

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

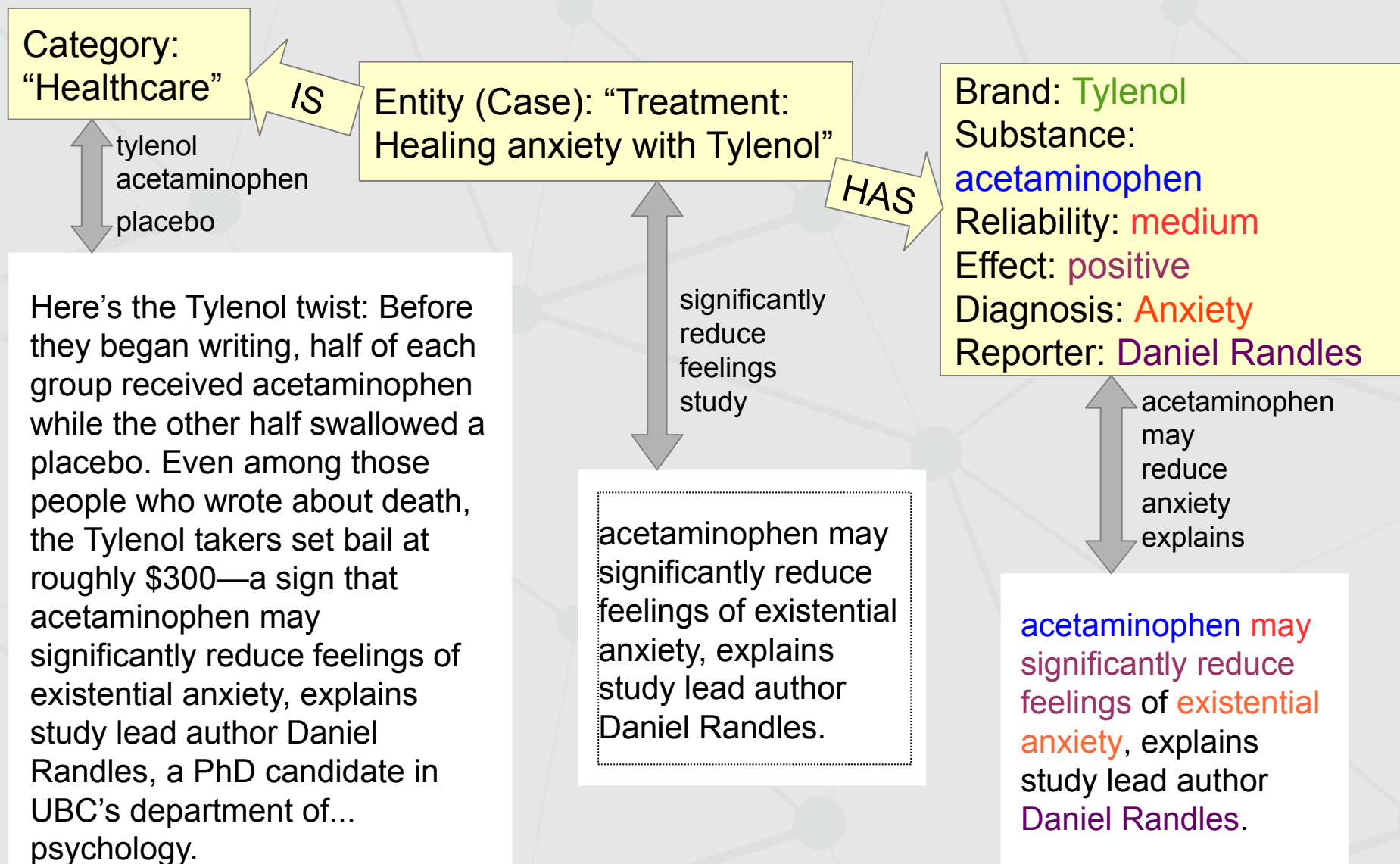
Anton Kolonin

[Webstructor project](#)

