Towards a New Science of Big Data Analytics, based on the Geometry and Topology of Complex, Hierarchic Systems

Fionn Murtagh
fmurtagh@acm.org

EGC 2013 Toulouse, 30 Jan. - 1 Feb. 2013

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

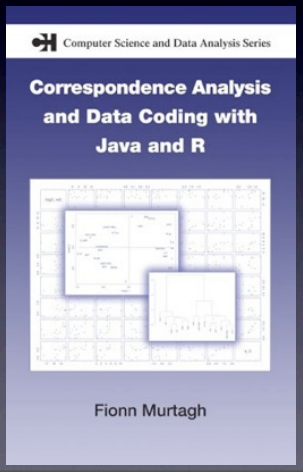
- Euclidean geometry for semantics of information.
- Hierarchical topology for other aspects of semantics, and in particular how a hierarchy expresses anomaly or change. A further useful case is when the hierarchy respects chronological or other sequence information.

Correspondence Analysis is A Tale of Three Metrics

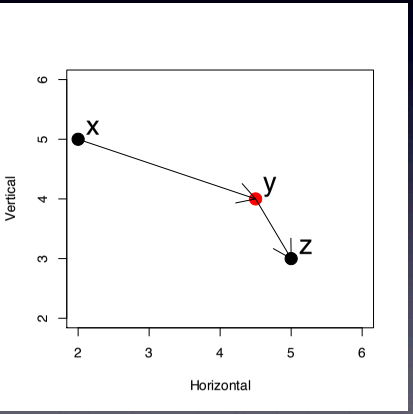
– Chi squared metric – appropriate for profiles of frequencies of occurrence

– Euclidean metric, for visualization, and for static context

– Ultrametric, for hierarchic relations and for dynamic context



Triangular inequality holds for metrics

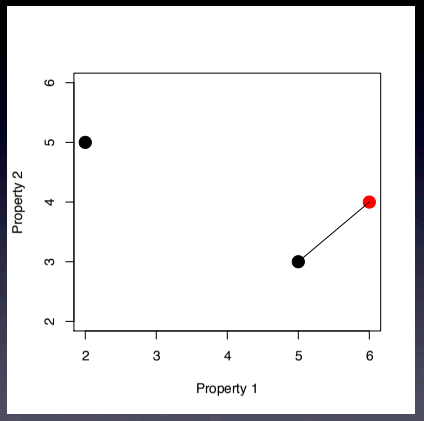


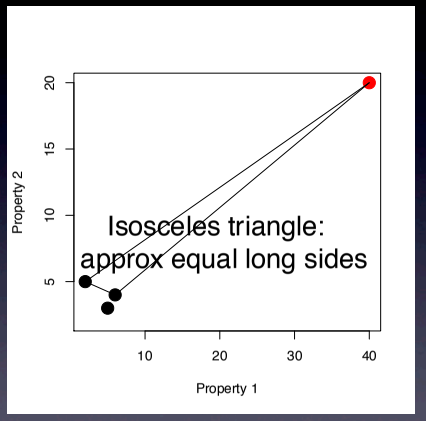
Example: Euclidean or "as the crow flies" distance

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

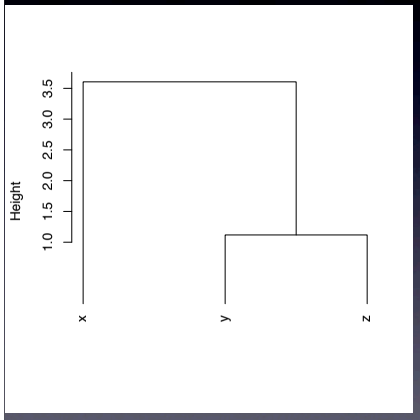
Ultrametric

- 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





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

Analysis of semantics: 1. Context - the collection of all interrelationships

- Euclidean distance makes a lot of sense when the population is homogeneous
- All interrelationships together provide context, relativities - and meaning

Analysis of semantics: 2. Hierarchy tracks anomaly and change

- 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

What is so special about hierarchy?

- Ultrametric spaces have interesting properties.
- Not just in data analysis and pattern recognition, but in physics at small scales, and in optimization.
- Ultrametric topology and p-adic number systems are closely associated.
- Next I will look at:
- (i) Quantifying inherent ultrametricity.
- (ii) Computational implications.

