# K-means 算法

# 主要内容:

- 数据挖掘简介
- 数据挖掘的任务简介
- 聚类算法简介
- K-means算法简介
- K-means算法的缺陷及改进

# 什么是数据挖掘？

定义：

数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中、人们事先不知道的、但又潜在有用的信息和知识的过程。

# 数据挖掘的主要任务

- 分类（Classification）
- 预测（Prediction）
- 聚类（Clustering）
- 关联规则(Association)
- 偏差检测（Deviation detection ）

●分类：指将数据映射到预先定义好的群组或类。1.从数据中选出已经分好类的训练集，在该训练集上运用数据挖掘分类的技术，建立分类模型，2.对于测试数据进行分类。

●预测：预测是通过分类或估值起作用的，也就是说，通过分类或估值得出模型，该模型用于对未知变量的预言。

●聚类：在没有给定划分类的情况下，根据信息相似度将信息分组。是一种无指导的学习。

●关联规则：揭示数据之间的相互关系，而这种关系没有在数据中直接表现出来。

●偏差检测：用于发现与正常情况不同的异常和变化。并分析这种变化是有意的欺诈行为还是正常的变化。如果是异常行为就采取预防措施。

# 聚类算法简介

1. 聚类的目标：将一组数据分成若干组，组内数据是相似的，而组间数据是有较明显差异。

2. 与分类区别：分类与聚类最大的区别在于分类的目标事先已知，聚类也被称为无监督机器学习

3. 聚类手段：传统聚类算法 ①划分法　②层次方法　③基于密度方法　④基于网络方法　⑤基于模型方法

