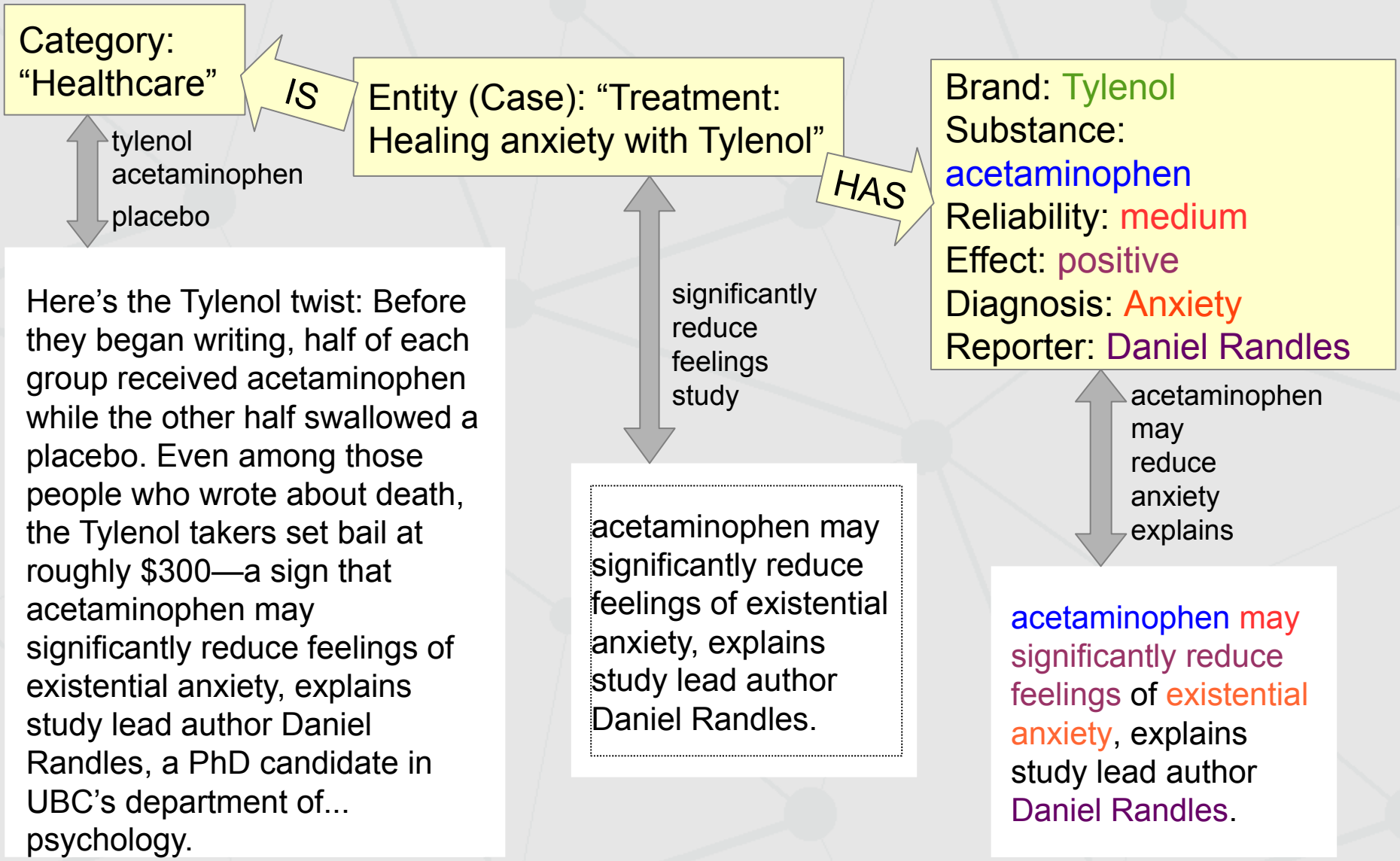
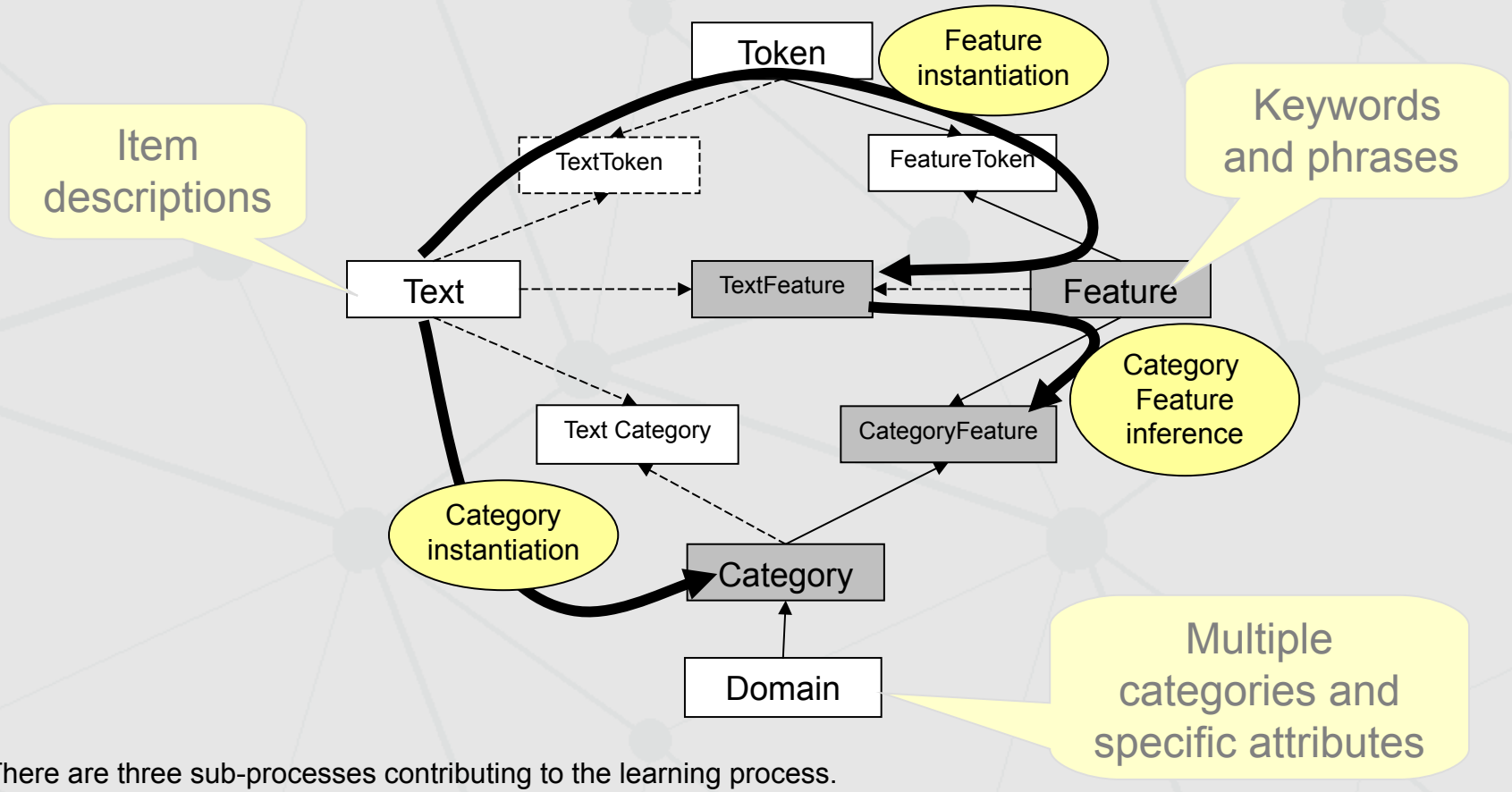
Automatic text classification and extraction of entities and their properties from the text

