



**UNIVERSITY OF WATERLOO**  
**FACULTY OF MATHEMATICS**  
David R. Cheriton School  
of Computer Science



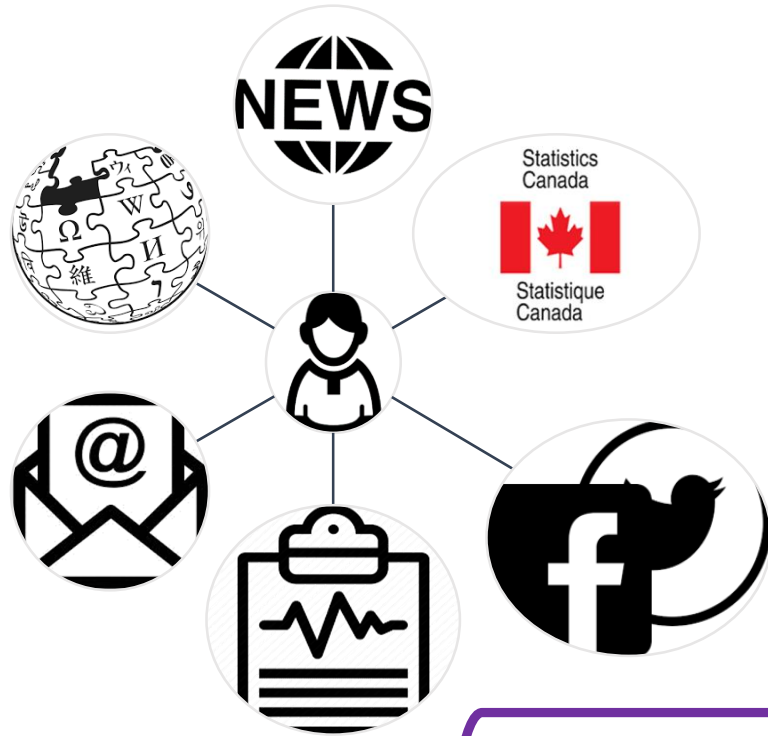
# Predictable and Consistent Information Extraction

Besat Kassaie

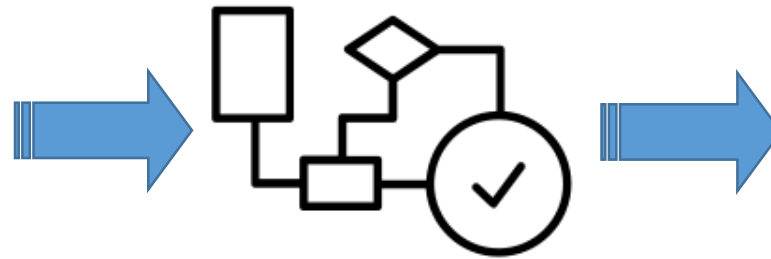
Frank Wm. Tompa

# Text is everywhere ...

## *Data Sources*



## *Algorithms*



## **Information Extraction**

## *Applications*

Sentiment Analysis

Customer Profiling

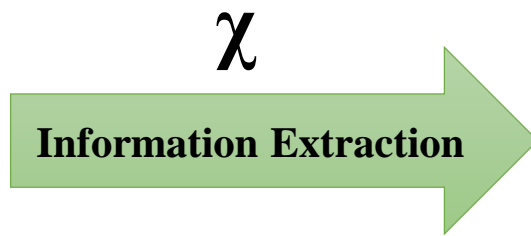
Cohort Identification

...

# Information Extraction



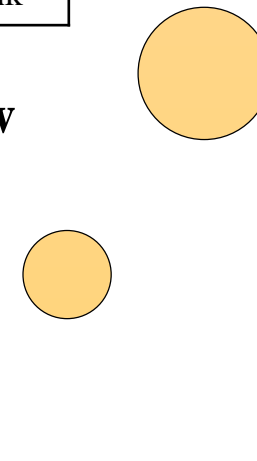
Source Documents



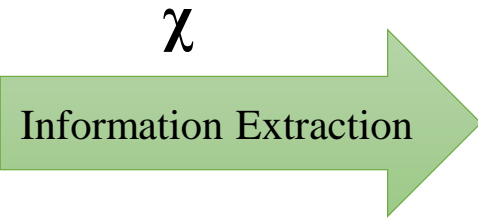
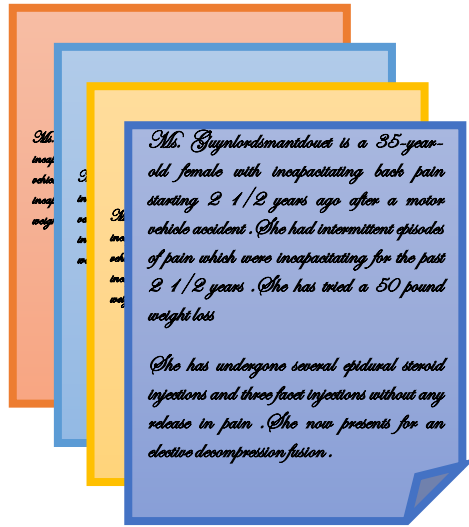
$A_1$	...	$A_n$
$V_{11}$	...	$V_{n1}$
$V_{12}$	...	$V_{n2}$
$V_{1k}$	...	$V_{nk}$

Extracted View

How to design extractors that generalize to extract accurate information from a diverse set of unseen sources?



