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

Predictable and Consistent Information Extraction

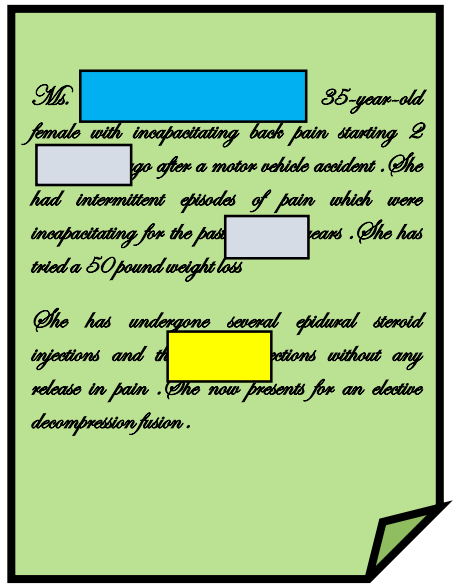
Besat Kassaie & Frank Wm. Tompa

PhD Seminar

May 2019

Information Extraction

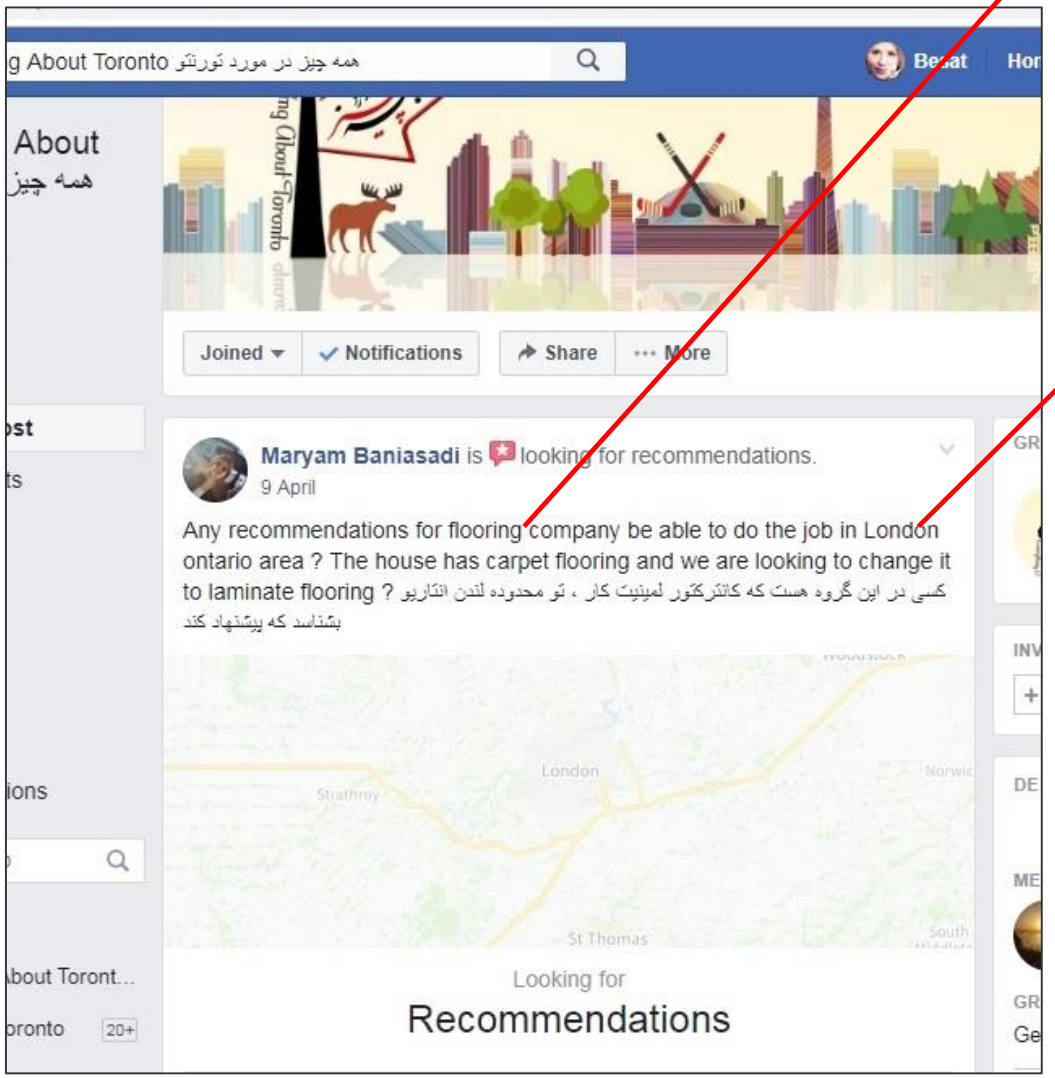
Information extraction identifies and isolates words and phrases within documents and presents them in relational tables

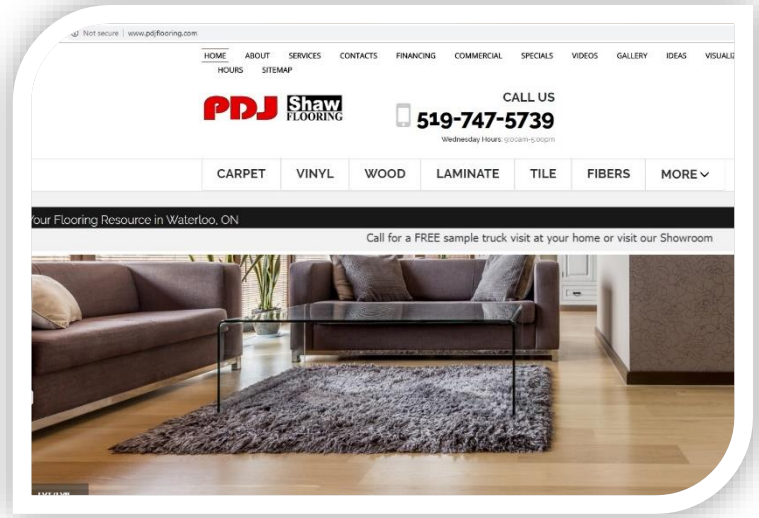


Information Extraction

Product: Flooring company | **Location:** London, Ontario

Social media sources





Information Extraction

Side effect: Short term memory loss | **Medication:** Depression

Medical records

Patient Name: [redacted]
Health Number: [redacted]
Date Of Birth: [redacted]
General Practitioner:
Exam Date: September 09, 2015
Practitioner: Hadley, K
Occupation:
[redacted] Cell)
Canada

General Information
Accompanied By: Alone
Driver's License (Type | Restrictions): G - Normal
Hobbies/Activities/Computer Use: want to get glasses mostly for reading

Chief Complaint Extended
Chief Complaint:
- GP suspect px has constricted peripheral vision
- lost her glasses from last year (readers) and bought drug store readers to read, and like this one better, cause they are stronger than the ones from UHV optical services
- safety glasses filled in too last year, but rarely used them
- occasional dryness OU, same as last year , esp after wake up
- px hit by a truck when she was 16yo, and had migraine and short term memory loss ever after, and taking depression meds.

Blur | Diplopia: No
Asthenopia | Headaches: No
Flashes | Floaters: No

Patient Ocular History

Current Medications
Other Prescription Medication:
- ALmotriptan tablets for migraines
- Fluoxetine HCL 20mg for anti-depression

Other OTC Medication: Meds for Allergy, low aspirin,

Family History
[?] Family History Unknown
Glaucoma: No
Blindness/Visual Impairment: No
Other Ocular Conditions: No
Other Health Conditions: DM | older brother

Habitual Rx
Rx Date last year | Lens Type Single Vision | Mat/Tint

	Sphere	Cyl	Axis	PD	ADD	PD
OD	+1.50					
OS	+1.50					

External Notes:

Visual Acuity

Pupils
Direct pupil reflex: 4+ Brisk | 4+ Brisk
Consensual pupil reflex: 4+ Brisk | 4+ Brisk
Accommodative Reflex: Present | Present
Pupil shape: Round | Round
Relative afferent pupil defect: Negative
Finding: PERRL & RAPD Negative

Confrontation Fields
Confrontation Fields: Full to finger count | Full to finger count

Pupillary Distance
OU PD (Dx | Nx): 65 | 60

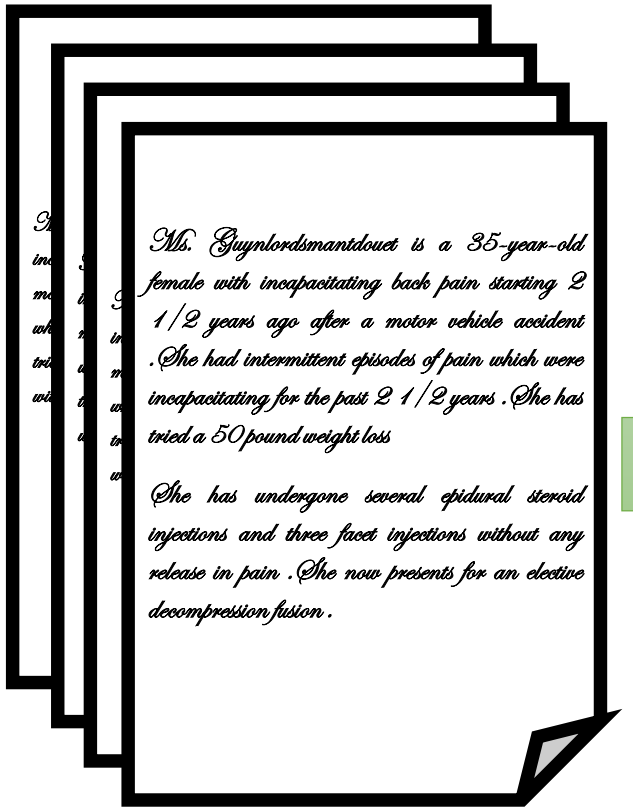
		Sphere	Cyl	Axis	Acuity
Static	OD	-0.25	Sph		6/6
	OS	Plano			6/6

Method Retinoscopy | User Y39HE

Sphere	Cyl	Axis	Acuity

Cohort identification

Information Extraction



Source Documents

χ

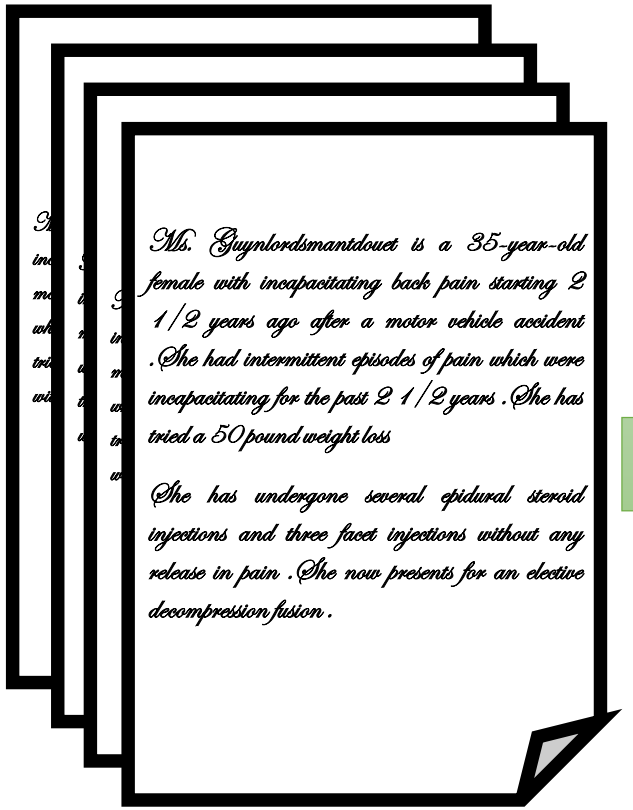
Information Extraction

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

Extracted View

Information Extraction

How to design extraction algorithms that generalize to extract accurate information from a diverse set of unseen sources.



Source Documents

χ

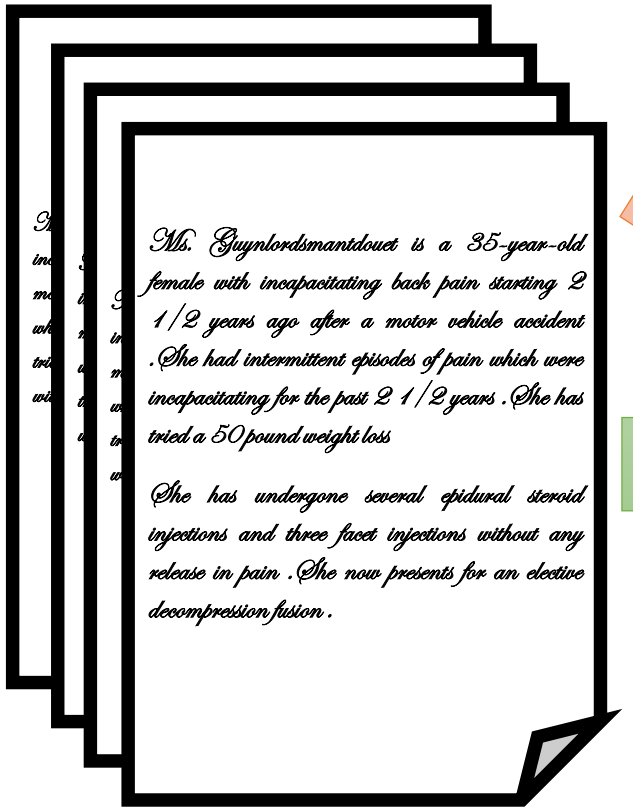
Information Extraction

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

Extracted View

Information Extraction

Immutable after extraction



Source Documents

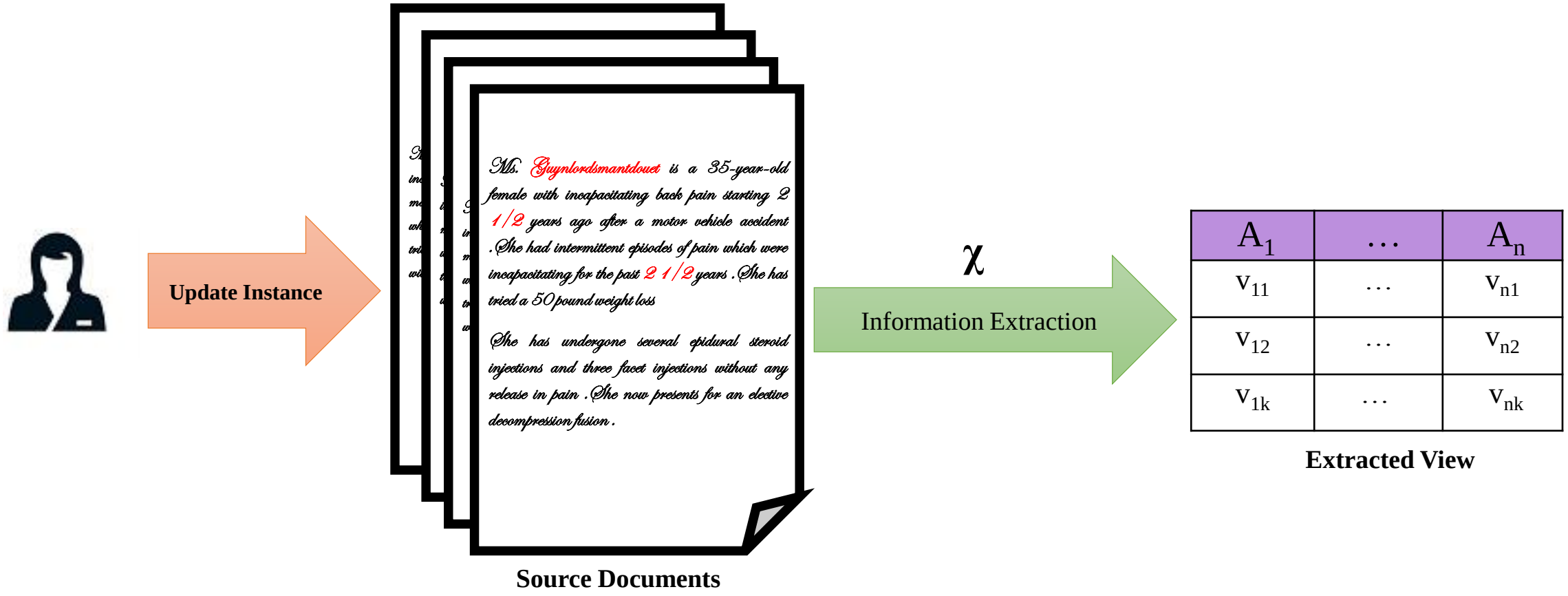
χ

Information Extraction

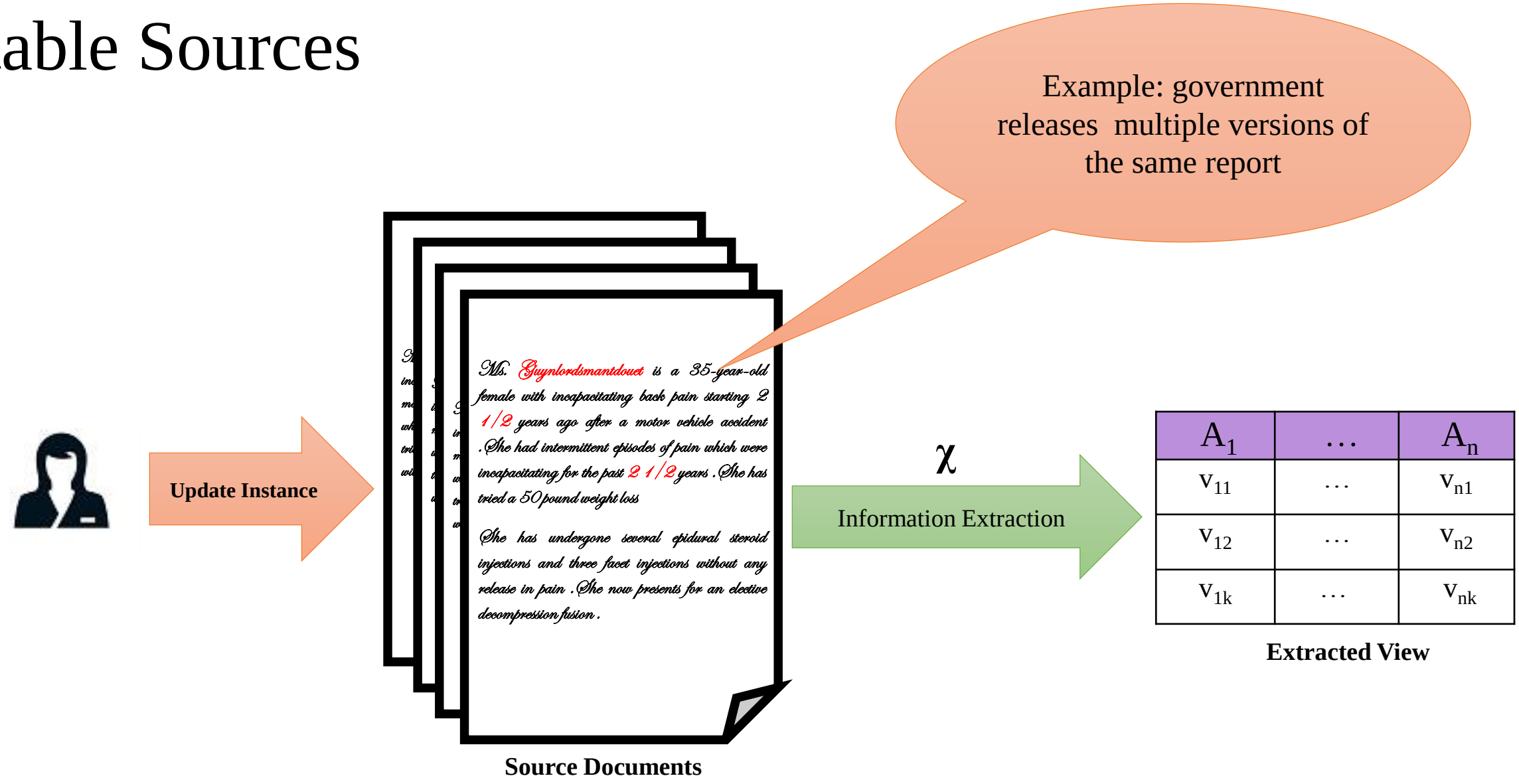
A_1	…	A_n
v_{11}	…	v_{n1}
v_{12}	…	v_{n2}
v_{1k}	…	v_{nk}

Extracted View

Mutable Sources



Mutable Sources



Statistics Canada Statistique Canada

Home → 57-003-X

Report on Energy Supply and Demand in Canada

2016 Revision

Release date: April 11, 2019

Highlights

Primary energy production in Canada increased 2.9% in 2016 to 19,709 petajoules. This followed a 1.2% increase in 2015.

Analysis

Energy supply and demand, 2016

Primary energy production in Canada increased 2.9% in 2016 to 19,709 petajoules. This followed a 1.2% increase in 2015.

Crude oil accounted for the largest proportion of primary energy production in Canada in 2016 at 45.1%, followed by natural gas (35.0%), primary electricity (9.3%), total coal (6.8%) and gas plant natural gas liquids (3.9%).

It was the seventh consecutive year in which crude oil accounted for the largest share of primary energy production.

Exports and imports increase

Exports of Canadian energy and energy products increased 2.4% in 2016 to 12,507 petajoules.

Canada exported 80.4% of its crude oil production in 2016, and 46.4% of its marketable natural gas.

Imports of energy increased 6.1% in 2016 to 3,659 petajoules. Crude oil accounted for 50.8% of imports, followed by natural gas (21.6%).

Energy consumption decreased

Canada's energy consumption decreased 0.8% in 2016 to 7,953 petajoules, following a 0.8% decrease in 2015.

