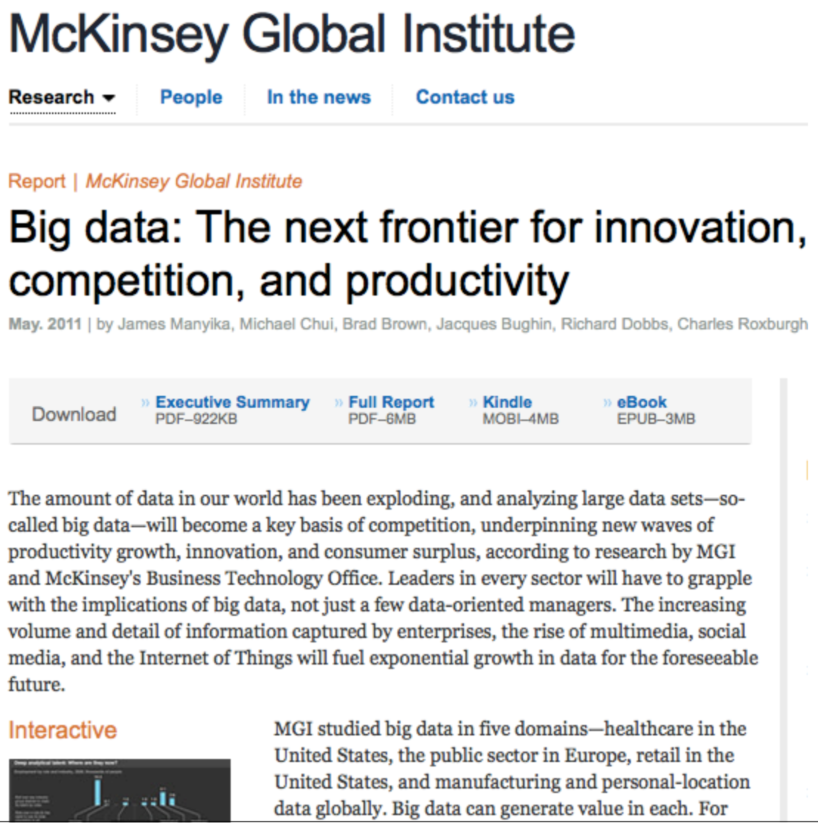
Lecture 3

The Future of Search and Discovery in Big Data Analytics: Ultrametric Information Spaces

Themes

1) “Big Data” and analytics: the potential for metric (geometric) and ultrametric (topological) analysis.
2) Baire distance, ultrametric and hierarchy, applied to astronomy data.
3) Chemoinformatics application: first, clustering and data analysis through modifying precision of the data; secondly, Baire distance, making use of random projections.
4) Finally, best match (nearest neighbour) searching using heuristics can be seen to be “stretching” the data in order to be ultrametric.



McKinsey Global Institute

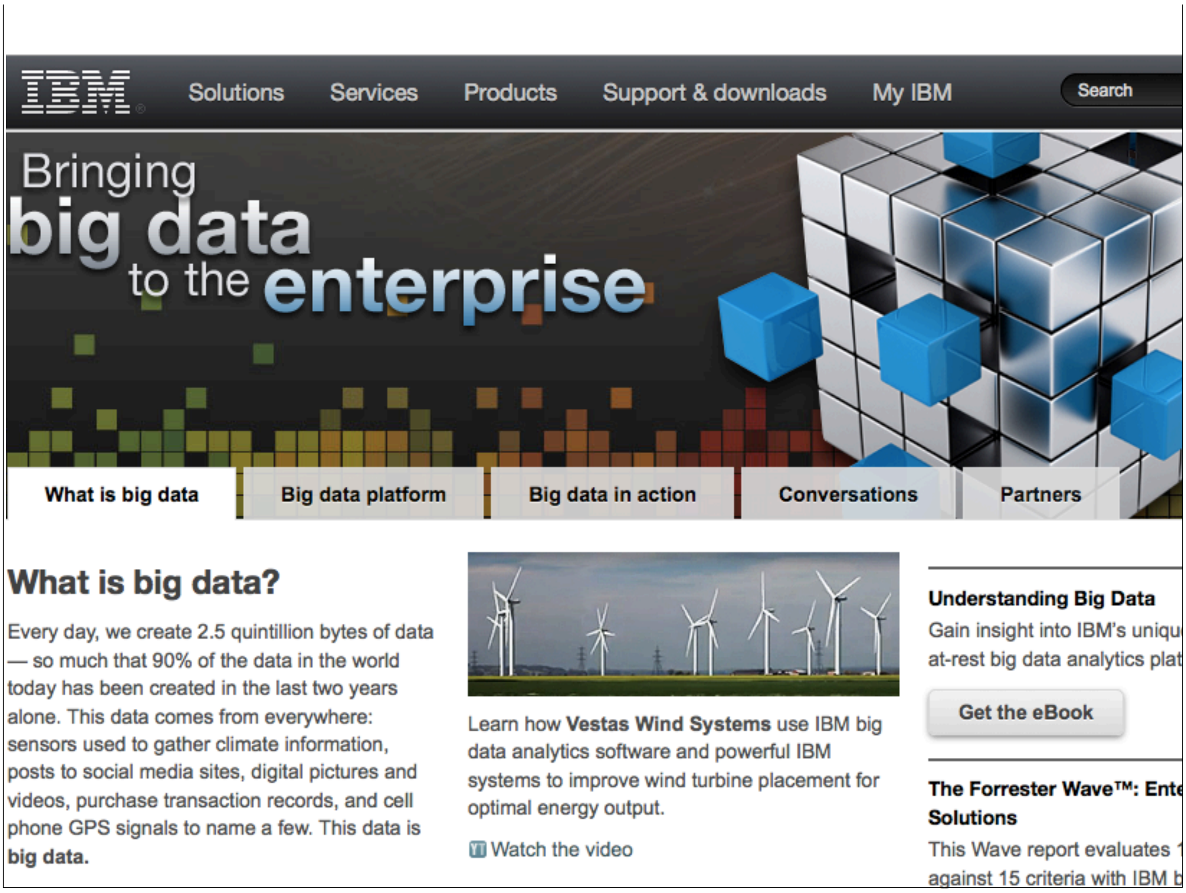
Report | *McKinsey Global Institute*

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



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

Overview

– 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 spectroscopic data.

–Hierarchical clustering via Baire distance using chemical compounds.

– Finally, understanding some other approaches to nearest neighbour or best match searching in terms of ultrametric "stretching".

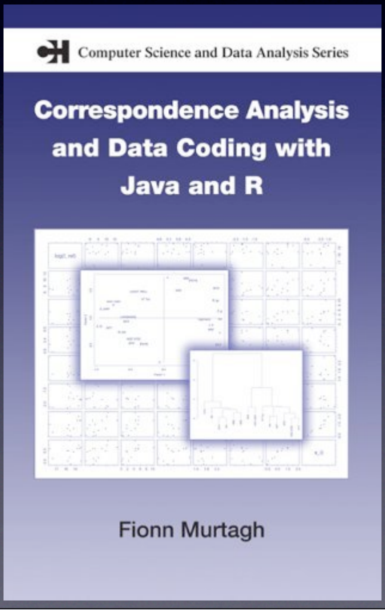
- Comments follow on ultrametrics - “distances on a hierarchical tree”.
- Ultrametric topology captures well the idea of novelty, exception, new.
- We are often interested in hierarchical clustering not just for deriving a partition (of compact clusters) but rather for more general information encapsulated in the hierarchy.

