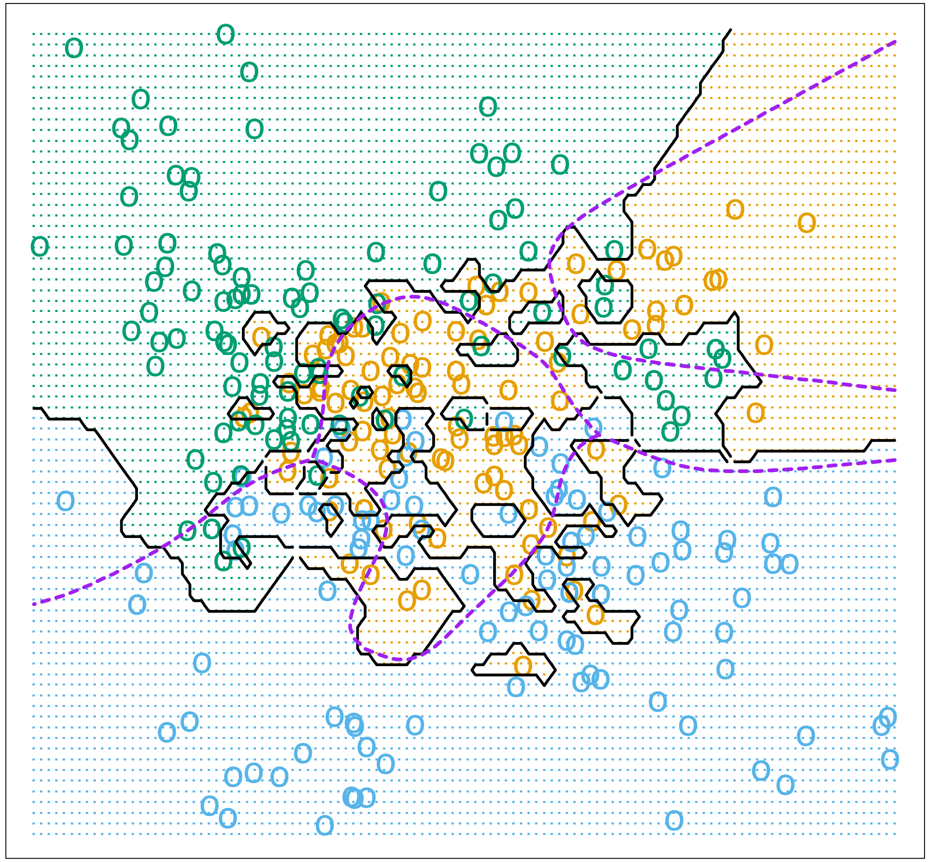


VALUE OF K

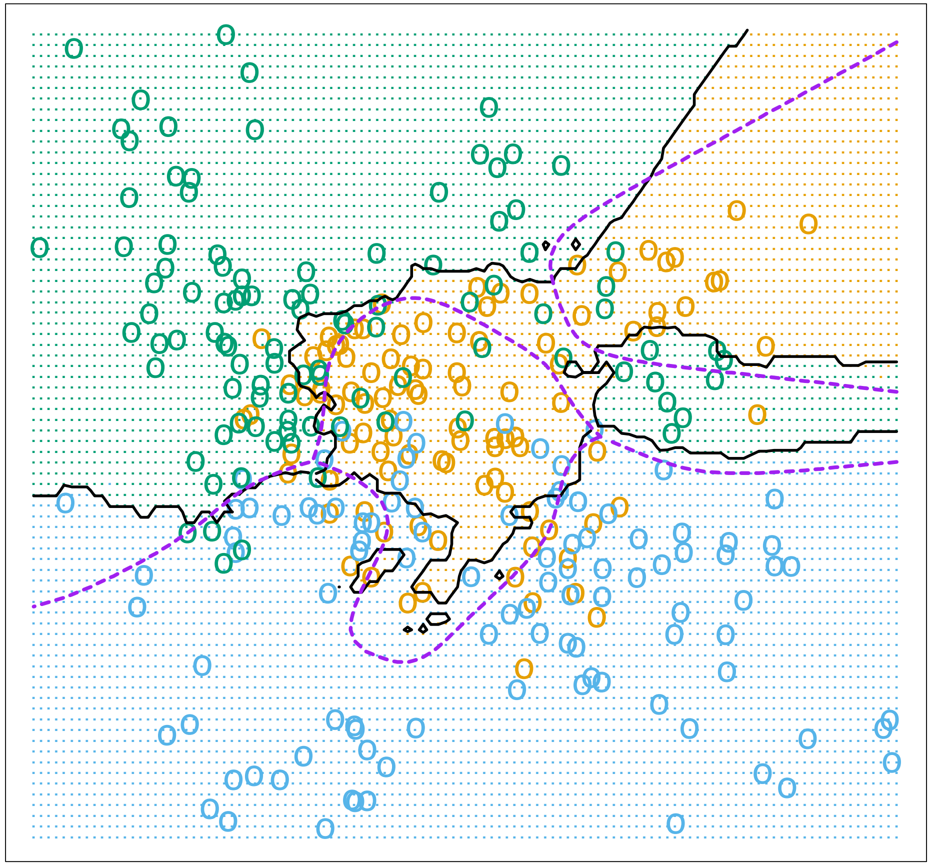

K = 1
K = All
Age
60
40
20
0
10
20
Number of Malignant Nodes
60
40
20
0
10
20
Number of Malignant Nodes

