CSIT5210

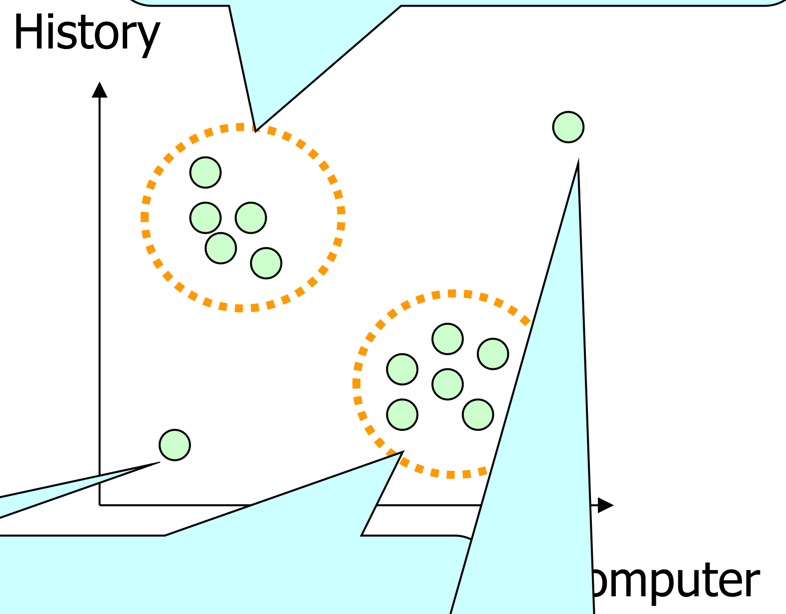
Outlier

Prepared by Raymond Wong
Presented by Raymond Wong
raywong@cse

Outlier

Clustering:

	Computer	History
Raymond	100	40
Louis	90	45
Wyman	20	95
…	…	…



Cluster 2
(e.g. High Score in History
and Low Score in Computer)

Outlier
(e.g. Low Score in Computer
and Low Score in History)

Cluster 1
(e.g. High Score in Computer
and Low Score in History)

Outlier
(e.g. High Score in Computer
and High Score in History)

Problem: to find all outliers

Outlier

- Applications
  - Fraud Detection
    - Detect unusual usage of credit cards or telecommunication services
  - Medical Analysis
    - Finding unusual response to various medical treatment
  - Customized Marketing
    - Customers with extremely low or extremely high incomes
  - Network
    - A potential network attack
  - Software
    - A potential bug

Outlier

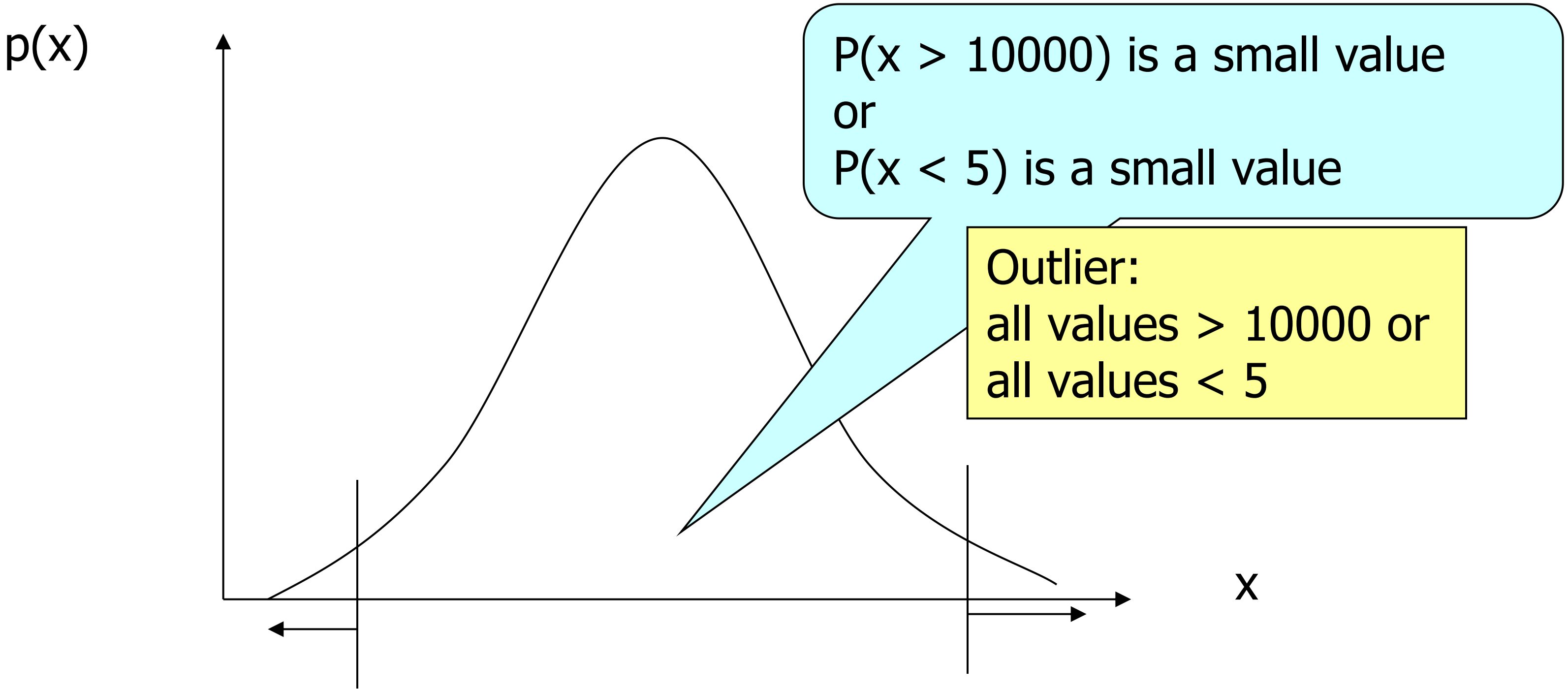
- Statistical Model
- Distance-based Model
- Density-Based Model

