Patinformatics, LLC®
Data Driven Decisions
Patent Strategy and Analytics Services

Spatial Concept Mapping as a Patent Landscaping Task

Anthony Trippe
Managing Director – Patinformatics, LLC
WIPO Regional Workshop on Patent Analytics
Intellectual Property of the Philippines (IPOPHL)
Manila, Philippines – 5 December 2013

Overview

- Mapping is related to clustering or classification exercises, where the systems involved take the document clusters or classes and arrange them in 2-dimensional space by considering the similarity of the documents relative to one another over the entire collection

- Documents that share elements in common are placed closer together spatially, while ones with less similarity are placed further away

Two most Often used Clustering Algorithms are K-means and Force-Directed Placement

- K-means – a method of cluster analysis, which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean
- Force-Directed Placement – At the most basic level the algorithm tries to place similar objects close together and dissimilar objects far apart. The process is achieved by moving the objects randomly around the solution space via a technique similar to 'simulated annealing'. The criterion for moving a node is the minimization of energy

How ThemeScape Works

- The text engine scans through the document collection and automatically determines the distinguishing words or "topics" within the collection, based upon statistical measurements of word distribution, frequency, and co-occurrence with other words
- Distinguishing words are those that help describe how each document in the dataset is different from any other document
- In a dataset where every document mentions nanotech, "nanotech" wouldn't be a distinguishing word either

How ThemeScape Works

- The text engine uses these distinguishing words to create a mathematical signature for each document in the collection

- Then it does a rough similarity comparison of all the signatures to create cluster groupings

- The tool compares the clusters against each other for similarity, and arranges them in high-dimensional space (about 200 axes) so that similar clusters are located close together. The clusters can be thought of as a mass of bubbles, but in 200-dimensional space instead of just 3

How ThemeScape Works

- That high-dimensional arrangement of clusters is then flattened down to a comprehensible 2-dimensions, trying to preserve a picture where similar clusters are located close to each other, and dissimilar clusters are located far apart. Finally, the documents are added to the picture by arranging each within the invisible "bubble" of their respective cluster.

Patinformatics, LLC®
Data Driven Decisions
Patent Strategy and Analytics Services

Some Tips on ThemeScape Maps

- Use the best, most standardized text possible for clustering
- Consider using Mechanism of Action and Therapeutic Use fields in DWPI
- Routinely use Advantage, Novelty and Use fields as opposed to source abstracts or even full DWPI abstracts
- Be prepared to change the labels or create call outs

Patinformatics, LLC®
Data Driven Decisions
Patent Strategy and Analytics Services

A Representative ThemeScape Map

How STN AnaVist Works

- Documents are compared using cosine similarity and then positioned using force-directed placement
- Based on this explanation, it is probably correct to say that this method of generating a spatial concept maps is not technically clustering, since the documents are not initially partitioned, but since it is an unsupervised machine learning method, it is usually placed in that category

How STN AnaVist Works

- Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them
- The cosine of 0° is 1, and it is less than 1 for any other angle
- Force-directed placement position the nodes of a graph in two-dimensional or three-dimensional space so that all the edges are of more or less equal length and there are as few crossing edges as possible, by assigning forces among the set of edges and the set of nodes, based on their relative positions, and then using these forces either to simulate the motion of the edges and nodes or to minimize their energy

How STN AnaVist Works

- The software was significantly enhanced to improve visualization results, utilizing the expertise of our database building staff and scientists:

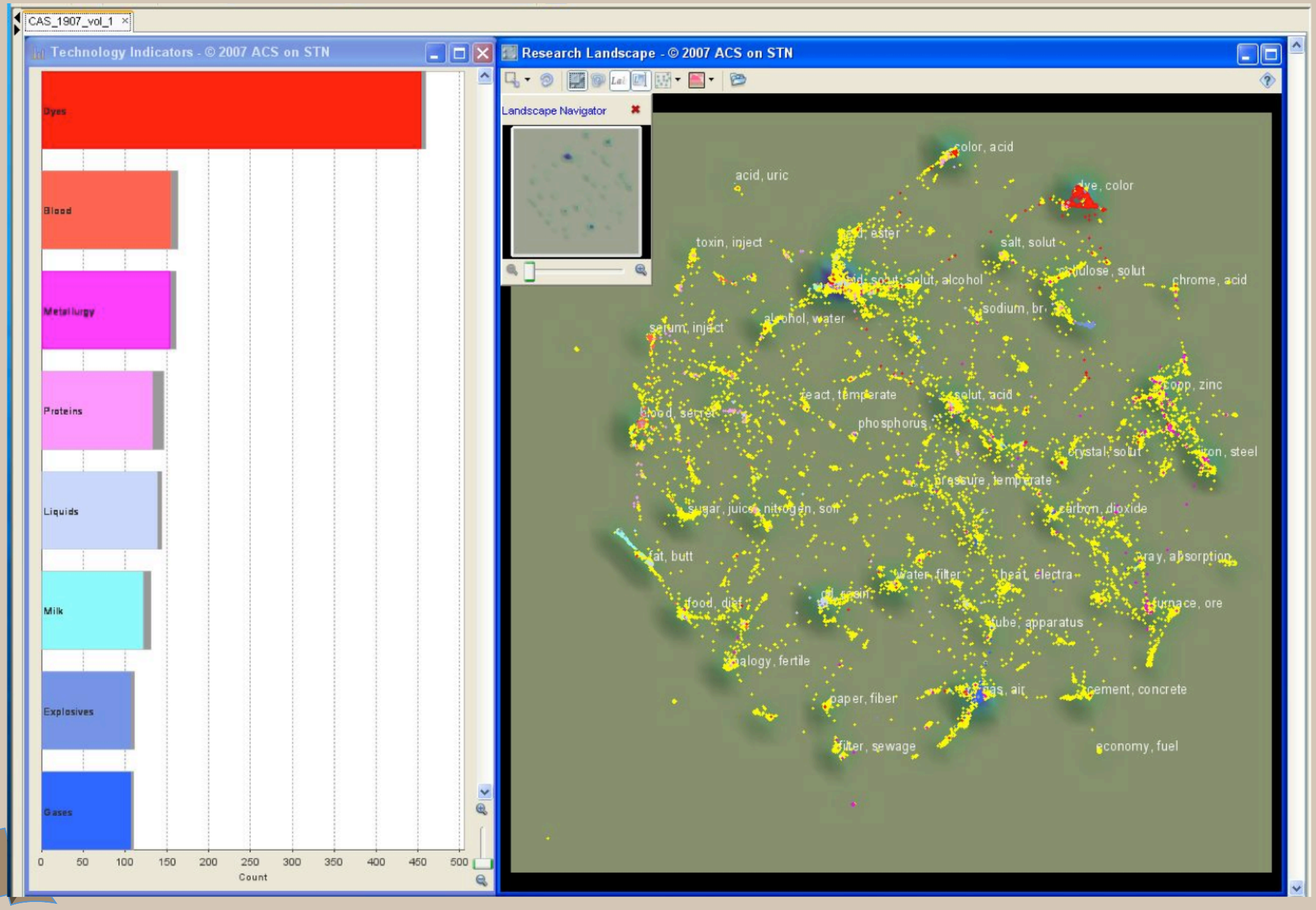
- CAS vocabulary to standardize the clustering concepts

- A stopword list to improve cluster results for sci-tech searches

- These enhancements allow for the software to produce more scientifically relevant clusters that are focused on scientific and intellectual property information

Patinformatics, LLC®
Data Driven Decisions
Patent Strategy and Analytics Services

How STN AnaVist Works



Regardless of Similarity Method All Tools Start with A Vector

- A vector is a mathematical concept, which represents the identifiers associated with each document that is going to be analyzed
- Theoretically, the total dimensionality of the vector is the number of distinct words occurring in the corpus, that are going to be used for comparing the documents
- If a discrete identifier occurs in a document, its value in the vector, that corresponds to the specific document, is non-zero
- Other identifiers, such as classification codes, or citations, can also be used, but for spatial concept maps the focus will be on words from the documents of interest

Term Frequency Inverse Document Frequency (tf-idf)

- tf–idf, term frequency–inverse document frequency, is a numerical statistic which reflects how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining
- The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others

Users can Impact Term Selection with Vectors with Stopwords

- Stopwords are also referred to as non-content bearing words, and they can adversely impact similarity measurements if they are included in the vector since they do not impart knowledge of the topic area
- Almost all mapping tools come with a list of standard stopwords, such as “the”, “and”, “a”, and other non-content bearing terms, but users can also look at initial results and identify words that do not add meaning to the technology being examined
- New words can be added to stopword lists within tools on a map-to-map basis, or permanently

