Iterative Extraction (ITEX SEFIT (1990)): Extensions of Principal Component Analysis

Boris Mirkin

School of Computer Science

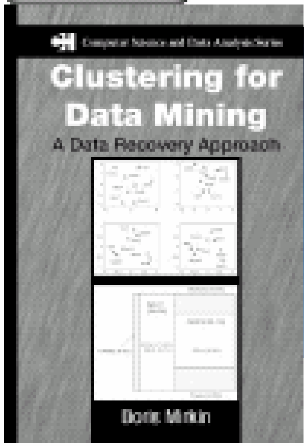
Birkbeck College

University of London

Special Issue: The Computer Journal, **Profiling Expertise and Behaviour**: Deadline 15 Nov. 2006

C5343 FL

NEW! *Get more than just ad hoc methods—discover a sound theoretical framework!*



Clustering for Data Mining
A Data Recovery Approach

Boris Mirkin
School of Computer Science and Information Systems
Birkbeck College, University of London, UK

- **WHAT IS CLUSTERING**; **WHAT IS DATA**
- **K-MEANS CLUSTERING**: Conventional K-Means; **Initialization of K-Means; Intelligent K-Means; Interpretation Aids**
- **WARD HIERARCHICAL CLUSTERING**: Agglomeration; **Divisive Clustering with Ward Criterion; Extensions of Ward Clustering**
- **DATA RECOVERY MODELS**: Statistics Modelling as Data Recovery;
  Data Recovery Model for K-Means; **for Ward; Extensions to Other Data Types; One-by-One Clustering**
- **DIFFERENT CLUSTERING APPROACHES**: Extensions of K-Means; Graph-Theoretic Approaches; **Conceptual Description of Clusters**
- **GENERAL ISSUES**: **Feature** Selection and **Extraction**; **Similarity on Subsets and Partitions**; Validity and Reliability

Talk's outline

- Data model and Pythagorean decomposition
- Principal component analysis as a data model
- Extension of PCA to clustering and K-Means
- Principal cluster analysis for clustering
- General ITEX strategy
- Examples of ITEX: hierarchical clustering, additive clustering, box clustering, contingency data aggregation

Pythagorean framework for data analysis methods

- Type of Data
  - Similarity
  - Temporal
  - Entity-to-feature
  - Co-occurrence
- Type of Model
  - Regression
  - Principal components
  - Clusters

Model:

Data = Model_Data + Residual

Pythagoras:

$$\mathbf{Data^2 = Model\_Data^2 + Residual^2}$$

Pearson's PCA: measuring talent

- **Given**: marks x_{iv} (i – student, v – subject)
- **Find**: talent score z_i and subject loading c_v
- $x_{iv} = c_v z_i + e_{iv}$ $$L^2 = \sum_{i \in I} \sum_{v \in V} e^2_{iv} = \sum_{i \in I} \sum_{v \in V} (x_{iv} - c_v z_i)^2$$
- **Solution**: $X^T z^* = \mu c^*$, $Xc^* = \mu z^*$, max μ
- **Properties:**
- P1: z^* **is lc of** $\boldsymbol{X}$ **columns**
- P2: $T(X) = \mu^2 + L^2$, $T(X) = \Sigma\, x_{iv}^2$ – *data scatter*

PCA as a data model

Data Model:

$$y_{iv} = \sum_{k=1}^{K} c_{kv} z_{ik} + e_{iv},$$

minimising L^2 over c and z

Properties:

- *[Z, M, C]=svd(Y),*
  - Thus z and c are lc of X
  - Can be done sequentially, one by one
- $T(Y) = \mu_1{}^2 + \mu_2{}^2 + \ldots + \mu_K{}^2 + L^2$

Extension of PCA to clustering

$$y_{iv} = \sum_{k=1}^{K} c_{kv} z_{ik} + \varepsilon_{iv},$$

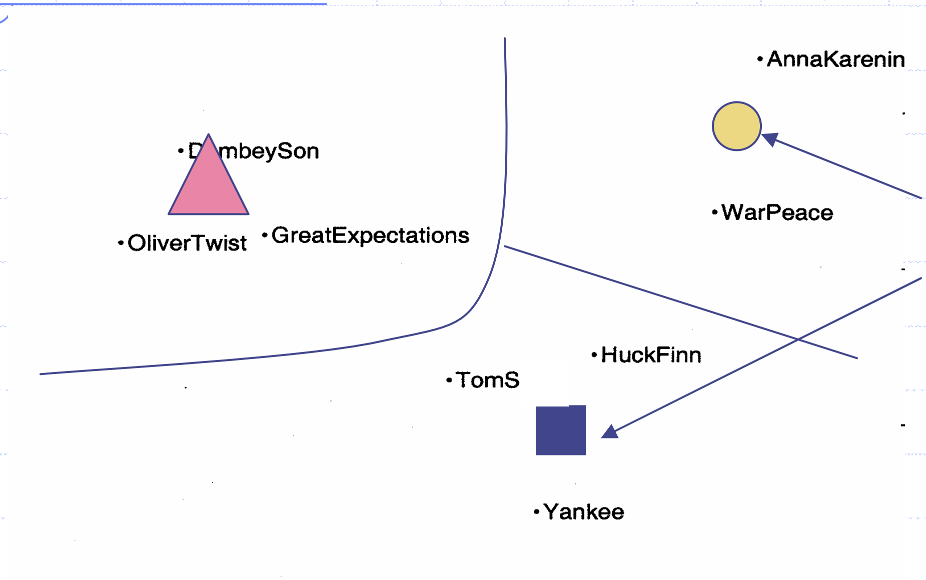
$$\sum_{i=1}^{N}\sum_{v=1}^{V} y_{iv}^2 = \sum_{v=1}^{V}\sum_{k=1}^{K} c_{kv}^2 N_k + \sum_{k=1}^{K}\sum_{i\in S_k}\sum_{v=1}^{V} (y_{iv} - c_{kv})^2$$

y – data entry, z – 1/0 membership

c **- cluster centroid,** ***N*** **– cardinality**

i **- entity,** ***v*** **- feature /category,** ***k*** **- cluster**

Representing a partition



Cluster k:

Centroid

c_{kv} (v - feature)