Energy use increased in three of six sectors including public administration (+3.1%), industrial (+1.9%), and agriculture (+1.0%). Residential (-6.8%), commercial and other institutional (-1.3%), and transportation (-0.5%) saw a decrease in energy use.

Within the industrial sector, energy consumption increased in forestry and logging and support activities (+17.1%), construction (+10.5%), mining and oil and gas extraction (+2.3%), and manufacturing (+0.8%).

Retail pump sales continued to represent the largest proportion of energy consumption in the transportation sector (63.2%), followed by road transport and urban transit (14.2%), airlines (9.6%), pipelines (7.1%), railways (3.1%), and marine (2.8%).

Refined petroleum products (39.8%) were the main source of energy consumed in Canada in 2016, followed by natural gas (33.8%) and electricity (22.7%).

Energy consumption trends across the country

Ontario, Alberta and Quebec continued to account for the majority of energy consumed in Canada. In 2016, their combined share of total energy consumption was 73.4%.

Six provinces recorded increases in energy consumption in 2016 compared with 2015. British Columbia (+3.9%) saw the greatest increase, followed by Prince Edward Island (+3.7%), New Brunswick (+2.9%), Manitoba (+0.8%), Newfoundland and Labrador (+0.5%), and Alberta (+0.3%).

Energy consumption decreased in 5 regions in 2016 compared to 2015. The largest decrease was in Nova Scotia (-3.3%) followed by Ontario (-3.0%), Saskatchewan (-2.6%), the Territories (-2.5%), and Quebec (-1.1%).

Note: The above text refers to the preliminary 2016 data.

Note to readers

Factors influencing revisions include late receipt of company data and revisions to previously estimated or reported data. The revised data are available in the appropriate tables.

Data for any period may be revised and included in subsequent issues (such revisions are incorporated in the database). Given that further revisions to submitted data are received after the publication issue of any given year, it should be borne in mind that the statistical series shown in this publication are not necessarily the same in every detail as those shown in other publications produced by the Energy Section of Statistics Canada. From time to time, revisions to previous years may be incorporated in the database; tables 25-10-0026-01 (www.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2510002601) , 25-10-0027-01 (www.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2510002701) , 25-10-0028-01 (www.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2510002801) , 25-10-0029-01 (www.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2510002901) , 25-10-0030-01 (www.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2510003001) and 25-10-0031-01 (www.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2510003101) .

Acknowledgements

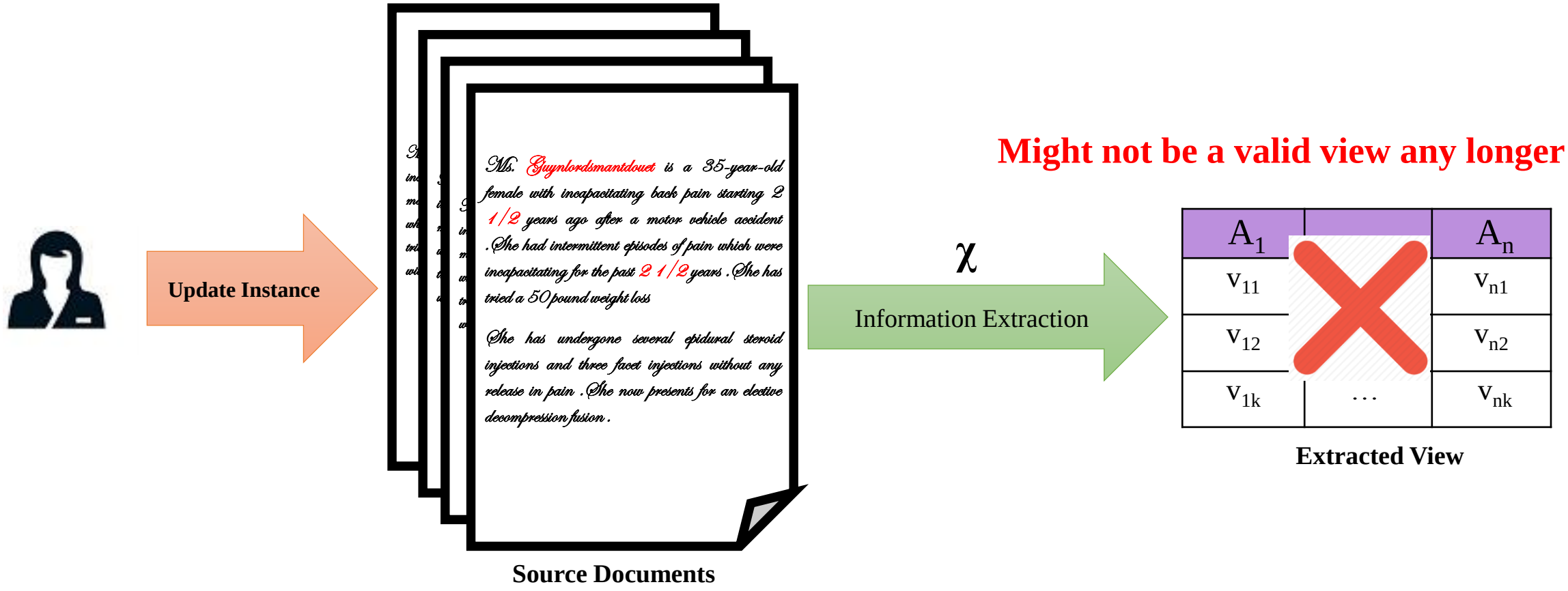
This publication was prepared in the Manufacturing and Energy Division under the direction of Kevin Roberts, Director, and Gabriel Gagnon, Section Chief.

Additional Information

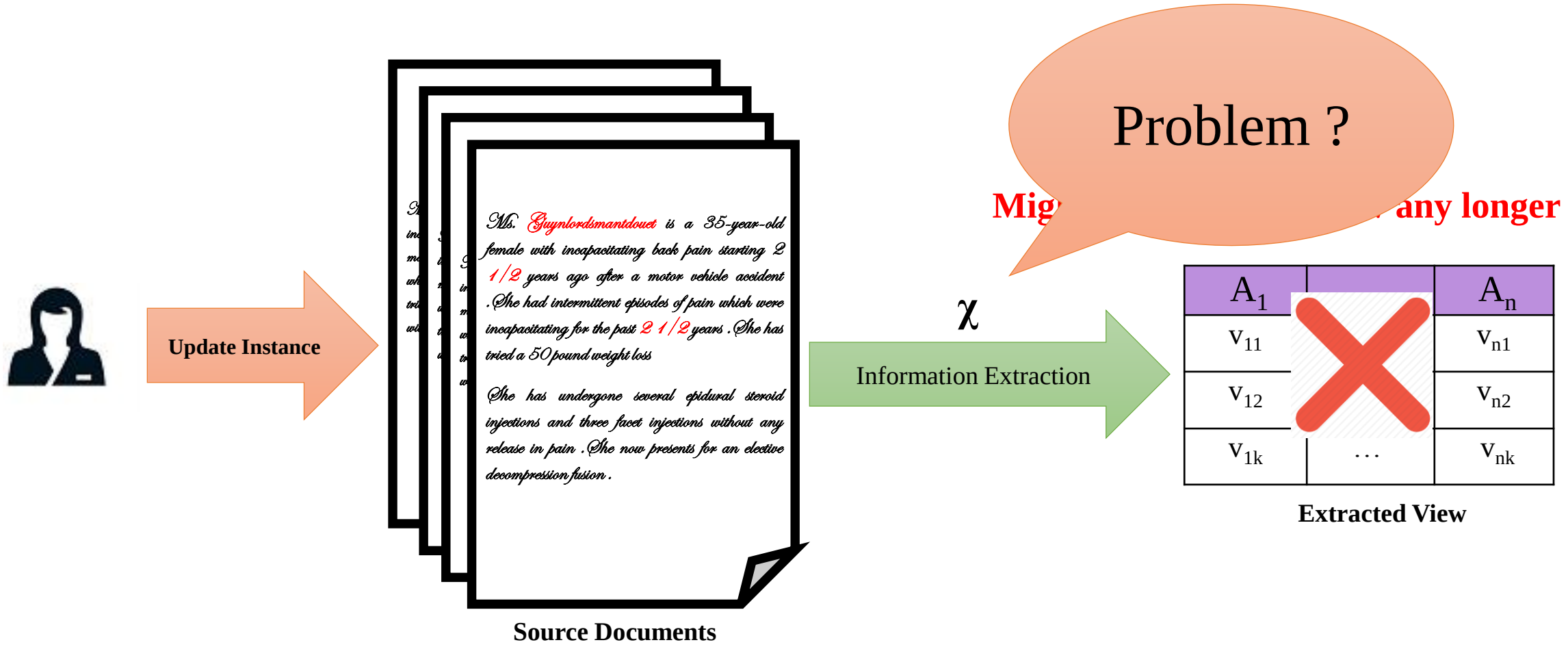
For information, please contact the Marketing and Dissemination Section (613) 951-9497 or toll-free (866) 873-8789; energ@statcan.gc.ca.

Next ➔

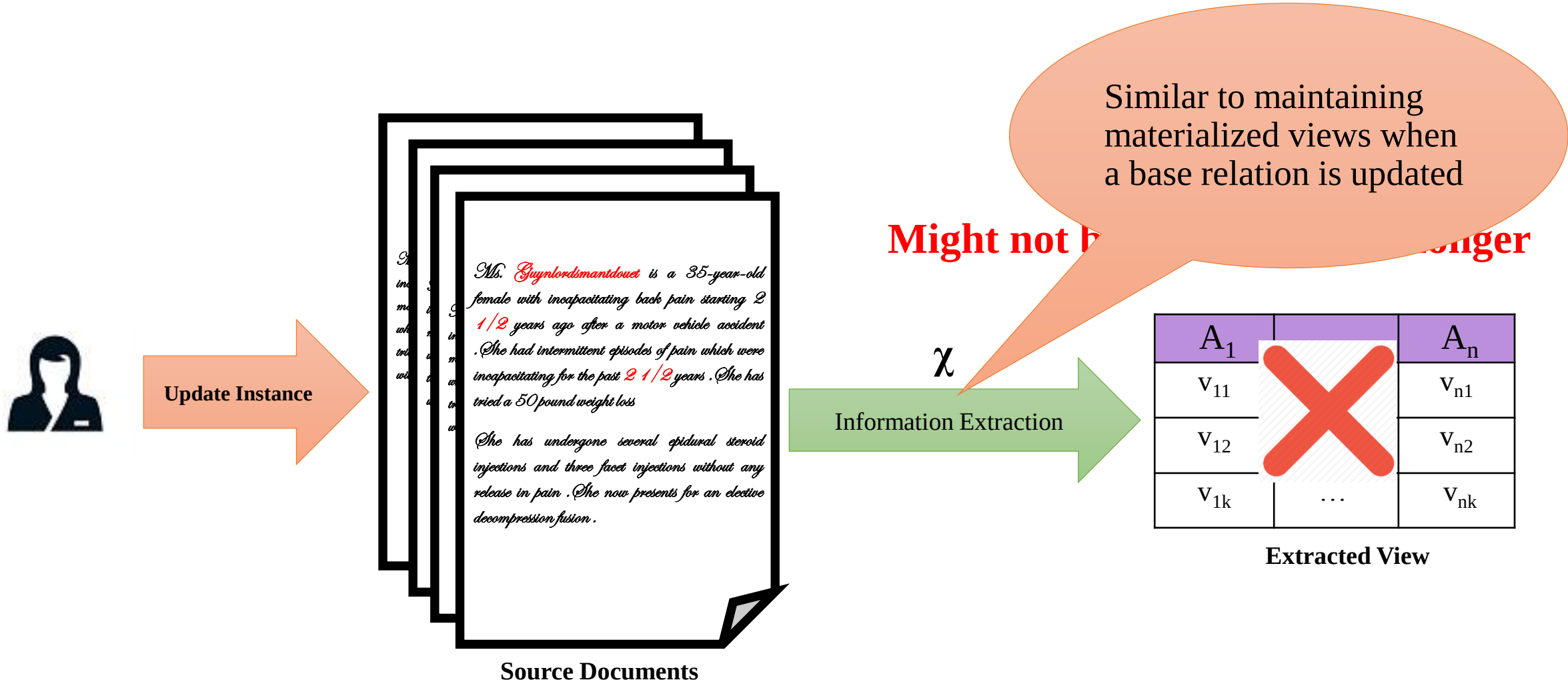
Mutable Sources



Mutable Sources



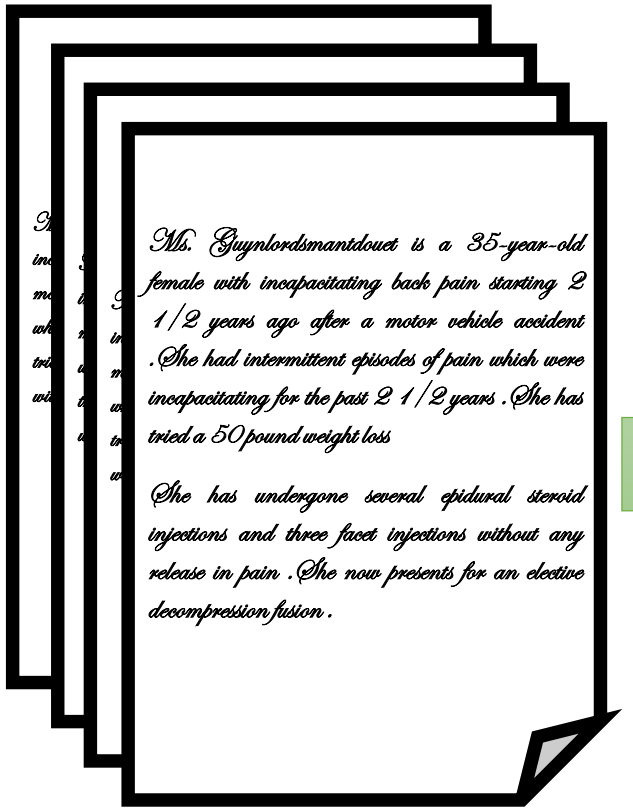
Mutable Sources



Problem Statement

Given an extraction specification and view/document update languages how can updates of source documents be translated to view update instance?

Information Extraction



Source Documents

χ

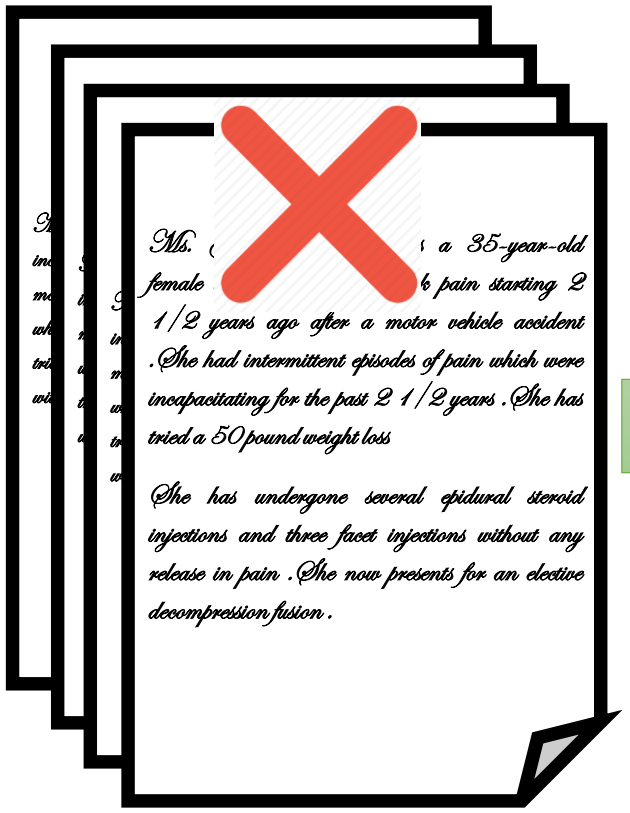
Information Extraction

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

Extracted View

Mutable Views

Might not be in a valid state any longer



Source Documents

χ

Information Extraction

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

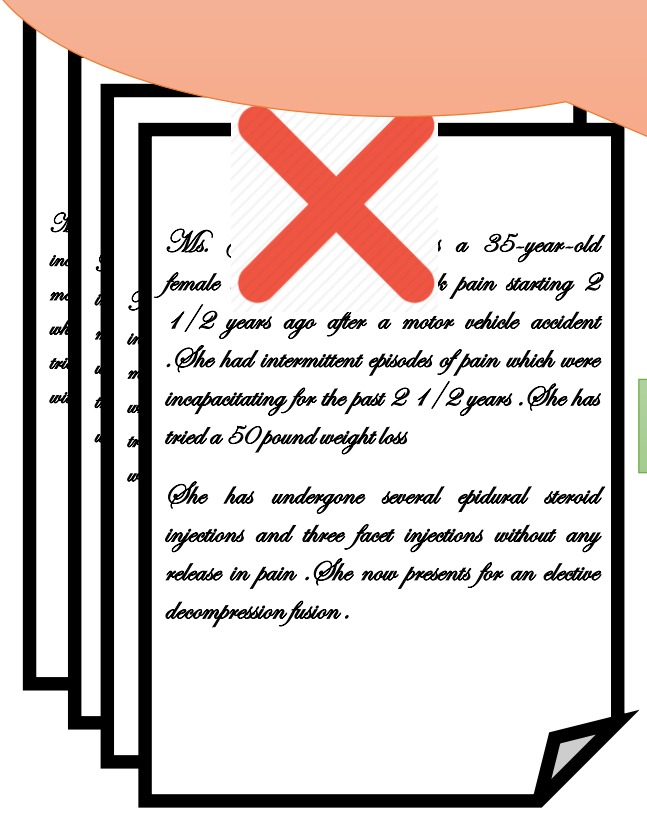
Extracted View

Update Instance

Mutable Views

Mig r

Problem ?



