Lecture 7

K-means

Find k points

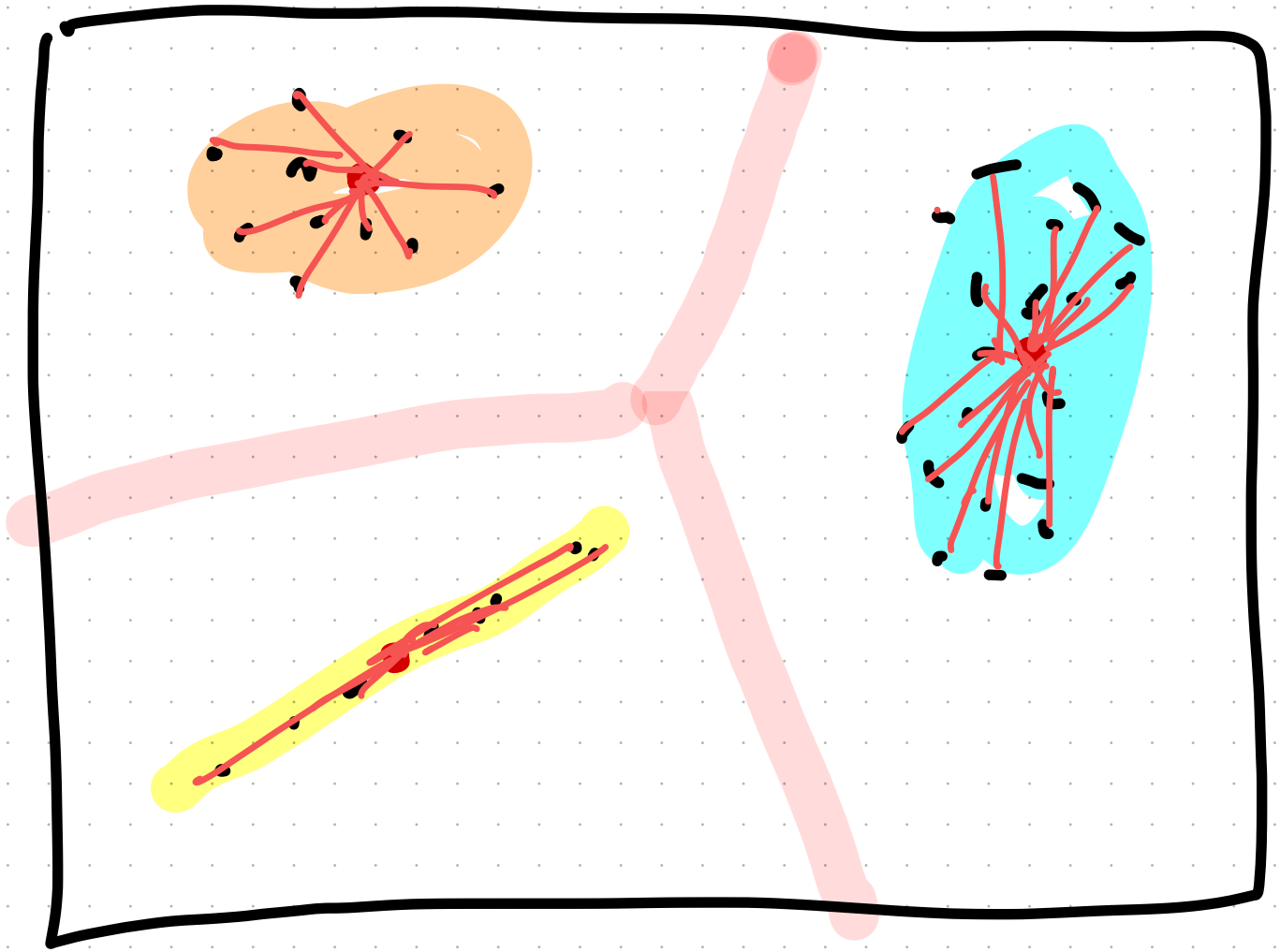
$m_1, m_2 \dots m_k$

That minimize

