COMP24111 lecture 2

Rugby players, Ballet dancers,

and the

Nearest Neighbour Classifier

A Problem to Solve with Machine Learning

Distinguish rugby players from ballet dancers.

You are provided with a few examples.

Fallowfield rugby club (16).
Rusholme ballet troupe (10).

Task
Generate a program which will correctly classify ANY player/dancer in the world.

Hint
We shouldn't "fine-tune" our system too much so it only works on the local clubs.

Taking measurements….

We have to process the people with a computer, so it needs to be in a computer-readable form.

What are the distinguishing characteristics?

1. *Height*
2. *Weight*
3. *Shoe size*
4. *Sex*

Terminology

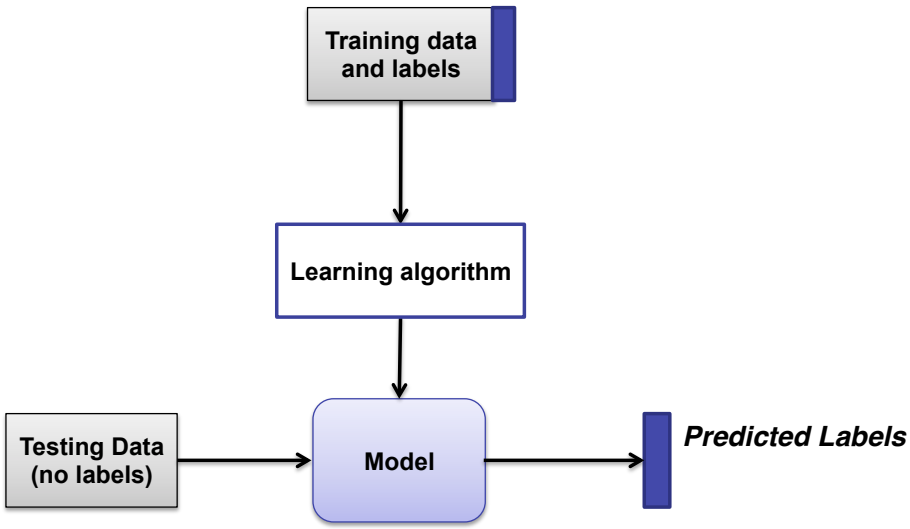
"Examples"

id	height	weight	shoe size	sex	Ballet?
1	70	64	3	1	1
2	23	86	5	0	1
3	56	49	5	1	0
4	50	88	3	0	0
5	12	50	1	0	1
6	56	66	2	1	0
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
N	56	1	5	0	0

"Features"

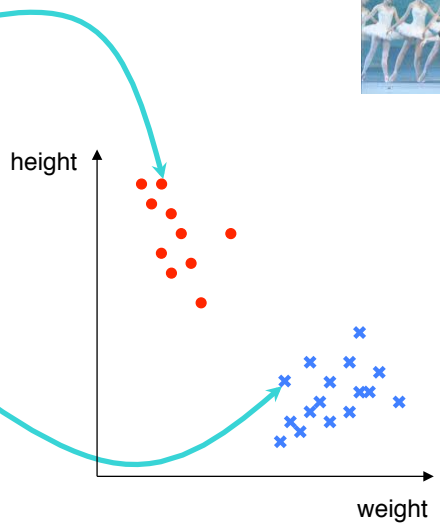
Class, or "label"

The Supervised Learning Pipeline

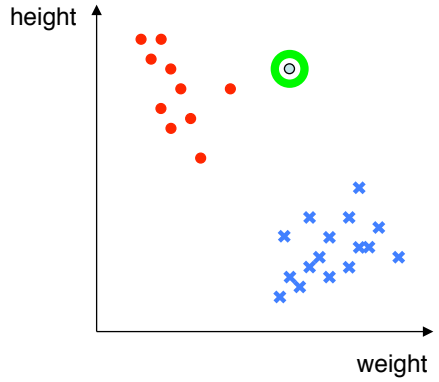


Taking measurements….

