ВВЕДЕНИЕ В МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА МНОГОМЕРНЫХ ОБЪЕКТОВ

ВСЕ, ЧЕГО НЕЛЬЗЯ ВЫРАЗИТЬ В ЦИФРАХ,
НЕ НАУКА, А ПРОСТО МНЕНИЕ

Роберт Хайнлайн

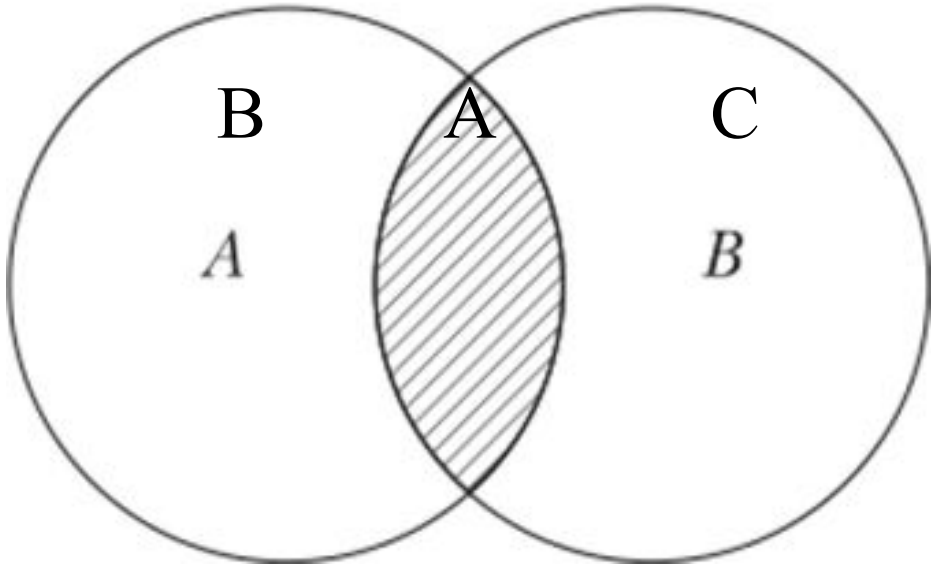
ЗАКОНЫ МАТЕМАТИКИ, ИМЕЮЩИЕ КАКОЕ-ЛИБО
ОТНОШЕНИЕ К РЕАЛЬНОМУ МИРУ, НЕНАДЕЖНЫ;
А НАДЕЖНЫЕ МАТЕМАТИЧЕСКИЕ ЗАКОНЫ
НЕ ИМЕЮТ ОТНОШЕНИЯ К РЕАЛЬНОМУ МИРУ

Альберт Эйнштейн

ГОРАЗДО ЛЕГЧЕ ЧТО-ТО ИЗМЕРИТЬ, ЧЕМ ПОНЯТЬ,
ЧТО ИМЕННО ВЫ ИЗМЕРЯЕТЕ

Дж. Салливен

Индексы сходства



$$I_{CS} = \frac{2a}{(a+b)+(a+c)}$$

$$I_J = \frac{a}{a+b+c}$$

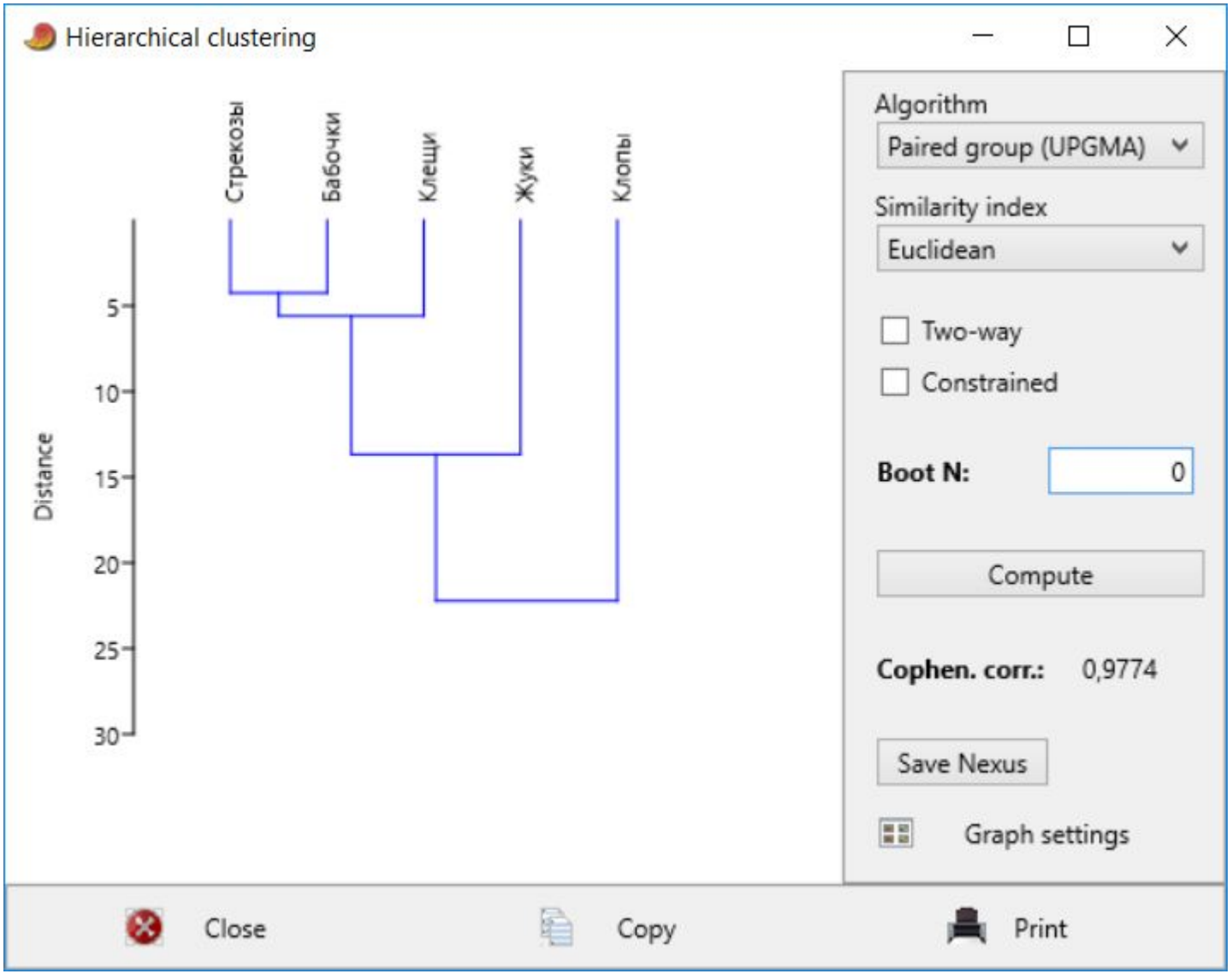
$$I_{OB} = \frac{a}{\sqrt{(a+b)(a+c)}}$$