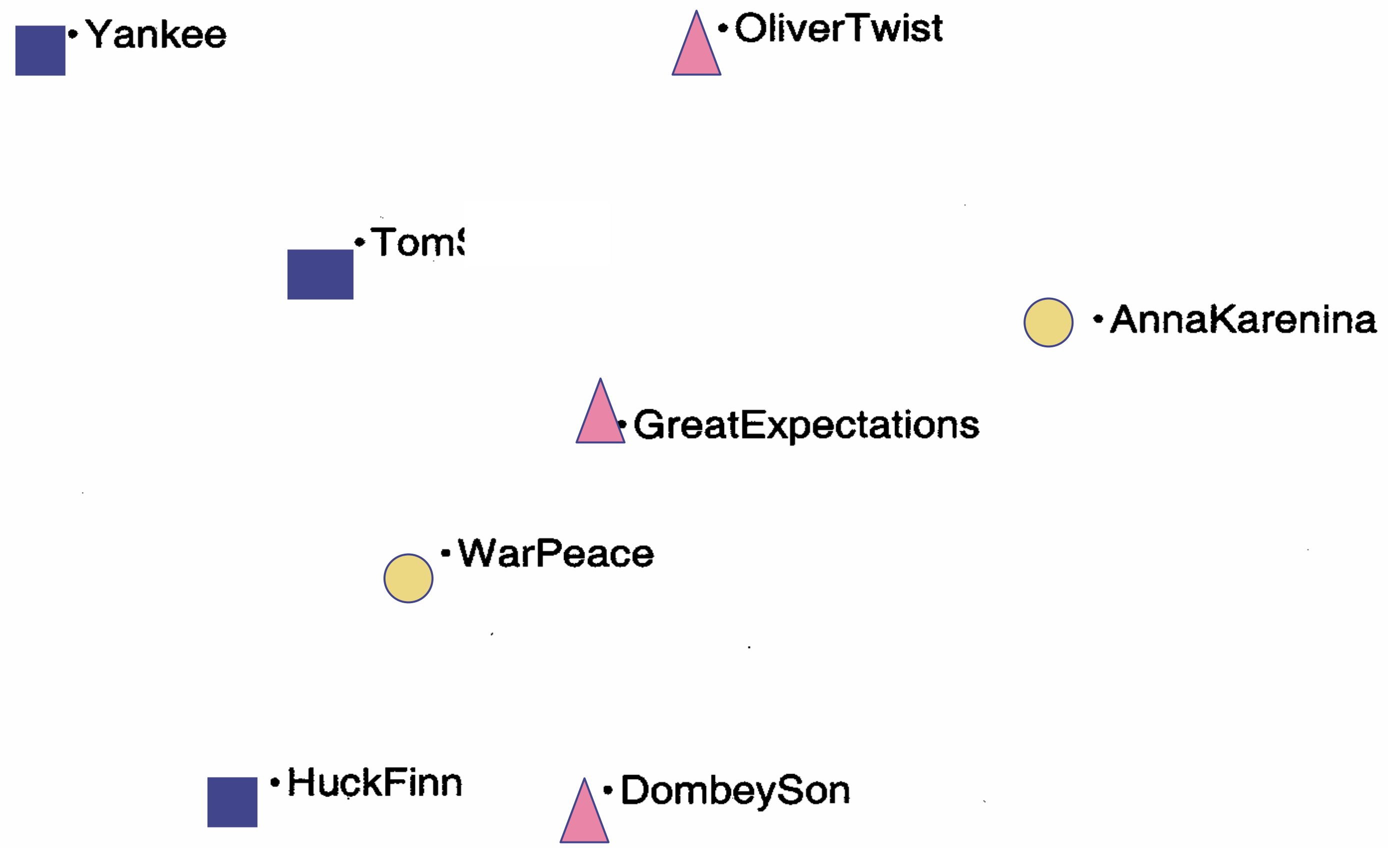
Binary 1/0 membership

z_{ik} (i - entity)

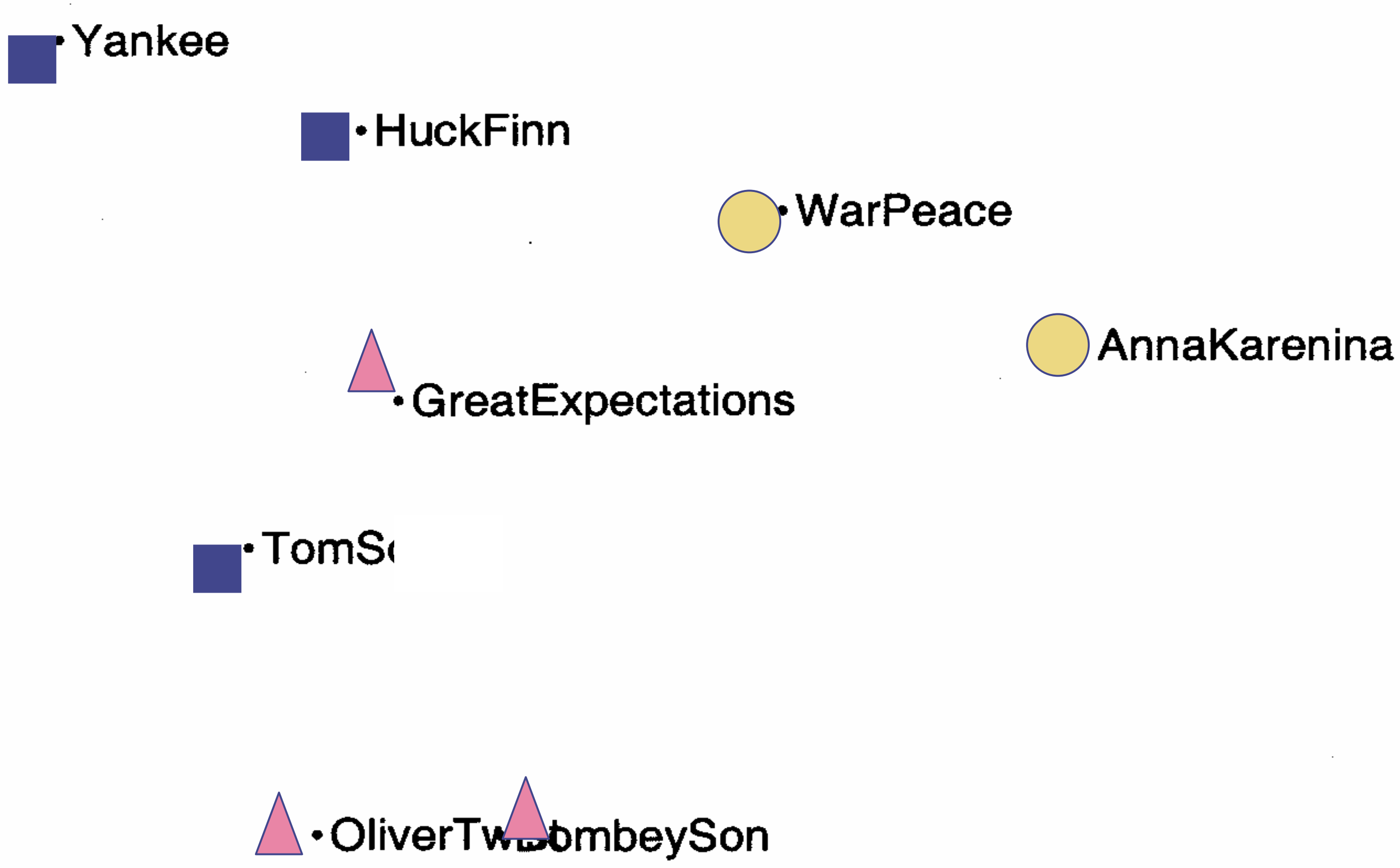
Standardisation of features

- $Y_{ik} = (X_{ik} - A_k)/B_k$
  - X - original data
  - Y – standardised data
  - i – entities
  - k – features
  - A_k – shift of the origin, typically, the **average**
  - B_k – rescaling factor, traditionally the **standard deviation**, but **range** seems better in clustering

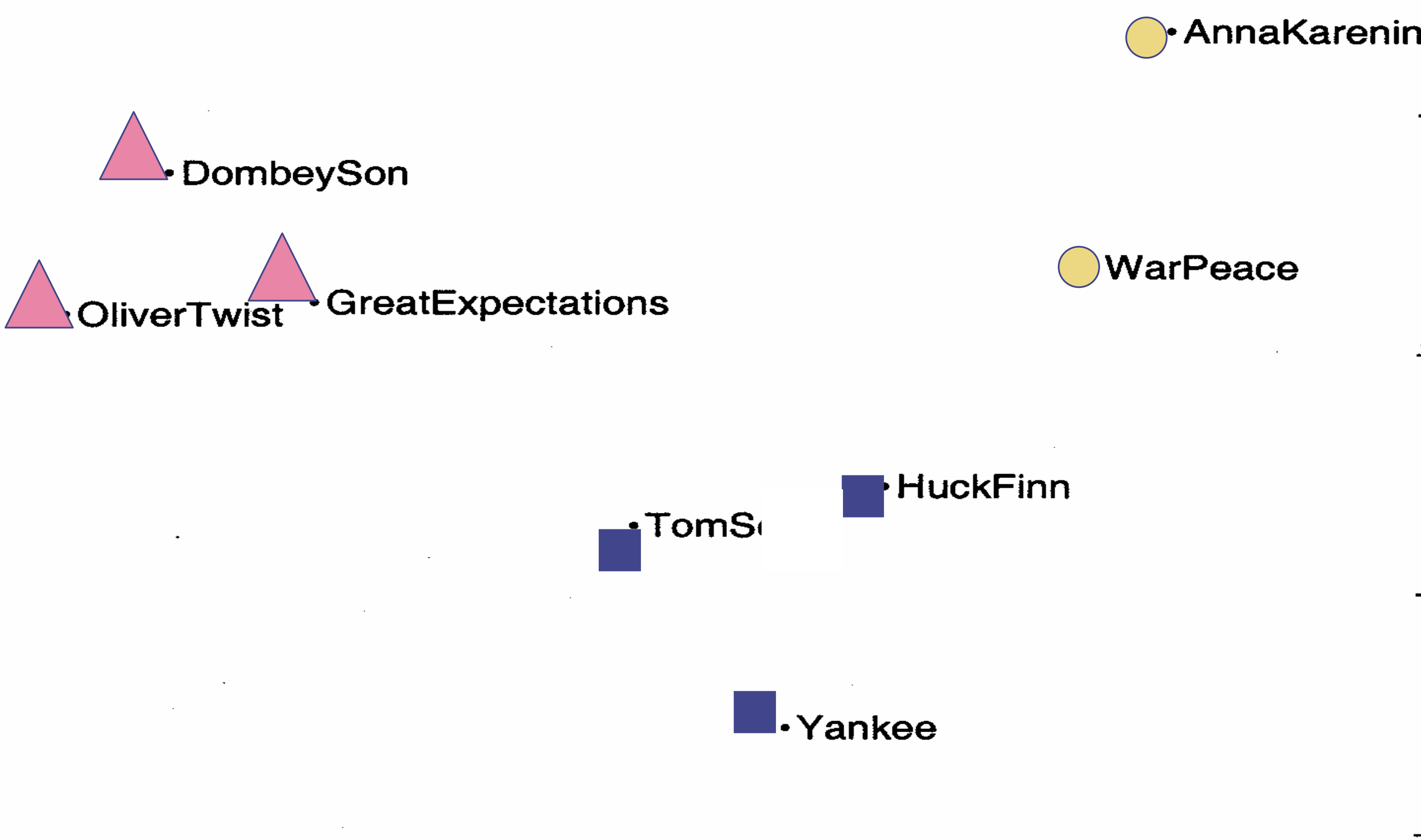
No standardisation



Z-scoring (scaling by std)