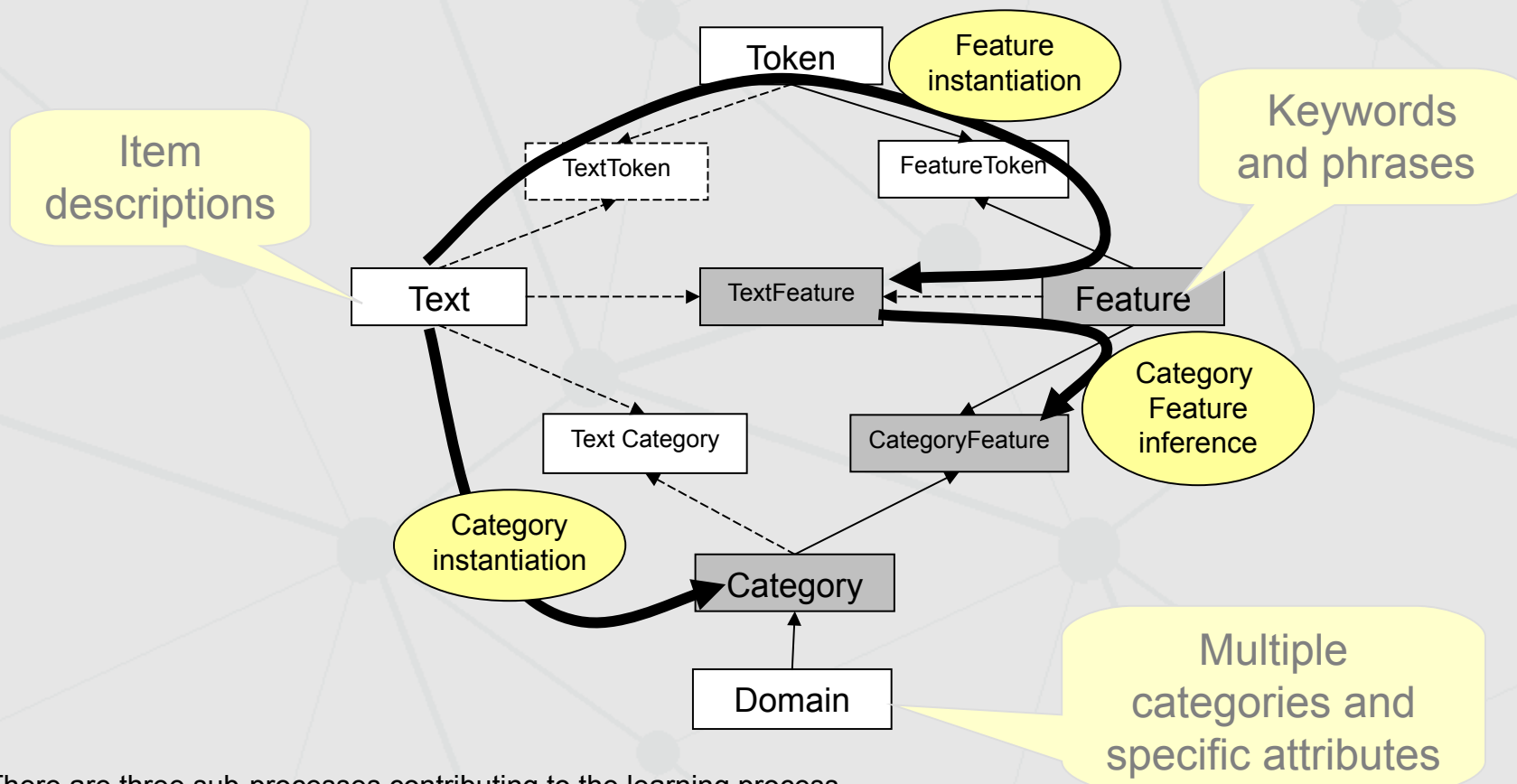
2015, SIBIRCON/SibMedInfo

<http://www.webstructor.net/>

## Unified approach : Different cases



## Classification with feature vectors : Training Process



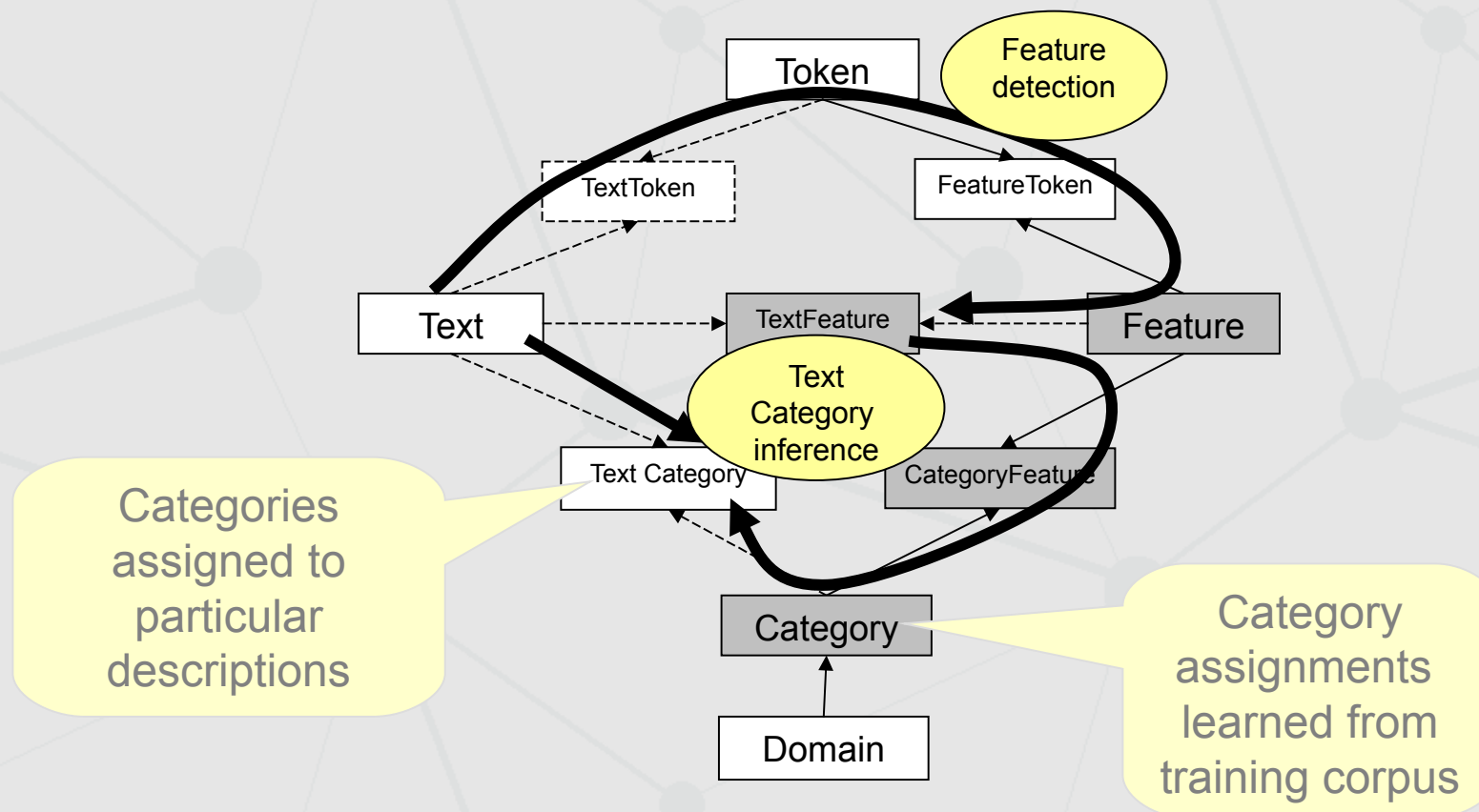
There are three sub-processes contributing to the learning process.

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 is 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.

# Automatic text classification and extraction of entities and their properties

## Webcat: Plain user interface

WebCat User [WebCat Expert](#)

Documents for Source:  ?

Abt Associates Inc., Cambridge, MA, USA. gseage@hsph.harvard.edu  
Adventist Health Studies, School of Public Health, Loma Linda University, Loma Linda, CA 92350, US.  
Air Pollution and Respiratory Health Branch, Division of Environmental Hazards and Health Effects, N  
Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Inst  
Applied Research Program, National Cancer Institute, Rockville, MD 20852, USA. LD120i@nih.gov  
Army Medical Surveillance Activity, Epidemiology and Disease Surveillance Directorate, US Army Cen  
Baron Edmond de Rothschild Chemical Dependency Institute, Beth Israel Medical Center, New York, I  
Behavioral Medicine Research Center, University of Miami, Miami, FL, USA.  
Biological Psychiatry Laboratory, McLean Hospital, 115 Mill Street, Belmont, MA 02478, USA. jhudso  
Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, N  
Biometry Branch, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD 20892-735  
Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD 20  
