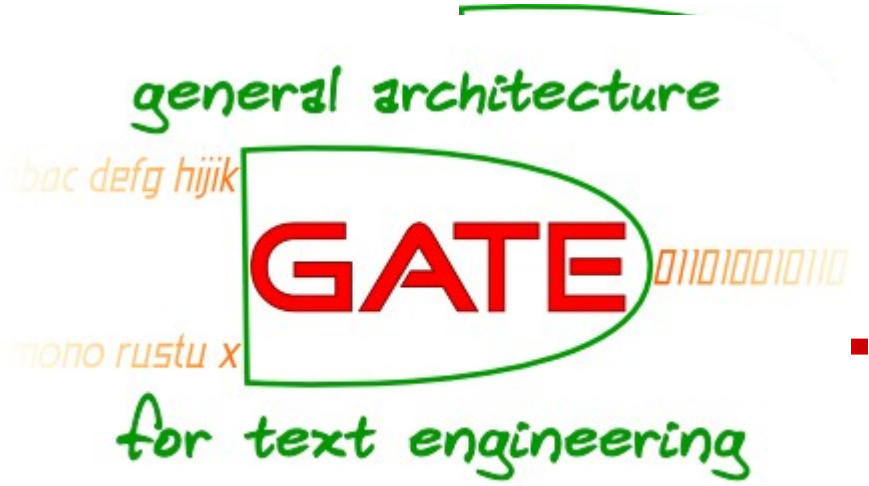



# Tackling text: mining unstructured data in the CRIS system

Angus Roberts
University of Sheffield


# Acknowledgements


- University of Sheffield
  - Natural Language Processing Group
  - GATE Team
- Biomedical Research Centre at SLAM / IOP
- Initial stages: Ontotext


# Outline

- Why tackle text?
- CRIS and GATE
- A clinical text mining landscape
- Two approaches
  - Pattern matching
  - Supervised machine learning
- GATE on the BRC cluster


# The challenge

- A lot of free text in mental health records
- Much information of value is not in the structured data
  - Few laboratory tests
  - Emphasis on relatively subtle symptomatology and overlapping diagnoses
- A clear need to extract structure from free text
  - Outcomes (e.g. cognitive function)
  - Context (e.g. education)
  - Presentations (e.g. symptoms)
  - Risk profiles (e.g. smoking)

# Free text vs structured data: MMSE coverage


| | Cases | Instanses |
|---|---|---|
| MMSE in structured data | 4000 | 5792 |

# Free text vs structured data: MMSE coverage


| | Cases | Instanses |
|---|---|---|
| MMSE in structured data | 4000 | 5792 |
| Text retrieved containing the string ”MMSE” | 16585 | 48805 |

# Free text vs structured data: MMSE coverage


| | Cases | Instanses |
|---|---|---|
| MMSE in structured data | 4000 | 5792 |
| Text retrieved containing the string ”MMSE” | 16585 | 48805 |
| MMSEs with dates text mined and validated | 15364 | 34871 |

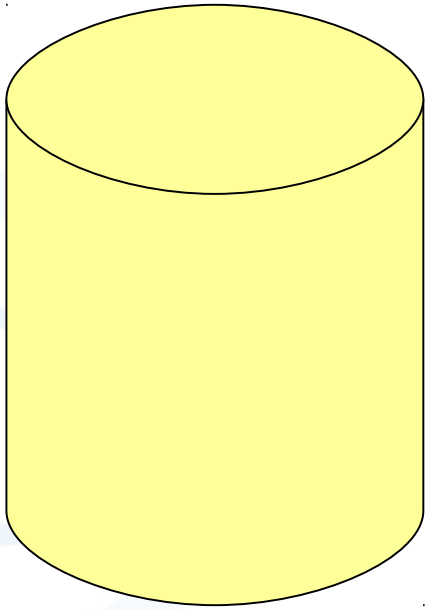

# Text mining in CRIS

- NIHR funded Biomedical Research Centre at South London and Maudsley NHS Trust / KCL IoP
- Part of a long term project to provide data for mental health epidemiology centred around a Case Register of previous cases - CRIS
- Specific targets in text
- But targets not known in advance
- Therefore required text mining with flexibility – a text mining capability
- Proof of concept started in 2009

# Context: the source mental health EHR system (SLAM )

National Institute for Health Research

**EHR**
**The Patient Journey System (PJS)**



**Coverage: Lambeth, Southwark, Lewisham, Croydon**

**Local population: c. 1.1 million**

**Clinical area: specialist mental health**

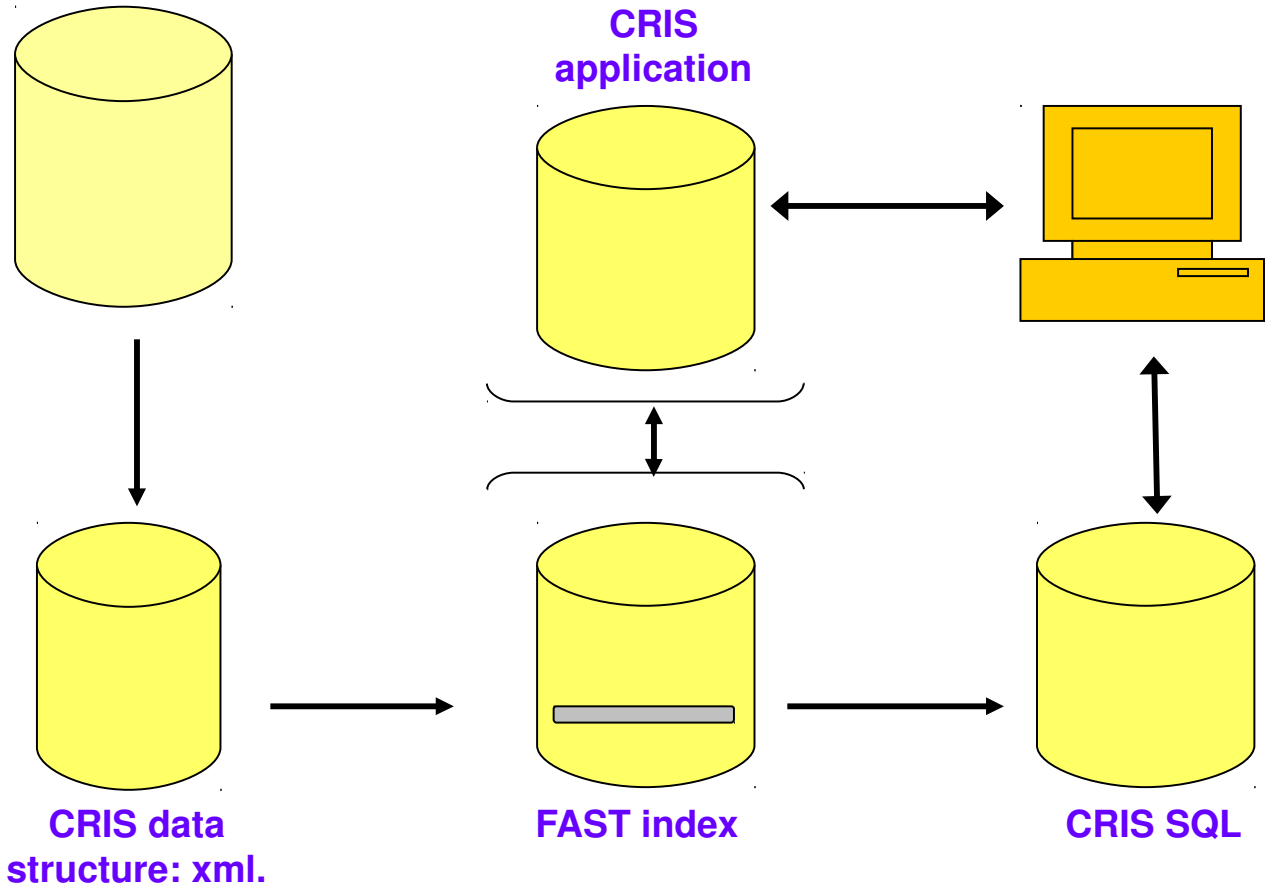
**Active patients: c. 35000**

**Total inpatients: c. 1000**

**Total records: c. 180000**


# CRIS architecture

EHR: The Patient Journey System (PJS)



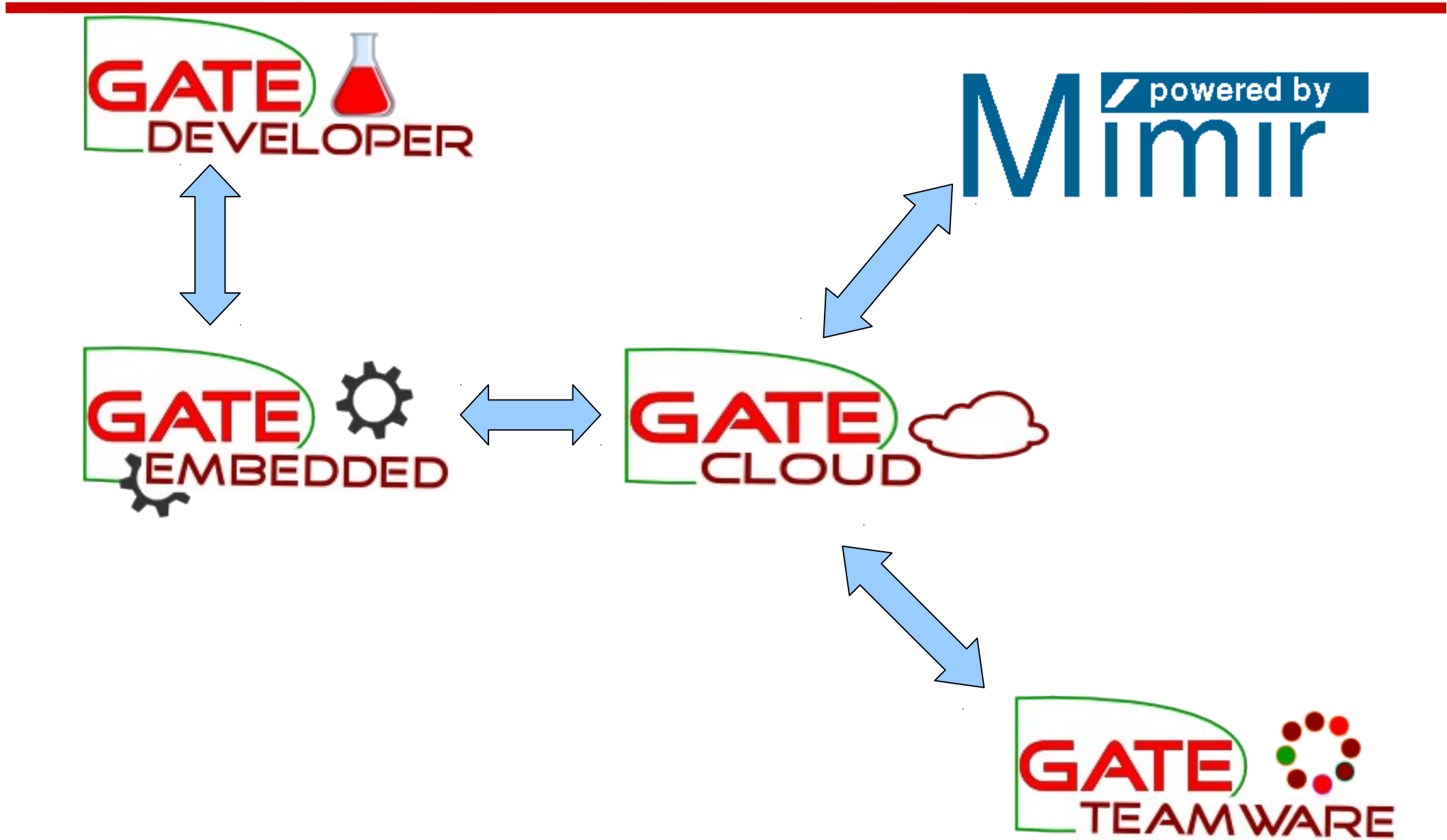
# The problem with free text search


- He burnt the toast and set off the **smoke** alarm
- Mother has **alzheimer's**
- We will do an **MMSE** next week
- **Two weeks ago MMSE** was **19/30**

# GATE: a framework for Human Language Technology


- A framework for language processing
- Open Source – a large community of users and developers
- Mature: over ten years old, currently at version 7.1
- Funded by a mix of EU, UK RC and commercial funding
- The most widely used toolkit of its kind, with 1000s of users at 100s of sites
  - BBC World Cup and Olympics sites; The Press Association; The National Archives; Elsevier; IBM and Oracle integration; various pharma; many other multi-nationals and SMEs
- Biggest single installation supports 10 000 concurrent users
- An architecture: simplifying the construction of NLP software.

# The GATE family



# GATE and medical records


- CaTIES
- HiTEX
- GATE systems often highly ranked in I2B2 challenges
- Commercial use
- University of Sheffield
  - CLEF – a Clinical e-Science Framework
  - German radiology reports
  - Obstetrics system
  - BRC at South London and Maudsley

# Types of IE systems


- Various types of IE system - GATE is agnostic and can act as an architecture for any
- Deep or shallow analysis
- Knowledge Engineering or Machine Learning approaches
  - Supervised
  - Unsupervised
  - Active learning

# Knowledge Engineering vs machine learning


## Knowledge engineering approach

- rule based
- developed by experienced language engineers
- Rules are bespoke
- make use of human intuition
- require only small amount of training data
- development can be very time consuming
- Can give good accuracy

# Knowledge Engineering vs machine learning

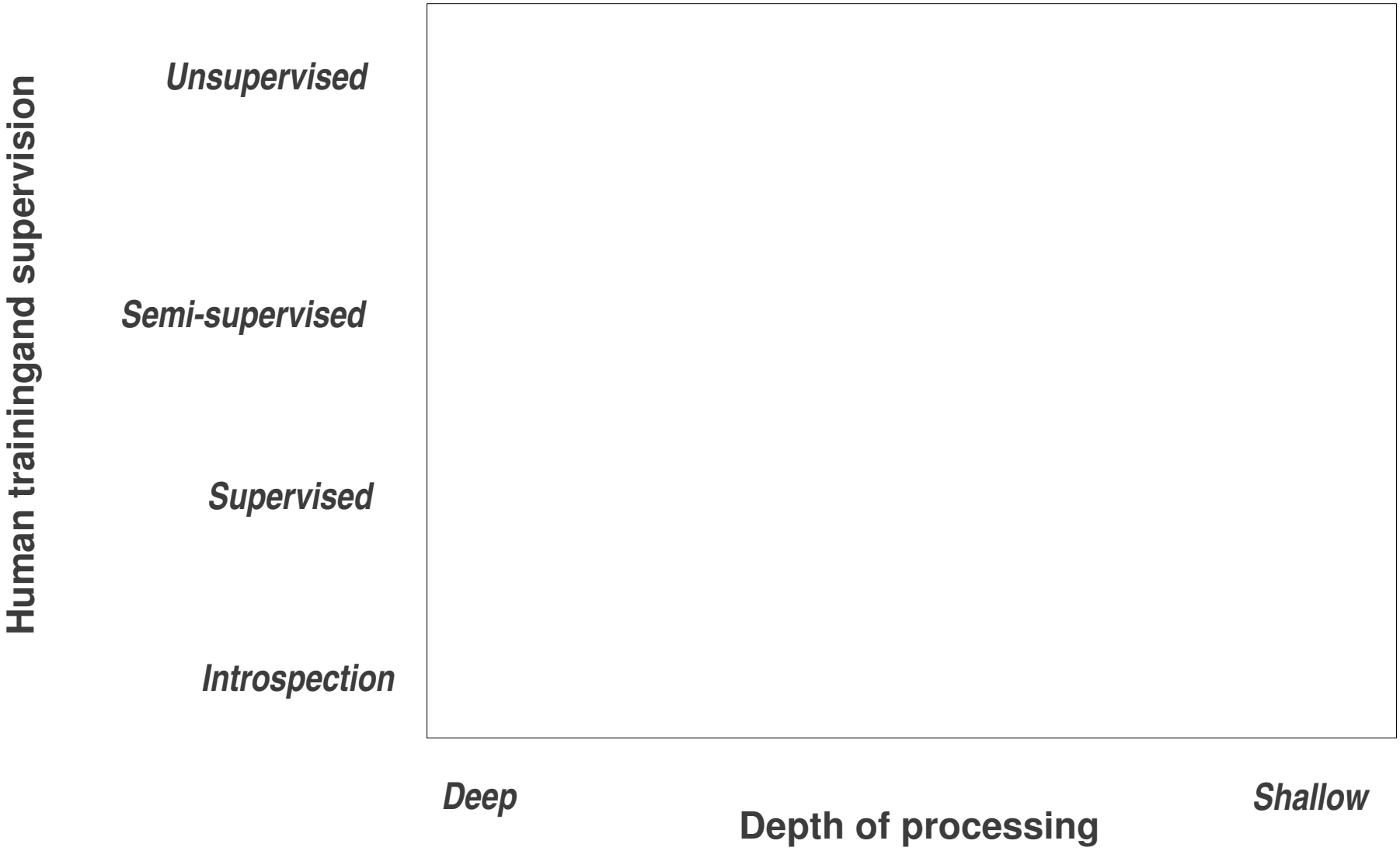

### Knowledge engineering approach

- rule based
- developed by experienced language engineers
- Rules are bespoke
- make use of human intuition
- require only small amount of training data
- development can be very time consuming
- Can give good accuracy

