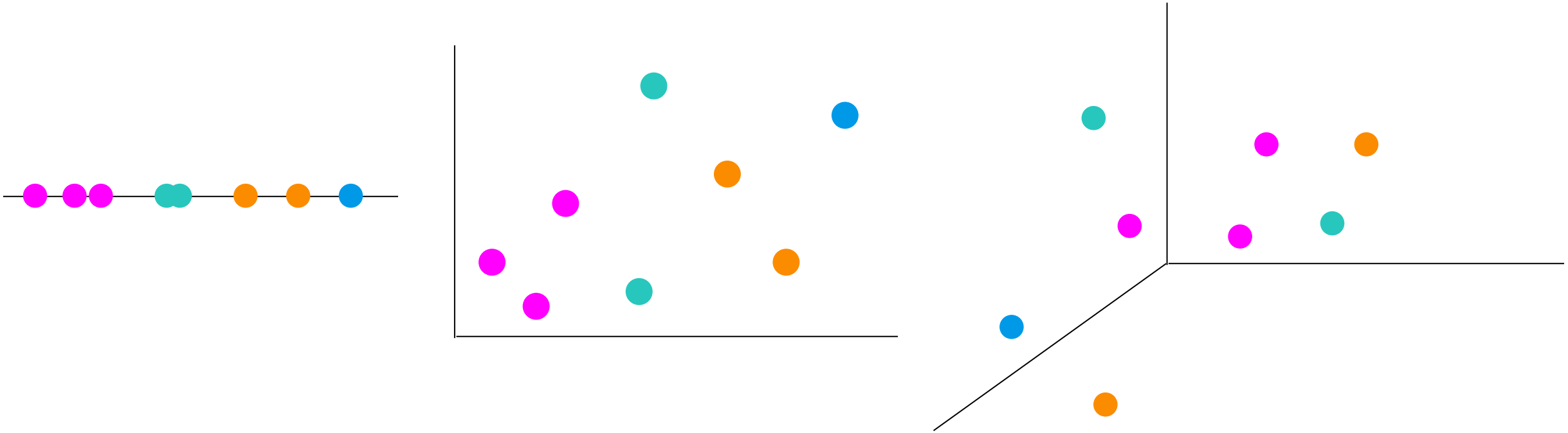
The Curse of Dimensionality

More data is always better, right?

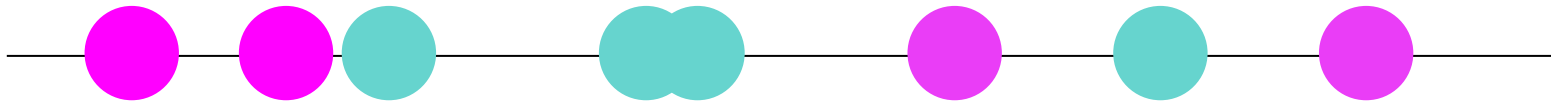
- Not always! Depends on whether you're adding **observations** or **dimensions**
- If we hold fixed the number of observations, adding more dimensions is not always good
- Issues from having too many dimensions:
  - Overfitting
  - Model that's harder to interpret
- This problem is called **the curse of dimensionality** and it's not necessarily intuitive
- This applies both to classification and regression, but it'll be easier to visualize for classification

Data is sparser in higher dimensions



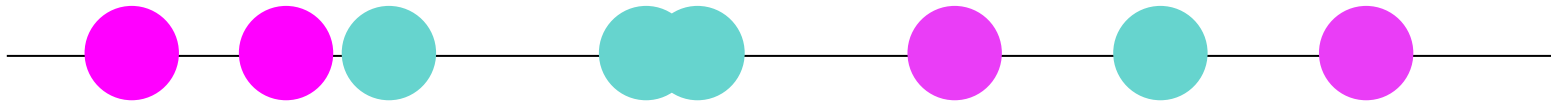
Several dimensions are better than one

- Let’s try a classification task in one dimension



Several dimensions are better than one

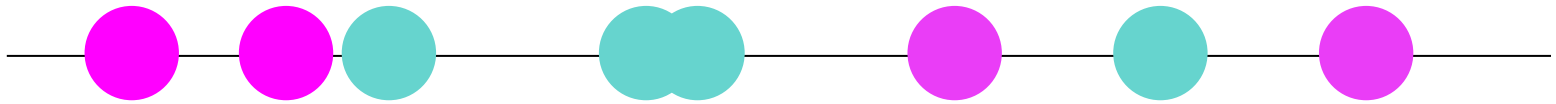
- Let’s try a classification task in one dimension



