

# **Наступление метрик: оценка научного вклада**

**Б. Г. Миркин** (совместная работа с М.А. Орловым)

**Факультет Компьютерных Наук, НИУ Высшая школа экономики,  
Москва РФ**

**МЛАВР, НИУ Высшая школа экономики, Москва РФ**

**Department of Computer Science, Birkbeck University of London UK**

# Contents

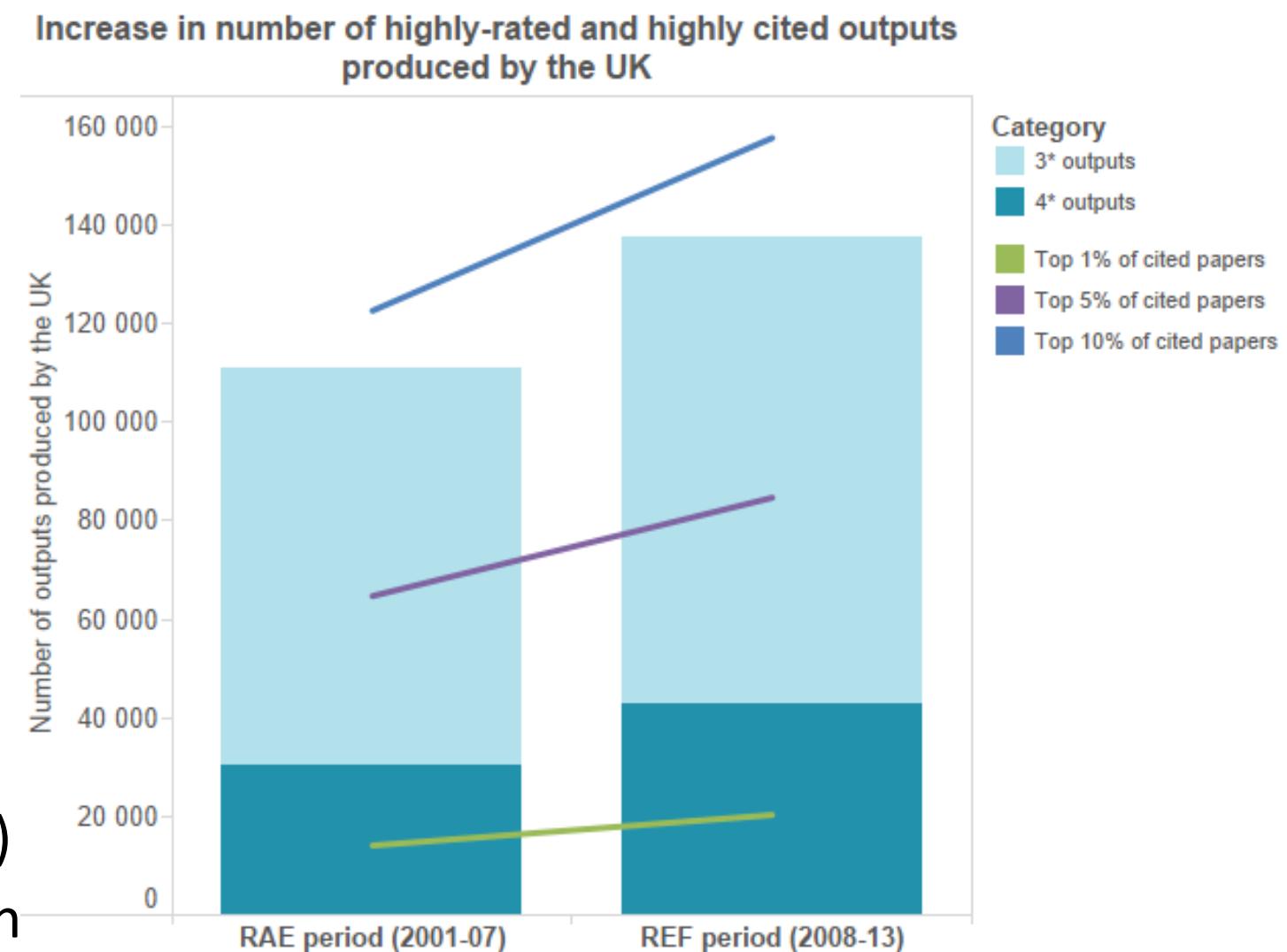
- Introduction: The problem of research impact assessment
- Method 1: Automatic aggregation of criteria
- Method 2: Using a domain taxonomy for assessment of quality of research results
- Application to the domain of Machine Learning/Data Analysis
- Conclusion: of developing a system for impact assessment

# Motivation

- University league tables
- Scientist citation indexes
- Desire of administration to follow formal (=objective?) procedures
- Research activities getting closer to other industries (because of twofold processes)

# Как это делается в Соединённом Королевстве

- RAE [REF from 2014]
- Every 5-6 years
- Over fields of science
- Evaluation of each UnDep by a high peer commission
- Final rates
  - P1% - 2 (National level)
  - P2% - 3 (International level)
  - P3% - 4 (Highest Internation)
  - P4% - 4\* (Groundbreaking)