Standardising by range & weight



Fitting the model with Straight K-Means Partitioning

Start:

* Presenting cases as multidimensional points
* Putting initial centroids (seeds)

Reiterated until no change:

* Collecting points into clusters around centroids
* Recalculating centroids as cluster prototypes

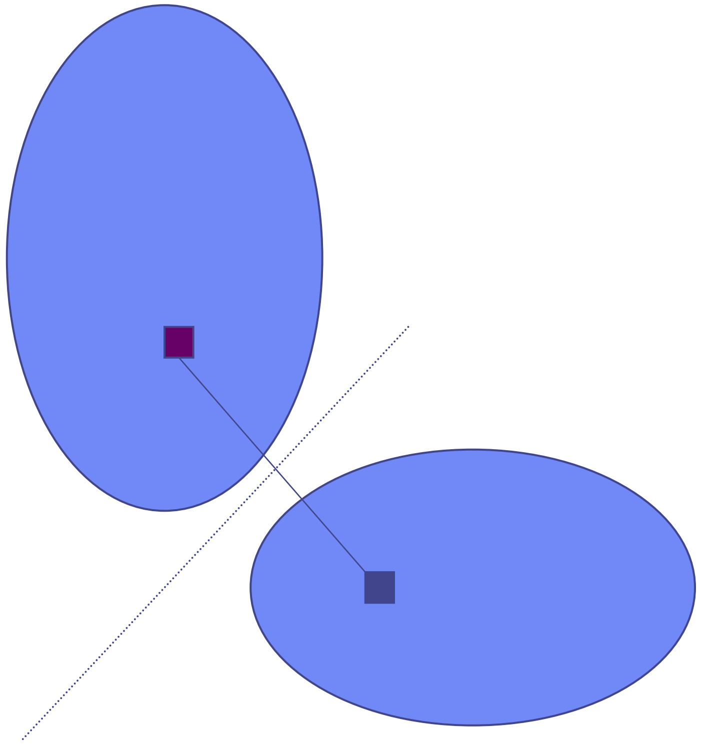
Advantages of K-Means

- Conventional:
  - **Models typology building**
  - **Computationally effective**
  - **Can be incremental, `on-line’**
- Unconventional:
  - **Associates feature salience with feature scales and correlation/association**
  - **Applicable to mixed scale data**

Drawbacks of K-Means

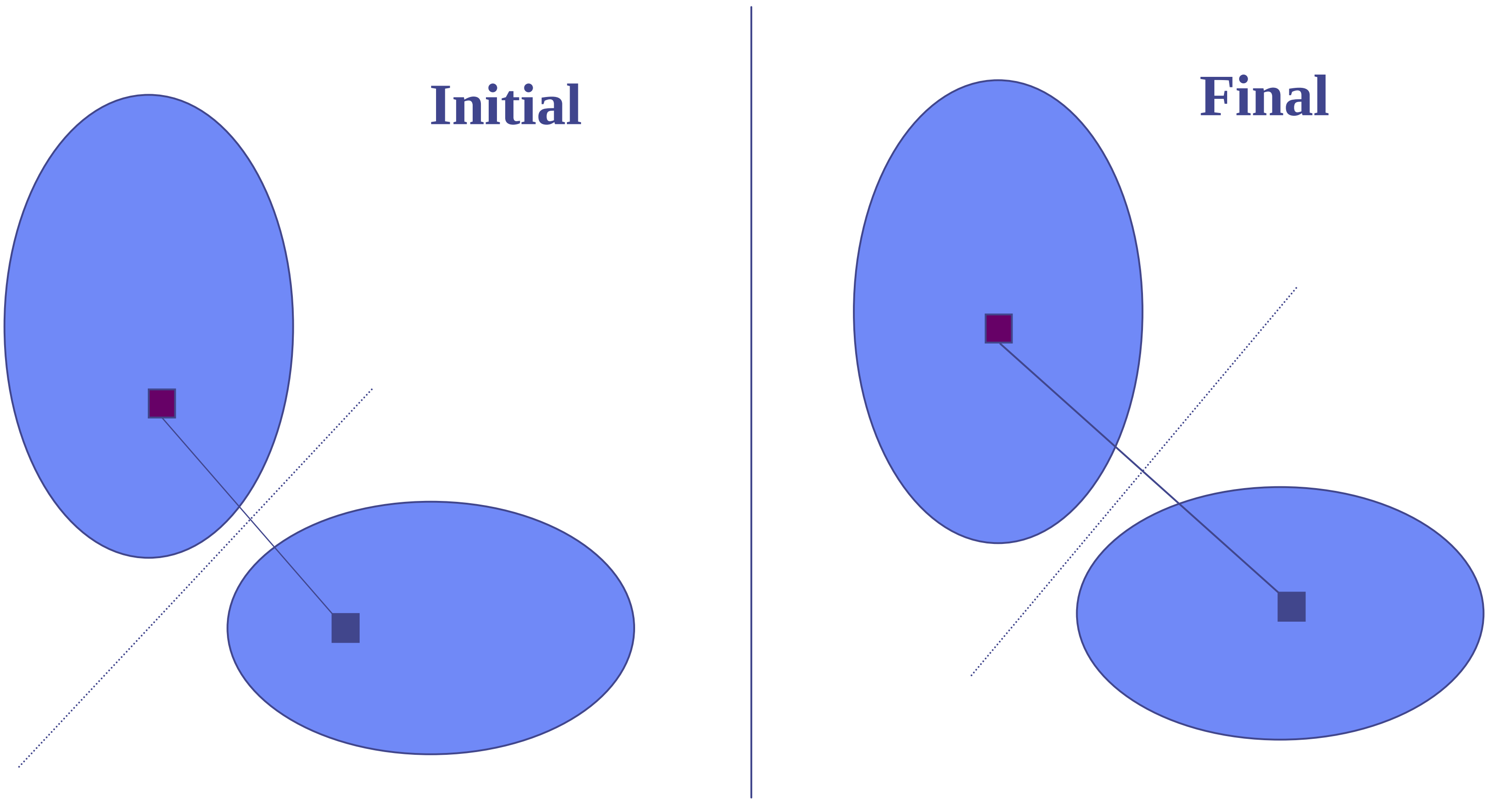
- No advice on:
  - Data pre-processing
  - Number of clusters
  - Initial setting
- Instability of results
- Criterion can be inadequate
- Insufficient interpretation aids

