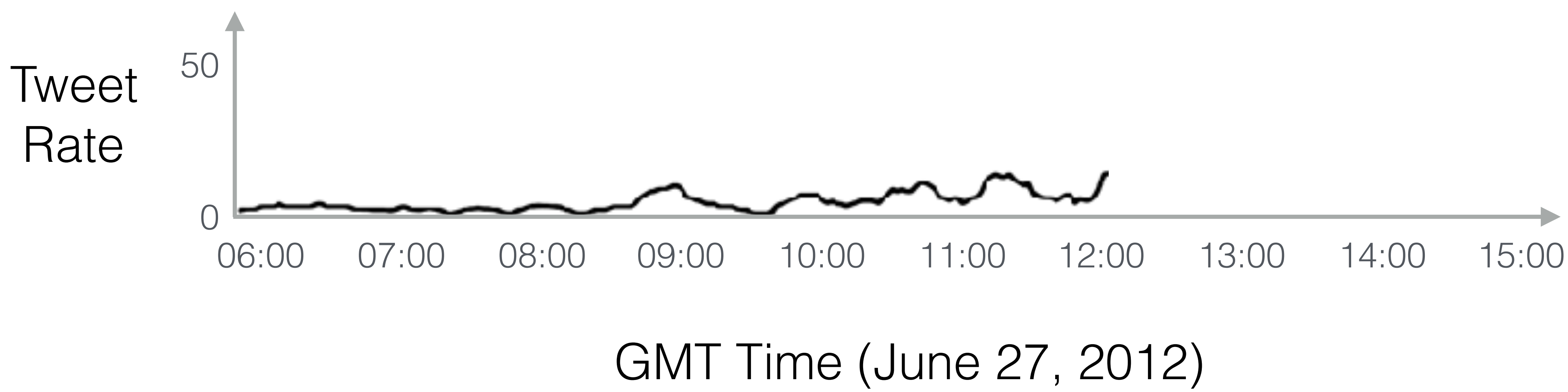
Adaptive Nearest Neighbor Classification and Regression Based on Decision Trees

slides by
George Chen
Carnegie Mellon University
Fall 2017

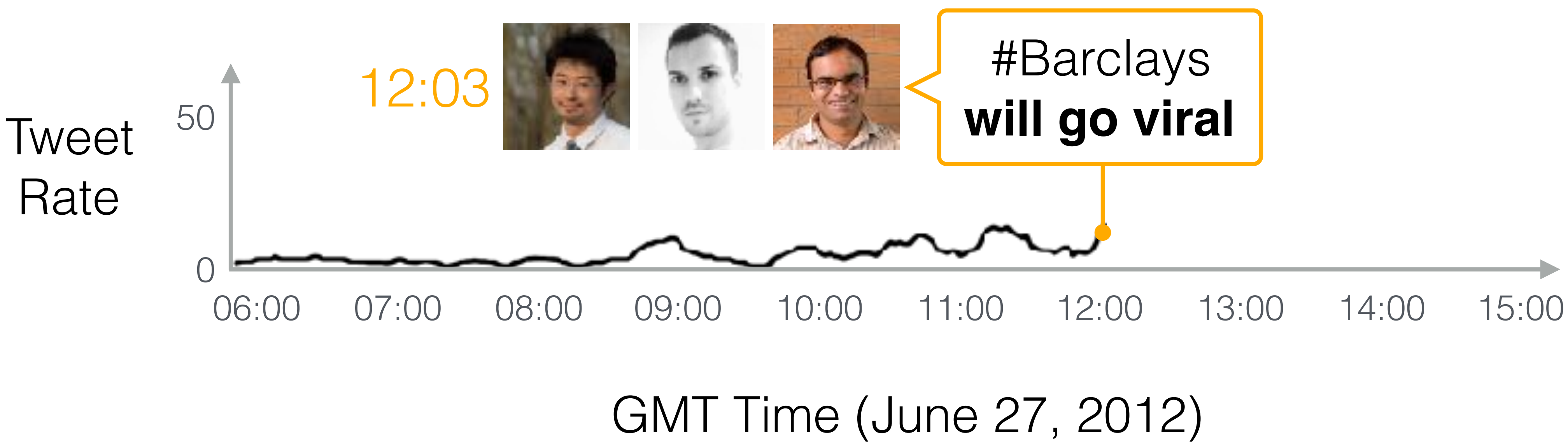
NN and Kernel Classification and Regression


BARCLAYS

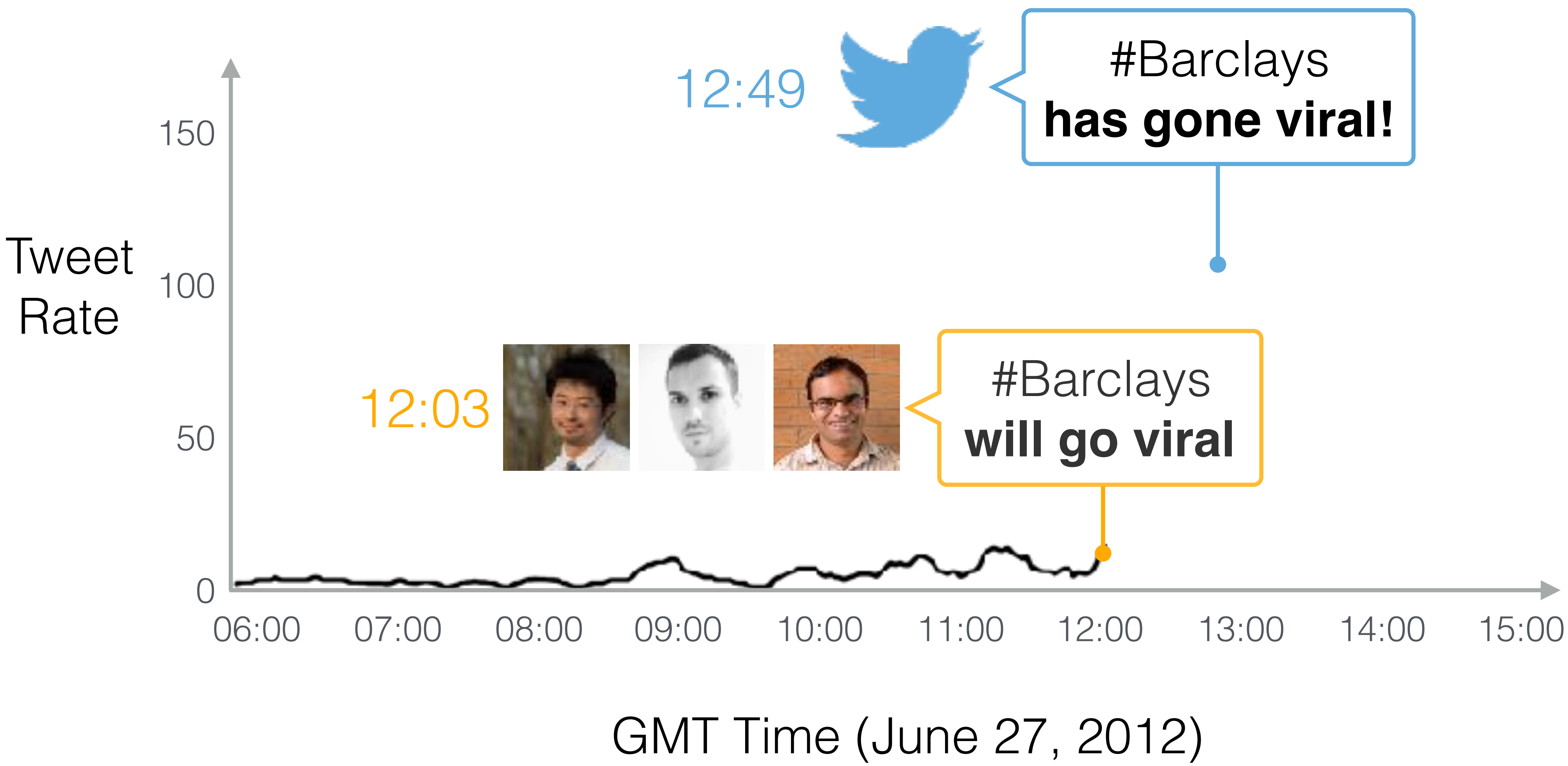
News Activity for #Barclays



News Activity for #Barclays



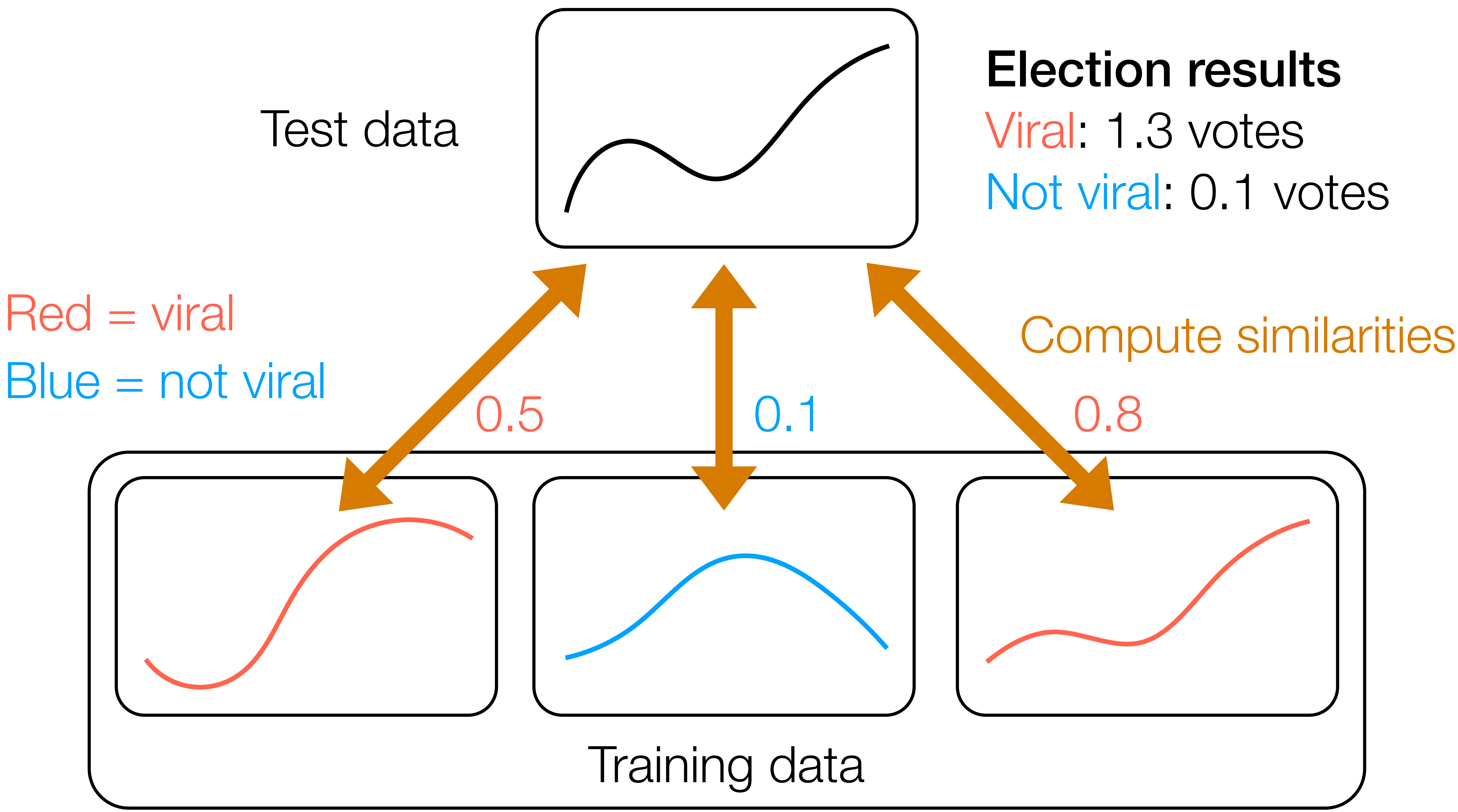
News Activity for #Barclays



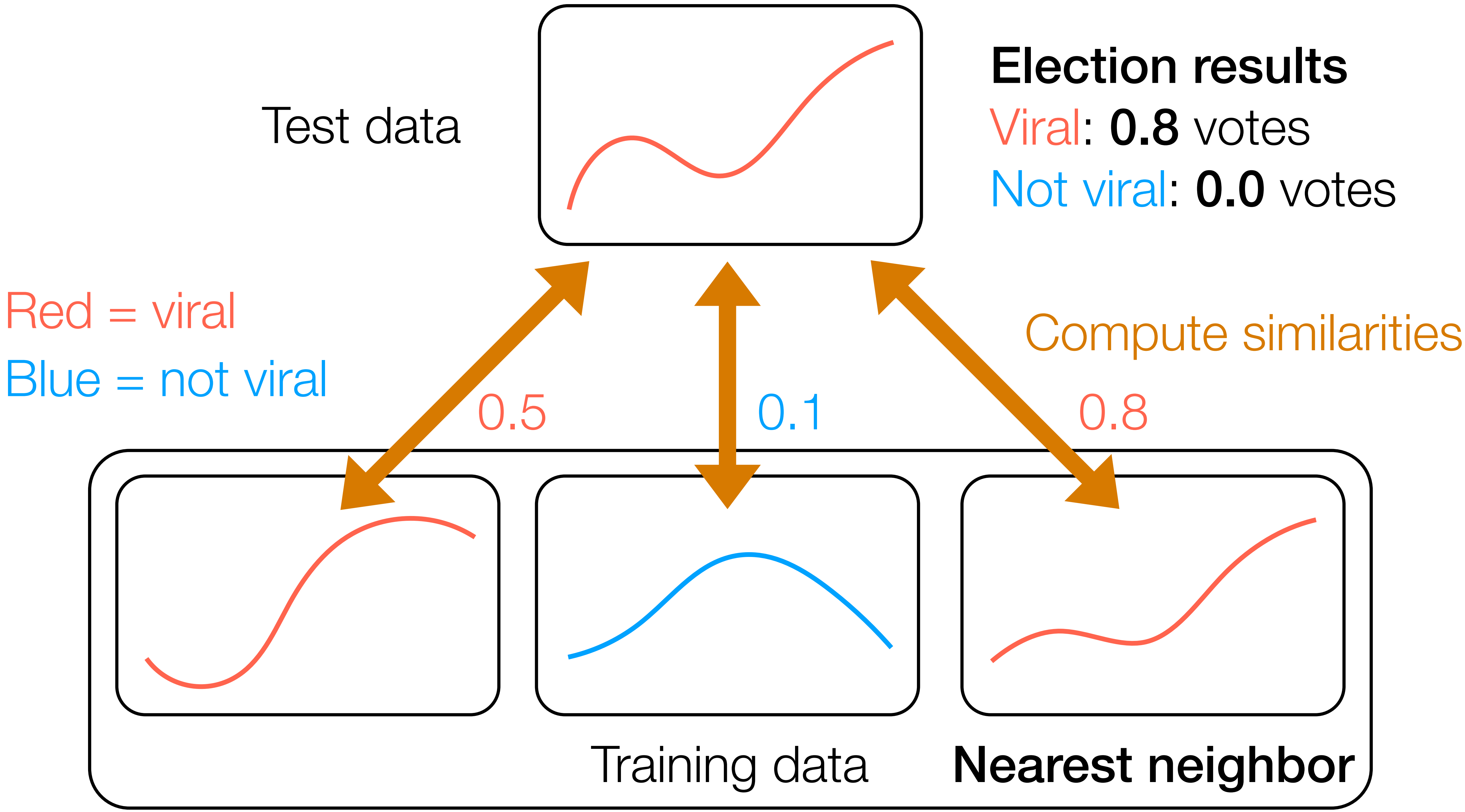
How we did this: **weighted majority voting**

Chen, Nikolov, and Shah. A Latent Source Model for Nonparametric Time Series Classification. NIPS 2013.

