Data Summarization at Clustering and Ranking

Boris Mirkin

Prof., Data Analysis & Machine Intelligence, Faculty of Computer Science, NRU Higher School of Economics, Moscow RF

Prof. (Emeritus), Computer Science & IS, Birkbeck University of London UK

Data Summarization at Clustering and Ranking: Outline 1

- A summarization data recovery model: PCA and SVD

- Extensions: Latent Semantic Analysis, Correspondence Analysis, Topic Allocation, …

- K-Means data recovery model and Anomalous clusters

- K-Means, Pythagoras, and Anomalous cluster criterion
- Anomalous cluster method and iK-Means
- Extending one-by-one anomalous clusters:
- Minkowski Weighted Features iK-Means;
- Delineating upwellings on temperature maps;

Data Summarization at Clustering and Ranking: Outline 2

Metric Tide: Ranking research results and impacts

- Automatic aggregation of criteria
- Domain taxonomy for ranking quality of research results
- Applying to Data Analysis domain

Conclusion

- Summarization versus Prediction
- Big Data
- Of a project in research ranking: work to do & outcome

Data recovery summarization: student marks I

#	Sen	OOP	CI	Average
1	41	66	90	65.7
2	57	56	60	57.7
3	61	72	79	70.7
4	69	73	72	71.3
5	63	52	88	67.7
6	62	83	80	75.0

F. Galton: Talent is inherited; let us measure it

K. Pearson: find student Talent score Tal(Stud), Subject loading Load(Subj)

Multiplicative Decoder

RecMark(Stud, Subj)= **Tal**(Stud)***Load**(Subj)

Criterion: summary squared error

|RecMark(Stud, Subj) - ObsMark(Stud, Subj)|²

Data recovery summarization: student marks 2

Summarization Data Recovery Model

ObsMark(i,v)= **Tal**(i)∗**Load**(v) + Error(i,v)

Criterion: summary squared error

$|\text{RecMark}(\text{Stud, Subj}) - \text{ObsMark}(\text{Stud, Subj})|^2$

Summarization Data Recovery Model

$$\mathbf{Mark}(i,v) = \mathbf{Tal}(i) * \mathbf{Load}(v) + E(i,v)$$

$$||\mathbf{E}||^2 \Rightarrow \min$$

Solution: Principal Component

Tal, Load, $||\mathbf{E}||^2$

Data recovery summarization: student marks 4

$$\mathbf{Mark}(i,v) = \mathbf{Tal}(i) * \mathbf{Load}(v) + E(i,v)$$

$$\|\mathbf{E}\|^2 \Rightarrow \min$$

Solution: Principal Component

$$\mathbf{Tal} = \mu^{1/2}\mathbf{z}, \quad \mathbf{Load} = \mu^{1/2}\mathbf{c}$$

Pythagorean: $\|\mathbf{X}\|^2 = \mu^2 + \|\mathbf{E}\|^2$ **(*)**

first singular triplet of mark matrix **(μ, z, c)**

$$\mathbf{Xc} = \mu\mathbf{z}, \quad \mathbf{X}^T\mathbf{z} = \mu\mathbf{c}$$

Data recovery summarization: PCA=SVD

$$X = Z * C^T + E$$

Find

Z **Entity × Hidden factor** rank p

C **Feature × Hidden factor** rank p

$$||E||^2 \Rightarrow \min$$

Solution: Principal Components = SVD

$$Z = M^{1/2} Z^*, \quad C = M^{1/2} C^*$$

SVD: $X = Z^* M C^T$ **(Orthonormal)**

Pythagorean: $||X||^2 = \mathrm{Sum}_k \mu_k^2 + ||E||^2$ **(*)**

Data recovery summarization: SVD methods

Principal Component Analysis (PCA)

- Hidden factor in organization systems
- Data reduction
- Data visualization
- Data interpretation

Latent Semantic Analysis (LSA)

- Information retrieval, tackling polysemy and homonymy

Correspondence Analysis (CA)

- Co-occurrence data; product design

Data recovery summarization: popular methods

Principal Component Analysis (PCA)

Data - entity × feature

Decoder **ZC**

Z - entity×hfactor

C - hfactor ×feature

Topic Allocation (LDA)

Data – Probability(word/text)

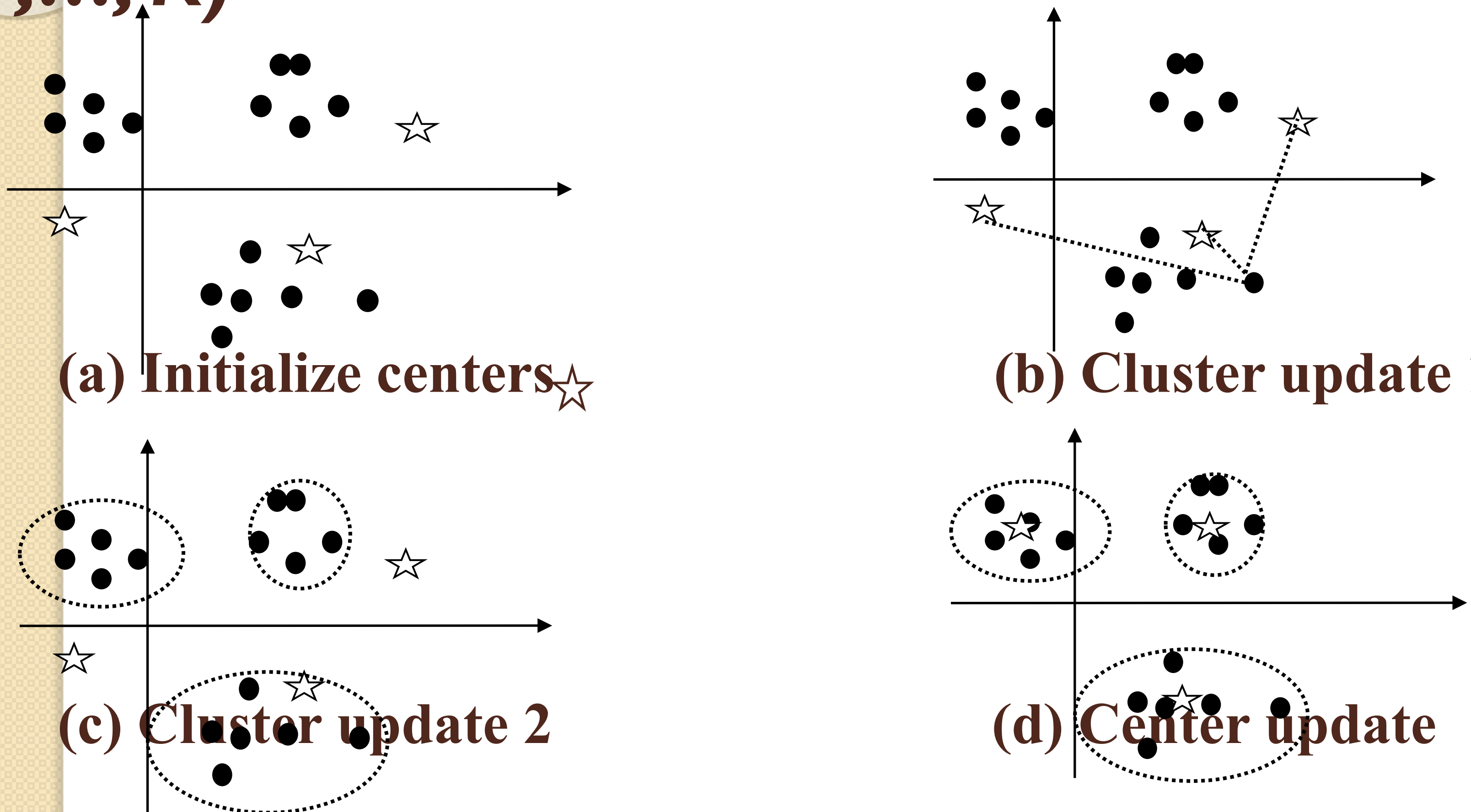
Decoder **ZC**

Z – Probability(word/htopic)