K-NN: SMOOTHING

1-Nearest Neighbor



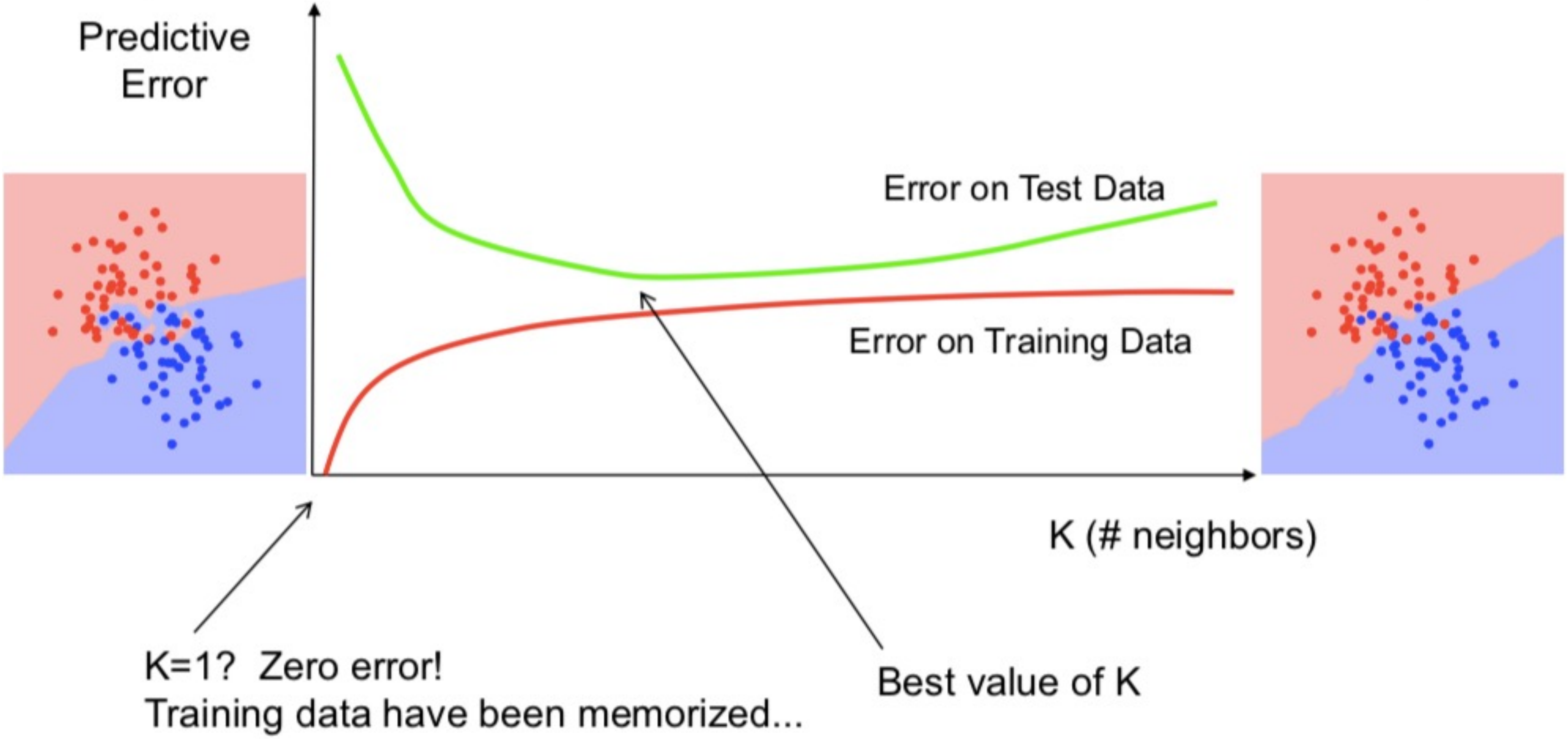
15-Nearest Neighbors



What is the training error when k=1?

Error rates and K

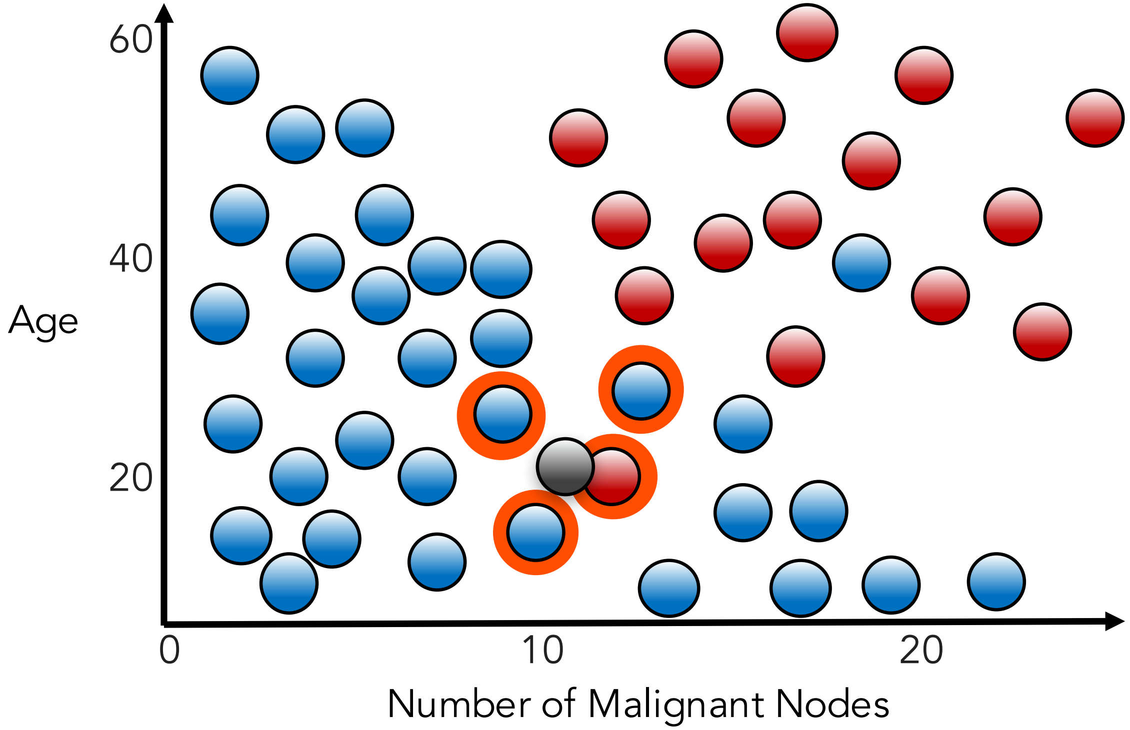
(More on model selection later)



https://medium.com/30-days-of-machine-learning/day-3-k-nearest-neighbors-and-bias-variance-tradeoff-75f84d515bdb

K-NN: PRACTICAL CHALLENGES

- How to pick k?
- What is the right measure of closeness/distance?



REVIEW: SUPERVISED LEARNING

- Learning a mapping from input to output, given a labeled set of input-output pairs, i.e. training dataset D

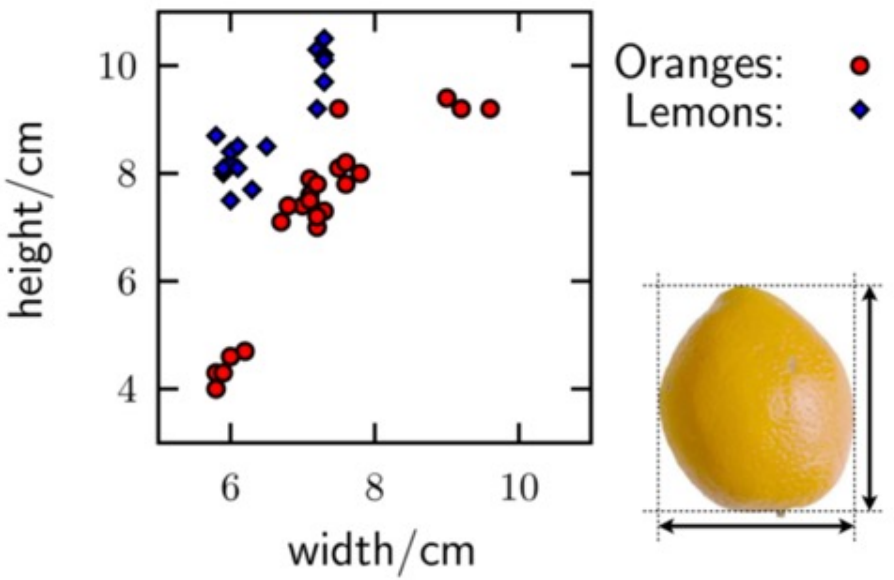
$$\{(\mathbf{x}_i, y_i)\}, i = 1, \cdots, N$$

- Each input instance represents an object/sample as a d-dimensional vector of features

