

Classification

March 19, 2020

Data Science CSCI 1951A

Brown University

Instructor: Ellie Pavlick

HTAs: Josh Levin, Diane Mutako, Sol Zitter

Today

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

Supervised vs. Unsupervised Learning

- Supervised: Explicit data labels
 - Sentiment analysis—review text -> star ratings
 - Image tagging—image -> caption
- Unsupervised: No explicit labels
 - Clustering—find groups similar customers
 - Dimensionality Reduction—find features that differentiate individuals

Classification

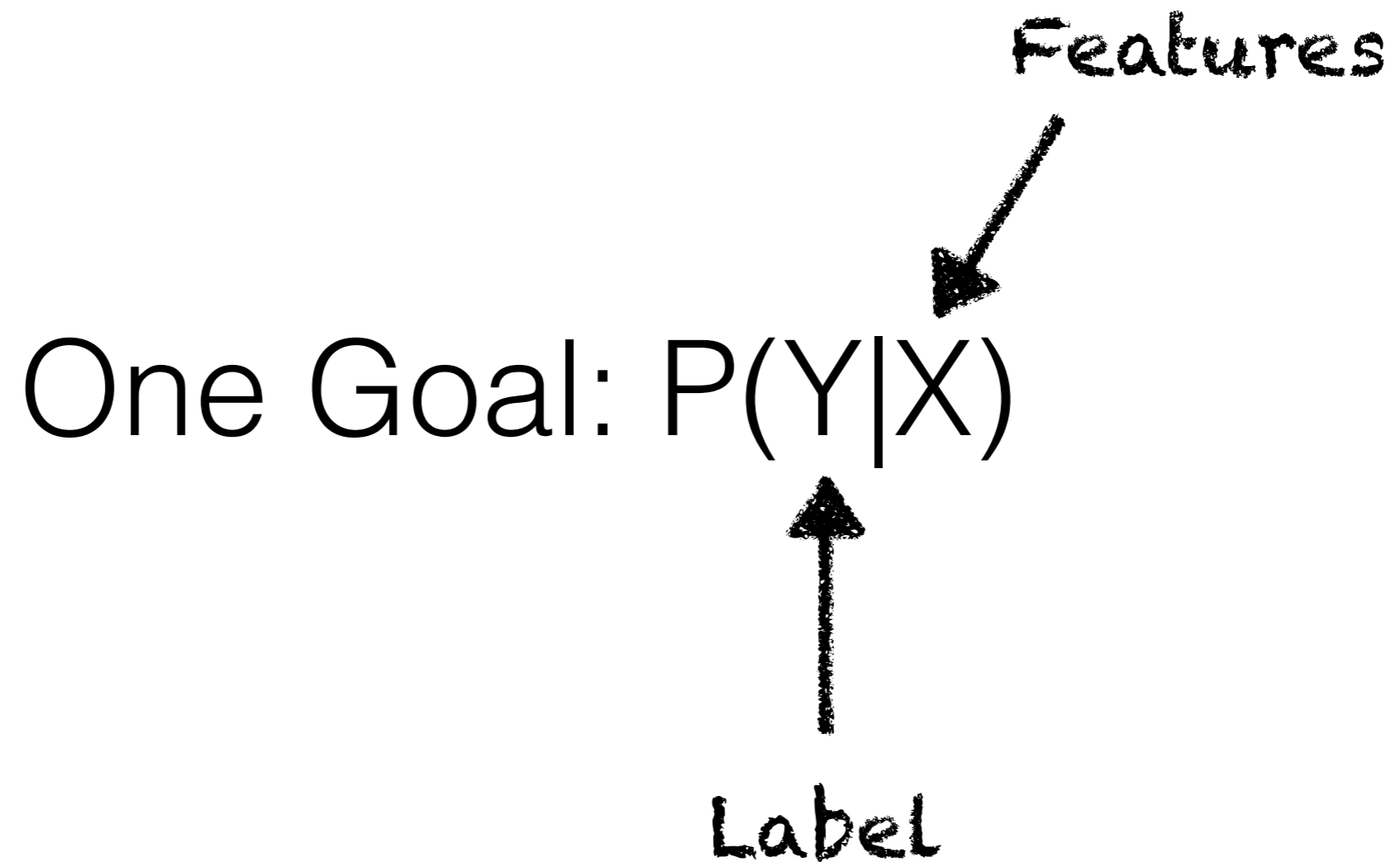
One Goal: $P(Y|X)$

Classification

One Goal: $P(Y|X)$

↑
Label

Classification

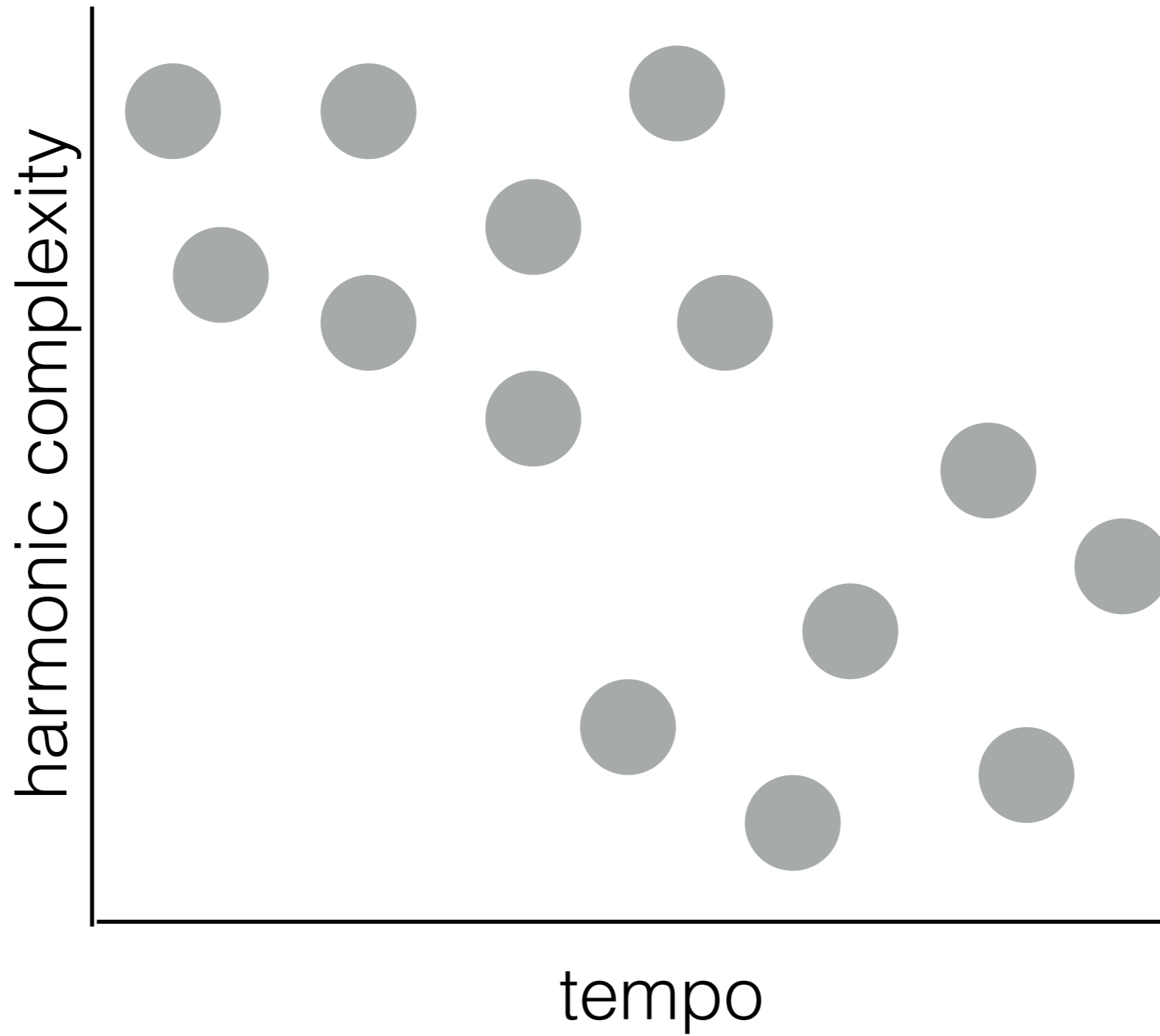


Classification

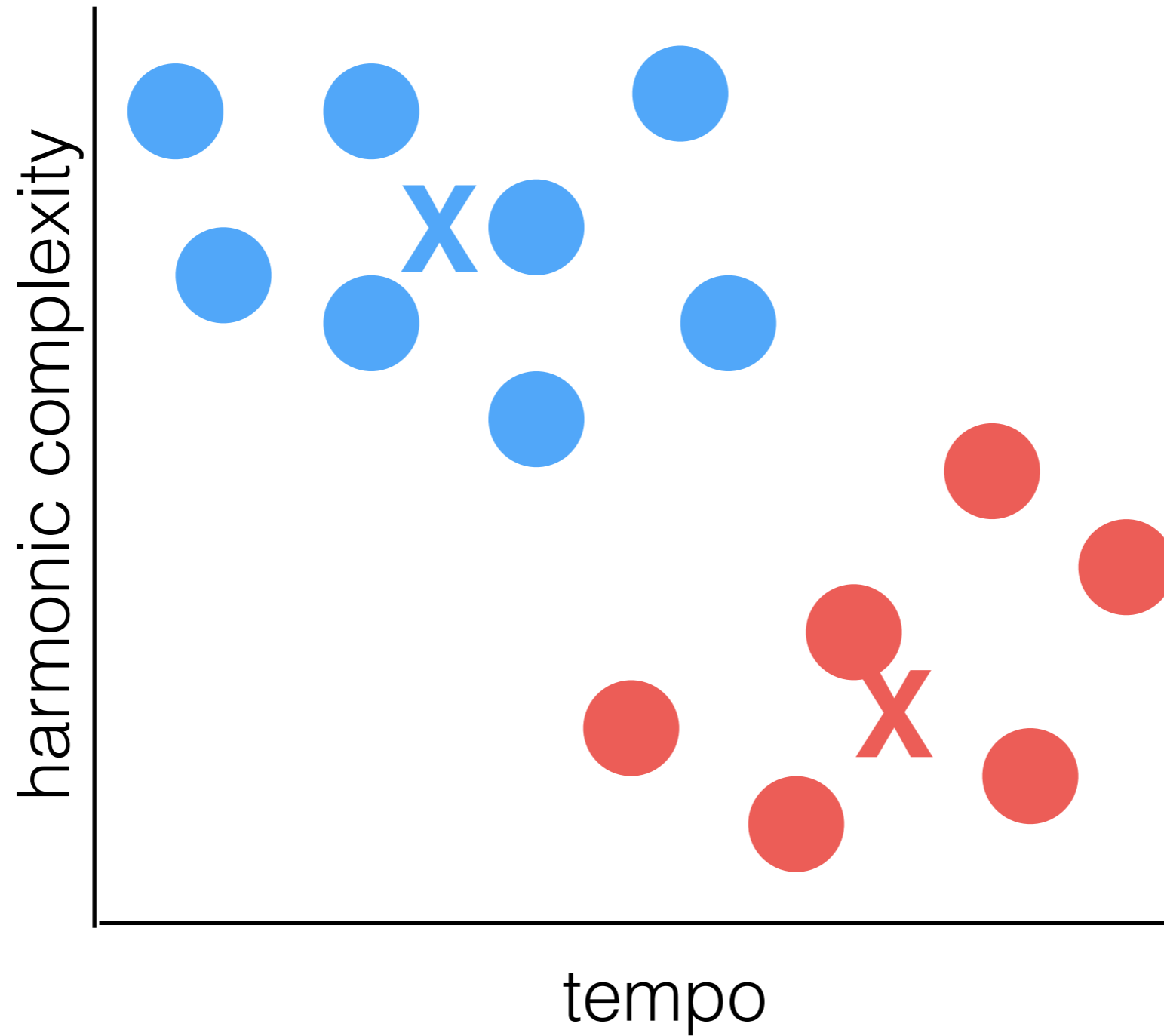
One Goal: $P(Y|X)$

$P(\text{email is spam} \mid \text{words in the message})$
 $P(\text{genre of song} \mid \text{tempo, harmony, lyrics...})$
 $P(\text{article clicked} \mid \text{title, font, photo...})$