Source Documents

χ

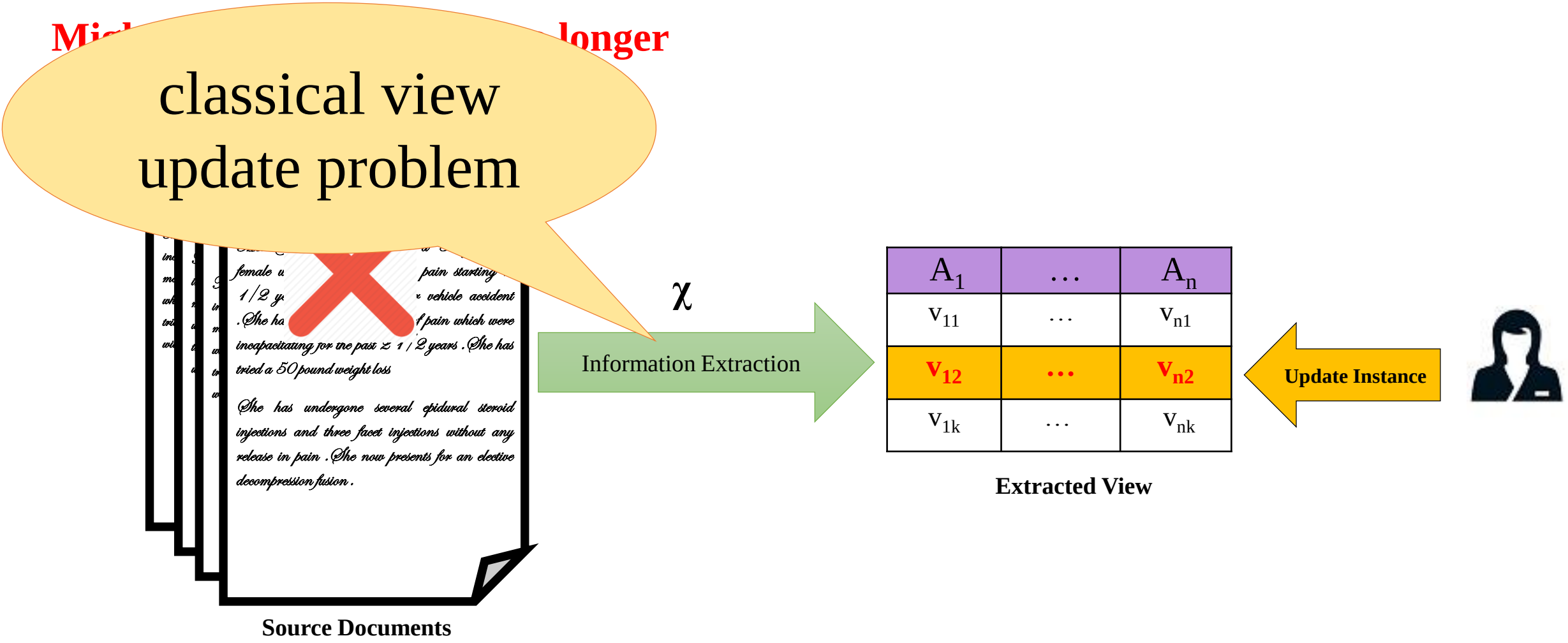
Information Extraction

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

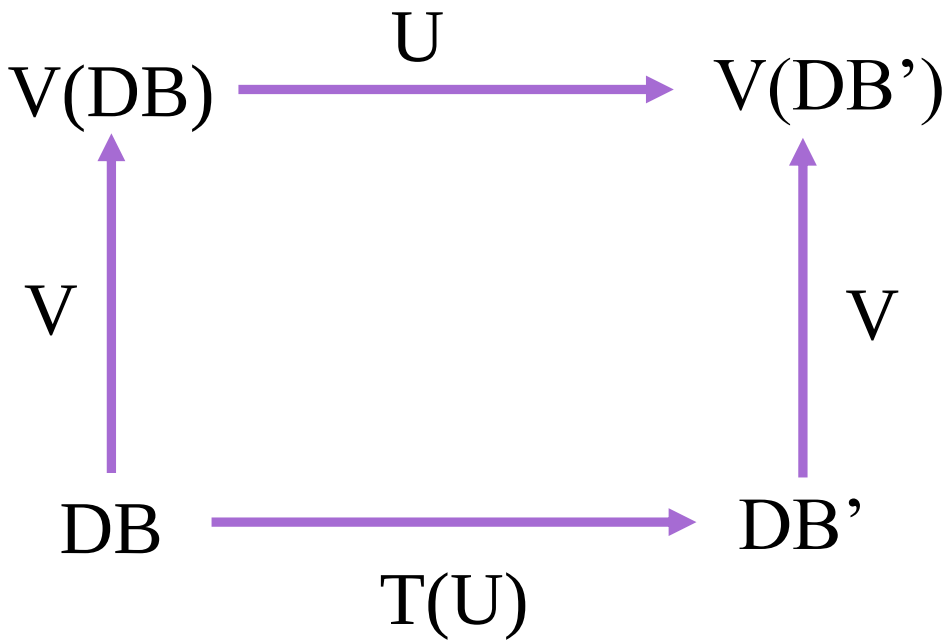
Extracted View

Update Instance

Mutable Views

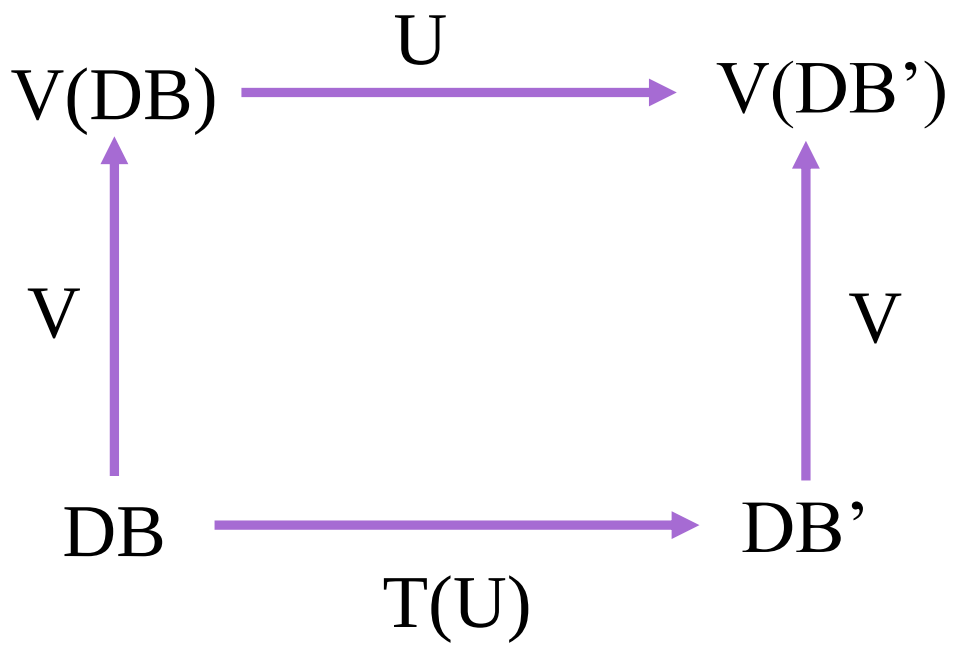


Classical View Update Problem



Classical View Update Problem

View is many-to-one relation, there may not be a unique translation



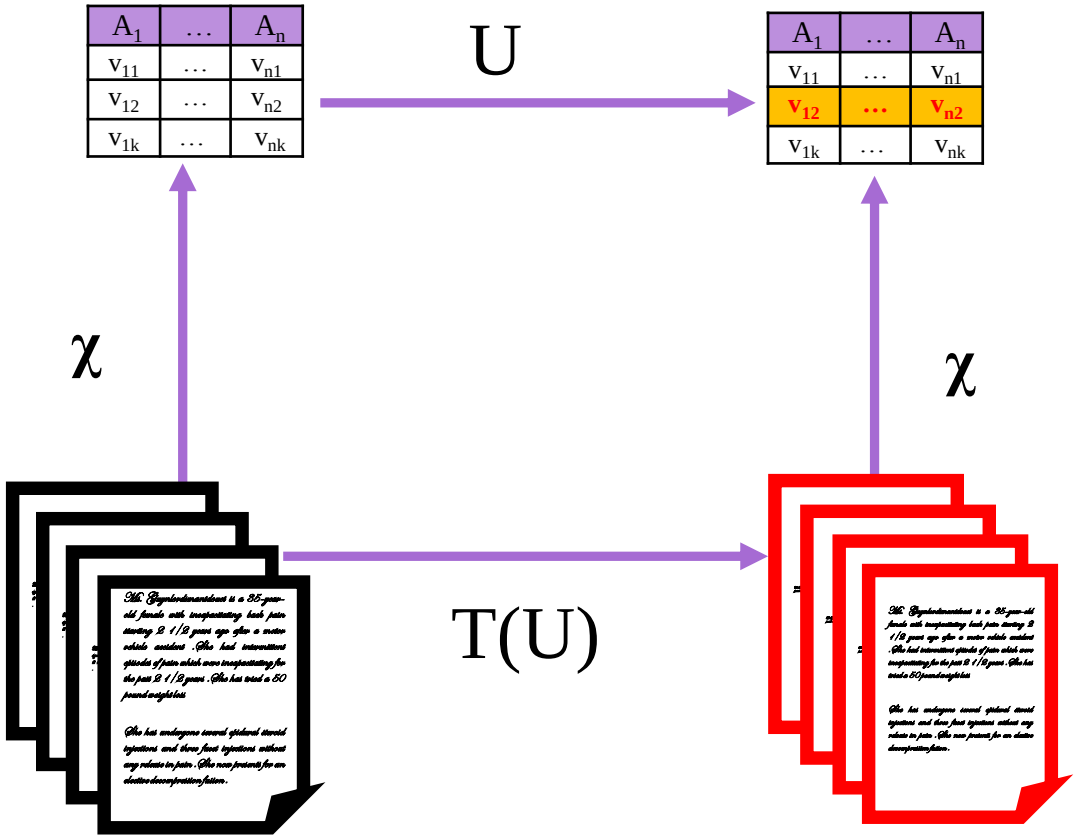
Translation may create inconsistencies in the database

Translation may not even exist

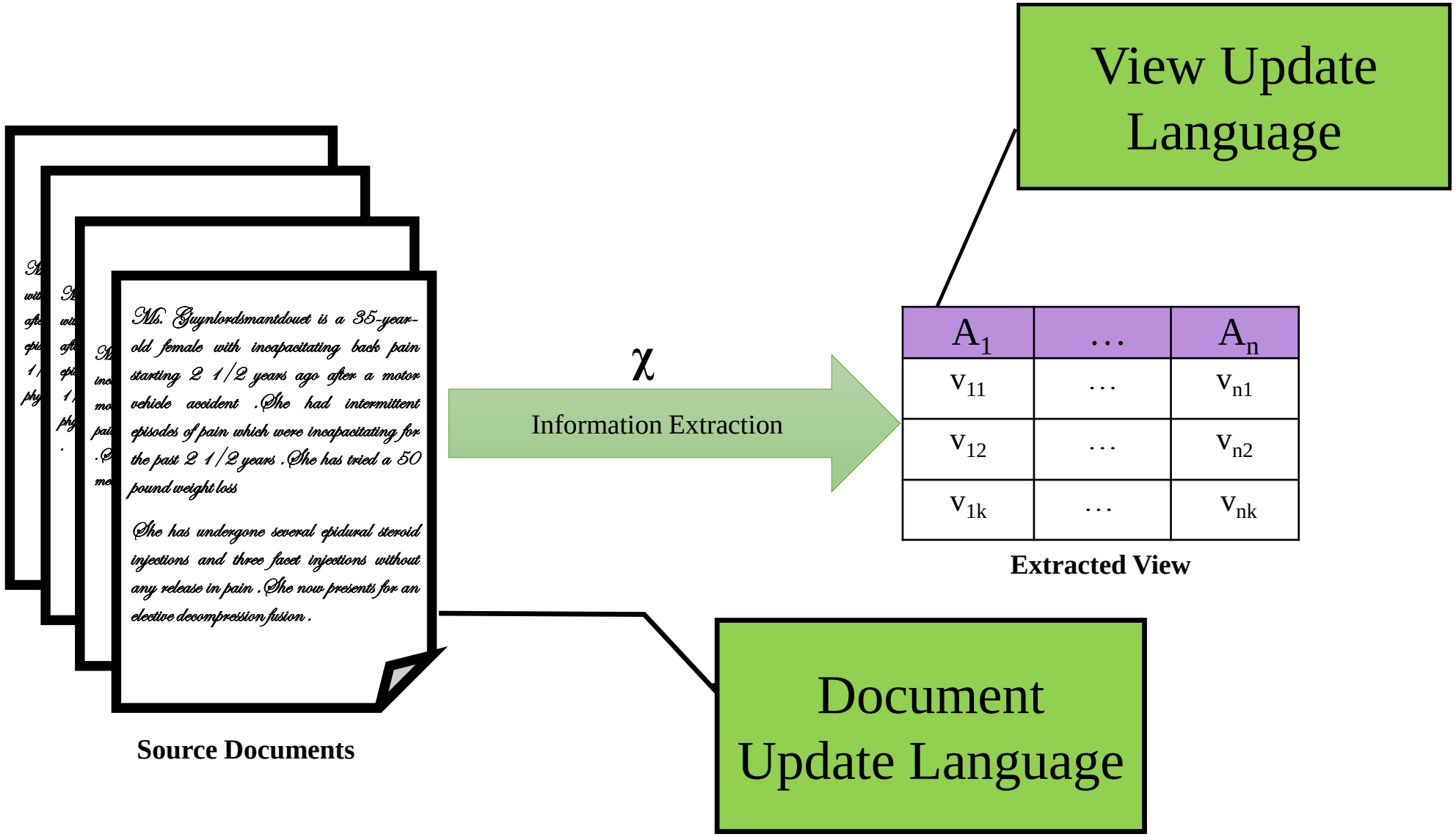
Problem Statement

Given an extraction specification and view/document update language how can source documents be updated to produce the modified view?

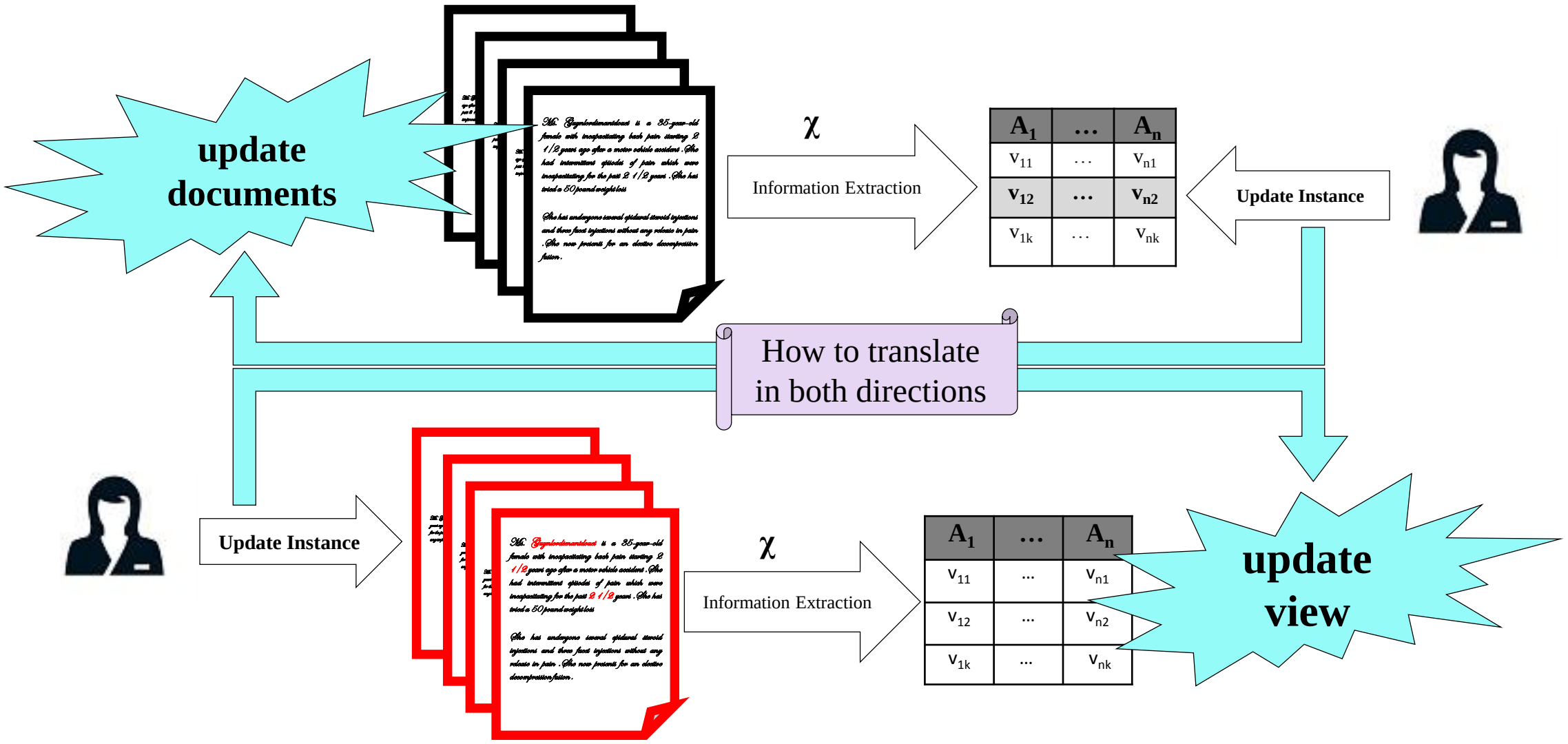
Extracted View Update Problem



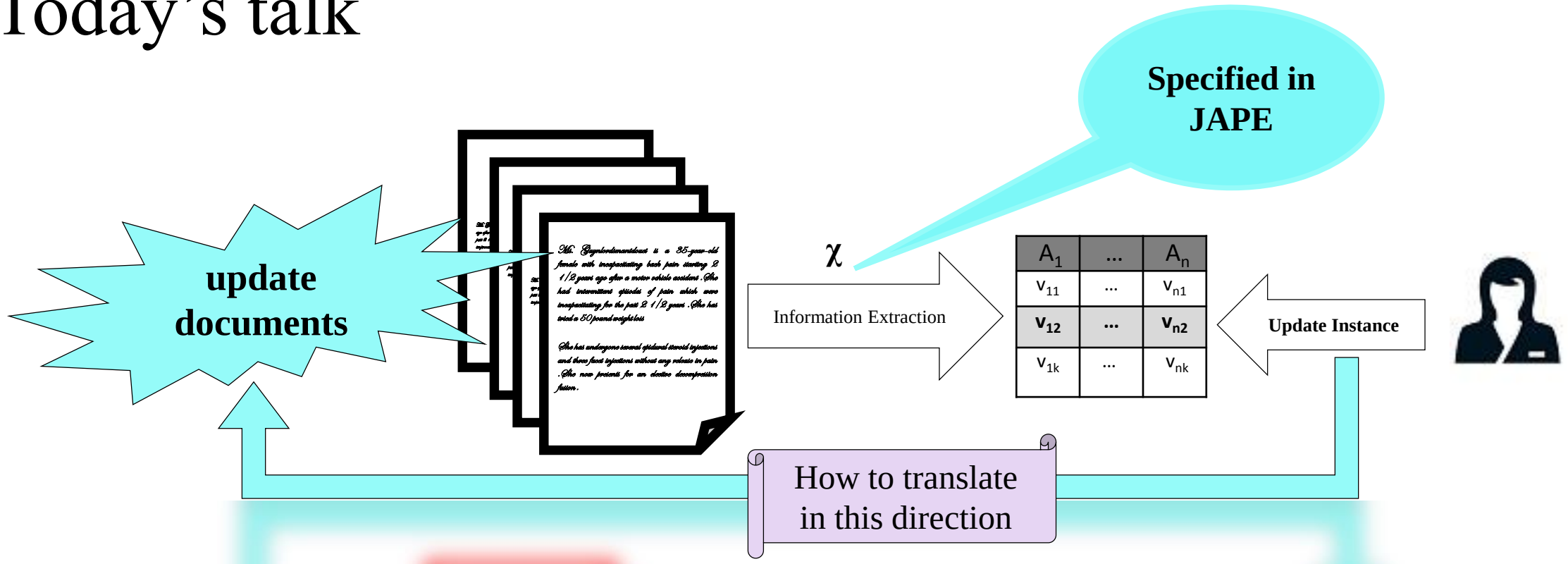
Information Extraction



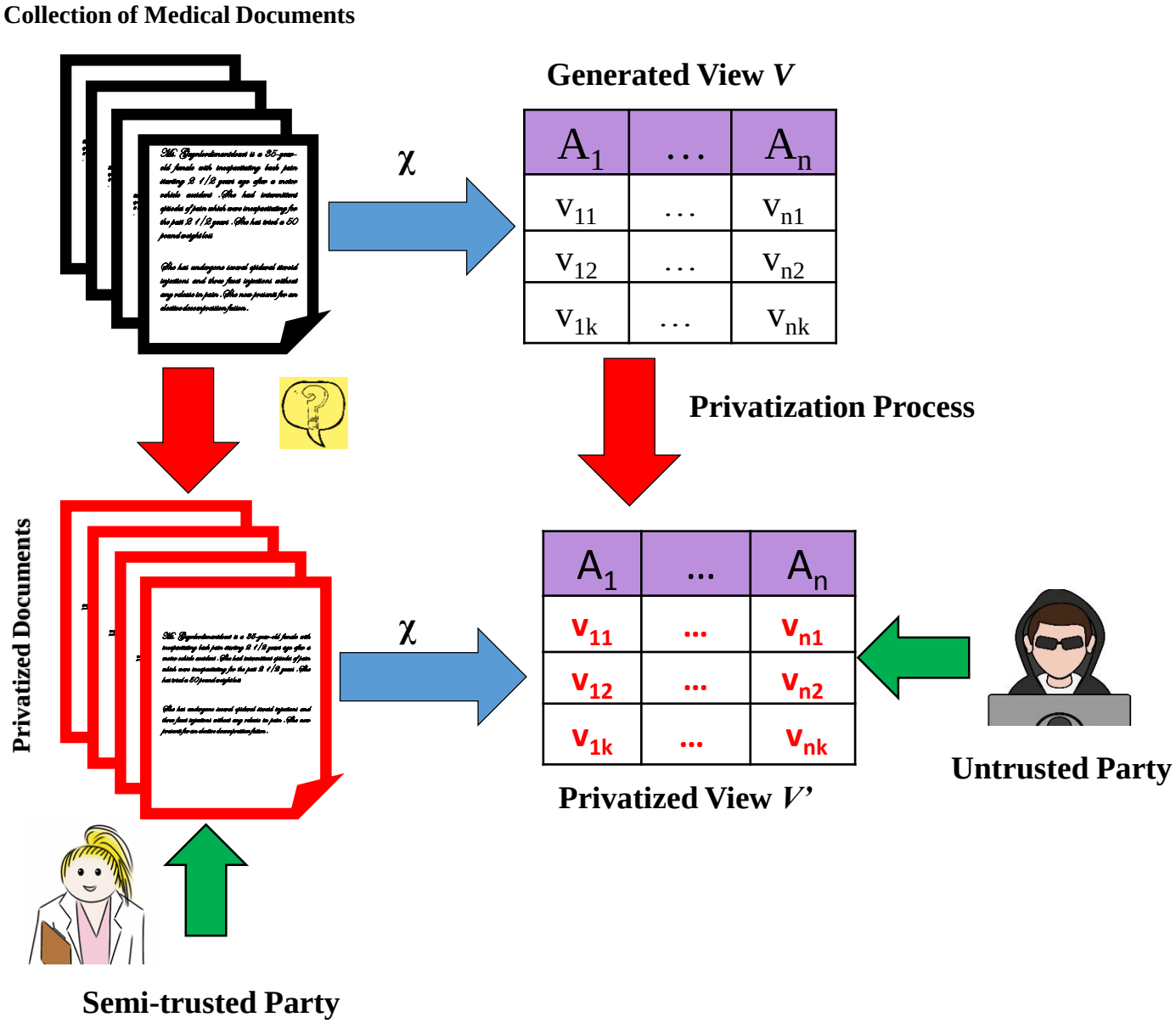
My research scope



Today's talk

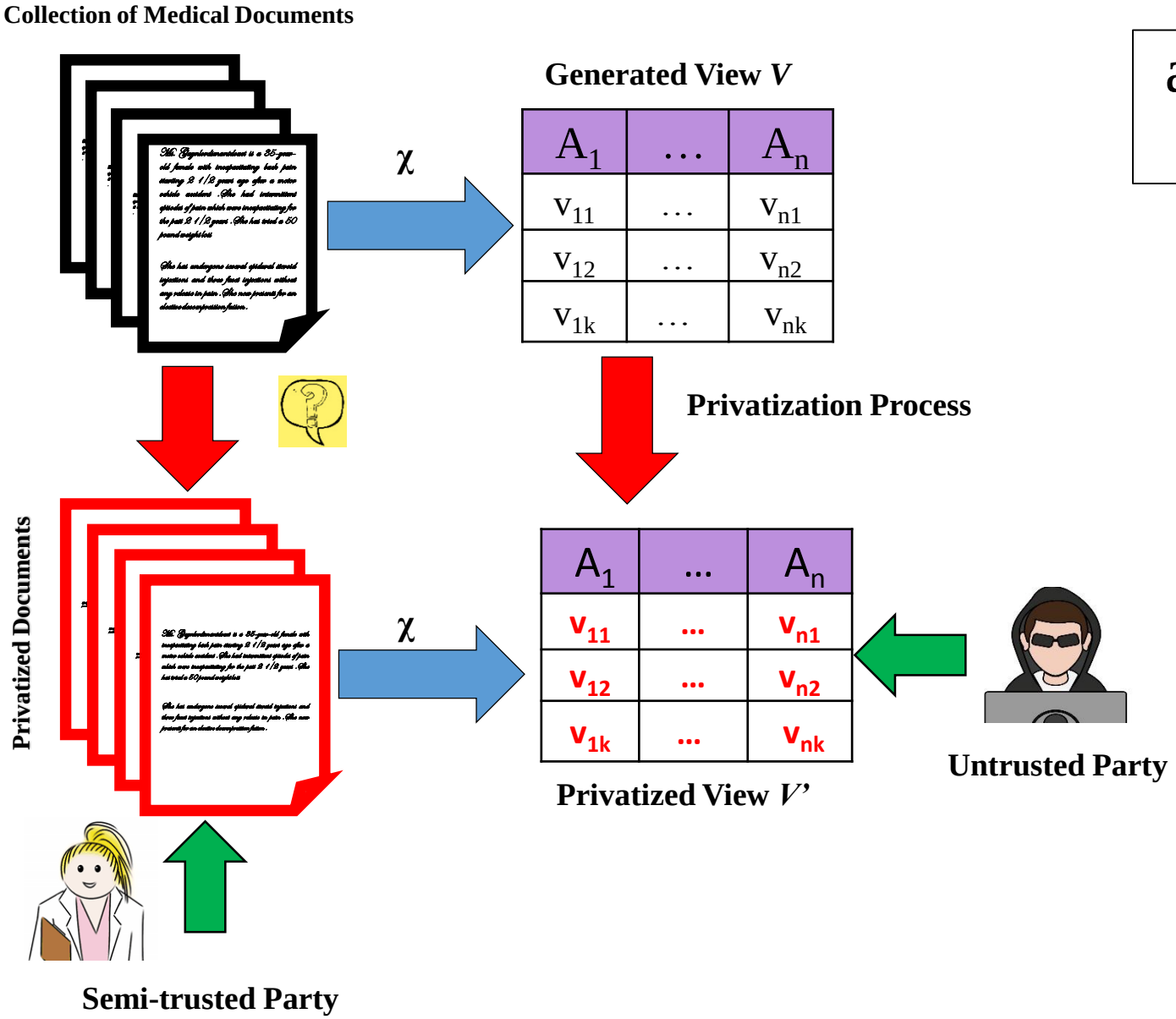

Specified in JAPE
update documents
Information Extraction
χ
A1 ... An
v11 ... vn1
v12 ... vn2
v1k ... vnk
Update Instance
How to translate in this direction

