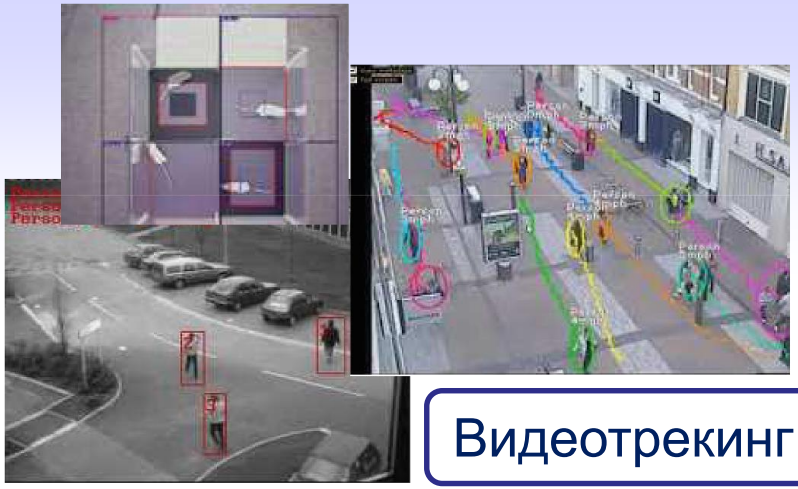
Алгоритм непараметрической кластеризации на основе комбинации сеточного подхода и процедуры среднего сдвига