- Classification: output is categorical (e.g. orange, lemon)

$$y_i \in \{1, \dots, C\}$$

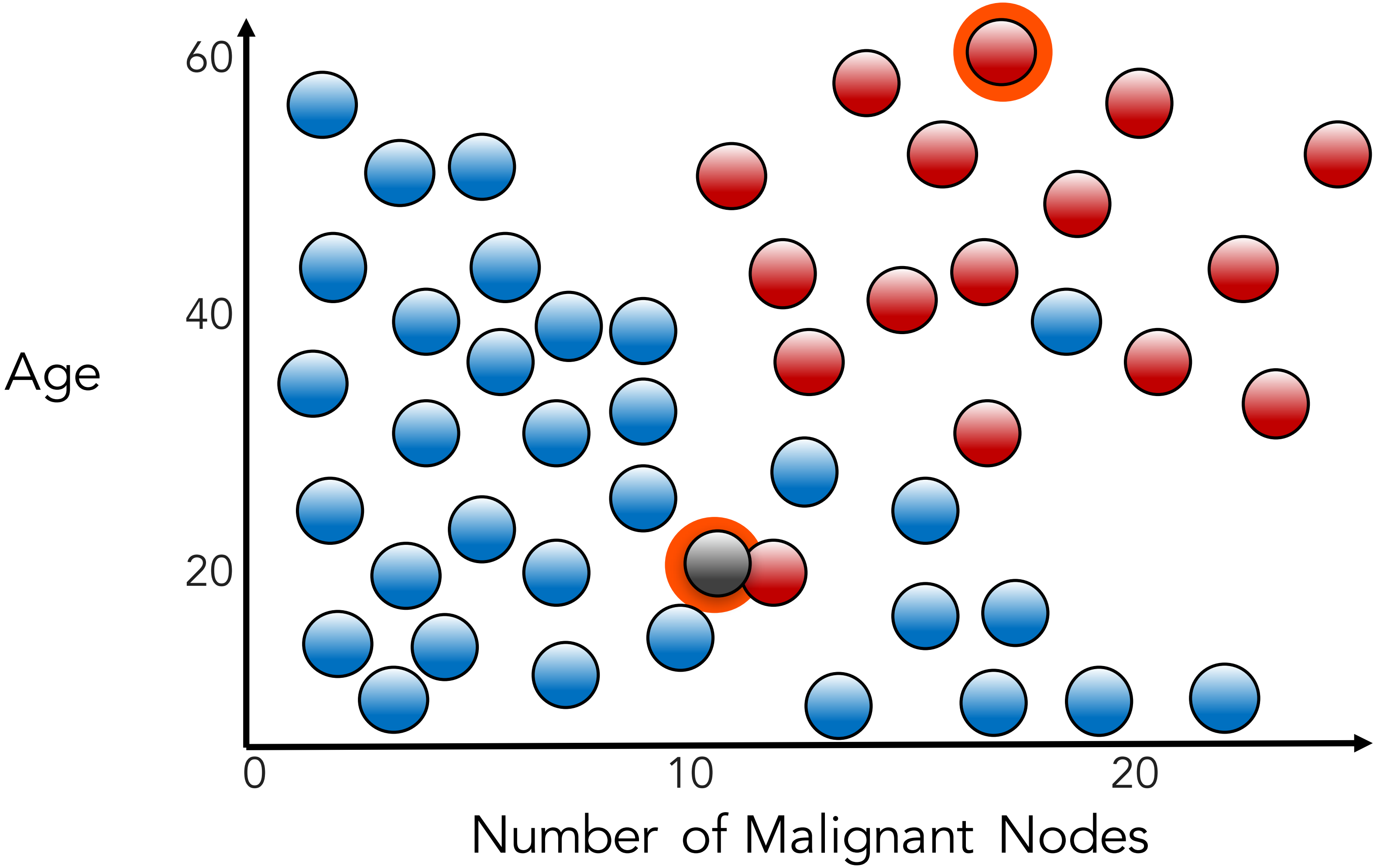
  - Binary vs. multiclass classification



http://homepages.inf.ed.ac.uk/imurray2/

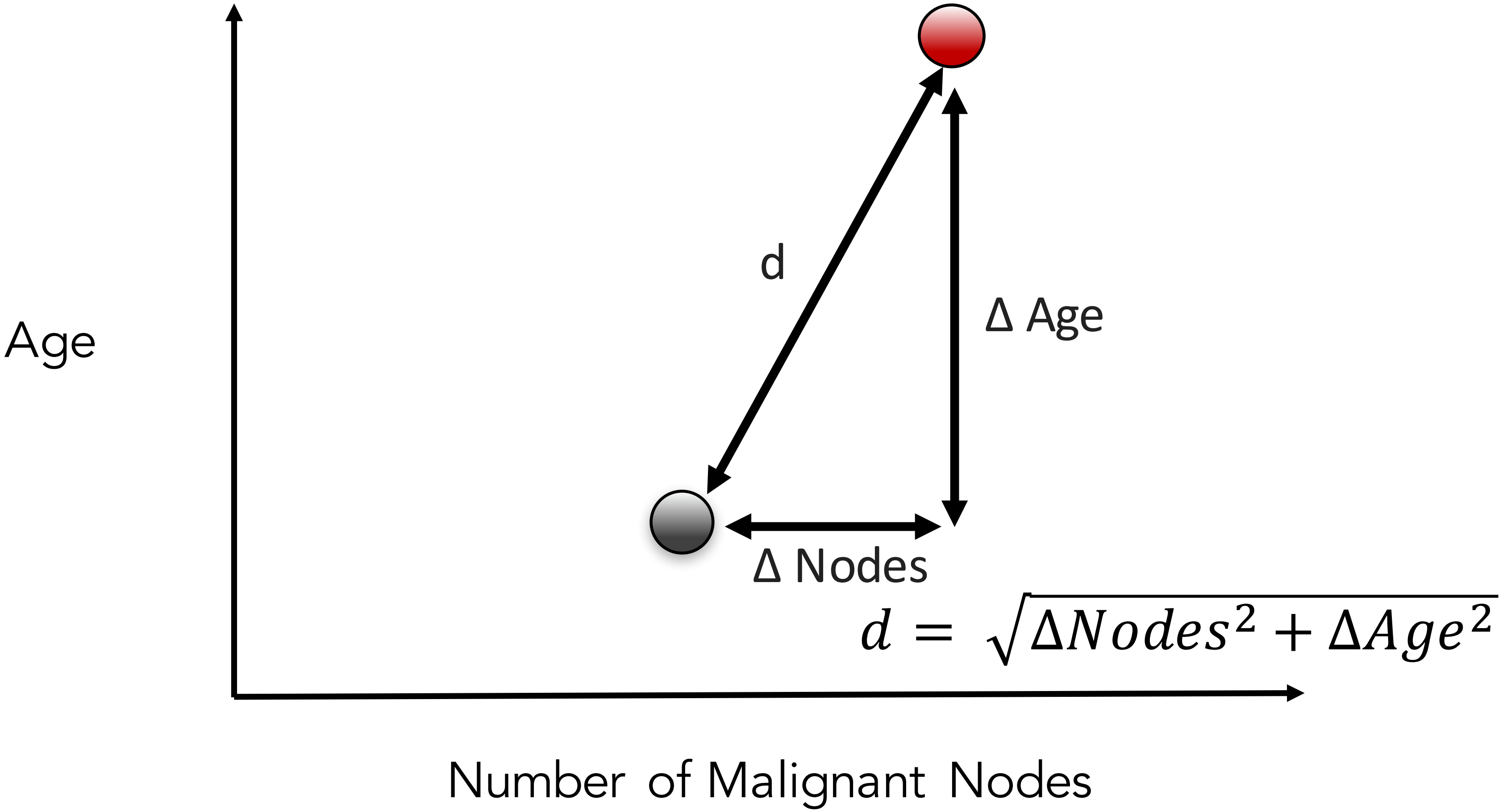
- Regression: output is real-valued

MEASUREMENT OF DISTANCE



What is the distance of these two points?