Motivation


Collection of Medical Documents
Generated View V
A1 ... An
v11 ... vn1
v12 ... vn2
v1k ... vnk
χ
Privatization Process
Privatized Documents
A1 ... An
v11 ... vn1
v12 ... vn2
v1k ... vnk
χ
Privatized View V'
Untrusted Party
Semi-trusted Party

Motivation

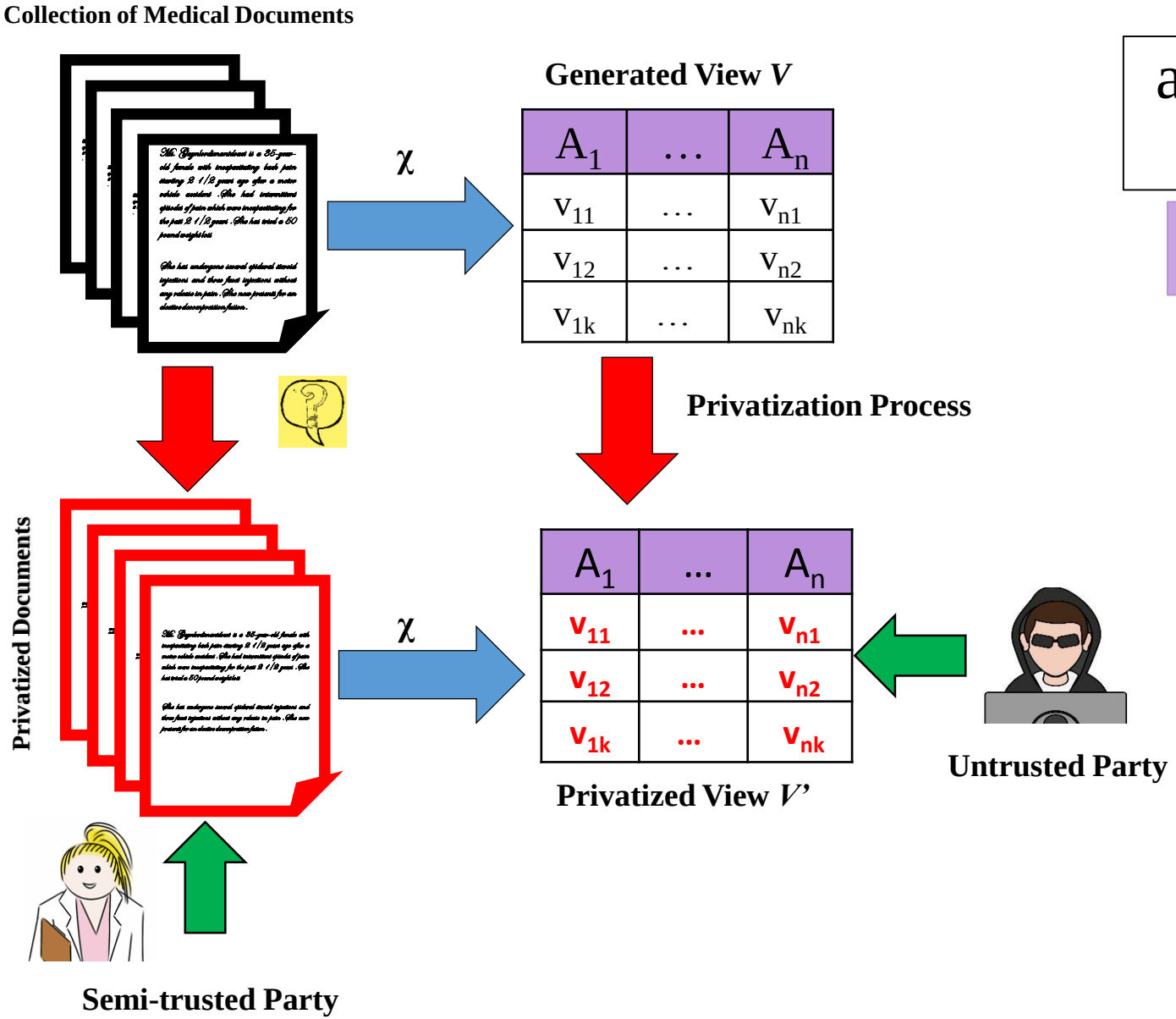
applying privacy transformations to medical documents



Motivation

applying privacy transformations to medical documents

Variant of Differential Privacy

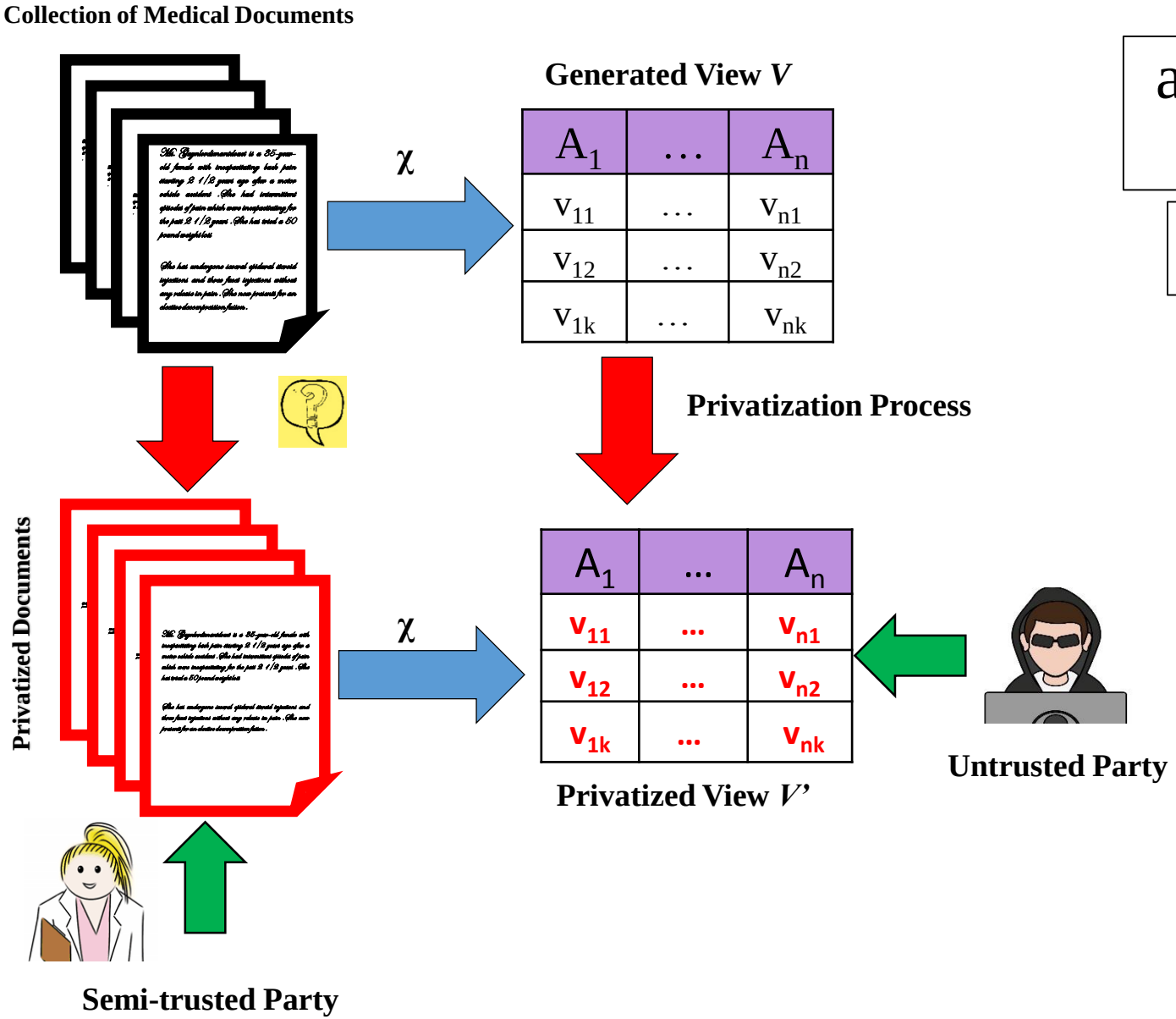


Motivation

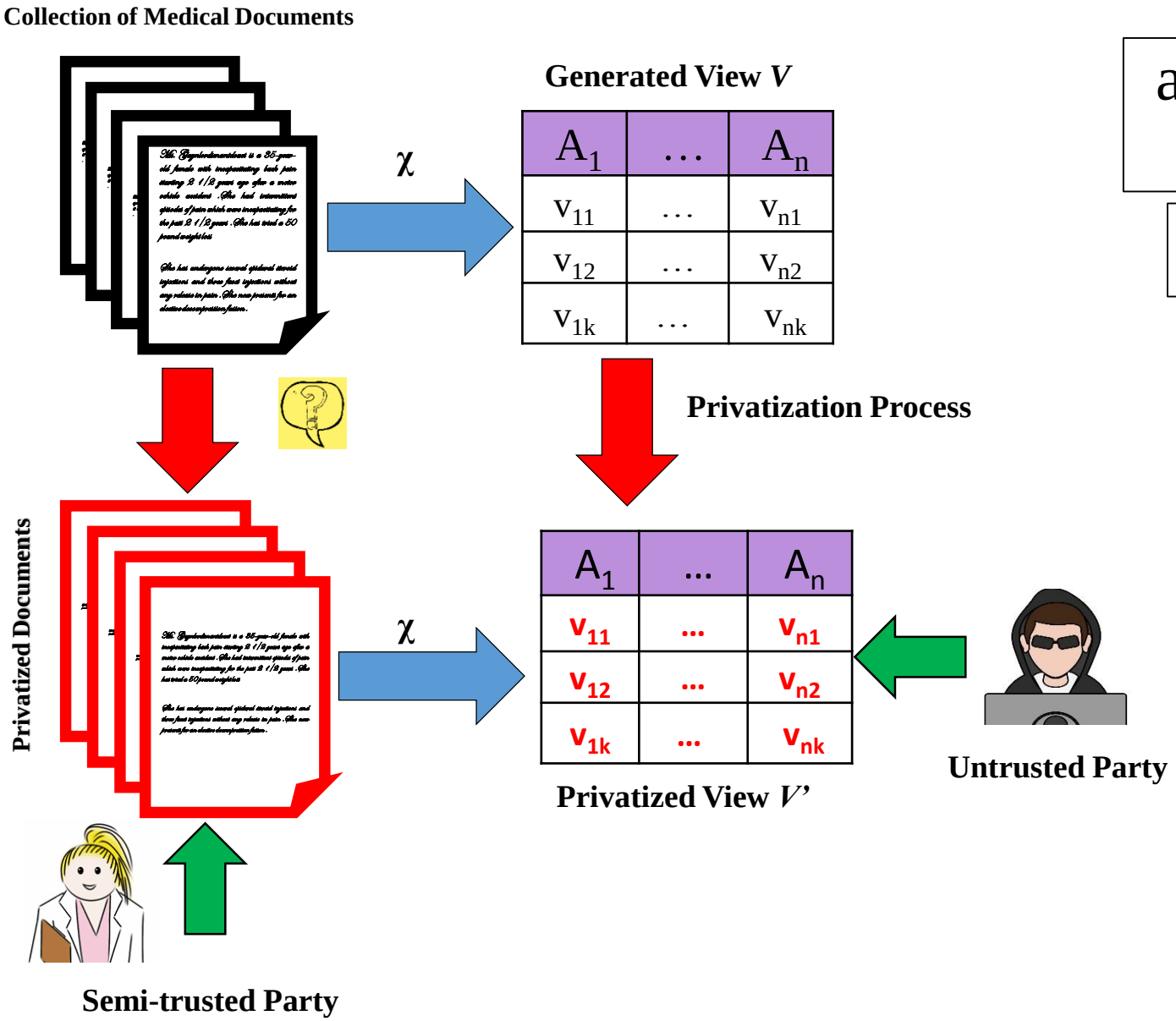
applying privacy transformations to medical documents

Variant of Differential Privacy

a randomized algorithm



Motivation



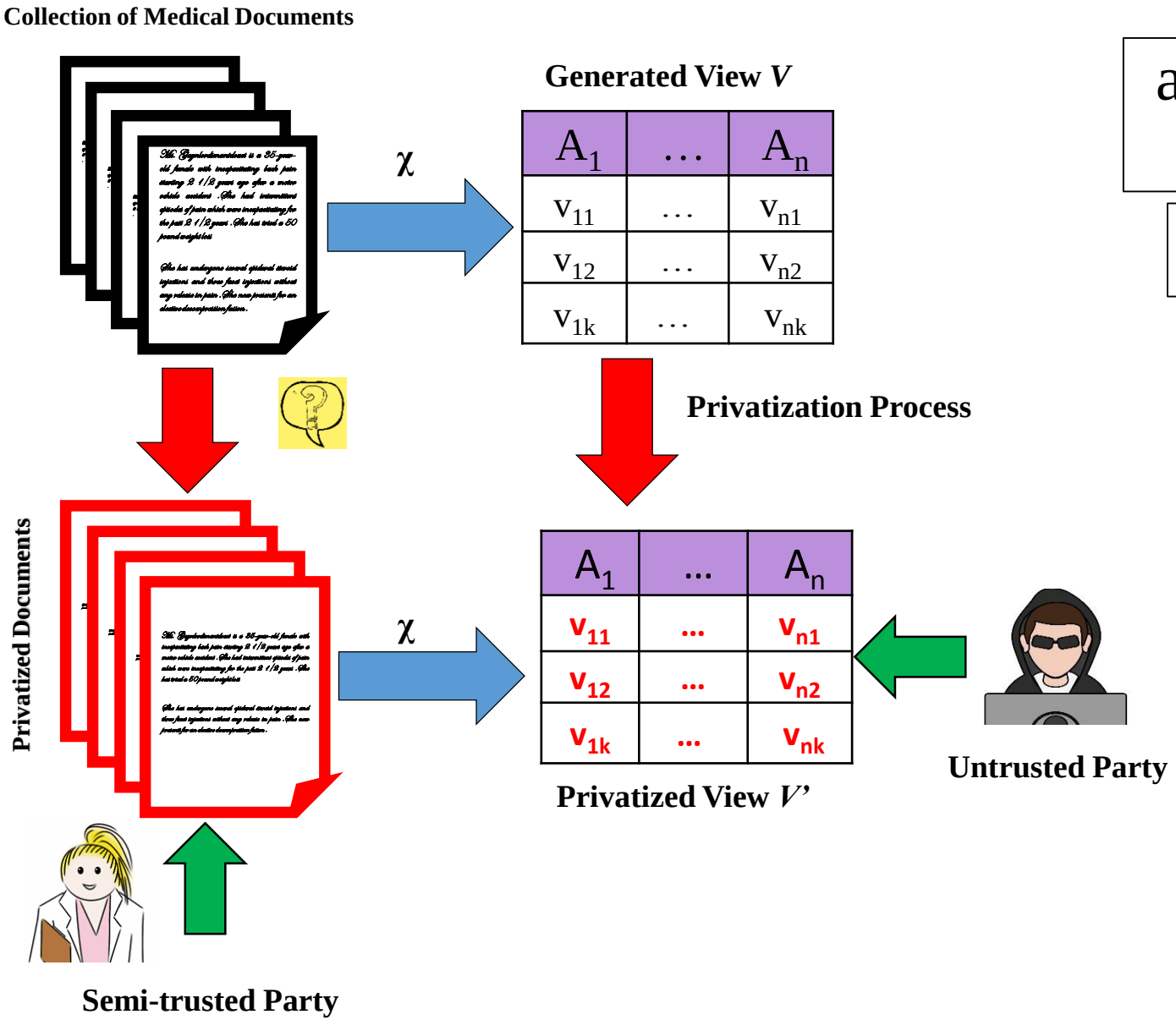
applying privacy transformations to medical documents

Variant of Differential Privacy

a randomized algorithm

maps records in table **T** to records in **T′**

Motivation



applying privacy transformations to medical documents

Variant of Differential Privacy

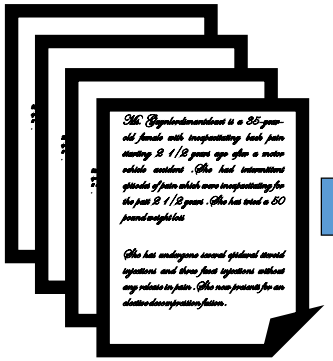
a randomized algorithm

maps records in table **T** to records in **T′**

T′ can be analyzed by untrusted parties without fearing the loss of privacy for individuals

