Кластер-анализ: средство анализа и интерпретации данных

Б.Г. Миркин, д.т.н.
Профессор

Отделение прикладной математики и информатики НИУ ВШЭ, Москва

Department of Computer Science, Birkbeck University of London, UK
(поддержано Программой фундаментальных исследований НИУ ВШЭ через грант «Учитель-ученики» 2011-2012, НУГ «Методы визуализации и анализа текстов» 2013, и Научную лабораторию ЛАВР (Москва) 2010-2013)

Кластер – это совокупность элементов, которые являются однородными или похожими в данной системе признаков.

Цели кластер-анализа:

(а) структуризация (представление общей структуры данных)

(б) описание кластеров в терминах тех или иных признаков

(в) установление взаимосвязи между различными аспектами явлений

(г) формирование обобщающих утверждений о свойствах данных и явлений

(д) визуализация данных в процессах принятия решений

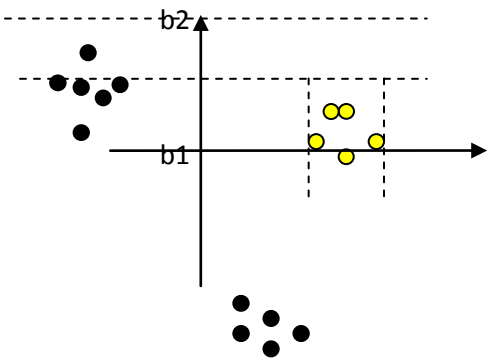
Технология:

Данные объект-признак (фрагмент)

Town	Pop	PS	D	Ho	Ba	Sst	Pet	DIY	Swi	Po	CAB	FM
Mullion	2040	1	0	0	2	0	1	0	0	1	0	0
So Brent	2087	1	1	0	1	1	0	0	0	1	0	0
St Just	2092	1	0	0	2	1	1	0	0	1	0	0
St Colum	2119	1	0	0	2	1	1	0	0	1	1	0
Nanpean	2230	2	1	0	0	0	0	0	0	2	0	0
Gunnisla	2236	2	1	0	1	0	1	0	0	3	0	0
Mevagiss	2272	1	1	0	1	0	0	0	0	1	0	0
Ipplepen	2275	1	1	0	0	0	1	0	0	1	0	0
Alston	2362	1	0	0	1	1	0	0	0	1	0	0
Lostwith	2452	2	1	0	2	0	1	0	0	1	0	1
StColumb	2458	1	0	0	0	1	3	0	0	2	0	0
Padstow	2460	1	0	0	3	0	0	0	0	1	1	0
Perranpo	2611	1	1	0	1	1	2	0	0	2	0	0
Kingsbri	5291	1	1	0	5	3	1	0	1	1	1	0
Wadebrid	5676	2	0	0	4	4	1	0	0	2	1	1
Dartmout	6466	4	1	0	8	4	4	0	1	3	1	0
Launcest	6929	2	1	1	7	2	1	0	1	4	0	1

