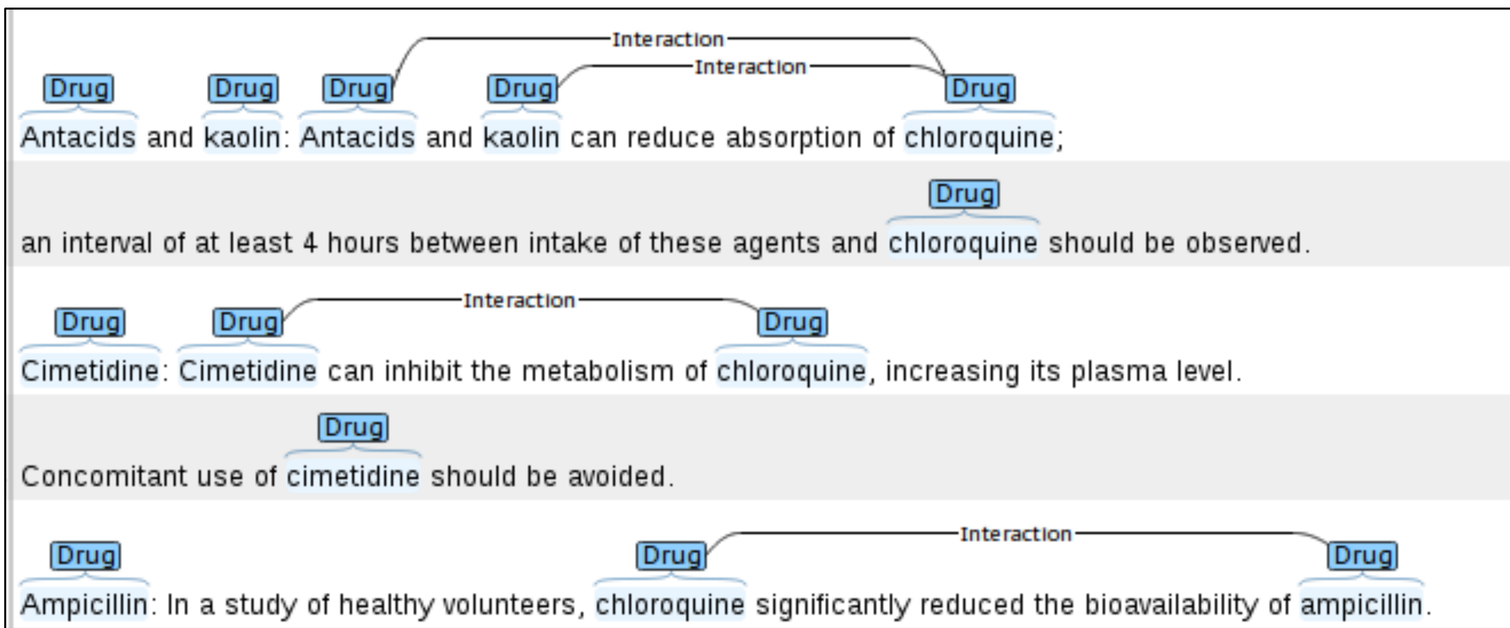
Named Entity Recognition

Renaud Richardet
BlueBrain
(former Ph.D. student)

Jean-Cédric Chappelier
LIA

Named Entity Recognition (NER)

- Named entity: an atomic element in text belonging to predefined categories
- E.g. names of persons, organizations, locations, proteins



Challenges in NER

- Variation of NEs – e.g. John Smith, Mr Smith, John.
- Ambiguity of NE types
  - John Smith (company vs. person)
  - Washington (person vs. location)
  - 1945 (date vs. time)
  - May (person vs. month)
- Ambiguity with common words, e.g. “may” or “the” is a protein name