# Обложка доклада комиссии REF (July 2015)



**Выводы:**

В настоящее время автоматизация оценки невозможна  
Увеличить финансирование разработок по теме

# Researcher's impact: sum of products in five areas

- Research and presentation of results (number, quality)
- Science functioning (journal editing, running research meetings)
- Teaching (knowledge transfer, knowledge discovery)
- Technology innovations (programs, patents)
- Societal interactions (popularization)

# Novel results 1: two methods, one application

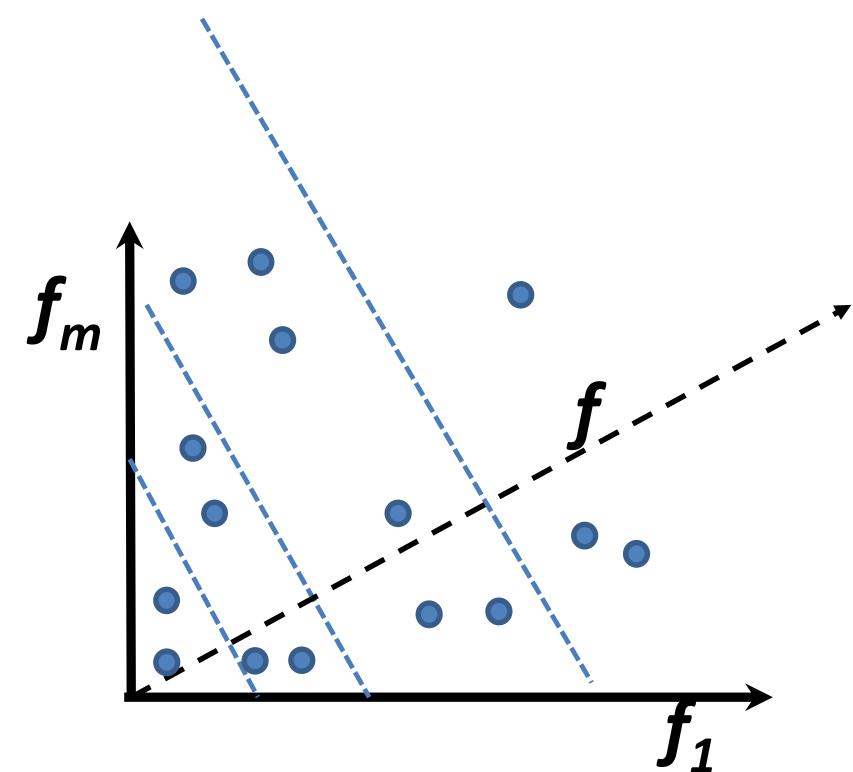
- research results – measuring by mapping to a taxonomy of the domain (manual)
- Science functioning (journal editing, running research meetings), teaching knowledge discovery (supervising) – unsupervised aggregation of criteria for stratification
- technology innovations and societal interactions – nothing, no data of academics
- Applying to the domain of data analysis

# Results 2: two methods, one application

- Method 1: Unsupervised aggregation of criteria for stratification Linstrat
- Method 2: Taxonomic rank (TaxRank) by mapping results to a taxonomy of the domain (manual)
- Application to the domain of data analysis
  - Taxonomy of Data analysis
  - Sample of Data analysis scientists
  - TaxRank scoring with a manual mapping
  - Aggregate Criteria: Panoramic[TaxRank, Citation[3], Merit[3]]

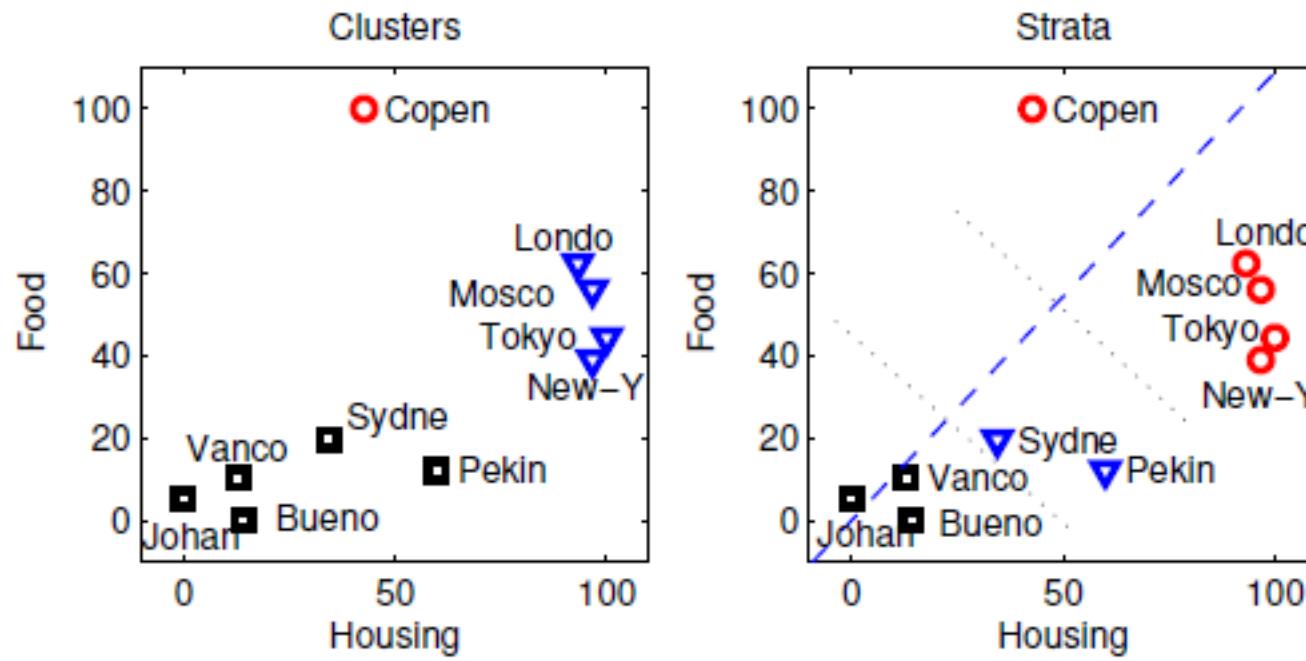
# Method 1: Convex combination of criteria