Proposed Approach

Collection of Medical Documents



χ

Generated View *V*

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

Privatization Process

Privatized Documents

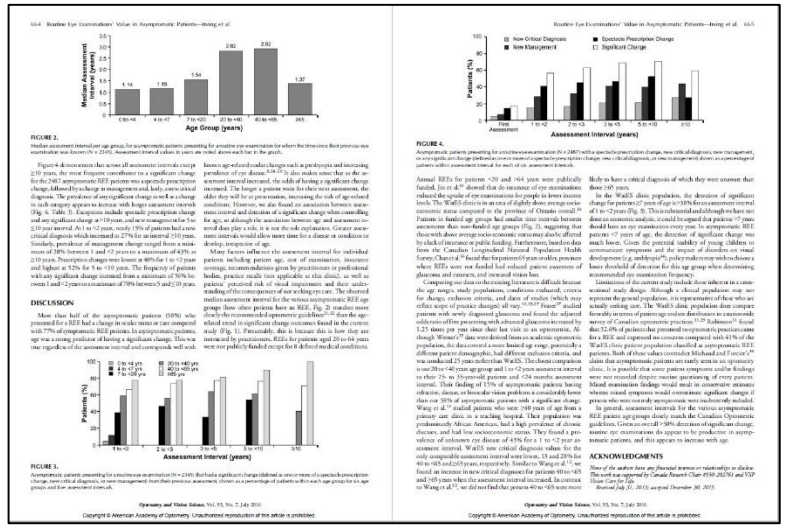
χ

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

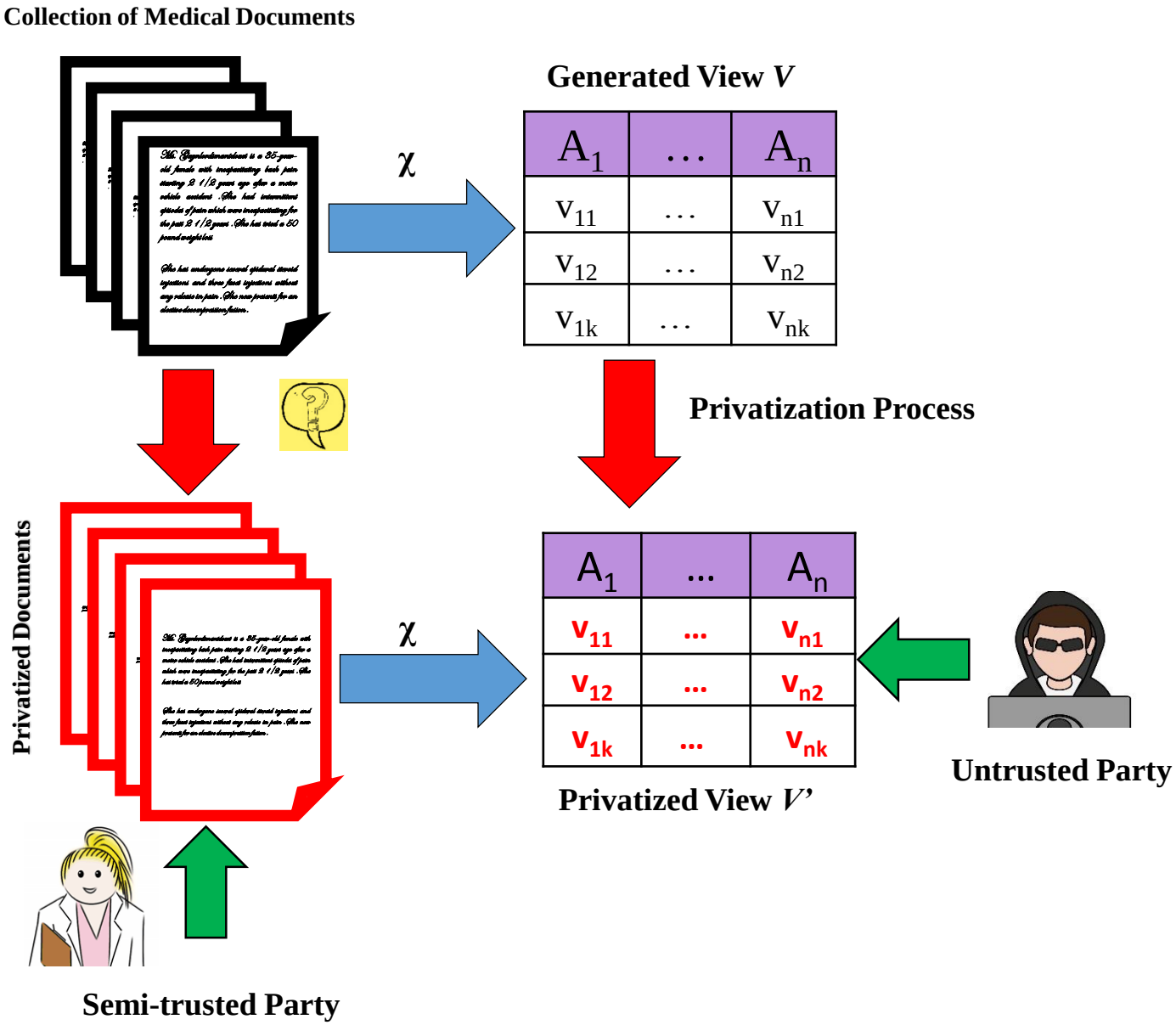
Privatized View *V'*

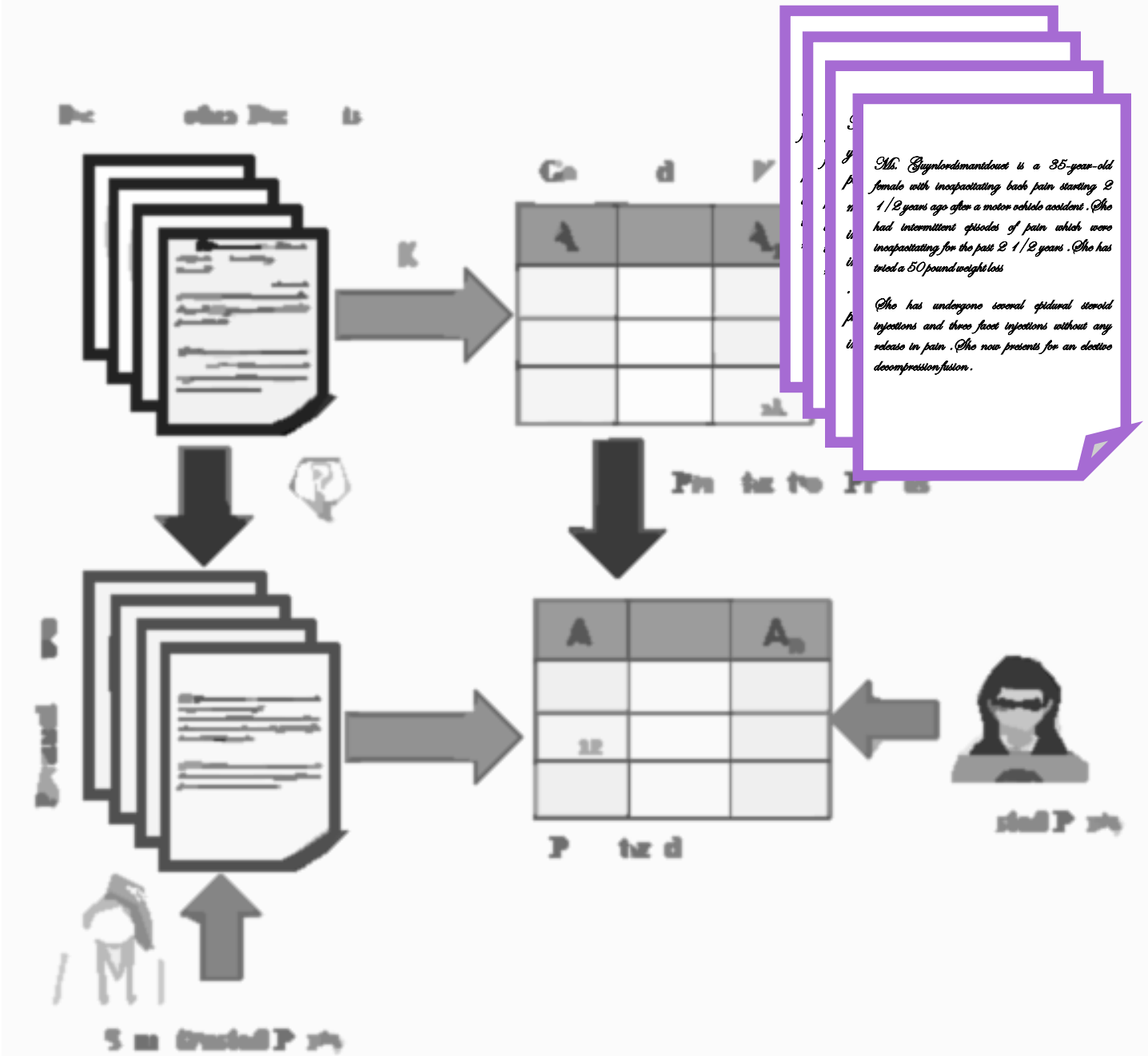
Untrusted Party

Semi-trusted Party



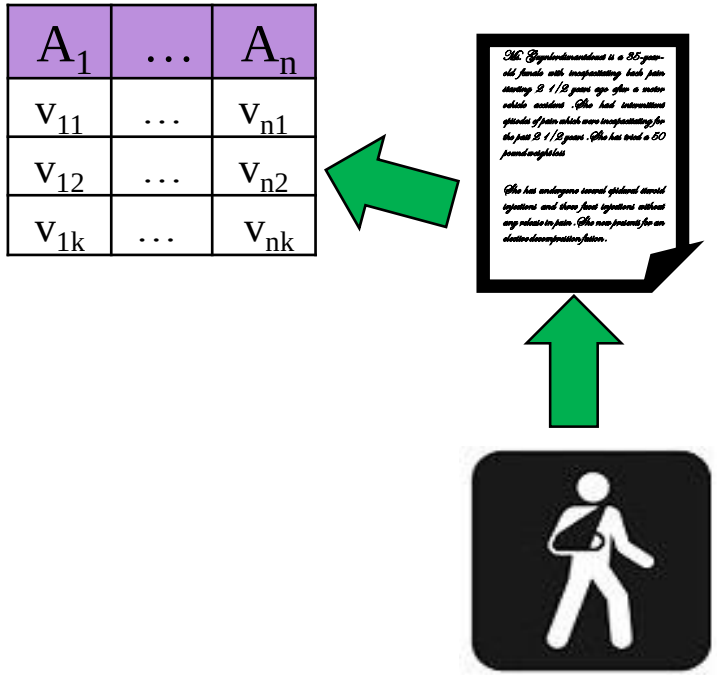
Assumptions



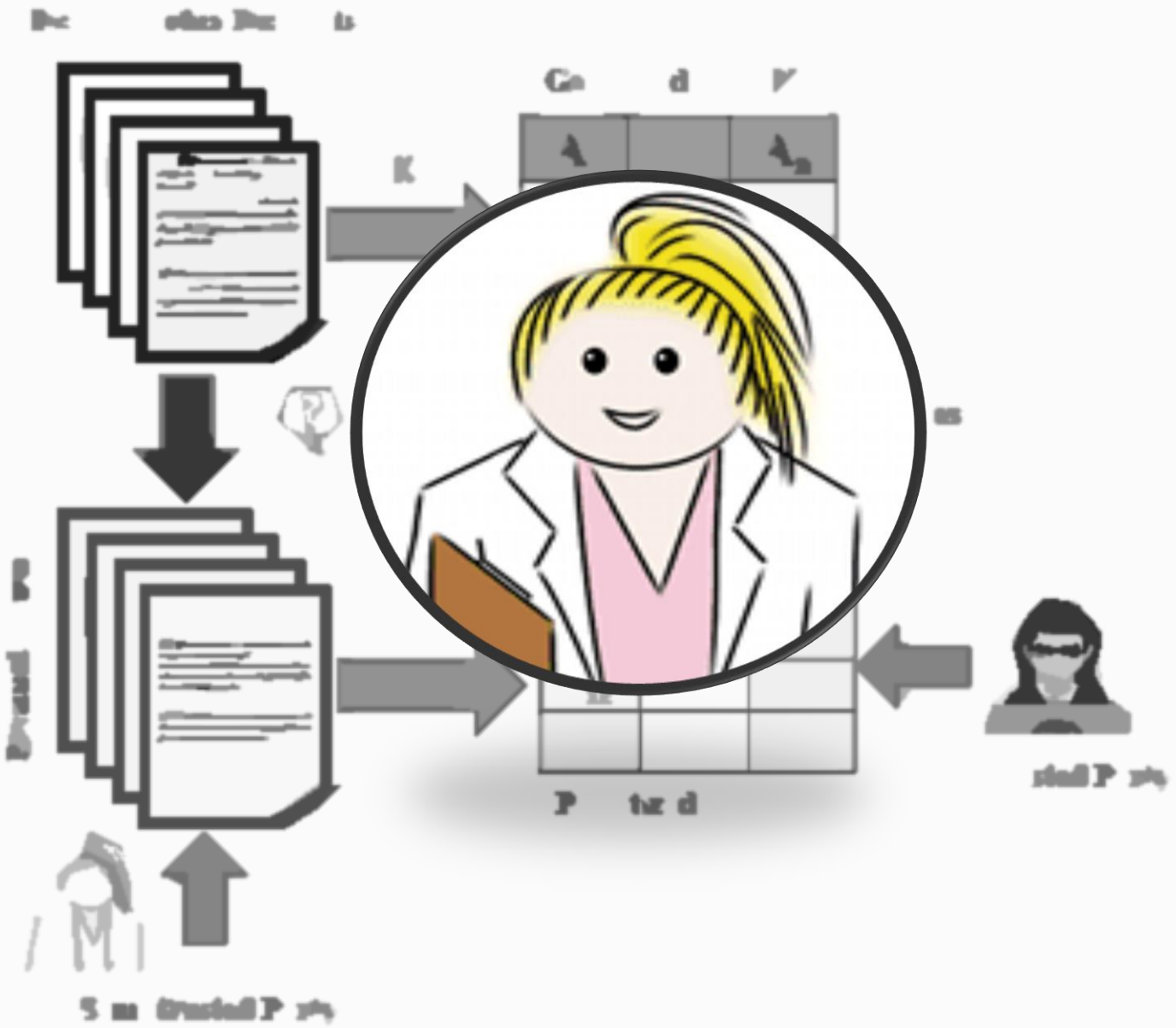


Assumptions

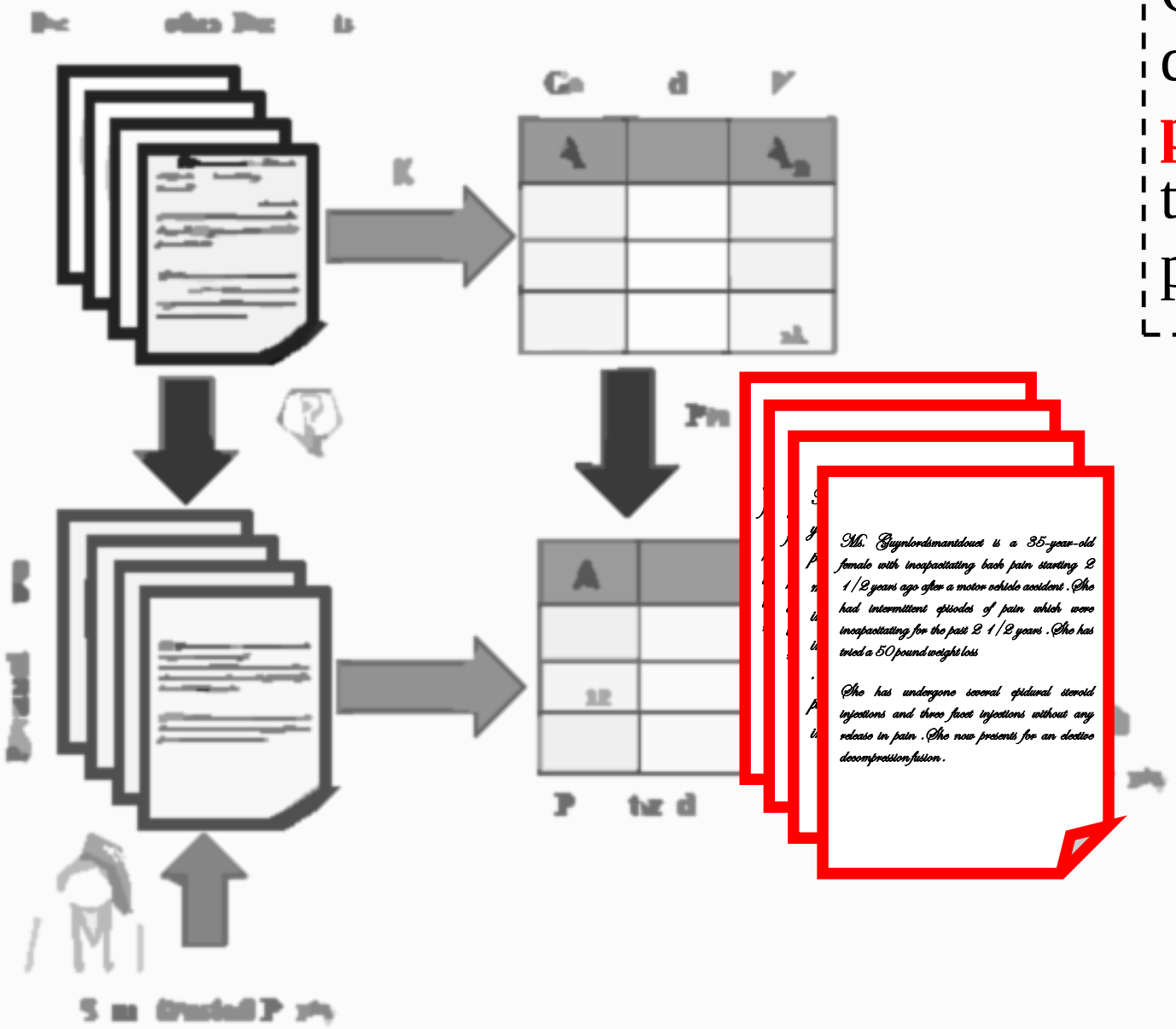
- Each document belongs to one individual
- Each document produces a single row



Assumptions



- Prepare tables for publication
- no wish to violate individuals' privacy, may do so unintentionally
- Want to read the documents to interpret and validate perturbed tables
- not experts on differential privacy



Our task is to prepare a set of documents that **would have produced** the modified table had the same information extraction procedure been applied.

Robust Extraction Algorithm

A given extraction algorithm is **robust** if purposeful modifications of text produce exactly the expected changes to the table with no other entries affected.

For all possible input documents, all entries in the extracted table, and all values in each entry's domain.

Characterization of Robust Information Extraction Programs

Notation and Terminology

W	set of all possible words
D	a sequence of words $D = \langle w_1, \ldots, w_{\mathcal{N}} \rangle$, $w_i \in W$
$D[a, b]$ or *span*	non-empty sequence of consecutive words $\langle w_a, \ldots, w_{a+m} \rangle$ $m \geq 0$, $b = a + m$ if $b < a$ represents an empty span
$\mathcal{D}$	set of all possible documents
$\mathcal{T}$-ary record	consists of $\mathcal{T}$ attributes, i^{th} attribute A_i has domain $W_i \subseteq \bigcup_{k=0}^{\infty} W^k$
T	$\mathcal{T}$-ary table, $T : (A_1 : W_1, \ldots, A_{\mathcal{T}} : W_{\mathcal{T}})$
$\mathcal{R}$	set of all possible records that could appear in T
$\mathcal{X} : \mathcal{D} \rightarrow \mathcal{R}$	extraction function

Notation and Terminology

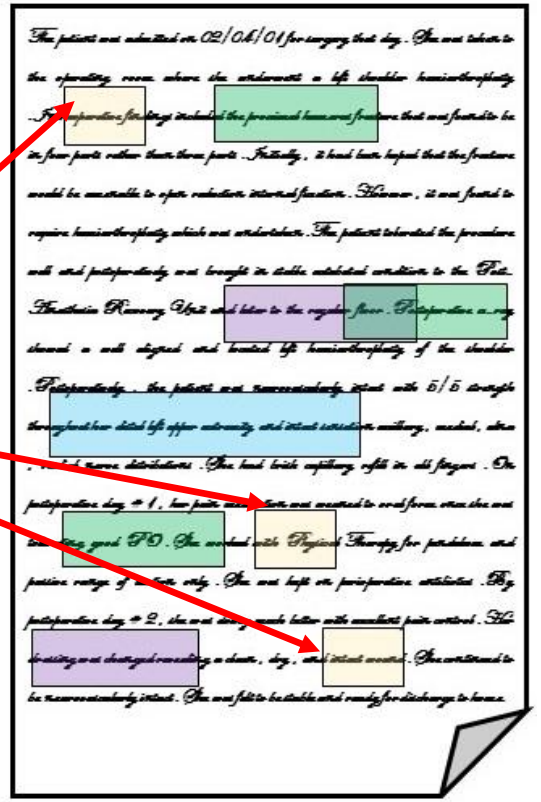
$\mathcal{F}$	indexed set of *domain preserving* functions $\mathcal{F} = \{f_i \mid f_i : W_i \to W_i\}$ $i \in [1 \ldots \mathcal{T}]$, W_i is the domain for attribute A_i
$F(r, j) = \langle v'_1, \ldots, v'_\mathcal{T} \rangle$,	$r = \langle v_1, \ldots, v_\mathcal{T} \rangle$ $v'_k = \begin{cases} f_k(v_k) & \text{if } k = j, \\ v_k & \text{otherwise.} \end{cases}$
$F(r) = \langle f_1(v_1), \ldots, f_\mathcal{T}(v_\mathcal{T}) \rangle$	

Properties of Extractors

Strict Extractor

For every possible input document, the set of extracted values in the corresponding record is 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



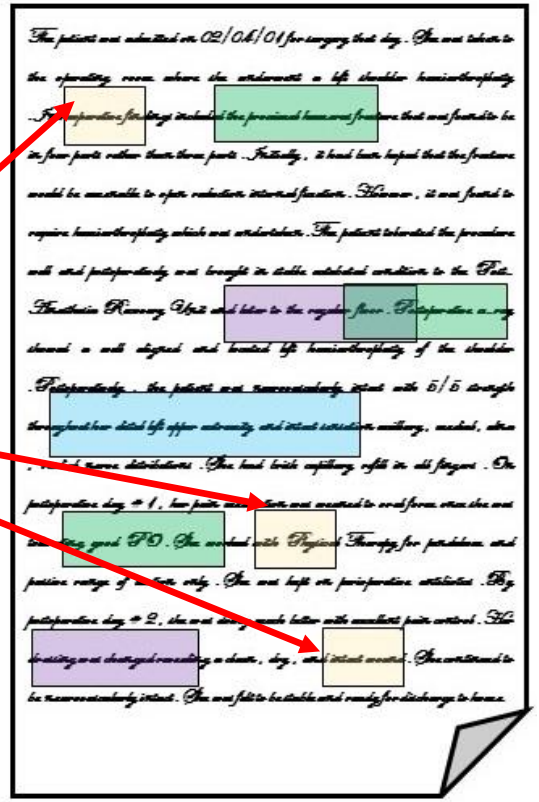
Strict Extractor

Formally, $\mathcal{X} : \mathcal{D} \rightarrow \mathcal{R}$ is *strict* if $\forall D = \langle w_1, \ldots, w_{\mathcal{N}} \rangle \in \mathcal{D}$,

$$\mathcal{X}(D) = \langle v_1, \ldots v_{\mathcal{T}} \rangle \implies$$

$$\{v_1, \ldots, v_{\mathcal{T}}\} \subseteq \{D[a, b] \mid 1 \leq a \leq b \leq \mathcal{N}\}.$$

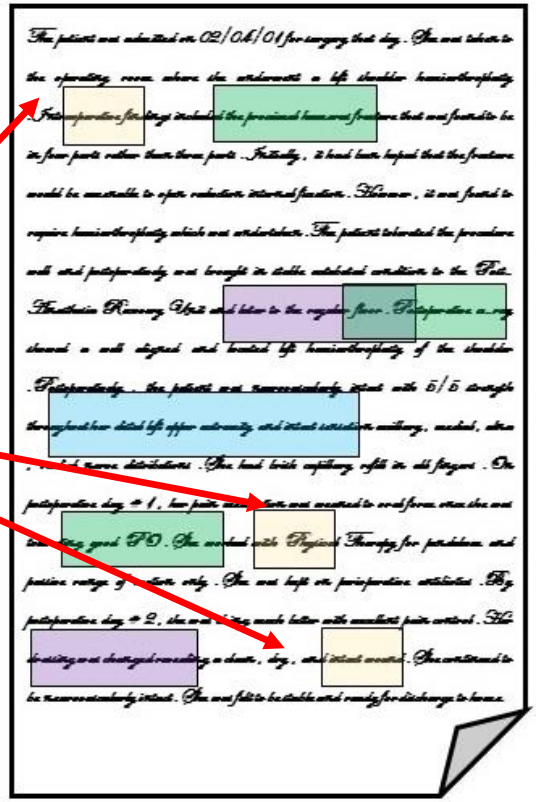
$A_1 : W_1$	$A_2 : W_2$	$A_3 : W_3$	$A_4 : W_4$
v_1	v_2	v_3	v_4



Strict Extractor

If $\mathcal{X}$ is a strict extractor, we use $P_{\mathcal{X}}(D, j)$ to denote the span(s) in input document $D \in \mathcal{D}$ from which v_j is extracted.