Statistical Model

- An outlier is an observation that is numerically distant from the rest of the data
- E.g.,
  - Consider 1-dimensional data
  - How is a data point considered as an outlier?

Statistical Model

- Assume the 1-dimensional data follows the normal distribution



P(x > 10000) is a small value
or
P(x < 5) is a small value

Outlier:
all values > 10000 or
all values < 5

Statistical Model

- Disadvantage
  - Assume that the data follows a particular distribution

Outlier

- Statistical Model
- Distance-based Model
- Density-Based Model

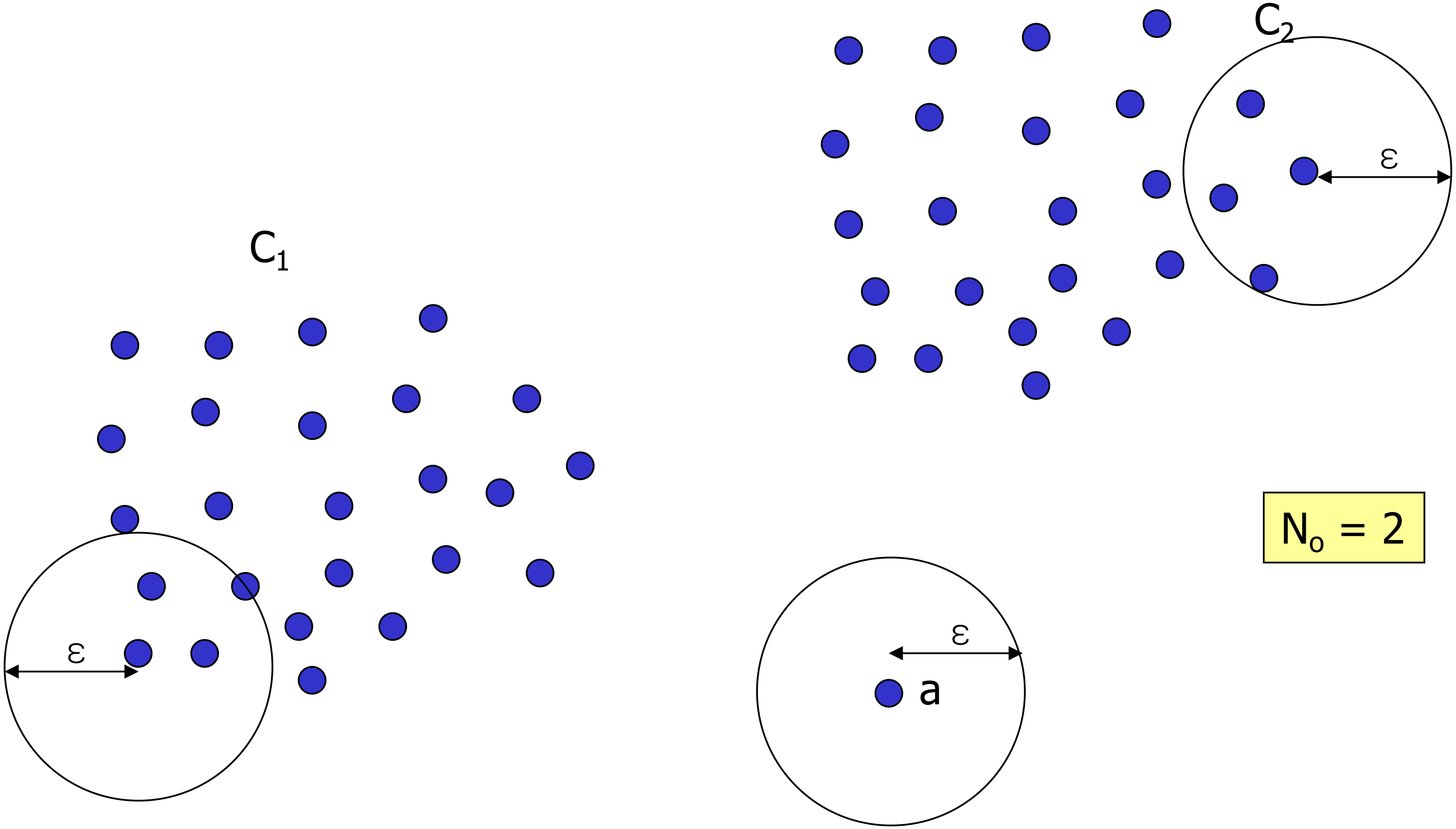
Distance-based Model

- Advantage
  - This model does not assume any distribution
- Idea
  - A point p is considered as an outlier if there are too few data points which are close to p