Person	Weight	Height
1	63kg	190cm
2	55kg	185cm
3	75kg	202cm
4	50kg	180cm
5	57kg	174cm
…	…	
16	85kg	150cm
17	93kg	145cm
18	75kg	130cm
19	99kg	163cm
20	100kg	171cm
…	…	



The Nearest Neighbour Rule



Person	Weight	Height
1	63kg	190cm
2	55kg	185cm
3	75kg	202cm
4	50kg	180cm
5	57kg	174cm
…	…	
16	85kg	150cm
17	93kg	145cm
18	75kg	130cm
19	99kg	163cm
20	100kg	171cm
…	…	

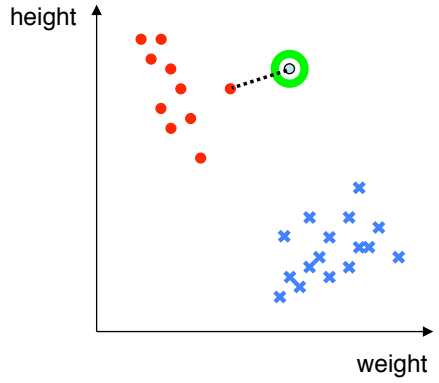
"TRAINING" DATA

"TESTING" DATA

Who's this guy?
- player or dancer?

height = 180cm
weight = 78kg

The Nearest Neighbour Rule



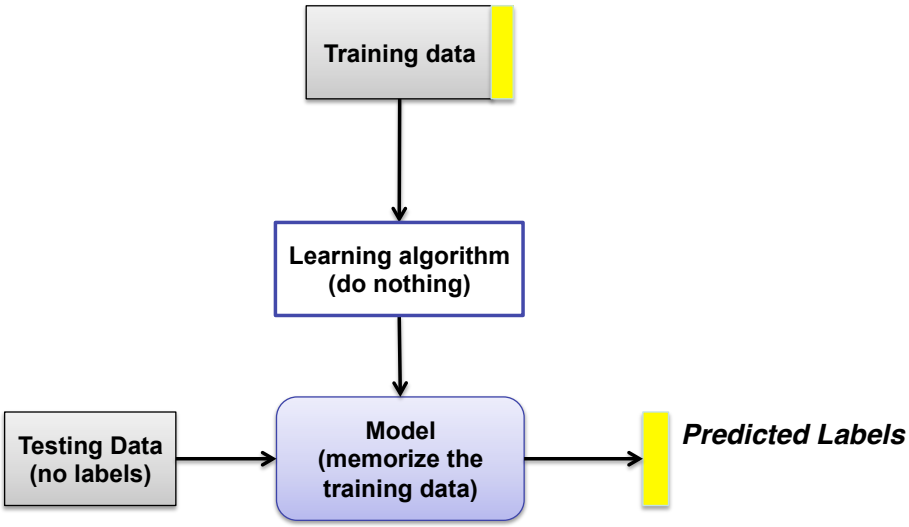
Person	Weight	Height
1	63kg	190cm
2	55kg	185cm
3	75kg	202cm
4	50kg	180cm
5	57kg	174cm
…	…	
16	85kg	150cm
17	93kg	145cm
18	75kg	130cm
19	99kg	163cm
20	100kg	171cm
…	…	

"TRAINING" DATA

height = 180cm
weight = 78kg

1. **Find nearest neighbour**
2. **Assign the same class**

Supervised Learning Pipeline for Nearest Neighbour

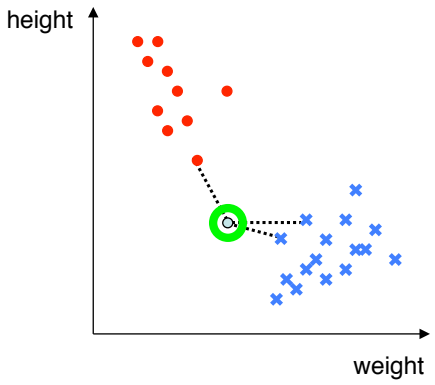


The K-Nearest Neighbour Classifier

```
Testing point x
For each training datapoint x'
   measure distance(x,x')
End
Sort distances
Select K nearest
Assign most common class
```