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 center; (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 – I

- Assume Euclidean space. Consider a triplet of points, that defines a triangle.
- Take smallest internal angle, a, in triangle ≤ 60 deg.
- … and, for the two other internal angles, b and c, if $| b - c | < 2$ deg. (arbitrary small angle),
- Then this triangle is ultrametric.
- We look for the overall proportion of such triangles in our data.

Quantifying ultrametricity – II

- So: we take all possible triplets, i, j, k
- We look at their angles, and judge whether or not the ultrametric triangle properties are verified
- If so: #UM-triangles++
- Having examined all possible triangles, our α measure is: #UM-triangles / #triangles
- All triangles respect these ultrametric properties implies $\alpha = 1$; no triangle does, then = 0
- For n objects, this is computationally prohibitive, so we sample i,j,k in practice (uniformly)

Other Ways of Quantifying Ultrametricity – III

- Relationship between subdominant ultrametric, and given dissimilarities.
- Rammal, Toulouse and Virasoro, Ultrametricity for physicists, Rev. Mod. Phys., 58, 765-788, 1986.
- Whether interval between median and max rank dissimilarity of every set of triplets is nearly empty. (Taking ranks provides scale invariance.)
- Lerman, Classification et Analyse Ordinale des Données, Dunod, 1981.

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; then: "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.

– References.

– 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:

$d_B(x, y) = 10^{-2}$

Base ($_B$) 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≠7

On the Baire (ultra)metric

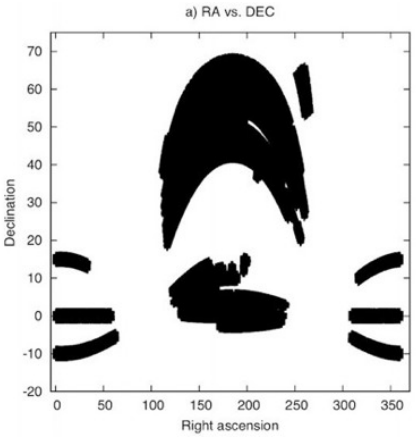
– Baire space consists of countable infinite sequences with a metric defined in terms of the longest common prefix *[A. Levy. Basic Set Theory, Dover, 1979 (reprinted 2002)]*

– 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.

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.

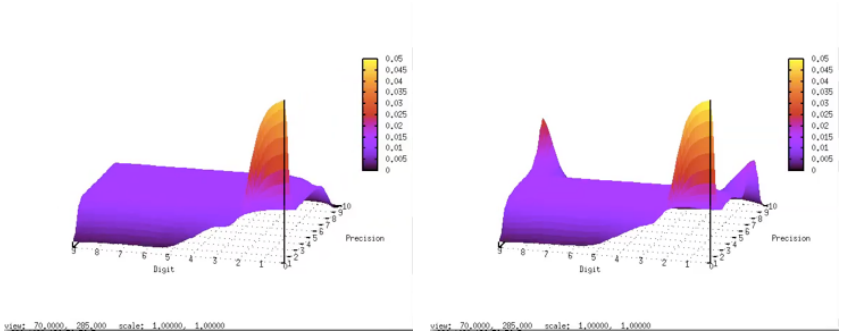
- Dec vs RA are shown in the figure.

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



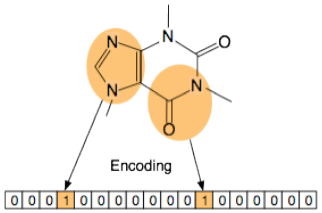
On the left we have z_spec where three data peaks can be observed.
On the right we have z_phot where only one data peak can be seen.

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?

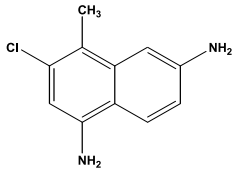
Baire Distance Applied to Chemical Compounds



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.

- 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

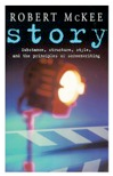
- We must consider complex web of relationships.
- Semantics include web of relationships - thematic structures and patterns. Structures and interrelationships evolve in time.
- Semantics include time evolution of structures and patterns, including both: threads and commonality; and change, the exceptional, 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.

Topics

- Casablanca - analysis of emotion in scene 43
- Two senses of semantics - all interrelationships, and change over time
- Associate these, resp., with Euclidean metric and ultrametric
- Text attributes and their significance (and feasibility of mapping these onto desired outputs)
- Characterizing structure and properties of CSI television episodes
- Text synthesis - supporting collaborative narrative construction - book writing

Movie Casablanca shot by Warner Brothers between May and August 1942

For McKee, composition of Casablanca is "virtually perfect".