Anton Kolonin
Webstructor project
2015, SIBIRCON/SibMedInfo

Unified approach : Different cases

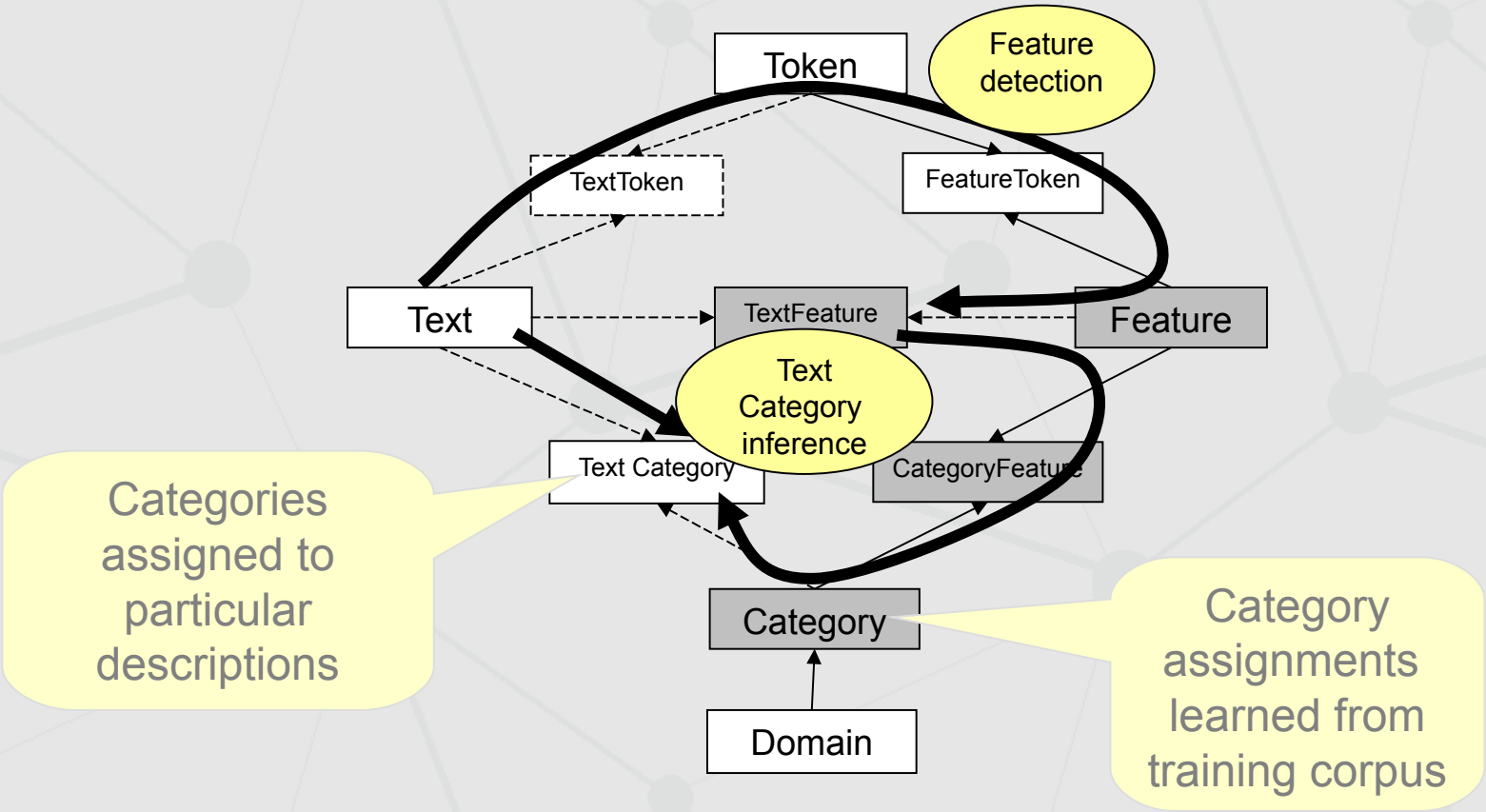


Classification with feature vectors : Training Process



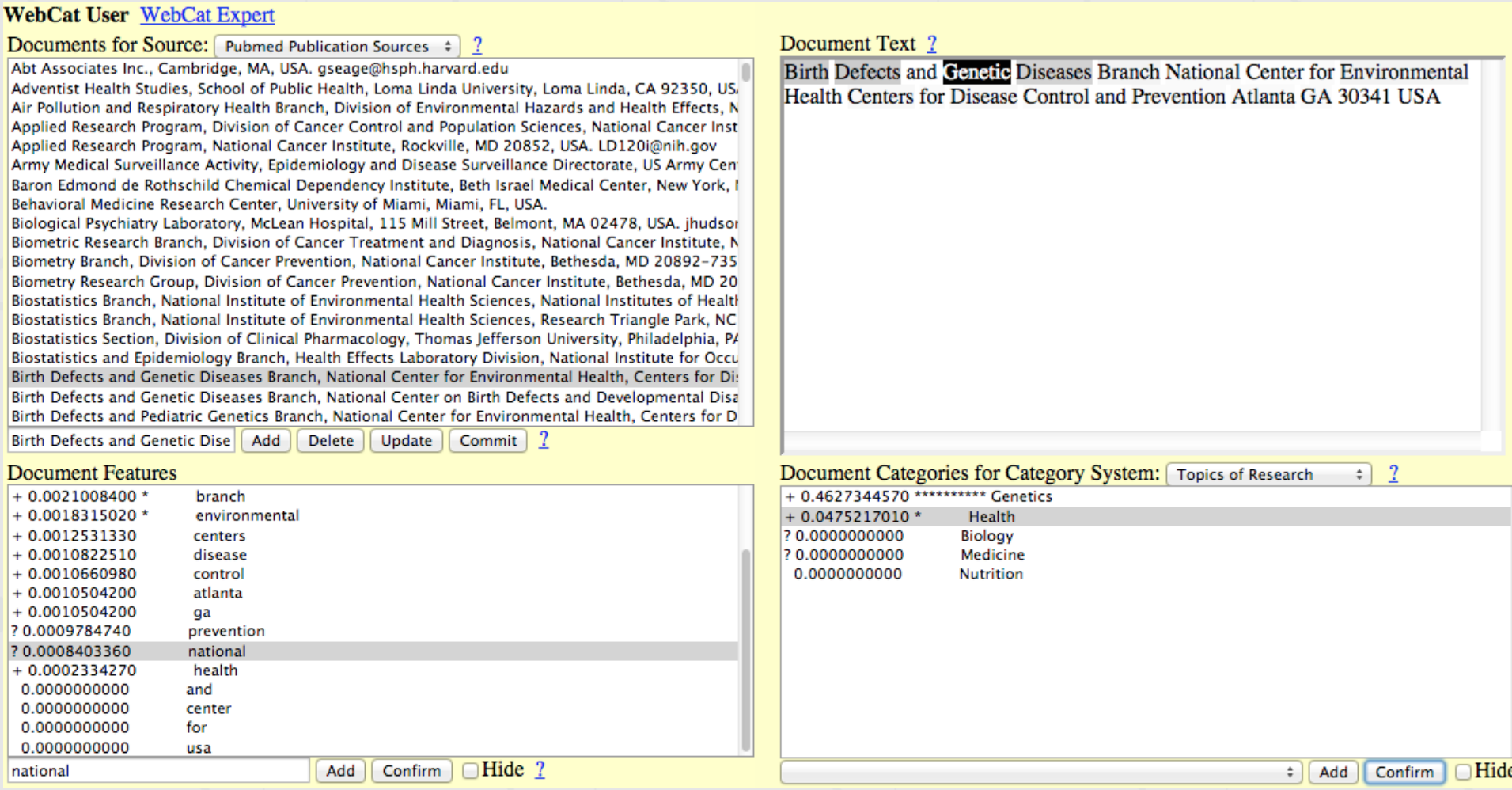
There are three sub-processes contributing to the learning process.
The **first process is Category instantiation** which takes the attributes defined for text in training corpus (either encoded in the text as tags or taken from respective database table fields) and creates categories for them, given the domain indicated by the attribute.
The **second process id Feature instantiation** which takes the text in training corpus and decomposes it into tokens and features accordingly to the parser, tokenizer and feature builder depending on the implementation.
The two processes above are independent, but they precede **the third process which is Category Feature inference**. It employs statistics to infer the relevance of features encountered in the texts to the categories associated with those texts.