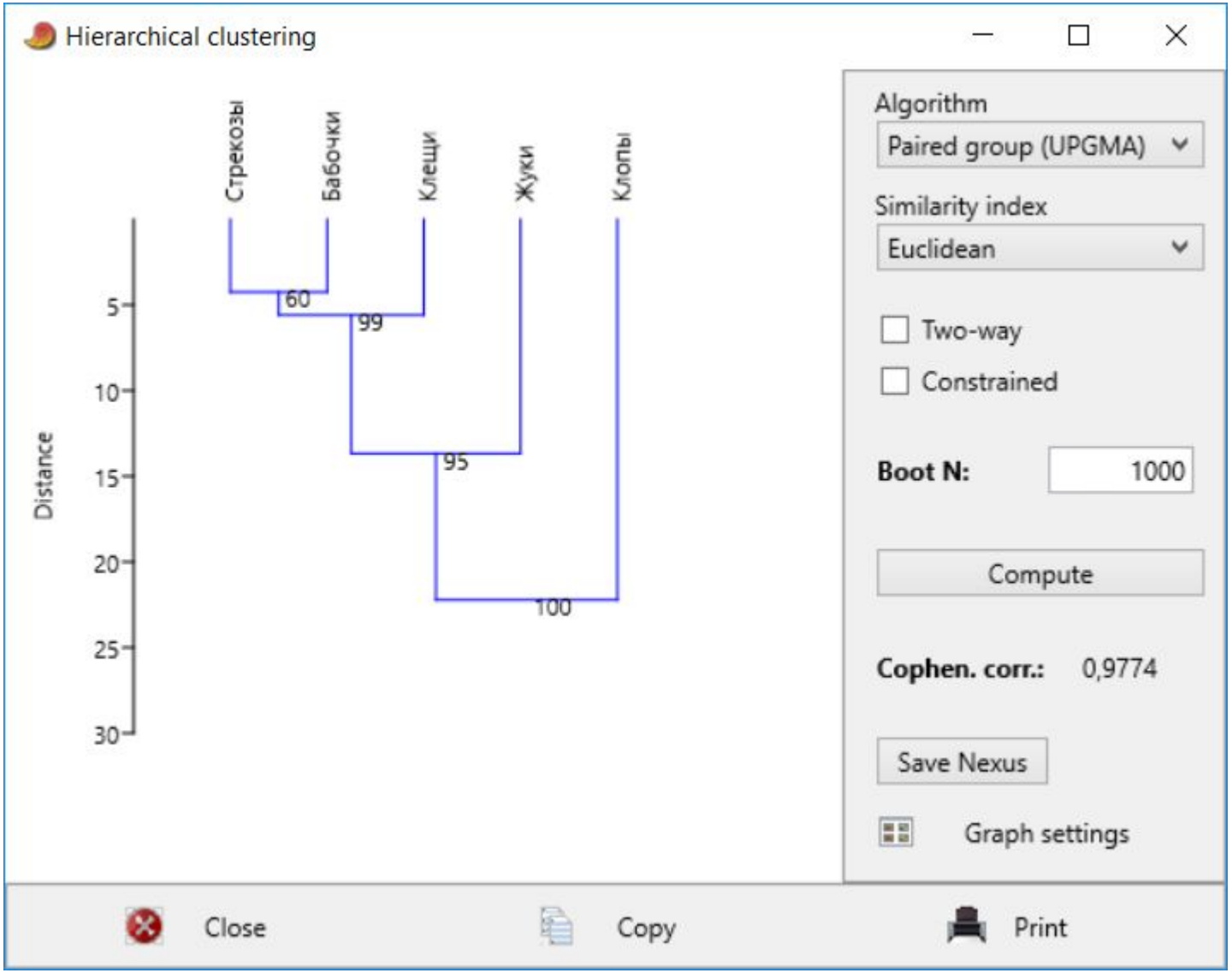
Бабочки	Стрекозы	Жуки	Клопы	Клещи
1	1	0	0	0
0	0	1	1	0
1	1	0	1	0
0	0	0	0	0
1	1	0	0	0
0	1	0	0	0
0	0	1	1	1
1	1	0	0	0
1	1	1	1	0
0	0	0	0	0
1	1	1	1	0
0	1	0	0	0
1	1	0	1	0
1	1	0	0	0
1	1	0	0	0
1	1	0	0	0
0	1	0	0	0
1	1	0	0	0
1	1	0	0	0
1	1	0	0	0
0	1	0	0	0

	Б	С	Ж	Кп	Кщ
Б	1				
С	0,95	1			
Ж	0,49	0,44	1		
Кп	0,63	0,58	0,68	1	
Кщ	0,10	0,09	0,19	0,19	1

Кластерный анализ

Бабочки	Жуки	Клопы	Стрекозы	Клещи
0,927	2,734	4,279	0,298	0,078
0,961	3,514	3,796	1,156	0,136
0,615	1,004	7,607	1,379	0,123
0,969	2,486	1,915	0,045	0,192
1,700	0,961	5,490	0,545	0,200
0,971	4,065	5,653	2,045	0,138
0,331	0,578	7,758	0,420	0,138
1,208	0,961	1,536	0,361	0,065
1,117	4,622	6,844	0,782	0,195
0,459	3,620	3,702	1,816	0,058
1,004	3,850	7,561	1,320	0,087
1,278	3,636	2,306	1,785	0,083
1,047	2,241	0,126	0,720	0,118
0,525	0,413	5,258	2,354	0,017
0,332	5,102	6,493	0,975	0,092
1,237	2,276	1,400	0,254	0,003
1,936	0,077	8,199	1,910	0,163
0,311	0,390	6,915	2,129	0,143
0,397	0,049	3,922	2,325	0,194
0,275	1,120	8,055	2,011	0,041
1,512	0,038	8,319	1,876	0,080


Hierarchical clustering
Стрекозы
Бабочки
Клещи
Жуки
Клопы
Distance
5
10
15
20
25
30
Algorithm
Paired group (UPGMA)
Similarity index
Euclidean
Two-way
Constrained
Boot N:
0
Compute
Cophen. corr.: 0,9774
Save Nexus
Graph settings
Close
Copy
Print


Hierarchical clustering
Стрекозы
Бабочки
Клещи
Жуки
Клопы
60
99
95
100
Distance
5
10
15
20
25
30
Algorithm
Paired group (UPGMA)
Similarity index
Euclidean
Two-way
Constrained
Boot N:
1000
Compute
Cophen. corr.:
0,9774
Save Nexus
Graph settings
Close
Copy
Print

ИНДЕКСЫ СХОДСТВА ДЛЯ КОЛИЧЕСТВЕННЫХ ДАННЫХ

- БРЕЯ-КЁРТИСА (BRAY-CURTIS):

$$S_{B-C} = 1 - \frac{\sum_R |n_{1i} - n_{2i}|}{\sum_R (n_{1i} + n_{2i})} \qquad \left(= \sum_R \min(p_{1i}; p_{2i})\right)$$

- ПИАНКИ (PIANKA)(ЧУВСТВИТЕЛЕН К РАЗЛИЧИЯМ **В ДОМИНАНТАХ**)

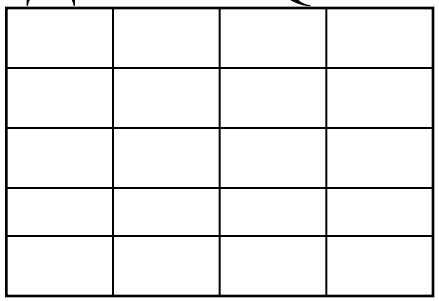
$$S_{PI} = \frac{\sum n_{1i} \cdot n_{2i}}{\sqrt{\sum n_{1i}^2 \cdot \sum n_{2i}^2}}$$

