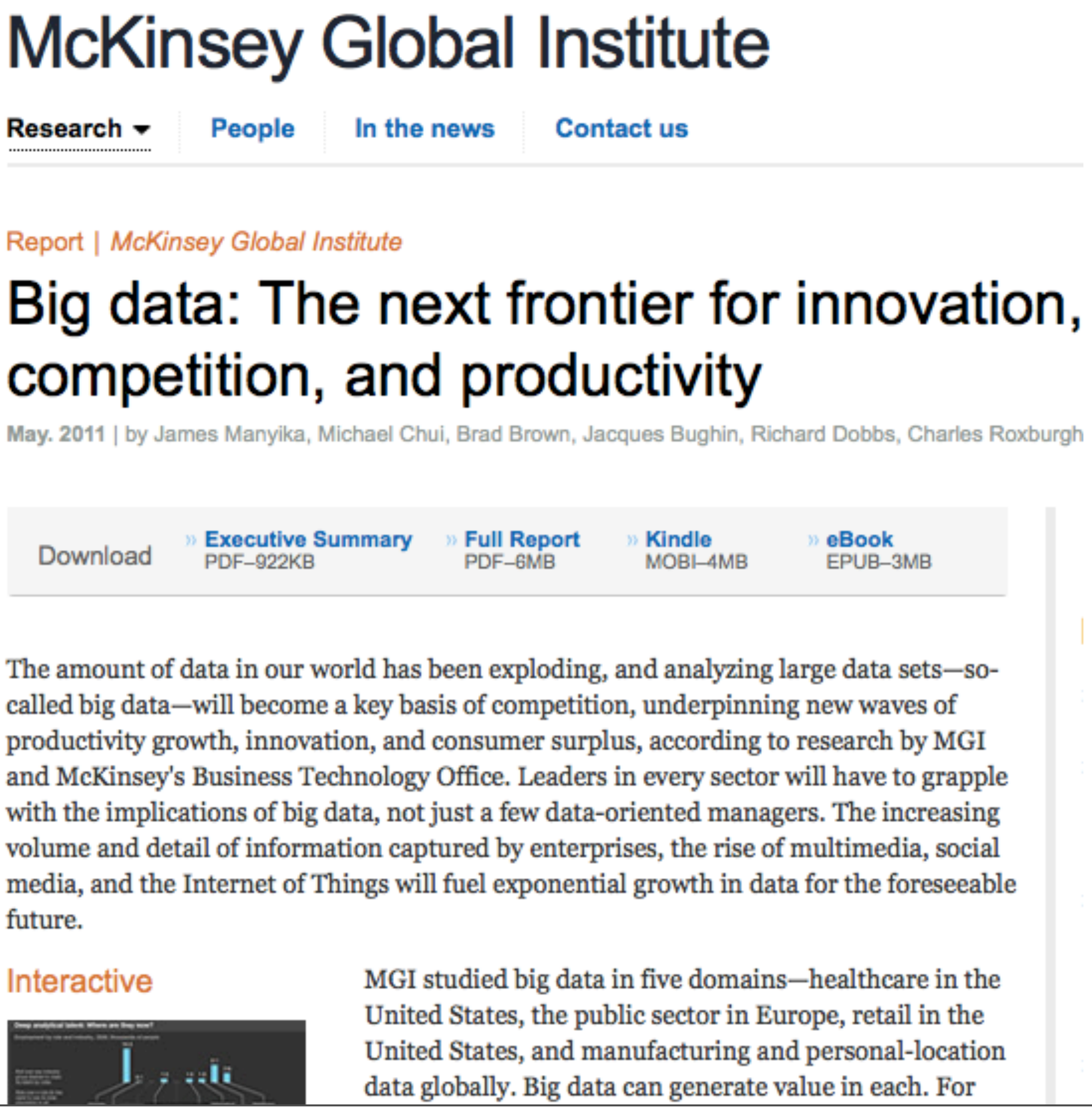
The New Science of Complex Systems Through Ultrametric and p-Adic Analysis: Application to Search and Discovery, to Narrative and to Thinking

Fionn Murtagh

International Workshop on p-Adic Methods for Modeling of Complex Systems
15-19 April 2013



Report | *McKinsey Global Institute*

Big data: The next frontier for innovation, competition, and productivity

May. 2011 | by James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh

The amount of data in our world has been exploding, and analyzing large data sets—so-called big data—will become a key basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus, according to research by MGI and McKinsey's Business Technology Office. Leaders in every sector will have to grapple with the implications of big data, not just a few data-oriented managers. The increasing volume and detail of information captured by enterprises, the rise of multimedia, social media, and the Internet of Things will fuel exponential growth in data for the foreseeable future.

MGI studied big data in five domains—healthcare in the United States, the public sector in Europe, retail in the United States, and manufacturing and personal-location data globally. Big data can generate value in each. For

Bringing big data to the enterprise

What is big data?

Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is **big data.**

Learn how **Vestas Wind Systems** use IBM big data analytics software and powerful IBM systems to improve wind turbine placement for optimal energy output.

Understanding Big Data

Gain insight into IBM's uniqu
at-rest big data analytics plat

**The Forrester Wave™: Ente
Solutions**

This Wave report evaluates 1
against 15 criteria with IBM b

Dominant Concerns in Regard to Applying Technologies: 3 Historical Phases

- Compute: Better computer infrastructure, including processor power and memory, up to the early 1990s.
- Network: especially from the release of the Mosaic web browser in early 1993. Followed later by search engines.
- Data: from the late 2000s.

- My talk: central role of geometric data analysis and I will be particularly focused on the role of hierarchical topology.

Basic ideas and definitions

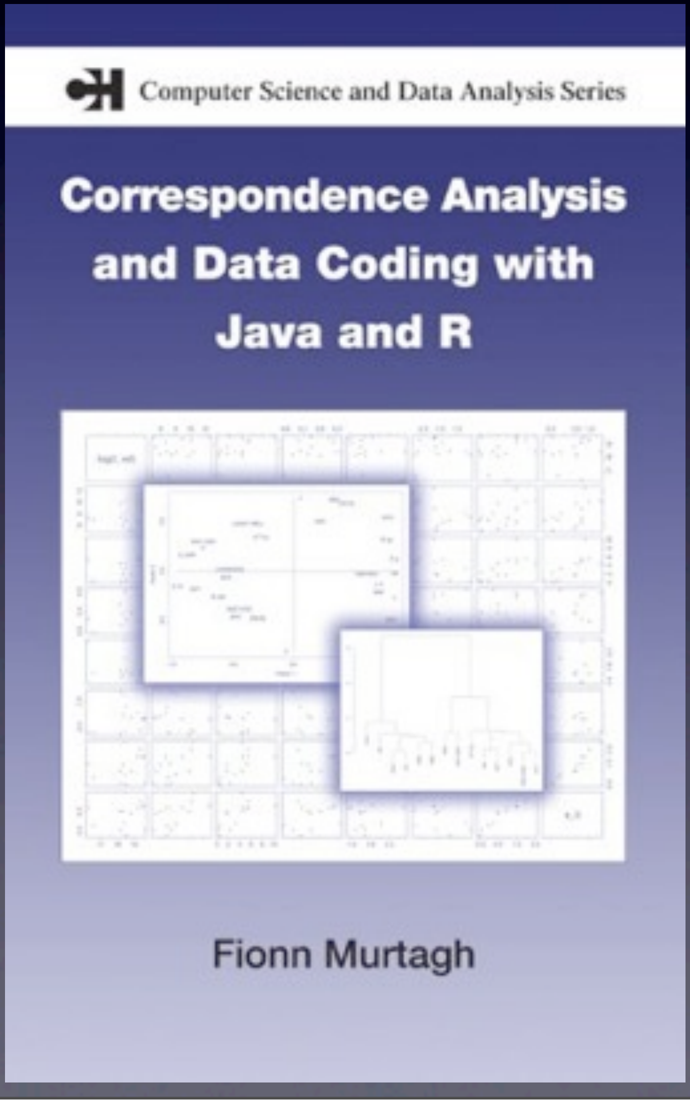
- Euclidean geometry for semantics of information.
- Hierarchical topology for other aspects of semantics, and in particular how a hierarchy expresses anomaly or change. A further useful case is when the hierarchy respects chronological or other sequence information.

