Classification

March 19, 2019

Data Science CSCI 1951A

Brown University

Instructor: Ellie Pavlick

HTAs: Wennie Zhang, Maulik Dang, Gurnaaz Kaur

Announcements

- 😟 MR regrades
- 😀 ML HW due Friday (Spring Break = 1 late day)

Today

- Generative vs. Discriminative Models
- KNN, Naive Bayes, Logistic Regression
- SciKit Learn Demo

Classification

One Goal: $P(Y|X)$

Classification

One Goal: P(Y|X)

Label

Classification

Features

One Goal: $P(Y|X)$

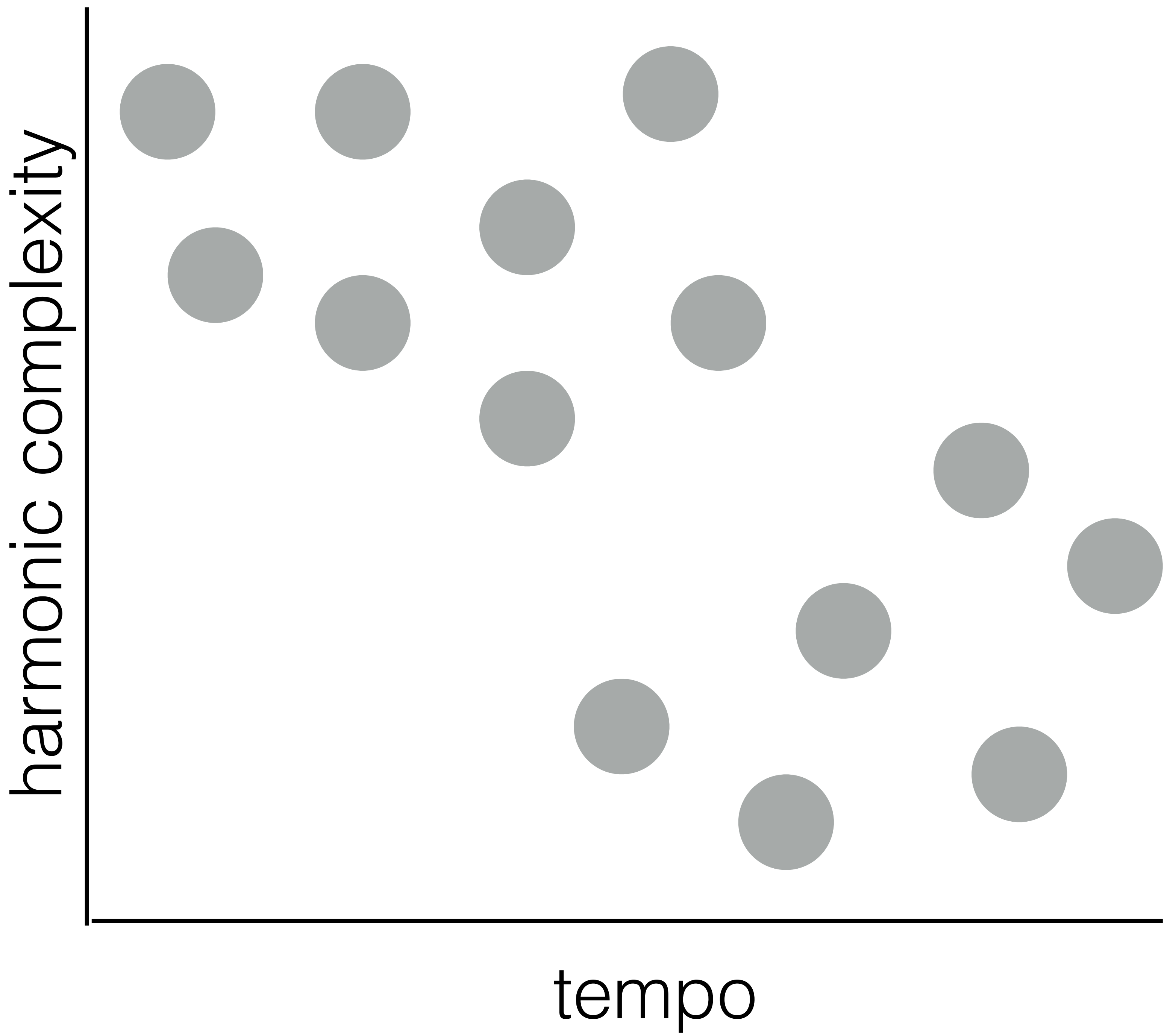
Label

Classification

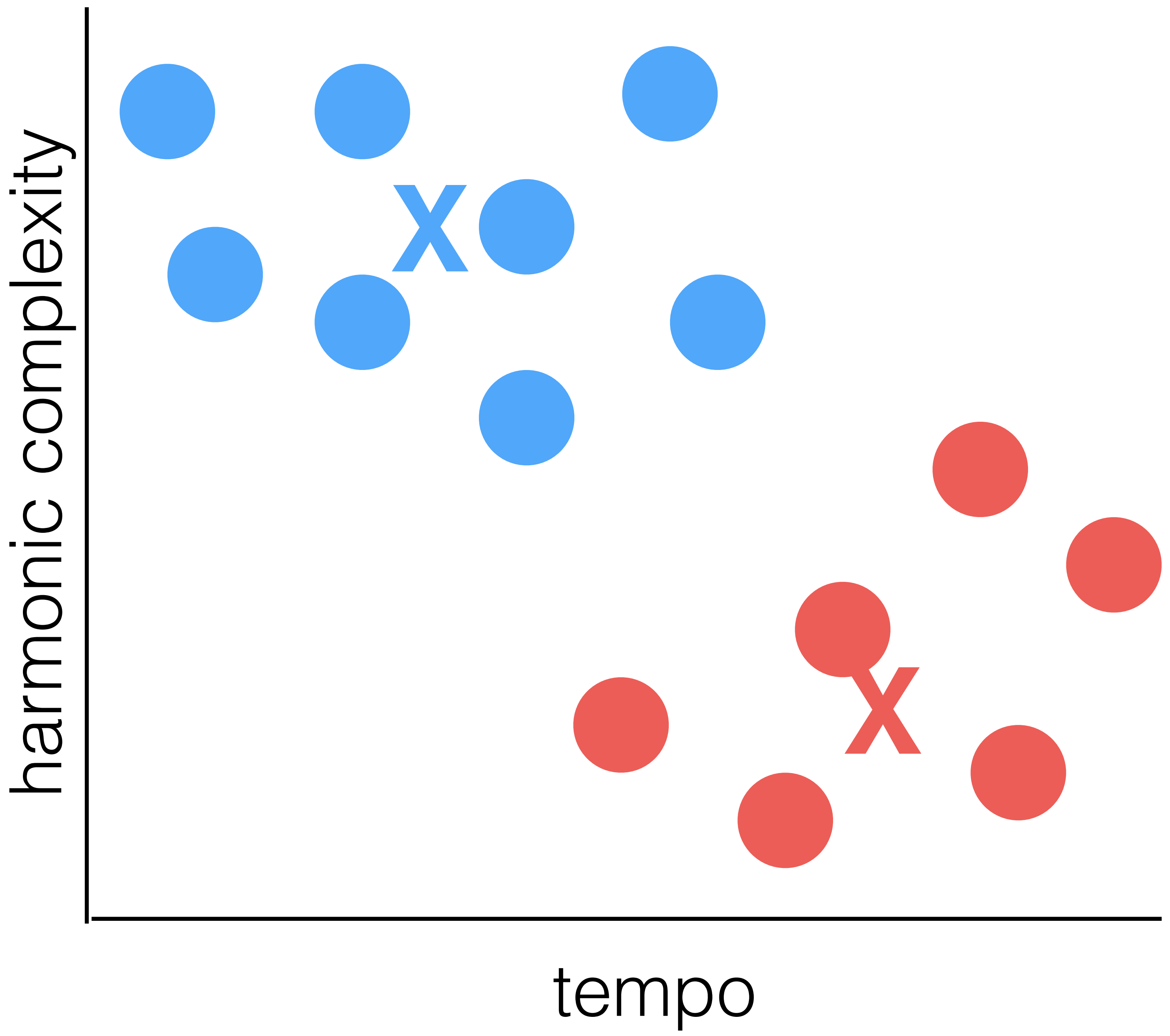
One Goal: $P(Y|X)$

P(email is spam | words in the message)
P(genre of song|tempo, harmony, lyrics…)
P(article clicked | title, font, photo…)