- Scene 43 in Casablanca (out of 77 scenes).
- Crucial mid-point scene. Following McKee, I will analyze 11 subscenes ("beats").
- Right, first three subscenes (in blue, brown, red).

EXT. BLACK MARKET - DAY

At the linen stall, Ilsa examines a tablecloth which an Arab vendor is endeavoring to sell. He holds a sign which reads "700 francs."

ARAB
You will not find a treasure like this in all Morocco, Mademoiselle. Only seven hundred francs.

Rick walks up behind Ilsa.

RICK
You're being cheated.

She looks briefly at Rick, then turns away. Her manner is politely formal.

ILSA
It doesn't matter, thank you.

ARAB
Ah, the lady is a friend of Rick's? For friends of Rick we have a small discount. Did I say seven hundred francs? You can have it for two hundred.

Reaching under the counter, he takes out a sign reading "200 francs", and replaces the other sign with it.

RICK
I'm sorry I was in no condition to receive you when you called on me last night.

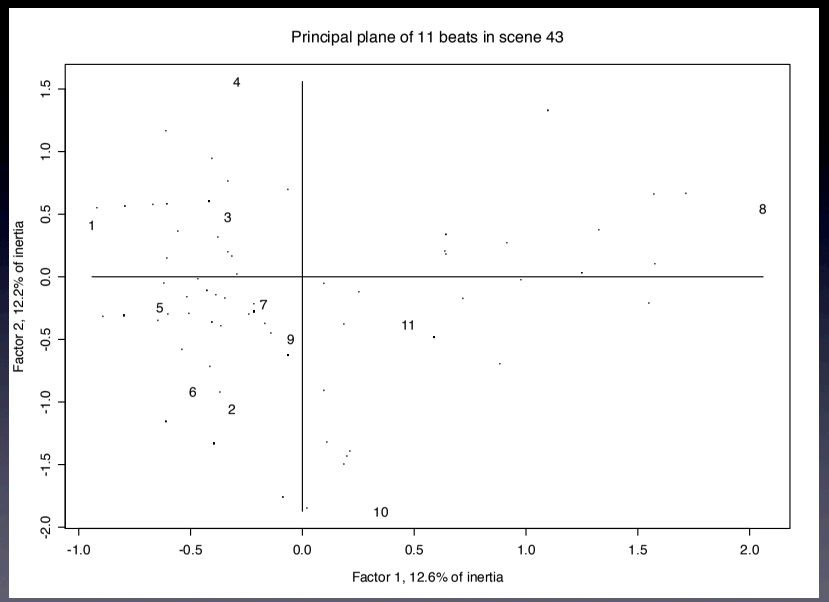
ILSA
It doesn't matter.

ARAB
Ah, for special friends of Rick's we have a special discount. One hundred francs.

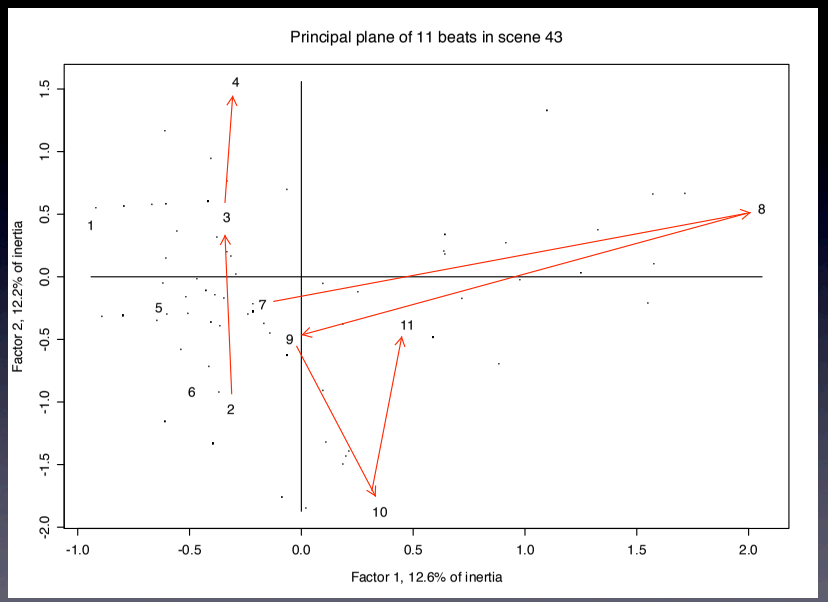
He replaces the second sign with a third which reads "100 francs."