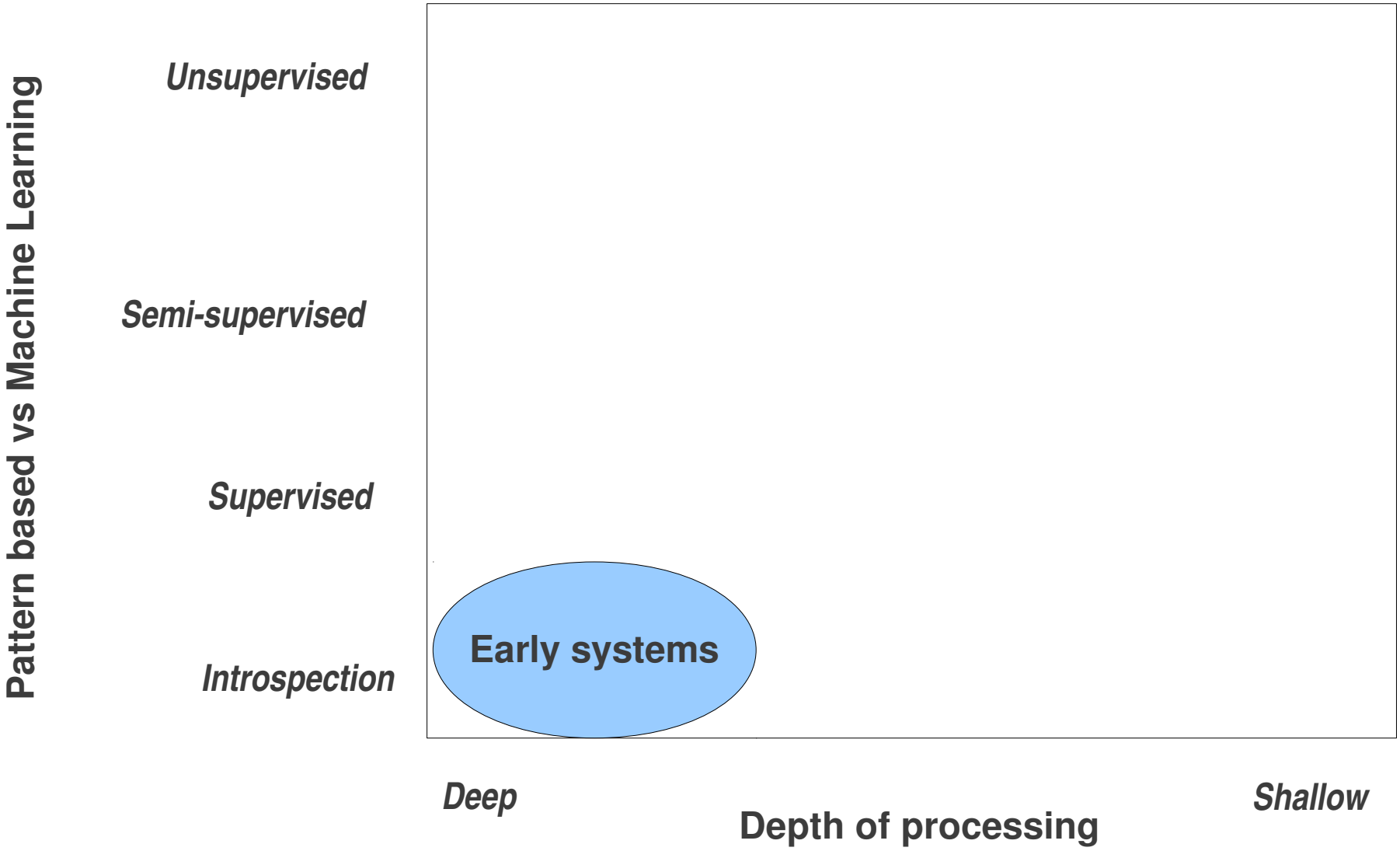
### Supervised Machine Learning approach

- use statistics or other machine learning
- developers do not need LE expertise
- Off the shelf software
- require large amounts of annotated training data
- some changes may require re-annotation of the entire training corpus
- Accuracy depends on complexity of target, amount and quality of training data
- Not transparent

