DMML, 14 Feb 2019

Clustering

- Top Down, K-means
- Bottom Up, Hierarchical

Distance

- Euclidean $\sqrt{\Delta x_1^2 + \cdots + \Delta x_n^2}$

  Minkowski $\left(\Delta x_1^m + \cdots + \Delta x_n^m\right)^{1/m}$

  $L_{m=1}$, Manhattan distance

- Differences in scale across attributes
  - Large magnitudes drown out smaller ones
  - Scale all attributes to a common range
- Non-numeric attributes
  - Categorical values (e.g. colour)
  - Binary values

Mixed attributes

- Some numeric, some not

Problem specific

Assume all attributes are non-numeric

- Simplest distance — fraction of agreement

x_1	x_2	x_3	x_4
4	11	11	*
y_1	y_2	y_3	y_4

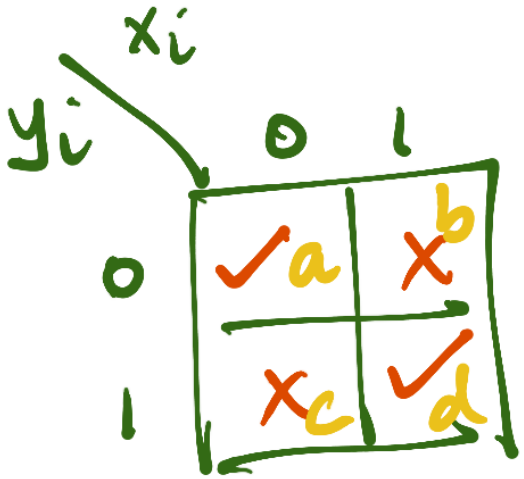
0.75 similar on a 0-1 scale

Distance = 1 − Similarity

Example

DNA sequences

Special case - binary



Each $x_i | y_i$ is in one of four categories

Similarity $\frac{\#a+\#d}{\#a+\#b+\#c+\#d}$

Asymmetric case

Boolean document model

Vocabulary $V = \{w_1, \ldots, w_N\}$

Document d is a vector $\{0,1\}^N$

$d_i = 1$ iff w_i occurs in d_i

Compare documents d & d'

Similarity $=$ $\dfrac{\#a + \#d}{\#a + \#b + \#c + \#d}$

$\#a$ — Words absent in d & d'

$\#d$ — Words in d & d'

N

Most words are absent in most documents

$N \approx 10^5$

Doc length $\approx$ 2500 (10 pages)

At least 97,500 of 100,000 words do not appear in d

$\#a \gg \#d$

Similarity ≈ 1 for any pair d, d'

