Example Application Principle Component Analysis

Frederik Mallmann-Trenn
6CCS3AIN

PCA Example

- Say you have a bunch of house listings and you would like to group them into student housing, regular and luxury

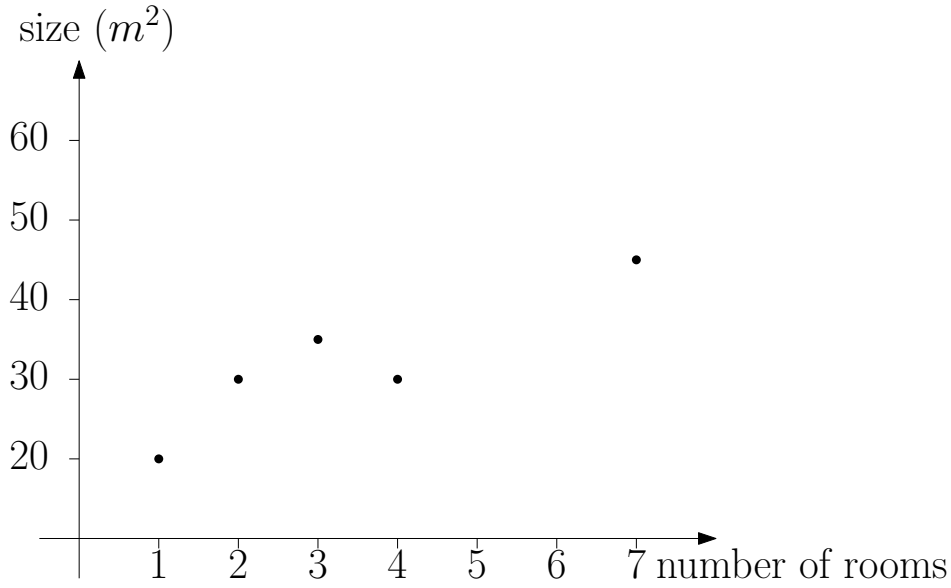
PCA Example

- Let's say we have the following features
  - Floor size (m^2)
  - Number of rooms
  - Distance supermarket
  - Distance King's
  - Hipster vibe
- Let's say we want to reduce to two features to have a nice visual representation.
- Can we reduce it to two features?

Why does PCA work?

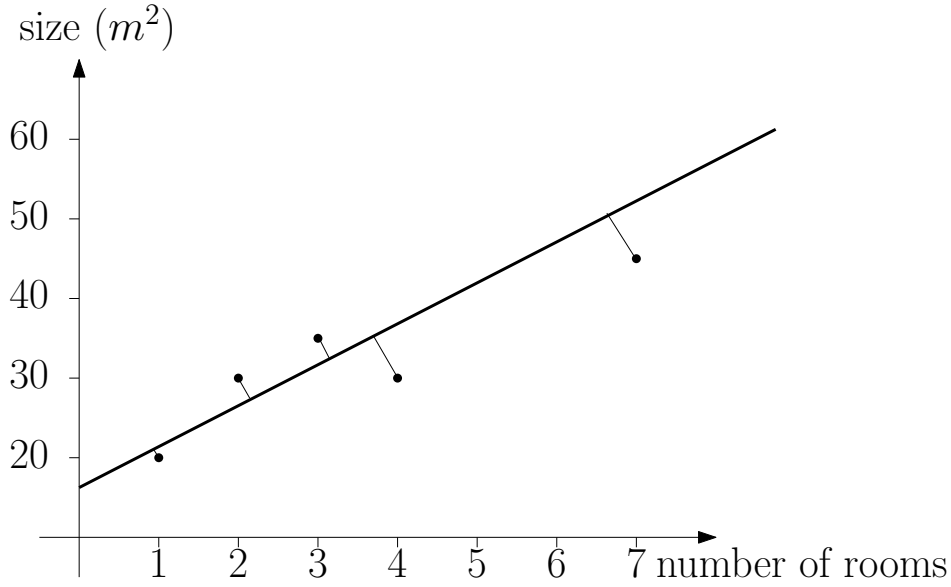
- Reduce to two or three features
  - Size
    - Floor size (m^2)
    - Number of rooms
  - Location
    - Distance supermarket
    - Distance King's
    - Hipster vibe
- Why does this make sense?