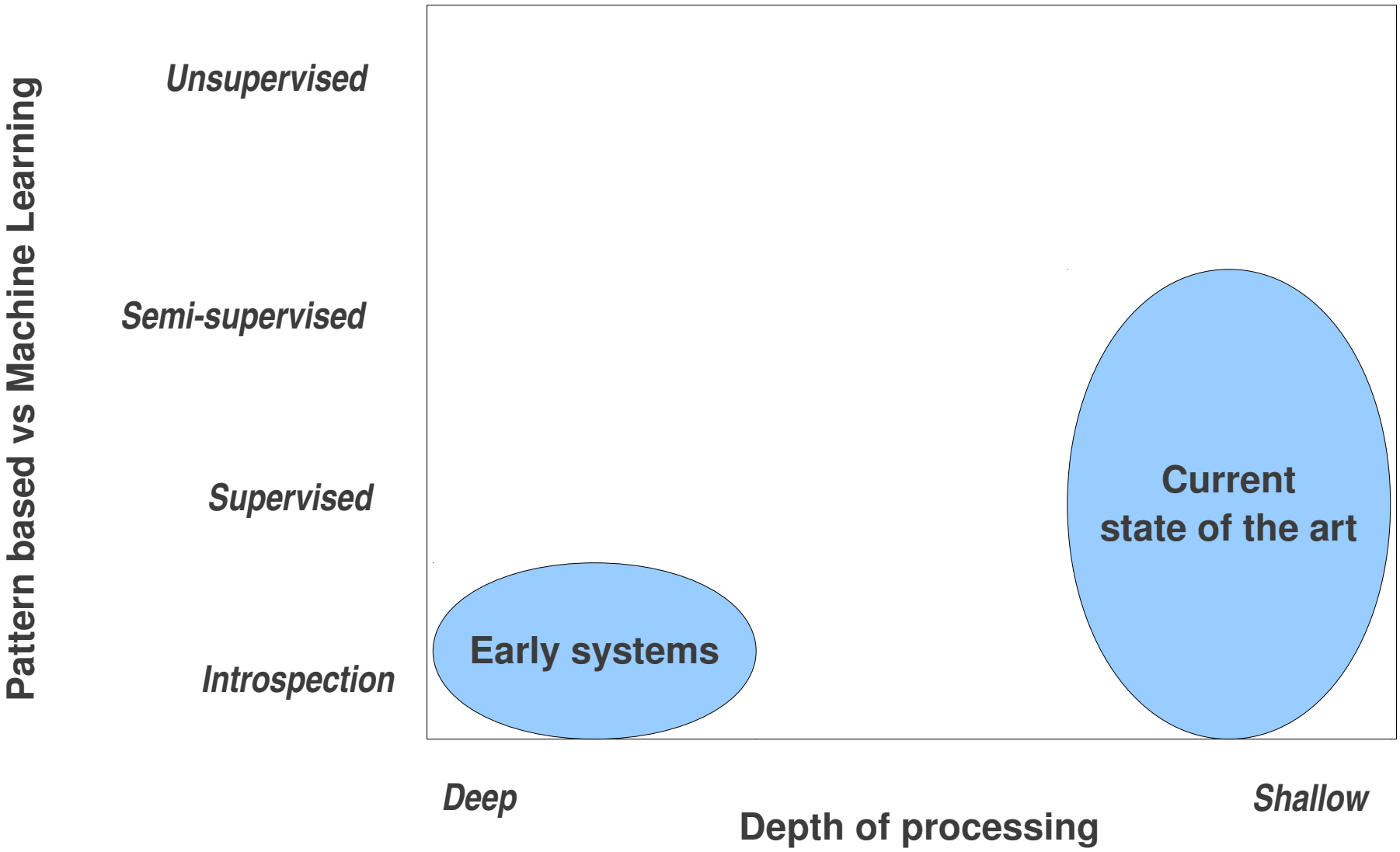
# Clinical text mining landscape



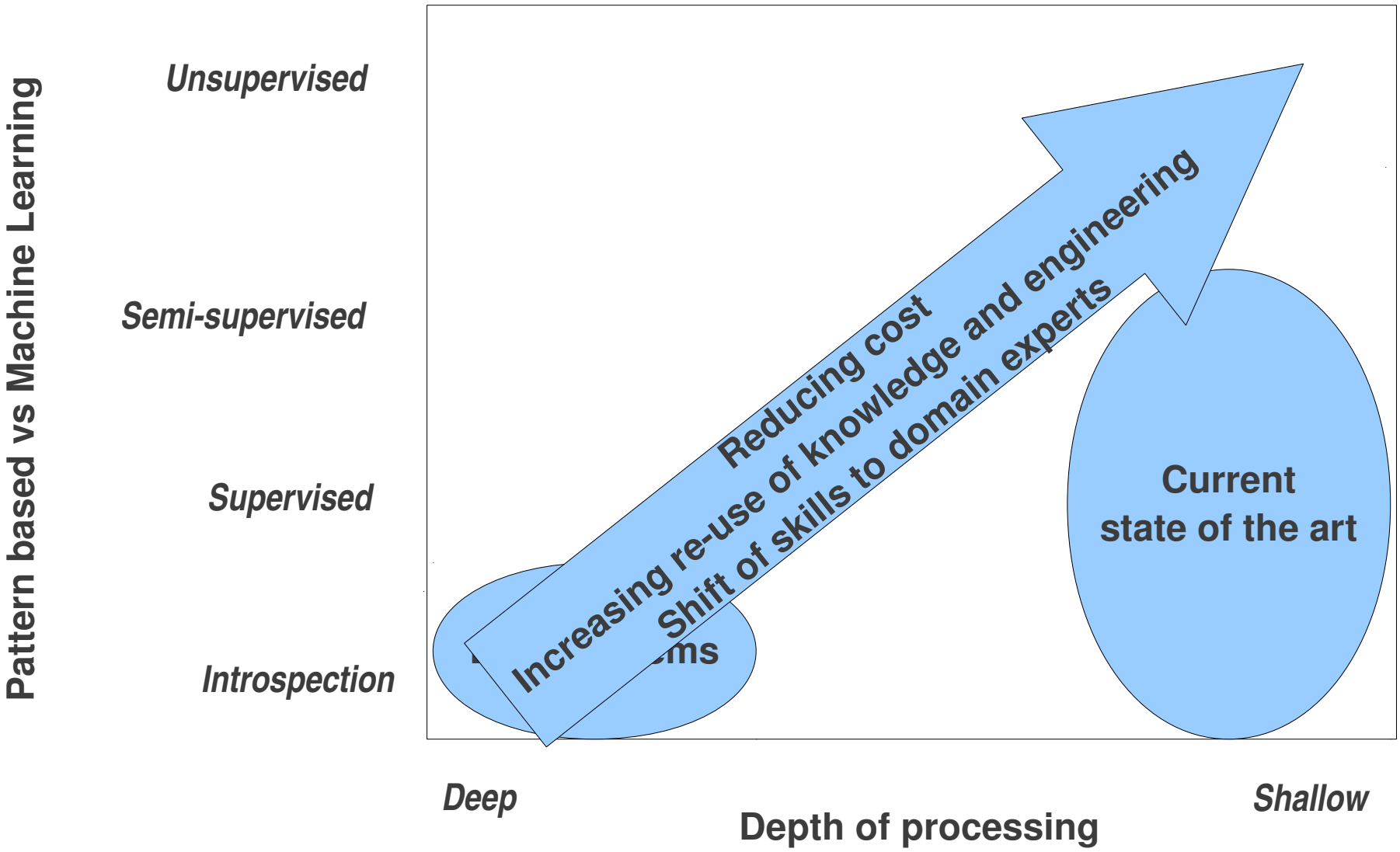
# Clinical text mining landscape



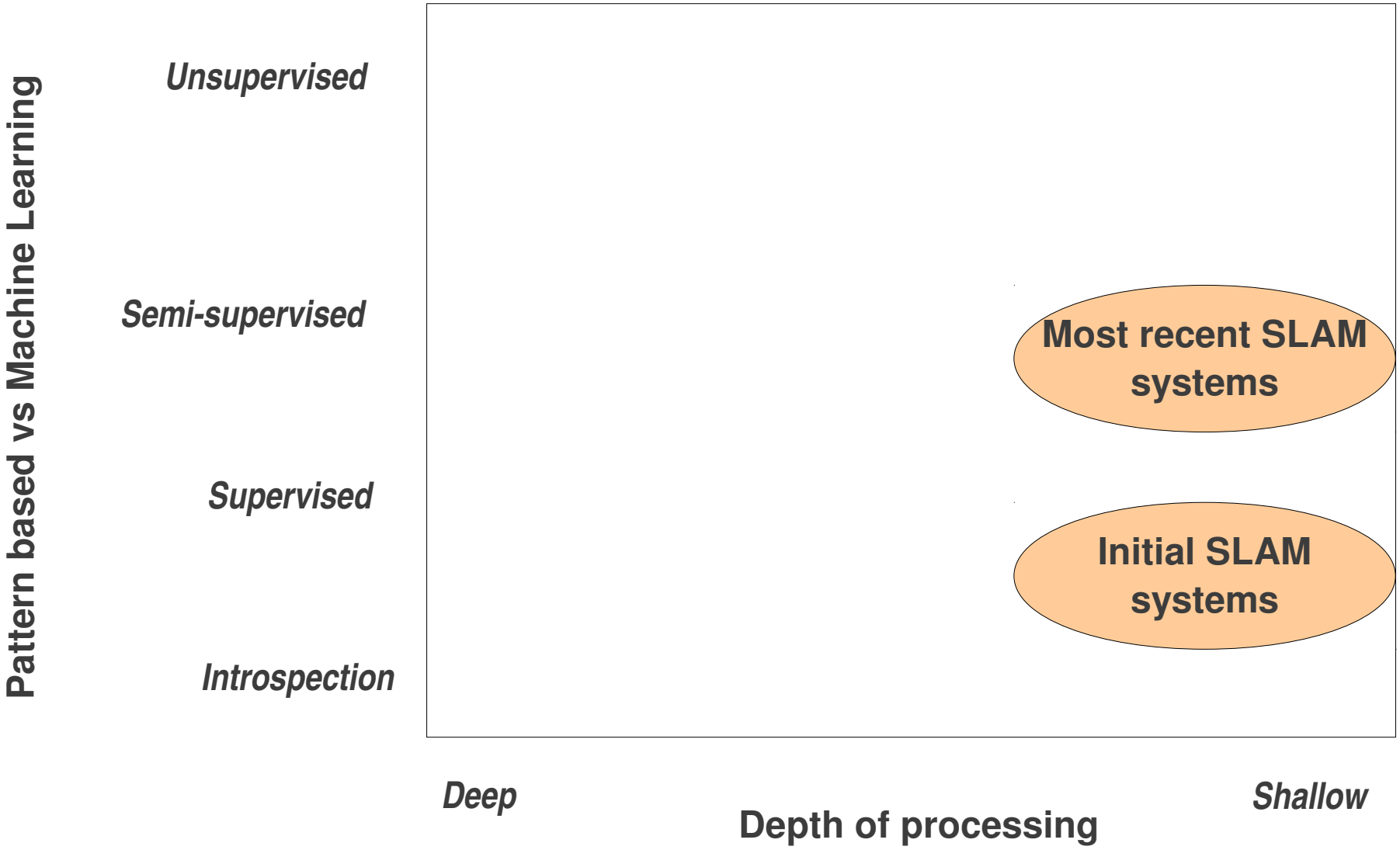
# Clinical text mining landscape

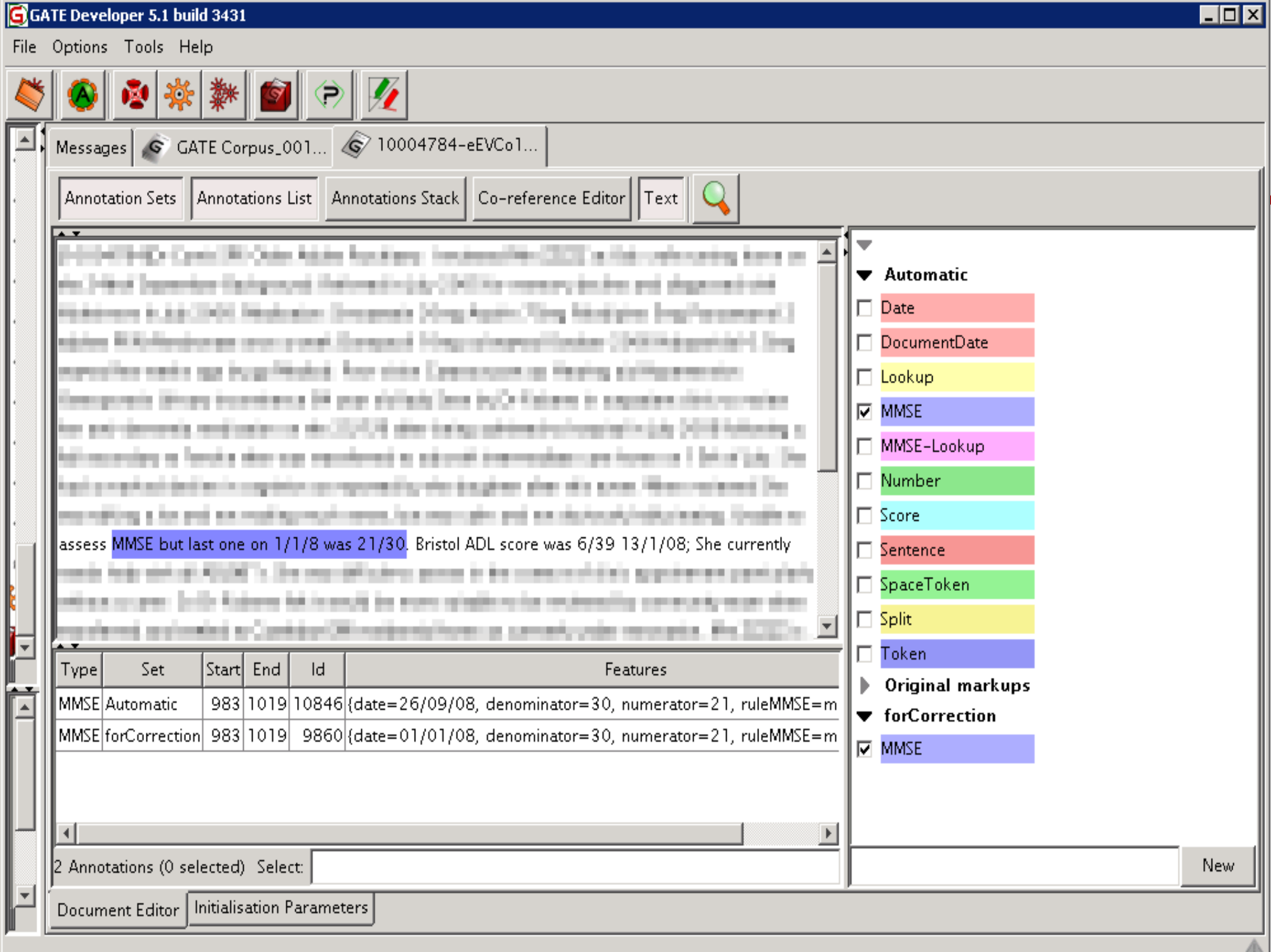


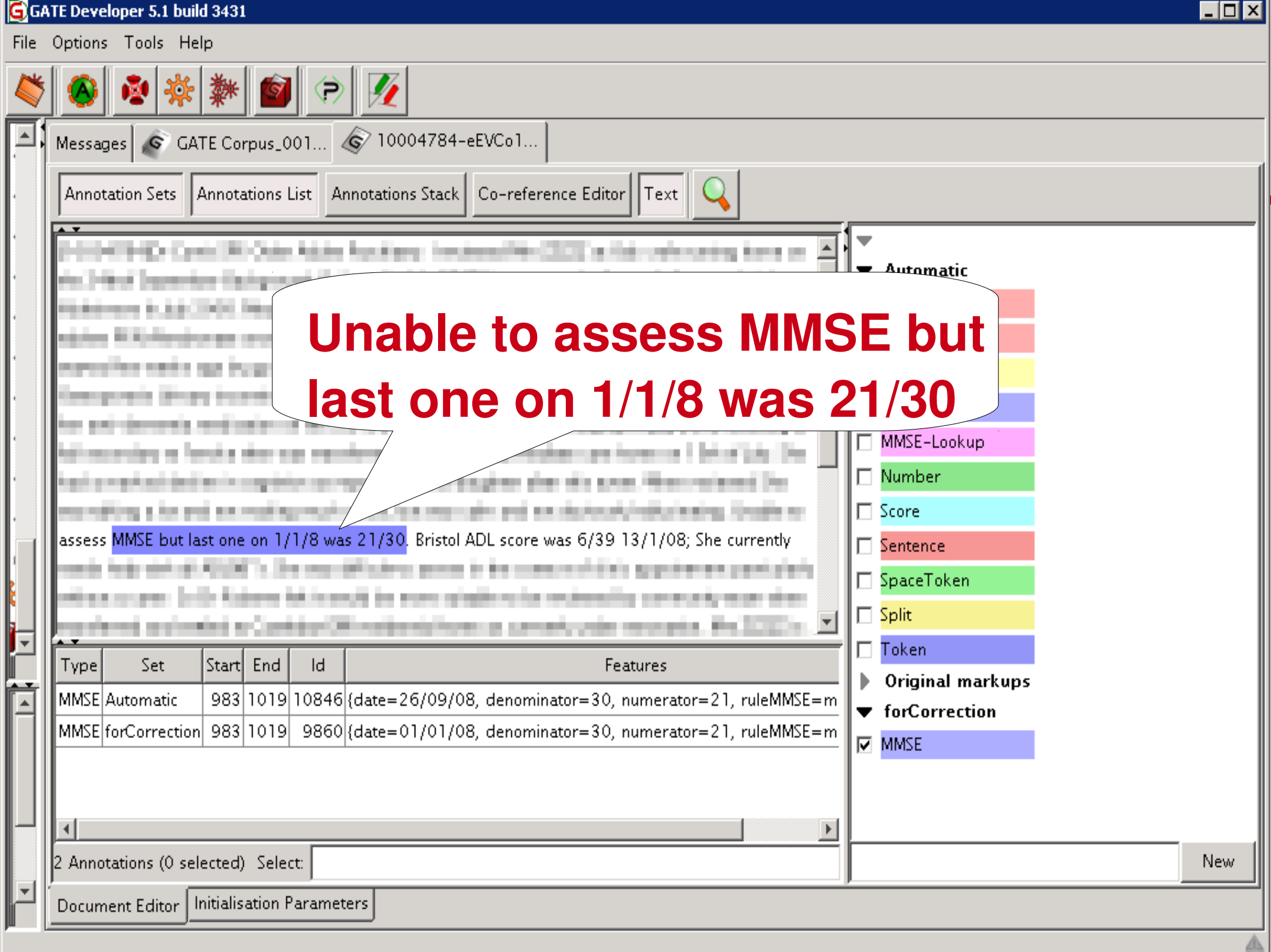
# Clinical text mining landscape

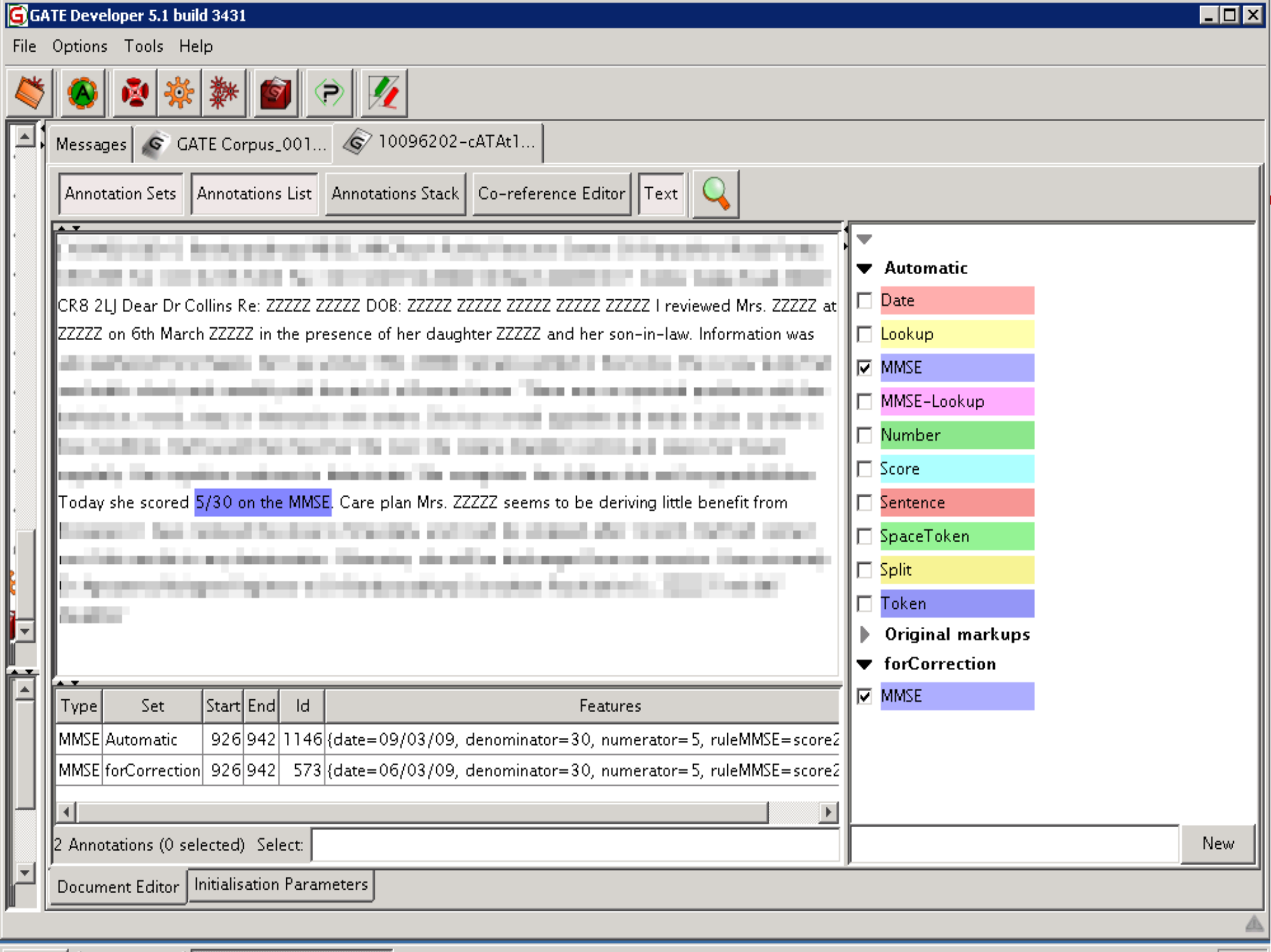


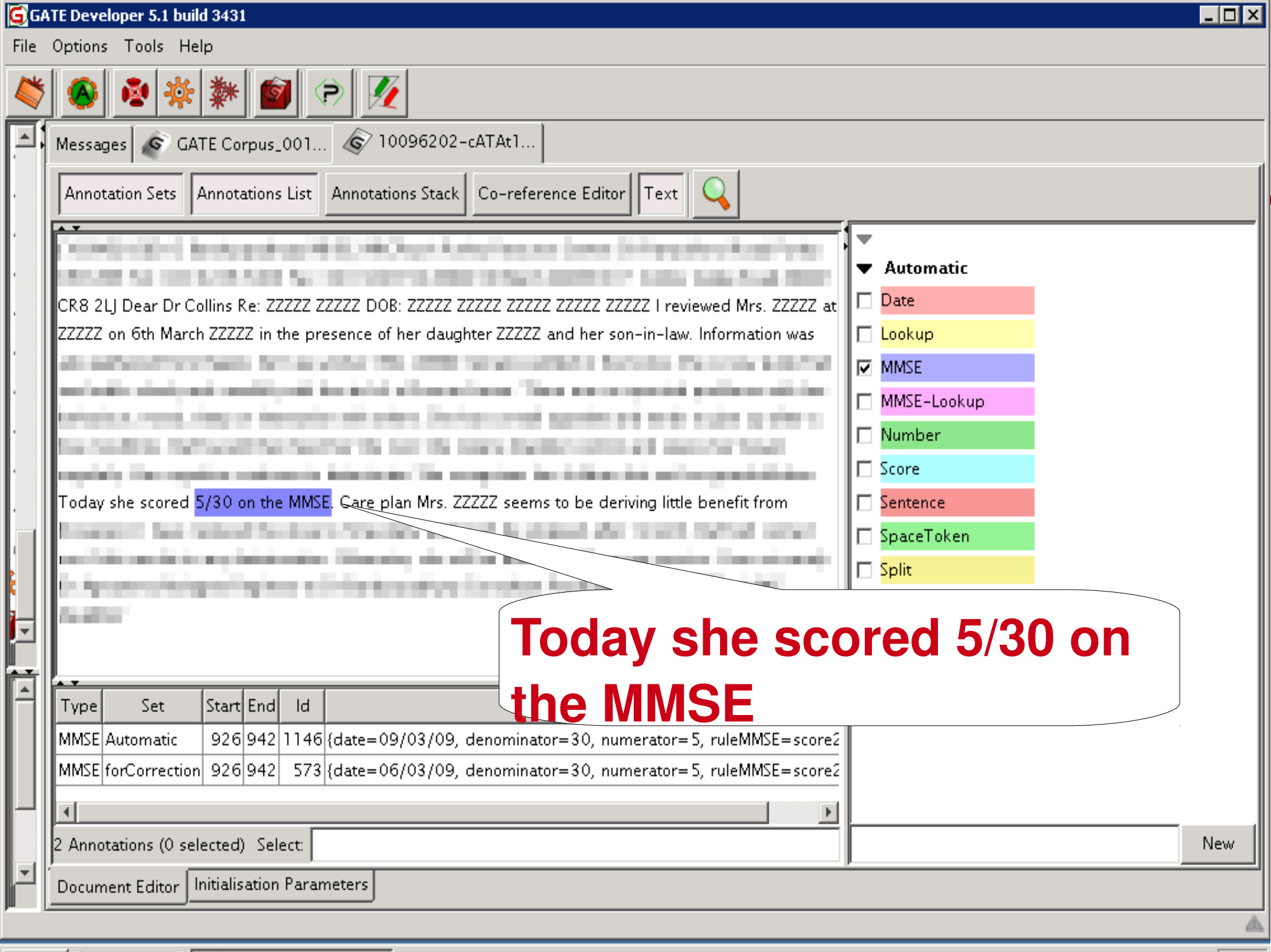
# Clinical text mining landscape

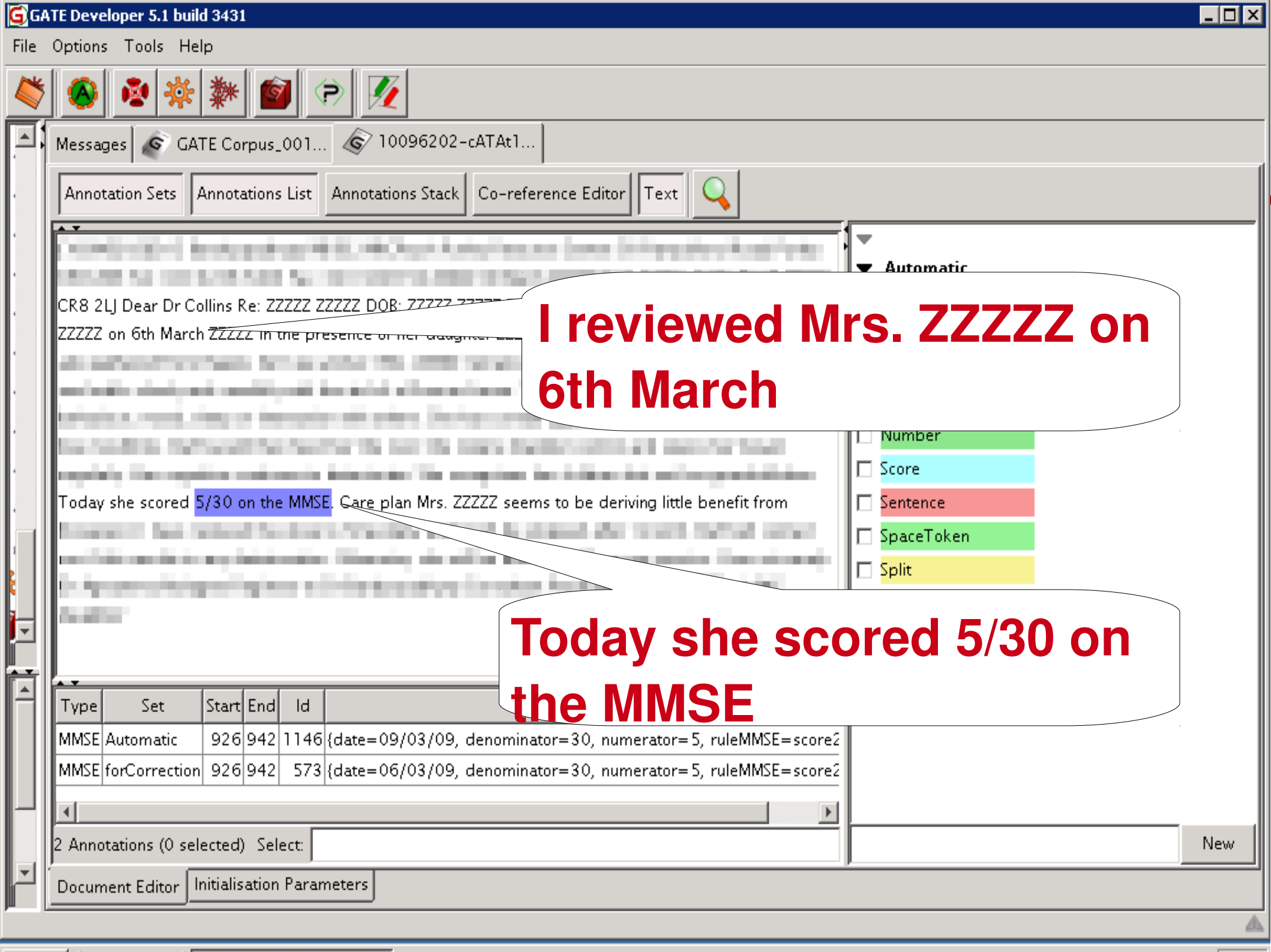



GATE Developer 5.1 build 3431
File Options Tools Help
Messages
GATE Corpus_001...
10004784-eEVCo1...
Annotation Sets
Annotations List
Annotations Stack
Co-reference Editor
Text
assess MMSE but last one on 1/1/8 was 21/30. Bristol ADL score was 6/39 13/1/08; She currently
Type Set Start End Id Features
MMSE Automatic 983 1019 10846 {date=26/09/08, denominator=30, numerator=21, ruleMMSE=m
MMSE forCorrection 983 1019 9860 {date=01/01/08, denominator=30, numerator=21, ruleMMSE=m
2 Annotations (0 selected) Select:
Automatic
Date
DocumentDate
Lookup
MMSE
MMSE-Lookup
Number
Score
Sentence
SpaceToken
Split
Token
Original markups
forCorrection
MMSE
New
Document Editor
Initialisation Parameters