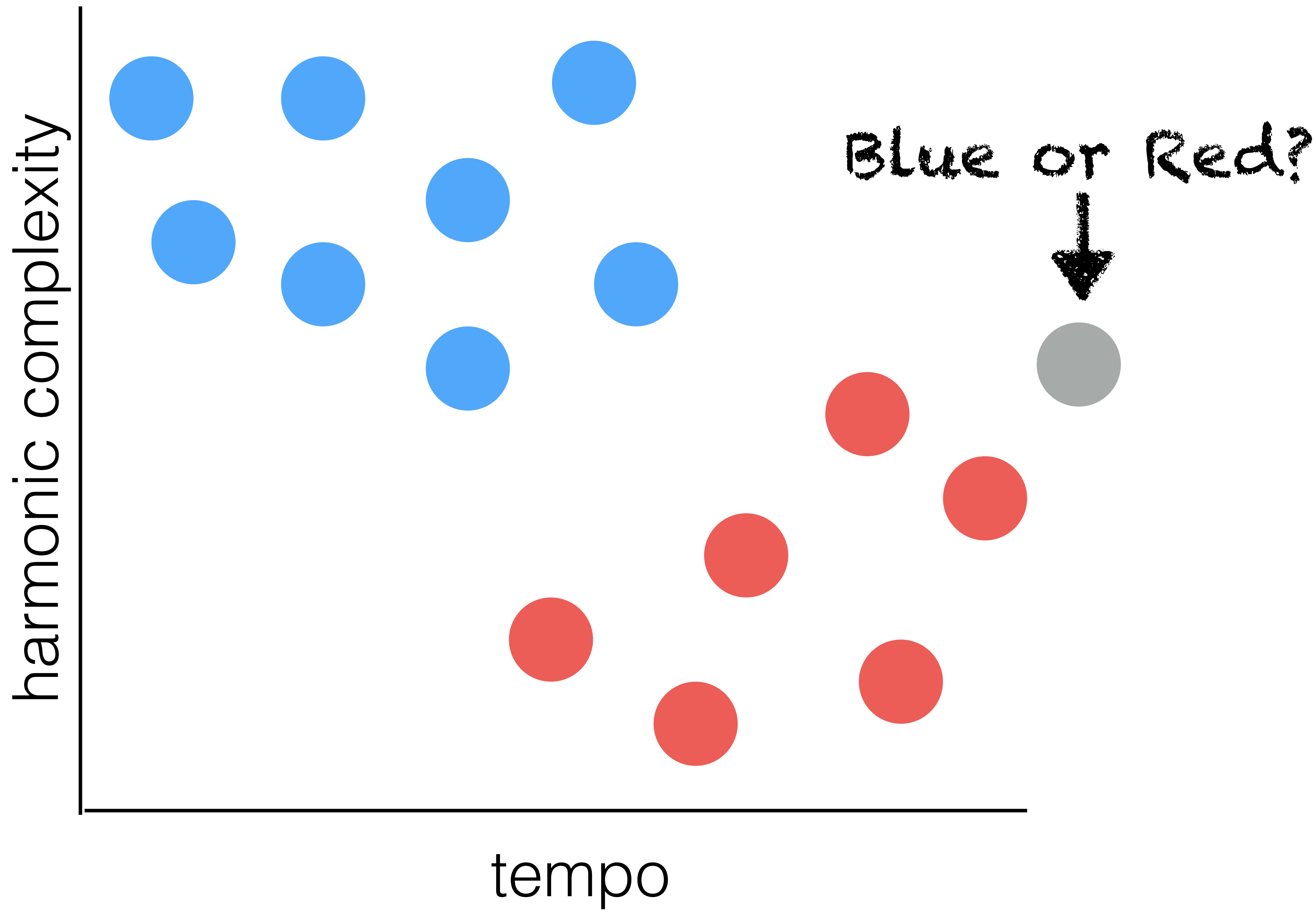
K Means



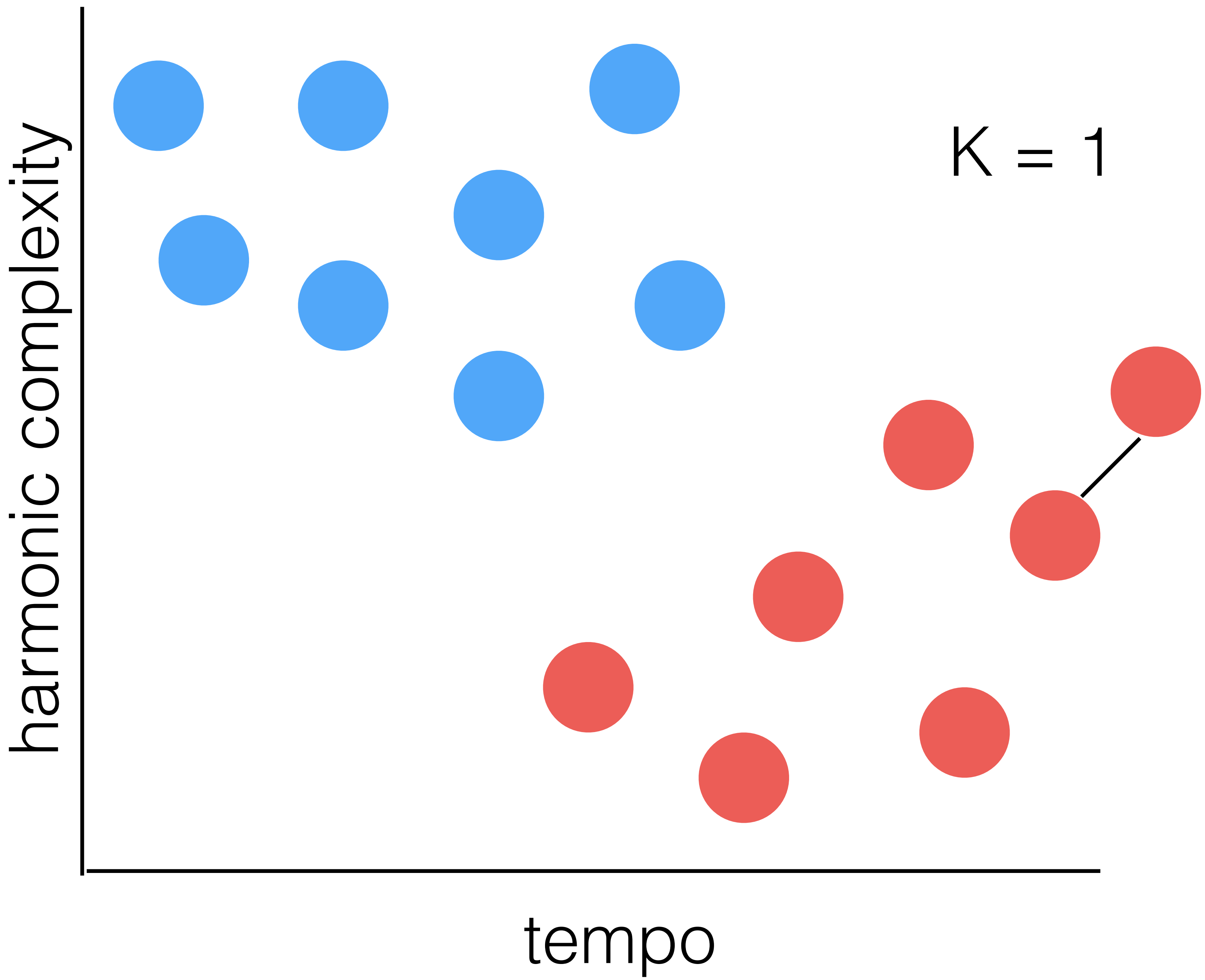
K Means



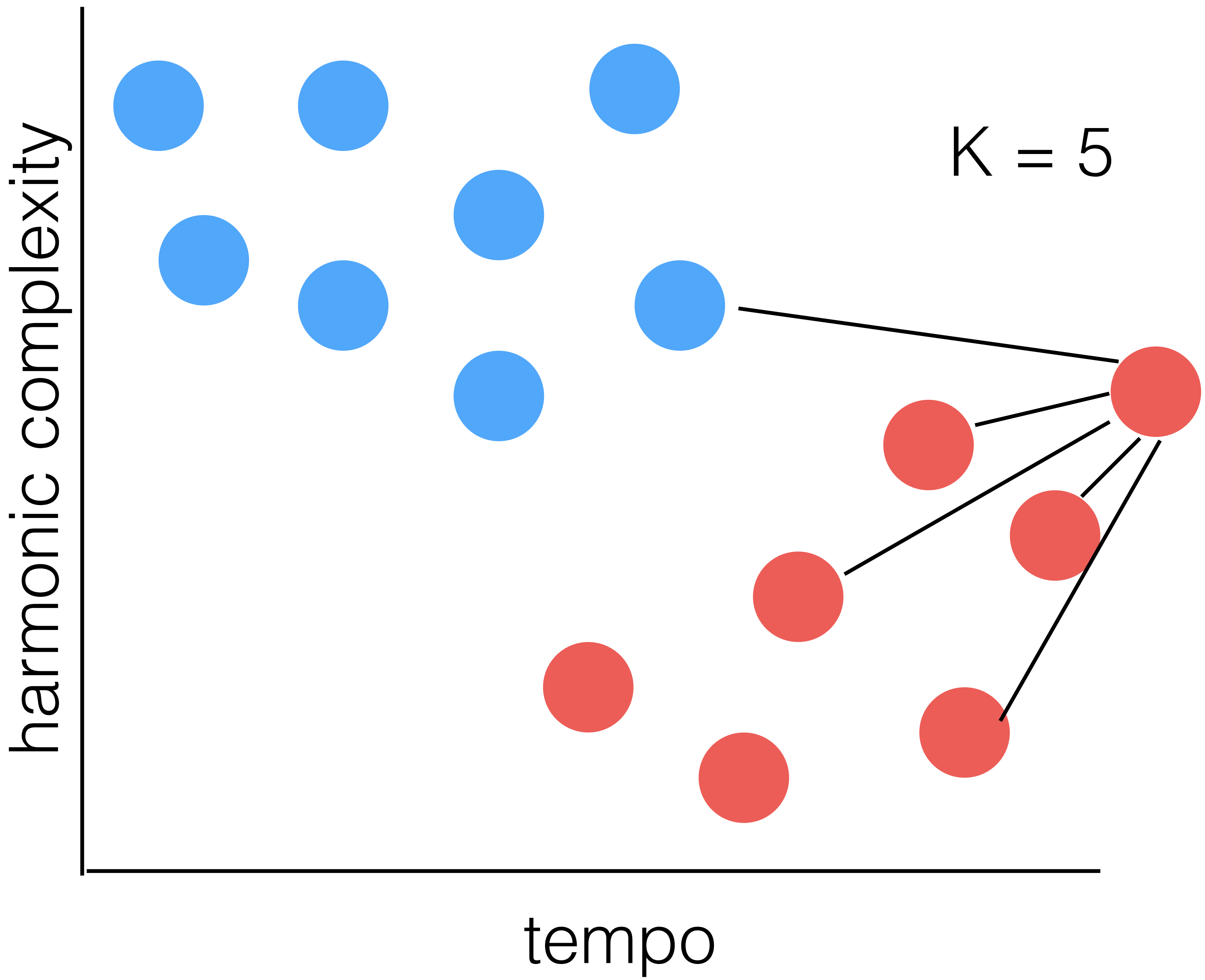
K Nearest Neighbors



K Nearest Neighbors



K Nearest Neighbors



K Nearest Neighbors

K Nearest Neighbors

- Arguably the simplest ML algorithm

K Nearest Neighbors

- Arguably the simplest ML algorithm
- “Non-Parametric” — no assumptions about the form of the classification model

K Nearest Neighbors

- Arguably the simplest ML algorithm
- “Non-Parametric” — no assumptions about the form of the classification model
- All the work is done at classification time

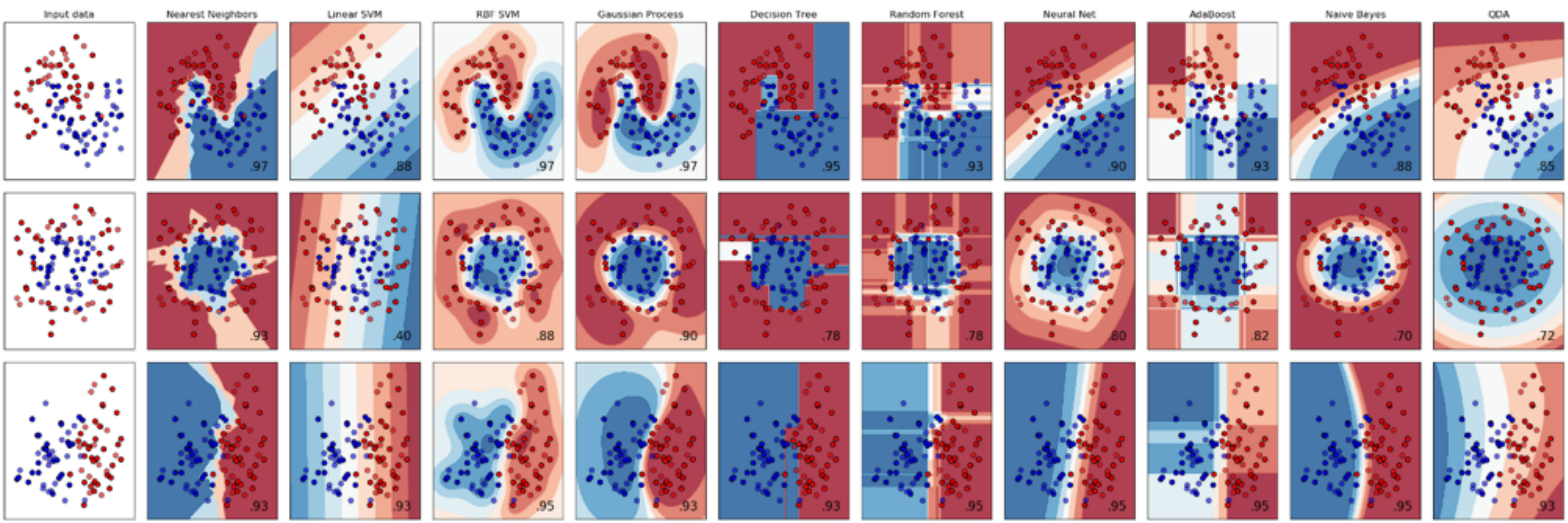
K Nearest Neighbors

- Arguably the simplest ML algorithm
- “Non-Parametric” — no assumptions about the form of the classification model
- All the work is done at classification time
- Works with tiny amounts of training data (single example per class)

K Nearest Neighbors

- Arguably the simplest ML algorithm
- “Non-Parametric” — no assumptions about the form of the classification model
- All the work is done at classification time
- Works with tiny amounts of training data (single example per class)
- The best classification model *ever*???

Supervised Classification



https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

Generative Models	Discriminative Models

Generative Models	Discriminative Models
estimate $P(X, Y)$ first	

Generative Models	Discriminative Models
estimate P(X, Y) first	estimate P(Y | X) directly

Generative Models	Discriminative Models
estimate $P(X, Y)$ first	estimate $P(Y \mid X)$ directly /no explicit probability model

Generative Models	Discriminative Models
estimate P(X, Y) first	estimate P(Y | X) directly /no explicit probability model
Can assign probability to observations, generate new observations	

Generative Models	Discriminative Models
estimate P(X, Y) first	estimate P(Y | X) directly /no explicit probability model
Can assign probability to observations, generate new observations	Only supports classification, less flexible

Generative Models	Discriminative Models
estimate P(X, Y) first	estimate P(Y | X) directly /no explicit probability model
Can assign probability to observations, generate new observations	Only supports classification, less flexible
Often more parameters, but more flexible	

Generative Models	Discriminative Models
estimate P(X, Y) first	estimate P(Y | X) directly /no explicit probability model
Can assign probability to observations, generate new observations	Only supports classification, less flexible
Often more parameters, but more flexible	Often fewer parameters, better performance on small data

Generative Models	Discriminative Models
estimate $P(X, Y)$ first	estimate $P(Y \mid X)$ directly /no explicit probability model
Can assign probability to observations, generate new observations	Only supports classification, less flexible
Often more parameters, but more flexible	Often fewer parameters, better performance on small data
Naive Bayes, Bayes Nets, VAEs, GANs	

Generative Models	Discriminative Models
estimate P(X, Y) first	estimate P(Y | X) directly /no explicit probability model
Can assign probability to observations, generate new observations	Only supports classification, less flexible
Often more parameters, but more flexible	Often fewer parameters, better performance on small data
Naive Bayes, Bayes Nets, VAEs, GANs	Logistic Regression, SVMs, Perceptrons

Generative Models	Discriminative Models
estimate P(X, Y) first	estimate P(Y | X) directly /no explicit probability model
Can assign probability to observations, generate new observations	Only supports classification, less flexible
Often more parameters, but more flexible	Often fewer parameters, better performance on small data
Naive Bayes, Bayes Nets, VAEs, GANs	Logistic Regression, SVMs, Perceptrons
	KNN

Generative Models	Discriminative Models
estimate P(X, Y) first	estimate P(Y | X) directly /no explicit probability model
Can assign probability to observations, generate new observations	Only supports classification, less flexible
Often more parameters, but more flexible	Often few parameters, better performance on small data
Naive Bayes, Bayes Nets, VAEs, GANs	Logistic Regression, SVMs, Perceptrons
	KNN

Supervised Classification

Supervised Classification

Good if not dramatic fizz. ***

Rubbery - rather oxidised. *

Gamy, succulent tannins. Lovely. ****

Provence herbs, creamy, lovely. ****

Lovely mushroomy nose and good length. *****

Quite raw finish. A bit rubbery. **

Supervised Classification

Lovely mushroomy nose and good length. 1

Gamy, succulent tannins. Lovely. 1

Provence herbs, creamy, lovely. 1

Good if not dramatic fizz. 0

Quite raw finish. A bit rubbery. 0

Rubbery - rather oxidised. 0

Superv … sification

Lovely mushroomy nose and good length. 1

Gamy, succulent tannins. Lovely. 1

Provence herbs, creamy, lovely. 1

Quite raw finish. A bit rubbery. 0

Good if not dramatic fizz. 0

Rubbery - rather oxidised. 0

Label	lovely	good	raw	rubbery	rather	mushroomy	gamy	…
1	1	1	0	0	0	0	0	…
1	1	0	0	0	0	0	1	…
1	1	0	0	0	0	0	0	…
0	0	0	1	1	0	0	0	…