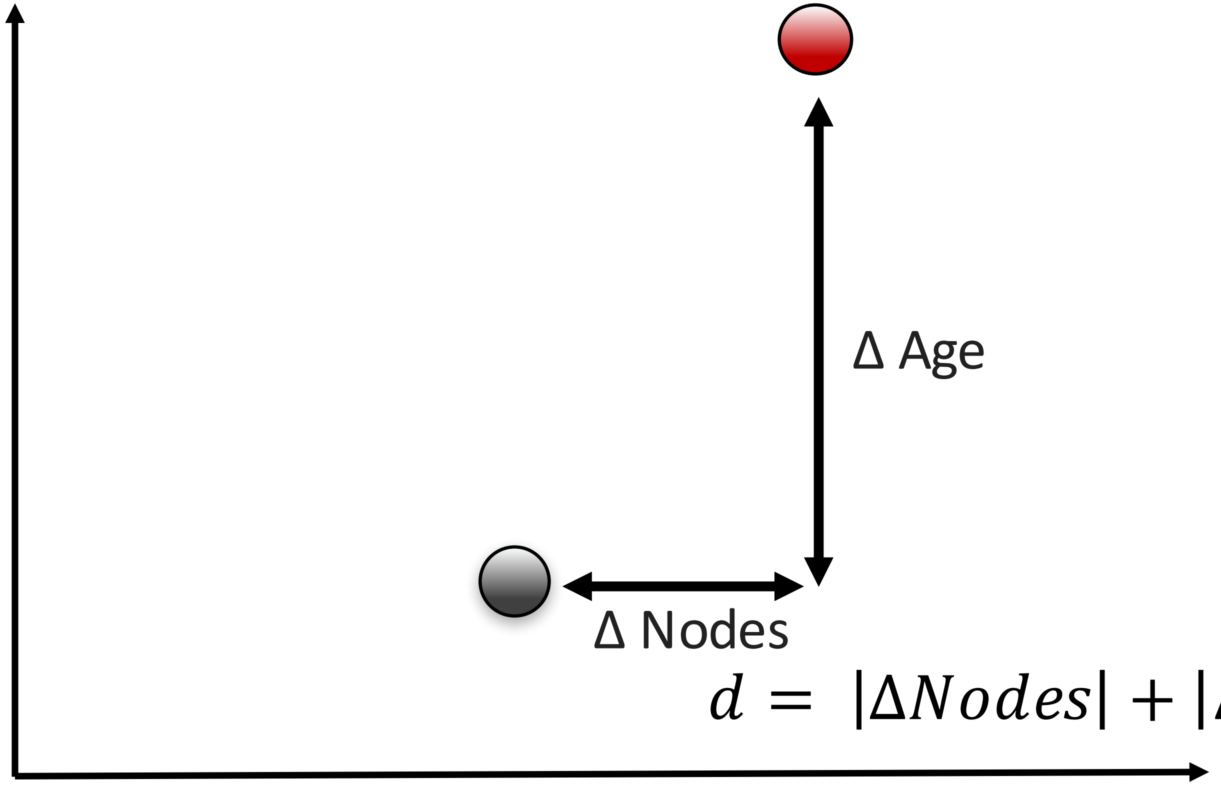
EUCLIDEAN DISTANCE



Also known as crow distance (green)
or L2 norm of the difference vector

MANHATTAN DISTANCE

Age



$$d = |\Delta Nodes| + |\Delta Age|$$

Number of Malignant Nodes

Also known as taxicab distance (purple)
or L1 norm of the difference vector

COMMON DISTANCE METRICS

Euclidean	$D(\mathbf{x}, \mathbf{z}) = \sqrt{\sum_{i=1}^{d} (x_i - z_i)^2}$
Manhattan	$D(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{d} \lvert x_i - z_i \rvert$
Minkowski	$D(\mathbf{x}, \mathbf{z}) = \left(\sum_{i=1}^{d} \lvert x_i - z_i \rvert^p \right)^{\frac{1}{p}}$

NORMS

$$\ell_p = \left(\sum_{i=1}^{N} |x_i|^p \right)^{1/p}, \text{for } p \geq 1$$

For $p = 1$, we get $\ell_1 = |x_1| + |x_2| + \ldots + |x_n|$

For $p = 2$, $\ell_2 = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$

For $p = 3$, $\ell_3 = \sqrt[3]{|x_1|^3 + |x_2|^3 + \ldots + |x_n|^3}$

For $p \rightarrow \infty$, $\ell_\infty = \max_i(|x_1|, |x_2|, \ldots, |x_n|)$

OTHER DISTANCE METRICS

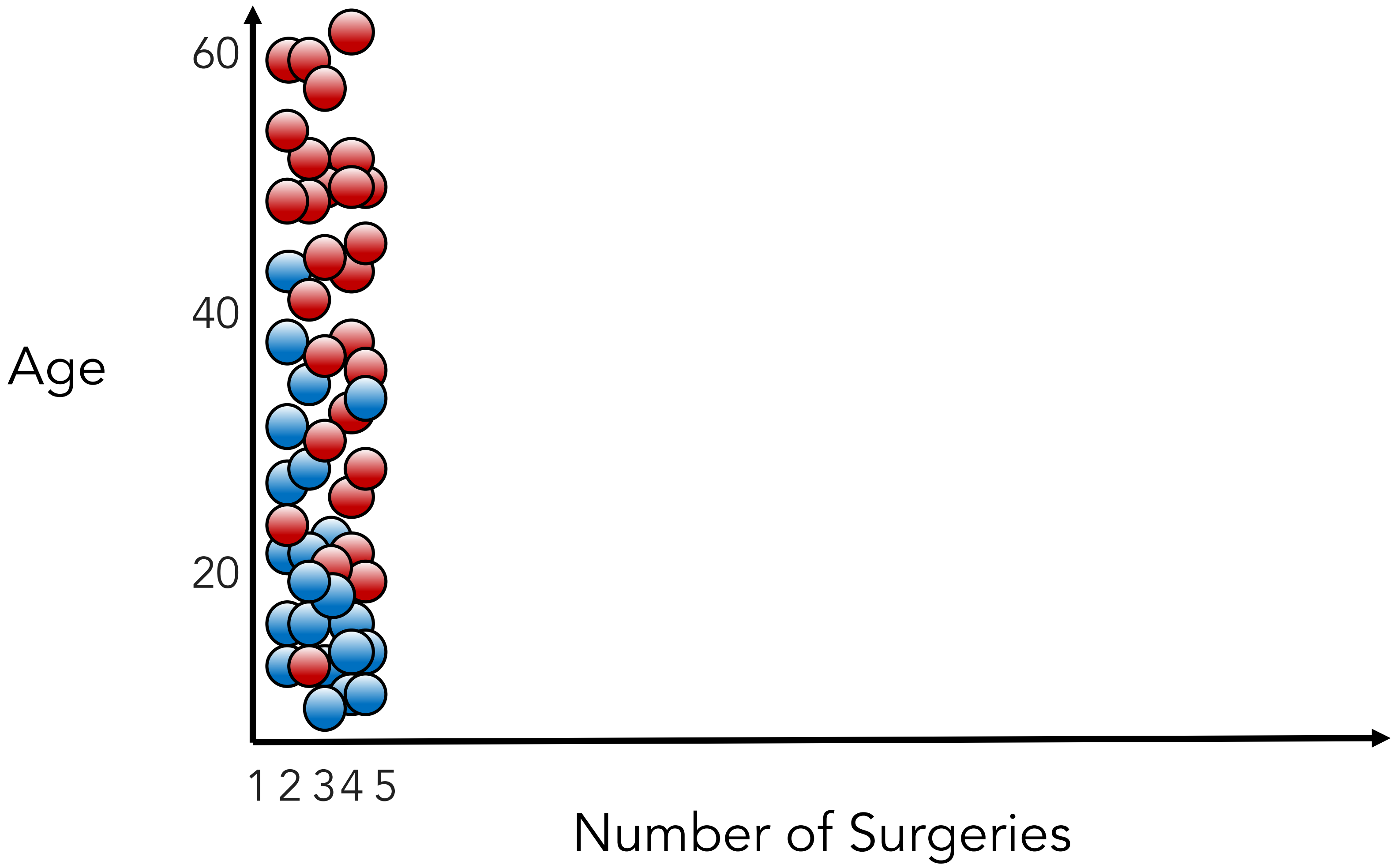
- Categorical/Integer-valued space
  - Hamming distance: $$D(\mathbf{x}, \mathbf{y}) = \frac{N_{\text{different}}(\mathbf{x}, \mathbf{y})}{N_{\text{total}}}$$
  - Canberra: $$D(\mathbf{x}, \mathbf{y}) = \sum \frac{|x_i - y_i|}{|x_i| + |y_i|}$$
- Boolean-valued space
  - Jaccard: $$D(\mathbf{x}, \mathbf{y}) = \frac{|\mathbf{x} \cap \mathbf{y}|}{|\mathbf{x} \cup \mathbf{y}|}$$