Biostatistics Branch, National Institute of Environmental Health Sciences, National Institutes of Health  
Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC  
Biostatistics Section, Division of Clinical Pharmacology, Thomas Jefferson University, Philadelphia, PA  
Biostatistics and Epidemiology Branch, Health Effects Laboratory Division, National Institute for Occu  
Birth Defects and Genetic Diseases Branch, National Center for Environmental Health, Centers for Di  
Birth Defects and Genetic Diseases Branch, National Center on Birth Defects and Developmental Disa  
Birth Defects and Pediatric Genetics Branch, National Center for Environmental Health, Centers for D  
Birth Defects and Genetic Dise     ?

Document Features

+ 0.0021008400 *	branch
+ 0.0018315020 *	environmental
+ 0.0012531330	centers
+ 0.0010822510	disease
+ 0.0010660980	control
+ 0.0010504200	atlanta
+ 0.0010504200	ga
? 0.0009784740	prevention
? 0.0008403360	national
+ 0.0002334270	health
0.0000000000	and
0.0000000000	center
0.0000000000	for
0.0000000000	usa

national    Hide ?

Document Text ?

Birth Defects and Genetic Diseases Branch National Center for Environmental Health Centers for Disease Control and Prevention Atlanta GA 30341 USA

Document Categories for Category System:  ?

+ 0.4627344570 *****	Genetics
+ 0.0475217010 *	Health
? 0.0000000000	Biology
? 0.0000000000	Medicine
0.0000000000	Nutrition

Hide

# Automatic text classification and extraction of entities and their properties

## Webcat: Expert user interface

WebCat User WebCat Expert

Categories for Category System: Topics of Research ?