Lovely mushroomy nose and good length. 1

Gamy, succulent tannins. Lovely. 1

Provence herbs, creamy, lovely. 1

Quite raw finish. A bit rubbery. 0

Good if not dramatic fizz. 0

Rubbery - rather oxidised. 0

y

Label	lovely	good	raw	rubbery	rather	mushroomy	gamy	…
1	1	1	0	0	0	0	0	…
1	1	0	0	0	0	0	1	…
1	1	0	0	0	0	0	0	…
0	0	0	1	1	0	0	0	…

Supervised classification

Lovely mushroomy nose and good length. 1

Gamy, succulent tannins. Lovely. 1

Provence herbs, creamy, lovely. 1

Quite raw finish. A bit rubbery. 0

Good if not dramatic fizz. 0

Rubbery - rather oxidised. 0

y	X							
Label	lovely	good	raw	rubbery	rather	mushroomy	gamy	…
1	1	1	0	0	0	0	0	…
1	1	0	0	0	0	0	1	…
1	1	0	0	0	0	0	0	…
0	0	0	1	1	0	0	0	…

Supervised Classification

Lovely mushroomy nose and good length. 1

Gamy, succulent tannins. Lovely. 1

Provence herbs, creamy, lovely. 1

Quite raw finish. A bit rubbery. 0

Good if not dramatic fizz. 0

Rubbery - rather oxidised. 0

y X

Label	lovely	good	raw	rubbery	rather	mushroomy	gamy	…
1	1	1	0	0	0	0	0	…
1	1	0	0	0	0	0	1	…
1	1	0	0	0	0	0	0	…
0	0	0	1	1	0	0	0	…
???	0	1	1	0	1	0	1	…

Bayes Rule

Bayes Rule

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Bayes Rule

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Bayes Rule

Label	lovely	good	raw	rubbery	rather	mushroomy	gamy	…
1	1	1	0	0	0	0	0	…

Bayes Rule

Label	lovely	good	raw	rubbery	rather	mushroomy	gamy	…
1	1	1	0	0	0	0	0	…

$$P(Y=1 \mid \text{lovely, good}, \ldots)$$

Bayes Rule

Label	lovely	good	raw	rubbery	rather	mushroomy	gamy	…
1	1	1	0	0	0	0	0	…

$$P(Y=1|\text{lovely, good,}\ldots)$$
$$=P(\text{lovely, good,}\ldots|Y=1)P(Y=1)$$

Bayes Rule

Label	lovely	good	raw	rubbery	rather	mushroomy	gamy	…
1	1	1	0	0	0	0	0	…

$$P(Y=1|\text{lovely, good,}\ldots)$$
$$=P(\text{lovely, good,}\ldots|Y=1)P(Y=1)$$
$$=P(Y=1,\text{ lovely, good,}\ldots)$$

Bayes Rule

Label	lovely	good	raw	rubbery	rather	mushroomy	gamy	…
1	1	1	0	0	0	0	0	…

$$P(Y=1|\text{lovely, good},\ldots)$$
$$=P(\text{lovely, good},\ldots|Y=1)P(Y=1)$$
$$=P(Y=1, \text{lovely, good},\ldots)$$
$$=P(\text{lovely}|Y=1, \text{good},\ldots)P(Y=1, \text{good},\ldots)$$

Bayes Rule

Label	lovely	good	raw	rubbery	rather	mushroomy	gamy	…
1	1	1	0	0	0	0	0	…

$$P(C|x_1, x_2, \ldots, x_k)$$
$$=P(x_1|x_2, \ldots, x_k, C)P(x_2|x_3, \ldots, x_k, C)\ldots P(x_k|C)P(C)$$

Naive Bayes

Label	lovely	good	raw	rubbery	rather	mushroomy	gamy	…
1	1	1	0	0	0	0	0	…

$$P(C|x_1, x_2, \ldots, x_k)$$
$$=P(x_1|x_2, \ldots, x_k, C)P(x_2|x_3, \ldots, x_k, C)\ldots P(x_k|C)P(C)$$

Assume features are independent!

Naive Bayes

Label	lovely	good	raw	rubbery	rather	mushroomy	gamy	…
1	1	1	0	0	0	0	0	…

$$P(C|x_1, x_2, \ldots, x_k)$$
$$=P(x_1|x_2, \ldots, x_k, C)P(x_2|x_3, \ldots, x_k, C)\ldots P(x_k|C)P(C)$$
$$=P(x_1|C)P(x_2|C)\ldots P(x_k|C)P(C)$$

Assume features are independent!

Naive Bayes

Lovely mushroomy nose and good length. 1

Gamy, succulent tannins. Lovely. 1

Provence herbs, creamy, lovely. 1

Quite raw finish. A bit rubbery. 0

Good if not dramatic fizz. 0

Rubbery - rather oxidised. 0

x	P(x|Y=1)	P(x|Y=0)
lovely	??	??
good	??	??
raw	??	??
rubbery	??	??

Clicker Question!

Naive Bayes

Quite mushroomy, a bit dramatic. ???

x	P(x|Y=1)	P(x|Y=0)
a	0.9	0.9
bit	0.2	0.4
dramatic	0.6	0.4
gamy	0.1	0.0
good	0.2	0.2
lovely	0.5	0.1
mushroomy	0.2	0.2
quite	0.7	0.8

Naive Bayes

Quite mushroomy, a bit dramatic. ???

x	P(x|Y=1)	P(x|Y=0)
a	0.9	0.9
bit	0.2	0.4
dramatic	0.6	0.4
gamy	0.1	0.0
good	0.2	0.2
lovely	0.5	0.1
mushroomy	0.2	0.2
quite	0.7	0.8

What do we do now?

Naive Bayes

Quite mushroomy, a bit dramatic. ???

x	P(x|Y=1)	P(x|Y=0)
a	0.9	0.9
bit	0.2	0.4
dramatic	0.6	0.4
gamy	0.1	0.0
good	0.2	0.2
lovely	0.5	0.1
mushroomy	0.2	0.2
quite	0.7	0.8

$$P(Y|X) = P(X|Y)P(Y)$$

Naive Bayes

Quite mushroomy, a bit dramatic. ???

x	P(x|Y=1)	P(x|Y=0)
a	0.9	0.9
bit	0.2	0.4
dramatic	0.6	0.4
gamy	0.1	0.0
good	0.2	0.2
lovely	0.5	0.1
mushroomy	0.2	0.2
quite	0.7	0.8

???

$$P(Y|X) = P(X|Y)P(Y)$$

Naive Bayes

Quite mushroomy, a bit dramatic. ???

x	P(x|Y=1)	P(x|Y=0)
a	0.9	0.9
bit	0.2	
dramatic	0.6	
gamy	0.1	0.0
good	0.2	0.2
lovely	0.5	0.1
mushroomy	0.2	0.2
quite	0.7	0.8

Domain knowledge
or estimate from data

$$P(Y|X) = P(X|Y)P(Y)$$

Naive Bayes

Quite mushroomy, a bit dramatic. ???

x	P(x|Y=1)	P(x|Y=0)
a	0.9	0.9
		0.4
		0.4
		0.0
good	0.2	0.2
lovely	0.5	0.1
mushroomy	0.2	0.2
quite	0.7	0.8

Decision rule:
argmax_y P(Y=y|X)

$$P(Y|X) = P(X|Y)P(Y)$$

Clicker Question!

Naive Bayes

x	P(x|Y=1)	P(x|Y=0)
a	0.9	0.9
bit	0.2	0.4
dramatic	0.6	0.4
gamy	0.1	0.0
good	0.2	0.2
lovely	0.5	0.1
mushroomy	0.2	0.2
quite	0.7	0.8

Naive Bayes

x	P(x|Y=1)	P(x|Y=0)
a	0.9	0.9
bit	0.2	0.4
dramatic	0.6	0.4
gamy	0.1	0.0
good	0.2	0.2
lovely	0.5	0.1
mushroomy	0.2	0.2
quite	0.7	0.8

Naive Bayes

x	P(x|Y=1)	P(x|Y=0)
a	**0.9**	0.9
bit	0.2	0.4
dramatic	0.6	0.4
gamy	0.1	0.0
good	0.2	0.2
lovely	0.5	0.1
mushroomy	0.2	0.2
quite	0.7	0.8

A … 0.9

Naive Bayes

x	P(x|Y=1)	P(x|Y=0)
a	0.9	0.9
bit	0.2	0.4
dramatic	0.6	0.4
gamy	0.1	0.0
good	0.2	0.2
lovely	0.5	0.1
mushroomy	0.2	0.2
quite	**0.7**	0.8

A quite … 0.63

Naive Bayes

x	P(x|Y=1)	P(x|Y=0)
a	0.9	0.9
bit	0.2	0.4
dramatic	**0.6**	0.4
gamy	0.1	0.0
good	0.2	0.2
lovely	0.5	0.1
mushroomy	0.2	0.2
quite	0.7	0.8

A quite dramatic … 0.38

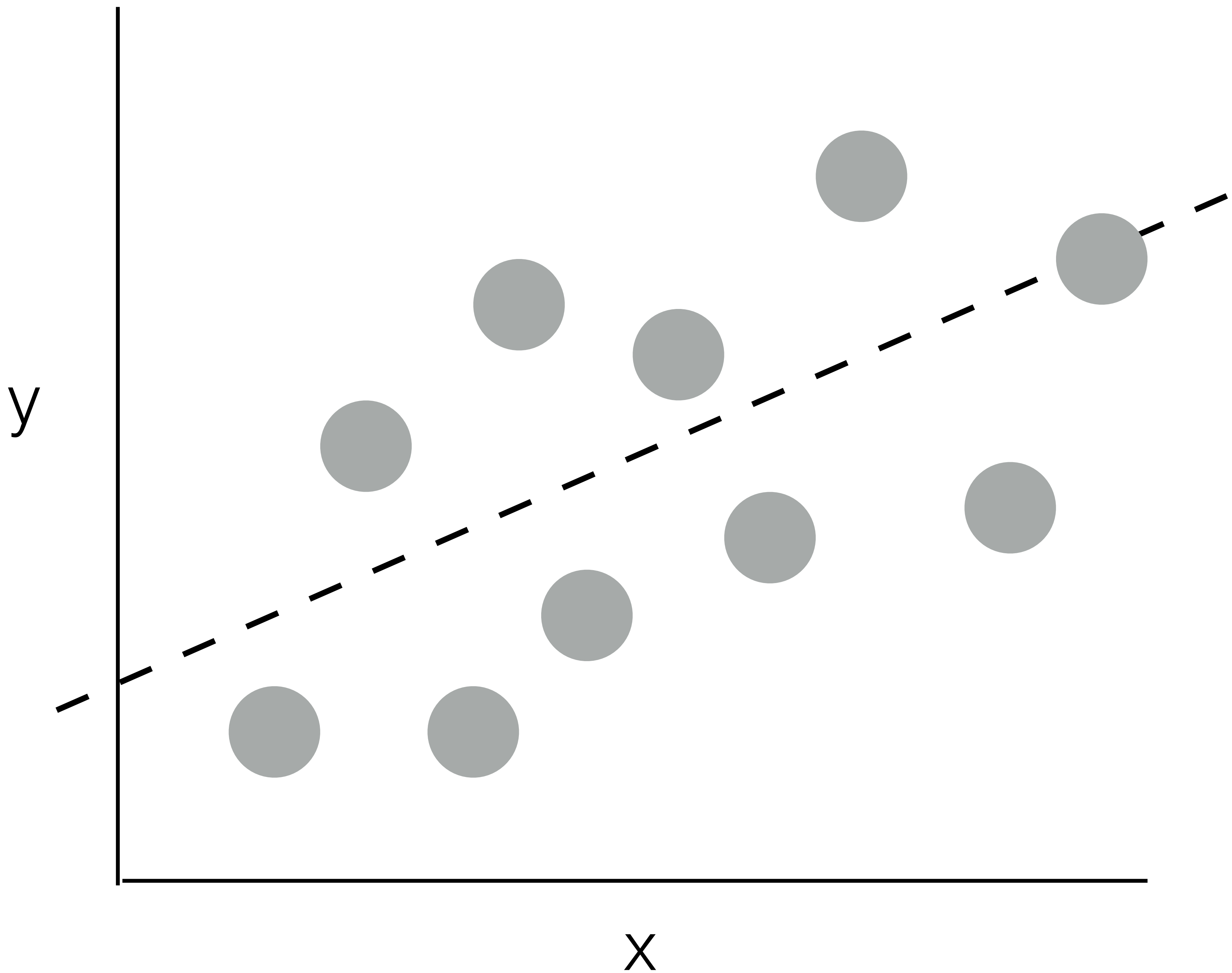
Naive Bayes

x	P(x|Y=1)	P(x|Y=0)
a	0.9	0.9
bit	0.2	0.4
dramatic	0.6	0.4
gamy	**0.1**	0.0
good	0.2	0.2
lovely	0.5	0.1
mushroomy	0.2	0.2
quite	0.7	0.8