Classification with feature vectors : Recognition process

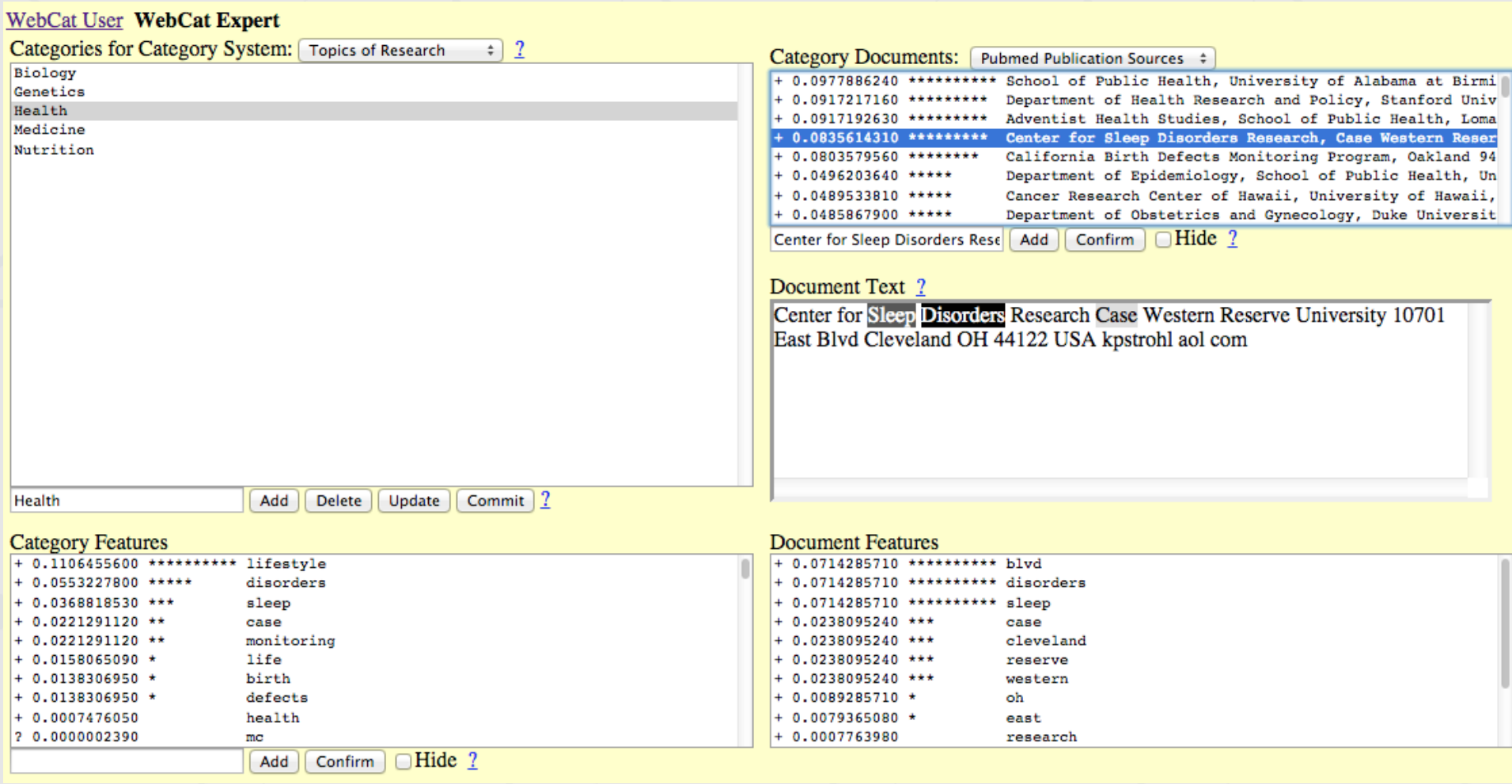


There are two sub-processes contributing to the rule applying process and the following process flow diagram depicts the dependency between the sub-processes and the data.
The **first process is Feature detection** which takes the text in novel data and decomposes it into tokens and features accordingly to the parser, tokenizer and feature builder depending on the implementation. This process is similar to Feature instantiation in the course of learning, but the key difference is that only the features instantiated earlier in the course of learning can be detected, no new features are instantiated.
The **second process is Text Category inference**. It employs statistics to infer the relevance of texts to the categories associated with those texts through the features detected in the texts and learned for those categories.