Initial Centroids: Correct

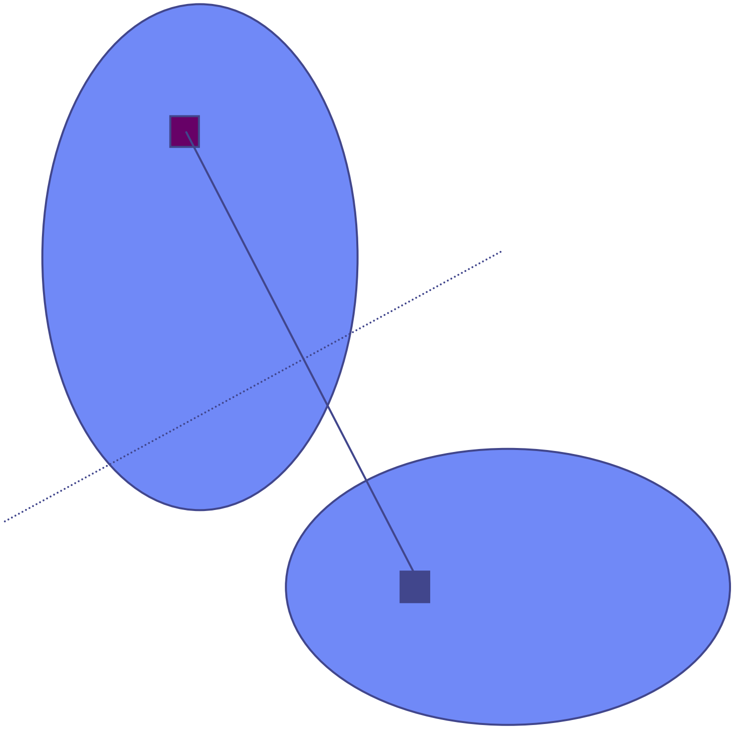


Two cluster case

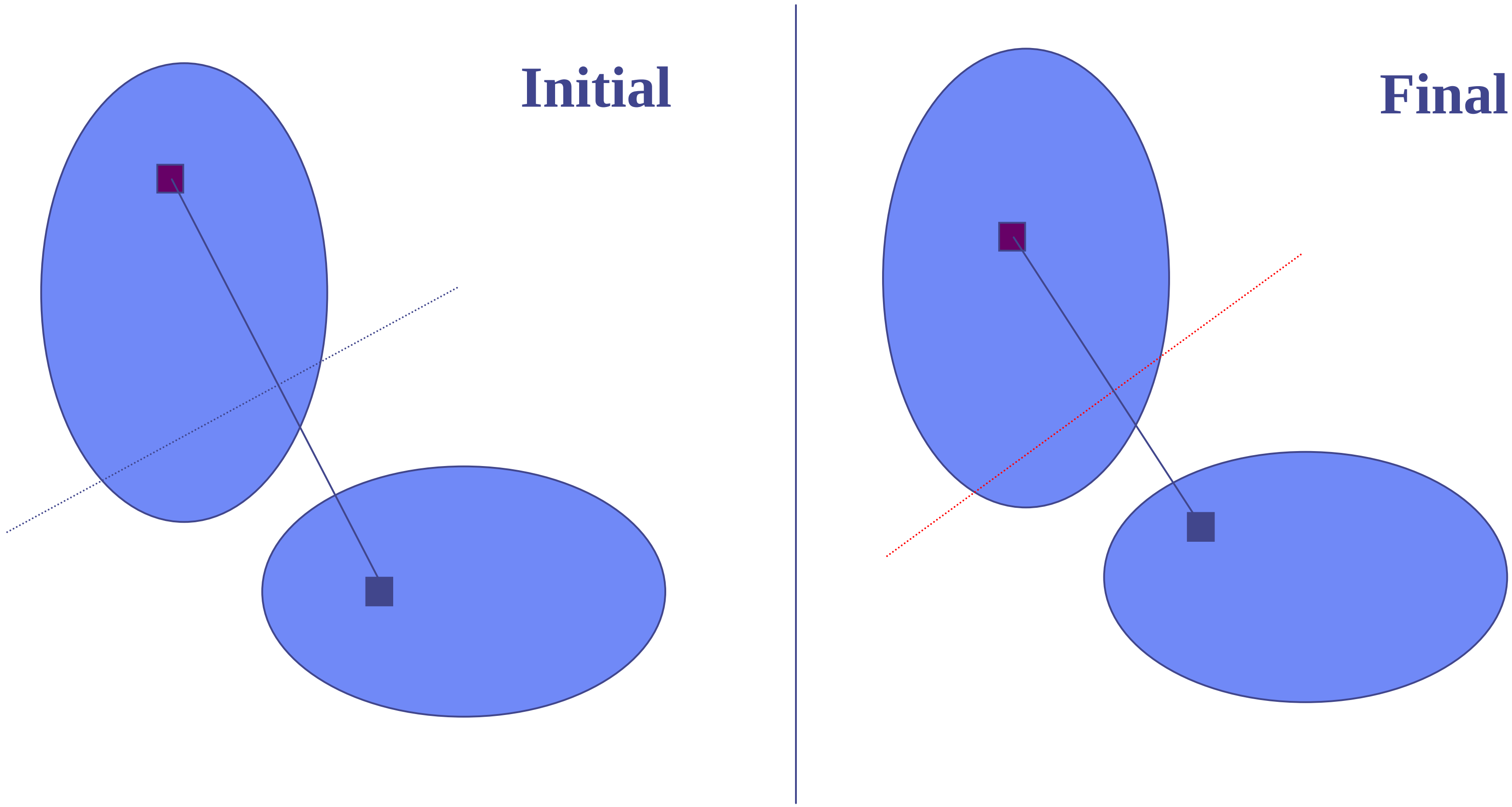
Initial Centroids: Correct



Different Initial Centroids



Different Initial Centroids: Wrong, even though in different clusters



Principal Cluster Analysis: One cluster at a time

$$y_{iv} = c_v z_i + e_{iv},$$

where $z_i = 1$ if $i \in S$, $z_i = 0$ if $i \notin S$

With Euclidean distance squared

$$\sum_{i=1}^{N}\sum_{v=1}^{V} y_{iv}^2 = \sum_{v=1}^{V} c_{Sv}^2 N_S + \sum_{i\in S}\sum_{v=1}^{V}(y_{iv} - c_{Sv})^2$$

$$\sum_{i=1}^{N} d(i,0) = d(c_S,0)N_S + \sum_{i\in S} d(i,c_S)$$

c_S must be **anomalous**, that is, interesting

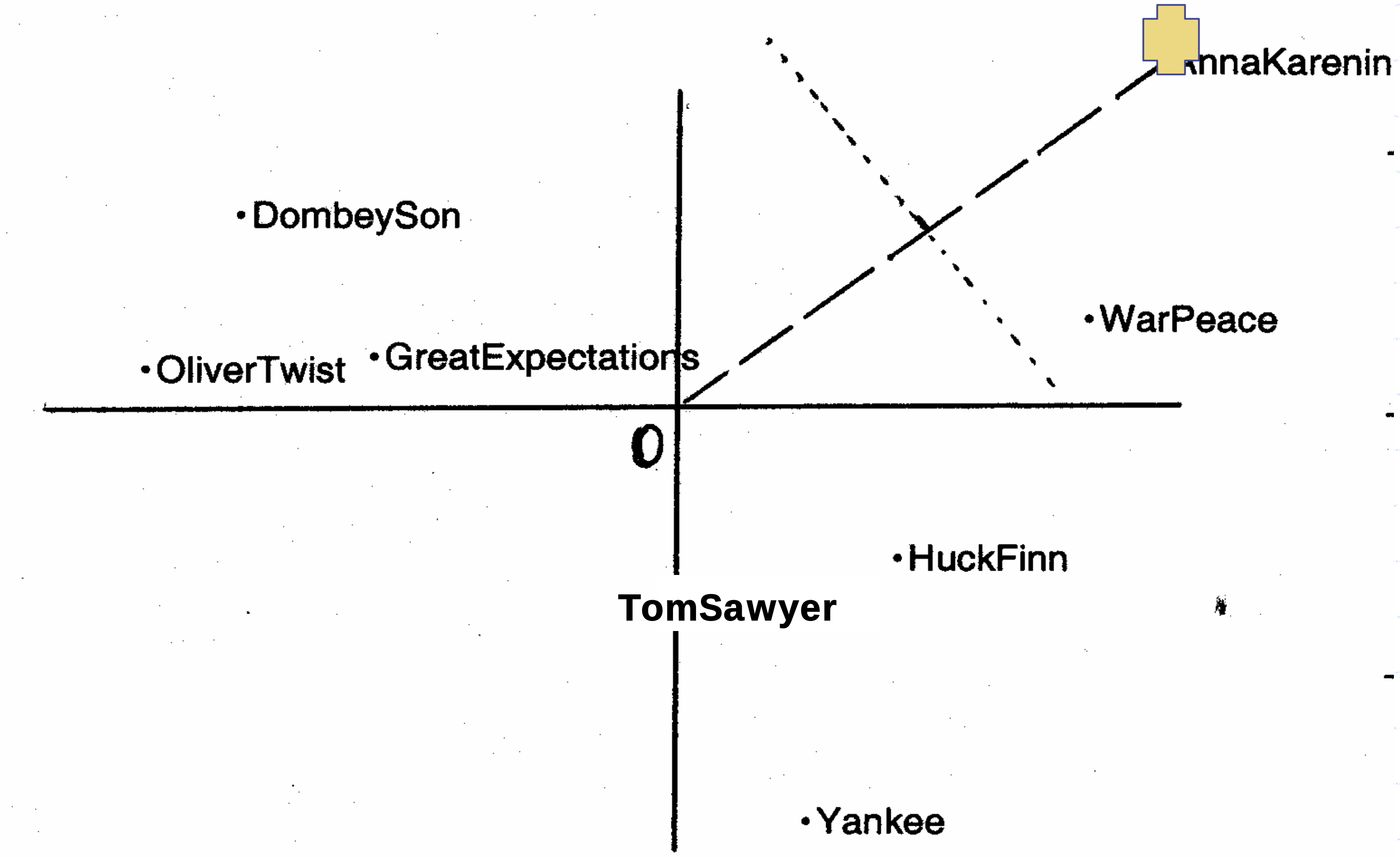
Principal Cluster Analysis: One cluster at a time

