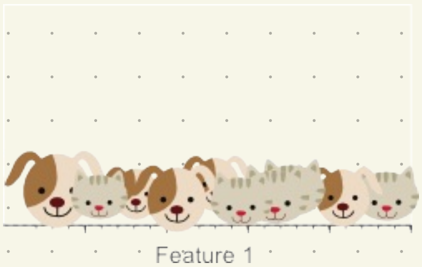
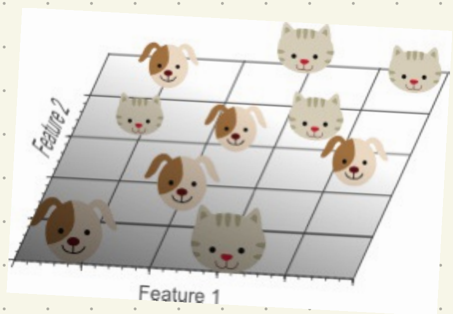
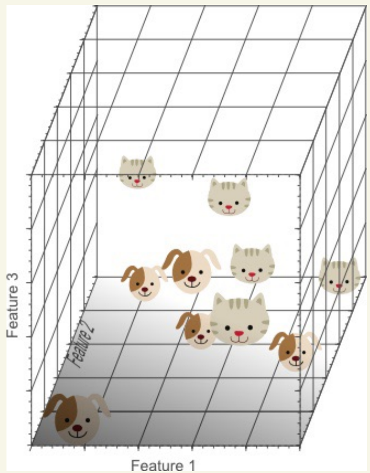
# Curse of dimensionality



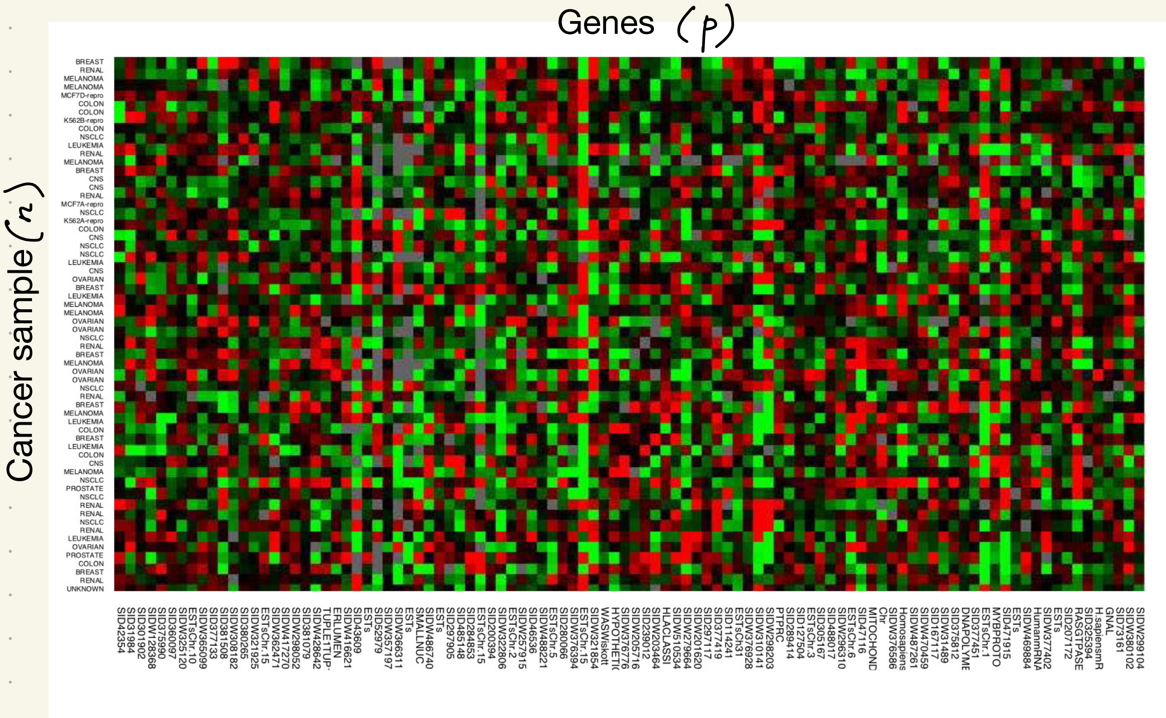




https://medium.com/@prasanNH/the-curse-of-dimensionality-in-data-analysis-b16ea6903611