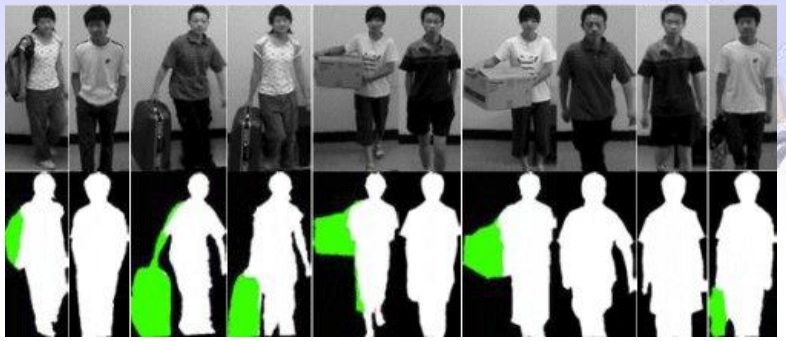
Сергей Рылов

ИВТ

ИВТ СО РАН, Новосибирск

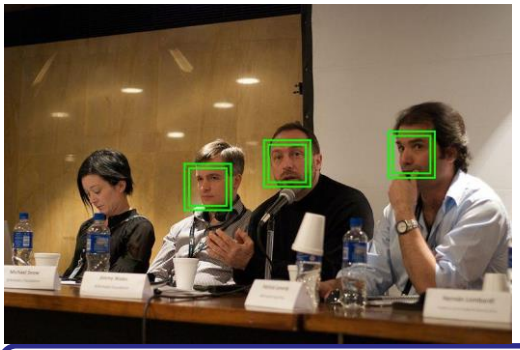


Видеотрекинг

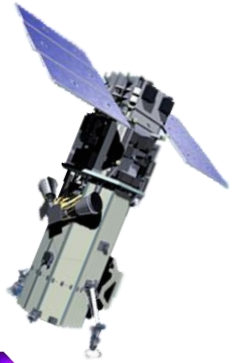


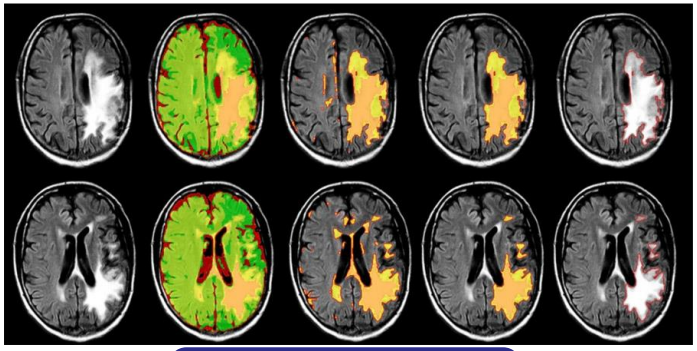
Распознавание образов

Сегментация изображений

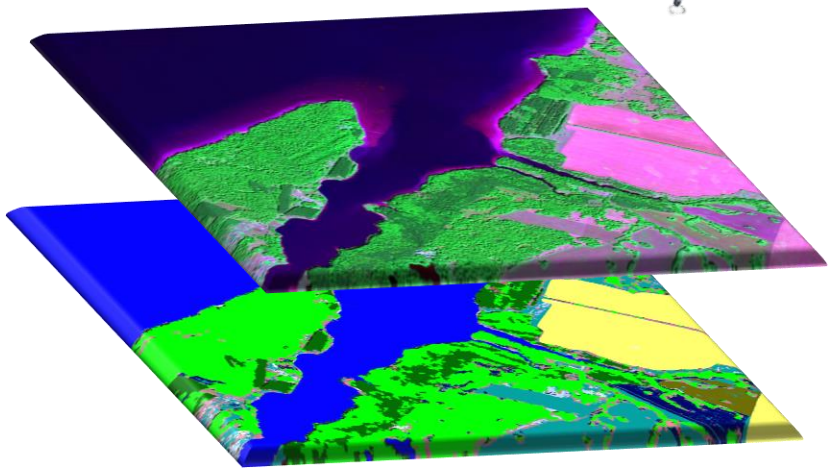


Распознавание лиц





Медицина



Обработка спутниковых снимков

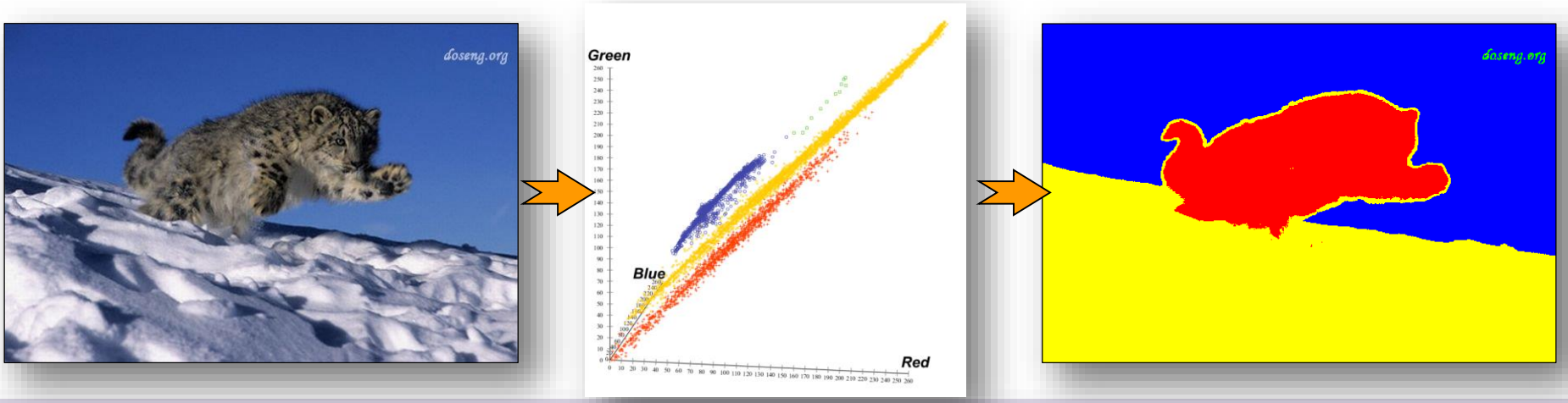
Задача кластеризации

Входные данные: набор классифицируемых объектов в пространстве признаков R^d.

$$X = \{x^{(1)}, \ldots, x^{(N)}\},\ x^{(i)} \in R^d$$

Требуется: согласно некоторому критерию схожести разбить выборку X на подмножества $C_i, i = 1, \ldots, M$:

1) $C_i \neq \varnothing, i = 1, \ldots, M$, 2) $X = \bigcup_{j=\overline{1,k}} C_j$, 3) $C_i \cap C_j = \varnothing \ (i \neq j)$.



Особенности кластеризации спутниковых изображений

- Большой объем обрабатываемых данных (~ 10^6 – 10^8 пикселей)
- Отсутствие априорной информации о количестве и вероятностных характеристиках искомых классов
- Наличие «шума» и выбросов в данных

Алгоритмы кластеризации

	Быстро-действие	Кластеры сложной формы	Устойчивость к шуму	Представители
Методы разбиений	+	–	+/–	K–means, ISODATA, FOREL, CLUSTER
Плотностные параметрические	–	–/+	+	EM, GMDD, MCLUST
Плотностные непараметрические	–	+	+	DBSCAN, OPTICS, Mean Shift
Сеточные	+	+	+	CLIQUE, STING, GRIDCLUS, AMR
Иерархические	–	+/–	–/+	SLINK, CURE, OPTICS, PHA
Нейронные сети	–/+	–	+	SOM, Neural Gas
Спектральные	–	+	+	SC, AFC, SCE

Плотностной подход
Алгоритм кластеризации Mean-shift

Алгоритм кластеризации Mean-shift (процедура среднего сдвига)

Построение непараметрической оценки плотности

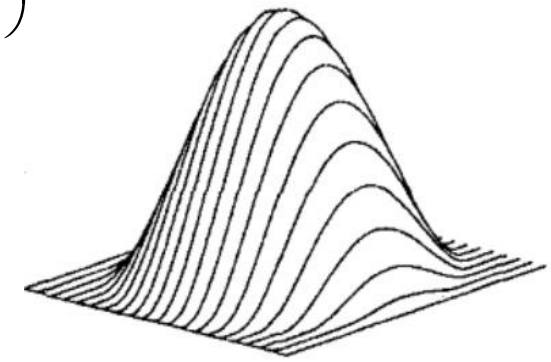
Плотность оценивается как суммарное влияние элементов выборки:

$$\hat{f}_h(x) = \frac{1}{Nh^d} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h} \right)$$

Вклад каждого элемента описывается с помощью колоколообразной функции (ядра) $K(x)$, зависящей от расстояния до этого элемента

Ядро Епанечникова: $K_{Ep}\left(\frac{x - x_i}{h} \right) = \left(1 - \frac{\|x - x_i\|^2}{h^2} \right) \cdot I\left(\|x - x_i\| \le h \right)$

Ядро Гаусса: $K_G\left(\frac{x - x_i}{h} \right) = \exp\left(-\frac{\|x - x_i\|}{2h^2} \right)$



где h – параметр сглаживания.

Алгоритм кластеризации Mean-shift (процедура среднего сдвига)

Процедура «среднего сдвига» (Mean-shift)

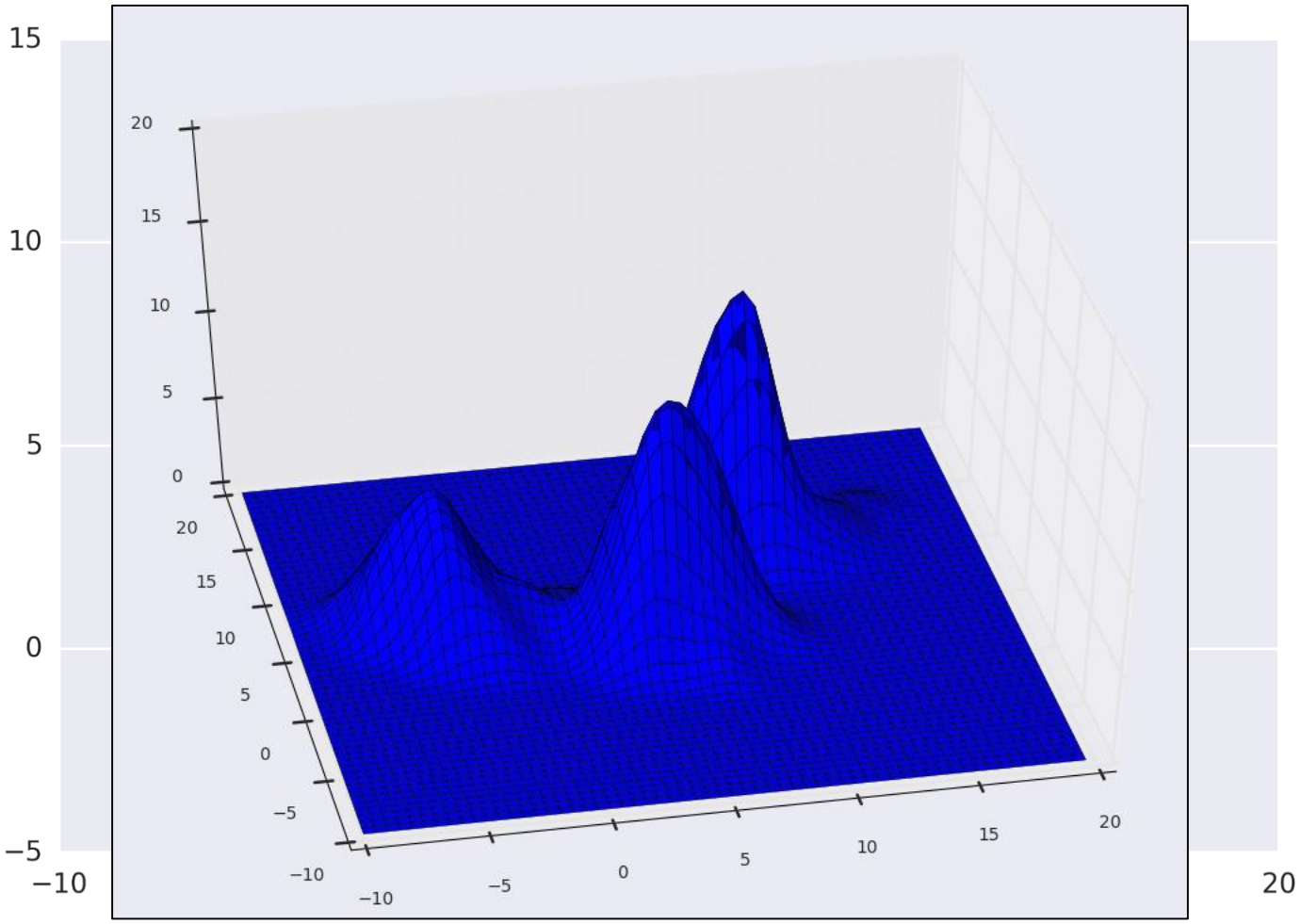
– итеративная процедура, начиная с точки x_0, последовательно перемещается в точку сдвига $x_{k+1} = m(x_k)$ вплоть до сходимости к локальному максимуму плотности.

