**Data Mining:** This term refers to the method of extracting useful models of data. Sometimes, a model are often a summary of the information, or it are often the set of most extreme features of the info .

**Bonferroni’s Principle:** If we are willing to look at as a stimulating feature of data something of which many instances are often expected to exist in random data, then we cannot believe such features being significant.

This observation limits our ability to mine data for features that aren't sufficiently rare in practice.

**TF.IDF:** The measure called TF.IDF lets us identify words during a collection of documents that are useful for determining the subject of every document. A word has high TF.IDF score during a document if it appears in relatively few documents, but appears during this one, and when it appears during a document it tends to seem repeatedly.

**Hash Functions:** A hash function maps hash-keys of some data type to integer bucket numbers. an honest hash function distributes the possible hash-key values approximately evenly among buckets. Any data type can be the domain of a hash function.

**Indexes:** An index may be a arrangement that permits us to store and retrieve data records efficiently, given the worth in one or more of the fields of the record. Hashing is a method to create an index.

**Storage on Disk:** When data must be stored on disk (secondary memory), it takes considerably longer to access a desired data item than if an equivalent data were stored in main memory. When data is large, it's important that algorithms strive to stay needed data in main memory.

**Power Laws:** Many phenomena obey a law which will be expressed as y = cxa for a few power a, often around −2. Such phenomena include the sales of the xth hottest book, or the amount of in-links to the xth most popular page.

**References:** Mining of Massive Datasets by Anand, Leskovec and Jeffrey