



Enterprise-level cheminformatics.

Clustering


Converting data to knowledge

John Bradshaw
Daylight CIS Inc

Thanks

- Jack Delany
- Daylight CIS Inc, Santa Fe, NM
- Ifat Noreen
 - MSc Dissertation submitted in part fulfilment of a Masters in Chemoinformatics, University of Sheffield, 2004
- John Holliday
 - University of Sheffield

Background

- There is a general perception that the only way to do clustering with Daylight software is to use similarity values using the Tanimoto index on standard Daylight fingerprints using the Jarvis-Patrick algorithm.
- This is illustrated in a recent book chapter, M. Stahl, M. Rarey, and G. Klebe, in *Bioinformatics: From Genomes to Drugs*, T. Lengauer Ed., VCH, Weinheim, 2001, pp. 229. “Screening of drug databases”

- Whilst none of this information is incorrect, it is counter to Daylight’s avowed intent to provide cheminformatics tools, rather than Hobsonian applications.
- This talk accounts for some of the work we are doing do correct this misconception.

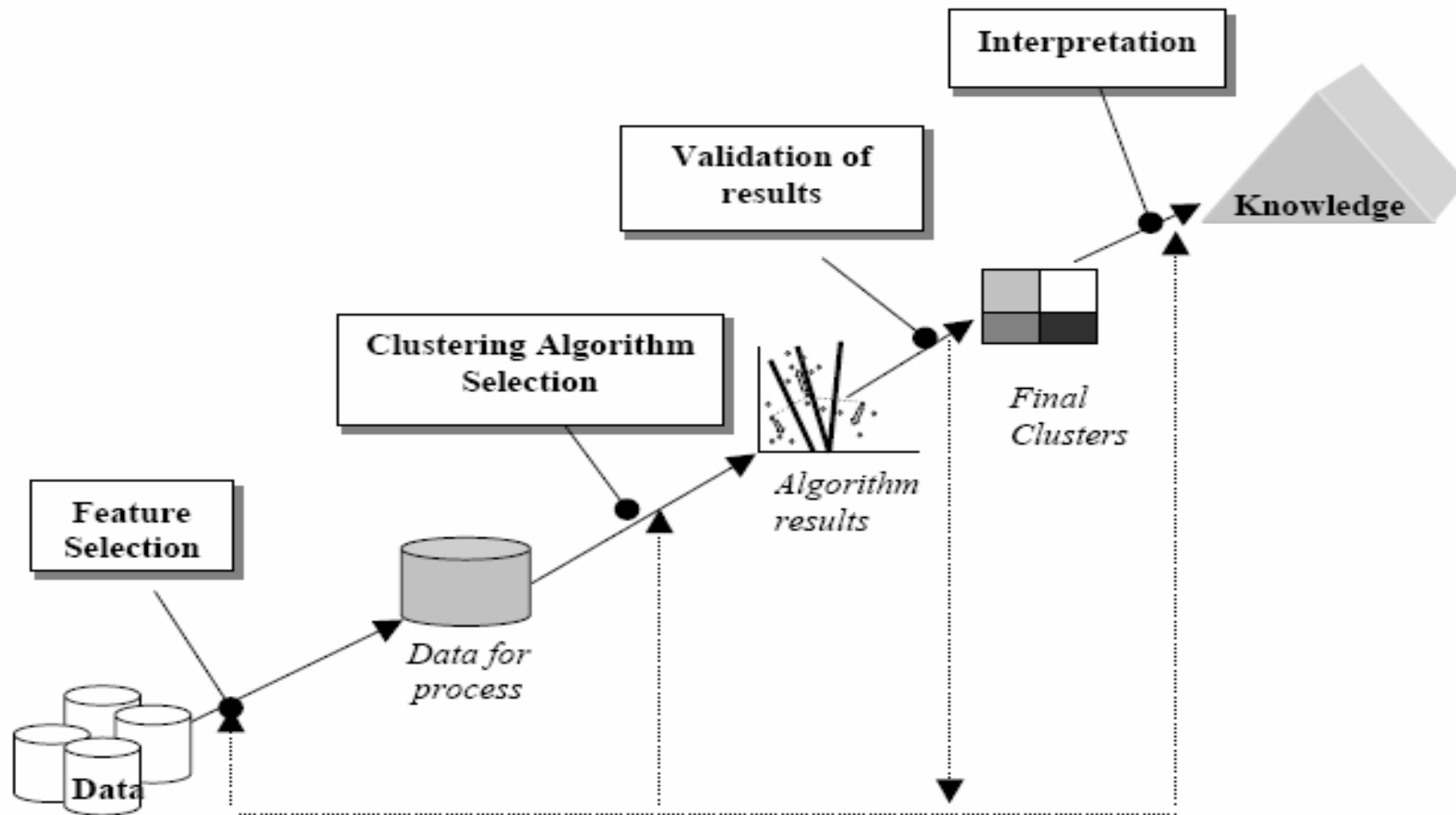
...*Daylight fingerprints* (DF) [25] are bit strings generated from bond paths of length zero to seven. The length of the fingerprints can be folded until a specified density of set bits is reached.

Alternatively, the fingerprint length can be set to a fixed value. The similarity index used is the Tanimoto coefficient, which is the number of bit positions set to 1 in both strings divided by the number of bit positions set to 1 in at least one of the strings. If a set bit is considered as a feature present in the molecule, the Tanimoto coefficient is a measure of the number of common features in two molecules [28, 29]....

M. Stahl, M. Rarey, and G. Klebe, in *Bioinformatics: From Genomes to Drugs*, T. Lengauer Ed., VCH, Weinheim, 2001, pp. 229. "Screening of drug databases"