GATE Developer 5.1 build 3431
File Options Tools Help
Messages
GATE Corpus_001...
10004784-eEVCo1...
Annotation Sets
Annotations List
Annotations Stack
Co-reference Editor
Text
Unable to assess MMSE but last one on 1/1/8 was 21/30
assess MMSE but last one on 1/1/8 was 21/30. Bristol ADL score was 6/39 13/1/08; She currently
Automatic
MMSE-Lookup
Number
Score
Sentence
SpaceToken
Split
Token
Original markups
forCorrection
MMSE
Type Set Start End Id Features
MMSE Automatic 983 1019 10846 {date=26/09/08, denominator=30, numerator=21, ruleMMSE=m
MMSE forCorrection 983 1019 9860 {date=01/01/08, denominator=30, numerator=21, ruleMMSE=m
2 Annotations (0 selected) Select:
New
Document Editor
Initialisation Parameters


GATE Developer 5.1 build 3431
File Options Tools Help
Messages
GATE Corpus_001...
10096202-cATAt1...
Annotation Sets
Annotations List
Annotations Stack
Co-reference Editor
Text
CR8 2LJ Dear Dr Collins Re: ZZZZZ ZZZZZ DOB: ZZZZZ ZZZZZ ZZZZZ ZZZZZ ZZZZZ I reviewed Mrs. ZZZZZ at ZZZZZ on 6th March ZZZZZ in the presence of her daughter ZZZZZ and her son-in-law. Information was
Today she scored 5/30 on the MMSE. Care plan Mrs. ZZZZZ seems to be deriving little benefit from
Automatic
Date
Lookup
MMSE
MMSE-Lookup
Number
Score
Sentence
SpaceToken
Split
Token
Original markups
forCorrection
MMSE
Type Set Start End Id Features
MMSE Automatic 926 942 1146 {date=09/03/09, denominator=30, numerator=5, ruleMMSE=score2
MMSE forCorrection 926 942 573 {date=06/03/09, denominator=30, numerator=5, ruleMMSE=score2
2 Annotations (0 selected) Select:
New
Document Editor
Initialisation Parameters


GATE Developer 5.1 build 3431
File Options Tools Help
Messages
GATE Corpus_001...
10096202-cATAt1...
Annotation Sets
Annotations List
Annotations Stack
Co-reference Editor
Text
CR8 2LJ Dear Dr Collins Re: ZZZZZ ZZZZZ DOB: ZZZZZ ZZZZZ ZZZZZ ZZZZZ ZZZZZ I reviewed Mrs. ZZZZZ at
ZZZZZ on 6th March ZZZZZ in the presence of her daughter ZZZZZ and her son-in-law. Information was
Today she scored 5/30 on the MMSE. Care plan Mrs. ZZZZZ seems to be deriving little benefit from
Automatic
Date
Lookup
MMSE
MMSE-Lookup
Number
Score
Sentence
SpaceToken
Split
Type Set Start End Id
MMSE Automatic 926 942 1146 {date=09/03/09, denominator=30, numerator=5, ruleMMSE=score2
MMSE forCorrection 926 942 573 {date=06/03/09, denominator=30, numerator=5, ruleMMSE=score2
2 Annotations (0 selected) Select:
New
Document Editor
Initialisation Parameters
Today she scored 5/30 on the MMSE


GATE Developer 5.1 build 3431
File Options Tools Help
Messages
GATE Corpus_001...
10096202-cATAt1...
Annotation Sets
Annotations List
Annotations Stack
Co-reference Editor
Text
CR8 2LJ Dear Dr Collins Re: ZZZZZ ZZZZZ DOB:
ZZZZZ on 6th March
Today she scored 5/30 on the MMSE. Care plan Mrs. ZZZZZ seems to be deriving little benefit from
Automatic
Number
Score
Sentence
SpaceToken
Split
I reviewed Mrs. ZZZZZ on 6th March
Today she scored 5/30 on the MMSE
Type Set Start End Id
MMSE Automatic 926 942 1146 {date=09/03/09, denominator=30, numerator=5, ruleMMSE=score2
MMSE forCorrection 926 942 573 {date=06/03/09, denominator=30, numerator=5, ruleMMSE=score2
2 Annotations (0 selected) Select:
New
Document Editor
Initialisation Parameters


# A shallow approach

- Pre-processing, including
  - morphological analysis
    - *"Patient was seen on" vs "I saw this patient on"*
  - POS tagging
    - *"patient was [VERB] on [DATE]"*
- Dictionary lookup
  - *"MMSE", "Mini mental", "Folstein", "AMTS"*
- Coreference
  - *"We did an MMSE. It was 23/30"*

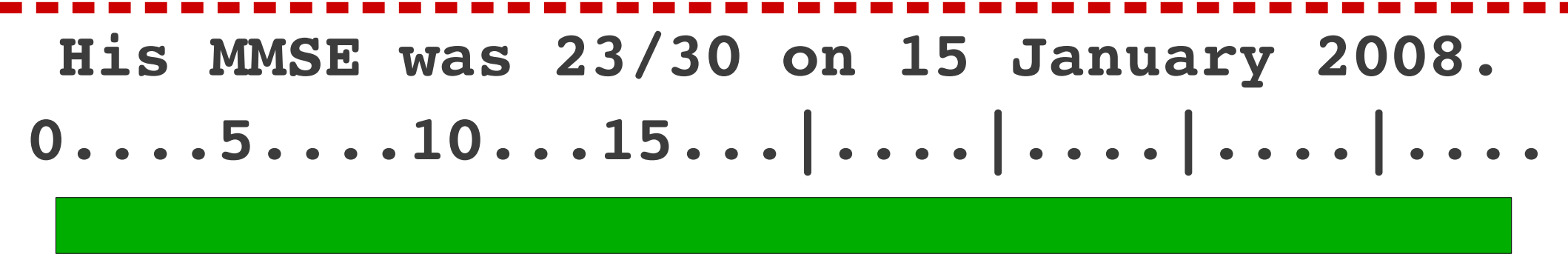
# Annotations


His MMSE was 23/30 on 15 January 2008.

# Annotations

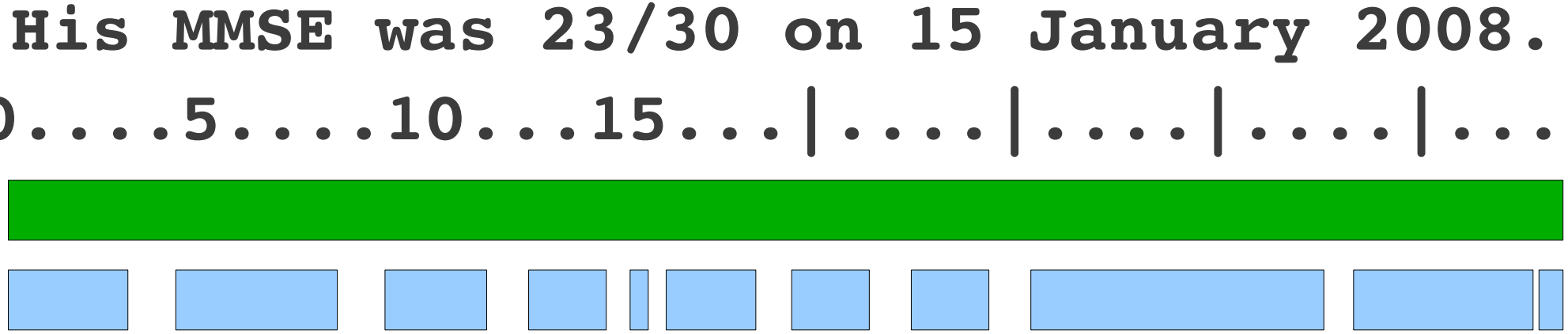
# Annotations

# Annotations



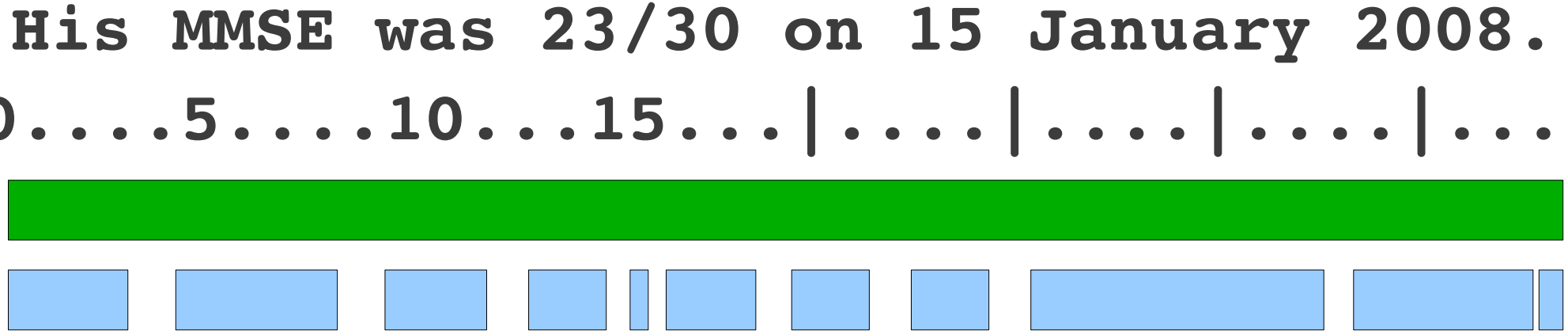

| Id | Type |
|---|---|
| 1 | sentence |

# Annotations



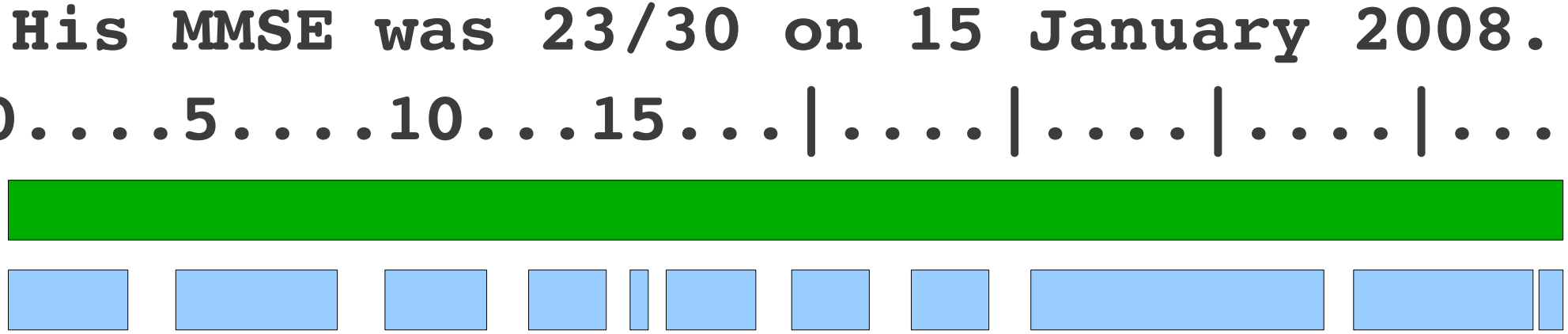

| Id | Type |
|---|---|
| 1 | sentence |

# Annotations



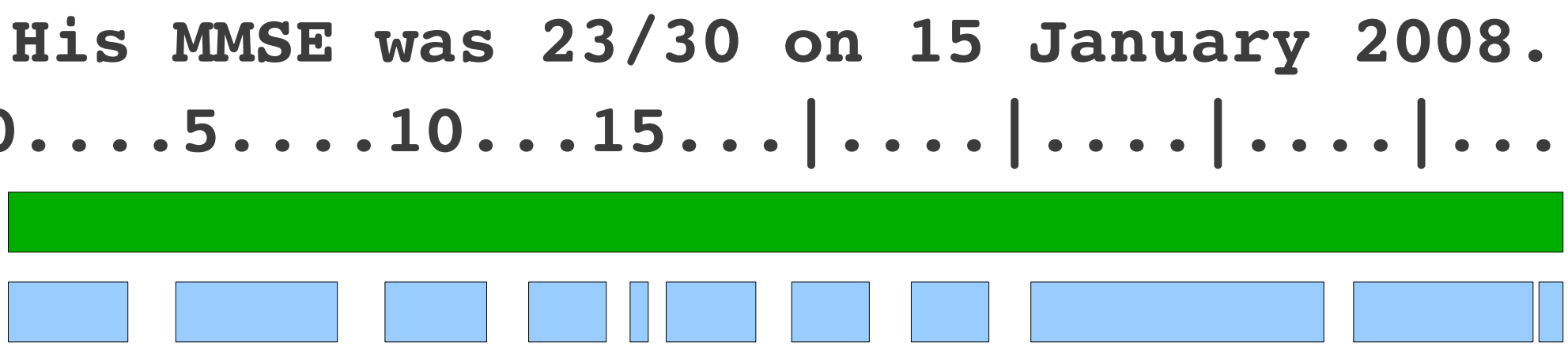

| Id | Type |
|---|---|
| 1 | sentence |
| 2 | token |
| 3 | token |
| 4 | token |
| 5 | token |
| 6 | token |
| 7 | token |

# Annotations



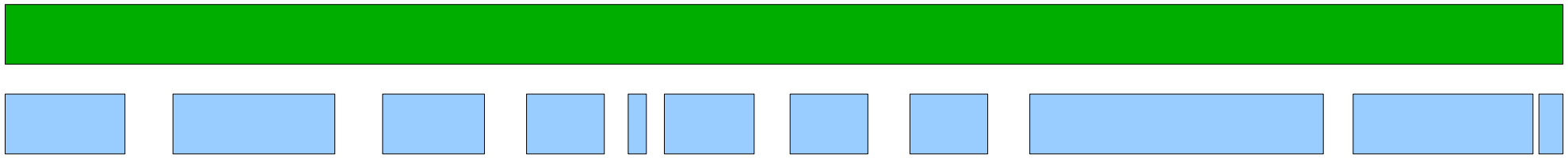

| Id | Type | Start | End |
|---|---|---|---|
| 1 | sentence | 0 | 39 |
| 2 | token | 0 | 3 |
| 3 | token | 4 | 8 |
| 4 | token | 9 | 12 |
| 5 | token | 13 | 15 |
| 6 | token | 15 | 16 |
| 7 | token | 16 | 18 |

# Annotations



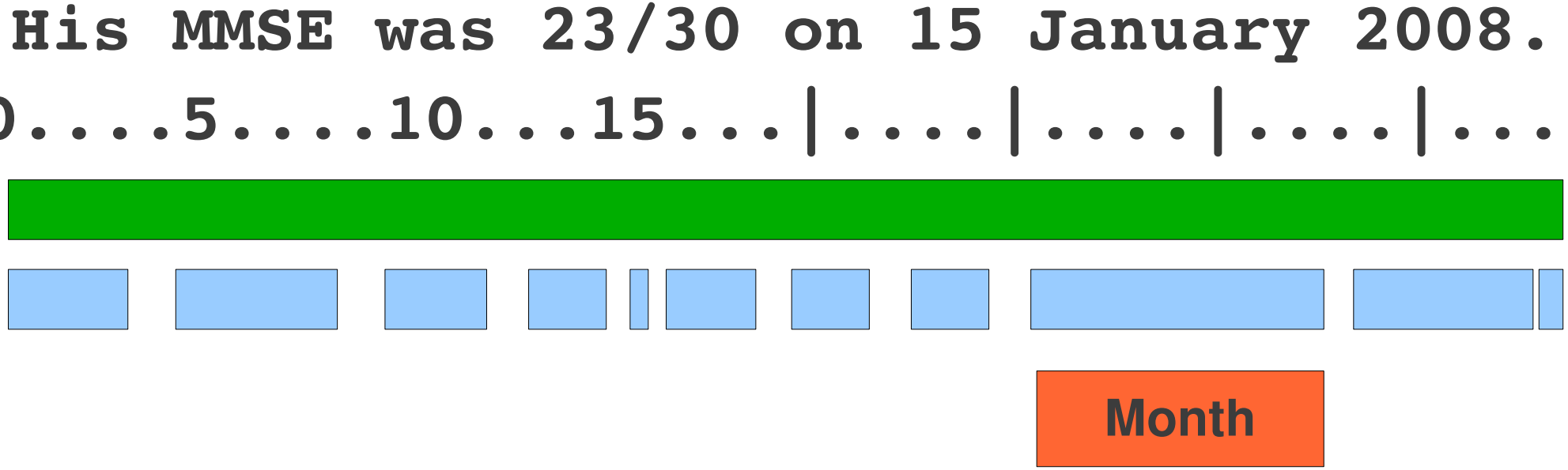

| Id | Type | Start | End | Features |
|---|---|---|---|---|
| 1 | sentence | 0 | 39 | |
| 2 | token | 0 | 3 | pos=PP |
| 3 | token | 4 | 8 | pos=NN |
| 4 | token | 9 | 12 | pos=VB |
| 5 | token | 13 | 15 | pos=CD |
| 6 | token | 15 | 16 | pos=SM |
| 7 | token | 16 | 18 | pos=CD |