C – Probability(htopic/document)

K-Means clustering as data recovery summarization: Algorithm

Partition with Clusters k: center c_k and set S_k ($k=1,\ldots,K$)



K-Means Clustering: Good

Advantages:

- K-Means computations model typology making
- Computation is intuitive
- Computation is fast and requires no additional memory
- Computation is easy to parallelize (big data)

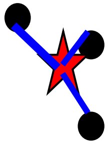
K-Means Clustering: Bad

Issues:

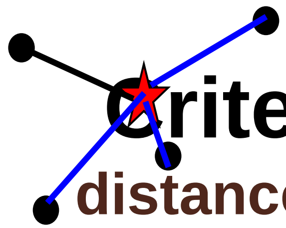
- ❑ Would the K-Means computation ever converge?
- ❑ Results depend on the initialization, how one should initialize?
- ❑ How number of clusters K should be chosen?
- ❑ Helpless against wrong/noise features.

K-Means clustering: Alternating minimization

Find partition *S* and centers *c* to **minimize**:



$$W(S,c)=\sum_{k=1}^{K}\sum_{i\in S_k} d(y_i,c_k)$$



Criterion: Sum of squared Euclidean distances between entities and centers of their clusters

K-Means: Alternating minimization of *W(S,c)*

K-Means: Equivalent criterion

How initial centers should be chosen? More theory

Minimize

$$W(S,c) = \sum_{k=1}^{K} \sum_{i \in S_k} d(y_i, c_k)$$

over S and c.

Equivalent criterion:

Maximize

$$B(S,c) = \sum_{k=1}^{K} |S_k| < c_k, c_k >$$

$<c_k, c_k>$ - *Euclidean squared distance between 0 and c_k*

Data scatter (sum of squared data entries) = = *W(S,c)+B(S,c)*

Data scatter is constant while partitioning

K-Means SVD-like **data recovery** clustering model

[Mirkin 87 (Rus), 90 (Eng)]

$$Y = ZC^T + E \qquad (*)$$

Y - $N \times V$ **data matrix**

Z - $N \times K$ **0/1 cluster membership**

C - $V \times K$ **center matrix**

E - $N \times V$ **residual matrix**

$$\min_{Z,C} [\|E\|^2 = W(S,c)] \qquad (**)$$

Criteria from (***) :

Minimize

$$W(S,c) = \sum_{k=1}^{K} \sum_{i \in S_k} d(y_i, c_k)$$

or Maximize

$$B(S,c) = \sum_{k=1}^{K} |S_k| < c_k, c_k >$$

over S and c.

Pythagorean decomposition

$$\|Y\|^2 = W(S,c) + B(S,c) \qquad (***)$$

K-Means : Anomalous criterion

Part 3: How initial centers should be initialized?, 5

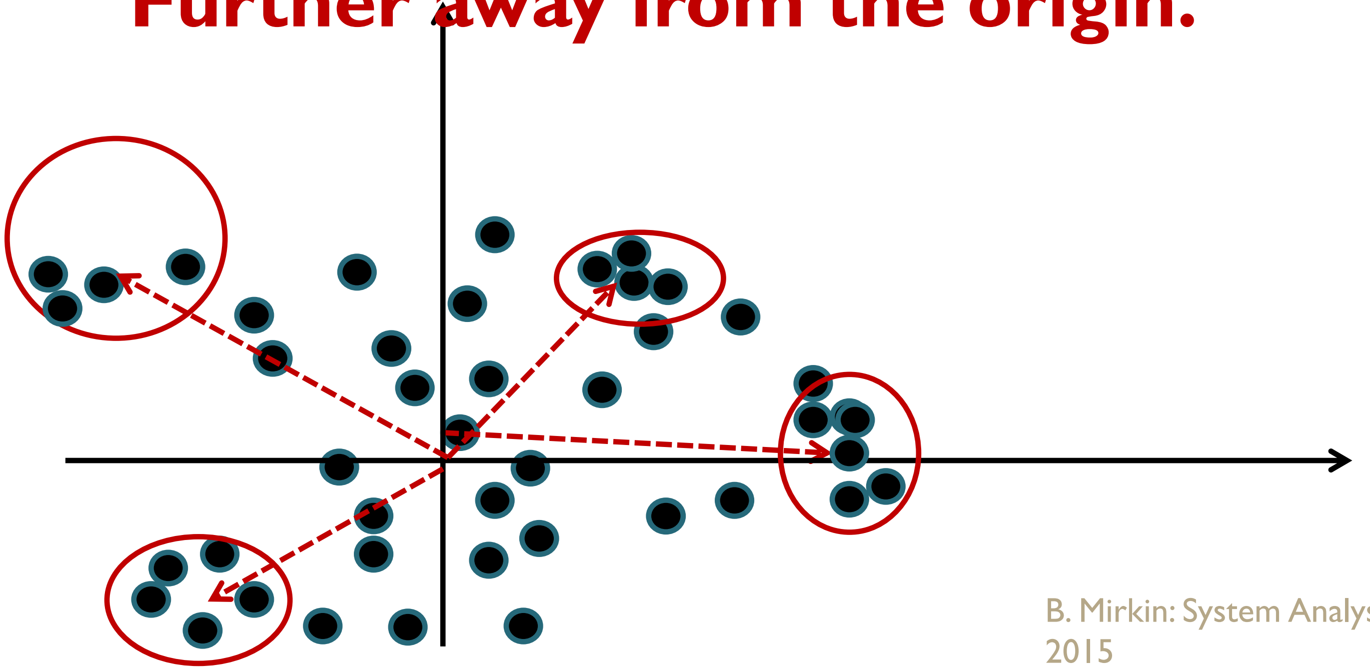
Maximize $B(S,c) = \sum_{k=1}^{K} |S_k| < c_k, c_k >$

Preprocess data by centering: 0 is grand mean

$<c_k, c_k>$ - *Euclidean squared distance between 0 and c_k*

Look for anomalous & populated clusters!!!

Further away from the origin.

