Introduction to Text Mining

Module 4: Development Lifecycle (Part 1)

Aims of this module

- Turning resources into applications: SLAM
- RichNews: multimedia application and demo
- Musing: Business Intelligence application
- KIM CORE Timelines application and demo
- GATE MIMIR: Semantic search and indexing in use
- The GATE Process

Semantic Annotation for the Life Sciences

Aim of the application

- Life science semantic annotation is much more than generic annotation of genes, proteins and diseases in text, in order to support search
- There are many highly use-case specific annotation requirements that demand a fresh look at how we drive annotation – our processes
- Processes to support annotation
  - Many use cases are ad-hoc and specialised
  - Clinical research – new requirements every day
  - How can we support this? What tools do we need?

Background

- The user
  - SLAM: South London and Maudsley NHS Trust
  - BRC: Biomedical Research Centre
  - CRIS: Case Register Information System
- February and March 2010
  - Proof of concept around MMSE
  - Requirements analysis, installation, adaptation
- Since 2010
  - In production
  - Cloud based system
  - Further use cases

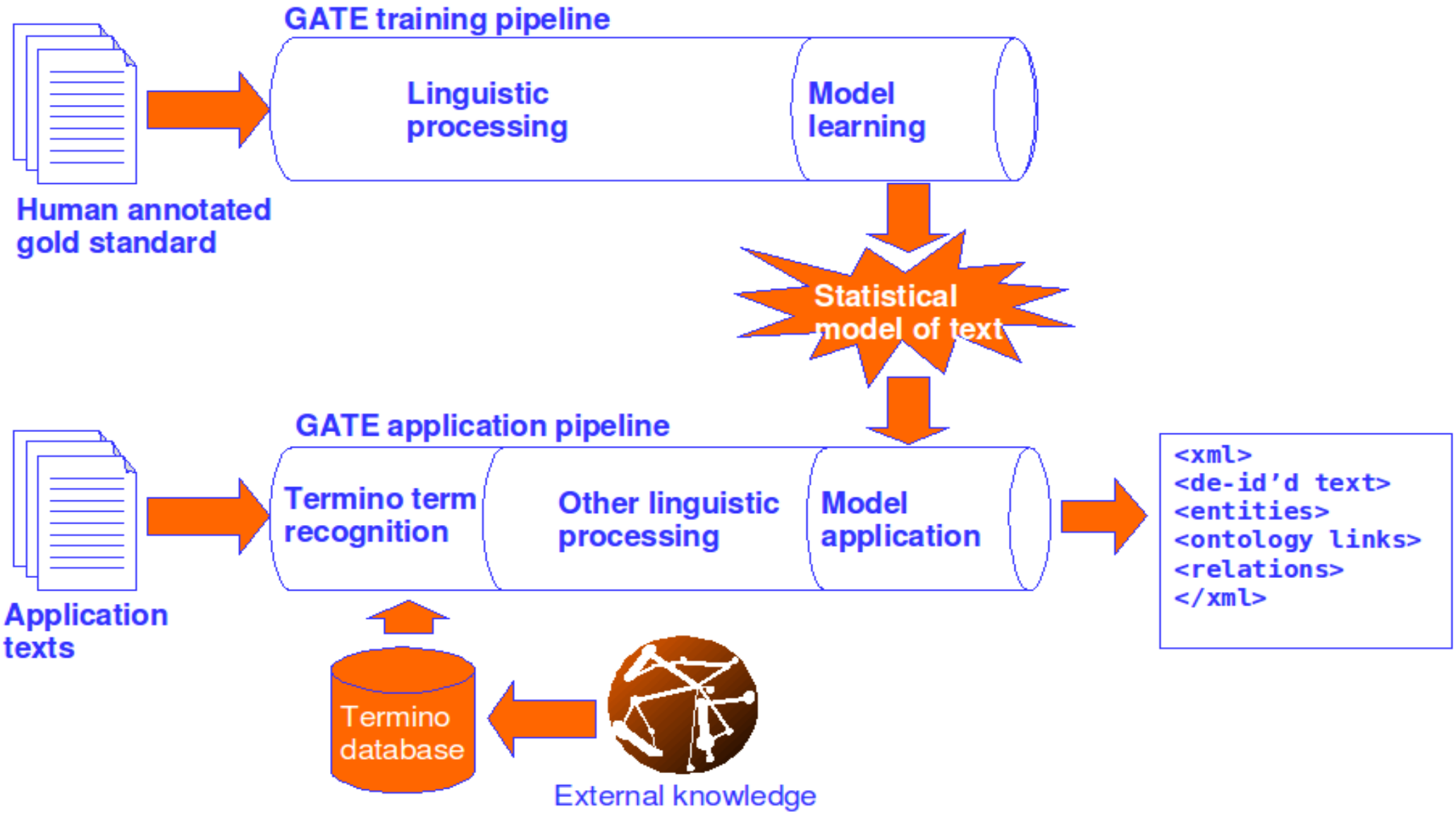
Clinical records

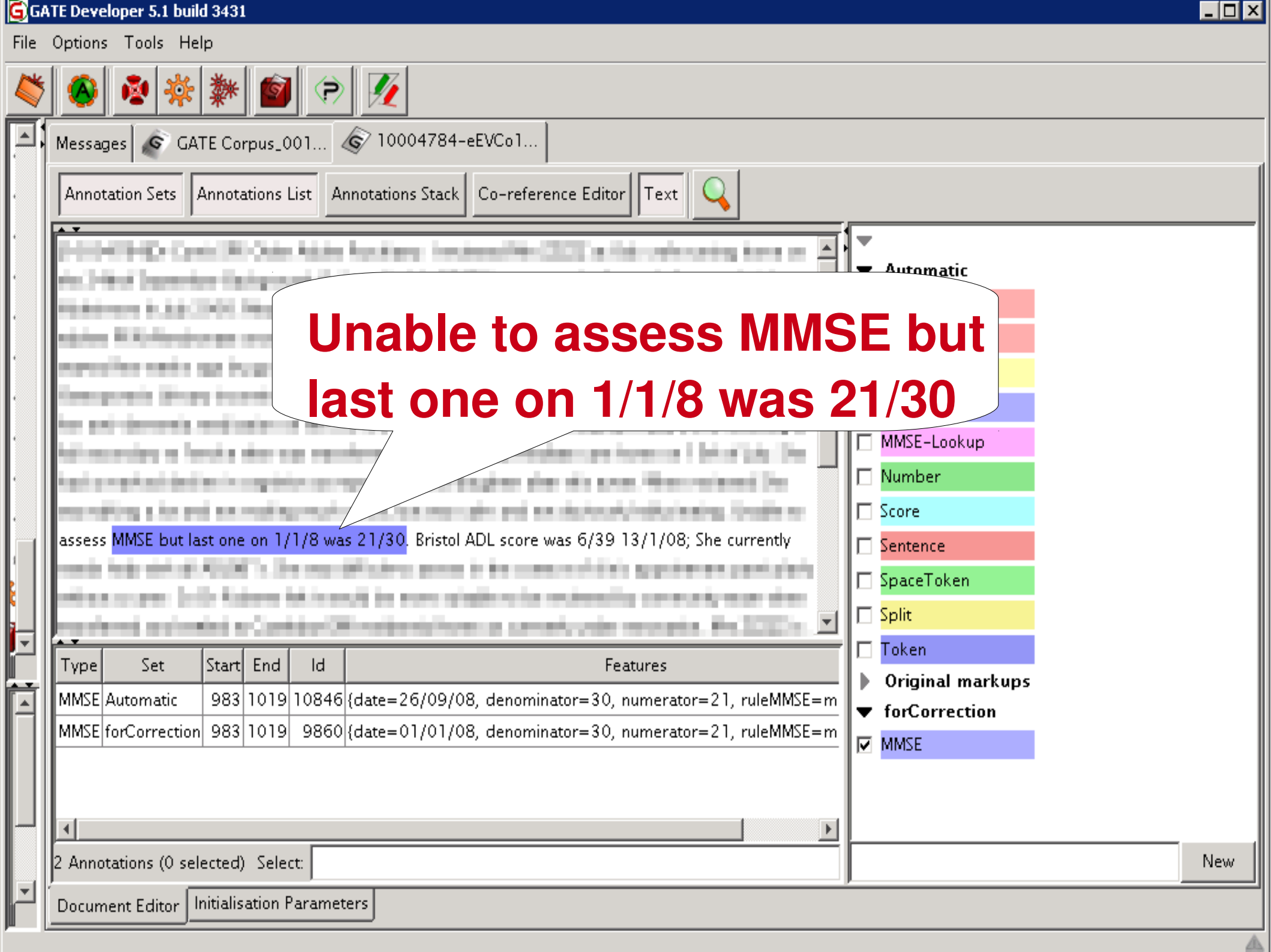
- Generic entities such as anatomical location, diagnosis and drug are sometimes of interest
- But many of the enquiries we have seen are more often interested in large numbers of very specific, and ad hoc entities or events
- This example is with a UK National Biomedical Research Centre
- An example – cognitive ability as shown by the MMSE score
- Illustrates a typical (but not the only) process

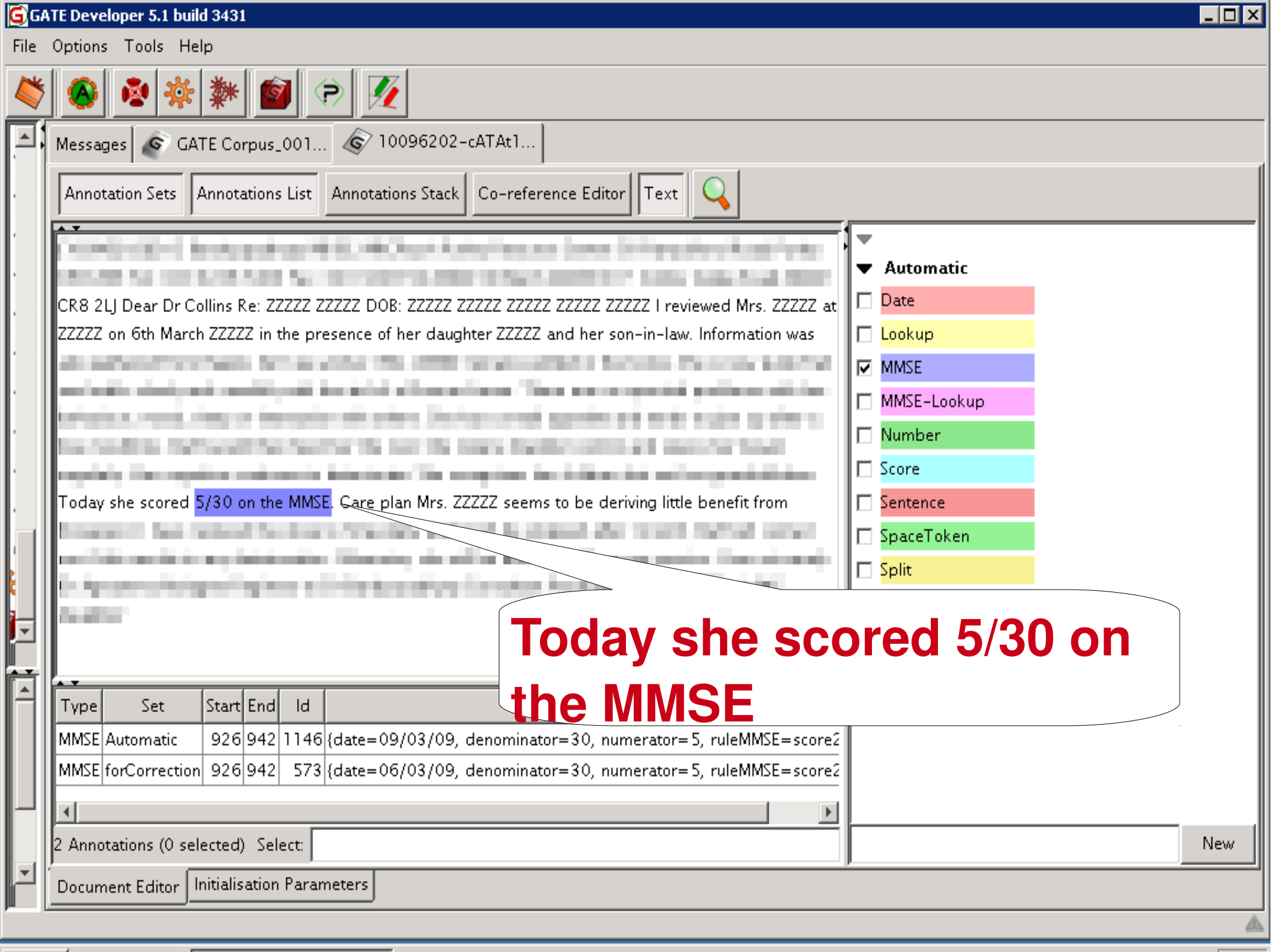
Types of IE systems

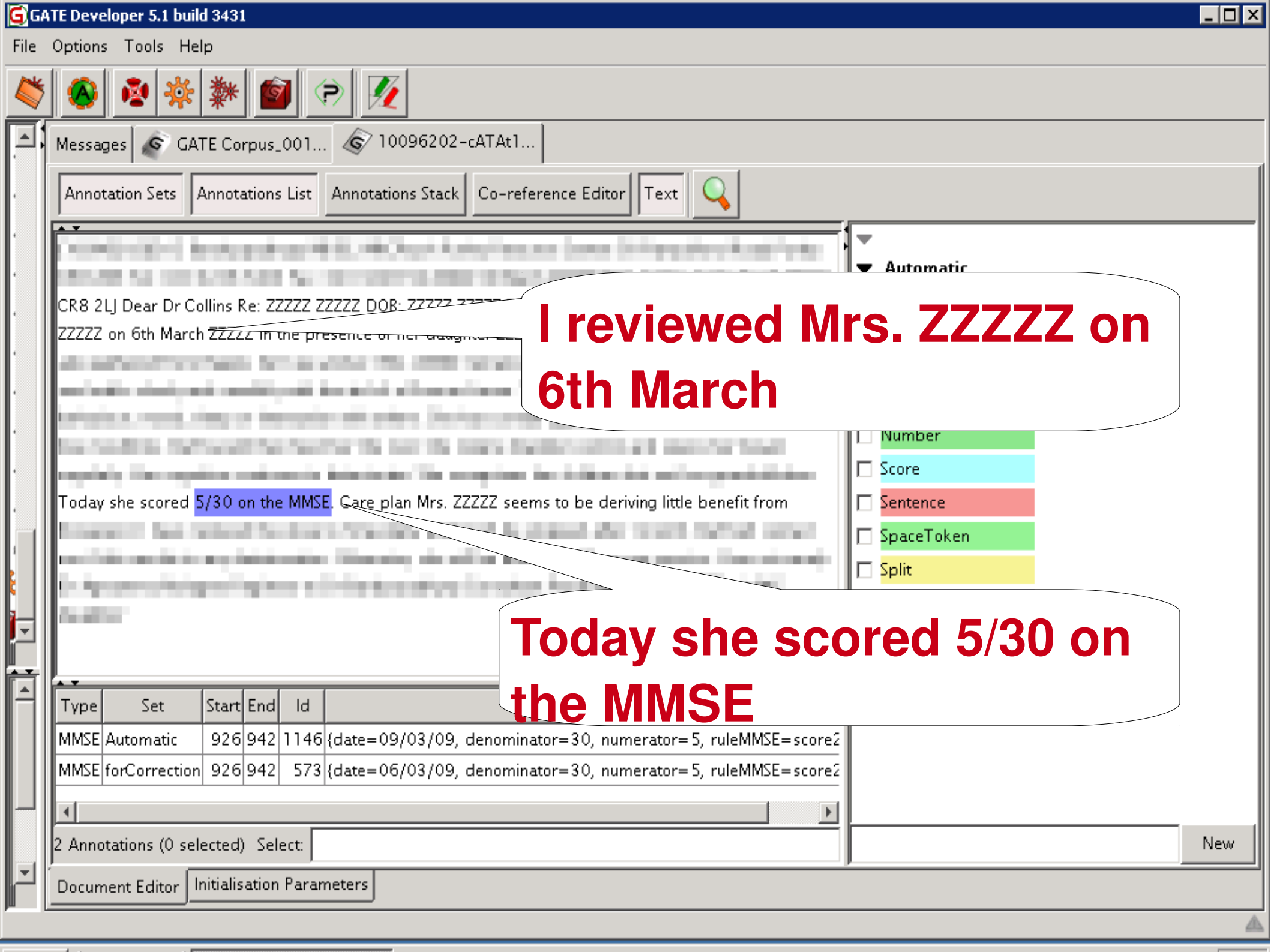
- Deep or shallow analysis
- Knowledge Engineering or Machine Learning approaches
  - Supervised
  - Unsupervised
  - Active learning
- GATE is agnostic

