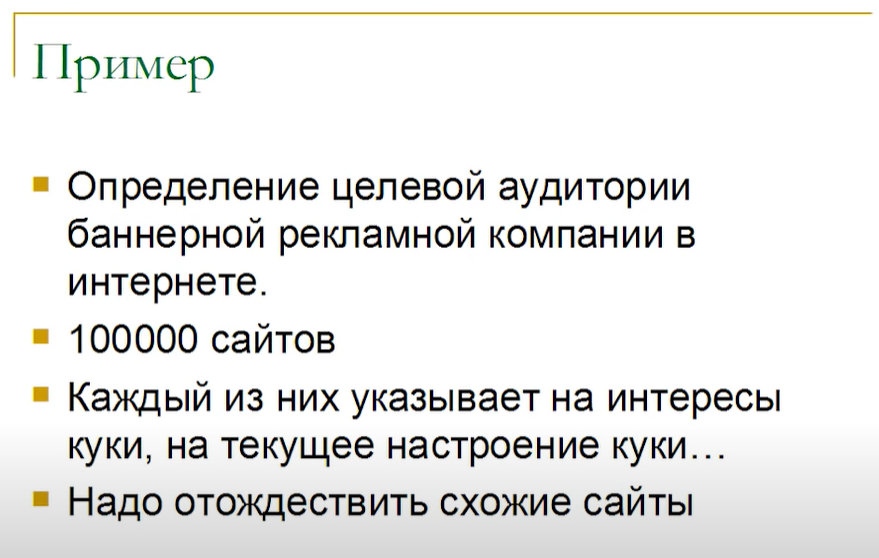


У нас есть информация о посещении людьми сайтов и этих сайтов 100000. Поэтому работать с таким большим объёмом данных будет очень неудобно и долго, даже скорее невозможно, причем почти все сайты одинаковые, только незначительные отличия. Поэтому, чтобы сократить размерность наших данных, мы объедениям сайты в группы (кластеры), тем самым решая множеств проблем.



*Метод К-средних.* Основная идея заключается в том, что на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе

*Коэффициент «силуэт»* вычисляется с помощью среднего внутрикластерного расстояния (a) и среднего расстояния до ближайшего кластера (b) по каждому образцу. Силуэт вычисляется как (b - a) / max(a, b). Поясню: b — это расстояние между a и ближайшим кластером, в который a не входит. Можно вычислить среднее значение силуэта по всем образцам и использовать его как метрику для оценки количества кластеров.