OTHER DISTANCE METRICS

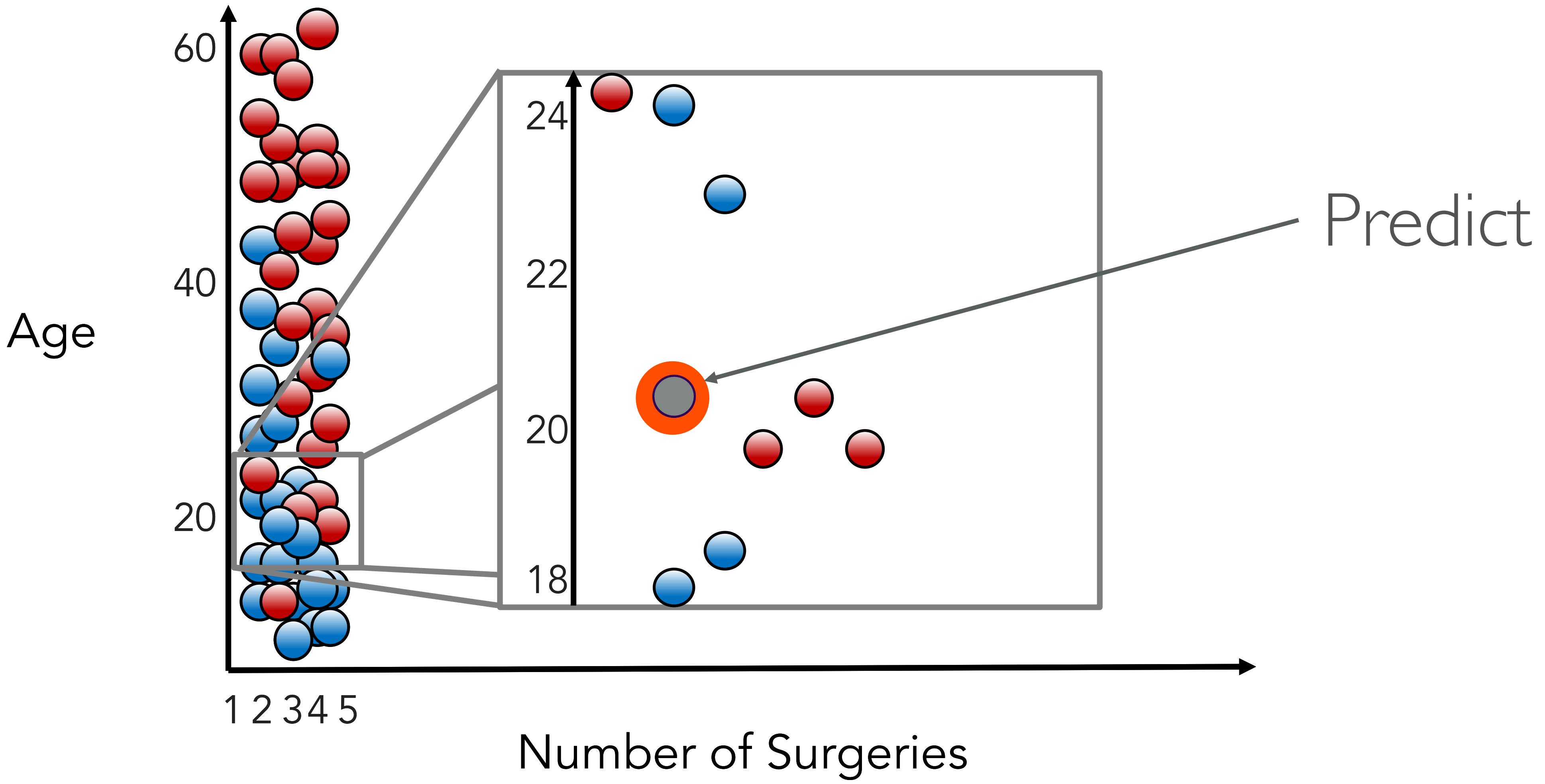
	Star Wars I: The Phantom Menace	Star Wars IV: A New Hope	Star Wars VII: The Force Awakens	Raiders of the Lost Arc	Casablanca	Singing in the Rain
Sam	3	4	3	4	1	2
Alice	4	5	5	4	2	1
Bob	1	2	3	2	5	3
Matt	2	3	3	1	4	4
Joyce	5	5	5	**?**	**?**	2

What's the hamming distance of the two records (considering column 1, 2, 3, 6)?

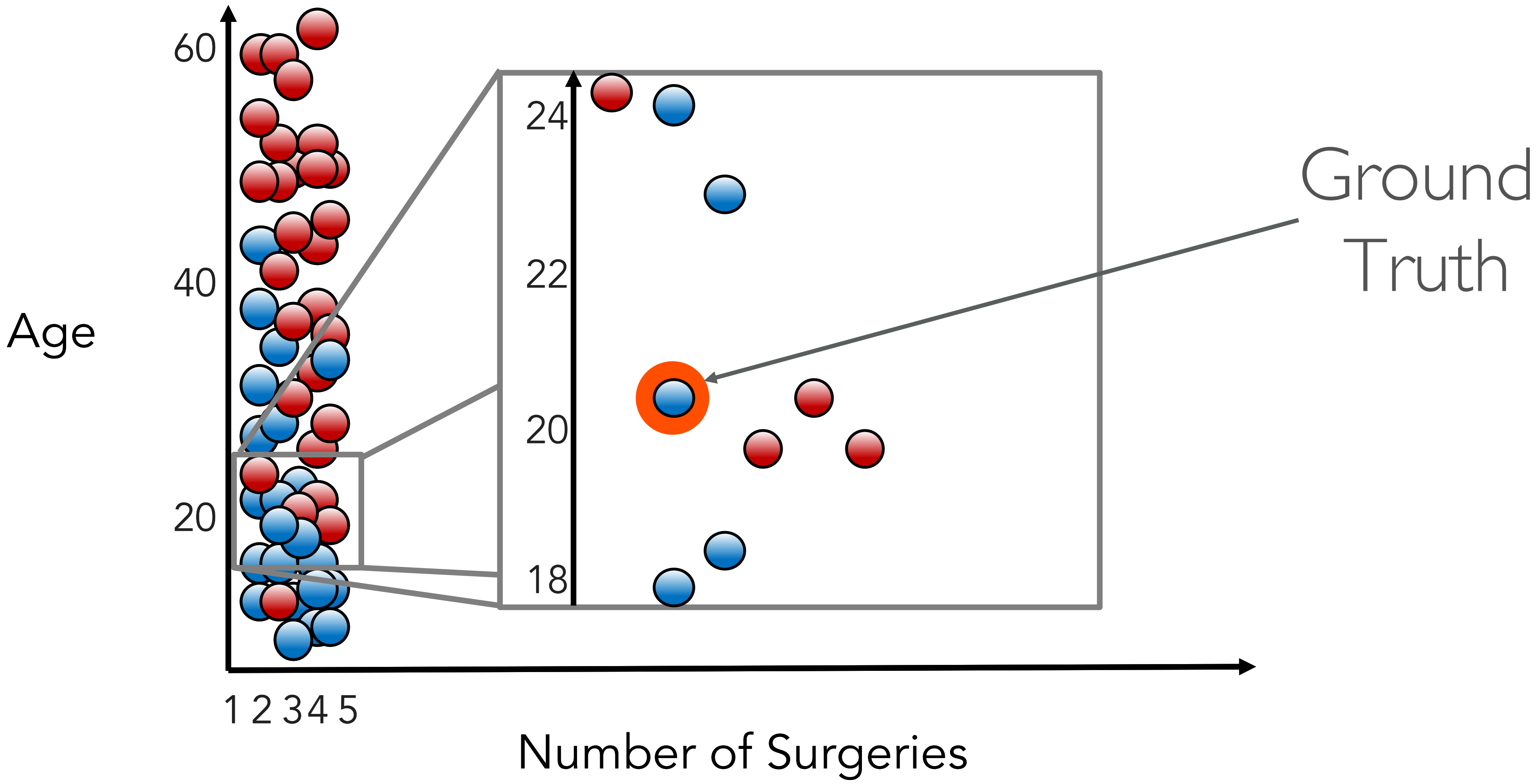
DOES SCALE MATTER?



