# Principle Component Analysis

# A professional nose

# Ex: Fine fragrance

- 12 experts were asked to rate 12 perfumes on 11 scent adjectives.

```
##  [1] "spicy"    "heady"    "fruity"   "green"    "vanilla"  "floral"
##  [7] "woody"    "citrus"   "marine"   "greedy"   "oriental"
```

- Each rating is on the scale 1 - 10.
- Ratings for each perfume were averaged across experts.

# Ex: Fine fragrance, cont.

```
head(experts)
```

```
## # A tibble: 6 x 12
##   perfume spicy heady fruity green vanilla floral woody citrus marine
##   <fct>   <dbl> <dbl>  <dbl> <dbl>   <dbl>  <dbl> <dbl>  <dbl>  <dbl>
## 1 Angel   3.22  8.26   1.9   0.133   7.75    2.09 1.05   0.142 0.125
## 2 Aromat… 7.41  8.17   0.575 0.35    1.75    3.71 3.39   0.375 0.0583
## 3 Chanel… 3.93  8.42   1.18  0.5     1.73    4.66 1.02   0.6   0.05
## 4 "Cin\x… 0.983 2.07   5.2   0.267   4.18    5.32 1.25   0.775 1.02
## 5 Coco M… 0.925 0.717  4.58  1.2     2.02    7.31 1.13   1.17  1.14
## 6 J'ador… 0.108 1.03   6.85  1.62    0.183   8.51 0.925  2.13  1.91
## # … with 2 more variables: greedy <dbl>, oriental <dbl>
```

# Learning structure

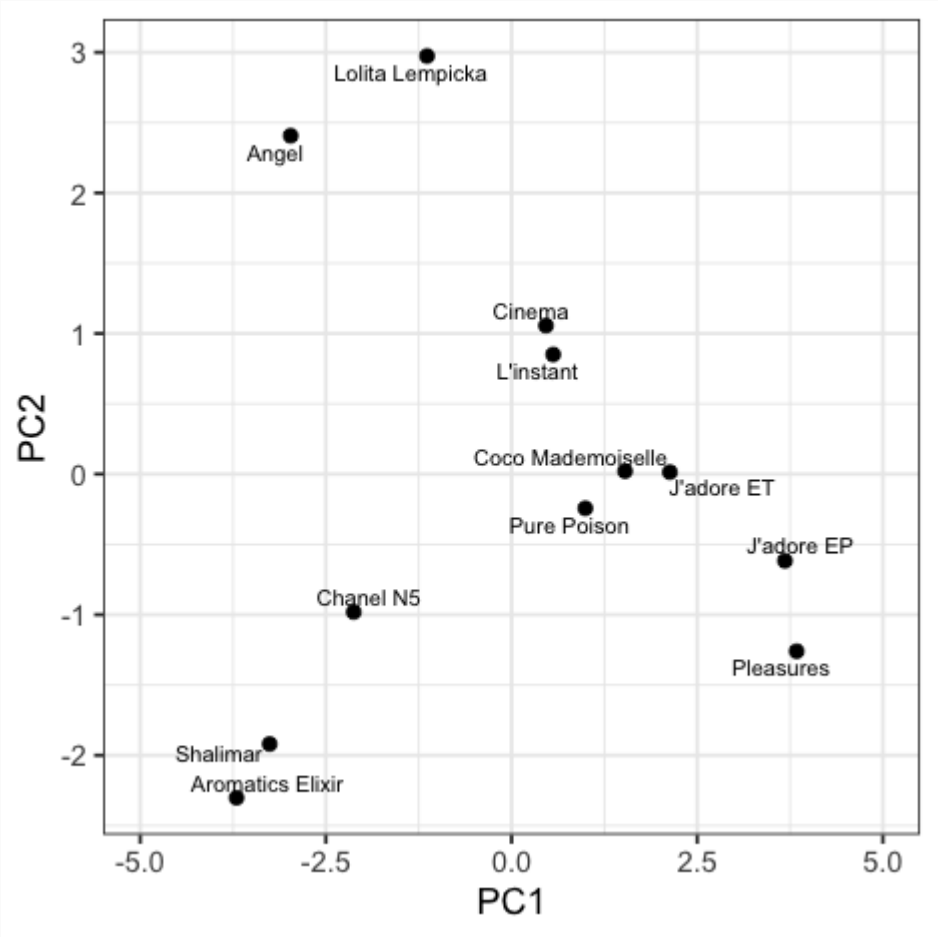
How do we visualize this data set?