# Annotations

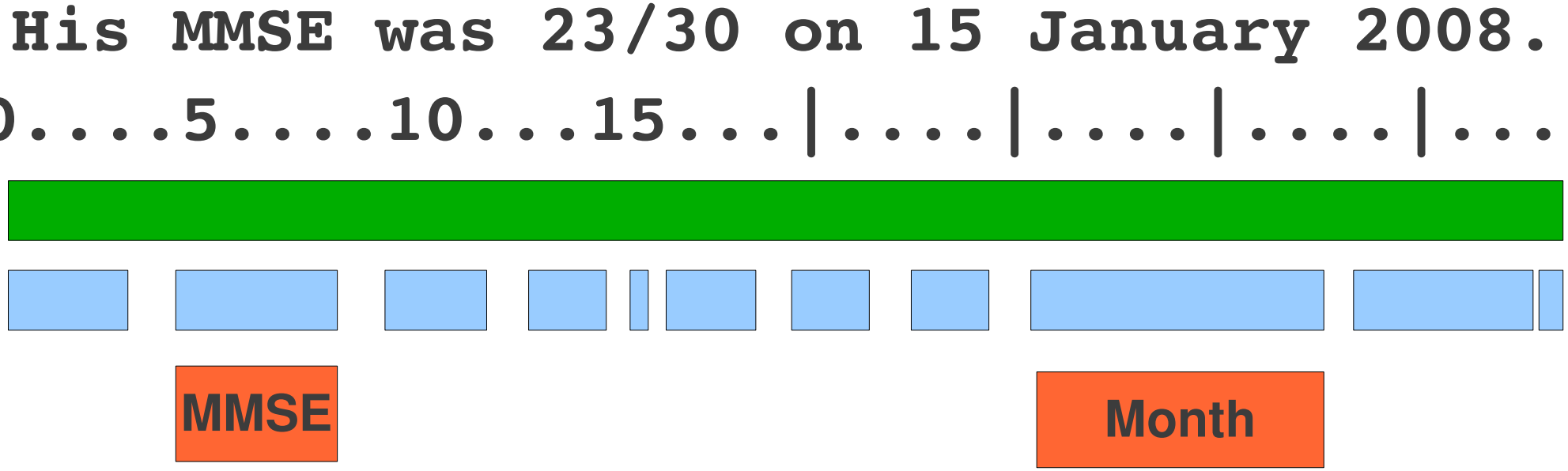



| Id | Type | Start | End | Features | |
|---|---|---|---|---|---|
| 1 | sentence | 0 | 39 | | |
| 2 | token | 0 | 3 | pos=PP | |
| 3 | token | 4 | 8 | pos=NN | |
| 4 | token | 9 | 12 | pos=VB | root=be |
| 5 | token | 13 | 15 | pos=CD | type=num |
| 6 | token | 15 | 16 | pos=SM | type=slash |
| 7 | token | 16 | 18 | pos=CD | type=num |

# Dictionary lookup



# Dictionary lookup



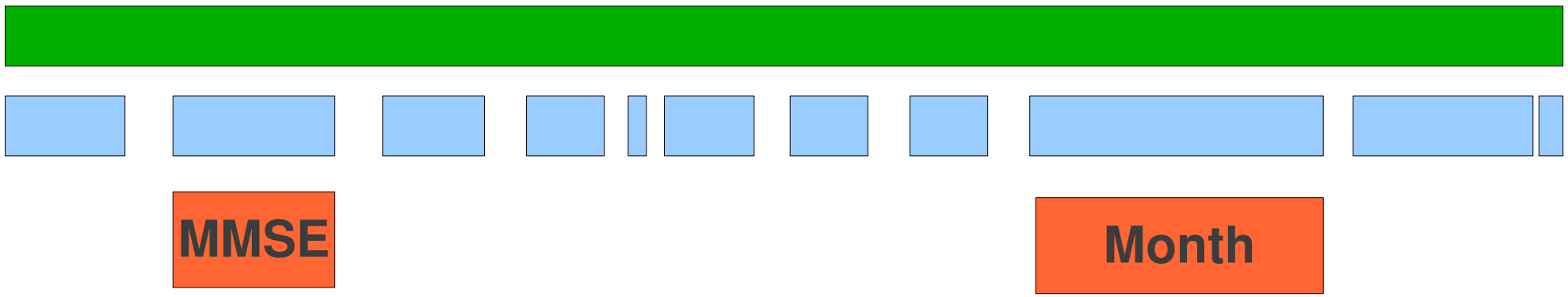

# Limitations of dictionary lookup

- Dictionary lookup is designed for finding simple, regular terms and features
- False positives
  - *“He may get better”*
  - *“Mother is a smoker”*
  - *“He often burns the toast, setting off the smoke alarm”*
- Cannot deal with complex patterns
  - For example, recognising e-mail addresses using just a dictionary would be impossible
- Cannot deal with ambiguity
  - I for Iodine, or I for me?


# Pattern matching

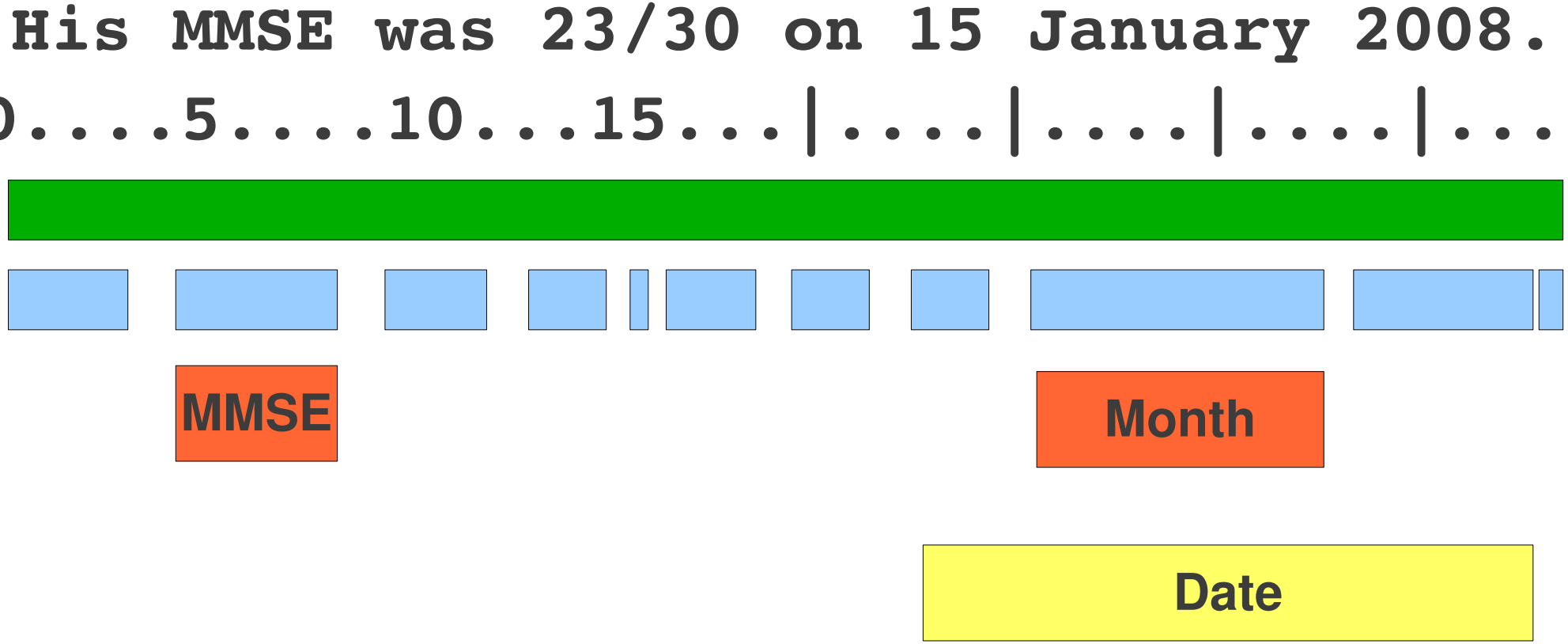
- The early components in a GATE pipeline produce simple annotations
  - Token, Sentence, Dictionary lookups
- These annotations have features
  - Token kind, part of speech, major type...
- Patterns in these annotations and features can suggest more complex information
- A pattern matching language, JAPE, is used to find these patterns