PCA



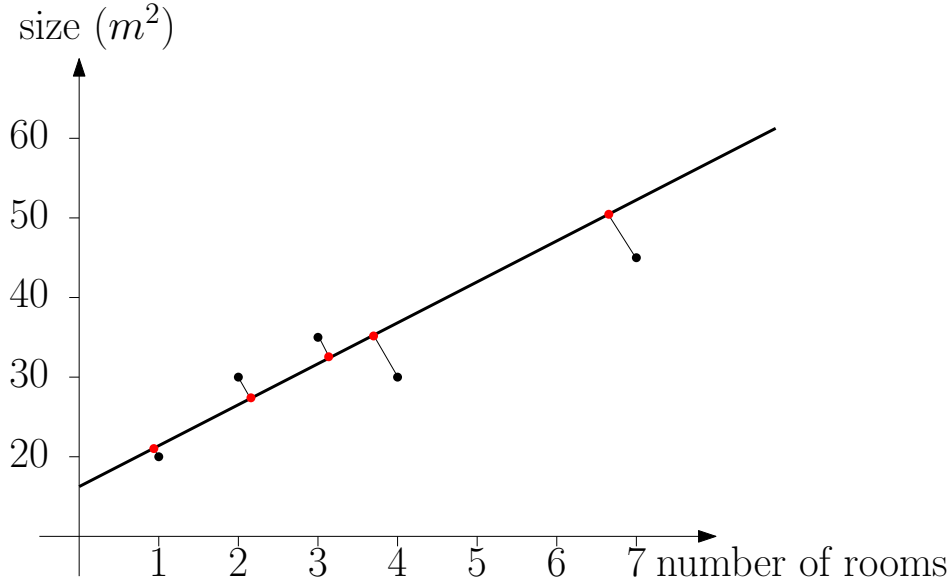
- For example the floor size and the number of rooms are often correlated
- Let's see how it would look like if we compressed both dimensions to one dimension

PCA



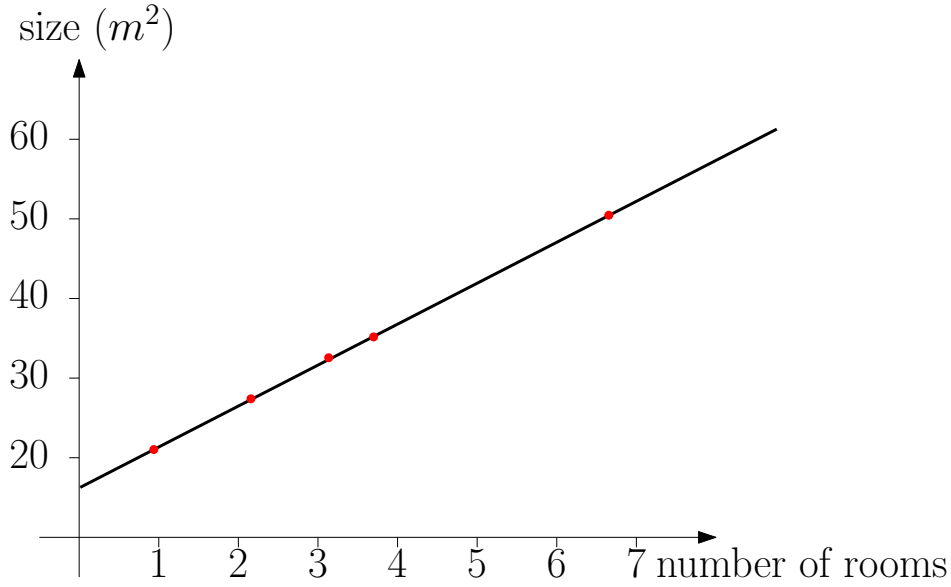
- If we take the line that minimises the Least Squares Distance, we get ...