Distance-based Model

- Given a point p and a non-negative real number ε,
  - the ***ε-neighborhood*** of point p, denoted by N(p), is the set of points q (including point p itself) such that the distance between p and q is within ε.
- Given a non-negative integer N_o and a non-negative real number ε
  - A point p is said to be an outlier if
    - $N(p) <= N_o$

Distance-based Model

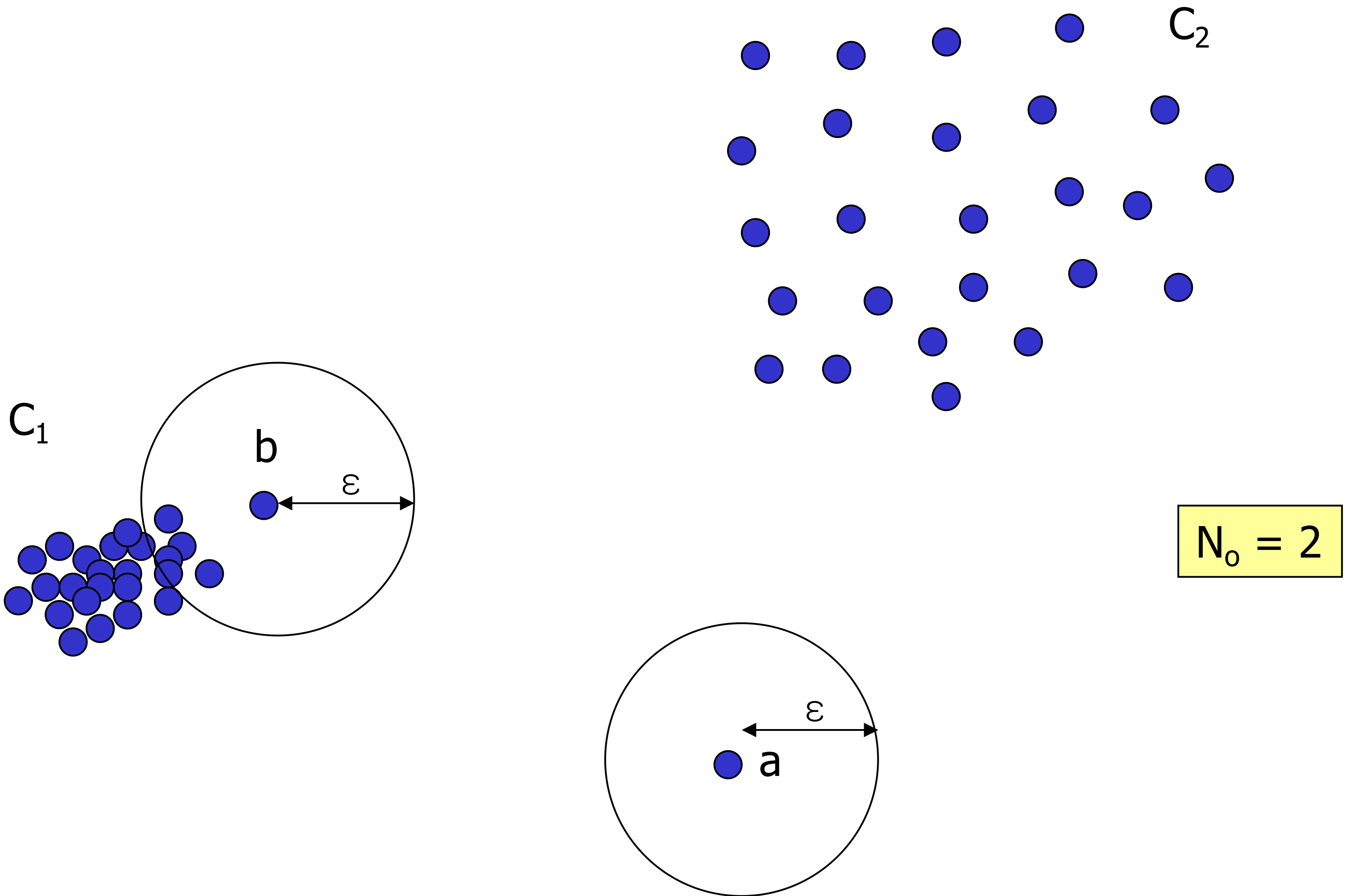


$N_o = 2$

Distance-based Model

- Is the distance-based model “perfect” to find the outliers?