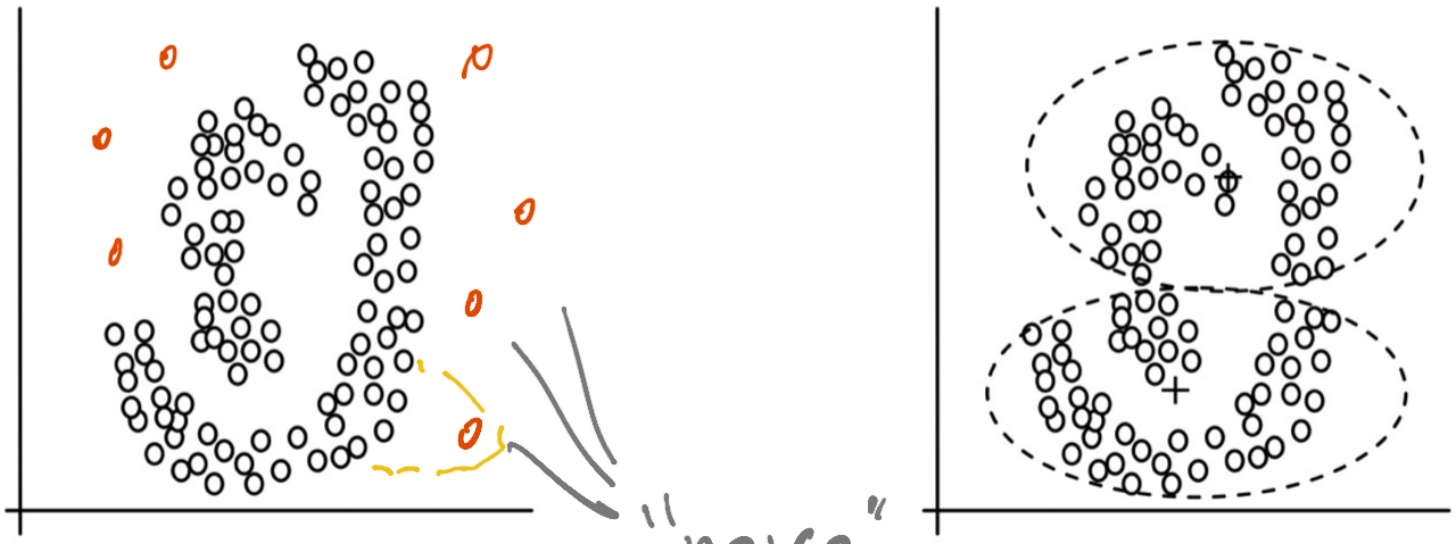
In such cases, discard #a

Focus on words that occur in $d \cup d'$

$$\frac{\#d}{\#b + \#c + \#d}$$

Jaccard Distance

Finding strangely shaped clusters



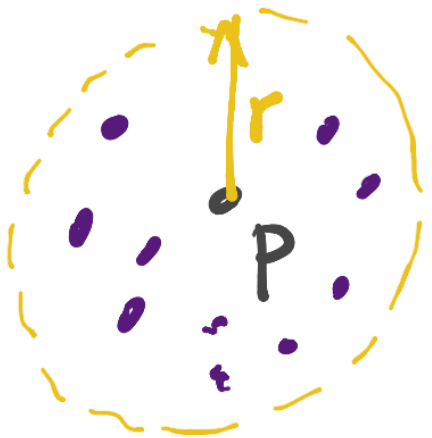
(A): Two natural clusters

(B): k-means clusters

Use "density" to define a cluster

Formalize

Fix a radius & count the points