Very often $p$(features) >> $n$(observations)

Exampee 1: DNA microarray data

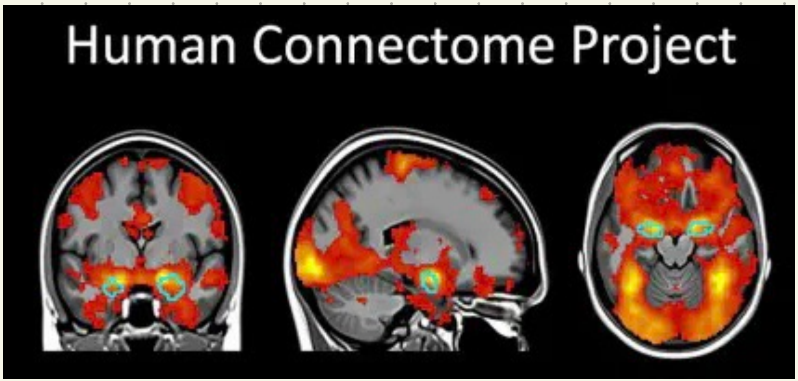



ESL II

green: underexpressed    red: overexpressed    grey: missing

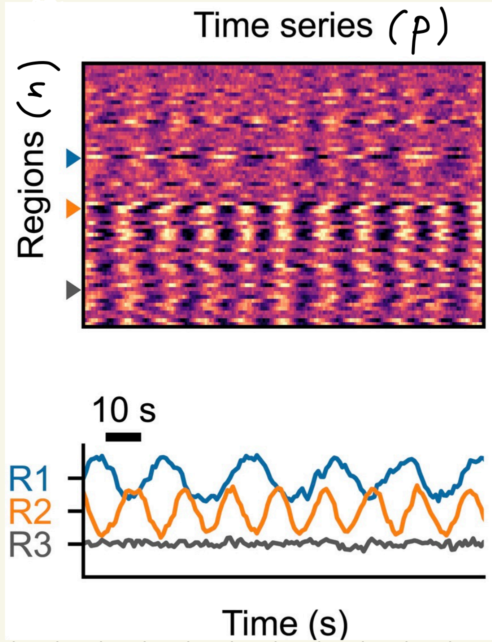
Very often $p$(features) $\gg$ $n$(observations)

Example 2: fMRI data




fMRI signal during emotion task

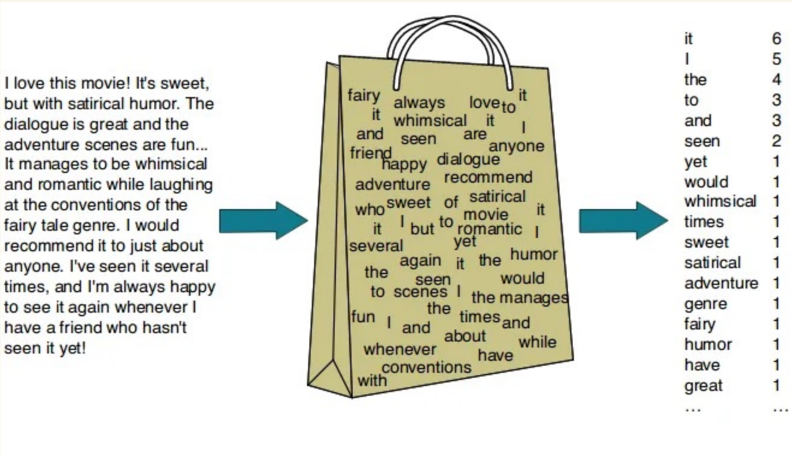
https://penghui-du.com/post/reading-notes-handbook-of-functional-mri-data-analysis-introduction/




https://www.science.org/doi/10.1126/sciadv.abq7547

Very often $p(\text{features}) \gg n(\text{observations})$

Example 3 : text data




Bag of words (BoW) representation

| | Movie reviews |
|---|---|
| **Review 1** | This movie is good. |
| **Review 2** | The movie is not good. |
| **Review 3** | I love this movie. Watch, you will love it too. |

