The Ultrametric Topology Perspective on Analysis of Massive, Very High Dimensional Data Stores

Fionn Murtagh

Science Foundation Ireland, Wilton Park House, Wilton Place, Dublin 2, Ireland
Dept. Computer Science, Royal Holloway, Univ. London, Egham TW20 0EX, UK

Contemporary Frontiers in High-Dimensional Statistical Data Analysis, Isaac Newton Institute for Mathematical Sciences, January 2008
Lessons

- Very high dimensional spaces are of very simple structure.
- It becomes easier to find clusters in high dimensions.
- We need to think differently when analyzing such data.
- The simple high dimensional structure is hierarchical.
- Ease of handling high dimensional data, e.g. reading off clusters, emulates the human perception system which similarly processes data with no evident latency.
- Symmetry principle applied, rather than sparsity.
Triangular inequality defines a metric: $d(x, z) \leq d(x, y) + d(y, z)$. 

---

**Metric**

Triangular inequality defines a metric: $d(x, z) \leq d(x, y) + d(y, z)$. 

---

**Fionn Murtagh**

Ultrametric Topology Perspective
Strong triangular inequality, or ultrametric, holds for tree distances:
\[ d(x, z) \leq \max \{d(x, y), d(y, z)\} \]
Here: \( d(x, z) = 3.5; d(x, y) = 3.5; d(y, z) = 1.0. \)
In an ultrametric space, all triangles are either isosceles with small base, or equilateral. Based on these triangle properties, we determine the extent to which we can embed a point cloud in an ultrametric space. (See Murtagh, JoC, 2004, for some details; and comparison with Lerman’s, 1981, ranks-based measure of ultrametricity.)
Bellman’s “curse of dimensionality”

Very high dimensional spaces are of simple structure

- Curse of dimensionality: no! In fact: Very high dimensional spaces are full of symmetries.
- In the (high dimensional) limit, everything has a simplex structure, which is a particular case of ultrametricity.
- For $n$ points, $m$ dimensions, we are concerned with (approximately) fixed $n$, but $m \to \infty$. See:
CC Aggarwal et al., “On the surprising behavior of distance metrics in high dimensional spaces”, Proc. 8th Intl. Conf. on Database Theory, 4-6 Jan. 2001, pp. 420–434: “Recent research results show that in high dimensional space, the concept of proximity, distance or nearest neighbor may not even be qualitatively meaningful.”

Breuel, 2007 (citn. next slide): if nearest neighbor point \( y \) for some query point \( x \) has distance \( d(x, y) \), then an \( \epsilon \)-approximate nearest neighbor \( y' \) is s.t. 
\[
d(x, y') \leq (1 + \epsilon)d(x, y).
\]

“... the relationship between approximation and 'cost' of a solution need not be linear. For example, the cost of picking an \( \epsilon \)-approximate nearest neighbor could be proportional not to the difference of distances between the optimal answer and the approximation, but to the volume of the shell between the two, that is, as \((1 + \epsilon)^{m-1}\), where \( m \) is the dimension of the space.”
From the curse of dimensionality to the crisis of dimensionality.

Rather than nearest neighbor, determine $\epsilon$-approximate one.

Difference between query and nearest neighbor is $1 + \epsilon$, compared to exact solution.

Great! But what is the cost of this approximation? Is it $\epsilon$, or is it related to the volume between two shells, $(1 + \epsilon)^{m-1}$?

Consider ratio of nearest to furthest neighbor. In high dimensions these tend to be equal! (We will see this in the following slides.)

If $1 + \epsilon >$ this ratio, then any point is an $\epsilon$-approximate neighbor.

Ahn et al.: “when $m >> n$, under a mild assumption, the pairwise distances between each pair of data points are approximately identical so that the data points form a regular $n$-simplex. In a binary classification setting, the training data from each class becomes two simplices ... any reasonable classification method will find the same [discriminant result] when $m$ becomes very large.”