$$\sum_{i=1}^{N}\sum_{v=1}^{V} y_{iv}^2 = \sum_{v=1}^{V} c_{Sv}^2 N_S + \sum_{i\in S}\sum_{v=1}^{V}(y_{iv}-c_{Sv})^2$$

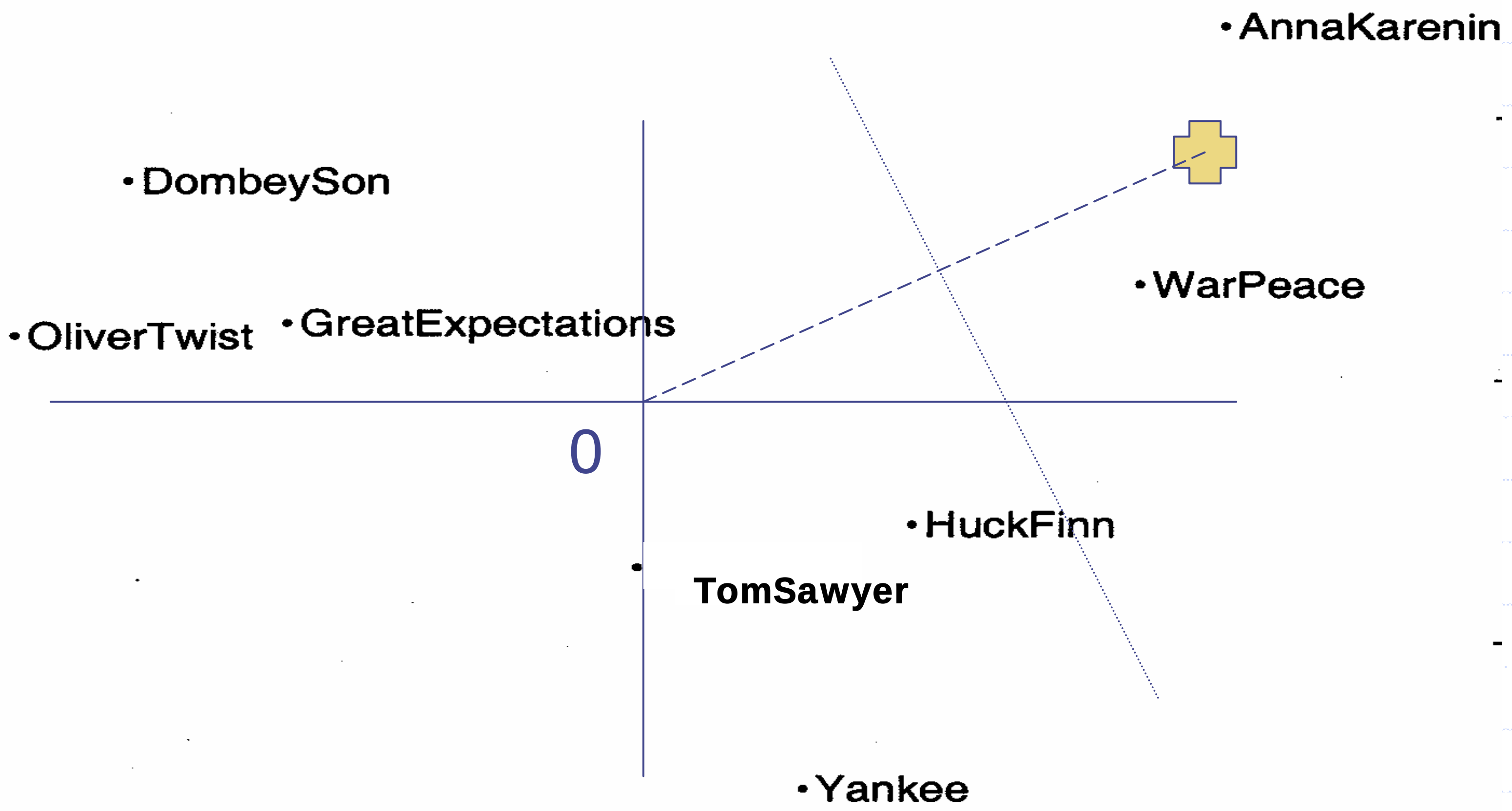
- Or, with Euclidean distance squared *d(,)*

$$\sum_{i=1}^{N} d(i,0) = d(c_S,0)N_S + \sum_{i\in S} d(i,c_S)$$

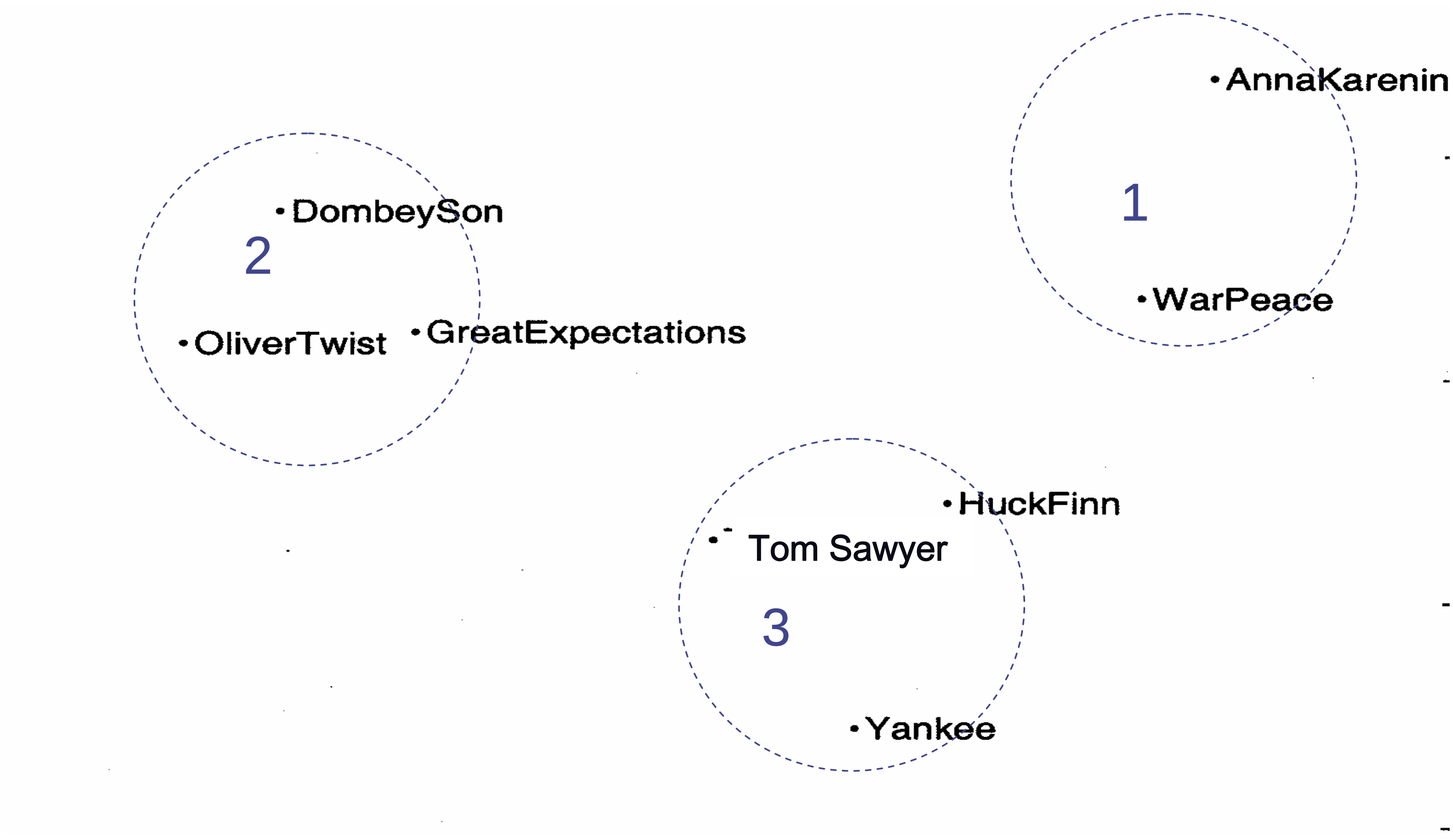
Initial setting with Anomalous Pattern (AP) clustering