$$m(x) = \frac{\sum_{i=1}^{N} x_i \cdot K(x - x_i)}{\sum_{i=1}^{N} K(x - x_i)}$$

Вектор $\left(m(x) - x\right)$ называется *вектором «среднего сдвига»*, его направление совпадает с направлением максимального роста плотности в точке x.

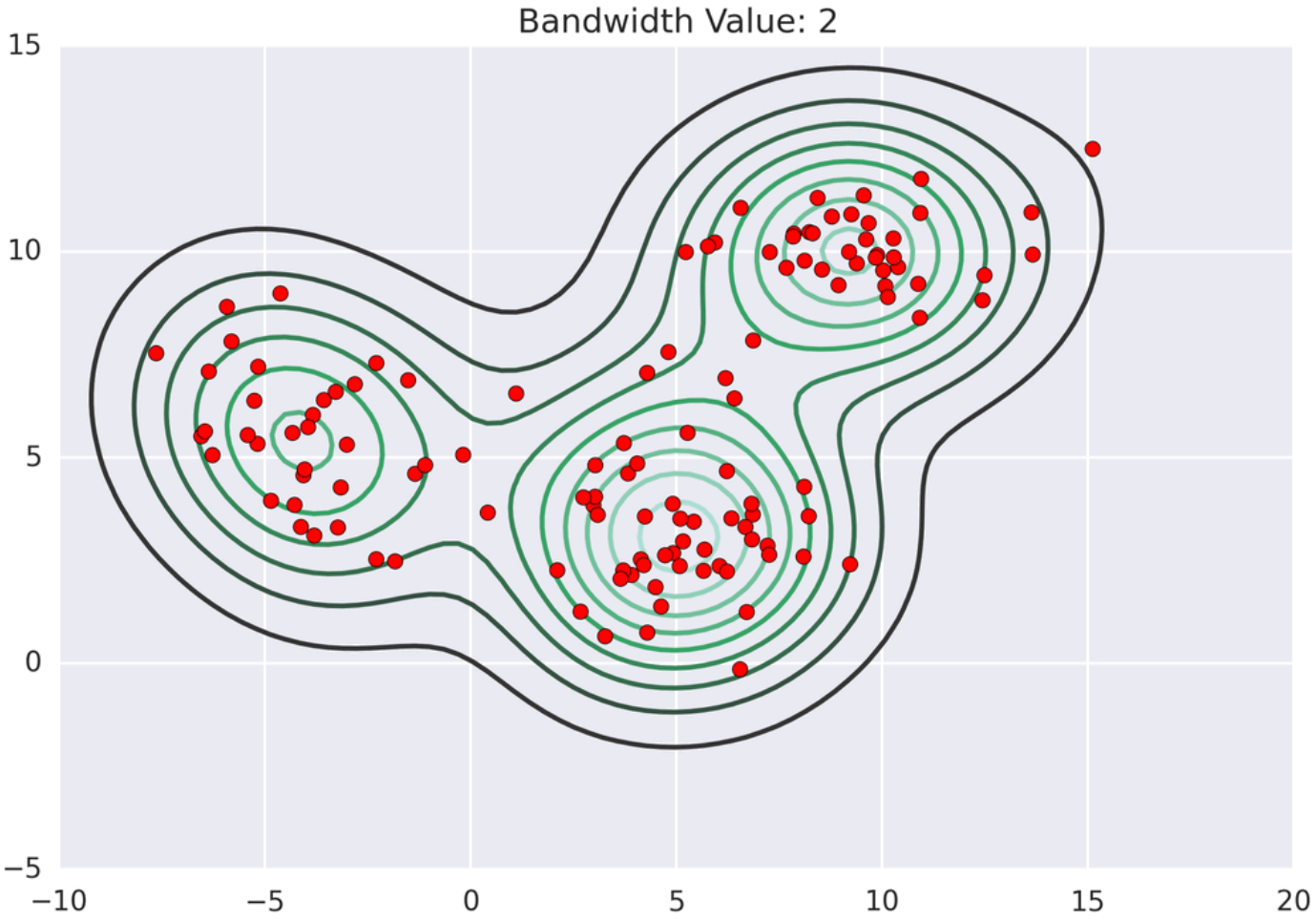
Кластеры соответствуют локальным максимумам функции оценки плотности (модам), к которым сходится процедура.