The mild condition for simplex structure formation, as $m \rightarrow \infty$ is that directionality of the Gaussian cloud is “diffuse”, defined in terms of eigenvalues:

$$
\sum_{j}^{m} \frac{\lambda_j^2}{\left( \sum_{j}^{m} \lambda_j \right)^2} \rightarrow 0 \text{ as } m \rightarrow \infty
$$

Then it is shown that the covariance matrix approaches a constant times the identity matrix.

For a Gaussian cloud, “not only are the points on the convex hull, but all reasonable-sized subsets span faces of the convex hull”.

Intuitively, if all points fly apart from one another as dimensionality grows, then (i) each point is a vertex of the convex hull of the cloud of points; (ii) each pair of points generates an edge of the convex hull; and (iii) sets of points form a regional face of the convex hull.

Conclude: “This is wildly different than the behavior that would be expected by traditional low-dimensional thinking.”
We have seen briefly:

- Ahn, Hall, Marron
- Aggarwal et al.
- Breuel
- Donoho and Tanner
Our general approach

- Why do we lay importance on the fact that the high dimensional simplex additionally defines an ultrametric topological embedding?
- Recall that ultrametric topology requires any triangle to be either (i) equilateral, or (ii) isosceles with small base.
- The equilateral case corresponds fine with the simplex structure.
- But it is useful to us to hang on to the isosceles with small base case, too, for inter-cluster relationships.
- We will now look at simple examples to support this.
Simulated Examples of Increasing Ambient Dimensionality

- We sample triangles because we cannot (computationally) look at all.
- Relative proportion of: isosceles with small base triangles, and equilateral triangles, gives us a measure of ultrametricity. $1 = \text{globally ultrametric}$.
- For triangles we need angles, hence scalar product. We may even make use of distances. Hence, to start with, we require a scalar product space embedding. (A fortiori: Euclidean.)
- We will look in turn at: (1) uniformly distributed points; (2) uniformly distributed hypercube vertices; (3) Gaussian cloud; (4) cases of 2, 3 and 4 Gaussian clusters, the latter very overlapping.
Data are uniformly generated on $[0, 1]^{\text{Dimen.}}$.

<table>
<thead>
<tr>
<th>No. points</th>
<th>Dimen.</th>
<th>Isosc.</th>
<th>Equil.</th>
<th>UM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>100</td>
<td>20</td>
<td>0.10</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>200</td>
<td>0.16</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>2000</td>
<td>0.01</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>20000</td>
<td>0</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Typical results, based on 300 sampled triangles from triplets of points. Dimen. is the ambient dimensionality. Isosc. is the number of isosceles triangles with small base, as a proportion of all triangles sampled. Equil. is the number of equilateral triangles as a proportion of triangles sampled. UM is the proportion of ultrametricity-respecting triangles ($= 1$ for all ultrametric).
Hypercube vertices are in $\{0, 1\}^\text{Dimen}$.

<table>
<thead>
<tr>
<th>No. points</th>
<th>Dimen.</th>
<th>Isosc.</th>
<th>Equil.</th>
<th>UM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypercube</td>
<td>Vertices</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>0.14</td>
<td>0.02</td>
<td>0.16</td>
</tr>
<tr>
<td>100</td>
<td>200</td>
<td>0.16</td>
<td>0.21</td>
<td>0.36</td>
</tr>
<tr>
<td>100</td>
<td>2000</td>
<td>0.01</td>
<td>0.86</td>
<td>0.87</td>
</tr>
<tr>
<td>100</td>
<td>20000</td>
<td>0</td>
<td>0.96</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Dimen. is the ambient dimensionality. Isosc. is the number of isosceles triangles with small base, as a proportion of all triangles sampled. Equil. is the number of equilateral triangles as a proportion of triangles sampled. UM is the proportion of ultrametricity-respecting triangles ($= 1$ for all ultrametric).
Gaussian cloud, $\mu = 0, \sigma = 1$.

