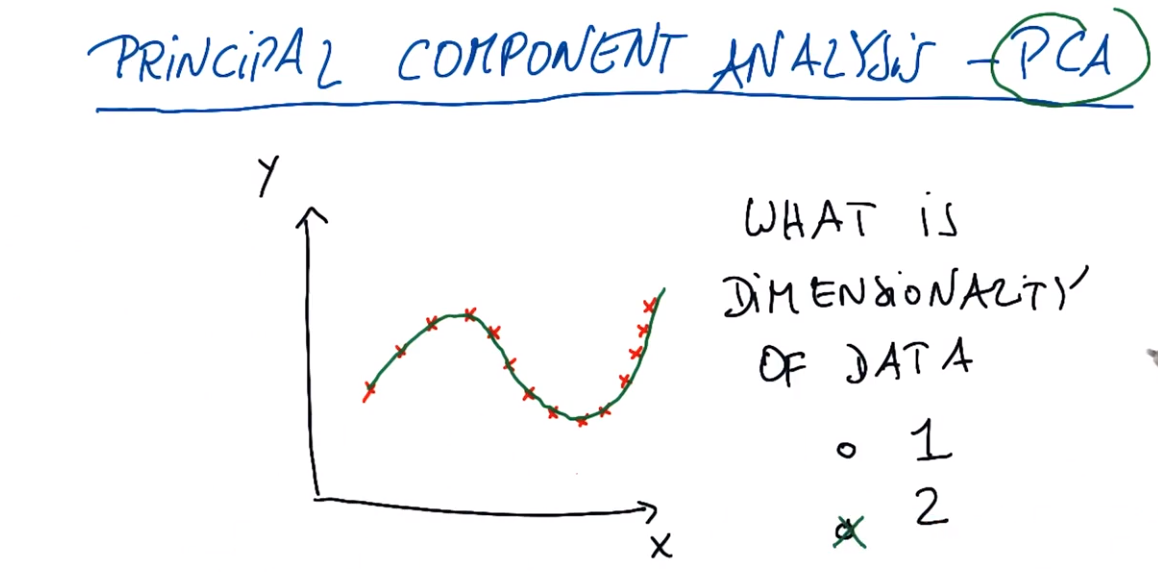
PCA主成分分析

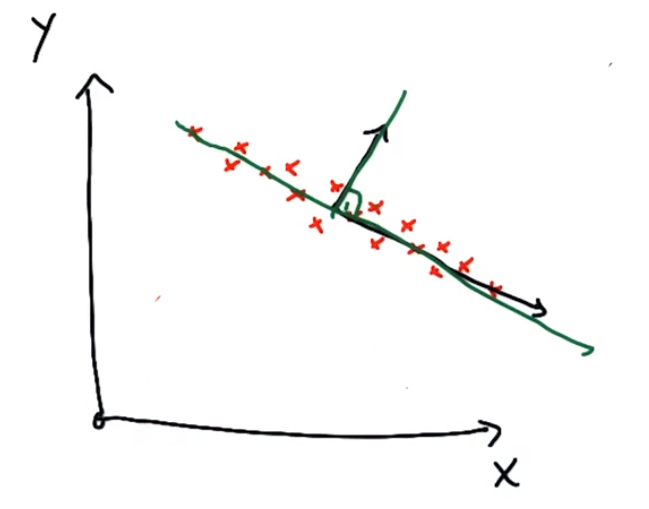
主成分分析（英语：Principal components analysis，PCA）是一种分析、简化数据集的技术。主成分分析经常用于减少数据集的维数，同时保存数据集中的对方差贡献最大的特征。