- Input: set of criteria  $f_1, f_2, \dots, f_m$  over an entity set  $I$
- Output: set of weights  $w = (w_1, w_2, \dots, w_m)$  so that  $I$  is divided in  $K$  strata over



$$f = \sum_{j=1}^m w_j f_j$$

# Method 1: Strata versus Clusters



# Method 1: Criterion for unsupervised stratification

**W to minimize the strata widths:** projections of entity points on  $f$  to fall as near to strata centers as possible

$$\min_{w,c,S} \quad \sum_{k=1}^K \sum_{i \in S_k} \left( \sum_{j=1}^M x_{ij} w_j - c_k \right)^2$$

such that  $\sum_{j=1}^M w_j = 1$

$$w_j \geq 0, j \in 1...M.$$

# Method 1. Ellistat - unsupervised K stratification

Minimize alternatingly:

- Initialise  $\mathbf{w}$  randomly
- given weights  $\mathbf{w}$ , find  $K$  centers  $\mathbf{c}_k$  and strata  $S_k$

- given  $\mathbf{c}_k$  and strata  $S_k$ , find  $\mathbf{w}$

$$\min_{w,c,S} \sum_{k=1}^K \sum_{i \in S_k} \left( \sum_{j=1}^M x_{ij} w_j - c_k \right)^2$$

$$\text{such that } \sum_{j=1}^M w_j = 1$$

$$w_j \geq 0, j \in 1 \dots M.$$

# Изменение критерия $\Rightarrow$ Изменение агрегации

Объект	Критерий А	Критерий Б	Агрегат $f_{AB}$	Страта
I1	0	1	0+0.9091	3
I2	0	2	0+1.8182	2
I3	0	3	0+2.7273	1
I4	10	0	0.9090+0	3
I5	20	0	1.8180+0	2
...	...	...	...	...

$$f_{AB} = 0.0909A + 0.9091B$$

объект	Критерий А	Критерий В	Агрегат $f_{AB}$	Страта
I1	0	80	0+7.7920	3
I2	0	90	0+8.7660	3
I3	0	100	0+9.7400	3
I4	10	0	9.0260+0	3
I5	20	0	18.0520+0	2
I6	30	0	27.0780+0	1

$$f_{AB} = 0.9026A + 0.0974B$$

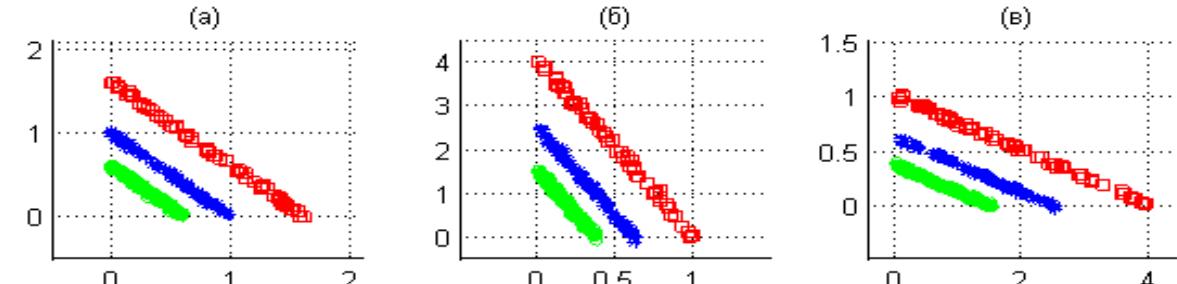
# Method 1: **Testing** **Linstrat** - Method for unsupervised **K** stratification The winner, at modest number of criteria (less than 20), not so wide strata

- Tested over synthetic datasets (accuracy)
- Tested over real datasets (centrality over KS-distance)
- Compared with other stratification heuristics (Pareto boundary extraction, linear program, etc.)