Supervised learning architecture




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
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

```
 His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|....
```

Annotations

```
 His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|....
```

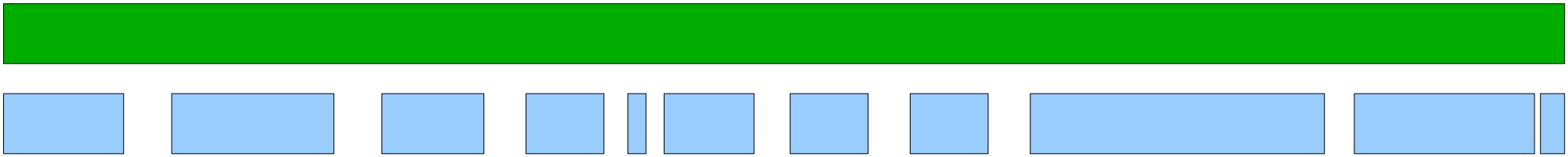
Annotations

```
 His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|....
```

Id	Type
1	sentence

Annotations

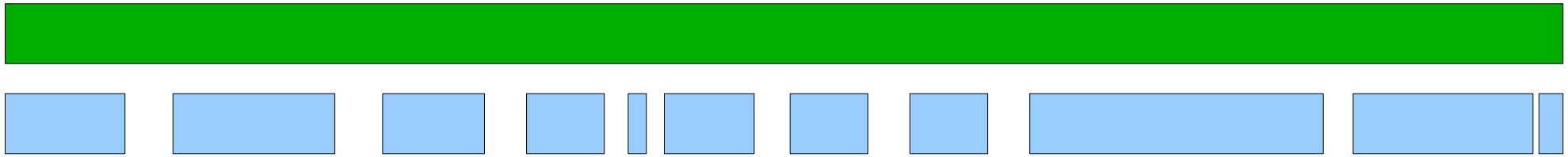
```
 His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|....
```



Id	Type
1	sentence

Annotations

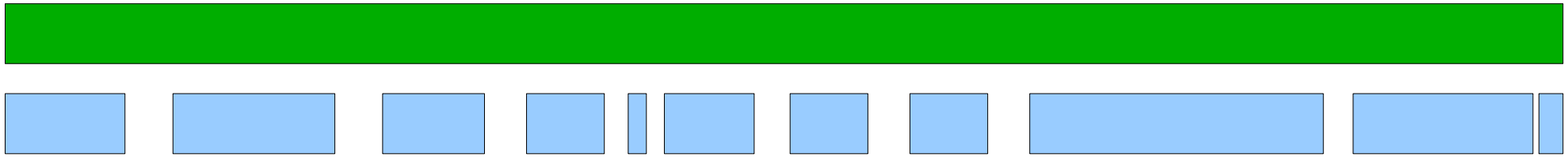
```
 His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|....
```



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

Annotations

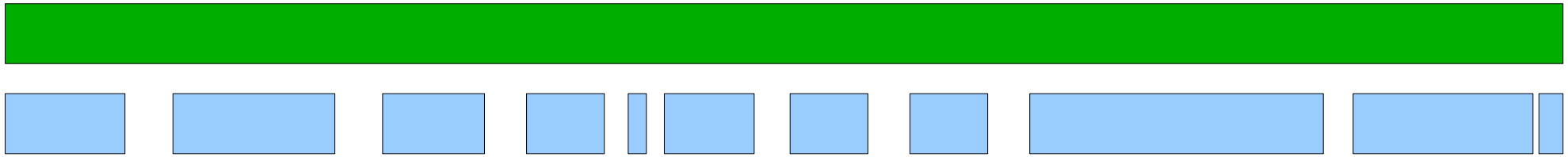
```
 His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|....
```



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

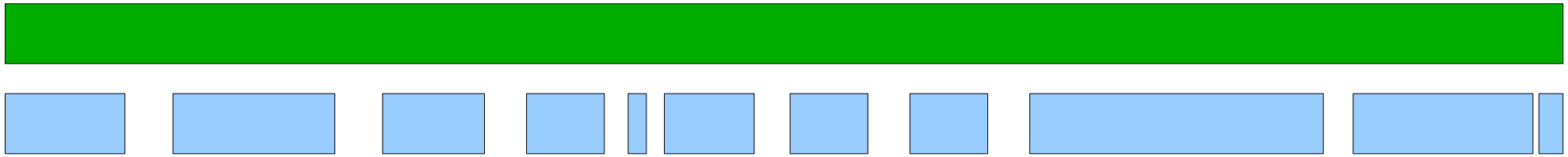
```
 His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|....
```



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

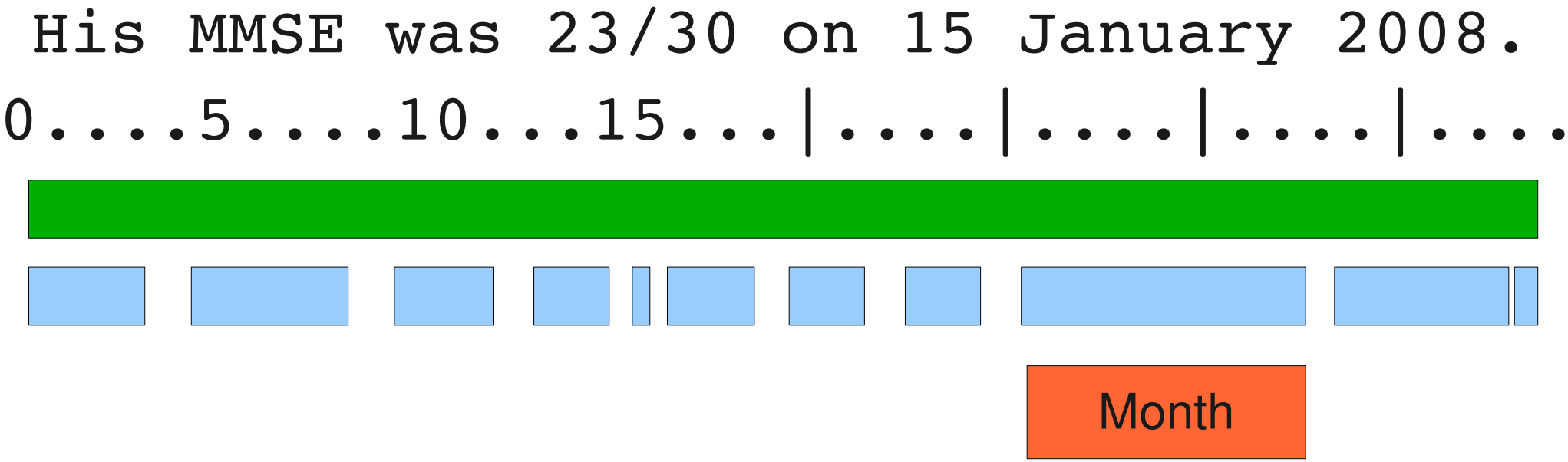
Annotations

```
 His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|....
```

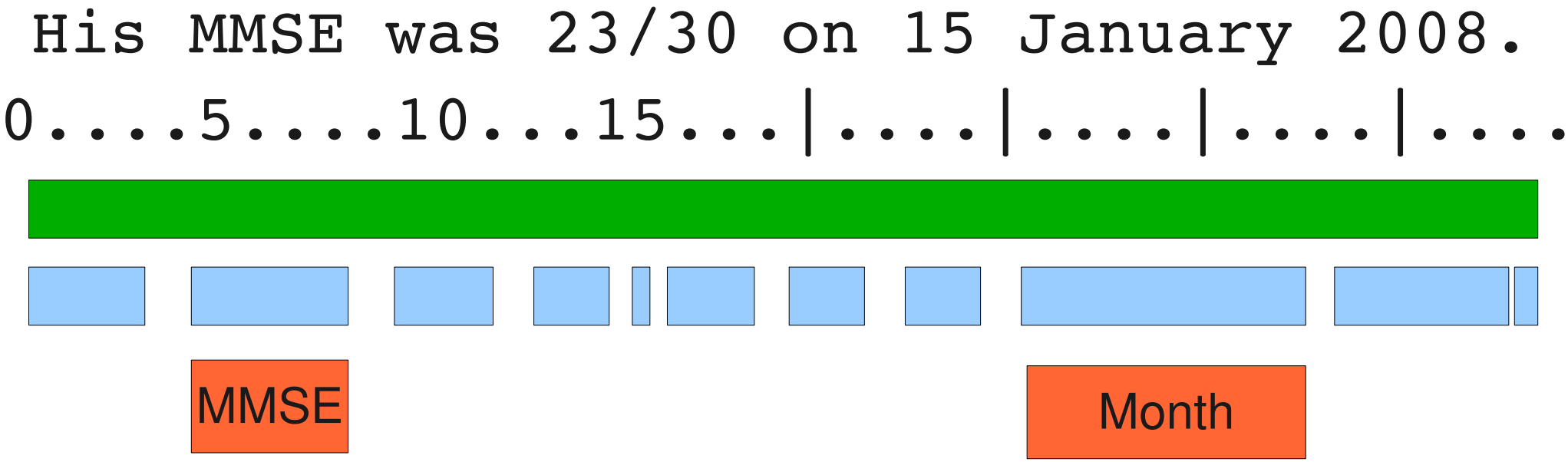


Id	Type	Start	End	Features	
1	sentence	0	39		
2	token	0	3	pos=PP	
3	token	4	8	pos=NN	
4	token	9	12	pos=VB	root=be
5	token	13	15	pos=CD	type=num
6	token	15	16	pos=SM	type=slash
7	token	16	18	pos=CD	type=num

Dictionary lookup



Dictionary lookup



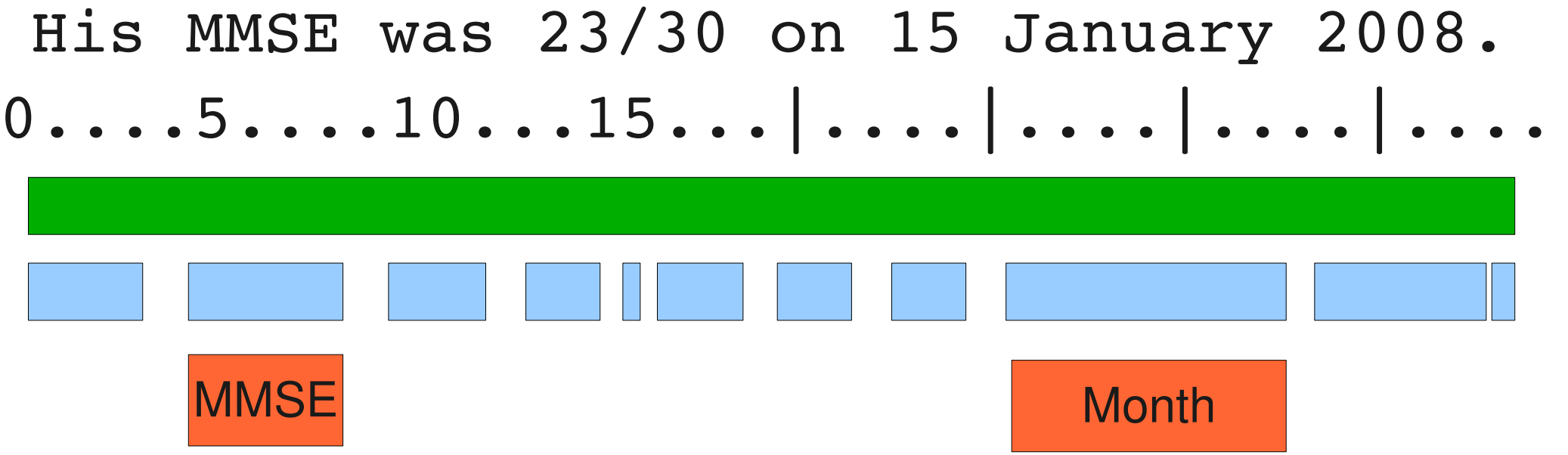
Limitations of dictionary lookup

- 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

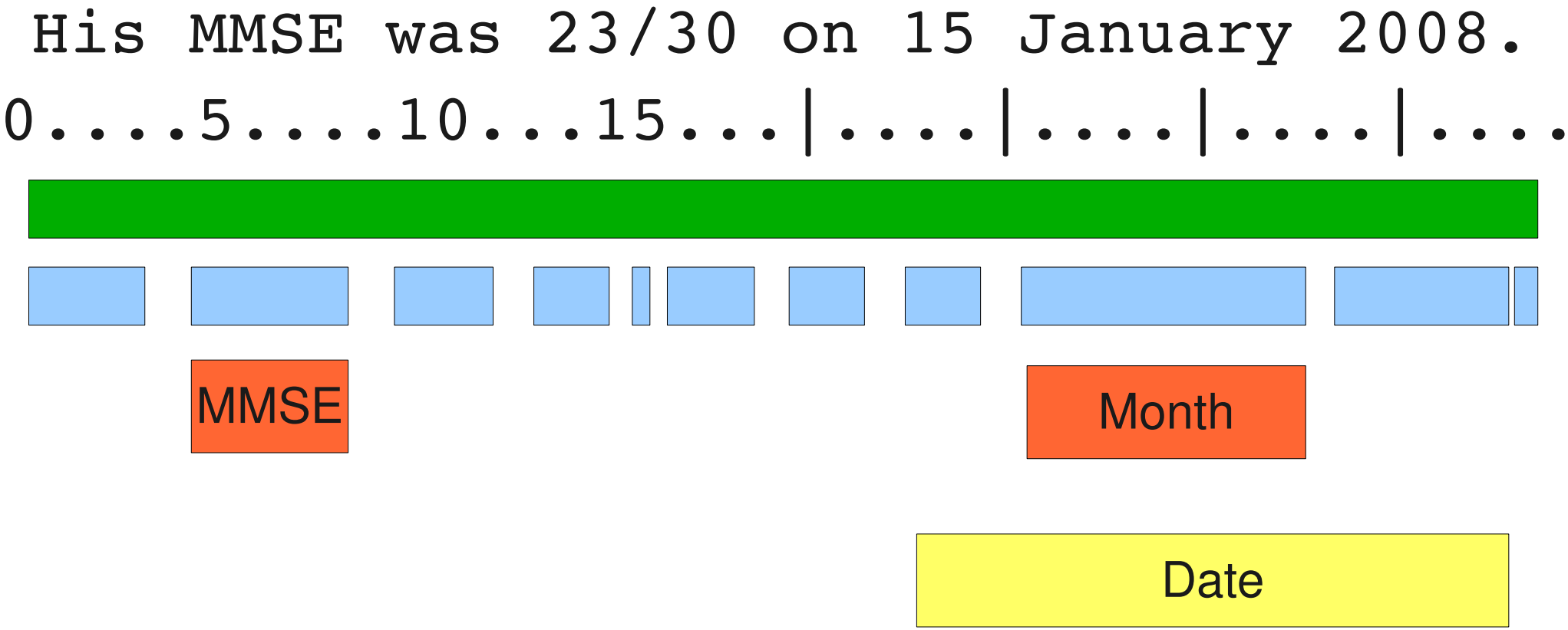
- The early components in a GATE pipeline produce simple annotations (Token, Sentence, Dictionary lookups)
- These annotations have features (Token kind, part of speech, major type...)
- Patterns in these annotations and features can suggest more complex information
- We use JAPE, the pattern matching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