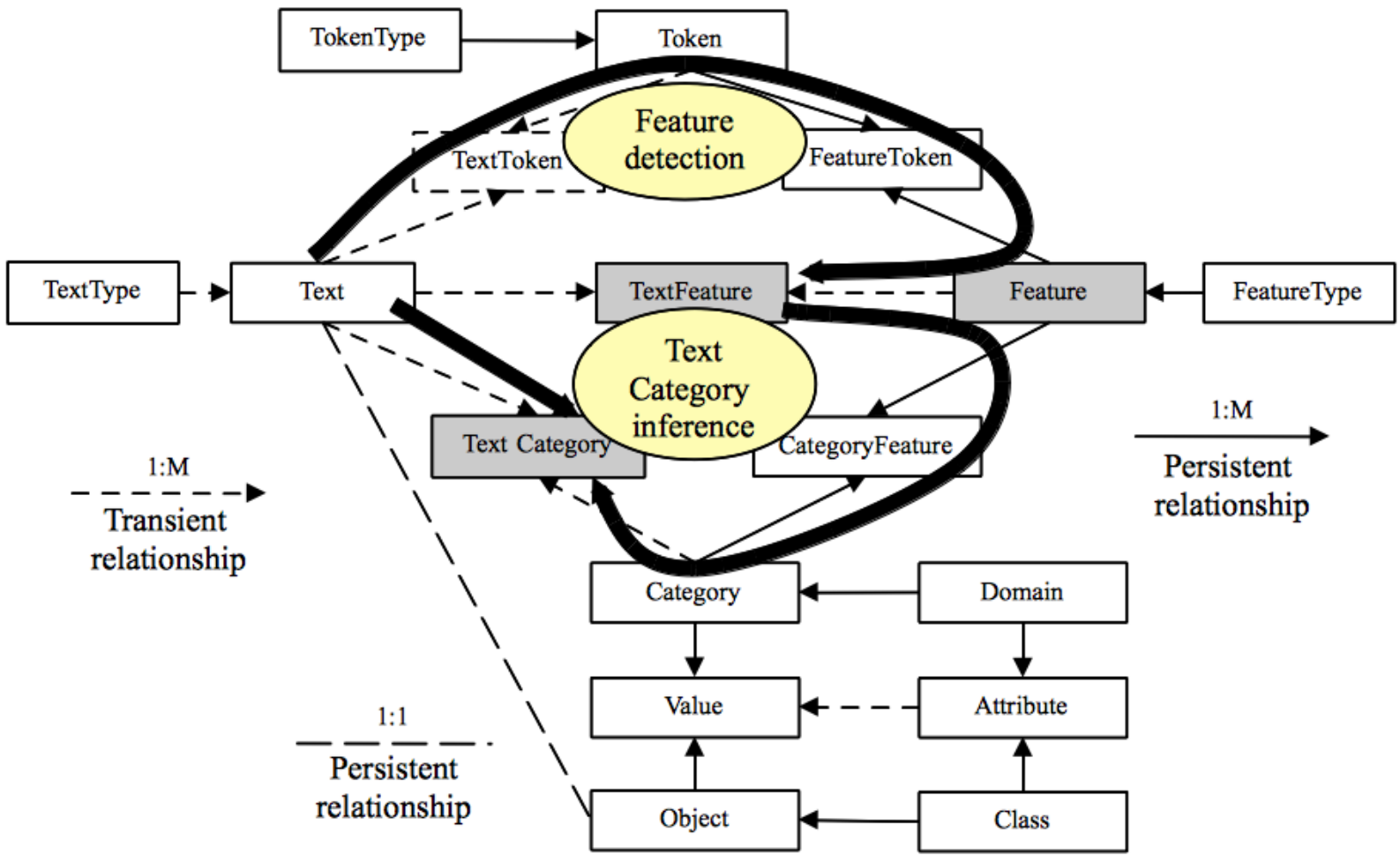
Webcat: Plain user interface



Webcat: Expert user interface



Webcat: Extracting entity properties



Webcat: Complex patterns and rules

Sparse N-gram:
"tylenol significantly reduce feelings of existential anxiety" = "tylenol ... reduce ... anxiety"

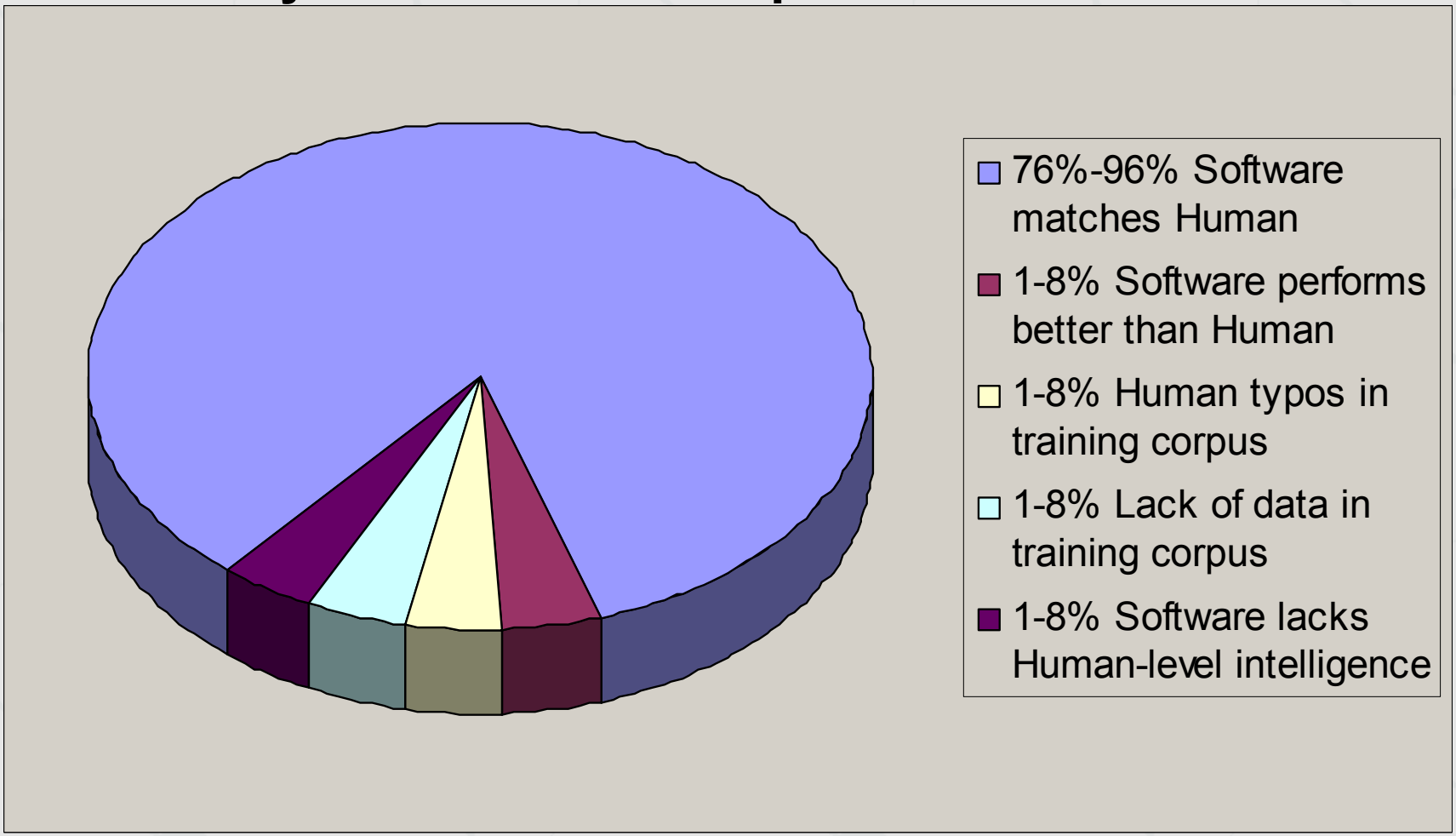
Priority on order:
"reduce ... feelings" *is more important than* "reduce AND feelings"