# Overview of Problem



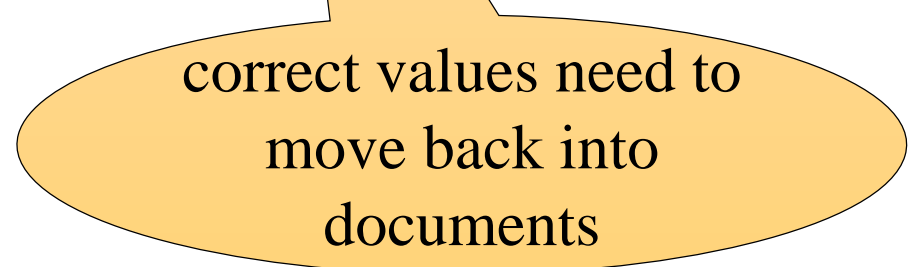
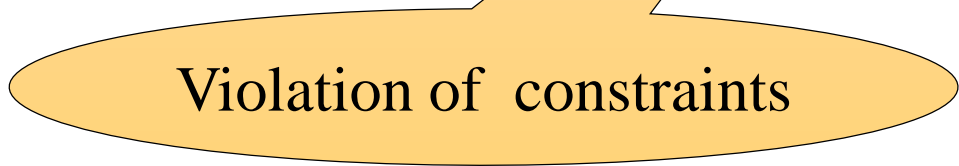
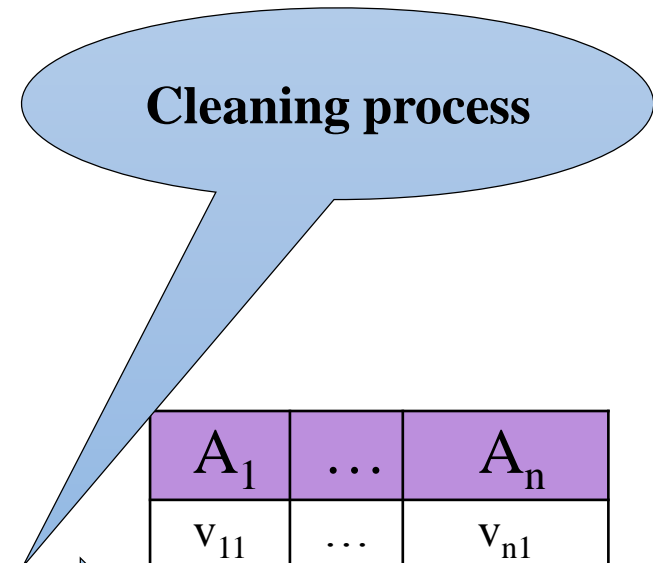
$A_1$	...	$A_n$
$v_{11}$	...	$v_{n1}$
$v_{12}$	...	$v_{n2}$
$v_{1k}$	...	$v_{nk}$

Extracted View



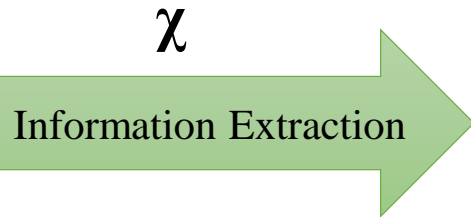
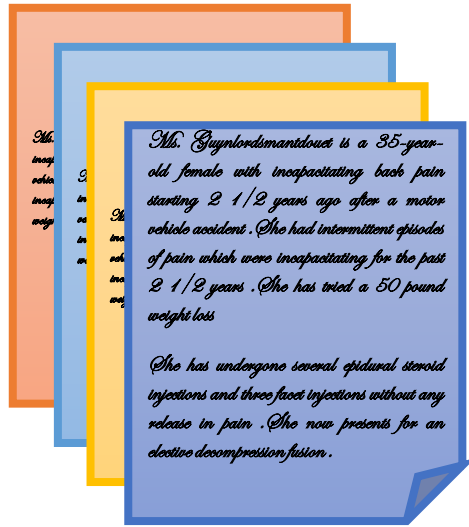
$A_1$	...	$A_n$
$v_{11}$	...	$v_{n1}$
$v'_{12}$	...	$v_{n2}$
$v_{1k}$	...	$v'_{nk}$

Updated View



# Overview of Problem

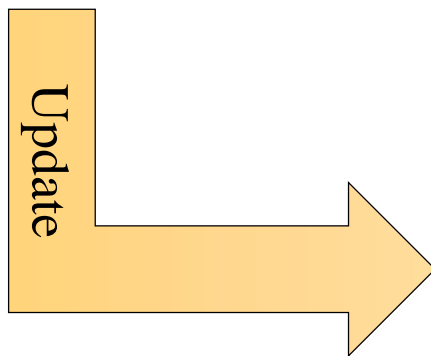
## Documents



$A_1$	...	$A_n$
$V_{11}$	...	$V_{n1}$
$V_{12}$	...	$V_{n2}$
$V_{1k}$	...	$V_{nk}$