**判断数据维度**



PCA 特别擅长处理,坐标系的移位和旋转问题，在这个例子中，你无法在原始坐标系中，通过移位和旋转得到一个坐标系，以使其获得一维性。

**用于数据转换的PCA**



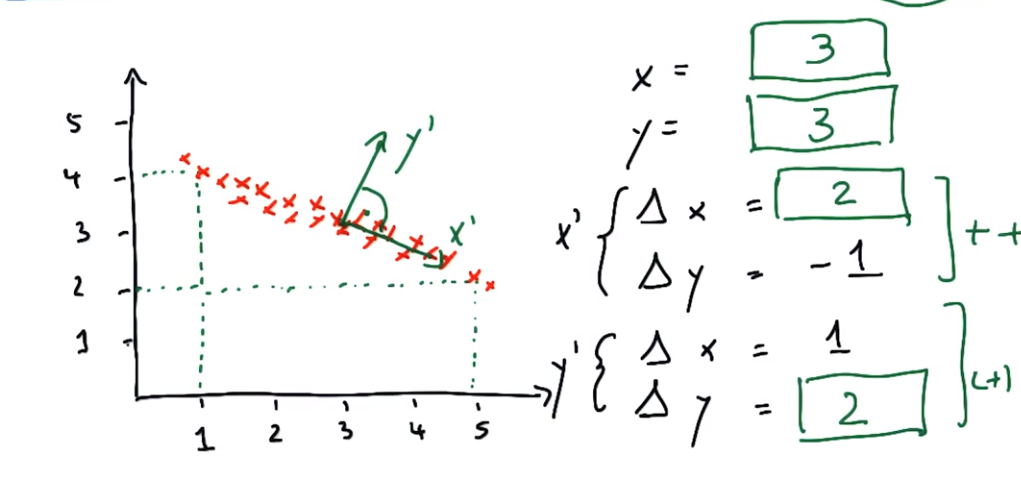
无论你收到任何形状的数据，比如此处的这个数据点云。

PCA都能从 旧坐标系统 仅通过 转化和轮换 获得 新坐标系统。PCA为你找到这些轴 并告诉你这些轴的重要性

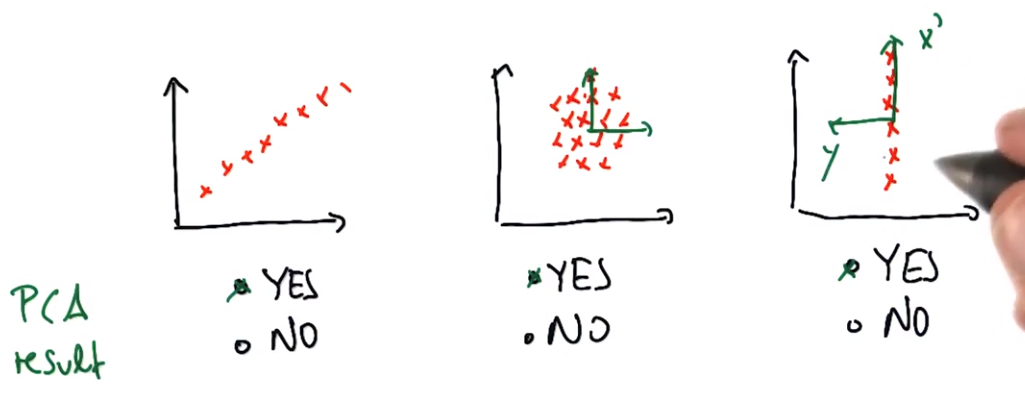