# Синтетические данные (Орлов 2014):

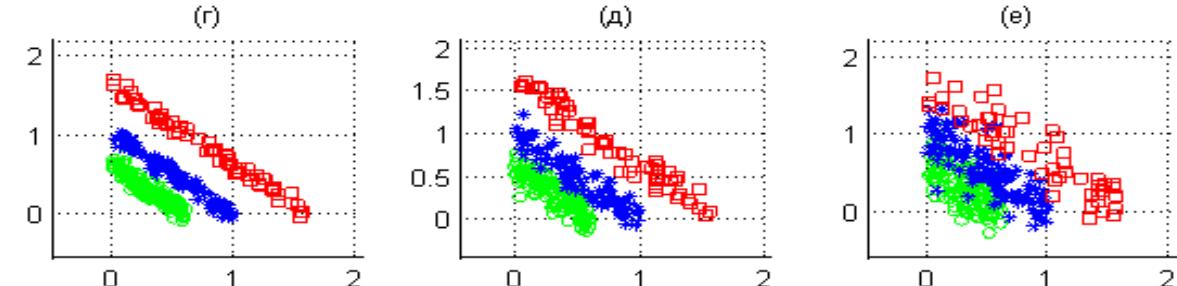
Страты от **ориентации**  $w$

- (а)  $w = (0.5, 0.5)$ , (б)  $w = (0.8, 0.2)$ ,  
(в)  $w = (0.2, 0.8)$



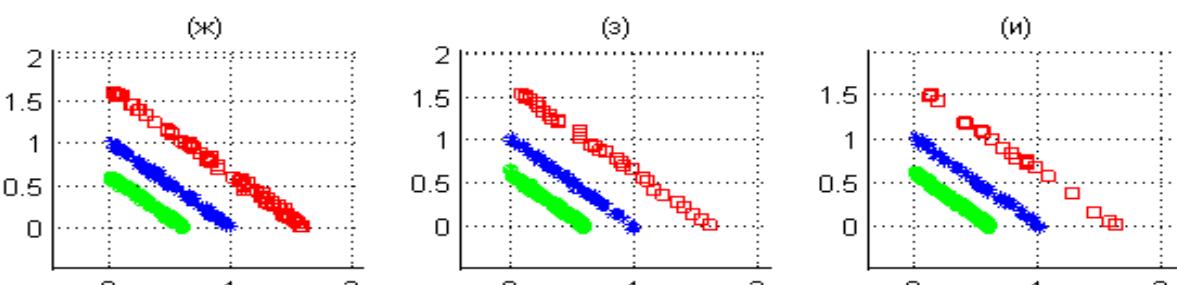
Страты от **толщины**

- (г)  $\sigma=0.05$ , (д),  $\sigma=0.1$  (е)  $\sigma=0.2$



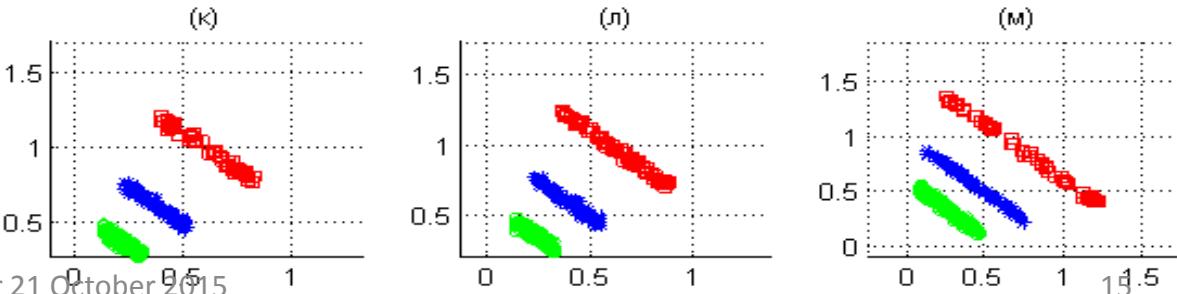
Страты от **интенсивности**:

- $\vartheta=(0.5, 0.3, 0.2)$  (з)  $\vartheta=(0.7, 0.2, 0.1)$   
(и)  $\vartheta=(0.8, 0.15, 0.05)$



Страты от **размаха**:

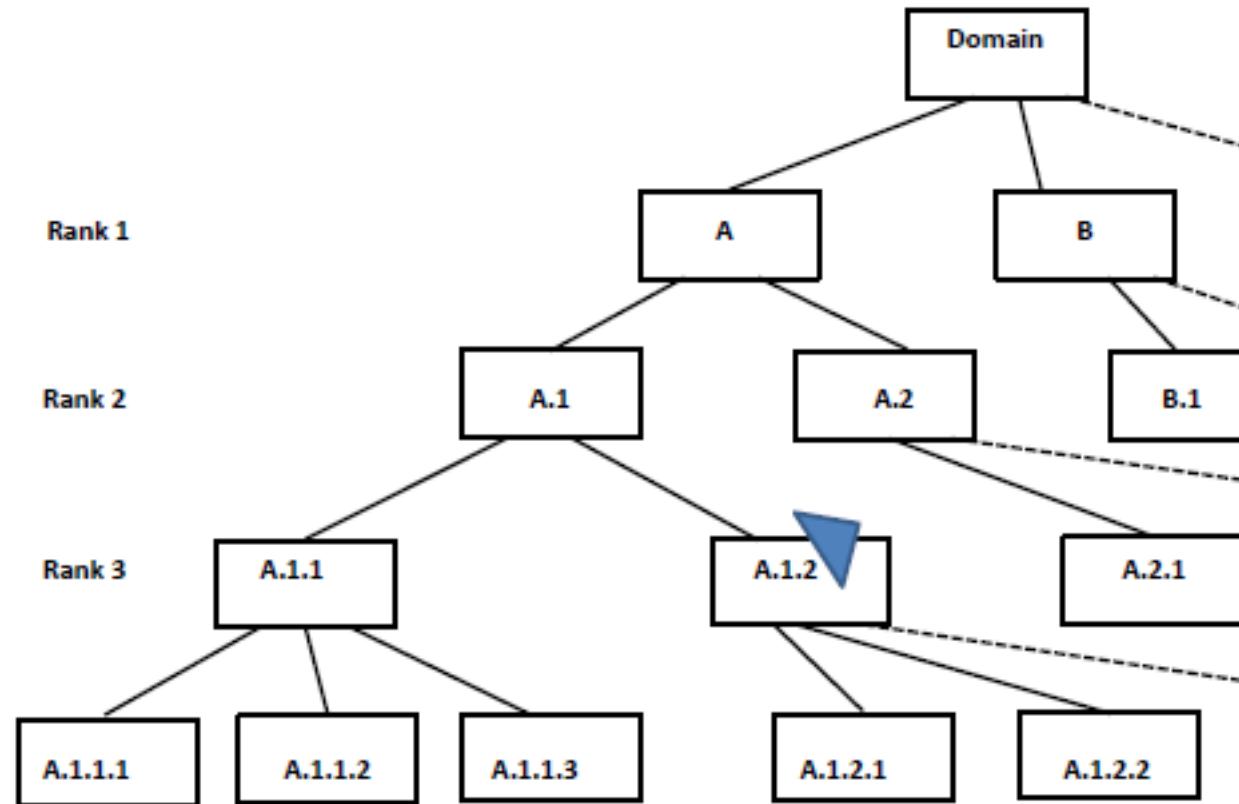
- (к)  $\varphi=0.05$ , (л)  $\varphi=0.1$ , (м)  $\varphi=0.5$ .