Vocabulary ($p$)

Reviews ($n$)

| | This | Movie | Is | The | Good | Of | Times | Not | I | Love | Watch | You | Will | It | too |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Review 1** | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Review 2** | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Review 3** | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 1 | 1 | 1 | 1 | 1 |

https://medium.com/@vamshiprakash001/an-introduction-to-bag-of-words-bow-c32a65293ccc

## Why do we care?

- interpretability and visualijation
- noise level
- Computational issues
- combinatorics

Suppose we have $n = 1000$ patients and $p$ symtoms
(categorical: "mild", "moderate", "severe")
$x_1, \dots x_n \in \{0,1,2\}^p$, each $x_i$ takes one of $3^p$ values.

$p=3$ $\quad 1000/27 \simeq 37$ patients / symptom combination

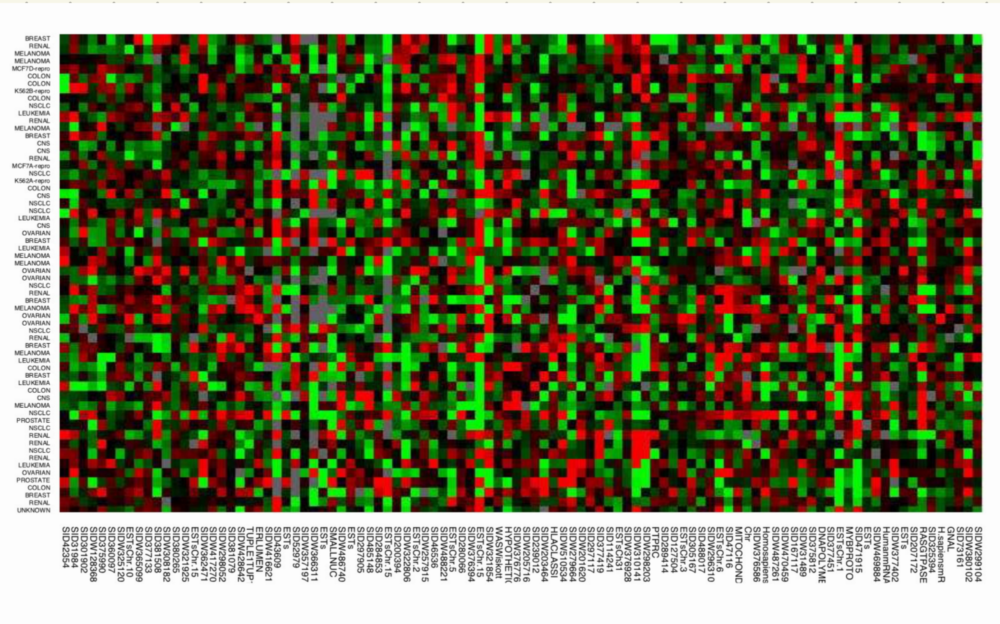
$p=6$ $\quad 1000/729 \simeq 1.4$ patients / symptom combination

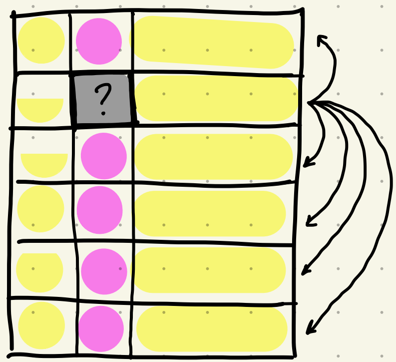
We need more observations to draw conclusions.