Boolean ranking:
"acetaminophen AND tylenol" *is more important than* "placebo", regardless of statistics

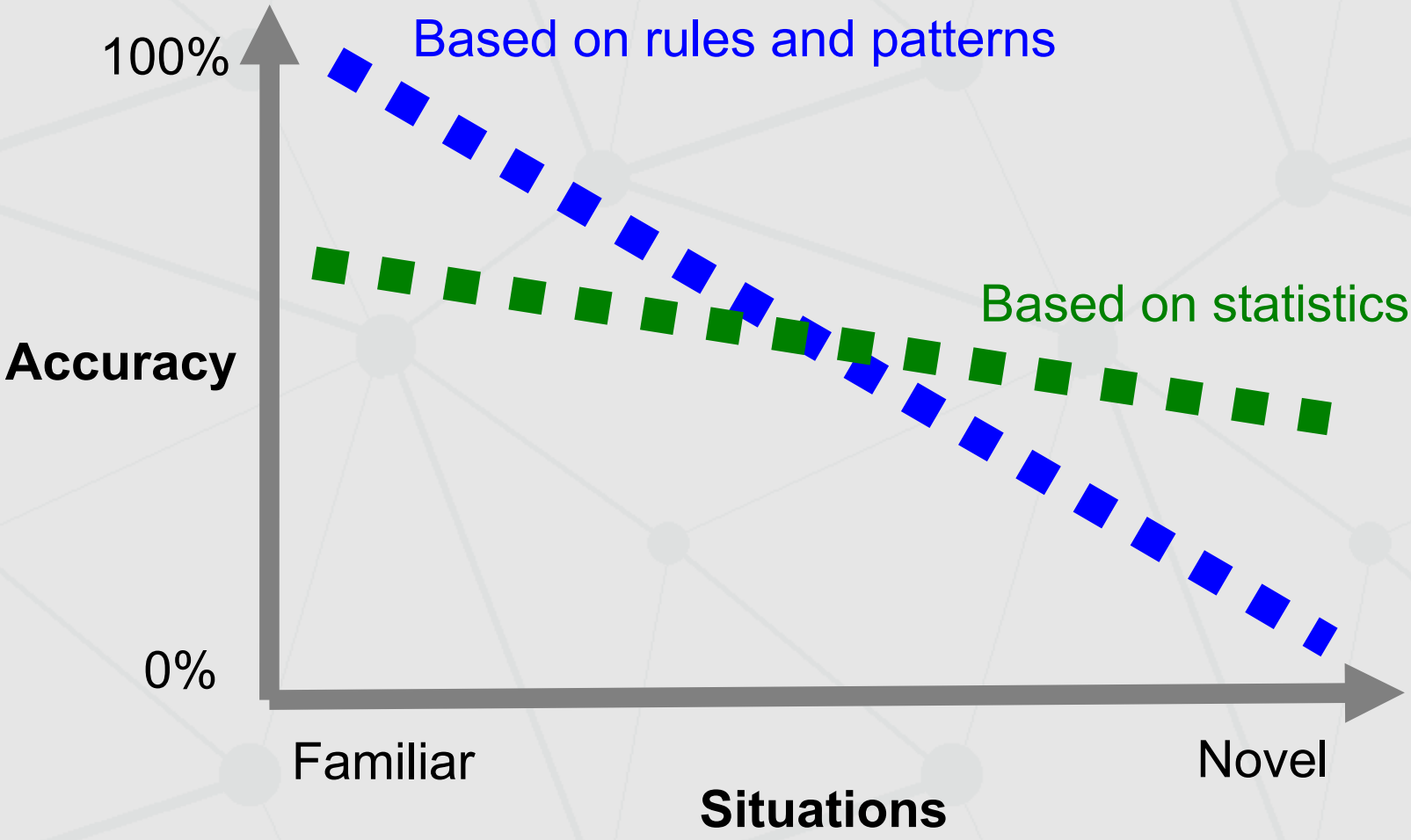
Contextual scoping:
"tylenol" *implies that* "may" *is* reliability measure, *not* month of the year

Compression of vector space:
"disease OR illness" is much faster than "disease OR illness OR ail OR blast OR sick" (even if little bit less accurate)