Correspondence Analysis is
A Tale of Three Metrics

- Chi squared metric – appropriate for profiles of frequencies of occurrence

- Euclidean metric, for visualization, and for static context

- Ultrametric, for hierarchic relations and for dynamic context

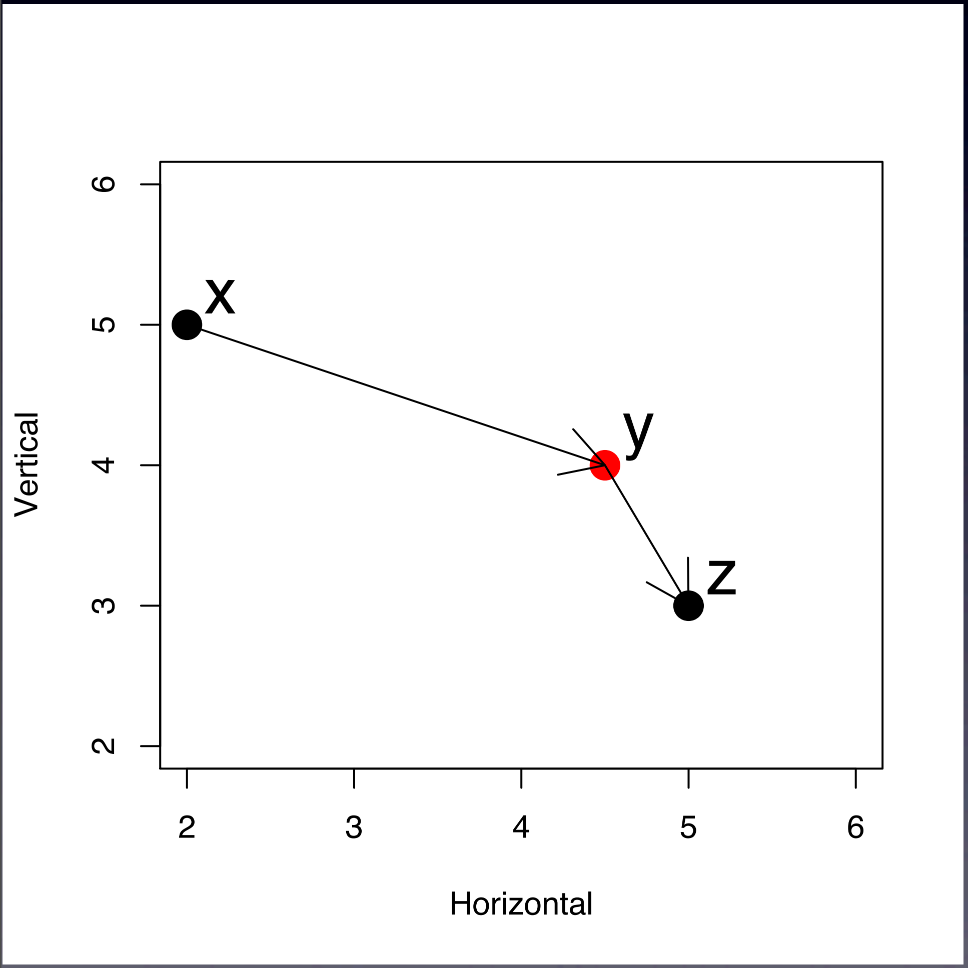


Computer Science and Data Analysis Series

Correspondence Analysis and Data Coding with Java and R

Fionn Murtagh

Triangular inequality holds for metrics

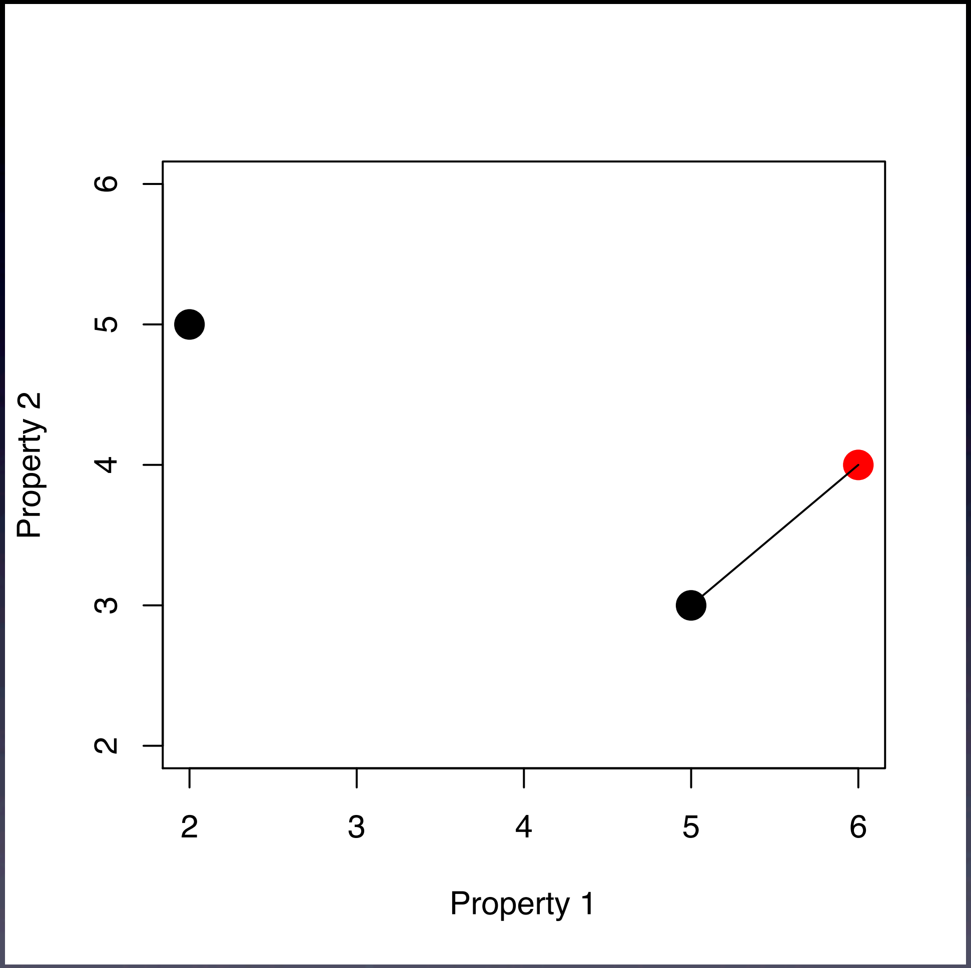


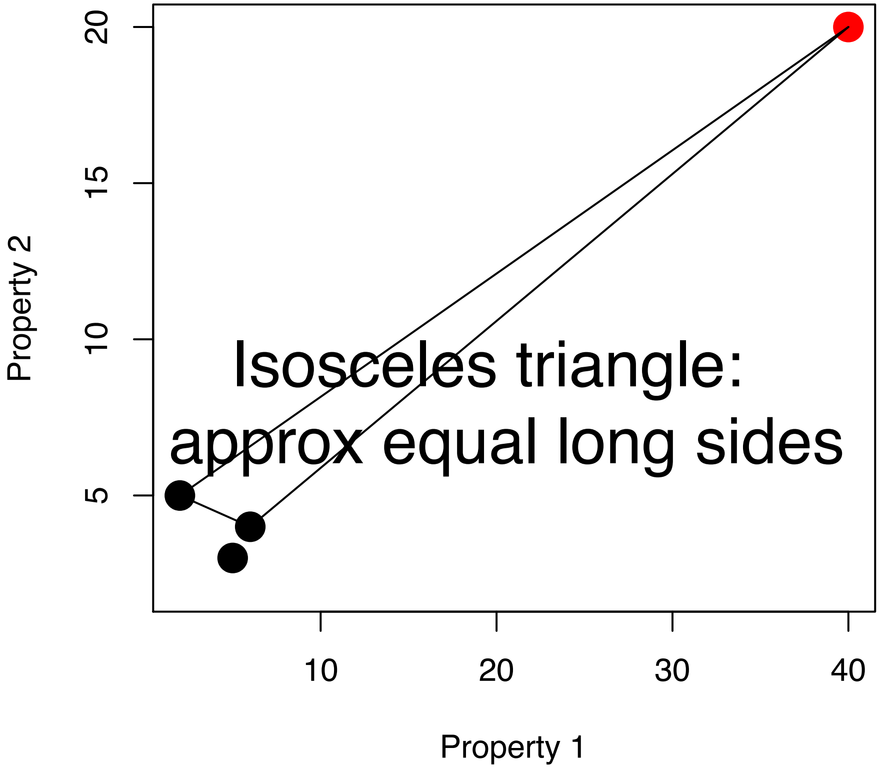
Example: Euclidean or "as the crow flies" distance

$$d(x,z) \leq d(x,y) + d(y,z)$$

Ultrametric

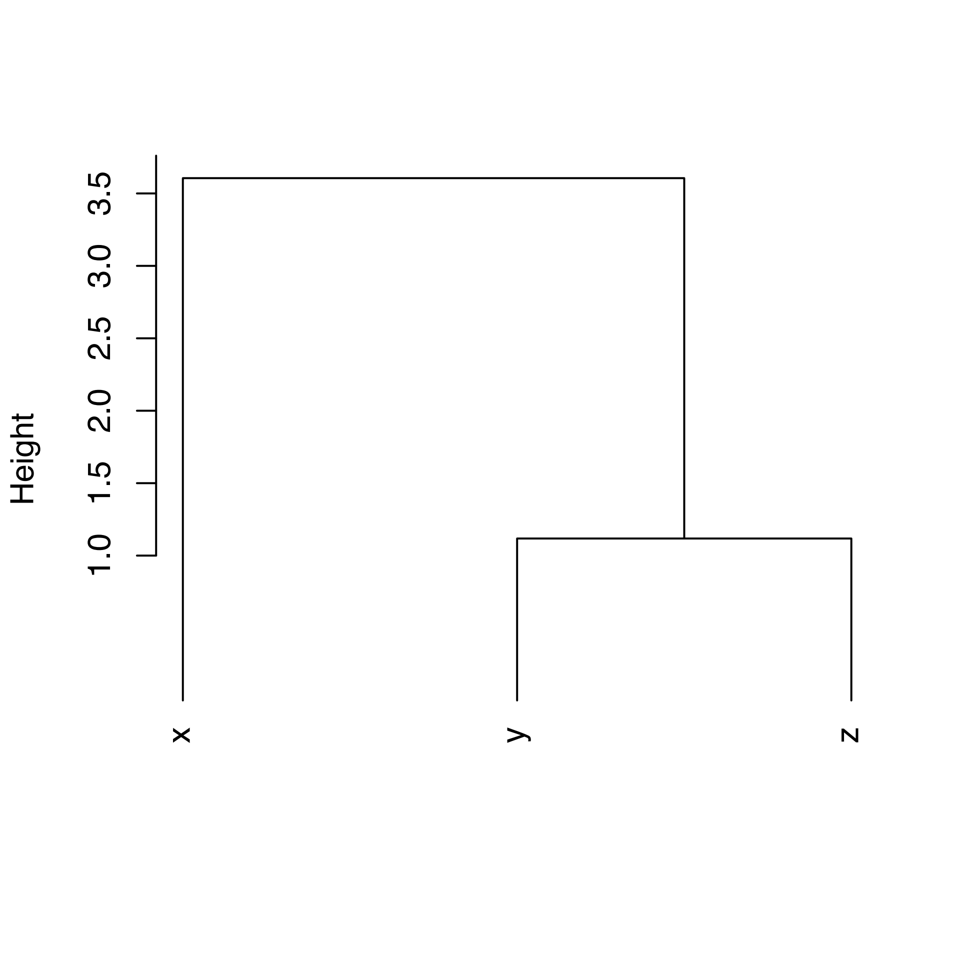
- Euclidean distance makes a lot of sense when the population is homogeneous
- Ultrametric distance makes a lot of sense when the observables are heterogeneous, discontinuous
- Latter is especially useful for determining: anomalous, atypical, innovative cases


Property 2
2
3
4
5
6
Property 1
2
3
4
5
6



Isosceles triangle:
approx equal long sides

Strong triangular inequality, or ultrametric inequality, holds for tree distances



$$d(x, z) \leq$$

$$\max\{d(x, y), d(y, z)\}$$

$$d(x, z) = 3.5$$

$$d(x, y) = 3.5$$

$$d(y, z) = 1.0$$

Closest common ancestor distance is an ultrametric

Some Properties of Ultrametrics

- The distance between two objects -- or two terminals in the tree -- is the lowest rank which dominates them. Lowest or closest common ancestor distance.
- The ultrametric inequality holds for any 3 points (or terminals):
- $d(i, k) \leq \max \{d(i,j), d(j,k)\}$
- Recall: the triangular inequality is: $d(i,k) \leq \{d(i,j) + d(j,k)\}$
- An ultrametric space is quite special: (i) all triangles are isosceles with small base, or equilateral; (ii) every point in a ball is its centre; (iii) the radius of a ball equals the diameter; (iv) a ball is *clopen*; (v) an ultrametric space is always topologically 0-dimensional.

Quantifying Ultrametricity

Examples of application to text collections and time series

- Take all triplets of points (or sample), check isosceles with small base, or equilateral, configurations, and determine a coefficient of the relative proportion of such triangles.
- On a scale of 1 = 100% ultrametric-respecting properties, 0 = 0%, find (averaged values): Grimm Brothers: 0.1147; Jane Austen: 0.1404; aviation accident reports: 0.1154; dream reports: 0.1933 (in the case of one person, 0.2603). Joyce's Ulysses (between the latter two).
- Adapted for 1D signals: FTSE, USD/EUR, sunspot, stock, futures, eyegaze, Mississippi, www traffic, EEG/sleep/normal, EEG/petit mal epilepsy, EEG/irreg. epilepsy, quadratic chaotic map, uniform. (Eyegaze data high, chaotic series low, in inherent ultrametricity.)

Pervasive Ultrametricity

- As dimensionality increases, so does ultrametricity.
- In very high dimensional spaces, the ultrametricity approaches being 100%.
- Relative density is important: high dimensional and spatially sparse mean the same in this context.

- See: F Murtagh, “On ultrametricity, data coding, and computation”, Journal of Classification, 21, 167-184, 2004
- Hall, P., Marron, J.S., and Neeman, A., “Geometric representation of high dimension low sample size data”, JRSS B, 67, 427-444, 2005
- F. Delon, Espaces ultramétriques, J. Symbolic Logic, 49, 405-502, 1984

Computational Implications

- Consider a dendrogram: a rooted, labeled, ranked, binary tree. So: *n* terminals, *n-1* levels.
- A dendrogram's root-to-terminal path length is $log_2 n$ for a balanced tree, and *n-1* for an imbalanced tree. Call the computational cost of such a traversal *O(t)* where *t* is this path length. It holds: $1 \geq O(t) \geq n-1$.
- Adding a new terminal to a dendrogram is carried out in *O(t)* time.
- Cost of finding the ultrametric distance between two terminal nodes is twice the length of a traversal from root to terminals in the dendrogram. Therefore distance is computed in *O(t)* time.
- Nearest neighbor search in ultrametric space can be carried out in *O(1)* or constant time.

Applications in Search and Discovery

– First, agglomerative hierarchical clustering; or: "hierarchical encoding" of data.

– Ultrametric topology, Baire distance.

– Clustering of large data sets.