Biology  
Genetics  
Health  
Medicine  
Nutrition

Health     ?

Category Features

+ 0.1106455600 \*\*\*\*\* lifestyle  
+ 0.0553227800 \*\*\*\*\* disorders  
+ 0.0368818530 \*\*\* sleep  
+ 0.0221291120 \*\* case  
+ 0.0221291120 \*\* monitoring  
+ 0.0158065090 \* life  
+ 0.0138306950 \* birth  
+ 0.0138306950 \* defects  
+ 0.0007476050 health  
? 0.0000002390 mc

Hide ?

Category Documents: Pubmed Publication Sources ?

+ 0.0977886240 \*\*\*\*\* School of Public Health, University of Alabama at Birmi  
+ 0.0917217160 \*\*\*\*\* Department of Health Research and Policy, Stanford Univ  
+ 0.0917192630 \*\*\*\*\* Adventist Health Studies, School of Public Health, Loma  
+ 0.0835614310 \*\*\*\*\* Center for Sleep Disorders Research, Case Western Reser  
+ 0.0803579560 \*\*\*\*\* California Birth Defects Monitoring Program, Oakland 94  
+ 0.0496203640 \*\*\*\*\* Department of Epidemiology, School of Public Health, Un  
+ 0.0489533810 \*\*\*\*\* Cancer Research Center of Hawaii, University of Hawaii,  
+ 0.0485867900 \*\*\*\*\* Department of Obstetrics and Gynecology, Duke Universit

Center for Sleep Disorders Rese    Hide ?

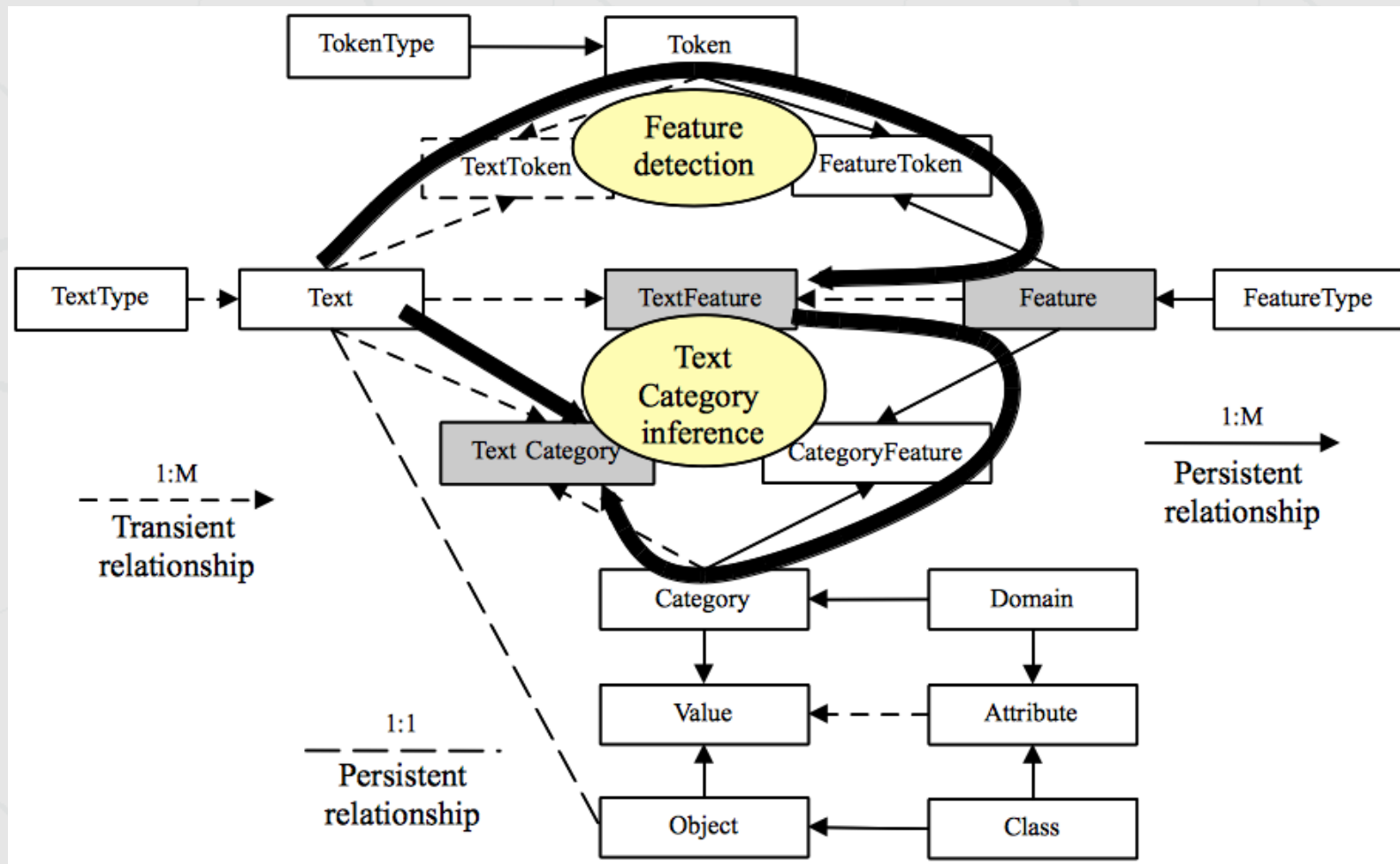
Document Text ?