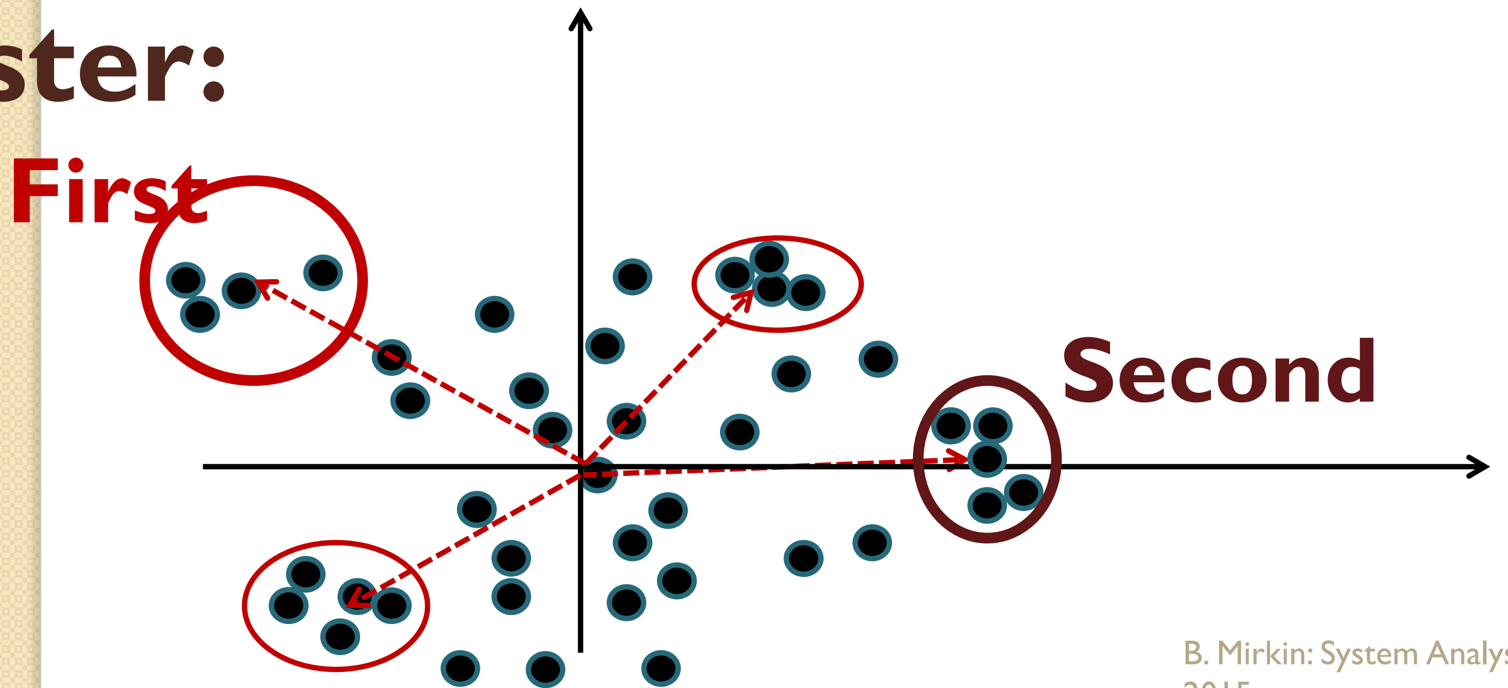


K-Means : Anomalous clusters and intelligent K-Means, I

Preprocess data by centering: 0 is grand mean

Look for anomalous & populated clusters!!!

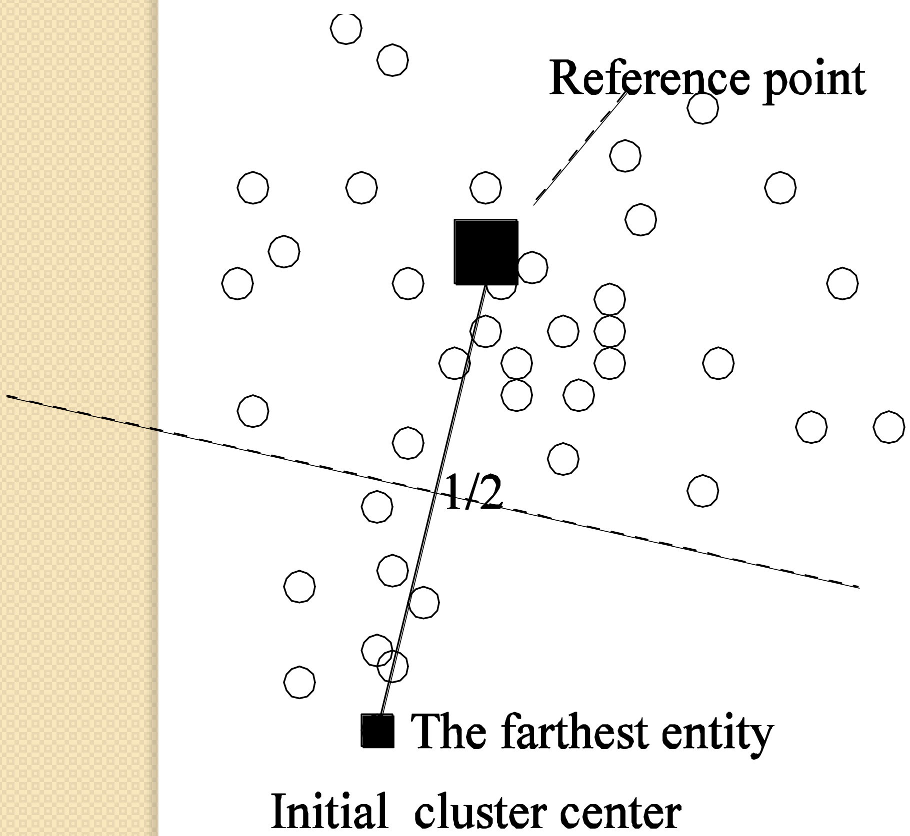
If *K* is unknown, do that cluster by cluster:

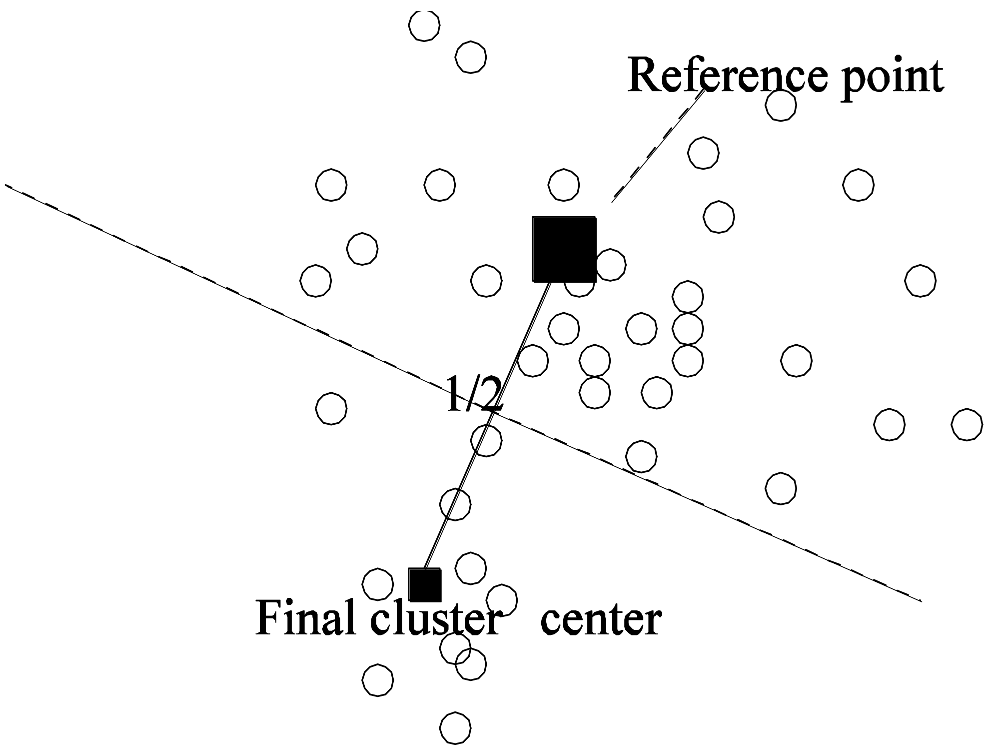


K-Means: Anomalous clusters and intelligent K-Means 2

Preprocess data by centering to Reference point.

Build just one Anomalous cluster.





K-Means: Anomalous clusters and intelligent K-Means,3

Preprocess data by centering to Reference point, typically grand mean. Build just one Anomalous cluster:

1. **Initial** center ***c*** is entity farthest away from **0**.

2. **Cluster update.** if $d(y_i,c) < d(y_i,0)$, assign y_i to **S**.

3. **Centroid update:** Within-***S*** mean **c'** if **c'** $\neq$ **c**. Go to 2 with **c**$\Leftarrow$**c'**. Otherwise, halt.

K-Means: Anomalous clusters and intelligent K-Means,4

Anomalous Cluster is (almost) K-Means up to:

(i) the number of clusters K=2: the "anomalous" one and the "main body" of entities around 0;

(ii) center of the "main body" cluster is forcibly always at 0;

(iii) a farthest away from 0 entity initializes the anomalous cluster.

K-Means: Anomalous clusters and intelligent K-Means,5

Anomalous Cluster applied to Iris (150×4) dataset just centered (no further normalization):

Initial center: the furthest away entity 132

c0=(1.8567 -0.4573 3.1420 1.1007)

- 27 entities are closer to c0 than to 0; their center

c1=(1.1641 0.0390 2.1716 0.9377)

- 47 entities are closer to c1 than to 0; their center

c2=(0.8865 -0.0361 1.8399 0.8156)

- 58 entities are closer to c2 than to 0; their center

c3=(0.7618 -0.0729 1.7023 0.7593)

- 60 entities are closer to c3 than to 0; their center

c4=(0.7600 -0.0773 1.6737 0.7407)

STABLE !

Anomalous clusters and intelligent K-Means,6

Anomalous Cluster at Iris, ITERATIVELY to those yet unclustered:

AnomClus 1		Center				Contribution
60 entities	c=(0.7600	-0.0773	1.6737	0.7407)		**34.6%**
AnomClus 2						
50 entities	c=(-0.8373	0.3707	-2.2960	-0.9533)		**51.5%**
AnomClus 3						
31 entities	c=(-0.1853	-0.4122	0.3872	0.0684)		**1.6%**
AnomClus 4	{67}	singleton				**0.2%**
AnomClus 5	5 entities					**0.6%**
AnomClus 6	{98}	singleton				Less 0.1%
AnomClus 7	{99}	singleton				Less 0.1%
AnomClus 8	{55}	singleton				Less 0.1%

iK-Means

iK-Means is superior in experiment (Chiang, Mirkin, Journal of Classification, 2010) over cluster recovery

Method	Acronym
Calinski and Harabasz index	CH
Hartigan rule	HK
Gap statistic	GS
Jump statistic	JS
Silhouette width	SW
Consensus distribution area	CD
Average distance between partitions	DD
Square error iK-Means	LS
Absolute error iK-Means	LM

Extending K-Means model 1: Feature weighting

K-Means is defenseless against noise features: all have equal weights in Euclidean distances

Extension of K-means iteration steps from two to three using Minkowski distances with feature rescale factors (weights):

(i) centers update
(ii) clusters update
(iii) feature weight update

Amorim & Mirkin (2012) record:

5 errors on Iris (with cluster-specific feature weights)

Extending K-Means I: MWK-Means results

Alternating Min W_p(S,c,w) [Amorim, Mirkin, 2012]

1. Weights may be cluster-specific. They reflect the level of dispersion of features *v* within clusters.

2. In experiments, cluster recovery much depends on the *p* value which is data dependent. At a right *p*, MWK-Means beats all other k-means versions.

3. i-MWK-Means implementing sequential anomalous clusters works well at medium data sizes.

Extending Anomalous cluster to temperature map data (Nascimento, Caska, Mirkin 2015)

