11回
クラスタリング

- クラスタリングの概要
- 階層的クラスタリング
- 非階層的クラスタリング

クラスタリングの概要

クラスタリング

- クラスタリング（Clustering）
  - データの集まりをデータ間の類似度に基づいて複数の部分集合（クラスタ）に切り分ける処理
  - 生物、化学、 経済学、社会学、考古学, 心理学等、様々な分野で使われる処理
- クラスタ（Cluster）
  - 群れ、集団、花やブドウ等の房といった意味

クラスタリングの例

- データの集まりをデータ間の類似度に基づいて複数のクラスタに切り分ける

様々なデータ

模様の違いによるクラスタリング

大きさの違いよるクラスタリング

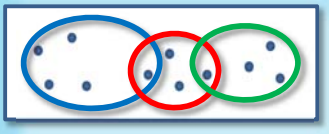
クラスタリング手法の分類（階層、非階層）

- 階層的クラスタリング（Hierarchical Clustering）
  1. 個々のデータを１つのクラスタとして設定
  2. クラスタ間の類似度を計算、最も類似しているクラスタを併合
  3. すべてのクラスタが１つのクラスタになるまで処理を繰り返す
- 非階層的クラスタリング
  (Non-Hierarchical Clustering Method； Partitional Clustering)
  - 分割最適化手法とも呼ばれる
  - 分割の状態を表す関数を使い、関数の値が最適解となるように探索を行う

クラスタリング手法の分類（ハード、ソフト）

- ハードクラスタリング
  - すべてのデータが、単一クラスタに属するようなクラスタリング

- ソフトクラスタリング
  - データが複数のクラスタに属することができる
  - ファジィ（Fuzzy）クラスタリングとも呼ばれる



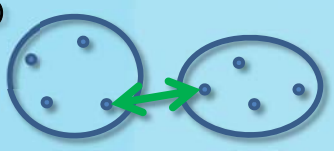
階層的クラスタリング

階層的クラスタリング

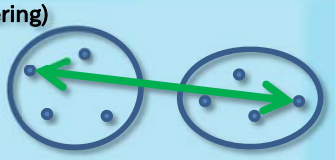
1．個々のデータを１つのクラスタとして設定
2．クラスタ間の類似度を計算
　　最も類似しているクラスタを併合
3．すべてのクラスタが
　　１つのクラスタになるまで処理を繰り返す

クラスタ間の類似度を求める方法

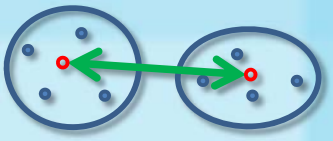
- **最近隣法** (Nearest Neighbor Method; Single Linkage Clustering)
  最も近いデータの距離を
  この２つのクラスタの距離と定義



- **最遠隣法** (Furthest Neighbor Method; Complete Linkage Clustering)
  最も遠くなるデータの距離を
  この２つのクラスタの距離と定義

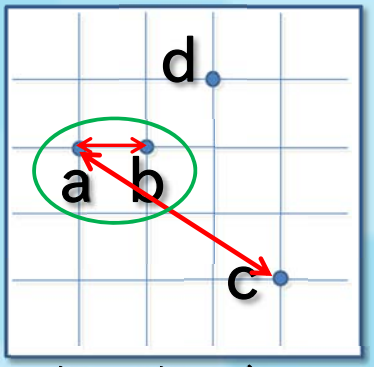


- **重心法** (Centroid Clustering)
  ２つのクラスタのそれぞれの重心を求めて
  重心間の距離をクラスタの距離と定義



階層的クラスタリング （例１-1）

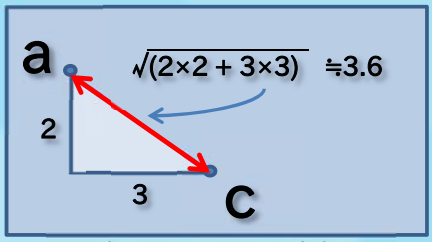
（例１）２次元空間にある４個のデータをクラスタリング



例：４個のデータ

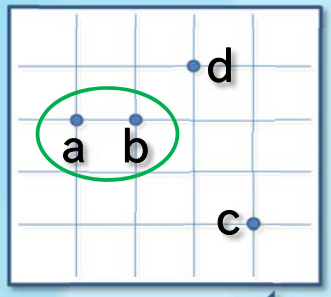
a～b ＝ 1.0
a～c ≒ 3.6
a～d ≒ 2.2
b～d ≒ 1.4
b～c ≒ 2.8
c～d ≒ 3.2