# Patterns

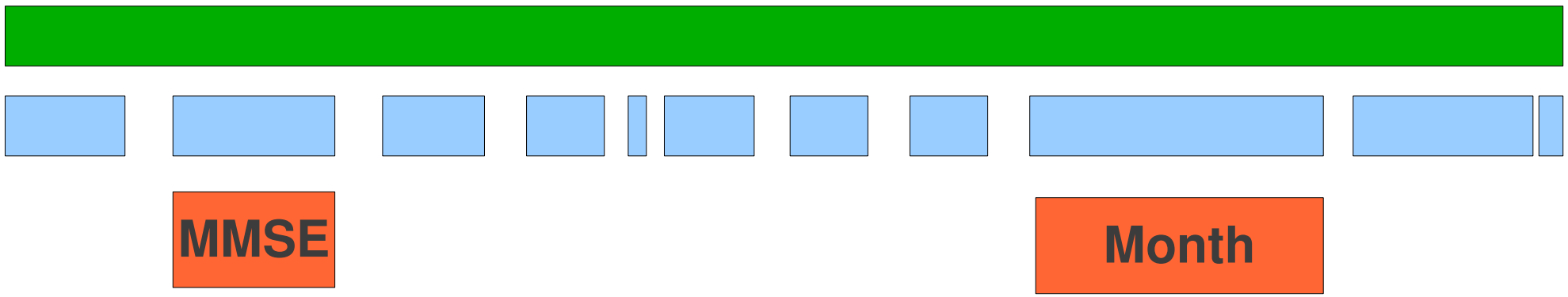



```
{number}{Month}{number}
```

# Patterns

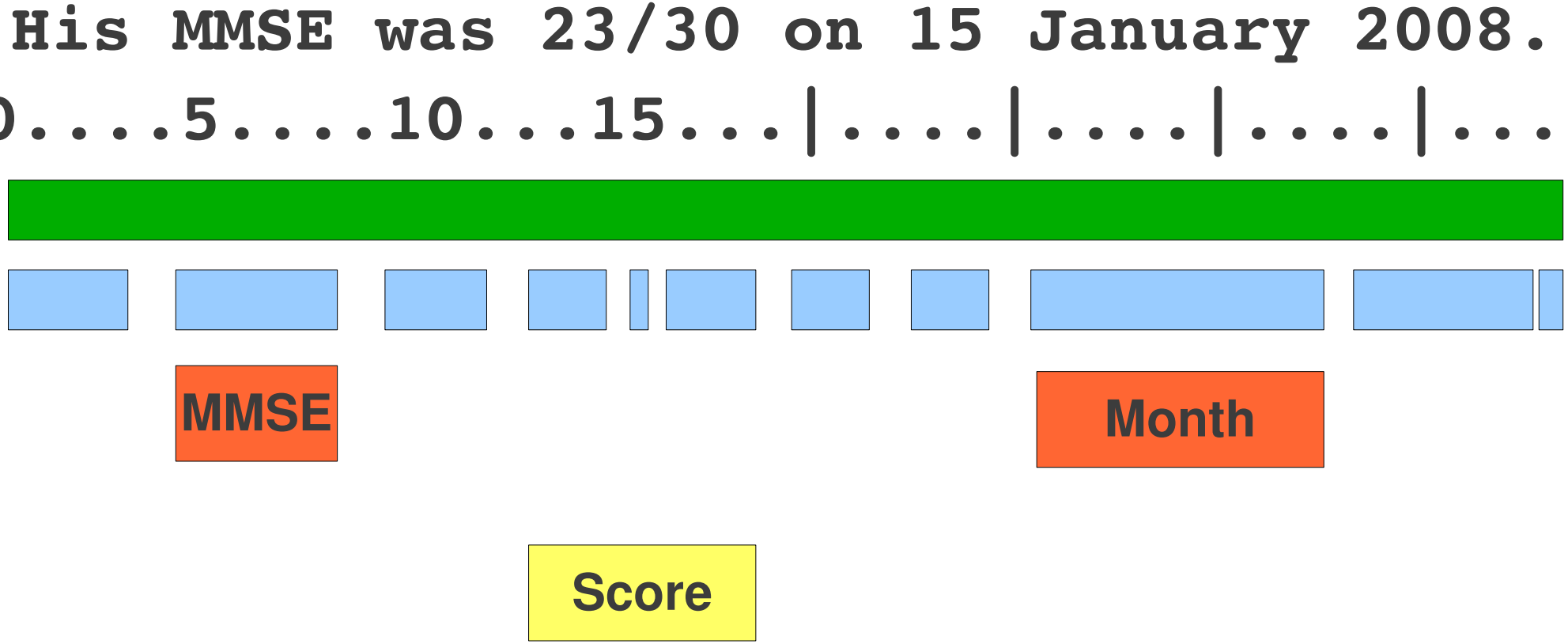


# Patterns

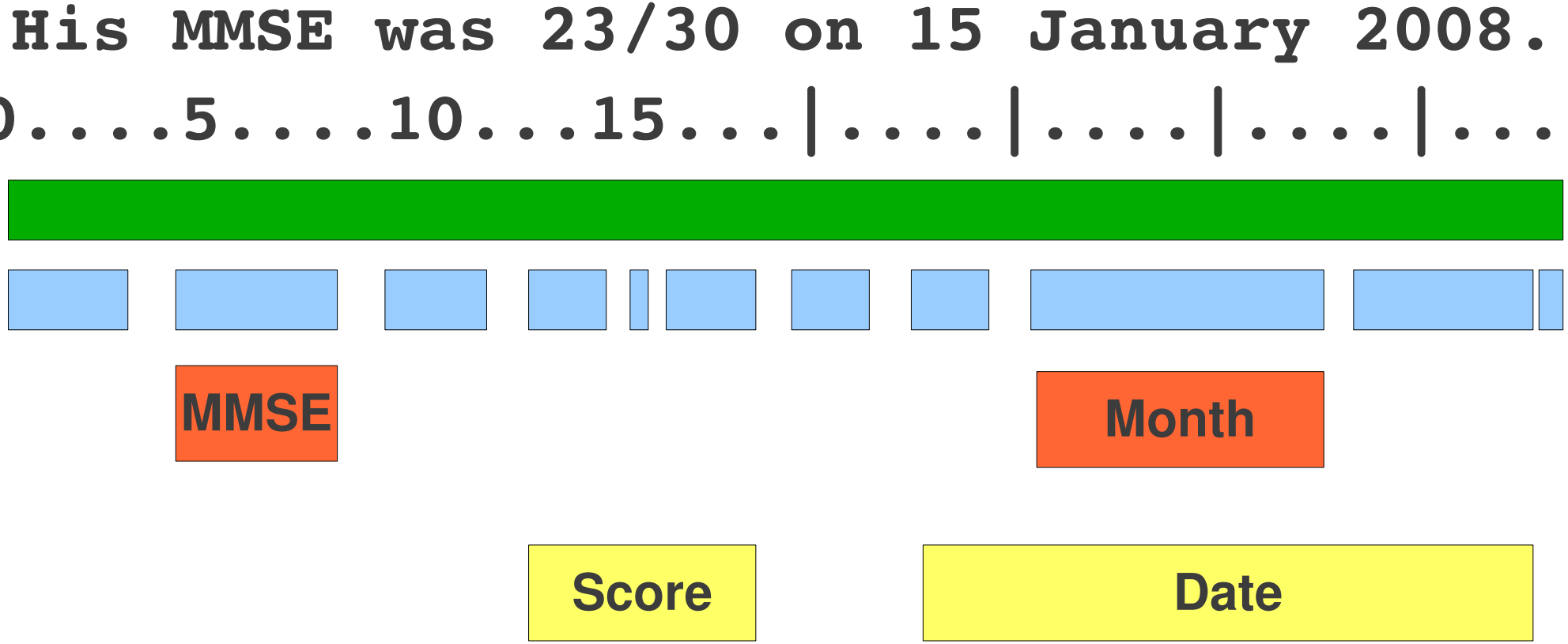



```
{number}{slash}{number}
```

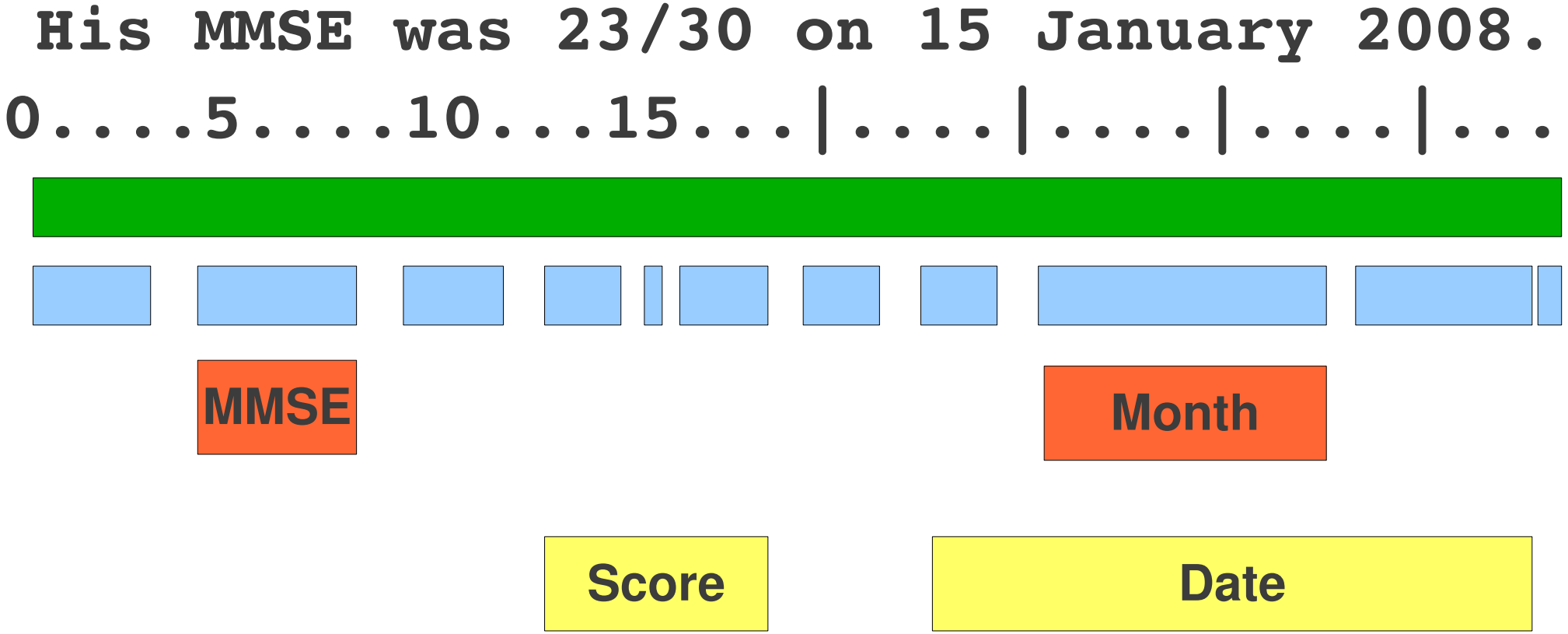
# Patterns



# Patterns

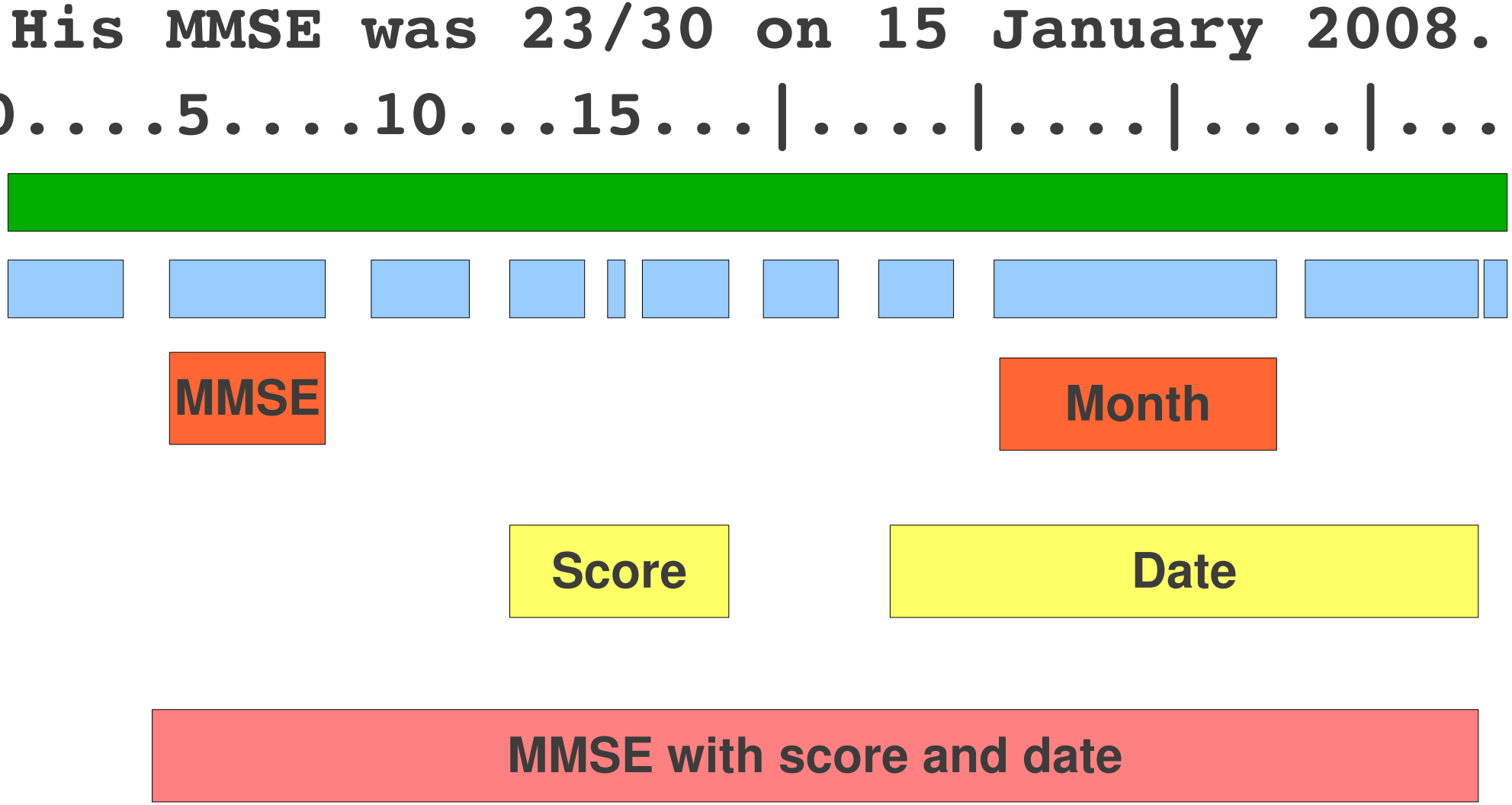

His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|....|....
MMSE
Month
Score
Date

# Patterns




```
{MMSE}{BE}{Score}{?}{Date}
```

# Patterns


His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|....|....
MMSE
Month
Score
Date
MMSE with score and date

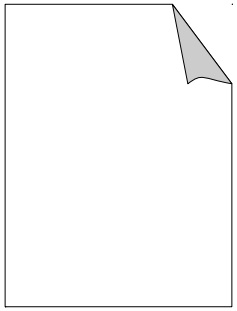
# Patterns are general


- MMSE was 23/30 on 15 January 2009
- Mini mental was 25/30 on 12/08/07
- MMS was 25/30 last week
- MMSE is 25/30 today
- With adaptation
  - MMSE 25 out of 30
  - Long range dependencies on dates

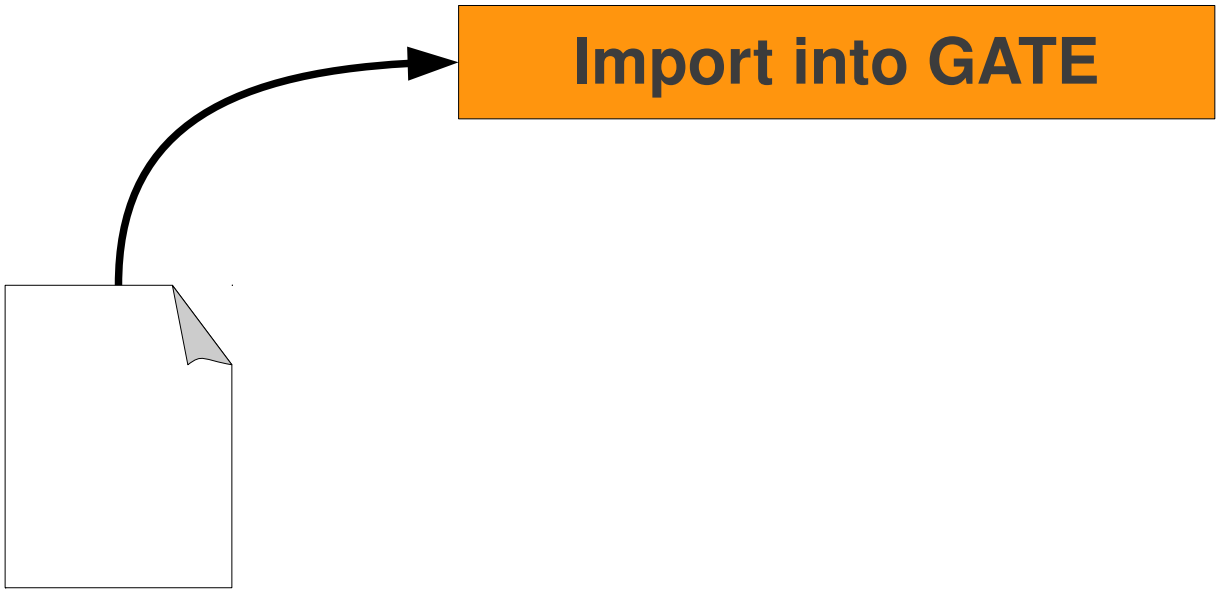

# A JAPE example

```
Phase: EMail
Input: Token SpaceToken
Options: control = appelt

Macro: WORD_OR_NUMBER
(
    ({Token.kind == word}|{Token.kind == number})
)

Rule: emailaddress
Priority: 50
(
   (WORD_OR_NUMBER)+
   ({Token.string == "."}(WORD_OR_NUMBER)+)*
   {Token.string == "@"}
   (WORD_OR_NUMBER)+
   ({Token.string == "."}(WORD_OR_NUMBER)+)*
)
:email -->
   :email.EMail= {rule = "emailaddress"}
```