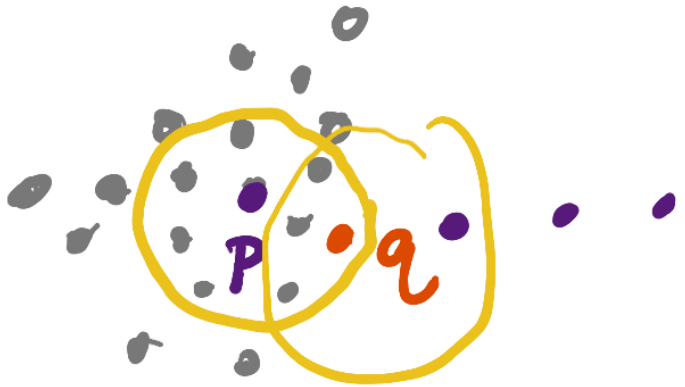


fixed radius

Fix p, $nbd_r(p)$

$q \in nbd_r(p)$

added to p's cluster



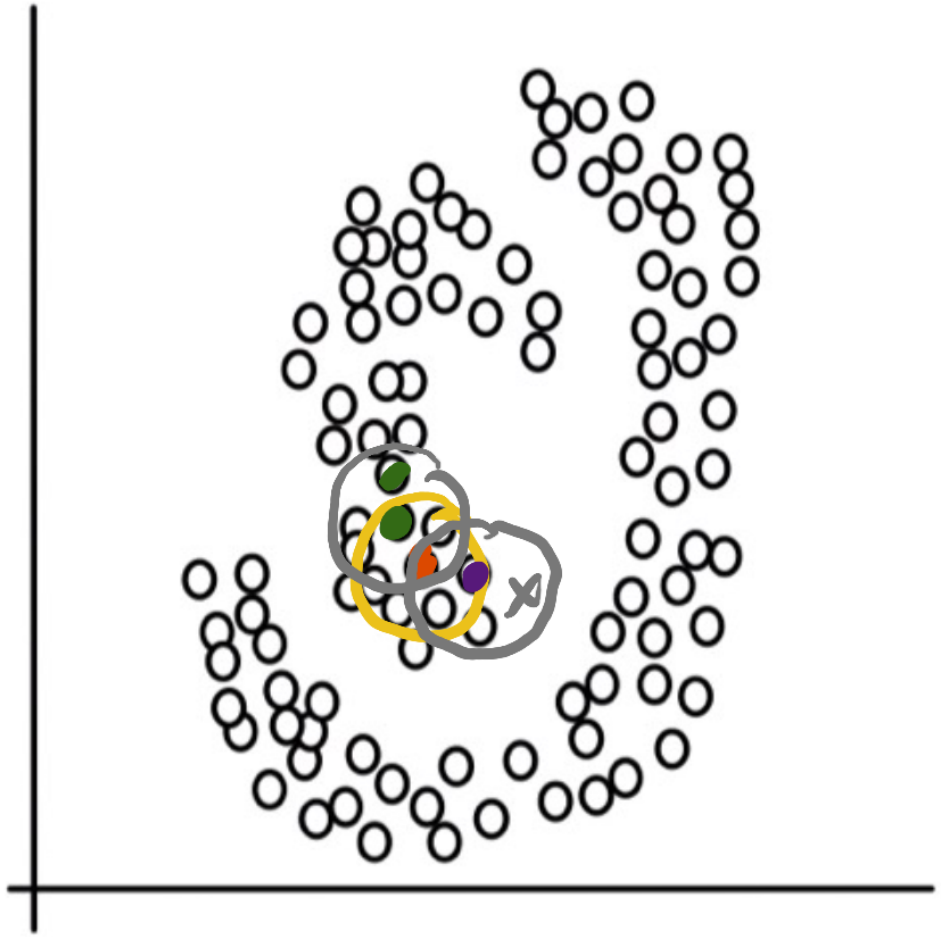
Should we continue the process with q?

Impose a density threshold m (minimum)

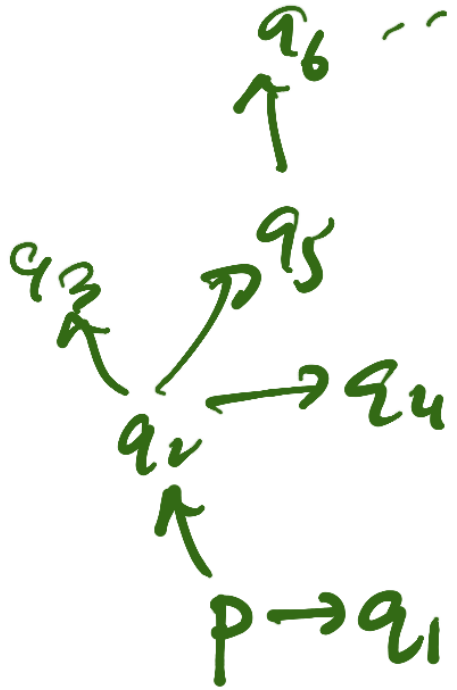
Explore $nbd_r(p)$ only if $nbd_r(p)$ has at least m points