データ間の類似度

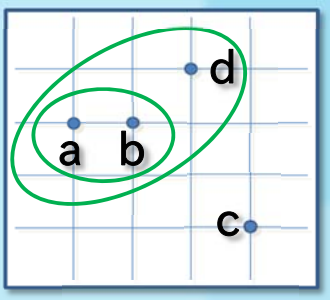


（ユークリッド距離）

階層的クラスタリング （例1-2）

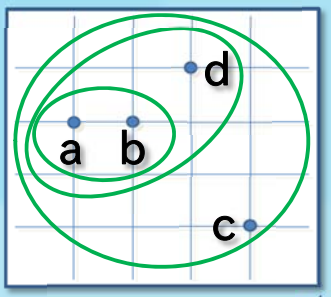


a～b = 1.0
a～c = 3.6
a～d = 2.2
b～d = 1.4
b～c = 2.8
c～d = 3.2

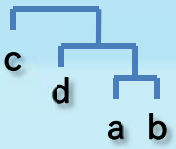


(ab)～c = 2.8
(ab)～d= 1.4
c～d=3.2

最近隣法で！



((ab)d) ～c= 2.8



階層的クラスタリング （例2-1）

	商品1	商品2	商品3	商品4	商品5
A店	79	80	58	45	52
B店	58	80	70	75	98
C店	67	88	60	50	70
D店	48	50	20	55	77
E店	57	78	80	60	72
F店	39	24	57	87	71

6つの店にある5種類の商品価格（円）

	A	B	C	D	E	F
A						
B	60					
C						
D						
E						
F						

店舗間の距離

A店とB店の類似度（ユークリッド距離）

$$\sqrt{(79-58)^2+(80-80)^2+(58-70)^2+(45-75)^2+(52-98)^2}$$
$$=\sqrt{441+0+144+900+2116}$$
$$=\sqrt{3601}$$
$$\approx 60.0$$

階層的クラスタリング （例2-2）

	A	B	C	D	E	F
A	0	60	24	63	40	83
B	60	0	41	66	32	67
C	24	41	0	59	27	79
D	63	66	59	0	67	56
E	40	32	27	67	0	67
F	83	67	79	56	67	0