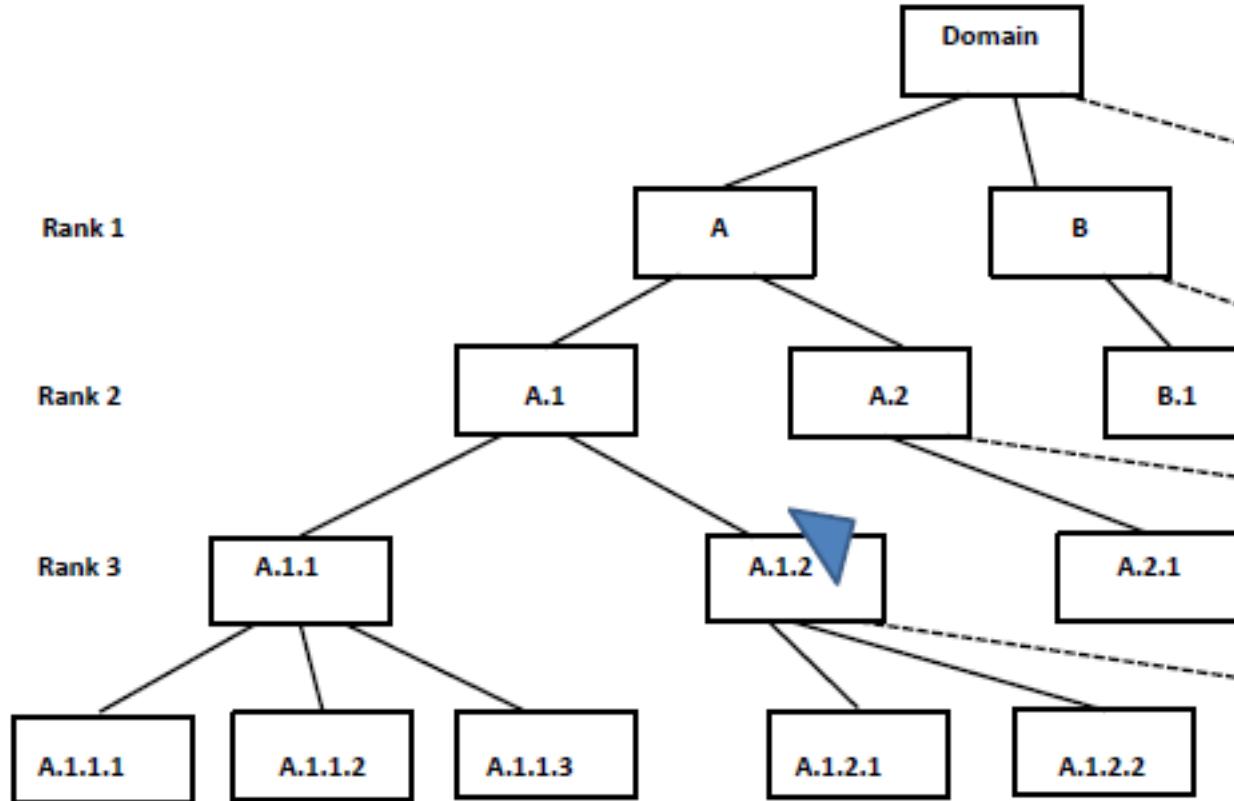
# Методы, участвовавшие в сравнении

Метод стратификации	Аббревиатура	Источник
Метод линейной стратификации Linstrat-Q с использованием квадратичного программирования	Linstrat_Q	[Орлов 2014]
Метод линейной стратификации Linstrat на основе эволюционной минимизации	Linstrat_E	[Миркин, Орлов 2013]
Стратификация с помощью правила Борда (Borda count)	BC	[Алескеров, Хабина, Шварц 2006]
Метод ABC- классификации на основе линейной оптимизации весов (Linear weights optimization)	LWO	[Ramanathan 2006]
Ранжирование по влиянию (Authority ranking)	AR	[Sun et. al 2009]
Стратификация объединением границ Парето (Paretostrat)	PS	[Миркин, Орлов 2013]

Method 2: rank of result is rank of the taxon in a Domain Taxonomy that has emerged or been drastically transformed because of it



Method 2: **rank** defined by taxa emerged or drastically transformed because of results:  
here rank **3**, or **2.9** if two taxa affected