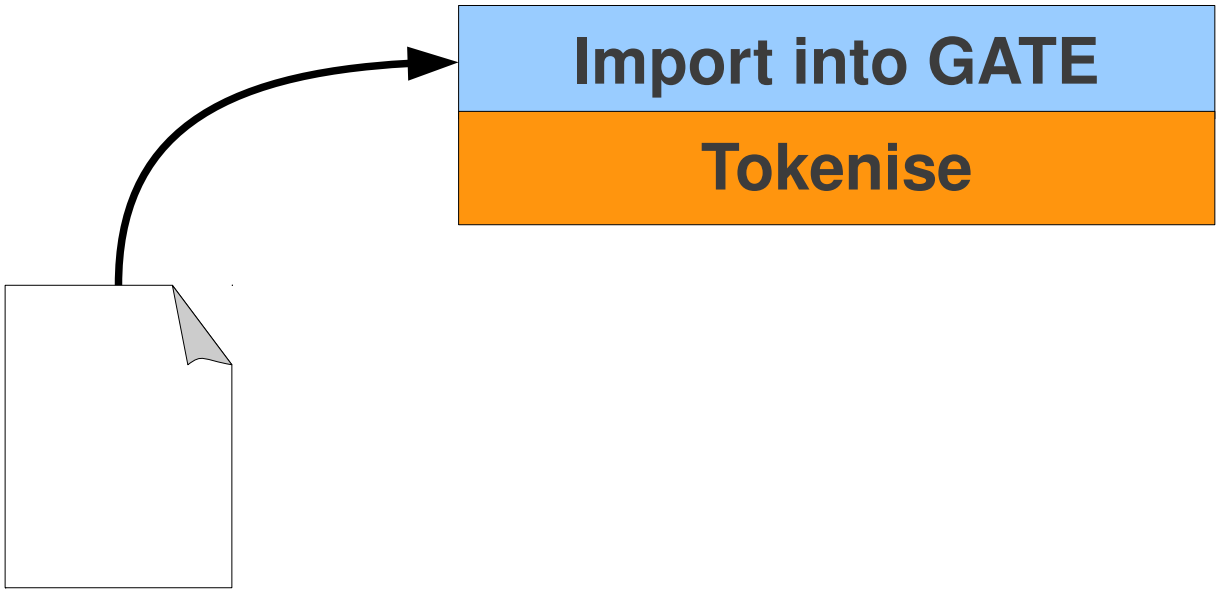

# MMSE pipeline



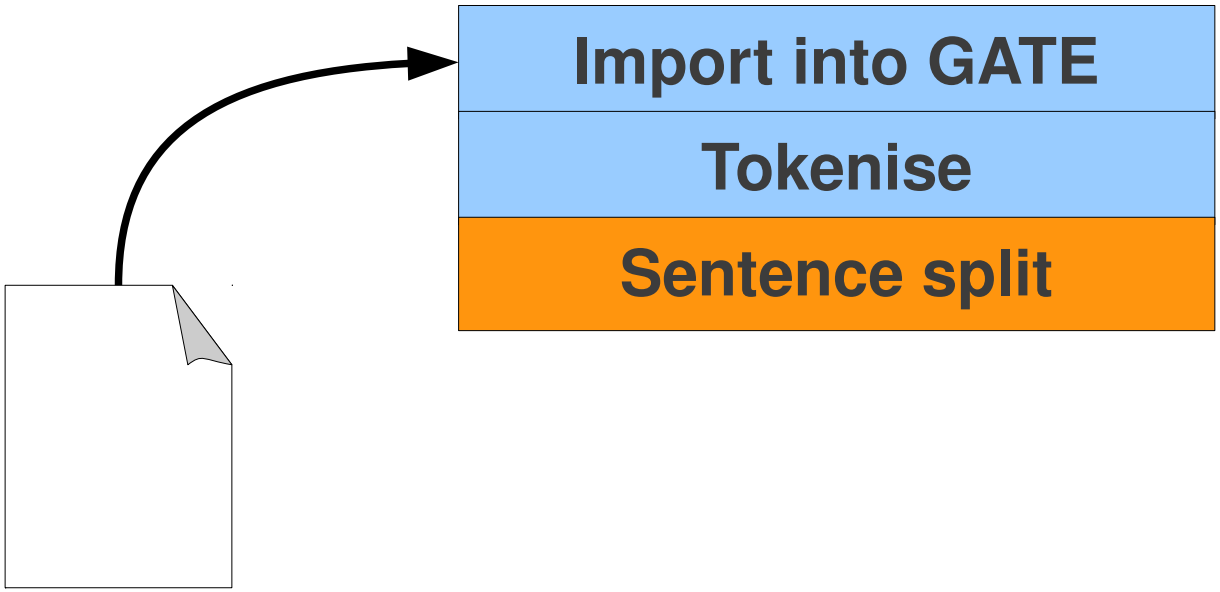
# MMSE pipeline



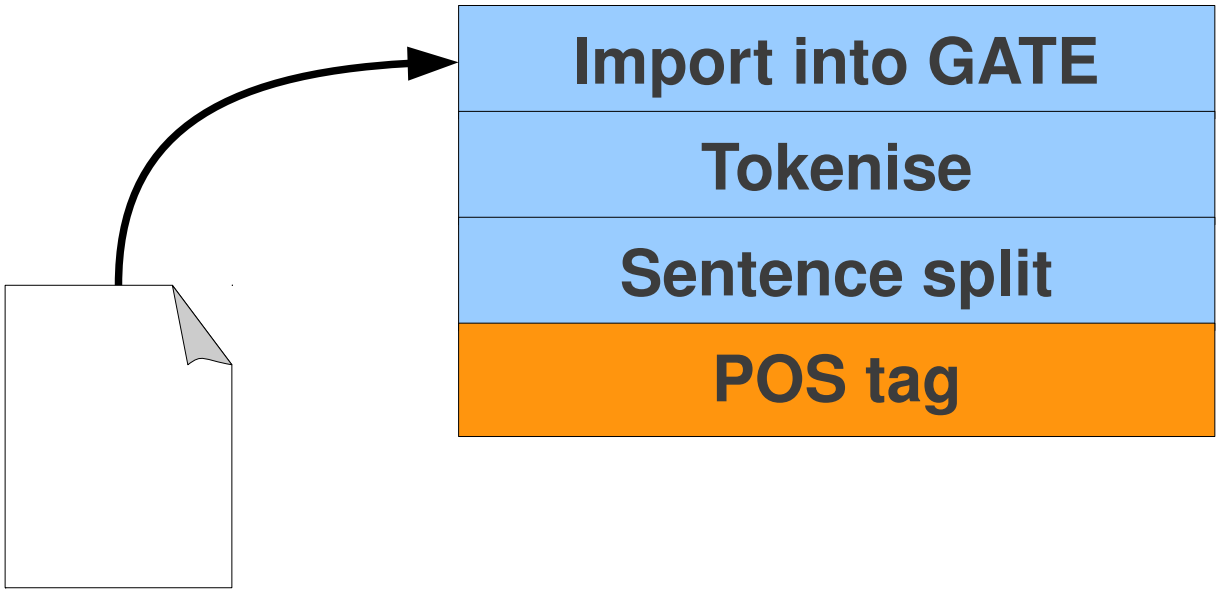
# MMSE pipeline



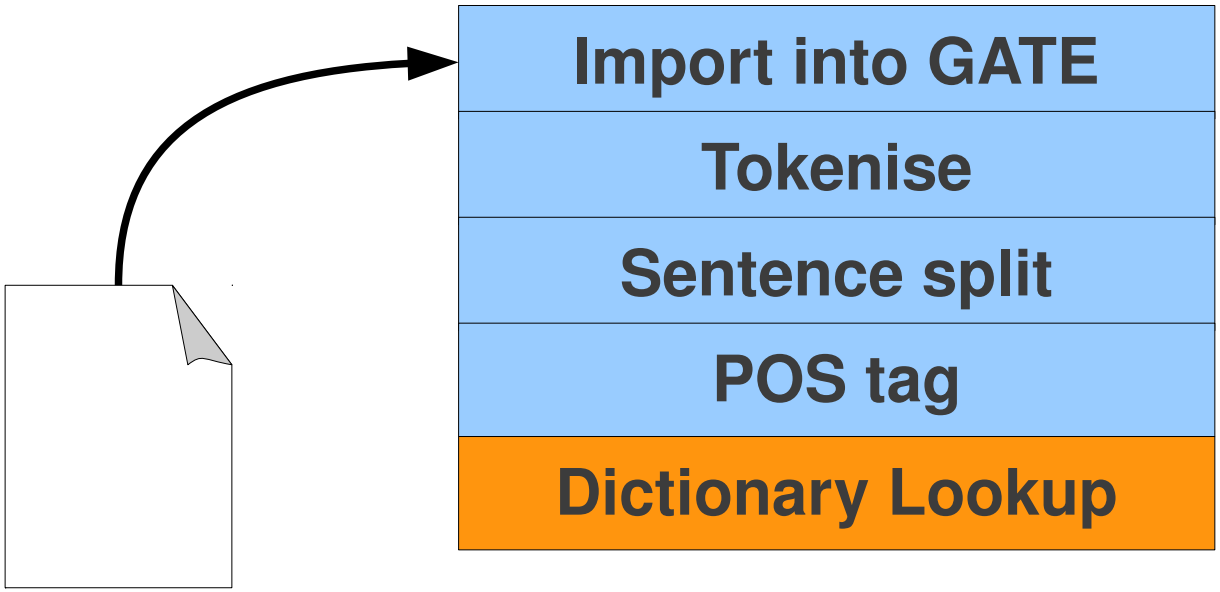
# MMSE pipeline



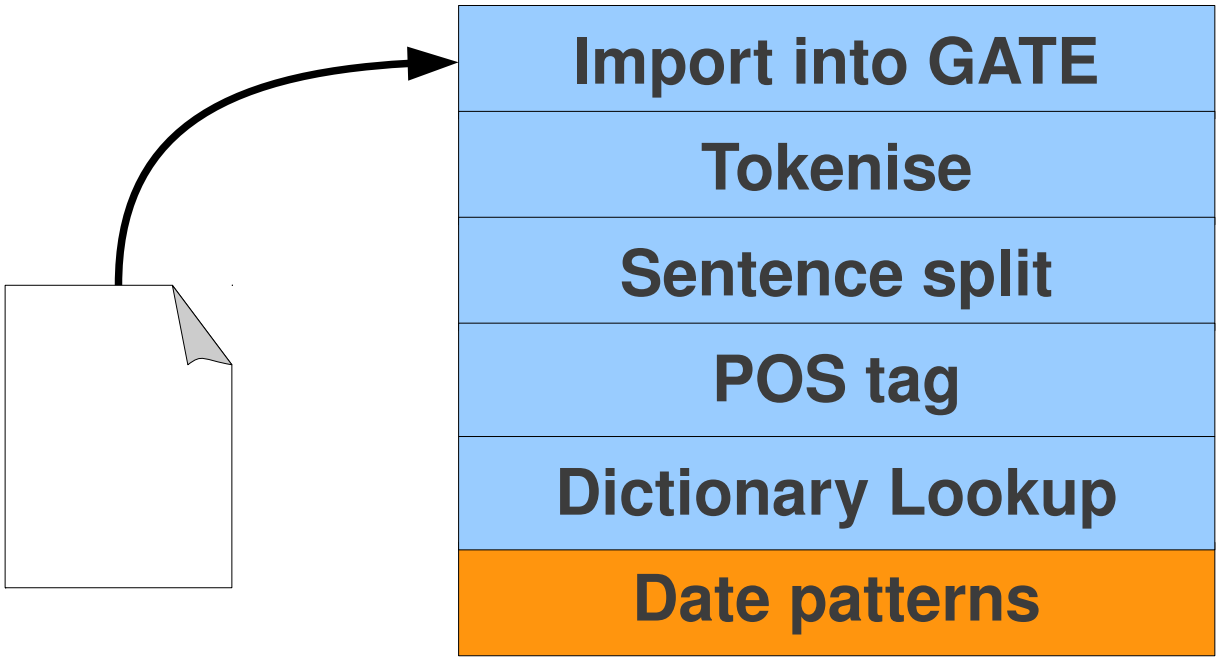
# MMSE pipeline



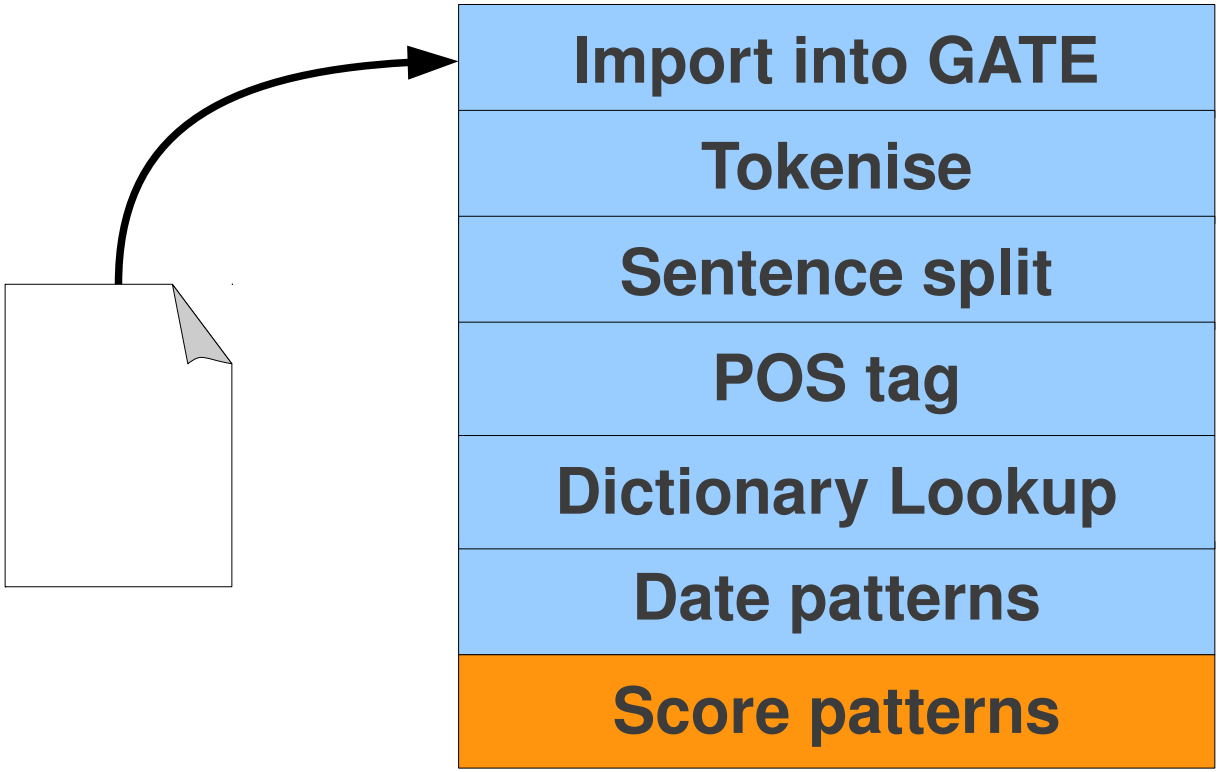
# MMSE pipeline



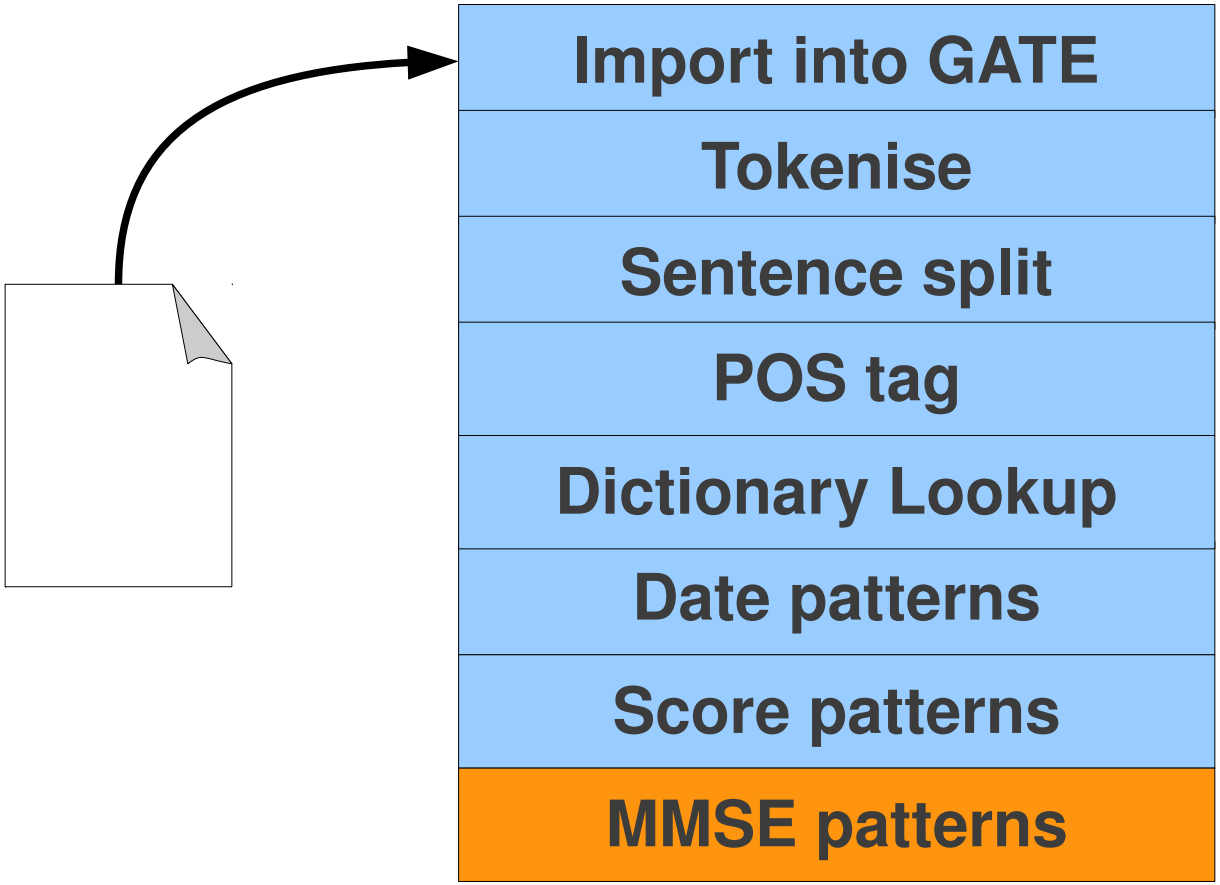
# MMSE pipeline



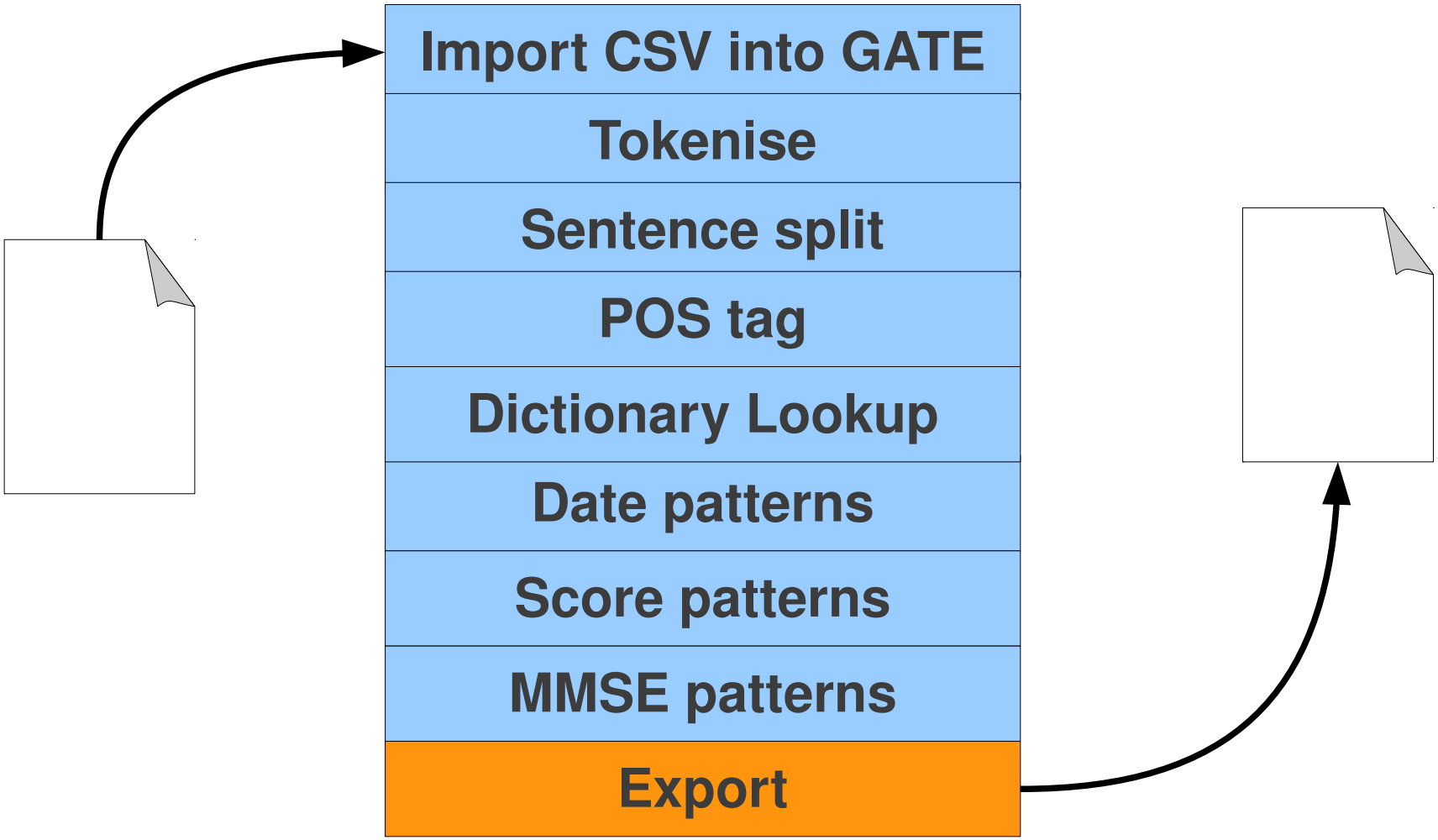
# MMSE pipeline



# MMSE pipeline



# MMSE pipeline




# Results

Accuracy for correctly identifying target text and features, measured against unseen data

| Application | Iterations | Recall | Precision |
|---|---|---|---|
| MMSE | 6 | 0.89 | 0.94 |
| Diagnosis text only | 6 | 0.46 | 0.50 |
| Smoking and status | 6 | 0.58 | 0.93 |


# Results

Accuracy of correctly identifying target text and features, measured against unseen data

| Application | Iterations | Recall | Precision |
|---|---|---|---|
| Medication | 7 | 0.62 | 0.9 |
| Dose, route, start, stop | 7 | 0.59 | 0.87 |
| Education level | 7 | 0.25 | 1.00 |
| Left school age | 7 | 1.00 | 1.00 |
| Lives alone | 2 | 1.00 | 0.93 |