- Representing *all* of the structure in the data requires $p = 11$ dimensions.

$$\binom{11}{2} = 55 \text{ possible scatterplots}$$

- Can we represent *most* of the structure using fewer dimensions?

# A particularly useful scatterplot



# Principle Component Analysis (PCA)

Produces a low-dimensional representation of a dataset. It finds a sequence of linear combinations of the variables that have maximal variance and are mutually uncorrelated.

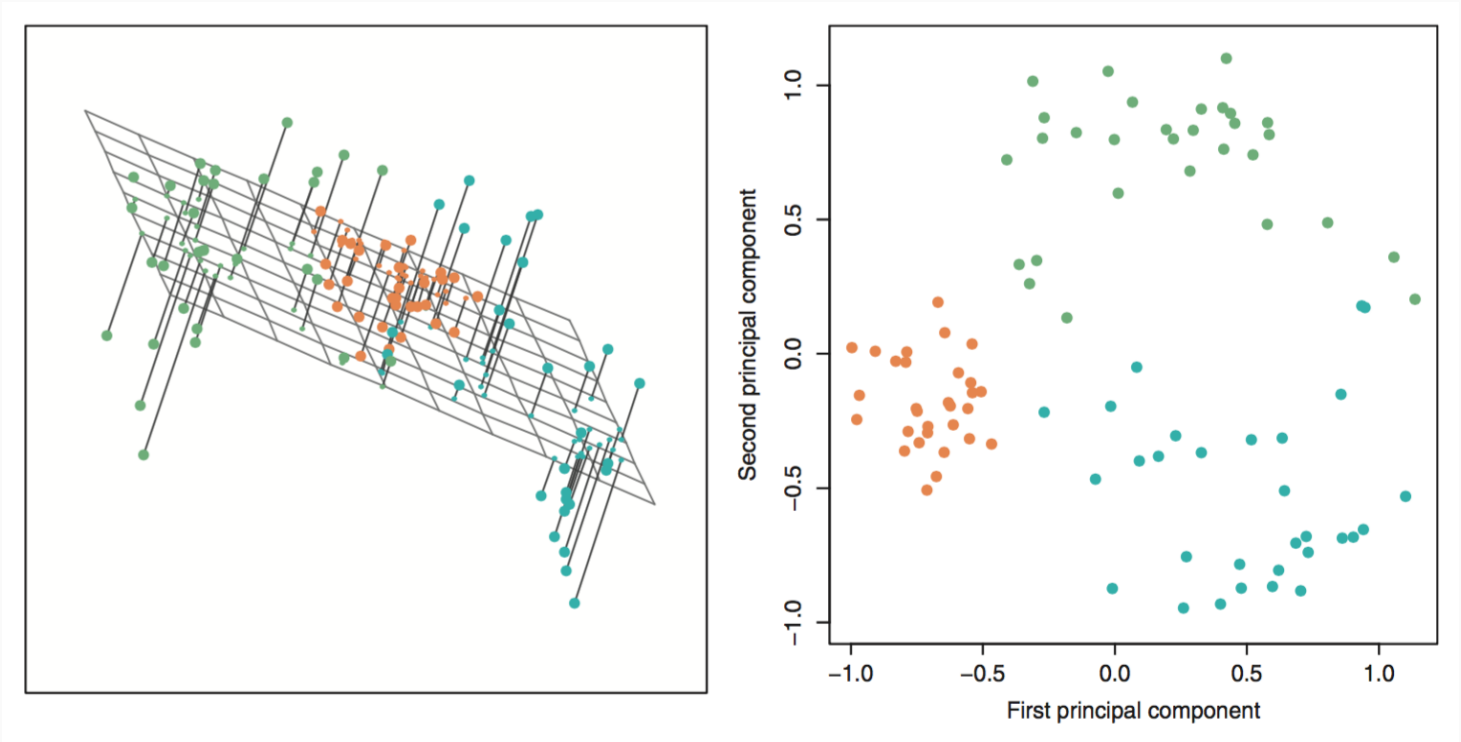
Used to:

- Visualize structure in data
- Learn about latent meta-variables
- Produce imputs for subsequent supervised learning

# boardwork

# Dimension Reduction

Reducing from $p = 3$ to 2 principal components.



# Finding PCs

For each component, we want the $\phi$ vector that solves the optimization problem:

$$\max \left( \frac{1}{n} \sum_{i=1}^{n} z_{i1}^2 \right) \text{ subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

where each $z_{i1} = \sum_{j=1}^{p} \phi_{j1} x_{ij}$.

Can be solved via an eigen decomposition ( $z_i$: eigenvectors of the covariance matrix of $X$).

