ClearTK: A Framework for Statistical Biomedical Natural Language Processing

Philip Ogren

Philipp Wetzler

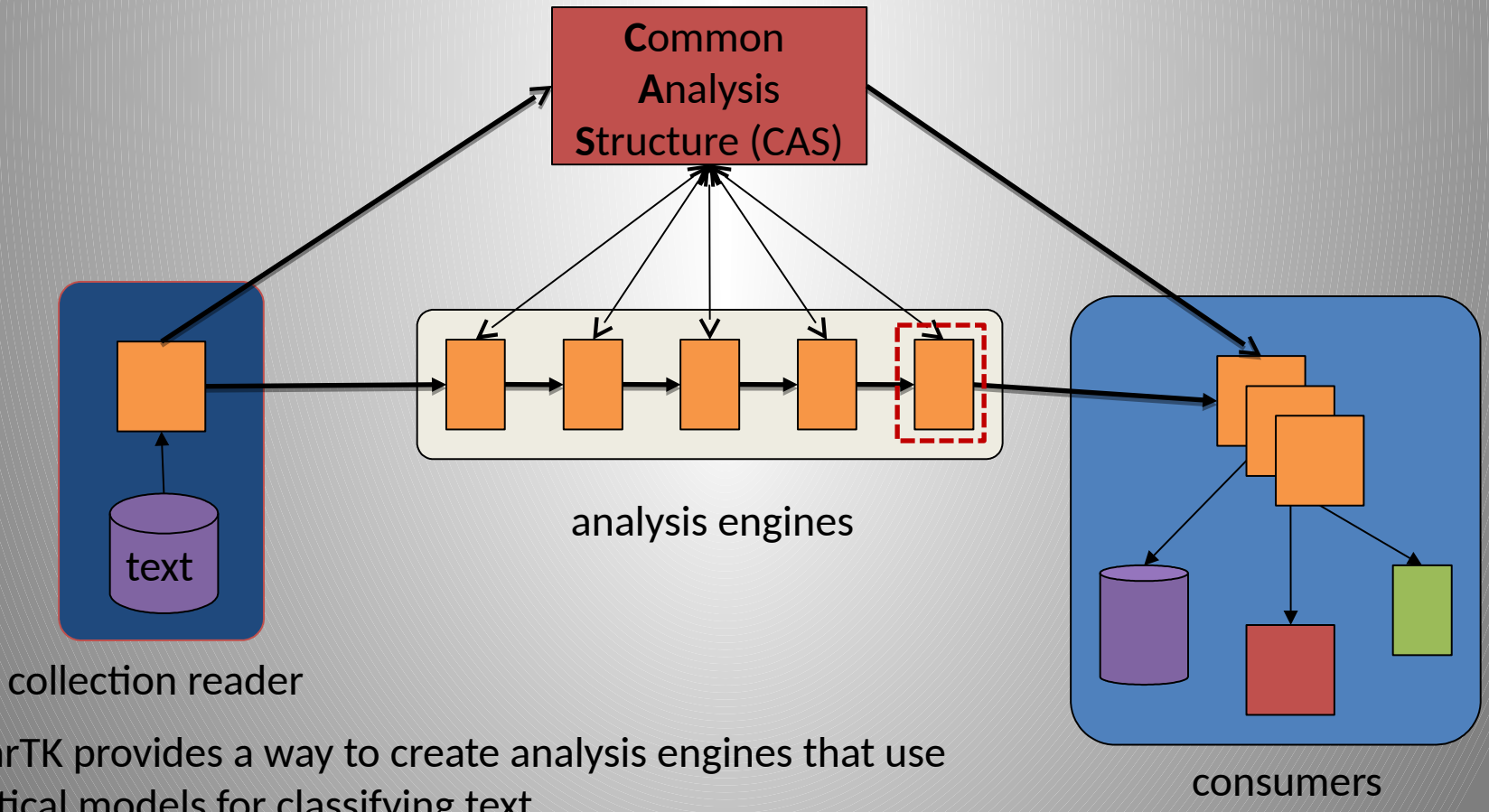
Department of Computer Science
University of Colorado at Boulder

Introduction

- ClearTK is a software package that:
  - facilitates statistical biomedical natural language processing
  - is written for UIMA
    - Java
  - Provides extensible feature extraction library
  - Interfaces with popular machine learning libraries
    - Maximum Entropy (OpenNLP)
    - Support Vector Machines (LIBSVM)
    - Conditional Random Fields (Mallet)
    - Misc. –e.g. Naïve Bayes (Weka)
- Available free for academic research

  (contact philip@ogren.info)

UIMA 101



- ClearTK provides a way to create analysis engines that use statistical models for classifying text.
- The structure of the CAS is defined by a **type system** determined by the development team.