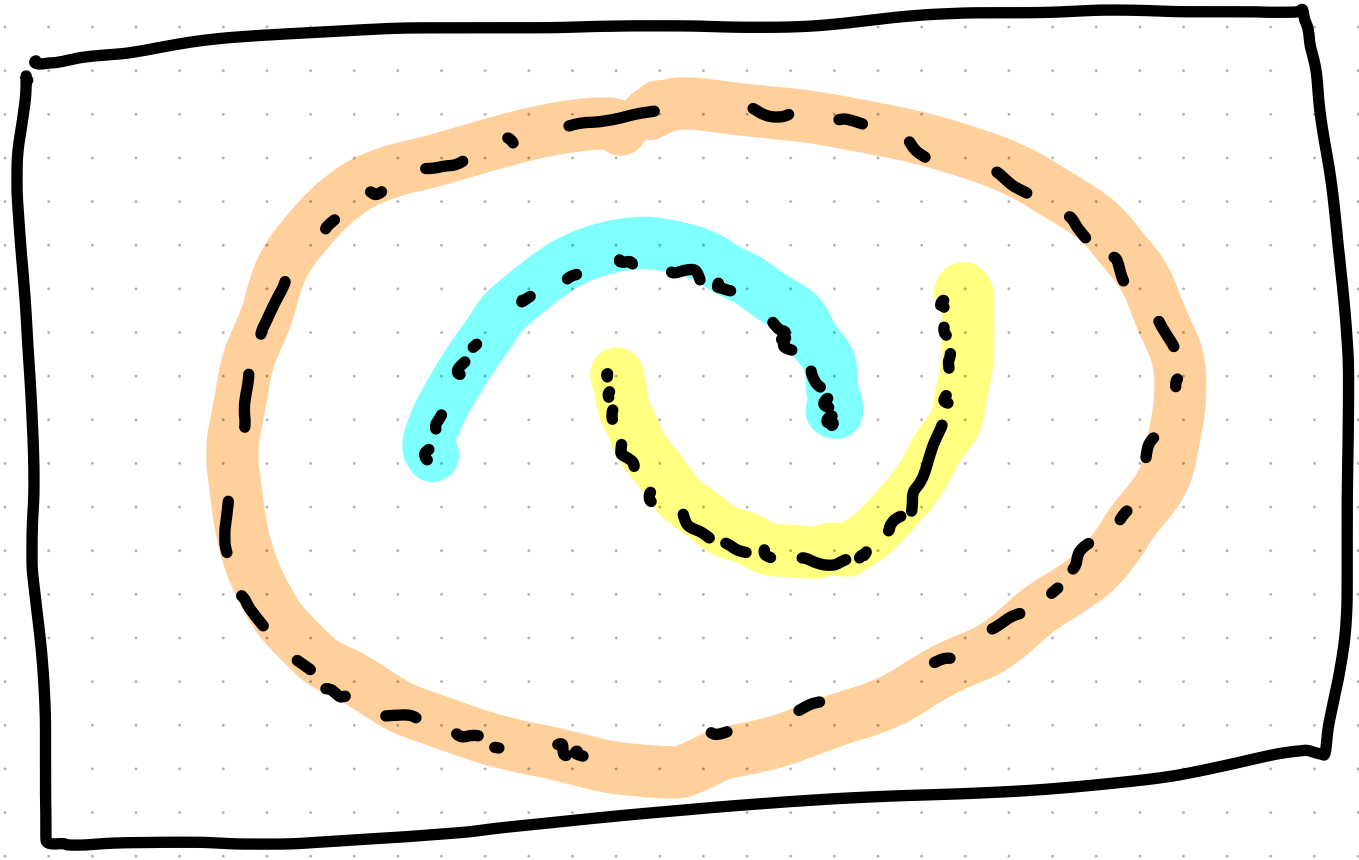
$$\sum d\left(x_i, \underset{M}{\text{closest}}(x_i)\right)$$

NP-Complete:

we think there is no $O(n^c)$ algorithm for any c



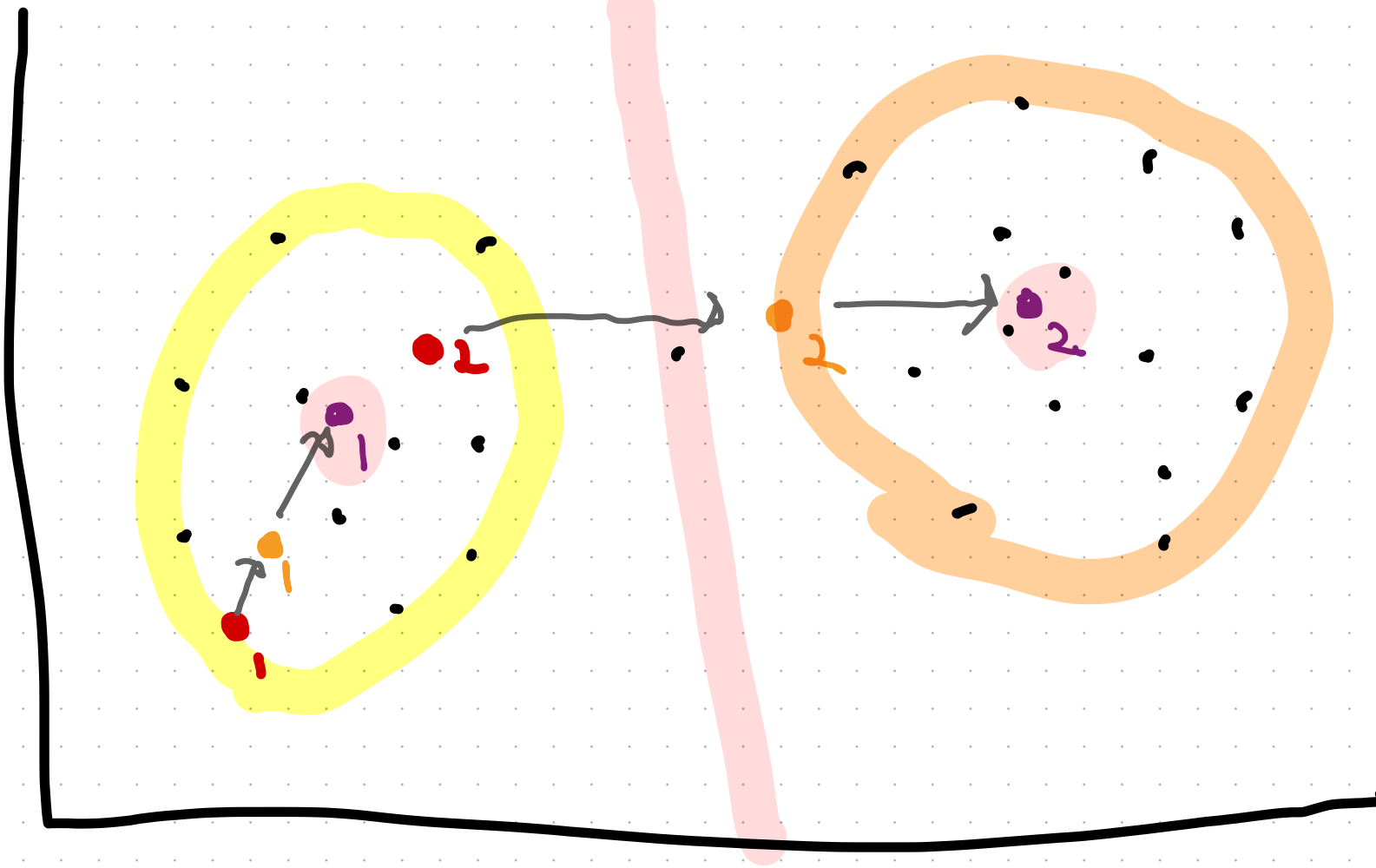
Clustering

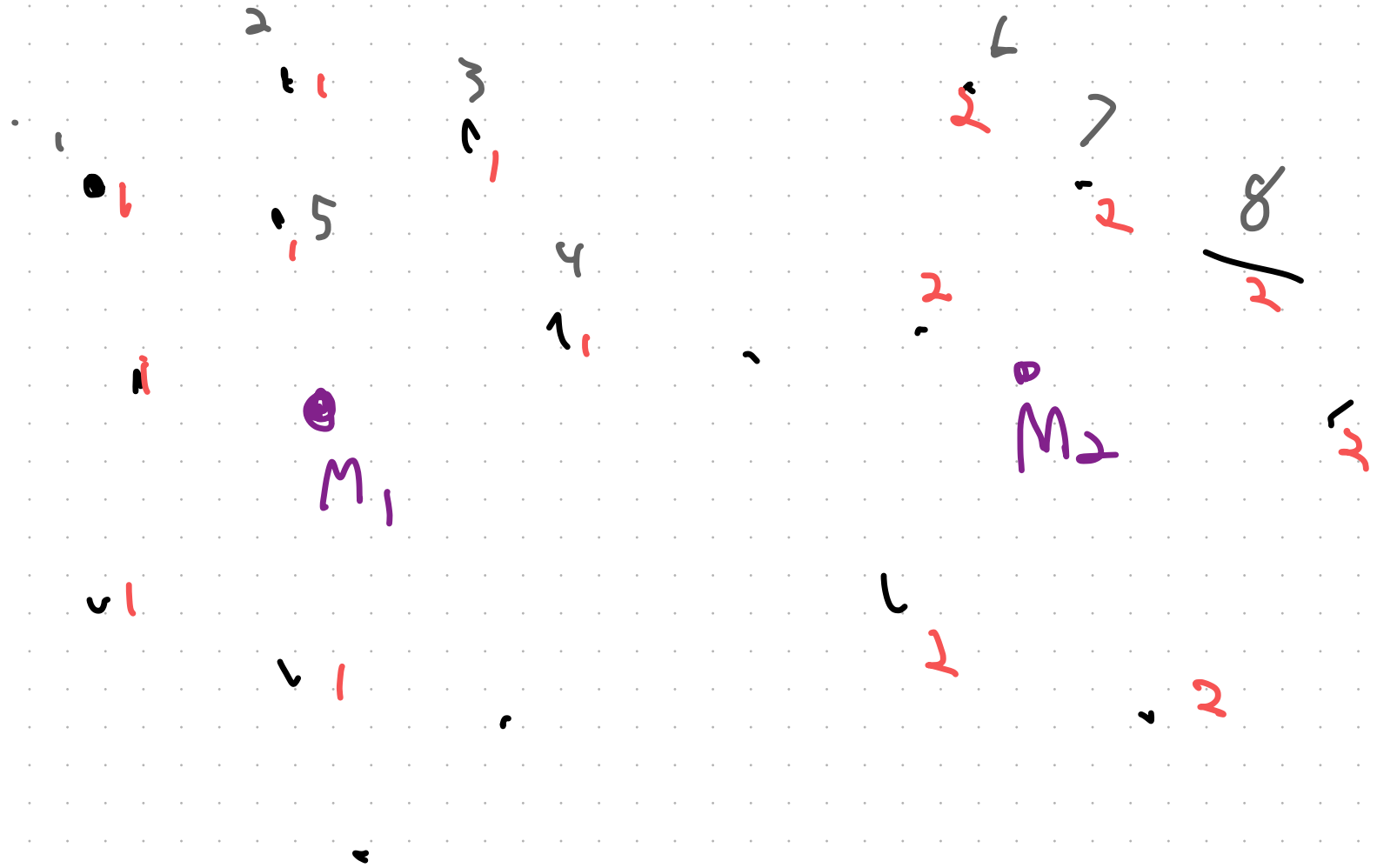


K=2

For each point
group according
to Clasest mean

Compute the
Centroid of
each group





P:

Runtime

n: Number of data points

d: Dimension