Center for Sleep Disorders Research Case Western Reserve University 10701  
East Blvd Cleveland OH 44122 USA kpstrohl aol com

Document Features

+ 0.0714285710 \*\*\*\*\* blvd  
+ 0.0714285710 \*\*\*\*\* disorders  
+ 0.0714285710 \*\*\*\*\* sleep  
+ 0.0238095240 \*\*\* case  
+ 0.0238095240 \*\*\* cleveland  
+ 0.0238095240 \*\*\* reserve  
+ 0.0238095240 \*\*\* western  
+ 0.0089285710 \* oh  
+ 0.0079365080 \* east  
+ 0.0007763980 research

## 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:

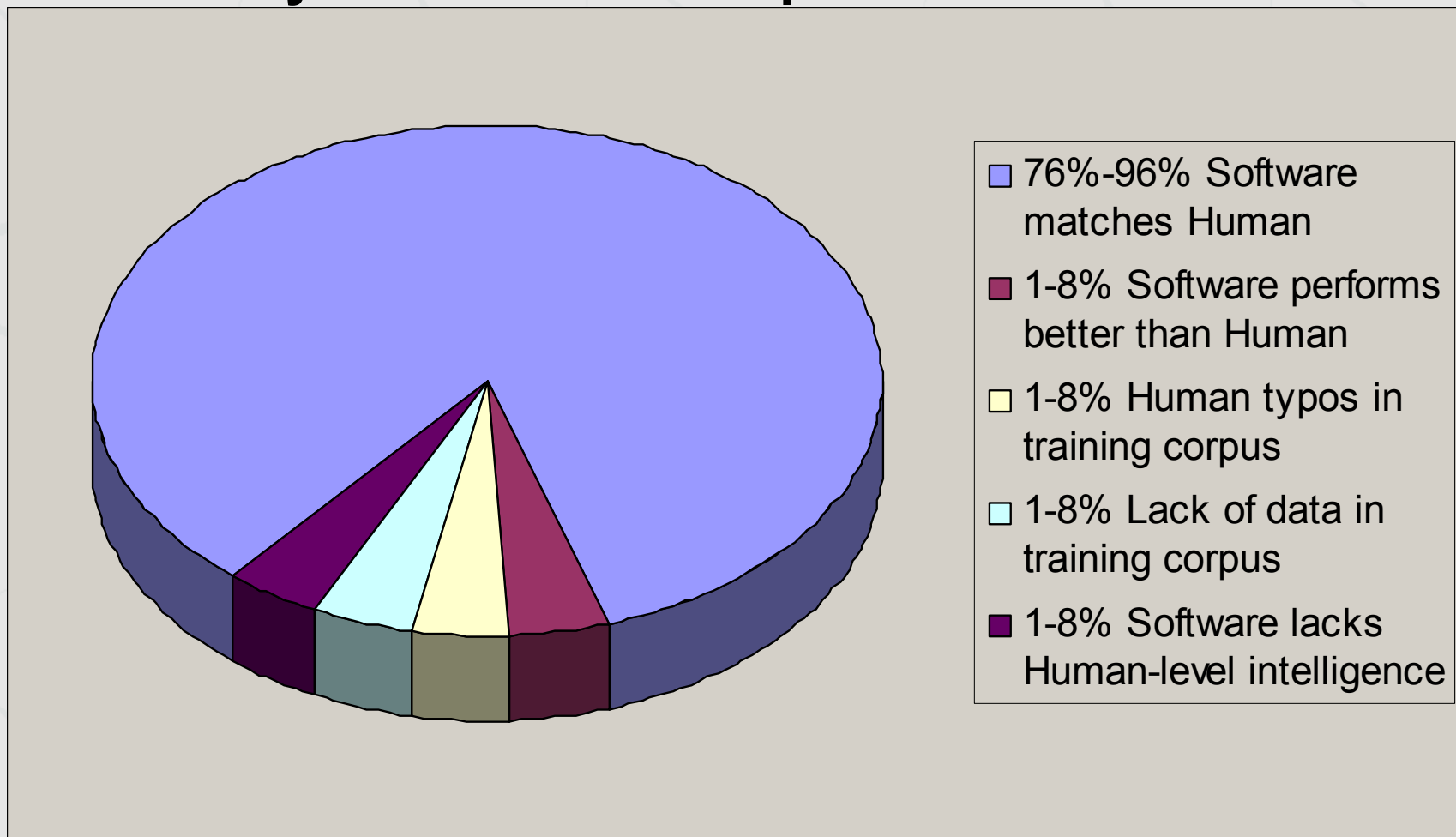
“tylenol” *implies that* “may” *is* reliability measure, *not* month of the year

