**Course: Advanced Bioinformatics**

**Module title: Steps for Creation of Biomedical Corpora**

**Module no. : 122**

**Steps for Creation of Biomedical Corpora**

Nature of data.

Standard datasets.

Formalism.

Users of dataset.

Evaluation measures.

**Data Mining**

Discovers unsuspected associations.

Combines & links facts and events.

Discovers *new* knowledge, finds new associations.

**IR:** yields all relevant Corpora; gathers, selects, filters documents that may prove useful.

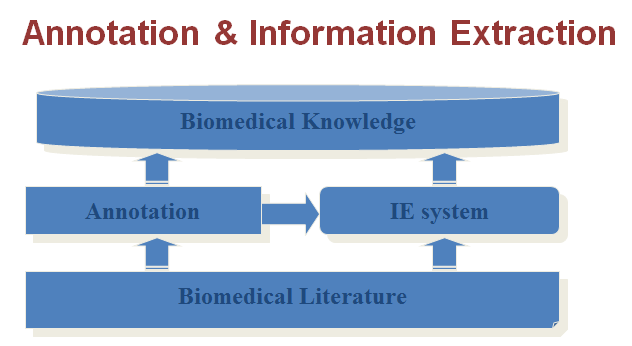
Finds what is known.

**IE:** extracts facts & events of interest to user.

Finds relevant concepts, facts about concepts.

Finds only what we are looking for.

**Annotation & Information Extraction**

****IE systems can be developed by referencing annotated corpus.

The performance of IE systems can be evaluated by being compared to the annotated corpus.

**Text Mining Pipelines**

* Text documents
* Retrieval/storage Index access relevant storage
* Text Processing: word Filters, Pattern filters, Lexicon matching, Ontology, NLP parsing etc.

**Feature Extractions:**

**Statistical:** Word counts, pattern extraction & counts, etc.

**Domain-specific:** Gene Name counts, etc.

**NLP-specific:** Phrase counts, etc.

**Data Mining:**

Classification, Clustering, Association, Statistical Analysis, Visual Analysis, etc. …