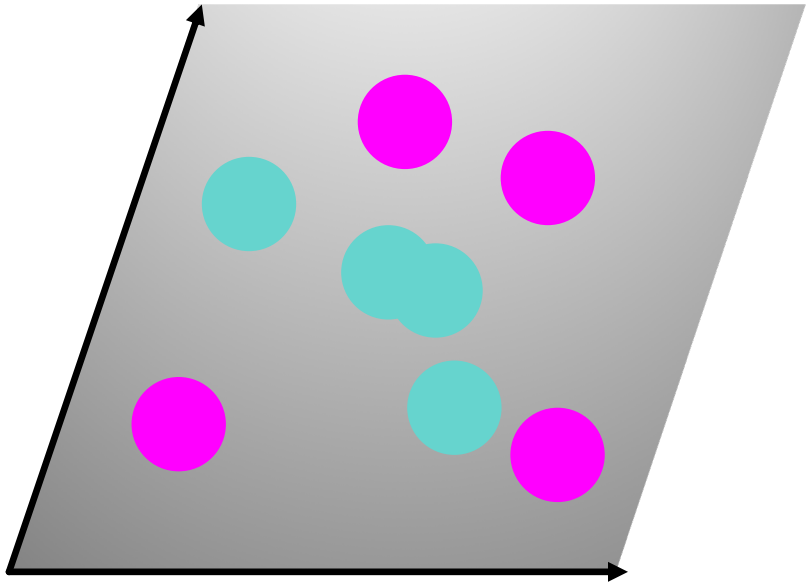
- We can’t get separability in this case
- Let’s try adding another dimension