### Compression of vector space:

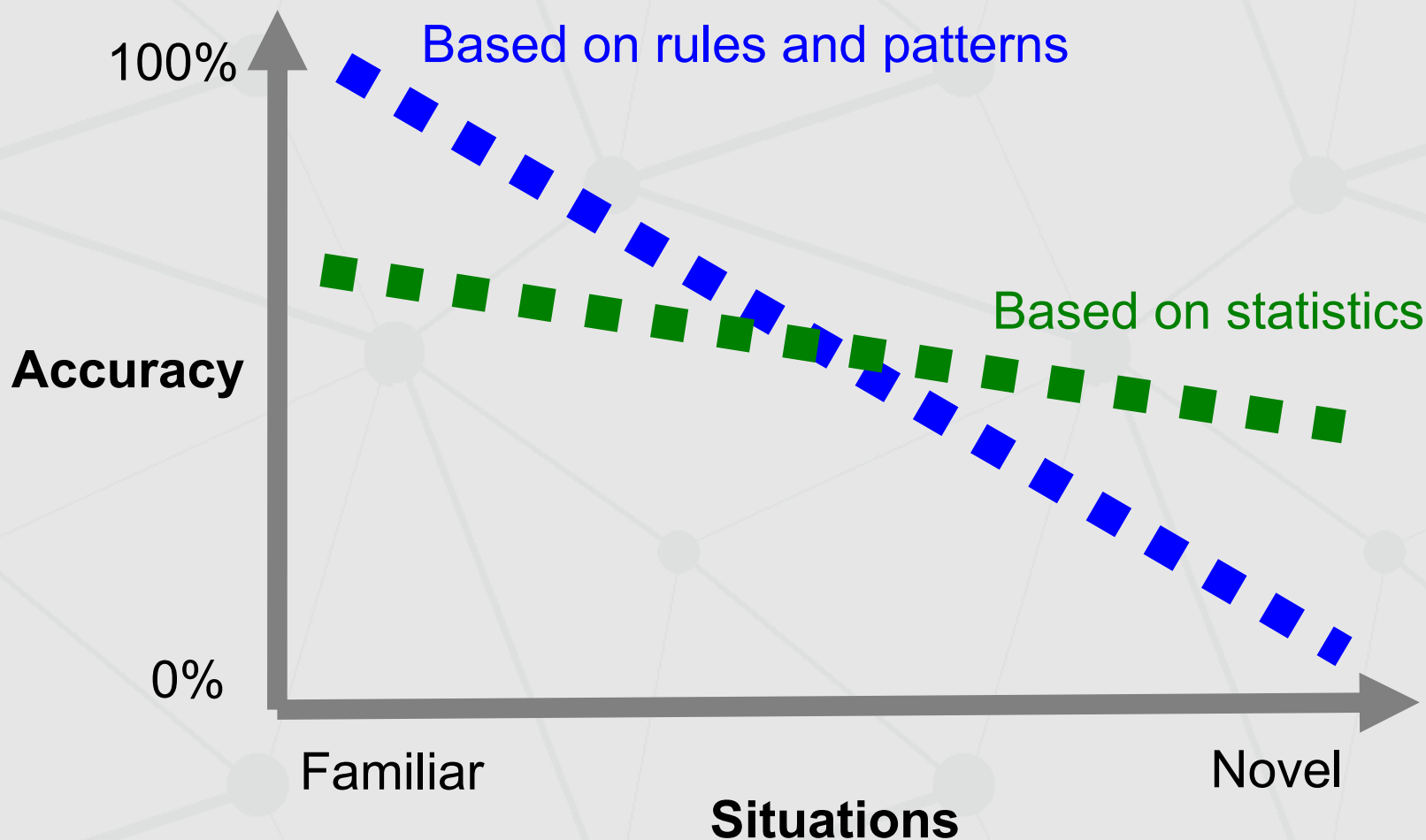
“disease OR illness” is much faster than “disease OR illness OR ail OR blast OR sick”  
(even if little bit less accurate)