Кластер – скопление объектов как точек многомерного пространства

Есть кластеры



Нет кластеров

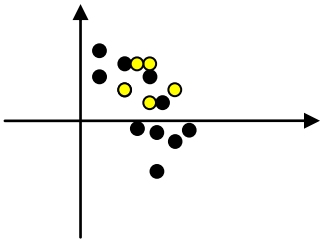
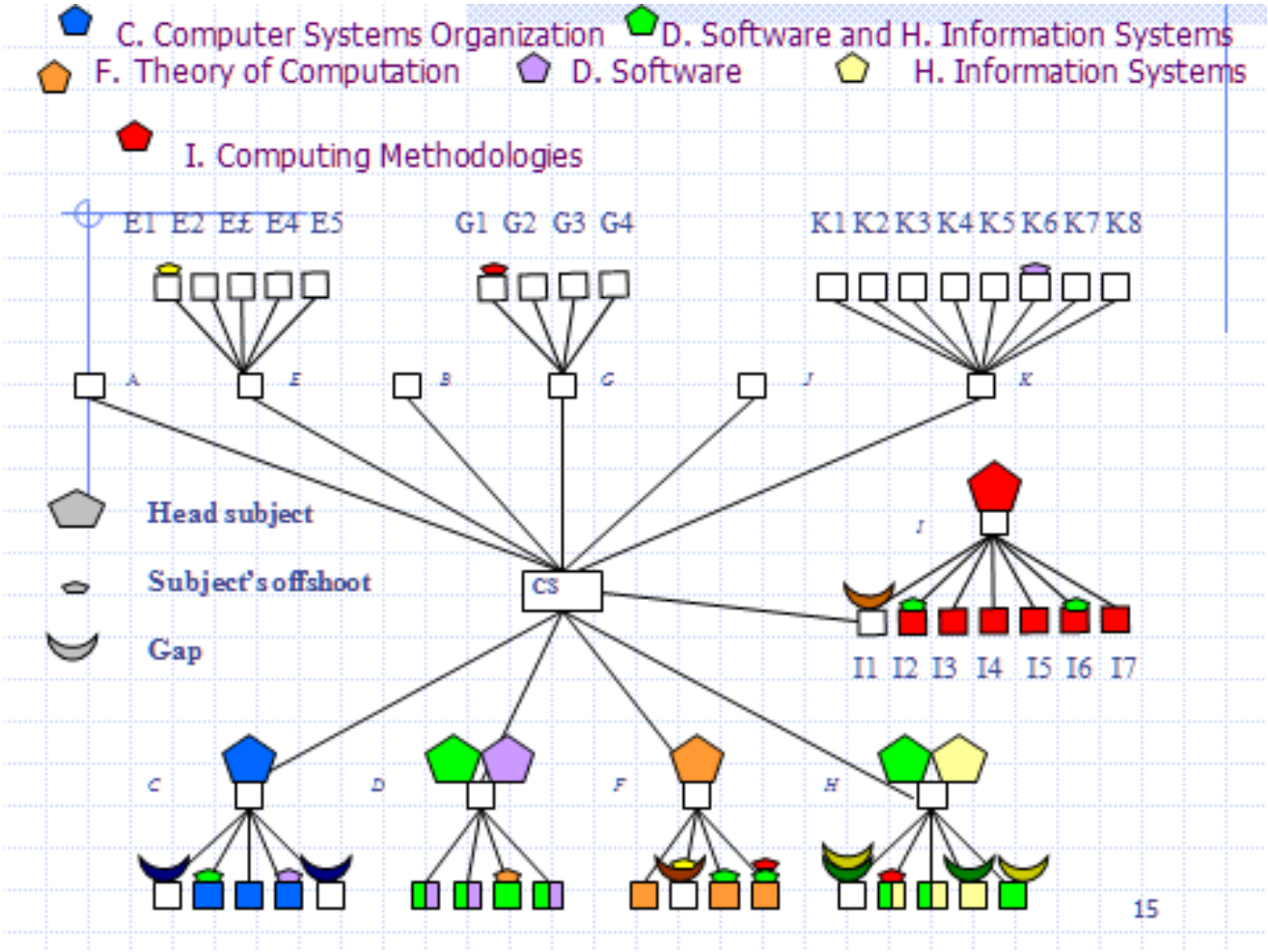