Statistical Biomedical Natural Language Processing 101

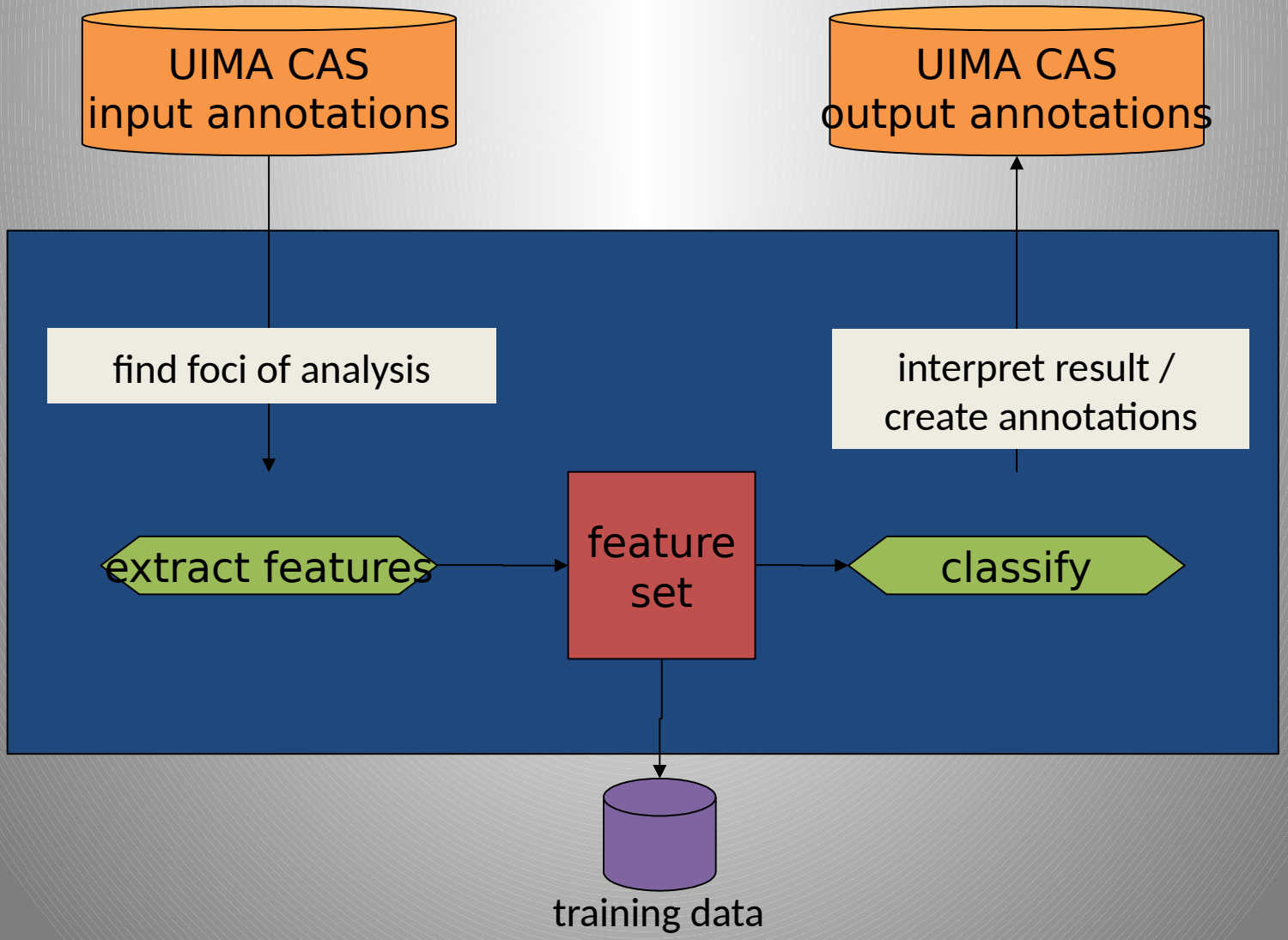
- Frame NLP task as classification task – e.g. For named entity recognition classify tokens as one of “B”, “I”, or “O”.

The concentration of **alpha 2-macroglobulin**, **alpha 1-antitrypsin**, **plasminogen**, **C3-complement**, **fibrinogen degradation products** (**FDP**) and fibrinolytic activity…

- Training
  - Manually annotate a bunch of data
  - Extract features from text *
  - Write out training data *
  - Train a model
- Run time
  - Extract features from unseen text *
  - Classify features with trained model*
  - Create annotations

* ClearTK facilitates these tasks

ClearTK Analysis Engine



ClearTK Summary

- Provides a framework that simplifies feature extraction and interfacing with a wide variety of machine learning libraries.
- Is not dependent on any specific type system
- Provides sophisticated feature extractors.
- Provides infrastructure supporting core library (i.e. collection readers, analysis engines, consumers, etc.)
- Well documented and unit tested.