Иллюстрация работы Mean-shift



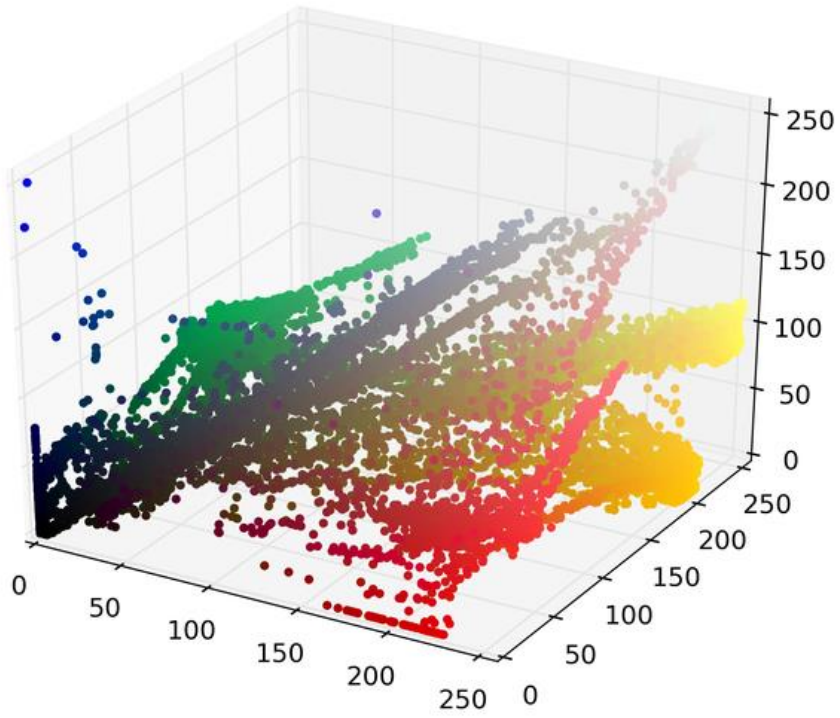
By Matt Nedrich: https://spin.atomicobject.com/2015/05/26/mean-shift-clustering/

Иллюстрация работы Mean-shift



By Matt Nedrich: https://spin.atomicobject.com/2015/05/26/mean-shift-clustering/

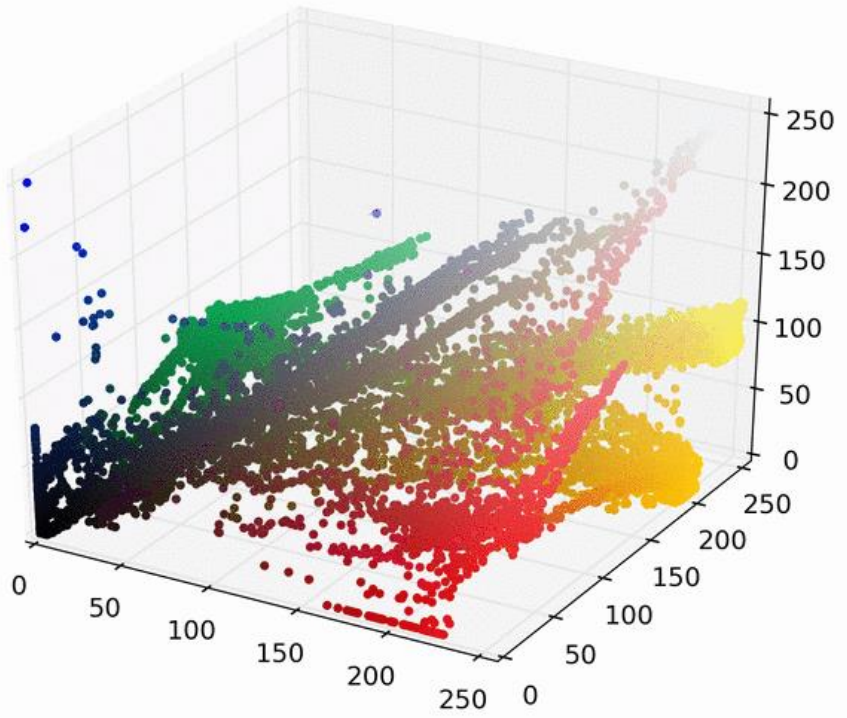
Иллюстрация работы Mean-shift



By Matt Nedrich: https://spin.atomicobject.com/2015/05/26/mean-shift-clustering/

Иллюстрация работы Mean-shift

➢ **Высокая вычислительная сложность!**



By Matt Nedrich: https://spin.atomicobject.com/2015/05/26/mean-shift-clustering/

Сеточный подход
Алгоритм кластеризации HCA

Сеточный алгоритм кластеризации NCA

Сеточная модель описания данных

✓ Высокая скорость работы и выделение кластеров сложной формы

Сеточная структура – разбиение d-мерного пространства признаков гиперплоскостями на *клетки*:

$$x^j = \frac{i}{m}\left(\max_{x_l \in X} x_l^j - \min_{x_l \in X} x_l^j\right) + \min_{x_l \in X} x_l^j$$

где m – параметр масштаба сетки

$$i = 0,\ldots, m, \quad j = 1,\ldots, d$$

Плотность клетки D_B определяется как число элементов, попавших в клетку B

Клетка B_i непосредственно связна с B_j: $B_i \to B_j \Leftrightarrow B_j = \arg\max_{B - \text{смежная с } B_i} D_B$

→ Вводится отношение связности (является отношением эквивалентности)

→ Порождает разбиение непустых клеток на компоненты связности

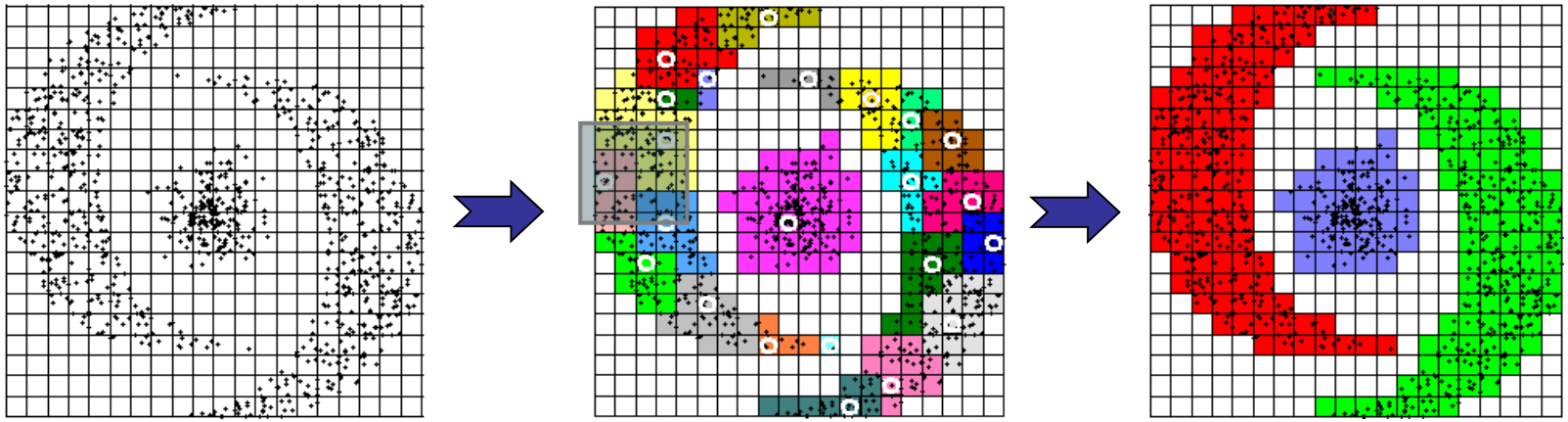
Клетка представитель компоненты связности G: $Y(G) = \max_k \left\{ B_k \mid D_{B_k} = \max_{B_l \in G} D_{B_l} \right\}$