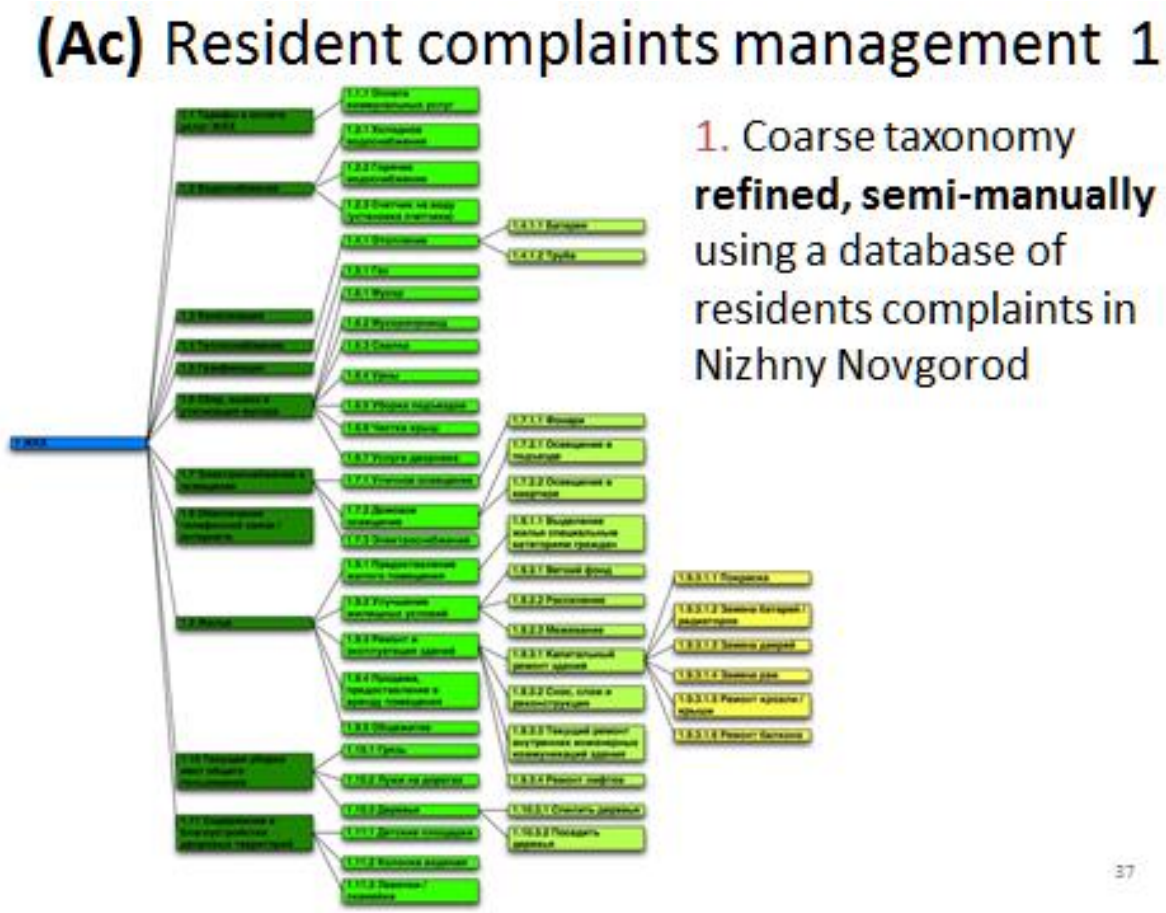
ИЛЛЮСТРАЦИЯ ЦЕЛЕЙ (примеры)

(а) структуризация (представление общей структуры данных)

- Структура научной тематики работ ЦЕНТРИА, Лиссабон (в таксономии АВМ)

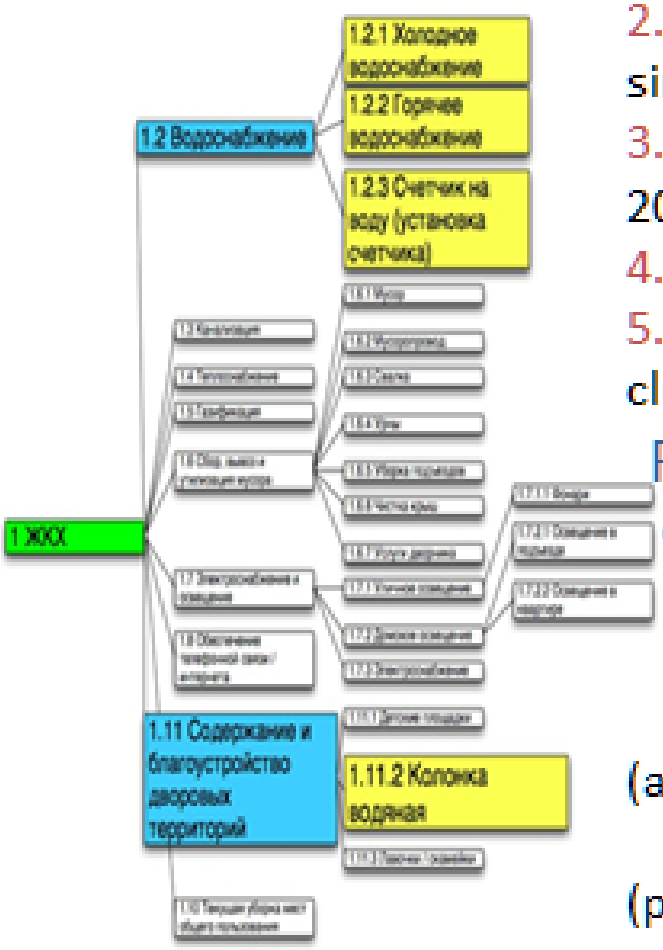


- **Жалобы жителей**: кластеры писем жителей, отображённые на таксономию тематики этих писем



