



APPLYING LOCAL DIFFERENTIAL PRIVACY TO TEXT

Besat Kassaie
December 2017

Optometry Sample Record

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

General Information Y39HE 09 Sep 2015@15:07
Y39HE 09 Sep 2015@15:19

Accompanied By
 Alone

Driver's License (Type | Restrictions)
 C - Normal

Hobbies/Activities/Computer Use
 want to get glasses mostly for reading

Chief Complaint Extended Y39HE 09 Sep 2015@15:07
Y39HE 09 Sep 2015@15:20

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 UW 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 18yo, 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 Y39HE 09 Sep 2015@15:07
Y39HE 09 Sep 2015@15:13

Last Full Eye Exam (Where | When | Outcome)
 PC 2013 May

Previous Injuries/Infections
 No

Previous Eye Surgery
 No

Amblyopia/Strabismus/VT Patching
 No

Other Eye Conditions
 No

Current or previous CL/spectacle wear
 No

Patient General Health Y39HE 09 Sep 2015@15:07
Y39HE 09 Sep 2015@15:16

Last Medical Exam
 Routine two months ago Dr Mclean

Health Conditions / Investigations
 No

Diabetes
 pre-DII controlled with exercise net sure about the blood sugar level

Hypertension | Heart Disease/Stroke
 no

COPD/Asthma | Allergies
 no dust allergy

Tobacco Use
 Never

Current Medications Y39HE 09 Sep 2015@15:07
Y39HE 09 Sep 2015@15:19

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

Other OTC Medication
 Meds for Allergy, low aspirin,

Family History Y39HE 09 Sep 2015@15:07
Y39HE 09 Sep 2015@15:18

Family History Unknown

Glaucoma
 No

Blindness/Visual Impairment
 No

Other Ocular Conditions
 No

Other Health Conditions
 DM older brother

Habitual Rx Y39HE 09 Sep 2015@15:07
Y39HE 09 Sep 2015@15:13

R_x 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 Y39HE 09 Sep 2015@15:07
Y39HE 09 Sep 2015@15:19

Acuity Test Used Distance
 Snellen

Near Method
 40cm

Unaided Y39HE 09 Sep 2015@15:07
Y39HE 09 Sep 2015@15:20

	OD	OS	OU	Nx	OD	OS	OU
Unaided	6/6	6/6	6/6				

Aided Y39HE 09 Sep 2015@15:07
Y39HE 09 Sep 2015@15:19

	OD	OS	OU	Nx	OD	OS	OU
Specs			0.4M		0.4M	0.4M	0.4M

Cover Test Y39HE 09 Sep 2015@15:07
Y39HE 09 Sep 2015@15:22

Cover Test Unilateral Alternating
 Distance Non-Strab Ortho
 Near Non-Strab 4 exo

Broad H Y39HE 09 Sep 2015@15:07
Y39HE 09 Sep 2015@15:22

Method
 Associated

Results
 Unrestricted

Pursuits Full Equal Smooth an Saccades

NPC Y39HE 09 Sep 2015@15:07
Y39HE 09 Sep 2015@15:22

NPC 6/6/6 cm

Pupils Y39HE 09 Sep 2015@15:07
Y39HE 09 Sep 2015@15:22

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
 FERRL & RAPD Negative

Confrontation Fields Y39HE 09 Sep 2015@15:07
Y39HE 09 Sep 2015@15:22

Confrontation Fields
 Full to finger count | Full to finger count

Pupillary Distance Y39HE 09 Sep 2015@15:07
Y39HE 09 Sep 2015@15:22

OU PD (Dx | Nx)
 65 | 60

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

Method Retinoscopy User Y39HE

Subjective	OD	Sphere	Cyl	Axis	Acuity
		-0.25	Sph		6/6
	OS	-0.25	Sph		6/6
	OU				6/6

Add Determination Y39HE 09 Sep 2015@15:07
Y39HE 09 Sep 2015@15:35

Method	Age	WD	cm
Add	Acuity	BMA	WD Range
OD +1.75		Low	30 cm
Add	Acuity	BPA	WD Range
OS +1.75		High	50 cm

Note

Final Add Y39HE 09 Sep 2015@15:07
Y39HE 09 Sep 2015@15:36

WD	cm	Final Add	Acuity
WD 40 cm		OD +1.75	0.4M
		OS +1.75	0.4M
		OU	0.4M

Trial Frame Y39HE 09 Sep 2015@15:07
Y39HE 09 Sep 2015@15:45

Method
 Working Distance
 40cm and 6m Snellen VA chart

Comments
 same Rx as the OTC reading glasses, and px liked it : px also like the clarity in distance with -0.25D OU

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

General Information Y39HE 09 Sep 2015@15:07
 Accompanied By: Alone
 Driver's License (Type | Restrictions): C - Normal
 Hobbies/Activities/Computer Use: [Redacted]

Current Medications Y39HE 09 Sep 2015@15:07
 Other Prescription Medication: ALmotriptan tablets for migraines, Fluoxetine HCL 20mg for anti-depression
 Other OTC Medication: Meds for Allergy, low aspirin

Chief Complaint Extended Y39HE 09 Sep 2015@15:07
Chief Complaint Y39HE 09 Sep 2015@16:20

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 UW 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 18yo, and had migraine and short term memory loss ever after, and taking depression meds.