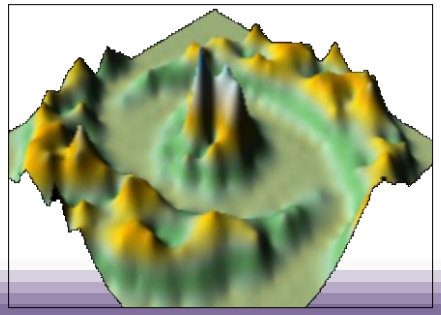
Компоненты связности – локальные сгустки плотности

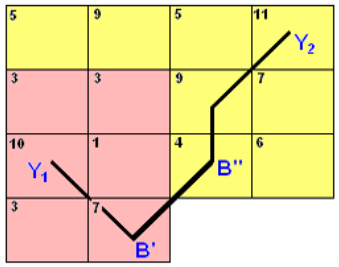
Сеточный алгоритм кластеризации HCA

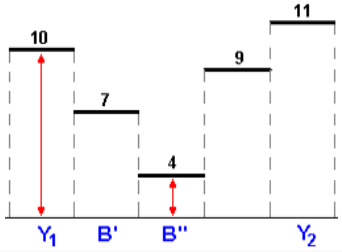
Основные этапы работы алгоритма HCA(m, T)

1) Формирование сеточной структуры. Для каждого вектора данных $x_i \in X$ определяется содержащая его клетка и вычисляются плотности всех клеток.
2) Выделение компонент связности $G_1,\ldots,G_S$ и их клеток представителей $Y_1,\ldots,Y_S$.
3) Построение иерархии на множестве компонент связности.