A quite dramatic gamy … 0.04

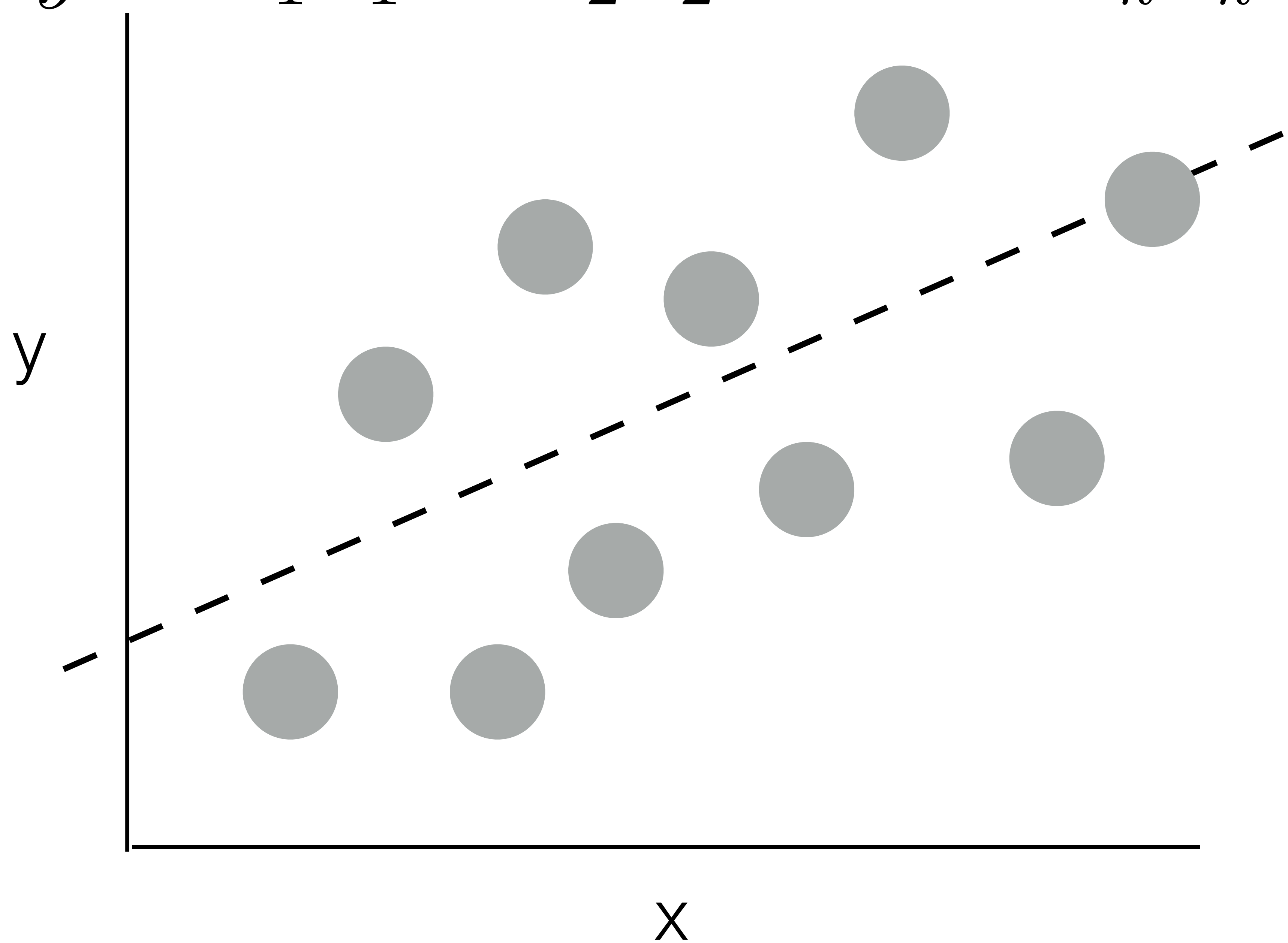
Linear Regression

Linear Regression



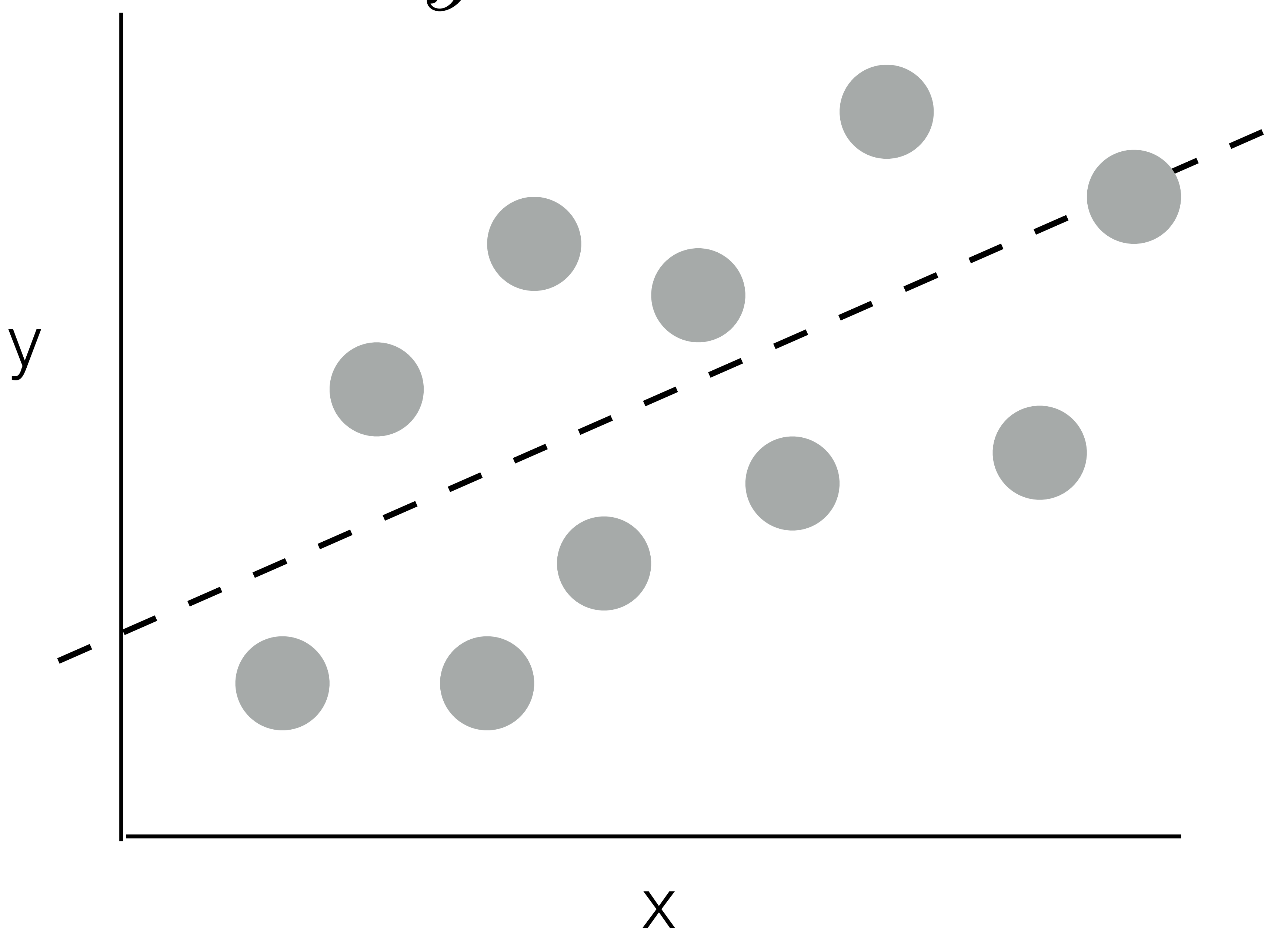
Linear Regression

$$y = w_1x_1 + w_2x_2 + \cdots + w_kx_k$$



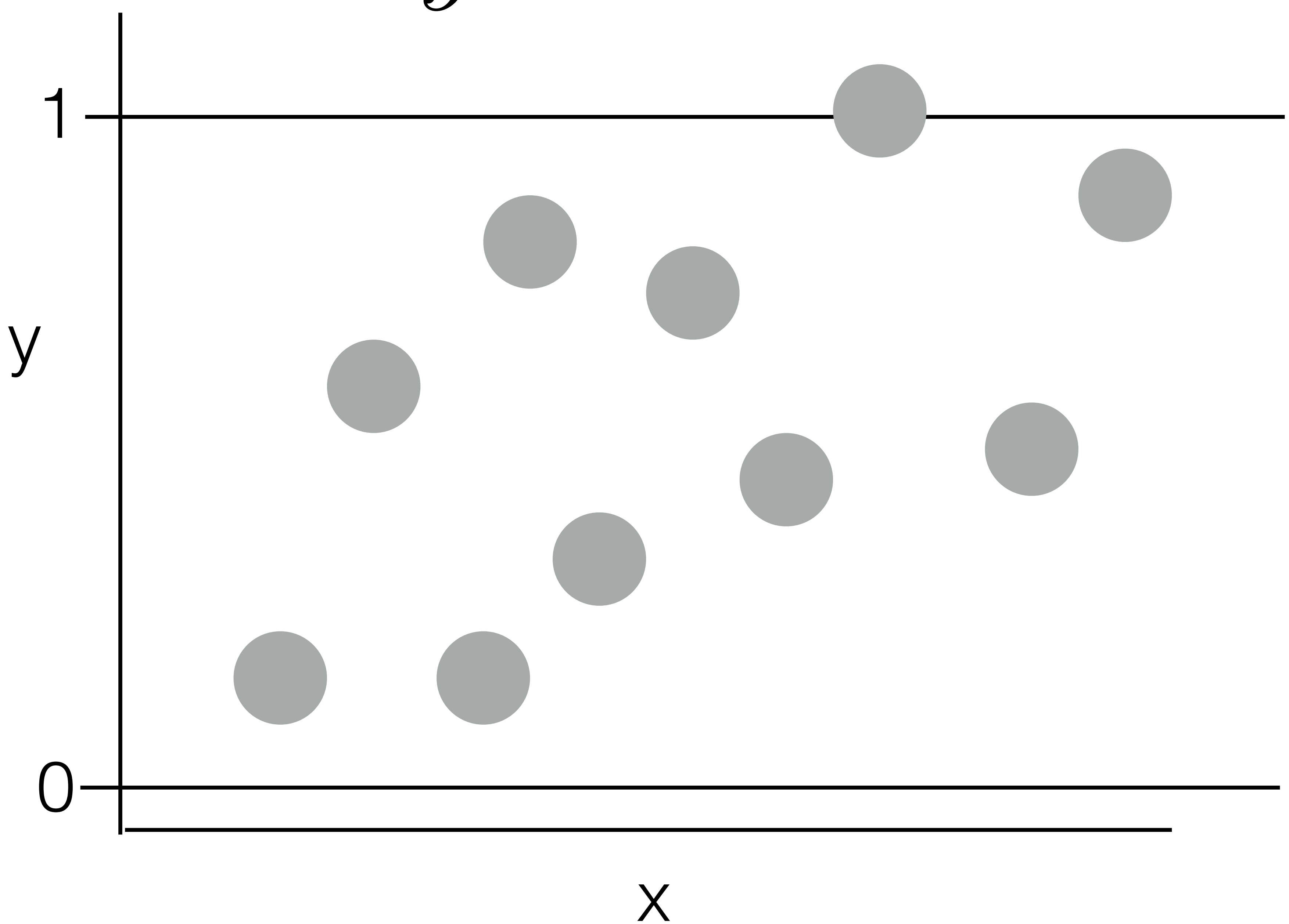
Linear Regression

$$y = \vec{w} \cdot \vec{x}$$



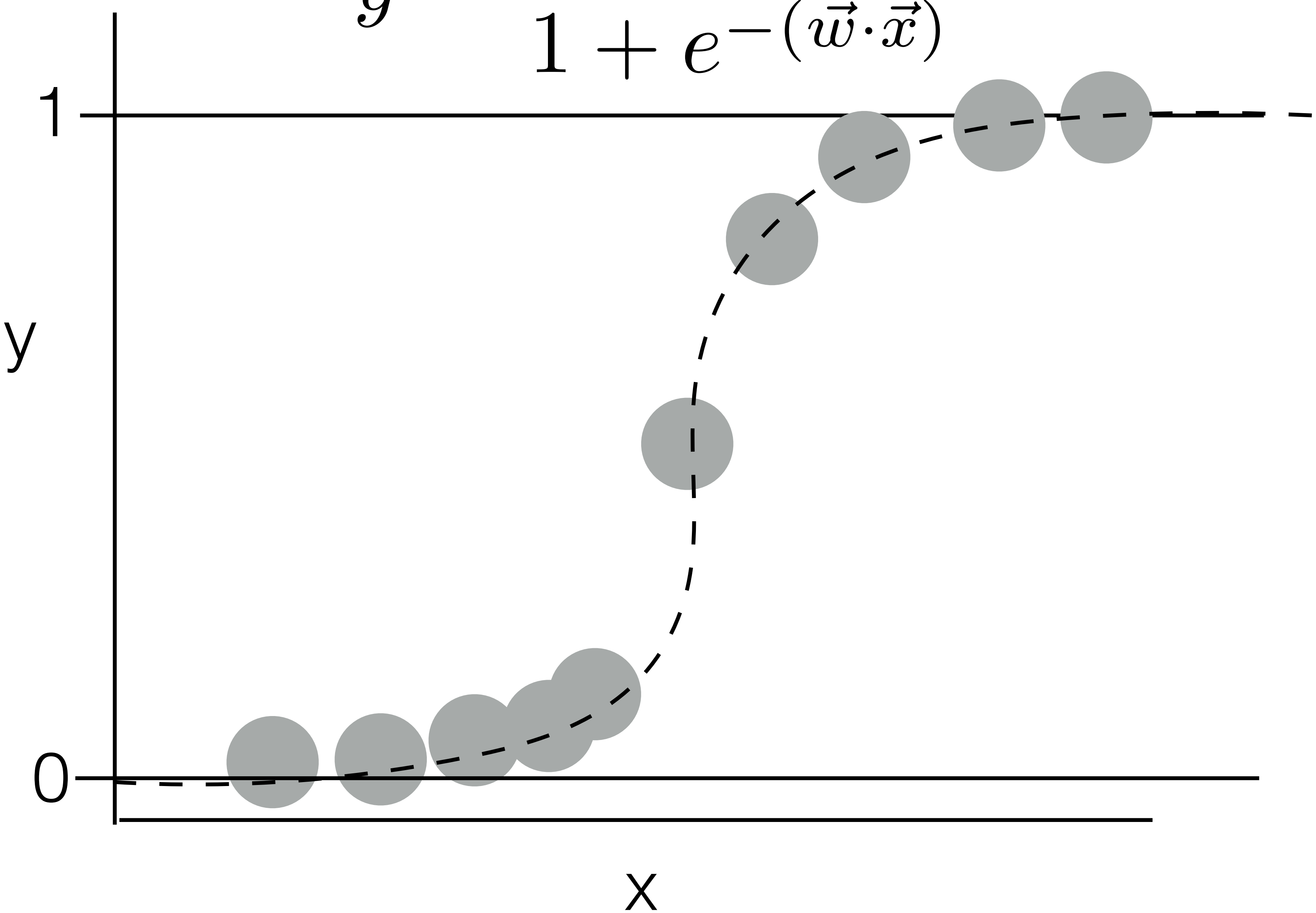
Linear Regression

$$y = \vec{w} \cdot \vec{x}$$



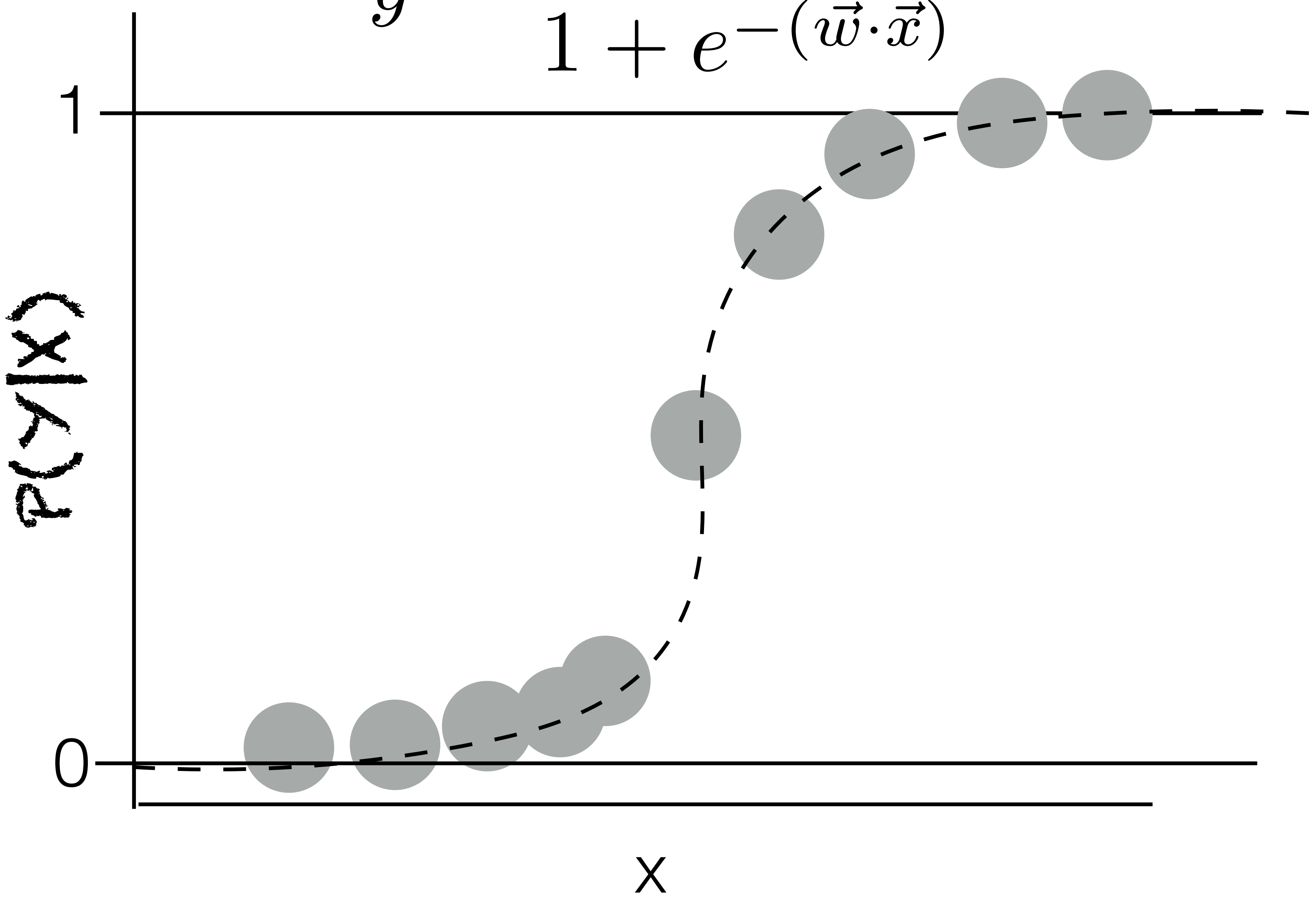
Logistic Regression