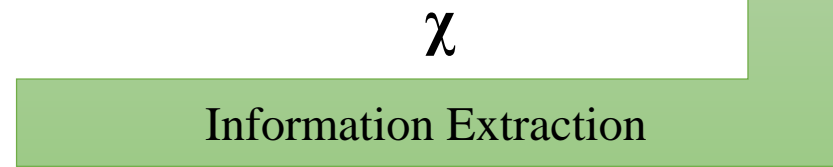


$A_1$	...	$A_n$
$V_{11}$	...	$V_{n1}$
$V'_{12}$	...	$V_{n2}$
$V_{1k}$	...	$V'_{nk}$



## Extracted View

## Updated View



# Overview of Problem

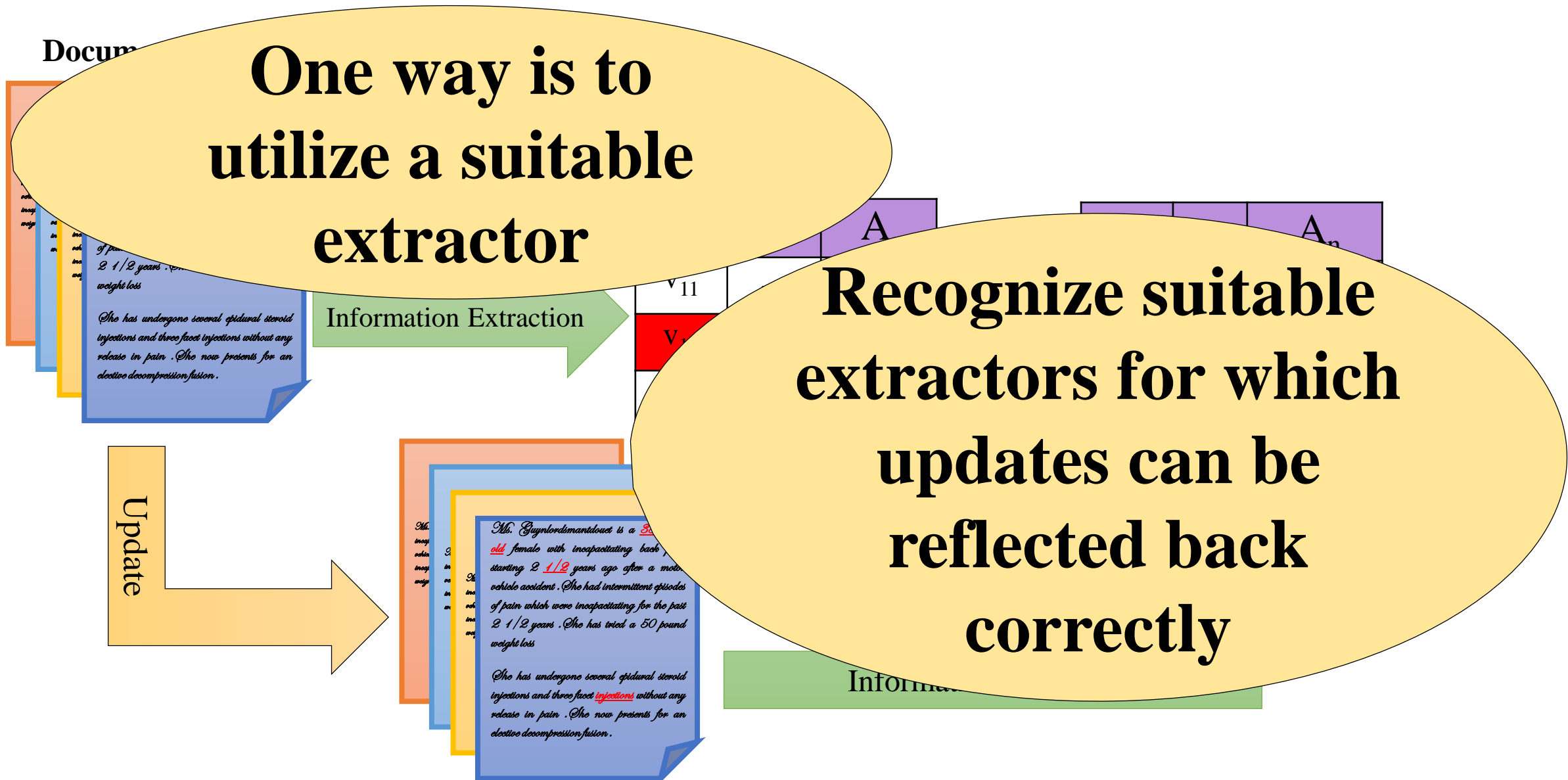
One way is to utilize a suitable extractor