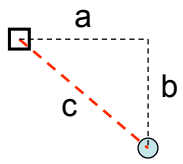
Person	Weight	Height
1	63kg	190cm
2	55kg	185cm
3	75kg	202cm
4	50kg	180cm
5	57kg	174cm
…	…	
16	85kg	150cm
17	93kg	145cm
18	75kg	130cm
19	99kg	163cm
20	100kg	171cm
…	…	

"TRAINING" DATA



Quick reminder: Pythagoras' theorem

```
. . .
   measure distance(x,x')
. . .
```

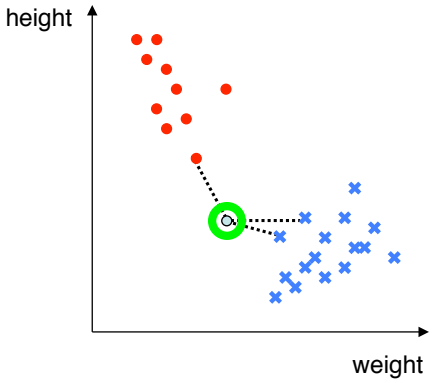


$$a^2 + b^2 = c^2$$

$$So.... \quad c = \sqrt{a^2 + b^2}$$

a.k.a. "Euclidean" distance

$$\text{distance}(x, x') = \sqrt{\sum_i (x_i - x'_i)^2}$$



The K-Nearest Neighbour Classifier

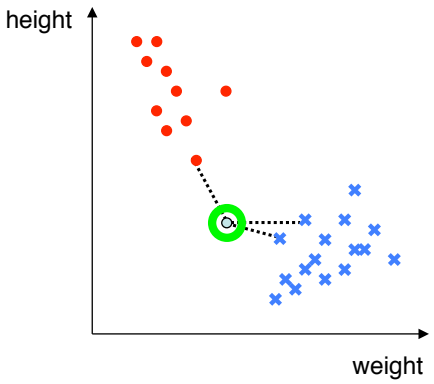
```
Testing point x
For each training datapoint x'
   measure distance(x,x')
End
Sort distances
Select K nearest
Assign most common class
```

Person	Weight	Height
1	63kg	190cm
2	55kg	185cm
3	75kg	202cm
4	50kg	180cm
5	57kg	174cm
…	…	
16	85kg	150cm
17	93kg	145cm
18	75kg	130cm
19	99kg	163cm
20	100kg	171cm
…	…	

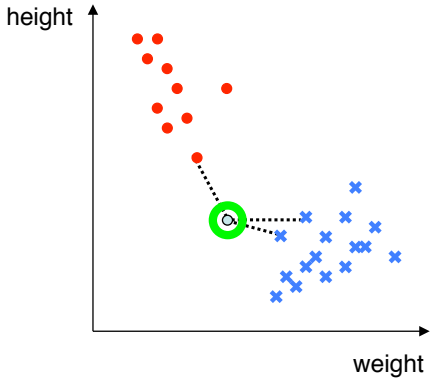
"TRAINING" DATA

Seems sensible.

But what are the disadvantages?



The K-Nearest Neighbour Classifier



Person	Weight	Height
1	63kg	190cm
2	55kg	185cm
3	75kg	202cm
4	50kg	180cm
5	57kg	174cm
…	…	
16	85kg	150cm
17	93kg	145cm
18	75kg	130cm
19	99kg	163cm
20	100kg	171cm
…	…	

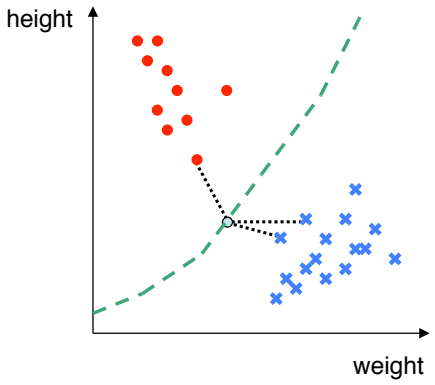
"TRAINING" DATA

Here I chose k=3.

What would happen if I chose k=5?

What would happen if I chose k=26?