$$y = \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x})}}$$



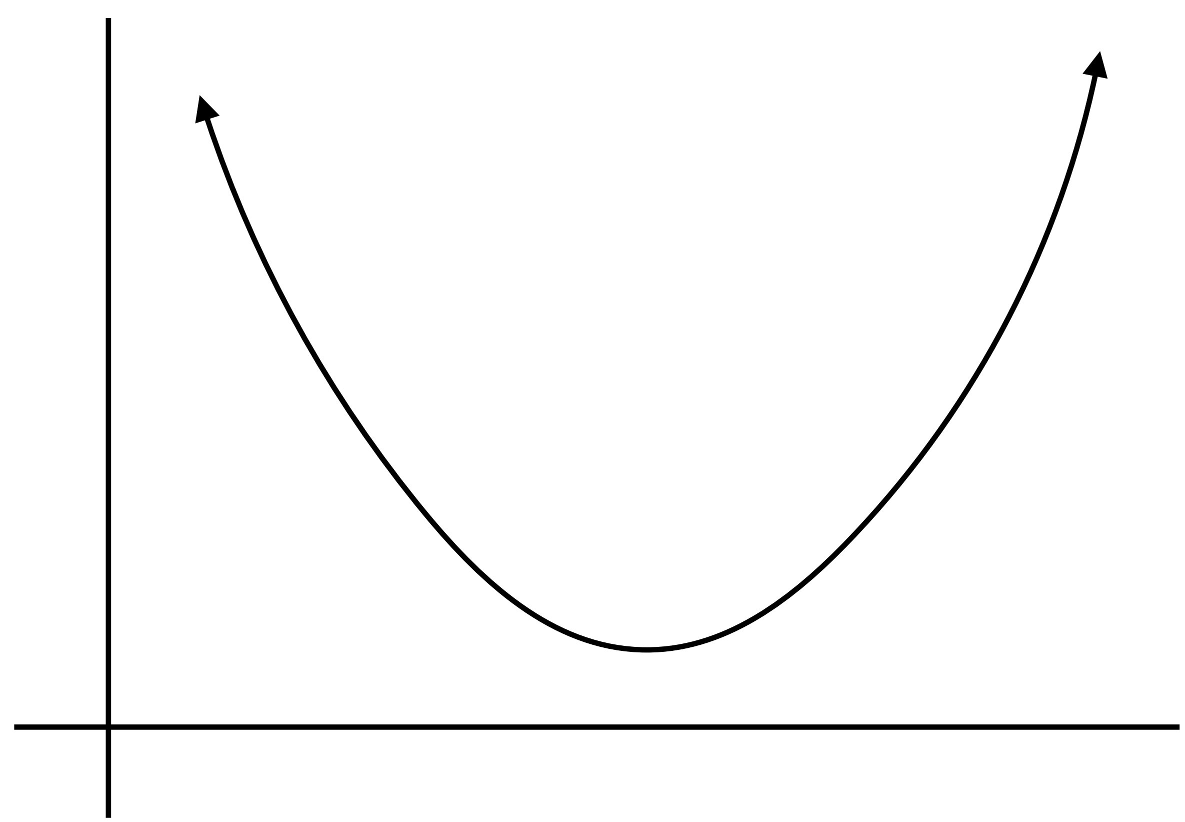
Logistic Regression

$$y = \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x})}}$$



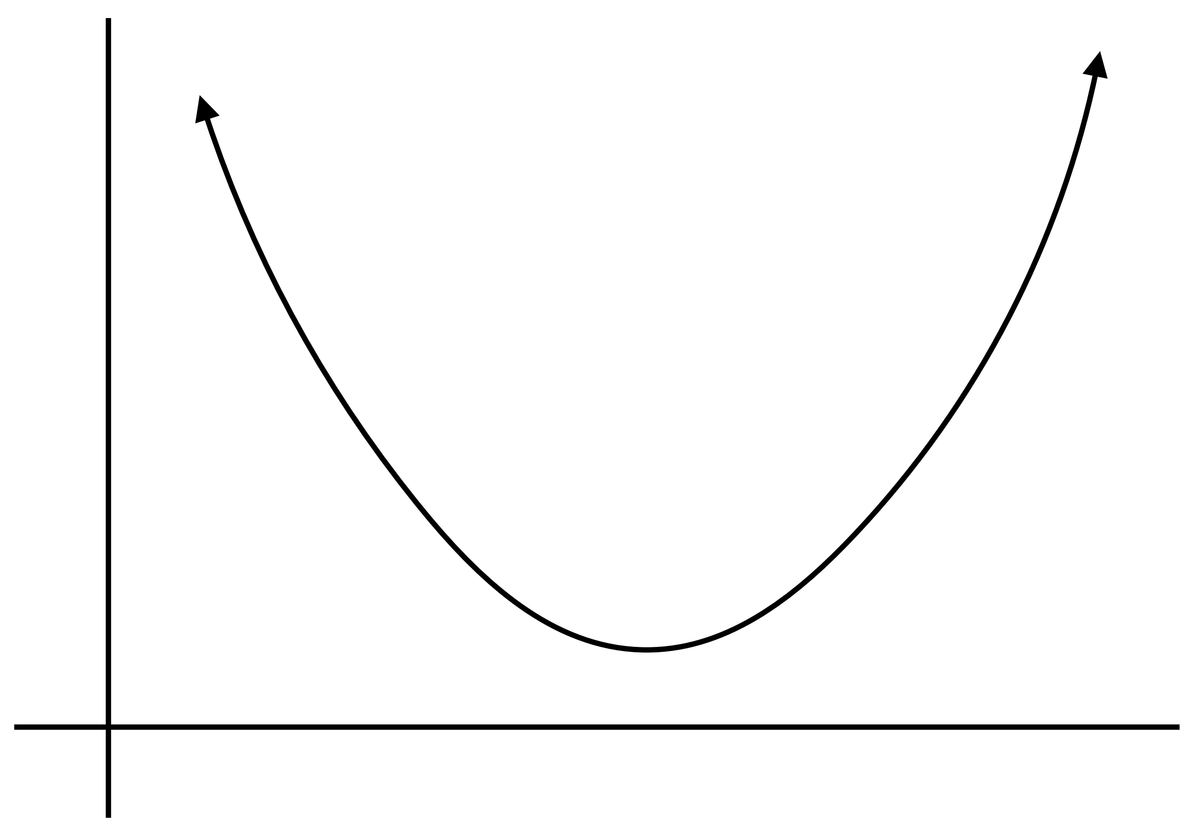
Linear Regression

minimize $$\sum_{i=1}^{n}(Y_i - \hat{Y})^2$$



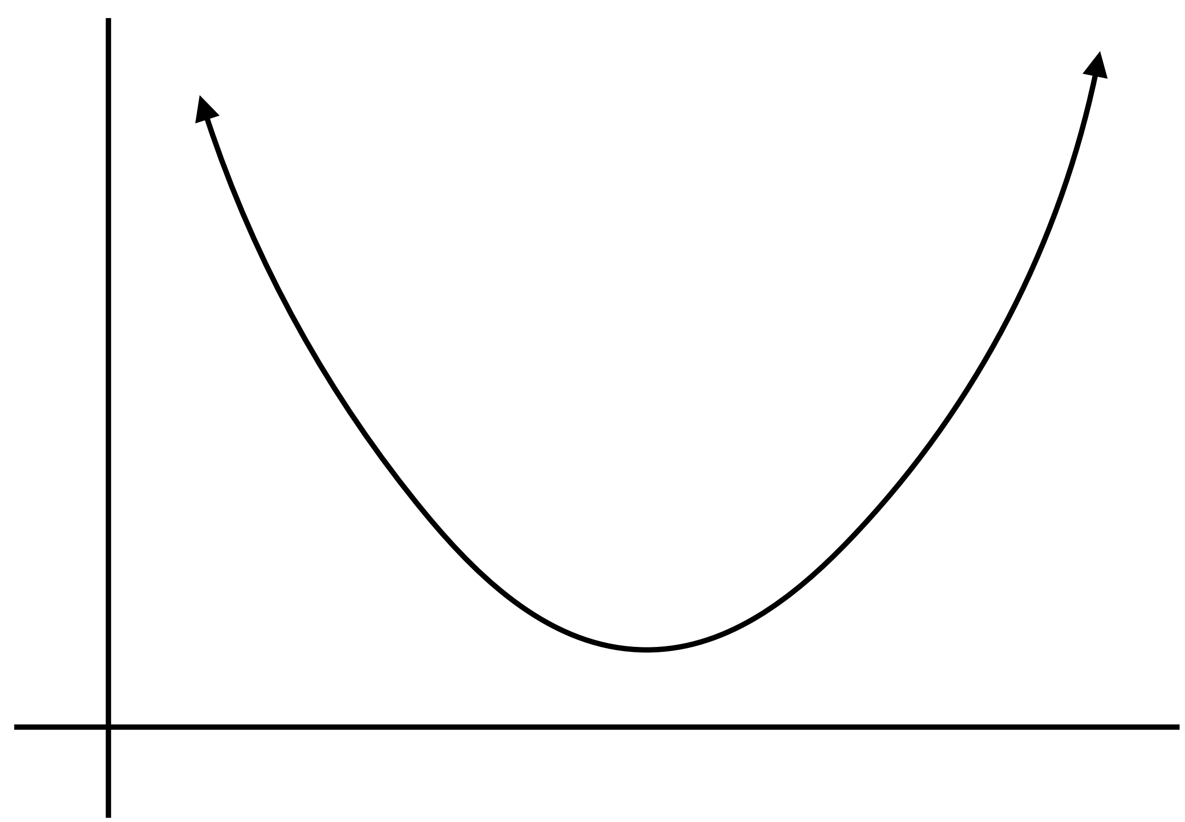
Logistic Regression

minimize $-logP(Y|\hat{Y})$



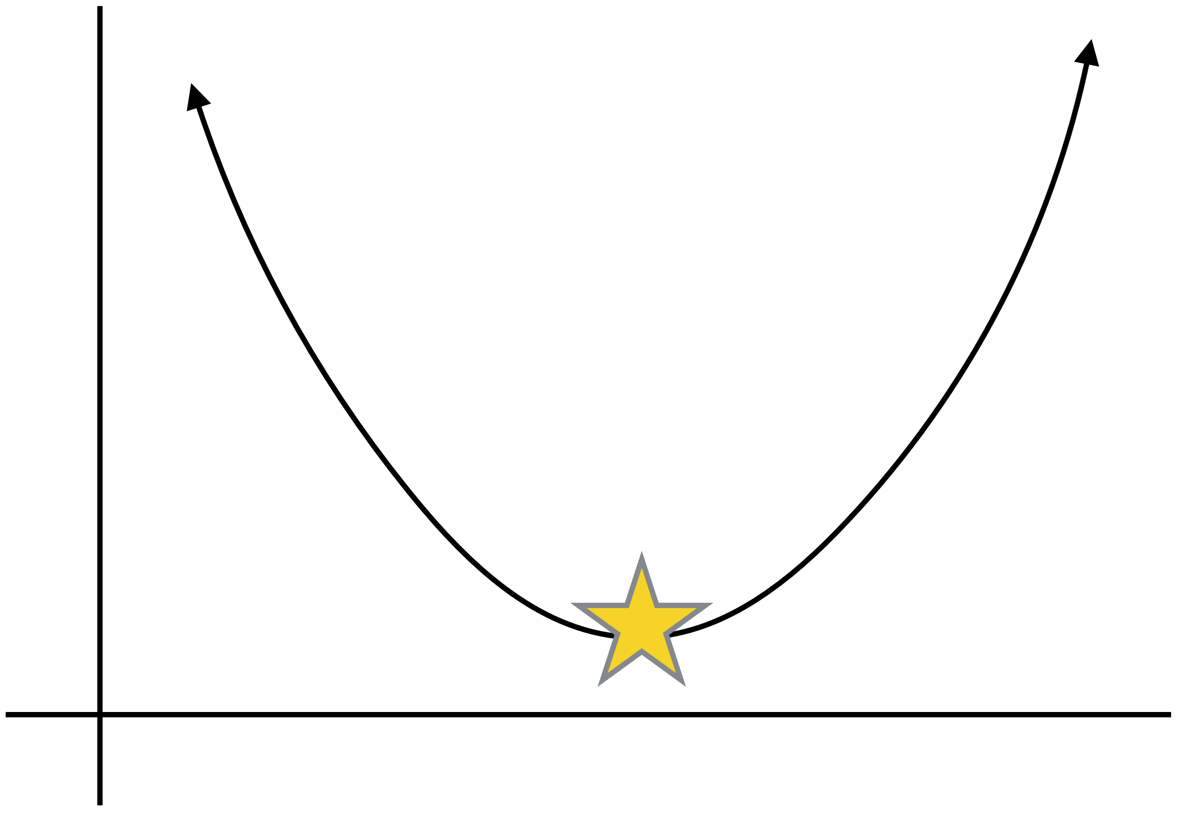
Logistic Regression

minimize $-Y log\hat{Y} + (1-Y)log(1-\hat{Y})$