$P_{\mathcal{X}}(D, j)$

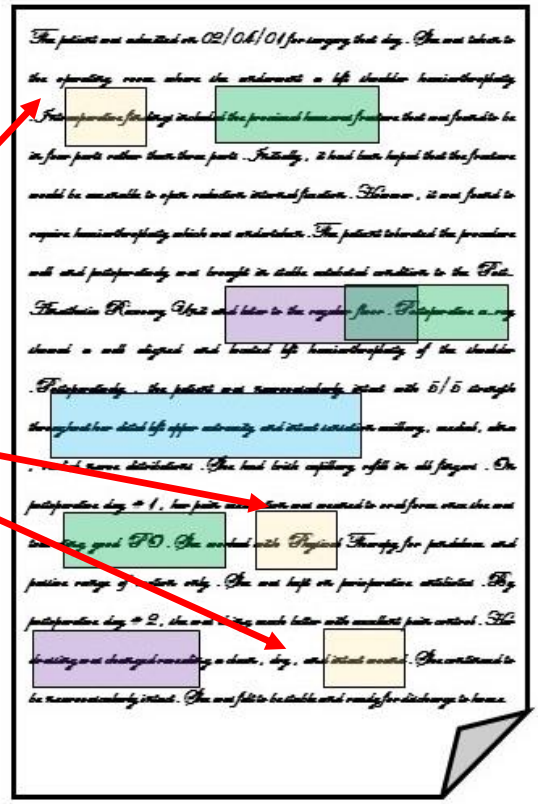


Strict Extractor

If $\mathcal{X}$ is a strict extractor, we use $P_{\mathcal{X}}(D, j)$ to denote the span(s) in input document $D \in \mathcal{D}$ from which v_j is extracted.

Data Lineage

$P_{\mathcal{X}}(D, j)$

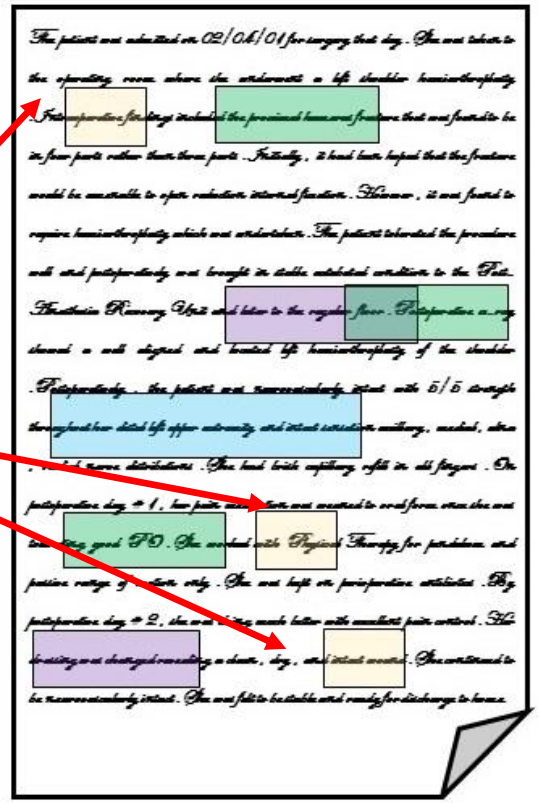


We assume that the spans in $P_{\mathcal{X}}(D, j)$ do not overlap.

Strict Extractor

If $\mathcal{X}$ is a strict extractor, we use $P_{\mathcal{X}}(D, j)$ to denote the span(s) in input document $D \in \mathcal{D}$ from which v_j is extracted.

$P_{\mathcal{X}}(D, j)$

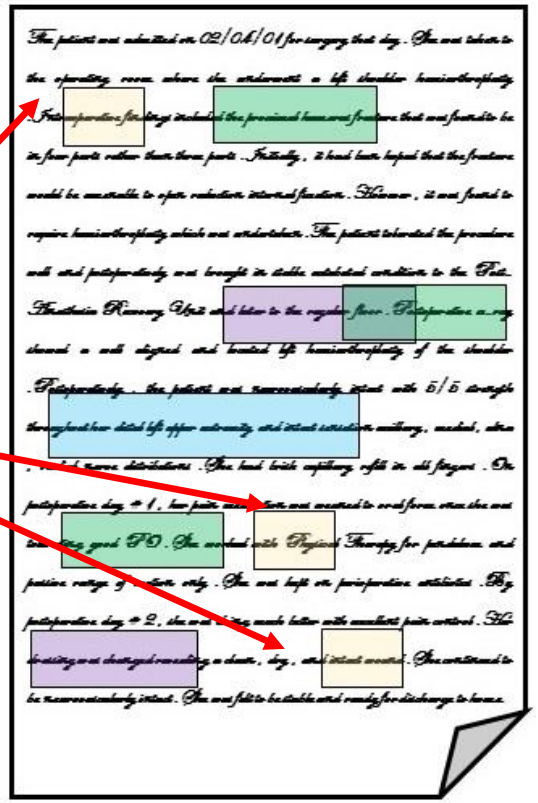


We assume that the spans in $P_{\mathcal{X}}(D, j)$ do not overlap.

Computable Extractor

A strict extractor is computable if for all possible input documents and corresponding extracted attributes, we have access to positions from which the attributes are extracted.

$$P_{\mathcal{X}}(D, j)$$

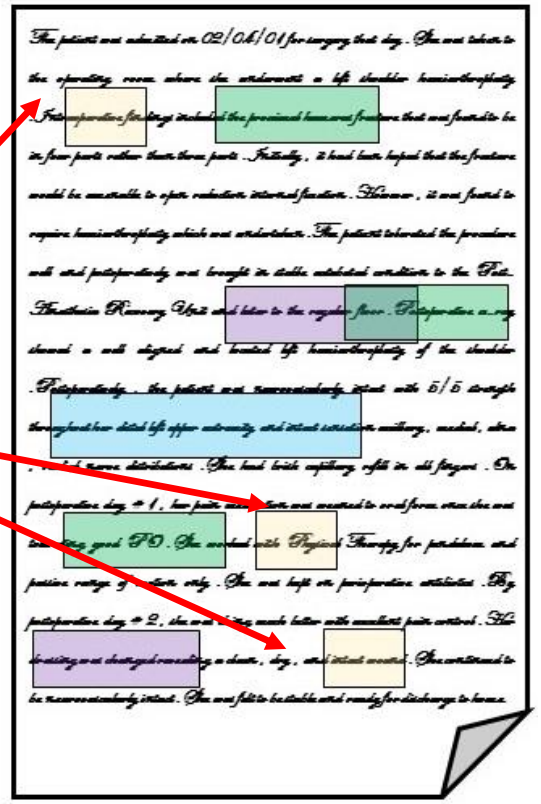


Data Lineage

Computable Extractor

Formally, $\forall j \in [1 \ldots \mathcal{T}]$ and $\forall D \in \mathcal{D}$, $\mathcal{X}$ is *computable* if $P_{\mathcal{X}}(D, j)$ is explicit (given) or provided as a side-effect of running the extractor on D.

$$P_{\mathcal{X}}(D, j)$$



Data Lineage

Stable Extractor

Let $\mathcal{X}$ be a strict and computable extractor

$D = \langle w_1, \ldots, w_{\mathcal{N}} \rangle$ be a document

A_j be an extracted attribute with value v_j

$P_{\mathcal{X}}(D, j) = \{\langle a_i, b_i \rangle\}$ be the set spans from which v_j is extracted

Assume that $1 \leq a_1 \leq b_1 < a_2 \leq b_2 < \ldots < a_k \leq b_k \leq \mathcal{N}$

Stable Extractor

We define a modified document as:

$$g(D,j) = D[1, a_1 - 1] \bullet f_j(v_j) \bullet D[b_1 + 1, a_2 - 1] \bullet f_j(v_j) \bullet$$
$$D[b_2 + 1, a_3 - 1] \bullet f_j(v_j) \bullet \cdots \bullet D[b_k + 1, \mathcal{N}]$$

Stable Extractor

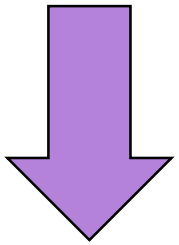
An information extraction algorithm is *stable* if

$\forall D \in \mathcal{D}$ and $\forall j \in [1 \ldots \mathcal{T}]$ we have $\mathcal{X}(D) = r \implies \mathcal{X}(g(D, j)) = F(j, r)$.

Stable Extractor

An information extraction algorithm is *stable* if

$\forall D \in \mathcal{D}$ and $\forall j \in [1 \ldots \mathcal{T}]$ we have $\mathcal{X}(D) = r \implies \mathcal{X}(g(D, j)) = F(j, r)$.



changing values in appropriate positions in a document affects only the expected attribute in the extracted record.

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
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 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

Robust Extraction Algorithm

Proposition. *For any strict, computable, and stable extractor* $\mathcal{X} : \mathcal{D} \to \mathcal{R}$, *there exists an algorithm* $A(\mathcal{F}, D, P_{\mathcal{X}}(D, j))$ *such that for all indexed sets of domain preserving functions* $\mathcal{F} = \{f_i | f_i : W_i \to W_i,$ *where* $i \in [1 \ldots \mathcal{T}]\}$ *and any document* $D \in \mathcal{D}$, $A(\mathcal{F}, D, P_{\mathcal{X}}(D, j))$ *produces* $D^{\mathcal{P}}_{\mathcal{F}}$ *in such way that* $F(\mathcal{X}(D)) = \mathcal{X}(D^{\mathcal{P}}_{\mathcal{F}})$.

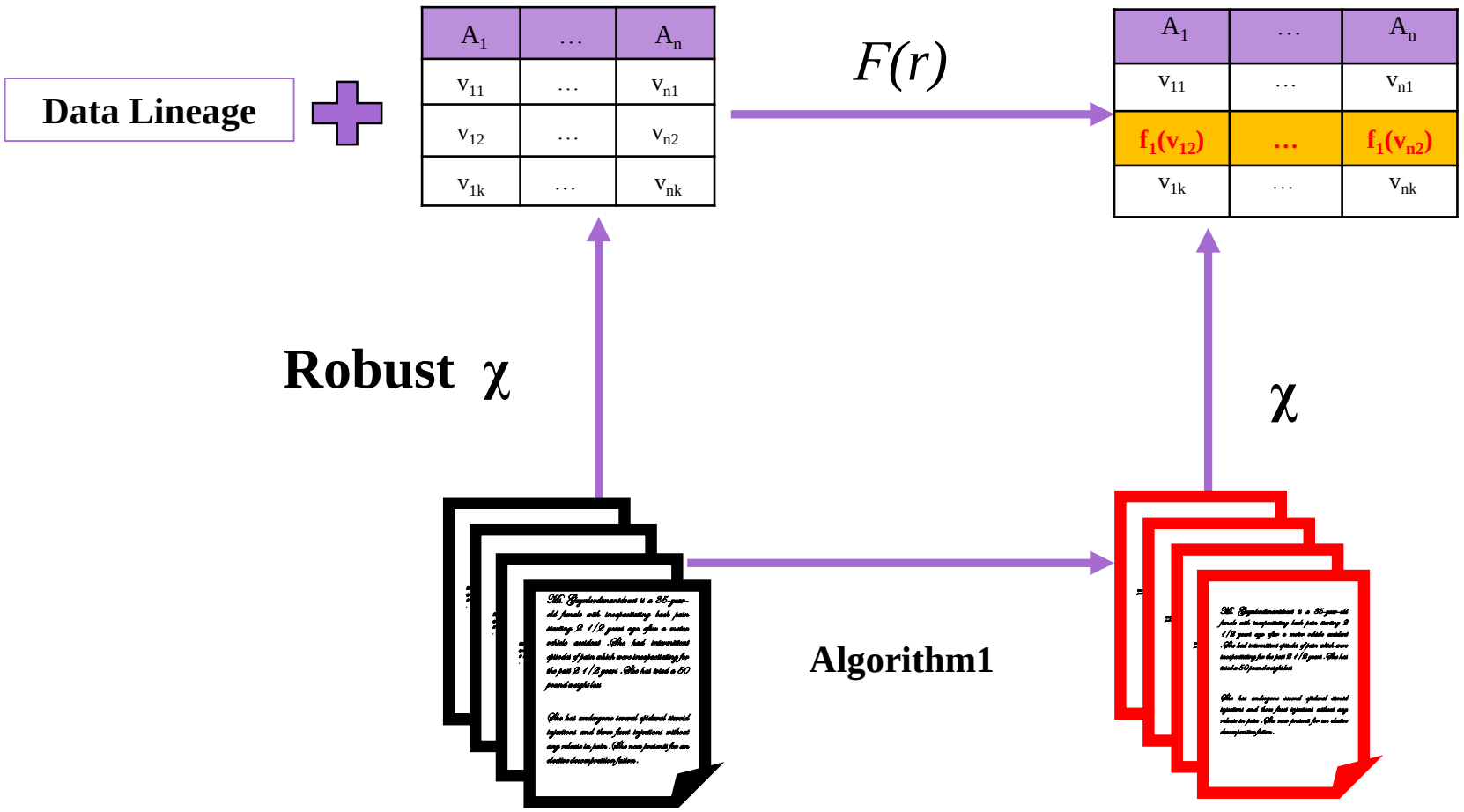
Robust Extraction Algorithm

Claim. *For any information extraction algorithm $\mathcal{X}$ having the aforementioned properties, Algorithm 1 produces $D^{\mathcal{P}}_{\mathcal{F}}$ in such a way that $F(\mathcal{X}(D)) = \mathcal{X}(D^{\mathcal{P}}_{\mathcal{F}})$.*

Input: $\mathcal{F}$, D, $j \rightarrow P_{\mathcal{X}}(D, j)$
Output: $D^{\mathcal{P}}_{\mathcal{F}}$
$D^{\mathcal{P}}_{\mathcal{F}} \leftarrow D$
for $j \in [1 \ldots \mathcal{T}]$ **do**
  for $\langle a, b \rangle \in P_{\mathcal{X}}(D, j)$ **do**
    replace $D[a, b] \in D^{\mathcal{P}}_{\mathcal{F}}$ by $f_j(D[a, b])$
  end
end
return $D^{\mathcal{P}}_{\mathcal{F}}$

Algorithm 1: Updating a document.

Extracted View Update Problem

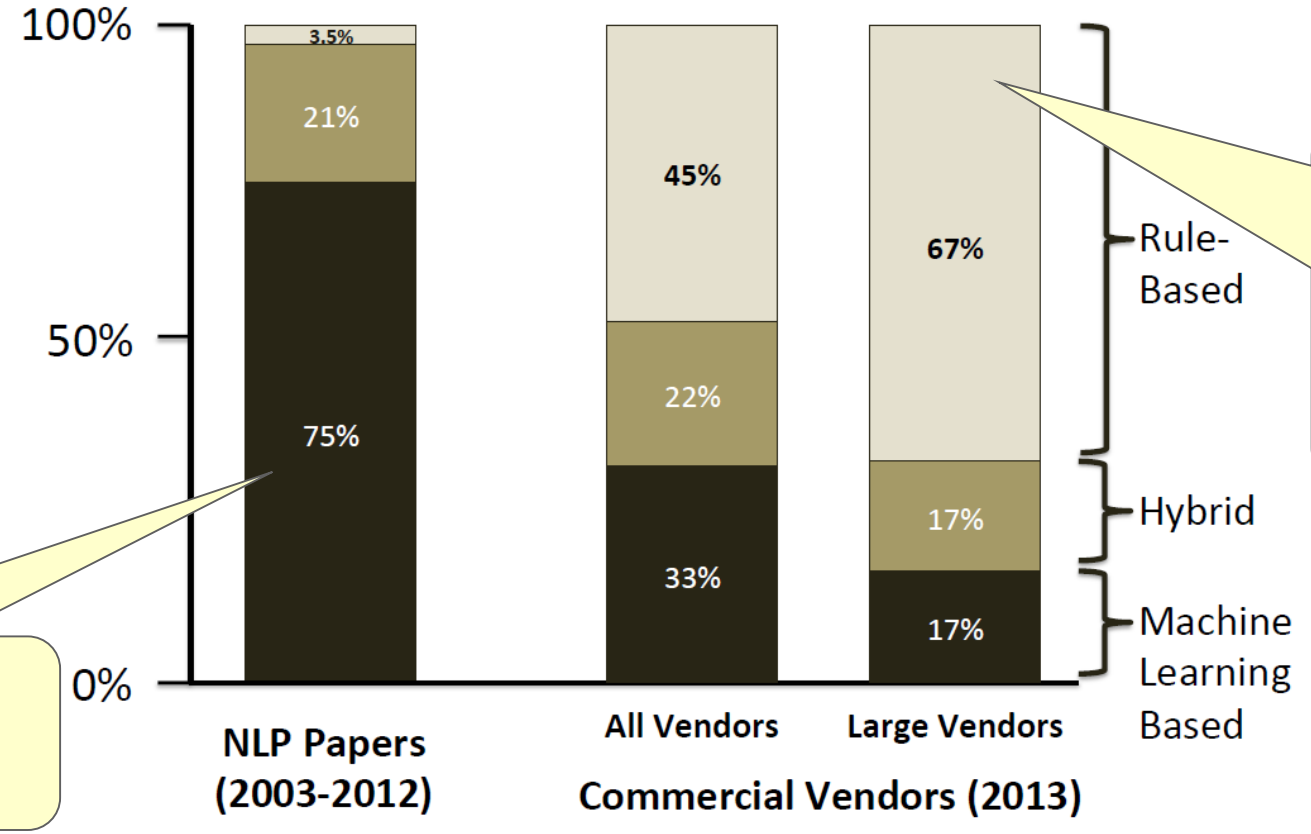


Verification


EXTRACTION PROGRAM
VIEW UPDATE INSTANCE
PROPERTY VERIFICATION PROCESS
ROBUST?