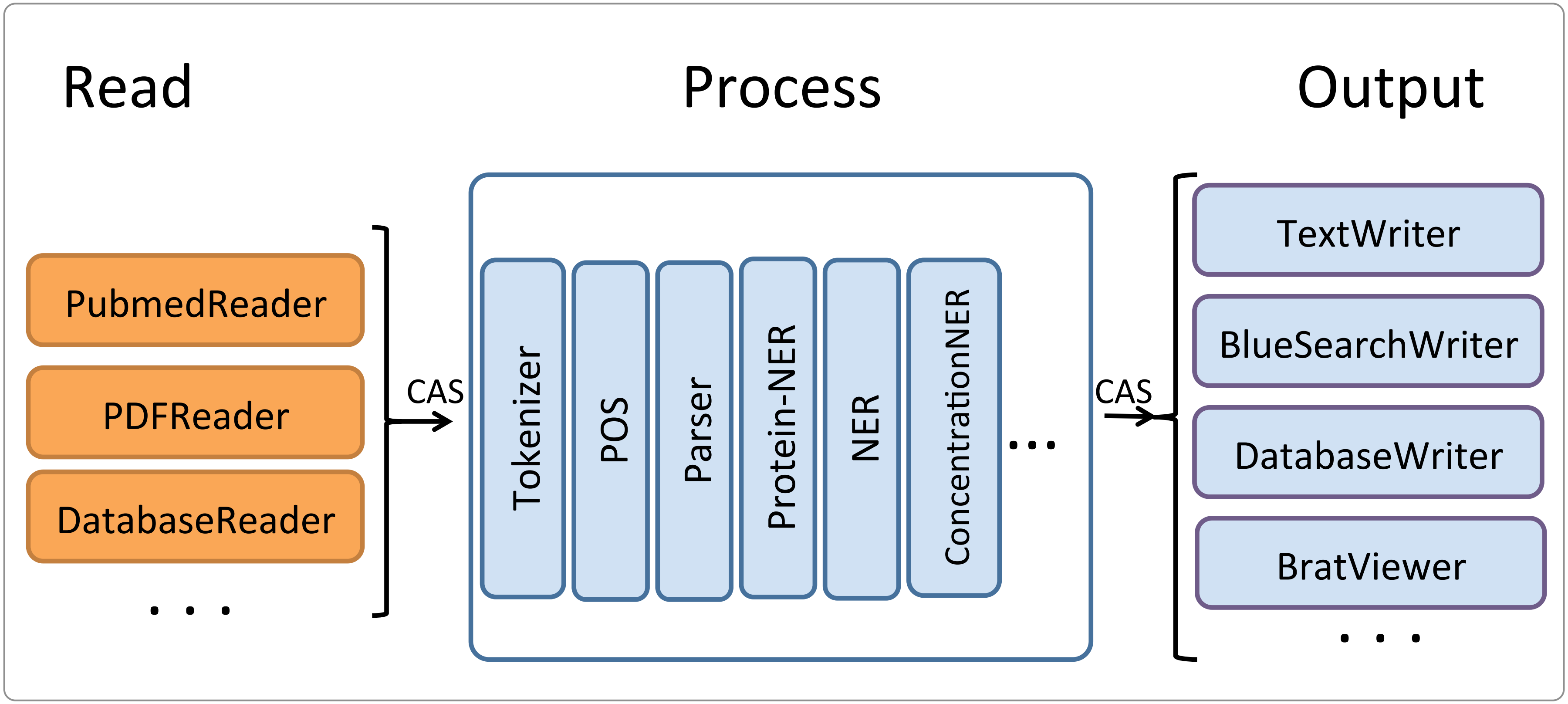
3 Methodological Approaches for NER

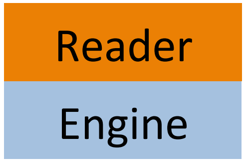
1) Regex	Hand-written regexes for complex but regular entities. Example: **measures** (units and numbers, ratios)
2) Lexicon	Matches occurrences of lexical entries in text. Some preprocessing (stemming, rewriting, synonyms) to increase matching. Example**: brain regions**
3) Machine Learning	ML models (MaxEnt, HMM, conditional random fields CRF), trained and evaluated on corpus. Feature selection based on domain knowledge. Requires (costly) annotated data Example: **brain region**

BioNLP

- NLP in the biomedical domain, e.g.:
  - identify NEs (proteins/gene, species, methods, …)
  - extract events (protein-protein interactions, …)
  - create a knowledge base of the concentrations of proteins in the different brain cell types
- Why is BioNLP important?
  - Most valuable knowledge in **text form** within papers
  - 1 new paper **each minute** on PubMed (average)
  - Synonyms, Homonyms, New terms

UIMA Workflow





CAS = Common Analysis Structure
= object model passed btw. modules

2^{nd} Approach: Lexicon-based NER to extract Brain Regions

1. *Whitetext* annotated corpus, for validation
2. *Neuronames* lexicon
3. UIMA dictionary lookup tool, to annotate corpus with lexicon

NeuroNames as a Lexicon

- integrated nomenclature for structures in the brain and spinal cord
- four species: human, macaque, rat and mouse
- > 15,000 neuroanatomical terms
- 550 primary structures with hierarchical relations to all other structures
- standard English and Latin names for > 850 structures
- acronym
- 9,000 synonyms
- alternate definitions (homonyms)