Several dimensions are better than one

- Let’s try a classification task in one dimension

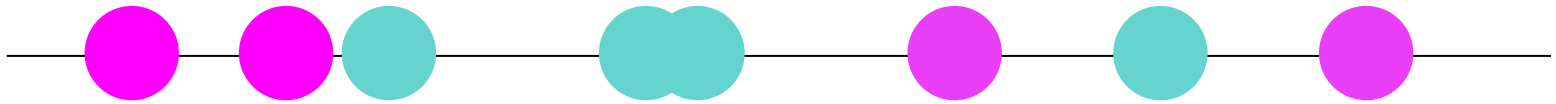


- We can’t get separability in this case
- Let’s try adding another dimension

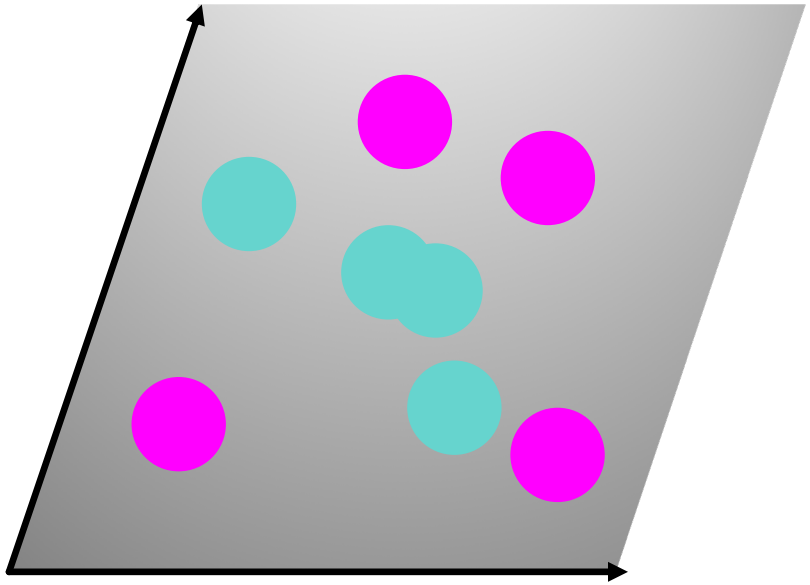


Several dimensions are better than one

- Let’s try a classification task in one dimension



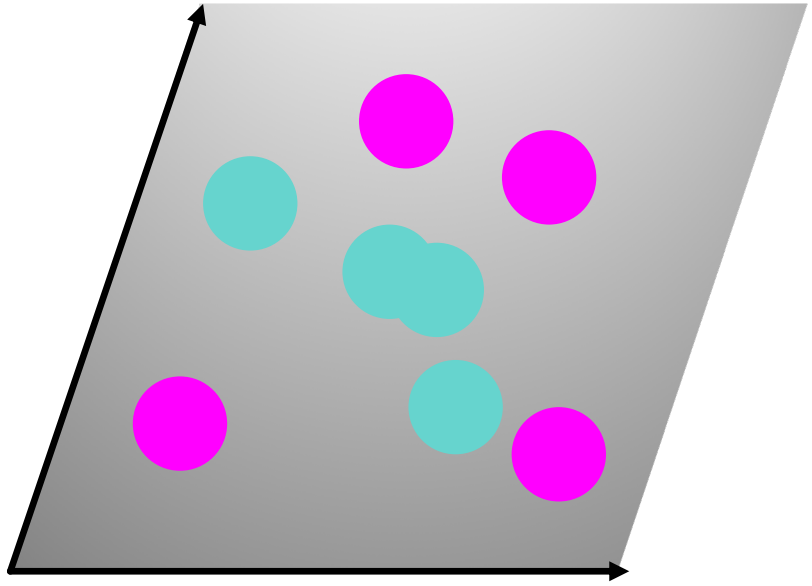
- We can’t get separability in this case
- Let’s try adding another dimension

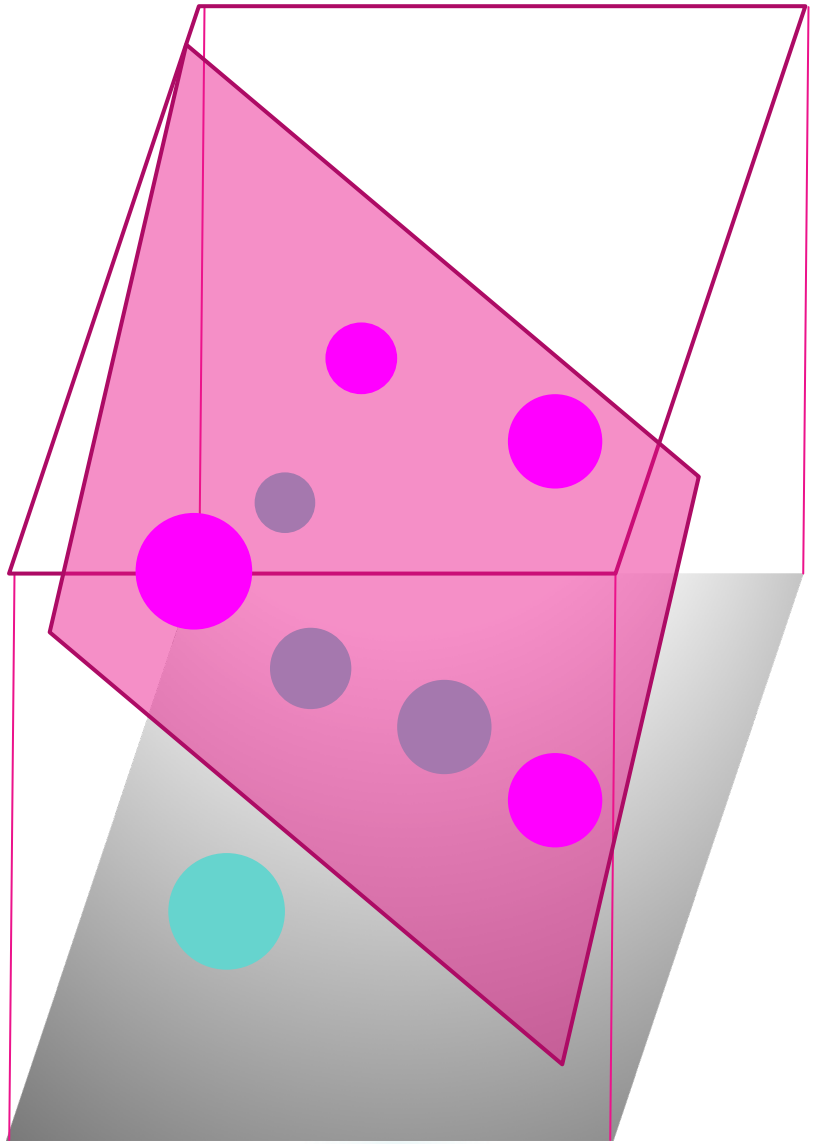


- Hmm, still not great
- What if we add a third dimension?