The K-Nearest Neighbour Classifier



Person	Weight	Height
1	63kg	190cm
2	55kg	185cm
3	75kg	202cm
4	50kg	180cm
5	57kg	174cm
…	…	
16	85kg	150cm
17	93kg	145cm
18	75kg	130cm
19	99kg	163cm
20	100kg	171cm
…	…	

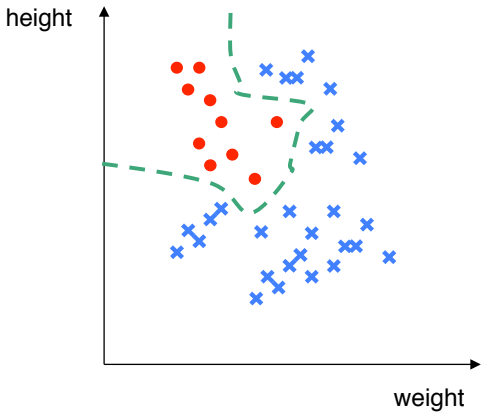
"TRAINING" DATA

*Any point on the left of this **"boundary"** is closer to the **red circles**.*

*Any point on the right of this **"boundary"** is closer to the **blue crosses**.*

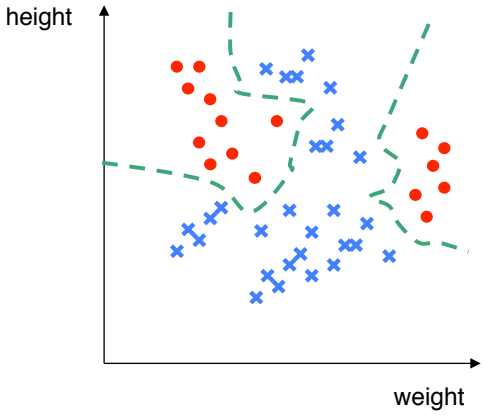
*This is called the **"decision boundary".***

Where's the decision boundary?



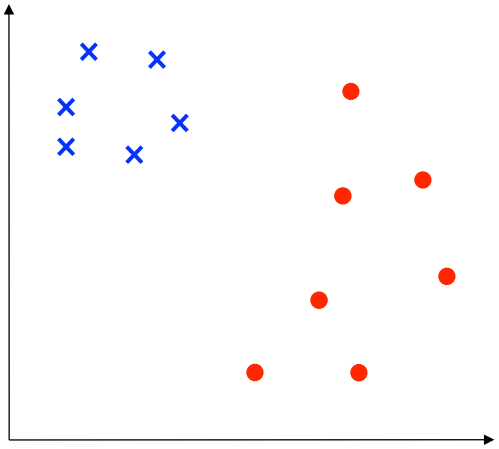
Not always a simple straight line!

Where's the decision boundary?



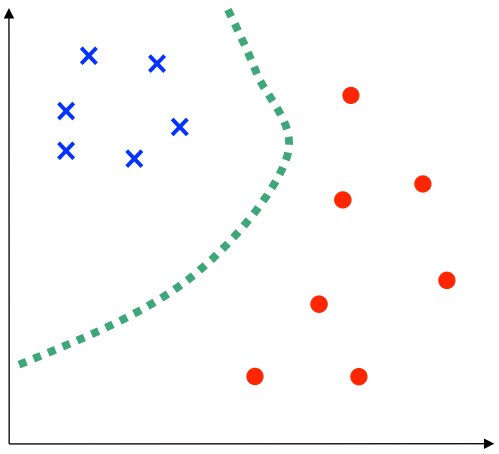
Not always contiguous!