Information Extraction

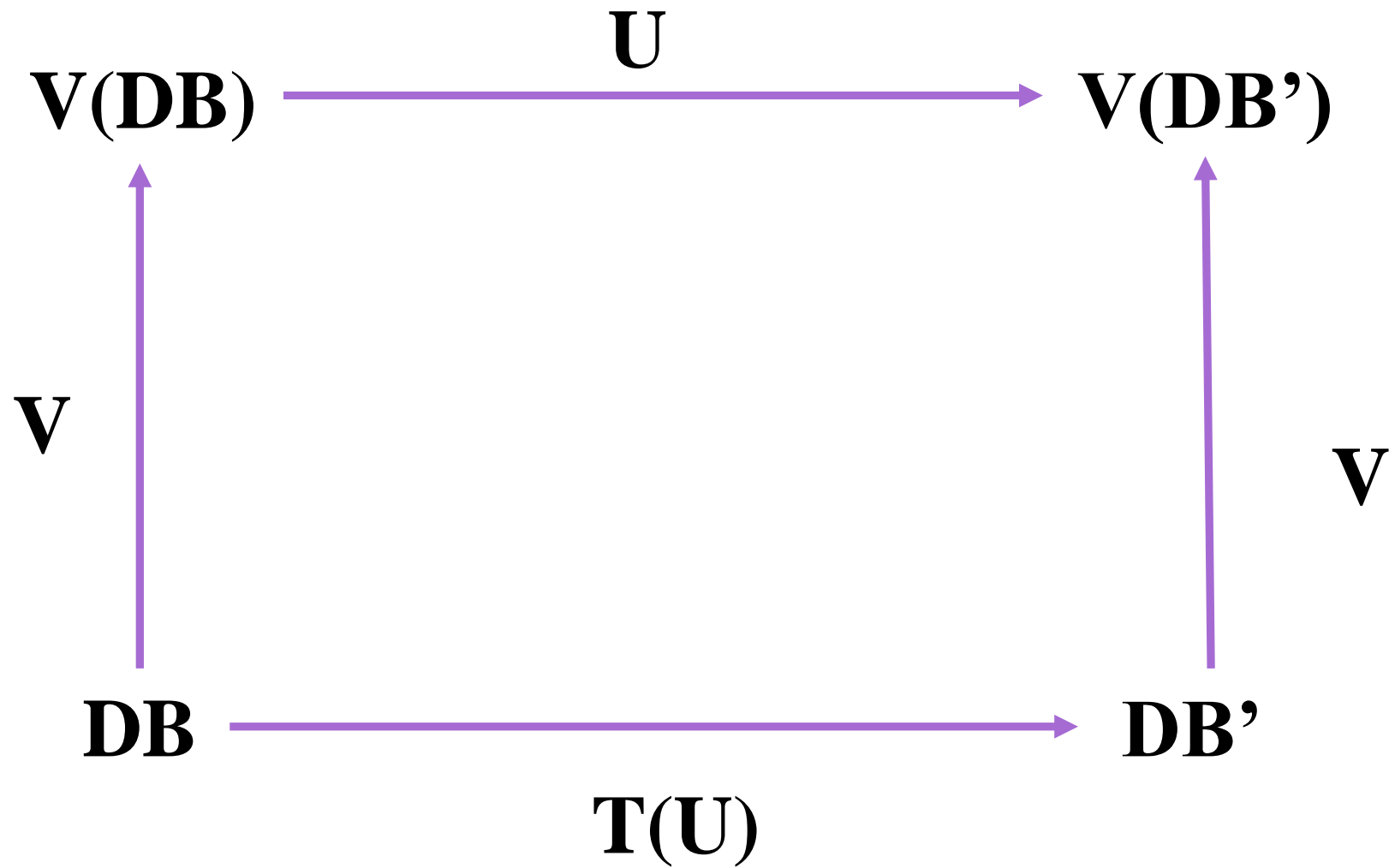
Recognize suitable extractors for which updates can be reflected back correctly

Update

Information



# Classical View Update Problem



# Extracted View Update Problem

$A_1$	...	$A_n$
$V_{11}$	...	$V_{n1}$
$V_{12}$	...	$V_{n2}$
$V_{1k}$	...	$V_{nk}$

U

$A_1$	...	$A_n$
$V'_{11}$	...	$V_{n1}$
$V_{12}$	...	$V'_{n2}$
$V_{1k}$	...	$V_{nk}$

$\chi$



Translate(U) ?