Patterns

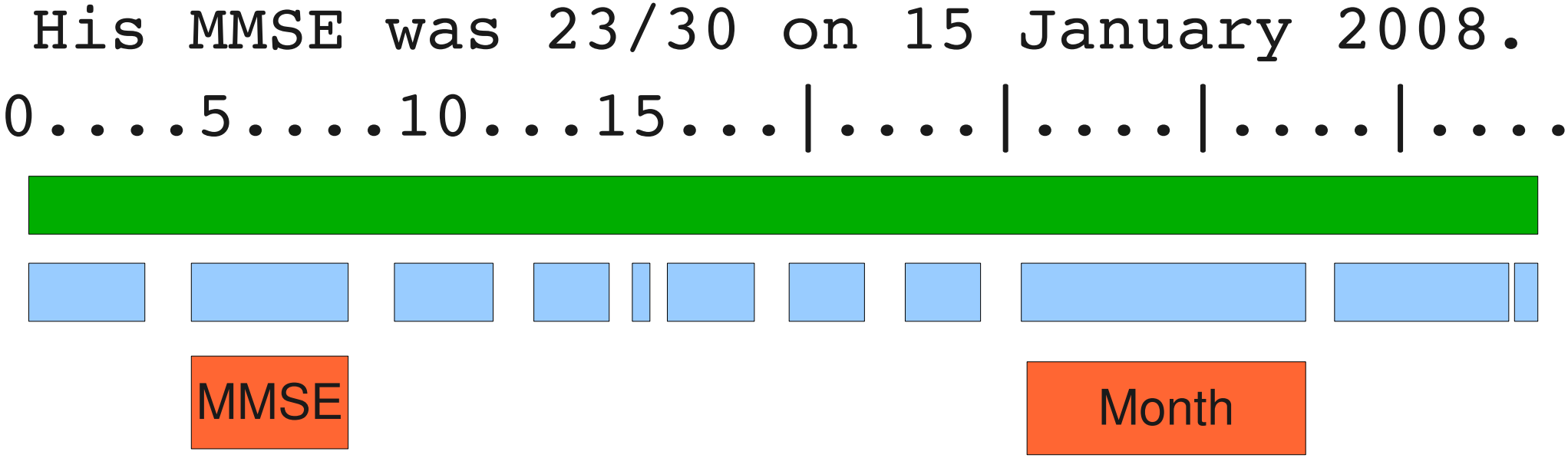


{number}{Month}{number}

Patterns

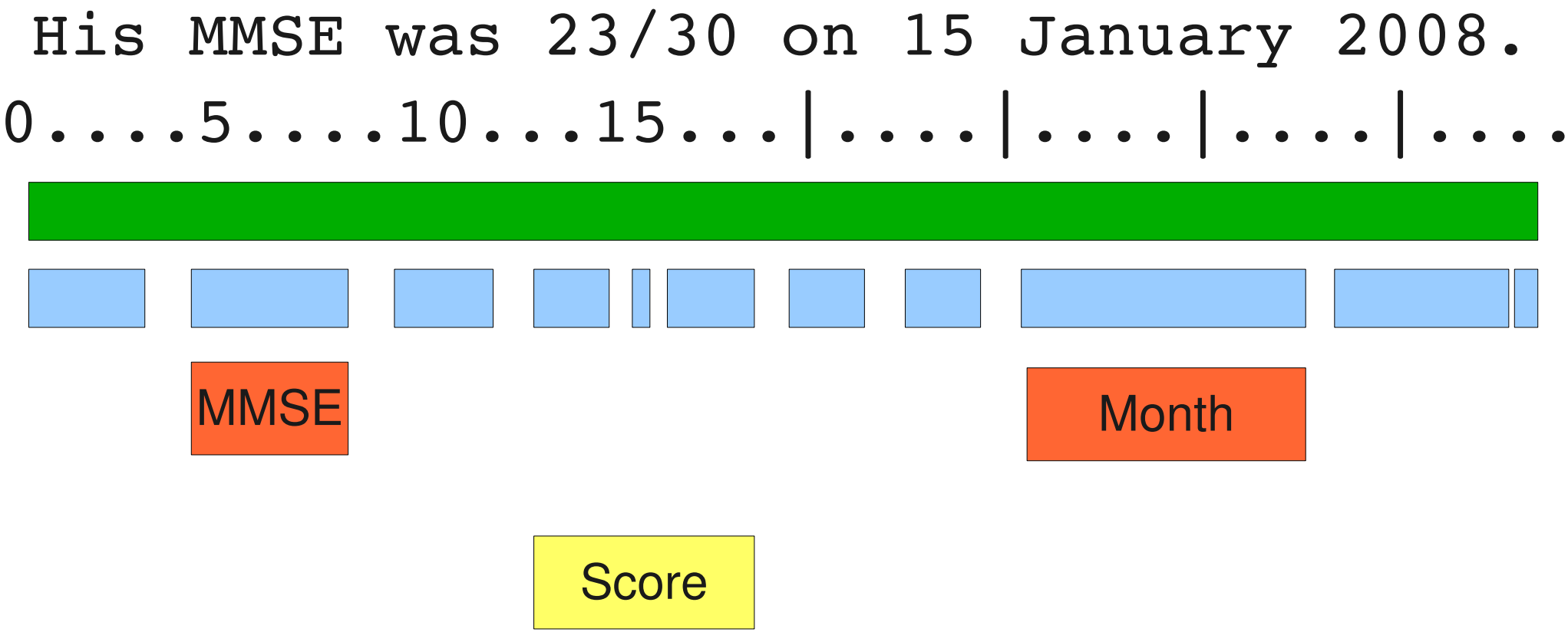


Patterns

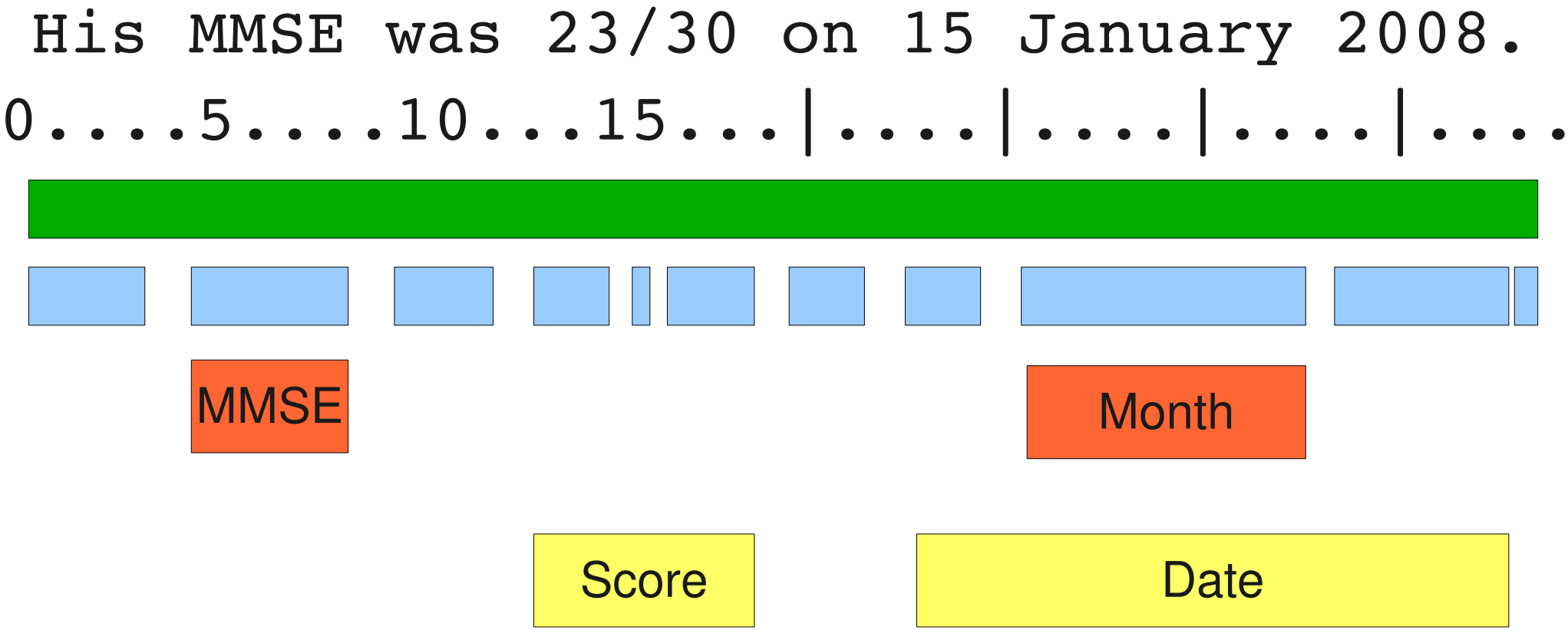


{number}{slash}{number}

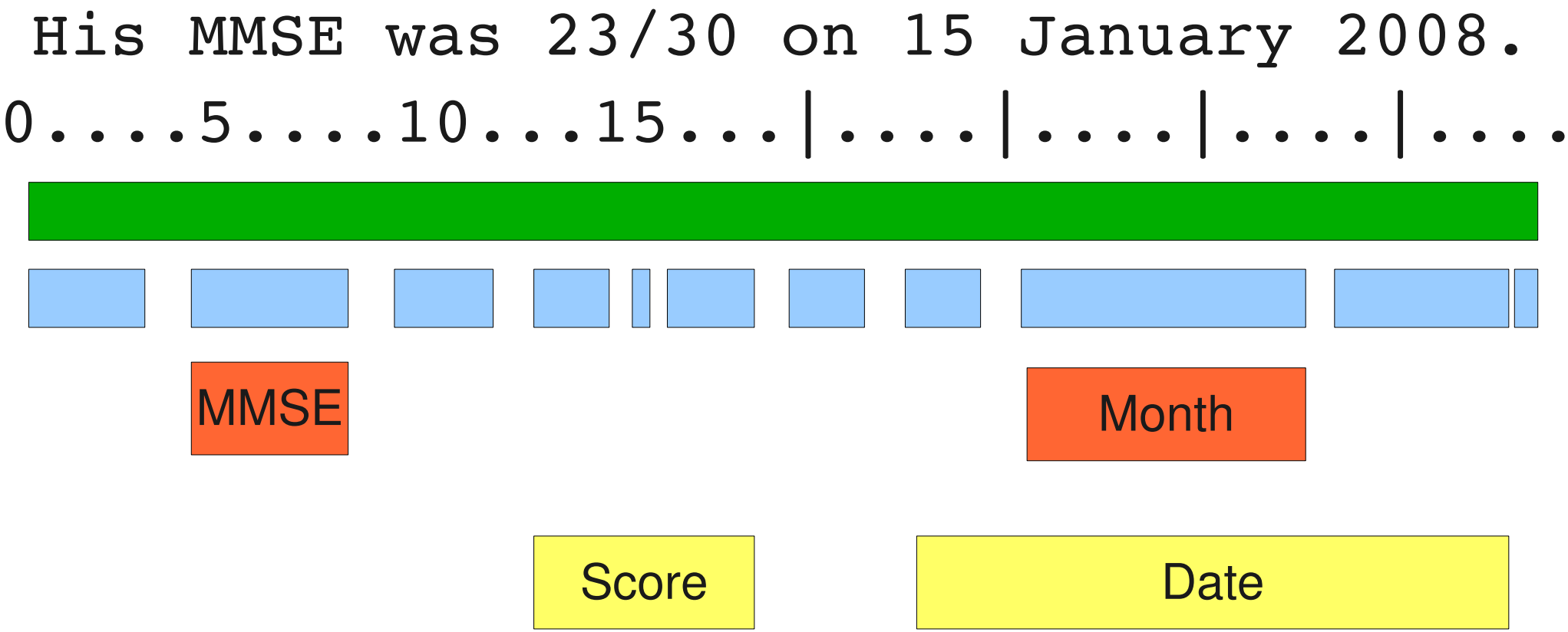
Patterns



Patterns

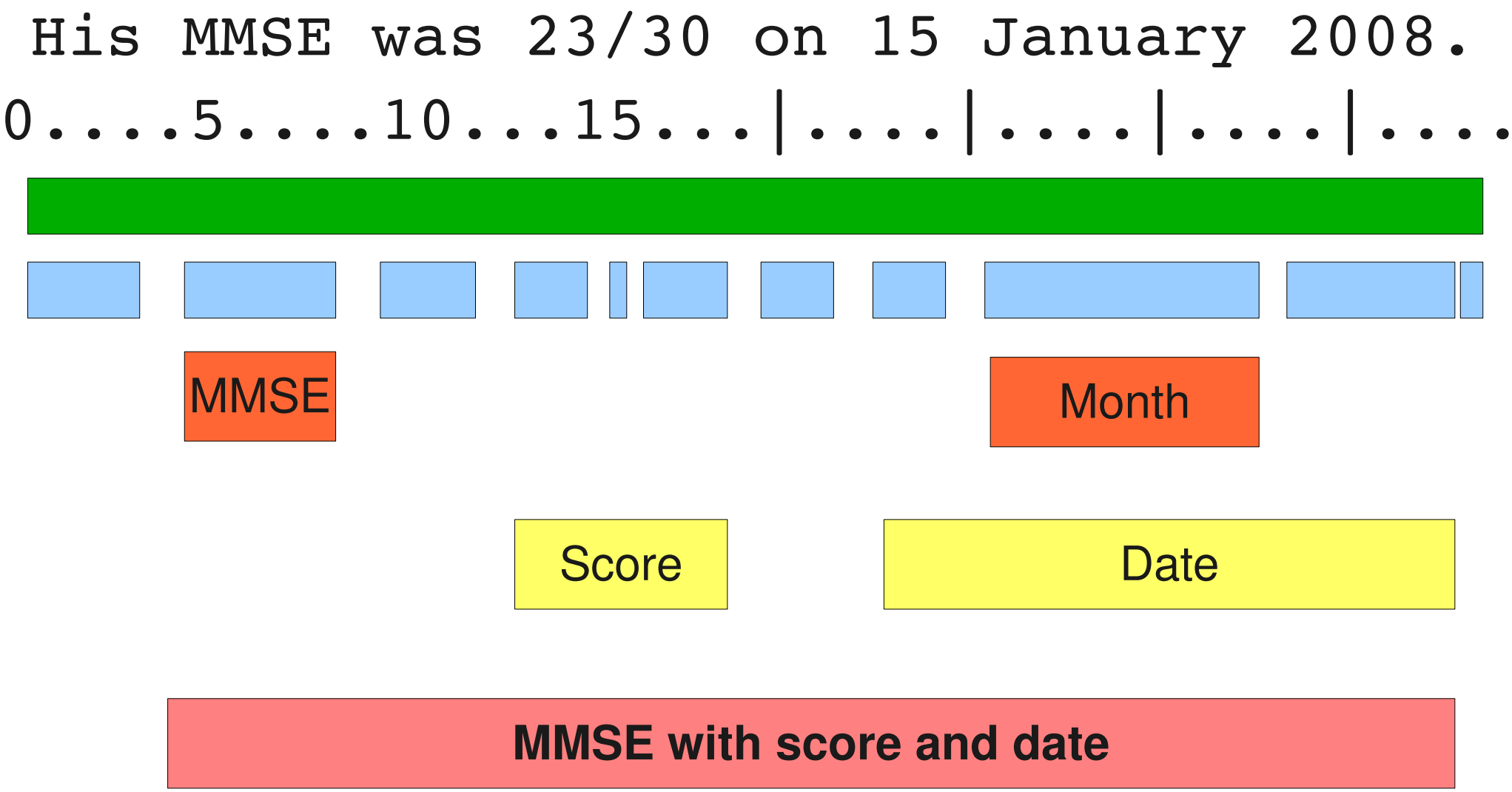

His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|....|....
MMSE
Month
Score
Date