The *most important* concept in Machine Learning



The *most important* concept in Machine Learning

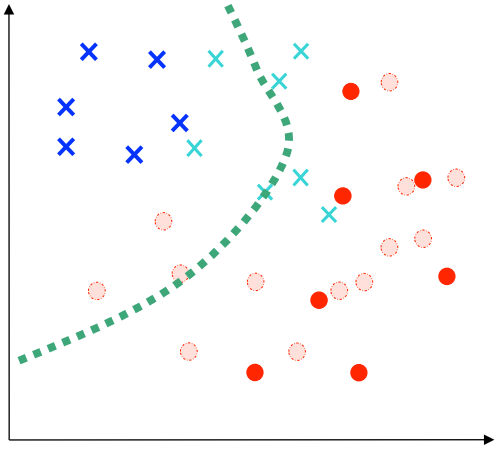
Looks good so far…



The *most important* concept in Machine Learning

Looks good so far…

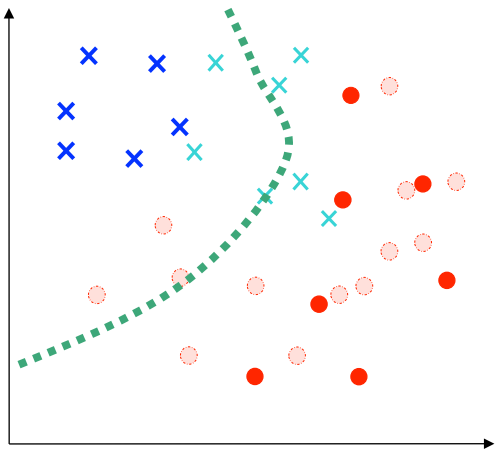
Oh no! Mistakes!
What happened?



The *most important* concept in Machine Learning

Looks good so far…

Oh no! Mistakes!
What happened?



We didn't have all the data.

We can never assume that we do.

This is called "OVER-FITTING" to the small dataset.

So, we have our first "machine learning" algorithm… ?

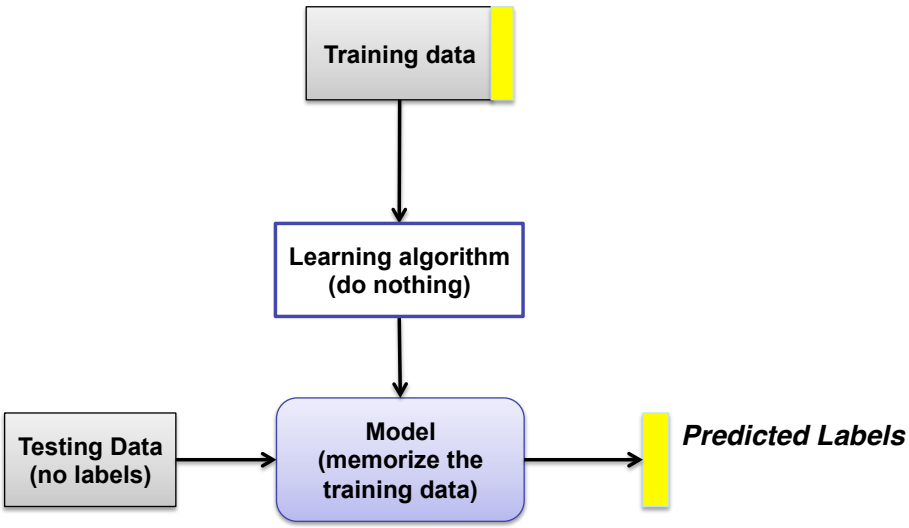
The K-Nearest Neighbour Classifier

```
Testing point x
For each training datapoint x'
   measure distance(x,x')
End
Sort distances
Select K nearest
Assign most common class
```

Make your own notes on its advantages / disadvantages.

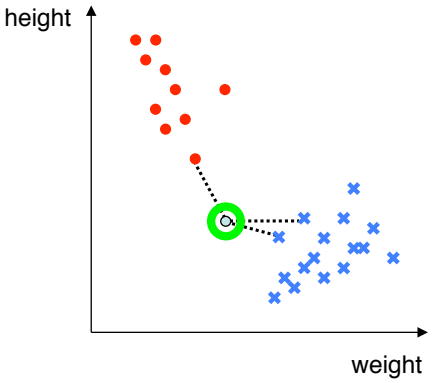
I will ask for volunteers next time we meet…..

Pretty dumb! Where's the learning!