$\chi$





# Trivial Document Update Approach

$A_1 : W_1$	$A_2 : W_2$	$A_3 : W_3$	$A_4 : W_4$
$v_1$	$v_2$	$v_3$	$v_4$



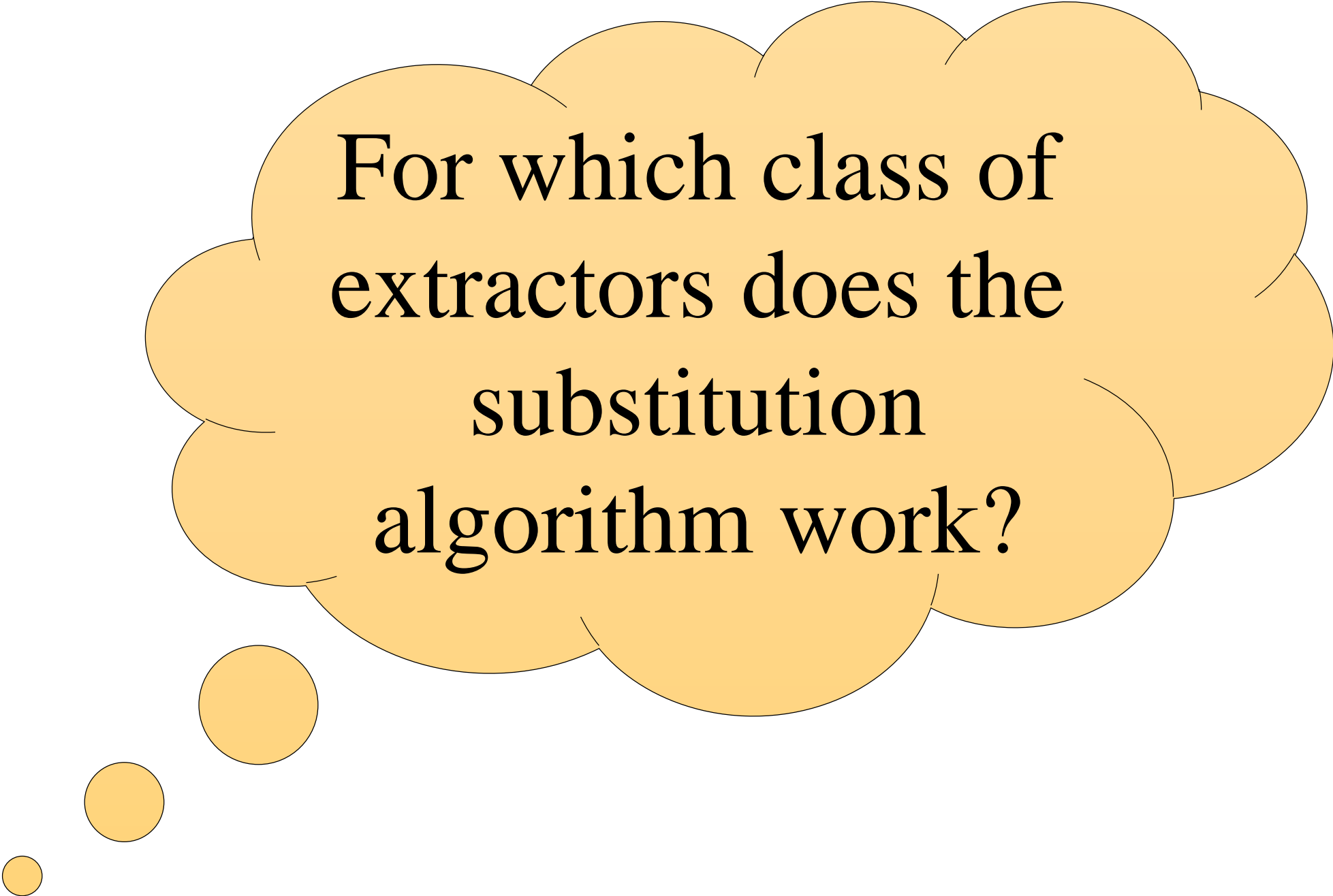
$A_1 : W_1$	$A_2 : W_2$	$A_3 : W_3$	$A_4 : W_4$
$v_1$	$v_2$	$v_3$	$v'_4$

**Substitution**

The patient was admitted on 02/04/01 for surgery that day. She was taken to the operating room where she underwent a left shoulder hemiarthroplasty. The patient's surgery included the proximal humeral fracture that was found to be in four parts rather than three parts. Initially, it had been hoped that the fracture would be amenable to open reduction internal fixation. However, it was found to require hemiarthroplasty which was undertaken. The patient tolerated the procedure well and postoperatively was brought in stable condition to the Post-Anesthetic Recovery Unit and later to the regular floor. Postoperative course showed a well-healed and limited left hemiarthroplasty of the shoulder. Postoperatively, the patient was successfully intubated with 5/5 strength throughout her left upper extremity and initial ventilation without incident, which reached normal parameters. She had back spiking with all fingers. On postoperative day #1, her vital numbers were good and she was able to tolerate good PO. She was started with Physical Therapy for pain relief and passive range of motion only. She was kept on paracetamol with good effect. By postoperative day #2, she was doing much better with excellent pain control. Her vital signs were stable resulting in a clean, dry, and intact wound. She continued to be successfully intubated. She was able to be stable and ready for discharge to home.

# Trivial Document Update Approach

Substitution does not work for all  
extractors!



For which class of  
extractors does the  
substitution  
algorithm work?

The set of extracted values in the corresponding record must be a subset of words and phrases appearing in the input.