Patterns



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

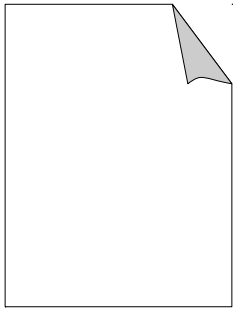
Patterns


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

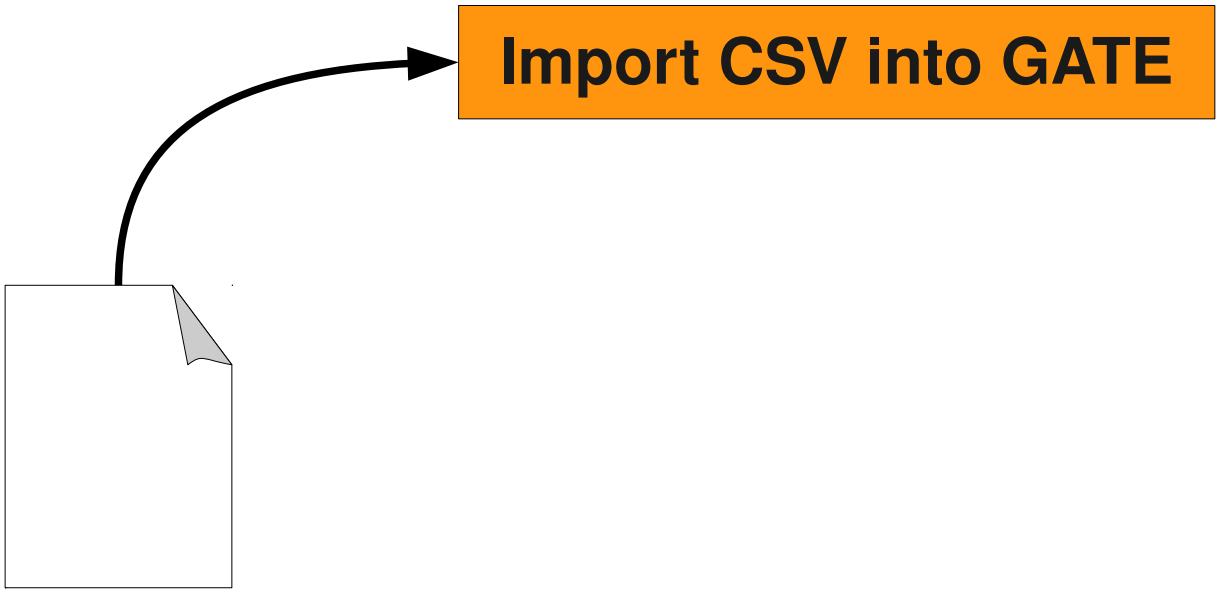
Patterns are general

- 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

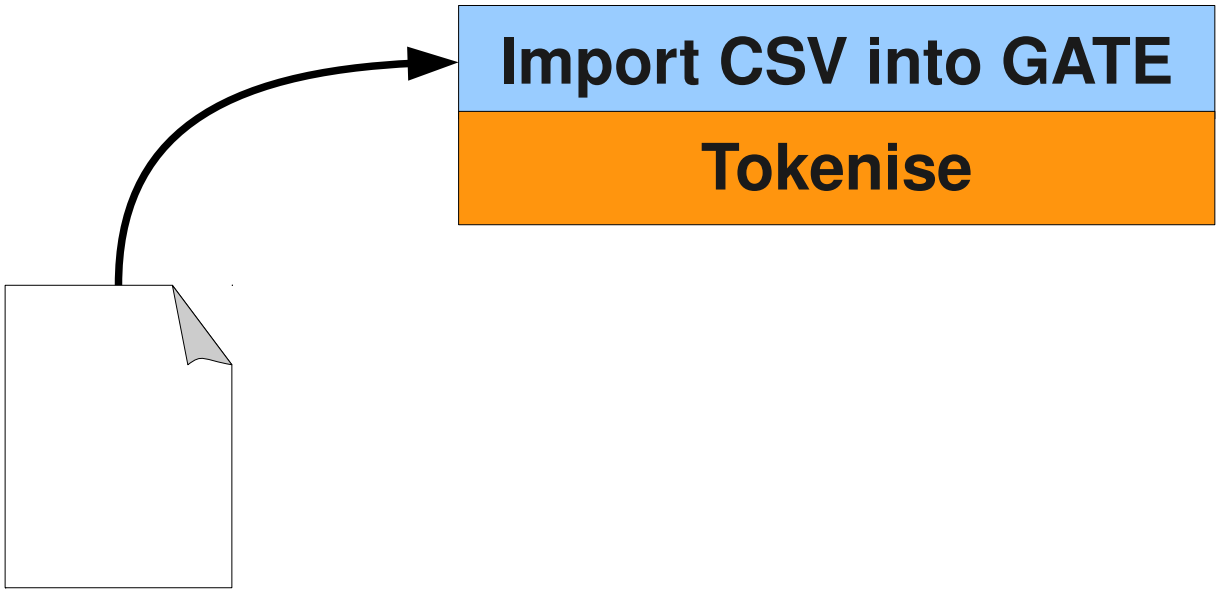
MMSE pipeline



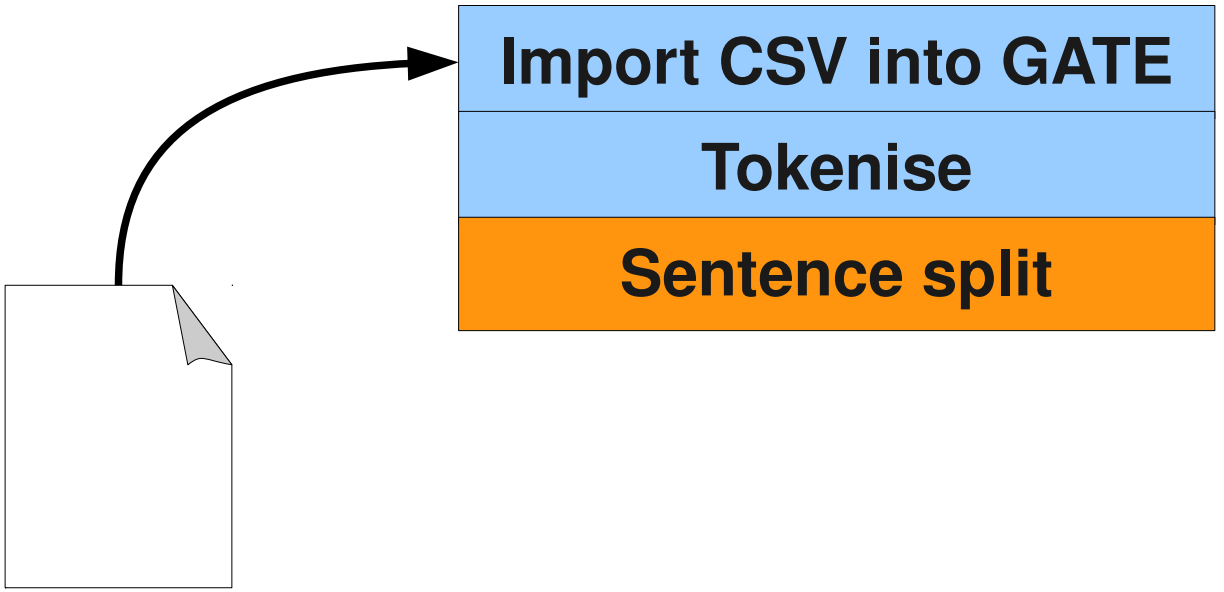
MMSE pipeline



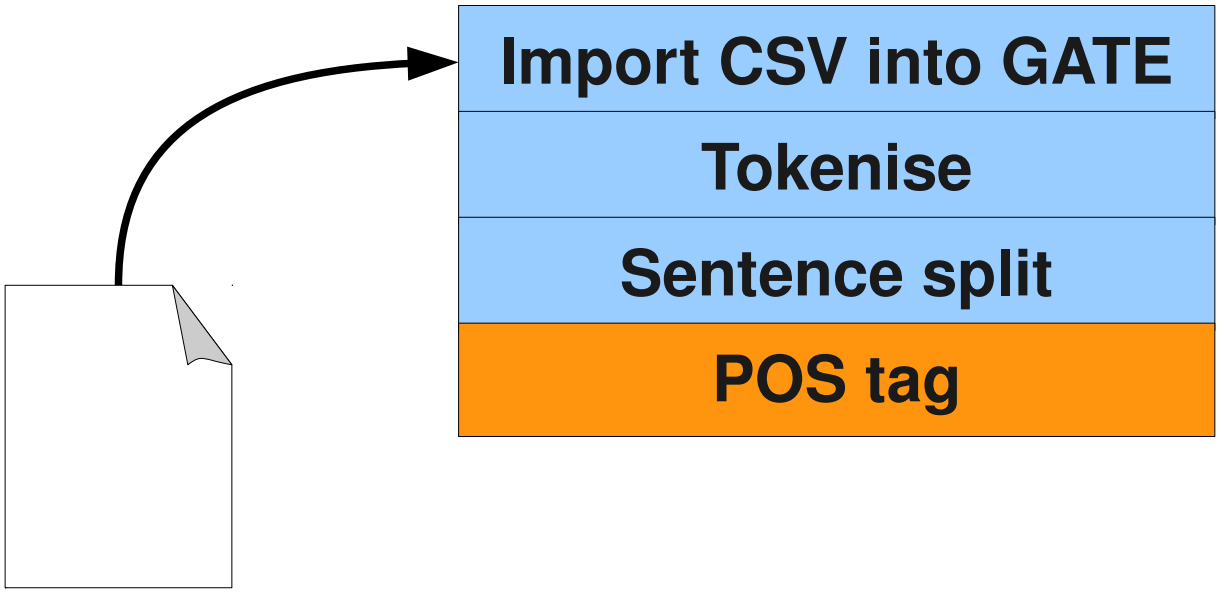
MMSE pipeline



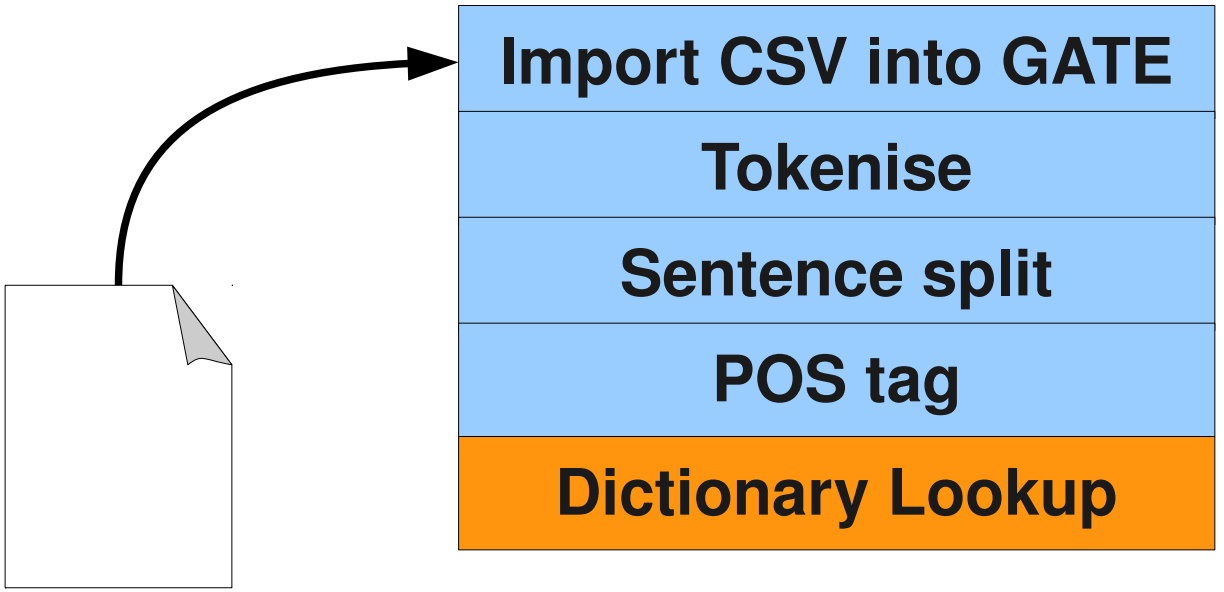
MMSE pipeline



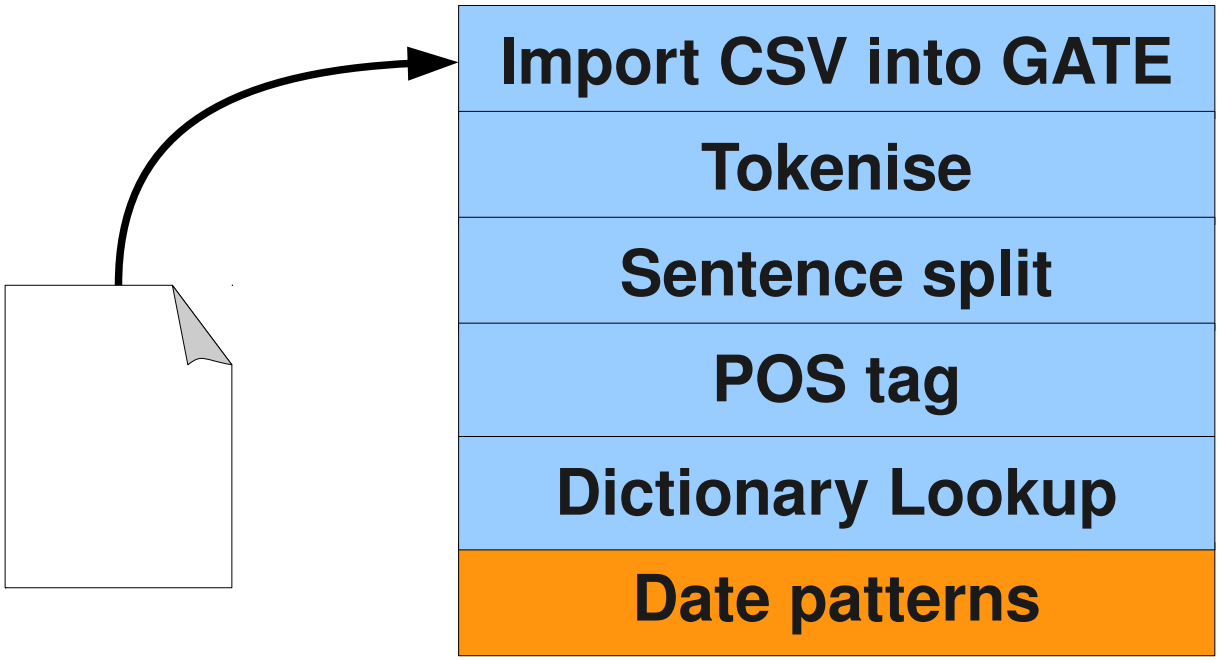
MMSE pipeline



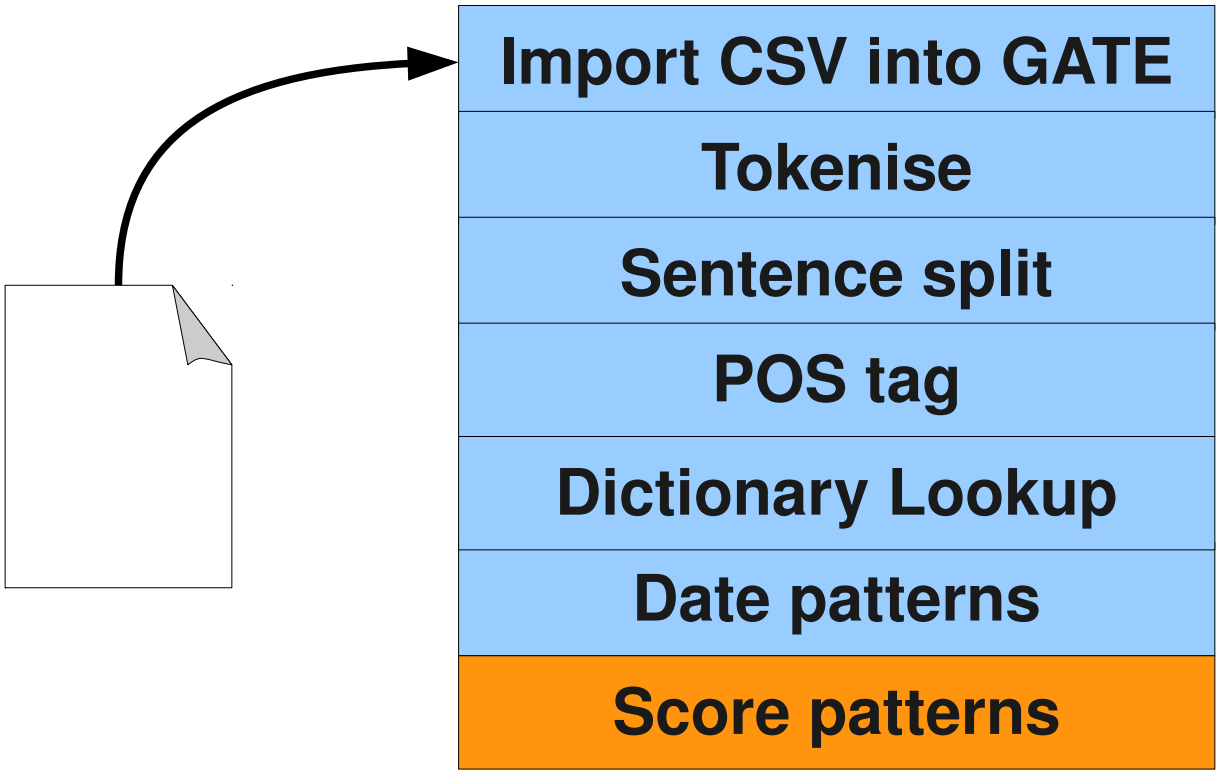
MMSE pipeline



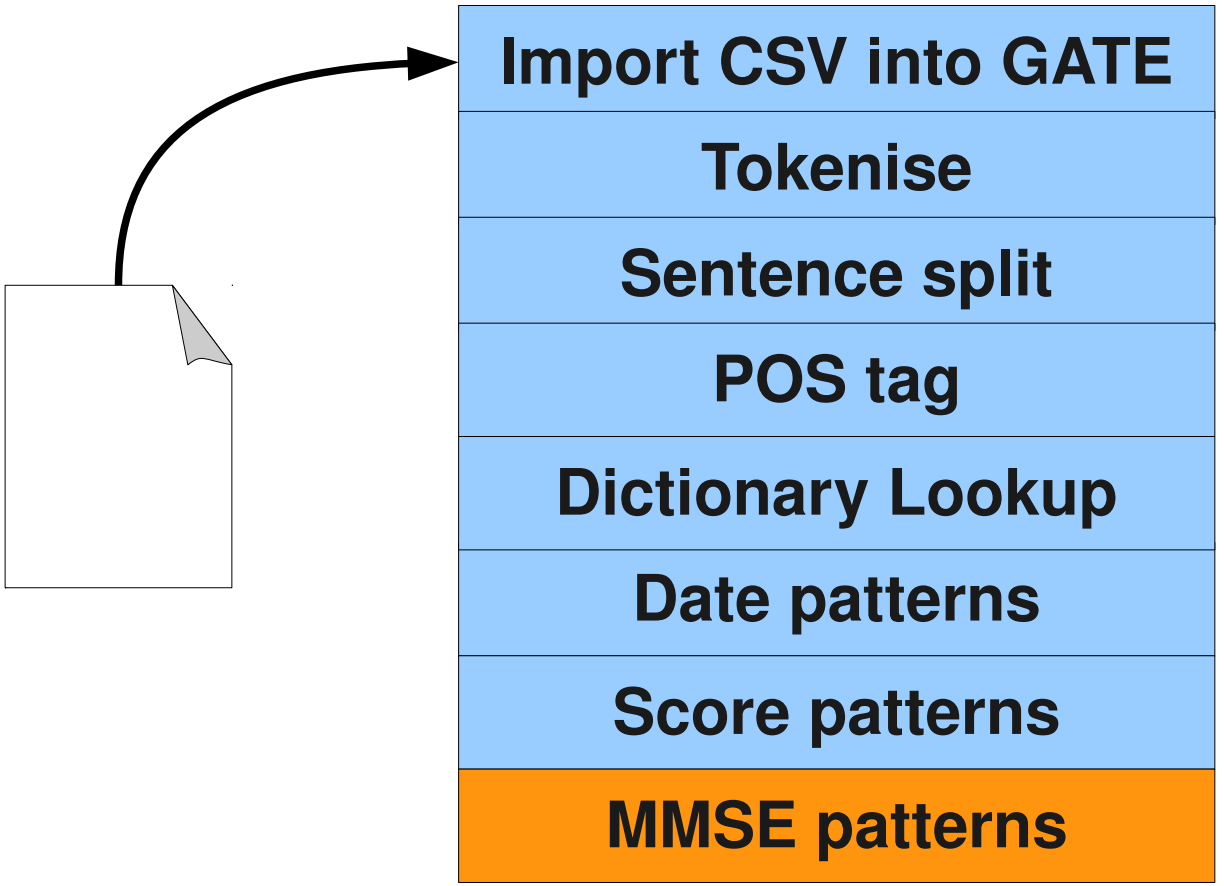
MMSE pipeline



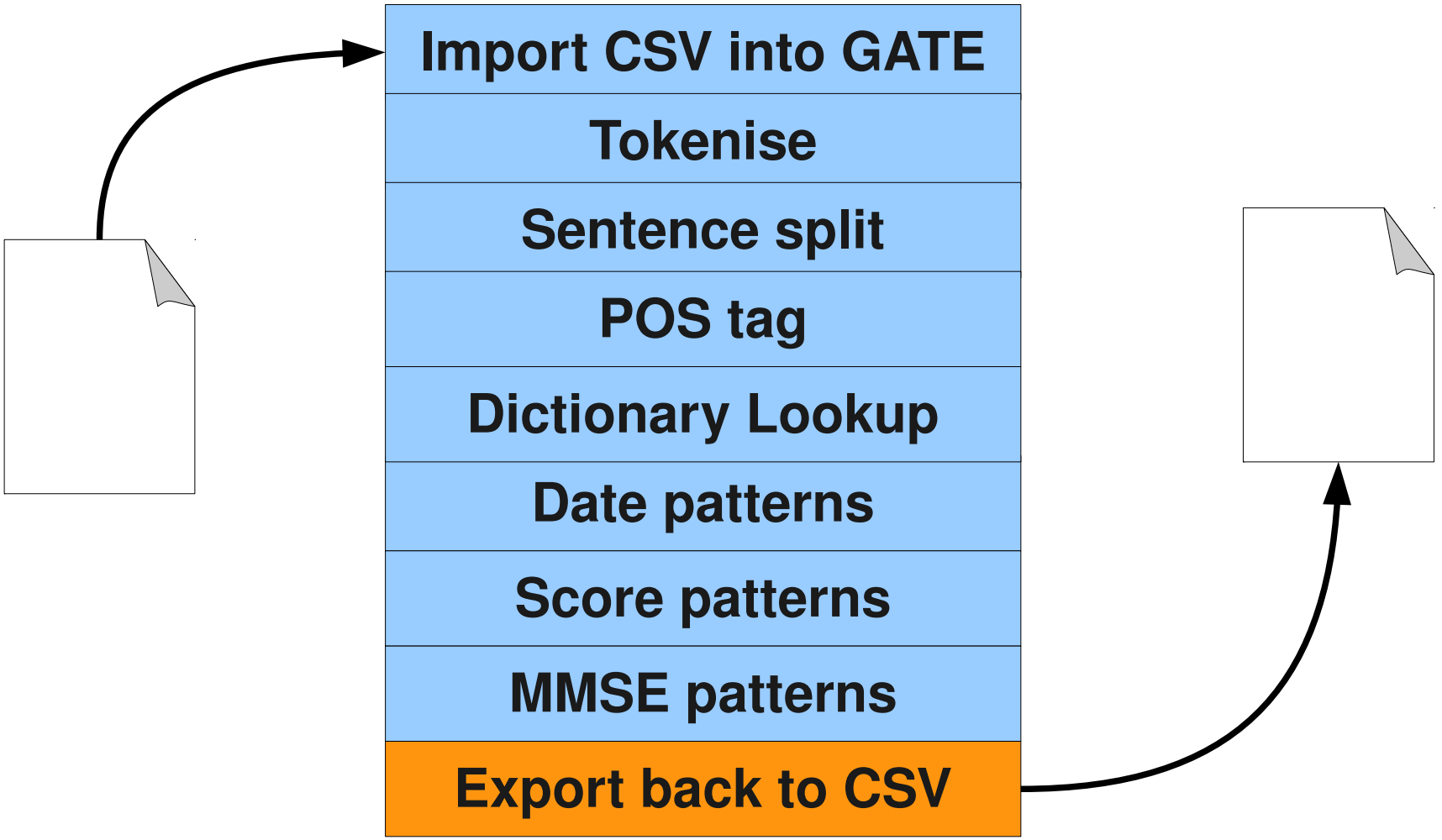
MMSE pipeline



MMSE pipeline



MMSE pipeline



Writing patterns

- Requires training
- Depending on time and skills, domain expert may take on some rule writing