NeuroNames as a Lexicon


```xml
<token canonical="lingual white matter" ref_id="2">
  <variant base="Substantia medullaris lingualis" />
  <variant base="lingual white matter" />
</token>
<token canonical="isthmus of cingulate white matter" ref_id="3">
  <variant base="Substantia medullaris isthmus cinguli" />
  <variant base="isthmus of cingulate white matter" />
</token>
<token canonical="supramarginal white matter" ref_id="4">
  <variant base="supramarginal white matter" />
  <variant base="Substantia medullaris supramarginalis" />
</token>
<token canonical="precentral white matter" ref_id="5">
  <variant base="pre-central white matter" />
  <variant base="Substantia medullaris precentralis" />
  <variant base="precentral white matter" />
</token>
<token canonical="posterior orbital gyrus" ref_id="6">
  <variant base="posterior orbital gyrus" />
  <variant base="Gyrus orbitalis posterior" />
</token>
```

WhiteText Annotated corpus

- annotated corpus of brain regions
- 1377 PubMed abstracts from Journal of Comp Neurology
- 17’585 brain region mentions; abbreviations expanded
- IAA (evaluated on subset of docs) 90.7% and 96.7% for strict and lenient matching respectively

WhiteText Annotated corpus


```xml
<?xml version="1.0" encoding="UTF-8" ?>
<PubmedArticles>
    <PubmedArticle>
        <PMID>1692855</PMID>
        <ArticleTitle>Intermediate and deep layers of the macaque <BrainRegion>
        superior colliculus</BrainRegion>: a Golgi study.</ArticleTitle>
        <AbstractText>We studied the intermediate and deep layers of the
        macaque <BrainRegion>superior colliculus</BrainRegion> by means of the
        Golgi technique in an attempt to better understand the structural
        features of this important oculomotor center. For this study, we
        examined the optic (<BrainRegion>stratum opticum</BrainRegion>, SO),
        intermediate gray (<BrainRegion>stratum griseum intermedium</
        BrainRegion>, SGI), intermediate white (<BrainRegion>stratum album
        intermedium</BrainRegion>, SAI), and deep gray (<BrainRegion>stratum
        griseum profundum</BrainRegion>, SGP) layers. These are the four
        layers in which neurons having saccade-related activity are localized.
        We identified eight neuronal types on the basis of differences in
        somatic and dendritic morphologies: large multipolar neurons (Type I);
        large pyramidal neurons (Type II); large fusiform neurons (Type III);
        medium fusiform neurons with spiny, radially oriented dendrites (Type
```