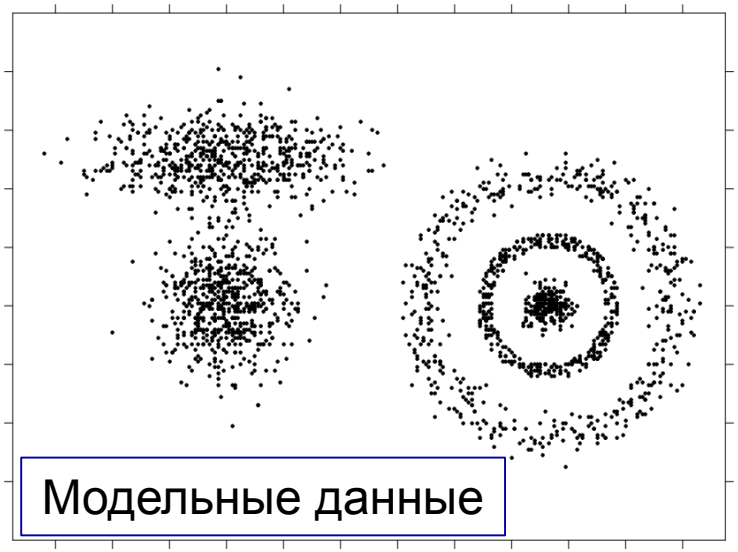


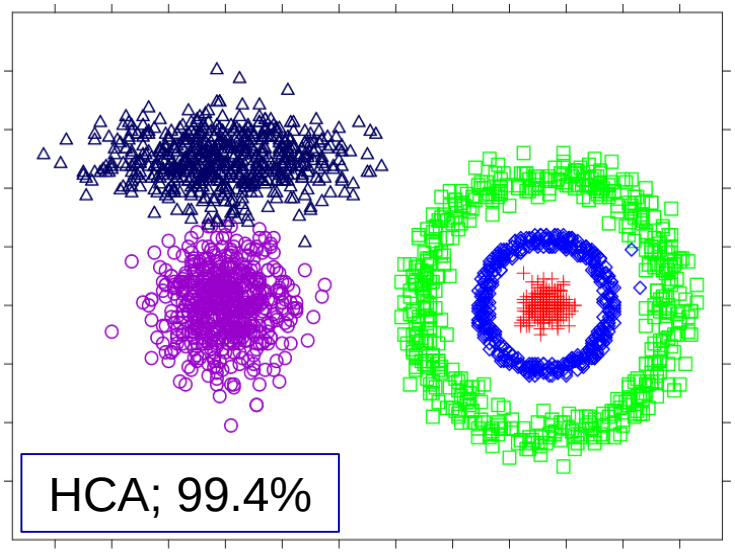


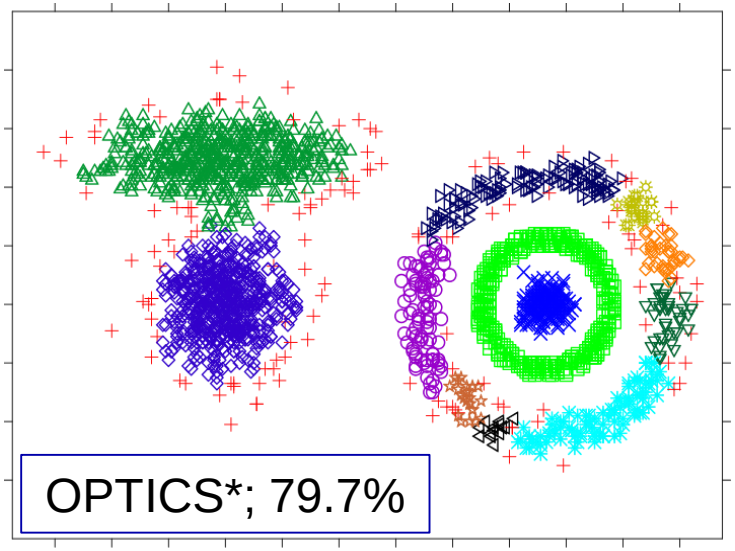


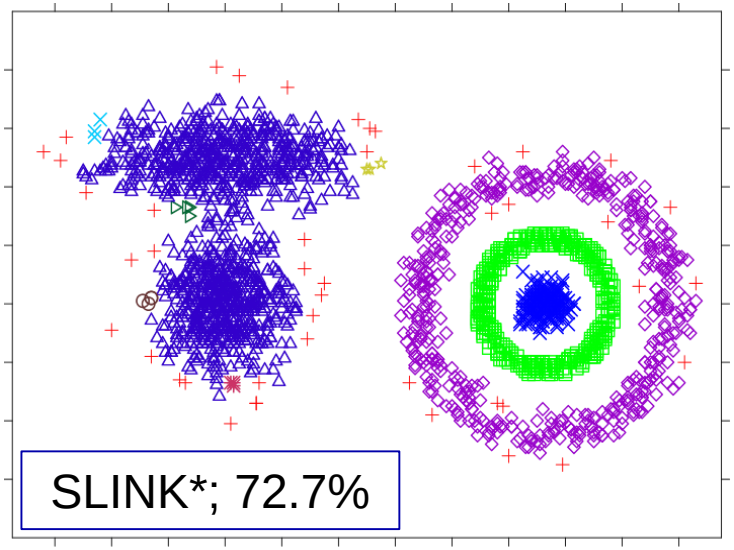
$$h_{1,2} = \frac{\min(D_{B'}, D_{B''})}{\min(D_{Y_1}, D_{Y_2})}$$

Сеточный алгоритм кластеризации HCA





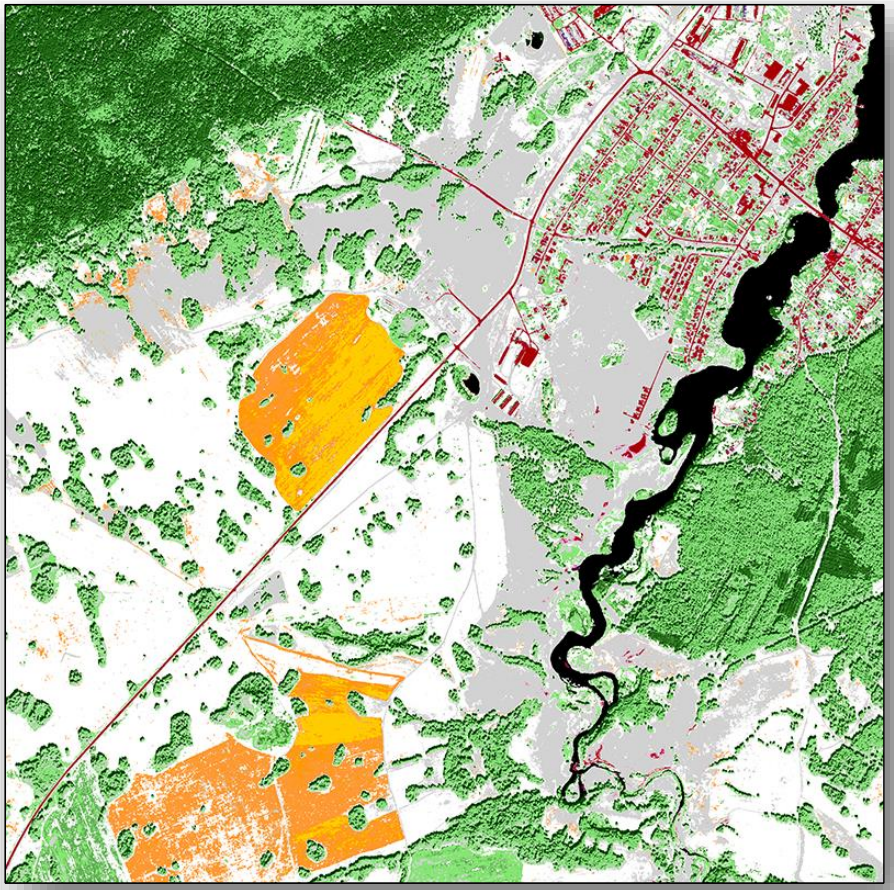




** Алгоритмы из программного пакета ELKI (elki.dbs.ifi.lmu.de)*

Сеточный алгоритм кластеризации HCA

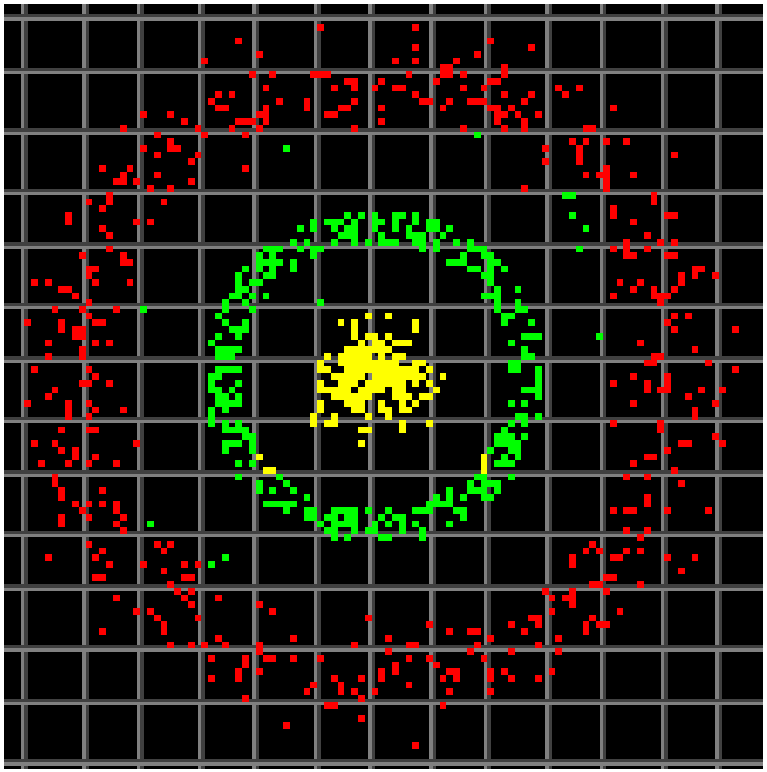
RGB-композит (каналы 5, 3, 2)
изображения WorldView-2
размера 2048 × 2048 пикселей