Distance-based Model



Outlier

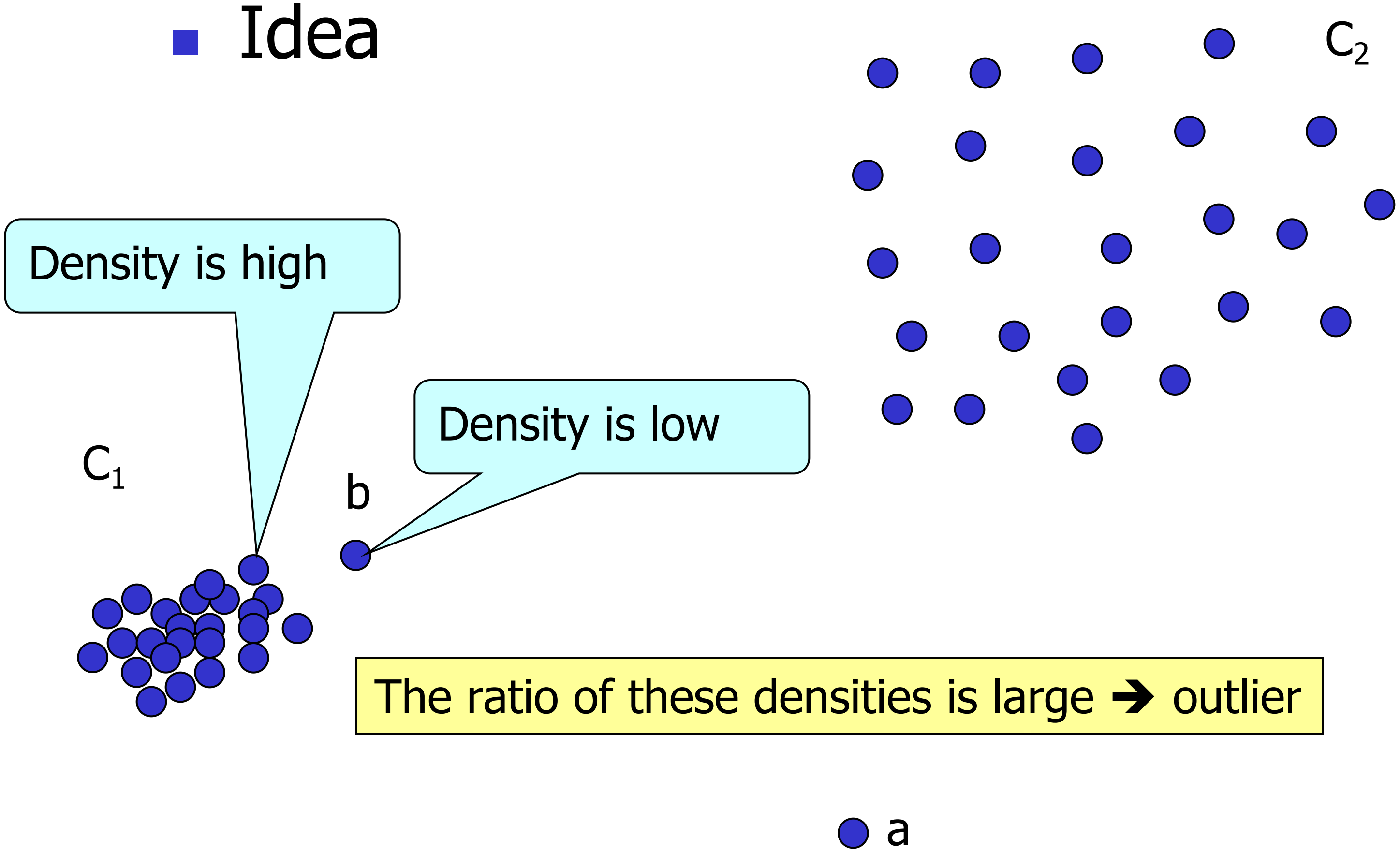
- Statistical Model
- Distance-based Model
- Density-Based Model