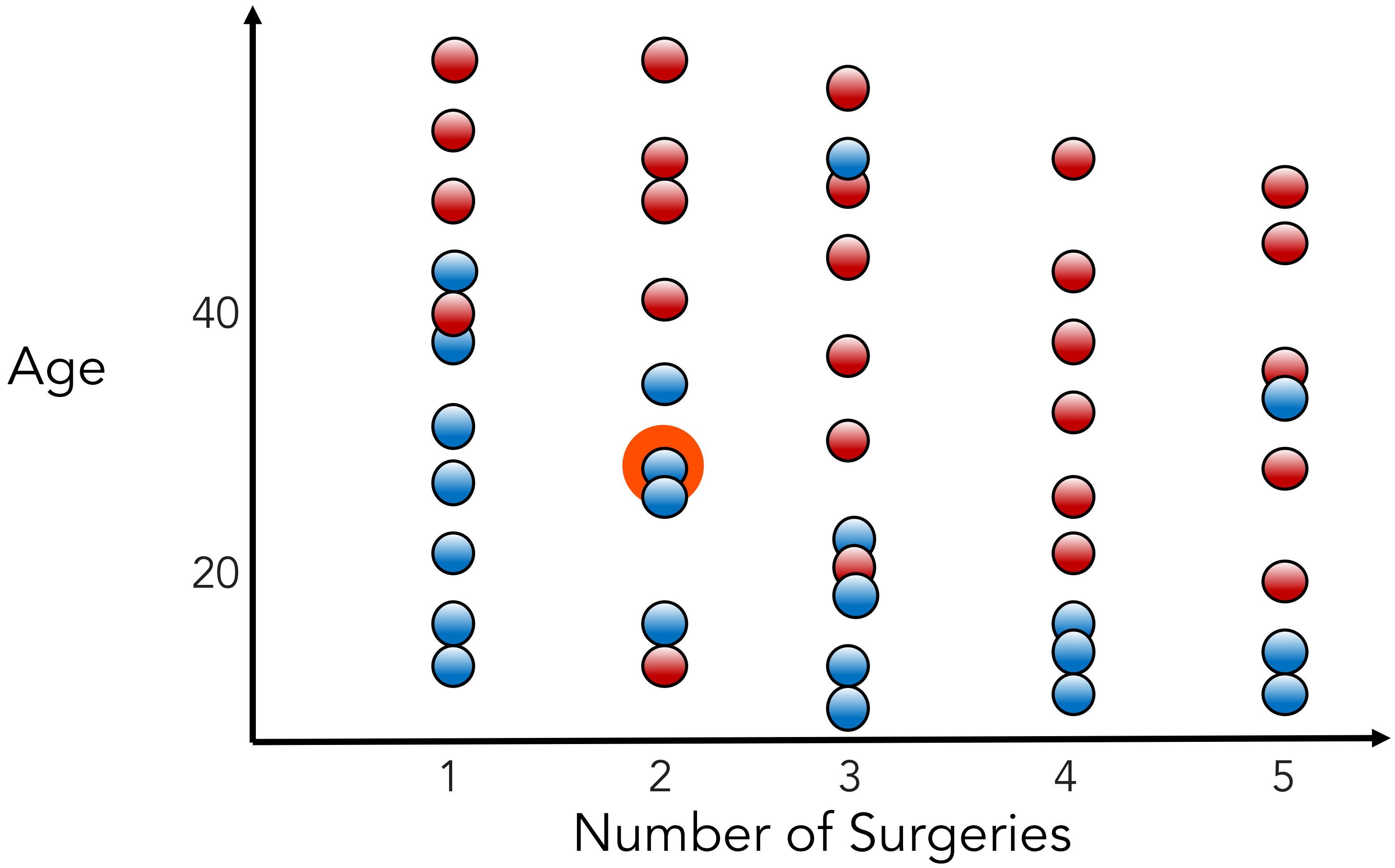
DOES SCALE MATTER?



DOES SCALE MATTER?



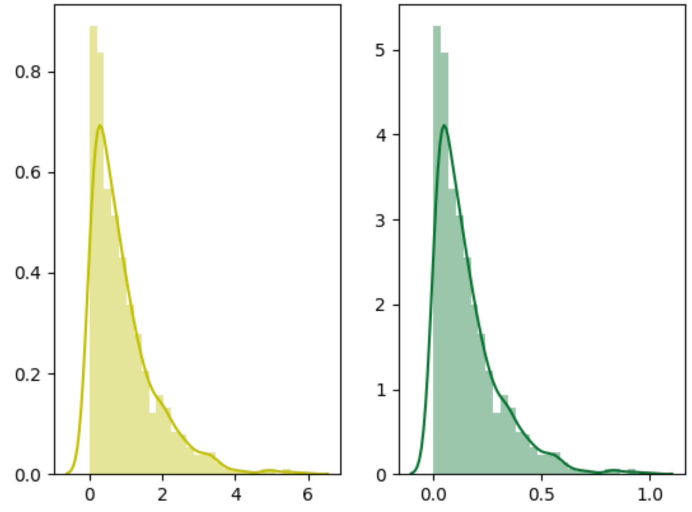
DOES SCALE MATTER?



FEATURE SCALING

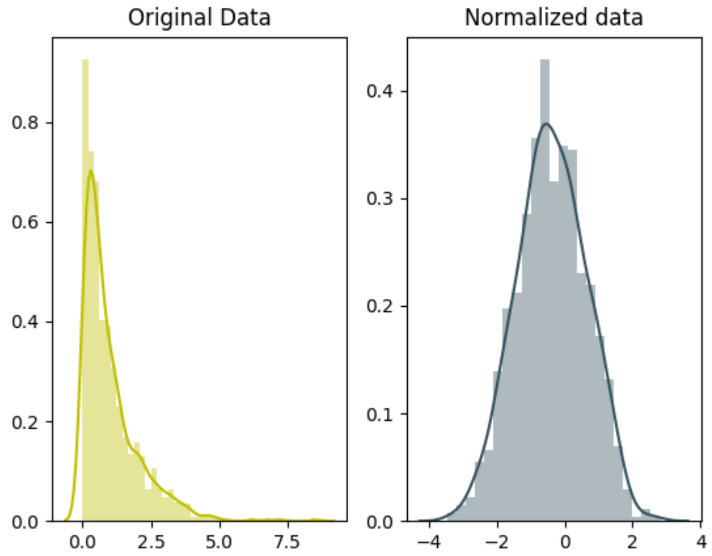
- Min-max normalization: Scale data to fixed range [0, 1] or [a,b]