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

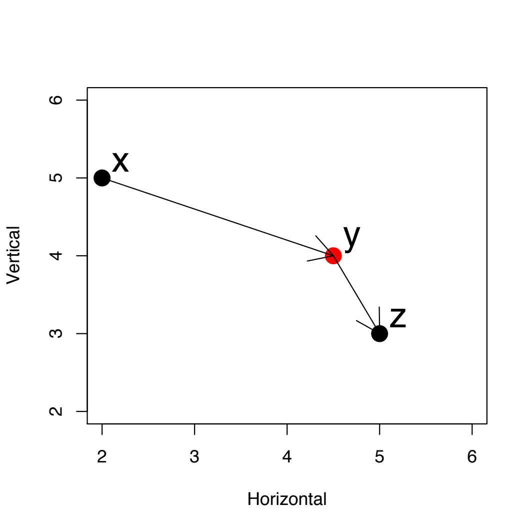


Computer Science and Data Analysis Series

Correspondence Analysis and Data Coding with Java and R

Fionn Murtagh

Triangular inequality holds for metrics

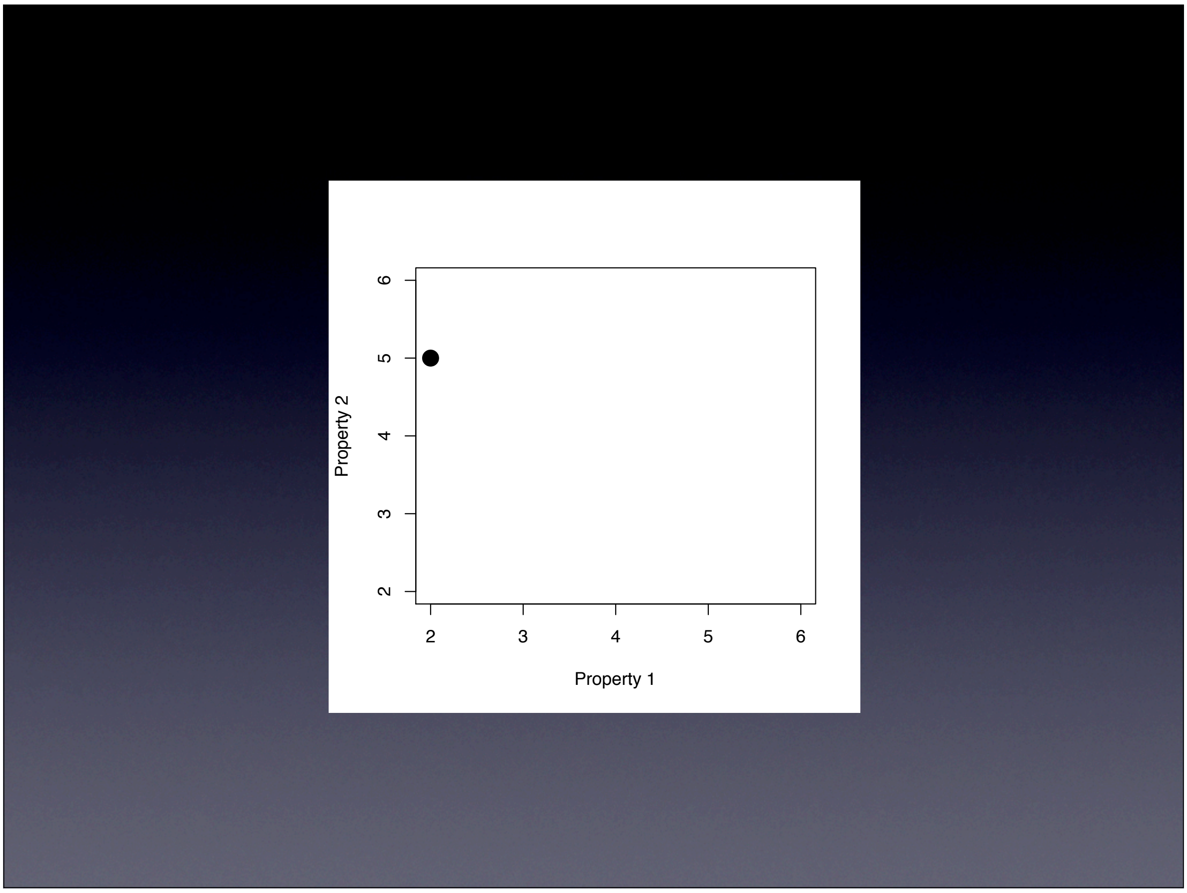


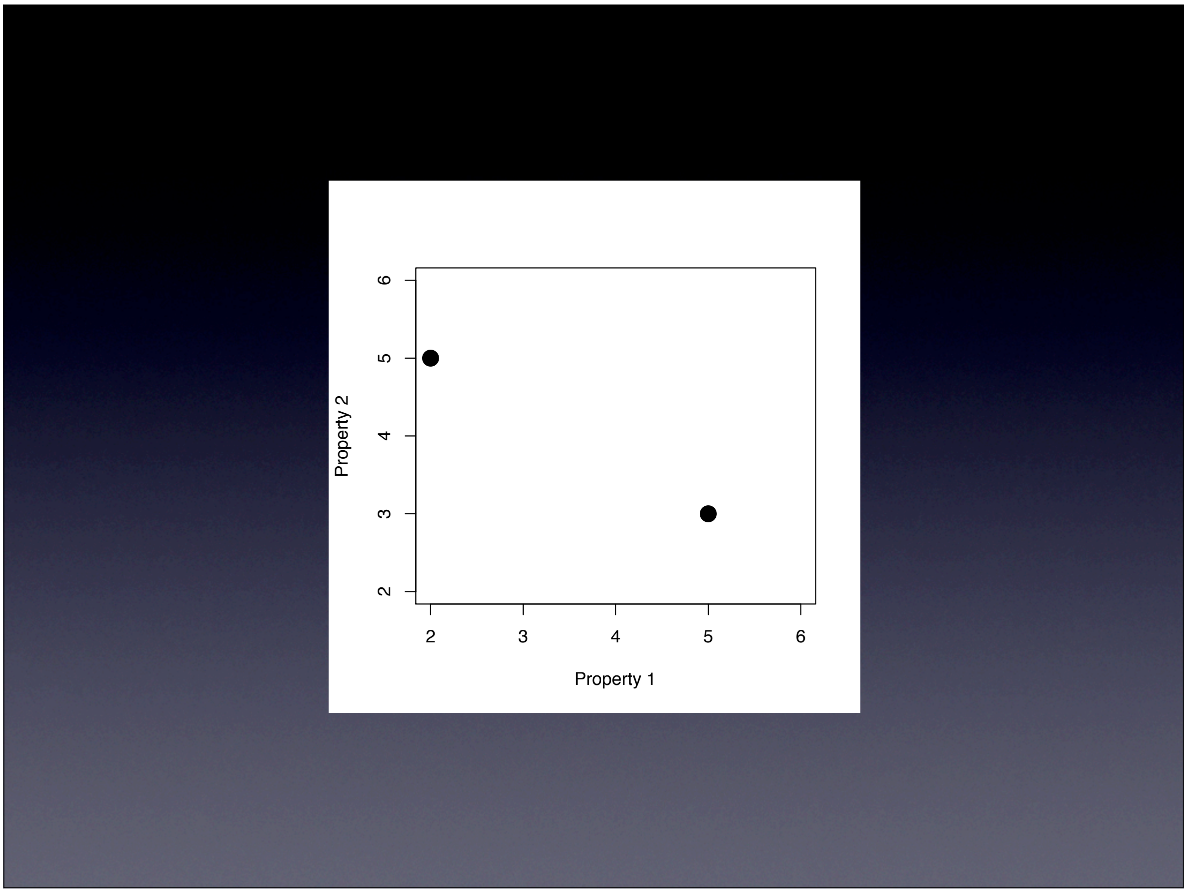
Example: Euclidean or "as the crow flies" distance