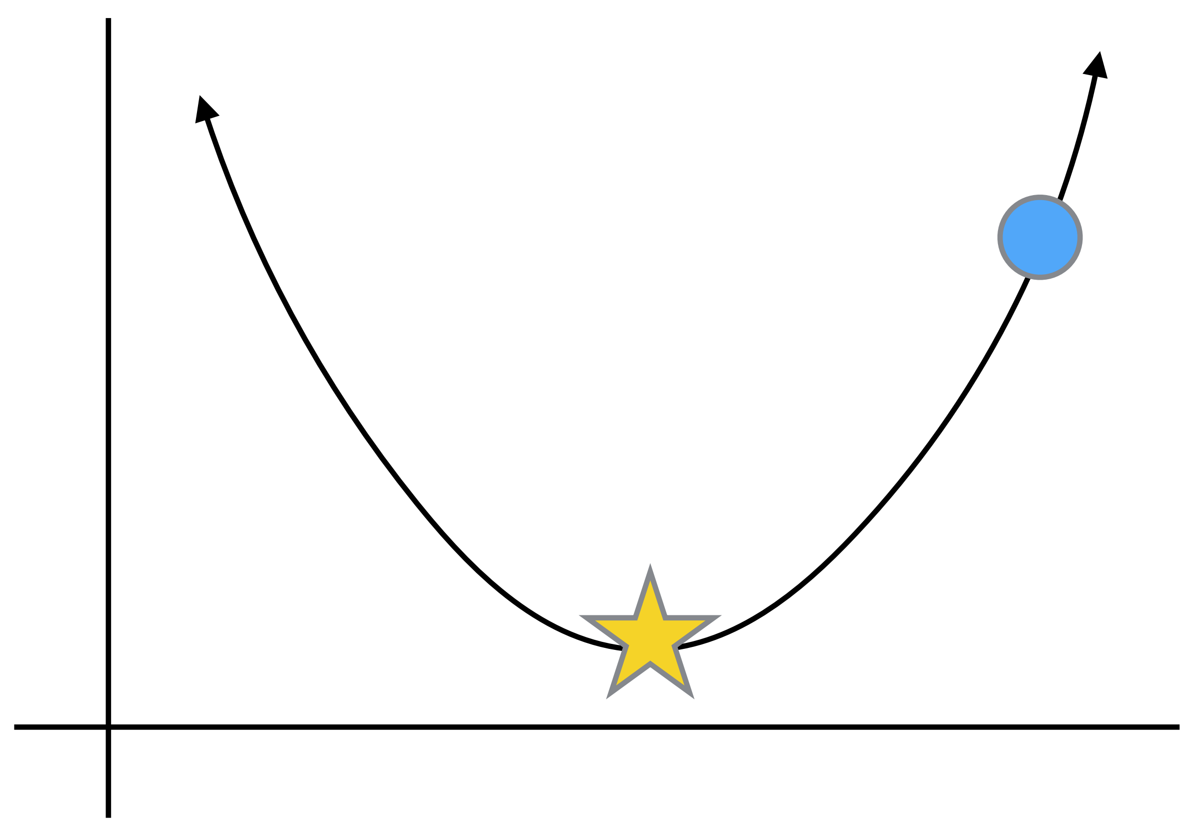
Logistic Regression

minimize $-Y log\hat{Y} + (1-Y)log(1-\hat{Y})$



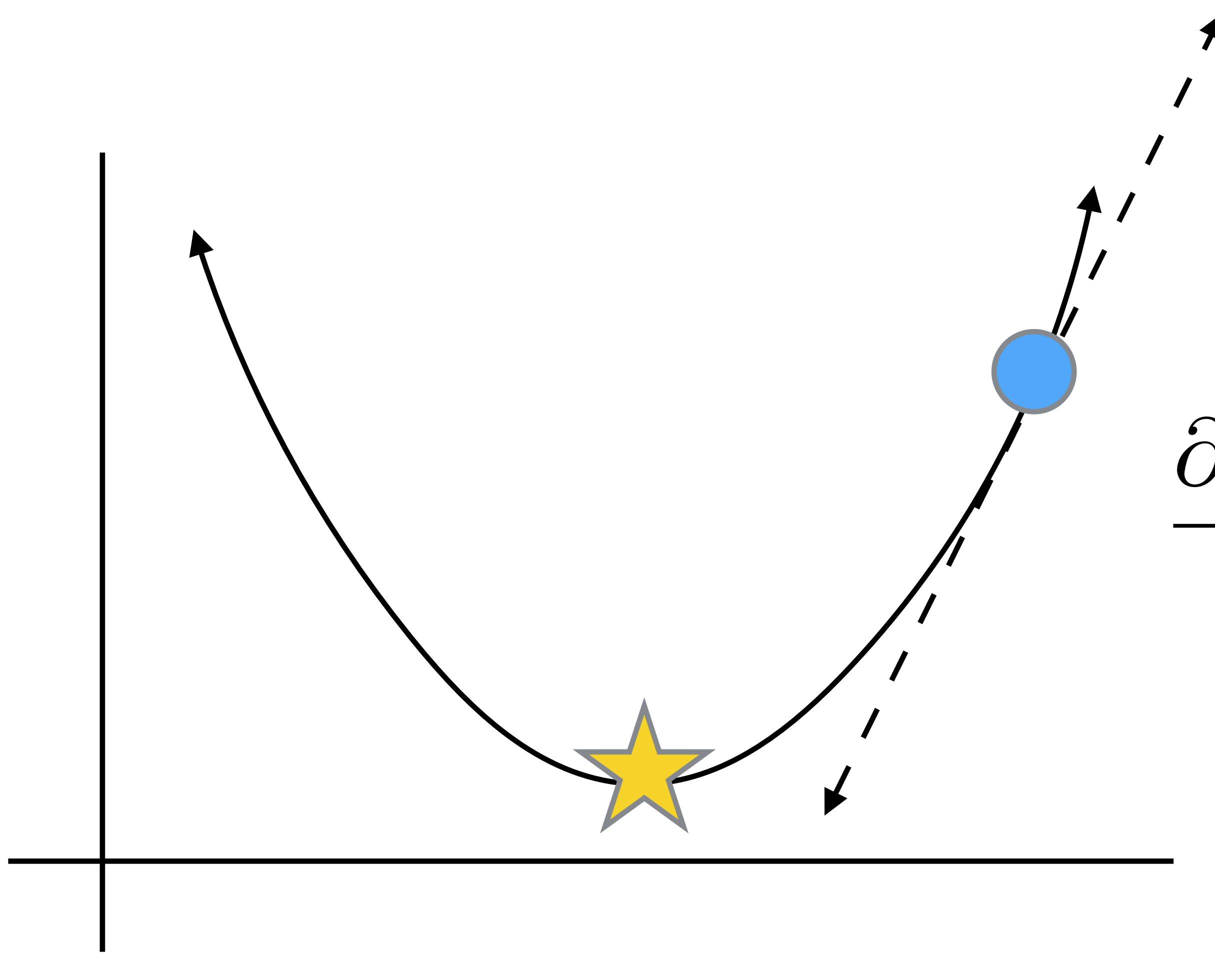
Logistic Regression

minimize $-Y log\hat{Y} + (1 - Y)log(1 - \hat{Y})$



Logistic Regression

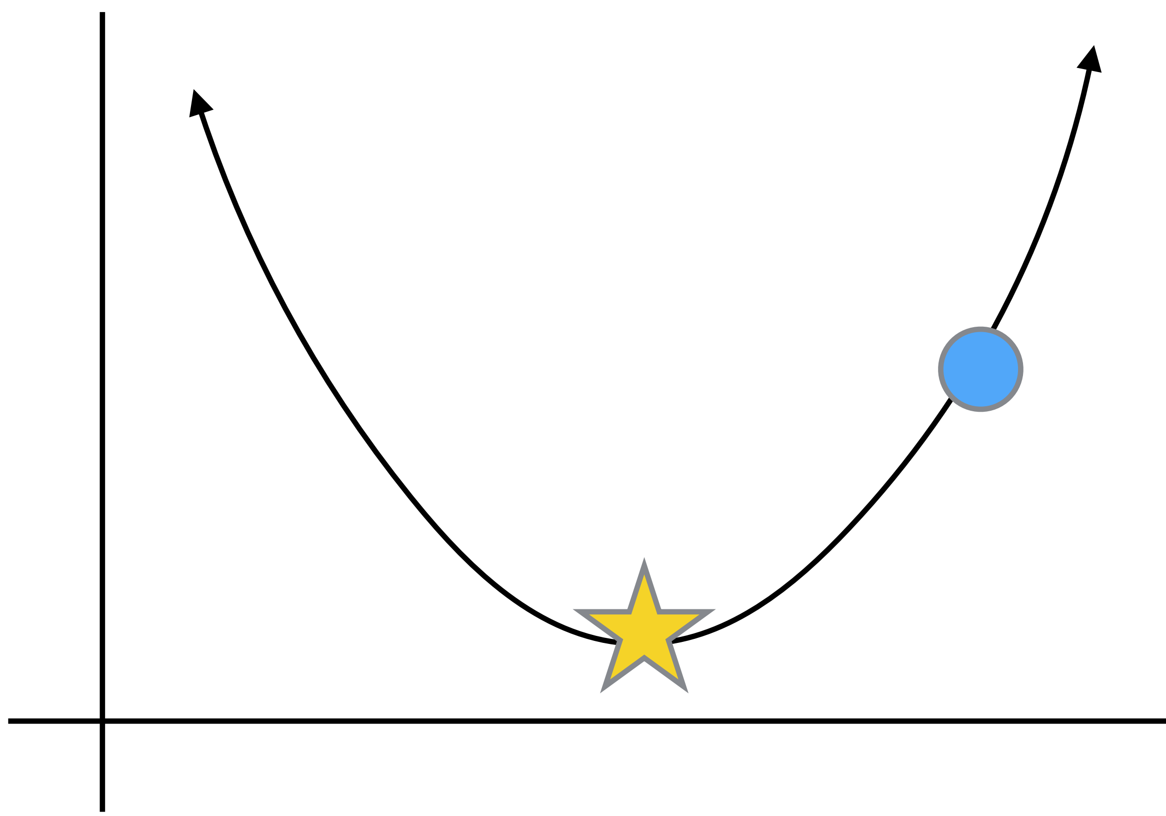
minimize $-Y log\hat{Y} + (1 - Y)log(1 - \hat{Y})$



$$\frac{\partial \text{loss}}{\partial w}$$

Logistic Regression

minimize $-Y log\hat{Y} + (1-Y)log(1-\hat{Y})$