<table>
<thead>
<tr>
<th>No. points</th>
<th>Dimen.</th>
<th>Isosc.</th>
<th>Equil.</th>
<th>UM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>0.12</td>
<td>0.01</td>
<td>0.13</td>
</tr>
<tr>
<td>100</td>
<td>200</td>
<td>0.23</td>
<td>0.14</td>
<td>0.36</td>
</tr>
<tr>
<td>100</td>
<td>2000</td>
<td>0.04</td>
<td>0.77</td>
<td>0.80</td>
</tr>
<tr>
<td>100</td>
<td>20000</td>
<td>0</td>
<td>0.98</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Dimen. is the ambient dimensionality. Isosc. is the number of isosceles triangles with small base, as a proportion of all triangles sampled. Equil. is the number of equilateral triangles as a proportion of triangles sampled. UM is the proportion of ultrametricity-respecting triangles (\(= 1\) for all ultrametric).
Up to now: triangle properties; now: histograms of Euclidean distances

How “symmetry” or “structure” becomes more pronounced as dimensionality increases. Shown are histograms of pairwise distances, using 3 sub-populations of Gaussian-distributed data, in ambient dimensions of 2000 and 20,000.
A more detailed simulation

Here, histograms of distances using 2 Gaussian clouds $\mu, \sigma = 0, 1; 10, 1$, in, resp., $\mathbb{R}^{20}$ and $\mathbb{R}^{20000}$. 

Fionn Murtagh  Ultrametric Topology Perspective
<table>
<thead>
<tr>
<th>No. points</th>
<th>Dimen.</th>
<th>Isosc.</th>
<th>Equil.</th>
<th>UM</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>20</td>
<td>0.08</td>
<td>0</td>
<td>0.08</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>0.19</td>
<td>0.04</td>
<td>0.23</td>
</tr>
<tr>
<td>200</td>
<td>2000</td>
<td>0.42</td>
<td>0.20</td>
<td>0.62</td>
</tr>
<tr>
<td>200</td>
<td>20000</td>
<td>0.74</td>
<td>0.22</td>
<td>0.96</td>
</tr>
<tr>
<td>200</td>
<td>20000</td>
<td>0.7</td>
<td>0.28</td>
<td>0.98</td>
</tr>
<tr>
<td>200</td>
<td>20000</td>
<td>0.77</td>
<td>0.21</td>
<td>0.98</td>
</tr>
<tr>
<td>200</td>
<td>20000</td>
<td>0.76</td>
<td>0.21</td>
<td>0.98</td>
</tr>
<tr>
<td>200</td>
<td>20000</td>
<td>0.75</td>
<td>0.24</td>
<td>0.99</td>
</tr>
<tr>
<td>200</td>
<td>20000</td>
<td>0.73</td>
<td>0.25</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Two Gaussian clusters, each of $n = 100$. $\mu, \sigma = 0, 1; 10, 1$. 
### Strongly overlapping clusters

<table>
<thead>
<tr>
<th>No. points</th>
<th>Dimen.</th>
<th>Isosc.</th>
<th>Equil.</th>
<th>UM</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>20</td>
<td>0.04</td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>0.11</td>
<td>0.05</td>
<td>0.16</td>
</tr>
<tr>
<td>200</td>
<td>2000</td>
<td>0.28</td>
<td>0.06</td>
<td>0.34</td>
</tr>
<tr>
<td>200</td>
<td>20000</td>
<td>0.5</td>
<td>0.08</td>
<td>0.58</td>
</tr>
<tr>
<td>200</td>
<td>200000</td>
<td>0.55</td>
<td>0.11</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Four Gaussian clusters, each \( n = 50 \). \( \mu, \sigma = 0, 1; 3, 2; 5, 1; 8, 3 \) on each dimension.
A projection onto one dimension, to illustrate the less than clearcut clustering problem addressed. There are four Gaussians here, each of 50 realizations, with means at 0, 3, 5 and 8, and with respective standard deviations of 1, 2, 1, 3.
Compaction of distances with rise in dimensionality: 4 clusters, substantially overlapping are the basis for the histograms of all pairwise distances. Top: ambient dimensionality 20. Bottom: ambient dimensionality 200,000.
Cluster identifiability

- Three clusters, Gaussians, with resp. means and standard deviations on all dims.: (10, 0.5); (0, 4); (40, 10). Expect: 3 intra-cluster distance peaks, and 3 inter-cluster distance peaks.