Pupils Y39HE 09 Sep 2015@15:07
 Direct pupil reflex: 4+ Brisk
 Consensual pupil reflex: 4+ Brisk
 Accommodative Reflex: Present
 Pupil shape: Round
 Relative afferent pupil defect: Negative
 Finding: FERRL & RAPD Negative

Confrontation Fields Y39HE 09 Sep 2015@15:07
 Confrontation Fields: Full to finger count

Pupillary Distance Y39HE 09 Sep 2015@15:07
 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
Subjective OD	-0.25	Sph		6/6
OS	-0.25	Sph		6/6
OU				6/6

Add Determination Y39HE 09 Sep 2015@15:07
 Method Age: WD 40 cm

Add	Acuity	BMA	WD Range
OD +1.75		Low 30	cm
OS +1.75		High 50	cm

Final Add Y39HE 09 Sep 2015@15:07
 WD 40 cm
 Final Add Acuity: OD +1.75 0.4M, OS +1.75 0.4M, OU 0.4M

Amblyopia/Strabismus/VT Patching: No
Other Eye Conditions: No
Current or previous CL/spectacle wear: No

Patient General Health Y39HE 09 Sep 2015@15:07
 Last Medical Exam: Routine two months ago, Dr Mclean

Health Conditions / Investigations: No

Diabetes: pre-DII, controlled with exercise, net sure about the blood sugar level

Hypertension | Heart Disease/Stroke: no

COPD/Asthma | Allergies: no, dust allergy

Tobacco Use: Never

Cover Test Y39HE 09 Sep 2015@15:07
 Cover Test: Unilateral, Alternating
 Distance Non-Strab: [Redacted] Ortho
 Near Non-Strab: [Redacted] 4 exo

Broad H Y39HE 09 Sep 2015@15:07
 Method: Associated
 Results: Unrestricted

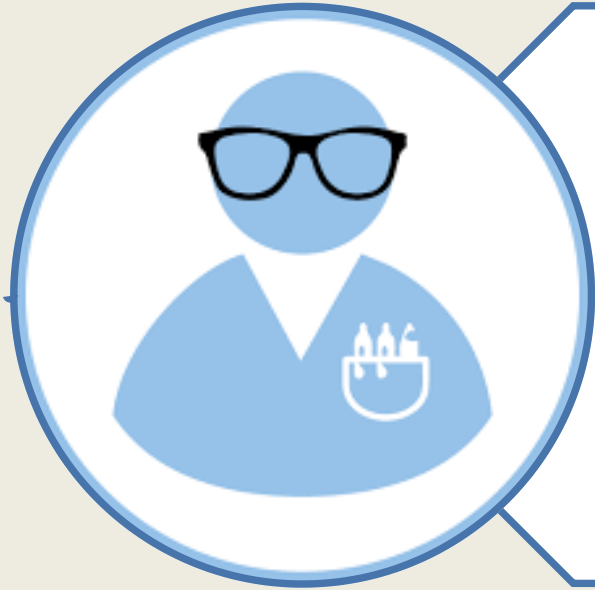
Pursuits Full Equal Smooth an Saccades

NPC Y39HE 09 Sep 2015@15:07
 NPC 6/6 cm

Trial Frame Y39HE 09 Sep 2015@15:07
 Working Distance: 40cm and 6m Snellen VA chart

Comments Y39HE 09 Sep 2015@15:45
 same Rx as the OTC reading glasses, and px liked it; px also like the clarity in distance with -0.25D OU

Electronic Health Record (EHR) For Medical Research



Researchers like to access all the data available in a clinical database to conduct accurate studies



Patients are cautious about their medical data which are considered as private information

Records with consent

Comprehensiveness

Accuracy

Records with no consent

Comprehensiveness

Accuracy

Privacy

Records with consent

Comprehensiveness

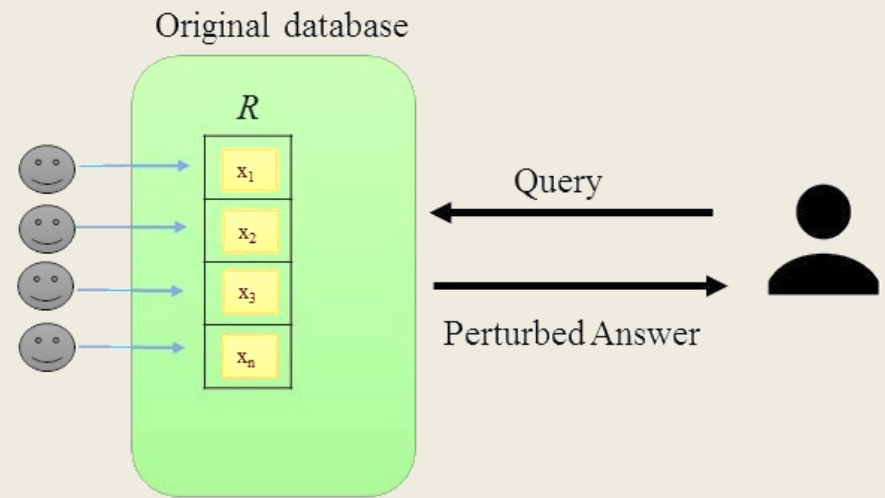
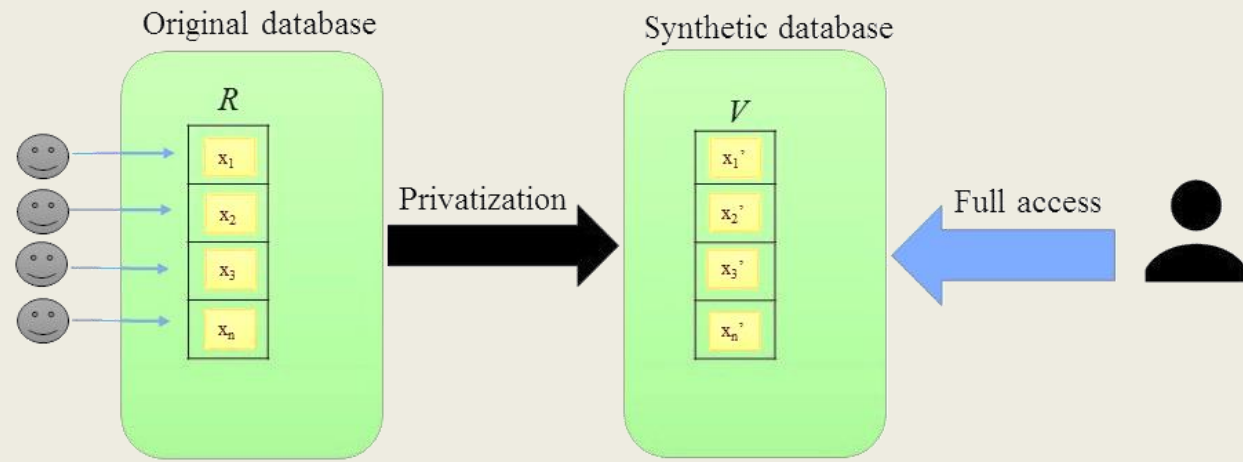
Accuracy