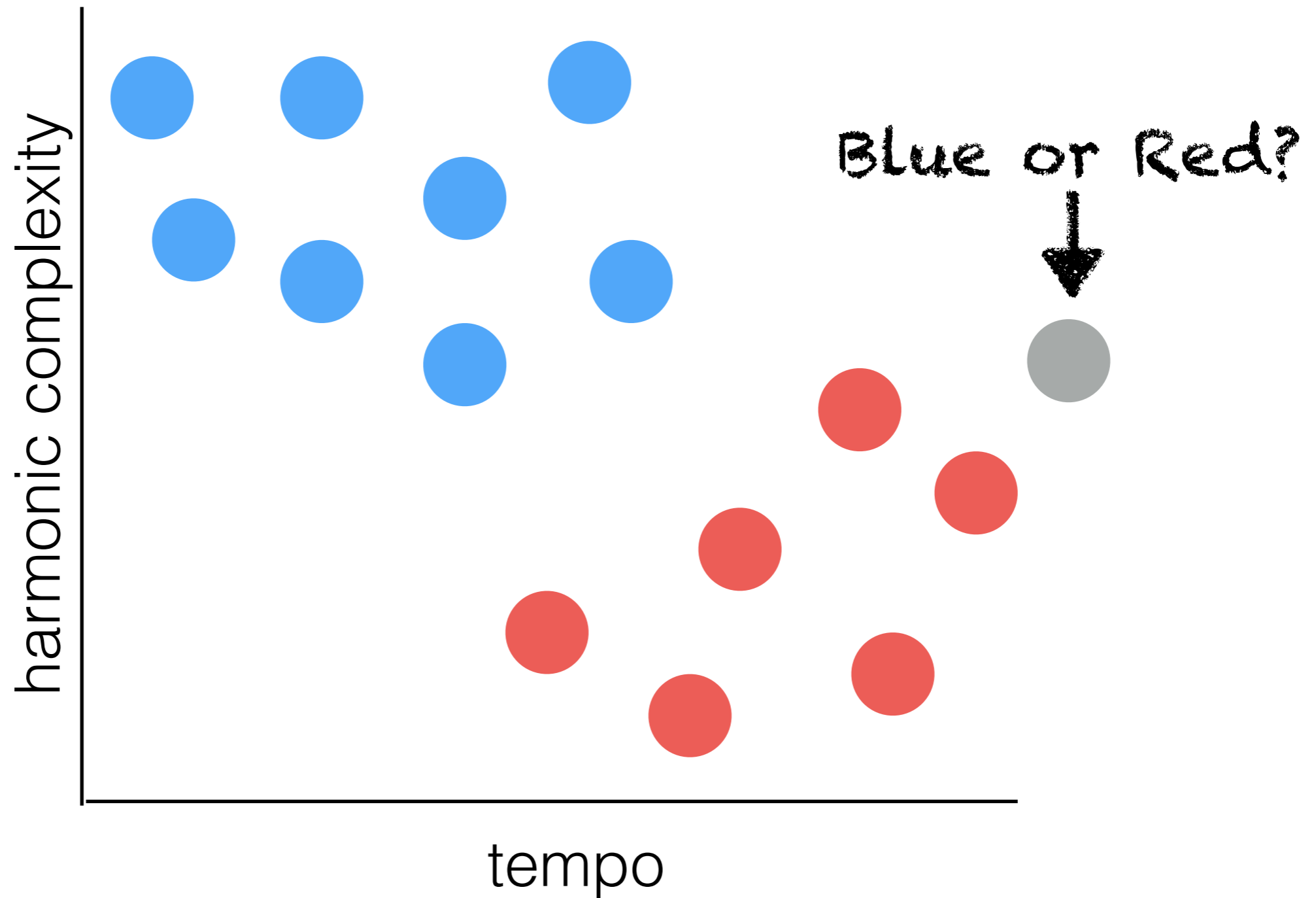
K Means



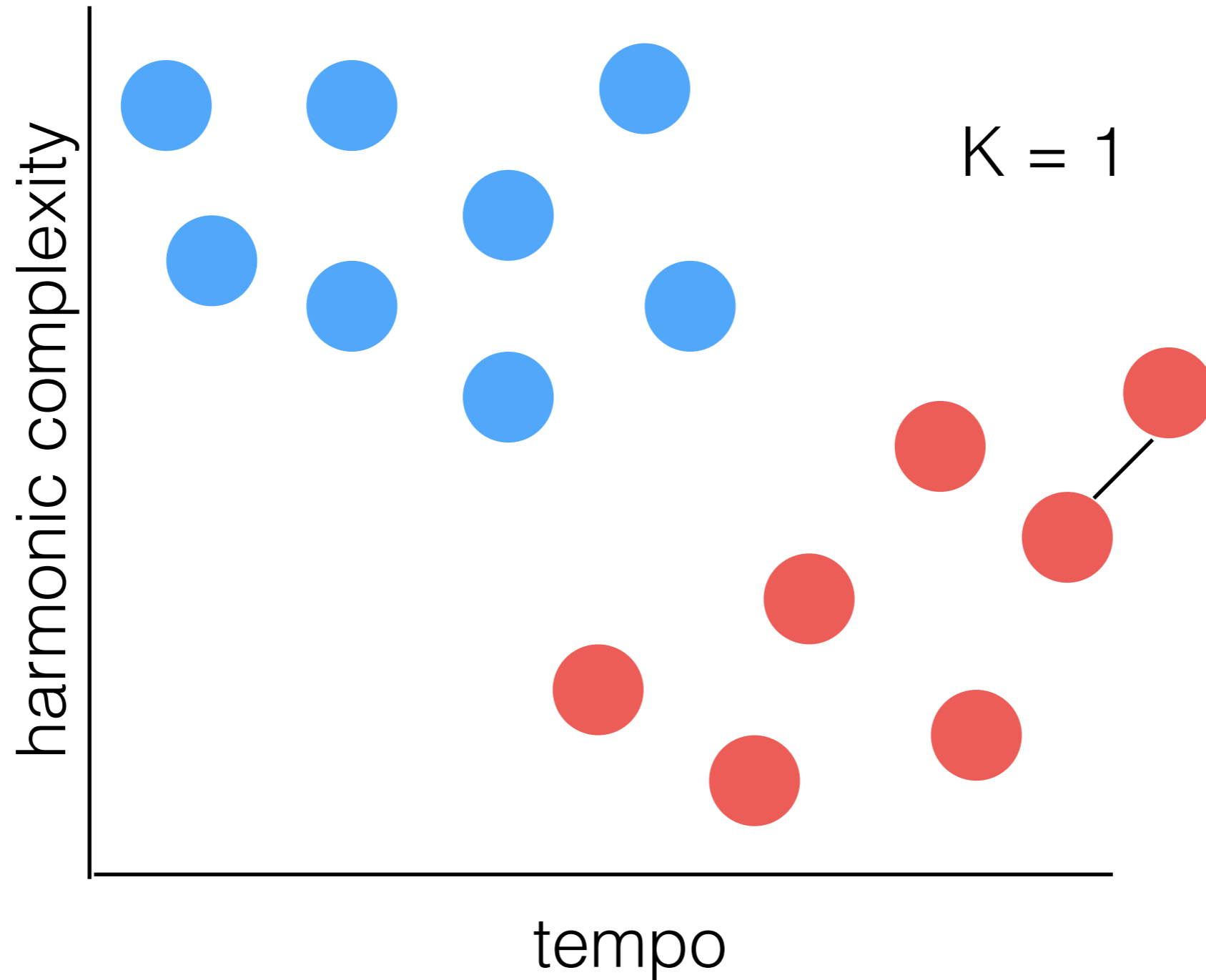
K Means



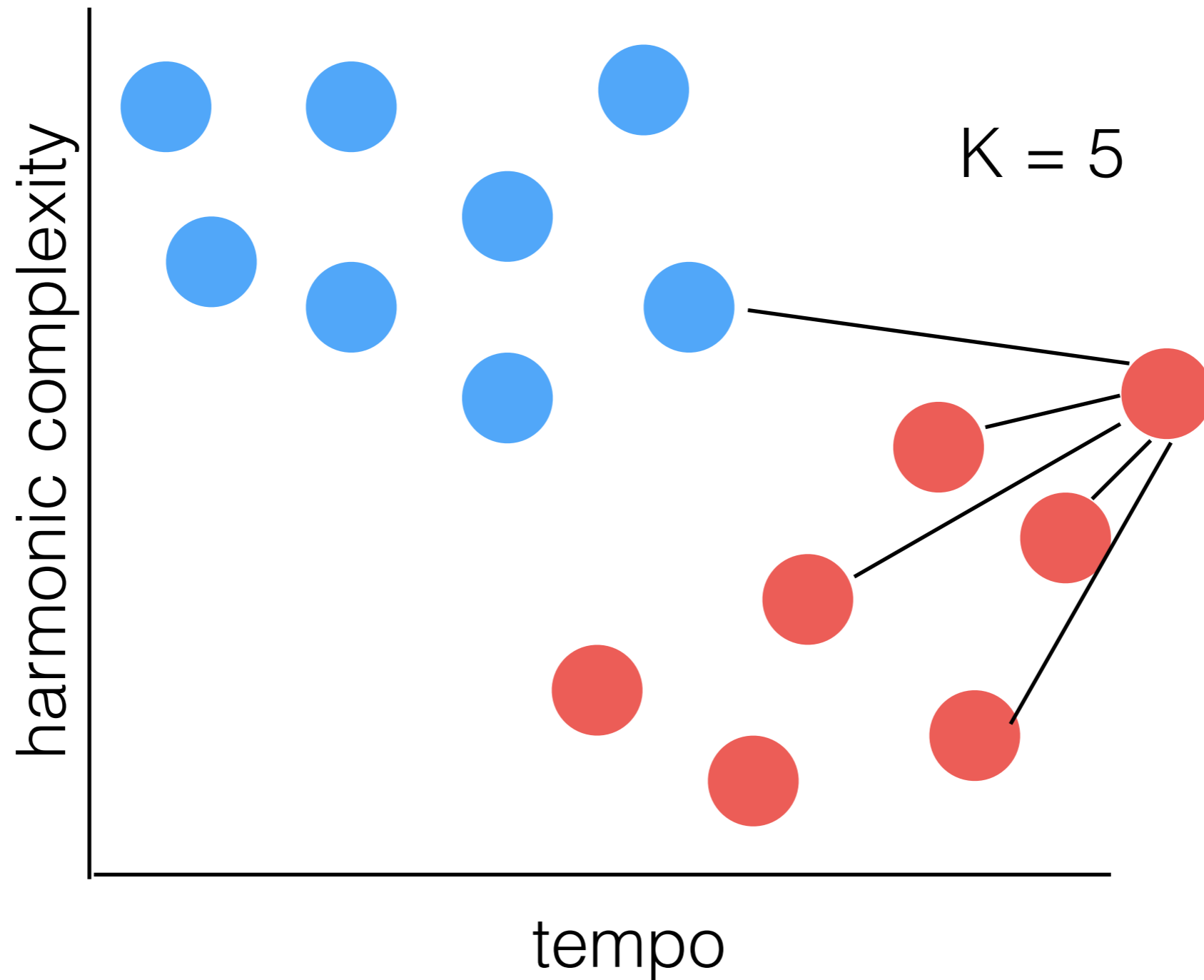
K Nearest Neighbors



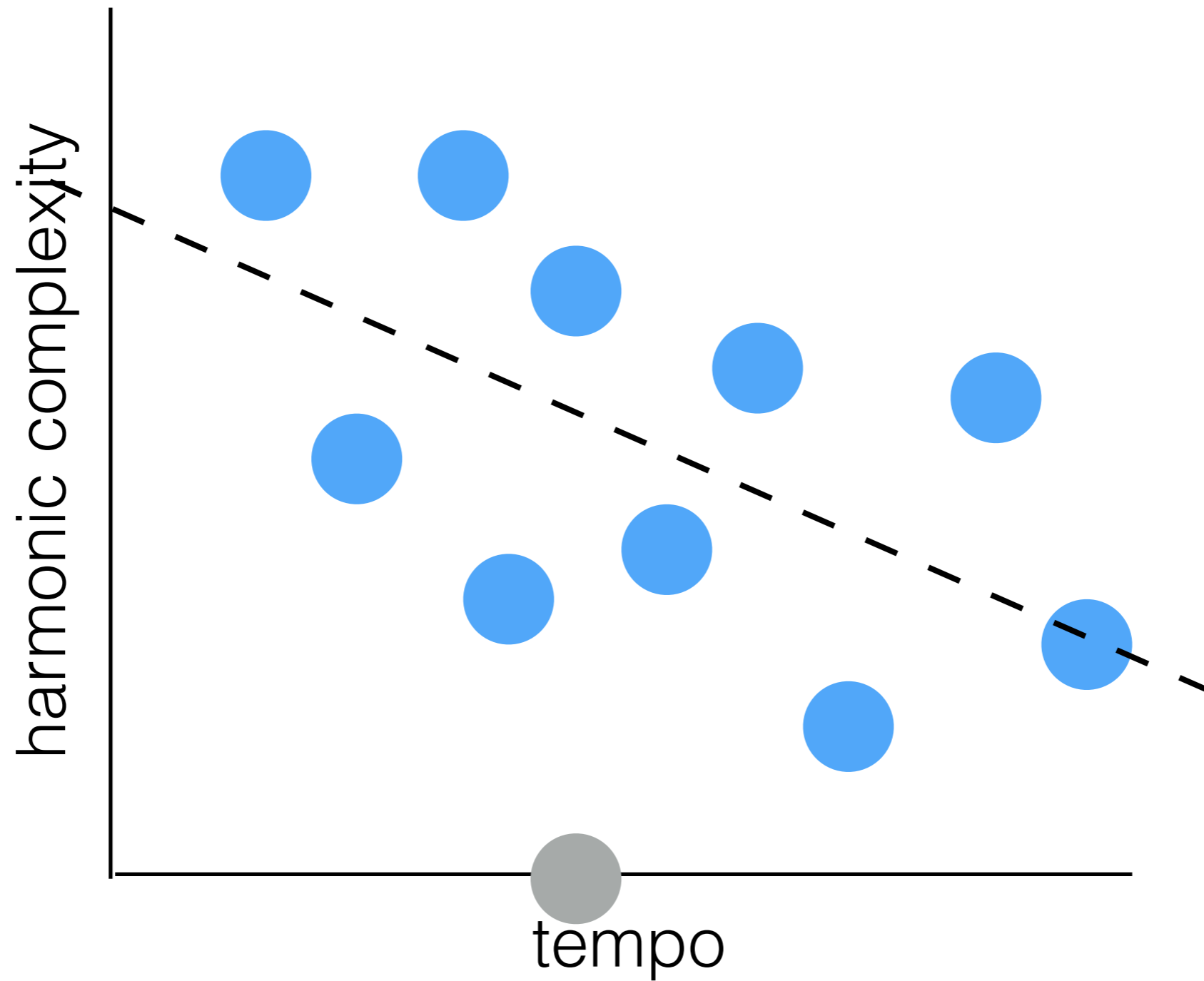
K Nearest Neighbors



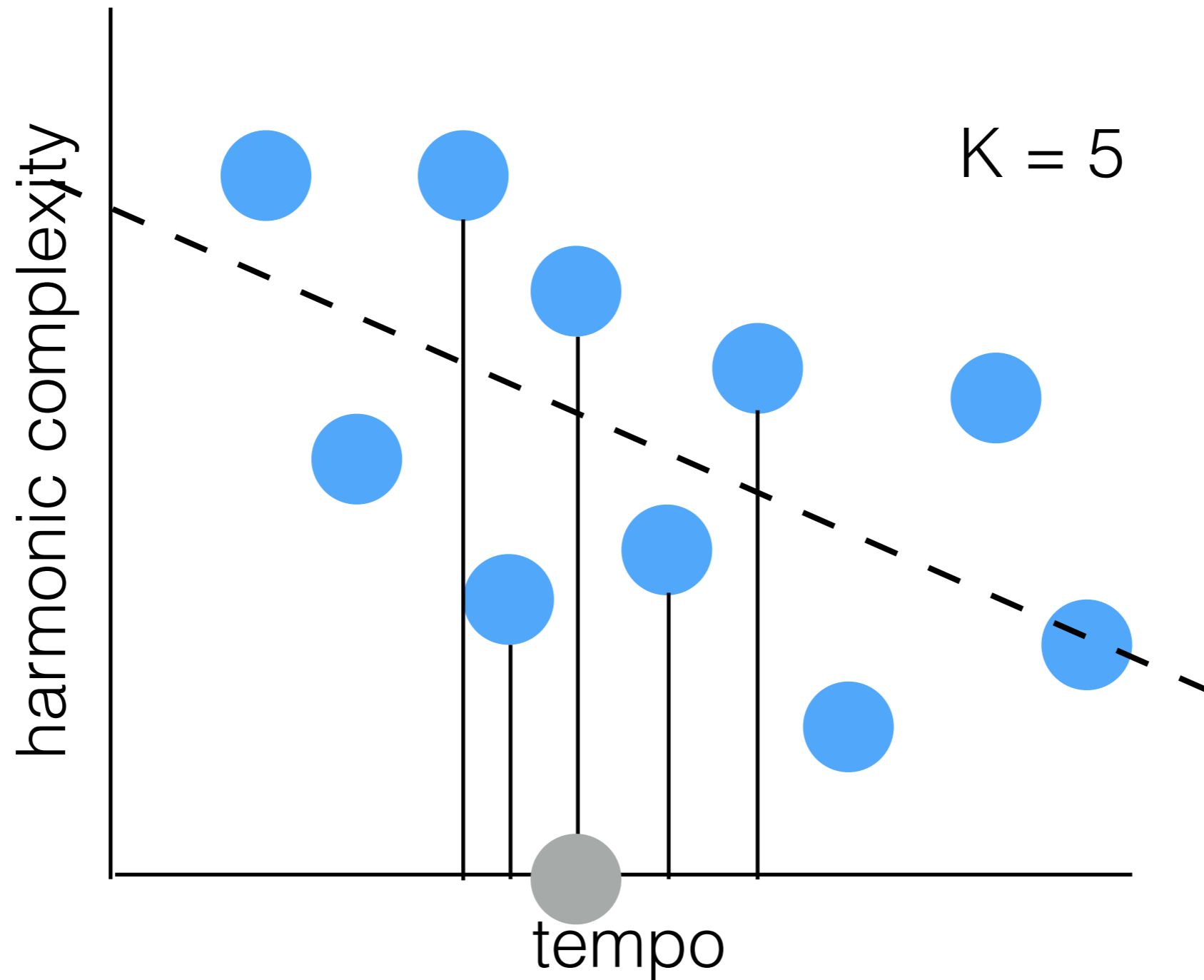
K Nearest Neighbors



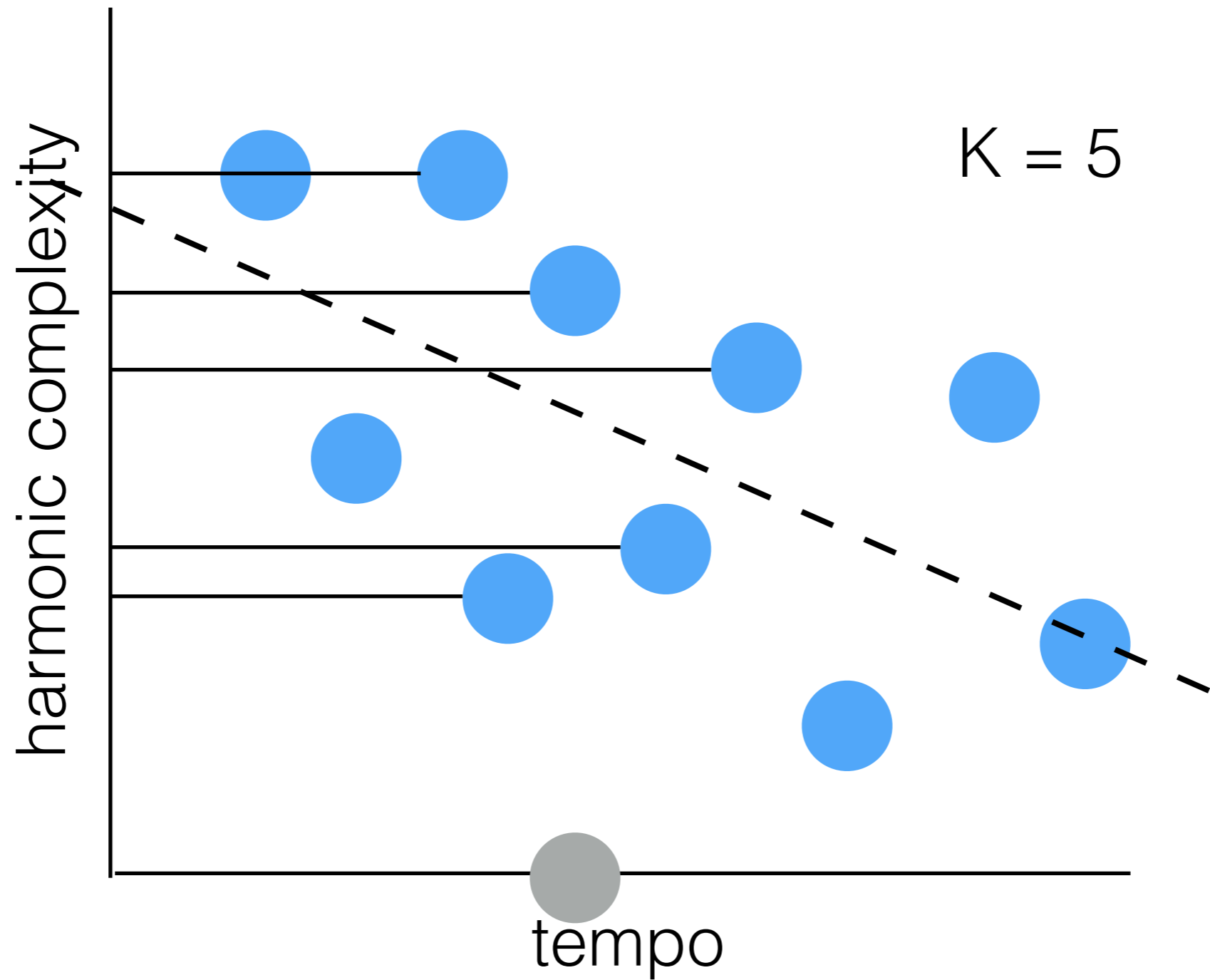
K Nearest Neighbors



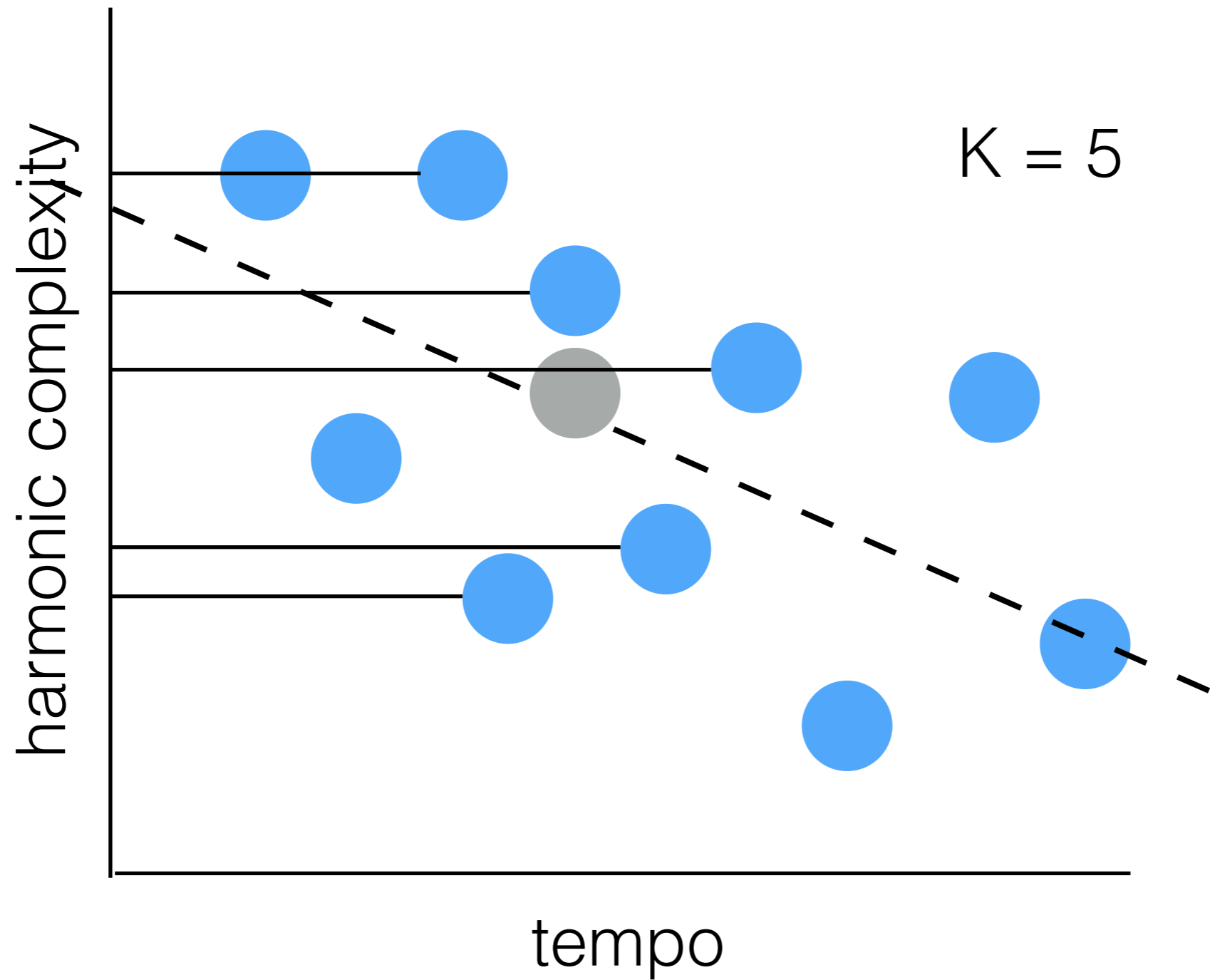
K Nearest Neighbors



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

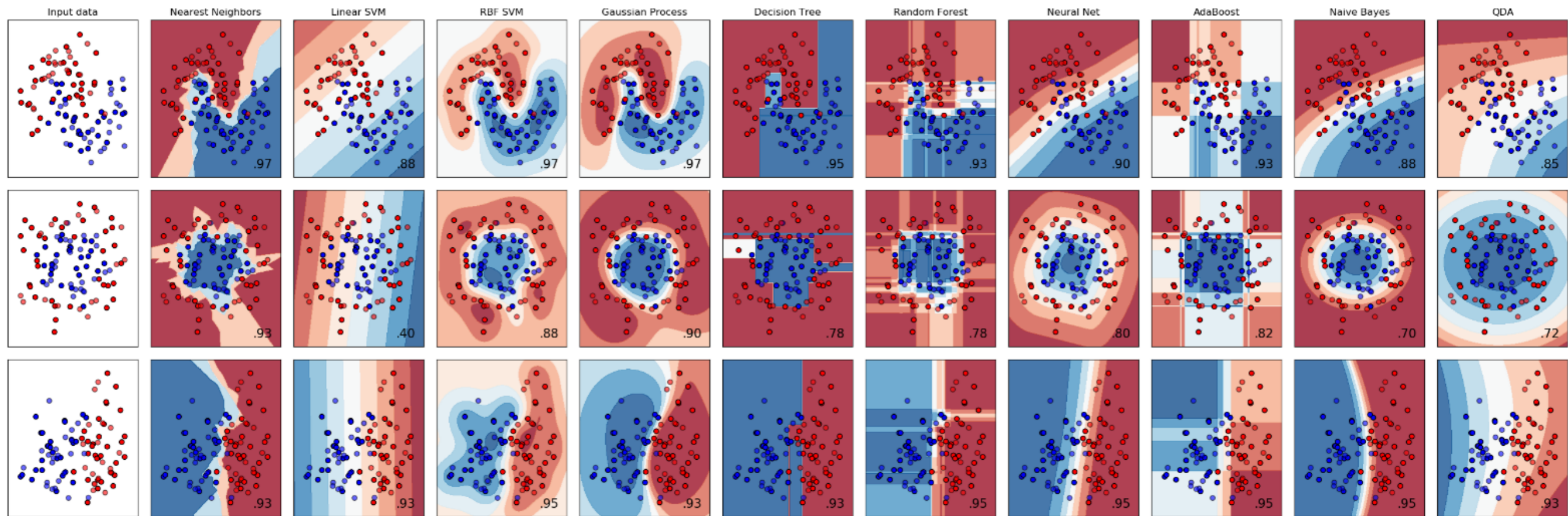
K Nearest Neighbors

- 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



Generative Models

Discriminative Models

Generative Models

Discriminative Models

estimate $P(X, Y)$ first

Generative Models

estimate $P(X, Y)$ first

Discriminative Models

estimate $P(Y | X)$ directly

Generative Models

estimate $P(X, Y)$ first

Discriminative Models

estimate $P(Y | X)$ directly
/no explicit probability model

Generative Models

estimate $P(X, Y)$ first

Can assign probability to observations, generate new observations

Discriminative Models

estimate $P(Y | X)$ directly
/no explicit probability model

Generative Models

estimate $P(X, Y)$ first

Can assign probability to observations, generate new observations

Discriminative Models

estimate $P(Y | X)$ directly
/no explicit probability model

Only supports classification,
less flexible

Generative Models

estimate $P(X, Y)$ first

Can assign probability to observations, generate new observations

Often more parameters, but more flexible

Discriminative Models

estimate $P(Y | X)$ directly
/no explicit probability model

Only supports classification, less flexible

Generative Models

estimate $P(X, Y)$ first

Can assign probability to observations, generate new observations

Often more parameters, but more flexible

Discriminative Models

estimate $P(Y | X)$ directly
/no explicit probability model

Only supports classification, less flexible

Often fewer parameters, better performance on small data

Generative Models

estimate $P(X, Y)$ first