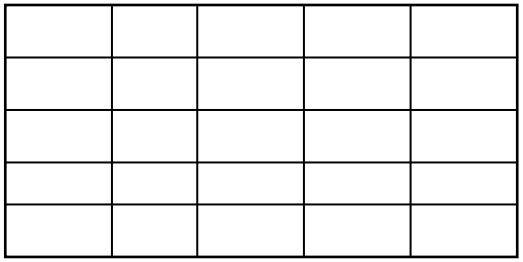
- ЭВКЛИДОВО **РАССТОЯНИЕ:**

$$D_{EU} = \sqrt{\sum (n_{1i} - n_{2i})^2}$$

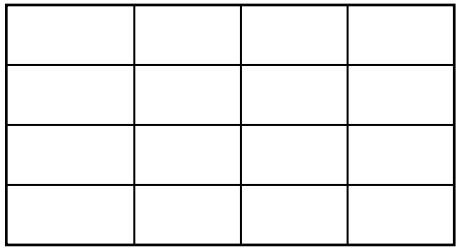
АНАЛИЗ СХОДСТВА

R ВИДОВ × Q ПРОБ

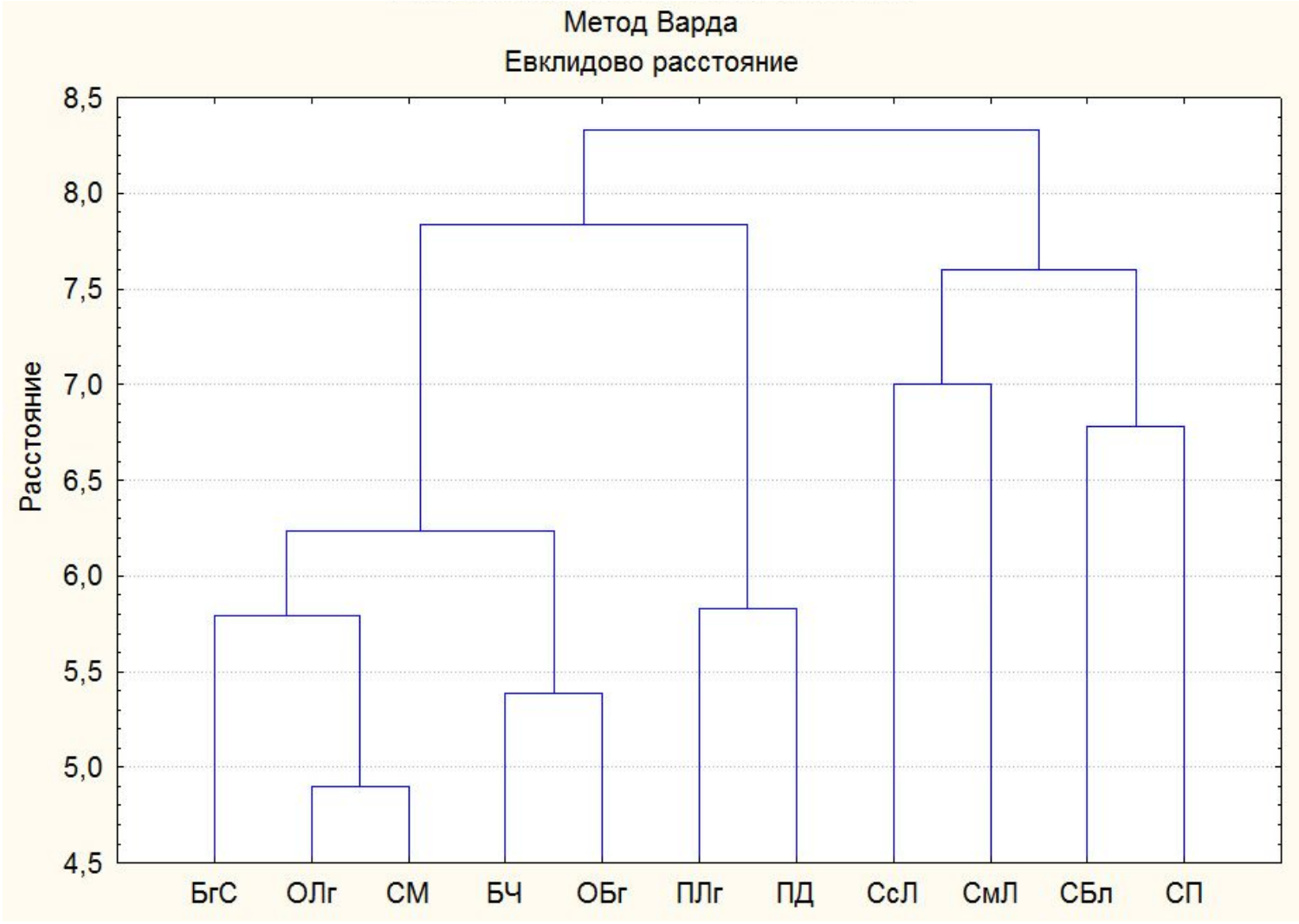




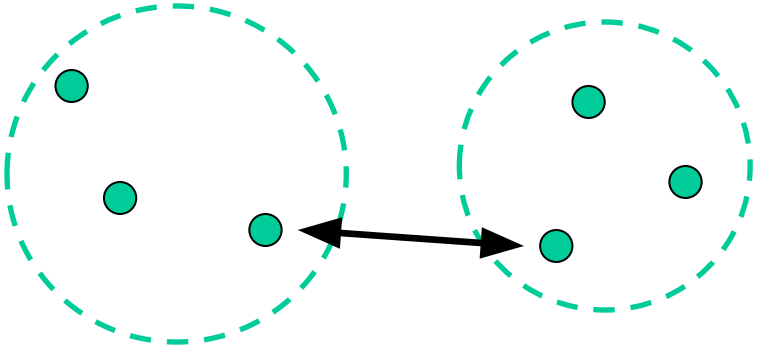
R × R ВИДОВ



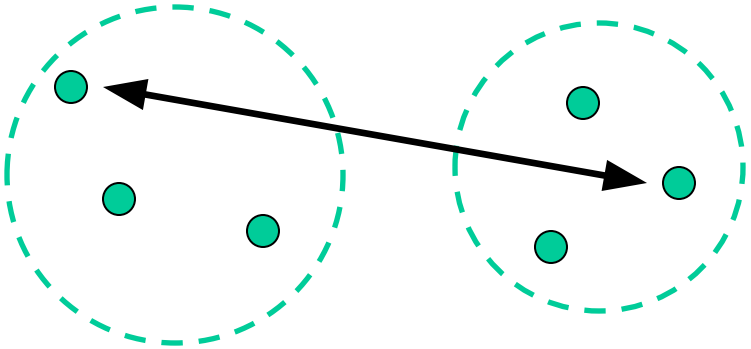
Q × Q ПРОБ


Метод Варда
Евклидово расстояние
Расстояние
8,5
8,0
7,5
7,0
6,5
6,0
5,5
5,0
4,5
БгС
ОЛг
СМ
БЧ
ОБг
ПЛг
ПД
СсЛ
СмЛ
СБл
СП

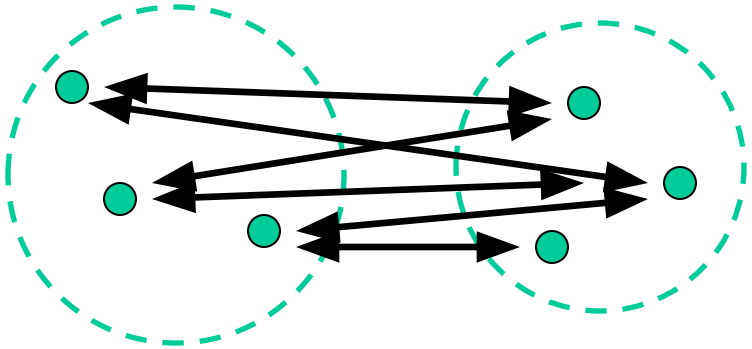
СПОСОБЫ ОБЪЕДИНЕНИЯ ГРУПП ОБЪЕКТОВ



МЕТОД
БЛИЖАЙШЕГО СОСЕДА
(SINGLE LINKAGE)



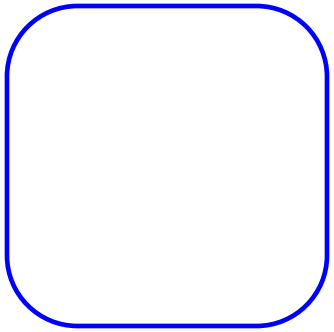
МЕТОД
ДАЛЬНЕГО СОСЕДА
(COMPLETE LINKAGE)

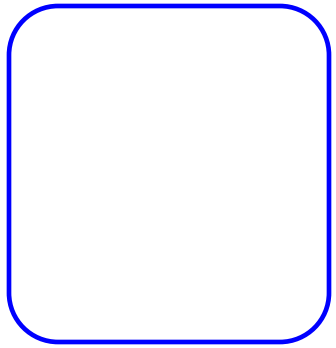


МЕТОД **СРЕДНЕГО ПРИСОЕДИНЕНИЯ**
(GROUP AVERAGE)