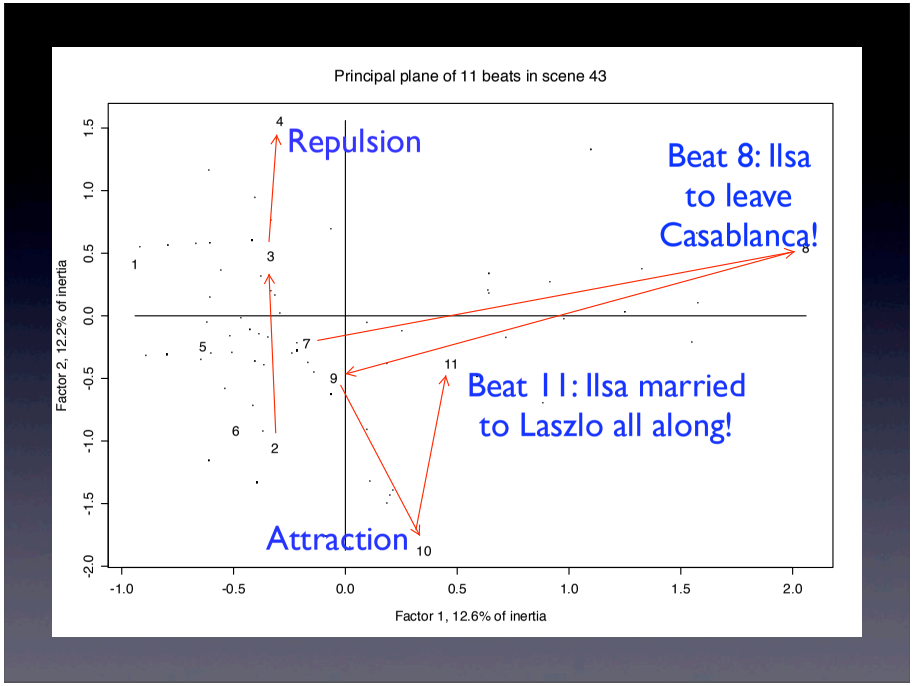
Analysis of Casablanca's "Mid-Act Climax", Scene 43 subdivided into 11 "beats" (subscenes)

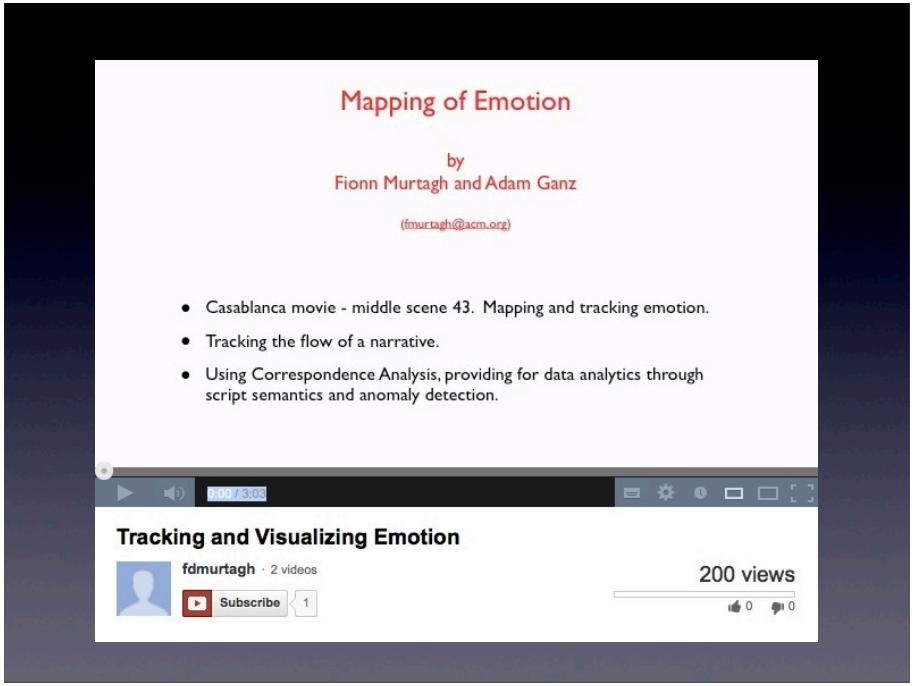
- McKee divides this scene, relating to Ilsa and Rick seeking black market exit visas, into 11 "beats"
- Beat 1 is Rick finding Ilsa in the market
- Beats 2, 3, 4 are rejections of him by Ilsa
- Beats 5, 6 express rapprochement by both
- Beat 7 is guilt-tripping by each in turn
- Beat 8 is a jump in 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