Can assign probability to observations, generate new observations

Often more parameters, but more flexible

Naive Bayes, Bayes Nets, VAEs, GANs

Discriminative Models

estimate $P(Y | X)$ directly
/no explicit probability model

Only supports classification, less flexible

Often fewer parameters, better performance on small data

Generative Models

estimate $P(X, Y)$ first

Can assign probability to observations, generate new observations

Often more parameters, but more flexible

Naive Bayes, Bayes Nets, VAEs, GANs

Discriminative Models

estimate $P(Y | X)$ directly
/no explicit probability model

Only supports classification, less flexible

Often fewer parameters, better performance on small data

Logistic Regression, SVMs, Perceptrons

Generative Models

estimate $P(X, Y)$ first

Can assign probability to observations, generate new observations

Often more parameters, but more flexible

Naive Bayes, Bayes Nets, VAEs, GANs

Discriminative Models

estimate $P(Y | X)$ directly
/no explicit probability model

Only supports classification, less flexible

Often fewer parameters, better performance on small data

Logistic Regression, SVMs, Perceptrons

KNN

Generative Models

estimate $P(X, Y)$ first

Can assign probability to observations, generate new observations

Often more parameters, but more flexible

Naive Bayes, Bayes Nets, VAEs, GANs

Discriminative Models

estimate $P(Y | X)$ directly
/no explicit probability model

Only supports classification, less flexible

Often fewer parameters, better performance on small data

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

Lovely mushroomy nose and good length. 1

Super, 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

Super, 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	1	0	0	0	0	0	...
1	1	1	0	0	0	0	0	1	...
1	1	1	0	0	0	0	0	0	...
0	0	0	0	1	1	0	0	0	...

Lovely mushroomy nose and good length. 1

Super, 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	...

Lovely mushroomy nose and good length. 1

Super, 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	...