– Hierarchical clustering via Baire distance using SDSS (Sloan Digital Sky Survey) spectroscopic data.

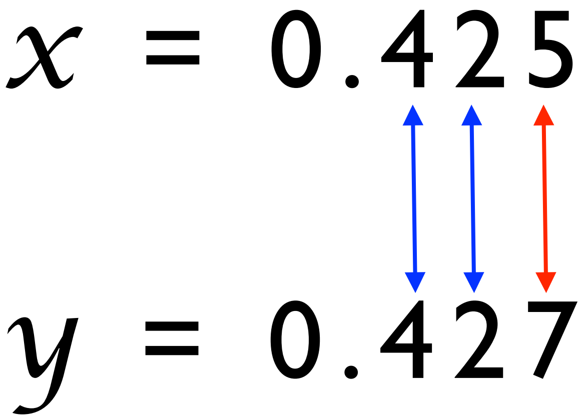
–Hierarchical clustering via Baire distance using chemical compounds.

– Then I will move to narrative analysis and synthesis.

Next: the Baire (ultra)metric

Baire, or longest common prefix distance – and also an ultrametric

An example of Baire distance for two numbers (x and y) using a precision of 3:



$x = 0.425$

$y = 0.427$

Baire distance between x and y:

$d_{\mathcal{B}}(x, y) = 10^{-2}$

Base ($\mathcal{B}$) here is 10 (suitable for real values)

Precision here = |K| = 3

That is:

k=1 -> $x_k = y_k$ -> 4

k=2 -> $x_k = y_k$ -> 2

k=3 -> $x_k \neq y_k$ -> $5 \neq 7$

On the Baire (ultra)metric

– Baire space consists of countable infinite sequences with a metric defined in terms of the longest common prefix *[A. Levy. Basic Set Theory, Dover, 1979 (reprinted 2002)]*

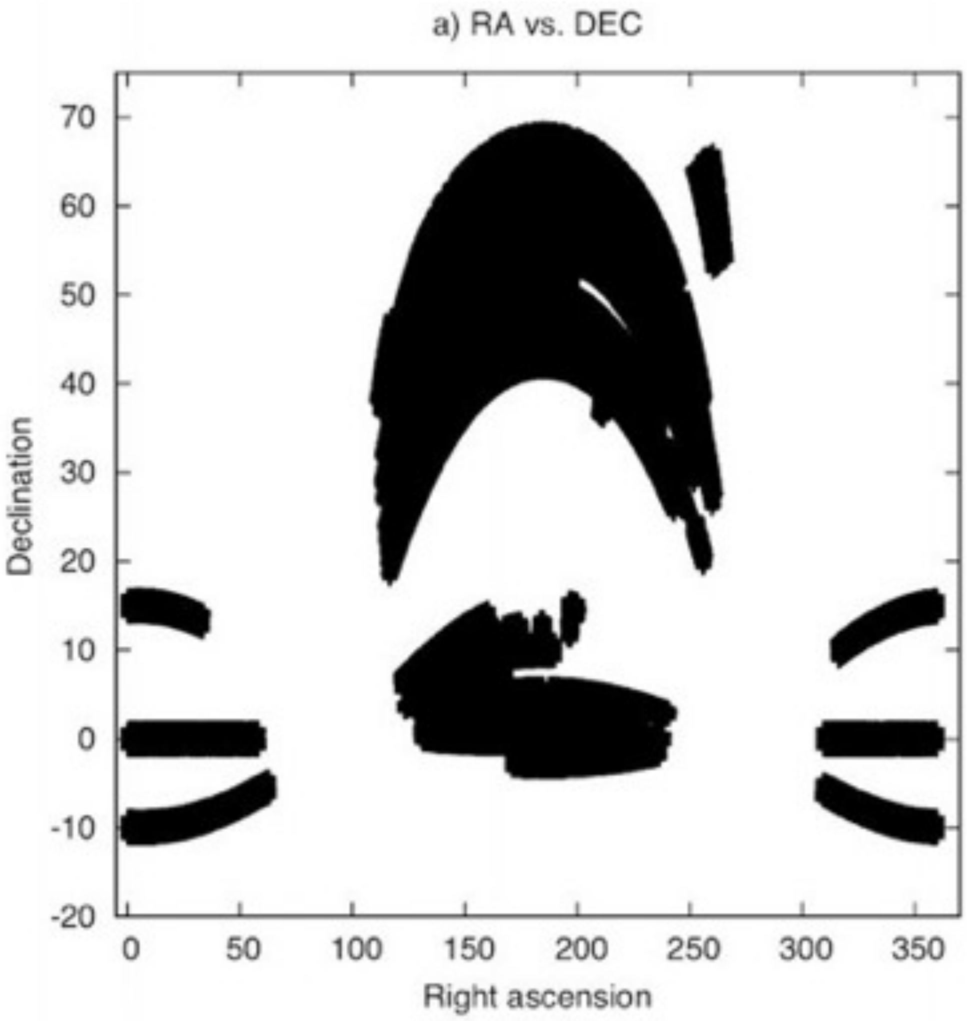
– The longer the common prefix, the closer a pair of sequences.

– The Baire distance is an ultrametric distance. It follows that a hierarchy can be used to represent the relationships associated with it. Furthermore the hierarchy can be directly read from a linear scan of the data. (Hence: hierarchical hashing scheme.)

– We applied the Baire distance to: chemical compounds, spectrometric and photometric redshifts from the Sloan Digital Sky Survey (SDSS), and various other datasets.

- A subset was taken of approximately 0.5 million data points from the SDSS release 5.
- These were objects with RA and Dec (Right Ascension and Declination, and spectrometric redshift, and photometric redshift). Problem addressed: regress one redshift (spectro.) on the other (photo.).
- Baire approach used, and compared with k-means.
- 1.2 million chemical compounds, each characterized by 1052 boolean presence/absence values.
- Random projections used on normalized compound/attribute values.
- Baire approach used; also another approach based on restricting the precision of the normalized compound/attribute values.

SDSS (Sloan Digital Sky Survey) Data



- We took a subset of approximately 0.5 million data points from the SDSS release 5 *[see D'Abrusco et al]*:

  -declination (Dec)

  -right ascension (RA)

  -spectrometric redshift

  -photometric redshift.

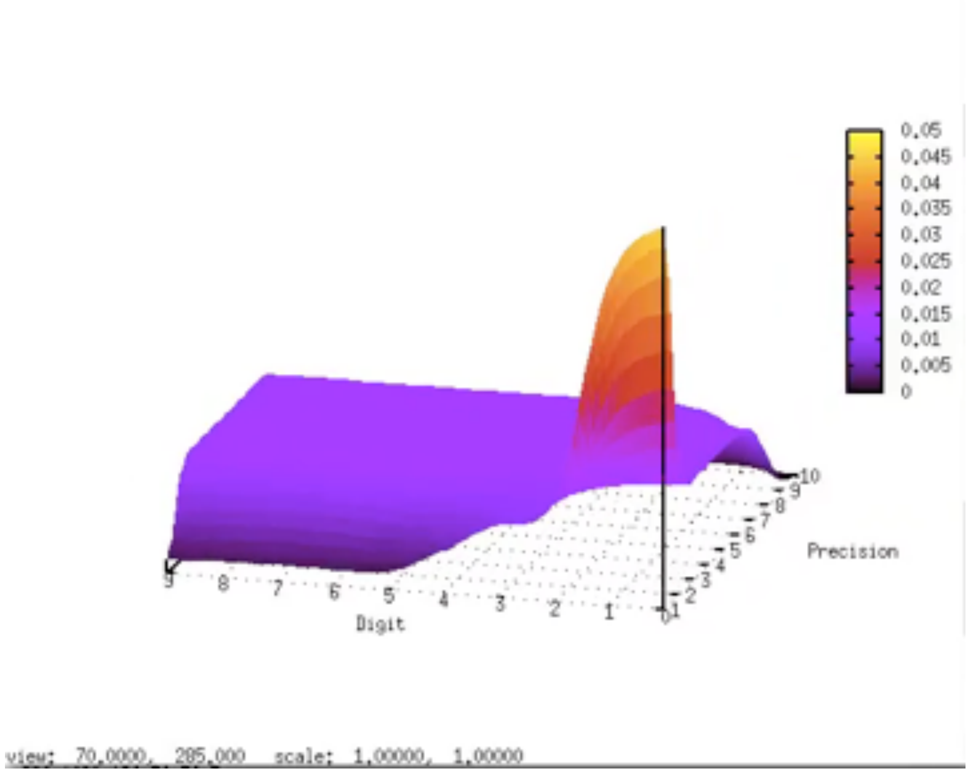
- Dec vs RA are shown in the figure.

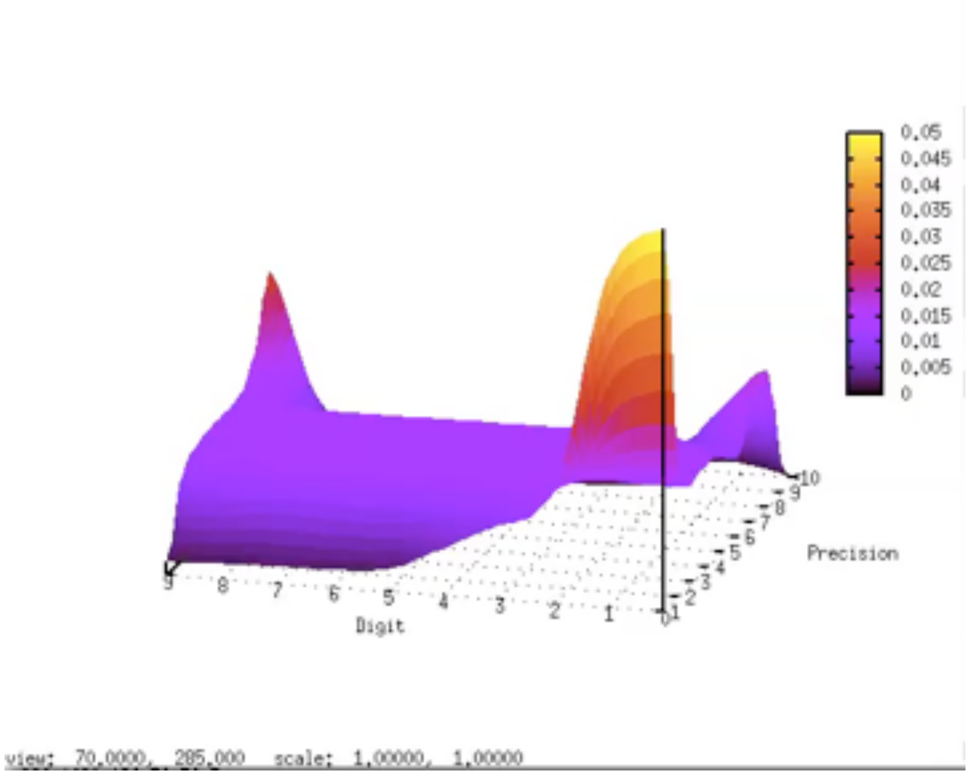
Data - example

RA	DEC	spec. redshift	phot. redshift
145.4339	0.56416792	0.14611299	0.15175095
145.42139	0.53370196	0.145909	0.17476539
145.6607	0.63385916	0.46691701	0.41157582
145.64568	0.50961215	0.15610801	0.18679948
145.73267	0.53404553	0.16425499	0.19580211
145.72943	0.12690687	0.03660919	0.06343859
145.74324	0.46347806	0.120695	0.13045037

- Motivation - regress z_spect on z_phot
- Furthermore: determine good quality mappings of z_spect onto z_phot, and less qood quality mappings
- I.e., cluster-wise nearest neighbour regression
- Note: cluster-wise not spatially (RA, Dec) but rather within the data itself

Perspective Plots of Digit Distributions





On the left we have z_spec where three data peaks can be observed. On the right we have z_phot where only one data peak can be seen.