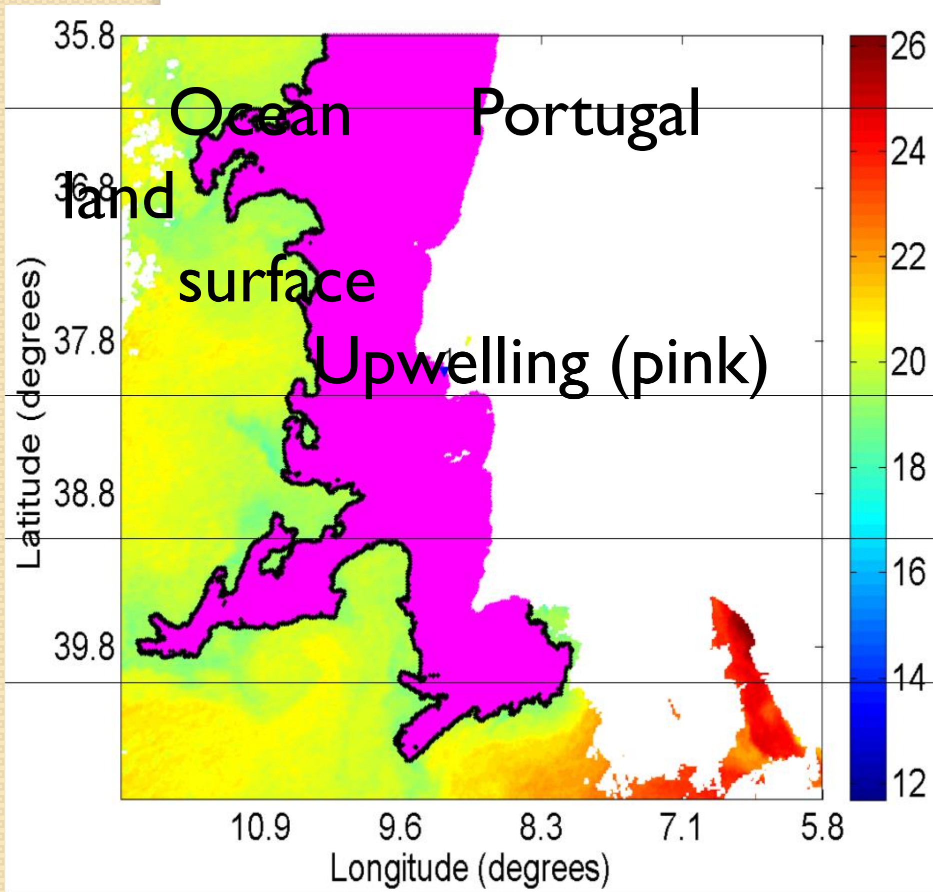
Given a temperature map

data over pixels i,

Find center **c** and

cluster of pixels ***S*** to

maximize



$$\mathbf{g}(S,\mathbf{c}) = |S| < \mathbf{c}, \mathbf{c} >$$

Extending Anomalous cluster to temperature map data, (NCM 2015),2

- Given a temperature x map data over pixels i

  find center **c** and cluster of pixels ***S*** to maximize

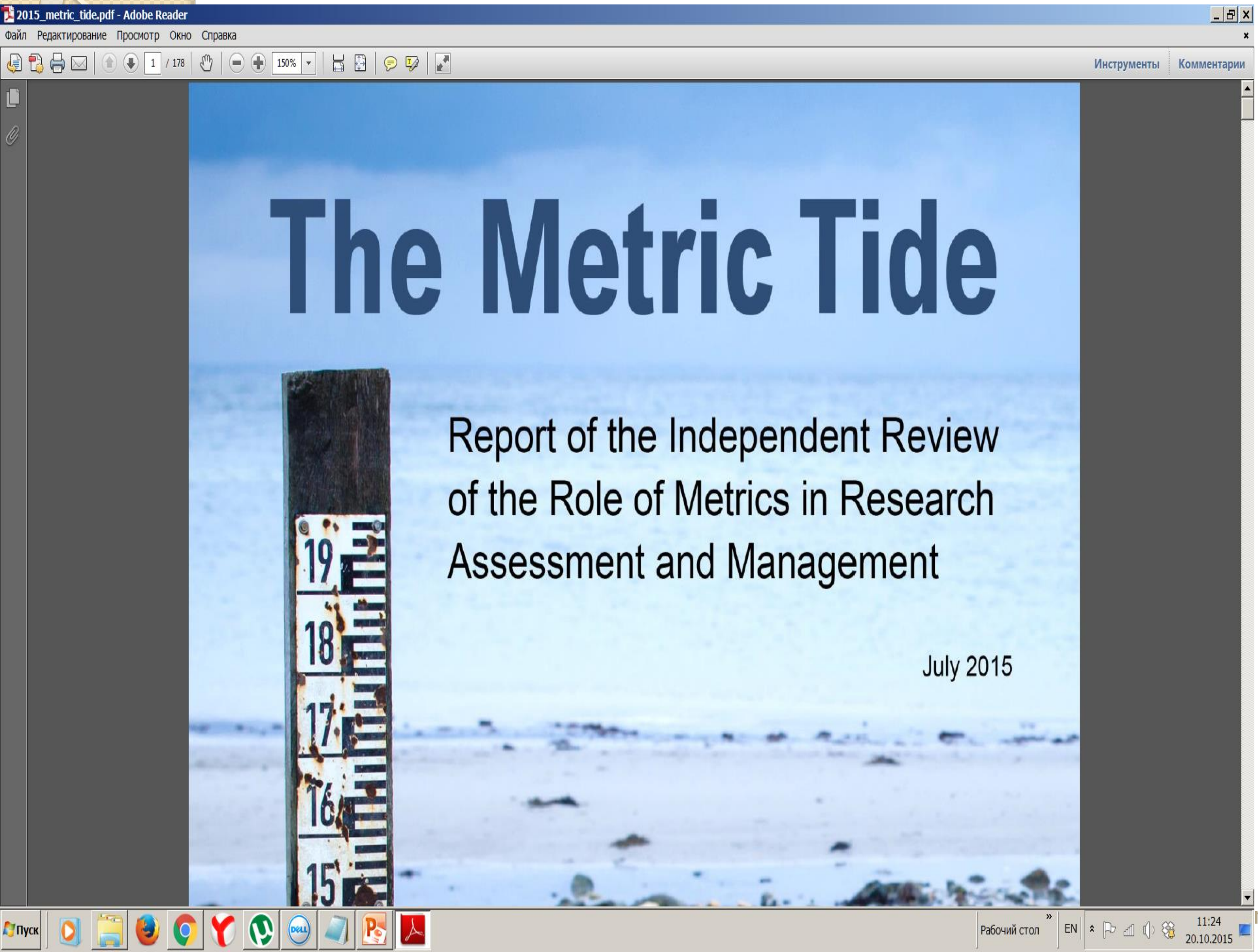
$$g(S,c)=|S| < \mathbf{c}, \mathbf{c} >$$

- Using a window size as a smoothing/restricting parameter

- **One by one adding/removing pixels is a Seed-Growing segment finding algorithm (with no other parameters, unlike the major seed-grwing algorithms)**

Summarization by ranking: Metric Tide in research assessment

Cover of report by a UK REF commission (July 2015)



Conclusions:

….

- Currently no automatic impact scoring is possible
- Financing projects on research impact should be opened in UK

…..

DORA Initiative
San Francisco Declaration on Research Assessment

Impact is not impact factor only

Citation makes use of publication activities, yet a comprehensive assessment should take into account other researcher's products as well

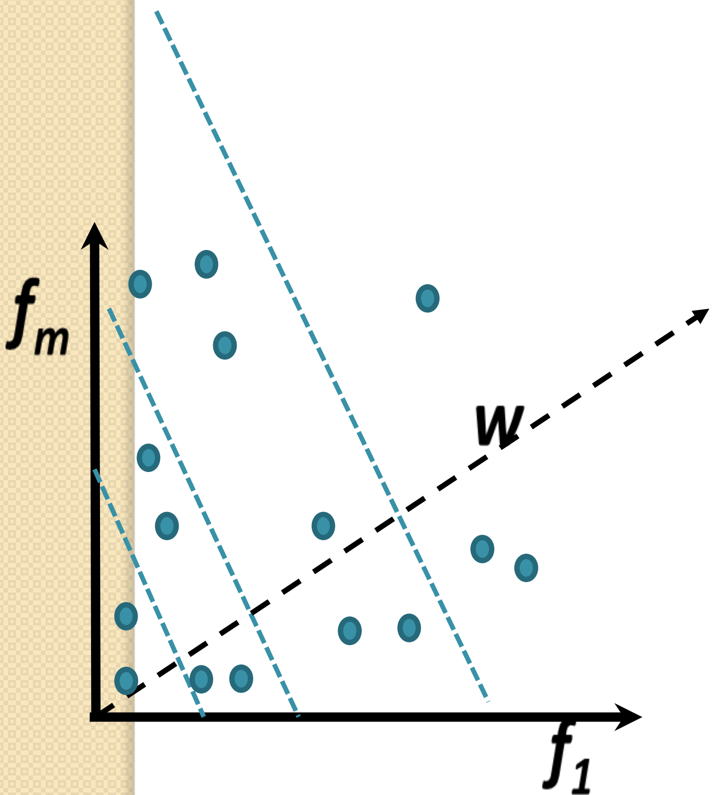
Research ranking: my contribution

- 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
- Essay on developing a system for impact assessment