# Interpretation

The weights corresponding to a PC, $\phi_{j1}$, are called the *loadings*.

**What does PC1 represent?**

```
##       spicy       heady      fruity       green     vanilla      floral
## -0.32374624 -0.35203437  0.34000811  0.30379083 -0.19229442  0.34403713
##       woody      citrus      marine      greedy    oriental
## -0.25228410  0.32991738  0.32175400 -0.08503758 -0.35323834
```

Bright vs Dark?

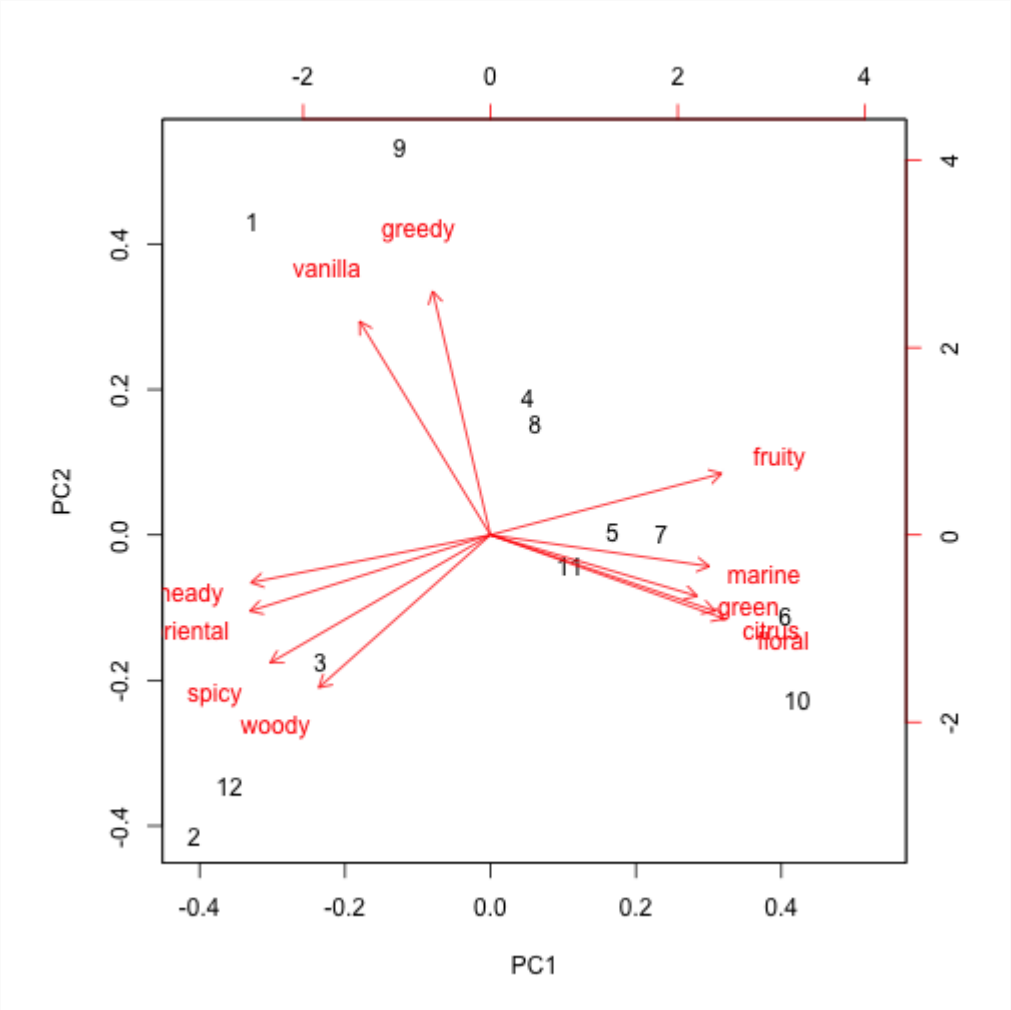
# Interpretation

## What does PC2 represent?

```
##       spicy       heady      fruity       green     vanilla      floral
## -0.30665034 -0.11432681  0.14713276 -0.14667649  0.51171193 -0.20104293
##       woody      citrus      marine      greedy    oriental
## -0.36590048 -0.18273912 -0.07533076  0.58413221 -0.18249284
```

Mellow vs piquant?

