

## Successive Approximation Source Coding and Image Enabled Data Mining

Christopher F. Barnes  
Georgia Institute of Technology

Key elements of active *data flows* are observable subjects, sensors measuring subject data, human experts uncovering information buried in measured data, computing machines, algorithms supporting expert human analysis, and information consumers. Properly designed *data warehouses* capture all essential data flow elements: 1) sensor data, processed data, and derived expert information; and 2) state information about subjects, sensors, analysts, algorithms and consumers; and 3) quantified valuations of information products (from consumers' perspectives). Fundamental elements of data warehouses are *data-tuples* containing data (sensor, processed and derived), state information (subject, sensors, analysts, algorithms and consumer) and user valuations. A data-tuple is generated each time a subject is observed. Data-tuples are stacked to form *tables* with *rows* of data-tuples and *columns* of single elemental data types (text, numeric, images, video, etc). Data mining steps are 1) query a table to find all rows relevant to the query question, and 2) do statistical or pattern discovery analysis on some or all columns of the sub-table or *aggregate* resulting from the query's answer.

Conventional query tools like SQL form logic statements able to question data warehouse table columns in text or numeric data formats. Conventional tools do not exist to query columns containing images. Query tools to find, for example, "all images that look like this image" are desirable. Source coding provides an approach for such queries. Source coded representations of archived images (or blocks of images) are (also) stored as data-tuple elements. New images (or blocks of images) are coded with the same source code and their coded representations are compared (using SQL) with coded representations of archive images to find similarities. Vector quantization (VQ) source codes can support image similarity queries on an image block basis.

Successive approximation source codes have advantages over fixed rate source codes. First, low rate source code comparisons are computationally efficient, so it is less expensive to search for block similarities on every possible (overlapping) block position in an input query image ("not-similar" query answers take fewer compute cycles with low rate representations). Second, moderate rate source code comparisons are able to find "general" similarities. Third, high rate source code comparisons are able to find "quite similar" images. Fourth, variable-rate, embedded, successive approximation source codes provide query returns consisting of sequences of aggregate data-tuples with image sets that collectively vary from "somewhat similar" to "very similar." Fifth, a data mining statistical analysis or pattern search over a sequence of aggregates provides a sequence of data mining answers that is desirably fuzzy. A sequence of answers resulting from a single query readily shows trends in non-image data-tuple elements, uncovers biases, and exposes outliers in the data mining analysis. Residual vector quantization (RVQ) provides a successive approximation source code with utility in image enabled queries in image data mining tasks, especially when RVQs have more than two stages.