Представление кластера **«жалобы на водоснабжение в квартирах»** в таксономии

(Ac) Resident complaints management 2



2. Complaint-to-Topic suffix tree based similarity table S
3. Clusters over S with iK-Means (Mirkin 2012) - Anomalous patterns one-by-one
4. Removal of small and large clusters
5. Parsimoniously lifting remaining clusters

Figure caption:

Cluster mapped to **1. Housing service**

1.2.1. Hot water problems

1.2.2. Cold water problems

1.2.3. Water meter problems

(all three are parts of **1.2. Water Supply**

1.11.2. Public water pump

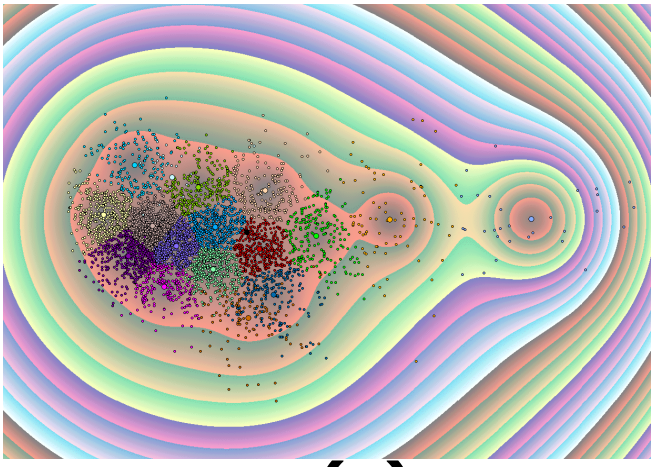
(part of **1.11. Urban landscaping and public amenities**)

39

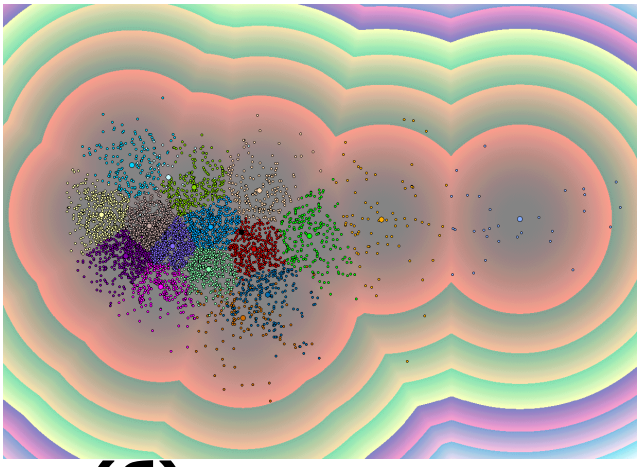
Вывод: Кластеры **не вписываются** в структуру городского хозяйства. Нужны комплексные «Центры услуг для горожанина»

(б) описание кластеров в терминах тех или иных признаков

- Принадлежность к совокупности 14 000 химических соединений (признаки структуры): Прогноз активности по регрессионному уравнению - только для тех соединений, что принадлежат!!!



(а)

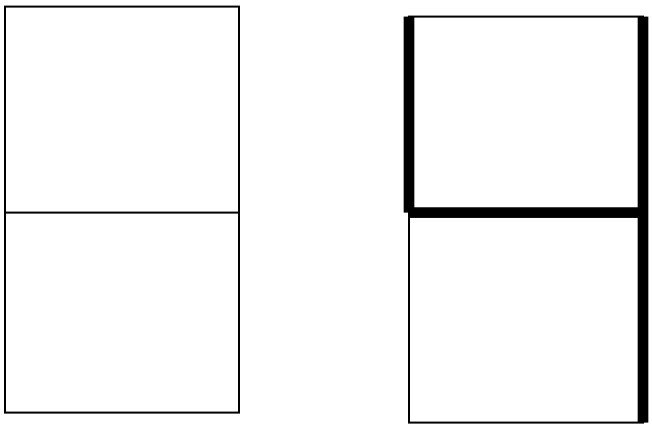


(б)