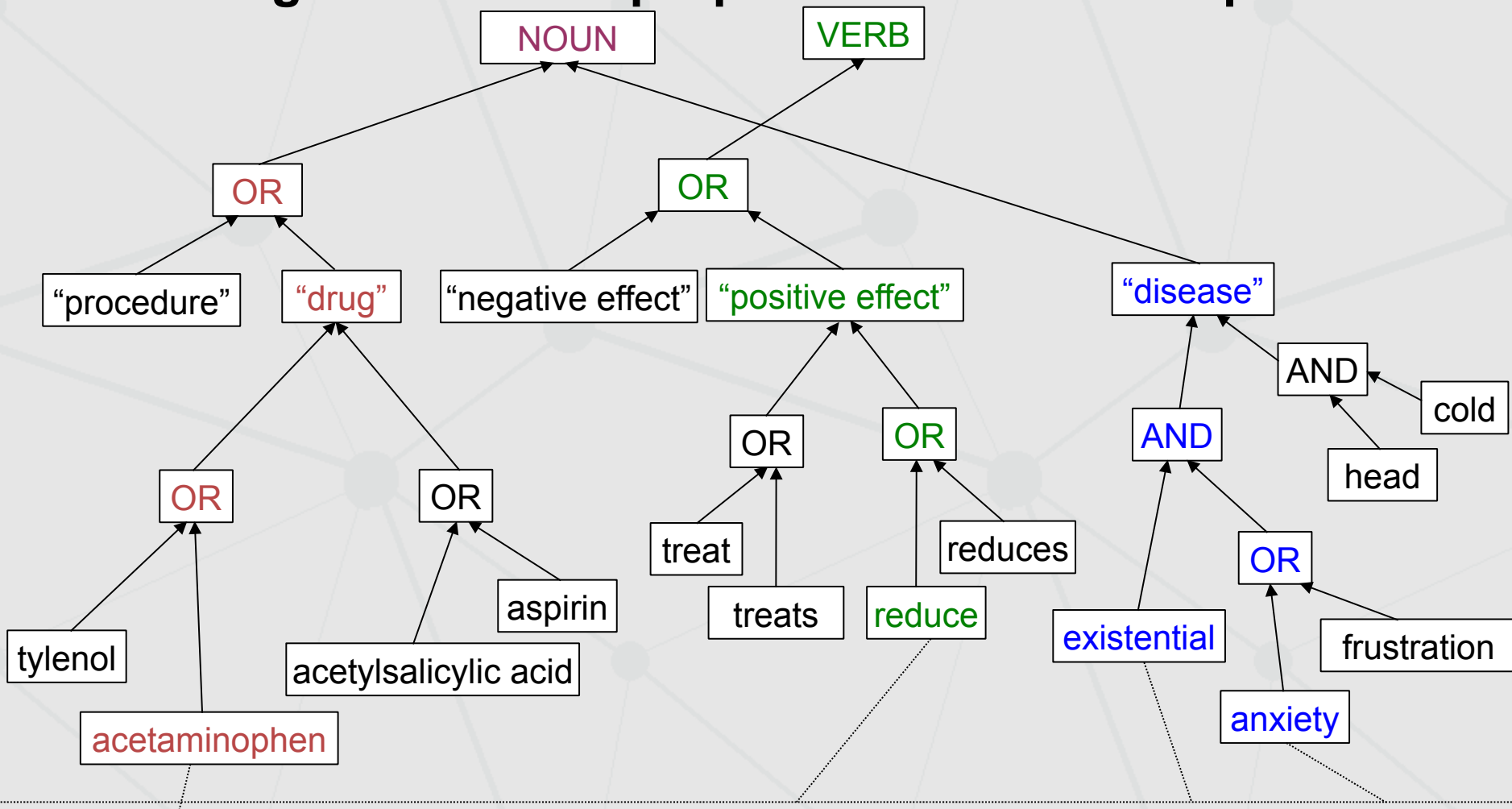
## Accuracy : Human vs. Computer : Sources of errors