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

Bayes Rule

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

$$\begin{aligned} &P(Y=1|\text{lovely, good,} \dots) \\ &=P(\text{lovely, good,} \dots|Y=1)P(Y=1) \end{aligned}$$

Bayes Rule

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

$$\begin{aligned} &P(Y=1|\text{lovely, good,} \dots) \\ &= P(\text{lovely, good,} \dots | Y=1)P(Y=1) \\ &= P(Y=1, \text{lovely, good,} \dots) \end{aligned}$$

Bayes Rule

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

$$\begin{aligned} &P(Y=1|\text{lovely, good,} \dots) \\ &=P(\text{lovely, good,} \dots|Y=1)P(Y=1) \\ &=P(Y=1, \text{lovely, good,} \dots) \\ &=P(\text{lovely}|Y=1, \text{good,} \dots)P(Y=1, \text{good,} \dots) \end{aligned}$$

Bayes Rule

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

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

Naive Bayes

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

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

Assume features are independent!

Naive Bayes

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

$$\begin{aligned} P(C|x_1, x_2, \dots, x_k) \\ &= P(x_1|x_2, \dots, x_k, C)P(x_2|x_3, \dots, x_k, C)\dots P(x_k|C)P(C) \\ &= P(x_1|C)P(x_2|C)\dots P(x_k|C)P(C) \end{aligned}$$

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!

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

$P(x|Y=1)$

$P(x|Y=0)$

good

??

??

(a) 1.0, 0.0

(b) 1/2, 1/2

(c) 1/3, 1/3

Clicker Question!

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

$P(x|Y=1)$

$P(x|Y=0)$

good

1/3

1/3

(a) 1.0, 0.0

(b) 1/2, 1/2

(c) 1/3, 1/3

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

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

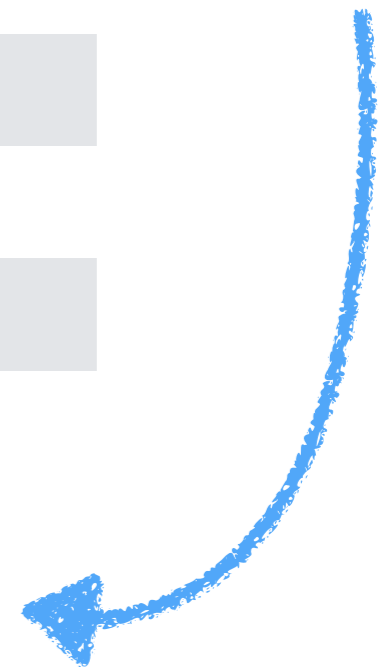
Naive Bayes

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
b	0.4	0.4
c	0.4	0.4
d	0.0	0.0
good	0.2	0.2
lovely	0.5	0.1
mushroomy	0.2	0.2
quite	0.7	0.8

Decision rule:

$\operatorname{argmax}_y P(Y=y|X)$

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

Clicker Question!

x	$P(x Y=1)$	$P(x Y=0)$		
a	0.9	0.9		
bit	0.2	0.4		
dramatic	0.6	0.4	$P(Y=1)$	$P(Y=0)$
gamy	0.1	0.0	0.3	0.7
good	0.2	0.2		
lovely	0.5	0.1		
mushroomy	0.2	0.2		
quite	0.7	0.8		

Quite mushroomy, a bit dramatic. ???

- (a) Positive
- (b) Negative

Clicker Question!

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=1)	P(Y=0)
0.3	0.7

Quite mushroomy, a bit dramatic. ???

$$0.9 \times 0.2 \times 0.6 \times 0.2 \times 0.7 \times 0.3 = 0.005$$

- (a) Positive
- (b) Negative

Clicker Question!

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=1)	P(Y=0)
0.3	0.7

Quite mushroomy, a bit dramatic. ???

$$0.9 \times 0.2 \times 0.6 \times 0.2 \times 0.7 \times 0.3 = 0.005$$

$$0.9 \times 0.4 \times 0.4 \times 0.2 \times 0.8 \times 0.7 = 0.016$$

- (a) Positive
- (b) Negative

Clicker Question!

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=1)	P(Y=0)
0.3	0.7

Quite mushroomy, a bit dramatic. ???