Framework for Fast Clusterwise Regression

- 82.8% of z_spec and z_phot have at least 2 common prefix digits.

- I.e. numbers of observations sharing 6, 5, 4, 3, 2 decimal digits.

- We can find very efficiently where these 82.8% of the astronomical objects are.

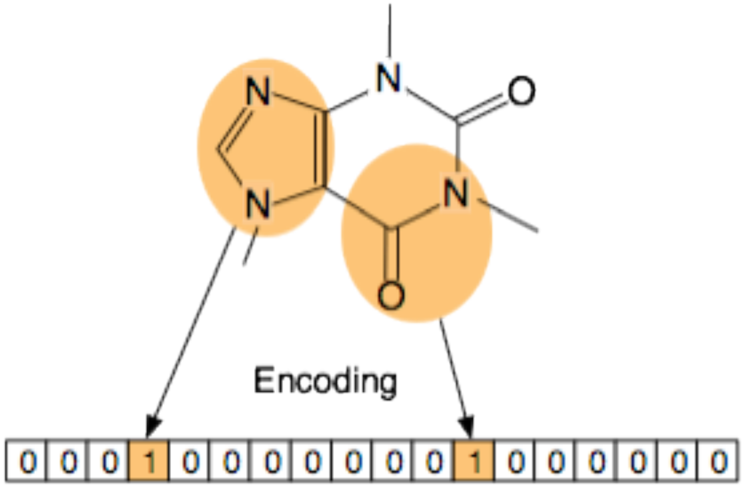
- 21.7% of z_spec and z_phot have at least 3 common prefix digits.

- I.e. numbers of observations sharing 6, 5, 4, 3 decimal digits.

- Next - another case study, using chemoinformatics - which is high dimensional.

- Since we are using digits of precision in our data (re)coding, how do we handle high dimensions?

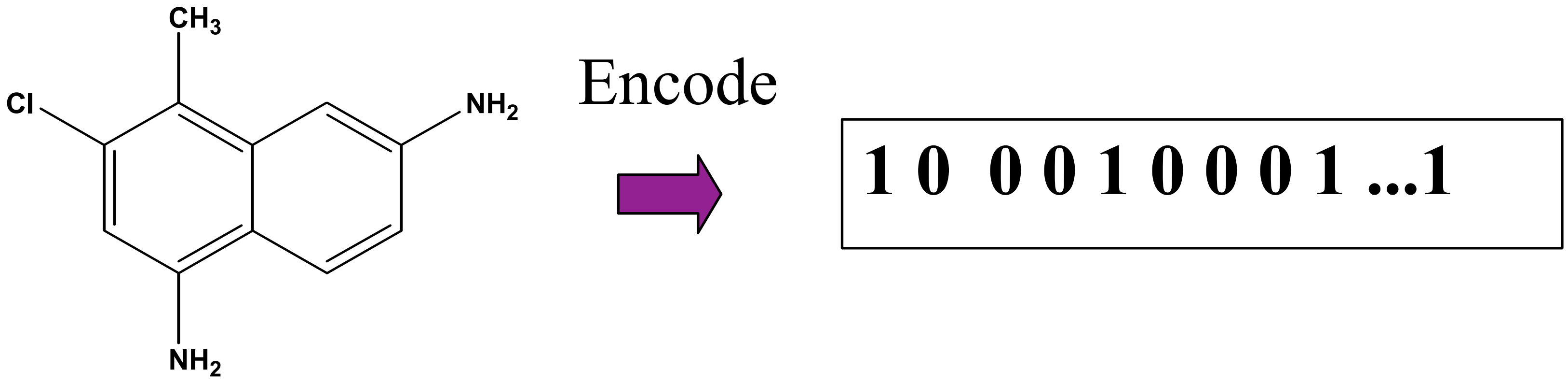
Baire Distance Applied to Chemical Compounds



Matching of Chemical Structures

- Clustering of compounds based on chemical descriptors or chemical representations, in the pharmaceutical industry.
- Used for screening large corporate databases.
- Chemical warehouses are expanding due to mergers, acquisitions, and the synthetic explosion brought about by combinatorial chemistry.

Binary Fingerprints



Encode

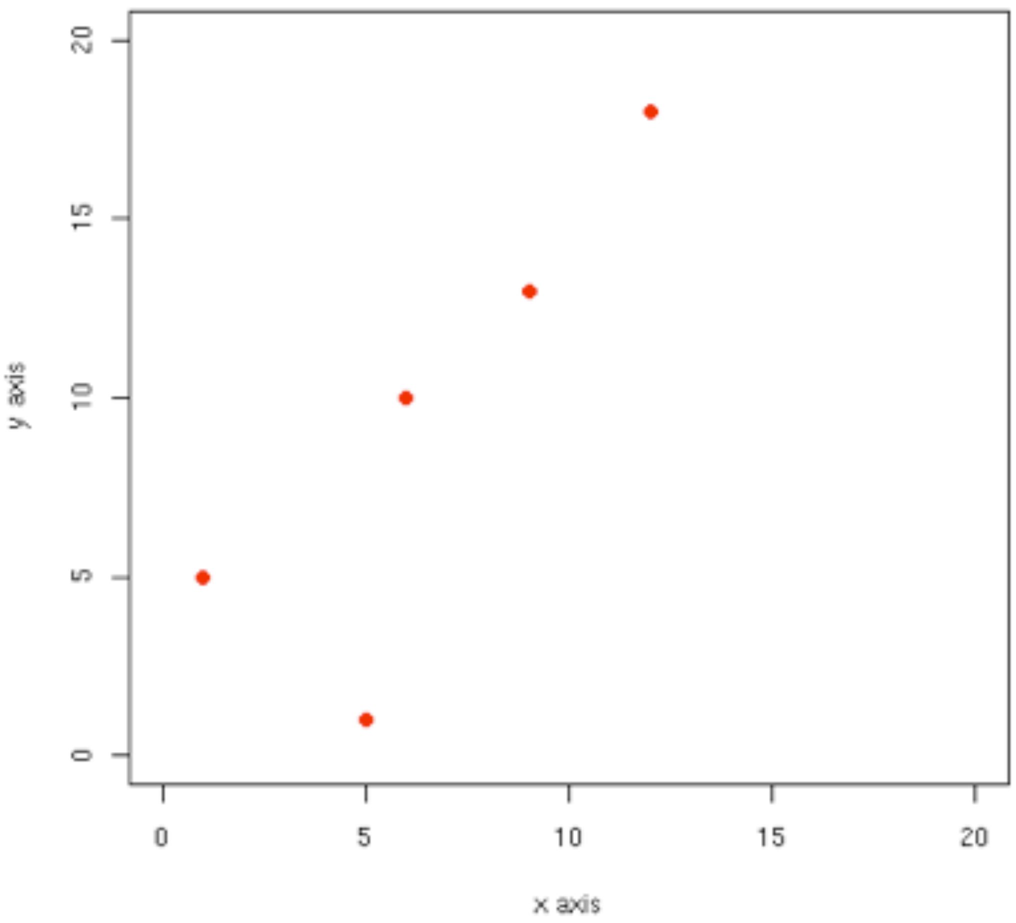
1 0 0 0 1 0 0 0 1 ...1

Fixed length bit strings such as
Daylight
MDL
BCI
etc.

Chemoinformatics clustering

- 1.2 million chemical compounds, each characterized by 1052 boolean presence/absence values.
- Firstly we note that precision of measurement leads to greater ultrametricity (i.e. the data are more hierarchical).
- From this we develop an algorithm for finding equivalence classes of specified precision chemicals. We call this: data “condensation”.
- Secondly, we use random projections of the 1052-dimensional space in order to find the Baire hierarchy. We find that clusters derived from this hierarchy are quite similar to k-means clustering outcomes.

Random projection and hashing

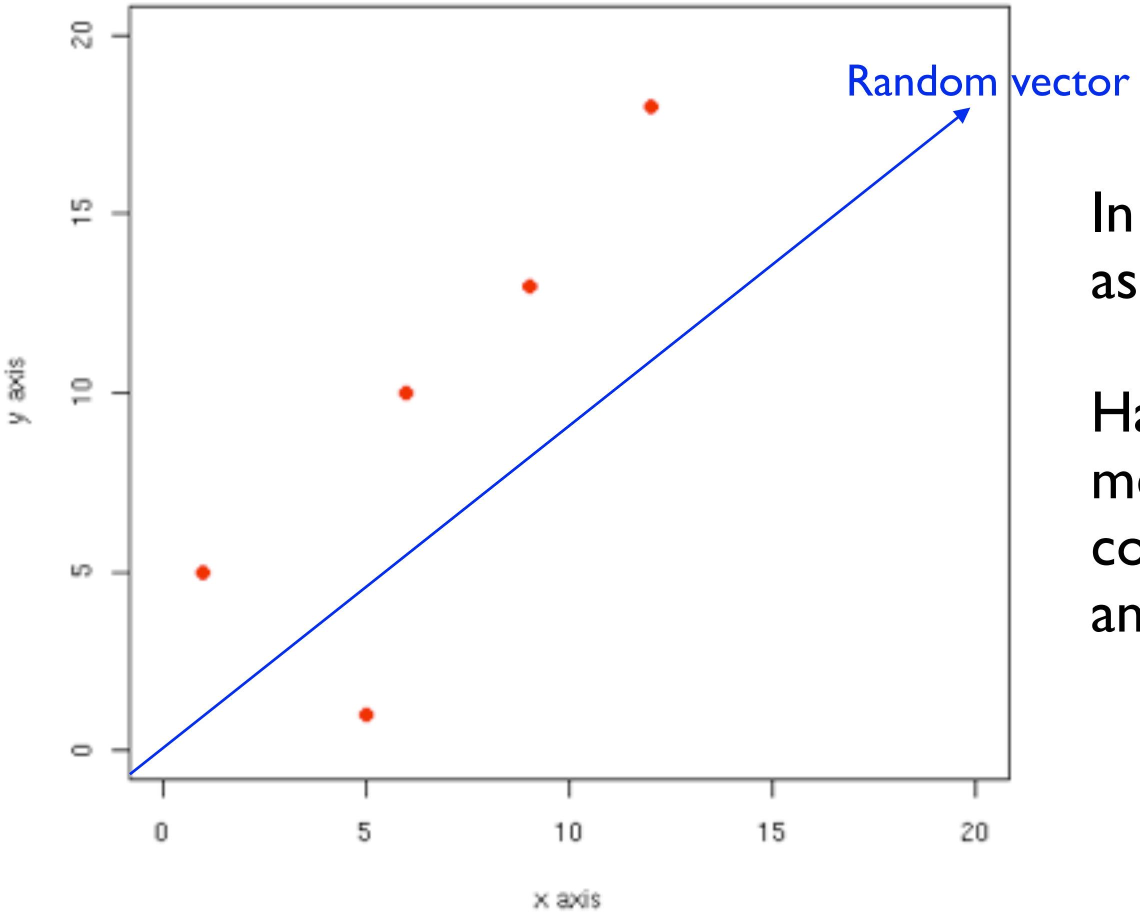


In fact random projection here works as a class of hashing function.

Hashing is much faster than alternative methods because it avoids the pairwise comparisons required for partitioning and classification.

If two points (p , q) are close, they will have a very small |p-q| (Euclidean metric) value; and they will hash to the same value with high probability; if they are distant, they should collide with small probability.