МЕТОД БЛИЖАЙШЕГО СОСЕДА

МЕТОД ДАЛЬНЕГО СОСЕДА

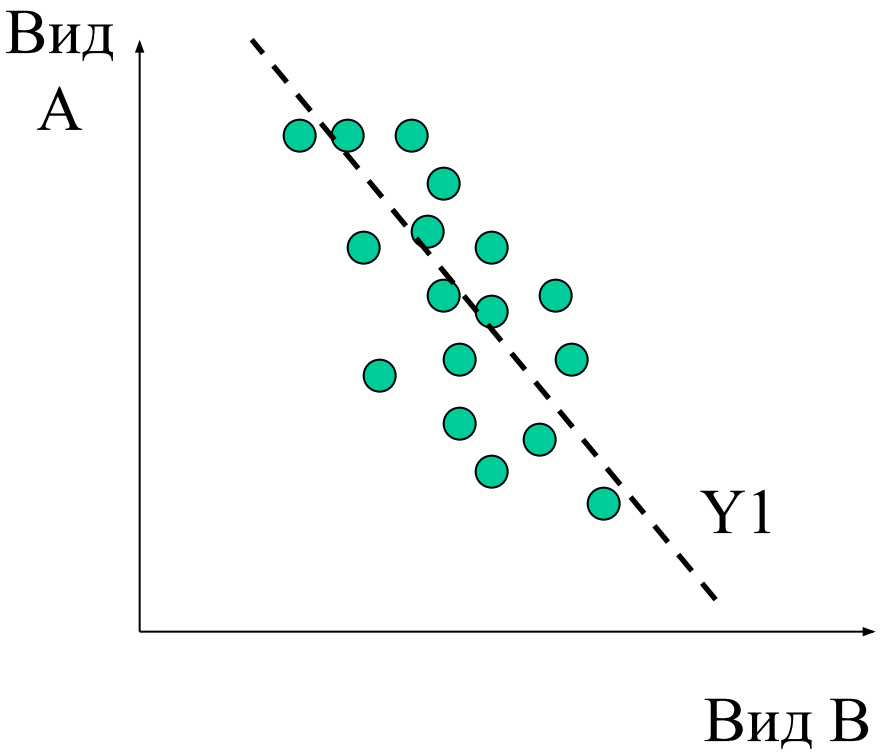




Внимание!
Далее будет по-настоящему многомерная статистика

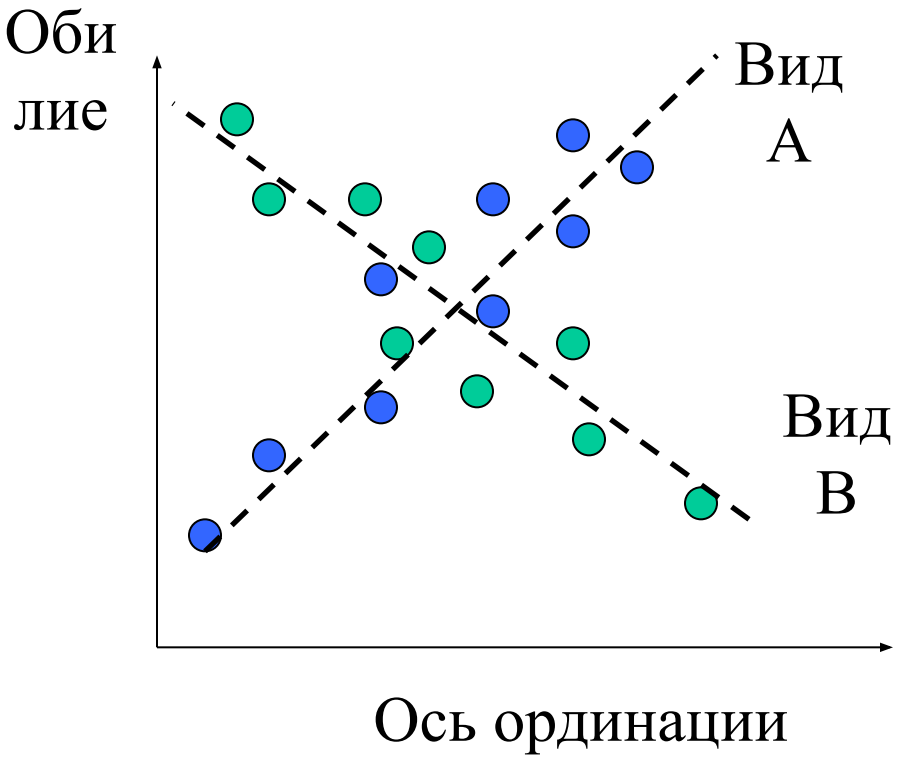
МЕТОД ГЛАВНЫХ КОМПОНЕНТ (PRINCIPAL COMPONENT ANALYSIS, PCA)

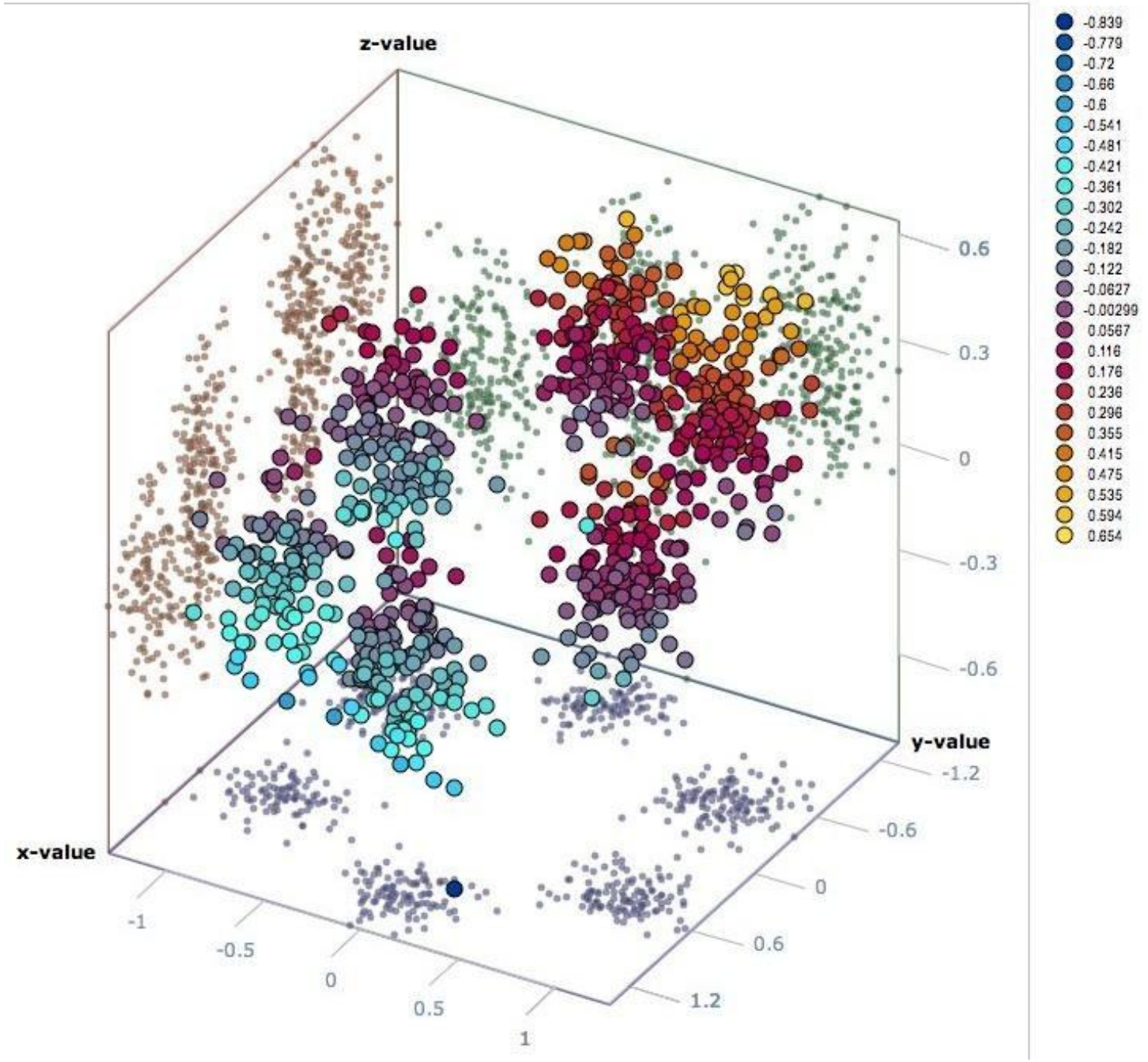
ПРЕДПОЛАГАЕТСЯ, ЧТО ПРИЗНАКИ СВЯЗАНЫ МЕЖДУ СОБОЙ ЛИНЕЙНО



ТОГДА ОСИ - ЛИНЕЙНЫЕ КОМБИНАЦИИ ПРИЗНАКОВ

$$Y1 = a_1 X_A + b_1 X_B$$



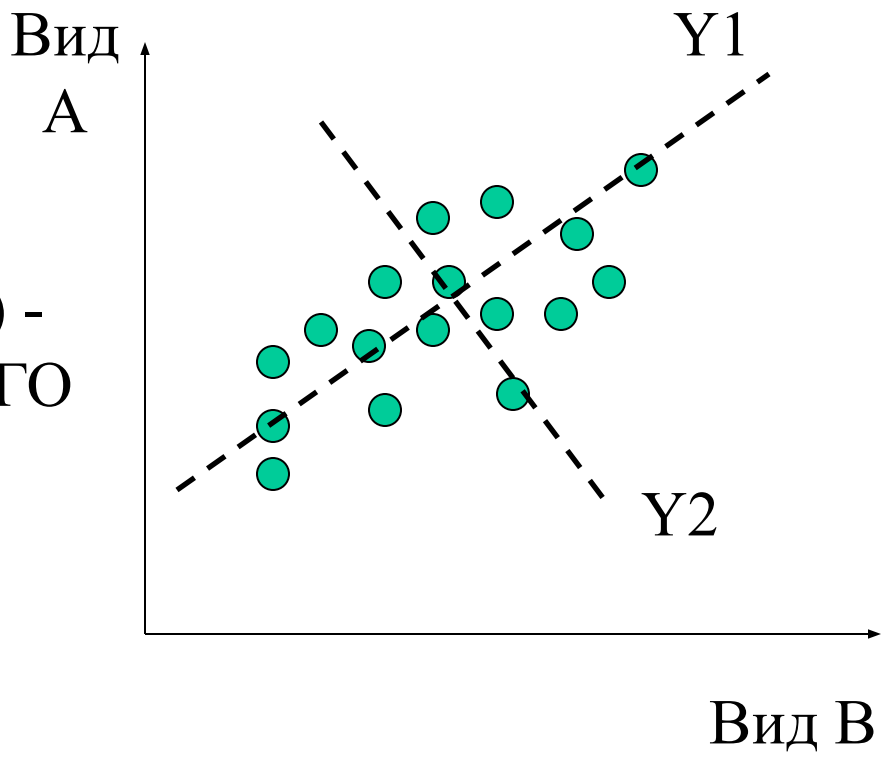

z-value
0.6
0.3
0
-0.3
-0.6
y-value
-1.2
-0.6
0
0.6
1.2
x-value
-1
-0.5
0
0.5
1
-0.839
-0.779
-0.72
-0.66
-0.6
-0.541
-0.481
-0.421
-0.361
-0.302
-0.242
-0.182
-0.122
-0.0627
-0.00299
0.0567
0.116
0.176
0.236
0.296
0.355
0.415
0.475
0.535
0.594
0.654