$$0.9 \times 0.2 \times 0.6 \times 0.2 \times 0.7 \times 0.3 = 0.005$$

$$0.9 \times 0.4 \times 0.4 \times 0.2 \times 0.8 \times 0.7 = 0.016$$

(a) Positive

(b) Negative

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

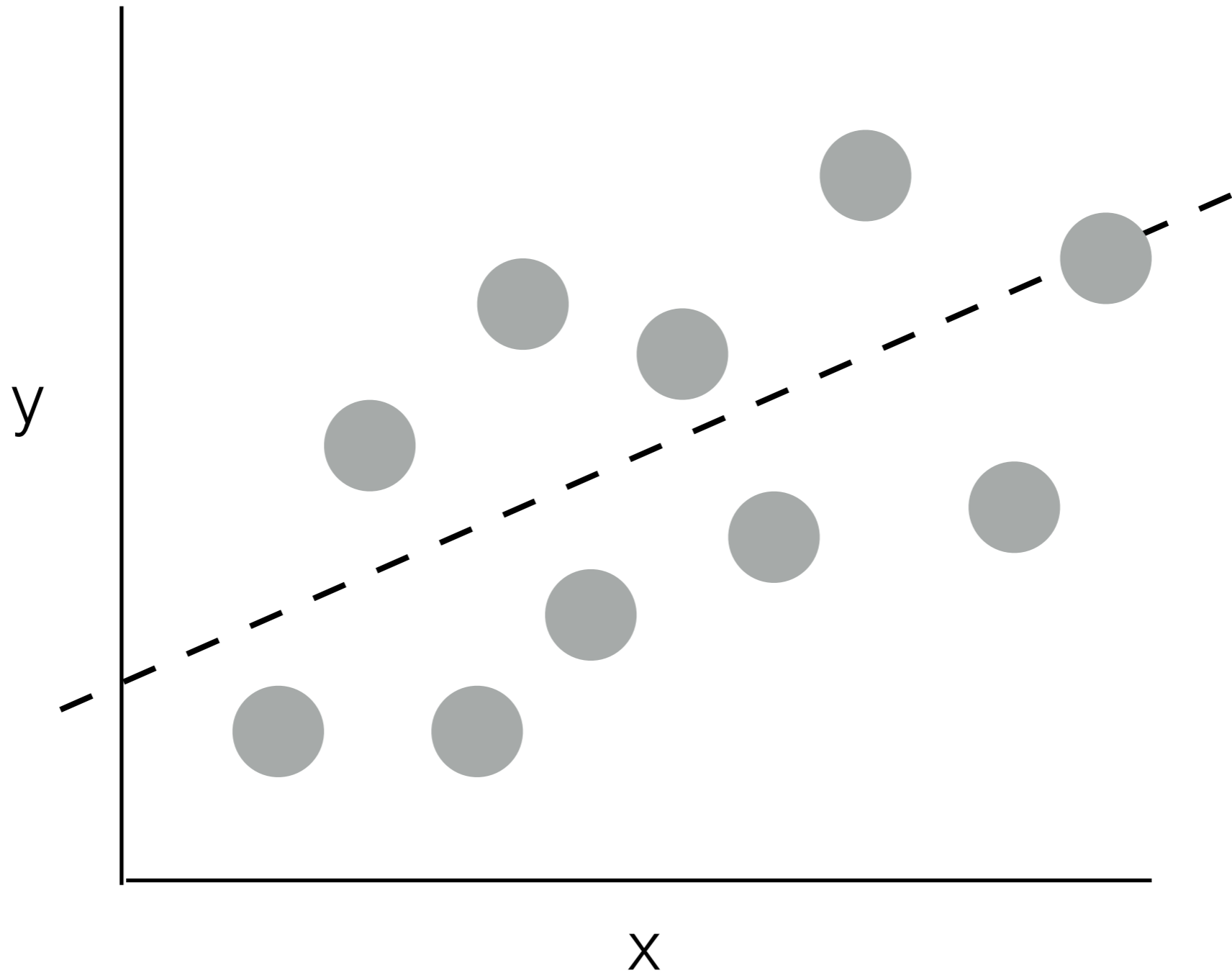


A quite dramatic gamy ... 0.04



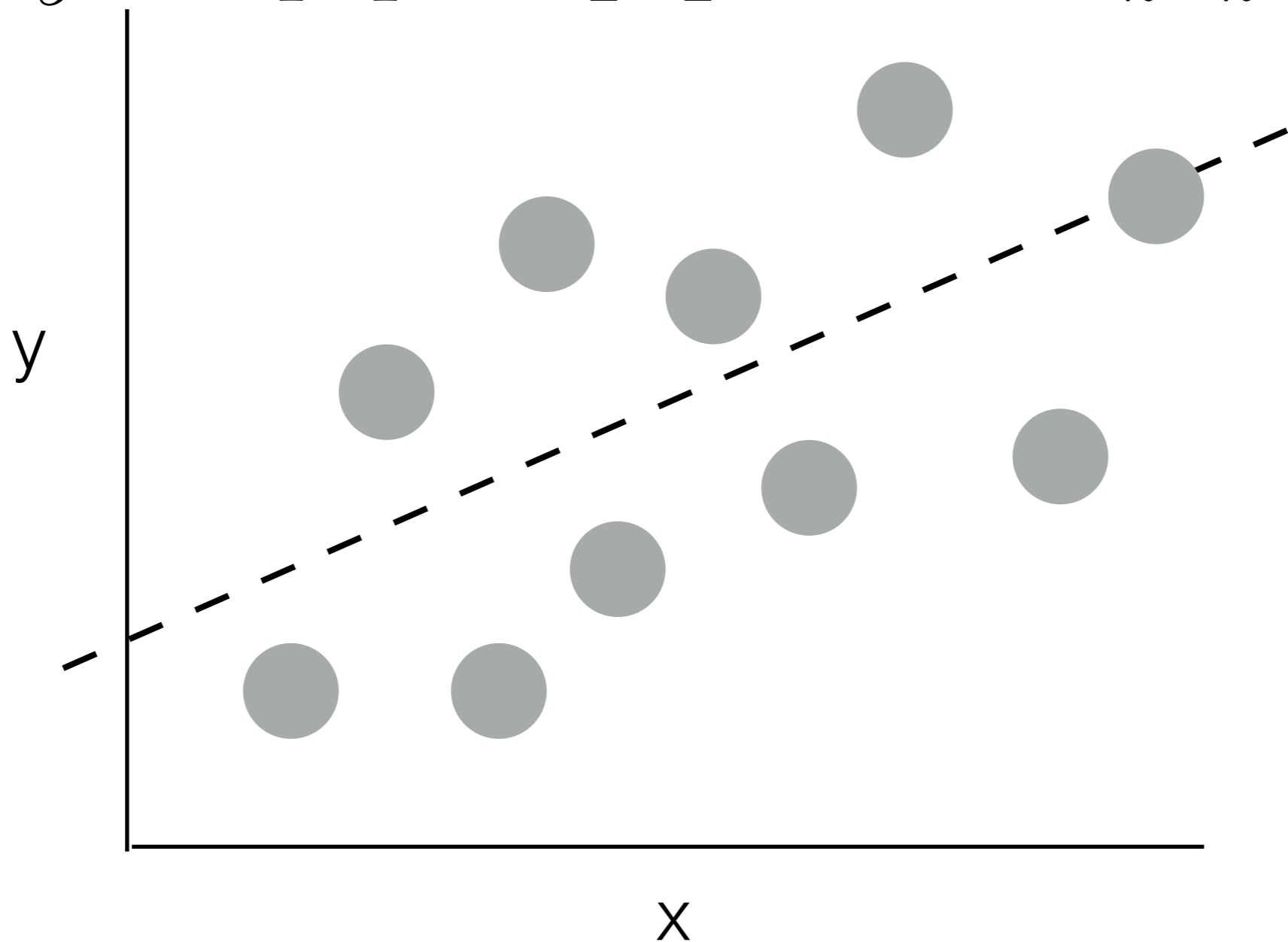
Linear Regression

Linear Regression



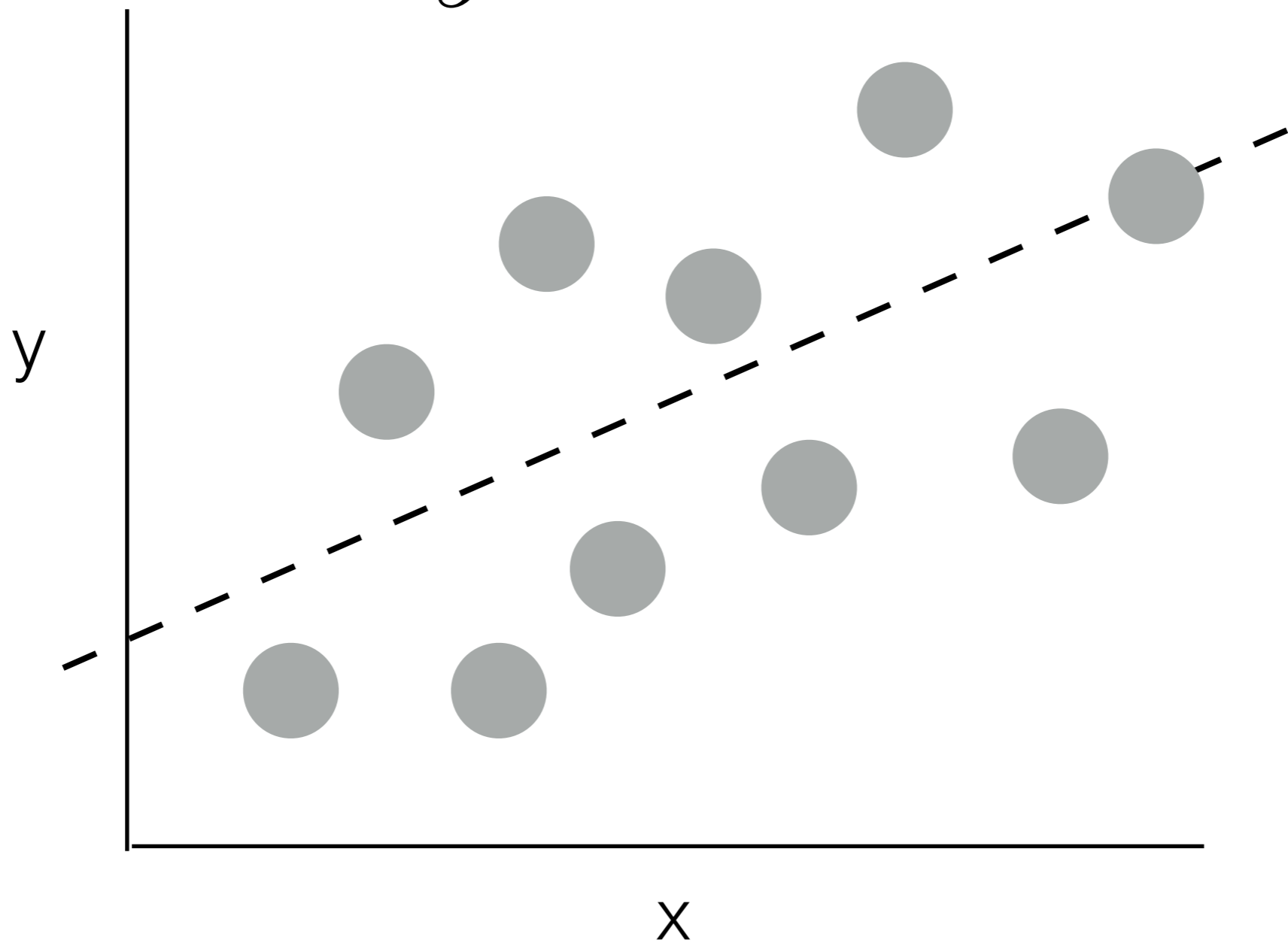
Linear Regression

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



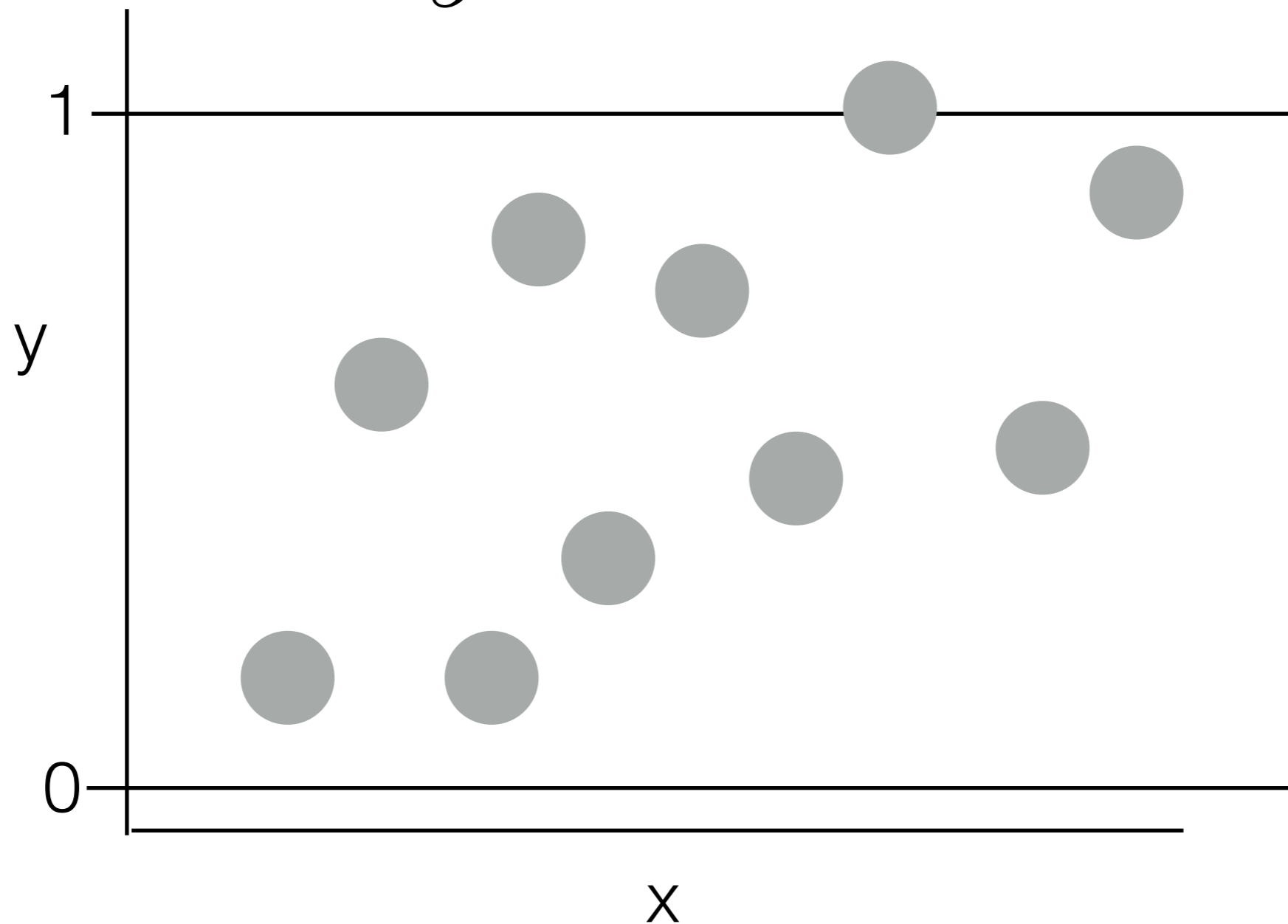
Linear Regression

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



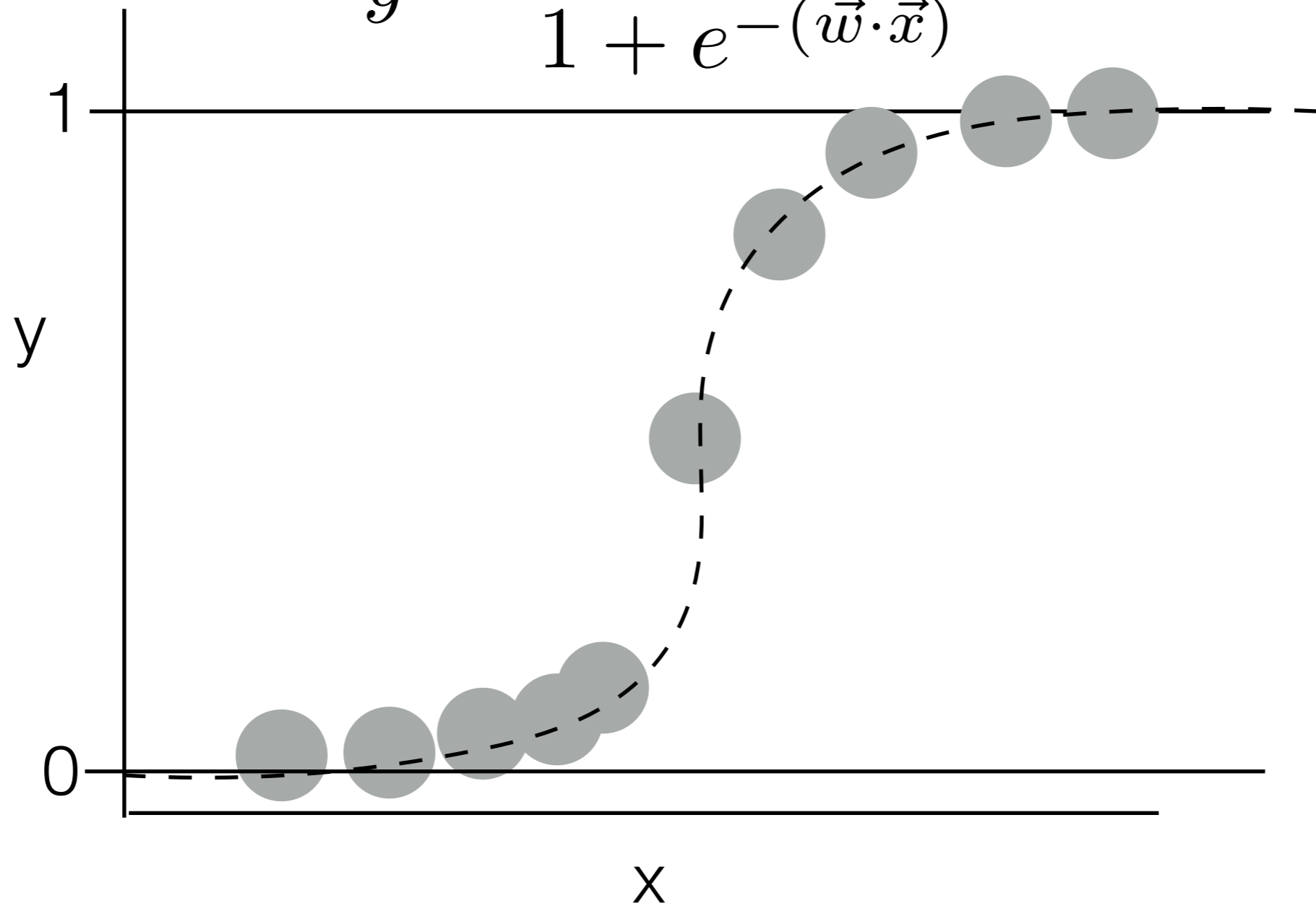
Linear Regression

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



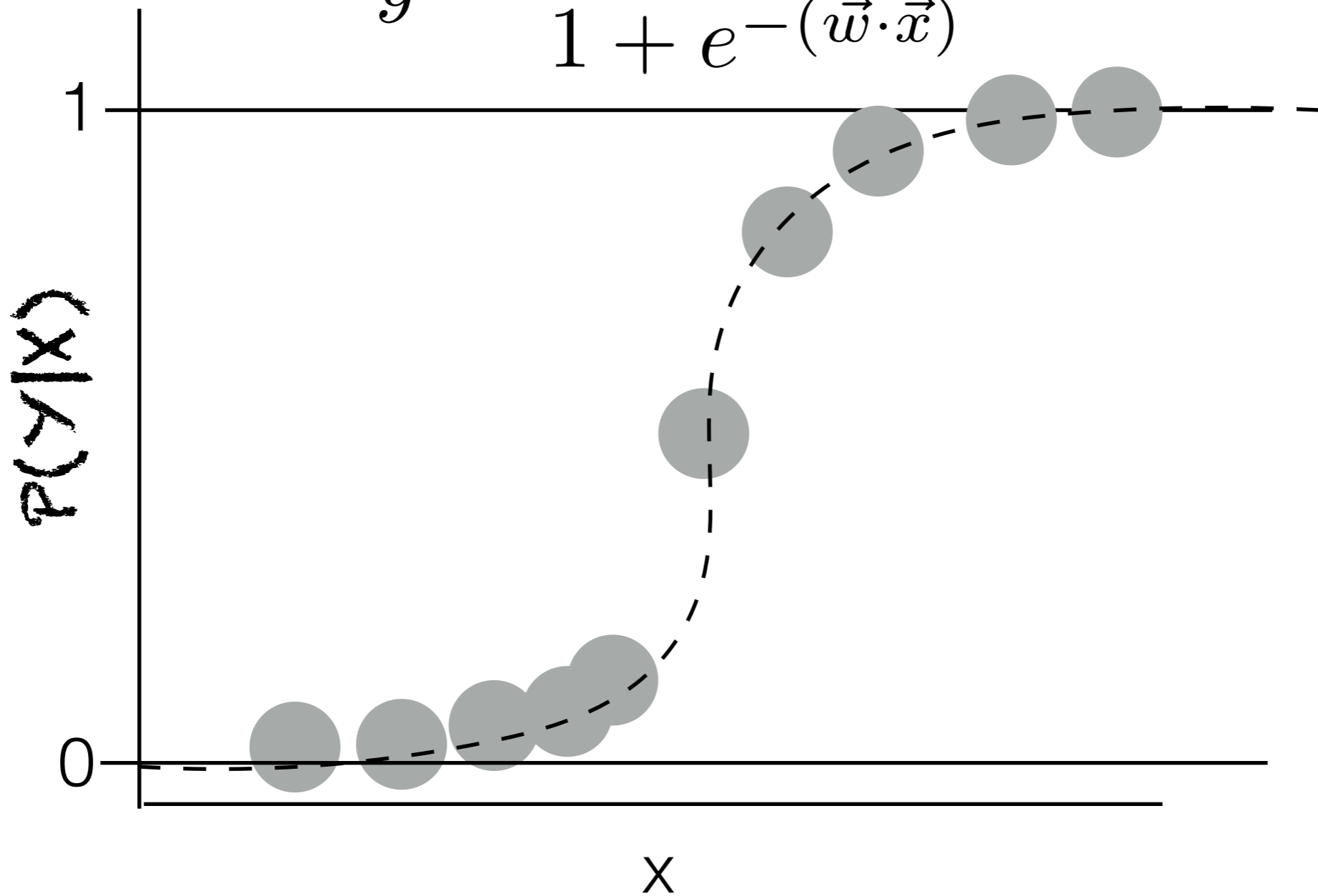
Logistic Regression

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



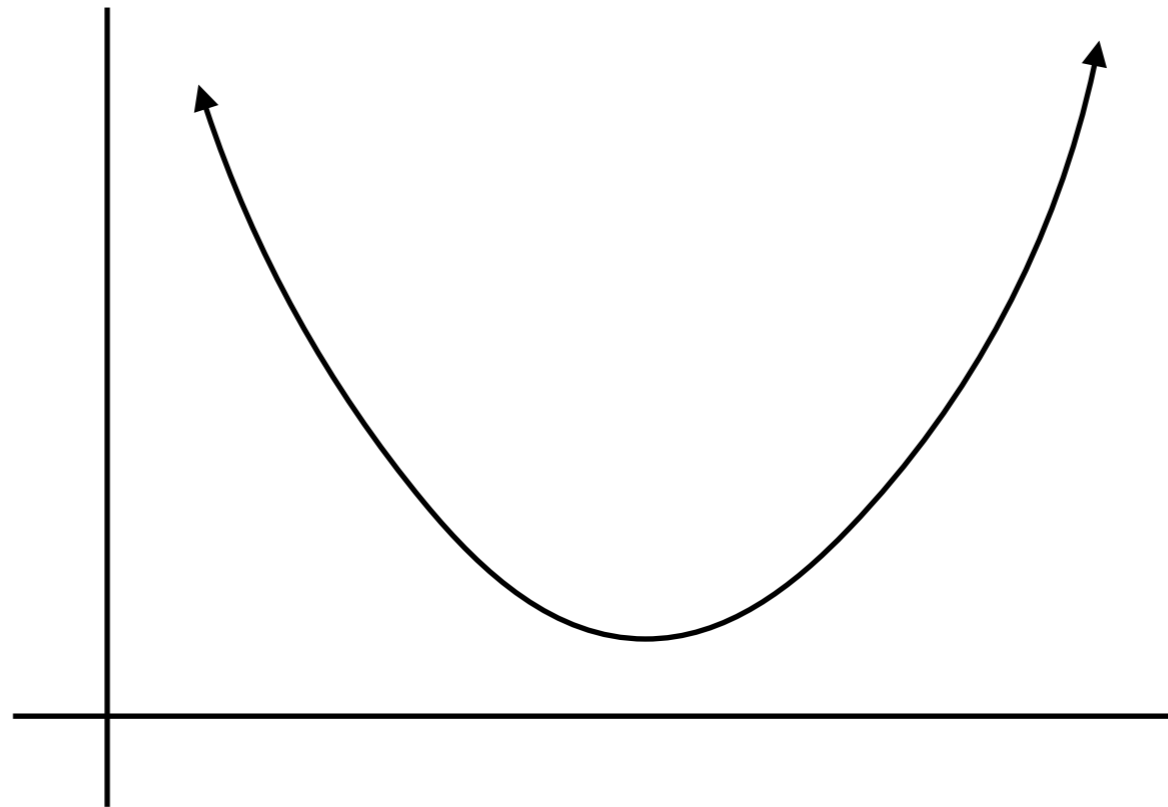