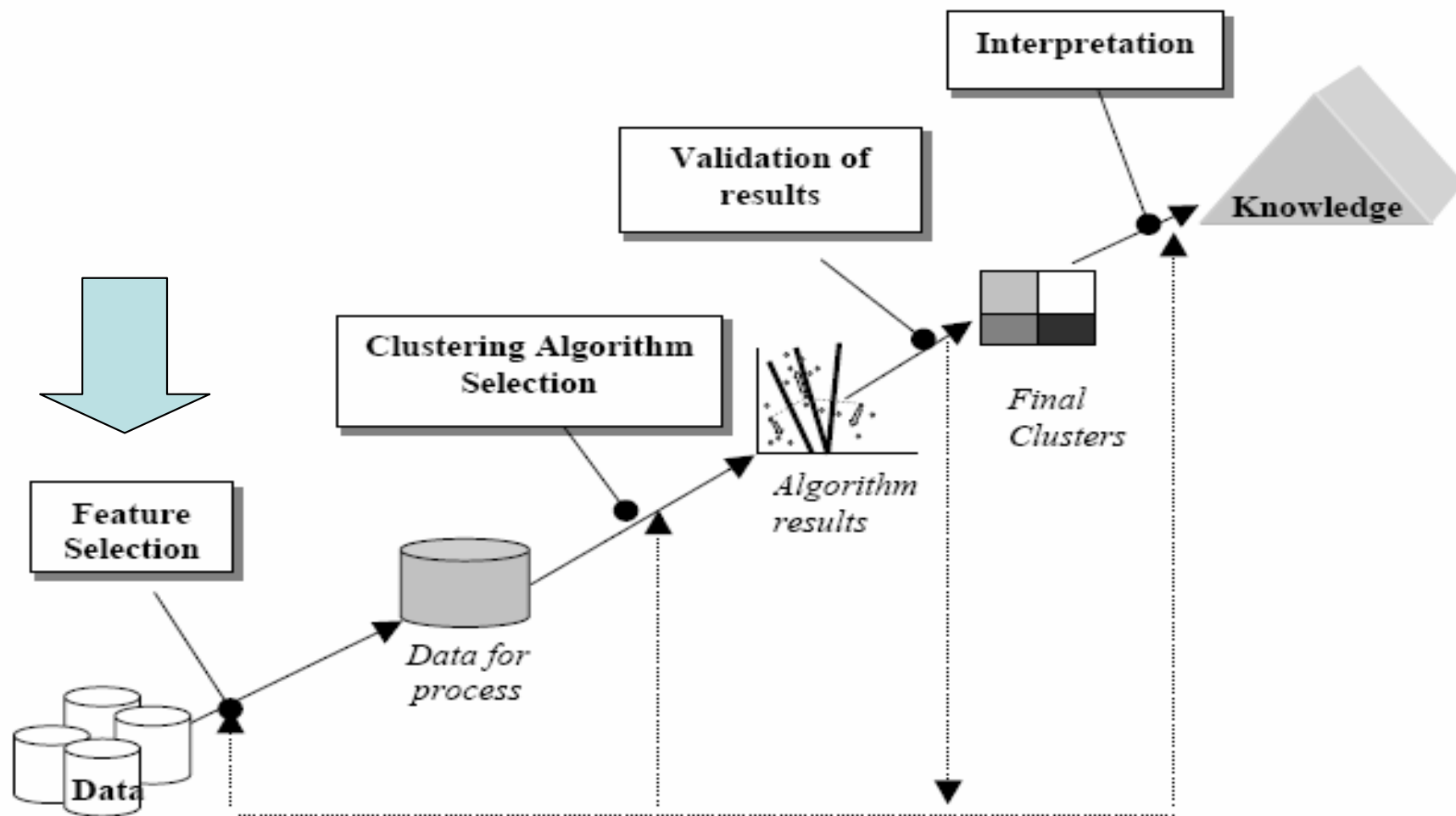
The process



Steps in the process

- Whilst the steps in the process are independent, the choices made early on can prune the choices available later.
- For example
 - The bit values in Daylight fingerprints are categorical data.
 - The binary values of 0 and 1 indicate the absence (0) or possible presence (1) of a particular path.
 - There is no sense in which any of the 0 values along a particular fingerprint can be equated or related.
 - The fact a snake possesses neither wheels nor legs allows us to say nothing about the relative value of wheels or legs.
 - The same is true for the 1 values. In Daylight fingerprints this situation is complicated further by the ambiguous nature of the meaning of a single set bit.
- This restricts the options available in clustering algorithm and similarity index.

The process



Feature Selection

- Feature selection is by far the most important step in this process.
- “...First it is important to note that our total database concerning a particular object ...is generally rich in content and complex in form. It includes appearance, function, relation to other objects, and any other property of the object that can be deduced from our general knowledge of the world. When faced with a particular task ...we extract and compile from our database *a limited list of relevant features on the basis of which we perform the required task*”

Tversky, A. (1977) *Psychological Review* **84**(4), 327

Fingerprints and feature keys

- The default object descriptor for molecules in Daylight is structure based.
- There are two main types of structure based descriptions.
 - Feature keys
 - + These map well to observations and to the class nature of organic chemistry.
 - However they require you know the classes up front to set the keys.
 - Potentially there are a large number of possible features.
 - Fingerprints
 - + These are graph based so do not rely on *a priori* classification.
 - + It is possible to pack them into a fixed width, irrespective of number of features.
 - There is no simple relationship between the pattern and the feature.

Daylight fingerprints

- Starting with each atom, traverse all paths, branches, and ring-closures up to a certain depth (typically 8). For each substructure, derive a hash-like number from unique, relatively-prime, order-dependent contributions of each atom and bond type. Critical properties of this number are that it is reproducible (each substructure produces a single number) and its value and graph are not correlated (a linear congruential generator is used to insure this).
- Map each resulting number into a large range (typically 2K-64K) to produce a redundant, large-scale, binary representation of the substructural elements. The resultant "fingerprint" contains a large amount of information at a low density.
- Iteratively "fold" the fingerprint by OR-ing the fingerprint in half until the bit-density reaches a minimum required value or until the fingerprint reaches a minimum allowable length. The resulting fingerprint now has a high information density with a minimal (and controllable) information loss.

OK. So what does that mean?

- For example, the molecule OC=CN would generate the following patterns:
 - *0-bond paths*: C O N
 - *1-bond paths*: OC C=C CN
 - *2-bond paths*: OC=C C=CN
 - *3-bond paths*: OC=CN
- The list of patterns produced is exhaustive: *Every* pattern in the molecule, up to the pathlength limit, is generated. For all practical purposes, the number of patterns one might encounter by this exhaustive search is infinite, but the number produced for any *particular* molecule can be easily handled by a computer.

Health warning

- Fingerprints (and also feature keys) were designed to act as filters in substructure and superstructure searches.
- If molecule A is a substructure of molecule B, all the patterns that exist in the fingerprint of molecule A must be present in the fingerprint of molecule B.
- In a fingerprint, created as described, all parts of the molecule are treated equally. Aliphatic carbon has the same weight as aromatic arsenic.
- Whilst the folding paradigm works well for filtering, in a similarity search the value is directional (more later)

Fingerprints are not...

- Representations of high dimensional Cartesian space.
- Appropriate input for a neural network or other data reduction techniques.
 - Paths are represented by a *pattern* of bits, so individual bits are not independent

- Unique

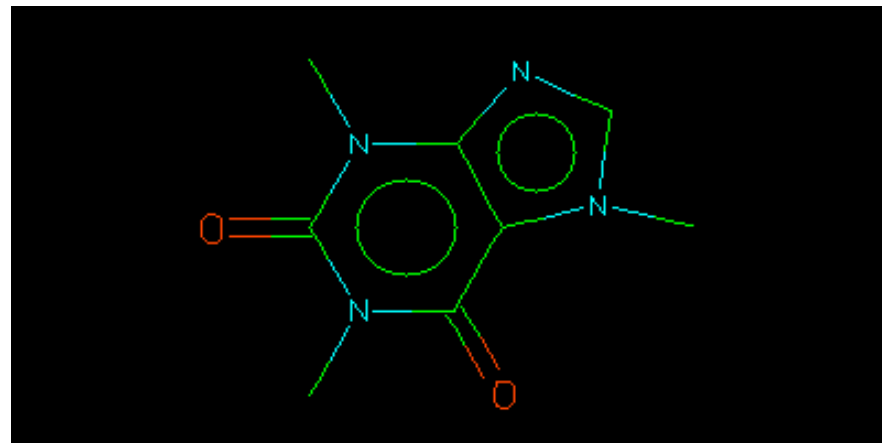
- Try

```
thorlist medchem02demo      \  
| grep `FP<`                \  
| sort                       \  
| uniq -c                   \  
| sort -nr                   \  
| more
```

- There is less duplication with unfolded fingerprints.