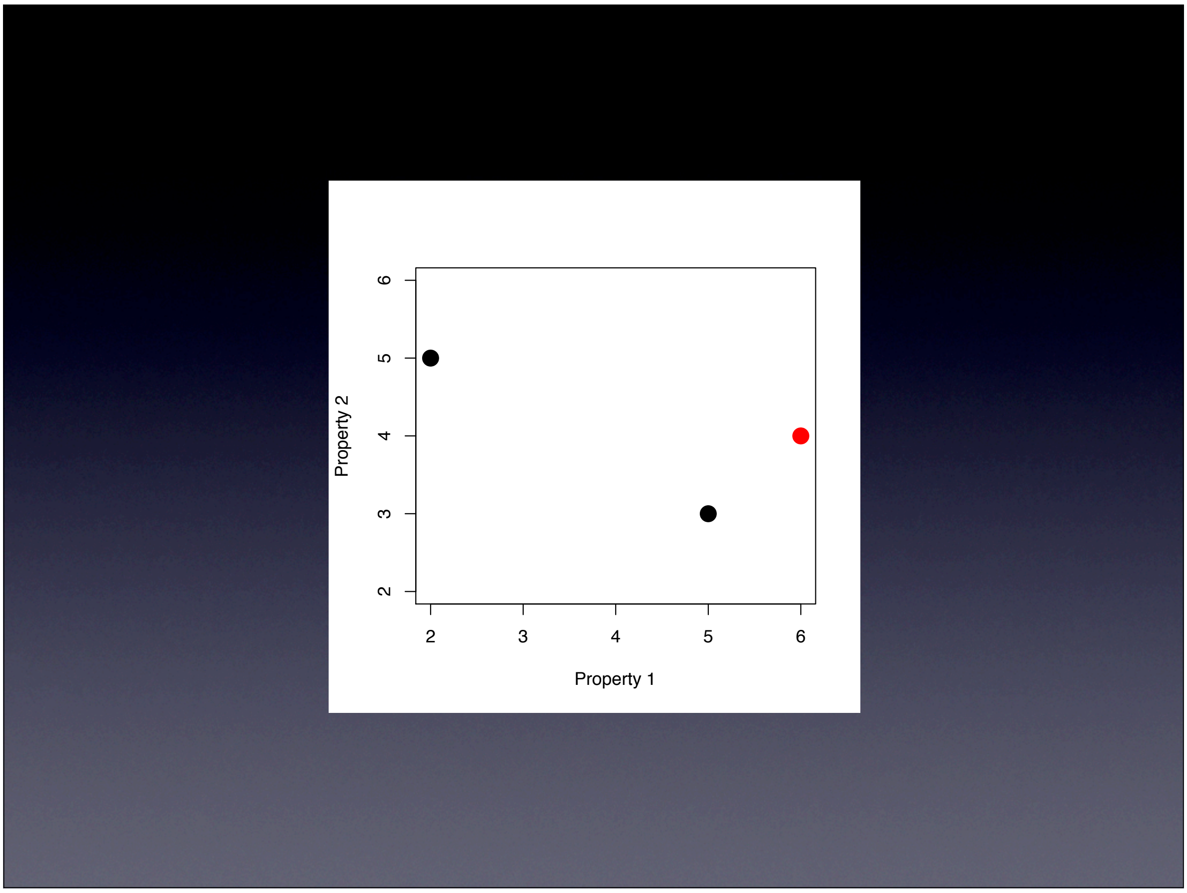
$$d(x, z) \leq d(x, y) + d(y, z)$$

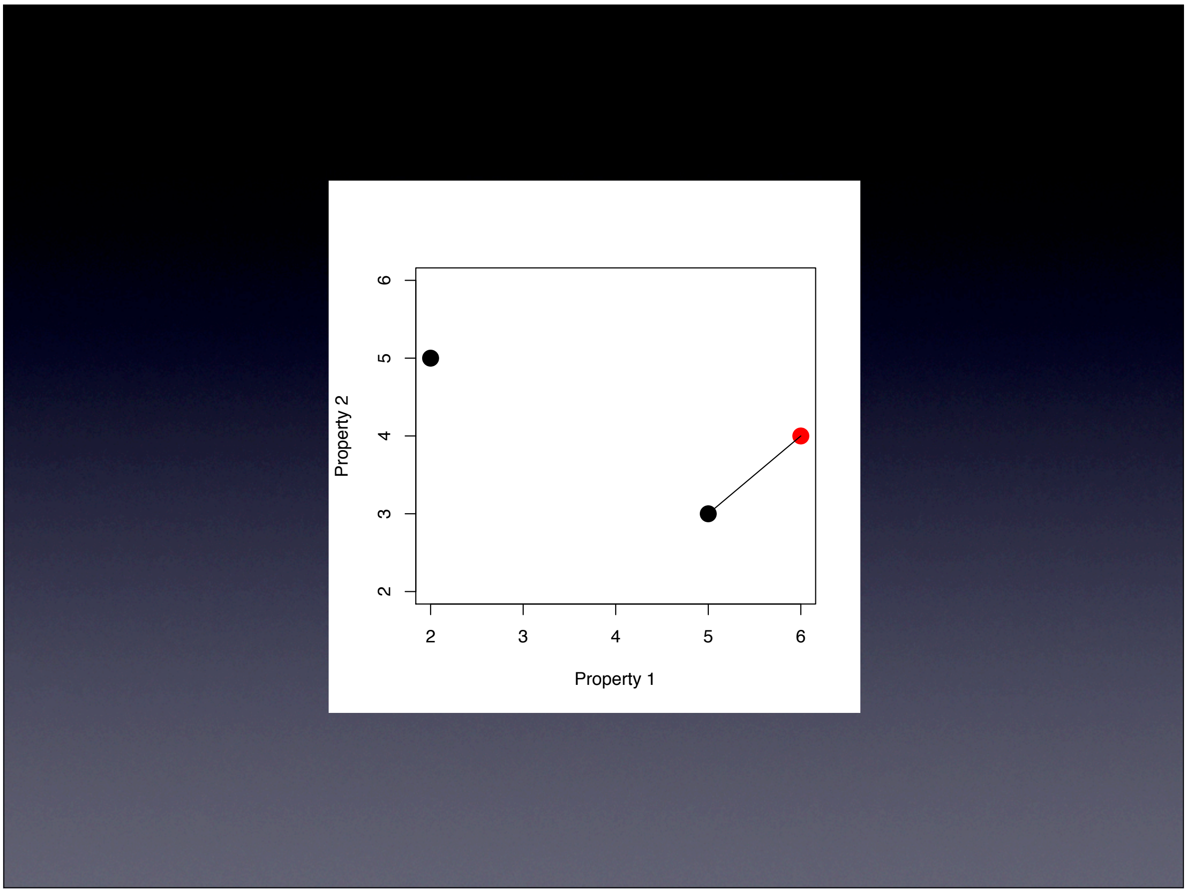
Ultrametric

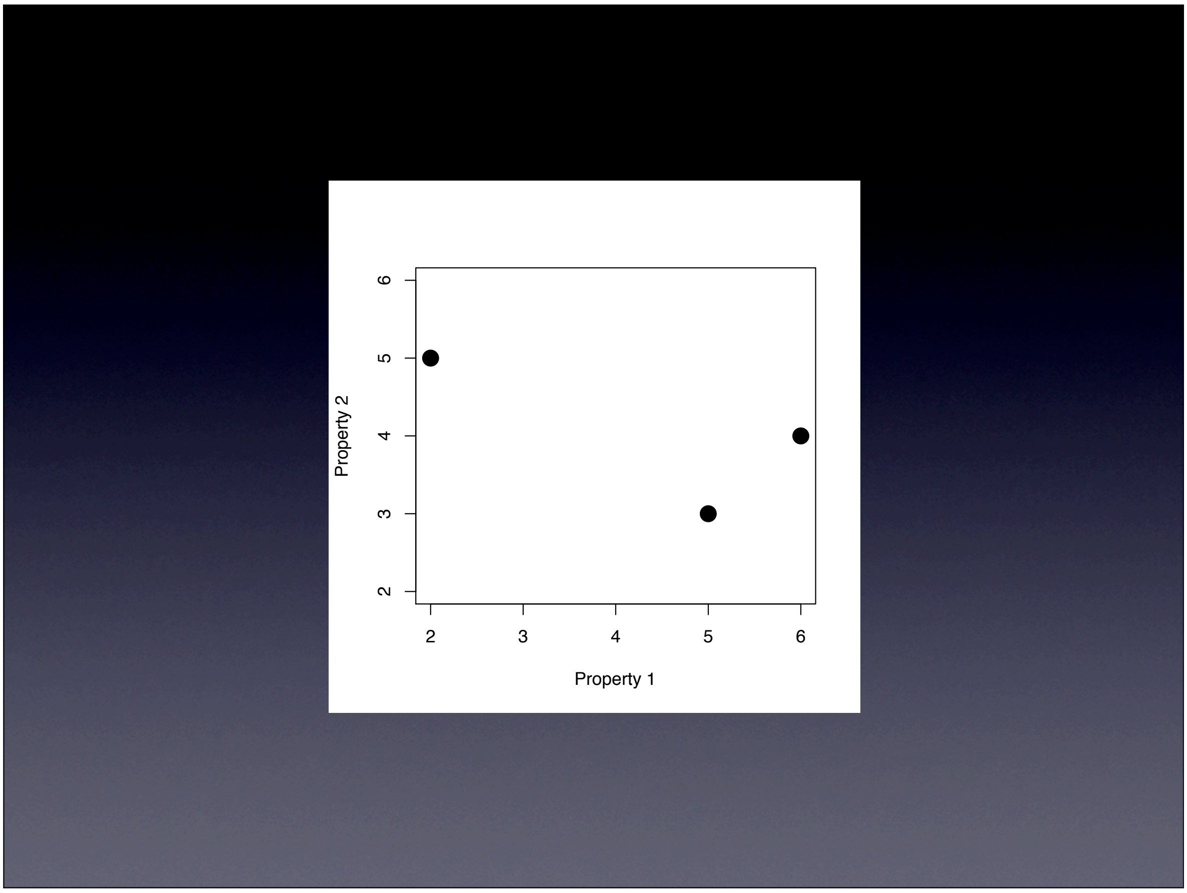
- Euclidean distances makes a lot of sense when the population is homogeneous
- Ultrametric distance makes a lot of sense when the observables are heterogeneous, discontinuous
- Latter is especially useful for determining: anomalous, atypical, innovative cases