Random projection and hashing

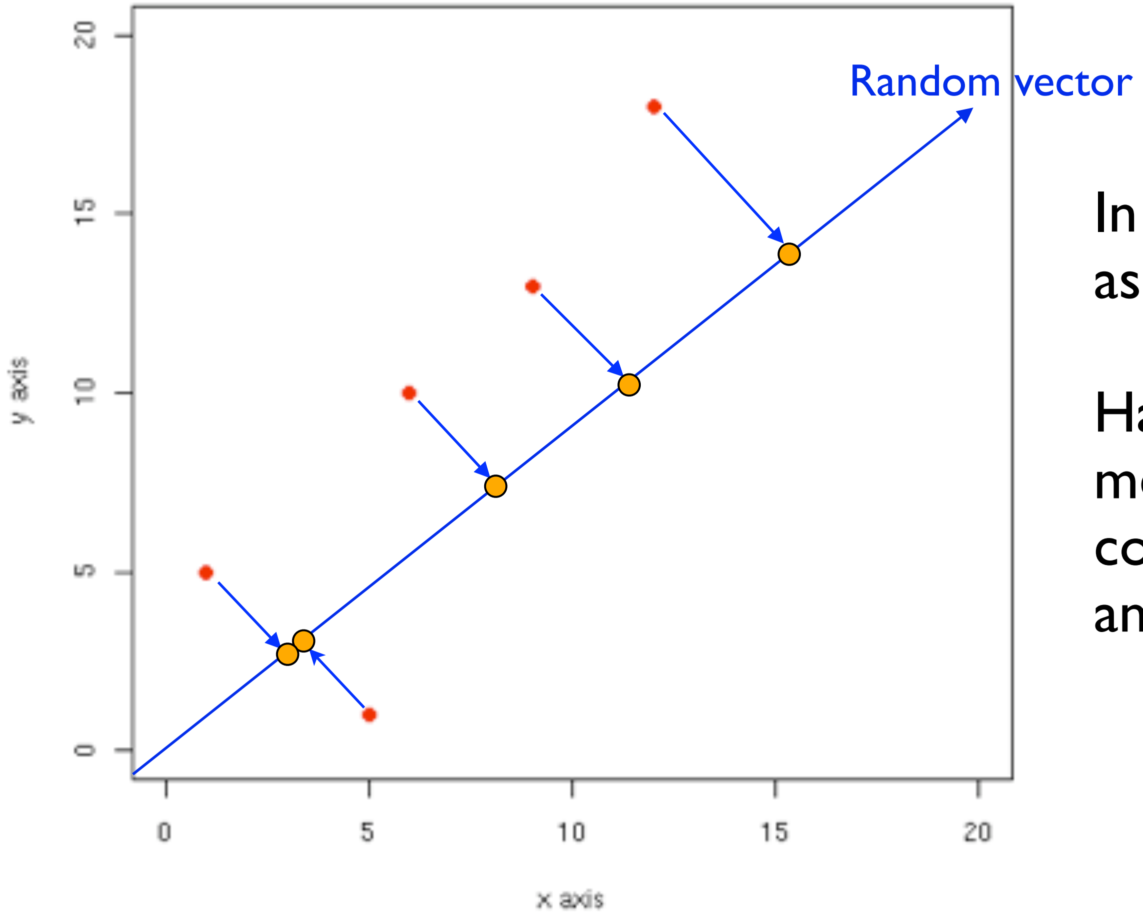


In fact random projection here works as a class of hashing function.

Hashing is much faster than alternative methods because it avoids the pairwise comparisons required for partitioning and classification.

If two points (p , q) are close, they will have a very small |p-q| (Euclidean metric) value; and they will hash to the same value with high probability; if they are distant, they should collide with small probability.

Random projection and hashing

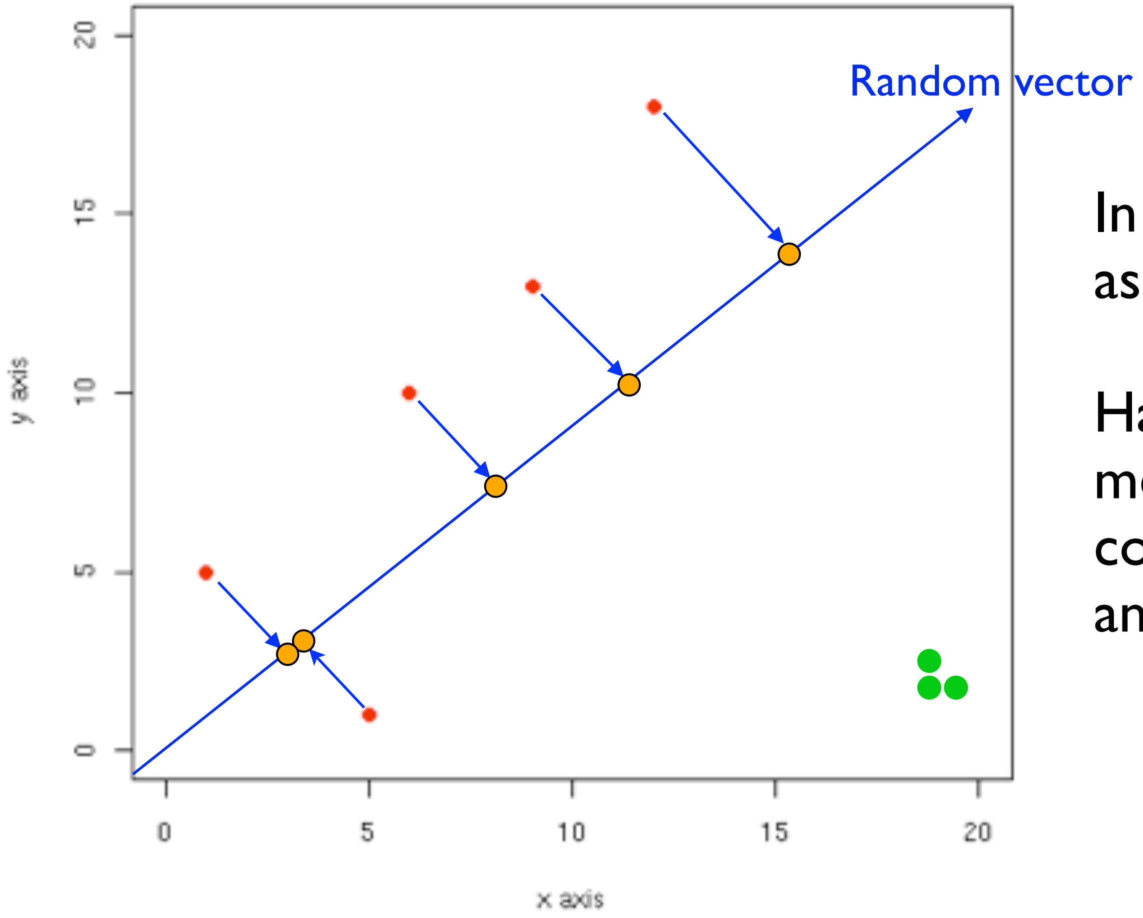


In fact random projection here works as a class of hashing function.

Hashing is much faster than alternative methods because it avoids the pairwise comparisons required for partitioning and classification.

If two points (p , q) are close, they will have a very small |p-q| (Euclidean metric) value; and they will hash to the same value with high probability; if they are distant, they should collide with small probability.

Random projection and hashing

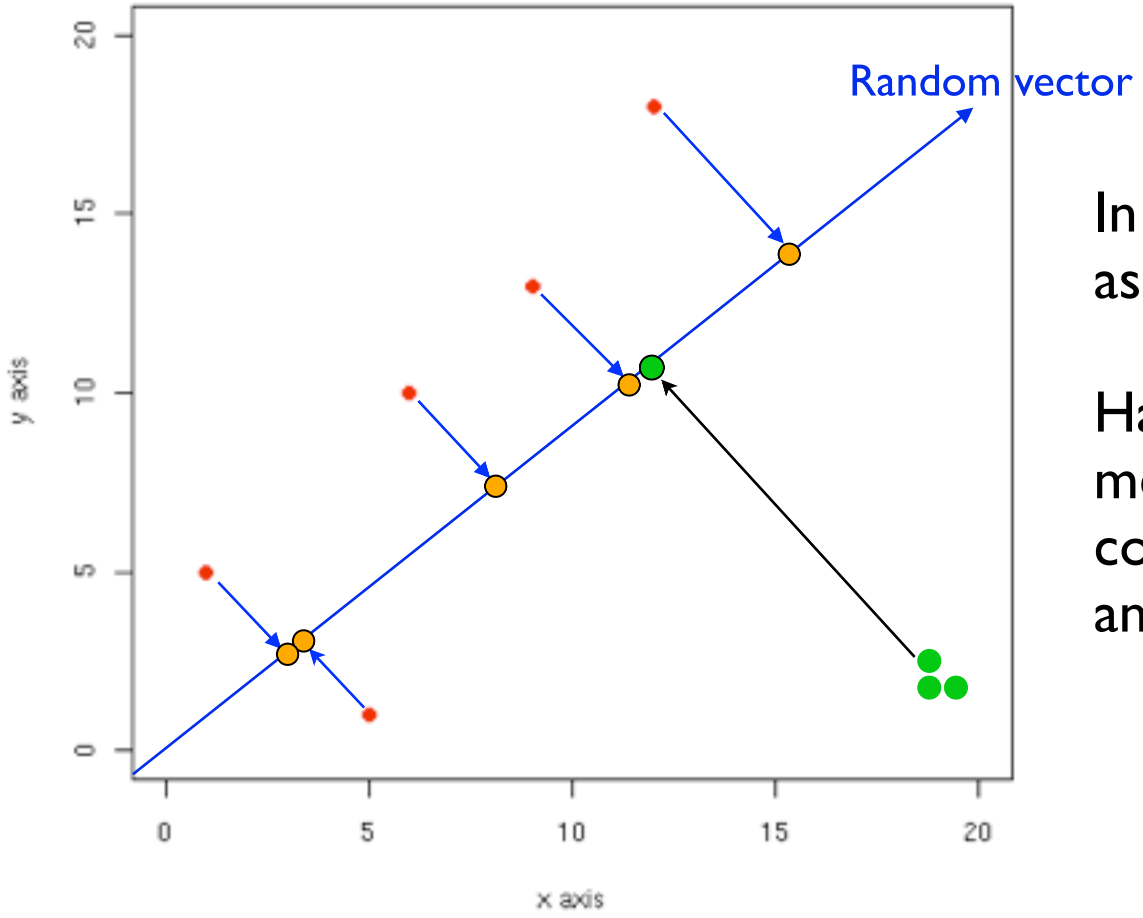


In fact random projection here works as a class of hashing function.

Hashing is much faster than alternative methods because it avoids the pairwise comparisons required for partitioning and classification.

If two points (p , q) are close, they will have a very small |p-q| (Euclidean metric) value; and they will hash to the same value with high probability; if they are distant, they should collide with small probability.

Random projection and hashing



In fact random projection here works as a class of hashing function.

Hashing is much faster than alternative methods because it avoids the pairwise comparisons required for partitioning and classification.

If two points (p , q) are close, they will have a very small |p-q| (Euclidean metric) value; and they will hash to the same value with high probability; if they are distant, they should collide with small probability.

- Normalize chemical compounds by dividing each row by row sum (hence “profile” in Correspondence Analysis terms).
- Two clustering approaches studied:
- Limit precision of compound / attribute values. This has the effect of more compound values becoming the same for a given attribute. Through a heuristic (e.g. interval of row sum values), read off equivalence classes of 0-distance compounds, with restricted precision. Follow up if required with further analysis of these crude clusters. We call this “data condensation”. For 20000 compounds, 1052 attributes, a few mins. needed in R.
- Second approach: use random projections of the high dimensional data, and then use the Baire distance.