**PCA查找新轴**

通过 PCA 还可得出一个重要值 就是X’轴的一个散布值

如果散布率较小 该散布值对于X'轴来说倾向于一个很大值，而对于Y’轴来说则小很多



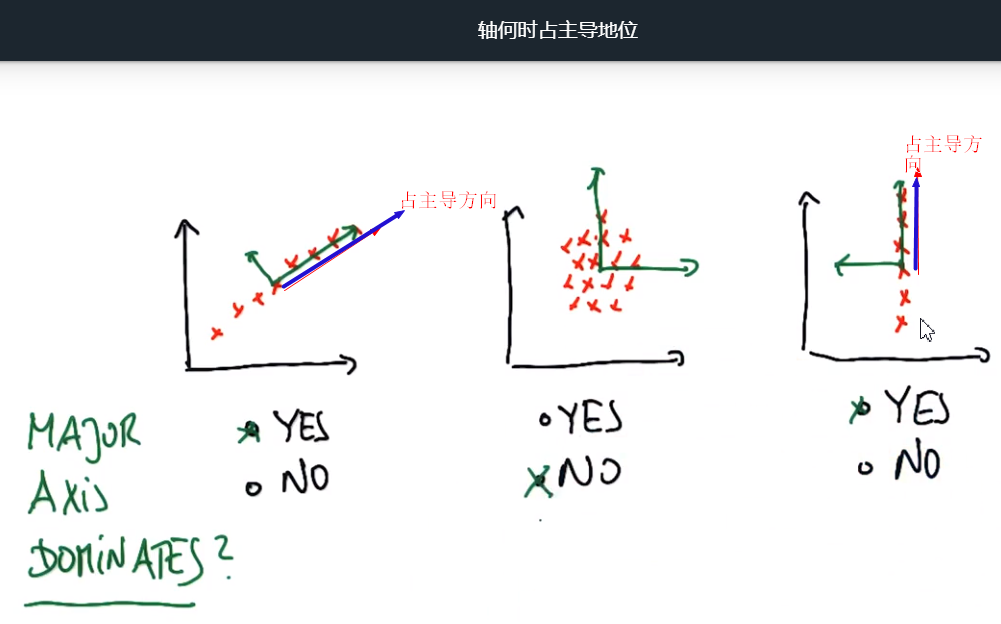
**PCA的新坐标系（那些数据可用于PCA）**



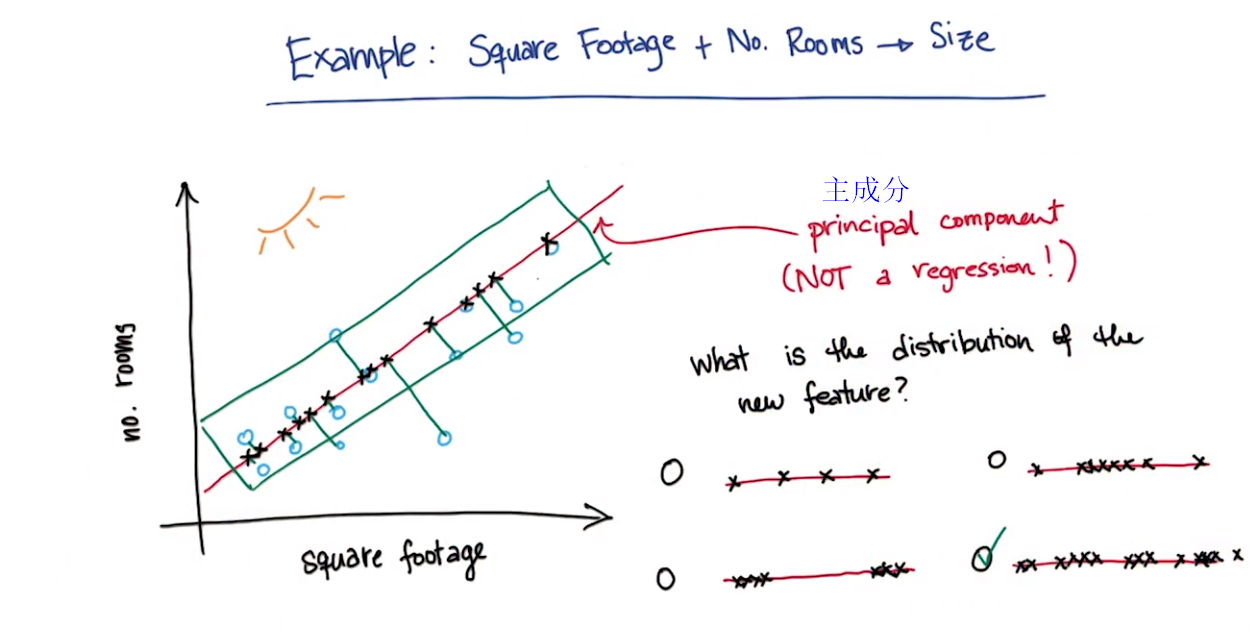
第三个例子比较特殊

因为，我们无法构建垂直的回归函数(既无法用函数 y = f(x) 来分割)