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

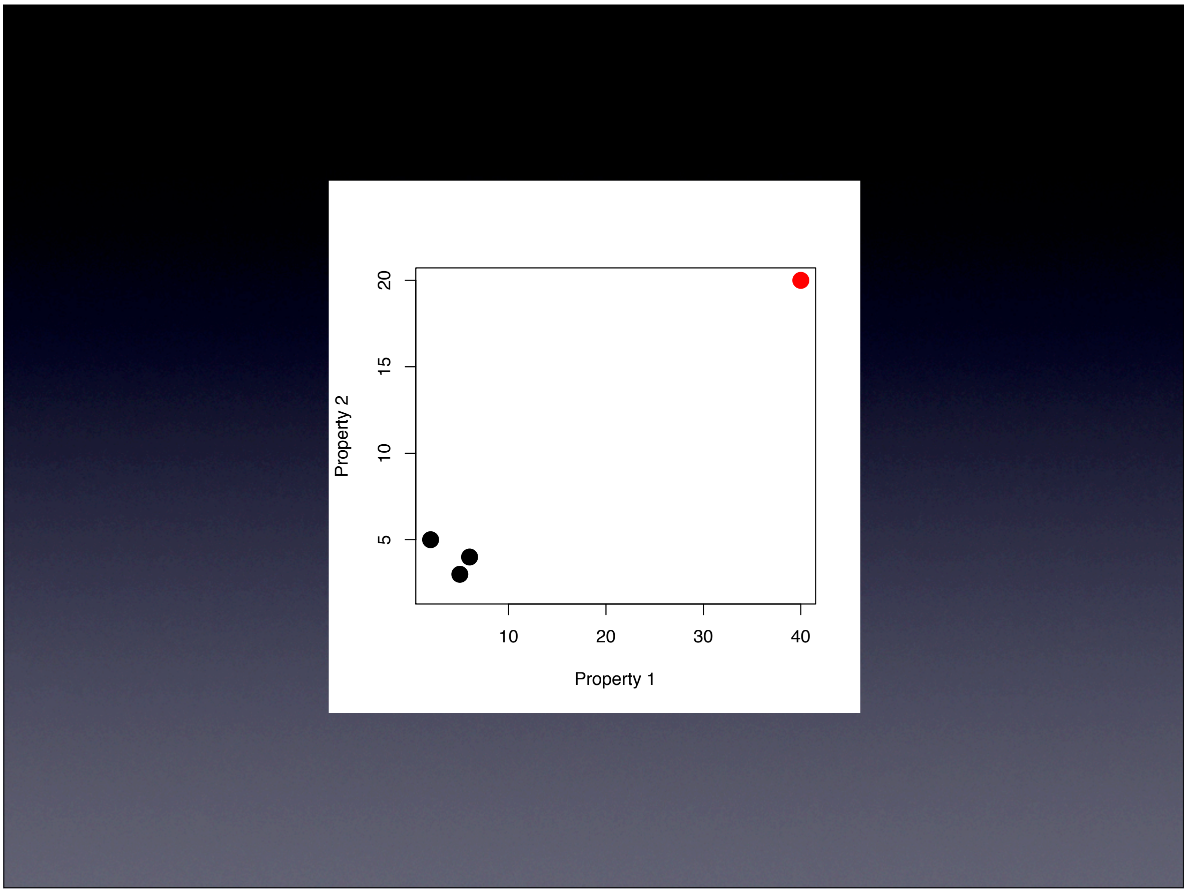

6
5
4
3
2
Property 2
2
3
4
5
6
Property 1


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