- And fourth cluster with: (25, 7). Expect: 4 intra-cluster peaks, and 6 inter-cluster.

- Four clusters, with equally-sized subclusters: (10, 0.5) and (0, 0.5); (0, 4) and (10, 4); (40, 10) and (0, 10); and all (25, 7). We can have up to 28 peaks.
90 points, 3 clusters, dim. 1000

90 points, 3 clusters, dim. 10000
120 points, 4 clusters, dim. 1000

120 points, 4 clusters, dim. 10000

Fionn Murtagh  Ultrametric Topology Perspective
240 points, 4 double clusters, dim. 1000

Distances

Frequency

0 200 400 600 800 1000 1200 1400

0 600

240 points, 4 double clusters, dim. 10000

Distances

Frequency

0 1000 2000 3000 4000

0 1000
Some preliminary conclusions

- A scalar product space becomes ultrametric in the high dimensional limit.
- Between cluster relationships contribute to the ultrametric case of isosceles with small base. Within cluster relationships contribute to the ultrametric case of equilateral triangles.
- Take Fisher’s iris data, i.e. 150 points in $\mathbb{R}^4$. Recode to complete disjunctive form as used in correspondence analysis, i.e. a boolean coding. We get 150 hypercube vertices in $\{0, 1\}^{123}$, hence in a 123-dimensional space. To begin with we measure ultrametricity as (very low) 0.017. The ultrametricity measure of the recoded data is 0.948.
- One reason why recoding data is so important in correspondence analysis is that ultrametricity is enhanced! Also correspondence analysis maps a space endowed with a $\chi^2$ metric into a Euclidean one, i.e., the factor space.
Consider clustering algorithms in terms of number of observables, $n$, and number of attributes, $m$.

By "large" is meant thousands upwards.

Large $n$, small $m$; as is fairly standard in astronomy.

Large $n$, large $m$; as is fairly typically the case in information retrieval.

And small $n$, large $m$; as is often the case in bioinformatics, and textual forensics.

Last is of interest to us. Called "HDLSS, high dimension, low sample size" case by Jeongyoun Ahn (Georgia), Steve Marron (CMU), Peter Hall (ANU).
Application to very high frequency data analysis

- Application: segmentation, model identification of e.g. engine vibration analysis, forex trades, bio-engineering, or meteorology.
- We use two ARIMA signals, with a Student t innovation with 5 degrees of freedom, providing “mildly longtailed” distributions.
First signal: sample (using first 2000 values) of a time series segment, based on the first ARIMA set of parameters. (Order 2 AR parameters: 0.8897, −0.4858, MA parameters: −0.2279, 0.2488.)
Second signal: sample (using first 2000 values) of a time series segment, based on the second ARIMA set of parameters. (Order 2 AR parameters: 0.2897, −0.1858, MA parameters: −0.7279, 0.7488.)
Sample of 2000 values of first signal, followed directly by 2000 values of second signal.
Histogram of sample (using first 2000 values) of first signal.
Histogram of sample (using first 2000 values) of second signal.
Point = embedding of “window”. So Dimen. = embedding dimensionality.