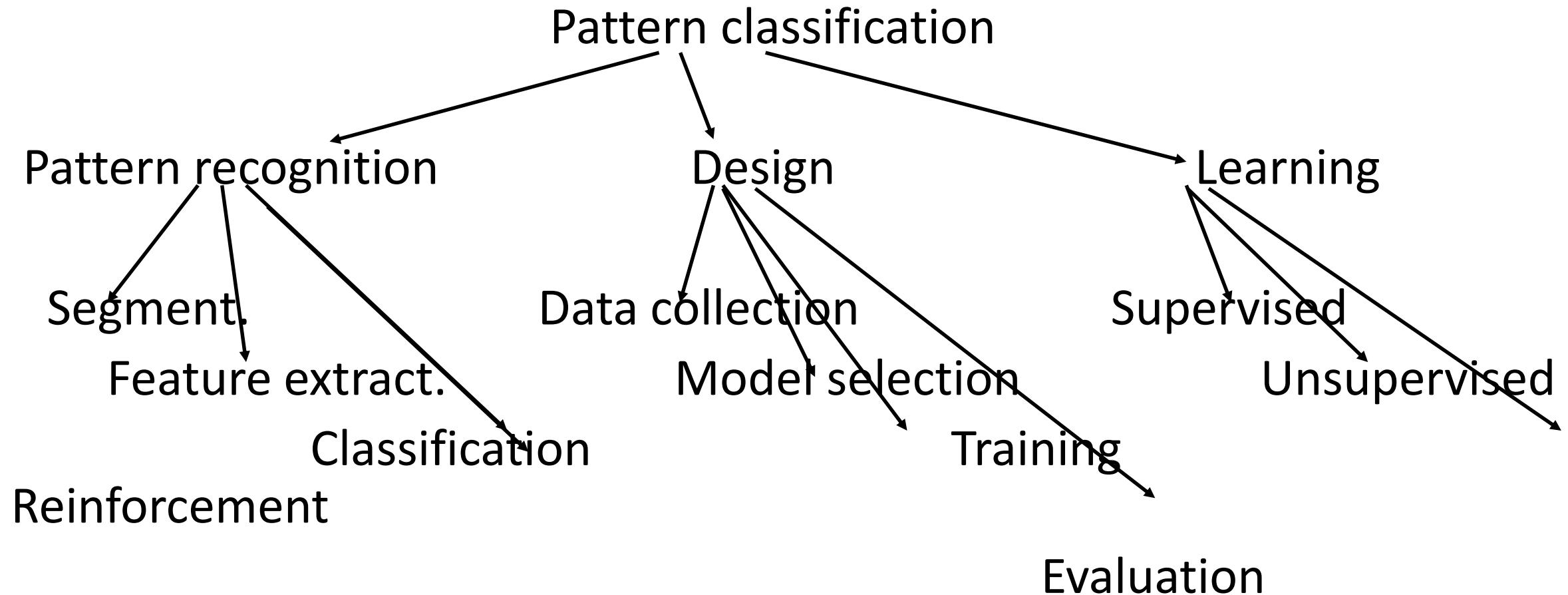


# Method 2: Scoring quality of research results

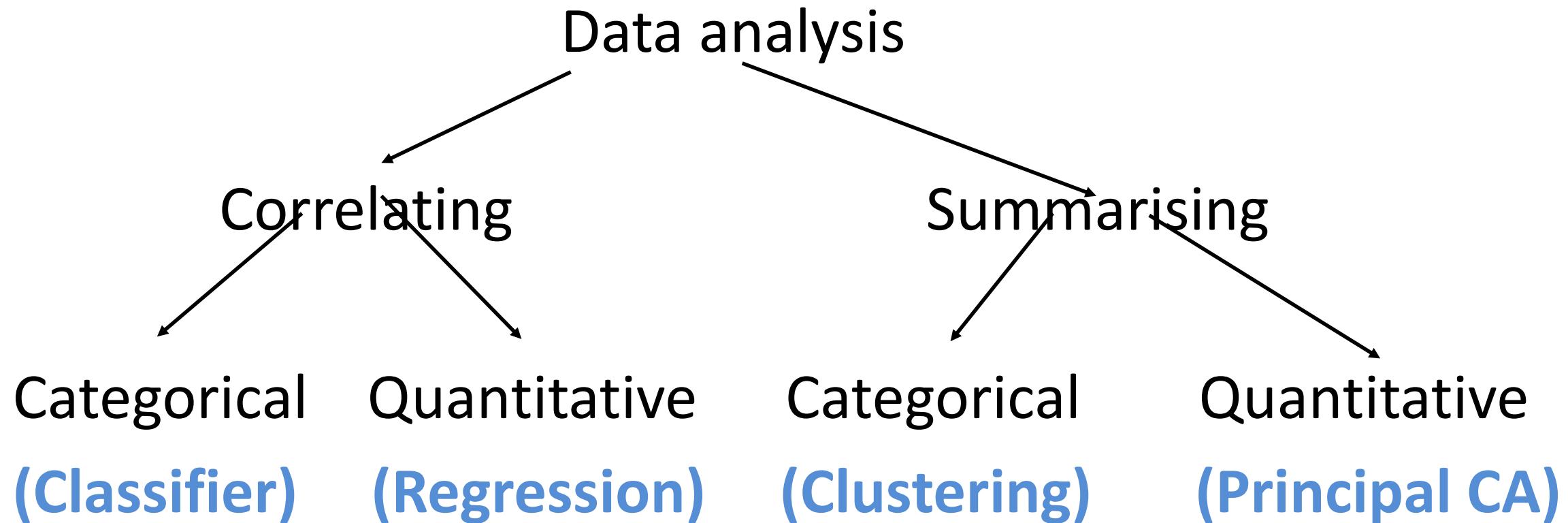
- **Choosing domain (Data analysis)**
- **Defining taxonomy**
- **Sampling scientists**
- **Choosing their results**
- **Mapping them to taxonomy**
- **Scoring Taxonomy rank of scientists**

# Defining Taxonomy: Pattern classification

## Duda/Hart/Stork (2001) - poor



# Defining Taxonomy: Core Data Analysis Mirkin (2011)- poor



# Defining Taxonomy: ACM CCS 2012 - good

Subject index	Subject name
1.	Theory of computation
1.1.	Theory and algorithms for application domains
2.	Mathematics of computing
2.1.	Probability and statistics
3.	Information systems
3.1.	Data management systems
3.2.	Information systems applications
3.3.	World Wide Web
3.4.	Information retrieval
4.	Human-centered computing
4.1.	Visualization
5.	Computing methodologies
5.1.	Artificial intelligence
5.2.	Machine learning

# Defining Taxonomy: ACM CCS 2012 Lower level good

3.2.1.	Data mining
3.2.1.1.	Data cleaning
3.2.1.2.	Collaborative filtering
3.2.1.2.1**	Item-based
3.2.1.2.2**	Scalable
3.2.1.3.*	Association rules
3.2.1.3.1**	Types of association rules
3.2.1.3.2**	Interestingness
3.2.1.3.3**	Parallel computation
3.2.1.4.	Clustering
3.2.1.4.1**	Massive data clustering
3.2.1.4.2**	Consensus clustering
3.2.1.4.3**	Fuzzy clustering
3.2.1.4.4**	Additive clustering
3.2.1.4.5**	Feature weight clustering
3.2.1.4.6**	Conceptual clustering
3.2.1.4.7**	Biclustering
3.2.1.5.	Nearest neighbor search

# Sampling scientists

- Data (from Google):
  - research publications/results
  - citation [total #, #10, Hirsch index]
  - “merit” [PhDs supervised, (co)-editing, plenary talks]
- 30 leading scientists in data analysis, data mining, knowledge discovery
- Diversity: About half are from the USA, 2-3 from each UK, Netherlands, China, Russia, etc.
- Diversity: From three-four thousand citations in Europe to a hundred thousand citations in the USA