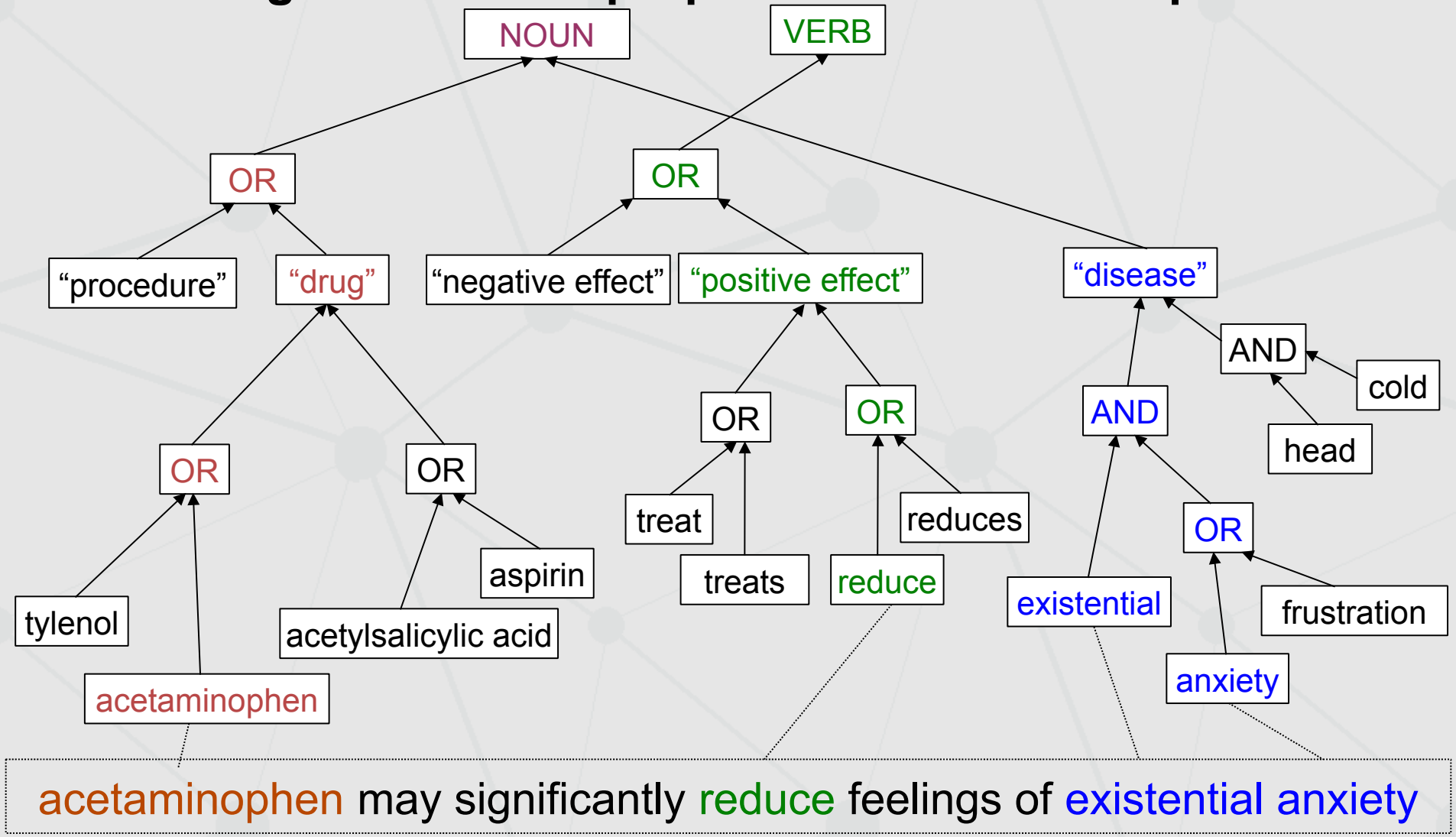
Accuracy : Human vs. Computer : Sources of errors



Statistical «fuzzy» learning vs. «rigid» patterns and rules



Finding entities with properties: Hierarchical patterns



Hierarchical patterns: Definition

<pattern> := <token> | <regexp> | <variable> | <set>
<set> := <conjunctive-set> | <N-gram> | <syn-set>
<conjunctive-set> := (<pattern> *)
<N-gram> := [<pattern> *]
<syn-set> := { <pattern> * }