Several dimensions are better than one

- With a third dimension, we can now linearly separate our classes



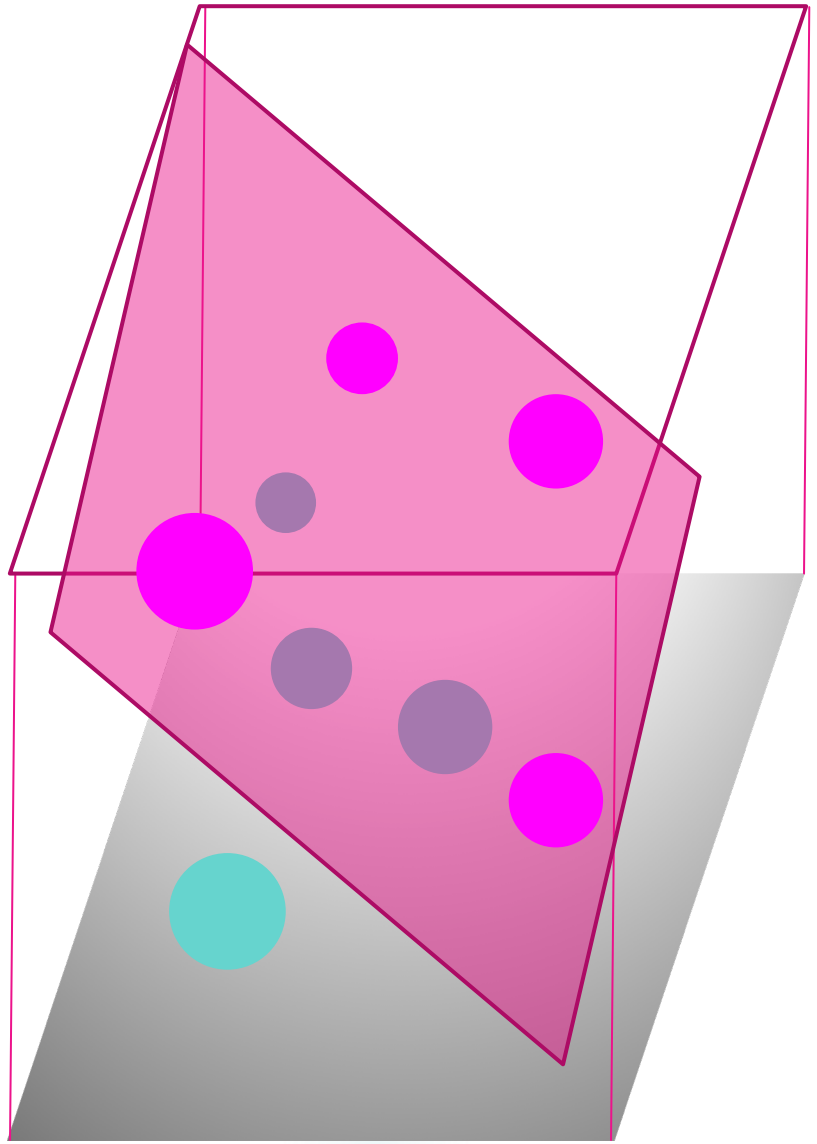


Several dimensions are better than one

- With a third dimension, we can now linearly separate our classes
- Does this mean we should keep adding dimensions forever?

Check for understanding

What are some things we notice about our observations as we add more dimensions?