Weighted Majority Voting



Nearest Neighbor Classification



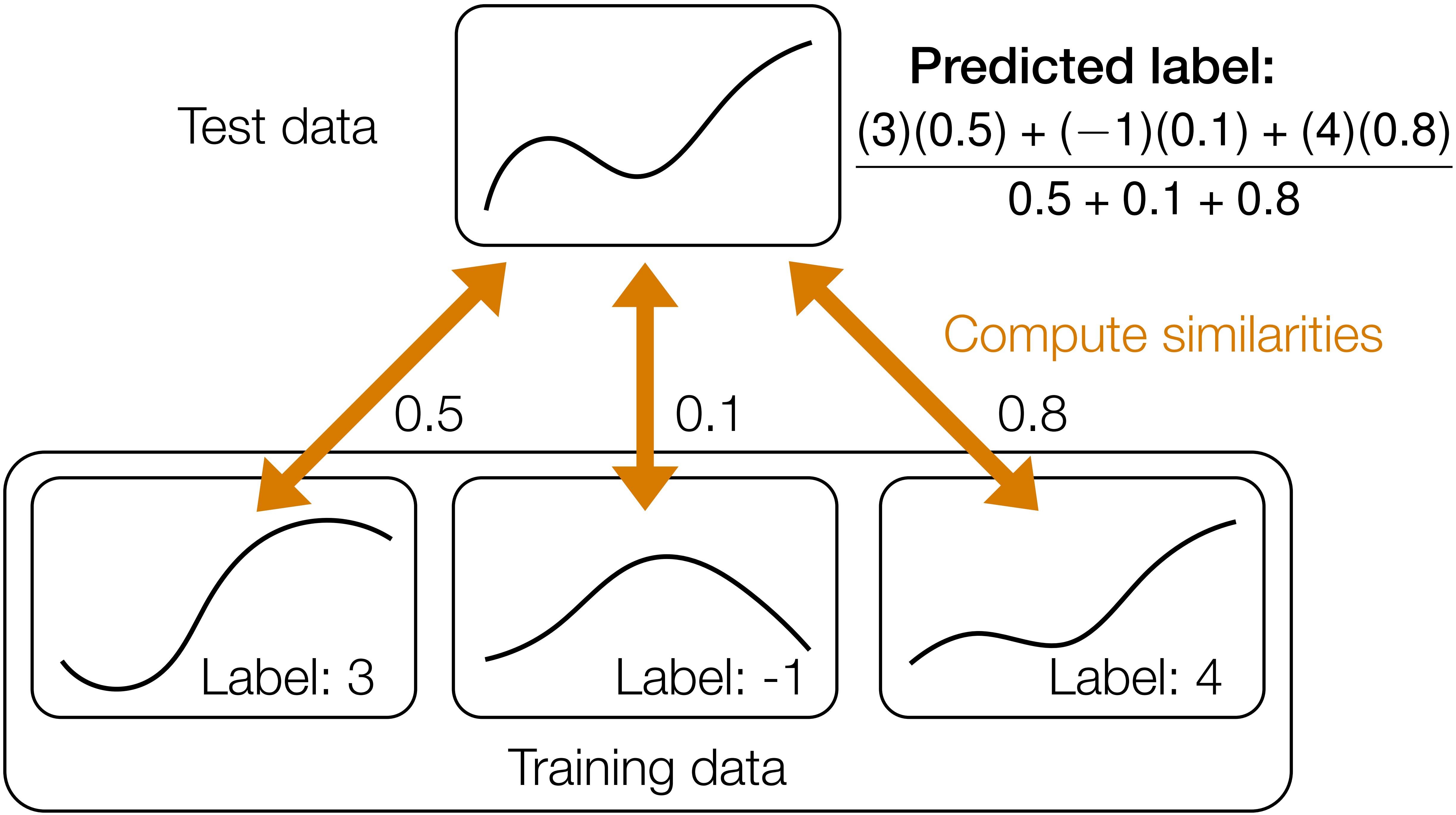
NN Classification Variants

- ***k*-NN classification:** consider k most similar training data to test data point
  - **Weighted:** when tallying up votes, use the similarities that we computed
  - **Unweighted:** when tallying up votes, have each of the k nearest neighbors have an equal vote of 1
    (usually k-NN classification refers to unweighted case)
- **Fixed-radius near neighbor classification:** consider all training data at least some similarity threshold close to test data point (i.e., use all training data distance $\leq h$ away)
  - Once again, can use weighted or unweighted votes

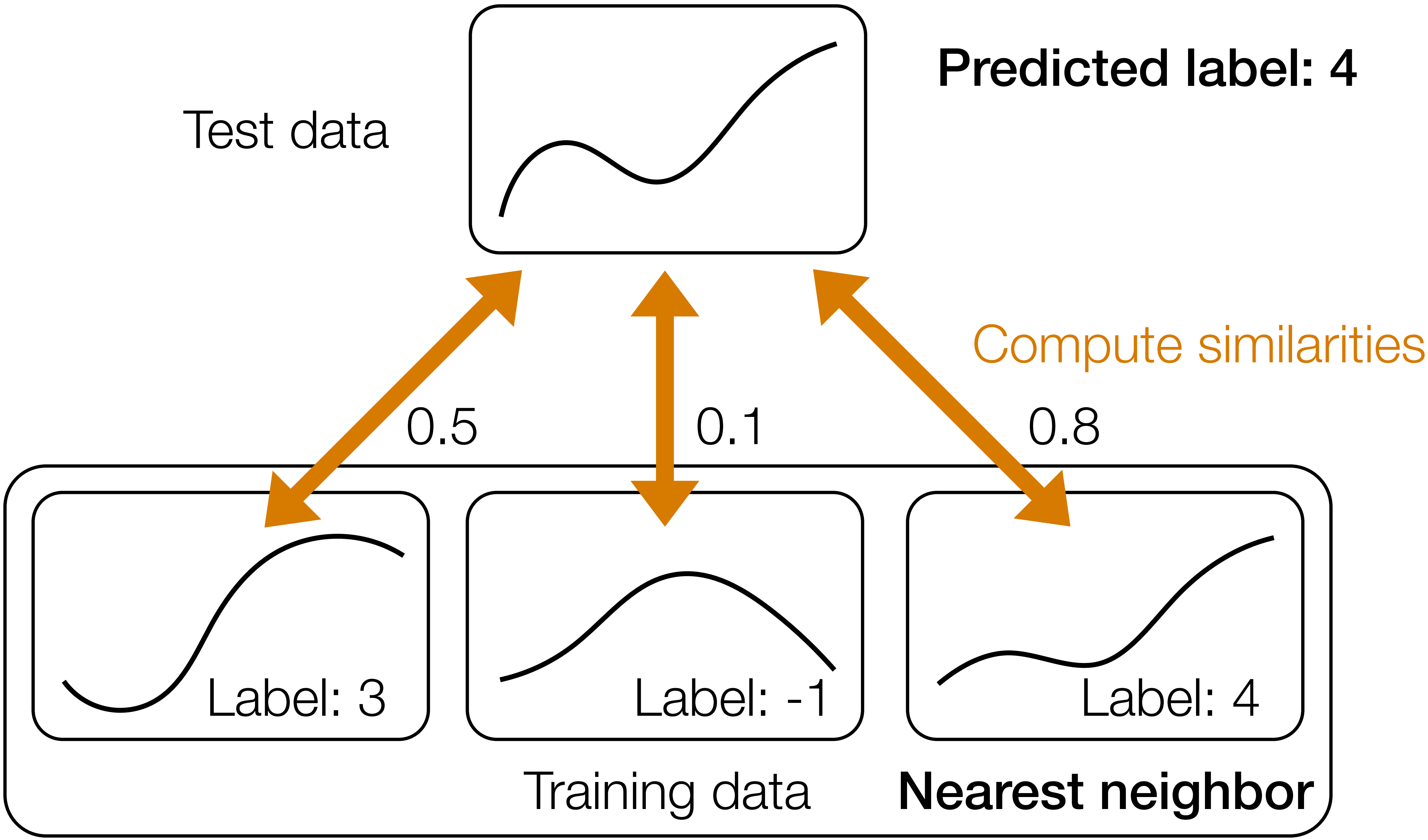
Regression: Each label is *continuous* instead of *discrete*

Kernel Regression

Weighted average instead of weighted majority vote



NN Regression

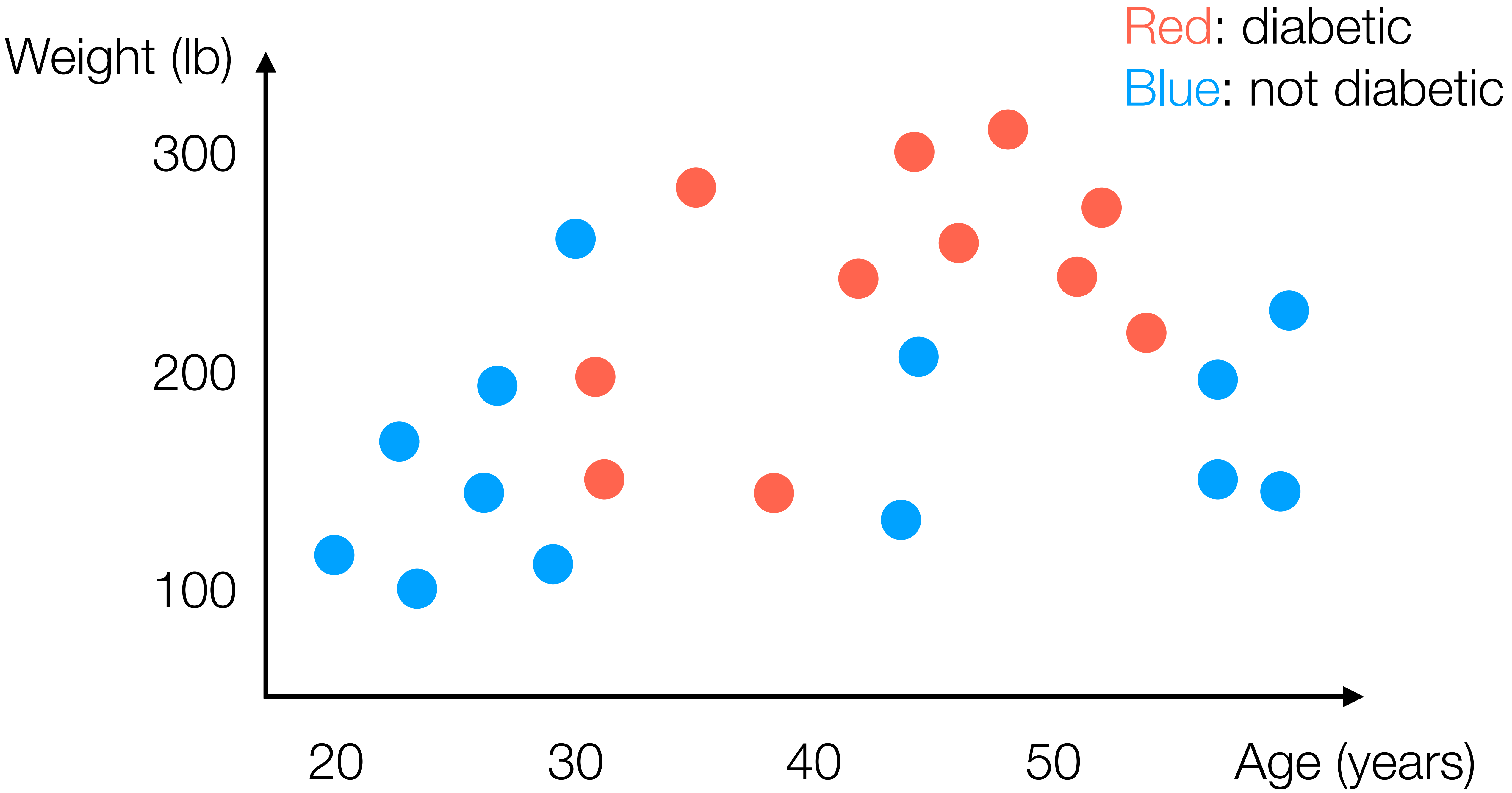


Just like classification: *k*-NN and fixed-radius NN variants, also weighted and unweighted