# 什么是Kmeans算法？

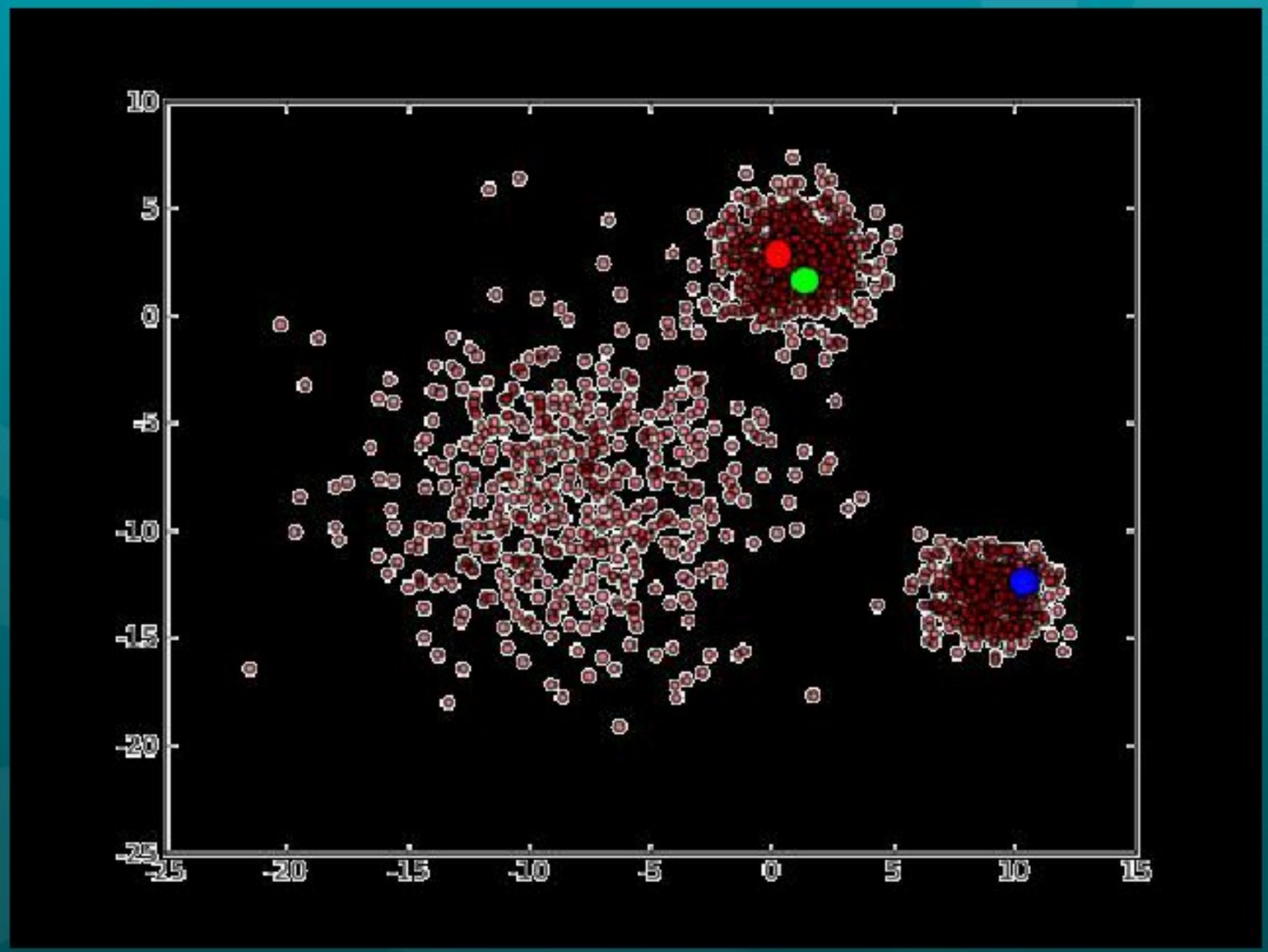
Q1：K是什么？A1：k是聚类算法当中类的个数。

Q2： means是什么？A2： means是均值算法。

**Summary： Kmeans是用均值算法把数据分成K个类的算法！**

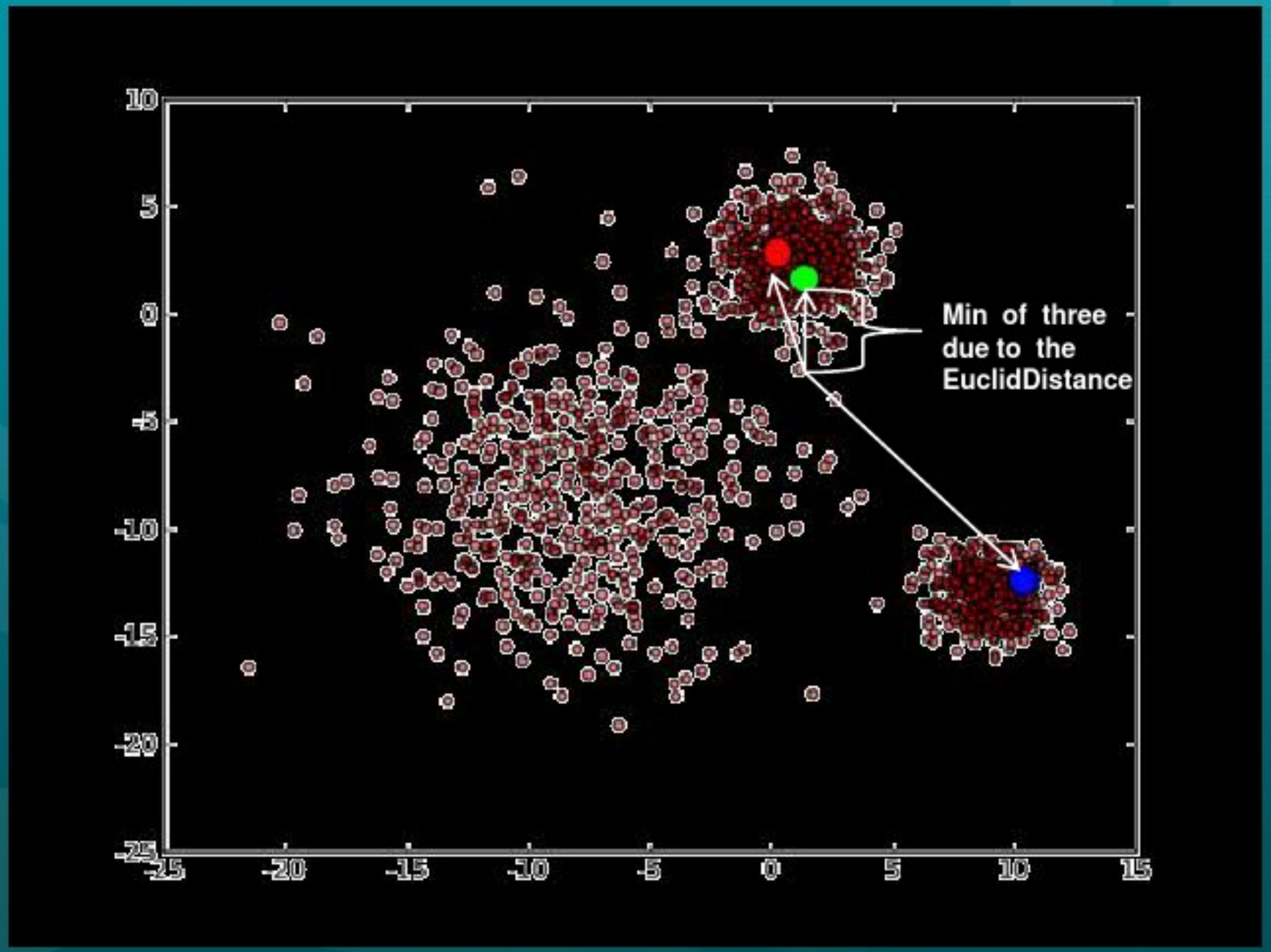
# Kmeans算法详解（1）

步骤一：取得k个初始中心点



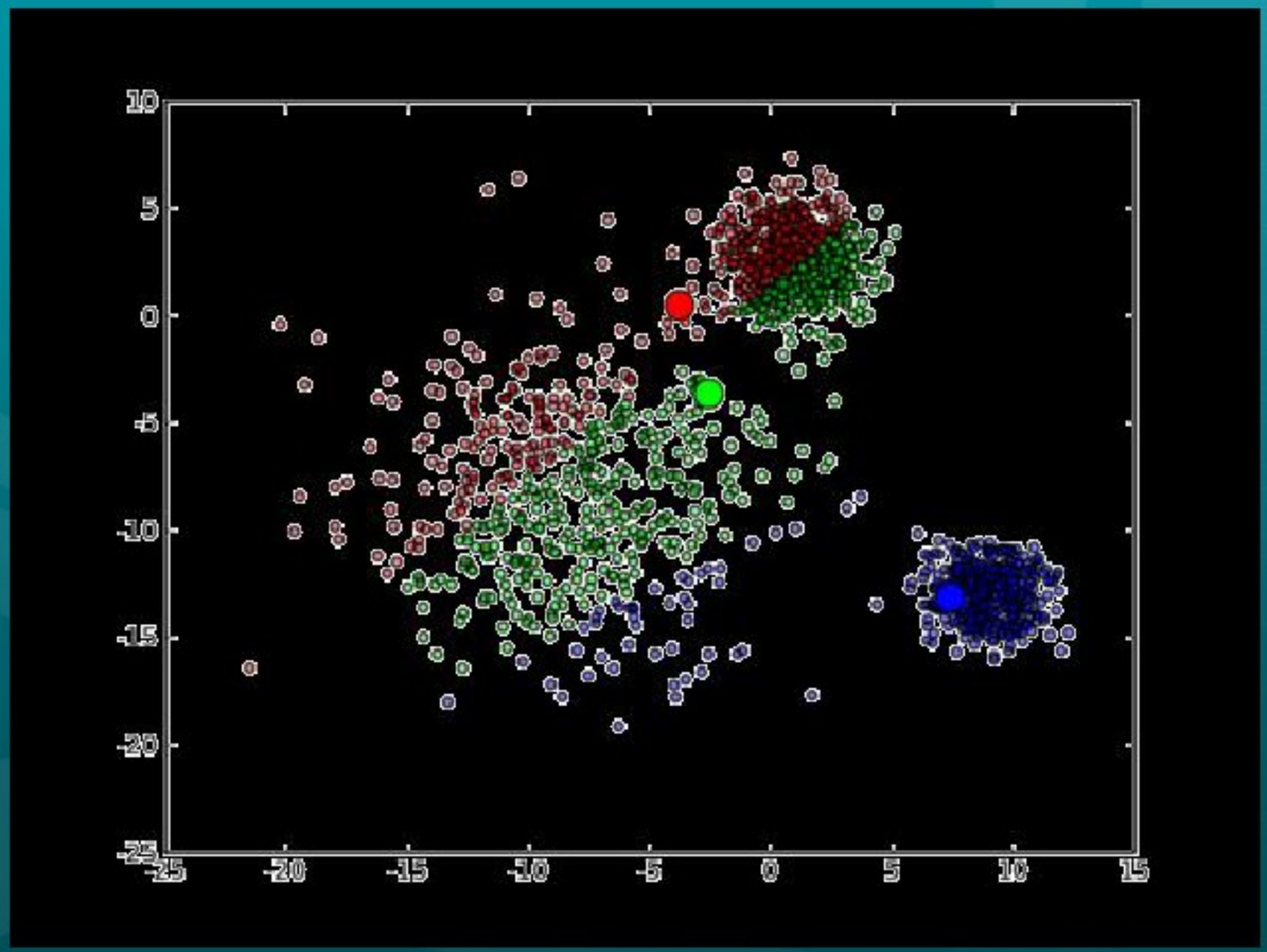