店舗間の距離（類似度）

※ 値が小さい程、類似しているので
非類似度とよぶ場合もある

	A	B	C	D	E	F
A						
B	60					
C	24	41				
D	63	66	59			
E	40	32	27	67		
F	83	67	79	56	67	

店舗間の距離（類似度）

簡略化した表現

階層的クラスタリング （例2-3）

	A	B	C	D	E	F
A						
B	60					
C	24	41				
D	63	66	59			
E	40	32	27	67		
F	83	67	79	56	67	

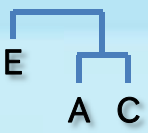
店舗間の距離（類似度）

①

	A	B	C	D	E	F
A						
B	60					
C	24	41				
D	63	66	59			
E	40	32	27	67		
F	83	67	79	56	67	

②

	(AC)	B	D	E	F
(AC)					
B	41				
D	59	66			
E	27	32	67		
F	79	67	56	67	



最近隣法で
クラスタリング！

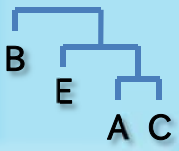
階層的クラスタリング （例2-4）

	A	B	C	D	E	F
A						
B	60					
C	24	41				
D	63	66	59			
E	40	32	27	67		
F	83	67	79	56	67	

店舗間の距離（類似度）

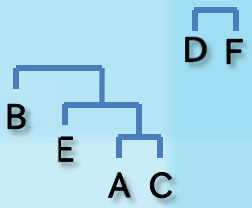
③

	((AC)E)	B	D	F
((AC)E)				
B	32			
D	59	66		
F	67	67	56	



④

	(((AC)E)B)	D	F
(((AC)E)B)			
D	59		
F	67	56	



最近隣法で
クラスタリング！

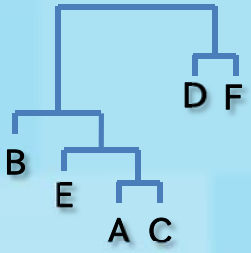
階層的クラスタリング （例2-5）

	A	B	C	D	E	F
A						
B	60					
C	24	41				
D	63	66	59			
E	40	32	27	67		
F	83	67	79	56	67	

店舗間の距離（類似度）

⑤

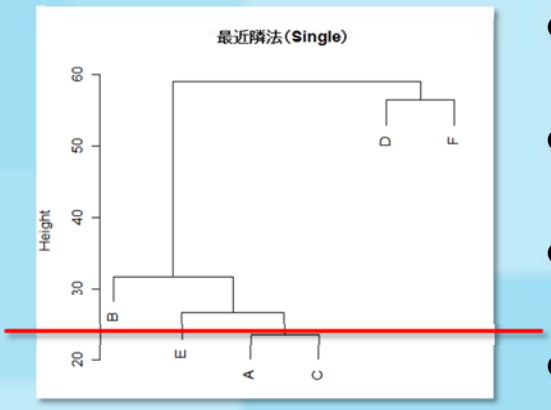
	(((AC)E)B)	(DF)
(((AC)E)B)		
(DF)	59	



デンドログラム
（樹形図）

- データが多いとクラスタリングの手計算は困難
- コンピュータソフトウェアでクラスタリングを行うのが一般的

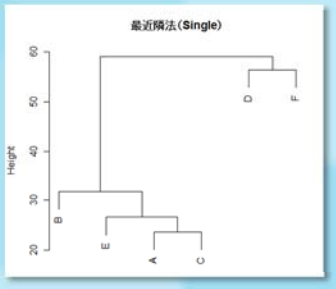
デンドログラム（Dendrogram;樹形図）



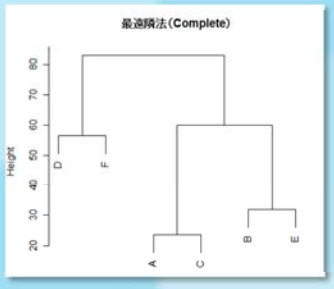
- クラスタリング結果を表示するグラフ
- クラスタが作成される過程を知ることができる
- グラフの下の方で結合しているデータ程、類似しているデータ
- グラフの高さはデータが結合した時の距離

デンドログラム（Dendrogram;樹形図）

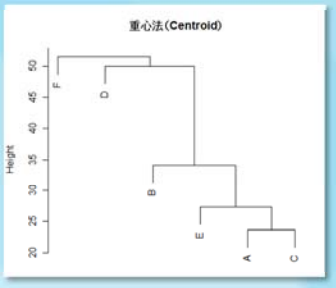
- クラスタリング手法によるデンドログラムの形状の違い
- 同じデータでも手法によりデンドログラムの形が異なる



最近隣法

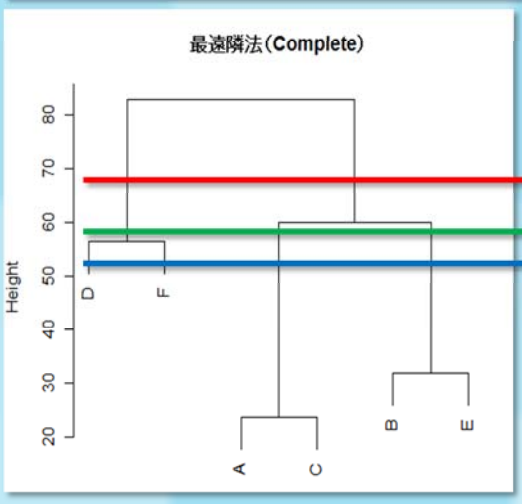


最遠隣法



重心法

デンドログラム（Dendrogram;樹形図）



2個
(DF) (ACBE)

3個
(DF) (AC) (BE)