Core point p has $\geq m$ neighbours in $nbd_r(p)$

$p \rightarrow q$ if p is a core point & $q \in nbd_r(p)$



(A): Two natural clusters



$p \to q$

to not

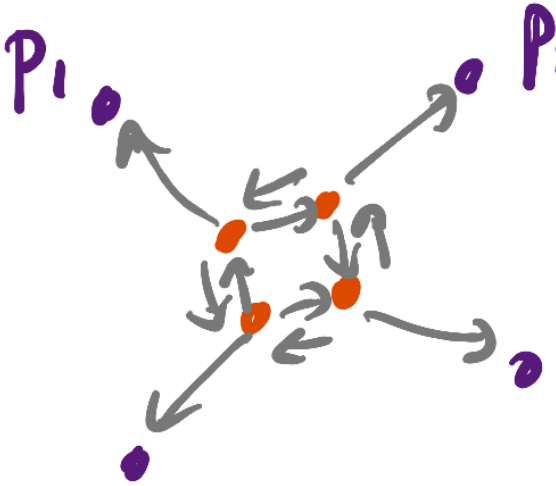
symmetric

Defines a directed graph on the points

p & p' are both core

and $p \to p'$ $\Rightarrow$ $p' \to p$

Connected component form clusters



Directed reachability

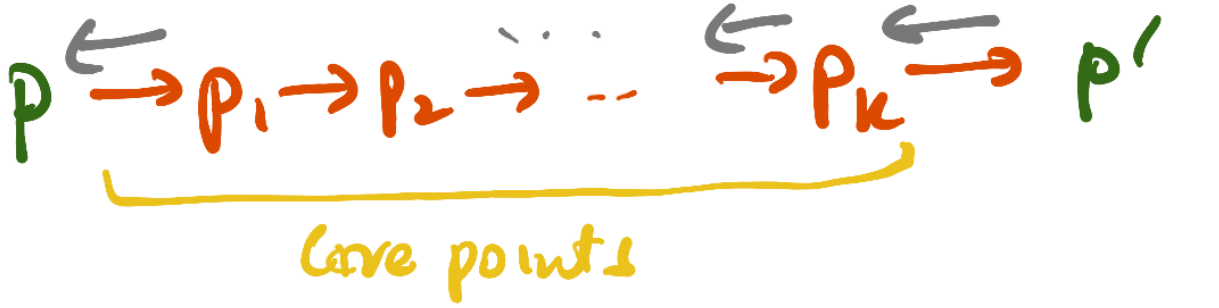
→ Undirected Connectivity

Strategy

Pick a point p in data set

- Explore all reachable points of p
  is a core point
- Identifies cluster containing p

Suppose p & p' are core point in same cluster

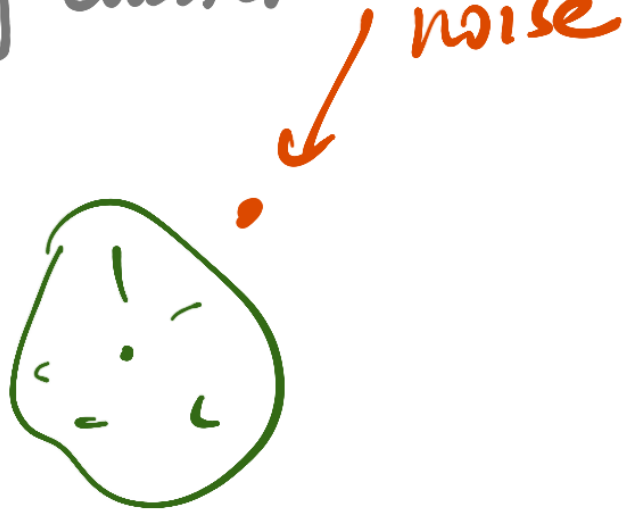


For all core nodes p

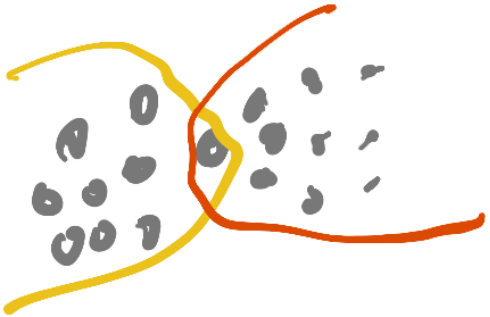
If cluster(p) not defined

explore from p & label cluster(p)