# What else?

Many ML / Stats methods are based on distances between observations.

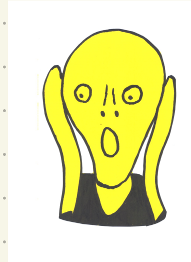
Example: impute missing values in the DNA data with the nearest neighbor.





Distance behaviour is wierd in high dimensions.

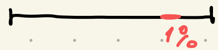
(1) In high-dimensional space nobody can hear you scream



Assume $n=100$ points live in a cube $[0,1]^p$

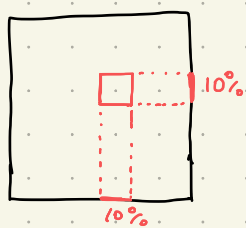
$p=1$ each observation takes $1/100 = 0.01$ of the length

i.e. 1% of the segment.



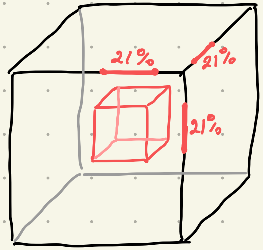

$p=2$ each observation takes $1/100 = 0.01$ of the area

i.e. a square of size $\sqrt{0.01} = 0.1 = 10\%$



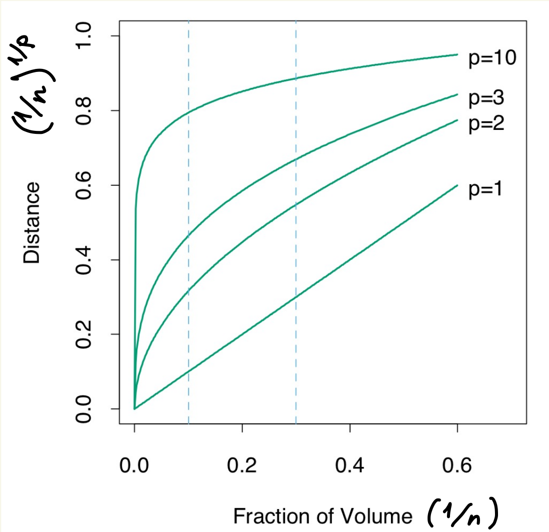
$p=3$ each observation takes $1/100 = 0.01$ of the volume

i.e. a cube of size $\sqrt[3]{0.01} = 0.21 = 21\%$

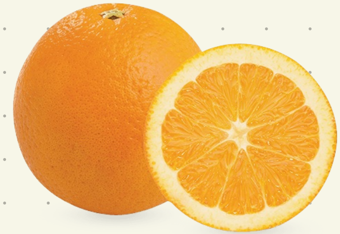



In general, to capture 1% of the volume we need a hypercube o size $\sqrt[p]{0.01}$.

$p>7$ The hypercube side is $> 50\%$.



## ② Orange peel

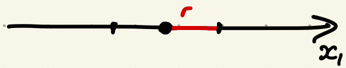


The ball of radius $r$ is

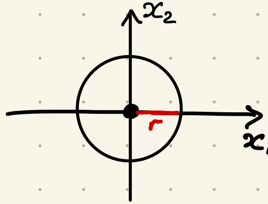
$$B_p(r) = \{ x \in \mathbb{R}^p : \|x\|_2 \leq r \}$$

Suppose $x_1 \dots x_n \overset{iid}{\sim} Unif(B_p(1))$

Then the median distance from the origin to the closest data point is



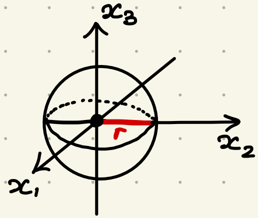

$$d(p,n) = \left(1 - \left(\frac{1}{2}\right)^{1/n}\right)^{1/p}$$




We want to find $r$ s.t.