(D) (F) (AC) (BE)
4個

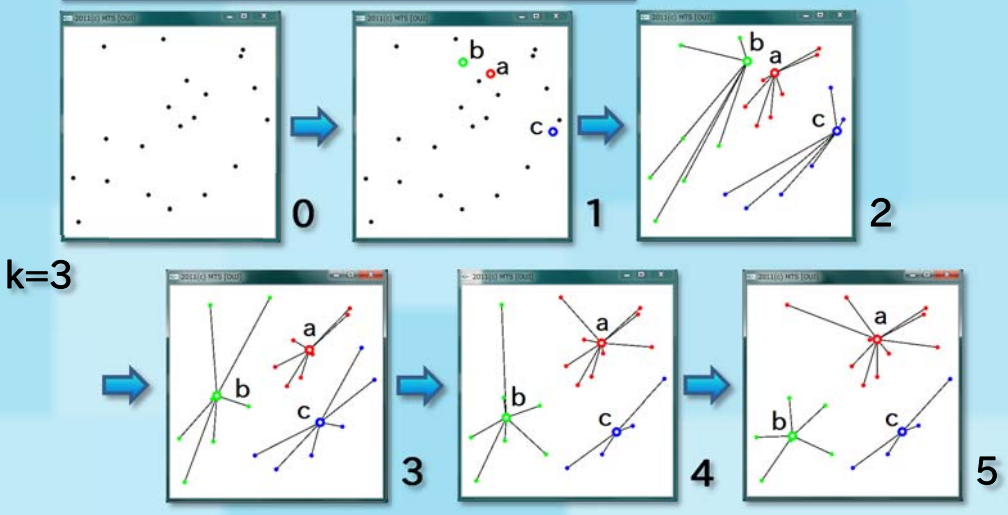
非階層的クラスタリング

非階層的クラスタリング

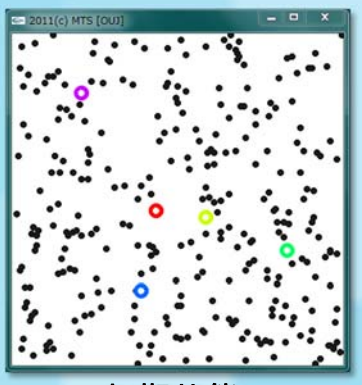
- k 平均法（k-means法）

1. 作成したいクラスタの数を設定（k の値を設定）
2. 各クラスタの重心の初期値を設定
3. 各データを最も近いクラスタに分類
4. 各クラスタの重心を計算
5. ３と４を重心の位置が動かなくなるまで繰り返す