МЕТОД ГЛАВНЫХ КОМПОНЕНТ (PRINCIPAL COMPONENT ANALYSIS)

ОСИ - ЛИНЕЙНЫЕ КОМБИНАЦИИ ПРИЗНАКОВ

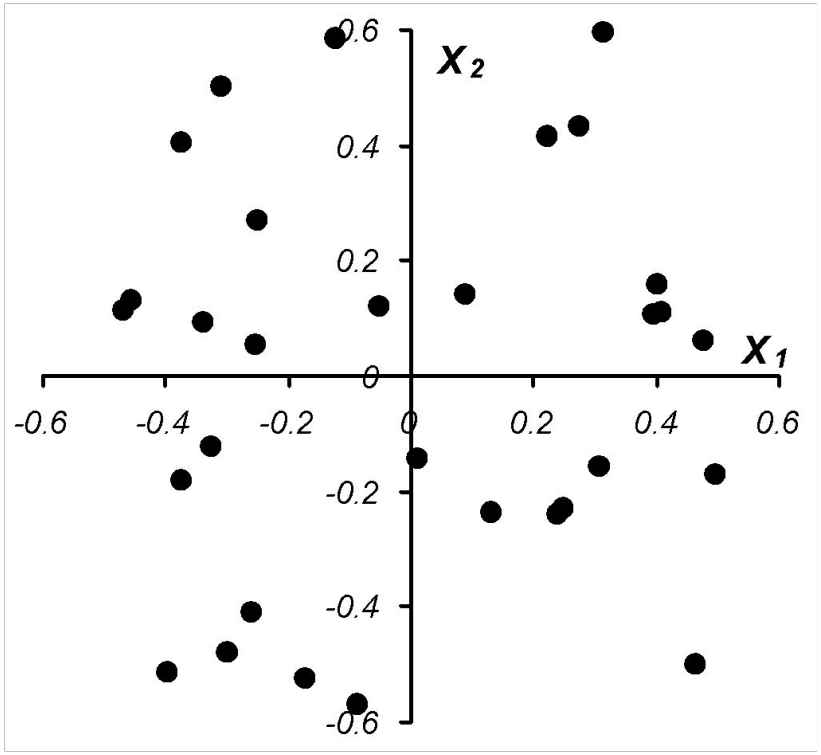
ПЕРВАЯ ОСЬ (КОМПОНЕНТА) - НАПРАВЛЕНИЕ НАИБОЛЬШЕГО РАЗБРОСА ТОЧЕК

$$Y1 = a_1 X_A + b_1 X_B$$

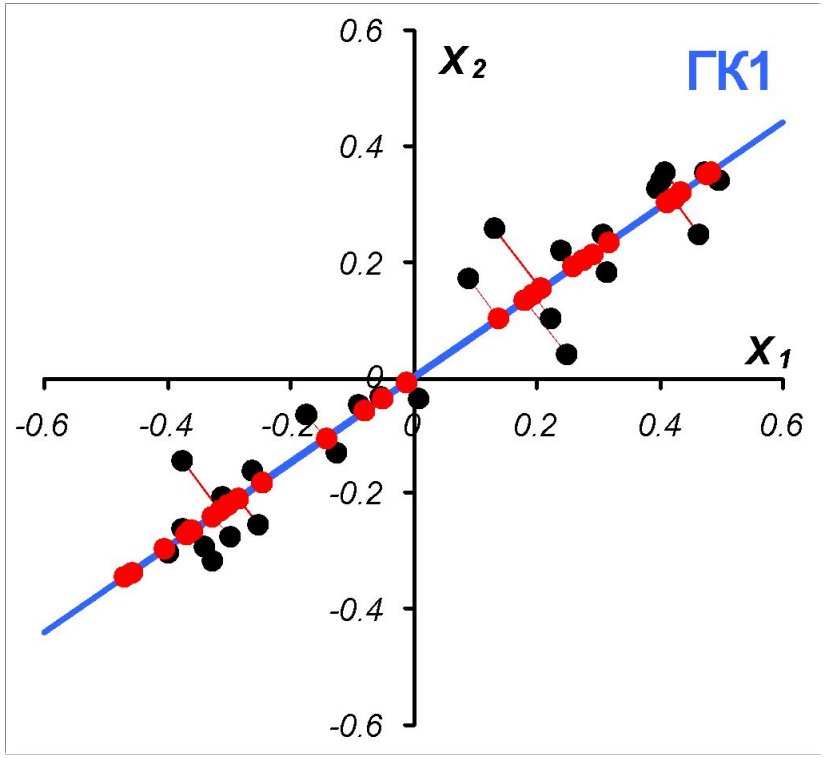


МЕТОД ГЛАВНЫХ КОМПОНЕНТ (PRINCIPAL COMPONENT ANALYSIS)