i: # of iterations to converge

t: # of different trials

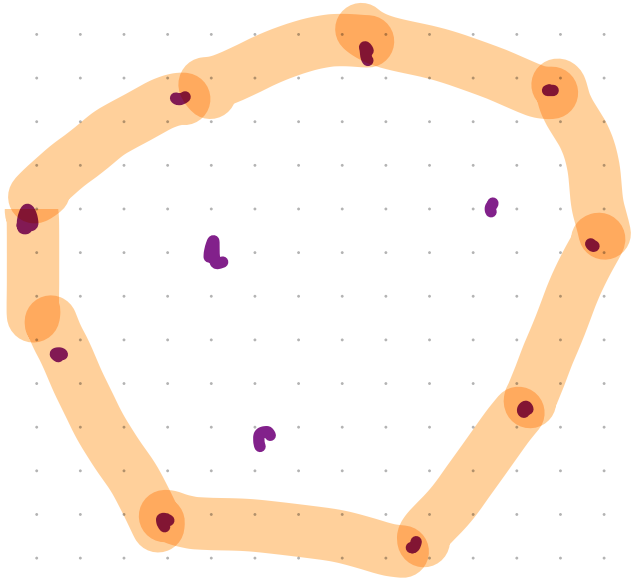
k: # of means

O(nkd):
- $n \cdot k \cdot d$ — for each point find closest mean
- $n \cdot k$ — for each mean group closest point
- $n \cdot d$ — Compute centroid

Not too big

Total: $O(NKdtI)$

"Curse of dimensionality"