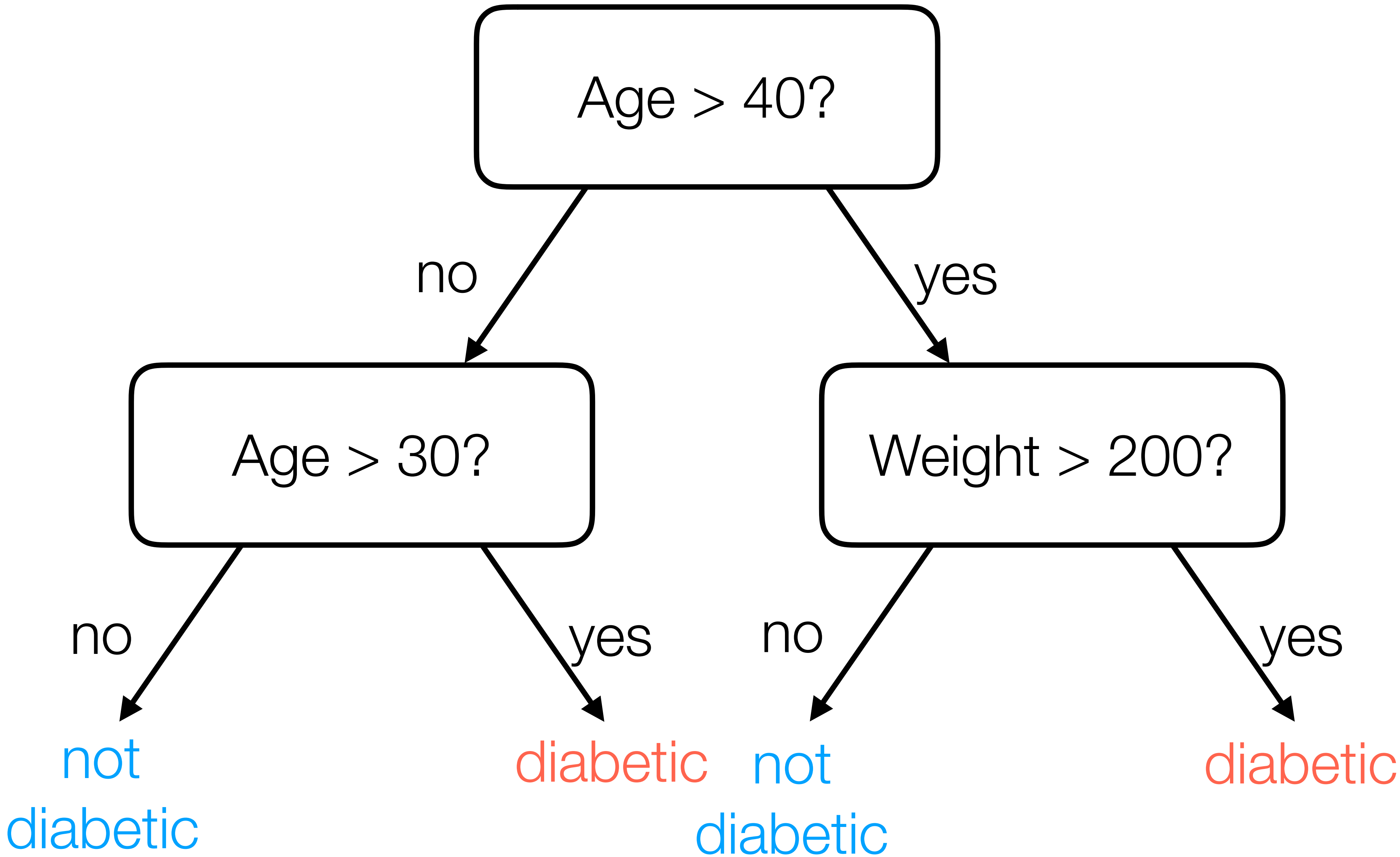
“Adaptive” nearest neighbors: learn the similarity function

Decision Trees

Example Made-Up Data



Example Decision Tree

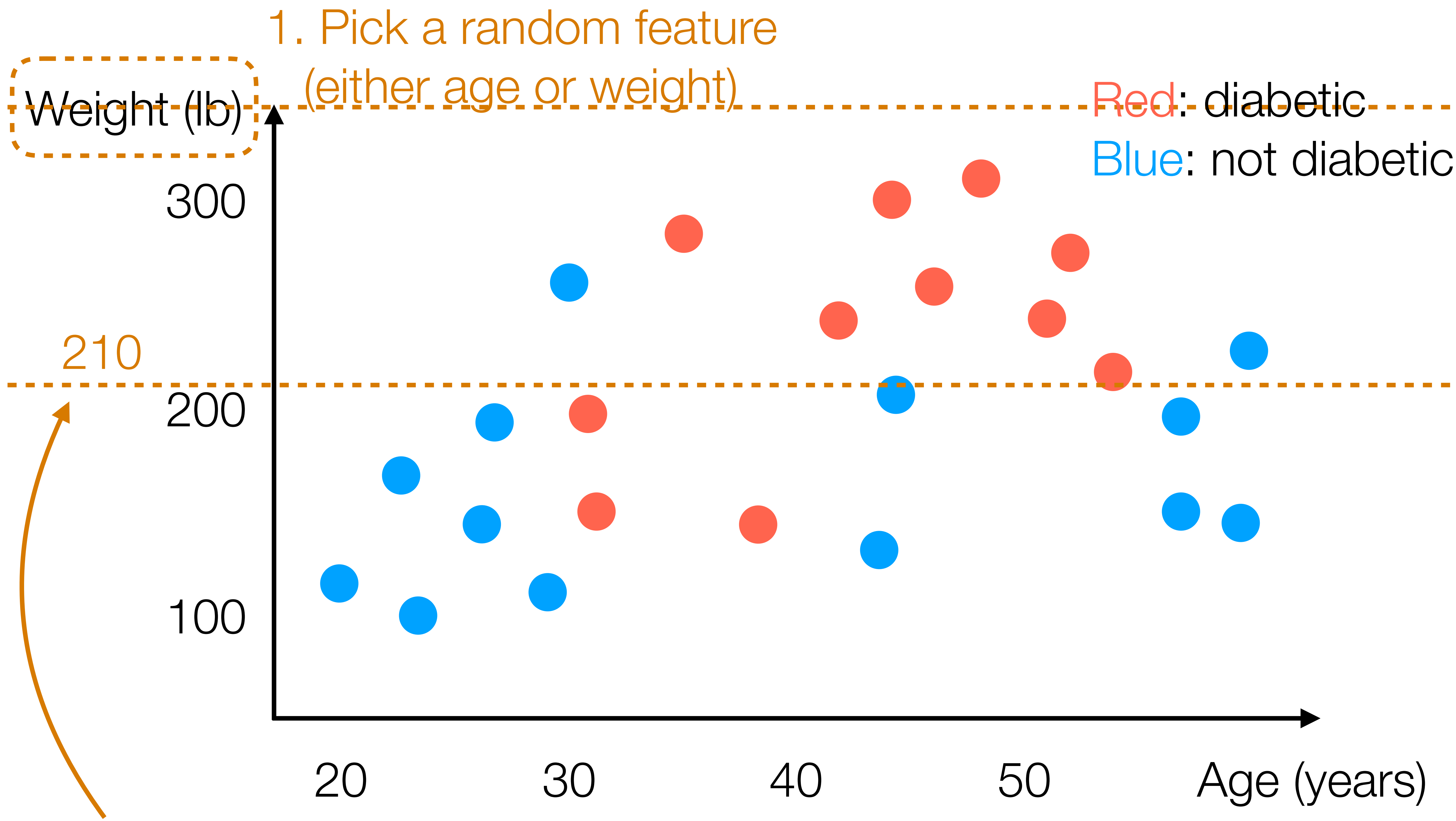


Learning a Decision Tree

- Many ways: general approach actually looks a lot like divisive clustering *but accounts for label information*

- I'll show one way (that nobody actually uses in practice) but it's easy to explain

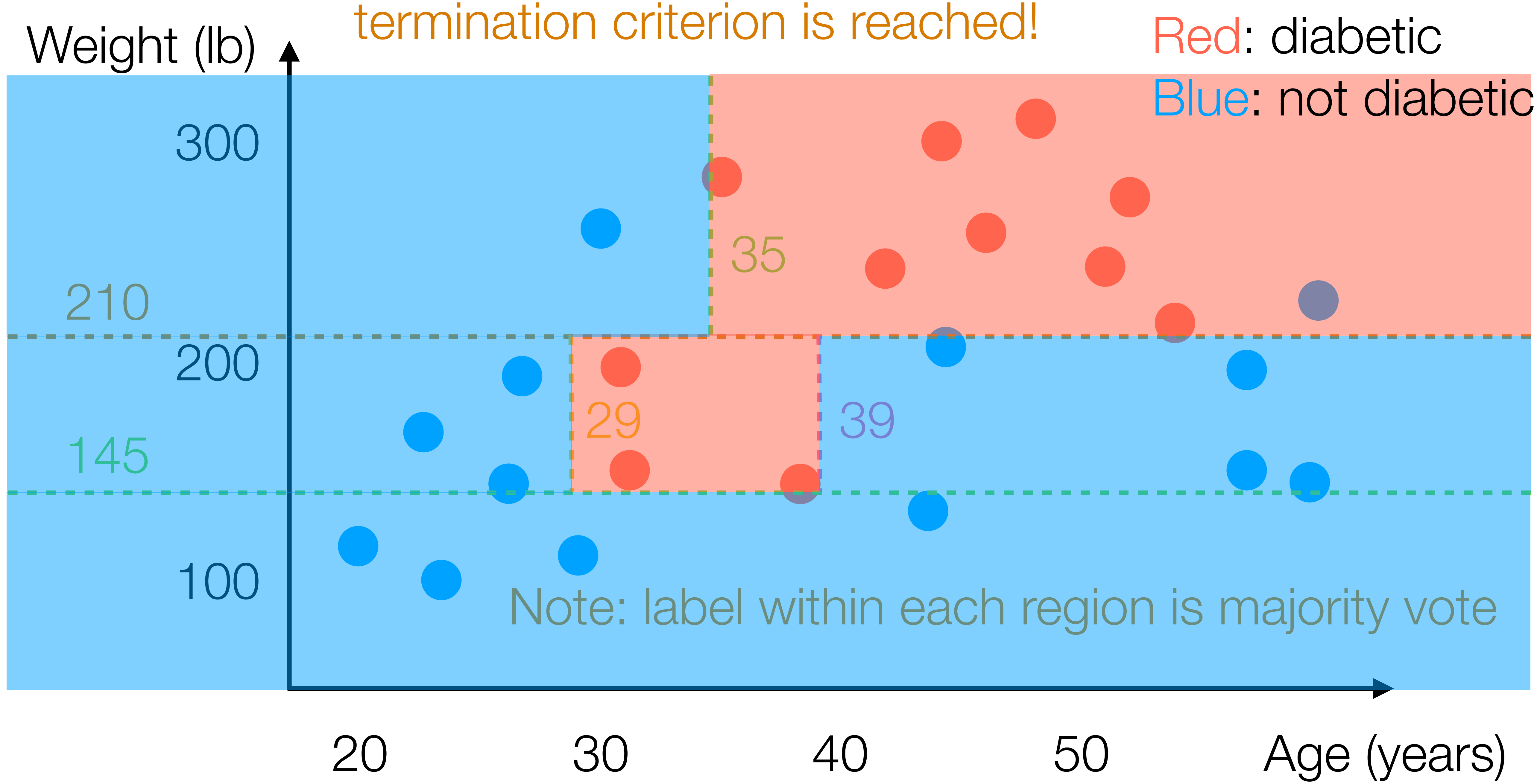
Learning a Decision Tree



2. Find threshold for which red and blue are as "separate as possible" (on one side, mostly red; on other side, mostly blue)

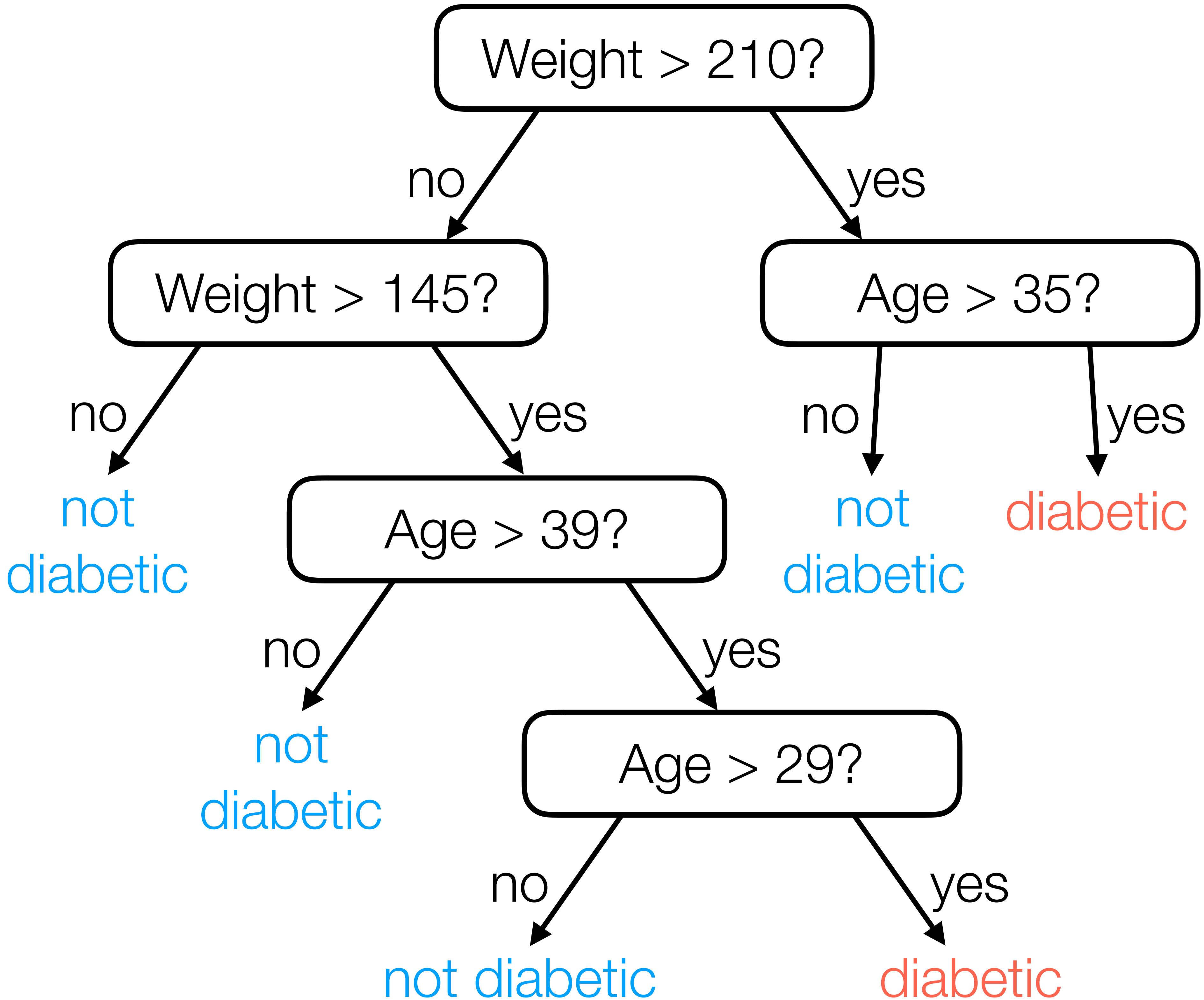
Learning a Decision Tree

Within each side, recurse until a termination criterion is reached!