# Results

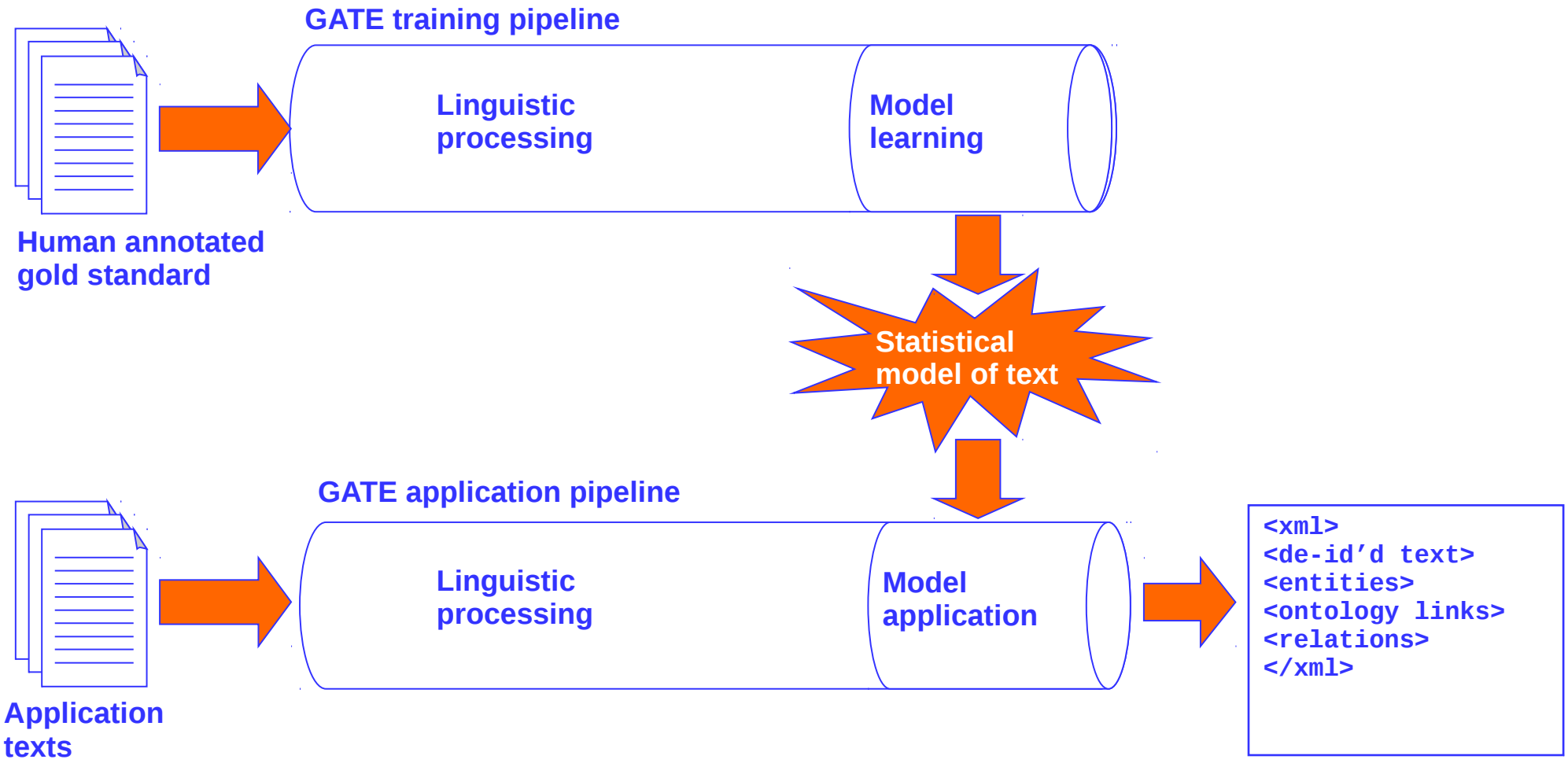
Accuracy for correctly identifying social care interventions and their currency (past, current, planned etc), measured against development data

| Application | Iterations | Recall | Precision |
|---|---|---|---|
| Care home | 5 | 0.73 | 0.82 |
| Generic care package | 5 | 0.79 | 0.78 |
| Day care | 5 | 0.89 | 0.79 |
| Home care | 5 | 0.96 | 0.88 |
| Meals on wheels | 5 | 0.89 | 1.00 |
| Respite care | 5 | 0.84 | 0.81 |
| Overall accuracy | | 0.82 | 0.82 |

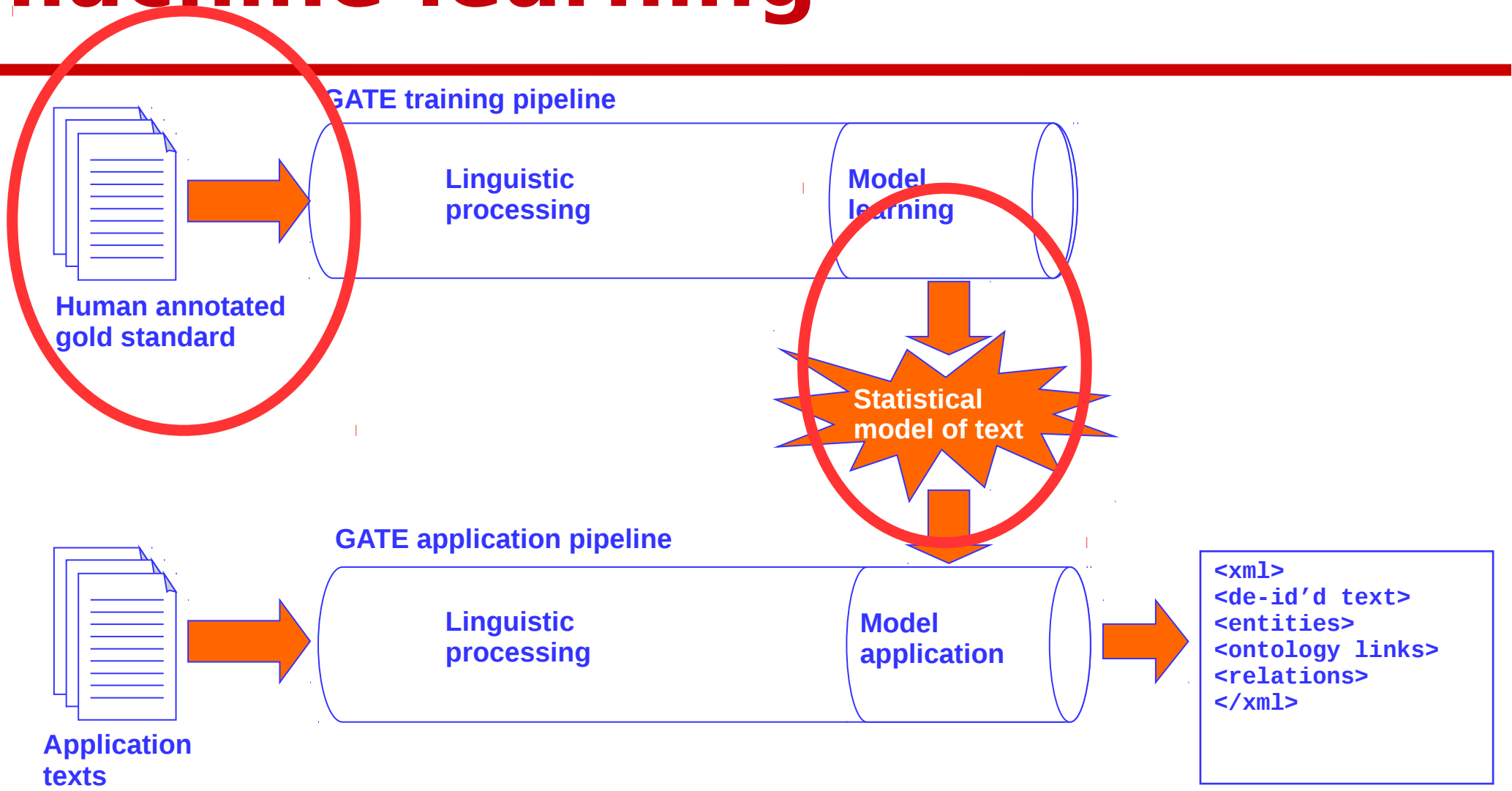

# PANSS

- Positive and negative syndrome scale
- Task is to find negative symptoms of schizophrenia to assist with assigning PANSS scores
  - Poverty of speech
  - Poor rapport
  - Emotionally withdrawn
- Symptomatology not recorded in structured data, and each symptom generally expressed in a single sentence.
  - Mr ZZZ is emotionally withdrawn
  - Eye contact was poor, and affect was blunted
  - Rapport was good

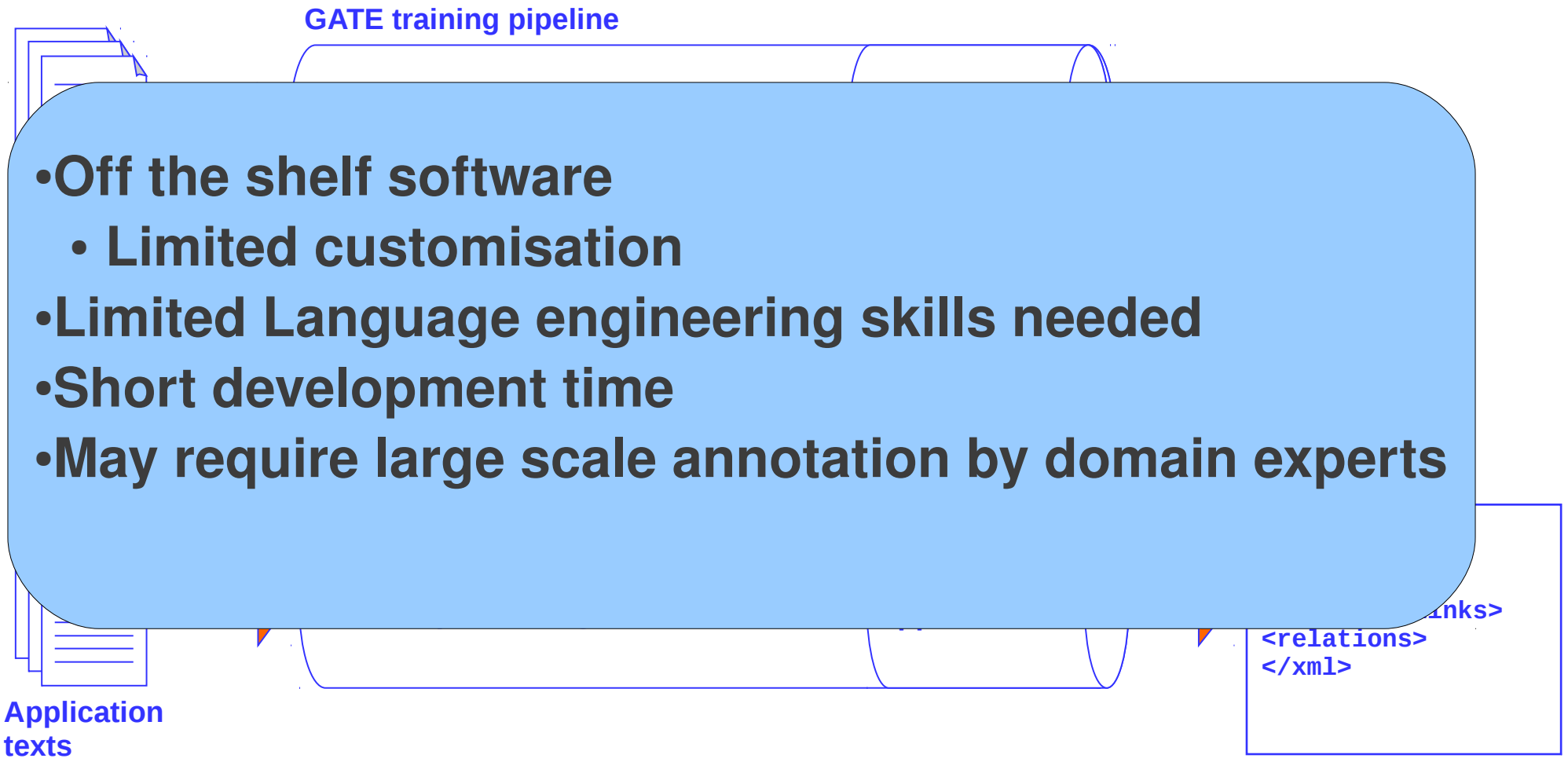
# Supervised machine learning



# Supervised machine learning



# Supervised machine learning




- Off the shelf software
  - Limited customisation
- Limited Language engineering skills needed
- Short development time
- May require large scale annotation by domain experts


# Classification

- Sentence classification
  - **POS:** *Affect was reactive*
  - **NEG:** *Seen in clinic today*
  - **NEG:** *No doubt this will affect her treatment*
- Train on manually classified example sentences
- Algorithms: Perceptrons and Support Vector Machines
- Feature set still under development
  - Unigrams on token string and POS
  - Unigrams on dictionary lookups


# Results

## Initial results against unseen data (k-fold cross validated)

| Application | Recall | Precision | F1 |
|---|---|---|---|
| **Abstract thinking** | **0.58** | **0.74** | **0.65** |
| **Affect** | **0.28** | **1.00** | **0.78** |
| **Apathy** | **0.89** | **0.60** | **0.72** |
| **Emotional withdrawal** | **0.06** | **0.50** | **0.10** |
| **Eye contact** | **0.54** | **0.82** | **0.65** |
| **Poverty of speech** | **0.27** | **0.83** | **0.40** |
| Rapport | 0.56 | 0.82 | 0.67 |
| Social withdrawal | 0.58 | 0.85 | 0.69 |

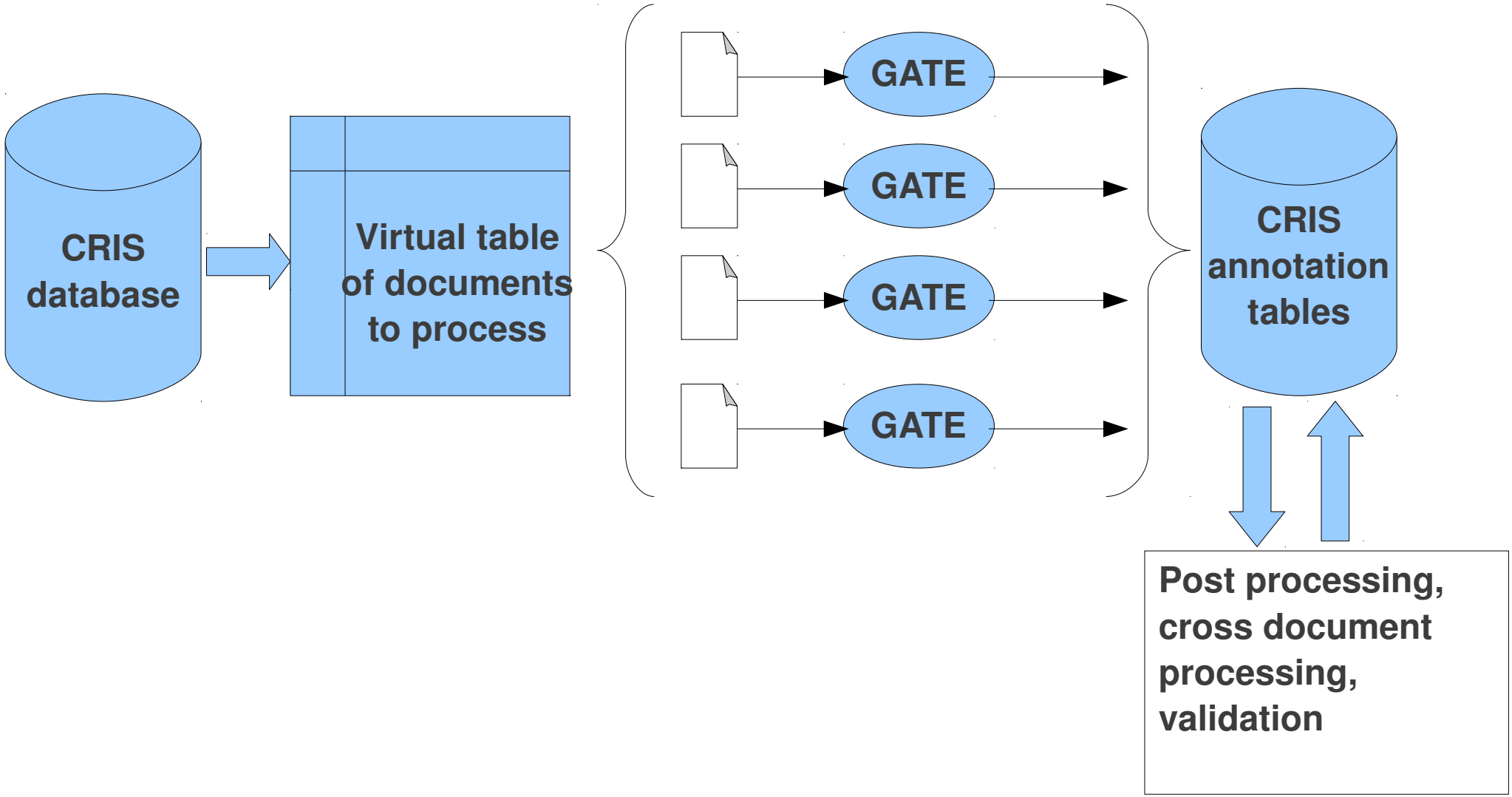

# Active learning

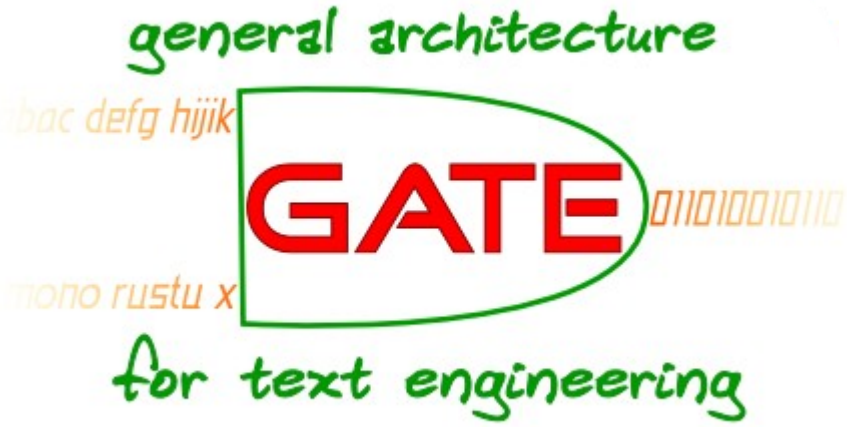
- The more examples machine learning sees, the more accurate the model
- How can we generate large numbers of training examples?
- Active learning
  - Run ML over unseen data
  - Split results into annotations for which ML has a high confidence, and those for which it has a low confidence
  - Manually correct only those annotations for which the ML has a low confidence
  - Re-train with both sets of annotations
  - Repeat
- Reduces annotator workload, iteratively improves model, and quickly creates large amounts of training data by concentrating on just the low confidence examples

# GATE on the BRC cluster


- GATE is regularly used to process 15 million documents
- At one document every 200ms this is 35 days of processing
- Expanding to 75 million documents
- The original proof of concept exported documents from the CRIS database to text files, and processed them on the BRC cluster using GATE's paralleliser
- Currently testing a CRIS-specific architecture, GATE interacting directly with the CRIS database

# GATE on the BRC cluster






# Thank you.
# Any questions?