Echo + баесовская кластеризация

Дмитрий Грошев, Ирина Горбунова

СПбАУ

04.06.2012

Краткое содержание предыдущих серий

ECHO

- http://uc-echo.sourceforge.net/
- эффективно исправляет ошибки

BayesHammer

- улучшенная версия утилиты hammer
- кластеризует k-меры перед исправлением ошибок
- эффективнее начальной версии

Bayecho

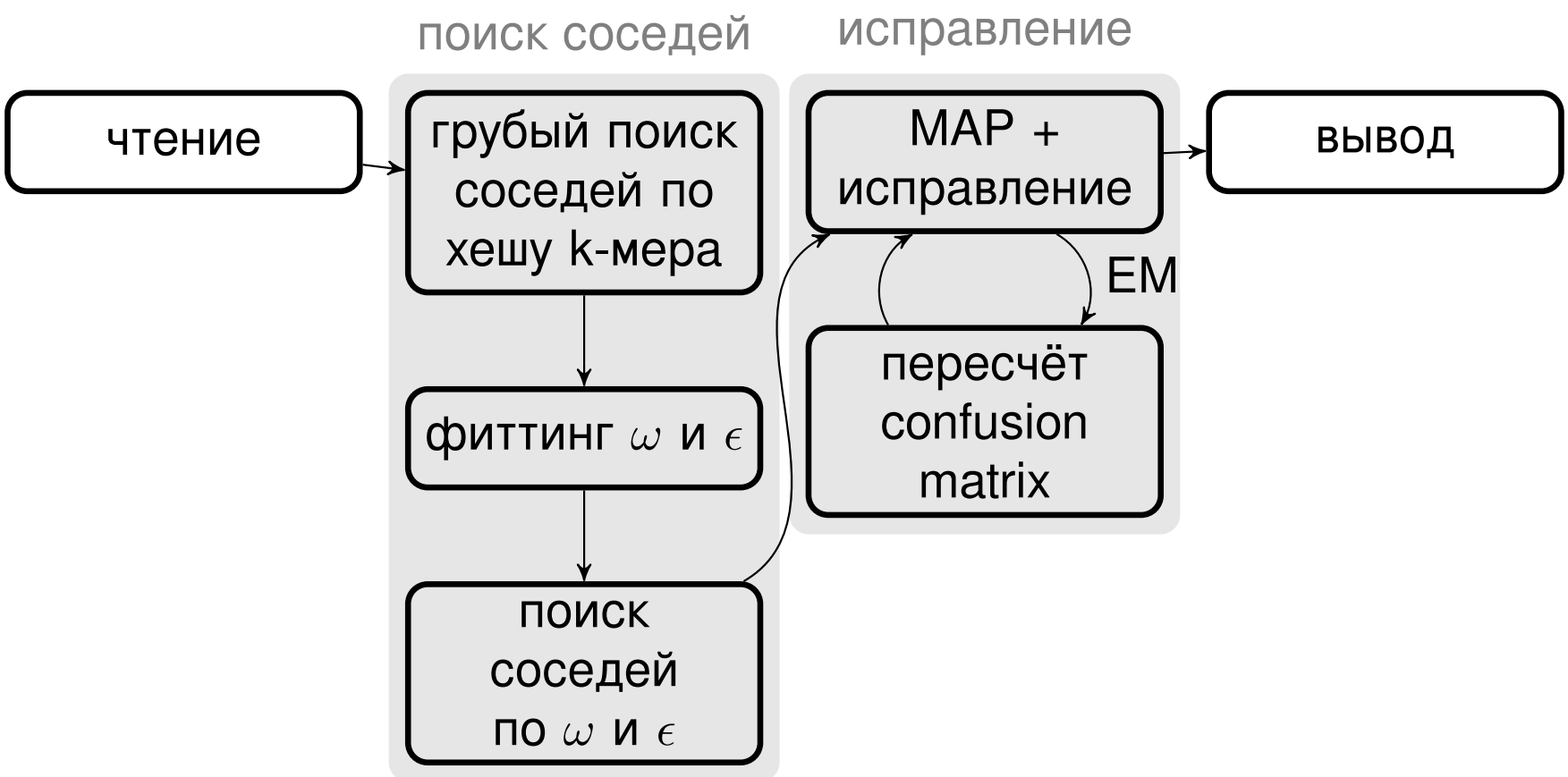
- попытка улучшить ECHO
- кластеризация похожих ридов