Some non core points may not be in any cluster



Two natural clusters that overlap at boundary



Not clear how to deal with overlapping boundary points

Preprocessing + efficient data structure to extract underlying graph from raw data

Outliers

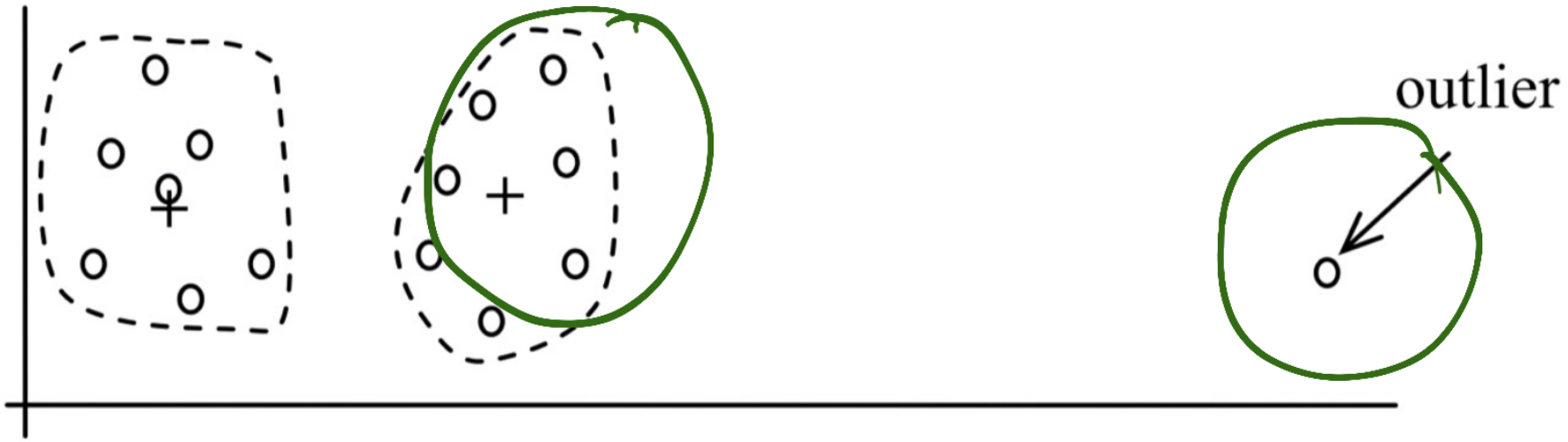
Points "far away" from others

Simple numeric value (single),