Density-Based Model

- Advantage:
  - This model can find some “local” outliers

Density-Based Model

- Idea



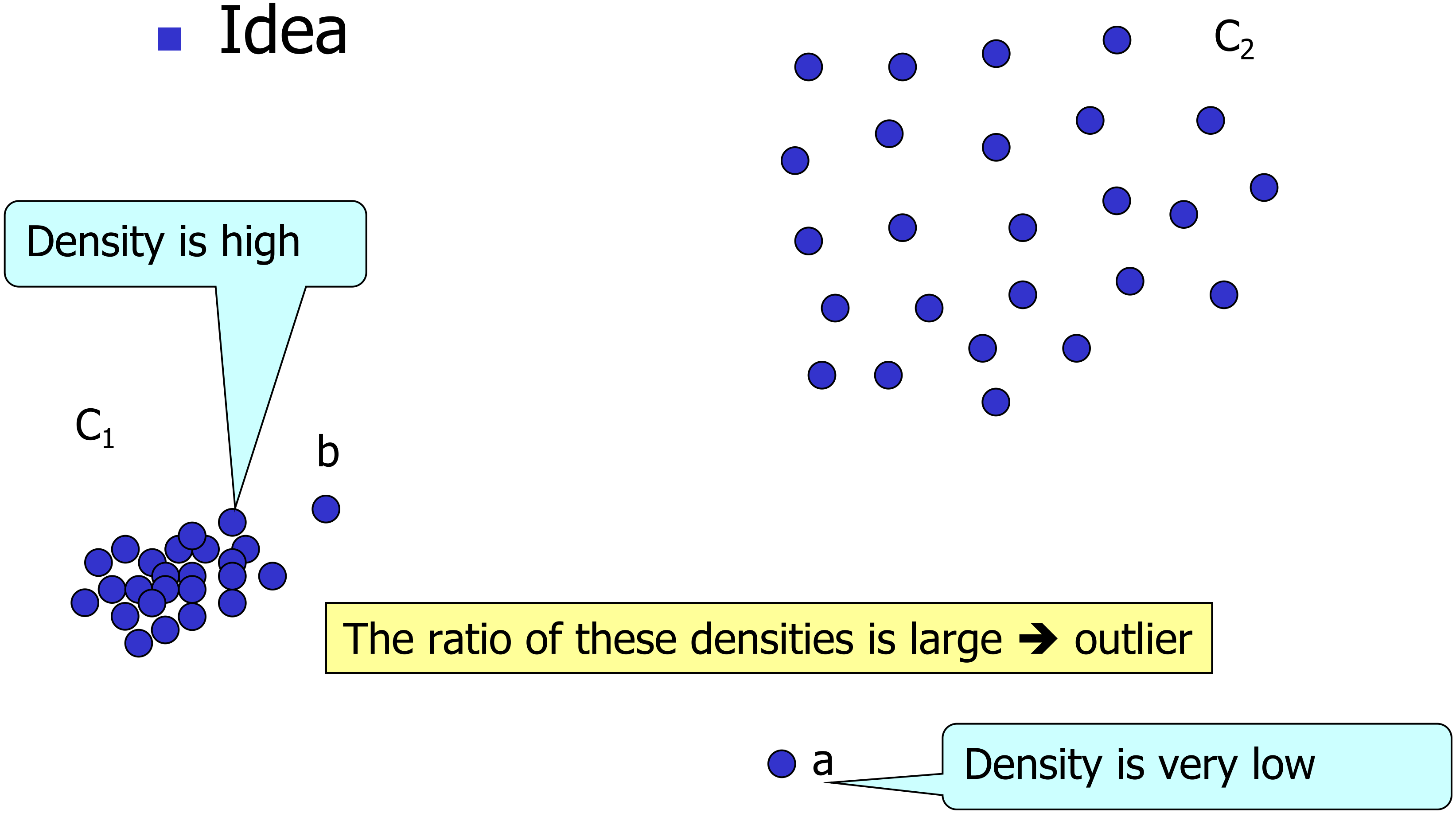
Density is high

Density is low

The ratio of these densities is large ➔ outlier

Density-Based Model

- Idea



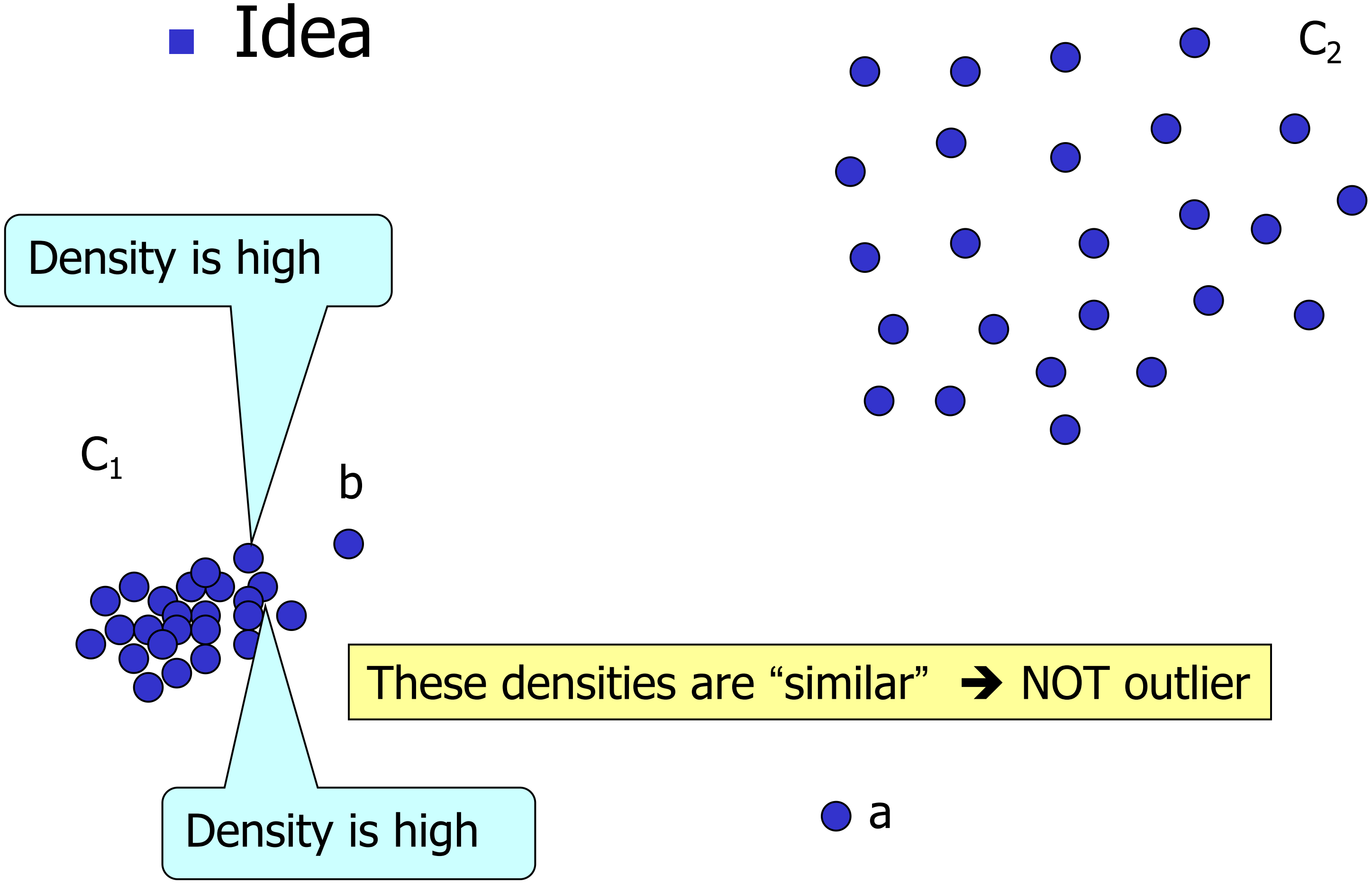
Density is high

The ratio of these densities is large ➔ outlier

Density is very low

Density-Based Model

- Idea



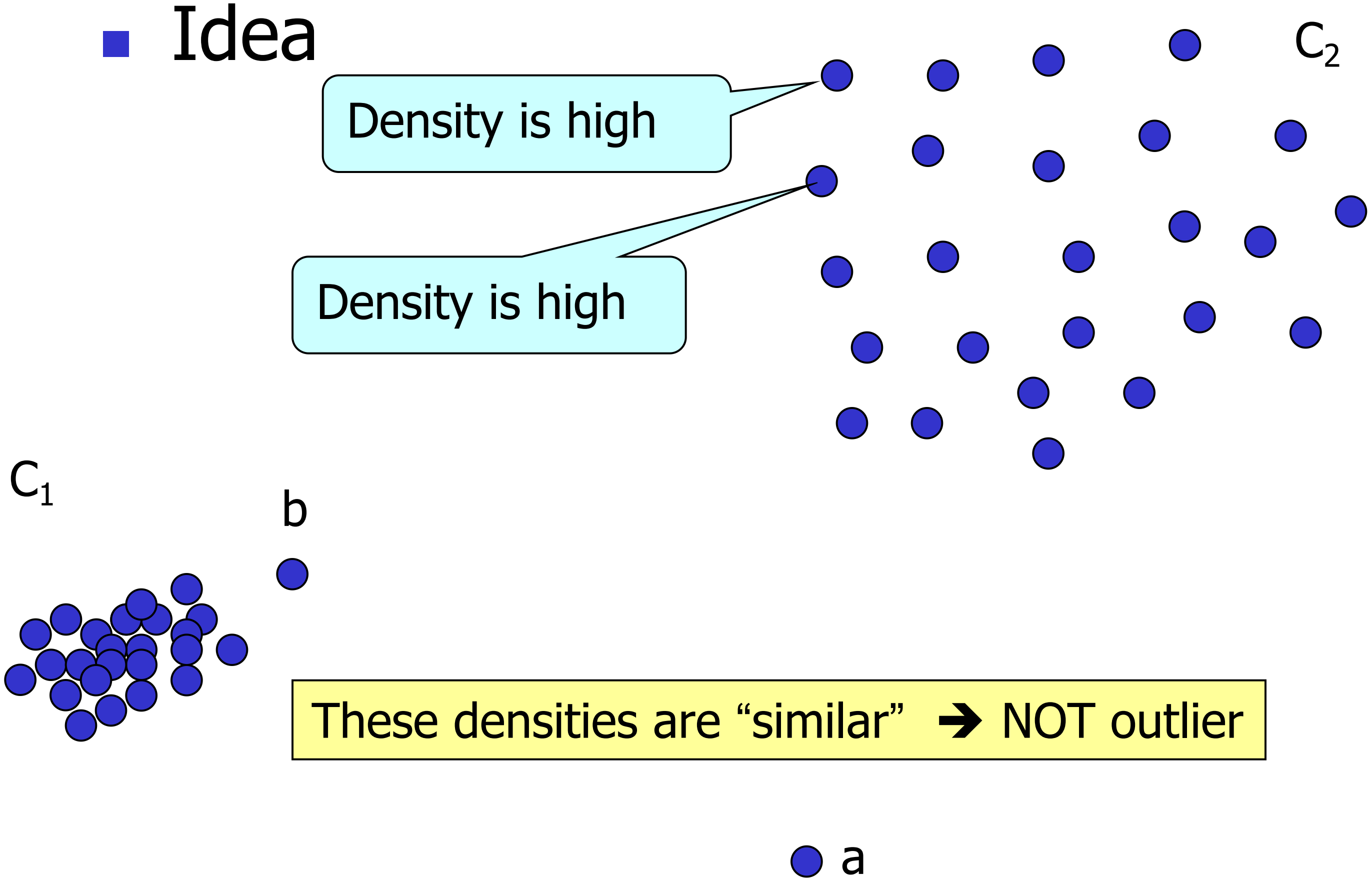
Density is high

Density is high

These densities are “similar” ➔ NOT outlier

Density-Based Model