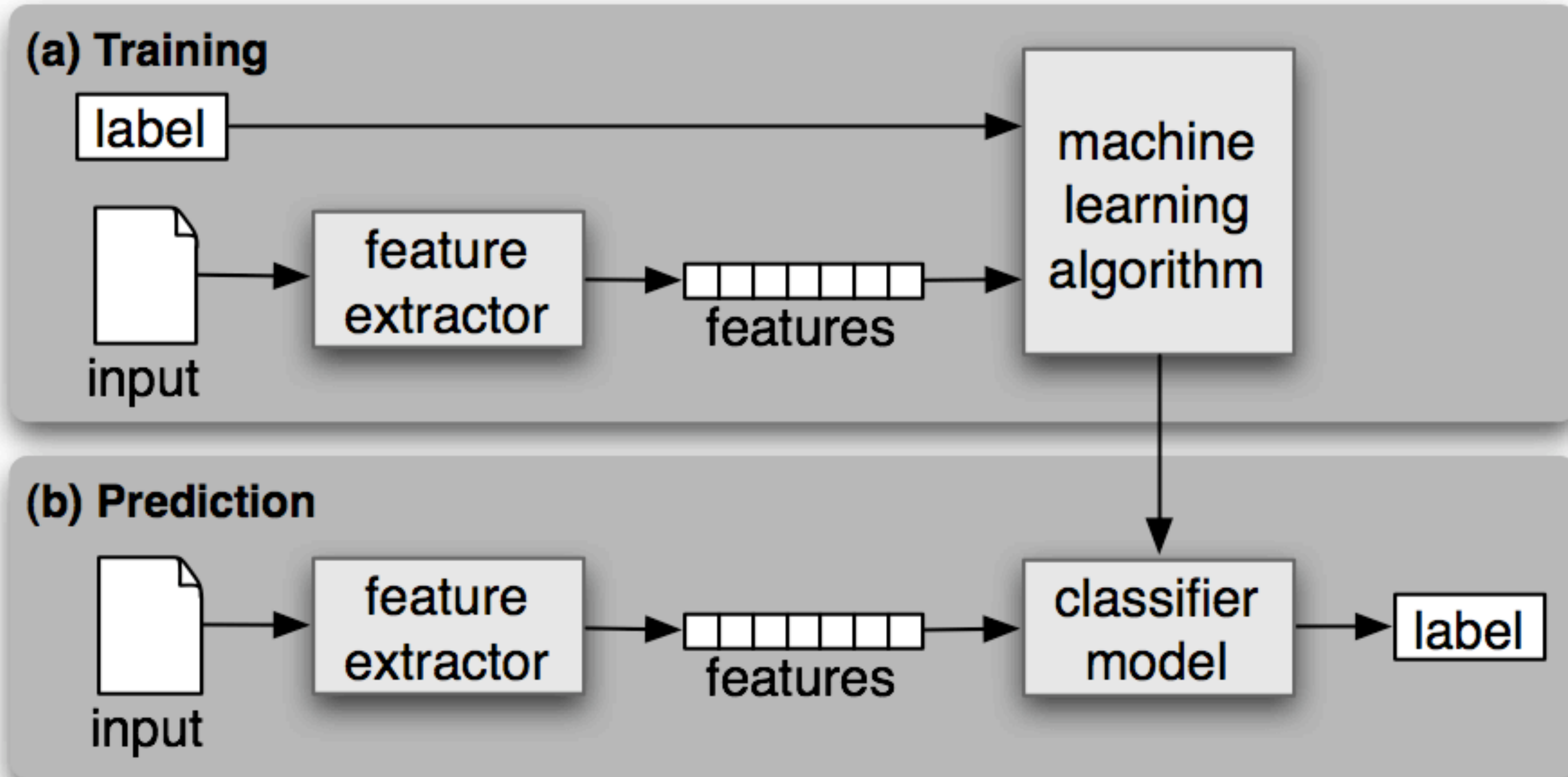
UIMA ConceptMapper

- dictionary lookup tool
- configuration (more)
  - caseMatch:
    - ignoreall - fold everything to lowercase for matching
    - insensitive - fold only tokens with initial caps to lowercase
    - digitfold - fold all (and only) tokens with a digit
    - sensitive - perform no case folding
  - StopWords: a list of words to be ignored in dictionary lookups
  - Stemmer: stemmer class to use before matching
  - OrderIndependentLookup: If "True", "foo bar" would equal "bar foo"
  - SearchStrategy:
    - ContiguousMatch - longest match of contiguous tokens
    - SkipAnyMatch - longest match of not-necessarily contiguous tokens
    - SkipAnyMatchAllowOverlap - longest match of not-necessarily contiguous tokens
  - FindAllMatches: If False, only the longest matches are found

3^{rd} Approach: CRF-based NER to extract Brain Regions

1. *Whitetext* annotated corpus, for validation
2. Mallet CRF library to train, evaluate and perform inference
3. *Feature engineering*

Supervised Classification

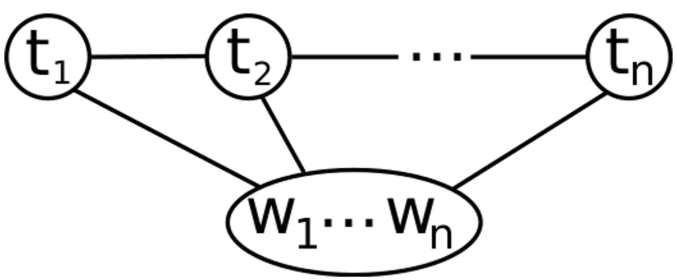


source: NLTK

CRF vs HMM

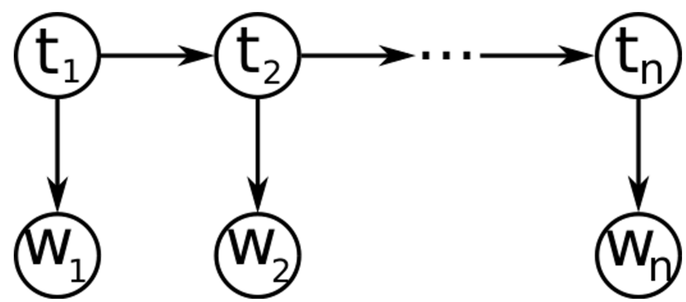
(linear) CRF are a (discriminative) generalization of HMM where "features" no longer needs to be state-conditionnal probabilities (less constraint):

$$P(T_1^n | w_1^n) = \prod_{i=2}^{n} P(t_i, t_{i-1} | w_1^n)$$



CRF

$$P(T_1^n, w_1^n) = \prod_{i=1}^{n} P(w_i | t_i)\, P(t_i | t_{i-1})$$



HMM

Feature Engineering

Text-based	Lexicon-based	Regexes-based
text	common brain region prefixes	areas, spine
lemma	neurological directions	substring
POS	stopwords	AllCaps
	TextPresso	Bracketed
	AllenBrainAtlas	Percent
	NeuroNames	
	BAMS	

Window: 2 before, 2 after