Summary Remarks on Search and Discovery

- We have a new way of inducing a hierarchy on data
- First viewpoint: encode the data hierarchically and essentially read off the clusters
- Alternative viewpoint: we can cluster information based on the longest common prefix
- We obtain a hierarchy that can be visualized as a tree
- We are hashing, in a hierarchical or multiscale way, our data
- We are targeting clustering in massive data sets
- The Baire method - we find - offers a fast alternative to k-means and a fortiori to traditional agglomerative hierarchical clustering
- At issue throughout this work: embedding of our data in an ultrametric topology

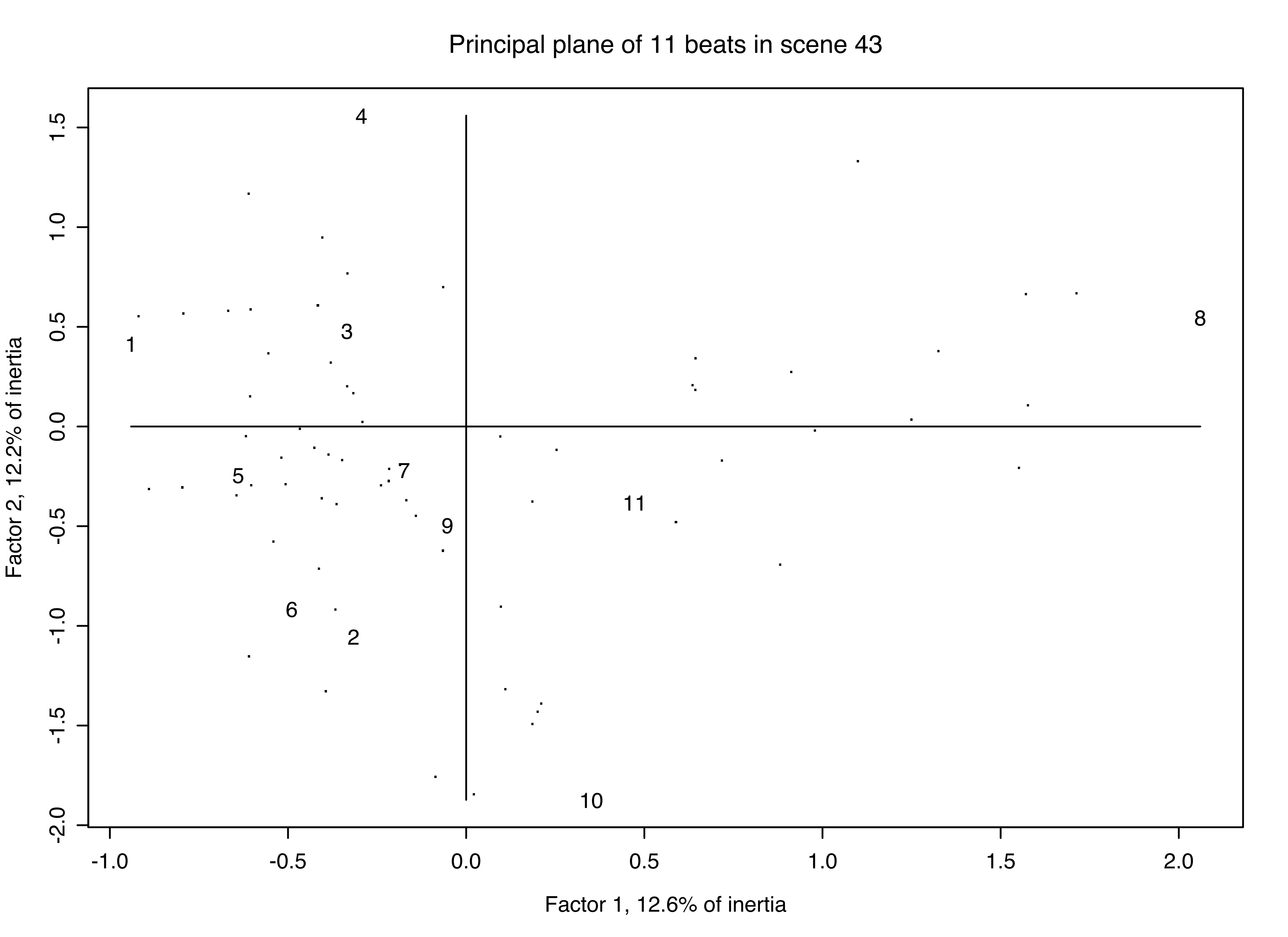
Analysis of Narrative

Technical Issues Addressed

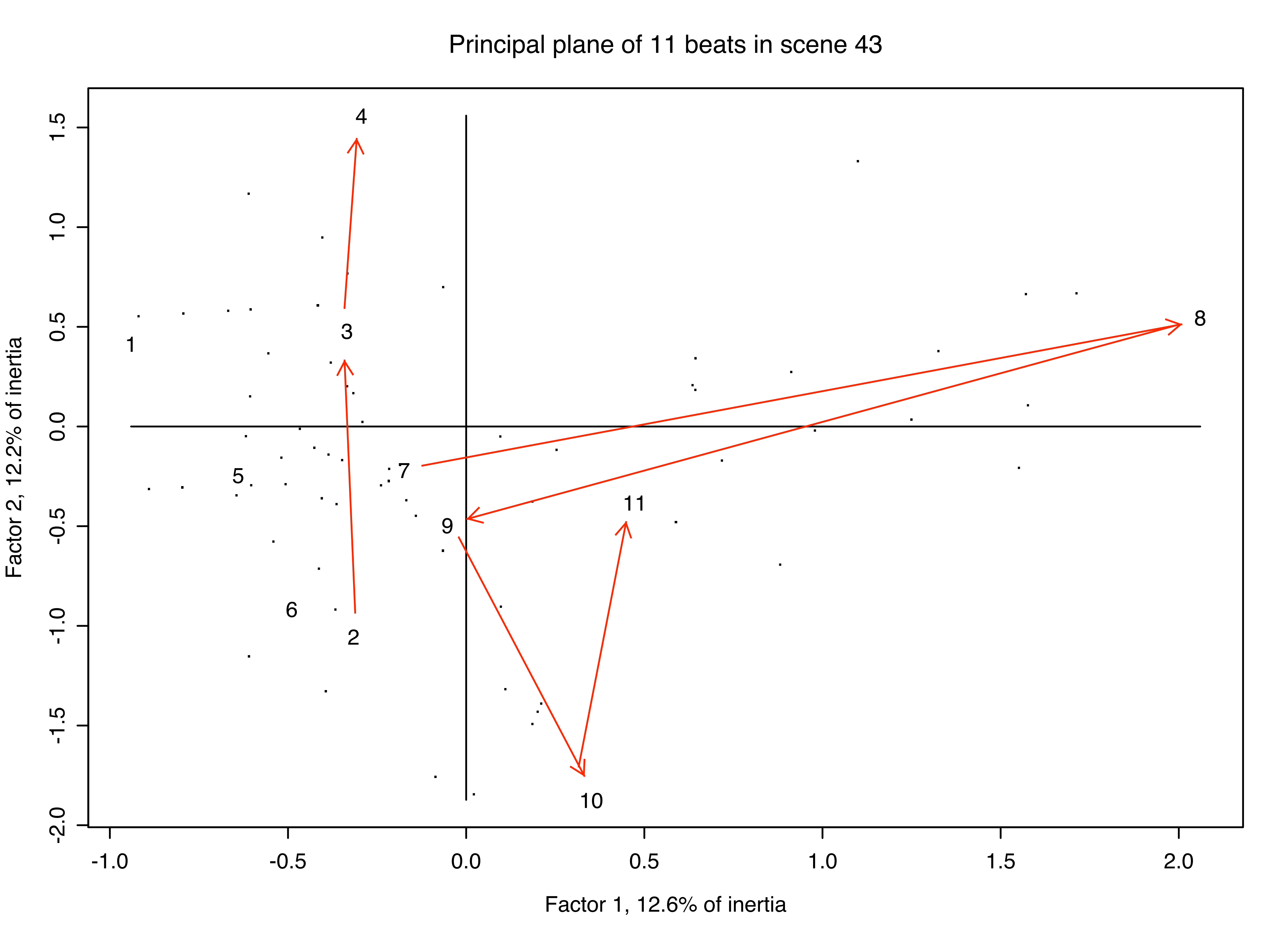
- We must consider complex web of relationships.
- Semantics include web of relationships - thematic structures and patterns. Structures and interrelationships evolve in time.
- Semantics include time evolution of structures and patterns, including both: threads and commonality; and change, the exceptional, the anomalous.
- Narrative suggests a causal or emotional relationship between events.
- A story is an expression of causality or connection
- Narrative connects facts or views or other units of information.

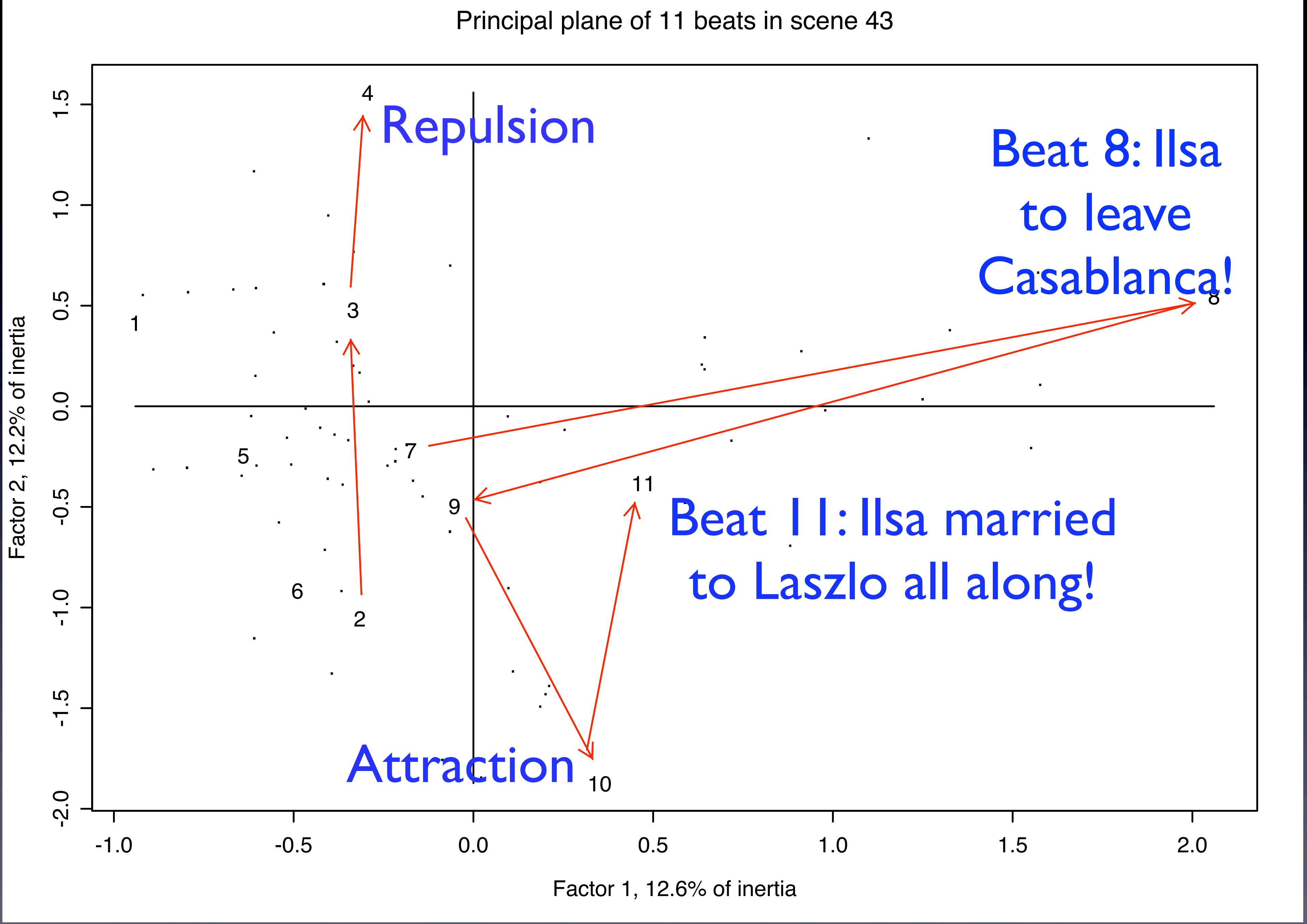
Analysis of Casablanca's "Mid-Act Climax", Scene 43 subdivided into 11 "beats" (subscenes)