k 平均法（例１）

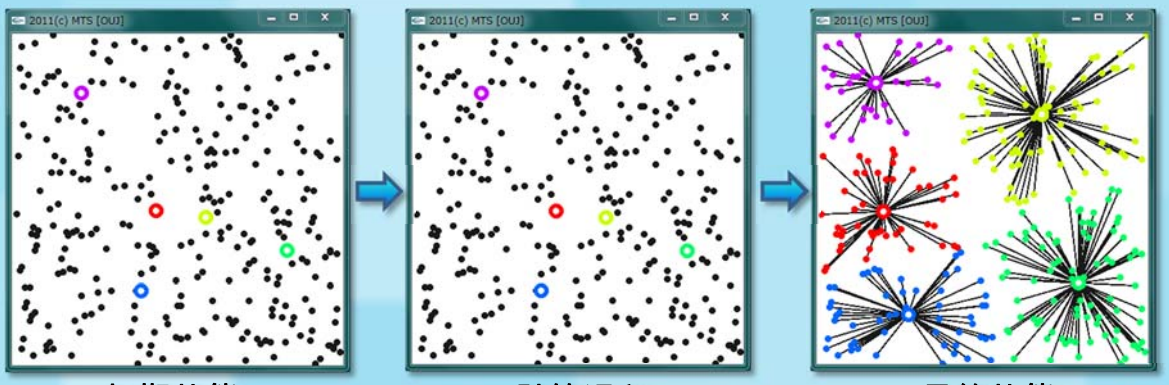


k 平均法（例2）



初期状態
データ数　：500個
クラスタ数：　5個

k 平均法（例2）

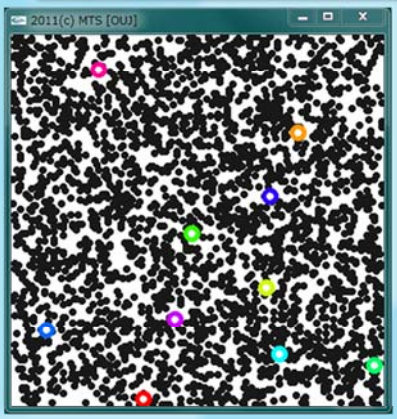


初期状態
データ数　：500個
クラスタ数：　5個

計算過程

最終状態

k 平均法（例３）

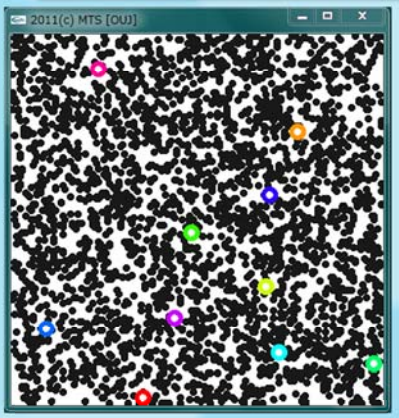


k 平均法の初期値設定
データ数　：2000個
クラスタ数：　10個

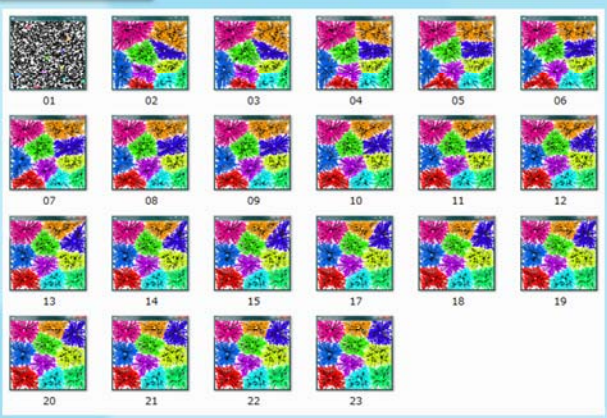
- データ数：　２０００個
- クラスタ数：　　１０個

- 初期の各クラスタの重心位置は乱数によって設定

k 平均法（例３）

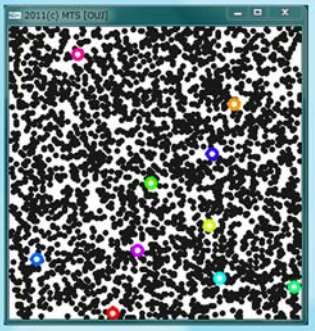


k 平均法の計算過程
データ数　：2000個
クラスタ数：　10個

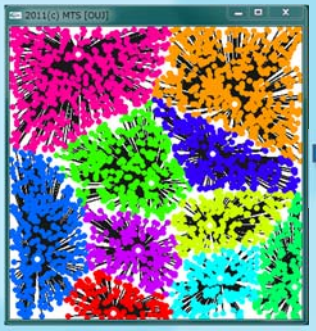


- 各クラスタの重心位置が移動していく

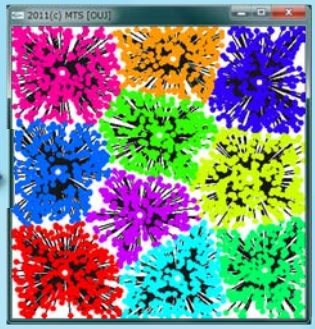
k平均法（例3）



初期値設定
データ数　：2000個
クラスタ数：　10個



ステップ1

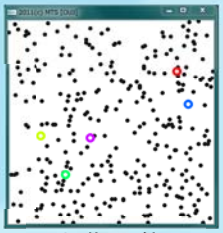


ステップ23（結果）

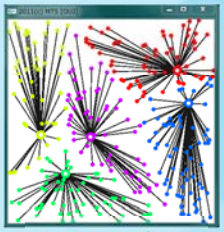
- k平均法では各クラスタのサイズが同じぐらいになる性質がある（データが一様に分布の場合）