Как работает ECHO

2 стадии

- поиск пересекающихся ридов (соседей)
- исправление ошибок

ECHO


поиск соседей
исправление
чтение
грубый поиск соседей по хешу k-мера
фиттинг ω и ϵ
поиск соседей по ω и ϵ
MAP + исправление
EM
пересчёт confusion matrix
вывод

Поиск соседей

- поиск точно совпадающего k-мера hashmap'ом

$$K = max(\frac{L}{5}, \log(\sqrt[4]{N * L}))$$

- подбор ω и ϵ (покрытие конкретного нуклеотида должно иметь распределение Пуассона)
- поиск неточных соседей по ω и ϵ

Исправление

- В группе соседей делается MAP по каждому нуклеотиду и confusion matrix
- На основе исправленных ридов пересчитывается confusion matrix
- Повторять до схождения

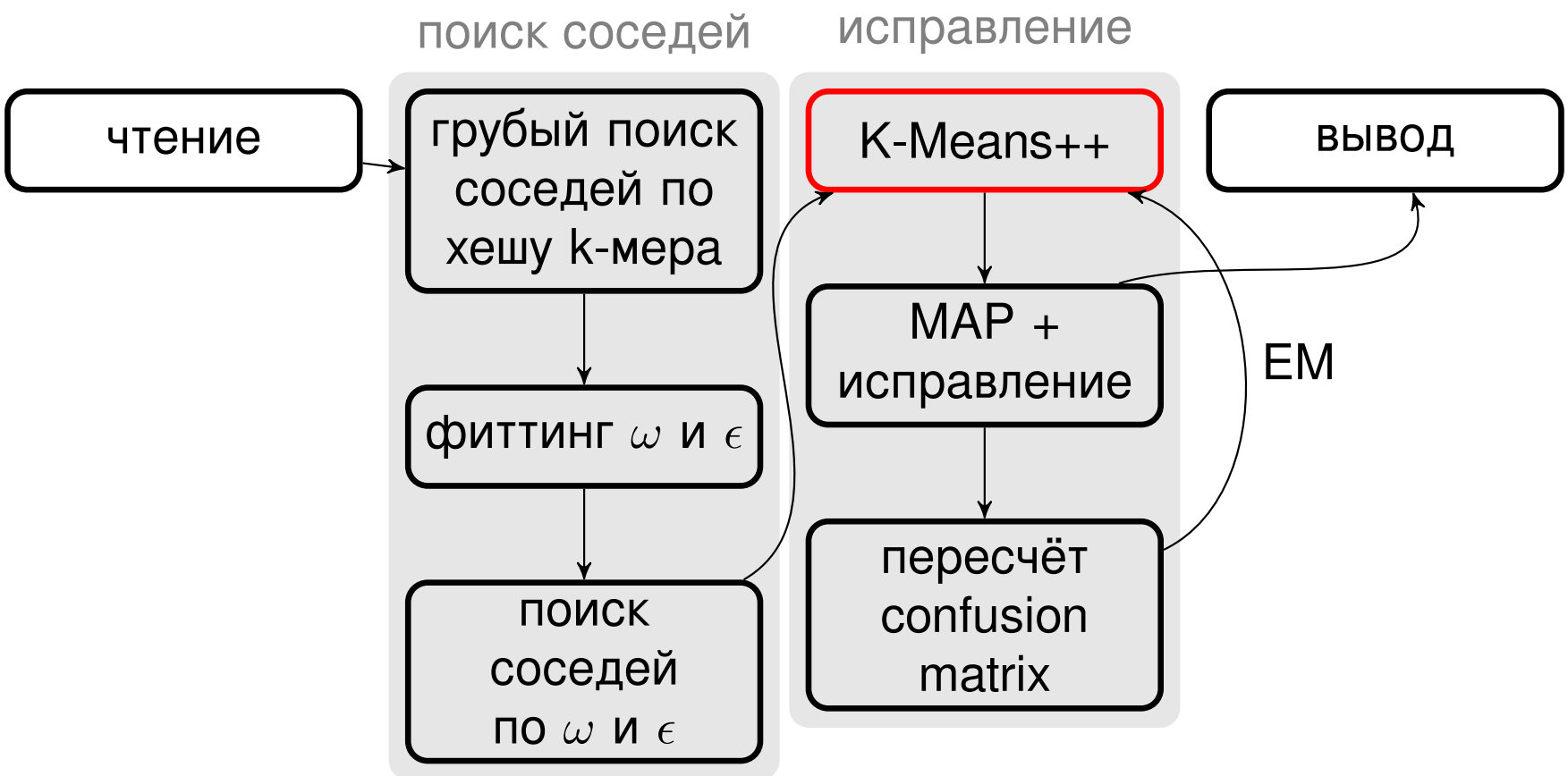
Это EM-алгоритм

Maximum A Posteriori

$$P(A_i) = \prod_{j=1}^{N} \Phi_{READS_{j,i},A}$$

$$\arg\max_{X=A,T,G,C} P(X)$$

Что сделали мы

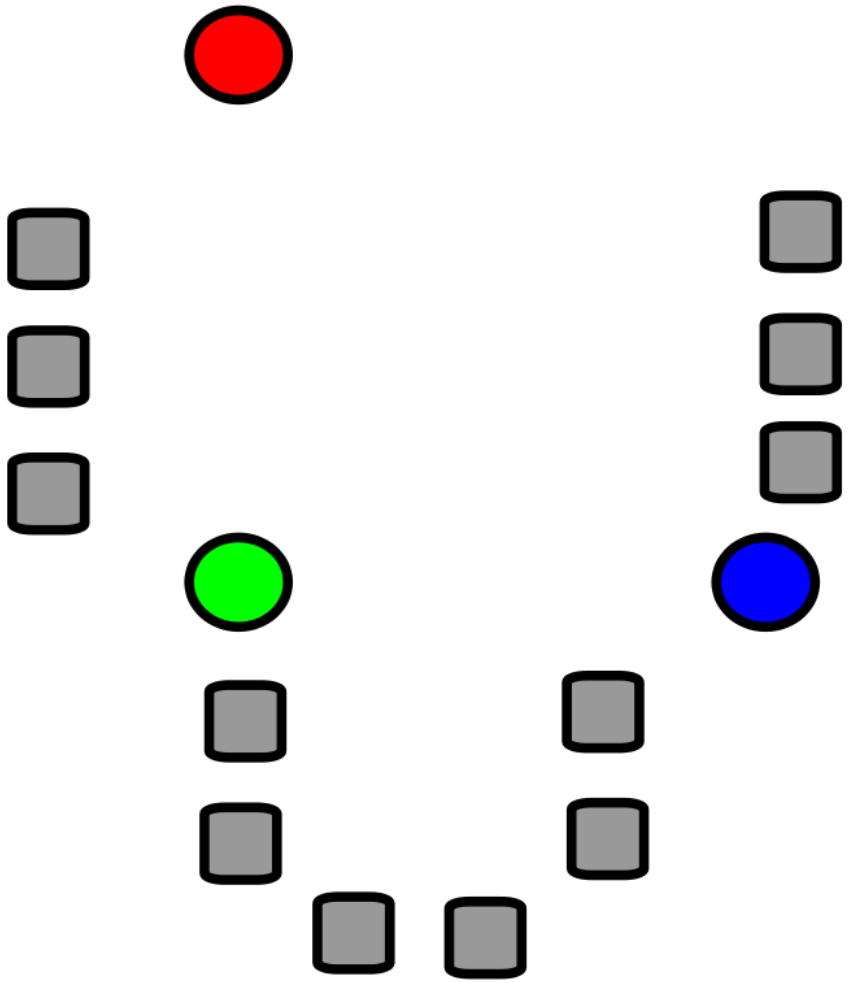

поиск соседей
исправление
чтение
грубый поиск соседей по хешу k-мера
фиттинг ω и ϵ
поиск соседей по ω и ϵ
K-Means++
MAP + исправление
пересчёт confusion matrix
EM
вывод