PCA



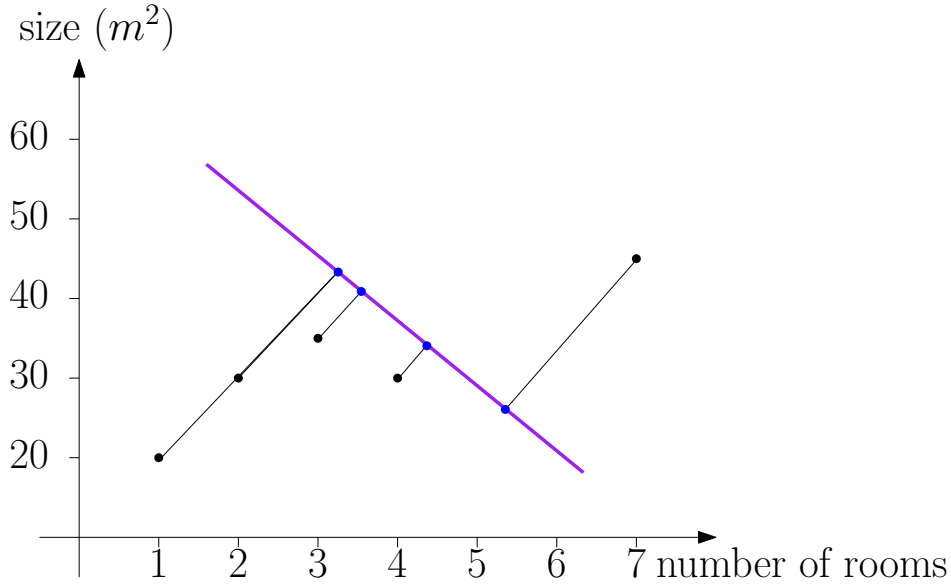
- If we take the line that minimises the Least Squares Distance, we get ...
- ... the following projection of the points.

PCA



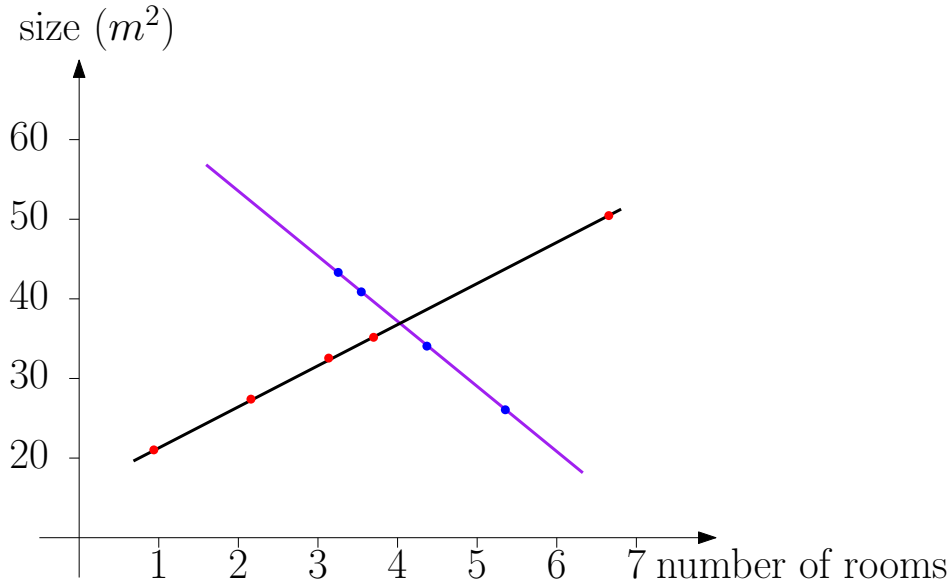
- After cleaning up, this is what we get