- McKee divides this scene, relating to Ilsa and Rick seeking black market exit visas, into 11 "beats"
- Beat 1 is Rick finding Ilsa in the market
- Beats 2, 3, 4 are rejections of him by Ilsa
- Beats 5, 6 express rapprochement by both
- Beat 7 is guilt-tripping by each in turn
- Beat 8 is a jump in content: Ilsa says she will leave Casablanca soon
- In beat 9, Rick calls her a coward, and Ilsa calls him a fool
- In beat 10, Rick propositions her
- In beat 11, the climax, all goes to rack and ruin: Ilsa says she was married to Laszlo all along. Rick is stunned



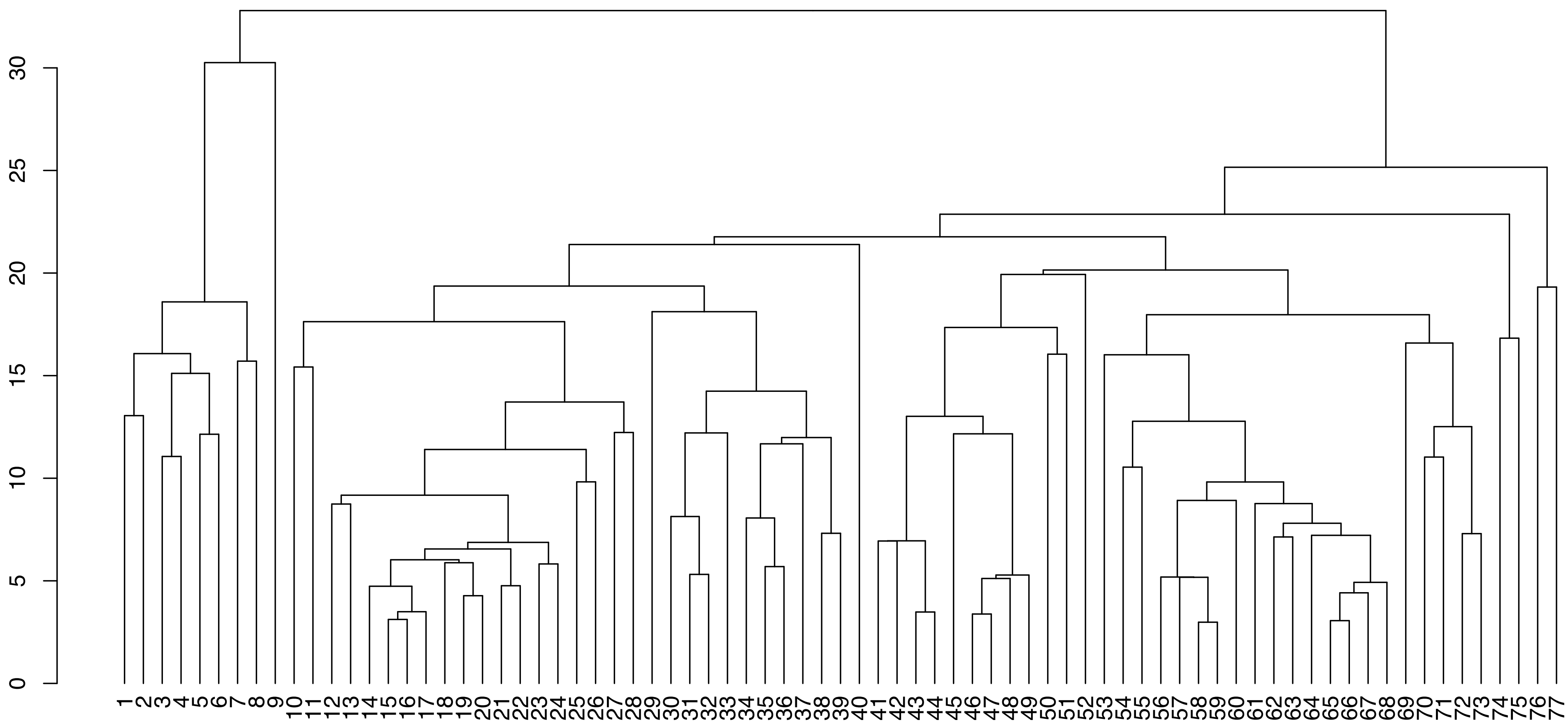
210 words used in these 11 “beats” or subscenes


Principal plane of 11 beats in scene 43
Factor 1, 12.6% of inertia
Factor 2, 12.2% of inertia
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
1
2
3
4
5
6
7
8
9
10
11


Principal plane of 11 beats in scene 43
Repulsion
Beat 8: Ilsa to leave Casablanca!
Beat 11: Ilsa married to Laszlo all along!
Attraction
1
2
3
4
5
6
7
8
9
10
11
Factor 1, 12.6% of inertia
Factor 2, 12.2% of inertia
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

Example: 77 scenes clustered - contiguity or sequence-constrained, complete link hierarchical clustering.

Shows up 9 to 10, and progressing from 39, to 40 and 41, as major changes.



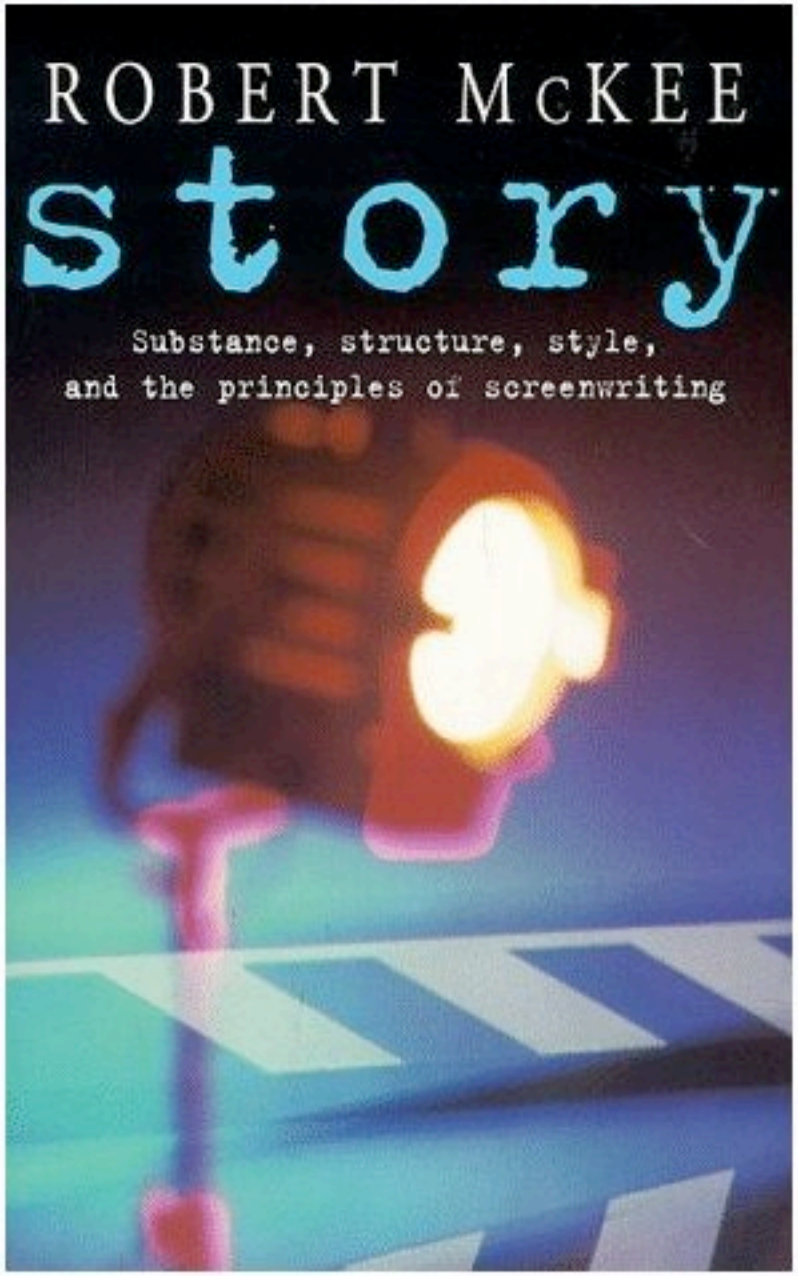
Foregoing example on YouTube (see www.narrativization.com)

- Back to a deeper look at Casablanca
- We have taken comprehensive but qualitative discussion by McKee (following slide) and sought qualitative and algorithmic implementation.
- For McKee: Text is the "sensory surface" of the underlying semantics.

McKee, Methuen, 1999

Casablance is based on a range of miniplots.

McKee: its composition is “virtually perfect”


ROBERT McKEE
story
Substance, structure, style,
and the principles of screenwriting

Style analysis of scene 43 based on McKee Monte Carlo tested against 999 uniformly randomized sets of the beats

- In the great majority of cases (against 83% and more of the randomized alternatives) we find the style in scene 43 to be characterized by:
- small variability of movement from one beat to the next
- greater tempo of beats
- high mean rhythm

Our way of analyzing semantics

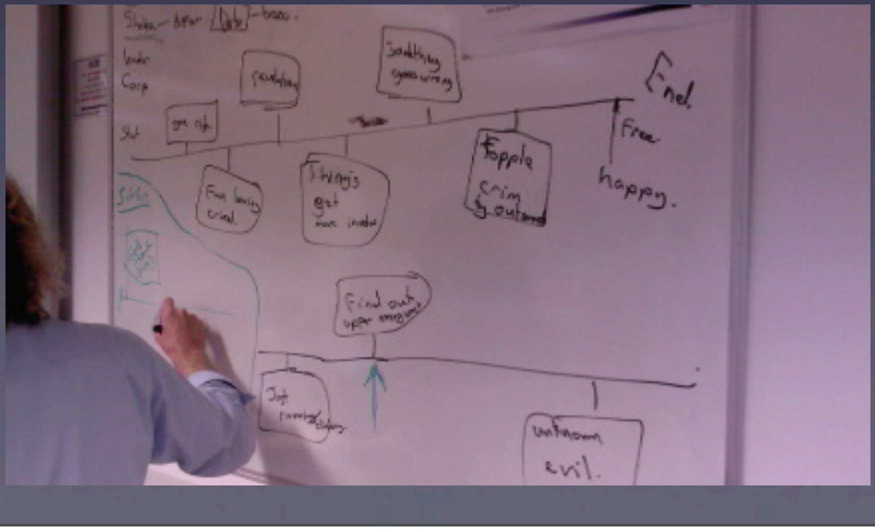
- We discern story semantics arising out of the orientation of narrative
- This is based on the web of interrelationships
- We examined caesuras and breakpoints in the flow of narrative
- With CSI (Crime Scene Investigation - Las Vegas - TV series) scripts: characterization

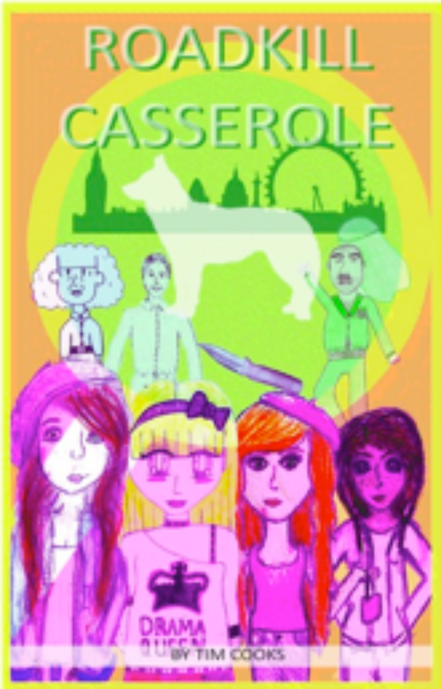
Support environment for collaborative, distributed creating of narrative