Карта химических соединений, автоматически сгруппированных в кластеры (41) на **(а)**

(в) установление взаимосвязи между различными аспектами явлений

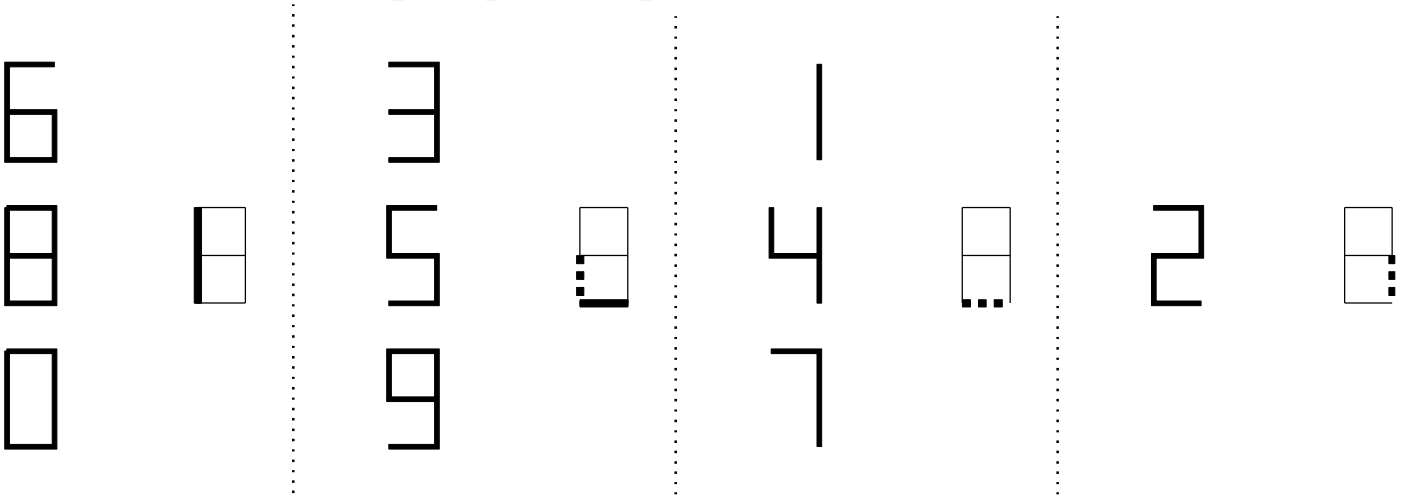
Стилизованные цифры и ошибки в их различении



Confusion between segmented numeral digits:

	1	2	3	**4**	5	6	**7**	8	9	0
1	**877**	7	7	**22**	4	15	**60**	0	4	4
2	14	782	47	4	36	47	14	29	7	18
3	29	29	681	7	18	0	40	29	152	15
4	**149**	22	4	**732**	4	11	**30**	7	41	0
5	14	26	43	14	669	79	7	7	126	14
6	25	14	7	11	97	633	4	155	11	43
7	**269**	4	21	**21**	7	0	**667**	0	4	7
8	11	28	28	18	18	70	11	577	67	172
9	25	29	111	46	82	11	21	82	550	43
0	18	4	7	11	7	18	25	71	21	818

Кластеры ошибок различения и признаки написания (справа)