$n^{\lfloor \frac{d}{2} \rfloor}$


quality
Elbow
K

Dimensionality Reduction

Points in high dimension

Points in a lower dimension

Preserve distance

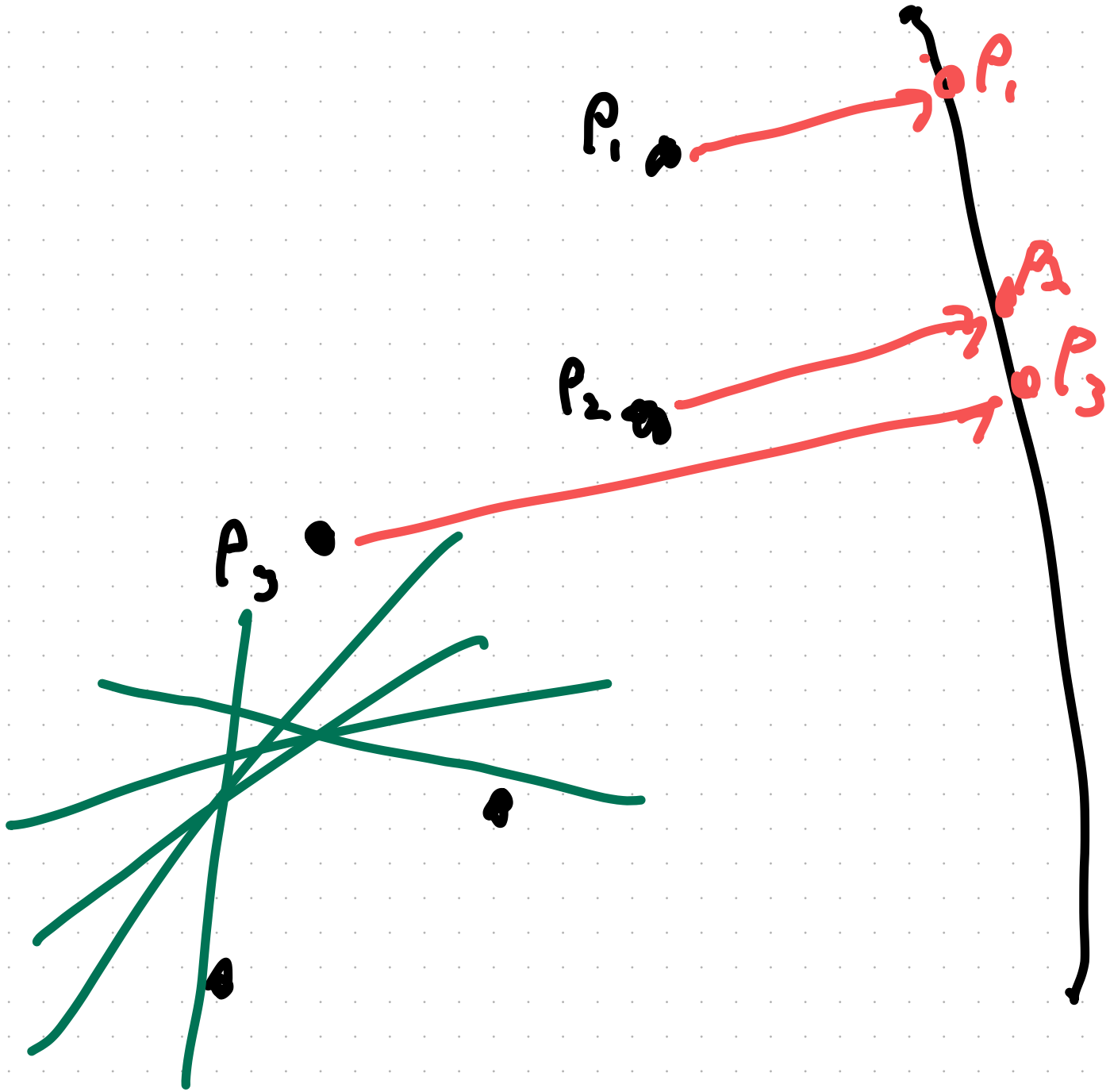
Johnson - Lindenstrauss (JL)