Example termination criteria: ≥90% points within region has same label, number of points within region is <5

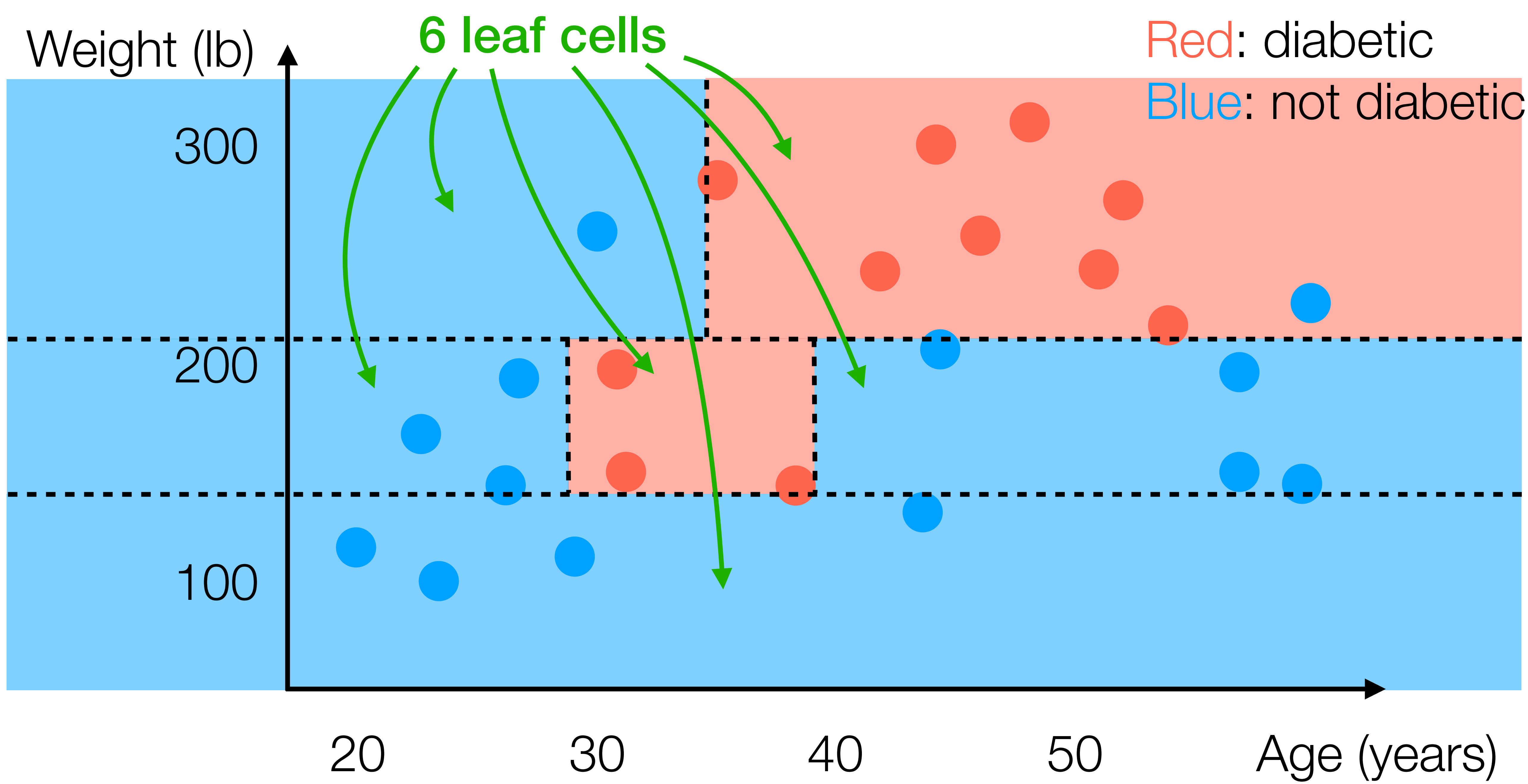
Decision Tree Learned



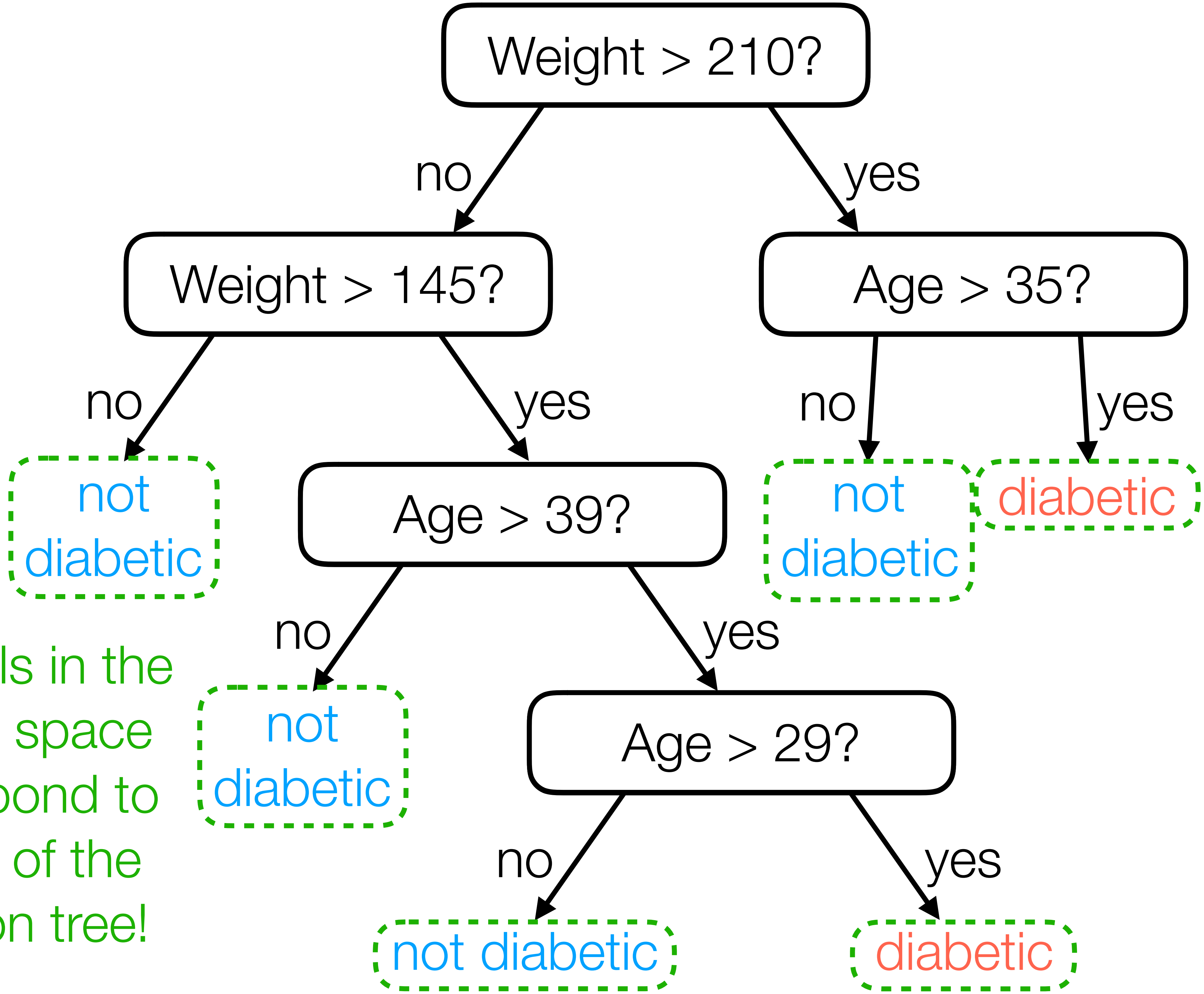
For a new person with feature vector (age, weight), easy to predict!

Nearest Neighbor Interpretation

Note: Each training data point lands in one “leaf cell”



Decision Tree Learned

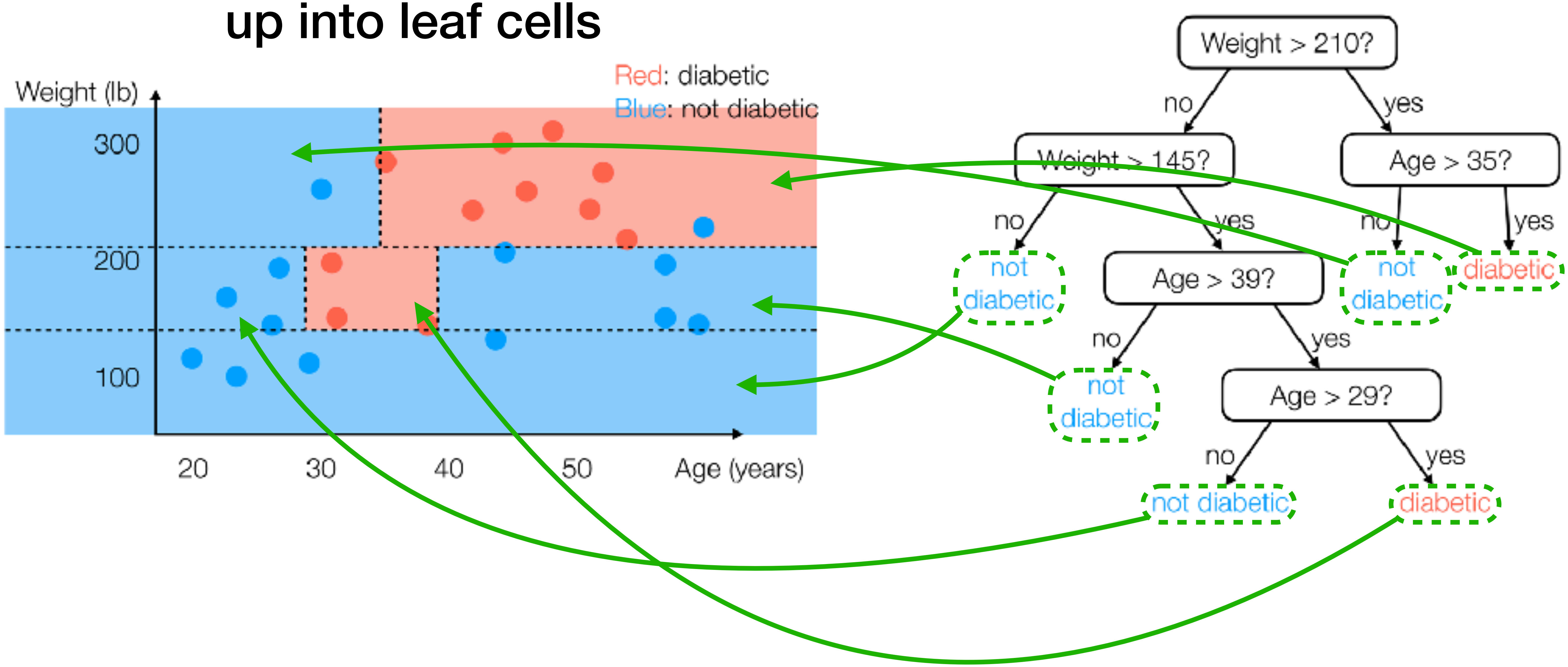


Leaf cells in the feature space correspond to leaves of the decision tree!

For a new person with feature vector (age, weight), easy to predict!