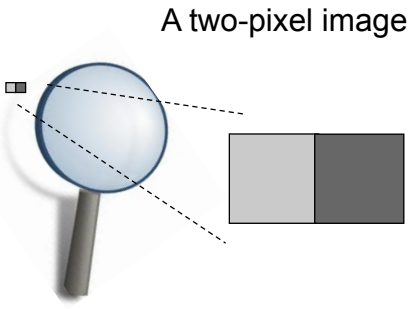


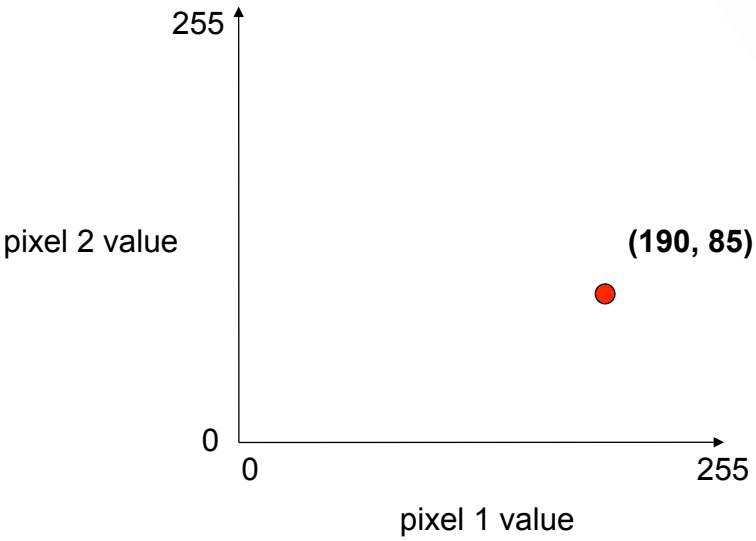
Now, how is this problem like

handwriting recognition?



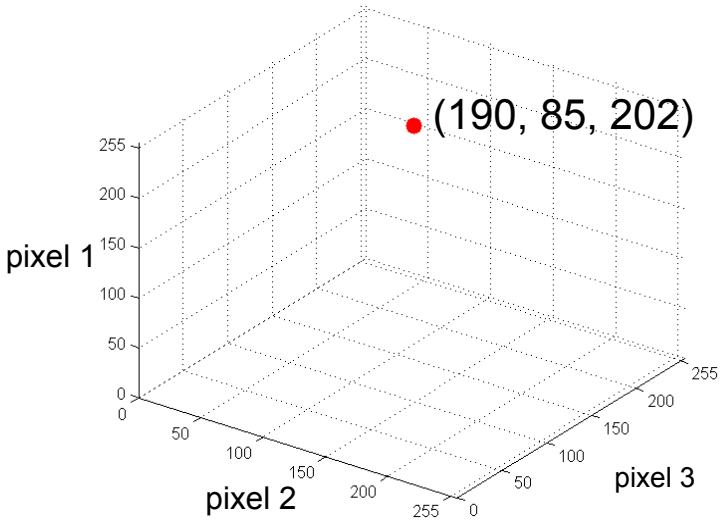
Let's say the measurements are pixel values.





Three dimensions…

A three-pixel image

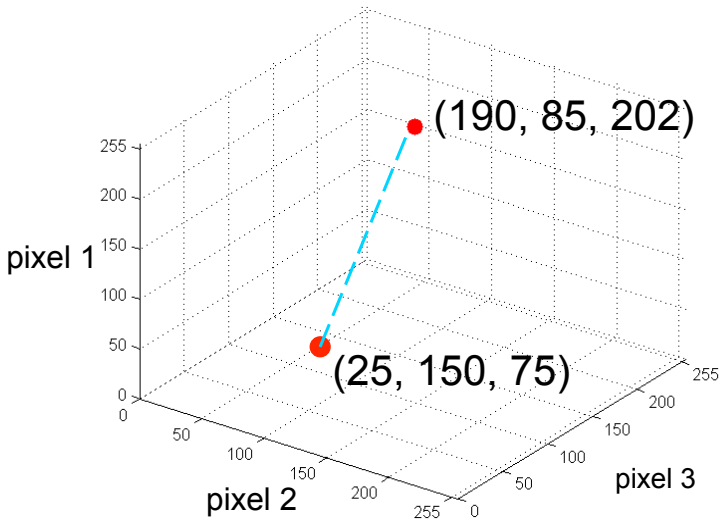


This 3-pixel image is represented by a **SINGLE** point in a 3-D space.