AP clustering: Iterate



iK-Means with Anomalous Single Clusters



Decomposing Data scatter

- The sum of standardised entries squired

$$D^2 = \sum_{i=1}^{N} \sum_{v=1}^{V} y_{iv}^2$$

- The sum of **contributions of features**

- Proportional to the summary variance

Contribution of a feature *F* to a partition

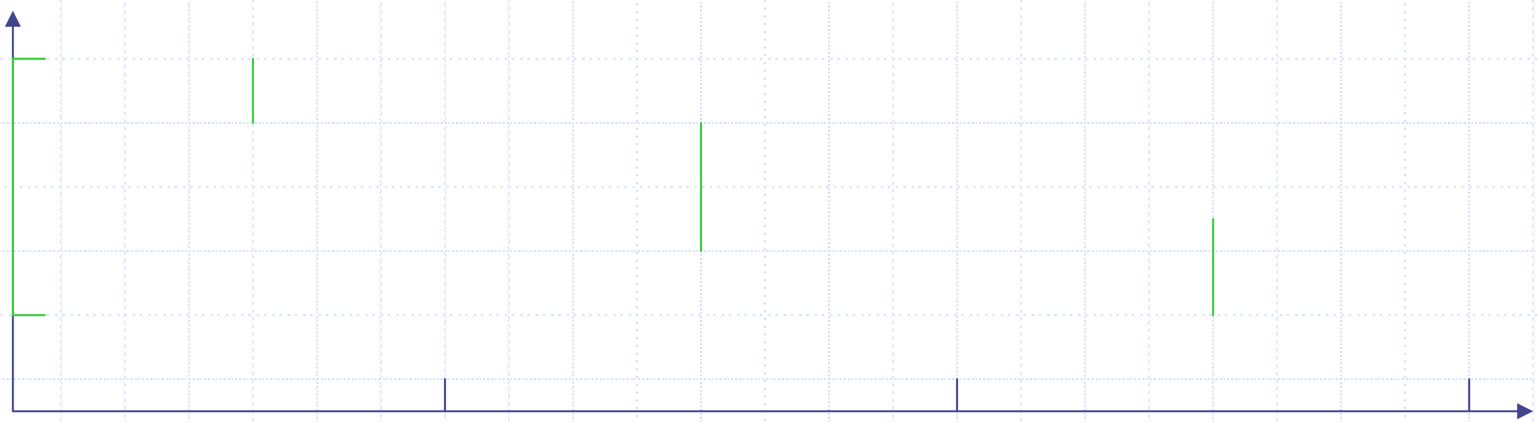
$$Contrib(F) = \sum_{v \in F} \sum_{k=1}^{K} c_{kv}^2 N_k$$

- Proportional to
  - correlation ratio η^2 if F is quantitative
  - a contingency coefficient if F is nominal
    - Pearson chi-square (Poisson normalised)
    - Goodman-Kruskal tau-b (Range normalised)

Contribution of a quantitative feature to a partition

$$N\eta^2 == N\sum_{k=1}^{K}(\sigma^2 - p_k\sigma_k^2)/\sigma^2$$

- Proportional to
  - correlation ratio η^2 if F is quantitative



Contribution of a nominal feature to a partition

$$NX^2 = N\sum_{k=1}^{K}(p_{ij} - p_i p_j)^2 / p_i B_j^2$$

- Proportional to a contingency coefficient
  - Pearson chi-square (Poisson normalised)

$$B_j = \sqrt{p_j}$$

  - Goodman-Kruskal tau-b (Range normalised)
    - $B_j=1$

Pythagorean Decomposition of data scatter for interpretation

ised as in Table 2.6 according to author based clusters.

Title	LenS	LenD	NChar	FCon	Pers	Obje	Dire	Cntr	Cntr,%
OTwist	-0.18	0.29	0.29	1.46	0.23	0.02	0.10	2.21	6.31
DombyS	0.36	0.06	0	1.46	0.23	0.02	0.10	2.22	6.34
GExpectations	0.08	0.12	0	1.46	-0.14	-0.03	0.10	1.58	4.51
Cl. 1 Cntr	0.26	0.47	0.29	4.38	0.32	0.01	0.29	6.01	17.17
TomSoyer	0.48	0.44	0.58	0.52	-0.03	-0.14	0.10	1.95	5.57
HuckFinn	-0.38	0.83	0	0.52	0.02	0.23	0.10	1.32	3.77
YankeeA	1.22	1.21	0.58	0.52	0.02	0.23	0.10	3.88	11.09
Cl. 2 Cntr	1.31	2.48	1.17	1.58	0.01	0.32	0.29	7.15	20.43
WarPeace	0.14	-0.23	1.31	0.52	0.18	0.18	0.88	2.97	8.49
Akarenina	0.47	1.42	2.62	0.52	0.18	0.18	0.88	6.26	17.89
Cl. 3 Cntr	0.61	1.19	3.94	1.05	0.35	0.35	1.75	9.23	26.37
Explained	2.18	4.14	5.40	7.00	0.67	0.67	2.33	22.39	63.97
Unexplained	4.82	2.86	1.60	0	1.66	1.67	0	12.61	36.03
Total	7.00	7.00	7.00	7.00	2.33	2.33	2.33	35.00	100.00

Contribution based description of clusters

- C. Dickens: FCon = 0
- M. Twain: LenD < 28
- L. Tolstoy: NumCh > 3 or Direct = 1

Simulation study of Number-of clusters methods(joint work with Mark Chiang):

- Variance based:
  - **Hartigan(HK)**
  - **Calinski & Harabasz (CH)**
  - **Jump Statistic (JS)**
- Structure based:
  - **Silhouette Width (SW)**
- Consensus based:
  - **Consensus Distribution area (CD)**
  - **Consensus Distribution mean (DD)**
- Sequential extraction of APs:
  - **Least Square (LS)**
  - **Least Moduli (LM)**

Data generation for the experiment

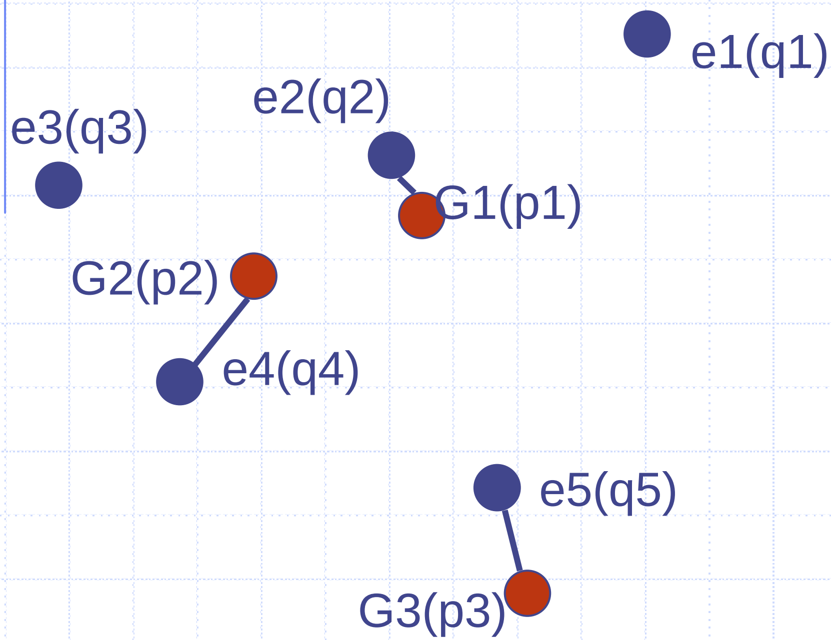
- Gaussian Mixture (6,7,9 clusters) with:
  - Cluster spatial size:
    - Constant (spherical)
    - k-proportional
    - k^2-proportional
  - Cluster spread (distance between centroids)