Rule-based vs. ML

- Entity extraction
- EMNLP, ACL, NAACL, 2003-2012
- 54 industrial vendors (Text Analytics, 2012)

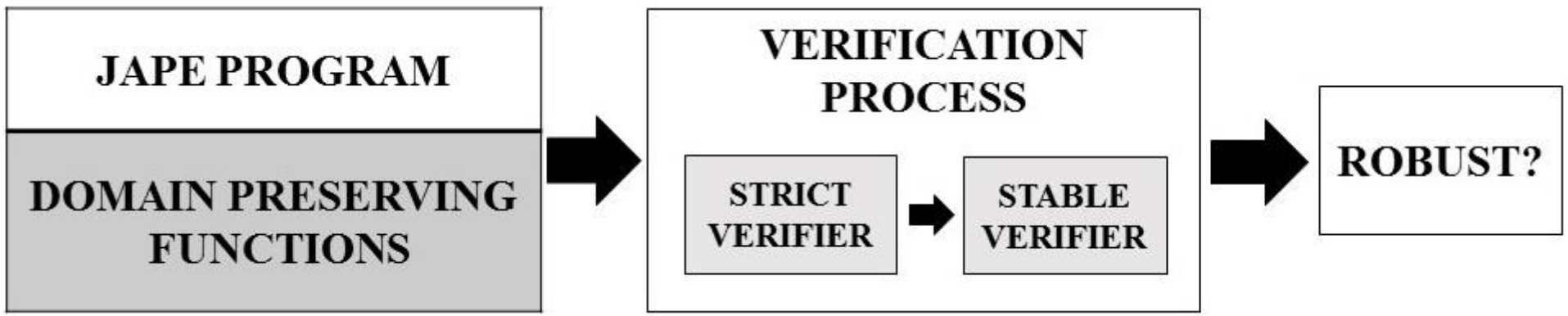


Fast development, fast adaptation, better results in limited time, sophistication

Easy to comprehend, maintain, debug & optimize performance; lower reliance on labeled data

[adopted from Chiticariu, Li, Reiss, EMNLP' 13]

Verification OF JAPE Programs

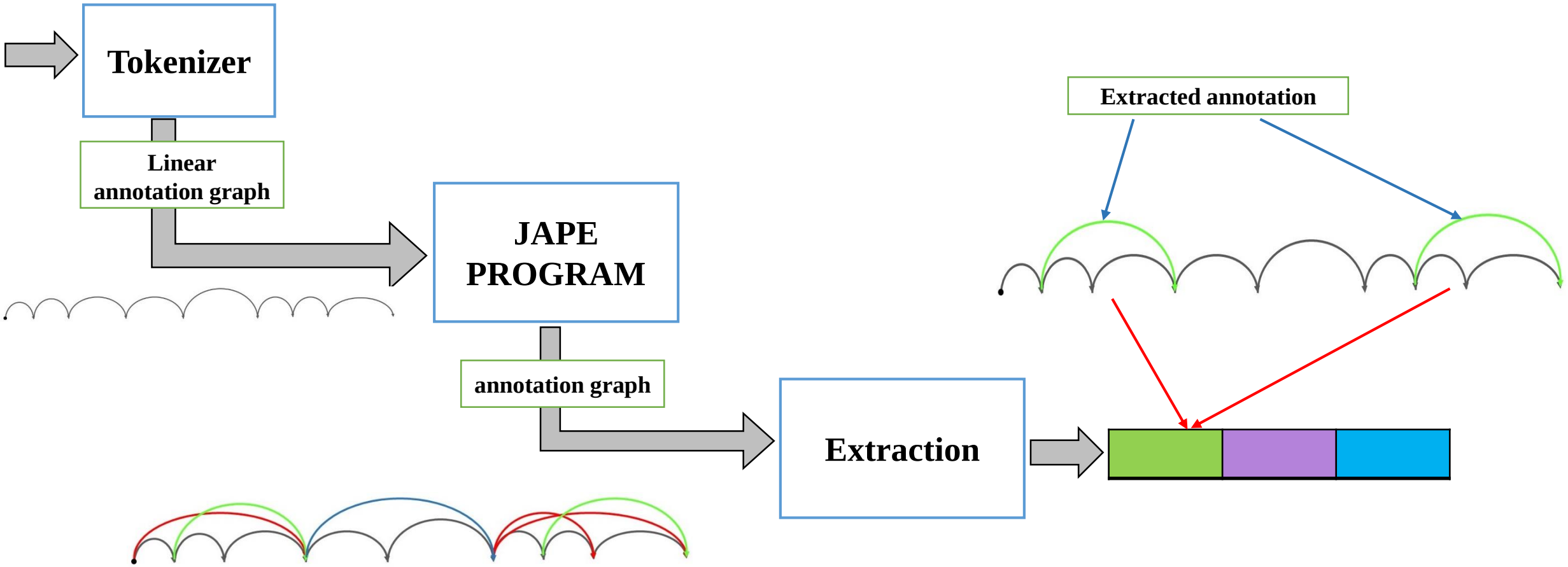

JAPE PROGRAM
DOMAIN PRESERVING FUNCTIONS
VERIFICATION PROCESS
STRICT VERIFIER
STABLE VERIFIER
ROBUST?

OVERVIEW OF JAPE

Commonly used rule-based information extraction system

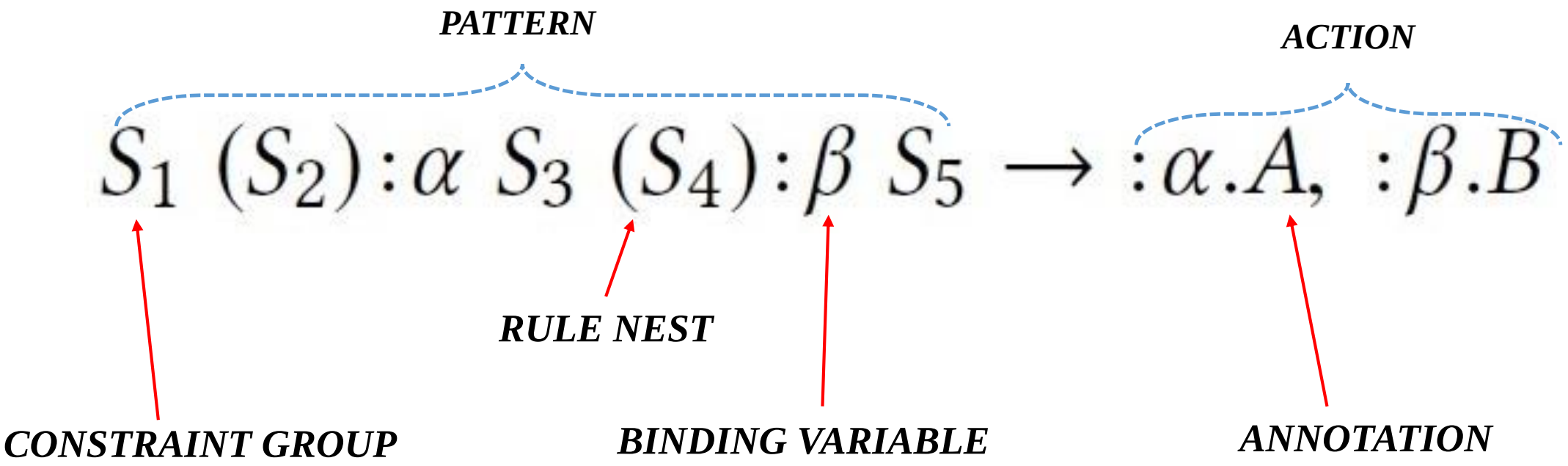
Rules are written in JAPE Language

JAPE Running Environment (simplified)



OVERVIEW OF JAPE

Simple JAPE Rule



OVERVIEW OF JAPE

$$S_1\ (S_2):\alpha\ S_3\ (S_4):\beta\ S_5 \rightarrow :\alpha.A,\ :\beta.B$$

PATTERN

describes a regular expression over annotations

consumes the input and assign spans of text to binding variables

OVERVIEW OF JAPE

```
Phase: p1
Input: Token
Options: control = appelt

Rule: name
({Token.orth==upperInitial})+:mark --> :mark.Name={rule="name"}

Rule: addr
({Token}{Token.string=="@"}{Token}{Token.string =="."}{Token}):mark -->
:mark.Addr={rule="addr"}

Rule: email
({Token.string =="email"}{Token.string =="address"}):mark -->
:mark.Email={rule="email"}

Rule: otherLower
({Token.orth==lowercase}):mark --> :mark.Lower={rule="otherLower"}
```

```
Phase: p2
Input: Lower Name Email Addr
Options: control = first

Rule: hasEmail
({Name}):person ({Lower})* {Email} ({Lower})* ({Addr}):contact -->
:person.Person= {rule = "hasEmail"}, :contact.Contact= {rule = "hasEmail"}

Rule: emailFor
({Addr}):contact ({Lower})* {Email} ({Lower})* ({Name}):person -->
:person.Person= {rule = "emailFor"}, :contact.Contact= {rule = "emailFor"}
```

Input

Only those edges in the annotation graph that are labelled by input types are visible to the rules in each phase.

OVERVIEW OF JAPE

```
Phase: p1
Input: Token
Options: control = appelt

Rule: name
({Token.orth==upperInitial})+:mark --> :mark.Name={rule="name"}

Rule: addr
({Token}{Token.string=="@"}{Token}{Token.string =="."}{Token}):mark -->
:mark.Addr={rule="addr"}

Rule: email
({Token.string =="email"}{Token.string =="address"}):mark -->
:mark.Email={rule="email"}

Rule: otherLower
({Token.orth==lowercase}):mark --> :mark.Lower={rule="otherLower"}
```

```
Phase: p2
Input: Lower Name Email Addr
Options: control = first

Rule: hasEmail
({Name}):person ({Lower})* {Email} ({Lower})* ({Addr}):contact -->
:person.Person= {rule = "hasEmail"}, :contact.Contact= {rule = "hasEmail"}

Rule: emailFor
({Addr}):contact ({Lower})* {Email} ({Lower})* ({Name}):person -->
:person.Person= {rule = "emailFor"}, :contact.Contact= {rule = "emailFor"}
```

Policy

Policy determines the strategy to be taken to pick a match when more than one span can be matched and when matches might overlap.

OVERVIEW OF JAPE

Phase: p1
Input: Token
Options: control = appelt

Rule: name
({Token.orth==upperInitial})+:mark --> :mark.Name={rule="name"}

Rule: addr
({Token}{Token.string=="@"}{Token}{Token.string =="."}{Token}):mark -->
:mark.Addr={rule="addr"}

Rule: email
({Token.string =="email"}{Token.string =="address"}):mark -->
:mark.Email={rule="email"}

Rule: otherLower
({Token.orth==lowercase}):mark --> :mark.Lower={rule="otherLower"}

Phase: p2
Input: Lower Name Email Addr
Options: control = first

Rule: hasEmail
({Name}):person ({Lower})* {Email} ({Lower})* ({Addr}):contact -->
:person.**Person**= {rule = "hasEmail"}, :contact.**Contact**= {rule = "hasEmail"}

Rule: emailFor
({Addr}):contact ({Lower})* {Email} ({Lower})* ({Name}):person -->
:person.**Person**= {rule = "emailFor"}, :contact.**Contact**= {rule = "emailFor"}

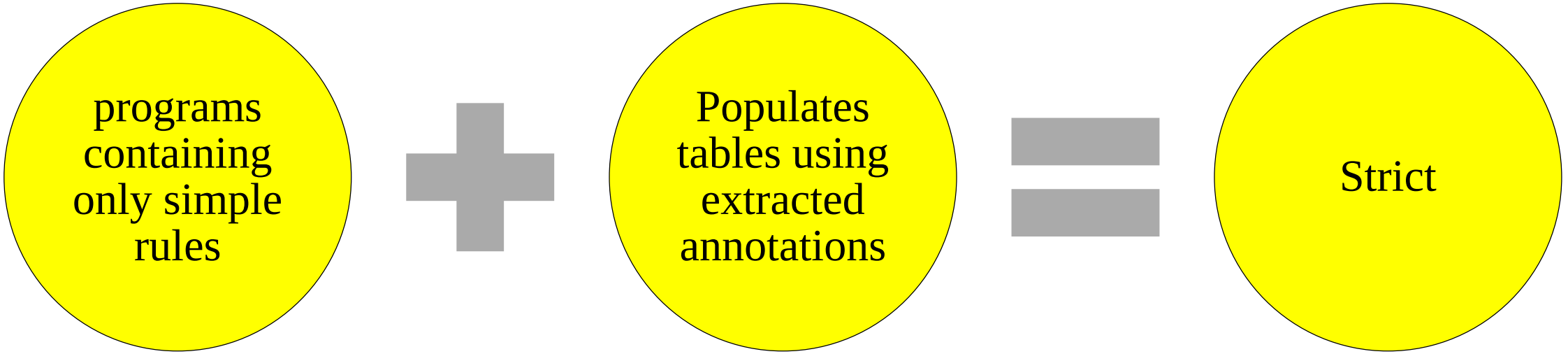
Annotated Text

... Note that John Doe has email address john@hotmail.com...
Person Contact

... jane@gmail.com is the email address for Jane ...
Contact Person

Verification OF JAPE Programs

Strict JAPE Program

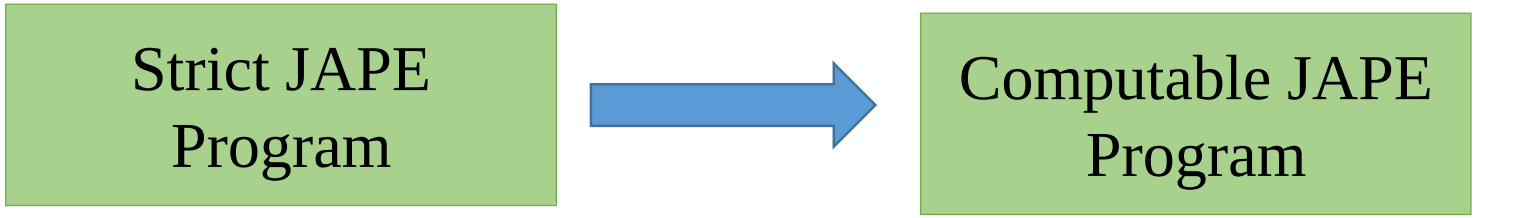


Strict verifier needs to examine each rule to determine whether it is simple.

$$S_1\ (S_2):\alpha\ S_3\ (S_4):\beta\ S_5 \longrightarrow :\alpha.A,\ :\beta.B$$

Computable JAPE Program

- unique binding variables within the scope of the rule
- binding variables provide the start and end offsets of the spans
- the offsets can be used to determine $P_X(D, j)$



$$S_1\ (S_2):\alpha\ S_3\ (S_4):\beta\ S_5 \longrightarrow :\alpha.A,\ :\beta.B$$

Stable JAPE Program

Running a computable program over $g(D, j)$ for all $j \in [1, \ldots, T]$, a stable program extracts a correctly modified record.

Stable JAPE Program

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 JAPE Program

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
50	3

Stable JAPE Program

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 .

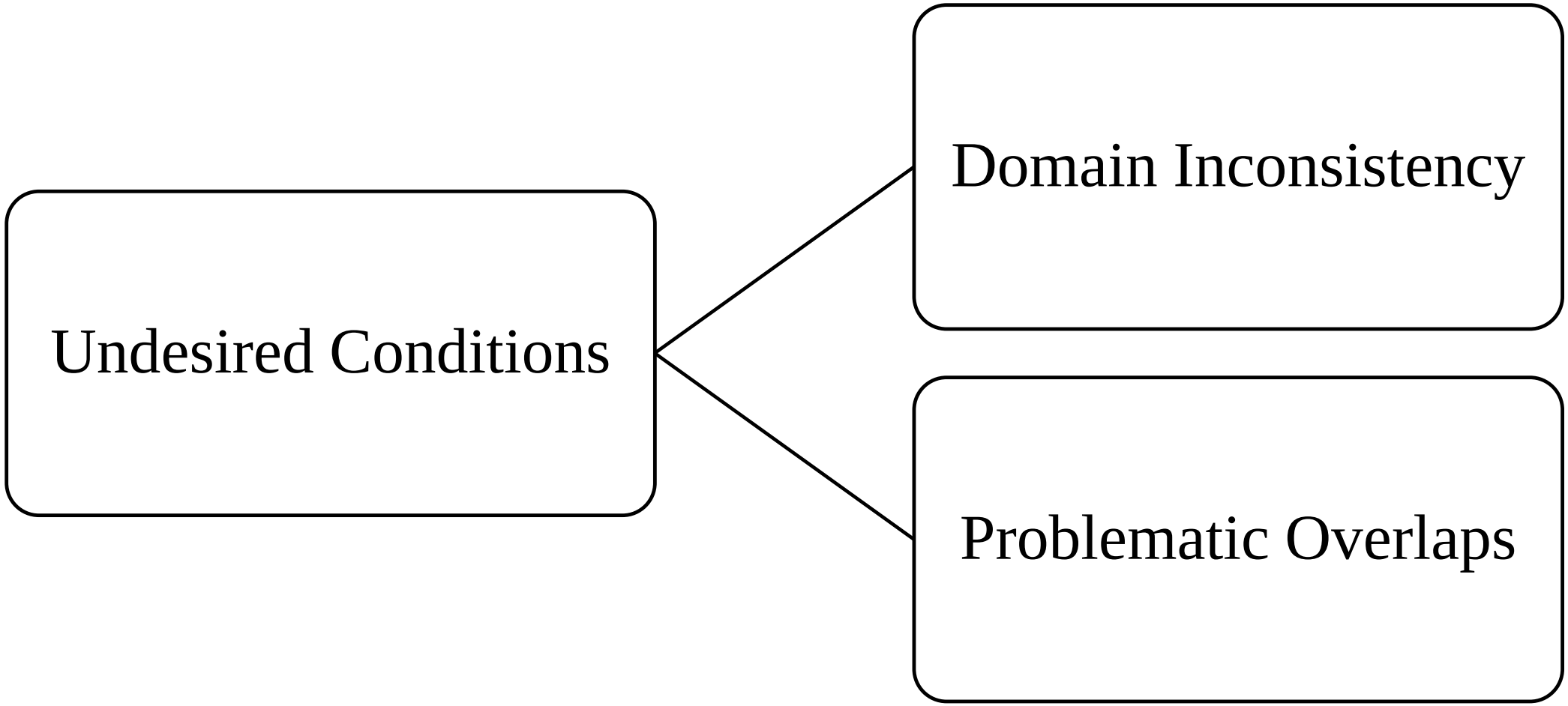
A_1	A_2
35	2 1/2

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 JAPE Program



Domain Inconsistency

```
Phase: p1
Input: Token
Options: control = appelt

Rule: name
({Token.orth==upperInitial})+:mark --> :mark.Name={rule="name"}

Rule: addr
({Token}{Token.string=="@"}{Token}{Token.string =="."}{Token}):mark -->
:mark.Addr={rule="addr"}

Rule: email
({Token.string =="email"}{Token.string =="address"}):mark -->
:mark.Email={rule="email"}

Rule: otherLower
({Token.orth==lowercase}):mark --> :mark.Lower={rule="otherLower"}
```

```
Phase: p2
Input: Lower Name Email Addr
Options: control = first

Rule: hasEmail
({Name}):person ({Lower})* {Email} ({Lower})* ({Addr}):contact -->
:person.Person= {rule = "hasEmail"}, :contact.Contact= {rule = "hasEmail"}

Rule: emailFor
({Addr}):contact ({Lower})* {Email} ({Lower})* ({Name}):person -->
:person.Person= {rule = "emailFor"}, :contact.Contact= {rule = "emailFor"}
```

Domain Inconsistency

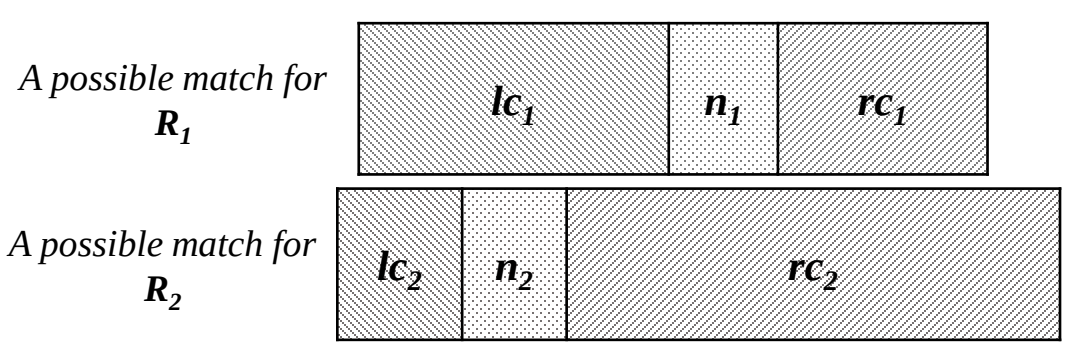
Domain of $fj \in F$ is a subset of the domain formed by the rule corresponding to attribute A_j, for $j \in [1, \ldots, T]$,

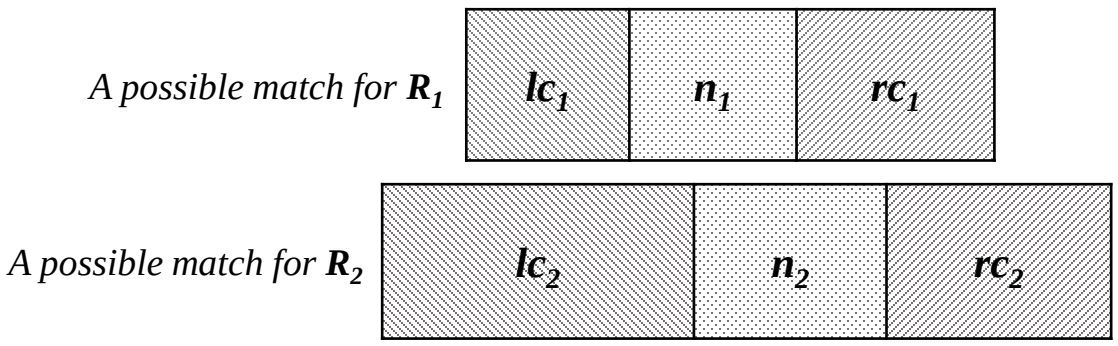
```
Phase: p1
Input: Token
Options: control = appelt

Rule: name
({Token.orth==upperInitial})+:mark --> :mark.Name={rule="name"}

Rule: addr
({Token}{Token.string=="@"}{Token}{Token.string =="."}{Token}):mark -->
:mark.Addr={rule="addr"}

Rule: email
({Token.string =="email"}{Token.string =="address"}):mark -->
:mark.Email={rule="email"}

Rule: otherLower
({Token.orth==lowercase}):mark --> :mark.Lower={rule="otherLower"}
```

```
Phase: p2
Input: Lower Name Email Addr
Options: control = first

Rule: hasEmail
({Name}):person ({Lower})* {Email} ({Lower})* ({Addr}):contact -->
:person.Person= {rule = "hasEmail"}, :contact.Contact= {rule = "hasEmail"}

Rule: emailFor
({Addr}):contact ({Lower})* {Email} ({Lower})* ({Name}):person -->
:person.Person= {rule = "emailFor"}, :contact.Contact= {rule = "emailFor"}
```

Problematic Overlaps

- If a rule's nest can serve as a different part of the same rule's pattern or as a part of some other rule's pattern.