Loss of information on folding

Size	Bits on
16384	176
8192	175
4096	173
2048	169
1024	161
512	148



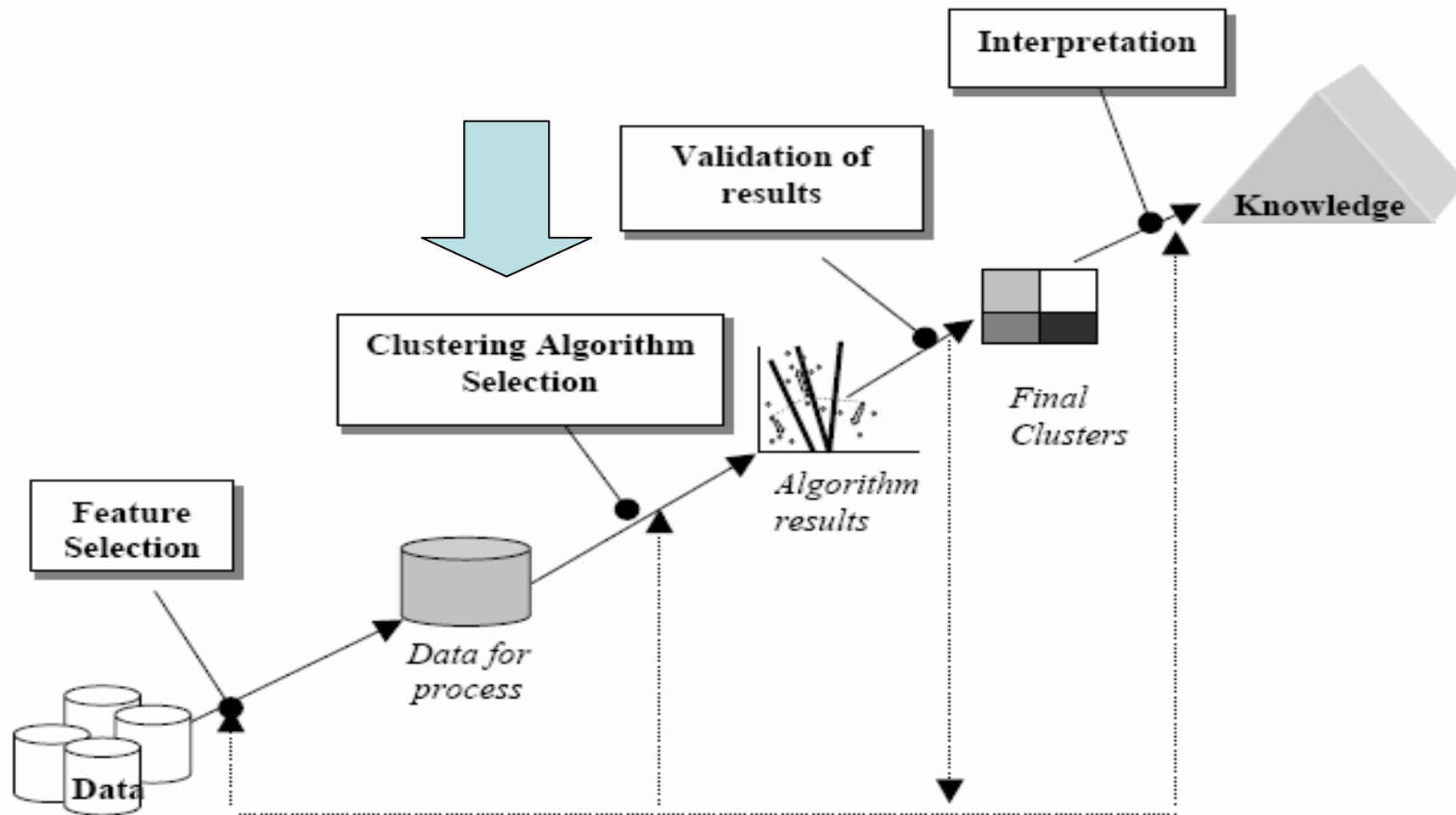
Using non-standard fingerprints

- In the default Daylight fingerprints all heavy-atom paths are equally weighted.
- However any Daylight object which can be streamed over to return a stream of atoms and bonds can be used to build a fingerprint.
- Of especial importance in this regard is the reaction object which can return a full or transform fingerprint.
 - What follows applies equally to a substance or reaction based ontology
- Daylight provides example code to fingerprint
 - Fragments
 - Rings
 - All but non-chain atoms C atoms.
- Users can use the toolkit to make other choices to more accurately reflect the critical parts of the molecule

Feature keys and other fingerprints

- Daylight fingerprint objects only have constraints on size, not on content or meaning.
- So any molecular descriptor which can be represented as a binary string whose size is a multiple of 8 and a power of 2 ($32 \leq N \leq 2^{31}$) can be used as a fingerprint in the toolkit and applications including DayCart™
- A talk program, **filter_fingertalk()** is available which will
 - read a file of SMARTS and return a fingerprint with bits set corresponding to the presence or absence of the patterns for each subsequent SMILES passed to the program
 - read a list of bits to turn-on and return a valid Daylight fingerprint. This fingerprint can be used as a screen with **fingertest()**.
- This has been designed to populate columns in Oracle.
- A thor equivalent is also available to create merlin pools.

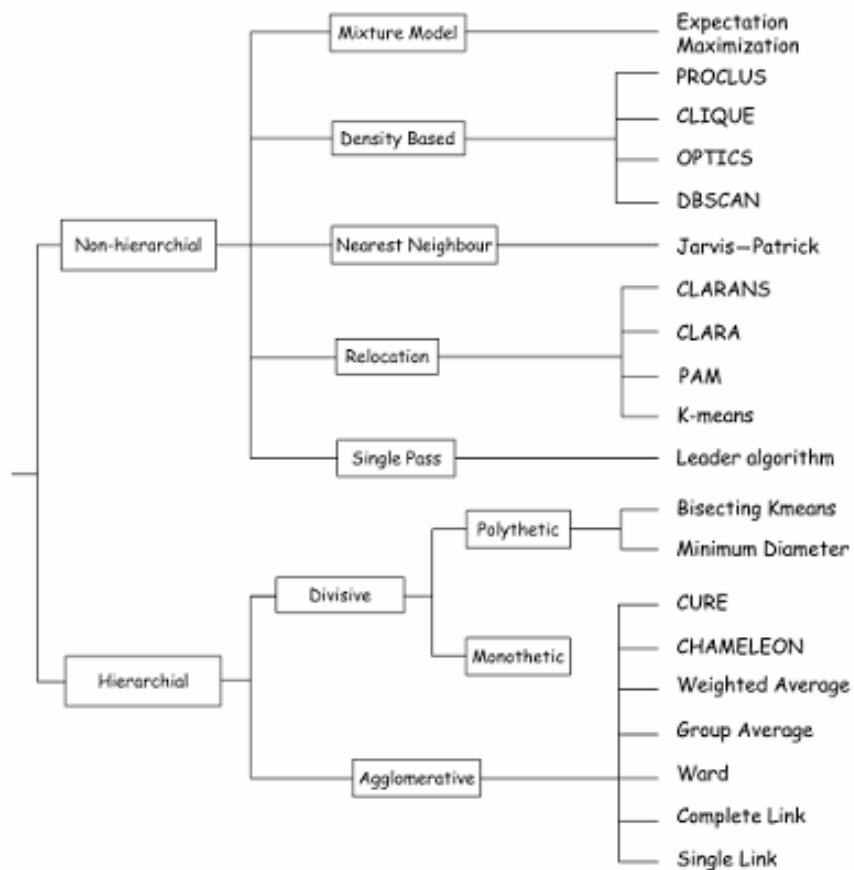
The process



Clustering Algorithm Selection

- The options open to users for clustering algorithms are enormous. The choice of features used to describe the molecules will restrict the choices.
- All cluster algorithms are founded on the premise that objects within a cluster are more similar to each other than to objects in other clusters.
- The choice of similarity coefficient we have made available is wide, see <http://www.daylight.com/meetings/mug04/Bradshaw/coefficients.html> represented by a standard nomenclature, see http://www.daylight.com/meetings/mug04/Bradshaw/bit_count.html

Cluster methods



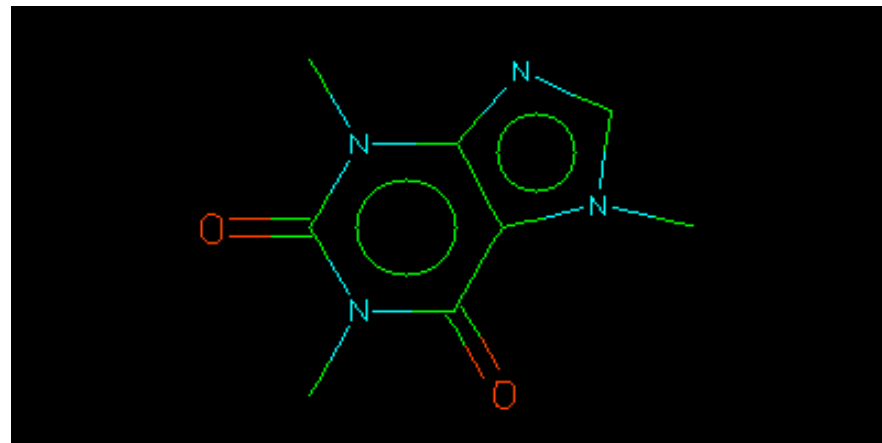
After Downs and Barnard

Daylight clustering algorithms

- Given the characteristics of the default Daylight fingerprints we have implemented appropriate non-hierarchical, non-parametric clustering algorithms.
 - Jarvis-Patrick
 - *k*-modes
 - Sphere exclusion
- All the named similarity functions are available plus user defined measures which are **f(a,b,c,d)**
- Whilst all options are *available*, it does not mean they all are *appropriate*.