Spread	Spherical	PPCA model	
		k-proport.	k^2-proport.
Large	2 (❶)	10 (❷)	10 (❸)
Small	0.2 (❹)	0.5 (❺)	2 (❻)

Evaluation of results: Estimated clustering versus that generated

- Number of clusters
- Distance between centroids
- Similarity between partitions

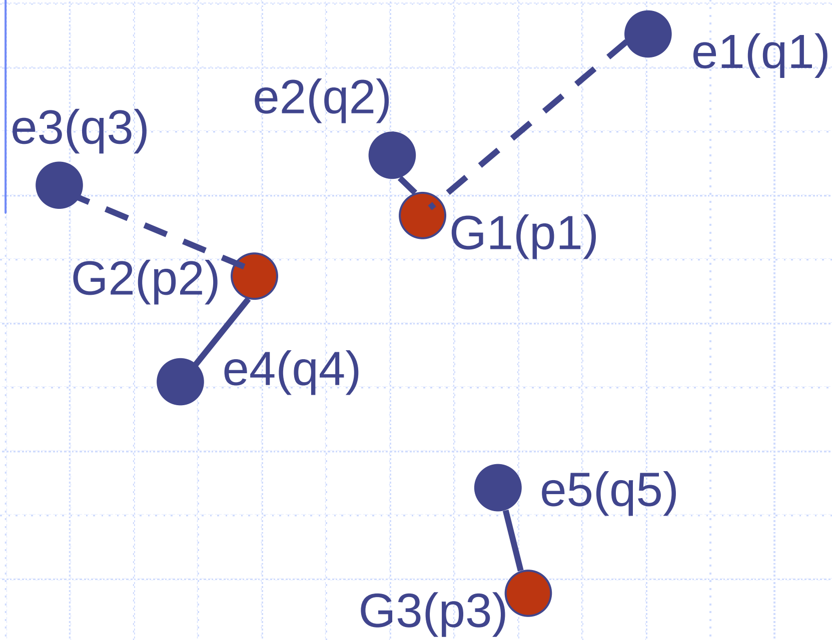
Distance between estimated centroids (o) and those generated (o)



Prime Assignment

g1-------e2
g2-------e4
g3-------e5

Distance between estimated centroids (o) and those generated (o)



Final Assignment

g1-------e2, e1
g2-------e4, e3
g3------e5

Distance between centroids: quadratic and city-block

g1(p1)------e1(q1), e2(q2)
g2(p2)------e3(q3), e4(q4)
g3(p3)------e5(q5)

1. Assignment
2. Distancing

d1=(q1*d(g1,e1)+q2*d(g1,e2))/(q1+q2)

d2=(q3*d(g2,e3)+q4*d(g2,e4))/(q3+q4)

d3=(q5*d(g3,e5)/q5

Distance between centroids: quadratic and city-block

p1*d1+p2*d2+p3*d3

1. Assignment
2. Distancing
3. Averaging

Similarity between partitions according to their confusion table

- Relative distance (Mirkin-Cherny 1970)
- Tchouprov coefficient (Cramer 1943)
- Adjusted Rand Index (Arabie-Hubert, 1985)
- Average Overlap (Mirkin 2005)

Results

at 9 clusters, 1000 entities, 20 features generated

	Estimated number of clusters		Distance between Centroids		Adjust Rand Index	
	Large spread	Small spread	Large spread	Small spread	Large spread	Small spread
HK						
CH						
JS						
SW						
CD						
DD						
LS						
LM						

Extending PCA to ITEX

- Iterative Extraction Elements:
  - Data X format: at PCA, entity-to-feature
  - Structure to extract; at t-th step set D(t): at PCA, a pair z and c;
  - Criterion to minimise, $\Phi(\varepsilon)$: at PCA, L2
  - Relation between D(t) and D(t+1): at PCA, same
  - Method for minimising, at step t, $\Phi(|X(t) - s|)$ over $s \in D(t)$ where $X(t)=X(t-1)-s(t-1)$, $X(0)=X$: at PCA, svd or AP clustering

Result: $X = \Sigma_t s(t) + \varepsilon$, along with Pythagorean decomposition of T(X)

Proof of (finite) convergence (Mirkin (1990, 1998))

ITEX examples:

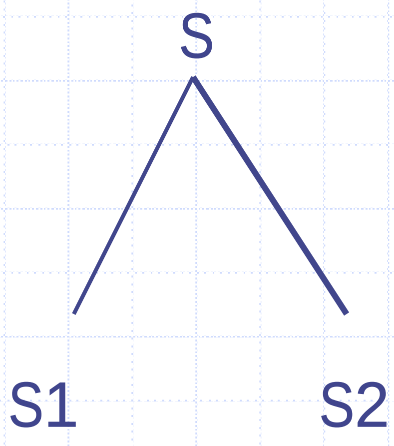
- Hierarchical clustering for conventional and spatial data
- Similarity clustering with additive clustering
- Similarity clustering with boxes (“plaid clustering”)
- Contingency data clustering and aggregation

Hierarchical clustering for conventional and spatial data

- Model: Same

$$y_{iv} = \sum_{k=1}^{K} c_{kv} z_{ik} + e_{iv},$$

- Cluster structure: 3-valued z's
- A split S=S1+S2 of a node S in children S1, S2:



z_i = 0 if $i \notin$S, = a if $i \in$S1
= $-b$ if $i \in$S2

If a and b taken to z being centred, the node vectors for a hierarchy form **orthogonal base** (an analogue to SVD)

Hierarchies, Wavelets, Haar base

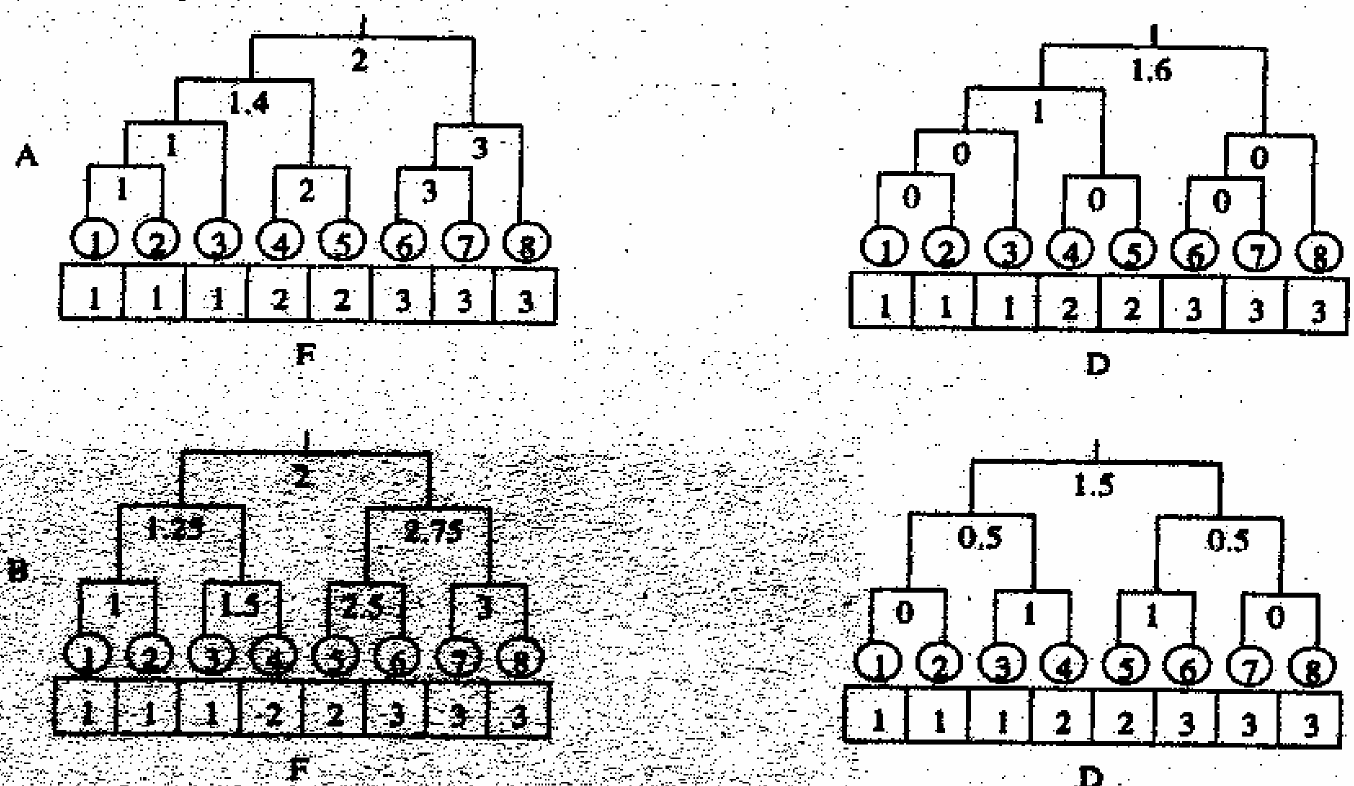


FIGURE 7. Compression and decompression of the boxed data with hierarchies A and B from Figure 6.