$A_1 : W_1$	$A_2 : W_2$	$A_3 : W_3$	$A_4 : W_4$
$v_1$	$v_2$	$v_3$	$v_4$

The patient was admitted on 02/04/01 for surgery that day. She was taken to the operating room where she underwent a left shoulder hemiarthroplasty. Preoperative findings included the proximal humeral fracture that was found to be in four parts rather than three parts. Initially, it had been hoped that the fracture would be amenable to open reduction internal fixation. However, it was found to require hemiarthroplasty which was undertaken. The patient tolerated the procedure well and postoperatively was brought in stable ambulatory condition to the Post-Anesthesia Recovery Unit and later to the regular floor. Postoperative course showed a well aligned and healed left hemiarthroplasty of the shoulder. Postoperatively, the patient was progressively intact with 5/5 strength throughout her distal left upper extremity and intact sensation, motor, and cranial nerve distributions. She had brisk reflexes, right in all fingers. On postoperative day # 1, her left ambulation was excellent and good when she was in the good PPD. She worked with Physical Therapy for painless and passive range of motion only. She was kept on pain medication. By postoperative day # 2, she was doing much better with excellent pain control. Her walking was changed resulting a short, dog, and intact wound. She continued to be progressively intact. She was felt to be stable and ready for discharge to home.

This needs to be true for every possible input document.

# Strict Extractor

$A_1 : W_1$	$A_2 : W_2$	$A_3 : W_3$	$A_4 : W_4$
$V_1$	$V_2$	$V_3$	$V_4$

The patient was admitted on 02/06/01 for surgery that day. She was taken to the operating room where she underwent a left shoulder hemiarthroplasty. Preoperative findings included the proximal humeral fracture that was found to be in four parts rather than three parts. Initially, it had been hoped that the fracture would be amenable to open reduction internal fixation. However, it was found to require hemiarthroplasty which was undertaken. The patient tolerated the procedure well and postoperatively was brought in stable ambulatory condition to the Post-Anesthesia Recovery Unit and later to the regular floor. Postoperative course showed a well aligned and healed left hemiarthroplasty of the shoulder. Postoperatively, the patient was neurovascularly intact with 5/5 strength throughout her distal left upper extremity and intact sensation, motor, and small nerve distributions. She had brisk capillary refill in all fingers. On postoperative day # 1, her postoperative course was unremarkable and she was doing well. She was good PO. She worked with Physical Therapy for pain and passive range of motion only. She was kept on pain medication. By postoperative day # 2, she was doing much better with excellent pain control. She was discharged on a clear, day, and intact wound. She continued to be neurovascularly intact. She was felt to be stable and ready for discharge to home.

We need to have access to positions from which the attributes are extracted.

$$P_{\chi}(D, j)$$

The patient was admitted on 02/08/01 for surgery that day. She was taken to the operating room where she underwent a left shoulder hemiarthroplasty. Postoperative findings included the proximal humeral fracture that was found to be in four parts rather than three parts. Initially, it had been hoped that the fracture would be amenable to open reduction internal fixation. However, it was found to require hemiarthroplasty which was undertaken. The patient tolerated the procedure well and postoperatively was brought in stable condition to the Post-Anaesthesia Recovery Unit and later to the regular floor. Postoperative course showed a well aligned and located left hemiarthroplasty of the shoulder. Postoperatively, the patient was successfully treated with 5/5 strength throughout her distal left upper extremity and intact sensation, motor, and radial nerve distributions. She had brisk capillary refill in all fingers. On postoperative day 1, her pain medications were weaned to oral form when she was tolerating oral PO. She worked with Physical Therapy for pain relief and passive range of motion only. She was kept on passive activities. By postoperative day 2, she was doing well better with excellent pain control. Her vital signs were stable and she was discharged on day 3, when she was well. She continued to be neurovascularly intact. She was felt to be stable and ready for discharge to home.

This needs to be true for all possible input documents and corresponding extracted attributes.

# Computable Extractor

$P_{\chi}(D, j)$

The patient was admitted on 02/04/01 for surgery that day. She was taken to the operating room where she underwent a left shoulder hemiarthroplasty. Preoperative findings included the proximal humeral fracture that was found to be in four parts rather than three parts. Initially, it had been hoped that the fracture would be amenable to open reduction internal fixation. However, it was found to require hemiarthroplasty which was undertaken. The patient tolerated the procedure well and postoperatively was brought in stable, unlabored condition to the Post-Anesthetic Recovery Unit and later to the regular floor. Postoperative x-rays showed a well aligned and located left hemiarthroplasty of the shoulder. Postoperatively, the patient was neurovascularly intact with 5/5 strength throughout her distal left upper extremity, and intact sensation, motor, and cranial nerve distributions. She had brisk capillary refill in all fingers. On postoperative day #1, her pain medications were weaned to oral form when she was tolerating oral PO. She worked with Physical Therapy for painless and passive range of motion only. She was kept on pain-inhibitor antibiotics. By postoperative day #2, she was doing much better with excellent pain control. Her dressing was changed resulting in clean, dry, and intact wound. She continued to be neurovascularly intact. She was felt to be stable and ready for discharge to home.

We need to be able to predict the effect on the extracted records when changing values in legitimate positions of document.



# Stable Extractor



# Stable Extractor

**D**

Ms. Smith is 35 years old with incapacitating back pain starting 2 1/2 years ago after a motor vehicle accident. She had intermittent episodes of pain which were incapacitating for the past 2 1/2 years. She has tried a 50 pound weight loss. She has undergone several epidural steroid injections and three facet injections without any release in pain. She now presents for an elective decompression fusion .