Logistic Regression

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



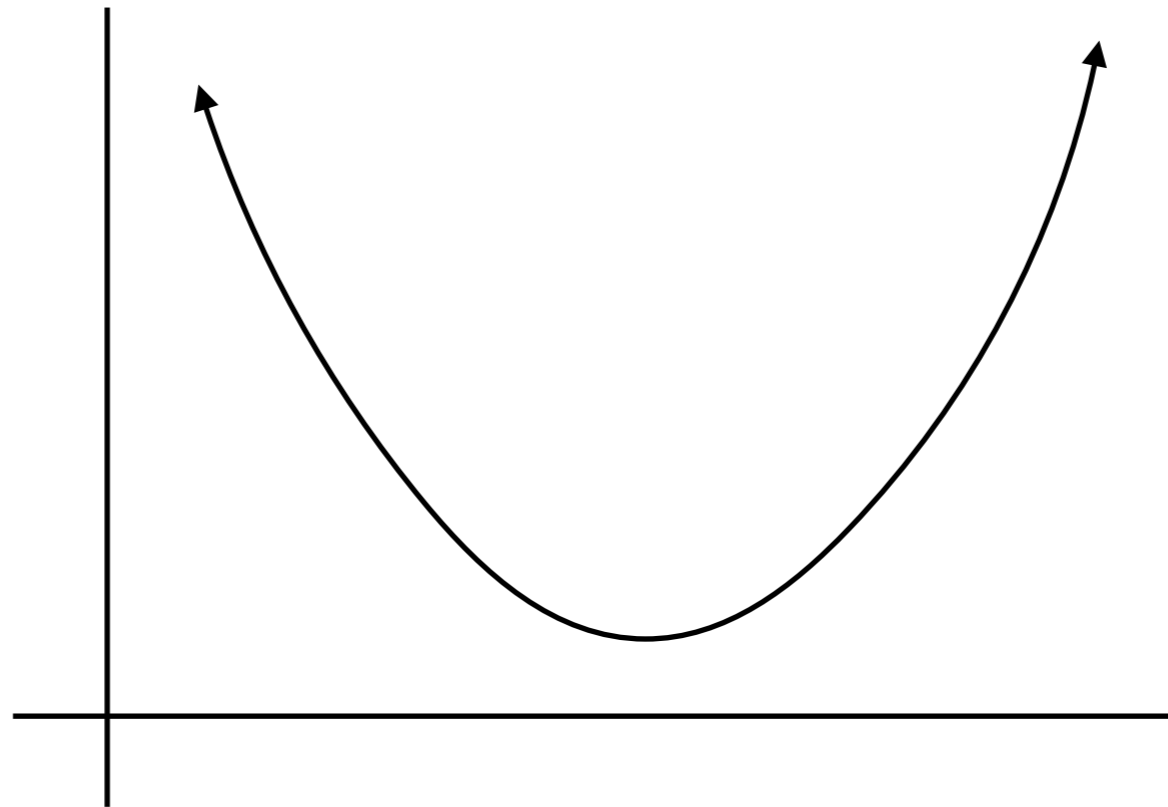
Linear Regression

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



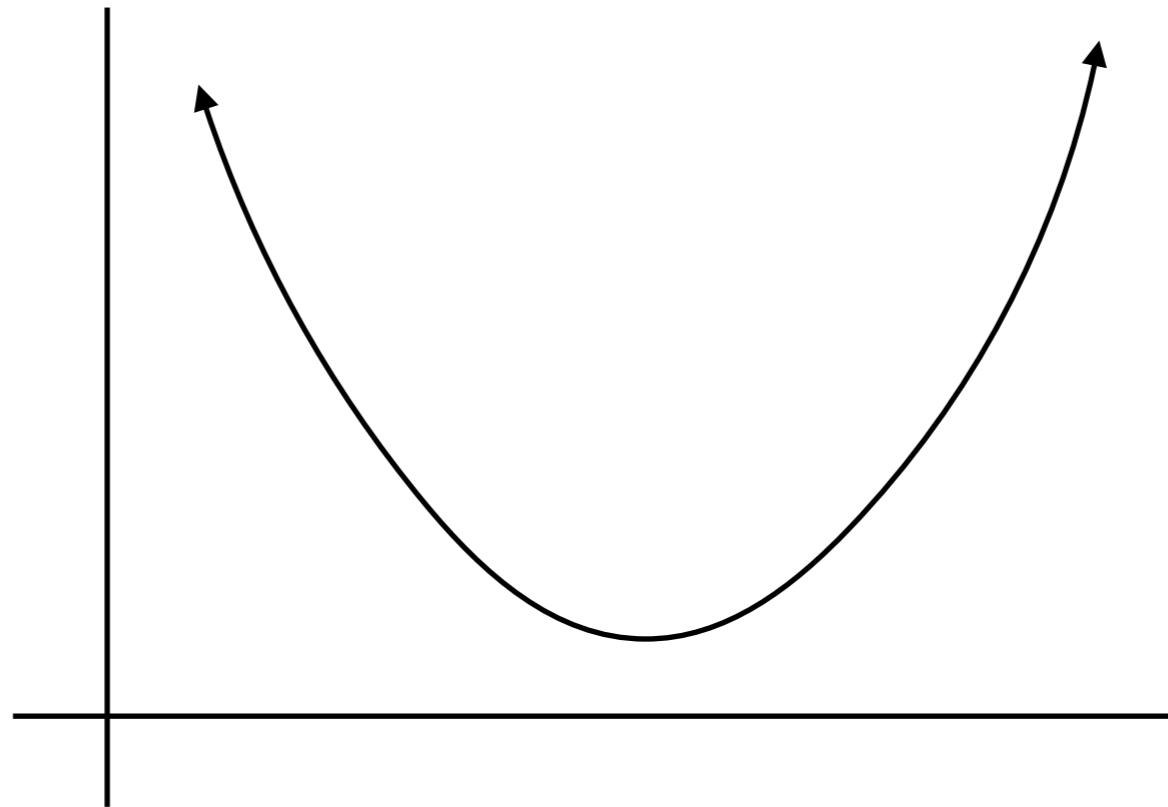
Logistic Regression

minimize $-\log P(Y|\hat{Y})$



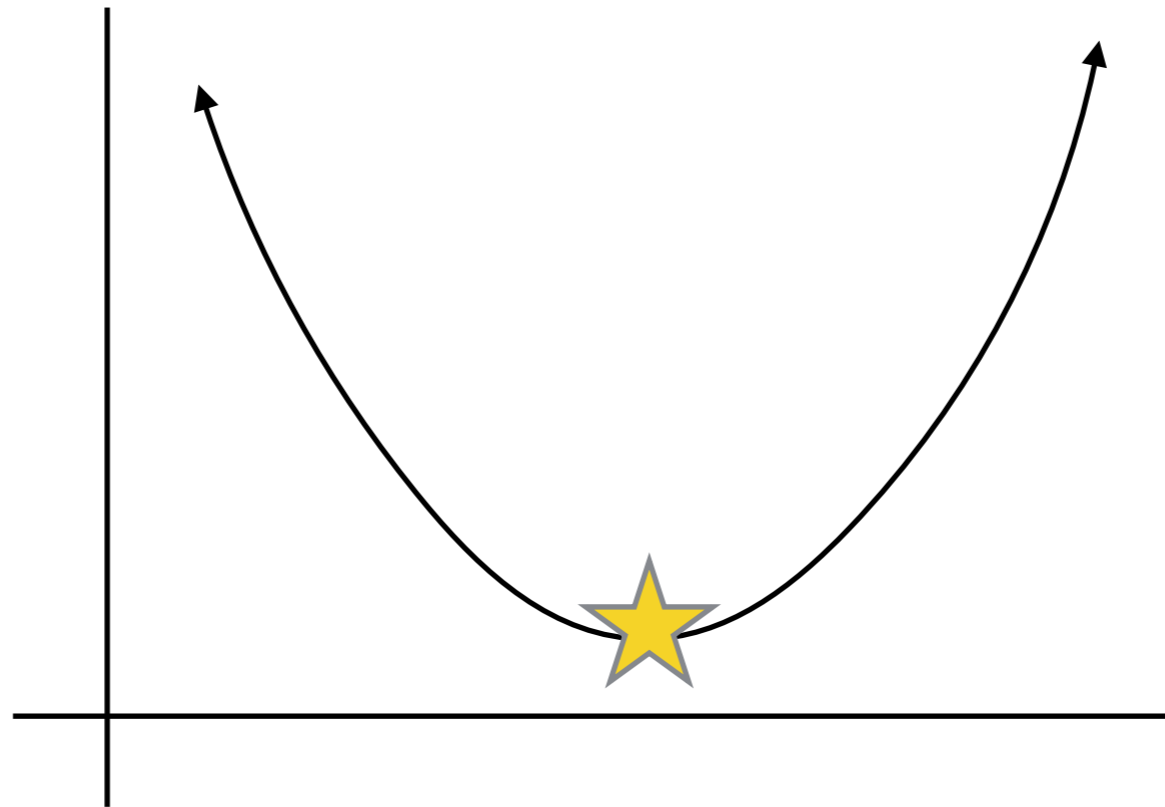
Logistic Regression

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



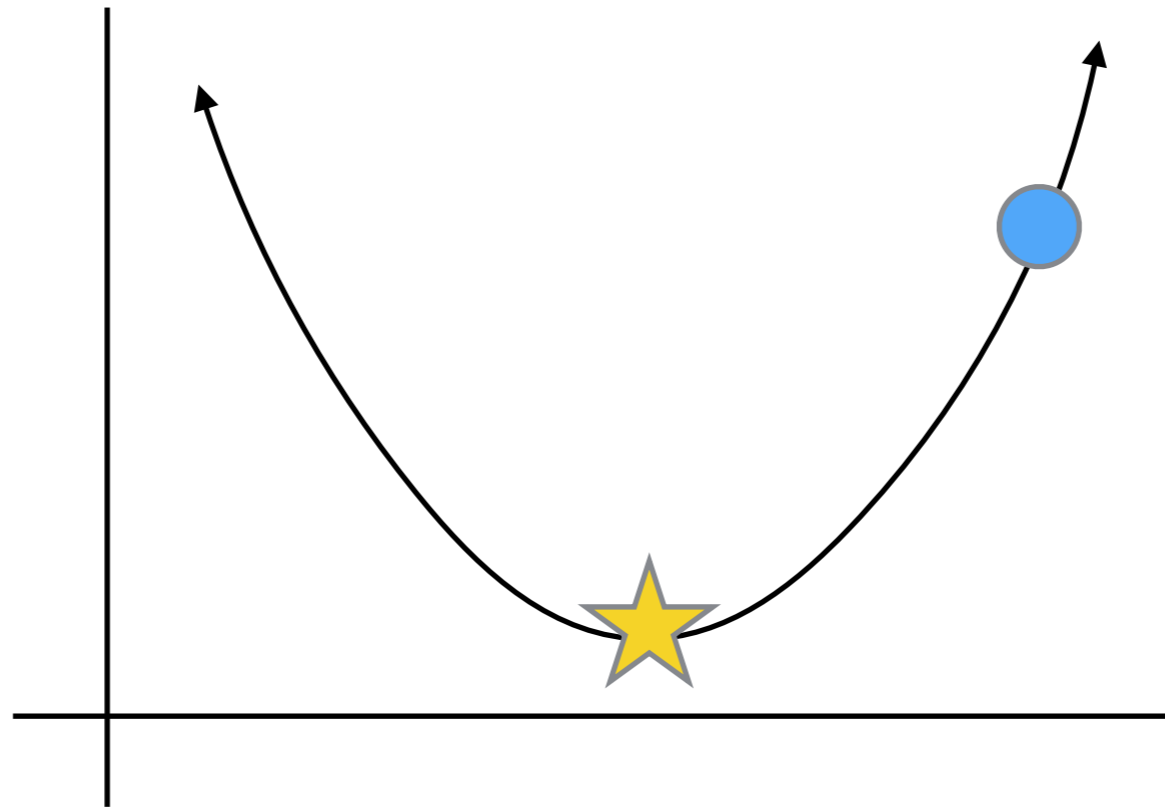
Logistic Regression

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



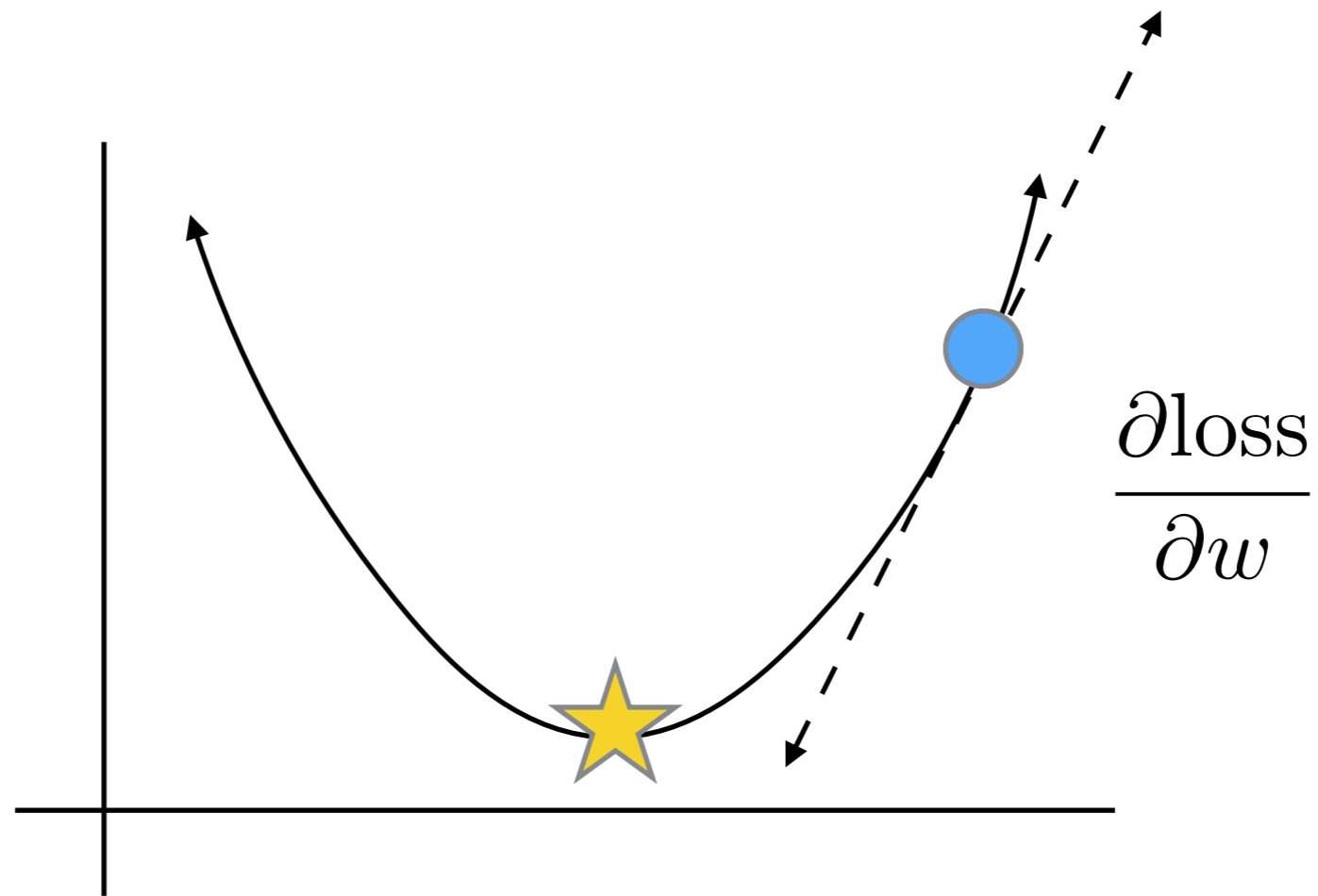
Logistic Regression

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



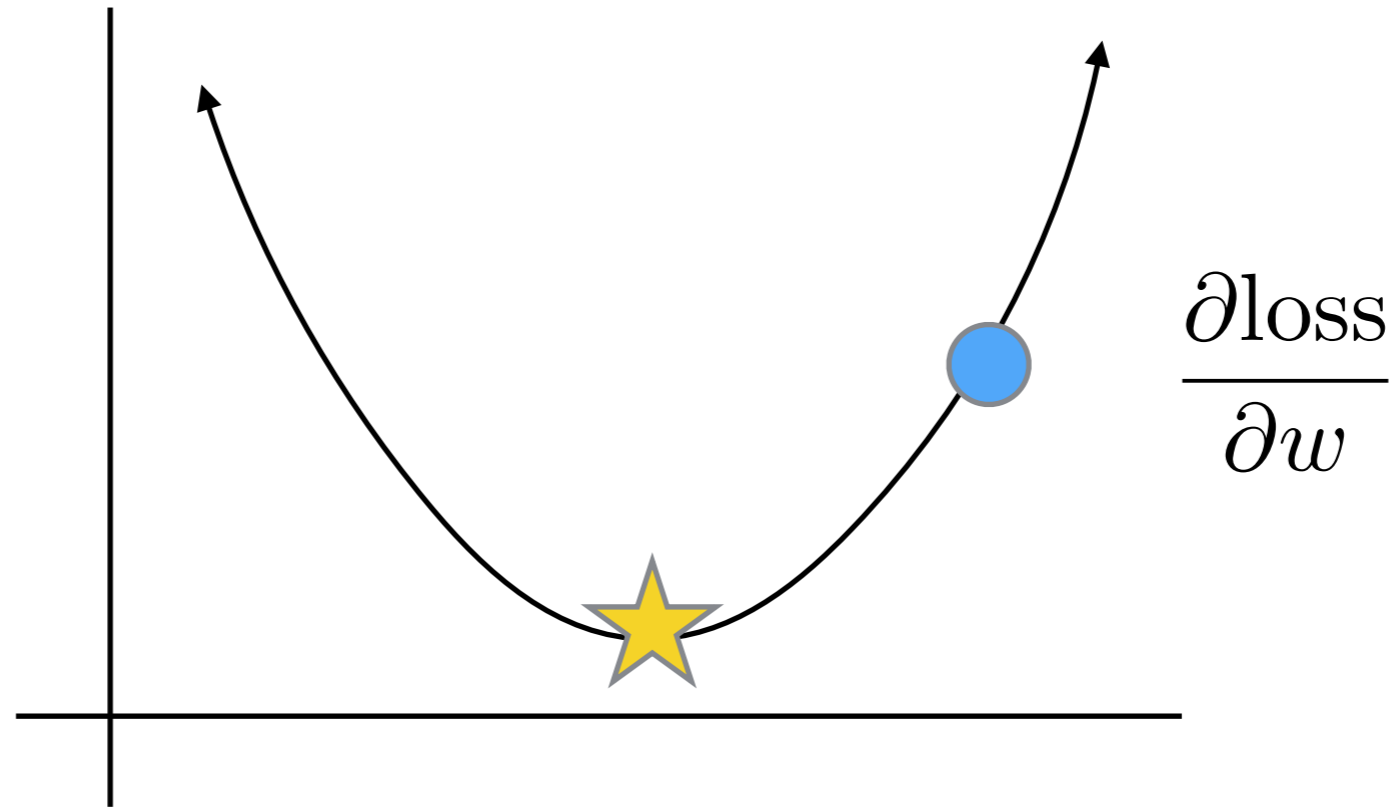
Logistic Regression

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



Logistic Regression

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



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!