$$\frac{\partial \text{loss}}{\partial w}$$

Logistic Regression

Naive Bayes

x	P(x|Y=1)
a	0.9
bit	0.2
dramatic	0.6
gamy	0.1
good	0.2
lovely	0.5
mushroo	0.2
quite	0.7

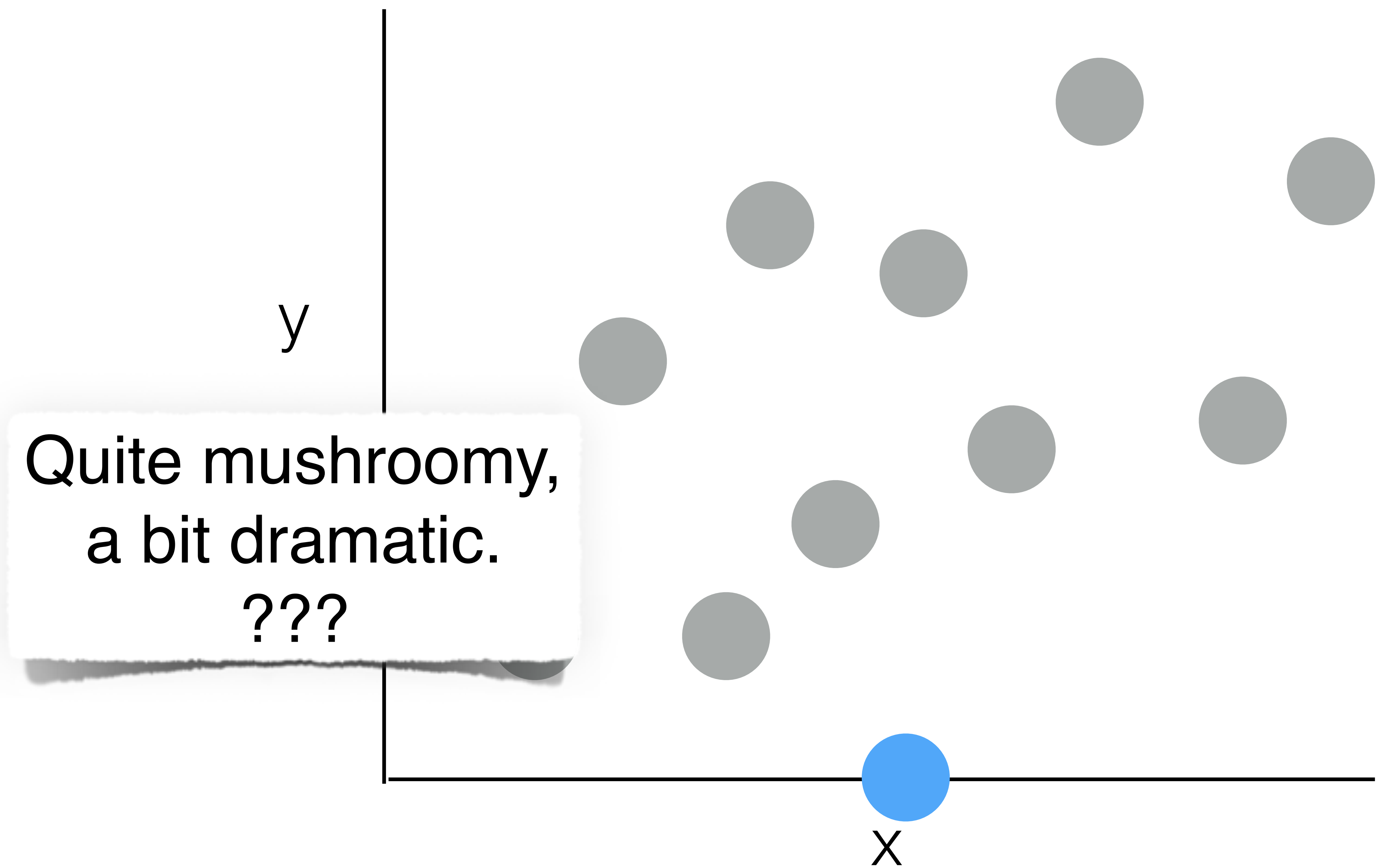
Logistic Regression

Logistic Regression

x	???
a	0.9
bit	0.4
dramatic	1.0
gamy	0.7
good	0.2
lovely	0.4
mushroom	0.8
quite	0.7

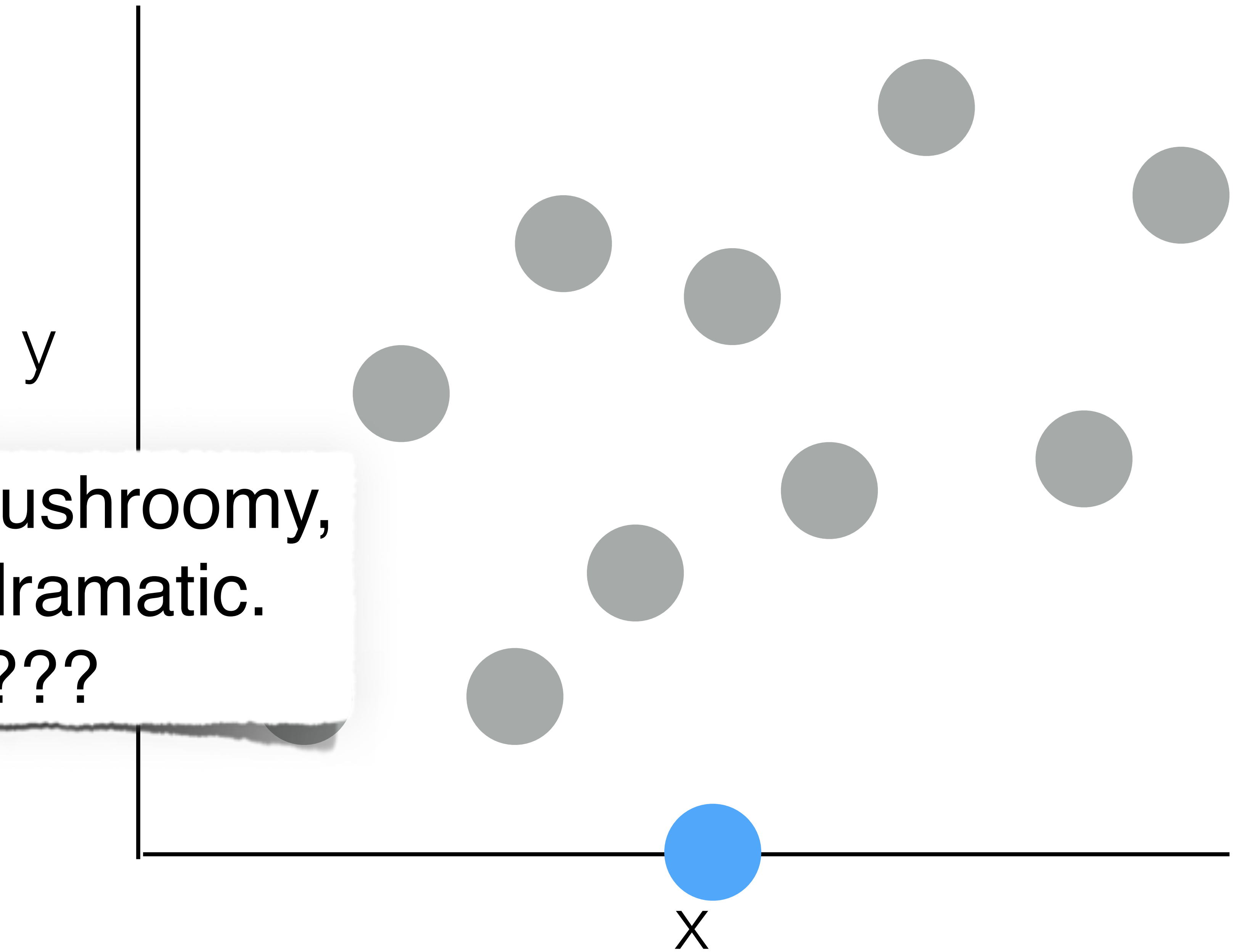
Clicker Question!

Logistic Regression



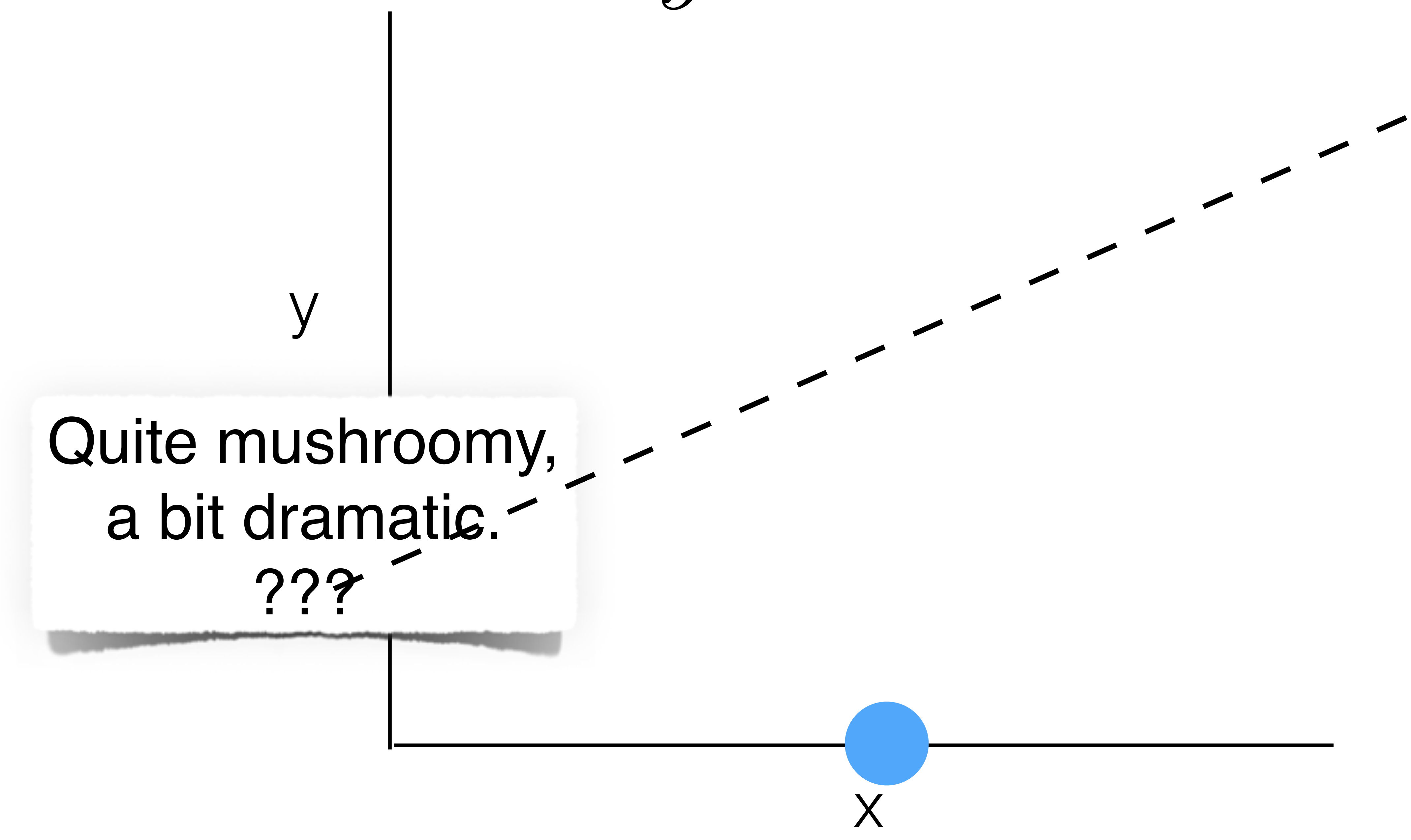
ogistic Regression

What do we do now?