Distance between images

A three-pixel image

Another 3-pixel image

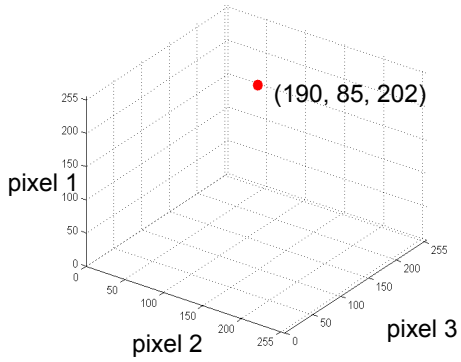


Straight line distance between them?

4-dimensional space? 5-d? 6-d?

A three-pixel image

A four-pixel image.

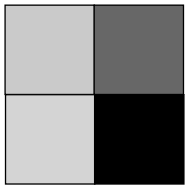


A five-pixel image

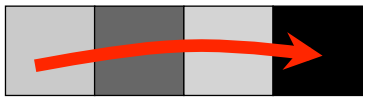
A four-pixel image.

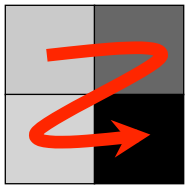
(190, 85, 202, 10)

A different four-pixel image.



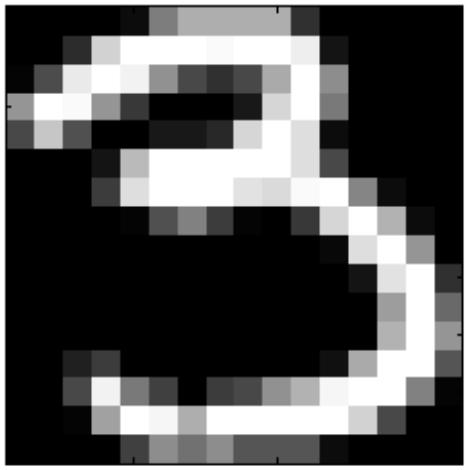
(190, 85, 202, 10)
Same 4-dimensional vector!



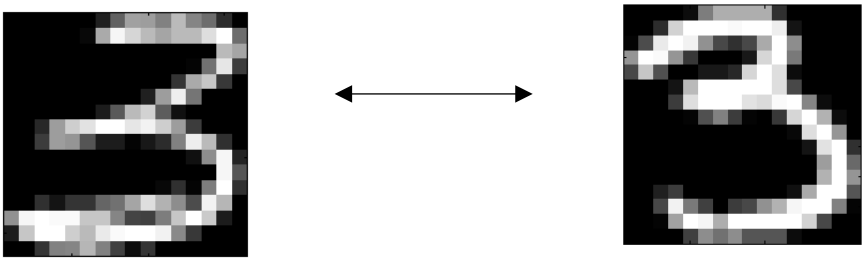


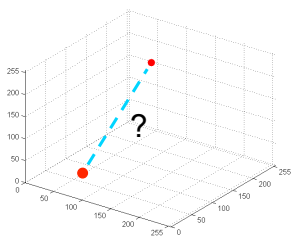
Assuming we read pixels in a systematic manner, we can now represent any image as a single point in a high dimensional space.

16 x 16 pixel image. How many dimensions?