<table>
<thead>
<tr>
<th>No. time series</th>
<th>Dimen.</th>
<th>Isosc.</th>
<th>Equil.</th>
<th>UM</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>2000</td>
<td>0.17</td>
<td>0.32</td>
<td>0.49</td>
</tr>
<tr>
<td>100</td>
<td>20000</td>
<td>0.15</td>
<td>0.5</td>
<td>0.65</td>
</tr>
<tr>
<td>100</td>
<td>200000</td>
<td>0.03</td>
<td>0.57</td>
<td>0.60</td>
</tr>
</tbody>
</table>

Results based on 300 sampled triangles from triplets of points. Two sets of the ARIMA models are used, each of 50 realizations.
Histogram of distances from 100 time series segments, using 50 segments each from the two ARIMA models, and using an embedding dimensionality of 200,000.
Model-based clustering in very high dimensions

- The histogram of inter- and intra-cluster distances has been shown, using a signal embedding of 200,000. As discussed, the peaks are compact because we are in a high dimensional space.

- No normalization is applied. We do not want to rescale. Also unweighted Euclidean distance is consistent with no modification of triangle invariants.

- Lower and higher histogram peaks belong to the two ARIMA components. The central peak belongs to the inter-cluster distances.

- So model-based clustering in very high dimensions is: estimate number of clusters, and their properties, from the distinct compact peaks. *The greater the dimensionality the better!*
A further interesting conclusion on this work follows.

Traditionally, clustering algorithms have generally been considered as distance-based or model-based.

The former is exemplified by agglomerative hierarchical clustering, or k-means partitioning.

The latter is exemplified by Gaussian mixture modeling. (One motivation: the difficulty of defining distance.)

The approach described in this work is both distance-based and model-based.

How do we define distance in general contexts, given frequencies, ranks, mixed qualitative/quantitative data as input? Correspondence analysis provides a general way to map data into a Euclidean, factor space.

Financial futures, circa March 2007, denominated in euros from the DAX exchange. Data stream is at the millisecond rate. Comprises about 382,860 records. Each record includes: 5 bid and 5 asking prices, together with bid and asking sizes in all cases, action.

Extracted one symbol (commodity) with 95,011 single bid values.
The signal used: a commodity future, with millisecond time sampling.
Embeddings were defined as follows.

- Windows of 100 successive values, starting at time steps: 1, 1000, 2000, 3000, 4000, . . . , 94000.
- Windows of 1000 successive values, starting at time steps: 1, 1000, 2000, 3000, 4000, . . . , 94000.
- Windows of 10000 successive values, starting at time steps: 1, 1000, 2000, 3000, 4000, . . . , 85000.

Note: for 10000-length windows, fully 90% of the values of the point in this embedding will overlap with values of the next point. So any clusters we find will necessarily be strongly overlapping.

- The histograms of distances between these windows, or embeddings, in respectively spaces of dimension 100, 1000 and 10000, are shown next.
Histograms of embeddings in dimensionalities 100, 1000, 10000. Respectively the numbers of embeddings are: 95, 95 and 86.
So what are the clusters arising out of the histograms shown?

We take the 10000-length window case.

We fit Gaussians to the data shown in the bottom histogram of previous figure.

Fit Gaussians, and determine best fit using BIC (Bayes information criterion).

See figure to follow. Best solution: 5 Gaussians.

Means: 517, 885, 1374, 2273 and 3908.

Standard deviations: 84, 133, 212, 410 and 663.

Cardinalities: 358, 1010, 1026, 911 and 350.
BIC values, for the distances between points in a 10000-dimensional embedding, indicating a best fit for 5 Gaussians.
While we have the segmentation of the distance histogram, we need the segmentation of the original financial signal.

One possibility: use principal coordinates analysis. In fact, a 2D mapping is very similar in its distance histogram to that seen using the full, 10000, dimensionality.

Knowing that 5 clusters are wanted, we will simply use an agglomerative hierarchical clustering, with the minimum variance criterion, to find them.

Reading off the cluster memberships gives a directly applicable result. We find: embedded points 1–10 to be in class 1; points 11–28 in class 2; points 29–49 again in class 1; points 50–75 in class 3; points 76–79 in class 4; and finally points 80–86 in class 5.