$$P\left(\min_{i=1..n} \|x_i\|_2 > r\right) = \frac{1}{2}$$

$$P(\|x_i\|_2 < r) = \frac{Vol(B_p(r))}{Vol(B_p(1))} \quad \text{where } Vol(B_p(r)) = \frac{\pi^{p/2}\, r^p}{\Gamma\left(\frac{p}{2}+1\right)}$$



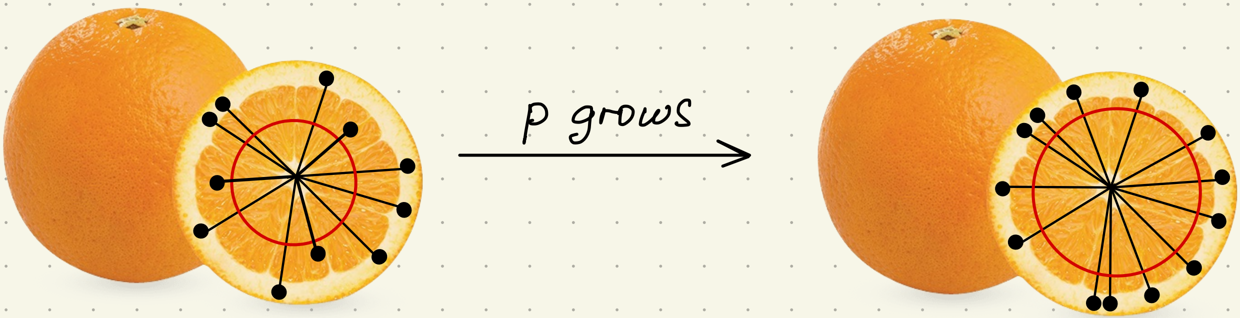

$$P(\|x_i\| < r) = r^p$$

$$P\left(\min_{i=1..n} \|x_i\|_2 > r\right) = \prod_{i=1}^{n} P(\|x_i\| > r) = (1 - r^p)^n = \frac{1}{2}$$

If $n=100$ then $d(p,n) > 0.5$ for $p > 7$.

Thus

- most points are close to the boundary
- the points in high-dim. space are isolated



③ In high dimensions all distances are similar

$$x \sim N_p(0, I) \qquad x = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}$$

- $E(\|x\|^2) = p$

| $E(\|x\|^2) = \sum_{i=1}^{p} E(x_i^2) = p \cdot E(x_i^2) = p$

- $Var(\|x\|^2) = 2p$

| $Var(\|x\|^2) = p \cdot Var(x_i^2) = p \cdot (\underbrace{E(x_i^4)}_{3} - \underbrace{E(x_i^2)}_{1})$

We can also show that:

$$|E(\|x\|) - \sqrt{p}| \leq \frac{1}{\sqrt{p}} \quad \text{and} \quad Var(\|x\|) \leq 2$$

That is, $\|x\| \longrightarrow \sqrt{p}$ and the spread is bounded

If $x, y \overset{i.i.d}{\sim} N_p(0, I)$ then $\underline{E}(\|x-y\|^2) = 2p$

$| \; E(\|x-y\|^2) = \sum_{i=1}^{p} E(x_i^2) - 2E(x_i y_i) + E(y_i^2) = 2p$

Moreover,

$|E(\|x-y\|) - \sqrt{2p}| \le \frac{1}{\sqrt{2p}}$ and $Var(\|x-y\|) \le 3$

More formally,

$$P(|\|x\| - \sqrt{p}| \ge \varepsilon) \le 2 e^{-c\varepsilon^2} \quad \forall \varepsilon \in [0, \sqrt{p}]$$

E.g. when $p \ge 100$ then $\|x\| \in \sqrt{p} \pm 10$

with probability 0.99.

④ In high dimensions two random vectors are orthogonal.

Let $x, y \sim N_p(0, I)$ and $a \in \mathbb{R}^p$ is constant.