k平均法（異なる初期値の場合）

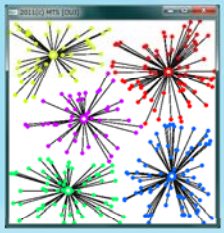
A



初期状態

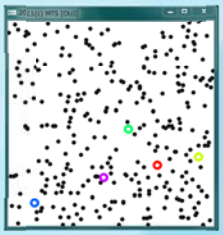


ステップ1

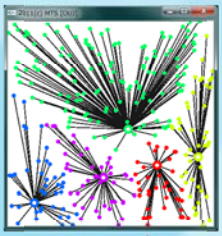


ステップ8

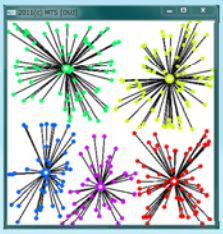
B



初期状態



ステップ1

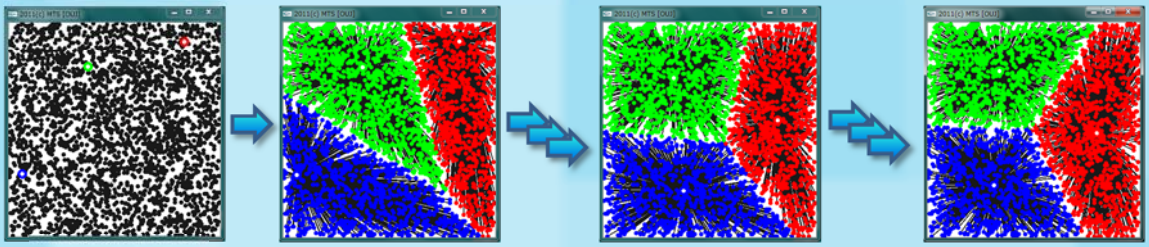


ステップ10

k 平均法（異なる初期値の場合）

- 設定する各クラスタの重心の初期値の違いによって
  クラスタリングの結果が変化する
- 重心の初期値の違いが
  クラスタリングに必要なステップ数にも
  影響をおよぼし
  処理が完了するステップ数も異なってくる

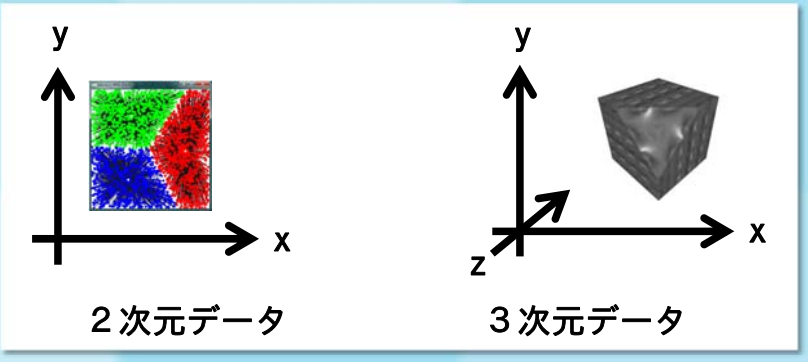
k 平均法（データ数が多い場合）



初期状態　ステップ1　ステップ10　ステップ23

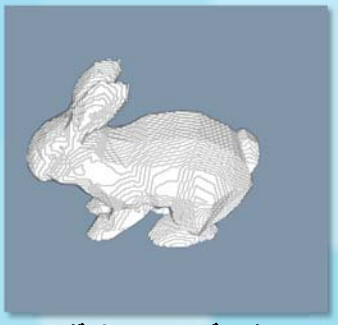
- データ（３０００個）、クラスタ（３個）のｋ平均法の例
- 例では２３回の重心の再計算が必要
- 多くのデータがあると、各クラスタの重心の移動の変化が無くなるまでに
  （重心位置の収束）多くの繰り返し計算が必要