# Biplot


-2
0
2
4
9
1
greedy
vanilla
4
8
fruity
5
7
11
marine
green
6
citrus
floral
3
spicy
woody
10
12
2
0.4
0.2
0.0
-0.2
-0.4
PC2
-0.4
-0.2
0.0
0.2
0.4
PC1
4
2
0
-2

# Ex. More Crime

This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

```
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

# PCA

```
pca1 <- prcomp(USArrests, scale = TRUE)
names(pca1)
```

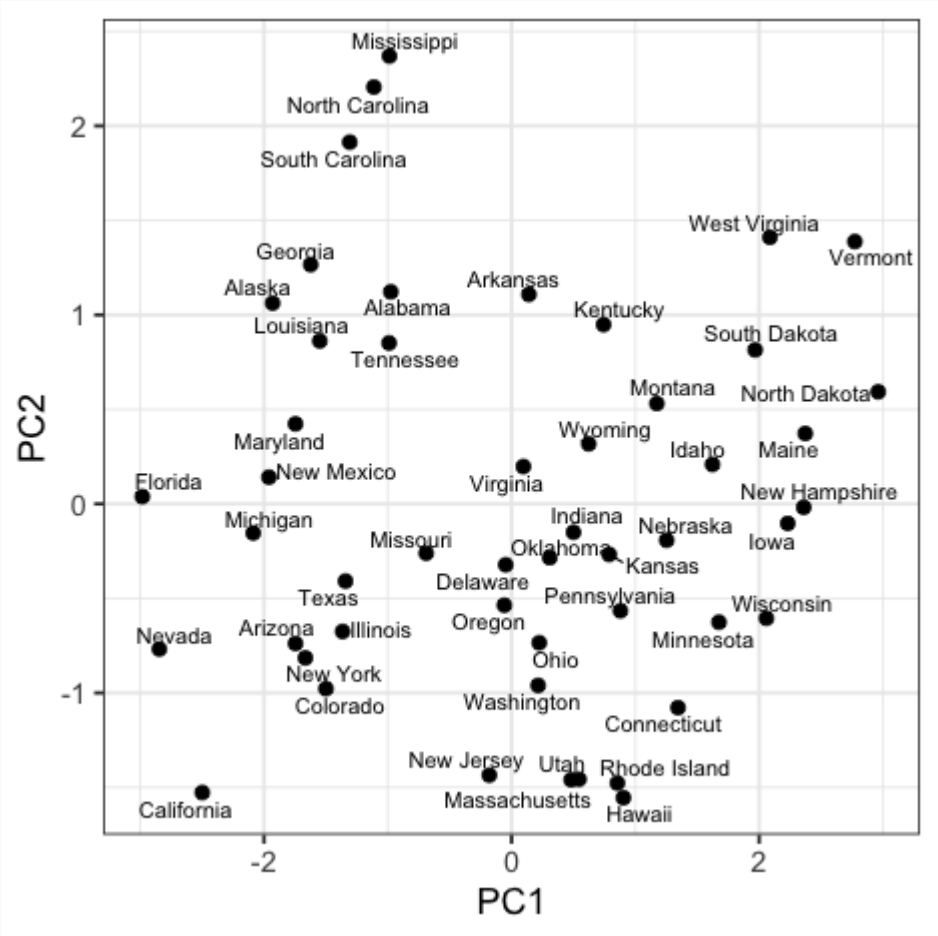
```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

- `rotation` holds the matrix of loadings; the $\phi$'s.
- `x` holds the scores for the principle components; the $z_{ij}$.

```
pca1$rotation
```

```
##                 PC1        PC2        PC3         PC4
## Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
## Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
## UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
## Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

# A particularly useful scatterplot



# Interpretation

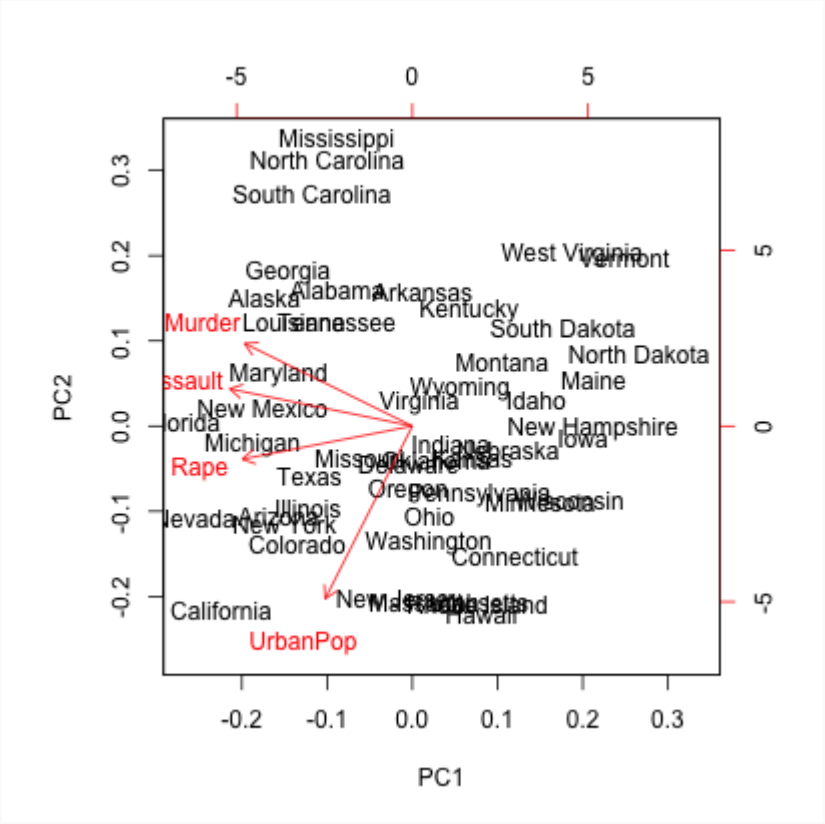
```
pca1$rotation
```

