**Course: Advance Bio Informatics**

**Module Title: MedTag**

**Module No: 123**

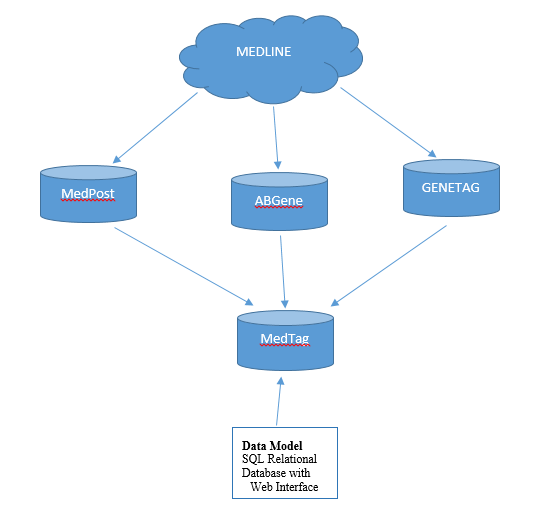
**MedTag**

* Collection of biomedical annotations
* MedTag is a database that combines three corpora:
  + MedPost
  + ABGene
  + GENETAG
* Data is available in flat files + software to facilitate loading data into SQL database

**MedPost:**  6700 sentences. Annotated with POS.

**ABGene:** 4000 sentences. Bio creative workshop. ABGene tagger to recognize single genes/protein in MEDLINE records

**GeneTag:** 20000 sent. Annotated for multiple gene/protein names. NER tagger



**Usefulness**

* Training and evaluating POS taggers
* Training and evaluating gene/protein NER taggers
* Usage in medical text
* Feature extraction for ML applications
* Entity annotation guidelines