Similarity additive (and hierarchical) clustering

Observed similarity matrix

$$\mathbf{B} = \lambda_1 \mathbf{z}_1 \mathbf{z}_1^T + \lambda_2 \mathbf{z}_2 \mathbf{z}_2^T + \lambda_K \mathbf{z}_K \mathbf{z}_K^T + \mathbf{E}$$

Problem: given **B**, find λs and **z**s to minimize **E**, the differences between **B** and summary clusters

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

Additive clusters: ITEX

Doubly greedy strategy

OUTER LOOP: One cluster at a time

Find real λ (intensity) and binary $\boldsymbol{z}$ (membership) to minimize $\boldsymbol{L(B, \lambda, z)}$.

Update $\boldsymbol{B} \Leftarrow \boldsymbol{B} - \lambda \boldsymbol{zz}^T$; and reiterate!

After K iterations, clusters S_k, of cardinality N_k,

$$T(B) = \lambda_1^{\,2} N_1^{\,2} + \lambda_2^{\,2} N_2^{\,2} + \ldots + \lambda_K^{\,2} N_K^{\,2} + L^2 \quad (\bullet)$$

INNER LOOP: maximise $\lambda_k N_k$

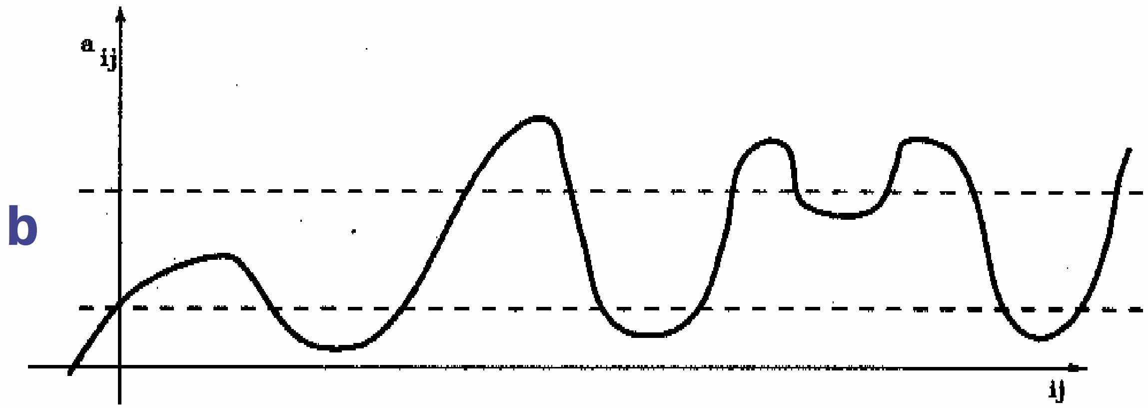
Algorithm: ADDI-S (Mirkin JoC 1987), a data approximation technique

- **To maximize Contribution to Data Scatter**, Average within-cluster similarity λ multiplied by the cluster's size ***#S***
- ***Algorithm ADDI-S:***
  - ***Take S={ j } for arbitrary j***
  - ***Given S, find λ =c(S) and similarities b(i,S) to S for all entities i in and out of S;***
  - ***Check the differences b(i,S)- λ /2. If they are consistent, change the state of a most contributing entity. Else, stop and output S.***
- **Resulting S: a tightness property**.
- Holzinger (1941) B-coefficient, Arkadiev&Braverman (1964, 1967) Specter, Mirkin (1976, 1987) ADDI family, Ben-Dor, Shamir, Yakhini (1999) CAST

Algorithm: ADDI-S a data approximation techniques

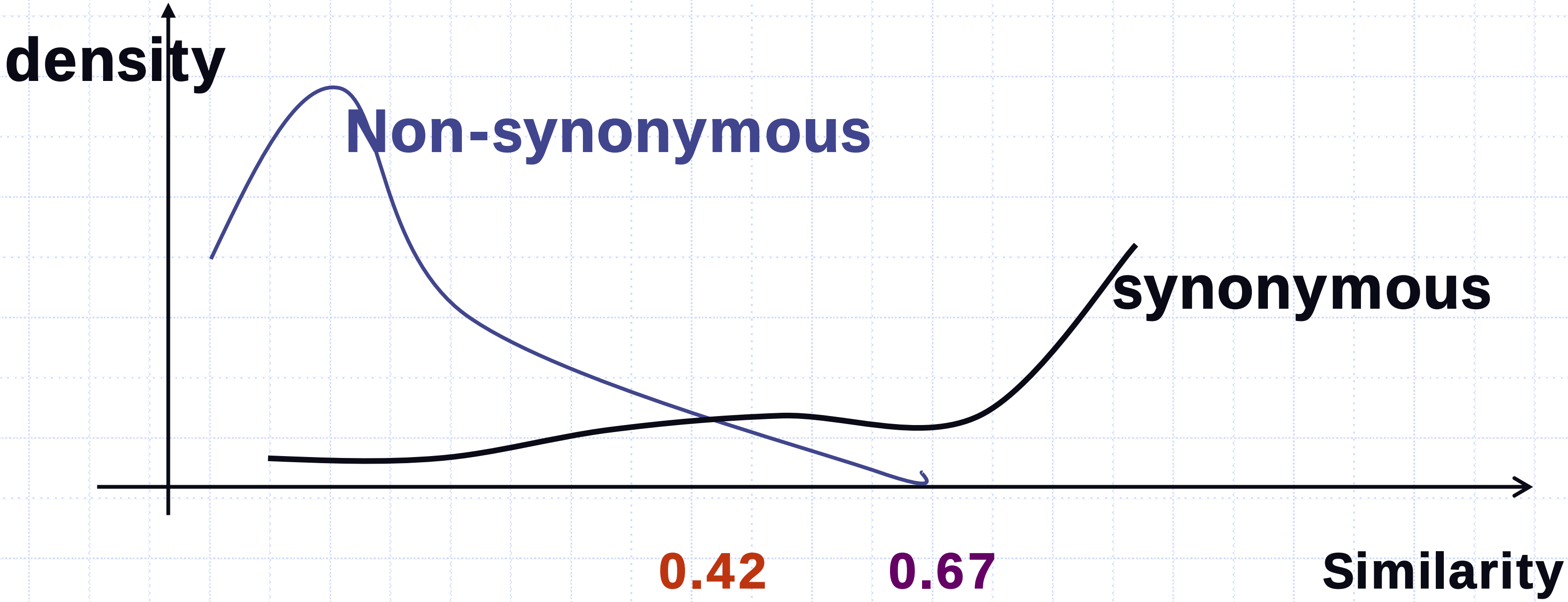
Number of clusters: Depends on similarity shift threshold b

$$b(ij) \Leftarrow b(ij) - b$$



Domain knowledge: Function is known at some HPFs

- 287 pairs of HPFs with known function of which 86 are SYNONYMOUS (same function)



Two values: Min error No non-synonymous

Hierarchical similarity clusters

Spectral clustering

Similarity clustering with boxes

Plaid clustering

Contingency data clustering and aggregation

- $P(I,J)=(p_{ij})$ non-negative and summable
- Correspondence Analysis rather than PCA
- Quetelet coefficients rather than p_{ij}

$$q_{ij} = p_{ij}/(p_{i+}\ p_{+j})-1 = [p(i/j)-p(i)]/p(i)$$

Let A partitions I and B partitions J: P(A,B) by summing up p_{ij} within blocks to approximate q_{ab} by the $p_{i+}\ p_{+j}$ weighted least-squares L^2:

Pythagorean

$$X^2(I,J)=X^2(A,B) + L^2$$

Conclusion

Looking forward to hear of further ideas for combining clustering and visualisation à la PCA