- Requirements not always clear, and users do not always understand what the technology can do
- Needs a process to support
  - Domain expert manually annotates examples
  - Language engineer writes rules
  - Measure accuracy of rules
  - Repeat

The process as agile development

- IE system development is often linear
  - Guidelines → annotate → implement
- This is similar to the “waterfall” method of software development
  - Gather requirements → design → implement
- This has long been known to be problematic
- In contrast, our approach is **agile**

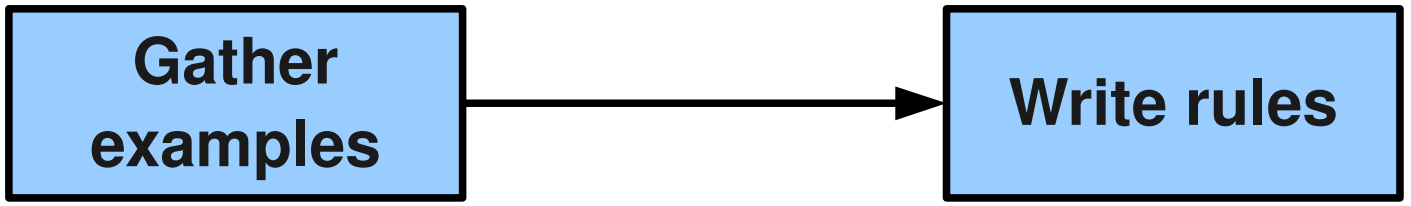
The process as agile development

- Recognise that requirements change
- Embrace that change
- Use it to drive development
- Developers and software engineers work alongside each other to understand requirements
- Early and iterative delivery
- Feedback to collect further requirements
- Reduces cost of annotation