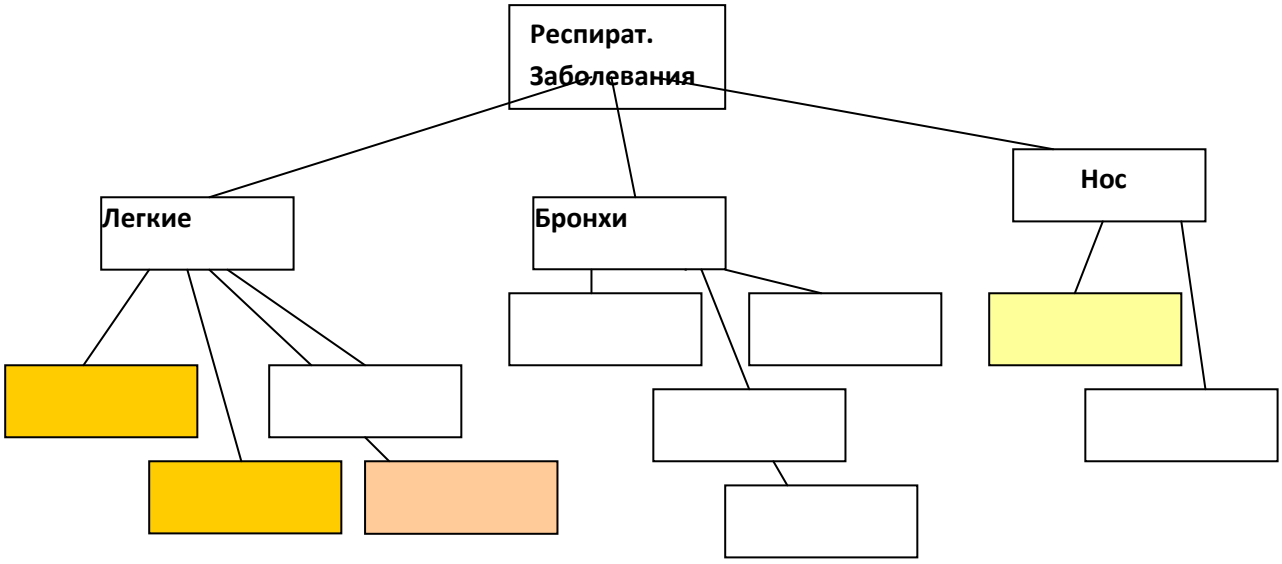


(г) формирование обобщающих утверждений о свойствах явлений

Ростовцев, Миркин, Шанин (1981):

исследование заболеваний органов дыхания и их факторов риска

50 000 анкет: 14 иерархических кластеров



Предполагаемые факторы риска:

Курение

Алкоголь

ФАКТОРЫ РИСКА ПО полученным нами кластерам:

Наличие заболевания в семье

Плохие жилищные условия

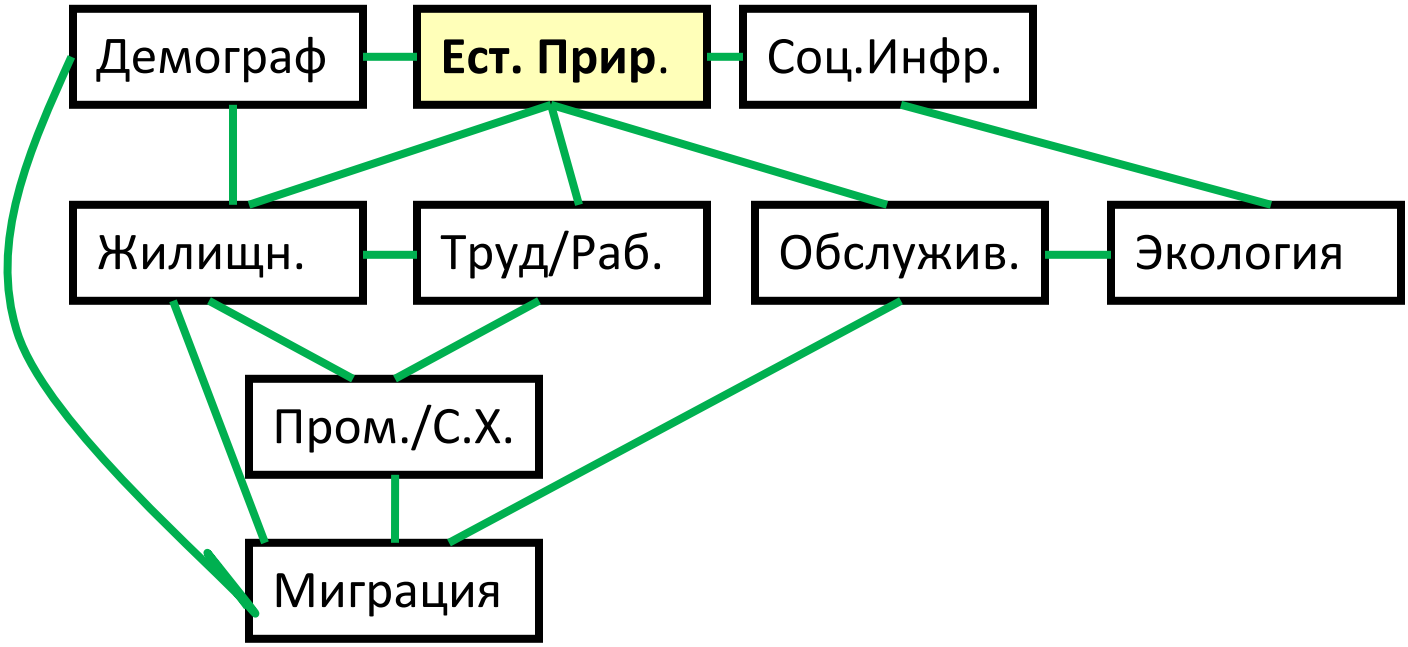
а Курение и Алкоголь **никак не связаны**