在PCA中 我们只有向量,所以 坐标系中的 X 轴有可能是垂直的，而 Y 轴则朝向左方 这就是该问题的 PCA 答案。



**复合特征**



使用回归算法，找出数据的大致方向，在尽可能减少信息损失的情况下，将数据 映射在 主成分 方向上。

假设我有一盏灯，它的光线垂直照射于主要成分的表面，这时 数据点投射出影子 都在一张纸上。

此时数据最初的 二维 ,变成现在的一维数据了.

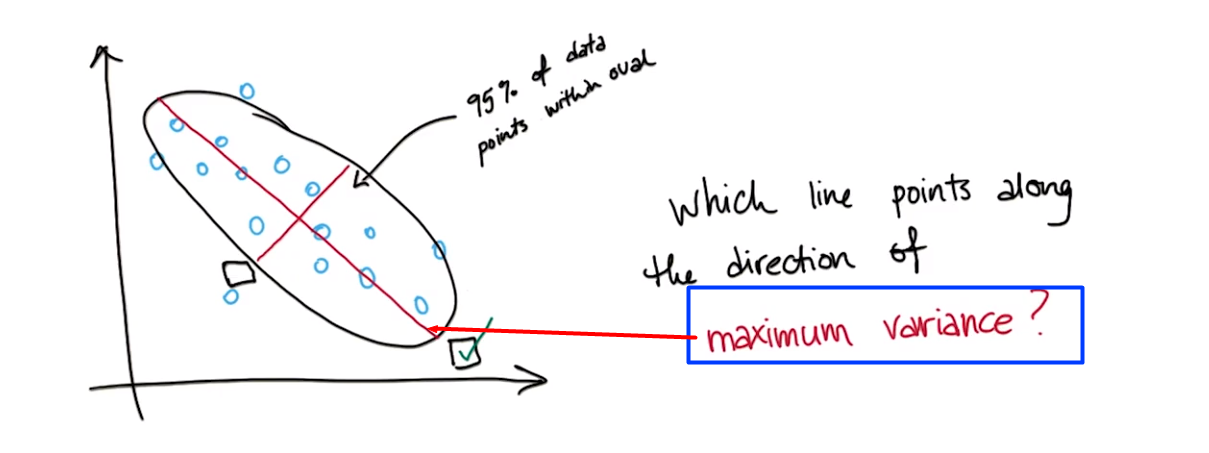
**最大方差**

1. 方差：