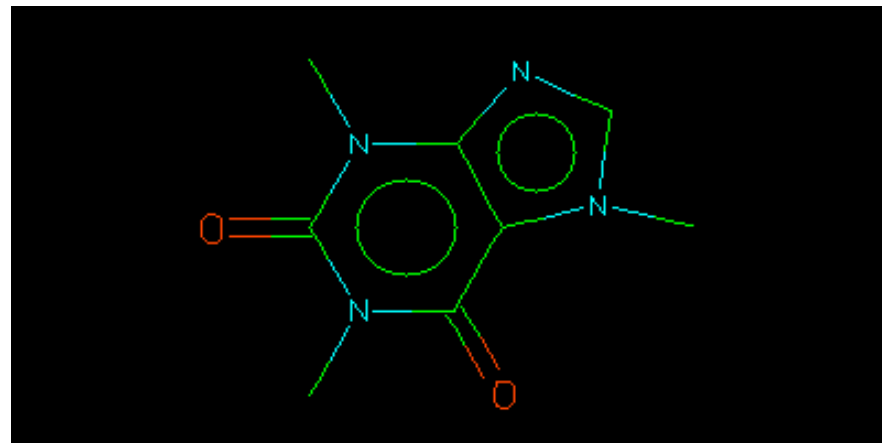
Loss of information on folding

Size	Bits on
16384	176
8192	175
4096	173
2048	169
1024	161
512	148



Loss of information on folding revisited

Δ -size	Δ -bits on
8192	1
4096	2
2048	4
1024	8
512	13
-	-



Counting matching off-bits

- As the number of off-bits in a Daylight fingerprint can be arbitrarily altered without significantly increasing the information content, the use of similarity coefficients which use a match of these off-bits, **d**, is to be discouraged.
- This may not be the case for fingerprints originating in other paradigms.
- This does not affect the use of folded fingerprints for sub/superstructure screening.
- Different rank similarity orderings are observed in DayCart and merlin, depending on fingerprint size.

Jarvis-Patrick

- There have been several improvements to the Jarvis-Patrick algorithm
 - Improvements to the handling of ties see <http://www.daylight.com/meetings/mug04/Delany/ties.html>
 - Availability of user-defined similarity measures
 - Use of user defined fingerprints

Sphere-exclusion

- Sphere exclusion, see <http://www.daylight.com/meetings/mug04/Delany/spherex.html> is a simple, intuitive selection method.
- Given a set of items, one proceeds by selecting items one at a time, usually at random, and excludes items which are near it from further consideration. Selection is complete once all items have either been selected or excluded.
- It is not strictly a clustering algorithm, however it does approximate to clustering where the information outcome is a sub-set of representative objects chosen as the representatives of clusters.

k-modes

- *k*-modes clustering
http://www.daylight.com/meetings/mug04/Bradshaw/why_k-modes.html is the non-parametric version of the well-known *k*-means algorithm. Modal Daylight fingerprints are used to typify the cluster.
- In the current algorithm the user needs to input the number of starting clusters. Other than in exceptional circumstances this is also the final cluster number. These starting points can be user-defined or assigned at random.
- Good methods are therefore required to get the starting set, “seeds” and to estimate the “right” number of clusters
 - Starting and stopping rules

Setting the seeds

- In order to compare the methods of seed selection one needs a method to measure the goodness/validity of the solution.
- There is more on this later, but what is described here is the criteria used for the Sheffield work.
- Given that most of the clustering work Daylight customers do is aimed at supporting drug design, there is an implicit assumption that there is a relationship between such drug activity and the structural descriptors.
- Any method which reproduces a known classification is “better” than any other method which reproduces the classification less well.

The Test Sets

- A set of 10453 compounds with 300 USAN activity classes which had at least four members. This was derived from medchem02.
- A much larger set derived from MDDR. This is fully described in Ifat's dissertation. We do not have rights to this database, so we cannot describe the results. Suffice it to say they were consistent with the smaller set results.

How good is it?

- The test was how well the method reproduced the known classes.
- The function was, for each class, to calculate the mean of the sum of proportions of clusters representing the class.
- This coefficient is 1.0 for a perfect match, even if the class is across multiple clusters.
- As we are summing the proportions it is susceptible to the Judas effect. See http://www.sciencemediacentre.org/downloads/communicating_risk.pdf

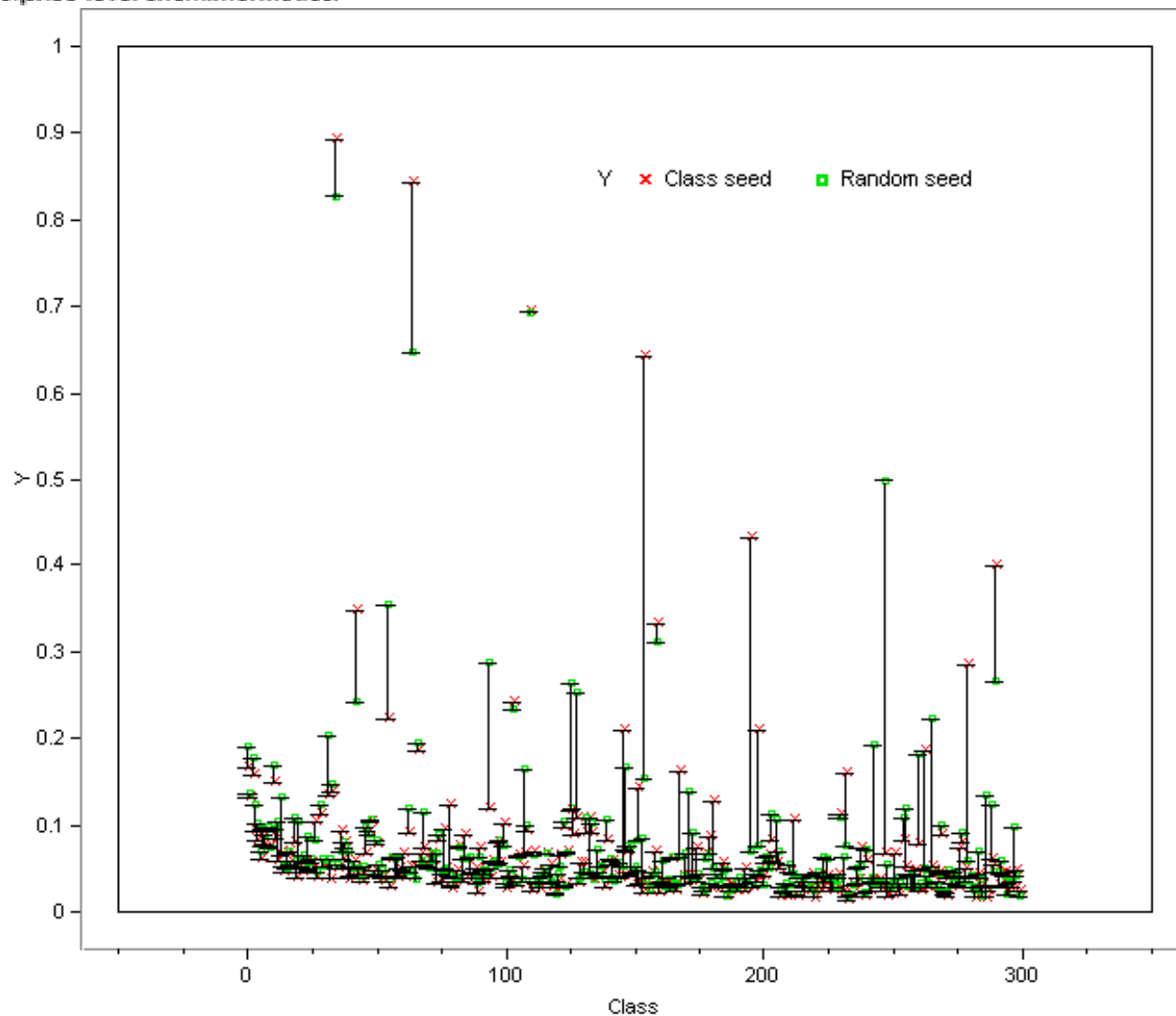
$$M_i = \sum_j \frac{\text{count_of_cluster}_j}{\text{size_of_cluster}_j} / \text{count_of_clusters_in_class}_i$$