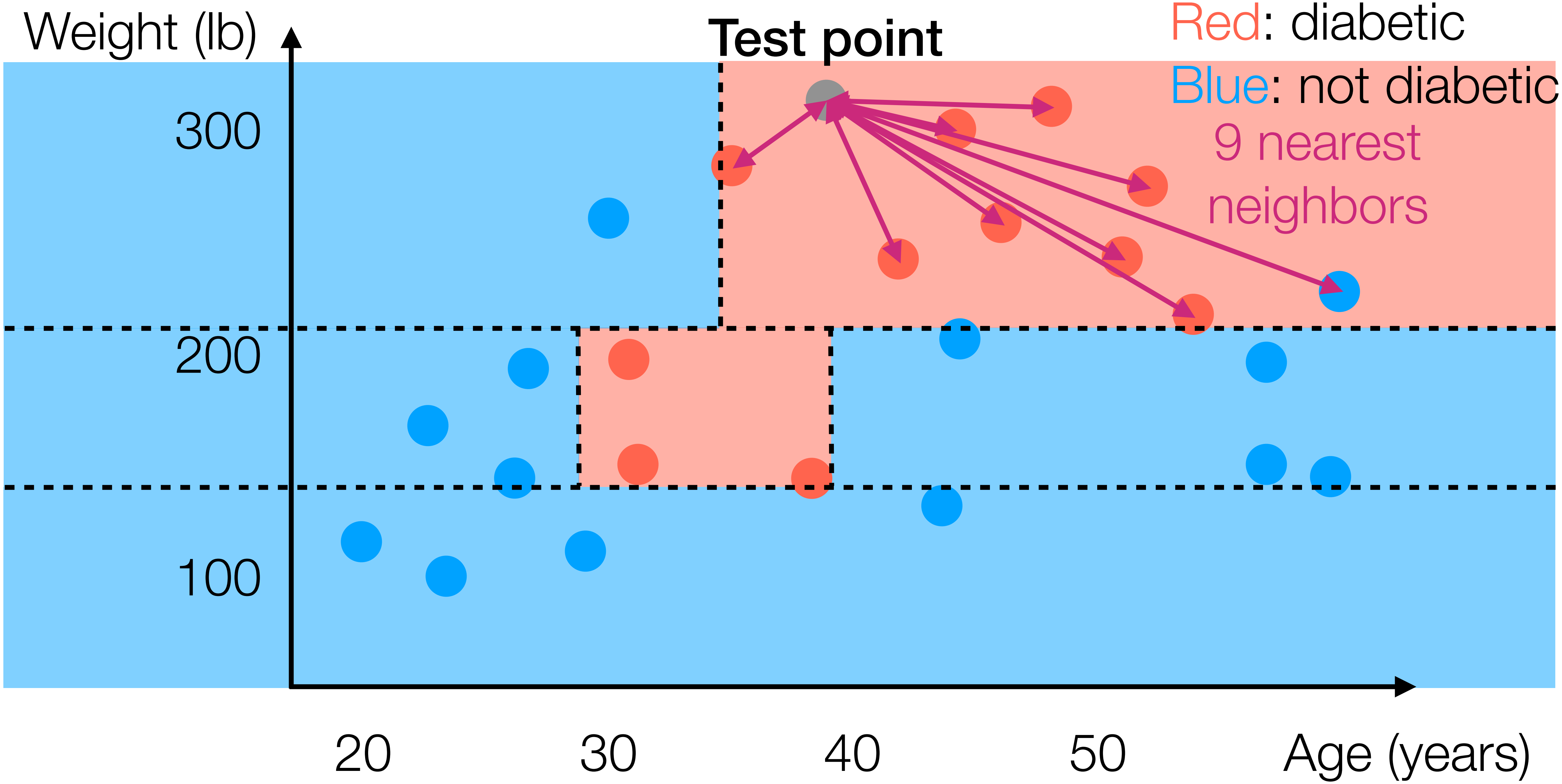
Feature space sliced up into leaf cells

Decision Tree



Nearest Neighbor Interpretation

Note: Each training data point lands in one "leaf cell"
Also: Any test data point lands in one leaf cell

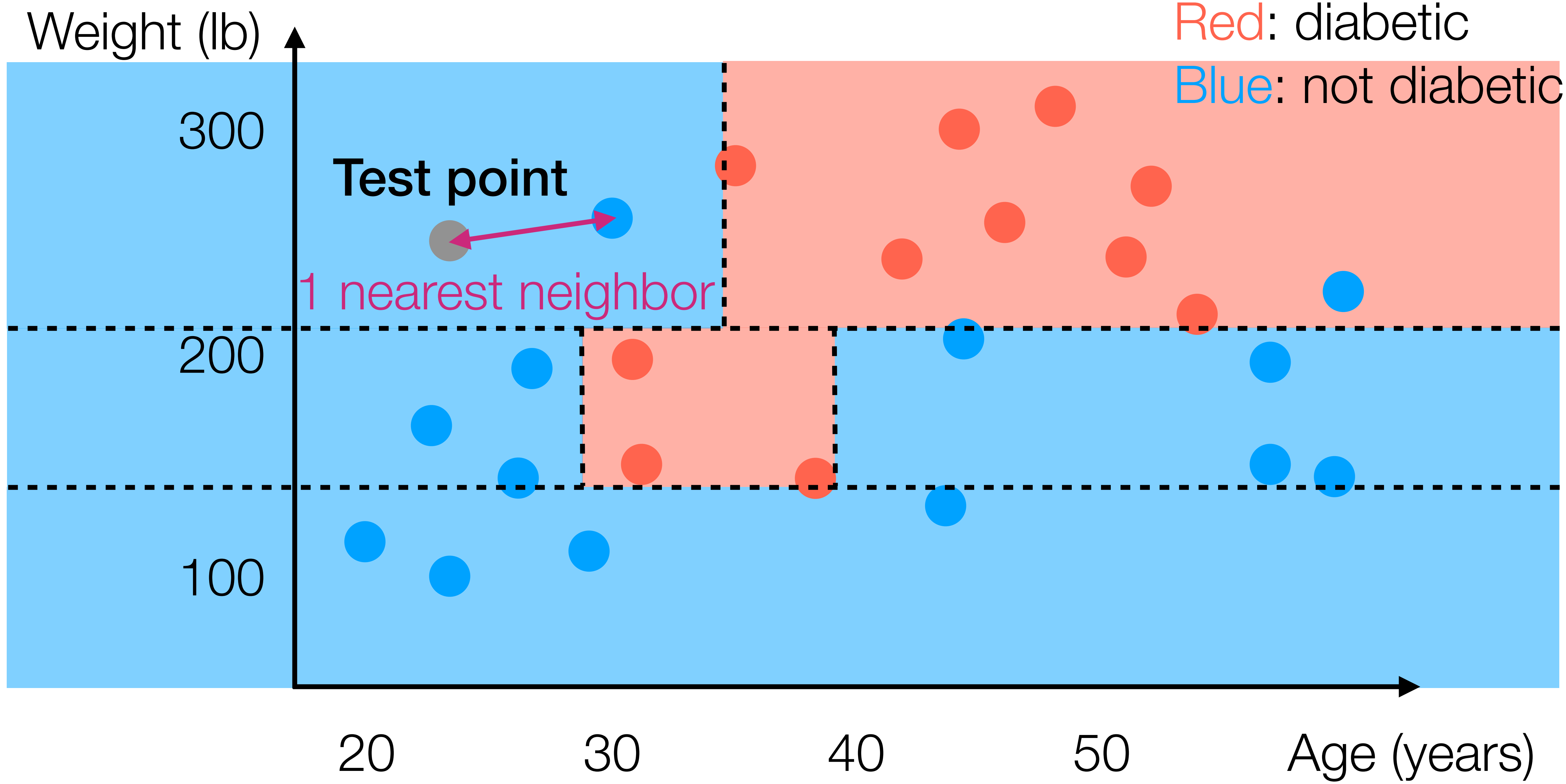


Prediction for test point: majority vote of training points in same leaf cell
(these training points act as nearest neighbors to the test point!)

Nearest Neighbor Interpretation

Note: Each training data point lands in one "leaf cell"

Also: Any test data point lands in one leaf cell



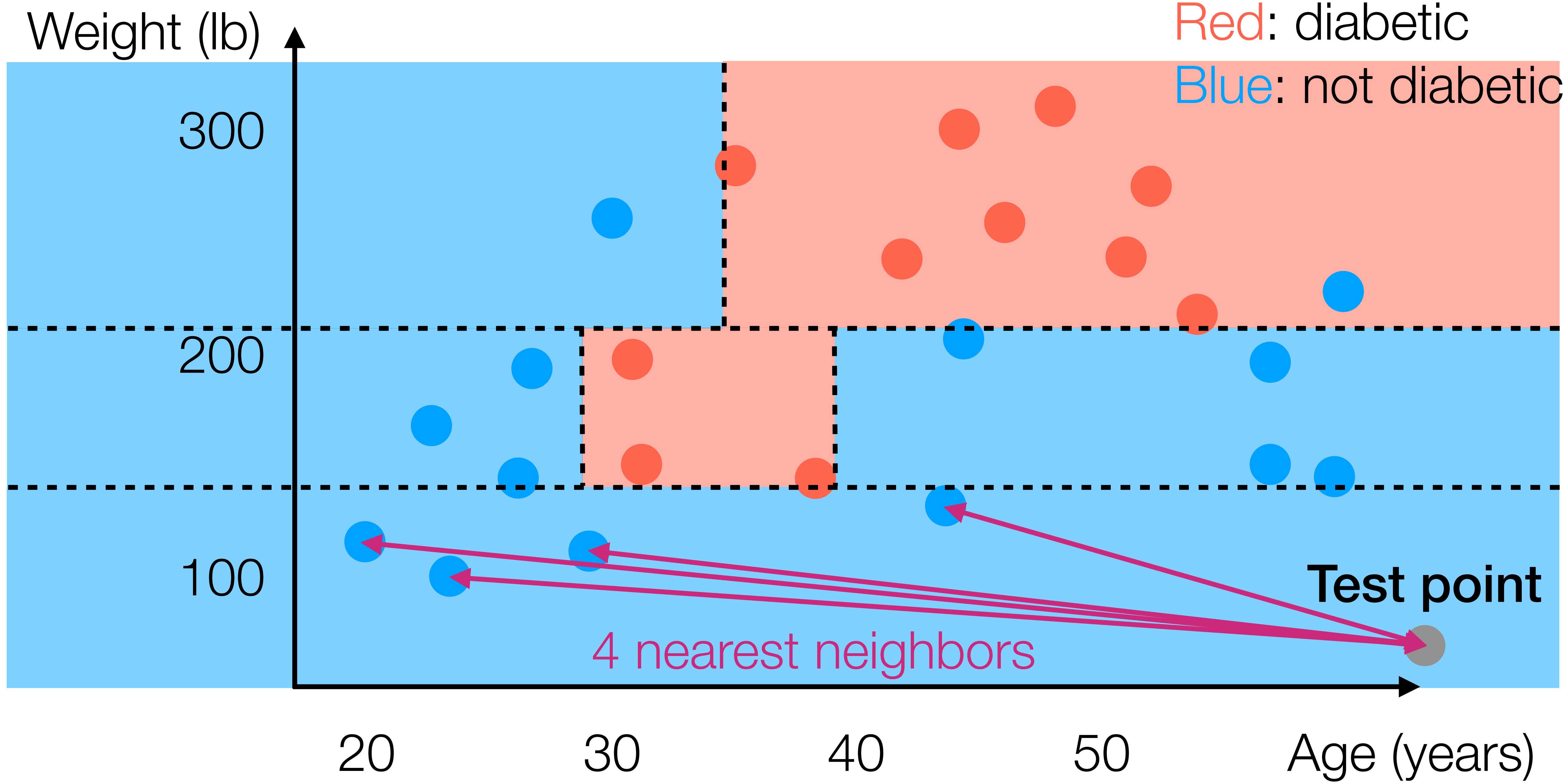
Prediction for test point: majority vote of training points in same leaf cell

(these training points act as nearest neighbors to the test point!)

Nearest Neighbor Interpretation

Note: Each training data point lands in one "leaf cell"

Also: Any test data point lands in one leaf cell



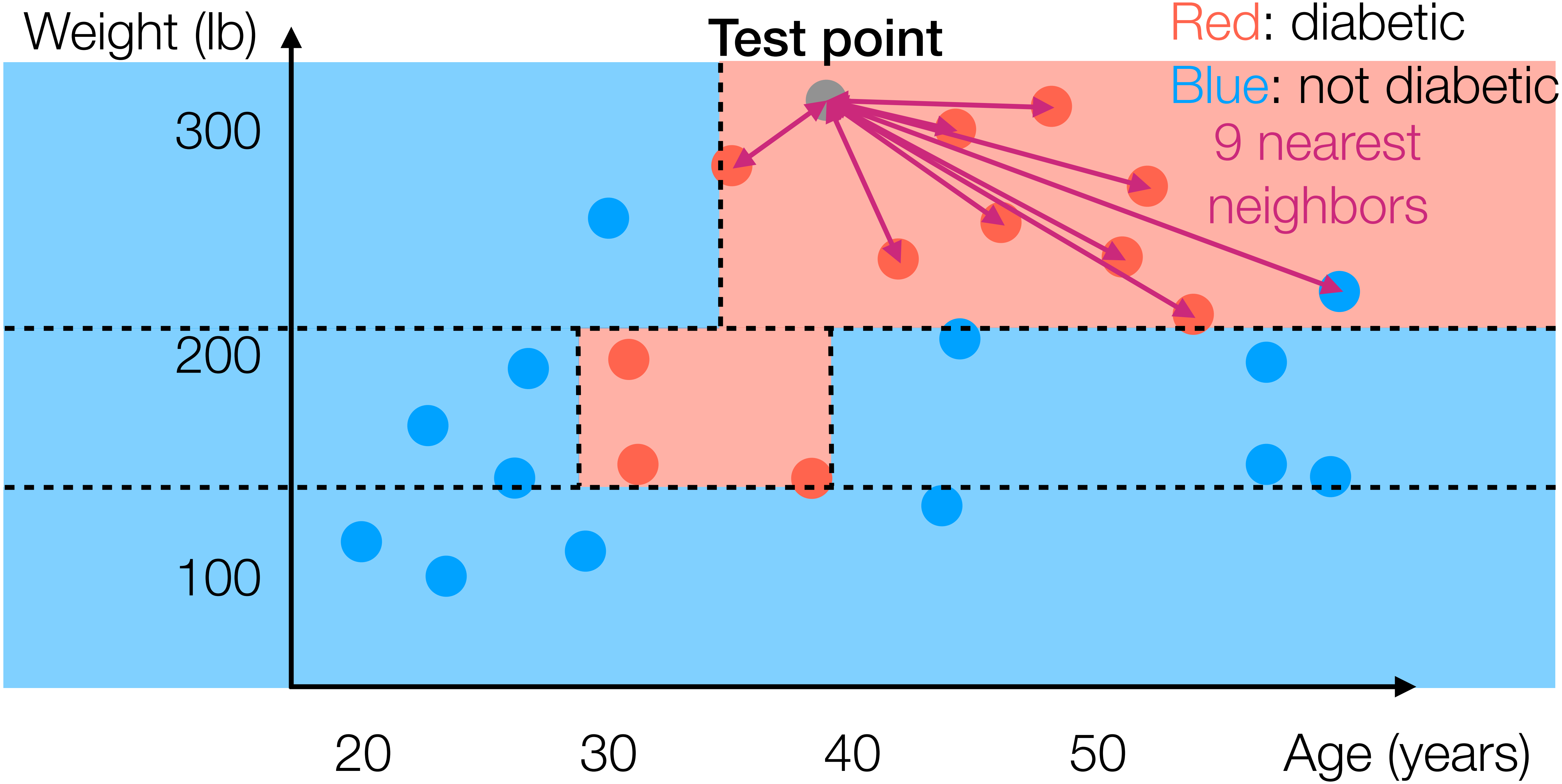
Prediction for test point: majority vote of training points in same leaf cell

(these training points act as nearest neighbors to the test point!)