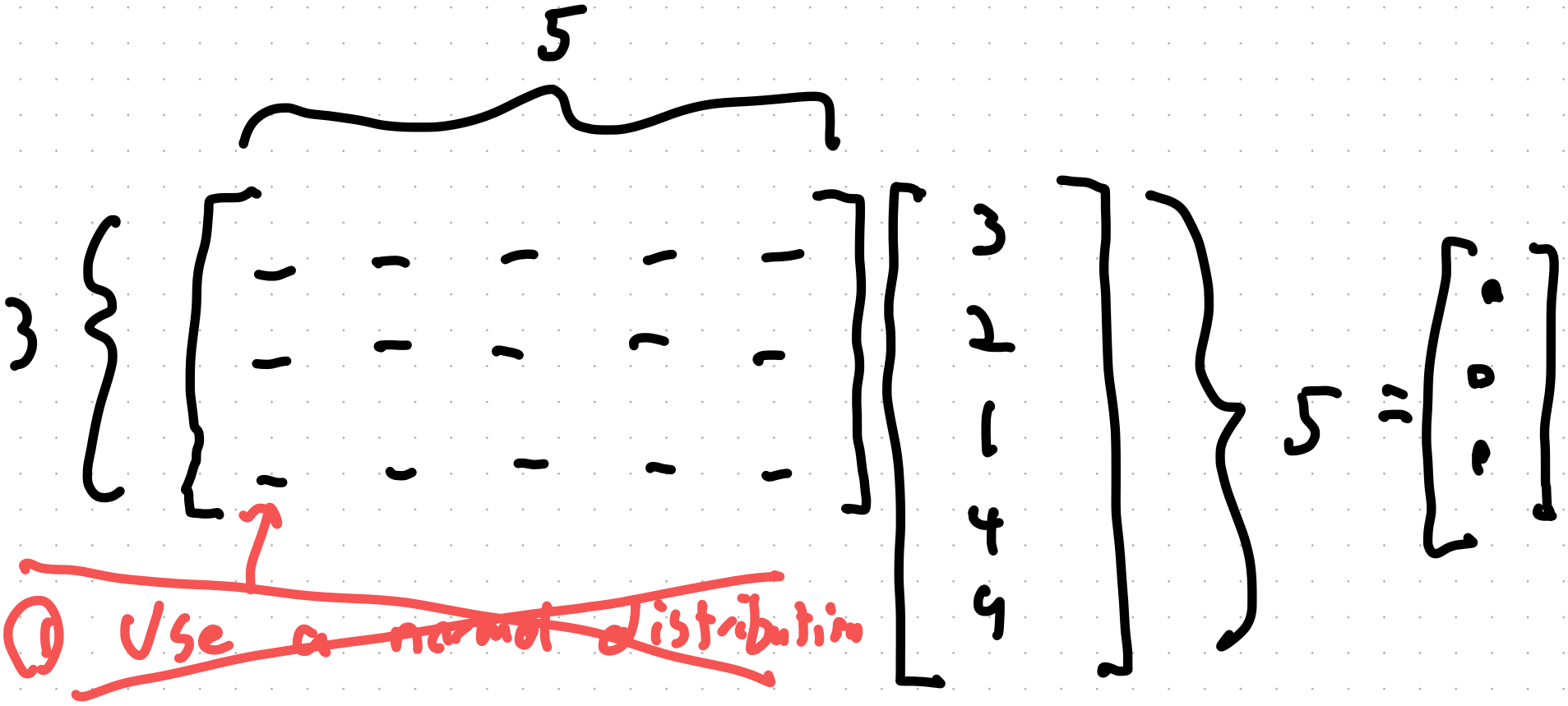
Have a set of S points in $\mathbb{R}^d$ (d dimensions)

Then there is a function $JL(x)$ such that
$JL(x)$ is K dimensional
and $K \geq \frac{20 \ln n}{\epsilon^2}$

for any x, y in S:

$$(1-\epsilon)d(JL(x), JL(y)) \leq d(x,y) \leq (1+\epsilon)d(JL(x), JL(y))$$


P_1
P_1
P_2
P_2
P_3
P_3



① Use a normal distribution

Pick -1 1 with 50% chance

if $k \geq \frac{4+2\beta}{\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{3}} \ln n$