Random versus Class

- A set of seeds was generated from the known classes by generating the modal fingerprint from the members of each class.
- A second set of seeds was generated using the random function in `kmodes()`. The input set was in thor-hashed order.
- Clustering was carried out exhaustively and the goodness measure calculated for each class.

Random versus Class

Enterprise-level cheminformatics.

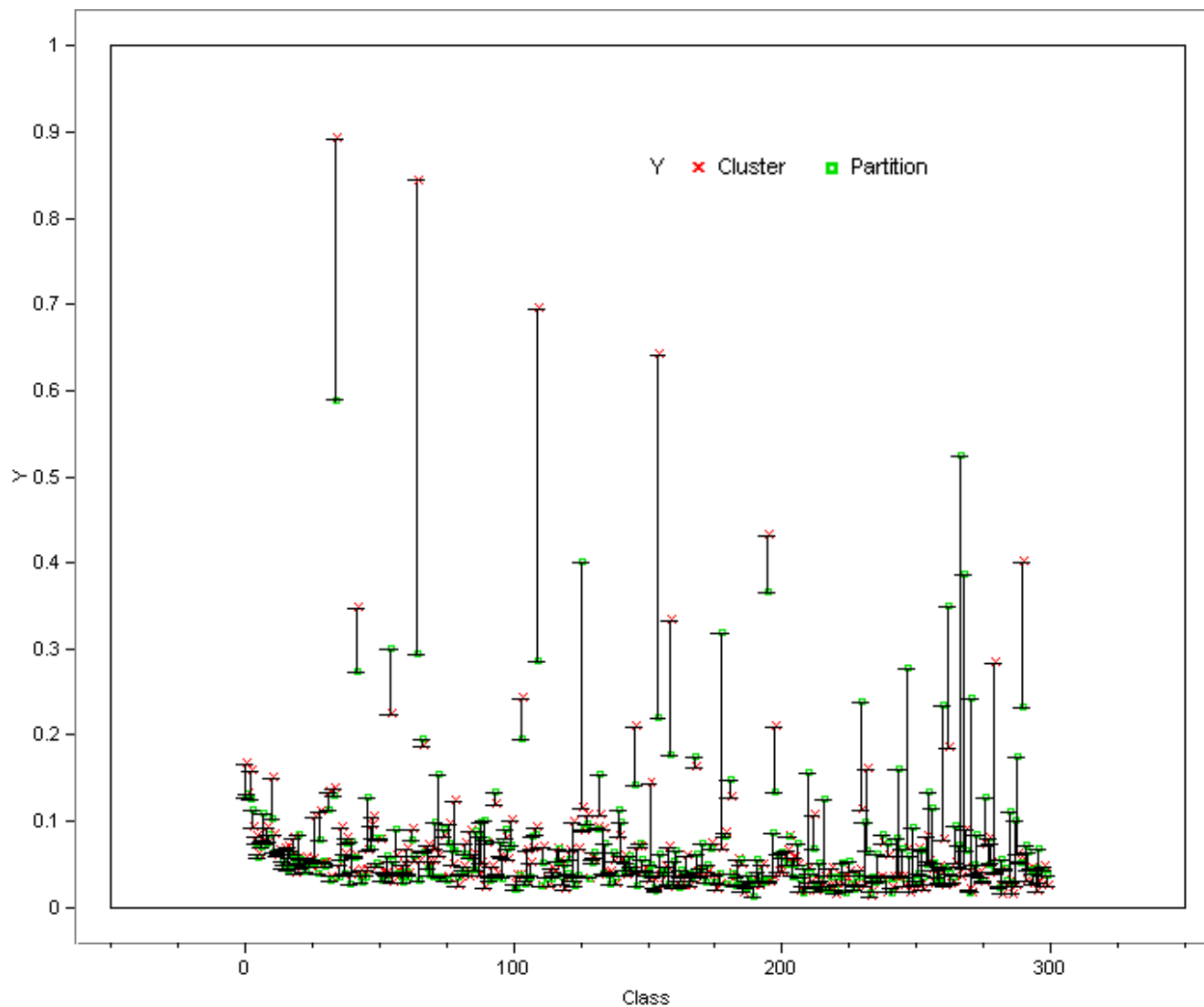


Random versus Class

- Neither method was particularly good at reproducing the known classes.
- There seemed no gain in having the *a priori* knowledge of class membership.
- In separate experiments where there was no relocation, simply partitioning to the nearest modal, there was little improvement in class recovery. Indeed clustering was marginally better
- This implies that compounds are more like the modals of other classes than the class they have been assigned to.
- Whilst this lack of selectivity accurately represents the real world, it is not what proponents of HTS would have one believe.