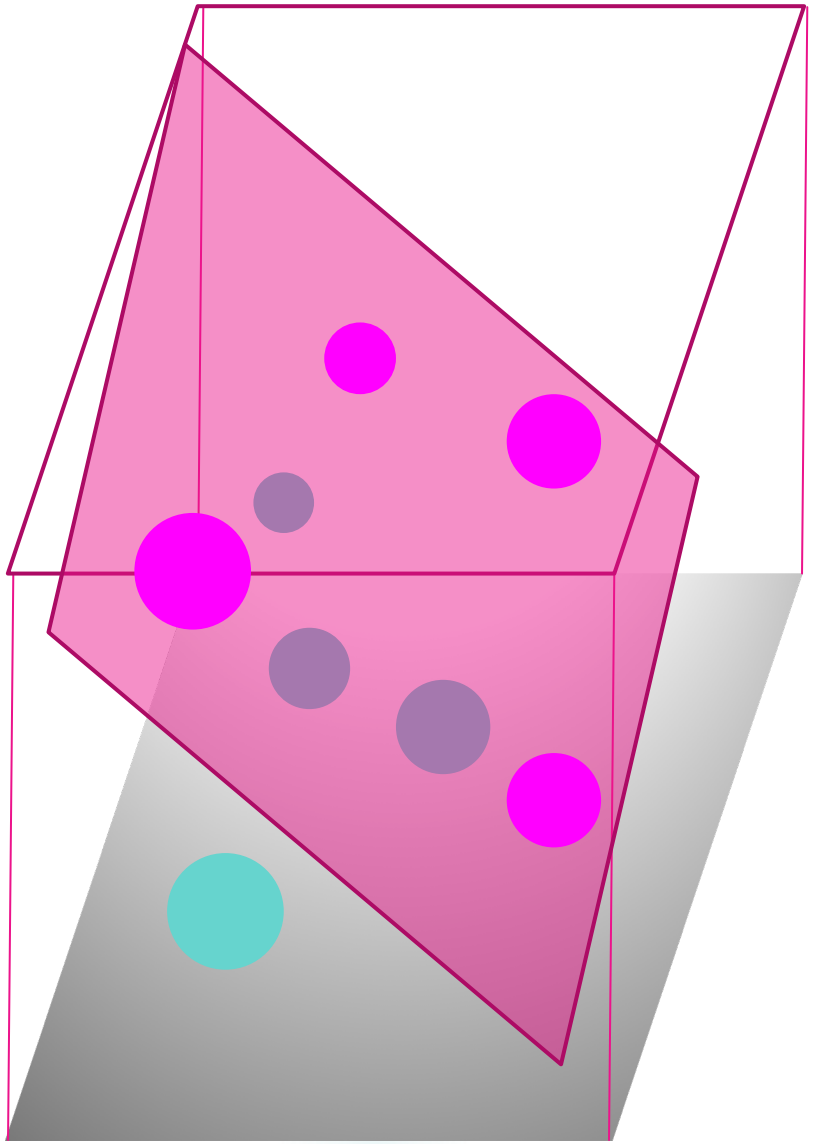
Several dimensions are better than one

- With a third dimension, we can now linearly separate our classes
- Does this mean we should keep adding dimensions forever?

Check for understanding

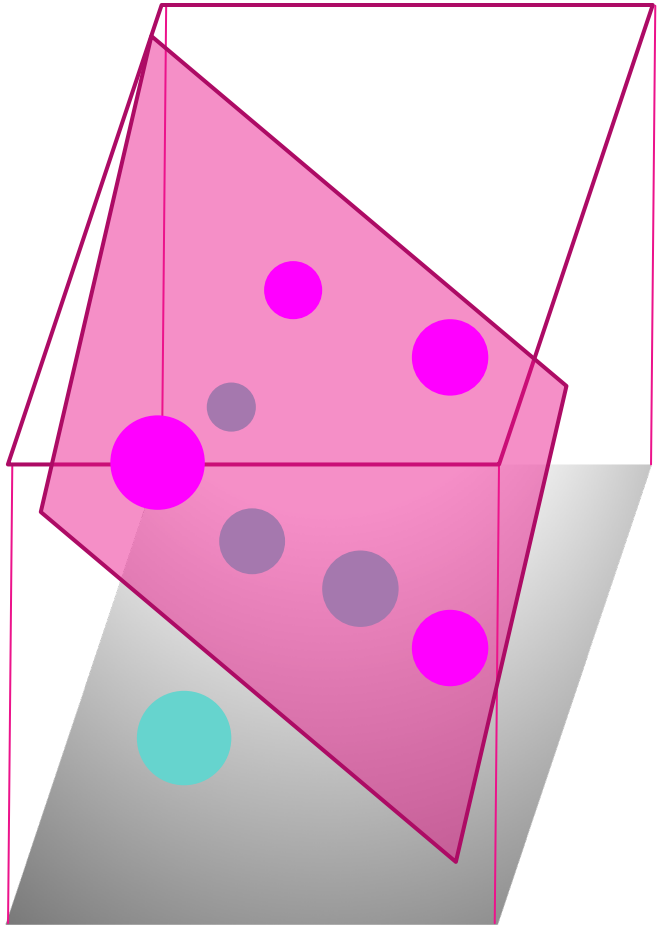
What are some things we notice about our observations as we add more dimensions?