PCA



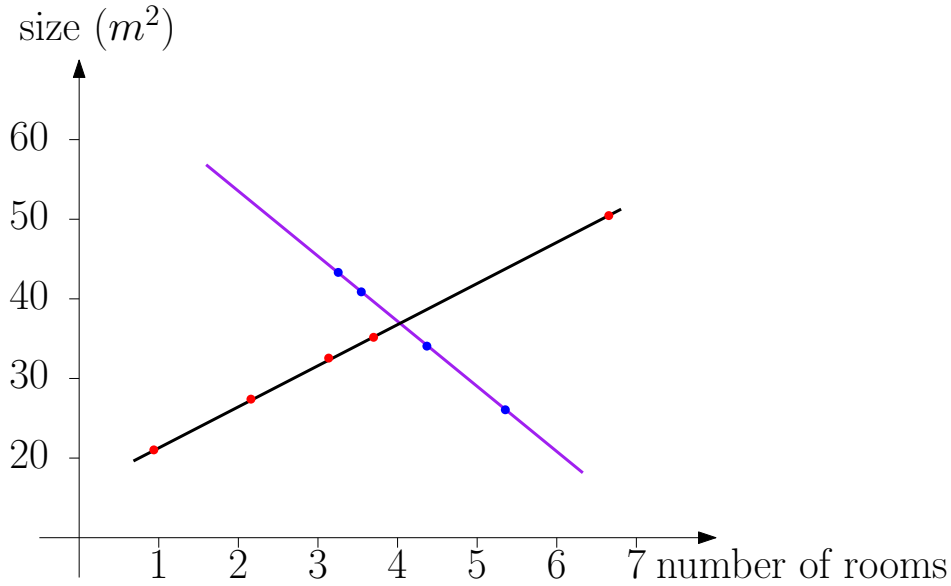
- What if we take a different line? (purple)?

PCA



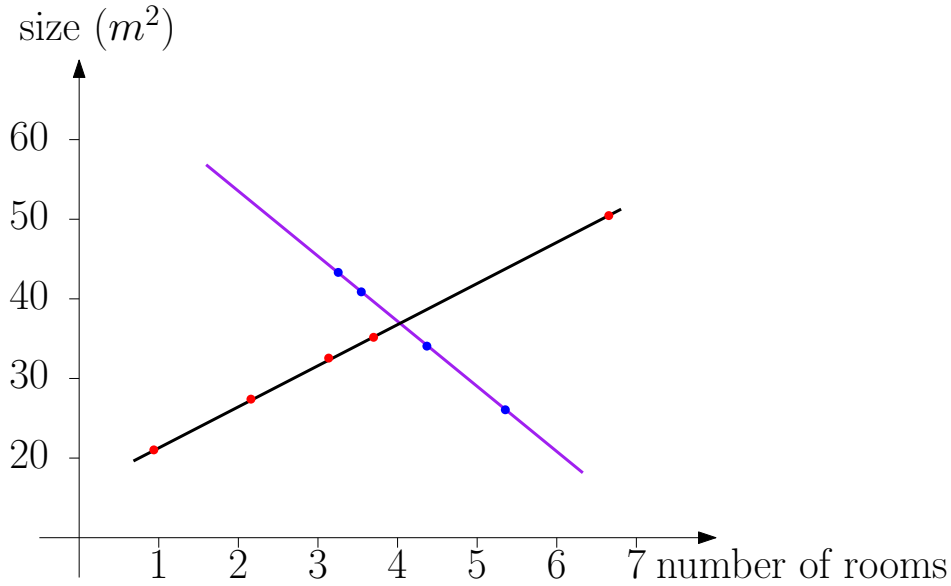
- The spread here is the variance of the data
- And we would like to maximise it.