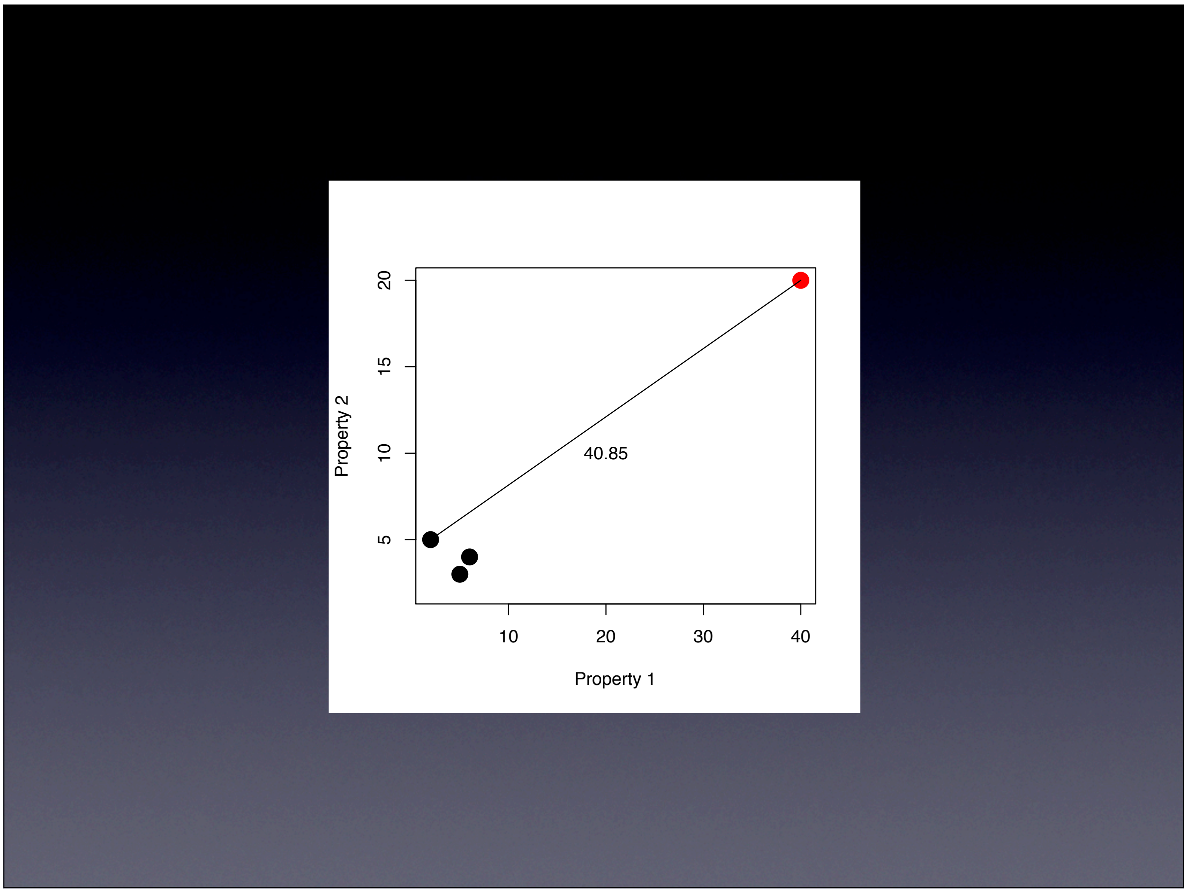

Property 2
Property 1
2
3
4
5
6
2
3
4
5
6

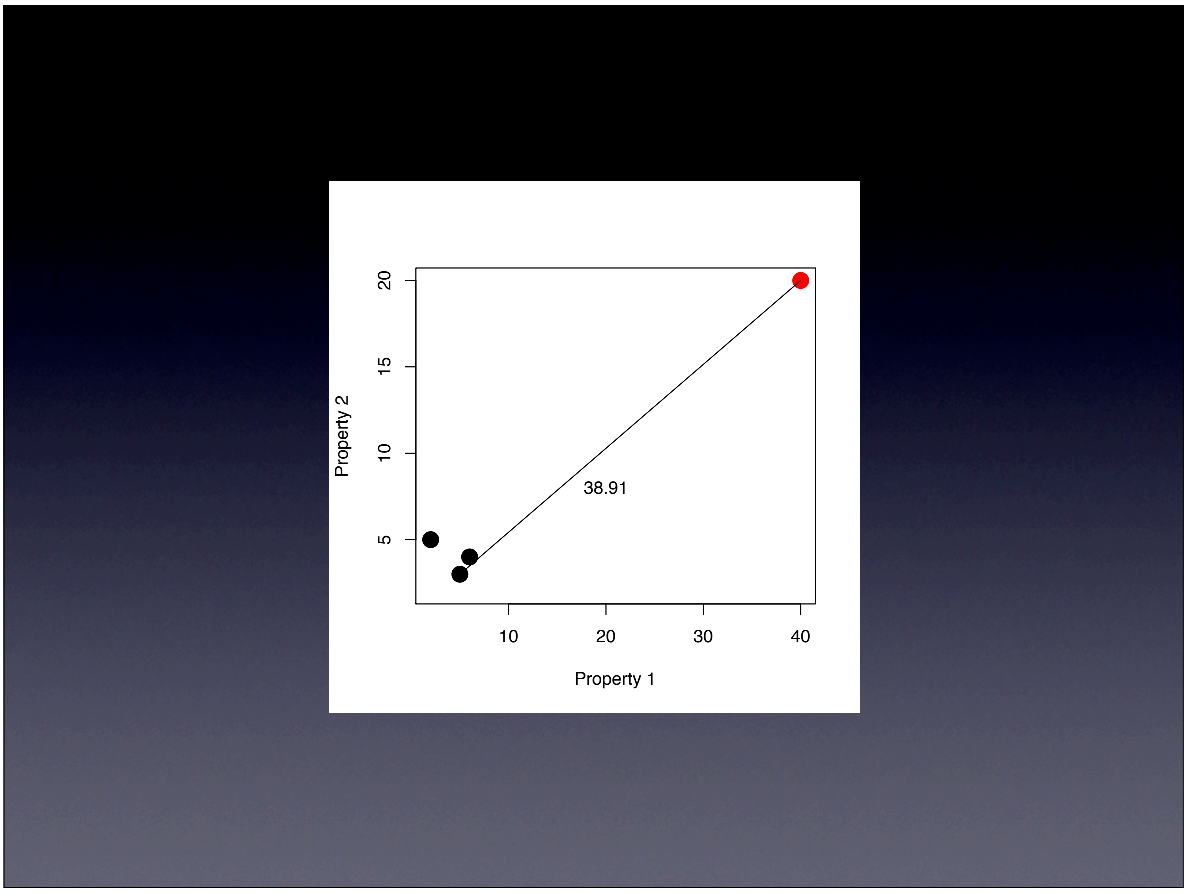


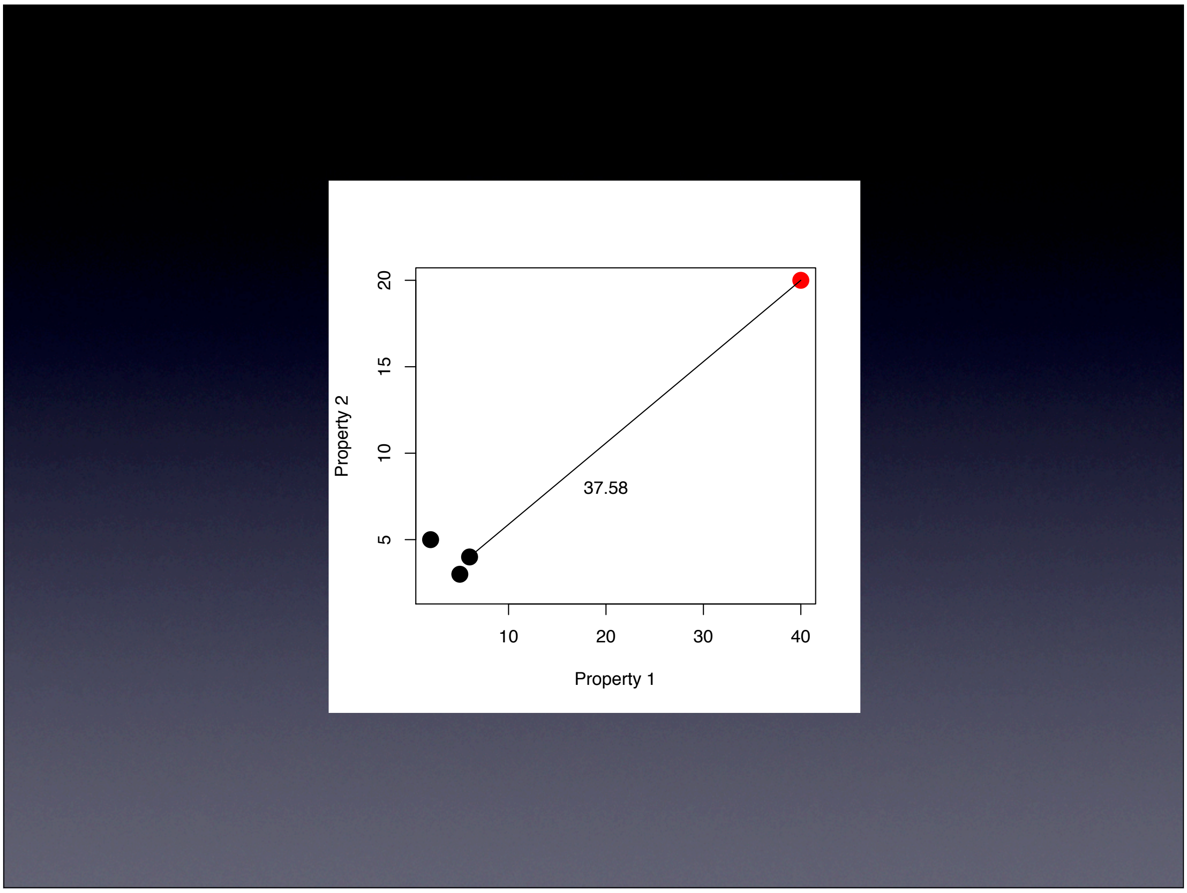
Property 2
6
5
4
3
2
2
3
4
5
6
Property 1