Records with no consent

Comprehensiveness

Accuracy

Privacy



Privacy Preserving Data Access

Anonymization

Removal of personally identifying information

Prone to linkage attack

Anonymization

Queries over Large Sets

Anonymization

Forcing all queries to be over large sets

Anonymization

Prone to differencing attacks

Anonymization

Queries over Large Sets

Query Auditing

Anonymization

Queries over Large Sets

Auditing the sequence of queries and refusing disclosive answers

Anonymization

Queries over Large Sets

Refusing to answer some queries is itself disclosive

Anonymization

Queries over Large Sets

Refusing to answer some queries is itself disclosive

Auditing can be computationally infeasible

Anonymization

Queries over Large Sets

Query Auditing

Differential Privacy

Learning useful
information about
a population



Learning almost
nothing about an
individual

Differential privacy ensures that the same conclusions will be reached independent of whether any individual is **present in** or **absent from** the data set.

Definition 1. *A randomized algorithm M is ϵ -differentially private if for all $S \subset \text{Range}(M)$ and for all $x, y \in \text{domain}(M)$ such that $\|x - y\| \leq 1$: $\Pr[M(x) \in S] \leq \exp(\epsilon)\Pr[M(y) \in S]$.*

Desirable properties of differential privacy

Provides a measure to compute the privacy loss

Immune to post processing

No assumptions about an adversary's background knowledge

Applying Differential Privacy to Optometry Records

DP for medical domain

```
graph TD; A[DP for medical domain] --- B[The importance of data utility in medical domain]; A --- C[Preserving correlations between data items]; A --- D[Finding and justifying the parameter values]; A --- E[Prevalence of unstructured data items];
```

The importance of
data utility in
medical domain

Preserving
correlations
between data items

Finding and
justifying the
parameter values

Prevalence of
unstructured data
items

DP for medical domain

The importance of
data utility in
medical domain

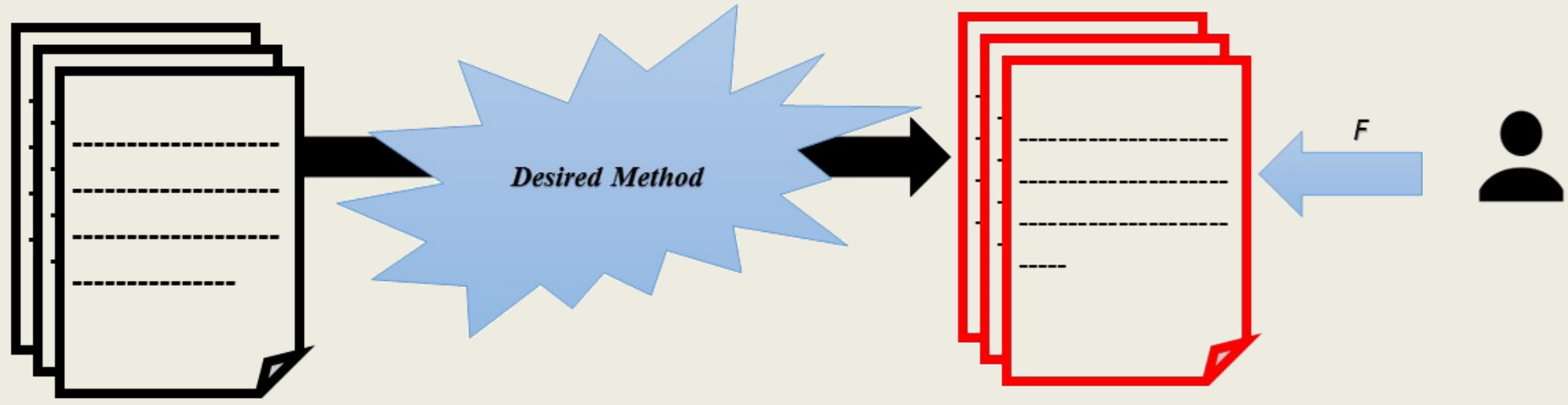
Preserving
correlations
between data items