```
##                    PC1         PC2         PC3          PC4
## Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
## Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
## UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
## Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

- PC1: crime
- PC2: urbanization

# Biplot

```
biplot(pca1)
```

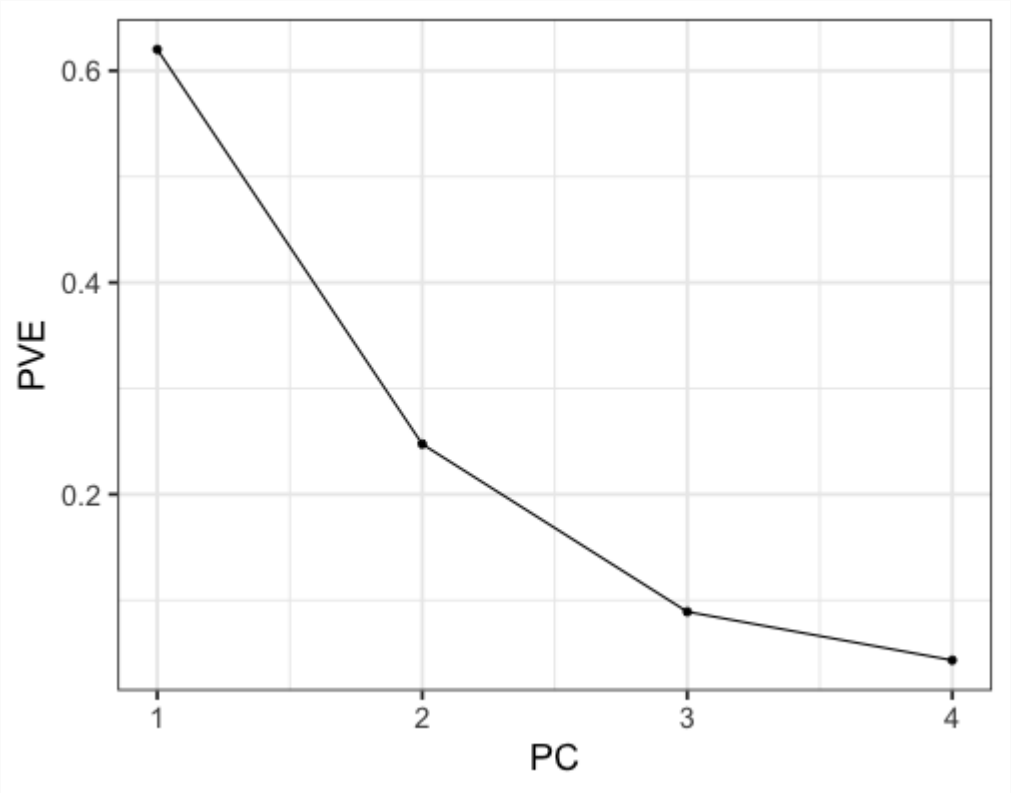


# Constructing a scree plot

Used to visualize the proportion of variance explained (PVE) by each PC.

```
d <- data.frame(PC = 1:4,
                PVE = pca1$sdev^2 /
                  sum(pca1$sdev^2))
ggplot(d, aes(x = PC, y = PVE)) +
  geom_line() +
  geom_point()
```

# Scree plot



# How many PCs?

- 1st PC: 62% PVE
- 1st + 2nd PC: 62 + 25 = 87% PVE

Usually most of the structure is in the first several principal components, but results may vary!

*Rule of thumb*: look for the elbow in the scree plot.