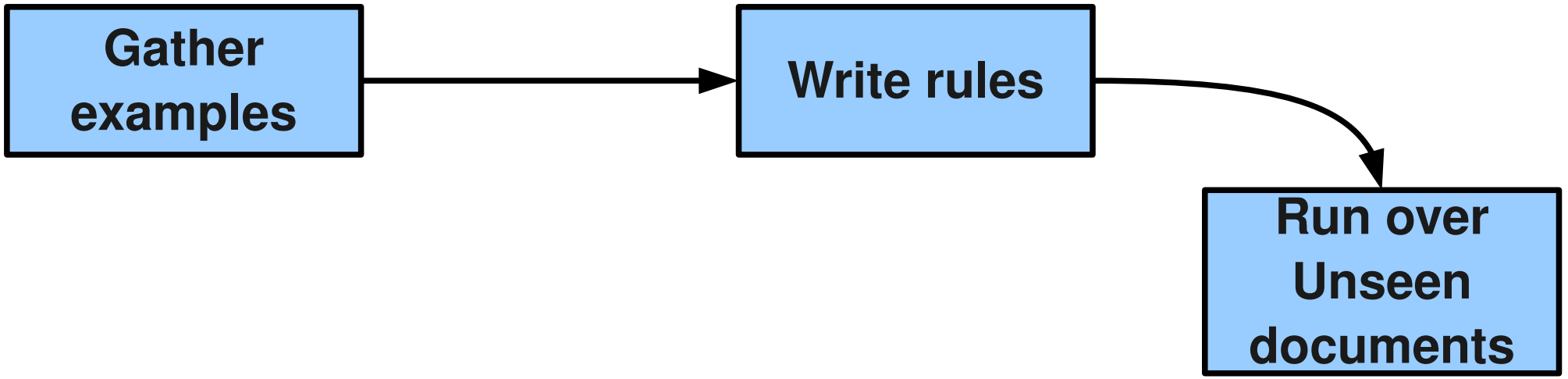
Annotation - a process

Gather examples

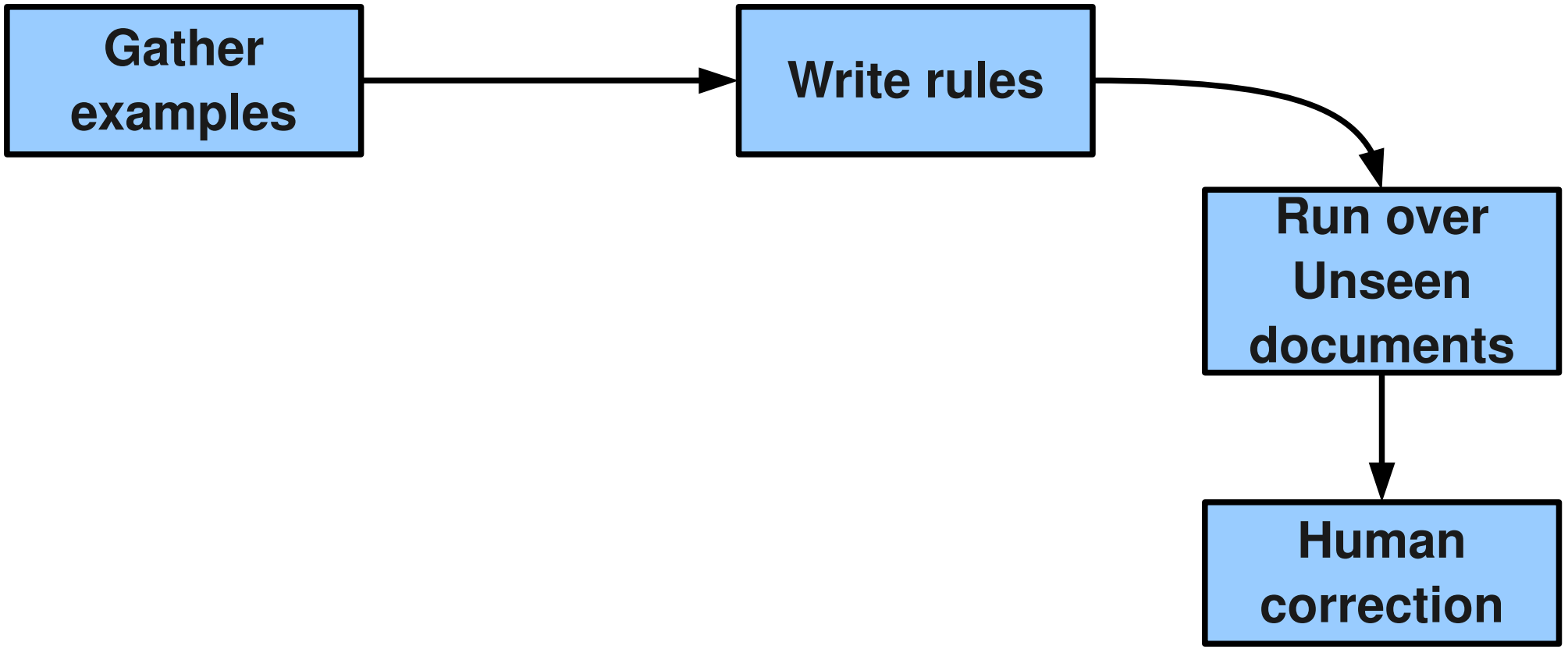
A process



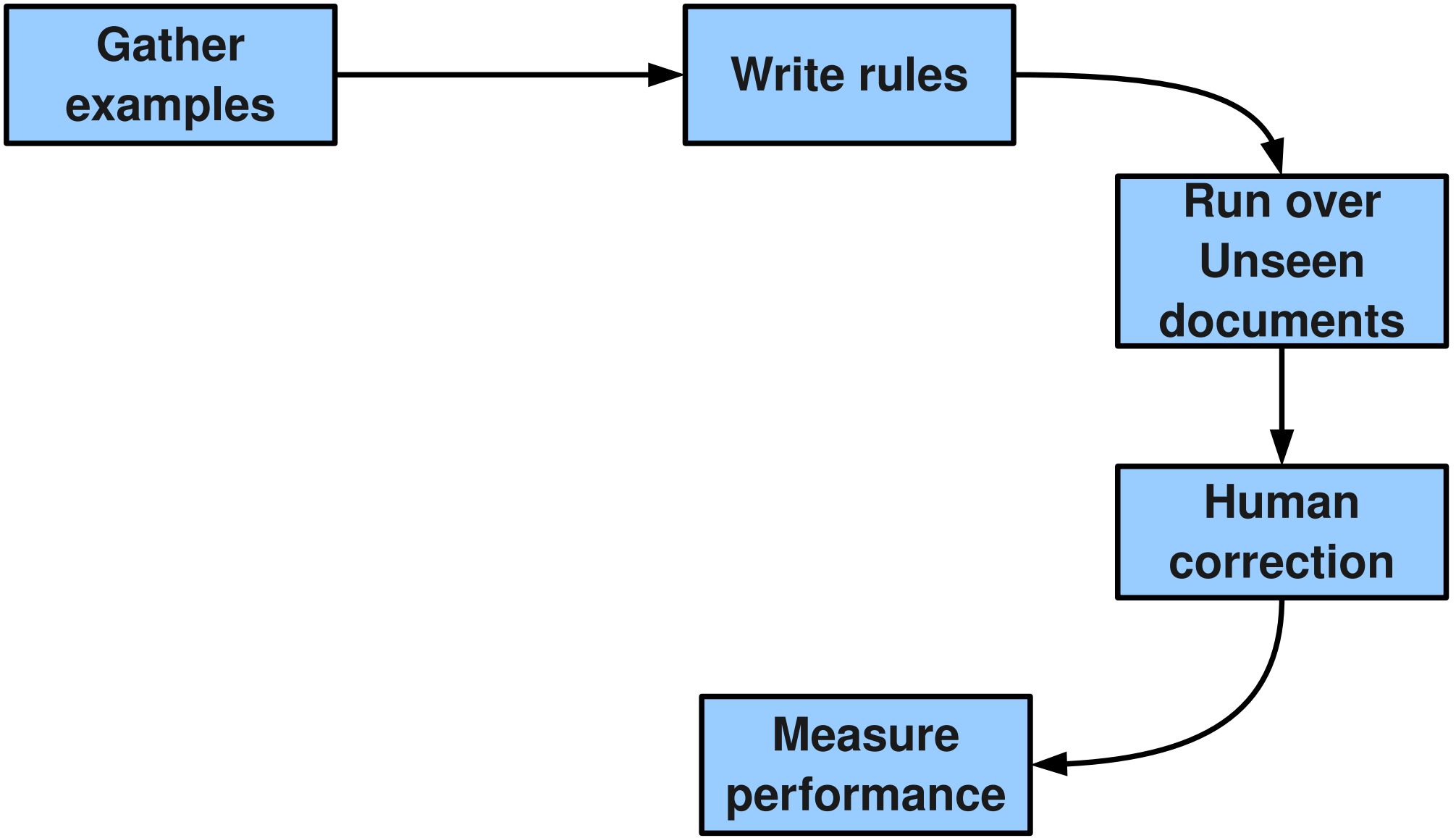
A process



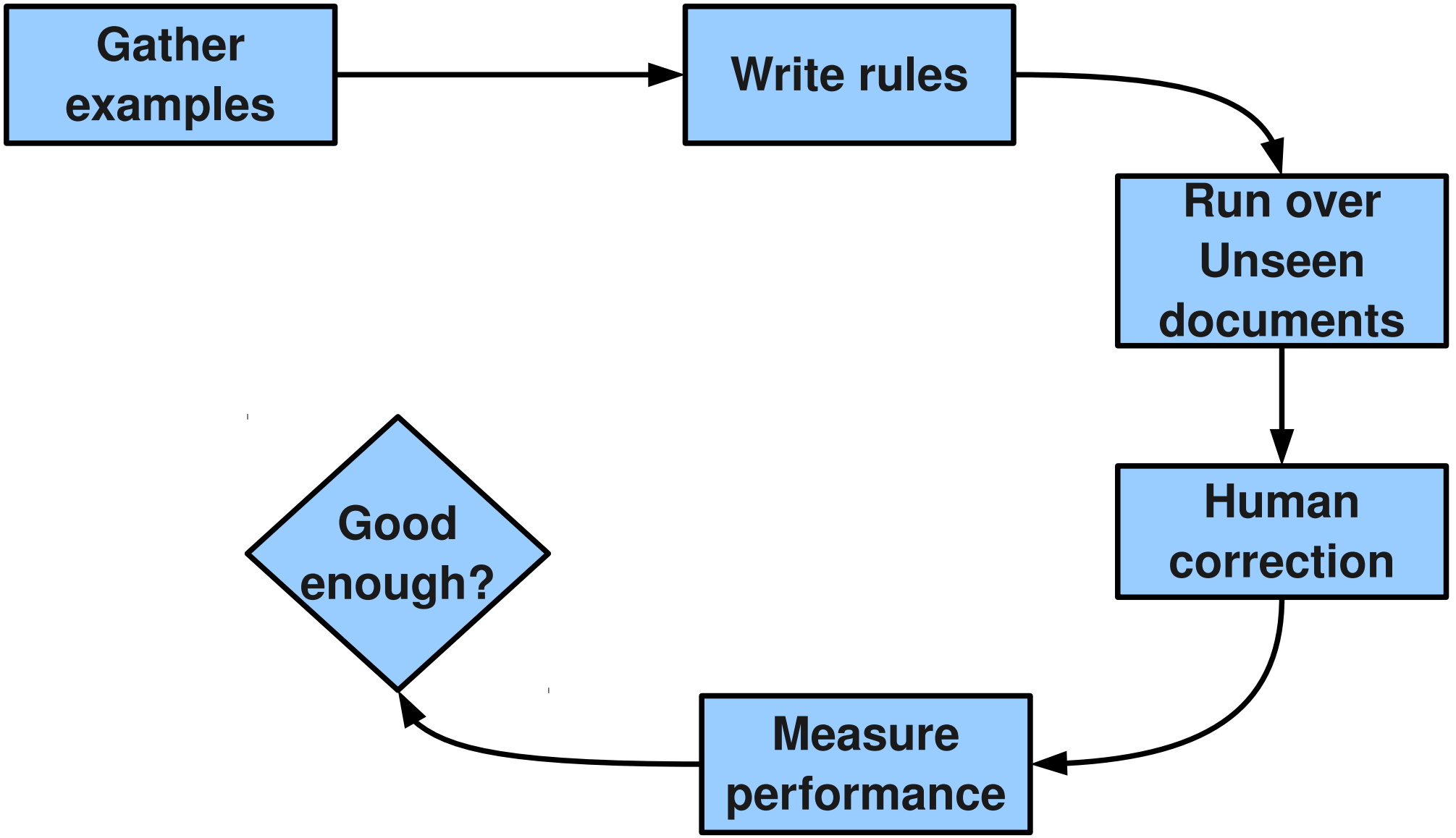
A process


Gather examples
Write rules
Run over Unseen documents
Human correction

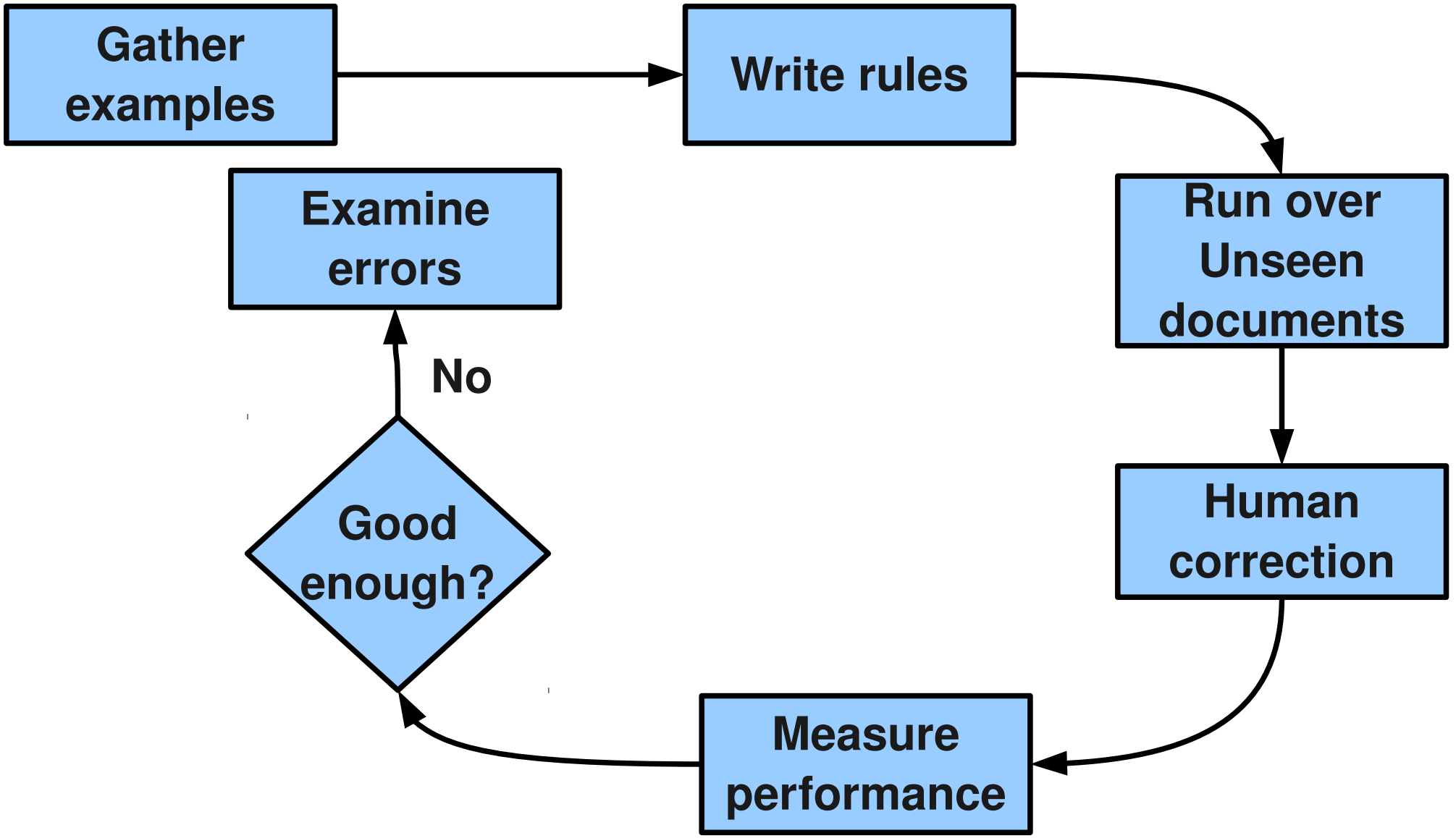
A process


Gather examples
Write rules
Run over Unseen documents
Human correction
Measure performance

A process


Gather examples
Write rules
Run over Unseen documents
Human correction
Measure performance
Good enough?