WTF does this mean?

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

- (a) There is a 1.0 probability of observing “dramatic” given $Y = 1$
- (b) There is a 1.0 probability that $Y = 1$ given we observe “dramatic”
- (c) 1 is the co-efficient on the “dramatic” variable in the best fit linear regression.
- (d) 1 is the co-efficient on the “dramatic” variable in linear regression that minimizes the log loss.

Clicker Question!

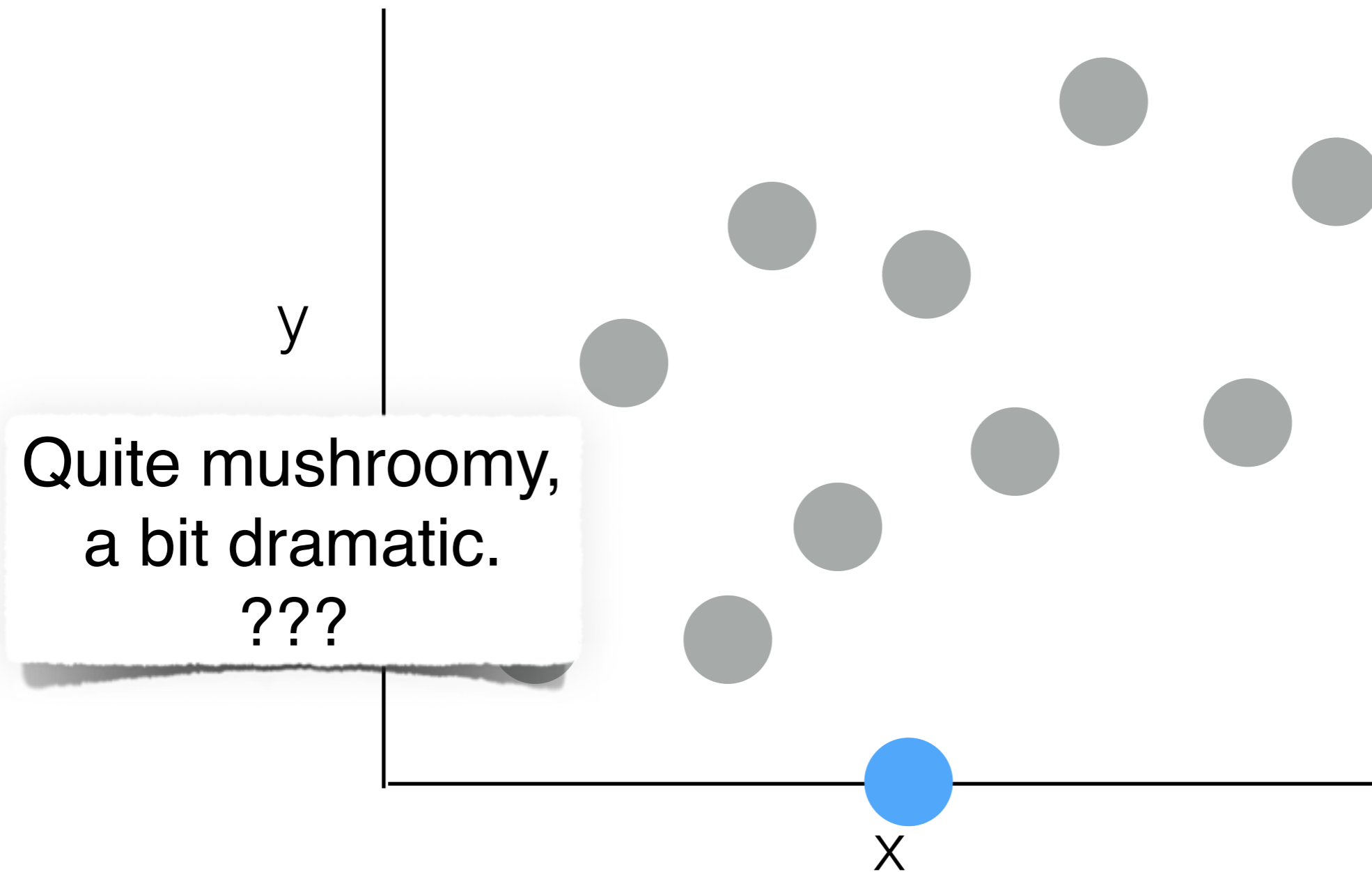
WTF does this mean?

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

- (a) There is a 1.0 probability of observing “dramatic” given $Y = 1$
- (b) There is a 1.0 probability that $Y = 1$ given we observe “dramatic”
- (c) 1 is the co-efficient on the “dramatic” variable in the best fit linear regression.
- (d) 1 is the co-efficient on the “dramatic” variable in linear regression that minimizes the log loss.