# Sample of scientists: anonymous

<a href="#">Alexander N. Gorban</a>	5,5,4	3,88	73
<a href="#">Andrew Mccallum</a>	4,4,4,4,4	3,50	100
<a href="#">Anil K. Jain</a>	5,5,5,5,5	4,50	29
<a href="#">Bernhard Scholkopf</a>	5,5,5,5,4,5	3,90	71
<a href="#">Boris Mirkin</a>	5,5,5,5,5	4,50	29
<a href="#">Christos Faloutsos</a>	4,5,5,4,5	3,77	81
<a href="#">Daphne Koller</a>	5,5	4,80	7
<a href="#">Fionn Murthagh</a>	5,5,5,5,5	4,50	29
<a href="#">Geoff McLachlan</a>	5,5,5,5,5	4,50	29
<a href="#">Geoffrey Hinton</a>	5,5	4,80	7
<a href="#">George Karypis</a>	4,5,5,5,5	3,86	74
<a href="#">Ian H. Witten</a>	5,4,6,5,5,5	3,86	74
<a href="#">Inderjit S. Dhillon</a>	5,4,5,5,5	3,86	74
<a href="#">Jiawei Han</a>	B. Mirkin Seminar 21 October 2015	4,90	0 25

# Results: Linstrat aggregate citation at 3 strata

**CITATION =**

**0.5\*Total\_citat+0.5\*Citat\_over10+0.0\*Hirsh\_Index**

# Results: Linstrat aggregate merit at 3 strata

MERIT =

**0.22\*#PhD+0.10\*Conf\_Ch+0.69\*Edit/AssocEJ**

## Results:

Aggregate **taxonomic rank , citation, merit** correlation

	Tr	Cit	Merit
Tr		-.12	-.04
Cit			.31
Merit			

Citation/Merit (.31): **Scientist's Popularity**

TaxR versus Cit|Merit: **No Correlation**

**Results:** Aggregate criterion **1**

**Panoramic =**

**0.80\*TaxRank + 0.04\*Citation + 0.16\*Merit**

## Results: Aggregate criterion **2**

2 версия = Changed from версии 1:

- Taxonomy drastically streamlined
- Results of 8-10 papers (not 4-5) taken into account
- 6 scientists from the sample have TaxRank drastically increased (Linstrat will decrease the weight)

# 1<sup>st</sup> Taxonomy: ACM CCS 2012 – 5 nodes of rank1

Subject index	Subject name
1.	Theory of computation
1.1.	Theory and algorithms for application domains
2.	Mathematics of computing
2.1.	Probability and statistics
3.	Information systems
3.1.	Data management systems
3.2.	Information systems applications
3.3.	World Wide Web
3.4.	Information retrieval
4.	Human-centered computing
4.1.	Visualization
5.	Computing methodologies
5.1.	Artificial intelligence
5.2.	Machine learning

## Вторая таксономия «Анализа данных» (очищенная от повторений) – 2 узла ранга 1

Индекс	Название предмета		
1.	Information systems		
1.1.	Information systems applications		
1.2.	World Wide Web		
1.3.	Information retrieval		
2	Computing methodologies		
2.1.	Artificial intelligence		
2.2.	Machine learning		
2.3	Visualization		

## Results: Aggregate criterion 2

2 = Changed from 1:

- Taxonomy drastically streamlined
- Results of 8-10 papers (not 4-5) taken into account
- 6 scientists from the sample have TaxRank drastically increased (Linstrat will decrease the weight)

Panoramic\_Second =

$$0.12 * \text{TaxRank} + 0.02 * \text{Cit} + 0.85 * \text{Merit}$$

Panoramic\_First =

$$0.80 * \text{TaxRank} + 0.04 * \text{Citation} + 0.16 * \text{Merit}$$

# CONCLUSION 1

- A method for assessment of quality of research involving
  - (i) ranking,
  - (ii) mapping to taxonomy
- This activity should be further pursued by us as innately related decision analysis and choice
- Should result in both **methods and substance**

# CONCLUSION 2: Researcher's products in 5 areas

- **Research and presentation of results**
  - Publications
  - Presentations
  - Funded and unfunded projects
- **Participation in Science functioning**
  - Journal editing
  - Running research meetings
  - Refereeing
  - Research cooperation
  - Research societies

# CONCLUSION 3: Researcher's products in 5 areas

- **Teaching**

- **knowledge**

- Lectures
    - Seminars
    - Projects
    - Consultation
    - Assessments and exams
    - Textbooks

- **knowledge discovery**

- PhD Students
    - Research students

# CONCLUSION 4: Researcher's products in 5 areas

- **Technology innovations**

- Programs
- Services
- Patents
- Industrial consultations

- **Societal interactions**

- Popular books
- Articles
- Blogs
- Networks

# CONCLUSION 5: Expected results

- In substance:
  - Developing a system for assessment of research impact
  - Maintaining the system
  - Taxonomy of field(s) of science
  - Cataloguing research results and researchers
  - Forum for discussing taxonomy and results
- In methods:
  - Improving the concept of Taxonomy
  - Methods for relating paper texts and taxonomy
  - Methods for taxonomy building using research texts
  - Methods for mapping research results to taxonomy
  - Ranking impact of results
  - Aggregate rankings