define in terms of distance from mean, etc

Multidimensional case

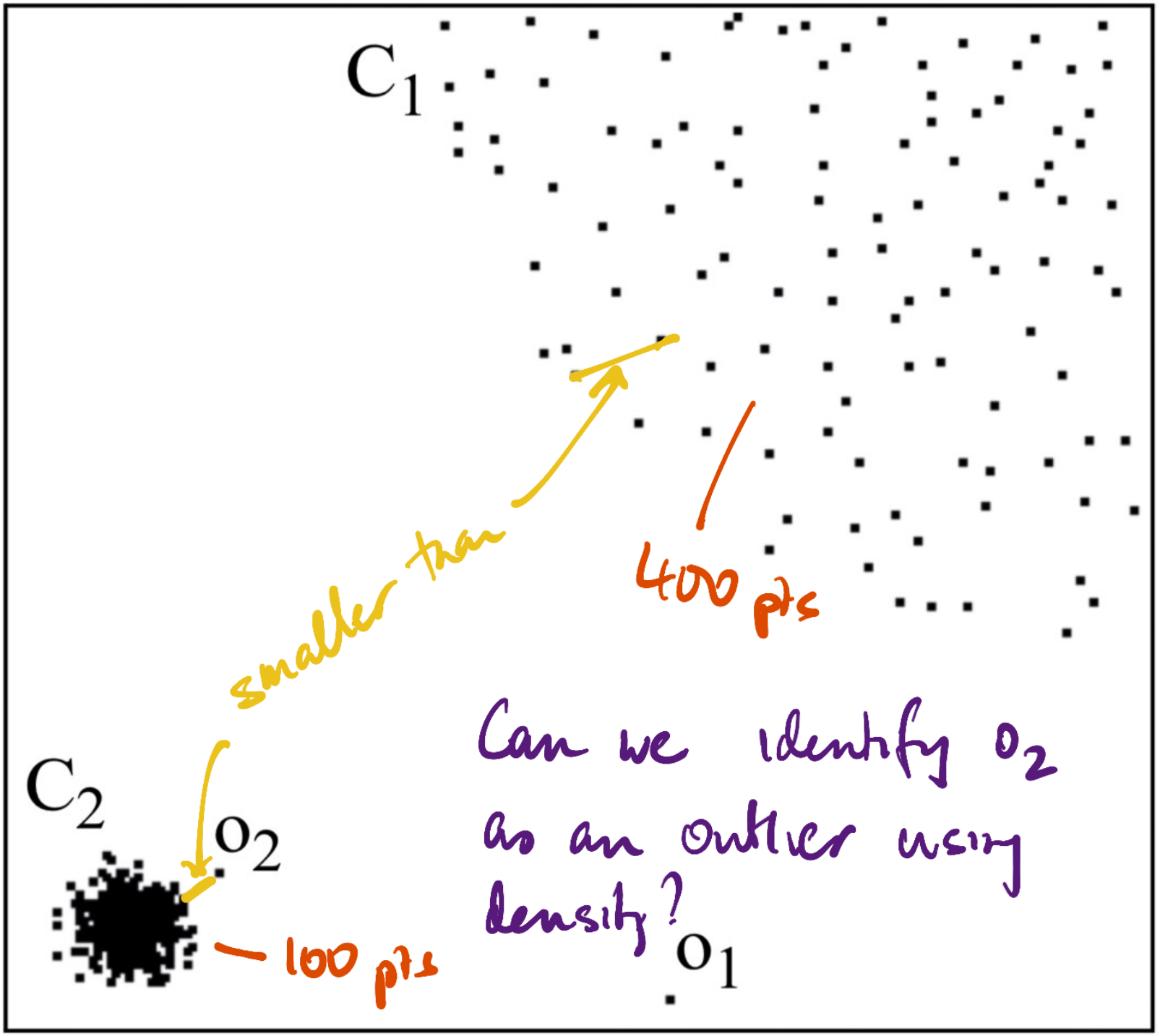
Use density



(B): Ideal clusters

An outlier will have a very sparse neighbourhood

- Fix radius r, threshold m
- Problem: r & m are global!

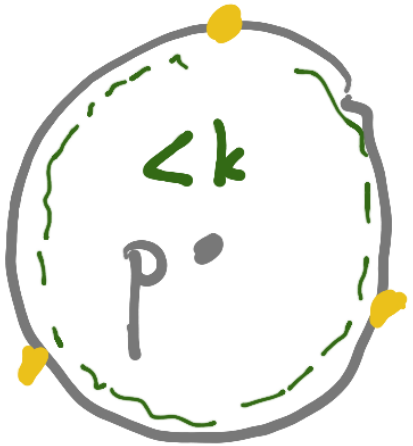

C_1
smaller than
400 pts
C_2
o_2
100 pts
Can we identify o_2 as an outlier using density?
o_1