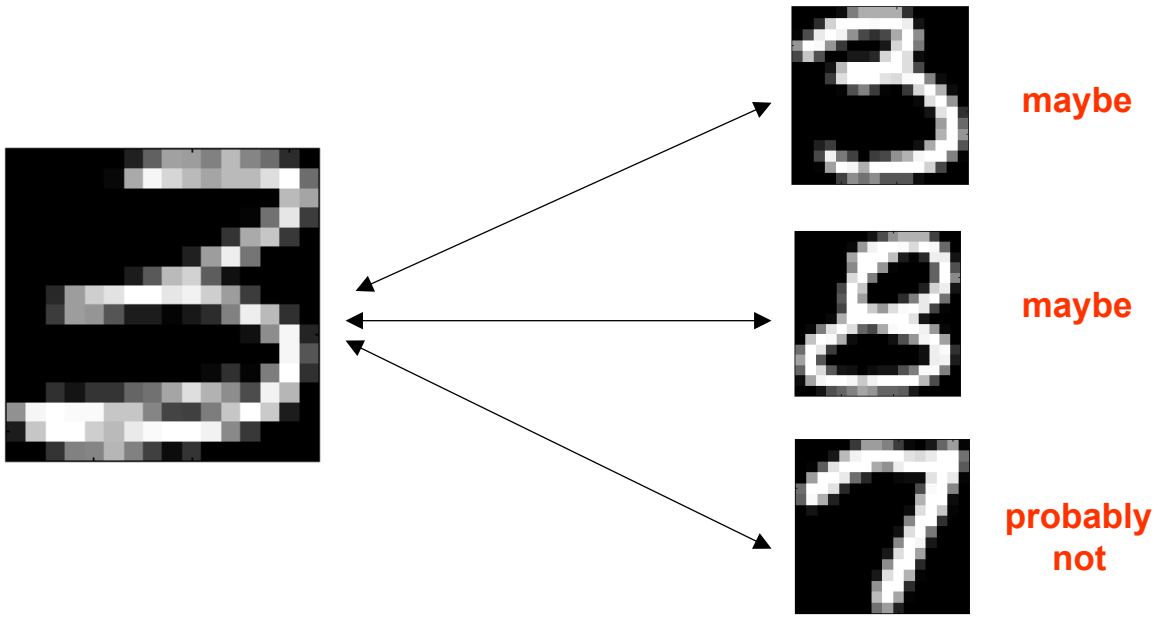
We can measure distance in 256 dimensional space.





$$\text{distance}(x, x') = \sqrt{\sum_{i=1}^{i=256} (x_i - x'_i)^2}$$

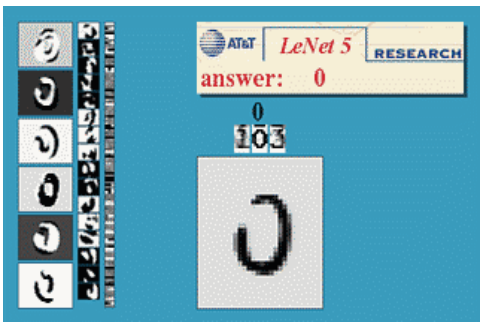
Which is the **nearest neighbour** to our ‘3’ ?



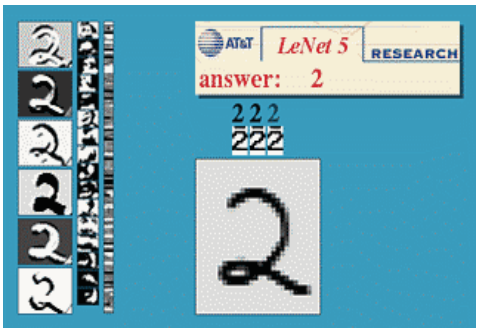
AT&T Research Labs

The USPS postcode reader – **learnt** from examples.

UNITED STATES POSTAL SERVICE



FAST recognition



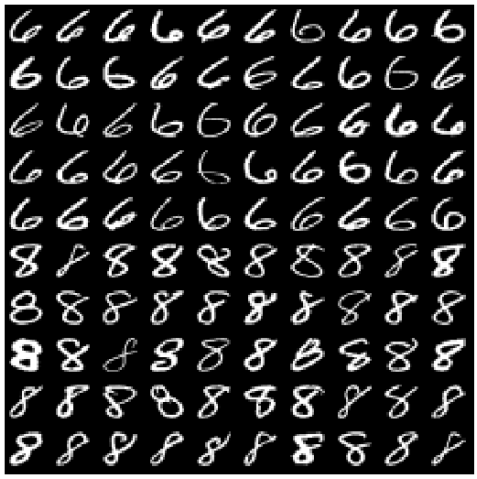
NOISE resistant, can **generalise** to future unseen patterns

Your lab exercise

Use K-NN to recognise handwritten USPS digits.

Final lab session before reading week.
(see website for details)

i.e. you have 4 weeks.



Biggest mistake by students in last year's class?

…starting 2 weeks from now!