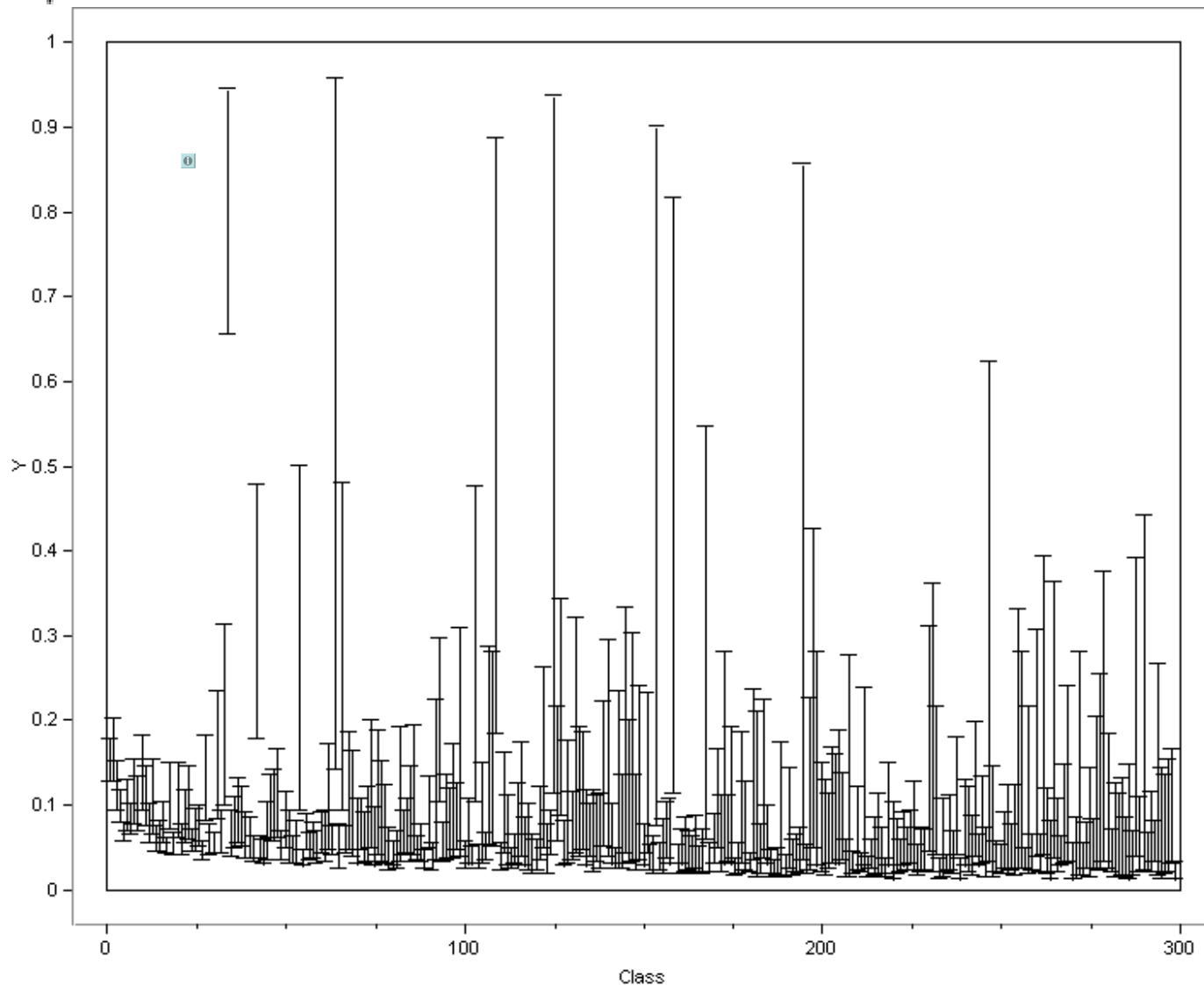
Partition versus cluster

Enterprise-level cheminformatics.



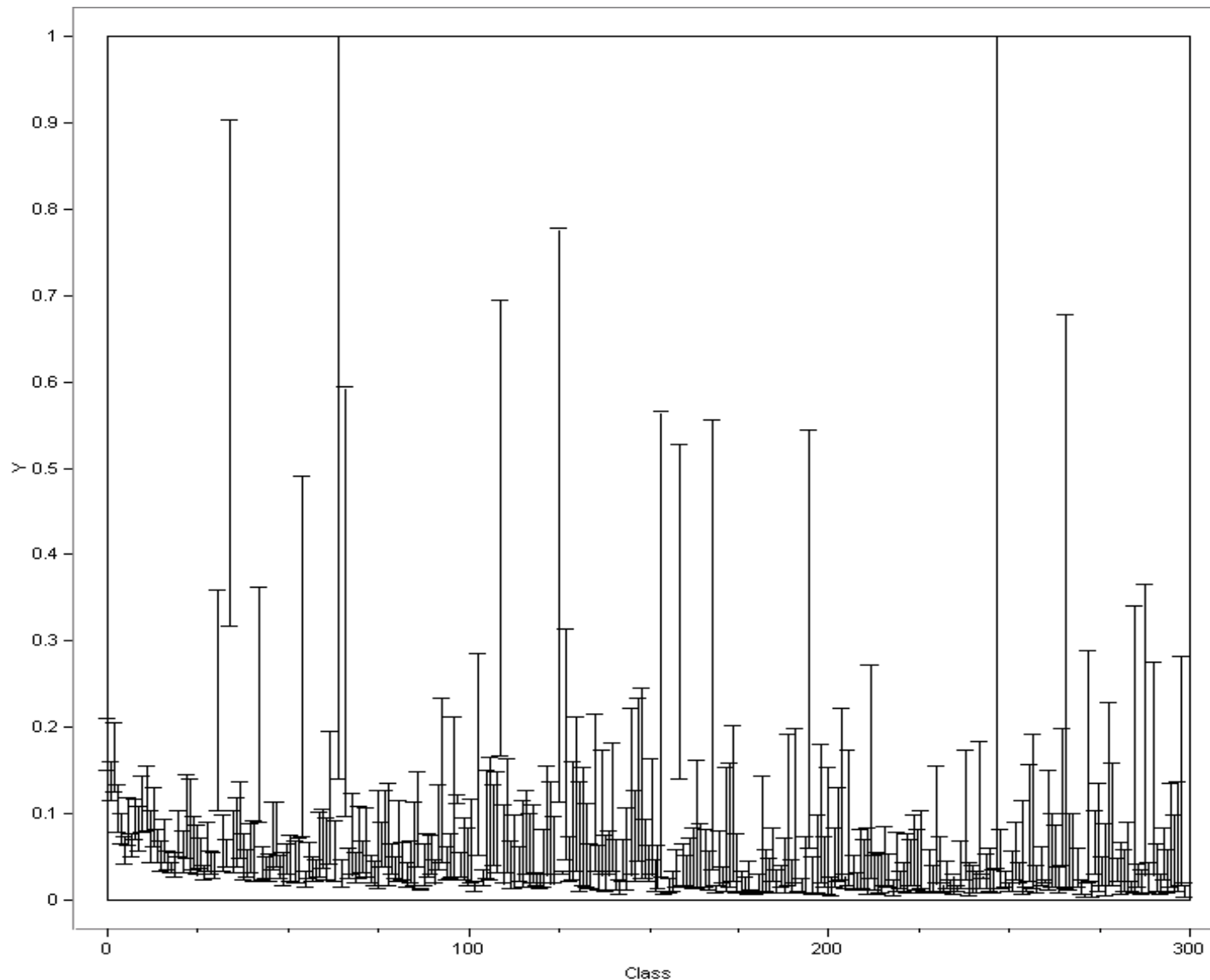
Try, try again. Results of 100 random starting sets

Enterprise-level cheminformatics.