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \qquad x' = a + \frac{(x - \min(x))(b - a)}{\max(x) - \min(x)}$$



- Standardization (Z-score normalization): Center data to zero mean and scale by unit variance

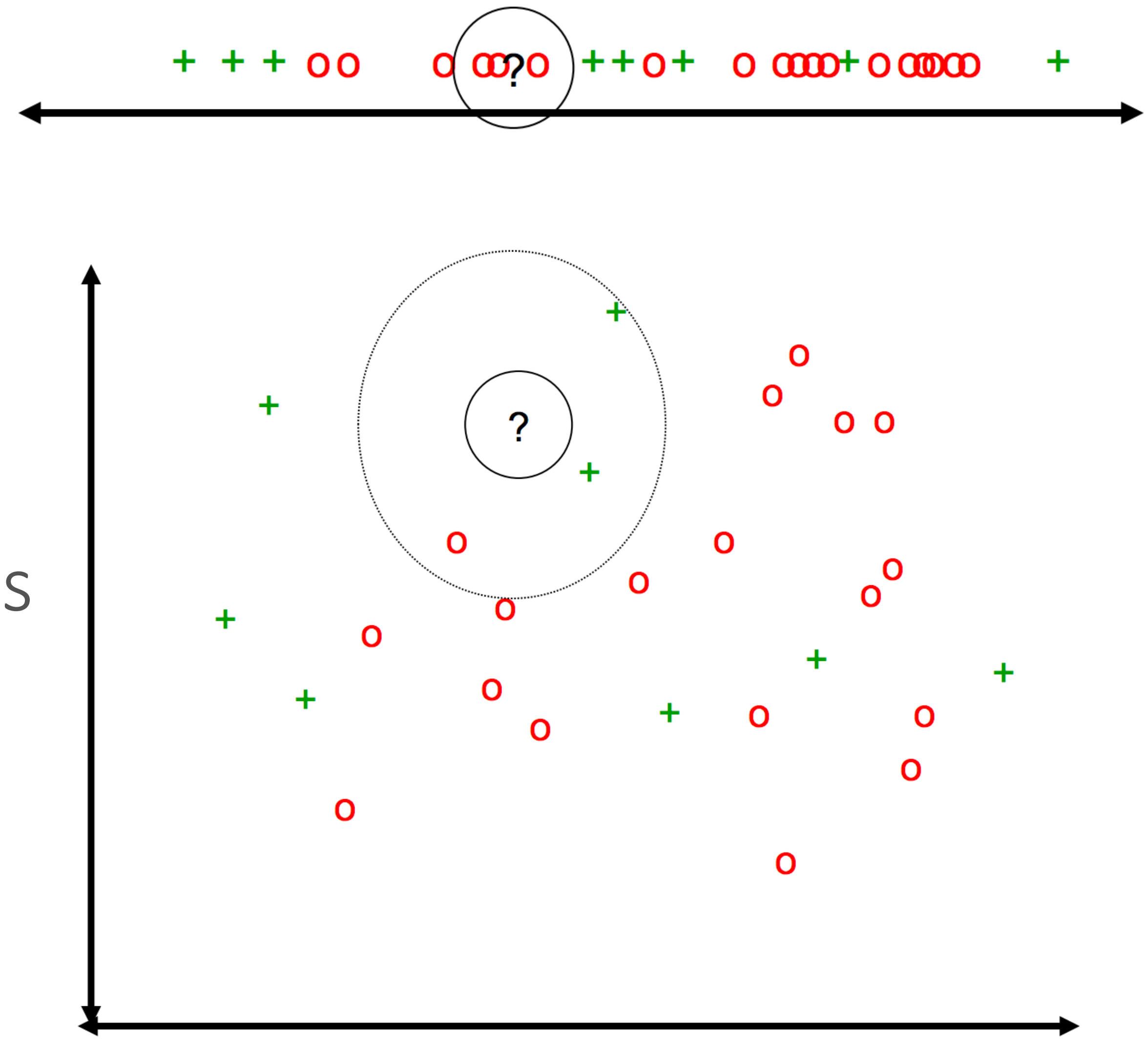
$$x' = \frac{x - \bar{x}}{\sigma}$$



K-NN: IRRELEVANT FEATURES

- Irrelevant / noisy features may hurt performance since it adds random perturbations to the distance measure

- Example: 1-D data, what happens if we add noisy attribute?



K-NN: CHARACTERISTICS

- Instance-based (lazy) learning (as vs. model-based eager learning)
- Non-parametric (as vs. parametric)
- Easy to understand and implement
- Can model complex decision boundaries quite well (depending on k)
- Memory intensive (needs to store all the data) – can use clustering
- Can be fooled by irrelevant features

ONLINE DEMO

http://vision.stanford.edu/teaching/cs231n-demos/knn/

NETFLIX PRIZE (2006-2009)

$1M prize for 10% improvement

Collaborative Filtering

- Estimate rating by user x for item I
- Collaborative filtering
  - User-based: find similar users to user x
- Select ***k***-nearest neighbors, compute the rating

	Star Wars I: The Phantom Menace	Star Wars IV: A New Hope	Star Wars VII: The Force Awakens	Raiders of the Lost Arc	Casablanca	Singing in the Rain
Sam	3	4	3	4	1	2
Alice	4	5	5	4	2	1
Bob	1	2	3	2	5	3
Matt	2	3	3	1	4	4
Joyce	5	5	5	?	?	2

Collaborative Filtering

- Estimate rating by user x for item I
- Collaborative filtering
  - User-based: find similar users to user x
  - Item-based: find similar items rated by user x
- Select ***k***-nearest neighbors, compute the rating

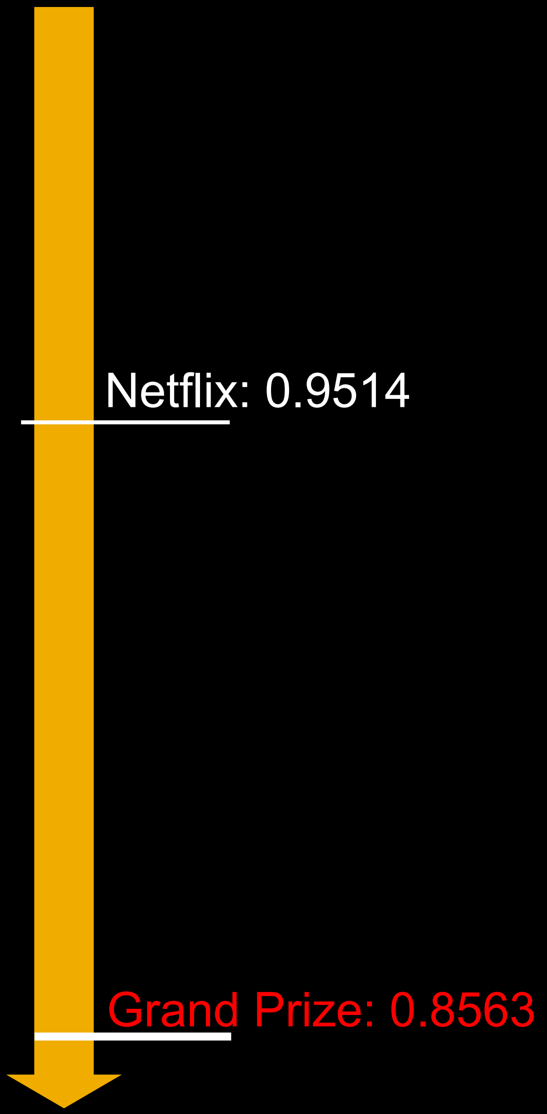
$$r_{xi} = \frac{\sum_{j \in N(i;x)} s_{ij} \cdot r_{xj}}{\sum_{j \in N(i;x)} s_{ij}}$$

s_{ij}… similarity of items ***i*** and ***j***
r_{xj}…rating of user ***u*** on item ***j***
N(i;x)… items similar to ***i*** rated by ***x***