Nearest Neighbor Interpretation

Note: Each training data point lands in one "leaf cell"

Also: Any test data point lands in one leaf cell



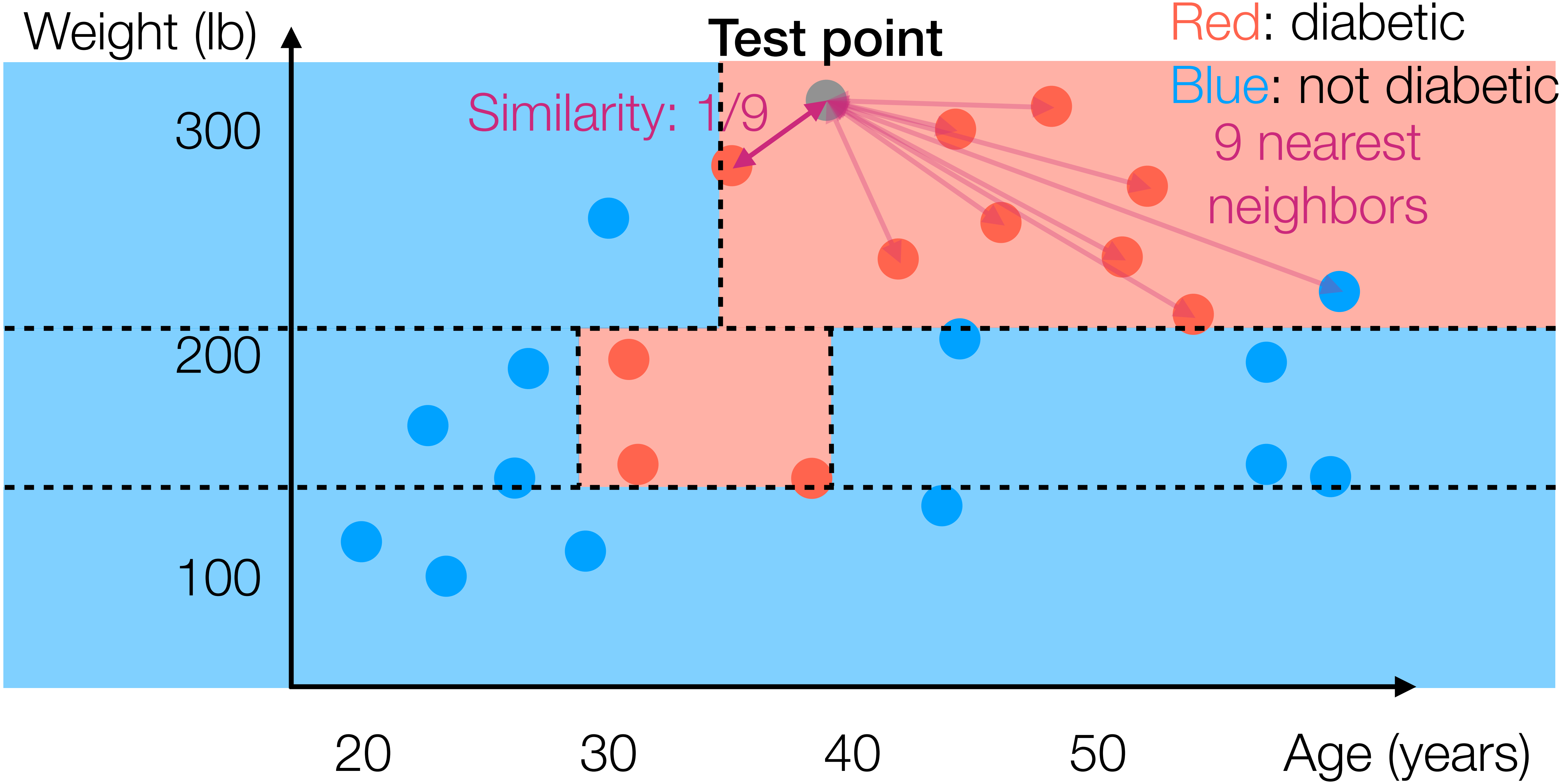
Prediction for test point: majority vote of training points in same leaf cell

(these training points act as nearest neighbors to the test point!)

Nearest Neighbor Interpretation

Note: Each training data point lands in one "leaf cell"

Also: Any test data point lands in one leaf cell

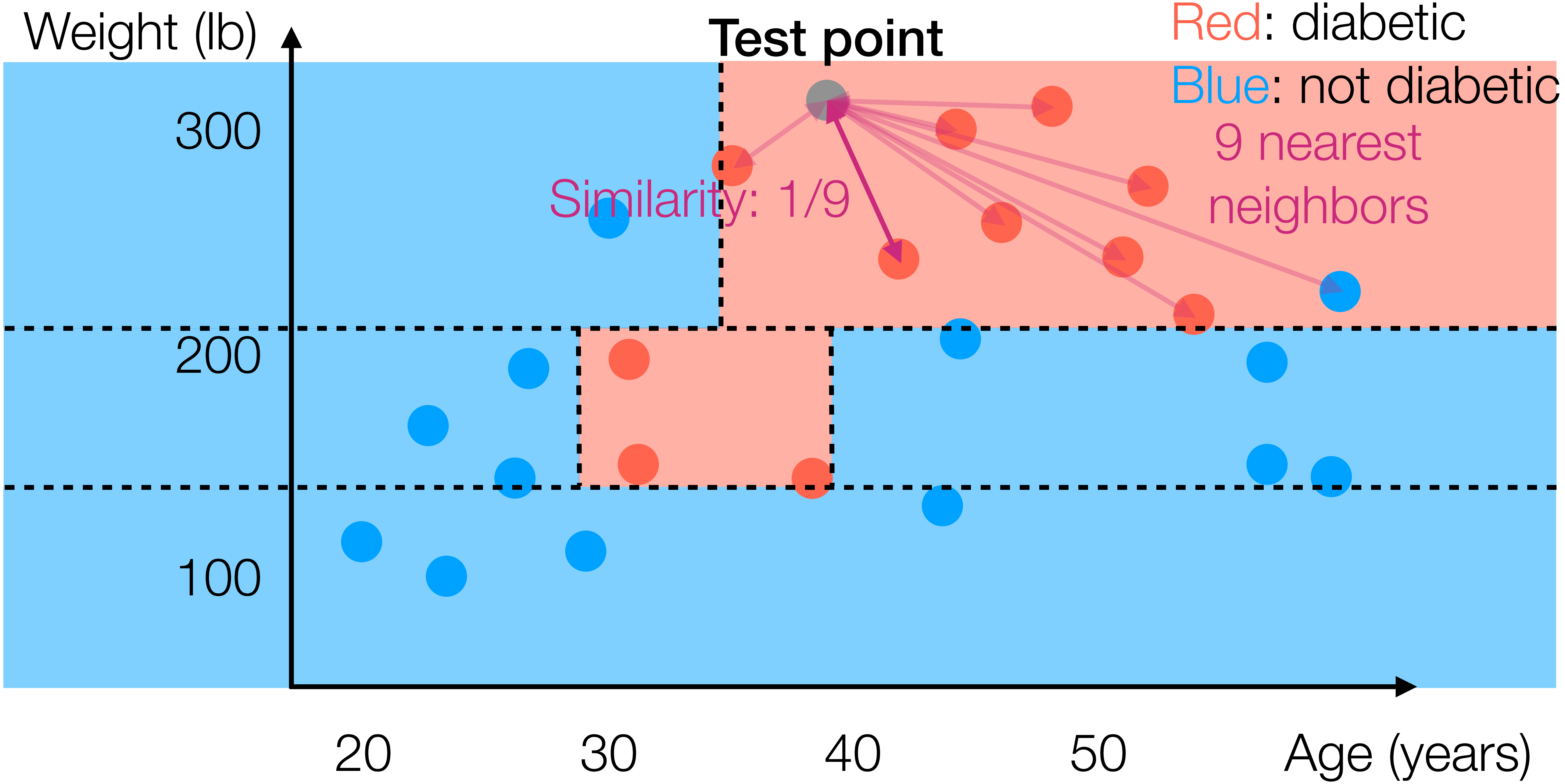


Prediction for test point: majority vote of training points in same leaf cell
(these training points act as nearest neighbors to the test point!)

Nearest Neighbor Interpretation

Note: Each training data point lands in one “leaf cell”

Also: Any test data point lands in one leaf cell

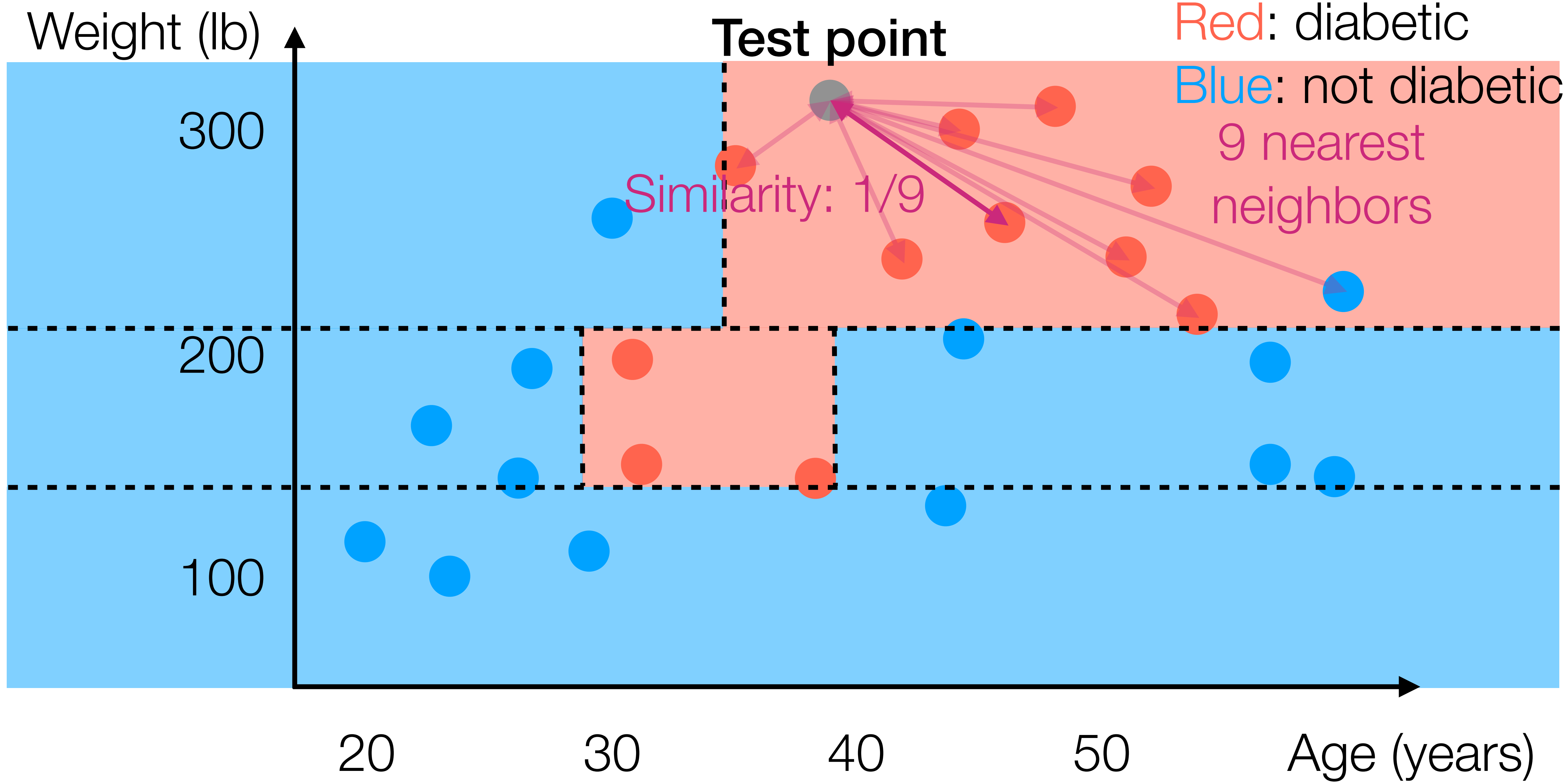


Prediction for test point: majority vote of training points in same leaf cell

(these training points act as nearest neighbors to the test point!)