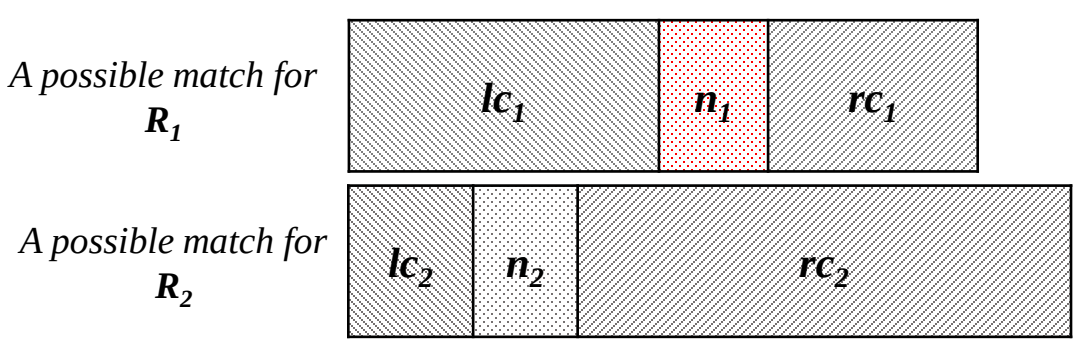


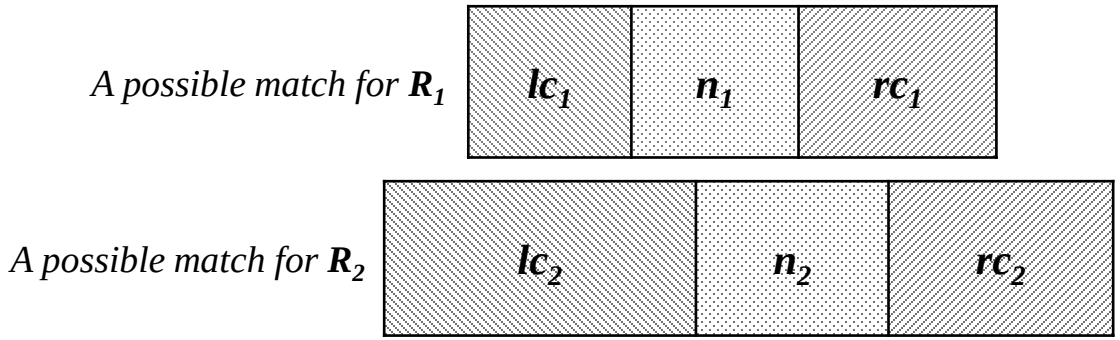


Problematic Overlaps

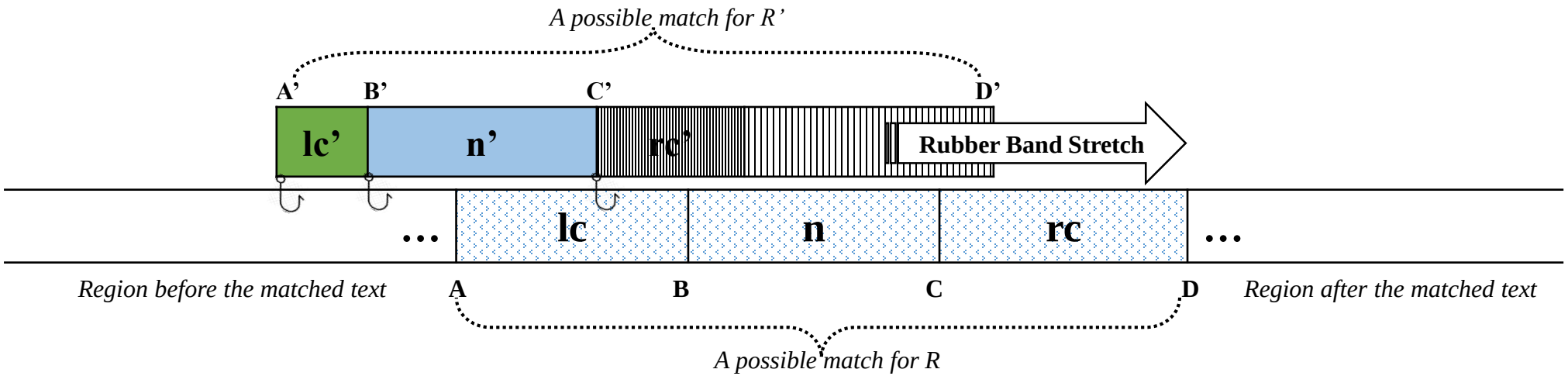
Not Problematic for First policy

Problematic for Appelt policy

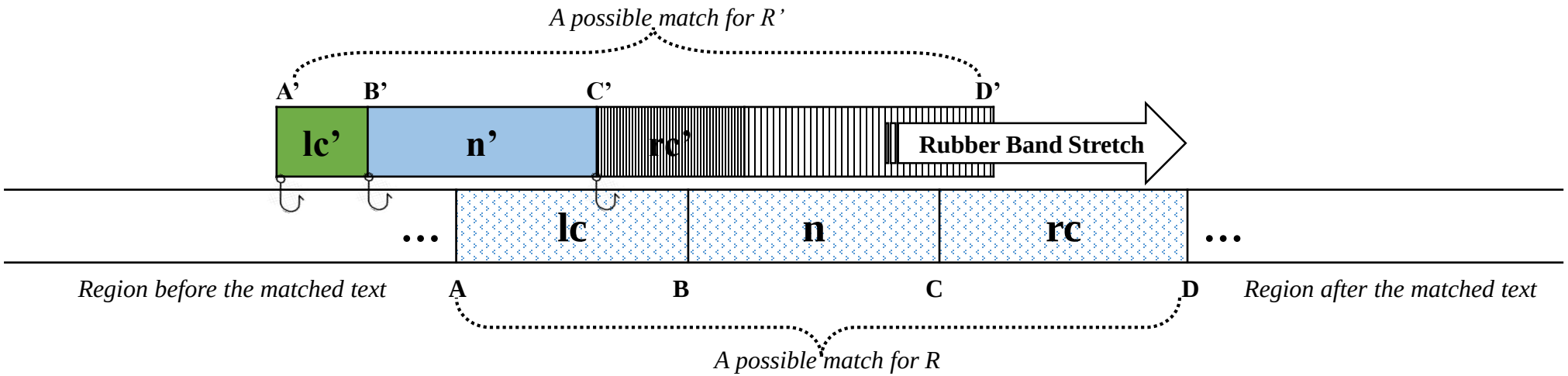


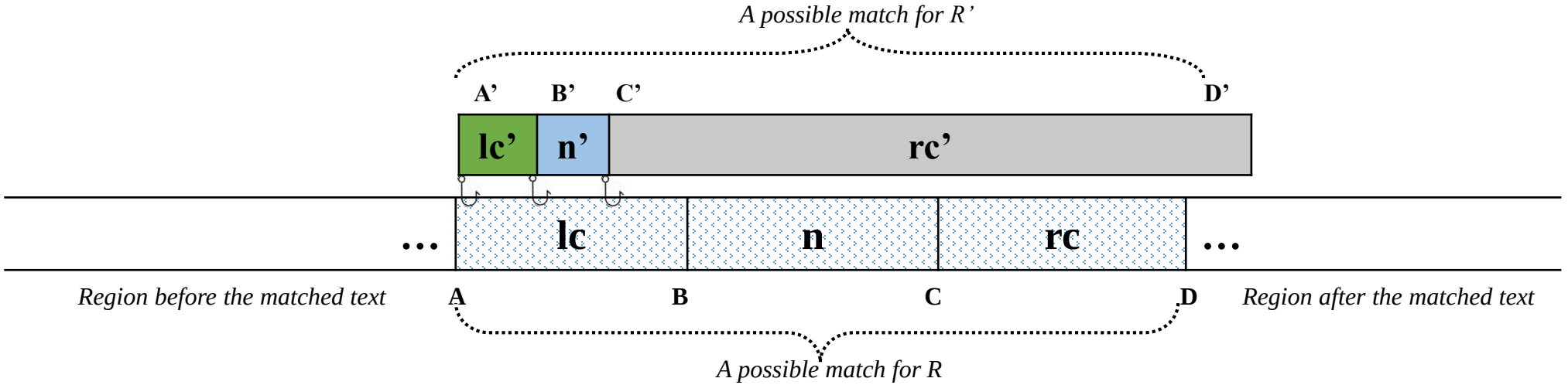


Rubber Band Analysis

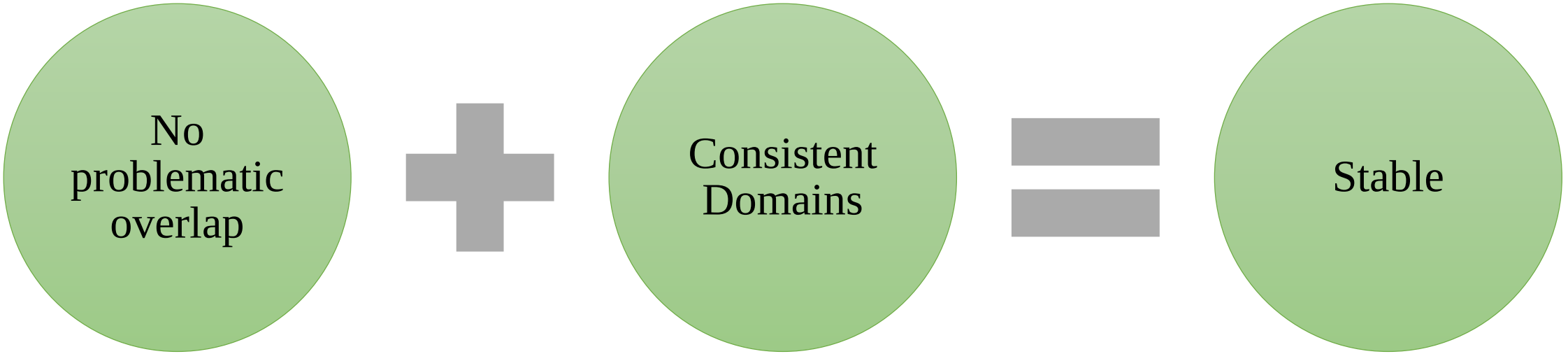


Rubber Band Analysis





Stable JAPE Program



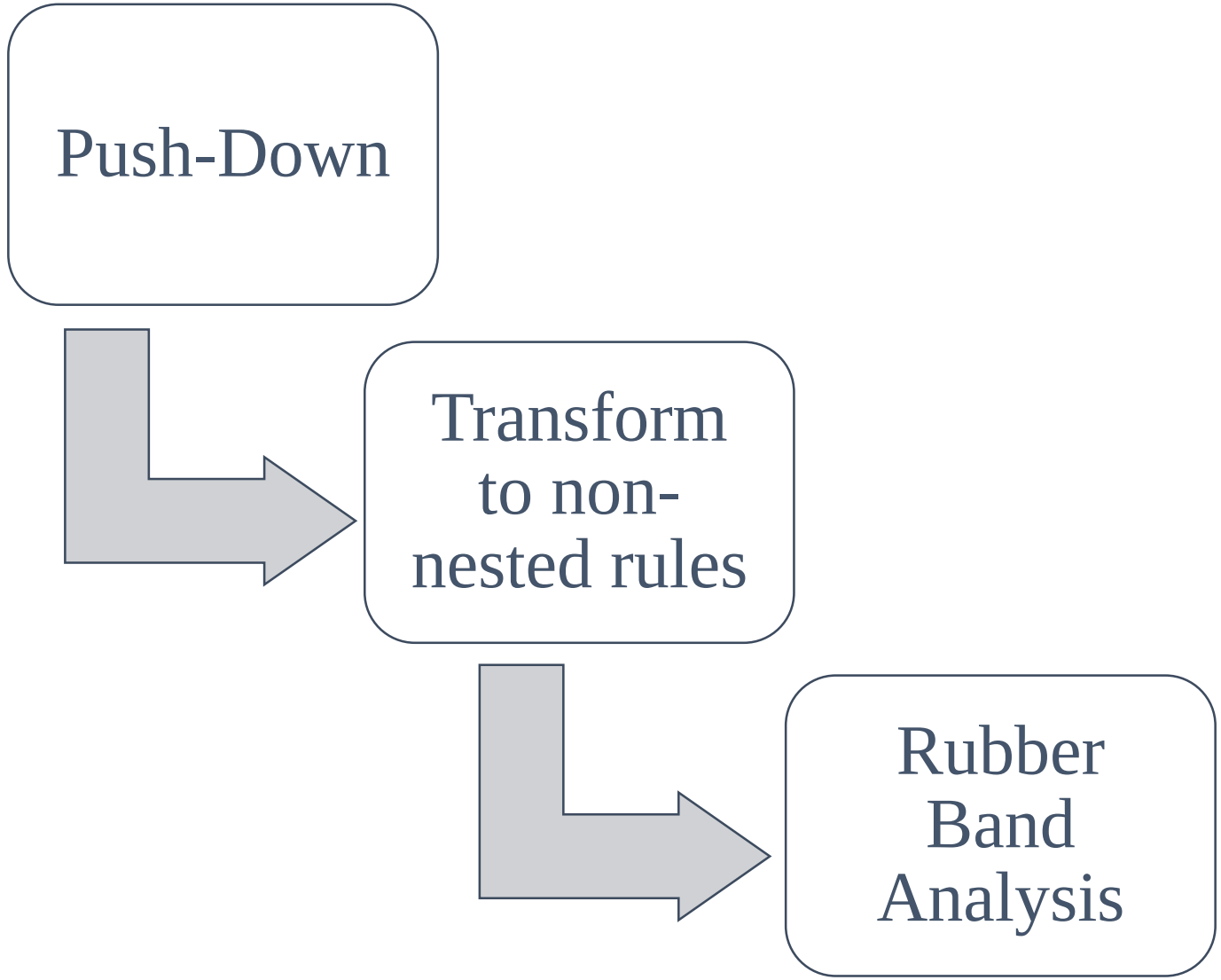
Rubber Band Analysis

- The rubber band analysis accepts a pair of simple un-nested rules.

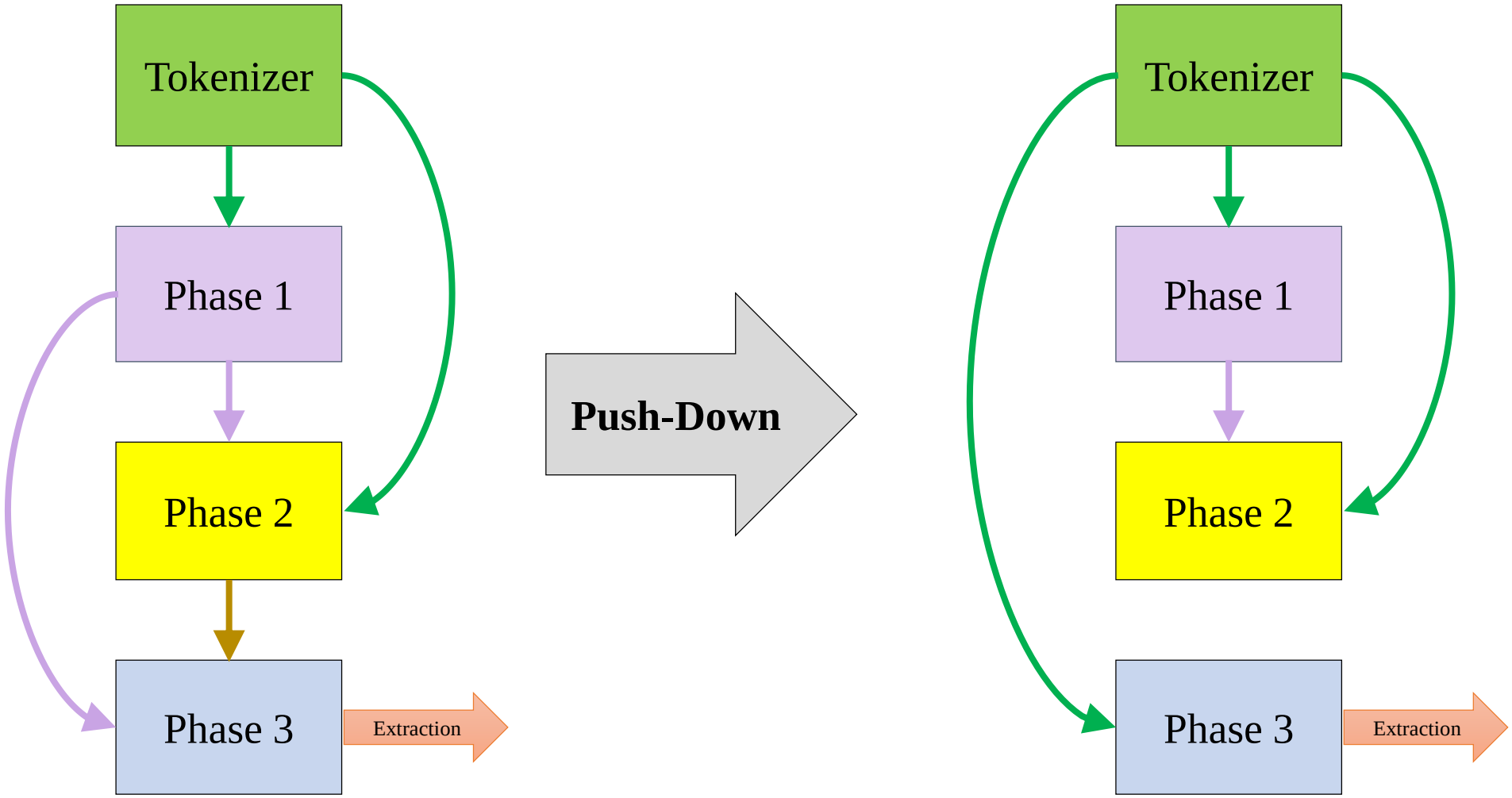
$S_1\ (S_2):\alpha\ S_3\ (S_4\ (S_5\ (S_6\ (S_7):\eta\ S_8):\gamma\ S_9):\theta):\beta \rightarrow$ annotate spans corresponding to $\alpha, \beta, \theta, \gamma$, and η.

- The rubber band analysis take into account the effects of prior rules

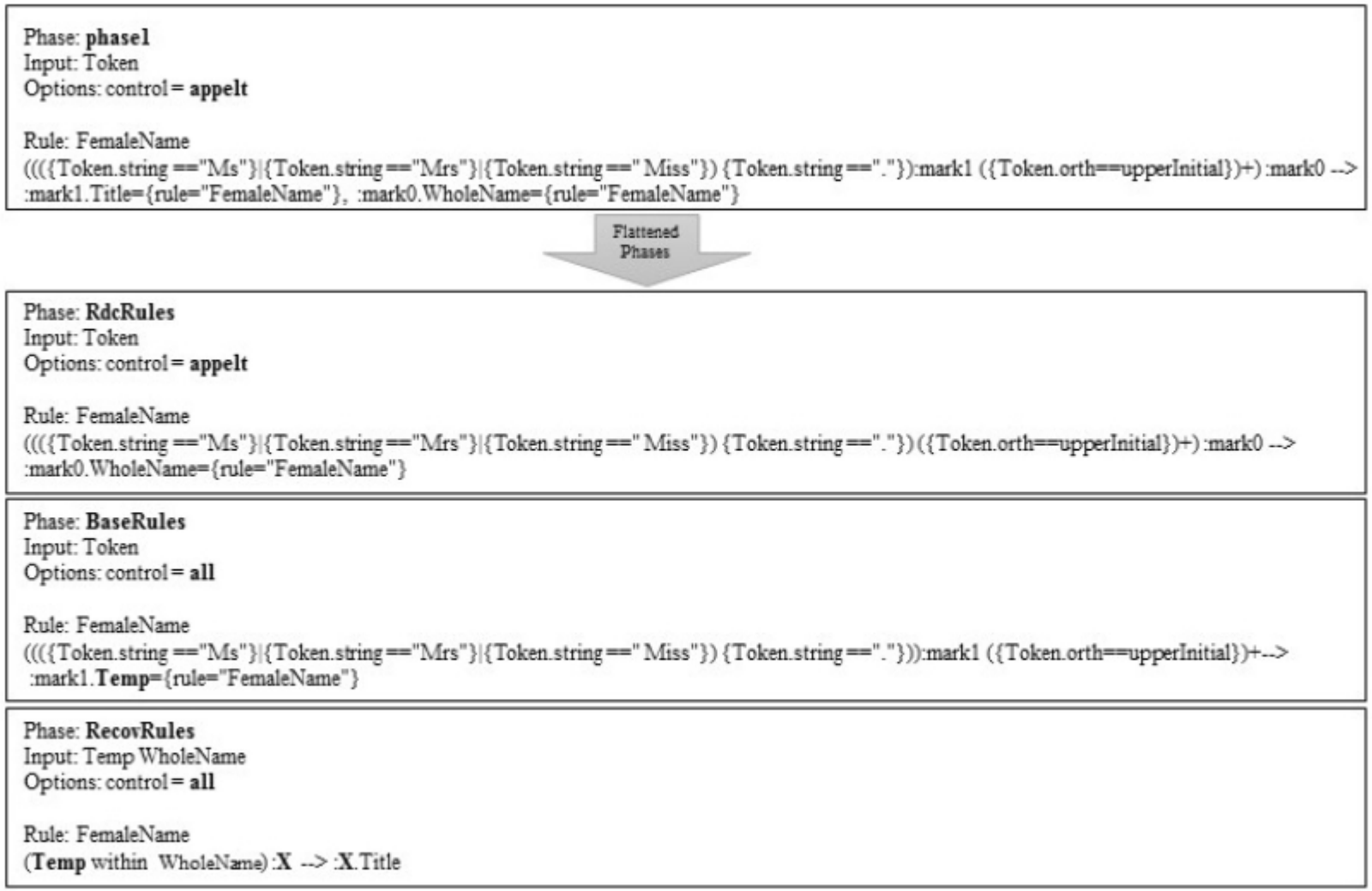
Rubber Band Analysis



Push-Down Algorithm



Transformation Algorithm



Work in Progress

- Exploring necessary properties
- Loosening simplifying assumptions such as independence between extracted attributes
- Developing verification tools for extractors based on machine learning techniques

Thank you!