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



## Finding entities with properties: Hierarchical patterns



acetaminophen may significantly reduce feelings of existential anxiety

## 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

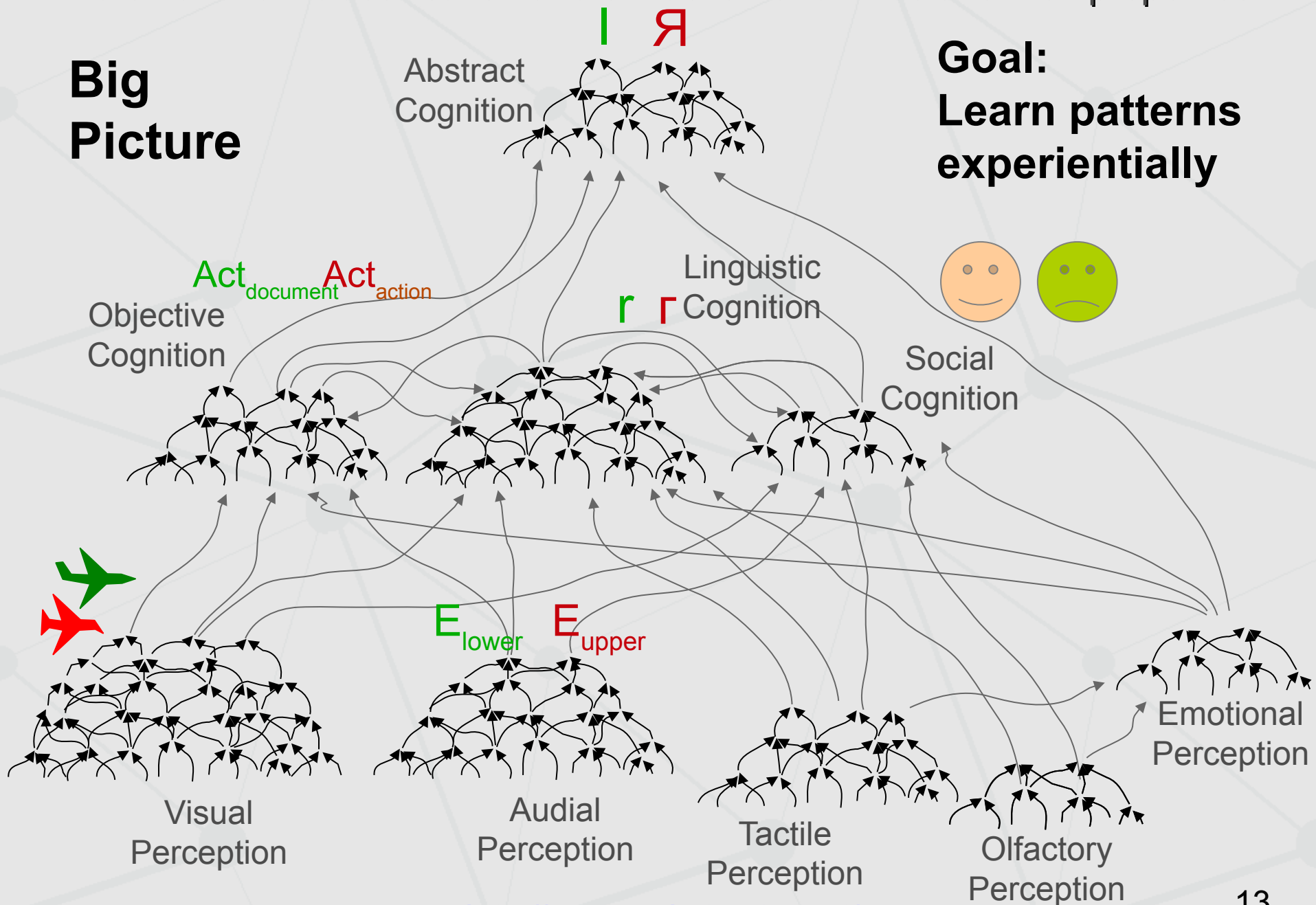
=

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

# Automatic text classification and extraction of entities and their properties

## 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

## Current implementation

The screenshot shows the website <https://aigents.com> with a navigation bar containing icons for 'Темы' (Topics), 'Сайты' (Sites), 'Новости' (News), 'Другие' (Others), and 'Беседа' (Chat). A notification bubble says 'Надо помочь?' (Need help?). The main content area displays a list of news items:

- today доллар сша руб . 60,3458 руб . ↑ 62,4677 <http://cbr.ru>
- 2015-07-31 apple campus will have an observation deck for visitors <http://wired.com>
- 2015-07-31 google street view cars now sniff pollution instead of wi-fi <http://wired.com>
- 2015-07-31 доллар сша руб . 58,9906 руб . ↑ 60,3458 <http://cbr.ru>
- 2015-07-31 北京当前温度 . . 22.8° 相对湿度 : 87% 24小时天气预报 风向风力 . 东北风 东北风 每日播报天气 pm2.5 : 29 优 <http://beijing.tianqi.com>
- 2015-07-30 доллар сша руб . 59,7665 руб . ↓ 58,9906 <http://cbr.ru>

At the bottom, there is a search bar with the text 'Введи текст для поиска' (Enter text for search) and a search button with a magnifying glass icon.

Авторские права 2015 Антон Колония, Aigents Group

The screenshot shows a mobile application interface with a top navigation bar containing icons for 'Темы' (Topics), 'Сайты' (Sites), 'Новости' (News), 'Другие' (Others), and 'Беседа' (Chat). The main content area displays a list of news items:

- yesterday доллар руб . 60,3458 руб . 62,4677 <http://cbr.ru>
- yesterday евро руб . 66,0002 руб . 68,5770 <http://cbr.ru>
- yesterday fortress among bidders for japan's simplex: sources .italy's exor wins battle to buy partnerre for \$6.9 billion <http://reuters.com>
- yesterday german carmakers buy nokia maps to fend off digital rivals .star trek-style home elevator could replace stairlifts | .eu regulators to decide on nxp <http://reuters.com>

At the bottom, there is a search bar with the text 'Поиск' (Search).

**Thank you for attention!**

Anton Kolonin  
[Webstructor project](#)  
2015, SIBIRCON/SibMedInfo