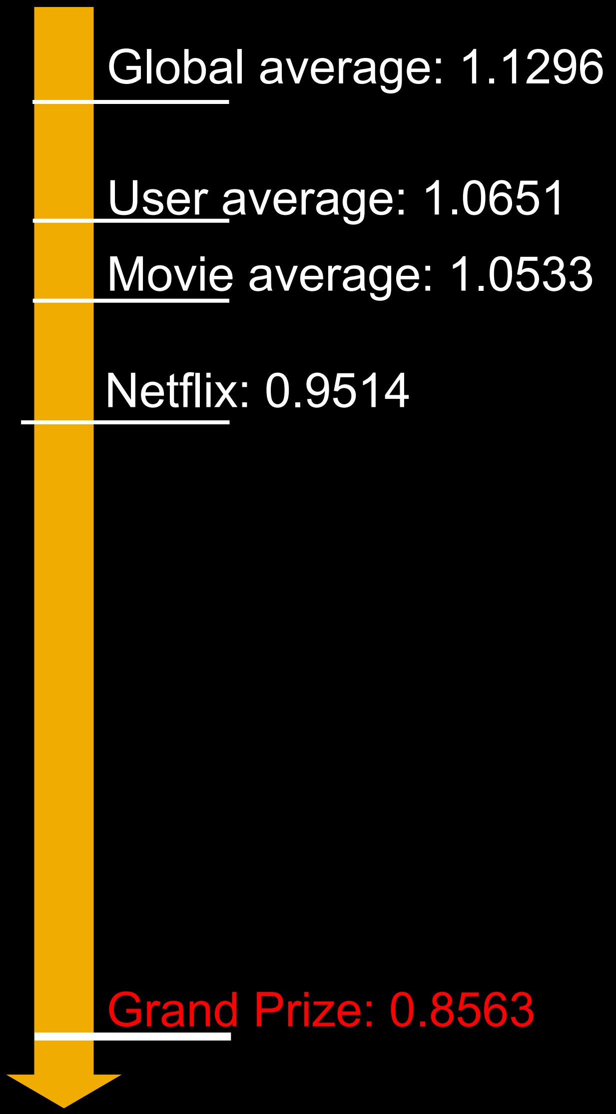
	Star Wars I: The Phantom Menace	Star Wars IV: A New Hope	Star Wars VII: The Force Awakens	Raiders of the Lost Arc	Casablanca	Singing in the Rain
Sam	3	4	3	4	1	2
Alice	4	5	5	4	2	1
Bob	1	2	3	2	5	3
Matt	2	3	3	1	4	4
Joyce	5	5	5	?	?	2

	Star Wars I: The Phantom Menace	Star Wars IV: A New Hope	Star Wars VII: The Force Awakens	Raiders of the Lost Arc	Casablanca	Singing in the Rain
Sam	3	4	3	4	1	2
Alice	4	5	5	4	2	1
Bob	1	2	3	2	5	3
Matt	2	3	3	1	4	4
Joyce	5	5	5	?	?	2

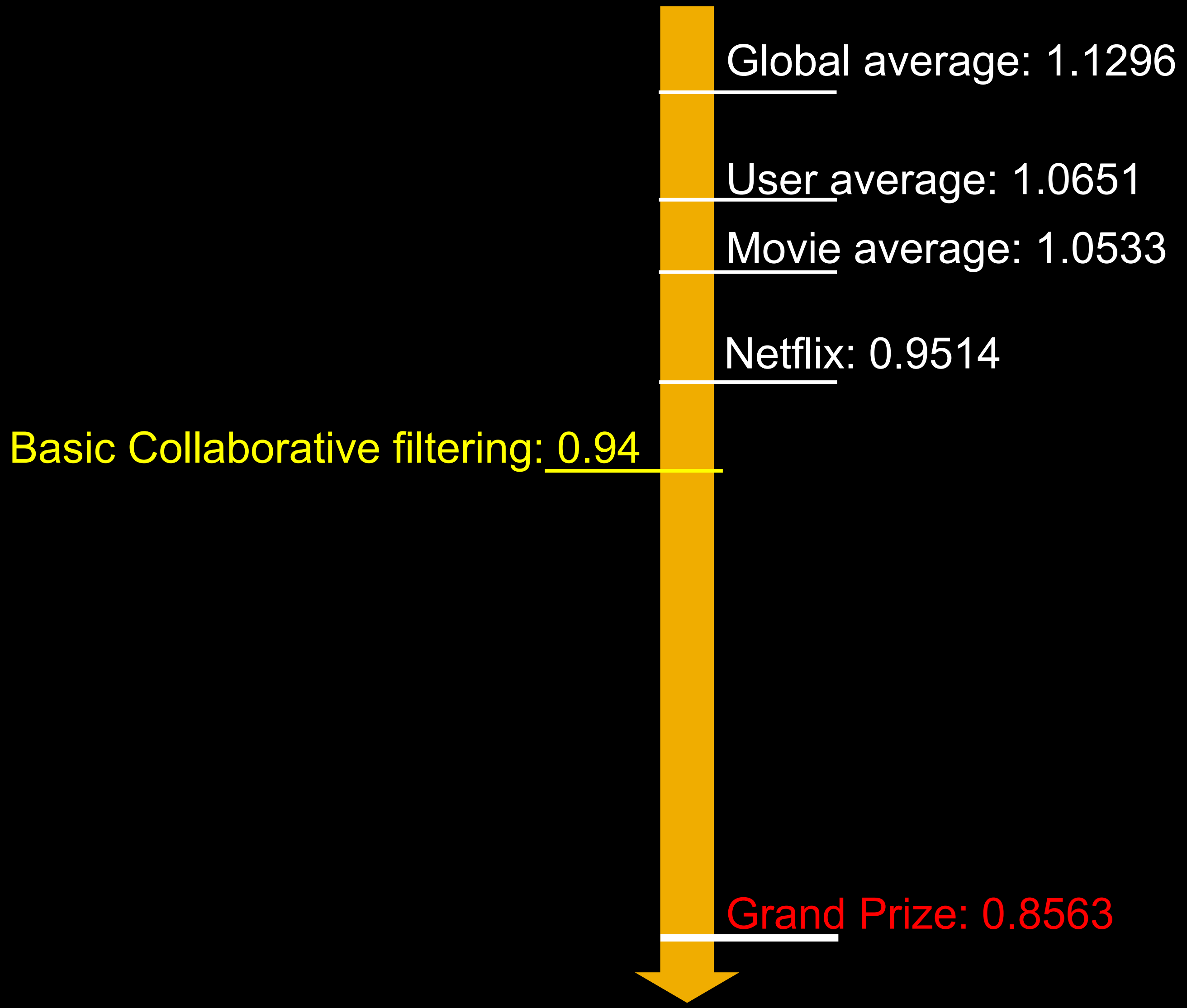
Netflix Prize



Netflix Prize



Netflix Prize



HOMEWORK #1 ANNOUNCEMENT

- Out 8/28, Due 9/12 @ 11:59 PM ET on Gradescope
- 4 questions
  - Q1-Q2: Get familiar with Python
    - Numerical programming (Numpy)
    - Dataset loading and visualization (Pandas and other libraries)
  - Q3-Q4: kNN
    - Implement kNN (use Numpy)
    - Evaluate kNN (use sklearn)