方差：数据离散程度的度量。也就是说：方差大：说明样本值分布范围广。方差小：，说明样本值 是紧密的聚集在一起。

2、主成分

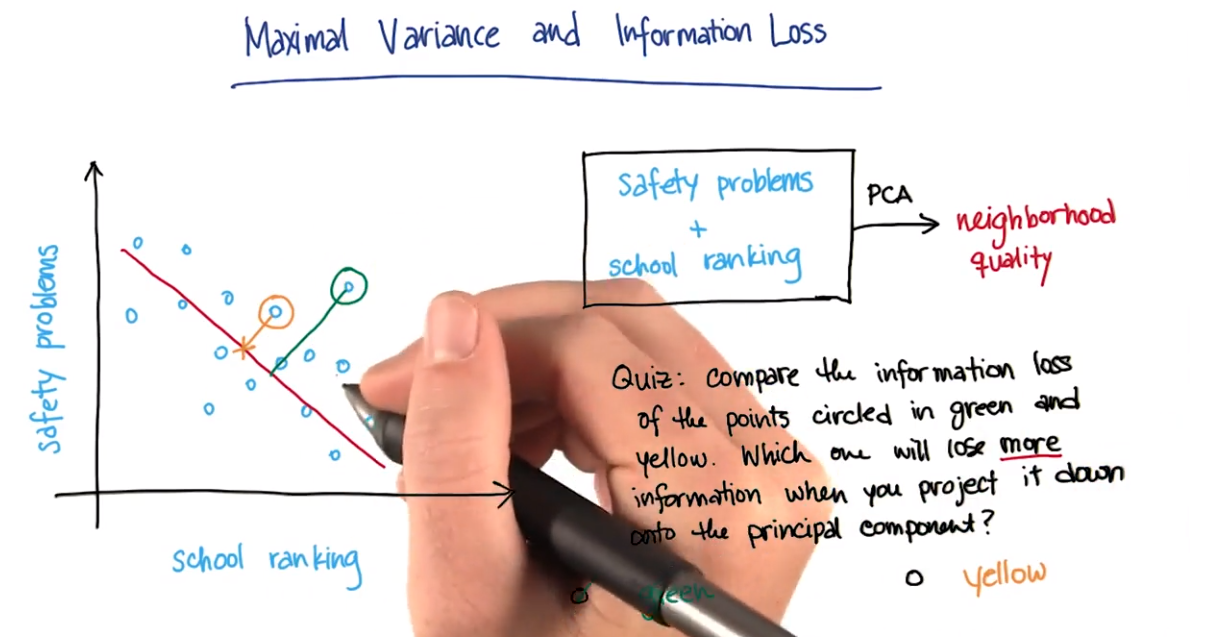
即 最大方差的方向，既图中的红线。



**为什么是数据集中 方差最大的方向 是主成分?**

是因为沿着 最大方差的维度 进行映射时 能够 最大程度保留 原始数据 中 最多的信息。

**3、信息损失**



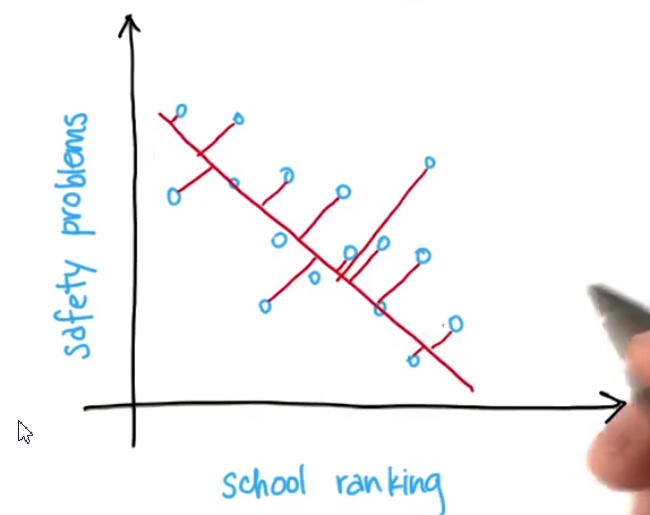
1.这条一维红线就是 主成分，当将这些二维的点投射到，主成分上时，就会出现信息损失。