- Idea



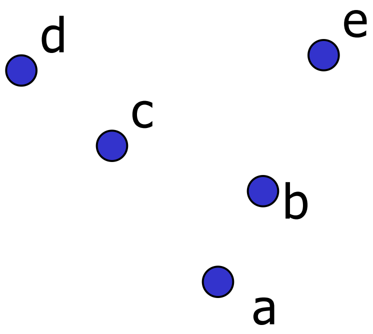
Density is high

Density is high

These densities are “similar” ➔ NOT outlier

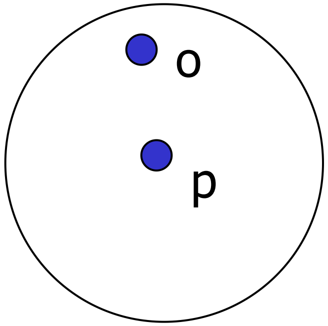
Density-Based Model

- Formal definition
  - Given an integer k and a point p,
    - $N_k(p)$ is defined to be the ε-neighborhood of p (excluding point p)
    - where ε is the distance between p and the k-th nearest neighbor



$N_1(a) = ?$

$N_2(a) = ?$



Density-Based Model

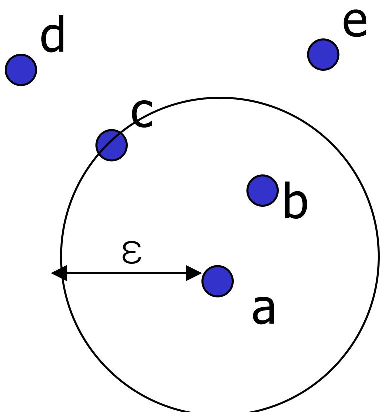
- Reachability Distance of p with respect to o