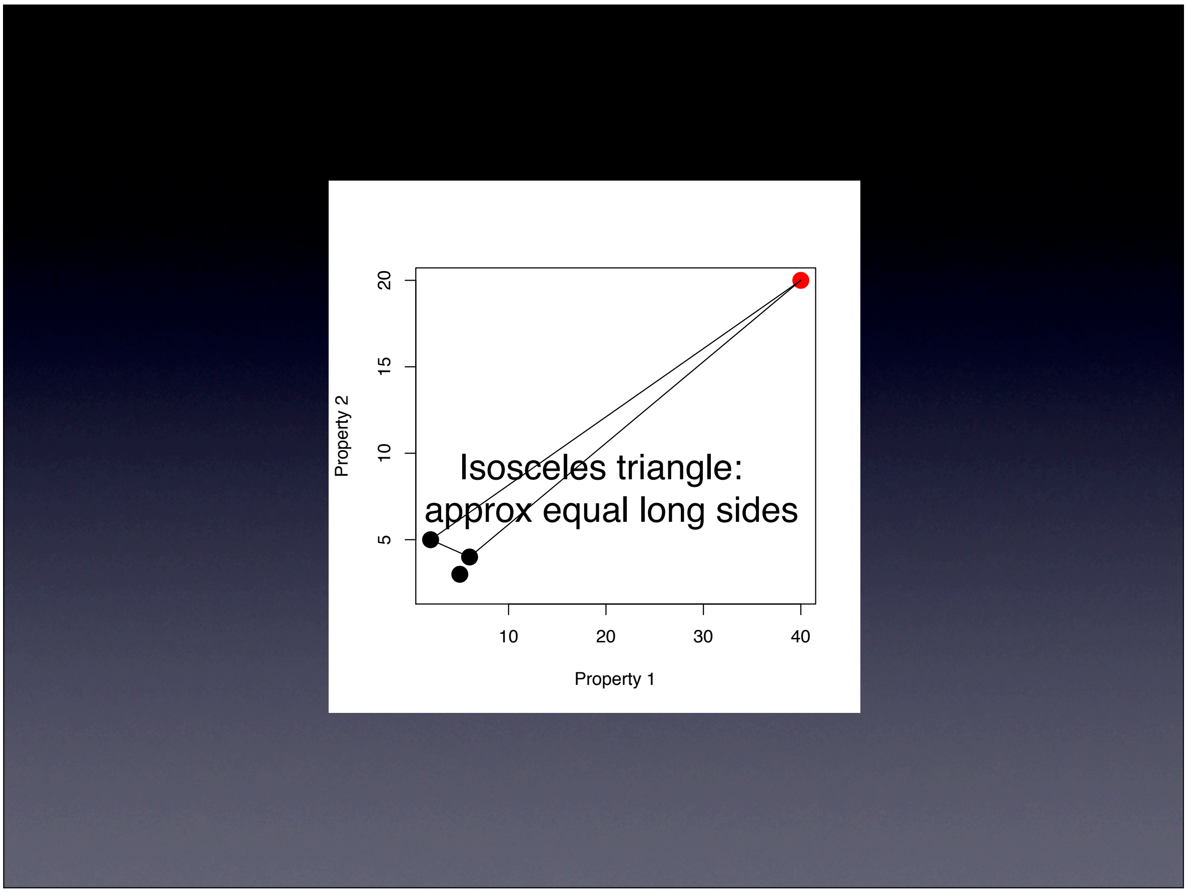


Property 2
20
15
10
5
10
20
30
40
Property 1


Property 2
20
15
10
5
40.85
10
20
30
40
Property 1

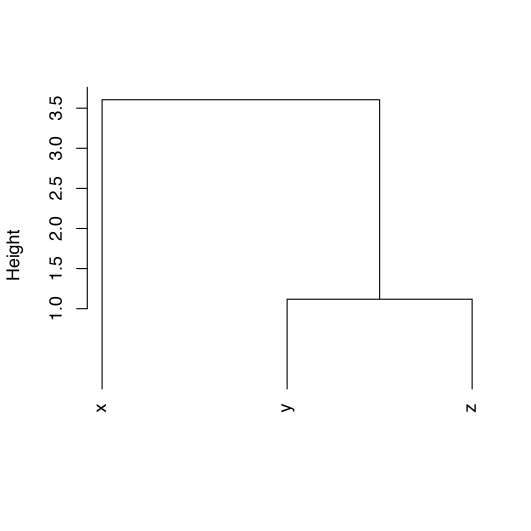

Property 2
20
15
10
5
38.91
10
20
30
40
Property 1





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) ≤ max {d(i,j), d(j,k)}
- Recall: the triangular inequality is: d(i,k) ≤ {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.

Shown by our measuring of 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

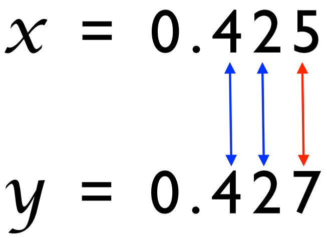
Computational Implications

- 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.

Next: the Baire (ultra)metric

Baire, or longest common prefix

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)]*

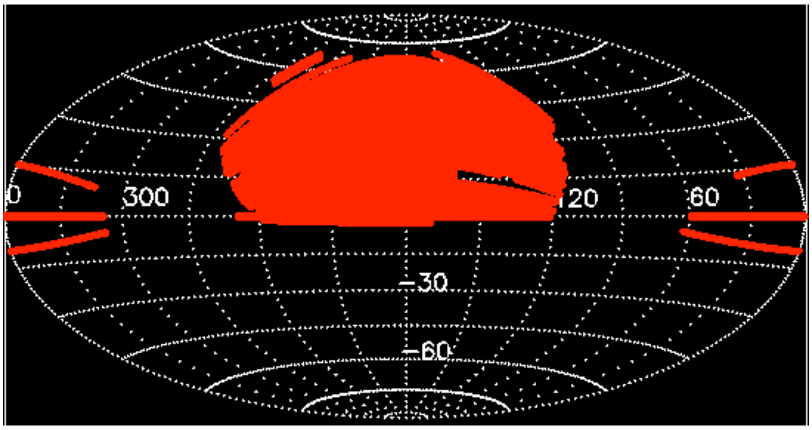
– The longer the common prefix, the closer a pair of sequences.

– 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.)

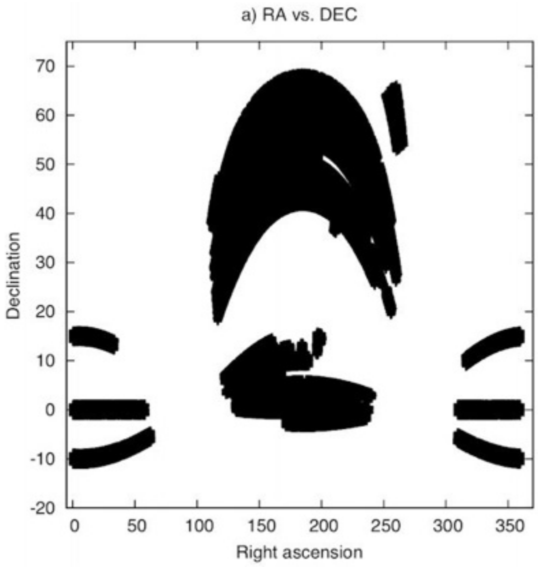
– We applied the Baire distance to: chemical compounds, spectrometric and photometric redshifts from the Sloan Digital Sky Survey (SDSS), and various other datasets.

Sloan Digital Sky Survey: redshifts (photometric, some spectroscopic) of galaxies, quasars and stars

- SDSS DR5 Imaging Sky Coverage
- (Aitoff projection of Equatorial coordinates)



SDSS (Sloan Digital Sky Survey) Data



- We took a subset of approx. 0.5 million data points from SDSS release 5.