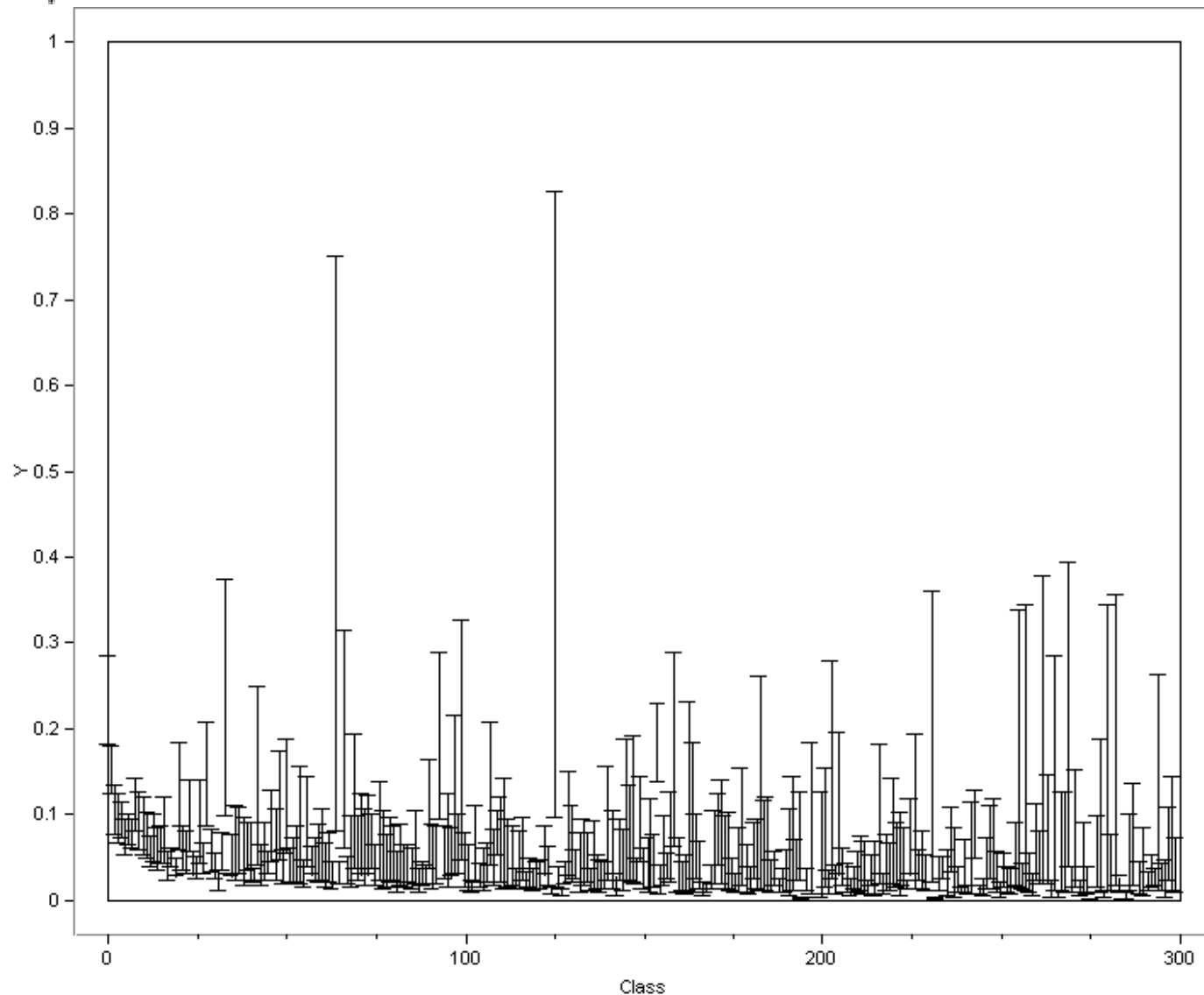
Fragment fingerprints

Enterprise-level cheminformatics.



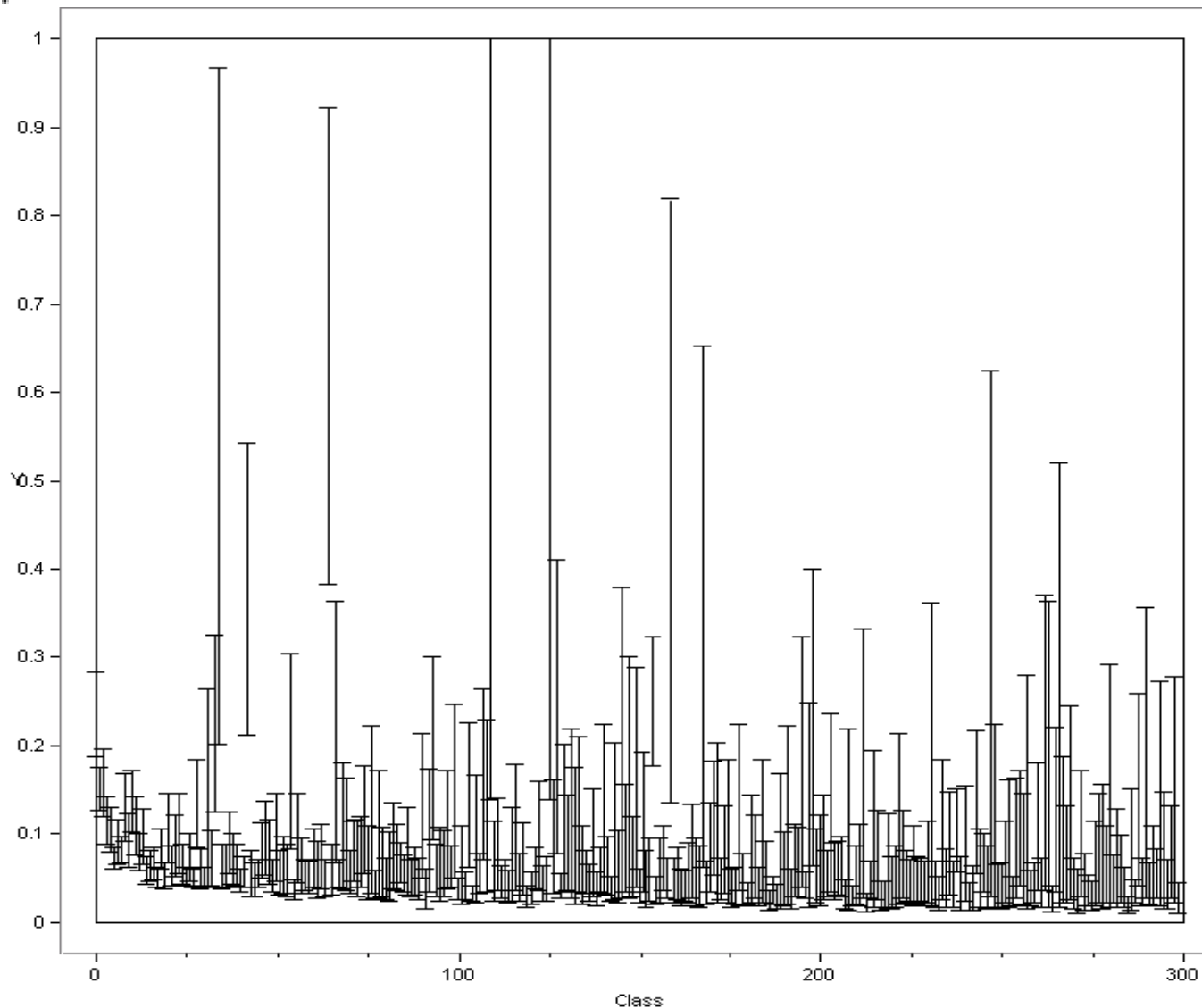
Ring only fingerprints

Enterprise-level cheminformatics.