Finding and
justifying the
parameter values

**Proposing a solution
to apply an
appropriate variant
of DP to text**

The problem statement

Collection of Documents

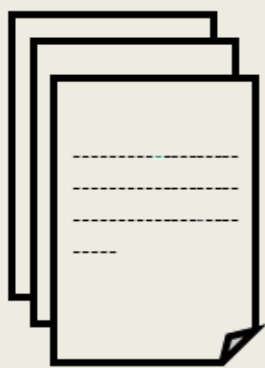


Collection of Privatized Documents

Information Extraction

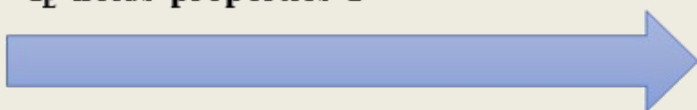
Automatic extraction of structured information such as entities and relationships between entities from unstructured sources.

Collection of Documents



Information Extraction : (I_E)

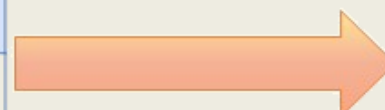
I_E holds properties P



Generated View V

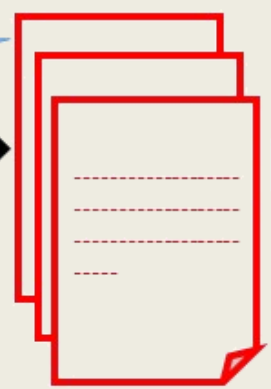
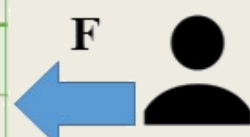
A_1	A_2	A_n
---	--	--
---	--	---

Privatization Process

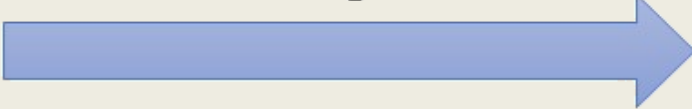


Privatized View V'