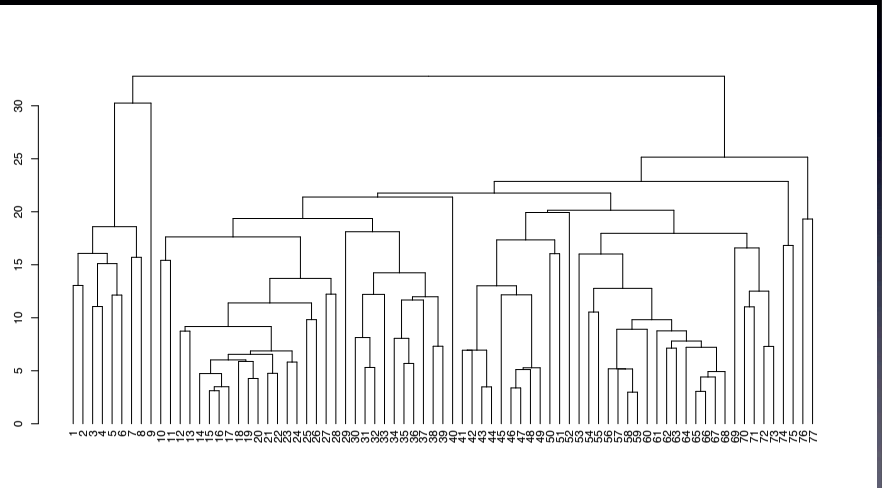
210 words used in these 11 "beats" or subscenes







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.



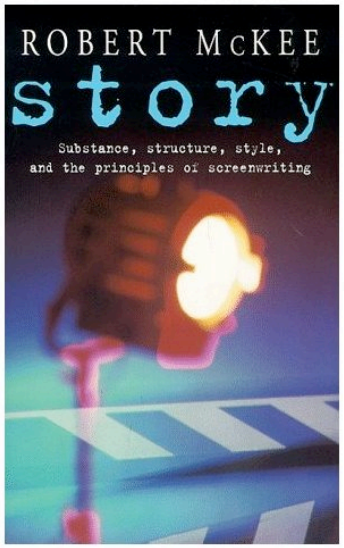
- Back to a deeper look at Casablanca
- We have taken comprehensive but qualitative discussion by McKee and sought qualitative and algorithmic implementation

McKee, Methuen, 1999

Casablance is based on a range of miniplots.

McKee: its composition is "virtually perfect"

Text is the "sensory surface" of the underlying semantics



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

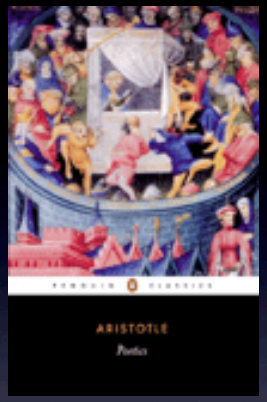
Our way of analyzing semantics

- 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 scripts: characterization

CSI: Crime Scene Investigation

Transcripts of 3 episodes first aired by CBS Oct. 2000

Text Synthesis

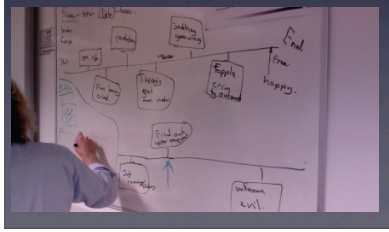


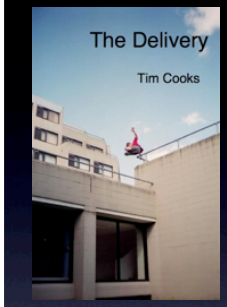
- Aristotle's Poetics (c. 350 BC)
- "Outlines and episodization" - "Stories ... should first be set out in universal terms ... on that basis, one should then turn the story into episodes and elaborate it."
- "... reasoning is the speech which the agents use to argue a case or put forwards an opinion"

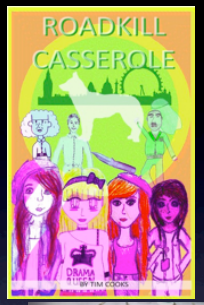
Support environment for collaborative, distributed creating of narrative