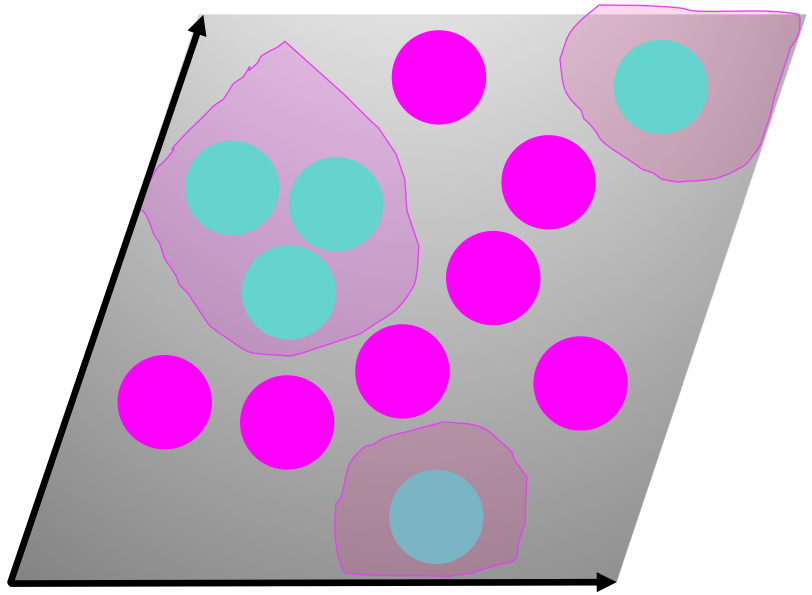
- Points are farther away from each other
- More dimensions make it possible to linearly separate our observations



Adding dimensions is good, up to a limit

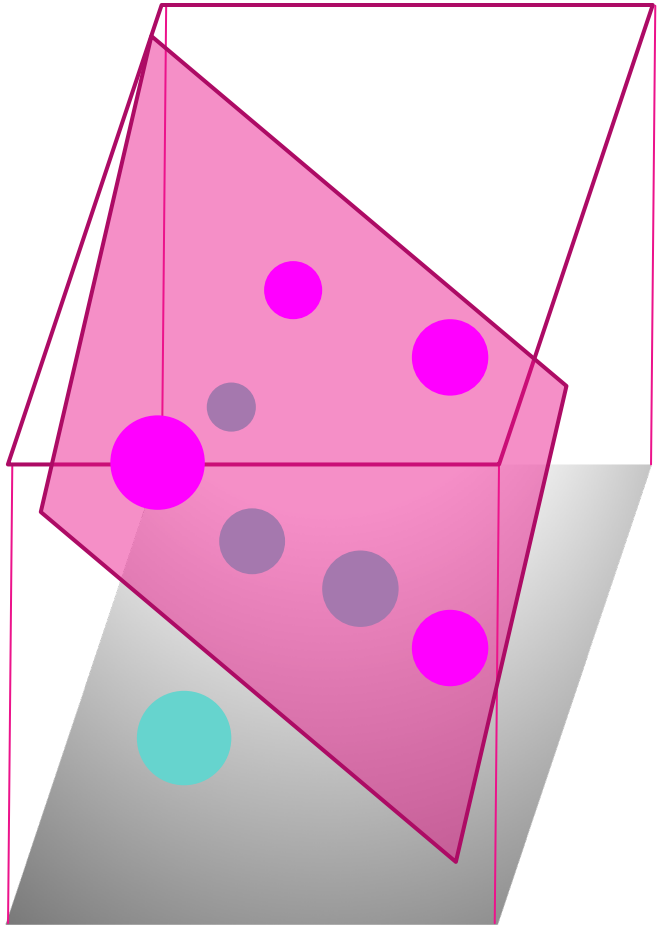
- What happens if we squish our observations and separating plane down to two dimensions?