This allows us to segment the original time series.
Hierarchical clustering, of the 86 10000-dimensional points. Due to the Gaussian mixture modeling, and BIC analysis, we are interested in the 5-cluster solution.
Boundaries found for segments. Note: lowest segment and 3rd segment are both cluster 1; 2nd segment is cluster 2; 4th segment is cluster 4; and highest segment is cluster 5.
A quite fascinating representation – a type of “return map” – found using a principal coordinates analysis of the 86 10000-dimensional points. Again a demonstration that very high dimensional structures can be very simple.
A different ultrametric embedding, compared to that of agglomerative hierarchical clustering.

Works directly on the data, rather than distances.

Can be viewed as taking the data to increasingly finer precision. In this respect, can be viewed as making use of a particular data coding.

Can also be viewed as a form of Formal Concept Analysis.
Chemical compound matching. Based on a fixed length descriptor. Here: Digital Chemistry bci1052 dictionary of fragments. Assumed: properties are associated with such structure.
Attributes per chemical are approximately Gaussian. Three different subsets of chemicals contribute to this histogram.
Chemicals per attribute are long-tailed. Shown: histogram of column sums.
Chemicals per attribute follow a power law. Find: probability of having more than $p$ chemicals per attribute to be approximately $\frac{c}{p^{1.23}}$ for large $p$ and for constant, $c$. 
1. Normalize column-wise, so each value is $\leq 1$ and $\geq 0$.

2. Determine a random projection: for $i$th chemical and attribute set $J$, $x_{iJ}$, a random projection is the scalar product $x_{iJ}^t w_J$, where $t$ denotes transpose, and $w_J \in [0..1]|J|$, is a uniformly distributed vector.

3. Based on data recoding: we will cluster data values at successively better levels of precision.

4. Consider chemicals $x$, $y$ on just one attribute, $x_K = 0.478$; and $y_K = 0.472$, where precision is $|K| = 3$.

5. 

$$d_B(x_K, y_K) = \begin{cases} 
1 & \text{if } x_1 \neq y_1 \\
\inf 2^{-n} & x_n = y_n, \ 1 \leq n \leq |K|
\end{cases}$$

6. For the example, $d_B(x_K, y_K) = 2^{-3}$.

7. We have: Greatest common prefix metric. Or 1-bounded ultrametric. (The Baire space is the space of countably infinite sequences.)
Results for 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).
Comparative evaluation: Results of k-means for the same three data sets used heretofore, each 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.
We note that the partitions in each case are dominated by a very large cluster, which is a direct consequence of the data used. In cases that do not give rise to such “imbalanced” cluster cardinalities, our Baire-related approach should perform even better, in that it will give rise to more equal cardinality clusters.

We are now pursuing work on Sloan Digital Sky Survey (SDSS) archive.

Using both (high quality, more costly to collect) spectroscopic and (lower quality, more readily available) photometric redshifts.

For best match (nearest neighbor) searching, typically in this case we are dealing with millions of objects in a low dimensional attribute space.

Our first goal, p-adic regression of photometric and spectroscopic redshifts.
Conclusions

- High dimensional spaces are different. They are of simple structure.
- An ultrametric topology embedding may be a convenient way to analyze such data.
- There are different ultrametric embeddings available to us, of course.
- To read further:
Ongoing Work

- Sloan Digital Sky Survey
- Pedro Contreras: best and/or exact match retrieval in PB (petabyte) sized archived data stores
- For information retrieval “out of context”, innovation-finding, – a hierarchy explains a lot, much more so than a metric embedding. Application to national research funding systems.
Q: High dimensional Gaussian clouds – surely just the Central Limit Theorem? All very well known?

A: No: the ultrametric framework is a new context and offers new insights.

Q: Feature selections (hence low dimensional mapping) is not handled in this framework.

A1: There are many applications where all attributes should be treated equally, e.g. text processing, or the chemo-informatics case study.

A2: Simple structure in high dimensions means that low dimensional embeddings are there – but are difficult to find. Cf. the forex case study.