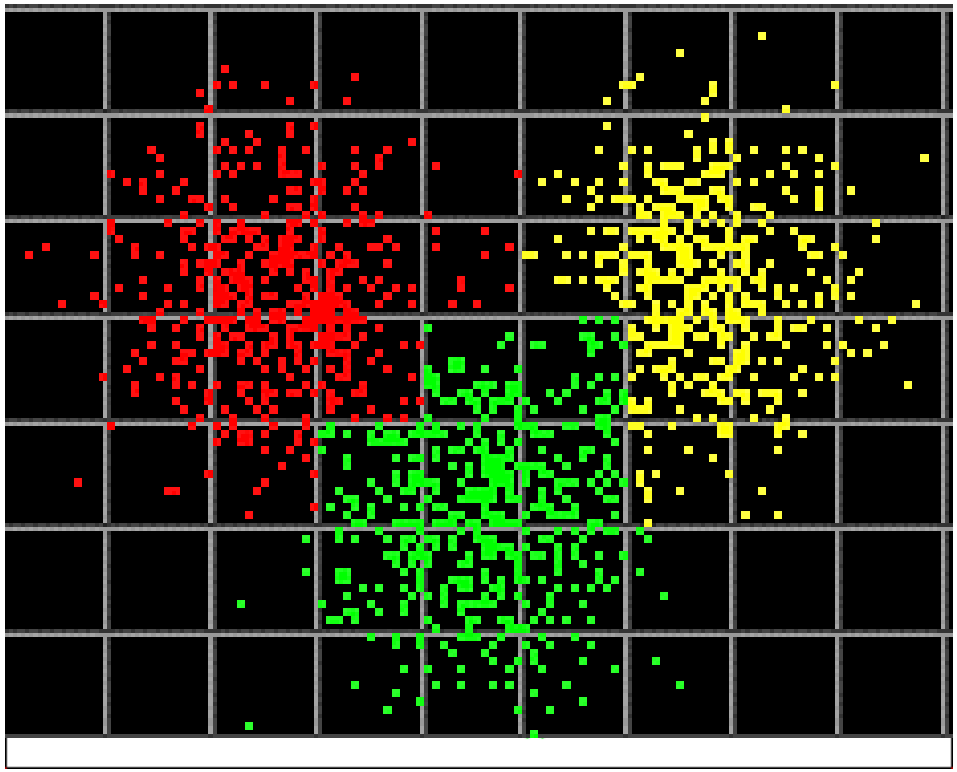
Кластеризация алгоритмом HCA
по четырем каналам: 1, 2, 4, 7;
время обработки – 0.3 с

Проблема сеточной структуры

- **Точность разделения кластеров зависит от сеточной структуры, может приводить к ошибкам, особенно при неудачном выборе параметра масштаба сетки**



m = 30



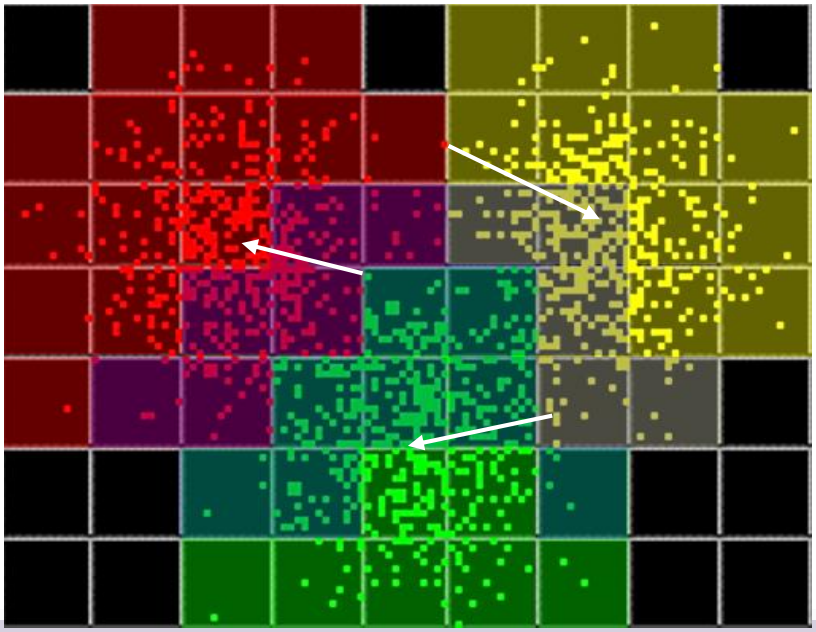
m = 20

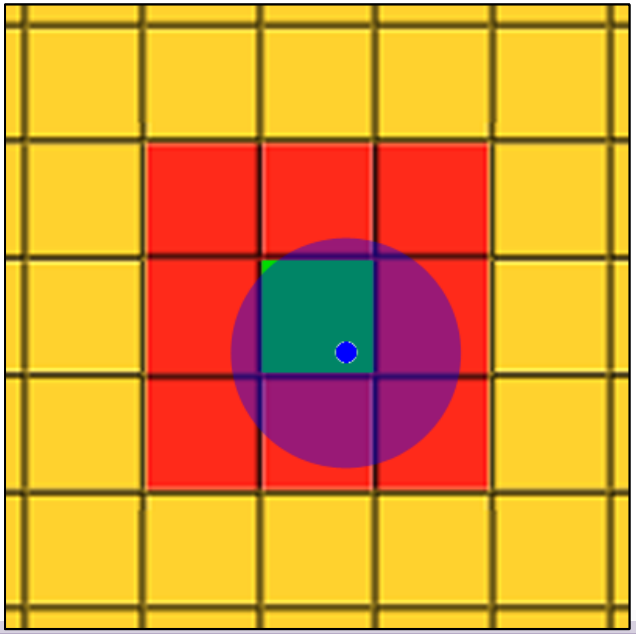
Комбинация сеточного подхода и процедуры среднего сдвига