  - Given two points p and o and an integer k,
    - $\text{Reach\_dist}_k(p, o)$ is defined to be $\max\{\text{dist}(p, o), \varepsilon\}$
    - where ε is the distance between p and the k-th nearest neighbor

$\text{Reach\_dist}_2(a, b)$ =?

$\text{Reach\_dist}_2(a, c)$ =?

$\text{Reach\_dist}_2(a, d)$ =?

$\text{Reach\_dist}_2(a, e)$ =?



k = 2

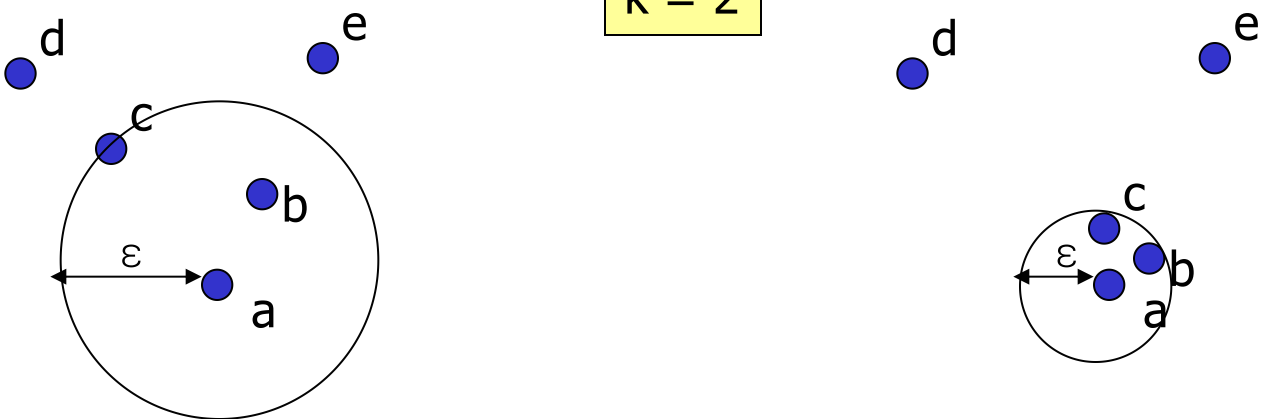
Density-Based Model

- The **average reachability distance** of p among all k nearest neighbors is equal to ε
  - where ε is the distance between p and the k-th nearest neighbor

Why?