2.损失的信息量 等于 某个点 与 主成分这条线 映射 的距离，因此绿色点 比 黄色点 信息损失更大。

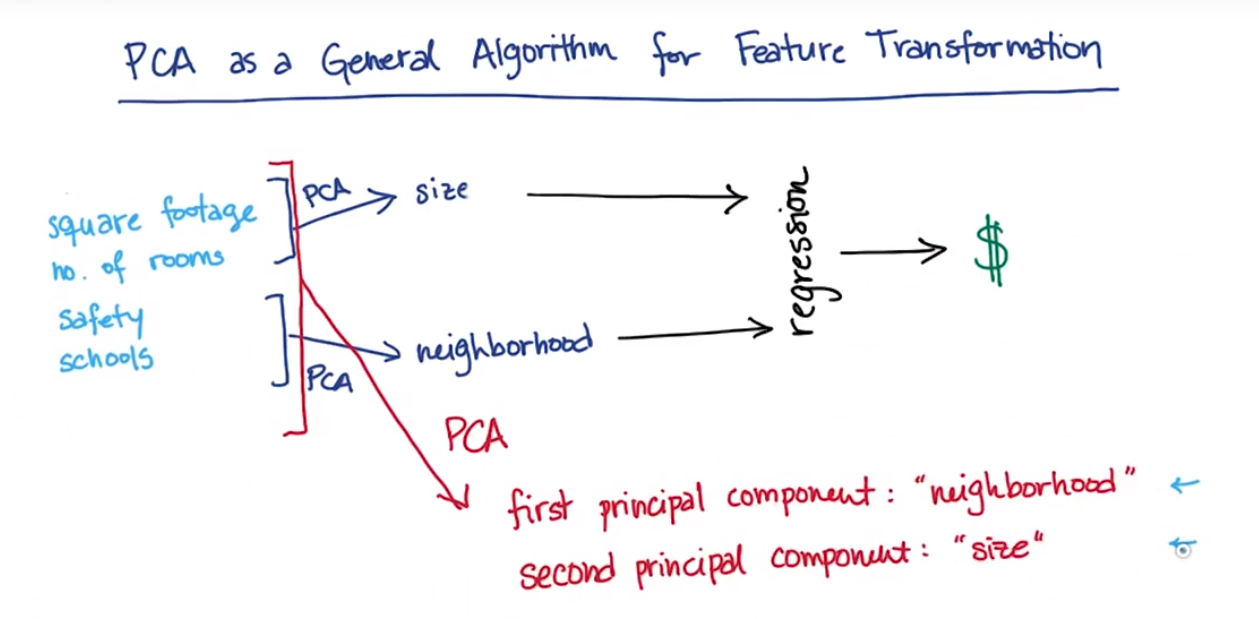
3.图中，绿色的数据点 的 信息损失最大，因为绿点到红线的距离 大于 黄点到红线的距离。

**4、信息损失与主成分**

将 方差最大化 时，就是 将 所有点 与主成分（图中红线）上的投影 距离 的 最小化。这样做是为了 将信息丢失的可能 降到最低。



**用于特征转换的PCA**



将所有这四个特征一块放入 PCA 中，它可以自动将这些特征结合成一些新的特征

并对这些新特征的相对能力进行 划分等级

如果案例中有两个隐藏特征 推动数据中大部分的变化 那么 PCA 将选出这些特征 并将其作为**第一个和第二个主成分（第一个主成分是混合产生的，可能包含所有特征或多或少的元素）**