Then $-1/+1$ works with prob $\geq 1-\beta$

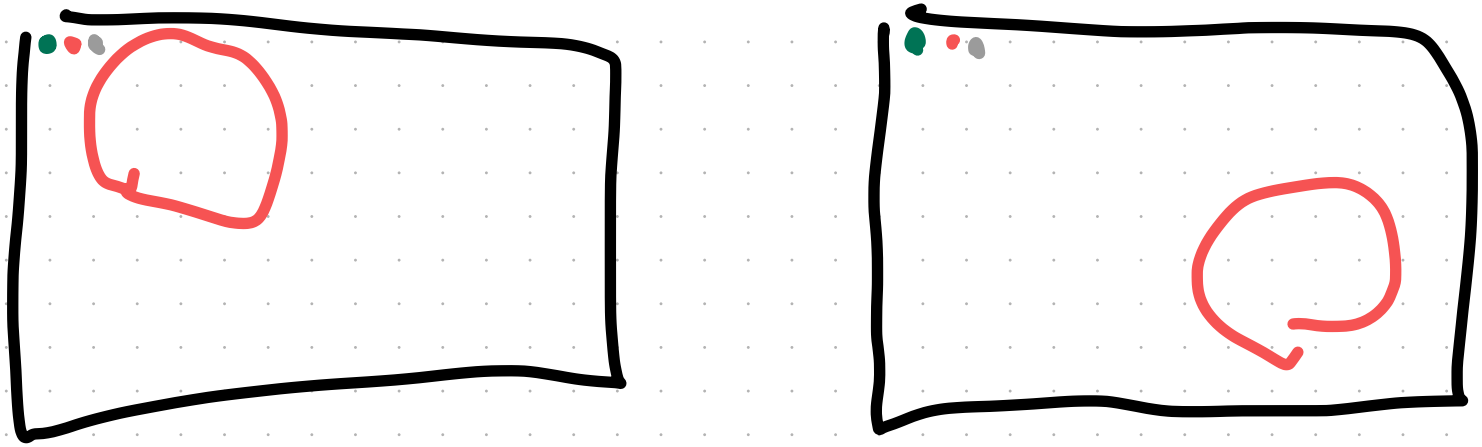


Image
Feature selection
few thousand dimensional vector

Distance Estimation

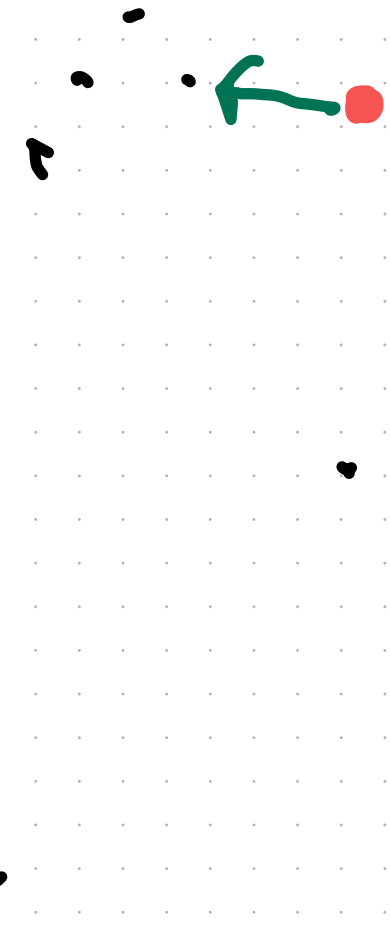
Approx Nearest Neighbor

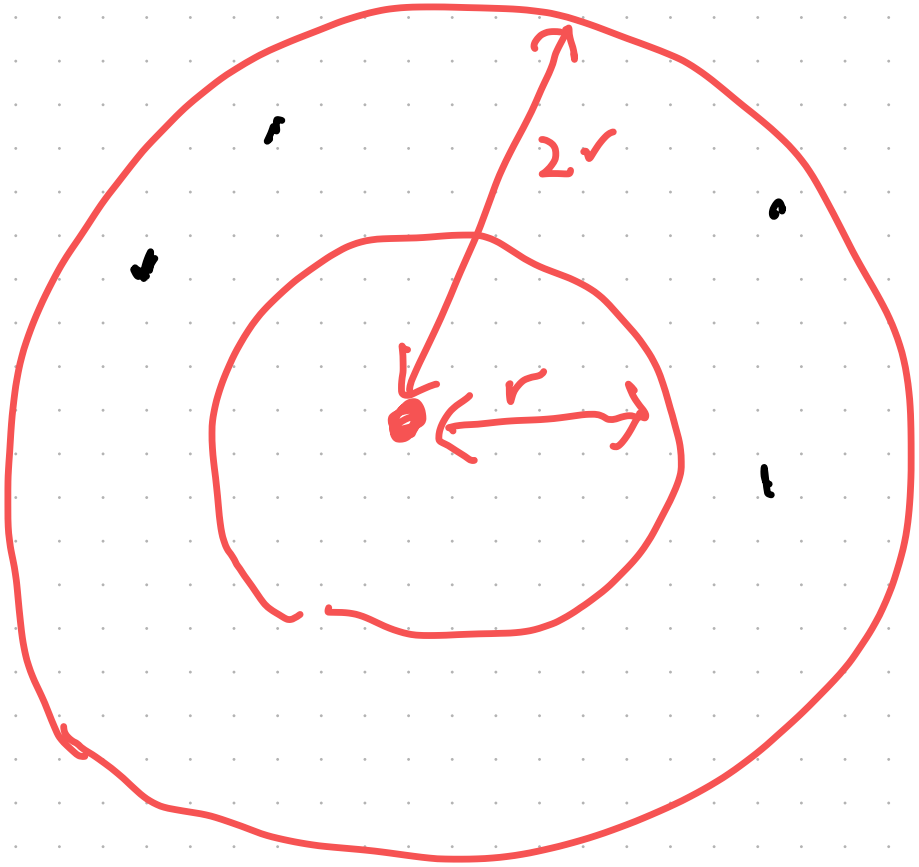
(r)