# Kmeans算法详解（2）

步骤二：把每个点划分进相应的簇



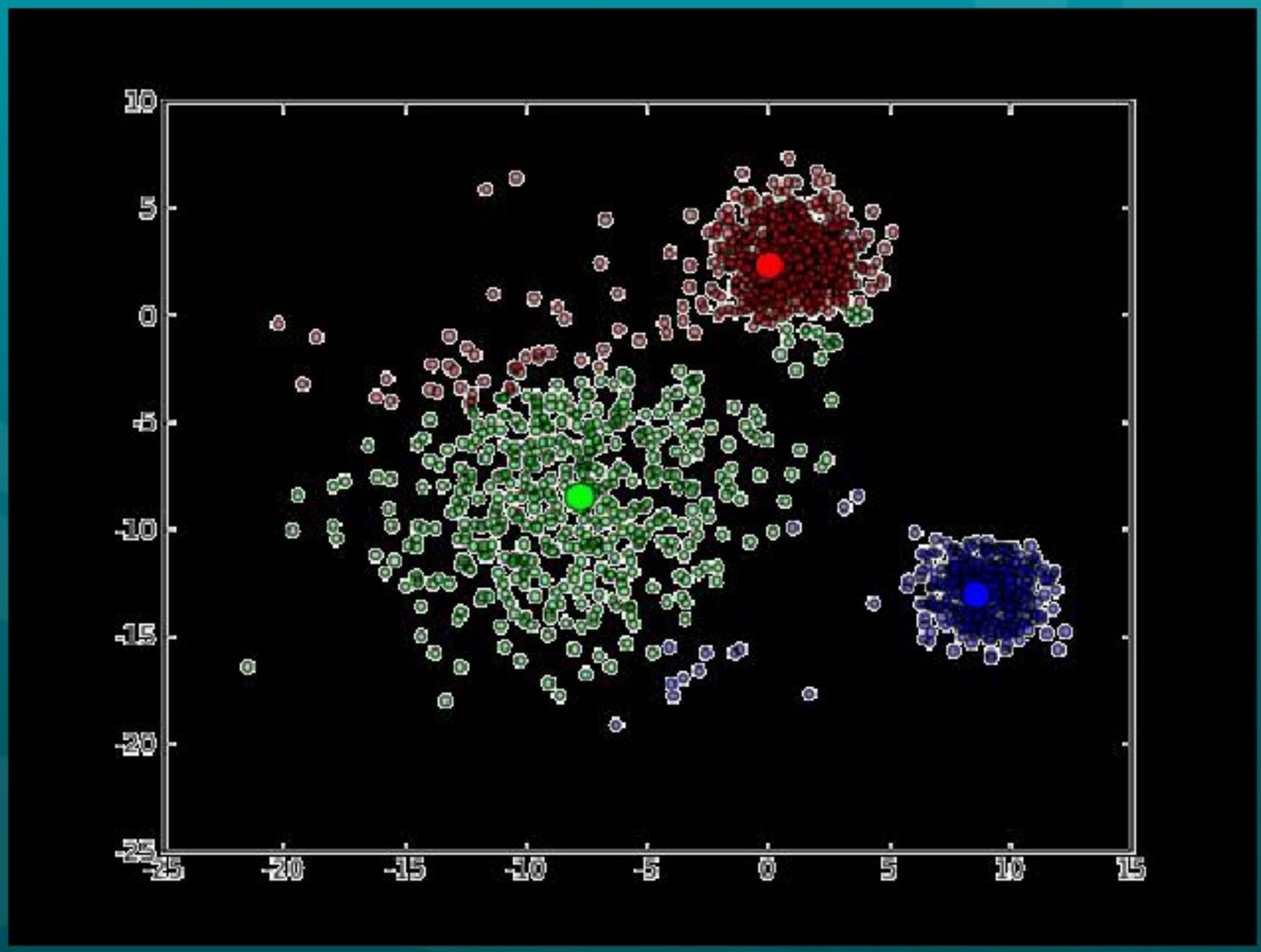

# Kmeans算法详解（3）

步骤三：重新计算中心点



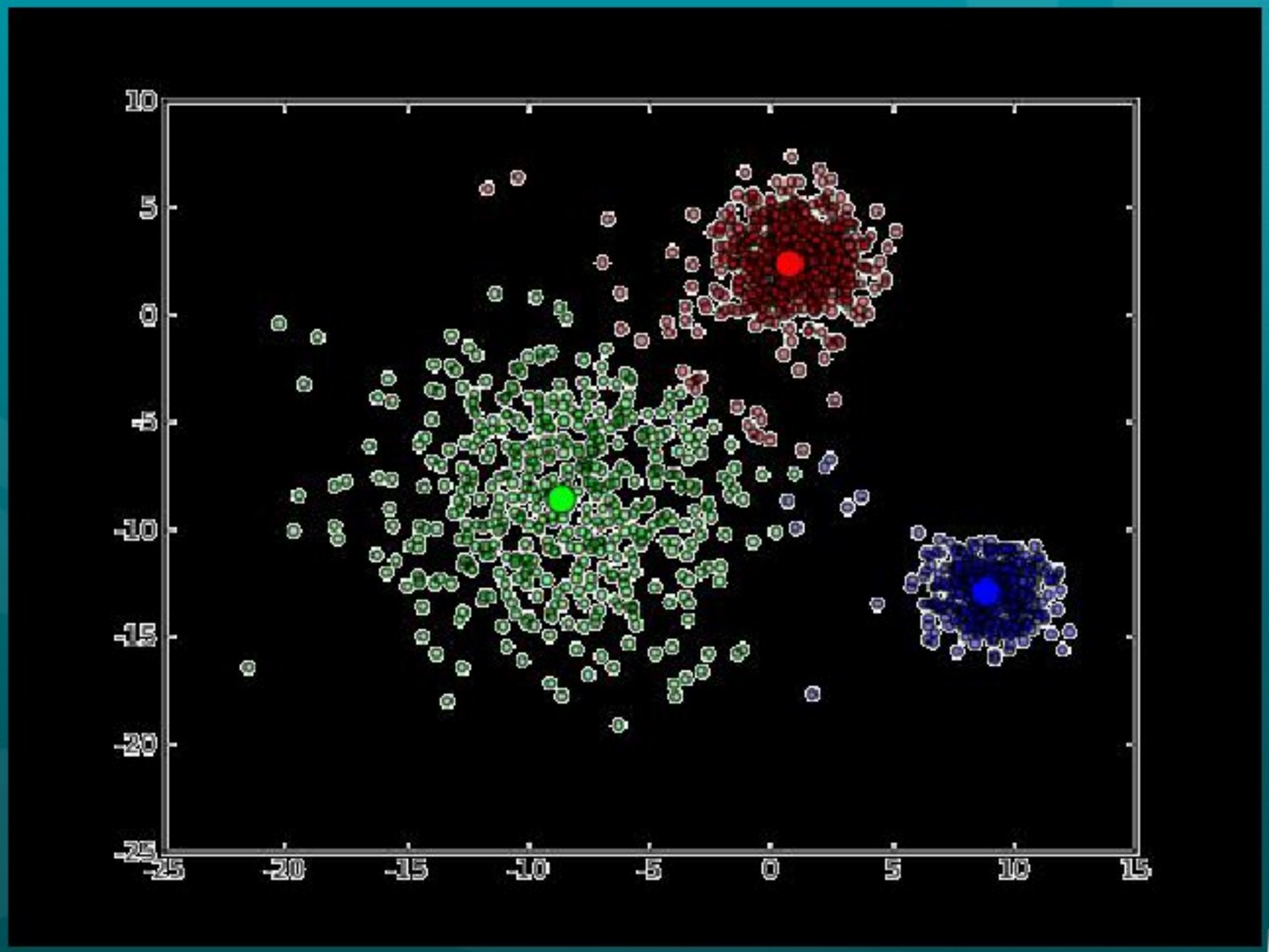

# Kmeans算法详解（4）

步骤四：迭代计算中心点




# Kmeans算法详解（5）

步骤五：收敛




# Kmeans算法流程

1.从数据中随机抽取k个点作为初始聚类的中心，由这个中心代表各个聚类
2.计算数据中所有的点到这k个点的距离，将点归到离其最近的聚类里
3.调整聚类中心，即将聚类的中心移动到聚类的几何中心（即平均值）处，也就是k-means中的mean的含义
4.重复第2、3步直到聚类的中心不再移动，此时算法收敛

# 决定性因素

## Input & centroids

①数据的采集和抽象
②初始的中心选择

## MaxIterations & Convergence

①最大迭代次数
②收敛值

factors?