Выводы были **отвергнуты (1981)**

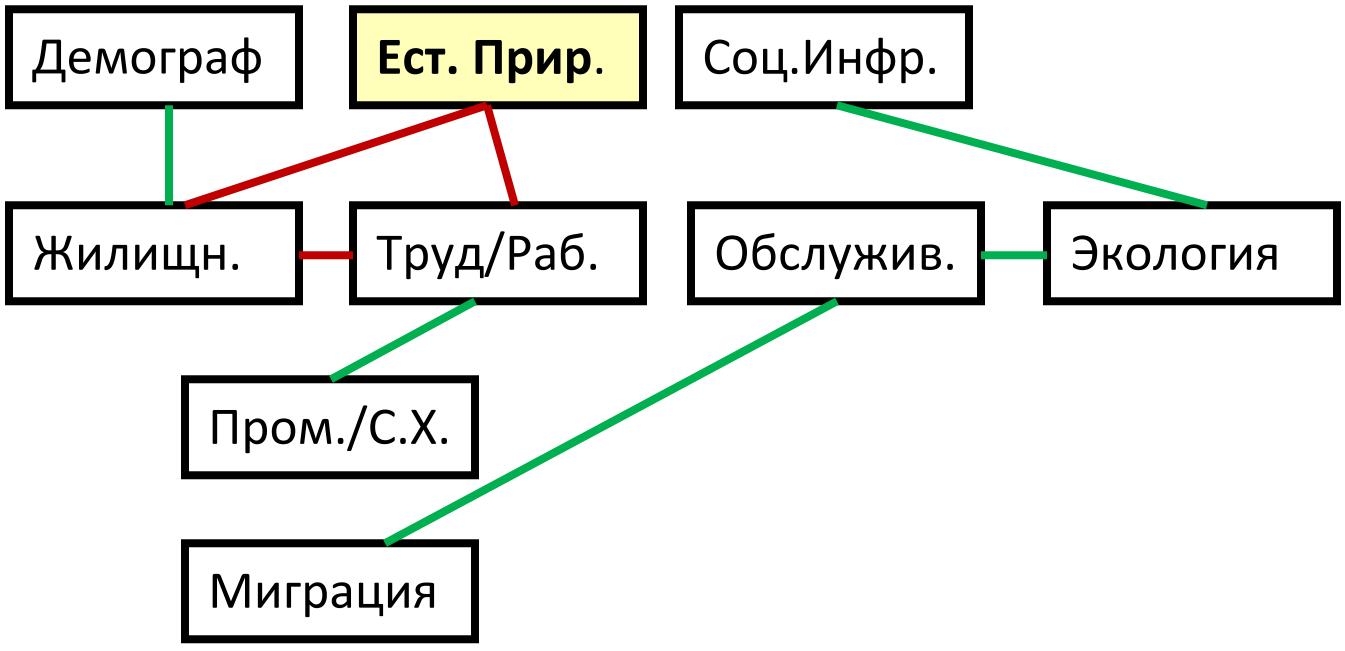
(д) визуализация данных в процессах принятия решений

Разрушение структуры факторов **прироста населения** Московской обл. (1979-88)