Spatial concept maps can also be made using classification methods

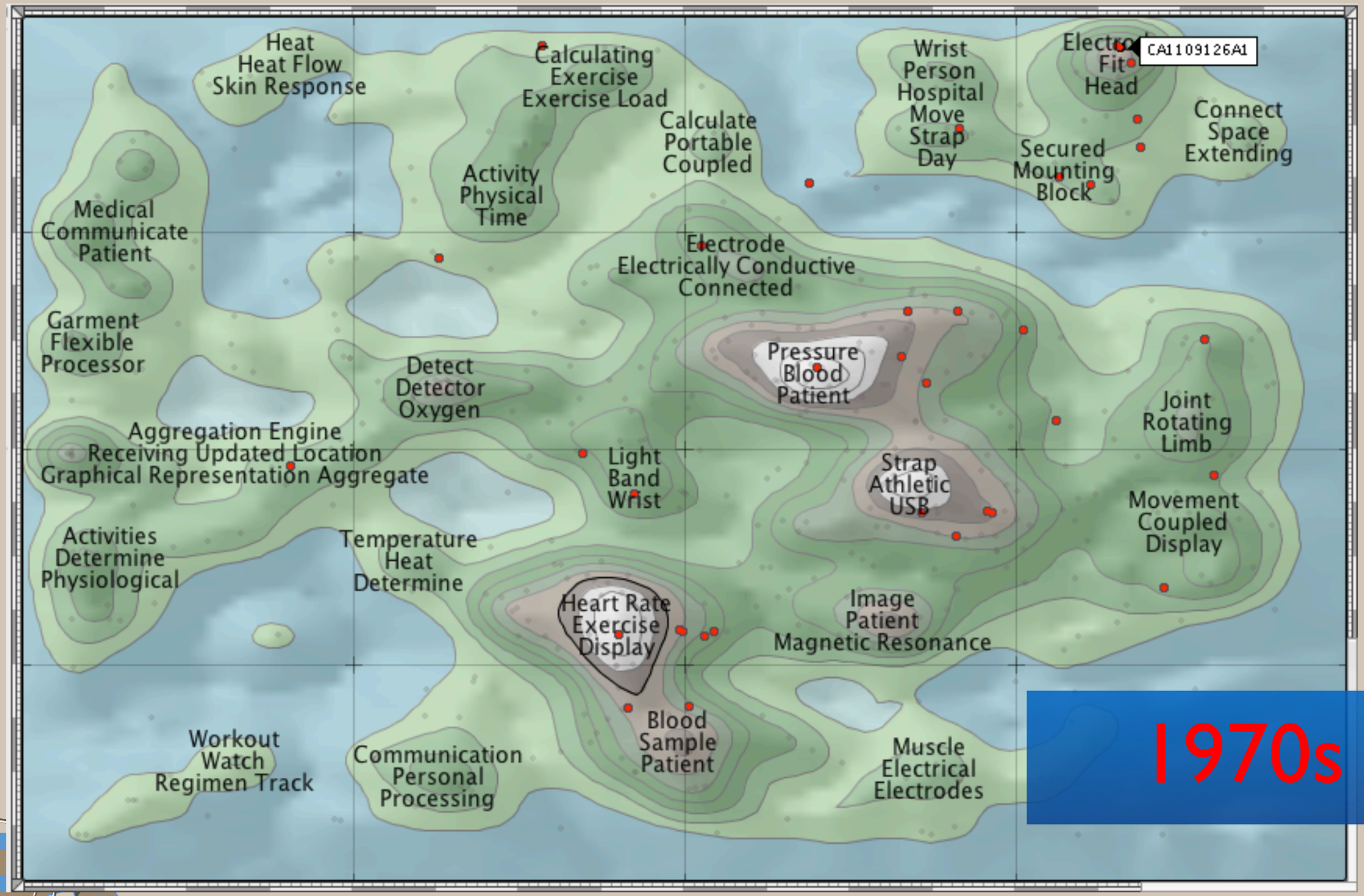
- Kohonen Self Organizing Maps – a type of artificial neural network (ANN) that is trained using unsupervised learning to produce a low dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map. Self-organizing maps are different from other artificial neural networks in the sense that they use a neighborhood function to preserve the topological properties of the input space

Layering Additional Information on Maps can add Context

- Most mapping systems provide a means to highlight, with the use of different colors, two or more patent assignee or periods of time within the collection used to generate the map
- These groupings are then laid over the existing map and can be used to provide context on when technology subsections were investigated, or which organizations were investing in different areas compared to others
- Try with Patent Assignees or Application Years

Thematic mapping of the domain can also highlight areas of focus

Thematic mapping of the domain can also highlight timing



Thematic mapping of the domain can also highlight timing

Document Density is Implied using the Third Dimension

- While the maps, and document organization, is provided in two-dimensions a third-dimension is often added, after the fact, by incorporating document density
- The number of documents, found in a cluster, or in reasonable proximity to one another, can be used to call out topics of higher interest than others in the collection
- On a topographical version of the spatial maps this is represented by an implied increase in peak heights on the map, visualized by a change in color

Contour Lines Do Not Normally Demonstrate Relationships

- Generally, these lines are drawn based on the distance between the document dots
- The distance between a dot and its nearest neighbor determines the boundaries of the lines
- Once the threshold is exceeded the line is drawn between the two dots
- It has been speculated that contour lines encompassing multiple groups on a map implies a relationship between the two groups, but generally, this is not the case, and the lines are simply based on the spread of the documents

Change Labels to Provide Project Specific Perspectives

- Most systems generate labels by looking at frequently used words, especially if they are unique to a particular cluster
- Sometimes this works well, but often the label terms are too generic and don't really reflect the contents of the cluster
- The clustering, in fact, may have been quite good, but a poor label may be the first, and only, thing that a client sees
- If the labels are poor, and don't reflect meaningful categories, the client can lose interest or believe that the map is not meaningful
- Labels can be changed within most mapping tools and should be done on a cluster-by-cluster basis by examining the titles of the individual documents within them

Maps are not Scatterplots and X & Y can not be Extrapolated

- There are no X and Y-axis associated with the spatial concept maps, and the distance between documents, usually represented by dots, are relative and based on similarity as previously discussed

- Since these distances are relative, and based on the contents of the collection, guesses cannot typically be made about what sort of document might occupy an empty space on the map