PCA



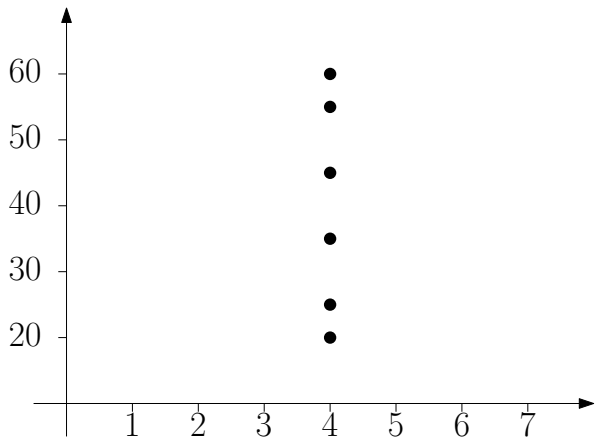
- The spread here is the variance of the data
- And we would like to maximise it.
- Intuitively, the more variance we capture, the better we can approximate the higher-dimensional space (here $d = 2$)

PCA

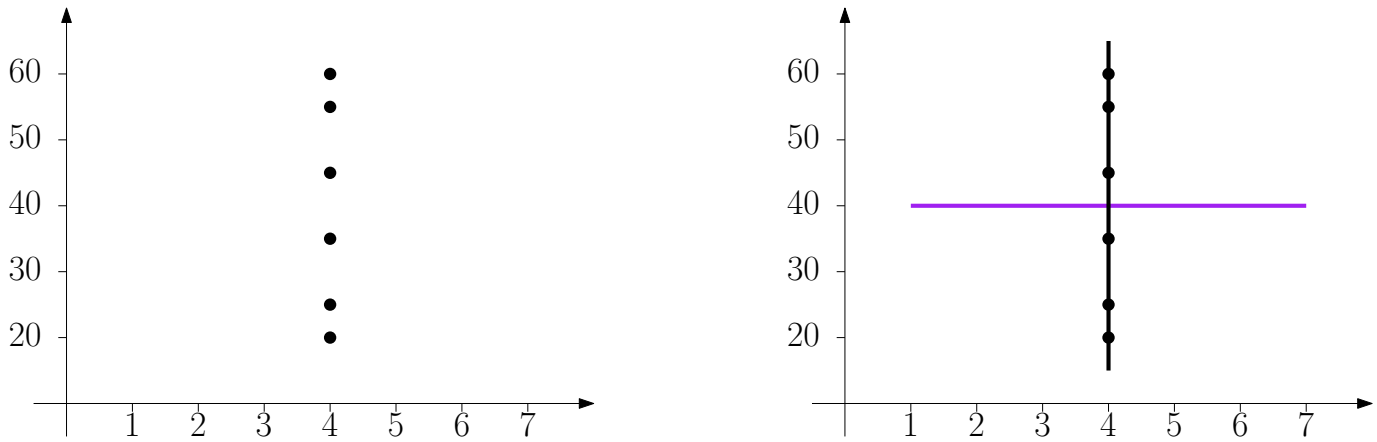


- The spread here is the variance of the data
- And we would like to maximise it.
- Intuitively, the more variance we capture, the better we can approximate the higher-dimensional space (here $d = 2$)
- If we compare them, we see that the points are less spread out on the purple line
- The black line actually maximises the spread and therefore is the best for approximating the higher-dimensional space

Why should we maximise the variance?

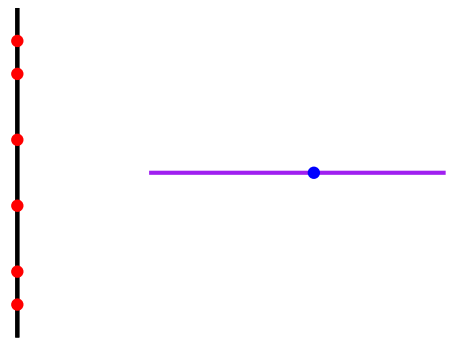


Why should we maximise the variance?