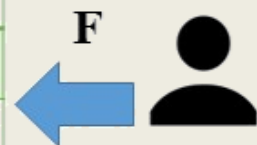
A_1	A_2	A_n
---	--	---
---	---	---



Same I_E

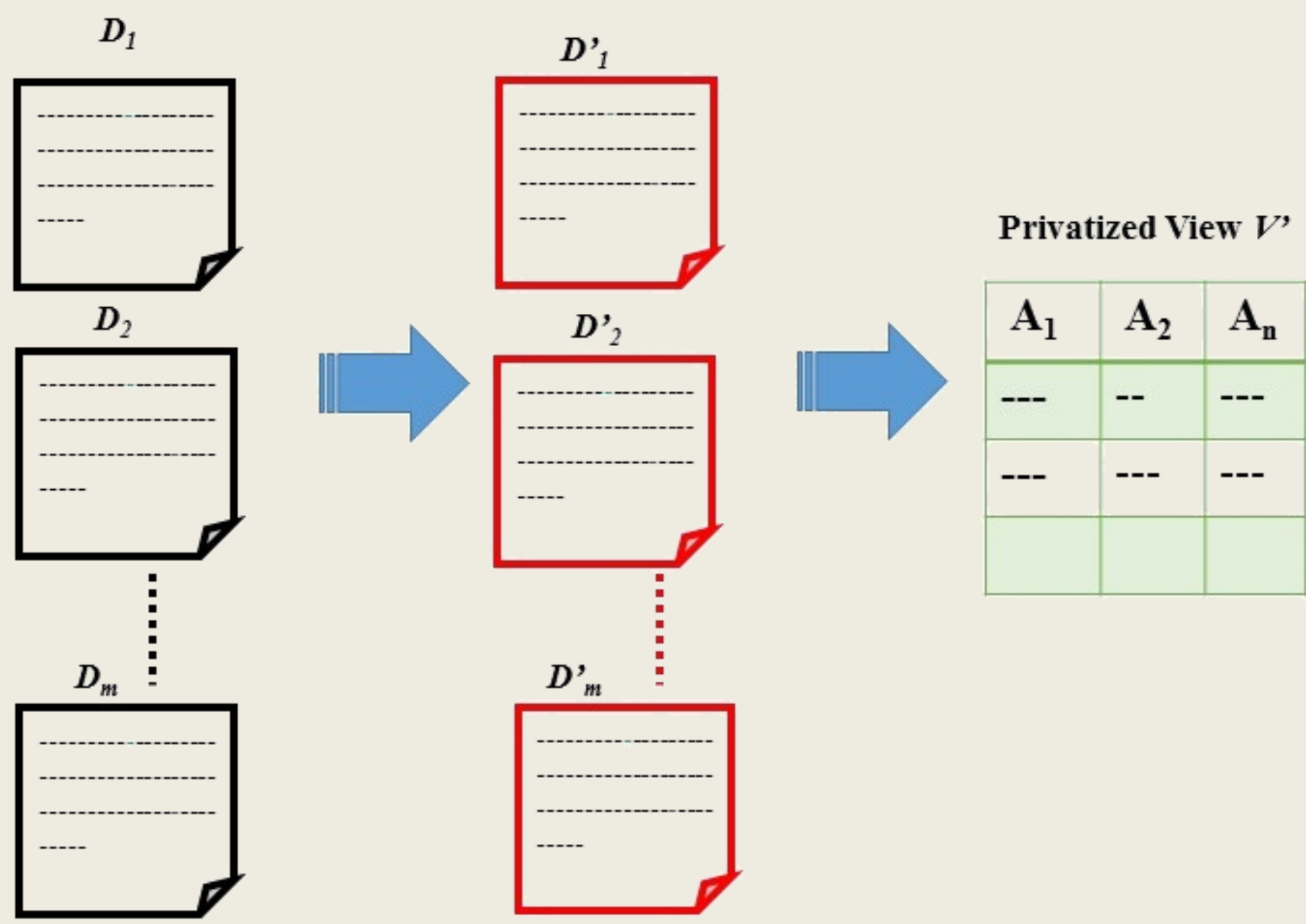


A_1	A_2	A_n
---	--	---
---	---	---



Privatized View V'

Collection of Privatized Documents

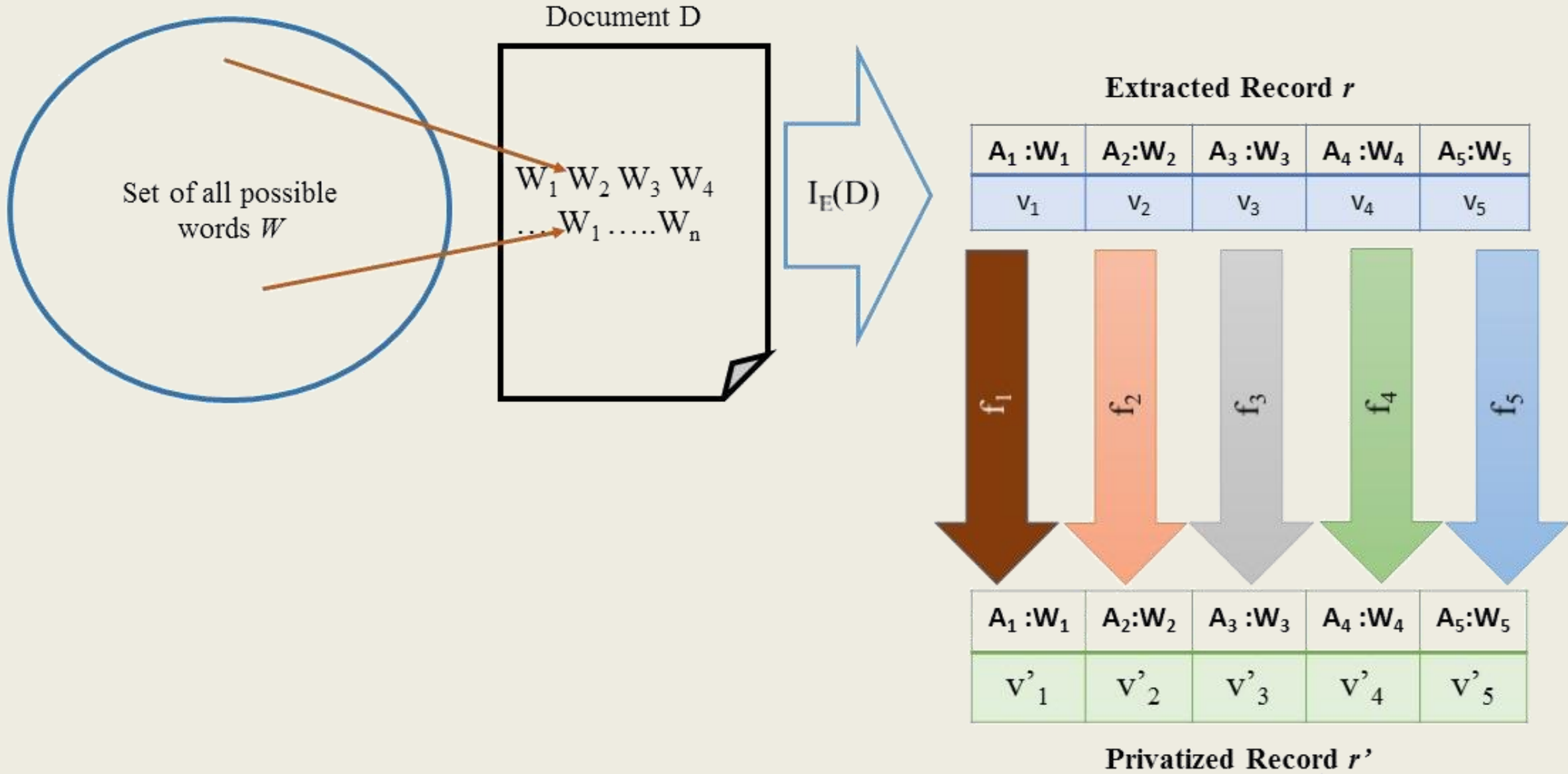


Privatized View V'