- $E(\langle x, a\rangle) = 0$ and $Var(\langle x, a\rangle) = \|a\|^2$

$$E(\langle x, a\rangle) = \sum_{i=1}^{p} a_i E(x_i) = 0$$

$$Var(\langle x, a\rangle) = \sum_{i=1}^{p} a_i^2 \, Var(x_i) = \|a\|^2$$

- $E(\langle x, y\rangle) = 0$ and $Var(\langle x, y\rangle) = p$

$\langle x, y\rangle = \frac{1}{2}\left(\|x\|^2 + \|y\|^2 - \|x-y\|^2\right)$ so

$$E(\langle x, y\rangle) = \frac{1}{2}(p + p - 2p) = 0$$

$$Var(\langle x, y\rangle) = p \cdot Var(x_i y_i) = p \, Var(x_i) Var(y_i) = p$$

Combining 3 and 4

$$\|x\|^2 \simeq p \pm \sqrt{2p} \qquad \|y\|^2 \simeq p \pm \sqrt{2p}$$

$$\langle x, y \rangle \simeq 0 \pm \sqrt{p}$$

The cosine between $x$ and $y$ is

$$\frac{\langle x, y \rangle}{\|x\| \, \|y\|} \simeq \frac{0 \pm \sqrt{p}}{\left(\sqrt{p \pm \sqrt{2p}}\right)^2} = \frac{0 \pm \sqrt{p}}{p \pm \sqrt{2p}} \longrightarrow 0 \quad \text{for } p \to \infty$$

Formally, $\forall \ \varepsilon > 0$ and $p \geq 1$

$$P\left( \left| \left\langle \frac{x}{\|x\|}, \frac{y}{\|y\|} \right\rangle \right| \geq \varepsilon \right) \leq \frac{2/\varepsilon + 7}{\sqrt{p}}$$

If $x \sim N_p(0, I)$ what is the distribution of $\frac{x}{\|x\|}$?

$p = 1$ then $\frac{x}{|x|} = \text{Sign}(x) \sim \text{Unif}(-1, 1)$

$p = 2$

$\frac{x}{\|x\|} \sim \text{Unif}(\text{circle})$

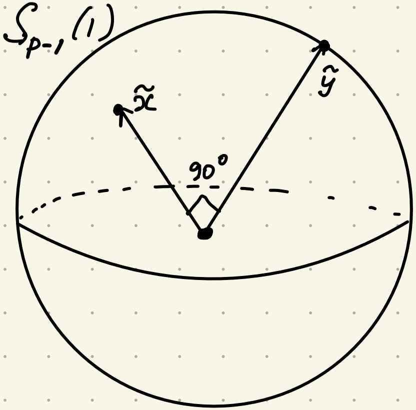
For generae $p$, if $x \sim N_p(0, I)$ the density $f(x) \propto e^{-\frac{1}{2}\|x\|^2}$ depend on the $\|x\|$ but not the direction.

Moreover, for any $Q$ orthogonal $Qx \sim N_p(0, I)$ thus $N_p(0, I)$ is rotation invariant and $x$ has the same probability to point in any direction.

Thus $\frac{x}{\|x\|} \sim \text{Unif}(S_{p-1}(1))$ where $S_{p-1}(r) = \{x \in \mathbb{R}^p : \|x\| = r\}$

$\tilde{x} = \frac{x}{\|x\|}$, $\tilde{y} = \frac{y}{\|y\|}$ are uniform on a sphere $S_{p-1}(1)$

where $S_{p-1}(r) = \{ x \in \mathbb{R}^p : \|x\| = r \}$




$$\frac{\langle x, y \rangle}{\|x\| \, \|y\|} = 0 \quad \text{means} \quad \tilde{x} \perp \tilde{y}$$

In other words, in high dimensions two random $x, y$ on a sphere are orthogonal

| | $\|x\|$ | $\|x-y\|$ | $\langle x, y \rangle$ |
|---|---|---|---|
| $x, y \sim N_p(0, I)$ | $\simeq \sqrt{p}$ | $\simeq \sqrt{2p}$ | $\simeq 0$ |
| $x, y \sim \text{Unif}(S_{p-1}(1))$ | $= 1$ | $\simeq \sqrt{2}$ | $\simeq 0$ |