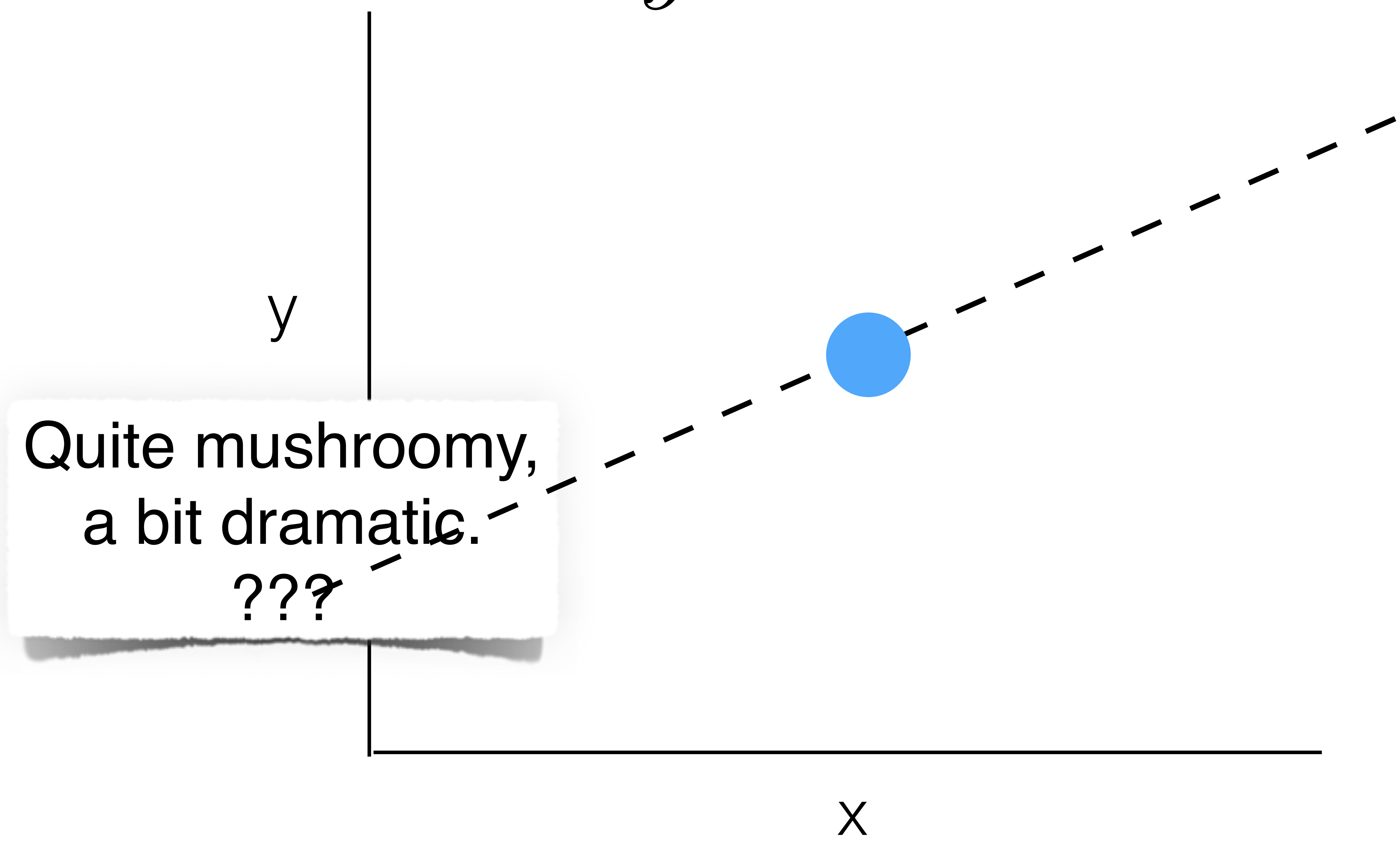
Logistic Regression

$$y = \vec{w} \cdot \vec{x}$$



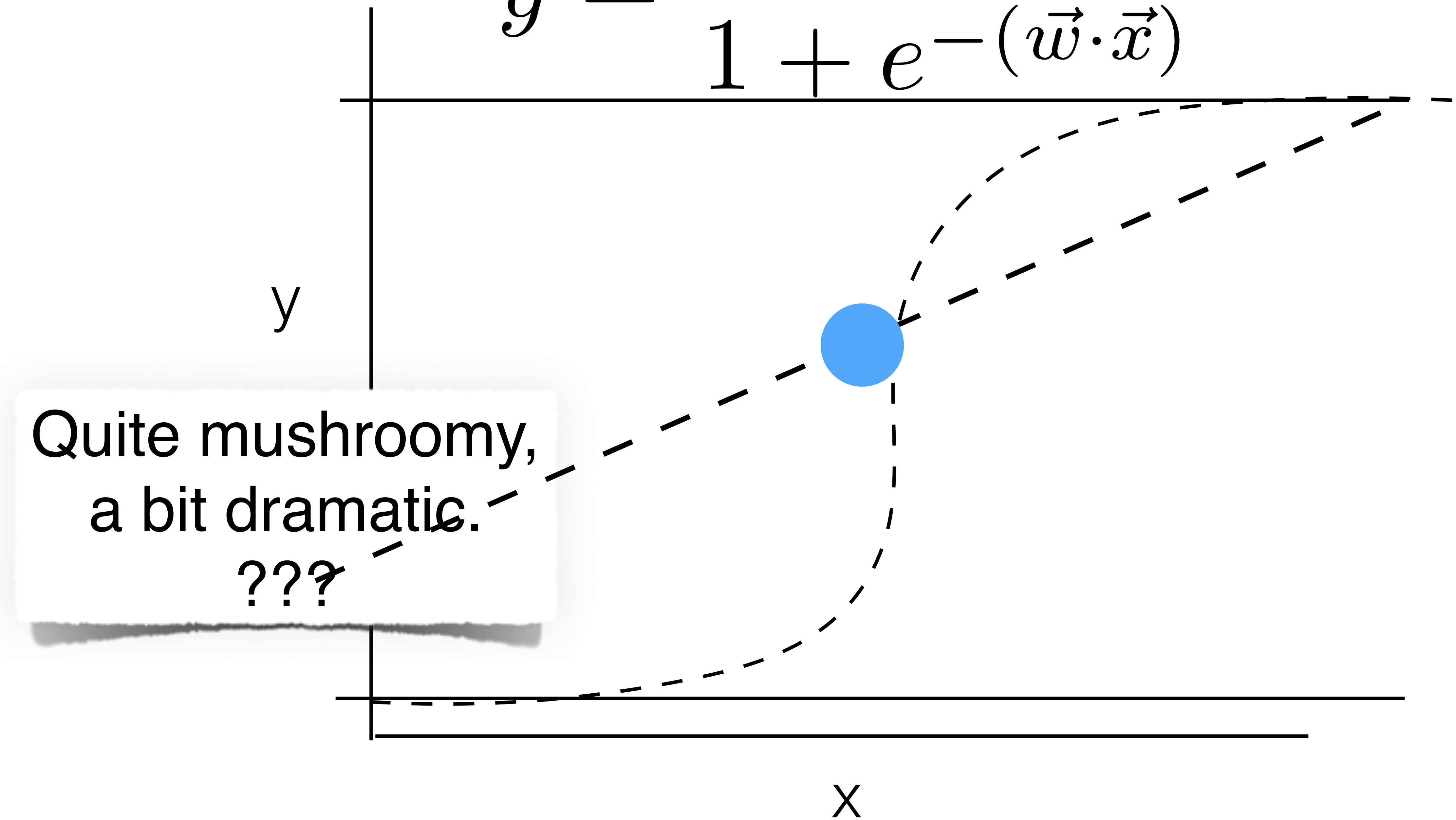
Logistic Regression

$$y = \vec{w} \cdot \vec{x}$$



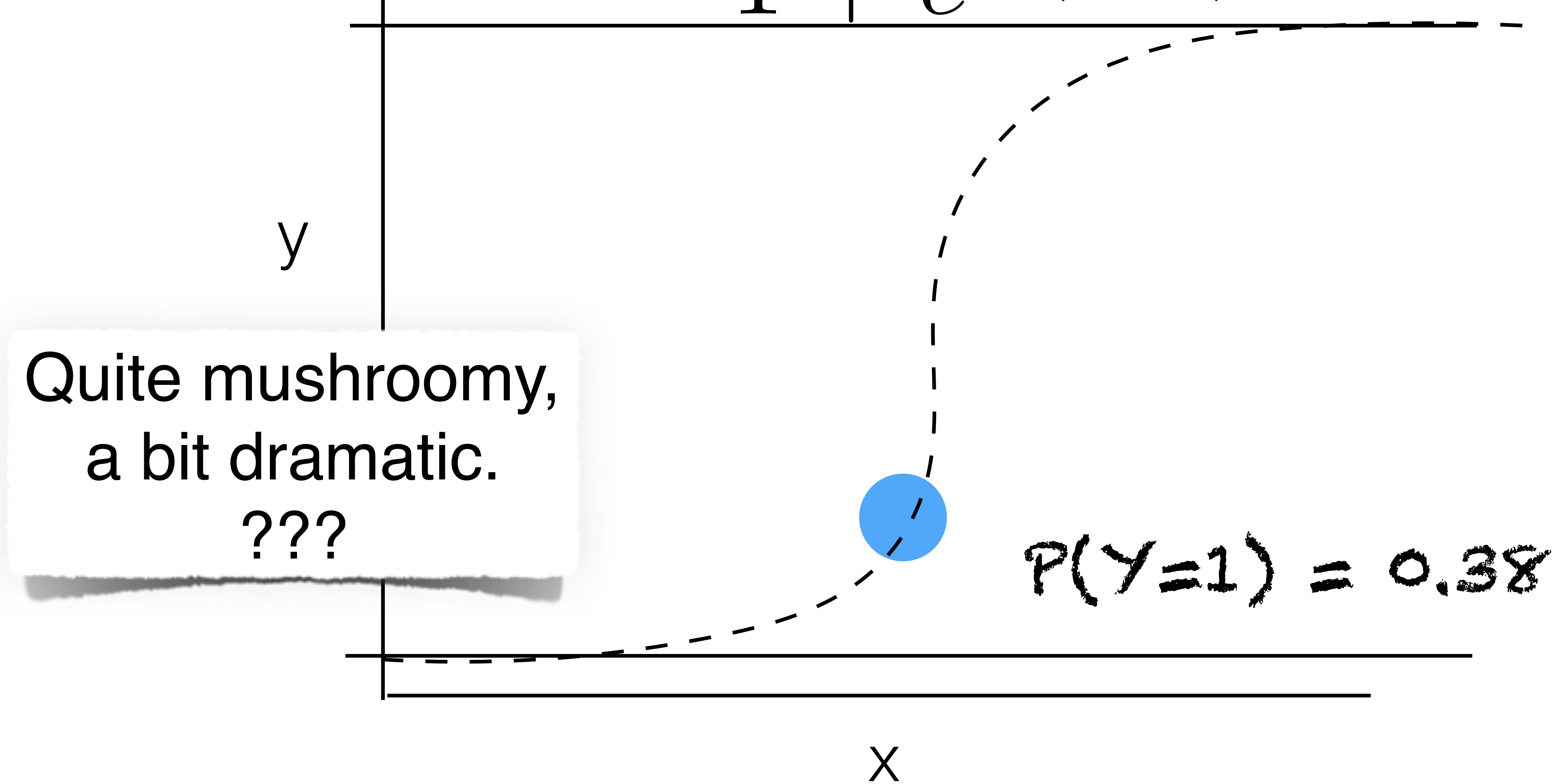
Logistic Regression

$$y = \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x})}}$$



Logistic Regression

$$y = \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x})}}$$



Code-along!

```
from sklearn.linear_model import LogisticRegression
```