Nearest Neighbor Interpretation

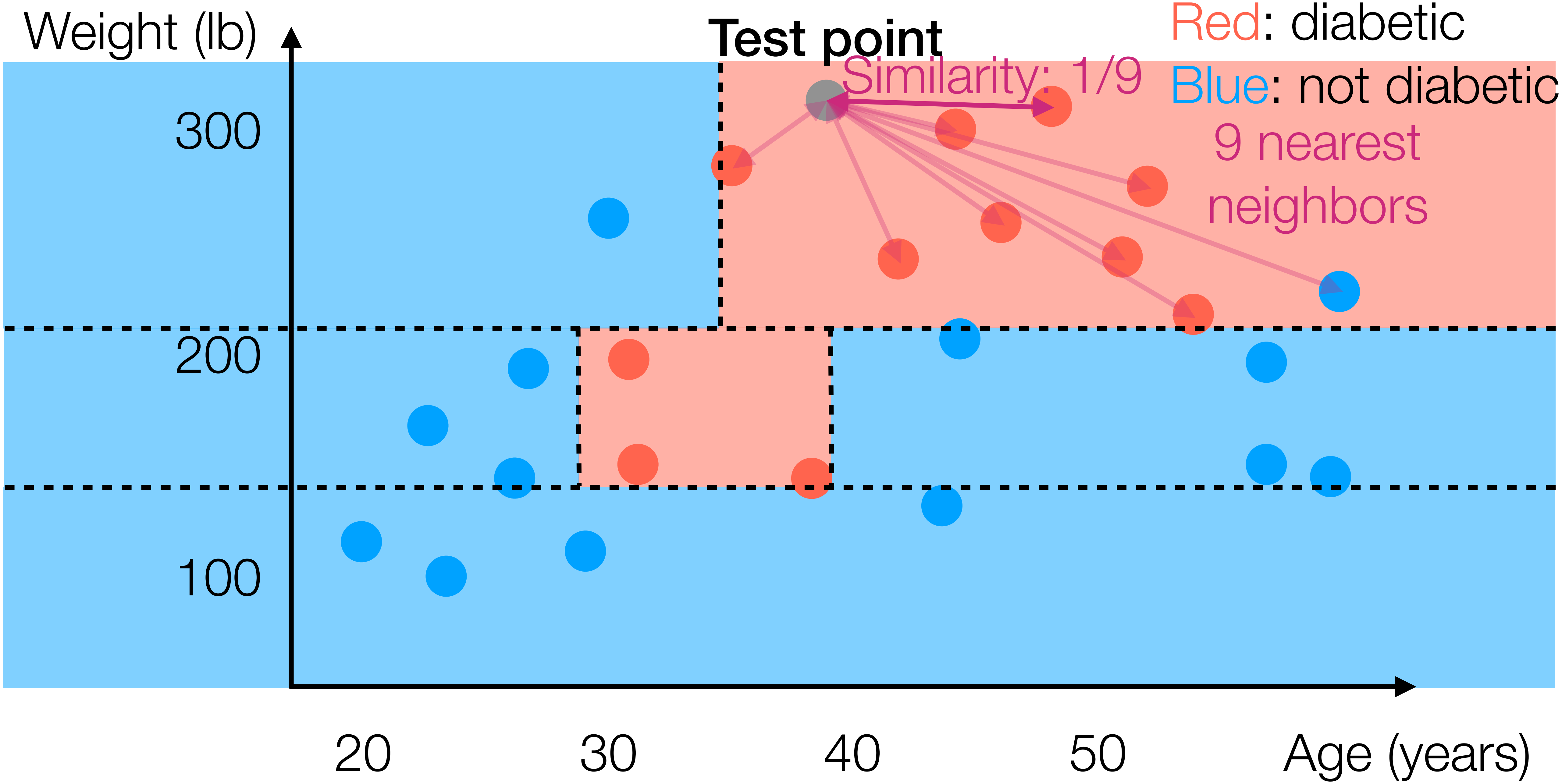
Note: Each training data point lands in one "leaf cell"
Also: Any test data point lands in one leaf cell



Prediction for test point: majority vote of training points in same leaf cell
(these training points act as nearest neighbors to the test point!)

Nearest Neighbor Interpretation

Note: Each training data point lands in one “leaf cell”
Also: Any test data point lands in one leaf cell

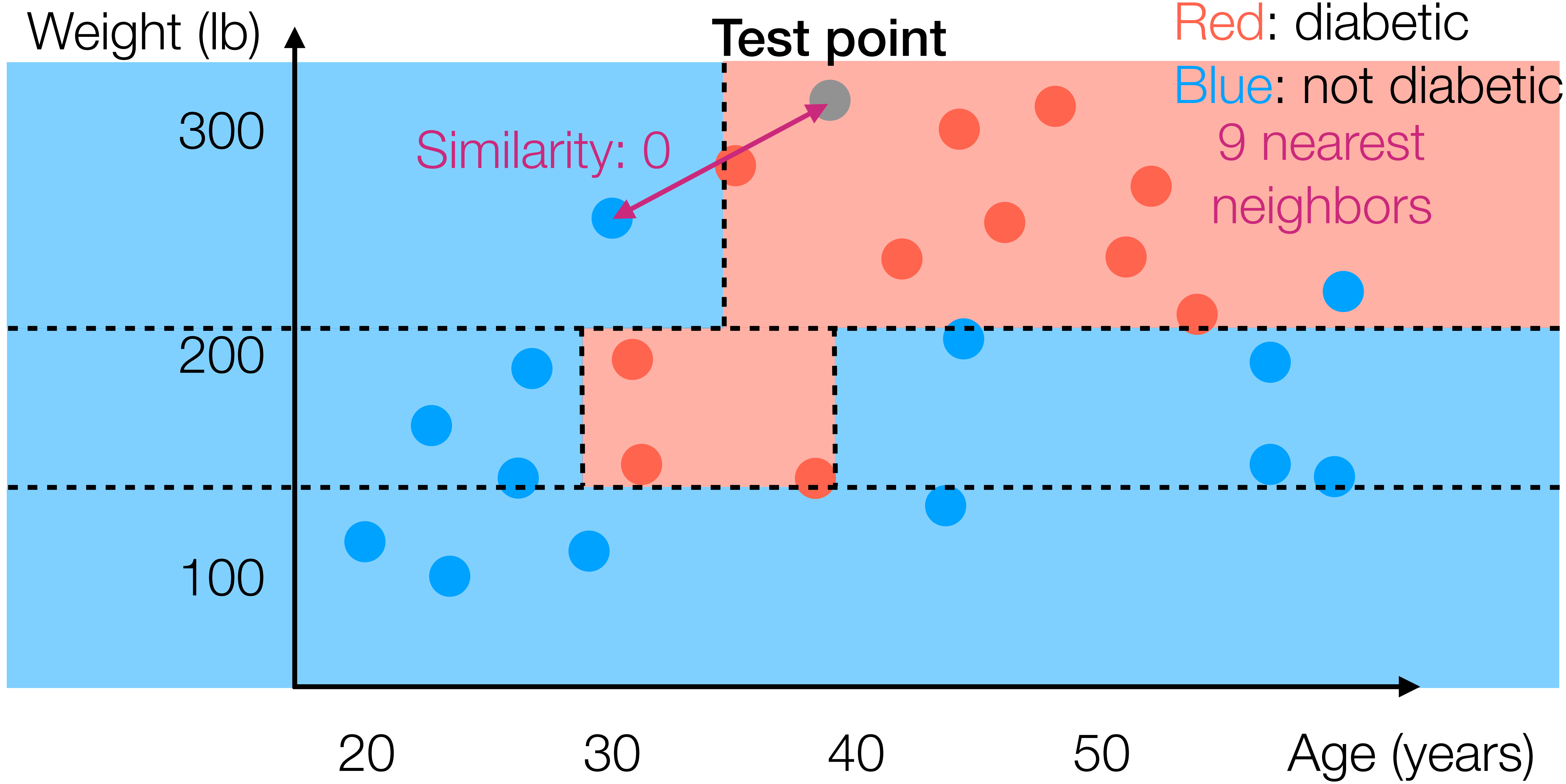


Prediction for test point: majority vote of training points in same leaf cell
(these training points act as nearest neighbors to the test point!)

Nearest Neighbor Interpretation

Note: Each training data point lands in one "leaf cell"

Also: Any test data point lands in one leaf cell



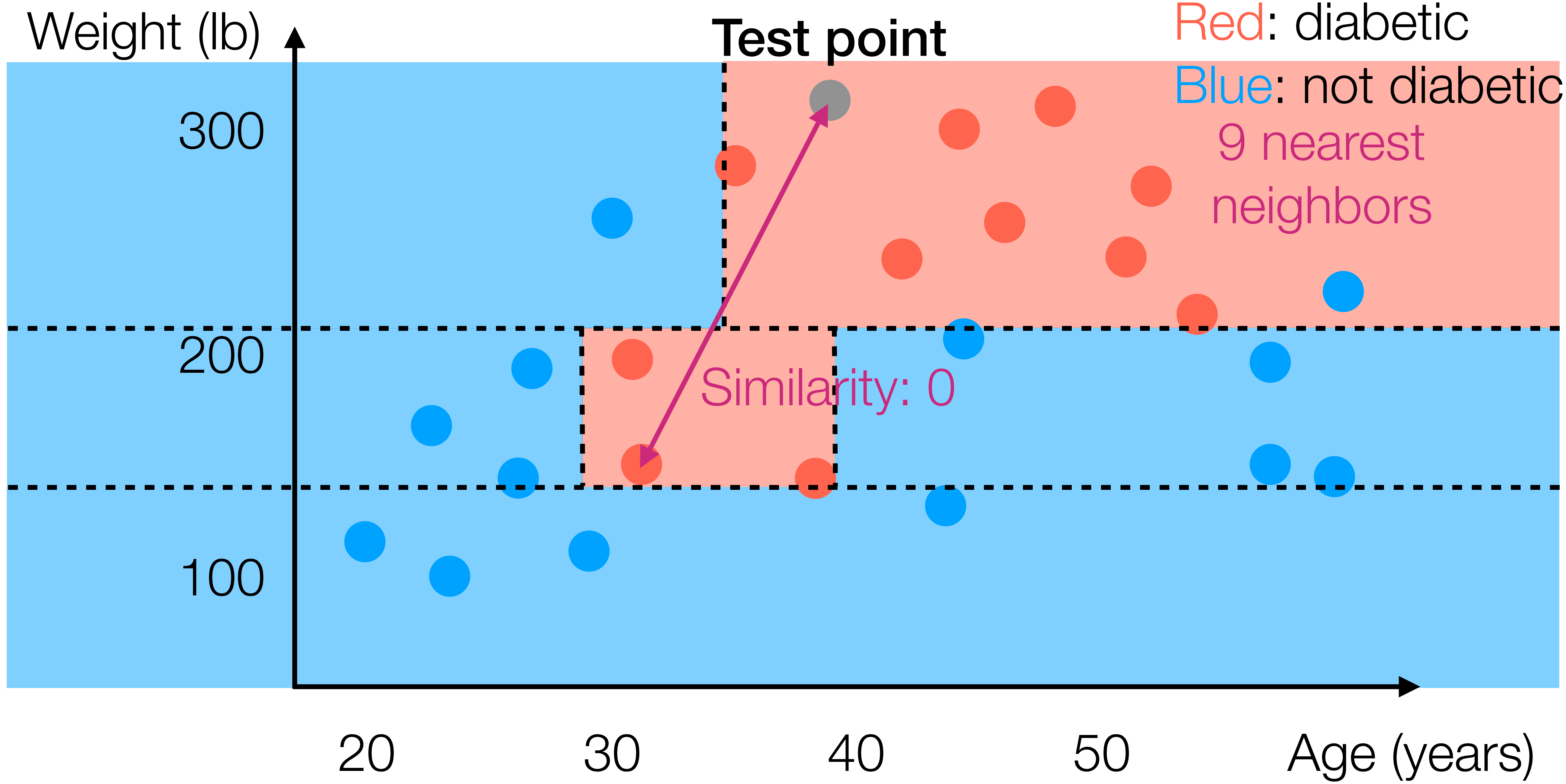
Prediction for test point: majority vote of training points in same leaf cell

(these training points act as nearest neighbors to the test point!)