## Selected k

① k值的选定

## Meassures

①度量距离的手段

# 主要因素

初始中心点

输入的数据及K值的选择

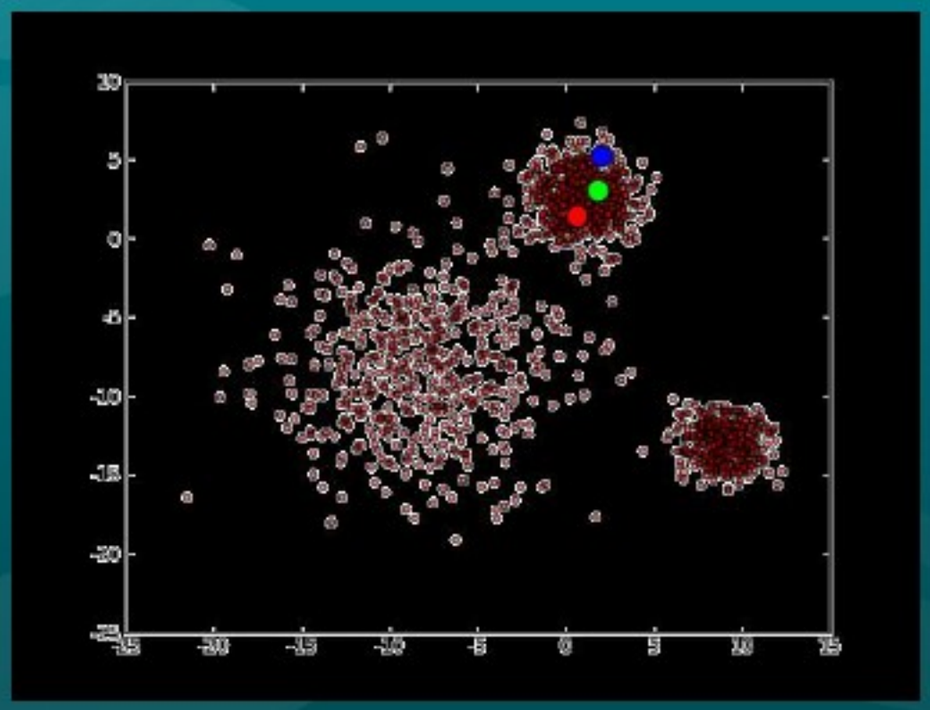
距离度量

主要三个方面因素。

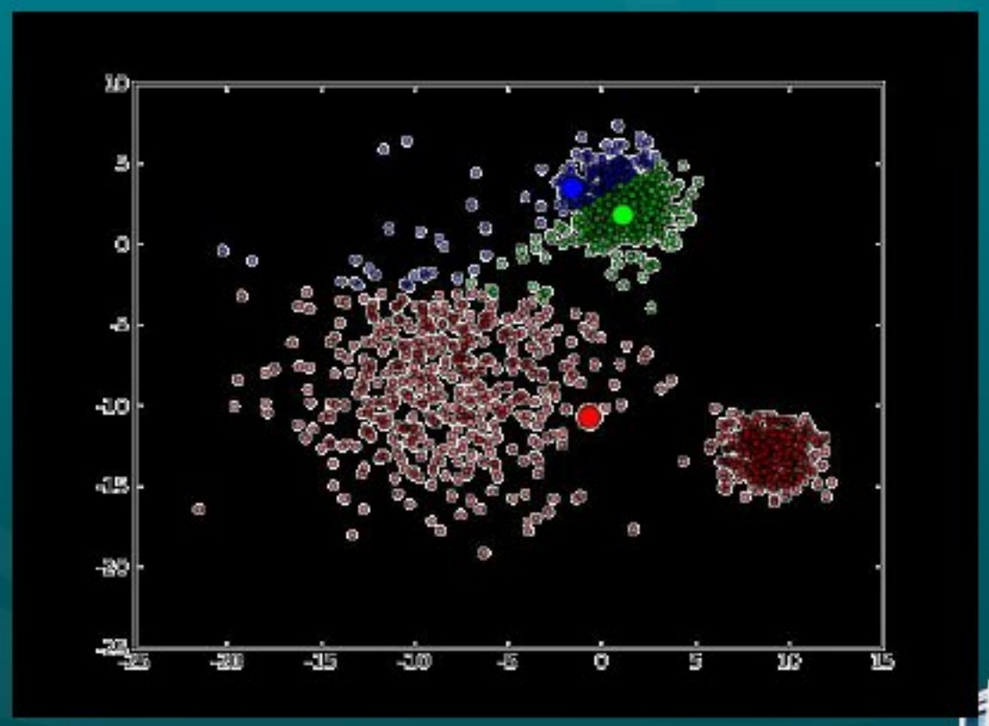
# 初始中心点的划分

讨论初始中心点意义何在？下面的例子一目了然吧？

初始中心点



收敛后



# 改进的算法——二分Kmeans算法

为了克服k均值算法收敛于局部的问题，提出了二分k均值算法。该算法首先将所有的点作为一个簇，然后将该簇一分为二。之后选择其中一个簇继续划分，选择哪个簇进行划分取决于对其划分是否可以最大程度降低SSE值。

伪代码如下：

将所有的点看成一个簇

**Repeat**

从簇表中取出一个簇

（对选定的簇进行多次二分实验）

**for** i=1 to实验次数 **do**

试用基本K均值（k=2），二分选定的簇

**end for**

从实验中选取总SSE最小的两个簇添加到簇表中

**Until** 簇表中包含K个簇

# 谢谢！