No good uniform values of m & r to classify O_2 accurately

Local notion of density

- Earlier: fix r, count points in $nbd_r(p)$
- Instead: fix k, find r s.t. $nbd_r(p)$ has k points

Smallest r s.t. $nbd_r(p)$ has $\geq k$ points, for any $r' < r$, $nbd_{r'}(p) \leq k-1$ pts



Given p, $\text{local-radius}_k(p) = lr_k(p)$

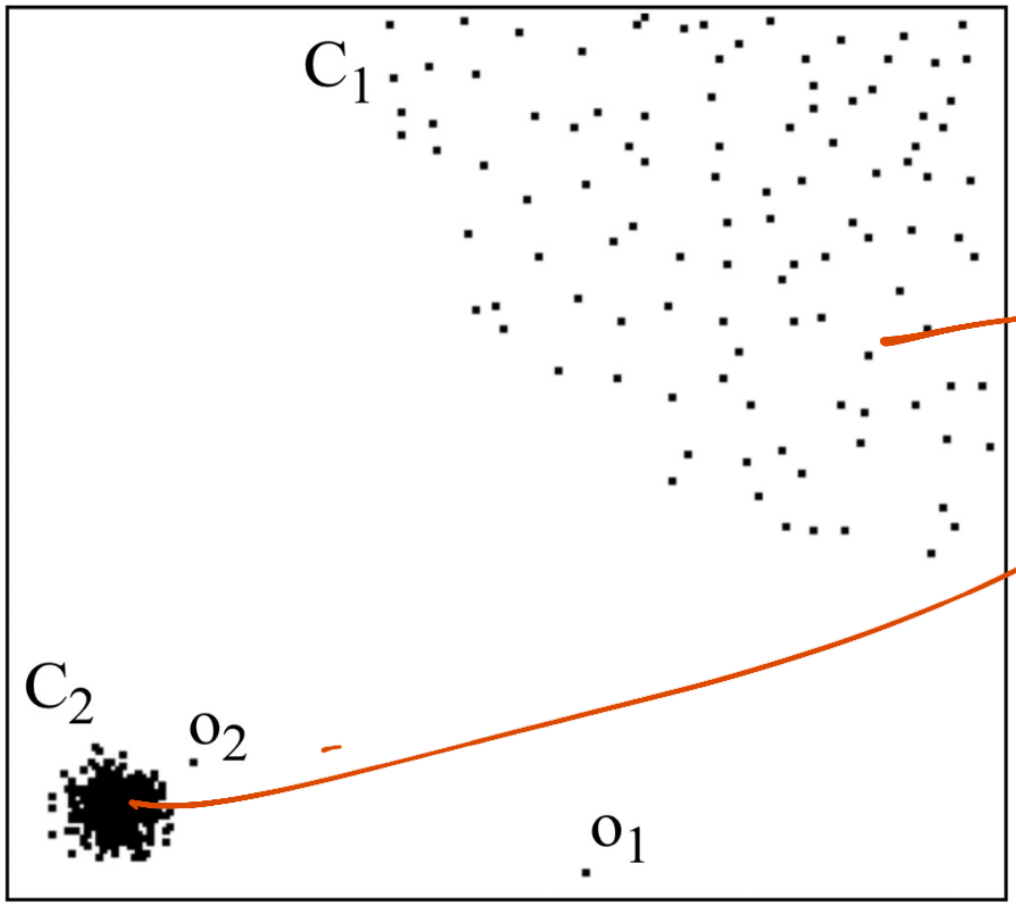
For all q in $lr_k(p) \rightarrow lr_k(q)$

Take average: $\frac{1}{\text{size}(knbd(p))} \sum_q lr_k(q)$

$$\frac{lr_k(p)}{\text{Avg } lr_k(q) \text{ of neighbours}} = \text{Local Outlier Factor of } p$$

Fix a threshold t (empirically)

Define p to be an outlier if $LOF(p) \geq t$



LOF
LOF
> ≥1 in both cases