Данные без структуры

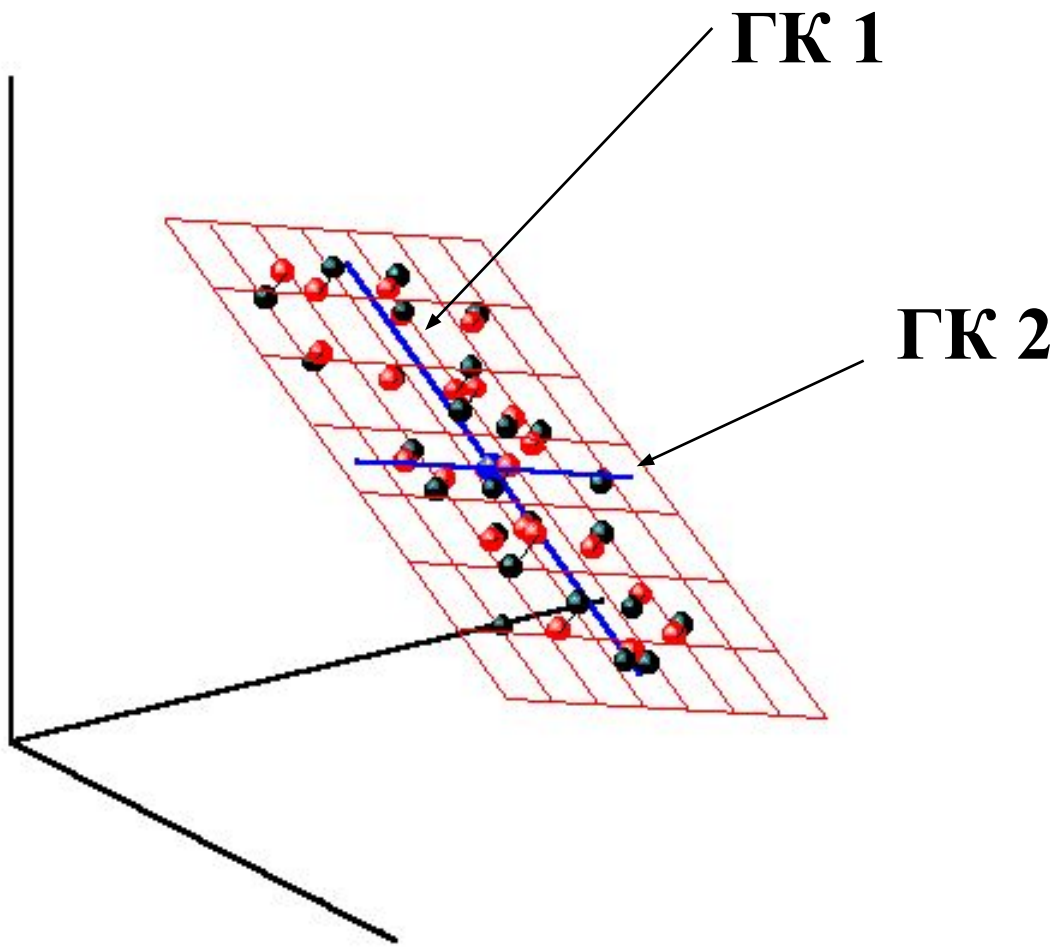


Данные со скрытой структурой

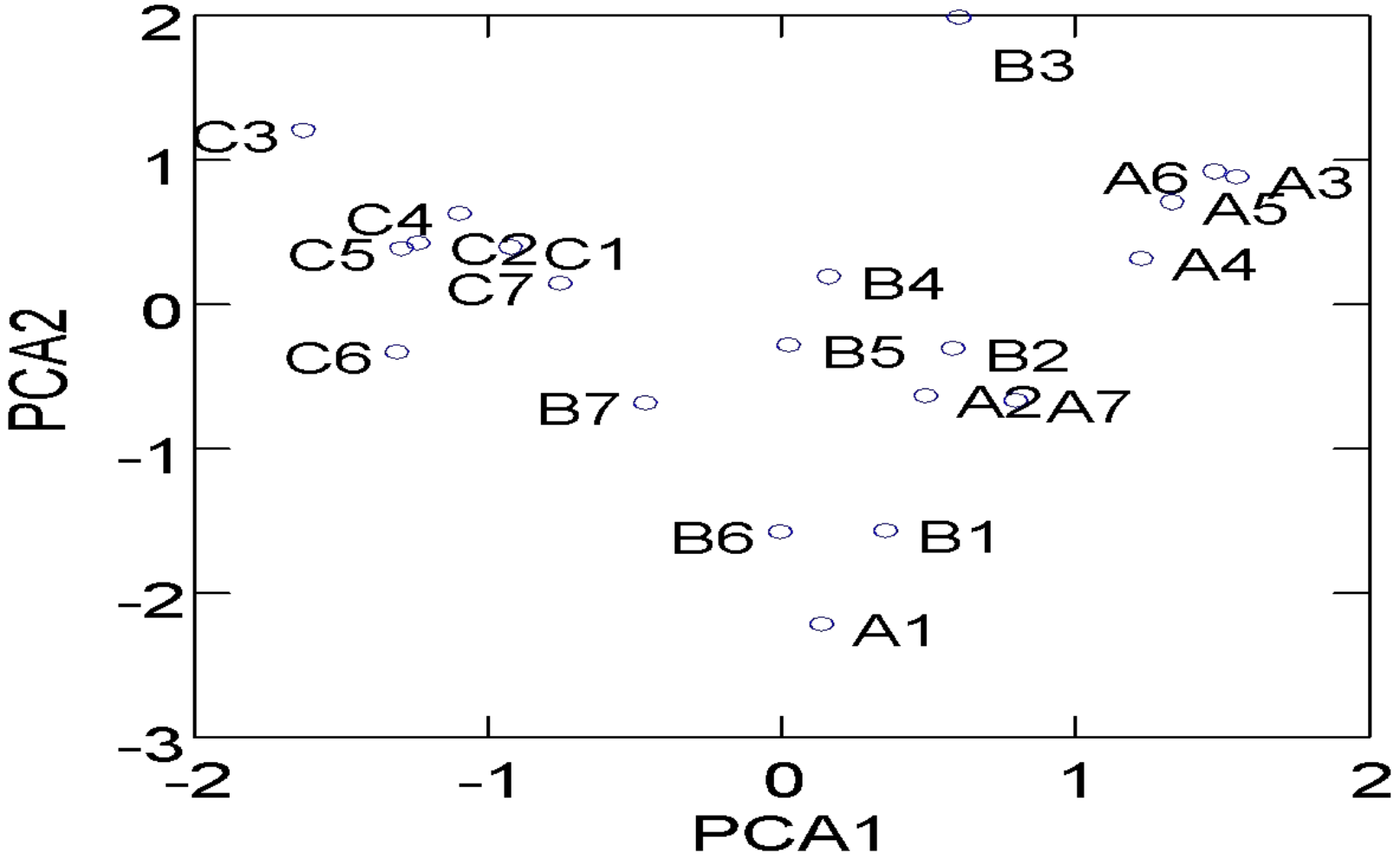


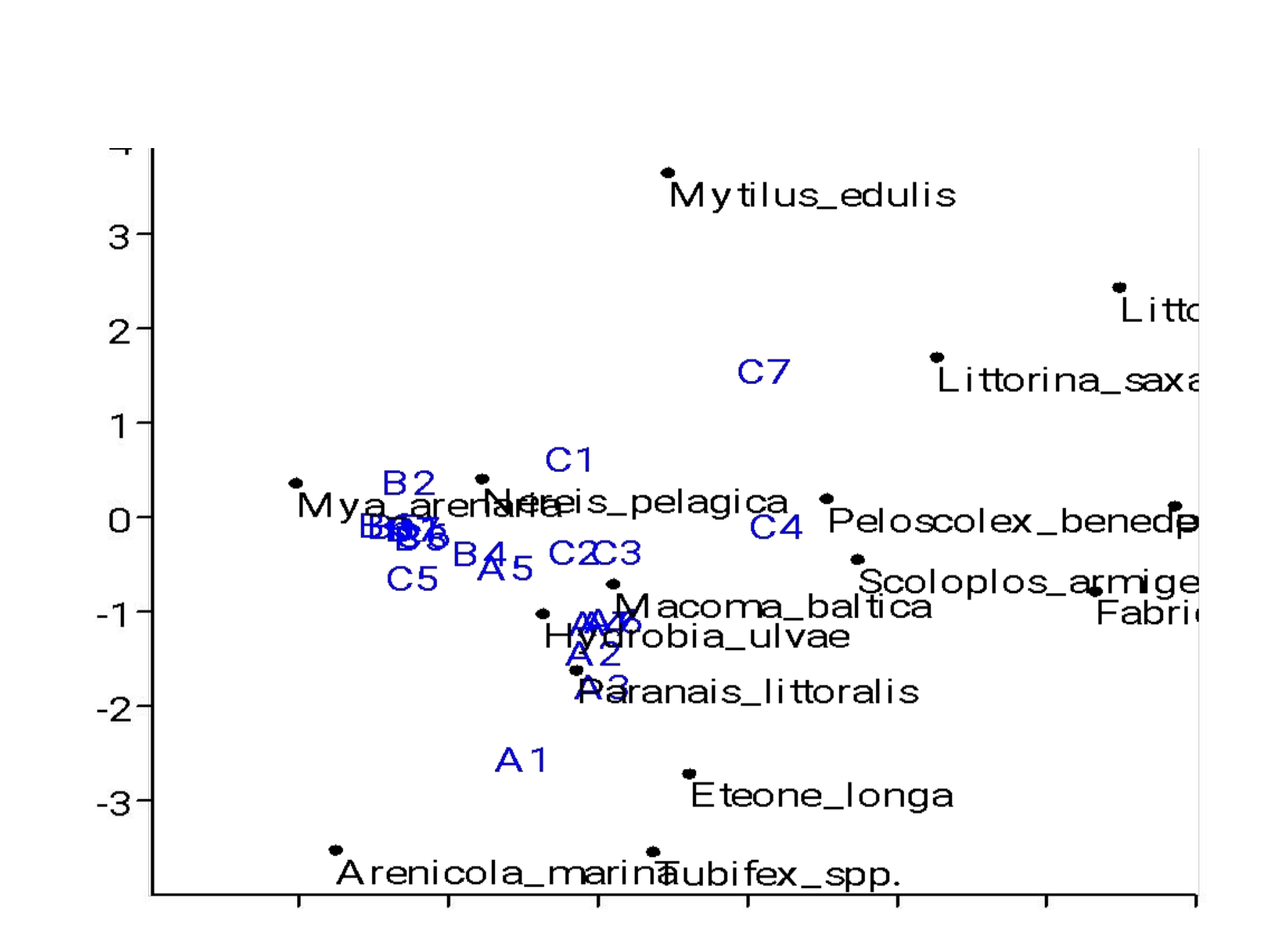
МЕТОД ГЛАВНЫХ КОМПОНЕНТ (PRINCIPAL COMPONENT ANALYSIS)

ВТОРАЯ ОСЬ - НАПРАВЛЕНИЕ НАИБОЛЬШЕГО РАЗБРОСА ТОЧЕК, ПЕРПЕНДИКУЛЯРНОЕ ПЕРВОЙ



ПРИМЕР:
ОРДИНАЦИЯ СТАНЦИЙ ПО ФАКТОРАМ СРЕДЫ




Mytilus_edulis
Litto
C7
Littorina_saxa
C1
B2
Nereis_pelagica
Mya_arenaria
C4
Peloscolex_benedp
B4
C2
C3
A5
C5
Scoloplos_armige
Macoma_baltica
Fabri
Hydrobia_ulvae
A2
A3
Paranais_littoralis
A1
Eteone_longa
Arenicola_marina
Tubifex_spp.
3
2
1
0
-1
-2
-3