**最大主成分数量**

sklearn 允许的范围内，PCA 实现方式所能找到的 主要成分 的 最大数量是多少？

因为它等于训练点数量和特征数量的最小值。

**PCA总结**

1.PCA 是将 输入特征 转化为其 主成分的 系统化方式。

2、这些主成分可当做新的输入特征。例如可将其用作回归或分类任务中的新特征。

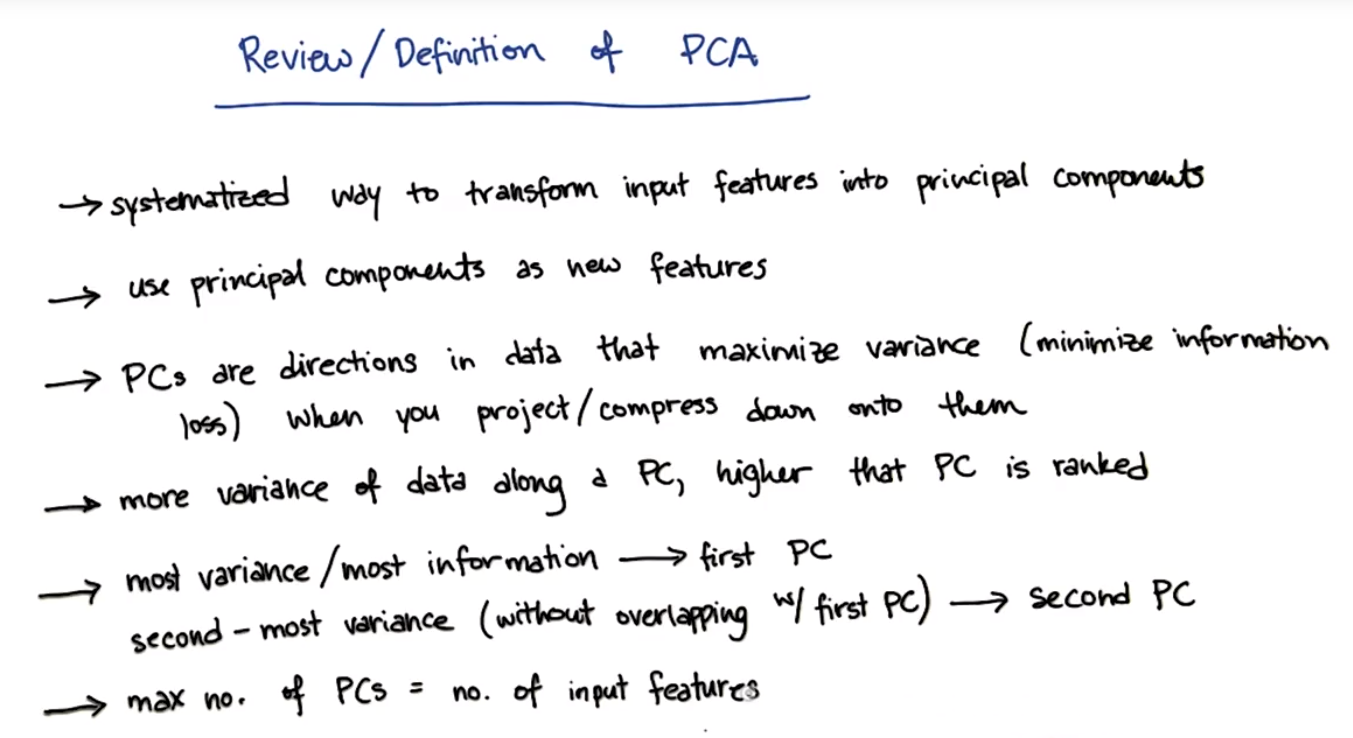
3、主成分的定义是 数据中会使用 方差最大化的方向，它可以在你对这些 主成分执行投影或压缩时 将出现信息丢失的可能性降至最低。