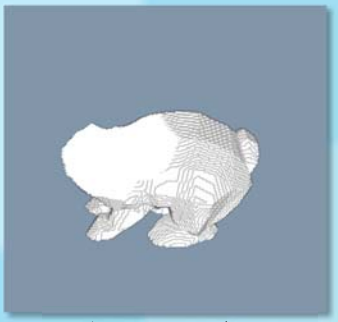
k 平均法（3 次元データの場合）



- k 平均法は、２次元データだけではなく、３次元データ、あるいはそれ以上の多次元データ（n 次元）のクラスタリングも行うことができる

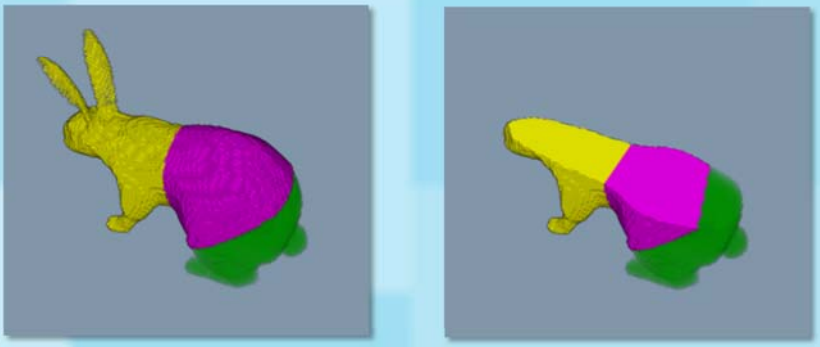
k 平均法（3 次元ボリュームデータの場合）


ボリュームデータ
128×128×128


ボリュームデータ
一部分をカット


ボリュームデータ
一部分をカット

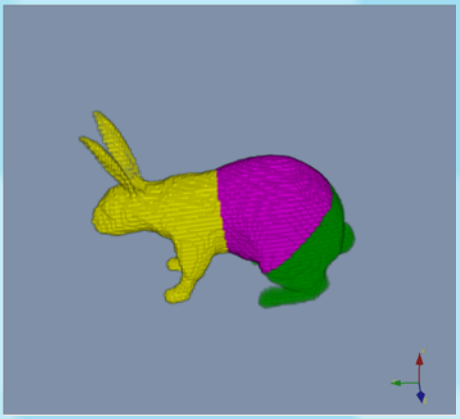
k 平均法（例：３次元ボリュームデータ、３クラスタ）



ボリュームデータ（128×128×128）３クラスタ

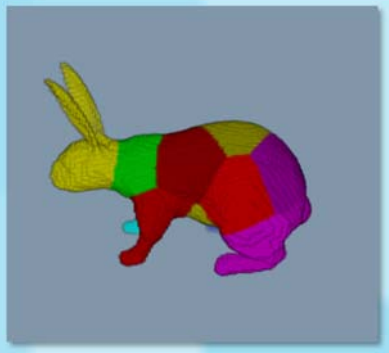
- 物体部分の各ボクセル間の距離を類似度と定義してクラスタリング
- 物体内部に初期重心（３個）をランダムに設定

k 平均法（例：３次元ボリュームデータ、３クラスタ）

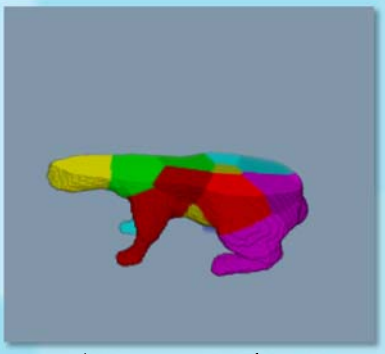


ボリュームデータ （128×128×128）３クラスタ

k 平均法（例：３次元ボリュームデータ、１６クラスタ）

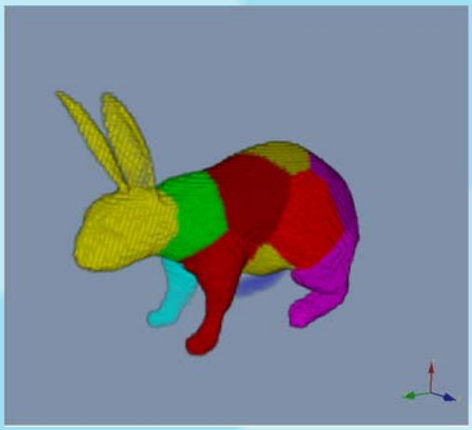


ボリュームデータ
（128×128×128）
１６クラスタ



ボリュームデータ
（128×128×128）
１６クラスタ
データ上部をカットした表示

k 平均法（例：３次元ボリュームデータ、１６クラスタ）



ボリュームデータ（128×128×128）１６クラスタ

まとめ（11回　クラスタリング）

- クラスタリングの概要
  - 階層的クラスタリングと非階層的クラスタリング
  - ハードクラスタリングとソフトクラスタリング
- 階層的クラスタリング
  - 最近隣法、最遠隣法、重心法
  - デンドログラム（樹形図）
- 非階層的クラスタリング
  - k平均法（k-means法）、クラスタ重心