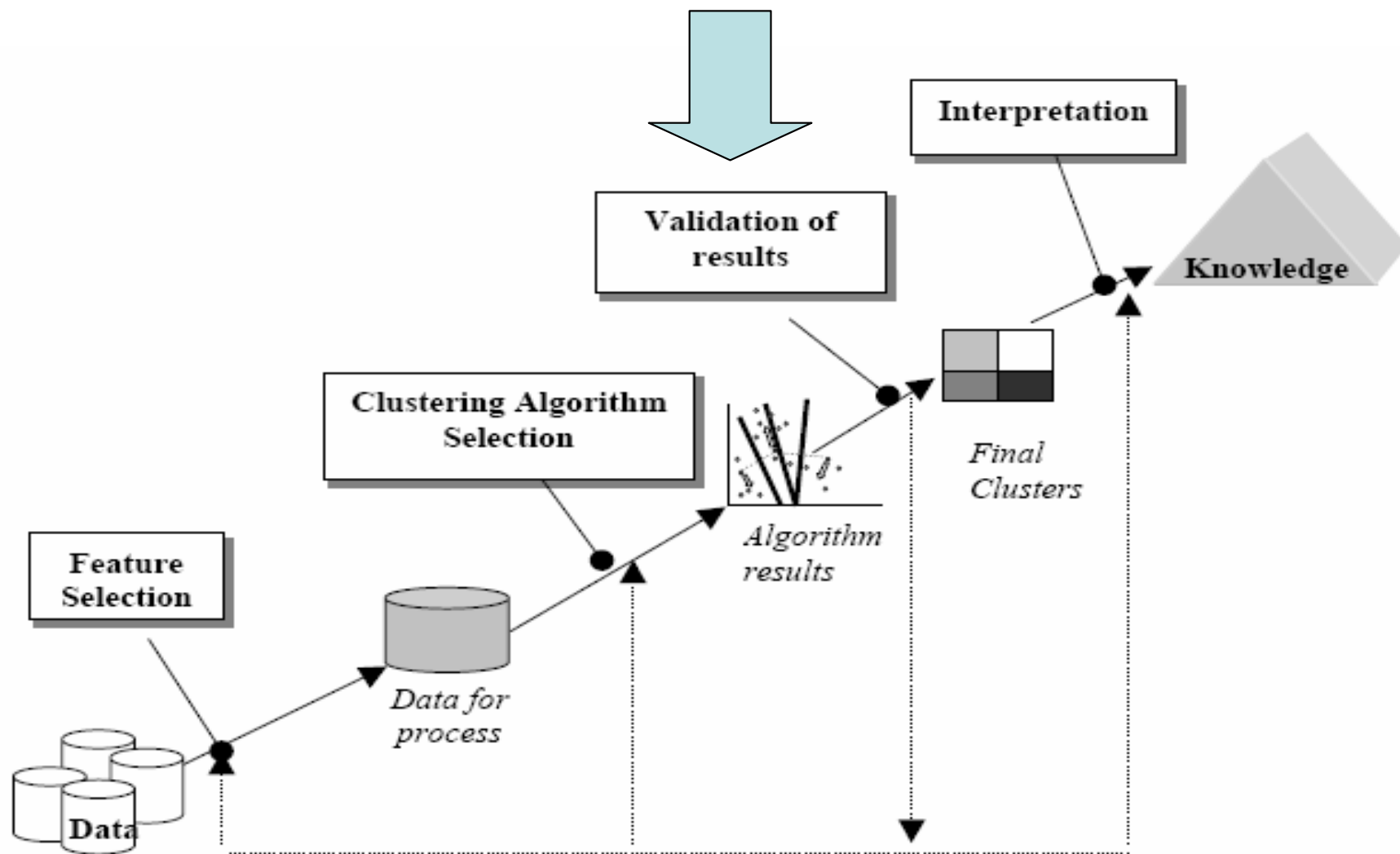


No_C_chain fingerprints

Enterprise-level cheminformatics.



The process



Validation of Results

- The major difficulty with validating results is that there is no independent measure of good clustering.
- Mostly there is an in-built assumption that the descriptors used are related to some property of interest. The ability to reproduce that property on test sets, becomes the only criterion.
- Aside from this, stopping rules which are algorithm dependent are quite often implemented.

k-modes stopping rules

- In the current version of *k*-modes the default stopping rule is when there are no more relocations.
- In additions users may
 - Stop when the relocations drop below a given percentage
 - Stop after the first allocation, with no relocations, partition mode

Comparison with classes

- The comparison of clustering and classification can be carried out in several ways.
- An asymmetric method driven by the classes has been described
- A more satisfying method is to calculate a ratio like

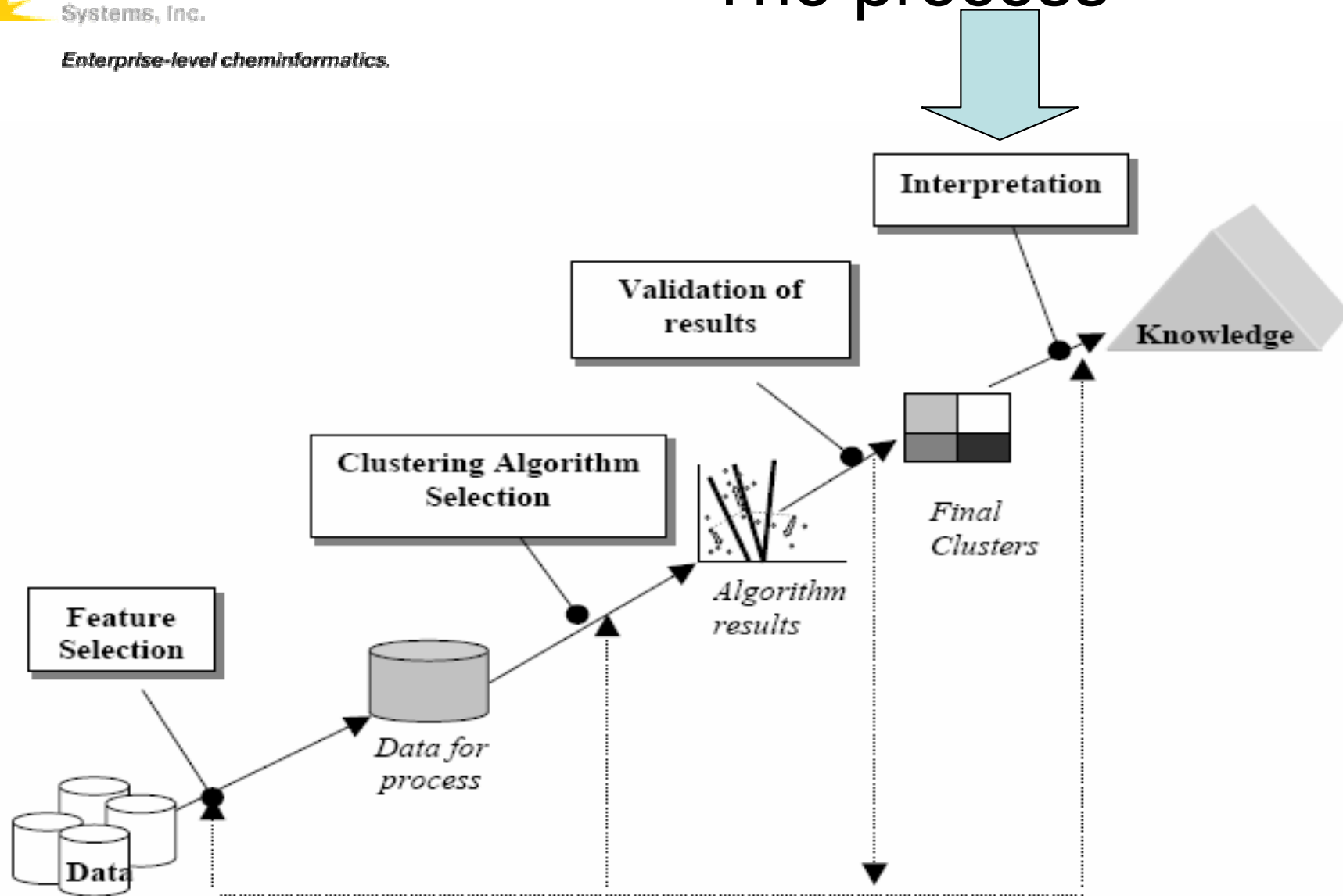
$$M_{ij} = \frac{\textit{compounds _ in _ common}_{class_i, cluster_j}}{\textit{compounds _ in _ class}_i _ \textit{or _ cluster}_j}$$

Effectively a Tanimoto/Jaccard coefficient

Comparison with classes

- Clearly one could get these values from “select” statements in Oracle, however...
- As we have tools to set bits in a Daylight fingerprint we could describe both the class and the cluster by the compounds they contain.
- **filter_fingertalk()** described earlier will read a space delimited set of integers and set the appropriate bits in a Daylight fingerprint.
- Pairwise similarity values between the clusters and classes can be determined using any of the coefficients one wished.
- Equally one could cluster the classes and clusters...

The process



Interpretation

- Much of the interpretation will be driven by the particular application.
- Quite often only a representative set is required.
- If, however, you co-cluster the clusters and classes, as described earlier, you could discover related classes which may have underlying common biology i.e. fresh knowledge.
 - Whilst in the test set described, compounds could only belong to one class, in general this will not be the case.
- Indeed using these tools one could cluster the data from multiple HTS runs against varying targets and investigate the relationships between the targets.
 - $2^{31} \approx 2.1$ billion exceeds all but the most ambitious screening set size
- The tools provided for the initial analysis can also be used to aid interpretation.

Summary

- Hopefully we have shown that clustering in Daylight is not Hobson's Choice. See <http://www.hobsonschoice.com/hobstory.html>
- By providing tools, in the Daylight tradition, rather than fixed applications, we believe users are now able to explore the full process of transforming data to knowledge.
- Over the next few releases we would hope to boost the tools available to users to handle sets of compounds and the relationships between them.