$A_1$	$A_2$
35	2 1/2

**g(D, 2)**

Ms. Smith is 35 years old with incapacitating back pain starting 3 years ago after a motor vehicle accident. She had intermittent episodes of pain which were incapacitating for the past 3 years. She has tried a 50 pound weight loss. She has undergone several epidural steroid injections and three facet injections without any release in pain. She now presents for an elective decompression fusion .

$A_1$	$A_2$
35	3



# Stable Extractor

**D**

Ms. Smith is **35** years old with incapacitating back pain starting **2 1/2** years ago after a motor vehicle accident. She had intermittent episodes of pain which were incapacitating for the past **2 1/2** years. She has tried a 50 pound weight loss. She has undergone several epidural steroid injections and three facet injections without any release in pain. She now presents for an elective decompression fusion .

$A_1$	$A_2$
35	2 1/2



**g(D, 2)**

Ms. Smith is **35** years old with incapacitating back pain starting 3 years ago after a motor vehicle accident. She had intermittent episodes of pain which were incapacitating for the past 3 years. She has tried a **50** pound weight loss. She has undergone several epidural steroid injections and three facet injections without any release in pain. She now presents for an elective decompression fusion .

$A_1$	$A_2$
35	50

# Stable Extractor

D

Ms. Smith is 35 years old with incapacitating back pain starting 2 1/2 years ago after a motor vehicle accident. She had intermittent episodes of pain which were incapacitating for the past 2 1/2 years. She has tried a 50 pound weight loss. She has undergone several epidural steroid injections and three facet injections without any release in pain. She now presents for an elective decompression fusion .

A <sub>1</sub>	A <sub>2</sub>
35	2 1/2



g(D, 2)

Ms. Smith is 35 years old with incapacitating back pain starting 3 years ago after a motor vehicle accident. She had intermittent episodes of pain which were incapacitating for the past 3 years. She has tried a 50 pound weight loss. She has undergone several epidural steroid injections and three facet injections without any release in pain. She now presents for an elective decompression fusion .

A <sub>1</sub>	A <sub>2</sub>
50	3

# Stable Extractor

D

Ms. Smith is 35 years old with incapacitating back pain starting 2 1/2 years ago after a motor vehicle accident. She had intermittent episodes of pain

pain. She now presents for an elective decompression fusion .

A <sub>1</sub>	A <sub>2</sub>
35	2 1/2

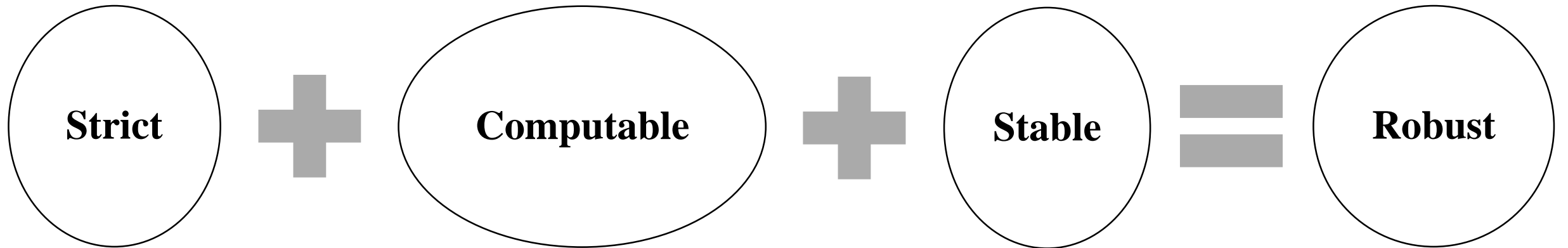
g(D, 2)

Ms. Smith is 35 years old with incapacitating back pain starting 3 years ago after a motor vehicle accident. She had intermittent episodes of pain which were incapacitating for the past 3 years. She has tried a 50 pound weight loss. She has undergone several epidural steroid injections and three facet injections without any release in pain. She now presents for an elective decompression fusion .