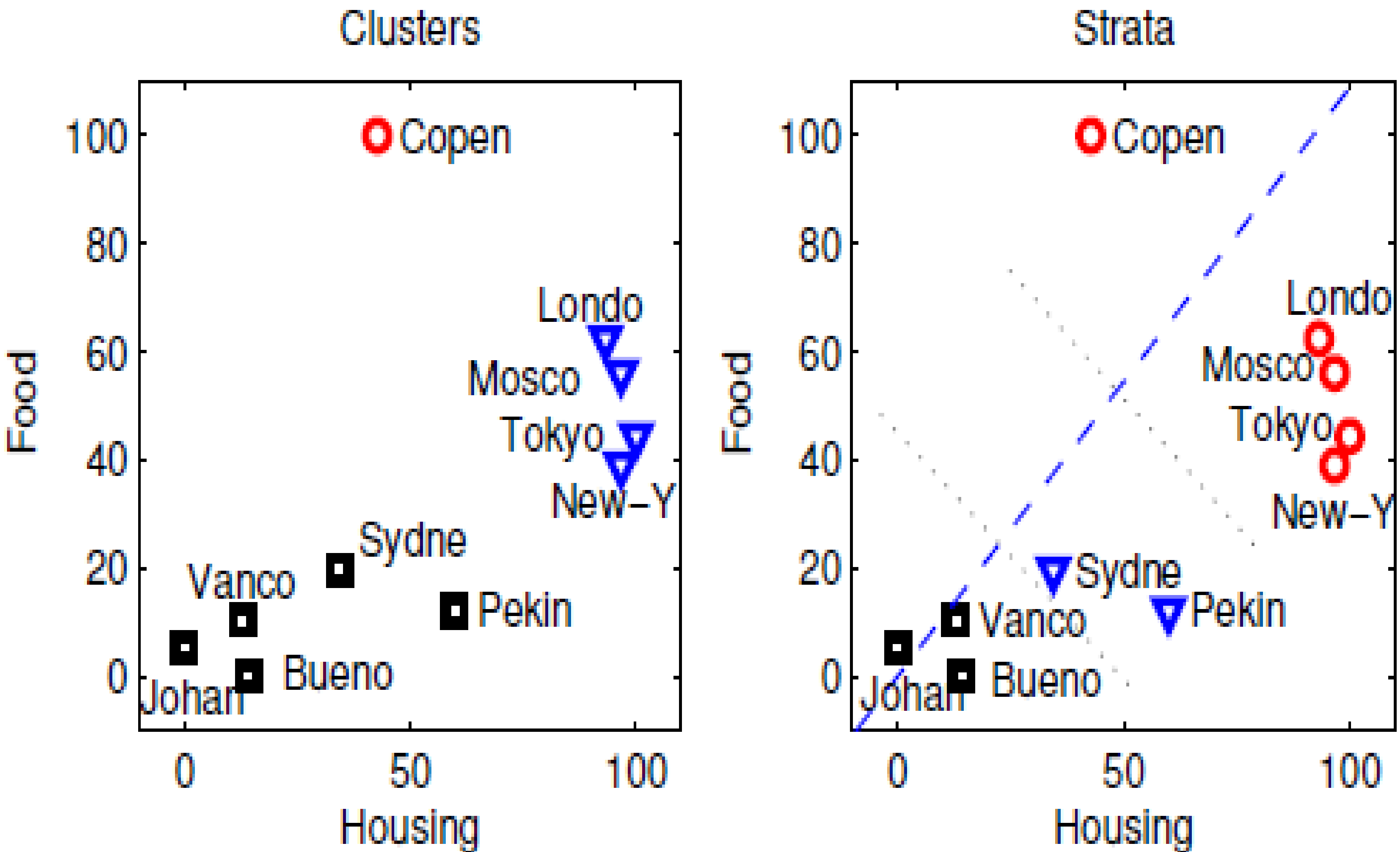
Method 1: **Convex combination of criteria**

- Input: set of criteria $f_1, f_2,..., f_m$ over an entity set I
- Output: set of weights $w=(w_1, w_2,..., w_m)$ 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:**

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

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

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

Method 1: Linstrat - unsupervised K stratification **Minimize alternatingly:**

- Initialise w randomly

- Given weights w, find K centers c_k and strata S_k

- Given c_k and strata S_k, find w

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

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

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

Ranking Method 1: Testing

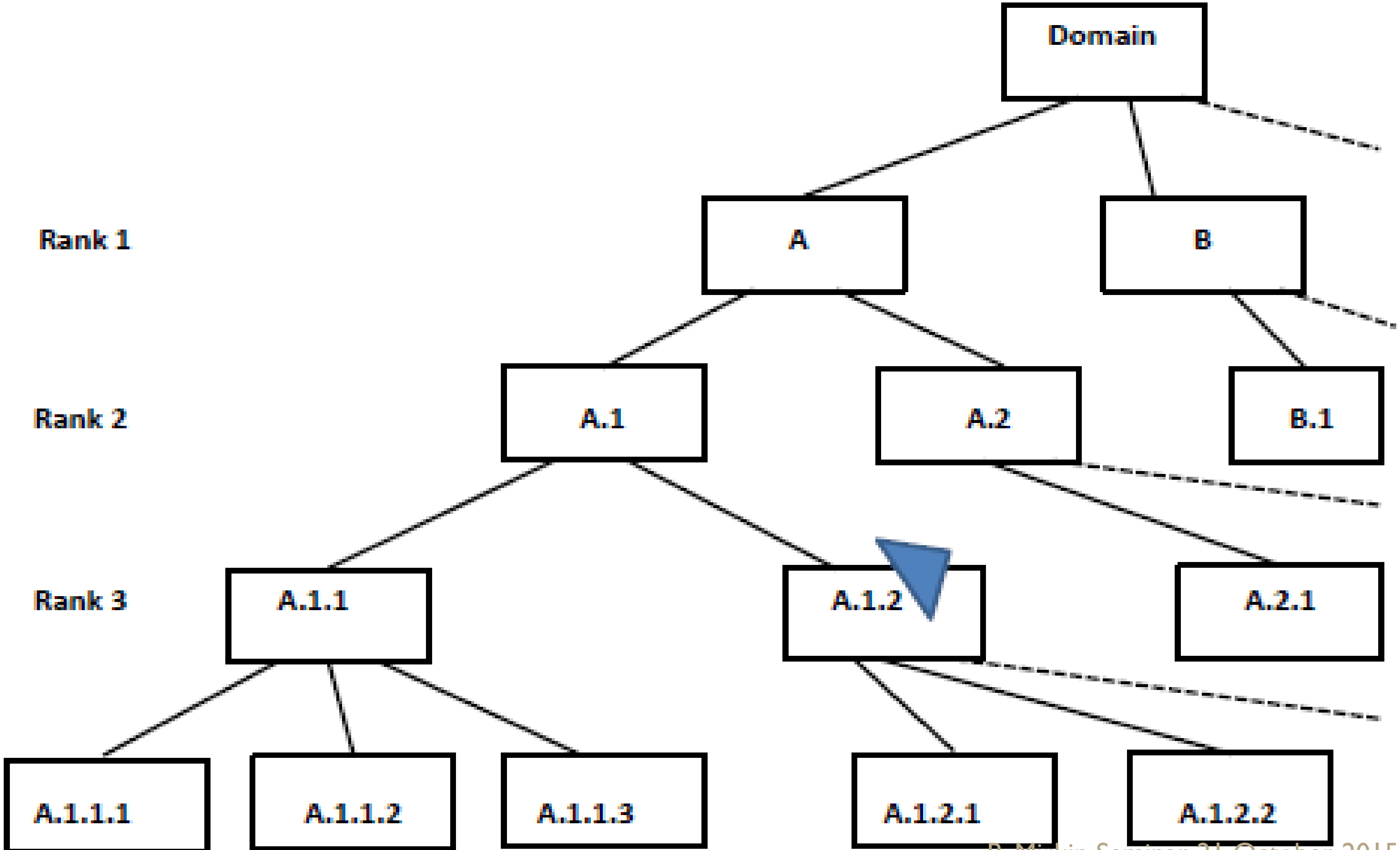
Linstrat - Method for unsupervised *K* stratification:

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.)**

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



Taxonomy for "Data analysis" from ACM CCS 2012, 1

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

Taxonomy for "Data analysis" from **ACM CCS 2012, 2**

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

Ranking: Experimental computation

- 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

Ranks of 4-6 results by scientists from our samplert of sample of scientists: anonymous

S1	5,5,4	3,88	73
S2	4,4,4,4,4	3,50	100
S3	5,5,5,5,5	4,50	29
S4	5,5,5,5,4,5	3,90	71
S5: Boris Mirkin	5,5,5,5,5	4,50	29
S6	4,5,5,4,5	3,77	81
S7	5,5	4,80	7
S8	5,5,5,5,5	4,50	29
S9	5,5,5,5,5	4,50	29
S10	5,5	4,80	7
S11	4,5,5,5,5	3,86	74
S12	5,4,6,5,5,5	3,86	74
S13	5,4,5,5,5	3,86	74
S19: Panos Pardalos	5,5,6,5	4,69	15

Results: Linstrat aggregate citation at 3 strata

CITATION =

0.5*Total_Cit+0.5*Cit_10+0.0*Hirsh

Results: Linstrat aggregate merit at 3 strata

MERIT =

0.22*#PhD+0.10*Conf_Ch+0.69*E/AssocEJ

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

	TaxR	Cit	Merit
TaxR		-.12	-.04
Citation			**.31**
Merit			

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

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

Results: Aggregate criterion

Panoramic =

0.80*TaxRank + 0.04*Citation + 0.16*Merit

Researcher’s products in 5 areas, 1

1 Research and presentation of results

- Publications
- Presentations
- Funded and unfunded projects

2 Participation in Science functioning

- Journal editing
- Running research meetings
- Refereeing
- Research cooperation
- Research societies

Researcher's products in 5 areas, 2

3 Teaching

- **knowledge**
  - Lectures
  - Seminars
  - Projects
  - Consultation
  - Assessments and exams
  - Textbooks
- **knowledge discovery**
  - PhD Students
  - Research students

Researcher's products in 5 areas, 3

4 Technology innovations

- Programs
- Services
- Patents
- Industrial consultations

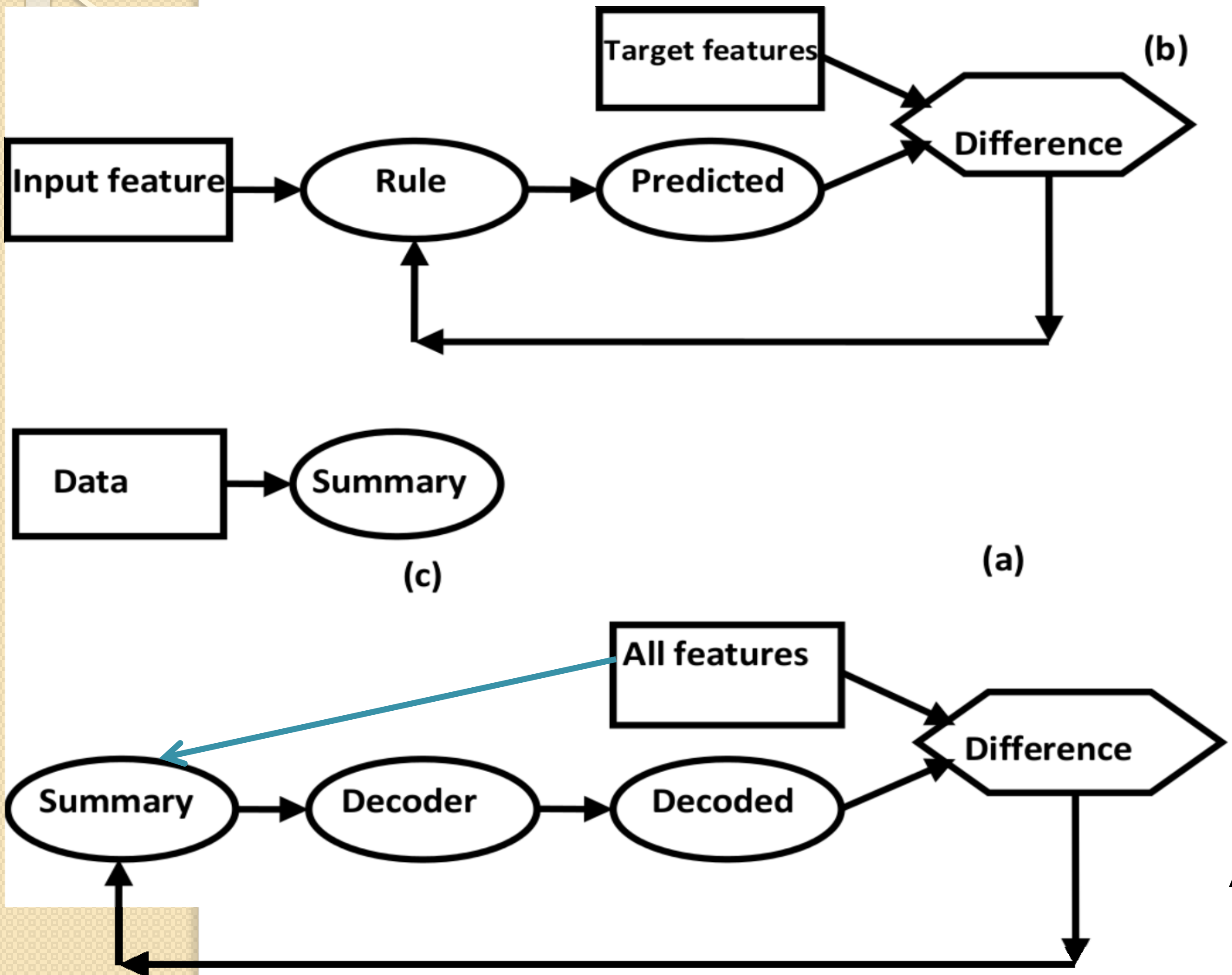
5 Societal interactions

- Popular books
- Articles
- Blogs
- Networks

Conclusion

- Summarization versus learning
- Extension to Big Data
- A ranking project in Systems Analysis