- The **local reachability density** of p (denoted by $\text{lrd}_k(p)$) is defined to be 1/ε

k = 2



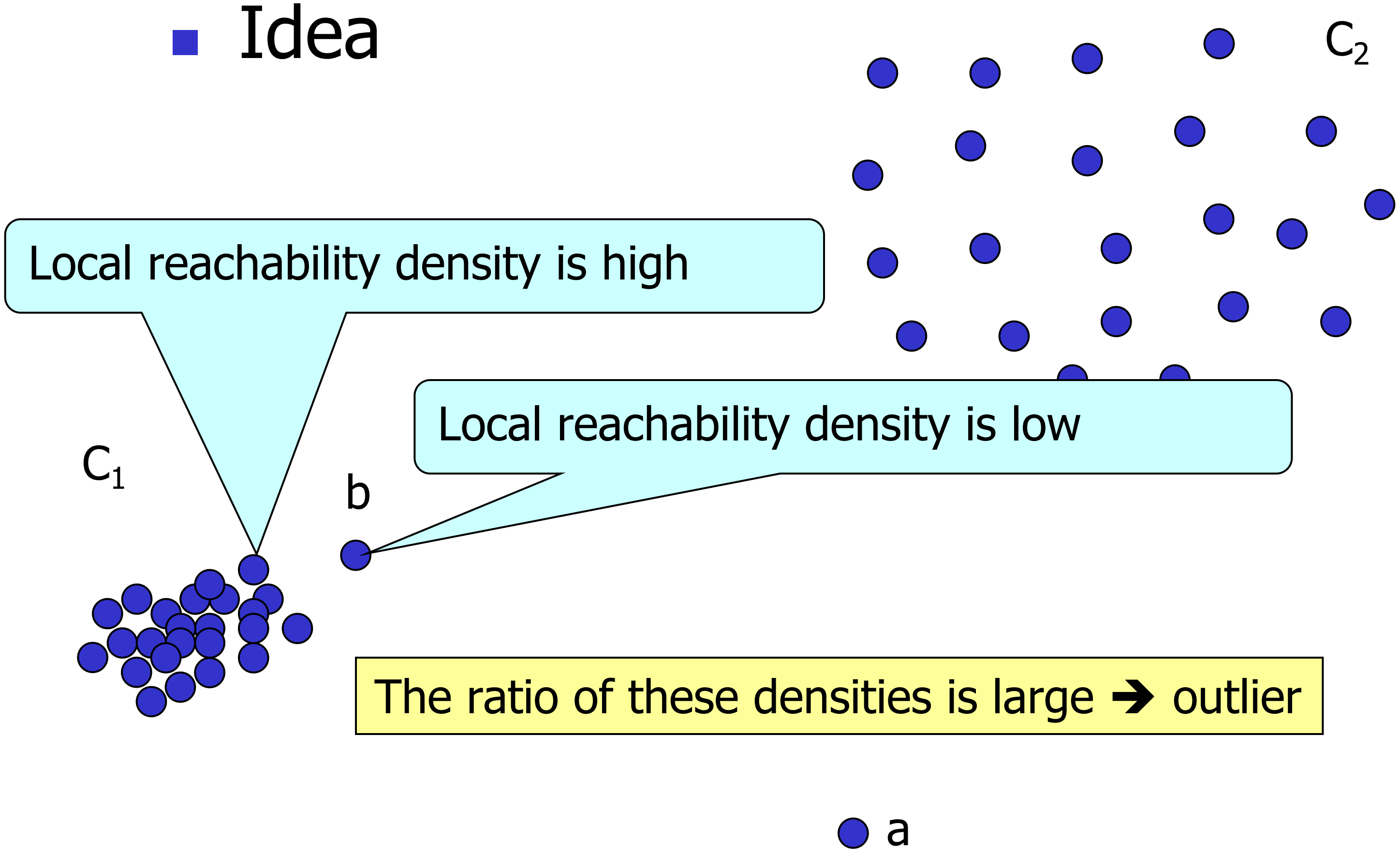
Density-Based Model

- The **local outlier factor (LOF)** of a point p is equal to

$$\frac{\sum_{o \in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{k}$$

Density-Based Model

- Idea



Local reachability density is high

Local reachability density is low

The ratio of these densities is large ➔ outlier