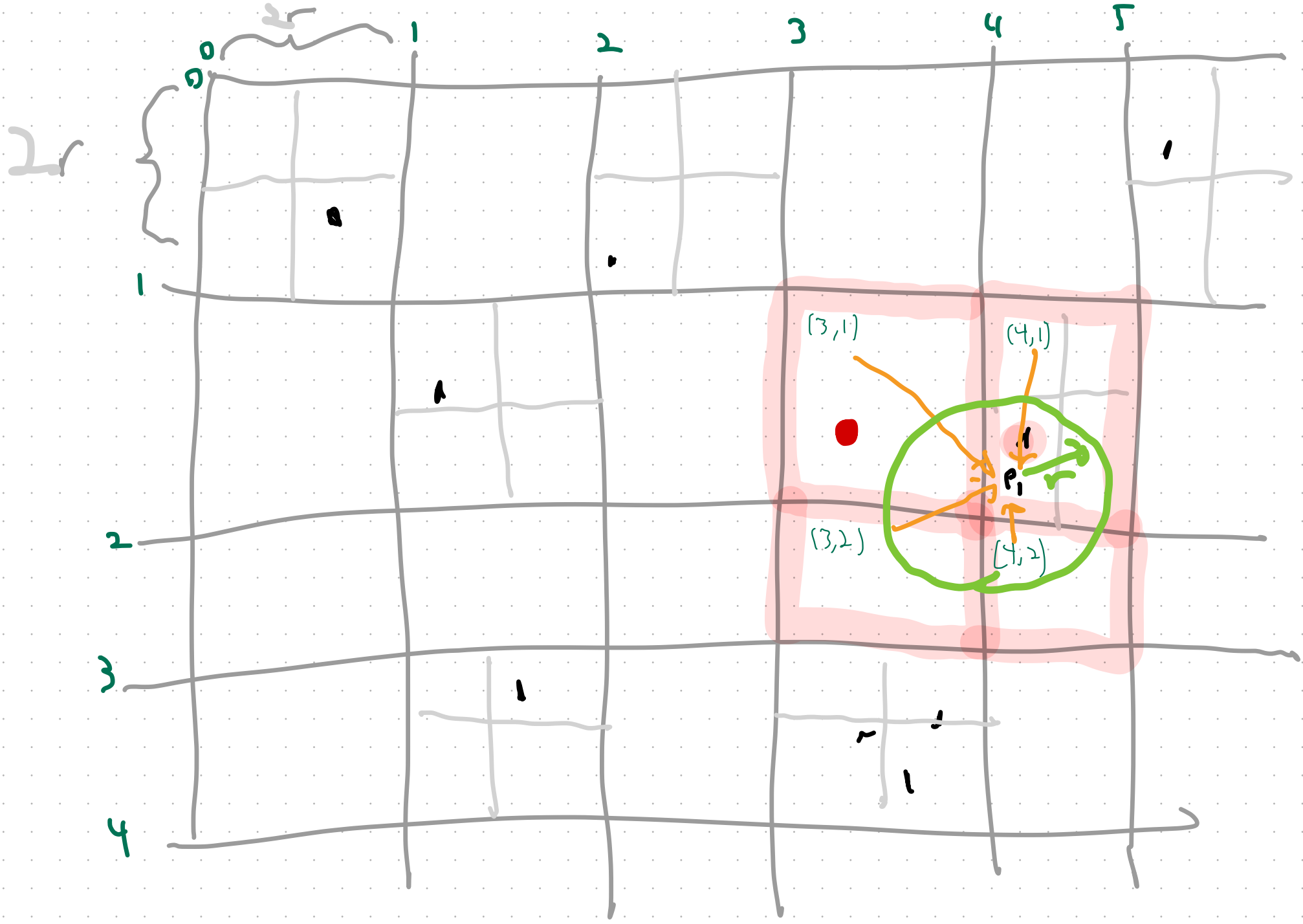
if there is a point within distance r find it

if there is a point $r \dots 2r$ perhaps find it

if there are no points $\leq 2r$ report that




2r
r


0
1
2
3
4
5
(3,1)
(4,1)
(3,2)
(4,2)
P1

$$2^d$$

$$\text{JL} \quad d = \log N$$

$$2^{\log N} = N$$