A process

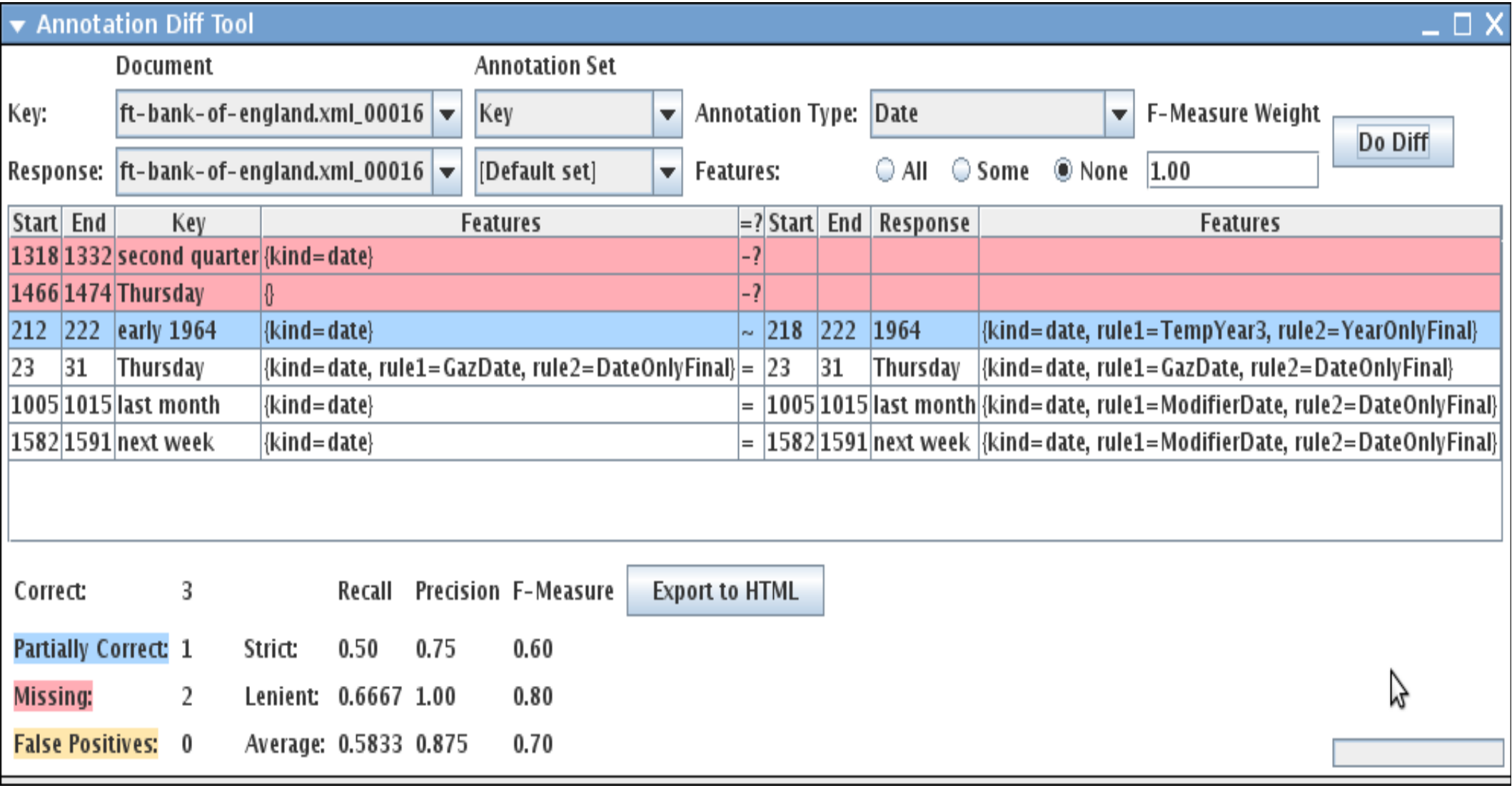

Gather examples
Write rules
Run over Unseen documents
Human correction
Measure performance
Good enough?
No
Examine errors

istics | Document statistics

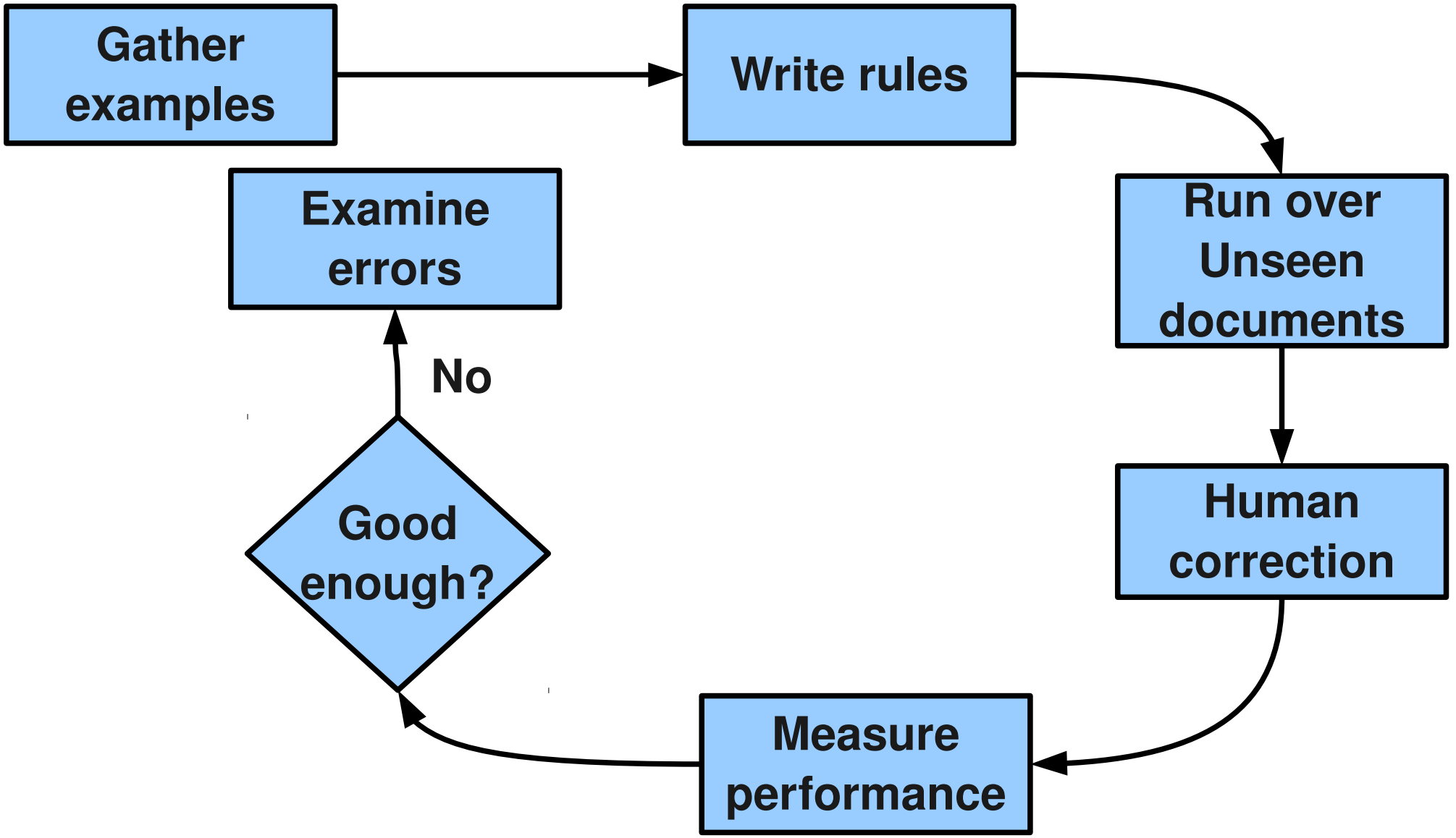
Document	Match	Only A	Only B	Overlap	Rec.B/A	Prec.B/A	F1-lenient
-eEVCo18_3.xml_00103	1	0	0	0	1.00	1.00	1.00
-eEVCo18_4.xml_00104	1	0	0	0	1.00	1.00	1.00
-eEVCo18_0.xml_00105	2	0	0	0	1.00	1.00	1.00
-eEVCo18_1.xml_00106	1	0	0	0	1.00	1.00	1.00
-eEVCo18_2.xml_00107	1	0	0	0	1.00	1.00	1.00
-eEVCo18_0.xml_00108	1	0	0	0	1.00	1.00	1.00
-eEVCo18_3.xml_00109	1	0	0	0	1.00	1.00	1.00
-eEVCo18_3.xml_0010A	2	0	0	0	1.00	1.00	1.00
-eEVCo18_4.xml_0010B	0	1	1	0	0.00	0.00	0.00
-eEVCo18_0.xml_0010C	1	0	0	0	1.00	1.00	1.00
-eEVCo18_1.xml_0010D	1	0	0	0	1.00	1.00	1.00
-eEVCo18_3.xml_0010E	2	0	0	0	1.00	1.00	1.00
-eEVCo18_0.xml_0010F	1	0	0	0	1.00	1.00	1.00
-eEVCo18_2.xml_00110	0	0	2	0	1.00	0.00	0.00
-eEVCo18_12.xml_00111	1	0	0	0	1.00	1.00	1.00
-eEVCo18_13.xml_00112	1	0	0	0	1.00	1.00	1.00
-eEVCo18_14.xml_00113	1	0	0	0	1.00	1.00	1.00
-eEVCo18_16.xml_00114	1	1	1	0	0.50	0.50	0.50
-eEVCo18_17.xml_00115	1	0	0	0	1.00	1.00	1.00

or | Initialisation Parameters | Corpus Quality Assurance

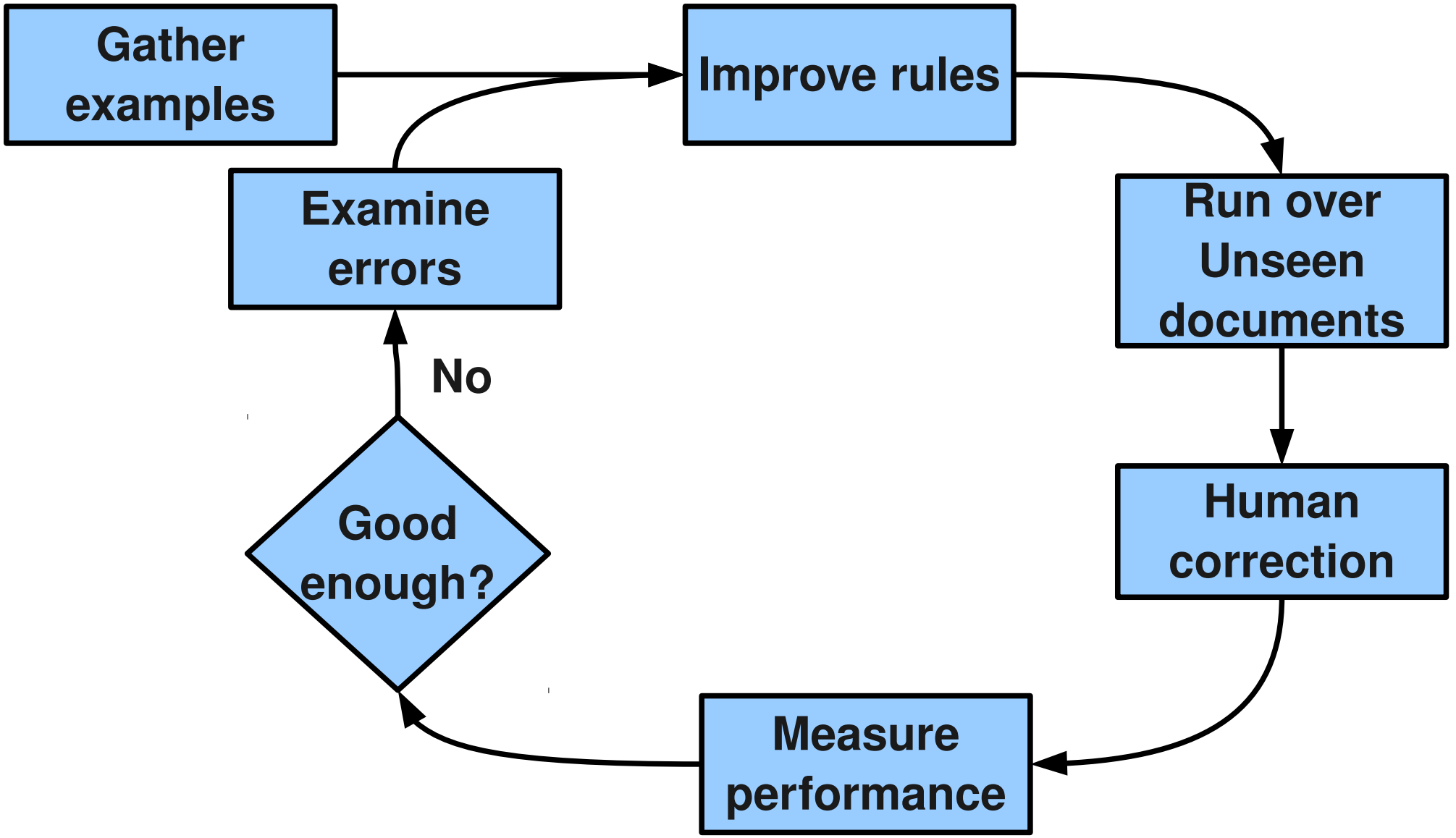
Annotation Diff



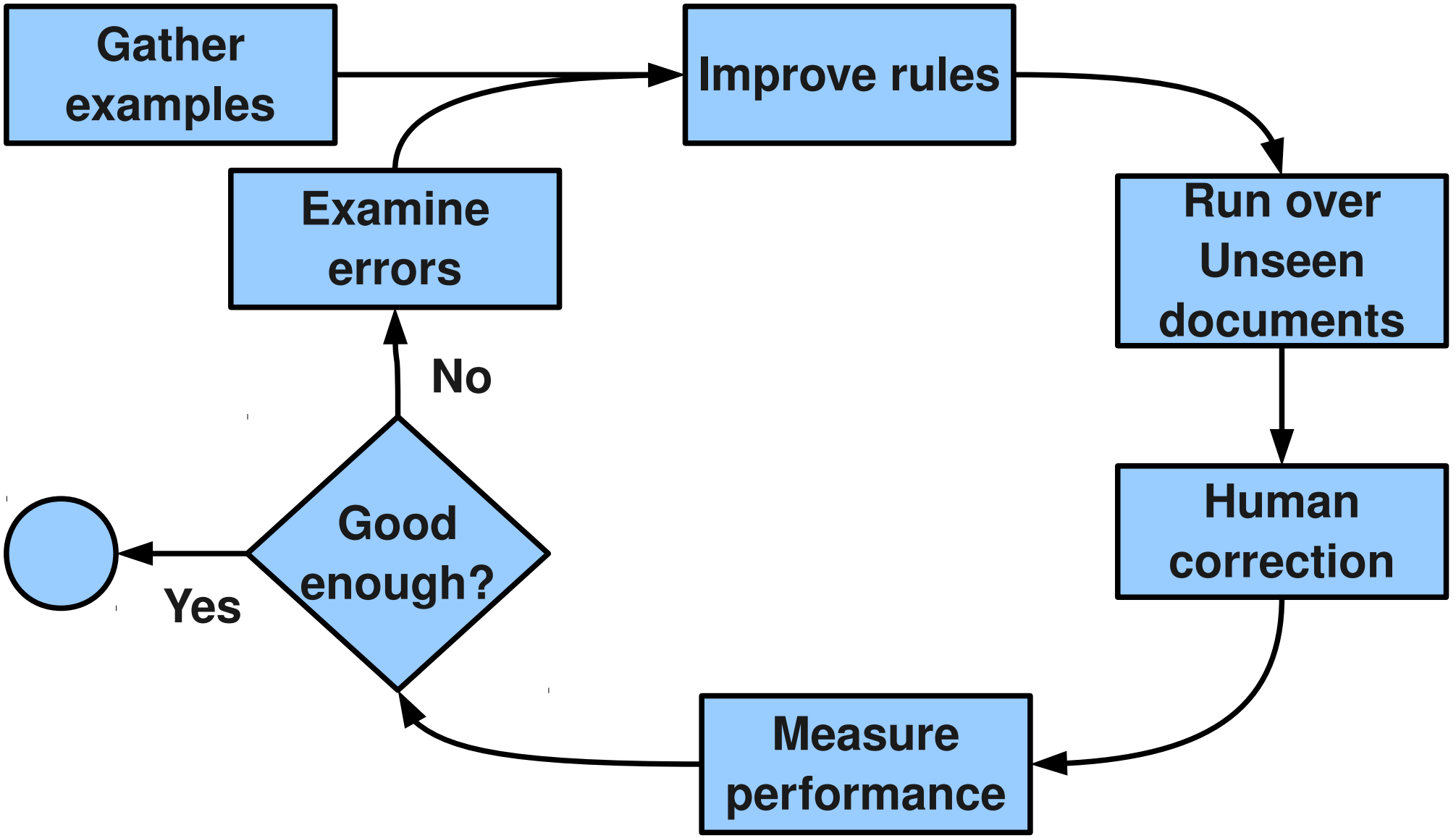
A process


Gather examples
Write rules
Run over Unseen documents
Human correction
Measure performance
Good enough?
No
Examine errors