Комбинация сеточного подхода и процедуры среднего сдвига

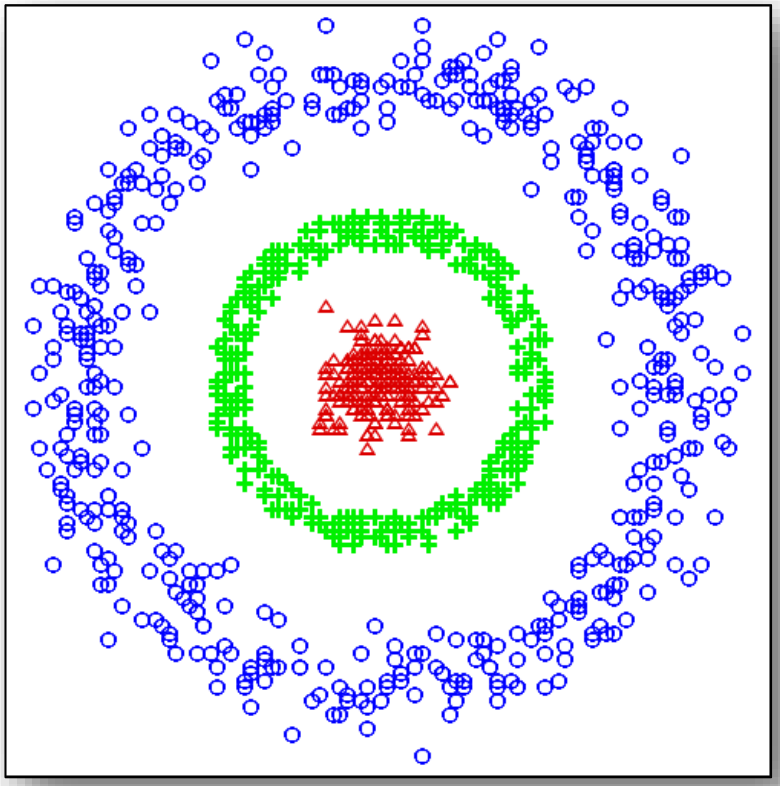
Новый алгоритм кластеризации HCA-MS

1. Выполнение алгоритма HCA (с заданным параметром сетки m).
2. Индексирование элементов данных по клеткам (быстрый доступ к списку элементов произвольной клетки).
3. К элементам граничных клеток применяется процедура «среднего сдвига» с ограниченным ядром (h = ширина клетки).

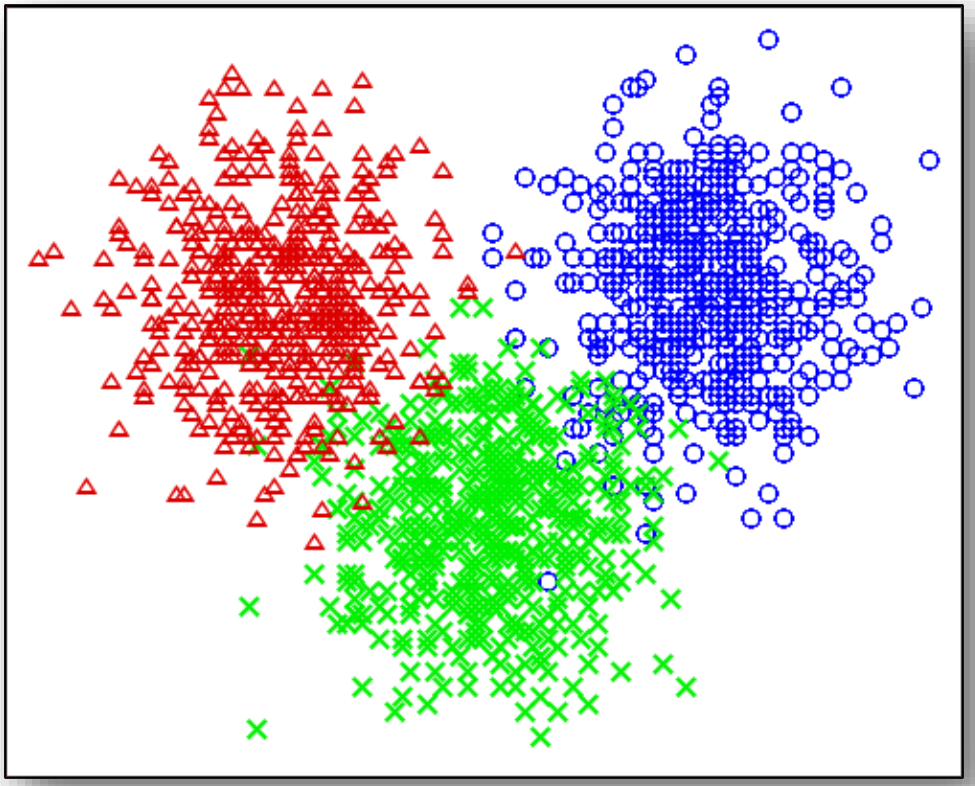




Алгоритм кластеризации HCA-MS
Экспериментальные исследования



Эталонное разбиение



Эталонное разбиение

Сравнение времени работы алгоритмов HCA, HCA-MS и Mean-shift

Размер изображения (МП)	Число каналов	HCA		HCA-MS		Mean-shift	
		m=25	m=32	m=25	m=32	h=20	h=25
1	3	0.05	0.05	0.7	0.3	52	58
5	3	0.1	0.11	1.2	0.7	67	90
14	3	0.2	0.2	7.3	3.7	388	563
4	4	0.2	0.3	71	27	4138	6009
12	4	0.4	0.5	458	350	62388	97121

Время указано в секундах

- ✓ Предложенный алгоритм позволяет повысить качество кластеризации сеточного алгоритма HCA
- ✓ HCA-MS обладает высокой вычислительной эффективностью, которая позволяет обрабатывать мультиспектральные спутниковые изображения большого размера

** Вычисления выполнялись на ПЭВМ с четырехядерным процессором Intel Core i5, 3.5 ГГц*


Спасибо за внимание!

Литература

1. *Пестунов И.А., Рылов С.А., Бериков В.Б.* Иерархические алгоритмы кластеризации для сегментации мультиспектральных изображений // Автометрия. – 2015. – Т. 51. – № 4. – С. 12-22.
2. *Рылов С.А.* Модельные данные для кластеризации [Электронный ресурс]. URL: https://drive.google.com/open?id=0ByK9GtU5ExExRnZwdFNmRHRWdFk.
3. *Рылов С.А., Пестунов И.А.* Использование графических процессоров NVIDIA при кластеризации мультиспектральных данных сеточным алгоритмом CCA // Интерэкспо ГЕО-Сибирь. – 2015. – Т. 4. – № 2. – С. 51-56.
4. *Рылов С.А., Новгородцева О.Г., Дубровская О.А., Пестунов И.А.* Иерархические алгоритмы кластеризации мультиспектральных изображений и их использование при создании тематических карт паводковой обстановки // Сборник трудов Всероссийской конференции «Обработка пространственных данных в задачах мониторинга природных и антропогенных процессов». – Новосибирск, 2015. – С. 165-171. [Электронный ресурс]. – URL: http://conf.nsc.ru/files/conferences/SDM-2015/294652/SDM-2015%20Thesis.pdf.
5. *Пестунов И.А., Синявский Ю.Н.* Анализ и синтез сигналов и изображений непараметрический алгоритм кластеризации данных дистанционного зондирования на основе grid-подхода // Автометрия. – 2006. – Т. 42. – №. 2. – С. 90-99.
6. *Cheng Y.* Mean shift, mode seeking, and clustering // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 1995. – Vol. 17. – No. 8. – P. 790-799.