- declination (DEC)

-right ascension (RA)

-spectrometric redshift

-photometric redshift

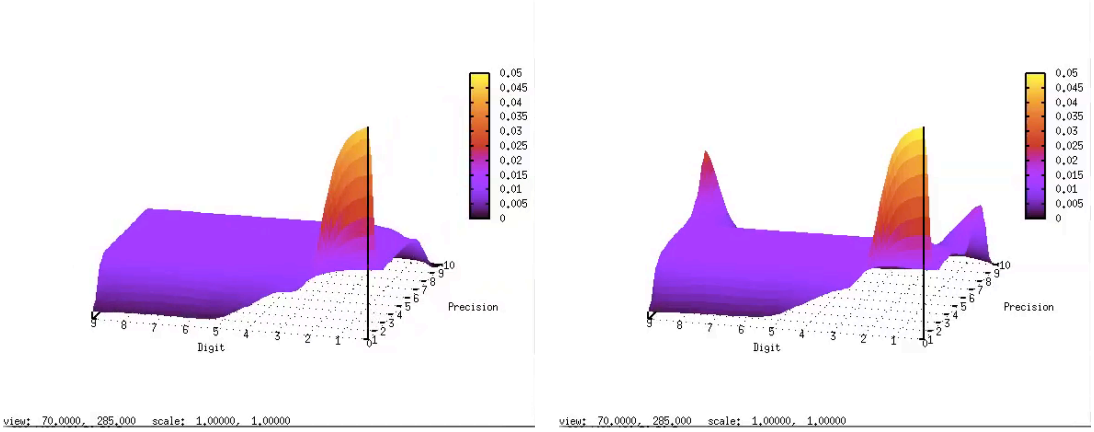
-Dec vs RA are shown in the figure

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

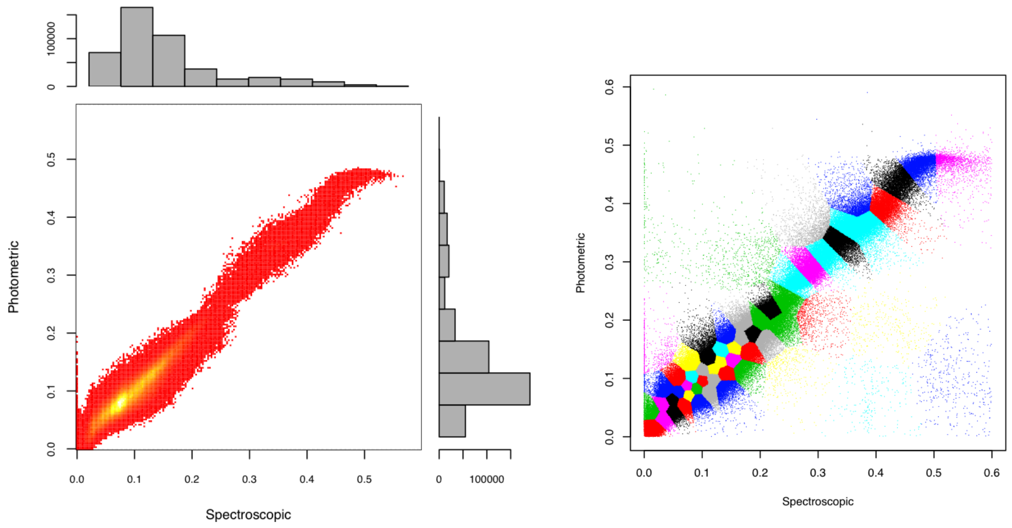


On the right we have z_spec where three data peaks can be observed.
On the left we have z_phot where only one data peak can be seen.

Framework for Fast Clusterwise Regression

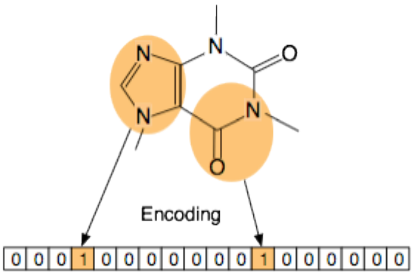
- 82.8% of z_spec and z_phot have at least 2 common prefix digits.
- I.e. numbers of observations sharing 6, 5, 4, 3, 2 decimal digits.
- We can find very efficiently where these 82.8% of the astronomical objects are.
- 21.7% of z_spec and z_phot have at least 3 common prefix digits.
- I.e. numbers of observations sharing 6, 5, 4, 3 decimal digits.

We find good consistency vis-à-vis k-means (right panel)



- 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

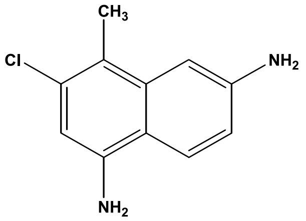


0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0

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 ⇨

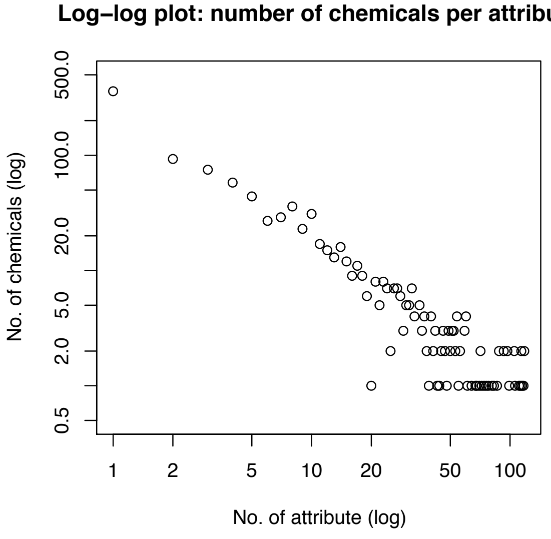
1 0 0 0 1 0 0 0 1 ...1

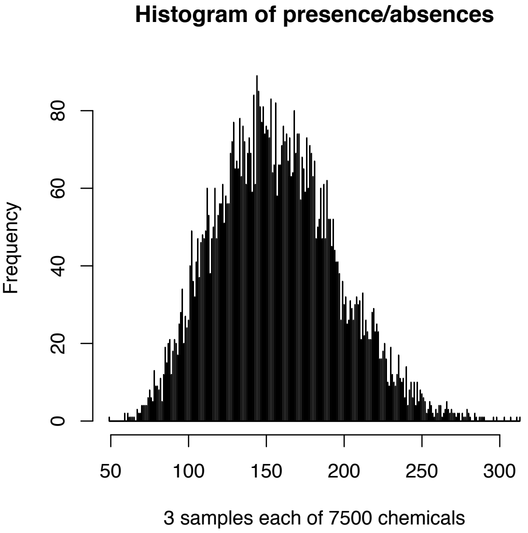
Fixed length bit strings such as
Daylight
MDL
BCI
etc.

Chemoinformatics clustering

- 1.2 million chemical compounds, each characterized by 1052 boolean presence/absence values.
- Firstly we show 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.

Data characteristics: 1.2M chemicals crossed by 1052 presence/absence attributes. Chemicals follow power law with exponent approx. 1.23. Attributes approx. Gaussian.





Dependence of ultrametricity on precision - II

- We have seen that significant numbers of chemicals are identical (0 distance)
- Normalize by dividing by column sums:

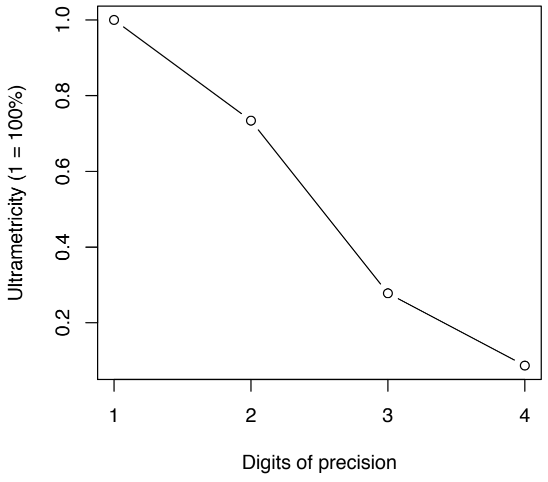
$$x_{IJ} \longrightarrow x^{J}_{IJ}, \text{ where } I, J \text{ are chemical, attribute sets },$$

$$x_J \text{ defines column or attribute masses , and we have: } x^{J}_{IJ} \circ x_J = x_{IJ}$$

- We limit the precision of all normalized values in a chemical's 1052-valued vector
- Then: with very limited precision, we get lots more identical (0 distance) chemicals
- And we find that local ultrametricity increases with limited precision

Dependence of ultrametricity, i.e. data inherently hierarchical, on precision - I

20,000 chemicals, normalized
2000 sampled triangles
Ultrametricities for
precisions 1,2,3,4,...
in all values.
Numbers of non-degenerate
triangles (out of 2000):
precision 1: 2
precision 2: 1062
precision 3: 1999
precision 4: 2000



- We now exploit what we have just observed - potentially high ultrametricity or inherently hierarchical properties in the data, ...
- ... arrived at through reducing the precision of our data values.
- We are looking at different sets of digit precision.
- Implicit here is the Baire distance.