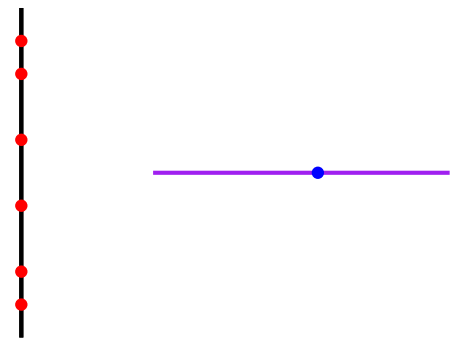
- Left: New input
- Right: Two potential lines onto which we can project.
- Consider projecting to a **horizontal** and a **vertical** line.

Why should we maximise the variance?



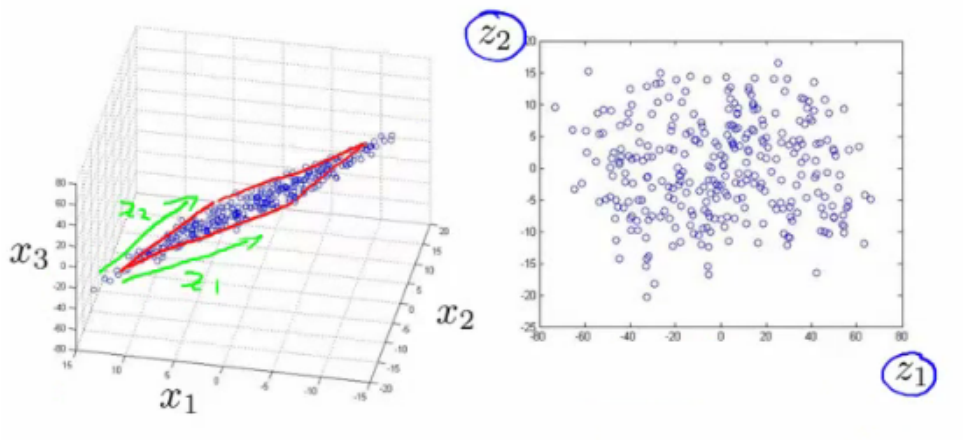
- This is how the output would looks like
- Which line retains more information?

Why should we maximise the variance?



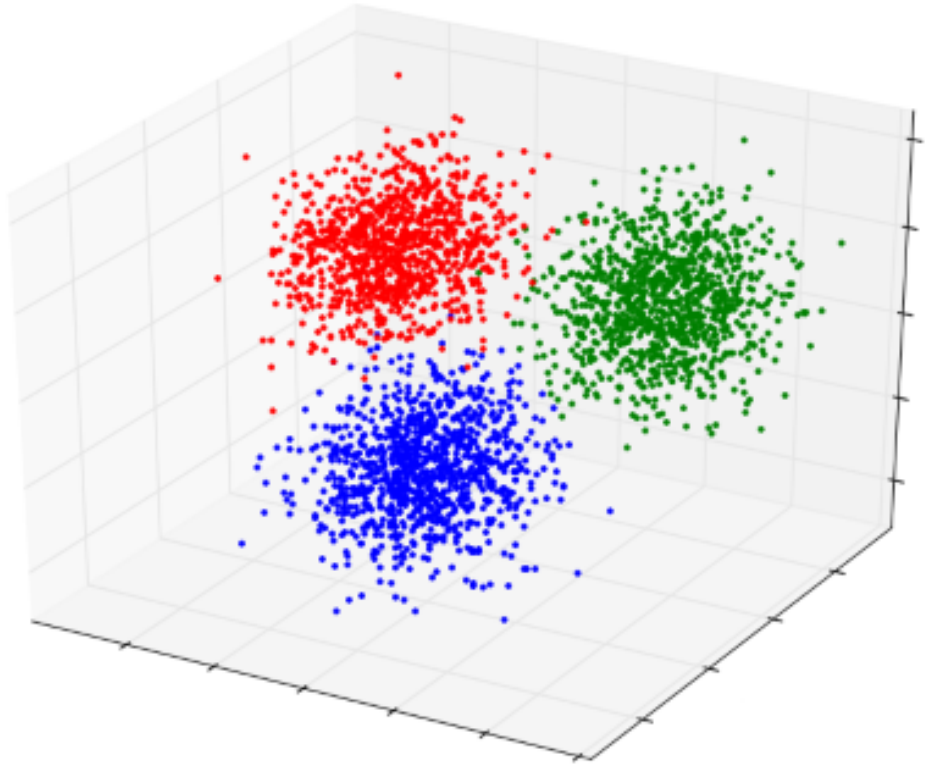
- This is how the output would looks like
- Which line retains more information?
- Clearly the black line, all points on the purple line are at the same location.