Logistic Regression

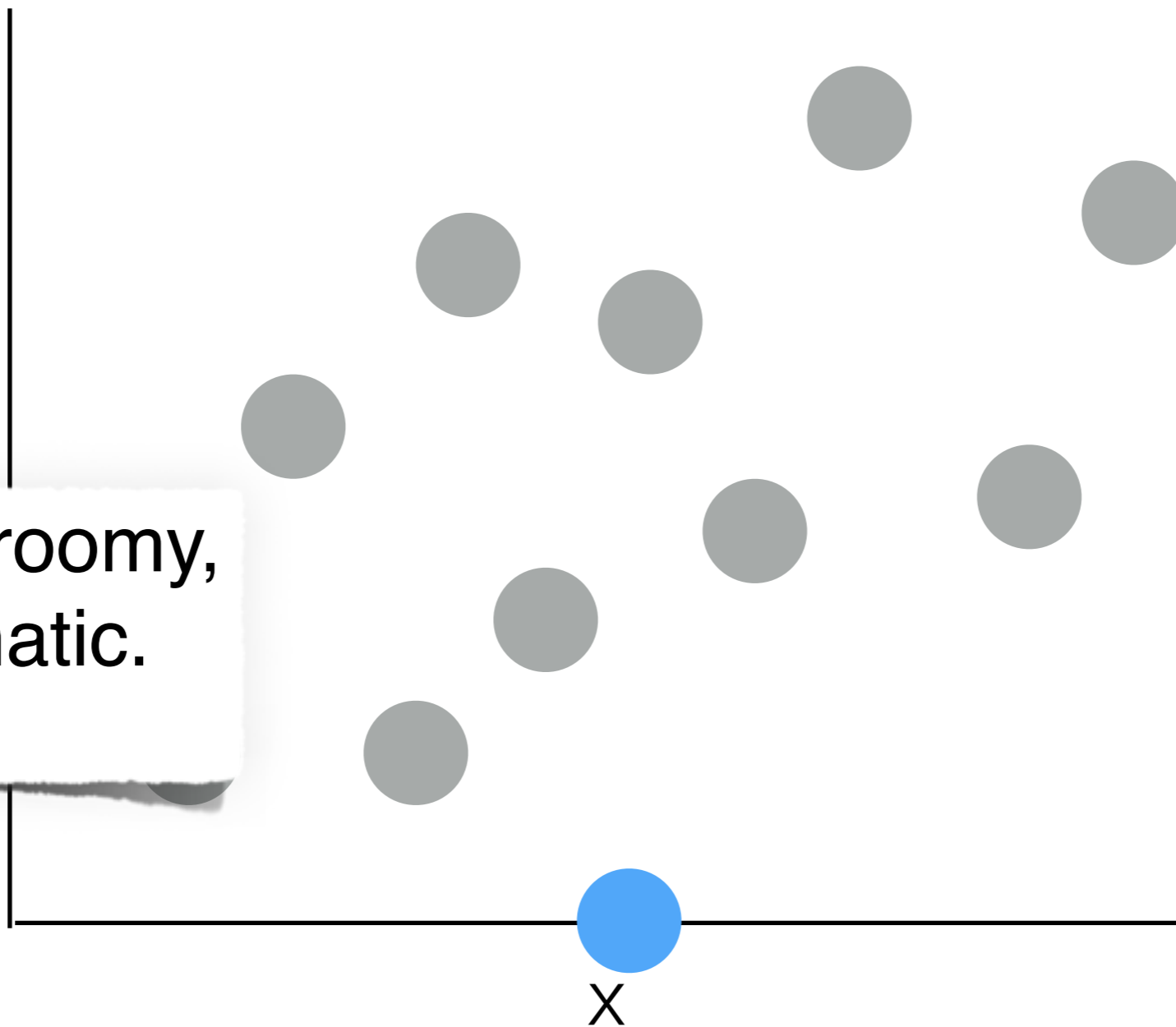




Logistic Regression

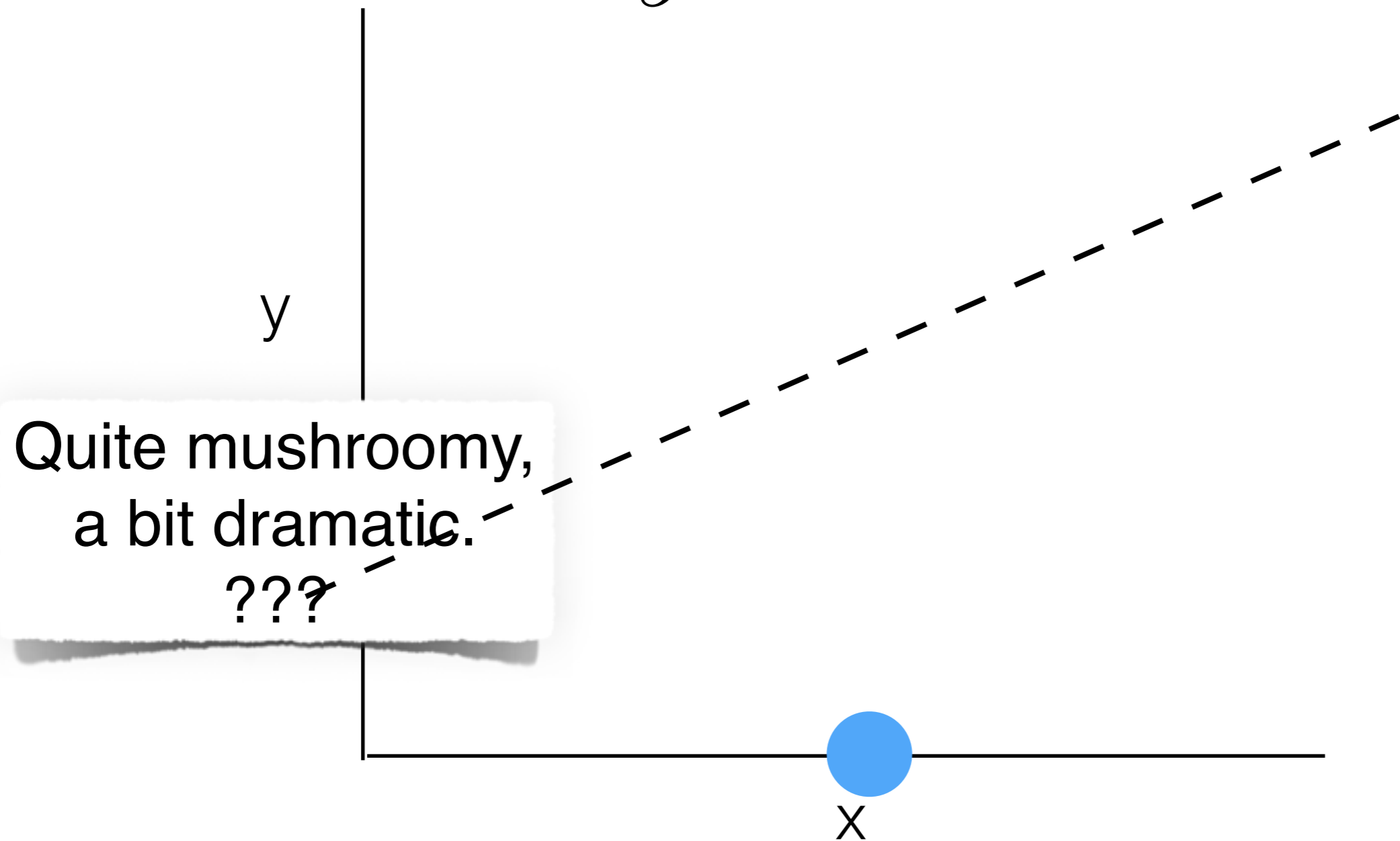
**What do we
do now?**

Quite mushroomy,
a bit dramatic.
???



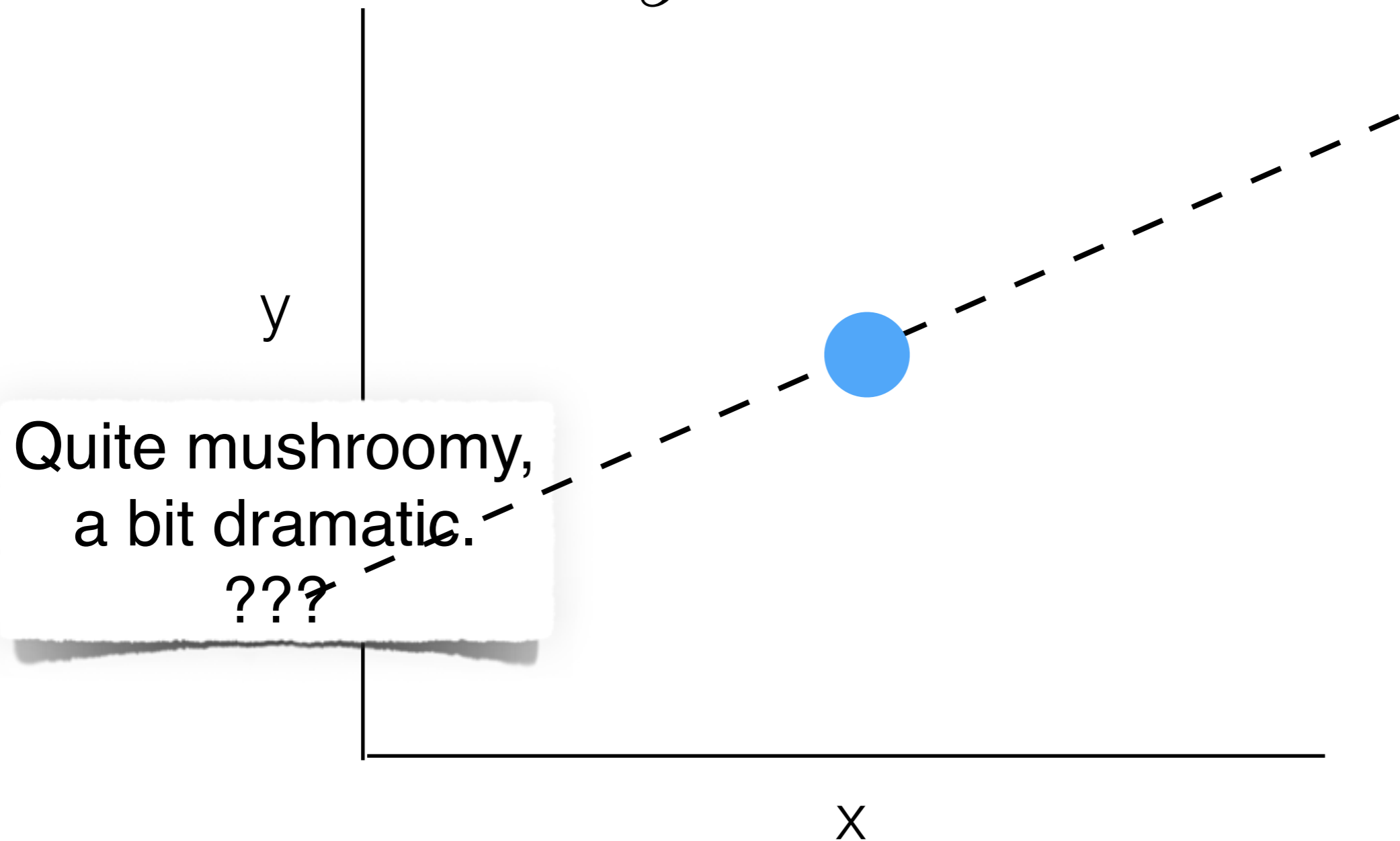
Logistic Regression

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



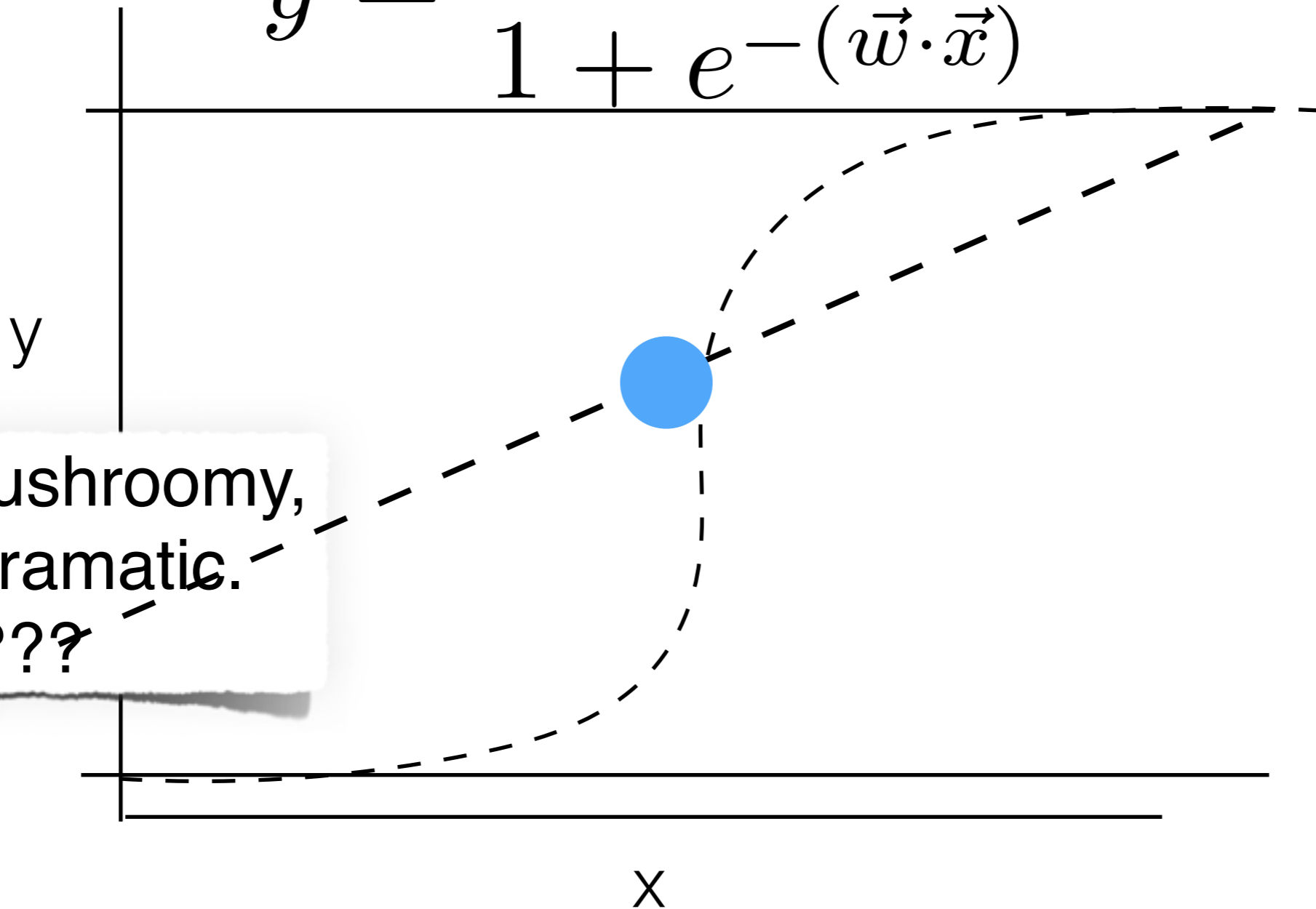
Logistic Regression

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



Logistic Regression

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



Quite mushy,
a bit dramatic.
???

Logistic Regression

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

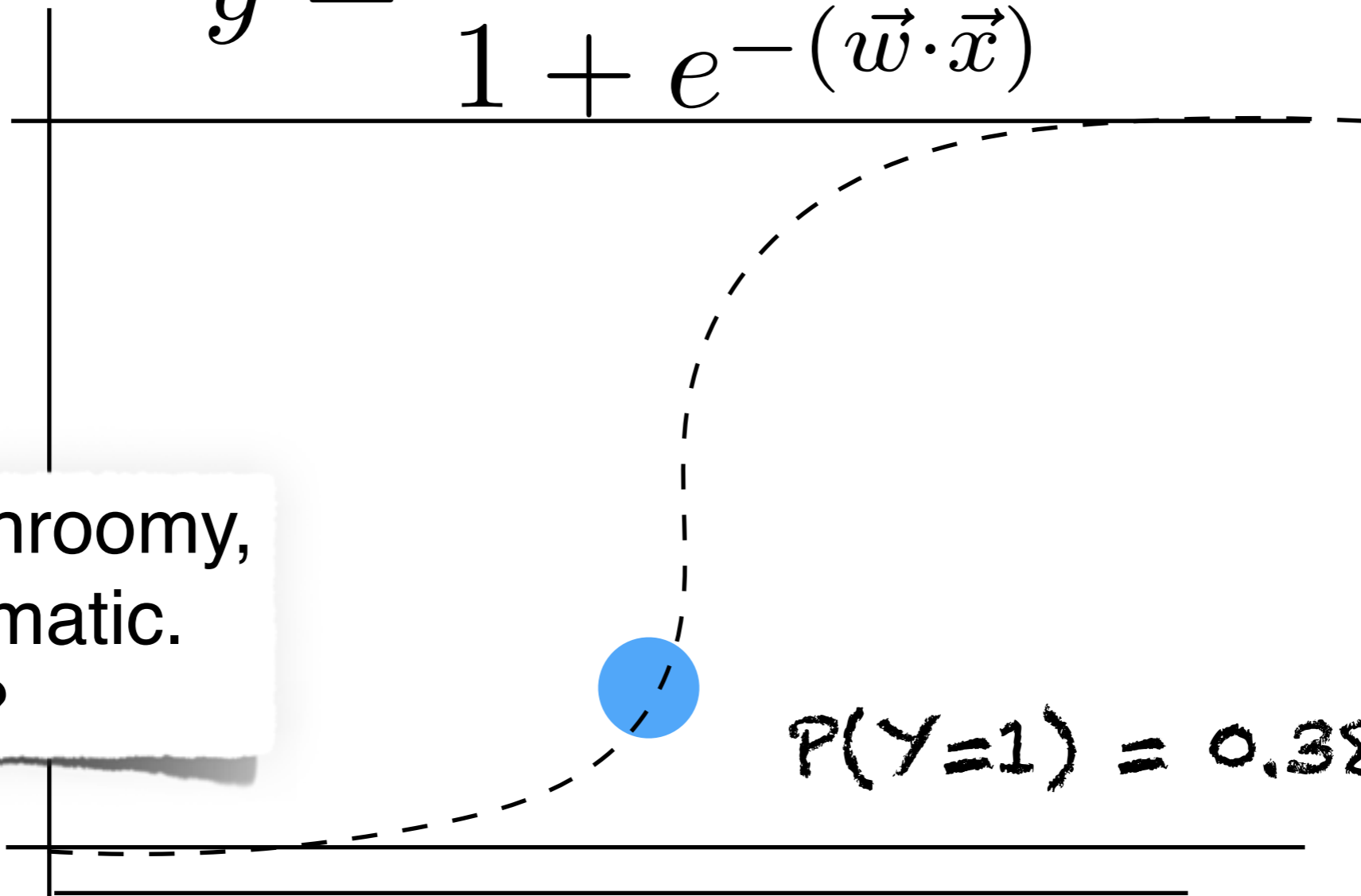
y

Quite mushy,
a bit dramatic.

???

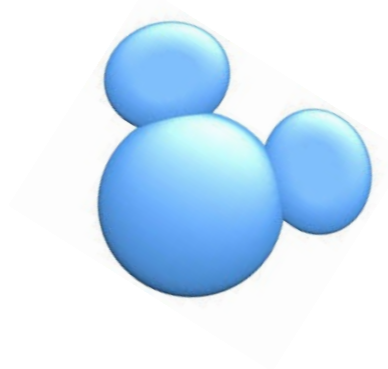
$$P(Y=1) = 0.38$$

x





Code-along!



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