Nearest Neighbor Interpretation

Note: Each training data point lands in one "leaf cell"

Also: Any test data point lands in one leaf cell

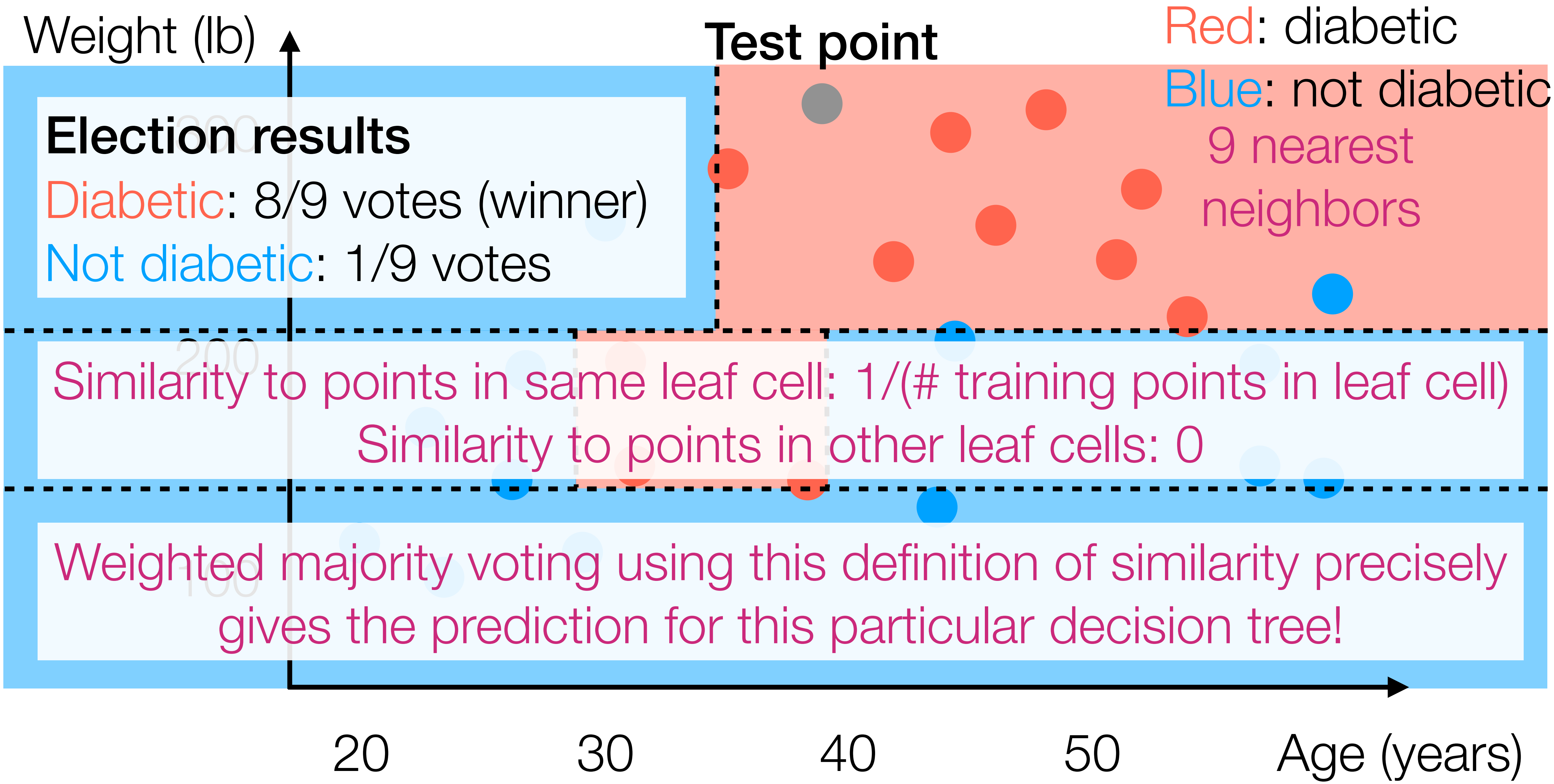


Prediction for test point: majority vote of training points in same leaf cell
(these training points act as nearest neighbors to the test point!)

Nearest Neighbor Interpretation

Note: Each training data point lands in one “leaf cell”

Also: Any test data point lands in one leaf cell



Prediction for test point: majority vote of training points in same leaf cell (these training points act as nearest neighbors to the test point!)

Decision Tree for Classification

- Many ways to learn (some popular ways: CART, C4.5)
- Extremely easy to interpret and to do prediction
- Nearest neighbor interpretation:
  - For each test point, look at leaf cell it falls into to find its nearest neighbors among the training data
    (note: # of nearest neighbors varies!)
  - Prediction for test point: majority vote of nearest neighbors' labels
- Learning a decision tree learns a similarity function (that depends on labels)

Decision Tree for ~~Classification~~ Regression

- Many ways to learn (some popular ways: CART, C4.5)

- Extremely easy to interpret and to do prediction

- Nearest neighbor interpretation:

  - For each test point, look at leaf cell it falls into to find its nearest neighbors among the training data (note: # of nearest neighbors varies!)

  - Prediction for test point: ~~majority vote~~ average of nearest neighbors' labels

- Learning a decision tree learns a similarity function (that depends on labels)

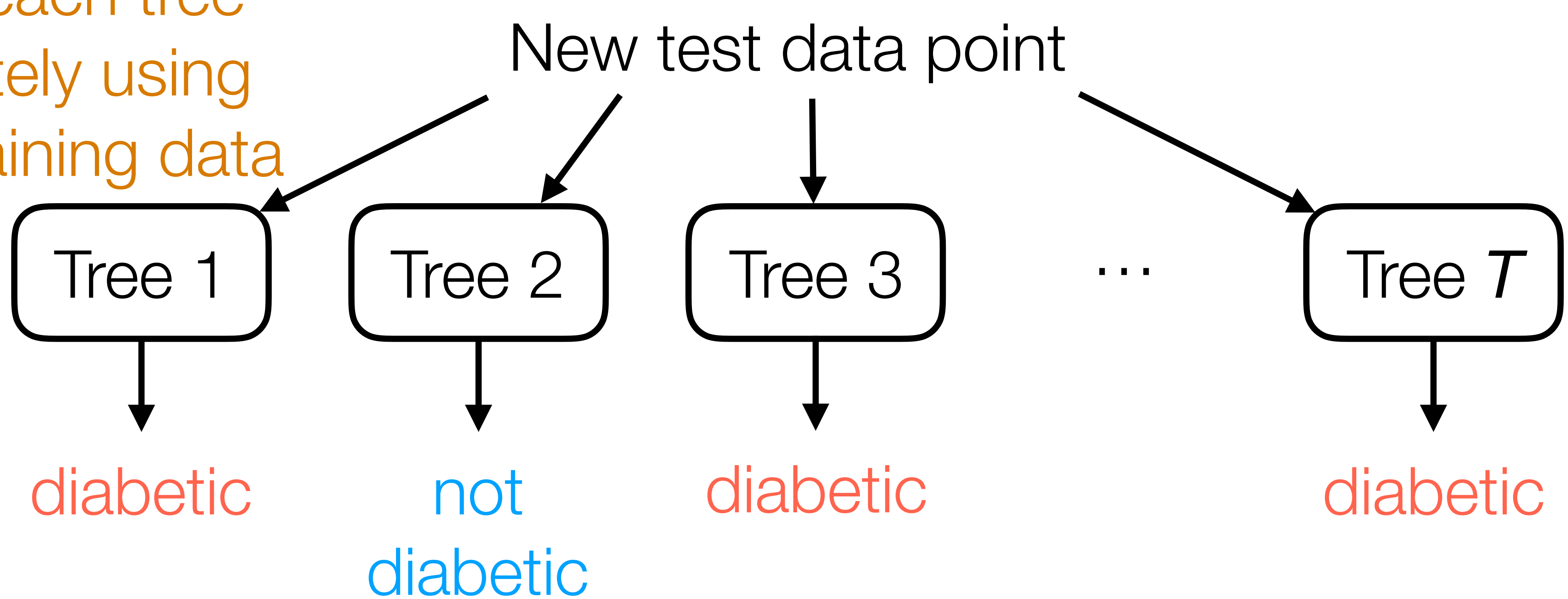
Decision Forest for Classification

- Typically, a decision tree is learned with randomness (e.g., we randomly chose which feature to threshold)

  ➔ by re-running the same learning procedure, we can get different decision trees that make different predictions!

- For a more stable prediction, use many decision trees

Learn each tree separately using same training data



Final prediction: majority vote of the different trees' predictions

This is not the only way to aggregate predictions!

Decision Forest for Classification

Learn each tree separately using same training data

New test data point

	Tree 1	Tree 2	Tree 3	⋯	Tree T
diabetic	8/9 votes	1/4 votes	5/7 votes		2/3 votes
not diabetic	1/9 votes	3/4 votes	2/7 votes		1/3 votes

Final prediction: sum up votes across trees to find winner of election!

Nearest neighbor interpretation:

For a specific test data point x and training data point x_i

$$\text{similarity}(x, x_i) = \boxed{\frac{1}{T}} \sum_{t=1}^{T} \boxed{\text{similarity}_t(x, x_i)}$$

makes overall similarity between 0 and 1

similarity function for ***t***-th tree

Decision Forest for ~~Classification~~ Regression

Learn each tree separately using same training data

New test data point

Tree 1 | Tree 2 | Tree 3 | … | Tree T

average label for tree 1 | average label for tree 2 | average label for tree 3 | average label for tree T

Average these values to get final prediction

Nearest neighbor interpretation:

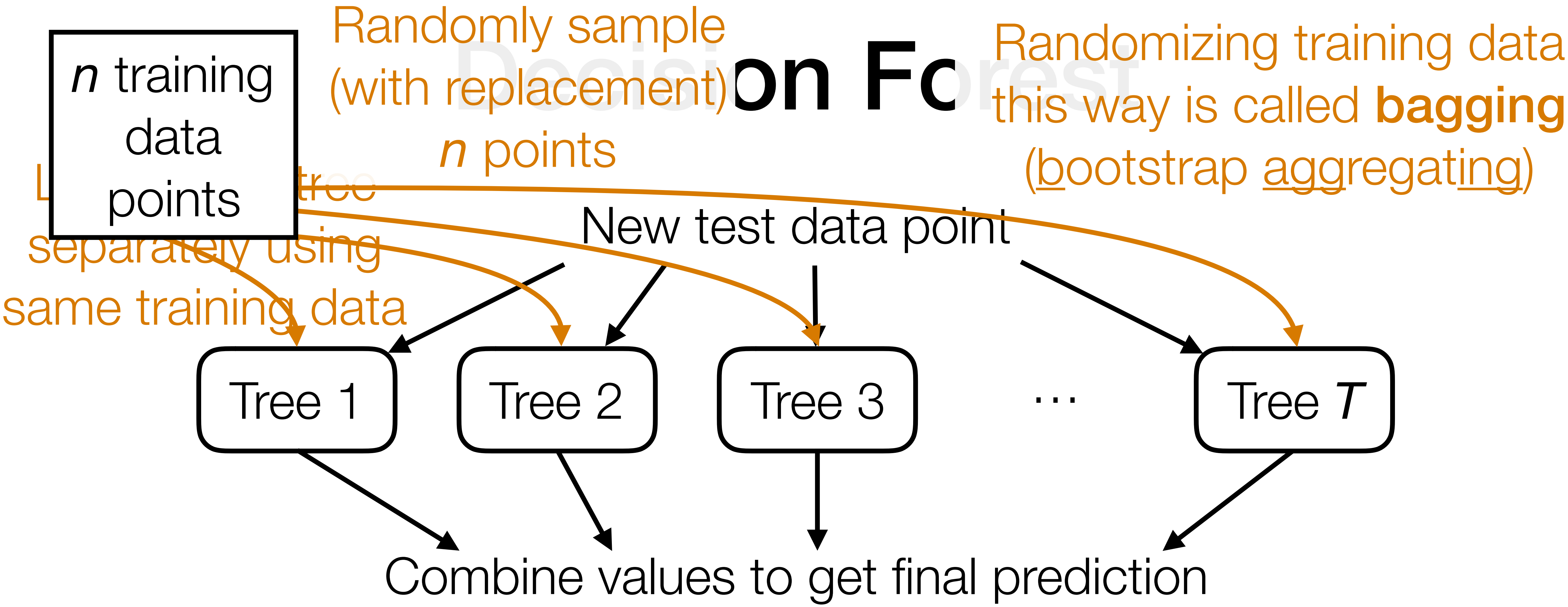
For a specific test data point x and training data point x_i

$$\text{similarity}(x, x_i) = \boxed{\frac{1}{T}} \sum_{t=1}^{T} \boxed{\text{similarity}_t(x, x_i)}$$

makes overall similarity between 0 and 1

similarity function for t-th tree

Decision Forest



Question: What happens if all the trees are the same?

Adding randomness can make trees more different!

- **Random Forest:** in addition to randomly choosing features to threshold, also randomize training data used for each tree
- **Extremely randomized trees:** further randomize thresholds rather than trying to pick clever thresholds

Boosting

I'll only sketch the general idea

Random decision forests learned each tree separately

Boosting: learn trees *sequentially*, and learn from previous trees' mistakes

If some trees are bad, we still weight them equally

Boosting: weight trees unequally so bad trees are down-weighted

Boosting

Training data

$w_1^{(1)} = 1$ $w_2^{(1)} = 1$ $w_3^{(1)} = 1$

Tree 1

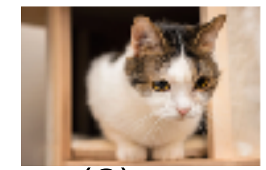
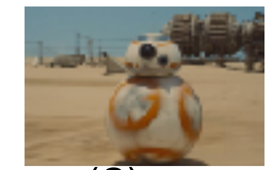
Predicted: cat, dog, shark

Actual: cat, cat, robot

Where did the errors appear?

Duplicate these training examples to emphasize them more when learning the next tree

Training data

$w_1^{(2)} = 1$ $w_2^{(2)} = 2$ $w_3^{(2)} = 2$

Tree 2

Predicted: cat, cat, donkey

Actual: cat, cat, robot

Where did the errors appear?

Duplicate these training examples to emphasize them more when learning the next tree

Boosting

Learn trees *sequentially* accounting for mistakes made previously

Training data: $x_1, x_2, \ldots, x_n$ $\quad$ $x_1, x_2, \ldots, x_n$ $\quad$ $x_1, x_2, \ldots, x_n$

Weights: $w_1^{(1)}, w_2^{(1)}, \ldots, w_n^{(1)}$ $\quad$ $w_1^{(2)}, w_2^{(2)}, \ldots, w_n^{(2)}$ $\quad$ $w_1^{(T)}, w_2^{(T)}, \ldots, w_n^{(T)}$

Tree 1 $\quad$ Tree 2 $\quad \cdots \quad$ Tree T

Predicted: $\widehat{y}_1^{(1)}, \widehat{y}_2^{(1)}, \ldots, \widehat{y}_n^{(1)}$ $\quad$ $\widehat{y}_1^{(2)}, \widehat{y}_2^{(2)}, \ldots, \widehat{y}_n^{(2)}$ $\quad$ $\widehat{y}_1^{(T)}, \widehat{y}_2^{(T)}, \ldots, \widehat{y}_n^{(T)}$

Actual: $y_1, y_2, \ldots, y_n$ $\quad$ $y_1, y_2, \ldots, y_n$ $\quad$ $y_1, y_2, \ldots, y_n$

Adjust for how much each tree's votes count

$$\text{similarity}(x, x_i) = \sum_{t=1}^{T} \alpha_t \text{similarity}_t(x, x_i)$$

weight for tree ***t***

Different ways to choose weights yield different boosting methods (e.g., AdaBoost, gradient tree boosting)

Still an adaptive NN method!