Data “Condensation” through Recoding - I

- We will look for identical chemicals (in the normalized 1052-valued attribute space).
- We will also take all attribute values to limited precision, thereby enabling many more chemicals to be identical.
- As a heuristic to find equivalence classes of identical chemicals, we use a spanning path.
- Path defined by row (chemical) marginal density. (Also looked at random projections, etc.)
- We find clusters of identical chemicals. But we may miss some; and we may have separate clusters that should be merged. For data condensation, unimportant.
- Dominant computational term: for n chemicals, O(n log n) to sort spanning path.

Data “Condensation” through Recoding - II

- Data set 1: form spanning paths, agglomerate identical, adjacent chemicals; repeat. Numbers of chemicals retained on successive passes:
- 20000; 8487; 8393; 8372; 8364; 8360.
- Data set 2:
- 20000; 6969; 6825; 6776; 6757; 6747.
- Similar for further data sets.

- Processing 20000 chemicals (characterized by 1052 normalized attributes) is fast: few minutes in R.

Data “Condensation” through Recoding - III

- Then remaining 8000-odd chemicals, out of 20000 started with (all characterized by the normalized 1052 attributes), are hierarchically clustered using traditional means - using a “commodity” clustering algorithm.
- Ward minimum variance method used.

Data “Condensation” through Recoding - IV

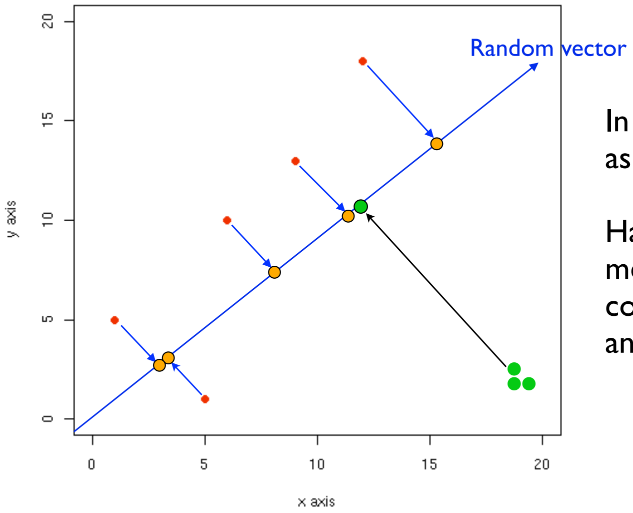
- Some comparative results, with no speed-up processing, from Geoff Downs (Digital Chemistry Ltd.) for clustering 15,465 chemical structures x 1052-bit descriptions:
- Ward 42.5 mins
- k-Means 19.5 mins
- Divisive k-Means 8 mins
- (4 year old PC used, 2.4MHz, 1Gb RAM, Windows XP SP2)

- 152,450 chemical structures x 1052-bit descriptions:
- k-Means 22 hrs
- Divisive k-Means 4.5 hrs

- Data “condensation” through recoding leads to a hybrid hierarchical clustering algorithm.
- It implicitly uses the Baire (ultra)metric in the first “condensation” phase.
- Now we will approach the same issue of finding clusters at increasing levels of refinement more explicitly, by using the Baire (ultra)metric.

- To handle high dimensional data, like the chemoinformatics data, we will use random projections.

Random projection and hashing

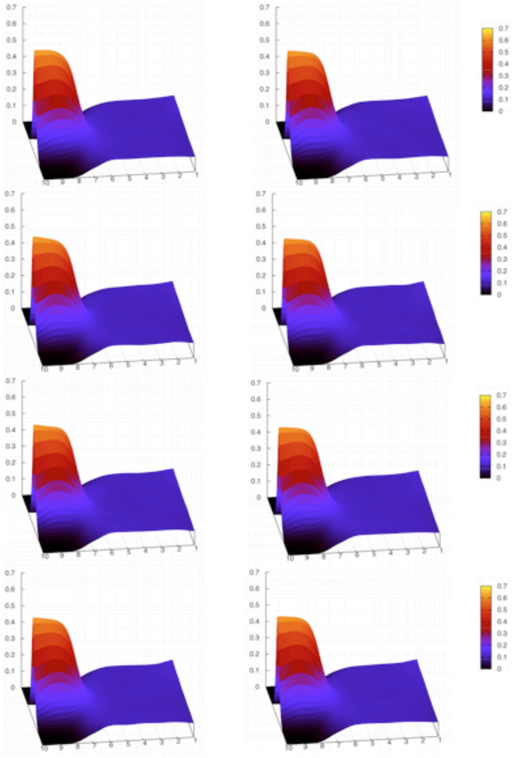


In fact random projection here works as a class of hashing function.

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



Is random projection a good method to reduce dimensionality?

Here we use different random vector to project the original data matrix.

It can be observed in that the shape is kept for the different projections. This implies that different random projections do not affect the resulting clusters.

Simple Clustering Hierarchy

Sig. dig. k	No clusters
4 4 4	6591 6507 5735
3 3 3	6481 6402 5360
2 2 2	2519 2576 2135
1 1 1	138 148 167

Results for the three different data sets, each consisting of 7500 chemicals, are shown in immediate succession. The number of significant decimal digits is 4 (more precise, and hence more different clusters found), 3, 2, and 1 (lowest precision in terms of significant digits).

Simple Clustering Hierarchy

Sig. Dig.	No. Clusters	No. discrep.	No. discrep. cl.
1 1 1	138 148 167	3 1 9	3 1 7

Comparative evaluation: Results of k-means using as input the cluster centres provided by the 1 sig. dig. Baire approach relating to 7500 chemical structures, with 1052 descriptors.

Sig. dig. : number of significant digits used.
No. clusters: number of clusters in the data set of 7500 chemical structures, associated with the number of significant digits used in the Baire scheme.
Largest cluster : cardinality.
No. discrep. : number of discrepancies found in k-means clustering outcome.
No. discrep. cl. : number of clusters containing these discrepant assignments.

- Next:
- Nearest neighbour or best match searching
- Using heuristic to make search in a coordinate space more efficient

Nearest neighbor finding through bounding: the unifying view of ultrametricity

- Feasibility bounds relating to nearest neighbors are an old idea (e.g. Fukunaga and Narendra, 1975)
- Chávez and Navarro (2000, 2003) show how bounds are used: they serve to "stretch the triangular inequality"
- What happens is: we look for a good approximation to a locally ultrametric configuration. From this we have a small and reliable candidate set of nearest neighbors.

- K Fukunaga and PM Narendra, A branch and bound algorithm for computing k-nearest neighbors, IEEE Trans. Computers, C-24, 750-753, 1975
- E Chávez and G Navarro, Probabilistic proximity search: fighting the curse of dimensionality in metric spaces, Information Processing Letters, 85, 39-46, 2003
- E Chávez, G Navarro, R Baeza-Yates and JL Marroquín, Proximity searching in metric spaces, ACM Computing Surveys, 33, 273-321, 2001

- Consider points *u* which we seek to discard when searching for nearest neighbors of query *q*, and we use pivots, *p*.
- Consider the situation of:

$$d(q,u) \leq d(u,p_i) \text{ and } d(q,u) \leq d(q,p_i)$$

- as being of interest. By the triangular inequality:

$$d(u,p_i) \leq d(u,q) + d(q,p_i) \text{ and } d(q,p_i) \leq d(q,u) + d(u,p_i)$$

- This gives the rejection rule: discard all *u* such that

$$|d(u,p_i) - d(q,p_i)| > r$$

- for a threshold *r*, and for some pivot *pi*.
- This gives a bound for the radius around *q* which could be relevant. This bound is in terms of pre-calculated distances.
- If $d(u,p_i) = d(q,p_i)$ then clearly we have no rejection at all of points *u*. But if *r* is small, i.e. $d(u,p_i) \approx d(q,p_i)$ then we have a small and reliable search neighbourhood. The smaller *r* is, *r > 0*, so much the better. But we can't allow it to be too small.
- From the foregoing observations, the triangle formed by *{q, u, pi}* is approximately isosceles with small base, or equilateral.

Concluding Remarks

- 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