**Box 1. Definitions**

**Edit distance.** The number of edits (character insertions, deletions, or substitutions) needed to turn one string into another.

**Scale (in time and space).** The amount of time or space a task takes as a function of the amount of data on which it must operate. A task requiring time directly proportional to the size of the data is said to scale linearly; for example, searching a database takes twice as long if the database grows by a factor of two.

**Distance metric.** A measure of distance that obeys several mathematical properties, including the triangle inequality.

**Covering spheres.** We define a set of spheres around existing points such that every point is contained in at least one sphere, and no sphere is empty.

**Metric entropy.** A measure of how dissimilar a dataset is from itself. Defined as the number of covering spheres.

**Fractal dimension.** A measure of how the number of points in contained within a sphere scales with the radius of that sphere.

**Information-theoretic entropy.** Often used in data compression as a shorthand for the number of bits needed to encode a database, or a measure of the randomness of that database.

**Pattern matching.** Refers to searching for matches that might differ in specific ways from a query, such as wildcards or gaps, as opposed to searching for all database entries within a sphere of a specified radius as defined by an arbitrary distance function.