Data summarization versus prediction



Prediction:
building rule
Target = F(Input)

Summarization:
Conventional view

Summarization:
Data recovery view - All features are target

Data recovery summarization: growth points

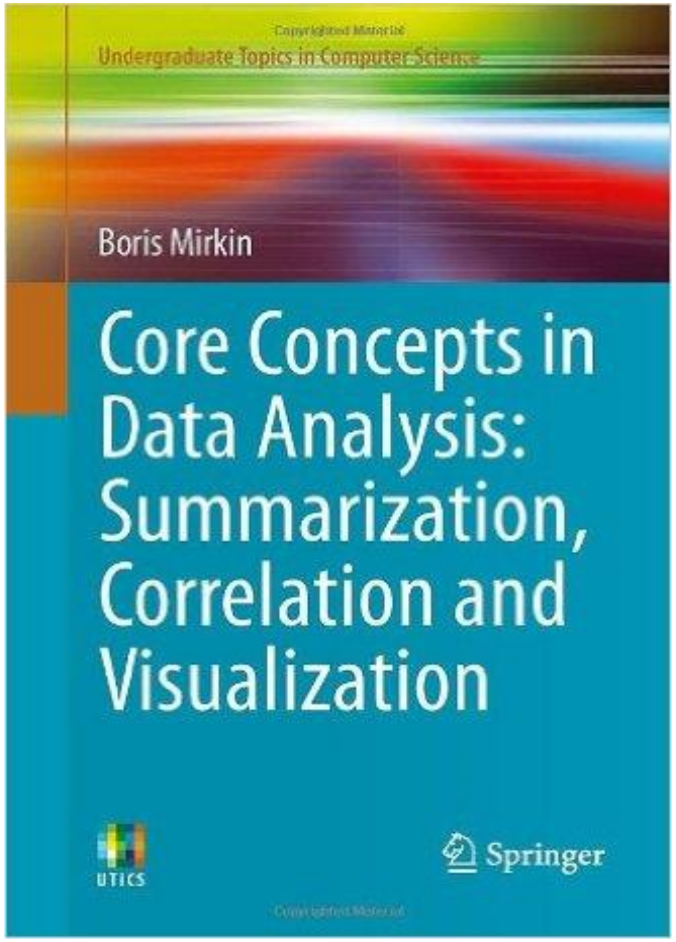
- Model

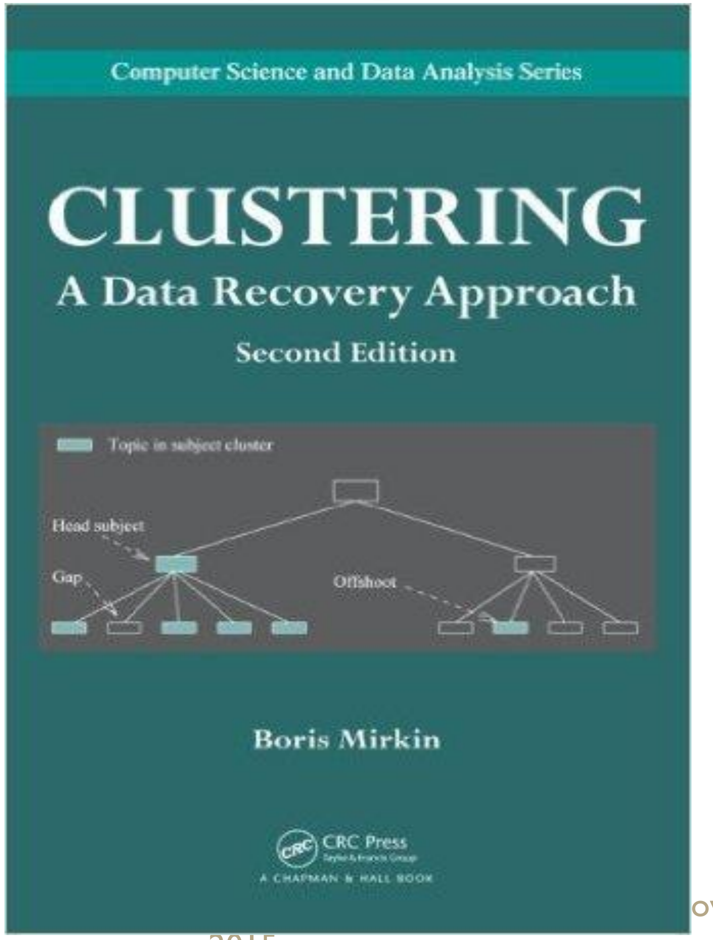
Data = Decoded(Model) + Residual

- More applications including in organization analysis
- Non-multiplicative decoders
- Different fitting criteria (advantages of using L1 and other non-linear criteria)
- Effects of noise added (a very new development)

Boris Mirkin's work on data recovery in clustering:

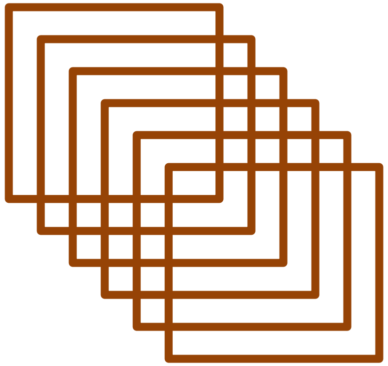
Text 2011 **Monograph 2012**





Extension to Big Data: example

- **Parallel computation for K-Means**



No data,
centers only

Zillion of local computers:

- **Keep local data**
- **Update clusters locally**
- **Compute local centers**

Central computer:

Updates centers by aggregating local centers

Can be done with MapReduce Technology:

(data, key)- format — **MAP**

data-format — **REDUCE**

Developing reasonable metrics for assessment of research impact 1

- Timeliness: Globalisation – science becomes a mass occupation while many others do involve research (banks, retailers, e-commerce, …)

- Stages of a project in assessment of systems analysis research
  - Defining and maintaining a comprehensive taxonomy of Systems Analysis domain

    (integrating 75 definitions)

Developing reasonable metrics for assessment of research impact, 2

- Stages of a project (continued):
  - Defining a scheme for research products and metrics for assessment of them, as well as committees to do the mapping
  - Maintaining a nomenclature of scientists and their metrics data
  - A working group on methods for integration of metrics and methods for automating extraction of metrics from internet data

Potential outcome, I

- In substance:
  - Developing a system for assessment of research impact
  - Maintaining the system
  - Taxonomy of the Domain
  - Cataloguing research results and researchers
  - Forum for discussing taxonomy and results

Potential outcome,2

- In methods:
  - Enhancing the concept of Taxonomy
  - Methods for relating research reports and taxonomy
  - Methods for taxonomy building using research reports
  - Methods for mapping research results to taxonomy
  - Ranking impact of results
  - Methods for combining rankings