A_1	A_2	A_n
---	--	---
---	---	---

Owner > Semi-Trustee Party > Not-Trustee Party

Characteristics of Information Extraction algorithms



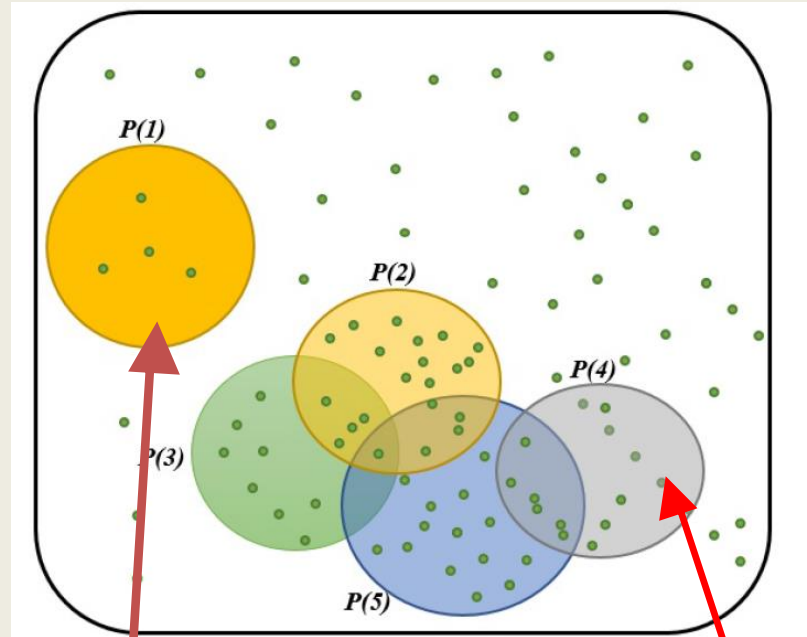
- ***Strict Extractor***

if $\{v_1, v_2, \dots, v_T\} \subseteq \{w_1, w_2, w_3, \dots, w_N\}$.

An Information Extraction algorithm is *Strict Extractor* if the set of extracted values in a record is a subset of words appearing in the corresponding document.

Let $P_D(j) \subseteq \{p | w_p = v_j\}$, i.e., a subset of positions in $D = \langle w_1, w_2, \dots, w_N \rangle$ where $w_p = v_j$ (the position(s) from which v_j is extracted).

Document D



Extracted Record r

$A_1 : W_1$	$A_2 : W_2$	$A_3 : W_3$	$A_4 : W_4$	$A_5 : W_5$
v_1	v_2	v_3	v_4	v_5

- ***Stable Extractor***

Let $g(D, j) = \langle w'_1, w'_2, w'_3, \dots, w'_N \rangle$ where:

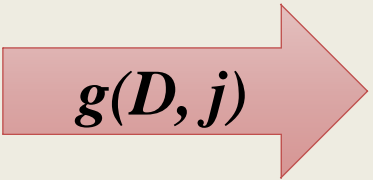
$$w'_k = \begin{cases} f_j(w_k), & \text{if } k \in P_D(j) . \\ w_k, & \text{otherwise.} \end{cases} \quad (2)$$

An Information Extraction algorithm is *Stable* if $\forall j \in [1 \dots \mathcal{T}] P_D(j) = P_{g(D,j)}(j)$ and $I_E(g(D, j)) = r'(j)$.

A *Stable Extractor* satisfies two conditions. First, $P_D(j)$ does not change when running the algorithm over a legitimately generated document, i.e., $g(D, j)$. Second, changing values in appropriate positions in a specific document affects only the expected attribute in the extracted record.

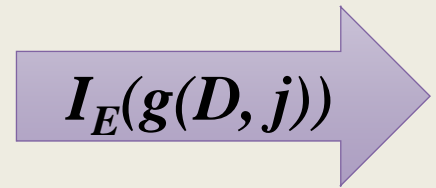
Document D

$W_1 W_2 W_3 W_4$
 $W_5 W_1$
 $W_1 W_{10} W_{20}$
 $W_4 W_1 W_3$



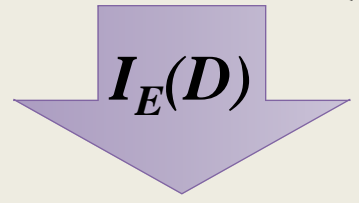
Document D'

$W_1 W_2 W_3 W_4$
 $W_5 f(W_1)$
 $f(W_1) W_{10} W_{20}$
 $W_4 W_1 W_3$



r'

$A_1: W_1$	$A_2: W_2$
$f(W_1)$	W_3



$A_1: W_1$	$A_2: W_2$
W_1	W_3

r

- *Computable Extractor*

$$if \begin{cases} P_D(j) \text{ is explicit(given) for all } j \in [1 \dots \mathcal{T}], \text{ and} \\ P_D(j) \text{ and } P_D(j') \text{ are pairwise disjoint for all } j, j' \in [1 \dots \mathcal{T}]. \end{cases} \quad (3)$$

Theorem. For any function $I_E : \mathcal{D} \rightarrow R$ having the aforementioned properties, there exists an algorithm $A(F, P_D(j))$ such that for an arbitrary set of functions $F = \{f_i | f_i : W_i \rightarrow W_i, i \in [1 \dots \mathcal{T}]\}$ and any document $D \in \mathcal{D}$, $A(F, P_D(j))$ produces $D_F^{\mathcal{P}}$ in such way that, $F(I_E(D)) = I_E(D_F^{\mathcal{P}})$.

Claim. For any function I_E having the aforementioned properties, Algorithm 1 produces $D_F^{\mathcal{P}}$ in such a way that $F(I_E(D)) = I_E(D_F^{\mathcal{P}})$.

```
Input:  $F, \{ P_D(j) \mid j \in [1 \dots \mathcal{T}] \}$   
Output:  $D_F^{\mathcal{P}}$   
for  $j \in [1 \dots \mathcal{T}]$  do  
  | for every  $i$  in  $P_D(j)$  do  
  |   | substitute  $w_i \in D$  with  $f_j(w_i)$   
  |   end  
end  
return  $D_F^{\mathcal{P}}$ 
```

Algorithm 1: PrivateGen

Work still in progress ...