4、你还可以对主成分进行划分等级。

数据因主成分而产生的方差越大，那么该主成分的级别越高。因此产生方差最大的主成分即为第一个主成分，产生方差第二大的则为第二个主成分，依次类推。

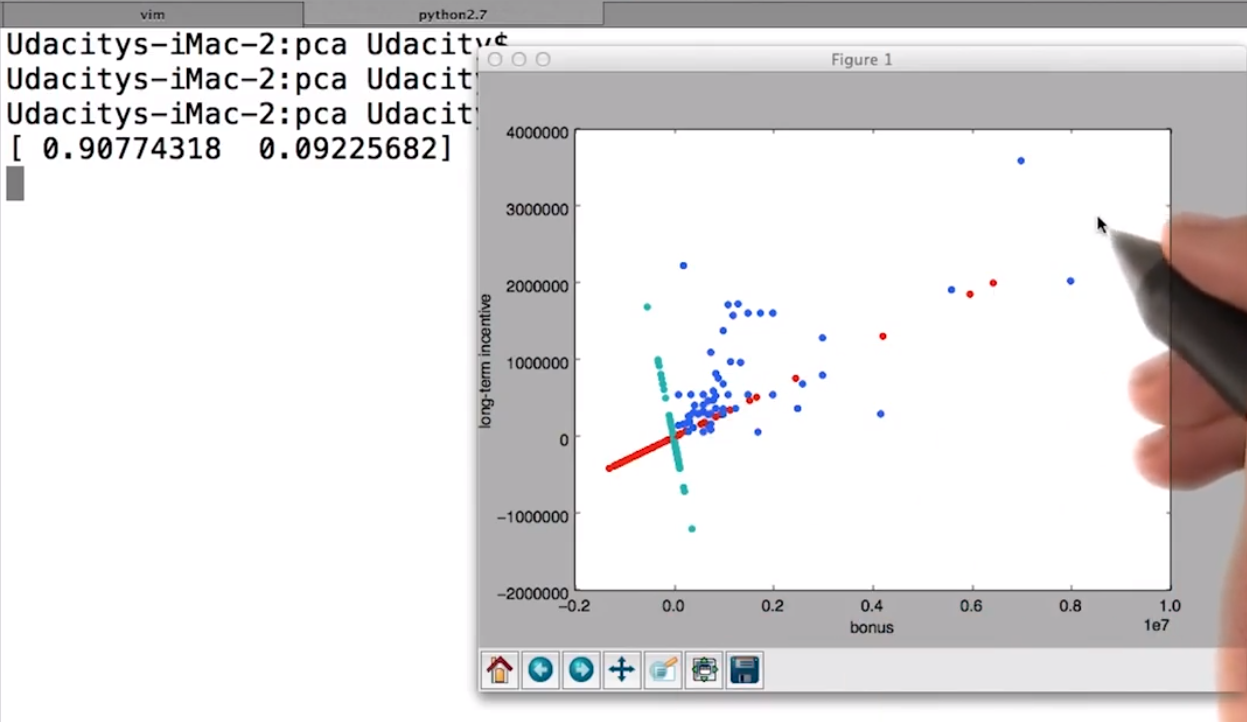
5、主成分在某种意义上是，相互垂直的，因此第二个主成分绝不会与第一个主成分重叠，第三个也不会通过第二个与第一个重叠等，因此在某种意义上您可以将它们当做单独的特征对待。

6、主成分数量是有上限的，该最大值 等于 数据集中输入的特征数量。通常情况下，只会使用前几个主成分。



**Sklearn 中的PCA**





第一个是已经介绍过的方差比，它其实是特征值的具体表现形式。

通过打印此处的该行内容，我借此认识到第一个主成分占数据变动的 90%，而第二个主成分的占比约为 9%

然后我要做的第二件事是研究第一个和第二个主成分。

无论出于何种目的，我都必须要执行数据转换这种操作。

第一行是红色。我将标绘出第一个主成分，该主成分所有点的位置以及主成分的方向。

然后我会访问蓝绿色或青色的第二个主成分，并将初始数据标绘成蓝色。

**何时使用PCA**

latent features driving the patterns in data

第一种情况是确定数据是否存在隐藏的特征,也就是想知道第一个主成分的大小。

例如：是否可以估量出 Enron 的大亨是谁。

dimensionality reduction

第二种 情况就是降维。

1、帮助可视化高维数据。

2、去除噪音（使用 PCA 降低输入特征的维数）。

比如分类算法，则算法的方差非常高，最终会被数据中的噪音同化，导致运行非常慢。

1. 是在使用另一个算法前使用 PCA 进行预处理，也就是归纳或分类任务。（例如人脸识别）

**用于人脸识别PCA**

什么PCA在人脸识别中有不错的应用呢？

也就是为什么这个问题与解决方案适用于每个人呢？

原因为：

1、人脸图片具有较高的维数。换言之 如果一幅人脸图片中有很多像素

那么其中任何像素对于辨别这张脸属于谁，都是重要的。

2、人脸具有一些一般性形态，这些形态 可以 以较小维数 的方式捕捉。

比如人一般都有两只眼睛，眼睛基本都位于接近脸的顶部的位置。

举例

以一百万像素的图片举例

1、一张一百万像素的人脸图片，那就有一百万个输入特征。因此缩减非常重要，SVM很难处理。

2、在两张人脸图像像中 并不是一百万个点都存在差异，而是只有几个主要的差异点。

可以利用 PCA 挑选出这些点，并发挥出它们最大作用。