Examples

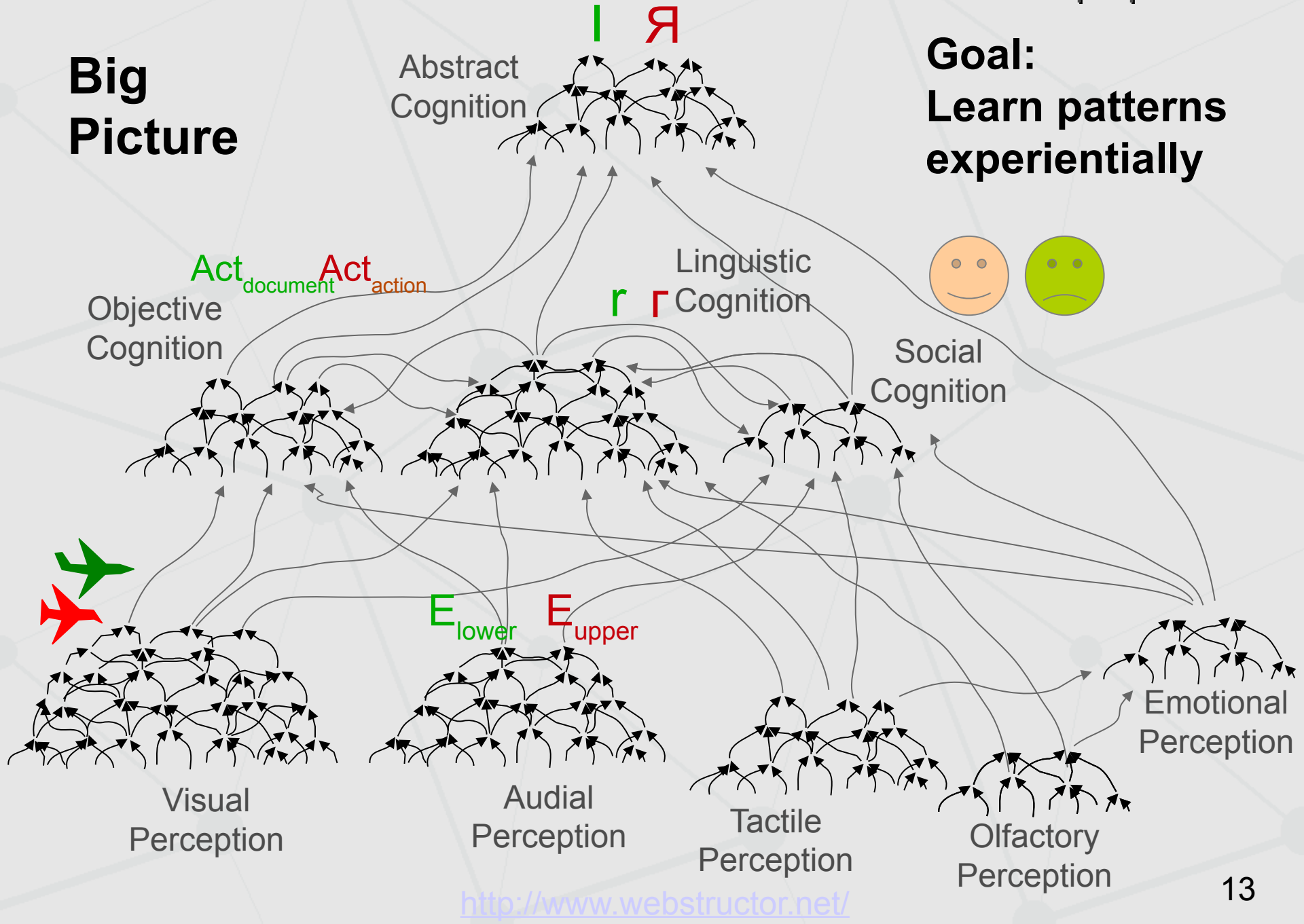
{[$description catheter] [$coating coating] [$inner-diameter
{diameter inner-diameter}] [$tip tip] [$pattern pattern]}
X

Convey Guiding Catheter. Unique hydrophilic coating.
Small atraumatic soft tip. Ultra-thin 1 × 2 flat wire braid pattern
=

{ coating : 'hydrophillic', description : 'convey guiding',
pattern : 'ultra-thin 1 × 2 flat wire braid', tip : soft }

Big Picture

Goal:
Learn patterns experientially



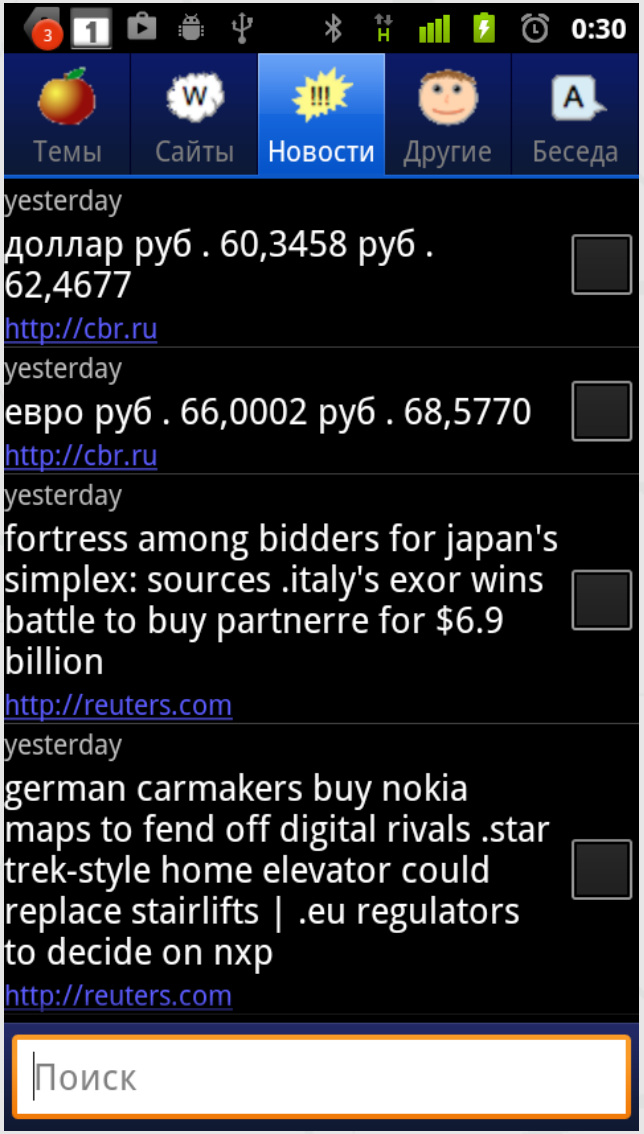
Part-of-speech tagging?
Need full semantic context to be precise

Какой (свойство зрения)?
Какой (состояние опьянения)?
Кто (профессия)?
Кто (имя, кличка)?
С чем?
Чем?
Что делал?
Где?
Как?

Косой косой косарь Косой с косой косой косил на косе косо

Automatic text classification and extraction of entities and their properties

Current implementation



Thank you for attention!

Anton Kolonin
Webstructor project
2015, SIBIRCON/SibMedInfo