Loosening assumptions:

- The assumption of independence between extracted attributes
- Having access to very limited information about the internal processes of the IE algorithm

Generalizing the proposed properties to cover common Information Extraction algorithms

Designing advanced algorithms to generate D'

Proposing property verification algorithms

Considering more complex relations between documents, individuals, and extracted records

Loosening assumptions:

- The assumption of independence between extracted attributes
- Having access to very limited information about the internal processes of the IE algorithm

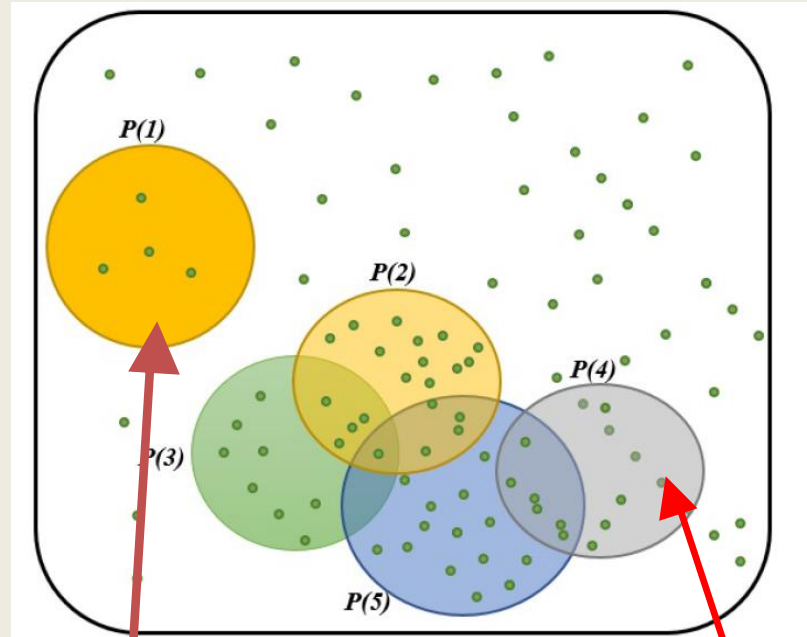
Generalizing the proposed properties to cover common Information Extraction algorithms

Designing advanced algorithms to generate D'

Proposing property verification algorithms

Considering more complex relations between documents, individuals, and extracted records

Document D



Extracted Record r

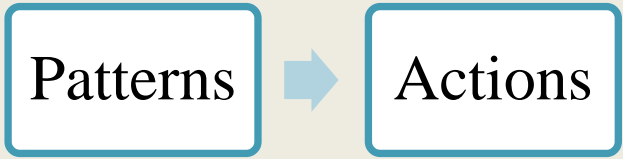
$A_1 : W_1$	$A_2 : W_2$	$A_3 : W_3$	$A_4 : W_4$	$A_5 : W_5$
v_1	v_2	v_3	v_4	v_5

Let $P_D(j) \subseteq \{p|w_p = v_j\}$, i.e., a subset of positions in $D = \langle w_1, w_2, \dots, w_{\mathcal{N}} \rangle$ where $w_p = v_j$. Let $Q_D(j) \subseteq \{p|p \in [1 \dots \mathcal{N}]\}$, i.e., a subset of positions in D . $C(Q_D(j))$ represents constraints for $Q_D(j)$ such as:

$$C(Q_D(j)) = \begin{cases} \text{Domain constraints} \\ \text{Inter-relation such as } q_1(w_i, w_j) = q_2(w_k) \\ \text{Value constraints such as } w_i = q(D) . \end{cases} \quad (5)$$

Rule-Based Information Extraction

Rule Representation



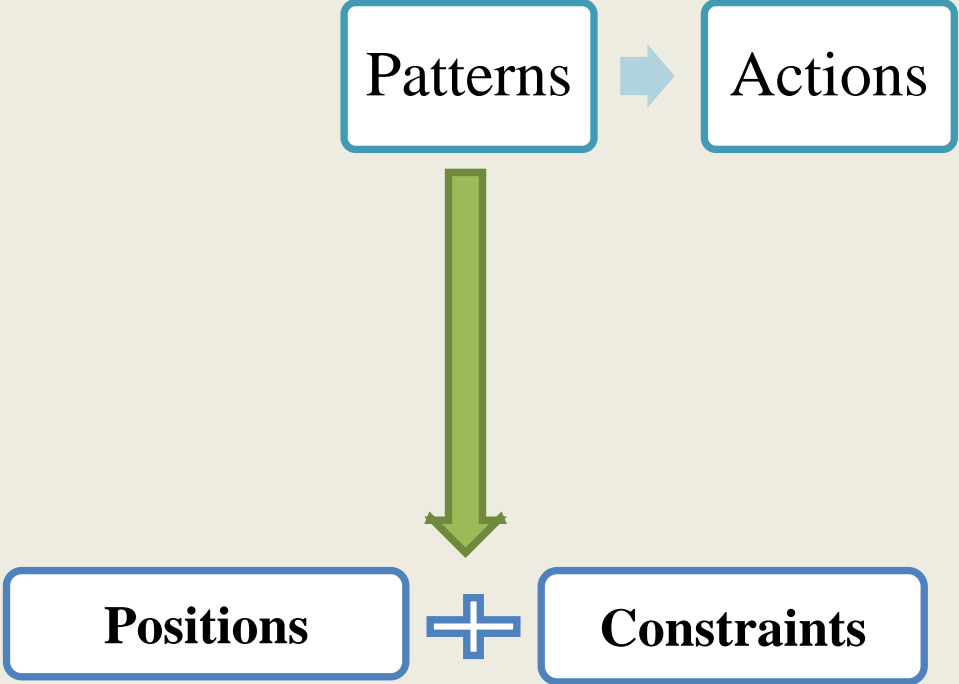
It was not immediately clear whether Trump had given final approval to the latest staff shakeup, but one of the officials said the president asked for the plan to be put together. Tillerson's long-rumored departure would end a troubled tenure for the former Exxon Mobil Corp chief executive who has been increasingly at odds with Trump over policy challenges such as North Korea and under fire for his planned cuts at the State Department. Tillerson was reported in October to have privately called Trump a "moron," something which the secretary of state sought to dismiss.