3D to 2D



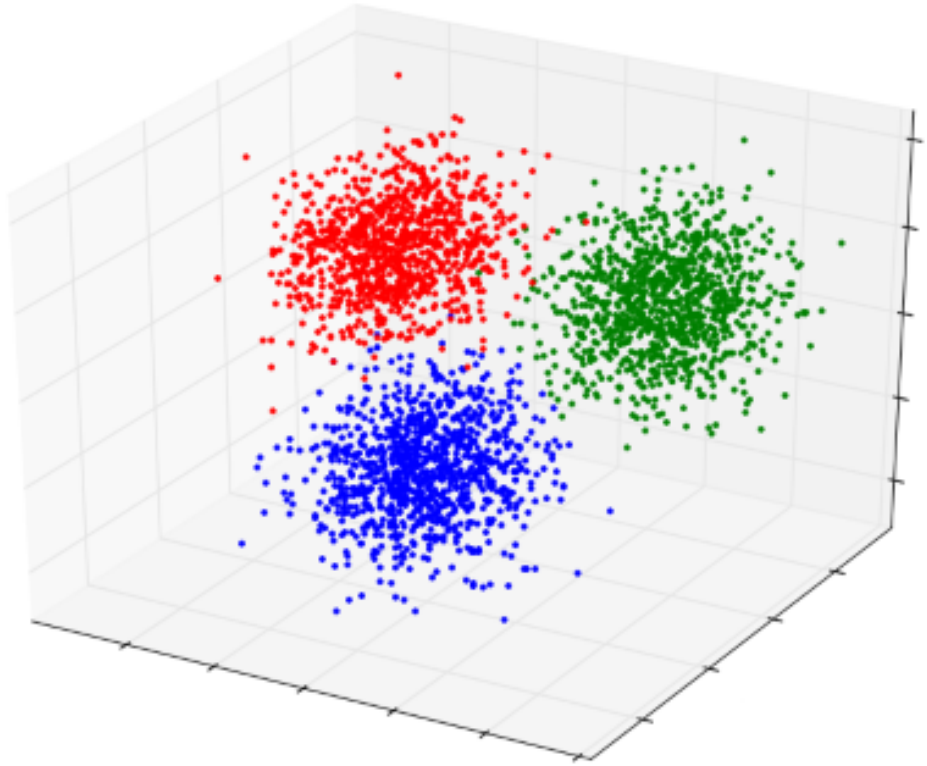
- Let's say our three dimensions (x_1, x_2, and x_3) are as on the l.h.s.
  - Distance supermarket
  - Distance King's
  - Hipster vibe
- Then after reducing it to $2D$ it looks like the r.h.s.
- We may also wish to reduce it to just a line $(1D)$, but we can see that this would be very lossy

5D to 3



- If we plot our 5D data using the components we found (1 for size and 2 for location)
- We get this 3D plot
- We can see that our different classes student housing, regular and luxury are well-separated.

5D to 3



- If we plot our 5D data using the components we found (1 for size and 2 for location)
- We get this 3D plot
- We can see that our different classes student housing, regular and luxury are well-separated.
- This is the whole point: reduce the information, but keep the important information!