K-Means

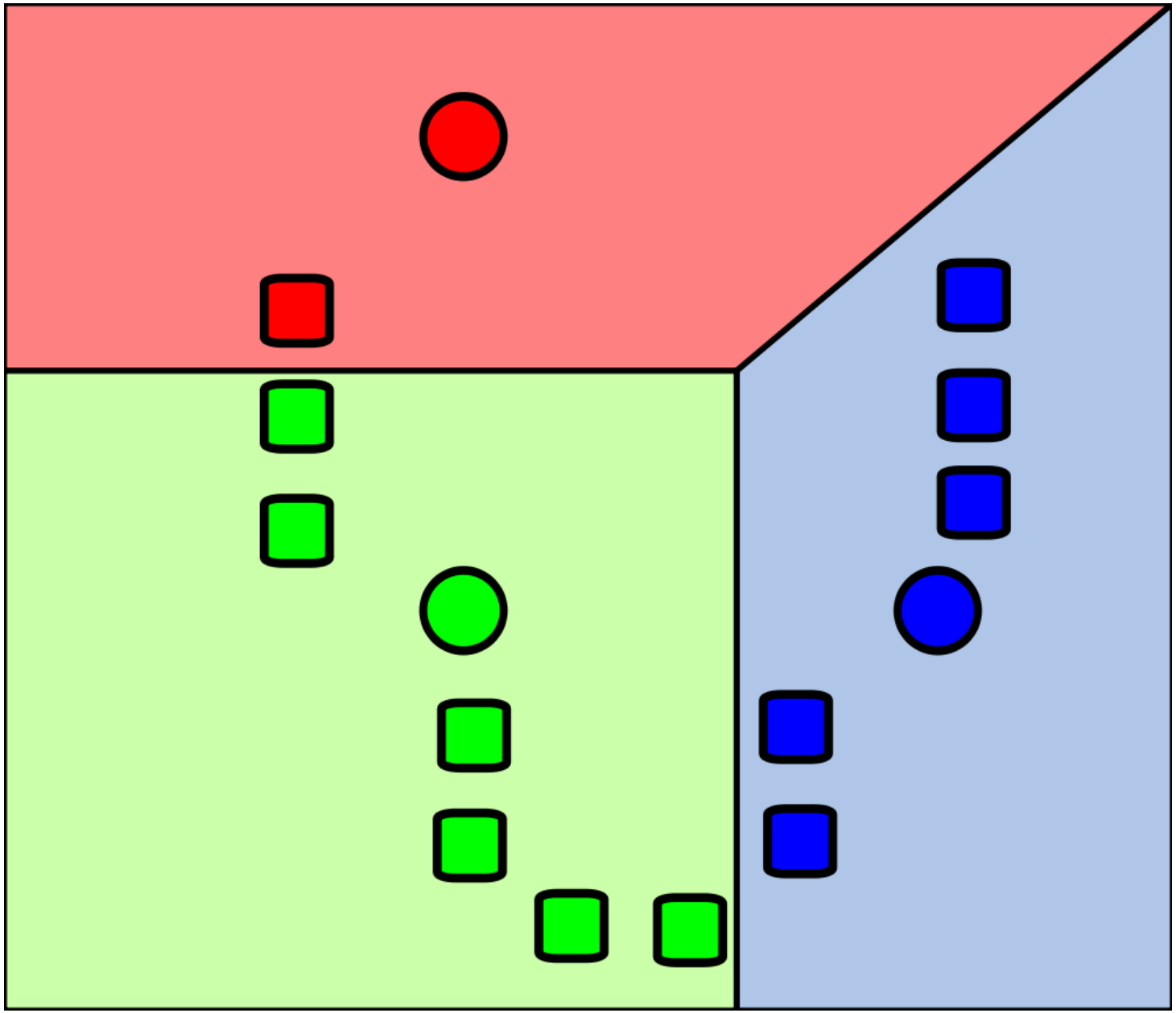
- выбираются произвольные центры
- для каждой точки ищется ближайший центр (диаграма Вороного)
- пересчитываются центры