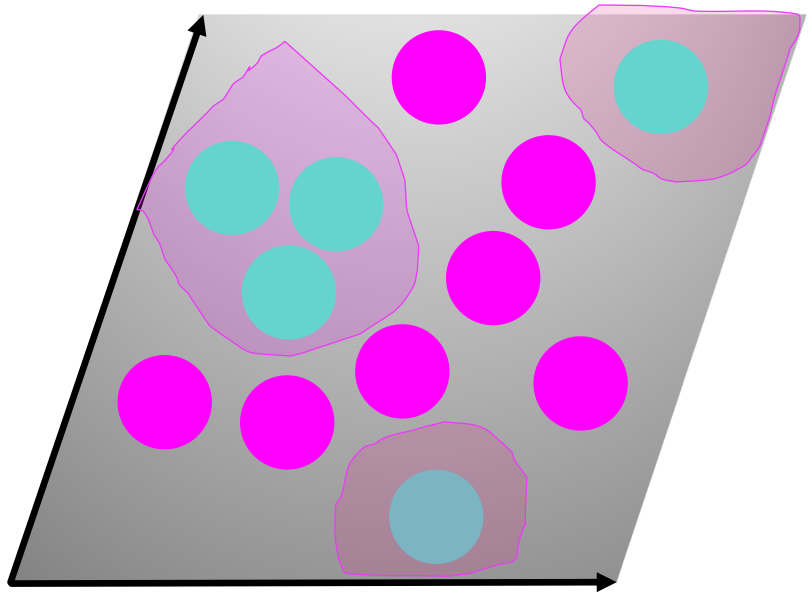




Adding dimensions is good, up to a limit

- What happens if we squish our observations and separating plane down to two dimensions?



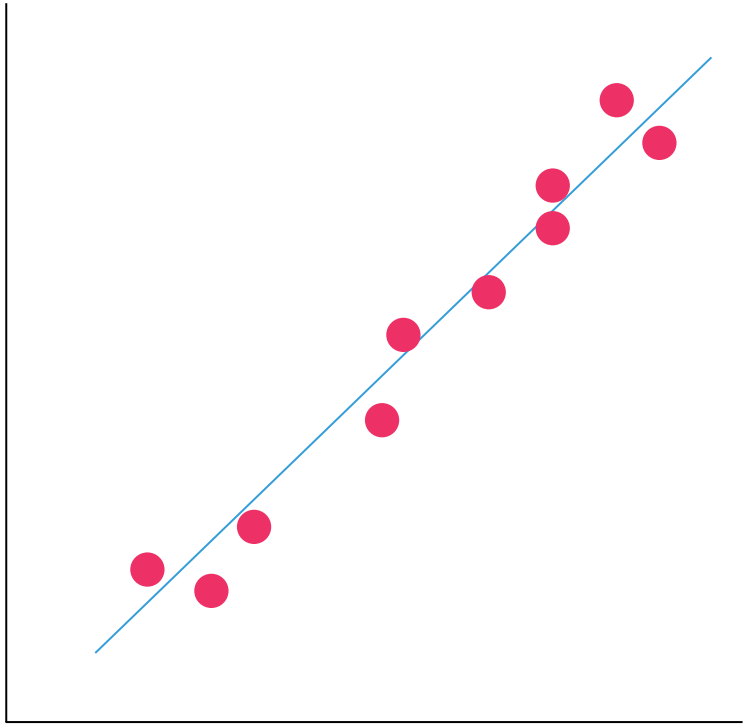


- Yikes, it looks like we've **over-fit**

The curse of dimensionality

- Observations are farther apart in higher dimensions
- As our data becomes more sparse, clustering and classification tasks become harder
  - As density decreases, there are fewer observations in any given neighborhood, meaning we have to travel farther to find a neighbor to be part of our cluster/class
  - If all of our neighbors are far away, what does it even mean to belong to the same cluster/class?
  - If our observations are spread out, we need more data to decrease our variance

Dimensionality reduction



- To deal with the curse of dimensionality, we can reduce our dimensions
- We could represent our pink dots in two dimensions, or we could collapse all points to be on the line and accept some minor error
- What we gain from reducing dimensionality:
  - Fewer dimensions to deal with, and comes at relatively cheap cost of minor error
  - Remove redundancy and noise
  - Can be used for visualization

More on this in the next lecture!

In the next lecture, we'll look at some methods we can use to fight the curse of dimensionality