1979

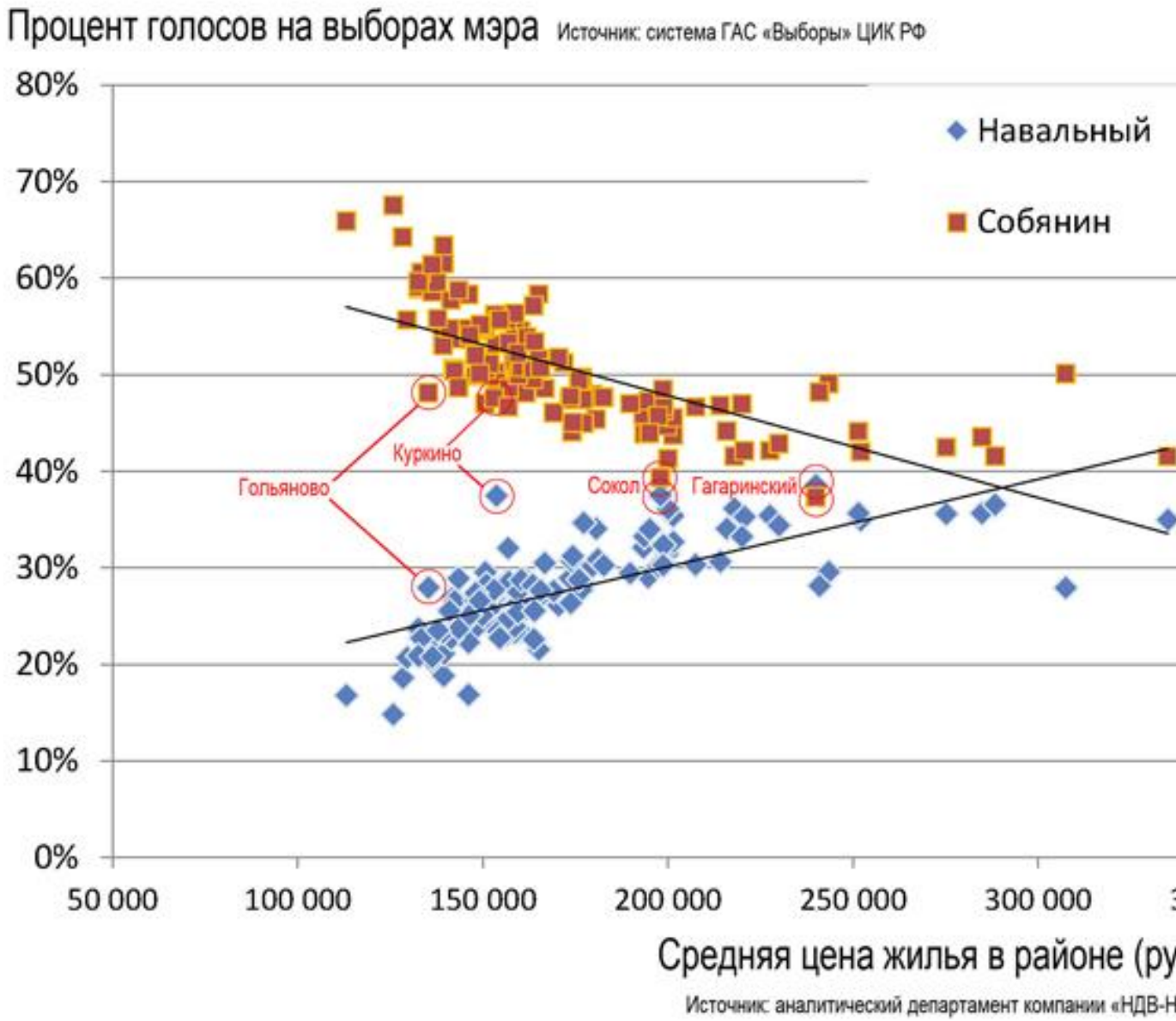


1988



- Выборы в Москве (В. Гурьянов, «Закон Бершидского: стоимость квадратного метра определила результаты выборов мэра», Квадратъ, №44, 16 Сентября 2013):

Два кластера в пространстве «стоимость жилья × % голосов за кандидата»



Вывод: кластеры важная часть автоматизации анализа данных