Это тоже EM

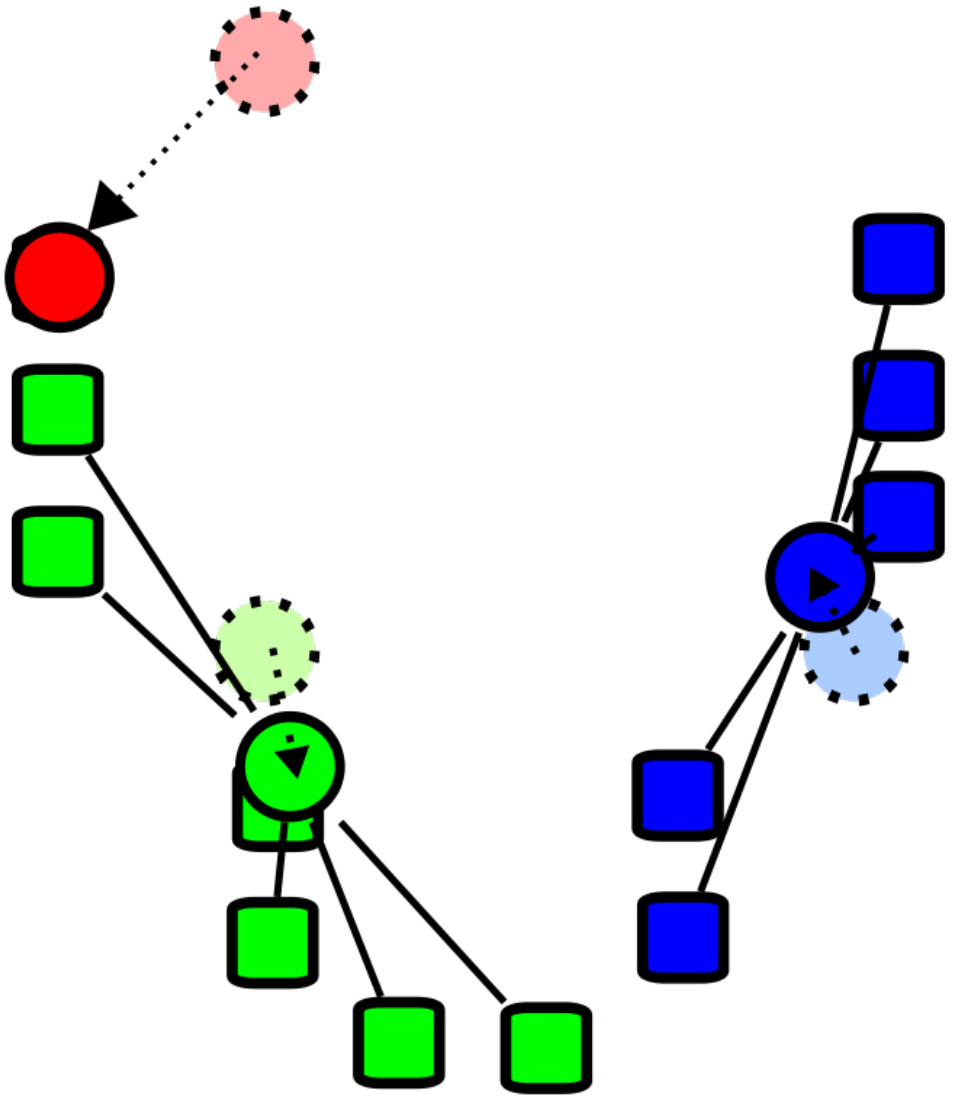
K-Means



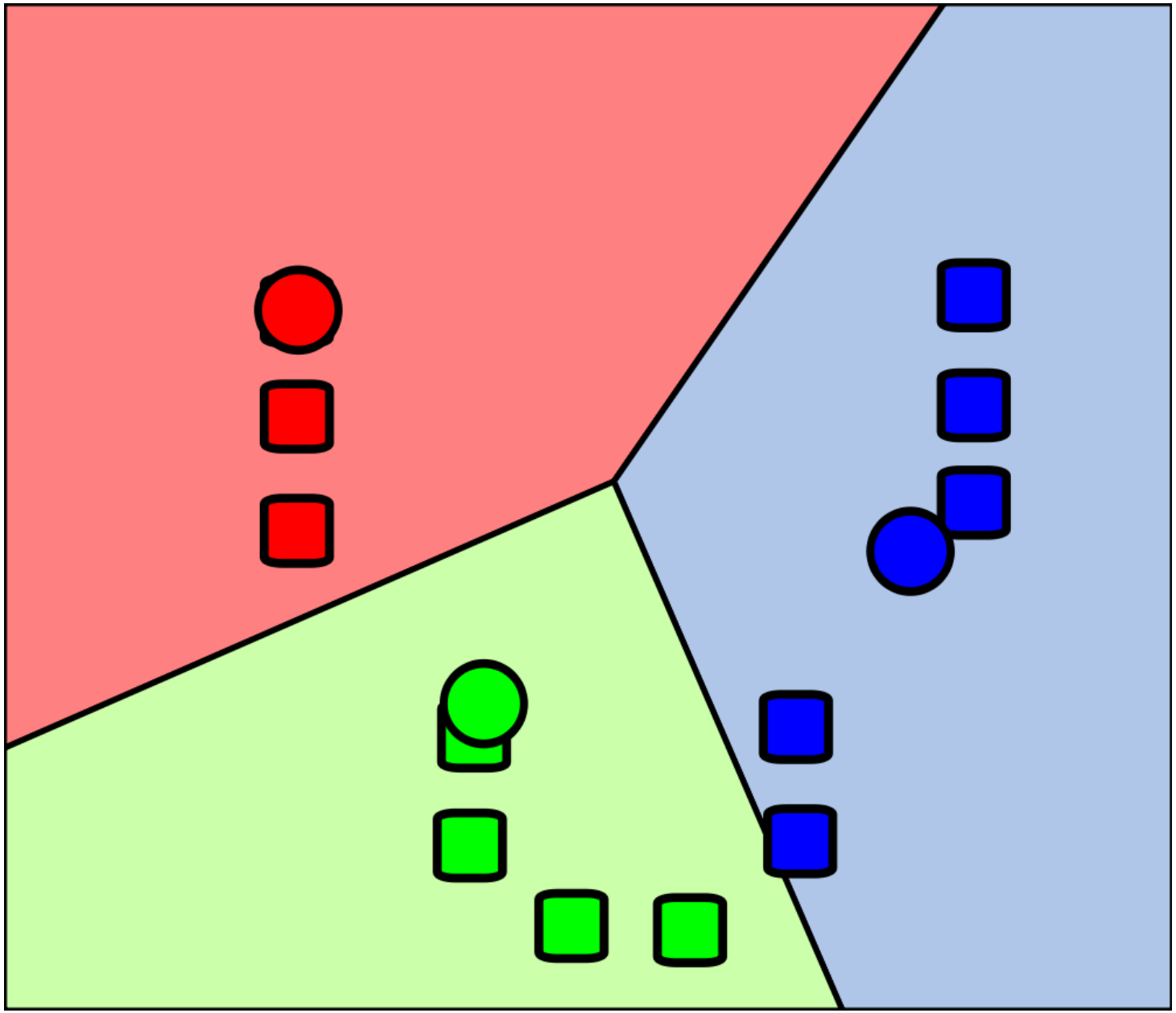
K-Means



K-Means



K-Means



K-Means++

Модификация K-Means с более сложной инициализацией

- центроидом выбирается случайная точка
- подсчитывается расстояние от каждой точки до ближайшего центроида
- новый центроид выбирается из оставшихся точек с вероятностью, пропорциональной квадрату расстояния от остальных центроидов

Проблема

Риды не полностью совпадают (в отличие от k-меров):

```
ATGCATGCA
   TCATGTACGG
 TGCATGCAC
```

Решение

Добавим спецсимвол:

```
ATGCATGCA___
__TCATGTACGG
_TGCATGCAC__
```

Расстояние от _ до любого символа = 0.

Инициализация

Если выбран неполный центроид, дополнить из ближайших (случайный нуклеотид пропорционально расстоянию)

Инициализация

```
___ATGC
GCGCTGC
GCGCTAC
GCGCT__   <---
TATATAT
TCTAT__
TATGTAT
```

Инициализация

```
___ATGC
GCGCTGC  *
GCGCTAC  *
GCGCT__  <---
TATATAT
TCTAT__
TATGTAT
```

Инициализация

```
   ___ATGC
  GCGCTGC
  GCGCTAC
[GCGCTGC] <---
  TATATAT
  TCTAT__
  TATGTAT
```

Поиск ближайшего центроида

Расстояние = 1 - likelihood

Поиск новых центроидов

Снова MAP:

```
group1     group2
___ATGC    TATATAT
GCGCTGC    TCTAT__
GCGCTAC    TATGTAT
GCGCT__
-------    -------
GCGCTGC    TATATAT
```

Результаты

- кластеризация кластеризует
- запускается
- на больших наборах данных ещё не запускали

Вопросы?