That followed a tweet by Trump a few days earlier that Tillerson should not waste his time by seeking negotiations with North Korea over its nuclear and missile program. Trump asked John Kelly, the White House chief of staff, to develop the transition strategy and it has already been discussed with other officials, one administration source said.

Type	Set	Start	End	Id	Features
Organization		259	275	673	{matches=[673, 579, 485], orgType=unknown, rule=OrgXEnding, ruleFinal=OrgFinal}
Organization		259	275	579	{matches=[673, 579, 485], orgType=unknown, rule=OrgXEnding, ruleFinal=OrgFinal}
Organization		259	275	485	{matches=[673, 579, 485], orgType=unknown, rule=OrgXEnding, ruleFinal=OrgFinal}
Location		368	379	674	{locType=country, matches=[674, 580, 486, 677, 583, 489], rule=Location1, ruleFinal=LocFinal}
Location		368	379	580	{locType=country, matches=[674, 580, 486, 677, 583, 489], rule=Location1, ruleFinal=LocFinal}
Location		368	379	486	{locType=country, matches=[674, 580, 486, 677, 583, 489], rule=Location1, ruleFinal=LocFinal}
Organization		423	439	675	{matches=[675, 581, 487], orgType=government, rule=GazOrganization, ruleFinal=OrgFinal}
Organization		423	439	581	{matches=[675, 581, 487], orgType=government, rule=GazOrganization, ruleFinal=OrgFinal}
Organization		423	439	487	{matches=[675, 581, 487], orgType=government, rule=GazOrganization, ruleFinal=OrgFinal}
Date		468	475	676	{kind=date, matches=[676, 582, 488], rule=GazDate, ruleFinal=DateOnlyFinal}
Date		468	475	582	{kind=date, matches=[676, 582, 488], rule=GazDate, ruleFinal=DateOnlyFinal}
Date		468	475	488	{kind=date, matches=[676, 582, 488], rule=GazDate, ruleFinal=DateOnlyFinal}
Location		699	710	677	{locType=country, matches=[674, 580, 486, 677, 583, 489], rule=Location1, ruleFinal=LocFinal}
Location		699	710	583	{locType=country, matches=[674, 580, 486, 677, 583, 489], rule=Location1, ruleFinal=LocFinal}
Location		699	710	489	{locType=country, matches=[674, 580, 486, 677, 583, 489], rule=Location1, ruleFinal=LocFinal}
Person		762	772	678	{firstName=John, gender=male, kind=fullName, matches=[678, 584, 490], rule=PersonLocAmbig, ruleFinal=PersonFinal, surname=}
Person		762	772	584	{firstName=John, gender=male, kind=fullName, matches=[678, 584, 490], rule=PersonLocAmbig, ruleFinal=PersonFinal, surname=}
Person		762	772	490	{firstName=John, gender=male, kind=fullName, matches=[678, 584, 490], rule=PersonLocAmbig, ruleFinal=PersonFinal, surname=}
Organization		778	789	679	{matches=[679, 585, 491], orgType=government, rule=GazOrganization, ruleFinal=OrgFinal}
Organization		778	789	585	{matches=[679, 585, 491], orgType=government, rule=GazOrganization, ruleFinal=OrgFinal}
Organization		778	789	491	{matches=[679, 585, 491], orgType=government, rule=GazOrganization, ruleFinal=OrgFinal}

- Date
- Location
- Lookup
- Organization
- Person
- Sentence
- SpaceToken
- Split
- Token
- Unknown

► Original markups



References

Adam, N. R., & Wortmann, J. C. (1989). Security-control methods for statistical databases: A comparative study. ACM Comput. Surv., 21 (4), 515556.

Aggarwal, C. C., & Yu, P. S. (Eds.). (2008). Privacy-preserving data mining – models and algorithms (Vol. 34). Springer.

Dankar, F. K., & Emam, K. E. (2012). The application of differential privacy to health data. In Proceedings of the 2012 joint EDBT/ICDT workshops, berlin, germany, march 30, 2012 (pp. 158166).

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9 (3-4), 211407.

Nabar, S. U., Kenthapadi, K., Mishra, N., & Motwani, R. (2008). A survey of query auditing techniques for data privacy. In Privacy-preserving data mining (pp. 415431). Springer.

Sarawagi, S. (2008). Information extraction. Foundations and Trends in Databases, 1 (3), 261377.

Thanks