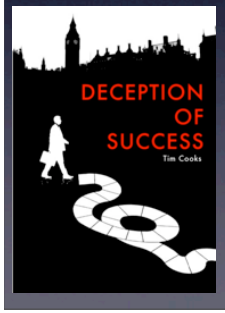
- Pinpointing anomalous sections
- Assessing homogeneity of style over successive iterations of the work
- Scenario experimentation and planning
- This includes condensing parts, or elaborating
- Similarity of structure relative to best practice in chosen genre

"Project TooManyCooks: Applying Software Design Principles to Fiction Writing"
Joe Reddington (RHUL, Comp. Sci.),
Doug Cowie (RHUL, English) and myself

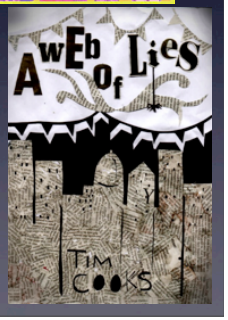








Books written collaboratively using support environment described here.
Upper left: RHUL English students; others: secondary school pupils.
Available for Kindle on Amazon.

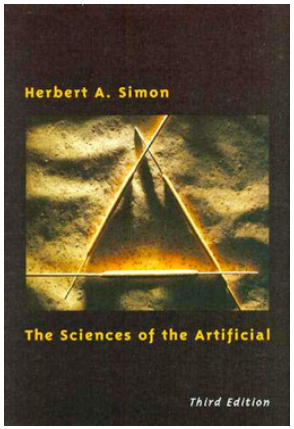


Hierarchy, as well as geometry (Euclidean factor space as in Correspondence Analysis) for both understanding and working in complex systems.
In this presentation: applications to search and discovery, information retrieval, clusterwise regression, knowledge discovery.
Then analysis and synthesis of narrative, 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

Thank You

- My collaborators:
- Pedro Contreras
- Adam Ganz
- Joe Reddington

References

- F. Murtagh, Correspondence Analysis and Data Coding with R and Java, Chapman and Hall/CRC Press, 2005. See chapter 5 on text analysis. Software at www.correspondances.info
- F. Murtagh, A. Ganz, S. McKie, J. Mothe and K. Englmeier, "Tag Clouds for Displaying Semantics: The Case of Filmscripts", Information Visualization Journal, forthcoming. 9, 253-262, 2010.
- F. Murtagh, "The Correspondence Analysis platform for uncovering deep structure in data and information", Sixth Boole Lecture, Computer Journal, 53, 304-315, 2010.
- F. Murtagh, A. Ganz and S. McKie, "The structure of narrative: the case of film scripts", Pattern Recognition, 42, 302-312, 2009. (See discussion in Z. Merali, "Here's looking at you, kid. Software promises to identify blockbuster scripts.", Nature, 453, p. 708, 4 June 2008.)
- F. Murtagh, A. Ganz and J. Reddington, "New methods of analysis of narrative and semantics in support of interactivity", Entertainment Computing, 2, 115-121 2011.

References

– F. Murtagh, "Thinking ultrametrically", in D. Banks, L. House, F.R. McMorris, P. Arabie and W. Gaul, Eds., **Classification, Clustering, and Data Mining Applications**, Springer, 3-14, 2004.

– F. Murtagh, "On ultrametricity, data coding, and computation", **Journal of Classification**, 21, 167-184, 2004.

– F. Murtagh, "Identifying the ultrametricity of time series", **European Physical Journal B**, 43, 573-579, 2005.

– F. Murtagh, G. Downs and P. Contreras, "Hierarchical clustering of massive, high dimensional data sets by exploiting ultrametric embedding". **SIAM Jnl. on Scientific Computing**, Vol. 30, No. 2, pp. 707–730. February 2008.

– P. Contreras and F. Murtagh. "Fast, linear time hierarchical clustering using the Baire metric". **Journal of Classification**, 29, 118-143, 2012.

– F. Murtagh and P. Contreras, "Fast, linear time, m-adic hierarchical clustering for search and retrieval using the Baire metric, with linkages to generalized ultrametrics, hashing, formal concept analysis, and precision of data measurement", **p-Adic Numbers, Ultrametric Analysis and Applications**, 4, 45-56, 2012.

– P. Contreras and F. Murtagh, "Linear time Baire hierarchical clustering for enterprise information retrieval", **International Journal of Software and Informatics**, 6, 363-380, 2012.