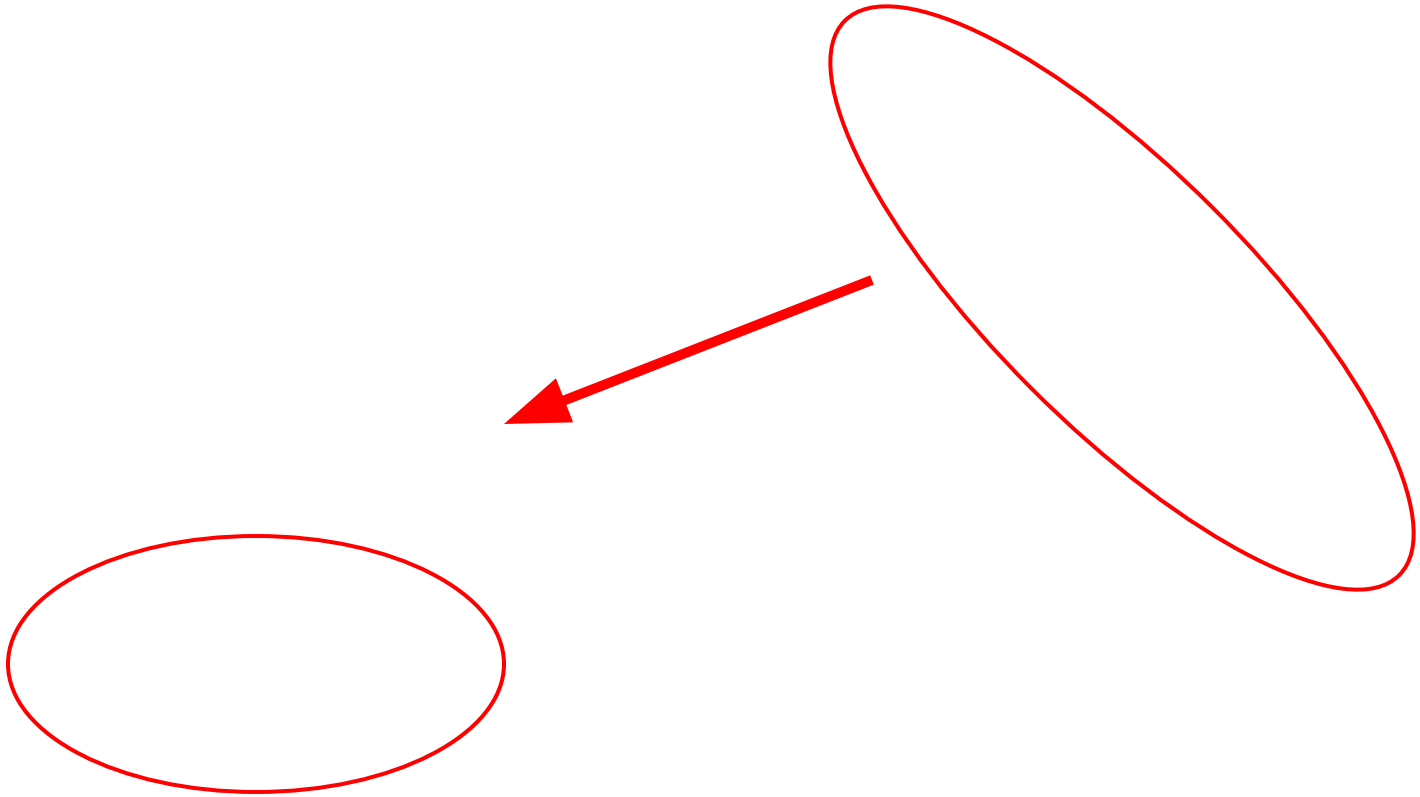
МНОГОМЕРНОЕ ШКАЛИРОВАНИЕ (MULTIDIMENSIONAL SCALING)

ЗАДАЧИ:

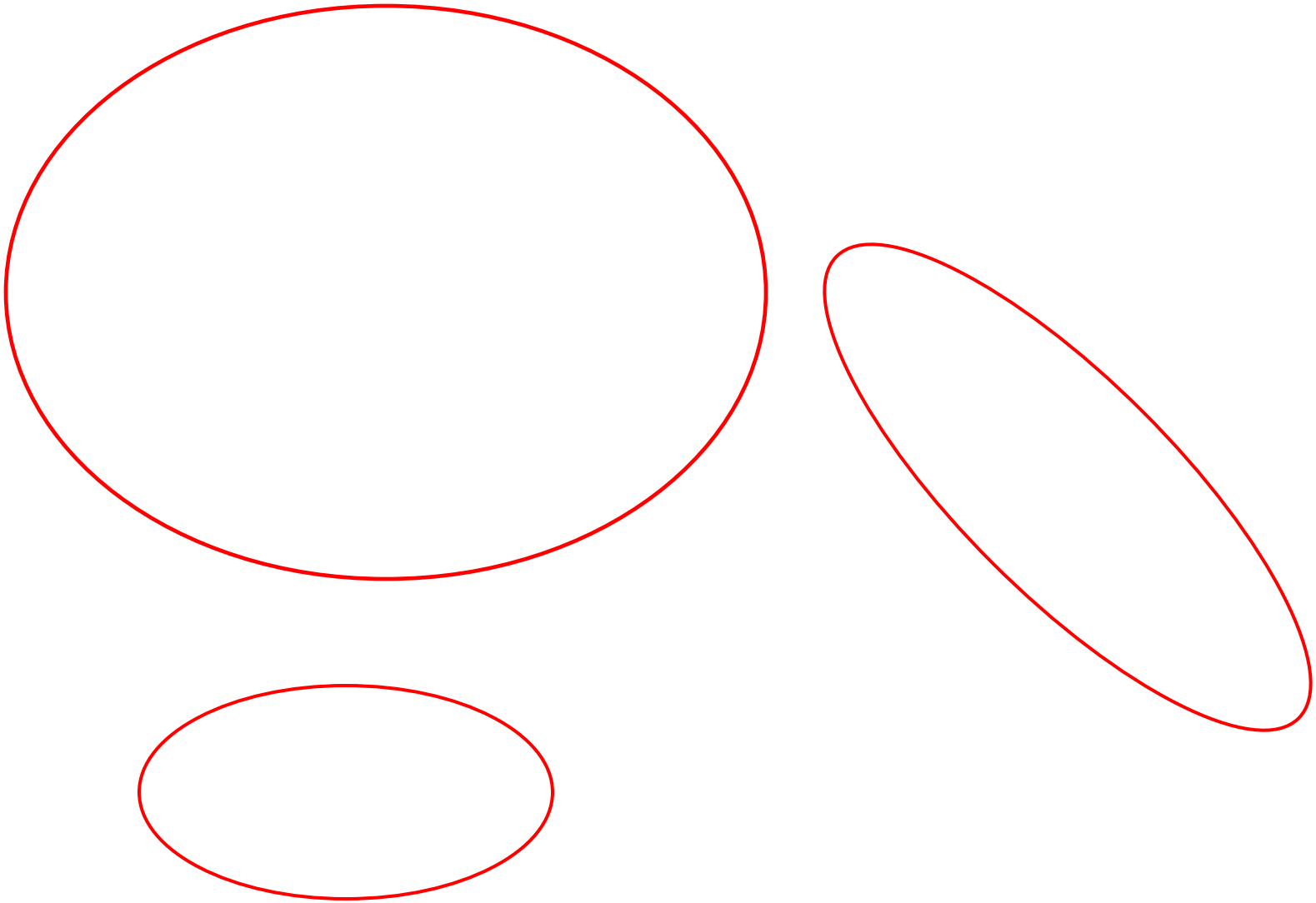
- ВИЗУАЛИЗАЦИЯ ДАННЫХ О СХОДСТВЕ
- УМЕНЬШЕНИЕ РАЗМЕРНОСТИ

РАСПОЛАГАЕТ ОБЪЕКТЫ ТАК, ЧТОБЫ РАССТОЯНИЯ МЕЖДУ НИМИ СООТВЕТСТВОВАЛИ ВЕЛИЧИНАМ НЕСХОДСТВА

ПРИМЕР МНОГОМЕРНОГО ШКАЛИРОВАНИЯ: ОРДИНАЦИЯ СТАНЦИЙ ПО ОБИЛИЮ ВИДОВ

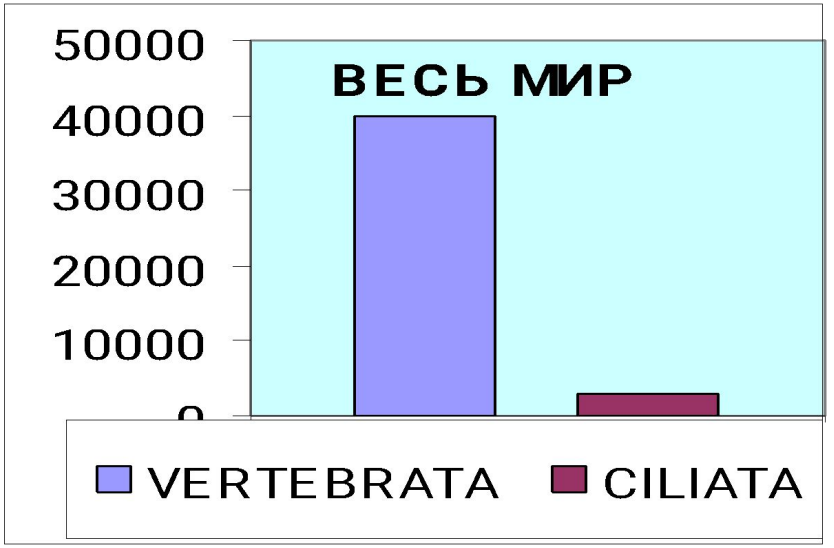


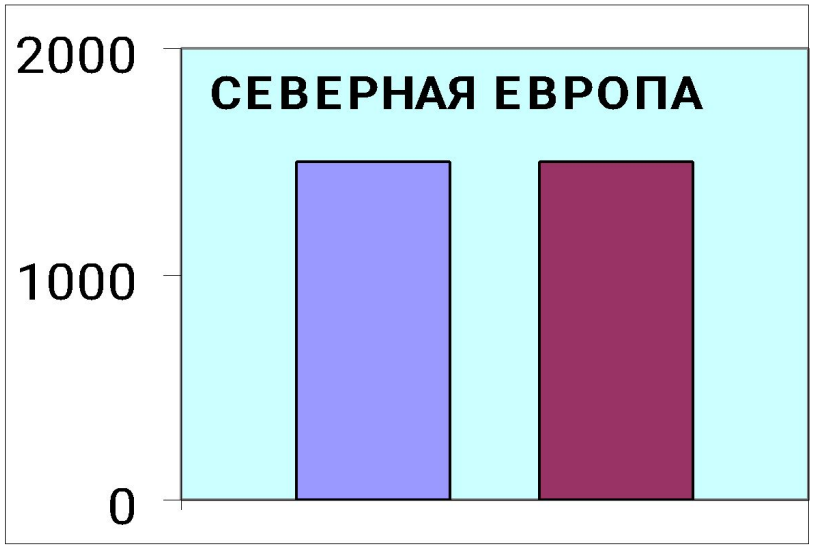
ОРДИНАЦИЯ СТАНЦИЙ ПО ОБИЛИЮ ВИДОВ (МНОГОМЕРНОЕ ШКАЛИРОВАНИЕ)

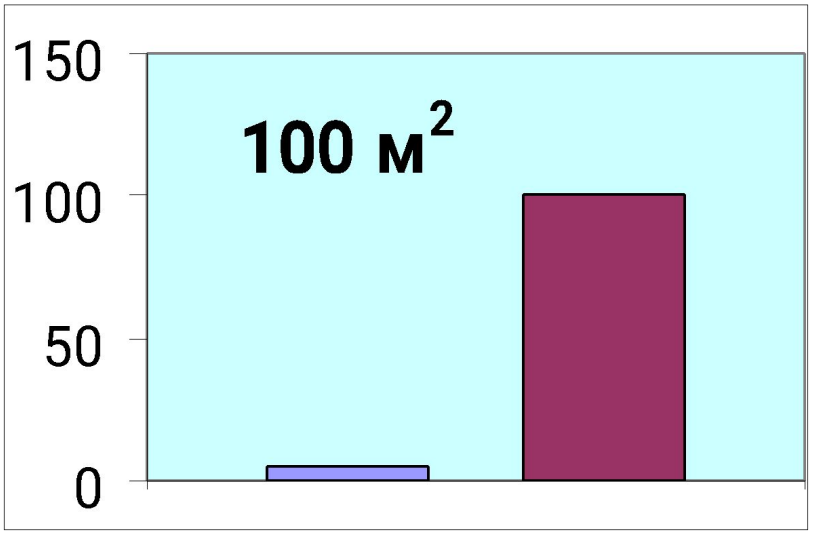


ВИДОВОЕ БОГАТСТВО: КАК ЕГО ОЦЕНИТЬ?

- **КАКАЯ ГРУППА БОГАЧЕ ВИДАМИ: ИНФУЗОРИИ ИЛИ ПОЗВОНОЧНЫЕ?**







ВИДОВОЕ БОГАТСТВО ЗАВИСИТ ОТ МАСШТАБА

ЕГО НУЖНО НОРМИРОВАТЬ. КАК?

- НА ПЛОЩАДЬ (НА $м^2$? НА ГЕКТАР?)
- НА РАЗМЕР ОСОБИ?
- НА ЧИСЛО ОСОБЕЙ?

НАКОПЛЕНИЕ ВИДОВ («КРИВАЯ СБОРЩИКА»)

Число видов
S

ЧИСЛО ВИДОВ:

- на пробу: S_{sample}
- на *n* особей: ES(*n*)

Объем выборки, N
(пробы или особи)

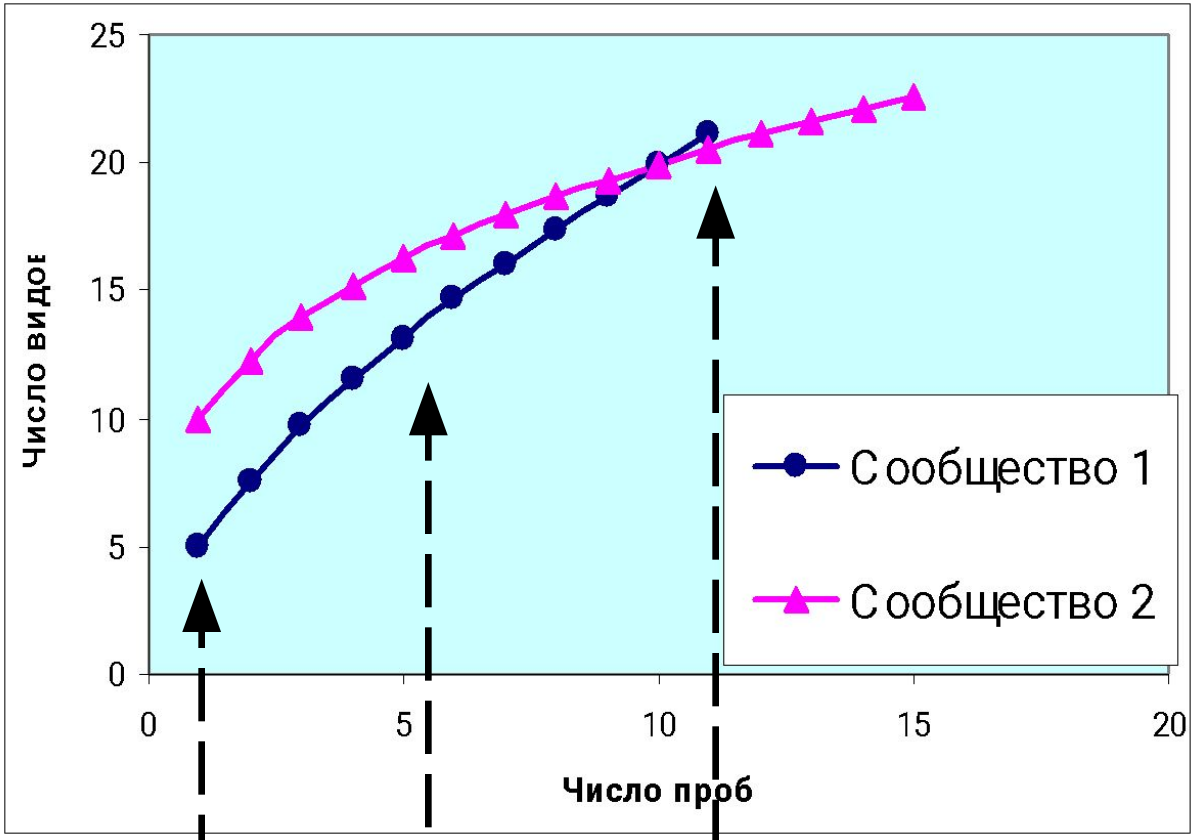
ВИДОВОЕ БОГАТСТВО

$$d_{MENH} = S / \sqrt{N}$$

$$d_{MARG} = \frac{S-1}{\log N}$$

$$d = \frac{\log S}{\log N}$$

СРАВНЕНИЕ КРИВЫХ НАКОПЛЕНИЯ ВИДОВ



РАСЧЕТ ПО СТЕПЕННОЙ АППРОКСИМАЦИИ:

$$S = a N^b$$

$$\log S = \log a + b \log N$$

$a = \alpha$;

$b = \beta$ (угол наклона)

$\alpha 2 > \alpha 1$

$\beta 1 > \beta 2$

$\gamma 1 = \gamma 2$ (при сопоставимых объемах выборки)

ОЦЕНКИ «ПОЛНОГО» ЧИСЛА ВИДОВ ПО ВЫБОРКЕ

Пусть взято N проб, вид найден в n проб.

Встречаемость такого вида (вероятность найти его в пробе): n/N

Вероятность НЕ найти его в пробе: $1-n/N$

Вероятность пропустить такой вид (не найти ни в одной из проб): $(1-n/N)^N$

Метод Chao (оценка полного числа видов с учетом «пропущенных»):

$$S_{ПОЛН} = S + \frac{S_1(S_1 - 1)}{2(S_2 + 1)}$$