A <sub>1</sub>	A <sub>2</sub>
50	3



# Robust Extractors



# Theorem

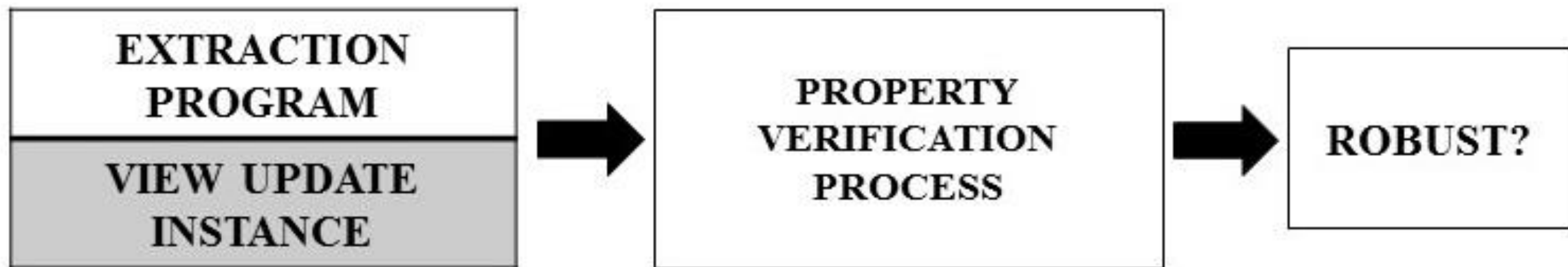
*Consider*

- *any strict, computable, and stable extractor  $X: D \rightarrow R$  producing a relation with  $\mathcal{T}$  attributes,*
- *any indexed set of domain preserving functions  $\mathcal{F} = \{f_i \mid f_i : \mathcal{W}_i \rightarrow \mathcal{W}_i, \text{ where } i \in [1 \cdots \mathcal{T}]\}$ ,*
- *and any document  $D \in \mathcal{D}$ .*

*For all  $i \in [1 \cdots \mathcal{T}]$ , substituting  $f_i(v_i)$  for  $v_i$  in all spans identified by  $\mathcal{P}_X(D, i)$  produces  $D_{\mathcal{F}}^{\mathcal{P}}$  in such way that  $\mathcal{F}(\mathcal{X}(D)) = \mathcal{X}(D_{\mathcal{F}}^{\mathcal{P}})$ .*

# Verification





# Verification OF JAPE Programs

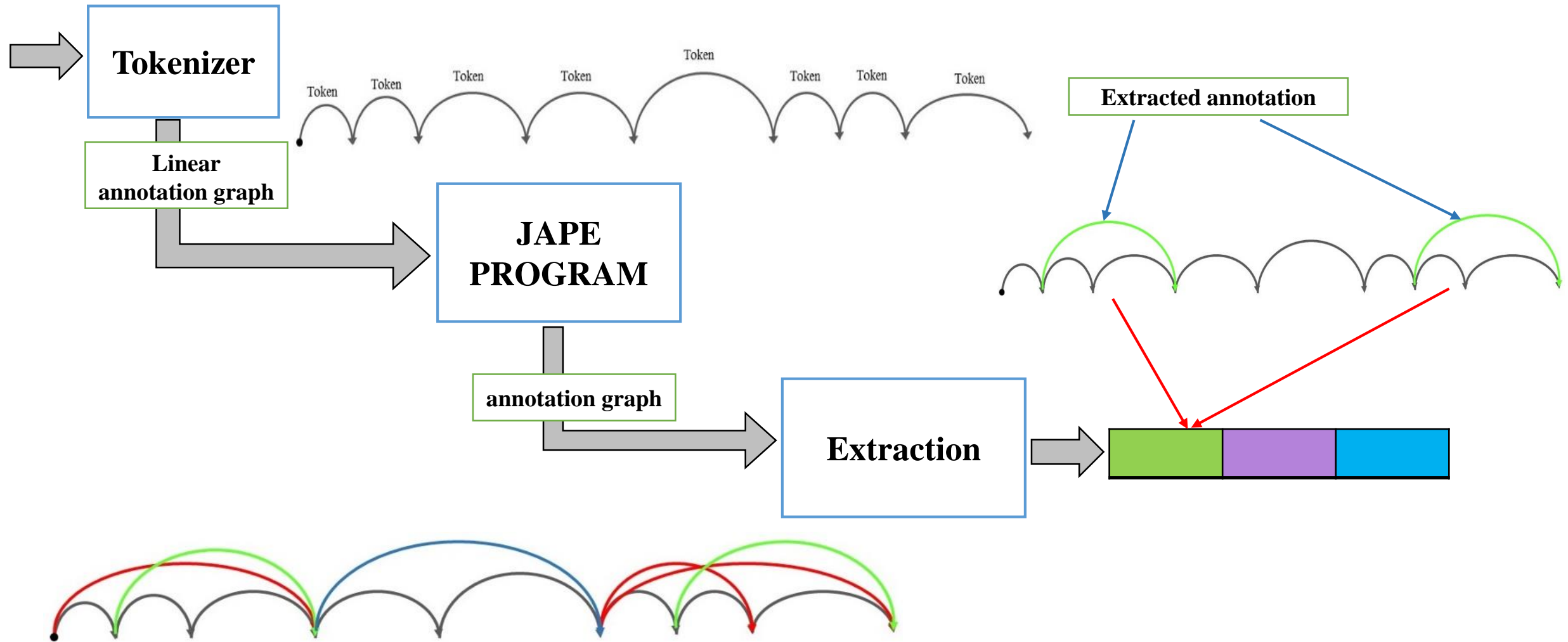
# OVERVIEW OF JAPE

In GATE rules are written in the JAPE language

GATE is a commonly used rule-based information extraction system



# JAPE Running Environment (simplified)



# Summary

- We observe that designing highly efficient and accurate extractors is not sufficient for new challenges we face.
- We introduce and formalize the extracted view update problem.
- We propose three sufficient properties for a robust extractor.
- We propose a verifier for testing these properties in JAPE programs.

Thank you.

Questions?