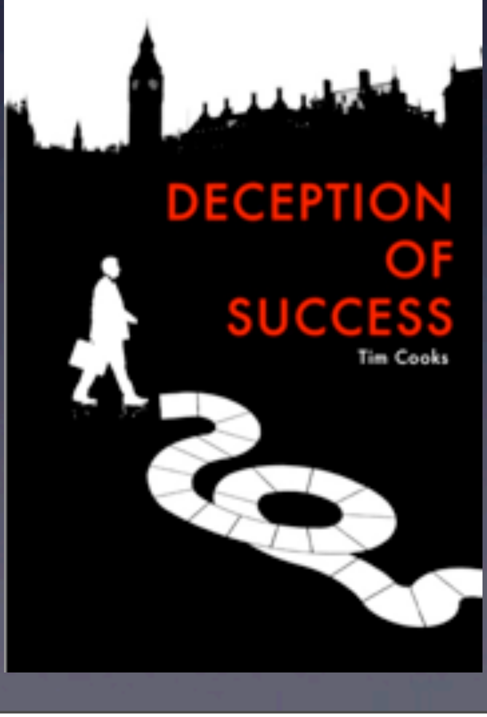
- Pinpointing anomalous sections
- Assessing homogeneity of style over successive iterations of the work
- Scenario experimentation and planning
- This includes condensing parts, or elaborating
- Similarity of structure relative to best practice in chosen genre

“Project TooManyCooks: Applying Software Design Principles to Fiction Writing”
Joe Reddington (RHUL, Comp. Sci.),

Doug Cowie (RHUL, English) and myself







Books written collaboratively using support environment described here.
Upper left: RHUL English students; others: secondary school pupils.
Available for Kindle on Amazon.

Hierarchy, as well as geometry (Euclidean factor space as in Correspondence Analysis) for both understanding and working in complex systems.

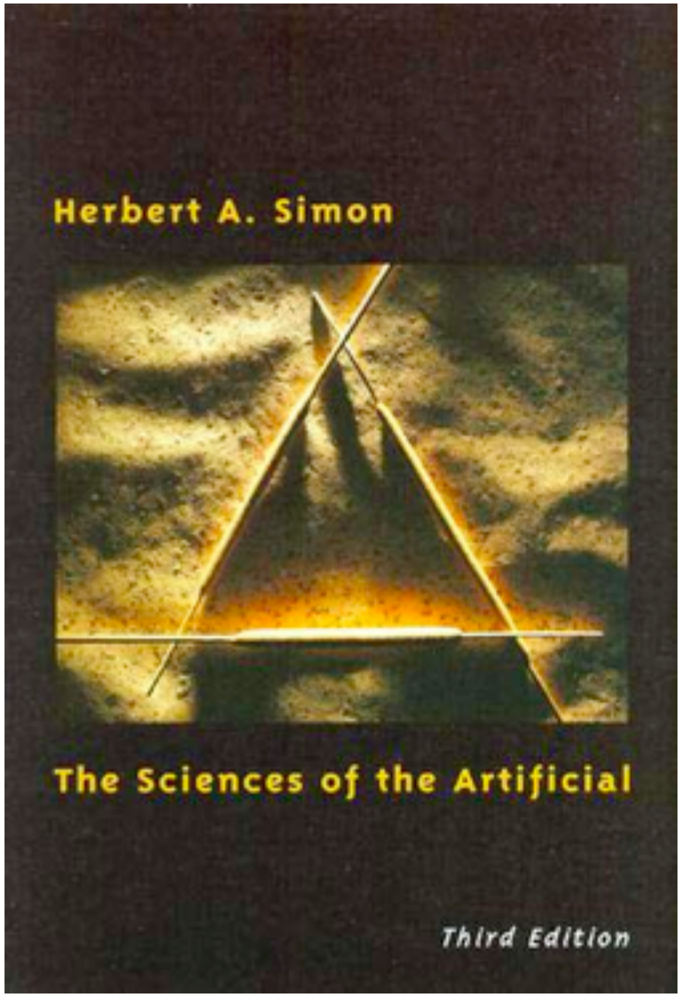
In this presentation: applications to search and discovery, information retrieval, clusterwise regression, knowledge discovery.

Then analysis and synthesis of narrative, using filmscript and literary texts.

- Following slide is of: Herbert A Simon (1916-2001), Nobel Prize in economics 1978. Coined: "bounded rationality", "satisficing", - and hierarchy as the architecture of complex systems. See: *The Sciences of the Artificial*, MIT Press.

Chapter titles include:

- The psychology of thinking
- Remembering and learning
- The science of design
- Social planning
- The architecture of complexity: hierarchic systems

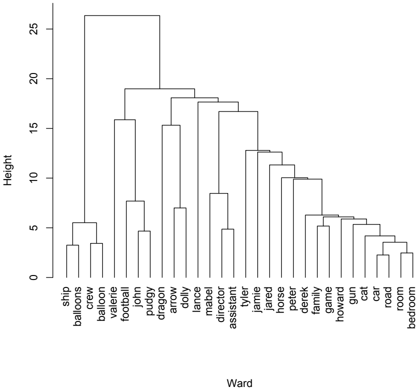


MIT Press, 3rd edn., 1996

Ultrametric Component Analysis: Application to Emotional Content (Or: Determining the Ideas of Andrei Khrennikov.)

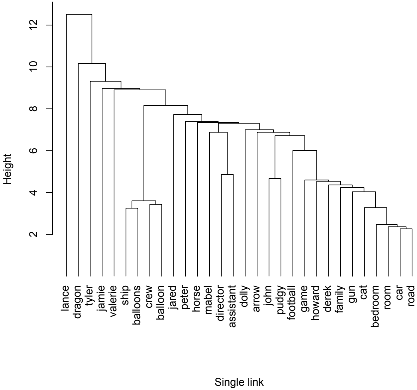
1. Determine the ultrametric parts of a data set, relative to the non-ultrametric, metric (or, indeed, non-metric) parts.
2. I.e., seek isosceles with small base configurations for triplets.
3. To avoid very distant points being considered, start by imposing hierarchical structure, i.e. a hierarchical clustering.
4. In fact, use more than one hierarchical clustering agglomerative criterion, in order to achieve greater – wider – consensus.
5. (1) Two hierarchical clusterings. (2) Their consensus based on analyzing all triplets. (3) Check relative to input data for isosceles with small base configuration.

Ultrametric Consensus – Hierarchy 1



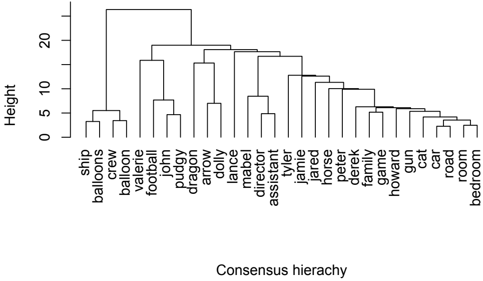
Hierarchical clustering using the Ward minimum variance agglomerative criterion. 30 terms used, in a 139-dimensional Euclidean space (factor space, resulting from Corr. An.).

Ultrametric Consensus – Hierarchy 2



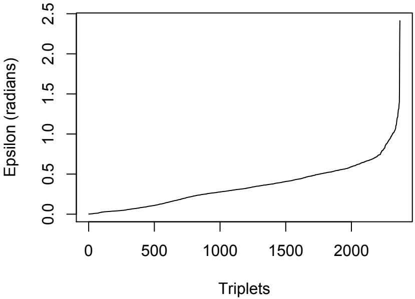
Hierarchical clustering using the single link agglomerative criterion (subdominant ultrametric). 30 terms used, in a 139-dimensional Euclidean space (factor space, resulting from Corr. An.).

Consensus Hierarchy of Hierarchies 1 and 2



Consensus hierarchical clustering using the Ward minimum variance and single link agglomerative criteria, with the Barbara Sanders data. Hence consensus of Hierarchies 1, 2 shown.

Varying ϵ: testing all triplets from consensus hierarchy

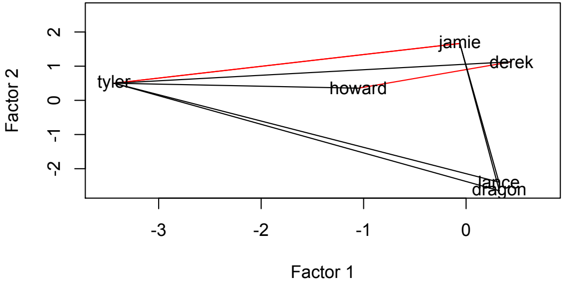


With reference to original points in a Euclidean space (30 points in a 139-dimensional space), how close are they to being inherently ultrametric? Typically we use $\epsilon = 0.034906585$.

Current work

1. Take all, or selected, ultrametrical relations, respecting ϵ-ultrametricity.
2. I have been doing this for dream reports (short texts) from the DreamBank online repository.

Example of three triplets from 163 (from consensus hierarchy)



Uses principal factor plane of Correspondence Analysis (original data were 138-dimensional) Red: small base of isosceles triangle.

Personages

- Howard: her ex-husband, divorced, died suddenly in 1997.
- Derek: a man she had an affair with from 1994, and broke with in 1996.
- Tyler (1980), “co-worker”
- Jamie (1981), “old friend”, “homosexual”
- dragon (1993) (no other names mentioned in this dream about a dragon)
- Lance (1992) “is black and used to be the city manager assistant to the disability rights city group”; “married”.

Marco Tonti, PhD, 2012

The ultrametric understanding of the subconscious is described in these terms:

"...an important idea is the one that sees the unconscious in terms of a topological semantic structure with specific features. If we imagine the experience of a life as encoded in a network of representations of facts, ideas, relations and so on, it can be hypothesized that some specific distance between the elements can be defined. In this novel interpretation [...] brought forward by Lauro-Grotto (called "ultrametric"), the fabric of this network is modified in a way that, in certain circumstances, the distance between a group G of otherwise distinct objects and a third object X is considered to be the same for each of them. If we consider the distance as the probability of going from X to each of the object in G, we should conclude that the probability of going from X to any of the objects in G is the same, i.e. they are structurally considered to be equivalent."

(*Emotions and the Unconscious: Modeling and Measuring the Affective Salience of the Mind*, Università del Selento.)

Application to Humour and Incongruity

T. Veale. *Exploding the Creativity Myth: The Computational Foundations of Linguistic Creativity*. Bloomsbury Academic, 2012.

"Recent work by cognitive scientists Jerey Loewenstein and Chip Heath shows that the AAB pattern in stories – which they call the repetition-break plot structure – is considered more enjoyable by readers than the equivalent AAA (unbroken repetition) or ABC (no repetition) patterns. Many narrative jokes use explicit repetition to enforce an AAA pattern in the minds of an audience, so that AAB repetition-break comes as an incongruous and potentially humorous surprise."

"There are whole genres of jokes involving a priest, a rabbi and an imam; or an Irishman, and Englishman and a Scotsman; or a trio of nuns, hookers, husbands or some other stock characters, in which two of the three act somewhat predictably while the zany actions of the third provide the humorous departure."

Status of This Work

1. Use text (rather than e.g. speech).
2. I have started with poems, where each individual line in succession is considered.
3. I consider the sequence-constrained hierarchy.
4. Then I look for two successive semantically close lines followed by a semantically distant third line in the sequence.
5. Results on a number of versions of a ballad at:
   http://thames.cs.rhul.ac.uk/∼fionn/lmaisry
   (user name and password: lmaisry l12maisry3)