A process


Gather examples
Improve rules
Run over Unseen documents
Human correction
Measure performance
Good enough?
No
Examine errors

A process


Gather examples
Improve rules
Run over Unseen documents
Human correction
Measure performance
Good enough?
Yes
No
Examine errors

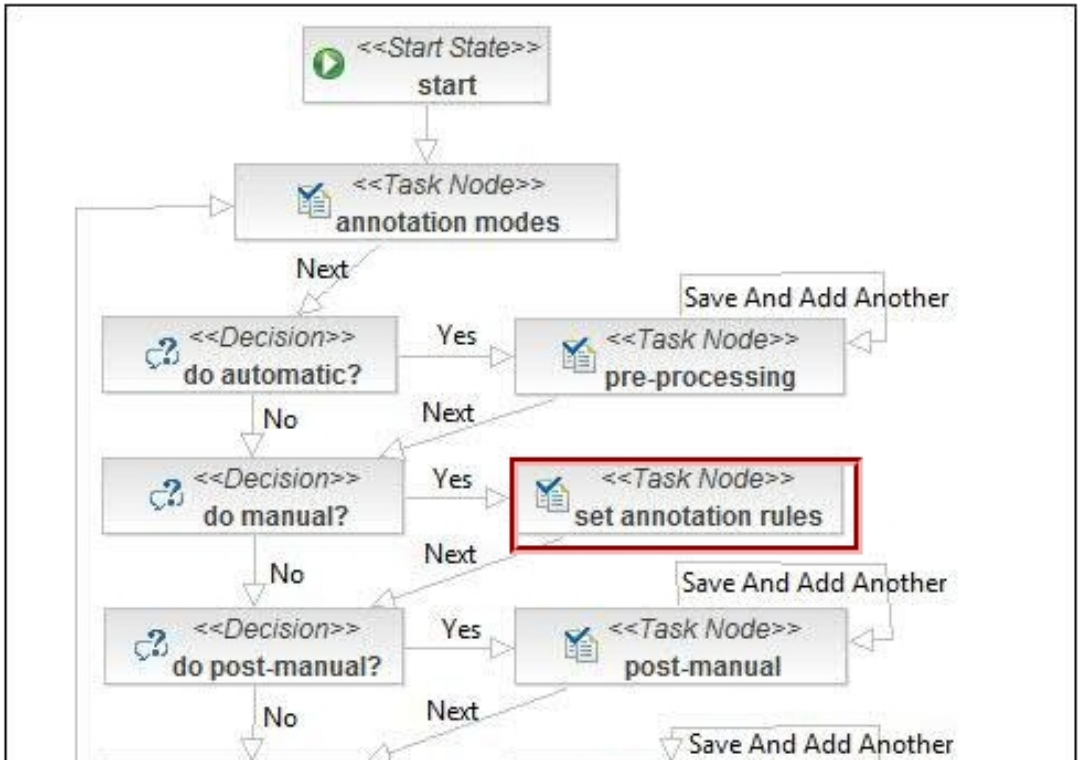
Supporting the process

- We need to train and implement the process
- We need tools to support this process
- Quality Assurance Tools
- Workflow: GATE Teamware
- Annotation pattern search: Mimir
- Coupling search and annotation: Khresmoi
  - Develop a pattern and run it over text
  - Human correction
  - Feedback

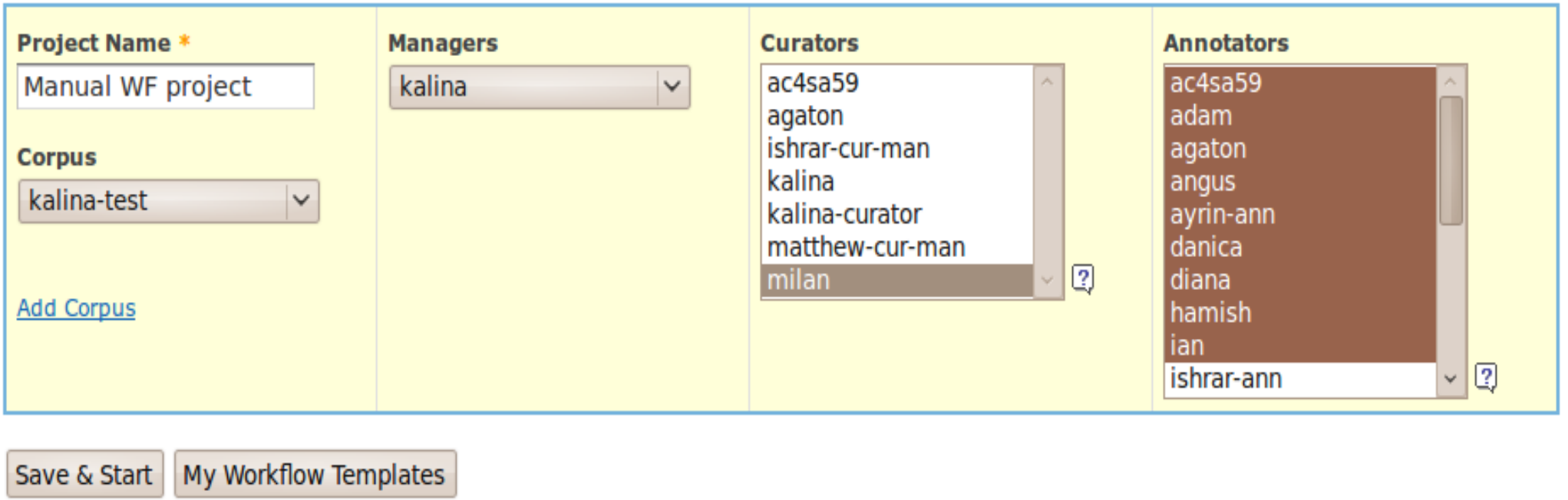
GATE Teamware: defining workflows

Annotators per Document:	2
Cancel Task Allowed:	☑
Anonymous Annotation:	☑
Annotation Schemas:*	OrganizationSchema.xm PersonSchema.xml AddressSchema.xml Non-FunctionalRequiren ConcordiaDataGroupsSc ConcordiaRequirementS
Add Schema:	Add Schema
Pre-Manual Service:*	None

Next Help Quit



GATE Teamware: managing projects



GATE Teamware: monitoring projects

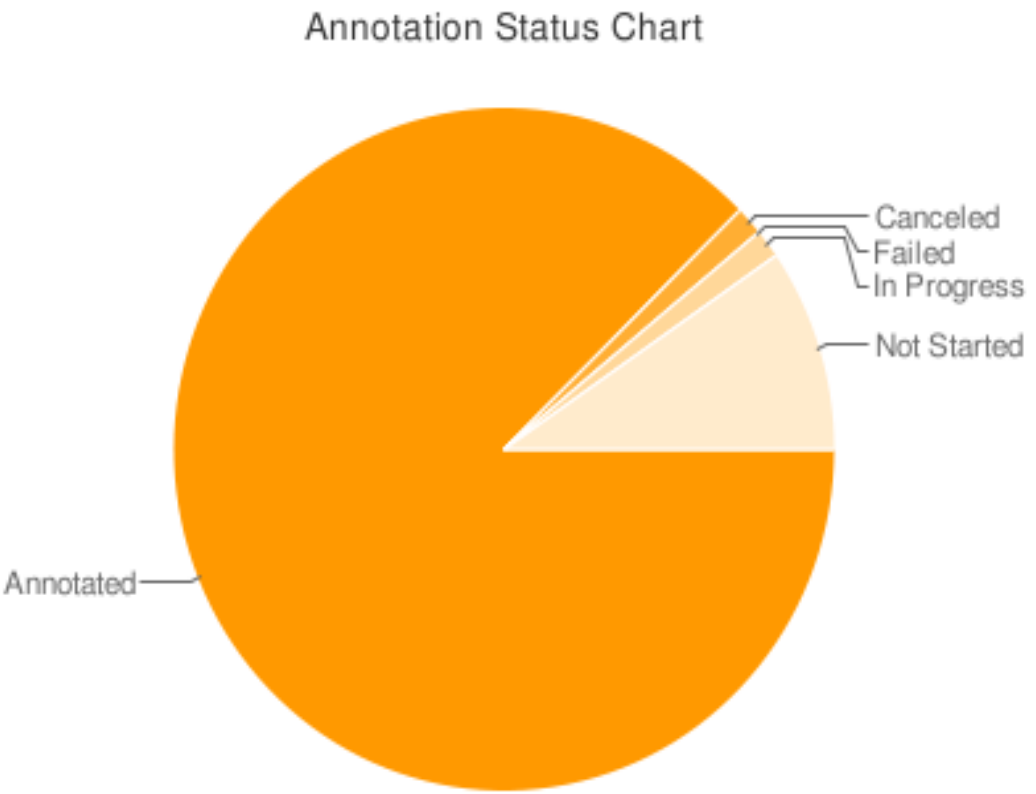
Process Monitoring: Annotation Status

Detailed View | Back to Project

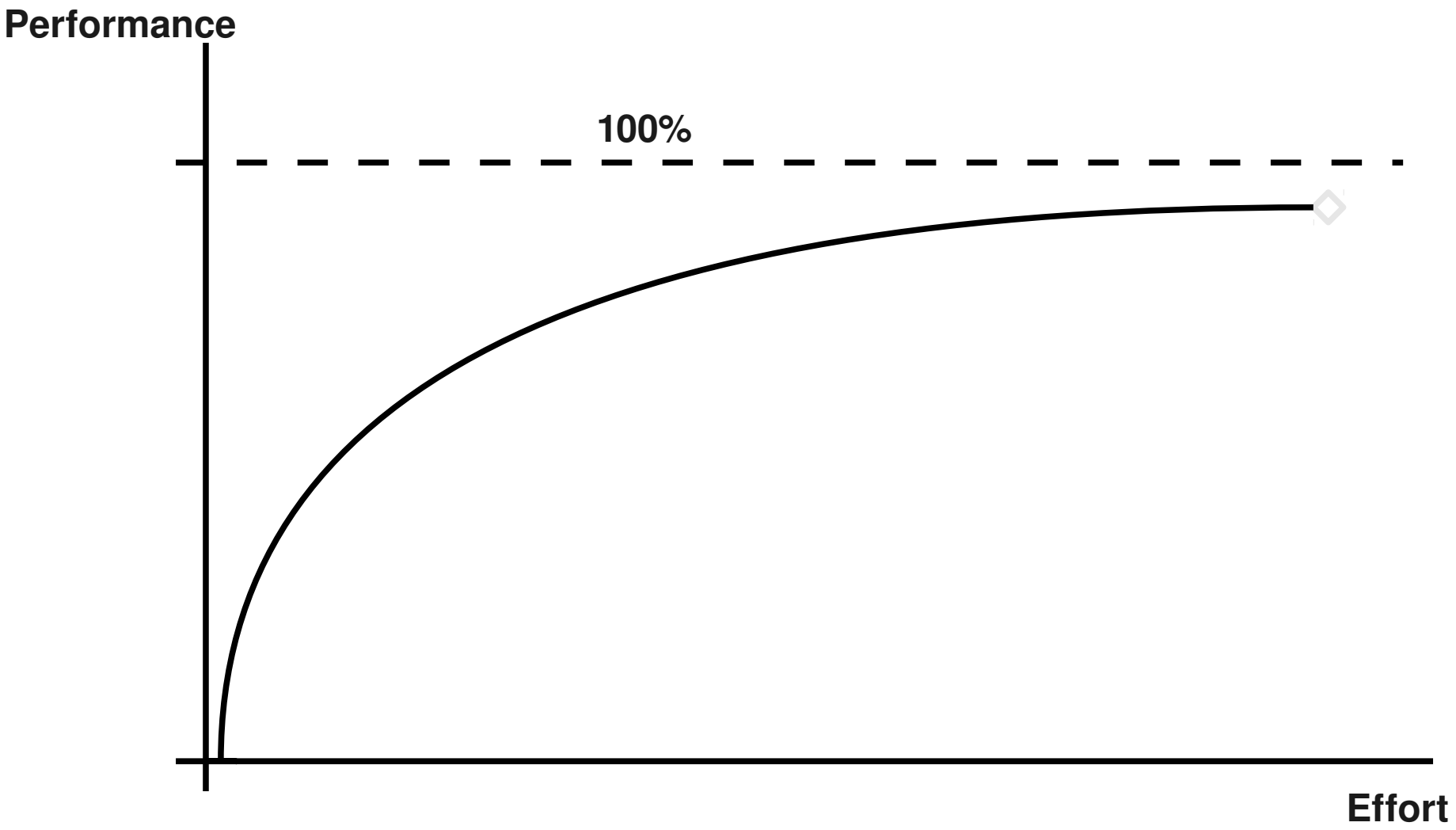
Status	#
Annotated	64
Canceled	1
Failed	0
In Progress	1
Not Started	7

Average Execution Time
627.609375

Detailed View | Back to Project



When is it good enough?



When is it good enough?

- Like disease
  - A small number of very common ones
  - Lots of rare ones
- For a straightforward use case,
  - 3 or 4 iterations
  - plateau at around 90%

Implications

- Ad-hoc annotation and search is as important an approach as generic annotation
- We need tools and processes to support this style of annotation
- An agile annotation process involves users, helps us to elicit their requirements, and reduces the cost of annotation